BigInsight

STATISTICS FOR THE KNOWLEDGE ECONOMY

ANNUAL REPORT 2017

sfi Centre for Research-based Innovation

The Research Council of Norway

BigInsight

Digital technology, despite its seeming ubiquity, has only begun to penetrate industries. As it continues its advance, the implications for revenues, profits, and opportunities will be dramatic.

(The case for digital reinvention - McKinsey Quarterly - February 2017)

CONTENT

SUMMARY

BigInsight is a Norwegian centre for research-based innovation, funded by the Norwegian Research Council and a consortium of private and public partners. We produce innovative solutions for key challenges facing our partners, by developing original statistical and machine learning methodologies. Exploiting complex, huge and unique data resources and substantial scientific, industrial and business knowledge, we construct personalised solutions, predict dynamic behaviours and control processes that are at the core of the partners' innovation strategies. Digitalisation of the Norwegian industry and society benefits from BigInsight that produces instruments for the analysis of data.

We discover radically new ways to target products, services, prices, therapies and technologies, towards individual needs and conditions, thus providing improved quality, precision and efficacy. We develop new approaches to predict critical quantities which are unstable and in transition, such as customer behaviour, patient health, electricity prices, machinery condition. This is possible thanks to the unprecedented availability of large scale measurements and individual information together with new statistical theory, computational methods and algorithms able to extract knowledge from complex and high dimensional data.

When we develop methods and algorithms we consider five principles: responsibility, explainability, accuracy, auditability and fairness. In the era of digitalization, BigInsight creates unique competence and capacity for the Norwegian knowledge-based economy, contributing to the development of a sustainable and better society.

Research at BigInsight will lead to value creation and will support our partners' leading position.

This is the annual report of the third year of BigInsight. Innovation results are highlighted, together with the broad spectrum of research projects.

> *BigInsight has an important role in Norway's digital revolution: our theories, methodologies, analysis and algorithms will create new business opportunities, accelerate solutions to global societal challenges, inform policymaking, and improve the environment, health, welfare and infrastructure.*
>
> (Arnoldo Frigessi, director of BigInsight)

VISION AND OBJECTIVES

Fulfilling the promise of the big data revolution, the center produces analytical tools to extract knowledge from complex data and delivers biginsight. Despite extraordinary advances in the collection and processing of information, much of the potential residing in contemporary data sources remains unexploited.

Digitalisation means producing data, organizing and storing data, accessing data and analyzing data. BigInsight works in the latter axis of digitalisation. There is a dramatic scope for industries, companies and nations – including Norway – to create value from employing novel ways of analysing complex data. The complexity, diversity and dimensionality of the data, and our partner's innovation objectives, pose fundamentally new challenges to statistical inference. We develop original, cutting-edge statistical, mathematical and machine learning methods, produce high-quality algorithms implementing these approaches and thereby deliver new, powerful, and operational solutions.

BigInsight's research converges on two central innovation themes:

- **personalised solutions:** to move away from operations based on average and group behaviour towards individualised actions
- **predicting transient phenomena:** to forecast the evolution of unstable phenomena for system or populations, which are not in equilibrium, and to design intervention strategies for their control

Our solutions are significantly better than the state-of-the-art, thanks to courageous and creative methodologies that extract knowledge from structure in complex data and integrate data from various sources. Generic methodology and their new applications are published in international scientific journals.

Through training, capacity building and outreach, BigInsight contributes to growth and progress in the private and public sector, in science and society at large, preparing a new generation of statisticians and machine learners ready for the knowledge based economy of the future.

10 Big Ideas for Future NSF Investments. Harnessing Data for 21st Century Science and Engineering:

> *"(we will build the) 21st-century workforce capable of working effectively with data ... creating innovations that drive the nation's economy and educating the next generation of scientists and engineers."*

(The National Science Foundation, USA, 2016.
https://www.nsf.gov/news/special_reports/big_ideas/harnessing.jsp)

Personalised solutions

The core operation of our partners involves interacting with many individual units: at Telenor, for example, millions of individual mobile phone customers are part of a communication network; at Gjensidige, a million policyholders share risks of contingent, uncertain losses; for DNB, customers transfer money and receive loans; at OUS, cancer patients need to be treated in the most effective personalized way; for DNV GL and ABB, hundreds of sensors register the functional state and operation of a ship at sea. There are many common characteristics:

- a high number of units/individuals/sensors;
- in some cases, massive data for each unit; in other cases, more limited information;
- complex dependence structure between units;
- new data types, new technologies, new regulations are available;
- in most cases, units have their own strategies and are exposed to their environment.

Each partner has specific management objectives for its units, but they share the goal to fundamentally innovate the management of their units, by recognising similarities and exploiting diversity between units. This will allow personalised marketing, personalised products, personalised prices, personalized recommendations, personalised risk assessments, personalised fraud assessment, personalised screening, personalised therapy, sensor based condition monitoring, individualised maintenance schemes, individualised power production and more – each providing value to our partner, to the individuals and to society: better health, reduced churn, strengthened competitiveness, reduced tax evasion, improved fraud detection and optimised maintenance plans.

Predicting transient phenomena

The modern measurement instruments, the new demands of markets and society and a widespread focus on data acquisition, is often producing high frequency time series data. As never before, we are able to measure processes evolving while they are not in a stable situation, not in equilibrium. A patient receiving treatment, a sensor on a ship on sea, a customer offered products from several providers, a worker who lost his job, the price of an asset in a complex market – all examples of systems in a transient phase. DNB, NAV, Skatteetaten, SSB, Telenor and Gjensidige are interested in the prediction of certain behaviours of their customers and service users, predicting churn or fraud activities. In the health area, the availability of real time monitoring of patients and healthcare institutions allows completely new screening protocols and treatment monitoring, real time prevention and increased safety. For ABB and DNV GL high dimensional times series are generated by sensors monitoring a ship, with the purpose of predicting operational drifts or failures and redesigning inspection and maintenance protocols. The objective is to predict the dynamics, the future performance and the next events. Importantly, real time monitoring of such transient behaviour and a causal understanding of the factors which affect the process, allow optimal interventions and prevention. While the concrete objectives are diverse, we exploit very clear parallels:

- systems operate in a transient phase, out of equilibrium and exposed to external forcing;
- in some cases, there are many time series which are very long and with high frequency; in other cases, short and with more irregular measurements;
- complex dependence structure between time series;
- unknown or complex causes of abnormal behaviour;
- possibilities to intervene to retain control.

BigInsight develops new statistical methodology that allow our partners to produce new and more precise predictions in unstable situations, in order to make the right decisions and interventions.

ORGANISATION

Board in 2017

Andree Underthun, ABB
Tron Even Skyberg, DNB, chairman
Bobbie Nicole Ray-Sannerud, DNV GL
Birgitte F. De Blasio, Folkehelseinstituttet
Erlend Willand-Evensen, Gjensidige
Ellen Charlotte Stavseth Paaske, Hydro
Ulf Andersen, NAV
Lars Holden, Norsk Regnesentral
André Teigland, Norsk Regnesentral
Peder Heyerdahl Utne, Oslo University Hospital
Marcus Zackrisson, Skatteetaten
Jørn Leonhardsen, SSB
Astrid Undheim, Telenor
Bård Støve, University of Bergen
Nadia Slavila Larsen, University of Oslo

Observer: Terje Strand, Research Council of Norway
The board had 2 meetings in 2017. All partners are represented in the Board.

Legal organisation

BigInsight is hosted by NR.
Legal and administrative responsible:
Managing director Lars Holden

Center Leader

Prof. Arnoldo Frigessi, UiO Director

Co-Directors

Ass. Research Director Kjersti Aas, NR
Prof. Ingrid Glad, UiO
Ass. Prof. Ingrid Hobæk Haff, UiO
Ass. Research Director Anders Løland, NR
Research Director André Teigland, NR

Principal Investigators

Kjersti Aas, NR
Magne Aldrin, NR
Arnoldo Frigessi, UiO
Ingrid Glad, UiO
Clara Cecilie Günther, NR
Ingrid Hobæk Haff, UiO
Alex Lenkoski, NR
Anders Løland, NR
Carlo Mannino, UiO
Magne Thoresen, UiO

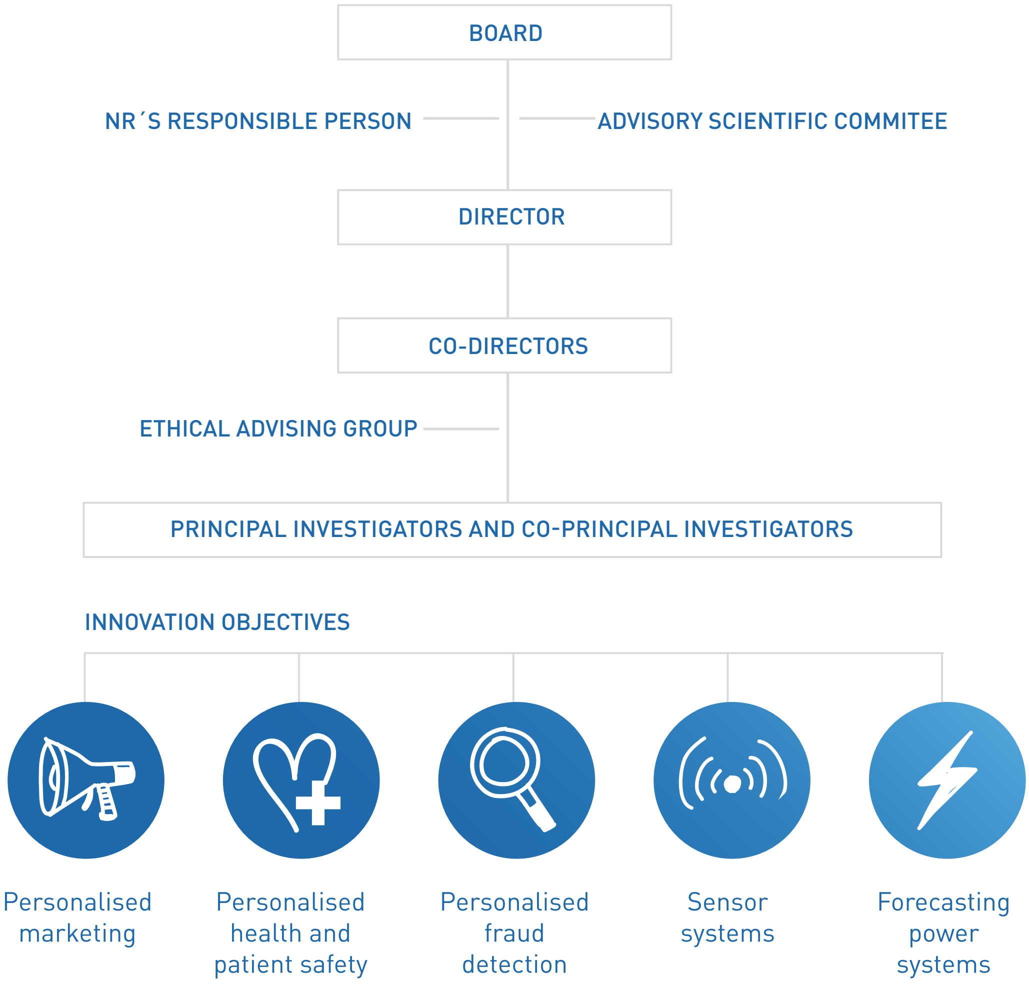
Administrative Coordinator

Unni Adele Raste, NR

Scientific Advisory Committee (SAC)

