[DS-01] Introduction to Data Science

Miguel-Angel Canela
Associate Professor, IESE Business School

Data Science and Data Mining

The expression **data scientist** is trending these days in job descriptions, referred to a mix of data analysis skills and a background of programming languages and databases. But it is, somewhat, a new name for an old job. Most of the methodology has been available for years, but, owing to the explosion in the amount of data at hand and the technology for processing these data, Data Science got hot a few years ago.

An ancestor of Data Science is **Data Mining**, a generic expression which applies to a heterogeneous set of methods, used to extract information from large data sets. The expression is understood as *mining knowledge from data*. Data Mining was born in the computer science field. The typical applications in management are related to Customer Relationship Management (CRM): market basket analysis, churn modeling, credit scoring, etc.

Also closely related to Data Science is **Machine Learning** (ML), a branch of **Artificial Intelligence** (AI). The objective of Machine Learning is the study of systems that can learn from data, but many of the methods were the same as those of Data Mining. Nowadays, Machine Learning and Artificial Intelligence are popular in the business world, due to the push of tech giants like IBM, Google and Facebook. Also, algorithms are now regarded as something common in many organizations, and the ability of developing, maintaining and optimizing them may be included in job requests.

Data Science in the computer

The user interacts with the Data Science software applications in three possible ways:

- Conventional menus.
- Programming code.
- Visual programming, based on flow charts which are a graphical translation of code.

This course is based on code. More specifically, it uses either R or Python, which are, currently, the leading choices of data scientists. Just a few years ago, Data Mining textbooks, such as Larose (2005), were using visual programming or menus in their examples, but, nowadays, almost all the Data Science books are based on R or Python, and the examples include code.

Many Data Science methods apply to data in structured form, that is, to data sets in tabular form, with rows and columns. The rows correspond to **instances** (also called observations, cases and records), such as individuals, companies or transactions. The columns correspond to **variables** (also called attributes and fields).

Typically, the variables are either **numeric**, as the amount paid in a transaction, or **categorical** (also called nominal), as gender. Nevertheless, there are also methods for dealing with **string** (text) data and dates. Categorical variables are typically managed through **dummies**, or attributes with 1/0 values.

In the computer implementation of Data Science, tabular data sets are managed as objects called **data frames**. Data frames were born in R, but have been adopted by other languages like Python and Scala. A data frame is a collection of data (column) vectors, all of the same length. All the data points in the same column are of the same type, but the data frame can contain vectors of different type.

What is a data scientist?

Data scientist is a broad job title coming in many forms, with the specific demands depending on the industry, the business and the role. So, certain skillsets suit certain positions better than others. Data scientists do not do anything essentially new. We have long had statisticians, analysts, and programmers. What is new is the way different skills are combined in a single profession.

It is capital for a data scientist to be able to ask the right questions. This is harder to evaluate than specific skills, but essential. It involves domain knowledge and expertise, coupled with the ability to see the problem, the available data, and match up them. It also requires empathy, neglected in most technical education programs.

They also must be able to communicate, and are valued for their ability to create narratives around their work. They do not live in an abstract, mathematical world; they understand how to integrate the results into a larger story, recognizing that if their results are not leading to action are meaningless.

In a survey at Quora, including the title of this section as one of the questions, Michael Hochster made a distinction which has made fortune. For Hochster, purpose is the biggest factor that dictates what form data science takes. He distinguishes between:

- **Type A** (analyst): focused on static data analysis. Essentially a statistician with coding skills. Very similar to a statistician, but knows all the practical details of working with data not taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on. Task example: business intelligence.
- **Type B** (builder): focused on building data products. Essentially a software engineer with knowledge in machine learning and statistics. Also a very strong coder. The type B data scientist is mainly interested in using data "in production". Task example: recommendation systems.

References

1. E Alpaydin (2016), *Machine Learning*, MIT Press.
2. DT Larose (2005), *Discovering Knowledge in Data*, Wiley.
3. F Provost & T Fawcett (2013), *Data Science for Business — What You Need to Know About Data Mining and Data-Analytic Thinking*, O'Reilly.