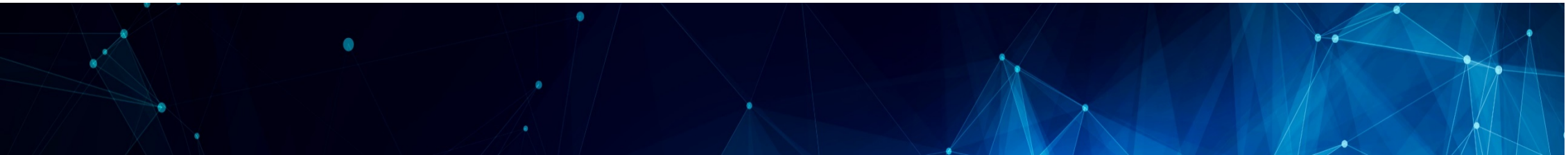
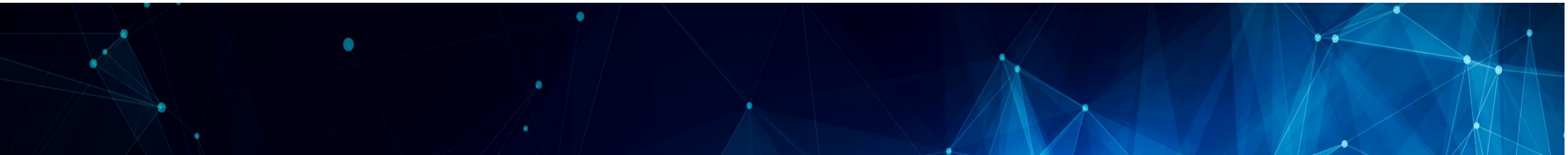
Fundamentals of Data Science

Data collection and management



Southampton Data Science Academy

Exploring and fixing data



Southampton Data Science Academy

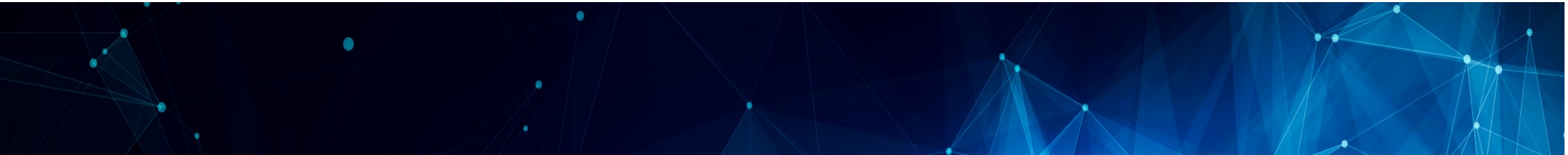
Data Pre-processing (1)

No quality data, no quality results!

- Quality decisions must be based on quality data. E.g., duplicate or missing data may cause incorrect or even misleading statistics
- Any data store needs consistent integration of quality data

Data extraction, cleaning, and transformation comprises the majority of the work of a data scientist

- Data extraction is the process of retrieving the data from storage
- Data transformation is the process of converting the initial format of the data into a different format to conduct analysis on



Data pre-processing (2)

Data in the real world are dirty…

- Incomplete lacking attribute values
  - Example: `occupation = “”`
  - Lacking certain attributes of interest, or containing only aggregate data
- Noisy: containing errors or outliers
  - e.g., `salary=”-10”`
- Inconsistent*:* containing discrepancies in codes or names
  - e.g., `Age=”42” Birthday=”03/07/1997”`
  - e.g., was rating `“1,2,3”`, now rating `“A, B, C”`

Data pre-processing (3)

Data needs to be summarised to ascertain its value and limitations

- Without a summary, we have no idea what the general content of the data is, leaving us unable to ascertain its value
- Providing an overview of your data is key, e.g., with descriptive statistics allows the features to be easily gathered

Fixing missing data

Option 1

Remove all records which have missing data

- This can be misleading, e.g., participants dropping out of a study
- May be acceptable if missing data are completely random

Option 2

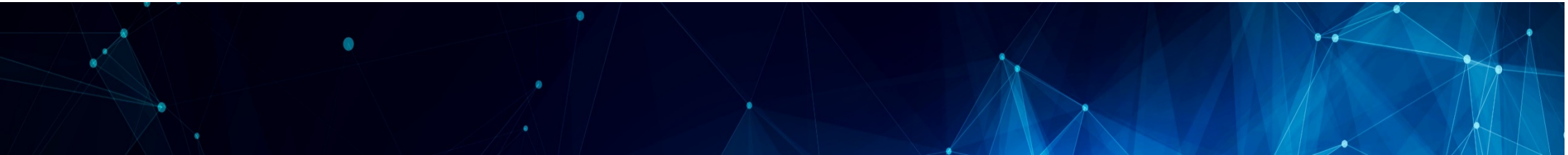
A global constant, e.g., “unknown”?

