

BigInsight

STATISTICS FOR THE KNOWLEDGE ECONOMY

ANNUAL REPORT 2021

```
collective_anom_list <- function(capacc_res) {  
  anom_list <- list()  
  m <- nrow(capacc_res$anom)  
  # Indices m correspond to m - 1 in the do.  
  while (m >= 1) {  
    if (capacc_res$anom[m, 2] == 0) {  
      # ...  
    } else if (capacc_res$anom[m, 2] == 1) {  
      end <- m - 1  
      start <- capacc_res$anom[m, 1]  
      J <- capacc_res$J[m]  
      means <- means(capacc_res$x[start:end, J], drop = FALSE)  
      anom_df <- data.frame('start' = rep(start, length(J)),  
                           'end' = rep(end, length(J)),  
                           'variate' = J,  
                           'mean_change' = means)  
      anom_list[[length(anom_list) + 1]] <- anom_df  
      m <- start  
    } else if (capacc_res$anom[m, 2] == 2) {  
      m <- capacc_res$anom[m, 1]  
    }  
  }  
}
```



sfi = Centre for
Research-based
Innovation

The Research Council of Norway

BigInsight

CONTENT

SUMMARY	3
VISION AND OBJECTIVES	4
PARTNERS	7
ORGANISATION	8
RESEARCH STRATEGY	10
METHODS	11
SCIENTIFIC HIGHLIGHTS	12
PERSONALISED MARKETING	14
PERSONALISED HEALTH AND PATIENT SAFETY	16
PERSONALISED FRAUD DETECTION	18
SENSOR SYSTEMS	20
FORECASTING POWER SYSTEMS	24
EXPLAINING AI	26
INTERNATIONAL COOPERATION	28
PHD GRADUATES IN 2021	33
ACTIVITIES AND EVENTS	38
TRAINING AND COURSES	42
COMMUNICATION AND DISSEMINATION ACTIVITIES	44
RECOGNITIONS	46
PERSONNEL	47
FINANCIAL OVERVIEW	52
PUBLICATIONS IN 2021	53



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 unique data resources and substantial scientific and domain specific 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 special situations. 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 warmly congratulate BigInsight for its impressive achievements during these years: the scale of its translational work, from the development of new methods to the implementation of results with the partners, is superb; the range of its activities, across diverse applied areas and industrial sectors, embracing many disparate scientific approaches and techniques, is outstanding.”

The BigInsight Scientific Advisory Committee February 2022

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 alone, 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, operational solutions. BigInsight's research converges on two central innovation themes:

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

Our solutions are courageous and creative, exploit knowledge and structure in complex data and integrate data 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: customers, users, patients, but also sensors, vessels, wind-turbines, etc. Beside their obvious differences, there are many common characteristics:

- the high number of units/individuals/sensors under consideration;

"What are the biggest challenges and opportunities facing statisticians? Big data, messy data, and complicated questions: when using data from more sources, it should be possible to make more finely-grained inferences and decisions."

Andrew Gelman and Aki Vehtari, April 2021



- in some cases, massive data for each unit; in other cases, more limited information per unit;
- complex dependence structure between units;
- new data types, new technologies, new regulations make their use innovative;
- in most cases, units have their own intelligence, their own strategies and are exposed to their specific environment.

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

Predicting transient phenomena

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. Our partners 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 health-care institutions allows completely new screening protocols and treatment monitoring, real time prevention and increased safety. 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 are exposed to external forces;
- 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;
- there is a complex dependence structure between time series;
- there can be unknown and complex causes of observed abnormal behavior;
- there is the 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 (NR) (host institute)
- University of Oslo (UiO)
- University of Bergen (UiB)
- ABB
- DNB
- DNV
- 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 2021, where all partners are represented. In addition to close cooperation with the researchers at NR and the universities, there have been seminar series 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 in 2021.

Big Insight has launched an Associated Partners program for companies and organisations who would like to explore the possibilities by being connected to the centre. In 2021 CoopX joined this program.

The annual BigInsight Day was arranged as a digital meeting in February. The program included an interesting debate on how to communicate statistical uncertainty, a central topic during the last years as BigInsight has been deeply involved in the Norwegian Covid-19-modeling under the lead of the Norwegian Institute of Public Health.



UiO : University of Oslo

UNIVERSITY OF BERGEN



ORGANISATION

Board in 2021

Karl Aksel Festø, DNB, chairman
 Stian Braastad, ABB
 Hans Anton Tvete, DNV
 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
 Alexander Bjerke, Skatteetaten
 Xeni Dimakos, SSB
 Kenth Engø-Monsen, Telenor
 Bård Støve, University of Bergen
 Nadia Slavila Larsen, University of Oslo

Observers: Siv Johansen Soriano / Terje Strand, Research Council of Norway

The board had 2 meetings in 2021.
 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 and OUS, 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 (chair)
 Prof. Samuel Kaski, Aalto University, Finland and University of Manchester, UK
 Prof. Geoff Nicholls, Univ. of Oxford, UK
 Prof. Marina Vannucci, Rice Univ., Houston, USA
 Prof. Veronica Vinciotti, Univ. of Trento, Italy





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 explainability 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 happens 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



Explaining AI

INNOVATION PARTNERS

DNB
Gjensidige
NAV
Skatteetaten
Telenor
SSB

DNV
Kreftregisteret
OUS
Telenor

DNB
Gjensidige
Skatteetaten

ABB
DNV
SSB

Hydro Energy

all partners

RESEARCH PARTNERS

NR
UiO
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

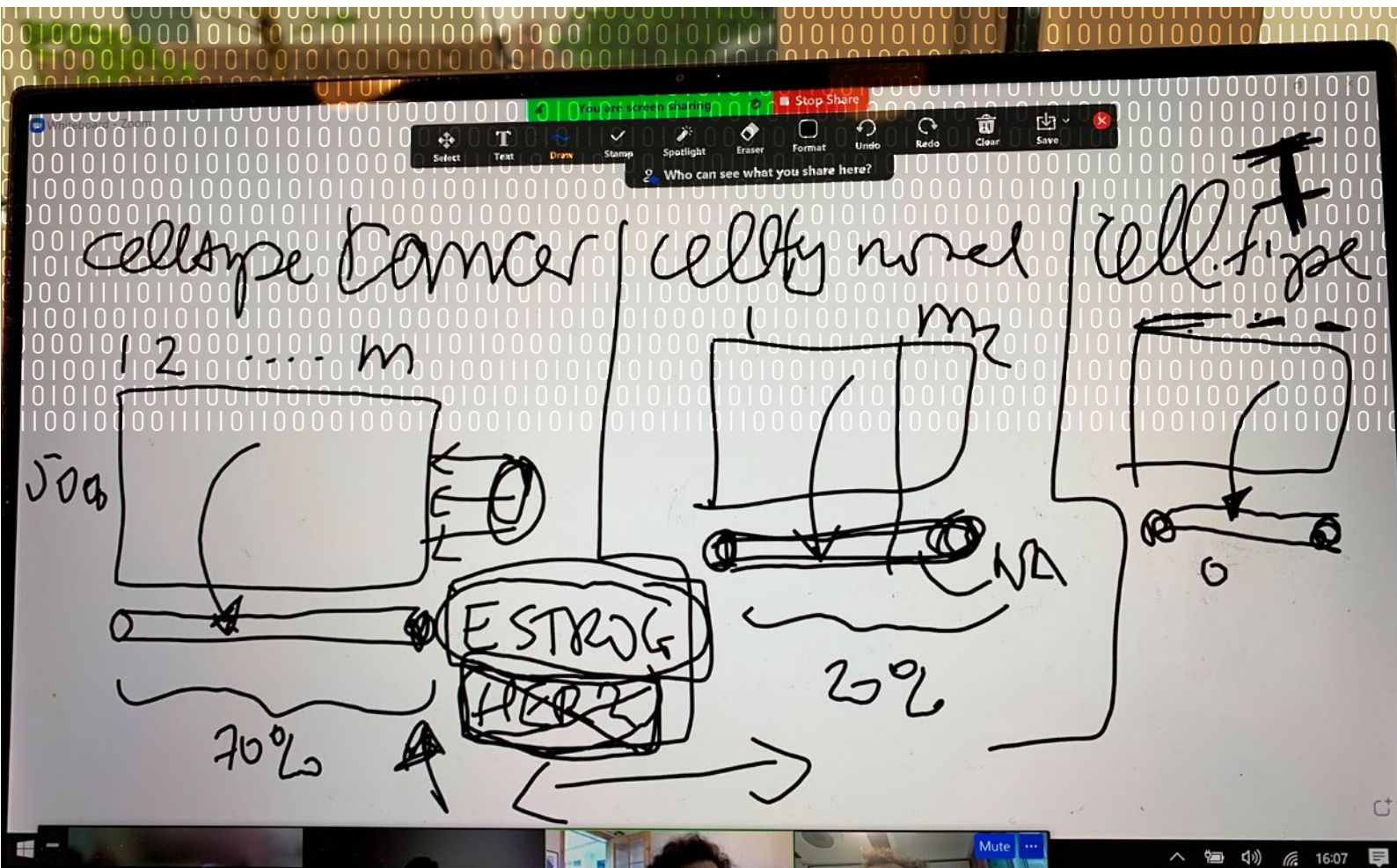
Alex Lenkoski
Carlo Mannino

Anders Løland
Arnoldo Frigessi

METHODS

We solve innovation challenges of our partners by developing new or applying state-of-the-art statistical, mathematical, and machine learning methods in these fields:

- Probabilistic Bayesian models and forecasting
- Complex dependence models
- Latent variable models
- Data integration and knowledge incorporation
- Knowledge based machine learning
- Multi-type and multi scale models
- Scalable approximation algorithms
- High dimensional data and time series
- Change and anomaly detection and prediction
- Local and global explanations of black box models
- Networks and graphical models
- Preference learning
- Mechanistic multi-type models in oncology
- Time-to-event models



SCIENTIFIC HIGHLIGHTS

Translational research develops, transforms, and converts generic methodological and theoretical results into outcomes that directly benefit society at large. Translational research is what we do at BigInsight. After six years of formidable work, we can start to collect our achievements, remarkable results that impact innovation processes at our partners. Without aiming to being complete, here are some of our highlights.

We have developed new neural network approaches for time-to-event data and proposed a way to predict mortgage defaults using only consumer transaction data, instead than traditional demography, ownerships, education and income. The method works, and DNB is evaluating it in practice, before deciding if it will be put in production. The methodological papers are published in top journals, Journal of Machine Learning Research and Expert Systems with Applications. The PhD student who was central in the project now works in an exciting investment firm.

Our work on preference learning and recommender systems has established Bayesian Mallows model as a new approach. Our R-package is used internationally, and we have been working on many improvements that make the method attractive for probabilistic recommendations. Two PhD students have written several papers, published in journals like Knowledge Based Systems and the Annual review of Statistics and its Applications. We developed Bayesian recurrent neural nets that act on time series of clicking data, and which scale to real world industrial settings. The method is published on Data Mining and Knowledge Discovery, a new open data set is launched at Recsys and a version of the method is currently in production at finn.no.

BigInsight has now a very strong research team in mathematical modelling and inference in oncology. Seven post-docs have started from Germany, Iran, Portugal, China, USA, Sweden, and Turkey, funded by our new European and Digital-Life-Norway projects. Focus is on personalised treatment in breast and blood cancers. For the first time, we have been able to model a number of cells that is 500 times larger than previously. We developed an algorithm, which we call DECIPHER, that allows us to estimate the number of subtypes of cancer cells that respond different to a drug, an important step towards precision medicine. We published one of the most flexible method and R-package (bayesynergy) to estimate the synergy of drugs used in combination in vitro, in a fully Bayesian setting. Thanks to our collaborations with OUS, there are many very exciting new results in the

pipeline, which will position us in the forefront internationally of mathematical oncology.

Of course we are very pleased of our Covid-19 research. What has been very special in an international context is the timeliness: we were among the first in the world able to estimate reproduction numbers and make predictions of hospital admissions at a regional level using Telenor's mobile phone data to capture real-time mobility of individuals. The reason is that BigInsight had worked on this theme for several years before March 2020, studying influenza epidemics in Bangladesh using mobile phone mobility data. In these years, we developed three major models (a stochastic compartmental metapopulation model with piecewise constant reproduction numbers; a stochastic Monte Carlo approach for daily changing reproduction numbers and an individual based model, a digital twin on covid-19 in Norway), that were used systematically, some weekly, to produce estimates, forecasts, scenarios for the Norwegian Institute of Public Health and the Norwegian Government. The Norwegian media have given a great attention to our work, and we have learned, also the hard way, how to communicate to the public. Our results, which are currently submitted for publications, have really helped decision makers, so that we can proudly affirm that it is also thanks to BigInsight that Norway succeeded to control the virus rather well.

