# Tackling big data in the Life Sciences: introduction to the Special Theme

Merks, R.M.H.; Sagot, M.F.

Introduction to the Special Theme

# Tackling Big Data in the Life Sciences

by the guest editors Roeland Merks and Marie-France Sagot

The Life Sciences are traditionally a descriptive science, in which both data collection and data analysis both play a central role. The latest decennia have seen major technical advances, which have made it possible to collect biological data at an unprecedented scale. Even more than the speed at which new data are acquired, the very complexity of what they represent makes it particularly difficult to make sense of them. Ultimately, biological data science should further the understanding of biological mechanisms and yield useful predictions, to improve individual health care or public health or to predict useful environmental interferences.

Biologists routinely measure DNA sequences, gene expression, metabolic profiles, and protein levels, infer molecular structures, visualize the positions and shape of cells in developing organs and whole embryos over time, rapidly screen the phenotypic effects of gene mutations, or they monitor the positions of species and individual animals and plants in whole ecosystems or herds, to name just a few.

The resulting datasets yield new insight into biological problems, while at the same time they pose new challenges to mathematicians and informaticians. How do we store these large bodies of data in an efficient manner and make sure they remain accessible in the future, while at the same time preserving privacy? How do we search through the data, spot interesting patterns and extract the biologically relevant information? How do we compare data of different conditions or species? How do we integrate data from different sources and across biological levels of organization to make new predictions? Can we then use the resulting patterns to solve the inverse problem and derive meaningful dynamical mathematical models from kinetic datasets?  How do we model complex and/or large biological systems, e.g., whole cells, embryos, plants or ecosystems? What are the challenges to multiscale modeling?

Given the great variety of topics that such general questions cover, this Special Theme of ERCIM News could only highlight a few, in a selection of nineteen papers.  These provide an overview of the latest techniques of data acquisition at the molecular level, address problems of data standardisation and integration, and propose sophisticated algorithms for analysing complex genetic features. Such algorithms help to infer the rules for predictive, dynamical modelling, by characterising genetic interactions, as well as molecular and cellular structures from noisy and incomplete data. They also provide key data for modelling multiscale systems at the multicellular or ecosystem level. Alongside data mining approaches, such dynamical models are a useful aid for proposing new control strategies, e.g., in personalised medicine or to help improve medical diagnostics. A further selection of articles discusses visualisation methods for big and noisy data, or suggest to make use of new communication techniques, such as Twitter, to help in quick management of health. Overall these approaches illustrate the breadth and the beauty of the mathematical and computational approaches that have been developed to cope with the challenges that biology poses to data science. These challenges will only continue to grow in the foreseeable future, as the volume,   quality and types of biological data keep on expanding.

**Please contact:**
Roeland Merks, CWI, The Netherlands
Roeland.Merks@cwi.nl

Marie-France Sagot, Inria, France
Marie-France.Sagot@inria.fr

# Networks to the Rescue – From Big “Omics” Data to Targeted Hypotheses

by Gunnar Klau

***CWI researchers are developing the Heinz family of algorithms to explore big life sciences data in the context of biological networks. Their methods recently pointed to a novel hypothesis about how viruses hijack signalling pathways.***

High-throughput technologies are generating more and more measurements of cellular processes in health and disease, leading to ever-increasing amounts of “omics” data. The main challenge is now to interpret and understand these massive data sets in order to ultimately answer important biological and biomedical questions.  This is a complex challenge, however, because important signals are hidden by high noise levels and heterogeneity of samples and diseases. Inspecting big life sciences data in the context of biological networks helps to address this challenge. To this end, researchers at CWI in Amsterdam are developing the Heinz family of algorithms to analyse and explore genome-scale measurements within biological networks.

## The Heinz family of algorithms

The Heinz project started in 2007 as a collaboration with University of Würzburg’s Biocenter. The first algorithmic prototype, Heinz 1.0 [1], was presented in 2008 at ISMB, the premier conference on computational biology. There, the work received the outstanding paper award for achieving a breakthrough in computing optimal subnetwork modules by introducing and exploiting a relation to graph theory. Since then, the method has been improved and several variations have been presented:

- Heinz, the workhorse method, is currently at version 2.0 [L1]. It takes as input a set of gene scores and finds an optimal active subnetwork module with respect to these scores. This is a connected subnetwork where the sum of the gene scores is maximal.  Finding such a module is an NP-hard problem. Heinz computes provably optimal modules using advanced techniques from mathematical optimization. While Heinz 1.0 exploited the close relation of the underlying Maximum-Weight Connected Subgraph (MWCS) problem to the Prize-Collecting Steiner Tree (PCST) problem and relied on PCST codes, Heinz 2.0 directly solves MWCS using a recursive graph-decomposition scheme into bi- and tri-connected components and a dedicated branch-and-cut algorithm.
- BioNet [L2] is an R package that provides an easy-to-use interface to Heinz. Provided with raw data, e.g., from RNA-Seq measurements, it generates the input score files needed by Heinz. The scores are based on a statistically sound decomposition of p-values describing the measurements into signal and noise components.
- Heinz has been adapted to answer a variety of research questions involving the interpretation of differential gene expression, GWAS, metabolomics, proteomics and metagenomics data.
- xHeinz [L3] is a recent addition that computes conserved cross-species modules. In a cooperation with the Netherlands Cancer Institute (NKI), xHeinz was used to provide evidence that the differentiation process of the recently discovered Th17 cell type, which plays an important role for the immune system, is conserved between mouse and human [2].
- eXamine, a visual analytics app for exploring Heinz results in the popular Cytoscape visualization platform,

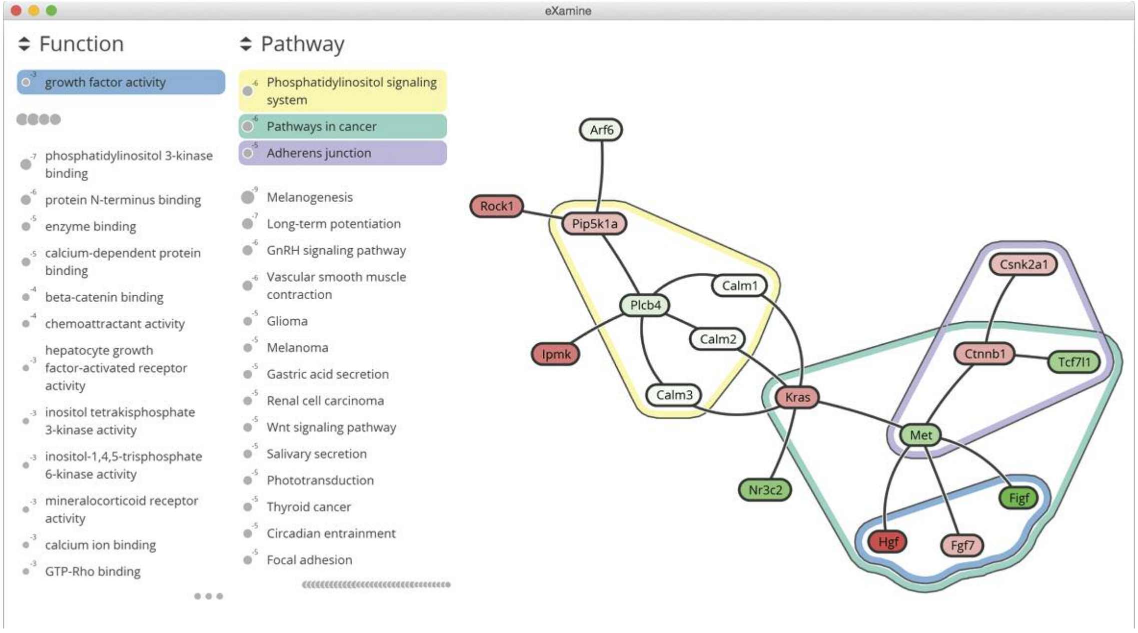

eXamine
Function
Pathway
-3 growth factor activity
-7 phosphatidylinositol 3-kinase binding
-6 protein N-terminus binding
-5 enzyme binding
-5 calcium-dependent protein binding
-4 beta-catenin binding
-4 chemoattractant activity
-3 hepatocyte growth factor-activated receptor activity
-3 inositol tetrakisphosphate 3-kinase activity
-3 inositol-1,4,5-trisphosphate 6-kinase activity
-3 mineralocorticoid receptor activity
-3 calcium ion binding
-3 GTP-Rho binding
-6 Phosphatidylinositol signaling system
-6 Pathways in cancer
-5 Adherens junction
-9 Melanogenesis
-7 Long-term potentiation
-6 GnRH signaling pathway
-6 Vascular smooth muscle contraction
-5 Glioma
-5 Melanoma
-5 Gastric acid secretion
-5 Renal cell carcinoma
-5 Wnt signaling pathway
-5 Salivary secretion
-5 Phototransduction
-5 Thyroid cancer
-5 Circadian entrainment
-5 Focal adhesion
Arf6
Rock1
Pip5k1a
Calm1
Plcb4
Ipmk
Calm2
Calm3
Kras
Nr3c2
Csnk2a1
Ctnnb1
Tcf7l1
Met
Figf
Hgf
Fgf7

*Figure 1: Optimal Heinz module along with enriched functional and pathway categories using eXamine.*

was developed in a cooperation with Eindhoven University of Technology. The tool makes it easy to explore annotated Heinz modules, for example, in the context of Gene Ontology or pathway enrichment.

All software in the Heinz family is open source.

### Case study on virally deregulated s ignalling

The Human Cytomegalovirus (HCMV) is a specific type of herpes virus with a high prevalence of 60% among humans. The interplay of HCMV infections with many diseases, including cancer, is an important topic of biomedical research. In a collaboration within the Amsterdam Institute for Molecules, Medicines and Systems, Heinz and eXamine were used to study a module that is activated by an HCMV-encoded G-protein coupled receptor. See Figure 1 for an illustration of the optimal Heinz module along with enriched functional and pathway categories using eXamine. Using the tools from the Heinz family, the researchers have been able to formulate a new hypothesis about deregulated signalling of β-catenin by viral receptor proteins. Parts of this new hypothesis have now been verified experimentally and have led to targeted follow-up studies, which are currently under way.

### Future

Current research includes the application to cancer genomics data. Here, the task is to extract subnetworks that show exclusive mutation patterns in the samples. A long term research goal is to move towards more dynamic descriptions of cellular mechanisms.

**Links:**

[L1] http://software.cwi.nl/heinz
[L2] http://bionet.bioapps.biozentrum.uni-wuerzburg.de
[L3] http://software.cwi.nl/xheinz

**References:**

[1] M. Dittrich et al.: "Identifying functional modules in protein-protein interaction networks: an integrated exact approach", Bioinformatics 24(13):i223-i231, 2008.

[2] M. El-Kebir et al.: "xHeinz: an algorithm for mining cross-species network modules under a flexible conservation model", Bioinformatics 31(19):3147-55, 2015.

[3] K. Dinkla et al.: "eXamine: Exploring annotated modules in networks", BMC Bioinformatics 15:201, 2014.

**Please contact:**
Gunnar W. Klau
CWI, currently Fulbright Visiting Professor at Brown University
Tel: +31 20 592 4012
E-mail: gunnar.klau@cwi.nl

# Interactive Pay-As-You-Go-Integration of Life Science Data: The HUMIT Approach

by Christoph Quix, Thomas Berlage and Matthias Jarke

***Biomedical research applies data-intensive methods for drug discovery, such as high-content analysis, in which a huge amount of substances are investigated in a completely automated way. The increasing amount of data generated by such methods poses a major challenge for the integration and detailed analysis of the data, since, in order to gain new insights, the data need to be linked to other datasets from previous studies, similar experiments, or external data sources. Owing to its heterogeneity and complexity, however, the integration of research data is a long and tedious task. The HUMIT project aims to develop an innovative methodology for the integration of life science data, which applies an interactive and incremental approach.***

Life science research institutes conduct high-content experiments investigating new active substances or with the aim of detecting the causes of diseases such as Alzheimer's or Parkinson's. Data from past experiments may contain valuable information that could also be helpful for current research questions. Furthermore, the internal data needs to be compared with other datasets provided by research institutes around the world in order to validate the results of the experiments. Ideally, all data would be integrated in a comprehensive database with a nicely integrated schema that covers all aspects of research data. However, it is impossible to construct such a solution since schemas and analysis requirements are frequently changing in a research environment. The a-priori construction of integrated schemas for scientific data is not possible, because the detailed structure of the data to be accumulated in future experiments cannot be known in advance. Thus, a very flexible and adaptable data management system is required in which the data can be stored irrespective of its structure.

The core idea of the HUMIT project is illustrated in Figure 1. Data management in life science is often file-based, as the devices (e.g., an automated microscope or a reader device) in a lab generate files as their main output. Further data processing is often done in scripting languages such as R or MATLAB by reading and writing the data from/to files. Therefore, the HUMIT system currently targets files as data sources, but database systems and other types of data sources can be integrated into the system later without changing the basic architecture. In a first step, a shallow extraction will be done for the input files in which the metadata of the source files is extracted and loaded into the data lake. "Data lake" is a new buzzword which refers to systems in which the data from the sources is copied, retaining its original structure, to a repository [3]. This idea is also applied in HUMIT:

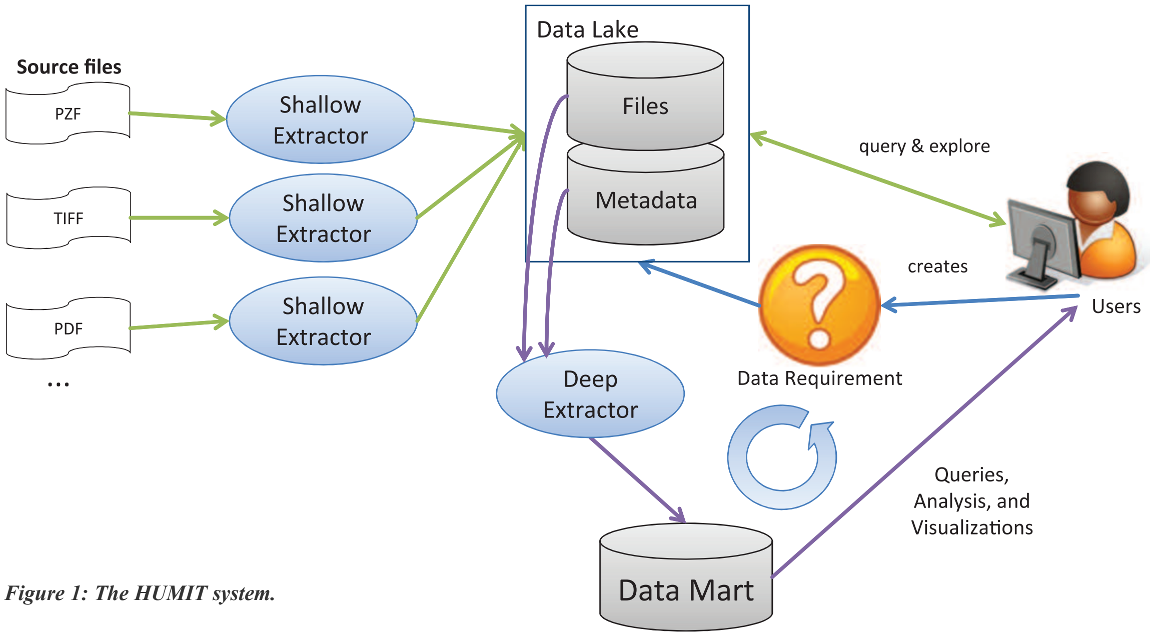


*Figure 1: The HUMIT system.*

the source files will be copied to the repository and stored with a bi-directional link to their metadata. The objects in the repository are immutable and only an append operation will be supported (no updates or deletes), as reproducibility and traceability is an important requirement for data management in life science.

The metadata will be provided to the user for exploration. By browsing through the metadata repository, the user can see what kind of data is available and incrementally construct a data mart in which data for her specific requirements are collected. Data integration takes place at this stage. This type of integration is referred to as "pay-as-you-go-integration" [2], since only the part of the data lake required for a particular application is integrated, and the integration is done while the user is exploring the data. The extraction of the detailed data from the source files (known as "Deep Extraction") will also be done at this stage if the data is required in the data mart. During the construction of the data mart, the user will be supported in the definition of data integration mappings by (semi-)automatic tools (e.g., for schema matching or entity resolution). Existing mappings can be also exploited for the definition of new mappings.

The requirements analysis for the project has been completed and a first version of the metadata management system is available. Metadata can be extracted from various file types and is stored in a generic metamodel, which is based on our experiences in developing the GeRoMe metamodel [1]. The framework for metadata extraction is extensible and can be easily adapted to new file types. The design of the interactive data exploration component is currently underway and will be one of the core components of the system.

The HUMIT project [L1] is coordinated by the Fraunhofer Institute for Applied Information Technology FIT and funded by the German Federal Ministry of Education and Research. Further participants are the Fraunhofer Institute for Molecular Biology and Applied Ecology IME and the German Center for Neurodegenerative Diseases (DZNE) as life science research institutes, and soventec GmbH as industry partner. The project started in March 2015 and will be funded for three years.

**Link:**
[L1] http://www.humit.de

**References:**

[1] D. Kensche, C. Quix, X. Li, Y. Li, M. Jarke: Generic schema mappings for composition and query answering. Data Knowl. Eng., Vol. 68, no. 7, pp. 599-621, 2009.
[2] A.D. Sarma, X. Dong, A.Y. Halevy: Bootstrapping pay-as-you-go data integration systems, in Proc. of SIGMOD, pp. 861-874, 2008.
[3] I. Terrizzano and Peter M. Schwarz and Mary Roth and John E. Colino: Data Wrangling: The Challenging Yourney from the Wild to the Lake, in Proc. of CIDR, 2015.

**Please contact:**
Christoph Quix, Thomas Berlage, Matthias Jarke
Fraunhofer Institute for Applied Information Technology FIT, Germany
E-mail:
christoph.quix@fit.fraunhofer.de,
thomas.berlage@fit.fraunhofer.de,
matthias.jarke@fit.fraunhofer.de

# Eliminating Blind Spots in Genetic Variant Discovery

by Alexander Schönhuth and Tobias Marschall

***Detecting genetic variants is like spotting tiny sequential differences among gigantic amounts of text fragment data. This explains why some variants are extremely hard to detect or have even formed blind spots of discovery. At CWI, we have worked on developing new tools to eliminate some of these blind spots. As a result, many previously undiscoverable genetic variants now form part of an exhaustive variant catalogue based on the Genome of the Netherlands project data.***

In 2007, the advent of "next-generation sequencing" technologies revolutionized the field of genomics. It finally became affordable to analyse large numbers of individual genomes, by breaking the corresponding DNA into fragments and sequencing those fragments, yielding “sequencing reads”. All of this is now happening at surprisingly – nearly outrageously – low cost and high speed. Advances in terms of cost and speed, paired with the relatively short length of the fragments (in comparison to “first-generation sequencing”) comes at a price, however. First, the rapid pile-up of sequencing reads makes for a genuine “big data” problem. Second, the reduced fragment length yields even more complex scientific riddles than in “first-generation sequencing” times. Overall, the resulting computational problems are now harder both from theoretical and practical points of view. Despite – or possibly owing to – the incredible mass of data, certain genetic variants stubbornly resist detection and form blind spots of genetic variant discovery due to experimental and statistical limitations.

Note that, in the absence of adequate methods to detect them, the first question to ask is: do these variants even exist in nature?

The presence of possible blind spots has not kept researchers from analysing these gigantic haystacks of sequence fragments. A prominent example of such an effort is the "Genome of the Netherlands" project [2], which has aimed at providing an exhaustive summary of genetic variation for a consistent population. Launched in 2010, it is both one of the earliest population-scale sequencing projects, and still one of the largest of its kind -- overall, the fragment data amounts to about 60 terabytes. The analysis of sequencing data is further enhanced by sequencing related individuals – either family trios or (twin) quartets – which allows the researchers to study transmission of variants and variant formation within one generation [3]. The resulting catalogue of variants establishes an invaluable resource, not only for the Dutch, but also for closely related European populations regarding association of disease risks with DNA sequence variation, and personalized medicine in general.

At CWI, as members of the Genome of the Netherlands project, we have succeeded in eliminating a prominent discovery blind spot, thereby contributing large numbers of previously undiscoverable genetic variants. We achieved this by reversing a common variant discovery workflow – usually, large amounts of seemingly ordinary looking sequence fragments are removed, which turns a big into a small data problem and renders fragment analysis a lot easier. In contrast, we process all data [1]: in other words, instead of removing large amounts of hay and, with it, considerable amounts of needles that are too tiny to be easily spotted, we rearrange the entire haystack such that even the tiny needles stick out. We have developed a "statistical magnet" that pulls the tiny needles to the surface.

The key to success has been the development of an ultra-fast algorithm that empowers the application of this

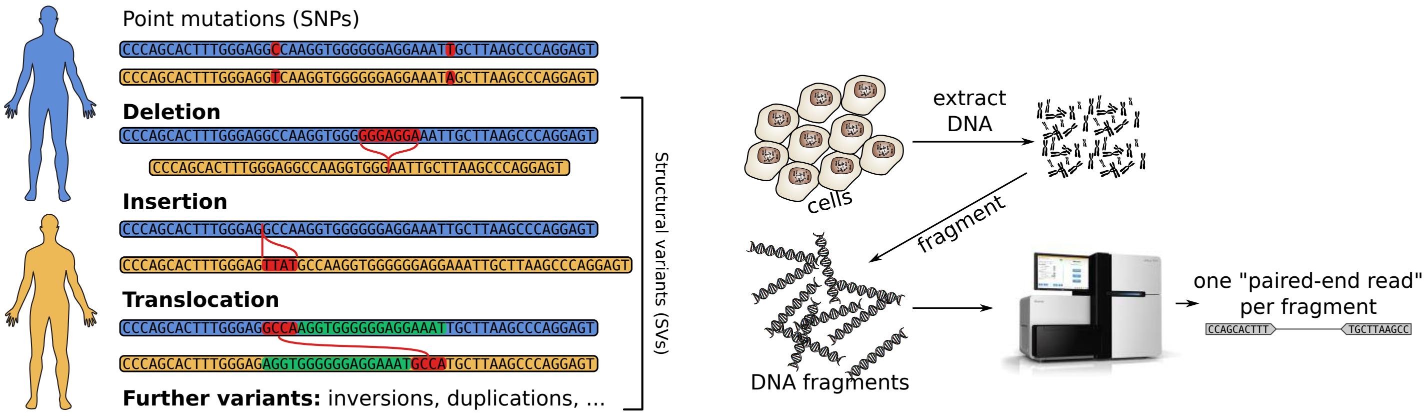

Point mutations (SNPs)
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT
CCCAGCACTTTGGGAGGTCAAGGTGGGGGGAGGAAATAGCTTAAGCCCAGGAGT
Deletion
CCCAGCACTTTGGGAGGCCAAGGTGGGGGAGGAAATTGCTTAAGCCCAGGAGT
CCCAGCACTTTGGGAGGCCAAGGTGGGAATTGCTTAAGCCCAGGAGT
Insertion
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT
CCCAGCACTTTGGGAGTTATGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT
Translocation
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT
CCCAGCACTTTGGGAGAGGTGGGGGGAGGAAATGCCATGCTTAAGCCCAGGAGT
Further variants: inversions, duplications, ...
Structural variants (SVs)
cells
extract DNA
fragment
DNA fragments
one "paired-end read" per fragment
CCAGCACTTT
TGCTTAAGCC

*Figure 1: Left: Different classes of genetic variants in human genomes. Right: Next-generation sequencing, only after breaking up DNA in small fragments, one can read the DNA – however, deletions and insertions of length 30-200 letters now are very difficult to spot. We have eliminated this blind spot in discovery by developing new algorithms.*

magnet even on such massive amounts of sequence fragments. In summary, the combination of a sound statistical machinery with a highly engineered algorithm allows for implementation of a reversed discovery workflow.

