ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ADA APPLIED DATA ANALYSIS

School Of Computer And Communication Sciences

Applied Data Analysis Summary

Prof. Catasta Michele
Distributed Information Systems Laboratory (LSIR)
michele.catasta@epfl.ch

June 10, 2016

Contents

1 Introduction

1.1 General information about the course

This course covers multiple topics in the data science field such as **Data Wrangling**, **Data Management**, **Data Mining**, **Machine Learning**, **Visualization**, **Statistics** and **Story telling**. It's about **breadth**, not depth. Indeed, Data science is evolving really quickly, hence learning in depth a specific tool won't pay off.

1.2 Data Science

When we talk about Data Science, we often use the term Big Data as the enormous amount of data that exist in the world. But Big Data is not only about collecting huge amount of data. It is challenging but not enough. The real value comes from the insights. The *internet* companies (Google, Facebook, etc.) understood this many years ago.

An accurate definition of Data Analysis is given by Wikipedia:

> **Analysis of data** *is a process of* **inspecting**, **cleaning**, **transforming**, *and* **modeling data** *with the goal of* **discovering useful information**, *suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names,* **in different business**, *science, and social science* **domains**. Wikipedia - Data Analysis

Therefore, a Data Scientist has to master different kind of skills such as **Mathematics** (for the Statistics), **Programming** and the **Domain Expertise**. Drew Conway's Venn diagram, Figure 1, shows the different combination man can obtain with these three skills.

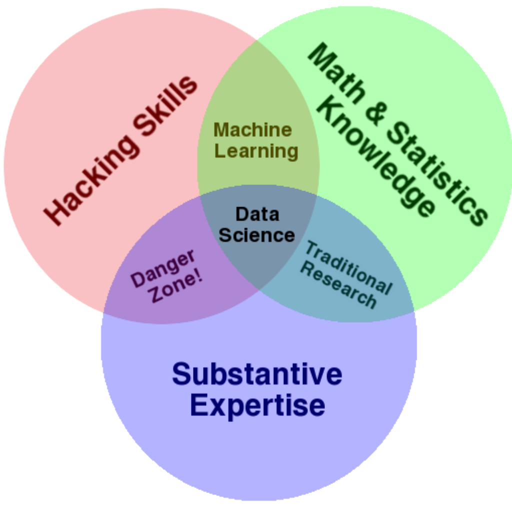


Figure 1: Venn Diagram describing the different combination of skills used by a Data Scientist (by Drew Conway)

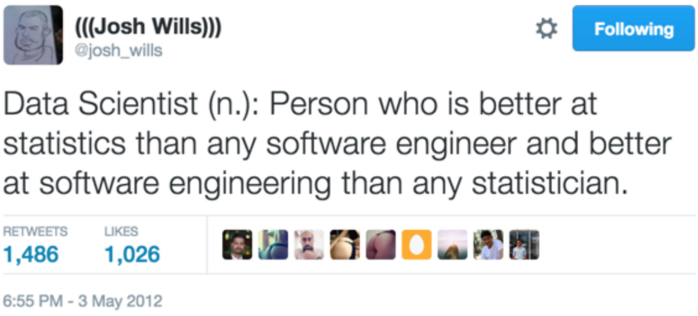
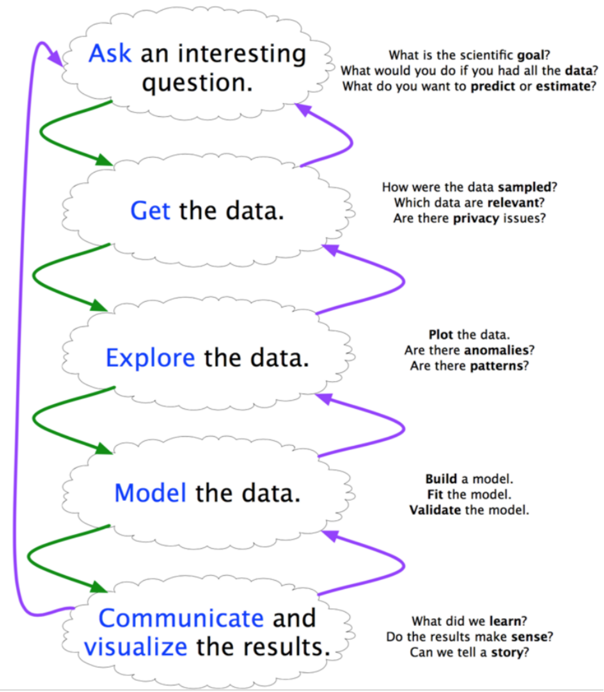


Figure 2: A tweet from Josh Wills, Data Scientist at Slack.

