Surveying the Data Science Skills Landscape in UK Government

What are the skills and capabilities needed for Data Science? Do they exist in UK Government?

The Data Science discipline draws from a broad set of skills and capabilities, from the IT, informatics, analytical and business expertise domains. Accordingly, there is no easy single definition of a 'Data Scientist', nor of the specific set of skills and capabilities they require to practice their craft.

Attempts have been made however, to define what sorts of skills and capabilities might be important and given priority. One source of evidence is the O'Reilly paper 'Analyzing the Analysers'[1]. This reports on a 2012 survey of 250 members of the Data Science community associated with a number of research and private sector organisations. The respondents were presented with a set of 22 skills and 11 roles and were asked to rank their skills and score their self-identification (strongly agree / agree / neutral etc.) with the roles described. The rank orders of the skills and the self-identification of the professions were then analysed to reveal the cluster patterns in the results: what skills tended to be ranked together and what self-identities were commonly seen grouped as high, or low preferences. Methodologically, this analysis approach (matrix factorisation) is from the same family of data science techniques that underpins the recommendation engines of Amazon and Netflix, taking user-assigned review scores and revealing the underlying patterns.

In December 2015 the Big Data and the GSS Capability teams re-ran this survey, mirroring the approach described above to data collection and analysis. Responses were solicited from Data Science community of interest, the GSS Data Science recruits and from the wider digital and policy professions to attempt to recreate the broad scope of the original survey. In total 290 valid responses were received, making this a valuable source of evidence and revealing some interesting patterns.

Who are the Data Science practitioners?

One of the first results from the survey is to explore who in government already identifies themselves as a Data Scientist. In the survey, respondents were permitted to nominate more than one profession, which explains why the results sum to more than the 290 respondents. What can be seen from the results is that two-thirds of Operational Research analysts and close to half of government statisticians already identify themselves as Data Scientists.

[1] http://cdn.oreillystatic.com/oreilly/radarreport/0636920029014/Analyzing_the_Analyzers.pdf

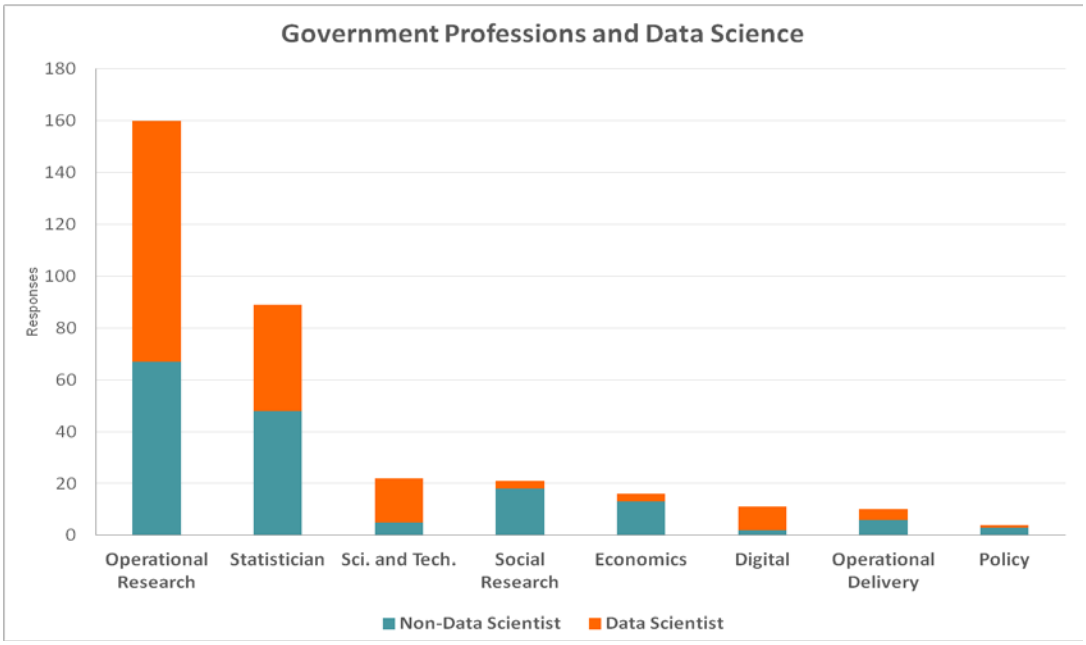

Figure 1 - Self-identified data scientists, by government professional group

How else do practitioners self-identify?

In the 2012 survey, the cluster analysis of the respondent's self-identification revealed the following four archetypes:

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

Figure 2 - Self-identity groups from the 2012 survey

The self-identification question asked of Government Analysts in 2015 used an identical approach but without the 'Hacker' category.

