Royal Statistical Society: Data Science Section

Ways to get in get in touch with the Data Science Section:

- Slack: https://rssdatascience.slack.com
- LinkedIN: https://www.linkedin.com/company-beta/11150048/
- Github: https://github.com/rssdatascience
- Twitter: https://twitter.com/RSS_DSS

The Industrialisation and Professionalisation of Data Science: 12 Questions

This document is available on Github: https://github.com/ivyleavedtoadflax/industrialisation/

Introduction

While the application of pattern recognition technology to large datasets has revolutionised the digital economy, outside the tech giants like Google and Amazon, Data Science (DS) is still a cottage industry with small teams of artisan DS crafting bespoke prototypes to their own standards. For DS is to fulfil its promise, it needs to industrialise and professionalise. By this we mean DS needs to develop a consensus about how organisations can evaluate, deploy and institutionalise DS to demonstrate sustainable value. Without this, decision makers in organisations that might benefit from DS will struggle to get the promised benefits and it risks being viewed as another IT fad which will pass into history. At the very least, without industrialisation the adoption of this new technology will be delayed, and the opportunity for UK to be at the heart of it may be lost.

The Royal Statistical Society (RSS) Data Science Section (DSS) is a forum to precipitate and participate in the conversation around industrialisation. As a first step we have proposed 12 key questions which should be addressed on the way to reaching a consensus across the various constituencies in the community of data science practitioners. This paper sets out these questions. They are not exhaustive. Nor are the questions listed in order of importance. They are simply a straw man to elicit debate. Are these the right questions? Have we missed anything obvious?

DS is for all sorts of organisations; businesses, government, not-for-profit and so on. Occasionally below we use the term 'business' but this is not intended to exclude the other use cases. It is just that some of the most well-known use cases of DS are currently in for-profit business, and it seemed natural to use that term there.

What does great data science look like?

DS is a new field, and its precise definition is yet to be generally agreed. DS integrates and adds to a spectrum of established disciplines (computer science, statistics, machine learning, business analytics, software engineering, etc). Likewise there is a spectrum of practitioners doing DS with a different balance of, and attaching a different importance to, these fields. This leads to debate over what is and is not real DS? Who is and is not a real Data Scientist?

The RSS promotes decision making informed by evidence and the effective use of statistics. It is a broad church that welcomes those with an interest in data and statistics from all backgrounds. But by forming the DSS, the RSS recognises that DS is more than just a branch of statistics. It seldom proves productive to assert the primacy of one aspect of DS over another. To be relevant, RSS DSS needs to bring different flavours of DS together, defining DS not as a wall to keep people out but as an exemplar to which all DS can aspire.

If we can't yet agree about what a data scientist is, we can perhaps discuss what great data science looks like. This is not obvious because DS is often difficult for a non-DS to understand. It incorporates ideas that few people in an organisation can review. The quality of DS may not be obvious from analytical results; they may look good but be unstable, non-generalisable, non scalable, difficult to refactor, etc. Since few DS themselves span all the constituent disciplines, is any one DS able to evaluate DS from all necessary standpoints?

What is perhaps required is a set of standards against which to evaluate DS. Examples of this might be:

- where statistical methods are used, these should be based on a rigorous understanding of the methodologies and the properties of the coefficients estimated;
- where results are to be deployed in an automated way, they should be based on best practice production software development techniques; a preference for simple solutions over complex ones, where models are interpretable and justifiable.

Key questions:

- By what criteria should we judge the quality and value of data science?
- How should a data scientist trade off these criteria in determining the course of the project?

What does a good data science workflow look like?

Data Scientists have to strike an unusual balance between discovering new things (science) and deploying them rapidly (engineering). They have to be simultaneously innovative and insightful while delivering reproducible results and production quality code at pace. Conventionally the research and development working practices necessary for exploration are incompatible from the way that production code is built.

Often DS projects are ill-defined throughout and you might not be sure that your application can even be built successfully until a late stage. DS must adopt approaches that reassure users that the final product will conform to the required standards even though during the build process visibility is low. This can be done by determining a workflow which will lead to design choices being made in a consistent and rational way, which requires a fine balance between controls and the freedom to build intellectual momentum.

Key questions:

- How do we do data science to deliver innovation, quality, insight and pace?
- How do we balance the sometimes competing need for exploration and production?
- Under what circumstances should Data Science be methods led? Data led? Science led?

What kind of problems can be addressed by data science?

To a DS, every problem looks like a DS problem. But as a community we are learning from experience that some problem areas have been fruitful (fraud detection, recommendation) some have been less so (these tend to be less publicised). Success can be due to factors other than the technical challenge of identifying patterns or productionising insight.

