DMI

Introduction to Data Mining

Data Mining and Data integration in Biomedicine
Master in Bioinformatics

Janet Piñero
Medbioinformatics Solutions SL
2025-2026

Outline

Brief description of the course

What is Data Mining?

Data

Common Data Mining Tasks

Best Practices in Data science projects

Hands-on sessions

Bibliography

Lectures / Hands-on sessions

1. Introduction to data mining and logistics
2. Exploratory Data Analysis (EDA)
3. Data Visualization
4. Unsupervised Learning (I)
5. Unsupervised Learning (II)
6. Supervised Learning (I)
7. Supervised Learning (II)
8. **Introduction to Agentic AI, Francesco Ronzano**
9. **Core capabilities of AI agents, Francesco Ronzano**
10. **Python-lab: building an AI agent, Francesco Ronzano**
11. Students Presentations (1) (16/02)
12. Students Presentations (2) (18/02)
13. **General Discussion**
14. **Invited speaker: Ferran Muiños, Nonparametric regression models and linear & mixed-effects models**

Evaluation

Work in pairs!

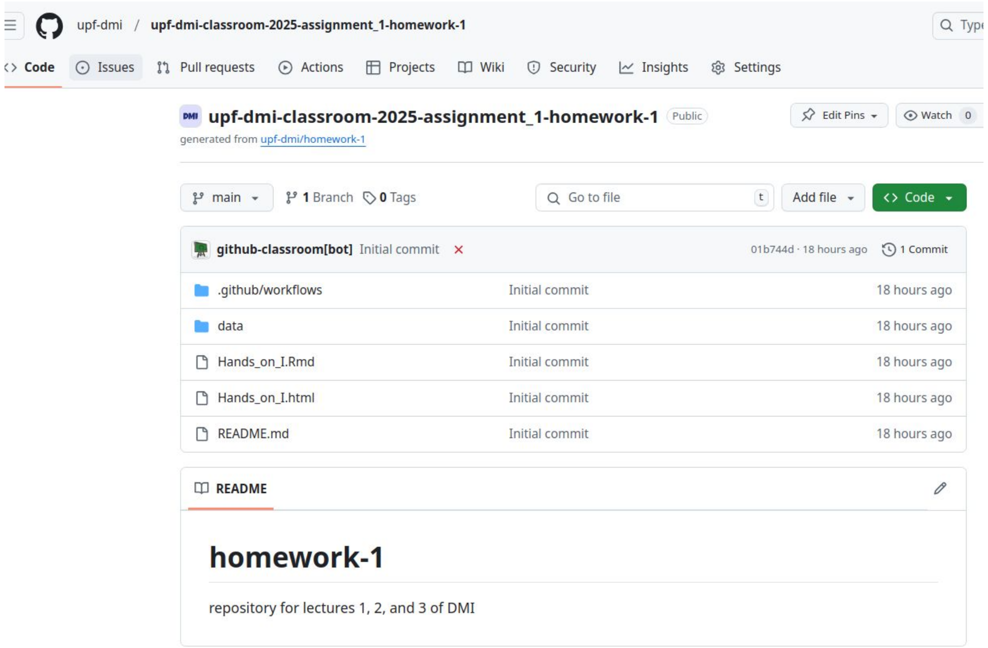
Exercises performed during the course -> 80 % (GitHub Classroom)

- Hands-on session 1: Lectures 1, 2 and 3 (Intro, EDA, Visualization) -> 25%
- Hands-on session 2: Lectures 4 and 5 (Unsupervised Learning) -> 25%
- Hands-on session 3: Lectures 6 and 7 (Supervised Learning) -> 25%

10 minute presentation of a scientific article -> 25 %

+5 minutes for questions

About the Hands-on sessions



upf-dmi-classroom-2025-assignment_1-homework-1

generated from upf-dmi/homework-1

github-classroom[bot] Initial commit ✕ 01b744d · 18 hours ago 1 Commit

.github/workflows	Initial commit	18 hours ago
data	Initial commit	18 hours ago
Hands_on_I.Rmd	Initial commit	18 hours ago
Hands_on_I.html	Initial commit	18 hours ago
README.md	Initial commit	18 hours ago

homework-1

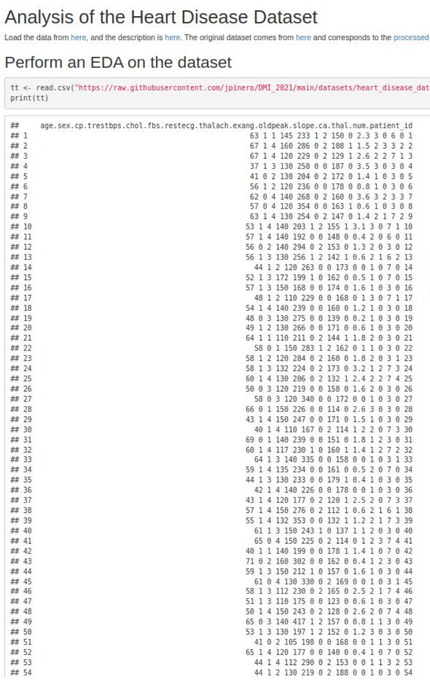
repository for lectures 1, 2, and 3 of DMI

Evaluation rubric for the hands-on sessions

DMI

- ✓ Does the GitHub repository contain all the necessary files (data and code)?
- ✓ Is the GitHub repository well-organized?
- ✓ **Does the R Markdown file (example, index.Rmd) and any additional code run the analysis without errors and generate the expected output (e.g., index.html)?**
- ✓ Does the analysis described in the resulting index.html file conform to the requested sectioning?
- ✓ Does the introduction clearly explain the question being addressed, the data used, and the number of observations and variables involved?
- ✓ Do the plots provide meaningful summaries of the data? Are the axes in the plots labeled in plain language and sized appropriately for readability?
- ✓ Does the R Markdown file (index.Rmd) and any other additional R Markdown files include R code interspersed with explanatory Markdown text? Is the code easy to read and understand?
- ✓ **Is the report clean and organized?**

Evaluation rubric for the hands-on sessions



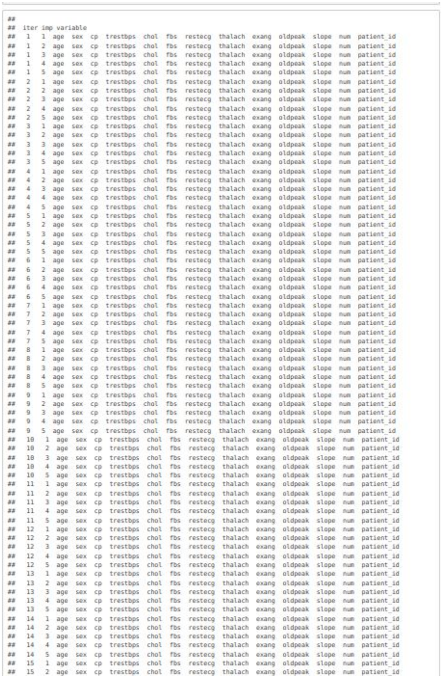
Analysis of the Heart Disease Dataset

Load the data from here, and the description is here. The original dataset comes from here and corresponds to the processed cleveland data

Perform an EDA on the dataset

```
tt <- read.csv("https://raw.githubusercontent.com/jpinero/DMI_2021/main/datasets/heart_disease_dataset.csv")
print(tt)
```

```
##     age.sex.cp.trestbps.chol.fbs.restecg.thalach.exang.oldpeak.slope.ca.thal.num.patient_id
## 1                                                    63 1 1 145 233 1 2 150 0 2.3 3 0 6 0 1
## 2                                                    67 1 4 160 286 0 2 108 1 1.5 2 3 3 2 2
## 3                                                    67 1 4 120 229 0 2 129 1 2.6 2 2 7 1 3
## 4                                                    37 1 3 130 250 0 0 187 0 3.5 3 0 3 0 4
## 5                                                    41 0 2 130 204 0 2 172 0 1.4 1 0 3 0 5
## 6                                                    56 1 2 120 236 0 0 178 0 0.8 1 0 3 0 6
## 7                                                    62 0 4 140 268 0 2 160 0 3.6 3 2 3 3 7
## 8                                                    57 0 4 120 354 0 0 163 1 0.6 1 0 3 0 8
## 9                                                    63 1 4 130 254 0 2 147 0 1.4 2 1 7 2 9
## 10                                                  53 1 4 140 203 1 2 155 1 3.1 3 0 7 1 10
## 11                                                  57 1 4 140 192 0 0 148 0 0.4 2 0 6 0 11
## 12                                                  56 0 2 140 294 0 2 153 0 1.3 2 0 3 0 12
## 13                                                  56 1 3 130 256 1 2 142 1 0.6 2 1 6 2 13
## 14                                                    44 1 2 120 263 0 0 173 0 0 1 0 7 0 14
## 15                                                  52 1 3 172 199 1 0 162 0 0.5 1 0 7 0 15
## 16                                                  57 1 3 150 168 0 0 174 0 1.6 1 0 3 0 16
## 17                                                    48 1 2 110 229 0 0 168 0 1 3 0 7 1 17
## 18                                                  54 1 4 140 239 0 0 160 0 1.2 1 0 3 0 18
## 19                                                  48 0 3 130 275 0 0 139 0 0.2 1 0 3 0 19
## 20                                                  49 1 2 130 266 0 0 171 0 0.6 1 0 3 0 20
## 21                                                  64 1 1 110 211 0 2 144 1 1.8 2 0 3 0 21
## 22                                                    58 0 1 150 283 1 2 162 0 1 1 0 3 0 22
## 23                                                  58 1 2 120 284 0 2 160 0 1.8 2 0 3 1 23
## 24                                                  58 1 3 132 224 0 2 173 0 3.2 1 2 7 3 24
## 25                                                  60 1 4 130 206 0 2 132 1 2.4 2 2 7 4 25
## 26                                                  50 0 3 120 219 0 0 158 0 1.6 2 0 3 0 26
## 27                                                    58 0 3 120 340 0 0 172 0 0 1 0 3 0 27
## 28                                                  66 0 1 150 226 0 0 114 0 2.6 3 0 3 0 28
## 29                                                  43 1 4 150 247 0 0 171 0 1.5 1 0 3 0 29
## 30                                                    40 1 4 110 167 0 2 114 1 2 2 0 7 3 30
## 31                                                  69 0 1 140 239 0 0 151 0 1.8 1 2 3 0 31
## 32                                                  60 1 4 117 230 1 0 160 1 1.4 1 2 7 2 32
## 33                                                    64 1 3 140 335 0 0 158 0 0 1 0 3 1 33
## 34                                                  59 1 4 135 234 0 0 161 0 0.5 2 0 7 0 34
## 35                                                  44 1 3 130 233 0 0 179 1 0.4 1 0 3 0 35
## 36                                                    42 1 4 140 226 0 0 178 0 0 1 0 3 0 36
## 37                                                  43 1 4 120 177 0 2 120 1 2.5 2 0 7 3 37
## 38                                                  57 1 4 150 276 0 2 112 1 0.6 2 1 6 1 38
## 39                                                  55 1 4 132 353 0 0 132 1 1.2 2 1 7 3 39
## 40                                                    61 1 3 150 243 1 0 137 1 1 2 0 3 0 40
## 41                                                    65 0 4 150 225 0 2 114 0 1 2 3 7 4 41
## 42                                                  40 1 1 140 199 0 0 178 1 1.4 1 0 7 0 42
## 43                                                  71 0 2 160 302 0 0 162 0 0.4 1 2 3 0 43
## 44                                                  59 1 3 150 212 1 0 157 0 1.6 1 0 3 0 44
## 45                                                    61 0 4 130 330 0 2 169 0 0 1 0 3 1 45
## 46                                                  58 1 3 112 230 0 2 165 0 2.5 2 1 7 4 46
## 47                                                  51 1 3 110 175 0 0 123 0 0.6 1 0 3 0 47
## 48                                                  50 1 4 150 243 0 2 128 0 2.6 2 0 7 4 48
## 49                                                  65 0 3 140 417 1 2 157 0 0.8 1 1 3 0 49
## 50                                                  53 1 3 130 197 1 2 152 0 1.2 3 0 3 0 50
## 51                                                    41 0 2 105 198 0 0 168 0 0 1 1 3 0 51
## 52                                                  65 1 4 120 177 0 0 140 0 0.4 1 0 7 0 52
## 53                                                    44 1 4 112 290 0 2 153 0 0 1 1 3 2 53
## 54                                                    44 1 2 130 219 0 2 188 0 0 1 0 3 0 54
```



```
##
##  iter imp variable
##   1   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   1   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   1   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   1   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   1   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   2   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   2   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   2   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   2   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   2   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   3   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   3   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   3   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   3   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   3   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   4   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   4   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   4   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   4   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   4   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   5   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   5   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   5   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   5   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   5   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   6   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   6   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   6   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   6   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   6   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   7   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   7   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   7   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   7   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   7   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   8   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   8   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   8   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   8   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   8   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   9   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   9   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   9   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   9   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   9   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   10   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   10   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   10   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   10   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   10   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   11   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   11   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   11   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   11   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   11   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   12   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   12   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   12   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   12   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   12   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   13   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   13   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   13   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   13   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   13   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   14   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   14   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   14   3  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   14   4  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   14   5  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   15   1  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
##   15   2  age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  num  patient_id
```