Data Analyst	Statistician		
Data Businessperson	Leader	Businessperson	Entrepreneur
Data Generalist	Jack of All Trades	Developer	Engineer
Data Researcher	Scientist	Researcher	

Figure 3 - Self-identity groups from the 2015 survey

The first category – which has been labelled Data Analyst – was an extreme group, where the cluster analysis revealed a particularly strong self-identity with Statistician, with no other self-identity coming even close. The identities of Engineer and Entrepreneur did not have strong affinities and there was almost no affinity with the Artist self-identity, which has therefore been excluded from the list.

What are the skills?

The approach to exploring skills is to ask respondents to rank the list of 22 skills in their own introspectively-determined order of ability. Although this means that skills can't be compared between individuals, the relative order of the skills and the sets that appear together, either as being present or absent, in people's self-assessment is revealing. In the 2012 survey, the following skill groups were identified:

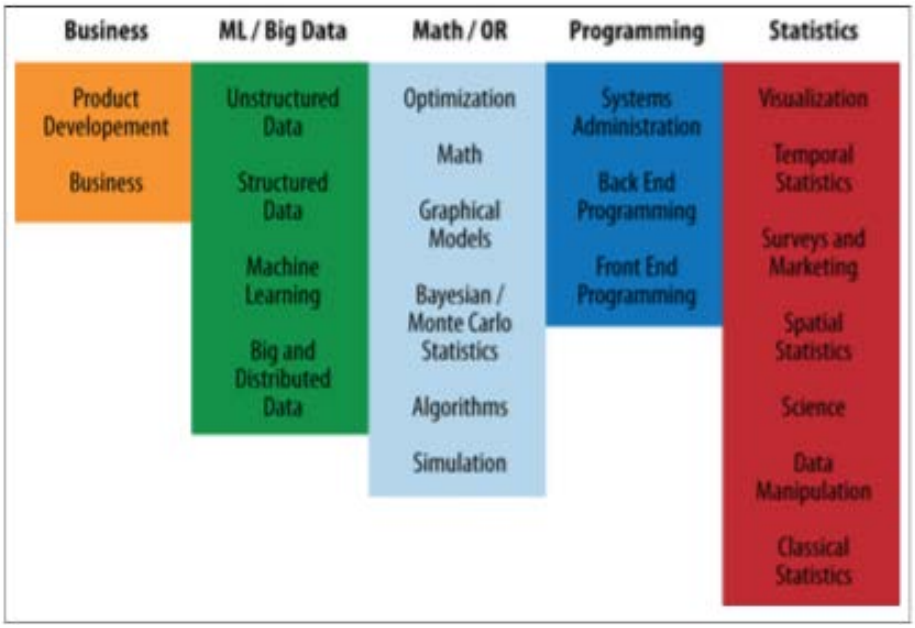

Figure 4 – Top-ranking skills in 5 cluster groups in the 2012 survey. ML = Machine Learning, OR = Operational Research

The 2015 results were subjected to the same analysis processes and are presented below, with the skills in rank order. For a small number of skills in the lower ranks, the placing is somewhat arbitrary – Machine Learning, for example, was ranked similarly in the Statistics, Data Manipulation and Maths and Model Groups and was only marginally the highest score in the latter.

Statistics	Data Management	Maths and Models	OR	Business
Data Manipulation	Structured Data	Science	Simulation	Product Development
Classical Statistics	Systems/RDBMS Administration	Maths	Optimisation	Business
Temporal Statistics	Back-End Programming	Front-End Programming	Temporal Statistics	
Visualisation	Spatial Statistics	Algorithms	Bayesian/Monte-Carlo Statistics	
Surveys and Marketing	Unstructured Data	Graphical Models		
		Machine Learning		

Figure 5 – Top-ranking skills in the 2015 survey in 5 cluster groups

Combining Skills and Self-Identification

The data on skills and self-identities can be combined to create a mosaic plot:

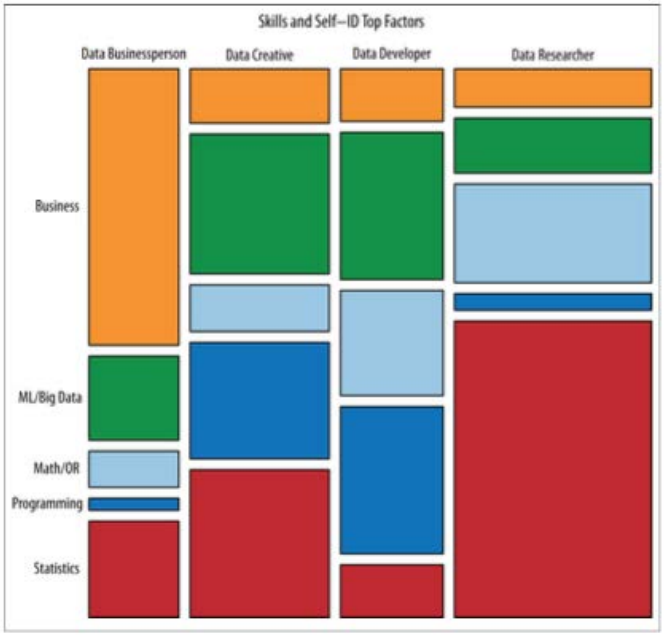

Figure 6 - Combination of self-identity and skills from a Data Science community in 2012

