IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Introduction to Data Science

- Datasets: scale, types
- Data Science pipeline

Data Science Toolkit & Applications
Lina Fahed & Laurent Brisson

Datasets: scale, types

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

What is a dataset?

A dataset is a collection of data **objects** described by **attributes**

Objects

the atomic elements from a dataset
e.g. records, prototypes, objects, cases, points, samples, instances

Attributes

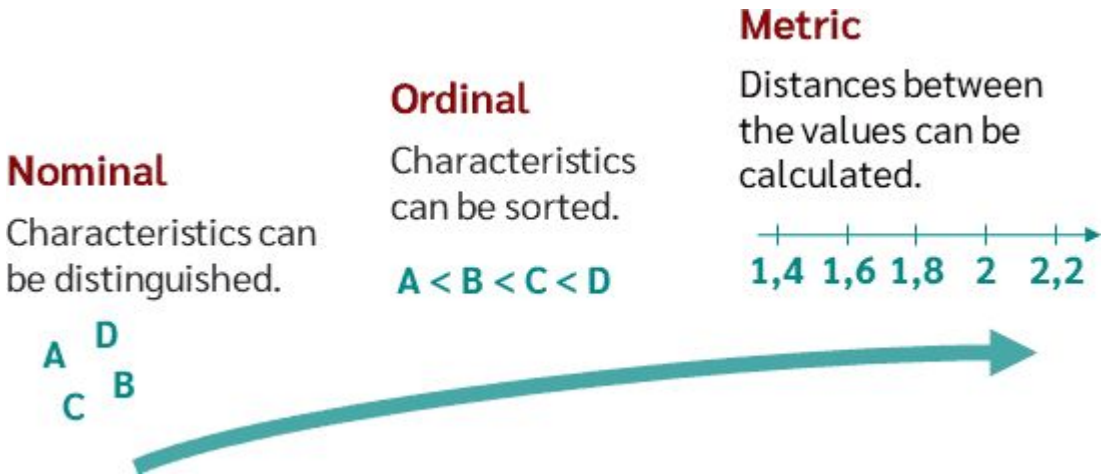
a property of an object
e.g. features, variables, fields, characteristics

Titanic dataset

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily	female	35.0	1	0	113803	53.1000	C123	S

Attribute Types: Scale/level of measurement

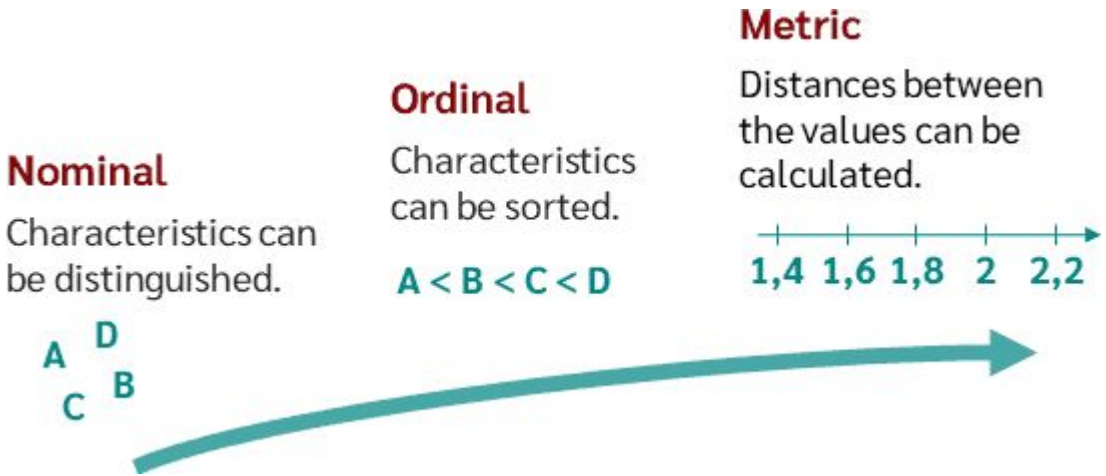
- tells us how our data can be collected, analyzed, and interpreted
- helps us understand what kind of information the data provides