As a result, the Genome of the Netherlands project is the first of its kind to exhaustively report on the corresponding class of genetic variants, previously termed “twilight zone deletions and insertions”, but which now enjoy somewhat more daylight.

In future work, we are also planning to eliminate this blind spot in somatic variant discovery, which will likely reveal large amounts of so far undetected cancer-causing genetic variants, and will hopefully shed considerable light on cancer biology as well.

**Links:**
http://homepages.cwi.nl/~as
http://www.nlgenome.nl

**References:**

[1] T. Marschall, I. Hajirasouliha, A. Schönhuth: “MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels”, Bioinformatics 29(24):3143-3150, 2013.
[2 ]The Genome of the Netherlands Consortium: “Whole-genome sequence variation, population structure and demographic history of the Dutch population”, Nature Genetics 46(8):818-825, 2014.
[3] W. Kloosterman, et al.: “Characteristics of de novo structural changes in the human genome”, Genome Research 25:792-801, 2015.

**Please contact:**

Alexander Schönhuth
CWI, The Netherlands
E-mail: A.Schoenhuth@cwi.nl

*Tobias Marschall was a postdoc at CWI from 2011-2014. Since 2014, he holds an appointment as assistant professor at the Center for Bioinformatics at Saarland University and the Max Planck Institute for Informatics in Saarbrücken, Germany*

# Computational Estimation of Chromosome Structure

by Claudia Caudai and Emanuele Salerno

***Within the framework of the national Flagship Project InterOmics, researchers at ISTI-CNR are developing algorithms to reconstruct the chromosome structure from "chromosome conformation capture" data. One algorithm being tested has already produced interesting results. Unlike most popular techniques, it does not derive a classical distance-to-geometry problem from the original contact data, and applies an efficient multiresolution approach to the genome under study.***

High-throughput DNA sequencing has enabled a number of recent techniques (Chromosome Conformation Capture and similar) by which the entire genome of a homogeneous population of cells can be split into high-resolution fragments, and the number of times any fragment is found in contact with any other fragment can be counted. In human cells, the 46 chromosomes contain about three billion base pairs (3 Gbp), for a total length of about 2 m, fitting in a nucleus with a radius of 5 to 10 microns. As a typical size for the individual DNA fragments is 4 kbp, up to about 750,000 fragments can be produced from the entire human genome. This means that there are more than 280 billion possible fragment pairs. Even if the genomic resolution is substantially lowered, the resulting data records are always very large, and need to be treated by extremely efficient, accurate procedures. The computational effort needed is worthwhile, however, as the contact data carry crucial information about the 3D structure of the chromosomes: understanding how DNA is structured spatially is a step towards understanding how DNA works.

In recent years, a number of techniques for 3D reconstruction have been developed, and the results have been variously correlated with the available biological knowledge. A popular strategy to infer a structure from contact frequencies is to transform the number of times any fragment pair is found in contact into the distance between the components of that pair. This can be done using a number of deterministic or probabilistic laws, and is justified intuitively, since two fragments that are often found in contact are likely to be spatially close. Once the distances have been derived, structure estimation can be solved as a distance-to-geometry problem. However, translating contacts into distances does not seem appropriate to us, since a high contact frequency may well mean that the two fragments are close, but the converse is not necessarily true: two fragments that are seldom in contact are not necessarily physically far from each other. Furthermore, we checked the topological consistency of the distance systems obtained from real data, and found that these are often severely incompatible with Euclidean geometry [1].

For these reasons, we chose to avoid a direct contact-to-distance step in our technique. Another problem we had to face when trying to estimate the chromosome structure was the above-mentioned size of the data record, and the related computational burden. The solution we propose exploits the existence of isolated genomic regions (the Topological Association Domains, or TADs) characterized internally by highly interacting fragments, and by relatively poor interactions with any other segment of the genome. This allows us to isolate each TAD and reconstruct its structure from the relevant data set, independently of the rest of the genome, then lower the resolution, considering each TAD as a single chain element, and then take the weaker interactions between TAD pairs into account, in a sort of recursive, multiresolution approach.

The result is an algorithm (CHROMSTRUCT [2]) characterized by:

- A new modified-bead-chain model of the chromosomes;
- A set of geometrical constraints producing solutions with consistent shapes and sizes;
- A likelihood function that does not contain target distances derived from the contact frequencies – in the present version, this likelihood is sampled by a Monte Carlo strategy to estimate a number of feasible structures for each TAD;
- A recursive framework to associate the structure of each reconstructed TAD with the shape and the size of a single bead in a lower-resolution chain, whose structure, in turn, is estimated on the basis of an appropriately binned data set;
- A recursive framework to build the final structure from the partial results at the different levels of genomic resolution.

So far, we have tested our algorithm on part of the human genome (29.2 Mbp from chromosome 1, at 100 kbp resolution, see Figure 1). The geometrical features of many of our results correlate positively with known functional features of the cells considered in our tests. To conclude our research, and to be able to assess our results against more detailed biological properties, we still need to remove the experimental biases from the raw data, and then try our strategy on larger parts of (or an entire) genome.

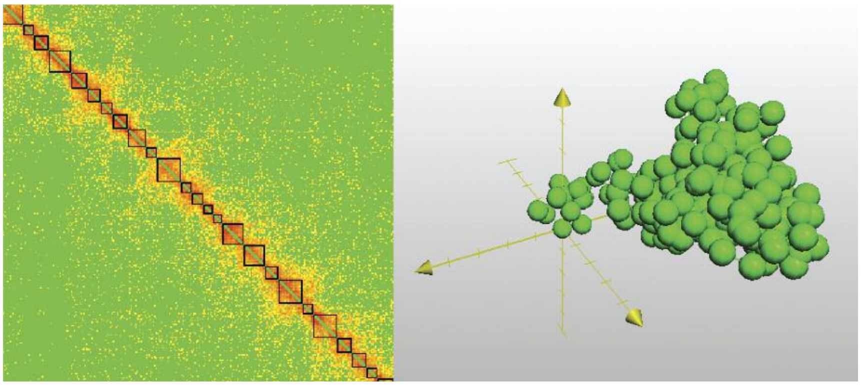


*Figure 1: Left: Contact frequency matrix for a segment of the long arm of chromosome 1 (q from 150.28 Mbp to 179.44 Mbp) from human lymphoblastoid cells GM06990, in logarithmic colour scale. Data from [3]; genomic resolution 100 kbp. The highlighted diagonal blocks define our maximum-resolution TADs. Right: one of our reconstructed structures, consisting of a chain with 292 beads.*

**Link:**
InterOmics Flagship Project:
http://www.interomics.eu/web/guest/home

**References:**

[1] C. Caudai, et al.: “A statistical approach to infer 3D chromatin structure”, in V. Zazzu et al. (Eds.), Mathematical Models in Biology, Springer-Verlag, to appear, DOI: 10.1007/978-3-319-23497-7_12.

[2] C. Caudai, et al.: “Inferring 3D chromatin structure using a multiscale approach based on quaternions”, BMC Bioinformatics, Vol. 16, 2015, p. 234-244, DOI: 10.1186/s12859-015-0667-0

[3] E. Lieberman-Aiden, et al., “Comprehensive mapping of long-range interactions reveals folding principles of the human genome”, Science, Vol. 326, 2009; pp. 289-293, DOI: 10.1126/science.1181369.

**Please contact:**

Claudia Caudai, Emanuele Salerno
ISTI-CNR, Italy
E-mail: claudia.caudai@isti.cnr.it, emanuele.salerno@isti.cnr.it

# Modelling Approaches to Inform the Control and Management of Invasive Seaweeds

by James T. Murphy, Mark Johnson and Frédérique Viard

***Invasive non-native plant and animal species are one of the greatest threats to biodiversity on a global scale. In this collaborative European project, we use a computer modelling approach (in association with field studies, ecological experiments and molecular work) to study the impact of an important invasive seaweed species (Undaria pinnatifida) on native biodiversity in European coastal waters under variable climatic conditions.***

The introduction of non-native species can transform habitats and disrupt ecosystems resulting in serious environmental and economic consequences. Non-native seaweeds represent one of the largest groups of marine invasive organisms in Europe. However, often the fundamental processes that affect their population dynamics and invasion success are poorly understood making it difficult to develop optimal management strategies at both a local and international scale.

The Asian kelp species *Undaria pinnatifida* (Wakame) has been nominated as one of the world's 100 worst invasive species according to the Global Invasive Species Database [1]. This species is the focus of a collaborative European research project supported by an Irish-Research Council-Marie Curie Actions co-funded Elevate international career development fellowship (2013-

16). It involves a close collaboration between the Ryan Institute at the National University of Ireland, Galway and the Station Biologique de Roscoff, France (jointly operated by the French National Centre for Scientific Research (CNRS) and the Pierre & Marie Curie University, Paris (UPMC)).

The aim of this project is to develop a robust, scalable agent-based modelling tool to investigate expansion rates, abundance and ecological functioning of invasive seaweeds. The agent-based (or individual-based) modelling paradigm is a useful approach for modelling algal populations as it allows the large-scale population dynamics to be derived from simple rules dictating the growth and interactions of the individual members of the population.

Some of the questions that we aim to explore, by combining a theoretical and field-based approach, include:
- What are the main life-history traits responsible for the invasive potential of seaweeds?
- Can we determine the dispersal potential and thus expansion rates and patterns of invasion?
- How does the interplay between natural and human-assisted dispersal affect the expansion rates of marine invaders?
- Can we predict the potential range of invasive seaweeds under present and future climatic conditions?

The model framework was developed in the C++ programming language and is fully parallelisable to take advantage of high performance computing resources [2]. It represents a robust and adaptable tool to simulate spatially and temporally heterogeneous phenomena in a coastal environment. A detailed model of the life cycle of *U. pinnatifida* (including the distinct microscopic gametophyte and macroscopic sporophyte stages) was then built upon this basic framework (see Figure 1).

Quantitative data from the literature on the responses of the individual life stages to environmental factors such as light and temperate were used in order to build up a mechanistic model of the overall population growth. For example, photosynthesis-irradiance curves are used to represent the effect of sunlight levels on growth, and thermal performance curves are incorporated to account for the effect of water temperature on development. This allows us to integrate the effects of different environmental conditions on growth dynamics and make predictions about potential expansion into new habitat ranges.

Simulations have been carried out, using environmental parameters representative of Brittany, France, in order to validate the model against field data collected by researchers at the Station Biologique de Roscoff, France. Initial results are promising and indicate that the model can accurately predict the growth dynamics of an *U. pinnatifida* population in situ and allow us to trace back the behaviour of the population to the complex interactions at the individual level. Future work will involve investigations into how changing climatic conditions may affect the range expansion of this species in the future.

The problem of biological invasion and resulting biodiversity loss in coastal habitats is a complex question that cannot be answered by purely theoretical or empirical means. This project aims to tackle this question using the combined toolsets of both fields of study and to integrate them into a coherent strategy for control and maintenance of biodiversity in coastal habitats. This type of multi-disciplinary research will help to inform decision processes in relation to the control/management of invasive species in the marine environment at both a national and regional scale.

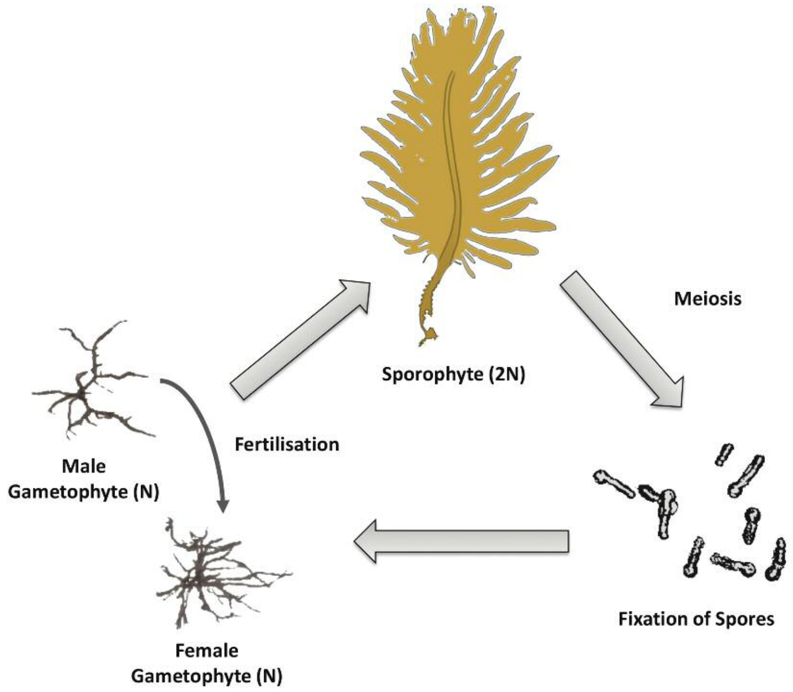


*Figure 1: A schematic representation of the annual life cycle of the invasive kelp species Undaria pinnatifida, which alternates between the microscopic gametophyte stages and the sporophyte stage (which can grow up to 3 metres in length).*

**Links:**
http://www.sb-roscoff.fr/en/divco-diversity-and-connectivity-coastal-marine-landscapes/team
http://www.ryaninstitute.ie

**References:**

[1] IUCN: "Global Invasive Species Database", ed: IUCN Species Survival Commission.

[2] J. T. Murphy, M. P. Johnson, F. Viard: "An agent-based modelling approach to biological invasion by macroalgae in european coastal environments", in Proc. of the European Conference on Complex Systems 2014, Springer, 2015.

**Please contact:**

James T. Murphy,
Adaptation & Diversity in Marine Environment (UMR 7144 CNRS-UPMC), Station Biologique de Roscoff, France
Tel: +33 2 98 29 56 57
E-mail: jmurphy@sb-roscoff.fr
https://www.researchgate.net/profile/James_Murphy35

# Kbdock – Searching and Organising the Structural Space of Protein-Protein Interactions

by Marie-Dominique Devignes, Malika Smaïl-Tabbone and David Ritchie

***Big data is a recurring problem in structural bioinformatics where even a single experimentally determined protein structure can contain several different interacting protein domains and often involves many tens of thousands of 3D atomic coordinates. If we consider all protein structures that have ever been solved, the immense structural space of protein-protein interactions needs to be organised systematically in order to make sense of the many functional and evolutionary relationships that exist between different protein families and their interactions. This article describes some new developments in Kbdock, a knowledge-based approach for classifying and annotating protein interactions at the protein domain level.***

Protein-protein interactions (PPIs) are fundamental biophysical interactions. Our understanding of many biological processes relies on the 3D-modelling of PPIs, and there is a growing need to be able to classify and analyse the structural repertoire of known PPIs using computational modelling and analysis techniques. There are currently over 105,000 experimentally determined 3D structures of proteins and protein complexes in the publicly available Protein Data Bank (PDB). Each entry is composed of the 3D coordinates of several thousands (sometimes even millions) of atoms belonging to one or more linear chains of amino-acid residues. By analysing multiple protein structures and their amino-acid sequences, it can be seen that many protein chains are modular, being composed of one or more structural domains. Thus, protein domains may be considered as “knowledge units” because they represent abstract classes that group together proteins with similar amino-acid subsequences and similar biological properties. We instigated Kbdock (Kb for “knowledge-based”) to address the problem of PPI modelling and classification at the domain level. In fact, as in many other complex scientific domains, big data approaches in the life sciences can only become viable by explicitly considering and making use of prior knowledge.

Essentially, Kbdock is a dedicated relational database which combines the Pfam domain classification with coordinate data from the PDB to analyse and model domain-domain interactions (DDIs) in 3D space. The detection of a DDI instance relies on the computation of the spatial distance between the surface residues of the domain instances found in each PDB entry. In the latest release of Kbdock, after duplicate or near-duplicate interactions are removed, a total of 5,139 distinct non-redundant DDIs involving two different domains have been identified from nearly 240,000 DDI instances extracted from the PDB. As illustrated in Figure 1A, the Kbdock resource can be queried by Pfam domain to visualise the DDI network involving this domain. Otherwise, Kbdock can return the list of DDI instances corresponding to a given pair of domains contained in two interacting proteins, even when there is little or no similarity with the query proteins. Moreover, calculating 3D super-positions of all DDI instances involving a given domain (Figure 1B) enabled us to perform a spatial clustering of all DDIs involving that domain, thereby identifying a discrete number of “domain family binding sites” (DFBSs) [1] on the domain of interest. This gives us a new and original kind of knowledge unit, which is essential when studying the structure and specificity of protein binding sites on a large scale [2]. The notion of DFBSs also led us to propose a case-based reasoning approach to the problem of how to retrieve the best

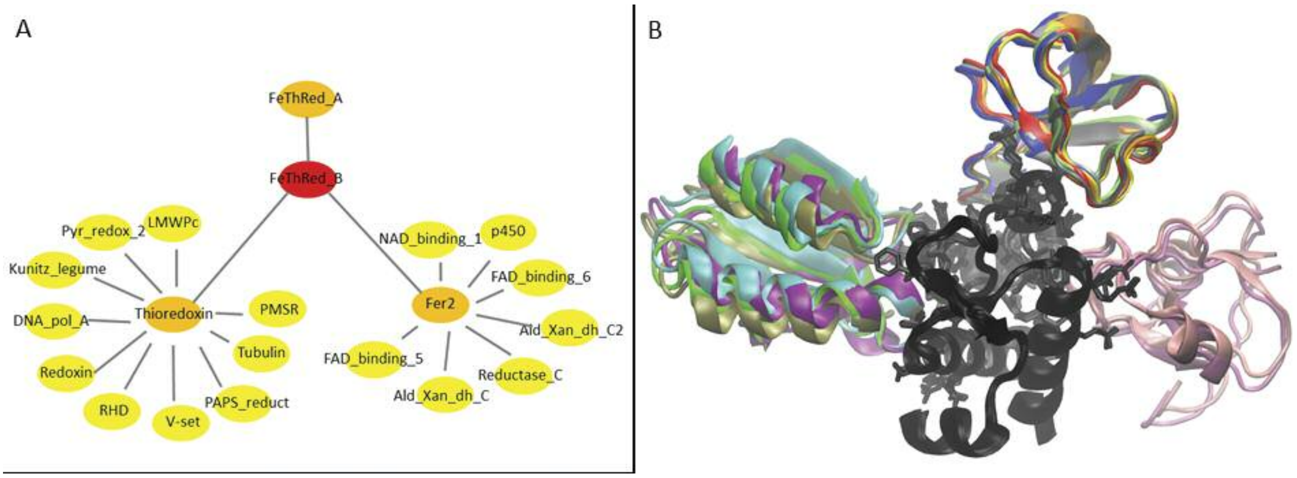


*Figure 1: Kbdock answers for a given domain (PF02943: FeThRed_B or Ferredoxin-Thioredoxin Reductase beta chain): (A) the graph of DDIs around this domain in Kbdock (depth = 2); (B) the superposed 3D DDI instances involving this domain.*

available template for modelling protein-protein "docking" interactions.

The Kbdock project is run as a collaboration between the Capsid and Orpailleur teams at the Loria/Inria research center in Nancy. It is funded and supported by Inria, the CNRS, and the University of Lorraine, as well as specific ANR ("Agence Nationale pour la Recherche") grants. The Kbdock program is available through its online interface. It may also be queried programmatically by expert users in order to execute complex or specialised queries. Recent developments to Kbdock make use of a novel protein structure alignment algorithm called "Kpax" that we have developed [3]. This allows queries in Kbdock to span structural neighbours of the retrieved DDIs, thus allowing Kbdock to search over more distant regions of protein structure space and to propose protein docking templates that cannot be found using conventional sequence-based or structure-based comparison techniques.

We are currently working to link KBdock's structural domain binding site classification with the widely used ExPASy Enzyme Classification scheme. In order to achieve this, we are developing efficient data-mining approaches to process the millions of sequence-function associations that are now available in large molecular biology databases, such as Swiss-Prot and TrEMBL, which together build the UniProt Knowledgebase at the European Bioinformatics Institute.

**Links:**
http://kbdock.loria.fr
http://kpax.loria.fr, http://hex.loria.fr
Protein Data Bank:
http://www.rcsb.org/pdb/home/home.do
Pfam domain classification:
http://pfam.xfam.org/
UniProt Knowledgebase:
http://www.ebi.ac.uk/uniprot
ExPASy: http://enzyme.expasy.org

**References:**

[1] A. W. Ghoorah et al.:"Spatial clustering of protein binding sites for template based protein docking", Bioinformatics, 2011 Oct 15; 27(20):2820-7.

[2] A. W. Ghoorah et al.: "A structure-based classification and analysis of protein domain family binding sites and their interactions", Biology (Basel), 2015 Apr 9; 4(2):327-43.

[3] D. W. Ritchie et al: "Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity", Bioinformatics. 2012 Dec 15; 28(24):3274-81.

**Please contact:**

David Ritchie
Inria, France
E-mail dave.ritchie@inria.fr

# The Source of the Data Flood: Sequencing Technologies

by Alberto Magi, Nadia Pisanti and Lorenzo Tattini

***Where does this huge amount of data come from? What are the costs of producing it? The answers to these questions lie in the impressive development of sequencing technologies, which have opened up many research opportunities and challenges, some of which are described in this issue. DNA sequencing is the process of "reading" a DNA fragment (referred to as a "read") and determining the exact order of DNA bases (the four possible nucleotides, that are Adenine, Guanine, Cytosine, and Thymine) that compose a given DNA strand. Research in biology and medicine has been revolutionised and accelerated by the advances of DNA and even RNA sequencing biotechnologies.***

The sequencing of the first human genome was completed in 2003 (though a draft was already complete by 2001). This achievement cost 100 million US dollars and took 13 years. Now, only 12 years later, the cost for the equivalent process has dropped to just over $1,000 and takes just a few days. As a consequence, sequencing the human exome (the coding part of the genome), for example, or the whole genome, has become common practice in medicine, and genomic sequencing of many other species has paved the way to new challenges and research fields in the life sciences.

The first breakthrough in the history of DNA sequencing was the development of Frederick Sanger's method using chain-termination with dideoxy nucleotides in 1977, which earned Sanger his second Nobel Prize in 1980. Sanger sequencing (SSeq) is a sequencing-by-synthesis technique since it requires DNA polymerase enzymes to produce the observable output by means of nucleotide radiolabelling or fluorescent dyes.

From around 2005-2006, second-generation sequencing (SGS) produced a dramatic cost reduction, and from that point on, we have seen a growing diffusion of sequencing machines that have revolutionised clinical research and practice, as well as molecular biology investigations in genomics [1]. The higher error rate (compared with SSeq) is balanced out by the impressive throughput of SGS platforms. Though they still rely on sequence-by-synthesis (using chemi- or photo-luminescence), these platforms are based on various implementations of cyclic-array sequencing. SGS methods (in particular Illumina and Roche) are  widely used for the investigation of the whole spectrum of genomic variants from single nucleotide variants (SNVs) to structural variants (SVs). Other implementations of SGS include SOLiD and Complete Genomics' nanoball sequencing. Notably, Ion Torrent was the first platform based on non-optical (i.e., electrochemical) methods

Today's low cost of sequencing allows any reasonably funded lab to sequence multiple genomes, thus raising new clinic and research issues. Sequencing several individuals of the same species, for example, is now common for personalised medicine and for under-

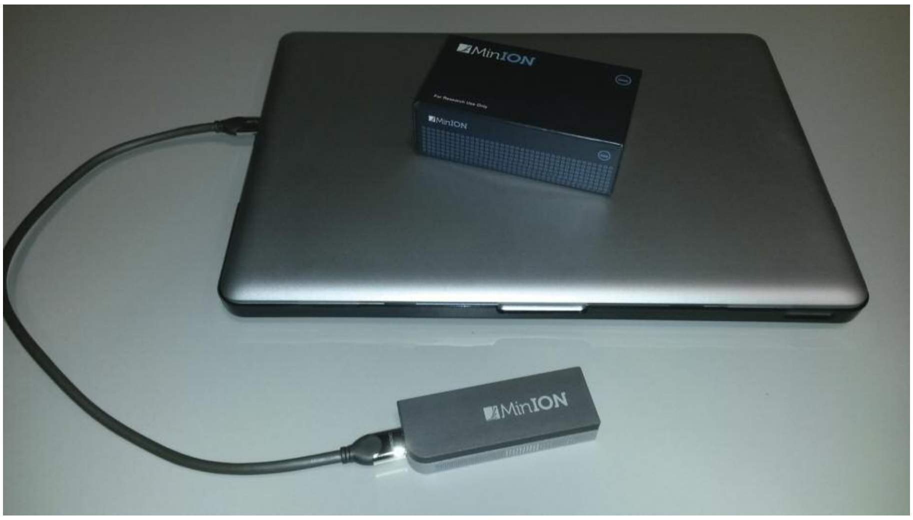


*Figure 1: USB-sized Nanopore MinION.*

standing how we differ genetically from each other. Also, RNA-Seq – that is sequencing genes that are transcribed for proteins synthesis – allows us to take a (possibly comparative) picture of which genes are expressed under certain conditions. Moreover, sequencing several strands of the same species allows us to investigate intra-species mutations that involve the mobile element of the genomes. Finally, "metagenomics" studies the microbiology of genetic material that is recovered from non-cultivated environments (e.g., soil, gut, sea-depths) and sequenced.

