



# DAGSTUHL REPORTS

## Volume 7, Issue 7, Juli 2017

Foundations of Wireless Networking (Dagstuhl Seminar 17271) <i>Christina Fragouli, Magnús M. Halldórsson, Kyle Jamieson, and Bhaskar Krishnamachari</i> .....	1
Citizen Science: Design and Engagement (Dagstuhl Seminar 17272) <i>Irene Celino, Oscar Corcho, Franz Hölker, and Elena Simperl</i> .....	22
Malware Analysis: From Large-Scale Data Triage to Targeted Attack Recognition (Dagstuhl Seminar 17281) <i>Sarah Zennou, Saumya K. Debray, Thomas Dullien, and Arun Lakhotia</i> .....	44
From Observations to Prediction of Movement (Dagstuhl Seminar 17282) <i>Mark Birkin, Somayeh Dodge, Brittany Terese Fasy, and Richard Philip Mann</i> ...	54
Resource Bound Analysis (Dagstuhl Seminar 17291) <i>Marco Gaboardi, Jan Hoffmann, Reinhard Wilhelm, and Florian Zuleger</i> .....	72
Topology, Computation and Data Analysis (Dagstuhl Seminar 17292) <i>Hamish Carr, Michael Kerber, and Bei Wang</i> .....	88
User-Generated Content in Social Media (Dagstuhl Seminar 17301) <i>Tat-Seng Chua, Norbert Fuhr, Gregory Grefenstette, Kalervo Järvelin, and Jaakko Peltonen</i> .....	110

ISSN 2192-5283

*Published online and open access by*

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at <http://www.dagstuhl.de/dagpub/2192-5283>

*Publication date*

March, 2018

*Bibliographic information published by the Deutsche Nationalbibliothek*

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

*License*

This work is licensed under a Creative Commons Attribution 3.0 DE license (CC BY 3.0 DE).



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

*Aims and Scope*

The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.

In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

*Editorial Board*

- Gilles Barthe
- Bernd Becker
- Stephan Diehl
- Hans Hagen
- Reiner Hähnle
- Hannes Hartenstein
- Oliver Kohlbacher
- Stephan Merz
- Bernhard Mitschang
- Bernhard Nebel
- Bernt Schiele
- Albrecht Schmidt
- Raimund Seidel (*Editor-in-Chief*)
- Arjen P. de Vries
- Klaus Wehrle
- Verena Wolf

*Editorial Office*

Michael Wagner (*Managing Editor*)  
Jutka Gasiorowski (*Editorial Assistance*)  
Dagmar Glaser (*Editorial Assistance*)  
Thomas Schillo (*Technical Assistance*)

*Contact*

Schloss Dagstuhl – Leibniz-Zentrum für Informatik  
Dagstuhl Reports, Editorial Office  
Oktavie-Allee, 66687 Wadern, Germany  
[reports@dagstuhl.de](mailto:reports@dagstuhl.de)  
<http://www.dagstuhl.de/dagrep>

Digital Object Identifier: 10.4230/DagRep.7.7.i

# Foundations of Wireless Networking

Edited by

Christina Fragouli<sup>1</sup>, Magnús M. Halldórsson<sup>2</sup>, Kyle Jamieson<sup>3</sup>, and  
Bhaskar Krishnamachari<sup>4</sup>

1 University of California at Los Angeles, US, [christina.fragouli@ucla.edu](mailto:christina.fragouli@ucla.edu)

2 Reykjavik University, IS, [mmh@ru.is](mailto:mmh@ru.is)

3 Princeton University, US & University College London, GB,  
[kylej@cs.princeton.edu](mailto:kylej@cs.princeton.edu)

4 USC - Los Angeles, US, [bkrishna@usc.edu](mailto:bkrishna@usc.edu)

---

## Abstract

This report documents the talks and discussions of Dagstuhl Seminar 17271 “Foundations of Wireless Networking”. The presented talks represent a wide spectrum of work on wireless networks.

**Seminar** July 2–7, 2017 – <http://www.dagstuhl.de/17271>

**1998 ACM Subject Classification** C.2.1 Network Architecture and Design, wireless communication

**Keywords and phrases** wireless networks

**Digital Object Identifier** 10.4230/DagRep.7.7.1

## 1 Executive Summary

*Christina Fragouli*

*Magnús M. Halldórsson*

*Kyle Jamieson*

*Bhaskar Krishnamachari*

**License** © Creative Commons BY 3.0 Unported license

© Christina Fragouli, Magnús M. Halldórsson, Kyle Jamieson, and Bhaskar Krishnamachari

Wireless communication has grown by leaps and bounds in recent decades, with huge social and societal impact. This is nowhere near saturation, especially with the coming Internet-of-Things on the horizon.

Underlying this technology are fundamental questions: how to efficiently configure and adapt communication links, organize access to the medium, overcome interference, and disseminate information. Unfortunately, the wireless medium is tricky, and the modeling of signal propagation and interference has proved to be highly involved, with additional challenges introduced by issues such as mobility, energy limitations, and device heterogeneity. New technologies such as cooperative MIMO, directional antennas, interference alignment, network coding, energy harvesting and motion control add another layer of complexity, and are yet to be well-understood. Deriving good algorithms and protocols is therefore a non-trivial task.



Except where otherwise noted, content of this report is licensed  
under a Creative Commons BY 3.0 Unported license

Foundations of Wireless Networking, *Dagstuhl Reports*, Vol. 7, Issue 7, pp. 1–21

Editors: Christina Fragouli, Magnús M. Halldórsson, Kyle Jamieson, and Bhaskar Krishnamachari



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## Communities

Different schools of thought have arisen to tackle these fundamental questions. These come from different backgrounds, involving different types of mathematical tools and different approaches and outlooks. It is not just a theory vs. practice split, but also splits within the theory and the practice camps. In addition to the more established information theory, there has been quite some work on network control theory, and also a budding algorithmic theory; on top of physical-layer hardware experimentation, we see also networking systems research and simulation studies.

We identified the following communities, which are not all exclusive:

**Information Theory** Characterized by an interest in fundamental information-theoretic capacity bounds; novel communication paradigms such as MIMO, network coding, interference cancellation, interference alignment; estimation and detection under known stochastic models of noise.

*Prime publication venues:* IEEE Trans. on Information Theory; IEEE International Symposium on Information Theory (ISIT).

**Algorithm Theory** Characterized by a focus on algorithms, their complexity and effectiveness, with emphasis on rigorous proofs and typically worst-case analysis.

*Prime publication venues:* PODC, DISC, STOC, SODA, ICALP (Track C)

**Experimental Mobile and Wireless Systems** Characterized by the design, implementation, and evaluation of practical wireless systems in real testbeds and real-world applications that both evaluate the efficacy of previously-known techniques and their combinations “in the wild” and add design insight by developing novel heuristic algorithms and architectures that are shown to perform well in practice.

*Prime publication venues:* MobiCom, MobiHoc, SIGCOMM, NSDI, SenSys, IPSN, BuildSys, MobiSys.

**Wireless Network Control and Optimization Theory** Characterized by formulation of various problems in wireless networks (typically focusing on medium access, network routing and flow rate control) as continuous control and optimization problems from both deterministic and stochastic perspectives; network utility optimization via primal-dual decomposition, game theory, stochastic decision theory, stochastic multi-armed bandits.

*Prime publication venues:* Infocom, SIGMETRICS, WiOpt, CDC, IEEE Trans. on Networking, IEEE Transactions on Automatic Control.

**Physical Layer and Hardware Design** Characterized by the design, implementation, and evaluation of new hardware, signal processing techniques. The theoretical members of this community have a lot of overlap with information theory, while the more experimental members of this community have a lot of overlap with the experimental wireless and mobile systems community in terms of the problems they consider and their solution approaches.

*Prime publication venues:* IEEE Trans. on Wireless, IEEE Trans. on Comms., IEEE Globecom, IEEE Vehicular Technology Conf.

## Goals of the Seminar

The goal of this Dagstuhl seminar is to bring together top researchers from the different wireless research communities to review and discuss models and methods in order to obtain a better understanding of the capabilities and limitations of modern wireless networks, and



to come up with more realistic models and new algorithm and protocol design approaches for future wireless networks that may then be investigated in joint research projects.

An important part of the workshop is to actively promote a dialog between different communities. As a result, we seek researchers that are by nature open to different perspectives and have enough self-confidence to welcome research of a different nature. The objective was for each participant to consciously reflect on the implicit values, identity, shared understandings and skill set that people in his/her community expect, and to articulate these issues to others in order to identify and appreciate commonalities and differences, and the potential gains from forming new bridges.

## Seminar Operation

The seminar had varied forms of activities during its operation. As people came from different communities, a major objective was to get to know each other.

**“Speak-and-spark” presentations** All the participants gave a brief, 5-10 minute pitch talk on a problem that they have been (or would to be) working on. Most of these were given on the first day, which provided a way of introducing one another, as well as a way to spark discussions that could be continued in private, in small groups, or in plenum.

**Survey presentations** Seven senior researchers were asked to give a one-hour survey talk on a topic of current interest. These were spread over the days excluding the first.

**Breakout sessions** Several topical issues were identified as particularly suitable for group discussions. The participants voted on the topics of their interest, after which three were selected. The groups were chosen so as to feature representatives from the different communities. Two such rounds of breakout sessions were organized on Tuesday. The discussions were summarized by the group leaders (often with the help of the scribes) for the whole audience on Wednesday morning, followed by discussions.

**“Important paper” pitches** The participants were encouraged to identify research paper(s) that open “new” research areas and/or pose questions in their community. This was also a means to articulating what researchers in that subfield found essential or influential. These were presented in 5-10 minutes, followed by open questions.

**Plenary discussions** Part of the last day’s morning was allocated to general discussions around the themes posed during the seminar, with the aim of identifying future problem directions and research areas.

Abstracts and summaries of these talks and discussions are given in the following sections.

## 2 Table of Contents

### Executive Summary

<i>Christina Fragouli, Magnús M. Halldórsson, Kyle Jamieson, and Bhaskar Krishnamachari . . . . .</i>	1
---	---

### Survey Talks

Breaking some communications paradigms: Spectrally efficient non-orthogonal signals concepts and practical implementation <i>Izzat Darwazeh . . . . .</i>	6
Developing Robust Wireless Network Algorithms (from a TCS perspective) <i>Fabian Daniel Kuhn . . . . .</i>	6
PHY Layer Design For Extreme Resource Sharing <i>Konstantinos Nikitopoulos . . . . .</i>	7
Cooperation in Wireless Networks – An Information-Theoretic Viewpoint <i>Michelle Effros . . . . .</i>	8
Some Topics and Problems in Wireless Networking <i>Muriel Médard . . . . .</i>	8
Survey on “Economics in Wireless” <i>Vijay Subramanian . . . . .</i>	8
Some interesting optimization problem formulations <i>Michele Zorzi . . . . .</i>	9

### Summaries of “Speak-and-Spark” Talks

Massive MIMO and Compressed Sensing: a marriage made in 5G <i>Giuseppe Caire . . . . .</i>	9
Simplicity and Robustness (or all about noisy beeping) <i>Seth Gilbert . . . . .</i>	10
Spatial Outage Capacity: The Missing Link between Average and Worst-Case Performance? <i>Martin Haenggi . . . . .</i>	10
Intelligence and Network <i>Longbo Huang . . . . .</i>	10
Fault-Tolerant Online Packet Scheduling on (Wireless) Channel(s) <i>Tomasz Jurdzinski . . . . .</i>	11
Some Thoughts on Wireless Networking Research <i>Holger Karl . . . . .</i>	11
Routing in Hybrid Networks with Holes <i>Christina Kolb . . . . .</i>	12
Enabling Extreme Resource Sharing in Future Wireless Communication Systems <i>Konstantinos Nikitopoulos . . . . .</i>	13
Towards a widely applicable SINR model for access sharing <i>Christian Scheideler . . . . .</i>	13

Wake-up Transceivers, Convergecast, Beamforming, and Beat Frequencies <i>Christian Schindelhauer</i> . . . . .	13
Distributed optimization for scheduling in wireless networks <i>Vijay Subramanian</i> . . . . .	14
Fine-grain Time / Spectrum Sharing <i>Patrick Thiran</i> . . . . .	14
Wireless Link Capacity under Shadowing and Fading <i>Tigran Tonoyan</i> . . . . .	15
Information Theoretic Caching <i>Daniela Tuninetti</i> . . . . .	15
Update or Wait: How to Get the Freshest and Most Accurate Data Through a Network <i>Elif Uysal-Biyikoglu</i> . . . . .	15
Storage Cost of Shared Memory Emulation <i>Zhiying Wang</i> . . . . .	16
PHY in Networks <i>Roger Wattenhofer</i> . . . . .	16
Problems in millimeter-wave communication: the difficult path from Gigabit links to Gigabit networks <i>Jörg Widmer</i> . . . . .	17
Wireless Networks: Dynamicity <i>Dongxiao Yu</i> . . . . .	17
Passive Communication with Ambient Light <i>Marco Zuniga</i> . . . . .	18
<b>Discussion groups</b>	
Discussion Group: Coding for New Channels <i>Michelle Effros, Tomasz Jurdzinski, Konstantinos Nikitopoulos, Patrick Thiran, Jörg Widmer, and Marco Antonio Zúñiga Zamalloa</i> . . . . .	18
Discussion Group: Interference and Scheduling <i>Patrick Thiran</i> . . . . .	19
<b>Participants</b> . . . . .	21

### 3 Survey Talks

#### 3.1 Breaking some communications paradigms: Spectrally efficient non-orthogonal signals concepts and practical implementation

*Izzat Darwazeh (University College London, GB)*

**License** © Creative Commons BY 3.0 Unported license

© Izzat Darwazeh

**Joint work of** Izzat Darwazeh, Safa Issam, Ioannis Kanaras, Ryan Grammenos, Tongyang Xu, John Mitchell, Spiros Mikroulis, Zhaohui Li, Waseem Ozan, Hedaia Ghannam, Kyle Jamieson

**Main reference** Tongyang Xu, Izzat Darwazeh: “Transmission Experiment of Bandwidth Compressed Carrier Aggregation in a Realistic Fading Channel”, IEEE Trans. Vehicular Technology, Vol. 66(5), pp. 4087–4097, 2017.

**URL** <http://dx.doi.org/10.1109/TVT.2016.2607523>

The talk details work done at University College London (UCL) and elsewhere on the general concepts, fundamentals and experimental validations of bandwidth compressed multicarrier waveforms for future wireless and wired systems. The proposed waveforms are derived from an existing non-orthogonal multicarrier concept termed spectrally efficient frequency division multiplexing (SEFDM) where sub-carriers are non-orthogonally packed at frequencies below the symbol rate. This improves the spectral efficiency at the cost of self-created inter carrier interference (ICI). In this presentation experiments are reported and testing is carried out in three scenarios including long term evolution (LTE)-like wireless link; millimeter wave radio-over-fiber (RoF) link and optical fiber links. In the first scenario, for a given 25 MHz bandwidth, the SEFDM testbed can provide 70 Mbit/s gross data rate while only 50 Mbit/s can be achieved for an OFDM system occupying the same bandwidth. For the millimeter wave experiment, occupying a 1.125 GHz bandwidth, the gross bit rate for OFDM is 2.25 Gbit/s and with 40% bandwidth compression, 3.75 Gbit/s can be achieved for SEFDM. Two experimental optical fiber links are described in this work; a 10 Gbit/s direct detection optical SEFDM system and a 24 Gbit/s coherent detection SEFDM system. The LTE-like signals and millimeter wave technologies are well suited to provide last mile communications to end users as both can support mobility in wireless environments. The lightwave signals delivered by optical fibers would offer higher data rates and support long-haul communications. The reported techniques, used individually or combined, would be of interest to future wireless system designers, where bandwidth saving is of importance, such as in 5G networks, aiming to provide high capacity and high mobility, simultaneously while saving spectrum. New signals, derived from SEFDM are discussed and new applications, including x-DSL like scenarios are outlined.

#### 3.2 Developing Robust Wireless Network Algorithms (from a TCS perspective)

*Fabian Daniel Kuhn (Universität Freiburg, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Fabian Daniel Kuhn

Over the last 30 years, we have seen an intensive effort to develop distributed algorithms and abstract models to deal with the characteristic properties of wireless communication. The models range from simple graph-based characterizations of interference to more accurate physical models such as the so-called signal-to-noise-and-interference (SINR) model.

As different as the typically considered models may be, most of them have one thing in common. Whether a node can successfully receive (and decode) a message is determined using some fixed, deterministic rule that depends on the structure of the network and some additional model parameters.

While in classical wired networks, assuming reliable communication might be a reasonable abstraction, this seems much more problematic in a wireless network setting. The propagation of a wireless signal depends on many diverse environmental factors and it does not seem to be realistic to explicitly model all of these factors or to exactly measure the properties of the wireless communication channels. In addition, the environmental factors might change over time and there can also be additional independent sources of signal interference that cannot be predicted or controlled by the network. Further, wireless devices might also be mobile so that we not only have unreliable communication channels, but potentially even almost arbitrary dynamically changing network topologies. Because the classic abstract wireless communication models do not capture such unpredictable behavior, many existing radio network algorithms might only work in the idealized formal setting for which they were developed.

In my talk, I describe ways to develop more robust wireless network algorithms. I will in particular show that complicated, unstable, and unreliable behavior of wireless communication can be modeled by adding a non-deterministic component to existing radio network models. As a result, any behavior which is too complex or impossible to predict is determined by an adversary. Clearly, such models lead to less efficient algorithms. However, they also lead to more robust algorithms which tend to work under a much wider set of underlying assumptions. Very often, such models also lead to much simpler algorithms. I will discuss several existing results and I will sketch some general ideas and possible directions for dealing with adversarial uncertainty and more generally dynamic wireless networks.

### 3.3 PHY Layer Design For Extreme Resource Sharing

*Konstantinos Nikitopoulos (University of Surrey, GB)*


**License** © Creative Commons BY 3.0 Unported license  
© Konstantinos Nikitopoulos

**Joint work of** Konstantinos Nikitopoulos, Georgios Georgis, Christopher Husmann, F. Mehran, C. Jaywardena, Kyle Jamieson, H. Jafarkhani, and R. Tafazolli

Future local area wireless communication systems shall be able to support very high peak user and network rates as well as very large numbers of connected devices, while keeping the latency requirements at very low levels. These needs have triggered a paradigm shift from orthogonal to non-orthogonal signal transmissions that enable extreme sharing of the available resources, including frequency and time. Such schemes include multi-antenna (MIMO) deployments for aggressive spatial multiplexing, as well as non-orthogonal multiple access schemes. In practice, to deliver the theoretical gains of such large non-orthogonal schemes, the mutually interfering information streams must be efficiently demultiplexed. However, the processing requirements of ML detection increase exponentially with the number of mutually interfering information streams, exceeding the processing capabilities of traditional processors. In this context we show that FlexCore; the first method to massively parallelize the ML detection problem in a nearly-independent manner, can overcome the current processing speed barriers. In addition, we show that new PHY approaches like out newly proposed Space-Time Super-Modulation can be efficiently used in the context of machine-type communications to enable joint medium access and rateless data transmission while reducing or even eliminating the need for transmitting preamble sequences.

### 3.4 Cooperation in Wireless Networks – An Information-Theoretic Viewpoint

*Michelle Effros (CalTech – Pasadena, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Michelle Effros

We consider the cost and benefit of cooperation in wireless communication networks. Here, cost is measured by the capacity added to the network to enable the cooperation, while benefit is measured by the amount that the network capacity increases due to the optimal cooperative strategy enabled by that addition. For wireline networks, whether the benefit can exceed the cost remains an open problem. For wireless networks, we show that benefit can exceed any polynomial function of the cost, that the cost-benefit curve exhibits an infinite slope at the limit of small cost, and that in some cases even an infinitesimal cost leads to a finite benefit resulting in a discontinuity in the cost-benefit curve.

### 3.5 Some Topics and Problems in Wireless Networking

*Muriel Médard (MIT – Cambridge, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Muriel Médard


We consider the interplay between information theory and practical networking in terms of a series of questions:

- What should the physical layer do?
- How do we deal with interference?
- How about cooperation?
- What about delay?
- What are the trade-offs?
- What about heterogeneity?
- Where does storage fit in?

We argue that equivalence theory points towards separating physical layer coding from the network, but that, in other respects, coding and networking are inextricably linked. In particular, while coding has generally been used to create a synthetic reliable pipe from lossy transmission media, it can be similarly used to synthesize different services, with varying delay and throughput trade-offs, over shared and often heterogeneous links, and over varied and distributed storage media, in a fluid approach.

### 3.6 Survey on “Economics in Wireless”

*Vijay Subramanian (University of Michigan – Ann Arbor, US)*


**License**  Creative Commons BY 3.0 Unported license  
© Vijay Subramanian

There are many topics where economics and game theory appear in wireless. The first and foremost is in investments and auctions of spectrum resources, which is a public resource in most countries that is auctioned to providers with defined property rights. Designing

computationally efficient, social-welfare maximizing, truthfully spectrum auctions in which providers voluntarily participate is challenging. Resource allocation games and their impact on market design is a topic where one can use general equilibrium theory based analysis to understand competition and sharing in wireless systems. Micro-payments using real-time distributed markets and auctions can be used for pricing, incentives and sharing in wireless systems. Finally, interference prices and best-response dynamics from the service providers can be used for scheduling and resource allocation.

### 3.7 Some interesting optimization problem formulations

*Michele Zorzi (University of Padova, IT)*

License  Creative Commons BY 3.0 Unported license  
© Michele Zorzi

In this talk, I will try to motivate the audience to understand some of the interesting and relevant stochastic optimization problems being addressed in the wireless networking area. This is in the spirit of trying to create a common understanding of each other's relevant research problems and possibly to stimulate cross-discipline collaboration or simply an appreciation of what others are doing. Towards this aim, I will present some problem formulations on which I have been working in the past few years, trying to describe how we approach problems, choose performance metrics, use tools and find solutions.

In particular, I will address the following areas: (i) derivation of an optimal transmission policy for an energy-harvesting device; (ii) derivation of an optimal transmission strategy for a secondary user in a cognitive radio scenario with ARQ and Interference Cancellation; (iii) study of optimal policies for wireless powered communication networks with doubly near-far effect; (iv) decentralized solutions for the multi-user MAC problem for energy-harvesting devices.

## 4 Summaries of “Speak-and-Spark” Talks

### 4.1 Massive MIMO and Compressed Sensing: a marriage made in 5G

*Giuseppe Caire (TU Berlin, DE)*

License  Creative Commons BY 3.0 Unported license  
© Giuseppe Caire

Massive MIMO channel estimation presents a number of problems ideally suited for compressed sensing. In 5 min I'd like to give an idea of why this happens and why compressed sensing can be very useful.

## 4.2 Simplicity and Robustness (or all about noisy beeping)

*Seth Gilbert (National University of Singapore, SG)*

Joint work of Seth Gilbert, Calvin Newport

We have been thinking about very simple forms of communication where entities can only communicate by beeping, such as either very simple sensors of biological networks. There are two basic questions we want to answer:

- Can we develop algorithms that are actually simple? (Or when we make the communication model simpler, do we inherently end up with more complicated algorithms?)
- Can we develop simple algorithms that are robust? (Especially in the context of biological systems, there is significant noise that can disrupt a computation; can we overcome it?)

To answer these questions, we have developed simple and robust algorithms as well shown lower bounds delineating the boundary of what is feasible.

## 4.3 Spatial Outage Capacity: The Missing Link between Average and Worst-Case Performance?

*Martin Haenggi (University of Notre Dame, US)*

License © Creative Commons BY 3.0 Unported license  
© Martin Haenggi

Joint work of Sanket Kalamkar and Martin Haenggi

Main reference S. S. Kalamkar and M. Haenggi, “The Spatial Outage Capacity in Wireless Networks”, arXiv:1708.05870v1 [cs.IT], 2017.

URL <https://arxiv.org/abs/1708.05870>

This brief presentation addresses a fundamental problem in wireless networks: what is the maximum density of reliable links that the network can support? This metric is called the spatial outage capacity (SOC). Reliability is captured using a constraint on the outage probability of a link, and the reliable link density is maximized over the node density  $\lambda$  and the transmit probability  $p$ . Surprisingly, this question has not been addressed before. Our analysis for Poisson networks with ALOHA shows that, surprisingly, that (unless the target reliability is small) the SOC is achieved by setting  $p = 1$ , i.e., by having all transmitters always active. This is counter-intuitive since one would think that in order to achieve high reliability, interference needs to be managed by allowing transmitters to be active only a fraction of time. In the high reliability regime, as the target outage probability goes to zero, we have obtained simple closed-form results on the SOC. It only depends on the rate-reliability ratio and the path loss exponent.

## 4.4 Intelligence and Network

*Longbo Huang (Tsinghua University – Beijing, CN)*

License © Creative Commons BY 3.0 Unported license  
© Longbo Huang

In this talk, I will survey some recent results we have on learning-aided stochastic network optimization. I will also introduce our recent effort on understanding the role of pricing and subsidies in the sharing economy and fast-convergence optimization algorithm design.



## 4.5 Fault-Tolerant Online Packet Scheduling on (Wireless) Channel(s)

Tomasz Jurdzinski (University of Wrocław, PL)

**License** © Creative Commons BY 3.0 Unported license  
© Tomasz Jurdzinski

**Joint work of** Pawel Garncarek, Tomasz Jurdzinski, Dariusz R. Kowalski, Krzysztof Lorys

**Main reference** Pawel Garncarek, Tomasz Jurdzinski, Krzysztof Lorys: “Fault-Tolerant Online Packet Scheduling on Parallel Channels”, in Proc. of the 2017 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2017, Orlando, FL, USA, May 29 - June 2, 2017, pp. 347–356, IEEE Computer Society, 2017.

**URL** <http://dx.doi.org/10.1109/IPDPS.2017.105>

This pitch talk discusses the model of online packet scheduling on (wireless) channel under adversarial errors [1] and its generalizations. An online algorithm is supposed to schedule dynamically arriving packets of various lengths, while transmissions of packets might be broken by an adversary. Focus on errors of transmission and various packets lengths reflect key features of contemporary wireless communication: high probability of errors on a communication channel and various requirements of services. In [3], an online algorithm for scheduling packets of arbitrary lengths is given, achieving optimal competitive throughput in  $(1/3, 1/2]$ . The model from [1] is generalized in [2] to the scenario with many parallel dependent channels. It is shown that the fact that errors have to appear simultaneously is beneficial for an online algorithm. Namely, an algorithm is designed which achieves the competitive throughput above  $1/2$  for  $p > 1$  channels, approaching  $3/4$  with increasing number of channels.

### References

- 1 A. F. Anta, C. Georgiou, D. R. Kowalski, J. Widmer, and E. Zavou. Measuring the impact of adversarial errors on packet scheduling strategies. In *Proc., of the 20th International Colloquium on Structural Information and Communication Complexity SIROCCO*, pages 261–273, 2013.
- 2 P. Garncarek, T. Jurdzinski, and K. Lorys. Fault-tolerant online packet scheduling on parallel channels. In *2017 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2017, 2017*, pages 347–356, 2017.
- 3 Tomasz Jurdzinski, Dariusz R. Kowalski, and Krzysztof Lorys. *Online Packet Scheduling Under Adversarial Jamming*, pages 193–206. In *Approximation and Online Algorithms – 12th International Workshop, WAOA 2014*, Springer, pages 193–206, 2014.

## 4.6 Some Thoughts on Wireless Networking Research

Holger Karl (Universität Paderborn, DE)

**License** © Creative Commons BY 3.0 Unported license  
© Holger Karl

1. NFV meets wireless multi-hop networks  
NFV has so far mostly been considered for fixed networks (core, mobile backhaul, mobile fronthaul networks, etc). Very little work has so far been invested in looking into the NFV concepts (and the algorithms that have emerged there) when applied to multi-hop networks. One possible application area might be acoustic sensor networks: microphones distributed in a network, non-trivial acoustic signal processing (e.g., speaker separation) at non-trivial data rates. The algorithmic challenge is to treat the placement (and possibly scaling) of individual function blocks in the network and the routing along interfering

paths as an integrated problem. It should result in many variations of this problem under plenty of different scenario assumptions.

See <https://cs.uni-paderborn.de/cn/research/research-projects/active-projects/acoustic-sensor-networks/> for more details

## 2. Control over wireless

Is there benefit by tighter integration between a wireless resource management system and a typical control application? Suppose we have a wireless cell, a controller which controls different sensors actuators over multiple wireless connections which have to share the resources in that link. The controller can send or request different amounts of data to each device under its control, obtaining more or less information or exerting more exact control. Links are assumed to undergo fading (independent? correlated? to be seen). In different situations of the controlled system, different actuators/sensors become urgent or irrelevant to be considered. Knowledge about this sits in the controller.

Goal is “best control performance” (e.g., minimize MSE) over a time-varying process. (Context: Throughput is usually not a problem; delay and error rates are much more relevant)

Question: how to best allocate wireless resources? How to design the interaction between controller and wireless resource management to that end?

See <https://cs.uni-paderborn.de/cn/research/research-projects/active-projects/nicci/> for more details.

## 3. Slicing – blessing or curse?

In 5G, “Slicing” is considered to be a magic wand: Resources are allocated to individual (groups of) users (e.g., “vertical industries” like car, manufacturing, etc.). Inside a slice, a lot of freedom regarding choice of protocols, PHY, MAC, ... exists.

Question: What is the tradeoff between increasing flexibility and customizing to specific, better defined needs (over all-purpose protocols) vs. the reduced diversity / multiplexing gains by dividing resources in a fixed fashion?


(Implementation complexity might be another consideration; but likely hard to capture in a theoretic concept)

## 4. Other, more generic rants:

- Why are we still bug-fixing broken protocols? CSMA/CA is not a promising tool for an AP-based scenario. There are reasons why cellular networks perform better, and their choice to do coordinated MAC is probably a big one of these reasons.
- More generally, let’s stop being in love with our models. Does the world really need more research on big mesh networks? Is there any single one

## 4.7 Routing in Hybrid Networks with Holes

*Christina Kolb (Universität Paderborn, DE)*

License  Creative Commons BY 3.0 Unported license

© Christina Kolb

Joint work of Christina Kolb, Christian Scheideler

Think of you having a contract with a smartphone provider. This contract offers a fixed Internet volume that is available for free, but that volume is rather small, so you want to use it as scarcely as possible. Fortunately, there are many other people nearby who are also interested in saving their Internet volume, including some of your friends. However, they are too far away from you to directly use WLAN communication, but in principle there

would be a route via other people along which your messages can be sent using WLAN communication to communicate with your friends. But how to find such a route efficiently? It is well-known in the research community that doing that just via WLAN communication does not scale well, but some of the Internet volume could be used in order to exchange control information, such as the current geographic positions of your friends or the location and dimension of radio holes (i.e., areas with no participants). Hence, we can set up a so-called hybrid communication network between the participants in order to find routes for WLAN communication more efficiently, but how much more efficiently would that be?

In this talk, we present the model and goals of the setting above.

## 4.8 Enabling Extreme Resource Sharing in Future Wireless Communication Systems

*Konstantinos Nikitopoulos (University of Surrey, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Konstantinos Nikitopoulos

There is a current paradigm shift from orthogonal to non-orthogonal signal transmissions that enables extreme sharing of the available resources, but it requires processing complexities far beyond the capabilities of traditional processors. In this context, massively parallel processing detection/decoding approaches are required to overcome the current processing speed barriers. In the same framework of extreme resource sharing, new PHY can be efficiently used to enable joint medium access and rateless data transmission while reducing or even eliminating the need for transmitting preamble sequences.

## 4.9 Towards a widely applicable SINR model for access sharing

*Christian Scheideler (Universität Paderborn, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Christian Scheideler

I propose an SINR model in which Background noise is controlled by an adversary and present a preliminary solution for it.

## 4.10 Wake-up Transceivers, Convergecast, Beamforming, and Beat Frequencies

*Christian Schindelhauer (Universität Freiburg, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Christian Schindelhauer

**Joint work of** Christian Schindelhauer, Thomas Janson, Amir Bannoura, Timo Kumberg  
**Main reference** Timo Kumberg, Christian Schindelhauer, Leonhard M. Reindl: “Exploiting Concurrent Wake-Up Transmissions Using Beat Frequencies”, *Sensors*, Vol. 17(8), p. 1717, 2017.  
**URL** <http://dx.doi.org/10.3390/s17081717>

This pitch talks highlights three topics concerned with collaborative signal transmission and wake-up receivers.

In his PhD thesis Thomas Janson [1] has shown that a multi hop unicast with simultaneous sending can send a message within  $O(\log \log n)$  hops and with energy  $O(\log d)$ , where single signal transmissions need linear time and energy with respect to the distance  $d$ . The open problem however is, whether such a beam can be widened for broadcast and how this affects time and energy.

Wireless Sensor Networks suffer from energy shortage and use duty cycling to reduce energy consumption and manage communication. Wake-up receivers overcome this problem. They always listens for own ID with some micro-Watts power consumption. Furthermore, every sensor node can send wake-up signals. However, they are short ranged and wake-up signals need additional time. In his PhD thesis Amir Bannoura [2] has shown several algorithms for broadcast and convergecast.

Combining these works Timo Kumberg et al. has considered the use of collaborative wake-up signals with different carrier frequencies such that the beat frequency produces the wake-up signal. Yet, it is not fully understood how such a construction scales for larger set of senders.

#### References

- 1 Thomas Janson *Energy-Efficient Collaborative Beamforming in Wireless Ad Hoc Networks*. PhD Thesis, University of Freiburg, 2015
- 2 Amir Bannoura, *Wake-Up Receivers Algorithms and Applications for Low Power Wireless Sensor Networks*. PhD Thesis, University of Freiburg, 2016

### 4.11 Distributed optimization for scheduling in wireless networks


*Vijay Subramanian (University of Michigan – Ann Arbor, US)*

License  Creative Commons BY 3.0 Unported license  
© Vijay Subramanian

Using the network utility maximization framework, we show how locally optimal scheduling and resource allocation can be performed using local updates, function approximation and consensus.

### 4.12 Fine-grain Time / Spectrum Sharing

*Patrick Thiran (EPFL – Lausanne, CH)*

License  Creative Commons BY 3.0 Unported license  
© Patrick Thiran

**Joint work of** Julien Herzen, Albert Banchs, Vsevolod Shneer, Patrick Thiran

**Main reference** Julien Herzen, Albert Banchs, Vsevolod Shneer, Patrick Thiran: “CSMA/CA in Time and Frequency Domains”, in Proc. of the 23rd IEEE International Conference on Network Protocols, ICNP 2015, San Francisco, CA, USA, November 10-13, 2015, pp. 256–266, IEEE Computer Society, 2015.

**URL** <http://dx.doi.org/10.1109/ICNP.2015.16>

It has recently been shown [see e.g. K. Tan et al, Proc. ACM SIGCOMM, 2010; S. Yun et al, Proc. ACM MobiCom, 2013] that flexible channelization, whereby wireless stations adapt their spectrum bands on a per-frame basis, is becoming feasible in practice. This offers the potential to strongly improve in the future the performance of MAC schemes, such as 802.11, which currently solve contention only in the time-domain. I briefly describe TF-CSMA/CA [J. Herzen et al, Proc. IEEE ICNP 2015], a first fully distributed algorithm for flexible channelization that schedules packets in time and frequency domains.

### 4.13 Wireless Link Capacity under Shadowing and Fading

*Tigran Tonoyan (Reykjavik University, IS)*

**License** © Creative Commons BY 3.0 Unported license  
© Tigran Tonoyan

**Joint work of** Magnús M. Halldórsson, Tigran Tonoyan

We consider the following basic spatial reuse problem in wireless networks: Given a set of communication links, find a maximum subset of links that can successfully transmit in the same channel and time slot. The problem has many components, including the way signal attenuation and reception is modeled. We consider three aspects: geometric pathloss, shadowing (spatial) and temporal fading. Towards the goal of understanding the general behavior of networks, we adopt stochastic models describing the latter two phenomena: stochastic shadowing (e.g., Lognormal shadowing) and Rayleigh fading. We present algorithms computing solutions under stochastic shadowing and fading, with arbitrary placement of links. We further show that Rayleigh (temporal) fading affects the size of the solution only by a constant factor, while stochastic shadowing can give significantly better results than deterministic geometric pathloss alone. These results are obtained under some independence assumptions on the distributions. The main question to be answered is: how to model dependencies between distributions in general networks?

### 4.14 Information Theoretic Caching

*Daniela Tuninetti (University of Illinois – Chicago, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Daniela Tuninetti

I give a very short summary of some results in centralized coded caching.

### 4.15 Update or Wait: How to Get the Freshest and Most Accurate Data Through a Network

*Elif Uysal-Biyikoglu (Middle East Technical University – Ankara, TR)*

**License** © Creative Commons BY 3.0 Unported license  
© Elif Uysal-Biyikoglu

An ever increasing amount of data on today's networks is in the form of status updates. Often, the source that generates the data (e.g. sensor) and the remote application that uses the data are connected through a network with varying quality and delay.

We study how to optimally manage the freshness of information updates sent from the source of the data to the destination via a network with random delay. A proper metric for data freshness at the destination is the Age of Information, or simply age, which is defined as the amount of time that elapsed since the freshest received update received so far was generated at the source. A reasonable update policy is the zero-wait policy, i.e., the source node submits a fresh update once the previous update is delivered and the channel becomes free, which achieves the maximum throughput and the minimum delay. Surprisingly, this zero-wait policy, which is work-conserving (and optimal with respect to throughput and

delay) does not always minimize the age. This counter-intuitive phenomenon motivates us to study how to optimally control information updates to keep the data fresh.

Next, we generalize the Age problem to consider a real-time sampling problem: Samples of a Wiener process are taken and forwarded to a remote estimator via a channel with random delay; the estimator forms a real-time estimate of the signal from causally received samples. The optimal sampling policy for minimizing the MMSE subject to a sampling-rate constraint is obtained exactly, which is determined by the signal, sampler, and channel in a simple form. Echoing the result of the Age problem, even in the absence of a restriction on sampling rate, there is a nonzero waiting time before taking the next sample. In fact, it is often optimal to sample below the maximum allowed sampling rate.

## 4.16 Storage Cost of Shared Memory Emulation

*Zhiying Wang (University of California – Irvine, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Zhiying Wang

**Joint work of** Viveck R. Cadambe, Nancy Lynch

**Main reference** Viveck R. Cadambe, Zhiying Wang, Nancy A. Lynch: “Information-Theoretic Lower Bounds on the Storage Cost of Shared Memory Emulation”, in Proc. of the 2016 ACM Symposium on Principles of Distributed Computing, PODC 2016, Chicago, IL, USA, July 25-28, 2016, pp. 305–313, ACM, 2016.

**URL** <http://dx.doi.org/10.1145/2933057.2933118>

**Main reference** Zhiying Wang, Viveck Cadambe, “Multi-Version Coding – An Information Theoretic Perspective of Consistent Distributed Storage”, arXiv:1506.00684v2 [cs.IT], 2015.

**URL** <http://arxiv.org/abs/1506.00684>

Shared memory emulation are important algorithms that bridge the two communication models of asynchronous distributed systems: the shared-memory model, and the message-passing model. Previous literature has developed several shared memory emulation algorithms based on replication and erasure coding techniques. In this talk, we present information-theoretic lower bounds on the storage costs incurred by such algorithms. We show that the storage cost is at least twice that of the Singleton-type of bound; and for a restricted class of write protocols, the storage cost grows approximately linearly with the number of servers failures and the number of concurrent writes in an execution. An interesting observation is that, when the number of concurrent writes is large, replication based algorithms have asymptotically optimal storage cost.

## 4.17 PHY in Networks

*Roger Wattenhofer (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license  
© Roger Wattenhofer

**Joint work of** Michael König, Roger Wattenhofer

I present some of our recent results regarding physical effects in networks, in particular (i) how to effectively capture attention using the capture effect, (ii) how to generate constructive interference using well-synchronized nodes, and (iii) how to share a medium between concurrent protocols without overhead.

## References

- 1 Michael König, Roger Wattenhofer, Effectively Capturing Attention Using the Capture Effect. SenSys 2016: 70–82
- 2 Michael König, Roger Wattenhofer, Maintaining Constructive Interference Using Well-Synchronized Sensor Nodes. DCOSS 2016: 206–215
- 3 Michael König, Roger Wattenhofer, Sharing a Medium Between Concurrent Protocols Without Overhead Using the Capture Effect. EWSN 2016: 113–124

## 4.18 Problems in millimeter-wave communication: the difficult path from Gigabit links to Gigabit networks

Jörg Widmer (*IMDEA Networks – Madrid, ES*)

**License** © Creative Commons BY 3.0 Unported license  
© Jörg Widmer

State-of-the-art wireless communication already operates close to Shannon capacity and one of the most promising options to further increase data rates is to increase the communication bandwidth. Very high bandwidth channels are only available in the extremely high frequency part of the radio spectrum, the millimeter wave band (mm-wave). Upcoming communication technologies, such as IEEE 802.11ad, are already starting to exploit this part of the radio spectrum to achieve data rates of several GBit/s. However, communication at such high frequencies also suffers from high attenuation and signal absorption, often restricting communication to line-of-sight (LOS) scenarios and requiring the use of highly directional antennas. This in turn requires a radical rethinking of wireless network design. On the one hand side, such channels experience little interference, allowing for a high degree of spatial reuse and potentially simpler MAC and interference management mechanisms. On the other hand, such an environment is extremely dynamic and channels may appear and disappear over very short time intervals, in particular for mobile devices. It is essential to take these characteristics into account to turn a collection of such very high speed but brittle links into an efficient, low latency, and reliable network. This talk will highlight some of the challenges of and possible approaches for mm-wave networking.

## 4.19 Wireless Networks: Dynamicity

Dongxiao Yu (*Huazhong University of Science & Technology, CN*)

**License** © Creative Commons BY 3.0 Unported license  
© Dongxiao Yu

**Joint work of** Dongxiao Yu, Yuexuan Wang, Tigran Tonoyan, Magnús M. Halldórsson

**Main reference** Dongxiao Yu, Yuexuan Wang, Tigran Tonoyan, Magnús M. Halldórsson: “Dynamic Adaptation in Wireless Networks Under Comprehensive Interference via Carrier Sense”, in Proc. of the 2017 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2017, Orlando, FL, USA, May 29 - June 2, 2017, pp. 337–346, IEEE Computer Society, 2017.

