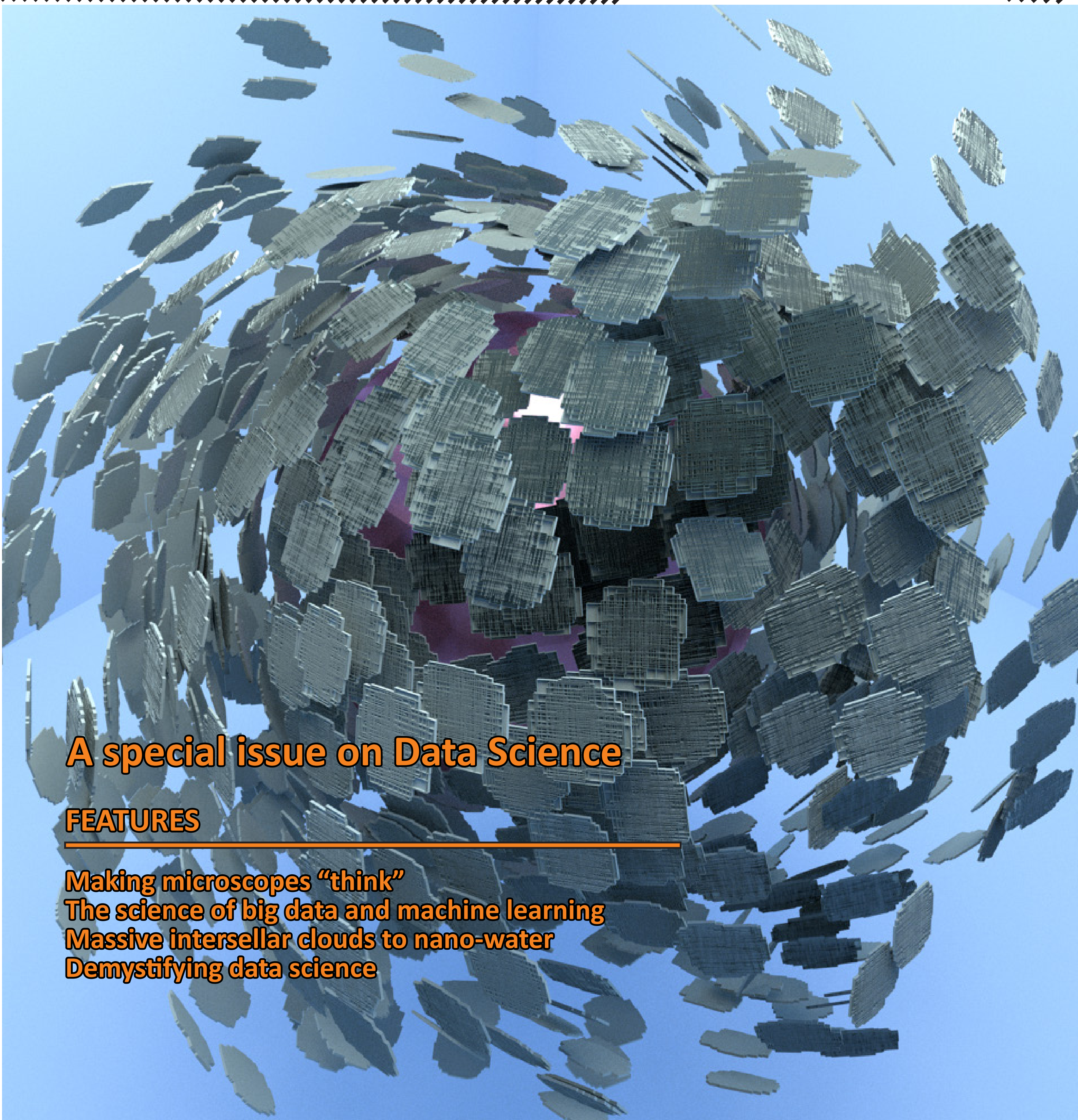
NUS National University of Singapore | Faculty of Science

Advances in Science

Volume 25 Number 1 June 2020



A special issue on Data Science

FEATURES

Making microscopes "think"
The science of big data and machine learning
Massive intersellar clouds to nano-water
Demystifying data science

Table of Contents

RESEARCH FEATURES

2 Making microscopes "think"
4 The science of big data and machine learning
6 Massive intersellar clouds to nano-water

PERSPECTIVE

8 Demystifying data science

NEWS ROUNDUP

10 A new library of atomically thin two-dimensional materials
10 New avian species discovered in little-explored islands of Wallacea

Advances in Science

The Faculty of Science conducts basic and applied experimental, theoretical and simulation research over a broad spectrum of science, mathematics and technology domains. We cover most of the key fields in biological sciences, chemistry, physics, pharmacy, food sciences, natural history, mathematics and statistics.

Advances in Science is published online twice a year. It is written for a broad scientific audience interested to keep up with some of the key areas of science pioneered by researchers at the Faculty of Science.