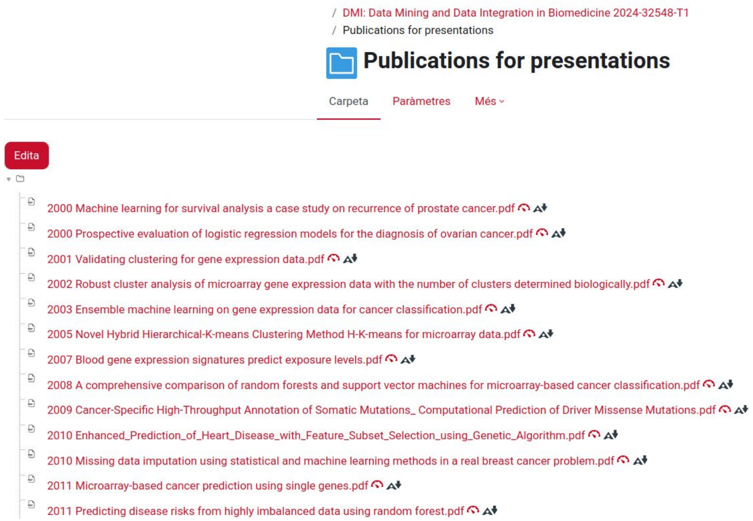
Presentation of a scientific article

Select wisely!

DMI

10 minute presentation of a scientific article -> 20 %

+5 minutes for questions



/ DMI: Data Mining and Data Integration in Biomedicine 2024-32548-T1
/ Publications for presentations

Publications for presentations

- 2000 Machine learning for survival analysis a case study on recurrence of prostate cancer.pdf
- 2000 Prospective evaluation of logistic regression models for the diagnosis of ovarian cancer.pdf
- 2001 Validating clustering for gene expression data.pdf
- 2002 Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically.pdf
- 2003 Ensemble machine learning on gene expression data for cancer classification.pdf
- 2005 Novel Hybrid Hierarchical-K-means Clustering Method H-K-means for microarray data.pdf
- 2007 Blood gene expression signatures predict exposure levels.pdf
- 2008 A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.pdf
- 2009 Cancer-Specific High-Throughput Annotation of Somatic Mutations_ Computational Prediction of Driver Missense Mutations.pdf
- 2010 Enhanced_Prediction_of_Heart_Disease_with_Feature_Subset_Selection_using_Genetic_Algorithm.pdf
- 2010 Missing data imputation using statistical and machine learning methods in a real breast cancer problem.pdf
- 2011 Microarray-based cancer prediction using single genes.pdf
- 2011 Predicting disease risks from highly imbalanced data using random forest.pdf

Presentation of a scientific article

Select wisely!

Guide for the presentations

- What is the objective of the article?
- What programming languages are used?
- What is the quality and size of the data? Is it publicly available?
- What preprocessing steps are applied to the data? How are they reported in the article?
- Regarding the algorithms employed: How are they described? Are the parameters clearly reported?
- For machine learning models, is the training process clearly described? How is the data split?
- How is the performance of the algorithm evaluated?
- Provide a critical assessment of the figures and tables. Are they clear, accurate, and well-labeled?
- How is the supplementary material used? What information is included in the supplementary material?
- Is the code available? Is the analysis reproducible based on the provided information?
- Are there any limitations or biases in the study? How are they addressed?

Presentation of a scientific article

The chosen scientific article should be sent via email (janet.pinero@upf.edu) no later than **February 01**, along with the names of the students. The list of the publications that have been already selected will be updated here here.

The presentation should be sent by email no later than February 15 (janet.pinero@upf.edu)

What is Data Mining?

Why do we need to analyze data?

- To understand what has happened or what is happening;
- To predict what is likely to happen, either in the future or in other circumstances we haven’t seen yet;
- To guide us in making decisions.

Drivers of Data Mining

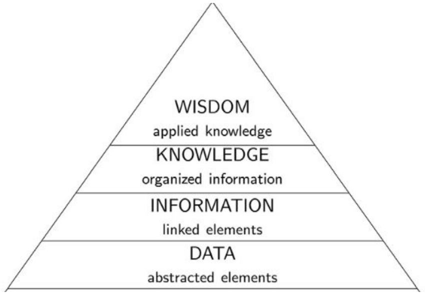
- The Explosive Growth of Data: from terabytes to petabytes
- Easy data collection and increase of data availability
- Automated data collection tools, database systems, Web, computerized society
- Advances in computing power

Abundant data, limited knowledge.

Data, information, and knowledge

Data – observations, signals, measurements: *Body temperature 103*
Information – a set of data in context, data with meaning: *The patient is having a fever*
Knowledge – justifiable beliefs based on data and information: i*f a patient temperature is > 100 F, he might a fever (or hyperthermia).*

Data – a patient's blood pressure is 100/50.
Information – that patient has a ten year history of blood pressures of 150/100.
Knowledge – the patient has a known history of coronary artery disease and is now experiencing chest pain. The sudden drop in blood pressure could indicate a serious myocardial infarction in progress.



The DIKW pyramid

Wisdom is acting on knowledge in an appropriate way

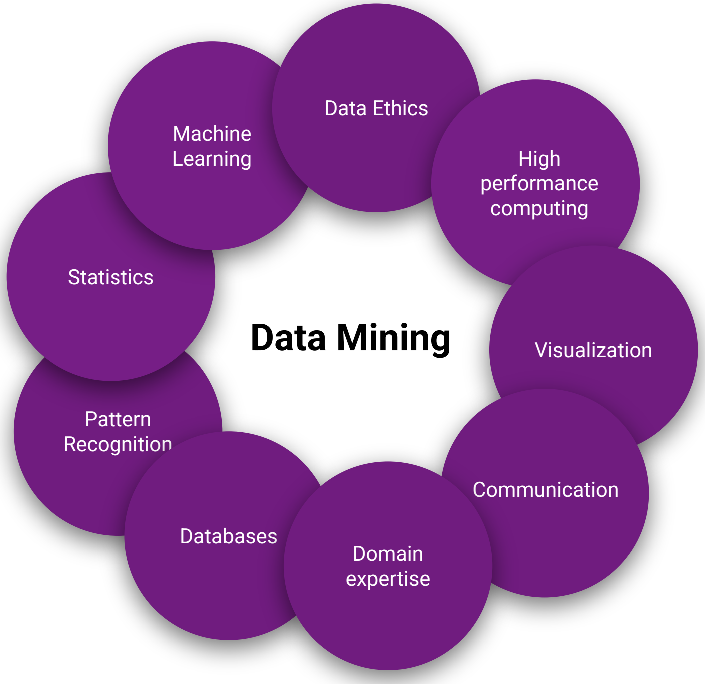
What is Data Mining?

- "**Data** mining is the science of extracting **useful knowledge** from huge **data** repositories"
- "**Data** Mining refers to a **set of methods** applicable to large and complex **databases** to eliminate the randomness and discover the hidden pattern. Data mining is about tools, methodologies, and theories for r**evealing patterns in data** — which is a critical step in **knowledge** discovery."
- "Data mining (aka knowledge discovery from data, or **data science**) consists on the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) **patterns or knowledge** from huge amount of **data**"
- Data mining is the **process** of deciphering **meaningful insights** from existing **datasets**
- "Finding hidden information in a **database**"

What is Data Mining?

- Many Definitions
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or **semi**-automatic means, of large quantities of data in order to discover meaningful patterns
- It’s a process, which means that it *can* be iterative

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

Tremendous amount of data

- ○ Algorithms must be *scalable* to handle big data

High-dimensionality of data

- ○ Microarray data may have tens of thousands of *dimensions*

High complexity and heterogeneity of data

- ○ Data streams and sensor data
- ○ Time-series data, temporal data, sequence data
- ○ Structure data, graphs, social and information networks
- ○ Spatial, spatiotemporal, multimedia, text and Web data
- ○ Software programs, scientific simulations

Need for new and sophisticated applications

Myths about Data Science

DMI

- Data mining is an automatic process
- Every data mining process needs big data and deep learning
- Modern data science tools are easy to use
- There is always a hidden gem in the data

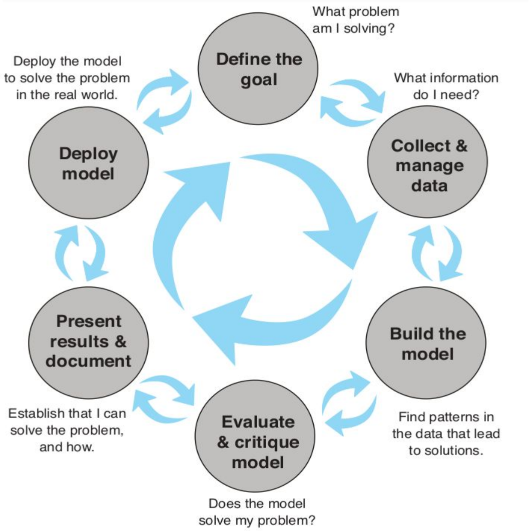
General steps of the data mining process

DMI

Most data science projects follow a common set of steps:

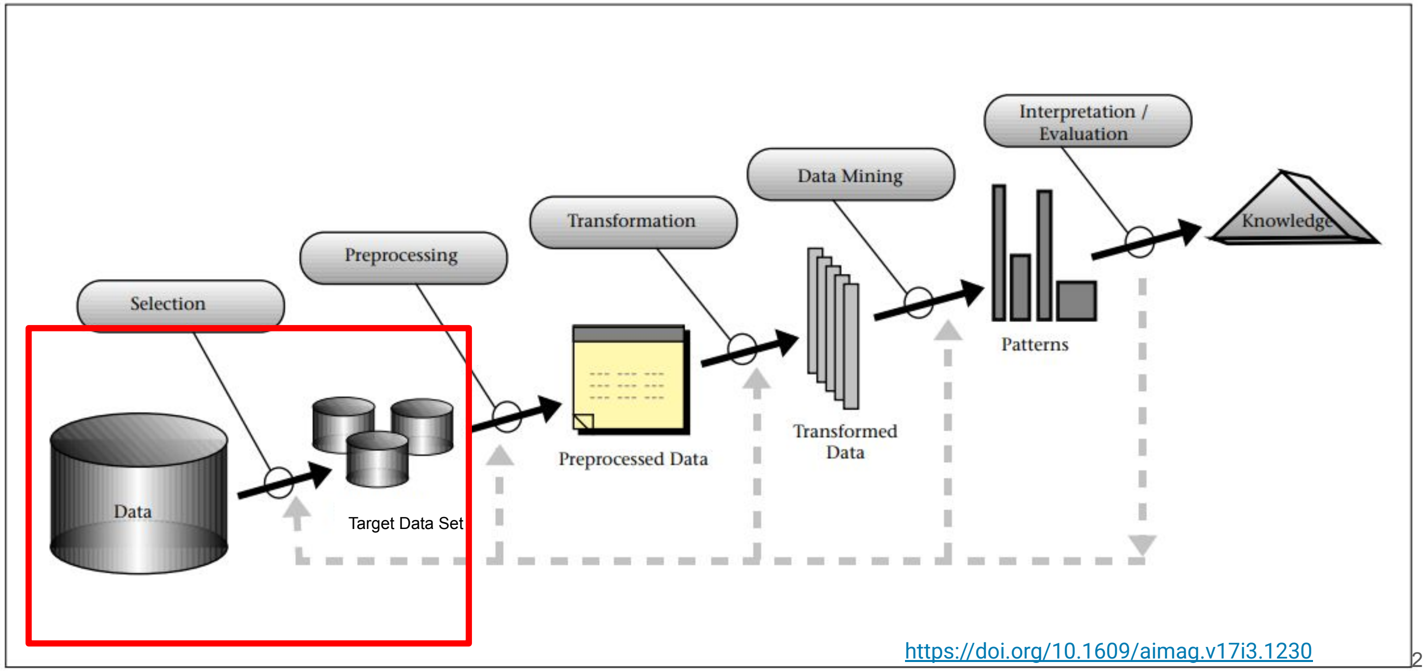
1. Define the question/goal 🤔
2. Data: Obtain the data 📊
3. **Exploratory data analysis** 😅
4. **Model**: Statistical prediction/modeling 😁
5. Interpret results 💡
6. Synthesize/write up results ✍️
7. Create reproducible code 😒

These steps might vary somewhat depending on the context of the project, but are remarkably consistent across projects in both industry and academia.



Define the goal
What problem am I solving?
What information do I need?
Collect & manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate & critique model
Does the model solve my problem?
Present results & document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

The Data Mining process



https://doi.org/10.1609/aimag.v17i3.1230

Data

“A map is not the territory it represents, but, if correct, it has a similar structure to the territory which accounts for its usefulness” Alfred Korzybski

Raw data

It is critical that you include the rawest form of the data that you have access to.

Some examples of the raw form of data are as follows.

1. The strange binary file your measurement machine spits out
2. The unformatted Excel file with 10 worksheets your colaborator sent you
3. The complicated JSON data you got from scraping the Twitter API
4. The hand-entered numbers someone collected looking through a microscope
5. The file that you downloaded from a particular resource
6. The supplementary material of a publication (if you are lucky!)

Data: Structured vs unstructured

Unstructured data sources

Structured data sources

Data: Structured vs unstructured

Unstructured data sources

- No logical data model, no particular logic and structure is defined to collect, store, and expose data

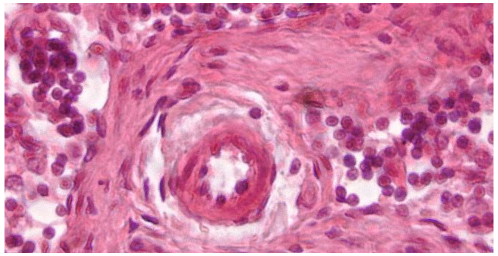
Structured data sources

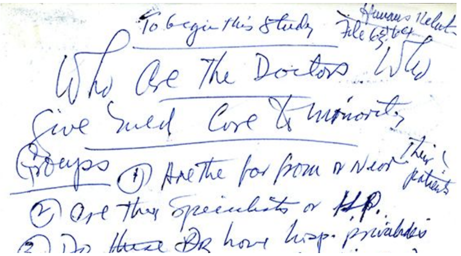
- High degree of organization. There is a specific data model

Data: Structured vs unstructured

Unstructured data sources

- No logical data model, no particular logic and structure is defined to collect, store, and present data:
  - Medical images: X-rays, MRIs, CT scans, pathology slides
  - Clinical notes: doctors’ narratives, discharge summaries
  - Biomedical signals: audio recordings of heart or lung sounds
  - Research articles: full-text biomedical papers





To begin this study

Humans Relat. File 63-64

Who are The Doctors Who Give Med Care to Minority Groups

(1) Are the far from or Near their patients

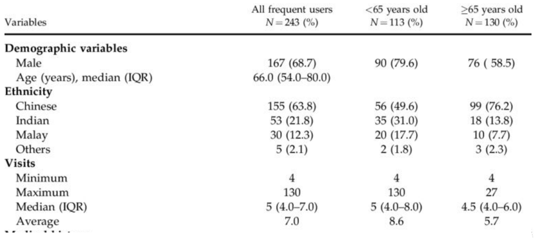
(2) Are they Specialists or H.P.

(3) Do ~~these~~ DR have hosp. privileges

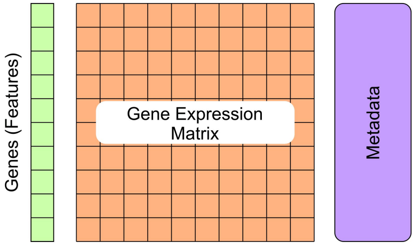
Data: Structured vs unstructured

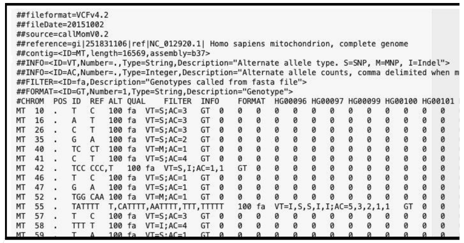
Structured data sources

- High degree of organization. There is a specific data model: **an R data frame is a typical example of structured data, where you can find columns and rows, and every column has a specific type of data among all records**



Variables	All frequent users N = 243 (%)	<65 years old N = 113 (%)	≥65 years old N = 130 (%)
Demographic variables			
Male	167 (68.7)	90 (79.6)	76 (58.5)
Age (years), median (IQR)	66.0 (54.0–80.0)		
Ethnicity			
Chinese	155 (63.8)	56 (49.6)	99 (76.2)
Indian	53 (21.8)	35 (31.0)	18 (13.8)
Malay	30 (12.3)	20 (17.7)	10 (7.7)
Others	5 (2.1)	2 (1.8)	3 (2.3)
Visits			
Minimum	4	4	4
Maximum	130	130	27
Median (IQR)	5 (4.0–7.0)	5 (4.0–8.0)	4.5 (4.0–6.0)
Average	7.0	8.6	5.7





```
##fileformat=VCFv4.2
##fileDate=20151002
##source=callMomV0.2
##reference=gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome
##contig=<ID=MT,length=16569,assembly=b37>
##INFO=<ID=VT,Number=.,Type=String,Description="Alternate allele type. S=SNP, M=MNP, I=Indel">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate allele counts, comma delimited when m
##FILTER=<ID=fa,Description="Genotypes called from fasta file">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO     FORMAT  HG00096 HG00097 HG00099 HG00100 HG00101
MT   10  .   T   C   100 fa  VT=S;AC=3   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   16  .   A   T   100 fa  VT=S;AC=3   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   26  .   C   T   100 fa  VT=S;AC=3   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   35  .   G   A   100 fa  VT=S;AC=2   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   40  .   TC  CT  100 fa  VT=M;AC=1   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   41  .   C   T   100 fa  VT=S;AC=4   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   42  .   TCC CCC,T   100 fa  VT=S,I;AC=1,1   GT  0   0   0   0   0   0   0   0   0   0   0   0
MT   46  .   T   C   100 fa  VT=S;AC=1   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   47  .   G   A   100 fa  VT=S;AC=1   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   52  .   TGG CAA 100 fa  VT=M;AC=1   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   55  .   TATTTT  T,CATTTT,AATTTT,TTT,TTTTT   100 fa  VT=I,S,S,I,I;AC=5,3,2,1,1   GT  0   0
MT   57  .   T   C   100 fa  VT=S;AC=3   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   58  .   TTT T   100 fa  VT=I;AC=4   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
MT   59  .   T   A   100 fa  VT=S;AC=1   GT  0   0   0   0   0   0   0   0   0   0   0   0   0
```

Data: Structured vs unstructured

Examples of structured and unstructured data in electronic health records (EHRs)

Time	dd/mm/yyyy 14 September 2017 Prescription start date: dd/mm/yy	‘Earlier today Mr X experienced chest pain’ ‘Operation scheduled on Tuesday’.

Data: Structured vs unstructured

Examples of structured and unstructured data in electronic health records (EHRs)

Diagnosis	The patient has diabetes without complications	C0271635

Data: Types of data sources

DMI

Examples of structured and unstructured data in electronic health records (EHRs)

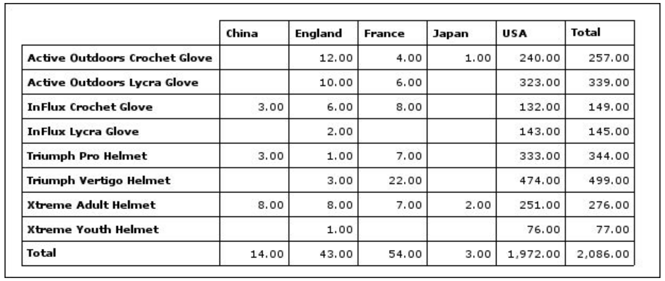
Data item	Structured	Unstructured
Time	dd/mm/yyyy 14 September 2017 Prescription start date: dd/mm/yy	‘Earlier today Mr X experienced chest pain’ ‘Operation scheduled on Tuesday’.
Symptoms	N242300 Neuropathic pain 1B1B.00 Cannot sleep—insomnia	‘…c/o shooting pain in upper right leg during the night, disturbing her sleep’
Diagnosis	C0271635 Type 2 diabetes without complication	‘The patient has diabetes without complications’
Prescription	01040200 (BNF code for codeine phosphate 60 mg tablets)	‘Px codeine 60 mg PO qid×7 days’
Referral	8H4D.00 Referral to psychogeriatrician	Rev 4 w ?refer pyscho ger
Test	43F1.00 Rheumatoid factor positive	Rheumatoid factor was 42 IU/mL which is a positive result

DMI

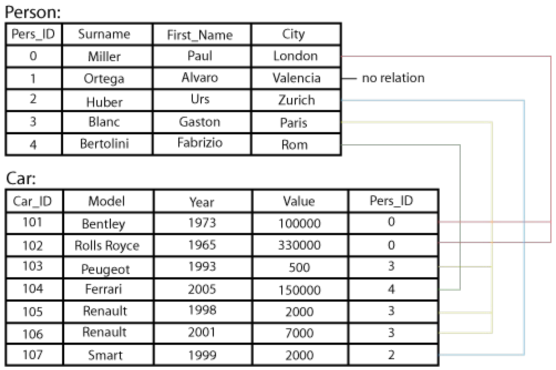
Types of Data Sets

Types of Data Sets: Record Data

- ❑ Relational records
  - ❑ Relational tables, highly structured
- ❑ Data matrix, e.g., numerical matrix



	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00



Person:

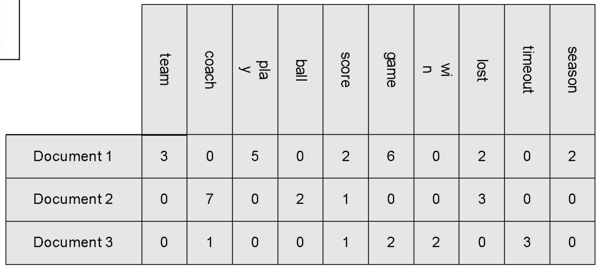
Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

