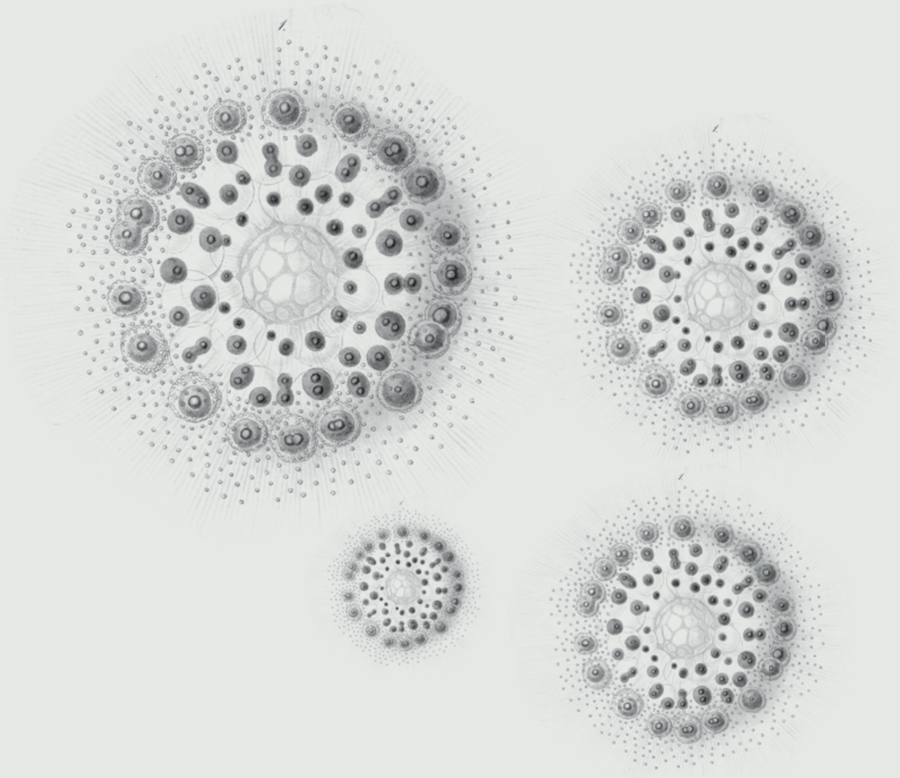
Hiring data scientists
Lessons from the trenches



ida
lab.

idalab GmbH | Alexanderstraße 7 | 10178 Berlin | idalab.de | Berlin, July 2015

Agency for Data Science

Machine learning & AI
Mathematical modelling
Data strategy

The challenge

Data scientist recruiting is a major pain point for many organizations, chiefly for three reasons: (a) vague and hard-to-assess skill set, (b) lack of data scientists on staff who are experienced with recruiting and candidate assessment and (c) no clear “ivy league” backgrounds that it is safe to hire from.

Companies who master the art of data scientist recruiting can turn this into a significant competitive advantage: getting hands on top data science talent helps them to sharpen their competitive edge as data scientists are the key resource for any organization willing to spur growth, gain traction and succeed in a digitalized world[1]. For service providers, such as management consultancies or digital agencies, excellence in data science recruiting may be a matter of survival – or a unique capability that commands a premium.

In a profession as new as data science, naturally, recruiting best practices are yet to emerge. This oftentimes leads to pre-mature hires and disappointments, when the new recruits do not perform as expected on the job – a situation that is typically detected too late, and not acted on decisively. What’s worse, this might lead to a loss of trust into the data science function as such and the organization as a whole might challenge the value of in-house data scientists if no tangible results are delivered straight away.

To avoid such unpleasant situations, companies need to carefully calibrate their recruiting process for data scientists[2]. The following five lessons learned can provide some guidance.

Assess your internal setup and clarify target roles

The first thing to consider before publishing the job posting is the intended organizational structure for deploying data science capabilities. There are, broadly speaking, two different working modes for data scientists at the moment. Data scientists either work “embedded” as an integral part of product teams or functions, or as a separate centralized unit, similar to an “inhouse consulting” function.

