U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Data Collection and Storage

Pedro G. Ferreira

pgferreira@fc.up.pt

Agenda

- Sources of Data
- Data Storage
- Data Pipelines
- Data Preparation

Sources of Data

Company Data

- Collected by companies
- Helps make them business decisions
- Types of data:
  - Web events
  - Financial Transactions
  - Survey data
  - Customer Data
  - Logistics Data

Open Data

- Provided in different repositories
  - UCI Machine Learning Dataset Repository
  - Kaggle
  - Google Datasets
- Data published from scientific papers or competitions
- Can be downloaded in different formats or accessed via APIs.
- Public Records
  - UN, WHO, Pordata

Data Storage

- To organize and store your data consider
- Location
  - Server or cluster to run locally
  - Cloud computing (AWS, Azure, GCP)
- Data types
  - Unstructured
    - Email, text, video and audio files, web pages, social media
    - Stored in Document Database
  - Tabular and Structured
    - Data organized as rows and columns
    - Relational Database
- Retrieval and Data Querying

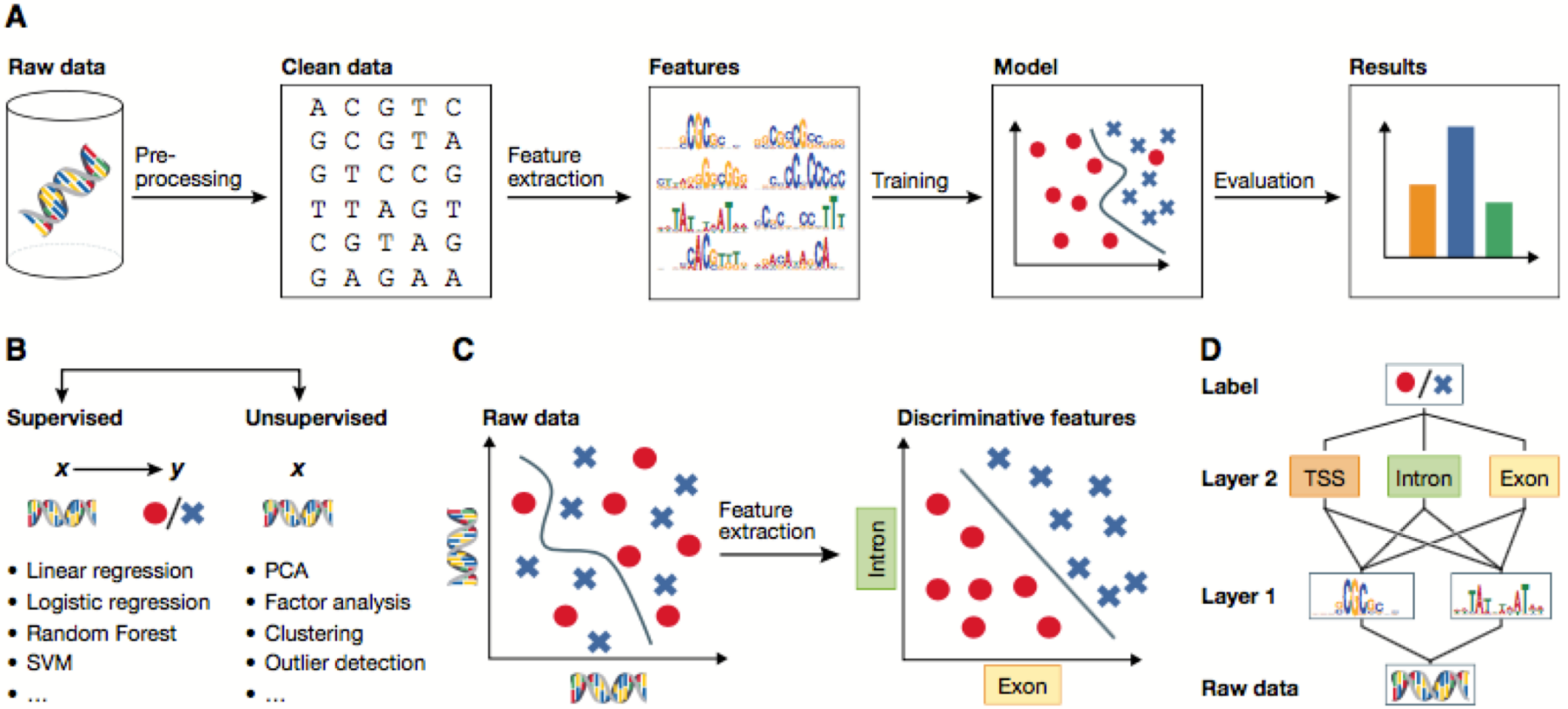
Data Type	Data baseType	Query Language
Unstructured	Document Database	NoSQL
Tabular	Relational Database	SQL

