Fantastic (data)-Beasts and Where to Find Them: Data Scientists and Data Engineers

Author, Francesco Corea

A Data Science Foundation White Paper

January 2019

----

www.datascience.foundation

What it takes to be a good data scientist (and how to become one)

I. A Philosophical Introduction

There are a great confusion and vagueness around what big data and AI really are, and the technicalities of the *data black box* have turned the people who analyze huge datasets into some kind of mythological figures. These people, who possess all the skills and the willingness to crunch numbers and providing insights based on them, are usually called ***data scientists***.

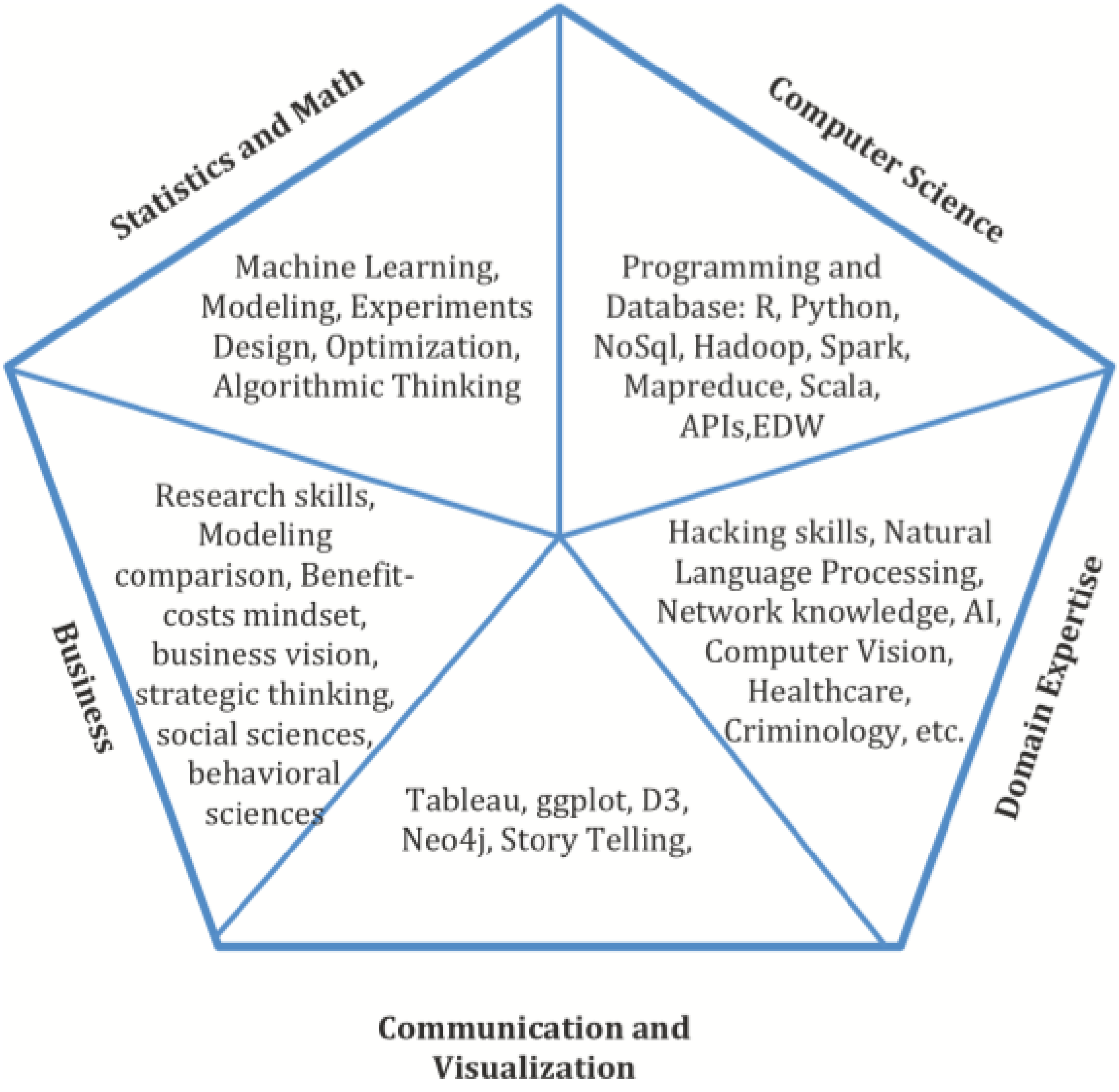
They have inherited their faith in numbers from the Pythagoreans before them, so it may be appropriate to fancily name them ***Datagoreans***. Their school of thinking, the Datagoreanism, encourages them to pursue the truth through data and to exploit blending and fruitful interactions of different fields and approaches for postulating new theories and identifying hidden connections.