Prof. Idris Eckley, Lancaster U, UK
Prof. Samuel Kaski, U. Helsinki, Finland
Prof. Geoff Nicholls, U. Oxford, UK
Prof. Marina Vannucci, Rice U, Houston, USA
Senior Lecturer Veronica Vinciotti, Brunel U of London, UK

BigInsight


BOARD

NR´S RESPONSIBLE PERSON

ADVISORY SCIENTIFIC COMMITEE

DIRECTOR

CO-DIRECTORS

ETHICAL ADVISING GROUP

PRINCIPAL INVESTIGATORS AND CO-PRINCIPAL INVESTIGATORS

INNOVATION OBJECTIVES

Personalised marketing

Personalised health and patient safety

Personalised fraud detection

Sensor systems

Forecasting power systems

PARTNERS

Partners

- Norsk Regnesentral (host institute) (NR)
- University of Oslo (UiO)
- University of Bergen (UiB)
- ABB
- DNB
- DNV-GL
- Gjensidige
- Hydro
- Telenor
- NAV (Norwegian Labour and Welfare Administration)
- SSB (Statistics Norway)
- Skatteetaten (Norwegian Tax Administration)
- OUS (Oslo University Hospital)
- Folkehelseinstituttet (Norwegian Institute of Public Health, NIPH)
- Kreftregisteret (Cancer Registry of Norway)

Cooperation between the partners of BigInsight

There have been two board meetings in 2017, where all partners are represented. In addition to close cooperation with the researchers at NR and the universities, there have been several meetings within the separate Innovation Objectives where partners have met and exchanged ideas.

In October the annual Big Insight Day was held at the premises of Skatteetaten. In this successful event a wide selection of our projects, ideas and first results were presented and discussed among representatives from the different partners and project members.

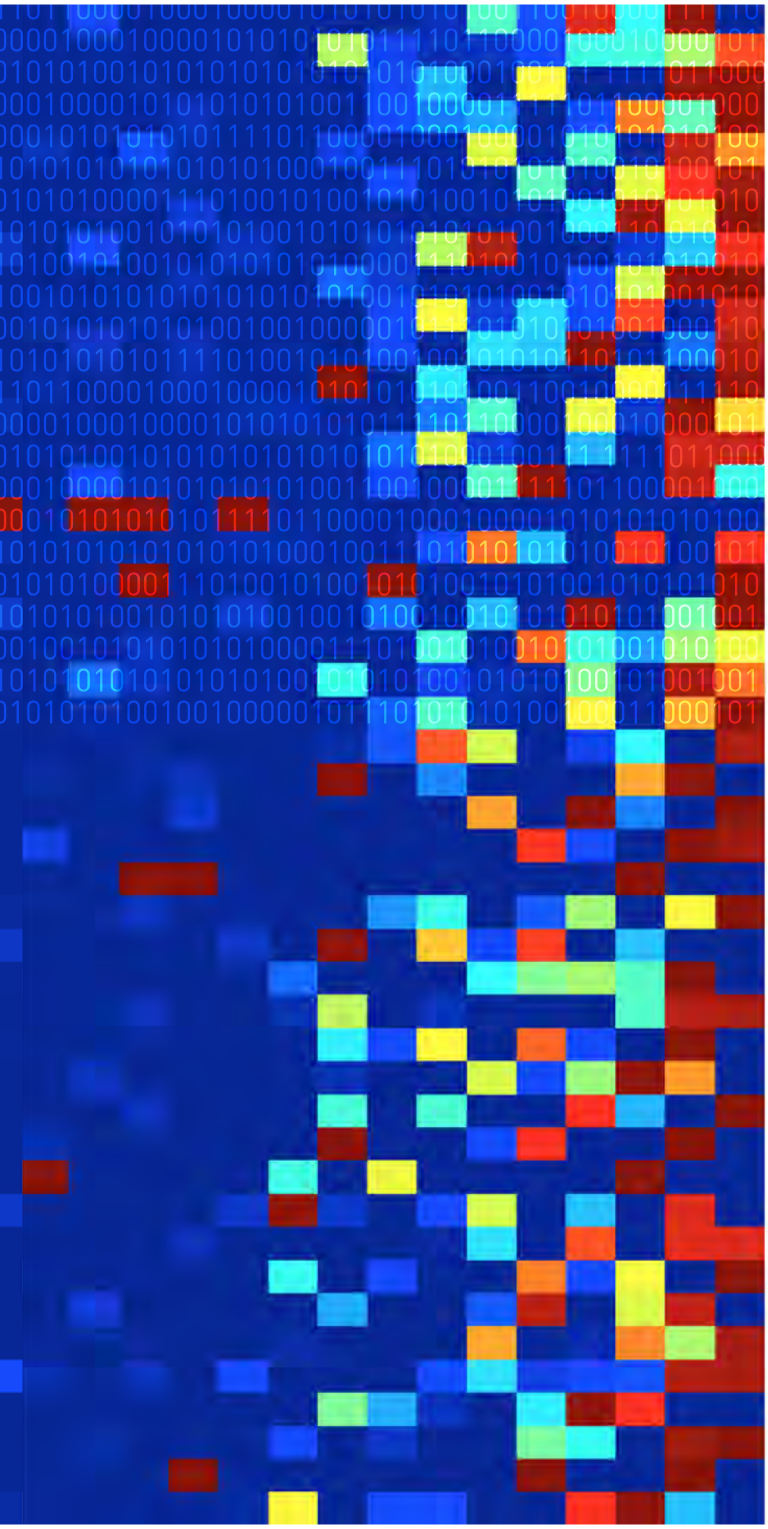
RESEARCH STRATEGY

BigInsight's research is organized in five innovation objectives. Each innovation objectives (IOs) is centered on a concrete innovation area: marketing, health, fraud, sensor, power. Most partners join more than one IO. Each IO has specific innovation aims related to outstanding open problems, which we believe can specifically be solved with new statistical, mathematical and machine learning methodologies. Our research projects deliver methods and tools for their solution. We aim to new, interesting and surprising solutions, which take the field and our partners ahead in their innovation agenda. Final transfer to partners' operations will happen both within and on the side of BigInsight.

INNOVATION OBJECTIVES	Personalised marketing	Personalised health and patient safety	Personalised fraud detection	Sensor systems	Forecasting power systems
INNOVATION PARTNERS	DNB Gjensidige NAV Skatteetaten Telenor SSB	DNV-GL Kreftregisteret OUS Telenor	DNB Gjensidige Skatteetaten	ABB DNV-GL SSB	DNV-GL Hydro Energy SSB
RESEARCH PARTNERS	NR UiO NIPH UiB	UiO OUS NR UiB NIPH	NR UiO UiB	NR UiO UiB	NR UiO UiB
PRINCIPAL INVESTIGATORS					
Principal Investigators:	Kjersti Aas	Magne Thoresen	Anders Løland	Ingrid Glad	Alex Lenkoski
co-Principal Investigators:	Arnoldo Frigessi	Clara Cecilie Günther	Ingrid Hobæk Haff	Magne Aldrin	Carlo Mannino

METHODS

We solve the innovation challenges of our partners by developing solutions which are based on new statistical, mathematical and machine learning methods.



Sampling bias and missing values take new dimensions
Data can be collected for one first purpose (say, billing) and then used in a different context (marketing): bias must be corrected. Data composed by several collections or collected in different periods can be inconsistent, something which has to be resolved. The unsolicited production of data on a volunteer basis can be utilised only if properly calibrated. More generally, data are often fragmented, with individuals entering and leaving cohorts and surveys at different time points and for different reasons, generating informative missingness. An arduous situation is encountered when data are only temporarily available, and are then deleted (because of regulations), or arrive in data streams, requiring the collection of appropriate statistics for future use.

There are new possibilities in using all data
Traditionally, data quality is strongly advocated, so that "bad" data might even be best omitted. For data rich areas this approach is too conservative, and we will design stochastic models to explore the information hidden in the complete mosaic of data. By integrating more data layers, we will be able to make more precise inference, combination of many weak signals and complex interactions. Bayesian borrowing of strength can compensate for data quality or uneven quantity. It also allows the rigorous integration of data with substantive knowledge, both hard facts as constraints and soft expertise as elicited prior models.

High frequency time series data allow intervention in real time
Sensors, machines, patients and customers generate very high dimensional time series which are analysed for motif discovery, anomaly detection and classification. The purpose is to automatically alert about shifts in trends, variability or extremes, about changing patterns of behaviour or about any potential deviations from the norm. Most time series capture the system in a transient phase and are therefore not stationary and with strong dependences. Sensor data, in industry and healthcare, have a time resolution which can be adaptively increased when anomalies appear as forthcoming or reduced in normal periods. In order to reduce false alarms, methods may effectively evoke sparsity assumptions. Measuring reactions to interventions, allows the design of new experimental plans, where groups of individuals are exposed to new situations (products, therapies), and observed in their transient phase.

Dimension reduction

Both the dimension of the data and of the unknown parameters space can be huge. In the first case, one might need to subsample the data or to distribute them to computers working in parallel, asynchronously. In the second case, one need to determine a reduced model which still allows efficient prediction. Penalised likelihood based approaches, or based on approximate models, but also boosting and other machine learning approaches are very useful.

Rapid computation and model approximation

Real time computations permit real time interventions. This is difficult, when large quantities of data need to be analysed or when the space of solutions is huge. Trading model accuracy and inferential precision for efficiency becomes imperative. Parallel computing, asynchronous and synchronous, on varying architectures, open new possibilities for big data analysis. We experiment with various solutions and are particularly interested in subdividing covariates (not samples) between machines in a cluster. Pseudolikelihood approximations should play a new role as approximate models, as they are efficiently estimated. Approximate Bayesian Computation and its variations might scale with problem dimension.

