# BigInsight

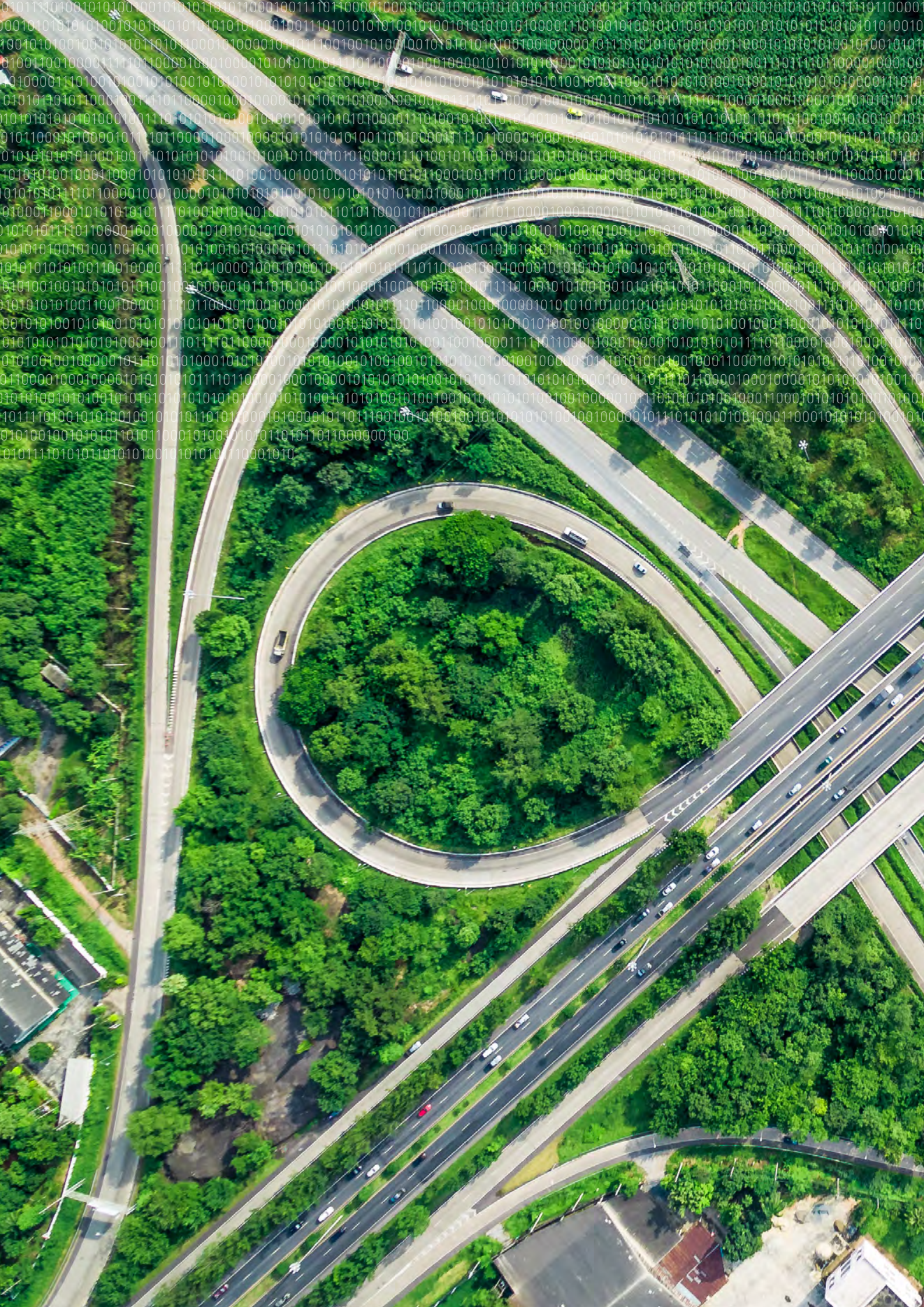
STATISTICS FOR THE KNOWLEDGE ECONOMY

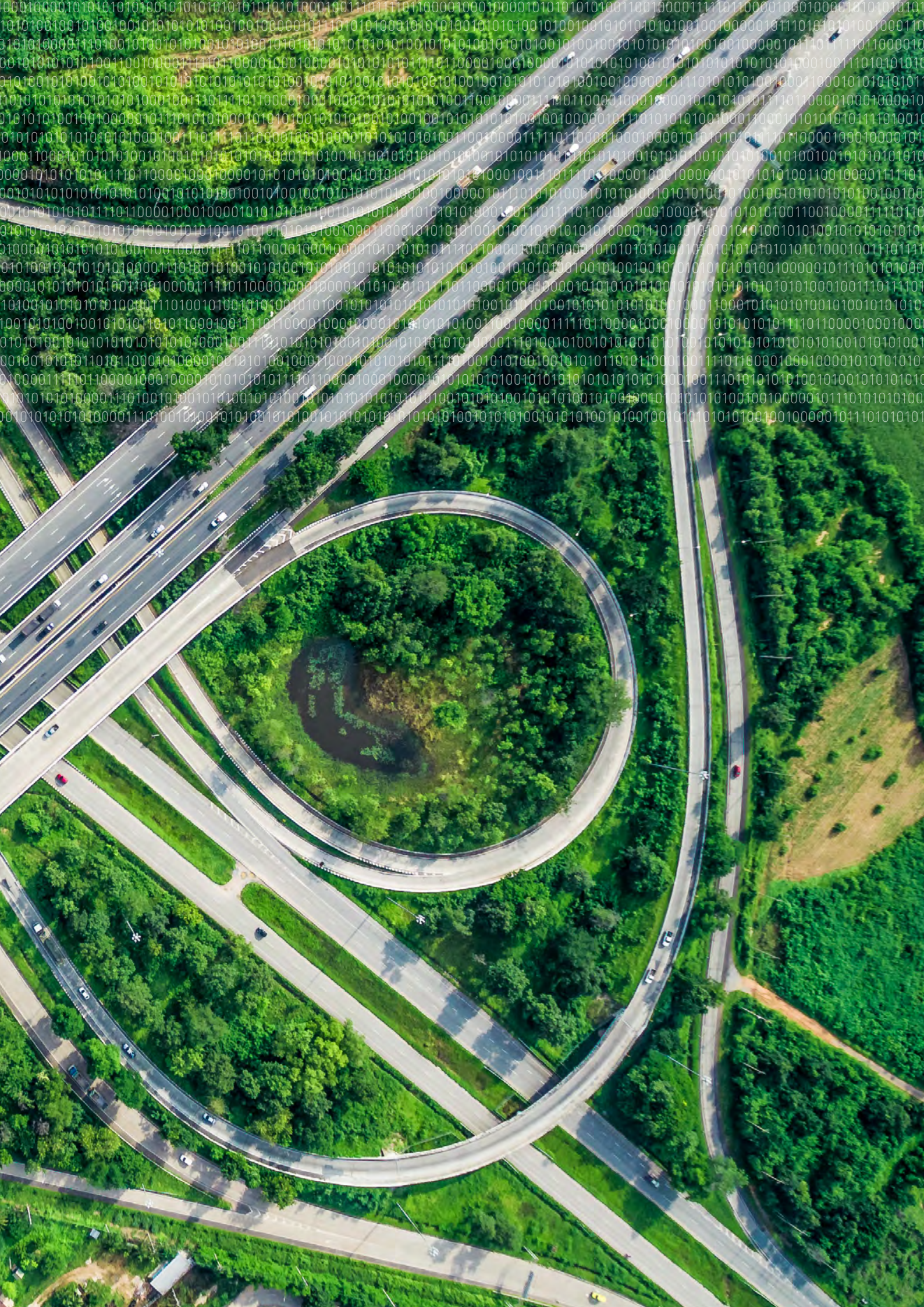
# ANNUAL REPORT 2019

sfi Centre for Research-based Innovation

The Research Council of Norway

BigInsight





# 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 problems 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, and more generally of contemporary AI. Digitalisation of the Norwegian industry and society benefits from BigInsight that produces powerful instruments for the analysis of data.

We discover radically new ways to target products, services, prices, therapies and technologies, towards individual needs and conditions. This provides improved quality, precision, value 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.

Research at BigInsight leads to value creation and strengthens our partners' leading position.
In the era of digitalization, BigInsight produces competence and capacity for the Norwegian knowledge-based economy, contributing to the development of a sustainable and better society.

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

*"If we interpret Artificial Intelligence as the development of methods and algorithms that make data driven predictions and discovers hidden processes, then BigInsight is historically the first Norwegian centre in AI. Proud of this!"*

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 big insight. Despite extraordinary advances in the collection and processing of information, much of the potential residing in contemporary data sources remains unexploited. The value does not reside in the data, which are often public, but in the methods to extract knowledge from them.

Digitalisation means producing data, organizing and storing data, accessing data and analyzing data. BigInsight works in this last direction. 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 statistics and machine learning. 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. Our solutions are explainable, fair and responsible.

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 courageous and creative, exploit knowledge and structure in complex data and integrate these from various sources.

Our research is open: we publish generic methodology and their new applications 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.

### 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 vessel at sea.

*"Artificial intelligence can be a great opportunity to accelerate the achievement of sustainable development goals. But any technological revolution leads to new imbalances that we must anticipate."*

Audrey Azoulay, UNESCO Director-General, 2019

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 objectives for and with their 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, personalised 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.

*"Because one thing is for sure: In the age of the algorithm, humans have never been more important."*

HELLO WORLD, Hannah Fry (2018)

# 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 2019, 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. The AI – Explanation & Law seminar series have brought all partners closer together and has resulted in more bilateral partner-to-partner cooperation across the Innovation Objectives, especially focusing on the field of explainable AI.

In November, we held the successful annual BigInsight Day at the premises of the Norwegian Institute of Public Health. The program included an overview based on short «fire talks», in-depth presentations and a debate about the AI-documentary iHuman to premier in March 2020.

# ORGANISATION

### Board in 2019

Marcus Zackrisson, Skatteetaten, chairman
Andree Underthun, ABB
Karl Aksel Festø, DNB
Hans Anton Tvete, DNV GL (from June)
Bjørn Johan Vartdal, DNV GL (until June)
Birgitte F. De Blasio, Folkehelseinstituttet
Erlend Willand-Evensen, Gjensidige
Plamen Mavrodiev, Hydro
Cathrine Phil Lyngstad, NAV
Lars Holden, Norsk Regnesentral
André Teigland, Norsk Regnesentral
Peder Heyerdahl Utne, Oslo University Hospital
Magnar Lillegård, SSB (from August)
Anders Holmberg, SSB (until August)
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 2019.
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
Arnoldo Frigessi, UiO
Ingrid Glad, UiO
Clara Cecilie Günther, NR
Ingrid Hobæk Haff, UiO
Alex Lenkoski, NR
Anders Løland, NR
Carlo Mannino, UiO
Hanne Rognebakke, NR
Magne Thoresen, UiO

### Administrative Coordinator

Unni Adele Raste, NR

### Scientific Advisory Committee (SAC)

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

BigInsight

BOARD

NR´S RESPONSIBLE PERSON — SCIENTIFIC ADVISORY COMMITEE

DIRECTOR

CO-DIRECTORS

PRINCIPAL INVESTIGATORS AND CO-PRINCIPAL INVESTIGATORS

INNOVATION OBJECTIVES

Personalised marketing

Personalised health and patient safety

Personalised fraud detection

Sensor systems

Forecasting power systems

Explaining AI

# RESEARCH STRATEGY

We aim to new, interesting and surprising solutions, which take the field and our partners ahead in their innovation strategy.

BigInsight's research is organized in six innovation objectives. Five innovation objectives (IOs) are centered on a concrete innovation area: marketing, health, fraud, sensor, power. The last IO is focusing on explanability of AI and data privacy.

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. Final transfer to partners' operations will happen both within and on the side of BigInsight.