However, the general consensus about who they are and what they are supposed to do (and internally deliver) is quite loose. By simply browsing job offers for data scientists one understands that employers do not often really know what they are exactly looking for, and this is probably one of the reasons of the apparent shortage of data scientists in the job market (Davenport and Patil, 2012).

II. Data Toolbox and Skill Set

In reality, data scientists as imagined by most do not exist because it is a completely new figure, especially for the initial degrees of seniority. However, the proliferation of boot camps and structured university programs on one hand, and the companies' increased awareness about this field on the other hand, will drive the job market towards its *demand-supply equilibrium*: firms will understand what they actually need in term of skills, and talents will be eventually able to provide those (verified) required abilities.

It is then necessary at the moment to outline this new role, which is still half scientist half designer, and it includes a series of different skills and capabilities, akin to the mythological *chimera*. An ideal profiling is then provided in the following table, and it merges basically five different job roles into one: the computer scientist, the businessman, the statistician, the communicator, and the domain expert.


Statistics and Math
Machine Learning, Modeling, Experiments Design, Optimization, Algorithmic Thinking
Computer Science
Programming and Database: R, Python, NoSql, Hadoop, Spark, Mapreduce, Scala, APIs,EDW
Domain Expertise
Hacking skills, Natural Language Processing, Network knowledge, AI, Computer Vision, Healthcare, Criminology, etc.
Communication and Visualization
Tableau, ggplot, D3, Neo4j, Story Telling,
Business
Research skills, Modeling comparison, Benefit-costs mindset, business vision, strategic thinking, social sciences, behavioral sciences

Data Scientist' core skills set

Clearly, it is very cumbersome if not impossible to substitute five different people with a single one. This consideration allows us to draw several conclusions. First, collapsing five job functions has a controversial effect on productivity because it might be:

i) ***efficient*** because the entire value and product chain is concentrated and not dispersed;
ii) ***risky*** because a single inpidual can sometimes be less productive than five different people working on the same problem at the same time.

Second, ***hiring one specialist should cost less than hiring five semi-specialists***, but much more than anyone of them singularly considered (because of his specialization, high-level knowledge and flexibility). Looking at some numbers, though, this does not seem to be reflected in the job market.

III. A Toy-Model for Data Jobs

Using Glassdoor.com, it is possible to notice that on average in 2015 in the United States (i) a computer scientist annually earns around $110,000, (ii) a statistician around $75,000, (iii) a business analyst $65,000, (iv) a communication manager $80,000, and finally, (v) a domain expert about $57,000. On the other side, a data scientist salary median is around **$100,000** according to the survey run by O'Reilly the same year (King and Magoulas, 2015).

From the survey it is possible to also notice that an average working week lasts often 40 hours, and ***they spend twice the time on ETL and cleaning data rather than running proper analysis or creating models*****.**

According to these statistics, and roughly (maybe incorrectly from a practitioner's point of view) assuming that the rest of their time is equally pided into the other three activities, a data scientist should earn around $92,000. This is, of course, a very approximate estimate, which does not take into account any seniority, differences across industries, etc., and where the domain expertise is computed as the average of marketing ($55,000), finance ($65,000), database ($57,000), network ($64,000), and social media ($41,000) specializations.

But it does convey a broad concept: **data scientists seem to be (almost) fairly compensated in absolute terms**, but their remuneration is definitely lower if compared to the **cost structure they face** to become such specialized figure.

It is really expensive in terms of education, effort, and opportunity costs to become a data scientist, and the average job market does not compensate enough a candidate for it.

Well, truth be told, ***the market is quickly becoming polarized:*** either you are a top scientist employed by huge companies (and you get paid a ton of money) or you don't get fairly compensated for the incredible work it took you to enter the data world.

