Foresight exercise report
Computational Science (Lastu)
Academy Programme

ACADEMY OF FINLAND

Contents

Foresight exercise report

Computational Science Academy Programme

Academy of Finland, February 2016

Editors:
Steering Group and Tommi Laitinen and Tuomas Katajarinne

1. INTRODUCTION

1.1. Brief summary

This report documents the objectives, process and results of a foresight exercise performed as part of the Academy of Finland's Computational Science Academy Programme (see www.aka.fi/lastu). The foresight exercise included a mapping and an analysis of important trends in the field of or related to computational science. The results of the exercise contain both the results of the group works provided as brief trend analyses (Appendix I) and the perception of the Steering Group (Section 2).

1.2. Objectives

The foresight exercise had three stated objectives. The first objective was to make the underlying trends visible to researchers in the field of computational science so that the trends could be taken appropriately into account by the researchers in their work. The second objective was to network researchers with each other and with other people attending the exercise, such as company representatives. The third objective was to generate data to be used by the Academy of Finland and by other organisations active in science policy.

1.3. Process of the foresight exercise

The foresight exercise was conducted as a four-step process in Autumn 2015.

Step 1: A questionnaire was sent to the principal investigators of the research projects of the programme. The questionnaire contained the following three questions:

1) Mention 1–3 important future trends in the field of computational science.
2) Mention 1–3 trends that have given rise to much discussion but that are not, in your opinion, significant future trends.
3) Mention 1–3 silent trends that may be unlikely but very significant, if realised.

Questions 1, 2 and 3 received 36, 17 and 18 responses, respectively. The responses were then discussed by the programme's Steering Group, which finally selected six trends for a more detailed analysis. The chosen trends were digital humanities, open data and open science, machine learning, big data, future supercomputing and artificial intelligence.

Step 2: The six trends were discussed in groups of 3–4 people (mostly professors and researchers in the field). The programme managers facilitated the discussion. The purpose was to identify the overall impacts of the trend on society and science over a period of 5 to 30, even 50 years. As a result of Step 2, the chair of each group formed a "trend tree" poster from the thoughts raised during the group discussion.

Step 3: During the closing seminar of the programme on 23 November 2015, the trend tree poster and the thoughts raised in Step 2 were exposed for discussion by a wider audience. Some 50 people, mainly researchers and professors in the field, attended the discussion in groups of 6–8 people, and the trend trees were further elaborated. Finally, the expanded trend trees were presented to all seminar attendants.

Step 4: The chairs of the groups delivered the final outcomes of the discussion in the form of a small report. The reports are presented in Appendix I.

2. PERCEPTION OF THE STEERING GROUP

One of the objectives of the foresight exercise was to generate information to be used by the Academy of Finland and by other organisations active in science policy. This section contains the key findings from the foresight exercise as seen by the Steering Group of the programme.

The findings presented here can affect science policy in many ways. The findings can help identify topics for new Academy Programmes or new programmes run by the Strategic Research Council. The findings can also be utilised by the research councils of the Academy of Finland, for example, to identify topics for targeted funding calls.

The perception of the Steering Group of the group reports is presented below.

Computational intelligence (CI)
The breakthrough of computational intelligence (CI) has been anticipated for decades, but only now is it about to realise. CI as a term is closely related to machine learning, owing to the emerging trends of digitalisation and big data and the increasing availability of large digital datasets.

CI has several applications in various fields of science such as biology, medicine, materials science, forestry and economics. However, the use of CI is in its infancy and it has only been marginally exploited so far. To support the renewal of science, there is a need to increase knowledge of the potential of CI in society and to explore its use in new applications realised through multidisciplinary research. More interaction is needed between machine learning experts and those who would like to apply the ideas. The potential risks should be also analysed, since the systems relying on CI must be robust and there are many ethical problems involved.

Digital humanities
Another insight from the results of the foresight workshop is that the potential for new scientific openings and important breakthroughs with computational methods seems to lie in social and human sciences. Computational methods are well used in the natural sciences such as physics and biology, where the datasets may be large but are typically also well structured. In social and human sciences, on the other hand, the existing big data are typically unstructured (e.g. shopping behaviour, public transport usage, social media usage). The computational problems are very complicated, such as in the case of analysing natural languages. Developing computational methods for analysing such unstructured data opens a wide range of new possibilities for novel multidisciplinary scientific work.

Cyber security and ethics
Security and safety are key elements of everyday life, especially in the field of computational sciences. Cyber security is under threat every day. Banks and government offices are attacked frequently, identities are stolen, even power plants and airfields are at risk, not to mention all the risks facing us with the Internet of Things. There are also more subtle risks, and ethical questions, related to companies and authorities influencing what information we are given on the internet, or computer systems making more and more decisions on behalf of users in cars and planes, for instance.

The importance of artificial intelligence and computational science is increasing in our lives, while the understanding of computational methodology is not. There is a huge demand for computational sciences to create methods and information for enhancing the ethical use of computing systems and helping decision-making processes in the open society. At the same time, we must not forget about making current systems safer and more secure by computational means.

Supercomputing

Supercomputing refers to computers that are among the fastest in the world, which also means that they have a large size and a high power consumption. Complex systems are growing in importance. Supercomputers have application, for example, in simulation and optimisation of complex systems in science, engineering and medicine. In meteorology, supercomputers improve long-term forecasts, and in communications they are used in centralised computing in data centres. Moore's law of charge-based electronics is expected to come to an end around 2020–2022, which implies that the energy efficiency and size of electronics will not improve as fast as before. An interesting possibility for special-purpose computing is quantum computers. The energy efficiency may be a problem, but quantum computers may have important applications in special areas such as cryptography.

3. APPENDIX I

3.1. ARTIFICIAL INTELLIGENCE

3.1.1. Participants of the working group

- Tapani Raiko, Assistant professor, Aalto University, School of Science (Chair)
- Juho Rousu, Associate professor, Aalto University, School of Science
- Alexander Jung, Assistant professor, Aalto University, School of Science
- Jouko Poutanen, Software Client Architect and Country Technical Leader, IBM Finland

3.1.2. Introduction

Artificial Intelligence (AI) is a rising trend in science. Recent developments in understanding the natural environment and interacting with people (computer vision, speech recognition, natural language processing, haptics) are opening up lots of new possibilities.