A new revolution is currently underway with "third-generation sequencing" (TGS) techniques [2]. These platforms are based on single molecule real time sequencing, a single DNA molecule sequencing approach. While PacBio platforms exploit an optical detection method, Oxford Nanopore Technologies are based on ionic current measurements [3]. Both platforms show high error rates, though the length of the reads produced is up to thousands of base-pairs. When we have a robust (i.e., low error rates) technology capable of such long reads – which will probably be soon – we will certainly witness new challenges in genomics.

The machine costs and sizes vary considerably, as does the throughput (that is, the amount sequenced base-pairs per hour), even within the same generation. In general, both price and size grow with the throughput. Machines range from the huge and expensive Illumina HiSeq (as big as a closet) to the smaller (desktop-sized) Illumina MiSeq and Ion Torrent, and the even smaller USB-sized Nanopore MinION, shown in Figure 1, passing through the desktop sized ones. Other performance parameters can instead be grouped according to generation: Table 1 reports the cost per base of sequencing, the length of fragments that can be output, and the error rate for each technology generation.

| Method | Cost per Base ($/Mbp) | Read Length | Error Rate |
|---|---|---|---|
| SSEQ | 400 (up to 2007) | 300-1000 | $10^{-5}$-$10^{-2}$ |
| SGS | 0.015 (2015) | O($10^{2}$) | $10^{-2}$ |
| TGS | 0.5 (PacBio), 0.65 (Nanopore) | O($10^{3}$) | $10^{-1}$ |

**References:**

[1] T. P. Niedringhaus et al.: "Landscape of next-generation sequencing technologies", Analytical chemistry 83.12 (2011): 4327-4341.

[2] C. S. Pareek, R. Smoczynski, A. Tretyn: "Sequencing technologies and genome sequencing", Journal of applied genetics 52.4, 2011, 413-435.

[3] G. F. Schneider, C. Dekker: "DNA sequencing with nanopores", Nature biotechnology 30.4, 2012, 326-328.

**Please contact:**
Nadia Pisanti
University of Pisa and Inria
E-mail: pisanti@di.unipi.it

# Big Data in Support of the Digital Cancer Patient

by Haridimos Kondylakis, Lefteris Koumakis, Manolis Tsiknakis, Kostas Marias and Stephan Kiefer

***The iManageCancer project is developing a data management infrastructure for a cancer specific self-management platform designed according to the patients' needs.***

A recent report by the eHealth Task Force entitled "Redesigning health in Europe for 2020" focuses on how to achieve a vision of affordable, less intrusive and more personalized care, ultimately, increasing quality of life as well as lowering mortality. Such a vision depends on the application of ICT and the use of data, and requires a radical redesign of e-health to meet these challenges. Two levers for change as identified by the report are "liberate the data" and "connect up everything". Fully exposing, integrating, linking and exploring health data will have a tremendous impact on improving the integrated diagnosis, treatment and prevention of disease in individuals. In addition it will allow for the secondary use of care data for research transforming the way in which care is provided [1].

The iManageCancer H2020 EU project, started in February 2015, aims to provide a cancer specific self-management platform designed according to the needs of patient groups while focusing, in parallel, on the wellbeing of the cancer patient with special emphasis on avoidance, early detection and management of adverse events of cancer therapy but also, importantly, on psycho-emotional evaluation and self-motivated goals. The platform will be centered in a Personal Health Record which will regularly monitor the psycho-emotional status of the patient and will periodically record the everyday life experiences of the cancer patient in terms of side effects of therapy, while different groups of patients and their families will share information through diaries, and clinicians are provided with clinical information.

The data collected in this context are complex, with hundreds of attributes per patient record that will continually evolve as new types of calculations and analysis/assessment results are added to the record over time (volume). In addition, data exist in many different formats, from textual documents and web tables to well-defined relational data and APIs (variety). Furthermore, they pertain to ambiguous semantics and quality standards resulting from different collection processes across sites (veracity). The vast amount of data generated and collected comes in so many different streams and forms — from physician notes, personal health records, images from patient scans, health conversations in social media (variability), to continuous streaming information collected from wearables and other monitoring devices (velocity). These data, if used to their full potential, may have a tremendous impact on healthcare, delivering better outcomes at a lower cost (value). As such, key questions to address include: How can we develop optimal frameworks for large-scale data-sharing? How can we exploit and curate data from various electronic and patient health records, assembling them into ontological descriptions relevant to the practice of systems medicine? And how can we manage the problems associated with large scale medical data?

The high level data management architecture the iManageCancer is shown in Figure 1. Within iManageCancer, Apache Cassandra is used as an instantiation of a "data lake" concept to store this vast amount of raw data in its native format including structured, semi-structured and unstructured data. Cassandra is an open-source, peer-to-peer and key value based store, where data are stored in key spaces, has built-in support for the Hadoop implementation of MapReduce and advanced replication functions and is currently considered state of the art for real-time big data analysis. On top of the Cassandra, the

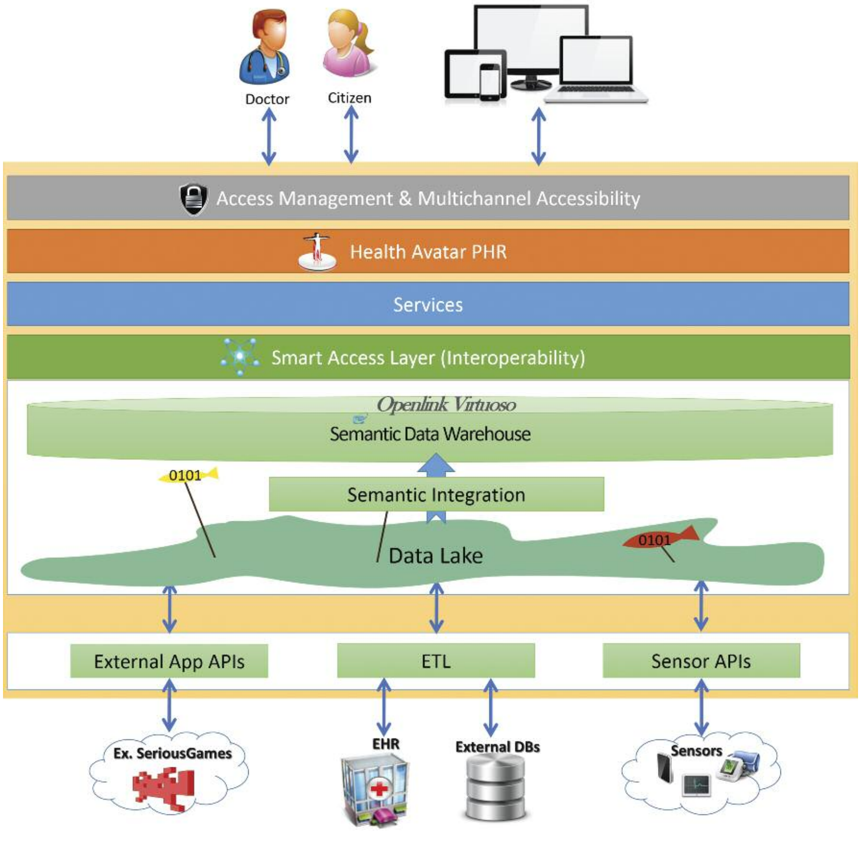


*Figure 1: The data management architecture of the iManageCancer platform.*

Semantic Integration Layer [2] pushes selected data to the Semantic Data Warehouse. Using this architecture we can select which of the available data should be semantically linked and integrated by establishing the appropriate mappings to a modular ontology [3]. Then these data are queried, transformed into triples and loaded to the Semantic Warehouse where they are available for further reasoning and querying. A benefit of the approach is that we can recreate from scratch the resulting triples at any time. However, for reasons of efficiency the data integration engine periodically transforms only the newly inserted information by checking the data timestamps.

In addition to off-line transformation, on-line transformation is also possible by issuing transformation events through an event bus. As such, our architecture adopts a variation of the command-query responsibility segregation principle where, in order to update information, one uses a different model to the model that one is using to read. We choose to store the original data using NoSQL technologies owing to their ability to handle enormous data sets and their “schema-less” nature. But the limited flexibility of their query mechanisms is a real barrier for any application that has not predetermined access use cases. The Semantic Warehouse component in the iManageCancer platform fills these gaps by effectively providing a semantically enriched and search optimized index to the unstructured contents of the Cassandra repository. Therefore, our approach tries to offer best of both worlds: efficient persistence and availability of heterogeneous data, and semantic integration and searching of the “essence” of the ingested information.

A key next step is to develop the eHealth services on top of this data management infrastructure and to test the whole platform in a real-world context in two clinical pilots, one for children starting in 2016 and one for adults starting in 2017. Big Data management is undoubtedly an important area for healthcare, and it will only become more critical as healthcare delivery continues to grapple with current challenges.

**Links:**
Redesigning health in Europe for 2020: http://ec.europa.eu/digital-agenda/en/news/eu-task-force-ehealth-redesigning-health-europe-2020
http://imanagecancer.eu/
http://www.myhealthavatar.eu/
http://cassandra.apache.org/

**References:**

[1] H. Kondylakis et al.: “Digital Patient: Personalized and Translational Data Management through the MyHealthAvatar EU Project”, EMBC, 2015.

[2] H. Kondylakis and D. Plexousakis: “Exelixis: Evolving Ontology-Based Data Integration System”, SIGMOD/PODS, 2011.

[3] H. Kondylakis et al.: “Agents, Models and Semantic Integration in support of Personal eHealth Knowledge Spaces”, WISE 2014.

**Please contact:**

Haridimos Kondylakis
FORTH-ICS, Greece
E-mail: kondylak@ics.forth.gr

# Towards an On-board Personal Data Mining Framework For P4 Medicine

by Mohamed Boukhebouze, Stéphane Mouton and Jimmy Nsenga

***A personal on-board data-mining framework that relies on wearable devices and supports on-board data stream mining can help with disease prediction, risk prevention, personalized intervention and patient participation in healthcare. Such an architecture, which allows continuous monitoring and real-time decision-making, can help people living with diseases such as epilepsy.***

The emergence of systems biology, big data tools and wearable computers is transforming medicine from a reactive paradigm that focuses on treating disease, to “P4 medicine” that focuses on predicting diseases, preventing the risks of developing diseases, personalizing interventions based on the patient context and participation of the patients in the management of their own health [1]. This new paradigm pledges to improve healthcare through early detection of disease and helping patients to make decisions about prevention and treatment.

For the P4 medicine approach to be successful, personal (patient-specific) data should be collected and mined at a large scale in order to discover hidden patterns and make effective decisions. The wearable devices can perform ubiquitous collections of a large subset of personal data and be used as interface to data mining algorithms. The latter are executed remotely on cloud-based infrastructure that provides high performance processing resources [2]. However, for P4 medicine to fully accomplish its goals, wearable devices must also execute offline data stream mining. Indeed, the on-board data mining execution allows continuous monitoring and real-time processing, which can help with early disease detection and prevention. Therefore, on-board execution could improve the mobility and safety of patients and enhance their quality of life.

Despite the great progress in wearable technology and ubiquitous computing, many issues must be addressed in order to implement efficient on-board data mining for P4 medicine, including:

- Resource limitation. Wearable devices are limited in terms of computational power, energy consumption, storage and memory capacity and bandwidth.
- Context and resource changes. Ubiquitous data stream mining has to deal with patient situation changes and device disconnection.

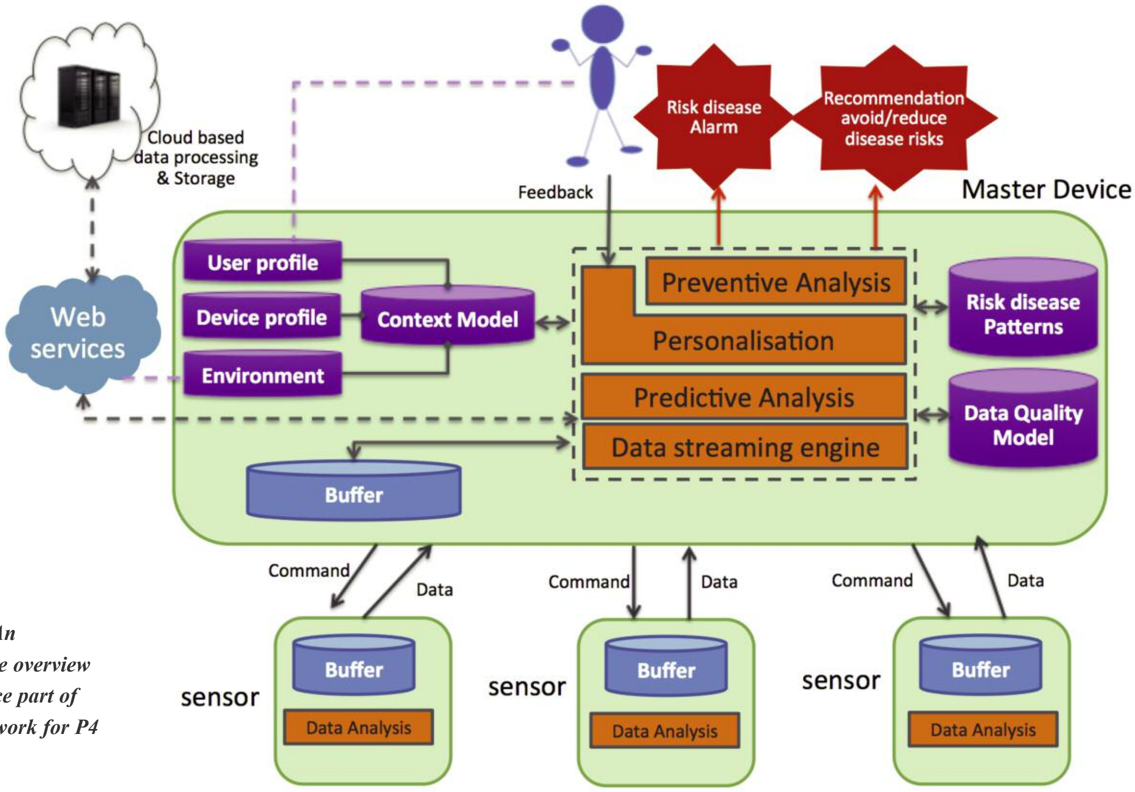


*Figure 1: An architecture overview of the device part of the Framework for P4 Medicine.*

- Data quality. Despite accrued sophistication of current sensors, accuracy of gathered data is not always ensured. Furthermore, variable data rates can cause problems [2].

To deal with these issues, we propose an on-board personal data-mining framework for P4 medicine, currently under development in our research centre, relying on a network of wearable devices to monitor the patient in real-time based on P4 medicine principles:

- Predicting the risk of developing diseases or faintness (e.g., epilepsy seizure, heart attack) based on well-defined patterns. The framework notifies patient once a risk is identified.
- Preventing certain diseases or faintness by providing recommendations that help to avoid or reduce the risk of developing them.
- Personalizing the prediction and prevention recommendations by taking into account the context of patients (profile, medical history, actual environment, etc.).
- Participation of the patients in the learning and refinement of the prediction and prevention models by providing feedback about issued notification and recommendations.

The proposed framework deals with the on-board processing issues by combining three approaches (see Figure 1):

- A distributed data mining approach is used to execute data mining algorithms concurrently over the network of wearable devices. This approach helps to address the computational constraints and energy efficiency issues by distributing data processing across devices.
- A context-aware and resource-aware adaptation strategy is used to detect changes of patient situation or devices and automatically adjust the parameters of data mining algorithms. This approach is also used to dynamically distribute the load of data mining algorithms based on resource availability.
- A probabilistic data mining approach is used to deal with data quality issues by computing data variation and uncertainty. In this way, risk prediction and prevention recommendations should be improved because the full range of possible outcomes can be taken into account.

To validate our proposed framework, we are considering epilepsy as a case study – in particular, refractory epilepsy patients who do not respond to conventional treatment. Thirty percent of people with epilepsy have refractory epilepsy, [3] a very debilitating form of the disease since the seizures are random and not controllable with medication. The proposed on-board data-mining framework can be used to continuously monitor refractory epilepsy patients and notify them of seizure risks based on early detection patterns that are developed by our academic partner LISA Laboratory – Université Libre de Bruxelles. These notifications contribute to improve the safety and quality of life of individuals affected by refractory epilepsy. Furthermore, the framework can provide context-aware recommendations to reduce seizure risk by analysing the triggering factors (e.g., stress level), consequently helping to reduce seizure frequency. Finally, the collected data could be used by epilepsy specialists to better identify appropriate treatment.

**Links:**
P4 Medicine Institute: http://www.p4mi.org
LISA Laboratory: lisa.ulb.ac.be

**References:**

[1] L. Hood and S. H. Friend: "Predictive, personalized, preventive, participatory (P4) cancer medicine," Nature Reviews Clinical Oncology, vol. 8, no. 3, pp. 184–187, 2011.
[2] M.M. Gaber, J.B. Gomes, and F. Stahl: "Pocket data mining". Springer Book, 2014, ISBN: 9783319027111, pp. 1–112.
[3] P. Kwan, and M.J. Brodie: "Early identification of refractory epilepsy." New England Journal of Medicine, 2000, 342(5), 314-319.

**Please contact:**

Mohamed Boukhebouze,
CETIC, Belgium
Mohamed.boukhebouze@cetic.be

# Can Data-driven Self-Management Reduce Low Back Pain?

by Kerstin Bach, Paul Jarle Mork and Agnar Aamodt

***A new Horizon 2020 research and innovation project will start the development of the SELFBACK decision support system for self-management of low back pain in January 2016.***

Low back pain is a common reason for activity limitation, sick leave, and disability. It is the fourth most common diagnosis (after upper respiratory infection, hypertension, and coughing) seen in primary care [1]. An expert group concluded that the most effective approach to manage non-specific low back pain is to discourage bed rest, to use over-the-counter pain killers in the acute stage if necessary (e.g., to be able to sleep), reassure the patient about the favourable prognosis, advise the patient to stay active, and advise strength and/or stretching exercise to prevent recurrence [2].

In January 2016 a new project funded by the European Commission's Horizon 2020 research and innovation programme will start the development of the SELFBACK decision support system for self-management of low back pain. The consortium includes the Norwegian University of Science and Technology (coordinator), the University of Glasgow and the Robert Gordon University in Aberdeen from the UK, the SMEs Kiolis from France and Health Leads from the Netherlands as well as the National Research Centre for the Working Environment in Denmark and the University of Southern Denmark.

Self-management in the form of physical activity and strength/stretching exercises constitute the core component in the management of non-specific low back pain; however, adherence to self-management programs is poor because it is difficult to make lifestyle modifications with little or no additional support. Moreover, the way the self-management advice is presented, and whether the advice is followed up by reinforcement have substantial impact on adherence to a self-management plan and progress of symptoms. In the SELFBACK project we will develop and document an easy-to-use decision support system to facilitate, improve and reinforce self-management of non-specific low back pain. The decision support system will be conveyed to the patient via a smart-phone app in the form of advice for self-management. Even though there are about 300 pain-related apps are available on the market, to the best of our knowledge, none of these apps have documented effects by scientific publications and none include an active decision support system. In contrast, we will document the effectiveness of the SELFBACK decision support system by conducting a nine-month randomized control trial.

The SELFBACK system will constitute a data-driven, predictive decision support system that uses the Case-Based Reasoning (CBR) methodology to capture and reuse patient cases in order to suggest the most suitable activity plans and goals for an individual patient. Based on personal preferences and clinical guidelines, the SELFBACK system will guide patients to adjust their lifestyle in order to optimize their management of low back pain, based on patient's activity data collected using a wristband. The incoming data will be analysed to classify the patient's activities and matched against the proposed activities in order to follow-up and advise the patient. The challenge of the data analysis is activity pattern detection, and matching it against existing patient profiles in order to suggest activity goals aiming at the most favourable outcome for the patient. Therefore we will enrich the data stream and the resulting patterns with the patient's symptom state, symptom progression and goal-setting.

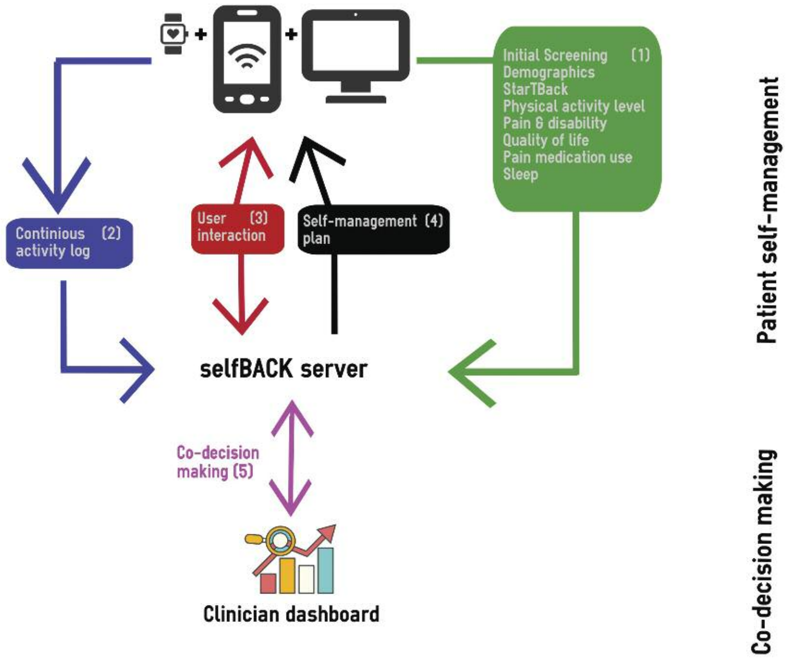


*Figure 1: SELFBACK data and decision process model.*

Figure 1 gives an overview of the SELFBACK proposed data and decision process model that contains five processing modules, of which four modules target self-management by the patient while the fifth module targets the possibility for co-decision making between the patient and a clinician.

The initial screening module (1) acquires basic patient data, which is sent to the SELFBACK server. This data will be collected through a web-page and will provide the starting point for assessing the patient's status and for running the self-management planning module (4) for the first time. The plan for self-management will be updated and revised as more patient data is added. The activity logging module (2) runs continuously since it provides data from the wristband regarding the patient's physical activity and sleep patterns. The main goal of this module is to verify whether the patient follows the suggestions in the recommended plan (e.g., for physical activity). As for the initial screening data, the activity log data is also sent to the server, and becomes input to the periodic run of the self-management planning module (4). The user interaction module (3) is basically a question/answering module, typically initiated by the patient when using the SELFBACK app and browsing through the given information.

Overall, the aim of the SELFBACK system is to solve the problems of reinforcement and adherence to a self-management plan by offering an evidence-based system that allows personalized follow-up and advice to the patient.