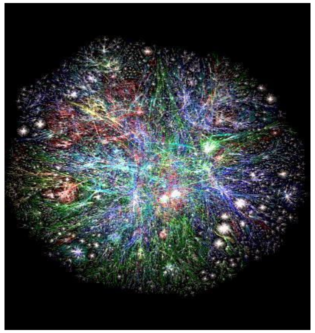
Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

- ❑ Document data: Term-frequency vector (matrix) of text documents

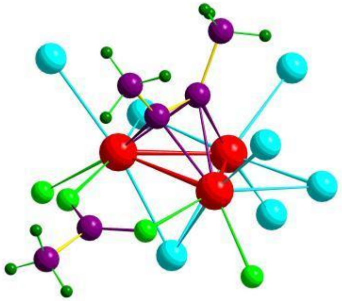


	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Types of Data Sets: Networks

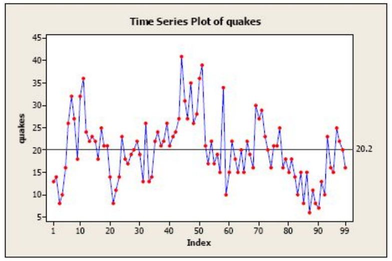


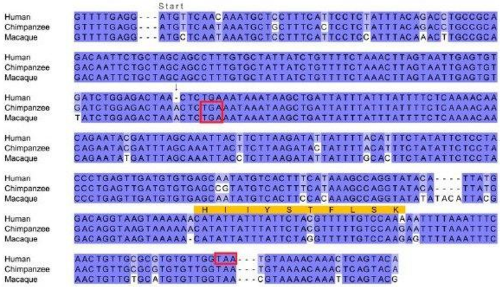
- ❑ Protein interaction networks
- ❑ Comorbidity networks
- ❑ Molecular Structures
- ❑ Social or information networks



Types of Data Sets: Ordered data

- ❑ Video data: sequence of images
- ❑ Temporal data: time-series
- ❑ Genetic sequence data




Start
Human GTTTTGAGG---ATGTTCAACAAATGCTCCTTTCATTCCTCTATTTACAGACCTGCCGCA
Chimpanzee GTTTTGAGG---ATGTTCAATAAATGCTGCTTTCACTCCTCTATTTACAGACCTGCCGCA
Macaque GTTTTGAGG---ATGCTCAATAAATGCTCCTTTCATTCCTCCATTTACAAACTTGCCGCA
Human GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Human GATCTGGAGACTAA-CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Chimpanzee GATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Macaque TATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Human CAGAATACGATTTAGCAAATTACTTCTTAAGATATTATTTTACATTTCTATATTCTCCTA
Chimpanzee CAGAATACGATTTAGCAAATTACTTCTTAAGATACTATTTTACATTTCTATATTCTCCTA
Macaque CAGAATATGATTTAGCAAATTACCTCTTAAGATATTATTTTGCACTTCTATATTCTCCTA
Human CCCTGAGTTGATGTGTGAGCAATATGTCACTTTCATAAAGCCAGGTATACA----TTATG
Chimpanzee CCCTGAGTTGATGTGTGAGCCGTATGTCACTTTCATAAAGCCAGGTATACA----TTATG
Macaque CCCTGAGTTGATGTGTGAGCAATATGTCACTTCCACAAAGCCAGGTATATATACATTACG
H I I Y S T F L S K
Human GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAAAATTTTAAATTTC
Chimpanzee GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAGAATTTTAAATTTC
Macaque GACAGGTAAGTAAAAA-CATATTATTTATTCTAGGTTTTTGTCCAAGAGTTTTAAATTTC
Human AACTGTTGCGCGTGTGTTGGTAA---TGTAAAACAAACTCAGTACA
Chimpanzee AACTGTTGCGCGTGTGTTGGTAA---TGTAAAACAAACTCAGTACA
Macaque AACTGTTGTGCATGTGTTGGTAA---CGTAAAACAAATTCAGTACG

Characteristics of Structured Data Sets

- ❑ Dimensionality: Number of attributes
- ❑ Sparsity: The number of empty cells
- ❑ Resolution: Patterns depend on the scale
- ❑ Distribution: Centrality and dispersion

Data Objects

- Data sets are made up of data objects
- A data object represents an entity
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, diseases, treatments
  - university database: students, professors, courses
- Also called **samples , examples, instances, data points, objects**

Data Attributes

- Data objects are described by attributes (also called features, dimensions)
- A data field, representing a characteristic or feature of a data object.
    - E.g., customer_ID, name, address, height, weight
- Correct representation of data:
    - Database rows ⟶ data objects
    - columns ⟶ attributes

Attribute types: scales of measurement

Qualitative (Categorical) Data

- Nominal:
    - Unordered set of categories identified only by name, no ranking, no magnitude
    - Hair color = {auburn, black, blond, brown, grey, red, white}, marital status, occupation, ID numbers, zip codes, blood type = {A, B, AB, O}
- Binary
    - A special case of nominal data with only two states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- Ordinal
    - Values have a meaningful order (ranking) but magnitude between successive values is not known
    - Size = {small, medium, large}, Stage of cancer (stage I, II, III, IV), hypertension categories = {mild, moderate, severe}

Attribute types: scales of measurement

Quantitative (Numeric) Data

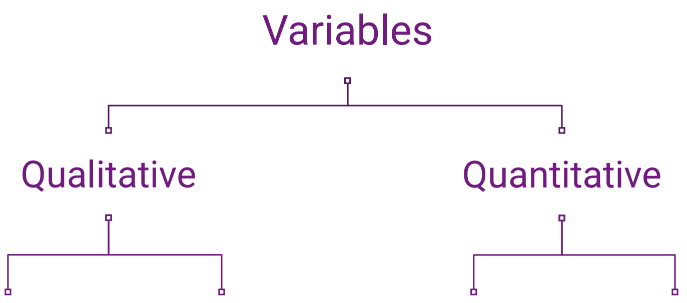
Interval

- Measured on a scale of **equal-sized units**
- Values have order
  - E.g., *temperature in C˚or F˚, calendar dates, IQ scores, z-scores*
- No true zero-point (zero doesn't mean the absence of value)

Ratio

- Inherent **zero-point** (represents the complete absence of any amount of the variable.)
- We can speak of values as being an order of magnitude larger than the unit of measurement
  - E.g., *length, counts, heart rate, age, weight, blood concentration measurements, dosage of drugs*

Summary



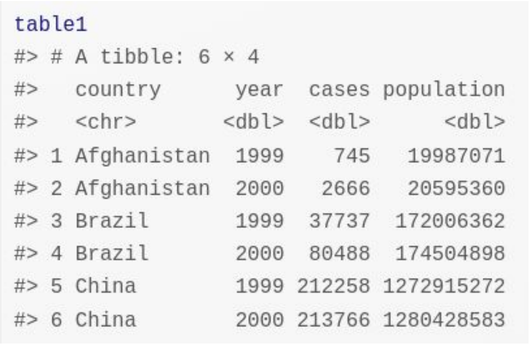
Variables

Qualitative

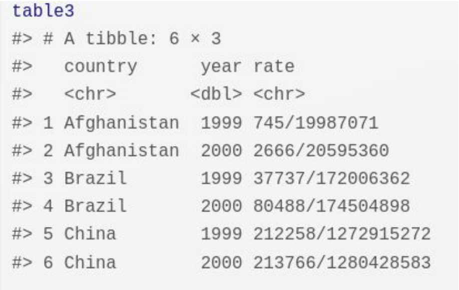
Quantitative

Categorical Data	Nominal	Ordinal	Interval	Ratio Discrete or continuous
Frequency distribution	✓	✓	✓	✓
Median and percentiles	✗	✓	✓	✓
Add or subtract	✗	✗	✓	✓
Mean, standard deviation	✗	✗	✓	✓
Ratios, coefficient of variation	✗	✗	✗	✓

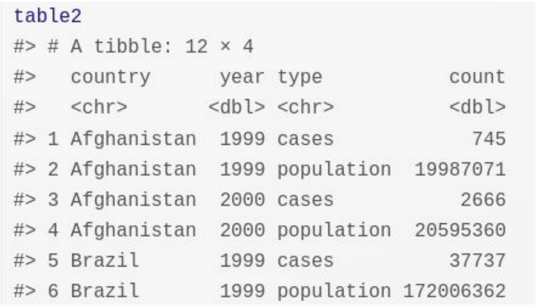
Tidy data



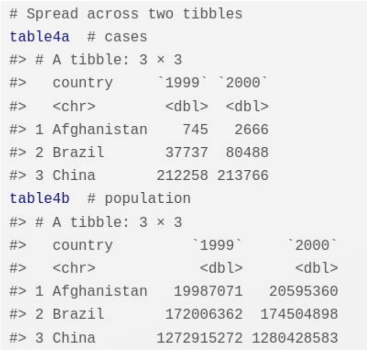
```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```



```
table3
#> # A tibble: 6 × 3
#>   country      year rate
#>   <chr>       <dbl> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```



```
table2
#> # A tibble: 12 × 4
#>   country      year type            count
#>   <chr>       <dbl> <chr>           <dbl>
#> 1 Afghanistan  1999 cases             745
#> 2 Afghanistan  1999 population   19987071
#> 3 Afghanistan  2000 cases            2666
#> 4 Afghanistan  2000 population   20595360
#> 5 Brazil       1999 cases           37737
#> 6 Brazil       1999 population 172006362
```

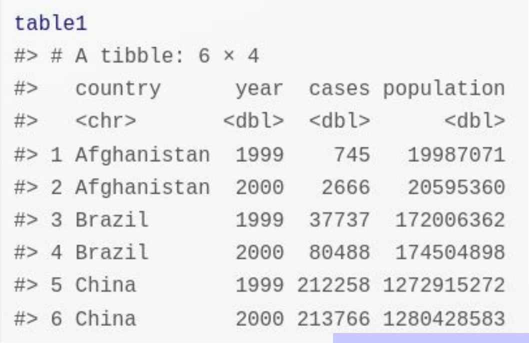


```
# Spread across two tibbles
table4a  # cases
#> # A tibble: 3 × 3
#>   country     `1999` `2000`
#>   <chr>        <dbl>  <dbl>
#> 1 Afghanistan    745   2666
#> 2 Brazil       37737  80488
#> 3 China       212258 213766
table4b  # population
#> # A tibble: 3 × 3
#>   country         `1999`     `2000`
#>   <chr>            <dbl>      <dbl>
#> 1 Afghanistan   19987071   20595360
#> 2 Brazil       172006362  174504898
#> 3 China       1272915272 1280428583
```

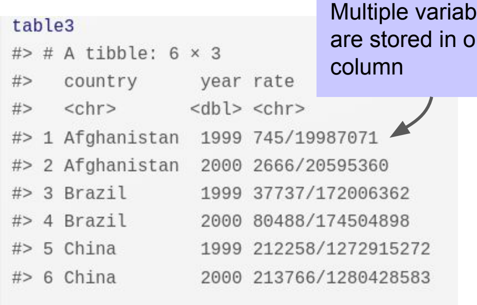
Each dataset shows the same values of four variables: country, year, population, and number of documented cases of TB (tuberculosis)

https://vita.had.co.nz/papers/tidy-data.pdf

Tidy data



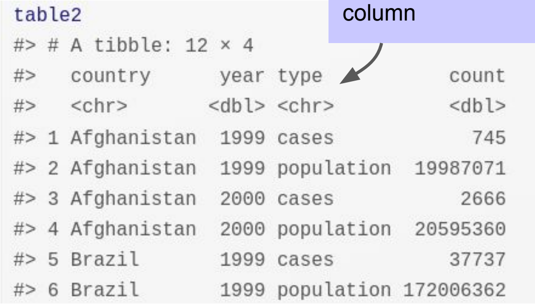
```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```



```
table3
#> # A tibble: 6 × 3
#>   country      year rate
#>   <chr>       <dbl> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

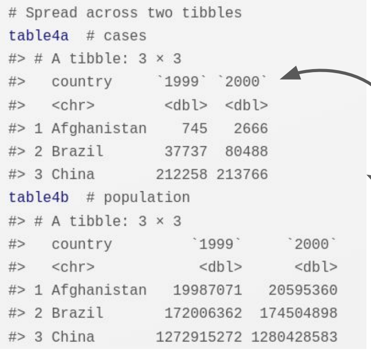
Multiple variables are stored in one column

Multiple variables are stored in one column



```
table2
#> # A tibble: 12 × 4
#>   country      year type            count
#>   <chr>       <dbl> <chr>           <dbl>
#> 1 Afghanistan  1999 cases             745
#> 2 Afghanistan  1999 population   19987071
#> 3 Afghanistan  2000 cases            2666
#> 4 Afghanistan  2000 population   20595360
#> 5 Brazil       1999 cases           37737
#> 6 Brazil       1999 population  172006362
```

“If you find yourself using a column to describe what another column means, your data probably isn’t tidy.”



