The University Of Sheffield.

Data Preparation: sources of data, cleaning up your data and preparing data structure

Marta Milo
University of Sheffield
Department of Biomedical Science

September 2019

Outline

- Sources of Data
- Cleaning your data
  - Missing data points
  - Removing ambiguous information
  - Imputation of missing data points
- Preparing your data for processing
  - Data matrix and tables
- Build your workflow

Where is the data coming from?

Data is coming from complex experimental procedures that are subject to random mistakes

They describe complex system often of many layers

The majority of quantitative data comes from Next Generation sequencing, both at bulk and single-cell level.

Life Science data is also image-rich, from microscopy or 3D assays

In clinical studies we have patients data that are a mix of quantitative data and qualitative data

High throughput screening of drug compounds

And more….

Sources of data: NGS data

In NGS data we have random base incorporation that are generated by the protocols for library preparation

For fluorescent assays there can be an artifact of the dye incorporation

NGS data particularly at single-cell level is not synchronised. It is a snap shot of a system that has different components.

The NGS are particularly sensitive and capture large amount of information, including noise

Given the complexity of the systems under study and their high sensitiveness, the data contain a level of complexity that cannot be fully explained but only approximated. This is where we approach analysis using Data Science methods.

Sources of variation in NGS data

Sampling variance: sequencing produces millions of reads, but these represent only a small fraction of the cDNA is actually present in the library. There is therefore a sampling variance in each experiments.

Technical variance: Library preparation and sequencing procedures involve a series of complex biochemistry that contributes to between sample variance.

Biological variance:. Even in the absence of sampling and technical variance, biological variance will always exist. We will always quantify an "approximate" picture of the biological process and is therefore important design experiment with this in mind.

Importance of having an experimental design that reflects the nature of the system and the data

Cleaning your data

Remove “noise”

- Remove all formatting that interfere with the data we want to input
- Cleaning from noise and technical outliers.
- Remove any ambiguous data point

Handling missing data points

Data that is incomplete: no data or no annotation

Ignoring and removing it is not a way to handle it

If you we a statistical software the choice is made for you: you might want to control it.

Cleaning your data (cont…)

In statistical approach a way of handling missing data is via **imputation:** *replacing the missing value with an estimate*

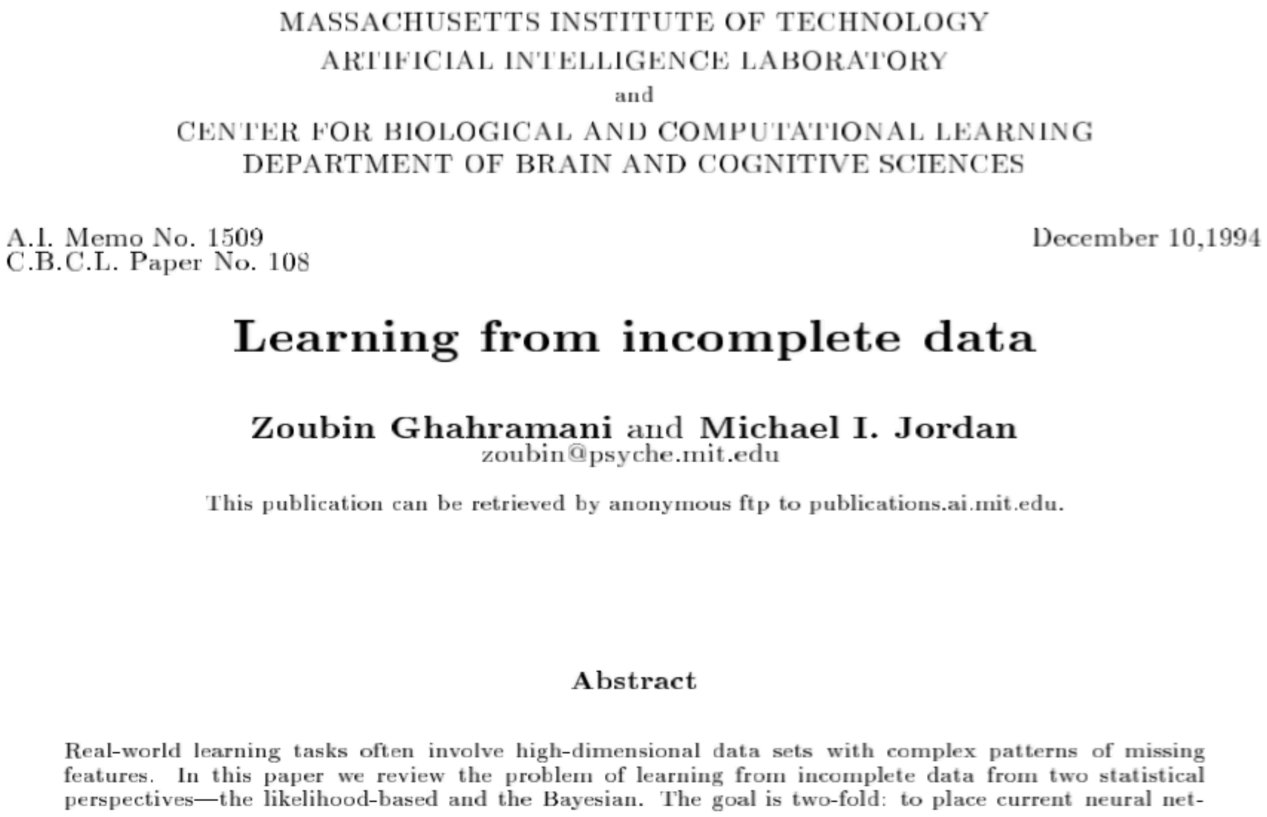
Mean imputation: the mean of all observed values. Not the best approach