We decided to look at the future of AI with two axes: One for time, and one representing positive and negative aspects. As an example of a negative aspect, powerful AI systems might concentrate power and influence to very few people that are controlling the swarms of journalist AIs or military robots. On the positive side AI combined with other technologies has tremendous potential to have positive impact on economies, organisations and individuals. We see that all industries will be affected by AI.

AI progress will accelerate because now the technical enablers have developed to new level: data, algorithms and computing. AI should maybe approached in a broader context, where the AI is a core technology, but closely dependent of big data management capabilities and extended with natural language processing (NLP), as well as touch and visual interface technologies to enable more natural interaction between humans. This broader construct is sometimes called cognitive computing.

3.1.3. Identified new focus areas and challenges

Near term

One key use for AI is to augment human capabilities in decision making, for example with help of cognitive assistants. We predict that a majority of Finnish consumers will utilize AI based or cognitive systems in their daily lives within the next 5-10 years. We also foresee that much of the AI technology will be ambiently embedded into the human environment, operating without explicit commands from the users. Indeed, the major benefit for the user will be the freedom from issuing explicit commands. Although, physical autonomous robots will take a large role in the future, we agree that most AI applications will be virtual.

A clear area where rapid development will take place is information retrieval and access. Computers will become much better in finding information relevant to the user and in bringing correct information to the user's attention in correct time. These technologies need to become clever in gathering implicit feedback from the user so that the models can be improved without user needing to give explicit feedback. The development will be the fastest in the areas of life where big open data is available.

A concrete, major near-term development is that of autonomous cars. This might change the society in various ways, such as people owning less cars and relying more on autonomous taxis, leading into less demand for parking places and less demand for new cars.

Near term research is needed to study how AI and cognitive computing can be utilized for teaching. What areas would potentially benefit from AI enabled and accelerated learning? What kind of content/data sets are required and what instances could produce them? Could there be some kind of AI / cognitive teachers or assistants? Potentially related technologies are already used in online gaming, where a player's skill level is automatically assessed and opponents with matching skill levels are sought for. Indeed, gamification is already a major trend in education - perhaps not yet fully embraced as an AI problem.

As the studied AI problems get more complex, the need for AI systems that can learn new knowledge domains autonomously grows. The area known as unsupervised learning is a challenging topic where Finland has a long strong tradition. Future AI systems will adopt greater unsupervised learning, which will require much less human interaction in the system training process. Lots of research focus is needed to enable breakthroughs in this area.

Societal Challenges

Education – A key challenge for the advancement of AI and cognitive computing will be the availability of skilled humans. Advancing AI capabilities and implementing systems require unique skill sets, such as those of machine learning experts and natural language processing scientists. These skills are currently in limited supply and high demand. Planning of appropriate curriculums and their implementation need proactive planning that should begin sooner than later. The current initiatives to integrate computer programming in to comprehensive and high schools is a welcomed, however, probably the realignment of mathematics curriculum should be undertaken, giving a larger emphasis on linear algebra, statistics and algorithms - the building blocks of modern AI systems.

Employment – Automation making jobs redundant has been going on since the industrial revolution and will continue. In particular, administrative jobs will greatly diminish as papers do not need to be moved around as much as thus far. So far there have been other new kind of jobs opening to compensate, but it is not clear whether this will continue. This might lead into massive unemployment, while most of the jobs are automated or be made significantly less time-consuming so that fewer people are needed to accomplish them. This is not as bad as it may sound, since all the work gets done. However, societal impacts such as fair distribution of wealth and the need of everyone to feel important and valued,need to be considered. It is likely that work will not maintain its current role as a "measure of human's self-worth" in the future.

Security and privacy – If all surveillance data is analyzed by facial recognition software etc. it will affect the security and privacy in a radical way. Automation of surveillance, media, and even military might lead into unwanted concentration of power and influence to those who control the automation.

In shorter term the concerns of privacy and security might also hinder the advancement of AI, if the reaction to the issues will be that of locking away data or making integration of different data sources on the level of individuals impossible. AI research needs to develop technologies for anonymization and privacy-preserving machine learning to make technology development feasible and acceptable for the general public.

Health & wellbeing, addictions – Entertainment (including social media), games, and porn are addictive already, but virtual reality and interactive AI systems have the potential to make them even more so. Digitalization in general and AI in particular, allows people to be more and more sedentary, when more and more of the daily tasks can be accomplished virtually. AI technology should be developed that frees gives incentives to be physically active. This is also linked to addiction: games and entertainment technology that make people do physical exercise "on the side" should be developed.