IV. Data Scientists' Personality

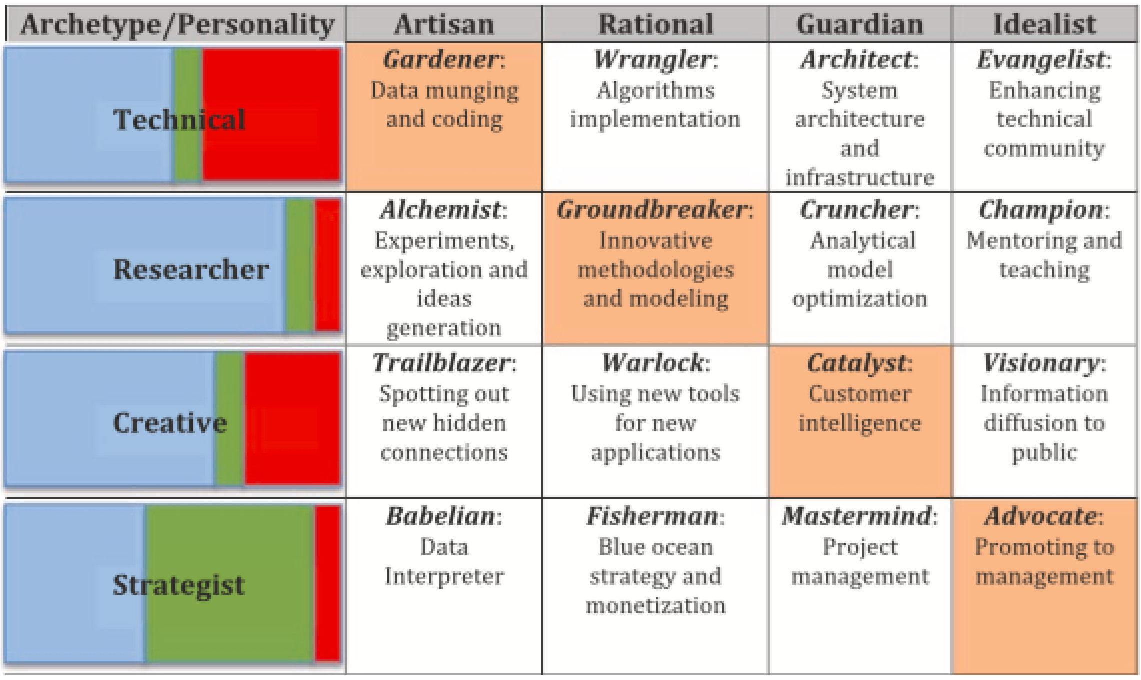
There is no scientist exactly alike another, and this is true for data scientists as well. Even if data science seems to mainly be a field run by American white male with a PhD (what I inferred from King and Magoulas, 2015), this is not conclusive at all on the ideal candidate to hire. The suggestion is ***to value the skills and capabilities more than titles or formal education*** (there are not many academic programs so well-structured to signal the right set of competencies to potential employers).

So far, in order to become a data scientist, the paths to be followed could have been unconventional and various, so it is important to assess the abilities instead of simply deciding based on the type of background or degree level. Never forget that one of the real extra-value added by data science is ***different field contaminations and cross-sectional applications***.

But there is also another relevant aspect to take into account in choosing the right candidate for your team, and that is ***Psychology***.

Not all the data specialists are the same (Liberatore and Luo, 2012; Kandel et al., 2012), and it is possible to cluster them in four different groups (Harris et al., 2013) and by four different personalities in order to reach a deeper granularity, based on their actual role within the company ("***Archetypes***") and on personal features ("***Personality***"—according to the *Keirsey Temperament Sorter*).

Correctly identifying the personality type of a data scientist is crucial to amplify his internal contribution and efficiency, as well as to maximize the resources employed to recruit him.



Archetype/Personality	Artisan	Rational	Guardian	Idealist
Technical	***Gardener***: Data munging and coding	***Wrangler***: Algorithms implementation	***Architect***: System architecture and infrastructure	***Evangelist***: Enhancing technical community
Researcher	***Alchemist***: Experiments, exploration and ideas generation	***Groundbreaker***: Innovative methodologies and modeling	***Cruncher***: Analytical model optimization	***Champion***: Mentoring and teaching
Creative	***Trailblazer***: Spotting out new hidden connections	***Warlock***: Using new tools for new applications	***Catalyst***: Customer intelligence	***Visionary***: Information diffusion to public
Strategist	***Babelian***: Data Interpreter	***Fisherman***: Blue ocean strategy and monetization	***Mastermind***: Project management	***Advocate***: Promoting to management

Data scientists' personality assessment and classification

In the table above, a full disentanglement of data scientists' types is provided. The color roughly represents the partition between three main skills they possess—based on the survey run by Harris et al. (2013)—that are *mathematics-statistic-modeling skills* (blue), *business competencies* (green), and *coding abilities* (red).

