Report on the Dagstuhl Perspectives Workshop 13342

ICT Strategies for Bridging Biology and Precision Medicine

Edited by

Jonas S. Almeida[1], Andreas Dress[2], Titus Kühne[3], and Laxmi Parida[4]

1 **University of Alabama at Birmingham, Dept. of Pathology, USA**
jalmeida@uab.edu

2 **Wiss. Zentrum, infinity^3 GmbH, Gütersloh, Germany**
andreas.dress@infinity-3.de

3 **German Heart Institute Berlin, Germany**
titus.kuehne@dhzb.de

4 **IBM T.J. Watson Research Center, Yorktown Heights, NY, USA**
parida@us.ibm.com

Abstract

The systems paradigm of modern medicine presents both, an opportunity and a challenge, for current Information and Communication Technology (ICT). The **opportunity** is to understand the spatio-temporal organisation and dynamics of the human body as an integrated whole, incorporating the biochemical, physiological, and environmental interactions that sustain life. Yet, to accomplish this, one has to meet the **challenge** of integrating, visualising, interpreting, and utilising an unprecedented amount of *in-silico*, *in-vitro* and *in-vivo* data related to healthcare in a systematic, transparent, comprehensible, and reproducible fashion. This challenge is substantially compounded by the critical need to align technical solutions with the increasingly social dimension of modern ICT and the wide range of stakeholders in modern healthcare systems.

Unquestionably, advancing healthcare-related ICT has the potential of fundamentally revolutionising care-delivery systems, affecting all our lives both, personally and – in view of the enormous costs of healthcare systems in modern societies – also financially.

Accordingly, to ponder the options of ICT for delivering the promise of systems approaches to medicine and medical care, medical researchers, physicians, biologists, mathematicians, computer scientists, and information–systems experts from three continents and from both, industry and academia, met in Dagstuhl Castle for a Dagstuhl Perspectives Workshop on *ICT Strategies for Bridging Biology and Medicine* from August 18 to 23, 2013, to thoroughly **discuss** this multidisciplinary topic and to **derive** and **compile** a comprehensive list of pertinent recommendations – rather than just to deliver a set package of sanitised powerpoint presentations on medical ICT. The recommendations in this manifesto reflect points of convergence that emerged during the intense analyses and discussions taking place at the workshop. They also reflect a particular attention given to the identification of challenges for improving the effectiveness of ICT approaches to Precision and Systems Biomedicine.

Seminar 18.–23. August, 2013 – www.dagstuhl.de/13342

1998 ACM Subject Classification I.2.1 Applications and Expert Systems: Medicine and science, H.3.5 Online Information Services: Web-based services, H.4 Information systems applications, J.3 Life and medical sciences: Medical information systems, etc

Keywords and phrases Systems biomedicine, precision medicine, healthcare-related information systems, biomedical workflow engines, medical cloud, patient participation, ICT literacy

Digital Object Identifier 10.4230/DagRep.3.8.87

ICT Strategies for Bridging Biology and Precision Medicine, *Dagstuhl Reports*, Vol. 3, Issue 8, pp. 87–135
Editors: Andreas Dress, Andreas Dress, Titus Kühne, and Laxmi Parida
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Jonas S. Almeida
Andreas Dress
Titus Kühne
Laxmi Parida

Water, water, everywhere, nor any drop to drink. So goes Coleridge's *Rime of the Ancient Mariner*. Until recently, the same went for data: everywhere, but not of much use so far, neither for deriving new medical insights nor for improving medical care.

However, three key developments currently help to overcome this problem: the rapid adoption of electronic medical records [1], the dramatic advances in molecular biology [2], and, just as dramatic, the growing pervasiveness of social computing environments combined with a new attitude towards participatory health management [3, 4, 5]. The result is an exciting medley of initiatives devoted to supporting healthcare related information flow ranging from patient-facing resources such as **PatientsLikeMe** [6] to initiatives such as **MD-Paedigree** [7] (EU's FP7) that provides a physician-centric sort of '**PatientsLikeMine**' analogue addressing treatment choices in paediatrics.

Managing the *creative deconstruction* [8] involved in advancing towards systems medicine requires fundamentally changing the use of ICT in both, healthcare and biomedical research. It requires in particular to take account of the new paradigm of **web-centric computing** which is a basic prerequisite for all these initiatives.

Reflecting these concerns, a Dagstuhl Perspectives Workshop on *ICT Strategies for Bridging Biology and Medicine* was held to discuss a wide range of fundamental and foundational issues. These ranged from architectural considerations to data-access policies including *Open/Linked Data, the Semantic Web, Pervasive Hardware Ecosystems, Medical Clouds, Patient-Participation Frameworks, 'Healthcare* 4.0*', Analytical Tools*, and *Medical Education* Clearly, the required changes can only be achieved by initiatives of a broader scale and scope than what can be accommodated within the existing academic organisations. They need to always involve **all** stakeholders in the healthcare environment. In response to these challenges, the discussions led to the following theses and postulates:

(i) An *open-data policy* for healthcare-related information systems is a fundamental and urgent imperative.
(ii) Following the *business-IT alignment* paradigm [9], healthcare should – on all levels – be supported by secure IT-platforms enabling clinical workflow engines that map healthcare-related processes while integrating pertinent data-analysis, visualisation, and engineering tools.
(iii) Such platforms should also take full advantage of advances provided by *cloud services*, *pervasive computing ecosystems*, and the *semantic web*.
(iv) The *participatory potential* of the Web should be exploited to advance new forms of partnership in the healthcare environment.
(v) The acquisition of *ICT literacy* must become a required part of biomedical education.
(vi) Specifically in Germany, the Bundesnetzagentur should be encouraged to setting up a Working Group *Medizinische Netze* to explore options for a *Medical Cloud* within the German healthcare environment.

References

1 Tracy D. Gunter and Nicolas P. Terry *The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions*. J Med Internet Res 7:1 (2005). DOI:10.2196/jmir.7.1.e3.

2 Susan Desmond-Hellmann et al. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Research Council (US), Committee on A Framework for Developing a New Taxonomy of Disease. The National Academies Press, Washington D.C., USA (2011).

3 Wikipedia. http://en.wikipedia.org/wiki/Social_computing

4 Barbara A. Israel et al. *Community-based participatory research: policy recommendations for promoting a partnership approach in health research*. Education for Health, 14:2 (2001):182–197.

5 Melanie Swan. *Emerging Patient-Driven Health Care Models: An Examination of Health Social Networks, Consumer Personalised Medicine and Quantified Self-Tracking*. Int. J. Environ. Res. Public Health 6 (2009):492–525. DOI:10.3390/ijerph6020492.

6 Ben Heywood et al. http://www.patientslikeme.com/.

7 Bruno Dallapiccola et al. http://www.md-paedigree.eu/. The European Commission.

8 Eric J. Topol. *The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care*. Basic Books, New York, NY, USA (2012).

9 Wim van Grembergen and Steven De Haes. *Enterprise Governance of IT: Achieving Strategic Alignment and Value*. Springer, New York Heidelberg Dordrecht London (2009).

2 Table of Contents

3 Overview of the Program

During the workshop, eight specific topics were addressed in regular sessions, complemented with impromptu evening presentations and discussions. For each session, one of the participants served as coordinator and one as rapporteur while every participant was invited to contribute – with some being particularly encouraged to contribute to specific sessions. The rapports for each were collected at http://bit.ly/dagict and formed the basis for this report.

Here is the list of the eight topics augmented by various relevant subtopics and the respective coordinators, contributors, and rapporteurs:

1. **Information and Communication Technology for Bridging Biology and Precision Medicine – Chances and Challenges**
   - From web appliances to web services
   - From home medicine to cloud-based health information systems
   - From laboratory systems to reference genomic atlases
   - Architectures and APIs for user-governed ICT

   Coordinator: Jonas S. Almeida
   Contributors: Hans Lehrach, Wolfgang Maass, ...
   Rapporteur: Mark Braunstein

2. **Big and Heterogenous Data**
   - What can Big Data tell us?
   - Genomic Medicine
   - The integration and modeling of heterogenous multi-omic and imaging data from disease and its implications for diagnosis and therapy
   - Fusing bioimaging data with clinical and molecular information (for enhancing a systems view of disease)

   Coordinator: Joel Saltz
   Contributors: Klaus Maisinger, Stefan Decker, Scott Kahn, ...
   Rapporteur: Alex Pothen

3. **How can ICT help us learning more about disease mechanisms?**
   - Architectures and APIs for user-governed ICT
   - Medical Clouds as platforms for annotation, exchange, and joint interpretation of healthcare data by medical experts (and patients?)
   - Statistics, machine learning etc.
   - Electronic health records

   Coordinator: Bernhard Balkenhol
   Contributors: Eric Neumann, Eric Gordon Prud'hommeaux, ...
   Rapporteur: David Gilbert

4. **Virtualisation, Computation, and AI in Medicine**
   - The status of structured (pathway, model etc) databases
   - The virtual oncology, diabetes, ... patient in medical practice
   - Mechanistic models
   - The vision of ITFoM

- How can we extract information from the scientific literature as well as from 'low-grade information' in the web (text mining, the semantic web in healthcare, search strategies in semantic webs)?
- Virtualisation in drug development

Coordinator: Hans Lehrach
Contributors: Laxmi Parida, Pietro Lio', Joel Saltz, ...
Rapporteur: Andrea Splendiani

5. **Molecular Systems Medicine and ICT I**
   - Assessing emergent properties of chronic diseases and disease mechanisms:
     – The dynamics of disease progression, implications for disease mechanisms
     – Parallel toponome decoding and genome sequencing
     – Cancer and the immune system
   - Family genome and tumor genome sequencing – the use of Next Generation Sequencing and its implications for therapy and disease stratification

   Coordinator: Peter Walden
   Contributors: Walter Schubert, Robert Burk, Markus Löffler, ...
   Rapporteur: Eric Gordon Prud'hommeaux

6. **Molecular Systems Medicine and ICT II**
   - A systems approach to diagnostics (including the use of proteins, mRNAs and miRNAs from blood and from tissue)
   - Biomedical engineering and systems-optimisation strategies in medical care
   - Assays for wellness
   - Precision medicine and guidelines for evidence-based medicine: complementary or incompatible?

   Coordinator: Ina Koch
   Contributors: Anja Hennemuth, Helena F. Deus, Susana Vinga, ...
   Rapporteur: Laxmi Parida

7. **The Stratification of Disease into Discrete Subtypes and its Implications for Science, 'Medical Ontology', Diagnosis, and Therapy**

   Coordinator: Markus Löffler
   Contributors: Mark Braunstein, Eric Prud'hommeaux, Peter Walden, ...
   Rapporteur: Susana Vinga

8. **Does the Potential of ICT in Medical Care Require New Forms and Levels of Medical Training, Medical Data Collection, Storage, and Reproducibility, Clinical Logistics, Clinical Trials, and Patient Participation?**

   Coordinator: Susana Vinga
   Contributors: David Gilbert, Jochen Dreß, ...
   Rapporteur: Ina Koch

4 The Participants' Introductions

The meeting began with the (then present) participants introducing themselves to each other:

Joel Saltz: Works at Emory, is moving to start a Biomedical Informatics Department at Stony Brook's Biotech Dept., and works also as a consultant for pharmaceutical companies.

Susana Vinga: Researcher at the TU Lisbon, works on modelling and control of dynamical systems relating to biological processes, runs projects on pharmacogenomics, computational biology, develops supporting infrastructure for the Biotech Deptartment.

Jonas Almeida: Started out as a plant biologist and then went to chemical engineering, machine learning, biostatistics, and bioinformatics, now applying machine learning to clinical informatics in the clinical context.

Wolfgang Maass: Chair for Information and Service Systems, Faculty of Business Administration, Saarland U, Germany. Developing and employing machine learning, cloud technology, Big-Data management strategies etc., works on embedding information and service systems into real-world systems.

Pietro Lio': Works at Cambridge U (UK) on genetics and complex systems, integrating bioinformatics and medical research, aiming to understand *comorbidities* and to link metabolic pathways with disease.

Klaus Maisinger: Works at Illumina, UK. Studied theoretical physics. As the world is now producing large amounts of sequence data (e.g., the National Health Service is funding full-genome sequencing of 100k patients), he wants to see what to do with these data and how we can you use them to affect clinical practice.

Bernhard Balkenhol: Academic Work on complexity problems in 'classical' information theory, then worked for Telefónica Deutschland, now builds SOAs that allow companies and their employees to deal with each other and their customers in a secure network, wants to investigate how his ICT procedures can be applied in the context of medicine.

Alex Pothen: Works at Purdue U on combinatorial and graph algorithms, sparse matrices, and optimisation models. In bioinformatics, he looked for disease biomarkers from mass spectra and for functional modules in biological networks. Recently, he developed algorithms for registering and analysing cell populations from flow-cytometry data and for dynamically classifying samples in high-dimensional high-throughput data in order to investigate immune-system responses by measuring protein concentrations on cell surfaces. He is also interested in high-performance computing and Big-Data management.

Mark Braunstein: Studied medicine and now works as a Professor of Practice and Associate Director at the Health Systems Institute of Georgia Tech where he is involved in fostering research and community outreach aimed at the wider and deeper adoption of health information technology to improve the quality and efficiency of care delivery. He developed Ambulatory Electronic Medical Records using the Massachusetts General Hospital Utility Multi-Programming System, started and run health IT over the next 30 years. He is interested in healthcare-process modelling, expects that some Health Information Systems are terrible because they don't involve process modelling, and wants to learn how such systems look around the world.

Jonas: Is there a *natural home* for such process-oriented ICT work?

Mark: Interactive computing is very applicable in this context. Performance folks are all talking about clinical applications. We want to use healthcare-specific modelling language to optimise the surgical units in a new surgical floor some hospital was planning to build.