**URL** <http://dx.doi.org/10.1109/IPDPS.2017.78>

In this talk, we will introduce our recent efforts in distributed algorithm design in dynamic networks. Dynamic behavior is an essential part of wireless networking, due to mobility, environmental changes or failures. We analyze a natural exponential backoff procedure to manage contention in a fading channel, in the presence of both node churn and link changes. We show that it attains a fast convergence, stabilizing contention from any state

in logarithmic time. We use it to obtain optimal algorithm for Local Broadcast that even improves known results for the static case. The results illustrate the utility of carrier sensing, a stock feature of wireless nodes.

## 4.20 Passive Communication with Ambient Light

*Marco Zuniga (TU Delft, NL)*

**License** © Creative Commons BY 3.0 Unported license  
© Marco Zuniga

**Joint work of** Domenico Giustiniano, Marco Zuniga, Qing Wang

**Main reference** Qing Wang, Marco Zuniga, Domenico Giustiniano: “Passive Communication with Ambient Light”, in Proc. of the 12th International on Conference on emerging Networking EXperiments and Technologies, CoNEXT 2016, Irvine, California, USA, December 12-15, 2016, pp. 97–104, ACM, 2016.

**URL** <http://dx.doi.org/10.1145/2999572.2999584>

In this work, we propose a new communication system for illuminated areas, indoors and outdoors. Light sources in our environments –such as light bulbs or even the sun– are our signal emitters, but we do not modulate data at the light source. We instead propose that the environment itself modulates the ambient light signals: if mobile elements ‘wear’ patterns consisting of distinctive reflecting surfaces, single photodiode could decode the disturbed light signals to read passive information. Achieving this vision requires a deep understanding of a new type of communication channel. Many parameters can affect the performance of passive communication based on visible light: the size of reflective surfaces, the surrounding light intensity, the speed of mobile objects, the field-of-view of the receiver, to name a few. In this work, we present our vision for a passive communication channel with visible light.

## 5 Discussion groups

### 5.1 Discussion Group: Coding for New Channels

*Michelle Effros (CalTech – Pasadena, US), Tomasz Jurdzinski (University of Wroclaw, PL), Konstantinos Nikitopoulos (University of Surrey, GB), Patrick Thiran (EPFL – Lausanne, CH), Jörg Widmer (IMDEA Networks – Madrid, ES), and Marco Antonio Zúñiga Zamalloa (TU Delft, NL)*

**License** © Creative Commons BY 3.0 Unported license  
© Michelle Effros, Tomasz Jurdzinski, Konstantinos Nikitopoulos, Patrick Thiran, Jörg Widmer, and Marco Antonio Zúñiga Zamalloa

In practice, there is a plethora of physical communication channels that share commonalities but also differences. Such channels include the traditional wireless communication channels, the newly proposed visible light communication (VLC) ones, the optical channels, and perhaps other communication channels, like underwater and biological ones. While some communications channels have been very well explored, other newly proposed, like some in the field of VLC, seem to be less well understood.

In our effort to understand new communications channels, it becomes apparent that there is an absence of a common framework that could allow us to analyze and utilize these communications channels in holistic manner and, therefore easily build upon prior knowledge. In this direction, such a framework should account for details about the transmission channel



(e.g., linear, or time (in)variant) as well as the corresponding noise characteristic, that should form the basis to produce holistic and generalized information theoretical results (both at a link and a network level), as well as produce efficient coding schemes. Such coding schemes can be further categorized according to their practicality vs. optimality.

In such a framework, some of the aspects that need to be better explored are:

- The role of transmission “imperfections”, as for example the non-linearities in optical systems. and how they affect both practicality, system design, practical codes and information theoretical results,
- Information theoretical results about the network aspects of such systems, including multiple access,
- Generic methodologies to achieve those limits,
- Coding approaches that are applied to structure of the transmitted information, instead of the traditional bit level coding (e.g., Space-Time-Super Modulation).

## 5.2 Discussion Group: Interference and Scheduling

*Patrick Thiran (EPFL – Lausanne, CH)*

License  Creative Commons BY 3.0 Unported license  
© Patrick Thiran

### 1. Background

Interference has a negative connotation, although information theory tells us that it does not need to be a bad thing. Wireless networks do not currently exploit these information theoretic findings, what are the steps needed to bring these advantages to reality?

### 2. Some fundamental challenges:

- When can interference be overcome, and when will it remain an unavoidable pain? What are the system parameters that delineate this border, not only under an information theoretic perspective, but also under a networking perspective?
- When and how to “reconfigure” the network so that it works with high SNR links in a regimes where interferences can be handled effectively?
- Global knowledge is often needed in the above setting but is not available in practice. What can we achieve with only a local knowledge? What is the control information that needs to be exchanged with other network users and which is the most effective for coordinating them, possibly using out-of-band communication, in hybrid communication schemes?

### 3. Practical approaches

None that was identified so far, but one can be felt is the translation of information-theoretic findings in results that can be exploited at the networking layers. Abstractions have played a key role in developing network protocols that work at a large scale, finding the proper abstraction model for these information-theoretic findings would be very helpful.

### 4. Different community perspectives

Coping with interference is viewed as a problem well solved for N users for the multiple access channel by information theoreticians, but not by the other communities.

### 5. Initial solution directions:

- Information theory provides strong results about handling interference for the multiple access channel model (packets from any user can be of interest for the receiver), but

the interference channel model (packets from only one user are of interest for a given receiver) remains open for  $N > 2$  users.

- Current network protocols discard packets that have collided, but recent work on random access codes show a promising direction to make use of these packets with different random overlaps at high SNR. With a sufficient number of retransmissions, the packet could be recovered even if all these retransmissions resulted in a collision (zig-zag codes).

## Participants

- Giuseppe Caire  
TU Berlin, DE
- Izzat Darwazeh  
University College London, GB
- Michelle Effros  
CalTech – Pasadena, US
- Anja Feldmann  
TU Berlin, DE
- Christina Fragouli  
University of California at  
Los Angeles, US
- Seth Gilbert  
National University of  
Singapore, SG
- Deniz Gündüz  
Imperial College London, GB
- Martin Haenggi  
University of Notre Dame, US
- Magnús M. Halldórsson  
Reykjavik University, IS
- Longbo Huang  
Tsinghua University –  
Beijing, CN
- Kyle Jamieson  
Princeton University, US &  
University College London, GB
- Tomasz Jurdzinski  
University of Wroclaw, PL
- Holger Karl  
Universität Paderborn, DE
- Christina Kolb  
Universität Paderborn, DE
- Bhaskar Krishnamachari  
USC – Los Angeles, US
- Fabian Daniel Kuhn  
Universität Freiburg, DE
- Muriel Médard  
MIT – Cambridge, US
- Konstantinos Nikitopoulos  
University of Surrey, GB
- Christian Scheideler  
Universität Paderborn, DE
- Christian Schindelhauer  
Universität Freiburg, DE
- Emina Soljanin  
Rutgers University –  
Piscataway, US
- Vijay Subramanian  
University of Michigan –  
Ann Arbor, US
- Patrick Thiran  
EPFL – Lausanne, CH
- Tigran Tonoyan  
Reykjavik University, IS
- Daniela Tuninetti  
University of Illinois –  
Chicago, US
- Elif Uysal-Biyikoglu  
Middle East Technical University  
– Ankara, TR
- Zhiying Wang  
University of California –  
Irvine, US
- Roger Wattenhofer  
ETH Zürich, CH
- Jörg Widmer  
IMDEA Networks – Madrid, ES
- Dongxiao Yu  
Huazhong University of Science  
& Technology, CN
- Michele Zorzi  
University of Padova, IT
- Marco Antonio Zúñiga  
Zamalloa  
TU Delft, NL



# Citizen Science: Design and Engagement

Edited by

Irene Celino<sup>1</sup>, Oscar Corcho<sup>2</sup>, Franz Hölker<sup>3</sup>, and Elena Simperl<sup>4</sup>

<sup>1</sup> CEFRIEL - Milan, IT, [irene.celino@cefriel.com](mailto:irene.celino@cefriel.com)

<sup>2</sup> Polytechnic University of Madrid, ES, [ocorcho@fi.upm.es](mailto:ocorcho@fi.upm.es)

<sup>3</sup> IGB - Berlin, DE, [hoelker@igb-berlin.de](mailto:hoelker@igb-berlin.de)

<sup>4</sup> University of Southampton, GB, [e.simperl@soton.ac.uk](mailto:e.simperl@soton.ac.uk)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17272 "Citizen Science: Design and Engagement". In this report, we briefly summarise the content of three invited keynote talks and two invited tutorials. We further outline the findings of five parallel working groups, which met on the first and third day of the workshop, in the areas of: sustainability, measuring success, community engagement, linking and quality.

**Seminar** July 2–5, 2017 – <http://www.dagstuhl.de/17272>

**1998 ACM Subject Classification** I.2 Artificial Intelligence, I.2.9 Robotics, Society, Human-computer Interaction

**Keywords and phrases** Citizen science, Crowdsourcing, Data Analytics, Gamification, Human Computation, Incentives Engineering, Online Community, Open Science

**Digital Object Identifier** 10.4230/DagRep.7.7.22

**Edited in cooperation with** Neal Reeves (University of Southampton)

## 1 Executive summary

*Irene Celino*

*Oscar Corcho*

*Franz Hölker*

*Elena Simperl*

**License** © Creative Commons BY 3.0 Unported license

© Irene Celino, Oscar Corcho, Franz Hölker, Elena Simperl

Citizen science is an approach to science that is enlisting the help of millions of volunteers across a range of academic disciplines to complete tasks that would have otherwise been unfeasible to tackle using expert time or computational methods [2]. While it is a popular and effective way to solve various problems, with many examples of incredible success [3, 1], there remains a number of ongoing challenges that must be addressed in order to ensure the validity of citizen science as a widespread approach to research. The aim of this workshop – organised in partnership with the SOCIAM<sup>1</sup> and Stars4All<sup>2</sup> projects – was to discuss and explore aspects of the future of citizen science, focusing on design factors and engagement strategies, although this naturally required a holistic assessment of citizen science projects, platforms and applications as a whole.

---

<sup>1</sup> <https://sociam.org/about>

<sup>2</sup> <http://stars4all.eu/>



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Citizen Science: Design and Engagement, *Dagstuhl Reports*, Vol. 7, Issue 7, pp. 22–43

Editors: Irene Celino, Oscar Corcho, Franz Hölker, and Elena Simperl



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

**References**

- 1 Jeehung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille et al. *RNA Design Rules from a Massive Open Laboratory*. Proceedings of the National Academy of Sciences, National Academy of Sciences, 111; 6, 2122–2127, 2014.
- 2 Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Bamford Steven, Daniel Thomas, Jordan M Raddick, Robert C Nichol, Alex Szalay and Dan Andreescu et al. *Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey*. Monthly Notices of the Royal Astronomical Society, Blackwell Publishing Ltd Oxford, UK, 389; 3, 1179–1189, 2008.
- 3 Chris J Lintott, Kevin Schawinski, William Keel, Hanny Van Arkel, Nicola Bennert, Edward Edmondson, Daniel Thomas, Daniel JB Smith, Peter D Herbert, Matt J Jarvis et al. *Galaxy Zoo: ‘Hanny’s Voorwerp’, a quasar light echo?*. Monthly Notices of the Royal Astronomical Society, Blackwell Publishing Ltd Oxford, UK, 399; 1, 129–140, 2009.

## 2 Table of Contents

### Executive summary

*Irene Celino, Oscar Corcho, Franz Hölker, Elena Simperl* . . . . . 22

### Introduction

#### Overview of Keynote Talks

Crowdsourcing for Smart Cultural Heritage: Harnessing Human Semantics at Scale  
*Lora Aroyo* . . . . . 26

Open Citizen Science: What, how and with whom?  
*Claudia Göbel* . . . . . 26

The state of the art of OpenStreetMap: technologies, community and research challenges  
*Maurizio Napolitano* . . . . . 27

#### Overview of Tutorials

Citizen Science as a New Way To Do Science  
*Marisa Ponti* . . . . . 27

Citizen Science as a Social Machine  
*Elena Simperl* . . . . . 27

#### Overview of Working groups

Working Group One - Sustainability . . . . . 28

Working Group Two - Measuring Success . . . . . 30

Working Group Three - Community Engagement . . . . . 31

Working Group Four - Linking . . . . . 34

Working Group Five - Quality . . . . . 34

Working Group Six - Manifesto for Citizen Science . . . . . 35

**Participants** . . . . . 36

**Appendix A - Lightning Talk Slides** . . . . . 37

### 3 Introduction

While amateur involvement in science began long before the establishment of modern academic institutions, the web and digital technologies have fundamentally revitalized and expanded the ways and scale in which untrained citizens can participate in scholarly research. These ‘Citizen Science’ projects have thus far enlisted the help of millions of volunteers in a wide array of scientific inquiries, ranging from the taxonomic classification of galaxies and the creation of an online encyclopedia of all living species on Earth, to the derivation of solutions to protein folding problems and the tracking and measuring the population and migratory patterns of animals in the Serengeti national park [2]. This new, more inclusive way of pursuing science is proving successful in many ways: it gives scientists around the world an effective, affordable way to collect and analyze large amounts of data in a short period of time, popularizes scientific topics to wider audiences, and encourages the formation of amateur scientific communities, which initiate their own projects and deliver notable results.

The seminar was arranged as a platform to discuss and explore the aspects of and to outline the future of the citizen science research, platforms, and applications. The seminar was organized around the following three perspectives:


1. A crowdsourcing perspective that views citizen science as a large-scale volunteer-driven human computation system. Relevant aspects include:
  - Task and workflow design
  - User experience design
  - Answer validation
  - Task assignment and contributor performance
  - Crowd learning, feedback, and tutorials
  - Gamification and rewards
2. An online community perspective that considers social and other communication and interaction activities that support task-centered efforts. This is related to quantitative and qualitative approaches for content and community analysis, including:
  - Analysis of discussion forum and chat activity
  - Social network analysis
  - Interplay with other community spaces such as social media
  - Analysis of community trajectories
  - Lurker behavior and more general contribution patterns
  - Conflict and collaboration
  - Surveys of motivation and incentives
3. An open science perspective that focuses on citizen science as an emerging model of collaborative research. In particular:
  - Open, participatory approaches to all stages of the scientific lifecycle
  - Crowdfunding for science
  - Open access publishing of research ideas and outcomes
  - Openness in data acquisition and sharing
  - Participation of young volunteers in citizen science activities
  - Scientific publishing for crowdsourced science

In order to discuss the design and engagement of citizen science, the workshop consisted of a number of talks and working groups that incorporated these differing perspectives throughout.

## 4 Overview of Keynote Talks

### 4.1 Crowdsourcing for Smart Cultural Heritage: Harnessing Human Semantics at Scale


*Lora Aroyo (Vrije Universiteit Amsterdam, NL)*

License  Creative Commons BY 3.0 Unported license  
© Lora Aroyo

The state of the art in machine learning and information extraction has advanced the detection and recognition of concepts and objects, like people, locations, and various other types of named entities. However, still there is various types of human knowledge that cannot yet be captured by machines, especially when dealing with wide ranges of real-world tasks and contexts. The key scientific challenge is to provide an approach to capturing human knowledge in a way that is scalable and adequate to real-world needs. Human Computation has begun to scientifically study how human intelligence at scale can be used to methodologically improve machine-based knowledge and data management. My research focuses on understanding human computation for improving how machine-based systems can acquire, capture and harness human knowledge and thus become even more intelligent. In this talk I will present use cases related to smart culture, e.g. enrichment of cultural heritage collections of artworks, videos, newspapers, etc. I will show how the CrowdTruth crowdsourcing framework <http://crowdtruth.org> facilitates data collection, processing and analytics of human computation knowledge. Processing real-world data with the crowd leaves one thing absolutely clear - there is no single notion of truth, but rather a spectrum that has to account for context, opinions, perspectives and shades of grey. CrowdTruth is a new framework for processing of human semantics drawn more from the notion of consensus than from set theory.

### 4.2 Open Citizen Science: What, how and with whom?

*Claudia Göbel (Museum für Naturkunde Berlin/ European Citizen Science Association, DE)*

License  Creative Commons BY 3.0 Unported license  
© Claudia Göbel

The presentation focused on unpacking relations of open and collaborative aspects in Citizen Science. On a conceptual level, I identified synergies and tensions between Citizen Science and Open Science by mapping agendas from the European research policy discourse against each other. What is done in Citizen Science practice was explored by looking at findings of an international stakeholder analysis on Citizen Science data interoperability and examples from other areas of Open Science. Finally, I sketched some cornerstones of an analytical perspective focusing on stakeholder networks and organizations for further analysis.



### 4.3 The state of the art of OpenStreetMap: technologies, community and research challenges

*Maurizio Napolitano (FBK - Centre for Information Technology/ Open Street Map, IT)*

**License** © Creative Commons BY 3.0 Unported license  
© Maurizio Napolitano

OpenStreetMap - OSM is known as the free world map created on a voluntary basis. But OSM is not just a map, it is much more: one of the biggest geo-referenced open data resources, a community of people who are able to give voice to a territory, an important resource for entrepreneurial initiatives, and much more. This talk introduced the project from its history and technological aspects in order to highlight what challenges science could help solving and the benefits deriving from it.

## 5 Overview of Tutorials

### 5.1 Citizen Science as a New Way To Do Science

*Marisa Ponti (University of Göteborg, SE)*

**License** © Creative Commons BY 3.0 Unported license  
© Marisa Ponti

**Main reference** Marisa Ponti, “Citizen Science as a New Way To Do Science”, SocArXiv, 2017.  
**URL** <http://dx.doi.org/10.17605/OSF.IO/KGXSQ>

Citizen science has received increasing attention because of its potential as a cost-effective method of gathering massive data sets and as a way of bridging the intellectual divide between layperson and scientists. Citizen science is not a new phenomenon, but is implemented in new ways in the digital age, offering opportunities to shape new interactions between volunteers, scientists and other stakeholders, including policymakers. Arguably, citizen science rests on two main pillars: openness and participation. However, openness remains unexploited if we do not create the technical and social conditions for broader participation in more collaborative citizen science projects, beyond collecting and sharing data with scientists. “Public participation” has too often accounted for the assumed ease with which hierarchies in science can be horizontalized, and economic and geographic barriers can be removed. However, public participation is a contested term, which should be problematized. The Scandinavian tradition of participatory design can help explore conceptually the challenges related to participation and to design for participation.

### 5.2 Citizen Science as a Social Machine

*Elena Simperl (University of Southampton, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Elena Simperl

Tim Berners-Lee envisaged the web as a platform for large-scale participation. Social machines are socio-technical constructs that enable all this. They define and support processes where people contribute creative work and algorithms carry out the more routine, predictable or “engineer-able” tasks to bring the results of that work together. Citizen science is itself an

example of a social machine: bringing groups and communities of volunteers together to help advance science by collecting or analysing data relevant to scientific experiments. To date, lots of the research has focused on showcasing how and where citizen science can help or studying specific properties of the systems using a variety of methods from social computing, HCI, or scientific communication. We need to appreciate that citizen science can be studied from several angles, and that more research is needed in understanding how citizen science can move away from being just an approach useful for the professional scientist to a fully-fledged social machine. In the same time, we need to design better tools and teach the scientists to understand how to improve and maintain community health and determine whether the volunteers are happy and producing enough data. This talk introduced a number of research questions and solutions, such as how to design platforms for meaningful contributions, how to manage the data produced by the volunteers, studying emerging communities, and designing incentives for participation. The ultimate question is ‘what makes citizen science successful?’ and we need a interdisciplinary, inclusive approach to solve it.

## 6 Overview of Working groups

Working groups were held on days 1 and 3 of the workshop in order to facilitate discussions on topics related to the areas outlined above. In the following sections we provide brief summaries of the main outputs of these working groups.

### 6.1 Working Group One - Sustainability

*Reported by Alessandro Bozzon.* This group discussed various dimensions of sustainability related to citizen science. These are presented in the table below, along with a series of challenges for each that must be overcome in order to ensure sustainable projects and practices. In addition the group discussed a number of developments that may help to increase sustainability, such as a ‘graveyard’ of projects that have run their course, addressed their hypothesis and wrapped up engagements with their community, along with a repository of workflows that have been annotated with what problems they’ve worked well on previously. This would allow other citizen science teams to see what has worked, and build from there. Other suggestions were a high-school level curriculum module on citizen science, to increase awareness of this from a younger age, and a re-usable set of personas to help in the design process.

1. Openness and Reuse of Data
  - Data and Metadata
    - Incentivising data sharing
    - Dealing with diverse standards
    - Licensing
    - Provenance and tracking
  - Technology
    - Creating a Github for open citizen science hardware and software
    - Designing projects for re-use
    - Describing workflows
    - Standardisation
    - Document deployment configurations and conditions

- Discovering existing initiatives
    - Creating a citizen science directory
    - <http://firstmonday.org/ojs/index.php/fm/article/view/5520/4194>
    - <https://scistarter.com/>
    - <http://vgibox.eu/>
    - <https://www.citizensciencealliance.org/>
    - [https://en.wikipedia.org/wiki/List\\_of\\_citizen\\_science\\_projects](https://en.wikipedia.org/wiki/List_of_citizen_science_projects)
  - Methodology
    - Co-creation and knowledge sharing for project design and implementation
    - People
    - Incentives
    - Tasks
    - Trade-offs
    - Lessons Learned
    - For example, repository of case studies, personas, design guidelines, etc.
- 2. Socio-economic aspects
  - Engagement/Participation
    - Limited number of potential participants
    - Reaching motivated participants
  - Fitness for purpose/availability
    - Evidence of required expertise
    - Trustworthiness
    - Spatial and temporal activity patterns
    - Responsiveness
  - Agency/Empowerment
    - Community tools for building citizen science
    - Co-creation of projects and research
  - Education
    - Accepted as curriculum
    - Hands-on and built on practice
    - Revisit pedagogical approaches
  - Social and Ethical goals
    - Encourage empowerment of under-represented groups (e.g., migrants)
    - Increase enthusiasm for science in lay public
- 3. Ecosystemics
  - Project Variety
    - Ensuring ecosystem is fit for different timescales and project cultures
  - Project Lifecycle
    - Funding - connecting funding for infrastructure and for doing science
    - End of Life - data archival, communities and connections
    - Creating bridges to other projects

## 6.2 Working Group Two - Measuring Success

*Reported by Paul Groth.* The success working group discussed a range of topics such as investigating what constitute ‘success’ in a citizen science context, the different types of outcome, and different frameworks for assessing various facets of success. The group then moved on to outline a range of grand challenges for this topic area. Each challenge was summarised with an impact statement of where, ideally, the state of play would be in ten years time.

### **Predict the success of a project before starting, based on a manual of best practices.**

For this challenge, the different criteria for different stakeholder groups were discussed, along with what features can be looked at to understand success. The scientists themselves, funders, and citizens may all have differing expectations about what success means in a citizen science project, and going even further there are planetary scale stakeholders that are concerned with the positive ecological impacts that many of these initiatives may have. Success may be measured using a huge range of metrics and features, going beyond measures of the number of classifications or volunteers to look at how many visitors the project has attracted, the number of scientists and developers engaged in the process, the quality of the input data, the number of countries reached, etc. As a grand vision, in ten years time, we will be able to predict successful citizen science projects and tell you why.

**Expand the diversity of contributors to citizen science.** In order for citizen science as an approach to be truly successful, there needs to be a diverse range of volunteers who participate in projects to ensure that there is no cultural, socio-economic, age, educational or language bias recorded in the data. We identified a number of steps required such as identifying what the current bias is, and eliciting methods of motivating more people to participate, from different backgrounds. There were also discussions around what automatic adjustments to design could be made so that an existing project or tasks could easily be replicated or made suitable for a different demographic group. The proposed impact statement for this group was: in ten years time, we will have citizen science projects that reflect the demographic breakdown of the world.

**Radically scaling up co-created citizen science.** Citizen science already reaches large numbers of volunteers in some cases, with this an essential aspect of many projects. However the role that the volunteers play in the design process is less established, and one of the areas of further potential is a true collaborative, co-creation process at all stages of the scientific process (as opposed to the majority of current projects that incorporate volunteer involvement in the data collection and/or analysis phase). Foundational work needs to be carried out to understand what is currently done in co-creation process so that experiences and lessons learned may be documented. A testing and development process for components of the scientific method workflow will then allow the creation of processes that work across multiple scientific domains, before over time things are scaled up so that performance may be tested against a baseline to determine the successfulness of running truly co-created citizen science projects. If it is possible for a completely citizen-led project to match or even beat the performance of professional scientists’ own research then citizen science will have reached a high level of success as an approach to executing the scientific method. In ten years time, a citizen science platform with over 100,000 people will beat the performance of professional scientists for the full scientific process.

### 6.3 Working Group Three - Community Engagement

*Reported by Oscar Corcho and Christopher Kyba.* Community engagement is crucial in order to ensure that volunteers continue to contribute to a project, as well as return to both the current project and future projects that may be of interest to them. This working group investigated two challenges as outlined below.

#### A framework and contextual conditions for citizen science projects

**The current status-quo.** Funding agencies have not adapted their funding strategies/calls to account for applying Citizen Science in any type of project. Either a project is submitted as a specific Citizen Science project for a Citizen Science call, or if it is a general project call then there is no clear space for the funding characteristics of Citizen Science projects.

- In some cases (e.g., some Ministries in Germany), there are no possibilities of having follow-up projects, and for a successful citizen science project this may not be adequate (you make a community orphan). The usual term for a research project (3 years) is not enough for building a successful community (which requires between 3 and 6 years).
- Private and public foundations (e.g., Wellcome Trust, Knight Foundation, World Development Bank, UN) in some countries are providing some smaller funding for this type of initiatives.
- There is not infrastructure-type of funding for Citizen Science projects, as it is done for other types of large-scale research infrastructure-related projects (e.g., ESFRI projects).

There are costs associated to Citizen Science projects that are not easy to claim for (e.g., creating a good brochure to reach a large set of citizens - 1 month of work instead of a couple of days needed for a scientific community -, to do scientific outreach/transfer) or that are not easy to get into the accounts of scientific organisations (e.g., problems with audits). How do you pay freelancers or crowdworkers?

Citizen Science projects are generally more risky (e.g., engagement with people may not work and they may not show up). We should be also more open to the fact that experiments may generate less (or more) data than originally expected, or with a different quality to the original expectations. So there must be more room for flexibility (even changing the hypotheses during the project execution), and some more knowledge inside research organisations of the characteristics of Citizen Science projects.

Research organisations have already some budget for scientific outreach (e.g., open days, education activities for secondary schools). However, Citizen Science projects are not normally considered as fundable under this budget, even if they cover some of these needs (creating awareness about Science for Society), but only fundable under the budgets for research projects.

**Challenges.** The following challenges were identified as necessary to overcome in order to advance this aspect of citizen science research:

- Change the funding (and accounting) mechanisms from funding agencies to adapt better to the funding needs for Citizen Science projects (e.g., to be more flexible)
- Expand the collection of organisations that can provide funding for these projects by providing a better explanation of what is done in Citizen Science projects and their potential benefits
- Connect Citizen Science to evidence-based decision making (e.g., fighting against fake news in newspapers)

- Create Citizen Science infrastructure helpdesks inside research organisations or at the national/international level.
- Create specific budgets inside research organisations for funding Citizen Science projects.
- Adapt administrative procedures - e.g. possibilities of paying volunteers for contributions or participation in the project; modalities of payment (paying upfront)
- Appoint Citizen Science coordinators in research organisations, so that they can help scientists to contact and engage citizens.

**Actions.** In the short term, we need to create a clear set of administrative procedures that can be applied by research institutions when paying volunteers for their contributions. To support this, helpdesks/support units could be created for Citizen Science projects at different levels, including online training and appointing Citizen Science coordinators.

In the medium term, there needs to be influence on public funding bodies in order to incorporate Citizen Science into their usual funding streams. In order to drive this, foundations should be engaged to disseminate information about the opportunities that citizen science can offer, and alongside this we should exploit the social corporate responsibility of many large companies as they may become potential funders. Future research should explore how citizen science could fit new business models for publishers, and there should be work to create budgets within organisations for funding citizen science projects. Finally, citizen science project involvement should be aligned to university curricula so that (for example) students can participate in projects and earn credits towards their degree.

Longer term, the group discussed the creation of citizen science ‘wallets’ where volunteers could collect remuneration for their contributions and develop a record of how they’ve helped in various science projects. Additionally, there is the need to create an ESFRI-like structure for large scale projects.

### Upscaling and diversity

**The current status-quo.** As discussed in the diversity section of the ‘Success’ working group, citizen science project participants are not currently representative of societies in general. They are often drawn from higher educated groups. To some extent this reflects self-selection, but this may be because of lack of awareness in other demographic groups. Not all projects collect participant information, so we don’t necessarily have the full picture of who is participating and aggregated statistical information about participant characteristics is not currently available or accessible. Many citizens don’t realise that they could be involved in science projects such as these.

Furthermore, in some projects the data often has very strong clustering - with some areas having a lot of data and others having little or none. Participatory approaches have a very strong bias towards Europe and North America and reflect cultural differences: within Europe there is considerable diversity in the tools that are available, and approaches such as bioblitz [1], (popular in e.g. Germany and UK, but not taking place in e.g. Spain). There are cultural differences particularly between Eastern and Western European countries with regards to participatory research. Even within countries, there are differences in participation rates among the population. In Spain for example, participation varies strongly between northern and southern Spain. Language creates a barrier for expanding participation across European countries and also to countries outside of Europe.

Research questions in projects that are labeled Citizen Science are far more commonly initiated by academic researchers, rather than co-created although this is less of the case in “digital social innovation”. Citizen science projects are however rarely being initiated by

citizens directly. Not all projects are being funded with a clear “closing strategy” for when the project will be handed off to e.g. a foundation, or how it will be properly ended (e.g. final communications, publications, etc.). As such, there is currently duplication of effort around these areas.

- Individual citizen scientists are very seldomly represented at discussions of citizen science
- Access to paywalled articles is strongly differentiated by class
- Participants are not necessarily getting enough feedback and education from the projects they participate in: missing a chance to expand understanding of critical thinking

**Challenges.** Subsequently, there are a number of challenges in terms of upscaling and increasing diversity in projects:

- Attraction of participants: how do you get them to know about your project?
  - Overcoming language barriers
  - There is no one “European media”, so getting messages out is difficult
  - Finding translators and funding for citizen science projects (apps, lesson plans, official communication and feedback)
  - How do you find “gatekeepers” who can promote your project to members of their group?
- Community management is very difficult with an international participant group
  - There can also be cultural issues in how to communicate (e.g. with students versus elderly people)
- If you want to expand a project into a neighboring country, how do you find who you need to contact?
- There is a lack of funding for community management, funders don’t realize that the true cost of a project is much larger than the cost of an app/platform.
- Citizen science funding is often put out in specific calls, and not more generally accepted within disciplinary funding programs
- How could you allow for citizen project initiation, while still dealing with questions of funding oversight?

**Actions.** To address these challenges, we first need to provide tools and resources to citizen scientists so that they can participate more easily in projects. Funding needs to be allocated to community management and towards a strong communication strategy in order to draw in a more diverse set of participants. To help with this further, online science literacy classes could be provided to give volunteers the skills they need in order to help out, while seminars could be given to scientists on how to interact with citizen volunteers. One larger-scale suggestion was to provide European funding for a citizen scientists conference where participants of projects can attend if they have demonstrated a level of participation in a project. There also needs to be efforts to share data among projects, and disseminate different engagement strategies.

In the medium term, this should go further and demographic information needs to be shared across projects so that we have a better understanding of what communities are being engaged. Citizen science practitioners also need to be up-to-date with what digital transformations are impacting citizen science.

Finally, longer term, there needs to be open access to research papers so that all interested parties can find out about the research opportunities and outcomes in citizen science. We discussed the idea of making citizen science part of the junior high school curriculum so that everyone becomes aware of its existence and so that students have the option to participate throughout their lives.

## 6.4 Working Group Four - Linking

*Reported by Andrea Wiggins.* The linking group discussed the complexities involved in linking communities across citizen science projects. Currently, many projects want to link and connect with others in order to enhance the wider citizen science network. However there are currently no incentives for such networking and it is not seen as innovation in citizen science design. There is little funding available for these activities and therefore most attempts generally start from the beginning rather than building on the records of previous attempts.

In the short term, the idea of networking communities within citizen science should be made clear to all projects so that existing and established communities may be used rather than always creating new ones. This should be detailed in the early stages of project formation, including the grant-writing process and project planning. There should also be an emphasis on sharing technology including the tools, user interfaces, data and profiles used within projects, and technologies such as OpenID that are easy to use and already have well documented APIs should be adopted.

In the longer term, there needs to be greater effort to network co-ordinators together. There needs to be an increase in funding allocated towards linking communities, consisting of tasks such as administration, arrangement of workshops and planning strategies. There should also be specific incentivisation mechanisms for linking so that this is not just tacked on to a project but instead planned appropriately. In order to encourage this, a change is required at the grant application stage so that policies and funding opportunities encourage linking to be included in project proposal and if this can be promoted internationally for cross-border sharing then the potential of citizen science will be realised faster.

## 6.5 Working Group Five - Quality

*Reported by Elena Simperl.* Data quality is a matter of primary concern for researchers who use citizen science. There are generally three stages at which data quality typically becomes a concern: the input data (e.g. the collection of objects or the results of a machine learning classifier), the data collected or produced by citizens, and the final results of the task. This working group considered mechanisms that projects employ to ensure quality in the data produced by these projects, and discussed frameworks and mechanisms for managing these - in particular a framework of 18 mechanisms presented by Wiggins et al. (2011) that are based on direct experience and existing surveys [3]. The working group took into consideration sources of error in the data, as well as different stages of the citizen science project process which may introduce such errors. Using this as a starting point for our discussions, we considered what the current debate in this area should be in citizen science.

There are different aspects of data quality that need to be considered. Relevance and trustworthiness are required in order to ensure that valid and reliable data has been collected. There then needs to be a representative presence covered by the data, which may be through the density of coverage (e.g. spatially). There also needs to be information to ensure that we know the degree to which a volunteer followed the research protocol - and therefore there needs to be a process of recording this so that it can be assessed. Finally, the data should be reproducible so that the same distributions of data can be obtained if the same process is followed again.

Common standards may help with ensuring a base level of quality across projects, but it is likely that there are too many different project types in order to develop a reliable standard in this area. As such it may be more appropriate to develop standards around the



tools, question types or methods used, and to also consider the domain specific concerns involved in specific projects as many of these already have existing standards. While some of these may be shared between domains and projects, there may also exist a conflict between what constitutes 'good' data.

Because of the complexity of devising a common benchmark for citizen science data quality, a number of issues were listed as things that need to be considered in order to proceed. In terms of data quality, it must be established exactly how citizen science differs from conventional science - and whether or not this requires a new way of thinking compared to traditional approaches. Furthermore, the different technologies and tools used by different projects will introduce unique data quality problems, and therefore an understanding of these needs to be shared so that project designers may anticipate and prevent potential problems. Noisy data may be caused by inconsistent or inaccurate sensors (either physical, or human), or through ambiguity and bias in the process. This may originate from the task design, where specific labelling requirements or semantics mean that bias is introduced into the process. It is therefore crucial to understand these so that future projects can be designed to alleviate these problems. Because of the impact that the citizens themselves have on the data, we must also consider the impact from using different participatory or engagement frameworks, and how these in turn determine the resulting data quality.

## 6.6 Working Group Six - Manifesto for Citizen Science

*Reported by Paul Groth, Edited by Neal Reeves* A final working group developed a manifesto for the future of Citizen Science, intended to support different stakeholders in producing projects and in publishing scientific results. This document can be viewed at <https://goo.gl/ojuZiR>.

### References

- 1 Cathy Lundmark. *BioBlitz: Getting into Backyard Biodiversity*. BioScience, JSTOR, 53;4, 329–329, 2003.
- 2 Alexandra Swanson, Margaret Kosmala, Chriss Lintott, Robert Simpson, Arfon Smith and Craig Packer. *Snapshot Serengeti, High-Frequency Annotated Camera Trap Images of 40 Mammalian Species in an African Savanna*. Scientific Data, Nature Publishing Group, 2, 150026, 2015.
- 3 Andrea Wiggins, Greg Newman, Robert D. Stevenson and Kevin Crowston. *Mechanisms for Data Quality and Validation in Citizen Science*. Proceedings of the 2011 IEEE Seventh International Conference on e-Science Workshops (ESCIENCEW '11), IEEE Computer Society, Washington, DC, USA, 14-19, 2011.




## Participants

- Lora Aroyo  
Free University Amsterdam, NL
- Alessandro Bozzon  
TU – Delft, NL
- Maria Antonia Brovelli  
Polytechnic University of Milan, IT
- Irene Celino  
CEFRIEL – Milan, IT
- Oscar Corcho  
Polytechnic University of Madrid, ES
- Dominic di Franzo  
Cornell University – Ithaca, US
- Claudia Göbel  
Museum für Naturkunde – Berlin, DE
- Esteban González Guardia  
Technical University of Madrid, ES
- Paul Groth  
Elsevier Labs – Amsterdam, NL
- Lynda Hardman  
CWI – Amsterdam, NL
- Franz Hölker  
IGB – Berlin, DE
- Tomi Kauppinen  
Aalto University, FI
- Christopher Kyba  
GFZ – Potsdam, DE
- Dave Murray-Rust  
University of Edinburgh, GB
- Maurizio Napolitano  
Bruno Kessler Foundation – Trento, IT
- Jasminko Novak  
Hochschule Stralsund, DE
- Christopher Phethean  
University of Southampton, GB
- Marisa Ponti  
University of Göteborg, SE
- Lisa Posch  
GESIS – Cologne, DE
- Gloria Re Calejari  
CEFRIEL – Milan, IT
- Neal Reeves  
University of Southampton, GB
- Sven Schade  
EC Joint Research Centre – Ispra, IT
- Sibylle Schroer  
IGB – Berlin, DE
- Elena Simperl  
University of Southampton, GB
- Alice Verioli  
BSDesign – Milan, IT
- Christopher A. Welty  
Google – New York, US
- Andrea Wiggins  
University of Nebraska – Omaha, US
- Amrapali Zaveri  
Maastricht University, NL



8 Appendix A - Lightning Talk Slides

Alessandro Bozzon, TU Delft




**Dr. Alessandro Bozzon**  
[www.alessandrobozzon.com](http://www.alessandrobozzon.com)

**Assistant Professors @ TU Delft**  
Crowd Computing, Information Retrieval,  
Web Data Management

**Principal Investigator @ AMS**  
Social Urban Data Lab

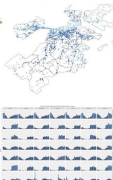

**Research Line Leader @ DDS**  
Social Data, Smart Cities

**Formalise Smart Citizens Initiatives**  
**From Need To Knowledge Model**



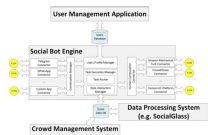
[http://bit.ly/FN2K\\_AMS](http://bit.ly/FN2K_AMS)

**Unlock Urban Knowledge**  
**Social Glass**



[www.socialglass.org](http://www.socialglass.org)

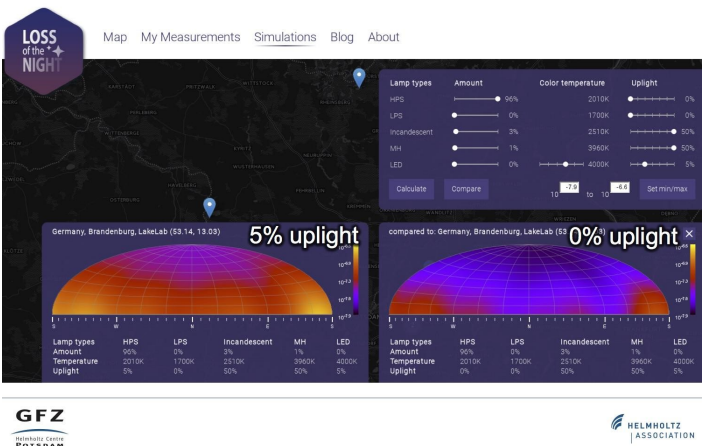
**New ways to engage with citizens**  
**AMS Social Bot**



<http://bit.ly/UrbanChatbotsVision>

Christopher Kyba, GFZ - Potsdam

Christopher Kyba, GFZ Potsdam



Maria Antonia Brovelli, Polytechnic University of Milan



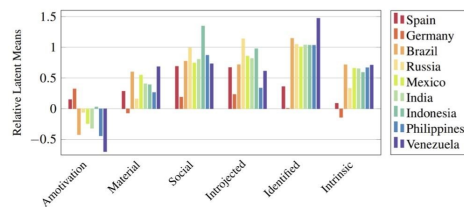
Lisa Posch, GESIS - Köln

gesis  
Leibniz Institute  
for the Social Sciences

## Crowdworker Motivations

Lisa Posch

- The Multidimensional Crowdworker Motivation Scale
- Cross-country and cross-income group comparison of crowdworker motivations



GESIS

paper in progress →



Sven Schade, EC Joint Research Centre

Sven Schade  
@innovatearth



<http://digitalearthlab.jrc.ec.europa.eu>

### Citizen Science and the European Commission



JRC SCIENCE FOR POLICY REPORT

Citizen Engagement in Science and Policy-Making

Two WS on *Citizen Engagement in Science and Policy Making*

Reflections and recommendations across the European Commission

Figuerola-Núñez, S., Corcho, S., Schade, S., Corcho-Pérez, A.

2018



2018 Workshop

**What policy makers think**

- ✓ The Environment Knowledge Community endorsed the outcome of the CS work
- ✓ DG RTD's promotes Open Science and Citizen Science
- ✓ SG now considering possible use of CS in stakeholder consultation

**Development of innovative applications** (mobile or web-based) in different domains addressing citizens' needs



Sibylle Schroer, IGB - Berlin

### Label for the quality of outdoor lighting companies – institutions – villages





Lamp design



Color



Intensity



Energy efficiency



Awareness



Night sky quality



Curfew



Photosensor



STARS+ALL



STARS+ALL



STARS+ALL





Andrea Wiggins, University of Nebraska

## Andrea Wiggins



### Data Quality: Behavior & Process

- iNaturalist data showed taxonomic favoritism/capacity impacted data validation
  - Anecdotally, aesthetics did too!
- Data from mobile app less likely to reach Research Grade
  - Documenting observation vs. crowdsourcing ID?
- Task-technology fit of mobile phones was poor for several taxa



### Projects & Systems: Design & Management

- Training
  - How is it being done?
  - What materials & resources support training? How are they provided & used?
- Evaluation
  - Science products inventory (under review) & field testing for practical implementation
  - Embedded assessment of skill development

Supported in part by NSF Grant CCF 1442668 & USGS Cooperative Agreement G16AC00267

Amrapali Zaveri, Maastricht University

### Accelerating Scientific Discovery through Crowdsourcing

#### Hypothesis

Can crowdsourcing, a combination of *expert* and *non-expert workers* can be used to (1) Design, (2) Discovery and (3) Validate scientific problems?