**Link:**
SELFBACK project:
http://research.idi.ntnu.no/selfback/

**References:**

[1] P. Wändell et al.: “Most common diseases diagnosed in primary care in Stockholm”, Sweden, in 2011. Fam Pract 30:506-513, 2013.

[2] M. van Tulder et al.: “Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care. Eur Spine J 15:s169-s191, 2006.

**Please contact:**

Kerstin Bach, Paul Jarle Mork, Agnar Aamodt
Norwegian University of Science and Technology
E-Mail: kerstin.bach@idi.ntnu.no,
paul.mork@ntnu.no,
agnar.aamodt@idi.ntnu.no

# Twitter can Help to Find Adverse Drug Reactions

by Mark Cieliebak, Dominic Egger and Fatih Uzdilli

***Drugs are great! We all need and use drugs every now and then. But they can have unwanted side-effects, referred to as “adverse drug reactions” (ADRs). Although drug manufacturers run extensive clinical trials to identify these ADRs, there are still over two million serious ADRs in the U.S. every year – and more than 100,000 patients in the U.S. die due to drug reactions, according to the U.S. Food and Drug Administration (FDA) [1]. For this reason, we are searching for innovative and effective ways to find ADRs.***

Identifying ADRs is an important task for drug manufacturers, government agencies, and public health. One way to identify them before a drug goes to market is through clinical trials. Governments worldwide also have diverse surveillance programs in order to identify ADRs once the drugs are in use by consumers. For example, official websites such as MedWatch allow both patients and drug providers to submit ADRs manually. However, only a very small fraction of all ADRs is submitted to these systems – experts estimate that over 90% of all reactions go unreported.

## Twitter can help!

On the other hand, there are millions of messages on Twitter that discuss medications and their side-effects. These messages contain data on drug usage in much larger test sets than any clinical trial will ever have. Inspired by this, research teams worldwide, including our team at Zurich University of Applied Sciences, are beginning to utilize these messages for ADR detection. The goal is to automatically find relevant messages, to “understand” their content, and to extract structured data about the drugs and (unwanted) reactions.

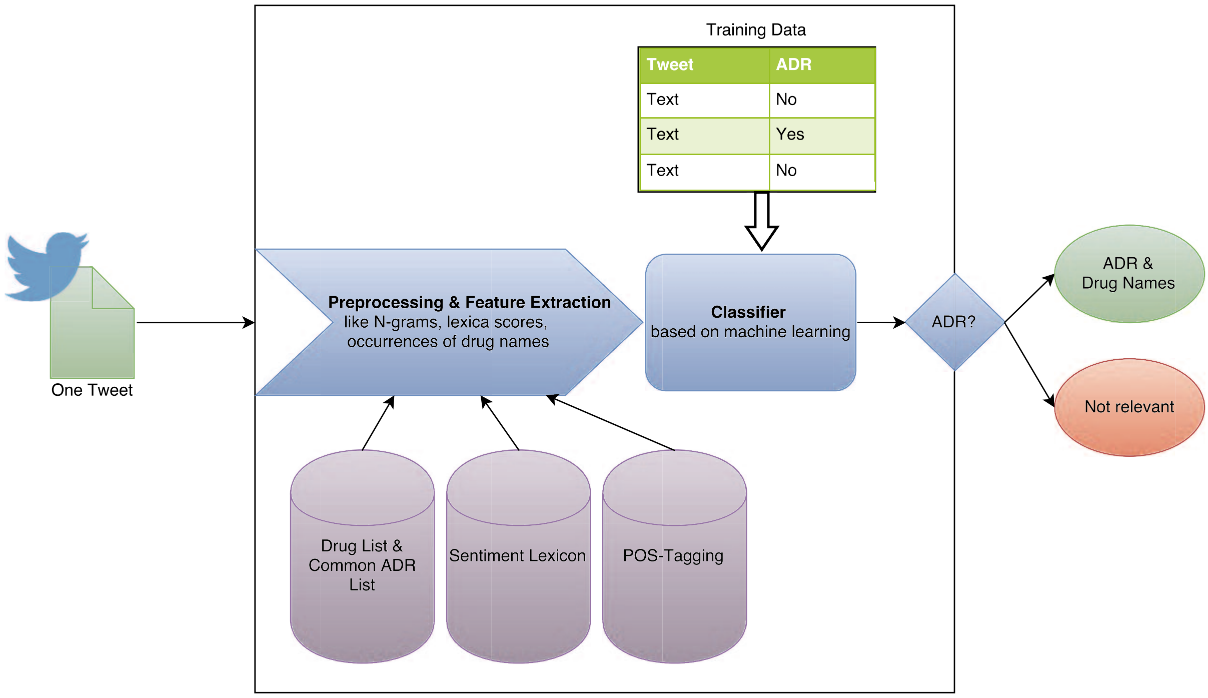
A typical approach for ADR detection uses Natural Language Processing (NLP) to analyze tweets automatically. Input for the system is the entire stream of Twitter messages. Each individual tweet is analyzed, using a classification system as shown in Figure 1: The tweet is preprocessed and a set of relevant properties (“features”) is extracted. Then, a classifier decides whether the tweet mentions an ADR. This classifier is based on machine learning and was trained beforehand on thousands of sample tweets that were tagged by humans. Finally, a system for named entity extraction is used to output a drug name and associated ADRs.

This approach is similar to technologies for sentiment analysis, which decide whether a tweet is positive or negative. Sentiment analysis is already successfully applied, for instance, in market monitoring, customer support and social media analysis.

## State-of-the-art

Our ADR system, which implements the technologies shown above, achieves a success rate of 32% (measured in F1-score). This is comparable to other academic ADR systems: in an open international competition this year, even the best systems achieved only a success rate of approximately 40% [2].

For a preliminary evaluation on real-world data, we applied our ADR system to the full Twitter stream. The low precision of the system resulted in 20% of all tweets being classified as ADR. This is way too high; there are not that many ADR tweets in the Twitter stream. For this reason, we pre-filtered the stream with a list of 1678 drug names. Out of about 50 million tweets, this resulted in 13,000 tweets referencing drugs. Using the ADR system on this reduced set yielded 2800 tweets. We expect to find


Training Data
One Tweet
Preprocessing & Feature Extraction
like N-grams, lexica scores, occurrences of drug names
Classifier
based on machine learning
ADR?
ADR & Drug Names
Not relevant
Drug List & Common ADR List
Sentiment Lexicon
POS-Tagging

| Tweet | ADR |
|---|---|
| Text | No |
| Text | Yes |
| Text | No |

*Figure 1: Typical system for ADR detection using machine learning.*

more than 60% of these to be true ADR tweets.

## Future improvements

Automatic detection of ADRs on Twitter (or other social media channels) is still a very young discipline, which only started some five years ago. There are only a few teams working on the topic at the moment, and a first large-scale benchmark dataset was only published in 2014 [3]. However, we expect a significant improvement in detection rates in the near future, owing in part to the existence of several new technologies in machine learning, such as word embedding and deep learning. These have already been successfully applied to other text analysis tasks and have improved existing benchmark scores there. Applying these technologies to ADR detection will probably help to increase the detection rate significantly. In addition, our team is working on a system that not only analyzes the text of a tweet, but also its context: the timeline of the user, other messages in the same geographic or temporal context etc. This will allow us to “step back” from an isolated event (a single tweet) and see the “whole picture” of the discourse on Twitter.

**References:**

[1] J. Lazarou, B.H. Pomeranz, and P.N. Corey: “Incidence of Adverse Drug Reactions in Hospitalized Patients: A Meta-analysis of Prospective Studies”, JAMA 279(15):1200-1205, 1998.

[2] Pacific Symposium on Biocomputing: http://psb.stanford.edu/workshop/wkshp-smm/

[3] R. Ginn et al.: “Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark”, BioTxtM, 2014.

**Please contact:**

Mark Cieliebak
School of Engineering
Zurich University of Applied Sciences (ZHAW)
Tel: +41 58 934 72 39
E-mail: ciel@zhaw.ch

# Trust for the “Doctor in the Loop”

by Peter Kieseberg, Edgar Weippl and Andreas Holzinger

***The "doctor in the loop" is a new paradigm in information driven medicine, picturing the doctor as authority inside a loop supplying an expert system with data and information. Before this paradigm is implemented in real environments, the trustworthiness of the system must be assured.***

The “doctor in the loop” is a new paradigm in information driven medicine, picturing the doctor as authority inside a loop with an expert system in order to support the (automated) decision making with expert knowledge. This information not only includes support in pattern finding and supplying external knowledge, but the inclusion of data on actual patients, as well as treatment results and possible additional (side-) effects that relate to previous decisions of this semi-automated system.

The concept of the "doctor in the loop" is basically an extension of the increasingly frequent use of knowledge discovery for the enhancement of medical treatments together with the “human in the loop” concept (see [1], for instance): The expert knowledge of the doctor is incorporated into "intelligent" systems (e.g., using interactive machine learning) and enriched with additional information and expert know-how. Using machine learning algorithms, medical knowledge and optimal treatments are identified. This knowledge is then fed back to the doctor to assist him/her (see Figure 1).

### Manipulation Security and Trust

The implementation of the doctor in the loop concept has met several challenges – both of a technical nature and in other areas. One challenge is gaining the acceptance of such systems by doctors themselves, who are often not researchers, but medical practitioners. While privacy and security in biomedical data applications have been discussed extensively in recent years (see [2] for an overview), the topic of trust has been neglected. Nevertheless, it is very important that the trustworthiness of the systems can be guaranteed in order to make the abstract concept of a doctor in the loop practicable.

In this medical context, it is absolutely crucial to ensure that the information sent by the doctor to the machine learning algorithms cannot be manipulated following its submission. In order to guarantee this, a new approach for protecting the underlying data based on a hash chain has been proposed explicitly for doctor in the loop systems (see [3]). This approach takes advantage of the fact that large amounts of data are typically stored in databases. While these are often identified with the abstract data model, they are in reality complex software systems providing a multitude of internal mechanisms that have various functions, such as enhancing performance. This approach utilizes the transaction mechanism for providing transaction safety (i.e., if a crash occurs, the database is brought back into a well-defined state) in order to protect the information sent to the system by the doctor against manipulation, even if the manipulation is carried out by the administrator of the database.

### Legal Issues

Legal issues are often overlooked by purely technical solutions, but are vital for any doctor that is actually participating in such an approach as expert. Important questions that need to be addressed include: What information can be shared between the doctor and the expert system, and what levels of data and privacy protection need to be applied? And who is responsible for the correctness of the results derived from the combination of human knowledge and machine learning algorithms? Although this is an important field, no guidelines are currently available. The issue is further complicated by the differences between national legislations even between member states of the European Union. Defining workflows that clinical doctors can reliably apply without the fear of prosecution lies thus in the focus of the planned RDA (Research Data Alliance) Workgroup “Security and Trust” that held its “Birds of Feather” session on the 6th RDA Plenary on September 24th in Paris [L1]. One of the major goals of this workgroup is to draw together a set of best practices and guidelines in the area of data driven medical research.

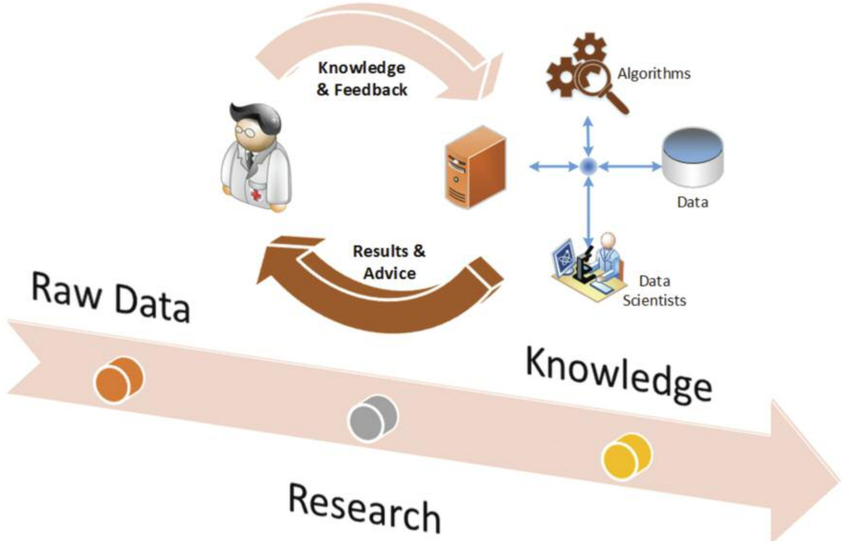


*Figure 1: The Doctor in the Loop.*

### Special focus in CBmed

CBmed [L2] is an interdisciplinary research centre, founded in 2014, with the aim of providing biomarker research, mainly in the areas cancer, metabolism and inflammation. One of the core features of this centre is the tight incorporation of ICT as a horizontal research area that provides data and techniques to all other projects. This does not only apply to applications associated with the doctor in the loop, but also other areas, such as data and privacy protection, the development of new data mining techniques and tools for the efficient analysis of large amount of “–omics” data.

For the CBmed consortium, the model of the doctor in the loop offers abundant possibilities, especially in the area of data driven research. Considering the recent surge in big data related research and developed tools, this approach is expected to be one of the major drivers in medical research in the years to come.

**Links:**

[L1]: https://rd-alliance.org/rda-working-group-data-security-and-trust-wgdst-p6-bof-session.html

[L2]: http://www.cbmed.org/en/

**References:**

[1] W. S. Levine, ed.: “The control handbook”, CRC press, 1996.

[2] P. Kieseberg, H. Hobel, S. Schrittwieser, E. Weippl, A. Holzinger, “Protecting anonymity in data-driven biomedical science”, in “Interactive Knowledge Discovery and Data Mining in Biomedical Informatics” (pp. 301-316), Springer Berlin Heidelberg, 2014.

[3] P. Kieseberg, J. Schantl, P.Fruehwirt and E. R. Weippl, A. Holzinger: “Witnesses for the Doctor in the Loop," in 2015 International Conference on Brain Informatics & Health (BIH), 2015.

**Please contact:**

Peter Kieseberg
SBA Research, Vienna, Austria
E-mail: pkieseberg@sba-research.org

# Big Data Takes on Prostate Cancer

by Erwan Zerhouni, Bogdan Prisacari, Qing Zhong, Peter Wild and Maria Gabrani

***Most men, by the time they reach 80 years of age, get prostate cancer. The treatment is usually an operation or irradiation, which sometimes has complications. However, not every tumour is aggressive, in which case there is no urgent need to remove it. Ascertaining whether a tumour is aggressive or insignificant is difficult, but analysis of big data shows great promise in helping in this process.***

Prostate cancer (PC) represents the second leading cause of cancer related deaths in the Western world. PC is typically diagnosed on the basis of increased levels of serum protein PSA (prostate specific antigen) together with digital rectal examination, and is confirmed by prostate needle biopsies. However, PSA and biopsies often fail to distinguish between clinically indolent and aggressive forms, leading to overtreatment such as unnecessary prostactectomies and irradiation that sometimes greatly deteriorates a patient's quality of life.

Currently, pathologists assess patient biopsies and tissue resections under a microscope, leading to diagnoses that are affected by subjective judgment and intra- and inter-observer variability. The procedure is time consuming and hence low-throughput, and hospitals may generate hundreds or even thousands of tissue samples per day. This number is expected to increase sharply, as the World Health Organization predicts the number of cancer diagnoses to increase by 70% in the next two decades. Moreover, novel staining technologies, such as immunohistochemistry (IHC) and in situ hybridization (ISH) enable the evidencing of molecular expression patterns through multicolour visualization. Such techniques are commonly used for targeted treatment response estimation and monitoring, increasing the need for high-throughput and unbiased digital solutions.

Emerging initiatives in moving hospitals into the digital era use bright-field and fluorescence scanners to convert glass slides of tissue specimens and needle biopsies to virtual microscopy images of very high quality. Slide images are huge, with several thousand pixels per axis, turning digital image analysis into a big data problem. Integration of digitized tissue specimens with an image analysis workflow allows objective and high-throughput evaluation of tissue slides, transforming surgical pathology into a truly quantitative and data-driven science.

We have developed a computational framework to extract unique protein signatures from immunohistochemistry (IHC) stained tissue images. IHC assays provide significant information with respect to cellular heterogeneity and disease progression [1], and are thus used to identify patients most likely to respond to targeted therapy. However, tissue morphology may have different molecular signatures (see Figure 1) owing to genetics and other parameters, such as environment and lifestyle. Currently, IHC image analysis focuses on the staining intensity performed mostly in a manual and thus low throughput and biased way. Emerging computational techniques use metrics, such as the H-score, or the Aperio metric [2]. Recent studies, however, show that to tailor a patient's treatment and to monitor treatment progression, the staining intensity needs to be correlated to the grade of disease; that is, to the morphological and cellular architecture that define cancer and many diseases.

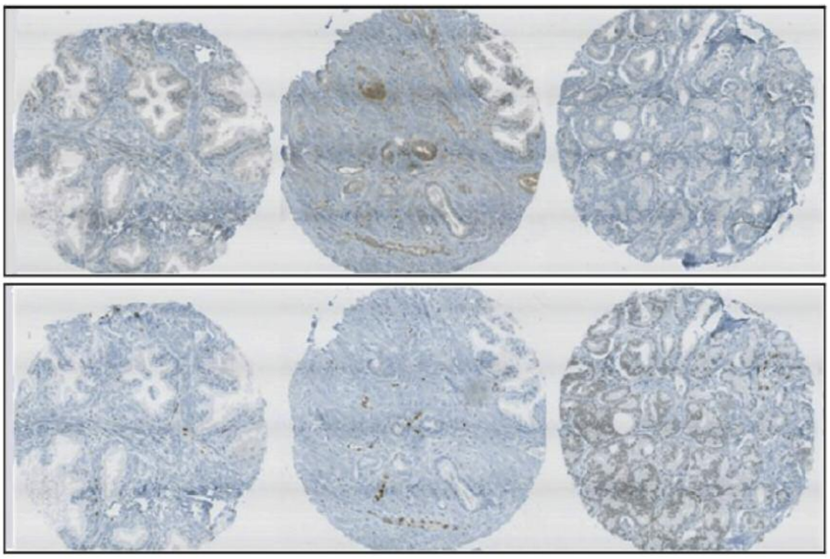


*Figure 1: IHC prostate images indicating the difficulty of sample analysis. Emphasizing the PTEN (top) and ERG (bottom) proteins. Left to right: normal, low, and intermediate grade stages of cancer development. Figure taken from Zerhouni et al [3].*

In the developed framework, we use a pre-trained convolutional network to extract features that capture both morphology and colour at image patches around identified stained cells. We generate a feature dictionary, whose size is data dependent. To represent tissue and tumour heterogeneity, we capture spatial layout information using a commute time matrices approach. We then use the commute time matrix to compute a unique signature per protein and disease grade. The goal of this work is to evaluate whether the expression of the proteins can be used as pathogenesis predicator and quantify the relative importance of each protein at the different disease grades. To this end, we evaluate them in the task of classification individually and in combination. For the individual evaluation, we use a random forest classifier. To evaluate the collective contribution of the proteins, we use a multiple kernel learning (MKL) approach. We tested the proposed framework on a PC tissue dataset and demonstrated the efficacy of the derived protein signatures for both disease stratification and quantification of the relative importance of each protein [3].

The technique already shows promise, and we expect even greater accuracy in our future tuning of these techniques. Owing to the limited number of images in our dataset, training a CNN was not feasible. To this end, we are currently investigating the use of a convolutional auto-encoder for unsupervised feature extraction. Furthermore, we plan to test the framework on larger datasets and more proteins.

**Links:**
http://www.zurich.ibm.com/mcs/systems_biology
http://www.wildlab.ch/

**References:**

[1] Q. Zhong et. al.: "Computational profiling of heterogeneity reveals high concordance between morphology- and proteomics-based methods in prostate cancer", DGP 2015.

[2] A. Rizzardi et al.: "Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring", Diagnostic Pathology 2012, 7:42.

[3] E. Zerhouni et al.: "A computational framework for disease grading using protein signatures", IEEE ISBI 2016 (accepted).

**Please contact:**

Maria Gabrani
IBM Research Zurich
E-mail: mga@zurich.ibm.com

# Mining Electronic Health Records to Validate Knowledge in Pharmacogenomics

by Adrien Coulet and Malika Smaïl-Tabbone

***Most of the state of the art in pharmacogenomics (PGx) is based on a bank of knowledge resulting from sporadic observations, and so is not considered to be statistically valid. The PractiKPharma project is mining data from electronic health record repositories, and composing novel cohorts of patients for confirming (or moderating) pharmacogenomics knowledge on the basis of observations made in clinical practice.***

Pharmacogenomics (PGx) studies how individual gene variations cause variability in drug responses. A state of the art of PGx is available and constitutes a basis for implementing personalized medicine, i.e., medical treatment tailored to each patient taking into account his/her genomic context. However, most of the state of the art in this domain is not yet validated, and consequently not yet applicable to medicine. Most information results from studies that do not fulfil statistics validation standards and are difficult to reproduce because of the rarity of gene variations studied (making it hard to recruit sufficiently large cohorts) and of the multifactorial aspects of drug responses [1]. The increasing use of electronic health records (EHRs) generates large repositories that offer new opportunities, such as composing patient cohorts for the study of clinical hypotheses hard to test experimentally. Typically, EHR repositories make it possible to assemble cohorts of patients to study the impact of gene variations on drug responses on the basis of practice-based data [2].

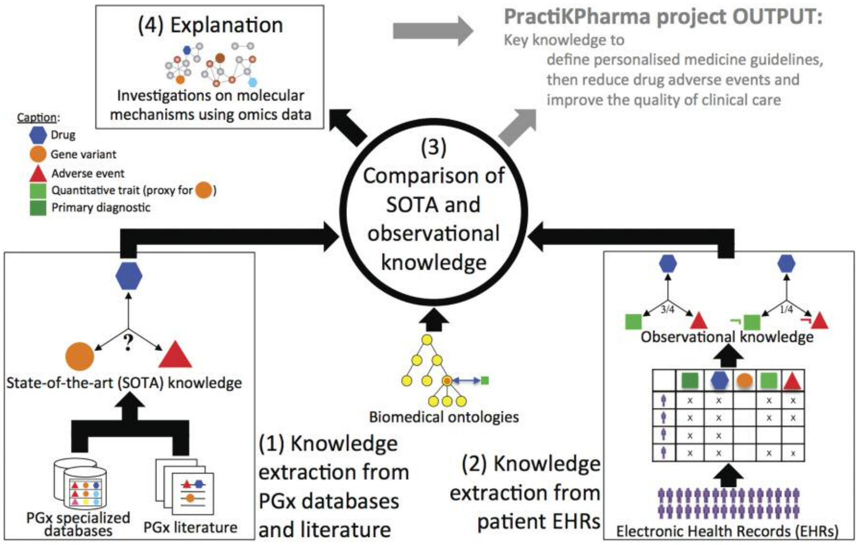

(4) Explanation
Investigations on molecular mechanisms using omics data
PractiKPharma project OUTPUT:
Key knowledge to
define personalised medicine guidelines,
then reduce drug adverse events and
improve the quality of clinical care
Caption:
Drug
Gene variant
Adverse event
Quantitative trait (proxy for )
Primary diagnostic
(3) Comparison of SOTA and observational knowledge
?
State-of-the-art (SOTA) knowledge
PGx specialized databases
PGx literature
(1) Knowledge extraction from PGx databases and literature
Biomedical ontologies
(2) Knowledge extraction from patient EHRs
3/4
1/4
Observational knowledge
Electronic Health Records (EHRs)

*Figure 1: Outline of the four objectives of the ANR PractiKPharma project.*

In October, 2015, The French National Research Agency (ANR) committed funding to a project called PractiKPharma (Practice-based evidence for actioning Knowledge in Pharmacogenomics). The project aims to validate or moderate PGx state-of-the-art (SOTA) knowledge on the basis of practice-based evidence, i.e., knowledge extracted from EHRs. Units of knowledge in PGx typically have the form of ternary relationships gene variant-drug-adverse event, and can be formalized to varying extents using biomedical ontologies. To achieve this goal, the PractiKPharma consortium will focus on four objectives, illustrated in Figure 1: (1) to extract SOTA knowledge from PGx databases and literature; (2) to extract observational knowledge (i.e., knowledge extracted from observational data) from EHRs; (3) to compare knowledge units extracted from these two origins, to confirm or moderate SOTA knowledge, with the goal of enabling personalized medicine; (4) Finally, to emphasize newly confirmed knowledge, omics databases will be investigated for molecular mechanisms that underlie and explain drug adverse events. This investigation will use and contribute to the biomedical Linked Open Data [3].

