Creating a Chemistry of Sciences with Big Data
Building the Data Science Institute at Imperial College London

Y. Guo, D. Johnson
Data Science Institute, Imperial College London
y.guo@imperial.ac.uk, david.johnson@imperial.ac.uk

Abstract

The Data Science Institute at Imperial College London launched in April 2014, and will provide a hub for data-driven research and education. Its mission is to provide a focal point for the College's capabilities in multidisciplinary data-driven research by coordinating advanced data science research for college scientists and partners, and educating the next generation of data scientists. We surveyed the data-driven research needs at Imperial College London to gain an understanding across all disciplines offered by the College, and analysed the responses to gain insights into scientific and engineering needs for data science research.

A clear message is that multidisciplinarity is essential for Big Data and data science research to enable a "chemistry of sciences": connecting all disciplines by integrating data. This paper presents our efforts to best understand data-driven research needs in a highly multidisciplinary research-intensive institution and describes our vision for the future of the Data Science Institute at Imperial College London.

1 Introduction

Data science is the discipline that deals with collecting, preparing, managing, analysing, interpreting and visualising large and complex datasets. Data science has its roots in the integration of statistics and computer science, where it is driving scientific and technological advancement in diverse areas such as astrophysics, particle physics, biology, meteorology, medicine, finance, healthcare and social sciences. Modern science typically involves big data, taking advantage of high-throughput data capture and high-performance computing capabilities. Data science is therefore becoming an essential element of all modern interdisciplinary scientific activities, acting as the glue to facilitating collaborative scientific discovery and involving the whole life cycle of data, from acquisition and exploration to analysis and communication of the results. Data science is not only concerned with the tools and methods to obtain, manage and analyse data: it is also about extracting value from data and translating it from asset to insight.

The establishment of the Data Science Institute (DSI) at Imperial College London aims to leverage the breadth and depth of data-driven research currently being carried out across the College. Its research spans the faculties of Medicine, Natural Sciences and Engineering, as well as the Imperial Business School. The DSI opened in April 2014, and will provide a hub for data-driven research and education. Its mission is to provide a focal point for the College's capabilities in multidisciplinary data-driven research by coordinating advanced data science research for college scientists and partners, and educating the next generation of data scientists.

As modern scientific research is largely data-driven, the DSI will conduct research on the core of data science to develop advanced theory, technology and systems that will contribute to the state-of-the-art in data science, and support world-class research. The DSI will act as a focal point for expertise in data-driven research within the College to help tackle grand challenges by encouraging the sharing of data and technologies for analysis and management. Data science aims to deliver tangible value from data assets in pursuit of data-driven innovation.

To act as the hub for data science research, the DSI will coordinate a set of research networks in the form of virtual or physical research laboratories. The focus is to support College-wide cross-faculty collaborative research programmes addressing data-driven scientific grand challenges. It will develop College-wide computational infrastructure for managing and processing scientific research data and will enable world-leading data science research at Imperial to strive to be a world leader in driving data science platform development. The Institute will devote effort to establishing College policy and strategy for building and strengthening its research data assets. It will offer technology support for data stewardship, software platform development, training and project-specific collaborations. It will provide a focal point for building a global alliance of academic and industrial partners to address major data science challenges and applications. The Institute aims to generate significant intellectual property and, through strategic partnerships, to translate this into social and economic impacts. Finally, the DSI plans to offer an advanced education programme to train a new generation of data scientists. The most up to date information can be found on our website shown in Figure 1.

To best understand the needs and driving factors for science and engineering disciplines, the DSI leverages Imperial College's highly diverse research areas to gain a representation of multidisciplinary needs throughout the College. The following section describes a survey on data-driven research needs at Imperial College London.