Choosing the right problem to tackle can be critical for the success of DS in an organisation. And being able to show to decision makers evidence that a particular use case has been addressed before successfully is crucial to give them the confidence to invest. This is hard in the absence of a decent corpus of business-school-style case studies. While there exist plenty of vendor-generated use cases, and the occasional conference slide deck, there are few solid and convincing studies of the process of implementing DS in organisations. This lack of shared experience is slowing the adoption of DS technologies.

Key questions:

- Is there a useful taxonomy of data science problem areas?
- What is it about those problem classes that make them suitable?
- Are there case studies of data science projects that have been tractable and valuable?

What are the characteristics of the ideal data scientist?

The characteristics of the ideal 'unicorn' data scientist have been widely mythologised. As with definitions of data science itself, these characteristics often bear a striking resemblance to whoever is writing them down.

Clearly a DS is someone with the right skills and aptitudes to do DS, which must include sufficient:

- Mathematics to understand the methodologies;
- Statistics to understand the properties of the estimators;
- Computer science to grasp the concepts of data structures and distributed computation;
- Software engineering to build and deploy DS applications;
- Analytics and visualisation to make sense of results;
- Commercial modelling experience to convert business problems into tractable DS problems;
- Organisational awareness and entrepreneurialism to get things done;
- Communication and interpersonal skills to interact with non-technical colleagues;
- etc. . .

While it is possible for one person to have all these skills, and one of our objectives is to help create a large cohort of these people over time, it has been widely observed that there are currently fewer people with the full range of skills than there are jobs requiring them.

Additionally you have to consider the type and level of qualification. Quantitative undergraduate degrees are generally agreed to be ideal, but views differ on the value of different levels and types of postgraduate education. Sometimes it seems that firms talk about the education level of their DS as a matter of prestige rather than as a result of a serious evaluation of what they need in order to be great DS.

Organisations cannot be expected to build a DS strategy if they believe that they will have to invest in some mythical transitory creature; they must have the confidence to build DS capability from the talent available to them now, and that may mean starting with teams with a range of complementary skills. What are the key skills, aptitudes and personality traits to start with, how can you put them together into a balanced team and how can that team be made to work together productively? Again the solutions to this conundrum are likely to differ depending upon the type of DS you will be doing and the kind of organisation you are. And this will change over time; as DS industrialises, tasks are standardised and functions are specialised, so the makeup and structure of the DS team will change.

Key questions:

- What skills, aptitudes and personality traits should a data scientist have?
- Are there different kinds of data science that require different kinds of data scientist?
- Given that you won't necessarily find all the right attributes in a single person, what makes a good data science team?

How should an organisation start a data science function?

Building a data science team is a leap in the dark for many organisations, particularly because the hiring manager may have no experience of the field. Often the decision to hire is based on no more than a boardroom conversation or an article in the Harvard Business Review. How should an organisation go about this?

The first steps can be difficult. It may take time to get the first DS project into production and not everyone in an organisation can be guaranteed to be supportive. But getting the first steps right is an important determinant of future adoption. And time may be short, particularly in industries facing a challenge from technology disruptors.

Typically there are three types of questions that organisations face.

- Strategy - What opportunity or challenge is an organisation hoping to address with data science? The kind of team you want to build will depend upon what you want to do with it.

- Scale - How much resource should be committed, given that returns to scale are not necessarily linear? You may just want to dip your toe in to begin with, but if your team is too small it may not be able to prove the case.
- Model - Should the team be permanent, contractor or outsourced, and what are the implications of each model? Cost, flexibility and time to market are important considerations, but so are continuity, quality and intellectual property.

There have been a number of high profile announcements (of huge investments and teams) not all of which have been successful. Sometimes it looks like these announcements are made for reasons of prestige rather than productivity. Yet without a consensus narrative around the way to start and build Data Science capability, it is possible that organisations could become disillusioned with DS and dismiss it as another IT fashion that has failed to deliver.

Key questions:

- How should organisations build a data science strategy?
- How should they build a team?
- What problems should they attempt to solve first?

How should data science fit into the structure of an organisation?

The cross-functional nature of DS becomes apparent when organisations think about where in their structure to put their new team of DS and how to manage them. For example, should DS sit in an IT function or within the main business functions? DS emphasis on the meaning of data makes them like analysts, but they tend to work more like software developers in building code to be run in production.

Then there is the balance that must be struck between grouping DS into a single centralised function, which encourages technical excellence, and embedding DS within different front-line functions, which ensures they understand the commercial task.