The PractiKPharma consortium comprises four academic partners: two computer science laboratories, the LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications, Nancy, France) in Nancy, the LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France) in Montpellier; and two University Hospitals, the HEGP (The Georges Pompidou European Hospital, Assistance Publique – Hôpitaux de Paris (AP-HP), France) in Paris, specialized in EHRs management and pharmacogenomics, and the CHU Saint-Etienne (University Hospital (CHU) of Saint Etienne, France), specialized in pharmacovigilance.

PractiKPharma will have impacts both in computer science and biomedicine, including:

- the development of novel methods for knowledge extraction from text and EHRs;
- enabling multilingual semantic annotation of EHRs;
- methods for representing and comparing SOTA and observational knowledge;
- a database that maps genotypes to quantitative traits to facilitate the study of PGx with EHRs;
- the completion and connection of Linked Open Data related to PGx;
- methods for hypothesizing on mechanisms of adverse events and validated PGx knowledge units.

Overall, the final goal of PractiKPharma is to provide clinicians with actionable PGx knowledge to establish guidelines that when implemented in personalized medicine will reduce drug adverse events, and improve the quality of clinical care.

The PractiKPharma consortium will hire two PhD students, two post-docs, one pharmacist and one research engineer to contribute to the project. In addition, PractiKPharma will foster collaboration and dissemination throughout the EU to take into consideration EHRs of various populations and to establish international projects.

**Link:**
PractiKPharma project:
http://practikpharma.loria.fr

**References:**

[1] J. P.A. Ioannidis: "To replicate or not to replicate: The case of pharmacogenetic studies", Circulation: Cardiovascular Genetics, 6:413–8, 2013.

[2] A. Neuraz et al.: "Phenome-wide association studies on a quantitative trait: Application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics, PLoS Computational Biology, 9(12), 2013.

[3] K. Dalleau, N. Coumba Ndiaye, A. Coulet: "Suggesting valid pharmacogenes by mining linked data", in Proc. of the Semantic Web Applications and Tools for the Life Sciences (SWAT4LS) Conference, 2015.

**Please contact:**
Adrien Coulet, LORIA, Nancy, France,
Tel: +33 3 54 95 86 38
E-mail: adrien.coulet@loria.fr
Malika Smaïl-Tabbone, LORIA,
Nancy, France
Tel: + 33 3 83 59 20 65
E-mail: malika.smail@loria.fr

# Modelling the Growth of Blood Vessels in Health and Disease

by Elisabeth G. Rens, Sonja E. M. Boas and Roeland M.H. Merks

***Throughout our lives our blood vessels form new capillaries whose insufficient or excessive growth is a key factor in disease. During wound healing, insufficient growth of capillaries limits the supply of oxygen and nutrients to the new tissue. Tumours often attract capillaries, giving them their own blood supply and a route for further spread over the body. With the help of biological and medical colleagues our team develops mathematical models that recapitulate how cells can construct new blood vessels. These models are helping us to develop new ideas about how to stimulate or stop the growth of new blood vessels.***

In collaboration with the two academic hospitals in Amsterdam our research group at CWI develops mathematical models to help unravel the growth of capillaries, the smallest blood vessels in our body. The building blocks of capillaries are endothelial cells, a cell type that covers the inner walls of all our blood vessels, and the extracellular matrix, a scaffolding of structural protein fibres that keeps cells and tissues together. Together with auxiliary cell types, the endothelial cells remodel the extracellular matrix to form blood vessels, and they secrete new matrix proteins to strengthen themselves. Our collaborators in the academic hospitals mimic angiogenesis, by culturing endothelial cells inside gels of extracellular matrix proteins. In response to growth factors and other signals in the extracellular matrix, the cells aggregate and organize into networks of blood-vessel-like sprouts.

The goal of our mathematical modelling project is to find out the rules that cells must follow in order to form biological structures, such as networks and blood vessel sprouts. Our models describe the behaviour of the individual cells: how they move, the strength of their adhesion to adjacent cells, their shape, and their response to signals from adjacent cells and from the extracellular matrix. If we understand how cells must behave to form a new blood vessel, perhaps then we can also learn how to use pharmaceuticals to fine-tune that behaviour so the cells do what we want them to do. If we could do this for a relatively simple mechanism like blood vessel

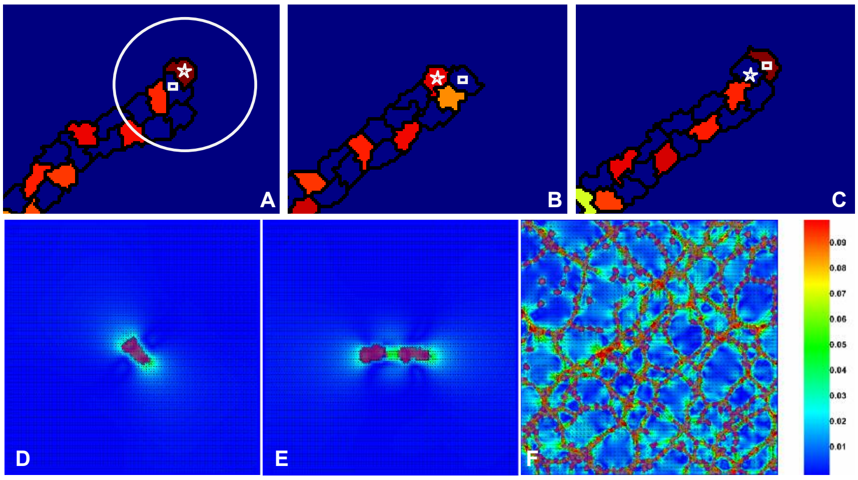


*Figure 1: Models of blood vessel growth. (A-C) Growth of a sprout and "overtaking" of the tip cell. Colour scale indicates state of the embedded genetic network models; red: high expression of tip cell marker Dll4; blue: low Dll4 expression. Figure modified from Ref. [1]. (D-F). Model of mechanical interaction between cells and extracellular matrix. Due to the cell-matrix interactions, the cell elongated (D), they align with adjacent cells (E), and they form networks (F). Figure by EGR, reprinted in modified form from "Merks. ITM Web of Conferences, in press" according to the model described in Ref. [2]. Figures A-F reproduced under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).*

growth, this might give us new ideas for other steps in the development of multicellular organisms.

In an initial set of models, we assumed that cells attract one another via chemical signals. It is not difficult to see that this rule will not suffice for forming networks: It simply makes the cells accumulate into small aggregates. To our surprise, however, the same rules could make cells form networks if the cells had an elongated shape. This makes the aggregates stretch out, and connect to adjacent aggregates such that a network emerges. But the cells did not need to be elongated: if instead the cells became insensitive to the attractive signals at the positions of their membranes where they touched one another, they also organized into networks.

Our recent work employs these two models to gain greater insight into the mechanisms of angiogenesis. Our initial models ignored that in actual angiogenesis, two types of endothelial cells are involved: the tip cells that lead the sprouts, and the stalk cells that form the body of the sprout by ``stalking'' the tip cells. Recent experimental observations have shown that one of the stalk cells regularly replaces the tip cell, much like migratory birds take turns in the leading position in front of the "V". Much experimental work has focused on the molecular regulation of this phenomenon. We discovered that such tip cell overtaking occurs naturally in our models, as a side effect of the cell behaviours responsible for sprouting. We integrated the molecular networks responsible for tip-stalk cell differentiation inside each of the simulated endothelial cells, and could reproduce many of the experimental observations that were taken as evidence for the molecular regulation of tip cell overtaking [1] (Figure 1). Thus our models suggest that the molecular networks may be there to ensure that the cell that happens to end up in front takes on the role of leader.

Despite the insights we acquired from our initial models, they ignored a key component in angiogenesis. Most cells in our body are surrounded by a fibrous or jelly material called the extracellular matrix, or "matrix" for short. It has a structural role – the matrix gives our bones their strength and keeps cells together – but recently it has become clear that it also acts as a medium for cell-cell signalling. Cells deposit growth factors into the matrix, which other cells find later on, as if cells leave "written" notes for one another. Also, cells apply mechanical forces on the matrix, and they respond to forces they feel in their surroundings. In this way the matrix can relay mechanical signals from cell to cell. Recently we have started to explore how such mechanical cell-cell signalling can help coordinate the behaviour of endothelial cells. We extended our models of angiogenesis with a finite-element model of the extracellular matrix [2]. The endothelial cells apply contractile forces on the matrix. The finite-element model then calculates the strains resulting from these forces, which, in turn, affects the membrane movements of the cells. These relatively few assumptions reproduced observations on how matrix stiffness determines the shape of single cells, how adjacent cells interact with one another, and how cells form vascular-like sprouts and networks (Figure 1).

Now that our models have given us some first clues on how cells build blood vessels, we are taking these insights back to our experimental collaborators. The basic understanding of angiogenesis captured by our models is employed in more specific models of diseases, e.g., diabetic retinopathy. The models give the experimentalists new ideas on how to "nudge" the endothelial cells to do, or not do, what they want. Of course, biology still has many surprises in store for us; rarely do experiments agree with our models in every detail. Frustrating as this may sound, this is actually when our models become most useful. Mismatches between model and experiment highlight the areas in vascular biology that are not yet fully understood.

**Links:**
Group website:
http://biomodel.project.cwi.nl

**References:**

[1] S. E. M. Boas and R. M. H. Merks: "Tip cell overtaking occurs as a side effect of sprouting in computational models of angiogenesis," BMC Systems Biology 9:86, 2015

[2 R. F. M. van Oers, E. G. Rens, D. J. LaValley, et al.: "Mechanical Cell-Matrix Feedback Explains Pairwise and Collective Endothelial Cell Behavior In Vitro", PLoS Computational Biol., 10(8): e1003774, 2014.

**Please contact:**

Roeland M.H. Merks
CWI, The Netherlands
E-mail: merks@cwi.nl

# Modelling? Using Standards Can Help You

by Brett G. Olivier and Bas Teusink

*Almost everything one does relies on the use of standards, from vehicle safety, to smartphone design. Here we use an idealised scenario to illustrate these standards for modelling.*

Standards play an important role in our everyday life; they define how our appliances communicate with one another and the address format of our last email. While ubiquitous in experimental sciences, the application of standards to mathematical modelling in the life sciences is relatively new. With its integral use of computer modelling many of these standardisation efforts have taken root in the relatively new field of systems biology.

With the field now moving to big data and whole-cell modelling, standardisation becomes even more important. The “whole cell model” will consist of many sub-models that are in themselves data that need to be exchanged between researchers and groups – in ways not originally envisioned by their original authors. Beyond model exchange, the use of standards enables the development and usage of interoperable simulation software. This allows ensembles of software with a diverse range of functionality to be combined together. To illustrate this, let us consider an “ideal” day in the life of a computational biologist asked to investigate yeast glycolysis in a “growing cell”.

Deciding that this model will combine a kinetic model of glycolysis with a genome-scale metabolic model, he first searches the Biomodels database and finds Teusink’s detailed kinetic model of glycolysis and downloads it in the Systems Biology Markup Language (SBML) Level 2 format.

SBML is a widely used model-encoding standard; using XML it aims to describe biological processes in a tool independent way. In addition to describing the model components (e.g., metabolites, reactions, kinetic laws, parameters and compartments) it also provides a framework for their detailed annotation.

Loading the model into the SBML compliant COPASI software allows him to view the model and its annotation and immediately start investigating its behaviour. Next he decides to use the consensus yeast metabolic network which he downloads in the SBML Level 3 format.

SBML Level 3 has a modular design and while the language core is analogous to SBML Level 2, its functionality can be extended through the use of packages. In this case the model makes use of the Flux Balance Constraints (FBC) package which adds elements used in constraint-based modelling such as flux bounds and objective functions [1]. In addition packages exist for hierarchical model composition as well as the definition of qualitative and spatial models.

To get a feel for his newly acquired model, he loads the model into the online CBM modelling system FAME and starts modelling. Deciding to combine both these models in a single analysis he loads them into two different SBML compatible, Python based tools. Using CBMPy for the genome-scale model and PySCeS for the kinetic model allows him to develop a new algorithm that combines both models and tools in a single analysis. Having succeeded, all that is left is to back up his work and send the analysis and result to his colleagues. For this he makes use of a COMBINE archive.

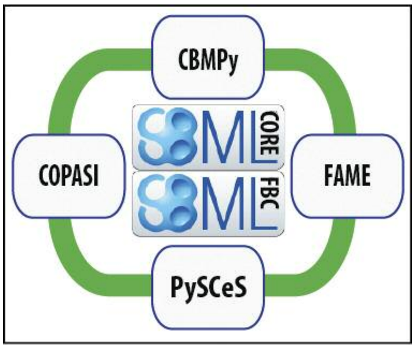


*Figure 1: The use of open standards for encoding models enables the integration of multiple models and tools.*

The COMBINE archive is a single file in the OMEX format [2]. It can contain multiple files, such as the two source models as well as the Python scripts used in the analysis. In addition it can include metadata about these files, the analysis itself as well as the software used to generate the results. The COMBINE archive is therefore an ideal way of storing and exchanging entire modelling experiments.

Unfortunately, this is not yet a typical day for the majority of modellers in systems biology. The challenge is to expand and implement support for open standards in as many tools as possible thereby enabling the development of a new in silico biology.

**Links:**

Modelling Software: http://teusinklab.nl/modelling-methods-and-tools
COmputational Modeling in BIology NEtwork (COMBINE): http://co.mbine.org
Biomodels database: http://biomodels.net
SBML: http://sbml.org
COPASI: http://copasi.org
Yeast metabolic network http://yeast.sourceforge.net
FAME: http://f-a-m-e.org/
CBMPy: http://cbmpy.sourceforge.net
PySCeS: http://pysces.sourceforge.net

**References:**

[1] B.G. Olivier and F.T. Bergmann: “The Systems Biology Markup Language (SBML) Level 3 Package: Flux Balance Constraints”, Journal of integrative bioinformatics, 12, 269, 2015.

[2] F. T. Bergmann et al.: “COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project”, BMC Bioinformatics, 15, 369, 2014.

**Please contact:**

Brett G. Olivier
Systems Bioinformatics
Vrije Universiteit Amsterdam,
The Netherlands
E-mail: b.g.olivier@vu.nl

# Understanding Metadata to Exploit Life Sciences Open Data Datasets

by Paulo Carvalho, Patrik Hitzelberger and Gilles Venturini

***Open data (OD) contributes to the spread and publication of life sciences data on the Web. Searching and filtering OD datasets, however, can be challenging since the metadata that accompany the datasets are often incomplete or even non-existent. Even when metadata are present and complete, interpretation can be complicated owing to the quantity, variety and languages used. We present a visual solution to help users understand existing metadata in order to exploit and reuse OD datasets – in particular, OD life sciences datasets.***

The Environmental Research and Innovation Department (ERIN) of the Luxembourg Institute of Science and Technology (LIST) conducts research and activities in environmental science (including biology, ecology and other areas of study) using advanced tools for big data analytics and visualization. ERIN's e-science unit is currently investigating how open data (OD) datasets, containing data pertaining to environmental science and related areas, may be reused. The quality of metadata that accompanies such datasets is often poor. Despite metadata being essential for OD reuse [1], it is often non-existent or of low quality. Sometimes metadata is defined, but not always. When it is defined, it may be incomplete. Furthermore, there is no single common standard used to specify metadata: each data provider can choose how metadata will be represented, and may or may not follow a specific metadata standard, for example: Dublin Core Metadata Initiative (DCMI), Data Catalogue Vocabulary (DCAT), Metadata Objects Description Schema (MODS). This lack of consistency increases the difficulty of exploiting metadata, especially when several data sources – potentially from different countries with metadata represented in different languages – are implicated.

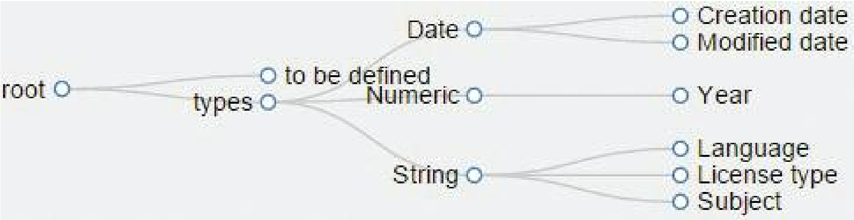


*Figure 1: Initial state of the metadata-mapper.*

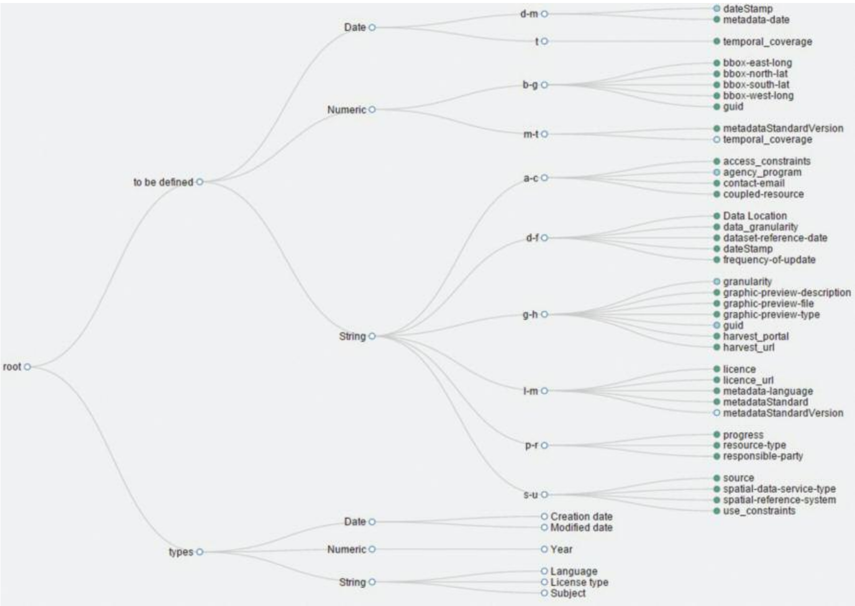


*Figure 2: The metadata of 543 datasets obtained from http://data.gov.au (search term: "biology").*

We propose a visual solution to avoid such problems and to help users understand metadata and to search for and find specific datasets. Why a visual solution? Data visualization can rapidly assimilate and recognize large amounts of information [2]. The datasets and metadata information are stored in a database previously filled by an OD dataset downloader script. We assume that this task has been successfully done: it is not the focus of our work. The metadata-mapper offers an overview of every piece of metadata obtained from the datasets. The way the information is displayed is organised in three levels:

- The first level organises the metadata according to whether or not the metadata are assigned to a given research criteria. Research criteria are necessary to search datasets, e.g., to enable the user to search a dataset by theme, by date of publication, by language, etc.
- The second level refers to the type of metadata value. Every metadata is formed by key-value pairs. Different types of metadata can exist. The value might be a number, a name, a date, an email-address, etc.
- Finally, to improve visualization and understanding of all metadata keys, the metadata are organised by name. Several groups exist: they contain the metadata whose key starts with a given letter belonging to the group.

In addition to providing a better understanding of existing metadata, our metadata-mapper solution permits the user to establish links between metadata and search criteria. This means that a specific metadata is used to apply a search based on a precise criterion, for instance: imagine we have a search criteria “Language”. The user can create a link between this criteria and metadata “Language” and “Country”. The solution will convert this relation into a query similar to: “SELECT * FROM Datasets WHERE Language = ? UNION SELECT * FROM Datasets WHERE Country = ?”.

Figure 2 demonstrates the result of the chart showing all the metadata related to the datasets found and downloaded from the http://data.gov.au OD portal after searching the term “biology”. Five hundred and forty three different datasets were found and downloaded. 17,435 metadata sets were obtained along with the datasets. These 17,435 metadata sets contain 34 different keys.

If it is impossible to understand and use the metadata associated with individual datasets, the datasets cannot be used, and are effectively useless. It is vital that we are able to unlock the potential and value of OD datasets. The metadata-mapper solution delivers a support tool in order to understand metadata delivered with datasets but also to link metadata with research criteria. This is only a first step towards making metadata more usable and to permit the reuse of OD datasets. This solution will give the user a global picture of all existing metadata, their meaning, and how they can be used to find datasets. The visual approach enables large amounts of metadata to be shown at one time. However, the solution has to be tested with large numbers of datasets in order to test whether the methodology can deal with massive amounts of metadata or whether further modifications need to be made.

**References:**

[1] N. Houssos, B. Jörg, B. Matthews: “A multi-level metadata approach for a Public Sector Information data infrastructure”, In Proc. of the 11th International Conference on Current Research Information Systems, pp. 19-31, 2012.

[2] S. Card, J. D. Mackinlay, B. Shneiderman: “Information visualization. Human-computer interaction: design issues, solutions, and applications, 181, 2009.

**Please contact:**

Paulo Carvalho
Luxembourg Institute of Science and Technology
E-mail: paulo.carvalho@list.lu

# Management of Big and Open Data in the Life Cycle Assessment of Ecosystem Services

by Benedetto Rugani, Paulo Carvalho and Benoit Othoniel

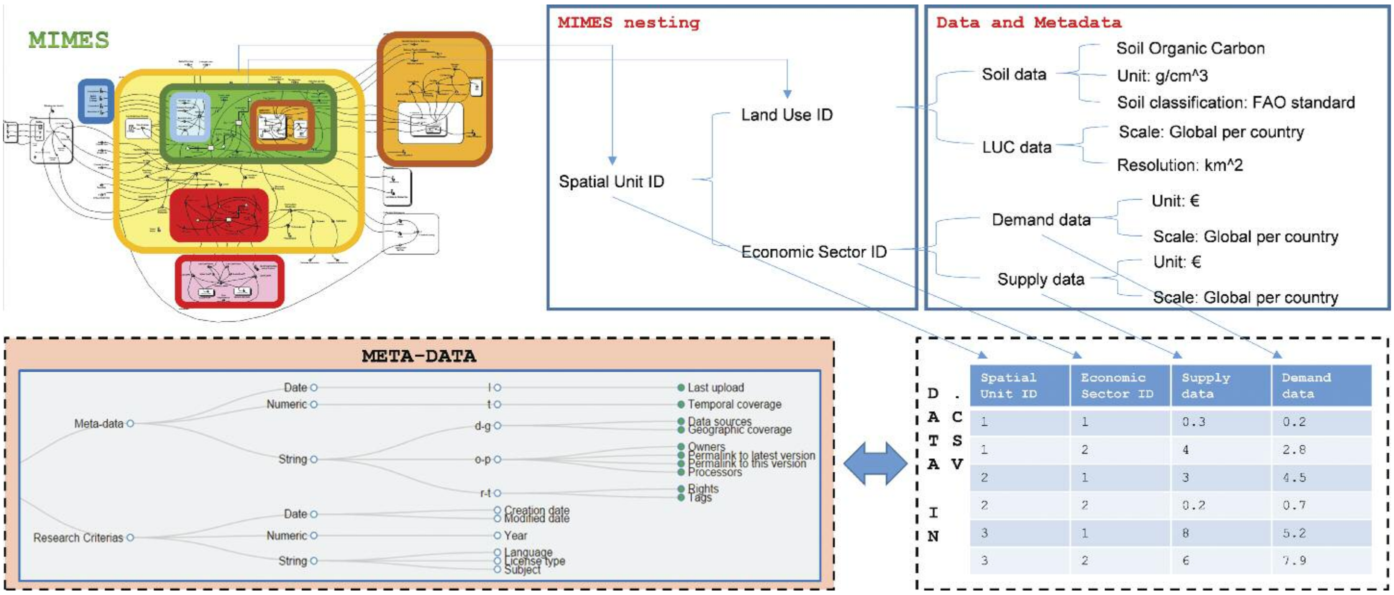
***When defined, metadata information that accompanies Big and Open Data (OD) datasets may be hard to understand and exploit. A visual approach can support metadata re-use in integrated ecological-economic modelling. A method that allows specific model datasets to be regularly and consistently updated may improve their readability for use in the Life Cycle Assessment (LCA) modelling of ecosystem services.***

“Ecosystem services” is a term used to describe the benefits humans get from the Earth’s natural capital (e.g., forest biomass, water purification, soil quality, etc.). These flows reflect the interaction among all capitals of biosphere and anthroposphere. Complex interdisciplinary modelling is necessary to study these interactions and to assess ecosystem service values, thus a big data (BD) collection and elaboration effort is often required. BD have also been put to good use in the Life Cycle Assessment (LCA) framework [1], which is one of the most robust and commonly applied methodologies to evaluate the impact of human activities on ecosystems. However, there is no consensus on how to evaluate ecosystem services in LCA. This is due, in part, to the lack of consensus on the standardization and use of data and models, which can be broadly retrieved from OD sources.

