Successful AI Data Products

Meninder Purewal

Fall 2019

"Whatever you do, do with all your might"

Marcus Tullius Cicero

Introduction

Artificial Intelligence could potentially deliver an additional $13 trillion of economic output by 2030, increasing global GDP by 1.2% per year [1]. In the last 5 years, interest in Machine Learning and Artificial Intelligence has increase tenfold [2]. To read the media or observe the myriad of conferences on the subject, it is easy to believe everyone is using it with great success. While there is no question of its untapped potential, adoption has been uneven at best and many companies have struggled extracting value out of their data science initiatives.

There are three essential factors required for an organization to build successful data products: the data scientist, the organization's data culture and data science project design.

This document is aimed to shed light on the components required to build successful data products and avoid the potential pitfalls. There is no standard, one-size fits all reason for why a data science project might fail [3, 4]. In industries where predictive analytics are embedded into the business objective, like trading, companies went through the growing pains when integrating their quants a couple of decades ago. While this is a very similar circumstance, the recent revolution is the result of institutions with data, which is to say, all companies - technical and non-technical ones. Even for organizations with advanced data strategies, the desire for data products are permeating parts of the organization that weren't traditionally in their original purview.

The first section covers the *data scientist* and what this person is. Having a firm understanding of them turns out to be non-trivial. Hiring a data scientist is usually a firm's first move into the

data science space. They are tasked with interpreting a business problem and converting into a quantitative one. Having the right type of person for the right type of use cases is the first challenge.

The second section covers an organization's *data culture*: its attitude towards data and their internal data science team. A data driven culture is essential for the make or break of the data strategy and the way the group is represented within the organization is how stakeholders will perceive it. The right approach can reduce the inherent friction and frustration that that occurs in collaboration and derails projects.

The last section covers the steps to create successful *data science projects*.

1 Successfully Hiring Data Scientists

1.1 What is a Data Scientist?

The term 'Data Scientist' is vague. The Data Scientist role can be comprised of several people and talents. Organizations typically begin with a vague notion that it needs a data strategy and react by hiring a lead data scientist to create and implement one, without considering what type they need. Indeed, over 40% of companies report their lack of analytical skills as a primary challenge [5]. Job postings for data scientists rose 75% between 2015-2018 in indeed.com [5], while Harvard Business Review says that a data scientist is the 'sexiest job of the 21st century' [6].

Depending on the organization, their data strategy can require an individual from a very broad spectrum of talent. Given this many-to-many relationship, it is no wonder there is considerable frustration from both sides. While much ink has been spilt on defining a Data Scientist, one thing people can agree on is that they are curious creatures that enjoy creatively dabbling into disparate fields.

One method is to define the role by its deliverables [7]:

- Ad Hoc Data Scientists: Their outputs are typically 'one-off' analysis that are supported with a table or graph that is not required to be reproduced regularly. They are characterized by their speed and breadth (not depth). Their analysis is typically a precursor to the type of analysis performed by the two other types of data scientists. An example an Ad Hoc Data Scientist's work at a bank would be to find and define a threshold for 'significant' balances for clients should be included in a report. Some basic techniques might show that under a certain balance, client account are typically stale - no operational expense or revenues - and therefore elimination in a report will give better average metrics for the bank as a whole. More sophisticated techniques might include unsupervised learning, like clustering, to show

that grouping clients by balance bands is a useful reporting format. Management would use this analysis output to have reports generated with those bands.

Sometimes called a data analyst, this is a very typical type of Data Scientist in many organizations [8] and in many cases, has existed for a long time and recently been re-labeled as a data scientist because of the title appeal. It is tactical in nature, not scalable and limited in impact by the person asking for the analysis.

- Strategic ('Strat') Data Scientists: Their outputs are generally readable by executives, heads of businesses and product managers in the form of dashboards, metrics (KPIs) and predictions for decisions. In financial institutions, this could be a desk strat helping a trader decide what positions to put on, an analyst helping the business understand client profitability, etc. The required skills are statistics, experimental mindset, analytical, excellent communication with non-technical people, ideally some domain knowledge and minimum of scripting level coding.

- Modeling Data Scientists: Their output are models and algorithms, such as a recommendation system, and is generally productionized directly into machines where the output is fed directly into the next decision point in a product pipeline. The required skills are in modeling and creating production-level code. They typically have a background in Computer Science and Machine Learning. Very often, organizations bring in a modeling data scientist with strong machine learning skills, but aren't quite ready for them. These data scientists want to build models, not translating business problems or handling politics. The main blockers in this failure case are not technical.