Anomaly detection and changepoint prediction allows control

Measures of surprise quantify the level of incompatibility of data with a given model, without any reference to alternatives. Surprise plays an important role in dynamic situations, where the reference is the past trajectory. There are connections to outlier theory; measures of Bayesian discrepancies between priors and posteriors; change point detection and we will need to extend and adapt these ideas to highly multivariate and non-stationary time series. A further aim is to develop models that allow change point prediction, rather than merely locating these after occurrence. We shall develop procedures to monitor parallel streams of data to detect anomalies and change-points.

Network based decision theory

Stochastic and dynamic networks appear in many innovation objectives, linking units by similarity, proximity or contact. The study of the mechanisms governing network growth is important and allows prediction and personalised intervention. We wish to develop a decision theory that exploits network structure, whenever decisions are taken according to local latent communities.

Large Scale Optimisation

Discrete and continuous optimisations are central tasks in decision making, management and inference. Highly efficient in the linear and convex case, optimisation becomes very hard in non-convex situations and in high dimensions. For example, many network based problems are large-scaled, NP-complete combinatorial optimisation problems which can only be tackled by suitable decomposition methods or efficient approximate algorithms. By using a mix of exact decomposition methods and heuristics, large optimization problems previously regarded as intractable are however now feasible.

Deep Learning

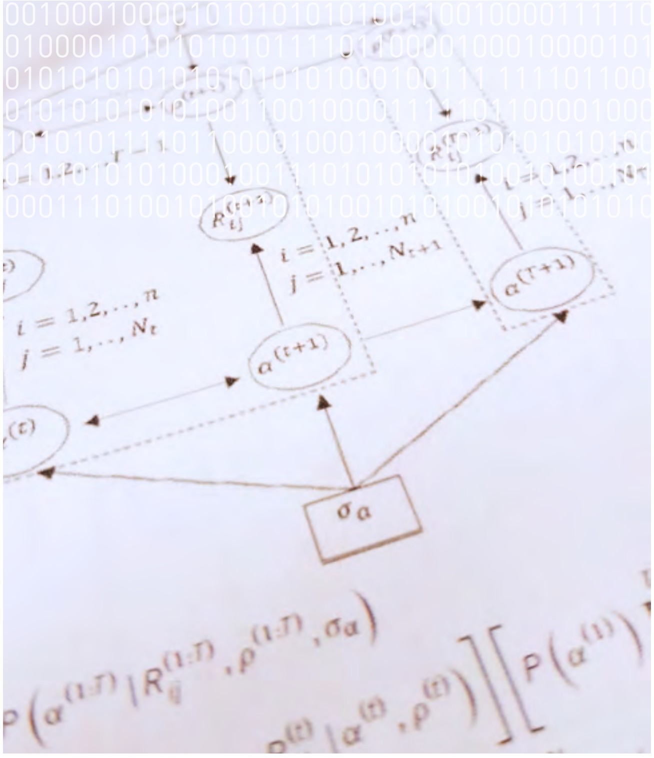
Deep learning allows precision classification in supervised problems, without the need to careful modelling. It allows optimising feature selection including interactions and non-linear effects. However, it requires very large amounts of labelled training data and computer power to train. We explore the use of deep learning in various situations, investigate power and scaling properties. We are also working on "opening the black box", in order to understand what interventions would be needed to change the classification of a user.

SCIENTIFIC ACTIVITIES

BigInsight researchers are working on several research projects, motivated by our partners' innovation needs. They cross disciplines and industrial sectors and challenge the available state-of-the-art. New methodology is developed and tested on specific innovation cases and data from the partners.

Each IO has a research team, with members from the relevant innovation and research partners. This includes senior and experienced staff as well as junior staff, postdocs, PhD and master students and international collaborators.

Two results from 2017 are highlighted:



An algorithm suggests that you should get a loan, because of good vibrations from your bank accounts

We have developed a deep learning approach for predicting mortgage default from consumer transaction data.

Deep learning is used as the description of the family of machine learning methods training different kinds of neural networks. In this work, we have used a type of deep neural networks denoted convolutional neural networks (CNNs). We think that we are the first to apply CNNs to consumers' account balances to predict mortage defaults. It is often hard for humans to figure out appropriate features for credit scoring. Using a CNN, the feature engineering may be automated in a systematic way, providing higher level abstractions of the low level time series signals. The obtained results are very promising. The low risk group increases from 80% with DNBs current model to 95% with the new model, enabling increased automatization of loan approvals. Moreover, the high risk group is better identified (and smaller), meaning that the manual resources may be focused on the more complex cases.

Kvamme, H., Sellereite, N., Aas, K. and Sjursen, S.: Credit scoring by deep learning on time series, accepted for publication in Expert Systems with Applications, February 2018.
https://www.sciencedirect.com/science/article/pii/S0957417418301179

Dagens Næringsliv, Aas, K., Sellereite, N., Kvamme, H. Roboten gir lån hvis den får gode vibrasjoner fra kontoen din.
https://www.dn.no/nyheter/2017/04/30/1817/Privatokonomi/roboten-gir-lan-hvis-den-far-gode-vibrasjoner-fra-kontoen-din

$$p_0 = g\left(\sum_{j=1}^{5} v_{j,0}\, f\left(\sum_{i=1}^{3} w_{i,j}\, x_i + b_j\right) + a_0\right)$$

$$p_1 = g\left(\sum_{j=1}^{5} v_{j,1}\, f\left(\sum_{i=1}^{3} w_{i,j}\, x_i + b_j\right) + a_1\right)$$

We need to be recommended: A personalized recommender system with uncertainty quantification

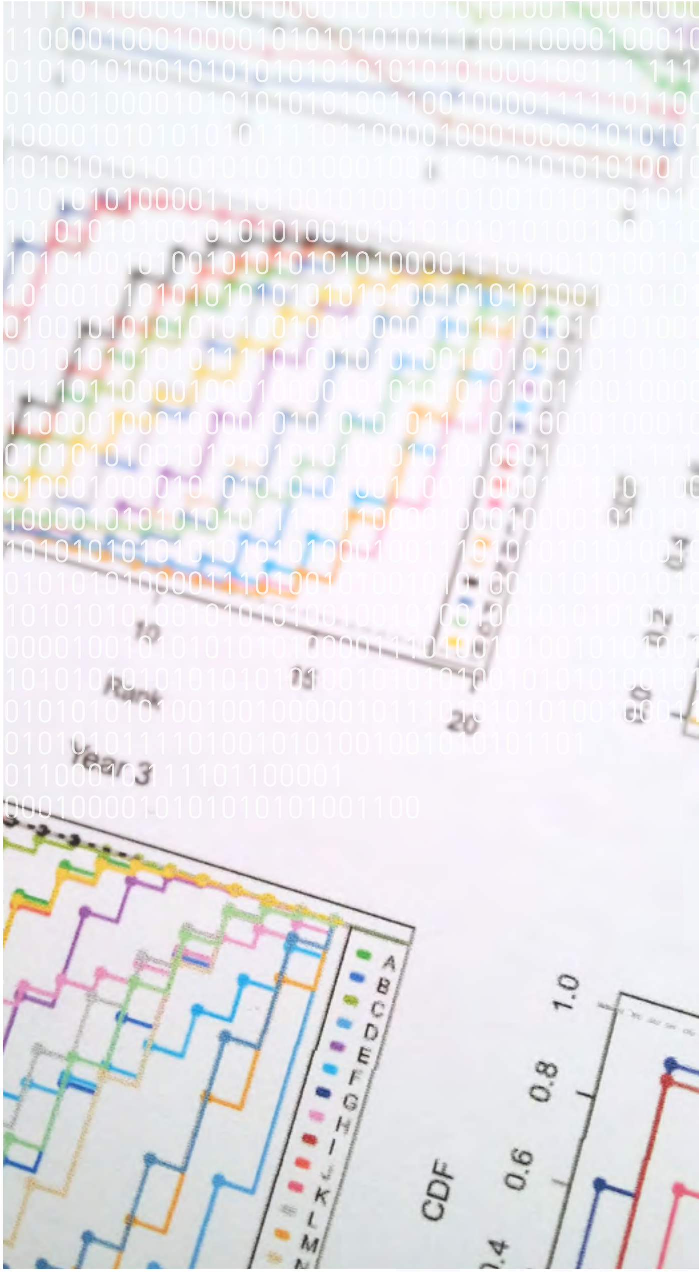
Users and customers are overloaded with choices. Sorting and searching is difficult and also boring. We need algorithms that do it automatically and personalized for us.

When using Netflix or Amazon, we get recommendations about movies or books we should watch or read. This is useful for finding out in the jungle of offers. The recommendations we get are based on what we have purchased or seen before; and what the other users have seen before. Collaborative filtering is the currently most used algorithm. But sometimes you get bad recommendations, When Netflix knows too little about what we prefer, because we are new to a webpage, because we change preferences, or because we share the same user within our families.

Our solution is a new statistical algorithm who better understands what we like, especially when not much is known about the preferences of a user. We developed a Bayesian model and a new algorithm which produces very good recommendations in test cases. It also quantifies how secure we are about the recommendations. If we are too unsure, it is probably best to wait instead than recommending inappropriately.

Vitelli V, Sørensen Ø, Crispino M, Frigessi A, Arjas E, Probabilistic preference learning with the Mallows rank model, accepted for publication in the Journal of Machine Learning Research, 2017.
https://arxiv.org/abs/1405.7945

Asfaw D, Vitelli V, Sørensen Ø, Arjas E, Frigessi A. Time varying rankings with the Bayesian Mallows model. Stat. 2017 Jan 1;6(1):14-30
http://onlinelibrary.wiley.com/doi/10.1002/sta4.132/full



$$
\begin{aligned}
& P\left(\alpha^{(1:T)} \mid R_{ij}^{(1:T)}, \rho^{(1:T)}, \sigma_\alpha\right) \\
\propto & \left[\prod_{t=1}^{T} P\left(R_1^{(t)}, \ldots, R_{N_t}^{(t)} \mid \alpha^{(t)}, \rho^{(t)}\right)\right]\left[P\left(\alpha^{(1)}\right) \prod_{t=2}^{T} P\left(\alpha^{(t)} \mid \alpha^{(t-1)}, \sigma_\alpha^2\right)\right] \\
= & \left[\prod_{t=1}^{T} \frac{1}{Z_n\left(\alpha^{(t)}\right)^{N_t}}\right] \exp\left[-\sum_{t=1}^{T} \frac{\alpha^{(t)}}{n} \sum_{j=1}^{N_t} d\left(R_j^{(t)}, \rho^{(t)}\right) - \sum_{t=2}^{T} \frac{1}{2\sigma_\alpha^2}\left(\alpha^{(t)} - \alpha^{(t-1)}\right)^2 - \lambda\alpha^{(0)}\right]\left[\prod_{t=1}^{T} 1_{\mathcal{P}_n}\left(\rho^{(t)}\right)\right].
\end{aligned}
$$

PERSONALISED MARKETING

We develop new methods, strategies and algorithms for individualised marketing, customer retention, optimised communication with users, personalised pricing and personalised recommendations or to maximise the probability of purchase of a product or other actions of the users. We exploit users' behavioural measurements in addition to their more standard characteristics and external data (including competitors' activity, market indicators, financial information, and geographic information). We exploit network topologies, informative missingness and temporal relations. A key point is to identify the actionable causes of customer behaviour.