Data Pipelines

- How do we scale the analysis?
  - Multiple data sources
  - Different data types
    - Unstructured data
    - Tabular data
    - Real-time streaming data

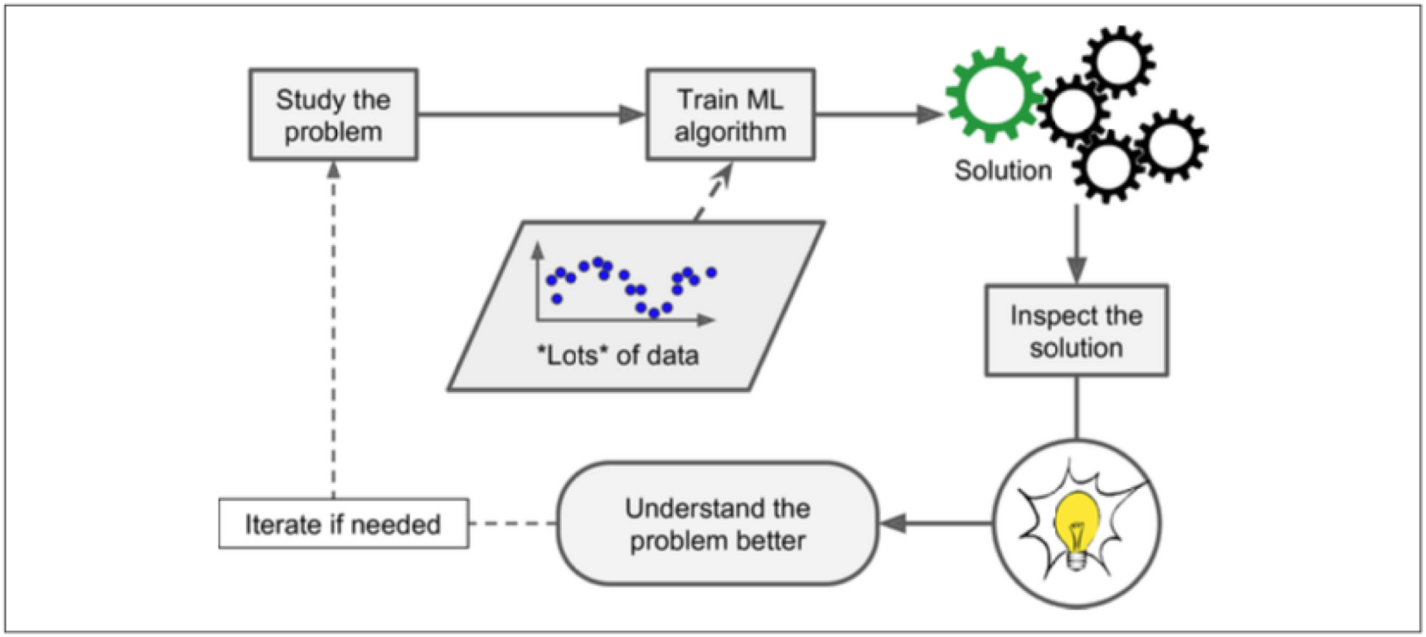
- Data pipeline
  - Moves data into defined stages
  - Automated collected and stored
    - Scheduled by frequency or triggered by an event
  - Monitored with generated alerts
  - Extract Transformation and Load (ETL)

Data Pipelines



- Copied from Angermueller C. *et al.* Deep learning for computational biology. Mol Syst Biol. 2016 Jul 29;12(7):878

Data Pipelines



- Adapted from Aurélien Géron, Hands-On Machine Learning with Scikit-Learn & TensorFlow

Data Preparation

- Why prepare data?
  - Real-life data is messy
  - Processing is done to prevent:
    - Errors
    - Biasing algorithms
    - Incorrect results
- Tidy data
  - Organize cases as rows
  - Features as columns

Adapted from DataCamp

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Data Preparation

- Remove duplicates

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

- Unique Identifiers

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

- Homogeneity

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

Data Preparation

- Data Types

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

- Missing values
  - Data entry
  - Error
  - Valid missing value

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien	28	1.80	"FR"

Handling Missing values
  - Impute (mean, max, median, ...)
  - Drop
  - Keep it if the algorithm handles it

Summary

- Data rarely comes in ready for analysis. Real-life data is messy and dirty.

- Preparing the data conveniently will save you time in later stages of analysis.

- Keep in mind the multiple steps of the analysis pipeline from retrieving the data to presentation of results.

- Most algorithms require data in tabular format without missing data or duplicates. Check that your data is in the right tabular format with cases as rows and features as columns; that you have no missing values; that data types are conveniently represented.