Another area where we have obtained several very important results is sensor systems. Methodologies are well published, for example on scalable anomaly detection in cross-correlated sensor data on the Annals of Applied Statistics, and then tested on cases from ABB and DNV, on condition monitoring and motor overheating. Our work on building testbeds for navigation algorithms which drive autonomous ships, is unique. Until now tests were essentially hand-designed. These tests are few, not realistic and too easy or too complex. In addition, they cover only some situations. We have developed a method and an algorithm that produces realistic test cases, with any wished level of complexity, and in any number. We use the AIS international data base of all movements of



all ships in the world. Our method is able to detect situations in the AIS data which are almost collisions or almost groundings. We developed a scoring system which allows to classify situations in various ways of complexity and type, so that one can search for test-cases as wished. In this way we can test algorithms in realistic situations, something that DNV is planning to use in its certification systems. Again, all our results are well published, for example on IEEE transactions on intelligent transportation systems. In fact, a very important by-product of our work is the digitalisation, in form of algorithms, of the international regulations for preventing collisions at sea, from 1972, and the critical evaluation of their implementation in practice. This is useful in the international revision of these conventions.

Our work on power systems, in collaboration with Norsk Hydro, continues to produce methods and computational tools which are be at the forefront. Our probabilistic price forecasts quantify uncertainty. Based on quantile regression and copula models, our algorithms are in operation at Norsk Hydro. We have particular attention to renewable energies, for example predicting the lifetime production of wind parks.

Several of our results on fraud detection are under evaluation at our partners, Gjensidige, DNB, NAV and Skatteetaten. We have focused on exploiting network relations between individuals or businesses and on sentiment analyses of text exchanges during claim related conversations. Insurance claims are submitted in text and sound. We estimated the sentiments present in an application, to improve fraud detection. Our sentiment predictor uses Natural Language

Processing, together with machine learning and statistical methods. Importantly, fraud data are highly unbalanced and we studied suspicious cases in addition to actual fraud cases. In general modelling dependence has been very advantageous. Several papers are published or submitted and more are coming.

We are among the first to develop new methodology to explain AI black box models (like deep learning) when the variables have complex dependencies. Black boxes determine a rule that links such variables (for example the genes of a patient) to an outcome (say if the patient has or not a certain disease). These rules are difficult to interpret. Our new approach allows to perform such interpretation also when the variables are very dependent on each other, as very often is the case. We have applied our methods to explain NAV's algorithms used to estimate the length of sick leave. This project is part of Datatilsynet's "sandbox for AI", where our methods and codes are now evaluated in legal terms. We published excellent papers, which have attracted international attention, and are positioning ourselves among the world experts on Shapley values.

And, let's admit: working on these (and other) projects, in this way, with these collaborators at BigInsight has been fun! We are harvesting from the work done in these exciting years, methodologically and in terms of innovation and impact. It is already time to plan celebrations!

Center Leader and Director
Arnaldo Frigessi

PERSONALISED MARKETING



We develop new methods, strategies and algorithms for individualised marketing, customer retention, optimised communication with users, personalised pricing, and personalised recommendations or to influence the probability of specific actions of the users. We exploit users' behavioural measurements in addition to more traditional characteristics, and external data (including competitors' activity, market indicators, financial information, 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 2021:

Bayesian methodology for recommender systems

BigInsight has developed two new Bayesian approaches to recommendations. The first is based on the Bayesian Mallows Model for preference learning: we have shown that it performs similarly to the best state-of-the-art approaches, while giving a higher level of diversity. This means that users get more interesting and less obvious recommendations, and the catalogue of offered items is exploited more. An important remaining issue has been to make the methodology scalable, in order for it to be production ready. In 2021, we finalized our Variational Bayesian approximation algorithm, which speeds up convergence drastically. Together with STORi, we have also developed a dynamic approach, which allows to simply extend current learning when new users and new data arrive, instead than restarting the whole procedure. This is now implemented in our R-package BayesMallows. We have also developed a new method that combines item selection with the Bayes Mallows approach, because one can often assume that learning the ranking of all items appears unnecessary, when only the top and bottom for example are informative.

The second approach is a joint project with finn.no and STORi on dynamic slate recommendation with gated recurrent units and Thompson Sampling, where we have considered the problem of recommending relevant content to users of an internet platform in the form of lists of items, called slates. We have introduced a variational Bayesian Recurrent Neural Net recommender system that acts on time series of interactions between the internet platform and the user, and which scales to real world industrial situations. The recommender system is used online by finn.no, carrying a good proportion of the recommendations.

In 2021 we have also made a unique offline dataset from finn.no available for research. This is the first publicly available


datasets which includes all the slates that are presented to users, as well as which items (if any) in the slates were clicked on. Such a data set allows to move beyond the common assumption that implicitly assumes that users are considering all possible items at each interaction.

Stochastic models for early prediction of viral customer behavior on networks

BigInsight has developed methodology for early prediction of the adoption of new products with viral potential on social networks. Our method accomplishes this task after having drawn inference from observing the early history of adoptions of the product on the social network of the customer. Our stochastic model is agent-based at the individual level and allows for influence both from viral word-of-mouth effects and external factors, such as marketing campaigns. Inference is by maximum likelihood, and prediction is performed by simulation with a computationally very efficient algorithm. In 2021 we have submitted two papers on this idea, one describing the methodology, investigating the performance on simulated data and showing successful results for real data on a Telenor product, and the other investigating the effect of having a partially observed network. We have shown that if less than half of the links are missing, the results are still qualitatively reasonable. This is important because in practice the full social networks cannot be observed, and in particular, observing all social links is not possible.

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 logged 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



Sequential dataset logging interactions, all viewed items and clicks/no-click for recommender systems research

Table 1. Dataset statistics

Description	Value
Number of interactions	37.5m
Time period of data collection	30 days
Unique Items	1.3m
Unique item groups	290
Unique Users	2.3m
Number of interactions per user	16.4
Percentage of interactions with <i>no-click</i>	24.4%
Average slate length	10.1 items
Percentage of slates which produced clicks	69.7%
Percentage of slates which did not produce clicks	30.3%

machine learning algorithms. The developed method and code have been transferred to DNB and are currently under evaluation: depending on results our methods might be implemented in the app.

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. We have developed further the methodology called "Shapley values". In addition, we focused on counterfactual explanations. Counterfactual explanations try to explain complex models by producing examples with a different outcome than the original one. The motivation is that individuals who are unhappy with their outcome will be interested in what they can do differently to attain the desired outcome/class. We have developed a counterfactual method that takes the dependence between explanatory variables into account. A paper describing the method has been submitted to the FaccT22-conference and we will study the applicability of our approach in cases from our partners.

Risk impact of weather conditions on car crash counts

What is the effect of weather conditions on claim frequencies for motor vehicles? Gjensidige wants to understand whether any of the regional differences and/or yearly changes observed for claim frequencies may be attributed

to weather conditions. Traditional Generalized Additive Model worked very well for the claims on monthly resolution, and we did not need to develop any new method in this case. Preliminary results show that the geographical variations to some extents are influenced by different weather conditions, while this is not the case for the year-to-year variations. We showed that including metrological information leads to better predictions, both out-of-sample and out-of-time.

Automatic index generation

Many of Statistics Norway's traditional surveys are time consuming and labour intensive. There is therefore a need for more efficient, semi-automatic and continuously updated statistics utilizing detailed streams of data collected automatically, analysed with adequate machine learning algorithms. To enable this exciting development, we have developed new machine learning methods suited to handle and impute missing values when estimating nutritional values from consumption data. Our methods will be applied in the forthcoming 2022 Survey of consumer expenditure from Statistics Norway.



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 2021:

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. We have developed an R package for Bayesian seemingly unrelated regression models for high-dimensional multivariate variable and covariance selection in linear regression (BayesSUR) and published the corresponding paper. In another paper we have proposed a new Bayesian approach for predicting survival outcomes, which allows sharing information across heterogeneous data sets by assuming a graph linking predictors within and across different cohorts.

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 (Bayesynergy). This provides the basis for ongoing work to identify biomarkers and develop prediction models for drug synergy.

Healthcare safety management

There is an extreme amount of information available in electronic health records that can be used to learn the behaviour of healthcare institutions, make predictions, guide treatment choices and so on. We have been working on electronic health record data from Akershus University Hospital (AUH) on a project to explore patients' movements

within the hospital and how these may affect the risk of spread of infections. We have been using network models and focused on the evolution of- and differences between networks, in time and space. The project is mostly descriptive and aims to inform decision makers. The idea could be further developed into a real time surveillance tool. The topic 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 but drug-resistant clone. To address this obstacle, we have developed 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. Our statistical platform, called DECIPHER, estimates the number of distinct clones present as well as how these clones respond to a specific drug, 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 inference for mixture models. Successful implementation of our method will potentially greatly aid in the management of different types of blood cancers, and potentially also solid cancers.

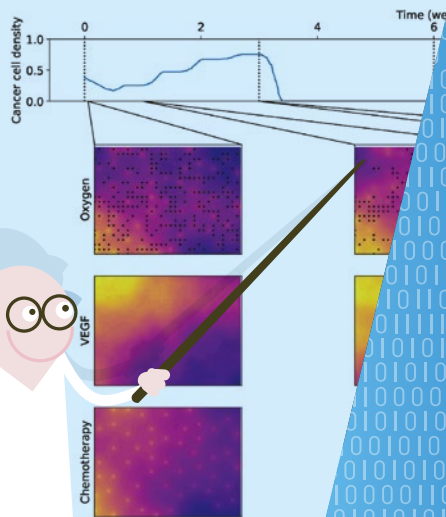
Mathematical models and Bayesian inference in personalised breast cancer therapy

Current personalized cancer treatment is based on biomarkers which allow assigning each patient to a subtype

3 | RESULTS

3.1 | Small-scale simulations

We first run personalized simulations of a patient analyzed previously on a non-parallel solver.⁹ These simulations correspond to a $200\ \mu\text{m} \times 300\ \mu\text{m}$ exact initial cell positions available from histopathological slides. We used previously published model parameters⁹ (see Table 1). As both the model and the data confirmed that the treatment outcome after 12 weeks strongly depends on the perfusion condition, we show a simulation representing a high perfusion condition estimate.



of the disease, for which treatment has been established. Such patient classification represents 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. Our strategy for personalized cancer therapy is *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 breast tumours collected in a recent neoadjuvant clinical phase II trial at OUS. This year, 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.

Case and anti-money laundering in Norway

```

    graph TD
        A[No alert (A)] --> B[No case (B)]
        A --> C[Alert]
        C --> D[Case]
        D --> E[Not reported (C)]
        D --> F[Reported (D)]
        F --> G[Money Laundering]
        F --> H[Legitimate]
    
```

Figure 1. Typical process of monitoring, investigating and reporting suspicious transactions in a bank

transaction reporting

In Section 2, let Y_i take the value 1 if transaction i was reported, given its associated explanatory variables (i.e. \mathbf{x}_i) and 0 if not. Let \mathbf{x}_i denote vectors containing numerical variables (i.e. $\mathbf{x}_i \in \mathbb{R}^d$) and categorical variables (i.e. $\mathbf{x}_i \in \mathbb{C}^d$) for some function f . This is usually called binomial or Bernoulli regression at

$$Y_i = f(\mathbf{x}_i) = \sigma(\mathbf{x}_i^T \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})} \quad (1)$$

What we did in 2021:

Network analysis for fraud detection

Fraud often involves more than a single individual. There could be groups of criminals all acting together, or one or more criminals that utilize businesses, financial services or other (innocent) individuals to perform the fraud. In such settings, network relations play a fundamental role. This is particularly the case for money laundering where both financial transactions, professional roles and social relations form networks relevant for modelling. Graph neural networks (GNN) is a model class which allows neural network models to be built on top of such graph data structures. A master student in the Data Science program at University of Oslo is working on this topic. The project uses data from DNB to model and detect money laundering with GNNs working on a heterogenous graph consisting of both transactions and professional role networks.

Statistical embeddings: A survey

Network analysis work the past few years 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 is under review. We believe that this comprehensive knowledge will inspire further work.

Detecting structuring or smurfing

Structuring is the act of parceling what would otherwise be a large financial transaction into a series of smaller transactions to avoid scrutiny by regulators and law enforcement. Criminal enterprises may employ several agents ("smurfs") to make the transaction. Structuring appears in money laundering and other financial crimes. Even though this is a known money laundering technique, methods for detecting smurfing are pretty scarce in the scientific literature. We are devising methods to search for and detect smurfing patterns, which can be used as rules directly or as complex features in a machine learning model. The usefulness of the approach applied on a large dataset will be evaluated in 2022 in cooperation with DNB.

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. We have developed a method to predict sentiments of Gjensidige insurance chats. Chats are instant messages that Gjensidige customers can use to

ask questions to customer service. Detecting 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 already being used by Gjensidige and there is interest from other BigInsight partners as well.

Copula regression

Traditional regression methods model the conditional probability of fraud given the covariates directly. In copula regression, which is an upcoming field, this conditional model is instead inferred from the joint distribution of the response and the covariates that are constructed with a copula. This allows for a lot of flexibility, especially for the modelling of interactions. However, the existing inference methods for copula regression handle only rather low dimensions. We have been working on new inference methods that are suitable for the dimensions we encounter in fraud detection. This approach can develop into a game changer.

Concept drift

Concept drift means that the statistical properties of the response variable, for example fraud or not, which the model is trying to predict, change over time in unforeseen ways. This causes problems for models because the predictions become less accurate as time passes. Concept drift is a typical challenge when modeling fraud. This can be due to new rules and regulations, time varying control strategies, new fraud methods, etc. Partly inspired by the Biofouling project in Sensor, we started an exploratory activity on modeling concept drift of VAT tax avoidance data from Skatteetaten.



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 and industrial sector. In addition, we consider the research activity with Statistics Norway as 'sensing' society.

