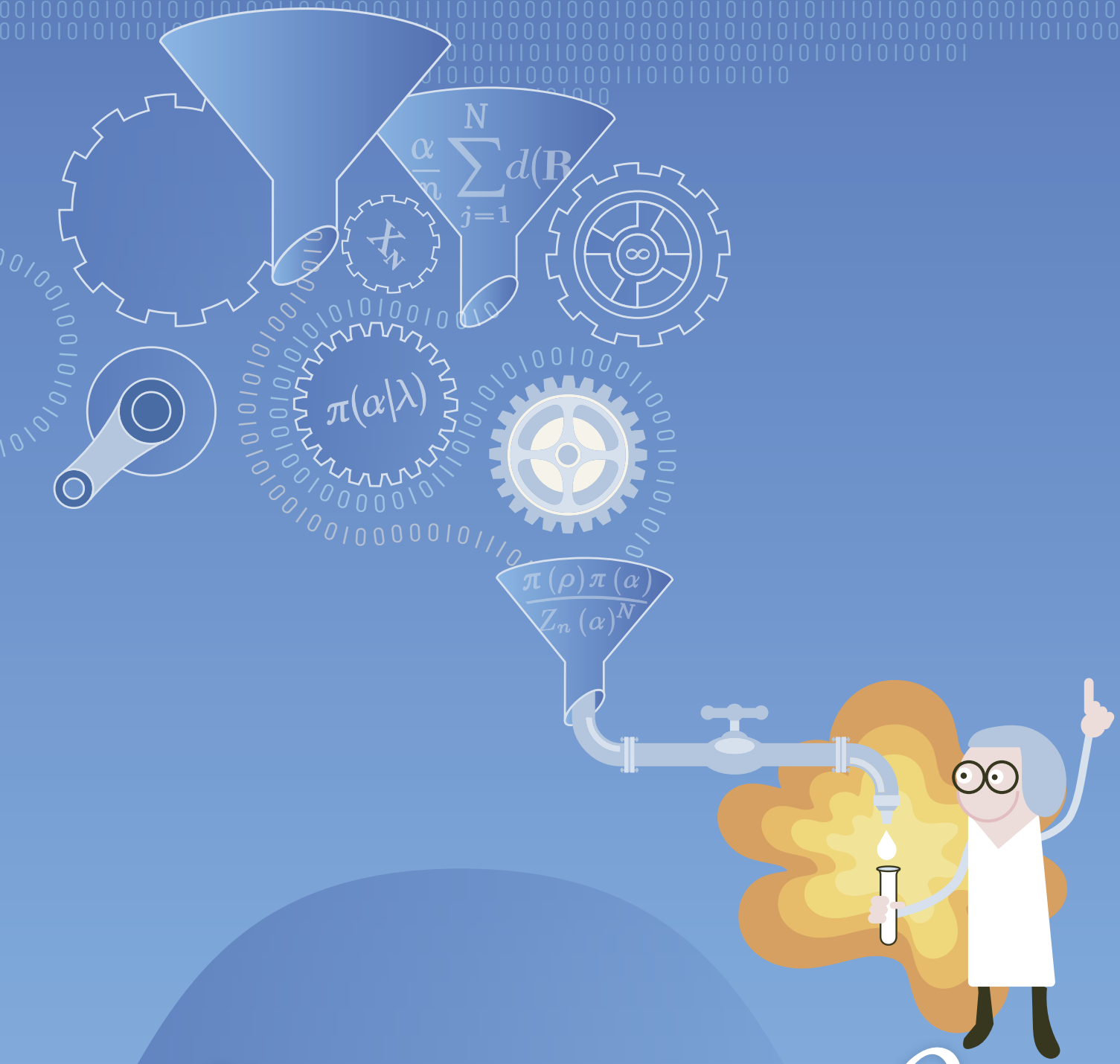
BigInsight

STATISTICS FOR THE KNOWLEDGE ECONOMY

ANNUAL REPORT 2020



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.

Methods and algorithms we develop and implement at BigInsight are explainable, accurate and fair, because we recognise our responsibilities. Our research is open. 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 sixth year of BigInsight. Innovation results are highlighted, together with the broad spectrum of research projects.

"We believe that AI will be a force multiplier on technological progress in our increasingly digital, data-driven world. This is because everything around us today, ranging from culture to consumer products, is a product of intelligence."

stateof.ai 2020

VISION AND OBJECTIVES

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

"Tell me about your business problem, do not tell me about your AI problem. It is my task to translate your business problem into an AI problem, which we can solve with machine learning. That's my expertise."

Andre Ng, professor Stanford University, Chief Scientist of Baidu, founder of Coursera. Lecture for the Royal Statistical Society, February 2021.

for DNV GL and ABB, hundreds of sensors register the functional state and operation of a vessel at sea.

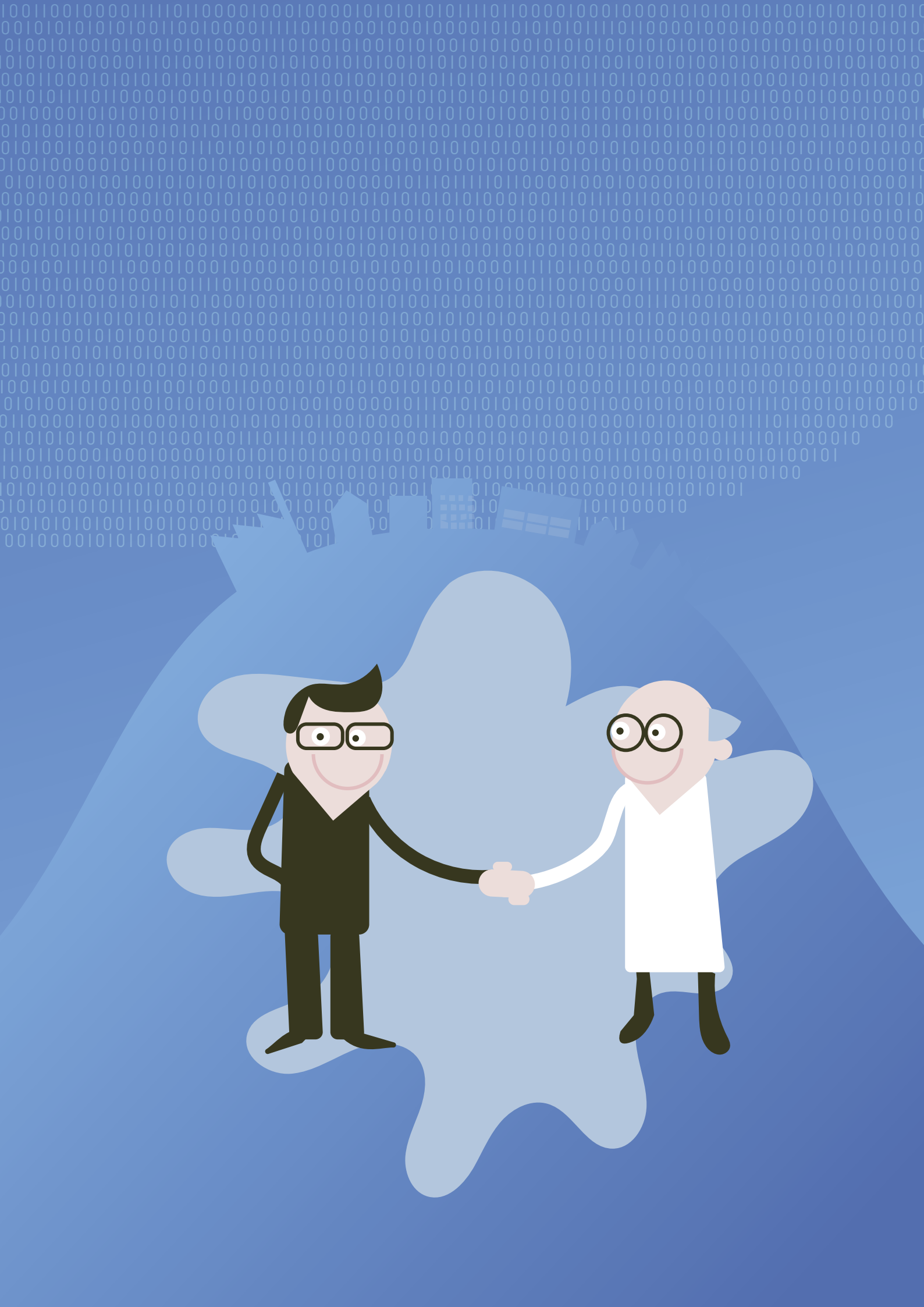
There are many common characteristics:
- a high number of units/individuals/sensors
- in some cases, massive data for each unit; in other cases, more limited information
- complex dependence structure between units
- new data types, new technologies, new regulations are available
- in most cases, units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

Each partner has specific 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 competitive-ness, 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.



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 2020, where all partners are represented. In addition to close cooperation with the researchers at NR and the universities, there have been several meetings on broader topics, like Explainable AI, where partners have met and exchanged ideas. This has resulted in more bilateral partner-to-partner cooperation across the Innovation Objectives. Due to the corona pandemic, most meetings, seminars and workshops have been digital on Zoom or Teams.

The annual BigInsight Day was postponed several times during 2020, as we were hoping to arrange a physical meeting. It was finally scheduled for February 2021. The program includes an interesting debate on how to communicate statistical uncertainty, a central topic during the last year as BigInsight has been deeply involved in the Norwegian corona-modeling through the Norwegian Institute of Public Health.

ORGANISATION

Board in 2020

Karl Aksel Festø, DNB, chairman (from November)
Marcus Zackrisson, Skatteetaten, chairman (until November)
Andree Underthun, ABB
Hans Anton Tvete, DNV GL
Birgitte F. De Blasio, Folkehelseinstituttet
Erlend Willand-Evensen, Gjensidige
Ellen Charlotte Stavseth Paaske, Hydro
Cathrine Phil Lyngstad, NAV
Lars Holden, Norsk Regnesentral
André Teigland, Norsk Regnesentral
Peder Heyerdahl Utne, Oslo University Hospital
Line Wilberg, Skatteetaten (from November)
Xeni Dimakos, SSB
Kenth Engø-Monsen, 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 2020.
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
Martin Jullum, NR
Alex Lenkoski, NR
Anders Løland, NR
Carlo Mannino, UiO
Hanne Rognebakke, NR
Ida Scheel, UiO
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
Prof. Veronica Vinciotti, University of Trento, Italy

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.

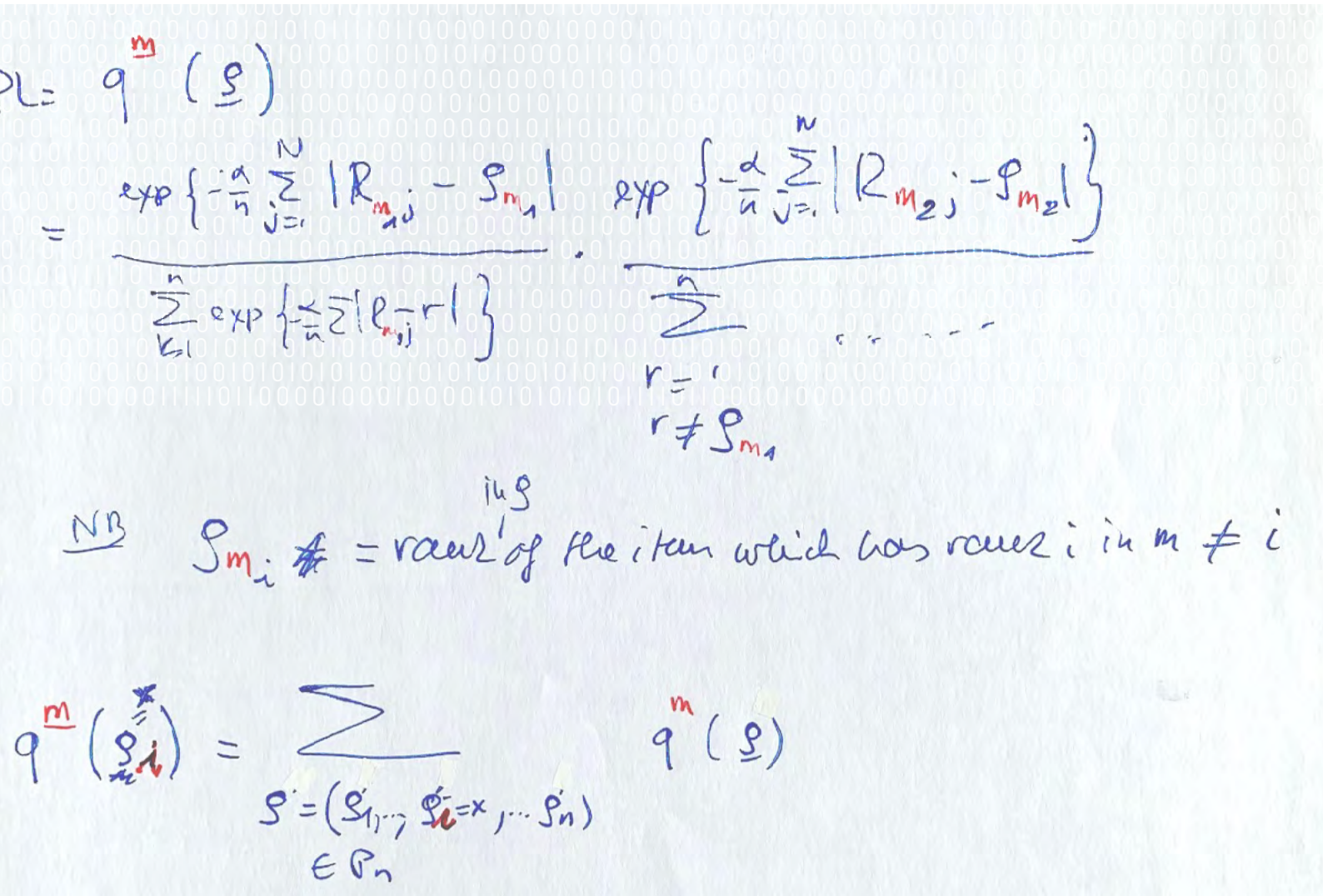
INNOVATION OBJECTIVES	 Personalised marketing	 Personalised health and patient safety	 Personalised fraud detection	 Sensor systems	 Forecasting power systems	 Explaning AI
INNOVATION PARTNERS	DNB Gjensidige NAV Skatteetaten Telenor SSB	DNV-GL Kreftregisteret OUS Telenor	DNB Gjensidige Skatteetaten	ABB DNV-GL SSB	DNV-GL Hydro Energy SSB	all partners
RESEARCH PARTNERS	NR UiO NIPH UiB	UiO OUS NR NIPH	NR UiO UiB	NR UiO	NR UiO	NR UiO
PRINCIPAL INVESTIGATORS						
Principal Investigators: co-Principal Investigators:	Kjersti Aas Ida Scheel	Magne Thoresen Clara Cecilie Günther	Anders Løland Martin Jullum	Ingrid Glad Hanne Rognebakke	Alex Lenkoski Carlo Mannino	Anders Løland 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 or other regularities, and with measurement error in covariates.
- 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 and inference for infectious disease processes, covid-19 in particular.
- General methods to describe uncertainty in predictions.
- Pair copula constructions for structure learning.
- Models for the forming of social networks and inference from data in time.
- Anomaly detection methods and algorithms.
- Modelling interactions between actors inspired by infection diseases mathematical models.
- Explaining black box models when covariates are dependent.
- Generation and ranking of situations of risk for autonomous ship navigation.
- Parallelisation of numerical solvers for stochastic differential equations and random cellular automata
- Time-to-event prediction using neural nets, when covariates are time dependent.



SCIENTIFIC ACTIVITIES

2020 was the first year of covid-19, the year when the pandemics started, and the world tried to control the virus. This was the year when the world experienced the closure of shops, hjemmekontor (the Norwegian word for teleworking), the complete stop of cultural events and entertainment, closed borders. We saw hospitals in critical conditions, first in Italy, and then in most European countries. Schools were closed in periods; teaching was often in digital form. We learned ZOOM and Teams, studied backgrounds, and saw our own face as never before, we could not interrupt each other, we could not see each other in the eyes. 2021 does not seem different, unfortunately, even if several vaccines are now used, but their production is slow, and it will take 8 months or so to vaccinate most Norwegians. In the meantime, new and more infectious variants of the SARS-CoV-2 virus started to be predominant. When this annual report goes into print, Norway –and Oslo in particular – is facing a new wave of the epidemics, which could be the worst one. But we also start to see the end of this period, maybe the autumn 2021 will be *normal*? And we wonder how *normality* will be.