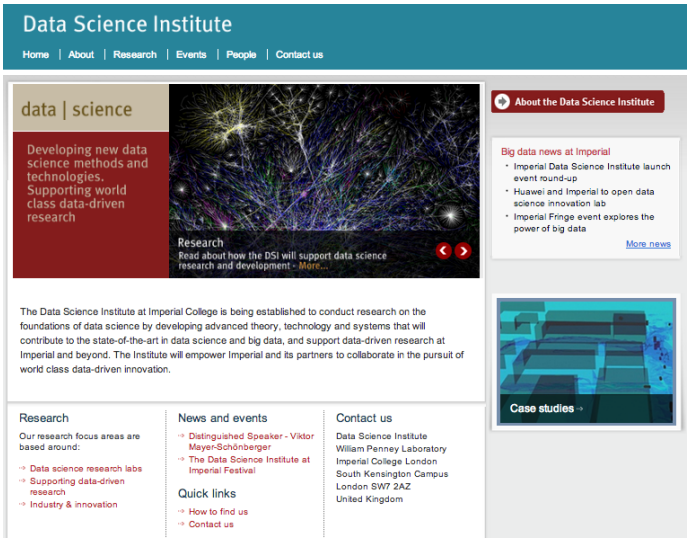

Figure 1: Screenshot of the homepage of the DSI website [6].

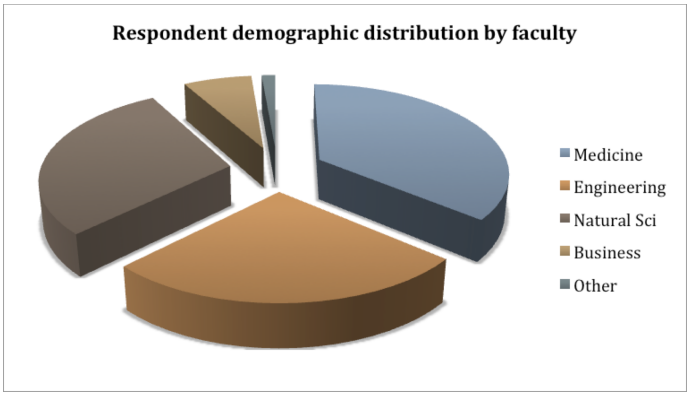

Figure 2: Pie chart illustrating the survey respondents by demographic distribution (faculty).

2 Survey of data-driven needs

In November 2013, we commissioned a College-wide survey to elicit information to help with building engagement, to assess where the main data science needs exist, and to investigate how the DSI could have value, assist and support research across the College. 147 people, drawn from the three faculties and the Imperial Business School completed the survey, representing a 2% return rate in proportion to the total number of staff in the College. Based on the responses to the survey, significant and broad support for the establishment of the Institute is observed, where there is a strong message for building greater collaboration and an active contribution to the College's research and engagement capacity.

The survey consisted of four open-ended questions exploring the current data-driven research needs of Imperial research and the impact of the establishment of the Institute across the faculties and business school. Respondent roles were broad, where staff involved in leading, supporting or undertaking deep and complex research in engineering, medicine, natural science and business returned the survey.

As shown in Figure 2 the largest group of respondents came from the Faculty of Medicine (n=54; 37%), closely followed by Natural Sciences (n=43; 29%) and Engineering (n=39; 27%) respectively, where the three faculties made up approximately 93% (n=136) of survey respondents. The remaining portion is distributed between the Business School (n=10; 7%) and other staff (n=1; ¡1%). The distribution between the three main faculties is relatively even, which indicates a broad interest in the survey and Institute. While respondent numbers from the business school was less, in proportion to the number of staff the response represents a 4.2% return rate, as opposed to 2%, 2.6% and 3.1% return rates from Medicine, Engineering and Natural Sciences respectively.

The survey also demonstrated a variety of respondent roles, where the majority of responses came from faculty, research, or teaching staff, as illustrated in Figure 3. Administrative and support staff were not envisaged to have any specific needs in relation to data science, although the survey was also sent to administrative staff to broaden the possible response profile. Notably, senior staff represents the largest set of respondents, where Senior Lecturers, Readers and Professors make up the majority (n=90; 61%), and of which Professors were the largest responding group (n=62; 42%). This indicates a strong strategic resonance for the establishment of the Institute by department and faculty leaders college-wide.