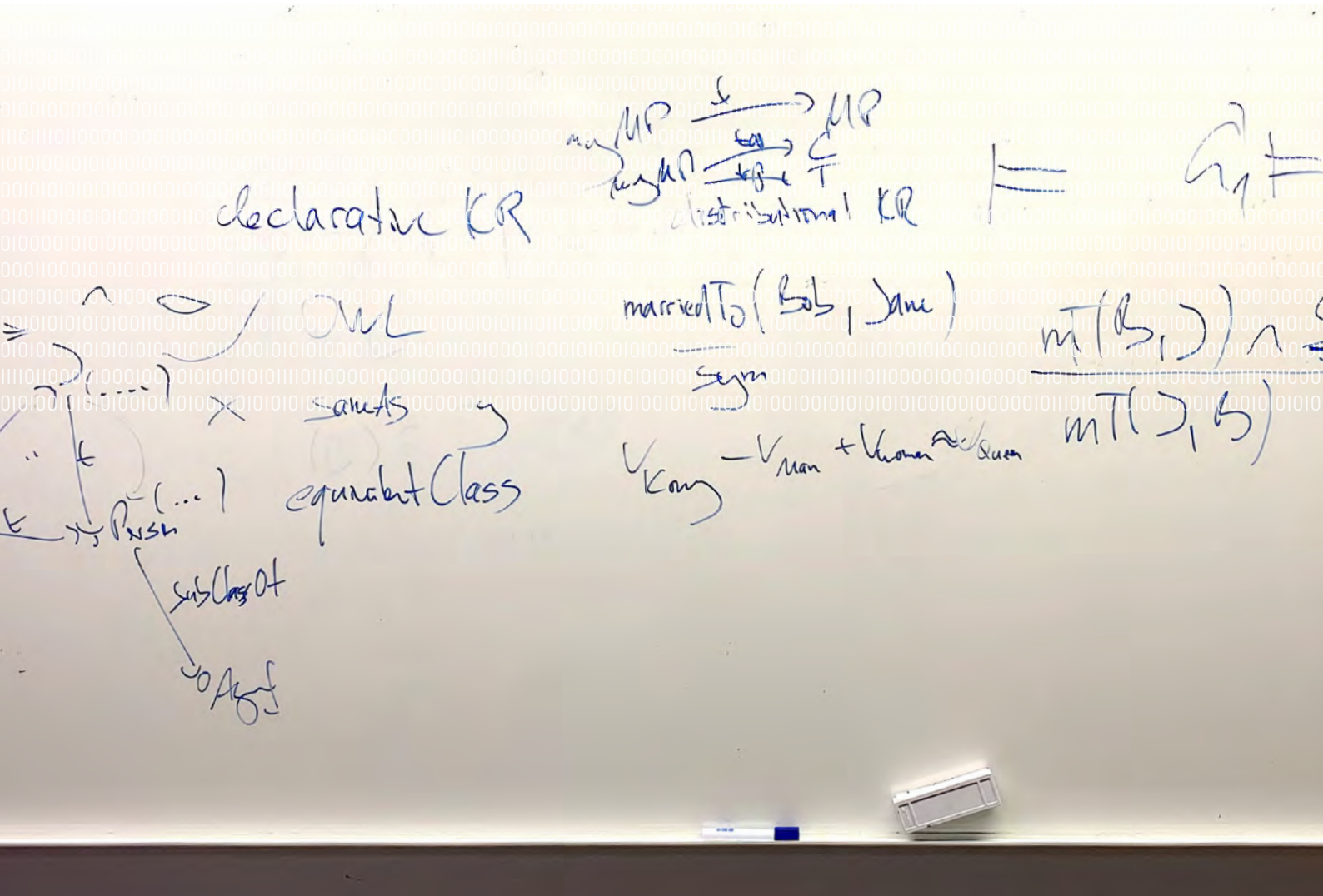
| | | | | | | |
|---|---|---|---|---|---|---|
| INNOVATION OBJECTIVES |  Personalised marketing |  Personalised health and patient safety |  Personalised fraud detection |  Sensor systems |  Forecasting power systems |  Explaning AI |
| INNOVATION PARTNERS | DNB<br>Gjensidige<br>NAV<br>Skatteetaten<br>Telenor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

# METHODS

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

**Our recent methodological results include:**

- Integrative analyses of complex multiple data sources, including integrative clustering and methods to investigate coordinated architectures across clusters in various data sets.
- High dimensional penalised regression, also assuming monotonicity, and with measurement error in covariates, with applications to genomics.
- Bayesian hierarchical models, including monotone multiple regression and cancer drug synergy prediction, with applications to insurance, drug screening, recommender systems and mortality data.
- Inference and prediction in multiscale models of stochastic differential equations in bio-mathematical models.
- Models for infectious disease processes, including the effect of urbanisation on epidemics and the use of mobility cellular phone data to describe the contact process.
- General methods to describe uncertainty in predictions, with applications to power market.
- Pair copula constructions for structure learning in Bayesian Networks for financial data.
- Models for the forming of social networks and inference from data in time.
- Anomaly detection algorithms with sequential build-up of anomaly evidence.
- Modelling interactions between actors inspired by infection diseases mathematical models.



# 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.

## BREAKING NEWS: BigInsight contributes to Covid-19 readiness and research

The world is shaken by the coronavirus disease (COVID-19) pandemic. In late February 2020, Norway experienced its first confirmed COVID-19 case. Since then, BigInsight is centrally involved in the COVID-19 modelling work at the Norwegian Institute of Public Health (NIPH). Together with the team at NIPH, we develop the models that allow understanding the present situation and allow predicting the future. Our models run every day on the SAGA cluster of USIT and on two specialized servers, we have been guaranteed. Ten researchers from UiO, NR, NIPH and Telenor are currently working full time, to help the Norwegian health authorities and the Norwegian government to make the right decisions. In addition, we have helped the press to understand the complexity of models: the reproduction numbers R0 that our models estimate are now a daily ingredient in the public discourse. The models that we are running today, would had never been ready, if BigInsight had not funded and started a project in 2017 with UiO, NIPH and Telenor as partners, on how to use mobile phone mobility data in the management of epidemics. Below we have printed this story, as reported recently in the media. For more information on BigInsight's work in the Covid-19 crises units, please contact Arnoldo Frigessi.

*"We see how important it is to be prepared. Without long term investments in research and development, nations, companies and services are not be able to act rapidly and efficiently enough."*

Arnoldo Frigessi, 15 April 2020,
WEBINAR: First Centre for Philosophy
and the Sciences Covid-19 Crisis Seminar

# Skal forutsi spredning av koronasmitte med mobildata

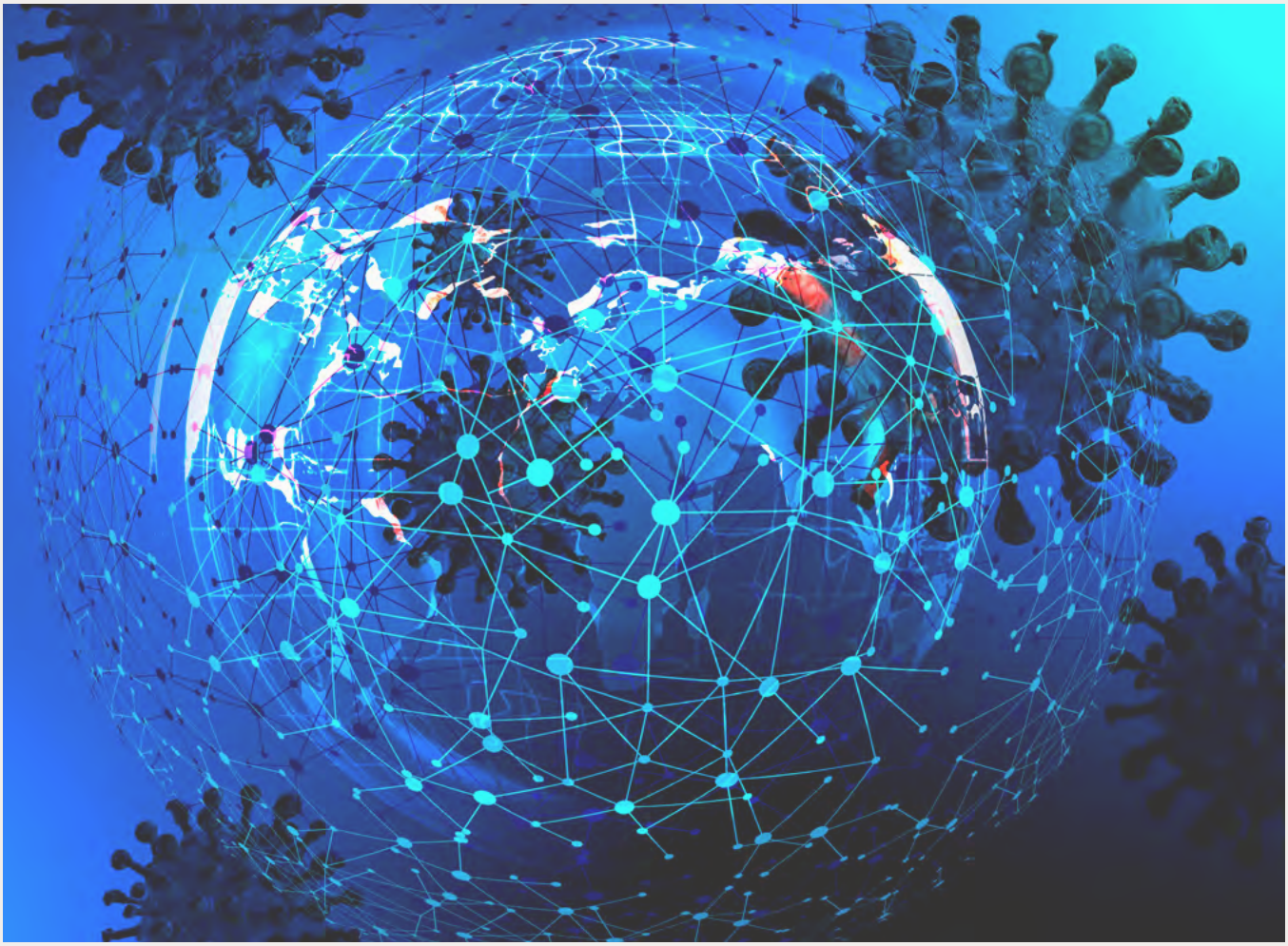
– Vi jobber på spreng for å lage en modell som kan forusti geografisk spredning av korona-viruset med utgangspunkt i mobiltdata, sier Arnoldo Frigessi, professor ved Universitetet i Oslo og Oslo Universitetssykehus. Han er statistiker og leder for Big Insight-senteret.

**Text: Siv Haugan, Forskningsrådet**. Published 16. march 2020

Big Insight-senteret er et Senter for forskningsdrevet innovasjon (SFI), finansiert av Forskningsrådet. Senteret samarbeider tett med næringsliv og offentlig sektor og utvikler modeller for statistisk analyse og maskinlæring.

Etter korona-utbruddet har de jobbet intenst med å finne en modell som ved hjelp av mobildata kan analysere folks bevegelsesmønstre og dermed forutsi hvor og når korona-viruset vil smitte. Modellen vil gjøre det mulig å vite hvor mange tilfeller det sannsynligvis vil bli på ulike steder i Norge fremover. Helsevesenet får dermed bedre tid til å forberede seg på belastninger som vil komme.

- Engbretsen, Frigessi og Kenth Engo-Monsen fra Telenor tok en midlertidig pause fra sine pågående prosjekter for å konsentrere seg fullt og helt om dette arbeidet, sammen med et større team på Folkehelsainstituttet ledet av professor Birgitte De Blasio (til venstre).



**Fra influensakartlegging i Bangladesh til korona i Norge**

– Sammen med Universitetet i Oslo, Folkehelseinstituttet og Telenor har vi samarbeidet om statistisk nettverksvitenskap. Vår tidlige ph.d.-student Solveig Engebretsen, som nå er forsker på Norsk regnesentral, er sentral i dette prosjektet. Vi har blant annet studert hvordan vi ved hjelp av mobildata kan forutsi spredning av influensa i tid og rom. I dette prosjektet fikk vi tilgang til å analysere data fra Bangladesh, forteller Frigessi.

– Vi var i gang med å utvikle en modell som kunne bruke mobiltelefondata til å forutsi spredningen av en normal influensa. Da covid-19 begynte å true også Norge, bestemte Telenor seg for å tilby oss mobilitetsdata fra Norge. Dette var et fantastisk mulighet, og vi er veldig taknemmelig til Telenor som gjør en stor jobb med å forberede data. Vi startet umiddelbart med å tilpasse Bangladesh-modellen til Norge og de nye mobiltelefondataene. Disse dataene er aggregerte i tid og rom for å garantere anonymitet, forklarer Frigessi.

Engbretsen, Frigessi og Kenth Engo-Monsen fra Telenor tok en midlertidig pause fra sine pågående prosjekter for å konsentrere fullt og helt om dette arbeidet, sammen med et større team på Folkehelseinstituttet ledet av professor Birgitte De Blasio.

– Det har vært et krevende arbeid med mye koding, svært komplekse data og mye usikkerhet. Modellen har nå blitt testet for første gang og Folkehelseinstituttet kan snart ta den i bruk sammen med flere andre modeller for å bidra til å ta besluninger for samfunnets beste.

Frigessi forteller at mobilitetsdata for mobiltelefoner er presise og gir sikre opplysninger om hvordan folk beveger seg.

Arnoldo Frigessi. Photo: Ola Sæther

**Sanntidsforskning**

– Allerede nå ser vi at folk beveger seg mindre, og dette er selvfølgelig gode nyheter for å bremse epidemiene. Mobilitetsdataene fra mobiltelefonene er unike, og gjør prognosene om spredning av covid-19 mye mer presise enn bruk av andre data for mobilitet, sier Frigessi.

For å gjøre beregninger starter vi fra de kjente covid-19-tilfellene, kalibrere de mot andre data fra covid-19 epidemien og så kjører vi tusenvis av simuleringer av fremtiden for å forstå statistisk hva vi kan forvente vil skje.

– Jeg har aldri likt skillet mellom grunnforskning og anvendt forsking, eller mellom nysgjerrighetdrevet forskning og resultat-drevet forskning. De fleste forskere er lidenskapelig nysgjerrige på ting de ikke vet, og samtidig enormt motivert av å bidra til en bedre verden. Forskningen vi nå holder på med er sanntidsforskning, med høy kvalitet, tidspress og er tverrfaglig, understreker Frigessi.

# 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 2019:

**Bankruptcy prediction using network data**

Company defaults may be seen as a disease, spreading directly or indirectly from one company to others. Network relations are therefore assumed to play a fundamental role. We have investigated the use of relational and transaction data for predicting defaults among DNBs customers. By using information about persons with key roles in companies and transactions between companies we create graphs and apply a method denoted weighted-vote relational neighbor to predict a score for each company based on the bankruptcy behaviour of its neighbours in the network. All the estimated network models are benchmarked against a base model using financial ratios as explanatory variables. We have also estimated an ensemble model combining network and financial information. The methodology and corresponding software have been transferred to DNB.

*"Artificial intelligence holds the key to delivering more human and relevant marketing experiences at scale"*

WPP May 2019

### Bayesian methodology for recommender systems

BigInsight has developed a new approach to recommendations, based on the Bayesian Mallows Model. The methods has been shown to perform as well as the industrial-state-of-the-art, but achieves a much higher level of diversity. In this way the catalogue of items is better exploited, something which is often very important. In addition, a more diverse personalization is often experienced positively by customers and users. In the last year, we have improved the basic algorithm, so that we can now perform recommendations at the needed scale. We have made progress on a Variational Bayesian approximation, which is very promising. The aim is to test the algorithm at finn.no or in another commercial context in the close future.