  For example, a data scientist is not able to create models because the data is 'too raw'. The analyst would like to work with number of website visits per user, but has access to tables of every action a user has ever taken. Spending weeks converting the raw data to features, potentially something a Data Engineer might do in a larger organization, is under-appreciated by the business, who simply sees long turnaround times.

  Communication, articulation of the issues and building a true strategy is beyond the scope of a single project. Modeling data scientists, who want to use their machine learning knowledge, are better served in later stage pipelines, where they can concentrate on hyper parameter tuning, scaling or more sophisticated modeling. In large organizations, they have documentation requirements and strong error handling/edge case concentration.

1.2 What are the skills of a data scientist?

1.2.1 Technical Skills

While Data Scientists can be defined by their output, the required skill set is varied and a source of confusion for many. In order to hire appropriately, understanding where the organization is in

terms of established data strategy and what type of investment is planned. A single person that spans the set of required skills is known as a *full-stack* data scientist. They are able to dabble in each of the technical skills below [9]. An alternative paradigm is known as a *coalition*, where a collection of teammates, each with a specific data science skill is brought together.

For organizations with a more mature data strategy and implementation, the pertinent skills required are around the machine learning modeler type data scientist, ready to create and integrate real AI/ML products. If there isn't any data recorded or the organization is unsure, then emphasis should be on the skills required to build the earlier phase of the pipeline, such as Data Engineering.

An important technical skill for data scientists to have is the confidence to step away from overly complicated, academically 'sexy' models, such as neural networks, when a simpler method would be sufficient for the problem or phase of the solution implementation. This is a common pitfall among junior data scientists who take pride in using obscure methodologies they recently read about. Another source is among aspiring data scientists that are curious about the hype or are trying to prove themselves. While commendable, it is important to veer away from over complicated solutions that might carry high technical debt, large investments, but little payoff to the business.

Title	**Description**	**Tools**
Data Architect	Forward thinking and experienced, has a holistic vision of the architecture. Design blueprints for data management systems.	Big data technologies, database architectures, data structures, data warehouses
Data Engineering	Develop, construct and test data architectures, generally consisting of robust and scalable infrastructure for moving bytes from A to B.	Many of the same as Data Architect, but more in depth, less breadth.
Data Quality	Finds and cleanses incorrect data and works with other teams to prevent data problems.	SQL, Data quality tools
Data Analytics Engineering	New role emerging from proliferation of vendor products, which has resulted in companies investing less in data architects and engineers, passing burden onto data scientists.	Strong software engineering skills, familiarity with vendor as-a-service products
Data Scientist	Building machine learning and quantitative models.	Python, Jupyter notebooks
Data Product Manager	New, specialized role understanding data scientists output and incorporating into products for internal and external customers.	Bridge between technical and business: communication and modest

1.2.2 Non-Technical Skills

Communication

An essential, but overlooked and undervalued skill is the ability to explain results to non-technical

stakeholders. As programming became more mainstream in the 1970's and 80s [10], responsibility for data visualization and story telling shifted from people with those specific skills to Quants and programmers, despite their lack of training in the area. Data scientists and quants are inherently interested in pursuing their craft; not improving soft skills. Below are types of miscommunication to consider when thinking of the type of communication skills to look for in your data scientist.

- Quant's curse: They have a insightful analysis, but their bad visualizations, story and connection to business impact loses their stakeholder audience.
- The Help: Stakeholder has an intuition about their day to day job and asks the data scientist to show quantitative evidence. The data scientist does the analysis and feels the insightful, valuable results show a more nuanced story that might help. However, he was unable to convince the stakeholder, who perceived the work as menial number crunching. They both walk away frustrated the other didn't understand.
- Dumb it down, Dumbass: The data scientist is partnered with a great communicator that completely simplifies the quantitative story to the point it is inaccurate. The output is compromised and the data scientist feels the work has been trivialized.
- Expectation Mismatch: Sales team engages the data scientist to perform customer segmentation who comes back with an amazing model of 10 segments. Sales has capacity for 3 segments resulting in neglect of the model, deflating the data scientist's morale.

