CSE 435 – Pattern Recognition

# DATA

Tanvir Azhar
Lecturer, EDU

# Data

- What is a data set?
- Examples
- Types
- Problems

# What is a data set?

the data set

$$\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$$

| Object | Shape | Shape colour | Leaf colour | Class label |
|---|---|---|---|---|
|  | Round | Blue | Blue | 1 |
|  | Square | Green | Blue | 1 |
|  | Square | Green | Green | 2 |
|  | Square | Red | Blue | 1 |
|  | Round | Red | Red | 1 |
|  | Square | Blue | Blue | 2 |
|  | Square | Red | Green | 1 |
|  | Round | Green | Red | 2 |
|  | Round | Blue | Blue | 2 |
|  | Round | Green | Blue | 2 |

the $N$ **objects**

(rows of the data matrix)

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \cdots & & & \\ z_{N1} & z_{N2} & \cdots & z_{Nn} \end{bmatrix} \qquad \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix}$$

the $n$ **features**

(columns of the data matrix)

labels

$y_j \in \Omega$

$|\Omega| = c$ classes

# Examples

The famous iris data set. https://archive.ics.uci.edu/ml/datasets/iris

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

## Iris Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Famous database; from Fisher, 1936

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 150 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 4 | **Date Donated** | 1988-07-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 3539941 |

The famous iris data set. https://archive.ics.uci.edu/ml/datasets/iris

$N = 150$ objects

$Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_{150}\}$

…

$n = 4$ features:

$x_1$ sepal length

$x_2$ sepal width

$x_3$ petal length

$x_4$ petal width

iris.txt - Notepad

File Edit Format View Help

| | | | | |
|---|---|---|---|---|
| 5.1 | 3.3 | 1.7 | 0.5 | 1 |
| 6.1 | 3.0 | 4.6 | 1.4 | 2 |
| 6.3 | 2.7 | 4.9 | 1.8 | 3 |
| 4.8 | 3.1 | 1.6 | 0.2 | 1 |
| 5.0 | 3.4 | 1.5 | 0.2 | 1 |
| 4.5 | 2.3 | 1.3 | 0.3 | 1 |
| 6.0 | 2.2 | 5.0 | 1.5 | 3 |
| 5.4 | 3.4 | 1.7 | 0.2 | 1 |
| 7.7 | 3.8 | 6.7 | 2.2 | 3 |
| 5.1 | 2.5 | 3.0 | 1.1 | 2 |
| 6.7 | 3.3 | 5.7 | 2.5 | 3 |
| 5.5 | 2.6 | 4.4 | 1.2 | 2 |
| 4.8 | 3.4 | 1.9 | 0.2 | 1 |

…

class label

The 4 features

$c = 3$ classes:

$\omega_1$ Iris Setosa

$\omega_2$ Iris Versicolour

$\omega_3$ Iris Virginica

Face recognition data set:
(toy)

$c$ = 2 ***classes***: Richard Armitage and Hugh Jackman

Each image is an *object* to classify

All images are scaled to the same dimension

300 × 325 pixels = 97500 pixels x 3 colours = $n$ = 292500 ***features***

$N$ = 13 objects

Face recognition data set:
(toy)

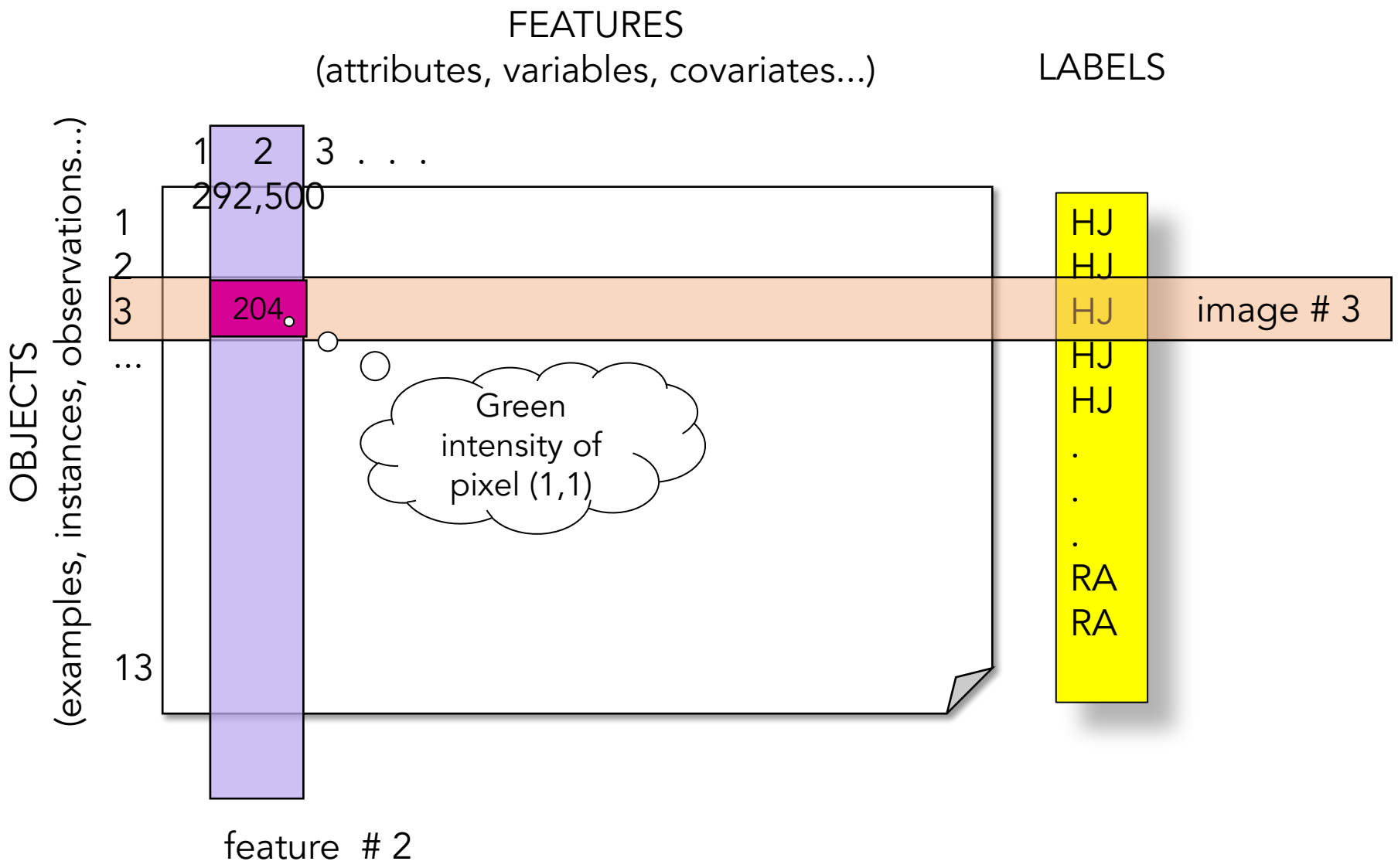
All images are scaled to the same dimension

300 × 325 pixels = 97500 pixels x 3 colours = 292500 *features*

Alternative/complementary *features* may be obtained by:

- finding important facial marks such as irises, eyebrows, nose, corners of mouth, etc., and measuring new, geometric features from these
- a pre-processing step extracting various colour and morphological features from the image

Face recognition data set:
(toy)



https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Thus far, $\mathbf{Z}$ is a **tabular** data set (rows = objects and columns = features)

There are many other types of labelled and unlabelled data sets, for example:

- images (e.g., emotion recognition)
- videos (e.g., action recognition)
- time series (e.g., note sequences, coordinates of handwriting samples)
- continuous streaming data (e.g., speech recognition, earthquake prediction)
- text (e.g., automatic translation, sentiment recognition, spam detection)

*Extracting tabular data from these is not easy!*

# ImageNet


- Over 14 million URLs of images have been hand-annotated by ImageNet to indicate what objects are pictured.
- In at least one million of the images, bounding boxes are also provided.
- ImageNet contains over 20 thousand categories; a typical category, such as "balloon" or "strawberry", contains several hundred images.
- Since 2010, the ImageNet project runs an annual software contest, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
- A dramatic 2012 breakthrough in solving the ImageNet Challenge is widely considered to be the beginning of the DEEP LEARNING revolution.

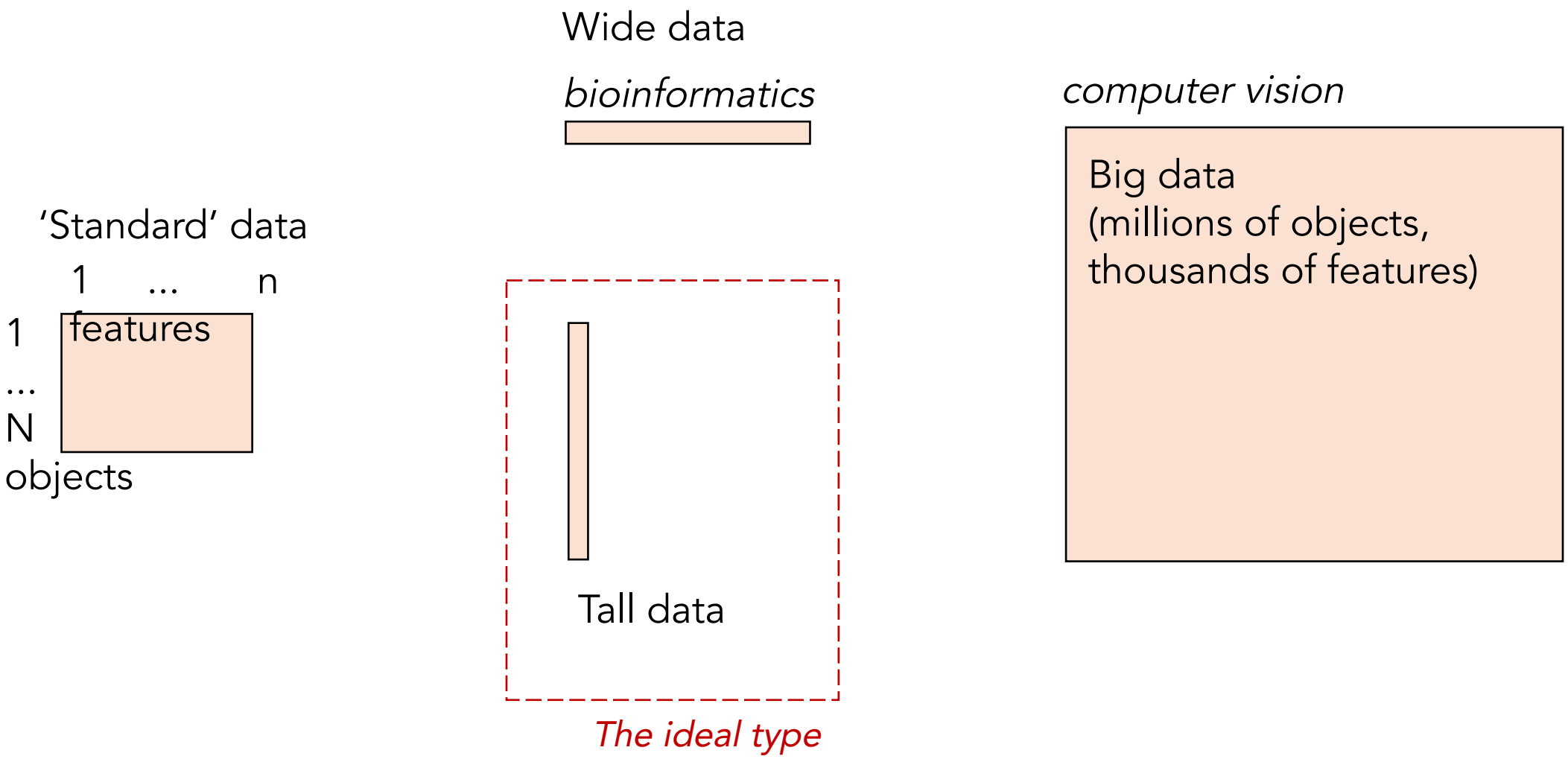
For ImageNet data set:

$N = 14{,}000{,}000$ objects (or more)

$c = 20{,}000+$ classes (trimmed to 1,000 classes for the competition)

$n = ???$ features (different image dimensions)

# Types of tabular data sets



Wide data

*bioinformatics*

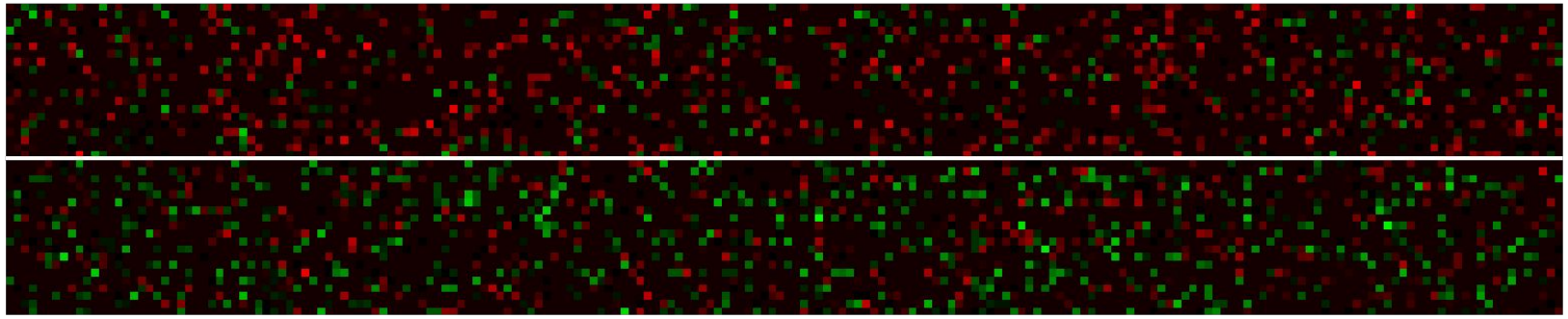
Gene expression data analysis

Wide data set

$n \gg N$

Genes (features)

Patients (objects)

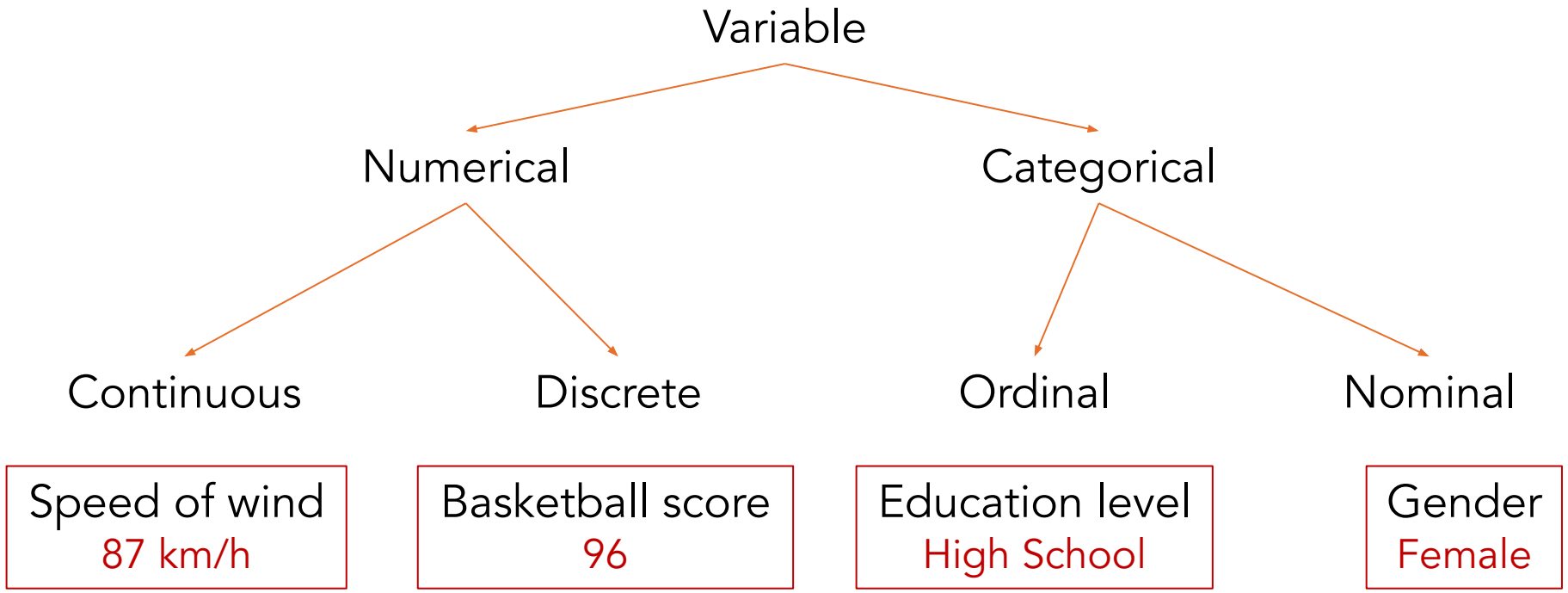


class 1

class 2

# Problems

- Mixed features: numerical, categorical




*Euclidean distance will not work*

Solutions to the problem of mixed data:

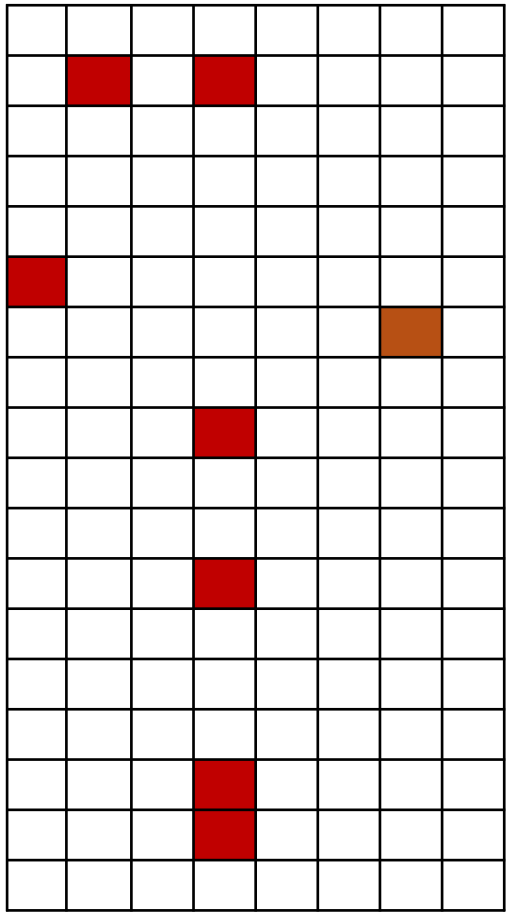
- Quantise numerical variables and use all as categorical. For example, replace $x_2$ with $y_2$: if $x_2 > 40, y_2 = Yes$, else $y_2 = No$.
- Replace a categorical variable with k binary variables, one for each value of the category. For example, replace "Fruit" with 3 binary variables:

| Fruit |
|---|
| apple |
| apple |
| pear |
| plum |
| apple |
| plum |

$\longrightarrow$

| Apple | Pear | Plum |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 0 | 1 |

# Missing values: measurement errors, sensor failure, unavailable information



Solutions to the problem of missing data:

- Remove columns
- Remove rows
- Remove both
- Replace with mean or median of the column
- Use correlation with another feature to predict a missing value

# Unavailable labels: when labelling is expensive / invasive / destructive.

A large volume of data is available but labelling or annotating are is not feasible

Solutions to the problem of unavailable labels:

- Use partial labelling (semi-supervised learning).
- Ask for labels only for “key” objects (active learning).
- Pour more money into the problem ☺️