### Stochastic models for early prediction of viral customer behavior on networks

Early prediction of the success or failure of the adoption of new products has important economic implications. We propose a probabilistic method that accomplishes this task after having drawn inference from observing the adoption of the product on the social network of the customer base. Our stochastic model is at individual level, governed by both peer-to-peer viral influence and external factors, such as personal interest or marketing campaigns. Inference is by maximum likelihood, and prediction is performed by simulation with a computationally very efficient algorithm, which we have successfully tested on a Telenor product. In 2019 we have improved the algorithm, and developed a new way to estimate the strength of the virality of the product. We have started the process towards a possible commercialization of the idea, jointly between UiO, Telenor and NR.

### 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, predict online purchases and a better user experience. We have used 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.

### Explanation of predictions from Black-Box models

In many applications, complex hard-to-interpret machine learning models like deep neural networks are currently outperforming the traditional regression models. Interpretability is crucial when a complex machine learning model is to be applied in areas such as fraud detection or credit scoring. In Big Insight we have an innovation area denoted "Explaining AI", which focuses on explaining black box models. However, due to limited resources in this innovation area and increasing interest from the Big Insight partners on this theme, we have decided to use a part of the budget for the "Personalised Marketing" for research on explainable AI also. In 2019 we have submitted a paper on incorporating dependence into an explanation method denoted kernelSHAP, and we have written an R package, "shapr", which can be found on GitHub.

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 projects.

## What we did in 2019:

### Personalized cancer statistics

National population based cancer registries publish survival statistics by cancer site, stage, gender and time period, using established epidemiological 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. Hence, more individualized prognosis become feasible. In this project, we have developed methodology for estimating several measures of individual prognosis for cancer survivors. The techniques we use is based on several earlier papers in survival analysis by Ryalen, Røysland and Stensrud. Most of these measures are tuned for answering questions like: when can I expect a risk of death that is similar to what non-cancer patients have? We have applied this methodology to data from the Norwegian Cancer Registry with several different cancer types. One paper is submitted for publication.

### Personalized cancer therapies: Modelling cancer drugs sensitivity and synergy in in-vitro screening

Cancer pharmacogenomic screens profile cancer cell lines versus many compounds to identify new combinations of existing drugs that have a high probability to work on individual patients. We work with data generated by our partners at Oslo University Hospital and public data to guide therapy based on the statistical prediction of how drugs will behave for individual tumor samples. To improve predictions, we developed a structured penalised regression and complementary composite low-rank model and we are further exploring structured priors in multivariate Bayesian models to incorporate prior knowledge about the dependence structure between drugs and between multi-omics profiles of cancer cell lines. For combinatorial treatments, prediction of likely synergistic effects is crucial to suggest efficient combinations. We developed a flexible Bayesian model for improved estimation of drug interaction surfaces with current focus on building corresponding user-friendly software. Finally, we laid the groundworks for a scalable Bayesian multiple kernel learning framework that will allow us to model many experiments in a high-throughput drug screen jointly, and thereby make prediction of synergistic effects for new drug combinations or new cell lines based on similar previously performed experiments possible. We also started to develop a mathematical model of the drug screening experiment at cell level, which we hope can allow estimating the clonal composition of a cell sample. One paper is accepted and two are submitted. In addition, we have published two R-packages where we have implemented our methods.

### Healthcare safety management

As reported on last year, we have been developing a method for intervenable predictions in order to use electronic health records and other sources of routinely collected data in hospital, to predict the likelihood or level of hospital acquired infections in a given ward in such a way that it allows for identification of potential targets for intervention. We have been struggling with access to data from OUS and hence, the project has been delayed. We have now received data from Akershus University Hospital and the project will soon be completed.

While waiting for these data, we have been working on another project of corresponding type. The idea is again that there is an extreme amount of information available in electronic health records that can be used to make predictions, guide treatment choices and so on. This is a field very much dominated by deep learning approaches. We have been working on a different approach. Our idea is to combine dynamic time warping with powerful tensor decomposition techniques to come up with a prediction model that is more interpretable. The method has been tested on publically available data and the results are promising. A paper has been submitted.

*"The dominant theme of the Roadmap of Mathematical Oncology is the personalization of medicine through mathematics, modelling, and simulation. This is achieved through the use of patient-specific clinical data to: develop individualized screening strategies to detect cancer earlier; make predictions of response to therapy; design adaptive, patient-specific treatment plans to overcome therapy resistance; and establish domain-specific standards to share model predictions and to make models and simulations reproducible."*

Rockne et al, Phys Biol. June 2019

**Network theory for health**

PhD student Solveig Engebretsen defended her thesis; *Contributions to network science in public health*. The project "Time-aggregated mobile phone mobility data are sufficient for modelling influenza spread: the case of Bangladesh" has turned out to be essential during the first period of the Covid-19 crisis. Human mobility plays a major role in the spatial dissemination of infectious diseases. We develop a spatio-temporal stochastic model for influenza-like disease spread based on estimates of human mobility. The model is informed by mobile phone mobility data collected in Bangladesh. We compare predictions of models informed by daily mobility data (reference) with that of models informed by time-averaged mobility data, and mobility model approximations. We find that the gravity model overestimates the spatial synchrony, while the radiation model underestimates the spatial synchrony. Using time-averaged mobility resulted in spatial spreading patterns comparable to the daily mobility model. We fit the model to 2014-2017 influenza data from sentinel hospitals in Bangladesh, using a sequential version of Approximate Bayesian Computation. We find a good agreement between our estimated model and the case data. We estimate transmissibility and regional spread of influenza in Bangladesh, which are useful for policy planning. Time-averaged mobility appears to be a good proxy for human mobility when modelling infectious diseases. This motivates a more general use of the time-averaged mobility, with important implications for future studies and outbreak control. A paper is submitted.

A second project focused on the spreading potential of gonorrhoeae among men having sex with men. A novel sequence type of Neisseria gonorrhoeae (ST-7827) with a worrisome resistance profile emerged rapidly in Norway in the period 2016-2018, circulating in sexual networks of men who have sex with men (SMS), primarily in the Oslo area. To understand the rapid dissemination, we performed phylogenetic analyses and transmission modeling using the Norwegian strain collection 2016-2019 and international ST-7827 isolates retrieved from databases (PubMLST/PathogenWatch, and recently published sequences. Phylogeographic inferences were made using stochastic character mapping, and by constructing bootstrap trees, from which a consensus

tree was generated. We then performed root-to-tip regression and temporal analyses to make inference about ancestor dating using a Bayesian method (BactDating). Finally, we reconstructed genome-based transmission trees with the time-stamped phylogeny as input. A total of 251 genomes belonging to ST-7827 were included of which 148 samples were isolates from Norway. The phylogeographic analyses demonstrated an Asian origin of ST-7827 with multiple importation events to Europe. We estimated a reproductive number of 1.2 (CI 1.10-131) for the whole outbreak. In total, there were 24 individual pairs with a direct transmission probability larger than 50%, implying that the transmission chains could not be completely resolved. However, we identified successfully groups of patients with few intermediates between them:  A small cluster (C1) containing samples from Spain and Italy within the internal branches, suggesting multiple introductions to Norway; one larger cluster (C2) that contains only Spanish samples on a single basal branch, pointing to a single introduction that has been sustained by local transmission. Due to a scarcity of European samples in recent years excluding Norway, it is difficult to assess how widespread ST-7827 is. Our findings, however, are highly suggestive of a hidden reservoir existing in Europe. We plan to employ the results obtained in this study further in the development of an individual-based model for MSM to evaluate the effectiveness of various intervention strategies for Neisseria gonorrhoeae.

### Mathematical models and Bayesian inference in personalised cancer therapy

Personalised therapy means to determine which drug combinations, in what doses and according to what schedule, is optimal for the one patient who has to be treated now. BigInsight has been developing an innovative approach based on (a) mathematical models of cancer, of pharmacodynamics and pharmacokinetics; (b) Bayesian inference of parameters based on the full integration of all data available from the one patient; (c) a highly optimized set of algorithms which allow to explore a multitude of potential therapies in useful time. After a very successful proof of concept in breast cancer, we have started to work on blood cancer, again with partners UiO and OUS and in collaboration with mathematicians from the University of Minneapolis. Our first result has been a method which is able to determine how complex a cancer is, in terms of how many different groups of cancer cells (clones) are present in the patient, each responding differently to a drug. The response can range from total resistance to rapid death at low doses. We use data from drug screening of patients cancer cells to perform estimation. We will test the method further before using this snapshot of the cancer to recommend which drugs appear to be most appropriate. We have received important additional funding in 2019, with two European H2020 projects (on breast cancer and on head and neck cancer), a Digital Life Norway NFR project on hematological cancers.

Principal Investigator
Magne Thoresen

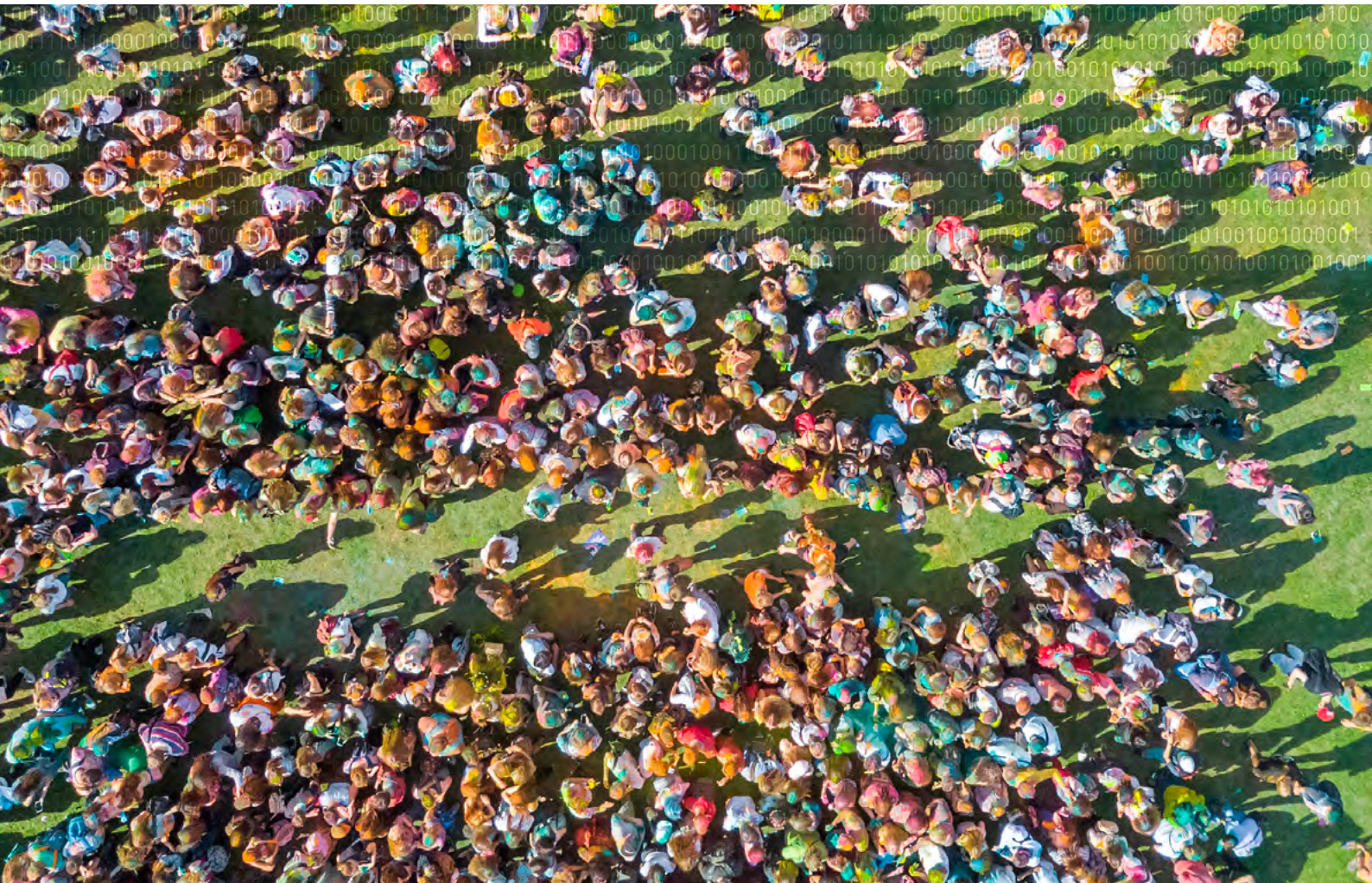
co-Principal Investigator
Clara Cecilie Günther

# 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 the 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 handle a diversity of data, including payment logs, relational networks, text and other available digital records, but under strict privacy protection regulations.

A further 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 accompanied by 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.



*"Combining supervised and unsupervised machine learning as part of a broader AI fraud detection strategy enables digital businesses to quickly and accurately detect automated and increasingly complex fraud attempts."*

FORBES August 2019

## What we did in 2019:

### Network analysis for fraud detection

Fraud can be seen as a disease, spreading directly or indirectly from one fraudster or one group of fraudsters to others. Network relations therefore play a fundamental role. The relations can be between people, businesses and groups thereof, often through financial transactions. The objective is to build these networks and extract useful variables from them so that statistical models can produce even better fraud forecasts and provide additional insight into how fraud spreads. We are pursuing methods based on word2vec/node2vec/metapath2vec framework, specifically for financial transactions in two data sets, namely 1) tax avoidance from Skatteetaten and 2) money laundering data from DNB, and the results are promising.

### 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 practice of excluding activities labelled as non-suspicious by experts leads to significantly worse performance. The method is being tested by DNB and a paper describing the approach has been published.