Accumulating consensus and gaining influence
The data scientist should have end-to-end accountability. They should create the problem statement with the business (not be handed one), engage in the pre-analysis work and in the post-analysis work. This typically requires a veteran data scientist to cut through the political and competing agendas and muddled thinking typical of group projects. They help create a clear problem statement. They should be involved in the measurement of the derived business value.

Many project fail in the *last mile*, where the analysis and insights are packaged and presented to the business. Explaining and convincing decision makers is critical to the success. This is where many projects fall down and it is driven by the communication gap.

2 Successfully building an organizational data culture

2.1 The people

In order top build a successful data product, the following partners are needed for buy-in and participation:

- Business SME partners: Integrate the results into the business/product line and provide feedback on usability of the solution.
- Data SME partners: Depending on the organization, data partners are *not* on the data science team. These are data engineers and software application engineers that facilitate the data scientist.
- Senior Leaders: Bring funding with the mindset that the investment isn't an add-on, but an integral part of the organization. They bring the political clout to get different teams to work together.
- Data Scientist: Assigned based on skill and talent required for project rather then role. If the project requires deep understanding of data lineage, data engineering skills are paramount. If the problem is well-defined and the technology in place, then someone with advanced machine learning knowledge can be leveraged to create the best predictive model.

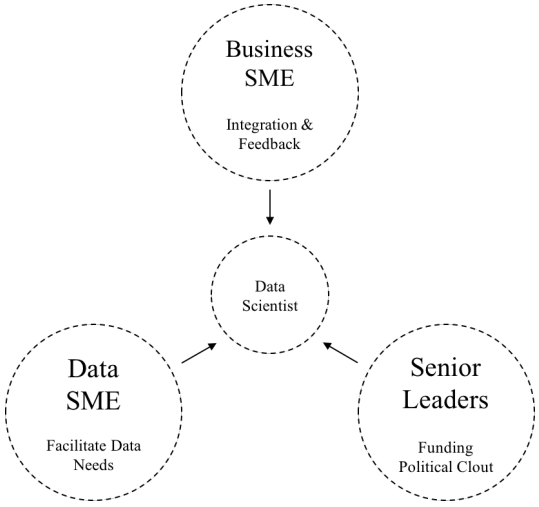

Figure 1: Partners for a successful data culture.

For each project, a lead should be well-defined and empowered to avoid the 'responsibility without authority' trap. Depending on the project and/or phase, this could be the business or a more technical person. For example, if the project is pipeline improvement in order to create a new, well-defined metric, then a more technical person could be assigned. On the other hand, if the project objective is to increase revenue by mitigating client attrition, then a business stakeholder familiar with the issue might be more appropriate. Once the problem is completed, the same teams can be templated and deployed onto other projects.

Talents

Talent	Task	Skill
Project Management	Manage timeline, marshal resources	Organization, people management
Data Wrangling	Find, clean and structure data	Coding, statistics
Data Analysis	Develop and test hypotheses	Statistics, scientific method
Subject Expertise	Define business objectives	Non-technical communication & functional knowledge, critical thinking, ability to define business outcomes
Design & Storytelling	visuals and stories	presentation, persuasive communication

2.2 The Data Science Team

2.2.1 Full-Stack vs. Coalition

There are two methods to achieving a personnel that can address your data science needs: full-stack and coalition [9], each with their trade-offs.

- Coalition: parallel, coordinated necessary, cost of iteration is high
- Full Stack: sequential, limited coordinated required, cost of iteration is low

There is generally the notion that the full-stack data scientist is an elusive unicorn. While a strong argument can be made in favor of the full-stack data scientist, finding or growing one within the organization can be very difficult. It is more typical to find 'half' stack data scientists that can do many, but not all the tasks. For example, a person can be very good at manipulating data very well to create features, building proof-of-concept machine learning models and writing production level code. However, it would be unreasonable to expect that person to focus on those tasks *and* build data engineering pipelines.

The concept of dividing a data science project into atoms aligned to individual talents comes at a cost. This approach to problem solving goes back to Adam Smith's idea of efficiency gains via division of labor [11]. Division of labor is beneficial when trying to achieve productivity gains from a known and established process, and small incremental improvements and functional excellence are being sought after. Data science typically has a step function impact where all new business capabilities are learned and discovered. It typically starts with an undefined task where one learns as they go and the product evolves during development. It isn't a set of well-defined tasks designed up front that are sequentially executed.