The pandemic has accentuated differences between those who had all the resources to maintain life standards and working conditions, and those who could not. In Norway, and possibly in other countries too, the pandemics has also generated a new feeling of belonging to the community. *Dugnad* is the Norwegian word. We have also understood that it is possible to slow down, even stop, economic growth, with the possibility this opens for controlling climate change, reduce poverty, and solve together other dramatic challenges the earth is facing.

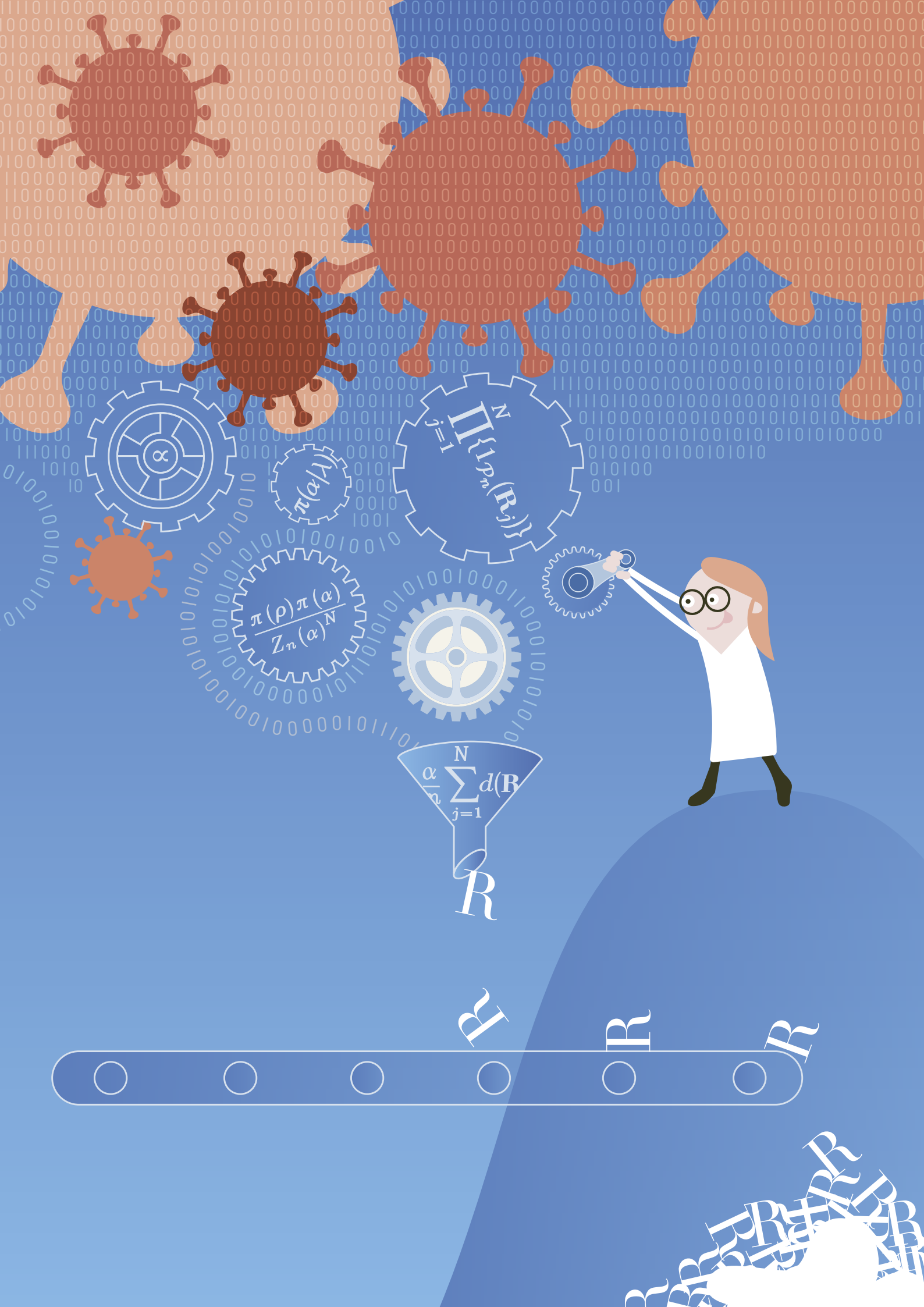
BigInsight has been central in the fight against the virus in Norway. Methodology developed by BigInsight has been in daily use at the Norwegian Institute of Public Health to model the spread of the pandemic in Norway. We estimated reproduction numbers, predicted number of new admissions to hospitals, and estimated the total number of infected, all with Bayesian 95% credibility intervals. The Monday 8 o'clock meeting was to decide exactly which models to run, the reports being ready Tuesday night. In the Thursday 8 o'clock meeting we planned the next model runs and discussed model development. We changed the model many times, improving its realism and inferential power. The prime minister cited from our reports a few hours after we had sent it off. Translated "på norsk", because all our reports have been in English. We have been in continuous contact with the Norwegian media, and hopefully have helped to understand the nowcasts and the forecasts.

During emergency, research has changed, from being "evidence-based" to becoming "evidence-making". Evidence-based science uses all available knowledge and data to produce results that are "final" in the actual situation. Research in this covid-19 year has shown that knowledge is changing fast, data are extending and improving. Questions of the public and decision makers have shaped research agendas. Models have always felt a starting point, in a dynamic process whose aim was to produce evidence about the pandemics; evidence-making. This new way of doing science is very exciting and it will be interesting to see how it will develop in the next few years.

Uncertainty is now a concept for the public. Nobody is surprised that predictions come with a confidence interval and estimates too. We have decided to call them confidence interval, even if they are very often Bayesian posterior credibility intervals, to ease communication. This is a major step forward, with broad consequences. Confidence intervals can be broad, nevertheless highly informative, and useful. The space they open is a space where discussion can start. It is a playground for democracy, as decisions are then influenced by other perspectives, political, ethical, economical, and so on.

In the landscape of covid-19, BigInsight has continued to produce solid scientific results which matter for innovation and create value for our partners. In this annual report we try to summarise our most important achievements. We highlight how our methodological results, new models and algorithms, innovative repurposing of methods in new areas, are giving interesting competitive advantages to our partners. They also bring about a fantastic enthusiasm in our research teams, something that has been very important, particular in 2020.

Arnoldo Frigessi.


α
$\pi(\alpha|\lambda)$
$\prod_{j=1}^{N}\{1_{\mathcal{P}_n}(\mathbf{R}_j)\}$
$\frac{\pi(\rho)\pi(\alpha)}{Z_n(\alpha)^N}$
$\alpha_n \sum_{j=1}^{N} d(\mathbf{R}$
R
R
R
R

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 2020:

Spatial modelling of risk premiums for water damage insurance

The Generalised Linear Model (GLM) is the common statistical tool used to estimate insurance risk premiums in nonlife insurance. In water damage insurance, the explanatory variables are typically different features characterising the size, age, and standard of the building as well as features describing the demographic status of the policyholder. In addition to such risk factors, it is of interest to take the risk associated with the policyholders' geographical location into account. This might be done by extending the GLM framework by modelling the latent geographic structure as random effects. In this sub project, we have evaluated four models that take the spatial variability into account: (1) the Intrinsic Markov Random Field (ICAR) model; (2) the Besag, York, Mollier (BYM) model, (3) the independent random effects model and (4) a spatial splines model. The models have been tested on a huge dataset from Gjensidige containing seven million observations of policyholders during the period 2011-2018. The work has been published in a paper submitted to the Scandinavian Actuarial Journal.

Bayesian methodology for recommender systems

BigInsight has developed a new approach to recommendations, based on the Bayesian Mallows Model. The methodology 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. We have made the methodology scalable by proposing a Variational Bayesian approximation. In 2020, we have worked on proving the optimal construction of this approximation algorithm. Furthermore, we are finalizing another Bayesian model, which is a fully probabilistic model of clicking histories, which captures user heterogeneity and item similarity. Again, we propose a Variational Bayesian approximation which scales to the needed dimensions in user and item space. With a good approximation of the posterior distribution of all parameters at hand, we investigate and compare three different recommender policies, which balance exploration and exploitation in a Bayesian context. One important aspect of our work is that we do not assume that the user has made her clicking decisions on the basis of a hypothetical comparison of all items, but only with the ones which have been offered to her in each interaction with the recommender system. The new method is under AB testing at finn.no.

Stochastic models for early prediction of viral customer behaviour 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 the 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. In 2020 we have submitted a paper which describes the methodology, investigates the performance on simulated data and shows the successful results for real data on a Telenor product. For most applications, the full social network will not be known. Instead, we have to use some sort of proxy network. For the study on the Telenor product the network was based on all tele communication between

their customers, which of course is an approximation to the full social network. In 2020, we have studied how sensitive the results of our methodology are to using such proxy networks in a simulation study.

Sales prediction

DNB Puls is an app for people running small or medium sized businesses. Among other things, it produces forecasts of future income based on time series of previous values. In this project, the aim is to improve the current predictions in DNB Puls. This is a very difficult problem, because the historical time series data are very noisy with irregular patterns and many missing values. We use a new combination of traditional time series methods and more recent machine learning methods. This is still work in progress, but first results are promising.

Explanation of predictions from Black-Box models

In some 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 where trust in the algorithm is required, like for example in clinical applications, fraud detection or credit scoring. In BigInsight we have an innovation area called "Explaining AI", which focuses on explaining black box models. There has been a big interest in explaining AI in the context of personalised marketing. Having in mind problems from this area, we have developed proper modelling of the dependence between explanatory variables in explainable postprocessing. In 2020 we have submitted one paper to the journal "Artificial Intelligence", and we are currently working on another paper that we plan to submit to the journal "Dependence Modelling". In addition, we have produced an R package, "shapr", which is available at CRAN.

Principal Investigator
Kjersti Aas

co-Principal Investigator
Ida Scheel

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 2020:

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. As reported last year, 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. Our paper is accepted for publication.

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

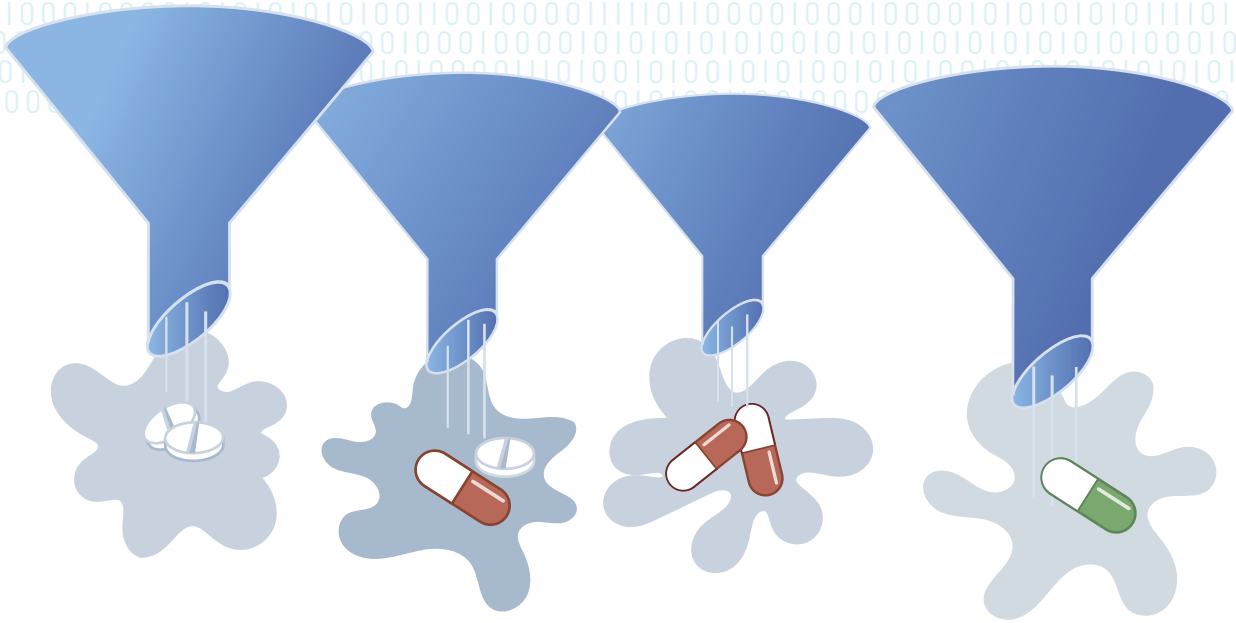
Cancer pharmacogenomic screens profile cancer cell lines versus many potential anti-cancer drugs to identify new combinations of drugs that have a high probability to work on individual patients. We work with data generated by our partners at Oslo University Hospital and public data to guide therapy based on the statistical prediction of how drugs will behave for individual tumor samples. To improve predictions, we are exploring both structured penalised regression models and 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 and corresponding software. Two papers have been published, one is accepted and two are submitted. In addition, we have published two mature R-packages where we have implemented our methods. BigInsight-affiliated PhD candidate Zhi Zhao has successfully defended his PhD thesis on "Multivariate structured penalized and Bayesian regressions for pharmacogenomic screens."

Healthcare safety management

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. As reported last year, 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 prediction models that are more interpretable. The method has been tested on publically available data, showing good results. One paper is published. Furthermore, we are working on electronic health record data from Akershus University Hospital (AUH) on a project to explore patients' movements within the hospital and how this affects the risk of spread of infections. The project is mostly descriptive and aims to inform decision makers about the amount of partly unnecessary movements that are taking place. The project is highly relevant in this period of the covid-19 pandemic and is made possible by the unique data from AUH.

Exploring clonal heterogeneity in blood cancers for personalised treatment.

Our goal is to develop a data-driven modelling framework to improve treatment strategies in blood cancers. BigInsight has strong clinical and experimental collaborations in blood cancer at OUS as well as access to unique datasets. One major obstacle to developing personalized medicine is the presence of cellular heterogeneity within the cancer cell population of each patient. This can lead to a common scenario where a therapy initially succeeds at reducing disease burden, but the cancer eventually rebounds due to the outgrowth of a minor drug-resistant clone. To address this obstacle, we have started to develop a new method to estimate and quantify the heterogeneity present in each



particular cancer. We use available high-throughput drug screening data to infer the subpopulation substructure. We will develop a statistical platform to help identify the number of distinct clones present as well as how these clones respond to specific drugs based on drug screens of patient samples. This information then feeds into evolutionary models of drug response to therapy, to predict the effect of a drug. We use a combination of mathematical modelling and mixture inference. Successful implementation of our method will potentially greatly aid in the management of different types of blood cancers, and potentially also solid cancers.

Mathematical models and Bayesian inference in personalised breast cancer therapy

Current personalized cancer treatment is based on biomarkers which allow assigning each patient to a subtype of the disease, for which treatment has been established. Such stratified patient treatments represent a first important step away from one-size-fits-all treatment. However, the accuracy of disease classification comes short in the granularity of the personalization: it assigns patients to one of a few classes, within which heterogeneity in response to therapy usually is still very large. In addition, the combinatorial explosive quantity of combinations of cancer drugs, doses and regimens, makes clinical testing impossible. We follow a new strategy for personalised cancer therapy, in silico, based on producing a copy of the patient's tumour in a computer, and to expose this synthetic copy to multiple potential therapies. We show how mechanistic mathematical modelling, patient specific inference and simulation can be used to predict the effect of combination therapies in a breast cancer. The model accounts for complex interactions at the cellular and molecular level and is able of bridging multiple spatial and temporal scales. The model is a combination of ordinary and partial differential equations, cellular automata, and stochastic elements. The model is personalised by estimating multiple parameters from individual patient data, routinely acquired, including histopathology, imaging, and molecular profiling. The results show that mathematical models can be personalized to predict the effect of therapies in each specific patient. The approach is tested with data from five breast tumours collected in a recent neoadjuvant clinical phase II trial. Recently we have been able to develop a numerical algorithm that allows the simulation of a full biopsy, exploiting parallel computing. This study is possibly the first one towards personalized computer simulation of breast cancer treatment incorporating relevant biologically-specific mechanisms and multi-type individual patient data in a mechanistic and multiscale manner: a first step towards virtual treatment comparison.

