Big Data in the physical sciences: challenges and opportunities

P. Clarke[1], P. V. Coveney[2], A. F. Heavens[3], J. Jäykkä[4], B. Joachimi[5,*], A. Karastergiou[6], N. Konstantinidis[5], A. Korn[5], R. G. Mann[1], J. D. McEwen[7], S. de Ridder[8], S. Roberts[9], T. Scanlon[5], E. P. S. Shellard[4], and J. A. Yates[5]

[1]School of Physics and Astronomy, University of Edinburgh, Kings Buildings, Mayfield Road, EH9 3JZ
[2]Centre for Computational Science, Department of Chemistry, University College London, 20 Gordon Street, London, WC1H 0AJ
[3]ICIC, Department of Physics, Imperial College London, Prince Consort Road, London, SW7 2AZ
[4]DAMTP, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA
[5]Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT
[6]Oxford Astrophysics, Denys Wilkinson Building, Keble Road, OX1 3RH
[7]Mullard Space Science Laboratory, University College London, Surrey RH5 6NT
[8]School of Mathematics, University of Edinburgh, The King's Buildings, Peter Guthrie Tait Road, Edinburgh EH9 3FD
[9]Information Engineering, University of Oxford, Oxford OX1 3PJ
[*] *Corresponding author*: b.joachimi@ucl.ac.uk

ABSTRACT

We provide a brief overview of challenges and opportunities in the physical sciences arising from data which is extreme in volume, acquisition rate, and heterogeneity. These require novel methodological approaches ranging from data management to analysis techniques to inference and modelling, affording close links with cutting-edge research in data science. We highlight past methodological breakthroughs with high impact, showcase selected current data science challenges in the physical sciences disciplines, and broadly discuss needs of the community in the era of Big Data. We argue that the recently founded Alan-Turing Institute (ATI) is ideally positioned to facilitate, and intensify, the mutually beneficial cross-fertilisation between core data science developments and their application in the physical sciences. Concrete measures are proposed to achieve these goals, critical for ensuring impact and continued research leadership in both fields.

1 Introduction

Big Data is not a new phenomenon in the physical sciences. Over centuries, chemistry, geosciences, physics, and their various sub-disciplines have generated and exploited among the largest and most complex datasets available to humankind. However, Big Data in the modern sense of the word encapsulates not just the sheer *volume* of data, but also its *velocity*, i.e. the data rate, as well as its *variety*, referring to the heterogeneity of datasets.[1] The defining quality of these three '*V*'s is that they prohibit a simple scaling of existing approaches to the acquisition, management, analysis, and interpretation of data, but demand new approaches to dealing with this data. The key to these new approaches is the algorithm used, including its practical implementation on a computing architecture.

The physical sciences abound in examples of Big Data. When it comes online, the Square Kilometre Array, a radio telescope network currently under construction in Australia and South Africa, will generate more data than the global internet traffic in a given period of time. The Large Hadron Collider at CERN discovered the elusive Higgs boson in data streams of particle collisions that were produced at a rate of GigaBytes per second. Meteorologists and seismologists routinely work with global sensor networks that are very inhomogeneous with regard to their spatial distribution, as well as the type, quantity, and quality of data produced.

The defining characteristics of Big Data are variously supplemented by additional *V*-words, which highlight aspects that become particularly acute when data is Big. *Veracity* is an issue in modern datasets in the physical sciences. The ratio of signals to the noise and/or the background is often small; low-fidelity signals may only provide biased estimates of desired quantities; incomplete data complicates or hinders the extraction of signals. Yet, important decisions are based on this data, often in an automated process (e.g. the triggering of tsunami warnings).

The *viability* of Big Data is not obvious any more but needs to be established via statistical techniques. This includes

questions such as if the data is able to answer the research question posed, and which combination of data is the most predictive. Finally, while the individual datum is practically worthless, Big Data contains enormous *value* which, depending on the physical sciences application, can be monetary (e.g. in oil exploration) but may also be measured in terms of human welfare (e.g. severe weather and earthquake predictions) or the advancement of knowledge (e.g. in fundamental physics and cosmology). Particular challenges in Big Data are the mining of a maximum of useful information and exploiting the discovery potential of datasets, by identifying and capitalising on unexpected information.

The challenges of the Big Data revolution have tremendous impact in many areas of the physical sciences, fundamentally changing the approaches to knowledge gain, modifying the structure and scope of scientific collaboration, and necessitating new career paths. As many areas of academia and commerce are affected in similar ways, the foundation of the Alan Turing Institute (ATI; https://turing.ac.uk) as the UK's national institute for data science in 2015 was a timely move to address these challenges. The ATI focuses on research in the core data science disciplines of mathematics, statistics, and computer science, but its mission statement includes a commitment to enabling "researchers from industry and academia to work together to undertake research with practical applications".[2]

To help inform the scope of the ATI and identify high-impact research areas, a number of Exploiter Summits were set up with the goal of gauging the needs and interests of the 'consumers' of data science. Physical scientists from the whole UK community, and from academia and industry, were invited to a one-day Summit on January 13th, 2016, at the Royal Society in London, entitled 'Big Data in the physical sciences' (https://indico.cern.ch/event/449964). Interest in the meeting was so great that not everyone could be accommodated due to the space limitations of the venue. In the end, around 85 physical scientists reviewed and discussed the challenges and opportunities of Big Data across the subject areas.

This paper provides a summary of the outcomes of the ATI Summit on physical sciences. In Section 2 we briefly review the long-standing cross-fertilisation between data science and the physical sciences, before highlighting a select few current examples that were also presented at the Summit in Section 3. The meeting featured several discussion sessions that were aimed at providing a synoptic view of the key data science problems of the near future in the physical sciences, which are presented in Section 4. In Section 5 we conclude on the requirements for the continued UK leadership in physical sciences research in the era of Big Data, and make suggestions on the role of the ATI in this endeavour.

2 The coevolution of data science and physical sciences

There is a long history of data science in the physical sciences. Indeed, the scientific method is founded on the analysis of data to validate physical theories. A compelling example that demonstrates the importance of data in driving the development of physical theory is the discovery of Kepler's Laws. Although Kepler initially thought circular motion was an elegant description of the orbits of the planets about the Sun, he was forced to reconsider this hypothesis due to the exquisite quality — at the time — of Tycho Brahe's data, leading to the elliptical planetary orbits of Kepler's Laws, which Newton later showed follow from his law of gravitation.

The analysis of data in the physical sciences is now ubiquitous and has transitioned from a supporting role to a symbiotic relationship, coupling the physical sciences with statistics, applied mathematics and computer science. A mutually beneficial feedback loop between fields has developed, with researchers from all fields adding contributions to the origin and motivation of new analysis methodologies, their theoretical foundations, algorithms for their practical use, their application, and the development of a deep understanding of their properties. Often the progression of development is not linear or ordered, resulting in a rich and sometimes chaotic simultaneous development and refinement of ideas. For example, in some cases supporting theory is developed at the beginning of the cycle, motivating subsequent development and application, while in other cases empirical results come first and motivate the need for deeper theoretical understanding.

A prime example of the symbiotic relationship between the physical and statistical sciences is the development of Markov Chain Monte Carlo (MCMC) methods. The evolution of MCMC methods began at Los Alamos after the end of World War II. Monte Carlo methods had recently been developed by physicists to simulate the behaviour of systems of large numbers of particles, exploiting the fact that the physical laws governing these systems were inherently probabilistic (e.g. thermodynamics, statistical physics, and quantum mechanics). The arrival of the first computers meant that large probabilistic systems could be simulated quickly, accurately and flexibly, for the first time. The extension from Monte Carlo simulations, based on standard probability distributions (e.g. uniform, Gaussian), to MCMC methods capable of sampling from other complex probability distributions relied on the use of Markov Chains. This development was marked by the seminal article of Metropolis,[3] published in 1953, which presented what came to be known as the Metropolis algorithm. However, it was not until some time later, in the 1960's and 1970's, that rigorous proofs of the convergence of the Metropolis algorithm to stable distributions appeared.[4,5] The generality of the Metropolis algorithm and its importance for the statistical sciences was not fully appreciated until Hastings made the connection in 1970,[5] showing its general applicability. Nevertheless, it was not for another 20 years that MCMC methods were adopted by the statistics community, triggered by the work by Gelfan and Smith.[6] Since, MCMC methods have been studied rigorously by the statistics community, leading to many theoretical developments, e.g., in

understanding convergence rates, burn-in periods, the impact of proposal distributions, and many others properties. MCMC methods are now prevalent across the statistical and physical sciences, and beyond, with important contributions also coming from applied mathematics and computer science. Nonetheless, MCMC methods remain an active area of research, with researchers across many fields contributing to further their theoretical foundations, their practical implementation and usage, and myriad applications.