### Local Gaussian discrimination with discrete and continuous variables

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

### Sentiment analysis for fraud detection

Sentiment analysis is the use of natural language processing or text analysis to systematically identify, extract, quantify, and study affective states and subjective information. In the case of fraud, certain sentiments, like "impatient" or "unsatisfied", or the transitions between them could be a signal of fraudulent behaviour. In 2019, we have developed a method to predict sentiments of Gjensidige insurance chats. Chats are instant messages that Gjensidige customers can use to ask questions to customer service. Predicting sentiments is a difficult problem, since even humans can disagree on which sentiment(s) that can be found in a specific text, but the results are promising and the method is being tested by Gjensidige.

### Fraud detection based on the fraud-loss

In fraud detection applications, the investigator is often required to efficiently allocate limited resources. This amounts to selecting a restricted number of cases that are most likely to be fraudulent, or most worthy of investigation. The set of cases to be investigated should be determined from the predicted probabilities from the chosen model. In this respect, we have a precise notion of what a good or bad model is for this purpose, namely one that lets us pick a certain number of cases, such that as many as possible of these are actual cases of fraud, and we term this notion fraud-loss. We have proposed a framework for choosing the best model according to the fraud-loss. The results are promising, and a paper will soon be submitted.

Principal Investigator
Anders Løland

co-Principal Investigator
Ingrid Hobæk Haff

# SENSOR SYSTEMS

Sensor data are multidimensional streams of observations from various sensor systems. In this IO we work mainly on sensor systems in the maritime sector, but as Statistics Norway has joined BigInsight last year, we consider their activity as ´sensoring´ society, and therefore include the research with SSB in this IO.

For maritime safety surveillance we develop new approaches based on the availability of large arrays of sensors, which monitor condition and performance of vessels, machinery and power systems. Sensor data are becoming increasingly available on global ship fleets, with efficient broadband connectivity to shore. We suggest new approaches to condition and/or performance 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. We also rely on other data sources such as AIS data for the study of maneuvers and collision avoidance.

## What we did in 2019:

### Scalable change and anomaly detection

Industrial-PhD Andreas Brandsæter at DNV-GL delivered his thesis "Data-driven methods for multiple sensor streams, with applications in the maritime industry" in the end of 2019. The thesis contains five papers, among which four are focused on sensor-based monitoring and classification. Among these, one paper on efficient on-line anomaly detection for ship systems in operation and one paper on unsupervised anomaly detection based on clustering methods were published in 2019. The thesis also contains one paper suggesting a new method aimed to evaluate how different subsets of training data explain individual black box predictions. This will be the first disputation in the Sensor Systems IO, taking place in March 2020.

Another PhD has worked mainly on the problem of detecting sparse changes in high dimensional sensor data. One paper on dimension reduction with PCA specifically for change point problems was published in 2019. This PhD student had a longer research visit to our collaborating center at the University of Lancaster in the autumn of 2019, starting to work on a related, new project on how to build dependencies between components (cross-correlated data) into a recent method for multivariate collective and point anomaly detection. Collective anomalies correspond to intervals where one or more of the data streams behave anomalously. These methods are highly scalable both in the number of sensors and in the number of time points and will be relevant for many sensor data sets in the Sensor systems IO.

### Combining AI and expert knowledge for more efficient monitoring

One of the 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 based on a case from ABB, the main problem is how to extract features from observed sequences (including time points) which are informative with respect to failures. This project has made strong progress in 2019. We are about to finalize a paper on fault prediction and classification for such categorical streaming data, along with one paper on the approximation of posteriors in Bayesian mixture models with an unknown number of components. The PhD student in this project is collaborating with ABB researchers on the implementation of these methods in the ABB solutions.

A master thesis started up in the end of 2019, where, on ship level, we will use all available sensor data in combination with logged error messages, to find out if it is possible to build a predictive detector for one specific event (critical

trip). Data are in this case labelled. We also used time in 2019 on a reading group preparing a new project focusing on weak labelling of sensor data using external information, which will start in 2020.

### Autonomous vessels test beds from AIS (traffic) data including collision avoidance rules

With autonomous vessel algorithms comes the fundamental need for realistic testing in complex navigational situations. In collaboration with DNV-GL, we have designed realistic navigation testbed scenarios from huge traffic data integrated with high resolution digital maps, vessel information registries, and digital nautical charts. In a first study, an adaptive ship-safety-domain with spatial risk functions was proposed to identify both multi vessel collision and grounding real-time risk based on motion and maneuverability conditions for all vessels. The algorithm is designed and validated through extensive amounts of Automatic Identification System (AIS) data over a large area and allows for real-time situation risk identification at a large-scale up to country level and up to several years of operation with very high accuracy. A first paper was accepted in 2019.

In a second study, collision avoidance rules (COLREGs) have been integrated in the algorithms, allowing all vessel-to-ground and vessel-to-vessel interactions to be efficiently analysed through a hierarchical method for collision and grounding conflicts, assessed with a 15-minutes prediction horizon. Relative risk is evaluated precisely over full periods of predicted close-quarters situations subject to physical limits and space availability for evasive maneuverers under COLREG rules and traffic separation restrictions. Spatial dependencies between multiple nested conflicts create complex momentary traffic situations which, through temporal dependencies, generate complex, realistic scenarios to be parameterized, filtered, classified and prepared for implementation as test beds. A paper is in preparation and various sample scenarios are under implementation by DNV-GL in Trondheim.

Also based on AIS data, we contribute to the compliance verification of the maritime collision avoidance rules using large amounts of historical navigation behavior data to make statistics on how the rules are practiced in real situations. A report has been written in collaboration with DNV-GL.

*"Sensors are rapidly bringing us to a place where we can gather, synthesize, and understand enormous amounts of data very quickly... and thus provide more accurate predictions and insights tied to the world around us"*

WIRED, The Sensor-Based Economy

**Towards zero emission vessels – li-ion battery diagnostics and prognostics**

A new PhD started in September 2019, opening a new project with DNV-GL and their collaborator Corvus Energy, which is a major producer of lithium ion battery systems for large maritime constructions, for example operating in several fully electric or hybrid ferries in Norwegian fjords. The aim of the project is to develop statistical and machine learning methods for data driven monitoring of various measures of battery health and remaining life, based on historical data from operating vessels, provided by the battery producer.

**Predicting propulsion motor heating using machine learning**

In 2019, the successful method for propulsion motor overheating prediction we have developed with ABB, has been extended in order to use training data across different types of vessels. We have also refined and retrained the method on new, high frequency data (seconds) and provided it with an adaptive cumulative sum (CUSUM) method for the anomaly detection part. Using data from a real fault case, the monitor is shown to alert between 30 to 60 minutes before the fault and is able to detect emerging faults at temperatures well below the current alarm limits.

We have also started a master project in the end of 2019 in order to investigate if it is possible to further improve the predictions using recurrent neural net algorithms (LSTMs).

**Fast and computationally cheap emulator for hull condition monitoring**

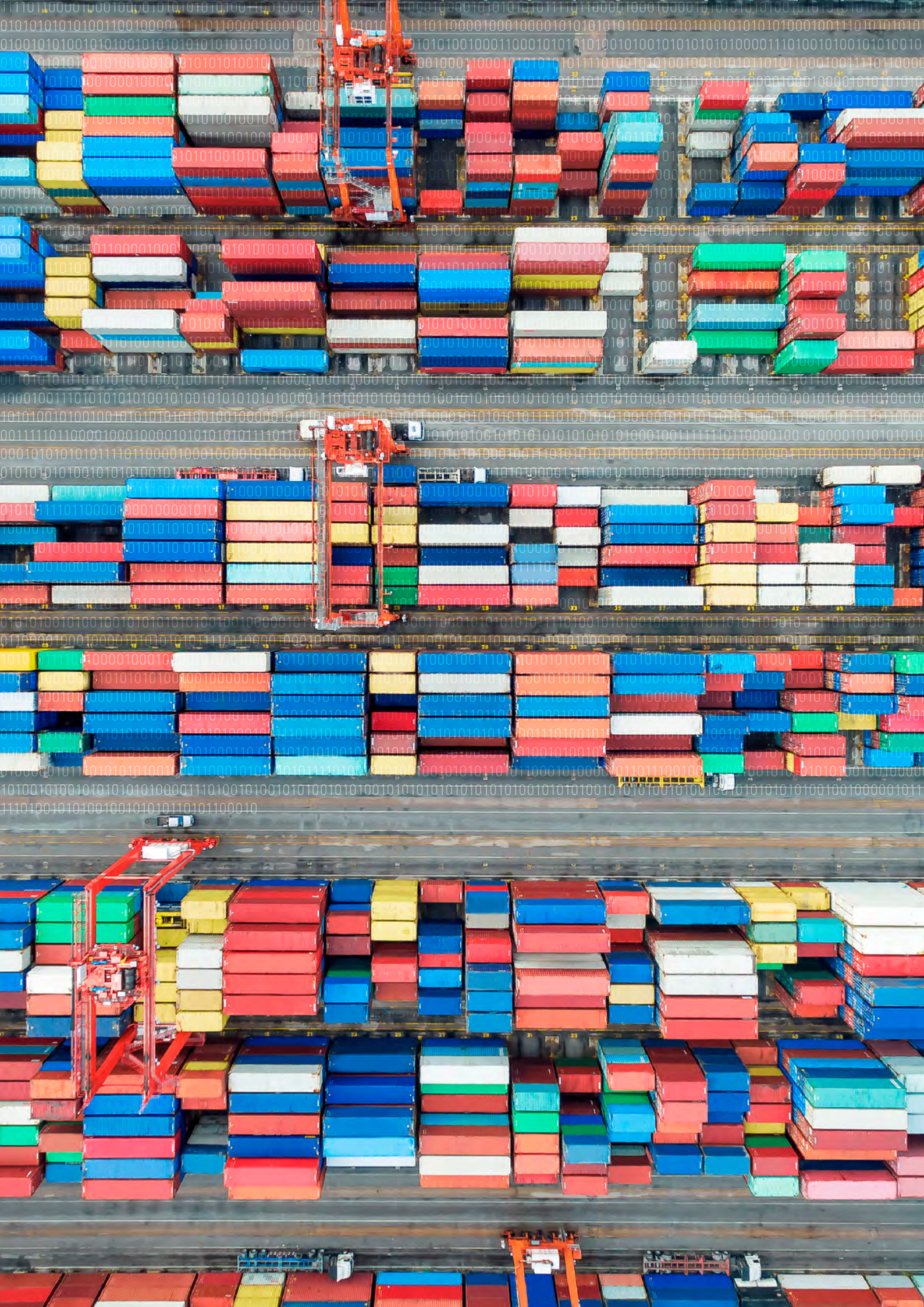
In 2019, we have finalized a project on statistical approximation to synthetic midship hull stress response with DNV-GL. Combined hydrodynamic and structural models are used to simulate structural responses on ship hulls in a seaway for design and risk assessment purposes. From a safety and inspection perspective, there is demand for continuously monitoring the ship hull conditions to estimate the structural utilization in real time. However, setting up the computer models and running the analysis are time consuming and costly, preventing such models from being used operationally. We have developed a statistical model that approximates the wave bending moment output from the computer model, which is computationally cheap and much faster to run than a hydrodynamic model, and may thus act as a virtual indicator sensor for structural condition monitoring. A journal paper was published in 2019.

**Protocol for combining data sources with misclassifications, maintaining privacy**

Due to problems with data availability, there has been limited progress in the commuting pattern estimation project we started in 2018 with partner Statistics Norway (SSB). In parallel, in 2019 we have worked on estimating the proportion of different, non-overlapping, classes in the population (currently employed/unemployed due to data availability, but could be all types of classes) using data from two parties, where both parties have data with misclassifications. The goal is to develop a "protocol" for how Statistics Norway can receive data from external companies in a way that safeguards privacy as much as possible. This work will continue in 2020.

Principal Investigator
Ingrid Glad

co-Principal Investigator
Hanne Rognebakke



# 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 2019:

### Finalized work on price spike warning systems

In 2018 we finalized a Bid/Ask model that enabled adaptive and highly non-symmetric price distributions to be created for electricity price forecasts. In developing this mode, we interacted extensively with our industrial partner Hydro. The Bid/Ask model is quite general, in that it issues predictive distributions and it is up to the users to decide what to subsequently do with these distributions. The most pressing matter that Hydro used these distributions for was to construct a "warning system" that issues a probability that prices will exceed some high threshold. We finalized this methodology in 2019 and have incorporated it into Hydro's systems

### Developed new models for electricity demand prediction

Predicting the total demand for electricity in every region of Nordpool, as well as neighboring markets such as Germany, is the key input into electricity price forecasts. The current system used at Norsk Hydro relies on a neural network estimation system that was developed in the 1990's.

We collected data on historical temperature forecasts (temperature being a main driver of electricity demand) and tested various modeling strategies for forecasting electricity demand. We found that a system based on principal components regression outperformed the current method used at Norsk Hydro.

Plans for 2020 include incorporating this new model inside the partner's forecasting system and writing an academic paper that outlines this methodology.

### Developed new models for renewable power generation forecasting

Another critical component of electricity price determination is the amount of power generated from renewable sources, in particular wind and solar power. Therefore, accurate forecasts of these quantities have a substantial impact on the quality of price forecasts.

We investigated models that use wind speed and solar irradiation forecasts to forecast renewable power generation. We found that ridge regression methods on a "cube" of weather forecasts outperformed more standard regression methodologies. We again plan on incorporating these

models into the price forecasting system used by Hydro as well as writing an academic article discussing these results.

**Developed new system for weather forecast trajectory correction**

As noted above, weather forecasts play a crucial role in forecasting electricity prices. These forecasts themselves are updated on a regular (6 hourly) basis. However, in the intervening period between successive model updates, observed weather can be used to update forecast trajectories, enabling market actors to more quickly update their forecasts of prices.

We developed a new methodology, called Rapid Adaptation of Forecast Trajectories (RAFT) which implemented this idea and showed that it can improve weather forecasts substantially. This has resulted in a paper accepted in the Quarterly Journal of the Royal Meteorological society, to be published in 2020.