The four open-ended questions in the survey that were analysed were:

1. What are your data-driven research problems?
2. How might the Data Science Institute help you?
3. How would you like to be involved in the Institute?
4. Would you like your website linked to the Institute website?

Firstly some background to current data-driven research across the college was sought (Question 1). This provides context for each response, as well as an opinion on whether or not a respondent perceives they are participating in data-driven research at all. The survey then asked how the establishment of the Institute could help (Question 2). This question elicits several different aspects. Firstly, it looks to gain an insight into whether or not current data science provision is fulfilled. Secondly, it seeks to determine if the participant has any prior knowledge of the Institute whatsoever. Thirdly, the question tries to gain an understanding of the perceived purpose of the Institute. Next, a question is posed to find out how the participant would like to be involved (Question 3). The insights gained by this question build on Question 2 on further eliciting current understanding of the Institute and its purpose, as well as acting as a gauge of what level, if any, of engagement can be expected. Finally, the survey asks if the participant wants to have their website linked to the Institute's (Question 4). While on the face of it the question seems quite discrete, the insights gained here are not so. The question seeks to elicit to what level of association participants wish to have with the Institute - a clear indicator of positive or negative perception.

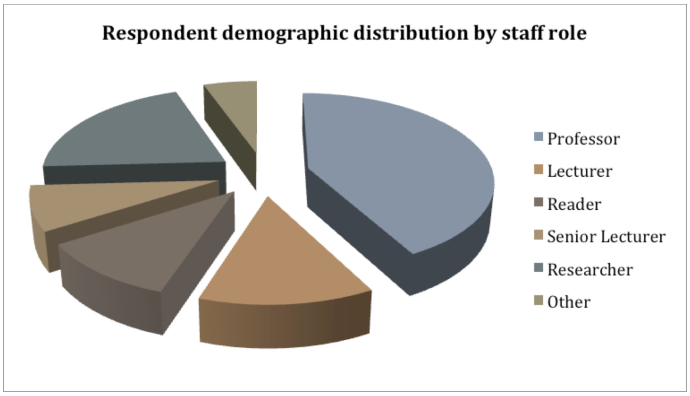

Figure 3: Pie chart illustrating the survey respondents by demographic distribution (academic role).

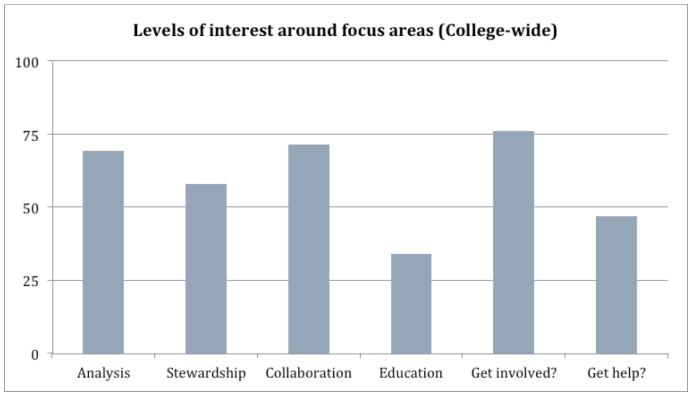

Figure 4: Histogram showing response levels on focus areas determined through qualitative analysis of returned surveys across all areas of College.