https://numigo.com/tutorial/level-of-measurement

Attribute Types: Scale/level of measurement

- determines the types of calculations and **statistical analyses** we can perform
- **mean and standard deviation:**
  - are **suitable** for metric data
  - **may be suitable** for ordinal data, but only if you know how to interpret the results correctly
  - definitely make **no sense** for nominal data



https://numigo.com/tutorial/level-of-measurement

Attribute Types: Scale/level of measurement

Nominal scale

- values are names or labels
- no ordering, no distance measure, no relation among values
- only equality tests and set membership can be performed
- boolean as a special case

Gender
1 = male
2 = female

Marital status
single
married
divorced
widowed

Preferred newspaper:
The Washington Post
The New York Times
USA Today
...

Attribute Types: Scale/level of measurement

Ordinal scale

- Values with an order
- Low level of measurement: few information
- contains nominal information, only with the difference that a ranking can be formed
- no distance measure (no relative size or degree of difference): distances between the values are not interpretable
- relation w.r.t. the order, addition does not make sens

Frequency of television:
1 = daily
2 = several times a week
3 = less frequently
4 = never

The government is doing a good job:
1 = agree with
2 = undecided
3 = disagree with

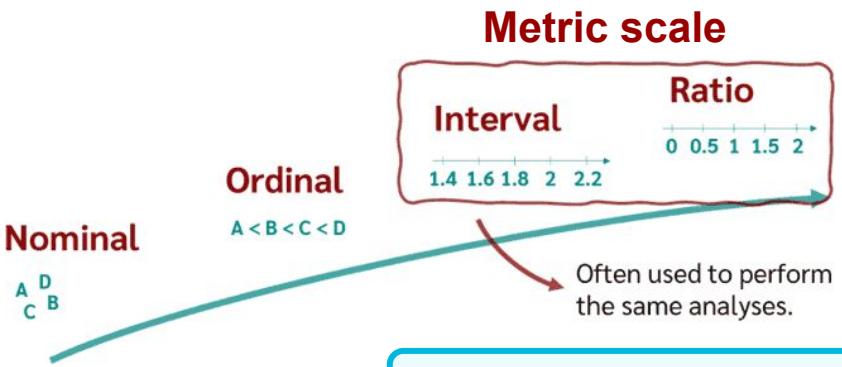
Attribute Types: Scale/level of measurement

Metric scale

- Values ordered and measured in fixed and equal units
- values can be compared and sorted
- distances between the values can be calculated

Income	Weight	Age	Electricity consumption
1820 $	81 kg	18 years	520 kWh
3200 $	70 kg	27 years	470 kWh
800 $	68 kg	64 years	340 kWh
...	...	...	...

Attribute Types: Scale/level of measurement



Ratio statement

"One value is twice as large as another" ⇒ absolute zero must be available as a reference

Marathon runners: time is measured

interval

- **stopwatch is forgotten** to start at the start of the marathon
- **only the differences** are measured starting from the fastest runner
- it can be said how big the interval between the runners is
- e.g. runner A is 22 minutes faster than runner B)
- it cannot be said that runner A ran 20 percent faster than runner B.

ratio

- statement: "fastest runner is **twice** as fast as the last runner"
- **absolute zero** is the starting time of the marathon

Attribute Types: Scale/level of measurement

Exampes

		Scale level
1	States of the USA	nominal
2	Product rating on a scale from 1 to 5	ordinal
3	religious confession	nominal
4	CO2 emissions in the year	metric, ratio scale
5	IQ-Score of students	metric, interval scale
6	examination grades from 1 to 5	ordinal
7	telephone numbers of respondents	nominal
8	care level of a patient	ordinal
9	Living space in m2	metric, ratio scale
10	job satisfaction on a scale from 1 to 4	ordinal