Principal Investigator
Alex Lenkoski

co-Principal Investigator
Carlo Mannino

*"The AI transformation in the energy industry will directly influence the international energy stability and economic prosperity"*

Eva Pongrácz, University of Oulu.

# EXPLAINING AI

At the intersection between artificial intelligence, transparency, privacy and law, there is a need for more research. This IO, which started up during 2018, now focuses on explaining AI or black box models and related issues.

Artificial intelligence, statistical models or machine learning models can often be seen as black boxes to those who construct the model and/or to those who use or are exposed to the models. This can be due to: a) Complicated models, such as deep neural nets, boosted tree models or ensemble models, b) Models with many variables/parameters and c) Dependencies between the variables.

Even simple models can be difficult to explain to persons who are not mathematically literate. Some models can be explained, but only through their global, not personalised, behaviour. There are a number of good reasons for explaining how a black box model works for each individual:

1. Those who construct or use the model should understand how the model works
2. Those who are exposed to the model should, and sometimes will, have the right to an explanation about a model's behavior, for example to be able to potentially contest its decision
3. It should be possible to detect undesired effects in the model, for example an unfair or illegal treatment of certain groups of individuals, or too much weight on irrelevant variables

*"As the financial implications and economic fallout of COVID-19 become more lucid around the world one thing is already clear: Many people will need loans to survive. And almost all loan decisions are determined using proprietary black box models. This is a problem."*

IMPACT, Duke University, April 2020.

Research at BigInsight can challenge some of the legal principles that govern data privacy, including the risk of re-identification of anonymised parties, the wish to minimise data made available to discover associations and causes and the uncertainty of the value created by big data research. The need for compromising between privacy protection and common good is particularly evident in medical research. Methods and algorithms should follow the five principles of responsibility, explainability, accuracy, auditability and fairness. How can these aspects be regulated, validated and audited?

## What we did in 2019:

### Seminar series

We organized two seminars, where BigInsight researchers and partners presented and discussed their views on themes related to anonymization and counterfactual explanations: "Three views on anonymization" and "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". Attendance and discussions were very god.
We also co-organised the one day seminar "Responsible use of Data and AI" at DNB.

### Correct explanations when there is dependence between the variables

In many real life models, some or many of the variables of interest are dependent. For example, income and age typically follow each other quite closely. Current approaches to individual explanations do not handle dependent variables at all or not very well, especially in terms of the computational burden needed even for a handful of variables. We have been constructing new methods to handle these situations and have submitted a paper on our new method. We have also written an R package – shapr – and a paper describing the package code has been published. We have started to improve our methods further in terms of efficiency, especially to handle categorical variables better.

### Practical testing of explanations on use cases

Even though the explanation methods we develop are mathematically sound and correct, it is not obvious that they are immediately useful for executive officers or end users. We will therefore investigate how test groups understand these explanations, to learn and further develop how the explanations can be explained or utilised better. We are working with NAV on these issues. Further BigInsight partners can follow suit in 2020.

### Other activities

We have been and will continue to be an important voice in the Norwegian AI debate. We have started a collaboration between DNB, OSLOMet, UiO Department of Public and International Law, which will lead to an industrial PhD. We are organizing a working group on methods for synthetic data. The group currently includes SSB, NAV, Skatteetaten, DNB, Lånekassen and Riksrevisjonen.

Principal Investigator
Anders Løland

co-Principal Investigator
Arnoldo Frigessi

# INTERNATIONAL COOPERATION

International Academic Partners are key resources for BigInsight. We collaborate in research and co-supervise PhD students. We organize joint workshops and events.

## International Academic Partners

**STOR-i, Statistics and Operational Research in partnership with Industry, University of Lancaster**

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 David Leslie co-supervise PhD students together with BigInsight staff, on recommender systems, reinforced learning, multivariate extremes, non-parametric isotonic spatial regression, Bayesian modelling, multivariate sensor data, pair copula models. BigInsight and STOR-i co-organise industrial statistics sessions in international conferences and exchange membership in each other's scientific advisory boards. STOR-i has recently been renewed until 2023, also thanks to the strong links to BigInsight. PhD student Martin Tveten spent 4 months at STORi to work with professors Eckley and Fearnhead. PhD student Simen Eide is co-supervised by professor Leslie. Frigessi is co-supervising STORi PhD student Anja Stein.

Christian Rohrbeck, who was PhD student at STOR-i, spent research time in Oslo, and now is lecturer at the University of Bath.


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

### 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, appli-cation 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 innova-tion, 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 molec-ular biology. The prestigious Aker Scholarship has been awarded to master student Camilla Lingjærde who starts in 2019 a PhD at the BSU. We have also involved the BSU in our collaboration with the University of Hawassa (Ethiopia). PhD student Zhi Zhao spent a research time at BSU in 2019.

Professor Sylvia Richardson, MRC Biostatistics Unit, Cambridge

Professor Sylvia Richardson was awarded with an honorary doctorate degree from the University of Oslo in 2017.

### The Department of Mathematics, University of Minneapolis, USA

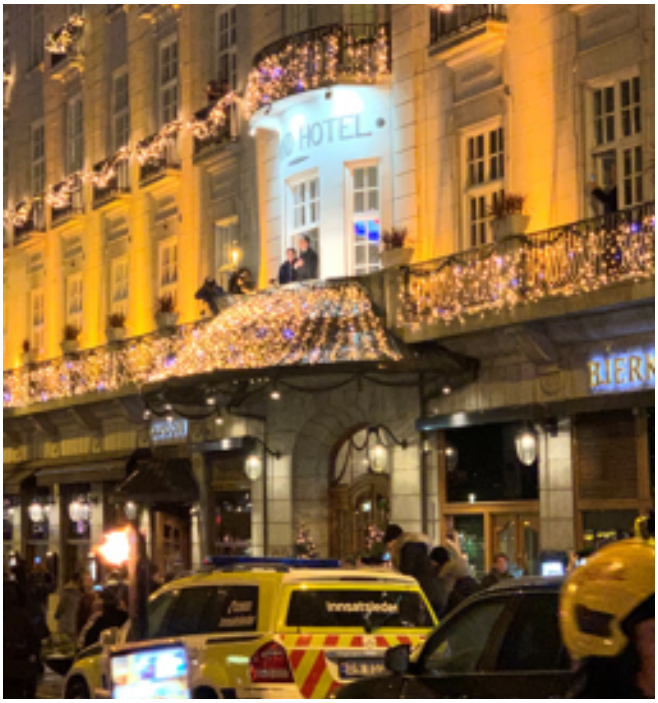
This collaboration started in 2018 when Professor Jasmine Foo and associate professor Kevin Leder spent a year at BigInsight, working at the interface between mathematics, cancer biology, clinical oncology, machine learning and statistics. The scientific core of this collaboration is the development of new methods for integrating patient data into mathematical models of cancer, contributing to better treatment for cancer patients. In addition we will develop new educational opportunities in mathematical modelling of cancer at the bachelor's, master's and PhD levels at UiO and UMN. The collaboration has recently been supported by an INTPART NFR funded project that BigInsight obtained.

Professor Jasmine Foo

University of Minneapolis

### University of Hawassa, Ethiopia

Funded by Norhed, the Norwegian Agency for Development Cooperation NORAD, NTNU, UiO and BigInsight are the Norwegian partners of a project with the Universty of Hawassa. We supervise PhD students who defend their thesis in Hawassa, but spend about a year at UiO. In 2019 we have welcomed PhD students Edossa Merga, Gezahegn Mekonnen and Tizita Geremew, whose research theme is extreme claims in insurance, clinical trial on liver diseases, clinical trial on pediatric heart pathologies. We also hosted two PhD students from the University of Jimma, funded by a different Norwegian development project: Teshome Kabeta and Henok Asefa, working on dietary diversity in rural Ethiopia and non-communicable diseases in Addis Abeba.




In 2019, the Ethiopian prime minister Abiy Ahmed, was awarded the Nobel Peace Prize, and the Ethiopian students were in Karl Johan lightning torches of celebration. Here Gezahegn Mekonnen and Ingrid Glad Frigessi.

### International guest programme

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

## In 2019 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 and recommender systems and health.

**Professor Jasmine Foo**, **University of Minnesota,** has a sabbatical year (2018-2019) at BigInsight. She collaborates on BigInsight projects related to personalized cancer treatment, with particular focus on mathematical models for blood cancers. Fulbright Scholar Grant.

**Ass. Professor Kevin Leder**, **University of Minnesota**, has a sabbatical year (2018-2019) at BigInsight. He works with BigInsight on projects related to the mathematical study of drug synergies for cancer, including important optimization aspects. Fulbright Scholar Grant.

Professor Gianpaolo Scalia Tomba

Professor Emeritus Elja Arjas

Professor Jasmine Foo,

Ass. Professor Kevin Leder

## Other International activities

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

In 2019 we welcomed: Polina Arsenteva master student from the University of Paris Descartes, studying for a master's degree in applied mathematics with the specialty in "Mathematical models in life sciences".

**BigInsight is partner and co-coordinator of the H2020 EU project:**

**RESCUER: RESISTANCE UNDER COMBINATORIAL TREATMENT IN ER+ AND ER- BREAST CANCER**

Breast cancer is the leading cause of cancer-related death in women. Breast cancer is classified into well-recognised molecular subtypes. Despite established molecular classification of tumour subtypes, only some patients benefit from administering drug combinations, which is an indication of tumour heterogeneity. The EU-funded RESCUER project aims to develop a new approach and identify mechanisms of resistance at systems level, exploring how the treatment is affected by patient- and tumour-specific conditions. The project will integrate longitudinal multidimensional data from ongoing clinical trials and novel systems approaches, which combine subcellular/cellular and organ-level in silico models to discover molecular signatures of resistance and predict patient response to combinatorial therapies. This new knowledge will be used to identify already approved drugs with a high curative potential of new personalised drug combinations.

**BigInsight is partner of the H2020 EU project:**

**BD4QoL: Big Data Models and Intelligent tools for Quality of Life monitorinBig Data Models and Intelligent tools for Quality of Life monitoring and participatory empowerment of head and neck cancer survivors**

The number of treatment options available for head and neck cancer (HNC) has increased in the last decade thanks to advanced technologies. While current post-treatment care plans focus on functional and health conditions, there are socioeconomic determinants of quality of life that also need to be addressed. The EU-funded BD4QoL project aims to improve HNC survivors' quality of life by developing a person-centred monitoring and follow-up plan. It will use artificial intelligence and Big Data collected from mobile devices, in combination with multi-source clinical and socioeconomic data and patients' reported outcomes. Analysis of the quality of life indicators collected over time will facilitate early detection of risks, prevent long-term effects of treatment, and inform patients and caregivers for personalised interventions.

**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 (NIPH) is nominated as deputy. This EU action started in 2016 and aims to facilitate interaction between diverse groups of statistical network modelers, establishing a large and vibrant interconnected and inclusive community of network scientists. This programme is ends in 2020.

## Scientific Advisory Committee of BigInsight

Scientific Advisory Committee of BigInsight has five international members.
The next meeting will be in 2020, though probably virtual because of COVID-19.

**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
- Director, Finnish Center for Artificial Intelligence FCAI, 2018-
- 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, Statistic Genetics, Geoscience, Linguistics and Archaeology

**Prof. Marina Vannucci**, Rice University, 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
- Former Editor-in-Chief for the journal Bayesian Analysis
- President, International Society for Bayesian Analysis

**Reader Veronica Vinciotti**, Brunel University of London, UK
- 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

# PHD GRADUATES 2019

In 2019 the following PhD students affiliated to BigInsight defended their PhD thesis:

**Xiaoran Lai** at Institute of Basic Medical Sciences defended the thesis "Modelling, inference and simulation of personalized breast cancer therapy" for the degree of PhD.

**Adjudication committee**

- Professor Miguel A. Herrero, Universidad Complutense, Madrid, Spain
- MRC Rutherford Research Fellow Kathleen Kit Curtius, Queen Mary University of London, UK
- Group leader, NCMM Marieke Kuijjer, Faculty of Medicine, University of Oslo

Trial lecture: "Mathematical Methods in Oncology"

Richard Xiaoran Lai presents his trial lecture with impeccable authority.

**Solveig Engebretsen** at Institute of Basic Medical Sciences defended the thesis "Contributions to network science in public health" for the degree of PhD.

**Adjudication committee**

- Professor Tom Britton, Department of Mathematics, Stockholm University, Sweden
- Professor Vittoria Colizza, Pierre Louis Institute of Epidemiology and Public Health, Sorbonne University, France
- Post doctoral Fellow Ernst Kristian Rødland, Faculty of Medicine, University of Oslo

Trial lecture: "Modelling Social Networks"


Solveig Engebretsen did not know that her thesis, successfully defended in November 2019, would become a fundamental part of the Norwegian surveillance of Covid-19: here at the Norwegian Institute of Public Health with colleagues Francesco Di Ruscio and Alfonso Diz-Lois Palomares.

# ACTIVITIES AND EVENTS

**2019 BigInsight Workshop**

The yearly BigInsight Workshop was held on November 14th at the Norwegian Institute of Public Health (NIPH). Thanks to all who participated with excellent "fire talks" (5 minutes each!), in-depth presentations. We showed some clips from the documentary iHuman, and had a debate about it also with 5 young bachelor students in philosophy and mathematics from UiO.

Camilla Stoltenberg, Director of NIPH, welcomed us with an interesting talk about the history of the institute and the centrality of statistical science for modern public health.

## Oslo Data Science Day

The University of Oslo Data Science Day 2019 on October 16th attracted more than 250 people and was a definite success with interesting presentations, stands from 14 companies, food, soft drinks and mingling.

The Vice Rector for Reseach and Innovation at UiO Per Morten Sandset gave an opening speech and SIRIUS PhD Summaya Mummtaz was chairing the whole event.