While the coalition model to some extent is unavoidable, this specialization comes at a cost:

- coordination time increases super-linearly with number of people [12]
- iteration is costly since coordination required up and down the chain
- wait time is increased as a result of reliance on other teams
- loss of accountability for the final product, as each person is a cog in a chain
- talented people get bored more quickly and become dull when they are required to stay in their lane
- context to each individual becomes more narrow

2.2.2 Centralized vs. Distributed

Once the data science team has been formed, where that team reports organizationally is an important consideration when trying to achieve successful data products. The team can align itself in a centralized way, distributed across all lines of business, or a combination.

- Centralized: One team has the benefit of cross-pollination of best practices among technical people, which is something data scientist types inherently crave. The drawback is they will miss out on the business domain knowledge that will make the data science project a success. Shielding the Data Scientist from the business has its benefits. Protecting and insulating the rare commodity from the day to day politics may limit turn over and allow the person to concentrate on their craft. There are also non-altruistic intentions: by burying the quants, stakeholders can take credit for the output.
- Distributed: Embedding them within the business will decrease the time it takes for them to understand the business and develop a rapport. The drawback is that they will have limited opportunities to discuss highly technical subjects. In addition, her colleagues will not always understand requests/needs, like why they need a cloud computing, and they may feel organizationally isolated.
- Combination: The best of both worlds. Collocation within the business allows them to grow domain knowledge. Organizationally, they have an outlet to express their technical requests, compare themselves with peers and gain career progression opportunities within their craft.

Ultimately, the goal should be to destroy silos. In any of the above cases, consider collocating project stakeholders while the project is active. Domain knowledge and curiosity about each member of the team's work is imperative.

2.3 The Data Culture

Data has changed our relationship to many fields and understanding data is no longer the sole job of quants. For example, athletes are highly aware of their advanced statistic metrics, given the impact on their salary and potential for performance. Data fluency across an organization enhances communication between the business and analytic teams and in turn creates 'better' requests [13]. The goal of the data team should be to go from 'doing everyone else's work' to providing tools and models that create 50x leverage and business impact. The appropriate organizational structure can help achieve this by closing the gap between existing non-technical and a data driven target state.

Gaps between business minded and technology types has existed for a long time. It has grown even more due to specialization. Data scientists, who are trained and enjoy quantitative subjects, are not typically adept at being communicators, politically savvy or business-minded. Organizations are not mindful of hiring to fill in the gap.

It is important that there is cross education of the data scientist skill set. To achieve this, all stakeholders should be included so soft skills permeate. Discussions that are required to become less technical force the data scientist to think in business terms and at a higher level of abstraction. The benefit for the business is that they become more data fluent.

Three methods for creating a data culture and improving chances of success in solving your data problems:

- Share tools: data science teams should not be the gatekeeper of trivial tasks. They should design tools and leverage an engineering team to implement them at scale for the end user to actively use and get familiar with.
- Spread data responsibility: all stakeholders should have access to and understanding of their data.
- Spread data skills: organizations can create a 'Data University', mentorship/pairings, sprint meetings

3 Successful Data Science Products

There is not a standard, one-size fits all answer to unlocking the value of your company's data. The factors that drive successful data science projects and the corresponding steps for success are:

1. Defining the business problems for the organization with appropriate business sponsorship.

2. Evaluation and prioritization of use cases based on a balance of complexity and value: quick wins help morale.

3. Creating an adoption strategy with sufficient resources for execution.

3.1 Defining business problems

Identifying the universe of potential problems

One way is to ask stakeholders across all lines of businesses to track all their activities and identify reoccuring ones [14]. If many across the organization have a common element of that repeated task, then identify if there are any prediction components in it. A non-quant sniff test to see if this is model-able, is to ask: if multiple colleagues were to perform this task, can I write down the instructions that most would agree on the outcome? Then begin to understand the data available. If the task was recently started, chances are there will be limited data. Then try to understand what the penalties for inaccuracy are and the tolerance for errors. If it is clear that the outcome has to be right 100% of the time, write about examples of recall/precision and the difficulties in modeling.

Criteria for each individual problem

For each problem, the following elements should be specified in a few pages by the relevant party or parties in cases where collaboration is required *prior* to selection.