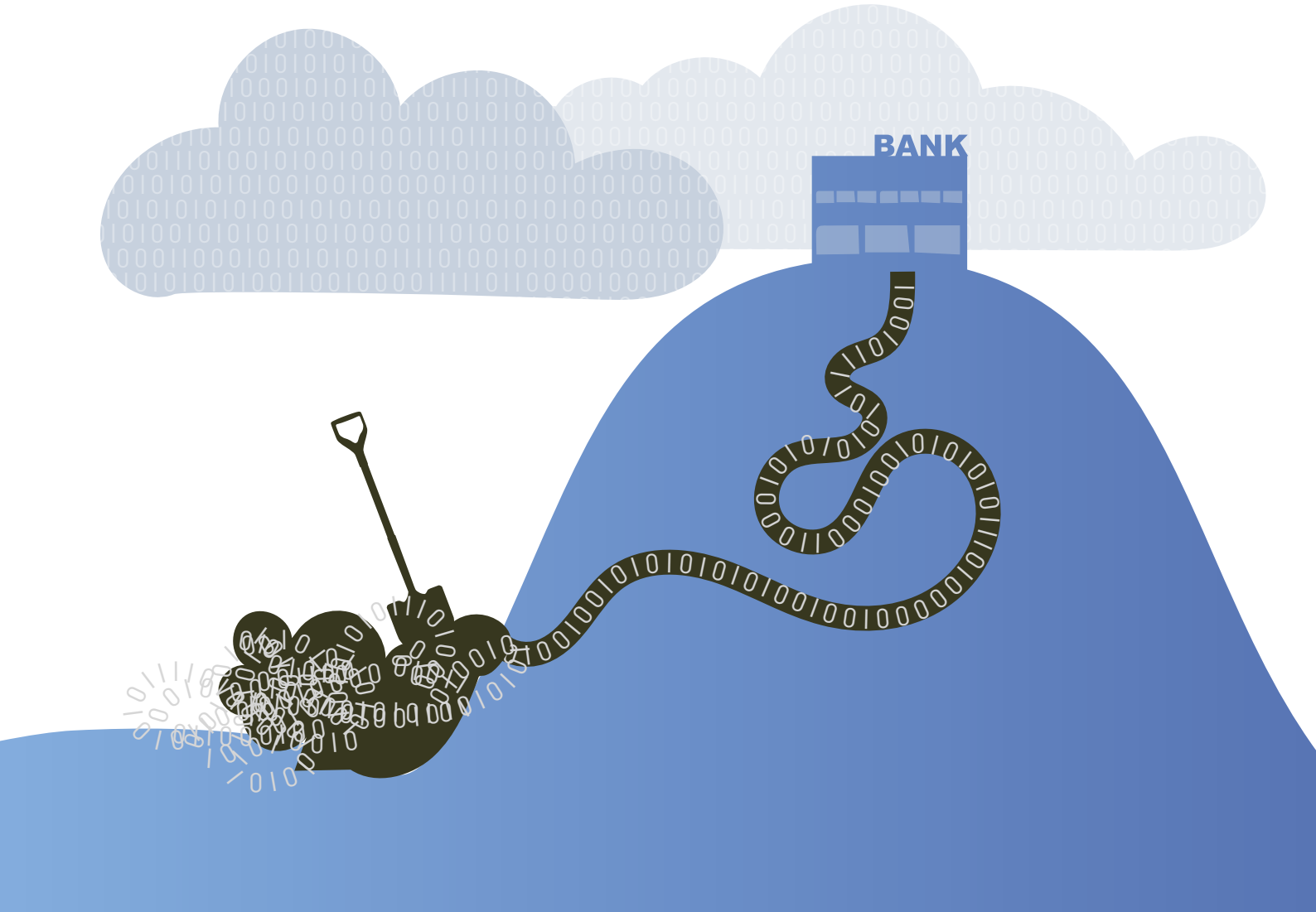
Principal Investigator
Magne Thoresen

co-Principal Investigator
Clara Cecilie Günther

PERSONALISED FRAUD DETECTION

Insurance fraud is expensive, affects insurance prices for all customers and is therefore important to detect and prevent. Soft fraud, the exaggeration of legitimate claims, is quite diffuse and difficult to spot. A sustainable welfare system requires implementation of effective measures to limit fraud, such as tax avoidance and tax evasion. Money laundering is also a serious threat to the global economy.

Fraud detection can be seen as a regression/forecasting problem, where fraud (true/false) is the response, possibly with a potential economic loss, and there are a great number of covariates connected to each case, especially if one considers interactions. Further, the data are class imbalanced, in the sense that the number of investigated fraud cases is generally low compared to the total number of cases. Another challenge is that the data are gathered over time, and that the quality may vary. In addition, only a small subset of the total number of cases is controlled. The objective is then to produce a trustworthy and interpretable probability of fraud for each new case, that can handle structured and unstructured data, including transactions, relational networks, and other available digital records in a privacy responsible setting.



What we did in 2020:

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 have been 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. The results are, however, so far not as promising as we wished, and we are evaluating new approaches.

Statistical embeddings: A survey

The network analysis work has spurred us to write a survey paper on statistical embeddings, from principal components, via non-linear embeddings and topological embeddings and topological data analysis to embeddings on networks. The paper will be submitted early 2021.

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 has been 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 2020, we have further 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. The method is being set into production by Gjensidige and there is interest from other BigInsight partners as well.

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. We term this notion fraud-loss and have proposed a framework for choosing the best model according to the fraud-loss. The results are promising, and a paper is submitted.

Principal Investigator
Anders Løland

co-Principal Investigator
Martin Jullum

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 at a later point, 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 manoeuvres and collision avoidance.

What we did in 2020:

Scalable change and anomaly detection

In March 2020, the first disputation in the sensor IO took place as industrial-PhD Andreas Brandsæter from DNV-GL defended his thesis "Data-driven methods for multiple sensor streams, with applications in the maritime industry". This defence was also the first fully digital PhD defence at the University of Oslo during the first lockdown due to covid-19.

In October 2020, BigInsight PhD Martin Tveten handed in his PhD thesis "Scalable change and anomaly detection in cross-correlated data", defended in January 2021. The thesis contains two papers on dimension reduction with PCA specifically tailored for change point detection, one on the motor overheating detection method developed for ABB, and the last on scalable changepoint and anomaly detection in cross-correlated data with an application to condition monitoring, with data from subsea equipment made available through DNV-GL.

The motor overheating detector of ABB was in 2020 successfully retrained on new, high frequency data (seconds). Furthermore, a master student has worked on investigating the possibility of further improving motor overheating detection, using recurrent neural net algorithms such as LSTMs. The master thesis will be finalized in spring 2021.

Combining AI and expert knowledge for more efficient monitoring

A typical monitoring system 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. Analysing this kind of data has been a great challenge but has led to both strong methodological contributions as well as innovation at ABB this year. We have finalized a paper on clustering and automatic labelling within time series of categorical observations, and a second paper is in preparation. The resulting clusters from the log files are highly informative and the BigInsight PhD student in this project has been part time hired by ABB in 2020 to help in the implementation of these methods in the ABB solutions.

A master student has been working in 2020 with combining all available sensor data and logged error messages for a certain vessel from ABB, to find out if it is possible to build

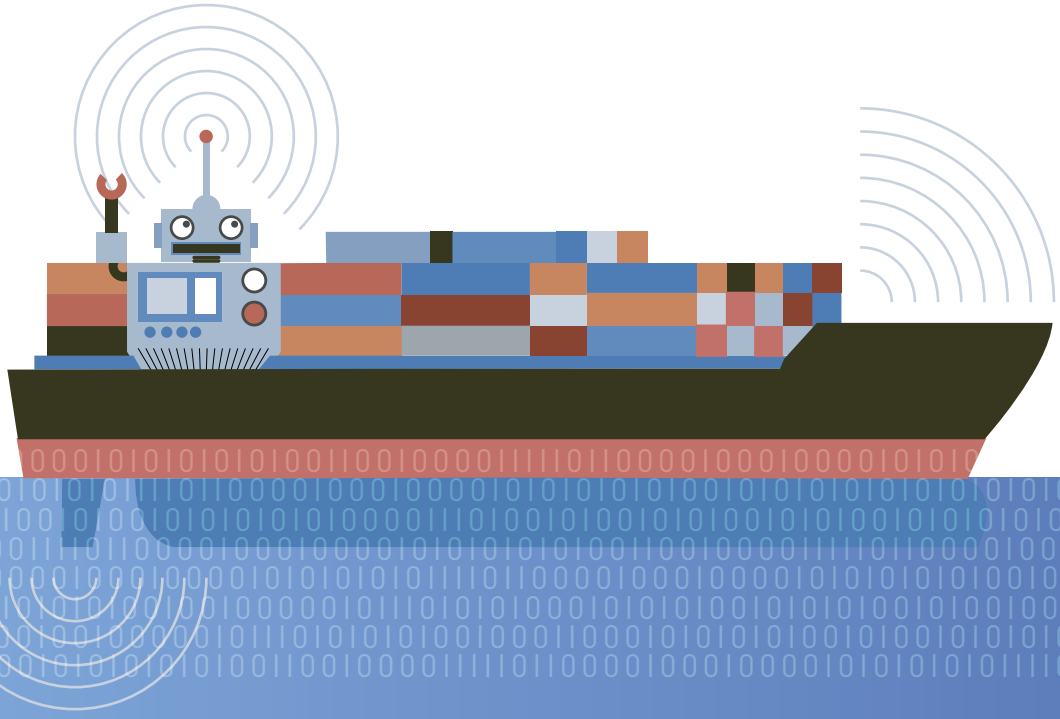
a predictive detector for one specific event (critical trip). Results are very promising, and the master thesis will be finalized in spring 2021.

The detection of incipient bearing faults as early as possible has great economic value in monitoring critical rolling element bearings (REBs) in industrial applications. A sudden failure of any bearing in the equipment results in huge financial losses. We have started to work on new methodology based on a cascade of precisely tuned filters, to detect a fault in a REB as early as possible by analyzing vibration signal measured at the bearing housing. A NæringsPhD with ABB is making very relevant progress in this direction.

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

Autonomous ships are run by algorithms. These algorithms are trained to be able to maneuver the ship according to international rules and in difficult situations. DNVGL will certify such ships, and therefore their algorithms. The International Maritime Organization has approved interim guidelines for autonomous ship trials. Consequently, major stakeholders have started test projects on sea. Simulation-based testing is an alternative and very attractive approach. The idea is motivated by the availability of digital-twin ship models and the software form of autonomous navigation algorithms. Simulation-based tests offer many advantages: the testing with numerous scenarios is faster, cheaper, and hazardless, the ability to control the tests and particularly their complexity. In collaboration with DNV-GL, we have designed realistic navigation testbed scenarios from huge historical traffic data integrated with high resolution digital maps, vessel information registries, and digital nautical charts. In short, our algorithm scans 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. Our method is able to detect situations in the AIS data which could have led to a collision or a grounding in a given time window ahead. Scoring situations allows to classify them in various ways of complexity and type, so that one can search for test-cases as wished.

In 2020, collision avoidance rules (COLREGs) have been carefully and successfully integrated in the algorithms, allowing all vessel-to-ground and vessel-to-vessel interactions to be efficiently analysed through a hierarchical method for identifying 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. This algorithm will be part of a large autonomous navigation vessels system (which for example will have to also certify sensors etc) owned by DNV-GL. Two papers are published and in revision and various sample scenarios are under implementation by DNV-GL in Trondheim.



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

Norway has decided that all ferries operating in Norwegian waters should be emission free within few years. Lithium-ion batteries are by far the most popular solution for electric ferries. During 2020, we have been working on a project with DNV-GL connected to sensor-based lithium-ion batteries diagnostics. We are partly working with publicly available data, and partly with operational data from DNV-GL´s collaborator Corvus Energy, which is a major producer of maritime lithium-ion batteries operating in several ferries in Norway. The aim is to develop methods for data driven monitoring of battery health, based on historical data from operating vessels provided by the battery producer. We have finalized a study of SOH (State of Health) degradation under dynamic conditions using publicly available data, comparing statistical models (fractional polynomials) and sequential deep learning (neural networks), with excellent prediction results. The reason for using public data for initial studies is that there are extremely few measurements of state of health in the operational data. In order to be able to do the same type of SOH degradation modelling on these data, we work on a method for extracting pseudo capacities that can be used instead.

Two master students have started with battery diagnostics master projects in the end of 2020. One of them performs a comparative study of the whole specter of machine learning methods for the SOH degradation with the publicly available data, while the other master student exploits error-in-variable models for total capacity estimation.

Inferring the effect of marine bio fouling on loss of performance

Marine biofouling on a ship's hull and propeller increases the resistance of the ship moving through water and may seriously influence the propulsion efficiency of the ship. Loss of performance means for example increased fuel consumption. DNV-GL has collected operational data from the fleet of a shipping company over several years, which we have used to study loss of performance due to bio growth (fouling) on hull and propellers. Data come from various ships, with timepoints for hull and propeller washing and a large amount of time series of relevant operational measurements. Since the amount of bio fouling itself is never measured, the modelling aims at inferring the "hidden effect" of this time varying process. We have used both a Bayesian approach based on Piecewise Splines and INLA, and a Random Forest. The results provide valuable insight into the bio fouling process and will be used by DNV-GL for assessing optimal hull and propeller cleaning strategies.

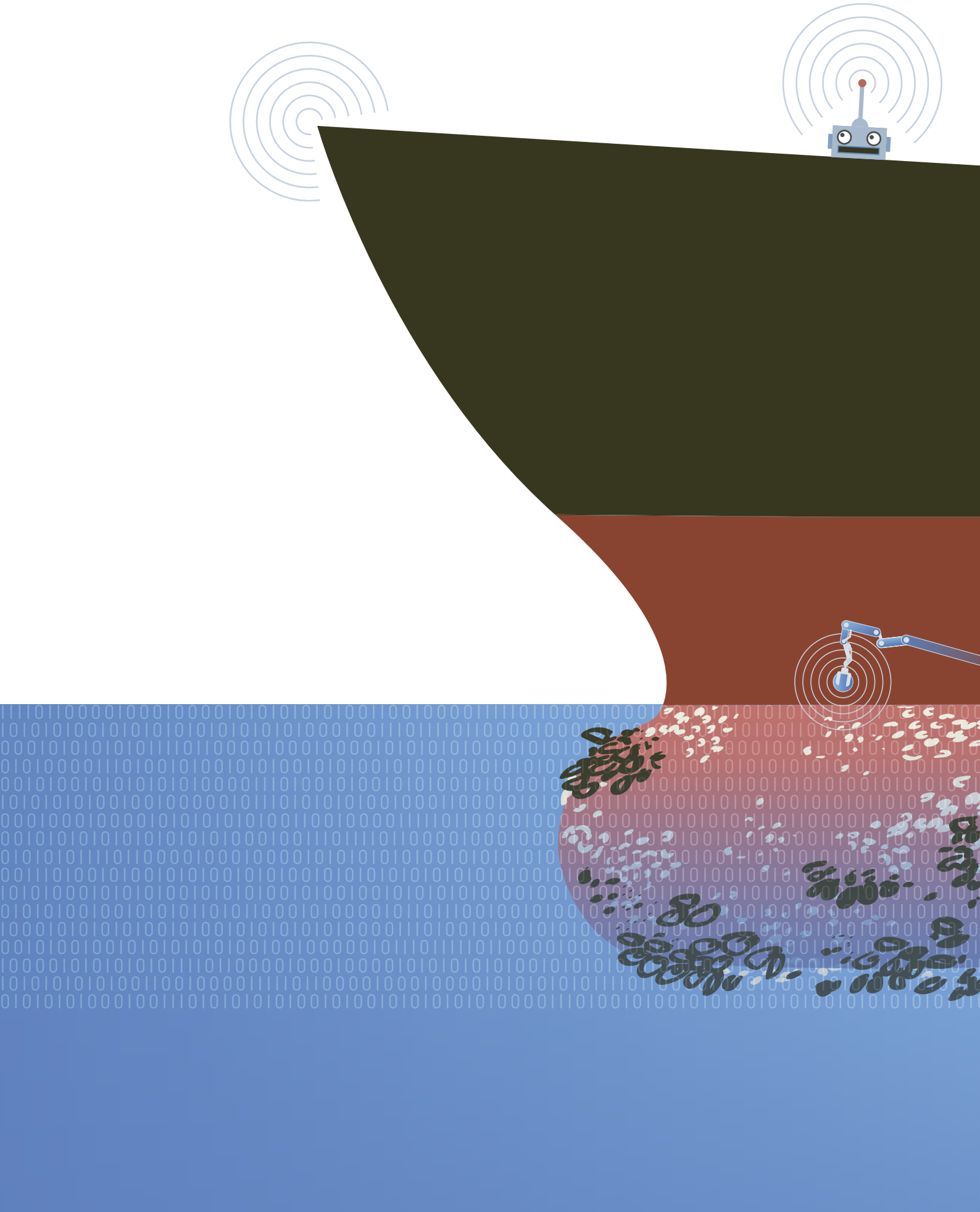
Protocol for combining data sources with misclassifications, maintaining privacy (SSB)

Integration with other data sources is often needed to overcome the various known deficiencies of administrative data. Our motivating problem is delay of administrative reporting that causes misclassification of register-based employed status, at a time immediately after the statistical reference month. The Labour Force Survey (LFS) provides an additional employed status, albeit aimed at a different definition of employment. Moreover, the LFS suffers from survey nonresponse, such that the LFS respondents from which both the measured variables are jointly observed is a nonprobability sample in reality. In 2020, we have written one paper where we develop models for adjusting two fallible classifiers jointly observed in a nonprobability sample. Comparisons are made to the results obtained from hidden Markov models (HMM). Our approach facilitates integrated use of available data, such that timely statistics can be produced with minimum cost. Unlike using the HMM models, our method requires only aggregated cell-level counts instead of individual-level data. It is more readily applicable to other big-data sources, such as a large nonprobability sample of employed status classified based on mobile phone movement data. A second paper, expanding the models above, is also in progress.