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. Our methodology is however very often of generic value. 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 2021:

Scalable change and anomaly detection

In January 2021, BigInsight PhD Martin Tveten defended his PhD thesis "Scalable change and anomaly detection in cross-correlated data" in a digital defence, due to Covid restrictions. The thesis contains two papers on dimension reduction with PCA specifically tailored for changepoint detection, one on a 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.

The work on real-time prediction of motor overheating with ABB was published and has gained attention within ABB internationally. There is a forthcoming company publication based on the paper, and the operational implementation at ABB of the methodology is currently ongoing. Furthermore, a master student successfully defended his thesis investigating temperature prediction using recurrent neural networks/LSTMs. The thesis developed new algorithms to handle the extensive data material (2.2 billion observations equal to one year of data at 1Hz resolution in 16 ships) and concluded that the RNNs could give substantial improvements in prediction error over the linear baseline model.

A new PhD student, funded through dScience, started working on developing new methods and theoretical results for fast, online, scalable change and anomaly detection. The project has special emphasis on contextual anomalies and changepoints, that is, detecting anomalies or changes when the baseline behaviour (context) changes over time.

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. In our earlier projects, we developed methods for clustering and automatic labelling of such messages using time series provided by ABB. In 2021 in collaboration with ABB, we studied more direct methods for prediction of messages using models for multivariate point processes combined with machine learning techniques. Here, the intensities of the point processes are modelled through earlier observed points (Hawkes processes) in a non-parametric setting. In particular, a method called Temporal Event Boosting has been proposed, a gradient boosting approach for multivariate point process data, where we discretize time and adapt boosting to counts of messages within time intervals. In another project, sensors are

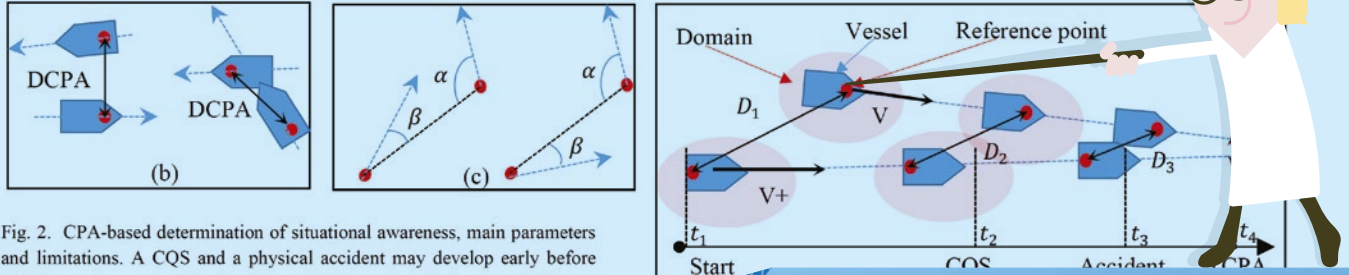
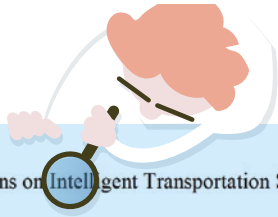


Fig. 2. CPA-based determination of situational awareness, main parameters and limitations. A CQS and a physical accident may develop early before TCPA as shown in (a). (b) shows DCPA safety limits have different interpretations depending on the type of encounter between vessels. Pose at CPA in (c) does not ensure accurate situation assessment with symmetric safety evaluation functions. CPA ignores vessel manoeuvres, encounter, and COLREGs-defined responsibility.



included in the modelling of processes. In this case, only a subset of the sensors is assumed to influence the point process. A Bayesian hierarchical approach has been taken.

A master student successfully defended his thesis on combining all available sensor data and logged error messages for a specific but anonymised vessel from ABB, in order to build a predictive detector for one specific, labelled event. Results were excellent and a paper based on the thesis (title: Multiple Instance Learning with Random Forest for Event-Logs Analysis and Predictive Maintenance in Ship Electric Propulsion Systems), has been accepted for publication.

This year, we have worked with ABB on integrating topology drawings with sensor data and textual log messages in order to automatically detect, classify and predict partial blackout of a ship power plant. The problem of detection and classification of such blackouts for several different ships has been completed successfully, and will work as a basis for further development towards prediction.

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. A nærings-PhD with ABB is dedicated to studying this topic. A first paper was published, based on an adaptive division of the vibration signal into a number of frequency bands, time-domain segmentation algorithm and high-resolution maximum likelihood frequency estimation to discover small vibration pulses excited by the defect in the bearing. We have also made progress on improved bearing fault detection by identification of different bearing degradation states and making use of prior information from previous measurement points.

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

We have finalized the work on creating realistic test beds for autonomous vessel algorithms. These will be part of a large autonomous navigation vessel system used for the future classification of such vessels performed by DNV. Combining collision avoidance rules, expert knowledge, historical AIS data, climate data and data bases for historical accidents, we describe virtual test beds where complexity and frequency of situations are realistic and at the same time span out the risks. Algorithms have been transferred to DNV in Trondheim and tested and developed further there, and we worked out an IP agreements between UiO and DNV, so that the methods can be safely used by DNV in their future developments and possible additional commercial advantages are properly shared.

Following up on our first paper from 2020, in 2021 we implemented and published a second and much more complex paper on scenario design, where collision avoidance rules (COLREGs) have been carefully 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.

A second paper, using fuzzy logic for collaborative decision-making analysis of autonomous ships in complex situations, was accepted for publication and will appear in 2022.

Also based on AIS data, we contributed to the future revision of maritime collision avoidance rules (COLREGs) by using large amounts of historical navigation behavior data to make statistics on how the rules are in fact practiced in real situations. A paper by UiO and DNV authors was published in 2021.

Our work on autonomous vessels test beds and collision avoidance rules is likely to be the most advanced in the world, and a game changer in the field.

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

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. We have been working on a project with DNV on sensor-based lithium-ion batteries diagnostics, partly working with publicly available data, and partly with operational data from DNV's collaborator Corvus Energy, which is one of the major producer in the world of maritime lithium-ion batteries. The aim is to develop methods for data driven monitoring of battery health, based on historical data from operating vessels.

A survey paper on this topic was written and published in 2021. We have also finalized a study of SOH (State of Health) degradation under dynamic conditions using publicly available data and explainable statistical models (fractional polynomials), with excellent prediction results.

SOH predictions from these models are computationally much faster, and more precise, than sequential deep learning (recurrent neural networks, LSTM) that we have also applied to this data set. A paper was finalised and submitted in 2021.

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 have worked on a method for extracting pseudo capacities that can be used instead, and this work was practically finished in the end of 2021. We are now ready to study how the extracted pseudo capacities can be used to learn an efficient predictive model.

Two master students worked with battery diagnostics master projects. One of them performed a comparative study of the whole spectra of machine learning methods for the SOH degradation using the publicly available data, while the other master student exploited error-in-variable models for total capacity estimation. The theses will be finished in the spring of 2022.

It is our ambition to contribute in a fundamental way to battery health diagnostics and prognostics by the end of BigInsight.

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 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 piecewise spline on a residual model based on GAM, which is a Bayesian generative model, and a Random Forest. The results indicate that estimates of the effect of biofouling can be obtained by carefully modelling the total shaft power of a ship in different conditions. This can provide useful information when deciding on a cleaning schedule for a ship, or for general assessments of ship performance. It may also be used to gain valuable insight into the bio fouling process. Results from the project will be presented on the

7th World Maritime Technology Conference in Copenhagen, April 2022. In addition, a more detailed paper has been submitted for publication.

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 non-probability sample in reality. We have developed models for adjusting two fallible classifiers jointly observed in a nonprobability sample. Comparisons are made with 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.

We have also extended our previous results on how a shock (like Covid-19) influences the data sources in different ways. This is solved using purposive sampling for the period of time where the shock has influence, before turning to standard methods when the shock period has passed.



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 2021:

Quantile regression and copula coupling for wind farm production

The goal was to create a methodology for issuing distributional forecasts of wind farm production based on the output of numerical weather prediction forecasts of wind speed. This involved using quantile regression to statistically post process weather forecasts and then copula methods for correlating estimates between individual wind turbines in a wind park. The underlying methodology and approach were investigated in a generic context. A separate contract with Hydro allowed us to implement the system in a production framework. It is now operational at Hydro, used daily and is currently being expanded to additional wind parks.

Principal component regression modelling for electricity consumption

Temperature forecasts drive forecasts of electricity consumption in the Nordic market and have formed the backbone of much market modelling for decades. However, temperature fields are extremely high dimensional and

previous approaches used unnecessary simplifications to accommodate this matter. This tool investigated the use of principal components analysis to reduce high-dimensional temperature fields to a lower dimensional feature set, followed by generalized additive models to forecast consumption. The underlying general methodology and approach were investigated with positive results, in particular in comparison to more involved machine learning methods proposed in the literature. The model was then put into production at Hydro.

Forecasts of country-level renewable energy production using ridge regression

The amount of renewable energy production has a substantial effect on electricity prices and therefore accurate forecasts of this quantity on a regional level are critical. However, weather forecasts of wind speeds and solar irradiation are subject to substantial dislocation issues; it is difficult to pinpoint the exact hour a wind gust will arrive, etc. In 2021 we finalized a model that used a ridge regression framework to substantially expand the covariate set of a renewable energy production forecasting system.

“What If” volume sensitivity model using generalized additive models

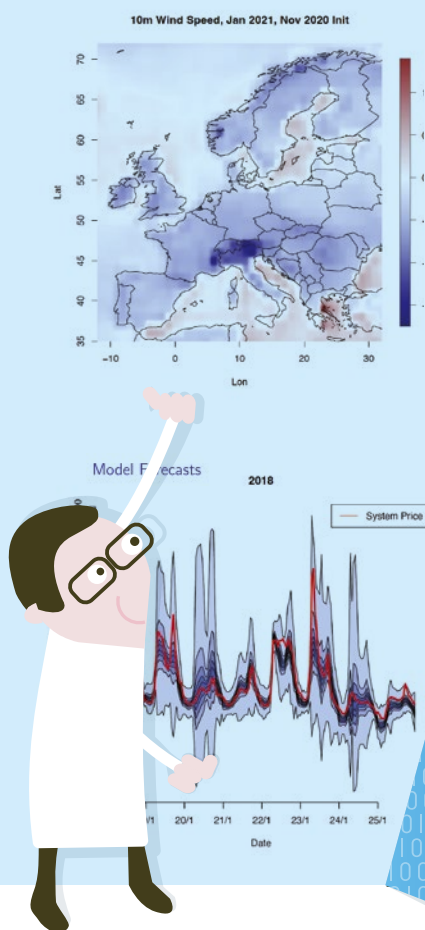
The industrial partner required a scenario analysis tool to test what the effect of adding or subtracting a substantial amount of volume to the system would be, if, for instance a large power plant suddenly went offline. They had been using the distributional price forecasting system to accomplish this, but this was inappropriate since such “shocks” are not implicit in the uncertainty bands. We investigated an additional layer using a generalized additive model to forecast the price effect from substantial, ex ante, deviations in market volume. The methodology was put into production at Hydro.



Principal Investigator
Alex Lenkoski



co-Principal Investigator
Carlo Mannino





EXPLAINING AI

At the intersection between artificial intelligence (AI), transparency, privacy and law, there is a need for more research. This IO focuses on explaining AI's black box models and related issues.

AI, statistical models and machine learning methods can often be seen as black boxes to those who construct the model and/or to those who use or are exposed to the methods. 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) complex 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 2021:

Seminar series

In 2021, we organized eight seminars, on the themes: "A decision-theoretic approach for model interpretability in Bayesian framework", "Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges", "True to the model or true to the data?", "SAGE, a principled approach to calculating global feature importance using Shapley values", "CF I – An introduction to counterfactuals from an explainable AI point of view (the non-technical version)", "CF II – A more in depth look at counterfactuals from an explaining AI point of view (the more technical version)", "CARE: Coherent Actionable Recourse based on Sound Counterfactual Explanations" and "MCCE: Monte Carlo sampling of realistic counterfactual explanations". Attendance and discussions were very good, and the seminar series continues into 2022 (with webinars and seminars in person). We have now built a scientific community on explainable AI, across the BigInsight partners.

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 our new method was in 2021 published in Artificial Intelligence. We continue to add new features to our R package – shapr, which in 2021 became available at CRAN (The Comprehensive R Archive Network). We also continued to improve our Shapley methods further 1) by using variational autoencoders to model dependent, mixed features better (paper submitted), 2) by grouping similar features for improved efficiency and interpretability (conference paper published), 3) by modelling the dependence between features with non-parametric vine copulas (paper published) and 4) by comparing Shapley with other explanation methods (conference paper published). In addition

to estimating Shapley values, we have devised a new counterfactual method we have called “MCCE”, which can be used to Monte Carlo sample realistic counterfactual explanations (conference paper submitted). Our work in this area has obtained significant attention, positioning us well internationally.

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 and 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 is providing very useful feedback on our shapr package and the new MCCE method. Further BigInsight partners can follow suit in 2022 as test beds.