Having this clear classification in mind may be argued to be a merely speculative and useless labeling exercise, but it is indeed extremely relevant to increase the team efficiency: identifying personal inclinations and aspirations ***would allocate the best people to the best job role***, and common complaints and problems such as the lack of time for doing analysis, the poor data quality, and the excessive time spent in collecting and cleaning data (King and Magoulas 2015), would be eliminated—or better, ***they would be assigned to the right people.***

This classification does not want to be, of course, a quality assessment of what type of data scientist is better than others but rather ***a framework for helping organizations to also identify the minimum team structure to start with:*** on the main diagonal there are indeed the basic figures needed in order to properly establish a fully-functional data science team.

The **Gardener** (usually known also as ***data engineer***) is in charge of maintaining the architecture and making the data available to the **Groundbreakers**, who are usually identified as the proper ***data scientist***, and that try to answer research questions and draw insights from data once they verified through models testing. The insights are then passed to **Advocates** (***business intelligence***) and **Catalysts** (***customer intelligence*** team), who respectively communicate that information to executives and use it to increase customers' satisfaction.

V. Final Considerations

All the considerations drawn so far point to a few suggestions for hiring data scientists: first of all, ***data science is a team effort***, not a solo sport. It is important to hire different figures as part of a bigger team, rather than hiring exclusively for inpidual abilities.

Moreover, if a data science team is a company priority, the data scientists ***have to be hired to stay and not simply on a project-base*** because managing big data is a marathon, not a 100 metres.

Second, data scientists come with two different DNAs: the scientific and the creative one. For this reason, they should be let free to learn and continuously study from one hand (the science side) and to create, experiment, and fail from the other (the creative side). They will never grow systematically and at a fixed pace, but they will do that organically based on their inclinations and multi-faceted nature. It is recommended to leave them with some spare time to follow their 'scientific inspiration'.

Finally, they need to be incentivized with something more than simply big money. The retention power of a good salary is indeed quite low with respect to interesting daily challenges, relevant and impactful problems to be solved, and being part of a scientific bigger community (i.e., being able to work with peers and publishing their research).

I am also aware I did not spend much time in this post discussing the differences between Data Scientist and Data Engineers because for the sake of this article I considered them as declinations of the same job-paradigm. ***Nonetheless, you might want to check this post to know more about those two different roles.***

References

Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679.

Davenport, T. H., & Patil, D. J. (2012). *"Data scientist: The sexiest job of the 21st century". Harvard Business Review, 90*(10), 70–76.

Dull, T. (2014). A Non-Geek's Big Data Playbook. SAS Best Practices White paper. Retrieved from http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/non-geeks-big-data-playbook- 106947.pdf.

Harris, H. D., Murphy, S. P., & Vaisman, M. (2013). *Analyzing the Analyzers*. O'Reilly Publishing.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise Data analysis and visualization: An interview study. In *Proeedings of IEEE Visual Analytics Science & Technology (VAST)*.

King, J., & Magoulas, R. (2015). *2015 data science salary survey*. United States: O'Reilly Media, Inc.

Liberatore, M., & Luo, W. (2012). ASP, the art and science of practice: A comparison of technical and soft skill requirements for analytics and or professionals. In *Interfaces 201343*, (vol. 2, pp. 194–197).

Note: the above is an adapted excerpt from my book "Big Data Analytics: A Management Perspective" (Springer, 2016). A new version of this article has been proposed in "Introduction to Data" (Springer, 2019).

Appendix I. Data Scientists Personality Test

I am not a psychologist, so I would suggest extra care and help in doing that, but I want to provide a basic test to understand the personality of the scientists belonging to your team.

The terminology used to classify into 16 subcategories the different kind of data scientists is given by the two-entry matrix exhibited in the table above. The terminology can be sometimes misleading if you have clear in mind the Keirsey Temperament Sorter (KTS), and this is why it is necessary to specify that the only categorization borrowed from KTS framework is the broader one, i.e. the *Artisan-Idealist-Rational-Guardian* partition.