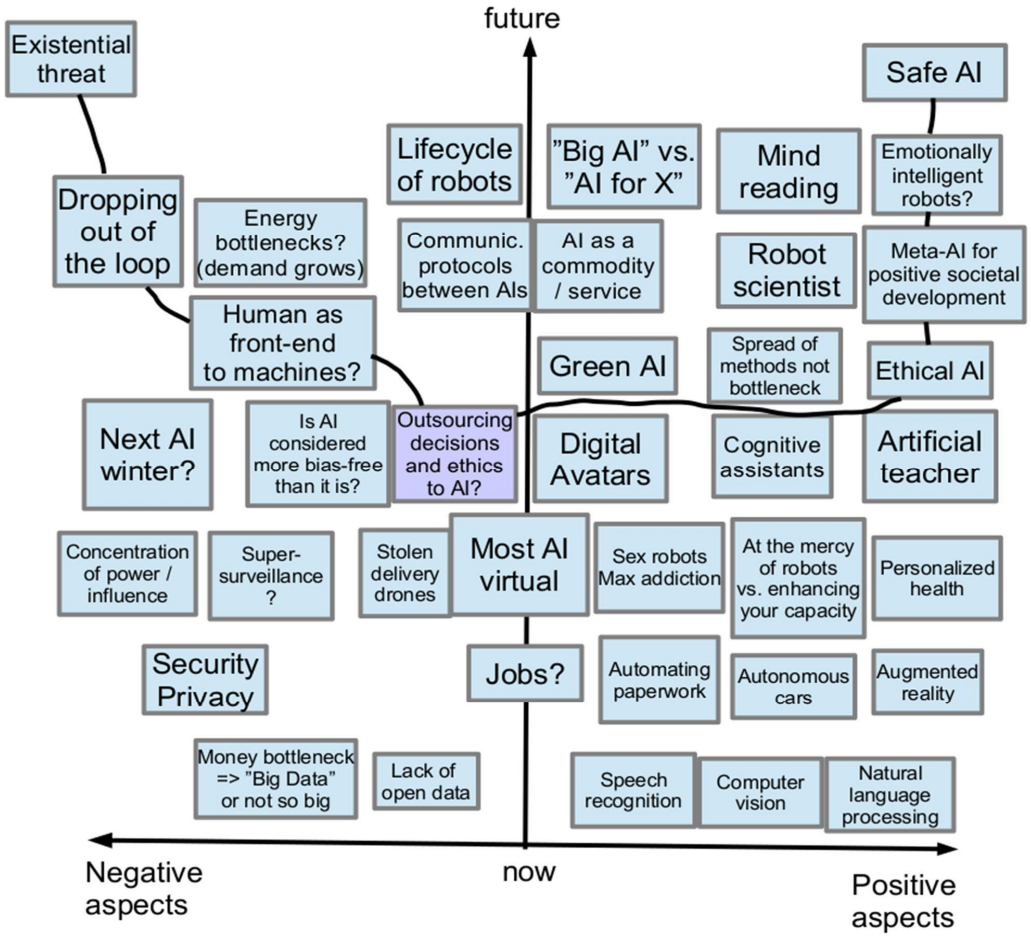
Ethics

At some point people might start to outsource decisions and ethics to AI. AI might be considered more fair and bias-free than its human counterparts. This has the potential to rid societies of corruption, inequality and conflicts. However, there is are risks of considering AI more bias-free than it is, and in the long term even humans dropping out of the loop. This emphasizes the longterm value of digital humanities.

Long term

Neuromorphic computing and quantum computing have tremendous potential to accelerate AI based technologies' output. Few companies are advancing in producing such new types of computers, but it is difficult to predict commercial and wide spread availability. With current development and traditional hardware, energy required for computing might become a bottleneck, if the demand for AI computing grows faster than computing efficiency.

We feel that even in the longer term, there will not be one big AI, but instead lots of solutions of the type "AI for X". As these systems become widely used, effective communication protocols between different AI systems becomes important. It should enabling autonomous communication, negotiation and even teaching between AIs. We feel that this is one area that has been so far largely neglected in research so far.



3.2. BIG DATA

3.2.1. Participants of the working group

- Jukka Heikkonen, Professor, University of Turku
- Tuomas Häme, Professor, VTT Technical Research Centre of Finland
- Pertti Koukkari, Professor, VTT Technical Research Center of Finland
- Petri Myllymäki, Professor, University of Helsinki

3.2.2. Introduction

Big data can be characterized as containing structured and unstructured data having the following four dimensions (4Vs): **volume, velocity, variety and variable veracity**. Big data does not commonly refer to any specific quantity, instead data may be called big data when it is extraordinary in one or multiple of the above dimensions: extreme volume, rapidly streamed and processed, wide variety of heterogeneous types of data with uncertainties. The following computational challenges in big data processing and especially in predictive and prescriptive analytics can be found: 1) Understanding the heterogeneity and commonality of high dimensional and large sample size data to find relevant and coherent information among multiple data subpopulations; 2) Ability to collect and handle data from multiple sources at different spatiotemporal resolutions; 3) Ability to handle missing data and errors in measurements, spurious correlations and incidental endogeneity without dramatic drop of performance; 4) Use of on-line feedback mechanisms such as crowdsourcing to improve the information content of the data in order to produce more accurate predictions; 5) Dealing with scalability and storage bottlenecks and computational costs especially in online real time use; 6) Dealing with performance estimation metrics of the big data models especially when dealing spatiotemporally autocorrelated data.

In the working group meeting (Heikkonen, Häme, Koukkari, Myllymäki) we decided to divide big data into four distinct streams each having their own state-of-the-art to mid and far future trends up to 50 years ahead. These streams with their corresponding trends are presented in the attached Figure 1.

The first stream "*Data sources*" covers new emerging data gathering technologies. In the near future advances in remote sensing, unmanned autonomous vehicles and internet of things are expected to produce growing number of data. For a longer perspective the society will become surrounded by multiple types of sensors and smart devices providing an excessive amount of data. Lifelogging and human augmentation will be normal part of our daily life. There will be also growing interest for symbiotic systems, a type of personal assistants that adapt to their user to support the human's life. In addition to data collection there will be urgent needs in data management such as data standardization, certification, privacy, security and legislative issues.

"*Platforms*" is closely related to big data processing infrastructure consisting of software and hardware solutions. In the future big data will be based on open-source data-processing platforms supporting highly distributed architectures and new levels of memory and processing power, such as graphical processing units today and quantum computers in the farther future. The increasing use of energy needed for data processing and rare materials used in processing hardware will raise the importance of sustainable computing awareness. The platform level would enable the co-operation of different data producers, research and methodology parties. It would also automate the extraction and management of the derived features from the primary ones. It would abstract away several case specific aspects like data queries. *"Processing"* deals with trends in big data analytics. It is clear that citizen data science and self-service delivery will emerge from open data and algorithms. This sets new challenges for education and research to prevent erroneous conclusions and causalities from starting to dominate public discussion due to limited understanding of the properties of the computational methodologies and the information context of the data itself. Prespective analytics will be needed. It is a type of predictive analytics but goes beyond by giving the likely outcome of each decision and recommending one or more courses of action. In the farther future to achieve human brain type of performance especially in sensory data processing we need human brain type of representations to understand semantic conceptions of information. This will help especially when building of symbiotic systems.