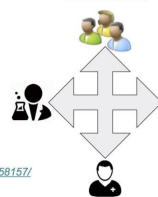
#### Significance

*Developing a new drug from original idea to the launch of a finished product is a complex process which can take 12–15 years and cost in excess of \$1 billion.\**

\*<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3058157/>

@AmrapaliZ

#### Innovation



amrapali.zaveri@maastrichtuniversity.nl

#### Impact



Maastricht University

Tomi Kauppinen, Aalto University

QUESTION

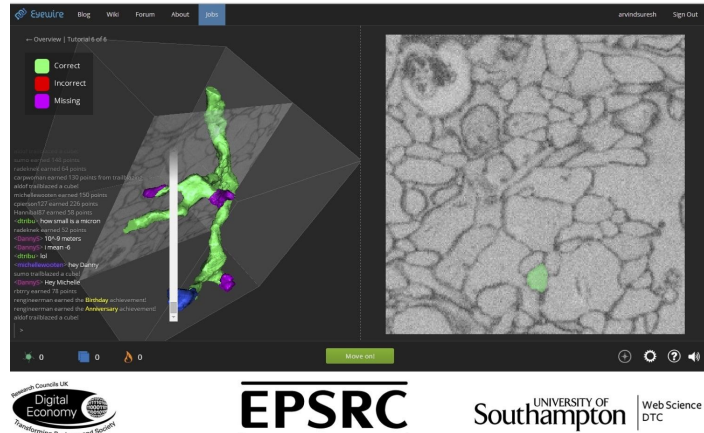
## HOW CAN WE CROWD-SENSE EXPERIENCES?

LINKED EARTH.ORG/EXPERIENCE/NS/

DATE 2017-07-02 WHO TOMI.KAUPPINEN@AALTO.FI

Neal Reeves, University of Southampton

## Neal Reeves – Sociality and Communication in Virtual Citizen science



Irene Celino, CEFRIEL

## Give and Take in Citizen Science

*Take* from citizen scientists

**Gathering** information about their surrounding environment (e.g. urban POIs in <http://bit.ly/urbanopolis>)

**Ranking/filtering** information based on "popularity" (e.g. cultural heritage in <http://bit.ly/indomilando>)

**Validation** of scientific data (e.g. Land Cover Game at <http://bit.ly/foss4game>)

**Classification** of images (e.g. ISS photos in <https://www.nightknights.eu/>)

*Give* back to citizen scientists

**Fun** (e.g. playing "pastime" games)

**Engagement** (e.g. competition with leaderboard and prizes)

**Recognition** (e.g. being part of a community, of something "bigger")

**KNOWLEDGE, INSIGHTS, LEARNING**

About their surrounding environment

About science and "grand challenges"

About themselves (e.g. quantified self, awareness, self-consciousness)

Human Computation, Gamification, Games with a Purpose, Serious Games, Data Analytics, ...

Irene Celino – [irene.celino@cefriel.com](mailto:irene.celino@cefriel.com) – [iricelino.org/publications](http://iricelino.org/publications) – [ninja-riders.eu](http://ninja-riders.eu)

DIGITAL, SEMINAR ON CITIZEN SCIENCE - JULY 2017

**Cefriel**

Gloria Re Calegari, CEFRIEL



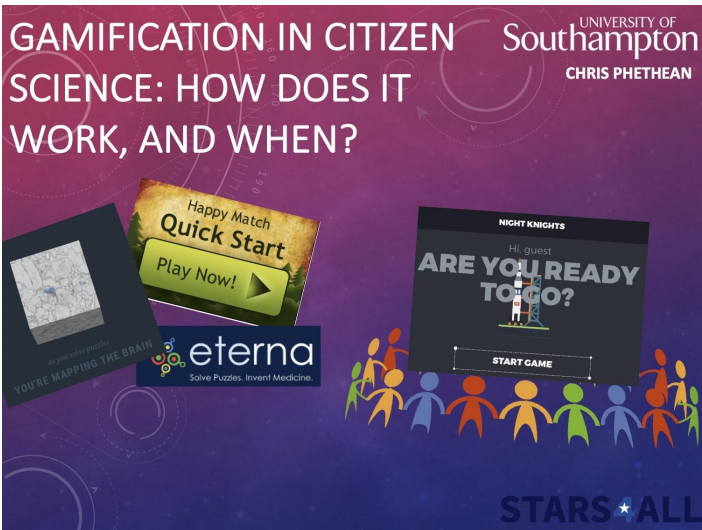
Gloria Re Calegari

**Cefriel**  
POLITECNICO DI MILANO

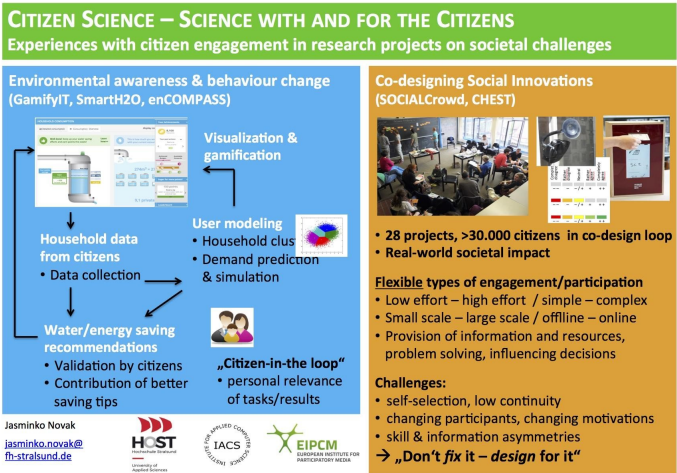




Chris Phethean, University of Southampton



Jasminko Novak, Hochschule Stralsund



# Malware Analysis: From Large-Scale Data Triage to Targeted Attack Recognition

Edited by

Sarah Zennou<sup>1</sup>, Saumya K. Debray<sup>2</sup>, Thomas Dullien<sup>3</sup>, and  
Arun Lakhotia<sup>4</sup>

- 1 Airbus – Suresnes, FR, [sarah.zennou@airbus.com](mailto:sarah.zennou@airbus.com)
- 2 University of Arizona – Tucson, US, [debray@cs.arizona.edu](mailto:debray@cs.arizona.edu)
- 3 Google Switzerland – Zürich, CH, [thomas.dullien@gmail.com](mailto:thomas.dullien@gmail.com)
- 4 University of Louisiana – Lafayette, US, [arun@louisiana.edu](mailto:arun@louisiana.edu)

---

## Abstract

This report summarizes the program and the outcomes of the Dagstuhl Seminar 17281, entitled “Malware Analysis: From Large-Scale Data Triage to Targeted Attack Recognition”. The seminar brought together practitioners and researchers from industry and academia to discuss the state-of-the art in the analysis of malware from both a big data perspective and a fine grained analysis. Obfuscation was also considered. The meeting created new links within this very diverse community.

**Seminar** July 9–14, 2017 – <http://www.dagstuhl.de/17281>

**1998 ACM Subject Classification** formal methods, program analysis

**Keywords and phrases** big data, executable analysis, machine learning, malware, obfuscation, reverse engineering

**Digital Object Identifier** 10.4230/DagRep.7.7.44

## 1 Executive Summary

*Sarah Zennou*

*Saumya K. Debray*

*Thomas Dullien*

*Arun Lakhotia*

**License** © Creative Commons BY 3.0 Unported license  
© Sarah Zennou, Saumya K. Debray, Thomas Dullien, and Arun Lakhotia

As a follow-up on the previous Dagstuhl Seminar 14241 on the analysis of binaries, the interest in attending this new seminar was very high. The attendance was very diverse, almost half academics and half practitioners.

Talks were arranged by topics and each day ended with an open discussion on one of the three topics: machine learning, obfuscation and practitioners’ needs.

Considering the given talks, it appears that the challenges in the realm of general binary analysis have not changed considerably since the last gathering. However, the balance between the topics shows that the academic interest is now more focused on machine learning than on obfuscation. On the contrary practitioners exhibited examples showing that the sophistication level of obfuscations has tremendously increased during this last years.

The open discussions were the most fruitful part of the seminar. The discussions enabled the academics to ask practitioners about the hypotheses that are relevant to build models



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Malware Analysis: From Large-Scale Data Triage to Targeted Attack Recognition, *Dagstuhl Reports*, Vol. 7, Issue 7, pp. 44–53

Editors: Sarah Zennou, Saumya K. Debray, Thomas Dullien, and Arun Lakhotia



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

for their analyses and the problems they face in their daily work. The practitioners gained awareness of the automated tools and techniques that they can expect to see emerge from research labs.

These informal exchanges will be gathered into a separate document and spread to the academic community.

Finally please note that not all people who presented have submitted their abstracts due to the sensitive nature of the content and/or the organization that the participants work for

## 2 Table of Contents

### Executive Summary

*Sarah Zennou, Saumya K. Debray, Thomas Dullien, and Arun Lakhotia* . . . . . 44

### Overview of Talks

Characterizing the Strength of Code Obfuscation Against Auto-MATED Attacks  
*Sebastian Banescu* . . . . . 47

Operation Avalanche: Not your average botnet take down  
*Thomas Barabosch* . . . . . 47

Deobfuscation: semantic analysis to the rescue  
*Sébastien Bardin* . . . . . 47

How Professional Hackers Understand Protected Code while Performing Attack Tasks  
*Bjorn De Sutter* . . . . . 48

Similarity Analysis in Verona & IMDEA  
*Roberto Giacobazzi, Mila Dalla Preda, and Niccolò Marastoni* . . . . . 48

The many Dimensions of Relationships  
*Tim Kornau-von Bock und Polach* . . . . . 49

A morphological approach to detect code similarities and to analyse x86 binaries  
*Jean-Yves Marion* . . . . . 49

Side-Channel Based Detection of Malicious Software  
*J. Todd McDonald* . . . . . 49

On the Availability of High-Quality Malware Data Sets  
*Daniel Plohmann* . . . . . 50

Measuring and Defeating Anti-Instrumentation-Equipped Malware  
*Mario Polino and Stefano Zanero* . . . . . 50

Formally Proved Security of Assembly Code Against Power Analysis  
*Pablo Rauzy, Sylvain Guilley, and Zakaria Najm* . . . . . 51

Advanced Semantics Based Binary Code Similarity Comparison  
*Dinghao Wu* . . . . . 51

Uroboros: Reassembleable Disassembling  
*Dinghao Wu* . . . . . 52

On The Unreasonable Effectiveness of Boolean SAT Solvers  
*Ed Zulkoski* . . . . . 52

**Participants** . . . . . 53

### 3 Overview of Talks

#### 3.1 Characterizing the Strength of Code Obfuscation Against Auto-MATEd Attacks

*Sebastian Banescu (BMW Group ITZ, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Sebastian Banescu

**Joint work of** Sebastian Banescu, Christian Collberg, Vijay Ganesh, Zack Newsham, Alexander Pretschner  
**Main reference** Sebastian Banescu, Christian Collberg, Vijay Ganesh, Zack Newsham, Alexander Pretschner, “Code obfuscation against symbolic execution attacks”, in Proc. of the 32nd Annual Conf. on Computer Security Applications, pp. 189–200, ACM, 2016.

**URL** <https://doi.org/10.1145/2991079.2991114>

There exist several obfuscation transformations and there are many ways in which these can be combined. This talk presents a method that is meant to aid a software developer in deciding which obfuscation transformations to employ in order to protect his/her own software against known automated man-at-the-end (MATE) attacks, for a certain amount of time, against attackers having a given resource cap. A case-study based on a symbolic execution attack is used to instantiate this method and to show its utility.

#### 3.2 Operation Avalanche: Not your average botnet take down

*Thomas Barabosch (Fraunhofer FKIE – Bonn, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Thomas Barabosch

The Avalanche network was one of the major cyber crime platforms offering services such as traffic obfuscation and money mule management. It hosted infamous malware families like Goznym, Matsnu and Urlzone. In late 2016, several international partners from law enforcement, industry and academia took down this network in a coordinated operation. Fraunhofer FKIE participated in this take down and carried out the technical analysis. I discuss the work we did prior to the take down, e.g. how we analyzed 130 TB of malware traffic, how we tracked 10+ malware families at once and what kind of obfuscations we faced in practice. This talk should be beneficial to participants since it gives insights into a four year long operation against a major cyber crime network.

#### 3.3 Deobfuscation: semantic analysis to the rescue

*Sébastien Bardin (CEA LIST, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Sébastien Bardin

**Joint work of** Sébastien Bardin, Robin David, Jean-Yves Marion  
**Main reference** Sébastien Bardin, Robin David, Jean-Yves Marion: “Backward-Bounded DSE: Targeting Infeasibility Questions on Obfuscated Codes”, in Proc. of the 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22–26, 2017, pp. 633–651, IEEE Computer Society, 2017.  
**URL** <http://dx.doi.org/10.1109/SP.2017.36>

Malware comprehension, i.e. deep malware analysis in order to understand their behavior, may be necessary in case of infection of a critical asset. Yet, in the case of targeted attack,

this analysis is highly challenging due to the strong obfuscation methods used for malware protection, and there is a clear need for more sophisticated and automated tools than currently available syntactic and dynamic approaches. In this talk, we present how semantic analysis coming from source-level safety analysis can be adapted to the context of binary-level deobfuscation, as well as their strengths and limitations.

### 3.4 How Professional Hackers Understand Protected Code while Performing Attack Tasks

*Bjorn De Sutter (Ghent University, BE)*

**License**  Creative Commons BY 3.0 Unported license  
© Bjorn De Sutter

**Joint work of** Mariano Ceccato, Paolo Tonella, Cataldo Basile, Bart Coppens, Bjorn De Sutter, Paolo Falcarin, Marco Torchiano


**Main reference** Mariano Ceccato, Paolo Tonella, Cataldo Basile, Bart Coppens, Bjorn De Sutter, Paolo Falcarin, Marco Torchiano: “How professional hackers understand protected code while performing attack tasks”, in Proc. of the 25th International Conference on Program Comprehension, ICPC 2017, Buenos Aires, Argentina, May 22-23, 2017, pp. 154–164, IEEE Computer Society, 2017.

**URL** <http://dx.doi.org/10.1109/ICPC.2017.2>

Code protections aim at blocking (or at least delaying) reverse engineering and tampering attacks to critical assets within programs. Knowing the way hackers understand protected code and perform attacks is important to achieve a stronger protection of the software assets, based on realistic assumptions about the hackers’ behaviour. However, building such knowledge is difficult because hackers can hardly be involved in controlled experiments and empirical studies. The FP7 European project ASPIRE has given the project researchers the unique opportunity to have access to the professional penetration testers employed by the three industrial partners. In particular, we have been able to perform a qualitative analysis of three reports of professional penetration test performed on protected industrial code. Our qualitative analysis of the reports consists of open coding, carried out by 7 annotators and resulting in 459 annotations, followed by concept extraction and model inference. We identified the main activities: understanding, building attack, choosing and customizing tools, and working around or defeating protections. We built a model of how such activities take place. We used such models to identify a set of research directions for the creation of stronger code protections.

### 3.5 Similarity Analysis in Verona & IMDEA

*Roberto Giacobazzi (University of Verona, IT), Mila Dalla Preda, and Niccolò Marastoni*

**License**  Creative Commons BY 3.0 Unported license  
© Roberto Giacobazzi, Mila Dalla Preda, and Niccolò Marastoni

We present the problems and main challenges in automated similarity analysis of high-level code. After an introduction to abstract interpretation and its precision as complete abstractions, we introduce the REHA, a tool for similarity analysis of Android applications. REHA scales well over large code. The talk ends with a discussion on the open questions and future developments in automated similarity analysis in malware detection and early threat detection.

### 3.6 The many Dimensions of Relationships

*Tim Kornau-von Bock und Polach (Google Switzerland – Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license  
© Tim Kornau-von Bock und Polach

**Joint work of** Tim Kornau-von Bock und Polach, Bruno Montalto

Function based similarity, detections, challenges and open questions.

In this talk the current and past research challenges about executable and function similarity will be discussed. We will discuss scale of executable / function similarity search at Google and its applicability to malware. We will demonstrate use cases from an executable and function similarity perspective to show where such a system is relevant in practice. We will highlight some future directions of research and also demonstrate failures during the development of this system.

### 3.7 A morphological approach to detect code similarities and to analyse x86 binaries

*Jean-Yves Marion (LORIA & INRIA – Nancy, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Jean-Yves Marion

Binary code analysis is a complex process which can be performed nowadays only by skilled cybersecurity experts whose workload just keeps increasing. Uses cases include vulnerabilities detection, testing, clustering and classification, malware analysis, etc... We develop a tool named Cyber-Detect, which is based on the reconstruction of an high level semantics for the binary code. Control flow graphs provide a fair level of abstraction to deal with the binary codes they represent. After applying some graph rewriting rules to normalize these graphs, our software tackles the subgraph search problem in a way which is both efficient and convenient for that kind of graphs. This technique is described as morphological analysis as it recognizes the whole shape of the malware. That being said, some pitfalls still need to be considered. First of all, the output can only get as good as the input data. And it is known that static disassembly cannot produce the perfect control flow graph since this problem is undecidable. As a matter of facts, malware heavily use obfuscation techniques such as opaque predicates to hide their payloads and confuse analyses. Dynamic analysis should then be used along with static disassembly to combine their strengths. Another dangerous pitfall feared by every expert is the so-called false positives rate : false alarms that make them waste indeed a precious time assessing the reality of the threat.

### 3.8 Side-Channel Based Detection of Malicious Software

*J. Todd McDonald (University of South Alabama – Mobile, US)*

**License** © Creative Commons BY 3.0 Unported license  
© J. Todd McDonald

**Joint work of** J. Todd McDonald, Joel A. Dawson, Patrick H. Luckett, Lee M. Hively

Timing and power are two potential indicators for the presence of malicious software execution. There is nascent research in using such indicators for training of intrusion or anomaly detection

systems. Of potential interest are non-linear methods that have a theoretic foundation and are adaptable to a wide range of anomaly detection problems, not just cyber. A nonlinear approach based on Taken's time delay embedding theorem has seen some early success in detecting rootkit execution. The approach relies on phasespace representation of the dynamics of a computer system and leverages graph-based difference to indicate a change in normal state. The technique is different from other traditional approaches that rely on statistical analysis methods based on machine learning and may provide a practical out-of-band approach for early indication of zero-day threats.

### 3.9 On the Availability of High-Quality Malware Data Sets

*Daniel Plohmann (Fraunhofer FKIE – Bonn, DE)*


**License**  Creative Commons BY 3.0 Unported license  
© Daniel Plohmann

In this presentation, the lack of availability of comprehensive, accurately labelled malware data sets is thematized. First, a typical sequence of steps an analyst may use during malware identification is outlined. This is then used to further motivate the knowledge fragmentation that is reportedly perceived by many malware analysts. Next, a selection of popular malware data sets is quickly examined, showing that there is no coherent corpus focusing on unpacked malware samples.

We then present our approach for a manually curated, high-quality malware corpus. This corpus focuses on providing coverage for as many distinct families as possible, while limiting itself to single unpacked samples for a given variant. The method of collaborative collection and inventorization is explained, along with a short evaluation of the current contents.

### 3.10 Measuring and Defeating Anti-Instrumentation-Equipped Malware

*Mario Polino (Polytechnic University of Milan, IT) and Stefano Zanero (Polytechnic University of Milan, IT)*

**License**  Creative Commons BY 3.0 Unported license  
© Mario Polino and Stefano Zanero

**Joint work of** Mario Polino, Andrea Continella, Sebastiano Mariani, Stefano D'Alessio, Lorenzo Fontana, Fabio Gritti, and Stefano Zanero

**Main reference** Mario Polino, Andrea Continella, Sebastiano Mariani, Stefano D'Alessio, Lorenzo Fontana, Fabio Gritti, Stefano Zanero: "Measuring and Defeating Anti-Instrumentation-Equipped Malware", in Proc. of the Detection of Intrusions and Malware, and Vulnerability Assessment - 14th International Conference, DIMVA 2017, Bonn, Germany, July 6-7, 2017, Proceedings, Lecture Notes in Computer Science, Vol. 10327, pp. 73–96, Springer, 2017.

**URL** [https://doi.org/10.1007/978-3-319-60876-1\\_4](https://doi.org/10.1007/978-3-319-60876-1_4)

Malware authors constantly develop new techniques in order to evade analysis systems. Previous works addressed attempts to evade analysis by means of anti-sandboxing and anti-virtualization techniques, for example proposing to run samples on bare-metal. However, state-of-the-art bare-metal tools fail to provide richness and completeness in the results of the analysis. In this context, Dynamic Binary Instrumentation (DBI) tools have become popular in the analysis of new malware samples because of the deep control they guarantee over the instrumented binary. As a consequence, malware authors developed new techniques,



called anti-instrumentation, aimed at detecting if a sample is being instrumented. We propose a practical approach to make DBI frameworks more stealthy and resilient against anti-instrumentation attacks. We studied the common techniques used by malware to detect the presence of a DBI tool, and we proposed a set of countermeasures to address them. We implemented our approach in Arancino, on top of the Intel Pin framework. Armed with it, we perform the first large-scale measurement of the anti-instrumentation techniques employed by modern malware. Finally, we leveraged our tool to implement a generic unpacker, showing some case studies of the anti-instrumentation techniques used by known packers.

### 3.11 Formally Proved Security of Assembly Code Against Power Analysis

*Pablo Rauzy (University of Paris VIII, FR), Sylvain Guilley, and Zakaria Najm*

**License** © Creative Commons BY 3.0 Unported license

© Pablo Rauzy, Sylvain Guilley, and Zakaria Najm

**Main reference** Pablo Rauzy, Sylvain Guilley, Zakaria Najm: “Formally proved security of assembly code against power analysis - A case study on balanced logic”, J. Cryptographic Engineering, Vol. 6(3), pp. 201–216, 2016.

**URL** <http://dx.doi.org/10.1007/s13389-015-0105-2>

In his keynote speech at CHES 2004, Kocher advocated that side-channel attacks were an illustration that formal cryptography was not as secure as it was believed because some assumptions (e.g., no auxiliary information is available during the computation) were not modeled. This failure is caused by formal methods’ focus on models rather than implementations. In this paper we present formal methods and tools for designing protected code and proving its security against power analysis. These formal methods avoid the discrepancy between the model and the implementation by working on the latter rather than on a high-level model. Indeed, our methods allow us (a) to automatically insert a power balancing countermeasure directly at the assembly level, and to prove the correctness of the induced code transformation; and (b) to prove that the obtained code is balanced with regard to a reasonable leakage model, and we show how to characterize the hardware to use the resources which maximize the relevancy of the model. The tools implementing our methods are then demonstrated in a case study in which we generate a provably protected PRESENT implementation for an 8-bit AVR smartcard.

### 3.12 Advanced Semantics Based Binary Code Similarity Comparison

*Dinghao Wu (Pennsylvania State University – State College, US)*

**License** © Creative Commons BY 3.0 Unported license

© Dinghao Wu

**Joint work of** Dinghao Wu, Jiang Ming, Dongpeng Xu, Yufei Jiang

Binary code comparison has many applications in malware analysis and software engineering. Previous semantics-based work focuses on extracting and comparing of semantics of basic blocks. This block-centric method has limitations on optimizations and obfuscations that merge or split basic blocks. To address the limitations, I will present a method that uses Equivalence Checking of System Call Aligned Segments. With two sequences of instructions, obtained from two traces with the same input on two programs, we first align the system

calls involved. With two aligned system calls (one in each sequence, or trace), we slice back on the data and control dependences from the call sites on their parameters. After slicing, we get two trace segments with only instructions related to these two system calls. Then we compute their weakest preconditions, and compare their equivalence using a constraint solver. Our experiments show quite promising results.

### 3.13 Uroboros: Reassembleable Disassembling

*Dinghao Wu (Pennsylvania State University – State College, US)*

**License** © Creative Commons BY 3.0 Unported license

© Dinghao Wu

**Joint work of** Dinghao Wu, Shuai Wang, Pei Wang

There are many disassemblers, but no one is able to disassemble an executable in a way that can be reassembled back to an executable. Uroboros is a tool we built for Reassembleable Disassembling. In this talk, I will discuss the challenges and methods to disassemble an executable into an assembly that can be reassembled it back into an executable, in a fully automated manner, and present our results.

### 3.14 On The Unreasonable Effectiveness of Boolean SAT Solvers

*Ed Zulkoski (University of Waterloo, CA)*

**License** © Creative Commons BY 3.0 Unported license

© Ed Zulkoski

**Joint work of** Ed Zulkoski, Vijay Ganesh, Jimmy Liang, Saeed Nejati, Zack Newsham

Modern conflict-driven clause-learning (CDCL) Boolean SAT solvers routinely solve very large industrial SAT instances in relatively short periods of time. This phenomenon has stumped both theoreticians and practitioners since Boolean satisfiability is an NP-complete problem widely believed to be intractable. It is clear that these solvers somehow exploit the structure of real-world instances. However, to-date there have been few results that precisely characterize this structure or shed any light on why these SAT solvers are so efficient.

In this talk, I will present results that provide a deeper empirical understanding of why CDCL SAT solvers are so efficient, which may eventually lead to a complexity-theoretic result. Our results can be divided into two parts. First, I will talk about structural parameters that can characterize industrial instances and shed light on why they are easier to solve even though they may contain millions of variables. Second, I will talk about internals of CDCL SAT solvers, and describe why they are particularly suited to solve industrial instances.

## Participants

- Radoniaina  
Andriatsimandefitra  
Rennes, FR
- Sebastian Banescu  
BMW Group ITZ, DE
- Thomas Barabosch  
Fraunhofer FKIE – Bonn, DE
- Sébastien Bardin  
CEA LIST, FR
- Konstantin Berlin  
SOPHOS – Fairfax, US
- Paul Black  
Federation University Australia –  
Mount Helen, AU
- Cory Cohen  
Carnegie Mellon University –  
Pittsburgh, US
- Christian Collberg  
University of Arizona –  
Tucson, US
- Sophia D’Antoine  
Trail of Bits Inc. – New York, US
- Mila Dalla Preda  
University of Verona, IT
- Robin David  
Quarkslab, FR
- Bjorn De Sutter  
Ghent University, BE
- Saumya K. Debray  
University of Arizona –  
Tucson, US
- Thomas Dullien  
Google Switzerland – Zürich, CH
- Roberto Giacobazzi  
University of Verona, IT
- Yuan Xiang Gu  
Irdeto – Ottawa, CA
- Tim Kornau-von Bock und  
Polach  
Google Switzerland – Zürich, CH
- Arun Lakhotia  
University of Louisiana –  
Lafayette, US
- Colas Le Guernic  
Direction Generale de  
l’Armement – Rennes, FR
- Jean-Yves Marion  
LORIA & INRIA – Nancy, FR
- Marion Marschalek  
G DATA Advanced Analytics  
GmbH – Bochum, DE
- J. Todd McDonald  
University of South Alabama –  
Mobile, US
- Xavier Mehrenberger  
Airbus – Suresnes, FR
- Michael Meier  
Universität Bonn, DE
- Bogdan Mihaila  
Synopsys Finland OY –  
Helsinki, FI
- Craig Miles  
Assured Information Security –  
Portland, US
- Asuka Nakajima  
NTT – Tokyo, JP
- Daniel Plohmann  
Fraunhofer FKIE – Bonn, DE
- Mario Polino  
Polytechnic University of  
Milan, IT
- Pablo Rauzy  
University of Paris VIII, FR
- Raphael Rigo  
Airbus – Suresnes, FR
- Radwan Shushane  
Columbus State University, US
- Natalia Stakhanova  
University of New Brunswick at  
Fredericton, CA
- Ryan Stortz  
Trail of Bits Inc. – New York, US
- Dinghao Wu  
Pennsylvania State University –  
State College, US
- Yves Younan  
Cisco Systems Canada Co. –  
Toronto, CA
- Stefano Zanero  
Polytechnic University of  
Milan, IT
- Sarah Zennou  
Airbus – Suresnes, FR
- Ed Zulkoski  
University of Waterloo, CA



# From Observations to Prediction of Movement

Edited by

Mark Birkin<sup>1</sup>, Somayeh Dodge<sup>2</sup>, Brittany Terese Fasy<sup>3</sup>, and  
Richard Philip Mann<sup>4</sup>

1 University of Leeds, GB, [m.h.birkin@leeds.ac.uk](mailto:m.h.birkin@leeds.ac.uk)

2 University of Minnesota – Minneapolis, US, [sdodge@umn.edu](mailto:sdodge@umn.edu)

3 Montana State University – Bozeman, US, [brittany@cs.montana.edu](mailto:brittany@cs.montana.edu)

4 University of Leeds, GB, [r.p.mann@leeds.ac.uk](mailto:r.p.mann@leeds.ac.uk)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17282 “From Observations to Prediction of Movement”. This seminar brought together researchers from Animal Behaviour, GIS, Computational Geometry, Data Science and other fields to exchange insights from these diverse fields. Presentations focused both on outstanding practical questions, as well as on fundamental mathematical and computational tools.

**Seminar** July 9–14, 2017 – <http://www.dagstuhl.de/17282>

**1998 ACM Subject Classification** I. Computing Methodologies, I.3 Computer Graphics, I.3.5 Computational Geometry and Object Modeling, I.5 Pattern Recognition, I.6 Simulation & Modelling, J.3 Life Sciences, and J.4 Social & Behavioral Sciences with one of the following sub-categories: Boundary representations Geometric algorithms, languages, and systems Physically based modeling

**Keywords and phrases** trajectory analysis, computational geometry and topology, GIS, animal movement, prediction, home range

**Digital Object Identifier** 10.4230/DagRep.7.7.54

## 1 Executive Summary

*Mark Birkin*

*Somayeh Dodge*

*Brittany Terese Fasy*

*Richard Philip Mann*

**License** © Creative Commons BY 3.0 Unported license

© Mark Birkin, Somayeh Dodge, Brittany Terese Fasy, and Richard Philip Mann

Dagstuhl Seminar 17282 took place at Schloss Dagstuhl from 9 to 14 July 2017. We had 29 participants and nine invited talks. The main theme of this seminar was the analysis and prediction of movement trajectories. In particular, we focused on the study of movement patterns of individuals, and the interactions of moving agents with each other and with the environment.

## Themes

Movement analysis is key to understanding the underlying mechanisms of dynamic processes. Movement occurs in *space* and *time* across *multiple scales* and through an embedding *context* that influence how entities move. The importance of spatiotemporal aspect of movement



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

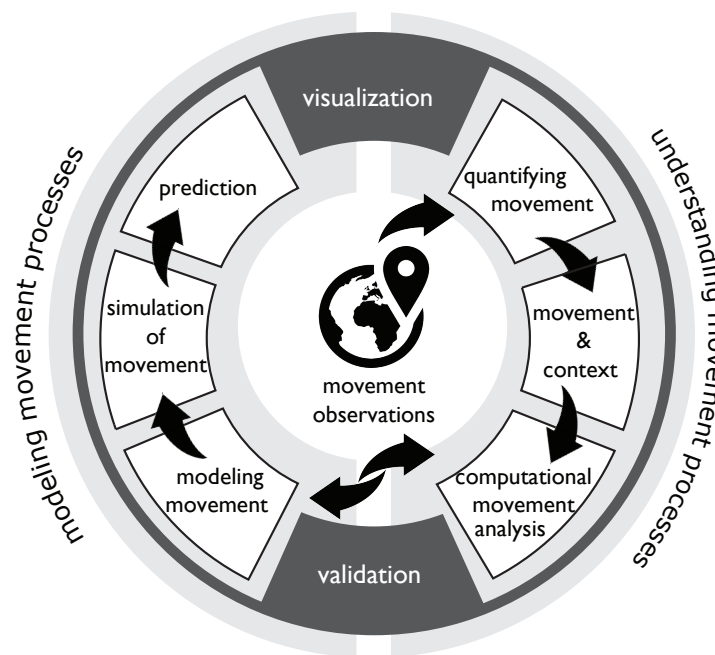
From Observations to Prediction of Movement, *Dagstuhl Reports*, Vol. 7, Issue 7, pp. 54–71

Editors: Mark Birkin, Somayeh Dodge, Brittany Terese Fasy, and Richard Philip Mann



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Movement research continuum.

has attracted a wide range of studies. Analysis of movement trajectories is a core element of Movement Ecology in Biology, as well as being important across disciplines as diverse as Geographic Information Sciences (GIS), Transportation, Criminology, Epidemiology, Computer Gaming, and Phylogenetics. Development of efficient algorithms for analyzing and predicting will be vital to realizing the hopes for new generation smart transport systems and smart cities. Furthermore, naturally generated trajectories provide a fascinating context for mathematical and computational study of Geometry and Stochastic Processes.

A trajectory is a time-stamped sequence of locations, representing the movement of entities in space and time. Trajectories are often created by sampling GPS locations, but they can also originate from RFID tags, video, or radar analysis. Time-series of locations can also be associated with other co-temporal data, such as pressure recordings for avian or aquatic animals, activity sensors and accelerometers to measure energy expenditure, or the myriad time-stamped data recorded by modern smartphones alongside GPS locations.

The study of movement involves development of concepts and methods to transform movement observations (trajectory data) to knowledge of the behavior of moving phenomena under known conditions. This knowledge is then used to calibrate simulation models to predict movement and behavioral responses in varying environmental conditions. Figure 1 illustrates a continuum encapsulating fundamental areas of movement research for (1) understanding movement processes through trajectory representation and computational movement analysis (the right side of Figure 1); and (2) modeling behavior of moving phenomena and prediction of their responses to environmental changes through modeling and simulation approaches (the left side of Figure 1). These two processes are tightly connected and feed into each other, often through a validation procedure on the basis of real trajectory observations.

During recent years computational movement analysis tools for trajectory data have been developed within the areas of GIScience and algorithms. Analysis objectives include clustering, similarity analysis, trajectory segmentation into characteristic sub-trajectories,

finding movement patterns like flocking, and relating patterns to context, and several others. Since these computations are mostly spatial, algorithmic solutions have been developed in the areas of computational geometry and GIScience. The basic analysis tasks for trajectory data are by now comparatively well understood and efficient algorithms have been developed to perform computational movement analysis. However, to be truly effective and to have real-world impact, trajectory analysis has to move beyond ‘understanding movement’ and tackle substantially more involved questions in ‘modeling, simulation, and prediction’ of movement responses to a changing environment or as results of (social) interactions.

Simultaneously, in the area of ecology the study of motion of animals has also become a topic of increasing interest. Many animal species move in groups, with or without a specific leader. The motivation for motion can be foraging, escape from predators, changing climate, or it can be unknown. The mode of movement can be determined by social interactions, energy efficiency, possibility of discovery of resources, and of course the natural environment. The more fascinating aspects of ecology include interaction between entities and collective motion. These are harder to grasp in a formal manner, needed for modelling and automated analysis. The basic analysis tasks for trajectory data are by now comparatively well understood and efficient algorithms have been developed to perform them. However, to be truly effective and have real-world impact, trajectory analysis has to move beyond these basic tasks and tackle substantially more involved questions, prime examples being (social) interaction and collective motion.

### Research Approach and Questions Addressed

Trajectories are mathematical objects with geostatistical properties. Movement is a process that occurs as a response to the state of a moving entity across multiple spatial and temporal scales. The state and resulting behavior of moving entities determine the characteristics and capacities of movement (e.g., speeds, directions, accelerations, path sinuosity), which are highly influenced by interaction with environment, geographic context, and other moving entities. Internal properties of the moving agent such as its propensity to explore, or its power and size, typically distinguish the trajectory from that of other agents. As such no element of a trajectory can truly be independent of its other parts. Therefore, we take a view of trajectory analysis that emphasizes the treatment of the whole trajectory as a unit, rather than a series of moment-by-moment steps.

Trajectories are generated by some underlying process, which is typically assumed to integrate both stochastic elements (such as Brownian motion) and more deterministic interactions between the moving agent and the external world. Many research questions can be posed about such processes, but in this seminar we will focus primarily on identifying the forms of interaction, both with other moving agents and with environmental stimuli, and on predicting the characteristics of future trajectories.

In the seminar, we explored the following questions:

- To what extent movement observations convey information on the underlying *behavior* of moving phenomena?
- How susceptible are behaviors of moving agents to environmental changes?
- To what extent changes in the behavior of moving phenomena indicate changes in the environment?
- What does it mean to *predict* a trajectory? Should we focus on predicting the spatial locations or the geometric properties?
- How can we assess a predictive model?
- How can computational geometry help movement prediction?

- What characteristics of motion are indicative of specific trajectory generating processes, and how can we compute these efficiently?
- What is the role of time in trajectory analysis? Where can we analyze the shapes of paths independently of the time stamps and where are these vital to understanding the underlying mechanisms?
- Can we build a classification of trajectory generating mechanisms and associated trajectory properties, such as navigation by waypoints, explorative foraging
- What is the *home range*? Can we have a concrete definition for home range or activity space?
- What is a *collective*?
- Can we make algorithms that work across scales?
- To what extent *goal oriented movement* can be inferred from *local movement patterns*?

## 2 Table of Contents

### Executive Summary

*Mark Birkin, Somayeh Dodge, Brittany Terese Fasy, and Richard Philip Mann* . . . 54

### Overview of Talks

Calibrating Agent Based Models as Multiple Scales <i>Sean Ahearn</i> . . . . .	60
Analysing and Predicting Movement – A Computational Geometry Perspective <i>Maike Buchin</i> . . . . .	60
Using Time Geography to Model Movement in Three Physical Dimensions <i>Urska Demšar</i> . . . . .	60
Using Prediction to Explore the Mechanisms and Consequences of Social Life <i>Damien Farine</i> . . . . .	61
From Fish to Worms: Spatial Cognition, Movement and Postures in Three-Dimensions <i>Robert Holbrook</i> . . . . .	61
Scalable Methods for Modelling Movement Patterns: from Areal to Road Segment Levels <i>Robin Lovelace</i> . . . . .	62
The Moving Across Places Study (MAPS): Public Transit, Physical Activity and Walking Route Choice <i>Harvey J. Miller</i> . . . . .	62
Random Trajectories in Movement Ecology: A Path to Crossing Scales in Movement Ecology? <i>Kamran Safi</i> . . . . .	63
On Language for Observation & Prediction <i>Jack Snoeyink</i> . . . . .	63
Collective Motion: More than the Sum of its Parts <i>Zena Wood</i> . . . . .	64

### Working groups

Computational Topology and Movement Data <i>Kevin Buchin, Maike Buchin, Brittany Terese Fasy, Kristine Pelatt, and Carola Wenk</i>	64
Defining Axes and Metrics to Characterise Collective Motion <i>Somayeh Dodge, Urska Demšar, Jed Long, Andrea Perna, Alexander Szorkovszky, Johan van de Koppel, and Zena Wood</i> . . . . .	65
Multi-scale Movement Modeling <i>Somayeh Dodge, Kevin Buchin, Urska Demšar, Harvey J. Miller, Kristine Pelatt, Alexander Szorkovszky, and Johan van de Koppel</i> . . . . .	66
Using and Explaining Non-Traditional Metrics in Biology Publications <i>Damien Farine, Robert Holbrook, Richard Philip Mann, Andrea Perna, and Kamran Safi</i> . . . . .	67



Learning Connections Between Landscape and Trajectories from Recorded Data  
*Richard Philip Mann, Maike Buchin, Robert Holbrook, and Nicholas Ouellette . . .* 67

Potential Applications of Ecology on Transport and the Implications on Policy  
*Samuel A. Micka, Mark Birkin, Maarten Löffler, Robin Lovelace, Richard Philip Mann, Kathleen Stewart, and Carola Wenk . . . . .* 68

Formalizing the Notions of “Activity Spaces” and “Homeranges”: Mathematical Definitions, Similarities, and Differences  
*Jack Snoeyink, Sean Ahearn, Samuel A. Micka, Harvey J. Miller, David Millman, and Frank Staals . . . . .* 68

Going Beyond the Level of the Individuals  
*Zena Wood, Maike Buchin, Brittany Terese Fasy, Jed Long, Jennifer Miller, and Nicholas Ouellette . . . . .* 69

**Fishbowl discussion . . . . .** 69

**Participants . . . . .** 71

### 3 Overview of Talks

#### 3.1 Calibrating Agent Based Models as Multiple Scales

*Sean Ahearn (City University of New York, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Sean Ahearn

The availability of GPS enabled devices has enhanced our ability to quantitatively analyze the movement and interaction of animals and people. In this analysis, we show how these data can be used to uncover behaviors at multiple spatial and temporal scales through segmentation and through the analysis of spatial-temporal usage of a tiger's home range. We use a time-geography approach to quantifying interaction between female tigers and analyze their boundary as a function of terrain characteristics (i.e. slope). It is suggested how these data are input into an agent-based model for calibration.

#### 3.2 Analysing and Predicting Movement – A Computational Geometry Perspective

*Maike Buchin (Ruhr-Universität Bochum, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Maike Buchin

Analysing movement data leads to geometric problems and hence is interesting from a computational geometry perspective. In the past 10 years much research has been done in this direction. I discuss results on two topics related to prediction of movement: analysing delays and segmenting and classifying trajectories. Segmentation and classification ask to split respectively group trajectories by their movement behaviour. Two different approaches have been followed for characterizing similar movement: by spatio-temporal criteria and by the parameter of a random movement model. I give an overview of algorithms and their application for these two settings.

#### 3.3 Using Time Geography to Model Movement in Three Physical Dimensions

*Urška Demšar (University of St Andrews, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Urška Demšar

**Joint work of** Demšar, Urška; Long, Jed

**Main reference** Demšar U, Long JA, "Time-Geography in Four Dimensions: Potential Path Volumes around 3D Trajectories", Short Paper Proceedings of GIScience 2016, Montreal, Canada, 27-30 Sept 2016.

**URL** <https://doi.org/10.21433/B3117gc866qs>

An increase in availability and accuracy of 3D positioning requires development of new analytical approaches that will incorporate the third positional dimension, the elevation and model space and time as a 4D concept. To address this we propose the extension of time geography into four dimensions. We generalise the time geography concept of a Potential Path Area into a Potential Path Volume around a 3D trajectory, present its mathematical definition and an algorithm for calculating these volumes around a set of given 3D trajectories. The algorithm was tested on simulated data and real 3D data from movement ecology.