Due to the pandemic situation in 2020, important arrangements for the student society such as Data Science Day at UiO, BigInsight Day and the yearly Klækken PhD workshop, usually sponsored and fully or partially arranged by BigInsight, have been cancelled. Instead, various series of digital seminars and meetings, journal clubs etc. have been tried out to keep in contact and scientifically updated. PhD students and postdocs have been prioritized in periods when access to the university campus has been limited.

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 2020:

Price Uncertainty via Quantile Arrays and the featurization of Bid/Ask Curves

Our primary efforts in have been to unlock the potential value from the published Bid Ask curves in the Nordpool market. These curves have proven to be an especially difficult data source to take advantage of, while their potential for improved price uncertainty forecasting is substantial.
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 model, we interacted extensively with our industrial partner Hydro. The Bid/Ask model was quite general, however it suffered from a number of side cases which would occasionally lead to non-sensical price distributions. After considerable experimentation, in 2020 we had a breakthrough which resolved most of the open issues in the previous methodology. We settled upon a two-step procedure that has finally proven to be a robust method for using the bid/ask data. In particular, we first found useful ways of "featurizing" the bid/ask data, converting the underlying data source from a high-dimensional and varying size curve into a small feature set. This feature set was then passed through a flexible array of quantile regressions. The improvement from this approach is substantial. Computation time has been dramatically reduced, mode performance has markedly improved and the final modelling framework is both easy to explain and diagnose. The new system has already been put in place in Norsk Hydro's production forecasting system and we interact with Hydro staff on the model performance regularly.

Developed new models for renewable energy production

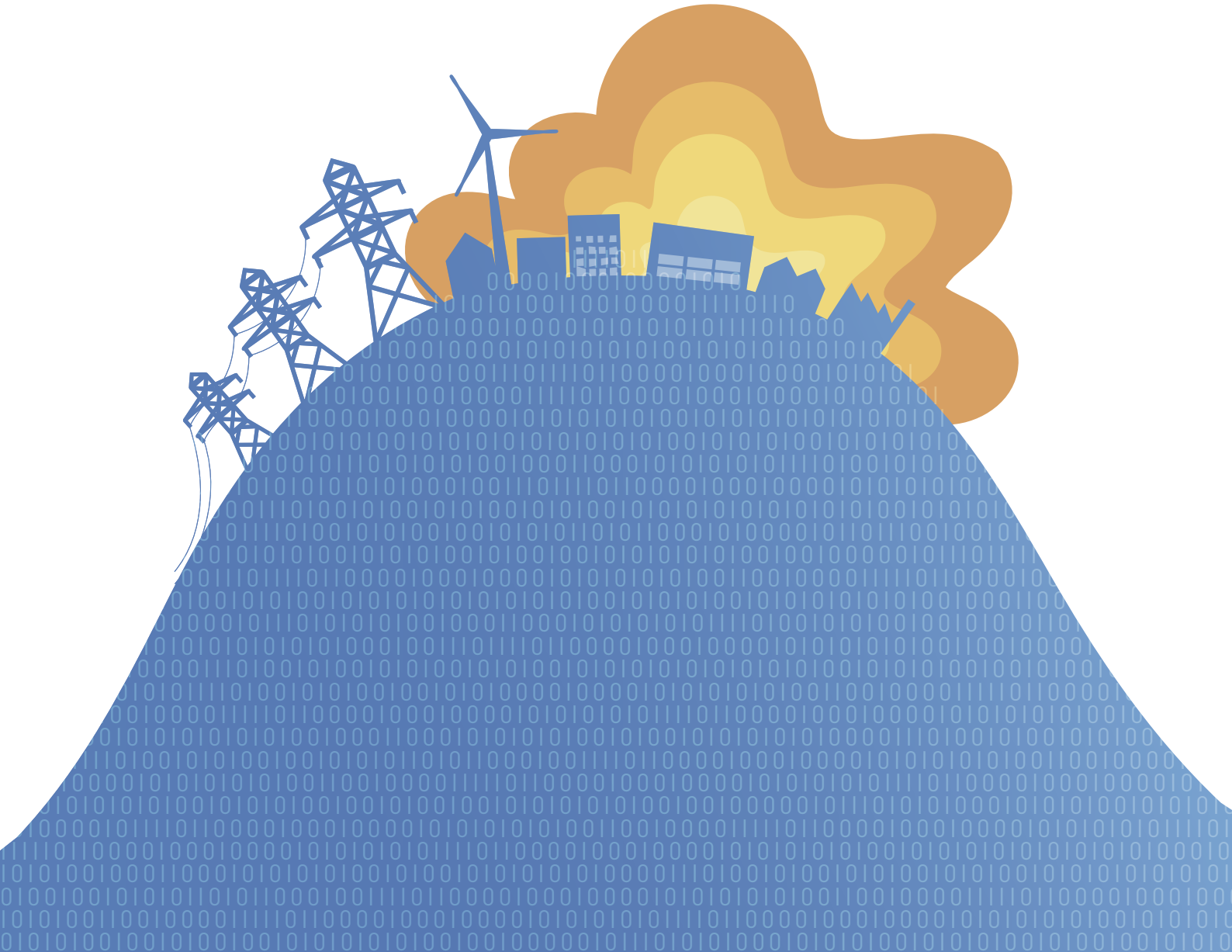
The amount of renewable energy to be produced in a given hour has a substantial and increasing effect on European electricity prices. Therefore, it of growing importance to accurately forecast this production in order make a high-quality electricity price forecast. We investigated a number of approaches for improving existing forecasts of renewable energy production. We found that a massive-dimensional ridge-regression framework substantially outperformed the current best practice. With substantial interaction with Hydro personnel, we implemented this new model in the Hydro production forecasting system.

Developed new models for wind speed power production using rapid forecast trajectory post processing

We worked with Hydro staff on a problem in statistical weather post processing. In particular, wind speed forecasts from the ECMWF model are issued every 6 hours for the next two weeks ahead. However, these models take time to run and disseminate, which means that a forecast run has not been published until 8 hours after the forecast was initialized. When a wind park has observational data at the site, it can be used to correct the stale wind speed forecast. We developed a statistical post processing technique was on the Rapid Adaptation of Forecast Trajectories (RAFT) approach that proved to be highly effective at correcting this feature of the ECMWF model output and led to improved wind production forecasts.

Principal Investigator
Alex Lenkoski

co-Principal Investigator
Carlo Mannino



EXPLAINING AI

At the intersection between artificial intelligence, transparency, privacy and law, there is a need for more research. This IO, which we started in 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:

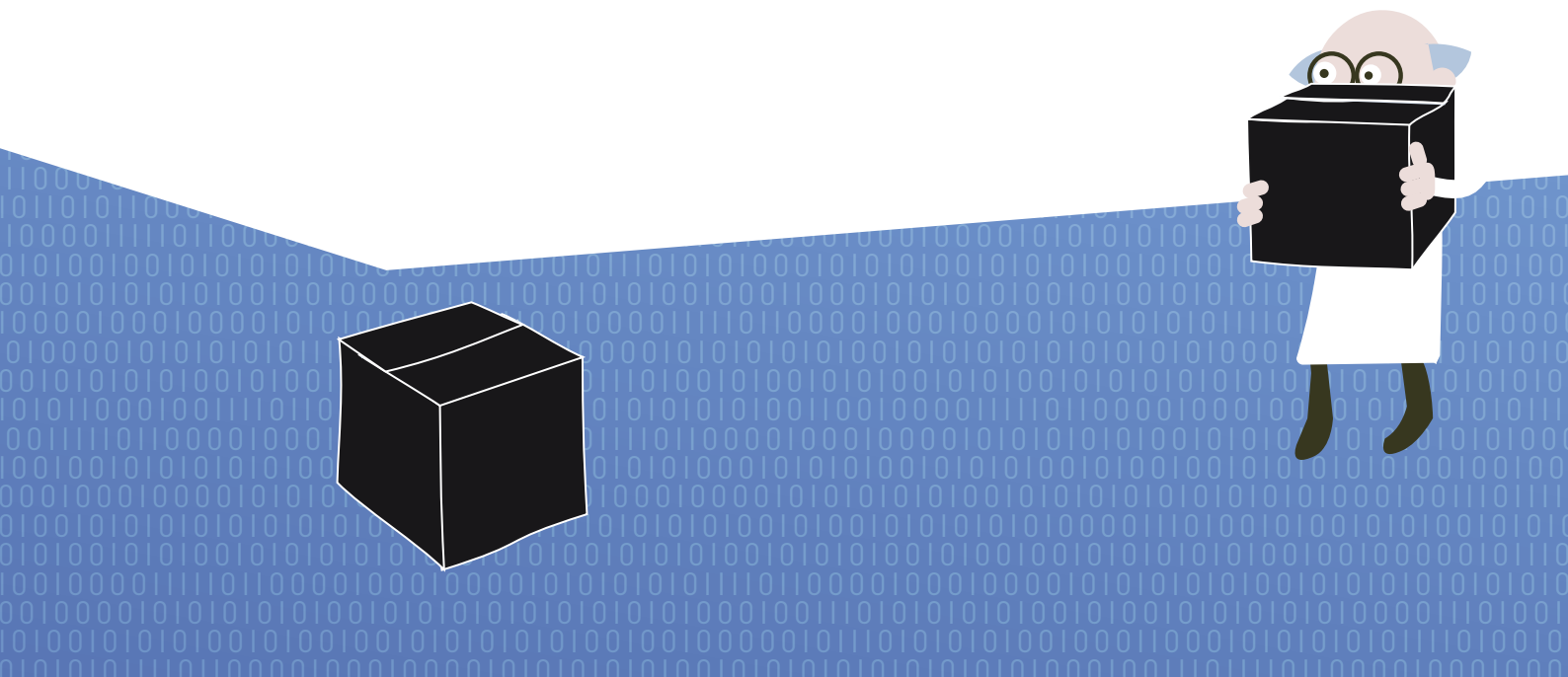
1. Those who construct or use the model should understand how the model works
2. Those who are exposed to the model should, and sometimes will, have the right to an explanation about a model's behaviour, for example to be able to contest its decision
3. It should be possible to detect undesired effects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

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 2020:

Seminar series

We organized two seminars, with the themes: "Explainable Artificial Intelligence: How Subsets of the Training Data Affect a Prediction" and "Explaining predictive models with mixed features using Shapley values and conditional inference trees". Attendance and discussions were very good, and the seminar series continues into 2021 (with very well attended webinars and, we hope, seminars in person as well). We also co-organised a very successful one day workshop "Responsible use of Data and AI" with 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 continue to add new features to our R package – *shapr*. We also continue to improve our methods further to 1) handle categorical variables better (conference paper published), 2) by grouping similar features for improved efficiency and interpretability (conference paper will be submitted), 3) by modelling the dependence between features with non-parametric vine copulas (paper submitted) and 4) using Shapley values to assess how subsets of the training data affect a prediction (paper submitted). Besides, in 2020 a PhD student joined the project, to further strengthen this line of research.

Practical testing of explanations on use cases

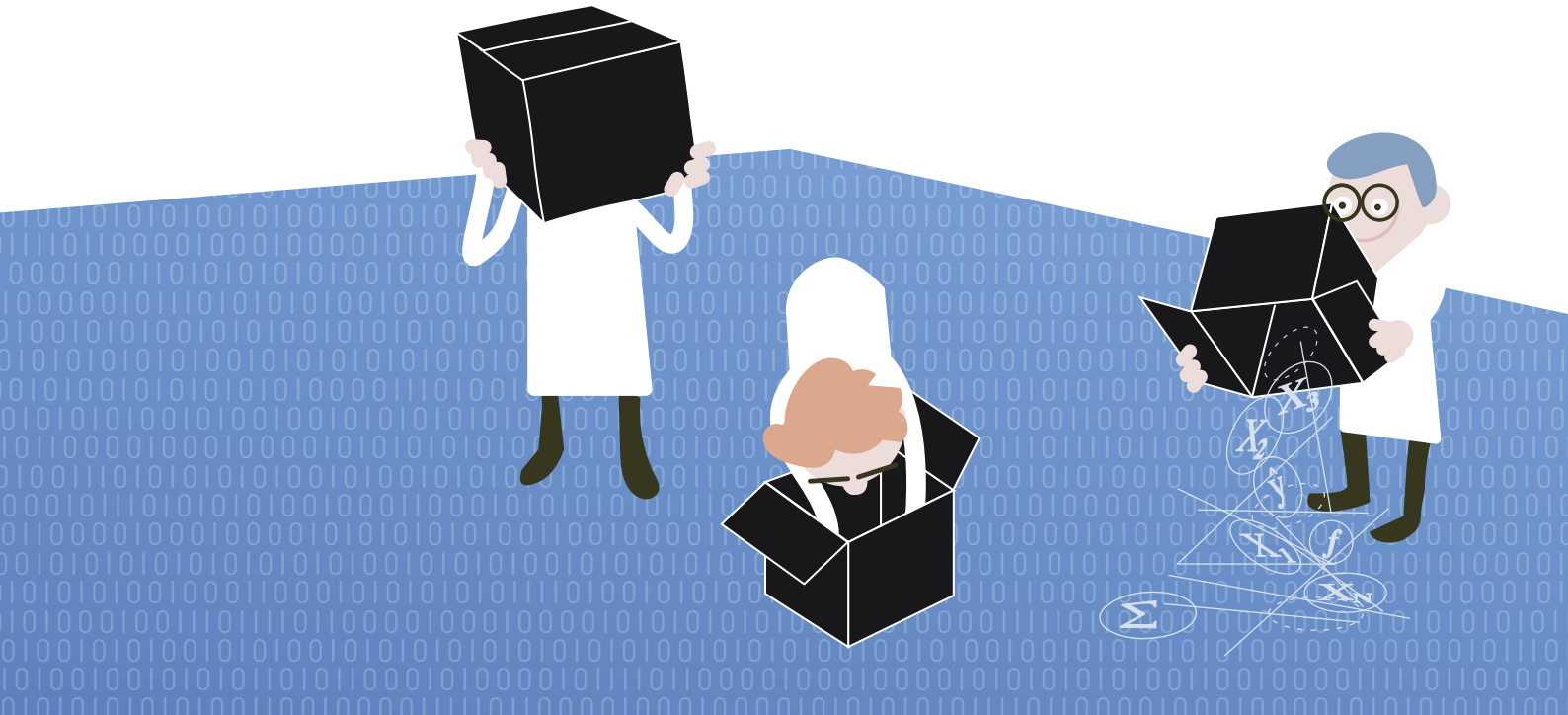
Even though the explanation methods we develop are mathematically sound and correct, it is not obvious that they are immediately useful for executive officers or end users. We will therefore investigate how test groups understand these explanations, to learn and further develop how the explanations can be explained or utilised better. We are working with NAV on these issues, who are using and providing very useful feedback on the aforementioned *shapr* package. Further BigInsight partners can follow suit in 2021.

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 and UiO (Department of Public and International Law), which will lead to an industrial PhD. We have been organizing a working group on methods for synthetic data. The group currently includes SSB, NAV, Skatteetaten, DNB, Lånekassen and Riksrevisjonen. Moreover, we have contributed to various XAI seminars and the course "Legal Technology: Artificial Intelligence and Law" at the Department of Public and International Law, UiO.

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. Professor 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.