"Applications" stream deals with big data application trends. There will be numerous trends in a wide varying of application sectors and it is impossible to cover all of them. One of the most important application sector will definitely be health and well-being as already today. However, to stay in a more general level we can identify the future needs for symbiotic systems that help human's in their daily life. In the far future robots/droids are making inventions based on the existing knowledge which also highlights the needs for semantic code; what the intelligent machines can do and what is illegal and will be penalized.

3.2.3. Identified new fields of science

In the discussion with the working group and at the Academy seminar three important emerging fields of science were identified as outlined below. All of them are very interdisciplinary in nature.

1. Symbiotic systems

Nowadays single assistant type of services, like Google, have important role in our daily life. In the foreseeable future we will be go well beyond single services in a form of symbiotic systems. The type of systems will be used as personal assistants in a variety of roles and environments. There will be varying forms of symbiotic systems appearances ranging from droids to other types of embedded systems. To build intelligent, adaptive, and assistive systems that will change our relation to the world and to each other and can overcome many our limitations, related to the memory, understanding and physical power, we need scientific progress in many fields of science and close collaboration activities between interdisciplinary fields.

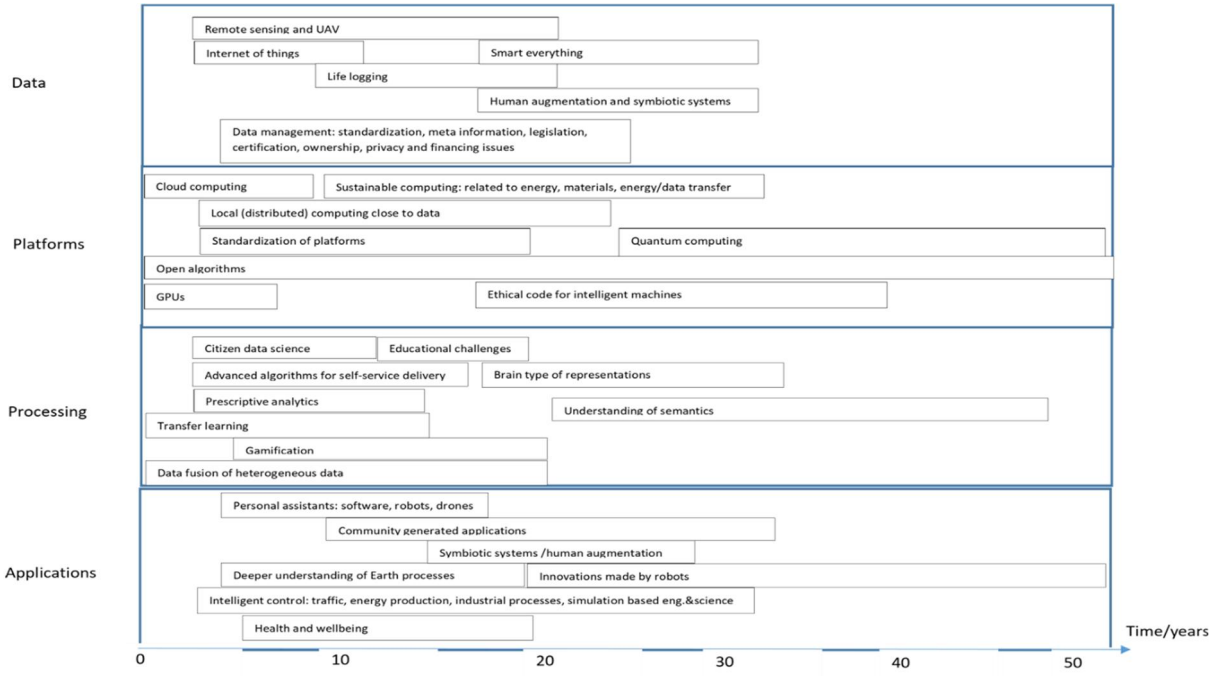
2. Ethics in computing and intelligent systems

Computers and information processing have already today a central role in today's society. In the future when emerging new computational technologies and intelligent smart systems will have a symbiotic type of role in our daily life, we are facing many ethical issues related to, e.g. social context of computation, intellectual property and privacy, risks and liabilities of systems, and computer crime. It should not be the responsibility of manufactures to define the ethical code for computations, on the contrary it should be the role of political decision makers.

3. Big data management

Efficient use of big data in the future requires data and platform standardizations and certifications. Especially meta data should be available in standard form to simplify the data usage when considering the representativeness and timeliness of the data for the current application. Big data management overlaps naturally with ethics in computing and intelligent systems related to the data ownership, privacy and legislative issues.

Figure 1. Big data research and development trends during the next 50 years



3.3. MACHINE LEARNING AND MATERIAL SCIENCE

3.3.1. Participants of the working group

- Kari Laasonen, professor, Aalto University, School of Chemical technology
- Ilpo Vattulainen, professor, Tampere University of Technology
- Jukka Corander, professor, University of Helsinki
- Sampsa Hautaniemi, professor, University of Helsinki, Faculty of Medicine

3.3.2. Introduction

We first discussed of machine learning (ML) and it applications to materials science in a group listed above. We agreed that ML is a very useful tool in various fields of science but it has not been used much outside of bioinformatics. In the following discussion we do not use much term “material science” since most of our observations are not limited to materials science. We also identified some problems why it will not be easy to use the ML in our fields of interest. We list some of these problems under the ‘Challenges’ below. Next we discussed the time development of this field and we collect our ideas afterwards with emails. The time-line ideas are in Appendix 2 in this document. We made a compact hand-out from our time-line ideas as background material for the Learning Cafe discussion. The time for discussion in the Learning Cafe was rather short so the discussion was not much controlled. The ideas from the Learning Cafe are collected to Appendix 1. To keep the document short we collected the key points to few recommendations.

3.3.3. Overview

The ML as an interesting tool to help several fields of sciences was clearly recognized. There is a lot of enthusiasm of ML but not very much knowledge outside the ML experts. Experts working in ML have done already quite a bit of collaboration with various research groups but mostly in a field of bioinformatics. So several application areas are untouched.