In the embedded setup, data scientist positions are tied to a product or function. Thus, data scientists work in close collaboration with the product manager, engineers and other employees; they focus on a specific area of application, will regularly oversee a whole lifecycle of a data science project,

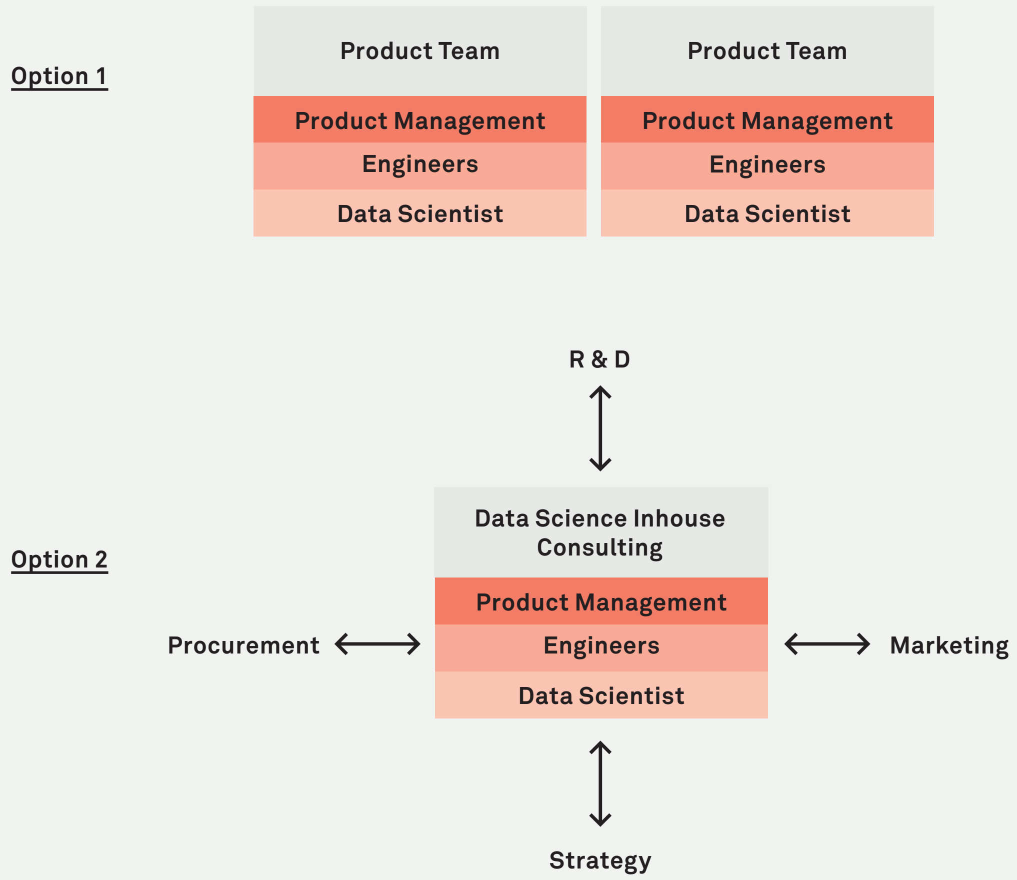
[1] See for example McFarland, Matt: “How Foursquare knew before almost anyone how bad things were for Chipotle”, The Washington Post, accessible at: https://www.washingtonpost.com/news/innovations/wp/2016/04/28/how-foursquare-knew-before-almost-anyone-how-bad-things-were-for-chipotle/

[2] Stanley, Jeremy: “How to Consistently Hire Remarkable Data Scientists”, accessible at: http://firstround.com/review/how-to-consistently-hire-remarkable-data-scientists/

and become highly familiar with aspects of production-grade engineering, maintenance and, most importantly, the specific application domain such a fraud prevention, recommendation or marketing.

In the inhouse model, a central team of data scientists engages with internal "customers" from strategy, R&D, marketing and other functions on project-basis. This type of organizational set-up is often chosen by organizations who can (a) allocate significant resources to build a centre of data science excellence and (b) have the aim to quickly explore the potential of data science across a large organization, and then, in a later stage, deploy data science capabilities within target departments for the long haul.