The aim of the VALUES project (see at: http://www.list.lu/en/project/values/) is to develop a novel approach to simulate the value of biodiversity and ecosystem services within the LCA methodology. This work is being conducted by researchers within the RDI Unit on Life Cycle Sustainability and Risk Assessment (LiSRA) at the Environmental Research & Innovation (ERIN) Department of the Luxembourg Institute of Science and Technology (LIST). The team is adapting MIMES (Multi-scale Integrated Model of Ecosystem Services) [2] to the LCA framework. The main challenge is to model the nexus among multiple elements of the biosphere and the anthroposphere at different geographical and temporal scales, generating a holistic integrated representation of the local (e.g., country scale of Luxembourg), multi-regional (e.g., Europe) and global relationships between economic and ecological systems.

One of the major tasks of VALUES is to solve the problem of collecting, elaborating, managing and running a large amount of data (mostly OD), such as georeferenced information, and ecological/biophysical and process life cycle data. These need to be stored in the model to describe all the relationships between modules, sub-models and compartments of MIMES, e.g., to describe the interactions between the emission of pollutants and wastes from energy production or transport systems and their effects on the receiving ecological and human compartments, i.e., air, soil, water, and the human body.

The global version of MIMES already contains thousands of variables and process modelling parameters. However, improving the existing MIMES relationships and model transparency requires the incorporation of additional “heavy” datasets and metadata layers, respectively. As a consequence, the number of model reference variables is expected to increase in the range of 40-70% by the end of



| Spatial Unit ID | Economic Sector ID | Supply data | Demand data |
|---|---|---|---|
| 1 | 1 | 0.3 | 0.2 |
| 1 | 2 | 4 | 2.8 |
| 2 | 1 | 3 | 4.5 |
| 2 | 2 | 0.2 | 0.7 |
| 3 | 1 | 8 | 5.2 |
| 3 | 2 | 6 | 7.9 |

*Figure 1: Simplified approach for data management in the VALUES project.*

VALUES in 2017. On one hand, a certain level of model aggregation needs to be maintained or achieved to avoid computational issues. On the other hand, a large amount of datasets and database sources needs to be harmonized in one unique modelling framework.

As shown in Figure 1, the complexity of sub-models in MIMES can easily be increased when amplifying the resolution of assessment or the detail of process connections. Moreover, a large amount of metadata information must be stored in the modular system of compartmental relationships across variables, and then regularly upgraded with information collected from the same or other BD-like sources (e.g., statistical datasets, climate measurements, biological engineering databases). Most of the metadata information is then screened, modified and linked to upstream .CSV files containing the original input datasets necessary to run the model.

However, metadata is often missing or difficult to understand, posing additional problems related to the transparency of the modelling approach. These complications limit the usability of the data, and the reproducibility of the models that need constant updates. When defined, metadata information may be set up in different languages, under various data types and using different keys to represent the same meaning (e.g., bottom-left box in Figure 1 represents the metadata existing for the EU dataset on the association of species and habitat types to ecosystems: http://www.eea.europa.eu/data-and-maps/data/linkages-of-species-and-habitat/). The LiSRA and e-Science RDI Units of ERIN/LIST are currently working to overcome such data modelling obstacles [3].

By considering MIMES as test-bed case to develop a methodology of OD management and metadata visualization, VALUES can be optimized to guarantee the success of modelling outputs and increased system transparency. Furthermore, the use of OD collection and visualization techniques in VALUES can support future expansions and cross-fertilizations of such IT approaches to access BD within the broader LCA framework. This project represents a tangible example that may pave the way towards more effective and transparent management of metadata in complex life cycle process unit databases, making a valuable contribution to current state-of-the-art practice and tools for LCA [1].

**Links:**
http://www.list.lu/en/project/values/
http://sourceforge.net/projects/mimes/
http://www.simulistics.com/

**References:**

[1] J. Cooper, et al.: "Big Data in Life Cycle Assessment", Journal of Industrial Ecology, vol. 17, pp. 796-799, 2013.
[2] R. Boumans et al.: "The Multiscale Integrated Model of Ecosystem Services (MIMES): Simulating the interactions of coupled human and natural systems", Ecosystem Services, vol. 12, pp. 30-41, 2015.
[3] P. Carvalho, P. Hitzelberger, and G. Venturini: "Understanding Open Data CSV File Structures for Reuse", in ERCIM News 100 (January 2015), ed ERCIM EEIG, 2015, p. 36.

**Please contact:**

Benedetto Rugani
Luxembourg Institute of Science and Technology (LIST)
Tel: +352 2758885039
E-mail: benedetto.rugani@list.lu

# WITDOM: Empowering Privacy and Security in Non-trusted Environments

by Juan Ramón Troncoso-Pastoriza and Elsa Prieto Pérez

***The WITDOM project (empoWering prIvacy and securiTy in non-trusteD environments) develops innovative technical solutions for secure and privacy-preserving processing of genomic and financial data in untrusted environments.***

The advent of outsourced and distributed processing environments like the Cloud is prompting fundamental transformations in whole ICT ecosystems, while bringing new opportunities to stakeholders in terms of the availability and rational use of physical resources with large-scale savings in IT investments. Conversely, it also poses new security challenges, especially for ensuring robust protection of privacy and integrity of personal information, which are essential for societal acceptance of new ICT schemes, services and solutions.

In this context, the WITDOM project focuses on developing innovative solutions for truly efficient and practical privacy enhancing techniques and efficient signal and data processing in the encrypted domain for outsourced environments. WITDOM's main goal is to produce a framework for end-to-end protection of data in untrusted environments, with a particular focus on data-outsourcing scenarios where new threats, vulnerabilities and risks due to new uses require end-to-end security solutions that will withstand progress for the lifetime of applications they support.

WITDOM aims to develop three main innovations in this area. Firstly, a novel framework for end-to-end security and privacy that will guarantee efficient and verifiable provision of privacy in the context of ICT services owned by third-party providers of distributed processing and storage. This framework will maximize independence from stated security and privacy commitments by respective providers, and minimize the current need for blind trust from clients, based solely on written consents. The initial contribution of this framework builds upon a requirements elicitation methodology, synergizing the results of the recently ended FP7 project PRIPARE with a co-design process to fuse privacy into the first stages of the systems design process, materializing a true Privacy-by-Design methodology.

The second dimension of innovations deals with the actual design and development of tools and technologies for efficient privacy protection of data outsourced and processed in untrusted environments. These techniques can be categorized according to the four main research areas addressed in WITDOM: efficient lattice cryptosystems for homomorphic processing, allowing for faster and more resource-efficient encrypted-domain processing [1]; accurate and effective secure signal processing to cope with the marriage between cryptography and the ubiquitous signal processing operations when dealing with sensitive signals [2]; efficient privacy-enhancing technologies for obfuscation, noise addition, anonymization and data masking, measuring and achieving quantitative guarantees that the processed data will be unidentifiable and the produced results unlinkable; and last, but not least, efficient and scalable integrity and consistency verification techniques to preserve fork-linearizability on data accessed and modified by several users on outsourced data stores [3]. The main strength in our approach is that these innovations are not applied independently or autonomously, but in an end-to-end secure and private architecture that defines a platform that can deploy privacy-preserving services on outsourced data with quantifiable and assessable technological guarantees. Promising initial results have already been produced within WITDOM's roadmap, which will continue to advance the state of the art in these fields for the next two years.

Finally, the third dimension of WITDOM's innovations deals with the instantiation of the developed framework, platform and tools in two carefully chosen use-case scenarios, for which privacy is essential owing to the sensitivity of the involved data, and where privacy and confidentiality constraints are a true barrier to using outsourced architectures and Cloud-based deployments. The first use-case is a health scenario that involves outsourcing genetic data processes and workflows for large research analyses and individual clinical analyses; genetic data is extremely sensitive, and genomic privacy has become a hot topic

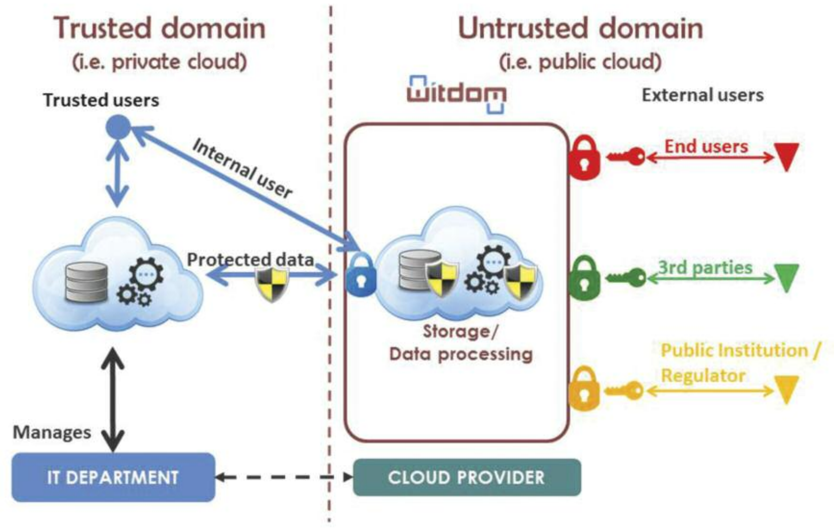


*Figure 1: WITDOM generic architecture.*

for research and innovation, to which WITDOM contributes by focusing on solutions for outsourced processing of genetic data. The second scenario deals with outsourced financial analyses based on the management of both customers' data and finance data, to enable credit risk calculations and card transaction fraud detection deployed as outsourced secure financial services over private and public Cloud environments.

Research and innovation in this field cannot ignore the impact of data protection regulations and directives on the evolution of Cloud-related environments and in the processing of personal and sensitive data. Therefore, WITDOM innovations are built upon a legal assessment and validation of the evolving European Data Protection Regulation, linking legal and ethical requirements with technological means to guarantee their enforcement.

The WITDOM project is a Research and Innovation Action funded by the European Commission's Horizon 2020 programme and the Swiss State Secretariat for Education, Research and Innovation. This project, started in January 2015, joins together a multidisciplinary consortium comprising Universities, research centres, strong industrial stakeholders, and end-users: Atos(Spain) as the project coordinator, the University of Vigo (Spain) as the technical coordinator, Katholieke Universiteit of Leuven (Belgium), IBM Research (Switzerland), Fondazione Centro San Raffaele (Italy), BBVA (Spain), Gradiant (Spain), and the Ospedale San Raffaele (Italy).

**Links:**
http://witdom.eu/
https://twitter.com/W1TD0M
https://www.linkedin.com/grp/home?gid=8257514

**References:**

[1] D.D. Chen et al.: "High-Speed Polynomial Multiplication Architecture for Ring-LWE and SHE Cryptosystems", IEEE Trans. on Circuits and Systems 62-I(1): 157-166 (2015).

[2] J. R. Troncoso-Pastoriza and F. Pérez-González: "Secure Signal Processing in the Cloud: enabling technologies for privacy-preserving multimedia cloud processing", IEEE Signal Process. Mag. 30(2): 29-41 (2013).

[3] M. Brandenburger, C. Cachin and N. Knežević: "Don't trust the cloud, verify: integrity and consistency for cloud object stores", ACM SYSTOR 2015: 16:1-16:11.

**Please contact:**

Juan Ramón Troncoso-Pastoriza
University of Vigo
E-mail: troncoso@gts.uvigo.es

Call for Participation

## ECCB 2016 – 15th European Conference on Computational Biology

The Hague, Netherlands,
3-7 September 2016

ECCB 2016 will welcome scientists working in a variety of disciplines, including bioinformatics, computational biology, biology, medicine, and systems biology. Participating in ECCB 2016 will be the perfect opportunity to keep pace with cutting edge research, and to network with members of ECCB community.

**More information:**
http://www.eccb2016.org/

## ECMTB / www.ecmtb2016.org

**10th European Conference on Mathematical and Theoretical Biology and Annual Meeting of the Society for Mathematical Biology**

Nottingham, 11-15 July 2016

Society for Mathematical Biology

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

ESMTB

# European Research and Innovation

## Secure and Privacy-Preserving Identity Management in the Cloud

by Bernd Zwattendorfer, Stephan Krenn and Thomas Lorünser

***CREDENTIAL is an EU H2020 funded research project that is developing, testing, and showcasing innovative cloud-based services for storing, managing, and sharing digital identity information and other highly critical personal data with a demonstrably higher level of security and privacy than other current solutions. This is achieved by advancing novel cryptographic technologies and improving strong authentication mechanisms.***

Digital identity management (IdM) is an essential tool for managing access to information technology (IT) resources and is an enabler for seamless interaction between systems, organizations, and end users in the future. However, in order to be fully and broadly accepted, IdM must involve secure identification and authentication processes and protect user privacy. This is especially true for high-assurance application domains such as e-Business, e-Government, or e-Health, which have a superior demand for security and privacy due to the harm a potential breach or identity theft could cause.

Identity management is currently experiencing a paradigm shift, and existing solutions fall short in many aspects when it comes to federated and heterogeneous environments. In the past, IdM was mainly a local issue and most organizations operated their own, custom-tailored identity management systems within the organization's domain boundaries. The use of external IdM systems was the exceptional case. Today, we often see mixed systems, mainly because of the increasing use of distributed and inter-connected applications integrating internal and external components, e.g., as in hybrid cloud applications. This situation leads to fragmented, non-standard authentication situations on the IdM level and causes high administrative costs compared with integrated solutions. Many "identity islands" have to be managed outside the corporate IT environment, and the advantages of integrated identity and access management (IAM) solutions are lost. Important features like single sign-on as well as easy and centralized provisioning/de-provisioning, audit, and control of identities are not possible anymore. For these reasons there exists a strong demand for the development and integration of trustworthy IAM systems. Ideally these systems would provide the necessary security and privacy guarantees aspired in federated business environments with the strongest guarantees possible, by cryptography.

The transformation in the identity management world goes hand in hand with the tremendous shift to cloud computing that has shaped the ICT world during recent years. By now, numerous IdM systems and solutions are available as cloud services, providing identity services to applications operated both in closed domains and in the public cloud. This service model is often referred to as Identity (and Access)

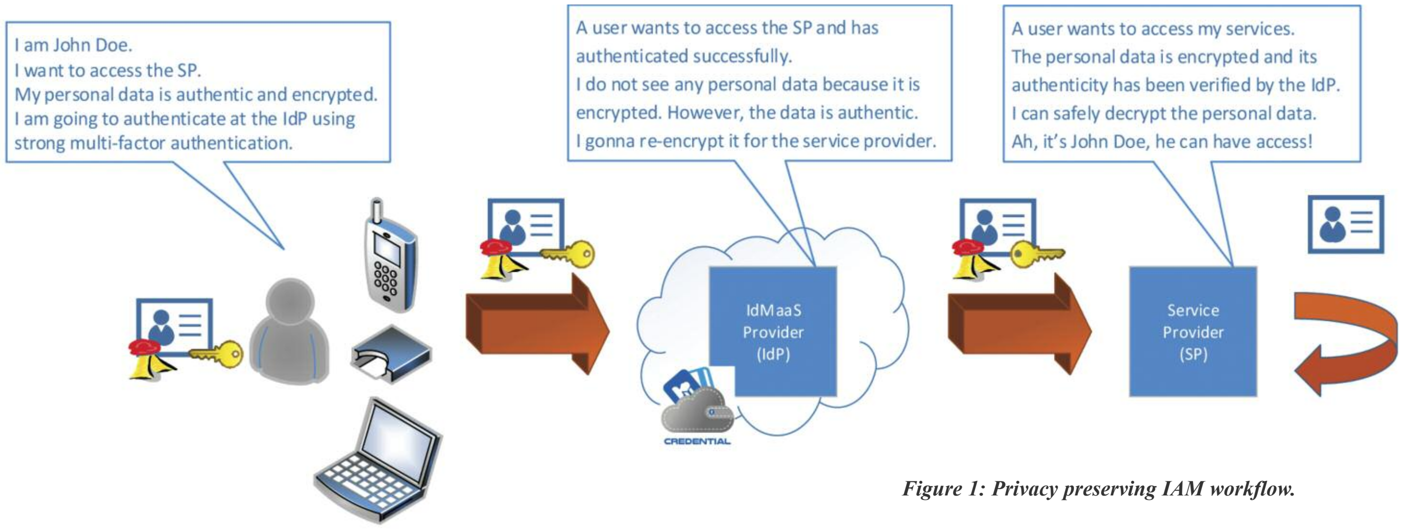

I am John Doe.
I want to access the SP.
My personal data is authentic and encrypted.
I am going to authenticate at the IdP using strong multi-factor authentication.
A user wants to access the SP and has authenticated successfully.
I do not see any personal data because it is encrypted. However, the data is authentic.
I gonna re-encrypt it for the service provider.
A user wants to access my services.
The personal data is encrypted and its authenticity has been verified by the IdP.
I can safely decrypt the personal data.
Ah, it's John Doe, he can have access!
IdMaaS Provider (IdP)
CREDENTIAL
Service Provider (SP)

*Figure 1: Privacy preserving IAM workflow.*

Management as a Service (IDMaaS). Popular examples for cloud IDMaaS providers are big companies from the sectors of social networks (Facebook, LinkedIn), search engines (Google), business solutions (Microsoft, Salesforce), and online retailers (Amazon). However, no satisfactory approaches currently exist which allow the storage and sharing of identity data by service providers in a privacy preserving manner – meaning without the identity provider learning the credentials and associated data.

The vision of CREDENTIAL is to fill this gap and develop a more trustworthy solution by combining secure and efficient identity management technologies with cryptography for cloud computing [1,2,3]. Users will be able to store identity data in a cloud-based IDMaaS system of an identity provider such that the confidentiality and authenticity of the data is upheld even from the provider. Now, if a user wants to access a specific service at a different provider or from the enterprise environment, she can advise the identity provider to select specific data items and re-encrypt them for the service provider such that (after transmission) the service provider remains the only party capable of accessing the data items in plain text.

In comparison to current IDMaaS systems, which have full access to the identity data they are hosting, the CREDENTIAL solution will significantly improve the privacy of cloud identity service users, whilst maintaining a high degree of usability in order to motivate secure handling of services. Data will be protected with end-to-end encryption, while the authentication of the users against the identity service provider will be secured with efficient and strong state-of-the-art multifactor authentication mechanisms within a consistent and holistic security approach.

CREDENTIAL had its kick-off in October 2015 and will last for three years. The consortium is currently doing a technology assessment and a requirements elicitation for a secure and privacy-preserving IDMaaS solution in the cloud. The CREDENTIAL consortium consists of a balanced team from seven EU countries, including six industry partners, two applied research organizations, three universities, and one SME.

**Links:**
Website: https://credential.eu/
Twitter: @CredentialH2020
http://twitter.com/CredentialH2020
LinkedIn: https://linkedin.com/in/credential
CORDIS:
https://cordis.europa.eu/project/rcn/194869_en.html

**References:**

[1] B. Zwattendorfer, D. Slamanig: "Design Strategies for a Privacy-Friendly Austrian eID System in the Public Cloud", Computers & Security 2014.

[2] D. Slamanig, K. Stranacher, B. Zwattendorfer: "User-Centric Identity as a Service-Architecture for eIDs with Selective Attribute Disclosure", in Proc. of SACMAT 2014.

[3] B. Zwattendorfer, A. Tauber: "Secure Cloud Authentication using eIDs", in Proc. of IEEE CCIS 2012.

**Please contact:**

Bernd Zwattendorfer
Graz University of Technology
Tel: +43 (0) 316 8735574
E-Mail: bernd.zwattendorfer@iaik.tugraz.at

Stephan Krenn
AIT Austrian Institute of Technology GmbH
Tel: +43 (0) 664 88256006
E-mail: Stephan.Krenn@ait.ac.at

# PRISMACLOUD – Privacy and Security Maintaining Services in the Cloud

by Daniel Slamanig, Agi Karyda and Thomas Lorünser

***The EU Horizon 2020 PRISMACLOUD research project is dedicated to enabling secure and trustworthy cloud-based services by improving and adopting novel tools from cryptographic research.***

Cloud computing is seen as a major growth area in ICT, with a 2013 forecast from International Data Corporation predicting that worldwide spending on public cloud services will exceed USD $107 billion in 2017. Despite these predictions, the practical adoption of cloud computing technologies may be greatly hindered by inadequate security and privacy protection. Some fundamental properties of (public) cloud computing, such as being an open platform, its anytime and anywhere accessibility, as well as the intrinsic multi-tenancy, introduce new security threats, leading to tremendous risk for personal and sensitive data. Studies from the EU Agency for Network and Information security (ENISA) show that security and privacy concerns represent major stumbling blocks for cloud adoption within European industry.

The need to protect the security and privacy of the data in the cloud is therefore a critical issue. The strongest level of protection can be achieved through cryptography. However, although many recent advances in cryptography have yielded promising tools for cloud computing, far more work is required to transform theoretical techniques into practical solutions for the cloud. PRISMACLOUD is contributing to this transformation by developing tools that allow the next generation of cryptographically secured cloud services to be built with security and privacy incorporated by design and from end to end.

PRISMACLOUD's research, and resulting developments [1], are based on the following objectives. On the one hand we focus on confidentiality of data, which is considered absolutely essential when outsourcing data into the cloud. In particular, we target the development of secure distributed cloud storage systems (i.e., the cloud-of-cloud approach [2]) as well as encryption and tokenization solutions for legacy applications already running in the cloud. Secondly, we are putting significant effort in verifiability features for the cloud. Thereby, we focus on cryptographic means (such as verifiable computing and malleable signatures [3]) to protect the integrity and authenticity of dynamic data in cloud-based workflows and computational tasks. Moreover, we are also focusing on cryptographic means (e.g., graph signatures) that allow auditors to attest or certify the cloud infrastructure and thus help providers to increase the transparency for customers without revealing internal information about the configurations. Thirdly, the privacy of users interacting with a cloud environment requires adequate protection. To protect user privacy, we apply privacy enhancing technologies (e.g., attribute-based anonymous credentials) to implement data minimization strategies and access privacy. In addition, we are interested in efficient data anonymization algorithms to anonymize large datasets to facilitate privacy-friendly data sharing and third party use.

To assure the practical relevance of the developments within PRISMACLOUD, the aforementioned efforts are accompanied by non-cryptographic research topics considered essential for the commercial success of the project results. We will provide secure and efficient software implementations of core technologies and showcase them in selected testbeds. Three different use-cases from different application domains will be used to demonstrate and evaluate the potential of the project outcome, i.e., demonstrate a measurable increase in service level security and privacy. Furthermore, novel human-computer interaction (HCI) guidelines, including HCI design patterns for usable cryptography and protocols for the cloud, will help to design services that respect the users' needs and will therefore maximize acceptance of the technology. In order to use the developed methods properly in novel application scenarios after the project, a holistic security framework and accompanying usage patterns will be prepared in support of service developers. Finally, a vital goal of the project is for the results to be incorporated into standards related to cloud security, and we will actively participate in various standardization bodies in the second phase of the project.