A national role

We organised an “Opening the black box” session in the 28th Nordic Conference in Mathematical Statistics, and we have contributed to various XAI seminars, conferences and the course “Legal Technology: Artificial Intelligence and

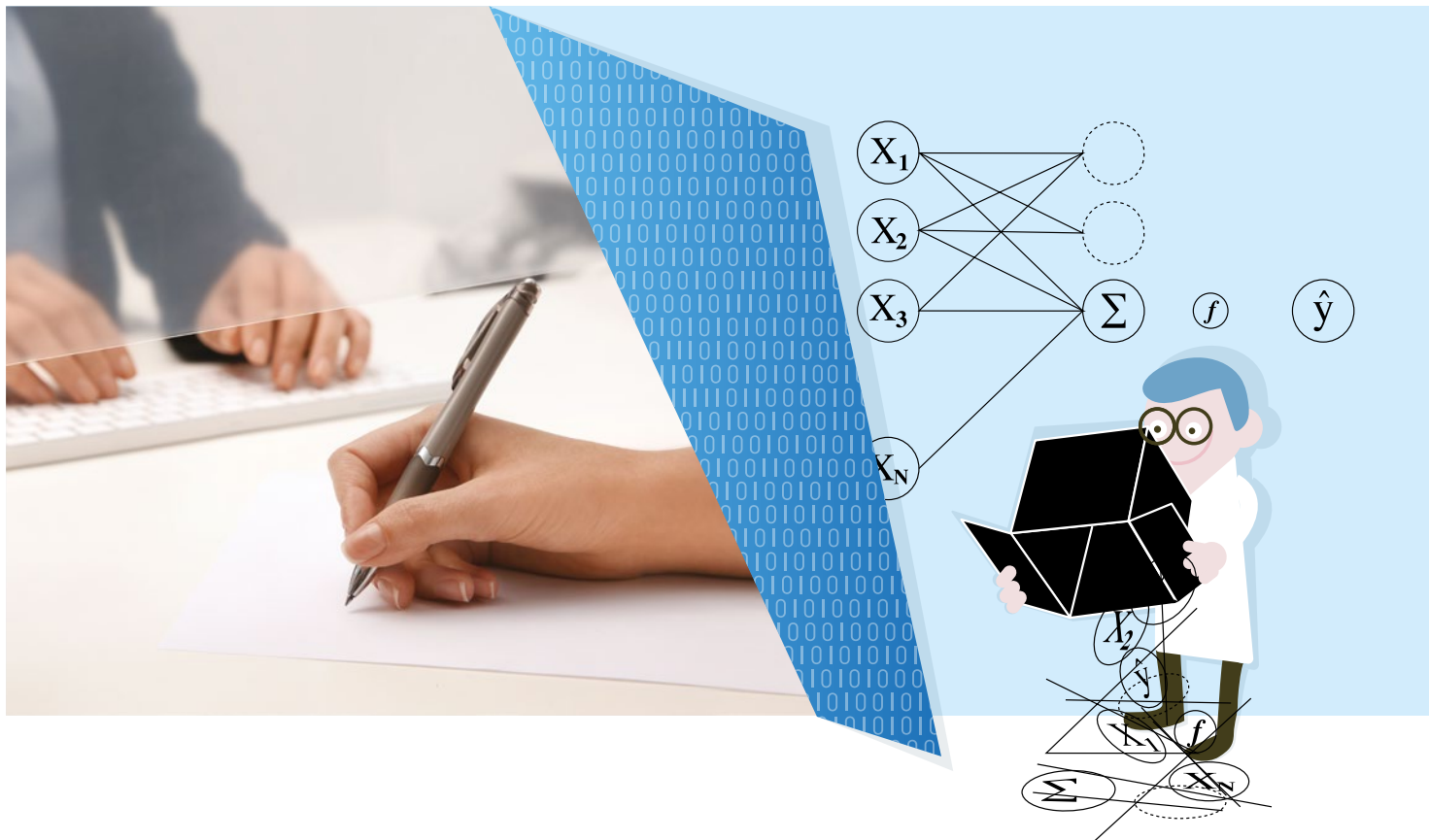
Law” at the Department of Public and International Law, UiO. We have been and will continue to be an important voice in the Norwegian AI debate. In 2021, we contributed to the Norwegian commission for data privacy, which will deliver a Norwegian Official Report in September 2022. During 2021, NAV has participated with its planned solution for prediction of absence due to illness (and accompanying explanations) in the Norwegian Data Protection Authority’s Sandbox for responsible artificial intelligence.



Principal Investigator
Anders Løland



co-Principal Investigator
Arnaldo 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 40 important UK industrial partners. The centre is at the forefront of international research effort in statistics and operation research, establishing an enviable track record of theoretical innovation arising from real world challenges. Professors Jonathan Tawn, professor Idris Eckley (who co-lead the centre) and professor David Leslie co-supervise PhD students together with BigInsight staff, on recommender systems, reinforced learning, multivariate extremes, non-parametric isotonic spatial regression, Bayesian modelling, multivariate sensor data, pair copula models. BigInsight and STOR-i co-organise industrial statistics sessions in international conferences and exchange chairing each other's scientific advisory boards.



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

The Medical Research Council Biostatistics Unit (BSU)

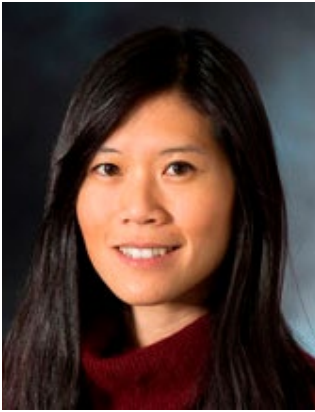
is part of the University of Cambridge, School of Clinical Medicine. It is a major centre for research, training and knowledge transfer, with a mission 'to advance biomedical science and human health through the development, application and dissemination of statistical methods'. BSU's critical mass of 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. BigInsight and the BSU have several joint projects in health and molecular biology. We are both partners in RESCUER, a H2020 project. We will continue our exchange programme, with one Oslo postdoc and one Oslo PhD student working at BSU.



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 develop new educational opportunities in mathematical modeling 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.



Jasmine Foo, Deputy Director, Institute for Mathematics and its Applications, Professor, School of Mathematics, University of Minnesota



Kevin Leder, Associate Professor, Industrial and Systems Engineering, University of Minnesota



Elisabeth Grisbauer, masterstudent, The Technical University of München

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. We now continue our collaboration with these universities, by supervising PhD students. Currently 5 PhD students are working with BigInsight staff, on themes spanning from infectious diseases modelling to diet diversity, from genetic vs environmental causes of non-communicable diseases to tests for hepatitis. We hope to be able to obtain new funding in the future for capacity building of data science in Ethiopia



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 and 2021, the programme could not host any visit because of the covid-19 pandemics. As soon as possible, we will start our guest programme again.

Students visiting BigInsight

PhD students from other universities spent periods of training and research collaboration at BigInsight. In the last two years, this programme has been working with remote access. We have hosted two students from ENSTA Paris, one of the best French engineering school in digital, computer science and mathematics, the Technical University of München, and the University of Milano.

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.



BigInsight is partner of Nordforsk project:

Data streams and mathematical modelling pipelines to support preparedness and decision making for COVID-19 and future pandemic

The goal of this programme is to, for the first time, create a joint Nordic long-term academic collaboration on pandemic preparedness using advanced mathematical modelling and systematically collected health data from a broad range of sources. To start off the programme involves Finland (Aalto), Norway (BigInsight), and Sweden (Stockholm), but our ambition is to also include Denmark, Iceland, and the Baltic countries later on. The programme participants comprise epidemiologists, statisticians, mathematicians, and computer scientists. The aim of the programme is to use public health data combined with real-time data streams representing social activity and human mobility, together with advanced mathematical modelling and computational methods to address several of the most urgent questions for COVID-19 and future pandemics: What

effects do community structure, individual heterogeneities, and spatial mobility have on reproduction numbers, community immunity, and the efficacy of different preventive measures? How can real-time data streams of social activity and human mobility combined with clinical health data aid in making more accurate predictions and more informed control decisions related to structurally and geographically targeted nonpharmaceutical interventions? How can Nordic health data and novel data streams of relevance for the ongoing COVID-19 and future pandemics be shared and published in a way that allows for better analyses without compromising data privacy of the individuals? The programme will develop methods, tools, and operational procedures for implementing cross-Nordic interoperable public health data pipelines, novel methodology published in international scientific journals, and support the national public health institutes in their aim to keep disease spreading low without causing too high burden on Nordic societies.



A second year of covid-19 has made a large impact. Here a home office desk.

Scientific Advisory Committee of BigInsight

Scientific Advisory Committee of BigInsight has five international members. A meeting was organised in January 2022.



Prof. Idris Eckley, Lancaster University, UK

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



Prof. Samuel Kaski, University of Helsinki, Finland

- Professor of Computer Science, Aalto University
- Director, Finnish Centre of Excellence in Computational Inference Research COIN, Aalto University and University of Helsinki
- Academy Professor (research professor), 2016-2020
- Director, Finnish Center for Artificial Intelligence FCAI, 2018-
- Statistical machine learning and probabilistic modeling



Prof. Geoff Nicholls, University of Oxford, UK

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



Prof. Marina Vannucci, Rice University, Houston, USA

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



Reader Veronica Vinciotti, University of Trento, Italy

- 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 2021

In 2021 the following PhD students affiliated to BigInsight defended their PhD thesis:



Sylvia Qinghua Liu, Department of Mathematics defended her PhD thesis “Bayesian Preference Learning with the Mallows Model” on Oct. 29, 2021.

Supervisor: Ida Scheel and Arnaldo Frigessi

Trial lecture: “Latent variable modelling for ranking data”

Adjudication committee

- Assistant Professor Cristina Mollica, Sapienza University of Rome
- Professor Brendan Murphy, University College Dublin
- Professor Nils Lid Hjort, University of Oslo

Summary

In our modern society, with the burgeoning of e-commerce and online streaming platforms, customers are overwhelmed by the choices. One important approach to solve this problem is recommender systems. Recommender systems learn customers’ preferences based on their past interactions with the website/platform, as well as the interactions data of other customers, to eventually provide a list of recommendations that is relevant to the customer.

In this work, the author studied the use of statistical models to learn customers’ preferences, with a focus

on the Bayesian Mallows Model. The author provided a new approach to learn personal preferences and make personalised recommendations from clicking data. Through experimentation, it was illustrated that the proposed method achieved good balance between recommending items that are closely related to what the customers previously interacted with, while not overlooking the issue of recommendation diversity: that is, recommending the items that are interesting, novel and surprising to the customer. The author also provided a new approach to achieve more computationally efficient preference learning.





**Martin Tveten, Department of Mathematics, defended his thesis
“Scalable change and anomaly detection in cross-correlated data”
on January 29, 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.



**Covariance matrix estimation in high
dimensions**

Martin Tveten

 **UiO** Department of Mathematics
University of Oslo



Salim Ghannoum, Institute of Basic Medical Sciences, defended his thesis “The role of Golgi fragmentation in breast cancer cell migration and tumor progression: an integrated experimental-computational approach” on October 25, 2021

Supervisor: Alvaro Köhn-Luque and Hesso Farhan

Trial lecture: “Towards a mathematically rooted approach to cancer treatment”

Adjudication committee

- Professor Ruth Baker, University of Oxford, UK
- Reader Bartłomiej Waclaw, The University of Edinburgh, UK
- Associate Professor Manuela Zucknick, University of Oslo

Summary

The malignancy of cancer is highly dependent on its ability to thrive and metastasize. Breast cancer is a highly heterogeneous disease originating from epithelial cells lining the milk ducts or lobules in the mammary gland. It is the most common type of cancer in women. Metastasis is the leading cause of breast cancer-related death. Motility is a crucial step towards escalating the metastatic capability of cancer cells, which can exhibit a broad spectrum of migration and invasion mechanisms. In normal mammalian cells the Golgi apparatus is a single-copy complex compartment often located adjacent to the centrosome in the juxtannuclear region. In many breast cancer cells, the Golgi is fragmented into many dispersed ministacks. Several pieces of evidence strongly indicate that Golgi fragmentation may open doors for fatal cascades to facilitate cancer progression and metastasis. However, neither the mechanism of the association between

Golgi fragmentation with tumor progression nor its consequences are yet known.

In this thesis I investigate the role of Golgi fragmentation in breast cancer cell migration and progression using an integrated experimental, quantitative and computational approach. The outcome of this thesis is four computational tools: DiscBIO, CellMAPtracer, cellmigRation and the mathematical modeling, in addition to highlighting potential roles for Golgi fragmentation in breast cancer progression. I consider this outcome as a starting point for further investigations.

Altogether, the results of this thesis point out the importance of cell migration for tumor progression and identify Giantin as a potential biomarker for cancer progression opening new doors with many directions for validating and further investigating the main findings.



