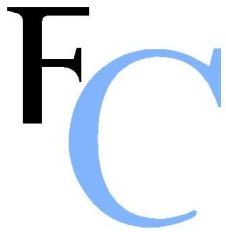
A Brief Introduction to Data Science

L. Torgo

ltorgo@fc.up.pt

Departamento de Ciência de Computadores / Faculdade de Ciências
Universidade do Porto

Feb, 2021



What is Data Science?

A possible definition:

> *Data Science is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner*
>
> *in* Principles of Data Mining (Hand et.al. 2001)

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

ForbesBrandVoice

TRANSFORMATIONAL TECH 6/26/2014 @ 11:00AM | 11,804 views

The Hottest Jobs In IT: Training Tomorrow's Data Scientists

EMC Contributor , EMC

Forbes

By Bob Violino

"The McKinsey Global Institute ... has predicted that by 2018 the United States could face a shortage of between 140,000 to 190,000 people with deep analytical skills, as well as a shortage of 1.5 million managers and analysts who know how to use the analysis of big data to make effective decisions."

So Data Mining/Science is...

At the end of the day Data Science is just
the Analysis of Data!

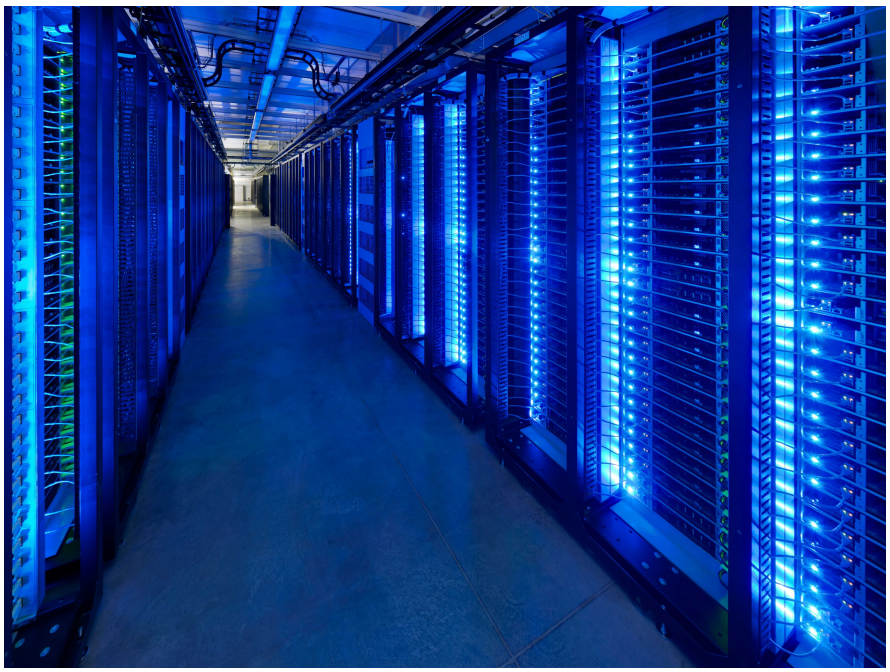
2012 - Data Science

2007 - Business Analytics

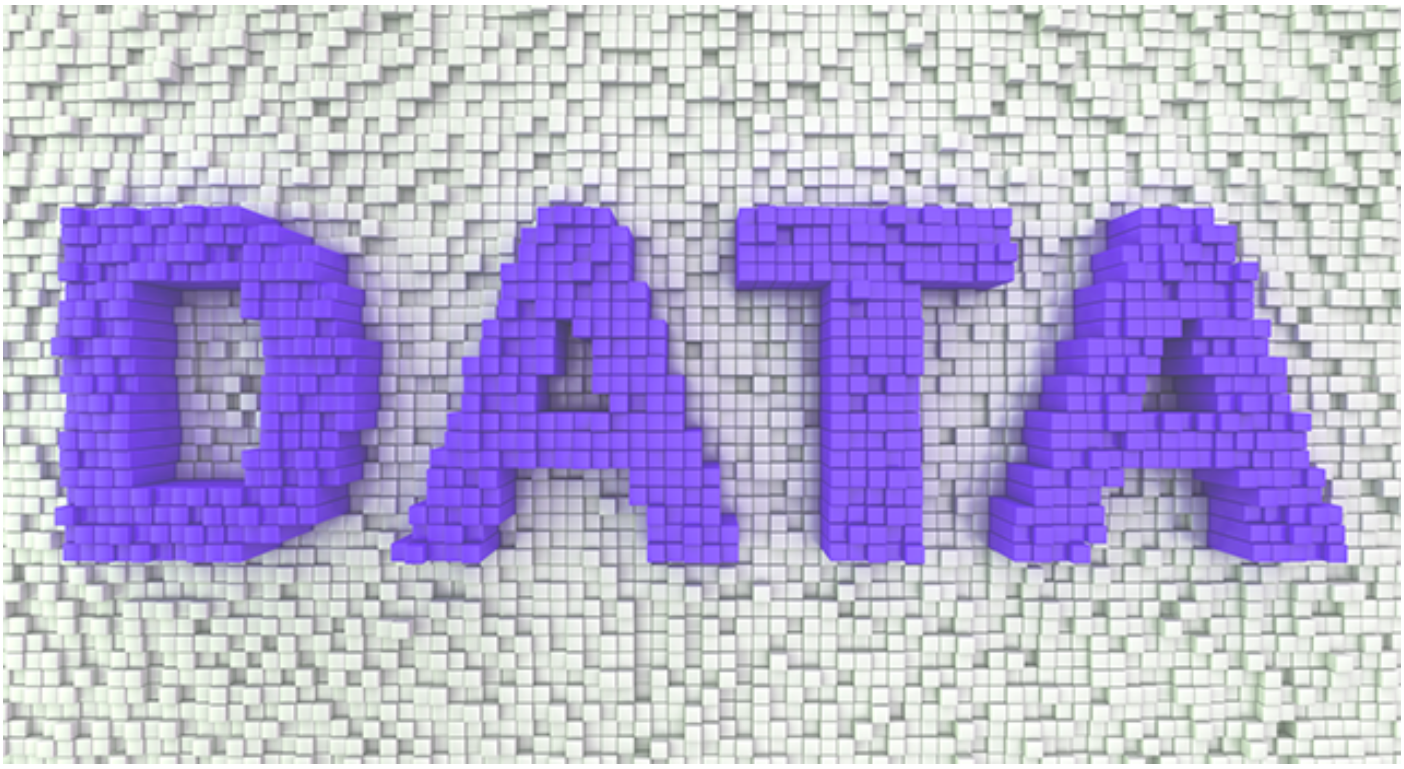
1990 - Data Mining / Knowledge Discovery in Databases (KDD)

? - surely yet another buzzword!

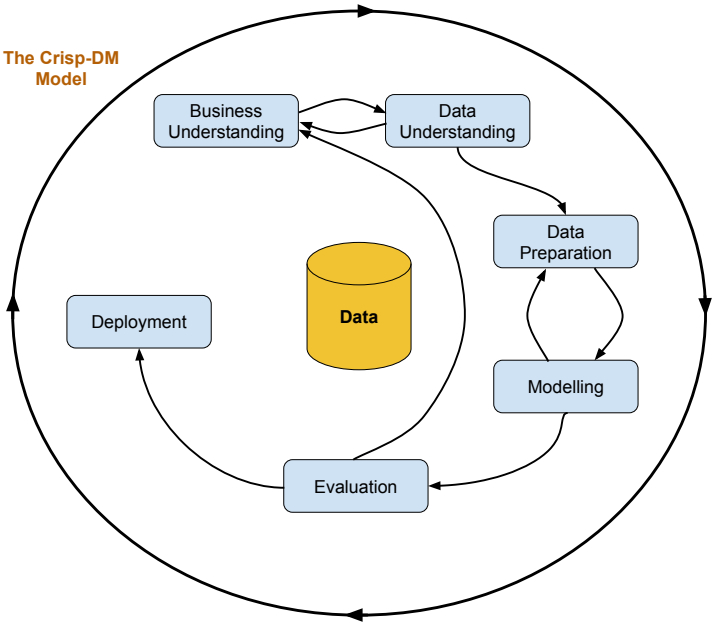
But is Data Science really “standard” data analysis?...