A practical definition of Data Science

Data Science is about the whole processing pipeline to extract information out of data. As such, a Data Scientist **understands and cares about the whole data pipeline**.



A **data pipeline** consists of 3 steps:

1. Preparing to run a model.
   Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping

2. Running the model

3. Communicating the results

A "good" Data Scientist will always go back and forth between the steps. The diagram on the left shows exactly what can happen.

In this course, you will develop the following skills:

data muning/scraping/sampling/cleaning in order to get an informative, manageable data setlength

data storage and management in order to be able to access data quickly and reliably during subsequent analysis

exploratory data analysis to generate hypotheses and intuition about the data

prediction based on statistical tools such as regression, classification, and clustering

communication of results through visualization, stories and interpretable summaries

2 Basic concepts

A data science student is attended to understand the **Grammar of Data Science**. Having some backgrounds in SQL concepts is also a good thing because, as it is very common, people loves to make example with it. Here is a brief refresh of some definitions and concepts about data science.

- **Data model** is a collection of concepts for describing data.
  - **Relational model** is one of the most common (SQL) and can handle most of the data. A counter exemple is facebook-like data, which requires **graph model**
- **Schema** is a description of a particular collection of data, using a given data model.
  - (Relational model) **Cardinality** is the #rows (number of items)
  - (Relational model) **Degree** or **Arity** is the #columns (number of attributes)
- A **JOIN** is a mean to combine tables based on shared attributes (most of the time some **IDs**). Despite its apparent simplicity beware of the many ways to compute a JOIN and check what is the default JOIN of a language before using it. The FIG 3 summurizes these possibilities.
- **Aggregation** or **reduction** is the action of reducing data with a common opperation (**sum**, **count**, **average**, ...) to summurize them.
  - (SQL command) GroupBy
- (Panda) **Series** is a named and ordered dictionary
  - keys are indexes
  - built on **numpy.ndarray** (so values can be any Numpy data type)
- (Panda) **DataFrame** is a table with named column
  - the columns are series
  - it is indeed a dictionary with (columnName → series)

2.1 Panda vs SQL

Panda is built to allow easy and fast **data exploration** and not to be a database manager, as SQL is. Thus there are benefits and drawbacks of using it.

Pros	**Cons**
Lightweight & fast	Tables stored directly in memory
Great expressivness (combine SQL + Python)	No post-load indexing functionality
Easy plot for data visualization (eg Matplotlib)	No transactions, journalings
	Large, complex joins are slower