Master theses at BigInsight

1. Daniel Piacek (2017): Detecting fraud using information from social networks [UiO – Fraud]
2. Jonas Fredrik Schenkel (2017): Collaborative Filtering for Implicit Feedback: Investigating how to improve NRK TV's recommender system by including context [UiO -Marketing]
3. Martin Tveten (2017): Multi-Stream Sequential Change Detection - Using Sparsity and Dimension Reduction [UiO - Sensor]
4. Kristin B. Bakka (2018): Changepoint model selection in Gaussian data by maximization of approximate Bayes Factors with the Pruned Exact Linear Time algorithm [NTNU - Sensor]
5. Simon Boge Brant (2018): Dynamic survival prediction for high-dimensional data [UiO - Health]
6. Tristan Hugh Curteis (2018): Focused Model Selection for Longitudinal Data [UiO - Health]
7. Amanda Haugnes Rygg (2019): GLM and GAM modelling of life insurance data [UiO - Marketing]
8. Jenine Gaspar Corrales (2019): Analyzing and Predicting Demographics of NRK's Digital Users [UiO - Marketing]
9. Eirik Halsteinslid (2019): Addressing collinearity and class imbalance in logistic regression for statistical fraud detection [UiO - Fraud]
10. Amirhossein Kazami (2019): A Semi-Supervised Approach to the Application of Sensor-based Change-Point Detection for Failure Prediction in Industrial Instruments [NTNU - Sensor]
11. Camilla Lingjærde (2019): Tailored Graphical Lasso for Data Integration in Gene Network Reconstruction [UiO - Health]
12. Vegard Stikbakke (2019): A boosting algorithm to extend first-hitting-time models to a high-dimensional survival setting [UiO - Health]
13. Oda Johanne Kristensen (2020): Scalable Markov Chain Monte Carlo by subsampling methods [UiO - Sensor]
14. Vera Haugen Kvisgaard (2020): Can undersampling boost fraud detection? Combining undersampling with stochastic gradient boosting for high-dimensional prediction of rare events [UiO - Fraud]
15. Elyas Dawod Mohammed (2020): Count time series with application to corporate defaults [UiB - Fraud]
16. Aleksander Njøs (2020): Multiple imputation for Cox regression with sampled cohort data [UiO - Health]
17. Lars H.B. Olsen (2020): Likelihood-Based Boosting: Approximate Confidence Bands and Intervals for Generalized Additive Models [UiO – Explain]
18. Hanne Tresselt (2020): Modelling Car Insurance Data with Individual Effects [UiO - Marketing]
19. Fredrik Wollbraaten (2020): Sequential Monte Carlo and twisted state space models [UiO - Sensor]
20. Edward F. Bull (2021): Introducing an Efficient Approach for Expressing Uncertainty in Deep Learning with Bayesian Neural Networks [UiO - Sensor]
21. Håkon Bliksås Carlsen (2021): Studying the application of semi-supervised learning for fraud detection [UiO – Fraud]
22. Bob Betuin Fjellheim (2021): Bayesian Plackett-Luce Models for Describing Consensus in Ranking Data - Review and applications to real data [UiO - Marketing]
23. He Gu (2021): Recurrent Neural Networks for predicting ship motor temperatures aiming to help prevent motor overheating [UiO - Sensor]
24. Nikola Kaletka (2021): Effects of prior information on monotonicity directions in additive monotone regression [UiO - Health]
25. Nicolay Bjørlo Kristensen (2021): Weakly Supervised Learning for Predictive Maintenance: An Extended Random Forest Approach using Imbalanced Event Data from Hybrid Ships [UiO - Sensor]
26. Anna Skovbæk Mortensen (2021): Fraud detection using copula regression [UiO - Fraud]

27. Øystein Skauli (2021): Modelling Short Term Changes in User Interest for Online Marketplaces [UiO - Marketing]
28. Peder Nørving Viken (2021): Sequential Monte Carlo and twisted state space models - Twisting models to reduce variance [UiO - Sensor]

Active students

29. Erik Holst Aasland (planned finished spring 2022): Shapley values for dependent features using divisive clustering [NTNU – Explain]
30. Christian Grindheim (planned finished 2022): Comparative study of machine learning methods for battery state of health estimation [UiO - Sensor]
31. Fredrik Johannessen (planned finished spring 2022): Finding Money Launderers Using Heterogeneous Graph Neural Networks [UiO – Fraud]
32. Meghana Kamineni (planned finished spring 2022): The effect of non-pharmaceutical interventions to control mobility during the Covid-19 pandemics [UiO - Health]
33. Anna Kejvalova (planned finished 2022): Total capacity estimation for marine batteries using measurement error regression [UiO - Sensor]
34. Shuijing Liao (planned finished 2022): Handling class imbalance within fraud detection [UiO - Fraud]
35. Arne Rustad (planned finished spring 2022): Model-based counterfactual explanations using tabular GAN or VAE [NTNU – Explain]
36. Jørn Frøysa Hole (planned finished 2023): Handling class imbalance within fraud detection using generative models [UiO – Fraud]
37. Elise Johannessen (planned finished 2023): Analysing and forecasting energy consumption in a building on a fine time scale [UiO - Power]
38. Anders Kielland (planned finished 2023): A comparison of machine learning and mechanistic models for predicting treatment response with combined aromatase and CDK4/6 inhibitors in breast cancer patients [UiO - Health]
39. Haakon Muggenud (planned finished 2023): Recommender systems for non-web-based commerce [UiO - Marketing]
40. Thomas Mullaly (planned finished 2023): Data science within insurance [UiB - Fraud]
41. Ingvild Riiser (planned finished 2023): Generating Synthetic Event Data for Labour and Welfare Studies [UiO - Explaining AI]
42. Eirik Sjøvik (planned finished 2023): Energy Demand Forecasting [UiO - Power]

ACTIVITIES AND EVENTS

2021 BigInsight Day

The annual BigInsight Day for 2021 was held via Zoom on February 19th 2021. This is the occasion for the larger BigInsight community to meet, and unfortunately this has not been possible in the last two years. The virtual BigInsight Day included short presentations of our own projects, a panel discussion with journalists and modellers about communicating science and uncertainty in the covid-19 pandemics, a discussion about the regulatory sandbox for AI and Privacy and a magical presentation of artificial reality. The next BigInsight Day 2022 is planned in the autumn at NAV.



Oslo Data Science Day

BigInsight, SIRIUS, DataScience@UiO and the newly established UiO Centre for Computational and Data Science, dScience, arranged the Data Science Day on October 27th 2021. It was physically held at the Science Library and Sophus Lie's auditorium with more than 700 registered participants.

The event was opened by an address from the Minister responsible for digitalisation, Bjørn Arild Gram, to celebrate the opening of the dScience centre, followed by a panel discussion with invited participants from academia and industry. BigInsight and SIRIUS' partners had booths for recruiting.

There were two exciting, invited speakers. Professor Lilja Øvrelid from the Language Technology Group at UiO, and Simen Eide, PhD of BigInsight and data scientist at finn.no, talked us through the various challenges and opportunities in data science research and its application to real life problems.

dScience
Senter for data- og beregningsvitenskap



Seminars

BigInsight's biweekly Wednesday lunch takes place at the Department of Mathematics and NR alternatingly. In 2021 9 lunches were organized, see our webside for a list of invited speakers. Due to covid-19 the lunches, except for one, were held via Zoom. 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 autumn, 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 all 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.





TRAINING AND COURSES

The University of Oslo established a new Master Program in Data Science in 2018. 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 time spent 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, but only around 25 are 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. See list above.

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 various series of digital seminars and meetings, journal clubs etc. have been organised to keep active contacts between students. PhD students and postdocs have been prioritized in periods when access to the university campus has been limited.



COMMUNICATION AND DISSEMINATION ACTIVITIES

Website

The website of the center is biginsight.no.

BigInsight outreach presentations

BigInsight has been present in the Norwegian media in 2021 because of our work in Covid-19 modelling and in explainable AI. We hold seminars and participate 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. We are also central to the UiO centre dScience and play an advising role in the building of the first Norwegian research district Oslo Science City. We were present in Arendalsuka, with two events:

- Kunstig intelligens – kan vi stole på den svarte boksen?
- Communication during a pandemic

BigInsight in the media (selection)

BigInsight Day 2021, 19.02.2021

Bakken, Jari; Johansen, Per Anders; Sandberg, Hallvard; Skille, Øyvind Bye; De Blasio, Birgitte Freiesleben; Engebretsen, Solveig; Britton, Tom; Frigessi, Arnaldo.

Panel debate: Mathematical models, statistical complexity, uncertainty and decision making in the Covid-19

pandemic: science, society and the public. A panel discussion with journalists and modellers.

Computerworld, 25.02.2021

Løland, Anders. **Utviklere av kunstig intelligens ber om klare rammer.**

NTNU Brain Talk, 03.2021

Arnoldo Frigessi

Covid-19 modelling and AI.

NRK, 16.03.2021

Peter Svaar. **Bekymret for sykehuskapasiteten: – I noen sykehus vil trykket bli veldig høyt.**

VG, 11.04.2021

Oda Leraan Skjetne.

Smitteekspert om R-tallet: Tror vi har sett toppen.

Sannsynligvis VIKTIG, NR's podcast, 30.04.2021

Løland, Anders; Abrahamsen, Petter; Dahle, Pål. **Fra verdifulle oljefelt til farlige løsmasser med COHIBA.**

Sannsynligvis VIKTIG, NR's podcast, 09.06.2021

Løland, Anders. **Hvor vil DNB med maskinlæring?** Med Karl Aksel Festø



Aftenposten, 18.06.2021

Aftenposten bommer om prognoser. Igjen og igjen.

Stoltenberg, Camilla; Aavitsland, Preben; Rø, Gunnar Øyvind Isaksson; de Blasio, Birgitte Freiesleben; Frigessi, Arnoldo; Engebretsen, Solveig.

Nordstat 2021, 21.06.2021

Frigessi, Arnoldo; Britton, Tom; Engebretsen, Solveig; Kuhlmann-Berenzon, Sharon; Leino, Tuija; Leskelä, Lasse. Panel debate: **The role of mathematics and statistics in the political and health political decisions in the covid crises.**

Arendalsuka, 17.08.2021

Paneldepat: Hva kan vi lære av runden med covid: hvordan håndterer vi usikkerhet og åpenhet?

Stoltenberg, Camilla; Frigessi, Arnoldo; Engebretsen, Solveig; Sandberg, Hallvard; Johansen, Per Anders; Ruhlin Gjuvsland, Elin.

IT Conference 2021 USIT, 09.2021

Arnoldo Frigessi

Data science for the Covid19 emergency.

digi.no, 21.09.2021

Løland, Anders. - **Fem grunner til at kunstig intelligens er på rett vei.**

Pint of Science, 11.2021

Arnoldo Frigessi

Data science for the Covid19 emergency.

Nordic Society of Human Genetics and Precision Medicine, 04.11.2021

Arnoldo Frigessi

The past and future of modelling in pandemics.

Kulturhuset i Oslo: Dataanalyser som redder verden – science not fiction, 10.11.2021

Engebretsen, Solveig. **Modellering av R-tallet i sanntid. Når PhD-arbeidet plutselig blir beslutningsgrunnlag under en pandemi.**

NRK radio P2, 12.11.2021

Debatt om utnyttelse av norske helsedata. Arnoldo Frigessi

Det Norske Videnskaps-Akademi, 18.11.2021

Ingrid Glad. **Hva er så stort med stordata?**

Bioteknologirådet, 29.11.2021

Alvaro Köhn-Luque, Vincenzo Politi, mfl

Hvor presis er egentlig presisjonsmedisin?

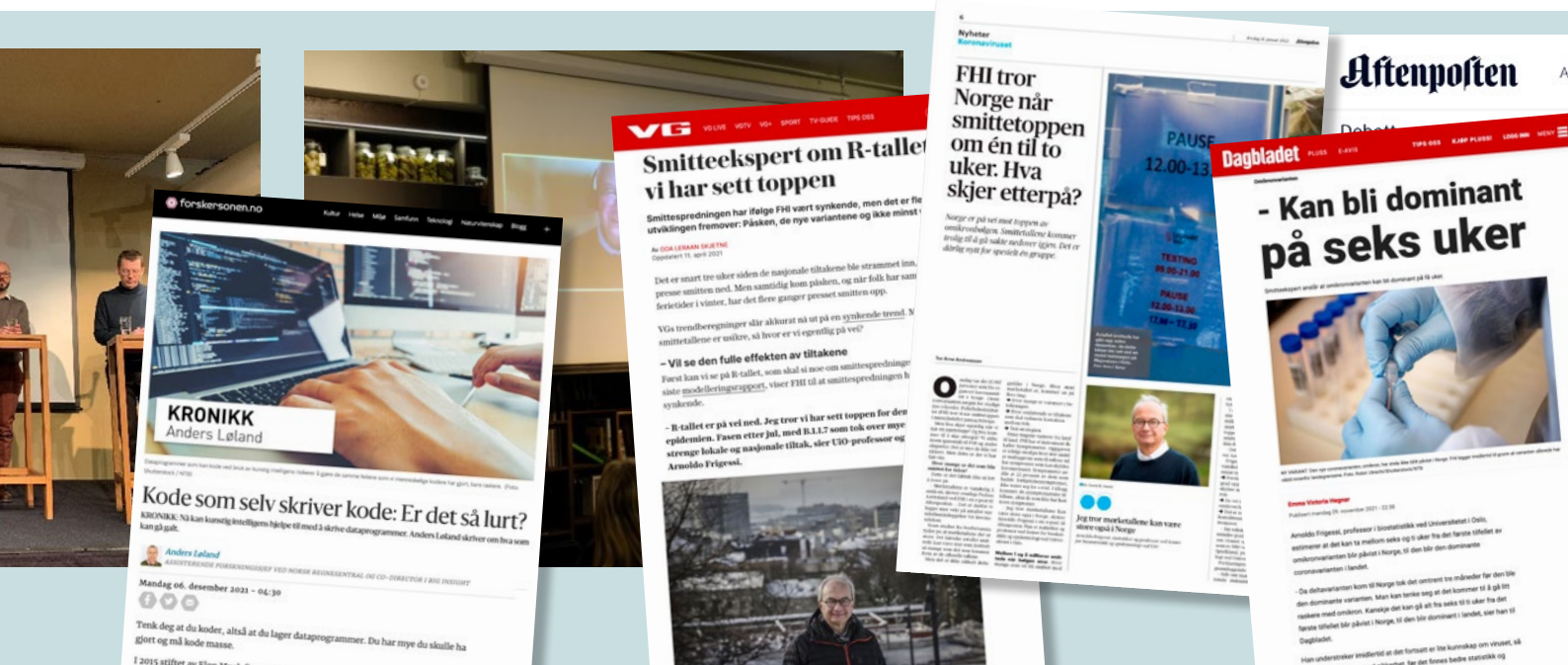
Dagbladet, 29.11.2021

Emma Victoria Hegnar

Kan bli dominant på seks uker.