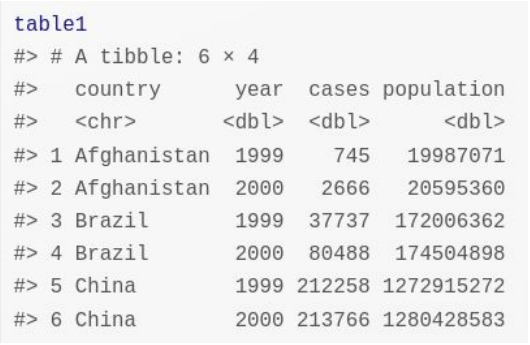
```
# Spread across two tibbles
table4a  # cases
#> # A tibble: 3 × 3
#>   country      `1999` `2000`
#>   <chr>         <dbl>  <dbl>
#> 1 Afghanistan     745   2666
#> 2 Brazil        37737  80488
#> 3 China        212258 213766
table4b  # population
#> # A tibble: 3 × 3
#>   country          `1999`     `2000`
#>   <chr>             <dbl>      <dbl>
#> 1 Afghanistan    19987071   20595360
#> 2 Brazil        172006362  174504898
#> 3 China        1272915272 1280428583
```

Column headers are values, not variable names

A single observational unit is stored in multiple tables

https://vita.had.co.nz/papers/tidy-data.pdf

Tidy data



```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

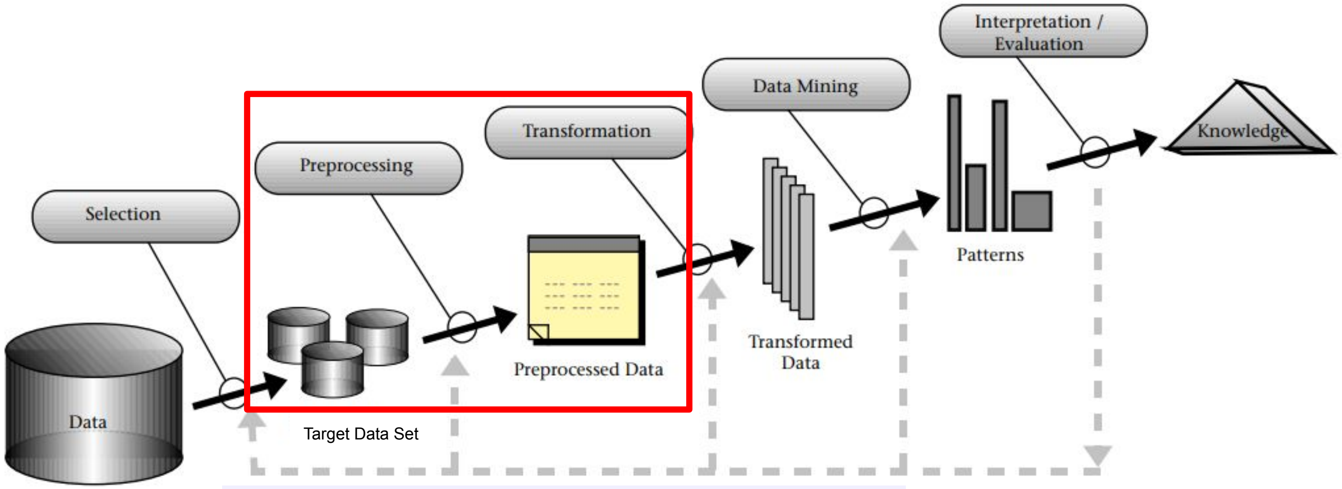
There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

“Happy families are all alike; every unhappy family is unhappy in its own way.” –– Leo Tolstoy
“Tidy datasets are all alike, but every messy dataset is messy in its own way.” –– Hadley Wickham

https://vita.had.co.nz/papers/tidy-data.pdf

The Data Mining process



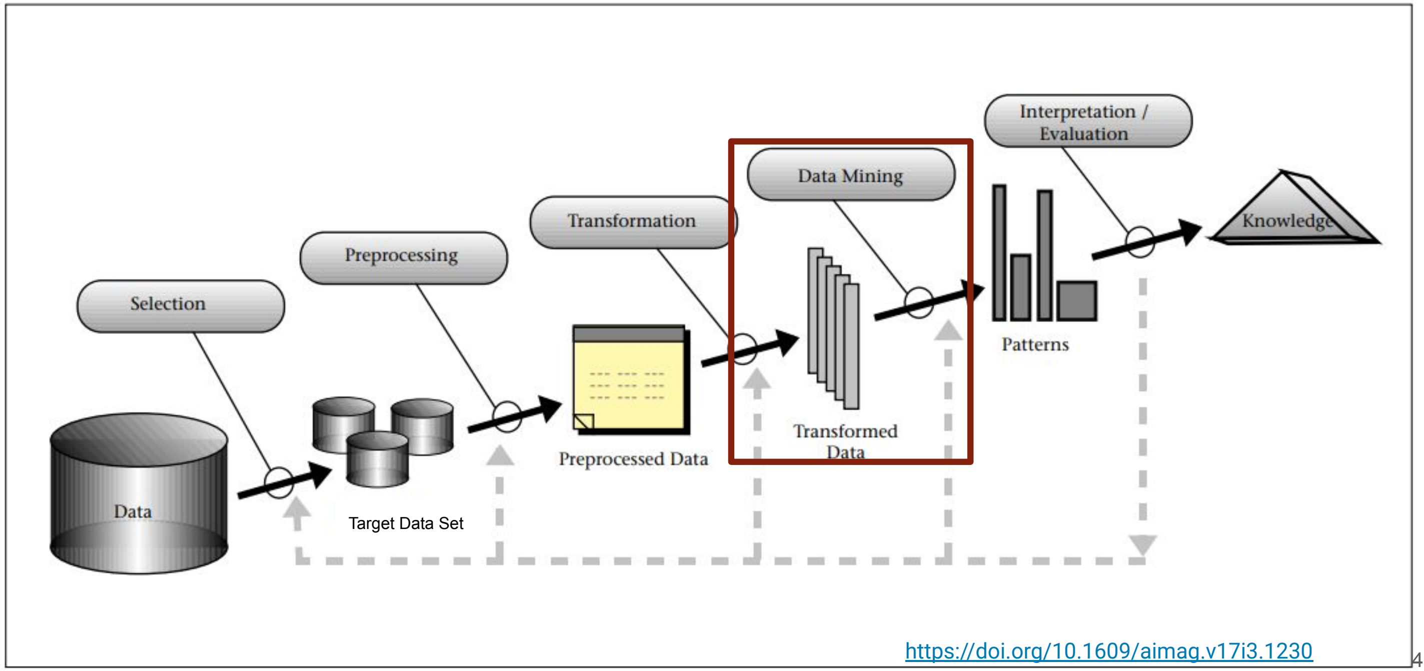
Raw data not always the best data set for data mining

https://doi.org/10.1609/aimag.v17i3.1230

Data preprocessing and transformation

- Dealing with missing data
  - replace all missing values with a single global constant (highly application dependent)
  - replace a missing value with its feature mean
  - replace a missing value with its feature mean for the given class (this approach is possible only for classification problems)
  - generate a predictive model to predict each of the missing values
- Outlier detection
  - Detect and eventually remove outliers as a part of the preprocessing phase
  - Develop robust modeling methods that are insensitive to outliers.
- Data transformation
  - Scaling, encoding, normalization, and selecting features

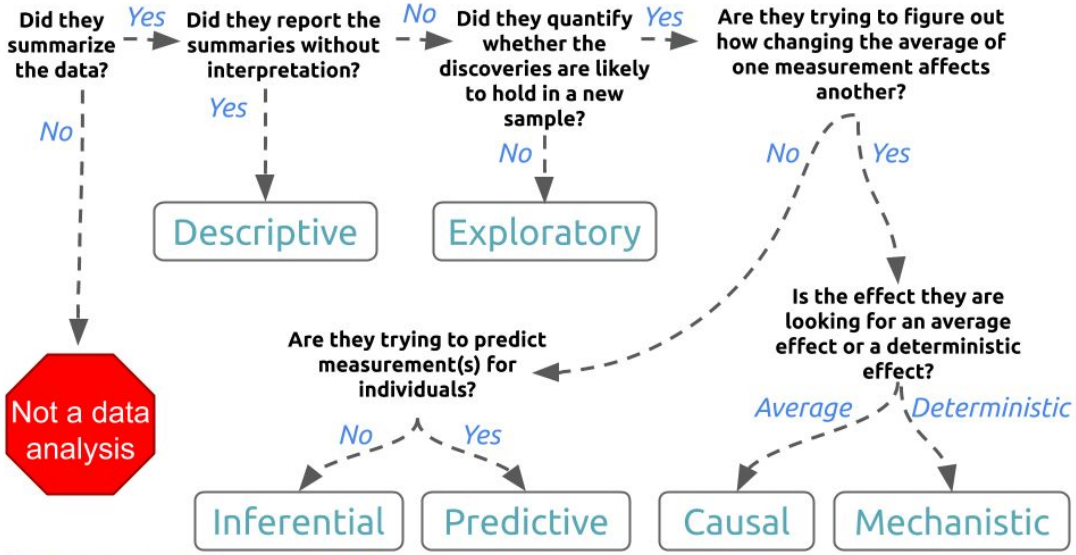
The Data Mining process



https://doi.org/10.1609/aimag.v17i3.1230

Types of analysis

patterns, trends, or relationships between variables.



Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
No
Yes
No
Yes
No
Descriptive
Exploratory
No
Yes
Is the effect they are looking for an average effect or a deterministic effect?
Are they trying to predict measurement(s) for individuals?
Not a data analysis
Average
Deterministic
No
Yes
Inferential
Predictive
Causal
Mechanistic

http://science.sciencemag.org/content/347/6228/1314

make a statement about something outside the dataset

Use a set of features to predict another feature on a given sample

what happens to one measurement if you make another change

not only understand that there is an effect, but how that effect operates

Data Mining Tasks

The two high-level primary goals of data mining tend to be description and prediction.

Descriptive Methods

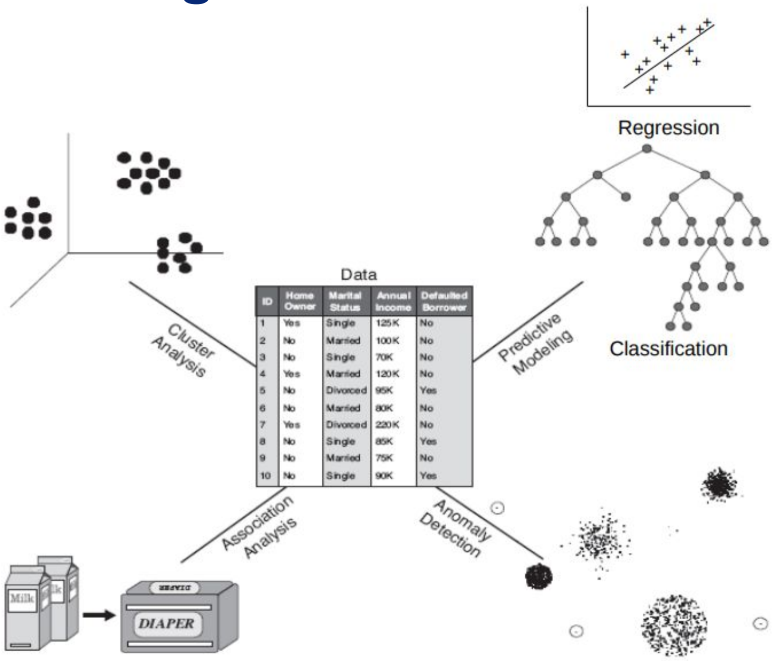
- Find human-interpretable patterns that describe the data (correlations, trends, clusters, trajectories, and anomalies).

Predictive Methods

- Use some features (variables) to predict and unknown or future value of other variable (target variable).

Many data analyses answer multiple types of questions.
The type of question you ask is determined in part by the data available to you

Basic data mining tasks

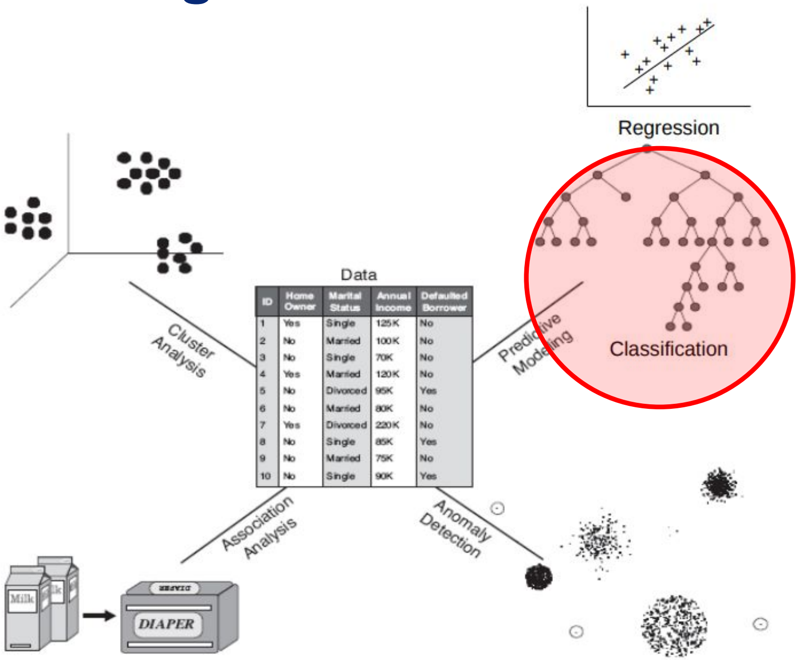


Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley,

Basic data mining tasks

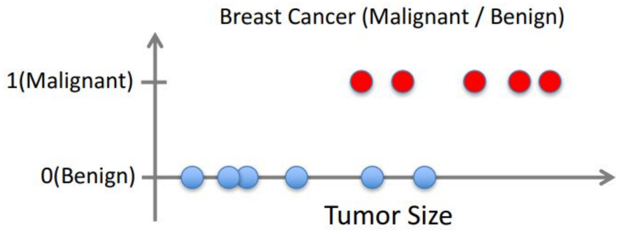


ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Predictive modeling: building a model for the target variable as a function of the explanatory variables.

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley,

Classification



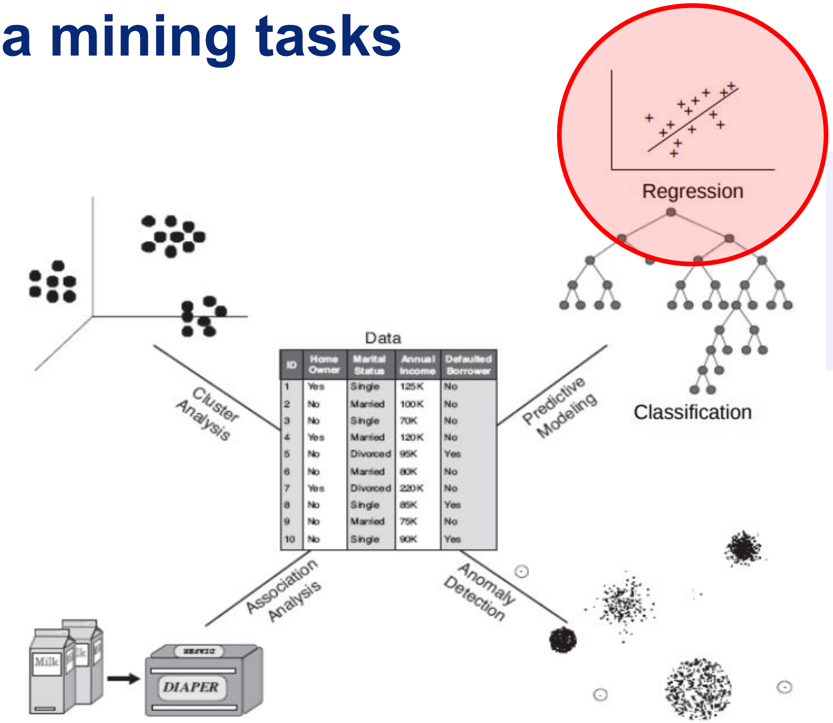
Classification is the process of predicting the class of given data points.
Best used when the output has finite and discrete values
Classes are sometimes called as targets/ labels or categories.

Types of Classification Algorithms

- Linear Models
  - **Logistic Regression**
  - Support Vector Machines

- Nonlinear models
  - **K-nearest Neighbors (KNN)**
  - Kernel Support Vector Machines (SVM)
  - Naïve Bayes
  - **Decision Tree Classification**
  - **Random Forest Classification**

Basic data mining tasks



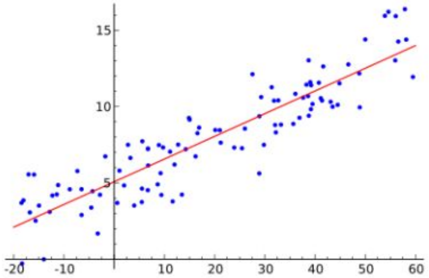
Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Predictive modeling: building a model for the target variable as a function of the explanatory variables.

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley,

Regression



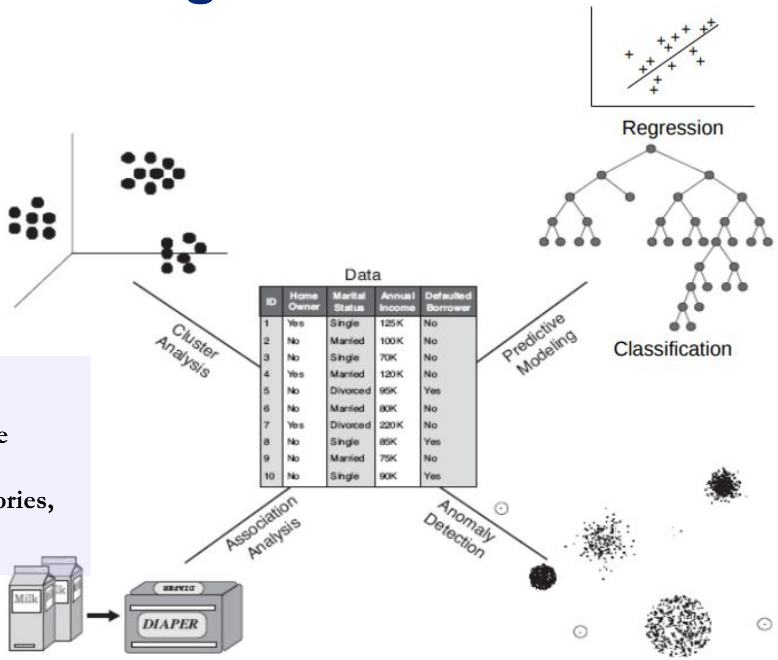
Set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x).

Examples of regression algorithms

- Simple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

Some algorithms can be used for both classification and regression with small modifications, such as decision trees and artificial neural networks.

Basic data mining tasks



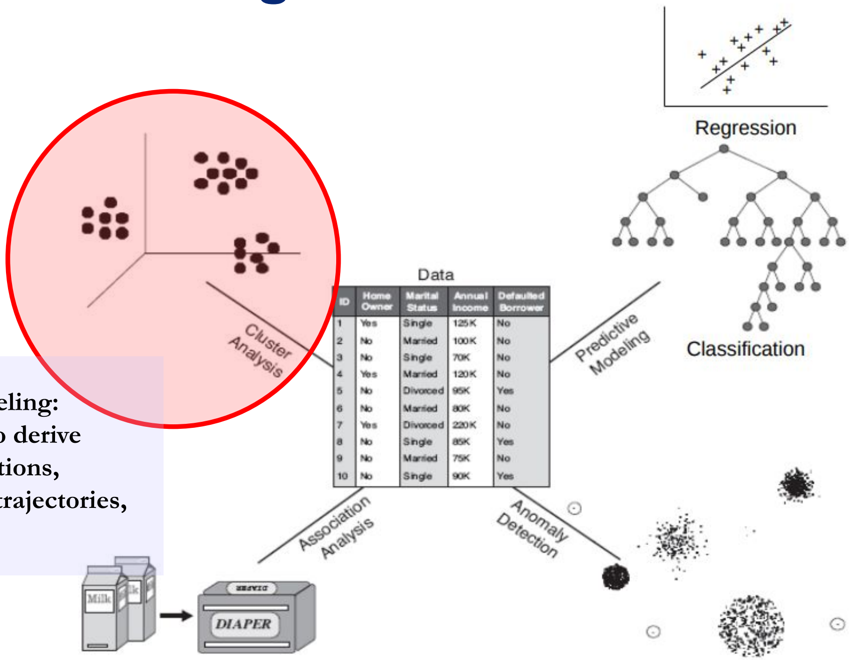
Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Descriptive modeling:
the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies)

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley,

Basic data mining tasks



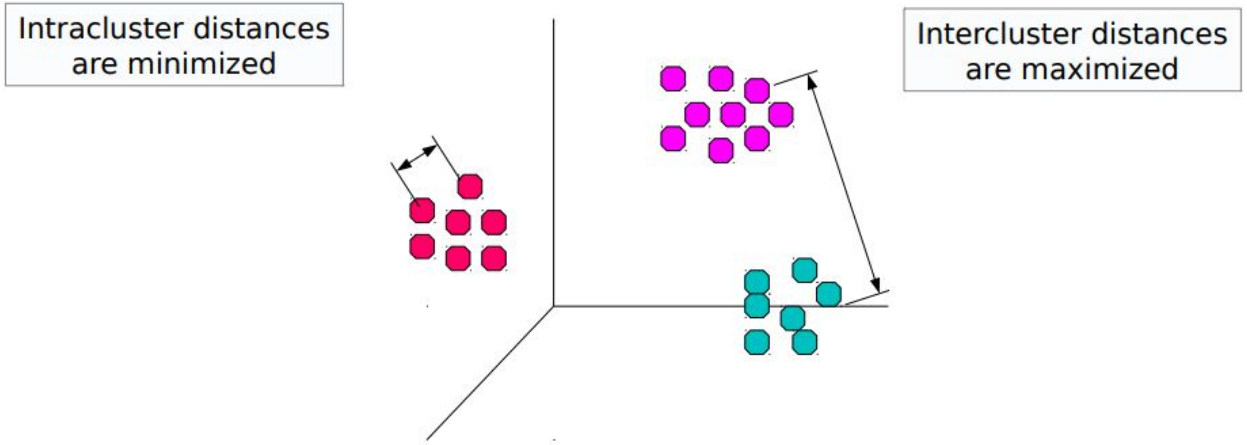
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Descriptive modeling: the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies)

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley,