So Why all the Fuss about Data Science?

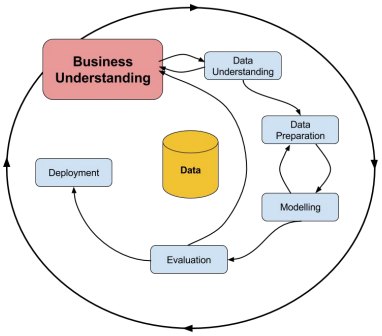


The Typical Data Science Workflow



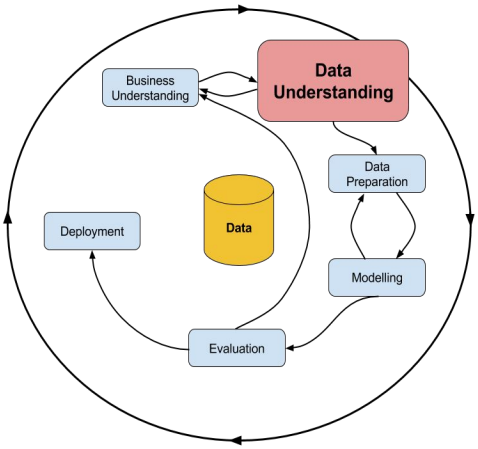
Shearer C.: *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22.

CrispDM - Business Understanding



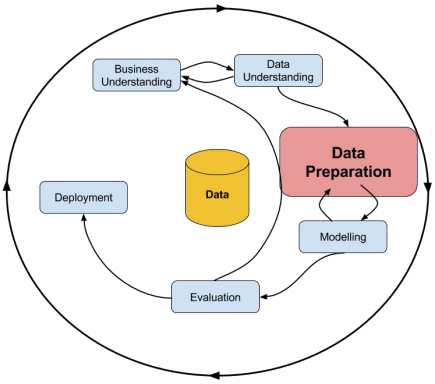
- Determine Business Objectives
  - Background
  - Business objectives
  - Business success criteria
- Assess Situation
  - Inventory of resources
  - Requirements, assumptions and constraints
  - Risks and contingencies
  - Terminology
  - Costs and benefits
- Determine Data Mining Goals
  - Data mining goals
  - Data mining success criteria
- Produce project plan
  - Project plan
  - Initial assessment of tools and techniques

CrispDM - Data Understanding



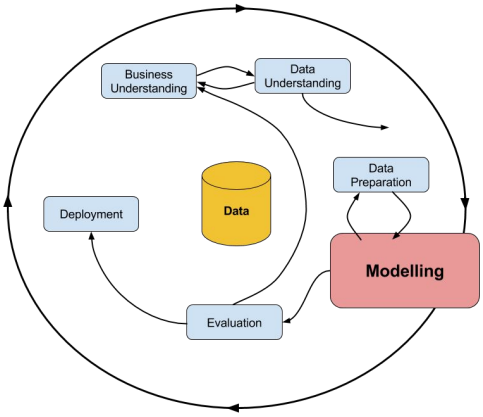
- Collect Initial Data
  - Initial data collection report
- Describe Data
  - Data description report
- Explore Data
  - Data exploration report
- Verify Data Quality
  - Data quality report

