Developing Analytics Solutions with the Data & AI Global Practice

How to deliver a data science project

★★★★★

September 4, 2018 by Michele Usuelli // 0 Comments

0 0

Author: Michele Usuelli, Lead Data Scientist, Microsoft Enterprise Services

What is a data science project? How does it get delivered? What are the key aspects?

There is a lot of hype around data science. The purpose of this article is to provide clarity and the content comes from the successful deliver of project with Microsoft Consultancy Services.

The topic is very broad and this article highlights some key questions rather than providing answers.

To start, let's have a look at three areas of focus, covered by the related sections:

- Who to work with and how: "Keep the key people involved".
- How to build the product: "Iteratively build the product".
- What are the dependencies and limitations: "Proactively highlight dependencies and limitations".

The assumptions are that the project already has a measurable business impact and a clear scope, and that the data is available and relevant.

Keep the key people involved

With "people" we are referring to specific roles in a project. There might be more than one person covering the same role whereas the same person could cover more than one role.

There are several people involved in a data science project and they belong to either of these two groups:

- **Delivery team**: people actively building the product. To keep the article simple and focused, we only mention the data scientist. Other aspects of the project, such as the IT infrastructure work, are equally important, but they're not covered for the sake of simplicity.
- **Everybody else who is involved**: who is related either to the context or to the outcome of the project. The key people are the stakeholder, the domain expert and the end user.

Depending on the project, the team structure might differ. However, there is always someone sponsoring the project, someone providing domain expertise and someone ultimately utilizing the product. The same person could cover more than one role, but it's easier to think in terms of people to structure how the data scientist captures different perspectives.

Let's have a closer look at the three key people:

- The **stakeholder** is ultimately impacted by the outcome of the project. Therefore, the data scientist should get constant feedback from the stakeholder to make sure that the project is heading towards the right direction and to tune the project objective in the most impactful way. Also, to keep the stakeholder engaged in the project, it's usually a good practice to provide them with quick results driving some value or at least showing useful insights.
- The subject-matter **expert** has domain knowledge related to the context of the problem and the meaning of the data. Therefore, the input of the expert should be the starting point of any data science project. Also, the data scientist will be able to discover new insights that will benefit the expert, and about which the expert will be able to provide useful advices. The checkpoints between the data scientist and the expert should be frequent.
- The **end user** is the person ultimately utilizing the product. To design a solution for success, a key aspect is to be aware of its implication. Therefore, the data scientist needs the input and feedback from the end user as well.

The key questions are:

- Do we have a stakeholder, an expert and a user actively involved in the project?
- How often is the delivery team communicating with them?

Iteratively build the product

We identified three key people the data scientist interacts with. This section highlights:

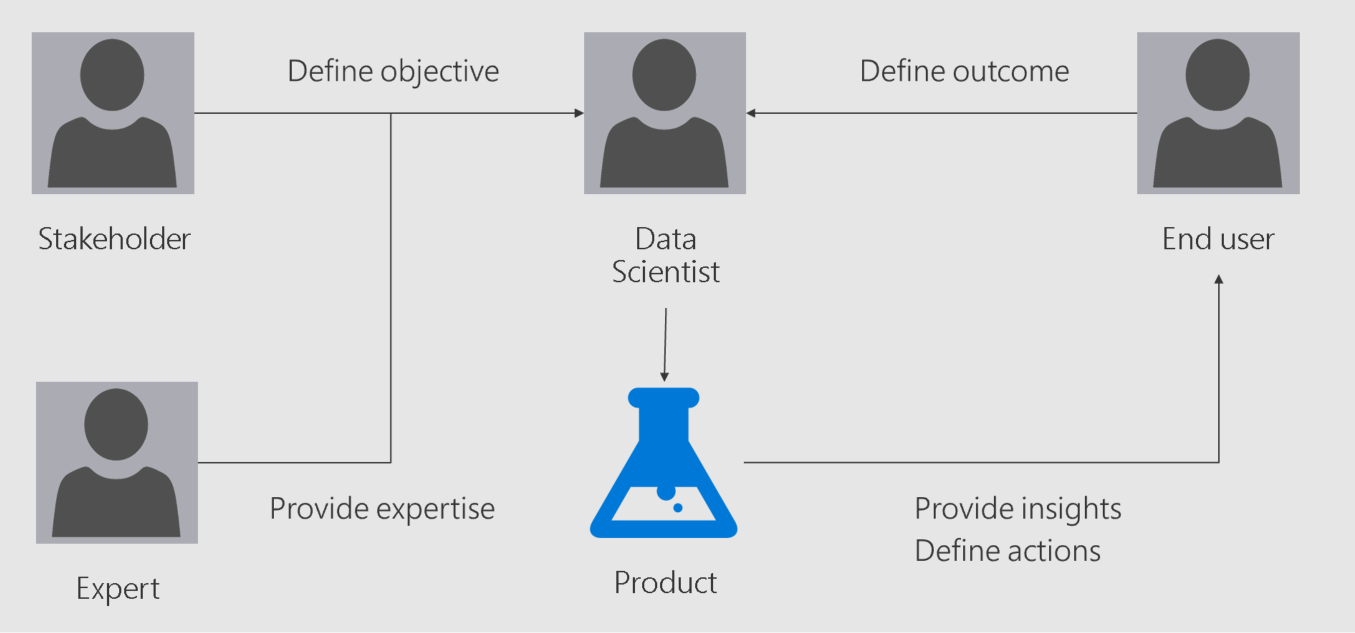
- What inputs the data scientist gets
- What outcomes the data scientist gives

Some relevant aspects are:

- Stakeholder management:
    - What questions are addressed by data science? What is the business impact if the questions remain unanswered?
    - Is there any quantifiable value or is the project mostly driven by hype?
    - Is data science the most appropriate methodology to answer the questions?
- Leverage the expert knowledge:
    - How much business expertise is the model capturing?
    - Do the results validate any prior business expertise? Do they highlight new areas that were previously unknown?
    - Is there anything contradictory in the results?

- Provide the end user with transparent results:
    - Is it possible to explain the overall flow of the solution in simple terms?
    - Can the end user trace the data sources and how they are being used?
    - If we weren't using a machine learning model, how would the solution design differ?

