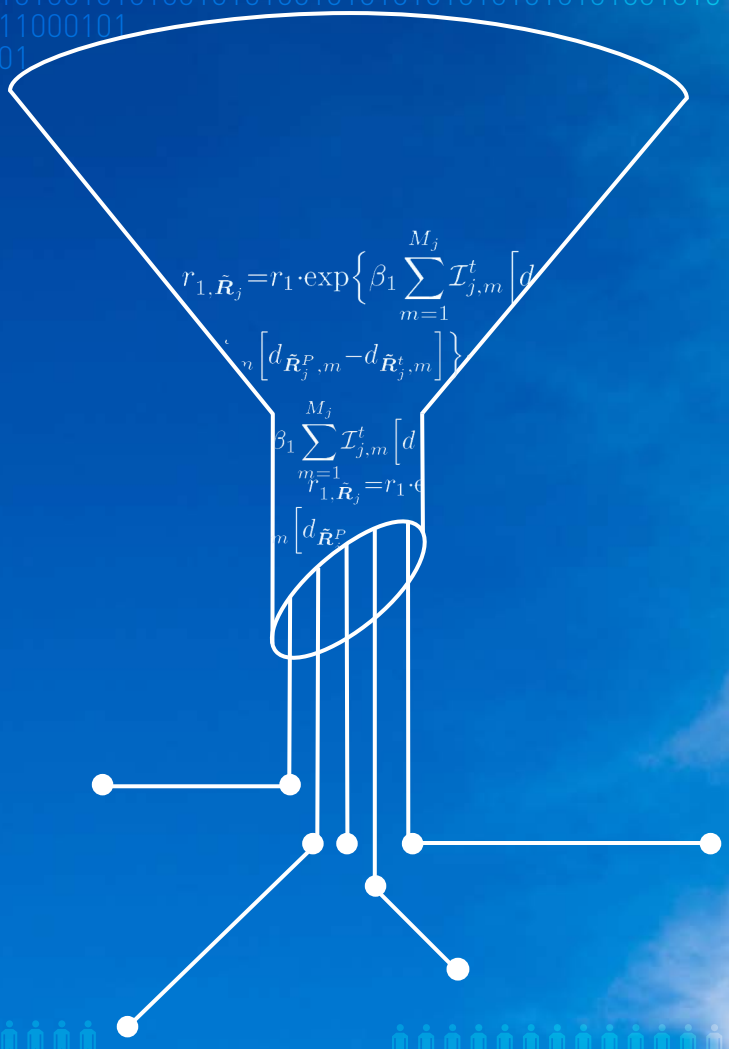
BigInsight

STATISTICS FOR THE KNOWLEDGE ECONOMY

ANNUAL REPORT 2015



Centre for Research-based Innovation

The Research Council of Norway

BigInsight

Without the right analytical methods, more data just gives a more precise estimate of the wrong thing

The Economist, Technology Quarterly

CONTENT

SUMMARY

BigInsight produces innovative solutions for key challenges facing a consortium of private and public partners, by developing and applying original statistical and machine learning methodologies. Exploiting huge and unique data resources and substantial scientific, industrial and business knowledge, we construct personalised solutions, predict dynamic behaviours and control processes that are the core of the partners' innovation. We shall discover radically new ways to target towards individual needs and conditions, products, services, prices, therapies, technologies, 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, etc. 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.

Research at BigInsight will lead to value creation and will help bring our partners to new leading positions in their global markets. BigInsight creates unique competence and capacity for the Norwegian knowledge-based economy, contributing to the development of a sustainable and better society.

BigInsight started in April 2015 as the first of the third generation Norwegian Centres for Research-based Innovation. It is funded by the Research Council of Norway and by fourteen partners and will operate until 2023. This is the annual report of the first year of the center.

A flood of data
is created every day. (...)

(We) are beginning to realise the potential for channeling these torrents of data into actionable information that can be used to identify needs,
provide services,
predict and prevent crises.

World Economic Forum

VISION AND OBJECTIVES

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

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 software 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 will be significantly better than the state-of-the-art, thanks to brilliant, courageous and creative generic methodologies to extract knowledge from complex data. Generic methodology and their new applications will be published on 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 economy of the future.

Analytics: are we data rich and insights poor?

Jo Boman, Research World, January/February 2015

Personalised solutions

The core business and operation of our partners involves interacting with many individual units: at Telenor, 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 between themselves using Vipps; at OUS, cancer patients need to be treated in the most effective personalized way; the Norwegian Insititute of Public Health works with individuals susceptible for infections or infectious themselves; NAV supports hundreds of thousands of people with special needs in relation to health or employment in challenging life situations; for Skatteetaten, millions of taxpayers need to be mentored to declare their incomes; for DNV GL and ABB, hundreds of sensors register the functional state of a ship at sea; for DNV GL and OUS, a multitude of sensors monitor safety in healthcare and patients' health records are digitalised. 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 intelligence, and are exposed to their environment.

Every partner has specific management objectives for its units, but they share the goal to deeply innovate the management of their units, by recognising similarities and exploiting diversity between units. This will allow personalised marketing, personalised products, personalised prices, 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 cancer treatment (OUS), a sensor on a ship on sea (ABB, DNV GL), a customer offered products from several providers (Telenor, Gjensidige, DNB), a worker who lost his job (NAV), the price of an asset in a complex market (Norsk Hydro) – are all examples of systems in a transient phase. DNB, NAV, Skatteetaten and Gjensidige are interested in the prediction of certain behaviours of their customers and service users, finding causes of churn, criminal financial or fraud activities, in order to step in with new prices, products, legal actions or investigations. For OUS and DNV GL, 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, thanks to prompt medical and nursing action. For ABB and DNV GL high dimensional times series are generated by sensors monitoring a ship or an industrial installation, 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.