Ina Koch: Works as the Head of the Bioinformatics Department at the Institute of Computer Science at Frankfurt U, started out as a theoretical chemist, yet already her thesis dealt with a topic from bioinformatics, currently interested in systems biology and medicine, cooperating e.g. with physicians on Hodgkins Disease: We need to be able to deal with raw data like images and mass spectrometry data. She also uses – and recommends to use – Petri Nets for systems for which one does not have sufficiently precise quantitative kinetic data.

David Gilbert: Works as the Head of the *School of Information Systems, Computing and Mathematics* at Brunel University in London, UK, and as co-director of its *Centre for Systems and Synthetic Biology.* Studied Computer Science and holds a PhD from Imperial College where he worked on computational logic and concurrent systems. Current topics include:

- In Computational Biology: *Protein structure at fold level.*
- In Systems Biology: *Static models of networks – metabolic pathways, dynamic aspects of signaling pathways (in e.g. cancer models), modelling qualitative and quantitative aspects of multidimensional and multiscale systems using Petri nets, ODEs, and stochastic systems, employing also model-checking procedures.*
- In Synthetic Biology: *Developing a common language and set of tools for synthetic biology research and, more specifically, modelling and programming cell-cell interaction as well as contributing to Malaria research.*

Motivated by the desire to get even stronger involved with biomedical application, he is also working on

- Freidreich's ataxia, integrating patient clinical data with biochemical data (human and animal models),
- precision-medicine modelling of interventions in terms of diet and genetic alterations,
- and on creating patient- and group-specific models of biochemical etc. processes for better predictions.

Stephan Decker: Director of the *Digital Enterprise Research Institute* at the National University of Ireland, Galway, devoted to semantic-web research in Ireland, turning the web from *a web of documents* to *a web of knowledge.* Tim Berners-Lee's insight was that it's not so much about technology, it's about how to get the rest of the planet to accept and use the technology and how to create a global agreement to amplify our own intelligence by taking a 'social-technical' approach. To learn how to create infrastructure that enables teams to share knowledge, biology is an ideal topic because it is complex and fragmented, and requires collaboration between groups.

Eric Neumann: Vice President responsible for 'Knowledge Informatics' at *Foundation Medicine*, a company that is commercialising cancer genome sequencing, matching results with targeted therapies and clinical trials. Originally a neurogeneticist, he was unimpressed with the single genome sequencing in the 1990's since it was over-sold to "cure all diseases", and decided to work at BBN Technologies on biological modeling platforms and synthetic biology for education goals. He saw the need for scientists to communicate more effectively about they findings, but saw the computer tools lagging behind. With the help of linked data, he is expecting research data will become a commodity for others, and that big data will increase in value once semantics has been included.

Peter Walden: Originally a biochemist (with a PhD in T-cell regulation), he works now at the Charité in Berlin on the cooperation and the interference of the human immune system and human disease, doing research on clinical cases in a clinical environment while trying to bridge research and clinical practice, yet respecting that physicians approach problems not like researchers and biologists, and that e.g. much mice knowledge does NOT apply to humans.

Titus Kühne: Works as a physician at the German Heart Institute in Berlin, studies causal relationships connecting systems biology and evidence-based medical knowledge, wants to know what he does not need to know.

Anja Hennemuth Works at the Institute for Medical Image Computing in Bremen on tissue/organ models based on image data for therapy planning, support for clinical studies, and auto-processing of image data, integrating this with management-system studies, and is therefore interested in how to manage different types of data and how to integrate image data in more complex models.

Jochen Dreß: Trained as physician (orthopaedics), moved to systems-biology modelling with Markus Löffler for 2 years, spent 8 years at a Contract Research Organisation supporting clinical trials, then got more involved in IT and software development and moved to the Cologne Center for Clinical Trials working on data-quality assurance and IT governance moving clinical research closer to care, works now at the German Institute of Medical Documentation building a new department providing access to public-health insurance data for researchers.

Helena (Lena) Deus: Started out as a marine biologist studying amino acids in sea urchins, was hired by Jonas to work on a database for methicillin-resistant Staphylococcus aureus bacteria, then went to the MD Anderson Cancer Center where she worked with oncologists. As IT folks weren't paying attention to their needs and their excel-file based IT infrastructure didn't let them exchange biomarker information, this created an opportunity for PhD work. So she built systems to exchange such information for decision support. Before SemWeb (an open-source library created by the semantic-web community) became available, data capture was based on tables. But the clinical space needed more variability. So, she went to the Digital Enterprise Research Institute (DERI) to work with Stephan on 'Linked Data' – a concept which is not just about building systems, but instead about viewing the whole web as a database: imagine asking a question and getting back answers that were as precise as the question you asked. While 'all models are wrong, but some are useful', SemWeb data enable injecting more knowledge to make models more useful. One of Foundation Medicine's goal is to support physicians wanting to help patients who are too weak for standard chemotherapy.

Mark: A classic problem is that the folks developing the IT systems aren't (currently) based on our conversations here.

Lena: Yes, that is the 'lost in translation' problem. Physicians issue requirements that the IT folks don't understand and ...

Mark: Physicians at the largest children's hospitals still grab data by hand to put them into excel.

Walter Schubert: Works as a physician, currently being the head of the Molecular Pattern Recognition Research (MPRR) group at the Medical Faculty of the Otto-von-Guericke-University, Magdeburg, teaching histology and mathematics-based data-analysis techniques, investigates proteins as building blocks of larger super-molecular complexes, founded 'Molecular Systems Histology' with a two-page paper describing how one can analyse whole protein systems without destroying the tissue, discovered the physical displacement of muscle tissue by (not attacking) T cells, and wants to learn more about Big-Data management. As he always deemed mathematics to be very important for his studies, he has a long-standing cooperating with Andreas.

Jonas: Some say Big Data are a waste of time. Biostatisticians are unwilling to go to the data.

Joel: Lots of areas are labeled 'Big Data' with a hope of funding. I've worked on spatial/temporal axes in GIS and was asked in this context 'Which methods scale the process of acquiring and assembling multi-resolution sensor data?'. My thesis: 'What is needed there, largely intersects with needs for assembling data for comorbidity discovery'.

Lena: The four Vs determining the *value* of Big Data are *Volume, Variety, Velocity*, and *Veracity*, see e.g. [1, 2, 3].

Joel: Those are necessary conditions, but the subclassifications will be important. You'll be able to build systems that apply to many distinct systems.

Walter: Probably invoking an identity crisis.

Zeng Zhenbing: Works in Shanghai, got his PhD in mathematics in Bielefeld, afterwords worked in the Computer Science Institute of the Chinese Academy of Sciences in Chengdu for 10 years, and then at the Institute for Software Engineering, East China Normal University, Shanghai, for another 10. Research topic: writing algorithms for discovering and proving mathematical theorems automatically. Started with bioinformatics in 2005 when Andreas came to Shanghai to establish the *CAS-MPG Partner Institute for Computational Biology* there. 'For the past years, I have learned a lot of biology, but not yet enough to do research – while, amongst my colleagues in Shanghai, I'm the comparative biology expert.'

Eric Prud'hommeaux: Works as 'data plumber' at the MIT, USA.

Andreas Dress: Studied Mathematics, after PhD and Habilitation in Kiel, he went for two formative years to the Princeton IAS, became founding member of the 'Fakultät für Mathematik' at Bielefeld U in 1969, and began to cooperate closely with Manfred Eigen on mathematical problems relating to phylogenetic reconstruction in 1976.

Laxmi Parida: Works at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA in computational biology and leads a Watson group on'Computational Genomics', with a trend towards solving problems through crowd-sourcing, also working on Pattern Discovery, Design and Analysis of Algorithms, Population Genomics, Next Generation Sequencing (NGS) Analysis, and Bioinformatics.

Robert Burk: Physician, works as a neurosurgeon at Johns Hopkins U, doing also research on the neural system and on the human papilloma virus and cervical cancer using genomic and epidemiological information, the human microbiome, the genetics of prostate cancer, and excessive sweating. Wants to learn more computer science and likes to reflect the practice of medicine.

We were later joined by

Scott Kahn: who earned his Ph.D. in Theoretical Organic Chemistry at UC Irvine, performed post-doctoral research at Cambridge University in the UK, and now serves as Vice President and Chief Information Officer of Illumina – San Diego, USA.

Hans Lehrach: who obtained his Ph.D. at the Max Planck Institute (MPI) for Biophysical Chemistry in Göttingen, worked as a postdoctoral fellow at Harvard University, was one of the first who initiated the human genome project, and now works as Director at the MPI for Molecular Genetics in Berlin with a focus on genetics, genomics, and systems biology.

Markus Löffler: who studied physics and medicine and now works as the director of the Institute for Medical Informatics, Statistics und Epidemiology (IMISE), U Leipzig, with

Figure 1 The *word cloud* for the opening session, reflecting the central focus on data, systems, models, and clinical work.

main emphasis on bioinformatics and computational biology, systems biology modelling, and cancer research.

Andrea Splendiani: who works as a computer scientist at the DERI with interests in the semantic web, open data, ontologies, taxonomies, bioinformatics, and systems biology, focusing on the interface between knowledge representation, user interaction, and analytics, mostly in a biomedical domain.

Marc Van Regenmortel: who obtained his Ph.D. in virology from the Medical School at Cape Town U and now works as emeritus Director at the School of Biotechnology, CNRS, Strasbourg U, on molecular recognition and the immune system.

5 ICT for Bridging Biology and Precision Medicine – Chances and Challenges

(The Discussions on Monday Morning)

Coordinator: Jonas S. Almeida
Contributors: Hans Lehrach, Wolfgang Maass, . . .
Rapporteur: Mark Braunstein

(i) Jonas starts moderating the session

Jonas: There is an issue of definitions: we don't know what this workshop is about, what is Big Data? . . .

People have identified data, not knowledge, as the start of the exercise. Apparently, there's a sharp logistical break between the *Resource Description Framework* (RDF) and the *Web Ontology Language* (OWL). E.g., the term 'same As' in OWL has been considered harmful to provenance [4].

Eric N: So, what's data vs. info?

Jonas: The medical side is ready to listen to you: this year has seen the first board exams for informatics relating to the broad area encompassed by data science, i.e. bioinformatics, computational biology, biomedical informatics, biostatistics, information science, quantitative biology, etc.; see also e.g. [5] for the Proceedings of a meeting on Genome-Era Pathology, Precision Diagnostics, and Preemptive Care.

Another question: How do I deal with governance of data to which I am entrusted or not entrusted?

A short demo shows how dragging data into system loads a data-specific interface and how, clicking on a cell, gives you tools to calculate e.g. volume, distance, etc., while `<view-source>` reveals the injected code.

Jonas: Perhaps we need DSL to convert to this code? I've done some analysis. How do I share that data? I've asked as pathologist and statistician. Now some professional dishonesty: I had an algorithm from Georgia Tech to do numerical processing of differential equations. Half-life of a bioinformatics app is 18 months. Can deliver matlab (like other systems) runtime server for free. But the runtime lengths change and folks need to hunt them down on the web. The MatLAB 2006 RT is the biggest download in our directory. Things like Angry Birds have changed the expectations. Now apps don't know about each other.

Two thirds of the data in Electronic Medical Records (EMRs) comes from Pathdata[1], but only about 5% of Pathdata make it into the EMR.

Now, when will we produce patient-facing interfaces allowing patients to log into the US Department of Health & Human Services, to create a family history, and to import data from social apps? Here, helpful advise can be found in [6], an important paper on fusion tables.

Mark: Soon, 5% of patients may have used already Data Share. The HeathVault, a web-based platform from Microsoft to store and maintain health and fitness information, is a good way to do this. The goal of patients integrating their clinical data makes sense, but it's been hard for my grad studs to even create a Google Health account.

Robert: Once you get invalid data, you've corrupted your value.

Mark: That's assuming the EMR data were sound in the first place.

Jonas: Often the person who understands the data about a patient is a relative. Using the governance model of e.g. Microsoft's HealthVault often is a disservice.

Eric N: The notion of a 'best copy' has been hurting us. Data is NOT a flat file though we often think of it that way. Data have multiple dimensions that need to be captured.

David: Patients look at doctors as well.

Jonas: The patient should be the ultimate governor of her/his own data.

Next, a '23andme' TV ad[2] was shown.

Jonas: The ability of authentication systems to delegate enables us to attach governance to data without moving the data. We can't e.g. guarantee patient anonymity.

Robert: People aren't set up to deal with all this information. In the US, it's very hard to get general care because of the changing systems.

[1] Cf. http://www.pds-europe.com/products/pathdata-system

[2] from http://www.youtube.com/watch?v=ToloqU6fCjw&feature=youtu.be

Joel: Some patients will want to know lots, some something, and some nothing. But many folks will need interpretation, transmitting granular and noisy data to patient-facing summaries.

Jonas: I agree, but also to physician-facing summaries because physicians need to see the same thing. I saw 'non-rectangular' patient data magically fixed by IMPUTE[3]. The web has the right characteristics for breaking these black boxes.

Walter: Address discrimination against elderly. Elders may live 80km away from where doctors are, they are excluded from healthcare – no internet, no phones.

Mark: I don't buy most patients getting sequenced and managing their data. Most patients (and physicians) don't know what to do with it. We need more input from the patients. In our breast-cancer project, the patients get a 150 USD android with all the apps they need for providing feedback on whether e.g. the pain medications are sufficient. Regarding the elderly, we try to see patients on a regular basis for managing chronic disease, but there is some hope in home devices.

Jonas: Who behind the system is making the judgment and how will this change decisions in disease management? We did a test with 50 people, editing various rules. The outcome was that knowledge bases have to evolve. The requirements will change with time, law, and person. We left it open and folks said 'this is an invitation to chaos'.