One clear message that came out from the Learning Cafe, was that there is a huge amount of knowledge and (physical) models in various fields that can be combined with ML. This knowledge should not be ignored. The deep knowledge from a given research field should be complemented with good and relevant data. This combination should be used to build field specific ML projects. The quality of data was raised and the different knowledge levels of ML were discussed from several points of view.

Also the ML field is developing rapidly. The increase of computer capacities and larger databases will help the usage of new and more complex models. So called Deep Learning methods are very interesting. Also likelihood-free reasoning will open new possibilities. We strongly believe that new problems, new data and new theoretical ideas will push the ML development further.

Time-line very shortly:
Short term: Get the ML experts and other scientist together. Formulate relevant research problems. Trialand-error phase. Some mistakes, some success. Teaching basics of ML to material scientists.

Medium term: More specific research projects, better mutual understanding. First new projects where ML really helps research projects where it has not been used before.

Long term: ML has become a natural and important part of several fields of science. We will have several significantly better ML methods and much more computational capacity.

3.3.4. Challenges

In many projects the **data quality** is a big problem. The data may exist but the researchers cannot use it, either due to privacy reasons or due to cost. Often the data is poor, it can be non-systematic, sparse, hard to find or expensive.

Knowledge gap: Even though the ML expertise exist in some departments, like mathematics and computer science, researchers working in other fields typically knows very little to nothing of ML. To bridge this gap is not easy. The ML experts need (rather) well posed questions and good data to be interested of collaboration projects. These rarely exist. Thus a lot of preparatory work is needed. In long run ML methods should be taught in other science programs. Clearly the ML concepts are easier to understand for persons that have physical science background but many participants saw significant potential of ML in fields far from physics and chemistry. These include linguistic, forest research etc.

The knowledge in different fields is not easy to utilize in ML but this also brings probably the most interesting future applications to ML.

3.3.5. Recommendations

We should go towards the Open Data ideas. We need to find means to have the data needed to train the ML algorithms. In the case of personal sensitive data we need think how the data should be modified that it cannot be connected to an individual person and it still would be useful for ML. Commercial data will be a problem since often it is too expensive for researchers. A lot of good quality data can be produced using simulation of physical models (like ab initio calculations, empirical force fields, weather models, etc.), but this data need to be tested towards real experimental data.

We need to get the ML experts and other scientist together. It would be very useful to have a funding program to do co-operative research with ML experts and other scientist. Bluntly, without money the cooperative projects will develop very slowly.

Utilize the vast knowledge accumulated in different research field to make tailored ML models. This will not be easy or fast but, in our opinion, it will be definitely worth of trying. Also new challenges and new data will push the ML research further.

3.4. DIGITAL HUMANITIES

3.4.1. Participants of the working group

- Erkki Oja, Professor emeritus, Aalto University, School of Science (Chair)
- Jussi Pakkasvirta, Professor, University of Helsinki, Faculty of Social Sciences
- Mika Pantzar, Research Professor, University of Helsinki, Consumer Society Research Centre

3.4.2. Introduction

Digital humanities is usually defined as a field of science in the intersection of humanities and computing. Humanities here entails things such as history, philosophy, linguistics, literature, art, archeology, music, cultural studies, and all social sciences. Computing means things such as hypertext-hypermedia, data visualization, information retrieval, data mining, statistics, text mining, and digital publishing. Some of these methodologies have been used in humanities for a long time, but notably the recent trend in big data analytics has brought a boost also in humanities research. For example, when more and more of the historic materials and texts are brought into easily accessible digital form, and computerized techniques can be used to help the researcher in the analysis of the materials, old hypotheses can be tested and new ones discovered. Another example is social network analysis: social media offer huge quantities of user-generated content from which social patterns and relations can be extracted by data mining tools. It is obvious that this offers big opportunities for humanities research.

In the preliminary brainstorming work by the working group (Oja, Pakkasvirta, Pantzar) we decided to divide the progress from present state-of-the-art to mid-term future (5 to 10 years) and further to far future (20-30 years) into three distinct streams. They are schematically presented in the attached Table 1 (see end of this report). The first stream, "*Theory and facilitating technologies*" presents, on one hand, basic digital techniques that are needed to advance the analysis of humanities materials. These include big data management and analysis in general as well as artificial intelligence and machine learning techniques. On the other hand, this stream includes basic questions from the humanities side, such as developing new educational platforms, legal and regulation issues for data such as ownership, security, and privacy, and finally structures in politics and education to support the needed change towards data- and knowledge-based society.

The second stream, "*Material and Methods*", deals with the central technologies through which humanities material in digital form can be processed and analyzed to help the humanities researcher reduce the needed manual labor. The central techniques are text analysis, multimedia analysis, and social media analysis. Through them, the stream advances to more "intelligent" automated analysis – always to be used just as tools for the human researcher. There are also more practical technical goals visible in this stream like advanced search engines perhaps with speech interfaces and natural language understanding. One important aspect here is "citizen science", meaning that with open data and easily accessible efficient analysis techniques, even laypeople could draw their own conclusions and make intelligible choices e.g. on social questions.

The third stream, "*Language and Communication*", targets some of the "grand challenges" in humanities research. Eventually, the outcomes of digital humanities must be communicated to researchers and citizens in understandable form. Some of these questions are social separation caused by Web communities with widely diverging views of the world; behavioral economics in which the starting point is true and measurable human behavior instead of

game-theoretic theories of rational market participants; understanding social change and human behavior by looking at the enormous data available over the Internet; and finally being able to better forecast and thus influence social and political effects.

3.4.3. Identified new fields of science

As the title “digital humanities” indicates, this is a inter-disciplinary effort of two separate fields: humanities and social sciences on one hand, and computational sciences on the other hand. In the Academy of Finland, this falls primarily into two research councils: the council of natural sciences and engineering, and the council of culture and society. The emphasis of digital humanities on the other two research councils (health, and biosciences and environment) is clearly smaller even if not insignificant.

Some of the research problems in Table 1 are already included in the present and near future Academy programs such as “the Human Mind”, “the Future of Learning, Knowledge, and Skills”, and especially the starting program “Digital Humanities”. In the latter program, the goal is to “address novel methods and techniques in which digital technology and state-of-the-art computational science methods are used for collecting, managing and analyzing data in humanities and social sciences research as well as for modeling humanities and social science phenomena”. This goal was also at the heart of the discussions leading to Table 1.