Substitution
Impute the value from a new data point

Hot/Cold deck imputation
A randomly/systematically chosen value from an individual in the sample who has similar values on other variables

Regression imputation
Estimate the missing value by regressing on other variables

Cleaning your data (cont…)



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1509
C.B.C.L. Paper No. 108

December 10,1994

Learning from incomplete data

Zoubin Ghahramani and **Michael I. Jordan**
zoubin@psyche.mit.edu

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

Real-world learning tasks often involve high-dimensional data sets with complex patterns of missing features. In this paper we review the problem of learning from incomplete data from two statistical perspectives—the likelihood-based and the Bayesian. The goal is two-fold: to place current neural net-

Machine learning handle the incomplete data by **learning its value** from the observed data. Associated uncertainty

This happens in many different applications like classification, decision trees

Preparing your data for processing

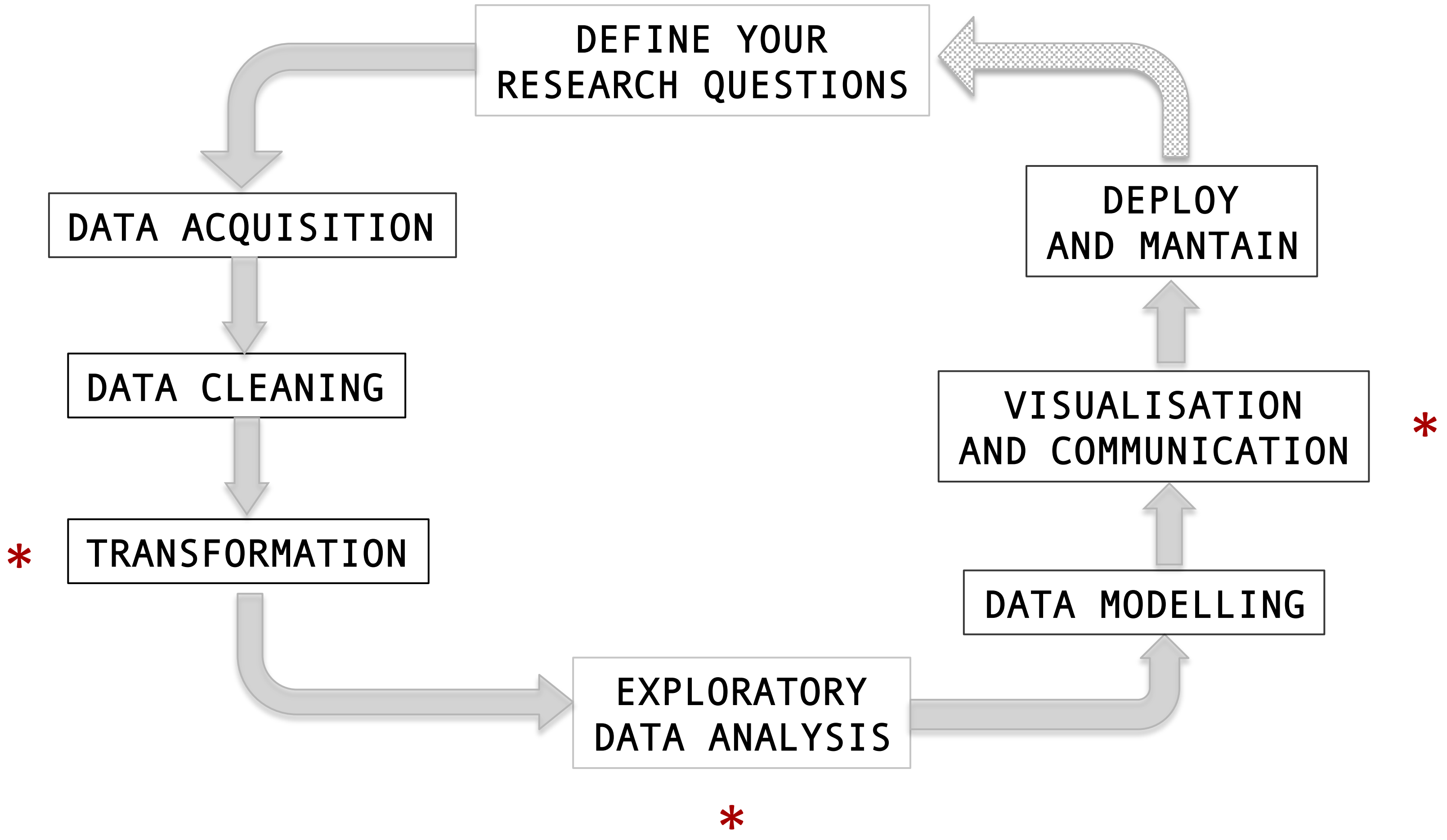
In most cases we need the data in a tabular format: data matrix

Decide the ***factors*** we want to analyse and their relations: often this is already established by experimental design

Transforming the data from their original values to standard values that will input the model for the processing depends on the model choice.

Establishing the data matrix is setting the variables of your model and their relationships

Establish your workflow



DEFINE YOUR RESEARCH QUESTIONS

DATA ACQUISITION

DATA CLEANING

* TRANSFORMATION

EXPLORATORY DATA ANALYSIS

*

DATA MODELLING

VISUALISATION AND COMMUNICATION *

DEPLOY AND MANTAIN

* = check points

Autumn School in Data Science, Cambridge 2019