For further information on the research in this newsletter, please contact:

Editor: SOH Kok Hoe (scisohkh@nus.edu.sg)
Deputy Editor: Janice QUAH (janice.quah@nus.edu.sg)
Consultant: Giorgia PASTORIN (scipg@nus.edu.sg)

Dean's Office, Faculty of Science
National University of Singapore
Blk S16, Level 5, Science Drive 2
Singapore 117546

For the latest research news, please refer to:
URL: www.science.nus.edu.sg/research/research-news

On the cover: Machine learning is used to recover the three-dimensional (3D) structure of small and fragile biomolecular particles. This schematic shows how it classifies many noisy and incomplete two-dimensional measurements (gray tiles clustered by orientation), each of which comes from a single particle in some random view, into a single 3D structure. Similar methods can also classify structural differences between different particles. [Image credit: Prof Duane Loh]

Making Microscopes "think"

Building the "visual cortex" for high-resolution microscopy

Seeing is believing, and it can be a problem

Optical microscopy grew from our abilities to recognise visual features. Before photography (1826), it was the eyes, brain, and hands of the microscopist that transferred images he/she saw through the microscope's lenses onto paper. These drawings recorded the discovery of cells (1665) and bacteria (1676), which forever changed biology and medicine. Despite its power and expressiveness, such human-guided microscopy was qualitative and subjective. There are obvious downsides to relying on human vision: simple examples can prove that our visual system is terrible with quantitation (see Figure 1). Unlike a camera with long exposure imaging capability, you might not be able to see details in a dimly lit street even if you stared long and hard at it. One might blame this on the fact that our vision is not quantitative; hence our brains do not "add" disparate images that we see together to produce less noisy "long exposure" images. *Put differently, our eyes can see but our brains measure poorly.*

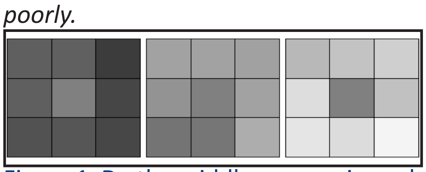


Figure 1: Do the middle squares in each 3x3 block have the same shade of gray?

From seeing to measuring to re-creating vision

Over the last millenia, "seeing" in microscopy gradually became measuring. This is tightly coupled with a strong desire to create a form of artificial visual intelligence, which was impractical until computers became affordable and prevalent. The algorithms that formed this artificial intelligence learned to automatically identify features that we sought. Soon they outdid us: they memorised differences between features on images far more precisely than we could, leading the way to complex image "arithmetic", nuanced segmentation, subtle feature detection, and statistical inference of hidden parameters.

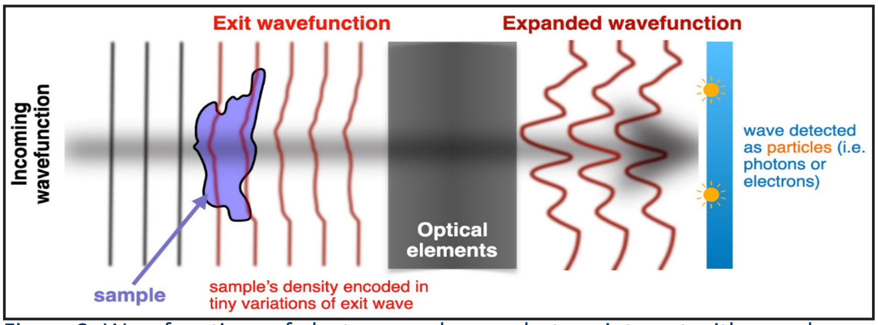


Figure 2: Wavefunctions of electrons and x-ray photons interact with samples.

In a literal sense, microscopy has become a recreation of our eyes (lenses), retina (detector), and visual cortex (fast algorithms), whose combined abilities extend beyond our biologically evolved vision. With this, we gained a unique window into the micrometre-size world. However, the further advancement into the nanometre-size world requires rethinking and reinvention.

Nanometer-scale microscopy with "just enough" x-rays and electrons

The imaging of nanometer-sized (10^{-9}m) features is far trickier than optical microscopy of micrometre-sized (10^{-6}m) ones. The physics of image formation demands imaging with probes of shorter wavelengths to resolve smaller features. For light microscopy, this means going to higher energies (e.g. x-rays) to resolve atoms. For electron microscopy, this means using high energy electrons as probes (e.g. 10-100 keV). The technology for atomic-resolution imaging of biomolecules and nanoparticles has advanced tremendously over the last century. Much of this advancement revolves around brighter and more coherent electron and x-ray sources. With these sources, we are now able to see atoms, molecules, and nanomachines, and infer how they self-organise [1]. But whether you like it or not, modern high-resolution imaging of nanometre-size objects almost always contains a trail of computational algorithm(s).

Most electrons or x-ray photons that "strike" any particular atom merely pass through it. To "know" if an atom was in the path of electrons or x-rays requires sending enough of either through the atom. Only from averaging the results of sufficient numbers/amounts of electron-atom or x-ray-atom interrogations can we determine the number and types of atoms. However, if you send too many electrons or x-rays through an atom, you will displace and/ or ionise it and its neighbouring atoms. In straightforward cases, this manifests as a loss of resolution; in trickier scenarios, you will not be able to tell if the features you see are due to the sample or induced by the beam.

Particle-wave duality: interact like a wave, detected like a particle

To understand how we can interpret electron or x-ray images, we need some basics about the image-formation mechanism. Quantum mechanics tell us that imaging electrons in a transmission electron microscope propagate from the electron gun as a wavefield down the microscope column. This wave-like description of a single electron is called a wavefunction. The description for x-ray microscopy is similar. It is like a "stadium wave" of enthusiastic sports fans.

As an electron wavefunction passes through a nanometre-size sample, the features of the sample are encoded in the subtle phase shifts that they impart onto this incident electron wavefunction (see Figure 2). In a typical transmission electron microscope, recovering these tiny spatial variations in weak phase shifts imparted on the electron wavefunction will reveal the sample's internal structure (i.e. density distribution). Mathematically, this is equivalent to recovering the complex-valued exit wavefunction. These tiny spatial variations on the exit wavefunction are then magnified, either using optics or just from diffraction in empty space, and intercepted by a detector.

The most advanced detectors only detect the arriving electron wavefunctions as particles localised somewhere in space. With a smattering of these arriving electrons, we have basically only half of the information needed to recover the exit wavefunction. More precisely, we only detect the probabilities of the wavefunction and not the critically missing phases.

