# A Data Scientist Perspective on Knowledge Graphs (Part 1): the Data-Driven Challenge

*Gautier Van Rossom*

*Photographer: Johannes Plenio*

This series of articles is a data scientist, and data engineer, perspective on knowledge graphs, which is intended not only for other data scientists and engineers, the nerdy role in the office that no one truly understands; but also for executives and business groups, who ultimately decide where to steer the organization, and are inundated with a multitude of use cases and business capabilities; as well as for project managers, who are tasked with leading a group of cross-functional teams to move their data projects into successful efforts.

The goal of this series of articles is not to describe what data science, engineering, or machine learning is, but it will ultimately depict what these are, and the reason why we hear about these distinct names, or roles, that intricately work together. Typically, these roles are executed by the same person in small teams, hence creating the confusion.

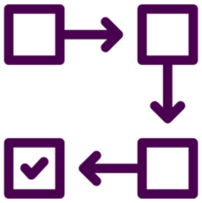
**Data Science is applying a scientific approach to solving a business challenge using enterprise data, or simply to provide an answer to a business question. In the same way, Data Engineering shall provide a holistic solution (e.g., via integrations with applications, data orchestration, etc.) to a business problem, or challenge. In short, data science will be more about prediction while engineering will be about automation.**

This article, part 1, will be focusing on the data scientist path from knowledge discovery to solving a business challenge. To understand their perspective, we will explore the challenges facing data scientists and discuss knowledge management as a solution. In the second article of this series, part 2, we will see how knowledge graphs apply to data science and machine learning in the enterprise or business context.

## A Scientific Approach to Business

The important word in data science, from a data scientist's perspective, is *science*, not data. And science starts with a question (or hypothesis). Data comes in secondly to validate or refute a hypothesis, significantly answering the question, or not.

From a business perspective, this approach means that what matters first is the initial question. Asking the wrong question will always lead to the wrong answer.



From our experience within the enterprise, it means that a data scientist might have a cross-functional role based on the question asked, or the problem exposed, as the data needed to answer the question, or solve the problem, might be spread among different units.

Simple problems usually require simple solutions. Data scientists might fit in a single department but we gather that they are often involved in more than one, sitting between departments.

The key role of a data scientist is to answer questions with data, find the model that best suits a problem, assess its performance by developing quality assurance tests (statistical tests), and determine what is needed (often better or more data) to improve the model.

Most of the research executed by a data scientist will consist of refining the initial question asked, or redefining the initial challenge exposed. Usually uncovering other questions, or challenges and iteratively making them more precise and more contextualized.

Here are a few examples of the questions or hypotheses that data scientists are confronted with:

- What will be our revenue next year?
- How can we maximize profits?
- How can we increase sales?
- Which products or services should we prioritize?
- Which marketing campaign brings in more customers?

As we can see, the extent to which these questions apply is vast and they are mainly business economics questions, although data science can apply to more operational or organizational questions as well:

- How can we improve our processes?
- How can we increase service uptime?
- How can we optimize tasks among employees?

Taking a scientific approach, the initial question that gave life to this series of articles shall be: *What is a data scientist's perspective on knowledge modeling and engineering in a business/enterprise context?*

Let's contextualize it in order to further refine this question.

## The Data-Driven Challenge

Businesses are more and more confronted with AI which is now becoming ubiquitous. We hear about data scientists and engineers, sometimes AI or machine learning (ML) engineers, automating business processes, developing predictive models, and many other algorithmic things.

Artificial intelligence is now involved in many, if not most, business processes. Indeed, there are questions to answer, and challenges to overcome, at all levels and in every department of a company. Executives have strategic problems – where to go, how to innovate? Businesses have business problems – how to earn or do more with less. A level lower, organizations have organizational problems – who needs to know what, what do we need where, etc.

> **"Companies must re-examine the ways that they think about data as a business asset of their organizations. Data flows like a river through any organization. It must be managed from capture and production through its consumption and utilization at many points along the way." -** Randy Bean, Harvard Business Review.

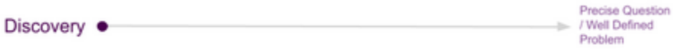
Businesses now have data lakes, because data is structurally siloed across their company. As the amount of data gathered is tremendous, they hired data scientists and engineers in order to make sense of it all.

In theory, everybody knows that. In practice, it is never as easy as it sounds, with people typically not knowing where to start. In the next section, we will dive into the data scientist's path forward.

*Photographer: Rodrigo Pederzini*

## The Data Scientist Path

Although business shall prevail over technology. The very empirical nature of economics forces the other way around when it comes to data science and engineering applied to business. The term *data-driven* depicts it best, but the issue here is that a data scientist can be left over with only data and a simple "find something" instruction. We will indicate this extreme case of pure discovery, at the very beginning, on the far left of our path, as shown in the following figure.



The more we move to the right side of the data science path depicted above, the closer we get to a precise business question, or challenge, and consequently the closer we get to an applicable solution.

Although its path might not be as straightforward, a data scientist strives to move from left to right.

The data scientist's path is similar for most questions or challenges and consists of data quality assessments (is the data appropriate, sufficient, accessible, and discoverable?), exploratory data analysis (can we extract patterns or trends from the data?), and feasibility checks (can we get actionable results or business insights?).

In practice, data assessment and exploratory data analysis are always about the same process, and the answers will depend on the initial question asked. As most cases follow the same path, the only difference is the time it takes and the end result, both of which depend on the question asked, or challenge definition.

Let’s have a first look at the two extreme scenarios, two very different questions, or problems, that are “find something” and “find this”, and see how they will differ in practice. We will also see how machine learning algorithms fit between these two. And finally, we will picture for each scenario what are the best and worst cases we can expect from them.