Forskning.no. 06.12.2021

Kode som selv skriver kode: Er det så lurt? Løland, Anders



RECOGNITIONS

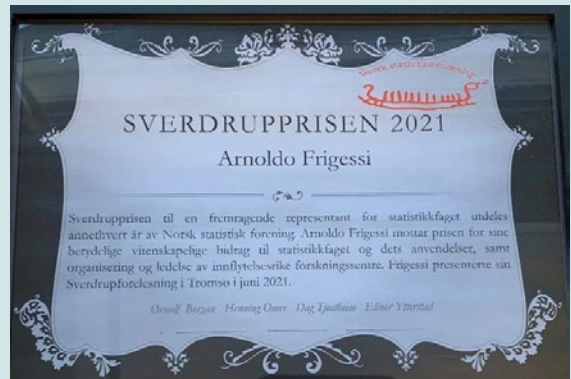
BigInsight PHD Ingrid Dæhlen received the 'NRs masterpris 2021'! The title of her master thesis is Empirical and Hybrid Likelihood and Nils L. Hjort was her supervisor.



Prof. Birgitte Freiesleben De Blasio has been elected member of Det Norske Videnskaps-Akademi in the nominations made in 2021



The Sverdrup Prize is the research prize of the Norwegian Statistical Association. Arnaldo Frigessi received the Sverdrup Prize at The 28th Nordic Conference in Mathematical Statistics in Tromsø on 23 June, 2021.



Prof. Arnaldo Frigessi awarded the honour of Knight of the Order of Merit of the Italian Republic, by the President of the Italian Republic in Oslo, 12 December 2021.



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
Jaroslav 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
Qinghua Liu	CoopX	Marketing
Christine Lunde	CoopX	Marketing
Lars Erik Bolstad	DNB	Fraud
Bjørn Ingeberg Fesche	DNB	Fraud
Karl Aksel Festø	DNB	Marketing
Andreas Bendixen Hovdenes	DNB	Marketing
Fredrik Johannessen	DNB	Marketing
Tobias Lillekvelland	DNB	Marketing
Marcus Nilsson	DNB	Marketing
Hodjat Rahmati	DNB	Marketing
Aiko Yamashita	DNB	Marketing, Explaining AI
Lars Holterud Aarsnes	DNV	Sensor
Øystein Alnes	DNV	Sensor
Ole Christian Astrup	DNV	Sensor
Håvard Nordtveit Austefjord	DNV	Sensor
Andreas Brandsæter	DNV	Sensor
Christos Chryssakis	DNV	Sensor
Øystein Engelhardtsen	DNV	Sensor
Ørjan Fredriksen	DNV	Sensor
Tom Arne Pedersen	DNV	Sensor
Gaute Storhaug	DNV	Sensor
Hans Anton Tvette	DNV	Sensor
Bjørn-Johan Vartdal	DNV	Sensor
Erik Vanem	DNV	Sensor, Power
Nikos Violaris	DNV	Sensor
Bendik Bull	Gjensidige	Marketing
Sindre Froyen	Gjensidige	Marketing
Jørgen Andreas Hagen	Gjensidige	Marketing
Mikkel Hinnerichsen	Gjensidige	Marketing

NAME	INSTITUTION	MAIN RESEARCH AREA
Hayat Mohammed	Gjensidige	Marketing
Gunnhildur Steinbakk	Gjensidige	Fraud
Geir Thomassen	Gjensidige	Fraud
Stefan Erath	Hydro	Power
Plamen Mavrodiev	Hydro	Power
Ellen Paaske	Hydro	Power
Peter Szederjesi	Hydro	Power
Birgitte De Blasio	NIPH	Health
Anja Bråthen Kristoffersen	NIPH	Health
Jonas Lindstrøm	NIPH	Health
Jørgen E. Midtbø	NIPH	Health
Alfonso D.L. Palomares	NIPH	Health
Francesco Di Ruscio	NIPH	Health
Gunnar Rø	NIPH	Health
Camilla Stoltenberg	NIPH	Health
Robindra Prabhu	NAV	Explaining AI
Cathrine Pihl Lyngstad	NAV	Explaining AI
Lars Sutterud	NAV	Explaining AI
Jon Vegard Sparre	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
Martin Tveten	NR	Sensor
Jens Christian Wahl	NR	Marketing, 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

NAME	INSTITUTION	MAIN RESEARCH AREA
Maria Serena Giliberto	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
Sygve Nakken	OUS	Health
Andrew Reiner	OUS	Marketing, Health
Leiv Arne Rosseland	OUS	Health
Fredrik Schjesvold	OUS	Health
Therese Seierstad	OUS	Health
Therese Sørлие	OUS	Health
David Swanson	OUS	Health
Dagim S. Tadele	OUS	Health
Kjetil Tasken	OUS	Health
Xavier Tekpli	OUS	Health
Anders Berset	Skatteetaten	Marketing, Fraud
Wenche Celiussen	Skatteetaten	Marketing
Rasmus Sjøholt Engelschiøn	Skatteetaten	Fraud
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
Kenth Engo-Monsen	Telenor	Marketing, Health, Sensor
Weiying Zhang	Telenor	Health
Dag Tjøstheim	UiB/NR	Fraud
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
Waldir Leoncio Netto	UiO	Marketing, Health
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
Alvaro Köhn Luque		Spain	2021-2023	M	Health

Postdoctoral researchers in BigInsight with financial support from other sources

Tugba Akman	Turkey Research Council	Turkey	2021-2022	F	Health
Theophilus Quachie Asenso	UiO	Ghana	2021-2023	M	Health
Youness Azimzade	UiO	Iran	2021-2023	M	Health
Fekadu Bayisa	UiO	Sweden	2020-2022	M	Health
Louis Hat Hin Chan	NIPH	China	2021-2023	M	Health
Neda Jalalil	NIPH	Iran	2021-2022	F	Health
Fatih Kizilaslaw	UiO	Turkey	2021-2023	M	Health
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
Leonard Schmiester	UiO	Germany	2021-2023	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
Ingrid Dæhlen		Norway	2021-2024	F	Several
Emanuele Gramuglia		Italy	2016-2022	M	Sensor
Even Moa Myklebust		Norway	2020-2023	M	Health
Riccardo Parviero		Italy	2018-2022	M	Marketing
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	2017-2021	M	Sensor
Fredrik Wollbraaten		Norway	2020-2023	M	Sensor

NAME	FUNDING	NATIONALITY	PERIOD	GENDER	TOPIC
PhD students in BigInsight with financial support from other sources					
Henok Asefa	Norad	Ethiopia	2021-2023	M	Health
Andrea Bratsberg	UiO	Norway	2021-2023	F	Health
Hilde K. Brustad	UiO/IMB	Norway	2021-2024	F	Health
Simen Eide	Finn.no, NæringslivPhD	Norway	2018-2021	M	Marketing
Emanuele Gramuglia	ABB	Italy	2016-2022	M	Sensor
Lars Petter Johnsen	UiO	Norway	2021-2023	M	Explaining AI
Teshome Kabeta	Norad	Ethiopia	2021-2023	M	Health
Torgeir Mo	OUS	Norway	2018-2021	M	Health
Per August Moen	dScience	Norway	2021-2025	M	Sensor
Jaroslav 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
Haifeng Xu	OUS/UiO	China	2019-2023	M	Health
Chi Zhang	NIPH	China	2016-2022	F	Health
Emilie Ødegård	UiO	Norway	2019-2023	F	Health, Marketing
Master degrees					
Håkon Bliksås Carlsen			2019-2021	M	Fraud
Bob Betuin Fjellheim			2019-2021	M	Marketing
Christian Grindheim			2020-2022	M	Sensor
He Gu			2019-2021	M	Sensor
Jørn Frøysa Hole			2021-2023	M	Fraud
Elise Johannessen			2021-2023	F	Power
Nikola Kaletka			2019-2021	F	Health
Meghana Kamineneni			2021-2022	F	Health
Anna Kejvalova			2020-2022	F	Sensor
Anders Kielland			2021-2023	M	Health
Nicolay Bjørlo Kristensen			2019-2021	M	Sensor
Vera Haugen Kvisgaard			2019-2021	F	Fraud
Anna Skovbæk Mortensen			2019-2021	F	Fraud
Haakon Mugggerud			2021-2023	M	Marketing
Ingvild Riiser			2021-2023	F	AI
Øystein Skauli			2019-2021	M	Marketing
Jonas Einar Thorsen			2021-2023	M	Fraud
Research assistant (vitas)					
Severin Schirmer			2021-2022	F	Health

FINANCIAL OVERVIEW

FUNDING	1000 NOK
The Research Council	12 238
Norwegian Computing Center (NR)	1 019
Research Partners*, in kind	10 035
Research Partners*, in cash	1 370
Enterprise partners**, in kind	2 235
Enterprise partners**, in cash	3 400
Public partners***, in kind	4 104
Public partners***, in cash	1 783
Sum	36 082
COSTS	
NR, research	9 628
NR, direct costs	275
Research Partners*, research	19 541
Enterprise partners**, research	2 135
Public partners***, research	4 504
Sum	36 082

*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 2021

Journal and peer-reviewed conference papers

Aas, Kjersti; Jullum, Martin; Løland, Anders. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence* (ISSN 0004-3702). 298 doi: [10.1016/j.artint.2021.103502](https://doi.org/10.1016/j.artint.2021.103502). 2021.

Aas, Kjersti; Nagler, Thomas; Jullum, Martin; Løland, Anders. Explaining predictive models using Shapley values and non-parametric vine copulas. *Dependence Modeling* (ISSN 2300-2298). 9(1) pp 62-81. doi: <https://doi.org/10.1515/demo-2021-0103>. 2021.

Aiken, John Mark; De Bin, Riccardo; Lewandowski, Heather; Caballero, Marcos Daniel. Framework for evaluating statistical models in physics education research. *Physical Review Physics Education Research* (ISSN 2469-9896). 17(2) doi: [10.1103/PhysRevPhysEducRes.17.020104](https://doi.org/10.1103/PhysRevPhysEducRes.17.020104). 2021.

Avella, Pasquale; Boccia, Maurizio; Viglione, Sandro; Mannino, Carlo. Practice Summary: Solving the External Candidates Exam Schedule in Norway. *INFORMS Journal on Applied Analytics (IJAA)* (ISSN 2644-0865). doi: [10.1287/inte.2021.1093](https://doi.org/10.1287/inte.2021.1093). 2021.

Bakdi, Azzeddine; Glad, Ingrid Kristine; Vanem, Erik. Testbed Scenario Design Exploiting Traffic Big Data for Autonomous Ship Trials Under Multiple Conflicts With Collision/Grounding Risks and Spatio-Temporal Dependencies. *IEEE transactions on intelligent transportation systems (Print)* (ISSN 1524-9050). 22(12) doi: [10.1109/TITS.2021.3095547](https://doi.org/10.1109/TITS.2021.3095547). 2021.

Carr, Ewan; Bendayan, Rebecca; Bean, Daniel; Stammers, Matt; Wang, Wenjuan; Zhang, Huayu; Searle, Thomas; Kraljevic, Zeljko; Shek, Anthony; Phan, Hang TT; Muruet, Walter; Gupta, Rishi K.; Shinton, Anthony J.; Wyatt, Mike; Shi, Ting; Zhang, Xin; Pickles, Andrew; Stahl, Daniel; Zakeri, Rosita; Noursadeghi, Mahdad; O'Gallagher, Kevin; Rogers, Matt; Folarin, Amos; Karwath, Andreas; Wickstrøm, Kristin E.; Köhn-Luque, Alvaro; Slater, Luke; Cardoso, Victor Roth; Bourdeaux, Christopher; Holten, Aleksander Rygh; Ball, Simon; McWilliams, Chris; Roguski, Lukasz; Borca, Florina; Batchelor, James; Amundsen, Erik Koldberg; Wu,

Xiaodong; Gkoutos, Georgios V.; Sun, Jiaying; Pinto, Ashwin; Guthrie, Bruce; Breen, Cormac; Douiri, Abdel; Wu, Honghan; Curcin, Vasa; Teo, James T.; Shah, Ajay M.; Dobson, Richard J.B. Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. *BMC medicine*, 19(1), pp.1-16. 2021.