Exploratory microscopes learn statistical reasoning

Here is where machine learning (ML) can help. ML models can learn prior knowledge about how general samples interact with wavefunctions, how these wavefunctions propagate through the optical elements, and the detection statistics of the electrons/ photons.

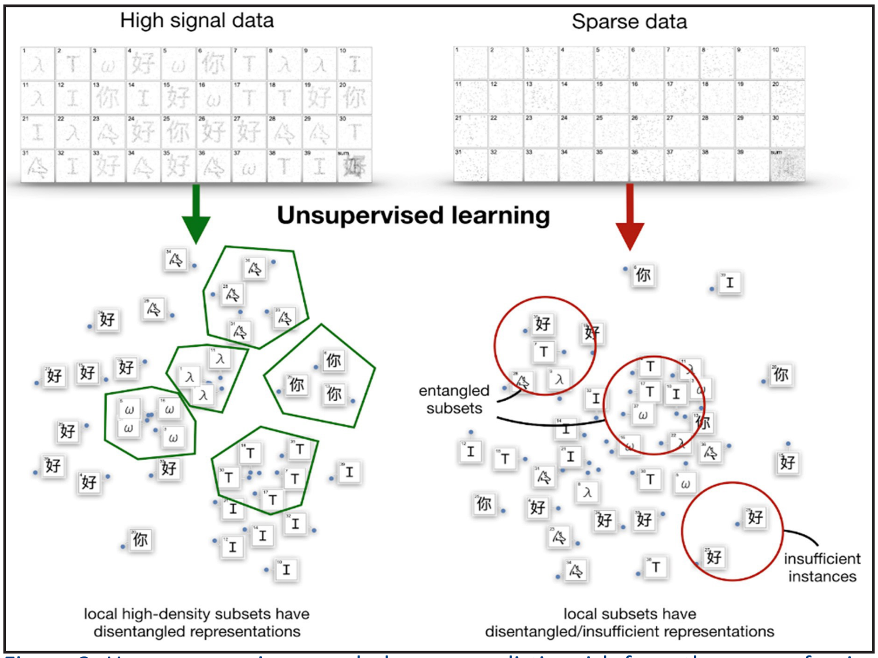


Figure 3: How many unique symbols can you distinguish from the array of noisy images (top row)? By imposing prior knowledge about the measurement process, unsupervised machine learning algorithms can help identify similar classes of noisy images. Such learning becomes more challenging as the images get noisier (left to right column), and similar symbols that used to be correctly grouped together (green polygons) now become increasingly mixed/entangled (red circles).

Using a ML framework, we can efficiently guide the search for the structures that can best explain a large body of observations on the detector. This basic idea can be extended to infer the structure and flexibility of biomolecular machines, recover ultrafast dynamics of magnetic domains, efficiently "pop-out" three-dimensional structures from two-dimensional projections, detect atomic defects that govern the efficacy of catalytic materials [2], and how nanoparticles grow from, or dissolve into solutions.

Microscopy is a powerful way to visually explore our world. Seeing and classifying objects for the first time is a potent way of understanding the mechanisms that organise and disorder our world. Exploratory microscopy, however, requires an ability to see, which is non-trivial for electron and x-ray microscopy.

This type of work ushers in a new form of probabilistic exploratory microscopy, where microscopists can expect a visual report of the structure and dynamics that are most likely present in their sample sorted by likelihood (see Figure 3). Adding domain knowledge about samples (e.g. physics, chemistry, materials science, biology, etc.) can further enhance the functions of these ML-powered microscopes. With ML, we can teach our microscopes yet another thing that humans tend to have trouble with: statistical reasoning.

Duane LOH is an Assistant Professor in the Department of Physics and the Department of Biological Sciences, NUS. He is also a Principal Investigator at the NUS Centre for Bio-imaging Sciences. His research combines statistical physics, optics, and machine learning to create novel "computational lenses" that help us see and navigate the fleeting and chaotic nanometre-size world.

References

[1] Loh ND; Sen S; Bosman M; Tan SF; Zhong J; Nijhuis CA*; Kral P*; Matsudaira P; Mirsaidov U*, " Multistep nucleation of nanocrystals in aqueous solution" NATURE CHEMISTRY Volume: 9 Issue: 1 Pages: 77-82 DOI: 10.1038/NCHEM.2618 Published: 2017.

[2] Dan J; Zhao X; Ning S; Lu J; Loh KP; Loh ND; Pennycook SJ, "A hierarchical active-learning framework for classifying structural motifs in atomic resolution microscopy" arXiv [cond-mat.mtrl-sci] 2020. Available: http://arxiv.org/abs/2005.11488.

The science of big data and machine learning

Can we trust Machine Learning to help in tackling Big Data in science?

Introduction

We live in an age deluged with data. How we extract information, understand and glean insights from ever-growing datasets is a challenging task. The volume, speed and variety of data accumulation present a game of "treasure hunting" (data mining). Data mining uses a broad array of expertise, knowledge, tools and methodologies to process Big Data to identify useful (and hidden) patterns and previously unseen connections, and to generate meaningful insights that enable data-driven decisions, sometimes within a given time frame. A higher level of investigation is to explore how Big Data can serve as a platform for new questions that we do not know we have. From a scientific perspective, such a possibility is indeed a prerequisite for novel discoveries. To achieve this, we require trustworthy and explainable tools.

The challenges of Big Data

Ploughing through the massive data jungle is complicated by the form and quality of data. With the myriad of different sources of data, it is difficult to link ("Are they the same?"), match ("Are they complementary?"), cleanse ("How do we separate the signal from the noise?") and transform ("How can we recast one dataset so that it is compatible with another dataset?") across systems. The quality (accuracy and relevance) of datasets also varies vastly, often with poorly understood and/or documented errors and uncertainties embedded within useful data. These are critical considerations when we use the datasets to make deductions and predictions.