## Scenario 1: No direction, purely discovery

The enterprise generates tremendous amounts of data, about customers, employees, or resources, and we want to make sense of it all, or simply extract some business insights out of it.

The data scientist gets confronted with an overly general demand: “Find something in our data.”

In machine learning terms, most models in use are of the unsupervised learning family that will end up in a classification, or categorization, problem. This is commonly called information or knowledge discovery or retrieval.

Regardless of the output, businesses end up with the same tricky question: how to ensure value out of it? And tackling this issue is fairly simple: define value.

Indeed, this scenario requires at least some business context and objectives, otherwise the team might dig into pointless directions.

**Worst case:** the project ends up in endless research or inapplicable findings, impacting the ability to retain the team with a value proposition. **Best case:** the project ends up in a classification problem, or categorization, leading to the next scenario; more precise questions, better defined challenges.

To avoid the worst case and achieve the best results, enterprises and managers should contextualize, or structure, our research and findings. This shall ultimately lead to more specific questions, which is the second scenario presented here after.

## Scenario 2: Precise Business question

The organization has a specific question, or goal. The data scientist’s job will be to study the feasibility of the question regarding data availability and quality, and the potential answers based on statistical significance and information availability (engineering), which together shall be able to conclude, potentially providing an answer to the initial question.

Because we don’t know in advance whether we shall be able to conclude or not on a question - often due to lack of or poor data. The result, the answer, to a question is uncertain.

Therefore, this scenario requires data scientists to have a list of (multiple) precise (well-defined) business questions that address specific business challenges or use cases. Having multiple questions increases the chance of gaining valuable insights from the analysis.

The machine learning models at play here are typically of the supervised learning family. Although many solutions require solely simple, more intricate, or time series regressions depending on the problem.
The result is a list of feasibility checks, prototypes, or proofs of concept.

**Worst case**: the data is not as good as expected, or is simply not available at the moment. The problem gets postponed to when data is available.

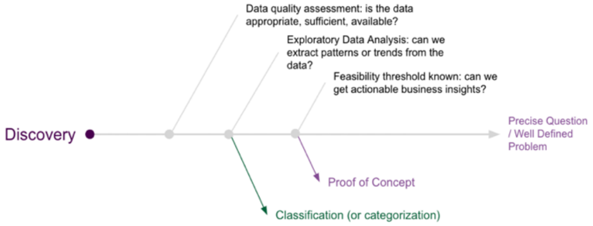
For example, a monthly prediction requires, due to yearly seasonality, at least 24 months of data.

**Best case**: the data is good and the model works. We have a proof-of-concept that can lead to an implementation or integration phase.

Again, to avoid the worst case and achieve the best results, enterprises and managers should contextualize, or structure, our questions and answers.

## Closing

The differences between the two scenarios discussed above are, of course the output, whether it shall end up in a classification or in a predictive problem, and how timely (and expensive) they are, both depending on the initial question, or challenge definition.


Data quality assessment: is the data appropriate, sufficient, available?
Exploratory Data Analysis: can we extract patterns or trends from the data?
Feasibility threshold known: can we get actionable business insights?
Discovery
Precise Question / Well Defined Problem
Proof of Concept
Classification (or categorization)

In reality, the difference between "find something" and "find this" is rather significant.
"Find something" can lead to unnecessary answers, such as solutions without a problem. We will see later in this paper how to avoid that situation.

"Find this", "Find why this is", "Find how this is", or "Find a solution to this", are already more precise and tangible questions but require "this" to be defined.

Companies will often place value in being data-driven, or following the data-driven approach, but we've seen here that an organization can be data-driven yet still ask the wrong questions. The value of a data science project is defined by the initial question. The data-driven approach is most valuable when the initial question is valuable to the business, meaning the answer to the question can be leveraged and have an impact on the enterprise.

A successful data science project starts with a good question. It does not necessarily mean that you will get a valid answer to your question, but rather that you will be able to answer the question with the data that you have.

Overall, we ensure value from a data science approach by generally being able to use information or knowledge extracted from discovery in order to tackle precise business questions or well defined problems.

The way our data scientist's path fits within the company will be the subject of the second article, part 2, of this series. We will first put our data scientist's path within the enterprise context, and second, we will see how knowledge graphs come in handy for that matter. Indeed, we will see how discovery in data science naturally leads to knowledge modeling and in turn how knowledge modeling helps define better, more precise, questions.

**Key Takeaways:**

- Data Science is the process of asking questions and answering these questions with data in a reproducible way.
- In practice it is about refining a question or problem.
- The first steps in a data science or machine learning approach will be an exploratory data analysis to assess data quality and capacity to answer the initial questions.
- Data science or machine learning ultimately ends up in a classification or a prediction problem, the output depending on the initial question asked.

Enterprise Knowledge (EK) is a services firm that integrates Knowledge Management, Information Management, Information Technology, and Agile Approaches to deliver comprehensive solutions. Our mission is to form true partnerships with our clients, listening and collaborating to create tailored, practical, and results-oriented solutions that enable them to thrive and adapt to changing needs.

Our core services include strategy, design, and development of Knowledge and Information Management systems, with proven approaches for Taxonomy Design, Project Strategy and Road Mapping, Brand and Content Strategy, Change Management and Communication, and Agile Transformation and Facilitation. At the heart of these services, we always focus on working alongside our clients to understand their needs, ensuring we can provide practical and achievable solutions on an iterative, ongoing basis.

info@enterprise-knowledge.com | 571-403-1109 | @EKConsulting