Clustering

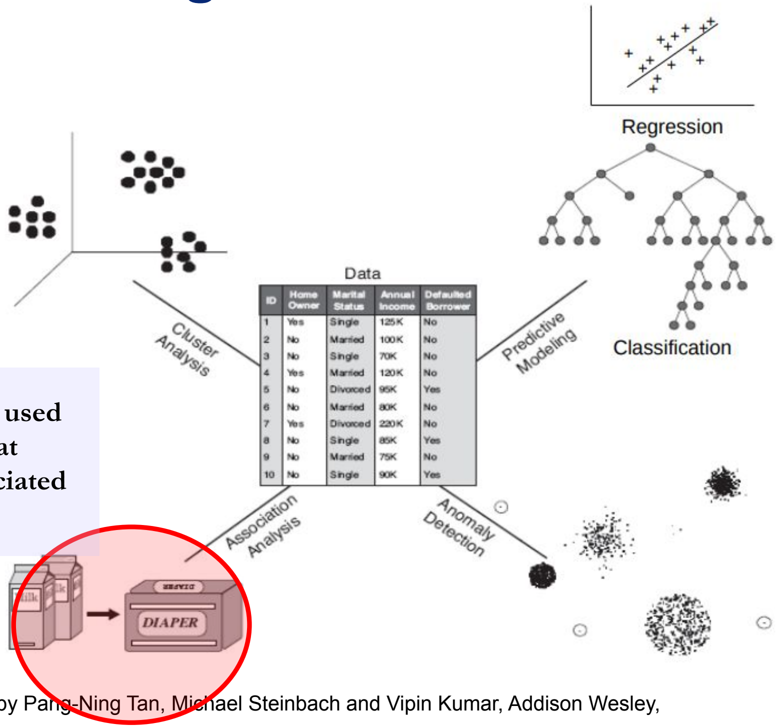
the goal of clustering is to create clusters (groups with the same or similar features)



Intracluster distances are minimized

Intercluster distances are maximized

Basic data mining tasks

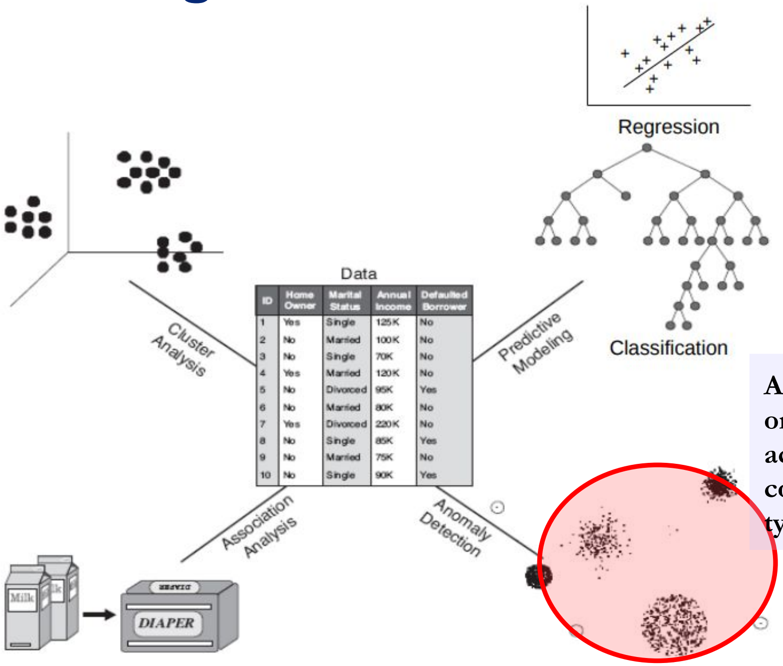


ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Association analysis is used to discover patterns that describe strongly associated features in the data.

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley,

Basic data mining tasks



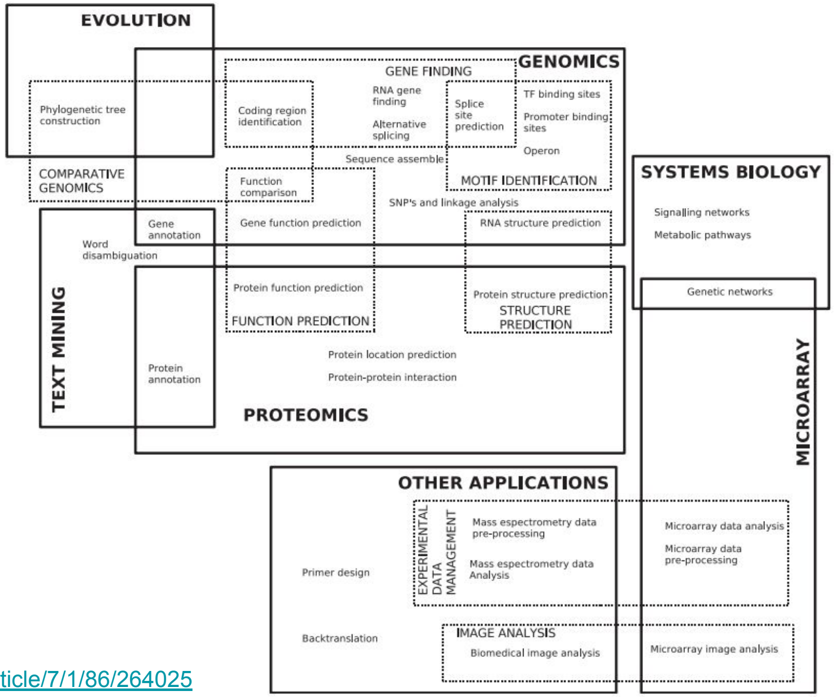
Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Anomaly detection focuses on finding deviations in data according to similar data collected in the past or by typical values of these data.

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley,

Applications of Data Mining in Biomedicine



EVOLUTION

GENOMICS

GENE FINDING

Phylogenetic tree construction

Coding region identification

RNA gene finding

Alternative splicing

Splice site prediction

TF binding sites

Promoter binding sites

Operon

Sequence assemble

MOTIF IDENTIFICATION

COMPARATIVE GENOMICS

Function comparison

SNP's and linkage analysis

SYSTEMS BIOLOGY

Signalling networks

Metabolic pathways

Gene annotation

Gene function prediction

RNA structure prediction

Word disambiguation

Genetic networks

TEXT MINING

Protein function prediction

FUNCTION PREDICTION

Protein structure prediction

STRUCTURE PREDICTION

Protein location prediction

Protein-protein interaction

Protein annotation

PROTEOMICS

MICROARRAY

OTHER APPLICATIONS

EXPERIMENTAL DATA MANAGEMENT

Mass espectrometry data pre-processing

Mass espectrometry data Analysis

Microarray data analysis

Microarray data pre-processing

Primer design

IMAGE ANALYSIS

Biomedical image analysis

Microarray image analysis

Backtranslation

https://academic.oup.com/bib/article/7/1/86/264025

DMI

Best Practices for Data Science Projects

Best Practices for Data Science Projects

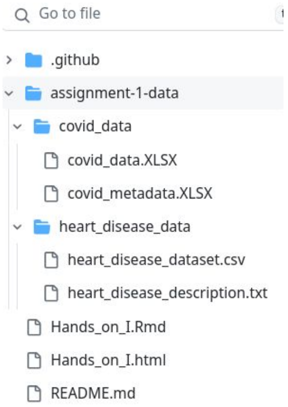
> "Your closest collaborator is you six months ago, but you don't reply to emails."

- Develop your own system
- Put everything in a common directory
- Be consistent
  - directory structure; names
- Separate raw from processed data
- Separate code from data
- It should be obvious what code created what files, and what the dependencies are.
- **No hand-editing of data files !!!!!**
- All aspects of data cleaning should be in scripts
- No manual steps! Each aspect of the analysis should be in scripts.
- Save your seeds for random number generation.
- Make regular backups of your data. In multiple locations.
- Use relative paths (../blah), not absolute paths (~/blah [or] /users/blah)

Best Practices for Data Science Projects

Organizing a data mining project

- README.md
- .gitignore
- project.Rproj
- data/
  - raw_data/
  - processed_data/
  - codebook.md
- code/
  - raw_code/
  - final_code/
- figures/
  - exploratory_figures/
  - explanatory_figures/
- products/
  - writing/



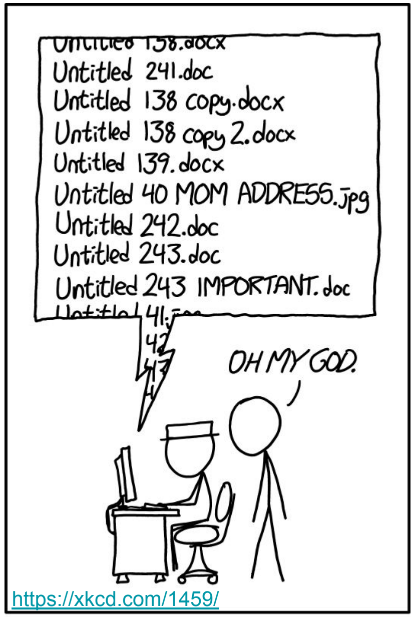
```
Go to file
> .github
v assignment-1-data
  v covid_data
      covid_data.XLSX
      covid_metadata.XLSX
  v heart_disease_data
      heart_disease_dataset.csv
      heart_disease_description.txt
  Hands_on_I.Rmd
  Hands_on_I.html
  README.md
```

"File organization and naming are powerful weapons against chaos."
— @JennyBryan

Dear past-Hadley: PLEASE COMMENT YOUR CODE BETTER. Love present-Hadley
— @hadleywickham

Best Practices for Data Science Projects

Naming files



https://xkcd.com/1459/

PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

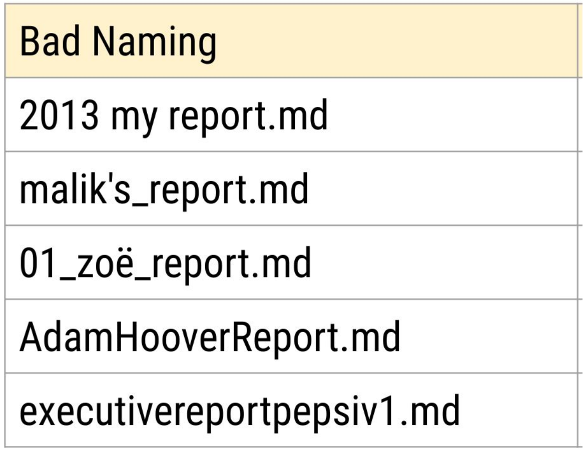
Best Practices for Data Science Projects

Naming files

- **Machine readable:** File names **shouldn't contain spaces, special characters**, and should include important pieces of information about the file contents (sometimes called slugs) separated by underscores or dashes. It is often easier if file names are entirely lowercase letters.
- **Human readable:** files should be labeled with names that make it easy for a person to follow along. So err on the side of **long file names** rather than abbreviations, numeric only file names, or obscure file names.
- **Be nicely ordered:** it is useful to see the files in the order you want them to be run. One way to do this is to name files alpha-numerically so that they will appear in the right order on your computer. One way to do this is to order the files and then append a number to the beginning of each file name (like 01_data_cleaning.R, 02_exploratory_analysis.R etc. so that the files will be ordered correctly)

Best Practices for Data Science Projects

Naming files



Bad Naming
2013 my report.md
malik's_report.md
01_zoë_report.md
AdamHooverReport.md
executivereportpepsiv1.md

Best Practices for Data Science Projects

Naming files

Bad Naming	Good Naming
2013 my report.md	2013_my_report.md
malik's_report.md	maliks_report.md
01_zoë_report.md	01_zoe_report.md
AdamHooverReport.md	adam-hoover-report.md
executivereportpepsiv1.md	executive_report_pepsi_v1.md

Naming variable and functions

Variable and function names should be lowercase. Use an underscore (_) or hyphen to separate words within a name. **Be consistent!** Generally, variable names should be nouns and function names should be verbs. Strive for names that are concise and meaningful (this is not easy!).