2.1 Key findings

Our analysis coalesced around four key focus areas: 'Analysis', 'Stewardship', 'Collaboration' and 'Education'. Each is defined as follows:

- *Analysis* is where the respondent indicates a clear need or current usage of mathematical, computational or other analysis techniques, such as visualisation, in their particular scientific or engineering domain;
- *Stewardship* is where the respondent indicates a need or current use of expertise, guidance, or service in managing their data or their data analysis processing;
- *Collaboration* is where the respondent indicates a clear need or potential for leveraging the Institute as a focal point for networking and connecting with others for the purposes of working together at a project-level or for searching for and obtaining funding for future projects;
- *Education* is where the respondent indicates a need or potential for the Institute to provide opportunities for training, either on an ad-hoc basis or as an organised degree, or for connecting to potential cross-disciplinary student supervision under current education provision.

Additionally, two other indicators were drawn out from our analysis of survey responses:

- *Get involved?* - The respondent gave a positive indication of active involvement with the Institute. Passive involvement, such as only consuming resources or wanting more information on the DSI was not considered as a positive indicator.
- *Get help?* - The respondent specifically requests help or assistance from the DSI in their on-going or future research projects.

Using these focus areas and indicators over the 147 responses, 608 discrete positive observations were made, a summary of which is depicted in the bar chart below.

Several trends are observed across the whole set of Imperial College respondents to the survey. Note that the results were also analysed by faculty, and found that the trends documented here are in general representative across the disciplines (Figure 4).

There seems to be a strong current theme of activity or need for data analysis as identified under the 'Analysis' focus area (n=102; 69%), which is reflected in both the descriptions of the data-driven research needs of respondents, and also in how the institute may help. Data analysis as part of on-going research, use of statistical analysis methods and computational modelling, and a need for better analytics tools are the most prevalent indicators in the written text of the returned survey responses. The Institute seems to be seen as somewhere that will bring together researchers with a strong focus on data analysis, where respondents either contribute their own expertise or would seek assistance or collaborations in analysis work. Several responses focused on sharing methods and software tools already developed by researchers within Imperial, or on a need for the development of new tools in collaboration with domain specific research. For example, seven respondents across all four faculties identified a need for "software for scientific computing" and "programming for science" to be either supported through a training programme to teach domain-specific researchers how to develop their own software, or through a dedicated scientific software development service.

It is observed that the 'Stewardship' focus also demonstrates a positive response (n=85; 58%) from across the college respondents. In particular, specific recurring needs that are present in the survey responses are not overwhelmingly for data storage and analysis capacity and capability, but rather for specific guidance and advice on how to appropriately store, process, and manage research data themselves. While a small number of respondents request computing capability, stewardship is required to make best use of existing resources college-wide, and this comes through as a recurring theme in the survey responses. Another issue that is especially of note is in respondents either needing advice on how to handle confidential or sensitive data (e.g. clinical patient data), or have indicated that confidentiality of their data is a barrier to using existing college facilities. One respondent even asked for specific help in this respect, on how to source and exploit data directly from the UK National Health Service. The DSI, as a hub of expertise in data man-

agement at Imperial, can provide this stewardship for the college in order to best utilise existing college capability, expertise, and resources.

The strongest positive response that falls in the four core focus areas is under 'Collaboration' (n=105; 71%). This is a clear signal that the majority of respondents feel that the Institute's potential as a centre for data-driven research could make it an important hub for Imperial's network of researchers, across faculties. Data science is multidisciplinary as it is part of all modern research disciplines, and it is made up of elements from disparate disciplines. This puts the Institute in a unique position within the college by being able to facilitate multidisciplinary collaboration as well as forging ahead with new research and developments within data science through collaborations. Specific recurring themes that are observed in the survey responses with respect to collaboration centre on forming partnerships for seeking and applying for large-scale strategic and collaborative research funding, using the Institute as a place for cross-fertilisation of ideas through networking events and research seminars, and as a means to strengthening links with industry.

The focus area that received the least significant observed levels of interest was 'Education' (n=50; 34%). It should be noted that although the response signal is less than the other of the four core focus areas, it does not necessarily represent any negativity in this focus area. It may be that the Institute's purpose is not perceived as having a strong focus on educational provision within the college, where excellence in research at Imperial is paramount and lays the foundation for excellent teaching. One theme that ran throughout the responses filtered under this focus area was for data science methodology and tooling training for specific domain teams. For example, a researcher in the business school responded,

> "[I would like the DSI to] provide support and training in data management and in programming for science."