Another excellent example of the symbiotic relationship between the physical and computational sciences comes from compressive sensing and sparse regularisation in applied mathematics. In this case, for the most part, empirical results drove the development of a rigorous underlying theory, after which both further theoretical and empirical developments exploded. Sparse regularisation techniques to solve inverse problems were developed in the 1970's and 1980's in astrophysics[7] and geophysics[8] and were demonstrated to be highly effective. The CLEAN algorithm for imaging the raw Fourier measurements acquired by radio interferometry telescopes, and its evolutions, still find use in radio interferometry today. In the 1990's, interest in sparsity, and ℓ_1 minimisation in particular, grew in the signal processing community (for finding sparse approximations of signals in overcomplete dictionaries[9,10]) and in the statistics community (for variable selection in regression, yielding the so-named Lasso approach[11]). While related theoretical developments can be traced back to Prony in 1795, who proposed a method for estimating the parameters of a signal comprised of a small number of complex exponentials, modern research on the theory of sparsity took shape in the 1990's and early 2000's, with many important contributions. However, it was not until the seminal works of Candès et al.[12] and Donoho[13] in the mid 2000's that the general theoretical underpinnings of compressive sensing were laid out, showing that a signal with a sparse representation can be recovered *exactly* from a small number of (incoherent) measurements. These theoretical foundations were motivated by prior empirical observations showing that in some cases signals could be recovered exactly from a small number of measurements by ℓ_1 minimisation. Following this foundational work, developments in both the theory of compressive sensing and also its practical application sky-rocketed and both remain active areas of research.

Recent work at the forefront of statistical data analysis has also had impact in the physical sciences, particularly in the particle and astrophysics communities, in which event discovery within the data is of core importance. Such approaches lie, for example, at the core of the detection of pulsars[14] and exoplanets.[15] In the latter examples not only have Bayesian approaches to inference been widely adopted, but recently non-parametric models have been widely used, not only for detection[16] but also for ascertaining and removing underlying (and unknown) numbers of systematic corruptions and artefacts[17,18] as well as more mainstream regression and classification methods, such as for the photometric redshift estimation requirements of the European Space Agency Euclid mission (`http://sci.esa.int/euclid`).[19] Indeed, there is a significant body of literature looking at whether techniques such as deep neural networks can be of value in the physical sciences, as they have proven to be in such areas as speech and language understanding.[20] Although this technology is as yet not fully understood, it is possible it will play a part in extracting physical insight from data in years to come.

These illustrative historical examples[1] highlight a deep connection between the physical sciences and the field known today as data science, which draws heavily on statistics, mathematics and computer science. A symbiotic relationship exists between data and physical sciences, with each field offering both theoretical developments and practical applications from which the other can benefit, which typically evolve through an interactive feedback loop. With the forthcoming emergence of larger and more complex datasets in the physical sciences, this symbiotic relationship is set to grow considerably in the near future.

3 Showcases

3.1 Current highlight from astrophysics

The basic principles of observational astrophysics dictate that weaker sources require larger photon collecting areas, where as the observation of detail in the sky requires a telescope of large diameter. Interferometry provides the technique to construct a telescope of enormous diameter (even on scales of thousands of kilometres), using a large number of connected small telescopes. When constructing a new instrument, this allows the flexibility to balance total aperture size, angular resolution, and cost, as dictated by the specific science objectives and available funds. One interesting aspect of interferometric telescopes, which have proven extremely successful in radio astronomy (astronomy conducted at wavelengths between millimetres and tens of metres), is that the data from a large number of receiving elements need to be combined to produce the desired data products, which are then analysed to address specific science problems. The Square Kilometre Array (SKA, see Fig. 1, `https://www.skatelescope.org/`) is the next generation radio interferometer that will produce ground-breaking results in several fields of astrophysics, using very large collecting areas made up of large numbers of individual elements. Specifically, the two telescopes comprising the first phase of the SKA, will be made up of around 200 parabolic dishes sampling frequencies above 350 MHz (mid frequency array), and approximately 130 000 dipole antenna elements, sampling the

[1] The brief reviews of historical developments presented here are far from complete but simply highlight some noteworthy contributions. Many researchers have made significant contributions to these fields, which it is unfortunately not possible to cover in detail here.

Figure 1. Artist's impression of the layout of parabolic antennas comprising the Square Kilometre Array. Credit: Swinburne Astronomy Productions, Swinburne University of Technology.

frequencies below 350 MHz (low frequency array). At each element of the interferometer, the data will be digitized at rates of order 1 Gsamples/s, generating broadband data streams that are combined digitally downstream, in two different ways. The first uses a device known as a correlator, which generates images of the radio sky. The second uses a device called a beamformer, to generate high time resolution time series of signals originating from individual and very localized positions in the sky.

The SKA will use beamformer technology to simultaneously generate 500 beams using the low frequency array, and 1500 beams using the mid frequency array. Each of these beams will carry information at a rate of approximately 2.5 Mbits/s, leading to a total data bandwidth of 1.3 Tbits/s and 4 Tbits/s for the low and mid frequency arrays. The primary purpose for generating these data streams is to search for pulsars and fast radio transients. Storing the data and processing offline is both prohibitively expensive and restrictive for the science; there is no opportunity for a rapid reaction to unique events detected in the data, which are often the most astrophysically interesting. Therefore the processing that is required to detect the interesting signals must be carried out in real time. As a consequence, the necessary algorithms, including specific filters for the correction of dispersion effects caused by interstellar propagation, and algorithms that search for periodic and quasi-periodic signals embedded in noise, require extreme optimization. A carefully designed combination of high performance computing hardware and optimized algorithms, will allow us to perform this unprecedented search in real time and within a tight power budget.

One of the main motivations for this enormous computing effort lies within the amazing properties of pulsars, and their capacity to continuously provide us with new insight into extreme processes and fundamental physics.[21] Pulsars are neutron stars, born in supernova explosions of massive stars. Their physical properties are extreme; they contain more than a solar mass within a sphere of 20 km, at densities that exceed nuclear density in their interiors, and with surface magnetic fields far greater than what can be generated in a laboratory (10^9 – 10^{15} Gauss). They rotate around their spin axis at periods of milliseconds to seconds, and maintain extremely stable rotational properties over years. For this reason, the radio pulses from pulsars can be used as ticks from extremely stable clocks, probing fundamental properties of space-time. The stability of pulsars suggests that we can use lines of sight towards multiple pulsars as arms of a Galactic Gravitational Wave detector, sensitive to (nHz) frequencies, many orders of magnitude lower than the recent discoveries of LIGO. In addition to stringent tests of General Relativity, pulsars allow us to study the properties of dense nuclear matter, high energy plasmas, and supernova explosions, as well as the structure of the intervening magneto-ionized interstellar medium. Pulsars are a treasure-trove for many areas of research in physics, and the SKA searches will reveal to us the majority of the Galactic pulsar population, feeding ground-breaking discoveries in the coming decades.

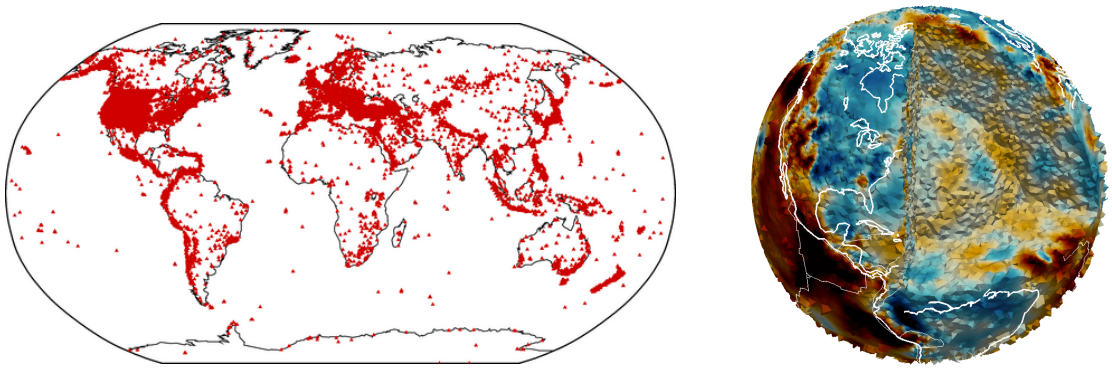

Figure 2. *Left*: Map of 22 000 seismograph station deployments since the 1960s (International Seismological Centre). Currently, 5 000 stations deliver data in real-time. *Right*: Model of seismic velocities in the Earth's interior (courtesy Kasra Hosseini).