Options for organizational set-up of data science capabilities


Option 1

Product Team
Product Management
Engineers
Data Scientist

Product Team
Product Management
Engineers
Data Scientist

Option 2

R & D

Procurement

Data Science Inhouse Consulting
Product Management
Engineers
Data Scientist

Marketing

Strategy

1.1

For the embedded data scientist (Option 1), the product team can inform the recruiting process with clear requirements regarding role and skill requirements (e.g. specifics of the application domain, technological requirements, etc.). The desired profile for members of an inhouse consulting unit, on the other hand, is harder to specify in terms of a technical skill set.

In general, candidates need to have a broader methodological skill set, work under higher pressure to succeed and deliver on-time in a more dynamic environment[3]. Moreover, the need to interface with an (internal) customer makes it essential to distinguish between several roles. The minimum viable inhouse consulting team has three people, whose skill sets must complement each other.

Within each team, there are three distinct roles, which one could hire for (singularly or in combination, depending on the scope of work).

(1) Interfacing with the internal clients (business line), translating business insights into actionable data science, back & forth communication
(2) Research and development focused data science work, departing with a clear objective into new innovative directions
(3) Maintenance and iterative improvement of algorithms in productions systems, continuous monitoring of performance and following-up on incidents

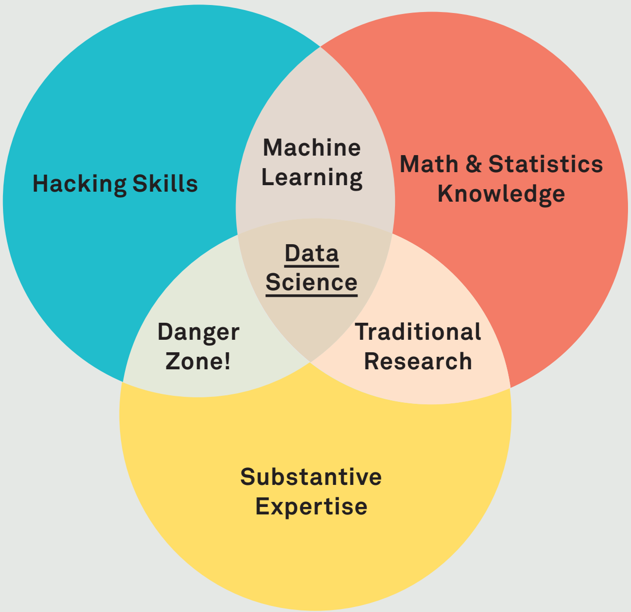
Even though it may not be able split these roles right from the beginning, it is essential to do so in the medium term.
Clarify the envisioned data scientist skill set

Clarify the envisioned data scientist skill set

Data science is positioned at the intersection of statistics, mathematics, information technology and domain expertise. Drew Conway's venn diagram[4] sketched this unique positioning in a very graphic and intuitive way.

[3] See Chu, Cheng-Tao: “Why building a data science team is deceptively hard”, accessible at: https://www.codecademy.com/blog/142
[4] See Conway, Drew: “The Data Science Venn Diagram”, accessible at: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science Venn diagram by Drew Conway



1.2

This is certainly helpful to get first orientation in the field of data science, interpret CVs and subsequently identify suitable background and experience. Still, what this diagram triangulates as a data science profile is a generic, domain-independent skill set: someone with a solid quant background, hacking skills – and substantive expertise in some domain.

But in what situations do you need to pay more attention to the specific challenges your company is facing, going beyond what this diagram suggests[5]? When do you need a particular type of data scientist?