Photo: Igor Stevanovic / Shutterstock.com

ORGANISATION

Board in 2015

Tone Sejnæs Eilertsen, Gjensidige, chairman
Tron Even Skyberg, DNB
Stefan Erath, Hydro
Bjørn Taale Sandberg, Telenor
Rune Braastad, ABB
Rune Torhaug, DNV GL
Ludvig Guldal, NAV
Marcus Zackrisson, Skatteetaten
Peder Heyerdahl Utne, Oslo University Hospital
John-Arne Røttingen Folkehelseinstituttet
Arne Bang Huseby, University of Oslo
Bård Støve, University of Bergen
Lars Holden, Norsk Regnesentral
André Teigland, Norsk Regnesentral
Observer: Terje Strand, Research Council of Norway
The board had 2 meetings in 2015.
All partners are represented in the Board.

Legal organisation

BigInsight is hosted by NR. The legal and administrative responsibility of BigInsight is with managing director Lars Holden, NR.

Scientific Advisory Committee (SAC)

To be appointed in 2016.

Director

Arnoldo Frigessi, UiO
Director

Co-Directors

Ingrid Glad, UiO,
Co-Director

Kjersti Aas, NR,
Co-Director

André Teigland, NR,
Co-Director

Håvard Rue, NTNU,
Co-Director

Principal Investigators

Magne Aldrin, NR
Arne Bang Huseby, UiO
Arnoldo Frigessi, UiO
Ingrid Glad, UiO
Clara Cecilie Günther, NR
Alex Lenkoski, NR
Anders Løland, NR
Håvard Rue, NTNU
Magne Thoresen, UiO
Kjersti Aas, NR

BigInsight

BOARD

NR´S RESPONSIBLE PERSON — ADVISORY SCIETIFIC 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

Distributed decision making (start 2017)

PARTNERS

Partners

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

Cooperation between the partners of BigInsight

The eight months of 2015 have been used to prioritise precise and concrete projects within each innovation objective, identifying methodological approaches and case studies suggested by the innovation partners. This has been a major process, involving all partners and many people. There have been brainstorming meeting with all interested innovation partners and researchers, and smaller working groups with individual innovation partners and few researchers. Precise research aims have then been identified and research teams have been set up. Innovation partners are very active. There are frequent joint working sessions and regular update and progress meetings. Coordination between Innovation Objectives is assured because several researchers are working in multiple projects, and PI's meet regularly for updates.

RESEARCH STRATEGY

Within our two innovation themes, we have identified five innovation objectives (plus one more to start in the next years) to harbour all research at BigInsight. Each innovation objectives (IOs) is centered on a concrete innovation area: marketing, health, fraud, sensor, power. Most partners join more than one IO (and changes will happen in the years to come, as focus and interests will develop further). Each IO has a few specific innovation aims related to outstanding unresolved problems, which we believe can specifically be solved with new statistical, mathematical and machine learning methodologies. Our research projects are designed to 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	DNV-GL	DNB	ABB	DNV-GL
Gjensidige	Kreftregisteret	Gjensidige	DNV-GL	Hydro Energy
Skatteetaten	OUS	NAV		
Telenor	Telenor	Skatteetaten		

RESEARCH PARTNERS

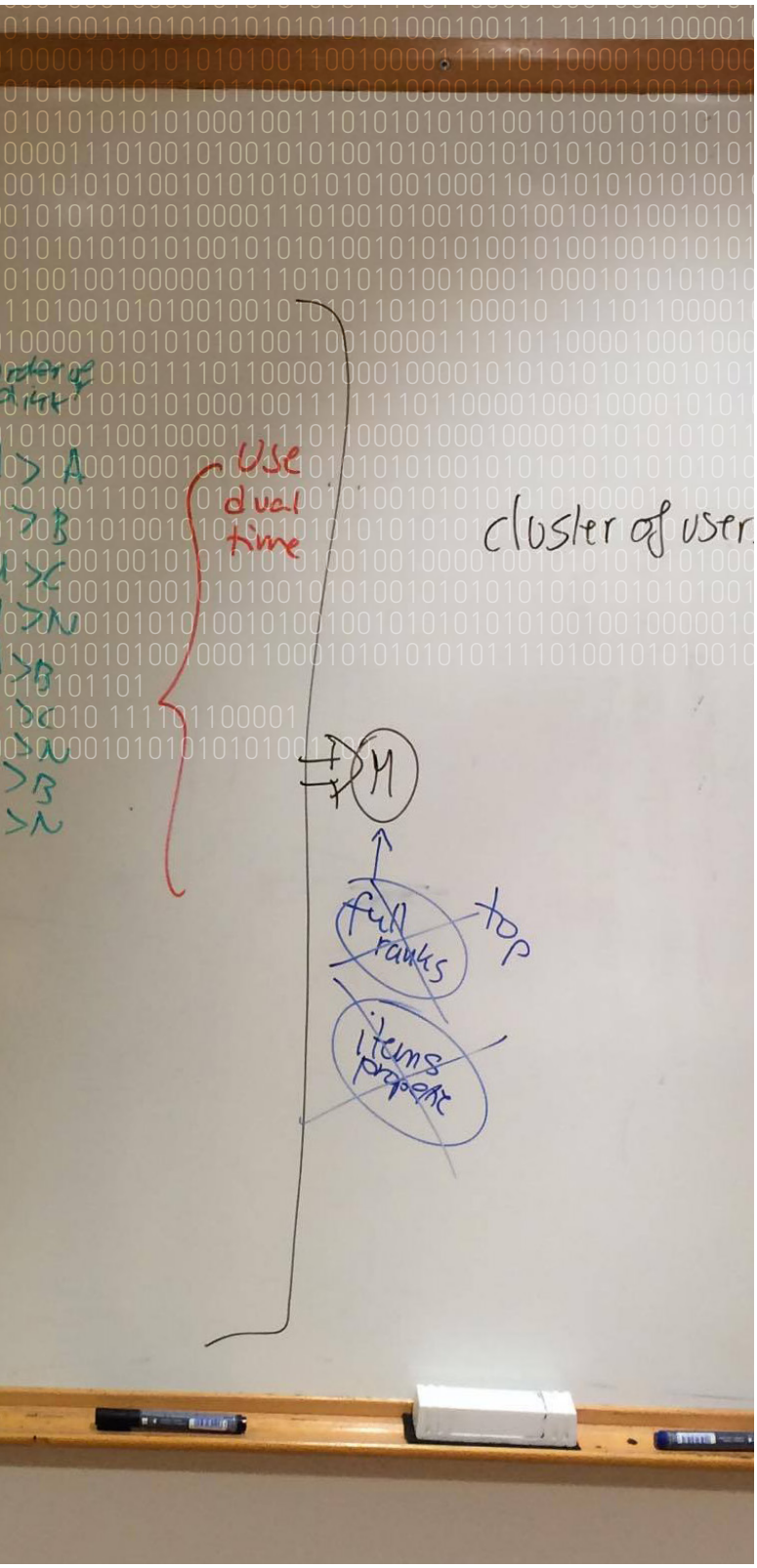
NR	UiO	NR	NR	NR
UiO	OUS	UiO	UiO	UiO
NIPH	NR	UiB	UiB	UiB
UiB	UiB			
	NIPH			

