Industrialisation and Professionalisation of Data Science - Working Paper: December 2018

Outcomes from the RSS Data Science Section workshops held in collaboration with the Institute and Faculty of Actuaries

Author: Leone Wardman (RSS Data Science Section committee)

Contact: datascience@rss.org.uk

Contents

Introduction

The RSS Data Science Section (DSS) was formed in 2017 to address the need for best practice and professional standards in the growing field of data science. In 2017 we formulated 12 questions that would help to industrialise and professionalise data science. The questions and rationale can be viewed here [1].

In 2018 the DSS chose four of these questions and in collaboration with the Institute and Faculty of Actuaries hosted a series of four workshops designed to collect feedback from data science professionals. This paper compiles the feedback from all four workshops, including where further questions were raised. General findings are discussed in the main body and a series of agreed principles and practices for each question are listed in Annex 1 through 4.

Key Findings

- Best practice for data science is dependent on the industry, organisational design, historical analytical workflows and availability of skills within data science teams
- The data science professionals we consulted broadly agreed several high-level principles and practices around workflows, ethics, what execs need to know and how to effectively place data science within the structure of an organisation.
- While agreement for good practice was found across many aspects, some of the more complex issues require more thought and input from the profession to resolve.

Next steps

The agreed principles and practices emerging from the workshop will be developed into RSS guidance. Where possible, unresolved issues will be addressed by the committee to develop formal guidance, for example through further consultation, research or peer review. There will be two main outputs from this work:

- Due to the need to embed ethics into the workflow, the practices for workflow and ethics will be developed into guidance in collaboration with the IFoA through a joint working party. This guidance will summarise the findings into a short 'cheat-sheet' for imbedding ethics that contains links to the full set of principles, further information and resources.
- The other two questions will be developed into guidance by the DSS. These may be better delivered as short videos. The committee thought guidance on what execs need to know could be further peer-reviewed by C-grade execs with experience of introducing and working with data science. The guidance on organisational structure could be further developed by including case-studies / examples of the different types of models and where they have been successful, or otherwise.

[1] https://github.com/rssdatascience/industrialisation/blob/master/industrialisation.md

Workshop aims and questions addressed

The aim of the workshops was to consult and gather feedback from senior data science professionals across the UK, and raise the profile of the issues with the broader data science profession through a series of talks and panel discussions. The four questions selected for the workshops were:

1. **What does a good data science workflow look like?**
    - 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?
2. **What is a data scientist's responsibility to wider society?**
    - Should there be an explicit data science code of ethics and behaviour?
    - What are appropriate and shared practical guidelines for using and storing data?
3. **How should data science fit into the structure of an organisation?**
    - 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?
4. **What do executives and managers need to know about data science?**
    - 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?

Workshop format

The workshops were held across the UK (Bristol, Manchester, London, Edinburgh) with participants from a broad range of industries and managerial levels, including practitioners, academics, senior managers CEOs/CDOs and consultants.

At the first three workshops there was a morning session where invited guests were asked to debate and provide responses to each of the four questions and an open afternoon session which presented the findings from the morning and invited further comment from the audience in a panel discussion. The final workshop focused more on the actuarial perspective and had just one session which focused on two of the first two questions (workflow and responsibility).

At the first two workshops (Bristol and Manchester) participants were given the same discussion material which consisted of the four questions as articulated in the '12 questions' paper. The feedback gathered was compared and as responses were reasonably consistent between the two groups the findings (along with unanswered questions) were summarised and presented to participants at the

London workshop. Likewise, feedback from London was added and presented at Edinburgh for discussion.