- Warning: The attribute mean could skew the data

Add data based on other data points,

- e.g., the mean for all samples belonging to the same class

Most probable value, inference-based such as Bayesian formula



Southampton Data Science Academy

Noisy data

Noise

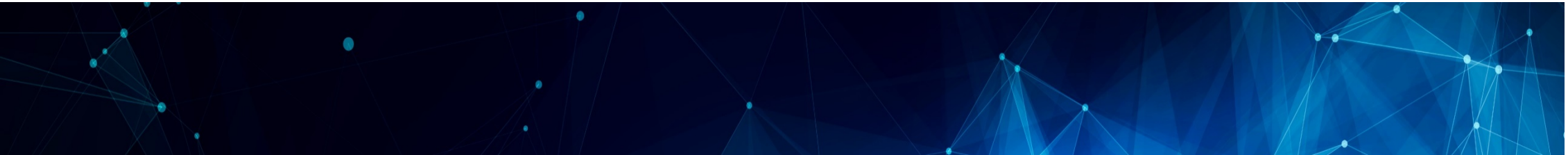
- Random error, variance, spam

Incorrect attribute values may be due to:

- Technological limitation (e.g., limited bandwidth)
- Changing data schema
- Wrong data collected (this actually happens a lot!)

Other data problems which require cleaning

- Duplicate records
- Incomplete or inconsistent data



Southampton Data Science Academy

Dealing with noisy data

Moving Average

- A fixed “subset” size is created, and the average is computed for each subset
- Each datum is modified by a subset of the same size, excluding the previous value

Local Regression

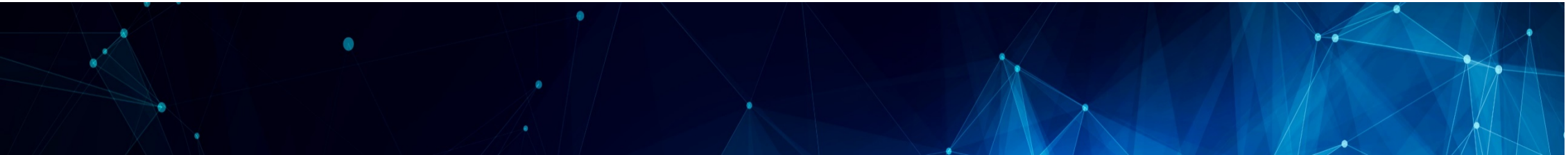
- Use a regression function which calculates the value of data points based on the data points around them

Clustering

- Combine the data into clusters, and remove outliers from those data

Combined computer and human inspection

- Manually check suspicious values - (takes time!)



Southampton Data Science Academy

Data integration

Combines data from multiple sources into a coherent store

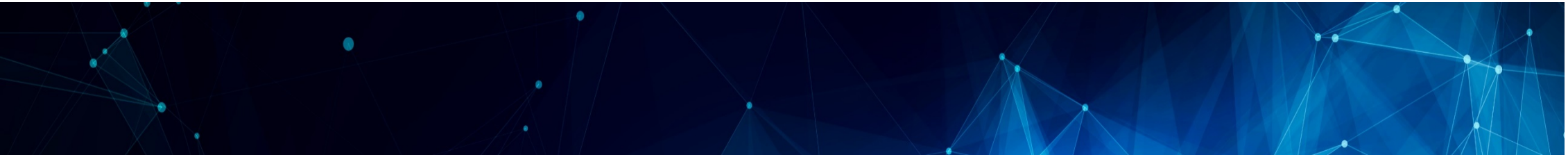
- Data may be coming from different sources, devices
- Data might not be collected at the same point in time

Integrate metadata from different sources

- Entity identification: identify real world entities from multiple data sources
- Imagine two databases tables A and B in different databases
- The customer ID might be defined as `A.user_id` and `B.customer_id`

Detecting and resolving data value conflicts

- For the same real world entity, attribute values from different sources are different
- Possible reasons; different representation, different scales,
  - E.g., metric vs. imperial units



Southampton Data Science Academy

Entity resolution

- Entity resolution is identifying which entities are the same when they come from different datasets
- This is a difficult problem to do automatically, because it inevitably has a quadratic computational cost: each entity has to be compared with every other entity
- Blocking groups similar entities into “blocks”, and performs comparisons within those blocks. Still quadratic, but on a far smaller size of data
- Will likely have to compromise on precision or recall for big datasets

Entity resolution

Despite best efforts, there could be duplication of data at the schema level as well as at the instance level

There are no perfect ways of identifying equivalent fields automatically, but some things can be checked for:

- Synonyms and translations
- Text similarity
- Similarity between comments or documentation of an entity
- Use additional information from related fields or instances to infer the equivalent field

Data integration: conflicts

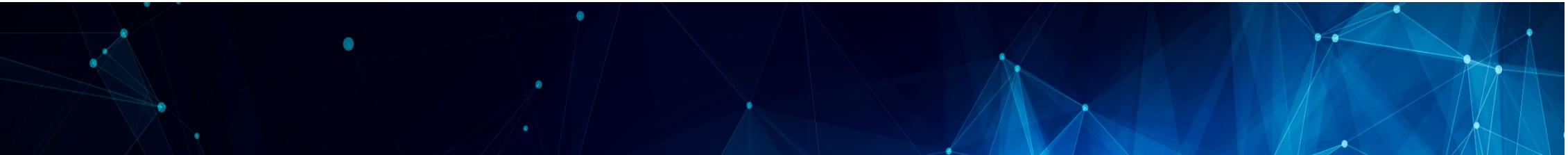
In the event that there are conflicting data, there is no easy solution

- Even if one source is trusted above other sources, it is not always correct
- Data evolves over time, so one may be out of date
- Data is often copied from one store to another, so mistakes can propagate

When resolving a conflict, these different issues must be balanced to get the one most probable to be correct

- The trustworthiness of the data
- The freshness of the data
- Considering any dependence between sources

There will have to be some manual inspection of the data no matter what!



Southampton Data Science Academy