What we did in 2017:

Stochastic customer growth dynamics

Understanding how networks of customers grow in time and topology is important. The Vipps transaction data may be viewed as a graph with users corresponding to the nodes and the financial transactions between the users defining the edges. With an advanced statistical model we have analysed the growth of this graph. Our experiments show that the intrinsic quality of the nodes plays an important part in the evolution of the network. This insight may be used to identify influential nodes for viral marketing. The approach will be published.

Predicting customer behavior from time series

Multiple time series, related to individual customers have been used to predict, as early as possible, whether the customer will pay back his loan or not. We have developed a new solution based on deep learning of time series, which gives excellent results on the DNB cases it has been tested on. The method has been transferred to DNB and has been accepted for publication in the journal Expert systems and Applications.

Bayesian methodology for recommender systems

In many important situations, we wish to recommend to each individual user or customer, the items she might be most interested in, or the ones she would benefit most. Starting from data where user either rate or compare items or click on them, we predict user's preferences of other items. We have invented a new Bayesian approach based on extensions of the Mallows model, which allows making individualized recommendations, equipped with a level of uncertainty. We are working on public data and NRK data and plan to analyse other industrial cases, including an AB testing experiment. The methodology has been published on the Journal of Machine Learning Research.

Stochastic models for early prediction of viral customer behavior on networks

Can we predict if a new service or product will be go viral in a market, or if it will be a flop? We look to markets where adopters are organised as nodes in a network, where links represent contacts which allow a user to "convince" a neighbor to adopt. We worked on a case from Telenor. We are able to predict, already after a few weeks, if the adoption process will spread virally (or not) on the network and how the adoptions will happen in the future. Our simulation based approach is generic and looks very promising.

Clustering of clickstream data

Web stream data are routinely collected to study how users browse the web or use a service. The ability to identify user behaviour patterns from such data may be very valuable for different businesses. It may help to produce better marketing strategies and a better user experience. We use model-based clustering to segment users based on web clickstream data from Skatteetaten and Gjensidige. Model-based clustering assumes that users' behaviours are generated by a set of probabilistic models and each model corresponds to a cluster.

Principal Investigator
Kjersti Aas

co-Principal Investigator
Arnoldo Frigessi

PERSONALISED HEALTH AND PATIENT SAFETY

The health system is producing data at an unrestrainable speed; data that can mean personalized therapy, patient safety, personalized cancer prognoses, better prevention and monitoring of epidemics. We show how such data can be exploited, with a series of innovative prototype projects.

What we did in 2017:

Personalized cancer statistics

National population based cancer registries publish survival statistics by cancer site, stage, gender and time period, using relative survival methods. As new clinical registries are established, more data on treatment and later events become available, in addition to information on comorbidity or income and educational level. To become more relevant both to the clinician and the patient, the survival statistics will be tailored to encompass more detailed information, in line with the tradition of risk prediction models. We have investigated a number of different methods for relative survival and formulated them in terms of differential equations. Based on such a formulation, we can compare the behavior of the methods. As a by-product, we are developing a new method for measuring the probability of being "statistically cured", i.e. having the same risk of disease as the general population.

Personalized cancer treatments

We have developed a new multivariate penalized regression method (IPF-tree-lasso) that improves prediction of drug sensitivity in large-scale screening experiments based on molecular characterization of cancer cell lines in two ways: (i) by a more efficient combination of several sources of molecular data using the Integrative lasso with Penalty Factors, which we extended to the IPF-elastic-net and (ii) by borrowing information from available data on similar drugs through a hierarchical (tree) structure in the penalty terms. In simulation studies and application to the Genomics of Drug Sensitivity in Cancer (GDSC) data, multivariate IPF methods outperformed others, and IPF-tree-lasso obtained the best prediction.

Different measures for synergistic (or antagonistic) effects in drug combinations exist based on different definitions of non-interaction. We propose a fully statistical approach to model drug interactions based on a flexible Bayesian regression model. This allows us to include relevant information into the model specification, as well as to perform prediction for quantities of interest (e.g., for predicting the effect of those combinations of drugs that have not yet been tested). Furthermore, we link the proposed model with standard methodologies in quantitative drug interaction studies, and explore their performance on both simulated and real-life datasets.

Healthcare safety management

We are developing a two-step lasso-type model for prediction of hospital acquired infections (HAI) based on time series data. The goal is to use high-dimensional data from health records and administrative databases, routinely acquired in hospitals and health institutions, to predict the occurrence of HAI, while at the same time being able to identify factors to intervene on, in order to prevent HAI from happening. Simulations indicate very interesting properties of the method. Subsequently, the method will be applied to data from Oslo University Hospital.

Network theory for health

We are interested in the spread of infectious diseases, with the goal of planning and prevention. We have developed a new model that allows the simulation of varying degree of population clusters within a country. The specific goal of this study is to understand the effect of restricting long distance travelling as a function of the level of clustering of the population, which can be seen as a level of urbanization of a country. In a different project, together with the Norwegian Institute of Public Health, we have investigated to what extent pain tolerance, as measured by three different tests, is correlated among friends in a social network. Our study exploits data from Tromsø on friendships among all high school students and their pain tolerance. We find a significant correlation between cold-pressor pain tolerance of an individual and the pain tolerance of the individual's friends.

Personalised cancer therapy guided by computer simulation

Mathematical modelling and simulation have emerged as a potentially powerful, time- and cost effective approach to personalised cancer treatment. In order to predict the effect of a therapeutic regimen for an individual patient, it is necessary to initialize and to parametrize the model so to mirror exactly this patient's tumor. We are developing a comprehensive approach to model and simulate a breast tumor treated by two different chemotherapies in combination or not. In the multiscale model we represent individual tumor and normal cells, with their cell cycle and others intracellular processes (depending on key molecular characteristics), the formation of blood vessels and their disruption, extracellular processes, as the diffusion of oxygen, drugs and important molecules (including VEGF which modulates vascular dynamics). The model is informed by data estimated from routinely acquired measurements of the patient's tumor, including histopathology, imaging, and molecular profiling. We implemented a computer system which simulates a cross-section of the tumor under a 12 weeks therapy regimen. We showed how the model is able to reproduce patients from a clinical trial of OUS, both responders and not. We show that other drug regimens might have led to a different outcome.

Principal Investigator
Magne Thoresen

co-Principal Investigator
Clara Cecilie Günther

PERSONALISED FRAUD DETECTION

Fraud is expensive, affects common resources and prices and is therefore important to detect and prevent. Soft fraud, the exaggeration of legitimate insurance claims, is quite diffuse and difficult to spot. A sustainable welfare system and efficient insurance operations require implementation of effective measures to limit fraud. Tax avoidance and tax evasion are other important types of fraud. We are also interested in money laundering detection. We develop adaptive tools that use "all data", including payment logs, relational networks and other available digital records, but under strict privacy protection regulations.

A further objective is to combine the multitude of fraud detection models in an optimal way, taking advantage of the strength of each predictor while blurring away weaknesses, and still obtaining coherent quantifications of the uncertainty in the fraud prediction. A similar objective is the development of new individualised anti-money laundering solutions. So far, the detection of suspicious transactions is based on labour-intensive semi-manual approaches and restricted to customers who significantly differ from the norm. Since the volume of banking transactions is steadily increasing, automated, intelligent tools are needed. The aim is to significantly increase the number of correctly identified money laundering transactions.

Fraud detection can be seen as a regression/forecasting problem, where fraud (true/false) is the response, possibly with a potential economic loss, and there are very many covariates. Including interactions, the number of covariates is huge. Generally, there are few fraud cases that are investigated, and a great number of undetected cases exist. The objective is to produce a trustworthy probability of fraud for each case.

What we did in 2017:

Ensemble methods for fraud detection

Combining results obtained by different statistical and machine learning procedures can be convenient. We constructed a toolbox for combining fraud forecasting models, exploiting the time series aspect of the data, available covariates and the probabilistic confidence in the classification obtained by each individual model. The toolbox has been tested on data from the Norwegian Tax Administration.

Combining dependent probability forecasts

Different fraud forecasting methods are however likely to be dependent. This is in particular the case if they are based on the same covariates. Therefore, ensemble methods which ignore such dependence will not be able to recognise the presence of the same forecast, just repeated several times. This will lead to spurious confidence in the ensemble forecast. The idea is to construct a joint model for the outcome (fraud/not fraud) and the forecasts, capturing dependences of these, based on a pair-copula construction. We developed the pair copula methodology in the last decade, and it is now used in many areas with great success. The new method will be tested on simulated and real data from our partners.

Network analysis for fraud detection

Fraud can spread directly or indirectly from one fraudster to another. Exploiting knowledge about social relations between users/customers can be useful to discover fraud. Understanding how such networks of users look and evolve over time is expected to significantly improve fraud detection models. We build these networks and extract useful characteristics to produce better fraud forecasts and provide additional insight into how fraud spreads. We are here working with insurance fraud data and later on tax avoidance and money laundering data.

A machine learning model for suspicious transactions

Most supervised anti money laundering methods assume that suspicious activities are labelled as such by experts, while legitimate activities are just randomly sampled from the complete population of activities. This is motivated by the fact that the chance of a random activity being suspicious is almost zero. We challenge this view by 1) modelling suspicious transactions directly instead of via accounts or parties, and 2) show that the current practise of excluding activities labelled as non-suspicious by experts leads to significantly worse performance. The method is being transferred to DNB and the approach will be published.

Local Gaussian discrimination with discrete and continuous variables

We generalise classical discriminant analysis (LDA and QDA) to local-Gaussian class distributions instead of regular Gaussian distributions. This lifts the variable dependence from globally pairwise to locally pairwise dependence. We are also able to combine discrete and categorical variables with the continuous variables by relying on pairwise dependence in a unified framework. The method will be evaluated on simulated and real data from one or several of the partners and the approach will be published.

