Data Science and Computing

From Morten Hjorth-Jensen

Oct 23, 2019

Vision

The new center should seek to transform the University of Oslo (UiO) to become a leader nationally and internationally in scientific discovery through large-scale computations and data-driven research. The aim is to position UiO as a leader in computational and data sciences by recruiting faculty whose expertise pertains to large-scale computing and mathematical foundations of data science - both generalists (algorithm/tool developers) and specialists (focused on specific disciplines).

Scientific computing and data science play a central role in scientific investigations and is central to innovation in most domains of our lives. It underpins the majority of today's technological, economic and societal feats. We have entered an era in which huge amounts of data offer enormous opportunities, but only to those who are able to harness them. By 2020, it is also expected that a large fraction of jobs in the STEM (Science, Technology, Engineering and Mathematics) fields will be in computing (Association for Computing Machinery, 2013).

Furthermore, the 3rd Industrial Revolution will alter significantly the demands on the workforce. To adapt a highly-qualified workforce to coming challenges requires strong fundamental bases in STEM fields. Computational Science can provide such bases at all stages. Most of our students at both the undergraduate and the graduate level are unprepared to use computational modeling, data science, and high performance computing – skills valued by a very broad range of employers. The new center will also focus on the development of courses in computational science and data science tailored to the needs of the rest of society, both for the private and the public sector.

These developments, needs and future challenges, as well as the developments which are now taking place within quantum computing, quantum information and data driven discoveries (data analysis and machine learning) will play an essential role in shaping future technological developments. Most of these developments require true cross-disciplinary approaches, approaches which normally cannot be accomplished within the realms of one single disciplinary-based department.

An older (Feb 2018) and longer version of this document is at https://computationalscienceuio.github.io/CCAD/doc/pub/whitepaper/html/whitepaper-bs.html.

Societal impact

The center should function as a hub for and coordinator of courses and study programs in data science and computing across disciplines. It should

1. Participate in the development of a comprehensive set of courses and degree programs at both undergraduate and graduate levels that will give students across the university exposure to practical computational methods, understanding how to analyse data and more generally to the idea of computers as problem-solving tools.

2. Facilitate the adoption of computational tools and techniques for both research and education across campus, through education and faculty collaboration.

3. Develop an all university PhD program in Computational Science and Data Science.

4. Develop courses and course modules in Computational Science and Data Science for the private and the public sectors.

5. Develop a new Bachelor of Science program in Data Science and Computing

6. Develop a Master of Science program and a PhD program in Computational Science and Data Science tailored to the needs of the private and the public sectors, allowing for students residing outside UiO to develop their knowledge about Computational Science and Data Science.

7. Be a driving force in the education of the next generation of scientists and teachers, with a strong focus on digital competences.

The new center will be the first in Norway to comprehensively treat computation as the triple point of algorithm development and analysis, high performance computing, and disciplinary knowledge with applications to scientific and engineering modeling and data science. There is no such center in Norway, The above paradigm shift recognizes computation as a new discipline rather than decomposed into isolated sub-disciplines, enabling application-driven computational modeling and data-driven discoveries, while also exposing disciplinary computationalists to advanced tools and techniques, which will ignite new transformational connections in research and education. This research nexus also gives rise to the educational opportunities driven by similar synergy, leveraging common resources among disciplines, and enabling joint programs and unique degrees across the entire computational space.

Why should we focus on developing a center in Data Science and Computing?

Modern problems in science and engineering bridge a vast range of temporal and spatial scales and include a wide variety of physical processes. The analysis of such problems is not possible, so one must turn to computation. To develop computational tools for such complex systems that give physically meaningful insights requires a deep understanding of approximation theory, high performance computing, and domain specific knowledge of the area one is modeling. National laboratories like SIMULA research lab have addressed the interdisciplinary nature of computing by having experts in numerical algorithms co-located with disciplinary experts who have a deep understanding of computation, and who use scientific computing to address key topics in science.