In the discussions within the working group and at the Academy seminar, several new or important possibly emerging fields of science were identified. Four of them are outlined below.

A. Towards citizen science

This program would develop technologies and skills that facilitate nonscientists to gather, use, and interpret collections of data. At the simplest level of citizen science is crowdsourcing, where a large group of people in an online community provide services or content to a given task, often on a voluntary basis. At a higher level, the laypeople take a more active role in participating in data collection and problem definition, possibly in collaboration with the professional scientists in the field. This is expected to establish a wholly new scientific culture. The educational impact can be large, as the volunteers acquire some understanding of scientific principles which leads to understanding generic evidence-based decision making.

Citizen science projects are not easy to design. There are several computing technologies needed, which deal with tools for data collection and data curation, as well as easy-to-use and welltested tools for data analytics. On the humanities side, research is needed on what kind of training the volunteers need in order to ensure consistency in data collection and analysis. Another wider research question is how to link research findings with management and decision making, as well as the effect of citizen science on societal norms and rules.

B. Language technology

In Table 1, stream 2, some central trends are social media analysis, multimedia analysis, text analysis, and content and topic analysis. The central research tools can be collected under language technology. This means methods in natural language processing, computational linguistics, and speech technology. One of the goals is to construct more human-friendly human-computer interfaces, using natural text or speech. Another goal is to facilitate the human-human communication by automatic language translation between different languages or within the same language between people with widely different backgrounds. Yet another goal is content and topic analysis, which is the key in turning raw text or speech documents, too extensive for manual processing, into collective knowledge.

In the stream 3, “Theory and facilitating technologies”, there are several growing trends, which however are not as interdisciplinary as the above two. One is

C. Affective computing and social robotics

In the research on artificial intelligence, the final goal is “human-like intelligence” which would be highly useful in various decision-making scenarios but also has its obvious drawbacks. Towards this goal, there is growing research on affective computing, meaning e.g. human-computer interfaces that would somehow take into account the emotional status of the human user. For social robots assisting people in tasks like household work or helping the elderly, this kind of behavior is very desirable. A social robot is supposed to interact and communicate with humans by following the social rules associated with its role and tasks, leading to complete new forms of social interactions. This raises several difficult research problems at the interface of computing and humanities. This research must be a joint effort of computer scientists and cognitive scientists.

Another trend in stream 3 that almost inevitably will grow in importance is

D. Big data regulation

Big data is expected to have a deep impact on society and business. In order for the data to be acceptable, the privacy of individuals must be protected. It is well known that by integrating data from various sources, highly sensitive results can be produced which may violate privacy laws. Data is also international. The question is how to balance the privacy laws and legal interests with the advantages that data analysis can bring. Another question is the ownership of data, related to the question of open data and innovative data sharing. In this research, expertise on both data science and sociology/political science/law/economy/ethics is needed.

Table 1.

FUTURE

GENUINE (DIGITAL) HUMANITIES WITH FULL INTEGRTION

-Understanding, explaining, forecasting, influencing -Understanding social change -Understanding human behaviour -Civilization in the digital world -Effects of digitalization on human brains and evolution -History -Regulation, democracy	-Unconstrained speech rec. -Intelligent Wed + search engines -Text, multimedia, social media understanding -Decision support -Citizen science -Open data and access -Democracy and public sphere	-Ubiquitous data -Human-like intell. -Structures, politics, state, education to support change

5-10 YEARS

CROSS-DISCIPLINARY TEAMS AND PROJECTS

-Developed game simulations -Behavioral economics -Unified vocabulary -New interfaces -Reading, writing in change -Social separation prevention -Meetings, games, play -Second Life	-Social media analysis -Multimedia analysis inc. speech -Text analysis, mining, machine translation, search engines -Content and topic analysis -Expanding context -Hypothesis testing in social science for big data	-Bigger data, knowledge management -Regulation for data: ownership, privacy -Artificial intelligence -Edu platforms -Big Data, cloud, etc.
LANGUAGE AND COMMUNICATION	***MATERIALS AND METHODS***	***THEORY AND FACILITATING TECHNOLOGIES***

PRESENT

3.5. SUPERCOMPUTING

3.5.1. Participants of the working group

- Dr. Susan Leerink, Aalto University (Chair)
- Dr. Timo Kiviniemi, Aalto University
- Dr. Ronan Rochford, Aalto University
- Mr. Jari Varje, Aalto University

3.5.2. Introduction

For a long time the challenges in the field of supercomputing were focusing mainly on building bigger machines with faster processors, leading to more computations per second. Now that exascale level computing is around the corner, the focus is slowly shifting towards solving a broader variety of problems, not necessarily only related to increasing computational power. For instance, the energy efficiency of the new computer clusters will become a big priority for future systems as one would need the power of a small city to operate them. Furthermore, as the amount of data increases rapidly with the size of the problems, more focus on effective memory handling, memory storage and (I/O) communication protocols is required in order to boost system performance. Computer networks will not get much faster as the chip clock speeds will not be increasing much. Instead a lot of speedup can be gained from reducing the data-access time by more efficient data handling and hierarchy of storage and memory. Many fields of research that use high performance computing will benefit a lot from more advanced memory, storage and communication methods as the problems at hand are often not optimized to be solved with only an increase in floating point operations.

So far the users of super computers are quite evenly spread over public/government users and users at private companies. As it is highly likely that computing will become more integrated in daily life (think for example about the self driving car or the use of robotics in daily life) there is a substantial chance that the amount of users of high performance computing in the private sector will become the dominant costumer. If this will indeed be the case, the architecture of future super computer systems will most likely be optimized to deal with specific problems as a private costumer usually has a very specific problem to solve. It will be interesting to see how the research field of machine learning could contribute to designing, building and programming future super computers to optimize the computer architecture for specific problems.

