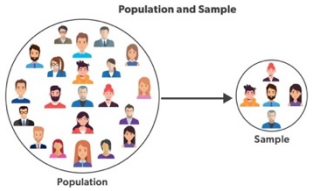
Data Literacy

The ability to explore, understand, and communicate with data

Sample vs Population



BI applications and implementations are built on data.
Data comes as registered facts about real objects and events.

We have collected only a little sample of all existing data but we still want to make conclusions that a valid for the whole population. It requires following some rules.

Observations vs Features

Our sample consists of multiple observations of individual items, data points.

In a table representation **`one observation = one row`** (like in SQL database).

Each observation is presented by same multiple features, or properties **`one feature = one column.`**

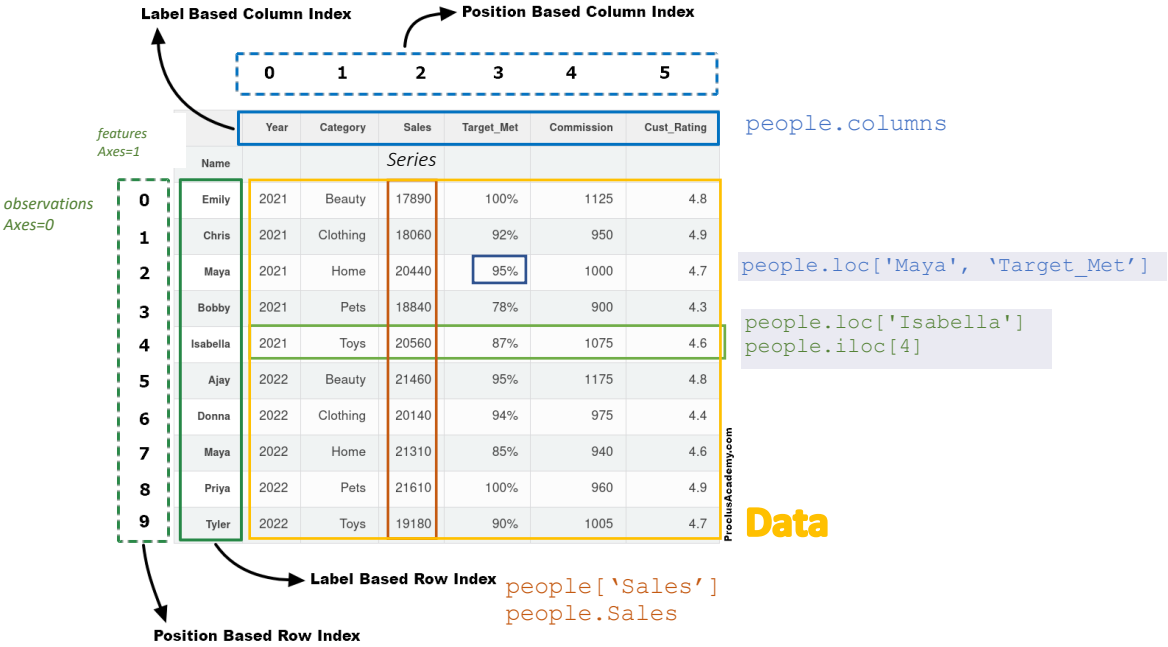
In Python programming, we process data frames by use of a software library called `pandas`. The data in a `pandas DataFrame` is actually a collection of `Series` – every column is one `Series` object.

The `Series` itself is a one-dimensional array with indices, a dictionary type of a key (index) and a value (column value) pairs.

Label vs Position

The data in the rows and columns can be addressed either by label or by position:

```
iloc[Position-based Row Indexes, Position-based Column Indexes]
loc[Labeled Row Indexes, Labeled Column Indexes]
```



Data Types

Numeric (quantities) = measures

- Numeric values measured in even intervals - *age, distance, temperature*
- Measurement needs a scale
  - Discrete – clearly distinguishable values, like *1, 2, 3*
  - Continuous – no strict border between two neighbor values, like between *1.1 and 1.2 are 1.11, 1.111, 1.11111111, etc*

Categorical (qualities) = labels, symbolic names

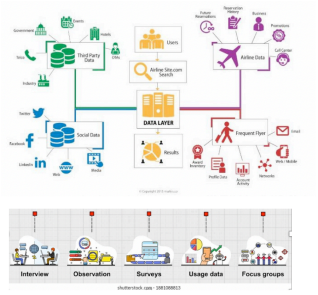
- Ordinal data
  - there is a logical rank-order relationship between the range of values – such as *{never, sometimes, mostly, always}; {good, better, best}*
- Nominal data
  - categories cannot be ranked - *Asia, Europe, America, Australia; male, female*

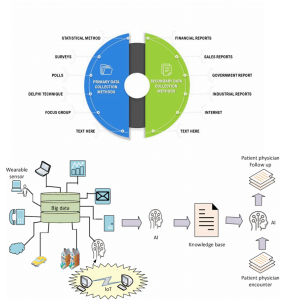
More examples:

Numeric Data	Ordinal Categorical	Nominal Categorical
Weight (10 kg. 35 kg, 100 kg)	Gold Silver Bronze	Africa Asia Europe
Cost ($500, $2.5 M, $4 B)	Excellent Good Poor	Alice Bob Chris
Discount (0%, 2.5%, 5%)	January February March	Wine Beer Water
*We need a **scale** to measure them* {arithmetic operations}	*Ordered in some way* {< >}	*Equivalent in meaning* {= ≠}

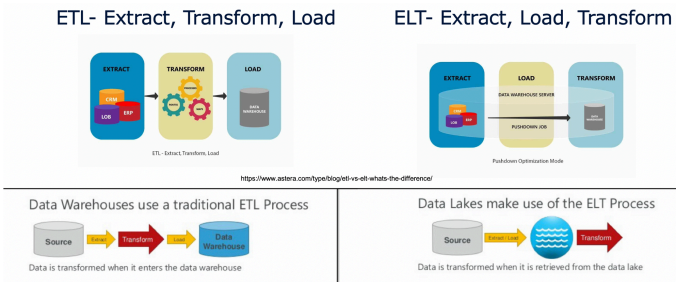
The different categories of data require different techniques of processing.

Data comes from everywhere, as illustrated here:

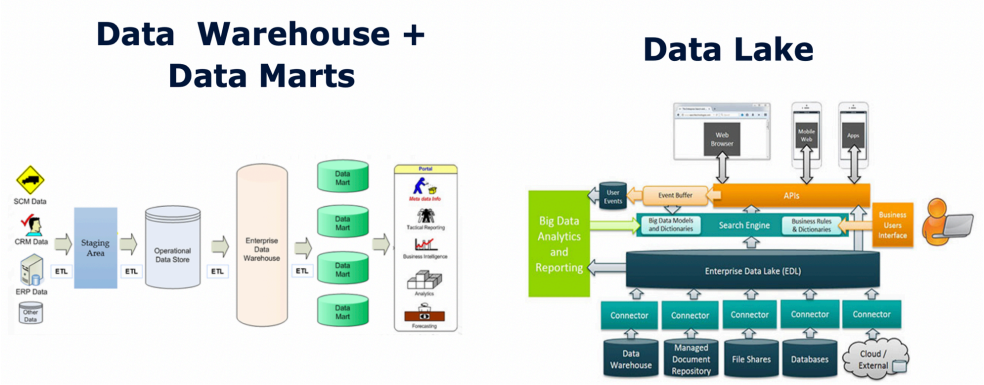




ETL vs ELT



Data Warehouse vs Data Lake



Data Ingestion vs Data Wrangling

Data Ingestion is the process of obtaining, importing, and processing data for later use or storage in a database. This can be achieved manually, or automatically using a combination of software and hardware tools designed specifically for this task (IBM).

Data Wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics (Wikipedia).

Data Quality

Good Data Quality

- Volume – larger amount data with lowest level of details provides better chance to find good answers
- History – longer-time collected data contains more clear patterns
- Variety – variety of observations and features reveals better findings
- Consistency – fitting the old and the new data
- Clarity – meaningful labels - better understanding
- Aggregation and segmentation – grouping by categories, as possible
- Transparency of data - clear origin, clear pre-processing
- Clean data – repairing noise and missing data
- Well-structured data – optimal for BI operations

Bad Data Quality

- too low number of observations
- too high number of features
- missing data
- distorted distributions, outliers, anomalous examples
- redundancy of information, duplication
- invalid data formats, special characters
- wrong data structures
- non-consistent labels

How to Prepare Good Data?

1. Obtain meaningful data, correctly measured and usefully labelled
2. Acquire sufficient data – may not be possible to tell in advance, but new data can be obtained or generated later
3. Shape the data in data frames or other appropriate structures
4. Pre-process it
    A. Clean the data
    B. Transform the data values as necessary
    C. Revise and reengineer the features

Data Preprocessing Operations

A. Data Cleaning

Finding and removing incorrect and inaccurate records from a set or a data source

- examples for garbage data: duplicate values, dummy values, missing data, and contradictory data
- examples for a cause: corruption in the technical systems – collection, transmission, storage

Data cleaning includes activities like

- removing typographical errors
- validating and correcting values
- harmonizing and standardizing data

B. Data Transformation

Converting and mapping data from one format to another format

- first, data is extracted from a data source in its raw format
- then, it is parsed into a predefined data structure or processed by an algorithm
- finally, it is stored in a storage unit for future use

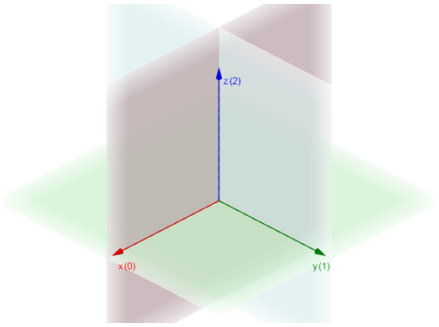
Different tools for data wrangling are available (see https://www.varonis.com/blog/free-data-wrangling-tools/)

C. Feature Engineering

Re-engineering the initially available attributes/features.

Can include:

- ignoring insignificant attributes
- calculation and generation of new attributes
- aggregating or merging attributes
- splitting attributes
- replacing all available attributes with a more appropriate set


z (2)
x (0)
y (1)