STRATEGIC RESEARCH AGENDA

for

Data to Intelligence (D2I)

Authors:
Petri Myllymäki, Jukka Ahtikari, Kai Puolamäki, Christer Carlsson,
Sami Sahala, Rauno Saarnio and Pentti Kurki

Version 1.0

6th June 2011

Contents

Figures

1 EXECUTIVE SUMMARY

The number of devices capable of automatically gathering and storing digital data is increasing fast: our mobile phones, home appliances, digital televisions, cars, industrial process monitoring systems, email clients, web browsers, social media applications, traffic and security cameras, and numerous other sources of digital information produce vast masses of data all the time. It is estimated that the amount of data produced globally in year 2010 was 1.2 Zettabytes, which equals the amount of digital information that would be created if every man, woman and child on Earth would tweet continuously, every hour of every day, for the next 100 years. What is more, in 10 years the annual amount of information produced is estimated to be over 40 times higher.

Global trend setters like Google, Yahoo, Netflix, Amazon and Autonomy have already shown that it is possible to transform data to economic value by producing novel, immensely popular and profitable services based on intelligent analysis of massive data sets. Nevertheless, new user-centric role based approaches and cooperative organization networks require ever more intelligent ways to utilize the available data. The content should be available automatically and be based, e.g., on user role, context requirements and process perspectives. This means that data sources are often crossing traditional organization borders and may be utilizing also open data reserves.

An additional important issue here is that the data is not only big, but it is also unstructured. As the data often emerges as a “side product” of our digital society, and not as a result of carefully designed and implemented data gathering activities, 95% of the emerging data is unstructured, consisting of not clean numerical data, but text, images, videos, audio and other forms of data that humans can process effortlessly, but that are most difficult to process automatically by computers. The magnitude of this data easily prevents straightforward human-assisted manual solutions where the data is enriched with “computer-friendly” tags or other forms of supplementary information.

The mission of the D2I SRA is to support the global trend, contribute to emerging ecosystems and boost Finnish international competitiveness through intelligent (context-sensitive, personalized, proactive) data processing technologies linked to new data-driven services that add measurable value, leading to increased knowledge, comfort, productivity or effectiveness. The target is reached by developing intelligent methods and tools for managing, refining and utilizing diverse data pools, and by creating new, innovative data-intensive business models and services based on these methods.

2 INTRODUCTION

The Sloan Digital Sky Survey that started its operations in 2000 has now – a decade later – stored 140 terabytes of information, while its successor, the Large Synoptic Survey Telescope that will come on stream in Chile in 2016, will collect that amount of information in only five days. The Large Hadron Collider at CERN generates 40 terabytes of data every second. The genome sequencing machines of hundreds of biomedical laboratories around the world create several terabytes of data per day, each, and with constant improvement of genome sequencing machines, the sequence output of the world is doubling every 9 months, outstripping Moore's law. These examples illustrate how quickly the amount of digital data in the world is increasing, and the emerging massive data sets require new, very effective analytics methods and an abundance of computing resources to make sense of the data and to possibly build new understanding of the complexities of our universe.

As impressive as these volumes are, it is perhaps even more important that we can easily recognize equally impressive examples in the industrial and business sectors as well: monitoring systems in industry, actors in the retail markets or in the public sector are now amassing information on a similar scale, and these massive amounts of information can form a basis for intelligent planning, effective problem solving and productive decision making. To support this more efficiently there is also increasing need for cross-industrial approaches. In addition, intelligent and automated data analytics enable new operational business models and innovative services.

But the trend does not end here, as the number of devices capable of automatically gathering and storing digital data is increasing fast: our mobile phones, home appliances, digital televisions, cars, industrial process monitoring systems, email clients, web browsers, social media applications, traffic and security cameras, and numerous other sources of digital information are estimated to produce globally over one zettabyte (one million petabytes) of new digital data in year 2010 alone (IDC 2010 Digital Universe Study), and by the year 2020, the size of our "digital universe" is expected to be 44 times as big as it was in 2009.

The key issue here is that unlike in science, where the data is typically generated using specific instruments designed for the purpose and the resulting data is well structured, 95% of the data generated by our digital environment is unstructured, consisting of not clean numerical data, but text, images, videos, audio and other forms of data that humans can process effortlessly, but that are most difficult to process automatically by computers. The magnitude of this data easily prevents straightforward human-assisted manual solutions where the data is enriched with "computer-friendly" tags or other forms of supplementary information. Automated computer-based solutions are difficult, but

companies like Google, Yahoo, Netflix, Amazon and Autonomy have already demonstrated that it is possible to produce novel, immensely popular and useful applications based on intelligent analysis of massive data sets.

The D2I SRA will aim at data-intensive, intelligent support and services utilizing the rapidly increasing masses of heterogeneous, unstructured digital data. Intelligence is the result of bringing computational methods and technologies to bear on data, information and knowledge. The target is to seize maximal benefit and added value of these services, leading to increased knowledge, comfort, productivity and effectiveness. This will be facilitated through a data processing pipeline where raw data is enriched through information and knowledge to be used as intelligence. The main concepts of this framework are

- **Data:** Elementary description of things, events, activities and transactions that are recorded, classified and stored but are not organized to convey any specific meaning; large sets of data is referred to as “big data”; algorithms can be designed and used to work on data to produce knowledge and intelligence.
- **Information:** Data organized so that they have meaning and value to the recipient; information is a collection of data; analytics models, algorithms and systems can be designed and used to work on information to produce knowledge and intelligence.
- **Knowledge:** Data and/or information organized and processed to convey understanding, experience, accumulated learning and expertise as they apply to a current problem or activity; knowledge is made up of different strands of information.
- **Intelligence:** Subsets of data, information and knowledge worked on with computational methods, algorithms and systems to produce focused insight in special topics and areas; intelligence built on data offers less insight than intelligence built on information [which is supported with analytics]; intelligence built on knowledge can be adaptive to the context and to the cognitive profile of the user; knowledge-based intelligence offers more effective user support, and guiding of the user.

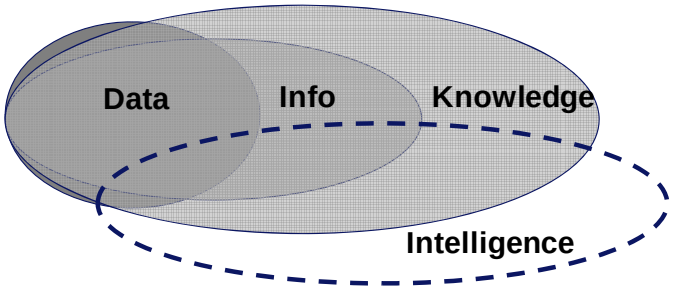

Figure 1. Data, information, knowledge and intelligence

Knowledge builds on data and/or information that is organized and processed to convey understanding, experience, accumulated learning and expertise; then with quickly growing volumes of both data and information we also have quickly growing volumes of knowledge in industry, in the retail markets, in the public sector and in the society. Knowledge offers more of a challenge than data and information, as it is the foundation for sustainable competitive advantage in industry and business, for fair and effective governance in the public sector and for intellectual growth in the society. Thus we need to develop theory, methods and tools for developing a better understanding of the results we get through processing and analyzing massive sets of data and information, for presenting the results in a form that is understandable for the users and relevant for the context in which the results are going to be used. The development of this type of methods has been fast over the last decade as the quickly growing computing power has offered the basis for developing more advanced and innovative algorithms. As an example of this trend, companies like SAS, SAP, IBM and several multinational consulting firms have successfully built a BUSD business around the business intelligence concept.

In summary, the Data to Intelligence SRA is about

- Identification of organization and user centric needs and (measurable) targets both on strategic and operative daily life perspectives.
- Making sense of large/huge amounts of data and information online and in real time on many somehow connected platforms.
- Creating, building and forming knowledge from multiple sources.
- Building new and utilizing existing frameworks (e.g. European Interoperability Framework, EIF) to support open architecture development.
- Searching for, finding, systematizing and activating latent knowledge, formalizing it and using it for planning, problem solving and decision making with sophisticated data analytics methods.

- Making knowledge operational for real-time use so that it is adapted to and made relevant for its context and for the cognitive profiles of the users.

- Utilizing the results in the development of new data-intensive services, processes and business models in cooperative ecosystems.

The Data to Intelligence SRA will develop a joint generic base in theory, methods and technology that will be worked out as an underlying core development in cooperation with other Tivit SRAs, other SHOKs, and the Academy of Finland research program. There are different contextual perspectives related to for example the public sector, industry, well-being, traffic and sales/retail that will be used for developing specific contributions to D2I in an iterative fashion over the duration of the D2I program.