Another respondent, a senior member of staff in the Faculty of Medicine wrote,

> "[I would like the DSI] as a place to train my team."

Another theme that was noted was requests for the Institute to provide student supervision at postgraduate-level (Masters or doctoral), or as a means to connecting to appropriate supervisors from other departments and faculties. An interesting observation under the theme of Education is the absence of references to specific degree course provision. Currently the impression is that training, on the level of professional or practice, may be sought from the Institute, rather than at an academic degree level. Whilst this reflects current understanding of the DSI's position within the college, it should not preclude future provision of postgraduate, or even undergraduate, data science courses or degree programmes from being developed.

On the additional indicators for involvement and help, it is observed that there is an overwhelming desire for involvement in some capacity with the Institute from survey respondents (n=112; 76%). This response from the survey demonstrates that the majority of those polled see added value in the establishment of the DSI. This is reflected across specific comments, in particular those in response to Question 3 ("How would you like to be involved in the Institute?"), for example:

> "Would welcome involvement at any level."

> "[I would like to get involved] by becoming an active participant of the Institute."

> "[I would like to be involved] actively, happy to act as a bridge to biological sciences and medicine."

This positive response is reflected widely in the number of respondents who agreed to have their websites linked to from the DSI's (n=85; 58%), where this is interpreted as being happy to be publicly associated with the Institute. This inference is supported by an association analysis[1] on the survey data (supp=0.58, conf=0.91)[2]. However it is also clear that while the majority of responses were positive and sought active involvement in the Institute, there is also significant lack of understanding in what the DSI offers in terms of opportunities for involvement. A senior staff member from the Faculty of Natural Sciences summarised this impression succinctly in their response:

> "Impossible to say [how I can get involved in the Institute] until I know more about how it will shape up, what its mission will be, and what it might offer me. This is a very hard question to answer at this stage - my whole career could be described as data science, as could many at College, so I have no real idea what the Institute will be about within that."

Other responses echoed these sentiments, or strongly indicated no desire for involvement at this time. This is not interpreted as a negative perception of the Institute, but rather that the potential opportunities for involvement in the Institute, and its purpose and mission, are not yet well understood across the college. To this end, work on college understanding and engagement with data science and the Institute should be priorities in the DSI's inaugural year of establishment. It will be made clear that involvement with the Institute does not constitute a commitment to membership - rather that the Institute aims to be inclusive by providing both a physical presence and virtual organisation that can scale to encourage participation and collaboration for mutual benefit as and when required.

Finally, on the indicator for help, where responses were analysed to determine if those polled required specific assistance from the Institute in their work, 47% of respondents (n=69) were observed to request and require some kind of help with respect to data-driven research. The kinds of

[1] *Association analysis* is a method for inferring relations between variables in large data sets. For this report the Apriori algorithm [1] was used to determine association rules within the discretized survey results.

[2] *supp* and *conf* refer to measures of support and confidence in association analyses respectively.

requests for help vary, with some respondents asking for the DSI to provide new hardware infrastructure for storage and computation, some asking for assistance in accessing data within the college and from public sources (e.g. from open initiatives such as data.gov.uk [5], a UK Government project to make available non-personal UK government data as open data, or MetaboLights [4], an open-access curated repository for metabolomic studies maintained by the European Bioinformatics Institute), help with selecting and training in appropriate tools for data analysis, and in training students how to handle data and build analysis pipelines. For example, an analyst and developer in the Faculty of Medicine wrote:

> "Help and training with techniques and tools that would allow us to analyse large datasets within a short time would save us a lot of resources and allow us to analyse the data early. It would make it easier to plan the studies and allow us to allocate the resources where they are needed most. [We would like] training in tools like Hadoop, NoSQL etc. and training in techniques that allow us to collect data from several different sources with minimal difficulty."