Speakers were BigInsight´er professor Geir Storvik (about Bayesian Machine Learning) and Workplace by Facebook´s Lauren Edelson, who gave a fascinating talk about biases in AI.

Nine BigInsight partners were present with their HR recruiting offices.


SIRIUS PhD Summaya Mummtaz was chairing the event

Data Science day at UiO is a unique event, with participants ranging from junior bachelor students to established professionals in the AI industry.

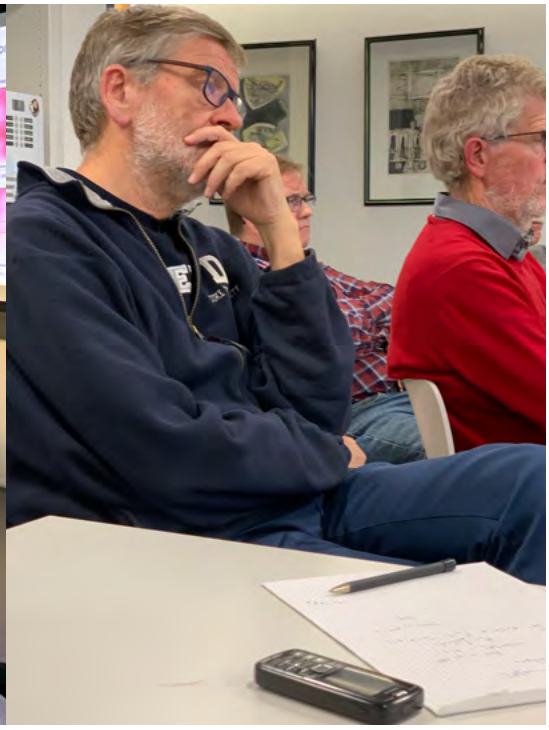
## Seminars

BigInsight's biweekly Wednesday lunch takes place at the Department of Mathematics and NR alternatingly. In 2019 18 lunches were organized, see our webside for a list of invited speakers. 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.

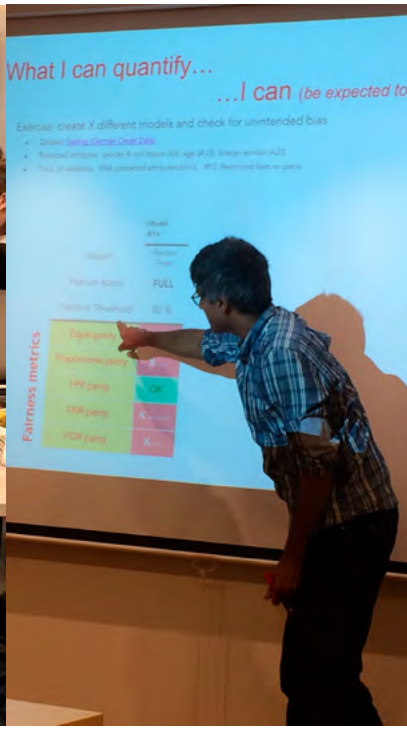
The Tuesday statistics seminar at the Department of Mathematics, co-sponsored by BigInsight, is a traditional semi-weekly seminar for the whole statistics community in the Oslo area. Speakers from abroad have been often invited.

The Biostatistics Seminar on Thursday is now merged with the Sven Furberg Seminars in Bioinformatics and Statistical Genomics. The seminars are hold at OCBE, the Department of Informatics and at NCMM. 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.

All events have seen an exceptionally large participation, so that we can proudly say that these Oslo seminars are among the best attended statistics seminar in Europe.



Socializing is an important activity at BigInsight, where we have the most wonderful colleagues



# TRAINING AND COURSES

BigInsight was instrumental in the establishment of the new Master Program in Data Science at UiO, which started up in 2018. Admission to this master program requires a bachelor with at least two statistics and two computer science courses, plus a solid mathematical foundation, and as such it is different from many other competing programs in Norway, which do not have such requirements. The focus of the master courses is on methods, algorithms and data analysis pipelines, with less focus on the use of available tools, because we believe that understanding the principles and foundations of data science is what will allow students to remain competent also in the future. There has been an immense interest for this program with many hundred applications both in 2018 and 2019, but only around 15 of these have been admitted each year. The first batch of Data Science masters will finish their degree in the summer of 2020, and some of them will move on to PhD work in BigInsight. BigInsight participates to the master program by teaching, master projects and industrial contacts.

BigInsight staff supervise MSc projects in data science, and also in the more traditional master in statistics. When possible, we couple these projects to an on-going PhD project, so that the PhD student can participate to the supervision.

Some PhD students work as teaching assistants, and in the final year also as teachers, in our courses, also at the Faculty of Medicine. Postdocs have teaching duties occasionally, and participate in supervision of master and bachelor students.

The Faculty of Mathematics and Natural Sciences started a pilot internship course for bachelor students across all fields of specialization in 2019. BigInsight staff has helped in the shaping of this course, using experience from contact with partners in the centre. The pilot was very successful in 2019, but has been cancelled for 2020 due to the corona virus situation.

Thanks to BigInsight, there is a large cohort of PhD students at the Department of Mathematics, and at the Oslo Centre for Biostatistics and Epidemiology, which allows organising more courses and activities for them. Supervision of PhD students includes experts from the partners and the students often have direct and continuous contact with the partners.

Of strong importance to the students society, BigInsight has fully or partly arranged and/or sponsored arrangements like Data Science Day 2019 at UiO, BigInsight Day 2019 at FHI and the yearly Klækken PhD workshop in 2019.

Many PhD students contribute to the advising services in statistics, biostatistics, bioinformatics and data science, which we offer to researchers at UiO and OUS. They follow an experienced advisor, before they advise on their own (with behind the scene support if needed). We offer a drop-in advising service and a more long term support. In this latter case, students are often co-authors of a research paper. These are very precious experiences. PhD students at OCBE typically use about 2-3 weeks per semester in advising, on average.

Junior researchers at NR are mentored and participate in on-going BigInsight projects. This gives them an overview of the centre and a valuable exposure to methods and applications.


PhD student Andrea Chi Zhang discussing her project on electronic health records with associate professor Signe Søvik, specialist in anesthesiology and Silje Bakken Jørgensen, infection protection consultant and specialist in microbiology, both from Akershus universitetssykehus

# COMMUNICATION AND DISSEMINATION ACTIVITIES

### Website

The website of the center is biginsight.no.

### BigInsight outreach presentations

Biginsight leadership and principal investigators hold seminars and participate to public events where they describe BigInsight's activities and research results, and contribute to the public debate about AI and digitalization. We maintain a list of our public appearances on our webpage. BigInsight participates, through UiO and NR, to the Norwegian *Artificial Intelligence* Research Consortium (NORA) and to the Norwegian Open *AI Lab*.

### The Digital Life Norway prize for "Transdisciplinary publication of the year" to a BigInsight paper

The paper "Toward Personalized Computer Simulation of Breast Cancer Treatment: A Multiscale Pharmacokinetic and Pharmacodynamic Model Informed by Multitype Patient Data", published in Cancer Research was awarded the prize: "The publication is an excellent example of how computation and modelling can be applied to clinical data and move towards clinical applications and in silico trials." https://digitallifenorway.org/gb/news/prize-for-transdisciplinary-publication-2019. **Congratulations to the authors!**

Alvaro Köhn-Luque trying to understand how cancer cells differ from normal cells, as part of transdisciplinary research

## BigInsight in the media

Dagens Næringsliv, Oslo. 13.12.2019, pp. 35
**«Det er algoritmens feil» – greit?** Anders Løland

Dagens Næringsliv, Oslo, 13.11.2019
**Kunstig intelligens og maskinlæring – hvordan bruker man data riktig?** Anders Løland. DN Fintech 2019.

Aftenposten (morgenutg. trykt utg.) 05.11.2019, pp. 27
**Vil vi ha robotlate dommere?** Anders Løland

Norsk Farmaceutisk Tidsskrift 7/2019 / farmatid.no 18.10.2019
**Lover verktøyet som skal gi fremtidens antibiotika**

Dagens næringsliv. pp 31. 23.07.2019.
**Fem grunner til at vi ikke lykkes med kunstig intelligens (ennå)**. Løland, Anders.

Computerworld nr. 6 - Juni 2019, pp 28
**Simulerer kreftbehandlingen din på datamaskinen**

Finansfokus 2/2019 23.04.2019
**Algoritmer skal avsløre svindlerne**

Bioteknologiradet.no, 14.02.2019
**Temaside til ungdomskoleprosjekt: 'Gen'-bokas hemmeligheter**, Arnoldo Frigessi

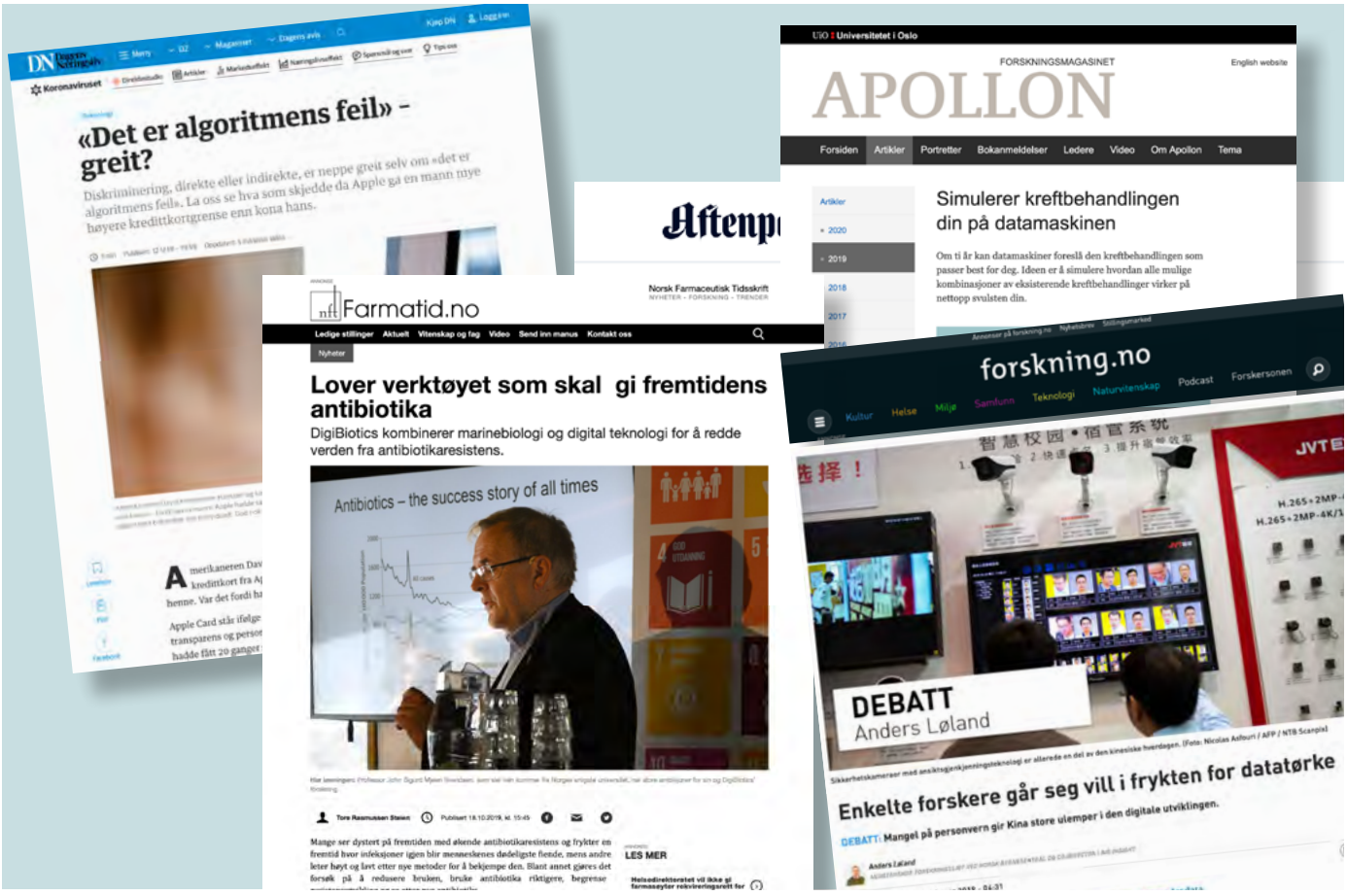
Forskning.no, 14.02.2019
**Enkelte forskere går seg vill i frykten for datatørke**, Anders Løland

Podcast LØRN.TECH #200 - 18.01.2019
**AI – Keiserens kunstige klær**, André Teigland

Dagens næringsliv. pp 39. 09.01.2019
**Skjult diskriminering med algoritmer?**, Anders Løland

Apollon 1/2019, pp 31
**Unngår skipskatastrofer med avansert statistikk**

Apollon 1/2019, pp 35
**Persontilpasset medisin: Simulerer kreftbehandlingen din på datamaskinen**



# 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.

**In 2019 the following people started at UiO, OUS and NIPH:**

| NAME | POSITION | FUNDING | RESEARCH AREA | AFFILIATION |
|---|---|---|---|---|
| Clara Bertinelli Salucci | PhD student | BigInsight | Sensor | UiO |
| Magnus Nygård Osnes | PhD student | NIPH | Health | UiO, NIPH |
| Emilie Ødegård | PhD student | UiO | Health/Marketing | UiO |
| Tero Aittokallio | Professor II | OUS | Health | UiO, OUS |
| Haifeng Hu | PhD student | OUS | Health | UiO, OUS |
| Johan Pensar | Ass. Professor | UiO | Health, methods | UiO |