Findings

In answering the four questions several common themes emerged across the workshops. These have been formulated into **principles** (what we should do) and **practices** (how we do it) and grouped into six broad **elements** of data science:

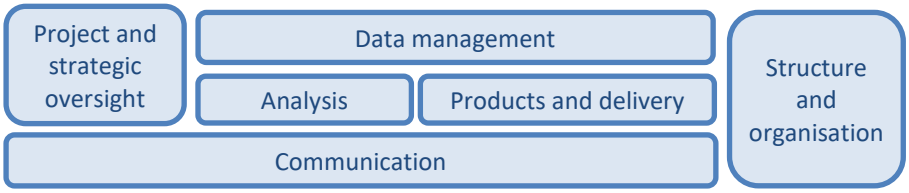

Figure 1: Broad elements of data science identified in previous workshops

The participants agreed there is no single 'good' data science workflow, structure, ethical governance or tool-kit for executives because the ways data science will emerge and develop is industry specific and will depend on:

- The maturity and size of the organisation, and where it is on its journey to becoming data driven
- The structure of the data science team within the organisation
- Who is included in the data science team and their skills (e.g. number of staff, variety of disciplinary backgrounds, level of technical vs soft expertise)
- The historical analysis workflow within the organisation
- Whether the data science implemented is primarily exploratory or focused on answering pre-defined questions/business objectives

For each question some agreed principles and practices emerged – the discussion around these is summarised in the following sections and the principles/practices are tabled in Annex 1 through 4. They represent a non-exhaustive list of options that might be appropriate depending on the factors listed above.

Question one: What does a good data science workflow look like?

There was debate whether data science should always start with a question, and whether it should always have a defined business benefit. Participants identified two different types of workflow:

- **Exploratory**: **research and development** - may focus on assessing potential value in existing operational data, or investigating new data sources. More relevant to organisations with large volumes of untapped / unmapped data, or in early stages of introducing data science

- **Focused**: **specified problem and outcome** – may focus on business problems, inefficiencies, automation of tasks. Appropriate for organisations where data science is the business, or organisations further along the journey to being data-driven / capable where data science is expected to provide measurable value

Participants discussed how project workflow (work is done) might translate into team workflow (who does what). There was agreement that:

- Workflow design needs to be appropriate for the organisational structure.
- An appropriate workflow design depends on the people and skills available in the team, and the way work flows in other teams in the organisation.
- Elements of project management should be incorporated, but these could be light touch or Agile to help keep flexibility and pace.
- When using exploratory workflows, it was felt it was important to have some basic questions, to challenge findings and ensure ethics are considered.
- Skills need to be embedded into the organisation so that the use of data science solutions is not dependent on the individuals / team designing them.

Two issues raised remained unresolved:

- Should the innovation and blue-sky data science projects workflow be different from the workflow for a standard data science project?
- If a good data science workflow starts with a good question, what makes a good data science question?

The agreed principles and practices for a good data science workflow are in Annex 1.

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

Participants felt data scientists have multiple and multifaceted responsibilities to wider society. These include responsibilities covered by existing data and business ethics, such as respecting privacy and upholding professional standards. Therefore, data science ethics might be thought as sitting between these, or perhaps encompassing them with some additional challenges to consider. There was a discussion about the need for a separate data analytics ethical code alongside the existing ethical codes for Data Ethics and Business Ethics.

There was a lot of discussion on algorithmic responsibility and who was ultimately responsible for the outcomes of automated AI. Generally, the participants agreed that executive management are responsible for taking algorithms into production, but it is up to data scientists to adequately communicate the methods and results so that execs can make informed decisions. We need to ask the right questions at all levels, discuss bias, and understand who might be harmed or disadvantaged.

There was discussion around bias and the thought that it was unavoidable, but the onus was on data scientists to understand the bias, mitigate where possible and be transparent about it.

The groups also identified the following techniques that could be used to ensure the highest standard of ethics within the profession. These included:

- Peer and bias review
- CPD in ethics
- Conference tracks focusing on ethics and responsibility
- Increasing diversity in data science teams
- An ethical toolbox for data scientists to use

Several issues raised remained unresolved:

- Should data science ethics draw on industry specific safeguards and boundaries, e.g. targeting problem gamblers, to help explore and address some of the ethical considerations unique to data science?
- Are we as data scientists making things better?
- Are there any use cases for holding ethical standards around automated decisions?

The common ethical principles and practices identified for acting responsibly are in Annex 2.