PRINCIPAL INVESTIGATORS

Principal Investigators:	Kjersti Aas	Magne Thoresen	Anders Løland	Ingrid Glad	Alex Lenkoski
co-Principal Investigators:	Arnoldo Frigessi	Clara Cecilie Günther	Håvard Rue	Magne Aldrin	Arne Bang Huseby

METHODS

We attack the innovation challenges by developing solutions which are based on 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.

Accounting for uncertainty of estimates is fundamental in decision making.

Industry, business, science and service make decisions in the presence of multiple sources of uncertainty. Recent research shows the benefit of probabilistic forecasting, where a full distribution is issued, over more classic point forecasts. Large levels of uncertainty are not just bad news, as they indicate the need for collecting more and sharper data. Combining a large number of models yields often an improved ensemble predictor.

Causal effects enable effective actions.

Causal statistical modelling is opening new avenues in data analysis, moving from pointing to association to estimating causes. Graphical, counterfactual and networks models allow to ask what would have happened if an intervention had not been made? While counterfactual cases cannot be observed, some knowledge about them can usually be inferred from the data collected. Time-dependent confounding is a particular challenge. The distinction between direct and indirect effects allows interpreting causes.

Rapid computation means 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.

Surprise 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 a procedure to monitor parallel streams of data for a set of (lagged) 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 optimisation 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 to optimise 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.

SCIENTIFIC ACTIVITIES

After a period of intense, inspiring and very creative brainstorming among all BigInsight partners, crossing disciplines, industrial sectors and areas of primary interest, we identified generic questions and challenges, motivated by innovative ideas at the core of our partners' activity. For each such specific question, we selected specific innovation cases from the partners, and started to work on them. Below is a snapshot of some of the themes that crystallized and which we are now working on.

Each IO has build 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.