It is observed that many responses that include specific requests for help are very much in line with training and education, where the help requested is to help staff who are carrying out data-driven research improve their on-going and future research, rather than the Institute directly getting involved with other's research. The overall picture given by the survey is a positive one, where amongst survey respondents there seems to be a college-wide appreciation for data-driven research, data science, and the establishment of the Institute. This College-wide survey helped us shape the mission and organisational mechanisms of the Institute to be as inclusive and representative as possible for the whole of Imperial College London's data-driven needs.

The DSI can play a vital role in enhancing and leveraging college-wide talent and state-of-the-art in data-driven research activity. Existing infrastructure can be more efficiently utilised through collaboration, where the DSI will act as a focal point for data science activity. The Institute will act as advisor and educator for data-driven research at Imperial College, as well as working with all within College to develop new tools and technologies at the forefront of data science. It is clear that the establishment of the Institute is seen as a positive step, however much work is beginning to further engage and inform the college, and public at large, of the DSI's mission and vision for the future. The rest of this paper describes this mission.

3 Mission

The DSI was formed to be a focal point of Imperial College in data science research. For College, it facilitates multidisciplinary research by providing support to data-driven scientific activities. For the outside world, it provides a uniform interface for Imperial when collaborating with international partners and governments to form research

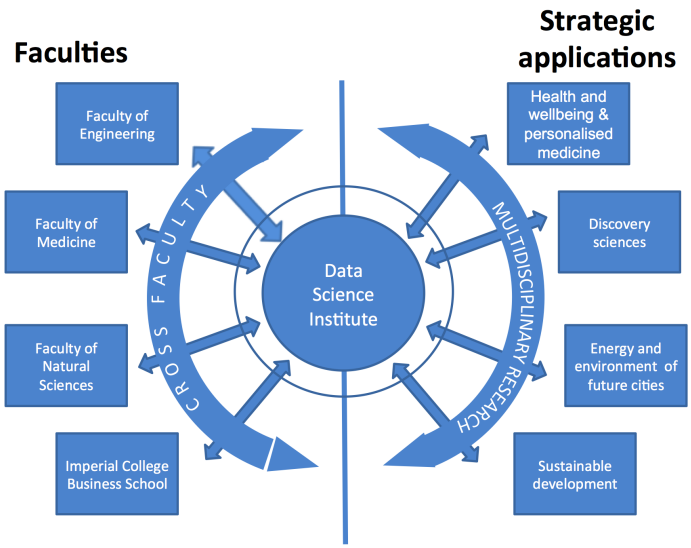

Figure 5: Diagram showing how the DSI will act as a focal point for College faculties, as well as for strategic Big Data application areas.

programmes addressing science and technology grand challenges (Figure 5). The DSI's mission can be summarised as the following five points:

1. catalysing multidisciplinary research;
2. building our capacity in data science;
3. educating a new generation data scientists;
4. collaborating with international partners;
5. translating innovation to commercial or social value.

3.1 Multidisciplinarity

A key mission of the DSI is to facilitate multidisciplinary research. Scientific research today is largely data driven. The pervasive use of sensors and other high throughput measurement devices create the phenomena of "datafication" [7]: the properties of any physical research subject can be quantified and those quantification data transforms all research activities into data analysis. In this sense, data science becomes the glue to link all the research activities to enable a "chemistry of sciences".

Information from genetics, molecular activity, physiology, lifestyle and environmental impact is needed to understand human disease and to predict how it behaves. Modern medicine is the best example of multidisciplinary research where data science provides a common base for its systematic study. It is our mission that the DSI can provide a platform where such cross-disciplinary research can be fruitfully built.