In addition there are certain D2I level cross-themes that cover certain issues e.g. in the area of data security/privacy and interoperable platforms with cross-operational re-usable technology-free modules and ICT based components. For example cooperation with other Tivit SRA development areas is linked to these D2I cross-themes.

3 BACKGROUND

The astronomers may be producing huge amounts of data and information about the universe with their advanced telescopes, but equal amounts are produced in much more mundane circumstances: Wal-Mart is registering more than 1M customer transactions every hour which is feeding a database of 2.5 petabytes; Facebook is home to more than 40 billion photos; data and information stored in O&M monitoring systems in the 20 largest industrial corporations in Finland is estimated to be on par with the data collected by Wal-Mart.

The problem we face is that we now have the technology to produce much more data than we can ever hope to make sense of. Wal-Mart feels that point-of-sales data could reveal customer preferences and shopping habits that could be used for promotional campaigns and marketing; Google has some hope to be able to turn the large collection of photos stored on Facebook into some BUSD business [but it is not yet apparent how]; Finnish industry knows that they will have to start using the O&M data collected to improve on the quality and cost effectiveness of their operations.

Together with the new produced data it is important also to utilize already existing "old" content for example located in different data reserves. This means e.g. more efficient ways of identifying and integrating current data to support new process and service development. These activities utilize naturally also new content innovatively from new data sources. These activities require understanding of organization, process and user centric service development needs and targets both on strategic and operative level.

The D2I SRA will build on a carefully studied state-of-the-art of leading breakthrough theoretical frameworks, methodologies, models and algorithms and technologies for implementation in various contexts. This is going to be continuously updated and evaluated against breaking new developments coming both through the EU FP7 and FP8 projects as well as commercial products and services that are being launched by major vendors.

The D2I SRA will build on a core of research in the field of data analytics and other fields working on the data-information-knowledge-intelligence pipeline, and the work is going to be partly built on collaborative projects run by Academy of Finland. This core of research will develop analytics tools to build information from data, work out methodological solutions for building knowledge from data and information, develop effective algorithms and tools to find latent knowledge in "big data" to support intelligence-based action programs and services, and to build intelligence from data, information and knowledge in support of better planning, more effective problem solving and more produc-

tive decision making. Intelligent and automated data analytics enable new operational business models and innovative services.

3.1 D2I as a horizontal solution for industry verticals

The research will be organized in themes that are built around the three "corner stones" of the D2I SRA: OS (Organizations, Services), MA (Models, Algorithms) and DT (Data, Technology) but which together will cover the research core.

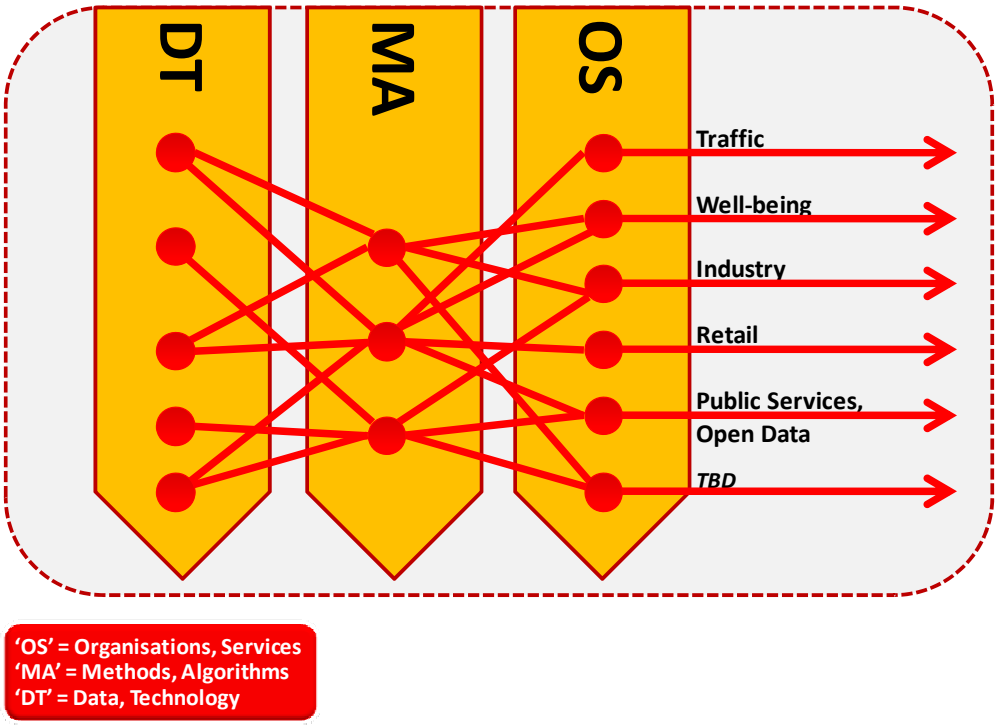

Figure 2. The overall structure of the D2I framework.

The research themes will be explored in cooperative work programs that address problem areas in one or several "vertical" areas where the research can find or develop (i) theory frameworks, (ii) analytics based methods, models and algorithms, (iii) methods and technologies for working on the data-information-knowledge-intelligence complex and (iv) theory frameworks and best practice cases for using D2I MA and DT solutions to build cost-effective and agile organizations and service systems.

The research process may (i) be initiated in one business area (cf. industry) starting with either OS, MA or DT; (ii) be jointly initiated by several application areas (cf. industry, retail/services and traffic) starting with either OS, MA or

DT; (iii) be initiated as core research on either OS, MA or DT [or a system of cooperative research processes on OS, MA and DT] and breakthrough results then tested, verified, validated and implemented for one or several of the business areas.

The D2I SRA activities will then focus on one or [jointly on] several areas:

- Identify from organization and user-centric point of view e.g. process and data/information needs [OS, DT]
- Define measurable development targets and organization/user based benefits [OS, MA]
- Make sense of large/huge amounts of data and information online and in real time on many somehow connected platforms [OS, MA, DT]
- Create, build and form new knowledge [MA, DT]
- Search for, find, systematize and activate latent knowledge, formalize it and use it for planning, problem solving and decision making with data analytics methods [MA]
- Make knowledge operational for real-time use that is adapted to and relevant for its context and for the cognitive profiles of the users – i.e. knowledge support that is relevant for the problems to be dealt with and given in such a form that the users will understand it and can carry out better action programs [OS, MA, DT]

The details will be worked out in the research program that is going to be developed with the partner organizations.

3.2 Finnish and global markets

D2I is open to activities in any business vertical where there is sufficient commercial interest, access to relevant data, a plan how to utilize the data, and the necessary expertise to make all this happen. Below we will shortly describe a few verticals that we feel offer a good starting point for this program.

Traffic

In the traffic sector a central driver is to make sure that the objectives defined for the traffic streams are obtained. In traffic planning the so-called Intelligent Traffic Strategy initiative is one of the most important. The planned service development (=business) is to a large extent based on the production and distribution of open data, which supports traffic management and driver decision making. The data collected about and around traffic streams is projected to grow exponentially over the next few years.

There are different sources of data and of different structures from which relevant information can be developed. The two important aspects are collection of traffic data from probe vehicles and ordinary road users. Moreover, the roadside monitoring units could be utilised in a more versatile manner than today. However, using the monitoring units calls for a market place type of a solution for sharing the captured data and providing business opportunities for companies dealing with transportation. The aim towards real-time data is offering a number of challenges for processing data and for building information based services. In addition these activities influence also to the development of "traditional" (service) business in various sectors.

In the public sector there are strong motives to promote stable and sustainable service development around intelligent traffic using business models that are built around multiple services. In the private sector similar movements are becoming visible among B2B customers where the driving forces are the consolidation of services, the development of new services that are introduced by regulations and the standardization (of also existing) services brought by EU regulations.

Well-being

In the well-being sector there is a growing need to produce data that can be used for a number of purposes despite the fact that data is stored in different data warehouses and according to different standards. The joint use of data warehouses is growing in importance, which requires better methods and technologies as the data sets grow quickly in volume and diversity.

There is a growing need to use data of different types and structures (numerical data, free text, video, audio, etc.) in business intelligence applications and the customers need to get access to better and better methods for extracting content from data and information. There is also a growing need to offer decision support systems that include more and more features of expert systems in order to produce enhanced analytical information in support of decision-making. In the future, emergence of personalised medicine and especially personal genome sequencing means that individuals will want to interpret their data in the context of data from other people. A huge market is likely to emerge in producing software solutions for interpreting personal health related data, aimed for common consumers instead of healthcare professionals.

In the well-being area increasingly more activity and cooperation is taking place between public and private sector organizations – e.g. in the area of home care etc. services provided people at home. In this environment intelligent and interoperable data can support flexible cross-organizational processes with customer caring services and simultaneously cost efficient operations.