We highlight the machine learning (ML) approach here. To put it simply, ML is a technology that exploits available computing capabilities to mine datasets. Through different "forms" (architectures) and "recipes" (algorithms), the computers can "learn" (either through hit-or-miss training or trial and error) and improve independently, without human intervention. The power of depth and efficiency of machine learning methods in data exploration is the key driver for its adoption: ML excels at processing data, extracting patterns from it in a fraction of time a human would take, and producing otherwise inaccessible insights.

Machine learning and scientific discovery

Increasingly, machine learning techniques are being used in scientific investigations. The scientific method dictates how we pursue investigations and discoveries: we formulate hypotheses based on observations and available data, build models that reflect the regularity and principles derived from past knowledge, and seek further verification (either through new experimentation and data collection, or simulation) and reproducibility by other researchers. Such extension to unexplored territories serves to shape and refine our theories of understanding.

Using machine learning in scientific investigations must consider these constraints and steps, and many, if not most, of the machine learning algorithms so far do not meet the requirements. As Judea PEARL (2012 Turing Award winner) emphasised in a keynote talk (2018):

"Current machine learning systems operate, almost exclusively, in a statistical or model-free mode, which entails severe theoretical limits on their power and performance … To achieve human level intelligence, learning machines need the guidance of a model of reality, similar to the ones used in causal inference tasks."

Another artificial intelligence (AI) researcher, Ali RAHIMI of Google, charged that machine learning algorithms have become a form of "alchemy" at an AI conference in 2018, and warned of the machine learning "culture that emphasises wins, most often demonstrating that a new method beats previous methods on a given task or benchmark … Yet, a moment of reflection recalls that the goal of science is not wins, but knowledge."

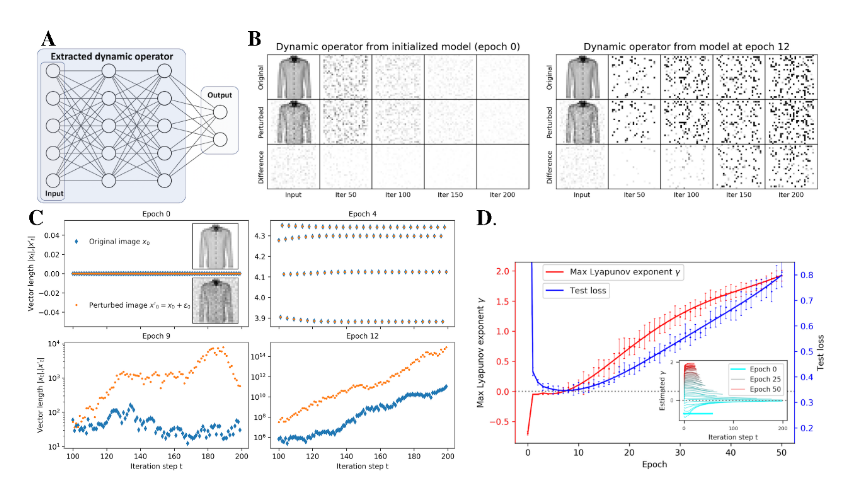
So, can we trust machine learning results? What is it that the machine is learning?

Recognising that data-driven machine learning methods are inherently data hungry, often hard to explain and generalise, it is useful to ponder what scientists have to bring to the table in this convergence of technology and science. To help shape the trend of thought, it is helpful to give a simple picture of what machine learning does. When machine learning algorithms are learning, they are actually searching in the hypothesis space defined by the choice of algorithm, architecture and configuration. (Think of the curve/surface fitting for a set of data points – the hypothesis space is the space of all possible parameter values.) This hypothesis space could be quite large even for a simple algorithm. Data is the only guide we use to look for the solution. But what if we can use our knowledge of the world, e.g. physics, together with the data to guide this search? This is the idea of physics-guided machine learning, with physical features incorporated into the machine learning algorithm, and physical consistency checks imposed in the predictions.

Physicists also hold dear symmetries and conservation laws, and they formulate fundamental principles upon which models are built and dynamics of physical systems are derived. Is/ Are there underlying principle(s) that form the basis of machine learning algorithms?

The critical brain as a guide

Artificial neural networks are often used in machine and deep learning. These are networks with multiple "neurons" in multiple layers that accept inputs and, through a chosen algorithm, make output predictions. Training data are used to adjust the interlayer link weights so as to achieve high prediction accuracy. To the extent that machine learning through these artificial neural networks mimics how human brains learn and make deductions and predictions, we can learn from developments in neuroscience and emerging understanding of how the brain network functions. Here, the premise is that the brain evolves and has its synaptic strengths between neurons adjusted for various functions. The concept of "criticality" or "edge of chaos" is being advocated as a "final state" development of the human brain. To put it simply, critical systems often exhibit optimal computational properties, and it has been suggested that criticality might have been selected evolutionarily as a useful condition of our brain system. A critical brain functions optimally, enabling signals to propagate long distances without generating excessive activity. More relevantly, a quasi-critical brain stays close to criticality but adapts to changing conditions.



Stability analysis of 2-hidden-layer fully connected neural network.

Figure 1: Some results from a preliminary study that points to optimal machine intelligence near the Edge of Chaos. Please refer to reference [2] for details.

The quasi-critical state is sometimes being referred to as the edge of chaos. Some comments about regular and chaotic behaviours of dynamical systems (systems that change with time according to some prescribed rules/ laws) may be helpful here.

Linear dynamical systems are ones in which the cause (input) and effect (output) are linearly (proportionally) related, and they have a well-defined regular behaviour: the simple (small-angle) pendulum motion and falling of an apple under the earth's gravitational pull are examples. Nonlinear dynamical systems have much more complex behaviour possibilities, depending on the system parameters. The system evolves in a well-behaved and regular manner for a certain range of parameter values; but for some particular (critical) values, the system becomes seemingly chaotic (apparently random, and hence "unpredictable") despite the fact that it is still governed by some deterministic evolution law. For such systems in the chaotic regime, small differences in the "cause" can result in dramatically different "effects".