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

It was recognised that the best structure will vary depending on the specific circumstances of the organisation. Several different models were identified:

> **The Consultancy Model:** the individuals or team act as internal consultants to business units, providing expertise and advice on how data science can add value in their unit. They should have or develop sector specific knowledge.
>
> **The R&D Model:** the individual or team focuses on innovative uses of data science and advanced analytics. In this model the team are under less pressure to deliver an immediate return on investment.
>
> **The Hub and Spoke Model:** there is a core data science team, complemented by individuals embedded within other business units.
>
> **The Federated Model:** Analytics teams are deployed within the business. In this model there is no core team or hub, but infrastructure like data management and storage may be centralised.

The Centralised Model: data science as a corporate resource, includes governance, management, analysis and project service teams. Builds a core of expertise, but neglects upskilling of workforce.

The hub and spoke model appeared to be the most commonly adopted approach by participants' organisations, and was consistently favoured. Examples of the federated model were also evident, with support for using this approach alongside some centralised functions for IT and data management.

Participants also discussed reporting lines. Examples included reporting into IT or a member of the C-suite (CDO, CIO, CEO or COO). A suggestion was to report to the CFO. Some participants felt reporting into IT may limit the ability to work effectively with the wider business.

Several issues raised remained unresolved:

- To be successful, data science teams need to work across the whole business, does being part of IT limit that ability to do that effectively?
- What is the best way for internal customers to engage with Data Science teams? What is the Data Science 'front door'?
- How should the structure of the team evolve as the organisation progresses towards data-led maturity?
- Should data scientists get out more?!". i.e. how much of their time should they spend 'embedded' in customer/stakeholder groups, vs back in their skills hub?
- What are the case studies we can draw from - are data scientists currently embedded and integrated into their organisations or are they isolated and siloed? Is the structure of the team different depending on the industry? What other factors influence how teams are set up?

The agreed principles and practices for structure and organisation are in Annex 3.

Question four: What do executives and managers need to know about data science?

The participants wanted executives and managers to know and understand enough about data science to make good decisions. This included:

- knowledge (e.g. what it is, who does it)
- skills (e.g. data literacy, formulating analytical questions)
- mindset (focus on problems and value, data as an asset, acknowledges limitations in knowledge)

There was emphasis on the importance of mindset and agreement that knowledge and skills were broad and high level. Also important, was asking the right questions as the organisation moved through the data science journey from inception to maturity.

Issues that remained unresolved:

- Which executives and managers need to understand data science?
- Does having a CDO change what other executives and managers need to learn?
- How much information is too much information for executives and managers to have?
- There was some difference in opinion around how much technical knowledge execs responsible for data science need to know, for example in Bristol it was thought they don't need to know what a p-value is, while in Manchester there was discussion about understanding error and different types of models. What other technical concepts would it be helpful to understand at a high level?

The common principles and practices of data science for execs to foster are in Annex 4.

Annex 1: Principles and practices for good data science workflows

Element	Principles - *What*	Practices - *How*	*Why?*
Project and strategic oversight Needs Questions Exploration	• Start with a **clear question**. Focused: articulate the business need and potential value. Exploratory: articulate the strategic aims, set goals • Design projects around **milestones and outcomes** (rather than processes and schedules), and have a **clear end-point** • **Be customer focused**, involve and partner with customers, stakeholders • Embed ethics and legal considerations • Create insight that leads to actionable change and **return on investment** • Be **flexible**, balance short and long-term activities	• Ensure questions /goals can be answered with data that you have, or can acquire • Scope with customers - help them to define the problem, goals and requirements • Build ethical assessment and check points into project plans; consider biases etc. at design / planning stage • Assess strategic alignment to business, readiness of business to implement • Be aware of changing priorities	Clear questions / goals ensure workflows deliver value. Focusing on outcomes and milestones is more flexible and provides natural review points (ability to fail early). Involving customers leads to clearer requirements, better outcomes and buy-in from wider business. Embedding ethics ensures it is not just a 'tag -on', becomes part of culture. Actionable change and ROI aids buy-in from execs, budget holders.