Industry – O&M sector

The advanced automated systems that are standard in all large-scale industrial processes is one of the corner-stones of Finnish engineering technology and a major competitive advantage for Finnish industry. In the D2I the industry is going to focus on Operations and Maintenance [O&M] where the objectives to improve cost effectiveness and the productivity of working time has cut down on the number of operators of the O&M systems to a level which has produced a growing number of cases of overload as fewer and fewer operators have to deal with growing sets of data and information from automated systems that run increasing numbers of sensors. Until recent years the industry has been able to rely on experienced operators who have decades of experience and – what is called – tacit knowledge about what ails operations or what should be done as maintenance to ensure optimal levels of operation. This generation of experienced operators is now retiring and the tacit knowledge is getting lost which will start to appear as unscheduled interruptions of highly advanced, very complex and very costly production processes; needless to say, this will be eroding the competitive advantages of main parts of Finnish industry. The solution that the industry proposes to find and implement through the D2I program is to build data and information analysis capabilities into the O&M systems to process the take from the automated monitoring systems and to build this as knowledge support for inexperienced and semi-experienced operators, both in Finland and to an increasing degree for Finnish plant operations in other countries.

Industrial processes in all major Finnish industries are operated with highly automated systems and operators are supported with advanced and complex monitoring systems. Experienced operators normally react to process problems very effectively and with optimal action programs, which are both cost effective and avoid creating larger (often in some ways disastrous) problems. The operation and monitoring systems now work with huge sets of data that are collected routinely and only parts of which can be analyzed in any meaningful way but which should provide a basis for developing more efficient operations. The older generation of experienced operators is retiring and the younger generation typically works in the same jobs for only 3-5 years, which is not enough to build tacit knowledge; Finnish industry is increasingly operating internationally which makes it even harder to collect and communicate tacit knowledge on operations. Proactive and preventive maintenance build on advanced knowledge of the processes and effective diagnostics systems that can identify upcoming problems. Effective maintenance systems are now crucial for many of the advanced production systems and key Finnish companies have created BEuro businesses internationally around effective maintenance systems. The D2I will introduce data, information, knowledge and intelligence producing entities in the operations and maintenance systems which will in-

crease both their internal productivity and cost effectiveness and the possibility to build export products and services.

Retail

In retail a lot of interesting personal data can in principle be gathered e.g. through the numerous loyalty card systems. Although these systems face certain serious privacy issues, customers seem to be willing to allow usage of this data provided that they will be offered better, personalized and context-sensitive services. Early examples of this type of services include the already existing services for monitoring what type of food the customer has been buying (e.g. ravintokoodi.fi), or simple recommender systems (e.g. the Viiniopas applet for iPhone).

To make this type of services more useful, they must be made more intelligent and more personalized so that they can provide useful user-specific information. A particularly interesting direction is to combine the user profile data with some contextual data, e.g., location data: this would allow time- and location-specific personalized advertisements and notifications, route optimization, navigation support etc. For indoor retail outlets this of course would require an indoor positioning system, but Finland has been a trailblazer in this area too. For business owners analysis of this data allows better understanding of the customers, and means to optimize the offered services, e.g. by tracking the supermarket customers it may be possible to organize the shelf layout so that the shopping experience will be improved, or in more general a mall might use this type of analysis for better planning of its shop layout.

Public services / Open data

Public sector produces routinely a large amount of data that has the potential to be used for developing new forms of services. Public services can be developed by combining data from different sources, by making better use of the data with better analytics methods and by improving impact measures in order to find out how useful the data is for various purposes. New business can be built in both the public and the private sectors.

The data produced by organizations in the public sector will be opened up to all actors as required by legislation (e.g. EU Inspire Directive). This will bring data sources of various kinds and different quality to the market; the challenge will be to find relevant data and information through processing of these data sources. The growth in open data and it becoming available for real-time use offer new challenges to data processing and to the design of useful services around the data.

The public sector aims at opening up their data sources for service producers to exploit and commercialize which will promote new business and help to en-

hance existing service business. It will also help to give direction and focus to well-being services for user groups where this will have the best impact. The B2B customers of the private sector have the same objectives as (again) there is a need to consolidate services, to develop new services that are introduced by regulations and the standardization (of also existing) services brought by EU regulations.

The current market size cannot be estimated precisely but for example in the public sector the estimated European market for information (incl. open data) is about 27 BEuro (Mepsir study 2006). The overall market size in general could be estimated from Cloud markets and data growth rate predictions.

3.3 Global state-of-the-art

The D2I is not the first time that attempts have been made to span the data, information, and knowledge continuum for management and business purposes. *Knowledge management* (KM) as a keyword collects 76 million hits on Google in less than 1 second (or 0.03 femtogalactic years as they put it). KM is recognized as a discipline since 1991; more recently, other fields have started contributing to KM research; these include information and media, computer science, public health, and public policy. KM is now described as generating an annual revenue of about 10 B€ worldwide, which makes it an important consulting and software market.

In practice, however, KM programs, many of which continue today, have been only marginally successful (Davenport and Glaser, 2002, 2008). In the area of KM systems for the sales force, it has been documented that around 70% of KM projects have been unsuccessful (Braganza and Möllenkrame, 2002). Moreover some researchers found that there is a systematic lack of evidence for the claims put forth about the alleged knowledge management success stories (see Ekbia and Hara, 2008). In his attack on the “nonsense of knowledge management”, Wilson (2002) reported a 2001 survey carried out by Bain & Company showing that only 35 percent of a worldwide sample of 451 companies reported satisfaction rating about 3.5 on a five-point scale, when it comes to their KM initiatives.

What went wrong? The literature on KM disappointments and failure revealed that firms stumbled by adopting an IT-driven approach (cf. Storey and Barnett, 2000; Braganza and Möllenkrame, 2002; BenMoussa, 2009). KM technology has been designed without a deep understanding of knowledge worker’s work. For instance the gap between implemented KM technologies and organizational needs is illustrated in the following quote from one senior manager who parti-

cipated in their study: "I am not quite sure…where a knowledge management system ends and a business system starts (Nevo and Chan, 2007, p.592).

Knowledge mobilisation is an enhancement of the knowledge management methods and technology and represents the next step in implementing new forms of information and communication technologies (ICT) in management processes. The use of ICT is commonly believed and accepted to help improve the quality of planning, problem solving and decision making as these processes are supported with relevant and updated knowledge. The introduction of knowledge mobilisation will shift the focus from a supply driven to a demand driven approach, which will overcome some of the obstacles which have slowed down the implementation of knowledge management [KM] in many organisations (cf. Keen-Macintosh (2001)) The major obstacles for KM implementation in real organisations are: (i) knowledge workers who develop relevant, useful and advanced knowledge are unwilling to give it away to others who are less knowledgeable and/or unwilling to spend as much time to build a knowledge base unless there is a good and effective reward system in place; (ii) knowledge becomes obsolete and there should be incentives for updating, enhancing and improving core elements of the knowledge; (iii) knowledge is partly tacit and difficult to represent and share with other knowledge users, and (iv) knowledge is difficult to distribute and use independently of the knowledge producer. Nevertheless, the less than effective use of available knowledge is recognised as a major flaw in usage of corporate resources in most multinational corporations and it is a problem also for research and scientific communities as well as for public administrations (which is made visible by the eGovernment movement).

Recently, much hope has been invested in the work on the Semantic Web (Berners-Lee et al, 2001) as a way to build the required data-information-knowledge processing pipelines. The vision is that people will be able to build data stores on a network, build vocabularies and write rules for handling data; linked data is created through several standards (and technologies built on them) like RDF, SPARQL, OWL and SKOS; this platform is useful for the D2I but will be most useful for structured data (while a majority of the data sets we will have to deal with is unstructured data). One of the reasons for this is that Semantic Web builds on classical ontology, which for unstructured data grows very large very quickly which makes it hard to manage and (above all) to maintain. Semantic Web has been tested in the Finnish industry for O&M applications and the results have not been too encouraging.

In contrast to this, the current success stories on handling large data sets for commercial applications (e.g. Google, Netflix, Yahoo, Autonomy) are based on analytics-based computational methods that process raw data masses automatically, with no or very little human intervention. These algorithms can typically handle noise and uncertainties with probabilities or other *computational*

intelligence methods that are now replacing the AI research of the last century. The research sub-fields of computational intelligence that will be relevant for D2I include [but are not limited to] areas like probabilistic graphical models, evolutionary algorithms, artificial neural nets, decision trees, association rules, support vector machines, swarm intelligence and soft computing.