3.2 Current highlight from Earth sciences

Big Data in the Earth sciences Big Data in the Earth sciences are characterized by the large variety of observables of Earth processes. For example, satellite observations, weather and buoy stations, permanently deployed GPS stations, tilt meters, and seismograph stations, to name just a few. These data help to address a wide range of the atmospheric and oceanographic sciences and geoscientific issues and questions, over time-scales as short as minutes (for example earthquakes) to and time-varying processes over human time scales, that are directly relevant to the existence of humanity for example global warming. Large simulations in model spaces with billions of parameters are employed to explain these data and unravel the governing physical systems and their properties.

Geographical Information systems (GIS) have transformed the Earth sciences and enabled geospatial analysis on big data. However, GIS relies on scientists to interpret and classify what information to input and map. Earth scientists employ a variety of algorithms to mine and merge datasets, and prepare input to GIS. However, due to the rapid increase of the amount of data available in the Earth sciences, the algorithms to interpret and classify information lag behind in efficiency and effectiveness. Moreover, often case they fall short of taking full advantage of the ability to analyse Earth scientific data on a global scale. One sub-field of the Earth sciences where algorithmic research takes a key role in data-mining and interpretation, is seismology. Earth scientists cannot observe the Earth's interior directly, all observations are made indirectly and the dominant way in which we learn about the Earth's interior is through seismic waves.

Algorithmic challenges and opportunities in seismology Seismology concerns the study of vibrations of the solid Earth. These vibrations propagate with seismic velocity, and variations in this velocity teach us about core and mantle composition, plate tectonics, pressures, temperatures, rock type, melts, fractures, faults, fluids in pore spaces, resources, underground storage, etc. The best-known sources of such seismic vibrations are earthquakes. But there are many more sources, including wind, traffic, volcanoes, (nuclear) explosions, and landslides. The data are used to study a wide range of phenomena, and are critical to a wide spectrum of Earth and environmental sciences. The primary use of seismic vibrations is to offer an opportunity to image the Earth's interior. The industry acquires seismic data with up to 1 000 000 sensors distributed extremely densely on a small patch of the Earth's surface, to support exploration and production of underground (usually hydrocarbon) resources.

Seismology on a global scale faces the challange of keeping big data with it's metadata, through a complicated workflow. During the last half-century more than 22 000 stations have been temporarily or permanently deployed (Fig. 2, left). Predominantly, these are deployed in Western countries (Europe and North America), and cover the other continents and oceans more sparsely. Although all stations record seismic vibrations, there are of a variety of instrument types that record at different frequencies under different site conditions. The collected data is interpreted for the properties of the Earth's interior. One method for interpretation is by inversion, which relies on large simulations of wave propagation. This enables us to form a map of seismic velocities in the Earth's interior (Fig. 2, right). Seismic velocities near the Earth's surface are known to change over time. Shedding light on time-variant subsurface processes, such as active volcanoes,[22] plate tectonics,[23] hydrocarbon production,[24] and water saturation.[25]

The advancement of dense networks of seismic stations with data available in real-time motivates research into algorithms that extract information on temporal variations that occur in the subsurface of the Earth. Since a few years, a rapidly growing number of seismic stations supply data in real-time. Currently over 5 000 such seismic stations are distributed around the globe. Software for real-time seismology must combine aspects of data mining, pattern recognition, and map reduction, to deal with

heterogeneous data efficiently, veraciously, and automatically. Elements of machine learning play a key role in overcoming these challanges. Because seismic data is used to constrain Earth processes that are poorly understood, it is a challenge to process the data without knowing exactly what signals to expect. Furthermore, even though we employ sophisticated systems of partial differential equations to explain the data, much of our data does not fit our simplified models – for example, complex waveforms of multiply-scattered long seismic waveforms, continuous background seismic noise, and processes at finer scales than we can model due to computational cost.

Seismological research benefits from close co-operation with industry who have permanently deployed dense seismic station arrays. The hydrocarbon industry faces similar challenges as the academic community, but their seismic data is more homogeneous in nature (although much larger in quantity).

The road ahead Although in the Earth sciences we are able to acquire vast quantities of data, only a small portion of the data is actually explained by sophisticated systems of partial differential equations. Large portions of the data remain unexplained and unused. To harness such 'unexplained' parts of the data we need data mining methods to extract patterns and reveal systematics from the data that would otherwise go unnoticed.

3.3 Current highlight from particle physics

The particle physics community has a long history at the forefront of big data analysis. This is not surprising considering the amount of data collected by particle physics experiments. For instance, in the Large Hadron Collider (LHC), particle collisions occur 40 million times per second. In each event (an event is the crossing of two bunches containing 10^{11} protons in the middle of one of the LHC experiments), there will be on average around 20 protons colliding and up to 160 million channels to read out, creating PB/s data rates. This data rate has to be reduced by a factor 10^6 within a few microseconds before it can be recorded. Several billion events will typically be recorded each year and many analyses need to identify only a few tens of interesting events. For example, in the recent discovery of the Higgs Boson the signal in the $H \to ZZ \to \ell\ell$ channel shown in Fig. 3 consisted of 13 events[26] from the billions recorded. To identify such events requires the particles from the proton-proton collisions to be precisely and accurately identified in a very complex environment, with the events of interest then selected from an overwhelming background. All these challenges have required cutting edge tools to be adapted and developed over the past few decades.

To store and process these large amounts of data, particle physics pioneered the development of new computing platforms. One such example is the deployment of the Grid,[28] now known as Cloud computing, which provides almost real-time access to the data to about 10 000 physicists from around the world. Novel approaches to increase the amount of processing power available, such as ATLAS@HOME,[29] which makes use of volunteers' idle computer time, are also being exploited. New machine architectures, like GPUs, are being explored. Lastly, an overused, yet illustrative, example is that the world wide web was developed at CERN for information exchange.

The high energy physics community has constantly developed advanced and intelligent methods for data analysis, pattern recognition, and model inference, thus regularly engaging with the field of data science, ranging from the development of advanced statistical inference[30] to machine learning techniques. In order to test and optimise such techniques, sophisticated event simulation techniques are also employed, including Monte Carlo techniques to not only simulate the particle collisions and underlying interactions, but also the passing of the particles through the active detectors and the detector response.

Multivariate and machine learning techniques are widely deployed in most stages of analysing the data, to identify the particles in the detector[31] and to achieve an efficient separation between background and interesting signal events.[27] The most widely used techniques currently include Artificial Neural Networks and Boosted Decision Trees. More advanced algorithms, such as machine learning techniques based on deep learning architectures are also being studied and ATLAS is collaborating with world experts to maximise the impact from such tools.[32] Such techniques have also been studied in the context of using detector images, rather than physically motivated feature driven approaches, to boost the performance.[33] A freely available software Toolkit for Multivariate Data Analysis (TMVA), which contains many of the most common machine learning techniques, has been developed and is widely used throughout the field.[34,35]

The Inter-Experimental LHC Machine Learning Working Group[36] is a common point of contact. This group has organised the Data Science LHC 2015 Workshop.[37,38] Several more dedicated workshops, for example the "ATLAS machine learning workshop"[39] and the "Heavy Flavour Data Mining workshop"[40] have been held. A number of particle physics challenge events have been organised over Kaggle, from finding the Higgs[41,42] to identifying signatures of heavy flavour particles.[43] The winning strategies have been further discussed and published.[44]

Over the next few years, the challenges faced in this field will greatly increase. For example, the LHC experiments will collect 100 times more data (the final datasets are envisioned to exceed 100 ExaBytes). The events will become a lot more complex (with up to 200 proton-proton collisions in each event as shown in Fig. 3) and the data needs to be fully exploited to maximise the chances of discovering new physics. The physics signatures being searched for also become more complicated. One of the next stages is looking for decays of the Higgs boson into two bottom quarks ($b\bar{b}$). This decay mode provides a