Every sub-category has instead to be taken as newly generated. Here it follows the personality test to sort data scientists into a specific box. It is composed of 10 questions, and for each one, a single answer has to be provided.

Again, this test is not a professional temperament test to fully understand inpiduals' personality (which I think to be almost impossible), but it is more a quick tool for managers to efficiently and consciously allocate the right people to the right task.

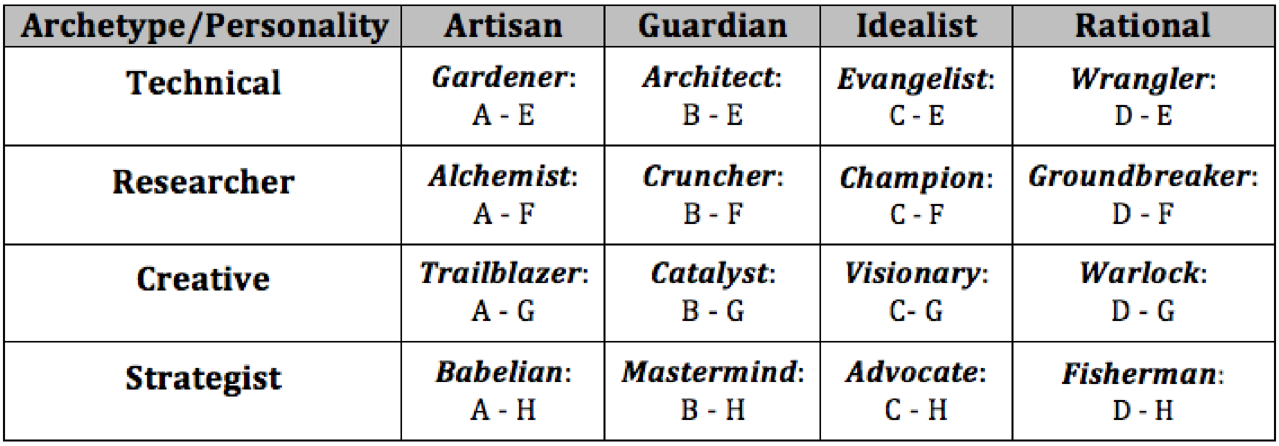
Questionnaire

1. **When you start working on a new dataset:**
   a. You start exploring immediately and querying the data
   b. Plan in advance how to tackle it
   c. You spent time in understanding the data, where they come from, and their meaning
   d. You identify a research question quickly and focus on designing a new improved method for analyzing your data
2. **In your team, people count on you for your:**
   a. Troubleshooting ability
   b. Organizational skills
   c. Capacity to reduce the problem complexity
   d. Strategic approach and conceptualization of the problem
3. **When facing a new data challenge, your first thought is:**
   a. Is what I am doing impactful and relevant?
   b. When do I have to deliver some results?
   c. How can this challenge make me a better scientist?
   d. What can I learn from this dataset?
4. **In a data analysis, which is the most important thing to you?**
   a. Results, no matter how you do achieve them, what strategy or technology you do employ
   b. To achieve a result in the correct way and with the right process or technology
   c. Attaining significant results in an ethical manner
   d. Reaching the outcomes through an accurate, replicable, and efficient procedure
5. **If you have finished your required daily work, you would:**
   a. Focus again on your analysis and try to find an alternative and innovative way to achieve your final goal
   b. Start with something else, even if this may involve staying longer at your desk
   c. Help a colleague in difficulty with his analysis
   d. Give suggestions and highlight weaknesses in your colleagues' works for the sake of the team and of the business development
6. **If you would have some spare time during your work, you would prefer to:**
   e. Optimize existing technology for the whole company
   f. Improve your analysis

   - **g.** Try to derive new insights from your previous analysis
   - **h.** Understanding how to maximize the value of your analysis
7. **It is your ‘*data-dream*’ of:**
   - **e.** Speaking about data with only engineers and IT teams
   - **f.** Teaching data related contents
   - **g.** Engaging with people who do not know anything about data science
   - **h.** Persuading and convincing the business team of the opportunities generated by big data
8. **You prefer to work with:**
   - **e.** Huge amount of structured data
   - **f.** Any kind of data that challenge me
   - **g.** Behavioral or social media data, or any unusual data
   - **h.** No data in particular
9. **If you would quit tomorrow your data science job, you would prefer to become:**
   - **e.** An IT manager or a software engineer
   - **f.** A professor
   - **g.** A consultant
   - **h.** An entrepreneur
10. **What characteristic of big data you value the most?**
    - **e.** Volume
    - **f.** Velocity
    - **g.** Variety
    - **h.** Value

Once each question has been addressed choosing a single answer, the result is given by pairing the reply chosen more often within the first five questions (a–d) with the answer that appears more often in the last five (e–f), as shown in the following table.