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

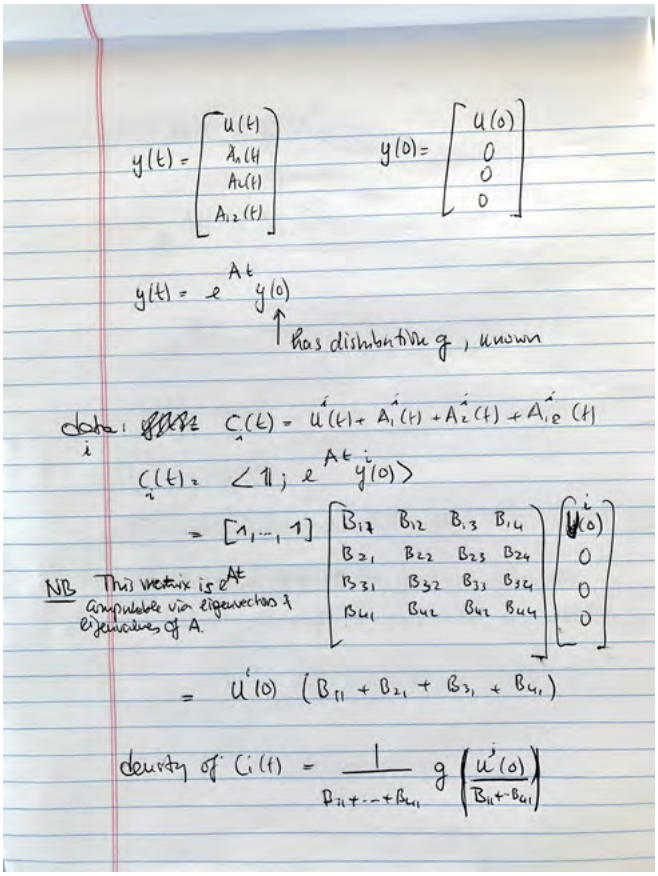
The Medical Research Council Biostatistics Unit (BSU)

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, application and dissemination of statistical methods'. BSU's critical mass of methodological, applied and computational expertise provides a unique environment of cutting edge biostatistics, striking a balance between statistical innovation, dissemination of methodology and engagement with biomedical and public health priorities. Professor Sylvia Richardson is director of the BSU and she has received an honorary degree of the University of Oslo. BigInsight and the BSU have several joint projects in health and molecular biology. We have also involved the BSU in our collaboration with the University of Hawassa (Ethiopia). PhD student Zhi Zhao spent a research time at BSU in 2019. We are partners in RESCUER, a H2020 project.

Professor Sylvia Richardson, MRC Biostatistics Unit, Cambridge

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 is also supported by an INTPART NFR funded project that BigInsight obtained.



University of Hawassa and University of Jimma, Ethiopia

Funded by Norhed, Norpart and the Norwegian Agency for Development Cooperation NORAD, and in partnership with NTNU and UiO, BigInsight concluded in 2020 a ten year project with the University of Hawassa. In 2020, we supervised 4 PhD students on 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: Teshome Kabeta and Henok Asefa, working on dietary diversity in rural Ethiopia and non-communicable diseases in Addis Abeba. We will apply for renewed funding in 2021.

International guest programme

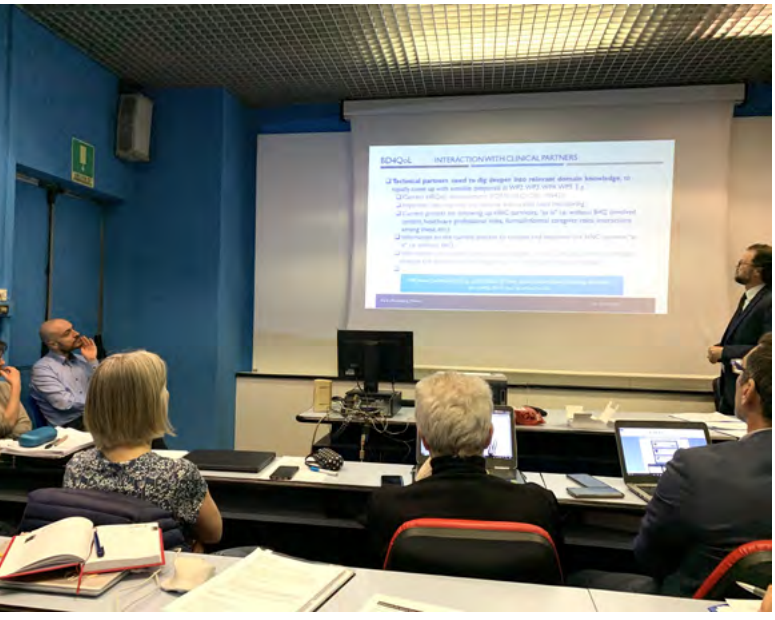
BigInsight has an international guest programme, funding from short visits up to long-term visiting and adjunct positions and a sabbatical visitor programme. In 2020, the programme could not host any visit because of the covid-19 pandemics. As soon as possible, we will start our guest programme again.

Other International activities

PhD students from other universities spent periods of training and research collaboration at BigInsight. This has however been impossible in 2020.

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 monitoring Big 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.

Scientific Advisory Committee of BigInsight

Scientific Advisory Committee of BigInsight has five international members. A meeting was planned in 2020, but could not be organized. We aim to a new try in 2021.

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 Cambrige, 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 IN 2020

In 2020 the following PhD students affiliated to BigInsight defended their PhD thesis:

Martin Tveten, Department of Mathematics, defended his PhD thesis "Scalable change and anomaly detection in cross-correlated data" on Jan. 21, 2021.

Supervisor: Ingrid Kristine Glad

Trial lecture: "Covariance matrix estimation in high dimensions"

Adjudication committee

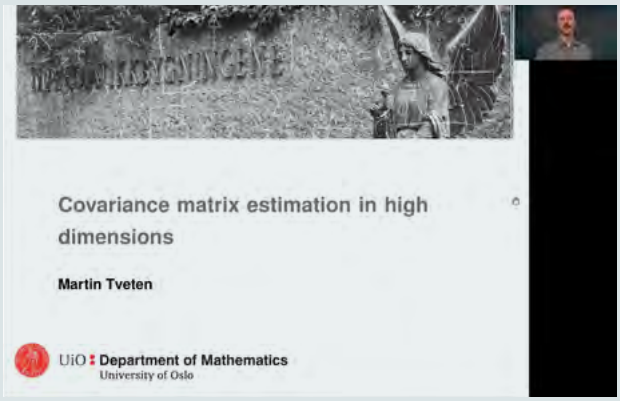
- Senior researcher Stéphane Robin, AgroParisTech/Université Paris-Saclay
- Senior lecturer Haeran Cho, University of Bristol
- Associate professor Johan Pensar, University of Oslo

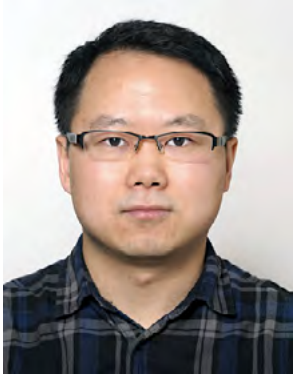
Summary

Both in science and industry, the sizes of data sets are growing. It is not uncommon to encounter sets containing millions or even billions of measurements. Without appropriate tools for turning such enormous amounts of data into insight, however, the data's value is severely limited.

Apart from consisting of many measurements, a common feature of big data sets is that some properties of the data change over time. Determining whether and when changes have taken place is important in many scientific problems. For example: Is the climate changing? Has the covid-19 reproduction number changed? Is the quality of manufactured cars stable? Moreover, monitoring changes in network traffic data can be used to detect cyber attacks.

Therefore, in this thesis, I have studied statistical methods for detecting changes and estimating when they have occurred. My collaborators and I have constructed efficient computer programs both for retrospective analysis of large data sets as well as for real-time analysis of streaming data. We have also demonstrated that detecting changes in a stream of data from temperature sensors could have prevented a costly and dangerous overheating event in a ship motor.





Zhi Zhao, OCBE, Institute of Basic Medical Sciences, defended his PhD thesis "Multivariate structured penalized and Bayesian regressions for pharmacogenomic screens" on Oct. 9, 2020.

Supervisor: Manuela Zucknick

Trial Lecture: «Model selection for time-to-event response variables in high-dimensional settings"

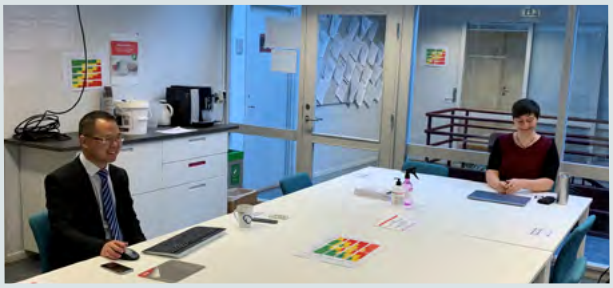
Adjudication committee

- Professor Anne-Laure Boulesteix, Ludwig-Maximilians-Universität München
- Associate professor Francesco Claudio Stingo, University of Florence
- Group leader Dr. Marieke Kuijjer, University of Oslo

Summary

Pharmacogenomic screens for personalized cancer therapy are the focused biomedical application in this thesis. Due to the complex relationships between targeted cancer drugs and high-dimensional genomic predictors, we have developed penalized likelihood methods and Bayesian hierarchical models to capture the complex structures in the pharmacogenomic data and to predict drug sensitivity.

The first part of the thesis proposed to address the correlations between drug sensitivity measures for multiple cancer drugs and the heterogeneity of multiple sources of genomic data in multivariate penalized likelihood methods with structured penalities. The proposed methods can improve the prediction performance of drug sensitivity. The second part of the thesis exploited Bayesian priors for the relationships between multiple drugs and relationships between drug sensitivity and the targeted pathways or genes of cancer drugs. Large pharmocogenomic screens may also include samples from multiple cancer tissue types. We employed random effects to address the sample heterogeneity in the proposed Bayesian model. The results have shown good structure recovery in the complex data and good prediction of responses by the new Bayesian models.



Håvard Kvamme, Department of Mathematics, defended his PhD thesis "Time-to-Event Prediction with Neural Networks" on May 28, 2020.

Supervisors:
- Ørnulf Borgan, University i Oslo
- Ida Scheel, Universitetet i Oslo
- Kjersti Aas, Norsk Regnesentral

Trial lecture: "Model-based vs. black box learning"

Adjudication committee

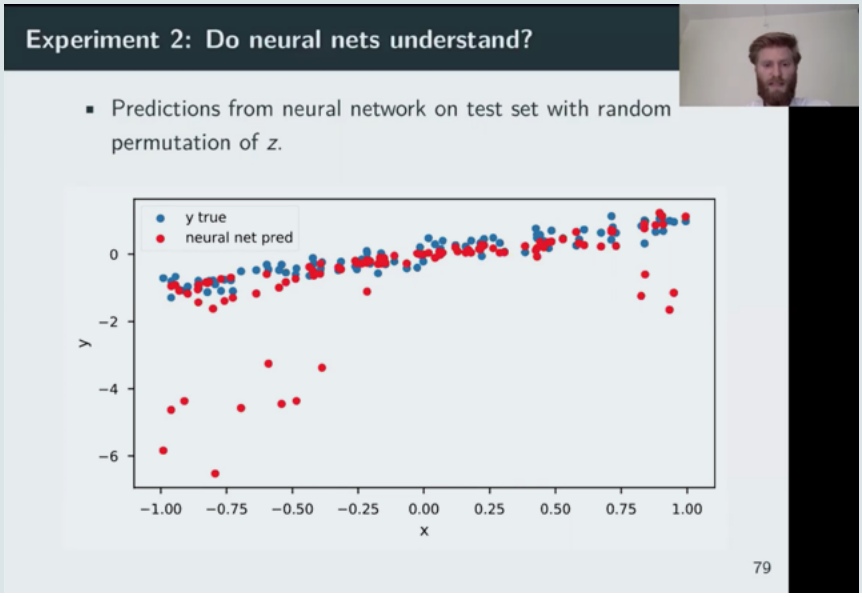
- Harald Binder, Albert-Ludwigs-Universität Freiburg
- Jan Terje Kvaløy, Universitetet i Stavanger
- Geir Storvik, Universitet i Oslo

Summary

In the last decades the analytical value of data has really become apparent and the amount of data collected has vastly increased. This enables us to approach problems in more data driven manners. In the thesis, I have combined recent developments in machine learning with statistical methods to better answer the question: "When in the future will a given event occur?"

The first part of the thesis was done in collaboration with the Norwegian bank DNB. We created new methods for predicting when in the future customers will default on their mortgage loans. By investigating the historical balances of the customers' checking accounts, savings accounts and credit cards, we found that we could improve on existing methods for predicting mortgage defaults.

In the second part of the thesis, our attention was directed toward more general methodology that may be applied to a number of problems. Our proposed improvements were illustrated using a selection of available datasets, ranging from how gene and protein expression profiles affect the mortality of breast cancer patients, to how customer information can help determine if customers are likely to continue to subscribe to a music streaming service.



Yinzhi Wang, Department of Mathematics, defended her PhD "Model selection and reinsurance optimization for general insurance" on Apr. 24, 2020.

Supervisors:
- Ingrid Hobæk Haff, Matematisk institutt, Universitetet i Oslo
- Erik Bølviken, Matematisk institutt, Universitetet i Oslo
- Arne Huseby, Matematisk institutt, Universitetet i Oslo

Trial lecture: «Machine learning in actuarial sciences"

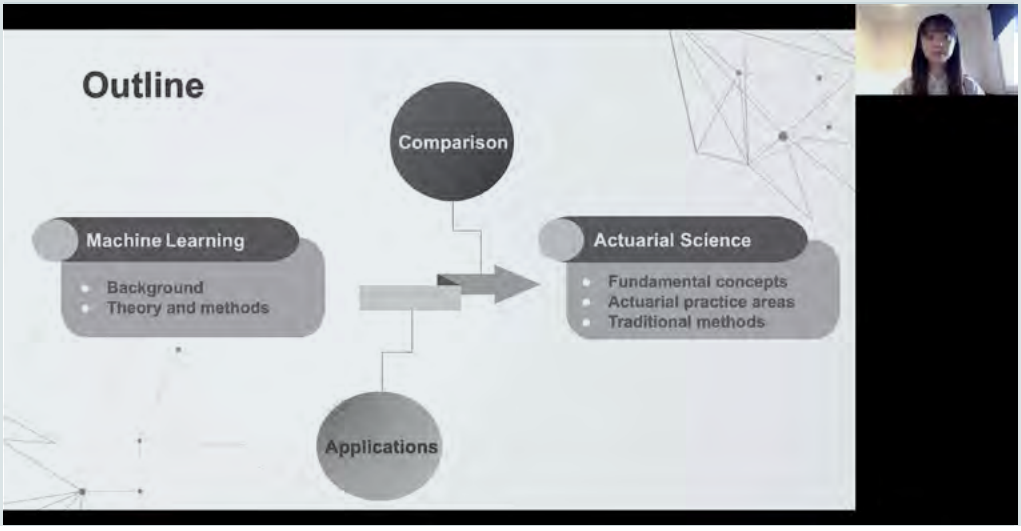
Adjudication committee
- Maria De Las Mercedes Ayuso Gutiérrez, Universitat de Barcelona
- Bård Støve, Universitetet i Bergen
- Ingrid Kristine Glad, Universitetet i Oslo

Summary

Insurance risk assessment is a core theme within general insurance. One of the major challenges centers around estimation and prediction of potential claims, since they have a big influence on how much the insurer will charge for the protection provided, and how much the insurer needs to reserve for future claims payments. In this work, I studied how the insurance risk is evaluated in the application of reserve estimation and reinsurance optimization, through pragmatic and efficient problem-driven approaches.

Finding good statistical models for losses due to claims is essential for insurers. Therefore, model selection methods for the claim severity distributions were investigated and compared under different sceneries, especially in the case of data limitation. A further question is how to model extreme claims. In this thesis, a variety of threshold selection methods for composite models were explored, indicating their strengths and weaknesses in real life situations.

The other major topic in this thesis is how much the optimal reinsurance solution is degraded by parameter and model error. Both asymptotics and numerical approaches were used to study how the errors behaved in the limiting case and with finite samples, separately. The results suggested that the shape of the claim severity distribution may not be of primary importance when designing an optimal reinsurance scheme. Moreover, it offered a simple solution for large portfolios, and addressed the practical importance of optimal reinsurance.



Andreas Brandsæter, Department of Mathematics, defended his PhD thesis "Data-driven methods for multiple sensor streams, with applications in the maritime industry" on Mar. 26, 2020.

Supervisors:
- Ingrid K. Glad, Matematisk institutt, Universitetet i Oslo
- Erik Vanem, DNV-GL
- Geir O. Storvik, Matematisk institutt, Universitetet i Oslo
- Magne Aldrin, Norsk regnesentral
- Arne B. Huseby, Matematisk institutt, Universitetet i Oslo

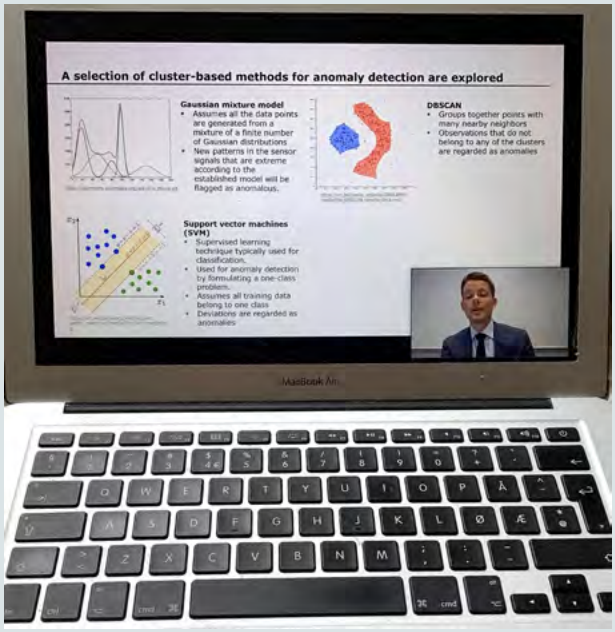
Trial Lecture: "Artificial intelligence for time series classification: theory and methods"

Adjudication committee
- Piero Baraldi, Politecnico di Milano
- Prasad Lokukaluge Perera, UiT Norges Arktiske Universitet
- Riccardo De Bin, Universitetet i Oslo

Summary

Når en datadrevet metode benyttes for å lage en prediksjon, er denne prediksjonen basert på et sett med historiske data. Vi kan for eksempel predikere ulykkesrisiko for et skip, og da er prediksjonen vanligvis basert på historiske data om andre skips ulykker, samt historiske data om det aktuelle skipet. Vi kaller ulykkeshistorikken for treningsdata, fordi vi bruker dette datasettet til å trene en modell som kan brukes til å produsere prediksjoner.