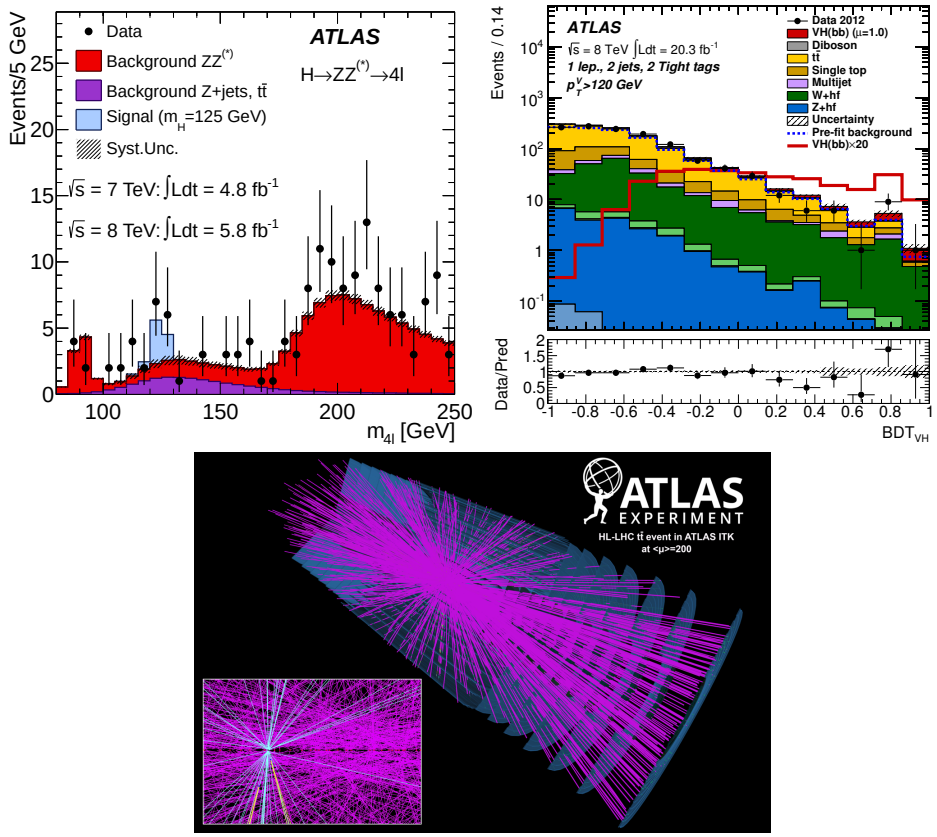

Figure 3. *Left*: The reconstructed mass of Higgs candidates, selected from the ATLAS data by requiring the presence of four leptons (pairs of electrons or muons). The mass, reconstructed from the four leptons, is shown for the collision data (black points), the simulated background expectation (red and purple filled histograms) and the simulated signal expectation (light blue filled histogram) for a SM Higgs with mH = 125 GeV.[26] A significant excess (thirteen events from the billions recorded) can be observered around a mass of 125 GeV which is consistent with a SM Higgs boson. *Right*: The output of a boosted decision tree (BDT) trained to identify collision events where the Higgs decays to a pair of bottom quarks (solid red histogram) against the multiple sources of background which can mimic the signal (other coloured histograms).[27] A comparison between the four lepton (left) and bottom-quark decay channels (right) illustrates the larger and much more complex background processes which will need to be overcome in future studies. *Bottom*: A simulated event containing a top quark pair at an average pile-up of 200 collisions per bunch crossing, illustrating the challenging conditions at the upgraded LHC. The bottom-left inset is a 2D r-z view of the interaction region. The vertical scale is 2.5mm and the horizontal one 12cm. The tracks coming from the $t\bar{t}$ vertex of interest are coloured in cyan. Two secondary vertices can be reconstructed and the tracks coming from them are highlighted in yellow.

particular challenge, as background processes generate $b\bar{b}$ pairs 9 orders of magnitude more frequently than the Higgs signal. Once this background is reduced, a complex mixture of difficult-to-model background processes also needs to be suppressed as demonstrated in Fig. 3[27] . These challenges will require another leap in our usage of Big Data techniques in terms of more advanced machine learning algorithms, pattern recognition and data handling/storage/processing solutions.

3.4 Current highlight from the co-design of computer architectures and algorithms

The physical sciences are facing unprecedented challenges brought about by huge increases in data from experiments ranging from the Euclid satellite observing the structure of the Universe, through Earth observations of climate, down to sub-nuclear particle interactions at the LHC. By 2023, new experiments in astronomy, chemistry, physics, and material science are projected to increase these data holdings to over 1EB. High-performance computing (HPC) systems remain key to these fruitful data exploitation efforts and, moreover, their scientific interpretation depends on simulations solving partial differential equations of matching resolution and precision; both Archer (www.archer.ac.uk) and DiRAC (www.dirac.ac.uk) are currently producing more than 10PB of research data for this purpose, with next generation systems set to increase this by a factor of

30-50 by 2020.

Future progress, however, is endangered by technological, algorithmic, and programming model revolutions brought about by the growth of power and complexity in computing architectures. The most recent computer processor architectures are making use of on-chip vectorisation to give Moore's Law improvements in chip performance. However, this form of on-chip parallelism is difficult to exploit, so while CPU and GPU architectures grow more parallel, there is no gain in serial processing speed at all, rendering most existing algorithms no more capable than they were several years ago. Compounding this, memory bandwidth also lags behind, meaning that CPUs typically receive data at rates orders of magnitude slower than they are capable of processing. No clear direction has emerged to deal with these key issues, which represent some of the greatest challenges in data-intensive computing.

To solve these Big Data problems, we need to adopt a cross-disciplinary approach combining new algorithmic advances with architectural co-design, in partnership with the technology industry.

Algorithms for nested and many core parallel architectures In order to utilise the computing power of new complex computer systems and advance our scientific understanding, we need to take a fresh look at key algorithms underpinning data analytics. We need specialist (parallel) research software developers collaborating with computer scientists and mathematicians to develop novel new algorithms and/or improved implementations of existing methods to exploit the latest many-core processor architectures. By undertaking this research in close collaboration with hardware vendors, the UK data science research community will have a unique opportunity to be at the forefront of future parallel computing technology. As an exemplar, we have recently shown how such nested and many-core algorithms can deliver orders of magnitude increases in the performance of scientific software for Planck satellite analysis.[45] This 100x speed-up was generated after a series of changes carried out in the process of porting the code MODAL to the Intel Xeon Phi coprocessor (Knights Corner); see Fig. 4, bottom panel. MODAL is used to analyse the Cosmic Microwave Background (CMB), a microwave frequency background radiation left over from the Big Bang. In analysing this data from the origin of the Universe and verifying it against theoretical models, we reconstructed the CMB bispectrum for the first time. In seeking to understand how the Universe emerged out of an intense period of expansion, called inflation, we found evidence of extra dimensions.

A production run using the original MODAL code (unoptimised, pure MPI) takes about six hours on 512 Intel Xeon E5-4650L cores of the COSMOS SGI supercomputer. The better code ran 80x faster on this x86 system. We were able to reduce the run time to 4-5 minutes using the new version of MODAL. This increased the size of parameter space and greatly reduced the errors in fitting to the data. It was these two factors that enabled new physics to be spotted in the Planck data.

Addressing new memory hierarchy and data volume challenges Modern and future computing platforms have complex memory hierarchies with a growing number of layers between slow long-term storage and the fastest CPU caches; this presents a great challenge both to hardware design, efficient platform usage, and to the software pipelines processing and analysing the data. New algorithms need to be developed and heterogeneous architectures considered as a solution to the increasing disparity between the ability to process the data (in the CPU or accelerator) and access the data (from the memory or disk). Particular attention needs to be paid to data locality even within clusters, nodes, and processors – *data movement and its flow through the system now has to be understood and modelled by programmers and numericists*. In addition, memory increases will be possible using larger on-processor caches and new non-volatile RAM technology which are soon becoming available. This provides the opportunity for new programming methodologies exploiting in situ analysis and "on the fly" post-processing without any intermediate data products stored. This was demonstrated at SC '15 for the real-time visualisation of 10TB early universe datasets; see the top panel of Fig. 4. This was done in collaboration with the Intel OSPRay team - see `https://youtu.be/yrw_11Tc8JQ` for a demonstration and explanation of this approach.

System architecture co-design for efficient data analytics The work outlined in the previous two paragraphs shows how productivity and the ability to interpret large complex data volumes has been greatly enhanced through engagement with hardware vendors in co-designing future data analytics platforms. This has been demonstrated many times in the past, with one of the most recent successes being the co-designed COSMOS@DiRAC heterogeneous shared-memory-Xeon Phi system. Such flexible heterogeneous architectures facilitate the rapid development of complex data analytic pipelines, ameliorating the impact of Amdahl's Law in regions where high efficiency remains to be achieved. The hybrid COSMOS platform enabled development of the scalable many-core Planck pipeline which recently received the HPCwire Readers' Choice Award for *Best Use of High Performance Data Analytics* (SC '15). Ongoing and future projects, like the Gaia Mission, will provide ground-breaking opportunities for technology demonstrators using new architectures. These developments go well beyond co-designed hardware to system software, requiring a broader vendor partnership on capabilities such as 'offload' and 'reverse offload' between computer subsystems, as well as adaptable resource scheduling.