3.2 Capability

The second mission of DSI is to strengthen Imperial College's data science research capability. This includes three key aspects: one is to support College's broad activities

in data analysis. The DSI will support the exchange and fertilisation of ideas across the College to lead to the development of new algorithms and methodologies. Second, is to put significant effort into building software infrastructure to support College's data-driven research. One example is the major effort led by the DSI core software team in building the eTRIKS system as the European standard platform for translational medicine research [2]. This system has been deployed and started to serve several research groups in the medical school in their translational medicine research. Software aside, and thirdly, the DSI will also build up its Big Data computing facilities. One recent development is the installation of a dedicated computing resource for medical informatics research, supported by UK Medical Research Council (MRC). With this support from the MRC, and with additional support from industry, the DSI will have soon the capacity of over 1000 machine cores, 50TB memory, and 3PB storage. Also the DSI has started to plan to work with some leading industrial partners to design and develop new specialised hardware for statistical computing. Such a R-machine, which will implement the statistical language R [8] on hardware, could be the DSI's major contribution to Big Data research and application.

3.3 Education

Education is clearly another mission. Data scientists have emerged from data science as a new profession. It is not straightforward profession. Data scientists certainly need to know data - and need to know how to process and manage data, in which case they need to be data engineers. Secondly, data scientists need to know the meaning of the data. They need to understand the context where data comes from and how it is to be applied. Domain knowledge of data is incredibly important. Therefore a data scientist needs to be an informatician. Finally, data scientists need to know how to use computing tools and technologies to extract knowledge from the data. Machine learning skills are therefore essential for an effective data scientist.

Data science is a fast growing discipline with data scientists today a highly sought after breed of professionals. The DSI will contribute to the College's efforts in training a new generation of data scientists by organising cross-faculty education activities such as regular research seminars, special training courses, and student projects. In particular, we will arrange various "data science challenge" competitions, organising industrial placements where students will be trained in our partner's organisation, and, over the long term, the DSI will establish its own education program to train postgraduate students. Some examples of the DSI's education efforts so far include having signed an agreement with Imperial College Press to publish a book series, called "New Frontiers of Data Science" (Figure 6), to turn the best PhD student theses from Imperial College into a series of data science monographs, and also we will soon launch our own Big Data challenge competition (Figure 7).

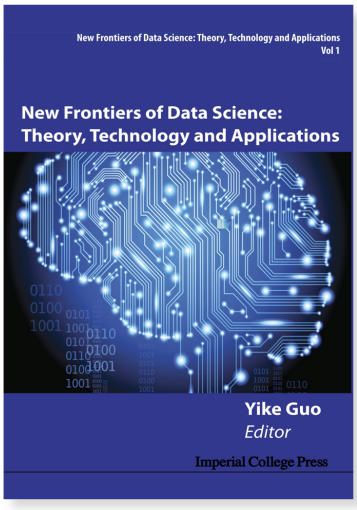

Figure 6: Front cover mock-up of the New Frontiers of Data Science book series, to be published by Imperial College Press.

Figure 7: Promotional poster for the Data Science Institute Summer Data Challenge competition. Photo used is by Doug8888 (Flickr) and licensed under CC BY 2.0.

3.4 Collaboration

A key mission of the DSI is collaboration. Imperial College has a very rich profile and extensive experience in data-driven research. For the launch of the DSI, we compiled and published a booklet that provides a taste of the depth and breadth of such activities at Imperial [3]. Such a rich culture and expertise in data science attracts collaborators from all over the world, from industrial to academic communities, to work with us. Such collaborations are clearly beneficial to further advance our research. The DSI acts as the focal point to facilitate, consolidate and coordinate such collaborations. We have developed several mechanisms to support collaborations, both internally and externally, including industry-funded labs, joint labs with academic partners, joint research programmes, and an academic visitor scheme.