(ii) Now, Wolfgang continued moderating the session

Wolfgang: Modes of identification are also at stake. You can give any kind of object an ID – semantic web people would call it an URL. The miniaturisation of hardware allows not only to take pills that are computers, also the price has fallen exponentially. Capturing ids from the real world went from labels to 2d barcode and now to radio-frequency identifiers (RFID). One of the goals: new information projected into an operating room. Other applications: Intelligent bathrooms that interact with you.
Q1. How to design these types of objects?
Q2. What are the technical architectures that enable it?

Eric P: What are the things that a physician would need to see during a surgery?

Wolfgang: Overall patient information from many sources.

Eric P: Such an information ubiquity is probably meant to inject enough context into the environment so that people can work at a higher level. Could be tested in command centers for emergency response or for search & rescue. Inject code into the sensors to make them smarter. This was used in DOLCE[4], but failed. It is now much more implicit in JavaScript than it was in DOLCE. Here, you can create custom objects like self-controlled plants with personalised irrigation. Now, replace 'plant' with 'patient' (though this statement, taken out of context, may not sound so good). Once planted, sensors communicate with edge nodes by representing data using semantic frameworks that process the data using object-oriented functional computing (as offered by JavaScript); then actuators act based on that information. With a million plants, data streams need to be considered. Computation gets pushed into a master node.

[3] A program for estimating ('imputing') unobserved genotypes in SNP association studies.

[4] Descriptive Ontology for Linguistic and Cognitive Engineering, c.f. http:/en.wikipedia.org/wiki/Upper_ontology_(computer_science)#DOLCE_and_DnS.

Mark: In the US, we have 8M patients with chronic disease who account for 50% of healthcare costs. Putting intelligence into the nodes makes sense. There's a company that took technology out of Argon labs and showed that they can build a model of patients to predict which patients are heading towards problems.

Wolfgang: So, we will collect a lot of data. Why to do it with plants: it is difficult to get data from healthcare. And plants allow testing the architecture.

(iii) Now, Hans continues moderating the session and refers to the ITFoM project [7]

Mark: What's going into the model today?

Hans: We're hoping to sequence across a pathology slide. If you put in the oncogenic RAS mutation, you already get the right results. If you put in the genome, your results get better. By definition, we're sequencing the dominant cell. Ideally you'd want to subdivide the tumor into cell types which can react differently.

Jonas: You enter only first-order kinetics?

Hans: In principle, you could enter anything you know about.

Mark: What's the cost?

Hans: 2 times the transcriptome, about 5K.

Mark: For the patient whose metastasis wouldn't be addressed by the drugs, what was the outcome?

Hans: There was another drug that should have addressed that metastasis which came out one month before she died.

6 Big and Heterogeneous Data

(The Discussions on Monday Afternoon)

Coordinator: Joel Saltz
Contributors: Klaus Maisinger, Stefan Decker, Scott Kahn, ...
Rapporteur: Alex Pothen

(i) Joel starts moderating the session and states:

In Data Science dealing with Big Data and in Biomedical Informatics, we need integrative analysis across multiple time scales, modalities, ...
We need clinical pathologists and computer scientists working together for the analysis of images of cancer pathologies.
We have to cope with

- multiple images across multiple modalities,
- exabytes of data from mice,
- the intersection of omics data, radiology, etc.,
- the commoditisation of imaging and sequencing data,
- in vivo high resolution pathology,
- 3D microscopy imaging,
- animal models,
- reconstruction of cellular structures etc.

Yet, in many application areas of multiscale imaging, there is a common core set of tools that could be used.
Extracting features and identifying clusters, three morphological groups have been identified by clustering data from 200 million nuclei provided by *The Cancer Genome Atlas* (TCGA). The big question is how will we cure cancer, how and when to administer stem cells, how to treat other diseases? Using *Random Forests* to construct an ensemble of *decision trees* for identifying the top ten percentile of high risk patients for readmission (in the next 30 days?), a clustering analysis of data regarding Emory's patients and 180 other hospitals from a University health consortium helped to better learn how to predict readmissions and survival, and how some treatment was working at a given moment. The data included clinical phenotypes (heart disease, ...), geographic distribution, and clinical outcomes. Due to what individual medical coders do, the data were also quite noisy.
There were a few terabytes of data including critical care data from sensors.
There is software now for predictive modelling to support healthcare-data analytics and spatiotemporal data management and analyses.
There is also experimental work employing in vivo microscopy at the MD Anderson Cancer Center on the horizon correlating imaging phenotypes with genomic signatures.

Jonas: It should be worthwhile to pool the data even when there are differences among the treatments. We should design a good physical health environment including smart bed towers able to track locations, analyze streaming physiological data, obtain info on smart phones, tablets, etc.
Where do you put genome information? Better put this in the healthcare data analytics rather than in the physical environment. What about other organisations that track people and obtain streams of sensor data that are reliably and flexibly interconnected?

Andreas: However, whatever hospital you go to today, you'd better have a super immune system!

(ii) After coffee break, Klaus Maisinger and Scott Kahn continued

Klaus: The main challenge of experimental data today is there volume. All data seem to be stored in central repositories.
Question: How will this model evolve? Most probably to distributed storage: keeping track of where the data come from and where they are.
What is the form of data? Many, many files, or linked data?

Jonas: Distributed storage using URLs of the data sources is sub-optimal creating a hodgepodge involving the organisations that pay for data.
TCGA data are not easy to use. The Memorial Sloan-Kettering Cancer Center makes subsets of data accessible. In practice, data is stashed where it is stashed.
And, to report on expensive experiments, caGrid uses Version 4.03 of the Globus Toolkit, produced by the Globus Alliance. It was built on presuming that its data federation would manage data security. It sank on its own weight having just too many features.

Robert: According to Wikipedia, the *Kaiser Permanente Center for Health Research* has 8.9 million health plan members, 167,300 employees, 14,600 physicians, 37 medical centers, and 611 medical offices. An important piece of its overall data-warehouse capacity is its Virtual Data Warehouse (VDW). The VDW is a set of standardised data and methods that enable researchers to produce comparable data, e.g. microbiome genetic data, across collaborating sites for proposing and conducting research.

The VDW is ‘virtual’ in the sense that the raw (‘real’) data remain at the local sites, meaning the VDW is not a multi-site physical database at a centralised data-coordinating center. It might also be called a distributed data warehouse.
What we are doing: coming up with a 10–12 page Dagstuhl Manifesto that could be presented to Government funding agencies, a Manifesto about data, medicine, and ICT. We should think about suggestions for using ICT properly beyond current Health Information Systems. For instance, there is no effort yet to standardise data collections.

Scott: There is a benefit to aggregation. We can store raw data. Should we propose a hierarchy of data products similar to a satellite bus[5]? By design, they provide

- a modular structure supporting a wide range of payload and launch vehicle interfaces,
- a redundancy architecture that can be tailored to meet mission needs,
- and their propulsion can be sized (or eliminated) to meet mission objectives.

Note also that there is the *Electronic Medical Records and Genomics Network* (eMERGE), a national consortium organised by the National Human Genome Research Institute with the task of combining clinical data with genomics.
To achieve interoperability, a higher-level abstraction standard at each institution should get us there about 95% of the time. But we have to look more at the interface. To make healthcare providers better, one needs data and interfaces with tools to use them.

Jonas: I am a bit surprised the linked-data folks in the audience are so quiet.

Scott: For those who are intrigued about this business of sequencing formats embedding the metadata within the file itself: note that FASTQ files from the NCBI/EBI Sequence Read Archive often include a description like e.g.

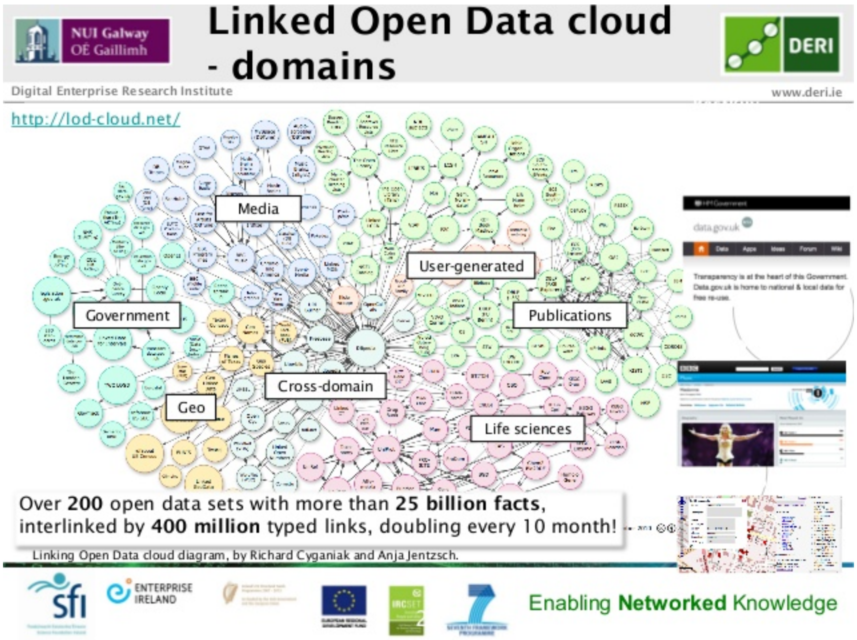
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

In this example taken from http://en.wikipedia.org/wiki/FASTQ_format, there is an NCBI-assigned identifier[6], and the description holds the original identifier from Solexa/Illumina (as described above) plus the read length. The VCF[7] specification can be downloaded from https://github.com/amarcket/vcf_spec/blob/master/VCFspec_4.2.pdf?raw=true. Yet, there is not yet a clear-cut ‘genomic’ definition of prostate cancer etc. versus breast cancer.

Jonas: Are linked-data models used at Illumina? Which architectures are used for raw big data, just collections of files in some VCF format?.

Scott: Illumina uses data for quality-control purposes as well as to search for allelic variations. A clear distinction is that one has raw data, and data that one needs to do analytics about. Also, confirming a result is different from a de novo result.
I am not sure that we deal with Big Data. We have no need to use Hadoop on a 1000 genomes. We only look at a subset of the data. So, tools for smaller data suffice. Genomic data is clean and consistent compared with clinical data.
Also, setting up a ‘Data Federation’ is not ‘One ring to rule them all’. One needs a strategy to ‘federate’ data. Some data may usefully put into data silos so that comparisons can be done across distributed data sets.

[5] Cf. http://en.wikipedia.org/wiki/Satellite_bus
[6] NBCI = National Center for Biotechnology Information.
[7] VCF = Variant Call Format.



Linked Open Data cloud - domains

Digital Enterprise Research Institute

www.deri.ie

http://lod-cloud.net/

Over **200** open data sets with more than **25 billion facts**, interlinked by **400 million** typed links, doubling every 10 month!

Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch.

Enabling **Networked** Knowledge

Figure 2 Linked open data cloud domains: a representation of over 200 open data sets.

Jonas: We should start marking data that is least valuable. A good indexing and annotation system could help solve this problem. E.g., we need to distinguish between missing variant transcriptome data, and the absence of variant data.
There is also data heterogeneity as data come from different sequencing methods.
Semantic data is tiny; e.g., the liquid vs solid tumor distinction is based just on where in the body they grow.
Functional genomics is also not yet as applicable as one might hope. E.g., for cutaneous lymphoma, there is no signature mutation present in all tumors. Instead, we have about 25 or so different kinds of such tumors. And all of them are different in different individuals.
Precision medicine and international genome efforts depend on interoperability of data (such as the TCGA data). The report by the National Academy of Sciences entitled *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* [8] asks to reclassify diseases based on molecular mechanisms that underlie them.
All that is a long way to go.

Robert: Medicine is going through changes, its work is not just disease-based, but also conditions-based. Precision might be unhelpful, and also expensive[8]. And Leroy Hood's P4-Medicine elaboration of precision medicine [9, 10, 11] wants medicine to become predictive, preventive, personalised, and participatory.

(iii) Finally, Stefan Decker presented his views on *knowledge pipelines*

He considers the semantic web rather a *social* than a *technical* initiative.
There are two ingredients:

[8] Four talks on practical efforts in delivering precision medicine can be found at http://www3.ccc.uab.edu/index.php?option=com_content&view=article&id=591&Itemid=300.

Figure 3 The word cloud relating to Stefan Decker's presentation and the ensuing discussions.

1. RDF, the Resource Description Framework, offering a graph-based data representation: URLs are objects, interlink information defines the relationships between objects.
2. Vocabularies (Ontologies): They provide *shared understanding* of a domain, organise knowledge in an *machine-comprehensive* way, and give an exploitable *meaning to the data.*

The linked open data cloud brings together information in various databases as described e.g. in http://lod-cloud.net referring to 200 open data sets (see Figure 2 above).
One goal is to almost 'automatically' transform data into knowledge: There are patients with breast cancer who respond well to the chemotherapy drug *Cisplatin* while others do not. Now comes a third patient: what should the doctor do? The patients' responses depend on their genome and its epigenetic regulation. In fact, there are microRNAs silencing tumor-suppressor genes that can be used to distinguish patients who respond well from those who do not. This could be done in a few minutes instead of weeks.

Stefan also presented a figure of Knowledge Pipelines and the three steps of a Workflow:
Step 1: Load publications into text index to create an overview.
Step 2: Load clinical trial data into RDF database and test queries.
Step 3 : Create a filter on publications for 'neoplasm' and load results into RDF database.

Another useful tool is 'Bio2RDF', a biological database using semantic-web technologies to provide interlinked life-science data supplying a data interface and a query interface designed to help biologists to focus on their discovery process rather than the IT effort.

Finally, Stefan presented a CloudSpace demo.

Eric N: When you make a Solr index[9], you are creating triples?

Stephan: Yes, and then feeding them into the next step of the pipeline (hence 'pipeline').

Jonas: I see how you use SPARQL[10], but will you allow code injection?

Stephan: Not now. Let's work on this together.

Now, Stefan presented a second CloudSpace demo.

[9] Solr is an open-source enterprise search platform for distributed search and 'index replication', a method for promoting data availability for searching while preventing data loss, cf. http://en.wikipedia.org/wiki/Apache_Solr.