### 3.4 Using Prediction to Explore the Mechanisms and Consequences of Social Life

*Damien Farine (MPI für Ornithologie – Radolfzell, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Damien Farine

Maintaining cohesion during movement is fundamental for animals to gain the benefits of living in groups. Yet studying the mechanisms underlying collective movement is challenging using traditional measurement frameworks. I demonstrate how movement prediction can inform the mechanisms that underpin movement, using case studies from baboons and predator-prey interactions.

### 3.5 From Fish to Worms: Spatial Cognition, Movement and Postures in Three-Dimensions

*Robert Holbrook (University of Leeds, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Robert Holbrook  
**Joint work of** Victoria Davis, Theresa Burt de Perera, Richard Mann, Mate Nagy, Dora Biro, Sarah Schumacher, Thomas Ranner, Felix Schafer, David Pertab, Ian Hope, Netta Cohen

The real world is spatially three-dimensional. Animals, including humans, live in and move through three-dimensions every day. However, the majority of the research on animal movement has focused on only two-dimensions. For some animals with only two translational degrees of freedom of movement, the complexity of the navigation task may not change between two- and three-dimensions significantly, but for those animals with an extra degree of freedom of movement such as those that fly or swim, accurate navigation becomes a much more difficult task. Here I show examples of how a fish, *Astyanax fasciatus*, navigate through three-dimensional mazes by separating out the vertical and horizontal components of space and then integrating these when they need to navigate. The separation of the vertical component is likely aided by the use of an additional cue – hydrostatic pressure. This cue can be used alone for successful navigation in the vertical dimension, with the rate in change of swim-bladder volume a likely candidate sensory system. Despite the importance of the vertical component during navigation, it appears that it is not so important when deciding whom to pay attention to while shoaling in three-dimensions, with at least one two-dimensional implementation of a model accurately predicting the behaviour of 3D fish shoals.

The nematode worm, *Caenorhabditis elegans*, also moves through three-dimensional environments, yet all the kinematic and biomechanical research to date has been done on two-dimensional flat plates. We may therefore be missing some important behavioural repertoire from the worm in its natural habitat. Here I attempt to rectify this by analysing worm trajectories and postures while it moves through a three-dimensional gelatin cube. The worm exhibits three-dimensional behaviour more often than planar behaviour, both in trajectories and postures. Importantly, there appears to be a helical gait motion that seems to be present for much of the time the worm is moving through higher viscosities of gelatin, a behaviour which has not yet been documented.

### 3.6 Scalable Methods for Modelling Movement Patterns: from Areal to Road Segment Levels

*Robin Lovelace (University of Leeds, GB)*

**License** © Creative Commons BY 3.0 Unported license

© Robin Lovelace

**Joint work of** Alan Wilson

**Main reference** Robin Lovelace, Anna Goodman, Rachel Aldred, Nikolai Berkoff, Ali Abbas, James Woodcock, “The Propensity to Cycle Tool: An open source online system for sustainable transport planning”, *Journal of Transport and Land Use*, 10(1), 2017.

**URL** <http://dx.doi.org/10.5198/jtlu.2016.862>

From: <http://rpubs.com/RobinLovelace/290584>

It is important to know where people travel for a number of reasons. Most important among these is the urgent need to transition away from fossil fuels: models of travel patterns can help identify the most effective interventions to make this happen.

This paper explores globally scalable methods for generating estimates of travel patterns that build on areal and point-based data to estimate movements down to the road network level currently, and under scenarios of the change. The presentation is based on my experience developing the Propensity to Cycle Tool (PCT) and scaling it across all areas and major cyclable roads in England (see [pct.bike](http://pct.bike)) and recent experiments extending it internationally with a case study in Seville, Spain.

Methodologically I will explore the possibility of extending the methods to be dynamic and multi-modal, themes that will be prominent during the summer school.

### 3.7 The Moving Across Places Study (MAPS): Public Transit, Physical Activity and Walking Route Choice

*Harvey J. Miller (Ohio State University, US)*


**License** © Creative Commons BY 3.0 Unported license

© Harvey J. Miller

The Moving Across Places Study (MAPS) is a quasi-experimental study of the impacts of light rail transit and street rehabilitation on physical activity and walking route choice. Participants (n=536) wore GPS recorders and accelerometers for one week before and after the construction of a light rail transit (LRT) line and walkability enhancements in a neighborhood of Salt Lake City, Utah, USA. We are able to demonstrate that these design interventions resulted in more physical activity and new physical activity time that did not draw from recreational physical activity time or cannibalize existing us ridership. We compare theory-driven and data-driven approaches to understanding walking route choice through this built environment. The theory-driven approach is easier to explain, but the data-driven approach fits the data better and also points more directly to actionable knowledge.

### 3.8 Random Trajectories in Movement Ecology: A Path to Crossing Scales in Movement Ecology?

*Kamran Safi (MPI für Ornithologie – Radolfzell, DE)*

License  Creative Commons BY 3.0 Unported license  
© Kamran Safi

With the advances of technological developments the granularity and volume of movement data is ever increasing raising the need for new analytic tools. Partially the problem lies buried in the way movement data is collected by discretising a continuous process, where with increasing resolution in time and space the discretisation suffers from increasing autocorrelation. The discretisation affects, in interaction with the underlying continuous movement process, almost all quantities usually derived from trajectories, such as speed or turning angle and the amount of autocorrelation detectable. The definition of Null models in movement ecology being inherently difficult task has become more challenging mainly as the assumption of independence in the data becomes more evidently violated. Different methods have been suggested to incorporate some formal continuous time movement model to integrate the autocorrelation structure. With the increasing volume and accessibility of movement data, however, another alternate path might open: the use of empirical distributions to create conditional random trajectories. In this talk I present the eRTG, the empirical random trajectory generator, which is based on using empirical joint distributions of speed and turning angle derived from discretised movement data to create random trajectories connecting a given start and end point maintaining the original geometry of the template. Finally, I use two case examples to highlight the potential of the eRTG to explore hypothesis and formulate hypothesis. First, the eRTG is used to show the difference in orientation task in the white stork when migrating along the Western or Eastern migratory flyways based on the fusion of movement data with banding recoveries. Based on the eRTG, the Eastern migratory flyway should pose higher orientation demands on the white storks than the western route population. I conclude with a study using eRTG to investigate the potential of wild waterfowl transporting avian influenza.

### 3.9 On Language for Observation & Prediction

*Jack Snoeyink (University of North Carolina at Chapel Hill, US)*

License  Creative Commons BY 3.0 Unported license  
© Jack Snoeyink

“Language shapes the way we think, and determines what we can think about.” I present four vignettes on how to think about the languages used in collaboration between movement science(s) and computational geometry:

1. Computer Scientists create languages – e.g., object programming creates nouns and attaches their verbs.
2. Languages that can aid ones thinking can still hinder communication if not shared.
3. Boundary objects, which can be described by each collaborator in their own way, facilitate collaboration.
4. When creating computational models, create scientific “unit tests” that use a language of features and their distributions to circumscribe desired behavior.

### 3.10 Collective Motion: More than the Sum of its Parts

Zena Wood (*University of Greenwich, GB*)

**License** © Creative Commons BY 3.0 Unported license  
© Zena Wood

To sufficiently analyse collective motion, and predict or simulate future motion, we need to look at more than just the level of the individual members. The level of the collective, and the environment where the motion takes place, must also be considered. This talk will illustrate how concepts from formal ontology and other disciplines, such as group organisation theory, can influence the analytical methods that we develop to analyse collective motion. The challenges and opportunities relating to the analysis of collective motion will also be discussed.

#### References

- 1 Wood, Zena, and Antony Galton. (2009) “A taxonomy of collective phenomena.” *Applied Ontology* 4.3-4: 267–292.
- 2 Wood, Zena. (2013). Profiling Spatial Collectives. In Max Bramer and Miltos Petridis, editors, *Incorporating Applications and Innovations in Intelligent Systems XXI Proceedings of AI-2013, The Thirty-third SGA International Conference on Innovative Techniques and Applications of Artificial Intelligence*: 95–108. Springer.
- 3 Wood, Zena., (2014). What can Spatial Collectives tell use about their environment? *IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2014)*, 329–336.
- 4 Galton, Antony, and Zena Wood. (2016) “Extensional and intensional collectives and the de re/de dicto distinction.” *Applied Ontology* 11.3: 205–226.

## 4 Working groups

### 4.1 Computational Topology and Movement Data

Kevin Buchin (*TU Eindhoven, NL*), Maike Buchin (*Ruhr-Universität Bochum, DE*), Brittany Terese Fasy (*Montana State University – Bozeman, US*), Kristine Pelatt (*St. Catherine University – St. Paul, US*), and Carola Wenk (*Tulane University, US*)

**License** © Creative Commons BY 3.0 Unported license

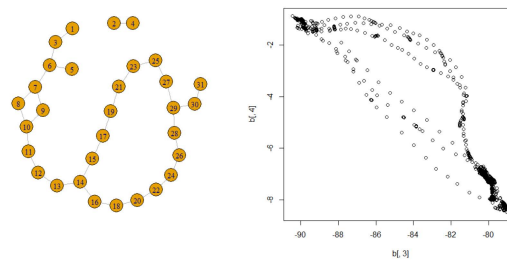
© Kevin Buchin, Maike Buchin, Brittany Terese Fasy, Kristine Pelatt, and Carola Wenk

**Main reference** Gurjeet Singh, Facundo Mémoli, Gunnar E. Carlsson: “Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition”, in *Proc. of the Symposium on Point Based Graphics*, Prague, Czech Republic, 2007. Proceedings, pp. 91–100, Eurographics Association, 2007.

**URL** <http://dx.doi.org/10.2312/SPBG/SPBG07/091-100>

Computational Topology has proved effective in a wide range of applications, but has so far only found few applications in movement analysis. The aim of this working group was to explore how existing software for topological data analysis can be used to analyze movement.

The working group focused on TDMapper, an R package for using discrete Morse theory to analyze a data set using the Mapper algorithm (Singh et al., 2007), and demonstrated it on Galapagos Albatross tracks; see Figure 2.



■ **Figure 2** Output of the mapper algorithm (left) for Albatross trajectory data (right).

## 4.2 Defining Axes and Metrics to Characterise Collective Motion

*Somayeh Dodge (University of Minnesota – Minneapolis, US), Urska Demšar (University of St Andrews, GB), Jed Long (University of St Andrews, GB), Andrea Perna (University of Roehampton – London, GB), Alexander Szorkovszky (Uppsala University, SE), Johan van de Koppel (Royal Netherlands Inst. for Sea Research – Yerseke, NL), and Zena Wood (University of Greenwich, GB)*

**License** © Creative Commons BY 3.0 Unported license

© Somayeh Dodge, Urska Demšar, Jed Long, Andrea Perna, Alexander Szorkovszky, Johan van de Koppel, and Zena Wood

In this working group we worked towards the identification of relevant axes along which to characterise, and ideally to quantify, collective motion phenomena.

We started from considering properties that define a group, such as the spatial proximity (or proximity in non-spatial dimensions), the differentiation of roles across group members, the coherence of mutual positions and of collective motion. From this, we moved to analysing the drives that determine group formation in terms of costs and benefits for the individuals that compose the group, and benefits for the entire group. At one extreme, animals can exhibit collective motion, in the form of an aggregation in a single place, without any form of interaction: this is the case of animals that aggregate for instance around a resource such as a source of food or water. In many examples of naturally occurring animal groups, the individuals form a group because they experience a direct benefit from being with other members of the group in terms of avoiding predation or gaining protection from natural phenomena (waves, low temperature etc.) We proposed that the costs and benefits of group formation could be characterised by using a classification similar to the one traditionally used to characterise ecological interactions such as predation and parasitism (whereby some members of the group benefit from the association, to the detriment of other members of the group), mutualism and symbiosis (in which both units participating in the association gain a benefit) and commensalism (whereby some individuals benefit from the association, with no detriment or benefit for the other individuals).

The natural next step would be defining axes that are more specific to the collective motion of animal groups, and not simply to the aggregation behaviour or to the characterisation of static groups. Both computer scientists working on ontologies and ecologists have been independently working on definition of the properties of (animal) groups and the possibility to exchange ideas between these two disciplines in this working group was particularly insightful.

### 4.3 Multi-scale Movement Modeling

*Somayeh Dodge (University of Minnesota – Minneapolis, US), Kevin Buchin (TU Eindhoven, NL), Urska Demšar (University of St Andrews, GB), Harvey J. Miller (Ohio State University, US), Kristine Pelatt (St. Catherine University – St. Paul, US), Alexander Szorkovszky (Uppsala University, SE), and Johan van de Koppel (Royal Netherlands Inst. for Sea Research – Yerseke, NL)*

**License** © Creative Commons BY 3.0 Unported license  
 © Somayeh Dodge, Kevin Buchin, Urska Demšar, Harvey J. Miller, Kristine Pelatt, Alexander Szorkovszky, and Johan van de Koppel

Movement occurs at multiple spatial and temporal scales, which can result in a range of embedded local to global movement patterns. The way in which context influences patterns of movement differs across scales. Movement ecology models have mainly involved random searches and step selection strategies at local scales. These models incorporate the environmental factors and social interactions that influence local movement choices of individuals. On the other hand, human mobility studies have traditionally focused on modeling macro-level patterns such as origin-destination flows. These models mainly consider global behavioral patterns and goal-oriented movement of humans. While both approaches are essential in the study of moving phenomena, there is a gap in methodology for tackling multiple scales of movement and their associations to the individual's behavior and environment. In a multi-scale movement modeling approach, global models should describe the process and local models should describe the local variabilities of movement. Our group discussed the differences between data-driven and theory-driven modeling approaches and ways in which they could potentially be used to integrate multiple scales of movement. As a data-driven approach, the group discussed the potential usage of topological modeling to filter data to the big trend and then extracting the details of local movement patterns. The theory-driven approach could potentially benefit from applying both local rules and global rules to model movement across multiple scales. The group also came up with the following relevant research questions/challenges:

- how can we generate algorithms that work properly across scales?
- how can we connect different scales of movement, in terms of both geographic scale and time scale?
- to what extends can we infer a goal-oriented movement from local movement patterns?
- do global objectives of movement emerge from local rules or does the global objective of movement influence local movement rules?

The group concluded that 'multi-scale modeling of movement' is a challenging research gap in the field and deserves more attention from the scientific community.



#### 4.4 Using and Explaining Non-Traditional Metrics in Biology Publications

*Damien Farine (MPI für Ornithologie – Radolfzell, DE), Robert Holbrook (University of Leeds, GB), Richard Philip Mann (University of Leeds, GB), Andrea Perna (University of Roehampton – London, GB), and Kamran Safi (MPI für Ornithologie – Radolfzell, DE)*

**License** © Creative Commons BY 3.0 Unported license

© Damien Farine, Robert Holbrook, Richard Philip Mann, Andrea Perna, and Kamran Safi

We were concerned with the issue of how to present metrics from movement studies in biology journals. These metrics do not typically come in the form of significance tests and p-values, but might for instance be the likelihood ratio between two models, comparisons of analyses between real and permuted data sets, or effect-sizes (with uncertainties)

Overall we agreed that a crucial aspect was the ability to communicate clearly the value and rigour of the alternative metric. This might be aided by having a shared resource that explains these metrics clearly, perhaps by analogy with more well-known measures. This could be in the form of a paper or web page. However, to create this would require referencing established literature where these metrics are used, and ideally also justified. To this end we should also seek to explain clearly why we use the metrics that we do, and why we do not follow established significance test methods, in our own papers. By doing this we can create cultural change that will make the use of new methods easier in future.

#### 4.5 Learning Connections Between Landscape and Trajectories from Recorded Data

*Richard Philip Mann (University of Leeds, GB), Maike Buchin (Ruhr-Universität Bochum, DE), Robert Holbrook (University of Leeds, GB), and Nicholas Ouellette (Stanford University, US)*


**License** © Creative Commons BY 3.0 Unported license

© Richard Philip Mann, Maike Buchin, Robert Holbrook, and Nicholas Ouellette

We discussed how data from navigating pigeons, combined with landscape images, could be used to understand how the landscape drives movement. This followed on from related discussions at the previous Dagstuhl seminar nr. 16022. We assumed that pigeons' trajectories should result from some sort of optimisation process, involving landscape characteristics under the path. We have data from early, learning flights to later consistent flights, which offers the possibility of understanding the dynamics of this optimisation. We initially considered constructing a 'energy'-potential that would define the 'energy' of any route, and combining this with localised improvements to the route to lower potentials, in a framework similar to step-selection. However, it became clear based on earlier trials of this idea and further discussion that this would not work – changes in routes do not appear to be local, but global. We therefore wondered how we could understand the process of exploring different trajectories and settled on ideas similar to markov-chain monte carlo or simulated annealing. However, we concluded that the data we had would probably be insufficient to infer the potential landscape within this regime. As such we determined that a fruitful next step would be to create simulated data from a known ground truth and assess how much/what type of trajectory data we would need to accurately infer the learning process and the potential landscape used.

## 4.6 Potential Applications of Ecology on Transport and the Implications on Policy


*Samuel A. Micka (Montana State University – Bozeman, US), Mark Birkin (University of Leeds, GB), Maarten Löffler (Utrecht University, NL), Robin Lovelace (University of Leeds, GB), Richard Philip Mann (University of Leeds, GB), Kathleen Stewart (University of Maryland – College Park, US), and Carola Wenk (Tulane University, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Samuel A. Micka, Mark Birkin, Maarten Löffler, Robin Lovelace, Richard Philip Mann, Kathleen Stewart, and Carola Wenk

The activity spaces of people provide important information about where they travel on a day-to-day basis. The aim of this working group was to identify similarities between the ecological term “home range”, which defines a similar concept for animals, and activity spaces. Drawing similarities between these ideas could help identify more information about the motivations behind human actions within their activity spaces. Ultimately, this information could determine where people are throughout the day, and why they are there. This sort of predictive model could provide input for route planning algorithms and city planners. However, the definitions of activity spaces and home ranges vary drastically in different contexts, making a relationship difficult to define. Despite these difficulties, we explored different types of data sets and how they may fit into a predictive model. These data sets included origin-destination pairings, trajectory data, survey data (where do you work, where do you live, etc.), and census data. We considered different models that would accept the different data types as input, such as dynamic graphs that could store the intent of trips.

## 4.7 Formalizing the Notions of “Activity Spaces” and “Homeranges”: Mathematical Definitions, Similarities, and Differences

*Jack Snoeyink (University of North Carolina at Chapel Hill, US), Sean Ahearn (City University of New York, US), Samuel A. Micka (Montana State University – Bozeman, US), Harvey J. Miller (Ohio State University, US), David Millman (Montana State University – Bozeman, US), and Frank Staals (Aarhus University, DK)*

**License**  Creative Commons BY 3.0 Unported license  
© Jack Snoeyink, Sean Ahearn, Samuel A. Micka, Harvey J. Miller, David Millman, and Frank Staals

Activity spaces and home ranges both generalize the trajectories representative of where humans and animals travel for their daily tasks and activities. These spaces, on the surface, appear as geometric summaries of these trajectories. However, cross-disciplinary understandings of home ranges and activity spaces differ, creating ambiguity in the definitions leading to inconsistent mathematical representations. To create cohesion between these fields we propose a space partitioning data structure which provides tunable rules to create home ranges/activity spaces from trajectory data. By offering a general data structure to geometrically represent these spaces, the definitions and respective representations will become more consistent and easily communicated in cross-disciplinary research.

## 4.8 Going Beyond the Level of the Individuals

*Zena Wood (University of Greenwich, GB), Maike Buchin (Ruhr-Universität Bochum, DE), Brittany Terese Fasy (Montana State University – Bozeman, US), Jed Long (University of St Andrews, GB), Jennifer Miller (University of Texas – Austin, US), and Nicholas Ouellette (Stanford University, US)*

**License** © Creative Commons BY 3.0 Unported license

© Zena Wood, Maike Buchin, Brittany Terese Fasy, Jed Long, Jennifer Miller, and Nicholas Ouellette

The group discussed the relationship between the individual members and the collective itself. Key questions/topics included whether a hierarchical structure exists; the relationship between roles and members; how roles are defined and identified within a spatiotemporal dataset; how a true collective can be identified; and, the metrics that can be applied to a collective but not the individual members.

To address how a true collective could be identified, and whether it be done independent to an application, the properties of collectives were discussed. Individual members know that they are part of a collective with each collective having an identity. Identifying changes in behaviour of an individual could be used to identify membership. Both collective and individual goals can be considered. Distinct roles can lead to individuals participating in a collective goal in different ways. Properties can be ascribed to the collective that cannot be ascribed to the individuals. It is not clear which properties would be considered meaningful given a dataset. We discussed methods that might prove useful in identifying collectives (e.g., connected components looking for persistent features). Instead of identifying true collectives, you could try to determine if something is not a collective using adversary detection.

Going forward there are lots of questions with no apparent answers. It is clear that we need to develop metrics and identify some examples of collective motion (e.g., examples where the individual goal is fundamentally different to the collective goal).

## 5 Fishbowl discussion

Over the course of an afternoon, a fishbowl conversation was used to encourage discussion. In this session, three attendees discussed a topic in front of the rest of the seminar. The positions in the center were vacated and refilled by others as people wanted to make contributions to the conversation. One moderator was responsible for asking questions and providing talking points. Here, we outline some of the major contributions and conclusions drawn from this session.

The first part of the session was centered around defining the characteristics of prediction of movement. Differences were highlighted between local and global behaviors, human and animal trajectories, and the purposes of predictions. Many speakers had different interpretations of prediction and what it could be used for. The conversation moved on the role of geometry in prediction of trajectories. Specifically, can deterministic methods be used to help predict real world movement? Despite predictive models being available for cell life and animal populations, the speakers decided that it would not be realistic to predict animal behavior deterministically. This topic led to the discussion of fundamental laws, such as the ones found in physics. The general consensus was that animals have goals, and use movement to achieve these goals. Some goals are predictable, but a universal predictive model is not feasible. Later, the idea of naïve movement was introduced. Naïve movement encapsulates

the idea of natural, predictable movement, like water flow in physics. Since each animal has a different interpretation of its surroundings, defining a universal a set of senses is difficult.

For many animals, home ranges encompass a large number of simple, and predictable, behaviors. This led to the discussion of how animals interpret their own homeranges, which again, was decided to be subjective. However, some animals possess a cognitive map of their surroundings, which could help in predictive models for certain behaviors and species. In particular, pigeons have an uncanny ability to navigate.

The role of mathematicians and computer scientists in this field of research also emerged as a topic. The desire for a common language to communicate animal behavior rose. The idea being that, if we can communicate these movements across disciplines in a way that everyone understands, we will be able to more easily develop models. One of the major problems with communicating these biological results with mathematicians and computer scientists is that, without a set of fundamental rules, how are methods verified? The speakers agreed that verification should be achieved through observation and professional opinions. However, with a lack of a deterministic model for animal movement, prediction is still a very animal-specific and difficult problem to approach.

## Participants

- Sean Ahearn  
City University of New York, US
- Mark Birkin  
University of Leeds, GB
- Kevin Buchin  
TU Eindhoven, NL
- Maike Buchin  
Ruhr-Universität Bochum, DE
- Urska Demšar  
University of St Andrews, GB
- Somayeh Dodge  
University of Minnesota –  
Minneapolis, US
- Damien Farine  
MPI für Ornithologie –  
Radolfzell, DE
- Brittany Terese Fasy  
Montana State University –  
Bozeman, US
- Robert Holbrook  
University of Leeds, GB
- Maarten Löffler  
Utrecht University, NL
- Jed Long  
University of St Andrews, GB
- Robin Lovelace  
University of Leeds, GB
- Richard Philip Mann  
University of Leeds, GB
- Samuel A. Micka  
Montana State University –  
Bozeman, US
- Harvey J. Miller  
Ohio State University, US
- Jennifer Miller  
University of Texas – Austin, US
- David Millman  
Montana State University –  
Bozeman, US
- Nicholas Ouellette  
Stanford University, US
- Kristine Pelatt  
St. Catherine University –  
St. Paul, US
- Andrea Perna  
University of Roehampton –  
London, GB
- Kamran Safi  
MPI für Ornithologie –  
Radolfzell, DE
- Jack Snoeyink  
University of North Carolina at  
Chapel Hill, US
- Frank Staals  
Aarhus University, DK
- Kathleen Stewart  
University of Maryland –  
College Park, US
- Daniel Strömbom  
Swansea University & Uppsala  
University
- Alexander Szorkovszky  
Uppsala University, SE
- Johan van de Koppel  
Royal Netherlands Inst. for Sea  
Research – Yerseke, NL
- Carola Wenk  
Tulane University, US
- Zena Wood  
University of Greenwich, GB



# Resource Bound Analysis

Edited by

Marco Gaboardi<sup>1</sup>, Jan Hoffmann<sup>2</sup>, Reinhard Wilhelm<sup>3</sup>, and  
Florian Zuleger<sup>4</sup>

- 1 University at Buffalo, US, [gaboardi@buffalo.edu](mailto:gaboardi@buffalo.edu)
- 2 Carnegie Mellon University – Pittsburgh, US, [jhoffmann@cmu.edu](mailto:jhoffmann@cmu.edu)
- 3 Universität des Saarlandes, Saarland Informatics Campus, DE,  
[wilhelm@cs.uni-saarland.de](mailto:wilhelm@cs.uni-saarland.de)
- 4 TU Wien, AT, [zuleger@forsyte.at](mailto:zuleger@forsyte.at)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17291 “Resource Bound Analysis”. Resource-bound analysis is studied in formal methods and programming languages at different levels of abstraction. The goal of the Dagstuhl seminar was to bring together leading researchers with different backgrounds in resource-bound analysis to address challenging open problems and to facilitate communication across research areas.

**Seminar** July 16–21, 2017 – <http://www.dagstuhl.de/17291>


**1998 ACM Subject Classification** C.3 [Special-Purpose and Application-Based Systems]: Real-Time and Embedded Systems, F.3.2 [Semantics of programming languages]: Program Analysis

**Keywords and phrases** quantitative analysis, resource-bound analysis, WCET

**Digital Object Identifier** 10.4230/DagRep.7.7.72

## 1 Executive Summary

Marco Gaboardi  
Jan Hoffmann  
Reinhard Wilhelm  
Florian Zuleger

**License**  Creative Commons BY 3.0 Unported license  
© Marco Gaboardi, Jan Hoffmann, Reinhard Wilhelm, and Florian Zuleger

*This seminar is dedicated to our friend and colleague Martin Hofmann (1965-2018). Martin’s vision and ideas have shaped our community and the way resource analysis is performed and thought about. We are grateful for the time we spent with him and we will sorely miss his ingenuity, kindness, and enthusiasm.*

There are great research opportunities in combining the three aforementioned approaches to resource bound analysis. The goal of the Dagstuhl seminar was to bring together leading researchers with different backgrounds in these three areas to address challenging open problems and to facilitate communication across research areas.

To this end, the program included seven tutorials on state-of-the-art techniques in the different communities, and short talks on concrete topics with potential for cross-fertilization. This included combining WCET analysis with higher-level bound analysis techniques, hardware-specific refinement of high-level cost models, and interaction of resource analysis with compilation. Additionally, the seminar included two tools sessions: the first



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Resource Bound Analysis, *Dagstuhl Reports*, Vol. 7, Issue 7, pp. 72–87

Editors: Marco Gaboardi, Jan Hoffmann, Reinhard Wilhelm, and Florian Zuleger



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

was a presentation of the aiT tool of AbsInt by Simon Wegener; the second was a session with presentations of different tools from different participants. Finally, the seminar included a discussion on open problems in the different areas as well as open problems for cross-fertilization.

The tutorials, the talks solicited from the participants, and the tool and discussion sessions allowed us to identify topics which are of common interest to the three different communities. Some of these topics are

- invariant and flow analysis,
- constraint solving and
- formalisms and logics for resource bounds.

Supporting information about program invariants and the possible control flow are often required by a resource analysis, e.g., the maximal value of a loop counter, or the infeasibility of a program path. The actual resource analysis is often reduced to solving a constraint system, e.g., using techniques from linear programming or recurrence equations. Verification logics for resource bounds as well as programming language formalisms are of common interest as they allow to specify or to guarantee that a program satisfies a required worst case resource bound.

We believe that the further study of these topics promises to increase the connections and to leverage the synergies between the different communities.

## 2 Table of Contents

### Executive Summary

*Marco Gaboardi, Jan Hoffmann, Reinhard Wilhelm, and Florian Zuleger . . . . .* 72

### Background

*Marco Gaboardi, Jan Hoffmann, Reinhard Wilhelm, and Florian Zuleger . . . . .* 76

### Overview of Tutorials . . . . . 76

Tutorial: WCET Analysis: A Primer

*Jan Reineke . . . . .* 76

Tutorial: Control Flow Analysis for WCET Analysis

*Björn Lisper . . . . .* 77

Tutorial: Automatic Amortized Analysis

*Martin Hofmann . . . . .* 77

Tutorial: Cost Analysis with Recurrence Relations (using CiaoPP) and its Applications

*Manuel Hermenegildo . . . . .* 78

Tutorial: From Implicit Complexity to Resource Bound Analysis

*Ugo Dal Lago . . . . .* 78

Tutorial: Complexity Analysis of Term Rewrite System

*Martin Avanzini . . . . .* 79

Tutorial: RTDroid: Toward dynamic real-time systems

*Lukasz Ziarek . . . . .* 79

### Tool demo session report . . . . . 79

### Overview of other Talks . . . . . 80

Parametric Timing Analysis

*Sebastian Altmeyer . . . . .* 80

Amortised Resource Analysis with Separation Logic

*Robert Atkey . . . . .* 80

Resource Analysis of Session Typed Programs

*Ankush Das . . . . .* 80

A proof system for relational cost analysis

*Deepak Garg . . . . .* 81

Compositional timing control in HLS

*Dan R. Ghica . . . . .* 81

Complexity for Java with AProVE

*Jürgen Giesl . . . . .* 81

Enabling Compositionality for (Multi-Core) Execution Time Analysis

*Sebastian Hahn . . . . .* 82

Energy Consumption Analysis and Verification

*Manuel Hermenegildo . . . . .* 82



Generating Invariants and Resource Bounds with Recurrence Analysis  
*Zachary Kincaid* . . . . . 82

Type systems for Energy Management  
*Yu David Liu* . . . . . 83

Weighted Automata Theory for Resource Analysis of Rewrite Systems  
*Georg Moser* . . . . . 83

Solving Recurrences Using Operational Calculus  
*Thomas W. Reps* . . . . . 83


ECA: Energy Consumption Analysis of software controlled systems  
*Marko van Eekelen* . . . . . 84

**Panel discussion and Open problems** . . . . . 84

**Participants** . . . . . 87

### 3 Background

Marco Gaboardi (University at Buffalo, US), Jan Hoffmann (Carnegie Mellon University – Pittsburgh, US), Reinhard Wilhelm (Universität des Saarlandes, DE), and Florian Zuleger (TU Wien, AT)

License  Creative Commons BY 3.0 Unported license  
© Marco Gaboardi, Jan Hoffmann, Reinhard Wilhelm, and Florian Zuleger

Resource-bound analysis is studied in formal methods and programming languages at different levels of abstraction. This ranges from concrete clock-cycle bounds on specific hardware (WCET analysis), to high-level symbolic bound analysis (recurrence relations, type systems, abstract interpretation, term rewriting), to logical characterizations of asymptotic complexity (linear logic, type systems, semantics). These are active areas of research and there has been significant progress in all of them over the past decade. However, these problems are often studied by different communities with little overlap. Methods close to the implementation level base their approaches on concrete hardware models and derive precise execution-time bounds, vulgo WCETs, in number of clock-cycles, and bounds on the extension of stack and heap-space. WCET analysis is a success story of formal methods and is meanwhile applied in industry. WCET methods usually work on the machine code level, often lack compositionality and are restricted to programs without complex control flow. Techniques used in WCET analysis include Abstract Interpretation, ILP solving, and Model Checking.


The field of *automatic symbolic resource-bound analysis* studies methods for automatically deriving bounds that are functions of the inputs of a program. The bounds are usually non-asymptotic and many techniques can automatically handle complex data structures and control flow. However, the cost models that are used are often simplistic and ignore the characteristics of modern high-performance architectures. The used techniques include abstract interpretation, recurrence relations, type systems, LP solving, and SMT solving.

Finally, there is an active community that studies *theoretical foundations of resource analysis*. This includes derivation of asymptotic bounds, mechanization and verification of bounds, completeness results, relational resource analysis, formal cost semantics, and soundness proofs. The studied techniques are often very powerful and general. However, the used cost models are fairly abstract and the used bound-analysis techniques are hard to automate. The used techniques include linear logic, affine and dependent type systems, proof assistants, and operational semantics.

### 4 Overview of Tutorials

#### 4.1 Tutorial: WCET Analysis: A Primer

Jan Reineke (Universität des Saarlandes, DE)

License  Creative Commons BY 3.0 Unported license  
© Jan Reineke

Worst-case execution time (WCET) analysis is concerned with computing an upper bound on the execution time of a program on a particular hardware platform.

After discussing the three main factors that influence the execution time of a program, 1. the program's inputs, 2. the state of the HW platform, and 3. interference from the environment, I discuss the structure of modern WCET analysis tools for single-core

architectures. Such tools essentially separately solve subproblems concerning (a) the possible paths through the program, and (b) the possible “microarchitectural” paths. The solutions to these subproblems can then be combined in an integer linear programming formulation to obtain an upper bound on the WCET.

## 4.2 Tutorial: Control Flow Analysis for WCET Analysis

*Björn Lisper (Mälardalen University – Västerås, SE)*

License © Creative Commons BY 3.0 Unported license  
© Björn Lisper

An important part of WCET analysis is the Control Flow Analysis (CFA), which in this context amounts to finding constraints on the program flow such as loop iteration bounds, and infeasible path constraints. We review the problem in the IPET setting, where the program flow constraints are expressed as linear inequalities on execution counters for basic blocks. We discuss the relation to CFA for functional programs. We then show some existing methods for CFA for WCET, and we discuss their respective pros and cons.

## 4.3 Tutorial: Automatic Amortized Analysis

*Martin Hofmann (LMU München, DE)*

License © Creative Commons BY 3.0 Unported license  
© Martin Hofmann

Amortized complexity was introduced by Tarjan to facilitate the runtime analysis of data structures that sometimes perform costly operations which pay off later. Typical examples include self-organising binary search trees and also many functional data structures. Initially motivated by ideas from implicit computational complexity (ICC), amortized complexity has been harnessed for the automatic analysis of runtime and other resources. The crucial idea is that with an appropriate choice of a potential function the amortized complexity of basic operations becomes constant in many cases which avoids the need for tracking sizes and shapes of data structures during the analysis. Furthermore, the compositionality of amortized complexity makes it suitable for integration with type systems. This research has led to several systems for the automatic analysis of functional and object-oriented programs. There have also been applications to resource usage certification, to the analysis of low-level systems code and to lazy functional programming and to term rewriting. The talk gives a gentle introduction to this topic beginning from its sources in implicit computational complexity and closes with recent developments and some open questions.

#### 4.4 Tutorial: Cost Analysis with Recurrence Relations (using CiaoPP) and its Applications

*Manuel Hermenegildo (IMDEA Software – Madrid, ES)*

License  Creative Commons BY 3.0 Unported license  
© Manuel Hermenegildo

We present in a tutorial fashion, how to perform cost analyses for a wide class of resources by setting up and solving recurrence relations on program data sizes and procedure costs. We follow the method we have proposed and developed for analysis of programs in Horn clause form and its implementation in the CiaoPP system, but the discussion is applicable in general to recurrence relation-based models.

We start by discussing a number of interesting applications, from our original motivation of task granularity control in automatic program parallelization to others such as resource verification, security, or static profiling, demonstrating them through examples run on the CiaoPP system. Based on the characteristics of the problem and the demands of these applications, we motivate the objective of obtaining upper and lower bounds, the requirement that these bounds be functions of program input sizes, and the need to infer both data size and procedure cost functions.

We continue by discussing how a good intermediate representation (in our case, Horn clauses) allows supporting multiple languages in a uniform way, showing several transformations from different languages into this intermediate form. We also present the concept of user-definable resources, via assertions, and demonstrate it through a number of simple examples run on the CiaoPP system. We also show how this concept is instrumental in supporting multiple languages.

We then present the method for performing cost analyses by setting up and solving recurrence relations on the program data sizes and procedures. We illustrate the method through a detailed worked example, deriving step by step the cost relations and the closed forms for both the sizes and the procedure cost. We also discuss a number of issues related to solving recurrence equations. We show through an example how to set up non-deterministic recurrence equations for inferring balanced costs (e.g., for obtaining quadratic bounds for quick-sort). To conclude we present the notions of accumulated cost and static profiling and also illustrate them with an example.

#### 4.5 Tutorial: From Implicit Complexity to Resource Bound Analysis

*Ugo Dal Lago (University of Bologna, IT)*

License  Creative Commons BY 3.0 Unported license  
© Ugo Dal Lago

Implicit computational complexity aims at giving precise, but machine-free, characterisations of complexity classes, like polynomial time or polynomial space computable functions. Many of the proposed systems can be turned into verification methodologies providing resource bounds on the input program. The intrinsically poor expressive power of these methodologies has been sometime overcome by making the underlying logic less implicit but more informative. We focus our attention on systems derived from Girard's linear logic.

## 4.6 Tutorial: Complexity Analysis of Term Rewrite System

*Martin Avanzini (Universität Innsbruck, AT)*

**License** © Creative Commons BY 3.0 Unported license  
© Martin Avanzini

Over the last decade, the rewriting community has introduced various novel techniques for the runtime complexity analysis of term rewrite systems. In particular, this research has also led to various tools in this context.

In this talk, I give a general overview on how the rewriting community tackles the problem.

## 4.7 Tutorial: RTDroid: Toward dynamic real-time systems

*Lukasz Ziarek (University at Buffalo, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Lukasz Ziarek

Time predictability is a requirement for computer systems that have deadlines to meet, but it is frustratingly difficult to achieve in the complex, layered, execution environments that are common place today. This talk will consider how to bring a degree of time predictability to Android applications.

Potential solutions include fundamental changes to the Android framework and the introduction of a new programming model, which focuses on the interplay between real-time activities and the rest of the system. This talk will detail the changes in the Android APIs which are required for developers to express the timeliness requirements of code and how well those requirements can be met on stock hardware in the presence of multiple, potentially interacting applications.

The talk will also cover some experimental data validating feasibility over several applications including UAV fight control, implantable medical devices, as well as a wind turbine monitoring device. Lastly, I will discuss future directions, including adding adaptivity to the system to achieve a dynamically defined real-time system.

## 5 Tool demo session report

During the seminar some time was dedicated to the presentation of tools for resource bound analysis. Participants of 7 tools presented during a tool demo session on Thursday afternoon. Each presenter had 15 minutes to present their tool. The list of presented tools and presenters is the following.

- GenE (Peter Wegemann)
- OTAWA (Hugues Cassé)
- SWEET (Björn Lisper)
- Absynth (Chan Ngo)
- RAML (Ankush Das)
- LazyAA (Pedro Vasconcelos)
- AProVE (Jürgen Giesl)
- CiaoPP (Manuel Hermenegildo)

## 6 Overview of other Talks

### 6.1 Parametric Timing Analysis

*Sebastian Altmeyer (University of Amsterdam, NL)*

License  Creative Commons BY 3.0 Unported license  
© Sebastian Altmeyer

Parametric timing analysis provides computes symbolic WCET estimates instead of numeric bounds. In this talk, I will present which analyses are needed to extend the standard numeric WCET analysis and how these analyses work. The symbolic path analysis is of particular interest, and I will provide two variants of it.

### 6.2 Amortised Resource Analysis with Separation Logic

*Robert Atkey (University of Strathclyde – Glasgow, GB)*

License  Creative Commons BY 3.0 Unported license  
© Robert Atkey

Type-based amortised resource analysis following Hofmann and Jost—where resources are associated with individual elements of data structures and doled out to the programmer under a linear typing discipline—have been successful in providing concrete resource bounds for functional programs, with good support for inference. In this work we translate the idea of amortised resource analysis to imperative pointer-manipulating languages by embedding a logic of resources, based on the affine intuitionistic Logic of Bunched Implications, within Separation Logic. The Separation Logic component allows us to assert the presence and shape of mutable data structures on the heap, while the resource component allows us to state the consumable resources associated with each member of the structure.

We present the logic on a small imperative language, based on Java bytecode, with procedures and mutable heap. We have formalised the logic and its soundness property within the Coq proof assistant and extracted a certified verification condition generator. We also describe an proof search procedure that allows generated verification conditions to be discharged while using linear programming to infer consumable resource annotations.

We demonstrate the logic on some examples, including proving the termination of in-place list reversal on lists with cyclic tails.

### 6.3 Resource Analysis of Session Typed Programs

*Ankush Das (Carnegie Mellon University – Pittsburgh, US)*

License  Creative Commons BY 3.0 Unported license  
© Ankush Das

In this work, we aim to measure the work done by session typed programs. For that, we define a potential based type system where the types provide an upper bound on the work performed by the processes. I will introduce and motivate session types, define a cost semantics to measure work, and design the type system. Finally, I will describe the soundness of our system, and conclude with an example and remarks on future directions.

## 6.4 A proof system for relational cost analysis

*Deepak Garg (MPI-SWS – Saarbrücken, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Deepak Garg

Formal frameworks for cost analysis of programs have been widely studied in the unary setting and, to a limited extent, in the relational setting. However, many of these frameworks focus only on the cost aspect, largely side-lining functional properties and value-sensitivity that are often required for cost analysis, thus leaving many interesting programs out of their purview. This talk shows how a simple, expressive proof system for costs (unary and relational) can be built using basic ingredients: a cost monad and higher-order refinements. The result is highly expressive. Besides several new examples, it can be used as a meta framework for embedding existing formal systems for cost analyses.

## 6.5 Compositional timing control in HLS

*Dan R. Ghica (University of Birmingham, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Dan R. Ghica

We examine how type systems are used to control space and time constraints in high-level synthesis. Syntactic Control of Interference, an affine restriction of Idealise Algol allows execution in constant space whereas a bounded-linear logic-like type system controls timings. However, this type system gives impractical timing constraints because of excessive required precision. We examine an alternative, simpler, notion of timing in type which is non-deductive, i.e. established directly from examining the semantic model, yet compositional.