The following chart summarizes the interactions

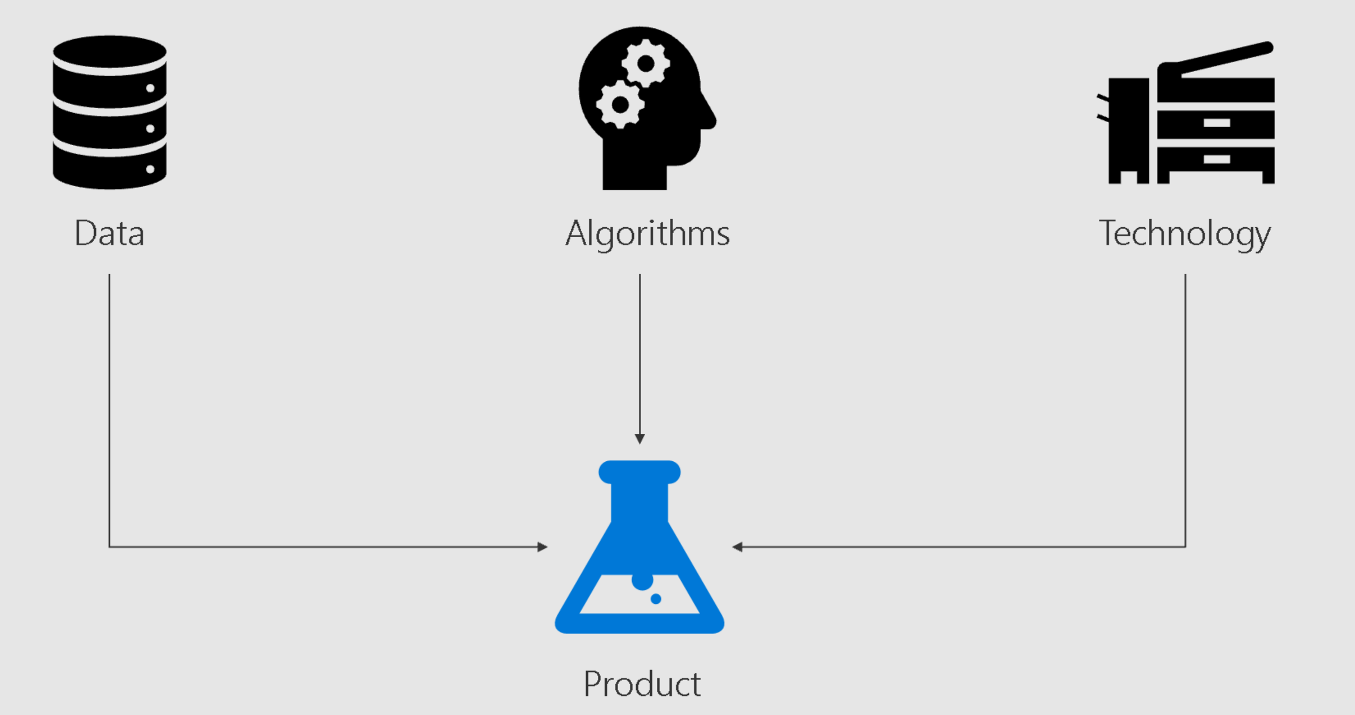


The information required to deliver the project often depends on the findings from the data. Typically, at the beginning of a project, the data scientist is not fully aware of all the required information. Therefore, it's crucial to keep all the three players engaged throughout the project to iteratively get more input and feedback. The product is built in small steps and in an agile manner. Having regular checkpoints with the key people ensures that the data scientist has the information needed to define the next steps.

Proactively highlight dependencies and limitations

To scope a project for success, it's crucial to proactively highlight dependencies and risks. Besides the input from the key players, data science projects require some technical input that can be summarized in these three areas:

- Data: the sources of information being used by the model.
- Techniques: the algorithms that discover patterns from the data.
- Technology: the technical infrastructure to store the data and run the algorithms.