[10] SPARQL is a Protocol and RDF Query Language, see e.g. http://en.wikipedia.org/wiki/SPARQL.

Figure 4 The *word cloud* for the session on big and heterogeneous data.

Andreas: Is that a real implementation or a demo?

Stephan: It is based on a script that consumed data from the Linked Clinical Trials data space at http://linkedct.org.

Andreas: You said ‘when you take the results’, how do you do that?

Stephan: We don’t have anything that automagically transforms data, but it sets up, for example, your hadoop structure and stuff like that.

Mark: What’s involved in getting the source into RDF?

Stephan: Identify the objects, e.g. – for clinical trials – patients, drugs, ... Once your data are stored in relational form, getting them into RDF format is standard.

Mark: So, if I had thousands of clinical records in SQL[11], ‘RDF-fying’ them would be serious work?

Eric P: You’d probably leave it in situ and create an R2RML[12] wrapper to provide a SPARQL view.

Wolfgang: How did you solve the vocabulary mapping problem?

Stephan: That came from composing the SPARQL query.

Andreas: The query was in terms of MESH terms in both data sources, so the mapping was ‘already done’.

7 How can ICT help us learning more about disease mechanisms?

(The Discussions on Tuesday Morning)

The topics of this session were

- Architectures and APIs for user-governed ICT,

[11] SQL = Structured Query Language, cf. http://en.wikipedia.org/wiki/SQL
[12] R2RML = Relational Database to RDF Mapping Language, cf. http://www.w3.org/TR/r2rml/

- Medical Clouds as platforms for annotation, exchange, and joint interpretation of healthcare data by medical experts (and patients?),
- Statistics, machine learning, etc.,
- Electronic Health Records.

Coordinator: Bernhard Balkenhol
Contributors: Eric Neumann, Eric Gordon Prud'hommeaux, . . .
Rapporteur: David Gilbert

The discussion started with some general remarks regarding the possible outcome of the conference.

Andreas: Could it make sense to use the Manifesto to push for the installation of a medical cloud in Germany?

Titus: What should our discussion focus on? There are illustrative cases of ICT use regarding oncology as well as neuro- and cardiovascular diseases. And there are various possible directions for ICT to push:
(i) mechanistic modelling (incl. visualisation) to improve the understanding of biomedical insights and their medical implications,
(ii) documenting patient outcome,
(iii) making data & tools accessible to stakeholders.

Next, Joel proposesd three topics on the blackboard:
1. A wider view of health (P4?),
2. Data first vs mechanism/understanding,
3. Data access / sharing.

There ensued a lively discussion on this during which the following statements were made:

A better understanding of disease is an essential component of patient care.

An informatics core common to both, (2) and (3). Yet, how do we make the data available to the different communities?

Grant issues lead to a different focus in different communities.

A better understanding of disease also leverages new science / materials. These have not yet been integrated into practice of medicine. We should focus on this gap.

There is also a gap between disease mechanisms & patients. One needs to make data available. Yet, are we the best consortium to do this?

Informatics, computation, and genomics are driving forces. They will be taken up by medicine in the future. What can we do to improve medicine?

We need to know where we are going. What will medicine look like 10 years in the future.

Titus: Future of medicine: prediction and prevention will be improved.
We need to better understand *pathophysiology* because there are so many diseases that are poorly understood.
Currently, there is interesting work on basic research with a pharmaceutical company at my lab in this direction.
We try out possible therapies on 'model patients'.
There are also fundamental changes in the market place – more smaller pharmaceutical companies. Is there also a change in the economics of the marketplace?

Here are some useful web sites. Regarding

- the current decline in the pharmaceutical industry: see http://www.nature.com/nrd/journal/v11/n3/full/nrd3681.html,
- personalised medicine – its future impact from a pharma perspective: see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2918032/,
- economics of personalised healthcare and prevention: see http://commonfund.nih.gov/pdf/[13].

Possible conclusion: personalised and precision medicine may be the key to economically develop treatments for diseases for which this is currently not feasible.
Pharmas buy up products from small companies, they have stopped doing basic research.

Eric P: We will have a discussion about the evolution of data standards this evening.

Finally, a video from the EU Flagship Proposal *IT Future of medicine* (ITfoM) was shown.

Tuesday Late Morning discussion:

Mark: We are a bit focussed on e.g. cancer and personalised medicine. But in the developed world, there are problems of ageing etc. which have to be taken into account by policy makers when planning health policies. If we give policy makers the impression that we are just focussed on cancer and personalised medicine, then policy makers will not really appreciate this.
To mention one example: Using modified Markov decision processes, a student made a model of clinical decision-making processes using different modes of care, and compared this with solutions provided by computation. The computer did better than standard models. Modified models have memory and, hence, may make better decisions about treatment.
So, how to design an optimal (real) clinic. We showed that for high-risk patients, the risk can be reduced substantially. We also gave Emory U a different model that could be offered to all 40,000 employees, with a 7% return on investment if the programme would run according to assumptions.

Eric: Are you certain that the assumptions were correct?

Mark: The surgical floor in a hospital can have far fewer beds if configured properly. Yet for stroke treatment, technology has increased costs. So, we need to be careful when suggesting that technology should be used.

David: There are 2 approaches in our discussion:

- data driven [genomics, proteomics etc] which can support basic scientific understanding of disease as well as best-treatment choices, and
- process driven – e.g. healthcare modelling.

Hans: ITfoM is not only just about modelling, but also about making data available. It provides ‘continuous’ modelling, from the genome of patient to the final outcome and the costs of healthcare.

Mark: So far, the conversation at this meeting is narrower than I would wish.

Others:

- If we make the focus too broad, then we will not make progress.
- Should we sidestep electronic medical records (EMRs)?
- *Enterprise EMRs* will become data repositories.

[13] and http://commonfund.nih.gov/pdf/HealthEconomics_Personalized_Health_Care_summary_REV_9-28-12.pdf for a related summary of health economics.

Figure 5 The word cloud for the Tuesday Morning session.

- We should talk to them, try to liberate healthcare data.
- How to move forward?

Robert: Breakout / work groups?

Andreas: Discuss after lunch.

Bernhard: How should healthcare-related ICT be designed? If you start a project, say, designing a medical cloud, there are several very simple questions to be asked:

- Why do we start the project: Because we would like to change something.
- What would we like to change: The options for hospitals to interact.
- How do we want to do this: By improving communication.
- And so on . . .

Data are restricted in use: local, country, international communities. There are resulting restrictions when combining data from different units, establishments, (parts of) organisations, or . . . .
Functions and/or roles with different security levels need to be conceived, designed, and codified, and rules defining access, activity options, and responsibilities need to be assigned and implemented.
What is the best design of the network topology – should it have a central hub?
How can a new entity join?
Who is responsible to enable access to the network?
Criteria need to be checked, and audits need to be performed regularly.
One option: Construct virtual private networks (VPNs) on application level. Yet who should do this? And how should this be done?

Discussion on different countries' policies regarding the physical location of cloud storage; USA doesn't care, Germany and many others do care.
Most people agree that this is an important issue.

Mark: In the US, there are problems with withheld information exchange. The solution was just to use encrypted emails for data exchange: The US national network for healthcare-data exchange is based on email.

After a coffee break, Roswitha Bardohl (Dagstuhl scientific officer) explains: The Dagstuhl Manifesto should be about 20–25pp long, it must have an Executive Summary and it should cover

- the state of the art,
- visions for the future,
- 6–7 main topics that could form the body of the manifesto.

It should say where research could go in 5–10 years, briefly address open problems. The *Informatik Spektrum* [a german journal] may publish a 2–3 page summary.

Audience: Should it address funding agencies in Germany, but also outside: European Commissions, NSF?
Would Dagstuhl like to collect suggestions regarding people who could / should receive the manifesto?
For the NSF, the executive summary could be 1–2 pages . A 25-pp document would be fine for an NSF working group. Who should be contact people in Germany?
While the *Informatik Spektrum* may publish a 2–3 page summary, should one also try to publish some summary in the US?

After this short explanation, Bernhard continues: The law defines how to deal with personalised data. Someone from security checks users and says whether or not that user is allowed to access certain personalised data. Permission is granted on a case-by-case basis for each data set.
Define what role a person may have at a certain moment and define what the person can do who steps into a role temporarily. Only a person willing to accept a certain role has the right to use some data.
Consider a manufacturing process, e.g. the process of building. There are groups who have rights and perform functions. They have to work out service-level agreements. This implies the need for a central entity for processing registrations.
There is also the problem that different domains in the building can have different languages and may use different words with the same meaning or the same word for different meanings. Can this be solved better by a bottom-up or a top-down service orientation?
We should build systems that are able to learn the meaning of data.
What is the financial consequence if we make changes? Someone may have to pay.
How can I convince two people to change their thinking: why is this or that better? This is the much more difficult problem.
What is specific for a medical hospital? In principle the infrastructure. Changes are data-driven.

Ina: What can we learn from the financial market systems?

Eric P: Discussion on protocols etc.: One should employ RDF.

Hans: George Church [Professor of Genetics at Harvard Medical School and Director of *PersonalGenomes.org*] is getting 1000's of people to put their sequences on-line. RDF documents need encryption, hiding of IP addresses. Yet, if excessive data protection would endangers the life of large numbers of people, this would be bad and would need to be highlighted.

Discussion on data protection, privacy.

Titus: At what time of life do we need access to data? Childhood or later? How often does one need to update the database? Screening – start early. If not done early, one will need to call an institute to do the data collection. Childhood or later? When should we start to trash data in the institute. Data more than 6 months old are not useful to the patient. But keep the genome!

Joel: There is the important history of re-admissions, needs preserving data.

Next, Eric P starts a slide presentation on *The landscape of Healthcare Data* , see http://www.w3.org/2013/Talks/0819-Dagstuhl-egp/, stolen from: http://www.w3.org/2013/Talks/0604-SemTech-egp/.
Standards and initiatives are collected at http://www.w3.org/wiki/HCLS/ClinicalObservationsInteroperability#Standards_Bodies describing

- levels of semantic interoperability,
- policies/initiatives,
- existing standards.

In the bottom-scale 'text soup', we have simple conjunction interpretation; sentiment analysis; NLP (Natural language processing).
At the top, we have complex input and interpretation.
The data size is growing as well as the length of the data vectors (where data are coming from organisations, remote clinical devices, web-based sources, . . .).
Warehouses, in particular clinical-data warehouses show varying coverage and levels of details, use anonymous selection.
i2b2 (Informatics for Integrating Biology and the Bedside) entertains wider networks, yet puts no effort at input coding, lookup tables are built into protocols, garbage in/garbage out.
There are three forms of standards: character level; semantic; semantic web-based.
What to want: Preventable death being prevented, accountability, data for far-sighted decisions, democratised innovation, personalised medicine (cheap enough to save populations of unimportant people), science (e.g. systems biology), P4 medicine [Leroy Hood].
Population wisdom: from bedside to bench and back and beyond.
Population-based conclusions from individual data instances; individual treatment options from population experience; common data appearing in multiple contexts, e.g. syndromic surveillance; outcomes associated with procedures, interventions, and substance administration.

Clinical care:

- US: Affordable Care Act mandates patient portability in the form of Health Level 7 (Consolidated) Clinical Document Architecture (HL7 C-CDA) standards;
- EU: The EU's Innovative Medicines Initiative focuses on clinical data re-use, supporting
  (i) Electronic Health Records for Clinical Research [EHR4CR],
  (ii) the SALUS project on post-market patient safety[14].

Clinical trials:
The European Medicines Agency (EMA) is pushing the publication of clinical-trial data for public health pursueing the goal of understandable clinical data formats.

[14] From the SALUS website: Pre-approval clinical trials cannot guarantee that drugs will not have serious side effects after they are marketed. Post-approval drug safety data studies aim to address this problem; however, their effectiveness is started to be discussed especially after recent examples of drug withdrawals. This is due to the fact that, current post market safety studies largely depend on the submission of spontaneous case reports where underreporting is a major problem. The need for a more proactive approach is apparent, where safety data from multiple sources are actively monitored, linked and analyzed. Effective integration and utilisation of electronic health records (EHR) can help to improve post-market safety activities on a proactive basis.

The US Food and Drug Administration (FDA) pursues the goal of improved use of submitted clinical trial data through 'Therapeutic Areas'.
Pharmaceutical companies have attempted to use submission standards to create study repositories.

Policies – how is it going:
Clinical terms: Too early to tell how SNOMED CT[15] will impact the medical world.
Clinical trials: In 2009, the Clinical Data Interchange Standards Consortium (CDISC) identified a need for intra- and inter-pharma sharing of trial metadata, articulating questions and value sets.
Now, GlaxoSmithKline have developed their own terminology system.
CDISC2RDF, a company developing semantic models for CDISC-based standard and metadata management, now part of the FDA/PhUSE Semantic Technology project, co-operated with AstraZeneca and Roche (Tucson, Arizona) to map SDTM data[16] to RDF while AbbVie, Boehringer Ingelheim, AstraZeneca, BMS, GSK, J&J, Lilly, Pfizer, Roche, and Sanofi formed *TranCelerate* to advance innovation and tackle inefficiencies in R&D, launching three new initiatives and expanding on two existing programmes in its second year of existence.

Clinical informatics: Developing clinical informatics is also considered to be critical to reform, presenting an information model in form of a general-purpose graph structure for medical information on

- activity times and states,
- mood: considered, planned, accomplished,
- factors commonalities in e.g. Prescription vs. Injection,
- acts: observations, substance administrations, procedures, . . .
- entities: people, organisations, places,
- roles: patient, nurse, surgeon.

The system tells you where to write, but not how to write defining terminology, data types, and value sets, and uses taxonomies for:

- problem statements,
- tests performed,
- medications administered,
- reactions observed,
- and diagnoses.