## 6.6 Complexity for Java with AProVE


*Jürgen Giesl (RWTH Aachen, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Jürgen Giesl

While AProVE is one of the most powerful tools for termination analysis of Java since many years, we now extend our approach in order to analyze the complexity of Java programs as well. Based on a symbolic execution of the program, we develop a novel transformation of (possibly heap-manipulating) Java programs to integer transition systems (ITSs). This allows us to use existing complexity analyzers for ITSs to infer runtime bounds for Java programs. We demonstrate the power of our implementation on an established standard benchmark set.

## 6.7 Enabling Compositionality for (Multi-Core) Execution Time Analysis


*Sebastian Hahn (Universität des Saarlandes, DE)*

License  Creative Commons BY 3.0 Unported license  
© Sebastian Hahn

The property of compositionality links the low-level (WCET) analysis, which computes characteristics of single tasks run in isolation, with schedulability analysis, which accounts for interference on shared resources caused by other tasks. In this talk, we explain the meaning of this compositionality property, demonstrate that it does not come for free even on simple microarchitectures, and show ways how to enable compositionality.

## 6.8 Energy Consumption Analysis and Verification

*Manuel Hermenegildo (IMDEA Software – Madrid, ES)*

License  Creative Commons BY 3.0 Unported license  
© Manuel Hermenegildo

We present our overall approach to the inference and verification of upper- and lower-bounds on the energy consumption of programs, as well as some results from our tools. We translate low-level program representations into a block-based intermediate form, expressed as Horn clauses, and compute abstract minimal models of such Horn clauses on abstract domains that include resource functions on data intervals and sized shapes for structured data. The computed abstract models include, for each procedure, and for each possible abstract call state and path to it, functions that return bounds on the corresponding energy consumed for any given data size, as well as the contribution of each procedure to the overall consumption (static profiling). These analysis results are compared with energy specifications for program verification or (performance) error detection. We will present results for the energy analysis of embedded programs on the XS1-L architecture, making use of ISA- and LLVM-level models of the cost of instructions or sequences of instructions, and compare them to the actual energy consumption measured on the hardware.

## 6.9 Generating Invariants and Resource Bounds with Recurrence Analysis

*Zachary Kincaid (Princeton University, USA)*

License  Creative Commons BY 3.0 Unported license  
© Zachary Kincaid

Compositional recurrence analysis (CRA) is an invariant generation technique that approximates the transitive closure of loops by extracting recurrence relations from the loop body and computing their closed forms. In this talk I will give an overview of CRA and show how it can be used to solve resource bound problems. I will focus particularly on symbolic methods for computing recurrence equations and inequations from a logical representation of a loop body's behavior.



## 6.10 Type systems for Energy Management

*Yu David Liu (SUNY Binghamton, USA)*

**License** © Creative Commons BY 3.0 Unported license  
© Yu David Liu

This talk attempts to bridge two largely disjoint areas of research – type system design and energy management – through typed programming language design for energy-efficient and energy-aware software development. The first part of the talk will focus on Energy Types, a static type system to enable phase-based and mode-based energy management. The second part of the talk will discuss the challenges in promoting proactive and adaptive energy management at the same time, and describe a new programming language called Ent with mixed static type checking and dynamic type checking for addressing these challenges. An open-source compiler built on the ideas of Energy Types and Ent has been implemented, and ported to x86 machines, Android phones, and Raspberry Pi.

## 6.11 Weighted Automata Theory for Resource Analysis of Rewrite Systems

*Georg Moser (Universität Innsbruck, AT)*

**License** © Creative Commons BY 3.0 Unported license  
© Georg Moser

Traditionally derivational and runtime complexity analysis for term rewrite systems (TRSs for short) is performed as restriction of termination methods for TRSs. One such technique is matrix interpretations, which is surprisingly versatile, but maybe costly. In the talk I presented the use of weighted automata theory (aka joint spectral radius theory) to bound the growth rate of matrix interpretations, which in turn yields sharp bounds on the complexity of TRSs.

## 6.12 Solving Recurrences Using Operational Calculus

*Thomas W. Reps (University of Wisconsin – Madison, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Thomas W. Reps  
**Joint work of** Thomas W. Reps, John Cyphert, Jason Breck, Zak Kincaid

This talk describes a tool/technique that is used to solve recurrence equations in the ICRA tool for finding program invariants in non-linear arithmetic. It is based on a variant of Mikusinski's operational calculus. The recurrences are transformed into equations in a field of sequences; the to-be-solved-for sequence is isolated by algebraic manipulations; and the result is transformed back to ordinary algebra. The technique is also applied to solve multivariate recurrences of the kind that arise in a loop that transforms multiple program variables on each iteration.

### 6.13 ECA: Energy Consumption Analysis of software controlled systems

Marko van Eekelen (*University of Nijmegen, NL*)

License  Creative Commons BY 3.0 Unported license  
© Marko van Eekelen

Energy consumption analysis of software controlled systems can play a major role in minimising the overall energy consumption of such systems both during the development phase and later in the field. ECA proposes such an energy analysis, analysing both software and hardware together, to derive the energy consumption of the system when executing. The energy analysis has the property of being adjustable both to the required precision and concerning the hardware used. In principle, this creates the opportunity to analyse which is the best software implementation for given hardware, or the other way around: choose the best hardware for a given algorithm.

The precise analysis is introduced for a high level language, that covers the essentials for control systems. Hardware is modelled as a finite state machine, in which the transitions are function calls that are made explicit in the source code. In these model state changes correspond to energy consumption level. For that reason timing is added to the finite state machines. All the transitions and states in the hardware models are annotated with energy consumptions, to account both for time-dependent and for incidental energy consumptions. A prototype of the ECA system has been developed. It is applied to a small case study.

## 7 Panel discussion and Open problems

During the seminar we had a panel discussion and open problems session on resource bound analysis. Here is our summary of this session with contributions by Jan Reineke and Tom Reps.

### Efficiently computed low and precise time bounds

The most fundamental open problem in the Timing-Analysis area connects the performance of architectures with the efficiency of timing-analysis methods and the precision of their results: how to design architectures that allow the efficient determination of precise and low execution-time bounds.

### Multi-core architectures

While the timing-analysis problem for single-core architectures is solved, there is currently no practically usable timing-analysis method for multi-core processors.

### Non-standard architectures

A completely open problem is how to determine safe execution-time bounds for GPUs, which are heavily used in computer-vision systems in cars, as well as in other types of accelerators. Similar challenges are given from heterogeneous systems including different kinds of architectures, e.g. mobile phones, wearable devices, data centers, etc.

**Other language paradigms**

How can one determine bounds on low-level execution-time and energy consumption for other types of languages, e.g. logic and functional languages and even higher order functional languages?

**Amortized analysis for WCET**

Is it useful to do amortized analyses in the determination of execution-time bounds, in particular for distributed systems, i.e. including communication?

**Side channels**

Modern execution platforms feature a multitude of shared resources, ranging from caches, branch predictors, and DRAM banks to shared functional units in case of hyperthreading. The execution of a program may leave traces on these resources, which side-channel attacks use to infer secrets, such as passwords or private keys. Techniques from resource-bound analysis may be used to quantify the amount of information that a program leaks through such side channels.

**Memory management**

Most modern languages have support for dynamic memory management. Dynamic memory management, however, is at odds with time predictability. While there have been efforts to build dynamic memory allocators and garbage collectors with bounded response times, almost all approaches ignore the memory hierarchy and in particular caches. This makes them unsuitable for hard real-time applications, where it is crucial to precisely account for the cache behavior of an application. It is an open problem to build a fully timing-predictable garbage collector taking into account microarchitectural effects.

**Legacy software**

Legacy software has for the most part been written to be executed on single-core platforms. Transitioning to multi-core platforms frequently uncovers race conditions between different software components. How can such legacy software be retrofitted to enable its predictable use on multi cores?

**Low-level vs high-level analyses**

Techniques for WCET traditionally are closer to the architecture, while techniques for high-level languages use abstract cost models. An open research question is how to integrate these two models. Specifically, how can we link the low-level cost to the high-level cost? Can we design models combining both low-level components and high-level analysis?

**Types as interfaces and specifications**

Several techniques for resource analysis for higher-level languages are based on types. Types can be seen as interfaces abstracting the behaviors of the different components but also as specifications for the components. Can we use types as interfaces and as specifications also for resource bounds? Can we design type-based techniques to combine precise resource bounds of different components whose resource analysis has been performed in isolation?

Can we use these techniques also to account for the cost of communications between the different components?

### **Resource analysis for effects**

Pure functional languages are amenable to simple resource bound analysis. Impure aspects of computations, like memory, I/O, communications, etc. are often encapsulated as general effects. These effects can change the control flow of the program, and add unpredictability to the program behavior. Providing precise resource bound analysis in presence of effects is a challenge.

### **Scalable benchmarks**

Benchmarks are important to evaluate and compare different resource bound analyses. Current benchmark suites include several important examples that are however often designed for the small scale of one machine. Can we also scale them to modern needs of multi-core architectures, heterogeneous computations, cloud computing, etc ?

### **Optimizing Compilers**

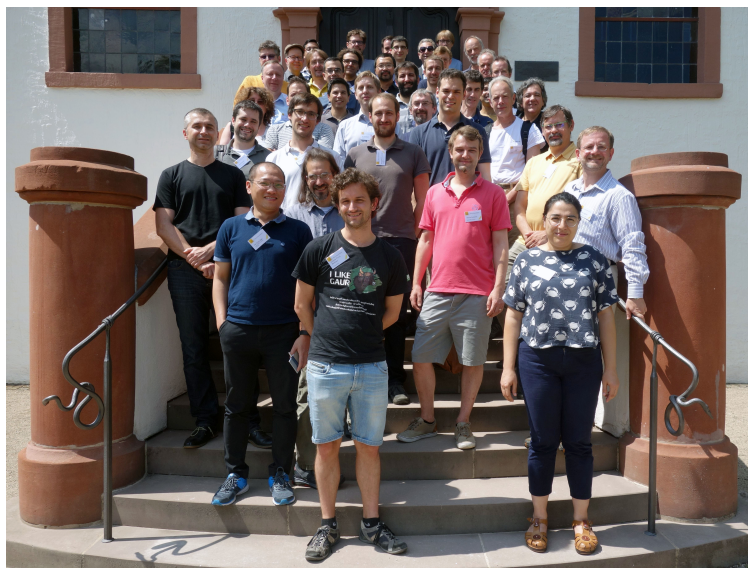
Can we create compilers/optimizers that can explain their results (including whether a given spec. was satisfied)?

### **Performance analysis and Performance Debugging**

Many performance problems are only found after a software has been released. Profilers are the most common technique to understand performance problems. An open problem is how techniques from resource bound analysis can help: Can static techniques be used as an alternative to dynamic profiling techniques? Can resource bound analysis be used to find performance bugs early on during the development process?

## Participants

- Sebastian Altmeyer  
University of Amsterdam, NL
- Robert Atkey  
University of Strathclyde –  
Glasgow, GB
- Martin Avanzini  
Universität Innsbruck, AT
- Gilles Barthe  
IMDEA Software – Madrid, ES
- Marc Brockschmidt  
Microsoft Research UK –  
Cambridge, GB
- Hugues Cassé  
University of Toulouse, FR
- Stephen Chong  
Harvard University –  
Cambridge, US
- Ezgi Cicek  
MPI-SWS – Saarbrücken, DE
- Ugo Dal Lago  
University of Bologna, IT
- Norman Danner  
Wesleyan Univ. –  
Middletown, US
- Ankush Das  
Carnegie Mellon University –  
Pittsburgh, US
- Marco Gaboardi  
University at Buffalo, US
- Deepak Garg  
MPI-SWS – Saarbrücken, DE
- Samir Genaim  
Complutense University of  
Madrid, ES
- Dan R. Ghica  
University of Birmingham, GB
- Jürgen Giesl  
RWTH Aachen, DE
- Reiner Hähnle  
TU Darmstadt, DE
- Sebastian Hahn  
Universität des Saarlandes, DE
- Kevin Hammond  
University of St. Andrews, GB
- Manuel Hermenegildo  
IMDEA Software – Madrid, ES
- Jan Hoffmann  
Carnegie Mellon University –  
Pittsburgh, US
- Martin Hofmann  
LMU München, DE
- Steffen Jost  
LMU München, DE
- Zachary Kincaid  
Princeton University, US
- Jens Knoop  
TU Wien, AT
- Björn Lisper  
Mälardalen University –  
Västerås, SE
- Yu David Liu  
Binghamton University, US
- Alexey Loginov  
GramaTech Inc. – Ithaca, US
- Hans-Wolfgang Loidl  
Heriot-Watt University –  
Edinburgh, GB
- Antonio Flores Montoya  
TU Darmstadt, DE
- Georg Moser  
Universität Innsbruck, AT
- Van Chan Ngo  
Carnegie Mellon University –  
Pittsburgh, US
- Jan Reineke  
Universität des Saarlandes, DE
- Thomas W. Reps  
University of Wisconsin –  
Madison, US
- Christine Rochange  
University of Toulouse, FR
- Claudio Sacerdoti Coen  
University of Bologna, IT
- Marko van Eekelen  
University of Nijmegen, NL
- Pedro B. Vasconcelos  
University of Porto, PT
- Marcus Völz  
University of Luxembourg, LU
- Peter Wägemann  
Universität Erlangen-Nürnberg,  
DE
- Reinhard Wilhelm  
Universität des Saarlandes, DE
- Lukasz Ziarek  
University at Buffalo, US
- Florian Zuleger  
TU Wien, AT



# Topology, Computation and Data Analysis

Edited by

Hamish Carr<sup>1</sup>, Michael Kerber<sup>2</sup>, and Bei Wang<sup>3</sup>

<sup>1</sup> University of Leeds, GB, [h.carr@leeds.ac.uk](mailto:h.carr@leeds.ac.uk)

<sup>2</sup> TU Graz, AT, [kerber@tugraz.at](mailto:kerber@tugraz.at)

<sup>3</sup> University of Utah – Salt Lake City, US, [wang.bei@gmail.com](mailto:wang.bei@gmail.com)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17292 “Topology, Computation and Data Analysis”. This seminar was the first of its kind in bringing together researchers with mathematical and computational backgrounds in addressing emerging directions within computational topology for data analysis in practice. The seminar connected pure and applied mathematicians, with theoretical and applied computer scientists with an interest in computational topology. It helped to facilitate interactions among data theorist and data practitioners from several communities to address challenges in computational topology, topological data analysis, and topological visualization.

**Seminar** July 16–21, 2017 – <http://www.dagstuhl.de/17292>

**1998 ACM Subject Classification** D.2.11 Software Architectures – Data abstraction, F.2.2 Non-numerical Algorithms and Problems – Computations on discrete structures, I 3.5: Computational Geometry and Object Modeling – Geometric algorithms, languages, and systems

**Keywords and phrases** computational topology, topological data analysis, Topological data visualization

**Digital Object Identifier** 10.4230/DagRep.7.7.88

## 1 Executive Summary

*Hamish Carr*

*Michael Kerber*

*Bei Wang*

**License**  Creative Commons BY 3.0 Unported license  
© Hamish Carr, Michael Kerber, and Bei Wang

The Dagstuhl Seminar titled *Topology, Computation and Data Analysis* has brought together researchers with mathematical and computational backgrounds in addressing emerging directions within computational topology for data analysis in practice. The seminar has contributed to the convergence between mathematical and computational thinking, in the development of mathematically rigorous theories and data-driven scalable algorithms.

## Context

In the last two decades, considerable effort has been made in a number of research communities into computational applications of topology. Inherently, topology abstracts functions and graphs into simpler forms, and this has an obvious attraction for data analysis. This attraction is redoubled in the era of extreme data, in which humans increasingly rely on tools that



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Topology, Computation and Data Analysis, *Dagstuhl Reports*, Vol. 7, Issue 7, pp. 88–109

Editors: Hamish Carr, Michael Kerber, and Bei Wang



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

extract mathematically well-founded abstractions that the human can examine and reason about. In effect, topology is applied as a form of data compression or reduction: topology is one of the most powerful forms of mathematical compression that we know how to apply to data.

Efforts to apply topology computationally to data, however, have largely been fragmented so far, with work progressing in a number of communities, principally computational topology, topological data analysis, and topological visualization. Of these, computational topology expands from computational geometry and algebraic topology to seek algorithmic approaches to topological problems, while topological data analysis and topological visualization seeks to apply topology to data analysis, of graphs and networks in the first case and of (usually) simulated volumetric data in the second. The research in these communities can roughly be clustered into theory (what are the underlying mathematical concepts), applications (how are they used for data analysis), and computation (how to compute abstractions for real datasets). It is crucial to advances in this area that these three branches go hand-in-hand, and communication between theoretical, applied, and computational researchers are therefore indispensable. On the other hand, there has been surprisingly little communication between the computational topology and topological visualization communities, mostly caused by the fact that each community has its own set of regular venues. As a consequence, the linkages in the two communities have been independent of each other, and results can take years to migrate from one community to the other.

## Vision

Our goal was therefore to soften the aforementioned rather strict separation between computational topology and topological visualization by establishing new inter-community ties. The seminar aimed to bring together cross sections of both communities, including researchers with theoretical, applied, and computational backgrounds. By reducing redundancy and accelerating cross-communication, we expected a significant boost to both areas, perhaps even leading to a singular more dynamic community. As a side effect, we also wanted to provide a communication platform within each community between theory and application.

## Topics

We identified specific research topics reflecting emerging trends in both communities. These topics were chosen to span the spectrum from the theoretical (category theory), to applicable theory (multidimensional persistent homology), and from applied theory (singularity theory and fiber topology) to the computational (scalable topological computation, applications) aspect.

**Category theory: theory and applications.** Category theory has recently gained momentum in computational topology, in particular through sheaves and cosheaves, which are extremely useful as an alternative foundation for level set persistence. Recent work has shown that the data of a Reeb graph can be stored in a category-theoretic object called a cosheaf, and this opens the way to define a metric for Reeb graphs known as the interleaving distance. Sheaves can also be used in deriving theoretical understandings between the Reeb space and its discrete approximations. Research into sheaves and their relationship with computation is, however, in its infancy, and would benefit from pooling



the resources of experts in category theory and topological data analysis, to address questions such as how to simplify theories in computational topology, how to reinterpret persistent homology, or how to compare topological structures.

**Multidimensional persistent homology.** The second area of active research, both mathematically and computationally, is the extension of unidimensional persistence to multidimensional persistence. Mathematically, the lack of a complete discrete invariant for the multidimensional case raises the theoretical question of identifying meaningful topological invariants to compute. Some earlier proposals have been complemented by recent approaches and raise the immediate question of computability and applicability. Besides the invariants themselves, other questions such as the comparison of multidimensional data, or the efficient generation of cell complexes suitable for the multidimensional case are crucial, but hardly studied questions in this context. Computationally, existing algorithms for topological constructs rely on filtrations to encapsulate a sweep order through the data, thus serializing the problem for algorithmic implementation. For multidimensional data, this serialization is hard to achieve, and progress in this area is, therefore, crucial for computational advances in the topological analysis of data.

**Singularity theory and fiber topology in multivariate data analysis.** Singularity theory and fiber topology both seek to extend Morse theory from scalar fields to multivariate data described as functions mapping  $f : \mathbb{X} \rightarrow \mathbb{R}^d$ . Since multivariate datasets are near-ubiquitous in scientific applications such as oceanography, astrophysics chemistry, meteorology, nuclear engineering and molecular dynamics, advances here are also crucial for topological data analysis and visualization. Methods from computational topology have been developed to support the analysis of scalar field data with widespread applicability. However, very few tools exist for studying multivariate data topologically: the most notable examples of these tools are the Jacobi set, the Reeb space, and its recent computational approximation, the Joint Contour Net. Here, we aim to bring together researchers in singularity theory, fiber topology and topological data analysis to develop new theory and algorithms driving a new generation of analytic tools.

**Scalable computation.** At the opposite pole from theory is the practical question: how do we apply topological analysis to ever-larger data sets? This question spans questions of algorithmic performance to the accuracy of representation: using the metaphor of compression, do we want lossy or lossless compression, how fast can we perform it, and what do we lose in the process? Moreover, the largest data sets are necessarily computed and stored on clusters, and scalability of topological computation therefore also depends on building distributed and parallel algorithms. For example, the standard algorithm for computing persistent homology is cubic in the number of simplices, but can be speeded up in theory and practice, and further improved by parallel computation. However, many challenges remain, including efficient generation, storage and management of simplicial complexes, streaming computation, I/O efficient computation, approximate computation, and non-simplicial complexes. Some of these approaches have already been applied in topological visualization, and cross-fertilization between the two communities is therefore of great interest.

## Participants, Schedule, and Organization

The invitees were chosen according to the topics, bringing together enough expertise for each topic and resulting in a representative subset of both communities. Out of the 37 invited



researchers in the first round, 28 accepted our invitation, pointing out the general interest for the seminar topic in both communities.

We decided for a mixed setup with introductory talks, contributed research talks and breakout sessions.

For the first day, we scheduled two overview talks per listed topic, which were delivered by Steve Oudot and Elizabeth Munch (Category theory), Michael Lesnick and Claudia Landi (Multidimensional persistent homology), Osamu Saeki and Julien Tierny (Singularity theory), and Yusu Wang and Valerio Pascucci (Scalable Computation). Further contributed talks by participants took place from Tuesday to Friday morning, resulting in a total of 19 contributed talks.

The afternoons of Tuesday and Thursday were used for breakout sessions. The format was different on the two days. Based on the discussions on Monday, we identified the topics “multivariate topology” and “scalable computation” as topics of general interest. We decided to let every participant discuss both topics, so we organized 4 discussion groups on multivariate topology in the early afternoon, and 3 discussion groups on scalable computation in the later afternoon (plus an alternative group with a different topic). We composed these groups mostly randomly, making sure that members of both communities are roughly balanced in each group. On Thursday afternoon, we let participants propose their topics of interest. 5 groups were formed discussing various aspects raised in contributed talks. On Wednesday and Friday morning, the outcomes of every discussion group were summarized and discussed in a plenary session.

Moreover, the majority of the participants joined an organized excursion to Trier on Wednesday afternoon.

## Results and Reflection

The participants gave the unanimous feedback that the breakout sessions were a full success (and several proposed more time for such discussions in possible upcoming seminars). We first let people from a mixed background to discuss rather vague topics on Tuesday, and asked for specific topics on Thursday. Such an organizational plan led to a stimulating working environment, and helped to avoid idle breakout sessions.

We believe that we have fully achieved the goal of softening the separation between the two communities involved in this seminar. We expect visible evidence of newly formed inter-community ties fostered by the seminar, for instance through joint research projects and/or survey articles summarizing major open problems on the interface of both communities. To the best of our knowledge, 3 working groups are being formed and at least 1 position paper is underway that will combine expertise from both communities to tackle key research questions raised during the seminar.

## 2 Table of Contents

### Executive Summary

<i>Hamish Carr, Michael Kerber, and Bei Wang</i> . . . . .	88
--	----

### Overview of Talks


Rips: Efficient Computation of Vietoris-Rips Persistence Barcodes <i>Ulrich Bauer</i> . . . . .	94
Interleaving Distance: Computational Complexity <i>Magnus Botnan</i> . . . . .	94
Dataflow EDSL: Parallel Topology Made Simple <i>Peer-Timo Bremer</i> . . . . .	94
What is Wrong with Time-Dependent Flow Topology? <i>Roxana Bujack</i> . . . . .	95
Reeb Spaces, Fiber Surfaces and Joint Contour Nets <i>Hamish Carr</i> . . . . .	95
A Discrete Gradient-Based Approach to Multivariate Data Analysis <i>Leila De Floriani</i> . . . . .	95
Representations of Persistence and Time-Varying Persistence: Past, Present and Future <i>Pawel Dlotko</i> . . . . .	96
Topology-Guided Visual Exploratory Analysis <i>Harish Doraiswamy</i> . . . . .	96
Persistence-Based Summaries for Metric Graphs <i>Ellen Gasparovic</i> . . . . .	97
Robust Extraction and Simplification of 2D Tensor Field Topology <i>Ingrid Hotz</i> . . . . .	97
The representation theorem of persistent homology revisited and generalized <i>Michael Kerber</i> . . . . .	98
Introduction to multidimensional persistent homology II <i>Claudia Landi</i> . . . . .	98
Topology Meets Machine Learning: How Both Fields Can Profit From Each Other <i>Heike Leitte</i> . . . . .	99
An Introduction to Multidimensional Persistent Homology I <i>Michael Lesnick</i> . . . . .	99
Introduction to categorical approaches in topological data analysis II <i>Elizabeth Munch</i> . . . . .	99
Feature-Directed Visualization of Multifield Data <i>Vijay Natarajan</i> . . . . .	100
Introduction to Categorical Approaches in Topological Data Analysis I <i>Steve Y. Oudot</i> . . . . .	100
A Stable Multi-Scale Kernel for Topological Machine Learning <i>Jan Reininghaus</i> . . . . .	101

Introduction to Singularity Theory and Fiber Topology in Multivariate Data Analysis	
<i>Osamu Saeki</i> . . . . .	101
A Topological Visualization Approach to Combinatorial Optimization	
<i>Gerik Scheuermann</i> . . . . .	102
Noise Systems and Multidimensional Persistence	
<i>Martina Scolamiero</i> . . . . .	102
Discrete Morse Theory and Simplicial Map Persistence	
<i>Donald Sheehy</i> . . . . .	102
Spectral Sequences for Parallel Computation	
<i>Primož Skraba</i> . . . . .	103
Topological Analysis of Bivariate Data	
<i>Julien Tierny</i> . . . . .	103
Generalizations of the Rips Filtration for Quasi-Metric Spaces with Corresponding Stability Results	
<i>Katharine Turner</i> . . . . .	104
Scalable Computation in Computational Topology	
<i>Yusu Wang</i> . . . . .	104
<b>Working groups</b>	
Discussion group on “Mean Reeb Graphs”	
<i>Ellen Gasparovic, Peer-Timo Bremer, Ingrid Hotz, Elizabeth Munch, Vijay Natarajan, Steve Y. Oudot, Julien Tierny, Katharine Turner, and Bei Wang</i> . . . . .	105
Discussion Group on “Multivariate Topology”	
<i>Michael Kerber, Harish Doraiswamy, Christoph Garth, Claudia Landi, Michael Lesnick, Jan Reininghaus, and Julien Tierny</i> . . . . .	106
Discussion Group on “Scalable Computation”	
<i>Michael Kerber, Peer-Timo Bremer, Leila De Florian, Michael Lesnick, Vijay Natarajan, and Primož Skraba</i> . . . . .	107
<b>Participants</b> . . . . .	109

### 3 Overview of Talks

#### 3.1 Ripser: Efficient Computation of Vietoris-Rips Persistence Barcodes

*Ulrich Bauer (TU München, DE)*

**License**  Creative Commons BY 3.0 Unported license  
© Ulrich Bauer

I will discuss the efficient computation of the Vietoris-Rips persistence barcode for a finite metric space. The implementation in the newly developed C++ code “Ripser” focuses on memory and time efficiency, outperforming previous software by a factor of more than 40 in computation time and a factor of more than 15 in memory efficiency on typical benchmark examples. The improved computational efficiency is based on a close connection between persistent homology and discrete Morse theory, together with novel algorithmic design principles, avoiding the explicit construction of the filtration boundary matrix.

#### 3.2 Interleaving Distance: Computational Complexity

*Magnus Botnan (TU München, DE)*

**License**  Creative Commons BY 3.0 Unported license  
© Magnus Botnan

The computational complexity of computing the interleaving distance for multi-parameter persistent homology is not known. I will discuss a special instance of the problem which I believe is NP-Hard.

#### 3.3 Dataflow EDSL: Parallel Topology Made Simple

*Peer-Timo Bremer (LLNL – Livermore, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Peer-Timo Bremer

Efficient and scalable implementations, especially of more complex analysis approaches, require not only advanced algorithms but also an in-depth knowledge of the underlying runtime. Furthermore, different machine configurations and different applications may favor different runtimes, i.e., MPI vs. Charm++ vs Legion etc., and different hardware architectures. This diversity makes developing and maintaining a broadly applicable analysis software infrastructure challenging. We address some of these problems by explicitly splitting the definition and implementation of analysis and visualization algorithms. In particular, we present an embedded domain specific language (EDSL) to describe an algorithm as a generic task graph, that can be executed with different runtime backends (MPI, Charm++, Legion). We demonstrate the flexibility and performance of this approach using three different large-scale analysis and visualization use cases, i.e., topological analysis, rendering and compositing dataflow, and image registration of large microscopy scans. Despite the unavoidable overheads of a generic solution, our approach demonstrates performance portability at scale, and, in some cases, outperforms hand-optimized implementations.

### 3.4 What is Wrong with Time-Dependent Flow Topology?

Roxana Bujack (*Los Alamos National Laboratory, US*)

**License** © Creative Commons BY 3.0 Unported license  
© Roxana Bujack

Vector field topology is a powerful visualization tool, because it can break down huge amounts of data into a compact, sparse, and easy-to-read description with little information loss. Visualization scientists struggle, because its generalization to time-dependent flow usually lacks a meaningful physical interpretation. We are looking for ways to overcome this problem.

### 3.5 Reeb Spaces, Fiber Surfaces and Joint Contour Nets

Hamish Carr (*University of Leeds, GB*)

**License** © Creative Commons BY 3.0 Unported license  
© Hamish Carr  
**Joint work of** Hamish Carr, Julien Tierney, Aaron Knoll, David Duke, Amit Chattopadhyay, Zhao Geng, Osamu Saeki, Kui Wu, Pavol Klacansky, Valerio Pascucci  
**Main reference** Kui Wu, Aaron Knoll, Benjamin J. Isaac, Hamish A. Carr, Valerio Pascucci: “Direct Multifield Volume Ray Casting of Fiber Surfaces”, *IEEE Trans. Vis. Comput. Graph.*, Vol. 23(1), pp. 941–949, 2017.  
**URL** <http://dx.doi.org/10.1109/TVCG.2016.2599040>

Recent work in topological visualization has developed a set of tools for bivariate functions of the form  $\mathbb{R}^3 \rightarrow \mathbb{R}^2$ . Here, the analog of an isosurface is a fiber surface – the 2-manifold pre-image of a 1-manifold curve in the range. From this, the Reeb graph extends naturally to the Reeb space, which for bivariate functions is a 2-cell complex. This talk will give a summary of these recent developments, including variations on fiber surfaces and Reeb spaces, and some of the application-oriented results arising from the Joint Contour Net – a quantized approximation of the Reeb space.

### 3.6 A Discrete Gradient-Based Approach to Multivariate Data Analysis

Leila De Floriani (*University of Maryland – College Park, US*)


**License** © Creative Commons BY 3.0 Unported license  
© Leila De Floriani  
**Joint work of** Leila De Floriani, Federico Iuricich, Riccardo Fellegara, Sara Scaramuccia, Claudia Landi, Kenneth Weiss  
**Main reference** Federico Iuricich, Sara Scaramuccia, Claudia Landi, Leila De Floriani: “A discrete morse-based approach to multivariate data analysis”, in *Proc. of the SIGGRAPH ASIA 2016*, Macao, December 5-8, 2016 - Symposium on Visualization, pp. 5:1–5:8, ACM, 2016.  
**URL** <http://dx.doi.org/10.1145/3002151.3002166>

In this talk, we present our recent work on topological analysis of big data based on discrete Morse theory, for applications to the efficient computation of (multi)persistent homology and to topology-based data visualization. In the first part of the talk, a new distributed data structure for simplicial complexes, the Stellar tree, is presented, which allows for an efficient generation and compact storage of the discrete Morse gradient and for an effective and efficient computation of the discrete Morse complex and its geometric embedding on very large data sets. Compactness and computational efficiency of the Stellar tree are demonstrated in comparison with state-of-the-art data structures for simplicial complexes. The second part

of the talk has been focused on the case of multivariate data, i.e., data equipped with a vector-valued function. Such problem is especially relevant for computing multipersistent homology efficiently on large data sets and for investigating and extracting critical features of multivariate data, such as Pareto or Jacobi sets. Specifically, a new approach based on a discrete gradient compatible with the vector-valued function is presented, which has been proven to generate a chain complex which has the same persistent homology as the original input complex. This allows to drastically reduce the time and space required to compute the multipersistent module, as the results of our experiments with the public domain tool for multipersistent homology computation. Moreover, our preliminary results show theoretically anticipated connections between the critical simplices associated with the discrete gradient and Pareto sets, and form the basis for the current developments of this research.

### 3.7 Representations of Persistence and Time-Varying Persistence: Past, Present and Future

*Pawel Dlotko (Swansea University, GB)*

**License**  Creative Commons BY 3.0 Unported license  
© Pawel Dlotko

Standard computational topology pipeline barely considers the problem of post-processing of persistence diagrams. Yet, in data analysis, this is an important, if not essential step. Classical tools that allow for basic analysis of persistence diagrams are restricted to Wasserstein and Bottleneck distances. Yet, to use persistence as an input for standard statistics and machine learning algorithms, one requires more: in addition to be able to compute a distance between diagrams, one may need to average them, compute their scalar products, confidence bounds and similar. Some of those operations can be performed on persistence diagrams, but a lot of them are ambiguous on persistence diagrams. To address this issue, we will introduce various representations of persistence diagrams that implement all the mentioned operations. We will speculate on general, data-dependent representations and kernels, and discuss the existing implementations, including the implementation in Gudhi library. At the end, we will generalize all the introduced representations for time-varying persistence diagrams.

### 3.8 Topology-Guided Visual Exploratory Analysis

*Harish Doraiswamy (New York University, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Harish Doraiswamy

**Joint work of** Alex Bock, Theodoros Damoulas, Harish Doraiswamy, Nivan Ferreira, Juliana Freire, Claudio Silva, Adam Summers

**Main reference** Harish Doraiswamy, Nivan Ferreira, Theodoros Damoulas, Juliana Freire, Cláudio T. Silva: “Using Topological Analysis to Support Event-Guided Exploration in Urban Data”, IEEE Trans. Vis. Comput. Graph., Vol. 20(12), pp. 2634–2643, 2014.

**URL** <http://dx.doi.org/10.1109/TVCG.2014.2346449>

Enormous amounts of data are being collected in different domains, from traditional ones such as biology to the more recent urban sciences. This has created new opportunities for using data-driven approaches to better support answering important questions that arise in these domains. Visualization and visual analytics systems have been successfully used to aid users obtain insight. However, manual (exhaustive) exploration of large data sets is not

only time consuming, but often becomes impractical. It is therefore necessary to also guide users during this exploration process. Furthermore, it is also important that these tools be designed in a way that they are usable and within reach of domain experts who often lack computer science expertise. In this talk, I will present a few examples of how techniques from computational topology used in conjunction with visualization has been instrumental in guiding domain experts in their analysis process.

### 3.9 Persistence-Based Summaries for Metric Graphs

*Ellen Gasparovic (Union College – Schenectady, US)*

**License** © Creative Commons BY 3.0 Unported license  
 © Ellen Gasparovic  
**Joint work of** Ellen Gasparovic, Maria Gommel, Emilie Purvine, Radmila Sazdanovic, Bei Wang, Yusu Wang, and Lori Ziegelmeier  
**Main reference** Ellen Gasparovic, Maria Gommel, Emilie Purvine, Radmila Sazdanovic, Bei Wang, Yusu Wang, Lori Ziegelmeier, “A Complete Characterization of the 1-Dimensional Intrinsic Cech Persistence Diagrams for Metric Graphs”, arXiv:1702.07379v2 [math.AT], 2017.  
**URL** <https://arxiv.org/abs/1702.07379>

In this talk, we focus on giving a qualitative description of information that one can capture from metric graphs using certain topological summaries. In particular, we give a complete characterization of the persistence diagrams in dimension 1 for metric graphs under a particular intrinsic setting. We also look at two persistence-based distances that one may define for metric graphs and discuss progress toward establishing their discriminative capacities.

### 3.10 Robust Extraction and Simplification of 2D Tensor Field Topology

*Ingrid Hotz (Linköping University, SE)*

**License** © Creative Commons BY 3.0 Unported license  
 © Ingrid Hotz  
**Joint work of** Ingrid Hotz, Bei Wang, Jochen Jankowai

In this work, we propose a controlled simplification and smoothing strategy for symmetric 2D tensor fields that is based on the topological notion of robustness. Robustness measures the structural stability of the degenerate points with respect to variation of the underlying field. We consider an entire pipeline for the topological simplification of the tensor field by generating a hierarchical set of simplified fields based on varying the robustness values. Such a pipeline comprises of four steps: the stable extraction and classification of degenerate points, the computation and assignment of robustness values to the degenerate points, the construction of a simplification hierarchy, and finally the actual smoothing of the fields across multiple scales. We also discuss the challenges that arise from the discretization and interpolation of real world data.

### 3.11 The representation theorem of persistent homology revisited and generalized

*Michael Kerber (TU Graz, AT)*

**License** © Creative Commons BY 3.0 Unported license  
© Michael Kerber

**Joint work of** Rene Corbet, Michael Kerber

**Main reference** Rene Corbet, Michael Kerber, “The representation theorem of persistent homology revisited and generalized”, arXiv:1707.08864v2 [math.AT], 2017.

**URL** <https://arxiv.org/abs/1707.08864>

The representation theorem by Zomorodian and Carlsson has been the starting point of the study of persistent homology under the lens of algebraic representation theory. In this work, we give a more accurate statement of the original theorem and provide a complete and self-contained proof. Furthermore, we generalize the statement from the case of linear sequences of  $R$ -modules to  $R$ -modules indexed over more general monoids. This generalization subsumes the representation theorem of multidimensional persistence as a special case.

### 3.12 Introduction to multidimensional persistent homology II

*Claudia Landi (University of Modena, IT)*

**License** © Creative Commons BY 3.0 Unported license  
© Claudia Landi

**Joint work of** Claudia Landi, Sara Scaramuccia, Federico Iuricich, Leila De Florian

Many scientific fields need to study multivariate data. Multivariate data can be represented by multiple real-valued functions  $f_1, f_2, \dots, f_n : M \rightarrow \mathbb{R}$  defined on the same domain  $M$ , giving rise to a vector-valued function  $f = (f_i) : M \rightarrow \mathbb{R}^n$ . A sublevel set  $M^u$  of  $f$  at  $u = (u_i) \in \mathbb{R}^n$  consists of those points  $p$  of  $M$  such that  $f_i(p) \leq u_i$  for every  $1 \leq i \leq n$ . Varying  $u$  in  $\mathbb{R}^n$  produces a multiparameter filtration of  $M$  by sublevel sets where  $u_i \leq v_i$  for every  $1 \leq i \leq n$  implies  $M^u \subseteq M^v$ . Multidimensional persistence detects the appearance and disappearance of homology features along this filtration with the multiparameter  $u$  varying in any increasing direction.


In the case when  $M$  is a smooth manifold and  $f$  is a smooth function, the values of the multiparameter  $u$  where homology features appear and disappear correspond to values taken at Pareto critical points. Intuitively, these are points where the gradients of the functions  $f_i$  disagree.

In the case when  $M$  is a simplicial complex and  $f$  is defined on its vertices and then extended to any simplex by taking the component-wise maximum over its vertices, a discrete gradient field compatible with the induced sublevel set filtration can be obtained by an algorithm based on homotopy expansion. The critical cells of such gradient field detect locations where homology classes are born and die along the filtration. In other words, they play, in the discrete setting, a role similar to that played by Pareto critical points in the smooth setting.



### 3.13 Topology Meets Machine Learning: How Both Fields Can Profit From Each Other

*Heike Leitte (TU Kaiserslautern, DE)*

**License**  Creative Commons BY 3.0 Unported license  
© Heike Leitte

Machine learning tries to reconstruct models from data, searching for underlying signals and patterns in it. A major hurdle is commonly noise and variations present in all real world data. Topological data analysis (TDA) provides a great set of tools to search for salient features in complex data, while filtering noise and short lived signals. In this talk, we will look at the major application fields of machine learning and see some examples of how they can be improved using modern TDA algorithms. We will also explore how visualisation can help to connect these two data analysis fields and make the results and the analysis process more easily accessible to the user.

### 3.14 An Introduction to Multidimensional Persistent Homology I

*Michael Lesnick (Princeton University, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Michael Lesnick

In topological data analysis, we often study data by associating to the data a filtered topological space, whose structure we can then examine using persistent homology. However, in many settings, a single filtered space is not a rich enough invariant to encode the interesting structure of our data. This motivates the study of multidimensional persistence, which associates to the data a topological space simultaneously equipped with two or more filtrations. The homological invariants of these “multi-filtered spaces,” while much richer than their 1-dimensional counterparts, are also far more complicated. As such, adapting the usual 1-dimensional persistent homology methodology for data analysis to the multi-dimensional setting requires some new ideas. In this talk, I’ll introduce multi-dimensional persistent homology and discuss some recent progress on this topic.

### 3.15 Introduction to categorical approaches in topological data analysis II

*Elizabeth Munch (Michigan State University, US)*


**License**  Creative Commons BY 3.0 Unported license  
© Elizabeth Munch

Arguably, the most beautiful mathematical idea coming from topological data analysis in the last decade is that of interleaving. An  $\epsilon$ -interleaving can be thought of as an approximate isomorphism between two persistence modules with  $\epsilon$ -allowed incorrectness. Using basic structures from category theory, one can think of a persistence module as a functor, then the  $\epsilon$ -interleaving is a set of natural transformations between  $\epsilon$ -shifted persistence modules which satisfy certain compatibility conditions. Once these ideas have been extended to category theory, we can then use the idea of interleaving for different choices of categories and functors

to obtain known metrics, including bottleneck distance for persistence modules, Hausdorff distance for sets, and  $L_\infty$  for points or functions; as well as create new metrics for many disparate objects including Reeb graphs, mapper graphs, and multi-dimensional persistence modules.

### 3.16 Feature-Directed Visualization of Multifield Data

*Vijay Natarajan (Indian Institute of Science – Bangalore, IN)*

**License**  Creative Commons BY 3.0 Unported license  
© Vijay Natarajan


**Main reference** Vidya Narayanan, Dilip Mathew Thomas, Vijay Natarajan: “Distance between extremum graphs”, in Proc. of the 2015 IEEE Pacific Visualization Symposium, PacificVis 2015, Hangzhou, China, April 14-17, 2015, pp. 263–270, IEEE Computer Society, 2015.

**URL** <http://dx.doi.org/10.1109/PACIFICVIS.2015.7156386>

Scientific phenomena are often studied through collections of related scalar fields generated from different observations of the same phenomenon. Exploration of such data requires a robust distance measure to compare scalar fields for tasks such as identifying key events and establishing a correspondence between features in the data. In this talk, I will pose the problem of designing appropriate distance measures to compare scalar fields in a feature-aware manner. Assuming that topological structures represent features in the data, what are good approaches towards the design of feature-aware distance measures between the scalar fields. In addition to provable properties, we will require the distance measure to be efficiently computable, and also interpretable.