# PERSONNEL

**Personnel affiliated with BigInsight for at least 10% of their time.**

| NAME | INSTITUTION | MAIN RESEARCH AREA |
|---|---|---|
| Arnoldo Frigessi | UiO/OUS/NR | Marketing, Health, Sensor, Explain AI |
| Jaroslaw Novak | ABB | Sensor |
| Morten Stakkeland | ABB | Sensor, Explain AI |
| Lars Erik Bolstad | DNB | Fraud |
| Johannes Lorentzen | DNB | Fraud |
| Fredrik Johannessen | DNB | Marketing |
| Karl Aksel Festø | DNB | Marketing |
| Marcus Nilsson | DNB | Marketing |
| Aiko Yamashita | DNB | Marketing, Explain AI |
| Nafiseh Shabib | DNB | Marketing |
| Hodjat Rahmati | DNB | Marketing |
| Roger Olafsen | DNB | Marketing |
| Geir Ånonsen | DNB | Fraud |
| Øystein Alnes | DNV-GL | Sensor |
| Ole Christian Astrup | DNV-GL | Sensor |
| Håvard Nordtveit Austefjord | DNV-GL | Sensor |
| Andreas Brandsæter | DNV-GL | Sensor, Explain AI |
| Øystein Engelhardtsen | DNV-GL | Sensor |
| Ørjan Fredriksen | DNV-GL | Sensor |
| Asun St. Clair | DNV-GL | Explain AI |
| Gaute Storhaug | DNV-GL | Sensor |
| Hans Anton Tvete | DNV-GL | Sensor |
| Bjørn-Johan Vartdal | DNV-GL | Sensor |
| Erik Vanem | DNV-GL | Sensor, Power |
| Anders Nyberg | Gjensidige | Marketing |
| Geir Thomassen | Gjensidige | Fraud |
| Ellen Paaske | Hydro | Power |
| Knut-Harald Bakke | Hydro | Power |

| NAME | INSTITUTION | MAIN RESEARCH AREA |
|---|---|---|
| Peter Szederjesi | Hydro | Power |
| Birgitte De Blasio | NIPH | Health |
| Francesco di Ruscio | NIPH | Health |
| Gunnar Ro | NIPH | Health |
| Ulf Andersen | NAV | Fraud |
| Robindra Prabhu | NAV | Fraud |
| Kjersti Aas | NR | Marketing, Explain AI |
| Magne Aldrin | NR | Sensor |
| Clara-Cecilie Günther | NR | Marketing, Health |
| Solveig Engetretsen | NR | Health, Marketing, Sensor |
| Ola Haug | NR | Marketing, Sensor |
| Marion Haugen | NR | Marketing |
| Kristoffer Herland Hellton | NR | Marketing, Sensor |
| Lars Holden | NR | Health, Fraud |
| Ragnar Bang Huseby | NR | Fraud, Power |
| Martin Jullum | NR | Fraud |
| Alex Lenkoski | NR | Power |
| Anders Løland | NR | Fraud, Power, Explain AI |
| Linda R. Neef | NR | Fraud, Explain AI |
| Annabelle Redelmeier | NR | Marketing |
| Hanne Rognebakke | NR | Marketing, Sensor |
| Nikolai Sellereite | NR | Marketing |
| Gunnhildur Steinbakk | NR | Fraud, Sensor, Power |
| André Teigland | NR | Explain AI |
| Ingunn Fride Tvete | NR | Health |
| Mette Langaas | NR/NTNU | Sensor, Health |
| Torsten Eken | OUS | Health |
| Thomas Fleischer | OUS | Health |
| Eivind Hovig | OUS | Health |
| Irena Jakopanec | OUS | Health |
| Vessela Kristensen | OUS | Health |
| Marissa LeBlanc | OUS | Health |
| Sygve Nakken | OUS | Health |
| Andrew Reiner | OUS | Marketing, Health |
| David Swanson | OUS | Health |
| Tonje Lien | OUS | Health |
| Bjørn Møller | OUS | Health |
| Jan Nygård | OUS | Health |
| Fredrik Schjesvold | OUS | Health |
| Therese Seierstad | OUS | Health |
| Tero Aittokallio | OUS/UiO | Health |
| Anders Berset | Skatteetaten | Marketing, Fraud |
| Wenche Celiussen | Skatteetaten | Marketing, Fraud |
| Øystein Olsen | Skatteetaten | Marketing |
| Nils Gaute Voll | Skatteetaten | Marketing, Fraud |
| Anders Holmberg | SSB | Marketing, Sensor |
| Øystein Langsrud | SSB | Marketing, Sensor, Explain AI |
| Kim Benjamin Boué | SSB | Marketing, Sensor |
| Li-Chun Zhang | SSB | Marketing, Sensor |
| Kenth Engo-Monsen | Telenor | Marketing, Health |
| Geoffrey Canright | Telenor | Marketing, Health |
| Bård Støve | UiB | Fraud |
| Dag Tjøstheim | UiB/NR | Fraud |
| Elja Arjas | UiO | Marketing |
| Ørnulf Borgan | UiO | Marketing |
| Jukka Corander | UiO | Health |
| Riccardo de Bin | UiO | Marketing |
| Ingrid K. Glad | UiO | Sensor |
| Ingrid Hobæk Haff | UiO | Fraud |
| Nils Lid Hjort | UiO | Fraud, Sensor |

| NAME | INSTITUTION | MAIN RESEARCH AREA |
|---|---|---|
| Carlo Mannino | UiO | Power |
| Johan Pensar | UiO | TBD |
| Kjetil Røysland | UiO | Health |
| Sven Ove Samuelson | UiO | Marketing |
| Ida Scheel | UiO | Marketing |
| Geir Storvik | UiO | Sensor |
| Øystein Sørensen | UiO | Health |
| Magne Thoresen | UiO | Health |
| Hadi Fanaee Tork | UiO | Health |
| Valeria Vitelli | UiO | Marketing, Health |
| Manuela Zucknick | UiO | Health |

| NAME | FUNDING | NATIONALITY | PERIOD | GENDER | TOPIC |
|---|---|---|---|---|---|
| **Postdoctoral researchers with financial support from BigInsight** | | | | | |
| Azzeddine Bakdi | | Algeria | 2018-2021 | M | Sensor |
| Pål Christie Ryalen | | Norway | 4 months | M | Health |
| **Postdoctoral researchers in BigInsight with financial support from other sources** | | | | | |
| Andrea Cremaschi | UiO/NCMM | Italy | 2016-2018 | M | Health |
| Alvaro Köhn Luque | UiO | Spain | 2016-2021 | M | Health |
| Christian Page | OUS/HSØ | Norway | 2016-2019 | M | Health |
| Richard Xiaoran Lai | UiO | UK | 2019-2022 | M | Health |
| Henry Pesonen | UiO | Finland | 2019-2022 | M | Health |
| **PhD students with financial support from BigInsight** | | | | | |
| Simon Boge Brant | | Norway | 2018-2021 | M | Fraud |
| Solveig Engebretsen | | Norway | 2016-2019 | F | Health |
| Emanuele Gramuglia | | Italy | 2016-2020 | M | Sensor |
| Brittany Rose | | USA | 2018-2021 | F | Health |
| Leiv Tore Salte Rønneberg | | Norway | 2018-2021 | M | Health |
| Clara Bertinelli Salucci | | Italy | 2019-2022 | F | Sensor |
| Riccardo Parviero | | Italy | 2018-2021 | M | Marketing |
| Jonas Schenkel | | Norway | 2018-2021 | M | SSB, Sensor |
| Martin Tveten | | Norway | 2017-2020 | M | Sensor |
| Andrea Chi Zhang | | China | 2016-2020 | F | Health |
| George Zhi Zhao | | China | 2019-2020 | M | Health |
| Xiaoran Lai | | UK | 2019-2019 | M | Health |
| **PhD students in BigInsight with financial support from other sources** | | | | | |
| Andreas Brandsæter | DNV-GL, NæringslivPhD | Norway | 2016-2020 | M | Sensor |
| Simen Eide | Finn.no, NæringslivPhD | Norway | 2018-2021 | M | Marketing |
| Håvard Kvamme | UiO | Norway | 2015-2020 | M | Marketing |
| Richard Xiaoran Lai | UiO | UK | 2016-2019 | M | Health |
| Sylvia Qinghua Liu | UiO/MI Innovation | China | 2017-2021 | F | Marketing |
| Andreas Nakkerud | UiO/MI Innovation | Norway | 2016-2020 | M | Power |
| Jaroslaw Nowak | ABB, NæringslivPhD | Poland | 2018-2021 | M | Sensor |
| Anja Stein | STORi, Lancaster | Norway | 2019-2023 | F | Marketing |
| George Zhi Zhao | UiO/IMB | China | 2016-2020 | M | Health |
| Yinzhi Wang | UiO/MI | China | 2017-2019 | F | Fraud |

| NAME | FUNDING | NATIONALITY | PERIOD | GENDER | TOPIC |
|---|---|---|---|---|---|
| Emilie Ødegård | UiO/IMB | Norway | 2019-2023 | F | Health, Marketing |
| Haifeng Hu | OUS/UIO | China | 2019-2023 | M | Health |
| Magnus Nygård Osnes | NIPH/UiO | Norway | 2019-2023 | M | Health |
| **Master degrees** | | | | | |
| Håkon Bliksås Carlsen | | | 2020-2022 | M | Fraud |
| Bob Betuin Fjellheim | | | 2019-2021 | M | Marketing |
| He Gu | | | 2019-2021 | M | Sensor |
| Eirik Halsteinslid | | | 2017-2019 | M | Fraud |
| Nicola Kaletka | | | 2019-2021 | F | Health |
| Nicolay Kristensen | | | 2019-2021 | M | Sensor |
| Camilla Lingjærde | | | 2017-2019 | F | Health |
| Vera Haugen Kvisgaard | | | 2019-2021 | F | Fraud |
| Øystein Skauli | | | 2019-2021 | M | Marketing |

# FINANCIAL OVERVIEW

| FUNDING | 1000 NOK |
|---|---|
| The Research Council | 16 063 |
| Norwegian Computing Center (NR) | 1 386 |
| Research Partners*, in kind | 10 426 |
| Research Partners*, in cash | 7 85 |
| Enterprise partners**, in kind | 4 233 |
| Enterprise partners**, in cash | 3 170 |
| Public partners***, in kind | 4 571 |
| Public partners***, in cash | 2 871 |
| **Sum** | **43 505** |

| COSTS | |
|---|---|
| NR, research | 10 386 |
| NR, direct costs | 967 |
| Research Partners*, research | 22 947 |
| Enterprise partners**, research | 4 233 |
| Public partners***, research | 4 971 |
| **Sum** | **43 505** |

*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 Institute (NIPH), Statistics Norway

# PUBLICATIONS IN 2019 (incl. up to Mar 2020)

**Journal and peer-reviewed conference papers**


Abdelmalek, Samir; Rezazi, Sarah; Bakdi, Azzeddine; Bettayeb, Maamar. **Voltage dips effects detection and compensation for doubly-fed induction generator based wind energy conversion system.** Revue Roumaine des Sciences Techniques - Serie Electrotechnique et Energetique (ISSN 0035-4066). 64(3) pp 199-204. doi: http://revue.elth.pub.ro/viewpdf.php?id=844. 2019.

Baadji, Bousaadia; Bentarzi, Hamid; Bakdi, Azzeddine. **Comprehensive learning bat algorithm for optimal coordinated tuning of power system stabilizers and static VAR compensator in power systems.** Engineering optimization (ISSN 0305-215X). pp 1-19 doi: https://doi.org/10.1080/0305215X.2019.1677635. 2019.

Bakdi, Azzeddine; Kouadri, Abdelmalek; Mekhilef, Saad. **A data-driven algorithm for online detection of component and system faults in modern wind turbines at different operating zones.** Renewable & Sustainable Energy Reviews (ISSN 1364-0321). 103 pp 546-555. doi: 10.1016/j.rser.2019.01.013. 2019.

Bakdi, Azzeddine; Bounoua, Wahiba; Mekhilef, Saad; Halabi, Laith M.. **Nonparametric Kullback-divergence-PCA for intelligent mismatchdetection and power quality monitoring in grid-connected rooftop PV.** Energy. doi: https://doi.org/10.1016/j.energy.2019.116366. 2019.

Bakdi, Azzeddine; Glad, Ingrid Kristine; Vanem, Erik; Engelhardtsen, Øystein. **AIS-based multiple vessel collision and grounding risk identification based on adaptive safety domain.** Journal of Marine Science and Engineering, 8(1), 5:1-19, 2020 https://www.mdpi.com/2077-1312/8/1/5.

Bakdi, Azzeddine; Glad, Ingrid Kristine; Vanem, Erik; Engelhardtsen, Øystein. **AIS-based multiple vessel collision and grounding risk identification based on adaptive safety domain.** In Proc. 18th Marine Traffic Engineering/ ISIS 2019 conference (MTE-ISIS 2019), Kołobrzeg, Poland, October 16-18 2019. http://www.irm.am.szczecin.pl/pl/mte-isis-2019

Banterle, M*; Zhao, Z*; Bottolo, L; Richardson, S; Lewin, A; Zucknick, M (2020). BayesSUR: Bayesian Seemingly Unrelated Regression. R package version 1.1-1. https://CRAN.R-project.org/package=BayesSUR. (* joint first authors)

Bounoua, Wahiba; Benkara, Amina B.; Kouadri, Abdelmalek; Bakdi, Azzeddine. **Online Monitoring Scheme Using PCA through Kullback-Leibler Divergence Analysis Technique for Fault Detection.** Transactions of the Institute of Measurement and Control. (ISSN 0142-3312). 189 doi: 10.1177/0142331219888370. 2019.

Brandsæter, Andreas; Vanem, Erik; Glad, Ingrid Kristine. **Efficient on-line anomaly detection for ship systems in operation. Expert systems with applications.** (ISSN 0957-4174). 121 pp 418-437. doi: 10.1016/j.eswa.2018.12.040. 2019.