The PRISMACLOUD project has been running since February 2015 and is a 42 month project that receives funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 644962. The project is coordinated by AIT and its consortium consists of 16 partners from academia and industry from nine different countries.

**Links:**
Website: https://prismacloud.eu
LinkedIn: https://linkedin.com/in/prismacloud
Twitter: @prismacloud, http://twitter.com/prismacloud
CORDIS: http://cordis.europa.eu/project/rcn/194266_en.html

**References:**

[1] T. Lorünser et al.: "Towards a New Paradigm for Privacy and Security in Cloud Services", Cyber Security and Privacy, Vol 530 of CCIS, Springer, 2015.
[2] T. Lorünser, A. Happe, D. Slamanig: "ARCHISTAR: Towards Secure and Robust Cloud Based Data Sharing", CloudCom 2015, IEEE, 2015.
[3] D. Derler et al.: "A General Framework for Redactable Signatures and New Constructions", 18th International Conference on Information Security and Cryptology, LNCS, Springer, 2015.

**Please contact:**

Thomas Lorünser, AIT Austrian Institute of Technology GmbH
Tel: +43 664 8157857, E-mail: Thomas.Loruenser@ait.ac.at

# TREDISEC: Towards Realizing a Truly Secure and Trustworthy Cloud

by Beatriz Gallego-Nicasio Crespo, Melek Önen and Ghassan Karame

***The main goal of the TREDISEC project is to increase trust in cloud technology by providing solutions that enhance data security and provide strong privacy guarantees, without hampering the efficiency-and-reduced-cost attractiveness of existing cloud systems.***

The cloud is the go-to technology for companies and businesses these days. Cloud adoption is growing exponentially each year, fueled by new functionalities and capabilities. Although it has many advantages, the cloud also has drawbacks and security-related issues that can make customers shy away from adopting it. Large-scale adoption of the cloud by enterprises and SMEs is hampered by serious concerns about the security and availability of data stored in the cloud. These concerns have been further exacerbated by recent data leak scandals (e.g., corporate data stolen [1], resulting in millions of dollars in losses in just one day) and privacy intrusions (e.g., users' private information illegally accessed and used for blackmail [2]). Moreover, companies are no longer happy to rely on standard cloud security solutions, instead demanding full control over the security and integrity mechanisms employed to protect their data across its lifecycle. As this data is typically a very valuable asset, security and privacy emerge as key features in cloud offerings. TREDISEC addresses these issues by providing a set of security primitives that enhance the resilience of existing cloud infrastructures against attacks and vulnerabilities, protecting data end-to-end, and thus making secure and trustworthy cloud systems a reality.

Achieving end-to-end security within the cloud environment is nevertheless not a straightforward task. Indeed, end-to-end security is at odds with current functionalities offered by the cloud, as shown in Figure 1. For example, while protecting cloud customers' data at rest usually requires data encryption solutions, such a security service inherently refrains cloud services from offering standard APIs for efficiently processing these encrypted data.

TREDISEC addresses security and privacy issues by analyzing, designing and implementing a set of cloud security primitives that are integrated naturally with existing cloud capabilities and functionalities, such as multi-tenancy or storage efficiency. Among other capabilities, these primitives support data reduction, enable secure data processing, enhance data availability and integrity and ensure user isolation and confidentiality in multi-tenant systems. Therefore, the innovation potential of TREDISEC covers:

- Deduplication on encrypted and multi-tenant data
- Means to verify the integrity and availability of multi-tenant data in presence of storage efficiency
- Secure deletion of multi-tenant data in presence of deduplication
- Storage efficiency in presence of securely outsourced database management services
- Secure outsourced analytics/processing in a multi-tenant environment
- Trustworthy, consistent access control for multi-tenancy settings
- Distributed enforcement of access control policies.

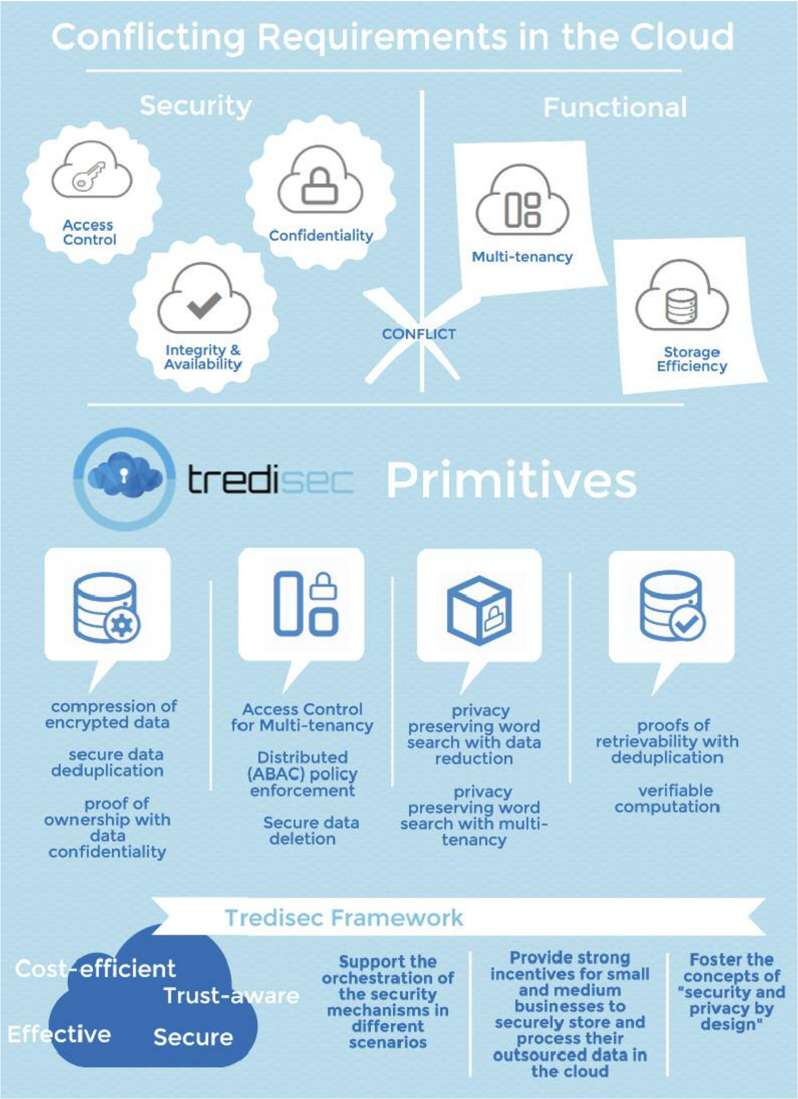

Conflicting Requirements in the Cloud
Security
Functional
Access Control
Confidentiality
Integrity & Availability
CONFLICT
Multi-tenancy
Storage Efficiency
tredisec
Primitives
compression of encrypted data
secure data deduplication
proof of ownership with data confidentiality
Access Control for Multi-tenancy
Distributed (ABAC) policy enforcement
Secure data deletion
privacy preserving word search with data reduction
privacy preserving word search with multi-tenancy
proofs of retrievability with deduplication
verifiable computation
Tredisec Framework
Cost-efficient
Trust-aware
Effective
Secure
Support the orchestration of the security mechanisms in different scenarios
Provide strong incentives for small and medium businesses to securely store and process their outsourced data in the cloud
Foster the concepts of "security and privacy by design"

*Figure 1: The TREDISEC approach.*

Addressing these innovations is only the first step of the project. The subsequent, and more challenging step, is to consider scenarios with multiple functional and security requirements to evaluate the effectiveness of our approach. These security primitives will be integrated and validated in real use case scenarios designed by four partners of the project, representing diverse business goals (for more detail about use cases see the Links section). Namely, our results will be validated onto (1) the Greek Research and Technology Network's computation and storage cloud (used by the University of Athens), (2) the cloud storage service provided by Arsys, (3) Morpho's identification solutions using biometric data that are stored in the cloud , and finally (4) SAP use case focuses on migration of legacy databases into a secure cloud, which is an important concern for small, mid-sized and large enterprises that wish to move their day-to-day business processes (e.g., ERP, CRM, HR) to the cloud.

We believe that the results of TREDISEC will have a big impact in business (both large companies and SMEs), allowing them to achieve greater business throughput, and lowering the barriers to enter new markets. We are fully aware of the problem of adoption of new technologies by companies and believe that TREDISEC will cover at least three of the top five barriers: security, data protection, trust, data access, and portability [3].

This paper is a joint work of the TREDISEC project consortium. The TREDISEC project is funded by the H2020 Framework of the European Commission under grant agreement no. 644412. In this project, nine renowned research institutions and industrial players with balanced expertise in all technical aspects of both security and cloud are working together to address the challenges of the project: Atos Spain (project coordinator), NEC Europe, IBM Research, ETH Zurich, Eurecom, Arsys Internet, Greek Research and Technology Network, SAP SE and Morpho (SAFRAN group).

**Links:**
TREDISEC official website: http://tredisec.eu
TREDISEC Use Cases Definition:
http://tredisec.eu/content/use-cases

**References:**

[1] Deadline – Sony hack: A Timeline (Last access: 05.11.15), http://kwz.me/QY
[2] Tripwire – The state of security. The Ashley Madison Hack – A Timeline (Last access: 05.11.15), http://kwz.me/QB
[3] European Commission:"Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Up-take", 2012, http://kwz.me/Qr

**Please contact:**

Beatriz Gallego-Nicasio Crespo
Atos, Spain
Tel: +34 912148800
E-mail: beatriz.gallego-nicasio@atos.net

# Smart Devices for the Home of the Future: A New Model for Collaboration

by Daniele Spoladore, Gianfranco Modoni and Marco Sacco

***An ontology based approach that aims at enhancing interoperability between home devices and services is being developed by the Italian "Design For All" research project.***

Contemporary design is characterized by a major paradigm shift: from traditional design focused on the "average man" to Universal Design, which takes into account a wide variety of possible human needs [1]. This paradigm is also applied to the field of Ambient Assisted Living (AAL) in the design of smart homes, living environments capable of anticipating and responding to the needs of their inhabitants through tailored services provided by diverse devices (e.g. sensors and actuators). However, these smart objects are managed by different software, based on their specific data format. Thus, the home of the future is currently characterized by a wide variety of data, which hinders efficient interactions between the devices involved. In order to address this issue, the Design For All research project, co-funded by the Italian Ministry for Education, University and Research within the cluster of initiatives for Technologies for Ambient Assisted Living, is developing semantic interoperability so that different systems can share and exploit the same information.

The main goal of the project is to develop a software architecture that supports the design phase for future smart homes suitable for any kind of user, enabling distributed devices to adapt to and react with the context. A platform denoted the Virtual Home Framework (VHF) [2] integrates knowledge about the domestic environment, the smart objects and the users. Semantic Web technologies have been adopted to formally describe this information (including the many linking elements) in an ontology, which is a "formal specification of a shared conceptualization" based on first-order logic languages (RDF and OWL). The ontology approach provides a holistic view of the smart home as a whole, considering the physical dimensions, the users involved, and their evolution over the time. It also allows the use of reasoning tools, able to derive new knowledge about the concepts and their relationships, thanks to inferencing rules specified in the Semantic Web Rule Language (SWRL).

The semantic model developed, called the Virtual Home Data Model (VHDM), provides a consistent representation of several knowledge domains; it is composed of several modules including: a) the Physiology model, to keep track of users' medical conditions over time; b) the Smart Object Model, which provides a description of the relationships between appliances and related functionalities; c) the Domestic Environment, which includes information on thermo-hygrometric conditions and air and light quality.

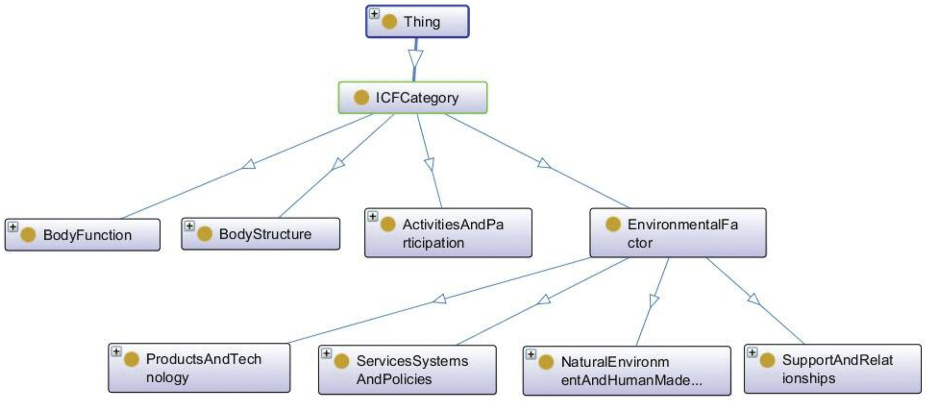


*Figure 1: A fragment of the taxonomy derived from ICF.*

A valid starting point for the design of the VHDM has been a study of the various reference models available in the literature covering the relevant knowledge domains. For example, the Physiology model is based on the International Classification of Functioning, Disability and Health (ICF) (Figure 1); the Smart Object model refers to the existing Units Ontology, which represents units of measurement and related concepts, in order to support measurements in several fields (e.g. physiological data on the user, environmental data and data related to the smart objects) (Figure 2).

A key topic was the selection of a valid database that could manage and reason over large amounts of semantic data. A survey of the state of the art of existing semantic repositories was carried out stipulating reasoning capability as the main criteria of evaluation. On the basis of this qualitative analysis [3], Stardog was finally adopted as the RDF-store and SPARQL end point for the validation of the framework, since, compared to other solutions, this allows a higher level of expressivity to represent the inferencing rules. The semantic repository has been installed on a cloud-based platform (Microsoft Azure), which guarantees an efficient management of the overall framework's horizontal scalability.

**Link:**
http://www.d4all.eu/en/

**References:**

[1] R. L. Mace, G. J. Hardie, J. P. Place: "Accessible Environments: Toward Universal Design", 1991.
[2] M. Sacco, E.G. Caldarola, G. Modoni, W. Terkaj: "Supporting the Design of AAL through a SW Integration Framework: The D4All Project", Universal Access in HCI, Springer, 2014.
[3] G. Modoni, M. Sacco, W. Terkaj: "A survey of RDF store solutions", in Proc. of ICE, 2014.

**Please contact:**

Daniele Spoladore, Gianfranco Modoni
ITIA-CNR, Italy
E-mail: {daniele.spoladore, gianfranco.modoni}@itia.cnr.it

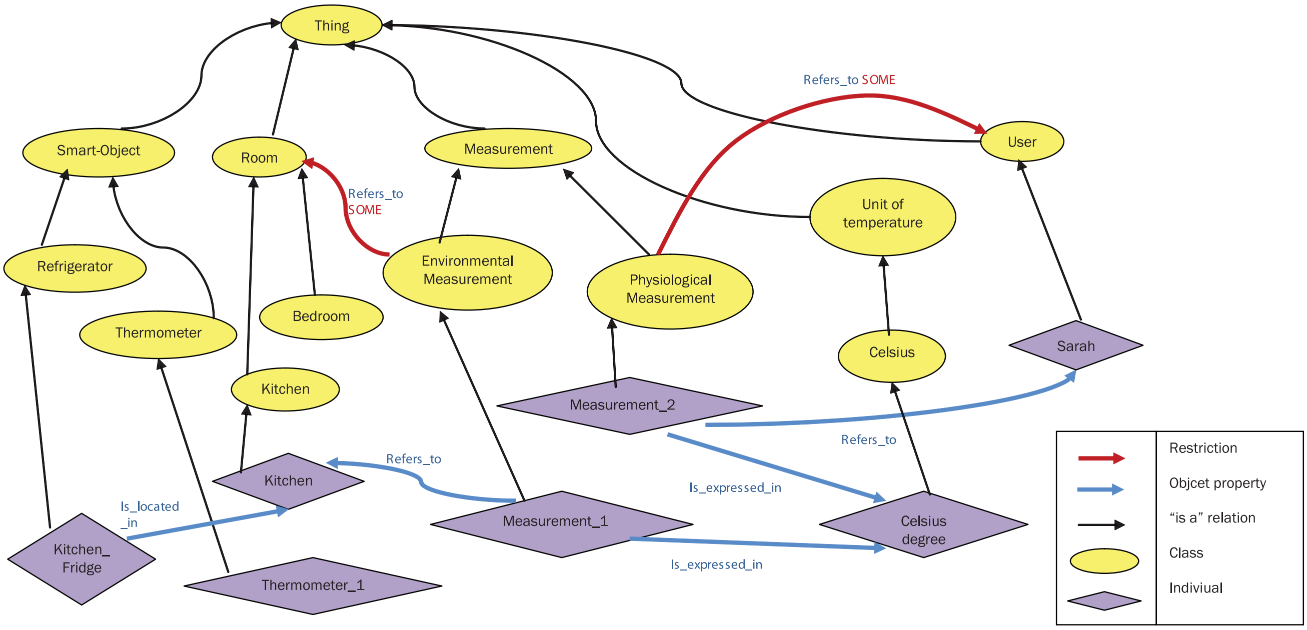


*Figure 2: An excerpt of the Smart Object model.*

# FOODWEB – Studying Food Consumption and Production Patterns on the Web

by Christoph Trattner, Tomasz Kuśmierczyk and Kjetil Nørvåg

***Food is a central part of our daily lives and is one of the most important factors that shape our health and well-being. Although research on individual food preferences has existed as a well-established field of study for several decades, there has been very little attention devoted to understanding recipe consumption and production patterns on the World Wide Web. To contribute to knowledge in this area, we recently initiated "FOODWEB", a research project that studies traces of users online with the goal of better understanding how we consume and produce food.***

Contrary to classic nutrition research, our methodology in the FOODWEB project is rather pragmatic. Rather than relying on time consuming and expensive user surveys, which are commonly used in nutrition research, we intend to mine traces of users online (e.g. from online food forums). We then analyze this data to draw conclusions about a population of users (e.g. in a country or a region) or at the individual level. The project started in October 2014 with support from ERCIM (in the form of an Alain Bensoussan Fellowship to the first author of this article, who was hosted by NTNU, Norway).

In order to conduct our research, we created a dataset by mining kochbar.de, one of the largest online food communities. Like other related websites, people can upload recipes, rate and comment on recipes, join or create groups to share cooking experiences or create social links to follow each other. What makes kochbar.de interesting as a test bed is not only their huge user base (and therefore the coverage of a huge population), but also the rich amount of metadata – which is usually not the case when conducting survey-based research. Interesting information available includes detailed descriptions of the ingredients used and the preparation steps needed to cook a meal, detailed nutrition facts per serving, and detailed user information, such as the user's gender and region. This dataset encompasses more than 400,000 recipes published between 2010 and 2014, and over 200,000 users.

Our project started with an exploratory analysis of the dataset and several research questions. Two initial papers were published at WWW'2015 [1] and TempWeb'2015 [2], with interesting findings on temporal effects in how we consume and produce food online. Interesting findings included temporal effects in terms of production and consumption over the course of a year (see Figure 1) and over a week. As a proxy for consumption we used ratings provided by the users to a given recipe, and "recipe uploads" were used a proxy for production. Another interesting finding was that some food categories are remaining significantly more interesting to the users over time than others – e.g. vegetarian recipes have a very short life time, while spring food stays interesting to people over a longer period. This finding is important since it suggests that temporal effects significantly influence food preferences, and researchers or engineers who are developing recommendation services in such platforms could, for example, take this into account.

Currently, we are in the process of obtaining data from other online community platforms. We are using this to investigate several interesting variables, including cultural differences, user characteristics – such as age and gender – and many more. The project is continuing to grow and current contributors include researchers from the Know-Center (Austria), NTNU (Norway), L3S (Germany), Freie University of Bolzano (Italy), and TU-München (Germany).

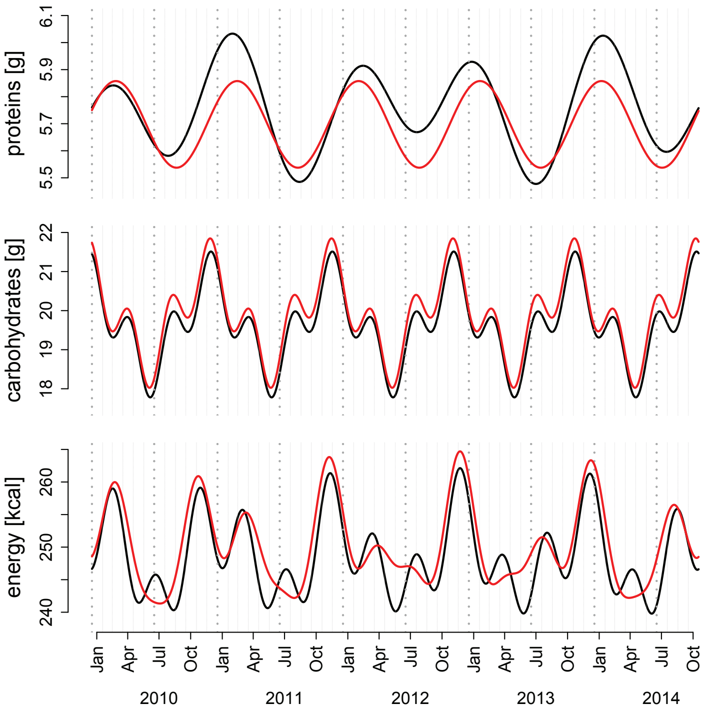


*Figure 1: Seasonal trends in online recipe production (red plot) and consumption (black plot) for proteins, carbohydrates and calories per 100g.*

**References:**

[1] T. Kuśmierczyk, C. Trattner, K. Nørvåg: "Temporality in Online Food Recipe Consumption and Production", In Proc. of WWW'15, 2015.

[2] T. Kuśmierczyk, C. Trattner, K. Nørvåg: "Temporal Patterns in Online Food Innovation", in Proc. of TempWeb'201, 2015.

**Please contact:**
Christoph Trattner
Know-Center, Graz, Austria
E-mail: trattner.christoph@gmail.com
*Christoph Trattner was an ERCIM Fellow hosted by NTNU*

# Robots Learn Actions and Cooperative Tasks by Imitation

by Maria Koskinopoulou and Panos Trahanias

***Imitation learning, which involves a robot observing and copying a human demonstrator, empowers a robot to "learn" new actions, and thus improve robot performance in tasks that require human-robot collaboration (HRC). We formulate a latent space representation of observed behaviors, and associate this representation with the corresponding one for target robotic behaviors. Effectively, a mapping of observed to reproduced actions is devised that abstracts action variations and differences between the human and robotic manipulators.***

Learning new behaviors enables robots to expand their repertoire of actions and effectively learn new tasks. This can be formulated as a problem of learning mappings between world states and actions. Such mappings, called policies, enable a robot to select actions based on the current world state. The development of such policies by hand is often very challenging, so machine learning techniques have been proposed to cope with the inherent complexity of the task. One distinct approach to policy learning is "Learning from Demonstration" (LfD), also referred in the literature as Imitation Learning or Programming by Demonstration (PbD) [1].

Our research in the LfD field is directed towards furnishing robots with task-learning capabilities, given a dataset of different demonstrations. Task learning is treated generically in our approach, in that we attempt to abstract from the differences present in individual demonstrations and summarize the learned task in an executable robotic behavior. An appropriate latent space representation is employed that effectively encodes demonstrated actions. A transformation from the actor's latent space to the robot's latent space dictates the action to be executed by the robotic system. Learned actions are in turn used in human-robot collaboration (HRC) scenarios to facilitate fluent cooperation between two agents – the human and the robotic system [2].