Slike datadrevne metoder er ofte svært komplekse og nærmest umulige å forstå og tolke. For å bedre brukerens forståelse av disse prediksjonene, har vi utviklet en metode for å kvantifisere effekten av ulike deler av treningsdatasettet. Dette lar brukeren for eksempel forstå hvordan treningsdata fra ulike tidsperioder påvirker prediksjonen.



Selv når vi ikke forstår bakgrunnen for en prediksjon kan den være presis. Vi utforsker en rekke forskjellige metoder for å sette modellene på prøve. Dette kan blant annet gjøres ved hjelp av gjentatte tester og såkalt kryssvalidering. Vi kan også gjøre små endringer i test og treningsdata, og undersøke hvordan dette påvirker resultatet. Et spesielt fokus er rettet mot utfordringer og muligheter innen autonom navigasjon av skip.

Metodene og problemstillingene vi presenterer og diskuterer er generelle, men anvendelsene er hovedsakelig hentet fra maritim næring. Vi anvender blant annet en rekke forskjellige datadrevne modeller for å predikere et skips fart i bølger. Videre undersøker vi ulike metoder for å detektere feil basert på analyse av sensordata.

Gražina Mirinavičiūtė, OCBE Institute of Basic Medical Sciences defended her thesis "Infections associated with varicella-zoster virus in Norway: disease burden and healthcare resource utilization" on Feb. 25, 2020.

Supervisors:
- Elmira Flem, Associate Director, Medical Affairs-Vaccines, MSD Norway
- Birgitte Freiesleben De Blasio, OCBE UiO, FHI

Trial lecture: "Facilitators and barriers for introduction of new vaccines in national immunization schedules"

Adjudication committee

- Pier Luigi Lopalco, University of Pisa, Italy
- Anette Siedler, Robert Koch Institute, Germany
- Professor Tron Anders Moger, University of Oslo

Summary

Nearly everybody is infected with varicella zoster virus (VZV), which causes varicella (chickenpox) and herpes zoster (shingles) (HZ). Varicella is usually a benign, but very itchy skin disease, predominantly occurring in childhood. HZ is a painful skin disease mainly affecting adults >50 years old. Both diseases can lead to serious complications. The burden of varicella and HZ is substantial and could be prevented by vaccination. Despite the availability of safe and efficacious vaccines, Norway currently does not implement vaccination programs against varicella and HZ.

The aims of the thesis were to characterize the healthcare burden of varicella and HZ in Norway in the pre-vaccine era by estimating rates of primary- and hospital care cases and assessing the levels of immunity against VZV in general population and among pregnant women. This aims to inform the national policy decision on the use of varicella and HZ vaccines, and guide current screening policies for varicella in obstetric populations.

Only 73% of the Norwegian population had immunity against VZV compared to >90% reported in most European countries. In addition, a small proportion of pregnant Norwegian women who were not immune to VZV got infected during their pregnancies, thereby increasing the risk of unfavourable health outcomes for themselves and for their offspring.

The data from national health registries showed that 10,881 varicella patients and 11,181 HZ patients were treated annually in primary care. Moreover, 361 varicella patients and 1,218 HZ patients were hospitalized annually. At least 47% of hospitalized HZ cases had complications. Very few patients were vaccinated against varicella and none against HZ during the study period 2008-2014.

Varicella and HZ cause considerable healthcare burden in Norway. There is an urgent need to develop robust knowledge-based national vaccine recommendations for both diseases and revise screening guidelines for VZV susceptibility in pregnancy.


Infections associated with varicella-zoster virus in Norway: disease burden and healthcare resource utilization

Grazina Mirinaviciute

University of Oslo, 2019

Thesis for the degree of Philosophiae Doctor (PhD)
Department of Infectious Diseases Epidemiology and Modeling, Division of Infection Control and Environmental Health, Norwegian Institute of Public Health

Institute of Basic Medical Sciences, Faculty of Medicine, University of Oslo

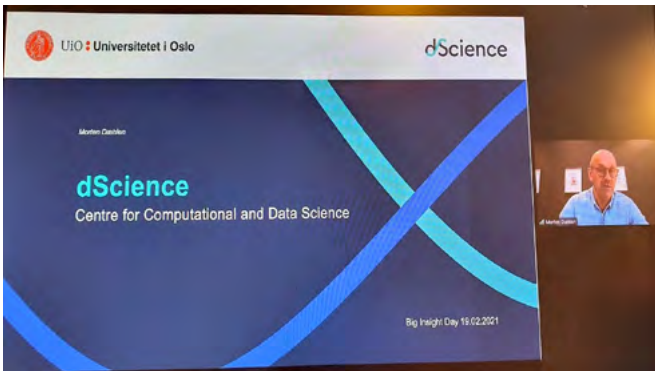
ACTIVITIES AND EVENTS

2020 BigInsight Day

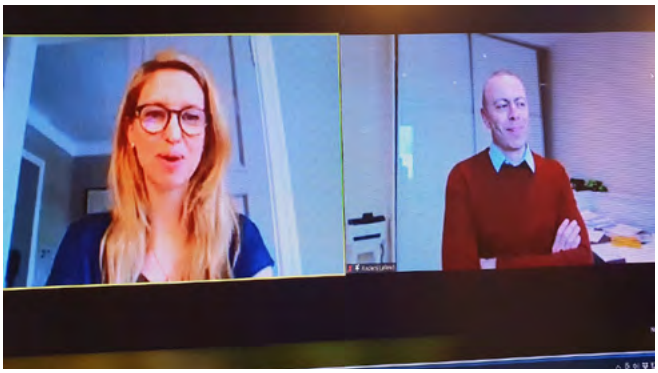
The annual BigInsight Day was postponed several times during 2020, as we were hoping for the possibility to arrange a physical meeting. It is finally scheduled for February 2021.

Oslo Data Science Day

The Oslo Data Science Day 2020 has been cancelled due to covid-19. We are planning the next Oslo Data Science Day 2021 in the autumn, with an exciting programme, which will include also the official opening of UiO's dScience centre for data and computational science.



Morten Dæhlen presents the new UiO centre dScience, which he leads.

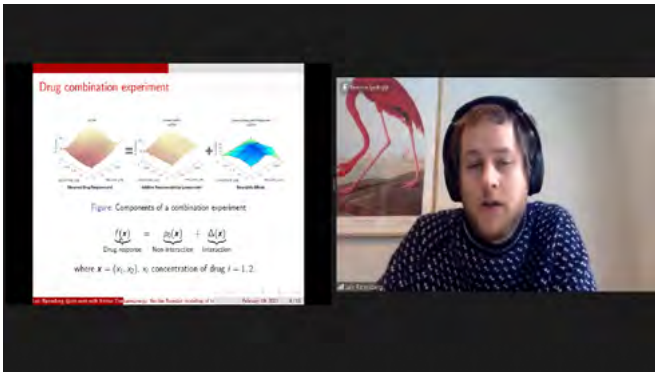


Kari Laumann, from Datatilsynet, introduces their regulatory sandbox, interviewed by Anders Løland.

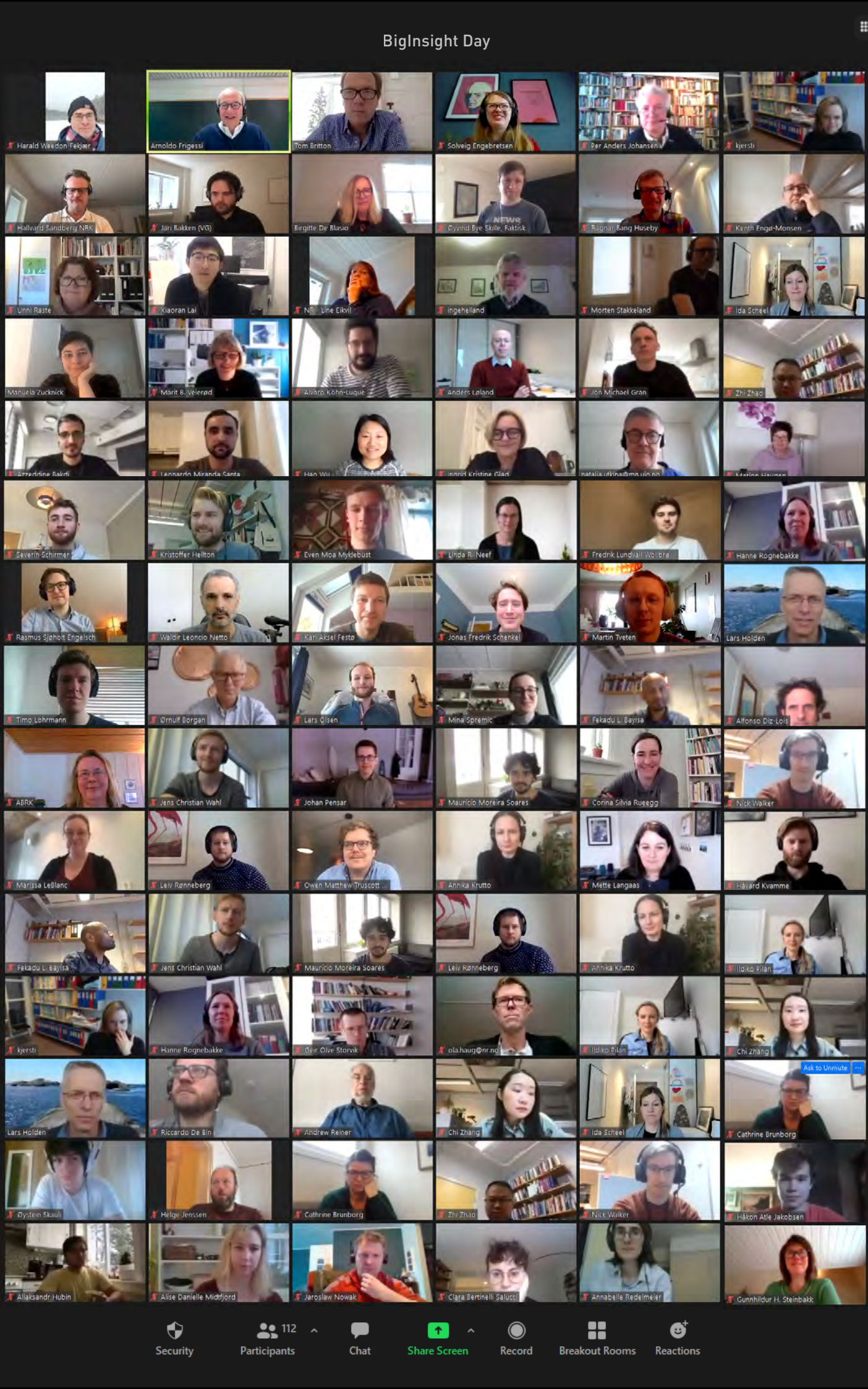
Intense discussion about models, uncertainty and communication between science and media with Jari Bakken (VG), Per Anders Johansen (Aftenposten), Hallvard Sandberg (NRK), Øyvind Bye Skille (faktisk.no), Tom Britton (Univ Stockholm), Solveig Engebretsen (NR), Birgitte De Blasio (NIPH). Arnoldo Frigessi led the debate.

Arnoldo Frigessi gave a status of BigInsight, with perspectives from science in emergence times.

Magician Giancarlo Scalia guided us in a new world of artificial intelligence, where intelligence does not help to understand what we see.



Leiv Rønneberg presents BayeSynergy


BigInsight Day
View
Harald Weedon Fekjær
Arnoldo Frigessi
Tom Britton
Solveig Engebretsen
Per Anders Johansen
kjersti
Hallvard Sandberg NRK
Jari Bakken (VG)
Birgitte De Blasio
Øyvind Bye Skille, Faktisk
Ragnar Bang Huseby
Kenth Engø-Monsen
Unni Røste
Xiaoran Lai
NR... Line Eikvil
ingehelland
Morten Stakkeland
Ida Scheel
Manuela Zucknick
Marit B. Veierød
Alvaro Köhn-Luque
Anders Løland
Jon Michael Gran
Zhi Zhao
Azzeddine Bakdi
Leonardo Miranda Santa
Hao Wu
Ingrid Kristine Glad
natalia.utkina@mn.uio.no
Severin Schirmer
Kristoffer Hellton
Even Moa Myklebust
Linda R. Neef
Fredrik Lundvall Wo. bra...
Hanne Rognebakke
Rasmus Sjoholt Engelsch
Waldir Leoncio Netto
Kari Aksel Festø
Jonas Fredrik Schenkel
Martin Tveten
Lars Holden
Timo Lohrmann
Ørnulf Borgan
Lars Olsen
Mina Spremic
Fekadu L. Bayisa
Alfonso Diz-Lois
ABRK
Jens Christian Wahl
Johan Pensar
Mauricio Moreira Soares
Corina Silvia Rueegg
Nick Walker
Marissa LeBlanc
Leiv Rønneberg
Owen Matthew Truscott ...
Annika Krutto
Mette Langaas
Håvard Kvamme
Fekadu L. Bayisa
Jens Christian Wahl
Mauricio Moreira Soares
Leiv Rønneberg
Annika Krutto
Ildiko Pilan
kjersti
Hanne Rognebakke
Geir Olve Storvik
ola.haug@nr.no
Ildiko Pilan
Chi Zhang
Lars Holden
Riccardo De Bin
Andrew Reiner
Chi Zhang
Ida Scheel
Ask to Unmute
Cathrine Brunborg
Øystein Skaug
Helge Jenssen
Cathrine Brunborg
Zhi Zhao
Nick Walker
Håkon Atle Jakobsen
Aliaksandr Hubin
Alise Danielle Midtfjord
Jaroslaw Nowak
Clara Bertinelli Salucci
Annabelle Redelmeier
Gunnhildur H. Steinbakk
Security
112
Participants
Chat
Share Screen
Record
Breakout Rooms
Reactions

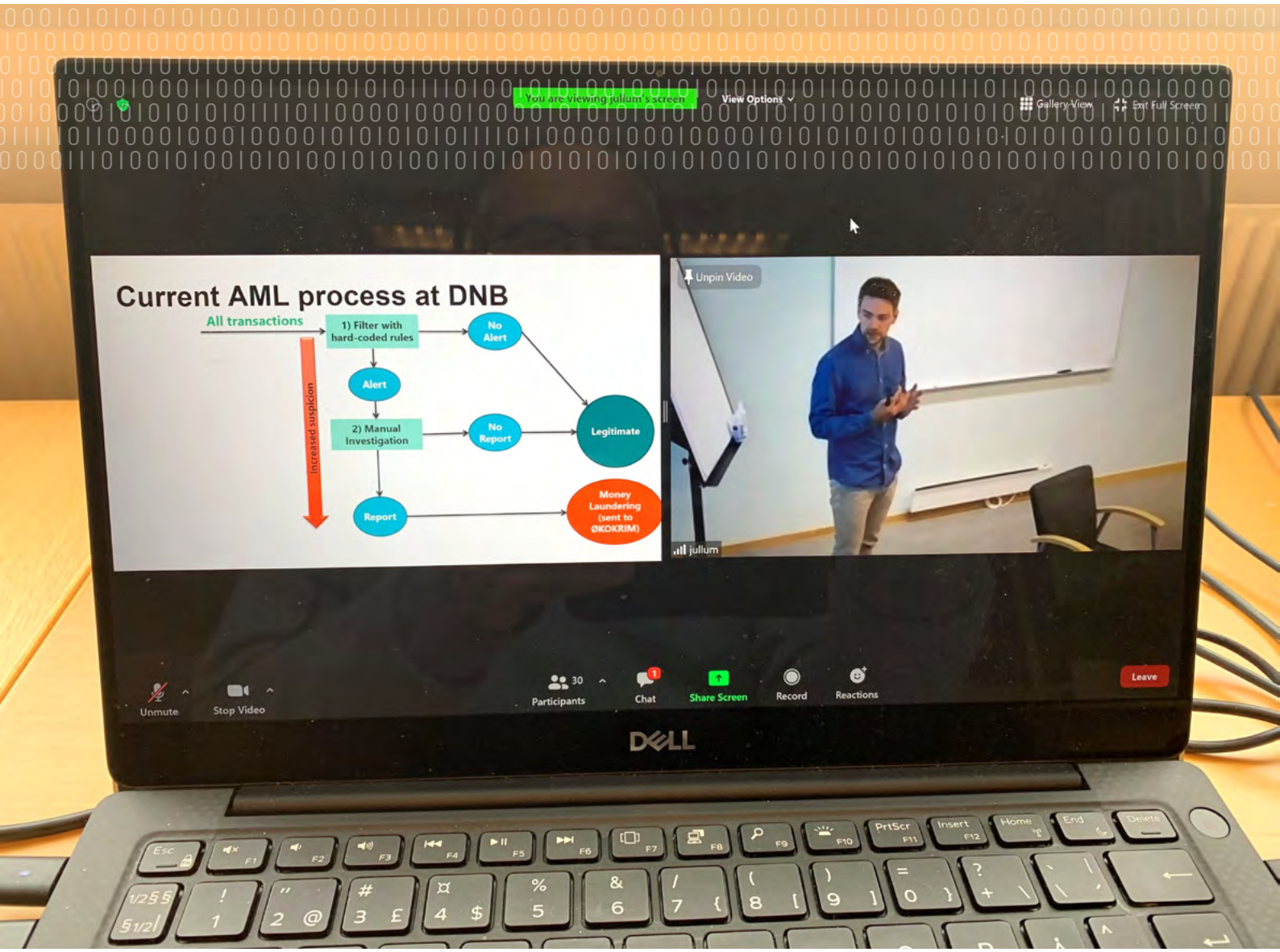
Seminars

BigInsight's biweekly Wednesday lunch takes place at the Department of Mathematics and NR alternatingly. In 2020 13 lunches were organized, see our webside for a list of invited speakers. Due to covid-19 the lunches were held via Zoom from mid March. 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. Also, these seminars have been via ZOOM, except for a few in summer, which were both in person and via ZOOM. The mixed audience worked fine.