Crispino, Marta; Arjas, Elja; Vitelli, Valeria; Barrett, Natasha; Frigessi, Arnoldo. **A Bayesian Mallows approach to nontransitive pair comparison data: How human are sounds?** Annals of Applied Statistics. 13(1) pp 492-519. doi: 10.1214/18-AOAS1203. 2019.

Engebretsen, Solveig; Glad, Ingrid Kristine. **Additive monotone regression in high and lower dimensions. Statistics Surveys.** 13 pp 1-51. doi: 10.1214/19-SS124. 2019.

Engebretsen, Solveig; Engø-Monsen, Kenth; Frigessi, Arnoldo; De Blasio, Birgitte Freiesleben. **A theoretical single-parameter model for urbanisation to study infectious disease spread and interventions.** PLoS Computational Biology. 15(3) doi: 10.1371/journal.pcbi.1006879. 2019.

Fanaee-T, H.; Thoresen, M. **Multi-insight visualization of multi-omics data via ensemble dimension reduction and tensor factorization.** Bioinformatics 2019, 35: 1625-1633. doi: 10.1093/bioinformatics/bty847. 2019.

Fanaee-T, H.; Thoresen, M. **Performance evaluation of methods for integrative dimension reduction.** Information Sciences 2019, 493: 105-119. doi.org/10.1016/j.ins.2019.04.041. 2019.

Ghosh, A.; Thoresen, M. **Consistent fixed-effects selection in ultra-high dimensional linear mixed models with error-covariate endogeneity.** Accepted for publication in Statistica Sinica. DOI: 10.5705/ss.202019.0421. 2019.

Grytten, I.; Rand, K. D.; Nederbragt, A. J.; Storvik, G. O.; Glad, I. K.; Sandve, G. K. (2019). **Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes.** PLoS computational biology, 15(2), e1006731.

Hentout, Abdelfetah; Maoudj, Abderraouf; Guir, Dallel; Saighi, Souhila; Harkat, Mohamed Aures; Hammouche, Mohamed Zakaria; Bakdi, Azzeddine. **Collision-free Path Planning for Indoor Mobile Robots Based on Rapidly-exploring Random Trees and Piecewise Cubic Hermite Interpolating Polynomial.** International journal of imaging and robotics. 19(03) 2019.

Gåsemyr, Jørund Inge; Scheel, Ida. **Alternatives to post-processing posterior predictive p-values.** Scandinavian Journal of Statistics https://www.mn.uio.no/math/personer/vit/idasch/NA. ISSN 0303-6898. 46(4), s 1252- 1273. 2019

Jullum, M.; Løland, A.; Huseby, R. B.; Ånonsen, G.; Lorentzen, J. (2020). **Detecting money laundering transactions with machine learning.** Journal of Money Laundering Control, 23(1), 173-186.

Karahasanovic, Amela; Zeh, Theodor; Eide, Aslak Wegner; Schittekat, Patrick; Swendgaard, Hans; Bakhrankova, Krystsina; Grantz, Volker; Kjenstad, Dag; Mannino, Carlo; Rokitansky, Carl-Herbert; Gräupl, Thomas. **Can Holistic Optimization Improve Airport Air Traffic Management Performance.** IEEE Aerospace and Electronic Systems Magazine. 34(5) pp 12-20. doi: 10.1109/MAES.2019.2918039. Institutional archive 2019.

Ko, Vinnie; Hjort, Nils Lid; Hobæk Haff, Ingrid. **Focused information criteria for copulae.** Scandinavian Journal of Statistics. 46 pp 1117-1140. doi: 10.1111/sjos.12387. 2019.

Kokko, Jan; Remes, Ulpu; Thomas, Owen; Pesonen, Henri; Corander, Jukka. **PYLFIRE: Python implementation of likelihood-free inference by ratio estimation.** Wellcome Open Research 2019. https://doi.org/10.12688/wellcomeopenres.15583.1

Kvamme, Håvard; Borgan, Ørnulf; Scheel, Ida. **Time-to-Event Prediction with Neural Networks and Cox Regression.** Journal of machine learning research. (ISSN 1532-4435). 20(129) pp 1-30. 2019.

Lai, Xiaoran; Geier, Oliver; Fleischer, Thomas; Garred, Øystein; Borgen, Elin; Funke, Simon Wolfgang; Kumar, Surendra; Rognes, Marie Elisabeth; Seierstad, Therese; Børresen-Dale, Anne-Lise; Kristensen, Vessela N.; Engebråten, Olav; Kohn Luque, Alvaro; Frigessi, Arnoldo. **Towards personalized computer simulation of breast cancer treatment: a multi-scale pharmacokinetic and pharmacodynamic model informed by multi-type patient data.** Cancer Research. 79(16) pp 4293-4304. doi: 10.1158/0008-5472.CAN-18-1804. 2019.

Lamorgese, Leonardo; Mannino, Carlo. **A non-compact formulation for job-shop scheduling problems in traffic management.** Operations Research. 67(6) doi: 10.1287/opre.2018.1837. 2019.

Liu, Qinghua; Reiner, Andrew Henry; Frigessi, Arnoldo; Scheel, Ida. **Diverse personalized recommendations with uncertainty from implicit preference data with the Bayesian Mallows model.** Knowledge-Based Systems. pp 1-12. doi: 10.1016/j.knosys.2019.104960. 2019.

Liu, Qinghua; Crispino, Marta; Scheel, Ida; Vitelli, Valeria; Frigessi, Arnoldo. **Model-Based Learning from Preference Data. Annual Review of Statistics and Its Application.** 6(1) pp 329-354. doi: https://doi.org/10.1146/annurev-statistics-031017-100213. 2019.

Mannino, Carlo; Lamorgese, Leonardo. **A non-compact formulation for job-shop scheduling problems in traffic management. Operations Research.** 67(6) doi: 10.1287/opre.2018.1837. 2019.

Maros ME, et al. (2020). **Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data.** Nature Protocols, 1-34.

Menden, Michael P.; Wang, Dennis; Mason, Mike J.; Szalai, Bence; Bulusu, Krishna C.; Guan, Yuanfang; Yu, Thomas; Kang, Jaewoo; Jeon, Minji; Wolfinger, Russ; Nguyen, Tin; Zaslavskiy, Mikhail; Jang, In Sock; Ghazoui, Zara; Ahsen, Mehmet Eren; Vogel, Robert; Neto, Elias Chaibub; Norman, Thea; Tang, Eric K.Y.; Garnett, Mathew J.; Di Veroli, Giovanni Y.; Fawell, Stephen; Stolovitzky, Gustavo; Zucknick, Manuela; Guinney, Justin; Dry, Jonathan R.; Saez-Rodriguez, Julio. **Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen.** Nature Communications. (ISSN 2041-1723). 10:2674 pp 1-17. doi: 10.1038/s41467-019-09799-2. 2019.

Moss, Jonas; Tveten, Martin. kdensity: **An R package for kernel density estimation with parametric starts and asymmetric kernels.** Journal of Open Source Software. doi: 10.21105/joss.01566. 2019.

Otneim, H., Jullum, M., Tjøstheim, D. (2020). **Pairwise local Fisher and naive Bayes: Improving two standard discriminants.** Journal of Econometrics, 216(1), 284-304.

Page, Christian M., Djordjilović, Vera; Nøst, Therese H.; Ghiasvand, Reza; Sandanger, Torkjel M.; Frigessi, Arnoldo; Thoresen, Magne; Veierød, Marit B. **Lifetime Ultraviolet Radiation exposure and DnA Methylation in Blood Leukocytes: the norwegian Women and cancer Study.** Scientific reports 10, no. 1 (2020): 1-8.

Pensar, Johan; Puranen, Santeri; Arnold, Brian; MacAlasdair, Neil; Kuronen, Juri; Tonkin-Hill, Gerry; Pesonen, Maiju; Xu, Yingying; Sipola, Aleksi; Sánchez-Busó, Leonor; Lees, John A; Chewapreecha, Claire; Bentley, Stephen D; Harris, Simon R; Parkhill, Julian; Croucher, Nicholas J; Corander, Jukka. **Genome-wide epistasis and co-selection study using mutual information.** Nucleic Acids Research, Volume 47, Issue 18, 10 October 2019, Page e112, https://doi.org/10.1093/nar/gkz656

Romeo, G.; Thoresen, M. **Model selection in high-dimensional noisy data: a simulation study.** Journal of Statistical Computation and Simulation 2019, 89: 2031-2050. doi/10.1080/00949655.2019.1607345. 2019.

Sellereite, N., Jullum, M. (2020). shapr: **An R-package for explaining machine learning models with dependence-aware Shapley values.** Journal of Open Source Software, 5(46), 2027.

Swanson, David; Lien, Tonje Gulbrandsen; Bergholtz, Helga; Sørlie, Therese; Frigessi, Arnoldo. **A Bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a breast cancer cohort.** Bioinformatics. doi:10.1093/bioinformatics/btz381. 2019.

Tekpli, X., Lien, T., Røssevold, A. H., Nebdal, D., Borgen, E., Ohnstad, H. O Kyte, J, Vallon-Christersson, MF, Due EU, Svartdal LG, Sveli MAT, Garred Ø, OSBREAC, Frigessi, A, Sahlberg KK, Sørlie T, Russnes HG, Naume B, Kristensen V. **An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment.** Nature Communications, 10(1), 1-14.

Tveten, Martin. **Which principal components are most sensitive in the change detection problem?** Stat. 8(1) doi: 10.1002/sta4.252. Institutional archive. 2019.

Vanem, Erik; Brandsæter, A. **Unsupervised anomaly detection based on clustering methods and sensor data on a marine diesel engine.,** Online first, Journal of Marine Engineering and Technology. https://www.tandfonline.com/doi/abs/10.1080/20464177.2019.1633223?journalCode=tmar20. 2019.

Vanem, Erik; Holterud Aarnes, Lars; Storhaug, Gaute; Astrup, Ole Christian. **Statistical Modelling and Comparison of Model-Based Fatigue Calculations and Hull Monitoring Data for Container Vessels** In proceedings of The 14th International Symposium on Practical Design of Ships and Other Floating Structures (PRADS 2019), Yokohama, Japan, September 22-26, 2019. https://prads2019.jp/

Zhang, L.-C. **On valid descriptive inference from non-probability sample.** Statistical Theory and Related Fields, 3:103-113. DOI:10.1080/24754269.2019.1666241. 2019

Zhao, Z; Zucknick, M. **Structured penalized regression for drug sensitivity prediction.** Journal of the Royal Statistical Society: Series C (Applied Statistics). First published: 23 February 2020, URL: https://doi.org/10.1111/rssc.12400.

### Reports and submitted papers

Cremaschi, A., Frigessi, A., Taskén, K., & Zucknick, M. (2019). **A Bayesian approach for the study of synergistic interaction effects in in-vitro drug combination experiments.** arXiv preprint arXiv:1904.04901.

Engebretsen, S., Engoe-Monsen, K., Aleem, M. A., Gurley, E. S., Frigessi, A., & de Blasio, B. F. (2020). **Time-aggregated mobile phone mobility data are sufficient for modelling influenza spread: the case of Bangladesh.** medRxiv.

Kvamme, Håvard; Borgan, Ørnulf. **Continuous and Discrete-Time Survival Prediction with Neural Networks.** Submitted https://arxiv.org/pdf/1910.06724.pdf (recently rejected from JMLR)

Kvamme, Håvard; Borgan, Ørnulf. **The Brier Score under Administrative Censoring: Problems and Solutions.** https://arxiv.org/pdf/1912.08581.pdf (submitted to JMLR)

Mai, TT*; Rønneberg, L*; Zhao, Z; Zucknick, M; Corander, J (2019). **Composite local low-rank structure in learning drug sensitivity.** Proceedings of CIBB 2019. arXiv preprint arXiv:1905.00095.
(* joint first authors)

Ryalen, Pål Christie; Møller, Bjørn; Laache, Christoffer; Stensrud, Mats; Røysland, Kjetil. **Prognosis of cancer survivors.** Submitted on 22 Mar 2020 to Biostatistics.

Sande, L.S. and Zhang, L.-C. (2019). **Design-unbiased statistical learning in survey sampling.** Submitted to Sankaya, special in celebration of C.R.Rao.

Steinbakk, Gunnhildur Högnadottir; Aarsnes, Lars Holterud; Aldrin, Magne; Astrup, Ole Christian; Haug, Ola; Storhaug, Gaute; Vanem, Erik. **Statistical approximation to synthetic mid-ship hull girder stress response. Fast track paper at Journal of Ship Research.** (Selve paper ikke ute ennå, men link til journalen her: https://www.onepetro.org/journals/Journal%20of%20Ship%20Research/64/01)

Thomas, Owen; Corander, Jukka. **Diagnosing model misspecification and performing generalized Bayes' updates via probabilistic classifiers.** arXiv.org > stat > arXiv:1912.05810
Submitted on 12 Dec 2019

Thomas, Owen; Dutta, Ritabrata; Corander, Jukka; Kaski, Samuel; Gutmann, Michael U. **Likelihood-free inference by ratio estimation.** arXiv.org > stat > arXiv:1611.10242
Submitted on 30 Nov 2016 (v1), last revised 16 Oct 2019 (v5)

Tveten, M., & Glad, I. K. (2019). **Online Detection of Sparse Changes in High-Dimensional Data Streams Using Tailored Projections.** arXiv preprint arXiv:1908.02029.

Wahl, Jens Christian; Sellereite, Nikolai; Aas, Kjersti. **Predicting probability of default for SMEs using relational and transaction data.** Norsk Regnesentral, NR-notat SAMBA/07/19. pp 52. 2019.

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

Design og lay-out: **designavdelingenkis.no**
Ved å bruke designavdelingenkis støtter man Kunst i Skolens arbeid med kunstformidling og kreativitetstrening for barn og unge.


BigInsight