CrispDM - Data Preparation



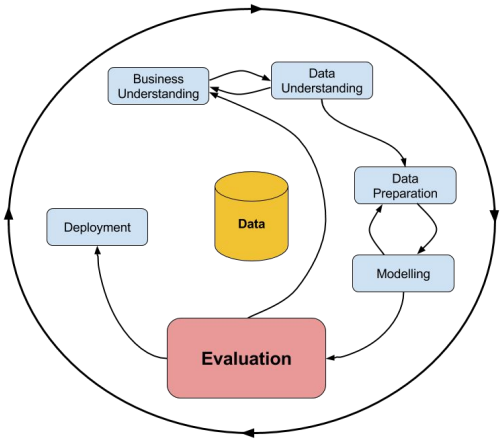
- Data Set
  - Data set description
- Select Data
  - Rationale for inclusion/exclusion
- Clean Data
  - Data cleaning report
- Construct Data
  - Derived variables
  - Generated records
- Integrate Data
  - Merged data
- Format Data
  - Reformatted data

CrispDM - Modelling



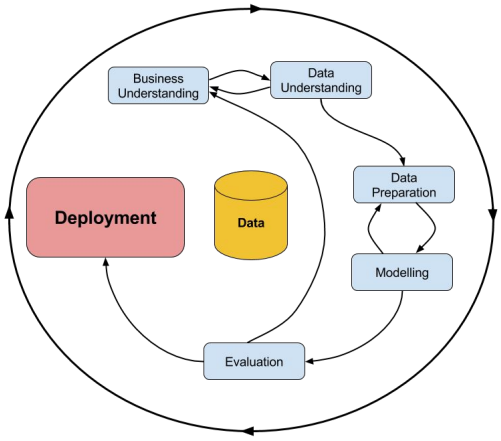
- Select modelling technique(s)
  - Modelling technique(s)
  - Modelling assumptions
- Generate Test Design
  - Test design
- Build Model(s)
  - Parameter settings
  - Models
  - Model despcription
- Assess Model(s)
  - Model assessment
  - Revised parameter settings

CrispDM - Evaluation



- Evaluate Results
  - Assessment of data mining results w.r.t. Business success criteria
  - Approved models
- Review Process
  - Review of the process
- Determine Next Steps
  - List of possible actions
  - Decision

CrispDM - Deployment



- Plan Deployment
  - Deployment plan
- Plan Monitoring and Maintenance
  - Monitoring and maintenance plan
- Produce Final Report
  - Final report
  - Final presentation
- Review Project
  - Experience documentation

Data Sets

Data Sets

- A data set is a collection of measurements taken from some environment.
- In the simplest case, we have p measurements for a set of n objects, i.e. a data matrix of dimension $n \times p$. The n rows represent the objects for which we have collected data. The p columns represent the measurements that were made for each object.
- The rows of the data matrix are also often named examples, instances, records or cases, while the columns are sometimes referred to as variables, features, fields or attributes.

An example of a data matrix

Age	Sex	Area	Income
45	m	insurance	85000
32	f	education	72500
24	f	services	97000
.	. . .	. . .	. . .

Table: An example of a data table (matrix)

Types of Measurements

- Quantitative measurements
  - Integer values
  - Real numbers
- Categorical measurements
  - Ordinal variables (implicit ordering among values - small, medium, large)
  - Nominal variables (no order - red, blue, yellow)

Types of Data Sets

- Simple data tables (the most common situation)
- Databases (multiple data tables related with each other)
- Data streams, time series
- Text
- Multimedia data (images, sound, etc.)
- etc.

Models

Type and structure of the models

- Global
- Local

- Mathematical formulae
- Logical formulae
- Black boxes
- etc.

Different models frequently lead to different compromises in terms of understandability and predictive accuracy

Examples of different models

- Logical formulae - decision rules

```
IF amount = high AND salary = low AND employment = short.term
THEN risk = high

IF amount = average AND salary = high
THEN risk = low
```

- Mathematical formulae

```
houseValue = 10.5 + 5.2 * nrRooms - 3.1 * distCenter + 2.6 * area
```

Tasks

Some of the Main Data Mining Tasks

- Exploratory Data Analysis
  - summarisation and visualisation tools
- Descriptive Models
  - probabilistic models
  - clustering models
  - association rules
  - anomaly and deviation detection
- Predictive Models
  - classification models
  - regression models