Principal Investigator
Anders Løland

co-Principal Investigator
Ingrid Hobæk Haff

SENSOR SYSTEMS

Maintenance and inspections of ships are traditionally based on a preventive scheme where components have been maintained according to a time schedule. This approach is based on the assumption that a component has a defined lifetime, after which its failure rate increases. However, estimates of lifetime have large uncertainties and a large percentage of failures are not age-related and are therefore not adequately addressed in this way. We develop new approaches based on the recent availability of large arrays of sensors, which monitor condition and operation of machinery and power systems. Sensor data are becoming available for the first time on global ship fleets, with efficient broadband connectivity to shore. We suggest new approaches to condition monitoring, which is the process of identifying changes in sensor data that are indicative of a developing anomaly or fault. In addition to using previous failure data and pattern recognition techniques to detect anomalies, we test model based approaches that exploit knowledge on the sensors and the conditions they assess. For the design of sensor monitoring systems, a key challenge is to determine the level of resolution in time and sensor density needed to have a precise dynamic picture of the actual health of the system. Borrowing strength across sensors and ships in a fleet is an important aspect, leading to increased safety of a whole fleet.

What we did in 2017:

Overview of data driven statistical methods for condition monitoring systems

With focus on data-driven methods, we have reviewed statistical methods specifically relevant to condition monitoring of ship machinery systems. A number of statistical methods and approaches which are relevant for diagnostics and prognostics of ship machinery systems based on sensor data have been identified and studied. This study has resulted in big insight and a paper published in the International Journal of Condition Monitoring in 2018.

Time-efficient on-line anomaly detection methodology for the maritime industry

Because the method based on AAKR and sequential testing for maritime sensor data that we published in 2016, is highly computer intensive, we have developed and tested several modifications of the methodology in order to increase time-efficiency and still produce robust results. Especially, we have modified the way training data is stored and used in the reconstruction step, by clustering the training data in AAKR, which leads to significant speed-up. The improved one-line anomaly detection schemes have been tested on several multidimensional maritime sensor data sets. A conference paper is published in the Proceedings of the IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control. A journal paper is in revision for publication.

Cluster based anomaly detection

We have developed unsupervised anomaly detection methods based on various clustering techniques. The idea is to identify clusters in sensor data in normal operating conditions, and assess whether new data belong to any of these clusters. Applied to sensor data from a marine diesel engine, this class of methods turns out to provide an efficient, completely unsupervised initial screening of data streams for anomalies.

Anomaly detection using dynamical linear models

As an alternative to AAKR for signal reconstruction under normal operations, we have used a number of multivariate dynamic linear models for the reconstruction, and combined with sequential testing of residuals. A challenge to this type of modeling is the heterogeneous correlation structure in time of our maritime data series. A conference paper is published in the Proceedings of the Annual Conference of the Prognostics and Health Management Society. A journal paper is accepted for publication in International Journal of Condition Monitoring and Diagnostic Engineering Management.

Anomaly/fault prediction from ship operation log files

One of the condition monitoring systems in a ship sends messages regarding the operational mode of the ship at irregular time points. This log file works on a finite alphabet of possible events, and the purpose is to analyze this log file to detect sequences of events which appear to be abnormal or correspond to failures, preferably giving a warning early enough to be able to take action. We have attacked this problem in different ways, building on feature extraction, penalized variable selection, hidden Markov models and machine learning methods from speech recognition. The results are promising and two papers are in preparation.

Sequential detection of changes using dimension reduction techniques

We have developed methods for sequential change detection in high dimensional streaming data, with the aim of detecting changes in the distribution of the data as soon as possible, keeping false positives as a minimal level. We have extended sequential detection methods for changes in mean level to yield also changes in variance and covariance between streams. Such changes are impossible to detect with univariate monitoring methods, of course, but might represent important failure modes in a ship monitoring sensor system. Furthermore, we have studied properties of dimension reduction techniques such as sketching and PCA in connection to sequential change detection, assuming sparsity, that is, that changes happen only in a few of the monitored streams. A master thesis on this topic was finalized in May 2017, and a paper is in preparation.

Change point detection in maritime sensor data using PELT

We are investigating the use of the PELT-methods (Pruned Exact Linear Time, developed by our collaborators in Lancaster) for multiple change point detection in maritime sensor data. A master thesis will be finalized in 2018 and we will then understand if the approach is useful for early change point detection in the marine context.

Hull condition monitoring

We have been working with the construction of an emulator for the fatigue rate of a ship, given its physical characteristics and operating history, including environmental and weather data. The emulator is compared to calculations from a complex physical model run by DNV-GL. The aim is to be able to use the emulator as a fast and precise supplement to the full model. First analyses comparing results from the physical model and the emulator show that the correspondence is adequate in most conditions. Also in connection to hull condition monitoring, we have collected sensor data on bending moments from several vessels together with their history. The aim is to evaluate the precision of an existing physical model for these quantities. First analyses comparing sensed bending and calculated bending for the same ships show that the correspondence is satisfactory in certain conditions, but not always.

Surprise detection for a motor cooling system

We study situations where interest lies in a multivariate response variable to be monitored in time and prone to anomalies which can be related to a set of covariate time series. Regression methods allow studying the residuals and investigating surprises appearing in these residuals. This approach has been applied to the case of a motor cooling system with good results. We are transferring the method to the partner.

Principal Investigator
Ingrid Glad

co-Principal Investigator
Magne Aldrin

FORECASTING POWER SYSTEMS

Electricity producers rely on forecasts of electricity prices for bidding in the markets and power plant scheduling. Markets are changing: A much tighter integration between European markets and a rise in unregulated renewable energy production, especially wind and photo-voltaic, call for joint probabilistic forecasts. Incorporating the transient interplay between productions from renewable sources is critical to power production and financial operations. Multivariate probabilistic forecasts of electricity prices in the short horizon are required.

Appropriately characterising multivariate uncertainty will enable more effective operational decisions to be made.

Conventional power grids add extra generation and distribution capacity. Smart grids actively match energy supply and demand and combine the needs of the markets with the limitations of the grid infrastructure. With the implementation of smart meters and grid sensors, enormous amounts of time series data are generated, with seconds resolution. Our objective is to develop new methods that extract the right information from data to optimise grid control and for real time operation.

What we did in 2017:

Error dressing spot price forecasts

Electricity markets "spike" and "crash" when volumes are respectively slightly higher or lower than typical and these extreme price swings make uncertainty quantification a critical part of forecasts. However, the limited degree to which these extremes are observed makes such constructions difficult. We have developed a system using published bid/ask curves that determine the final price, to construct realistic distributional price forecasts that embed this extreme behaviour. We employ the concept of "error dressing" by using the curves to translate residual behaviour of market volume forecasts into price uncertainties.

This system has proven highly useful for our industrial partner Hydro and is now already embedded in their nightly production system. The output of this model informs operational considerations both on the production and the trading sides of NHY's operations. An academic article outlining the approach and highlighting some of the results has been drafted and is nearly ready for submission to an applied statistics journal.

High-Dimensional demand prediction via principal component analysis

The approach that underlies our error dressing methodology depends on point forecasts of demand and supply and improves as the quality of the underlying models increase. Some of the underlying models that are currently used in practice are decades old. We therefore took this occasion to revisit these models and investigate whether they too could use a reconsideration.

In particular, we found that the model for predicting electricity demand, which is currently based on neural networks, was in need of improvement. Electricity demand in the Nordic countries is primarily forecasted by projected weather conditions, which are a highly correlated, both across the Nordic region and across time. We conducted a preliminary investigation and learned that models based on principal component analysis (PCA) yield a 30% improvement in demand forecasts over the current model in use at Hydro.

The size of this improvement has led us to move quickly towards implementing the new model for Hydro, in tandem with writing an academic paper that explores the use of PCA to filter weather forecasts in the demand prediction problem. Our early 2018 plans will include a roll out of this methodology for Hydro and a draft paper that discusses the results of this investigation.

Research into Vintage Adjustment of Forecast Trajectories

Our work in the power system domain is high-dimensional both in the input space (factors used to make predictions) as well as the output space (the number of quantities simultaneously forecasted is also large). It is frequently the case that a set of forecasts, all issued at the same time, have different timeframes for resolution. Likewise, each observed outcome can be associated with a large number of different forecasts, each issued on a different date.

We call the set of forecasts issued on the same date a vintage. When the members of a vintage have different resolution dates, it was our belief that performance early in the vintage could help adjust later forecasts. Preliminary research into this topic, conducted in 2017, has proven highly promising. In particular, we found that taking a two-day ahead forecast of electricity demand and adjusting for the performance of the one-day ahead forecast from the same vintage outperforms a newly issued one-day ahead forecast for the same quantity by a considerable degree.

The potential for use of vintage adjustment is vast and seemingly relatively unexplored. We plan on continuing the research into this general methodology and creating a high quality general purpose implementation of the method in R. While there are a great number of applications of vintage adjustment in power systems, the scope of this methodology is far reaching and we plan on expanding our considerations to applications in meteorology, finance and natural language processing.

Optimisation based curve disaggregation methods for regional bid/ask curve construction

At present, the bid/ask curves published by Nordpool are aggregated across regions. While this is helpful in modeling the overall system price of the Nordpool region, operations are ultimately conducted on the regional level. However, regional level bid/ask curves are not published. Thus, a methodology that was capable of using market level bid/ask curves to construct their regional level equivalents would be highly valuable. This methodology corresponds to an assignment problem where each point on the market-level bid/ask curve is assigned to a given region.

Fortunately, regional price and volume information is published and it is possible to verify whether a given regional assignment yields clearing prices and volumes that roughly match the observed regional price and volume distribution. Thus, we have begun constructing a constrained optimization system that performs assignments of points in market-level bid/ask curves to the regional level.

This is a massive-dimensional optimization problem with an extremely large number of potential solutions. We have begun researching various methods that couple integer-programming routines with market simulators and statistical estimation procedures. A successful methodology here will have a substantial impact on Hydro's ability to conduct fine-tuned and localized price forecasts.

Principal Investigator
Alex Lenkoski

co-Principal Investigator
Carlo Mannin

INTERNATIONAL COOPERATION

International Academic Partners

International Academic Partners contribute to place BigInsight in the center of the global data science community. We collaborate in research and co-supervise PhD students. We organize joint workshops and events.