Element	Principles - *What*	Practices - *How*	*Why?*
Data management Acquire Assess Wrangle	• Fully **assess for purpose** and review project feasibility: Focused - assess against requirements; Exploratory - assess against strategic goals • Expect to spend majority of time on data provenance, error checking, wrangling • **Harness domain knowledge** of data generators / custodians / analysts to build understanding • **Be curious** – explore data fully, spot other opportunities • **Invest** in **data quality** and **infrastructure**, consider future uses	• Establish the what, where, how of data sources (meta data) • Quality check and articulate limitations, ask is the data good enough? Consider collecting better / additional data • Clean and enhance, eg. errors, inconsistencies, missing • Strong collaborations with domain experts, embedded working • Think about alternate uses	Success of data science is dependent on having the right data. Data scientists may not have domain knowledge - must therefore communicate with those closest to the data. Often using messy operational data, not designed for analysis. To become data-driven requires treating data as an asset.
Analysis Algorithms Prototypes	• Clarify requirements and expectations • Apply **statistical rigour** - but do not be rigid • Be transparent about limitations and uncertainty	• Articulate what models are meant to do (before training) • Validate models • Openly discuss and report error, bias	Data science should be underpinned by sound statistical methods.

Element	Principles - *What*	Practices - *How*	*Why?*
Products and delivery Deployment Automation Pipelines	• Build simple but flexible solutions • **Productionise outputs**, move quickly from prototypes to production • **Calibrate** to business resources and skills • Maintain human oversight	• Rapid iterations, agile working, sprints, build in scalability • Engage with cross functional teams (eg. IT, product owners) • Test driven development • Agree review processes	Simple solutions are more transparent, adaptable. Productionising refines needs, finds issues early. Solution needs to be accessible to the people who will use it, and maintained over time.
Communicate Report Share Disseminate	• Communicate the results, insight and value to the business • **Build public trust;** be transparent about limitations**,** use of data, methods, ethics, privacy • **Manage expectations**, with customers and executives • **Encourage challenge** from everyone in the workflow	• Explain methods and results in plain language • Articulate benefits in real terms (time, costs, quality etc.) • Highlight any limitations • Be realistic about delivery and outcomes	Helps business and execs understand the value of data science, buy-in. Business can spot further opportunities, better articulate problems Strong partnerships and working relationships

Annex 2: Principles and practices for ethically responsible data science

Element	Principles - *What*	Practices - *How*	*Why?*
Project and strategic oversight	• **Embed** ethics into the workflow • Ensure **checks and balances** are in place for the business	• Understand the potential harm of projects through risk assessment • Have ethical governance procedures in place • Provide execs enough information to make decisions • Educate the workforce	Embedding ethics ensures it is not just a 'tag -on', becomes part of culture. Highlights risks for customer / public relations. Limits risk of unacceptable ethical issue at end of project. Guides workforce towards consistent working practices.
Data management Acquire Assess Wrangle	• Use '**green data**' (ethically sourced) • Uphold **Data Ethics** principles (eg. privacy, security) • Engage with **ethical bodies** (ICO, ISO27001)	• Understand the data source, how it was collected, whether there was legal / informed consent • Keep data secure, consider the impact of deriving demographics or linking with other data	General good practice for any data use. Meets legal obligations.

Element	Principles - *What*	Practices - *How*	*Why?*
Analysis Algorithms Prototypes	• Consider the potential **impact** of models **on decisions**, especially in relation to people • Favour **explainable models** if they provide similar accuracy	• Understand biases, errors, assumptions and risks inherent in predictive modelling • Invite peer and bias review • Articulate the consequences for different groups	Minimises the risk of negative consequences for individuals, society.
Products and delivery Deployment Automation Pipelines	• Maintain **human oversight** in implementation of automated solutions	• Implement model governance • Agree responsibility for models in production (approval, reviews, longer term QA)	Monitors potential harm, bias over time. Changes in data can impact relative bias / consequences.
Communicate Report Share Disseminate	• Encourage transparent working • Communicate with the public and **build trust**	• Publish and share methods and the limitations • Honestly report results, outcomes, bias, uncertainty • Articulate ways data are used and the benefits to customers / wider public	Empowers individuals to make decisions about the data they give. Invites peer review and scrutiny. Breaks down myths / concerns about data science.