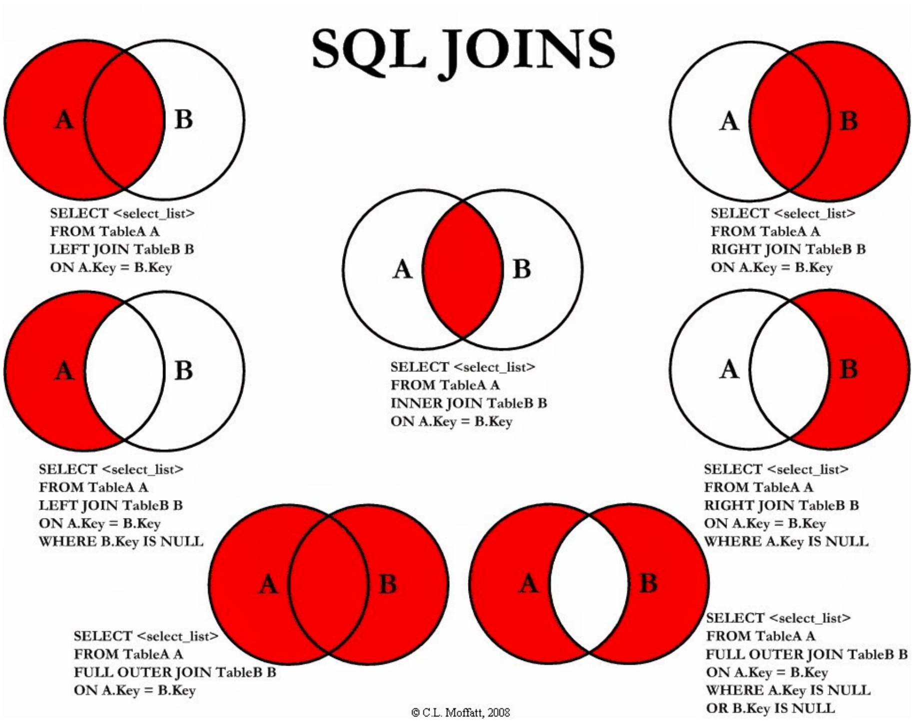


Figure 3: Different ways to join two tables and the related SQL command.

2.2 OnLine Analytical Processing (OLAP cubes)

OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives. Conceptually, it is like an n-dimensional spreadsheet (a cube) on which we can apply various opperations to take decisions.
OLAP cubes are an other way to see data table and are contructed based on them, as shows FIG 4.
Operations on OLAP cubes are the following and are illustrated on FIG 5

- **Slicing** fixes one or more variable

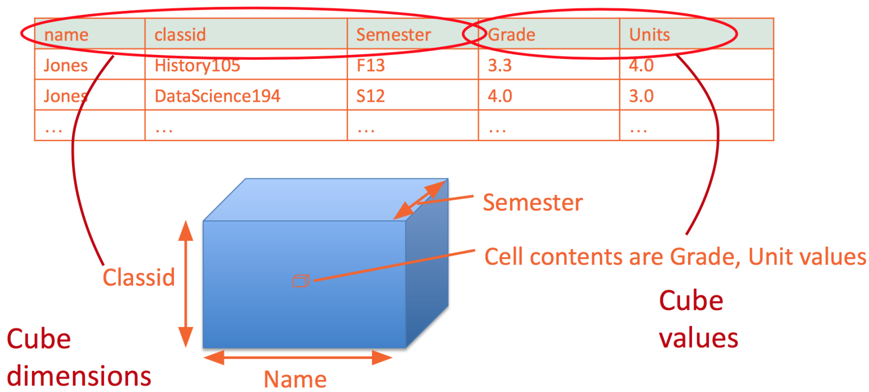
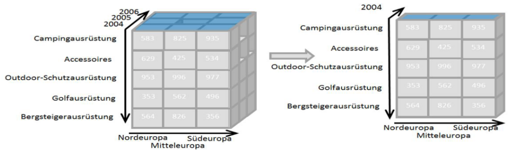
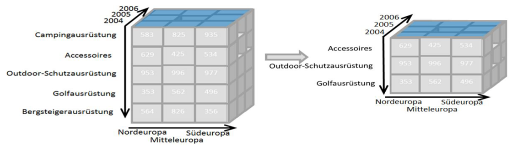


Figure 4: Construction of an OLAP cube from a table.

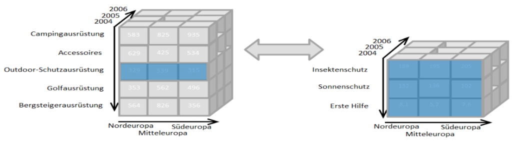
- **Dicing** selects a range of one or more variable
- **Driling up/down** changes levels of a hierarchically-indexed variable, ie "zoom" on a variable and see the sub-categories it contains.
- **Pivoting** change the point of view of the cube. Swap an aggregated variable an a detailed one.



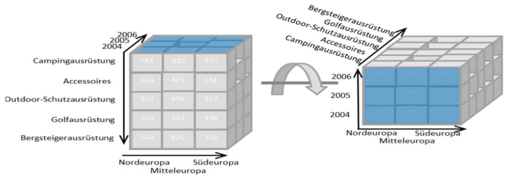
(a) Slincing



(b) Dicing



(c) Driling up/down



(d) Pivoting

Figure 5: Operations on OLAP cubes

Pros	Cons
The main adventage of OLAP cubes is that their are **conceptualy simpler** to understand by a non-scientist person, eg a business man who have to take day-to-day decisions based on company's data. Aggregations are limited but cover the main common cases that we can encounter.	Because of the "on-line" behaviour of this approach, all type of aggregation must be pre-calculated amoung all combination of axis which is very **expensive in memory and in time** (when updating the data)