A rapidly emerging field of science with possible future applications to super computers is the field of quantum computing. Although the development of quantum computers is still in a early scientific stage, the potential of using quantum mechanical phenomena to perform operations on data has huge potential in speedup compared to some of the best known classical algorithms. Many national governments, military agencies and private companies are funding research in quantum computing and the first quantum computing company, D- wave systems, selling computational resources on a quantum computer is a fact. Nevertheless there are still a number of technical challenges to overcome before quantum computers will be able to enter the marked on a broad scale, including but not limited to the scalability and readability of qubits, optimization of the decoherence time and the initialization of qubits. At the same time discoveries of new materials via molecular design or the development of efficient algorithms for classical computer architectures could boost the conventional computer market, making the need for quantum computers to speedup computations less critical.

3.6. OPEN SCIENCE

3.6.1. Participants of the working group

Pirjo-Leena Forsström, PhD, Development Director (chair)
Antti Honkela, D.Sc. (Tech.), Academy Research Fellow, University of Helsinki
Samuli Ollila, PhD, Researcher, Aalto University
Tuuli Toivanen, Professor in Geoinformatics, University of Helsinki

3.6.2. Introduction

Open science commonly refers to efforts to make the methods, processes and output of publicly funded research more widely accessible in digital format to the scientific community, the business sector, or society more generally. Openness has for a long time been at the heart of scientific process. In 1942, Robert King Merton, an American sociologist of science, described a set of ideals that characterised modern science and to which scientists are bound. First and foremost is the notion of common ownership of scientific discoveries, according to which the substantive findings of science are seen as a product of social collaboration and are assigned to the community.

Strength of science is in the capacity for self-correction. Publication of scientific theories, and the data and methods on which they are based, permits others to validate the results, to identify errors and to reject or refine theories. Reuse of data strengthens further understanding and knowledge. Generally speaking, when a work has been published, it is accepted that anyone is free to use the information thereby made available as a basis for further work, with, of course, proper acknowledgement.[1]

The kind of open science enabled by the digitalisation of the research process has become a globally significant way to promote both science itself and its societal impact. Although openness has always been and will be a fundamental principle of science and research, these new open operating models will make science more democratic than ever before. Open science[2] creates opportunities, and its benefits extend to whole society. For researchers and research groups, openness conserves resources, improves the quality of research, and potentially offers increased credits and opportunities for cooperation. [3]Greater access to scientific inputs and outputs can improve the effectiveness and productivity of the scientific and research system, by: reducing duplication costs in collecting, creating, transferring and reusing data and scientific material; allowing more research from the same data; and multiplying opportunities for domestic and global participation in the research process.

In the preliminary brainstorming work by the working group (Forsström, Honkela, Ollila, Toivonen) it was decided to address the foreseen change in the perspective of 10 years (2025 as the far horizon). The decision was based on the increasing development towards openness, as there is growing evidence that open science has an impact on the research process. Scientists and academics are not the only groups that can benefit from greater open science efforts. For example, usage data from PubMedCentral (the online repository of the US National Institutes of Health) show that 25% of the daily unique users are from universities, 17% from companies, 40% are individual citizens and the rest are from government or in other categories (UNESCO, 2012) [4].

In the brainstorming work, we identified that that the transition to Open Science needs to be approached holistically; involving incentives, policies, legislation, infrastructures and support services. Based on this, we created a map for Open Science development. This was further enhanced in the Academy Insight seminar 23.11.2015.

[1] E.B.Wilson Jr: An Introduction to Scientific research. 1951. Dover Publications, New York. 375 pp.

[2] The Royal Society, (2012), Science as an open enterprise, The Royal Society Science Policy Centre report 02/12

[3] http://www.oecd-ilibrary.org/docserver/download/5jrs2f963zs1.pdf?expires=1450168844&id=id&accname=guest&checksum=F87860E6E67CF1851786DC97370EC029

[4] https://www.innovationpolicyplatform.org/content/open-science

3.6.3. Open science promotes change

Open science promotes all fields of science. The societal impact of science is boosted not only by better dissemination of scientific results but also by participation and transparency through trust. The following changes can be associated with open science: x **Creating opportunities** for everyone to participate in scientific advancement and enabling more effective utilization of research results, and promoting awareness and trust in science.

- **Facilitating a move from one field of science to another,** as it promotes mutual understanding (for example on semantic layer: unified vocabularies) and combined analytics (combined text- and data mining, fusion of heterogeneous data.

- **Strengthening collaboration and co-creation** to tackle multi-science problems by o Shared platforms for distribution and utilization of scientific results: diffusion of knowledge, easy uptake of ideas
    - o Sharing research infrastructures. Big investments that are knowledge hubs have a big potential in knowledge dissemination and uptake
    - o Exchange of expertise: easier career moves between scientific fields and to and from science o Multiplication of opportunities for domestic and global participation in the research process o Bringing together top-down and bottom-up approaches
- **Strengthening of good scientific practices** (data management, version management) and makes validation of results easier for example with permanent identifiers. By defining a unique permanent identifier to each article, dataset, infrastructure, code etc. it is easy to connect and refer to the, and link them accordingly. For example linked open data cloud provides a powerful tool to study interlinked relationships. Increasing reliability, transparency and quality in the research validation process, by allowing a greater extent of replication and validation of scientific results.

Anticipated changes in some indicators of science practice are shown in Figure 1. The more open the scientific process, the more collaboration is possible. The generation difference diminishes as the skills grow. The top-down approach transforms to bottom-up process as the skills grow.