The insight here is that regular behaviour is predictable, but chaotic behaviour magnifies small differences and can have surprisingly huge responses to external stimuli, achieving maximum information processing capability.

A research team in the Faculty of Science, NUS is investigating the unique advantage of this "edge of chaos" state in the context of artificial neural networks. In particular, we study in detail how this concept influences the training processes and performance of various neural network architecture (see Figure 1). Through these efforts, we explore novel training schemes based on this principle. Some preliminary results are encouraging, and we hope to have more definitive conclusions in the near future.

LAI Choy Heng is a Professor with the Department of Physics, NUS. He is also Deputy Director of the Centre for Quantum Technologies, NUS. He graduated with undergraduate and graduate degrees from The University of Chicago, with a Ph.D. thesis on muon pair production in neutrino interactions carried out at Fermilab (Fermi National Accelerator Laboratory.). He joined NUS after 2 years of postdoctoral research at the Niels Bohr Institute in Copenhagen.

References

[1] Hesse J*; Gross T, "Self-organized criticality as a fundamental property of neural systems" FRONTIERS IN SYSTEMS NEUROSCIENCE Volume: 8 Article Number: 166 DOI: 10.3389/fnsys.2014.00166 Published: 2014.

[2] L Feng, CH Lai, "Optimal machine intelligence near the edge of chaos" arXiv [cs.LG] 2019. Available: https://arxiv.org/abs/1909.05176.

Massive intersellar clouds to nano-water

Studying the chemistry of giant gas clouds between the stars to basic units of life using computational approaches

Interstellar chemistry

Most people will not imagine that chemistry and astronomy would cross paths. As an undergraduate, I recall my astronomy Professor sharing this when I queried him on a line from our astronomy textbook. The textbook mentioned the presence of massive clouds of gas and dust in interstellar space, which can be seen clearly in optical photos of stellar fields in our own galaxy or even in other galaxies. Molecules had been observed in these dust clouds, so it follows that chemical reactions must be taking place to form them. In some of these clouds (precursors to regions where stars and planets form), relatively large and exotic carbon, nitrogen, oxygen and hydrogen containing chemicals had been detected unambiguously with radio telescopes. Such detection was only possible due to laboratory measurements of the same molecules on Earth using microwave spectroscopy. How did these molecules get there? How were they formed? Why were some molecules seen and not others? Why is the molecular composition in some clouds quite different compared to others?

These questions can be answered by the interstellar chemist who models the thousands of chemical reactions taking place under extreme conditions in space. The models involve hundreds of exotic chemical species and predict the amount of chemicals present over 10s of thousands to 100s of thousands of years of chemical evolution within the clouds. Information on the speed of the chemical reactions obtained from experimental measurements are incorporated into the models. In cases where experimental results are lacking, theoretical predictions of their reaction rates must be made. This was my research in the early days of my career before and after joining NUS.

When two molecules meet

To accurately model individual interstellar chemical reactions from first principles and therefore predict their outcomes and reaction rates, a mathematical function called the “potential energy surface”, or PES, must be known. This function is notoriously difficult to obtain accurately even for just a few atoms. Moreover, the complexity involved is exponentially increased by the inherently large dimensionality that is far more than just three dimensions, which the surface must faithfully reproduce for a chemical reaction to be modelled correctly.

Nevertheless, we have developed methods and algorithms to deal with these issues, and have even constructed the PES “on the fly” during simulated reactions between two molecules. The application uses machine learning to construct the PES as it discovers secret and subtle, yet important, aspects of the surface hidden in the manifold of dimensionality that was completely unanticipated by its human creators (i.e. us!).

The PES is also required for accurate treatment of the way single molecules move. When measuring molecules using spectroscopic methods, light atoms or groups of atoms within the molecules can teleport through potential walls. Such quantum weirdness manifests in experimentally measured spectra that can be extremely difficult to interpret without the assistance of accurate theoretical predictions. These measurements are crucial to understand the interactions of light with gas phase matter. They also have far reaching impact not just in the realm of interstellar and intergalactic space, but also within the Earth’s atmosphere where even a subtle spectral feature can significantly alter the balance of sunlight striking, or light reflected from the Earth’s surface.

Understanding such spectra is therefore important for other seemingly unrelated areas, such as enabling telecommunication wavelengths to be free of atmospheric absorption. Developing highly accurate bound-state PES and predicting complex spectra are other areas we have explored in our theoretical and computational work at NUS.

Pushing through the size limitations of accuracy

While developing highly accurate and efficient methods for PES construction, we were constantly hampered by the limitation on how large a system could be studied computationally by accurately solving the Schrodinger equation. Systems containing more than about seven or at eight atoms could not be feasibly studied to the high level of accuracy required for our predictions. How could we push through these limitations so that much, much larger molecules could be accurately studied from first principles? We tried a few different ideas, but eventually came upon a promising approach, “divide and conquer”, which is also under development by other research groups. Hence, we turned our attention to developing our own divide and conquer method. Put simply, the “trick” involved taking a large molecule, say dozens of atoms in size, and fragmenting it into overlapping pieces, then using the “inclusion-exclusion principle”, reconstruct the larger molecular system. There are additional subtleties that need to be considered to obtain high accuracy, but this summarises the gist of our approach.

How does breaking a larger molecule apart into smaller fragment molecules push through the size limitations of accuracy?