### 3.17 Introduction to Categorical Approaches in Topological Data Analysis I

*Steve Y. Oudot (INRIA Saclay – Île-de-France, FR)*

**License**  Creative Commons BY 3.0 Unported license  
© Steve Y. Oudot

The mathematical theory underlying topological data analysis, which is known as persistence theory, works at two different levels: the topological level, where it deals with nested families of topological spaces, as inspired from Morse theory; the algebraic level, where it deals with diagrams of vector spaces and linear maps, as inspired from quiver representation theory. While the objects involved in these two levels are very different in nature, they can be thought of as functors from partially ordered sets to some target categories. Category theory appears then as the right tool to build an abstraction of persistence, in which both levels can be cast and analyzed in tandem. This talk is therefore naturally divided into two parts: first, an introduction to the basics of category theory; second, an introduction to 1-dimensional persistence theory and its foundational results (decomposition, stability) from a categorical point of view.

### 3.18 A Stable Multi-Scale Kernel for Topological Machine Learning

*Jan Reininghaus (Siemens Industry Software GmbH – Wien, AT)*

**License** © Creative Commons BY 3.0 Unported license  
© Jan Reininghaus

**Joint work of** Jan Reininghaus, Stefan Huber, Ulrich Bauer, Roland Kwitt

**Main reference** Jan Reininghaus, Stefan Huber, Ulrich Bauer, Roland Kwitt: “A stable multi-scale kernel for topological machine learning”, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 4741–4748, IEEE Computer Society, 2015.

**URL** <http://dx.doi.org/10.1109/CVPR.2015.7299106>

Topological data analysis offers a rich source of valuable information to study vision problems. Yet, so far we lack a theoretically sound connection to popular kernel based learning techniques, such as kernel SVMs or kernel PCA. In this work, we establish such a connection by designing a multi-scale kernel for persistence diagrams, a stable summary representation of topological features in data. We show that this kernel is positive definite and prove its stability with respect to the 1-Wasserstein distance. Experiments on two benchmark datasets for 3D shape classification and texture recognition show considerable performance gains of the proposed method compared to an alternative approach that is based on the recently introduced persistence landscapes.

### 3.19 Introduction to Singularity Theory and Fiber Topology in Multivariate Data Analysis

*Osamu Saeki (Kyushu University – Fukuoka, JP)*

**License** © Creative Commons BY 3.0 Unported license  
© Osamu Saeki

**Joint work of** Yamamoto, Takahiro; Kawashima, Masayuki; Hiratuka, Jorge T.

**Main reference** Osamu Saeki, “Theory of singular fibers and Reeb spaces for visualization, Topological Methods in Data Analysis and Visualization IV – Theory, Algorithms, and Applications”, Proc. Topology-Based Methods in Visualization 2015, pp. 3–33, Springer, 2017.

**URL** <https://doi.org/10.1007/978-3-319-44684-4>

In this talk, we consider generic differentiable maps between differentiable manifolds, and propose a mathematical formulation of fibers from the viewpoint of singularity theory. In fact, this formulation is shown to be essential also for visualization purposes. Then, classification results of fibers for certain dimension pairs are presented. We also present results on local characterizations of Reeb spaces. Our study of fibers for maps of 4-dimensional manifolds into surfaces indicates a (possibly new) concept of a Reeb diagram, which is expected to be a source of new problems. Some computational problems will also be presented from mathematical viewpoints.


#### References

- 1 O. Saeki, Topology of singular fibers of differentiable maps, Lecture Notes in Math., Vol. 1854, Springer Verlag, 2004.
- 2 O. Saeki, Theory of singular fibers and Reeb spaces for visualization, Topological Methods in Data Analysis and Visualization IV – Theory, Algorithms, and Applications, H. Carr, C. Garth, T. Weinkauff (Eds.), Proc. Topology-Based Methods in Visualization 2015, pp. 3–33, Springer, 2017.
- 3 O. Saeki and T. Yamamoto, Singular fibers of stable maps of 3-manifolds with boundary into surfaces and their applications, Algebraic and Geometric Topology 16, 1379–1402, 2016.

- 4 D. Sakurai, O. Saeki, H. Carr, Hsiang-Yun Wu, T. Yamamoto, D. Duke, and S. Takahashi, Interactive visualization for singular fibers of functions  $f : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ , IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 945–954, 2016.

### 3.20 A Topological Visualization Approach to Combinatorial Optimization


*Gerik Scheuermann (Universität Leipzig, DE)*

License  Creative Commons BY 3.0 Unported license  
© Gerik Scheuermann

Topological data visualization has been applied to many different domain. In this talk, we take a look at a discipline that has rarely been studied by topological visualization – despite a very close relation to topology. We look at combinatorial optimization. In this discipline, there are many topological considerations, but hardly any topological visualization approaches. Obviously, optimization is hard if there are many local extrema, otherwise it is easy. Therefore, a topological study of the problem can provide insight. We show that for enumerable problems (even with millions or billions of points), topological visualization allows to visually study algorithmic behavior which is not possible with typically used visualizations. Thus, while these problems are still very small, parameter tuning and testing of optimization algorithms is simplified. For larger instances, we show that sampling of the landscape is a promising direction to go.

### 3.21 Noise Systems and Multidimensional Persistence


*Martina Scalamiero (EPFL – Lausanne, CH)*

License  Creative Commons BY 3.0 Unported license  
© Martina Scalamiero

In this talk I will introduce a framework that allows to compute a new class of stable discrete invariants for multidimensional persistence. In doing this, we generalise the notion of interleaving topology on multidimensional persistence modules by using noise systems. A filter function is usually chosen to highlight properties we want to examine from a dataset. Similarly, our new topology allows some features of datasets to be considered as noise.

### 3.22 Discrete Morse Theory and Simplicial Map Persistence

*Donald Sheehy (University of Connecticut – Storrs, US)*

License  Creative Commons BY 3.0 Unported license  
© Donald Sheehy

One efficient way to compute the persistent homologous of simplicial maps involves converting the sequence of complexes into a proper filtration. In this talk, I will show that this approach follows naturally from discrete Morse theory on the mapping telescope.

### 3.23 Spectral Sequences for Parallel Computation

*Primož Skraba (Jozef Stefan Institute – Ljubljana, SI)*

**License** © Creative Commons BY 3.0 Unported license  
© Primož Skraba

Spectral sequences represent a family of incremental algorithms for computing topological invariants. They are a fundamental tool used by both algebraic topologists and homological algebraists, most often to compute (co)homology, although in some cases also more difficult invariants such as homotopy groups. Often they are difficult to follow due to extensive notation and because they are generally applied to difficult problems (making the examples themselves difficult). In our case, however, the Mayer-Vietoris spectral sequence, a special case of the Leray spectral sequence (which is itself a special case of the Grothendieck spectral sequence), provides an algorithm for computing (co)homology. In this talk, we introduced what the spectral sequence is from an algorithmic point of view. We showed how to set it up and the operations which must be efficiently implemented in order to make the algorithm as a whole efficient. We concentrated on the structure which indicates how much parallelization can be achieved with this approach as well as discuss the obstacles which remain in order to extend this to persistence.

### 3.24 Topological Analysis of Bivariate Data

*Julien Tierny (CNRS-UPMC – Paris, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Julien Tierny


**Main reference** Julien Tierny, Guillaume Favelier, Joshua A. Levine, Charles Gueunet, Michael Michaux: “The Topology ToolKit”, IEEE Trans. Vis. Comput. Graph., Vol. 24(1), pp. 832–842, 2018.

**URL** <http://dx.doi.org/10.1109/TVCG.2017.2743938>

Multivariate scalar data sets are becoming increasingly popular in scientific visualization applications, since modern numerical simulations and acquisition devices have now the ability to simultaneously track a large number of variables on a single geometrical domain. Thus, the topological methods developed over the last twenty years for the analysis and visualization of univariate scalar data need to be completely revisited in that setting. The bivariate case is an appealing first step in this generalization effort, in particular since users often tend to project multivariate functions to the (two-dimensional) screen in the form of 2D scatterplots for visualization purposes. This talk reviews recent algorithms for the extension to the bivariate case of the notions of level sets (to fibers), critical points (to Jacobi sets) and Reeb graphs (to bivariate Reeb spaces). Applications to continuous scatterplot peeling, silhouette simplification, medial structure computation and feature similarity estimation are discussed. Finally, I will present current research problems, including the question of Jacobi set simplification, for which solutions are expected to eventually enable a wide adoption of bivariate topological data analysis in scientific visualization applications.

### 3.25 Generalizations of the Rips Filtration for Quasi-Metric Spaces with Corresponding Stability Results


*Katharine Turner (Australian National University – Canberra, AU)*

License  Creative Commons BY 3.0 Unported license  
© Katharine Turner

Rips filtrations over a finite metric space and their corresponding persistent homology are prominent methods in Topological Data Analysis to summarize the “shape” of data. For finite metric space  $X$  and distance  $r$  the traditional Rips complex with parameter  $r$  is the flag complex whose vertices are the points in  $X$  and whose edges are  $\{[x, y] : d(x, y) \leq r\}$ . From considering how the homology of these complexes evolves we can create persistence modules (and their associated barcodes and persistence diagrams). Crucial to their use is the stability result that says if  $X$  and  $Y$  are finite metric space then the bottleneck distance between persistence modules constructed by the Rips filtration is bounded by  $2d_{GH}(X, Y)$  (where  $d_{GH}$  is the Gromov-Hausdorff distance). Using the asymmetry of the distance function we construct four different constructions analogous to the persistent homology of the Rips filtration and show they also are stable with respect to the Gromov-Hausdorff distance. These different constructions involve ordered-tuple homology, symmetric functions of the distance function, strongly connected components and poset topology.

### 3.26 Scalable Computation in Computational Topology

*Yusu Wang (Ohio State University – Columbus, US)*

License  Creative Commons BY 3.0 Unported license  
© Yusu Wang

Recent years have witnessed a tremendous amount of growth in the field of computational topology. In addition to significant theoretical and algorithmic developments, topological methods have been used in various application domains, including in visualization. With the rapid increase in the number of applications and in the scale of data sizes, it is important that topological methods are scalable and can handle the challenge of mass data sizes. In this talk, I will survey some of the algorithmic efforts for topological methods, with a special focus on the computation of persistent homology (in various settings). I will also briefly touch on the computation of the Reeb graph and related structures.

## 4 Working groups

### 4.1 Discussion group on “Mean Reeb Graphs”

*Ellen Gasparovic (Union College – Schenectady, US), Peer-Timo Bremer (LLNL – Livermore, US), Ingrid Hotz (Linköping University, SE), Elizabeth Munch (Michigan State University, US), Vijay Natarajan (Indian Institute of Science – Bangalore, IN), Steve Y. Oudot (INRIA Saclay – Île-de-France, FR), Julien Tierny (CNRS-UPMC – Paris, FR), Katharine Turner (Australian National University – Canberra, AU), and Bei Wang (University of Utah – Salt Lake City, US)*

**License** © Creative Commons BY 3.0 Unported license

© Ellen Gasparovic, Peer-Timo Bremer, Ingrid Hotz, Elizabeth Munch, Vijay Natarajan, Steve Y. Oudot, Julien Tierny, Katharine Turner, and Bei Wang

The goal of our working group was to make precise the notion of a *mean Reeb graph* and discuss how to compute it. We agreed that such a mean should be a descriptive statistic that lends itself readily to topological and geometric interpretation.

We began by asking ourselves many questions, including:

- If we obtain the same Reeb graph from two different functions, should the mean be the same independent of the functions giving rise to them?
- Do we want the metric we use to depend on the original functions?
- Should the mean depend on the application?
- Should we use a feature-based distance and/or a spatial distance?
- Would some sort of *augmented* or *labeled* Reeb graph be desirable?
- Should we consider a whole set of Fréchet means from different metrics?
- Is this related to the notion of “tensor swelling”? Or perhaps uncertainty visualization for contour trees?

We then discussed possible metrics that one can define on the space of Reeb graphs, so that the associated Fréchet means of sets of Reeb graphs have nice properties. Possibilities included the functional Gromov-Hausdorff distance, interleaving and functional distortion (FD) distances [1, 2], as well as the persistence distortion distance [4]. We decided to first look at several simple examples involving pairs of Reeb graphs, figure out what we thought the means should be in those instances, and then find a distance that would yield the desired means.

Later in the week, we focused on the induced intrinsic bottleneck distance  $\hat{d}_B$  of [3], i.e., given Reeb graphs  $R_f$  and  $R_g$ , we have

$$\hat{d}_B(R_f, R_g) := \inf_{\gamma} |\gamma|_B$$

where  $\gamma$  ranges over all paths  $\gamma : [0, 1] \rightarrow \text{Reeb}$  ( $\gamma(0) = R_f$  and  $\gamma(1) = R_g$ ) that are continuous in  $d_{FD}$ ,  $|\gamma|_B = \sup_{n, \Sigma} \sum_{i=1}^{n-1} d_B(\gamma(t_i), \gamma(t_{i+1}))$  ( $n \in \mathbb{N}$  and  $\Sigma$  ranges over all partitions  $0 = t_0 \leq t_1 \leq \dots \leq t_n = 1$  of  $[0, 1]$ ), and  $d_B$  is the usual bottleneck distance. This distance has many nice properties, including the fact that it is globally equivalent to the similarly defined intrinsic version of the functional distortion distance,  $\hat{d}_{FD}$ , which implies that they both induce the same topology on *Reeb* [3].


The next step is to prove that the bottleneck distance between Reeb graphs is locally intrinsic in a certain sense, in the same way as Carrière and Oudot showed that it is in the space of persistence diagrams.

## References

- 1 Ulrich Bauer, Xiaoyin Ge, and Yusu Wang. *Measuring distance between Reeb graphs*. Proc. of the 30th Symposium on Computational Geometry, 2014.
- 2 Ulrich Bauer, Elizabeth Munch, and Yusu Wang. *Strong Equivalence of the Interleaving and Functional Distortion Metrics for Reeb Graphs*. Proc. of the 31st Symposium on Computational Geometry, 2015.
- 3 Mathieu Carrière and Steve Oudot. *Local Equivalence and Intrinsic Metrics between Reeb Graphs*. Proc. of the 33rd Symposium on Computational Geometry, 2017.
- 4 Tamal K. Dey, Dayu Shi, and Yusu Wang. *Comparing Graphs via Persistence Distortion*. Proc. of the 31st Symposium on Computational Geometry, 2015.

## 4.2 Discussion Group on “Multivariate Topology”

*Michael Kerber (TU Graz, AT), Harish Doraiswamy (New York University, US), Christoph Garth (TU Kaiserslautern, DE), Claudia Landi (University of Modena, IT), Michael Lesnick (Princeton University, US), Jan Reininghaus (Siemens Industry Software GmbH – Wien, AT), and Julien Tierny (CNRS-UPMC – Paris, FR)*

**License**  Creative Commons BY 3.0 Unported license

© Michael Kerber, Harish Doraiswamy, Christoph Garth, Claudia Landi, Michael Lesnick, Jan Reininghaus, and Julien Tierny

The discussion group started by brainstorming a list of possible questions to address. The list of topics includes:

- Computing the matching distance exactly or finding better ways of approximating it
- Simplification of Jacobi sets/Reeb spaces (in a PL setting)
- Good ways to depict Reeb spaces
- Stable kernels for multi-dimensional persistence (with connection to machine learning)
- Approximate Reeb space in non-manifold PL-setting
- Definition of a critical point in a multi-filtration that works in a PL-domain/combinatorially
- More (convincing) application scenarios (e.g., ensemble classifications in weather simulations)

Since this list was impossible to cover in the short time, the discussion focussed on specific aspects related to these questions:

- For the visualization community, the denoising aspect of persistent homology is important, but there are two issues: measuring importance and localization. Moreover, the point was raised that localization is not stable, as the pairing of critical points obtained from persistent homology can change a lot after small perturbations (even though the persistence diagram is stable). Do such effects also occur in the case of Reeb spaces?
- For the case of time-varying data, the common approach is to simplify each time step separately, but the major problem is to link these time frames. A mapper-style approach for time-varying data has been used in visualization in the past.
- A problem of current approaches is that they depend on or favor the chosen coordinate directions. A point was raised with respect to whether randomized constructions could help to eliminate such effects.
- For 1-dimensional data sets, a simplicial simplification based on persistence has been proposed by Bauer et al. To extend this to higher dimensions, the question is whether one can pick compatible slices. This relates to the talk by Claudia Landi, and the discussion reviewed some aspects of her talk.



- While in Claudia Landi’s talk, the setup was with two functions and two dimensions, a question came up as what happens if there are more functions than dimensions. Based on Osamu’s talk, it seems that everything becomes more difficult in that case. From the perspective of a researcher in visualization, it is a strange phenomenon that adding one additional function removes the “niceness” from the problem.
- A question was discussed as what complexes can arise as 2-dimensional Reeb spaces (with a generic manifold input). There seemed to be an agreement that it is not a manifold in general. It was discussed whether it is a pure complex (without any definite answer).

As a wrap-up, it was commonly agreed that being able to compare multi-dimensional modules is essential to visualization applications and to many of the discussed questions. A kernel for multi-dimensional persistence would add to the range of applications. Moreover, a first implementation for the (approximate) matching distance is currently in preparation.

### 4.3 Discussion Group on “Scalable Computation”

*Michael Kerber (TU Graz, AT), Peer-Timo Bremer (LLNL – Livermore, US), Leila De Floriani (University of Maryland – College Park, US), Michael Lesnick (Princeton University, US), Vijay Natarajan (Indian Institute of Science – Bangalore, IN), and Primoz Skraba (Jozef Stefan Institute – Ljubljana, SI)*

**License** © Creative Commons BY 3.0 Unported license

© Michael Kerber, Peer-Timo Bremer, Leila De Floriani, Michael Lesnick, Vijay Natarajan, and Primoz Skraba

The group started with a discussion on parallelizing the computation of persistent homology. First, it was discussed whether existing approaches for merge trees extend to higher-dimensional homology in a simple way. The conclusion was that this is not the case, mostly because it appears hard to obtain global information (like homology) based on local computations.

For the existing approaches to parallel computation, it is usually the case that one node does a substantial amount of work in the end. The question is whether this can be avoided. Again, the fact that persistence computation is, in fact, a linear algebra problem prevents an easy application of local computation.

Moreover, the possibility of a GPU implementation for persistent homology was briefly discussed.

It was established that “scalable” has different meanings in different communities: for the visualization community, it mostly means parallelizable algorithms (which were also the main topic of discussion in this group), but in the field of algorithm design, it also means asymptotically faster algorithms. It seems that for point cloud data, the latter aspect should currently be in focus because current approaches are too slow even if they would be parallelized. For the case of cubical data, however, the algorithms appear optimal from an asymptotic point of view, and hence parallelization becomes important for performance.

Another point of discussion was the computation of high-dimensional persistent homology. Indeed, there are application domains where high-dimensional simplices arise quite naturally, for instance in robotics. The question of whether the homology information in dimension 4 (or higher) is useful without interpretability was a controversial point of discussion. Some participants referred to machine learning, where sometimes the learned features can also not be interpreted in many cases. It was questioned by others whether this is the right way to go.

The group agreed that more work should go into the practical efficiency of low-dimensional Rips complexes with a low distance threshold. This includes topics such as the distributed creation of the Rips complex and distributed nearest-neighbor queries.

Finally, there was a request about a state of the art report on the sizes of data sets that can be handled by current approaches.

## Participants

- Ulrich Bauer  
TU München, DE
- Magnus Botnan  
TU München, DE
- Peer-Timo Bremer  
LLNL – Livermore, US
- Roxana Bujack  
Los Alamos National  
Laboratory, US
- Hamish Carr  
University of Leeds, GB
- Leila De Floriani  
University of Maryland –  
College Park, US
- Pawel Dlotko  
Swansea University, GB
- Harish Doraiswamy  
New York University, US
- Brittany Terese Fasy  
Montana State University –  
Bozeman, US
- Christoph Garth  
TU Kaiserslautern, DE
- Ellen Gasparovic  
Union College – Schenectady, US
- Hans Hagen  
TU Kaiserslautern, DE
- Ingrid Hotz  
Linköping University, SE
- Michael Kerber  
TU Graz, AT
- Claudia Landi  
University of Modena, IT
- Heike Leitte  
TU Kaiserslautern, DE
- Michael Lesnick  
Princeton University, US
- Elizabeth Munch  
Michigan State University, US
- Vijay Natarajan  
Indian Institute of Science –  
Bangalore, IN
- Steve Y. Oudot  
INRIA Saclay –  
Île-de-France, FR
- Valerio Pascucci  
University of Utah –  
Salt Lake City, US
- Jan Reininghaus  
Siemens Industry Software  
GmbH – Wien, AT
- Osamu Saeki  
Kyushu University –  
Fukuoka, JP
- Gerik Scheuermann  
Universität Leipzig, DE
- Martina Scolamiero  
EPFL – Lausanne, CH
- Donald Sheehy  
University of Connecticut –  
Storrs, US
- Primož Skraba  
Jozef Stefan Institute –  
Ljubljana, SI
- Julien Tierny  
CNRS-UPMC – Paris, FR
- Katharine Turner  
Australian National University –  
Canberra, AU
- Bei Wang  
University of Utah –  
Salt Lake City, US
- Yusu Wang  
Ohio State University –  
Columbus, US



# User-Generated Content in Social Media

Edited by

Tat-Seng Chua<sup>1</sup>, Norbert Fuhr<sup>2</sup>, Gregory Grefenstette<sup>3</sup>,  
Kalervo Järvelin<sup>4</sup>, and Jaakko Peltonen<sup>5</sup>

1 National University of Singapore, SG, chuats@comp.nus.edu.sg

2 Universität Duisburg-Essen, DE, norbert.fuhr@uni-due.de

3 IHMC – Paris, FR, ggrefenstette@ihmc.us

4 University of Tampere, FI, kalervo.jarvelin@uta.fi

5 Aalto University & University of Tampere, FI, jaakko.peltonen@uta.fi

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17301 “User-Generated Content in Social Media”. Social media have a profound impact on individuals, businesses, and society. As users post vast amounts of text and multimedia content every minute, the analysis of this user generated content (UGC) can offer insights to individual and societal concerns and could be beneficial to a wide range of applications. In this seminar, we brought together researchers from different subfields of computer science, such as information retrieval, multimedia, natural language processing, machine learning and social media analytics. We discussed the specific properties of UGC, the general research tasks currently operating on this type of content, identifying their limitations, and imagining new types of applications. We formed two working groups, WG1 “Fake News and Credibility”, WG2 “Summarizing and Story Telling from UGC”. WG1 invented an “Information Nutrition Label” that characterizes a document by different features such as e.g. emotion, opinion, controversy, and topicality; For computing these feature values, available methods and open research issues were identified. WG2 developed a framework for summarizing heterogeneous, multilingual and multimodal data, discussed key challenges and applications of this framework.

**Seminar** July 23–28, 2017 – <http://www.dagstuhl.de/17301>

**1998 ACM Subject Classification** H Information Systems, H.5 Information Interfaces and Presentation, H.5.1 Multimedia Information Systems, H.3 Information Storage and Retrieval, H.1 Models and principles, I Computing methodologies, I.2 Artificial Intelligence, I.2.6 Learning, I.2.7 Natural language processing, J Computer Applications, J.4 Social and behavioural sciences, K Computing Milieux, K.4 Computers and Society, K.4.1 Public policy issues

**Keywords and phrases** social media, user-generated content, social multimedia, summarisation, storytelling, fake-news, credibility, AI

**Digital Object Identifier** 10.4230/DagRep.7.7.110

**Edited in cooperation with** Nicolás Díaz Ferreyra



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

User-Generated Content in Social Media, *Dagstuhl Reports*, Vol. 7, Issue 7, pp. 110–154

Editors: Tat-Seng Chua, Norbert Fuhr, Gregory Grefenstette, Kalervo Järvelin, and Jaakko Peltonen



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Executive Summary

*Norbert Fuhr*

*Tat-Seng Chua*

*Gregory Grefenstette*

*Kalervo Järvelin*

*Jaakko Peltonen*

**License** © Creative Commons BY 3.0 Unported license

© Norbert Fuhr, Tat-Seng Chua, Gregory Grefenstette, Kalervo Järvelin, and Jaakko Peltonen

Social media play a central role in many people's lives, and they also have a profound impact on businesses and society. Users post vast amounts of content (text, photos, audio, video) every minute. This user generated content (UGC) has become increasingly multimedia in nature. It documents users' lives, revealing in real time their interests and concerns and activities in society. The analysis of UGC can offer insights to individual and societal concerns and could be beneficial to a wide range of applications, for example, tracking mobility in cities, identifying citizen's issues, opinion mining, and much more. In contrast to classical media, social media thrive by allowing anyone to publish content with few constraints and no oversight. Social media posts thus show great variation in length, content, quality, language, speech and other aspects. This heterogeneity poses new challenges for standard content access and analysis methods. On the other hand, UGC is often related to other public information (e.g. product reviews or discussion of news articles), and there often is rich contextual information linking, which allows for new types of analyses.

In this seminar, we aimed at discussing the specific properties of UGC, the general research tasks currently operating on this type of content, identifying their limitations and lacunae, and imagining new types of applications made possible by the availability of vast amounts of UGC. This type of content has specific properties such as presentation quality and style, bias and subjectivity of content, credibility of sources, contradictory statements, and heterogeneity of language and media. Current applications exploiting UGC include sentiment analysis, noise removal, indexing and retrieving UGC, recommendation and selection methods, summarization methods, credibility and reliability estimation, topic detection and tracking, topic development analysis and prediction, community detection, modeling of content and user interest trends, collaborative content creation, cross media and cross lingual analysis, multi-source and multi-task analysis, social media sites, live and real-time analysis of streaming data, and machine learning for big data analytics of UGC. These applications and methods involve contributions from several data analysis and machine learning research directions.

This seminar brought together researchers from different subfields of computer science, such as information retrieval, multimedia, natural language processing, machine learning and social media analytics. After participants gave presentations of their current research orientations concerning UGC, we decided to split into two Working Groups: (WG-1) Fake News and Credibility, and (WG-2) Summarizing and Storytelling from UGC.

### WG-1: Fake News and Credibility

WG-1 began discussing the concept of Fake News, and we arrived at the conclusion that it was a topic with much nuance, and that a hard and fast definition of what was fake and what was real news would be hard to define. We then concentrated on deciding what

elements of Fake (or Real) News could be calculated or quantified by computer. This led us to construct a list of text quality measures that have or are being studied in the Natural Language Processing community: Factuality, Reading Level, Virality, Emotion, Opinion, Controversy, Authority, Technicality, and Topicality. During this discussion, WG-1 invented and mocked up what we called an Information Nutrition Label, modeled after nutritional labels found on most food products nowadays. We feel that it would be possible to produce some indication of the “objective” value of a text using the above nine measures. The user could use these measures to judge for themselves whether a given text was “fake” or “real”. For example, a text highly charged in Emotion, Opinion, Controversy, and Topicality might be Fake News for a given reader. Just like with a food nutritional label, a reader might use the Information Nutritional Label to judge whether a given news story was “healthy” or not.

WG-1 split into further subgroups to explore whether current status of research in the nine areas: Factuality, Reading Level, etc. For each topic, the subgroups sketched out the NLP task involved, found current packages, testbeds and datasets for the task, and provided recent bibliography for the topic. Re-uniting in one larger group, each subgroup reported on their findings, and we discussed next steps, envisaging the following options: a patent covering the idea, creating a startup that would implement all nine measures and produce a time-sensitive Information Nutritional Label for any text submitted to it, a hackathon that would ask programmers to create packages for any or all of the measurements, a further workshop around the Information Nutrition label, integration of the INL into teaching of Journalists, producing a joint article describing the idea. We opted for the final idea, and we produced a submission (also attached to this report) for the Winter issue of the SIGIR (Special Interest Group on Information Retrieval) Forum<sup>1</sup>.

## WG-2: Summarizing and Storytelling from UGC

WG-2 set out to re-examine the topic of summarization. Although this is an old topic, but in the era of user-generated content with accelerated rates of information creation and dissemination, there is a strong need to re-examine this topic from the new perspectives of timeliness, huge volume, multiple sources and multimodality. The temporal nature of this problem also brings it to the realm of storytelling, which is done separately from that of summarization. We thus need to move away from the traditional single source document-based summarization, by integrating summarization and storytelling, and refocusing the problem space to meet the new challenges.

We first split the group into two sub-groups, to discuss separately: (a) the motivations and scopes, and (b) the framework of summarization. The first sub-group discussed the sources of information for summarization including, the user-generated content, various authoritative information sources such as the news and Wikipedia, the sensor data, open data and proprietary data. The data is multilingual and multimodal, and often in real time. The group then discussed storytelling as a form of dynamic summarization. The second group examined the framework for summarization. It identified the key pipeline processes comprising of: data ingestion, extraction, reification, knowledge representation, followed by story generation. In particular, the group discussed the roles of time and location in data, knowledge and story representation.

---

<sup>1</sup> <http://sigir.org/forum/>

Finally, the group identified key challenges and applications of the summarization framework. The key challenges include multi-source data fusion, multilinguality and multimodality, the handling of time/ temporality/ history, data quality assessment and explainability, knowledge update and renewal, as well as focused story/ summary generation. The applications that can be used to focus the research includes event detection, business intelligence, entertainments and wellness. The discussions have been summarized into a paper entitled “Rethinking Summarization and Storytelling for Modern Social Multimedia”. The paper is attached along with this report. It has been submitted to a conference for publication.

## 2 Contents

### Executive Summary

<i>Norbert Fuhr, Tat-Seng Chua, Gregory Grefenstette, Kalervo Järvelin, and Jaakko Peltonen . . . . .</i>	111
---	-----

### Overview of Talks

NLP Approaches for Fact Checking and Fake News Detection <i>Andreas Hanselowski . . . . .</i>	116
User-Generated Content and Privacy Risks: From Regrets to Preventative Technologies <i>Nicolas Diaz-Ferreira . . . . .</i>	116
Social Media Retrieval with Contradictions and Credibility <i>Norbert Fuhr . . . . .</i>	117
Multilingual Aspects of User-Generated Content <i>Tatjana Gornostaja . . . . .</i>	117
Spam in User-Generated Content <i>Gregory Grefenstette . . . . .</i>	118
Cross-Modal Recommendation: From Shallow Learning to Deep Learning <i>Xiangnan He . . . . .</i>	118
Altmetrics and Tweeting Behavior of Scientists <i>Isabella Peters . . . . .</i>	119
People Analytics with User-Generated Content <i>Rianne Kaptein . . . . .</i>	119
Detecting Malicious Activities in Community Question Answering Platforms <i>Yiqun Liu . . . . .</i>	120
Social Media and e-Commerce <i>Marie-Francine Moens . . . . .</i>	120
Learning from Social Media and Contextualisation <i>Josiane Mothe . . . . .</i>	121
Machine Learning for Analysis of Hierarchical Conversation Forums <i>Jaakko Peltonen . . . . .</i>	121
Multimedia in Data Science: Bringing Multimedia Analytics to the Masses <i>Stevan Rudinac . . . . .</i>	122
User- and Culture-Aware Models for Music Recommender Systems <i>Markus Schedl . . . . .</i>	122
Changing our Mind: Correlations of Media in Online Collaboration Systems <i>David Ayman Shamma . . . . .</i>	123
Social Media: A Narcissic Form of Lifelogging? <i>Alan Smeaton . . . . .</i>	123
User-Generated Content: An Adversarial Perspective <i>Benno Stein . . . . .</i>	124



An Anatomy of Online Video Popularity <i>Lexing Xie</i> . . . . .	124
--	-----

### Working groups


An Information Nutritional Label for Online Documents <i>Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Järvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein</i> . . . . .	125
Rethinking Summarization and Storytelling for Modern Social Multimedia <i>Stevan Rudinac, Tat-Seng Chua, Nicolas Diaz-Ferreyra, Gerald Friedland, Tatjana Gornostaja, Benoît Huet, Rianne Kaptein, Krister Lindén, Marie-Francine Moens, Jaakko Peltonen, Miriam Redi, Markus Schedl, David Ayman Shamma, Alan Smeaton, and Lexing Xie</i> . . . . .	141

Participants . . . . .	154
------------------------	-----

### 3 Overview of Talks

#### 3.1 NLP Approaches for Fact Checking and Fake News Detection

*Andreas Hanselowski (TU Darmstadt, DE)*


**License**  Creative Commons BY 3.0 Unported license  
© Andreas Hanselowski

In the past couple of years, there has been a significant increase of untrustworthy information on the web, such as fake news articles. In order to validate the false information circulating on the web, fact-checking became an essential tool, and today, there are numerous websites such as fullfact.org, politifact.com, and snopes.com devoted to the problem. Although manual fact-checking is an important instrument in the fight against false information, it cannot solve the problem entirely. The large number of fake news articles being generated at a high rate cannot be easily detected or debunked by human fact checkers. Many of the upcoming issues of manual claim validation can be addressed by automated fact-checking, as it would allow a large number of articles to be validated in real time as they appear on the web. The problem of tackling false information on the web can be divided into two different problem settings, that is, into fake news detection and automated fact-checking.

In the seminar, machine learning approaches for both problem-settings have been presented. For fake news detection, the problem of stance classification was discussed, as it was posed in the Fake News Challenge. For the solution of the problem a multilayer perception, which is based on linguistic features, was introduced. For the automated fact-checking, a comprehensive framework was presented, in which the problem is divided into a number of steps. In the first step, web documents are retrieved, which contain information required for the resolution of a given claim. Next, evidence in the web documents is identified, which explicitly support or contradict the claim. In the last step, the actual claim validation is performed, whereby the identified evidence and web documents serve as a basis.

#### 3.2 User-Generated Content and Privacy Risks: From Regrets to Preventative Technologies

*Nicolas Diaz-Ferreyra (Universität Duisburg-Essen, DE)*


**License**  Creative Commons BY 3.0 Unported license  
© Nicolas Diaz-Ferreyra  
**URL** [https://doi.org/10.1007/978-3-319-66808-6\\_7](https://doi.org/10.1007/978-3-319-66808-6_7)

User-generated content very often enclose private and sensitive information. When such information reaches an unintended audience, it can derivate in unwanted incidents for the users like job loss, reputation damage, or sextortion along with a feeling of regret. Preventative technologies aim to help users to bypass the potential unwanted incidents of online-self disclosure by raising awareness on the content being shared. However, in order to engage with the users, such technologies should follow some basic design principles. First, preventative technologies should be adaptive on the users' privacy attitudes and intentions. Second, they should generate a visceral connection between users and their private data. Finally, such technologies should provide supportive guidance to the users informing about possible actions that can help them to protect their privacy. This talk aim to discuss the role of regrettable user-generated content in the development of adaptive, visceral and

supportive preventative technologies. Particularly, how privacy heuristics can be extracted from regrettable experiences and integrated later on into the design of awareness mechanisms.

### 3.3 Social Media Retrieval with Contradictions and Credibility

Norbert Fuhr (*Universität Duisburg-Essen, DE*)

License  Creative Commons BY 3.0 Unported license  
© Norbert Fuhr

User comments in Social Media – e.g. reviews of products or services – often contradict each other, and they also may vary in terms of credibility. In order to aggregate these comments for the purpose of retrieval, we propose to apply a number of logic-based concepts: 1) While today's retrieval methods mostly use an implicit closed world assumption, an open world assumption allows to distinguish between missing information and explicit negation. 2) For handling contradictions, a four-valued logic also contains the truth values 'inconsistent' and 'unknown'. 3) A possible worlds semantics models an agent's belief over possibly contradictory statements as a probability distribution over different worlds, where this distribution can be used for representing credibility of statements. While these logic-based formalisms are well developed, a major challenge in their application is the development of appropriate indexing methods for creating the representations needed for the application of the logic-based models.

### 3.4 Multilingual Aspects of User-Generated Content

Tatjana Gornostaja (*tilde – Riga, LV*)

License  Creative Commons BY 3.0 Unported license  
© Tatjana Gornostaja

My name is Tatjana Gornostaja and I present on behalf of the company Tilde I have been working for more than 10 years with my background in terminology (knowledge management) and translation (human and automated). Tilde is specialising in natural language data (text and speech) processing with the focus on small languages with scarce resources and rich grammar, developing innovative products and services of machine translation, terminology management, speech analysis, virtual assistants and operating in the three Baltic countries – Estonia, Latvia, Lithuania with the headquarters in Riga.

A huge amount of content is generated by users on the Internet (YouTube, Facebook, Instagram, Twitter, WordPress etc.) in different languages. According to statistics, more than 40% of Europeans speak only their native language and more than 60% do not speak English well enough to consume the content published on the Internet, which is predominantly in English. However, if you talk to a man in a language he understands – that goes to his head, if you speak to him in his own language – that goes to his heart (Nelson Mandela). This is our motto for the products and services we provide to our users (public administration, business, academia – locally and globally) to help them to communicate successfully.

The latest advancements have been integrated into popular platforms recently:

- best AI-powered machine translation for Latvian on Twitter
- virtual assistants (multiplication table for children, currency exchange and travel guides for adults) on Facebook Messenger and Skype

- speech and text processing systems for government services, including the support to the European Presidency in Latvia (past), Estonia (ongoing), Bulgaria and Austria (upcoming)
- speech processing mobile applications for Latvian for people with vision impairment and dyslexia.

Being proud and honoured for the invitation and participation in the Schloss Dagstuhl seminar, I am grateful to its organisers for this excellent event with outstanding presentations and inspiring discussions. With our competences, expertise and experience in more than 25 international research, development and innovation projects, with a wide range of more than 60 partners worldwide we are open for collaboration and new ideas to connect people speaking different languages across borders.

### 3.5 Spam in User-Generated Content

*Gregory Grefenstette (IHMC – Paris, FR)*

**License**  Creative Commons BY 3.0 Unported license  
© Gregory Grefenstette

User Generated Content (UGC) provides rich data for understanding language use, user opinion, and many other uses. But researchers should be aware of the wide variety of spam that appears in this data. False blogs can generate well structured but random text to hide pointers to money-making sites. Comments may contain generic messages also to create links to spam pages. It is easy to create false users, false reviews, false likes for any social network. We examine some of these problems and show some ways to detect spam and false users, so that researchers know that this noise exists in their UGC.

### 3.6 Cross-Modal Recommendation: From Shallow Learning to Deep Learning

*Xiangnan He (National University of Singapore, SG)*

**License**  Creative Commons BY 3.0 Unported license  
© Xiangnan He

Recommender systems play a central role in the current user-centered Web. Many customer-oriented online services rely on recommender systems or advertising systems to earn money. The de-facto technique to build a personalized recommender system is collaborative filtering, which predicts a user's preference based on the historical user-item interactions. Besides the interaction data, there are also rich side information available, such as user demographics (e.g., gender, age), item attributes (e.g., textual description and visual images), and various contexts. The key research here is how to effectively leverage all relevant information available to build a better recommender system.

In this talk, I first present the existing and widely used techniques for building a generic recommender that model various information, including Logistic Regression (plus GBDTs), Factorization Machines (FMs), and Tensor Decomposition. Then, I briefly introduce recent deep learning solutions for recommendation, such as the Google's Wide&Deep and Microsoft's Deep Crossing. Lastly, I introduce our recently proposed neural recommendation solutions, Neural FMs and Attentional FMs.

### 3.7 Altmetrics and Tweeting Behavior of Scientists

*Isabella Peters (ZBW – Dt. Zentralbib. Wirtschaftswissenschaften, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Isabella Peters

The talk introduced altmetrics, the field of research evaluation by means of social media signals for scholarly products. Examples for ongoing research were given – for example the relationships between citations, the journal impact factor and altmetrics were discussed. It was shown that altmetrics are influenced by characteristics of the scholarly product, e.g., its popularity, the language in which it was published, and whether it was authored by known authors. Moreover, altmetrics are strongly dependent on the context the social media was used. The talk concluded that responsible use of altmetrics calls for better understanding of how, why, when and by whom social media signals are produced in order to draw correct conclusions from pure numbers.

Further reading at DBLP: [conf/isiwi/NurediniP15 journals/it/HausteinLTAP14](https://dblp.org/conf/isiwi/NurediniP15/journals/it/HausteinLTAP14)

### 3.8 People Analytics with User-Generated Content

*Rianne Kaptein (Crunchr – Amsterdam, NL)*


**License** © Creative Commons BY 3.0 Unported license  
© Rianne Kaptein

The relatively new field of people analytics has become a hot topic in organizations of all sizes. People Analytics, also known as HR analytics, refers to the method of analytics that can help managers and executives make decisions about their employees or workforce. Organizations are reaching out to learn more about predictive analytics in order to improve organizational effectiveness. Significant resources are devoted to identify talented employees and to develop them into future leaders. However, current people management processes are subjective and mostly retrospective. For example, emerged talent is identified in hindsight and valuable years of development are missed out on. Talented employees perceive this as being undervalued and might eventually leave the company. Other people analytics tasks include: Workforce planning, Succession planning, Recruitment optimization, Team composition, Predicting employee turnover and Employee engagement analysis. The main sources for user generated content on HR data are LinkedIn and Glassdoor. These sites include information on: resumes, work experience, skills, connections, company reviews, etc. Challenges when using this content are privacy, identity resolution and data sparseness. To overcome these challenges we do not identify individuals, but only analyze on the company level or anonymized data.

User generated content on Social Media often consist of a large number of short texts (e.g. survey question responses, tweets, Facebook Posts, forum posts). HR staff has problems with information overload – often there are too many messages to read. So we aim to give an overview or summary of the relevant/interesting responses. These summaries can also be coupled with sentiment analysis.