Parts of the UK data science community already have a strong track record in developing new algorithms hand-in-hand with new computer architectures in order to tackle some of the most demanding and pressing computational challenges facing the

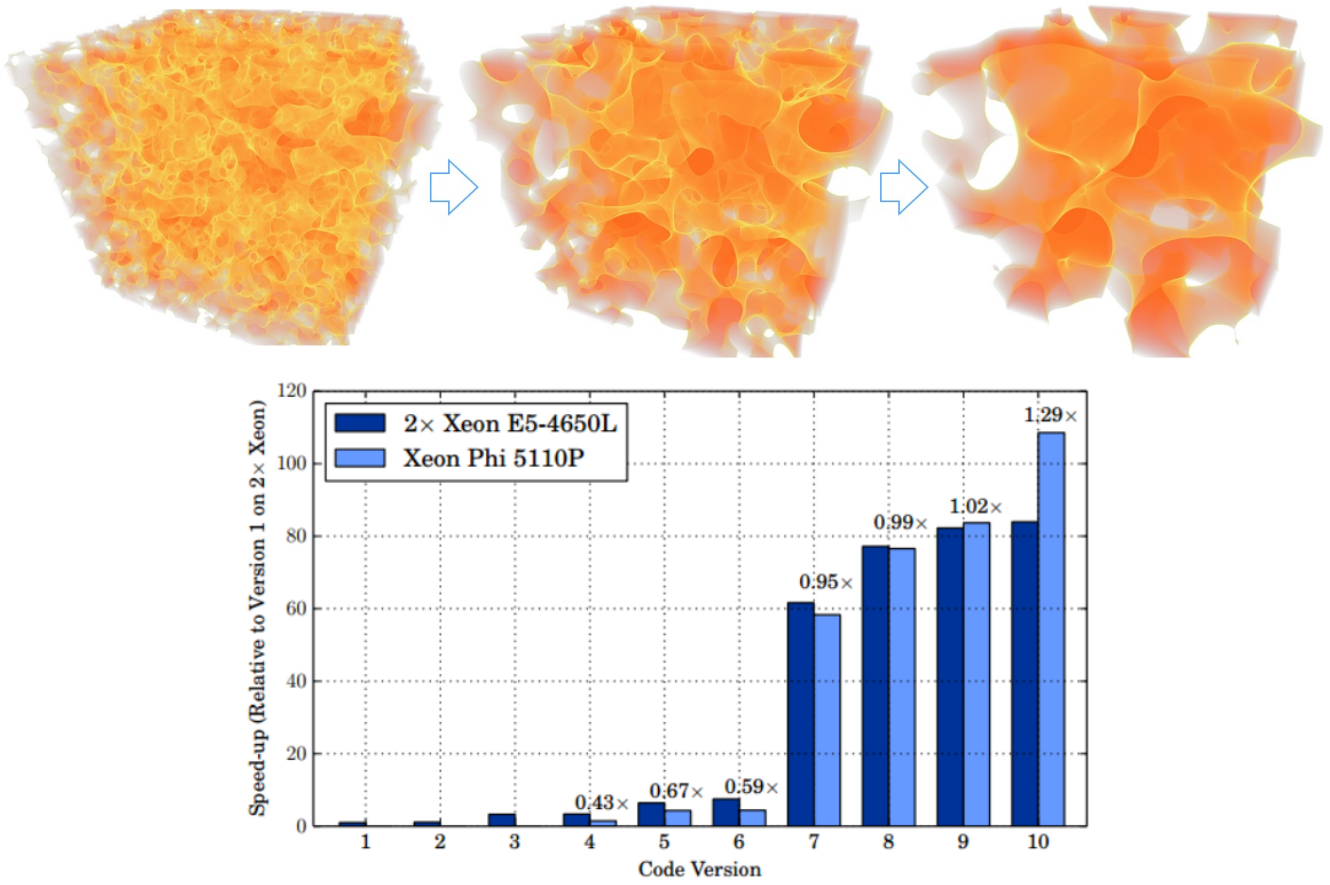

Figure 4. *Top:* Real-time visualisation demonstration at SC'15 of a huge 10TB dataset of early universe domain walls. This work is part of a COSMOS IPCC collaboration on Intel's new many-core and in-situ OSPRay ray-tracing application. *Bottom:* Plot showing the speed-up for the code Modal on the Xeon x86 and Xeon Phi chips. This speed-up was largely as a result of code changes and using better algorithms. This was required to get the performance offered by the cheaper Xeon Phi. However, these changes also greatly sped up the code on the x86 chips as well.

physical sciences today. We foresee the ATI as an unprecedented cross-disciplinary environment in which to jointly advance both algorithms and architectures, in collaboration with key industry players like Intel.

3.5 Current highlight of impact beyond the physical sciences

A very common problem in astronomy and other fields is that the dataset is large, but the amount of information that is sought is small. Can we exploit this imbalance to find efficient analysis methods to extract the information? In some cases, the answer is yes, and massive compression of the original data can be performed, in a way that is designed to preserve essentially all of the desired information. An example that was devised for astronomy, but which has been transferred to the medical imaging field, is the MOPED algorithm (Massive Optimised Parameter Estimation and Data compression[46]). MOPED began with the Sloan Digital Sky Survey (SDSS; `www.sdss.org`), which collected spectra of around a million galaxies, each of which had around 2000 spectral measurements. The light from each galaxy was the sum of the light from many stars, each of different ages and compositions, and which is partly obscured by dust. The problem was then to search a modest (typically 15-)dimension parameter space to find the best fit and errors for each of the 15 numbers, and to repeat a million times. These parameters characterise the star formation history, the composition, and the amount of dust.

The question which MOPED addressed, and solved, is 'Are there linear combinations of the data that can be constructed, in such a way that they (the compressed data) contain as much information about the parameters as the original compressed dataset?'. It turns out, that for Gaussian-distributed data, where the information is contained in the mean of the data, it is possible. The resulting compression is not strictly lossless, but almost so, in the sense that the shape of the likelihood surface around the most probable parameter set is the same - the expected curvature is the same as when the whole dataset is used.

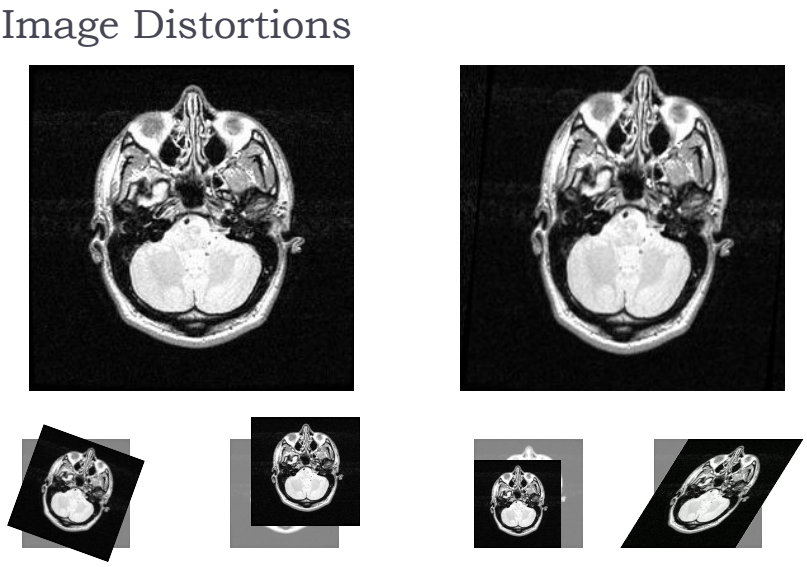

Figure 5. MRI scans of a brain. This illustrates a 12-parameter (in 3D) affine transformation, with more distortions than a rigid body allows. Scanners can introduce such additional distortions.

The result is that the sampling of the parameter space is much faster; likelihood functions have 15 operations instead of 2000, and are two orders of magnitude faster to compute. After an overhead of working out the optimal compression, the likelihood evaluation time scales with the number of parameters, not the size of the dataset. A crucial advantage for SDSS was that the overhead is done once, and the fast algorithm then applied to all the galaxies.

MOPED has been used for registering medical images, such as MRI brain scans. For this problem, the size of the dataset may be some tens of millions of voxels, and the task is to align two or more scans, very fast. Here the assumption is that the images are (almost) identical, but the position of the head in the scanner is different. To line the images up for comparison, 3 translations and 3 rotations need to be determined (in the simplest rigid-body case) - 6 numbers only. MOPED can compress the millions of voxels to 6 numbers, and find the relevant transformation quickly, apply the transformation and allow the specialist to compare rapidly the areas of concern. A spinout company, Blackford Analysis (www.blackfordanalysis.com), has taken this idea, extended it and developed new algorithms to treat a host of medical registration and other imaging problems, being able to compare images in different modalities (CT vs MRI, for example), and to register soft-tissue images where a rigid body transformation would not be appropriate.

4 Current and forthcoming data science challenges

4.1 Extracting meaning from Big Data

Big Data as it pertains to the physical sciences does not relate solely to the size of datasets and generic definitions of Big Data (i.e. the *V*'s) but is in itself multi-faceted. Big Data in the physical sciences encompasses big theoretical models and simulations, which typically pose extremely challenging computational problems, the big parameter spaces of such models, and the big algorithms required to solve theoretical models, implement them efficiently, and to explore high-dimensional parameter spaces. Big scientific discoveries are increasingly also requiring big collaborations, presenting social and organisation challenges, and providing scope for larger engagement with the general public through outreach and with industry. While all of these factors are critical in extracting meaning from Big Data and realising its potential, the most pertinent are those related to specific computational and methodological challenges.