The proposed organization with algorithmic scientists and disciplinary scientists in STEM fields as well as other fields is what facilitates the exploration of challenging multi-disciplinary and interdisciplinary topics that could not otherwise be addressed. This key observation motivates the model for the proposed center - a place where we will attack the critical problems facing us in the 21st century, problem which require the development of computing skills across disciplines, from the traditional STEM fields to the Humanities, Law, Educational Science and the Social Sciences. Furthermore, this center would strive to use computing as a critical tool to explore fundamental scientific questions in subjects as diverse as the physics of specific materials, evolutionary biology and data-driven economic forecasting. In addition, the synergy of data-driven computational modeling, combining aspects of traditional scientific computing with data science and data mining, is an exciting topic that this new unit will be uniquely suited to address. This is a rapidly emerging field that touches many of the STEM disciplines but also Medicine, Education, the Humanities and the Social Sciences, and attracting world-leading talent in this area can be greatly facilitated by the introduction of the new center. Furthermore, the development of the center has the potential to catapult UiO into the position of being a leader in this critical new field, and will open doors to new scientific challenges as well as new Center-level funding opportunities.

The new center will enable new science through these unique interdisciplinary collaborations and will become a focal point for data science and computational science research at UiO, bringing researchers in computational and data sciences together with domain experts in astrophysics, bioinformatics, chemistry, geoscience neuroscience, subatomic physics, materials science, life science, the Humanities, economy, Education and many more.

Strengths, Possibilities and Synergies. The University of Oslo has within several of the STEM fields strong research and educational activities, exemplified through for example:

- Several Centers of excellence in research where Computational Science plays a major role

- A newly established center of excellence in education research
- Newly established Master of Science programs in Computational Science and Data Science
- Several excellent groups in STEM fields that do Computational Science and Data Science
- Computational topics are included in all undergraduate STEM programs, with the possibility to develop a bachelor program in Computational Science and Data Science for all university colleges
- Several educational prizes and awards related to computational science
- Strong links with research laboratories like SIMULA research lab
- UiO has the potential to develop cross-college educational programs in Computational Science and Data Science, from undergraduate programs to PhD programs that serve also the public and the private sectors
- The courses to be developed can be offered to train employees and students outside UiO, serving thus the coming needs of for example Machine Learning for the public and the private sectors

With the new center we have the possibility to really position UiO as the leading Norwegian and perhaps European institution within Computational Science and Data Science.

New and transformational science that can be enabled by the new center

Here are four major overarching themes which encompass many of the present and scientific directions of interest in data science and computing. Note that these overarching themes represent research interests of several of our MN-fak departments. There are also several links between all these main overarching topics.

- **Data Science and Statistical analysis**, from algorithm development to applications (essentialll departments as well as the humanities, social sciences, law, medicine etc). This includes Machine Learning. Should we also include Artificial Intelligence in some way?
- **Quantum information theory and quantum computing technologies** (Mathematics, Physics, Chemistry, Informatics, Materials Science), herein also traditional methods for quantum mechanical studies. It would bring in essential transformational technologies to UiO.
- **Computational modelling of high-dimensional, nonlinear and complex systems**. This overarching theme covers topics that span all departments, from math studies of say partial differential equations to applications in mechanics, meteorology, hydrology, astrophysics, life science etc etc.

- **High-Performance computing**, from algorithms to applications.

What follows here is a selected list of possible fields of applications across departments and colleges.

Computational life science. The Life Sciences is transforming with the explosion of High-throughput data generation technologies and the need for integrating these across all the levels of the biological hierarchy. A 'system-dynamic' (Systems Biology) approach will dominate research in the coming decades. Here, computational modelling to integrate the various data types and data sets will be a driving force. Similar developments are seen in the field of molecular image analysis, where computational methods to integrate the image streams will become essential to make sense of the growing amount of data. Translational Bioinformatics, bridging the gap between the laboratory, computer and the clinic, with the ultimate goal of personalised medicine, is an important, and exciting new dimension. Many of these developments require cross-disciplinary thinking, and often breakthroughs in Bioinformatics/Computational Life Science start with creatively adjusting and implementing algorithmic or computational solutions originally developed for other fields. A center that has multidisciplinarity as its founding principle, and brings computationally skilled researchers from many fields together, will provide a solid foundation for researchers working towards these developments.