We conducted experiments with an actual robotic arm in scenarios involving reaching movements and a set of HRC tasks. The tasks comprised a human opening a closed compartment and removing an object from it, then the robot was expected to perform the correct action to properly close the compartment. Types of compartments tested included: drawers, closets and cabinets. The respective actions that the robotic systems should perform consisted of "forward pushing" (closing a drawer), "sideward reaching" (closing an open closet door) and "reaching from the top" (closing a cabinet). Furthermore, the robots were trained with two additional actions, namely: "hello-waving" and "sideward pointing". Those additional actions were included in order to furnish the robots with a richer repertoire of actions and hence operate more realistically in the HRC scenarios.

In these experiments, the robotic arm is expected to reproduce a demonstrated act, taught in different demo-sessions. The robotic arm should be able to reproduce the action regardless of the trajectory to reach the goal, the hand configuration, the trajectory plan of a reaching scenario or the reaching velocity, etc. The end effect of the learning process is the accomplishment of task-reproduction by the robot. Our experiments are developed using the robotic arm JACO designed by Kinova Robotics [L1], and have also been tested on the humanoid robot NAO by Aldebaran [L2].

Actual demonstrations may be performed in different ways. In our study, we recorded RGB-D videos of a human performing the respective motion act. From each video, we

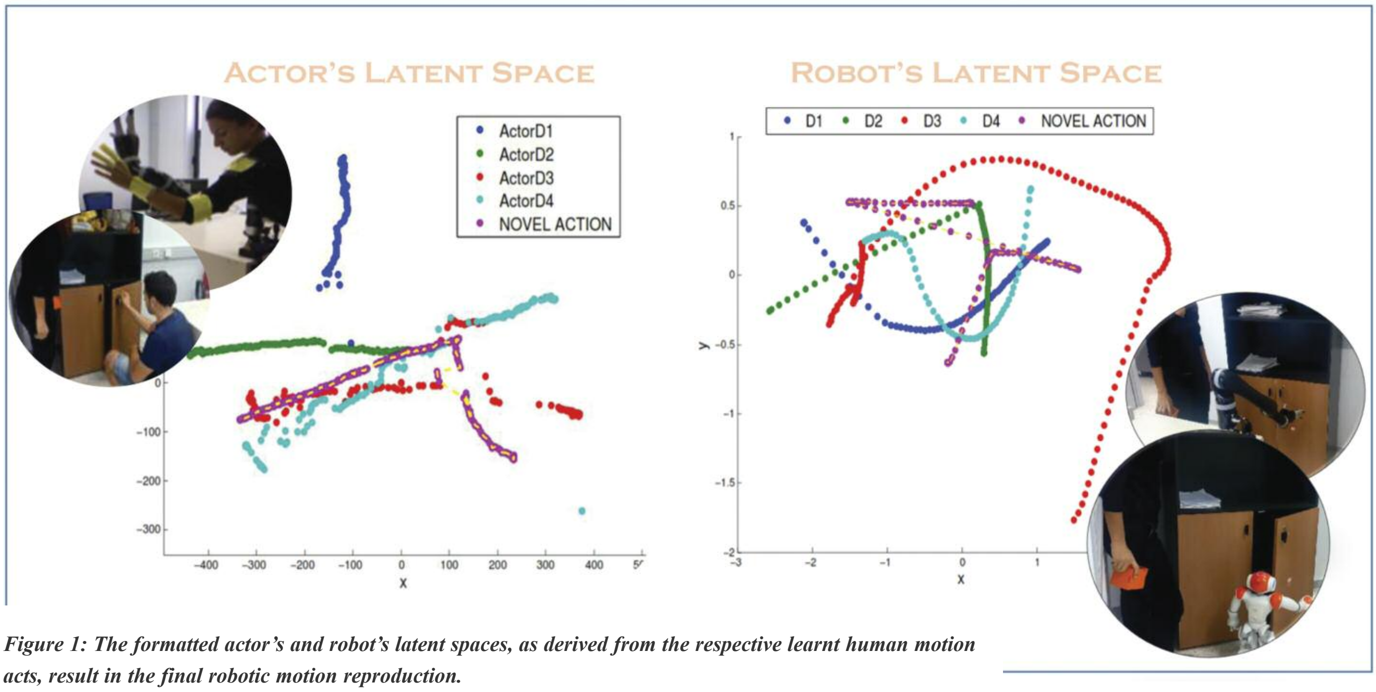


*Figure 1: The formatted actor's and robot's latent spaces, as derived from the respective learnt human motion acts, result in the final robotic motion reproduction.*

extract the {x,y,z} coordinates of each joint, previously marked with prominent (yellow) markers. By tracking three joints of the demonstrator, we form a 9D representation of the observed world. The derived trajectories of human motion acts are subsequently represented in a way that is readily amenable to mapping in the manipulator's space. The Gaussian Process Latent Variable Model (GP-LVM) representation is employed in order to derive an intermediate space, termed latent space, which abstracts the demonstrated behavior [3]. The latent space features only the essential characteristics of the demonstration space: those that have to be transferred to the robot-execution space, in order to learn the task. Effectively, the mapping of the observed space to the actuator's space is accomplished via an intermediate mapping in the latent spaces (Figure 1).

Our initial results regarding the latent space representation are particularly encouraging. We performed many demonstrations of left-handed reaching and pushing movements. Interestingly, the resulting representations assume overlapping trajectories that are distinguishable, facilitating compact encoding of the behavior. We are currently working on deriving the analogous representation for actions performed by the robotic arm, as well as associating the two representations via learning techniques.

In summary, we have set the ground for research work in Robotic LfD and have pursued initial work towards building the latent representation of demonstrated actions. In this context, we're aiming at capitalizing on the latent space representation to come up with robust mappings of the observed space to the robot–actuator's space. Preliminary results indicate that the adopted representation in LfD tasks is appropriate.

**Links:**
[L1] http://kinovarobotics.com/products/jaco-robotics/
[L2] https://www.aldebaran.com/en/humanoid-robot/nao-robot

**References:**

[1] S. Schaal, A. Ijspeert, A.Billard: "Computational approaches to motor learning by imitation" Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 537–547, 2003.

[2] M. Koskinopoulou, S. Piperakis, P. Trahanias: "Learning from Demonstration Facilitates Human-Robot Collaborative Task Execution", in Proc. 11th ACM/IEEE Intl. Conf. Human-Robot Interaction (HRI 2016), New Zealand, March 2016.

[3] N.D. Lawrence: "Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data", Advances in Neural Information Processing Systems 16, 329–336, 2004.

**Please contact:**

Maria Koskinopoulou and Panos Trahanias
ICS-FORTH,, Greece
E-mail: {mkosk, trahania}@ics.forth.gr

# Unlocking the Secrets of the Dynamics of Earth's Interior

by Costas Bekas

***"Mantle convection" is the term used to describe the transport of heat from the interior of the Earth to its surface. Mantle convection drives plate tectonics – the motion of plates that are formed from the Earth's crust and upper mantle – and is a fundamental driver of earthquakes and volcanism. We modelled mantle convection using massively parallel computation.***

In 2013, a group of IBM and university scientists won the ACM Gordon Bell Prize for work on cavitation [1]. Completely against conventional wisdom, their work demonstrated that, with advanced algorithmic re-engineering, it is possible to achieve close to theoretical peak performance on millions of cores for computational fluid dynamics simulations. In the meantime, Professors Omar Ghattas (University of Texas at Austin), Michael Gurnis (California institute of Technology), and Georg Stadler (New York University) and their groups were collaborating on creating a very high resolution/high fidelity solver for convection in Earth's mantle. The breakthrough represented by the Bell prize was the spark that brought together the Texas-Caltech-NYU team with the group of Costas Bekas at IBM to form a collaboration aimed at achieving unprecedented scaleout of the mantle convection solver.

Extreme scaleout would be a decisive step towards the goals of the original research, which aims to address such fundamental questions as: What are the main drivers of plate motion – negative buoyancy or convective shear traction? What is the key process governing the occurrence of earthquakes – the material properties between the plates or the tectonic stress? [2].

A series of key advances in geodynamic modelling, numerical algorithms and massively scalable solvers that were developed over the last several years by the academic members of the team was combined with deep algorithm re-engineering and massively parallel optimization knowhow from IBM Research, to achieve unparalleled scalability. This result offers geophysicists powerful tools to improve their understanding of the forces behind the origin of earthquakes and volcanism [2]. More broadly, the combination of advanced algorithms and the massively parallel optimizations demonstrated in the work applies much more generally and will enable scientists in multiple fields to make new discoveries and make it possible for industries to greatly reduce the time it takes to invent, test, and bring new products to market – including, for example, new materials and energy sources.

The problem of modeling mantle convection at a global scale with realistic material properties is notoriously difficult. The problem is characterized by an extremely wide range of length scales that need to be resolved—from less than a kilometre at tectonic plate boundaries to the 40,000 kilometre circumference of the Earth. Similarly, the rock that comprises the mantle has a viscosity (which represents resistance to shearing) that varies by over six orders of magnitude. The grid on which the computations were carried out thus had to be adapted locally to the length scale and viscosity variation present. Finally, the partial differential equations describing mantle convection are highly nonlinear. The result was a model with 602 billion equations and variables – the solution of which could be undertaken on only the most powerful supercomputers.

Rather than using simpler explicit solvers, which are easier to scale on large supercomputers (since they involve only nearest neighbour communication) but at the same time are extremely inefficient (as they require a very large number of iterations to converge), the team achieved extreme scaleout to millions of cores for an innovative implicit solver, previously developed by the Texas and NYU members of the team. This required extensive algorithmic implementation re-engineering and massively parallel adaptation as unforeseen bottlenecks arise when millions of cores are used. This success allowed the simulation of mantle convection with an unprecedented level of fidelity and resolution. Global high-resolution models such as this are critical for understanding the forces driving and resisting global tectonic plate motion.

Combining advances in mathematics, algorithms and massively parallel computing, simulations with the optimized mantle convection model demonstrated super-efficiency in running on two of the most powerful high-performance computers in the world [2]. Development and science runs utilized the Blue Gene/Q system at CCI, Rensselaer Polytechnic Institute and the Stampede system at the Texas Advanced Computing Center; Initial massive scaling runs were made on the Juqueen Blue Gene/Q supercomputer at Juelich Supercomputing Centre in Germany and then further scaled up to run on Sequoia, a Blue Gene/Q machine owned by the U.S. Lawrence Livermore National Laboratory. Sequoia comprises 1.6 million processor cores capable of a theoretical peak performance of 20.1 petaflops. A huge challenge is efficiently scaling to all of the1,572,864 processor cores of the full Sequoia. The overall scale-up efficiency using all 96 racks was 96% compared to the single rack run. This is a record-setting achievement and demonstrates that, contrary to widely-held beliefs, numerically optimal implicit solvers can indeed almost perfectly scale to millions of cores, with careful algorithmic design and massive scaleout knowhow.

Understanding mantle convection has been named one of the “10 Grand Research Questions in Earth Sciences” by the U.S. National Academies. The authors of [2] hope their work provides earth scientists with a valuable new tool to help them pursue this grand quest. The work won the 2015 ACM Gordon Bell Prize.

Many colleagues contributed to this work, too many to be included as authors here but all are heartily acknowledged: Johann Rudi, A. Cristiano I. Malossi, Tobin Isaac, Georg Stadler, Michael Gurnis, Peter W. J. Staar, Yves Ineichen, Costas Bekas, Alessandro Curioni and Omar Ghattas.

**Links:**

http://www.zurich.ibm.com/mcs/compsci/parallel/
http://www.zurich.ibm.com/mcs/compsci/engineering/
https://www.youtube.com/watch?v=ADoo3jsHqw8

**References:**

[1] D. Rossinelli et al.: “11 PFLOP/s simulations of cloud cavitation collapse”, Proc. of SC2013, Paper 3, Denver, Colorade, 2015.

[2] J. Rudi et al.: “An Extreme-Scale Implicit Solver for Complex PDEs: Highly Heterogeneous Flow in Earth’s Mantle”, in Proc. of SC2015, Paper 5, Austin Texas, 2015.

**Please contact:**

Costas Bekas, IBM Research – Zurich, Switzerland
E-mail: bek@zurich.ibm.com

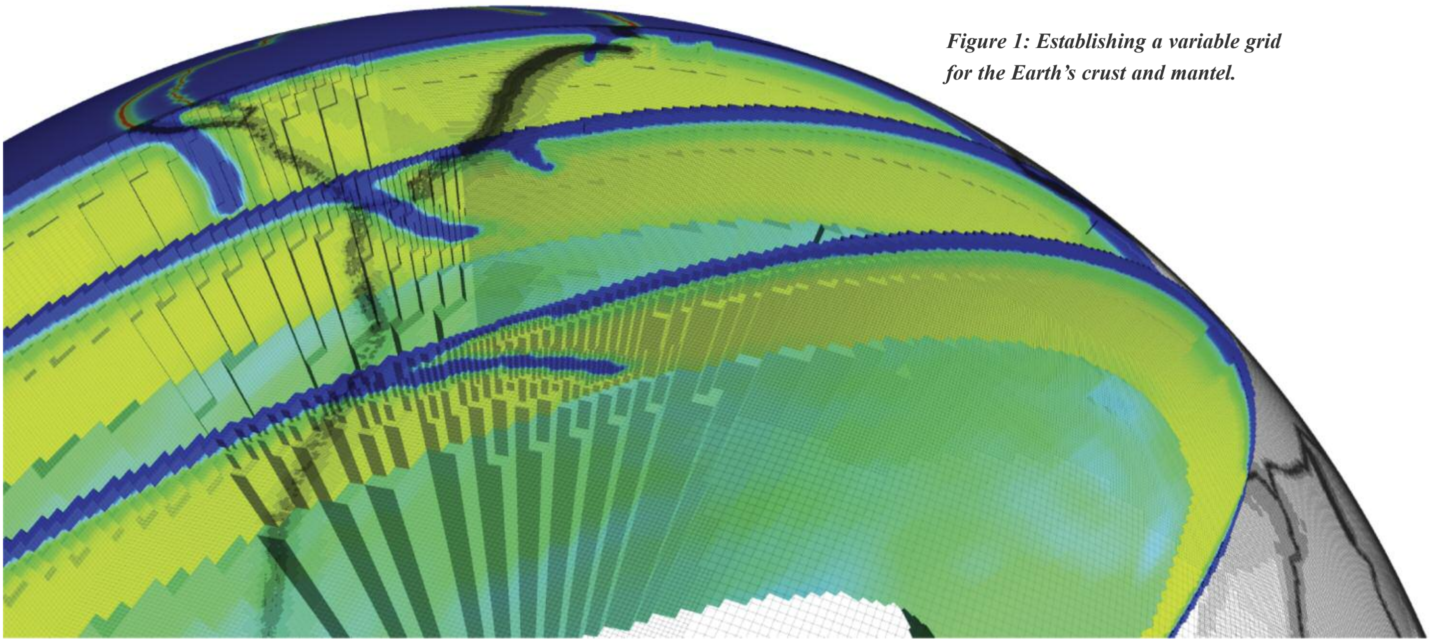


*Figure 1: Establishing a variable grid for the Earth’s crust and mantel.*

Call for Papers

## EMC$^2$ Summit in conjunction with the CPS Week 2016

2016, Vienna, Austria, 11 April 2016

EMC$^2$ – “Embedded Multi-Core systems for Mixed Criticality applications in dynamic and changeable real-time environments” is an ARTEMIS/ECSEL Joint Undertaking project in the ARTEMIS Innovation Pilot Programme “Computing platforms for embedded systems” (see http://www.artemis-emc2.eu/). EMC$^2$ will present work in progress and intermediate results of the ongoing project and invites researchers and industry to contribute with papers from their current work in the areas addressed and related projects. Topics of interest include, but are not limited to:

- architectures and platforms for embedded (cyber-physical) systems
- application Models and Design Tools for Mixed-Critical, Multi-Core CPS
- dynamic runtime environments and services
- multi-core hardware architectures and concepts
- system design platform, tools, models and interoperability
- applications of multi-core cyber-physical systems: avionics, automotive, space, cross-domain and other applications
- safety and security co-engineering in open dynamic CPS
- next generation embedded/cyber-physical systems
- standardization, qualification and certification issues of complex critical CPS

The EMC$^2$ Summit is scientifically co-sponsored by the ERCIM Dependable Embedded Software-intensive Systems Working Group and EWICS TC7 (European Workshop on Industrial Computer Systems, TC7, Safety, Reliability and Security), and co-hosted by the ARTEMIS projects ARROWHEAD, CRYSTAL and the standardization Innovation Action CP-SETIS.

Deadlines:

- Full paper submission: 7 February 2016
- Paper acceptance notification: 28 February 2016
- Camera ready paper submission: 10 March 2016

**More information:**
http://www.artemis-emc2.eu/events/2016041116-emc2-summit-2016-at-cps-week-2016-vienna-austria-call-for-papers/

**Please contact:**
Erwin Schoitsch, AIT Austrian Institute of Technology GmbH
Erwin.Schoitsch@ait.ac.at

## More on IFIP TC6’s Open Digital Library

As announced in the July issue of ERCIM News, the IFIP (International Federation for Information Processing) Technical Committee 6’s (TC6’s) open digital library: http://dl.ifip.org/ is up and running. TC6 deals with Communication Systems and organizes a number of conferences in this area each year. Proceedings of some 30 TC6 conferences are available in the library. The library also gives access to many other IFIP conferences so that the overall number of conferences covered is more than 140. In contrast to other digital libraries, IFIP TC6 does not charge authors for getting their papers published therein.

At the time of writing the most downloaded paper is from IFIP’s Technical Committee 2’s conference on Open Source Systems which took place in Como in June, 2006: “Evolution of Open Source Communities” by Michael Weiss, Gabriella Mouroiu and Ping Zhao. (http://dl.ifip.org/db/conf/oss/oss2006/WeissMZ06.pdf).It has been cited 35 times, according to Google.

During the month of November, the most accessed conference was CNSM2015, the 11th International Conference on Network and Service Management which was held in Barcelona from 9 to 13 November.

**Links:**
http://dl.ifip.org/
http://dl.ifip.org/db/conf/cnsm/index.html

**Please contact:**
Harry Rudin, Swiss Representative to IFIP TC6
hrudin@sunrise.ch

Book Announcement

## Online Social Networks: Human Cognitive Constraints in Facebook and Twitter Personal Graphs

V. Arnaboldi, A. Passarella, M. Conti, R.I.M. Dunbar

ELSEVIER
COMPUTER SCIENCE REVIEWS AND TRENDS
NETWORKING-COMMUNICATIONS
ONLINE SOCIAL NETWORKS
HUMAN COGNITIVE CONSTRAINTS IN FACEBOOK AND TWITTER PERSONAL GRAPHS
VALERIO ARNABOLDI | ANDREA PASSARELLA | MARCO CONTI | ROBIN I.M. DUNBAR

This book provides new insights into the structural properties of personal online social networks and the mechanisms underpinning human online social behavior.

As the availability of digital communication data generated by social media is revolutionizing the field of social networks analysis, the text discusses the use of large- scale datasets to study the structural properties of online ego networks, to compare them with the properties of general human social networks, and to highlight additional properties. Users will find the data collected and conclusions drawn useful during design or research service initiatives that involve online and mobile social network environments.

The book is intended for researchers and students active in the field of social network analysis, and professionals in the field of Online Social Networking service design.

Elsevier, Software Engineering and Programming, October 2015, 116 pages. ISBN: 9780128030233

# Lynda Hardman President Informatics Europe

Lynda Hardman (CWI) has been elected President of the Board of Informatics Europe, the European association of computer science departments and research laboratories, starting from January 2016. ERCIM and Informatics Europe have agreed to increase cooperation on activities that are relevant to both organisations. Hardman succeeds Carlo Ghezzi (Politecnico di Milano, Italy). The election took place at the annual General Assembly during the 11th European Computer Science Summit in Vienna, Austria.

Photo: Bob Bronshoff

Hardman has been serving as member of the Informatics Europe Board since 2012. She founded the Informatics Europe Working Group on Women in Informatics and was responsible for the publication of the booklet "More Women in Informatics Research and Education". She regularly speaks at events to inspire and encourage a new generation of female talent in computer science.

"Informatics research is a driving force of European economy and society", Hardman says. "It is my honour to be able to serve the community in my role as president of Informatics Europe. At a European scale we can increase the quality of informatics research, ensure that educational innovations can be more easily disseminated, and also strengthen the development of commercial cooperation across Europe."

Lynda Hardman is a member of the management team at CWI, a senior researcher in the Information Access research group and coordinator of the CWI Information theme. She is a professor in Multimedia Discourse Interaction at Utrecht University. Her research focuses on human-computer interaction. In 2014 she was named Distinguished Scientist by ACM, for her individual contribution to the field of computing.

**More information:** http://kwz.me/Rm

# QuSoft Research Center for Quantum Software Launched

QuSoft, the first research center dedicated to quantum software, was launched on 3 December 2015. This joint initiative of CWI, the University of Amsterdam and VU University will be located at Amsterdam Science Park in The Netherlands. QuSoft's mission is to develop new protocols, algorithms and applications that can be run on small and medium-sized prototypes of a quantum computer. Its main focus is the development of quantum software, requiring fundamentally different techniques and approaches from conventional software. QuSoft will specifically target few-qubit applications, quantum testing and debugging, quantum architecture and quantum cryptography.

Directors are Harry Buhrman (CWI and UvA) and Kareljan Schoutens (UvA). The new center will be hosted by CWI. It aims to grow to 35 to 40 researchers and will be home to four research groups, each with numerous PhD students and postdocs. The QuSoft initiative has been warmly welcomed internationally and enjoys strong support from the European and North American quantum computing communities. The structural funding of QuSoft allows for further strengthening of the existing ties with quantum computing research organizations worldwide and researchers from leading computer industry partners, including Google, Microsoft and ID Quantique.

**More information:** http://www.qusoft.org

# ERCIM "Alain Bensoussan" Fellowship Programme

ERCIM offers fellowships for PhD holders from all over the world. Topics cover most disciplines in Computer Science, Information Technology, and Applied Mathematics. Fellowships are of 12-month duration, spent in one ERCIM member institute. Fellowships are proposed according to the needs of the member institutes and the available funding.

**Application deadlines: 30 April and 30 September.**

**More information:** http://fellowship.ercim.eu/

ERCIM
European Research Consortium
for Informatics and Mathematics

**ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development, in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.**

**ERCIM is the European Host of the World Wide Web Consortium.**

Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
http://www.iit.cnr.it/

Czech Research Consortium
for Informatics and Mathematics
FI MU, Botanicka 68a, CZ-602 00 Brno, Czech Republic
http://www.utia.cas.cz/CRCIM/home.html

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
http://www.cwi.nl/

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
http://www.fnr.lu/

FWO
Egmontstraat 5
B-1000 Brussels, Belgium
http://www.fwo.be/

F.R.S.-FNRS
rue d'Egmont 5
B-1000 Brussels, Belgium
http://www.fnrs.be/

Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
http://www.ics.forth.gr/

Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
http://www.iuk.fraunhofer.de/

INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal

Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
http://www.inria.fr/

I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
http://www.isi.gr/

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

SBA Research gGmbH
Favoritenstraße 16, 1040 Wien
http://www.sba-research.org/

SICS Swedish ICT
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

Spanish Research Consortium for Informatics and Mathematics
D3301, Facultad de Informática, Universidad Politécnica de Madrid
28660 Boadilla del Monte, Madrid, Spain,
http://www.sparcim.es/

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
http://www.sztaki.hu/

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
http://www.cs.ucy.ac.cy/

University of Southampton
University Road
Southampton SO17 1BJ, United Kingdom
http://www.southampton.ac.uk/

Universty of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
http://www.mimuw.edu.pl/

Universty of Wroclaw
Institute of Computer Science
Joliot-Curie 15, 50–383 Wroclaw, Poland
http://www.ii.uni.wroc.pl/

VTT Technical Research Centre of Finland Ltd
PO Box 1000
FIN-02044 VTT, Finland
http://www.vttresearch.com

**Subscribe to ERCIM News and order back copies at http://ercim-news.ercim.eu/**