From the computational perspective, in many instances, data are simply to large to hold in memory or to store altogether, stream processing is challenging, visualisation, while often insightful is difficult, new and diverse programming models require specialised skills, failures can be problematic, and energy consumption is an increasingly important concern that is not always given sufficient attention. To mitigate these challenges, new approaches are required, such as memory centric computing, where compute power is brought to memory rather than the reverse, accelerator rich computing (e.g. GPUs, FPGAs, Xeon Phis), new failure-recovery modes, alternative programming models to consolidate the diverse set of current frameworks, and many others. Although single solutions to these issues are not likely to be forthcoming, exemplars should be sought and developed to provide

general guidance.

From the methodological perspective, while Big Data is rich, it is also complex, which poses many challenges. The general methodological challenges outlined by Fan et al.[47] all apply: namely, challenges associated with heterogeneity of data, error accumulation, spurious correlations, and incident endogeneity (i.e. chance correlations between signals and noise). Moreover, the validation of methodological approaches and results, and the quality of data and meta-data, is critical. In some cases, model-agnostic exploratory science is required, while in other cases model-based consolidatory science is needed. The ability to make probabilistic statements is also important. Additional challenges arise due to the specifics of experiments, and their associated data acquisition systems, and of observational signals of interest. Space-based observations must be resilient to hardware failures, the geometry of the domain of observational signals may be non-trivial (e.g. observations over the celestial sphere live on spherical manifolds) and, at a fundamental level, experiments may not be repeatable (e.g. we have only one Universe to observe). There are a host of methodological approaches for dealing with these challenges, from high-dimensional MCMC analyses, hierarchical probability models, variable selection, and experimental design, to optimisation, sparsity, and compressive sensing, to artificial intelligence approaches such as machine learning and recent evolutions of deep learning. All of these areas continue to undergo substantial developments, but challenges remain in all areas, and there is no methodological panacea.

While direct solutions to mitigate computational and methodological challenges are proceeding at pace, the most effective solution is sociological. To make more effective progress in-depth, multi-disciplinary collaboration is required between researchers in the physical sciences and in statistics, applied mathematics and computer science. In-depth collaboration is currently hampered by the different language and jargon adopted in different fields and a lack of a deep understanding of the fundamental problems, and solutions, offered in different fields, which essentially stems from a lack of interaction and communication. This sometimes results in idealised solutions that may not be suitably robust for practical application. Nevertheless, such solutions can lay the groundwork for subsequent robust approaches. The ATI has a unique opportunity to overcome these challenges by establishing a forum for fostering deep interaction between researcher across a broad range of fields in order to drive meaningful interdisciplinary collaboration. These aims could be promoted through workshops, seminars, hackathons, tutorials, discussion groups, and space and time for collaborative projects. Creating a taxonomy of problems, methods, their properties, and best practices, validated by researchers across a diverse range of fields, would be an excellent first step in promoting deeper and more meaningful discussion.

4.2 Challenges arising from data-centric large-scale simulations

The state-of-the-art projects that are generating, or will soon generate, Big Data in the physical sciences heavily rely on large-scale simulation effort. The physical processes under investigation are for the most part inherently non-linear, tightly coupled, and have large dynamic range. The same applies to instrumental, selection, and systematic effects. This makes precise numerical modelling of the processes and experiments a necessity, not least because the complexity and cost of Big Data projects require detailed dry runs with simulated data. Moreover, simulations enable the numerical evaluation of complex theoretical models, creating Big Data before a single experiment has been conducted.

Metadata Synthetic Big Data has specific issues, e.g. with regard to its viability for certain science questions, which necessitates careful data management of large-scale simulations. Accurate metadata is vital in this case, and – although the data can in principle be readily recreated – simulations in the Big Data era are worth curating due to the vast arrays of data involved, arising from multiple sources. The ATI could define open metadata standards to aid reproducibility and transparency across disciplines, and address questions relating to what happens if metadata standards change over time. It should play a key role in developing best practice relating to the curation of data.

Error control and reproducibility Researchers need to be confident that running a simulation again will return the same results (such as in cases where coding bugs have influenced results), and that numerical and methodological errors are under control. Developing mechanisms to test code validity, as well as mandating proper data-centric software engineering methodologies could help mitigate such problems. The confidence to believe that some synthetic data does exactly what it says on the label (i.e. in the metadata) should be an achievable goal, and is a specific curation issue that the ATI could assist with, which is especially important when using data from multiple sources.

Efficiency Algorithmic challenges arise in the efficiency of computation, and the need to balance efficiency and accuracy. An interesting open question relates to the efficiency of methodologies which employ ensembles of short simulation runs versus the opposite route of developing dedicated hardware to efficiently perform long runs. Both approaches are currently followed in physical sciences applications, but which one gains more accuracy in less wall clock time is not clear. The ATI could take a role in helping to standardize such methodologies, and to building platforms to parametrise simulations. Training from the ATI on latest testing techniques would be useful in this respect.

Multiscale modelling Multiscale simulations often rely on multiple codes/pipelines, and there is a need to validate automated versions of such pipelines while maintaining reproducibility, e.g. via machine learning techniques. While current hybrid multiscale modelling techniques exist, the ATI could partake in developing frameworks to link models of different scales, and bring data descriptions of other scales into the development of models. Moreover, the ATI could assist in the development of novel frameworks, building hybrid models that incorporate observation. Such data-constrained modelling, which incorporates observational data, is of great interest when reconstructing partially observed physical systems. Various applications exist in geosciences, astrophysics, and meteorology, e.g. updating weather forecast models by incorporating weather observations. The ATI could help facilitating collaboration across disciplines with researchers working on multiscale modelling in the UK.

4.3 Data management, indexing, and selection

Big Data creates challenges in the area of low-level data management, indexing, and selection. There is a natural a-priori presumption that pure low-level data management services (e.g. the high throughput reliable data movement services such as FTS http://egee-jra1-dm.web.cern.ch/egee-jra1-dm/FTS) are predominantly an engineering problem, and not an algorithmic one. Nevertheless, we identified several topics, commented on individually below, which are appropriate for an ATI focus and warrant further investigation.

Data indexing In some areas data recording and storage on a per file basis is the norm (e.g. particle physics). Each file consists of a large number of "events", each with some characteristics. Historically the file size chunk has been the analysis unit (i.e. one CPU job reads one or more files). However, as data rates increase, this looks less and less like a good strategy, and work is needed to understand how to very efficiently index each event. Examples exists, such as within the LHCb experiment at CERN, where an internet search engine company has invested time in this area. The ATI can assist in efficient event-based indexing algorithms.

Data replication strategy Many of the data-intensive science sectors are typified by very large international collaborations owning very large datasets that require multiple copies to be stored on a federated distributed computing infrastructure. The size of such datasets is now so great that an naive ad-hoc "a few copies here and there" approach is no longer justifiable in terms of storage resource costs. Therefore much more intelligent replication strategies are required, which need to include data popularity (how often are these data accessed), just in time on disk strategies, etc. Of crucial importance to this topic are data *deletion* strategies, which can be the hardest thing to do, resulting in much dark data occupying storage. The ATI can assist in characterisation of data usage and its use in global replication strategies.

Data federations This topic is related to the above. Driven by the perceived rising importance of storage costs, and therefore the need to minimise copies, several large data-intensive science areas choose to de-couple the CPU processing power form the data location by making much more use of the network. Specifically a data federation deploys strategies for making the data transparently visible across the network where there is no local copy (e.g. http://iopscience.iop.org/article/10.1088/1742-6596/513/4/042005/pdf). There may also be scope for understanding how to better incorporate network availability metrics (the network weather map) into decision making. This may also be relevant to software defined networking (SDN). Presumably there is much synergy here with commercial warehouse sales structures, using diverse delivery mechanisms. The ATI can assist in algorithms for logistics decision making for delivery, given distributed storage centres, and a variable delivery mechanism.

Fast Distributed Databases / Analysis in databases Sky survey databases are growing at such a rate as to challenge astronomers' traditional practice of downloading subsets of data from a database for analysis on local resources. For the next generation of surveys, the scientifically useful chunks of data will often be too large for that to be readily done, so that astronomers expect they will have to send to a data centre not only the query needed to identify the data of interest but also a description of the data analysis task to be run on it. This throws up a number of practical problems – e.g. the sandboxing of user-supplied code – but also more fundamental issues of how best to store large datasets in order to facilitate particular classes of analysis. For example, astronomers typically use relational databases to store their data, but many of their large-scale statistical analyses centre on density estimation tasks that are effectively implemented using k-d tree or similar data structures not typically found in conventional storage engines. NoSQL databases offer more flexibility than the relational model, but tend to be optimised for particular access patterns, so there is likely to be a generic requirement within data science for study of data structures supporting statistical analysis of large subsamples of data selected from larger datasets of high dimensionality.