These situations are rare, but when they exist it pays to search for a specialist, even though this might prolong the search process. Bringing a generalist data scientist up to speed is always an option, but in the following domains should companies should aim to recruit specialists:

- Natural language processing
- Time-series analysis
- Computer vision

[5] Dubey, Ashu: „How to Build A Powerful Data Science Team Without A Data Scientist“, Forbes, accessible at: http://www.forbes.com/sites/theyec/2016/03/01/how-to-build-a-powerful-data-science-team-without-a-data-scientist/#515ea5c73f41

Data Science domains and specificity of methodology

Area	Methodology (examples)
Natural language processing	Relation extraction, topic modelling, sentiment analysis, question answering, information retrieval
Time-series analysis	Filtering, auto-regressive models, segmentation, hidden markov models, time-frequency analysis
Computer vision	Object recognition, facial recognition, OCR, scene classification

1.3

Besides those applications, people with solid backgrounds (such as mathematics, physics or quantitative computer science degrees) can learn almost any kind of additional skill in the relevant area. Oftentimes, high quality candidates even see this as a great advantage of a position.

Bottom line: In a young and emerging field like data science, it remains crucial to hire for potential, intelligence and curiosity. These are the ingredients for continuous development – which will eventually turn into a major competitive advantage, as the field is so dynamic that it requires fast learning and adaption, even of those who seem to be ahead of the curve at the moment.

The absence of senior data scientists and how to deal with it

While hiring for potential is a good guideline, having an experienced data scientist with a proven leadership track record on board is essential to bring data science initiatives to sustainable fruition. Data science projects oftentimes have direct executive management oversight, requiring clear communication and a high level of project management skills. At the same time, data science projects are filled with uncertainty, trade-off decisions and complexity – and only experience allows for navigation.

While senior data scientists can be found at tech companies like LinkedIn, Twitter and other valley startups, the field is still lagging behind in Europe and there is a severe shortage of senior data scientists[6]. How should companies cope with this situation?

[5] See Davenport, Thomas; Patil, D.J: “Data Scientist: The Sexiest Job of the 21st century”, Harvard Business Review, accessible at: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Attracting senior data scientists needs an exquisite, hard to concoct, cocktail of significant remuneration, interesting product, strong brand (to some extent) and attractive location. Realistically, few companies will be able to bring this to the table, in particular when they are early on in their data science, or even digitalisation initiative. In this situation, companies should (a) focus on at least not losing potential senior data scientists in the hiring funnel and (b) hiring for leadership potential. The former can be achieved by fast-tracking promising candidates immediately after the CV-screening step, and get into wooing-mode. Leadership development, on the other hand, should be assessed at later stages of the funnel and – if proven on the job – those candidates should then receive additional training to be brought up to speed for senior positions.

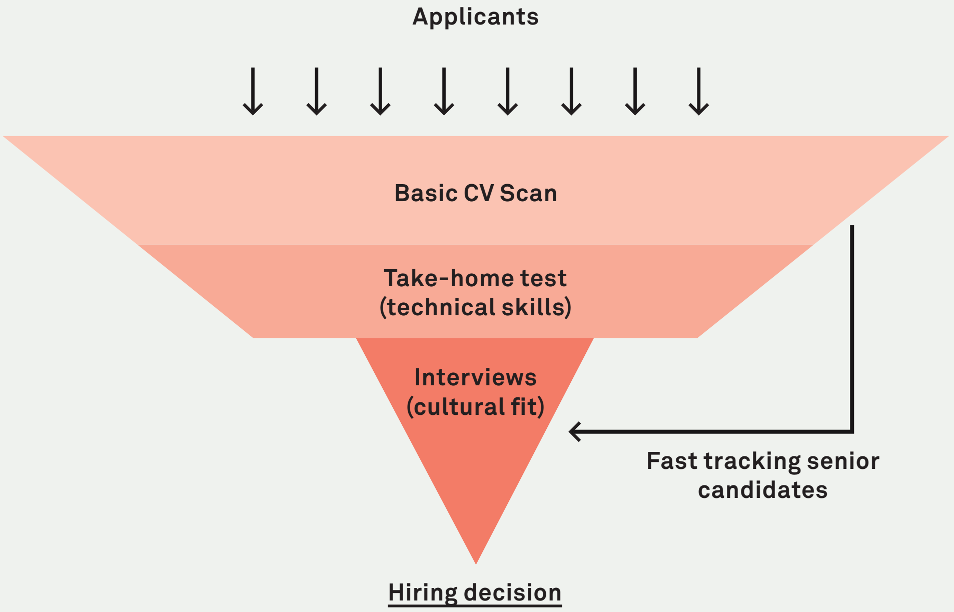
CVs are meaningless, there is no "ivy league" of data science

Getting the requirements and the job profile right is one part of the art of hiring a data scientist. Designing the actually process accordingly, is the the second – and just as important.