Annex 3: Principles and practices for data science organisational structure and design

Element	Principles - *What*	Practices - *How*	*Why?*
Structure and organisation Reporting lines Teams People	• Be driven by and have **buy-in from senior executives** and managers • Enable **strong links** to the **business areas** generating data science needs and generating / controlling the data • Enable development of domain knowledge • **Adapt** organisational structure **over time** to meet the changing needs of the business	• Choose a useful reporting line, consider how the function will be financed • Apply good change management practices • Embed or second data scientists into the business • Set up communities of practice to help transfer knowledge	The responsible exec should champion data science. Understanding the business enables more effective and value generating outcomes across all elements of data science.

Annex 4: Principles and practices for what execs need to know about data science

Element	Principles - *What*	Practices - *How*	*Why?*
Project and strategic oversight	• Understand the **strategic direction** for data science and how it links to wider business goals • Be problem focused • Understand **how** data science can **add value** and **what** sorts of **problems are solvable** (and unsolvable)	• Mindset: That success can be early failure – in other words, if fail, fail early • Knowledge: Be aware of good examples and use-cases • Knowledge: Understand how data science compliments other analysis •	Know which projects to endorse and prioritise.

Element	Principles - *What*	Practices - *How*	*Why?*
Data management Acquire Assess Wrangle	• Treat data as an asset • Champion good data governance and ethics • Be data savvy	• Mindset: Invest in data quality and building appropriate infrastructure • Mindset: Invest in identifying any risks of harm in use of data • Knowledge: Understand broad concepts of data privacy, legal framework • Knowledge: 90% of data science is data preparation (cleaning, formatting, merging etc) • Skills: Maintain or develop own fundamental data literacy	Good data science needs good data. Good data governance can provide savings in the future by reducing need for wrangling. Need to know enough about data to make ethical and business decisions about its use.
Analysis Algorithms Prototypes	• Ensure the **tools** and **environment** needed for good data science are **available** • Encourage and **support scientific rigour** within data science • Support use of models where **risks are balanced** with expected benefits	• Mindset: have humility around extent of own knowledge and skills • Knowledge: the steps and timelines in a data science project • Knowledge: difference between rule based models (like RPA) and probabilistic models – with a distinction between linear (e.g. logistic) and non-linear (e.g. random forests). • Skills: can critically evaluate solutions at a high level – data quality, do models make sense, is error acceptable, what are the risks? • Skills: ability to reason in the presence of uncertainty and evaluate alternatives	Execs need to be able to ask the right questions about analysis to decide which solutions to take forward, assess risks, and justify their decisions (to the wider business, boards, CEO etc.)

Element	Principles - *What*	Practices - *How*	*Why?*
Products and delivery Deployment Automation Pipelines	• Lead and endorse the **cultural shift** towards **data-led decisions** (automated) • Know **when to ignore** an automated decision	• Mindset: invest in the people and skills required to implement data science products • Knowledge: understand risks of products, and how they work	Investment in infrastructure is required to implement robust analytical pipelines. Need to understand when automated solutions go beyond appetite for risk, or are outside ethical principles
Communicate Report Share Disseminate	• Articulate **benefits**, while acknowledging **limitations** • Help **manage expectations** across the business	• Mindset: Recognise value, discourage hype	Execs support will help get buy-in from wider business.
Structure and organisation Reporting lines Teams People	• Align structure with the current needs and strategic objectives of the business • **Adapt** structures and models as data science matures	• Mindset: View as a continuous evolution and journey • Knowledge: who data scientists are and what they do • Knowledge: the skills required in data science	Successful data science teams will have a range of skills that align to business needs, and be positioned within the organisation where they can identify opportunities.