Besides handling large data sets we will also have to work on handling large sets of information and knowledge. Again, ontology-based methods are the most effective for handling these sets as long as they are well-structured and the entities are cleaned of noise, errors, imprecision and uncertainties. We know now with experience from industrial cases that we will have to deal with imperfect data and information using e.g. probabilistic methods (a relevant, although somewhat outdated report on this is given by Myllymäki and Tirri, 1998) or soft computing (Carlsson and Fullér, 2002).

For the last 10-20 years, the companies mentioned above, and other major actors in the field, have been aggressively recruiting the world's best experts in data analytics, machine learning, and related fields, obviously preparing themselves to be able to better utilize the new opportunities emerging with the new massive data reserves. In this respect, the following topics can be recognized as the most promising research directions, and will be addressed in D2I:

- *Collective intelligence*, where networked and distributed expert knowledge is fused for problem solving and end-user guidance.
- *Proactive intelligence* where monitoring data is analyzed for diagnosis, performance forecasts and predicted behavior.
- *Human-system joint intelligence* for which human performance is modeled and used for guidance to better performance.
- *Context-awareness* that is built around data on process states, skill levels of operators, data validity and uncertainties, etc. to decide what will be necessary and sufficient support.
- *Knowledge mobilization* that builds new knowledge from inconsistent data and information sources.

3.4 Finnish state-of-the-art

As argued above, utilization of massive, heterogeneous data sets requires sophisticated analytics methods; luckily in this respect Finland is in a very good position: in the Evaluation of Computer Science Research in Finland 2000-2006 (Academy of Finland, 8/07), the field of "machine learning and probabilistic methods" was recognized as "arguably the strongest single area of computer science in Finland", and many Finnish data analysis researchers are internationally recognized leaders of their field.

Finnish industry has co-operated with several research groups from Finnish universities and developed a number of breakthrough results utilizing intelligent data analysis methods, and this has led even to some commercial successes based on knowledge transfer from the academia to technology-driven companies like Xtract, M-Brain and Ekahau. However, in general the existing top-level expertise is too rarely successfully utilized in the Finnish industry. One reason for this is lack of knowledge: the relevant parties just cannot locate each other. Another problem is that the activities in Finland have been often focused very deeply in the technology development, with too little or even no activities concerning how to process the developed solutions into profitable products. It is evident that Finland still has to face significant challenges before it can create truly successful service-oriented data-driven businesses like for example Google and Facebook have done.

The objective of this SRA is to address these problems by recognizing the commercially most promising business opportunities for the available methodological solutions, and by bringing in together the best researchers in data analytics with the key companies having the best ideas for new data-driven services. There may be opportunities e.g. in the open data area for Finnish corporate and research partners to challenge leading international actors and to gain a leading position. In addition the size of Finland should support to build new innovative ecosystems capable to scale also international perspectives.

In D2I, one of the key aspects is transferability: we have to make sure that the insight and knowledge (including methods, models, algorithms and technologies) built within one business vertical such as process industry, can be successfully applied in other business verticals like traffic, well-being, retail and public services. This will of course require adaptation of the results from one application area to another – the contexts will be rather different in some of the cases – but this is where the D2I has a major role to play: it is envisioned as a horizontal program that cuts across a number of application areas, and it can be argued that Finland may create some sustainable competitive advantages by using this type of “cross-fertilization” of knowledge and expertise between key sectors of the economy; the competitive advantages will be useful for building better competitive and market positions for major actors in the Finnish economy.

A key part of the D2I will be to find international cooperative networks for quickly scaling up the results developed in Finland; these networks already exist in most of the application areas and part of the analytics methods, the models and the technologies are already being developed in international partnerships offering platforms for effective commercialization.

3.5 Global trends

In the data to intelligence area the key trendsetter is Google. At the moment it seems that mobile data will be collected with operating systems developed by Google, Apple and Microsoft. In addition there will be sector-specific actors developing independent solutions.

Open Data has a growing role in various public sector organizations globally. Its role covers both the internal and external development perspectives of the organizations. Open Data will influence the development of Web 2.0 based thinking and practical solutions. This area is also partly related to Open Source development.

After 2013, cloud computing will not bring any added value as a platform technology unless D2I produces solutions that will use cloud-based technology in significantly new ways and as a means for producing intelligent content.

Increasing cross-border service based solutions will increase the importance of identifying key data privacy legislation issues, e.g. on the national and EU level. Here for example the European Commission's interoperability activities with the European Interoperability Framework will obviously have an increasing role in the future activities.

The development cycle of D2I should be adapted to and follow relevant application areas in developing economies (e.g. China, India, South America, Africa); this could be one of the keys to a fast scale-up of business possibilities offered through the D2I results.

Gartner Symposium 2010, the world's most important gathering of CIOs and senior IT executives, listed top 10 technologies and trends that will be strategic for most organizations in 2011, "with the potential for significant impact on the enterprises in the next three years." Of these 10 technologies, we will address directly 7, which we have here combined to three key areas (and leaving out of our scope only the three more hardware-oriented technologies):

- *Next Generation Analytics:* We will identify how business can support operational decisions; we will develop and implement new analytics methods that combine analysis theory with computational intelligence and that will go beyond the state-of-the-art. This covers also business and service based development.

- *Context-Aware Computing; Ubiquitous Computing; Mobile Applications and Media Tablets:* We will define user and real-time intelligence needs; we will develop methods and technologies to make information and knowledge retrieval context-sensitive; we will work out how to make

these technologies ubiquitous and mobile and context adaptive to our application areas.

- *Social Communications and Collaboration; Social Analytics; Video:* We will identify people's daily life needs and routines and develop and implement methods and technologies to integrate heterogeneous data sources, including also audio, video and social media data.

Gartner predicts that "by 2013, more than half of Fortune 500 companies will have context-aware computing initiatives and by 2016, one-third of worldwide mobile consumer marketing will be context-awareness-based". Furthermore, David Cearley, vice president and distinguished analyst at Gartner, says: "Companies should factor these top 10 technologies in their strategic planning process by asking key questions and making deliberate decisions about them during the next two years." Consequently, we believe it to be evident that the timing for this SRA is exactly right and that the results of this SRA can have a substantial impact on different Finnish industries within the next few years.

3.6 Strategic challenges & opportunities

In (Economist, 2010), it is pointed out that "if you really want to transform health care, you basically build a sort of health care economy around the data that relate to people" [Mundie of Microsoft and Schmidt of Google on a presidential task force to reform American health care]. This is an example of how to transform data, information and knowledge into strategic assets by turning the cost- problems with handling large amounts of data into opportunities. "The data-centered economy" – continues Mundie of Microsoft – "is just nascent, you can see the outlines of it, but the technical, infrastructural and even business-model implications are not well understood right now".

Li & Fung Inc of Guangzhou in Southern China is one of the largest supply-chain operators in the world. It orchestrates a network of 12 000 suppliers in 40 countries with a turnover of 14 BUSD in 2008. Orders flow through a web portal and bids are solicited from pre-qualified suppliers; agents audit factories in real time with handheld computers; clients are able to monitor the details at every stage of an order from the start of the production run to the shipping; videoconferencing turned out to be an important break-through to allow buyers and manufacturers to examine material or the quality of a product in real time; data flowing through the Li & Fung network exceeds 1 terabyte per day.

Google handles around half of the world's internet searches, answering around 35 000 queries every second. Google is proving that metadata is potentially a lucrative business – metadata controls the pathways and means of finding in-

formation and knowledge and owners of metadata can (and will for sure) charge for their services; with more advanced tools for building and maintaining metadata that covers increasing large sets of data, information and knowledge the price and the revenue streams will grow, probably exponentially.

The D2I SRA builds on the strategy that we (i) identify sufficiently challenging D2I problems with national industrial and public sector partners, (ii) then develop analytics, data modeling and handling and knowledge building and using methods and technology for solving the problems, (iii) find ways to extract generic model and technology constructs from the results, and then (iv) find problem solving processes that can be applied in different contexts, also internationally. This requires flexible and vital ecosystems which are creative and also radical to build new business supportive service and technology solutions.

Advanced problem solving technologies developed for solving national problems can in this way be developed to export products and services, e.g. analytics expertise, horizontal sector crossing problem solving using knowledge from one sector in other sectors, consulting and international service production based on intelligence content developed with the models and the knowledge development technology.

This vision is more challenging to realize than what first appears to be the case. We should be aware that the numerous contexts we will have to deal with may be different even if they appear to be the same. Insights need to be verified through comparisons and benchmarking and the necessary data, information and knowledge cannot be assumed to be available in other contexts in Finland or in the same context in the other countries. Nevertheless, we belive that the D2I methods and technology will form good platforms for making use of advanced Finnish know how in many industrial and public sectors both nationally and internationally.