Unlike other recruiting processes, only minimal time should be invested into CV screening. At best, use it as a junk filter to get rid of applicants, which do not fit the requirements – and to spot data scientists with years of experience. There is no "ivy league" of data science, so it naturally would be unfitting to filter based on proxies such as university and grades. Beware also of supposed brands such as "CERN" or "Ph.D. in quantum chaos theory" – they are meaningless. While top-end consultancies oftentimes have a list of "target universities" from which they primarily hire, such practice does not improve recruiting for data scientists. In line of this thought, you should also skip time-consuming CV interviews. They only tie up resources, but provide little insight into the candidate's skills.

As a rule of thumb, always try to narrow the funnel significantly through a technical take-home test. This should filter your applicant base and allow you to focus on the 20 %, where your time is best spent. Once the candidate has passed the technical hurdle and explained the solutions accordingly, interviews should solely focus on exploring the cultural fit and getting a better understanding of the candidate's expectations and career path preferences.

Data scientist recruiting funnel


Applicants
Basic CV Scan
Take-home test (technical skills)
Interviews (cultural fit)
Fast tracking senior candidates
Hiring decision

1.4

Show me your work: the right way to assess

As the take-home data science assessment is a key component in the recruitment process, it is important to design the test accordingly.

Interestingly, take-home exams rarely scare away applicants. On the contrary, historically the drop-out rates have been extremely low. Good candidates actually enjoy the challenge as a way to enhance and showcase their skills. Moreover, an interesting take-home test is an opportunity to sell the position to the applicant – and signal that they will be evaluated on substance, not only shallow interview performance.

Any kind of test should be compiled with a healthy balance of coding and short essay questions. Just focusing on code makes it oftentimes tough to follow the candidate's thoughts. The essay questions are thus complementary and oftentimes help to understand why the candidate has chosen a certain approach in tackling the respective challenge. Furthermore, the qualitative section is a safety net for missing out on great talent and intellect, just due to the attribution of too much weight to hard coding skills. Coding can be learned and improved, the logic and intellect behind it not so easily. In the instructions, make it clear that the essay questions are as important as the code.

The coding challenge itself ideally resembles a real-life scenario, just like the candidate would have to tackle in his or her future job. Thus, work with data sets and expect actual coding. Keep the challenge as narrow as possible and provide adequate guidance, but do expect the candidates to put in eight hours to solve the challenge.

When assessing the challenge, make sure you check the code against stackoverflow copy & paste code snippets. Also, let your intuition be guided by answers in the short essay questions. If these are confusing and irritating, chances are that your candidate has also tried to copy code, without a clear modelling strategy. Taking code from other sources is not a bad thing (not at all actually), but if undertaken, it needs to be paired with a solid understanding for the context and application.

Bottom-line

Hiring a data scientist is one of the biggest challenges in recruiting, and will remain so until the size of the talent pool has increased significantly and the first cohort of data science leaders has grown up.

In order to avoid disappointments, companies should:

- Invest significant time into the compilation of the necessary skill set for the target role, and, whenever possible, avoid getting unnecessarily specific in the absence of a clear, well understood need for a certain technology or methodology
- Don’t waste time searching for a senior data scientist, hire for potential and fast-track accordingly
- In the recruitment process focus on testing technical skills through take-home tests, use CV-screening for fast-tracking and “junk” filtering only – not to assess performance
- Take-home assessments with a healthy mix of coding and essay questions are the best way to assess skills: candidates like them, and companies get a holistic view on the candidate’s skills

Contact details
Dr. Paul von Bünau
paul.buenau@idalab.de
+49 (30) 814 513-14