# DATA SCIENTIST SKILL SET

d4t4science.com | Philipp M. Diesinger, June 7th, 2016

## 1 BACKGROUND

Data science is first and foremost a talent-based discipline and capability. Platforms, tools and IT infrastructure play an important but secondary role. Nevertheless, software and technology companies around the globe spend significant amounts of money talking business managers into buying or licensing their products which often times results in unsatisfying outcomes that do not come close to realizing the full potential of data science.

Talent is key - but unfortunately very rare and hard to identify. If you are trying to hire a data scientist these days you are facing the serious risk of recruiting someone with the wrong or an insufficient skill set. On top of things, talent is even more crucial for small or medium-sized companies whose data science teams are likely to stay relatively small. Wasting one or two head counts on wrong profiles might render an entire team inefficient.

The demand for data scientists has risen dramatically in recent years [1, 2, 3, 4, 5]:

- **New technologies** significantly improved our ability to manage and process data; including new data types of data as well as large quantities of data.
- A **shift in mind set** in business environments took place [6] regarding the utilization of data: from data as a reporting and business analytics necessity towards a valuable resource to enable smart decision making.
- Last but not least exciting **new intellectual developments**
- Last but not least exciting **new intellectual developments** have taken place in relevant related academic disciplines like machine learning [7, 8] or natural language processing.

Due to high demand, the term ‘data scientist’ developed into a recruiting buzz word which is broadly being abused these days. Experienced lead data scientists share a painful experience when trying to fill a vacant position: Out of a hundred applicants, typically only a handful matches the requirements to qualify for an interview. Some candidates feel already qualified to call themselves ‘data scientist’ after finishing a six-week online course on a statistical computing language. Unqualified individuals often times end up being hired by managers who themselves lack data science experience - leading to disappointments, frustration and an erosion of the term ‘data science’.

## 2 WHO IS A DATA SCIENTIST?

The data scientist skill set described in the following is based on the idea that it fundamentally rests on three pillars, each representing a skill set mostly orthogonal to the remaining two.

Following this idea, a solid data scientist needs to have the following three well-established skill sets:

1. Technical skills,
2. Analytical skills and
3. Business skills.

Although technical skills are often times the focus of data science role descriptions, they represent only the basis of a data scientist’s skill set. Analytical skills are much harder to acquire (and to test) but represent the crucial core of a data scientist’s ability to solve business problems utilizing scientific approaches. Business skills enable a data scientist to thrive in corporate environments.