4.4 Big Data, algorithms, and architectures

It is recognised that we are currently in an age where IT technology is evolving at such a rate it is having disruptive effects on system design and software. The main drivers of this change are

1. robust many-core (XeonPhi, GPU, FPGA) and multi-core (Xeon, Power) processors, with on-chip vectorisation;

2. larger cache and RAM images per core;
3. improved network latency and bandwidth both on processors, between processors, between servers, and between the servers and IO platforms;
4. much lower latency storage options such as Non-Volatile RAM or Solid State Disks;
5. multi-level memory hierarchies with distributed memory, slow and fast local memory, and multi-level caches;
6. high Read and Write rates onto disks and faster parallel file systems;
7. robust cloud/grid technologies;
8. improved programming languages for GPU (OpenMP4.1) and FPGA (OpenCL);
9. recognition of the role of research software engineers in producing readable, reliable, and efficient codes.

As we better understand, and are able to quantify, the data movement

- on processor,
- between processors,
- from node to parallel file system,
- from parallel file system repository/archive,
- to web browser/cloud powered workflow,

we can better design systems and codes. Work by Boyle (Edinburgh) and Shellard (Cambridge) has shown that this approach can produce systems and scientific codes that increase productivity by 1-2 orders of magnitude for a fraction of the cost incurred by simply scaling up existing technologies and designs. These changes allow us to produce bigger, better, and more complex data from simulations, as well as allowing experiments to generate much more data, which can be analysed in a timely fashion. In addition, we are also seeing that the dividing line between High Performance Computing (HPC) and High Throughput Computing (HTC), which essentially was the quality of the intra-cluster network to allow parallel jobs, is becoming blurred, because the same network is now used for high IO data parallel work. Perhaps HPC should now stand for High Performance Components?

The research drivers are our quest to better understand the world around us and be able to predict outcomes based upon robust models and/or measurements. In a 4-d spatial-temporal model increasing the resolution by x2 along each axis brings an increase in the grid size of x16. Increasingly, new physical processes need to be added to models as resolution increases, which is another major resource requirement. There is a very real tension between our need to model and explore data at the level of detail we require and the fact that Moore's Law gives us 4x increase in floating point operations every 3 years. We therefore rely on greater parallelism and better code and workflow efficiency to generate the insight that we need. Thus the need for better algorithms that use this parallelism and increase basic code efficiency.

A particular example is in the area of data analytics, which relies on thousands of models to be run in order to fit models to data, discern missing physical processes and understand systematic effects. This is becoming a dominant activity in HPC projects, such as DiRAC (www.dirac.ac.uk). DiRAC ran the numerical models used to interpret the recent gravitational wave detection by LIGO and has performed a similar role for Planck and Lattice QCD. The use of HPC has been shown to greatly improve research outcomes and time to science.

When faced with these super Moore's Law requirements, we are seeing various approaches being undertaken to make sure the "computational" performance tracks the increase in requirements and yet still is affordable. Important work is now being done on improving libraries for cheaper processors such as GPUs, Xeon Phi, ARM processors, and FPGAs (e.g. Boyle's work on maths libraries for Xeon Phi); designing heterogeneous architectures for offloading calculations to appropriate architectures (Jäykkä, Cambridge); and better management of problems that suffer from load balancing (Theuns, Durham). At the heart of this work is the return to the algorithm; that which translates our problem into a set of instructions that can solve our problem.

The return to the algorithm should lead to new forms of abstraction, which allows computer/architecture-aware source code and executables to be generated. This should allow us to use sensible symbolic languages to define our problems algorithmically and let other tools generate source codes for the appropriate system architectures and, increasingly, sets of architectures. The ATI can play a key role in encouraging the development and use of new forms of abstraction.

Looking ahead, the physical sciences are rapidly approaching the ExaByte era. These data sizes and rates are presenting us with severe algorithmic and architectural challenges. To extract meaning, we need to make use of algorithms which are

tuned to those architectures capable of solving particular aspects of the problem in a reasonable time. The need to exploit other computational architectures, low latency memory/storage, and other data resources lead us to the notion of "in situ data reduction" in which the simulation and data reduction are contained within the same program and/or system. Such a program is able to exploit different parts of the node/cluster/cloud in order to arrive at a timely solution; the program is aware of its architectural environment and the options open to it. The ATI should help drive the development of tools to translate a problem to the algorithm, then to source code and then to an architecture aware program. Such tools are key to our ability to extract meaning from data in the ExaByte era.

Specifically the ATI can assist with

1. the development and use of hardware and software technologies that both speed up, and control, data movement;
2. the development of data exploration tools that can discover meaning in PetaByte to ExaByte size datasets in a meaningful time frame;
3. the development of tools that can federate and correlate data of different types;
4. the development of tools that allow the confrontation of experimental data and simulations at the PetaByte scale.

5 Conclusions & recommendations

Within the limited space of this short paper, we provided a broad-brush overview of the challenges arising in the physical sciences in the beginning era of Big Data and highlighted a select few specific examples in the different disciplines. From these representative showcases, as well as the discussions at the ATI Summit, it has become evident that there is strong interest, and indeed an increasing need, to boost the cross-fertilisation between data science and the physical sciences. This effort is ideally facilitated by the ATI and will generate countless opportunities for excellent research and high-impact applications.

Why do we need close links between data science and the physical sciences? Many disciplines in the physical sciences become increasingly overwhelmed by extremely high data volumes and data rates, generating some of the largest datasets on the planet. Hidden within these datasets is scientifically and commercially valuable information that is often difficult to identify and to extract robustly. The precision and accuracy of measurements need to be quantified in a probabilistic manner, creating big inference problems. The reproducibility and replicability of experiments with Big Data become a challenge in their own right, requiring the standardisation of algorithms and computation to ascertain credibility of the data analysis.

In response to these challenges forefront research areas in the physical sciences are undergoing a profound transition, with increasingly large collaborations that pool expertise, including a significant fraction of roles focussed on data science. New jobs and career paths have emerged in this area, but academic training lags behind, with no formal degrees in the UK yet that merge expertise in both data science and physical science disciplines. These changes reflect that the traditional modularity of the scientific endeavour – the scientific experimentation and interpretation, the application of out-of-the-box data analysis and statistical inference tools, and the provision of codes to perform such operations on the data – breaks down in the era of Big Data. Now, the data analysis tools have to be tailored towards the dataset and the questions asked on it, and implementations of the algorithms have to become aware of the specifics of the computational architecture available.

To ensure continued UK world leadership in many physical science disciplines in the imminent age of Big Data, we therefore need a productive flowdown of innovation from the principles of data science to applications in the physical sciences, and on to commercial exploitation in industry, impact on societal concerns, and new routes to public engagement with fundamental science. This will unlock synergies between the different areas of the physical sciences that face similar data-related challenges, e.g. via common cross-disciplinary tools as well as common data and computing standards.

Why should the ATI engage with the physical sciences? The physical sciences will be able to provide the ambitious data scientist with the most challenging applied data science problems and the most extreme datasets in existence; for instance: data volumes that will dwarf global internet traffic, data at PetaByte per second rates, and global sensor networks whose data are heterogeneous in time, space, data rate, and data type – pushing the boundaries of the defining characteristics of Big Data. Contrary to many other fields, this Big Data is generally free, not burdened with ethical restrictions, and readily accessible.

By engaging with the Big Data-related challenges and opportunities in the physical sciences, the ATI will not only support a wide range of forefront research and UK leadership therein, but also generate impact inside and outside academia. Acting as a catalyst for the aforementioned innovation flowdown of modern data science concepts, the ATI can play a key role in pushing methodological and algorithmic developments in the physical sciences, one of the main pathways to impact in industry and society. Another major output of physical science research is a highly trained workforce of problem solvers who go on to work in finance, consulting, R&D departments, and various other fields. Since the problems to be solved increasingly relate to Big Data, it is essential that junior physical scientists receive expert training in cutting-edge data science. The ATI could be a central hub for these training efforts.

The ATI as a hub to foster the symbiosis between data science and the physical sciences Both fields will benefit from a close working relationship, facilitated by the ATI. Innovation in the core data science disciplines is quickly and effectively developed towards applications in the physical sciences; in turn, conceptually new, practical data science problems are reported back efficiently and can trigger fundamental research in data science. These feedback loops have worked slowly but successfully in the past and will dramatically increase in importance in the era of Big Data.