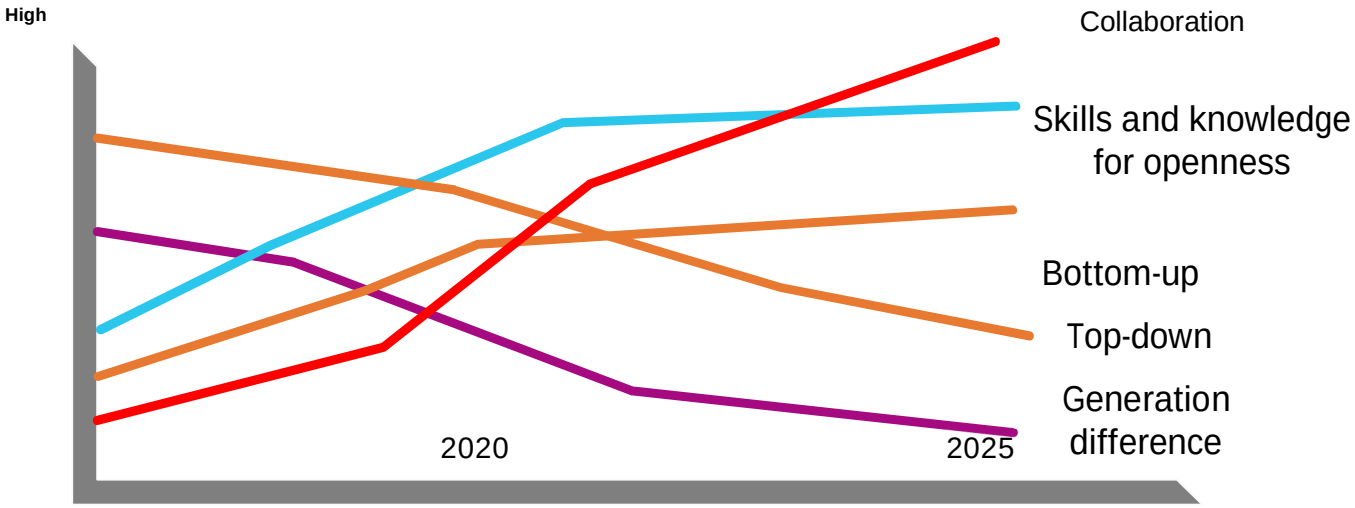


Figure 1: Foreseen changes in scientific practice by 2025.

3.6.4. Identified new fields of science

Several research problems can be addressed more effectively by open science. Some important emerging possibilities are listed below:

- **Data analytics and machine learning:** Data analytics tools (such as machine learning, pattern recognition techniques and statistical analysis) are increasingly used by scientists to generate new hypothesis, develop new models to gain knowledge of phenomena and to test or validate models. Accumulating large-scale data sets allow computer-based experiments and simulations, even in those fields where traditional lab experiments were impossible or too difficult to organize. Openness boosts the availability of large-scale data sets for computer-based experiments and simulations. With sufficiently large and open data sets, advanced computational methods can detect complex patterns and relationships that difficult to find in smaller data sets. Large data sets enable large scale computational validation or predictions on large samples, which can help mitigate the current reproducibility crisis in science. Openness helps in exploiting big data as an opportunity to develop new and more efficient algorithms for data analytics, to collect, manage and perform computations using large amounts of varied data. These are used for example to serve large-scale web applications and vast sensor networks.

- **Text and data mining (TDM):** Text and data mining is used in science and other disciplines to analyse and extract new insights and knowledge from the exponentially increasing store of digital data using linguistic, statistical and machine learning techiques. TDM needs open interfaces, open data, open publication. TDM is likely to become more important as researchers acquire the skills and the technology to address and investigate data sets of increasing size, complexity and diversity in all media: text, numbers, images, audio files and all other forms.

- **Semantic computing:** development of semantics (for example unified and shared vocabularies), as also standardization of data, metadata and interfaces facilitate content and topic analysis as never before. The semantic layer could evolve to be more intelligent with the growing common semantics. A true information layer on Internet is an addition to big data possibilities.

- **Computational science** will benefit the most from availability of open methods and code because computational methods can be exactly reproduced essentially for free. Open methods will greatly accelerate the progress of computational science as other researchers can directly build upon previously published work without the need to reimplement it. Progress in this area will critically depend on availability of funding also for maintaining existing code and other resources instead of just developing new ones as is currently mostly the case.

- **Gamification** is the use of game mechanics and game design techniques in non-game contexts. Open science can strengthen gamification works by making infrastructures and technologies available via open interfaces. The use of infrastructures and APIs can be made more engaging, by encouraging users to engage in desired behaviors, by showing a path to mastery and autonomy, by helping to solve problems and not being a distraction, and by taking advantage of humans' psychological predisposition to engage in gaming.

- **Big Data Science:** needs skills to collect, manage and perform computations using large amounts of varied data. These are used for example to serve large-scale web applications and vast sensor networks.

- **Meta-analysis and holistic science**: The aim in meta-analysis is to increase the power of statistical inferences by pooling results from multiple related studies. In addition to providing more reliable estimates of the quantities of interest, meta-analysis has the capacity to contrast results from different studies and identify patterns among study results, sources of disagreement among those results, or other interesting

relationships that may come to light in the context of multiple studies. Meta-analysis can be thought of as "conducting research about previous research."

3.6.5. Projection of open science future

The development towards open science process is depicted in Figure 2. The purple arrow is the growing openness in science, and the purple ovals represent stages and phases of openness. The orange ovals represent elements and structures needed in the science process. The deep blue ovals represent the changes in the science culture. The marine green ovals represent the fields affected and/or supporting open science. For example, legislation and research ethics affect strongly good scientific conduct, but are in turn touched by the open science movement.

Figure 2: Projection map of open science future.

By 2020, we identify that that Open Data, Open methods and Open Interfaces are common in scientific infrastructures and practice. By 2025, the scientific process has change towards open collaboration and co-learning, involving society at large.

3.6.6. Glossary

Open science **–** There is no formal definition of open science. In this text, the term refers to efforts by governments, research funding agencies or the scientific community itself to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction as a means for accelerating research; these efforts are in the interest of enhancing transparency and collaboration, and fostering innovation.

Open access **–** Unrestricted online access to scientific articles. Access can occur via a number of channels, such as institutional repositories, journal publishers' websites, researchers' webpages, etc.

Open data – Open data are data that can be used by anyone without technical or legal restrictions. The use encompasses both access and reuse. Whether such openness exists from the legal perspective depends on the applicability of possible legal restrictions (or otherwise, whether the restrictions are removed by a free licence). *Metadata* – are detailed descriptions of the data sets and documentation of the workflow needed to access these resources; they are often necessary for the usage of the data itself.