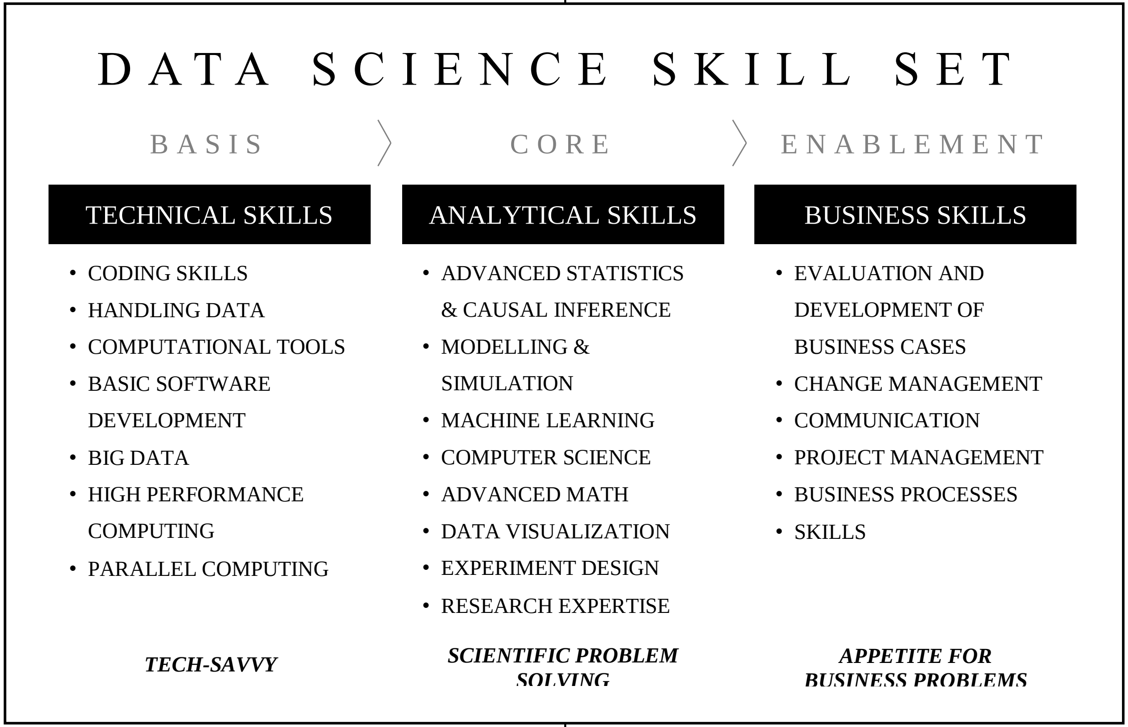
### 2.1 TECHNICAL SKILLS | BASIS

Technical skills are the basis of a data scientist’s skill set. They include coding skills in languages such as R or Python, the ability to handle various computational architectures, including different types of data bases and operating systems but also other skills such as parallel computing or high performance computing.

The ability to handle data is a necessity for data scientists. It includes data management, data consolidation, data cleansing and data modelling amongst others. As there is often times a high demand for these skills in corporate environments, it comes with the risk of focusing data scientists on data management tasks - thus distracting them from their actual work.

Almost more important than a candidate’s current technical skill set is their mind set. A key factor is intellectual agility providing candidates with the ability to adapt to new computational environments in a short amount of time. This includes learning new coding languages, dealing with new types of data bases or data structures or keeping up with current technological developments like moving from relational databases to object-analytical approaches. A data scientist with a static technical skill set will not thrive for long as the discipline requires constant adaption and learning. Strong candidates show a healthy appetite for developing their technical skills. When a candidate focusses on a tool discussion during an interview it can be an indication of a narrow technical comfort zone with firm constraints.

Unfortunately, data science job profiles are often times narrowly focused on technical skills; caused by a) the misperception that a successful data scientist's secret lies exclusively in the ability to handle a specific set of tools and b) a lack of knowledge on the hiring manager's end as to what the right skill set looks like in the first place. Focusing on technical skills when evaluating candidates renders a significant risk.



DATA SCIENCE SKILL SET

BASIS 〉 CORE 〉 ENABLEMENT

TECHNICAL SKILLS

- CODING SKILLS
- HANDLING DATA
- COMPUTATIONAL TOOLS
- BASIC SOFTWARE DEVELOPMENT
- BIG DATA
- HIGH PERFORMANCE COMPUTING
- PARALLEL COMPUTING

***TECH-SAVVY***

ANALYTICAL SKILLS

- ADVANCED STATISTICS & CAUSAL INFERENCE
- MODELLING & SIMULATION
- MACHINE LEARNING
- COMPUTER SCIENCE
- ADVANCED MATH
- DATA VISUALIZATION
- EXPERIMENT DESIGN
- RESEARCH EXPERTISE

***SCIENTIFIC PROBLEM SOLVING***

BUSINESS SKILLS

- EVALUATION AND DEVELOPMENT OF BUSINESS CASES
- CHANGE MANAGEMENT
- COMMUNICATION
- PROJECT MANAGEMENT
- BUSINESS PROCESSES
- SKILLS

***APPETITE FOR BUSINESS PROBLEMS***

## 2.2 ANALYTICAL SKILLS | CORE

Scientific problem solving is an essential part of data science. Analytical skills represent the ability to succceed at this complex and highly non-linear discipline. Establishing throrough analytical skills requires a high amount of commitment and dedication (which is a limiting factor contributing to the global shortage of data scientists).

Analytical skills include expertise in academic disciplines like computer science, machine learning, advanced statistics, probability theory, causal inference, artificial intelligence, feature extraction and others (including strong mathematical skills). The list can be extended almost infinetely [9, 10, 11] and has been subject to many debates. Covering all potentially usefull analytical disciplines is a life-time achievement for any data scientist and not a requirement for a successful candidate. Rather, a data scientist needs to have a healthy mix of analytical skills to succeed. For instance, an expert on Markov chains and an expert on Bayesian networks might both be able to develop a solution for the very same business problem although utilizing their respective strengths and thus fundamentally different methods.

Analytical skills are typically beeing developed through pursuing excellence in a highly quantitative academic field such as computer science, theoretical physics, computational math or bioinformatics. These skills are trained in academic institutions through exposure to hard, unsolved research problems that require a high level of intellectual curiosity and dedication to tackle and eventually solve. This is typically done over the course of a PhD.