### 3.9 Detecting Malicious Activities in Community Question Answering Platforms

*Yiqun Liu (Tsinghua University – Beijing, CN)*

License  Creative Commons BY 3.0 Unported license  
© Yiqun Liu

With Community Question Answering (CQA) evolving into a quite popular method for information seeking and providing, it also becomes a target for spammers to disseminate promotion campaigns. Although there are a number of quality estimation efforts on the CQA platform, most of these works focus on identifying and reducing low-quality answers, which are mostly generated by impatient or inexperienced answerers. However, a large number of promotion answers appear to provide high-quality information to cheat CQA users in future interactions. Therefore, most existing quality estimation works in CQA may fail to detect these specially designed answers or question-answer pairs. In contrast to these works, we proposed two methods for detecting spamming activities on CQA platforms. For individual spamming activity detection, we focus on the promotion channels of spammers, which include (shortened) URLs, telephone numbers and social media accounts. Spammers rely on these channels to connect to users to achieve promotion goals so they are irreplaceable for spamming activities. We propose a propagation algorithm to diffuse promotion intents on an “answerer-channel” bipartite graph and detect possible spamming activities. A supervised learning framework is also proposed to identify whether a QA pair is spam based on propagated promotion intents. Experimental results based on more than 6 million entries from a popular Chinese CQA portal show that our approach outperforms a number of existing quality estimation methods for detecting promotion campaigns on both the answer level and QA pair level. For collusive spamming activity detection, we propose a unified framework to tackle the challenge. First, we interpret the questions and answers in CQA as two independent networks. Second, we detect collusive question groups and answer groups from these two networks respectively by measuring the similarity of the contents posted within a short duration. Third, using attributes (individual-level and group-level) and correlations (user-based and content-based), we proposed a combined factor graph model to detect deceptive Q&As simultaneously by combining two independent factor graphs. With a large-scale practical data set, we find that the proposed framework can detect deceptive contents at early stage, and outperforms a number of competitive baselines.

### 3.10 Social Media and e-Commerce

*Marie-Francine Moens (KU Leuven, BE)*

License  Creative Commons BY 3.0 Unported license  
© Marie-Francine Moens

The lecture has focused on representation learning of social media content and was illustrated with two use cases: Bridging the language of consumers and product vendors and bridging language and vision for cross-modal fashion search.

In the first use case, we have focused on linking content (textual descriptions of pins in Pinterest to webshops). We have explained the problem of linking information between different usages of the same language, e.g., colloquial and formal “idioms” or the language of consumers versus the language of sellers. For bridging these languages, we have trained a multi-idiomatic latent Dirichlet allocation model (MiLDA) on product descriptions aligned with their reviews.

In the second use case, we have proposed two architectures to link visual with textual content. The first architecture uses a bimodal latent Dirichlet allocation topic model to bridge between these two modalities. As a second architecture, we have developed a neural network which learns intermodal representations for fashion attributes. Both resulting models learn from organic e-commerce data, which is characterised by clean image material, but noisy and incomplete product descriptions. We have demonstrated two tasks: 1) Given a query image (without any accompanying text), we retrieve textual descriptions that correspond to the visual attributes in the visual query; and 2) given a textual query that expresses an interest in specific visual characteristics, we retrieve relevant images (without leveraging textual metadata) that exhibit the required visual attributes. The first task is especially useful to manage product image collections by online stores who might want to automatically organise and mine predominantly visual items according to their attributes without human input. The second task allows users to find product items with specific visual characteristics, in the case where there is no text available describing the target image.

### 3.11 Learning from Social Media and Contextualisation

*Josiane Mothe (University of Toulouse, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Josiane Mothe

**Main reference** Idriss Abdou Malam, Mohamed Arziki, Mohammed Nezar Bellazrak, Farah Benamara, Assafa El Kaidi, Bouchra Es-Saghir, Zhaolong He, Mouad Housni, Véronique Moriceau, Josiane Mothe, Faneva Ramiandrisoa: “IRIT at e-Risk. CLEF”, Working Notes of CLEF 2017, CEUR-WS.org, 2017.

**URL** [http://ceur-ws.org/Vol-1866/paper\\_135.pdf](http://ceur-ws.org/Vol-1866/paper_135.pdf)

Social media can be a rich source of information either to extract some trends (models) or peculiarities (weak signals). We focused in this talk on early depression detection from social media posts using machine learning techniques and presented some results. We also proposed to use the same type of model to detect and extract locations from short posts when user localisation is not available. Finally, we mentioned our current work on tweet contextualisation that aims helping users to understand short texts.

### 3.12 Machine Learning for Analysis of Hierarchical Conversation Forums

*Jaakko Peltonen (Aalto University, FI)*

**License** © Creative Commons BY 3.0 Unported license  
© Jaakko Peltonen

In many domains, user generated textual data arise as hierarchically organised document sets. In particular, online discussion often occurs as conversation threads in online message forums and other social media platforms having a prominent hierarchical organisation, with multiple levels of sections and subsections. Modeling the online discussions is important for studies of discussion behavior, for tracking trends of consumer interests, and analysis of brands and advertising impact.

Machine learning methods can help to analyse latent topics within such content, which can then be used for predictive and exploratory tasks, such as analysis of trends, analysis of

user interest and sentiment across different types of content, recommendation of discussion content or targeting of advertising, and for intelligent interfaces to browse and participate in discussions. The hierarchical structure of online forums is designed to cover a subset of prototypical user interests, and could help build better models of content; however, most available machine learning methods cannot fully take the hierarchy into account in modelling.

In our recent work we have developed methods for taking this hierarchical structure into account in modelling latent topics of discussion, including probabilistic nonparametric topic models of the hierarchical content and interactive exploratory interfaces based on dimensionality reduction that reveal how the variety of discussion content is related to the hierarchy. We are further developing methods to model how individual users and populations of users visit content across the hierarchy.

### 3.13 Multimedia in Data Science: Bringing Multimedia Analytics to the Masses

*Stevan Rudinac (University of Amsterdam, NL)*

License  Creative Commons BY 3.0 Unported license  
© Stevan Rudinac

In this talk, using urban computing as a case study, we advocate that Multimedia community should embrace data science. Increased availability of open data in relation to various neighbourhood statistics such as demographics, transportation and services, made analysing and modelling processes in the city significantly easier. However, useful information about the problems a city is facing with may be also extracted from spontaneously captured social multimedia, participatory data and wearable technology. The examples from our work on interactive venue recommendation, discovering functional regions of the city, analysing liveability of the neighbourhoods and empowering local urban communities, demonstrate that multimedia analytics can be successfully deployed on such heterogeneous data for solving important societal problems. We further show that multimedia analytics and, in particular, interactive learning can be facilitated even on very large collections with 100 million user-generated images and associated annotations. Perhaps more importantly, it may be a possible solution for the imperfections in the automatic analysis techniques and facilitate easier technology adoption by keeping the user in control. Finally, we touch the topics of data reliability, privacy and ethics.

### 3.14 User- and Culture-Aware Models for Music Recommender Systems

*Markus Schedl (Universität Linz, AT)*

License  Creative Commons BY 3.0 Unported license  
© Markus Schedl

Nowadays, music aficionados generate millions of listening events every day and share them via services such as Last.fm or Twitter. In 2016, the LFM-1b dataset <http://www.cp.jku.at/datasets/LFM-1b> containing more than 1 billion listening events of about 120,000 Last.fm users has been released to the research community and interested public. Since then, we



performed various data analysis and machine learning tasks on these large amounts of user and listening data. The gained insights helped to develop new listener models and integration them into music recommender systems, in an effort to increase personalization of the recommendations. In this talk, the focus is on the following research directions, which we are currently pursuing: (i) analyzing music taste around the world and distilling country clusters, (ii) quantifying listener and country mainstreamness, (iii) music recommendation tailored to listener characteristics, and (iv) predicting country-specific genre preferences from cultural and socio-economic factors.

### 3.15 Changing our Mind: Correlations of Media in Online Collaboration Systems

*David Ayman Shamma (CWI – Amsterdam, NL)*

**License** © Creative Commons BY 3.0 Unported license  
© David Ayman Shamma

As humans, we create a lot of data and we change our minds. Sometimes we learn and grow; sometimes we were just wrong. In particular, we see these edits and changes in online user generated social systems. However, rarely are these changes accounted for when we index, recommend, and classify. In this talk, I illustrate, using historical Wikipedia associations, how community use and abuse changes the semantics and meaning of the images we use. Further, I assert we need to know why people make and change annotations as it changes how we build artificial intelligence systems for user-generated media.

### 3.16 Social Media: A Narcissic Form of Lifelogging?

*Alan Smeaton (Dublin City University, IE)*

**License** © Creative Commons BY 3.0 Unported license  
© Alan Smeaton


**Main reference** Cathal Gurrin, Alan F. Smeaton, Aiden R. Doherty: “LifeLogging: Personal Big Data”, Foundations and Trends in Information Retrieval, Vol. 8(1), pp. 1–125, 2014.

**URL** <http://dx.doi.org/10.1561/15000000033>

I present the state of work in lifelogging, first person digital ethnography, using off the shelf wearable sensors coupled with data from public sources. The talk focuses a lot on using wearable cameras and the various kinds of behaviour and activities can than be automatically extracted from such wearable camera images. I also present our lab’s work on new kinds of wearable sensors for glucose levels or sweat composition. The final argument of the presentation is that narcissic social media posts can be replaced by something extracted from human lifelogs.

### 3.17 User-Generated Content: An Adversarial Perspective

*Benno Stein (Bauhaus-Universität Weimar, DE)*

License  Creative Commons BY 3.0 Unported license  
© Benno Stein

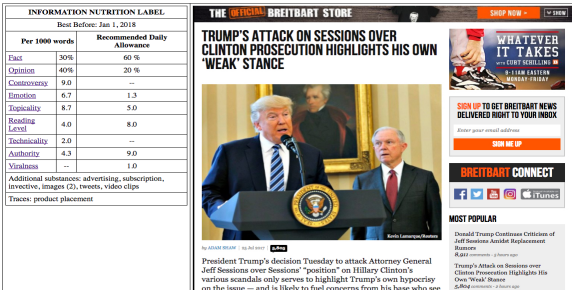
In this talk I will present research questions and results of selected adversarial analytics where our research group has been working on in the recent past: Hyperpartisan news (in Blogs), Clickbait (in Twitter) [80], Wikipedia Vandalism [81], Offensive Language in comments. The talk will point out the different natures of the adversarial incidents, which in turn give rise for different countermeasures: clickbait can be addressed with automatically formulating and submitting suited questions queries whose results are presented along the clickbait text. Similarly, “alternative” facts or fake news can be countered by consulting an argument search engine [82]. Aside from theoretical backgrounds the talk will provide demonstrations of recently developed technology.

### 3.18 An Anatomy of Online Video Popularity

*Lexing Xie (Australian National University – Canberra, AU)*

License  Creative Commons BY 3.0 Unported license  
© Lexing Xie

How did a video go viral? Or will it go viral, and when? These are some of the most intriguing yet difficult questions in social media analysis. I will cover a few recent results from my group on understanding and predicting popularity, especially for YouTube videos. I will start by describing a unique longitudinal measurement study on video popularity history, and introduce popularity phases, a novel way to describe the evolution of popularity over time. I will then discuss a physics-inspired stochastic model that connects exogenous stimuli and endogenous responses to explain and forecast popularity. This, in turn, leads to a set of novel metrics for forecasting expected popularity gain per share, the time it takes for such effects to unfold, and sensitivity to promotions.



■ **Figure 1** Mockup of the envisaged information nutrition label.

**4 Working groups**

**4.1 An Information Nutritional Label for Online Documents**

Norbert Fuhr (Universität Duisburg-Essen, DE), Anastasia Giachanou (University of Lugano, CH), Gregory Grefenstette (IHMC – Paris, FR), Iryna Gurevych (TU Darmstadt, DE), Andreas Hanselowski (TU Darmstadt, DE), Kalervo Järvelin (University of Tampere, FI), Rosie Jones (Microsoft New England R&D Center – Cambridge, US), Yiqun Liu (Tsinghua University – Beijing, CN), Josiane Mothe (University of Toulouse, FR), Wolfgang Nejdl (Leibniz Universität Hannover, DE), Isabella Peters (ZBW Dt. Zentralbib. Wirtschaftswissenschaften, DE), and Benno Stein (Bauhaus-Universität Weimar, DE)

**License** © Creative Commons BY 3.0 Unported license  
© Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Järvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein

With the proliferation of online information sources, it has become more and more difficult to judge the trustworthiness of news found on the Web. The beauty of the web is its openness, but this openness has lead to a proliferation of false and unreliable information, whose presentation makes it difficult to detect. It may be impossible to detect what is “real news” and what is “fake news” since this discussion ultimately leads to a deep philosophical discussion of what is true and what is false. However, recent advances in natural language processing allow us to analyze information objectively according to certain criteria (for example, the number of spelling errors). Here we propose creating an “information nutrition label” that can be automatically generated for any online text. Among others, the label provides information on the following computable criteria: factuality, virality, opinion, controversy, authority, technicality, and topicality.

**4.1.1 Introduction**

The 2016 American presidential elections were a source of growing public awareness of what has been termed “fake news”. In a nutshell, the term is used to describe the observation that “in social media, a certain kind of ‘news’ spread much more successfully than others, and, that these ‘news’ stories are typically extremely one-sided (hyperpartisan), inflammatory, emotional, and often riddled with untruths” [71].

Claims in news can take various forms. In the form of a verifiable assertion (“The density of ice is larger than the density of water.”) we have a fact checking situation, which can be clarified given access to online dictionaries or encyclopedias. In the form of a non-verifiable or not easily verifiable assertion (“Hillary Clinton is running a child sex ring out of a D.C.-area

pizza restaurant.”, “Marijuana is safer than alcohol or tobacco.”) one has to take a stance, i.e., the reader has to decide whether she believes the claim or not. Such a decision can neither universally nor uniquely be answered by means of a knowledge base but is to be clarified on an individual basis and may undergo change over time.

To help the online information consumer, we propose an Information Nutrition Label, resembling nutrition fact labels on food packages. Such a label describes, along a range of agreed-upon dimensions, the contents of the product (an information object, in our case) in order to help the consumer (reader) in deciding about the consumption of the object. The observations above however imply technical, but in particular self-imposed ethical limitations of our envisaged concept:

*(manifest) It is not our intention to say what is true or what is fault, right or wrong, and in particular not what is good or bad. That is, an Information Nutrition Label is not a substitute for a moral compass.*

Thus, as technical consequence, we neither propose a system that would state what is true or what is false, right or wrong, and in particular not what is good or bad. Ultimately, it is up to the consumer to consult the information nutrition label and to decide whether to consume the information or not. Aiming at aiding the consumer’s decision making process, we see various technical uses as well as societal impacts of our Information Nutrition Label:

- personalized relevance ranking for search engine results
- information filtering according to personal preferences
- machine-based fake news detection
- learning and teaching of information assessment
- raising awareness and responsibility about deciding what to read.

#### 4.1.2 An Information Nutrition Label

Of course, the assessment of information is not a new discipline—recall the large body of research related to the concept of “information quality”, for which Levis et al. provide a useful overview [66]. While there is no unique definition for the concept, information quality is usually interpreted in terms of utility, namely as the “fitness for use in a practical application” [78]. Note that our paper will neither reinterpret nor extend this quality concept; instead, we are aiming at a practical means to ease information consumption and meta reasoning when given an online document by breaking down a quality judgment into smaller, measurable components.

We consider the Wikipedia quality endeavour as the most related precursor to our proposal. Aside from its rather informal quality guidelines, Wikipedia has formalized its quality ideal with the so-called featured article criteria<sup>2</sup>, and, even more important, distinguishes more than 400 quality flaws to spot article deficits [53]. In particular, the machine-based analysis of Wikipedia articles to detect flaws in order to assess article quality [54] corresponds closely to our idea of computing the separate dimensions of an information nutrition label. However, because of our use case, the nutrition label dimensions as well as their computation differs from the Wikipedia setting.

The following subsections describe measurable qualities that may be included in such an information nutrition label and that we consider valuable in order to assess the

---

<sup>2</sup> Wikipedia, “Featured articles,” last modified February 19, 2017, [http://en.wikipedia.org/wiki/Wikipedia:Featured\\_articles](http://en.wikipedia.org/wiki/Wikipedia:Featured_articles)

nutrient content of an information object. Each of these categories have been the subject of experimentation in the natural language processing, information retrieval, or web sciences communities:

- factuality
- readability
- virality
- emotion
- opinion
- controversy
- authority / credibility / trust
- technicality
- topicality

In the next subsections we will define these aspects and describe the relationship they have with an information nutrition label. Moreover, tasks and methods explain what practical steps may need to be taken into account to automatize the measurement of these qualities for insertion into the nutrition label. We consider the task descriptions as broad avenues from which further research may take off.

### 4.1.3 Factuality

#### 4.1.3.1 Task for Factuality Assessment

The task of determining the level of commitment towards a predicate in a sentence according to a specific source, like the author, is typically addressed as factuality prediction ([73]). Lexical cues, such as modals *will*, *shall*, *can* indicate the confidence of the source whether a proposition is factual. However, in contrast to a binary decision, the underlying linguistic system forms a continuous spectrum ranging from factual to counterfactual ([73]). Thus, for the assessment of the factuality for the whole document, one needs to compute the average factuality of all the propositions contained in the text.

Since we are not planning to judge the truthfulness of the statements in a given text, as it is attempted in the domain of automated fact checking, we are only interested in determining whether a statement is factual from the perspective of the author. The issue of whether the statements in the documents are controversial and may therefore not be reliable, is discussed in subsection 4.1.8 about Controversy.

#### 4.1.3.2 Methods for Factuality Assessment

For factuality prediction rule-based approaches as well as methods based on machine learning have been developed.

The De Facto factuality profiler [74] and the TruthTeller algorithm [68] are rule-based approaches, which assign discrete scores of factuality to propositions. In the process, dependency parse trees are analyzed top-down and the factuality score is altered whenever factuality affecting predicates or modality and negation cues are encountered.

A machine learning based approach has been applied to factuality prediction in [65]. The authors used a support vector machine regression model to predict continuous factuality values from shallow lexical and syntactic features such as lemmas, part-of-speech tags, and dependency paths.

The rule-based approach has been combined with the machine learning based method in [76]. Thereby, the outputs from TruthTeller were used as linguistically-informed features for a support vector machine regression model in order to predict the final factuality value.

#### 4.1.3.3 Data sets for Factuality Assessment

There are a number of annotation frameworks, which have been suggested to capture the factuality of statements. On the basis of the suggested annotation schemes, a number of data sets have been constructed.

Fact-Bank [73] is a corpus which was annotated discretely by experts according to different classes of factuality: Factual, Probable, Possible, Unknown. In this corpus, factuality has been assessed with respect to the perspective of the author or discourse-internal sources.

The MEANTIME corpus was introduced in [69] and was also annotated discretely by expert annotators. The propositions have been classified as Fact / Counterfact, Possibility (uncertain), Possibility (future) with respect to the author's perspective.

The UW corpus [65] was annotated on the basis of a continuous scale ranging from -3 to 3. The annotation was performed by crowd workers who judged the factuality score from the author's perspective.

In [76], the annotation schemes of the three different corpora have been merged in order to combine the three data sets into one single large corpus. For this purpose, the discrete scales used for the Fact-Bank and MEANTIME corpora have been mapped to the continuous scale of the UW corpus.

#### 4.1.3.4 Further reading for Factuality Assessment

1. Nissim Malvina, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. "Cross-linguistic annotation of modality: a data-driven hierarchical model." In Proceedings of the 9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, pp. 7-14. 2013.
2. O'Gorman Tim, Kristin Wright-Bettner, and Martha Palmer. "Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation." In Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016). 2016.
3. Ghia Elisa, Lennart Kloppenburg, Malvina Nissim, Paola Pietrandrea, and Valerio Cervoni. "A construction-centered approach to the annotation of modality." In Proceedings of the 12th ISO Workshop on Interoperable Semantic Annotation, pp. 67-74. 2016.
4. Guggilla Chinnappa, Tristan Miller, and Iryna Gurevych. "CNN-and LSTM-based Claim Classification in Online User Comments." In Proceedings of the COLING 2016.
5. Szarvas György, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. "Cross-genre and cross-domain detection of semantic uncertainty." Computational Linguistics 38, no. 2 (2012): 335-367.

### 4.1.4 Readability

#### 4.1.4.1 Task for Readability Measurement

Readability is defined as "the ease with which a reader can understand a written text" [Wikipedia].

Readability can be measured by the accuracy of reading and the reading speed for the reader. Readability depends mainly on three categories of factors: writing quality, targeted audience and presentation.

*Writing quality* refers to the grammatical correctness of the text (morphology, syntax) such as taught in elementary schools [79]. Readability also depends on the *target audience* or in other words the level of educational background the reader needs to have to understand the text content (the complexity of its vocabulary and syntax, the rhetorical structure).

Finally, the *presentation* refers to typographic aspects like font size, line height, and line length [56] or visual aspects like color [64].

#### 4.1.4.2 Methods for Readability Measurement

Collins-Thompson provides a recent state of the art summary of automatic text readability assessment [58]. Two main factors are used in readability measures: the familiarity of semantic units (vocabulary) and the complexity of syntax.

Automatic readability measures estimate the years of education or reading level required to read a given body of text using surface characteristics. The current measures are basically linear regressions based on the number of words, syllables, and sentences [70] [58].

Wikipedia presents a number of readability tests in their eponymous article that usually involve counting syllables, word length, sentence length, and number of words and sentences.<sup>3</sup>

Crossley et al. developed Coh-Metrix, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis [59]. De Clercq et al. proposed to use the crowd to predict text readability [61].

Automatic methods have been developed for different languages as for Arabic [52], French [63], Polish [57], or Spanish [75] to cite a few.

#### 4.1.4.3 Data sets for Readability Measurement

There are a number of data sets and sample demos for readability measurement. The data sets include:

- Text Exemplars and Sample Performance Tasks in Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects (183 pages). [Copyright and Permissions] Includes examples with different grades, genres (English only).<sup>4</sup>
- [62] mentions a collection of Weekly Reader extracts that may still be available.
- Math Webpage Corpus with Readability Judgments<sup>5</sup>

#### Sample demos:

- Readability<sup>6</sup> implemented by Andreas van Cranenburgh (*andreasvc* on github) calculates a number of standard reading level features, including Flesch, Kincaid and Smog (a descendent of an `nlTK_contrib` package<sup>7</sup>). This package expects sentence-segmented and tokenized text. For English, van Cranenburgh recommends “tokenizer”.<sup>8</sup> For Dutch, he recommends the tokenizer that is part of the Alpino parser<sup>9</sup>. There is also *ucto*<sup>10</sup>, a general multilingual tokenizer. One can also use the tokenizer included in the Stanford NLP package.

<sup>3</sup> [https://en.wikipedia.org/wiki/Readability#Popular\\_readability\\_formulas](https://en.wikipedia.org/wiki/Readability#Popular_readability_formulas), last access Oct. 11, 2017

<sup>4</sup> [http://www.corestandards.org/assets/Appendix\\_B.pdf](http://www.corestandards.org/assets/Appendix_B.pdf)

<sup>5</sup> [https://web.archive.org/web/\\*/http://wing.comp.nus.edu.sg/downloads/mwc](https://web.archive.org/web/*/http://wing.comp.nus.edu.sg/downloads/mwc)

<sup>6</sup> <https://pypi.python.org/pypi/readability>

<sup>7</sup> [https://github.com/nltk/nltk\\_contrib/tree/master/nltk\\_contrib/readability](https://github.com/nltk/nltk_contrib/tree/master/nltk_contrib/readability)

<sup>8</sup> <http://moin.delph-in.net/WeSearch/DocumentParsing>

<sup>9</sup> <http://www.let.rug.nl/vannoord/alp/Alpino/>

<sup>10</sup> <http://ilk.uvt.nl/ucto>

**Test cases:**

```
$ ucto -L en -n -s '' 'CONRAD, Joseph - Lord Jim.txt' | readability
[...]
readability grades:
Kincaid:                4.95
ARI:                    5.78
Coleman-Liau:           6.87
FleschReadingEase:      86.18
GunningFogIndex:        9.4
LIX:                    30.97
SMOGIndex:              9.2
RIX:                    2.39
```

**Other tools:** Benchmark Assessor Live<sup>11</sup>, and also see Further Reading, below.

**4.1.4.4 Further reading for Readability Measuring**

1. Flesch and Kincaid Readability tests, <sup>12</sup> and the Wikipedia article on Readability<sup>13</sup> (for several other readability formulas)
2. Heilman, Michael, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. “om-bining lexical and grammatical features to improve readability measures for first and second language texts.” In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 460-467. 2007.
3. Collins-Thompson, Kevyn. “Computational assessment of text readability: A survey of current and future research.” *ITL-International Journal of Applied Linguistics* 165, no. 2 (2014): 97-135.
4. De La CHICA, Sebastian, Kevyn B. Collins-Thompson, Paul N. Bennett, David Alexander Sontag, and Ryen W. White. “Using reading levels in responding to requests.” U.S. Patent 9,600,585, issued March 21, 2017.
5. Vajjala, Sowmya, and Detmar Meurers. “On the applicability of readability models to web texts.” In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pp. 59-68. 2013.
6. Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. “Frequent words improve readability and short words improve understandability for people with dyslexia.” In *IFIP Conference on Human-Computer Interaction*, pp. 203-219. Springer, Berlin, Heidelberg, 2013. (excerpt: ... To determine how much individual queries differ in terms of the readability of the documents they retrieve, we also looked at the results for each query separately. Figure 4 shows the mean reading level of the Top-100 results for each of the 50 search queries...)
7. Newbold, Neil, Harry McLaughlin, and Lee Gillam. “Rank by readability: Document weighting for information retrieval.” *Advances in multidisciplinary retrieval* (2010): 20-30. (“...Web pages can be, increasingly, badly written with unfamiliar words, poor use of syntax, ambiguous phrases and so on...”)

<sup>11</sup> <https://www.readnaturally.com/assessment-tools>

<sup>12</sup> [https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid\\_readability\\_tests](https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests)

<sup>13</sup> <https://en.wikipedia.org/wiki/Readability>



8. Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. “A comparison of features for automatic readability assessment.” In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 276-284. Association for Computational Linguistics, 2010.

#### 4.1.5 Virality

In analyses of information objects and information flows on the internet the notion of “virality” is often stressed. This is especially true when discussed in the context of marketing and advertisement. Virality means that “information objects spread the way that viruses propagate. [Hence,v]irality has become a common way to describe how thoughts or information move through a human population, or the internet, social network sites in particular”<sup>14</sup>. The metaphor of the virus supports consideration of different properties that may influence the spread of information but that can also be used to quantify virality.

##### 4.1.5.1 Task for Virality Detection

For the detection of virality in texts or other information objects four types of property sets have to be taken into account: a) the sender, b) the information object, c) the recipient, and d) the channel. The combination of these sets influences the speed with which a virus spreads and also determines how far it can reach. The major factors on the sender side are their popularity and authority, the size of their network, but also the amount of trust they receive from recipients. The recipient must be able to receive the information object and should not be immune to it, e.g. because they had the information object before. The information object itself is often admissible to many different types of recipients, for example, because of its short topical distance to knowledge the recipients already hold. The channel offers varying functionalities and allows for different ease of use to further spread the information object. Higher ease of use encourages the sharing of information objects, e.g., retweeting a tweet on Twitter. Moreover, the environment in which the information object spreads is of interest, too. It may have been influenced by a frame setting activity, i.e. bringing certain information to the awareness of many recipients, that increases the probability of recipients getting infected, e.g. because they search for this type of information. The virality of information objects could also be subject to within-platform as well as cross-platform properties.

##### 4.1.5.2 Methods for Virality Detection

The determination of virality needs to operationalize all of these factors, especially with regard to the graph-like structure of the information flow. In social media, many signals can be used for this, e.g., number of likes, retweets, and comments, characteristics of followers, communities, or hashtags, or time of posting. Those factors build the ground for virality measurement. However, it is not only the quantity of these signals that may determine virality but also the speed with which information objects spread and how far they reach (e.g., when different communities are infected by the same information object).

---

<sup>14</sup> [https://en.wikipedia.org/wiki/Viral\\_phenomenon](https://en.wikipedia.org/wiki/Viral_phenomenon)

#### 4.1.5.3 Tools and Data for Virality Detection

Examples for existing software that visualizes the spread of claims (i.e. Hoaxy) or that follows memes are provided by the Indiana University Network Science Institute (IUNI) and the Center for Complex Networks and Systems Research (CNetS)<sup>15</sup>.

There are also several data sets available that can be used for training, for example viral images<sup>16</sup> or tweets<sup>17</sup>, see also Weng *et al.* in the Further Reading subsection.

#### 4.1.5.4 Further reading for Virality

1. Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. “Virality prediction and community structure in social networks.” *Scientific reports* 3 (2013): 2522.
2. Weng, Lilian, and Filippo Menczer. “Topicality and impact in social media: diverse messages, focused messengers” *PloS one* 10, no. 2 (2015): e0118410.
3. Guerini, Marco, Carlo Strapparava, and Gözde Özal. “Exploring Text Virality in Social Networks.” In *ICWSM*. 2011.
4. Guille, Adrien and Hacid, Hakim and Favre, Cecile and Zighed, Djamel A. “Information diffusion in online social networks: A survey.” *ACM Sigmod Record* 42.2 (2013): 17-28.

### 4.1.6 Emotion

#### 4.1.6.1 Task for Emotion Detection

One characteristic of Fake News is that it may make an inflammatory emotional appeal to the reader. Emotional arguments often employ words that are charged with positive or negative connotations (such as *bold* or *cowardly*). Such language also appears in product and movie reviews.

The task here is to detect the sentences which are emotive in a document, and to calculate the intensity, the polarity and the classes of the affect words found there. The emotional impact of a document can either be averaged over the number of words, or be calculated by using some maximum value encountered [55].

#### 4.1.6.2 Methods for Emotion Detection

As a sample method, an emotion detection method can include the following steps:

1. Divide document into sentences
2. Extract words, terms, negations, intensifiers, emoticons, parts of speech, punctuation from the sentence
3. Use these extracted items as features to classify the sentence
4. Identify which sentences carry emotion, and what emotion
5. Combine measures from all sentences to create a single emotion rating of the document.

#### 4.1.6.3 Data sets for Emotion Detection

Data resources for emotion detection include sentiment lexicons and test/training data sets. Some of the former are:

---

<sup>15</sup> <http://truthy.indiana.edu>

<sup>16</sup> <https://github.com/ArturoDeza/virality>

<sup>17</sup> <http://carl.cs.indiana.edu/data/#virality2013>

- A list of Affect Lexicons<sup>18</sup> maintained by Saif Mohammad
- SenticNet<sup>19</sup>
- AFINN<sup>20</sup>
- List of affect resources<sup>21</sup> maintained by Bing Liu
- Affective Norms for English Words (ANEW) is a set of normative emotional ratings for 2,476 English words. We use the “valence” rating considering positive (respectively, negative) the ratings above (respectively, below) the mean.
- General Inquirer is a list of 1,915 words classified as positive, and 2,291 words classified as negative.
- MicroWNOp is a list of 1,105 WordNet synsets (cognitive synonyms) classified as positive, negative, or neutral.
- SentiWordNet assigns to each synset of WordNet (around 117,000) a positive and negative score determined by a diffusion process.
- Bias Lexicon is a list of 654 bias-related lemmas extracted from the edit history of Wikipedia [72]. Sentiment words are used as contributing features in the construction of this bias lexicon.

Test and training data sets include: Reviews;<sup>22</sup> Twitter in 15 languages;<sup>23</sup> Twitter and emotions;<sup>24</sup> Twitter tweets;<sup>25</sup> Blog sentences;<sup>26</sup> Facebook statuses, CNN, the New York Times, Guardian, BBC news, ABC news;<sup>27</sup> three emotional dimensions (Valence, Arousal and Dominance)<sup>28</sup>

#### 4.1.6.4 Further reading for Emotion Detection

1. Valitutti, Alessandro, and Carlo Strapparava. “Interfacing WordNet-affect with OCC model of emotions.” In *The Workshop Programme*, p. 16. 2010.<sup>29</sup>
2. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. “Sentiment analysis algorithms and applications: A survey.” *Ain Shams Engineering Journal* 5.4 (2014): 1093-1113.
3. Giachanou, Anastasia, and Fabio Crestani. “Like it or not: A survey of twitter sentiment analysis methods.” *ACM Computing Surveys (CSUR)* 49, no. 2 (2016): 28.
4. Cambria, Erik. “Affective computing and sentiment analysis.” *IEEE Intelligent Systems* 31, no. 2 (2016): 102-107.
5. Tripathi, Vaibhav, Aditya Joshi, and Pushpak Bhattacharyya. “Emotion Analysis from Text: A Survey.”<sup>30</sup>

<sup>18</sup> <http://saifmohammad.com/WebPages/lexicons.html>

<sup>19</sup> <http://sentic.net/downloads/>

<sup>20</sup> [http://www2.imm.dtu.dk/pubdb/views/publication\\_details.php?id=6010](http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010)

<sup>21</sup> <https://www.cs.uic.edu/liub/FBS/sentiment-analysis.html>

<sup>22</sup> <https://www.cs.uic.edu/liub/FBS/sentiment-analysis.html#datasets>

<sup>23</sup> <https://www.clarin.si/repository/xmlui/handle/11356/1054>

<sup>24</sup> <http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html>

<sup>25</sup> <http://www.sananalytics.com/lab/twitter-sentiment/>

<sup>26</sup> <https://inclass.kaggle.com/c/si650winter11/data>

<sup>27</sup> <https://github.com/minimaxir/interactive-facebook-reactions/tree/master/data>

<sup>28</sup> <https://github.com/JULIELab/EmoBank/tree/master/corpus>

<sup>29</sup> <https://source.opennews.org/articles/analysis-emotional-language/>

<sup>30</sup> <http://www.cfil.itb.ac.in/resources/surveys/emotion-analysis-survey-2016-vaibhav.pdf>

■ **Table 1** Examples of Fact vs Opinion sentences as taught to US Elementary School Children<sup>a</sup>, along with a score which could be computed from them.

Sentence	Label
The first amendment includes the most misused freedom in our country, which is the freedom of the press.	<b>Opinionated</b>
The 18th amendment to the constitution prohibited the manufacture, sale, or transportation of alcohol.	<b>Fact</b>
The 16th amendment gave congress to collect taxes from American citizens, and they have been collecting way too many taxes ever since	<b>Opinionated</b>
Result	Opinion-Ratio = 2/3

<sup>a</sup> <http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postest.html>

#### 4.1.7 Opinion

Opinion is an element of the text which reflects the author's opinion, and readers' opinions may differ. The output is a percentage, based on the fraction of words or sentences which are opinion, in contrast to facts. Authors of opinionated text may be surreptitiously pushing a certain viewpoint which is not explicitly expressed in the text.

##### 4.1.7.1 Task for Opinion Detection

For the Information Nutrition Label, our task is to detect sentences that are opinionated, and calculate the percentage of opinionated sentences for entire text. Table 1 gives some examples of opinionated and factual sentences.

##### 4.1.7.2 Existing Methods for Opinion Detection

There is software available for opinion detection. Here are some:

- OpenNER<sup>31</sup> “aims to be able to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on the texts<sup>32</sup>...”
- Opinion Finder<sup>33</sup>, see Wilson et al, in Further Readings below.
- Opinion Sentence Finder<sup>34</sup>. See also Rajkumar et al., below.
- NLTK opinion lexicon reader<sup>35</sup>.

##### 4.1.7.3 Data sets for Opinion Detection

There are also data sets for opinion detection:

- Fact vs. opinion as taught to US Elementary School Children.<sup>36</sup> These examples have answers<sup>37</sup>, too. The overall output score is the percent of sentences which contain opinions.

<sup>31</sup> <http://www.opener-project.eu/>

<sup>32</sup> <http://www.opener-project.eu/getting-started/#opinion-detection>

<sup>33</sup> [http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder\\_2/](http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/)

<sup>34</sup> <http://cse.iitkgp.ac.in/resgrp/cnerg/temp2/final.php>

<sup>35</sup> [http://www.nltk.org/\\_modules/nltk/corpus/reader/opinion\\_lexicon.html](http://www.nltk.org/_modules/nltk/corpus/reader/opinion_lexicon.html)

<sup>36</sup> <http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postest.html>

<sup>37</sup> <http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postans.html>

- Bitterlemon collection 594 editorials about the Israel-Palestine conflict, 312 articles from Israeli authors and 282 articles from Palestinian authors.
- Opinion lexicon<sup>38</sup>
- Multi perspective question answering lexicon<sup>39</sup> corpus contains news articles and other text documents manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.).
- Arguing Lexicon<sup>40</sup>: includes patterns that represent arguing.

#### 4.1.7.4 Further reading for Opinion Detection

1. Fact vs opinion as taught to US Elementary School Children<sup>41</sup>
2. Paul, Michael J., ChengXiang Zhai, and Roxana Girju. “Summarizing contrastive viewpoints in opinionated text.” In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 66-76. Association for Computational Linguistics, 2010.
3. Yu, Hong, and Vasileios Hatzivassiloglou. “Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences.” In Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 129-136. Association for Computational Linguistics, 2003. “classify sentences as fact / opinion using word n-grams, word polarity”
4. Liu, Bing, Mingqing Hu, and Junsheng Cheng. “Opinion observer: analyzing and comparing opinions on the web.” In Proceedings of the 14th international conference on World Wide Web, pp. 342-351. ACM, 2005.
5. Wilson, Theresa, David R. Pierce, and Janyce Wiebe. “Identifying opinionated sentences.” In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4, pp. 33-34. Association for Computational Linguistics, 2003.
6. Rajkumar, Pujari, Swara Desai, Niloy Ganguly, and Pawan Goyal. “A Novel Two-stage Framework for Extracting Opinionated Sentences from News Articles.” In TextGraphs@EMNLP, pp. 25-33. 2014.

#### 4.1.8 Controversy

Controversy is a state of prolonged public dispute or debate, usually concerning a matter of conflicting opinion or point of view. The word was coined from the Latin *controversia*, as a composite of *controversus* – “turned in an opposite direction,” from *contra* – “against” – and *vertere* – to turn, or *versus* (see *verse*), hence, “to turn against.” The most applicable or well known controversial subjects, topics or areas are politics, religion, philosophy, parenting and sex (see Wikipedia articles in Further Reading, as well as Aharoni et al.) History is similarly controversial. Other prominent areas of controversy are economics, science, finances, culture, education, the military, society, celebrities, organisation, the media, age, gender, and race. Controversy in matters of theology has traditionally been particularly heated, giving rise to the phrase *odium theologicum*. Controversial issues are held as potentially divisive in a

<sup>38</sup> <https://www.cs.uic.edu/liub/FBS/sentiment-analysis.html#lexicon>

<sup>39</sup> [http://mpqa.cs.pitt.edu/corpora/mpqa\\_corpus/](http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/)

<sup>40</sup> [http://mpqa.cs.pitt.edu/lexicons/arg\\_lexicon](http://mpqa.cs.pitt.edu/lexicons/arg_lexicon)

<sup>41</sup> <http://teaching.monster.com/training/articles/2589-k-5-fact-versus-opinion>

given society, because they can lead to tension and ill will, and as a result they are often considered taboo to be discussed in the light of company in many cultures.

Wikipedia lists some 2000 controversial issues.

#### 4.1.8.1 Task for Controversy Detection

In its simplest form, for the Information Nutrition Label, we can calculate the number of controversial subjects in the text. A more evolved form would calculate the density of controversial subjects in the text.

#### 4.1.8.2 Methods for Controversy Detection

One method we can suggest for calculating the controversy of a text would be to look at those papers that implement Wikipedia featured article detection: they have to address the controversy flaw (the developed technology has parts that apply to non-Wikipedia articles as well). For topics that are covered by Wikipedia, determine the portion of reverts (after article editing), the so-called “edit wars” in Wikipedia. See the coverage measure (essay articles) below. Compute a number of features that hint controversy: topicality, retweet number and probability, query logs.

#### 4.1.8.3 Data sets for Controversy Detection

Data Sources: Aharoni *et al.* (see further reading) describes a novel and unique argumentative structure dataset. This corpus consists of data extracted from hundreds of Wikipedia articles using a meticulously monitored manual annotation process. The result is 2,683 argument elements, collected in the context of 33 controversial topics, organized under a simple claim-evidence structure. The obtained data are publicly available for academic research.

The paper by Dori-Hacohen and Allan below also has a data set.

Test cases

Balance the number of pro and con arguments, using an argument search engine<sup>42</sup>. For queries/documents, which contain one of the controversial topics listed on the Wikipedia page, search/find documents that discuss (essay-like style) a topic. Choose documents appropriate for a specific reading level/background. Extract keywords/concepts and measure the overlap with controversial topics list (Wikipedia), debate portals, and the like.

#### 4.1.8.4 Further reading for Controversy Detection

1. Wikipedia “Controversy” article <sup>43</sup>
2. Wikipedia list of controversial issues <sup>44</sup>
3. Examples of discussions of controversial topics can be found in the Scientific American<sup>45</sup> and on Plato<sup>46</sup>.
4. Aharoni, Ehud, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Rutu Rinott, Dan Gutfreund, and Noam Slonim. “A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics.” In *ArgMining@ACL*, pp. 64-68. 2014.

<sup>42</sup> <http://141.54.172.105:8081/>

<sup>43</sup> <https://en.wikipedia.org/wiki/Controversy>

<sup>44</sup> [https://en.wikipedia.org/wiki/Wikipedia:List\\_of\\_controversial\\_issues](https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues)

<sup>45</sup> <https://www.scientificamerican.com/article/fact-or-fiction-vaccines-are-dangerous/>

<sup>46</sup> <https://plato.stanford.edu/entries/creationism/>

5. Dori-Hacohen, Shiri, and James Allan. “Detecting controversy on the web.” In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp. 1845-1848. ACM, 2013. “... Our approach maps a webpage to a set of Wikipedia articles, and uses the controversiality of those ... used two stop sets, the 418 INQUERY stop set [4] or a short, 35 term set (“Full” vs. ... 3. Handling non labeled data: We use two alternatives to “fill in the blanks” when labeled data ...”

#### 4.1.9 Authority / Credibility / Trust

For the Information Nutrition Label, we consider trust and authority as synonyms that refer to a property of the source of a message, while credibility is an attribute of the message itself. On the Web, trust is assigned to a web site, while the different pages of the web site may be different in terms of credibility. When looking at a single document, users are most interested in its credibility; on the other hand, even experienced users judge credibility mainly based on their trust of the source. In the same way, for a system, however, it is easier to estimate the authority of a source (based on the information available), while there might be little document-specific evidence concerning its credibility.

##### 4.1.9.1 Task for Authority

The task is to determine the authority or trust of the source of a document. Here we focus on Web sites and social media as sources.

##### 4.1.9.2 Methods for Authority

For Web sites, a large number of methods for estimating authority have been proposed, of which we mention just a few:

- PageRank (Further Reading 1) is the most popular method for computing the importance of a Web site.
- Kleinberg’s HITS algorithm (Further Reading 2) distinguishes between hub and authority scores.
- BrowseRank (Further Reading 3) computes the importance of a Web site by analysing user behavior data.
- Alexa Rank<sup>47</sup> measures Web site’s popularity based solely on traffic to that site, in the form of a combined measure of unique visitors and page views of a website.