Dal Sasso, Veronica; Lamorgese, Leonardo Cameron; Mannino, Carlo; Onofri, Andrea; Ventura, Paolo. The Tick Formulation for Deadlock Detection and Avoidance in railways traffic control. *Journal of Rail Transport Planning & Management*. doi: [10.1016/j.jrtpm.2021.100239](https://doi.org/10.1016/j.jrtpm.2021.100239). 2021.

De Leo, Francesco; Besio, Giovanni; Briganti, Riccardo; Vanem, Erik. Non-stationary extreme value analysis of sea states based on linear trends. Analysis of annual maxima series of significant wave height and peak period in the Mediterranean Sea. *Coastal Engineering* (ISSN 0378-3839). 167 doi: [10.1016/j.coastaleng.2021.103896](https://doi.org/10.1016/j.coastaleng.2021.103896). 2021.

Divino, Fabio; Belay, Denekew Bitew; Keilman, Nico; Frigessi, Arnaldo. Spatial modelling of Lexis mortality data. *Spatial Statistics* (ISSN 2211-6753). 44 doi: [10.1016/j.spasta.2021.100532](https://doi.org/10.1016/j.spasta.2021.100532). 2021.

Eide, Simen; Leslie, David; Frigessi, Arnaldo; Rishaug, Joakim; Jenssen, Helge; Verrewaere, Sofie. FINN.no Slates Dataset: A new Sequential Dataset Logging Interactions, all Viewed Items and Click Responses/No-Click for Recommender Systems Research. In: *RecSys '21: Fifteenth ACM Conference on Recommender Systems*. (ISBN 978-1-4503-8458-2). doi: [10.1145/3460231.3474607](https://doi.org/10.1145/3460231.3474607) 2021.

Främling, Kary; Westberg, Marcus; Jullum, Martin; Madhikermi, Manik; Malhi, Avleen Kaur. Comparison of Contextual Importance and Utility with LIME and Shapley Values. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, *Lecture Notes in Computer Science (LNCS)* (ISSN 0302-9743). 12688 pp 39-54. doi: [10.1007/978-3-030-82017-6_3](https://doi.org/10.1007/978-3-030-82017-6_3). 2021.

Ghannoum, Salim; Antos, Kamil; Netto, Waldir Leoncio; Gomes, Cecil; Köhn-Luque, Alvaro; Farhan, Hesso. CellMAPtracer: A User-Friendly Tracking Tool for Long-Term Migratory and Proliferating Cells Associated with Fucci Systems. *Cells* (ISSN 2073-4409). 10(2) pp 1-18. doi: [10.3390/cells10020469](https://doi.org/10.3390/cells10020469). 2021.

Ghannoum, Salim; Netto, Waldir Leoncio; Fantini, Damiano; Ragan-Kelley, Benjamin; Parizadeh, Seyedamirabbas; Jonasson, Emma; Ståhlberg, Anders; Farhan, Hesso; Köhn-Luque, Alvaro. DiscBio: A User-Friendly Pipeline for Biomarker Discovery in Single-Cell Transcriptomics. *International Journal of Molecular Sciences* (ISSN 1661-6596). 22(3) doi: [10.3390/ijms22031399](https://doi.org/10.3390/ijms22031399). 2021.

Ghosh, Abhik and Thoresen, Magne. A robust variable screening procedure for ultra-high dimensional data. *Statistical Methods in Medical Research*, 30(8), pp.1816-1832. 2021.

Ghosh, Abhik and Thoresen, Magne. Consistent fixed-effects selection in ultrahigh-dimensional linear mixed models with error-covariate endogeneity. *Statistica Sinica*, 31, pp.1-30. 2021.

Gramuglia, Emanuele; Storvik, Geir Olve; Stakkeland, Morten. Clustering and automatic labelling within time series of categorical observations - with an application to marine log messages. *The Journal of the Royal Statistical Society, Series C (Applied Statistics)* (ISSN 0035-9254). 70(3) pp 714-732. doi: [10.1111/rssc.12483](https://doi.org/10.1111/rssc.12483). 2021.

Hellton, Kristoffer Herland; Tveten, Martin; Stakkeland, Morten; Engebretsen, Solveig; Haug, Ola; Aldrin, Magne Tommy. Real-time prediction of propulsion motor overheating using machine learning. *Journal of Marine Engineering & Technology* (ISSN 2046-4177). doi: [10.1080/20464177.2021.1978745](https://doi.org/10.1080/20464177.2021.1978745). 2021.

Hemerik, Jesse; Thoresen, Magne; Finos, Livio. Permutation testing in high-dimensional linear models: an empirical investigation. *Journal of Statistical Computation and Simulation* (ISSN 0094-9655). 91(5) pp 897-914. doi: [10.1080/00949655.2020.1836183](https://doi.org/10.1080/00949655.2020.1836183). 2021.

Hrafinkelsson, Birgir; Siegert, Stefan; Huser, Raphaël; Bakka, Haakon C.; Jóhannesson, Árni. Max-and-Smooth: A Two-Step Approach for Approximate Bayesian Inference in Latent Gaussian Models. *Bayesian Analysis* (ISSN 1936-0975). 16(2) doi: [10.1214/20-BA1219](https://doi.org/10.1214/20-BA1219). 2021.

Hubin A.; Storvik G.; Frommlet F. Flexible Bayesian Nonlinear Model Configuration. *Journal of Artificial Intelligence Research*. Nov 22;72:901-42. 2021.

Kloster, Oddvar; Mannino, Carlo; Riise, Atle; Schittekat, Patrick. Scheduling vehicles with spatial conflicts. *Transportation Science*. 2021.

Jullum, Martin; Redelmeier, Annabelle Alice; Aas, Kjersti. Efficient and simple prediction explanations with group-Shapley: A practical perspective. *CEUR Workshop Proceedings* (ISSN 1613-0073). 3014 2021.

Kvamme, Håvard; Borgan, Ørnulf. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis* (ISSN 1380-7870). 27(4) pp 710-736. doi: [10.1007/s10985-021-09532-6](https://doi.org/10.1007/s10985-021-09532-6). 2021.

Lai, Xiaoran; Taskén, Håkon Austlid; Mo, Torgeir; Funke, Simon Wolfgang; Frigessi, Arnaldo; Rognes, Marie Elisabeth; Köhn Luque, Alvaro. A scalable solver for a stochastic, hybrid cellular automaton model of personalized breast cancer therapy. *International Journal for Numerical Methods in Biomedical Engineering* 2021 s. 1-21. 2021.

LeBlanc, M., Rueegg, C.S., Bekiroğlu, N., Esterhuizen, T.M., Fagerland, M.W., Falk, R.S., Frøslie, K.F., Graf, E., Heinze, G., Held, U. and Holst, R., 2022. Statistical advising: Professional development opportunities for the biostatistician. *Statistics in medicine*, 41(5), pp.847-859.

Lindstrøm, Jonas Christoffer; Engebretsen, Solveig; Kristoffersen, Anja Bråthen; Rø, Gunnar Øyvind Isaksson; Diz-Lois Palomares, Alfonso; Engø-Monsen, Kenth; Madslie, Elisabeth Henie; Forland, Frode; Nygård, Karin Maria; Hagen, Frode; Gantzel, Gunnar; Wiklund, Ottar; Frigessi, Arnaldo; de Blasio, Birgitte Freiesleben. Increased transmissibility of the alpha SARS-CoV-2 variant: evidence from contact tracing data in Oslo, January to February 2021. *Infectious Diseases* (ISSN 2374-4235). 54(1) pp 72-77. doi: [10.1080/23744235.2021.1977382](https://doi.org/10.1080/23744235.2021.1977382). 2021.

Lingjærde, Camilla; Lien, Tonje G.; Borgan, Ørnulf; Bergholtz, Helga; Glad, Ingrid K.. Tailored graphical lasso for data integration in gene network reconstruction. *BMC Bioinformatics* (ISSN 1471-2105). 22(1) doi: [10.1186/s12859-021-04413-z](https://doi.org/10.1186/s12859-021-04413-z). 2021.

- Madjar, Katrin; Zucknick, Manuela; Ickstadt, Katja; Rahnenführer, Jörg. Combining heterogeneous subgroups with graph-structured variable selection priors for Cox regression. *BMC Bioinformatics* (ISSN 1471-2105). 22(1) pp 1-29. doi: [10.1186/s12859-021-04483-z](https://doi.org/10.1186/s12859-021-04483-z). 2021.
- Mancisidor, R.A.; Kampffmeyer, M.; Aas, K. and Jenssen, R. Learning latent representations of bank customers with the Variational Autoencoder. *Expert Systems with Applications*, 164, p.114020. 2021.
- Mannino, Carlo; Avella, Pasquale; Boccia, Maurizio; Viglione, Sandro. Solving the external candidates exam schedule in Norway. *INFORMS Journal on Applied Analytics (IJAA)*. 2021.
- Maros, Máté E.; Capper, David; Jones, David T.; Hovestadt, Volker; von Deimling, Andreas; Pfister, Stefan M; Benner, Axel; Zucknick, Karola Manuela; Sill, Martin. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nature Protocols* (ISSN 1754-2189). 15 pp 479-512. doi: [10.1038/s41596-019-0251-6](https://doi.org/10.1038/s41596-019-0251-6). 2020.
- Mo, T., Brandal, S.H.B., Köhn-Luque, A., Engebraaten, O., Kristensen, V.N., Fleischer, T., Hompland, T. and Seierstad, T., 2022. Quantification of Tumor Hypoxia through Unsupervised Modelling of Consumption and Supply Hypoxia MR Imaging in Breast Cancer. *Cancers*, 14(5), p.1326.
- Moss, Jonas; De Bin, Riccardo. Modelling publication bias and p-hacking. *Biometrics* (ISSN 0006-341X). doi: [10.1111/biom.13560](https://doi.org/10.1111/biom.13560). 2021.
- Picchini, U., Simola, U. and Corander, J., 2022. Sequentially guided MCMC proposals for synthetic likelihoods and correlated synthetic likelihoods. *Bayesian Analysis*, 1(1), pp.1-31.
- Roos, Malgorzata; Hunanyan, Sona; Bakka, Haakon C.; Rue, Håvard. Sensitivity and identification quantification by a relative latent model complexity perturbation in Bayesian meta-analysis. *Biometrical Journal* (ISSN 0323-3847). doi: [10.1002/bimj.202000193](https://doi.org/10.1002/bimj.202000193). 2021.
- Ryalen, Pål Christie; Møller, Bjørn; Laache, Christoffer Haug; Stensrud, Mats Julius; Røysland, Kjetil. Prognosis of cancer survivors: estimation based on differential equations. *Biostatistics* (ISSN 1465-4644). doi: [10.1093/biostatistics/kxab009](https://doi.org/10.1093/biostatistics/kxab009). 2021.
- Rønneberg, Leiv; Cremaschi, Andrea; Hanes, Robert; Enserink, Jorrit.M. and Zucknick, Manuela. bayesynergy: flexible Bayesian modelling of synergistic interaction effects in in vitro drug combination experiments. *Briefings in bioinformatics*, 22(6), p.bb251. 2021.
- Simões, Rui F.; Pino, Rute; Moreira Soares, Mauricio; Kovarova, Jaromira; Neuzil, Jiri; Travasso, Rui; Oliveira, Paulo J.; Cunha-Oliveira, Teresa; Pereira, Francisco B.. Quantitative analysis of neuronal mitochondrial movement reveals patterns resulting from neurotoxicity of rotenone and 6-hydroxydopamine. *The FASEB Journal* (ISSN 0892-6638). 35(12) doi: [10.1096/fj.202100899R](https://doi.org/10.1096/fj.202100899R). 2021.
- Suotsalo, Kimmo; Xu, Yingying; Corander, Jukka; Pensar, Johan. High-dimensional structure learning of sparse vector autoregressive models using fractional marginal pseudo-likelihood. *Statistics and computing* (ISSN 0960-3174). 31(73) doi: [10.1007/s11222-021-10049-z](https://doi.org/10.1007/s11222-021-10049-z). 2021.
- Thomas, Owen., Dutta, R., Corander, Jukka, Kaski, S. and Gutmann, M.U., 2022. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 17(1), pp.1-31.
- van Niekerk, Janet; Bakka, Haakon C.; Rue, Håvard. A principled distance-based prior for the shape of the Weibull model. *Statistics and Probability Letters* (ISSN 0167-7152). 174 doi:[10.1016/j.spl.2021.109098](https://doi.org/10.1016/j.spl.2021.109098). 2021.
- van Niekerk, Janet; Bakka, Haakon C.; Rue, Håvard. Stable Non-Linear Generalized Bayesian Joint Models for Survival-Longitudinal Data. *Sankhya A, The Indian Journal of Statistics* (ISSN 0976-836X). doi: [10.1007/s13171-020-00233-0](https://doi.org/10.1007/s13171-020-00233-0). 2021.
- van Niekerk, Janet; Bakka, Haakon C.; Rue, Håvard. Competing risks joint models using R-INLA. *Statistical Modelling* (ISSN 1471-082X). pp 1-16. doi: [10.1177/1471082X20913654](https://doi.org/10.1177/1471082X20913654). 2021.
- Vanem, Erik; Aarsnes, Lars Holterud; Storhaug, Gaute; Astrup, Ole Christian. Statistical Modelling and Comparison of Model-Based Fatigue Calculations and Hull Monitoring Data for Container Vessels. In: Practical Design of Ships and Other Floating Structures. Proceedings of the 14th International Symposium, PRADS 2019, September 22-26, 2019, Yokohama, Japan- Volume II. Springer. (ISBN 978-981-15-4671-6). pp 517-536. doi: https://doi.org/10.1007/978-981-15-4672-3_33. 2021.