We acknowledge that funding is limited and that few physical scientists to date have the background to directly engage with the ATI, e.g. via its secondment programme. However, momentum and enthusiasm with regard to Big Data challenges is widespread in the physical sciences community, as demonstrated by the strong interest in the ATI Summit. Below we propose a few exemplary measures with high value-to-cost ratio which would allow the ATI to engage with this large community:

- Build a pool of points of contact for various key data science subjects. These researchers would be interested in engaging with physical science applications and contribute to projects as interpreters between the languages of data science and physical sciences, providing ideas and advice.
- Provide hot desks and meeting spaces for cross-disciplinary collaborations with a clear Big Data focus, with usage ranging from single days to several weeks. This could include short-term funded secondments to the ATI on similar time scales.
- Organise workshops, seminar series, schools, and similar formats that promote knowledge exchange and collaboration between the core ATI activities and physical scientists.
- Offer joint PhD studentships with physical sciences departments at the ATI nodes, with training and research programmes that feature both core data science topics and its applications.

These measures would allow both communities to tap into the vast array of opportunities offered by Big Data in the physical sciences, ensuring continued UK leadership in research and innovation, with lasting societal and economic impact.

Acknowledgements

This document has benefited greatly from contributions by the participants of the ATI Summit on physical sciences, through discussions at the meeting and through direct feedback on this paper. We are grateful for the financial support of the Summit by UCL and by Nature Publishing. BJ acknowledges support by an STFC Ernest Rutherford Fellowship, grant reference ST/J004421/1. JDM acknowledges support from STFC (grant reference ST/M00113X/1) and EPSRC (grant references EP/M011852/1, EP/M011089/1, and EP/M008886/1). TS acknowledges support from the Royal Society, via his Royal Society Research Fellowship.

References

1. Laney – 3D Data Management: Controlling data volume, velocity, and variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-\Velocity-and-Variety.pdf. Accessed: 2016-03-02.
2. The Alan Turing Institute. https://turing.ac.uk. Accessed: 2016-03-03.
3. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**, 1087–1092 (1953).
4. Hammersley, J. M. & Handscomb, D. C. Monte carlo methods. *Methuen, London* 36 (1964).
5. Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**, 97–109 (1970). URL http://biomet.oxfordjournals.org/content/57/1/97.abstract. http://biomet.oxfordjournals.org/content/57/1/97.full.pdf+html.
6. Gelfand, A. E. & Smith, A. F. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* **85**, 398–409 (1990).
7. Högbom, J. A. Aperture synthesis with a non-regular distribution of interferometer baselines. *AAPS* **15**, 417 (1974).
8. Taylor, H. L., Banks, S. C. & McCoy, J. F. Deconvolution with the ℓ_1 norm. *Geophysics* **44**, 39–52 (1979).
9. Chen, S. S., Donoho, D. L. & Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM JOURNAL ON SCIENTIFIC COMPUTING* **20**, 33–61 (1998).
10. Mallat, S. *A wavelet tour of signal processing* (Academic press, 1999).

11. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288 (1996).

12. Candès, E., Romberg, J. & Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure and Appl. Math.* **59**, 1207–1223 (2006). arXiv:math/0503066.

13. Donoho, D. Compressed sensing. *IEEE Transactions on Information Theory* **52**, 1289–1306 (2006).

14. Brook, P. *et al.* Emission-rotation correlation in pulsars: new discoveries with optimal techniques. *Monthly Notices of the Royal Astronomical Society* **456**, 1374–1393 (2015).

15. Gregory, P. Bayesian exoplanet tests of a new method for MCMC sampling in highly correlated model parameter spaces **410** (2010).

16. Gibson, N. P. *et al.* A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy. *Monthly Notices of the Royal Astronomical Society* **419**, 2683–2694 (2012).

17. Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S. & Roberts, S. A Gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society* **452**, 2269–2291 (2015).

18. Aigrain, S., Hodgkin, S., Irwin, M., Lewis, J. & Roberts, S. Precise time series photometry for the Kepler-2.0 mission. *Monthly Notices of the Royal Astronomical Society* **447**, 2880–2893 (2015).

19. Almosallam, I., Lindsay, S., Jarvis, M. & Roberts, S. A sparse Gaussian process framework for photometric redshift estimation. *Monthly Notices of the Royal Astronomical Society* **455**, 2387–2401 (2015).

20. Sadowski, P. J., Collado, J., Whiteson, D. & Baldi, P. Deep learning, dark knowledge, and dark matter. In Sadowski, P. J., Collado, J., Whiteson, D. & Baldi, P. (eds.) *HEPML@NIPS*, vol. 42 of *JMLR Workshop and Conference Proceedings*, 81–87 (JMLR.org, 2014). URL http://dblp.uni-trier.de/db/conf/nips/hepml2014.html#SadowskiCWB14.

21. Kramer, M. & Stappers, B. Pulsar Science with the SKA. *Advancing Astrophysics with the Square Kilometre Array (AASKA14)* 36 (2015). 1507.04423.

22. Gret, A., Snieder , R., Aster, R. & K., K. Monitoring rapid temporal changes in a volcano with coda wave interferometry. *Geophysical Research Letters* **32**, 4 (2005).

23. Brenguier, F. *et al.* Postseismic relaxation along the San Andreas fault in the Parkfield area investigated with continuous seismological observations. *Science* **321**, 1478–1481 (2008).

24. de Ridder, S., Biondi, B. & Clapp, R. Time-lapse seismic noise correlation tomography at Valhall. *Geophysical Research Letters* **41**, 7 (2014).

25. Landro, M. Discrimination between pressure and fluid saturation changes from time-lapse seismic data. *Geophysics* **66**, 836–844 (2001).

26. ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the {ATLAS} detector at the {LHC}. *Physics Letters B* **716**, 1 – 29 (2012). URL http://www.sciencedirect.com/science/article/pii/S037026931200857X.

27. ATLAS Collaboration. Search for the Standard Model Higgs boson produced in association with a vector boson and decaying to a *b*-quark pair with the ATLAS detector. *Phys. Lett.* **B718**, 369–390 (2012). 1207.0210.

28. Gridpp: Distributed computing for data-intensive research. https://www.gridpp.ac.uk/. Accessed: 2016-02-22.

29. Atlas@home. http://lhcathome.web.cern.ch/projects/atlas. Accessed: 2016-02-22.

30. Feldman, G. & Cousins, R. A unified approach to the classical statistical analysis of small signals. *Physical Review D* **57**, p3873–3889 (1998). physics/9711021.

31. ATLAS Collaboration. Expected performance of the ATLAS b-tagging algorithms in Run-2 (2015). URL https://cdsweb.cern.ch/record/2037697.

32. Atlas short term assoication of pierre baldi, peter sadowski, gregor urban. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/AtlasPolicyDocuments. Accessed: 2016-02-23.

33. de Oliveira, L., Kagan, M., Mackey, L., Nachman, B. & Schwartzman, A. Jet-Images – Deep Learning Edition (2015). 1511.05190.

34. Toolkit for multivariate data analysis with root. http://tmva.sourceforge.net/. Accessed: 2016-02-22.

35. Hoecker, A. *et al.* TMVA: Toolkit for Multivariate Data Analysis. *PoS* **ACAT**, 040 (2007). physics/0703039.

36. Inter-experimental lhc machine learning working group. http://iml.cern.ch/tiki-index.php. Accessed: 2016-02-22.

37. DataScienceLHC2015. http://cern.ch/DataScienceLHC2015. Accessed: 2016-02-22.

38. Data Science@LHC 2015 Workshop. https://indico.cern.ch/event/395374/other-view?view=standard. Accessed: 2016-02-22.

39. Atlas machine learning workshop 29-31 march 2016. https://indico.cern.ch/event/483999/. Accessed: 2016-02-22.

40. Heavy flavour data mining workshop. https://indico.cern.ch/event/433556/. Accessed: 2016-02-22.

41. Higgs boson machine learning challenge. https://www.kaggle.com/c/higgs-boson. Accessed: 2016-02-22.

42. Atlas higgs challenge 2014. http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014. Accessed: 2016-02-22.

43. Flavours of physics: Finding $\tau \to \mu\mu\mu$. https://www.kaggle.com/c/flavours-of-physics. Accessed: 2016-02-22.

44. Cowan, G. e. a. NIPS 2014 Workshop on High-energy Physics and Machine Learning. *Journal of Machine Learning Reasearch* **Workshop and Conference Proceedings 42** (2014). URL http://jmlr.org/proceedings/papers/v42/.

45. Briggs, J. e. a. *J. Comp. Phys.* **310**, 285 (2016).

46. Heavens, A. F., Jimenez, R. & Lahav, O. Massive lossless data compression and multiple parameter estimation from galaxy spectra. *MNRAS* **317**, 965–972 (2000). astro-ph/9911102.

47. Fan, J., Han, F. & Liu, H. Challenges of big data analysis. *National science review* **1**, 293–314 (2014).