Develop data-driven discovery research programs utilizing recent developments in machine learning. **Machine Learning** plays nowadays a central role in the analysis of large data sets in order to extract information about complicated correlations. This information is often difficult to obtain with traditional methods. For example, there are about one trillion web pages; more than one hour of video is uploaded to YouTube every second, amounting to 10 years of content every day; the genomes of 1000s of people, each of which has a length of 3.0×10^9 base pairs, have been sequenced by various labs and so on. This deluge of data calls for automated methods of data analysis, which is exactly what machine learning provides. Developing activities in these frontier computational technologies is thus of strategic importance for our capability to address future science problems. The applicability of big data, data-driven discoveries, data-driven modeling and machine learning covers basically all disciplines and fields, with applications spanning from materials science, mechanics, medicine, applied mathematics, economic forecasting etc. Machine learning and big data concepts are being exploited in more and more fields. The big data challenge will be in the forefront of biology and life science research in the next few years. In materials science machine learning allows us to parametrize results from quantum mechanical calculations in terms of classical interactions. These interactions are in turn suitable for large scale molecular dynamics simulations of complicated systems spanning from subatomic physics to materials science and life science. To develop a multiscale science program starting with the smallest constituents and moving to larger systems can most likely only be done with the development and application of machine learning algorithms. Economists and policy makers need up-to-date information on the state of the economy to formulate effective policies. Variables such as GDP, Gini factors, unemployment rates, quality of life data etc are normally used as key indicators. These data are often only available with delays between collection and availiability to analysts, making it thus difficult to asses properly their relevance. Machine learning algorithms have the potential to deliver improved predictions as well as correlations and proper error estimates. The examples discussed here represent just a few of the possible applications of Machine Learning algorithms that the new center can aid in developing. To develop these research lines will be achieved most effectively within the new multidisciplinary center.

Develop research programs in Quantum Computing and Quantum Information theory. Enabling simulations of large-scale quantal many-particle systems is a long-standing problem in scientific computing. Quantum many-particle interactions define the structure of the universe, from nucleons and nuclei, to atoms, molecules, and even stars. Since the discovery of quantum mechanics, a lot of progress has been made in understanding the dynamics of certain many-particle systems. While some of our insight comes from a small set of analytically solvable models, numerical simulations have become a mainstay in our understanding of many-particle dynamics. The progress in numerical simulations has accelerated in the last few decades with the advent of modern high performance computing (HPC) and clever developments in classical simulation algorithms such as, quantum Monte Carlo,large-scale diagonalization approaches, Coupled-Cluster theory and other renormalization schemes. Despite the monumental advances, classical simulation techniques are reaching fundamental limits in terms of the size of the quantum systems that can be processed. Fortunately, the disruptive new field of quantum simulations has emerged, promising to enable simulations far beyond those which are classically tractable. In particular, scientific applications concerned with simulations of interacting fermions on a lattice are poised to reap the benefits of quantum simulations. Mathematical models of interacting fermions naturally extend to describe vastly different physics such as that of correlated electronic and the correlated nuclear systems.

Recent progress in quantum computing as well as digital and analog Quantum Algorithms (QAs) promise to enable the exciting possibility of performing simulations that are beyond the reach of all existing and future classical supercomputers. Despite the progress, there is still a gap between the resources required by state-of-the-art QA and the resources offered by available and near-future quantum hardware. It may take decades of quantum hardware development and engineering before the current QAs will outperform classical exascale class simulations. Therefore, to impact scientific computing on a more relevant time scale, improving the scalability and efficiency of quantum simulation algorithms is of the highest importance. Developments in quantum information algorithms and their mathematical properties, as well as their applications will play a critical role in studies of relevance for a wide variety of fields, from the design and studies of new materials to our basic understanding of systems of interest in chemistry and physics. The new center, in close collaboration with disciplinary experts, can play an essential role in developing this field by hiring world-leading experts in quantum information theory and quantum computing.