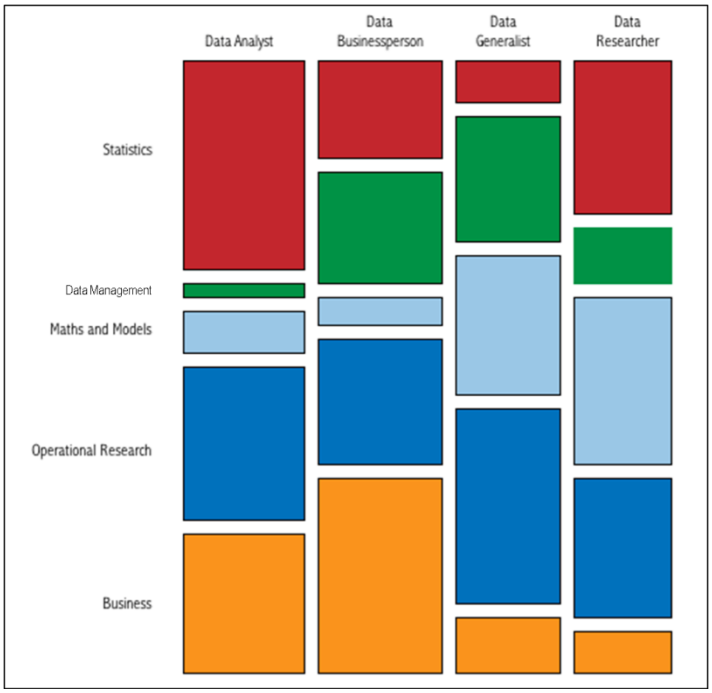

Figure 7 - Combination of self-identity and skills from the government analysts in 2015

Key findings here are that the self-identity groups in 2015 were much more evenly split numerically and with fewer distinct contrasts in the relative proportion of their skills. The Data Businessperson archetype for example, is much less 'business-like' in their skills mix than their counterpart in the wider community. It would seem that respondents within this group are keen to hang on to their analytical skills.

Conclusions

What this survey has provided is a source of useful evidence, from a relatively large sample, as to the current skills and professional composition of data science practitioners in government. By re-using an earlier method, comparisons with a wider community of practitioners is possible. Key messages that emerge include:

A good foundation for Data Science – around half of the respondents already identified themselves as a Data Scientist. This is indicative of receptivity to data science.

Good analytical skills – in the current government analytical community, there is good representation of a number of key analytical with the 'Maths and Models' and 'OR' skills groups strongly represented across the board. The skills consistently ranked the highest were:

- Data Manipulation
- Classical Statistics
- Maths
- Visualisation
- Science
- Structured Data
- Temporal Statistics

Lack of creativity and curiosity – the biggest deficit in government, compared to the wider community, is in the self-identity category 'Artist' and also in the relatively low rankings for Engineer and Entrepreneur.

Key gaps in the technical skills – across the skill groups, the five skills that ranked consistently the lowest scoring were:

- Unstructured Data
- Graphical Models
- Back-End Programming
- Big and Distributed Data
- Systems and RDBMS Administration

These conclusions provide us with a foundation on which to build the training and recruitment work in the Government Data Science Partnership and across the analytical professions more widely.

Bill Oates
January 2016
bill.oates@ons.gov.uk

Skills List Used in the Survey

Algorithms (ex: computational complexity, CS theory)
Back-End Programming (ex: JAVA/Rails/Objective C)
Bayesian/Monte-Carlo Statistics (ex: MCMC, BUGS)
Big and Distributed Data (ex: Hadoop, Map/Reduce)
Business (ex: management, business development, budgeting)
Classical Statistics (ex: general linear model, ANOVA)
Data Manipulation (ex: regexes, R, SAS, web scraping)
Front-End Programming (ex: JavaScript, HTML, CSS)
Graphical Models (ex: social networks, Bayes networks)
Machine Learning (ex: decision trees, neural nets, SVM, clustering)
Math (ex: linear algebra, real analysis, calculus)
Optimization (ex: linear, integer, convex, global)
Product Development (ex: design, project management)
Science (ex: experimental design, technical writing/publishing)
Simulation (ex: discrete, agent-based, continuous)
Spatial Statistics (ex: geographic covariates, GIS)
Structured Data (ex: SQL, JSON, XML)
Surveys and Marketing (ex: multinomial modeling)
Systems Administration (ex: *nix, DBA, cloud tech.)
Temporal Statistics (ex: forecasting, time-series analysis)
Unstructured Data (ex: noSQL, text mining)
Visualization (ex: statistical graphics, mapping, web-based data-viz)