4 VISION 2016

We have developed in the D2I SRA the necessary intelligent methods and tools for managing, refining and utilizing diverse data pools. The results enable innovative data-intensive business models and services with supportive ecosystems.

This vision requires us to (i) identify organisation/user centric needs and make sense of large/huge amounts of data and information online and in real time on many somehow connected platforms; (ii) create, build and form knowledge from multiple sources; (iii) search for, find, systematize and activate latent knowledge, formalize it and use it for management with data analytics methods; (iv) make knowledge operational for real-time use and adapted to and made relevant for the context and the cognitive profiles of the users; (v) utilize the results in the development of new data-intensive services, processes and business models

The work will be organized through the following common D2I framework, or a solution space, covering three interoperable perspectives:

- OS: Organizations, users, processes and services
- MA: Methods, models and algorithms
- DT: Data, structures and technologies

This common framework has a key role in all D2I activities to support cooperation and roadmap based (gradual) development in all vertical and horizontal approached projects.

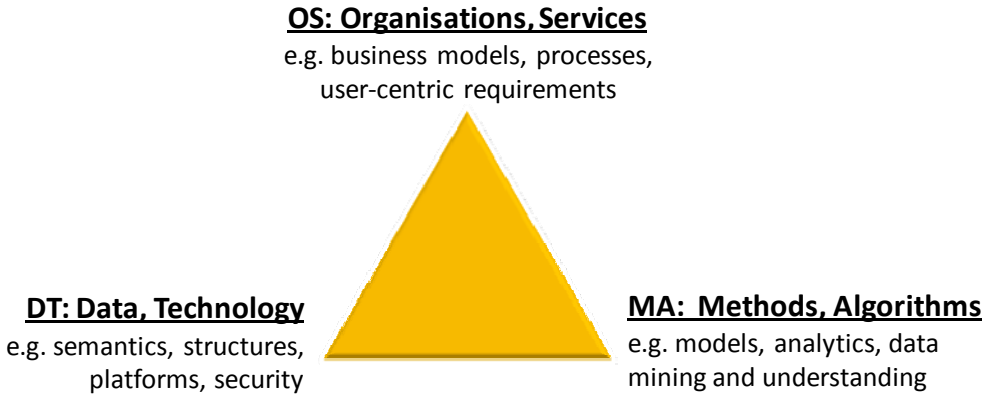

Figure 3. The D2I framework / solution space.

5 BREAKTHROUGH TARGETS

Our mission is to boost Finnish international competitiveness through intelligent (context-sensitive, personalized, proactive) data processing technologies and services that add measurable value.

The main breakthrough target is reached through the following sub-goals:

- Develop new data-intensive services improving productivity, cost effectiveness, comfort or knowledge.
- Create new data-driven business and earning models.
- Identify interoperability opportunities enabled by modern key technologies.
- Design, implement and test user-centric automated services utilizing big data analytics.
- Build re-usable technology-free service modules and technological components.
- Develop methods for big data analytics that
    - handle complexity through fusion of heterogeneous data sources, and
    - use adaptivity, context-sensitivity, scalability, and user relevance as the main methodological objectives.
- Create common (open) frameworks for intelligent data processing methods.

6 RESEARCH STRATEGY

6.1 Research methods

The goal of the D2I SRA is to link the world-class data analytics research done in Finland with innovative business opportunities so that the collaboration produces new intelligent, data-driven services, or improves the existing services or processes. Hence the work is driven by the industrial needs, which means that a substantial involvement by the Finnish industry is required, supported by the best methodological experts in the country. The collaboration between the academia and industry is based on gradual roadmap development approach which can be divided into three phases:

Phase one of D2I (years 1-2) has a general goal target to increase awareness of the possibilities in the most promising areas. For researchers, this means to increase understanding of the specific user requirements in various recognized problem settings, and what type of methodological solutions are needed and what type of modifications needs to be implemented on top of existing methods and algorithms. For the industrial participants, this means better understanding of the state-of-the-art: what types of methodological solutions already exist, what types of services they enable, and what is not currently possible. During this phase development projects are already taking place. They can be rapid pilots based e.g. on traditional production environments or flexible networked living labs. Because of increasing awareness there are already possibilities to utilize good practices in different themes (OS, MA, DT) and also combine them in new ways. For example, measurable targets and reuse of existing sector specific data/expertise/solutions in new verticals can be defined and a horizontal service oriented ecosystem can be built. For more advanced development targets and projects increasing activity is taking places in the phase two.

In the second phase (years 2-3) we will continue the work started in phase one, but at the same time try reach beyond state-of-the-art and identify the most promising business opportunities that cannot be reached with the current methodological solutions, or their minor modifications, but require new technological solutions. Already – in the phase one – active development projects can continue their work in this phase. For example continuing service oriented development with new earning models and identifying their possibilities to cooperate with other D2I projects e.g. starting in the phase two.

In phase three (years 3-4), the identified technological problems are solved and resulting methods implemented as parts of larger pilots that serve as test

beds for actual industrial applications and full-scale implementation projects for example supporting the roll-out of earlier executed rapid pilot results.

D2I is a horizontal, enabling an SRA aiming at generic, re-usable tools and methods that can be applied in any problem domain with a recognized need for intelligent, data-intensive services, and access to data. Although we have in Section 2 identified some promising business verticals as starting points for our work, we will not restrict ourselves to these areas alone, but look for new opportunities in other business sectors as well, especially those actively researched in other ICT SHOKs, or other SHOKs in general. To support these activities, we strongly encourage development of data markets, toolbox libraries and other platforms, standards and technology-free modules and technological components that support cross-sector transfer and re-use of results.

We strongly encourage open innovation and open standards, but at the same time acknowledge the privacy and security issues raised by working with large data pools collected by corporate and public sector actors. Although privacy and security issues may not be directly subjects of the research done in D2I, we will monitor the developments in these areas closely, and take into account in all stages of our operations the current and future trends in legislation, standards and policies regarding management of large data reserves.

6.2 International co-operation

The D2I SRA does not aim to establish a single, program-wide international initiative covering the whole D2I, but strongly encourages D2I partners to establish more focused international activities that still share the vision and goals of D2I. This type of activities include researcher mobility, research networks, standardization bodies and other forms of collaboration, including joint research projects funded by e.g. EU as part of their FP7 (FP8) or Artemis program. The focus here is global so that well-motivated activities with the best non-European partners are also highly welcome, but at the same time a special focus will be put on the recently established ICT Labs KIC of the European Institute of Innovation and Technology (EIT), where Finland is one of the key players.

Although utilization of large data reserves is not directly identified as one of the three research action lines of ICT KIC, D2I clearly supports two of them: Computing in the Cloud, which is an "emerging computing paradigm where applications, data and infrastructures are provided as a service that can be ubiquitously accessed from any connected devices over the Internet", and ICT-mediated Human Activity, which aims at "multimodal and embodied interaction, augmented and mixed reality, interaction with mirror worlds, and through intelligent information and media access". Also in this context D2I can be seen as a horizontal program supporting the five thematic action lines of the ICT Labs KIC.

7 RESEARCH THEMES

> *"Big data are datasets that grow so large that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analytics, and visualizing. This trend continues because of the benefits of working with larger and larger datasets allowing analysts to "spot business trends, prevent diseases, combat crime". […] Data sets also grow in size because they are increasingly being gathered by ubiquitous information-sensing mobile devices, "software logs, cameras, microphones, RFID readers, wireless sensor networks and so on."* (Wikipedia)

Automated collection of all this digital data that routinely takes place in increasingly many problem domains is producing vast collections of data that offer intelligent data analysis techniques numerous new potential application areas as this type of methods enable new intelligent data-intensive services. At the same time the nature of this type of data obviously presents new methodological challenges that will be addressed in the D2I program.

In addition it is essential to mention that Tivit has included data reserves as part of its strategic plan as it found them to offer great opportunity for Finland and its economical development. Data reserves also have a key role in the D2I development.

In Section 3.1 we divided the work in the D2I SRA under three *horizontal* thematic areas (see Figure 2) that form a pipeline processing first the raw data under the “DT” theme into enriched, integrated forms which are then analyzed and modeled under the “MA” theme so that the results enable fast, proactive, context-sensitive, personalized, and understandable technological solutions utilized in the intelligent services innovated, developed and studied in the “OS” theme.

Below we will address these *cross-sector* thematic areas in more detail. Although the resulting “data-information-knowledge-intelligence-intelligent services” process can be seen as a ("chronological") pipeline DT-MA-OS, the motivational driving forces behind the D2I SRA are supporting an idea of a cycle OS-MA-DT-MA-OS. Here the primary objectives are defined in the OS theme, which consequently determine the sub-goals of the theme MA, which in a similar manner motivates and guides the work done in the DT theme. This cycle takes place partly in parallel.