STOR-i, Statistics and Operational Research in partnership with Industry,
is a joint venture between the Departments of Mathematics & Statistics and Management Science of the University of Lancaster. STOR-i offers a unique interdisciplinary PhD programme developed and delivered with important UK industrial partners. The centre is at the forefront of international research effort in statistics and operation research, establishing an enviable track record of theoretical innovation arising from real world challenges. Professors Jonathan Tawn, professor Idris Eckley (who co-lead the centre) and professor Paul Fearnhead co-supervise PhD students together with BigInsight staff, on particle filters, multivariate extremes, non-parametric isotonic spatial regression, Bayesian modelling, multivariate sensor data. BigInsight and STOR-i exchange membership in each other's scientific advisory boards.

Professors Idris Eckley,
Jonathan Tawn and
Kevin Glazebrook,
leading STOR-i at
University of Lancaster

STOR-i
excellence with impact

The Medical Research Council Biostatistics Unit (BSU)
is part of the University of Cambridge, School of Clinical Medicine. It is a major centre for research, training and knowledge transfer, with a mission 'to advance biomedical science and human health through the development, application and dissemination of statistical methods'. BSU's critical mass of methodological, applied and computational expertise provides a unique environment of cutting edge biostatistics, striking a balance between statistical innovation, dissemination of methodology and engagement with biomedical and public health priorities. Professor Sylvia Richardson is director of the BSU and she has received an honorary degree of the University of Oslo. BigInsight and the BSU have several joint projects in health and molecular biology. We have also involved the unit in our collaboration with the University of Hawassa (Ethiopia).

Professor Sylvia Richardson,
MRC Biostatistics Unit,
Cambridge

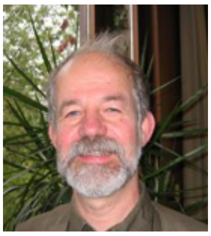
MRC Biostatistics Unit

International guest programme

BigInsight has an international guest programme, which includes all from short visits to long term visiting and adjunct positions and a sabbatical visitor programme.

In 2017 we hosted the following longer visits:

Professor **Gianpaolo Scalia Tomba**, University of Roma Tor Vergata, visits Oslo regularly in collaboration with NIPH. He is interested in models for infectious diseases and antibiotics resistance.



Professor Emeritus **Elja Arjas**, University of Helsinki, has a 20% adjunct position at BigInsight and collaborates in projects on recommender systems and health.

Professor **Per Mykland**, University of Chicago, spent a sabbatical year (2016-2017) at BigInsight. He has been working on multivariate time series models for high frequency sensor data.

Professor **Lan Zhang**, University of Illinois at Chicago, spent a sabbatical year (2016-2017) at BigInsight. She has been collaborating with the Sensor team.

Postdoc **Henri Pesonen**, University of Helsinki, spent several months at BigInsight, working on personalized cancer therapy simulation.

International training programme

PhD students from other universities spent periods of training and research collaboration at BigInsight.

In 2017 we welcomed:

Marta Crispino, University Bocconi, Milano (Recommendation systems)

Derbachew Asfaw, Hawassa University (Recommender systems)

Emma Simpson, STORi Lancaster University, (Multivariate Extreme Value Theory for Vines and Graphical Models)

Marta Crispino

First PhD in Statistics in Ethiopia

BigInsight is a partner of the Norwegian Programme for Capacity Development in Higher Education and Research for Development (NORHED) project at Hawassa University (Ethiopia), together with NTNU.

In June 2017, for the first time ever in Ethiopia, a PhD degree in Statistics has been awarded by an Ethiopian university. Denekew Belay, Markos Erango, Negussie Yohannes defended their theses with success. All three students had spent at least a year in Oslo and were part of the BigInsight student community. Magne Aldrin and Arnoldo Frigessi were among the supervisors. We are proud to have been able to contribute to statistical capacity building in Ethiopia.

https://www.norad.no/en/front/funding/norhed/projects/hawassa-university--phd-programme-in-mathematical-and-statistical-sciences/

International Programmes and Funding

BigInsight is partner of the COST Action CA15109 "European Cooperation for Statistics of Network Data Science (COSTNET)". Professor Arnoldo Frigessi is a member in the Management Committee and professor Birgitte Freiesleben de Blasio is nominated as deputy. This EU action started in 2016 and aims to facilitate interaction between diverse groups of statistical network modellers, establishing a large and vibrant interconnected and inclusive community of network scientists. The second workshop took place in Spain.
http://www.cost.eu/COST_Actions/ca/CA15109

From left to right: Markos Abiso, Dechassa Obsi, Tadele Tesfa, Denekew Bitew, Arnoldo Frigessi, Negussie Yohannes

Network of Big Data Centers of Excellence in Europe

Big Data National Centers of Excellence in Europe join forces for better networking and collaboration. BigInsight is part of the consortium. The main focus of the network is research itself and how research can be transferred into relevant industries. Current activities include: collect best practices and key achievements of each center, define big challenges, and align with other European initiatives.

Scientific Advisory Committee of BigInsight

Prof. Idris Eckley

Prof. Samuel Kaski

Prof. Geoff Nicholls

Prof. Marina Vannucci

Senior Lecturer Veronica Vinciotti

Scientific Advisory Committee of BigInsight

Scientific Advisory Committee of BigInsight has five international members. The next meeting will be in autumn 2018.

Prof. Idris Eckley, Lancaster University, UK

- Until 2007 Statistical Consulant at Shell Global Solutions
- Co-Director of the EPSRC-funded STOR-i Centre for Doctoral Training
- Within STOR-i he leads the Centre's industrially-engaged research activity
- Co-Director of the Data Science Institute DSI@Lancaster: Lancaster's new world-class, multidisciplinary Data Science Institute.
- Leads the EPSRC programme StatScale: Statistical Scalability for Streaming Data

Prof. Samuel Kaski, University of Helsinki, Finland

- Professor of Computer Science, Aalto University
- Director, Finnish Centre of Excellence in Computational Inference Research COIN, Aalto University and University of Helsinki
- Academy Professor (research professor), 2016-2020
- Statistical machine learning and probabilistic modeling

Prof. Geoff Nicholls, University of Oxford, UK

- Professor in Statistics and Head of Department of Statistics
- PhD in particle physics in the Department of Applied Mathematics and Theoretical Physics in Cambridge, University of Auckland in New Zealand
- Bayesian inference, Computational Statistics, Statistical Genetics, Geoscience, Linguistics and Archaeology

Prof. Marina Vannucci, Rice U, Houston, USA

- Professor and Chair of the Department of Statistics
- Adjunct faculty member of the UT M.D. Anderson Cancer Center
- Rice Director of the Inter-institutional Graduate Program in Biostatistics
- Honorary appointment at the University of Liverpool, UK
- NSF CAREER award in 2001
- Editor-in-Chief for the journal Bayesian Analysis

Senior Lecturer Veronica Vinciotti, Brunel U of London, UK

- Senior Lecturer in Statistics, Department of Mathematics, Brunel University
- Ph.D in Statistics, Imperial College, London
- Research in statistical classification methods in credit scoring and in statistical genomics
- Co-director of the European Cooperation for Statistics of Network Data Science

ACTIVITIES AND EVENTS

2017 BigInsight Workshop

The yearly BigInsight Workshop was held on Tuesday 31st October at Skatteetaten, Oslo. More than hundred researchers and innovators from all BigInsight partners participated to the workshop, during which a wide selection of projects were presented.

The day ended with a public chat among friends. On the scene four BigInsight PhD students discussed, while having a drink, their current lives, their expectations, their worries and wishes. Great success with Solveig, Sylvia Qinghua, Martin and Emanuele, from China, Italy and Norway.

TRAINING AND COURSES

New master programme in data science at UiO
The University of Oslo launches a completely new, unique and timely master program in Data Science in 2018.
"We welcome the first generation of students to a future-oriented study, which is the first of its kind in Norway, and which has many common features with Data Science at Harvard, for example" says the head of the program Geir Storvik from the Department of Mathematics at UiO.

In the era of digitization, this master program will educate specialists with very valuable expertise at the intersection of statistics and computer science, specialized in handling and analyzing large and complex data sets.

More than just being able to use different tools for big data analytics, our students will have a solid methodological foundation and possess a deeper understanding of the methods and algorithms. Skills in statistics, machine learning and data structures are absolutely fundamental for business and government to be able to utilize their data fully and correctly, and to develop new methods that provide competitive advantage and new opportunities. Ethical issues, privacy and data security are also important elements of the new study.

BigInsight will provide projects and master student supervision.

http://www.uio.no/english/studies/programmes/datascience-master/index.html

illustration: uio.no

COMMUNICATION AND DISSEMINATION ACTIVITIES

Website

The website of the center is biginsight.no. In 2017 we have redesigned the pages.

Seminars

BigInsight co-organises the traditional Tuesday Statistical Seminar (at the Department of Mathematics) and the Thursday Biostatistics Seminar (at OCBE).

From 2017 we also join forces with the Centre for Molecular Medicine Norway (NCMM) and launched the seminar series Sven Furberg Seminars in Bioinformatics and Statistical Genomics. The monthly seminars are joint bioinformatics-biostatistics catalyst events promoting scientific excellence and triggering collaborations on computational and statistical research projects related to molecular biology. The seminars are organized in three parts. First, a PhD student briefly presents their research. Second, the guest speaker gives a lecture on computational and/or statistical methods applied to molecular biology and medicine. Third, the audience gathers around pizza and refreshments. As part of the events, invited guest speakers meet local PIs and trainees.

BigInsight Wednesday Lunch Seminars are taken place every second week, alternating between the lunch room at NR and the eight floor of the Department of Mathematics. While we share a good lunch, we listen to an invited lecture. Our speakers help us to understand global trends of data science developments of statistics, machine learning, operations research, optimisation, computer science and mathematics in the era of high dimensional data

BigInsight in the media

forskning.no, 19.01.2017 "Forskere vil se i sykejournaler for å finne svindlere" by Anders Løland

Dagens Næringsliv, 30.04.2017 "Roboten gir lån hvis den får gode vibrasjoner fra kontoen din" Kjersti Aas, Nikolai Sellereite, Håvard Kvamme. An algorithm suggests that you should get a loan, because of good vibrations from your bank accounts.