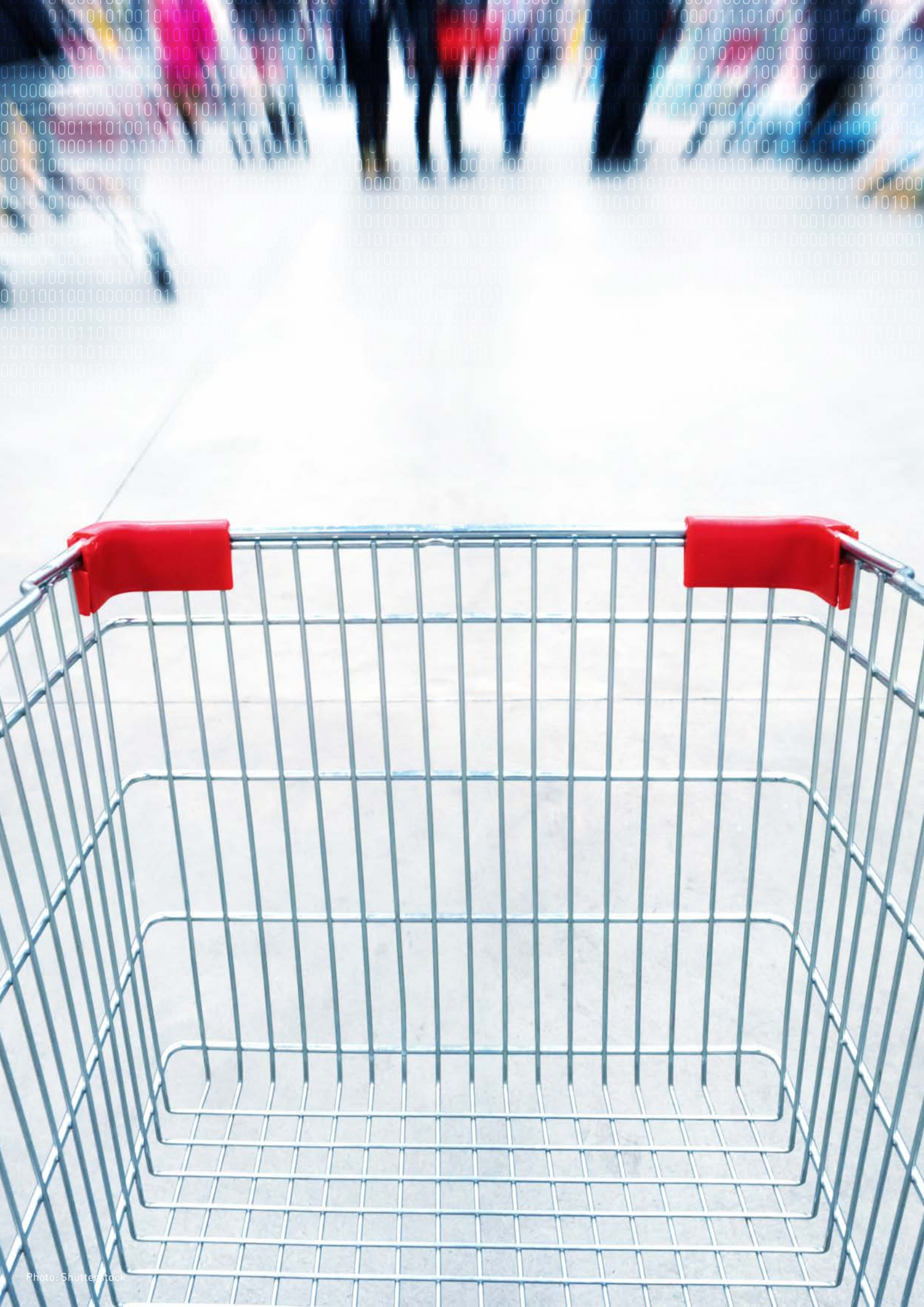


Photo: Shutterstock

PERSONALISED MARKETING

We develop new methods, strategies and algorithms for individualised marketing, customer retention, optimised communication with users, personalised pricing, personalised products 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, geographic information). We design and perform comparative trials with well-designed user cohorts, exploiting for example network relations, and exploit observational data in counterfactual (what-if) market studies. A key point is to identify the actionable causes of customer behaviour.

Personalised communication

There are many communication channels that customers (or users) can actively choose to take contact with the company (or service). And vice-versa, the company (or administration) can use diverse channels to communicate to the customers (users). The purpose of communication varies from successful sale to correctness in tax self-reporting. The objective is to determine, for each customer, which is the optimal communication channel to maximise the probability of success. Also important is how to organise the communication channels (content, operation, design) so that customers find by themselves the best channel. Data include standard structured data, as well as unstructured data (text, phone call records, chats, and for web based services, click histories) for each customer. Each channel has a cost and it is interesting to consider the probability of success per unit cost of communication.

Personalised marketing

The purpose is to intervene strategically on individual customers, to increase the chances for specific actions and behaviour (buy, use a new product; pay a loan; etc). We develop new models which describe customer behaviour, depending on own characteristics and external exposures and multiple interactions. The detection of key factors is crucial and network relations play a fundamental role. Once customer behaviour is appropriately modelled, we study by simulation which interventions among several possible ones, and on which customers, are optimal. Understanding how networks of customers grow in time and topology; how accumulation of ratings and rankings of products allow learning individual preferences; how sales spread virally on the network of customers - allows to experiment with new individualised marketing strategies. Data include structured data on each customer and on the product characteristics, time series of individual behaviour and their networking, as well as external data, including products and prices from competitors.

Activities 2016

- Bayesian methodology for recommendation systems
- Stochastic models for early prediction of viral customer behaviour on networks
- Stochastic customer network growth dynamics
- Disentangle the effect of competing advertising and marketing actions on customer behaviour
- Understand the factors leading to success or failure in tax-self-reporting

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 and screening, better prevention and monitoring of epidemics. We show how such data can be exploited, with a series of innovative prototype projects.

The organized cancer screening programs in Norway have so far used a generic "one-size-fits-all" approach for screening individuals. We develop new screening strategies using genetic markers and molecular profiles together with other risk factors, in order to individualize screening intensity, optimize early detection of cancer and increase cost-effectiveness of screening. Furthermore, national population based cancer registries routinely publish survival statistics. To become more relevant both to the clinician and the patient, the survival statistics should preferably be tailored to encompass more detailed information, moving towards personalized cancer statistics.

Standard cancer treatment consists of a combination of drugs, at various dosages and in different order, implying a huge number of possible cocktails. We develop a new approach to therapy optimization, based on simulation of cancer growth. Also, we develop methods to predict the synergy between drugs based on cancer cell line data. This allows exploring new approaches to personalized cancer treatments.

Patient safety is critical in healthcare. The amount of data collected in healthcare is vast and rapidly expanding, including electronic health records (EHR) and health care system/organization data. We develop new data analytic methods to predict and control risk in healthcare organizations at system level.

We use data on mobile phone locations and their movements to describe movement of and contacts between people, with the purpose to monitor and predict the development of infectious diseases. Simulating various vaccine plans allows to evaluate cost-benefits.

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.

Personalized cancer treatments

We develop a new mathematical, statistical and computational methodology to improve current ways of predicting treatment response for breast cancer patients. We design theoretical and computational models that capture key molecular and cellular mechanisms of the breast cancer, reproducing the effect of specific treatments used in actual clinical trials. The models are personalized to specific subgroups by calibrating the computer simulations to actual patient data. This allows personalized treatment guidance, by simulating an array of possible treatment schedules for a given patient to find the optimal regime.

Healthcare safety management

This project harvest information lying in the collections of high dimensional health records, routinely acquired in hospitals and health institutions, which carry a preventive signal relative to a potential harm to patients. This signal is used to predict the occurrence of a possible harm in an automatic way at aggregated health institution/ward level. It can also be used to define new control rules on hazards, to prevent and mitigate the risk of harm.

Telecom data for epidemics control

We study the spread of infectious diseases, by observing social mixing and mobility patterns of susceptible and infectious individuals, who are key drivers of the spatial dissemination of infectious diseases. The use of mobile

phone data containing geo-temporal information from individuals gives an accurate, real-time descriptions of population movements. This in turn allows accurate predictions of epidemic spread.