Vestre, Arnstein; Bakdi, Azzeddine; Vanem, Erik; Engelhardt, Øystein. AIS-based near-collision database generation and analysis of real collision avoidance manoeuvres. *Journal of navigation* (ISSN 0373-4633). 74(5) pp 1-24. doi: [10.1017/S0373463321000357](https://doi.org/10.1017/S0373463321000357). 2021.

Wahl, Jens Christian; Aanes, Fredrik L; Aas, Kjersti; Froyn, Sindre; Piacek, Daniel. Spatial modelling of risk premiums for water damage insurance. *Scandinavian Actuarial Journal* (ISSN 0346-1238). doi: [10.1080/03461238.2021.1951346](https://doi.org/10.1080/03461238.2021.1951346). 2021.

Wallisch, C.; Dunkler, D.; Rauch, G.; De Bin, R. and Heinze, G. Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Statistics in medicine*, 40(2), pp.369-381. 2021.

Wickstrøm, Kristin E.; Vitelli, Valeria; Carr, Ewan; Holten, Aleksander R.; Bendayan, Rebecca; Reiner, Andrew H.; Bean, Daniel; Searle, Tom; Shek, Anthony; Kraljevic, Zeljko; Teo, James; Dobson, Richard; Tonby, Kristian; Köhn-Luque, Alvaro; Amundsen, Erik K. Regional performance variation in external validation of four prediction models for severity of COVID-19 at hospital admission: An observational multi-centre cohort study. *PloS one*, 16(8), p.e0255748. 2021.

Zhang, Chi, Fanaee-Tork, Hadi and Thoresen, Magne. Feature extraction from unequal length heterogeneous EHR time series via dynamic time warping and tensor decomposition. *Data Mining and Knowledge Discovery*, 35(4), pp.1760-1784. 2021.

Zhao, Zhi; Banterle, Marco; Bottolo, Leonardo; Richardson, Sylvia; Lewin, Alexandra; Zucknick, Manuela. BayesSUR: An R package for high-dimensional multivariate Bayesian variable and covariance selection in linear regression. *Journal of Statistical Software* (ISSN 1548-7660). 100(11) pp 1-32. doi: [10.18637/jss.v100.i11](https://doi.org/10.18637/jss.v100.i11). 2021.

Reports and submitted papers

Brandsæter, A. and Glad, I.K., 2020. Explainable artificial intelligence: How subsets of the training data affect a prediction. *arXiv preprint arXiv:2012.03625*.

Brant, S.B. and Haff, I.H., 2021. The fraud loss for selecting the model complexity in fraud detection. *arXiv preprint arXiv:2101.11907*.

Eide, S., Leslie, D.S. and Frigessi, A., 2021. Dynamic Slate Recommendation with Gated Recurrent Units and Thompson Sampling. *arXiv preprint arXiv:2104.15046*.

Eliseussen, E., Fleischer, T. and Vitelli, V., 2021. Rank-based Bayesian variable selection for genome-wide transcriptomic analyses. *arXiv preprint arXiv:2107.05072*.

Engbretsen, S., Palomares, A.D.L., Rø, G., Kristoffersen, A.B., Lindstrøm, J.C., Engø-Monsen, K., Chan, L.Y.H., Dale, Ø., Midtbø, J.E., Stenerud, K.L. and Di Ruscio, F., 2021. Regional probabilistic situational awareness and forecasting of COVID-19. *medRxiv*.

Hubin, A., Frommlet, F. and Storvik, G., 2021. Reversible genetically modified mode jumping MCMC. *arXiv preprint arXiv:2110.05316*.

Jalali, N., Brustad, H.K., Frigessi, A., MacDonald, E.A., Meijerink, H., Feruglio, S.L., Nygård, K.M., Rø, G.Ø.I., Madslie, E.H. and De Blasio, B.F., 2022. Increased household transmission and immune escape of the SARS-CoV-2 Omicron variant compared to the Delta variant: evidence from Norwegian contact tracing and vaccination data. *medRxiv*.

Jullum, M., Redelmeier, A. and Aas, K., 2021. group-Shapley: Efficient prediction explanation with Shapley values for feature groups. *arXiv preprint arXiv:2106.12228*.

Köhn-Luque, A., Myklebust, E.M., Tadele, D.S., Giliberto, M., Noory, J., Harivel, E., Arsenteva, P., Mumenthaler, S.M., Schjesvold, F., Taskén, K. and Enserink, J.M., Lader, K., Frigessi, A., Foo, J. 2022. Phenotypic deconvolution in heterogeneous cancer cell populations using drug screening data. *bioRxiv*.

Lachmann, J., Storvik, G., Frommlet, F. and Hubin, A., 2022. A subsampling approach for Bayesian model selection. *arXiv preprint arXiv:2201.13198*.

Mancisidor, R.A., Kampffmeyer, M., Aas, K. and Jenssen, R., 2021. Discriminative Multimodal Learning via Conditional Priors in Generative Models. *arXiv preprint arXiv:2110.04616*.

Midtjord, A.D., De Bin, R. and Huseby, A.B., 2021. A Machine Learning Approach to Safer Airplane Landings: Predicting Runway Conditions using Weather and Flight Data. *arXiv preprint arXiv:2107.04010*.

Olsen, L.H.B., Glad, I.K., Jullum, M. and Aas, K., 2021. Using Shapley Values and Variational Autoencoders to Explain Predictive Models with Dependent Mixed Features. *arXiv preprint arXiv:2111.13507*.

Pesonen, H., Simola, U., Köhn-Luque, A., Vuollekoski, H., Lai, X., Frigessi, A., Kaski, S., Frazier, D.T., Maneesoonthorn, W., Martin, G.M. and Corander, J., 2021. ABC of the Future. *arXiv preprint arXiv:2112.12841*.

Redelmeier, A., Jullum, M., Aas, K. and Løland, A., 2021. MCCE: Monte Carlo sampling of realistic counterfactual explanations. *arXiv preprint arXiv:2111.09790*.

Salucci, C.B., Bakdi, A., Glad, I.K., Vanem, E. and De Bin, R., 2021. Multivariable Fractional Polynomials for lithium-ion batteries degradation models under dynamic conditions. *arXiv preprint arXiv:2102.08111*.

Salucci, C.B., Bakdi, A., Glad, I.K., Vanem, E. and De Bin, R., 2021. Simple statistical models and sequential deep learning for Lithium-ion batteries degradation under dynamic conditions: Fractional Polynomials vs Neural Networks. *arXiv preprint arXiv:2102.08111*.

Storvik, G., Palomares, A.D.L., Engebretsen, S., Rø, G.Ø.I., Engø-Monsen, K., Kristoffersen, A.B., de Blasio, B.F. and Frigessi, A., 2022. A sequential Monte Carlo approach to estimate a time varying reproduction number in infectious disease models: the Covid-19 case. *arXiv preprint arXiv:2201.07590*.

Tjøstheim, D., Jullum, M. and Løland, A., 2021. Statistical embedding: Beyond principal components. *arXiv preprint arXiv:2106.01858*.

Tveten, M., Glad, I.K. and Hjort, N.L., 2021. Scalable change and anomaly detection in cross-correlated data. *arXiv preprint unit.no*

Zhao, Z., Banterle, M., Lewin, A. and Zucknick, M., 2021. Structured Bayesian variable selection for multiple correlated response variables and high-dimensional predictors. *arXiv preprint arXiv:2101.05899*.

Open source published software

Banterle, M., Zhao, Z., Bottolo, L., Richardson, S., Leoncio, W., Lewin, A. and Zucknick, M., 2021. Package 'BayesSUR'. [\[CRAN\]](#)

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

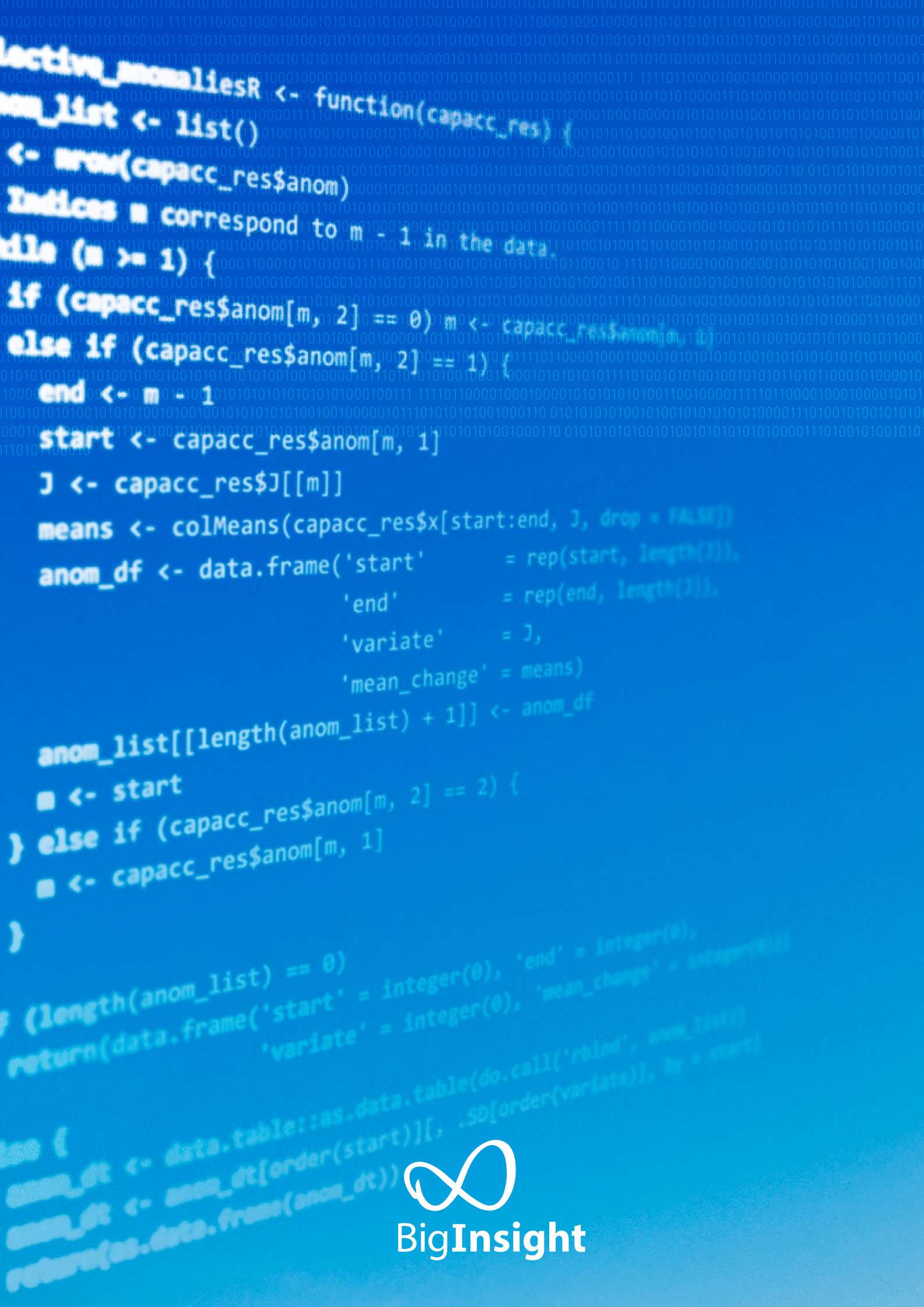
BigInsight
Oslo Center for Biostatistics and Epidemiology (OCBE)
Oslo University Hospital
Klaus Torgårdsvei 3
Sogn Arena, 2nd floor
0372 Oslo

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



```
detect_anomaliesR <- function(capacc_res) {  
  anom_list <- list()  
  <- mrow(capacc_res$anom)  
  Indices m correspond to m - 1 in the data.  
  while (m >= 1) {  
    if (capacc_res$anom[m, 2] == 0) m <- capacc_res$anom[m, 1]  
    else if (capacc_res$anom[m, 2] == 1) {  
      end <- m - 1  
      start <- capacc_res$anom[m, 1]  
      J <- capacc_res$J[[m]]  
      means <- colMeans(capacc_res$x[start:end, J, drop = FALSE])  
      anom_df <- data.frame('start' = rep(start, length(J)),  
                           'end' = rep(end, length(J)),  
                           'variate' = J,  
                           'mean_change' = means)  
      anom_list[[length(anom_list) + 1]] <- anom_df  
      m <- start  
    } else if (capacc_res$anom[m, 2] == 2) {  
      m <- capacc_res$anom[m, 1]  
    }  
  }  
  if (length(anom_list) == 0)  
    return(data.frame('start' = integer(0), 'end' = integer(0),  
                      'variate' = integer(0), 'mean_change' = integer(0)))  
  else (  
    anom_dt <- data.table::as.data.table(do.call('rbind', anom_list))  
    anom_dt <- anom_dt[order(start)][, .SD[order(variate)], by = start]  
    return(as.data.frame(anom_dt))  
  )  
}
```