Teknologiradet.no, 05.05.2017 "Fire punkt du må ha på plass før du set i gang med kunstig intelligens" Robindra Prabhu

Dagens Næringsliv, 28.07.2017 "Den svarte boksen kan åpnes" Anders Løland

Klassekampen, 15.12.2017 "Smake sin eigen medisin" interview with Arnoldo Frigessi

DN Dagens Næringsliv

Tron Even Skyberg, divisjonsdirektør og leder for risikokvantifisering i DNB håper å få god nytte av forskningen til Steffen Sjursen, prosjektleder for risikomodellering i DNB, Nikolai Sellereite, forsker Norsk Regnesentral, Håvard Kvamme, stipendiat UiO, og Kjersti Aas, prosjektleder i Norsk Regnesentral. Foto: Gunnar Lier

Nyheter Privatøkonomi

Roboten gir lån hvis den får gode vibrasjoner fra kontoen din

KLASSEKAMPEN Venstresidas dagsavis

Smake sin eigen medisin

«I framtida har du kanskje ein chip med DNA-et ditt på»

BigInsight outreach presentations

EVENT / **ORGANISER**
AI & Robotics FINANCE / **Relevent**
GDPR for designere og utviklere: Innebygget personvern / **Netlife Design**
Geomatikkdagene 2017 / **GeoForum**
Avdelingsdag / **Statistisk Sentralbyrå**
Nordic Agenda / **Skatteetaten**
Kunstig intelligens – Overlater vi for mye til maskinene? / **Tekna Ung/Tekna Big Data**
Beslutstödsdagen 2017 / **Copperberg Stockholm**
Innovation@altinn / **Altinn**
Stordata og offentlige tjenester / **Kommunal- og moderniseringsdepartementet**
Partnermøte Senter for Forskningsdrevet Innovasjon Foods of Norway / **NMBU**
Årsmøte for faglige medarbeidere / **Tidsskriftet for Den norske legeforening**
Internasjonale personverndagen frokostseminar / **Datatilsynet og Teknologirådet**
DNB Business Intelligence Day 2017 / **DNB**
Analyseforum / **SpareBank 1 Forsikring**
Faglig møte / **OUS Intervensjonssenter**
Seminar om cancer modelling – Livermore & Oslo / **Kreftregisteret**
Faglig Seminar / **Folkehelsa**
Høring / **Statistikklovutvalget 2017**
Seminar / **Datatilsynet**
Smart Analytics/Big Data-seminar/ **Aktuarforeningen**
Big Data Analytics for ingeniør- og realfag/ **TEKNA**
Frokostseminar i "Customer Analytics" / **BearingPoint**
Workshop on Social Security Fraud – Analytical Insights and Data Mining Approaches / **Udbetaling Danmark**
Yearly congress, invited talk / **European Network for Business and Industrial Statistics**
Åpent møte om lærende maskiner i offentlig sector / **Teknologirådet**
Faglig møte / **Bigmed NFR prosjekt**

RECRUITMENT

BigInsight's partners recruit researchers, postdocs, PhD students, Master students and summerstudents, in order to staff our projects. This happens with funding both from BigInsight and associated projects.

Those who started in 2017 were:

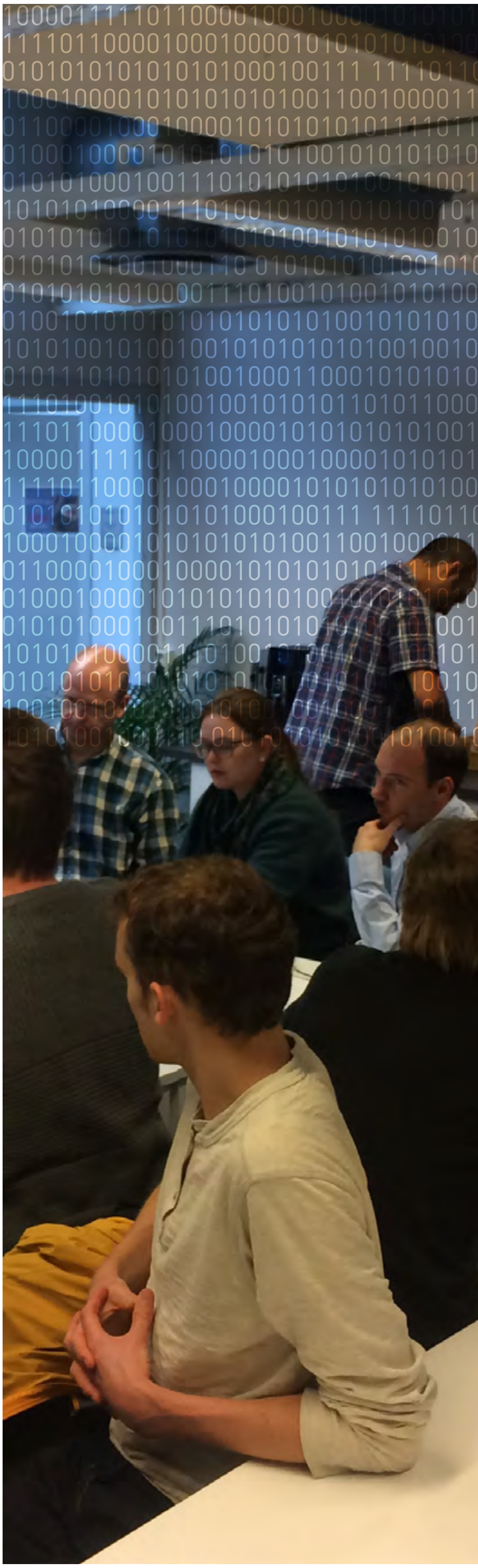
NAME	POSITION	FUNDING	RESEARCH AREA	AFFILIATION
Marta Crispino	Postdoc	BigInsight	Marketing	UiO, OCBE
Valeria Vitelli	Postdoc	UiO	Health, Marketing	UiO, OCBE
Christoffer Haug Laache	PhD student	BigInsight	Health	UiO, Cancer Registry
Daniel Piacek	PhD student	BigInsight	Fraud	UiO, Dep. Mathematics
Martin Tveten	PhD student	BigInsight	Sensor	UiO, Dep. Mathematics
Kristin Bakka	Master student	UiO	Sensor	UiO, Dep. Mathematics
David Swanson	Researcher	OUS	Health	OUS, OCBE
Håkon Taskèn	Summerstudent	BigInsight	Health	UiO, OCBE
Jenine Gaspar Corrales	Master	UiO	Marketing	Uio, Dep. Mathematics

PERSONNEL

Personnel affiliated with BigInsight

NAME	INSTITUTION	MAIN RESEARCH AREA
Rune Braastad	ABB	Sensor
Stian Braastad	ABB	Sensor
Børre Gundersen	ABB	Sensor
Petter Hausler	ABB	Sensor
Kenneth Nakken	ABB	Sensor
Jaroslaw Novak	ABB	Sensor
Morten Stakkeland	ABB	Sensor
Stian Torkildsen	ABB	Sensor
Andree Underthun	ABB	Sensor
Frank Wendt	ABB	Sensor
Bettina Kulle Andreassen	CRN	Health
Bjørn Møller	CRN	Health
Jan Nygård	CRN	Health
Hege Bolsø	DNB	Marketing
Thor Aage Dragsten	DNB	Marketing
Heidi Fredriksen	DNB	Marketing
Sven Haadem	DNB	Marketing
Katharina Henriksen	DNB	Marketing
Johannes Lorentzen	DNB	Fraud
Roy Oma	DNB	Fraud
Tron Even Skyberg	DNB	Marketing, Fraud

NAME	INSTITUTION	MAIN RESEARCH AREA
Mette S. J. Snilsberg	DNB	Marketing
Fredrik Strand	DNB	Fraud
Jørn Ødegård	DNB	Marketing
Geir Ånonsen	DNB	Fraud
Lars Holterud Aarsnes	DNV-GL	Sensor
Ole Christian Astrup	DNV-GL	Sensor
Håvard Nordtveit Austefjord	DNV-GL	Sensor
Theo Bosma	DNV-GL	Power
Ervin Bossanyi	DNV-GL	Power
Andreas Brandsæter	DNV-GL	Sensor
Christos Chryssakis	DNV-GL	Sensor
Frédéric Courivaud	DNV-GL	Health
Marcel Eijgelaar	DNV-GL	Power
Muhammad Jafar	DNV-GL	Power
Irena Jakopanec	DNV-GL	Health
Lars Landberg	DNV-GL	Power
Bobby Ray-Sannerud	DNV-GL	Health
Gaute Storhaug	DNV-GL	Sensor
Elizabeth Traiger	DNV-GL	Power
Bjørn-Johan Vartdal	DNV-GL	Sensor
Nikos Violaris	DNV-GL	Sensor
Randi Jule	Gjensidige	Fraud
Anders Nyberg	Gjensidige	Marketing
Marte Olstad	Gjensidige	Marketing
Thea Margrethe Skouen	Gjensidige	Marketing
Rune Sørensen	Gjensidige	Fraud
Geir Thomassen	Gjensidige	Fraud
Stefan Erath	Hydro	Power
Plamen Mavrodiev	Hydro	Power
Ellen Paaske	Hydro	Power
Håkon Otneim	NHH	Fraud
Birgitte De Blasio	NIPH	Health
Christopher S Nielsen	NIPH	Health
Ivar Areklett	NAV	Fraud
Eigil Johnsen	NAV	Fraud
Bjørn Atle Sundsback	NAV	Fraud
Morten Tholander	NAV	Fraud
Kjersti Aas	NR	Marketing
Magne Aldrin	NR	Sensor
Line Eikvil	NR	Fraud
Clara-Cecilie Günther	NR	Marketing, Health
Ola Haug	NR	Marketing, Sensor
Marion Haugen	NR	Marketing
Kristoffer Herland Hellton	NR	Marketing, Sensor
Lars Holden	NR	Health, Fraud
Marit Holden	NR	Health
Ragnar Bang Huseby	NR	Fraud, Power
Martin Jullum	NR	Fraud
Alex Lenkoski	NR	Power
Anders Løland	NR	Fraud, Power