3 Data Wrangling

Before any analysis, data need to be transformed from "dirty" to clean and processable data.
Data comes from different sources (excel or SQL?), sometime collected through different methods over time, with different conventions (space or NaN?), etc ... Data wrangling's goal is to **extract and standardize these raw data**. The best way to do it is to **combine automation with visualizations** in order to find outliers.
Data's problem can come from (non-exhaustive):

- Missing data
- Incorrect data

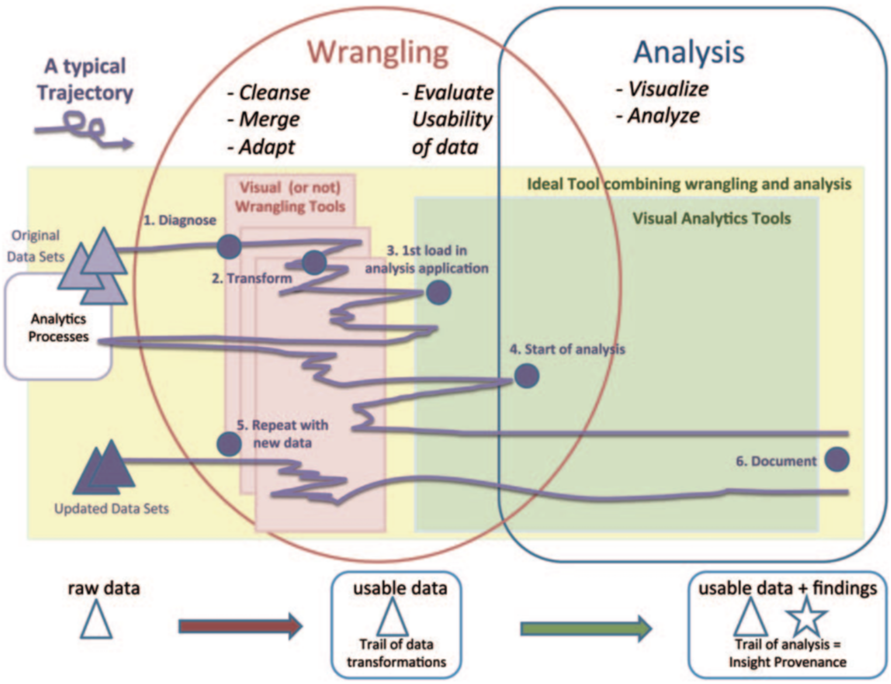


Figure 6: Things do not always happen as expected...

- Inconsistent representations of the same data
- Non-standardized data (centimeter or inches? farenheit or celsuis ?)
- Duplicated data

About 75% of theses problem will need **human intervention** to be corrected (by the data-scientist or by crowdsourcing).
Even if it seems really dirty, **beware not to over-sanitize the data!**. Applying what we can call "defensive programming" is not a good idea because we risk to lose any interesting data, keeping only the ones that fit perfectly in our model.

3.1 Dealing with missing values

Values can often miss from the data we have, because of various events (war, fire, ...). We must detect and correct these values with different method according with the domain we are working in.
Whatever the method used, it's good to keep track of these changes to know which are original data and which are modified ones.

- Set values to zero FIG 7a
- Interpolate based on existing data FIG 7b
- Omit missing data FIG 7c
- Interpolation with track kept 7d

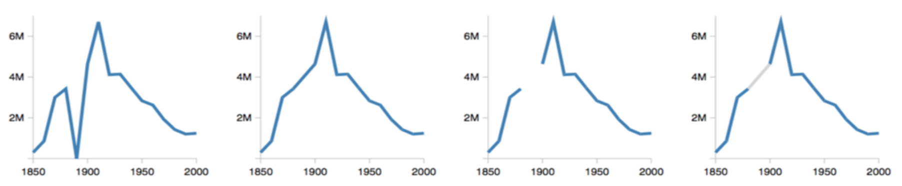


Figure 7: To deal with missing values.

3.2 General procedure

Once the data are well wrangled and before trying to analyse them we must take care of two more steps:

1. **Deal with uncertain data** (can arise from measurement errors, wrong sampling strategies, etc.)

2. **Parse/trasform data** (with aggregation and reduction techniques) to obtain meaningful records

It's always ideal to have the code and/or the documentation about the dataset you are analyzing (provenance)