Mastering a quantitative research question that nobody else has solved before is a non-linear process inadvertedly accompanied by failing over and over again. However, this process of scientitic problem solving shapes the analytical mind and builds the expertise to later succeed in data science. It typically consists of iterative cycles of

a) implementing and adapting an analytical approach
b) applying it and observing it fail, then
c) investigating the problems and
d) building an understanding why it failed and where the limitations of the approach lie
e) to come up with a better more refined approach.

These iterations are acompanied with key learnings and represent small steps towards the project goal thus effectively zig-zagging towards the final solution.

A key requirement for analytical excellence is the right mind set: A data scientist needs to have an intrinsic, high level of curiosity and a strong appetite for intellectual challenges. Data scientists need to be able to pick up new methods and mathematical techniques in a short amount of time to then apply them to a problem at hand - often times within the limited time frame of an ongoing project.

A good way to test analytical skills during an interview process is to provide potential candidates with a business problem and real data to then ask them to spend a few of hours working on it remotely. Discussing the code they wrote, the approach they chose, the solution they built and the insights they generated is a great way to evaluate their potential and at the same time provide the candidates with a first feeling for their potential new tasks.

### 2.3 Business Skills | Enablement

Business skills enable data scientists to thrive in a corporate environment.

It is important for data scientists to communicate effectively with business users utilizing business lingua and at the same time avoiding a shift towards a conversation that is too technical. Healthy data science projects start and end with the discussion of a business problem supported by a valid business case.

Data scientists need to have a good understanding of business processes as it will be required to make sure the solution they build can be integrated and ultimately consumed by the respective business users. Careful and smart change management almost always plays a role in data science projects as well. A solid portion of entrepreneurship and out-of-the-box thinking helps data scientists to consider business problems from new angles utilizing analytical methods that their business partners do not know about. Last but not least, many big and successful data science projects that ultimately lead to significant impact were achieved through ‘connecting the dots’ by data scientists who built up internal knowledge by working on different projects across departments and functions.

Candidates who come with strong technical and analytical skills are often times highly intelligent individuals looking for intellectual challenges. Even if they have no experience in an industry or in navigating a corporate environment, they can pick up required business skills in a short amount of time - given that they have a healthy appetite for solving business cases. Building strong analytical or technical skills takes orders of magnitude longer.

When trying to determine whether a candidate has an intrinsic interest in business questions or whether he or she would rather prefer to work in an academic setting, it can help to ask yourself the following questions:

- How well can the candidate explain data science methods like deep learning to business users?
- When discussing a business problem can the candidate communicate effectively in business terms while thinking about potential mathematical or technical approaches?
- Will the business users collaborate with the data scientist in the future respect him or her as a partner at eye-level?
- Would you feel comfortable sending the candidate on their own to present to your manager?
- Do you think the candidate will succeed in your business environment?

## 3 Recruiting

Data science requires a mix of different skills. In the end, this mix needs to be adapted to the requirements and the situation at hand, and the business problems that represent the biggest potential value for your company. Big data for instance, is a strong buzz word but in many companies data is under-utilized to a degree that a data science team can focus on low hanging fruit for one or two years in the form of small and structured data sets and at the same time already have a strong business impact.

A key characteristic of candidates that has not been mentioned so far and which can be hard to evaluate is attitude. Hiring data scientists for business consultant positions will require a different mindset and attitude than hiring for integration into an analytics unit or even to supplement a business team.

## 4 References

[1] NY Times, *Data Science: The Numbers of Our Lives* by Claire Cain Miller http://nyti.ms/1TfCFmX

[2] TechCrunch: How To Stem The Global Shortage Of Data Scientists http://tcrn.ch/1TUIqsB

[3] Bloomberg: Help Wanted: Black Belts in Data http://bloom.bg/1Xt8bTO

[4] McKinsey on US opportunities for growth http://bit.ly/1WAonmD

[5] McKinsey on big data and data science http://bit.ly/1VXQJHD

[6] Big Data at Work: Dispelling the Myths, Uncovering the Opportunities; Thomas H. Davenport; Harvard Business Review Press (2014)

[7] Andrew Ng on Deep Learning http://bit.ly/1Tg3g74

[8] Andrew Ng on Deep Learning Applications http://bit.ly/1Wza02H

[9] Data scientist Venn diagram by Drew Conway http://bit.ly/1Xd6MAn

[10] Swami Chandrasekaran’s data scientist skill map: http://bit.ly/1ZUGUIF

[11] Forbes: The best machine learning engineers have these 9 traits in common. http://onforb.es/1VXR9Og