Instead of doing one accurate calculation on a large molecule, one must now perform accurate calculations on many small fragment molecules. How does this help? The key is that the computational expense (both in computer time and resources) associated with solving the Schrodinger equation from first principles increases with high accuracy, by as much as the size of the system to the seventh power or worse! This means that doubling the size of the system of interest results in a calculation requiring about 128 times longer and vastly more resources. The problem is so severe that even with technology improving rapidly, it is still impossible to accurately solve the Schrodinger equation from first principles for large molecules like proteins with a single calculation now or in the foreseeable future. However, breaking a large molecule into many single fragment molecules enables us to harness the power of parallel computing. This problem then all but disappears – first principle calculations on systems as large as proteins now become "doable". We demonstrated the feasibility of this by performing such a calculation on the neuraminidase tetramer (a surface glycoprotein of the influenza virus) which is about 24,000 atoms in size! In addition, our fragmentation approach could be used to obtain accurate first principle energies of large molecules, as well as to predict other important properties from the calculations.

Our continued work on our fragmentation approach drew international recognition. My colleague and I were asked to write a review on the various fragmentation methods rapidly under development around the world. The review was in one of the most prestigious and highly cited chemical reviews journal in the world, *Chemical Reviews* [1], for which articles can only be submitted by invitation.

Bulk water from nano-water

While our calculation for neuraminidase may seem impressive, the original work was just a demonstration of the power of fragmentation. As it did not include water which is used as a solvent, the result was not particularly realistic. We continued to develop our fragmentation method to include water, or any solvent, around the molecules of interest by using an "implicit" model (represented as a continuous medium) for water. While this is commonly done, important effects like hydrogen bonding are not accounted for in such a treatment – which demonstrates a major limitation. This issue led us to apply fragmentation to accurately simulate "explicit" water (represented as individual molecules) as a solvent. In this approach, first principle calculations on water are performed by considering the quantum mechanical effects of individual water molecules.

Water has been, and continues to be extensively studied both theoretically and experimentally. The most commonly applied models for "explicit" water which can handle thousands or even tens of thousands of water molecules are those that involve very simple empirical treatments. Such models can be made to represent molar amounts of water molecules well, so dealing with only a few thousand, or even a few hundred water molecules is not a limitation. Their main major drawbacks are:

(a) the complete neglection of any quantum effects and

(b) "many-body" or cooperative effects are entirely missing from the models.

Many researchers have proposed "workarounds" for these issues, but such "fixes" ignore the origins of the problem due to the effects of quantum mechanics.

Our team is currently working on some of the issues using fragmentation treatment. Initially, it appeared as if cooperative effects in water were substantial and would seriously hamstring most fragmentation treatments of bulk water. However, our careful analysis of large water clusters pointed to an artifact in the quantum mechanical treatment of water that causes the apparent significant cooperative effects. We found that very large clusters of water molecules could be accurately modelled by considering, at most, up to only four water molecules at any one time. However, under most circumstances, only three water molecules need to be considered to account for the vast majority of the cooperative effects to achieve a highly accurate model of bulk water. Such trimers are roughly a nanometer across, making it look like nano-water (monomers, dimers and trimers of water). Nevertheless, whether we can achieve this goal is still a matter of research and therefore remains to be seen.

Ryan BETTENS is an Associate Professor with the Department of Chemistry, NUS. After completing a Ph.D. at Monash University in 1992, he embarked on eight years of postdoctoral research at three laboratories in the fields of Interstellar Chemistry, Millimeter- and Microwave Spectroscopy and Theoretical and Computational Chemistry. The labs were in the ETH, Zurich Switzerland, Ohio State University, U.S.A. and the ANU. In 2000 he joined the NUS as an Assistant Professor.

Reference

[1] Collins MA*; Bettens RPA*, "Energy-based molecular fragmentation methods" CHEMICAL REVIEWS Volume: 115 Issue: 12 Pages: 5607-5642 Special Issue: SI DOI: 10.1021/cr500455b Published: 2015.

PERSPECTIVE

Demystifying data science

My Data Science Journey - Yesterday, Today & Tomorrow

Introduction

Let me start off by answering some simple questions such as, what is data science, why is data science important and how we apply data science. Data Science is the study of data using a combination of skills in statistics, mathematics, computer science and programming. The aim of data science applications is to gain actionable insights and knowledge from data to support informed decision making. Data science is important because it assists businesses to better understand key factors relating to business problems. This in turn enhances operational efficiency and business competitiveness. For example, data science can mitigate risk and fraud by identifying anomalies in data patterns to create alerts to signal when unusual business activity is detected.

Having an impact and making a difference

My passion for data science took a new direction in 2002, while I was working as a biostatistician at the National Medical Research Council at the Clinical Trials Centre in Sydney, Australia. New tests and treatments are not offered to the public immediately. They need to be studied. Clinical trials are usually conducted in humans in various phases. My colleagues and I worked on cancer clinical trials and my appreciation for data analysis grew as I could see how data science contributed to supporting new treatments to benefit cancer patients.

Some of the key questions I had to consider when designing the clinical trials were, the number of treatment groups, the sample size, when to stop the trial, when to increase the sample size, etc. These were critical decisions that would impact human health and well-being, so I made every effort to ensure that my decisions were well-supported with statistical theories and rigour.

I stayed awake many nights thinking about the clinical trial design I was proposing, versus alternative possible designs. Good communication skills are required to present ideas and analytical results to stakeholders in a convincing and easy to understand manner. The next day, I would have marathon discussions with the oncologist, surgeon and my supervisor to ascertain and validate that my proposed clinical design was the right one to use.

The first clinical trial I worked on was related to breast cancer. In this study, we sought to determine whether invasive surgical procedures could be reduced while still ensuring successful treatment for the cancer. This would contribute significantly to the patient's quality of life.