Intersecting these, one gets information models, terminology models, and so on that are connected to list complex observations with respective evidence.

The five biggest challenges are outreach & education, architecture, validation, knitting/gluing, and building a physician interface (hardest task!).

Finally, Eric N takes over throwing in a number of unconventional thoughts and catchwords:
There were three important breakthroughs: Kurt Goedel, Ernst Mayr, Eleanor Franklin:

- Kurt Goedel – incompleteness theorems,
- Ernst Mayr – modern synthesis, teleonomic processes,
- Rosalind Franklin (with Watson, Crick, Wilkins) DNA.

[15] SNOMED CT or SNOMED Clinical Terms is a systematically organised computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world.

[16] SDTM = Study Data Tabulation Model, it defines a standard structure for human clinical trial (study) data tabulations.

Figure 6 The word cloud relating to the contributions by the two Erics.

Where are we today? One-directional view of biology: the genomic age – dramatic results. But do they yield the right representation?

Causal formalisms: useful in modelling & predicting biological systems, can also identify flaws & weaknesses in the system . . . .

Yet, living systems are not *control systems*.

The teleonomic principle answers not ‘how’, but ‘why’.

Teleonomic processes owe their goal directedness to the operation of a programme.

Normal evolution (species → organism → functions).

Basic drivers: energy partitioning, reproduction, mutation, selection (survival).

Cancer emergence (degenerative organismic functions → neoplastic functions).

Data ↔ Models, always separated.

Pathways, systems biology, causal networks.

Meta-mathematics – completeness & consistency.

Basic Diagonal Lemma [key to Goedel’s theorem]. Goedel’s incompleteness theorems: If a system is provably consistent, then it cannot be complete.

What does biological consistency mean?

True statements are unprovable, false statements are disprovable.

The Dilemma. Inconsistency.

ω-consistency: If a theory T is not ω-inconsistent, then it is ω-consistent.

β-consistency: within a biological system S, if a bio-function F defined by S occurs for most inputs x to S, but can be shown not to occur for some inputs z, then it is β-inconsistent.

incompleteness is an immune-related problem. ‘inconsistent’ means that you should not have let it stay.

Immune-related cancer avoidance:

Goedel type 1 errors – biosystem incompleteness [false foreigns],

Goedel type 2 errors – β-inconsistency [false selfs],

goedelian augmentation [GA]. Done by generating new DNA.

Medicine 2020, goedelian loop. Goedel’s incompleteness theorem has possible implications in the limitation of immune and cell regulatory functions

8 Virtualisation, Computation, and AI in Medicine

(The Discussions on Tuesday Afternoon)

The topics of this session were

- The status of structured (pathway, model etc) databases,
- The virtual oncology, diabetes, ... patient in medical practice,
- Mechanistic models,
- The vision of ITFoM,
- How can we extract information from the scientific literature as well as from 'low-grade information' in the web (text mining, the semantic web in healthcare, search strategies in semantic webs)?
- Virtualisation in drug development.

Coordinator: Hans Lehrach
Contributors: Laxmi Parida, Pietro Lio', Joel Saltz, . . .
Rapporteur: Andrea Splendiani

To start the discussion, Hans asked:
How much of the way we look at biology is shaped by the way we use to organise research? The publication process likes simple stories (referees don't like the truth: it is too messy and too complicated). Biology is complex: biological systems are built of parts that are selected with no goal or criteria behind it. There is no penalty against complexity.
There are basically two ways to study biology: statistics and modeling.
Modeling has a 'bad reputation' because of some application areas and because of the lack of incentives to validate models in academic research. In 20 years, medicine predicted via analytics will be much better.
Computational approaches to medicine exist in a continuum: from simple statistics to mechanistic modeling based on combining biomolecular insights, genetics and clinical-trial results (e.g. subpopulation performance). As more and more parameters can be estimated, mechanistic approaches will play bigger roles.
One problem is access to parameter values. However, the parameter space can be restricted by observations on model behaviour (feedback). Modeling results should not only be used for prediction. Their deviation from real outcomes should be put back in the modeling process. Models evolve (and their evolution can be to some extent automated).

Andreas: For a good model, systematically comparing the effects of parameter *perturbations* with real outcomes often is particularly useful.

Hans: The medicine of the future will be a self-learning system based on the underlying assumption that all diseases have a mechanistic base.
Of course, not all diseases are amenable to a model-driven approach. Cancer is ideal as you can take samples on which to test a model for intermediate time points. Yet, mechanistic models can also have 'soft parts'.
We have two types of information:

- Open information: what is known about the effects of X on Y (from any source, experimental or inferred). Linked Data is an interesting way to access this knowledge.
- Information that is specific to each patient. Any patient could have an IP number: this would be an interesting way to accumulate information on a patient over her/his

lifespan. There could be several data sources (hospitals, fitbits[17], blood pressure, food, environment, ...). A continuum from medical application to lifestyle.

How do we get all that information together? Not all information is certain.
Artificial intelligence is in a sense artificial. Statistics could have larger weight. The relation between Bayesian statistics and RDF should be explored.

Then, Laxmi started her contribution:
This will be a Big-Data free presentation on *IBM's Watson: The smartest machine on earth*
IBM's *Watson* is good at playing Jeopardy, a game rewarding general knowledge of participants, formulated in complex natural language.
Ultimate test in NLP: won at Jeopardy playing against the best two players in the world. Note that *Watson* had to go through an audition. The audition was not a Turing test (not blind, the presence of the machine could have influenced other participants). After the introduction of machine learning (with baseline NLP + Data Mining), there was a big improvement in performance. It has a massively parallel, probabilistic, evidence-based architecture, required 150 man-years, relies on known facts, and cannot (that is, it was not built to) query the web online. There were three main issues:
Issue 1: Ingest and understand information,
Issue 2: Finding information from a question,
Issue 3: Confidence in answering the question.

Confidence is relevant: wrong answers are penalised.
Watson in summary: Interprets and understands natural language questions, analyzes large volumes of data strored in its memory, generates and evaluates hypothesis and quantifies confidence in its answers, adapts and learns to improve results over time.

To cope with problems in current medicine, IBM is now adapting Watson technologies to healthcare.
Content: Samples of medical texts.
Training: 1300 diagnosis and treatment questions with known answers

History of *Watson*:

- 2006 IBM Research,
- 2011: Jeopardy,
- 2011: healthcare,
- 2012: finances and others.

Watson can also be used to predict the phenotype-genotype mapping, in particular simple-trait mapping.
Evaluation on (part of) the test training set showed a 5–10% of improvement of prediction over the base line which is considered to be a big improvement.
Prediction methods: rrBLUP, a software for genomic prediction, (standard) linear models [untypeable], other standard methods for genomic regression: Lasso, Ridge Regression, Bayes A/B/Cpi, SVM. More exotic methods: Random forests, Markov Nets, Neural Networks, Boosting, ...
MINT: A mutual-information based software for transductive feature selection based on genetic trait prediction, considering interaction between markers (2 markers together may

[17] i.e., wireless-enabled wearable devices for activity tracking, measuring data such as the number of steps walked, quality of sleep, and other personal metrics

group
Lehrach
change
Systems reality Saltz natural
linear doctors Model privacy used
produce together variation process Virtualisation
omics system studies link
IBM also Mechanistic
population
trait training compute use interesting Watson two
participants access correlation better approaches practice markers processes months
take
disease related Networks models care results effects biology years
certain medicine key patients health possible experimental Pietro
extract specific
language confidence online
report way patient mechanistic clinical information Modeling
healthcare questions conformation sample
schools understanding parameters trial gene medical Prediction improve intelligence Joel GWAS statistics
summary
Jeopardy
important improvement machine data model research semantic
epigenetics prediction question literature Issue difference Bayesian
well Lio e.g
web answers test complex via Medicine independence free methods
osteoporosis One learning
continuum need application performance modeling known asked
text drug
slides causal general treatment base slide time genomic trials diseases knowledge DNA line
part Hans one diabetes simple feedback
first
his/her VPH Nets

Figure 7 The word cloud relating to IBM's Watson.

boost a trait). It led to improvements of 20-25%, may help to find the missing causes for heritability of complex diseases by correlating multi-gene small effects with traits[18].

Mark: Watson is used at the Memorial Sloan-Kettering Cancer Center. It was trained for about 1 year on-site on one disease: adenocarcinoma (all literature + case studies). It is now in clinical use. A doctor sends the entire EMR of a patient to Watson (no genomic data), about 80% is free text. Watson gives a summary (links to papers) and about 60% of the time is confident enough to suggest a treatment. Other facilities in New York are using it as web-server.

The amount of training information specific to a domain is key to confidence (not applicable to all domains in healthcare).

Next, Pietro started his contribution on *From data to models*:

The *Virtual Physiological Human* (VPH) is a methodological and technological framework that, once established, will enable collaborative investigation of the human body as a single complex system.

It is developing a roadmap for the *Digital Patient* and has the potential to fund research projects.

While one needs to integrate all clinical and molecular (omics) data, it is difficult to follow all technologies, and groups will often follow only one type of omics.

The most important information (upstream) is related to DNA conformation (chromosomal territories). Chromosomes change, not only in Leukemia, but also in Diabetes and other pathologies.

It is possible – and may be enlightening – to compute the spatial functional statistics of loci distances in chronic diseases. In this context, tools are important.

Epigenomics reveals heart fatigue. There is a strong correlation between epigenetics and DNA conformation.

Also the gene expression process is linked to epigenetics.

[18] See e.g. Visscher, McEvoy, and Yang: *From Galton to GWAS: quantitative genetics of human height*, Genet Res (Camb). 92 (2010):371-9. doi: 10.1017/S0016672310000571. See also http://www.fimm.fi/en/research/research_groups/group_ripatti/ for a report on studies of Finnish biobank samples that provide a much more precise description of our genomic, transcriptomic, metabolomics, proteomic and other high throughput variation and its relation to complex traits and diseases.

GWAS studies cannot explain more than 10% of the observed variation. A long tail of common variants including more omics data could improve this rate.

Multi-Morbidities: A drug to fight an infection could differ depending on multiple conditions (Metformin targets NKFb. It's pro diabetes, but against Alzheimer). Multi-morbidity makes patient stratification very hard. Omics for diagnostic and prognostic markers needs to be developed.
Classical models are organ-centric. We need to move to process-oriented models.
Processes are mechanistically related. The website http://diseaseome.eu presents maps of diseases related via co-shared pathways.
Important are also multi-modeling methods – including parameter estimation.
There are many modelling approaches: Process Algebra, Compartment-based and Rule-Based Systems, State Charts, Hybrid Systems, Boolean Networks, Petri Nets, Agent- and Lattice-Based Models.
Here is an interesting challenge for multi-scale modeling:
TGFb is anti-cancer under a certain concentration and pro-cancer for higher concentrations (apoptosis inhibitor). Metastasis is first in the bone and produces osteoporosis. Other diseases as well produce osteoporosis.
Our first-order linear model includes a feed-forward loop external to the cell.
Geometry affects modeling. Think of what a few cells that detach integrin can do at the level of the tissue. From that, one can model effects on the bone. One cannot use a single model as there are different time scales.
Model checkers and other formal methods can be used to analyse how a program (model) fits results that can be observed.
What we need is

- multi-omics integration,
- understanding co-morbidity,
- identification of key parameters,
- process modeling,
- causal analysis.

There is resistance toward the adoption of models by clinicians (though examples as insulin pumps prove the contrary). You can imagine that, in a few years, patients could take their own data to train a basic model.
Issue: Where are data for machine learning coming from? There are data that are not in official trials. Results that can be derived are questionable (example: tumor reappearing 4 months after treatment). People debate whether it is appropriate or not to use such data. There is a difference between first-line care and last (least worst) solution. And there are also many other considerations.

Finally, Joel Saltz addressed *Pragmatic Clinical Trials*:

Randomised clinical trials are quite expensive and affected by combinatorial explosion if all factors are to be taken into account (race, age, gender, ...).

Basic concept: If you want to do a clinical trial in reality, you need lots of people, EHRs (not always report forms), even though you may have interest in understanding the impact of a variation of only one factor (A/B comparison), e.g. intervention type (measuring quality of care and other parameters).
Patients populations form a living experimental laboratory.
What sort of recommendation can be given, in particular when you need not only access to data, but access to processes.

Biomedical informatics: In many medical schools, biomedical informatics is done by a group of middle-ranked people with CS skills (should know a bit about machine learning). People doing the technical work are not in the position to change processes (e.g., add a few more questions to be asked).
We need to find mechanisms for making data available for the research side of medical schools (in compliance with privacy) and provide feedback to clinical practice.
PatientsLikeMe: Encompasses 2000 communities about serious diseases. There was a report in the Italian literature that Lithium was helping in a specific disease. They asked doctors for lithium prescription and posted results within 3-4 months on the net (not viable).
Most of privacy laws can be side-stepped if patients volunteer.
Wallgreens, America's big online pharmacy, can find patients matching any clinical trial specification.
Are causal Bayesian models a better statistical model to extract information from 'causal' resources?
One reference for conditional independence testing (à la Judea Pearl who developed Bayesian networks) is http://en.wikipedia.org/wiki/Conditional_independence and on Bayesian networks in general see http://en.wikipedia.org/wiki/Bayesian_networks. And for a reference to the learning healthcare system from the Academy of Medicine, see http://iom.edu/Reports/2012/Best-Care-at-Lower-Cost-The-Path-to-Continuously-Learning-Health-Care-in-America.aspx.

9 Molecular Systems Medicine and ICT I

(The Discussions on Wednesday Morning)

The topics of this session were

- Assessing emergent properties of chronic diseases and disease mechanisms:
  – The dynamics of disease progression and implications for disease mechanisms,
  – Parallel toponome decoding and genome sequencing,
  – Cancer and the immune system.
- Family genome and tumor genome sequencing – the use of Next Generation Sequencing and its implications for therapy and disease stratification.