Furthermore, since DS teams tend to be small, should they be part of a large function, which would given them organisational heft but might subsume them, or should they work independently, which could give them visibility but might make it hard for them to effect change.

There is no straightforward universal answer to these questions because, as an integrative technology, DS naturally cuts across functional and hierarchical structures. But an awareness of the issues raised by each choice will be crucial to the successful deployment of DS in an organisation.

Key questions:

- In what function should data science sit: IT, Data, Business?
- Should data science be a centralised function, emphasising technical expertise, or embedded within frontline functions, emphasising product knowledge?
- What is the right internal structure for a team of data scientists?
- How should data scientists interact with other existing technical functions. In what sense is data science similar and different to these functions?

How should business practices change to make a success of data science?

The most obvious way to implement DS is to try to apply DS techniques to existing business practices. This can incrementally improve performance, but it is not harnessing the full opportunity of data and DS. Instead, businesses might do well to think about how they might adapt their business to DS. While DS is a technology it is not just a support function: it can become a core front-line product. Think how Amazon has moved from retailing to recommendation. They didn't achieve this by thinking about DS as an add-on to optimise their existing business.

Key questions:

- In what way should an organisation adapt to optimise the impact of data science?
- In particular, what changes in organisational behaviour are implied by data science?
- What resources (incl data availability) are required to deliver data science applications?

What do executives and managers need to learn about data science?

If DS is to become a core product of a business, executives and business managers will need to understand it in the same way as they need to understand their core product. Decisions about DS might have just as much impact as whether to launch a new product or to change pricing. In organisations that are good at data science the specifics are not delegated to back-office technicians; managers are able to contribute to the decision as peers of their tech colleagues.

And yet managers can be to be ill-equipped to do this. Businesses, particularly large corporations, have historically rewarded process management above technical prowess. In some organisations, displaying technical knowledge implies that one has yet to graduate to a generalist managerial role. If DS is to be front and centre of business strategy, managers will not just have to acquire some new information, but relearning how to think technically and get over the stigma of being thought a technician (and therefore 'not commercial').

So what should managers actually know? Obviously those directly managing the DS team must be expert DS themselves. But not everyone in the entire organisation can be a DS, even in a data-driven company. For example, how much should a product manager understand about what the DS team is doing on her project? And how much should the IT Director or an HR manager understand? And crucially, how what should the CEO know about the data science that serves their customers and powers their profitability?

Although a manager's role will remain to ask the right questions and to hold DS to account, knowing what questions to ask and being able to critically evaluate the answer in the context of the commercial challenge requires considerable domain knowledge. How can they decide if an accuracy metric is plausible and persuasive? Or distinguish between a promising new DS application and snake oil? How can they specify business requirements so that their DS use case is successful? Without this, how can executives make the right decisions?

Developing a curriculum for general business managers is as important as developing one for data scientists themselves and will require the engagement of business schools, not just statistics and computer science faculties.

Key questions:

- As the products of companies increasingly become the data it generates, can managers continue to remain generalists?
- What should managers know? How can they find out?

How can an organisation build a coherent data science capability from a collection of data science projects?

To generate sustainable value from DS, organisations need to move from the initial excitement of building and deploying cool DS products, to maintaining and improving those products, to rolling out those technologies to other use cases in the organisation, to creating a DS capability which has value beyond the value of the products themselves. A car company is not valued by the expected sales of their current range, but by their capability to develop and sell cars in the future.

This is a challenge because the field is changing so fast. Applications that were built last year may be old fashioned or even redundant. DS that developed the required domain knowledge or mission-critical methodology may have moved on. Foundational technologies may have become peripheral. For example the speed at which the archetypal big data technology, Hadoop Mapreduce has been replaced by Apache Spark is remarkable. Likewise the adoption of deep learning methodologies. How can an organisation build a

capability when everything is changing so fast? In particular how can they take advantage of open source software when procurement is set up for over 5-year-plus life cycles?

Key questions:

- How should an organisation manage the choice of technology and methodology in a rapidly changing environment?
- How should organisations maintain the inevitable legacy technology created as a data science function develops?
- How should businesses manage operational risk as profitability becomes increasingly reliant on a set of data science applications?
- How should an organisation change its approach to data science as the technology becomes established?

What career paths are available to a data scientist?

There are all sorts of forecasts about the relative supply and demand of data scientists in the economy and how that might change over the next five years. The consensus is that there are fewer than are needed, so the DS industry needs to address how to create more and lose fewer, and one part of that is to agree a career development plan.