Developing future data scientists

After my first position in industry as a biostatistician, I developed and took charge of multiple innovative projects in the Pharmaceutical, Healthcare, Telecommunications, and Fast-Moving Consumer Goods (FMCG) Industry. After nine years in industry, I returned to academia where I was able to combine my industry knowledge with theoretical foundations to help groom future data scientists. My first position in academia was as Chief of Business Analytics at the National University of Singapore (Institute of Systems Science) from 2011 to 2016. During this period, I designed an approved Master of Technology in Enterprise Business Analytics Degree for postgraduate working professionals who sought to combine their current skills with analytics as the demand for data science grew. I also implemented five new short executive business analytics courses for corporate professionals. These courses were designed for business executives to appreciate the value of analytical solutions in their day–to-day business operations.

In 2017, I was appointed as the Director of the Data Analytics Consulting Centre at the National University of Singapore and Associate Professor jointly in the Department of Statistics & Applied Probability and the Department of Mathematics. My position jointly in academia and consulting now enables me to bring industry professionals and data science students together to apply data science to workplace problems. I seek to develop data science courses that are practical and provide experiential learning for our data science students. To make the data science experience as real as possible for my students, I provide real-world problem solving projects, where teams of students present and communicate their solutions to real-world problems to the whole class. I invite data science industry professionals as guest lecturers, so that students can learn first-hand current business challenges and how these challenges are overcome through data science.

To add further value to industry, I also address stakeholders through conferences, run in-house conferences and provide consulting services.

Editorial leadership

One of the students I supervised is Ms TAN Shuen Lin, for her final year project "A Machine Learning Overbooking Algorithm for Enhancing Clinic Efficiency" [1]. Her paper analyses patient no-shows, which limits access to healthcare for other patients. The focus of this paper was to evaluate our proposed overbooking algorithm and compare it to the reference scenario

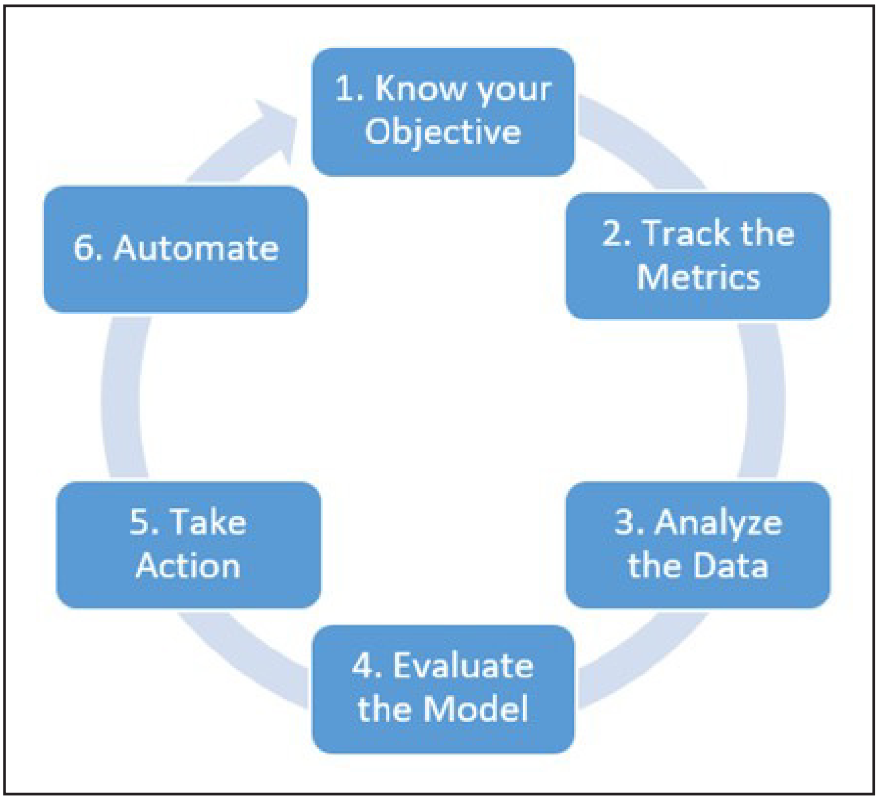


Figure 1: Six steps to building a powerful customer analytics system for your organisation. When businesses better understand their customers' purchasing behaviour and lifestyles, they can classify their customers more accurately and make targeted predictions that are meaningful in meeting customer needs.

of no-overbooking, as well as blind overbooking. The criteria used for the evaluation of the three methods was clinic efficiency and clinic profitability.

Our findings conclude that predictive overbooking brings about significant improvements in several aspects of clinic efficiency as compared to both no-overbooking and blind overbooking. We also observed a significant increase in clinic profits from the simulation of predictive overbooking compared to no-overbooking by almost 100%. Clinic profits under predictive overbooking were also greater than with blind overbooking, by 19%. This paper was published in the International Journal of Advances in Science Engineering and Technology [1].

Another area I have a strong interest in is Customer Analytics. Business success and sustainability are heavily dependent on customer satisfaction. Many businesses have lots of consumer data which is merely stored. This is a lost opportunity to draw insights and knowledge from the data for making better business decisions. In my article [2] last year, "6 Steps to Building a Powerful Customer Analytics System", I sought to provide a guide for business managers who are not data scientists through six steps for building a powerful customer analytics system (see Figure 1).

I also frequently author book chapters and publish my research. One of my book chapters, "Healthcare Analytics: A Case Study Approach using the Framingham Heart Study" [3], seeks to educate clinicians and healthcare professionals on the value of healthcare analytics. This chapter demonstrates how the analysis of health data, such as blood cholesterol, blood pressure, smoking, and obesity, can identify patients at high risk of heart attacks, and how the proactive management of patient lifestyles and use of medication can prevent heart attacks.