Activities 2016

- Variable selection in highly dimensional regressions
- Multi-state Markov models with frailties
- Personalized survival statistics
- Stochastic model of breast cancer growth
- First pilot study of hazard/harm prediction at system level for a hospital ward
- Mobile movement data in Bangladesh for infectious diseases prediction
- Drug synergy prediction in cancer cell lines

Principal Investigator
Magne Thoresen

co-Principal Investigator
Clara Cecilie Günther

Photo: Shutterstock

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 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 types of fraud. We also are interested in money laundering detection. We develop adaptive tools that use "all data", including payment logs, relational networks and other available digital records.

A further objective is to combine in an optimal way the multitude of fraud detection models, 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.

Ensemble methods for fraud detection

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. With interactions, the number of covariates is huge. The number of investigated fraud cases is generally low. In addition a great number of undetected cases exist. The objective is to produce a trustworthy probability of fraud for each case. Many statistical and machine learning methods already exist. Combining results produces better results than a single method can. We construct a toolbox for combining fraud forecasting models, exploiting both the time series aspect of the data and variables in the data beyond the probabilities stemming from each individual model.

Text-mining for fraud detection

In addition to ordinary variables (age, demography, background, behaviour, etc...), a potential fraud case can be accompanied with a variable amount of text, for example where a customer describes a claim or an officer summarises a case. These texts are informative for a human eye searching for fraud. The objective is to exploit recent advances in text-mining to produce text related features that can be used in the statistical models, and to investigate the added value of these.

Network analysis for fraud detection

Fraud is viral, spreading directly or indirectly from one fraudster to others. Exploiting knowledge about social relations can be useful. 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.

Activities 2016

- Select relevant covariates, factors and time series, out of a very large number
- Model and exploit network relations between individuals and businesses
- Text mining
- New toolbox for fraud detection based on ensemble methods
- Misreporting of VAT and risk for future misreporting of VAT
- Detection of money laundering,
- Detection of insurance fraud
- Detection of social security fraud.

Principal Investigator
Anders Løland

co-Principal Investigator
Håvard Rue

Photo: Shutterstock

SENSOR SYSTEMS

Maintenance and inspections of ships are traditionally based on a preventive scheme where components have been overhauled or maintained according to a time schedule. This philosophy 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 by preventive scheduled maintenance. We develop radically new approaches based on the recent availability of large arrays of sensors, which monitor condition and behaviour 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 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 network of 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.

Activities 2016 (with corresponding deliverables)**:**

- New methodologies for sensor based condition monitoring.
- Effects of weather, sea conditions, climate and other external noise factors on the sensor series
- Surprise detection
- Simulated data and actual ship data
- Prediction of change points in sensor series

Condition monitoring: Fault and anomaly detection and prediction

We develop multisensor, multiscale statistical approaches to detect, as rapidly as possible, faults and anomalies in multiple sensor time series and to predict these before they occur. Quantification of uncertainty of anomaly detection and prediction is important, with implication on intervention and maintenance strategies. Statistical models and algorithms will need to run in required real time. Our approach is data driven, inspecting composite sensor features drifting in time, or knowledge driven, representing first principal physical and functional relations.

Sesnor data filtering

Sensor data are affected by the ways the ship is operated and by the weather and ocean conditions. These components of the signal need to be filtered out from the data, in order to be able to detect actual abnormalities, signal predicting faults and faults themselves.

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 photovoltaic, 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.

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 use 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.

Using river inflow projections to augment electricity price spot forecasts

River inflow is a critical quantity that impacts the bidding strategies of hydroelectric operators. We use newly available data on river inflows for Norway and Sweden to build a river inflow component into spot price forecasting system.

Power Matching Problem for the Smart Grid

The Power Matching Problem amounts to determine the optimal (according to user preferences) e-profile, that is the utilization of the available energy during the time horizon by the appliances in the households so that demand and supply meet and all time constraints are satisfied. As such, this can be seen as a Resource Constrained Scheduling Problem with Time Windows and in presence of uncertainty. We can use stochastic optimization, where one minimizes the expected value of a given objective function, based on some probability distribution attached to a set of possible input scenarios. Alternatively, robust optimization does not need such a probability distribution, as just ranges of operations are needed.

Activities 2016

- New forecasting methodology using bid/ask curves for error dressing
- Nordpool and the German (EEX) market.
- River inflow model for the Nordpool spot market.
- Power Matching Problem for the Smart Grid

Principal Investigator
Alex Lenkoski

co-Principal Investigator
Arne Bang Huseby

INTERNATIONAL COOPERATION

Strategic International Academic Partners (SIAP)

SIAPs contribute to place BigInsight in the center of the global data science community. SIAP's staff collaborates in research and co-supervise PhD students. Joint workshops and training events will be organised.

Our first SIAP is STOR-i, Statistics and Operational Research in partnership with Industry, 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, recommendation systems, multivariate sensor data. Additional SIAPs are planned for 2016.

STOR-i
excellence with impact

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 2015 we hosted the following longer visits:

Professor Sylvia Richardson,
MRC Biostatistics Unit, Cambridge
(Cancer genomics, Bayesian models and computations)

Professor EljaArjas,
University of Helsinki
(Causality,Bayesian models, Recommendation systems)

Professor Idris Eckley,
University of Lancaster
(Change points, wavelets, sensor data)

Professor Mike West,
Duke University
(Bayesian dynamical systems)

Professor Gianpaolo Scalia Tomba,
University of Roma Tor Vergata
(Infectious diseases)

Professor MichaelSchimeck,
University of Graz
(Rank data)

Professor Elja Arjas has a 20% adjunct position at BigInsight.

In 2016 **Professor Per Mykland** and **professor Lan Zhang**, University of Chicago, will spend a sabbatical year at BigInsight; **professor Peter Müller** (University of Texas) will spend the autumn semester at BigInsight.