```
# Good ✔
day_one
day_1
```

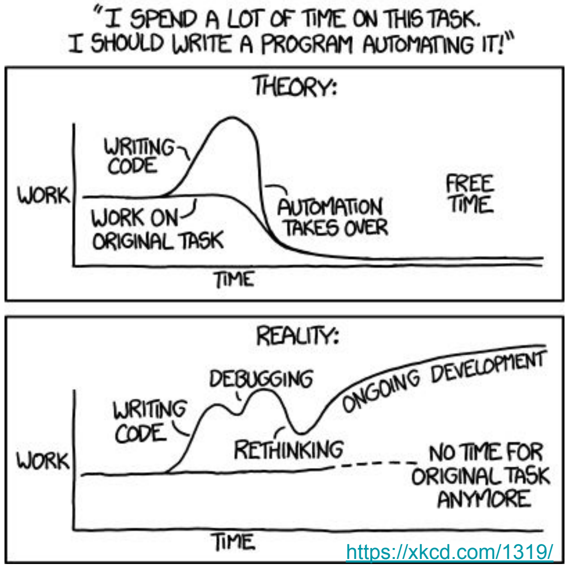
```
# Bad ✘
first_day_of_the_month
DayOne
dayone
djm1
```

Where possible, avoid using names of existing functions and variables. Doing so will cause confusion for the readers of your code.