Coordinator: Peter Walden
Contributors: Walter Schubert, Robert Burk, Markus Löffler, ...
Rapporteur: Eric Gordon Prud'hommeaux

First, Peter showed two images: Is this an example of a patient who will benefit from intervention?

Question: Patient or model?

Peter: First image was patient, 2nd was a patient-tailored model.

Jonas: What routes do you see for ICT in your work?

Peter: ICT is moving into clinics, but usability is low. So, we may need to breed a new generation of people who can use these tools.
Also, modelling techniques haven't entered the clinic because of usability issues.
The next slide showed heterogeneous cutaneous melanoma with an aggressive immune response attacking the tumor resulting in vitiligo (i.e., whitish skin due to depigmentation), visually-apparent melanoma clones, vascularisation, doomed.

One could not have predicted this just from omics data.
The transformation occurs at the interface between aggressive growth and regression zone.
An immune reaction is killing off one clone, thus making space for a more aggressive one.

Jonas: Was having to wait for a physician the mistake? Should we have a home device?

Peter: Absolutely. If your tool causes unnecessary surgery, that's no harm.

On the next slide showing melanoma metastasis, there were far more (tumor-specific) T cells than skin cells. Successful regression is not indicated by more T cells, just by T cell distribution.

On the next slide, the effect of therapeutic vaccination is demonstrated. Deep pigmentation indicates effective immune response.

Lena: There are iPhone apps for sending pics of skin issues (cf. e.g. https://angel.co/goderma-gmbh-3). Are they helpful? What's the balance?

Peter: You can always consult with a physician. We don't know what separates those who will respond to vaccination from those who will not. Anti-tumor responses are, by and large, autoimmune responses.

The next slide shows transplantation-associated cancers. Remarkably, melanoma risk increases 2 to 3 times in transplant subjects. Blood is a poor reflection of what's going on in the tumor. The T cell response is divided into tumor and inflammation response.

Eric N: If there are different groupings, you may see multi-modal correlations after genetic classification.

Peter: Not yet, but tried using many distinct markers. Coinhibitor responses need to be regulated, or you get immune pathologies.

The next slide shows examples of excessive immune response and leads to a discussion of anti-CD28 melanoma trial responses.

Question: What about this claim from Stanford regarding the over-expression of a certain gene in at least 20 kinds of cancer.

Peter: I can only say we know very little about it. We need to integrate omics data with clinical observations. This brings in spatio-temporal resolution, missing from classical omics. Clinical trials are experiments which should feed back into models.

Joel: Lots of folks are working on relationships between omics and pathways. As expected, both are significant. You can get extra precision from omics, but this is currently only good to predict outcome. Characterizing immune response is a 'key 3rd leg'.

Pietro: Predictors of outcome, mortality, and morbidity; what's the difference between omics and classical forms?

Peter: Omics seem to be a tool to get into mechanisms.

Robert: As we move from subjective to objective, we improve. Slides are subjective.

Pietro: If you had a choice between omics vs. patient-participation tools?

Peter: Cancer patients are always willing to participate.

Jonas: Can you push molecular studies to patient devices?

Robert: Sometimes, we can't even do non-invasive glucose monitoring.

Peter: Diagnosis for infectious diseases was always invasive. It moved from spleen biopsy to testing tears.

Markus: Why are we not doing trials on really toxic products? We need a product that makes sense to everyone.

Figure 8 The word cloud relating to molecular systems medicine.

Now, Walter starts his presentation:
We can't draw conclusions on the basis of the data we have. We may need co-mapping to get data, then built a model. There were elegant, logical studies that failed to translate into clinical trials. One needs to investigate mutations and functionality in parallel. Every disease has it's own 'cathedral' (super-molecule).
Next, Walter shows a demo on cross-communication between 100 proteins translated into to sound. The richness [sounds like chaos – Eric P] of the later sample indicates health.
Doing imaging of those 100 proteins also shows reduced functionality in the diseased tissue. Imaging at 1nm resolution revealed 6 layers in separate membranes of which only 3 were known.
An image of a T cell speared by an epithelial neighbor illustrates other remarkable processes. We know by volumetric analysis that killing 300–500g of T cells will kill you.
The whole system matters.

Next, Robert presents his views: Medicine has migrated from historical to evidence-based medicine. We gather a bunch of information going into a cloud. Epidemiologists have pathogenesis knowledge which has developed into action.

Scott: How is that knowledge codified and shared?

Robert: Everyone knows we need to integrate the data and, therefore, need to figure out the impediments: government, insurances and policies direct the data input. We will also have to address the microbiome (totality of bacteria and viruses within our body).

Eric N: We're the transport mechanism for a microbiome.

Robert, presenting a 'heatmap' showing percentages of different reads of body cavities, explains: Advances in microbiome analysis come from NGS. There are two ways to search for it:

- 16S ribosomal RNA in procaryotes, doesn't classify species to phylum, so they talk about Operational Taxonomic Units or 'OTUs'. You amplify a region common to most things and try to guess what's there. Limited to detecting biota already in our database.
- For a gold standard, you can sequence everything.

One can associate ecological measurements of complexity (e.g. Shannon entropy) with disease state. We can handle high variability if it's a strong predictor of disease. Disease in a microbiome is called 'Dysbiosis'.

Scott: Do we know how variable the biome is in healthy people?

Robert: Our microbiome is fairly stable in an individual over time. One is not thinking about simple risk alleles, but instead about interactions.
Regarding cervical cancer, it was thought to be associated with Herpes. It was difficult to prove connection with the human papillomavirus (HPV). With a proper study, the odds ratio is nearly infinite because it's a direct etiological agent. Understanding the molecular mechanism for disease was required for implementing the test.
A study by the *Kaiser Permanente Center for Health Research* involving 1 million people checking for HPV(+/-), Cytology (pap smear), and outcome (CIN3) showed a clear association of cervical cancer with HPV. But it was a great challenge to develop a corresponding effective treatment[19]. The complexity is too much for physician. 20 experts argued about treatment strategy. There is need to use IT to enable physicians without taking away their role as healers, and to query longitudinal data with a better interface to data capture.

Question: I heard that pretty much all males develop prostate cancer, but mostly non-aggressive variants.

Robert: From studies of soldiers, we see lots of stuff that 'looks' like prostate cancer, but we don't know what is going to happen to them.

Mark: The EMR systems were designed with billing in mind, not with what we know we need to know.

Scott: It's not just data capture; two 'epic' systems can't exchange.

Mark: But capture is critical. There are glimmers of hope, but the topic does not get much focus. The systems are ignorant of clinical processes, data modeling, and evidence production.

Finally, Markus takes over:
There are nice models for swarm formation: It suffices to introduce two simple rules to govern behaviour. Gene expression data won't reveal these rules. Models are rules; only informed by data. 'There is no clinical medicine unless in the light of models.'
There is a statistical (non-molecular) model that describes the development of Hodgkin's Lymphoma. Its predictions, validated by a study, established *the* treatment standard that was tried to transfer to non-Hodgkin's Lymphoma.

There are more tested/confirmed models, e.g., the model regarding the growth of epithelial crypts. The old belief was that the large granular lymphocytes are the drivers for the crypt. Our belief: They have no effect, plasticity overcomes. This is supported by a model of exchange between aggressive and dormant lymphona clones.
Personalised medicine started with brain-skull trepanation thousands of years ago. Personalised *molecular* medicine has some good use for splitting into groups, not to get distracted by irrelevant complexity.

Jonas: Suppose someone stratifies your results by similar outcomes, would it be an improvement?

Markus: We need to test if biomarker classifiers are good. There are tests for Burkitt's lymphoma based on *Burkitt-ness classifiers*[20]. However, this detection is not that important because all lymphomas are treated the same way.

[19] Cf. *Studies of Cancer in Humans*, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, http://monographs.iarc.fr/EricNG/Monographs/vol90/mono90-7.pdf

[20] Cf. http://www.nejm.org/doi/full/10.1056/NEJMoa055351.

Robert: Regarding such classifications, we may need to have experts curate our inference machines.

Markus: Question: How do we design trials? The *National Cancer Institute* has a protocol for a phase-3 marker test design. In tests that split out subgroups, you need to assure that the performance in derived subgroups is significantly distinct. After implementing a strategy to differentiate folks for whom a particular antibody treatment would apply, we hoped for 10% and were surprised by 20%, highlights the value of this specialisation.

Question: How does the mechanics of the diagnosis work?

Markus: We need molecular testing to discover whether the target for the antibodies exists in the tumor. There is a consistent protocol for producing self-tailored vaccines.
We are currently designing a trial for adaptive HIV therapies (which find the optimal balance between the 30 available single-drug based therapies in order to control the HIV virus and it's many mutations).

Jonas: From the FDA's perspective, your algorithm would be the drug. Do you get a stamp at the end?

Markus: No, they haven't worked out their protocol yet.

Jonas: How do you feel about the data supporting these trials?

Markus: This is not just an observational trial, it's a well-designed prospective cohort study.

Mark: How would you support blinding, i.e., keep docs from knowing they've been augmented by a machine?

Markus: Our plan is to create two teams making recommendations on the same patients. The treatment allocated to the patient is selected randomly from these two recommendations.

Andreas: There is Ingo Althöfer's *Triple Brain* Chess Team[21] involving two chess programs that propose moves and a third person (or program) that selects one of those two. This idea worked quite well.

Markus: It would indeed be good to feed back clinical trial data into a public repository. We've discussed the technical side. We also need to make the algorithms available and need protocols around the biobanks.

Jonas: Would you change the language in a patient consent form?

Markus: Many current consent forms wouldn't enable public sharing. One would need to change language from just 'research' to 'research and . . . '.

Robert: The fundamental basis of clinical trials is checking one or two variables. With this, you're introducing many variables. How do you compare them?

Markus: The brute-force design only tests the efficacy of the overall system, not the specifics.

10 Molecular Systems Medicine and ICT II

(The Discussions on Thursday Morning)

The topics of this session were

- A systems approach to diagnostics,
- Biomedical engineering and systems-optimisation strategies in medical care,

[21] Now also integrated in *Shredder*, cf. https://chessprogramming.wikispaces.com/Shredder

- Assays for wellness,
- Precision medicine and guidelines for evidence-based medicine: complementary or incompatible?

Coordinator: Ina Koch
Contributors: Anja Hennemuth, Helena F. Deus, Susana Vinga, . . .
Rapporteur: Laxmi Parida

Ina Koch reported on *Some aspects of image analysis of Hodgkin's Lymphoma* and presented some slides explaining her point of view on the subject matter.

Jonas: Your distance is not symmetric. Is that a problem from the biological perspective?

Walter: No. In tumor histology, this is fine. Within abnormal cells, things can be not symmetric.

Alex: How is the distance computed? Euclidian?

Walter: Normal lymphnodes will lead to the same distances.

Ina: Look at which size and shape tumor cells in the neighbourhood exhibit? Still some confusion remains about how the distance is computed.

David: Can one map individuals to these distance maps? Individualised maps?

Ina: Yes. But this is not the main goal. Pathologists still do it by eye.

Andreas: Would you like to eventually replace pathologists?

Ina: NO, we just want to help in complicated cases and learn more about the cell neighborhoods and the distribution of cells.

Anja reports on *From images to models to therapy*:

In contrast to Ina's work, her group uses a top down approach based on *Fractal mathematics*. Can this be used to explain the spread of vasculatures? It seemed that the program could be useful for surgery and, thus, was given to surgeons. So, they became an ICT institute providing surgeons from all over the world with software simulating the vasculature in the liver (virtual liver) to be used for tumor surgery and other therapies. Tissue properties regarding heat distribution were added as well as different biomarkers for stroke etc. to be used for diagnosis and therapy.
As a contribution to system medicine, this went from clinical to complex modeling and personalised and precision medicine[22].

Next, Lena presents slides on *Gene Silencing*.

Eric P: This is just for molecular therapy?? Must have some other uses as well.

Lena: TCGA data grow.

Jonas: Not just 'grows': TCGA data grow exponentially, doubling every 7 months.

Mark: Scale? For 1000 months?

Lena: No, just days, including many preliminary data.

Eric P: Isn't there a large amount of winnowing that happens in the drug industry.

[22] Media attention to this work is documented at http://9to5mac.com/2013/08/21/liver-surgery-now-theres-an-ipad-app-for-that/ and at the *Grand Challenges in image analysis* website http://www.grand-challenge.org/index.php/Main_Page.

Eric N: Pharma companies resist, only use small information at a time, everything has to be presented.

Susana: There is an important reference IEEE paper on *Grand Challenges in Interfacing Engineering with Life Sciences and Medicine* [12]. It emphasises that both, mechanistic and learning models, are important.

Marc: It is indeed a great paper, systems-engineering people are nice – they are problem solvers. It presents the medicine institutes' view as well.

Mark: Point-of-care sensor technologies are a very important and dynamic area.

Robert: Longitudinal studies – a lot of data has been collected, but not used. There is need for new innovative ways.

Walter: Isn't life spatial?

Susana, Joel, Anja: Indeed!

Joel: One needs a combination of radiology, pathology, and omics data to perform *spatial engineering*.

Walter: Does there exist any system that can be used?

Anja: There exist systems for gathering and assimilating such information.

Walter: We need a 3D approach to life. Which system is sufficiently far advanced so that it can be used?

Anja: You can combine our vasculature-systems model with other topics for a modeling approach. The VPH projects aim at this.

Mark, David: Model mashups – different models that come together, cf. http://en.wikipedia.org/wiki/Mashup_%28web_application_hybrid%29!

Mark: Job of systems engineers.

Susana: Largely ignored – ICT is needed for bridging these models.

Mark: THink of a whirlpool engineer, modern washing machines considers chemicals, mechanical etc. aspects. Is there a lesson for biology?

Anja: We are working in this area. Yet, already agreeing on terminology is difficult.