NAME	INSTITUTION	MAIN RESEARCH AREA
Linda R. Neef	NR	Fraud
Hanne Rognebakke	NR	Marketing, Sensor
Nikolai Sellereite	NR	Marketing
Gunnhildur Steinbakk	NR	Fraud, Sensor, Power
André Teigland	NR	Marketing, Fraud, Power
Dag Tjøstheim	NR	Fraud
Ingunn Fride Tvete	NR	Health
Tor Arne Øigård	NR	Fraud, Power
Mette Langaas	NR/NTNU	Sensor, Health
Torsten Eken	OUS	Health
Lars Åke Hall	OUS	Health
Eivind Hovig	OUS	Health
Vessela Kristensen	OUS	Health
Marissa LeBlanc	OUS	Health
Egil Lingaas	OUS	Health
Sygve Nakken	OUS	Health
Andrew Reiner	OUS	Marketing, Health
Kjetil Sunde	OUS	Health
David Swanson	OUS	Health
Harald Weedon-Fekjær	OUS	Health
Anders Berset	Skatteetaten	Marketing, Fraud
Wenche Celiussen	Skatteetaten	Marketing, Fraud
Øystein Olsen	Skatteetaten	Marketing
Nils Gaute Voll	Skatteetaten	Marketing, Fraud
Fatima Yusuf	Skatteetaten	Marketing
Anders Holmberg	SSB	Marketing, Sensor, Power
Øystein Langsrud	SSB	Marketing, Sensor, Power
Jørn Leonhardsen	SSB	Marketing, Sensor, Power
Kenth Engo-Monsen	Telenor	Marketing, Health
Geoffrey Canright	Telenor	Marketing
Gro Nilsen	Telenor	Marketing
Astrid Undheim	Telenor	Marketing
Bård Støve	UiB	Fraud
Elja Arjas	UiO	Marketing, Health
Ørnulf Borgan	UiO	Marketing
Jukka Corander	UiO	Health
Riccardo de Bin	UiO	Fraud, Health
Ingrid K. Glad	UiO	Sensor
Ingrid Hobæk Haff	UiO	Fraud
Nils Lid Hjort	UiO	Fraud, Sensor
Vinnie Ko	UiO	Fraud
Carlo Mannino	UiO	Power
Kjetil Røysland	UiO	Health
Geir Kjetil Sandve	UiO	Health
Ida Scheel	UiO	Marketing
Geir Storvik	UiO	Sensor
Magne Thoresen	UiO	Health
Marit Veierød	UiO	Health
Manuela Zucknick	UiO	Health
Erik Vanem	UiO/DNV-GL	Sensor, Power
Arnoldo Frigessi	UiO/OUS/NR	Marketing, Health, Sensor

NAME	FUNDING	NATIONALITY	PERIOD	GENDER	TOPIC
Postdoctoral researchers with financial support from BigInsight					
Marta Crispino		Italy	2017-2018	F	Marketing
Alvaro Köhn Luque		Spain	2016-2020	M	Health
Postdoctoral researchers with financial support from other sources					
Andrea Cremaschi	UiO/NCMM	Italy	2016-2018	M	Health
Gudmund Hermansen	UiO	Norway	2016-2018	M	Sensor
Christian Page	OUS/HSØ	Norway	2016-2018	M	Health
Owen Thomas	UiO	UK	2017-2019	M	Health
Valeria Vitelli	UiO	Italy	2017-2020	F	Health, Marketing
PhD students with financial support from BigInsight					
Solveig Engebretsen		Norway	2016-2019	F	Health
Emanuele Gramuglia		Italy	2016-2019	M	Sensor
Andrea Chi Zhang		China	2016-2019	F	Health
Christoffer Haug Laache		Norway	2017-2020	M	Health
Daniel Piacek		Slovenia	2017-2017	M	Fraud
Martin Tveten		Norway	2017-2020	M	Sensor
PhD students in BigInsight with financial support from other sources					
Derbachew Asfaw	UiO	Ethiopian	2016-2018	M	Marketing
Andreas Brandsæter	DNV-GL, NæringslivPHD	Norway	2015-2018	M	Sensor
Marta Crispino	Bocconi	Italy	2016-2017	F	Marketing
Vinnie Ko	UiO	Netherlands	2016-2019	M	Fraud
Håvard Kvamme	UiO	Norway	2015-2018	M	Marketing
Richard Xiaoran Lai	UiO	UK	2015-2018	M	Health
Sylvia Qinghua Liu	UiO/MI Innovation	China	2016-2020	F	Marketing
Andreas Nakkerud	UiO/MI Innovation	Norway	2016-2020	M	Power
Zhi Zhao	UiO/IMB	China	2016-2018	M	Health
Yinzhi Wang	UiO/MI	China	2016-2018	F	Fraud
Master degrees					
Kristin Bakka			2017-2018	F	Sensor
Jenine Gaspar Corrales			2017-2018	F	Marketing
Daniel Piacek			2016-2017	M	Fraud
Jonas Schenkel			2016-2017	M	Marketing
Martin Tveten			2016-2017	M	Sensor

FINANCIAL OVERVIEW

FUNDING	1000 NOK
The Research Council	9 111
Norwegian Computing Center (NR)	1 767
Research Partners*, in kind	9 727
Research Partners*, in cash	585
Enterprise partners**, in kind	3 475
Enterprise partners**, in cash	4 989
Public partners***, in kind	3 397
Public partners***, in cash	1 720
Sum	**34 771**

COSTS	
NR, research	11 207
NR, direct costs	813
Research Partners*, research	15 480
Enterprise partners**, research	3 475
Public partners***, research	3 797
Sum	**34 771**

* Research partners: UiO, UiB
** Enterprise partners: Telenor, DnB, Gjensidige, Norsk Hydro, DNV-GL, ABB
*** Public partners: Norwegian Tax Administration (Oslo), University Hospital HF, NAV, Public Health Insitute (NIPH)

PUBLICATIONS IN 2017

Journal and peer-reviewed conference papers

Aure, Miriam Ragle; Vitelli, Valeria; Jernström, Sandra Johanna; Kumar, Surendra; Krohn, Marit; Due, Eldri Undlien; Haukaas, Tonje Husby; Leivonen, Suvi-Katri; Vollan, Hans Kristian Moen; Luders, Torben; Rødland, Einar Andreas; Vaske, Charles; Zhao, Wei; Møller, Elen Kristine; Nord, Silje; Giskeødegård, Guro F.; Bathen, Tone Frost; Caldas, Carlos; Tramm, Trine; Alsner, Jan; Overgaard, Jens; Geisler, Jürgen; Bukholm, Ida Rashida Khan; Naume, Bjørn; Schlichting, Ellen; Sauer, Torill; Mills, Gordon B.; Kåresen, Rolf; Mælandsmo, Gunhild; Lingjærde, Ole Christian; Frigessi, Arnoldo; Kristensen, Vessela N.; Børresen-Dale, Anne-Lise; Sahlberg, Kristine Kleivi; OSBREAC, Oslo Breast Cancer Consortium.
Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. Breast Cancer Research 2017;19(1):44

Belay, Denekew Bitew; Kifle, Yehenew Getachew; Goshu, Ayele Taye; Gran, Jon Michael; Yewhalaw, Delenasaw; Duchateau, Luc; Frigessi, Arnoldo.
Joint Bayesian modelling of time to malaria and mosquito abundance in Ethiopia. BMC Infectious Diseases. 17(1) doi: 10.1186/s12879-017-2496-4. 2017.

Brandsæter, Andreas; Vanem, Erik; Glad, Ingrid Kristine.
Cluster Based Anomaly Detection with Applications in the Maritime Industry. In: Proceedings of the 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control SDPC 2017. pp 328-333. doi: 10.1109/SDPC.2017.69. 2017.

Crispino, Marta; Arjas, Elja; Vitelli, Valeria; Frigessi, Arnoldo.
Recommendation from intransitive pairwise comparisons. CEUR Workshop Proceedings.1688 2016.

Fleischer, T, Tekpli, X, Mathelier, A, Wang, S, Nebdal, D, Dhakal, H, Sahlberg, K, Schlichting, E, OSBREAC Oslo Breast Cancer Research Consortium, Børresen-Dale, AL, Borgen, E, Naume, B, Eskeland, R, Frigessi, A, Tost, J, Hurtado, A, Kristensen, V.
DNA methylation at enhancers identifies distinct breast cancer lineages, Nature Communications, 8, 1379 (2017)

Hellton KH, Thoresen M.
When and Why are Principal Component Scores a Good Tool for Visualizing High dimensional Data?. Scandinavian Journal of Statistics. 2017 Sep 1;44(3):581-97.

Løland, Anders; Berset, Anders; Hobæk Haff, Ingrid.
Er maskinlæring framtida i Skatteetaten? Praktisk økonomi & finans. (3) pp 344-352. doi: /10.18261/issn.1504-2871-2017-03-06. 2017.

Vanem, Erik; Brandsæter, Andreas; Gramstad, Odin.
Regression models for the effect of environmental conditions on the efficiency of ship machinery systems. I: Risk, Reliability and Safety: Innovating Theory and Practice: Proceedings of ESREL 2016 (Glasgow, Scotland, 25-29 September 2016). CRC Press 2017 ISBN 9781138029972.

Vanem, Erik; Storvik, Geir Olve.
Anomaly Detection Using Dynamical Linear Models and Sequential Testing on a Marine Engine System. In: PHM 2017 Proceedings of the Annual Conference of the Prognostics and Health Management Society 2017. PHM Society. pp 185-200. 2017.

Reports and other publications

Crispino M, Arjas E, Vitelli V, Barrett N, Frigessi A.
A Bayesian Mallows approach to non-transitive pair comparison data: how human are sounds?. arXiv preprint arXiv:1705.08805. 2017 May 24.

Kvamme, Håvard; Sellereite, Nikolai; Aas, Kjersti.
Credit Scoring using Deep Learning of Time Series data. Oslo: Norsk Regnesentral 2017 43 s. NR-notat(SAMBA/11/17)

Vanem, Erik.
Dynamical Linear Models Applied to Ship Sensor Data. Høvik, Norway: DNV GL 2017 (ISBN 978-82-515-0320-4) 79 s.

BigInsight

Biginsight.no

Postal address
PO box 114 Blinderen
NO-0314 Oslo
Norway

Visiting addresses
BigInsight
Norsk Regnesentral
Gaustadalléen 23a
Kristen Nygaards hus, 4th floor
0373 Oslo

BigInsight
Oslo Center for Biostatistics and Epidemiology (OCBE)
University of Oslo
Sognsvannsveien 9
Domus Medica
0372 Oslo

BigInsight
Department of Mathematics
University of Oslo
Moltke Moes vei 35
Niels Abel Hus, 8th floor
0316 Oslo

BigInsight
Oslo Center for Biostatistics and Epidemiology (OCBE)
Oslo University Hospital
Klaus Torgårdsvei 3
Sogn Arena, 2nd floor
0372 Oslo

Email contacts
Arnoldo Frigessi frigessi@medisin.uio.no
Ingrid Glad glad@math.uio.no
Lars Holden lars.holden@nr.no
Ingrid Hobæk Haff ingrihaf@math.uio.no
André Teigland andre.teigland@nr.no
Kjersti Aas kjersti.aas@nr.no
Anders Løland anders.loland@nr.no

Phone contact
Arnoldo Frigessi +47 95735574
Norsk Regnesentral +47 22852500