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 but due to covid-19 most of the seminars were virtual. The seminars are usually 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.

Usually, our seminars have seen a large participation, so that we can proudly say that these Oslo seminars are among the best attended statistics seminar in Europe.



During and just after the summer 2020, meetings in person were allowed, with 1 meter distance. We were happy to see each other. Otherwise, most seminars and meetings happened on ZOOM.

TRAINING AND COURSES

The University of Oslo established a new Master Program in Data Science in 2018, and the first batch of Data Science masters finished their degree in the summer of 2020. We are proud that two of them have moved on to PhD work in BigInsight in the autumn of 2020. All other started exciting jobs in the industry. Admission to this Data Science 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 per year (500 in 2020), but only around 15-20 of these have been admitted each year. BigInsight participates to the master program by teaching, master projects and industrial contacts. We also contribute to the UiO Honours-programmet (bachelor) with some teaching.

BigInsight staff supervise MSc projects in data science and 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.

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.

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 coauthors 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. Co-supervision of BigInsight master students together with university staff is also excellent training for young researchers at NR.

Due to the pandemic situation in 2020, important arrangements for the student society such as Data Science Day at UiO, BigInsight Day and the yearly Klækken PhD workshop, usually sponsored and fully or partially arranged by BigInsight, have been cancelled. Instead, various series of digital seminars and meetings, journal clubs etc. have been tried out to keep in contact and scientifically updated. PhD students and postdocs have been prioritized in periods when access to the university campus has been limited.

«Noreg er blant dei leiande landa i verda når det gjeld grunnleggande digital kompetanse i befolkninga, mens vi heng litt etter når det gjeld tilgang på IT-spesialistar. Dei siste åra har vi likevel sett ein positiv trend: Sidan 2015 har regjeringa særleg prioritert IKT-relaterte utdanningar. Dette har gjort at nesten 1600 fleire studentar kan starte på IKT-studium kvart år. Likevel er det framleis stor etterspørsel etter IT-spesialistar. Gjennom Utdanningsløftet 2020 er det derfor etablert om lag 1500 nye studieplassar innan matematisk-naturvitskaplege fag med vekt på informatikk, og teknologiske fag med vekt på IKT."

Vår nye digitale kvardag,
Kommunal- og moderniseringsdepartementet,
januar 2021.

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*.

BigInsight in the media (selection)

Adventspraten, [Internet] 11.12.2020: **Solveig Engebretsen regner ut R-tallet.**
Solveig Engebretsen

Forskning.no, 11.12.2020: **Vi må snakke om algoritmeimport.** Anders Løland

Forskning.no, 27.11.2020: **For enkelt om kunstig intelligens: – Diskriminerende og fordomsfull AI er ikke alltid lett å løse.** Riegler, Michael; Lison, Pierre; Strümke, Inga; Løland, Anders.

Intervju NRK, [TV] 16.11.2020: **Solveig Engebretsen og Kenth Engø-Monsen**

NORDE (podcast) Internet, 15.10.2020: **Forklarbar kunstig intelligens.** Anders Løland

Aftenposten, 07.09.2020, pp. 7: **Nye FHI-anslag: Lang flere smittede i Norge enn antatt**
Intervju med Arnoldo Frigessi

TV2 Nyhetskanalen, 29.06.2020: **Intervju med Solveig Engebretsen**

Apollon 2/2020, pp 5: **Kartlegg koronaspreiing med mobiltelefonar**

Dagens Næringsliv, Oslo. 18.06.2020, pp 30: **Algoritmer er mer enn bare data – de kan diskriminere og gjøre urett.** Løland, Anders; Strümke, Inga.

Sciencenorway.no, 27.05.2020: **Data on how Norwegians move around allow for a finetuned model of calculating the spread of coronavirus.** Solveig Engebretsen

Forskning.no, 25.05.2020: **Smittetallet R: Slik regner Folkehelseinstituttet ut hvordan korona-viruset sprer seg.** Solveig Engebretsen

Tekna magasinet, 21.05.2020: **Årets ledestjerne?** Solveig Engebretsen

Intervju NRK, 18.05.2020: **Nye FHI-beregninger: Viruset er mindre utbredt, men mer dødelig.** Birgitte Freiesleben De Blasio og Peter Svaar

D2, 08.05.2020: **Koronajegeren.** Solveig Engebretsen

Dagens Nærlingsliv, Oslo. 07.05.2020: **Vi får nok ikke selvkjørende lastebiler riktig ennå.** Anders Løland

Intervju 05.05.2020; De Blasio, Birgitte Freiesleben; Christian, Birk. **Mens danske myndigheder mørklægger centrale beregninger om coronavirussen, er stilen en helt anden i Norge og Sverige: "Det mener jeg, er den rigtige måde at gøre det på".**

Bergens Tidende, 03.04.2020: **Takk for at vi får dele data – la oss få dele enda mer!** Anders Løland

UiO, Det matematisk-naturvitenskapelige fakultet, 02.04.2020: **Jakter viruset med mobildata.** Arnoldo Frigessi og Solveig Engebretsen

Digi.no, 24.03.2020: **Dokumentarfilmen iHuman – forvirrende, men nyttig?** Anders Løland

NRK-nyheter intervju 22.03.2020: **Mobilsporingen de håper skal hjelpe mot koronaviruset.** De Blasio, Birgitte Freiesleben.

Morgenbladet, 18.03.2020: **Frykter politikerne ikke holder tritt med teknologisk utvikling.** Anders Løland

Morgenbladet, 13.03.2020: **Jakter viruset med mobildata.** Intervju med Arnoldo Frigessi og Solveig Engebretsen

Dagens Næringsliv, Oslo. 10.01.2020, pp. 35: **Blåøyd algoritmeoptimisme?** Anders Løland

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.

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, Explaining AI
Stian Braastad	ABB	Sensor
Børre Gundersen	ABB	Sensor
Petter Häusler	ABB	Sensor
Jaroslaw Nowak	ABB	Sensor
Morten Stakkeland	ABB	Sensor
Stian Torkildsen	ABB	Sensor
Andree Underthus	ABB	Sensor
Frank Wendt	ABB	Sensor
Bjørn Møller	CRN	Health
Jan Nygård	CRN	Health
Lars Erik Bolstad	DNB	Fraud
Karl Aksel Festø	DNB	Marketing
Andreas Bendixen Hovdenes	DNB	Marketing
Fredrik Johannessen	DNB	Marketing
Tobias Lillekvelland	DNB	Marketing
Marcus Nilsson	DNB	Marketing
Roger Olafsen	DNB	Marketing
Hodjat Rahmati	DNB	Marketing
Nafiseh Shabib	DNB	Marketing
Aiko Yamashita	DNB	Marketing, Explaining AI
Lars Holterud Aarsnes	DNV-GL	Sensor
Øystein Alnes	DNV-GL	Sensor
Ole Christian Astrup	DNV-GL	Sensor
Håvard Nordtveit Austefjord	DNV-GL	Sensor
Andreas Brandsæter	DNV-GL	Sensor
Christos Chryssakis	DNV-GL	Sensor
Øystein Engelhardtsen	DNV-GL	Sensor
Ørjan Fredriksen	DNV-GL	Sensor
Tom Arne Pedersen	DNV-GL	Sensor
Gaute Storhaug	DNV-GL	Sensor
Hans Anton Tvete	DNV-GL	Sensor
Bjørn-Johan Vartdal	DNV-GL	Sensor
Erik Vanem	DNV-GL	Sensor, Power
Nikos Violaris	DNV-GL	Sensor
Sindre Froyn	Gjensidige	Marketing

NAME	INSTITUTION	MAIN RESEARCH AREA
Mikkel Hinnerichsen	Gjensidige	Marketing
Anders Nyberg	Gjensidige	Marketing
Daniel Piacek	Gjensidige	Marketing
Gunnhildur Steinbakk	Gjensidige	Fraud
Geir Thomassen	Gjensidige	Fraud
Stefan Erath	Hydro	Power
Plamen Mavrodiev	Hydro	Power
Ellen Paaske	Hydro	Power
Knut-Harald Bakke	Hydro	Power
Peter Szederjesi	Hydro	Power
Birgitte De Blasio	NIPH	Health
Christopher S Nielsen	NIPH	Health
Ulf Andersen	NAV	Fraud
Robindra Prabhu	NAV	Explaining AI
Cathrine Pihl Lyngstad	NAV	Explaining AI
Kjersti Aas	NR	Marketing, Explaining AI
Magne Aldrin	NR	Sensor
Solveig Engebretsen	NR	Health
Clara-Cecilie Günther	NR	Marketing, Health
Ola Haug	NR	Marketing, Sensor
Kristoffer Herland Hellton	NR	Marketing, Sensor
Lars Holden	NR	Health, Fraud
Marit Holden	NR	Health
Ragnar Bang Huseby	NR	Fraud, Power
Martin Jullum	NR	Marketing, Fraud, Explaining AI
Alex Lenkoski	NR	Power
Pierre Lison	NR	Fraud
Anders Løland	NR	Fraud, Power, Explaining AI
Linda R. Neef	NR	Fraud, Marketing
Ildikó Pilán	NR	Fraud
Annabelle Redelmeier	NR	Marketing, Explaining AI
Hanne Rognebakke	NR	Marketing, Sensor
André Teigland	NR	Explaining AI
Ingunn Fride Tvete	NR	Health
Jens Christian Wahl	NR	Marketing, Power
Tor Arne Øigård	NR	Power
Mette Langaas	NR/NTNU	Sensor, Health
Tero Aittokallio	OUS	Health
Pilar A. Duran	OUS	Health
Torsten Eken	OUS	Health
Jorrit Enserink	OUS	Health
Harald Fekjær	OUS	Health
Thomas Fleischer	OUS	Health
Maria Serena Giliberto	OUS	Health
Lars Åke Hall	OUS	Health
Robert Hanes	OUS	Health
Eivind Hovig	OUS	Health
Irena Jakopanec	OUS	Health
Vessela Kristensen	OUS	Health
Marissa LeBlanc	OUS	Health
Tonje Lien	OUS	Health

NAME	INSTITUTION	MAIN RESEARCH AREA
Egil Lingaas	OUS	Health
Sygve Nakken	OUS	Health
Andrew Reiner	OUS	Marketing, Health
Fredrik Schjesvold	OUS	Health
Therese Seierstad	OUS	Health
Kjetil Sunde	OUS	Health
Therese Sørlie	OUS	Health
David Swanson	OUS	Health
Dagim S. Tadele	OUS	Health
Kjetil Tasken	OUS	Health
Anders Berset	Skatteetaten	Marketing, Fraud
Wenche Celiussen	Skatteetaten	Marketing
Øystein Olsen	Skatteetaten	Marketing
Nils Gaute Voll	Skatteetaten	Marketing, Fraud
Kim Benjamin Boué	SSB	Marketing, Sensor, Power
Xeni Dimakos	SSB	Marketing
Boriska Toth	SSB	Marketing
Øyvind Langsrud	SSB	Sensor, Power
Li-Chun Zhang	SSB	Marketing, Sensor, Power
Geoffrey Canright	Telenor	Health, Sensor
Kenth Engo-Monsen	Telenor	Marketing, Health, Sensor
Jørgen Eriksson Midtbø	Telenor	Health
Kristian Lindalen Stenerud	Telenor	Health
Weiqing Zhang	Telenor	Health
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	Sensor
Carlo Mannino	UiO	Power
Kjetil Røysland	UiO	Health
Sven Ove Samuelson	UiO	Marketing
Geir Kjetil Sandve	UiO	Health
Ida Scheel	UiO	Marketing
Geir Storvik	UiO	Sensor
Øystein Sørensen	UiO	Health
Magne Thoresen	UiO	Health
Marit Veierød	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
Haakon C. Bakka		Norway	2020-2023	M	Fraud
Annika Krutto		Estonia	2020-2023	F	Health
Postdoctoral researchers in BigInsight with financial support from other sources					
Alvaro Köhn Luque	UiO	Spain	2016-2021	M	Health
Richard Xiaoran Lai	UiO	UK	2019-2022	M	Health
Henry Pesonen	UiO	Finland	2019-2022	M	Health
Vandana Ravindran	UiO/OUS	India	2020-2023	F	Health
Leonardo Santana	UiO	Brasil	2020-2023	M	Health
Mauricio M. Soares	UiO	Brasil	2020-2023	M	Health
George Zhi Zhao	OUS	China	2021-2023	M	Health
PhD students with financial support from BigInsight					
Simon Boge Brant		Norway	2018-2021	M	Fraud
Emanuele Gramuglia		Italy	2016-2021	M	Sensor
Even Moa Myklebust		Norway	2020-2023	M	Health
Riccardo Parviero		Italy	2018-2021	M	Marketing
Brittany Rose		USA	2018-2021	F	Health
Leiv Tore Salte Rønneberg		Norway	2018-2021	M	Health
Clara Bertinelli Salucci		Italy	2019-2022	F	Sensor
Jonas Fredrik Schenkel		Norway	2018-2021	M	SSB, Sensor
Martin Tveten		Norway	2017-2020	M	Sensor
Fredrik Wollbraaten		Norway	2020-2023	M	Sensor
Chi Zhang		China	2016-2020	F	Health
George Zhi Zhao		China	2019-2020	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
Haifeng Xu	OUS/UiO	China	2019-2023	M	Health
Håvard Kvamme	UiO	Norway	2015-2020	M	Marketing
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
Lars H.N. Olsen	MatNat/UiO	Norway	2020-2024	M	Explaining AI
Magnus Nygård Osnes	NIPH/UiO	Norway	2019-2023	M	Health
Anja Stein	STORi, Lancaster	Norway	2019-2023	F	Marketing
Emilie Ødegård	UiO	Norway	2019-2023	F	Health, Marketing
Master degrees					
Håkon Bliksås Carlsen			2020-2022	M	Fraud
Bob Betuin Fjellheim			2019-2021	M	Marketing
Christian Grindheim			2020-2022	M	Sensor
He Gu			2019-2021	M	Sensor
Nicola Kaletka			2019-2021	F	Health
Anna Keivalova			2020-2022	F	Sensor
Nicolay Kristensen			2019-2021	M	Sensor
Vera Haugen Kvisgaard			2019-2021	F	Fraud
Øystein Skauli			2019-2021	M	Marketing

FINANCIAL OVERVIEW

FUNDING	1000 NOK
The Research Council	14 618
Norwegian Computing Center (NR)	1021
Research Partners*, in kind	9485
Research Partners*, in cash	1452
Enterprise partners**, in kind	4434
Enterprise partners**, in cash	4800
Public partners***, in kind	4895
Public partners***, in cash	2806
Sum	**43513**

COSTS	
NR, research	11446
NR, direct costs	204
Research Partners*, research	22274
Enterprise partners**, research	4434
Public partners***, research	5154
Sum	**43513**

*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 2020

Journal and peer-reviewed conference papers

Abdelmalek, Samir; Dali, Ali; Bakdi, Azzeddine; Bettayeb, Maamar. **Design and experimental implementation of a new robust observer-based nonlinear controller for DC-DC buck converters.** Energy (ISSN 0360-5442). 213 doi: 10.1016/j.energy.2020.118816. 2020.