Marc: Sensors are underrated, ON and OFF rates need to be studied. Biological OFF rate can be studied, OFF rate is the time for A and B to stay attached. Then, they dissociate. Should attach long enough and go to the membrane for reaction.

David: There are indeed grand challenges in the IEEE area. Yet, we need to be cautious.

Mark: Progress in the field will require multidisciplinary.

Alex: Flow Cytometry data analysis is a real case of Big-Data robots producing much larger data than can be processed at the moment. They use the same technology as IBM's ink-jet printer from 50 years ago, high dimensional and high throughput, measuring the intensity of the fluorescence (proportional to the number of certain proteins). Instruments are different. There is a problem with continuous data. Compare 'clustering' results across hospitals.

Walter: Different times of the day gives different readings

Alex: Yes, has been taken into account.

Andreas: Regarding the distance between samples: You have several hundreds of samples?

Ina: Do you use different stainings?

Alex: We measure all simultaneously.

Andreas: Do you see any clusters?

David: Or use hierarchical clustering all the time?

Alex: We use hierarchical clustering for community detection (like physicists), the data are collected and aggregated over many cells.

Andreas: Has this still anything to do with medicine?

Alex: Indeed: These are the actual protein measurements.

Joel: This is useful in transplants. A set of drugs that activate/suppress can provide information about compatibilities.

Ina Koch: So, this relates to *complexome* analysis?

Walter: Can the results be compared with normal data? What defines its homeostasis ? Can disbalance be predicted?

Marc: In some, there is age dependency, and in others , there is no age dependency.

Peter: Regarding the composition of the cells, there exist records over a period of time. This could be used for prediction.

Alex: SONY is getting interested in this area. They have announced a new machine that can perform many more observations – since the noise is mixed (gaussian – biological – poisson – electronic), they can measure 50 proteins at a time. The CyTOF Mass Cytometry and Mass Spectrometry platform is monitoring flu spreading geographically. It can also use tissues.

Andreas: Could you pick parts of healing tumors from different regions?

Peter: Cells must be in suspension. Tissues are difficult and tissue cells much larger.

Marc: CD4 is a remarkable receptor in the membrane; HIV attaches itself to this receptor; the number dramatically increases (to 350) and is ready for AIDS – then anti–retro viral drugs interfere. This number is important to decide when an AIDS infection is about to occur. CD4 is a crucial molecule and could be a good target to design a block for attachment to the CD4 target. When antibodies attach, they change the antigen conformation – there is *synergy*.

Alex: In the example above, it should be decreased to 350 cells.

11 The Stratification of Disease into Discrete Subtypes and its Implications for Science, 'Medical Ontology', Diagnosis, and Therapy

(The Discussions on Thursday Afternoon)

Coordinator: Markus Löffler
Contributors: Mark Braunstein, Eric Prud'hommeaux, Peter Walden, ...
Rapporteur: Susana Vinga

Markus: There are not so many prepared talks – anyone wants to contribute?

Mark: Is healthcare at a tipping point? According to Lena's graph, there is a Big-Data revolution in healthcare. Computing stretches from quantitative to qualitative. When you are in a middle of a *disruption*, it is difficult to see it. The first information revolution in medicine was the digitalisation of files.
1864 – Florence Nightinghale
1964 – Lawrence Weed: intelligible hospital records (cf. his paper on 'Medical records, patient care, and medical education)

2001 – crossing the quality chasms – IT must play a central role.

Americans think healthcare is like in TV, US has the best quality worldwide. But: Comparing mortality rates and costs for chronic diseases, it looks not so well.
68%: Chronic diseases costs. Why? Current healthcare delivery system is designed for treating acute illnesses.
For obesity, there are different system of care in the US.
The *Primary Care Physician* – network in the US claims that averaging of multi-chronic disease patients challenges patient records. Life expectancy is higher in Europe.
Federal Approach: Bush appointed some 20 – 30 billion to reimburse cost of EHRs.
Federal Strategy starts with EHR certification for meaningful use (doctors should use them that way). In April 2013, 75% of eligible providers have enrolled, also eligible hospitals are committed to the EHR system.

David: Do they have to put everything in electronic format?

Mark: Not old things. Yet, there are

- opportunities: expand the capacity to generate new knowledge , etc.,
- challenges: no strong incentives, privacy, fragmented EHR,
- perverse incentives: shouldn't be on how many visits or tests.

Shows a map illustrating the geographic distribution of 2012 medicare organisations.
Question: None in Alaska? Answer: Just 3 people there.
Political issues (sausage analogy: if you see how they are made...).
Walgreens (the smartest pharmacy in the states for decades) became the first national ACO (Accountable Care Organisation).
Interoperability, data-standards issues: MD can enter free text and it's exported to *SNOMED*, guarantees accuracy, IBM Watson is already addressed.
Projects/Examples:

- Earlier diagnosis of CHF (cardiac heart failure) by machine learning.
- Artificial intelligence framework combining Markov decision processes and dynamic decision networks to infer the most cost-effective treatment/approach.
- Paediatric chest pain: how to manage and evaluate, infer 'red flags', list of symptoms/features?

MDs echo data, compare younger with older doctors. Information from 2000 patients: what are the most important factors. With logistic regression, the result is simple. Does not work so well for major problems. Only hoisting red flags (based on expert knowledge) proved to be useful after all. Involved are the General Electric disease network, Microsoft Amalga, Optum Labs, Mayo Clinic, ...

David: How is that regulated, e.g. Walgreens' expansion to diagnosing and treating patients for chronic conditions such as asthma, diabetes and high cholesterol.

Mark: Well defined context, training, nurses can do that.

Robert: Be careful: For diabetes: yes, for hypertension: more difficult.
Involved are companies and researchers. The 2*net* device, a wireless hub, measures physiological data from home and sends it to the cloud employing qualcomm chips.
What type of data? E.g. in Georgia Tech, they used radar technology developed by the government (not classified anymore) to collect acustic, respiratory, and movement data without touching the patient.
The algorithm for diabetes was developed by an MD student.

The *vg-bio* company tries to improve healthcare through personalised predictive analytics using a new platform providing the earliest insight into deteriorating patient health. They apply software developed to detect airplane failures now to patients: which of them are the sickest.
Future opportunities:

- 4 levelS: clinical practices (people),
- delivery operations (processes),
- system-structure choices (organisations),
- healthcare ecosystems (society).

There is a Georgia Tech and Emory project on 'Clinic logistic predictions'.
Its *Patient-Centered Medical Home* program compares data collected at home to optimise e.g. the time for the next visit.

Wolfgang: Who designed the process model?

Mark: Students.

Jonas: We need to include process data in the medical record in the future.

Mark: Initially, these records were created only to support billing.

David: These issues should also be discussed with the participants of the security seminar in the other room

Wolfgang: Process exceptions: are they considered, when something goes wrong?

Mark: Yes, they are accounted for.

Joel: There is some experience at Stony Brook on how to characterise processes without black boxes.

Mark: Consider an example: In 60 rooms, there are 60 groups, each is focused on its issue. Then, there is no process as a whole.

Andreas: Bernhard's company's software is addressing *processes*, more specifically, business processes.

David: Focused on US processes. How are things in Europe?

Marc: France: Our card has all medical data and is totally portable.

Robert: All Scandinavian countries and biobanks work that way.

Mark: In the US, the federal system makes this more difficult.

Markus: Are there models/prototypes of processes that could be published?

Mark: Yes.

Markus: Even if there are companies involved? As we have Virtual Physiological Human, would it make sense to have also Virtual Hospital Process?

Jonas, Mark: There is *process mining.* However, data need to be free. This should be a key point for the manifesto.

Eric Prud'hommeaux starts by presenting slides on information models used in Clinical Informatics:

- the Health Level 7 (HL7) *Reference Information Model* (RIM):
- http://www.hl7.org/implement/standards/rim.cfm,
- the Continuity of Care Document (CCD): what institutions need to report: http://en.wikipedia.org/wiki/Continuity_of_Care_Document,
- the Clinical Document Architecture (CDA): http://en.wikipedia.org/wiki/Clinical_Document_Architecture.

Example: Description of a clinical record under XML.
No open access to data: Georgia Tech, Emory, . . . , or other consented research data from patients.
He mentions various initiatives providing software for converting XML to RDF called *XML2RDF* or *xml2ref*, and refers in this context to the *Reference Information Model* and the National Academy of Engineering (NAE) report on *Engineering the Health Care Delivery System* that can be downloaded from https://www.nae.edu/File.aspx?id=7417.

Stefan Decker: Starts a short discussion about the role of abstraction in the life sciences based on a number of observations and a conversation with Marc. Stefan says that, in Marc's presentation, the reduction to very few and often just one causation and the lack of abstractions was mentioned which could help to focus on observable behaviour instead of getting lost in too many factors.
Stefan mentions that CS likes to generate abstraction layers in order to manage complexity. These abstraction layers are not naturally given, but usually artificial constructions aiming to reduce the cognitive load of computer scientists (or programmers) when looking for solutions. He suggests that, maybe, the life sciences could benefit from similar abstraction layers that focus on useful behaviours and functions, and make them available as a concept for e.g. data integration and analysis.

Thanks to the efficient handling of Skype technology by the Dagstuhl offices, we could close the day with a Skype discussion with Lee Hood (Seattle) on Education and Training. Lee begins by explaining

The Systems Medicine initiative is being developed to reach all the way to healthcare, and seeks to validate a new paradigm that includes physicians and other healthcare agents. Taking the bottom-up route is advisable in this context. MDs are usually conservative.

Andreas: We need the medical community to support these news ideas, the question is how (explains the goals of the planned Manifesto).

Lee: 6 to 7 years ago, I contacted top medical schools including Johns Hopkins: There were mixed reactions towards training.

David: We understand the paradigm change of P4 medicine. Yet, writing just a white paper is not enough to change how a medical school teaches.

Peter: Explains plans of the Charité in Germany.

Lee: Who is going to be the leader of the (new) Institute will be critical when creating such a new institute.

David: Initiatives in the UK are directed towards medical information systems.

Lee: That is a very important component, particularly if it includes the various omics. The challenge there is to make sure that all of these data, including patients, falls in the right 'fold'. IT for healthcare is enormously welcome, as the Obama Health Care reform demonstrates. Based on a 3-page document, I am scheduled to talk to a parliamentary group in the UK later this year.

Andreas: Can you us send that 3-page document?

Lee: Major changes in health IT requirements are absolutely needed to catalyze these changes. IT for healthcare is the key component in the engineering of this data-acquisition exercise to identify stratification in the patient population at the molecular level.

Andreas: Can you tell us about medical clouds?

Lee: The first stage of data aggregation is to collect data from individual patients even if health IT may not be ready to handle it right now.

Andreas: We have semantic-web experts here. They think that they can handle this aggregation even though this is beyond current health IT.

Stephan then made a summary of the challenges the semantic web faces in bringing institutions to what is the edge of a new data enterprise.

Eric P: The W3C standards initiatives (cf. http://www.w3.org/TR/) work on *Open Web Platforms* for application development that have an unprecedented potential to enable developers to build rich interactive user interfaces buttressed by vast data stores that will be available on any device. He mentions that patients are part of this process and that standards are in fact advancing in this direction.

Lee: Teaching the CIOs about the advantages of these approaches will trickle to the CEOs and others. Patient advocacy and social networks also offer key mechanisms to educate institutions in moving towards systems biomedicine!
Initiatives such as PatientsLikeMe are key in this regard. This was very clear in the AIDS case where new treatments were driven by the patient associations, not by the medical institutions which, in effect, opposed them.

Eric P: Second-level underwriters play a role here, too, because they make care cheaper. How can they be brought to the table?

Lee: The insurance systems in the EU and the US are very different. In the US, they are going to be squeezed by the healthcare reform. So, their participation in a systems-biomedicine initiative would have to be driven by a clear cost-saving proposition. Maybe starting with small insurance providers is a good way to go.
Let me tell you about our 100M initiative in Luxembourg where this is supposed to be advanced: Bringing people to Luxembourg is not ideal because of locally diffuse healthcare systems and the lack of a Medical School. So, alternatives are being considered. On the other hand, the size of Luxembourg is appealing. So, that is also something that creates unique opportunities. Maybe, the Manifesto should mention these issues and the opportunities to advance systems approaches to healthcare.

David: What is systems medicine?

Lee: It's a systems approach to medicine and to disease. Disease is caused by two types of processes. One is genetic in nature, the other is environmental. Almost all processes leading to chronic diseases display a mixture of these two. Another aspect of systems medicine is that the patients will soon be surrounded, each of them, by *data clouds*.
Generating the analytical tools that can reach the healthcare environment is an absolute necessity. Machine-learning approaches capable of doing a good job at extracting signal from noise is another critical component of systems medicine. The network of networks that are documented by the patient-data cloud goes all the way to the social networks that connect people. When folks are perturbed by diseases, all sorts of subnetworks get involved in both, understanding them and treating (or preventing) them. This is a relevant discussion to have.

Stefan: There is a requirement and need for data integration. How do we generate knowledge? Play with and explore the data. This needs to be developed.

Lee: Our Institute has many threads of pipelines. Challenges will be to convince the society that this will be society data (for grandchildren): If the society generates the tools, it is the patients' obligation to give back that information to generate better medicine in the future.

Andreas: Pietro Lio' here works on co-morbidities

Lee: Disease classification (as proposed by the report of the National Academy of Sciences mentioned above) does not make sense and does not take into account systems-medicine approaches.

■ **Figure 9** The word cloud relating to medical education.

Walter: In the fluorescence-microscopy based *Toponome* approach, we take account of the spatial distribution of dozens of proteins in a single cell (or a small piece of tissue). From what we learned, we became convinced that this will be basic for any systems-medicine approach as well as for disease classification. In one specific case, our approach revealed 6 layers in separate membranes of which only 3 were known.

Lee: Studying molecular networks is revolutionising single-cell analysis.