International training programme

PhD students from other universities can spend periods of training and research collaboration at BigInsight. In 2015 we welcomed

- Marta Crispino, University Bocconi, Milano (Recommendation systems)
- Denekew Belay Bitew, Hawassa University (Malaria models)
- Derbachew Asfaw, Hawassa University (Dynamic rank data)
- Jamie-Leigh Chapman, University of Lancaster (Sensor data)
- Alvaro Kohn Loque, Technische Universität Dresden (Cancer personalised therapy, numerical models)
- Vendula Svendowa, University of Graz (Rank data)

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

Hawassa University

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 nominated as a Member in the Management Committee and professor Birgitte Freiesleben de Blasio is nominated as deputy. This EU action starts 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.
http://www.cost.eu/COST_Actions/ca/CA15109

BigInsight applied in 2015 for a MARIE SKŁODOWSKA-CURIE ACTIONS Innovative Training Networks, Data Science Meets Semantic Technology in Industry together with University of Lübeck, Dresden, Oxford and several industrial partners, including DNV-GL. We expect to know about the funding in 2016.

Other International activities

BigInsight has a team joining the international competition DREAM Challenge AstraZeneca-Sanger Drug Combination Prediction
https://www.synapse.org/#!Synapse:syn4231880/wiki/235645

The AstraZeneca-Sanger Drug Combination Prediction Challenge

We are working towards a project with the International Centre for Diarrhoeal Disease Research, Bangladesh and other health authorities in Bangladesh on prevention and control of infectious diseases, exploiting mobile location networks of Grameenphone Bangladesh.

COMMUNICATION AND DISSEMINATION ACTIVITIES

- **biginsight.no** is the webpage of the center.

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

- **BigInsight Wednesday Lunch Seminars** are take place every second week, alternating between the lunch room at NR and the newly refurnished "Erling Sverdrups plass" on 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 presentations:

EVENT	ORGANISER
Software 2016, Oslo	Dataforeningen
Partnerforums høstkonferanse 2015, Osllo	Partnerforum
Oslo University Hospital Stabsmøte, Oslo	OUS
SFI Food of Norway kick-off, Aas	FoodofNorway
BeslutningsstøtteDagen 2015 Norge, Stockholm	Allchar
Big Data & Analytics Insight 2015, Oslo	Business Analytics For All
NFR SFI-forum, Oslo	NFR
Skatteetatens analysesamling 2015, Oslo	Skattedirektoratet
ABB Marine key manager annual conference, Holmsbu	ABB Marine
Apollon lansering sommer 2015, Oslo	Apollon
Big Data Value Association seminar, Oslo	Sintef

- BigInsight in the media
  - Apollon, 16.11.2015 **De svakes ridder, av Amalie Kvame Holm**
  - Universitas, 2.11.2015 **Facebook kan redde livet, av Andreas Løhren**
  - ComputerWorld, September 2015: **"Innovasjon vs. personvern", Stig Øyvann, 3. september 2015**
  - Apollon, Forsknigsmagasin, August 2015 **"Her er likheten mellom kreft og skattejuks", Yngve Vogt, 14. aug. 2015.** See also WWW.FORSKNING.NO, 13.1.2016
  - Digi.no **"Lærer å utnytte dine dataspor", Marius Jørgenrud, 3. desember 2014**
  - Tidsskriftet Stat & Styring **«Stordata presser grensene», Mads A. Danielsen, 19.10.2015**

Partnerforum


BigInsight
BigInsight: Statistics for the knowledge economy
News and events
2016 - year two of BigInsight
Future events
Wednesday Lunch: Mette Langaas
Visitors 2016
TUESDAY STATISTICS SEMINAR: Manuela Zucknick


Bredbånd: Altibox leder an
Bra at regjeringen vil ha mer spill
COMPUTERWORLD
Helseforskerne omfavner
Big Data
PURITY
Problemer med it-helsa


APOLLON
Dataeksplosjonen
fremmer vitenskapen
Protein vekter genene våre
Fattige skaper sårbar hvit rike
Norsk rett diskriminerer innvandrerkvinner
UiO : Universitetet i Oslo

RECRUITMENT

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

The planned recruitment in 2016 includes 6 PhD students, 1 postdoc, 3 professors, 1 associate professor. Professor Jukka Corander, coming from University of Helsinki, starts 1.4.2016 at UiO, Oslo Center for Biostatistics and Epidemiology.

In 2015 started:

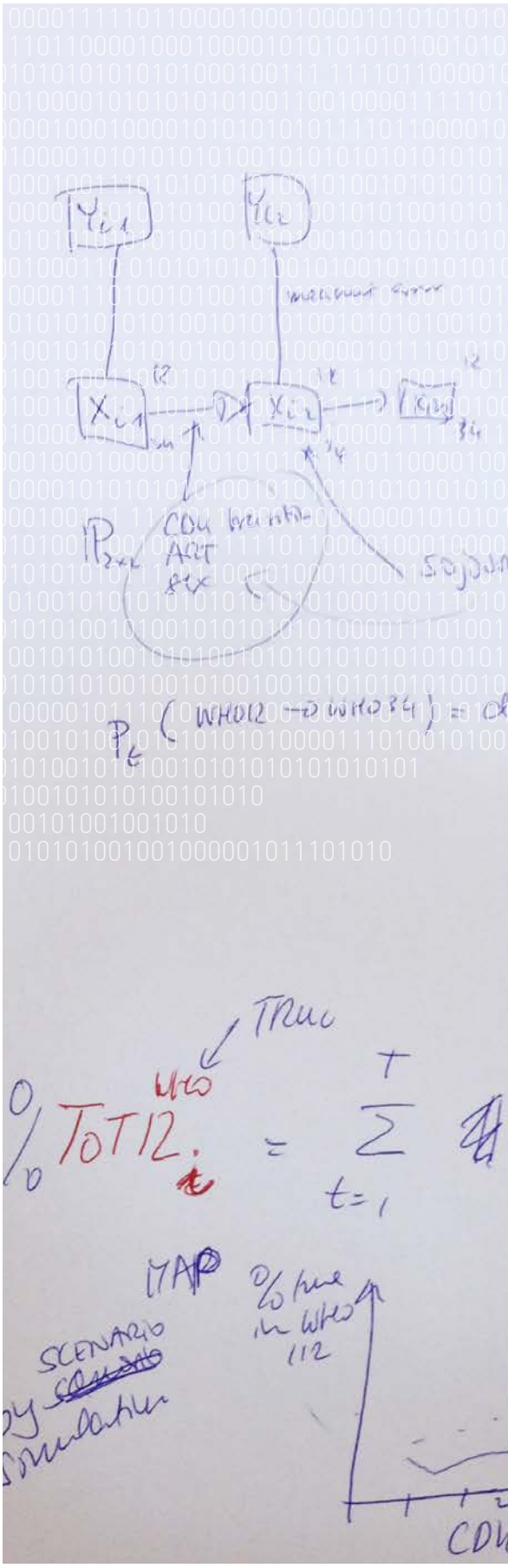
NAME	POSITION	FUNDING	RESEARCH AREA	AFFILIATION
Alvaro Kohn Loque	Postdoc	UiO/ERC	Personalised cancer treatment	UiO, Oslo Center for Biostatistics and Epidemiology
Andreas Brandsæter	PhD student	NFR Industrial-PhD and DNV-GL	Sensor, condition monitoring	UiO, Dept Mathematics and DNV-GL
Carlo Mannino	Professor II	BigInsight	Power, smart grid	UiO, Dept Mathematics
Dag Tjostheim	Professor II	BigInsight	Fraud	NR and UiB
Elja Arjas	Professor II	BigInsight	Personalised Marketing and Health	UiO, Oslo Center for Biostatistics and Epidemiology
Erik Vanem	Associate Professor II	BigInsight	Sensor, condition monitoring	UiO, Dept Mathematics and DNV-GL
Giacomo Meanti	Summer student	BigInsight	Hadoop computing	UiO, Oslo Center for Biostatistics and Epidemiology
Håvard Kvamme	PhD student	UiO	Personalised marketing	UiO, Dept Mathematics
Håvard Rue	Professor II	BigInsight	Co-director, fraud	NR and NTNU
Ingrid Hobæk Haff	Associate Professor	UiO	Fraud	UiO, Dept Mathematics
Jamie Chapman	PhD student	Lancaster Univ.	Sensor, change points	Lancaster Univ.
Jon Michael Gran	Researcher	OUS	Health, Personalised marketing	OUS
Marissa Le Blanc	Researcher	OUS	Health	OUS
Martin Jullum	Researcher	NR	Fraud	NR
Metter Langaas	Professor II	BigInsight	Health, fraud	NR and NTNU
Nikolai Sellereite	Researcher	NR	Personalised marketing	NR
Tor Arne Øigård	Researcher	NR	Fraud, Power	NR

PERSONNEL

Personnel working in BigInsight

NAME	INSTITUTION	MAIN RESEARCH AREA
Rune Braastad	ABB	Sensor
Stian Braastad	ABB	Sensor
Maryam Hessareyeh	ABB	Sensor
Stig Leira	ABB	Sensor
Kenneth Nakken	ABB	Sensor
Jaroslaw Novak	ABB	Sensor
Gunnar Prytz	ABB	Sensor
Bengt Sørensen	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
Thorstein Danielsen	DNB	Fraud
Thor Aage Dragsten	DNB	Marketing
Heidi Fredriksen	DNB	Marketing
Sven Haadem	DNB	Marketing
Katharina Henriksen	DNB	Marketing
Roy Oma	DNB	Fraud
Steffen A. S. Sjursen	DNB	Marketing
Fredrik Strand	DNB	Fraud
Jørn Ødegård	DNB	Marketing
Mette S. J. Snilsberg	DNB	Marketing
Trond Tangen Simonsen	DNB	Fraud
Øystein Alnes	DNV-GL	Sensor
Theo Bosma	DNV-GL	Power
Ervin Bossanyi	DNV-GL	Power
Frédéric Courivaud	DNV-GL	Health
Marcel Eijgelaar	DNV-GL	Power
Odin Gramstad	DNV-GL	Sensor
Muhammad Jafar	DNV-GL	Power
Lars Landberg	DNV-GL	Power
Stephen Leyshon	DNV-GL	Health
Gabriele Manno	DNV-GL	Sensor
Bahman Raeissi	DNV-GL	Sensor
Bobby Ray-Sannerud	DNV-GL	Health
Elizabeth Traiger	DNV-GL	Power
Erik Vanem	DNV-GL & UiO	Power, Sensor
Bjørn-Johan Vartdal	DNV-GL	Sensor
Tone Sejnæs Eilertsen	Gjensidige	Marketing
Anders Nyberg	Gjensidige	Marketing
Marte Olstad	Gjensidige	Marketing
Thea Margrethe Skouen	Gjensidige	Marketing
Geir Thomassen	Gjensidige	Fraud