7.1 Theme OS: Organisations, Services

7.1.1 Description

In modern organizations there are several reasons and targets to build modern intelligence based on new data and also utilize existing data. These reasons can be presented thorough following questions:

- How can data and information support business operations and processes from both the users' and the organizations' points of view?
- How can we set clearer and more measurable targets for organizations with the help of valuable data and intelligence?
- How should organizations take notice of the opportunities and challenges that new legislation creates in form of new directives such as Inspire, and of the tightening security demands?
- How is it possible to build cooperation in networks and ecosystems where cross-organizational data reserves have the key supporting role?

In addition to the above mentioned issues there is an increasing role for the raw source data. It is about to come more freely available and more cost efficient, that development hasn't yet had any impact on service development. The inevitable challenges of coping with exponential amounts of data shall be tackled, both with results from DT and MA themes but also by adjusting service development itself to a new world.

Development has taken place in the area of data and intelligence in different projects. Though one of the main challenges of organization and it's development has been the apparent disability to produce viable business out of the otherwise successful pilots. Developing especially new services has turned out to be fairly expensive or at least time taking. Sometimes also service oriented development has not been possible because of missing ecosystem. For example, utilizing data without proper algorithms and especially without proper understanding of the critical organization development targets and user (business) needs, the modern innovative service approach can not take place.

Service development should be easier, faster and need a lot less investment or initial 'homework'. Too often the major portion of a project is to redevelop something that has been solved many times already; these tools and aids need to be provided to developers as the starting point, therefore a refreshed focus on reusable technology-free modules (covering e.g. process and data architecture definitions). Developing and providing new data-intensive services raises questions on privacy and data security, which are to be considered cross-theme issues, but eventually will focus on the end-user services.

The above mentioned OS theme issues are part of the targets for the D2I results. They are linked to other D2I themes (MA, DT) as described earlier in the development cycle OS-MA-DT-OS.

7.1.2 Focus areas

The OS Theme will focus on four main areas:

Data management will result in new ways to identify the data-intensive service value chain; stakeholders, roles and their data requirements, and also how the data within the service can be transferred securely throughout the value chain. It will also define the data content and open interfaces of national data reserves and public data archives

Context-sensitivity and personalization focus area aims at creating services for defining the ecosystem, roles and data requirements, and that help define user-centric, role-based data integration to enable data at right time and right place and in right context. New models for recycling methods of data profiling are also needed.

Reusable service modules will create new tools for utilizers and end users of the data, whether they are organizations or individual people. This creates new business opportunities for ICT service providers and developers. Open data has an important role in this development too.

Pro-activeness enabling services consist of several areas:

- Data management throughout value chain, where also the end user / customer is a source of data.
- Proactive services that are context sensitive and are able to analyze total result per user's requirements
- Predictive actions that are to be measured constantly based on information
- Controlling actions based on the pro-active information
- Data users are motivated to voluntary pro-activeness by making tools and information available to support decision making
- Offering services promoting wellbeing
- User-centric automated services utilizing big data

7.1.3 Goals

Main target is to seize maximal benefit and added value of the to be developed services, and assess them from several viewpoints. This breaks down to several sub-targets.

Increasing knowledge

Refining the data to a more understandable form, finding the essential information from the data and refining the information to knowledge

Improving comfort

By easing routine tasks, bringing new information and services within reach and accepting ICT systems' ubiquitous presence in everyday life, end users' actions are made significantly easier thus creating more comfort in users' life.

Better productivity

The services to be developed will enhance productivity of both customer and the good or service produced

Impressiveness

The services to be developed help to enhance services quality, user experience and impact on both user/customer and society. With these services customer will act more as a decision maker and an active participant in the service. Impressiveness will be enhanced by enabling knowledge-based models for proactive operations.

Adding value

The services to be developed will produce significant, measured added value to all parties in the ecosystem. This development both creates and utilizes new business and earning models.

7.1.4 Results

D2I and it's projects will bring stronger data and intelligence focused support for organizations existing operations and generate new, innovative, groundbreaking services that are both businesswise viable and produce measurable added value to their users and customers. Here the creation of new ecosystems has a critical role to support the business and organization scalability and viability.

These services either create or promote new business and earning models, often relying in networked producers and complex value chains.

Results will be measured on the added value but also on their reusable service components and their ability to generate or support new business and it's underlying framework and architecture.

7.2 Theme MA: Methods, Models, Algorithms

7.2.1 Description

This thematic area addresses three crucial methodological problems. The first problem is caused by the sheer magnitude of the data: handling of this type of astronomical data sources calls for new, more efficient data analytics as currently available solutions become infeasible with terabyte or petabyte level data sets that require computationally extremely efficient algorithmic solutions, and in many cases completely new, on-line methods that can process and model the data sequentially at the same time as it is collected.

The second problem is that once the information in the data has been extracted and compiled into higher-level models, we need to be able to access quickly the relevant data or information that is most useful to the user in the current context. An additional difficulty is caused by the fact that the data is often not only big, but it is also parceled, consisting of potentially several data sources that may contain heterogeneous data types. The nature of this type of data makes it very difficult to retrieve relevant pieces of data or information in a given context, in particular when the links between the different data elements in different data sources are not explicit, as is the case in traditional multi-view learning, but implicit, and have to inferred with the help of the constructed models.

The third problem is that the data is not only big (and potentially parceled), but it is often also extremely high-dimensional, which makes it very difficult to understand the underlying phenomena. What is needed is a rich toolbox of methods for representing the information extracted from the raw data in such a manner that the results help the user to understand the domain better, and support decision-making processes by helping in drawing conclusions about future events and in estimating their probabilities.

7.2.2 Focus areas

Big data analytics

According to the Wikipedia, "Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set. [...] Big Data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times."

In this focus area we will address methodological data analysis and modeling challenges related to issues like scalability, new data structures and algorithms, architectures for massive data storage and efficient computing methods. We will work not only in the traditional static, off-line setting, but also in dynamic, on-line settings: in many cases there is a need for new, on-line methods and algorithms that can process the data sequentially at the same time as it is collected, either because the size of the data just does not allow traditional "batch" style iterative processing of the data, or because the inherently sequential nature of the modeling task calls for this type of methods. Here we are concerned with issues like real-time processing of streaming data and on-line estimation and prediction. Special emphasis will be put on predictive analysis, and the results provide a technological platform on which one can build methodological solutions needed in the two focus areas below.

Context-sensitive information retrieval

In contextual information retrieval the description of what is relevant to the user needs to be inferred from the context and the implicit and explicit feedback in users' behavior, and this requires (data-driven) user and context models. Contextual retrieval contains also personalized retrieval, where the user's identity can be considered stable context. An important concept here is relevance: accessing of massive multi-source parceled data sets needs to begin by inferring potentially relevant patterns from the data, and for this we need the big data models constructed in the focus area above. What is retrieved with the help of these models can be elements of (raw) data, or higher-level information or knowledge. As the context or the user needs are typically inferred and not explicitly defined by the user, the resulting methods can be used as enablers for proactive data-intensive services. This interactive nature of the process poses hard computational requirements for the methods developed in the two focus areas above: when a human is in the loop, in most cases the response times of the computer system cannot be very long.

Visual and interactive analytics

When relevant information has been extracted from raw data, it needs to be represented to the user in a manner that helps him/her in understanding the domain better, in drawing conclusions, estimating probabilities and making informed decisions. This is especially important in areas where the data is so high-dimensional that traditional analysis methods become useless. The overall goal here is to provide the final steps of the cross-thematic pipeline that processes raw data first to information (by discovering models, patterns, regularities), and then even to higher-level (tacit) knowledge providing support for intelligent decision-making. The whole process is often interactive so that the results are first represented to the user (typically in a visual form), who then

inspects and interprets the results, and then modifies the analysis process by re-formulating the problem, changing focus, removing anomalies, etc.

The relevant methodological problem areas include issues like representation and visualization of information, model-based, domain-specific visualizations, 3D-reconstructions, visual tools, data summarization and compression methods, and methods for detecting and highlighting anomalies and weak signals; for more on recent developments in this area, see e.g. (Keim at al., 2010).

7.2.3 Goals

The goal is to develop novel data, information and knowledge modeling techniques that work in multi-source scenarios with heterogeneous data sources even when the connections between the different data elements are implicit and also have to be inferred, and scale up to handle massively large data sets. The resulting methods offer fast retrieval of relevant (context-sensitive, user-dependent) information, predictions about future events, and provide tools for increasing our understanding of the complex phenomena underlying the data. The developed methods have to be versatile so that that they can be easily transferred from one problem domain to another.