- Business overview with problem statement
- Business value of proposed solution
- Requirements from an end user perspective
- Data sources
- Resources and team
- Timelines, key dependencies, and risks

3.2 Evaluation and prioritization of use cases

For each use case problem statements [15]:

- Define what a solution might looks like and the objective function for success. This isn't r^2, accuracy, or other data science mathematical metrics. It is business outcomes, such as revenue increase, headcount saved, operational expenses reduced, etc. The emphasis is less on the next best algorithm, wins can be achieved with simple models. Simple models allow you to demonstrate effectiveness in a decreased time to create an end to end solution. Once the problem is well-defined, model complexity can be addressed by a data scientist that is highly specialized.
- Identify how solution will impact performance and operations. The derived business value is generally in one of the following buckets: automation, improved decision making, growing business share.
- Define the levers that drive value and assign KPIs. This allows for monitoring of the data product success.

Now that these are defined for each use case, understand the relative cost, time to value, complexity and business value and alignment.

- Cost: Just because a use case might seem expensive, doesn't mean it should be discarded if it can be foundational to other use cases. Identify the connections between use cases to determine foundational solutions. These are the projects to begin digging deeper into to assess the opportunity. Projects must scale and have long term value across multiple applications.
- Time to value: Focusing too much on near term results compromises medium and long term investments. On the other hand, time to value is a consideration - quick wins can improve morale.
- Complexity: understanding the ML/AI feasibility, data availability and implementation details.
- Business value and alignment: building data products *requires* the business to be involved, so their resourcing availability and participation is an important factor. In addition, the data product objective should align with the organizational strategy. A business being phased out should focus more objectives aligned with automation, while a growing business might require use cases to grow client base.

3.3 Create an adoption strategy

Once it has been decided which data product will be build, it cannot be assumed the organization will adopt the product and integrate it into their workflow. Organizational inertia [16] tends to

favor existing methods, rather than adopt new, more efficient ones. Two parts of an adoption strategy should be pre-defined: 1) Proper execution of model development and deployment 2) Organizational communication and teach-ins.

Execution

The model should be built with a staged execution framework in mind. Lightweight models are faster to ship and thus create a tighter more agile feedback loop. Deep learning models are powerful, but not usually the first place to demonstrate gains. Hand curation, or manual work, at the beginning (AAI: artificial artificial intelligence) is not necessarily a bad thing when it serves a purpose to motivate towards a target state. This also brings the business along for the journey.

A specific plan for deployment with an adherence strategy should indicate existing solutions/models will be phased out, requiring the business to adopt and ask questions early.

Organizational

- Define the value proposition to the business stakeholders. The value of the solution may be apparent to the data scientists or project manager, but ultimately, the business needs to understand and believe in the value. Time should be spent convincing them of the gain.
- Communication and training plan
- Tracking and feedback
- Empower a single stakeholder.

References

[1] McKinsey & Company. Notes from the AI Frontier Modeling the Impact of AI on AI on the World Economy. Link.

[2] Google Trends

[3] Kaylan Veeramachaneni, MIT Labs. *Why You're Not Getting Value From Your Data Science.* HBR.

[4] Thomas C Redman, Data Quality Solutions. *Do Your Data Scientist Know the 'Why' Behind Their Work?* HBR.

[5] 7 Reasons to Fall in Love with a Data Scientist.

[6] Data Scientist: The Sexiest Job of the 21st Century

[7] Yael Garten, Apple - Siri. *The Kinds of Data Scientists.* HBR.

[8] Cassie Kozyrkov, Google. *What Great Data Analysts Do - and Why Every Organization Needs Them.* HBR.

[9] Eric Colson, StitchFix. *Why Data Science Teams Need Generalists, Not Specialists.* HBR.

[10] Scott Berinato, Good Charts. *Data Science and the Art of Persuasion.* HBR.

[11] Wikipedia: Adam Smith

[12] Steiner, I.D. (1972). Group Processes and Productivity. Academic Press.

[13] Jonathan Cornelissen, DataCamp. *The Democratization of Data Science.* HBR.

[14] Kathyrn Hume, integrate.ai. *How To Spot A Machine Learning Opportunity.* HBR.

[15] Hillary Mason, Cloudera. *How to Decide Which Data Science Projects to Pursue.* HBR.

[16] Clark G. Gilbert. The Academy of Management Journal. Vol. 48, No. 5 (Oct., 2005), pp. 741-763