Hence, if for instance in the first five answers **b** emerges as the predominant answer, while in the last five **f** is the median, the person considered is a Cruncher.



Archetype/Personality	Artisan	Guardian	Idealist	Rational
Technical	***Gardener***: A - E	***Architect***: B - E	***Evangelist***: C - E	***Wrangler***: D - E
Researcher	***Alchemist***: A - F	***Cruncher***: B - F	***Champion***: C - F	***Groundbreaker***: D - F
Creative	***Trailblazer***: A - G	***Catalyst***: B - G	***Visionary***: C- G	***Warlock***: D - G
Strategist	***Babelian***: A - H	***Mastermind***: B - H	***Advocate***: C - H	***Fisherman***: D - H

Appendix II. **Code of Professional Conduct—Instructions**

1. Terminology

Terms as data, data scientist, big data, have to be defined.

2. Working Relationship

This section highlights the relevance of defining the scope of the relationship, and the means through which the scope has to be reached.

It has to guarantee as well professionalism, competences, independence and objectivity (Boyd and Crawford 2012).

A subparagraph has to explain how to proceed in case of misrepresentation, misconduct, or fraud.

Finally, a section on the quality of the analysis and results have to be presented, as well as how a data scientist should act in case the results he achieved are after- wards misrepresented or misused. He has to use diligence, scientific method, replicability of process and analysis, and not provide evidences he knows to be false or incomplete.

3. Conflict of Interests

It shall explain the policy regarding disclosure of conflicts and limitations. Exceptions have to be listed.

4. Duties to Clients

The data scientist has to present correctly his results, preserve the confidentiality of the agreement, and act for the benefit of his clients before his own or his employer's one.

It should include a section related to the communication with clients, as well as the disclosure of risks on relying on the (data) results obtained. The results should also be evaluated with reasonable diligence and explained to the extent of allowing the client to reach a decision by his own.

It would apply to current and prospective clients, with a series of further confidentiality in the second case (not revealing information from a prospective client, measures to avoid conflict of interests, etc.).

5. Duties to Employers

In this section it has to be deal with themes as loyalty to the employer, and supervising responsibilities.

6 Confidential Information

This paragraph should de ne confidential information, setting the guidelines for protecting them, when the confidentiality can be breached (fraud, in order to prevent death, etc.), as well as the final return at the end of the project.

Appendix III. Data scientist Extended Skills List

Programming: R, Python, Scala, JavaScript, Java, Ruby, C++, C#

Statistics and Econometrics: probability theory, ANOVA, MLE, regressions, time series, spatial statistics, Bayesian Statistics (MCMC, Gibbs sampling, MH Algorithm, Hidden Markov Model), Simulations (Monte Carlo, agent-based modeling, NetLogo)

Scientific approach: experimental design, A/B testing, technical writing skills, RCT

Machine Learning: supervised and unsupervised learning, CART, algorithms (Support vector Machine, PCA, GMM, K-means, Deep Learning, Neural Networks)

Mathematics: Matrix algebra, relational algebra, calculus, optimization (linear, integer, convex, global)

Big Data Platforms: Hadoop, Map/Reduce, Hive, Pig, Spark, Storm

Text mining: Natural Language Processing, LDA, LSA, Part-of-speech tagging, Parsing, Machine Translation

Visualization: graph analysis, social networks analysis, Tableau, ggplot, D3, Gephi, Neo4j, Alteryx

Business: business and product development, budgeting and funding, project management, marketing surveys, domain/sector knowledge

Systems Architecture and Administration: DBA, SAN, cloud, Apache, RDBMS

Dataset Management:

- **Structured Dataset:** SQL, JSON, BigTable
- **Unstructured Dataset:** text, audio, video, BSON, noSQL, MongoDB, CouchDB

- **Multi-structured Dataset:** IoT, M2M

Data Analysis: feature extraction, stratified sampling, data integration, normalization, web scraping, pattern recognition

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation