Data collection and preparation

Fenna Feenstra, Msc. Data Science for Life Sciences
Hanze University of Applied Sciences Groningen

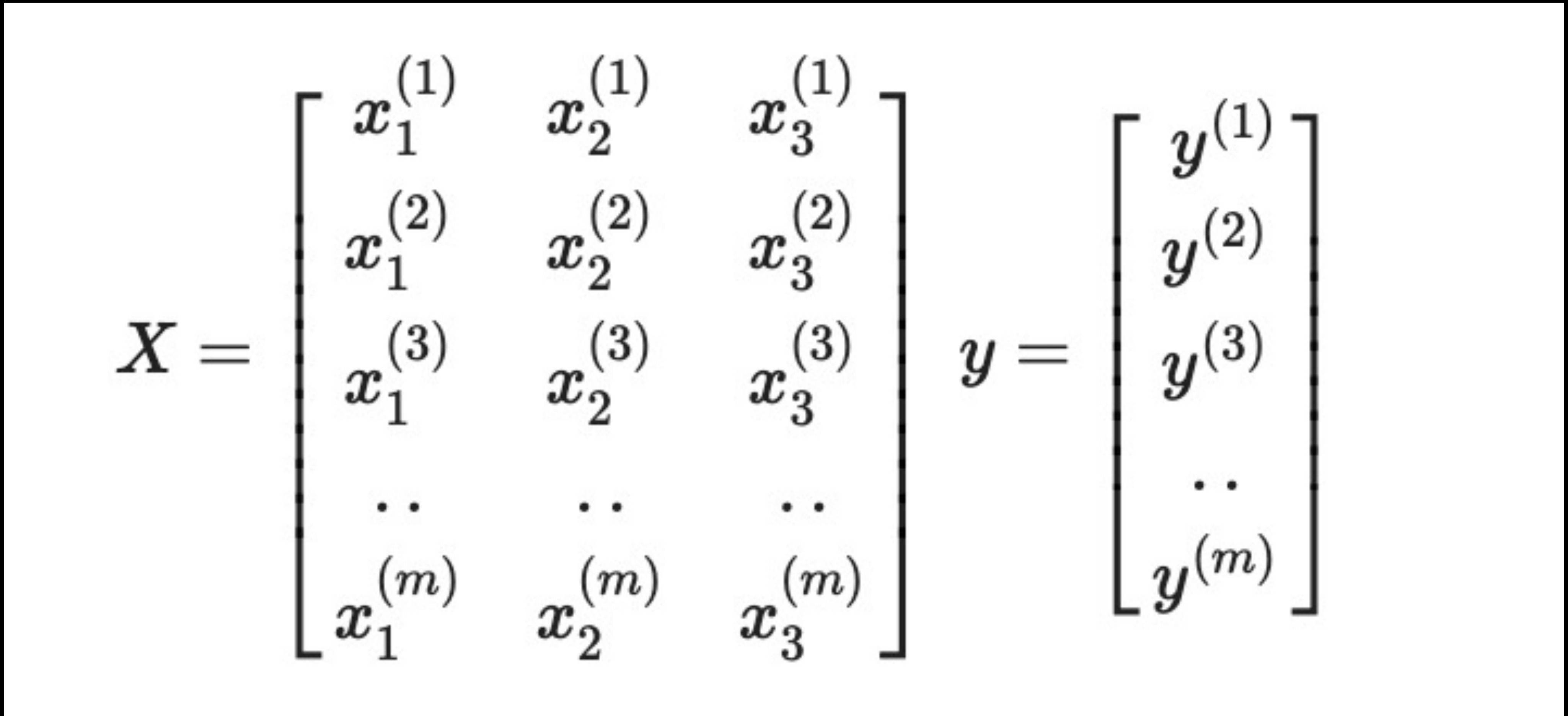
Data collection and preparation

- AI learns from examples, the quality and quantity of examples is of importance and needs to be known
- AI algorithms need the data in a specific format

Dictionary

- Features: independent factors that drive dependent factor (number of sigarets a day drives lung capacity)
- Label: A classification (obese versus not obese)
- Dataframe: tabular table with rows and column containing data
- Preprocess: make data suitable to feed to machine learning algorithm
- Imputation: fill empty spots with a estimated value (for instance mean)
- Observation: the features measurements of one subject

AI algorithms need data in a specific format

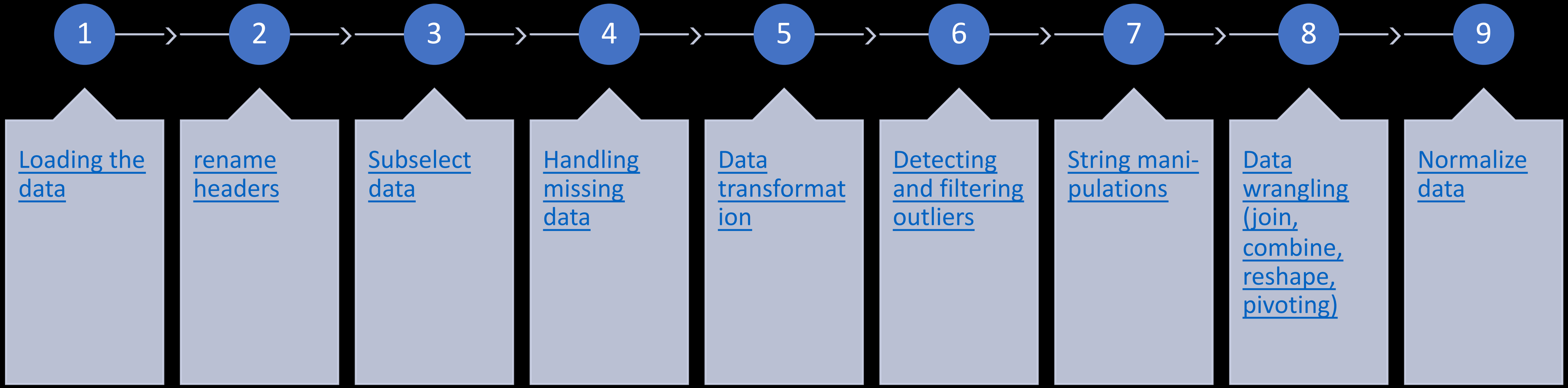


$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} \\ x_1^{(3)} & x_2^{(3)} & x_3^{(3)} \\ .. & .. & .. \\ x_1^{(m)} & x_2^{(m)} & x_3^{(m)} \end{bmatrix} \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ .. \\ y^{(m)} \end{bmatrix}$$

Exploratory data analysis

- Do we know what the headers represent (name, meaning, units, how collected)?
- Is the data in the correct data type?
- Do we need to reshape the data?
- Is imputation needed to handle missing data?
- Are there outliers and how might they effect the model?
- What is the sample size and how might that effect the (statistical) calculations in the model to apply?
- How many features does the dataset contain in relation to the sample size. Is feature selection required?
- Are features correlated or derived from each other? Is covariance occuring?
- How are features distributed? Is transformation to normal distribution needed? Is a resample strategy needed?
- How is the label distributed? Is a resample strategy needed?
- Is normalization of the data needed?

Common preprocessing steps



1. Loading the data
2. rename headers
3. Subselect data
4. Handling missing data
5. Data transformation
6. Detecting and filtering outliers
7. String manipulations
8. Data wrangling (join, combine, reshape, pivoting)
9. Normalize data