Computational Social Science. Survey data, the engine of the behavioral revolution of the social sciences is about to run its course, with low response rate and poorly representative samples being the norm rather than the exception. Fortunately, vast amount of new information from social media, via digitalized governmental archives, to population registries are opening up new exiting avenues for innovative social science research, such as paternity leave and children's performance in school, extent of censorship in Chinese online new reporting, or conditions for receptiveness to fake news. Moreover, the new data availability in combination with tools from machine-learning has spurred an interest in prediction and sophisticated policy-recommendations, ranging from optimize relocation of immigrants given their skill-set and local labor market needs, via probabilistic detection of election fraud, to forecasting of popular unrest and civil war. The undertaking of such research questions was, until recently, outside the realm of social science. There are however limits to the amount of new insights that can be obtained purely from richer data and "black-box" import of machine-learning tools. More robust, new insights require similar steps to be taken in the development of applied, testable, theoretical models to facilitate direct empirical evaluations of the model dynamics and the consistency of the model with the data. Such a step requires a solid grounding in computing.

Computational Geoscience. Geoscience has long been a computationally-intensive area. A typical climate simulation, used for example in the future projections discussed by the Intergovernmental Panel on Climate Change (IPCC), can generate a petabyte (1 million gigabytes) of data. Weather forecasts involve suites of complex simulations, which are then averaged to assess the probability of different scenarios. These models simulate not only the atmosphere, but the important interactions with the ocean, land and vegetation. Sophisticated models are also used for studying tectonic continental shifts, to understand the geology and climate of previous epochs, thereby informing our understanding of prehistoric life. And similar models are used to simulate hydrological reservoirs and the melting occurring at the base of major glaciers. Computation is so central to the geosciences that it is impossible to imagine the study without it.

The computational approaches relevant for geoscience can be grouped in two classes: simulation and analysis. Geoscientific computation demands advance programming techniques and optimized simulations, to ensure the fast calculations. Changes in the global ocean circulation can take tens of thousands of years, demanding the most rapid simulations possible. High performance computing approaches, for example using graphical processing units (GPUs), are now being applied to climate models, greatly increasing performance. The large amount of data generated by geophysical simulations is also a challenge and is well-suited for big data techniques. Machine learning is beginning to be used in weather forecasting and in climate simulations. This has led to the identification of weather patterns missed by researchers and to the identification of extreme events like cyclones and “atmospheric rivers”, on par with that of human analysts. Computational geoscience is an exciting and developing field, and one which will make major inroads to the earth sciences in the future.

Computational Psychology. Large files of audio/video are currently unused since data is in a form that is unavailable for quantitative analysis (such as video of weekly clinical interviews from multi-center trials of treatment for thousands of patients). Analysis of prosody can shed light on change processes, and should automatic transcription reach a sufficiently good level, this will, in combination with natural language processing, open up many interesting research questions.

Accumulated data from online use already provides measurements of quantities such as personality, attitudes, skills or mental disorders which in many cases have proven to approach the level of the best instruments we have. Here one obtain much more, especially since clinical treatment will increasingly be supplemented by electronic registrations in the future, as well as being able to disconnect data from sensors in smart devices. Present instruments in use generate relatively large amounts of data (from for example EEG, ERP, and fMRI), and newer methods of pattern recognition/classification can shed light on a number of research questions.

Machine learning in education research. Quantitative education research has historically been done at the micro-scale (classrooms) and the macro-scale (K-12, baccalaureate degree programs, etc.). Micro-scale research has been done using traditional correlational statistics with data gathered from surveys, conceptual tests, classroom observations, etc. With the advent of the *digital classroom* student behavior can now be examined in fine grain. Students access of online homework platforms, video lectures, and interactions with peers via online course forums has created new data sources for education researchers. New technologies such as computer textual analysis can pick apart student conceptual understanding of hard concepts in science and mathematics. Intelligent tutors can provide real time feedback to students as they solve problems. At the macro-scale students’ career decisions within their programs can be modeled. What courses they choose to take, who they choose to take courses from, and their comments on said courses, form new data sets which can be used to predict student decisions and provide timely feedback to students and faculty advisers. Ultimately these data sets can form a high dimensional picture of student learning painted by machine learning.