Recently, there also have been some approaches addressing the credibility of social media messages:

- Tweetcreed [4] is a Chrome browser extension computing a credibility score for a tweet using six types of features: meta-data, content-based simple lexical features, content-based linguistic features, author, external link URL’s reputation, and author network.
- Sharrieff et al. (Further Reading 5) aimed at estimating credibility perception of Twitter news considering features such as reader demographics, news attributes and tweet features.
- Popat et al. (Further Reading 6) presents a method for automatically assessing the credibility of claims in a message, which retrieves corresponding articles and models their properties such as the stance language style, their reliability, time information as well as their interrelationships.

---

<sup>47</sup> <https://www.alexa.com>

#### 4.1.9.3 Data sets for Authority and Trust

- Kakol et al. (Further Reading 7) provides a manually annotated dataset that can be used for credibility prediction <sup>48</sup>.
- Popat et al. (Further Reading 6) collected data from Wikipedia and snopes.com<sup>49</sup>

#### 4.1.9.4 Further reading for Authority and Trust

1. Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
2. J. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, 46:604–632, 1999.
3. Yuting Liu , Bin Gao , Tie-Yan Liu , Ying Zhang , Zhiming Ma , Shuyuan He , Hang Li, BrowseRank: letting web users vote for page importance, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, July 20-24, 2008, Singapore, Singapore [doi>10.1145/1390334.1390412]
4. Gupta, Aditi, Ponnuram Kumaraguru, Carlos Castillo, and Patrick Meier. “Tweetcred: A real-time Web-based system for assessing credibility of content on Twitter.” In Proc. 6th International Conference on Social Informatics (SocInfo). Barcelona, Spain. 2014.
5. Shafiza Mohd Shariff, Xiuzhen Zhang, Mark Sanderson. “On the credibility perception of news on Twitter: Readers, topics and features.” Computers in Human Behavior 75 (2017) 785-794.
6. Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. “Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media.” In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1003-1012. International World Wide Web Conferences Steering Committee, 2017.
7. Kakol, Michal, Radoslaw Nielek, and Adam Wierzbicki. “Understanding and predicting Web content credibility using the Content Credibility Corpus.” Information Processing & Management 53, no. 5 (2017): 1043-1061.

#### 4.1.10 Technicality

An article may be well written and grammatically understandable, but its content may cover concepts only understandable to people learned in a certain domain. These documents may deal with a technical issue or use a large proportion of technical terms.

##### 4.1.10.1 Task for Technicality Measurement

For our Information Nutrition Label, we want to calculate a technicalness score, or technicality, for a document that indicates how hard it would be to understand for someone outside the field.

---

<sup>48</sup> <https://github.com/s8811/reconcile-tags>

<sup>49</sup> <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/>



#### 4.1.10.2 Methods for Technicality Measurement

Similar to Readability, but more related to content than form, Technicality is a property of a document capturing the proportion of the domain-specific vocabulary used by the document. Style-based features are already captured by the readability score.

#### 4.1.10.3 Data sets for Technicality Measurement

Data Sources:

- Terminology extraction software<sup>5051</sup>
- Further tools are available from<sup>52</sup>
- In Wikipedia, external links provide a set of freely available tools under “Terminology Extraction”<sup>53</sup>
- Word frequency information<sup>54</sup> (English), in German<sup>555657</sup>, in other languages<sup>58</sup>

Test cases and benchmarks:

- ACL RD-TEC<sup>59</sup>. QasemiZadeh, Behrang, and Anne-Kathrin Schumann. “The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods.” In LREC. 2016.
- GENIA Corpus<sup>60</sup> is a popular corpus that has been used to evaluate various ATE algorithm for the last decade. In JATE2, instead of using the annotation file “GENIAcorpus302.xml”, the ‘concept.txt’ containing a breakdown list of GENIA concepts and relations (more like ontology) are used as the “Gold Standard” (GS) list.

#### 4.1.10.4 Further reading for Technicality Measurement

1. Justeson, John S., and Slava M. Katz. “Technical terminology: some linguistic properties and an algorithm for identification in text.” *Natural language engineering* 1, no. 1 (1995): 9-27.
2. Dagan, Ido, and Ken Church. “Termight: Identifying and translating technical terminology.” In *Proceedings of the fourth conference on Applied natural language processing*, pp. 34-40. Association for Computational Linguistics, 1994.
3. Pazienza, Maria, Marco Pennacchiotti, and Fabio Zanzotto. “Terminology extraction: an analysis of linguistic and statistical approaches.” *Knowledge mining* (2005): 255-279.

#### 4.1.11 Topicality

Topical documents are documents which cover topics that are in the current zeitgeist.

<sup>50</sup> <https://github.com/termsuite/termsuite-core>

<sup>51</sup> <https://github.com/texta-tk/texta>

<sup>52</sup> <https://github.com/search?utf8=%E2%9C%93&q=terminology+extraction>

<sup>53</sup> [https://en.wikipedia.org/wiki/Terminology\\_extraction](https://en.wikipedia.org/wiki/Terminology_extraction)

<sup>54</sup> <http://www.wordfrequency.info>

<sup>55</sup> <http://corpus.leeds.ac.uk/frqc/internet-de-forms.num>

<sup>56</sup> <http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html>

<sup>57</sup> <http://wortschatz.uni-leipzig.de/de/download>

<sup>58</sup> [https://web.archive.org/web/\\*/http://wortschatz.uni-leipzig.de/html/wliste.html](https://web.archive.org/web/*/http://wortschatz.uni-leipzig.de/html/wliste.html)

<sup>59</sup> <https://github.com/language-recipes/acl-rd-tec-2.0>

<sup>60</sup> <https://github.com/ziqizhang/jate/wiki/Evaluation-and-Dataset>

#### 4.1.11.1 Task for Topicality Detection

Topicality detection here means to decide whether the document is of current interest or not. One of the salient points of the negative effect of fake news was to falsely influence thinking about things in the current news cycle.

#### 4.1.11.2 Methods for Topicality Detection

Extract the salient terms (keyterms) and entities of the document. Compare those terms to the terms found in recent news or publications, or search engine queries.

Tools:

- Text Mining Online<sup>61</sup>
- KeyPhrase Extraction<sup>6263</sup>

#### 4.1.11.3 Data sets for Topicality Detection

Current topics can be found on these sites, for example, ABC News<sup>64</sup>, or lists of current events<sup>65</sup>. Current news and compiled multilingual lists of entities can be found at the UE-funded EMM NewsExplorer<sup>66</sup>

#### 4.1.11.4 Further reading for Topicality Detection

1. Zafar, Muhammad Bilal, et al. Zafar, Muhammad Bilal, Parantapa Bhattacharya, Niloy Ganguly, Saptarshi Ghosh, and Krishna P. Gummadi. “On the Wisdom of Experts vs. Crowds: Discovering Trustworthy Topical News in Microblogs.” In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 438-451. ACM, 2016
2. Wu, Baoning, Vinay Goel, and Brian D. Davison. “Topical trustrank: Using topicality to combat web spam.” In Proceedings of the 15th international conference on World Wide Web, pp. 63-72. ACM, 2006.
3. Diakopoulos, Nicholas, and Arkaitz Zubiaga. “Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance.” In ICWSM. 2014.

---

<sup>61</sup> <http://textminingonline.com/how-to-use-stanford-named-entity-recognizer-ner-in-python-nltk-and-other-programming-languages> Keyphrase extraction

<sup>62</sup> <https://github.com/luffycodes/KeyphraseExtraction> , <https://github.com/Gelembjuk/keyphrases>

<sup>63</sup> <https://github.com/snkim/AutomaticKeyphraseExtraction>

<sup>64</sup> <http://abcnews.go.com/topics/>

<sup>65</sup> <http://libguides.umflint.edu/topics/current> or <http://www.libraryspot.com/features/currentevents.htm>

<sup>66</sup> <http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html>

## 4.2 Rethinking Summarization and Storytelling for Modern Social Multimedia

*Stevan Rudinac (University of Amsterdam, NL), Tat-Seng Chua (National University of Singapore, SG), Nicolas Diaz-Ferreyra (Universität Duisburg-Essen, DE), Gerald Friedland (University of California – Berkeley, US), Tatjana Gornostaja (tilde – Riga, LV), Benoit Huet (EURECOM – Sophia Antipolis, FR), Rianne Kaptein (Crunchr – Amsterdam, NL), Krister Lindén (University of Helsinki, FI), Marie-Francine Moens (KU Leuven, BE), Jaakko Peltonen (Aalto University, FI), Miriam Redi (NOKIA Bell Labs – Cambridge, GB), Markus Schedl (Universität Linz, AT), David Ayman Shamma (CWI – Amsterdam, NL), Alan Smeaton (Dublin City University, IE), and Lexing Xie (Australian National University – Canberra, AU)*

**License** © Creative Commons BY 3.0 Unported license

© Stevan Rudinac, Tat-Seng Chua, Nicolas Diaz-Ferreyra, Gerald Friedland, Tatjana Gornostaja, Benoit Huet, Rianne Kaptein, Krister Lindén, Marie-Francine Moens, Jaakko Peltonen, Miriam Redi, Markus Schedl, David Ayman Shamma, Alan Smeaton, and Lexing Xie

Traditional summarization initiatives have been focused on specific types of documents such as articles, reviews, videos, image feeds, or tweets, a practice which may result in pigeonholing the summarization task in the surrounding of modern, content-rich multimedia collections. Consequently, much of the research to date has revolved around mostly toy problems in narrow domains and working on single-source media types. We argue that summarization and story generation systems need to refocus the problem space in order to meet the information needs in the age of user-generated content in different formats and languages. Here we create a framework for flexible multimedia storytelling. Narratives, stories, and summaries carry a set of challenges in big data and dynamic multi-source media that give rise to new research in spatial-temporal representation, viewpoint generation, and explanation.

### 4.2.1 Introduction

Social Multimedia [44] has been described as having three main components: content interaction between multimedia, social interaction around multimedia and social interaction captured in multimedia. Roughly speaking, this describes the interaction between traditional multimedia (photos and videos), mostly textual annotations on that media, and people interacting with that media. For almost a decade, fueled by the popularity of User-Generated Content (UGC), the bulk of research [18, 3, 40, 20, 1, 14, 9] has focused on meaningful extraction from any combination of these three points. With modern advancements in AI and computational resources [27, 19], we now realize that multimedia summarization and story telling has worked in isolated silos, depending on the application and media (object detection, video summarization, Twitter sentiment, etc.); a broader viewpoint on the whole summarisation and reduction process is needed. Consequently, this realization gives rise to a second set of research challenges moving forward. In this paper, we revisit and propose to reshape the future challenges in multimedia summarization to identify a set of goals, prerequisites, and guidelines to address future UGC. Specifically, we address the problems associated with increasingly heterogeneous collections both in terms of multiple media and mixed content in different formats and languages, the necessity and complexities of dense knowledge extraction, and the requirements needed for sense making and storytelling.

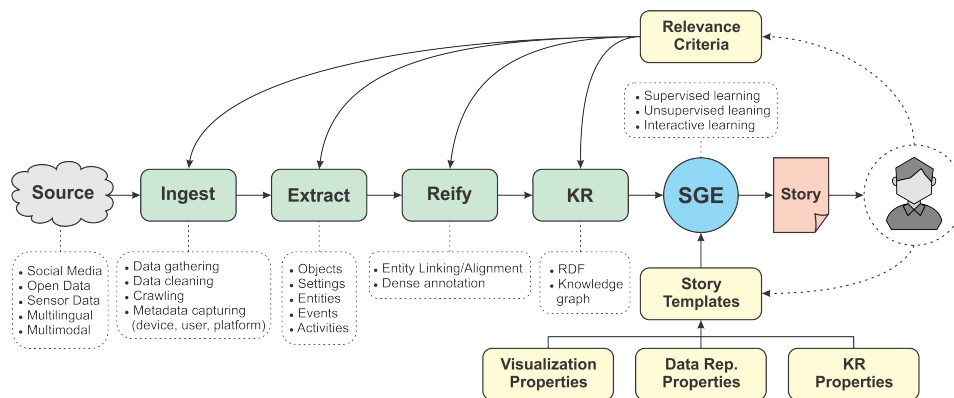
#### 4.2.2 Related Work

**Summarization problems.** Content summarisation problems arise in different application domains, and are a long-standing interest of the natural language processing, computer vision, and multimedia research communities. Summarising long segments of text from a single or multiple documents is often done with extractive techniques, on which extensive surveys exist [28]. The problem of summarising image collections arises when there are e.g. large amounts of images from many users in a geographic area [36, 35], or about a particular social event [49], or when it is necessary to generate a summarizing description (caption) [9].

Similarly, it is often needed to shorten or find alternative presentation for long video sequences. Automatic story boards were probably first introduced by the CMU Informedia project [2] and video manga system is another early example of video-to-image summarisation in a comic-book layout [45]. Summarizing videos based on both audio and visual channels involved e.g. joint optimisation of cross-modal coherence [41], or matching of audio segments [16]. Summarization of ego-centric or surveillance videos attracted much attention recently, with the example approaches including finding important objects and actions [26] or constructing a map of key people in a known environment [50]. Many multimedia summarization problems are driven by real-world events at different time scales [49] and in the last decade there is also an increasing focus on large-scale social events reported online [16, 48]. This position paper examines the summarization problem more broadly, taking a step back from one particular media format to be summarized, and targeting a large range of applications.

**Relevance criteria for summarization.** Early approaches to information retrieval (IR) and summarization focused on relevance of the content presented to the user. However, by the end of 90s the community realized that users prefer diversified search results and summaries instead of results lists produced based on relevance criterion only [2]. While the application domains varied, since then most summarization approaches focused on finding a balance between relevance, representativeness and diversity. The Informedia project is one of the best known early examples following such paradigm in addressing, amongst others, the problem of video summarization [47]. However, as users may be more sensitive to irrelevant than (near) duplicate items, enforcing diversity without hurting relevance is very challenging. This is witnessed by a large body of research on e.g. image search diversification [17, 46, 23, 37, 15]. Social multimedia summarization has further found its way in diverse applications ranging from personalized tweet summarization [34] to visual summarization of geographic areas and tourist routes [36, 35, 12, 17] for POI recommendation and exploration. With the increased availability of affordable wearables, in recent years lifelogging has started gaining popularity, where the goal is to generate a diary or a record of the day's activities and happenings by creating a summary or a story from the video/image data gathered [11, 22]. Progress has been made in summarizing heterogeneous user-generated content with regards to relevance, representativeness, and diversity [2]. However, relevance criteria and their interplay may be much more complex than commonly assumed [36] and, in case of visual content, include additional factors such as content popularity, aesthetic appeal and sentiment. Thus we call for rethinking the foundations of summarization and storytelling.

**Benchmarks and Formalization Efforts.** For almost two decades, common datasets, tasks, and international benchmarks fuelled research on summarization and storytelling [13, 5, 31]. A typical task involved automatically generating a shorter (e.g. 100-word) summary of a set of news articles. TRECVID BBC Rushes summarization was probably the first systematic effort in the multimedia and computer vision communities focusing on video summarization [30]. The task involved reducing a raw and unstructured video captured during the recording of a



■ **Figure 2** Pipeline of our proposed framework for generating narratives, stories and summaries from heterogeneous collections of user generated content and beyond.

TV series to a short segment of just a couple of minutes. Another well-known example is the ImageCLEF 2009 Photo Task, which revolved around image diversification [23]. The participants were expected to produce image search results covering multiple aspects of a news story, such as the images of “Hillary Clinton”, “Obama Clinton” and “Bill Clinton” for a query “Clinton”. Image search diversification has also been a topic of an ongoing MediaEval Diverse Social Images Task, run annually since 2013 [15].

Although many people intuitively understand the concept of summarization, the complexity of the problem is best illustrated by the difficulties in even unequivocally defining a summary [33]. So, instead of focusing on strict definitions, most benchmarks took a pragmatic approach by conducting either intrinsic or extrinsic evaluation [13]. In intrinsic evaluation an automatically generated summary is compared directly against a “standard”, such as summaries created by the humans. On the other hand, extrinsic evaluation measures the effectiveness of a summary in fulfilling a particular task as compared with the original set of documents (e.g. text, images or videos). Over the years many interesting metrics for evaluating (text) summaries were proposed, such as recall-oriented understudy for gisting evaluation (ROUGE) [25], bilingual evaluation understudy (BLEU) [32] and Pyramid Score [29]. Some of these were later on successfully adapted to the visual domain [24, 36]. These initiatives had an impact on the progress in the field of summarization. However, their almost exclusive focus on a single modality (e.g. text or visual) or language and the traditional tasks (e.g. text document and video summarization or search diversification) does not reflect the richness of social multimedia and the complex use cases it brings.

#### 4.2.3 Framework Overview

First, we take a step back and look at a media-agnostic birds-eye view of the problem (see Figure 2). We therefore imagine a generic framework that follows the requirements as driven by the user, instead of the technology. Figure 2 shows an overview of the concept, which follows the standard pattern of a media pipeline along the “ingest,” “extract,” “reify” paradigm. The goal of the framework is to create a story for the user, who is querying for information using a set of relevance criteria. Before doing that, we assume the user has configured the framework somehow, e.g. to choose some visualization template and define basic properties of the story. We then assume a tool that would query a set of sources from the Internet or elsewhere, download (“ingest”) the data, “extract” relevant

information and then “reify” it in a way that it can be added into some standardized Knowledge Representation (KR). The knowledge representation would then, in connection with the initial configuration, be used to create the final story. We will next discuss technical and other challenges to be addressed by the community in order to put flesh onto our bare bones framework.

#### 4.2.4 Challenges and Example Application Domains

A framework for holistic storytelling brings a new set of research challenges and also reshapes some of the more traditional challenges in UGC. We identify these as *storytelling challenges* which include handling of time/temporality/history, dynamic labeling of noise, focused story generation, tailoring to impartiality or a viewpoint, quality assessment and explainability as well as *UGC challenges* which include ethical use, multi source fusion, multilinguality and multimodality, information extraction, knowledge update and addition of new knowledge, staying agnostic to specific application, supporting various types of open data, portability and finding a balance between depth and breadth. We now describe a set of application domains that illustrate some of the aforementioned challenges.

**Smart urban spaces.** Increased availability of open data and social multimedia has resulted in the birth of urban computing [51] and created new possibilities for better understanding a city. Although spontaneously captured, social multimedia may provide valuable insights about geographic city regions and their inhabitants. For example, user-generated content has been used to create summaries of geographic areas and tourist routes in location recommendation systems [35, 36]. Sentiment extracted from social multimedia, in combination with neighborhood statistics was also proven invaluable for a more timely estimation of city livability and its causes [8]. Similarly, when looking for signs of issues such as neighborhood decay or suboptimal infrastructure, city administrators are increasingly monitoring diverse UGC streams, ranging from social media and designated neighborhood apps to data collected by mobile towers and wearables. Efficient approaches to summarization and storytelling are needed to facilitate exploration in such large and heterogeneous collections.

**Field study/survey.** Consumer-produced multimedia contains data which is not only relevant to the reason for creating and sharing it but also for other applications. As a side effect this information could be used for field studies of other kinds, if it can be retrieved in a timely fashion. The framework we propose and especially the kind of tools that it leads to should enable empirical scientists of many disciplines to leverage this data for field studies based on extracting required information from huge datasets. This currently constitutes a gap between the elements of what multimedia researchers have shown is possible to do with consumer-produced big data and the follow-through of creating a comprehensive field study framework supporting scientists across other disciplines. To bravely bridge this gap, we must meet several challenges. For example, the framework must handle unlabeled and noisily labeled data to produce an altered dataset for a scientist — who naturally wants it to be both as large and as clean as possible. We must also design an interface that will be intuitive and yet enable complex search queries that rely on feature and statistics generation at a large scale.

**Business intelligence.** User generated content is a valuable source of information for companies and institutions. Business information can be obtained by analyzing what the public is saying about a company, its products, marketing campaigns and competitors. Traditionally business intelligence relied on facts and figures collected from within the organisation, or

provided by third-party reports and surveys. Instead of surveys, direct feedback can be obtained by listening to what people are saying on Social Media, either directed at their own social circle, or directly at the company in the case of web care conversations. Content can consist of textual messages or videos, for example product reviews. Besides the volume of messages, the sentiment of messages is important to analyze into positive and negative aspects. The amount of user generated content can easily add up to thousands of messages on a single topic, so summarization techniques are needed to efficiently process the wealth of information available [6].

**Health and Wellness.** There is a wealth of data about our health and wellness which is generated digitally on an individual basis. This includes genomic information from companies like 23andme<sup>67</sup> which uses tissue samples from individuals to generate information about our ancestry as well as about our possible susceptibility to a range of inherited diseases. We also have information about our lifestyles which can be gathered from our social media profiles and information about our physical activity levels and sports participation from any fitness trackers that we might wear or use. When we have health tests or screening we can have indications of biomarkers from our clinical tests for such things as cholesterol levels, glucose levels, etc. We have occasional once-off readings of our physiological status via heart and respiration rates and increasingly we can use wearable sensors to continuously monitor glucose, heart rate etc. to see how these change over time. From all of this personal sensor data there is a need to generate the “*story of me*”, telling my health professional and me how well or healthy I am now, whether my health and wellness is improving or is on the slide, and if there’s anything derived from those trends that I should know.

**Lifelogs.** In this use case a large amount of first-person ethnographic video or images taken from a wearable camera over an extended period of days, weeks, months or even years, has been generated. Such a collection may be augmented and aligned with sensor data such as GPS location or biometric data like heart rate, activity levels from wearable accelerometers or stress levels from galvanic skin response sensors. There is a need to summarize each day’s or week’s activities to allow reviewing or perhaps to support search or browsing through the lifelog. Summaries should be visual, basically selecting a subset of images or videos, and applications could be in memory support where a summary of a day can be used to trigger memory recall [7]. In this case the visual summary should incorporate events, objects or activities which are unusual or rare throughout the lifelog in preference to those which are mundane or routine like mealtimes, watching TV or reading a newspaper which might be done every day [21].

**Entertainment.** Multimedia summarization and storytelling can also serve to fulfill a pure entertainment need. Respective approaches could, for instance, support event-based creation of videos from pictures and video clips recorded on smart phones. To this end, they would automatically organize and structure such user-generated multimedia content, possibly in low quality, and subsequently determine the most interesting and suited parts in order to tell the story of a particular event, e.g., a wedding. The multimedia material considered by such an event-based storytelling approach is not necessarily restricted to a single user, but could automatically determine and select the best images / scenes from the whole audience at the event, or at least those choosing to share material.

---

<sup>67</sup> <http://www.23andme.com>

#### 4.2.5 Use Cases

When rethinking the requirements, we primarily analyzed two types of use cases: summarization and storyboarding. Summarization has traditionally involved document summarization, i.e. reducing one or more pieces of text into a shorter version and video summarization where multiple or long videos are reduced to a shorter version. Summaries could include an abridged report of an event or a how-to instruction with the main points to perform a task. As data is increasingly available in many modalities and languages, it is possible to generate a summary from and in multiple modalities and languages according to the user's information request. Large events such as elections or important sports competitions are covered by many channels, including traditional media and different Social Media including text messages, images and video. New directions for summarization include interactive summaries of UGC opinions or sentiment-based data visualization, and forecasting including prediction of electoral results, product sales, stock market movements and influenza incidence [39].

While users of music streaming services are often drowning in the number of music pieces to choose from, getting an overview of a certain music genre or style, which serves an educational purpose, is barely feasible with current recommender systems technology. Addressing this, we need algorithms that automatically create consistent and entertaining summaries of the important songs of a given genre or style considering the genre's evolution over time. Such approaches need to identify the most representative and important music, use automatic structural music segmentation techniques [43, 10, 4], decide on the most salient parts, and present them in a way that connects them. Ideally, the approaches should also consider cultural perspectives to take into account that the meaning of genres such as folk music may change depending on the cultural background of the listener such as the country, among other aspects [38].

A storyboard is a summary that conveys a change over time. This may include a recount of the given input in order to tell an unbiased story of an event, e.g. the Fall of the Berlin wall or the Kennedy murder. It may also aim to present or select facts to persuade a user to perform a particular action or change opinion, e.g. pointing out the likely murderer in the Kennedy case. If the input is open-ended, the summary may be structured by background information, e.g. a composite clip giving a visual summarization of an event (such as a concert, a sports match, etc.) where the summarized input is provided by those attending the event but the story is structured according to a timeline given by background information.

#### 4.2.6 Prerequisites

Once user generated content has been gathered, extracted, and reified, it should be expressed in a KR. This is a step prior to the generation of stories and summaries which aims to describe the information of interest following a representation formalism. Some of the knowledge representation formalisms widely adopted in the multimedia community are Resource Description Framework (RDF) and Knowledge Graph. The selection of one approach over the others is tightly connected with the purpose of the summary/story and the technique used for its construction. This means that knowledge must be represented using a language with which the Story Generation Engine can reason in order to satisfy complex relevance criteria and visualization requirements (templates) specified by the users. These relevance criteria and visualization requirements imply a set of desired properties on the data and KR, as well as the end result presented to the user, which are fundamental for summarization and storytelling.



■ **Table 2** Properties data representation should have for facilitating effective summarization and storytelling.

Data Representation Properties		
<i>Location</i>	<i>Time</i>	<i>Observed</i>
Single $\Rightarrow$ Distributed	Scheduled $\Rightarrow$ Unplanned	Entity-driven $\Rightarrow$ Latent
Physical $\Rightarrow$ Virtual	Short $\Rightarrow$ Long	
Personal $\Rightarrow$ Public/Shared	Recurrent $\Rightarrow$ One-off	
Independent $\Rightarrow$ Cascaded		

#### 4.2.6.1 Representation Properties

Complex user information needs and the relevance criteria stemming from them require novel (multimodal and multilingual) data representations. In Table 2 we list some critical prerequisites they should fulfill.

**Time:** The “events” described by a story could have very different properties. For example, an event could be *scheduled* (e.g. Olympic Games) or *unplanned* (e.g. a terrorist attack). In the former case relevance criteria and the visualization templates could be easier to foresee, but an effective data representation should accommodate the latter use case as well. Similarly, the events could have a *longer* (e.g., studies abroad) or *shorter* (e.g., birthday) duration. Finally, data representation should ideally accommodate both *recurrent* and *once off* events.

**Location:** Although multimedia analysis has found its way in modeling different aspects of geographic locations [15, 35, 42], most related work addressed specific use cases and little effort has been made in identifying general “spatial” criteria underlying data representations should satisfy. In this regard, the representation should account for the events occurring at a *single* (e.g. rock concert) or *distributed* location (e.g. Olympic Games). In both cases those locations can be further *physical* or *virtual*. On the other hand, the events of interest can be *personal* or *public/shared*. While in the former case the content interpretation and relevance criteria may have a meaning for a particular individual only, the later is usually easier to analyze due to a higher “inter-user agreement”. Finally, data representation should be designed with the awareness that the aforementioned types of events could additionally be *independent* or *cascaded*.

**Observed:** In many analytic scenarios the summaries and stories presented to the user contain well-defined named *entities*, i.e. topics, people, places and organizations. An example would be a well-structured news article covering a political event. Yet the topics of interest may be *latent*, which is particularly common in social media discussions. For example, a public servant sifting through millions of social media posts in an attempt to verify an outbreak of a new virus may be interested in various unforeseen and seemingly unrelated conversations, which together provide conclusive evidence. Therefore, a good data representation should ideally provide support for both.

#### 4.2.6.2 Representation Properties

Building on best practices from the semantic web community, the results of ingestion, extraction and reification (cf. Figure 2) should be further organized in a knowledge representation. Example candidates include RDFs and knowledge graph. The KR should be flexible enough to allow for *temporal*, *spatial* and *observed* properties of the events discussed in subsection 4.2.6.1. It should further support both *implicit* and *explicit* relations between the items, as well as

■ **Table 3** Properties a knowledge representation should have.

**Knowledge Representation Properties**

Implicit $\Rightarrow$ Explicit	Independent $\Rightarrow$ Correlated/Casual
Uniqueness/Representativeness	Support for different semantic levels

■ **Table 4** Story properties that should be facilitated by the story generator engine.

**Story Properties**

Functional $\Rightarrow$ Quality	Modality-preserving $\Rightarrow$ Cross-modal
Self-contained $\Rightarrow$ Stepping-stone	Static $\Rightarrow$ Dynamic/Interactive
Succinct $\Rightarrow$ Narrative	Factual $\Rightarrow$ Stylistic
Abstractive $\Rightarrow$ Generative	Generic $\Rightarrow$ Personalized

their modification “on the fly” (cf. Table 3). The events and their building blocks could further be *independent* and *correlated/casual*. To facilitate a wide range of possible relevance criteria as well as their complex interplay, the KR should also include notions of importance, representativeness and frequency. Finally, the content interpretation and user information needs can be specified at different semantic levels, which in case of multimedia range from e.g. term or pixel statistics, semantic concepts, events and actions to the level of semantic theme and complex human interpretations involving aesthetic appeal and sentiment. Supporting a wide range of relevance is therefore a necessary condition for facilitating creation of effective and engaging summaries and stories.

#### 4.2.6.3 Properties

Given the content, data and KRs and the user information needs, the output of the pipeline depicted in Figure 2 is the story (or summary) presented to the user. Below we enumerate a number of criteria an ideal set of “story templates” should satisfy (see Table 4). A story should satisfy both *functional* (e.g. fulfilling a purpose) and *quality* (e.g. metrical) requirements [13]. The importance of a particular requirement should ideally be learned from user interactions. The system should further support *self-contained/interpretable* and *stepping-stone/connector* type of summaries. While the former by itself provides an insight into a larger multimedia item or a collection, the later serves a goal further on the horizon, such as faster collection exploration. Additionally, the design should accommodate both *succinct* and *narrative*, as well as *abstractive* and *generative* stories. With regard to the input and output modalities and languages, support should be provided for *modality-preserving* and *cross-modal* and/or *cross-lingual* use-cases. In many scenarios, user information needs can be satisfied with a *static* story. However, the size and heterogeneity of a UGC collection as well as the complexity of user information needs make *interactive* summarization and storytelling increasingly popular. Depending on the information needs, a *factual* or *stylistic* summary may be desirable, which is why the system should support both flavors and perhaps allow for interactive learning of their balance. Finally, while a *generic* story may be sufficient for some, *personalization* should also be supported.

#### 4.2.7 Conclusion

Motivated by an observation about discrepancies between state of the art research on the one hand and the increasing richness of user generated content and the accompanying complex user information needs on the other, we revisit the requirements for multimedia summarization and

storytelling. We reiterate the importance of summarization and storytelling for facilitating efficient and appealing access to large collections of social multimedia and interaction with them. Our proposed framework identifies a set of challenges and prerequisites related to data and KR as well as the process of their creation, i.e. ingestion, extraction and reification. We further make an inventory of the desirable properties a story should have for addressing a wide range of user information needs. Finally, we showcase a number of application domains and use cases that could serve as the catalyst for future research on the topic.

## References

- 1 J. Bian, Y. Yang, H. Zhang, and T. S. Chua. Multimedia summarization for social events in microblog stream. *IEEE Trans. Multimedia*, 17(2):216–228, Feb 2015.
- 2 Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for re-ordering documents and producing summaries. In *ACM SIGIR '98*, pages 335–336, New York, NY, USA, 1998. ACM.
- 3 Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *ACM IMC '07*, pages 1–14, 2007.
- 4 W. Chai. Semantic segmentation and summarization of music. *IEEE Signal Process. Mag.*, 23(2), 2006.
- 5 Hoa Trang Dang. Overview of DUC 2006. In *DUC '06*, 2006.
- 6 Lipika Dey, Sk Mirajul Haque, Arpit Khurdiya, and Gautam Shroff. Acquiring competitive intelligence from social media. In *MOCR AND '11*, page 3. ACM, 2011.
- 7 Aiden R. Doherty, Steve E. Hodges, Abby C. King, Alan F. Smeaton, Emma Berry, Chris J.A. Moulin, Siân Lindley, Paul Kelly, and Charlie Foster. Wearable cameras in health. *Am J Prev Med*, 44:320–323, March 2013.
- 8 Joost Boonzajer Flaes, Stevan Rudinac, and Marcel Worring. What multimedia sentiment analysis says about city liveability. In *ECIR '16*, pages 824–829, 2016.
- 9 Tatjana Gornostay (Gornostaja) and Ahmet Aker. Development and implementation of multilingual object type toponym-referenced text corpora for optimizing automatic image description generation. In *Dialogue '09*, 2009.
- 10 Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Trans. Audio, Speech, Language Process.*, 14(5):1783–1794, Sept 2006.
- 11 Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. Lifelogging: Personal big data. *Found. Trends Inf. Retr.*, 8(1):1–125, June 2014.
- 12 Qiang Hao, Rui Cai, Xin-Jing Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. Generating location overviews with images and tags by mining user-generated travelogues. In *ACM MM '09*, pages 801–804, New York, NY, USA, 2009. ACM.
- 13 Donna Harman and Paul Over. The DUC summarization evaluations. In *HLT '02*, pages 44–51, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.
- 14 Richang Hong, Jinhui Tang, Hung-Khoon Tan, Chong-Wah Ngo, Shuicheng Yan, and Tat-Seng Chua. Beyond search: Event-driven summarization for web videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7(4):35:1–35:18, December 2011.
- 15 Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. Result diversification in social image retrieval: a benchmarking framework. *Multimed Tools Appl*, 75(2):1301–1331, Jan 2016.
- 16 Lyndon Kennedy and Mor Naaman. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *WWW '09*, pages 311–320. ACM, 2009.

- 17 Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *ACM WWW '08*, pages 297–306, 2008.
- 18 Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. *Twitter Sentiment Analysis: The Good the Bad and the OMG!*, pages 538–541. AAAI Press, 2011.
- 19 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *NIPS '12*, pages 1097–1105. Curran Associates, Inc., 2012.
- 20 Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *ACM WWW '10*, pages 591–600, 2010.
- 21 Hyowon Lee, Alan F Smeaton, Noel E O'Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin. Constructing a SenseCam visual diary as a media process. *Multimedia Syst*, 14(6):341–349, 2008.
- 22 Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE CVPR '12*, pages 1346–1353, June 2012.
- 23 Monica Lestari Paramita, Mark Sanderson, and Paul Clough. *Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009*, pages 45–59. 2010.
- 24 Yingbo Li and Bernard Merialdo. Vert: Automatic evaluation of video summaries. In *ACM MM '10*, pages 851–854, 2010.
- 25 Chin Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL '04 Workshop*, pages 74–81, 2004.
- 26 Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE CVPR '13*, pages 2714–2721, 2013.
- 27 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS '13*, pages 3111–3119. Curran Associates, Inc., 2013.
- 28 Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. *Mining text data*, pages 43–76, 2012.
- 29 Ani Nenkova and Rebecca J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152, 2004.
- 30 Paul Over, Alan F. Smeaton, and George Awad. The TRECVID 2008 BBC rushes summarization evaluation. In *ACM TVS '08*, pages 1–20, 2008.
- 31 Karolina Owczarzak and Hoa Trang Dang. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *TAC '11*, 2011.
- 32 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL '02*, pages 311–318, 2002.
- 33 Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Comput. Linguist.*, 28(4):399–408, December 2002.
- 34 Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. Personalized time-aware tweets summarization. In *ACM SIGIR '13*, pages 513–522, 2013.
- 35 S. Rudinac, A. Hanjalic, and M. Larson. Generating visual summaries of geographic areas using community-contributed images. *IEEE Trans. Multimedia*, 15(4):921–932, June 2013.
- 36 S. Rudinac, M. Larson, and A. Hanjalic. Learning crowdsourced user preferences for visual summarization of image collections. *IEEE Trans. Multimedia*, 15(6):1231–1243, Oct 2013.
- 37 Mark Sanderson, Jiayu Tang, Thomas Arni, and Paul Clough. *What Else Is There? Search Diversity Examined*, pages 562–569. 2009.
- 38 Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *J Intell Inf Syst*, 41:523–539, December 2013.
- 39 Harald Schoen, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, and Peter Gloor. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013.

- 40 David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. Tweet the debates: Understanding community annotation of uncollected sources. In *ACM WSM '09*, pages 3–10, 2009.
- 41 Hari Sundaram, Lexing Xie, and Shih-Fu Chang. A utility framework for the automatic generation of audio-visual skims. In *ACM MM '02*, pages 189–198. ACM, 2002.
- 42 Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016.
- 43 Mi Tian and Mark B. Sandler. Towards music structural segmentation across genres: Features, structural hypotheses, and annotation principles. *ACM Trans. Intell. Syst. Technol.*, 8(2):23:1–23:19, October 2016.
- 44 Yonghong Tian, Jaideep Srivastava, Tiejun Huang, and Noshir Contractor. Social multimedia computing. *Computer*, 43(8):27–36, August 2010.
- 45 Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. Video manga: generating semantically meaningful video summaries. In *ACM MM '99*, pages 383–392. ACM, 1999.
- 46 Reinier H. van Leuken, Lluís Garcia, Ximena Olivares, and Roelof van Zwol. Visual diversification of image search results. In *ACM WWW '09*, pages 341–350, 2009.
- 47 H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens. Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52, May 1996.
- 48 Lexing Xie, Apostol Natsev, John R Kender, Matthew Hill, and John R Smith. Visual memes in social media: tracking real-world news in youtube videos. In *ACM MM '11*, pages 53–62. ACM, 2011.
- 49 Lexing Xie, Hari Sundaram, and Murray Campbell. Event mining in multimedia streams. *Proc. IEEE*, 96(4):623–647, 2008.
- 50 Shou-I Yu, Yi Yang, and Alexander Hauptmann. Harry potter’s marauder’s map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *IEEE CVPR '13*, pages 3714–3720, 2013.
- 51 Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):38:1–38:55, September 2014.
- 52 Abdel Karim Al Tamimi, Manar Jaradat, Nuha Al-Jarrah, and Sahar Ghanem. Aari: automatic arabic readability index. *Int. Arab J. Inf. Technol.*, 11(4):370–378, 2014.
- 53 Maik Anderka. *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia*. Dissertation, Bauhaus-Universität Weimar, June 2013.
- 54 Maik Anderka, Benno Stein, and Nedim Lipka. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 981–990. ACM, August 2012.
- 55 Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742–753, 2012.
- 56 Michael L Bernard, Barbara S Chaparro, Melissa M Mills, and Charles G Halcomb. Comparing the effects of text size and format on the readability of computer-displayed times new roman and arial text. *International Journal of Human-Computer Studies*, 59(6):823–835, 2003.
- 57 Bartosz Broda, Bartłomiej Niton, Włodzimierz Gruszczyński, and Maciej Ogrodniczuk. Measuring readability of polish texts: Baseline experiments. In *LREC*, pages 573–580, 2014.

- 58 Kevyn Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.
- 59 Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493, 2008.
- 60 Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493, 2008.
- 61 Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325, 2014.
- 62 Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.
- 63 Thomas François. An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107. Linköping University Electronic Press, 2014.
- 64 Richard H Hall and Patrick Hanna. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23(3):183–195, 2004.
- 65 Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. Event detection and factuality assessment with non-expert supervision. 2015.
- 66 Mary Levis, Markus Helfert, and Malcolm Brady. Information quality management: Review of an evolving research area. 01 2007.
- 67 Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the tenth conference on computational natural language learning*, pages 109–116. Association for Computational Linguistics, 2006.
- 68 Amnon Lotan, Asher Stern, and Ido Dagan. Truth-teller: Annotating predicate truth. 2013.
- 69 A-L Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, MGJ van Erp, AM Schoen, CM van Son, et al. Meantime, the newsreader multilingual event and time corpus. 2016.
- 70 Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics, 2008.
- 71 Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *CoRR*, abs/1702.05638, 2017.
- 72 Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659, 2013.
- 73 Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227, 2009.
- 74 Roser Saurí and James Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.
- 75 Sanja Stajner and Horacio Saggion. Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *IJCNLP*, pages 374–382, 2013.
- 76 Gabriel Stanovsky, Judith Eckle-Köhler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 352–357, 2017.



- 77 Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4):217–248, 2013.
- 78 Richard Y. Wang and Diane M. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.
- 79 Beverly L Zakaluk and S Jay Samuels. *Readability: Its Past, Present, and Future*. ERIC, 1988.
- 80 M. Potthast, S. Köpsel, B. Stein, and M. Hagen. Clickbait Detection. In Nicola Ferro et al., editors, *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 16)*, volume 9626 of *Lecture Notes in Computer Science*, pages 810–817, Berlin Heidelberg New York, March 2016. Springer.
- 81 S. Heindorf, M. Potthast, B. Stein, and G. Engels. Vandalism Detection in Wikidata. In Snehasis Mukhopadhyay et al., editors, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 16)*, pages 327–336. ACM, October 2016.
- 82 H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, and B. Stein. Building an Argument Search Engine for the Web. In *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 17)*, EMNLP 2017, Copenhagen, September 2017.

## Participants

- Tat-Seng Chua  
National University of  
Singapore, SG
- Nicolas Diaz-Ferreira  
Universität Duisburg-Essen, DE
- Gerald Friedland  
University of California –  
Berkeley, US
- Norbert Fuhr  
Universität Duisburg-Essen, DE
- Anastasia Giachanou  
University of Lugano, CH
- Tatjana Gornostaja  
tilde – Riga, LV
- Gregory Grefenstette  
IHMC – Paris, FR
- Iryna Gurevych  
TU Darmstadt, DE
- Andreas Hanselowski  
TU Darmstadt, DE
- Xiangnan He  
National University of  
Singapore, SG
- Benoit Huet  
EURECOM –  
Sophia Antipolis, FR
- Kalervo Järvelin  
University of Tampere, FI
- Rosie Jones  
Microsoft New England R&D  
Center – Cambridge, US
- Rianne Kaptein  
Crunchr – Amsterdam, NL
- Krister Lindén  
University of Helsinki, FI
- Yiqun Liu  
Tsinghua University –  
Beijing, CN
- Marie-Francine Moens  
KU Leuven, BE
- Josiane Mothe  
University of Toulouse, FR
- Wolfgang Nejdl  
Leibniz Universität  
Hannover, DE
- Jaakko Peltonen  
Aalto University, FI
- Isabella Peters  
ZBW – Dt. Zentralbib.  
Wirtschaftswissenschaften, DE
- Miriam Redi  
NOKIA Bell Labs –  
Cambridge, GB
- Stevan Rudinac  
University of Amsterdam, NL
- Markus Schedl  
Universität Linz, AT
- David Ayman Shamma  
CWI – Amsterdam, NL
- Alan Smeaton  
Dublin City University, IE
- Benno Stein  
Bauhaus-Universität Weimar, DE
- Lexing Xie  
Australian National University –  
Canberra, AU