Dataset types

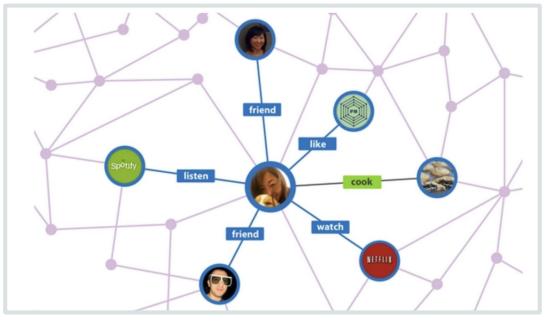
record
- data matrix
- document data
- transaction Data

Titanic dataset

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily	female	35.0	1	0	113803	53.1000	C123	S

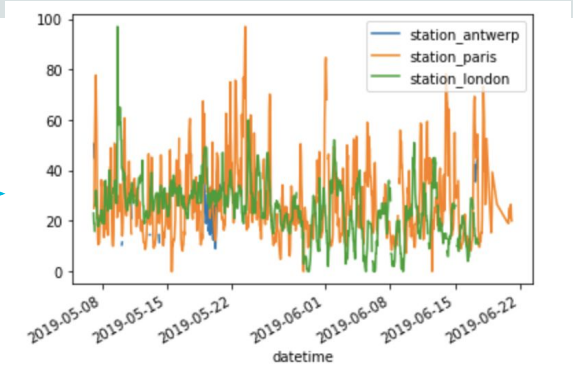
graph
- World Wide Web
- molecular structures

Social networks dataset



Dataset types

ordered
- spatial data
- temporal data
- sequential data
- genetic sequences



Text
- Structured: Spreadsheets, XML, JSON
- Unstructured: Social networks, emails
- Semi-structured: HTML…
- Labeled/unlabeled
- Multilingual

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፩፰ ዓመት ያሳለፈውን ዝንጣን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Nigeria sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke nọgoro afọ iri na asatọ na Naijiria
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Émìà ilú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàìjíríà

Dataset characteristics

- **dimensionality**: number of attributes
- **class label unbalance**: prior class probabilities
- **size**: number of samples
- **quality**: missing values, outliers, precision
- **sparsity**: presence of attributes values

most algorithms only work well on data in large quantity, with a reasonable number of attributes

high-unbalanced class label: bad performance of classifiers for the minority class