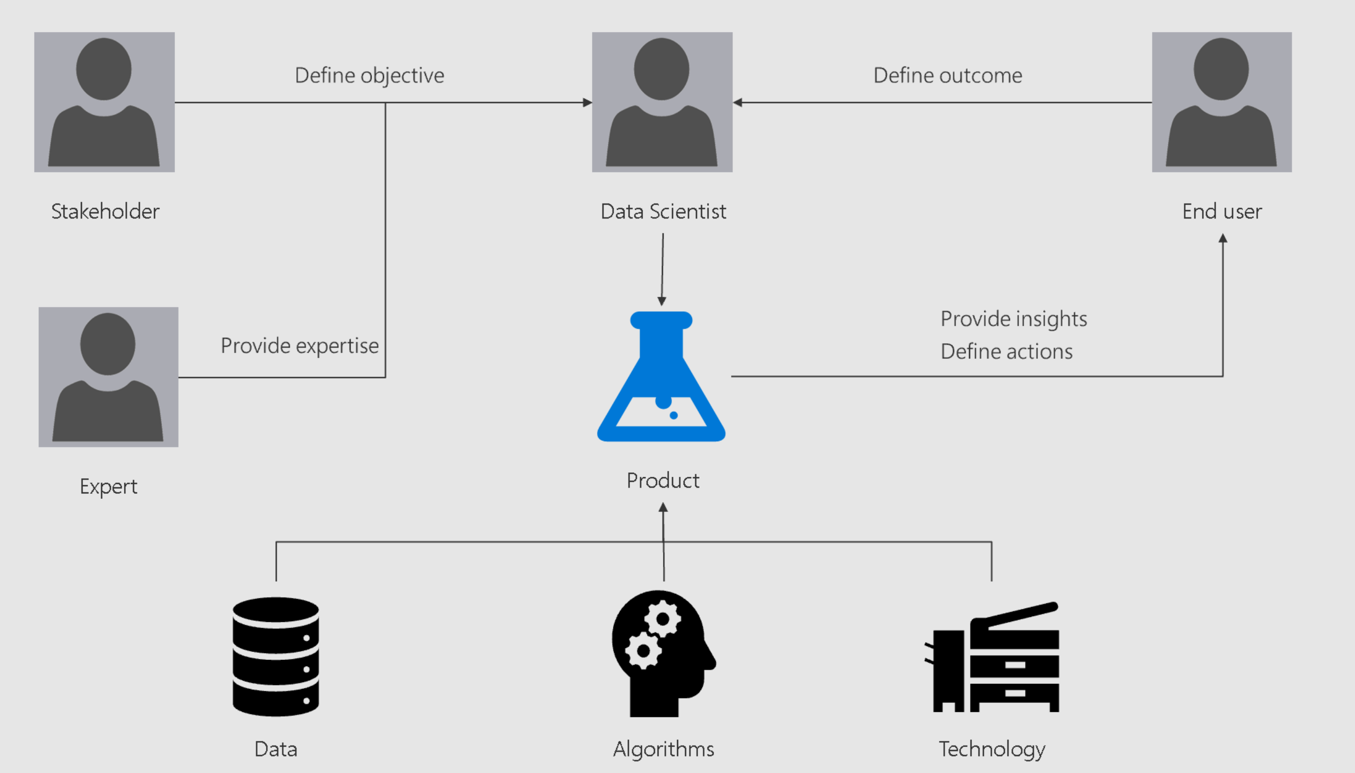
Each area has a strong impact on the goodness and performance of the solution. Some key questions are:

- Data
  - Is the data consistent and complete? Is it possible to match the data sources?
  - Does the data capture most of the relevant information? What is missing?
  - Do we have enough history of data?
- Techniques
  - Are the techniques explainable?
  - Are the techniques simple and generic?
  - Do the techniques perform consistently if applied on another dataset?
- Technology
  - Can all the data be stored and analysed?
  - How fast can the solution be built and applied?
  - Can the data volume be handled?

Data, techniques and technology define the boundaries of the performance. To have a reliable and trusted solution, it's crucial to highlight the boundaries since the very beginning of the project and to regularly track them.

Conclusions

The chart below summarizes the article.



This article hasn't covered all the aspects. Other important topics are the IT infrastructure work and the attribut

Popular Tags

#ArtificialIntelligence #DataScience AI AML Analytics Apache Spark APS Artificial Intelligence Azure Data Factory Azure Key Vault AzureML Azure SQL Data Warehouse Basket analysis Center of Excellence Centre of Excellence CoE Data Data Science Data Scientist Role Decision forests Event Hubs k-fold Machine Learning Measurability Model Goodness Modelling Power BI PowerShell ML Scoring R Random Projection R Services Scikit SQL Server 2016 SQL Server R Services Stream Analytics Visualizations Whitepaper

Archives

November 2018 (1)
All of 2018 (4)
All of 2017 (3)
All of 2016 (11)
All of 2015 (11)

Tags #ArtificialIntelligence #DataScience

Join the conversation

Add Comment

Dev centers

Windows

Office

Visual Studio

Nokia

Microsoft Azure

More...

Learning resources

Microsoft Virtual Academy

Channel 9

Interoperability Bridges

MSDN Magazine

Community

Forums

Blogs

Codeplex

Support

Self support

Programs

BizSpark (for startups)

DreamSpark

Imagine Cup

Newsletter Privacy Terms of use Trademarks