NAME	INSTITUTION	MAIN RESEARCH AREA
Stefan Erath	Hydro	Power
Plamen Mavrodiev	Hydro	Power
Ivar Areklett	NAV	Fraud
Eigil Johnsen	NAV	Fraud
Bjørn Atle Sundsback	NAV	Fraud
Morten Tholander	NAV	Fraud
Jonas Andersson	NHH	Fraud
Birgitte De Blasio	NHI	Marketing, Health
Kjersti Aas	NR	Marketing, co-director, PI
Magne Aldrin	NR	Sensor, PI
Line Eikvil	NR	Fraud
Clara-Cecilie Günther	NR	Health, Marketing, PI
Martin Jullum	NR	Fraud
Ola Haug	NR	Marketing, Sensor
Marion Haugen	NR	Marketing
Lars Holden	NR	Health
Marit Holden	NR	Health
Ragnar Bang Huseby	NR	Power, Fraud, PI
Alex Lenkoski	NR	Power, PI
Anders Løland	NR	Fraud, Power, PI
Linda R. Neef	NR	Fraud
Hanne Rognebakke	NR	Marketing, Sensor
Nikolai Sellereite	NR	Marketing
Gunnhildur Steinbakk	NR	Fraud, Power
André Teigland	NR	Co-director
Ingunn Fride Tvete	NR	Health
John Enok Vollestad	NR	IT
Tor Arne Øigård	NR	Power, Fraud
Mette Langaas	NTNU	Health
Håvard Rue	NTNU	Fraud, co-director, PI
Ragnhild Sørum Falk	OUS	Health
Harald Fekjær	OUS	Health
Jon Michael Gran	OUS	Marketing, Health
Eivind Hovig	OUS	Health
Marissa LeBlanc	OUS	Health
Sygve Nakken	OUS	Health
Anders Berset	Skatteetaten	Marketing, Fraud
Wenche Celiussen	Skatteetaten	Marketing, Fraud
Øystein Olsen	Skatteetaten	Marketing
Audun Solli	Skatteetaten	Marketing, Fraud
Fatima Yusuf	Skatteetaten	Marketing
Geoffrey Canright	Telenor	Marketing, Health
Humberto Castejon	Telenor	Marketing
Kenth Engo-Monsen	Telenor	Marketing, Health
Anita Fjuk	Telenor	Marketing
Gro Nilsen	Telenor	Marketing
Hai Thanh Nguyen	Telenor	Marketing
Massimilano Ruocco	Telenor	Marketing
Astrid Undheim	Telenor	Marketing
Jan Bulla	UiB	Marketing
Bård Støve	UiB	Fraud
Dag Tjøstheim	UiB	Marketing, Sensor



NAME	INSTITUTION	MAIN RESEARCH AREA
Elja Arjas	UiO	Marketing, Health
Ørnulf Borgan	UiO	Marketing
Arnoldo Frigessi	UiO	Marketing, Health, Sensor, PI, director
Ingrid K. Glad	UiO	Sensor, co-director, PI
Ingrid Hobæk Haff	UiO	Fraud
Arne Bang Huseby	UiO	Sensor, Power, PI
Carlo Manino	UiO	Power, Sensor
Giacomo Meanti	UiO	Summerstudent
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
Sylvia Richardson	BSU-MRC	Health
Idris Eckly	Lancaster	Sensor

	FUNDING	NATIONALITY	PERIOD	GENDER	TOPIC
Postdoctoral researchers with financial support from BigInsight					
Alvaro Köhn Luque	(co funding ERC Scientia Fellow)	Spain	2015-2017	M	Health
Postdoctoral researchers in BigInsight with financial support from other sources					
Valeria Vitelli	UiO/Kreftforening	Italy	2015-2017	F	Marketing, Health
Kristoffer Hellton	UiO/NFR Focustat	Norway	2015-1017	M	Health
Gudmund Hermansen	UiO/NFR Focustat	Norway	2016-2018	M	Sensor
PhD students with financial support from BigInsight					
Håvard Kvamme		Norway	2015-2019	M	Marketing
PhD students in BigInsight with financial support from other sources					
Derbachew Asfaw	UiO/Norhed	Ethiopia	2015-2016	M	Marketing
Marta Crispino	Univ Bocconi Milano	Italy	2015-2016	F	Marketing
Xiaoran Lai	UiO	UK	2015-2018	M	Health
Andreas Brandsæter	DNV-GL, NFR NæringsPHD	Norway	2015-2018	M	Sensor
Jamie-Leigh Chapman	Univ Lancaster	UK	2015-2017	F	Sensor
Master degrees					
Martin Tveten				M	Sensor
Jonas Schenkel				M	Marketing

FINANCIAL OVERVIEW

FUNDING	1000 NOK
The Research Council	4 333
Norwegian Computing Center (NR)	1 285
Research Partners*, in kind	6 812
Research Partners*, in cash	500
Enterprise partners**, in kind	2 242
Enterprise partners**, in cash	2 216
Public partners***, in kind	3 578
Public partners***, in cash	667
Sum	21 633

COSTS	
NR, research	6 700
NR, direct costs	515
Research Partners*, research	8 199
Enterprise partners**, research	2 242
Public partners***, research	3 978
Sum	21 633

* 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 2015

(Not much is published yet, as this is our first year of activity... :))

Scientific journal papers

Sørensen, Øystein; Frigessi, Arnoldo; Thoresen, Magne. Measurement error in Lasso: impact and likelihood bias correction. Statistica Sinica 2015; 25. s. 809-829

Reports

Lenkoski, Alex; Løland, Anders; Hobæk Haff, Ingrid; Neef, Linda Reiersølmoen. Calibrated Probabilities and the Investigation of Soft Fraud in Automobile Insurance Claims. Oslo: Norsk Regnesentral 2015 25 s. NR-note (SAMBA/41/15)

Vitelli, Valeria; Sørensen, Øystein; Frigessi, Arnoldo; Arjas, Elja. Probabilistic preference learning with the Mallows rank model. : ARXIV 2015 45 s.

BigInsight

biginsight.no

Postal address:
BigInsight
Pbox 114 Blindern
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
Songsveien 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
Håvard Rue hrue@math.ntnu.no
André Teigland andre.teigland@nr.no
Kjersti Aas kjersti@nr.no

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

Design: Designavdelingen, Kunst i Skolen