robustness, complexity issues

quality of the outputs

mining of associations among attribute sets only works when values are present

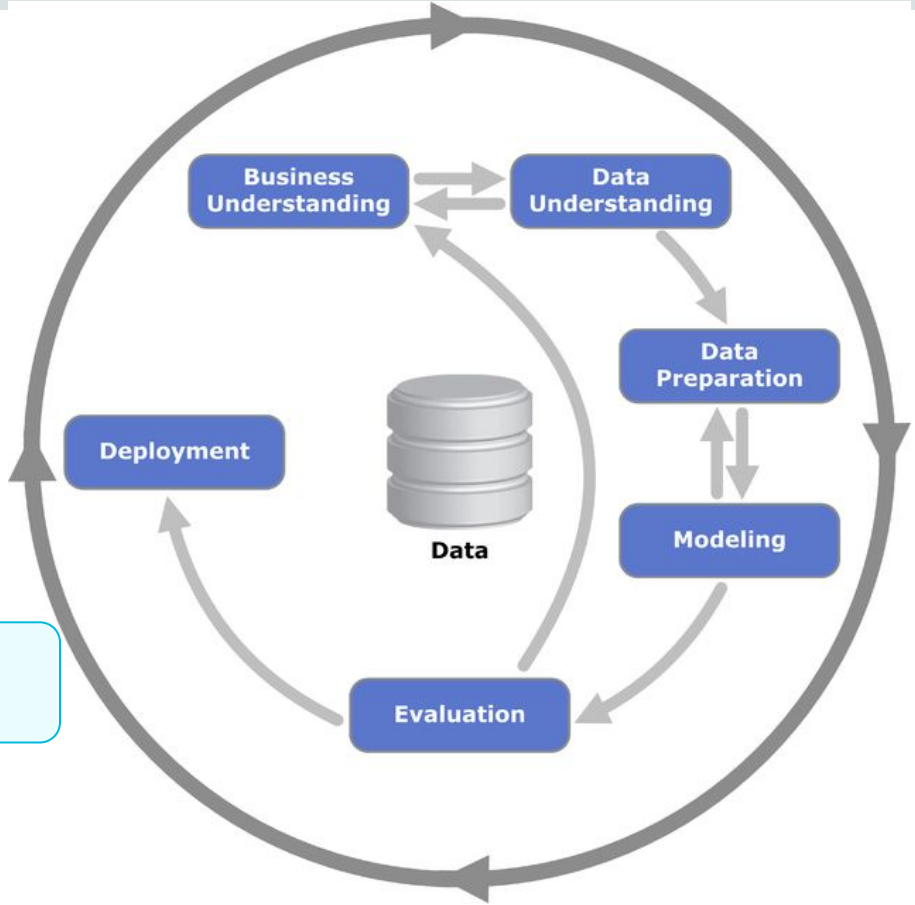
Data science pipeline: CRISP-DM

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

CRISP-DM: Cross Industry Standard Process for Data Mining

- A step-by-step data science guide
- Designed to solve the existing data science problems in the industries
- The most common methodology for data mining, analytics, and data science projects



Other methods

- KDD (Knowledge Discovery in Databases)
- SEMMA (Sample, Explore, Modify, Model, Assess)

CRISP-DM: 1. Business Understanding

- Determine business objectives:
  - understand, from a **business perspective**, what the customer really wants to accomplish
  - define business success criteria
- Assess situation:
  - Determine resources availability, project requirements
  - Assess risks and contingencies, and conduct a cost-benefit analysis
- Determine data science goals:
  - Define what success looks like from a technical data mining perspective
- Produce project plan:
  - Select technologies and tools and define detailed plans for each project phase

While many teams hurry through this phase, establishing a strong business understanding is like building the foundation of a house – absolutely essential.

CRISP-DM: 2. Data Understanding

- Collect initial data: Acquire the necessary data
- Describe data: Examine the data and its properties like data format, number of records, or field identities
- Explore data: Dig deeper into the data. Query it, visualize it, and identify relationships among the data
- Verify data quality: How clean/dirty is the data? Document any quality issues

CRISP-DM: 3. Data Preparation

- Select data: Determine which data sets will be used
- Clean data: correct, impute, or remove erroneous values
- Construct data: Derive new attributes that will be helpful. For example, derive someone's body mass index from height and weight fields
- Integrate data: Create new data sets by combining data from multiple sources
- Format data: Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations

CRISP-DM: 4. Modeling

- Select modeling techniques: Determine which algorithms to try (e.g. regression, clustering, ...).
- Generate test design: You might need to split the data into training, test, and validation sets.
- Build model: As glamorous as this might sound, this might just be executing a few lines of code like *“reg = LinearRegression().fit(X, y)”*.
- Assess model: Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design

Machine Learning

CRISP-DM: 5. Evaluation

- Evaluate results: Do the models meet the **business success criteria**? Which one(s) should we approve for the business?
- Review process: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
- Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

CRISP-DM: 6. Deployment

A model is not particularly useful unless the customer can access its results.

- Plan deployment: Develop and document a plan for deploying the model
- Plan monitoring and maintenance: Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-proj phase) of a model
- Produce final report: The project team documents a summary of the project which might include a final presentation of data mining results
- Review project: Conduct a project retrospective about what went well, what could have been better, and how to improve in the future