7.2.4 Results

We will develop a library of multi-purpose data analysis tools that serve as technological enablers for the services developed in the OS theme. Each tool is integrated as part of business pilots, and empirically validated in realistic scenarios related to the pilots. The results will be used as starting points for development of new, data-intensive services, which create new business opportunities for companies capable of offering this type of intelligent software or data analytics as a service.

7.3 Theme DT: Data, Technology

7.3.1 Description

The data can come in many forms such as:

- customer data and other data related to a person's behavior,
- business data and business architecture,
- genetic data and other biodata,
- multimodal data (video, audio, text etc.),
- GIS data,
- internet data (packet, log files), and
- process and sensor measurement data.

While the list above is not exclusive it is however indicative of the potential data sets to be tackled in the SRA.

Data is often not well structured nor labeled; hence methods developed for structured data are not directly applicable. The data needs to be refined from raw (often unstructured) data into something useful. The purpose of this section is to describe the initial part of this “data processing pipe” and to prepare the data into such a form that it can be further processed using the methods described in the previous section.

Some data sets are relatively small and can even fit into the memory of the workstation, but the data sets can also be extremely large with sizes up to many petabytes. Storing and processing data requires distributed solutions and cloud computing.

Data may be continuously produced and it is often necessary to analyze it in real time. At least any obvious anomalies that may be results of, e.g., broken sensors should be detected in as early stage as possible. Also, the amount of data may be increasing more rapidly than the available storage space (Economist 2010, IDC Study 2010). This means that early processing and analysis may be necessary also to decide what to store and in which form, because some of the data will be necessarily lost. It is therefore important to understand how the data will be used in order to store it in appropriate format and not to lose information that could be useful at a later stage.

Data must be preprocessed and managed so that it is readily usable for whatever purpose necessary, keeping in mind that there will probably be more uses for the data than originally planned. The data should be stored in a standard format, cleaned, and the missing values should be treated consistently. It is also beneficial to attach any metadata to the data at this stage. All this sets requirements for the management and mining of these massive data sets.

The data must be stored somewhere and an access must be provided for it. For large data sets the computation is a challenge and closely coupled with data management. Approaches such as MapReduce can be used to implement the data analysis methods within the data storage.

As discussed above, the data is in many cases no longer in a clean, structured numerical format, but may consist of multiple sets of more complex types of data, like text, images, audio, video, location information, and sensor data. Although analyzing this type of data is very difficult, it produces interesting new business opportunities with great potential: if one manages to recognize non-trivial links between multiple data sources, this very often reveals interesting information that would have been very hard to extract by traditional, single-source data analysis methods. If the dependencies between the data

sets are modeled we can obtain insights that would not be extremely difficult to find by looking at each of the data sources alone. As a very simple example, consider combining spatio-temporal information (the user's location at a given time) with semantic description of an event: the relevance of the data is very different if the user is at the Hartwall Arena (at the time when there is an ice-hockey game, or perhaps figure skating practice session), or at the home address of a very good friend on his/her birthday. In this simple example the link between the data sources was easy to construct through commonly share time stamps and location information, but in more complex cases the links are indirect and implicit, and have to be inferred from the data masses. As a more complex example, consider the genes of mice and men (CSCnews 4/2005). When the dependencies between functioning of the same genes are studied together using probabilistic modeling (ignoring information that is specific to only mice or to only men) we can, e.g., find the heart-specific clusters that are similar for both organisms, while the areas that correspond to brain homology in the two organisms differ. The functional regions would have been very difficult to find if the data sets would have been studied independently of each other. The similarities would also have been difficult to find without a principled probabilistic approach that takes the uncertainties in the data properly into account.

A classic example of the systems demonstrating the above properties are the Internet search engines, which continuously crawl the web and update their database to provide timely responses to queries by users. It is necessary not only to store data, but to pre-process it and combine several data sources (such as textual content of web sites, images found in them, the search behavior of the users, the profiles of the users in different geographic regions etc.) for the search engine to be able to give the results that are customized for a specific user (in a given geographic region with a certain search and click history) in an instant. To do all of this advanced data management (due to large amounts of data) and scalable algorithmic and probabilistic techniques (due to unstructured and noisy data and many data sources) are needed. One of the objectives of the D2I is to implement and generalize this approach also to other organizations and services.

The relevant methodological problem areas include methods for analyzing this type of complex data sets, recognition of (implicit) links or common patterns between heterogeneous data sets, and this area involves issues like data fusion, analysis of multi-resolution data, multi-view learning, and information retrieval. The ultimate goal is to make this type of methods fully automated and self-sustained, but at the first stage even semi-automated solutions supported by some manual effort may prove successful. Similar methodologies can be extremely useful always when there are several data sets of the same phenomena. This obviously sets prerequisites also for the data storage and

management: the data management should support such integration of information and whatever methods are used to fuse the data sets.

7.3.2 Focus areas

Data cleaning and pruning

The preprocessing and cleaning of the data are always an important part of any process that utilizes data.

Often analytical methods can be used to simplify and enrich the data (to extract the essential information from the data). Consider for example traffic measurements over time (maybe over several years) at some location. The probabilistic modeling of the time series might for example reveal that we can estimate the traffic by simple model according to which at a given time by the average traffic profile during a weekday (Mon-Fri) and during weekend (Sat-Sun). The representation has several advantages: the data is summarized and compressed, which makes further processing computationally and conceptually less demanding. The estimation of missing data is easier (we just estimate the missing data to be the estimate obtained from the model) - in the extreme case the model may be so good it may even be unnecessary to measure anything (whether we can skip measurements depends of course of the task at hand). This approach also makes it straightforward to spot anomalies, i.e., if the measured traffic at a given time is different from the expected average.

Another example is summarization of multimodal data (video, speech etc.): it may be much simpler to use summaries of videos instead of original video stream in further analysis. For example, finding parts of videos with cars in them is easier if the videos have been automatically annotated with keywords. Notice that in the D2I framework probabilistic modeling and video annotation would belong to the methods and algorithms part, but as demonstrated in the examples, the methods and algorithms can also be an integral part of the platform used to manage the data (e.g., video data is stored in conjunction with the annotations found by automatic annotation system).

In some cases the amount of data produced exceeds our ability to store data. In such cases the preprocessing and pruning of the data is extremely important and keeping the intended use of the data in mind, because otherwise essential information may be lost forever. An extreme example of this phenomenon is the particle Large Hadron Collider (LHC) experiment at CERN which produces roughly 15 petabytes of data annually (CERN, 2008). The LHC does heavy data preprocessing and pruning: the input rate of 10^9 interactions every second must be reduced by a factor of 10^7 to about 100 Hz (Smith, 2002). It would not be possible to store all of the data that this particle physics experiment produces.

Improving data quality, anomaly detection

Data should be preprocessed and anomalies detected. Preprocessing should however not loose information that could be useful for later analysis.

It is important to detect anomalies at the early stage of data collection. This may be a sign of an error (e.g., a broken sensor) or of some unusual phenomena which may warrant further study.

Enriching data with metadata and ontologies

In addition to the classic "Web of documents", the W3C organization is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

Various technologies allow you to embed data in documents (RDFa, GRDDL) or expose what you have in SQL databases, or make it available as RDF files. At times it may be important or valuable to organize data. Using OWL (to build vocabularies, or "ontologies") and SKOS (for designing knowledge organization systems) it is possible to enrich data with additional meaning, which allows more people (and more machines) to do more with the data.

These approaches are traditionally based on manual enriching of data. An interesting new development is to combine the semantic web technologies with intelligent data analysis techniques that can provide automatic annotation of data.

Management of data, standards, and open data

The data management should be designed keeping in mind the methods that will be used to analyze the data. For example, large data sets can be analyzed using distributed computing in which case the storage and management of data is closely coupled to the analysis methods. The storage of data should allow for missing values and expression of uncertainties. The storage should not loose essential information. While it may be beneficial to simplify the data (e.g., by classification or discretization) the original data should be available (if needed and if feasible).

An important topic is how to store and manage data, e.g., in cloud, such that it is readily available and in optimal format for further processing using methods

and algorithms described in Section 7.2. Special focus will be put on interoperable platforms and cross-operational re-usable modules/components. D2I will work together with other Tivit SHOK programs to create an open framework that can be used to manage and store data and apply methods and algorithms. Lots of data will be freely available and could be used in D2I (open data) and this will in itself create many business opportunities. Another side of the coin is the information security: the bulk of the D2I data will be confidential and non-public, which means that we must take the information security into account. This focus area also addresses other important cross-thematic issues like data privacy and data anonymization.