The best DS are passionate autodidacts and this has enabled the industry to grow to its current state without a formal career structure. But as DS matures, and DS become commonplace in businesses and government, we cannot rely solely on the natural supply of extraordinarily individuals: We must find ways of creating them.

RSS DSS may have a role here, to set career expectations and provide some guidance on career development and accreditation.

Key questions:

- What backgrounds would data scientists typically come from?
- What is an appropriate way to develop a pool of data scientists?
- How can institutions (incl. universities and government) develop programmes to support? And what would the curriculum include?
- How can data scientists progress in their careers? Is there an alternative to the management track?
- Is there a role for an industry body to accredit data scientists and represent their interests?

How can data scientists measure the value they create?

Each organisation embarking on DS strategy will have different objectives for and expectations of the effect of DS on their business. For a strategy to be expanded, DS will need to show that it is meeting those expectations. When those objectives are directly linked to observable quantitative outcomes this is relatively straightforward. For example, if following the implementation of a recommendation system sales increase, it may be reasonable to attribute this value to DS. But if DS is used to build less tangible or less immediate value, more sophisticated value measurement methodologies may need to be employed. For example, if DS was to be used by a bank to identify financially vulnerable customers (and maybe then not offer them an additional loan), then this could result in lower sales for the organisation in the advancement of a citizenship agenda. DS may have generated huge value which cannot be measured directly or immediately.

This matters because, as a new discipline DS must prove its worth in a way not expected of more established functions such as marketing or finance. And given the subjectivity of the choice of methodologies and the strategic nature of the competition for resources, consensus among DS about how to measure value will smooth the path to adoption.

It is also important to DS themselves because if they are to be generating all this value,they might feel they should participate in a fair proportion of it. As noted above, the DS skill set is both broad and deep, and

keeping up with the field requires a constant investment of time and effort. Yet it has been observed that, particularly in UK, it is difficult for technical professionals to enjoy the benefit of their skills and innovation. When firms do well, it is general management whose share options swell, with the engineers that equally create that value seldom enjoying commensurate increases in salary. Information products driven by DS offer the unusual opportunity for technical professionals to demonstrate their direct personal impact on business value. The linked opportunity to build a comfortable lifestyle from excellence in DS may even in time attract graduates away from the established professions such as investment banking, consulting and the law.

Key question:

- How can data scientists measure and benefit from the value they create?

What is a data scientist's responsibility to wider society?

The term DS implies that, as a science, it is pursuing knowledge for its own sake. Scientists often view their work as morally neutral and perhaps DS can be accused (at least historically) of having had a similar attitude. After all the average DS is primarily interested in uncovering the power of hidden patterns and can be less focussed on wider frameworks of evaluation. The limitations of society's understanding of the impact of DS has allowed the field to progress unchecked until recently. But now the public discourse has caught up. While DS themselves often remain uninterested in issues of data privacy, data ownership, algorithmic ethics, the newspaper editorials are fascinated by the idea that the internet has stolen your data and is spying on you. This has led to a low quality of public debate where hysterical tendentious headlines dominate debate. If we lose public confidence and strong regulation is imposed, this could threaten the tremendous value that DS promises to create.

If the debate is currently simplistic, it is only a matter of time before the public debate extends to more complex areas such as perpetuating biases, where for example a career-recommendation tool might suggest a home-maker career to a woman. A recommendation engine is a descriptive tool only masquerading as a normative one, but the public may not understand this, and moreover may not accept it. In this kind of debate there are all sorts of trade offs and subtleties which only DS can communicate. If DS is to grow it needs to leave Eden, and engage in the conversation wholeheartedly or it will end up the subject of uneducated opinion.

In addition to engaging in the public discourse, DS must put their house in order and show that they do behave ethically with their data. But what guidelines exist are mostly philosophical, legal, abstract documents that are hard to put into practice. For example, in data privacy there is the concept of a "motivated intruder" who could acquire your data, combine it with an arbitrary set of third-party data, and thereby uncover personally identifiable information. This is a nice thought experiment, but it is hard to think of a DS process to test whether a data set is exposed to motivated intruder risk. DS is a practical field and needs practical guidelines designed to help them behave well, rather than yardsticks by which to judge if they have behaved badly.

All of this relates back to professionalisation and to establishing principles of good practice. Given the present attention from the media and in government, we should expect regulators to intervene if the DS community does not take action to address these kinds of issues themselves.

Key questions:

- Should there be an explicit data science code of ethics and behaviour?
- What are appropriate and shared practical guidelines for using and storing data?