Abdelmalek, Samir; Dali, Ali; Bettayeb, Maamar; Bakdi, Azzeddine. **A new effective robust nonlinear controller based on PSO for interleaved DC–DC boost converters for fuel cell voltage regulation.** Soft Computing - A Fusion of Foundations, Methodologies and Applications (ISSN 1432-7643). 24 pp 17051-17064. doi: 10.1007/s00500-020-04996-4. 2020.

Bakdi, Azzeddine; Bounoua, Wahiba; Guichi, Amar; Mekhilef, Saad. **Real-time fault detection in PV systems under MPPT using PMU and high-frequency multi-sensor data through online PCA-KDE-based multivariate KL divergence.** International Journal of Electrical Power & Energy Systems (ISSN 0142-0615). doi: 10.1016/j.ijepes.2020.106457. 2020.

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 (ISSN 2077-1312). 8(1) doi: 10.3390/jmse8010005. 2020.

Belhechmi, S., De Bin, R., Rotolo, F., & Michiels, S. (2020). **Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models.** BMC bioinformatics, 21(1), 1-20.

Bergholtz, Helga; Lien, Tonje Gulbrandsen; Swanson, David; Frigessi, Arnoldo; Daidone, Maria Grazia; Tost, Jörg; Wärnberg, Fredrik; Sørlie, Therese. **Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions.** NPJ breast cancer (ISSN 2374-4677). 6 doi: 10.1038/s41523-020-0167-x. 2020.

Bounoua, Wahiba; Bakdi, Azzeddine. **Fault detection and diagnosis of nonlinear dynamical processes through correlation dimension and fractal analysis based dynamic kernel PCA.** Chemical Engineering Science (CES) (ISSN 0009-2509). doi: 10.1016/j.ces.2020.116099. 2020.

Engebretsen, Solveig; Engø-Monsen, Kenth; Aleem, Mohammad Abdul; Gurley, Emily Suzanne; Frigessi, Arnoldo; de Blasio, Birgitte Freiesleben. **Time-aggregated mobile phone mobility data are sufficient for modelling influenza spread: the case of Bangladesh.** Journal of the Royal Society Interface (ISSN 1742-5689). 17(167) doi: 10.1098/rsif.2019.0809. 2020.

Engebretsen, Solveig; Glad, Ingrid Kristine. **Partially linear monotone methods with automatic variable selection and monotonicity direction discovery.** Statistics in Medicine (ISSN 0277-6715). 39(25) pp 3549-3568. doi: 10.1002/sim.8680. 2020.

Fanaee-T, Hadi, and Magne Thoresen. **Iterative Multi-mode Discretization: Applications to Co-clustering.** International Conference on Discovery Science. Springer, Cham, 2020.

Halkola, A. S., Parvinen, K., Kasanen, H., Mustjoki, S., & Aittokallio, T. (2020). **Modelling of killer T-cell and cancer cell subpopulation dynamics under immuno-and chemotherapies.** Journal of theoretical biology, 488, 110136.

Hemerik, Jesse, Magne Thoresen, and Livio Finos. **Permutation testing in high-dimensional linear models: an empirical investigation.** Journal of Statistical Computation and Simulation (2020): 1-18.

Huang, Yeran; Mannino, Carlo; Yang, Lixing; Tang, Tao. **Coupling time-indexed and big-M formulations for real-time train scheduling during metro service disruptions.** Transportation Research Part B: Methodological (ISSN 0191-2615). 133 pp 38-61. doi: 10.1016/j.trb.2019.12.005. 2020.

Hubin, A., Storvik, G., & Frommlet, F. (2020). **A Novel Algorithmic Approach to Bayesian Logic Regression (with Discussion).** Bayesian Analysis, 15(1), 263-333.

Ianevski, A., Giri, A. K., & Aittokallio, T. (2020). **SynergyFinder 2.0: visual analytics of multi-drug combination synergies.** Nucleic acids research, 48(W1), W488-W493.

Jullum, Martin; Løland, Anders; Huseby, Ragnar Bang; Ånonsen, Geir; Lorentzen, Johannes P. **Detecting money laundering transactions with machine learning.** Journal of Money Laundering Control (ISSN 1368-5201). 23(1) pp 173-186. doi: 10.1108/JMLC-07-2019-0055. 2020.

Ledo, Alicia; Smith, Pete; Zerihun, Ayalsew; Whitaker, Jeanette; Vicente-Vicente, José Luis; Qin, Zhangcai; McNamara, Niall P.; Zinn, Yuri L.; Llorente, Mireia; Liebig, Mark; Kuhnert, Matthias; Dondini, Marta; Don, Axel; Diaz-Pines, Eugenio; Datta, Ashim; Bakka, Haakon C.; Aguilera, Eduardo; Hillier, Jon. **Changes in soil organic carbon under perennial crops.** Global Change Biology (ISSN 1354-1013). doi: 10.1111/gcb.15120. 2020.

Lingjærde, Camilla, Lien, Tonje, Borgan, Ørnulf & Glad, Ingrid K. (2020). **Tailored Graphical Lasso for Data Integration in Gene Network Reconstruction.** To appear BMC Bioinformatics.

Mancisidor, R. A., Kampffmeyer, M., Aas, K., & Jenssen, R. (2020). **Deep generative models for reject inference in credit scoring.** Knowledge-Based Systems, 196, 105758.

Mannino, Carlo; Huang, Yeran; Yang, Lixing; Tang, Tao. **Coupling time-indexed and big-M formulations for real-time train scheduling during metro service disruptions.** Transportation Research Part B: Methodological (ISSN 0191-2615). doi: 10.1016/j.trb.2019.12.005. 2020.

Máté E. Maros, David Capper, David T. W. Jones, Volker Hovestadt, Andreas von Deimling, Stefan M. Pfister, Axel Benner, Manuela Zucknick & Martin Sill. **Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data.** Nature Protocols, 1-34.

Otneim, Håkon; Jullum, Martin; Tjøstheim, Dag Bjarne. **Pairwise local Fisher and naive Bayes: Improving two standard discriminants.** Journal of Econometrics (ISSN 0304-4076). 216(1) pp 284-304. doi: 10.1016/j.jeconom.2020.01.019. 2020.

Page, C. M., Djordjilović, V., Nøst, T. H., Ghiasvand, R., Sandanger, T. M., Frigessi, A., Thoresen M. & Veierød, M. B. (2020). **Lifetime ultraviolet radiation exposure and DNA methylation in blood leukocytes: The Norwegian Women and Cancer Study.** Scientific reports, 10(1), 1-8.

Patone, Martina; Zhang, Li Chun. **On Two Existing Approaches to Statistical Analysis of Social Media Data.** International Statistical Review (ISSN 0306-7734). doi: 10.1111/insr.12404. 2020.

Pensar, J., Talvitie, T., Hyttinen, A., & Koivisto, M. (2020, April). **A Bayesian approach for estimating causal effects from observational data. In Proceedings of the AAAI Conference on Artificial Intelligence** (Vol. 34, No. 04, pp. 5395-5402).

Pensar, J., Xu, Y., Puranen, S., Pesonen, M., Kabashima, Y., & Corander, J. (2020). **High-dimensional structure learning of binary pairwise Markov networks: a comparative numerical study.** Computational Statistics & Data Analysis, 141, 62-76.

Pladsen AV, Gro Nilsen, Oscar M. Rueda, Miriam R. Aure, Ørnulf Borgan, Knut Liestøl, Valeria Vitelli, Arnoldo Frigessi, Anita Langerød, Anthony Mathelier, OSBREAC, Olav Engebråten, Vessela Kristensen, David C. Wedge, Peter Van Loo, Carlos Caldas, Anne-Lise Børresen-Dale, Hege G. Russnes & Ole Christian Lingjærde. **DNA copy number motifs are strong and independent predictors of survival in breast cancer.** Communications biology 3.1 (2020): 1-9.

Pulkkinen, O. I., Gautam, P., Mustonen, V., & Aittokallio, T. (2020). **Multiobjective optimization identifies cancer-selective combination therapies.** PLOS Computational Biology, 16(12), e1008538.

Redelmeier, Annabelle Alice; Jullum, Martin; Aas, Kjersti. **Explaining Predictive Models with Mixed Features Using Shapley Values and Conditional Inference Trees.** In: Lecture Notes in Computer Science (LNCS). (ISBN 978-3-030-58805-2). pp 117-137. doi: 10.1007/978-3-030-57321-8_7. 2020.

Sanguiao-Sande, Luis; Zhang, Li Chun. **Design-Unbiased Statistical Learning in Survey Sampling.** Sankhya A, The Indian Journal of Statistics (ISSN 0976-836X). doi: 10.1007/s13171-020-00224-1. 2020.

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), 2020.

Steens, Anneke; De Blasio, Birgitte Freiesleben; Veneti, Lamprini; Gimma, Amy; Edmunds, W. John; Van Zandvoort, Kevin; Jarvis, Christopher I.; Forland, Frode; Robberstad, Bjarne. **Poor self-reported adherence to COVID-19-related quarantine/isolation requests, Norway, April to July 2020.** Eurosurveillance (ISSN 1025-496X). 25(37) pp 1-6. doi: 10.2807/1560-7917.ES.2020.25.37.2001607. 2020. Institutional archive

Sørensen, Ø., Crispino, M., Liu, Q., & Vitelli, V. (2020). **BayesMallows: An R Package for the Bayesian Mallows Model.** The R Journal, 12(1), 324-342.

Zhang, Li-Chun. **Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics.** Journal of the Royal Statistical Society: Series A (Statistics in Society) (ISSN 0964-1998). doi: 10.1111/rssa.12632. 2020.

Zhao, Zhi; Zucknick, Manuela. **Structured penalized regression for drug sensitivity prediction.** The Journal of the Royal Statistical Society, Series C (Applied Statistics) (ISSN 0035-9254). 69(3) pp 525-545. doi: 10.1111/rssc.12400. 2020. Institutional archive

Reports and submitted papers

Asgari, F., Alamatsaz, M. H., Vitelli, V., & Hayati, S. (2020). **Latent function-on-scalar regression models for observed sequences of binary data: a restricted likelihood approach.** arXiv preprint arXiv:2012.02635.

Banterle, M., Zhao, Z., Bottolo, L., Richardson, S., Leoncio, W., Lewin, A., & Zucknick, M. (2020). **Package 'BayesSUR'**. On CRAN

Brandsæter, Andreas, and Ingrid K. Glad. **Explainable Artificial Intelligence: How Subsets of the Training Data Affect a Prediction.** arXiv preprint arXiv:2012.03625 (2020).

De Blasio, Birgitte Freiesleben; Di Ruscio, Francesco; Rø, Gunnar; Engebretsen, Solveig; Diz-Lois Palomares, Alfonso; Kristoffersen, Anja Braathen; Engø-Monsen, Kenth; Lindstrøm, Jonas; Zhang, Chi; Frigessi, Arnoldo. **Covid-19 situational awareness and forecasting for Norway, weekly report, Norwegian Institute of Public Health,** https://www.fhi.no/sv/smittsomme-sykdommer/corona/koronavirus-modellering/ (2020)

Djordjilović, Vera; Hemeri, Jesse; Thoresen, Magne. **On optimal two-stage testing of multiple mediators.** arXiv preprint arXiv:2007.02844 (2020).

Ghannoum, S., Antos, K., Netto, W. L., Kohn-Luque, A., & Farhan, H. (2020). **CellMAPtracer: A user-friendly tracking tool for long-term migratory and proliferating cells.** bioRxiv.

Ghosh, Abhik; Thoresen, Magne. **Robust Sure Independence Screening for Non-polynomial dimensional Generalized Linear Models.** arXiv preprint arXiv:2005.12068 (2020).

Ghosh, Abhik; Thoresen, Magne. **A robust variable screening procedure for ultra-high dimensional data.** arXiv preprint arXiv:2004.14851 (2020).

Hubin, A., Storvik, G. O., Grini, P. E., & Butenko, M. A. (2020). **A Bayesian binomial regression model with latent Gaussian processes for modelling DNA methylation.** arXiv preprint arXiv:2004.13689.

Hubin, Aliaksandr, Storvik, Geir; Frommlet, Florian. **Flexible Bayesian Nonlinear Model Configuration.** arXiv preprint arXiv:2003.02929 (2020).

Madjar, K., Zucknick, M., Ickstadt, K., & Rahnenführer, J. (2020). **Combining heterogeneous subgroups with graph-structured variable selection priors for Cox regression.** arXiv preprint arXiv:2004.07542.

Nowak, Jaroslaw. **Machine learning: believe it or not?** https://new.abb.com/news/detail/67724/machine-learning-believe-it-or-not, ABB White paper, 2020.

Ponzi, E., Thoresen, M., Nøst, T. H., & Møllersen, K. (2020). **Integrative analyses of multi-omics data improves model predictions: an application to lung cancer.** bioRxiv.

Saarela, O., Rohrbeck, C., Arjas, E. **Non-parametric ordinal regression under a monotonicity constraint.** https://arxiv.org/pdf/2007.01390.pdf. 2020

Steinbakk, Gunnhildur Högnadóttir; Langsrud, Øyvind; Løland, Anders. **A brief overview of methods for synthetic data for official statistics.** Norsk Regnesentral, . NR-notat SAMBA/23/20. pp 22. 2020.

Suotsalo, K., Xu, Y., Corander, J., & Pensar, J. (2020). **High-dimensional structure learning of sparse vector autoregressive models using fractional marginal pseudo-likelihood.** arXiv preprint arXiv:2011.01484.

Tveten, Martin & Glad, Ingrid K. (2020). **Online Detection of Sparse Changes in High-Dimensional Data Streams Using Tailored Projections.** arXiv:1908.02029.

Viinikka, J., Hyttinen, A., Pensar, J., & Koivisto, M. (2020). **Towards Scalable Bayesian Learning of Causal DAGs.** arXiv preprint arXiv:2010.00684.

Wahl, Jens Christian; Aanes, Fredrik L; Aas, Kjersti. **Spatial modelling of risk premiums for water damage insurance.** Norsk Regnesentral, NR-notat SAMBA/33/20. pp 33. 2020.

Open source published software

BayesMallows: Bayesian Preference Learning with the Mallows Rank Model [CRAN] [GitHub] [R Journal]

bayesynergy: An R package for Bayesian semi-parametric modelling of in-vitro drug combination experiments [GitHub]

DIscBIO: A user-friendly R pipeline for biomarker discovery in single-cell transcriptomics [GitHub]

hdme: High-Dimensional Regression with Measurement Error [CRAN] [GitHub] [Journal of Open Source Software]

kdensity: An R package for kernel density estimation with parametric starts and asymmetric [CRAN] [GitHub] [Journal of Open Source Software]

pycox: Survival analysis with PyTorch [GitHub] [PyPI]

shapr: Explaining the output of machine learning models with more accurately estimated Shapley values [CRAN] [GitHub] [Journal of Open Source Software]

spread: An R package that contains different infectious disease spread models [CRAN] [GitHub]

tpca: automatically selecting the principal components most sensitive to changes [GitHub]

tpcaMonitoring: performing TPCA change detection [GitHub]

twl: Two-Way Latent Structure Clustering Model [CRAN]

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

Illustrasjoner: Ellen Hegtun. Design og lay-out: designavdelingenkis.no
Ved å bruke designavdelingenkis støtter man Kunst i Skolens arbeid med kunstformidling og kreativitetstrening for barn og unge. Tusen Takk!

BigInsight

$$\pi(\rho) = (n!)^{-1} 1_{\mathcal{P}_n}(\rho)$$

$$\pi(\alpha|\lambda) = \lambda e^{-\lambda\alpha} 1_{[0,\infty)}(\alpha)$$

$$\lambda = 0.1 \text{ or } \lambda = 0.05$$

$$P(\rho, \alpha | \mathbf{R}_1, \ldots, \mathbf{R}_N) \propto \frac{\pi(\rho)\pi(\alpha)}{Z_n(\alpha)^N} \exp\left\{ -\frac{\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \rho) \right\}$$

$$P(\mathbf{R}_1, \ldots, \mathbf{R}_N | \alpha, \rho) \propto \frac{1}{Z_n(\alpha)^N} \exp\left\{ -\frac{\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \rho) \right\} \prod_{j=1}^{N} \{1_{\mathcal{P}_n}(\mathbf{R}_j)\}$$