7.3.3 Goals

The part of the objective of the D2I related to this section is to create enabling technologies that can be used to process unstructured data from various sources into format that is readily usable by the methods and algorithms described in the previous section. The methods used here should be standardized and freely available, and it should be possible to use the same methodology across application areas. The technology created should make integrating information from several data sources a straightforward task.

7.3.4 Results

The methods will help to manage and understand the big data of sizes up to several petabytes. The methods developed here will provide solutions for the problem of having more data than storage space: advanced preprocessing and online knowledge extraction methods will be created. As a result, we should have a framework and a pilot system in which the methods of the Section 7.2 are easy to implement. These developments are valuable in themselves and they may lead to unforeseen innovations also outside the D2I.

8 WHAT WILL BE CHANGED

8.1 Research ecosystem

Implementation of this SRA brings academia much closer to industry. Currently there has not been enough ways to disseminate academic results into Finnish industry to the extent that is needed.

Also collaboration between researchers will increase. Scattered research resources will be combined to tackle greater challenges. This creates critical mass and enables sharing of methods and data. Attacking the same grand challenges with different teams having different backgrounds will lead to new insights and even initiate new research topics or fields of research.

Increased level of collaboration in the SRA themes increases also international visibility and international research collaboration. This SRA can create a positive spiral bringing researchers to Finland.

Currently data analytics research has been typically concerned with a single data set at a time, focusing on structured data. After this SRA the focus will be shifted towards more complex (and realistic) situations with several large distributed unstructured data sets. D2I will create new knowledge on how to manage and analyze such data sets, and advances will be reached in, e.g., computational methods and algorithms (regarding both computational efficiency and practical speed constant complexity, relying on current state-of-the-art and developing new statistically principled approaches for this problem domain); software platforms; standards; methods to handle large data sets and unstructured heterogeneous data; analytics.

For researchers, new working opportunities will be created through spin-offs, entrepreneurship, employment opportunities in participating companies, and international employment.

8.2 Business ecosystem

In general the implementation of this SRA means adding monetary value to the data assets of the companies.

The competitiveness of the participating partner companies will increase when they are able to implement and use state-of-the-art research results. Increased awareness of research results is already helpful by itself, but especially useful is their utilization in the new business opportunities to be identified and developed in this SRA. The Finnish business ecosystem will also be strengthened by starting new companies, via networking, and through creation of

new business alliances. SME ecosystems can be strengthened and their role in supporting larger corporations can be clarified and deepened.

Working together with other industry verticals will create new, previously unrecognized opportunities leading to positive surprises, requiring e.g. out-of-the box thinking; restructuring of old business models and business process re-engineering; creating mew business models; creating new areas of business; creating new kinds of services. Currently, there is not sufficient amount of mechanisms for cross-pollination ideas between industry verticals, and this "horizontal" SRA will help in breaking down the silos. It is also beneficial for the companies to merge their data sets in order to create more complex value chains and add value to their businesses.

An example of a promising new business area comes from creating services based on opening public data and linking it with other data sources.

8.3 Increasing Tivit research program cooperation

The D2I program supports through both its research and business ecosystems increasing cooperation for example between Tivit's Cloud Software and Next Media programs. Cloud Software and Data to Intelligence programs will establish an integrated way of working. In the search for prosperous ecosystems that will be active in both programs, the cooperation will take place in two different forms.

Firstly, the Cloud Software program, being mainly targeted for software development organizations, will make its inventions and emerging technologies available for use in the Data to Intelligence program, provided that suitable business drivers arise. For Cloud Software, such experiments will be treated as business pilots associated with those work packages and organizations that are best suited for implementing the pilot, similarly to other business pilots that have so far been mostly emerging from within the project itself. In other words, Data to Intelligence can feed new business drivers to Cloud Software, which will then analyze the requirements and use the data in its activities.

Secondly, the Data to Intelligence program similarly actively seeks opportunities for using inventions of the Cloud Software in order to reuse ready-made solutions when such already exist. Consequently Cloud Software can act as a technology provider for Data to Intelligence when the goals of the programs are similar.

The collaboration will be actively monitored, and its success in terms of the creation of new business and research ecosystems will be one of the criteria that will be used to assess the success of the programs.

8.4 Societal impact

Implementation of this SRA will also create societal impact on many levels. Firstly it should be emphasized that as basically every decision in this field requires hands-on involvement with relevant data, this increases our understanding of the underlying phenomena, and helps us to better handle the constantly increasing complexity of our everyday lives. An important area where these tools can make a big difference is e-democracy, as the popularity of the election candidate selection machines ("vaalikone") demonstrates. Other examples of this type of activities include the GapMinder (www.gapminder.org) and the Helsinki Region Infoshare (http://www.hri.fi) run by Forum Virium. It should also be emphasized that the same tools can also have a great impact on the society by increasing the understanding of the public organizations and officials about what and how to regulate, and overall in supporting better informed decision-making.

8.5 Paradigm shift: From verticals towards re-usable horizontals

The D2I SRA creates disruptive technology – while a lot of data already exists, and it obviously has a lot of value, this is currently not being properly recognized and utilized: those who have access to the data do not have the technology to utilize it, and those who may have the necessary know-how, do not have access to the data. We will build tools for extracting latent knowledge that is relevant for the context and user, thus identifying the value of the data. We believe that the most unpredictable (disruptive) aspect of this SRA will be the value created by combining different types of data, and innovative ways to utilize open data reserves.

9 REFERENCES

BenMoussa, C (2009) "Moving beyond the traditional approach of knowledge management: A demand-based approach as a key success factor" *Journal of Knowledge Management Practice,*Vol. 10 No. 3.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), The Semantic Web, *Scientific American*, May 17, 2001

Braganza, A and Möllenkrame, G.J (2002). Anatomy of a Failed Knowledge Management Initiative: Lessons from PharmaCorp's Experiences", *Knowledge and Process Management*, Volume 9 Number 1 pp 23–33.

Carlsson, C. and R. Fullér, *Fuzzy Reasoning in Decision Making and Optimization*, Studies in Fuzziness and Soft Computing Series, Springer-Verlag, Berlin/Heidelberg, 2002, 340 p.

CERN (2008): Worldwide LHC Computing Grid.
http://public.web.cern.ch/public/en/LHC/Computing-en.html

CSCnews 4/2005, 24-26: Mice and men and yeast, and dependency exploration.
http://www.csc.fi/english/csc/publications/cscnews/back_issues/CSCnews4_2005

Computer Science Research in Finland 2000-2006. Academy of Finland, 8/07.
http://www.aka.fi/Tiedostot/Tiedostot/Julkaisut/8.07%20Computer%20Scienceverkko.pdf

Davenport, T.H and Glaser, H (2002). Just-in-time Delivery Comes to Knowledge Management, pp-107-111, *Harvard Business Review*, and July.

Davenport, T.H., Prusak, L., Strong, B (2008) Putting ideas to work, *The Wall Street Journal*, March 10.

Economist (2010), Special Report on "*Data, data everywhere*", February 27, 2010.

Ekbia, H and Hara, N (2008) The quality of evidence in knowledge management research: practitioner versus scholarly literature, *Journal of Information Science*, Vol. 34, No. 1, pp 110-126.

European Commission, Commission's Communication on Interoperability, December 16, 2010. *http://ec.europa.eu/isa/strategy/index_en.htm*

IDC 2010 Digital Universe Study: A Digital Universe Decade – Are You Ready?
http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

Keen, P.G.W. and Mackintosh, R. (2001) *The Freedom Economy: Gaining the mCommerce Edge in the Era of the Wireless Internet*, New York: Osborne/McGraw-Hill.

Keim, D., Kohlhammer, J., Ellis, G. and Mansmann, F., editors (2010), Mastering the Information Age: Solving Problems with Visual Analytics, Eurographics Association. Available at
http://www.vismaster.eu/book/

MEPSIR study, June 2006. http://ec.europa.eu/information_society/policy/psi/actions_eu/policy_actions/mepsir/index_en.htm

Myllymäki, P. and Tirri, H. (1998), Bayes.verkkojen mahdollisuudet. Tekesin teknologiakatsaus 58/98. http://www.cs.helsinki.fi/u/myllymak/bvmahd.pdf

Smith, Wesley H. Triggering at LHC experiments. Nuclear Instruments and Methods in Physics Research A 478 (2002) 62-67.

Storey, J. and Barnett, E (2000). Knowledge Management Initiatives: Learning From Failure". *Journal of Knowledge Management*, Vol.4, No.2, pp 145-156.

Wilson, T. (2002, October). The nonsense of knowledge management, Information Research, available at http://informationr.net/ir/8-1/paper144.html¸ Last accessed September 10, 2008.