Data scientists also contribute to infectious disease modelling for challenging times such as the COVID-19 pandemic. My colleague, Dr WU Chengyuan and I are working on a COVID-19 Genomics Research Project. The objective of our project is to analyse and study the novel coronavirus COVID-19, as well as related coronaviruses in humans and also animals. Currently, experts are divided on where the virus comes from (e.g. bats, snakes, pangolin, etc). Knowledge of the origin of the virus can potentially help the development of vaccines/cures and/or shape future public policy. These examples illustrate how data science plays a crucial role to enable industries and businesses across many sectors to make powerful data-driven decisions based on facts, statistical numbers and trends.

Carol HARGREAVES is Director of the Data Analytics Consulting Centre at NUS and holds an Associate Professor position jointly in both the Department of Statistics & Applied Probability and the Department of Mathematics. Her research area includes artificial intelligence and machine learning applications in the healthcare and financial industry. She is also a member of the Singapore University of Social Sciences Mathematics Programme Advisory Committee.

References

[1] CA Hargreaves; Tan SL, "A machine learning overbooking algorithm for enhancing clinic efficiency", INTERNATIONAL JOURNAL OF ADVANCES IN SCIENCE ENGINEERING AND TECHNOLOGY ISSN(p): 2321 –8991, ISSN(e): 2321 –9009, Volume: 8 Issue: 1 Published: 2020.

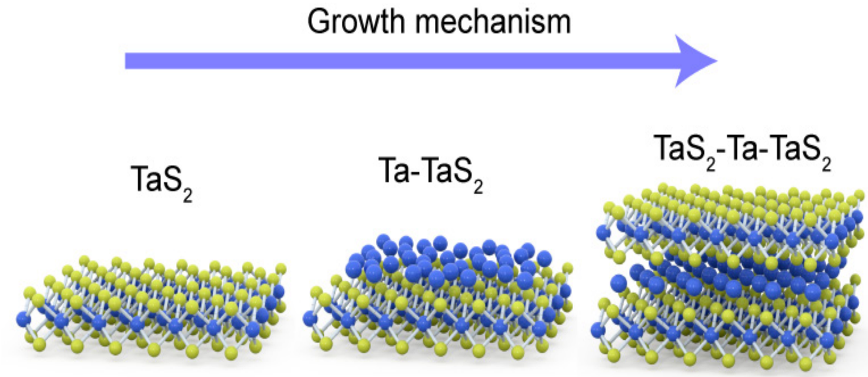
[2] https://blogs.oracle.com/datascience/6-steps-to-building-a-powerful-customer-analytics-system

[3] Data Science: Theory, Analysis and Applications. Chapter 7, "Healthcare Analytics: A Case Study Approach using the Framingham Heart Study", Carol Anne Hargreaves. Taylor and Francis. Page 159-172.

A new library of atomically thin two-dimensional materials

NUS researchers have created a whole new library of atomically thin two-dimensional (2D) materials, christened "ic-2D", to denote a class of materials based on self-intercalation of native atoms into the gap between the layers of crystals.

Atomically thin two-dimensional (2D) materials offer an excellent platform to explore a wide range of intriguing properties in confined 2D systems. However, compositional tuning of transition metal dichalcogenides to make new materials other than the standard binary or ternary compounds is challenging. In the past, theoreticians have tried to predict new properties based on combining atoms into a crystal structure where metal and chalcogen atoms sit in covalently bonded sites within the basic building block (unit cell). However, their theories did not address the situation when the same metal atom sits in between two unit cells (filling the van der Waals gap).



Researchers at NUS Chemistry, and Materials Science and Engineering have fabricated a whole new library of ic-2D materials by filling the van der Waals gap in (two-dimensional) 2D materials. Schematics showing the step-by-step growth of a typical Ta_7S_{12} ic-2D material.

Now, research teams led by Prof Kian Ping LOH from the Department of Chemistry, Faculty of Science, NUS and collaborator Prof Stephen J. PENNYCOOK from the Department of Materials Science and Engineering, Faculty of Engineering, NUS, have synthesised and characterised for the first time, an atlas of wafer-scale atomically thin ic-2D materials based on inserting the same metal atoms between the van der Waals gap of transition metal dichalcogenides. The researchers' results were published in *Nature* on 13 May 2020.

Next, the teams plan to incorporate this new library of materials into memory devices, for practical applications, and intercalate foreign atoms into the van der Waals gap and exploit novel functionalised ic-2D materials.

Reference: Zhao X, *et.al.*, "Engineering Covalently Bonded 2D Layered Materials by Self-Intercalation" NATURE DOI: 10.1038/s41586-020-2241-9.

New avian species discovered in little-explored islands of Wallacea

Birds are the best known class of animals, and since 1999, only five or six new species have been described each year on average. Recently, a joint research team from NUS and the Indonesian Institute of Sciences (LIPI) made a quantum leap in the discovery of cryptic avian diversity by uncovering five bird species and five subspecies new to science.

Visual showing the three new species found on Taliabu island (from left), the Taliabu Grasshopper-Warbler, the Taliabu Myzomela and the Taliabu Leaf-Warbler.

Photo credit: (left) Robert O. HUTCHINSON, (middle and right) James EATON/Birdtour Asia.

The team, led by Prof Frank RHEINDT from Department of Biological Sciences, NUS found the birds in three small island groups off Sulawesi, Indonesia. The islands are situated in Indonesia's Wallacea region, an archipelago at the interface between the Oriental and Australian biogeographical realms, named after Sir Alfred Wallace, who was the most famous historical collector exploring the area.

The results of the study, which were published in the journal *Science* on 10 January 2020, provide evidence that our understanding of species diversity of complex areas such as Wallacea remains incomplete even for relatively well-known groups such as birds. The findings also suggest that modern exploration to find undescribed species diversity can be targeted to areas of high promise.

Reference: Rheindt FE*, *et. al.*, "A lost world in Wallacea: description of a novel montane archipelagic avifauna" SCIENCE DOI: 10.1126/science.aax2146.