In the months preceding the launch of the Institute, the

DSI established some exciting new partnerships. One notable achievement is the establishment of the Huawei Data Science and Innovation Lab, where Huawei has set up a fund to support Imperial and Huawei research collaborations in data-driven research. Huawei will also provide the DSI with a new state-of-the-art data centre based within the College. We have already identified a number of concrete projects, where Huawei will work with Imperial scientists to pursue research in the areas of high-performance learning algorithms, network information propagation models, and in-memory analytics technology. Another example of industrial collaboration is in the area of business analytics. The DSI is now setting up an advanced analytics centre within the Institute in collaboration with our Business School and one of the top professional services companies. This centre will allow us to participate in research within the context of real-world business applications. Also, this collaboration will lead to the enhancement of the DSI's capacity, in particular in the area of data visualisation.

Apart of collaborating with industry, the DSI is also devoting much effort to establishing close relationships with global academic partners. One example is an agreement with Zhejiang University of China to develop a joint applied data science research laboratory. This laboratory will have offices in both the UK and in China that will be the base for joint research projects, and also for doctoral training with a dedicated scholarship scheme. This is a very interesting model of academic collaboration and we have already identified a few areas as our initial focus, including neuroscience and complex system modelling.

3.5 Translation

Last but not least, is translation of research into practical applications. Imperial has already had strong emphasis on translating innovations into commercial successes. For data science, such a translation is even more challenging because the innovation comes not just from technology, but also from new business models. We are planning to work closely with the Imperial Business School on studying data economies and investigating new data-driven business models. Also, we will set up an "Idea Lab" to enable our staff and students to practice new ideas and develop prototypes for research that is at a stage that is close-to-market. We are building mechanisms for realising industrial on-ramps through a structured translation. For example, we are currently organising an Idea Lab project on neurofeedback-based content broadcasting, where a TV viewer's cognitive activity is monitored by electroencephalogram (EEG) and eye tracking, where this activity is sent to the cloud for real-time analysis. Such analysis will use each individual's emotion model to score their emotional status and, based on this, television content can be retrieved and broadcasted to influence a change in desired emotional status. This kind of innovative research combines neurotechnology, cloud computing, machine learning and cognitive science, and has real commercial and social value. Idea Lab projects such as this have clear potential for industrial translation to generate immediate impact.

4 Conclusions

The five missions of the DSI - catalysing multidisciplinary research, building capacity, educating data scientists, collaborating partners, and translating innovations - determine that the DSI needs a flexible structure, where laboratories can be formed as the base of collaborative research projects. The DSI's core research and development activities will focus on providing scientific and technological support to these projects. Our survey across Imperial College London demonstrates the value of data-driven activity and supporting technological and methodological research across a broad range of subject areas. Being an inclusive multidisciplinary institute is key, to act as the glue for all the main sections of a multidisciplinary academic institution. In this paper, we present our experiences in setting up such an institute, and hope that they can be used as a model for new research institutes to follow in the future.

We would like to conclude by presenting you a statement of the vision of the Data Science Institute at Imperial College: *"Making data the soil of great scientific innovation for a better world"*. This is the vision we all share at the Institute, and we look forward to putting data at the core of innovation and research, creating a *chemistry of sciences* with Big Data.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington DC, USA, pages 207-216, 1993.

[2] eTRIKS consortium. eTRIKS website, May 2014. `http://www.etriks.org`.

[3] Y. Guo, D. Johnson, and C. Griffiths. Big data for better science (booklet), Apr. 2014. `http://bit.ly/1nDPQCO`.

[4] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck. Metabolights - an open-access general-purpose repository for metabolomics studies and associated metadata. Nucleic Acids Research, 41(Database-Issue):781-786, 2013.

[5] HM Cabinet Office. Government launches one-stop shop for data, Jan. 2010. `http://bit.ly/SWUjnI`.

[6] Imperial College London. Data Science Institute website, May 2014. `http://www.imperial.ac.uk/data-science`.

[7] V. Mayer-Schönberger and K. Cukier. Big data a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, Boston, 2013.

[8] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for

Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.