```
c <- 10 ✘
mean <- function(x) sum(x) ✘
```

Best Practices for Data Science Projects

Automating analysis



https://xkcd.com/1319/

Automate a pipeline

- to reproduce previous results.
- to recreate results deleted by mistake
- to rerun the pipeline with updated software.
- to run the same pipeline on a new data set.

Best Practices for Data Science Projects

Version Control

"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
JORGE CHAM © 2012
Track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc

> "Version control is a truly vital concept that has unfortunately been implemented by madmen."

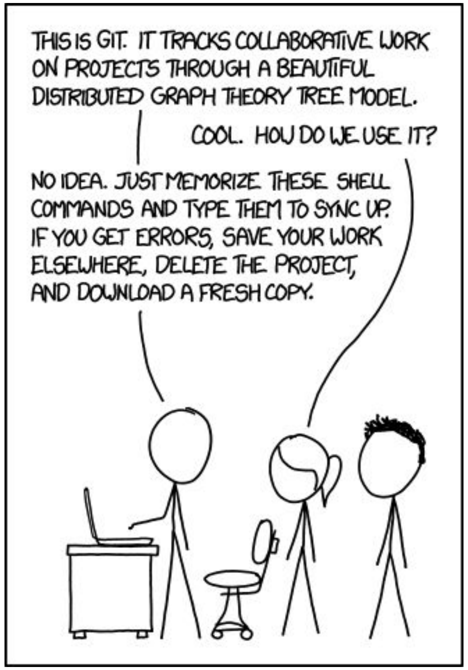
http://phdcomics.com/comics/archive.php?comicid=1531

Best Practices for Data Science Projects

DMI

Version Control: Git

- A version control system: manages the evolution of a set of files – called a repository – in a sane, highly structured way (“the ***Track Changes*** features from Microsoft Word on steroids.” Bryan, 2017)
- Originated to help groups of developers work collaboratively on big software projects.
- Re-purposed by the data science community to manage source code, and other files such as of data, figures, and reports.
- Help communicating and collaborating with other people.


THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.
COOL. HOW DO WE USE IT?
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.

Best Practices for Data Science Projects

Version Control: Github

- GitHub complements Git by providing a slick user interface and distribution mechanism for Git repositories.
- Git is the software you will use locally to record changes to a set of files. GitHub is a hosting service that provides a Git-aware home for such projects on the internet.
- GitHub is like DropBox or Google Drive, but more structured, powerful, and programmatic.
- Available at https://github.com
- Create your account if you don't have one yet!

Best Practices for Data Science Projects

Version Control: Git

In the root directory, create a file called .gitignore

And write in the file:

.Rproj.user

.Rhistory

.RData

.Ruserdata

data/ -> for other cases, not the practical sessions of DMI!!

Best Practices for Data Science Projects

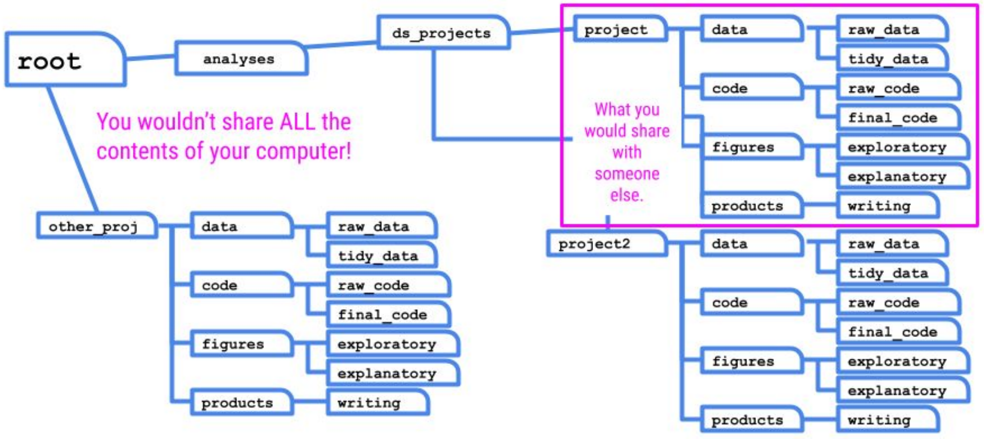
Version Control: Git

```
git add Hands_on_I.Rmd

git add Hands_on_I.html

git commit –m “first commit”

git status

git push
```

To check repository status and history:
git status ->shows if there are changes to any files that have not been committed
git diff file -> shows differences between last snapshot and the current file
git restore file -> discards local changes

Best Practices for Data Science Projects

Dealing with paths

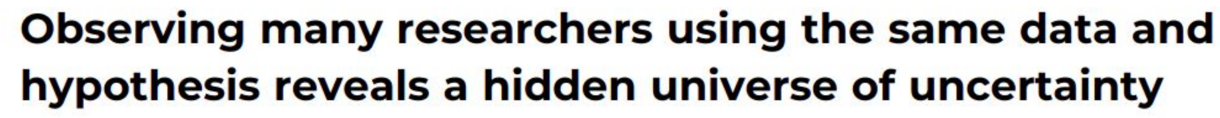


You wouldn't share ALL the contents of your computer!

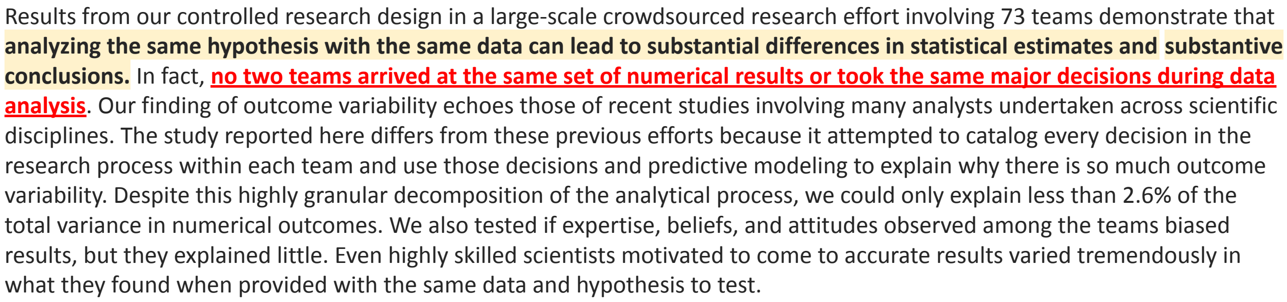
What you would share with someone else.

Absolute paths like `/root/analyses/ds_projects/project/` wouldn't make sense on someone else's computer !

Handling expectations



Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty



Results from our controlled research design in a large-scale crowdsourced research effort involving 73 teams demonstrate that **analyzing the same hypothesis with the same data can lead to substantial differences in statistical estimates and substantive conclusions.** In fact, **no two teams arrived at the same set of numerical results or took the same major decisions during data analysis**. Our finding of outcome variability echoes those of recent studies involving many analysts undertaken across scientific disciplines. The study reported here differs from these previous efforts because it attempted to catalog every decision in the research process within each team and use those decisions and predictive modeling to explain why there is so much outcome variability. Despite this highly granular decomposition of the analytical process, we could only explain less than 2.6% of the total variance in numerical outcomes. We also tested if expertise, beliefs, and attitudes observed among the teams biased results, but they explained little. Even highly skilled scientists motivated to come to accurate results varied tremendously in what they found when provided with the same data and hypothesis to test.

https://doi.org/10.1073/pnas.2203150119

DMI

What distinguishes data science from statistics?

What distinguishes data science from statistics?

Data Mining	Statistics
Explore and gather data first, builds model to detect patterns and make theories.	Usually starts with a hypothesis
Data used is Numeric or Non numeric.	Data used is Numeric.
Inductive Process (Generation of new theory from data)	Deductive Process (Does not involve making any predictions)
Data collection is less important	Data collection is more important.
Data cleaning is important	Clean data is used to apply statistical method.
Suitable for large data sets	Suitable for smaller data sets
Algorithm “learns” from data, no explicit programming rules	Formalization of the relationships in data as a mathematical equation
Examples: Classification, Clustering, Neural network, Association, Estimation, Sequence based analysis, Visualization	Examples: Descriptive Statistical, Inferential Statistical

Bibliography

Introduction to Data Mining, book by Michael Steinbach, Pang-Ning Tan, and Vipin Kumar

The Elements of Statistical Learning, book by Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie

Excuse me, do you have a moment to talk about version control? (2017) preprint by Jennifer Bryan

Tidy Data, paper by Hadley Wickham

Github assignments

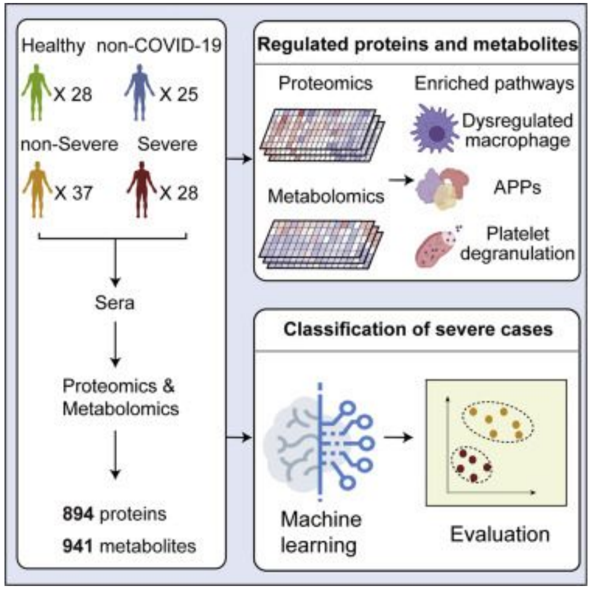
Go to this page and accept the assignment

Go to this sheet and enter your GitHub handle

Choose your scientific article with your partner and update it in the sheet

Hands-on sessions

We will work with the publication **Proteomic and Metabolomic Characterization of COVID-19 Patient Sera**


Healthy
non-COVID-19
X 28
X 25
non-Severe
Severe
X 37
X 28
Sera
Proteomics & Metabolomics
894 proteins
941 metabolites
Regulated proteins and metabolites
Proteomics
Enriched pathways
Dysregulated macrophage
Metabolomics
APPs
Platelet degranulation
Classification of severe cases
Machine learning
Evaluation