12 Does the Potential of ICT in Medical Care Require New Forms and Levels of Medical Training, Medical Data Collection, Storage, and Reproducibility, Clinical Logistics, Clinical Trials, and Patient Participation?

(The Discussions on Friday Morning)

Coordinator: Susana Vinga
Contributors: David Gilbert, Jochen Dreß, ...
Rapporteur: Ina Koch

For part of this session, we were joined by the participants of the security seminar that met in Dagstuhl for the same week.

Susana starts the discussion with a presentation of Carol Goble's ISMB 2013 slides[23] entitled *results may vary – reproducibility, open science and all that jazz.*

Robert: Recall the rather harsh Flexner Report from 1910. The Report called on American medical schools to enact higher admission and graduation standards, and to adhere strictly to the protocols of mainstream science in their teaching and research. Many American medical schools fell short of the standards advocated in the Flexner Report and, subsequent to its publication, nearly half of such schools merged or were closed outright.

Eric P: Government driven?

[23] Cf. http://www.slideshare.net/carolegoble/ismb2013-keynotecleangoble

Robert: This was driven by the American Medical Association (AMA): It determined the history of medical education in the US from 1910 to now. They developed a systems-based curriculum with a one-end examination after three years. The curricula in Germany and the US/Canada have four requirements including chemistry, physics, and biochemistry.

Walter: In Germany, a doctorate is not necessary.

Eric P: Students who grew up with computers might want to use it also during their study.

Alex: For many students, the doctorate just opens a second pathway.

Jochen: There should be more science included in the curriculum. This is not a question of technology, but of the way we learn.

Robert: According to Wikipedia[24], evidence-based medicine is the *conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients* or, more specifically *the use of mathematical estimates of the risk of benefit and harm, derived from high-quality research on population samples, to inform clinical decision-making in the diagnosis, investigation or management of individual patients.* So, it needs ICT-literate teachers and students.

Mark: Education of medical students is still mainly descriptive, we have to drive the market to adopt ICT.

Andreas: Yet, one should keep in mind that just understanding and using p-values is not necessarily proof of scientific soundness.

Robert: But, medical students need evidence-based probabilities.

Jonas: Informatics rotations, based on fellowships in informatics, are just starting to teach CS to medical students.

Mark: Formal CS is still more an exception in their education.

Jochen: Special courses are not necessary. Rather, ICT aiming to help patients should be integrated into the normal study.

Eric P: Put CS in different classes. According to a discussion with Jonas, there are three questions for the students.

Robert: We need to transform the CS-based lessons in a more structured way.

David: New forms are indeed needed to transmit the definition of terms: systems biology, systems medicine, healthcare systems, ...

Eric N: So, we need to present sort of a collection of new ways to think: thinking in terms of systems has changed it all.

Mark: So far, a healthcare system has never been engineered, they just came about. But now, we need an engineered healthcare system.

Ina: In chemistry, we have synthetic chemistry and analytic chemistry, and both profit from each other. So, synthetic biology and systems biology should be expected to also profit from each other.

David: Polyclinics and other *points of care* feature specialisations at different levels. In-service training is going to be important.What is going on in Australia, or in developing countries? According to a BBC news-channel report, data science and engineering are producing many

[24] Cf. http://en.wikipedia.org/wiki/Evidence-based_medicine#Assessing_the_teaching_of_evidence-based_medicine

different devices that should be taken account of, and a computer-science orientation will become decisive.

Jonas: How should young people be introduced to ICT? Can you start teaching from Kindergarden to 12?

David: In the UK, education is currently shying away from ICT.

Lena: Kids can learn very early. So, teach ICT to young kids.

Mark and David: We will also have to address computer-human interaction.

Mark: All countries have the same problems, e.g. ageing.

David: So, let us look at the global context, 1st/2nd /3rd world, including China.

Zhenbing: In China, all is very good in the big cities, but not so in the countryside.

Wolfgang: Sometimes, there is more flexible data handling in Africa, and technologies are coming back from the 3rd to the 2rd world.

Jonas: Which healthcare systems will arise in Africa and China?

Mark: Vaccination policies will be crucial.

Jochen: The changing climate brings new diseases.So, we just *need* to interact and share knowledge with 2nd and 3rd world countries.

David: There are many stakeholders, more than one can imagine – however, most are not yet involved.

Eric N: There was a public effort for introducing medical records – once pharmaceutical company got involved. Clinical trials organise stepwise. We have to build mechanism to make sharing easier.

David: Look at drugs: in the sports industry, in personalised medicine, in healthcare, ... Drug provision through the internet affects life style.

Teach formal methods: focus on process stuff (modeling, mining, . . .), workflow platforms, optimisation, event logs.

Wolfgang: Focus on the main process is necessary, you loose flexibility by optimisation.

Mark: Accurate time measurements, sensors, and measurement devices are increasingly becoming important for healthcare. They are already introduced in Scandinavia for personalised drug delivery.

Input from the other group: Early this year, the Newton institute ran a program on *Data Linkage and Anonymisation.*

Robert: How do we figure out risk factors? What is the risk of anonymisation?

David: How can you share data? Already for two data sets, there may be process problems, but also visualisation problems.

Mark: Two anonymised datasets put together reveals all.

Jochen: To avoid risks, we are not allowed to freely mix data in my office. If people need it, they first have to ask for permission.

Jonas: So, one needs to develop a 'culture' on how to deal with anonymity.

Jochen: For gene therapy and personalised medicine, this is indeed quite urgent as reproducibility is a basic principle for every scientific method (cf. Carol Goble's presentation). The Taverna workflow-management system should help turning knowledge into action. Yet, why are there so many retractions of scientific papers.

Data-sharing rules are adhered to by only 50% of journals. The G8 open data charter should provide motivations for science. Sharing results is not a goal in itself. And there have been reasons given why scientists don't want to share their data.

Jonas: Yet, scientific journals are the place where applications are disseminated.

Eric P: There are further problems with runtime environments. One needs to store old versions of e.g. matlab in order to run old code.

Walter: Most software institutions in Germany are funded by the taxpayer. Thus, their products are published as open source and full of bugs. Affects e.g. the biological community who uses this buggy software.

Mark: Greenway Medical Technologies is developing and offering apps that allow users to get access to data. I would like to see a design of an app platform that is ubiquitous – has access to all health data.

Jochen: Look at the Meta-manifesto from Carol Goble (also presented at ISMB 2013 in Berlin).

Robert: Universities want tech transfer via patents, this infringes on academic freedom.

Input from the other group: Algorithms are patentable but not 'copyrightable'; programs have copyright.

Eric P: W3 – specifications are royalty free. Critical algorithms can be used for certain purposes. HL7 is open to avoid the trouble associated with semi-closed knowledge based on software such as the Unified Medical Language System (UMLS), the SNOMED component.

Walter: 'Proprietariness' is a problem, I can't find out what is happening in the software. Is this relevant to our Manifesto? Should we include this as an ethical issue.

Andreas: Maple or Mathematica is proprietary, but we can publish the results obtained by using them.

Susana: What does 'free' mean?

Input from the other group: Brazil moved from Microsoft to the Linux Foundation. This was more expensive, but they got the source code. So, they can be certain that they will always be able to read their documents.

Walter: Software must also be validated.

Input from the other group: Software comes with proprietary formats – and the user is tied in to the software. There is a need for open formats. Open formats are beneficial – and have more business if they are used. The physics community has open access. In our community, editors make fortunes by deciding what will be published, even for PLoS One. Should the community self-publish?

Jonas: Here is Tim Berners-Lee's 5-star system for Linked Open Data:

- 1* = on the web,
- 2* = machine readable data,
- 3* = non-proprietary format,
- 4* = RDF standards,
- 5* = linked RDF.

There is an argument relating to scalability. How many stars should our data have for interoperability. Maximisation of secondary uses is the main motivation for adopting the 5-star system.

Andrea: We need to be clear about our motivation for open data.

Walter: Also ethical issues need to be taken into account.

Jonas: See e.g. the document from the Executive Office of the President: *Open Data Policy – Managing Information as an Asset*[25].

Eric P: Or the Yosemite Manifesto on RDF as a Universal Healthcare Exchange Language[26].

Walter: If patients read their record online without visiting their doctor, this may not be so good, the patients may misunderstand.

Mark: Only 15% of patients understand what they are told by medics. Our Dagstuhl Manifesto could either point to these other manifestos, or create its own set.

Mark: The Office of the National Coordinator for Health Information Technology (ONC) may have some meaningful use here[27].

13 Conclusions

As mentioned already, the discussions led to the following six claims:

(i) An *open-data policy* for healthcare related information systems is a fundamental and urgent imperative.

(ii) Following the *business-IT alignment* paradigm [13], healthcare should – on all levels – be supported by secure IT-platforms enabling clinical workflow engines that map healthcare related processes while integrating pertinent data-analysis, visualisation, and engineering tools.

(iii) Such platforms should also take full advantage of advances provided by *cloud services*, *pervasive computing ecosystems*, and the *semantic web*.

(iv) The *participatory potential* of the Web should be exploited to advance new forms of partnership in the healthcare environment.

(v) The acquisition of *ICT literacy* must become a required part of biomedical education.

(vi) Specifically in Germany, the Bundesnetzagentur should be encouraged to setting up a Working Group *Medizinische Netze* to explore options for a *Medical Cloud* within the German healthcare environment.

References

1 McKinsey Global Institute. *Big data: The next frontier for innovation, competition, and productivity.* http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (2011).

2 Ixchel M. Faniel and Ann Zimmerman.*Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse*. International Journal of Digital Curation, 6.1 (2011):58–69.

3 Steve Baunach. *Three Vs of Big Data: Volume, Velocity, Variety*. http://www.datacenterknowledge.com/archives/2012/03/08/three-vs-of-big-data-volume-velocity-variety/ (2012), see also http://dashburst.com/infographic/big-data-volume-variety-velocity/.

[25] Cf. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

[26] Cf. http://goo.gl/mBUrZ

[27] cf. http://www.healthit.gov/policy-researchers-implementers/meaningful-use-stage-2, also check http://www.ted.com/talks/dave_debronkart_meet_e_patient_dave.html.

Figure 10 The *word cloud* for the combined rapports of all regular sessions and evening presentations, reflecting the central focus on open data, nucleating discussions about medical information systems, and disease modeling.

4 James McCusker and Deborah L. McGuinness. *owl:sameAs Considered Harmful to Provenance*. The Tetherless World Constellation http://www.slideshare.net/jpmccusker/owlsameas-considered-harmful-to-provenance (2010).

5 P. J. Tonellato, J. M. Crawford, M. S. Boguski, J. E. Saffitz. *A national agenda for the future of pathology in personalised medicine: report of the proceedings of a meeting at the Banbury Conference Center on genome-era pathology, precision diagnostics, and preemptive care: a stakeholder summit.* Am J Clin Pathol. 135.5(2011):668–72. doi:10.1309/AJCP9GDNLWB4GACI.

6 Hector Gonzalez et al. *Google Fusion Tables: Web–Centered Data Management and Collaboration.* Presented at SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA.

7 Hans Lehrach et al. http://www.itfom.eu, MPI for Molecular Genetics, Berlin, Germany (2012).

8 National Research Council (US). Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* National Academies Press, Washington D.C., USA, 2011.

9 Leroy Hood and Stephen H Friend. *Predictive, personalised, preventive, participatory (P4) cancer medicine.* Nature Reviews Clinical Oncology 8.3 (2011):184–187.

10 Leroy Hood et al. *Revolutionizing medicine in the 21st century through systems approaches* https://www.systemsbiology.org/blog-topics/p4-medicine. Biotechnology Journal 7.8 (2012):992–1001. DOI:10.1002/biot.201100306.

11 https://www.systemsbiology.org/sites/default/files/Hood_P4.pdf

12 HE Bin et al. *Grand challenges in interfacing engineering with life sciences and medicine.* IEEE Trans Biomed Eng. 60 (2013):589–98.

13 Wim van Grembergen and Steven De Haes. *Enterprise Governance of IT: Achieving Strategic Alignment and Value.* Springer, New York Heidelberg Dordrecht London (2009).

Participants

- Jonas S. Almeida
University of Alabama at Birmingham, US
- Bernhard Balkenhol
infinity^3 GmbH – Gütersloh, DE
- Mark Braunstein
Georgia Tech – Atlanta, US
- Robert Burk
Yeshiva Univ. – New York, US
- Stefan Decker
National University of Ireland – Galway, IE
- Helena F. Deus
Foundation Medicine, Inc. – Cambridge, US
- Andreas Dress
Shanghai Institutes for Biological Sciences, CN & infinity^3, DE
- Jochen Dreß
DIMDI – Köln, DE
- David Gilbert
Brunel University, GB
- Anja Hennemuth
Fraunhofer MEVIS – Bremen, DE
- Scott Kahn
Illumina – San Diego, US
- Ina Koch
Goethe-Universität Frankfurt am Main, DE
- Titus Kühne
Deutsches Herzzentrum – Berlin, DE
- Hans Lehrach
MPI für Molekulare Genetik – Berlin, DE
- Pietro Lio’
University of Cambridge, GB
- Markus Löffler
Universität Leipzig, DE
- Wolfgang Maaß
Universität des Saarlandes, DE
- Klaus Maisinger
Illumina – United Kingdom, GB
- Eric Neumann
Foundation Medicine, Inc. – Cambridge, US
- Laxmi Parida
IBM TJ Watson Res. Center – Yorktown Heights, US
- Alex Pothen
Purdue University, US
- Eric Prud’hommeaux
MIT, US
- Joel Saltz
Emory University, US
- Walter Schubert
Universität Magdeburg, DE
- Andrea Splendiani
DERI – Galway, IE
- Marc Van Regenmortel
IREBS – Illkirch, FR
- Susana Vinga
Technical Univ. – Lisboa, PT
- Peter Walden
Charité – Berlin, DE
- Zhenbing Zeng
East China Normal University – Shanghai, CN