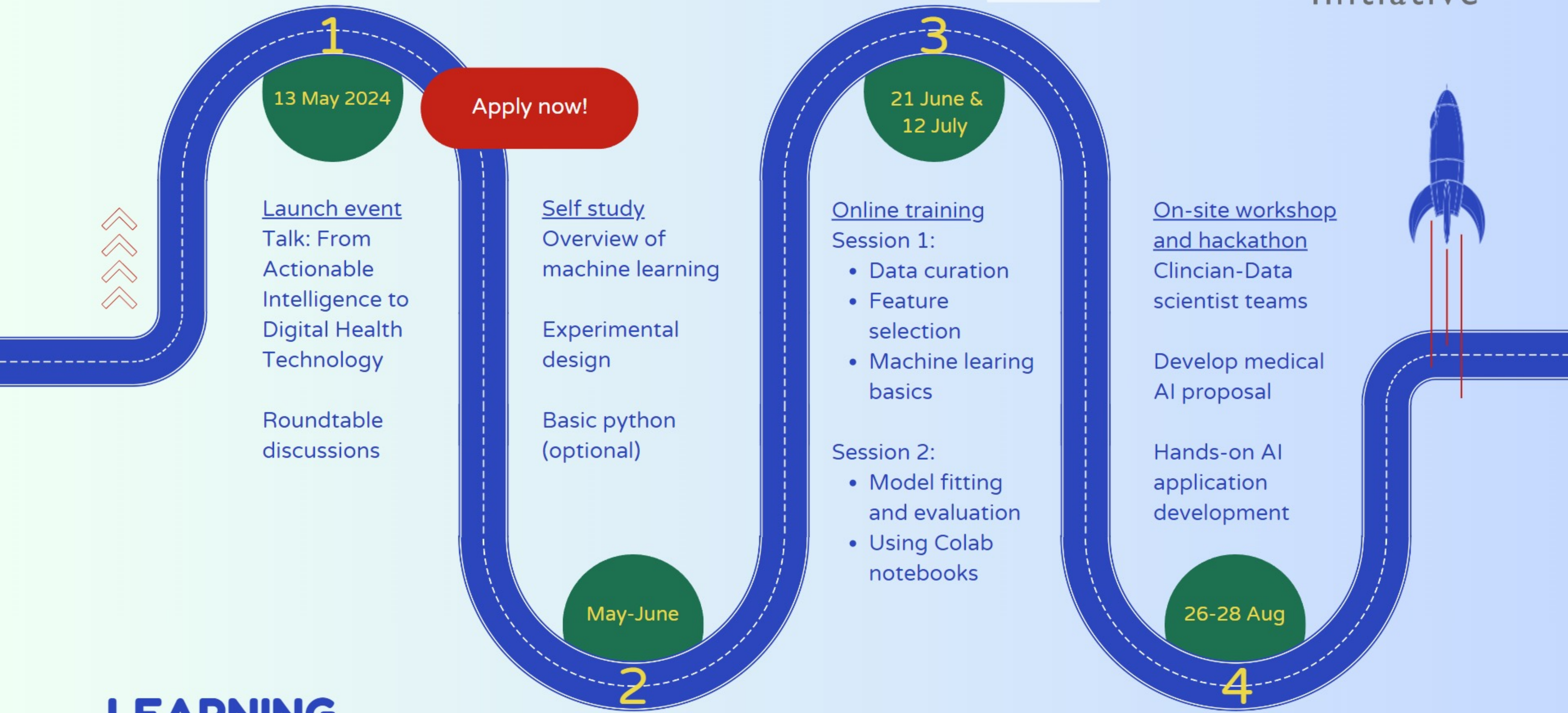
ThAIMed Initiative
Session 1

21st June, 2024

Igor Tatarnikov



1

13 May 2024

Apply now!

Launch event

Talk: From Actionable Intelligence to Digital Health Technology

Roundtable discussions

2

May-June

Self study

Overview of machine learning

Experimental design

Basic python (optional)

3

21 June & 12 July

Online training

Session 1:

- Data curation
- Feature selection
- Machine learing basics

Session 2:

- Model fitting and evaluation
- Using Colab notebooks

4

26-28 Aug

On-site workshop and hackathon

Clincian-Data scientist teams

Develop medical AI proposal

Hands-on AI application development

LEARNING JOURNEY

Outline/Plan

- Introduction to data types, structure, and representation
- Data quality and data cleaning
- Data transformation
- Feature selection/reduction
- Introduction to machine learning

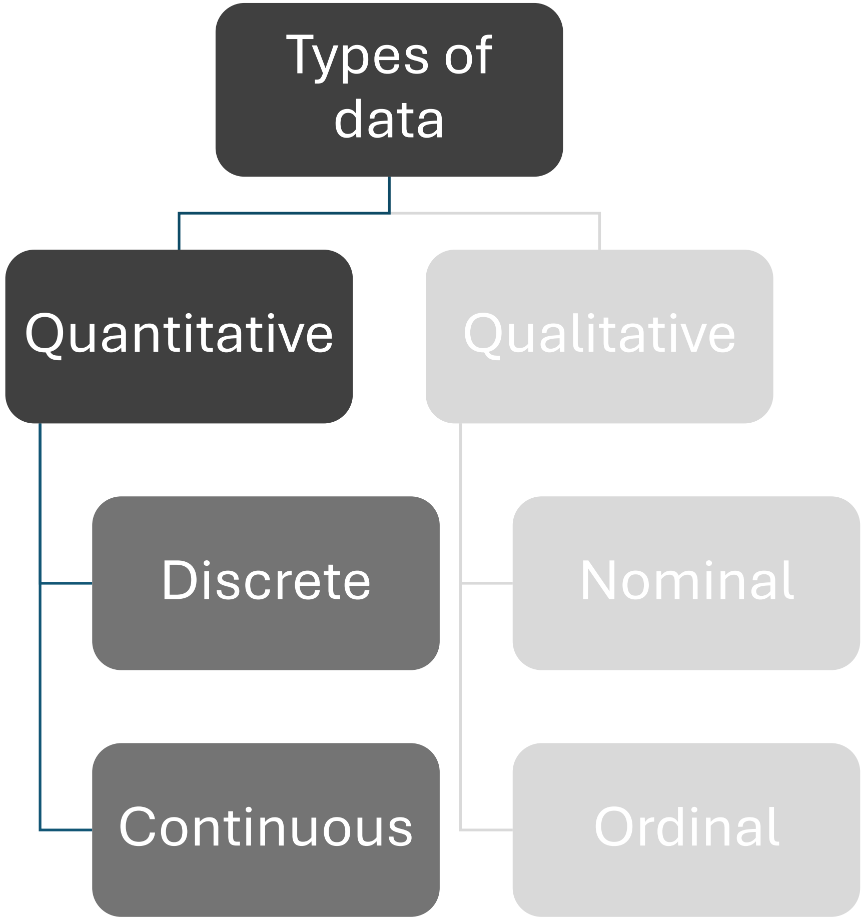
Data and modeling

statista

Sad loser, CC BY-SA 4.0, via Wikimedia Commons

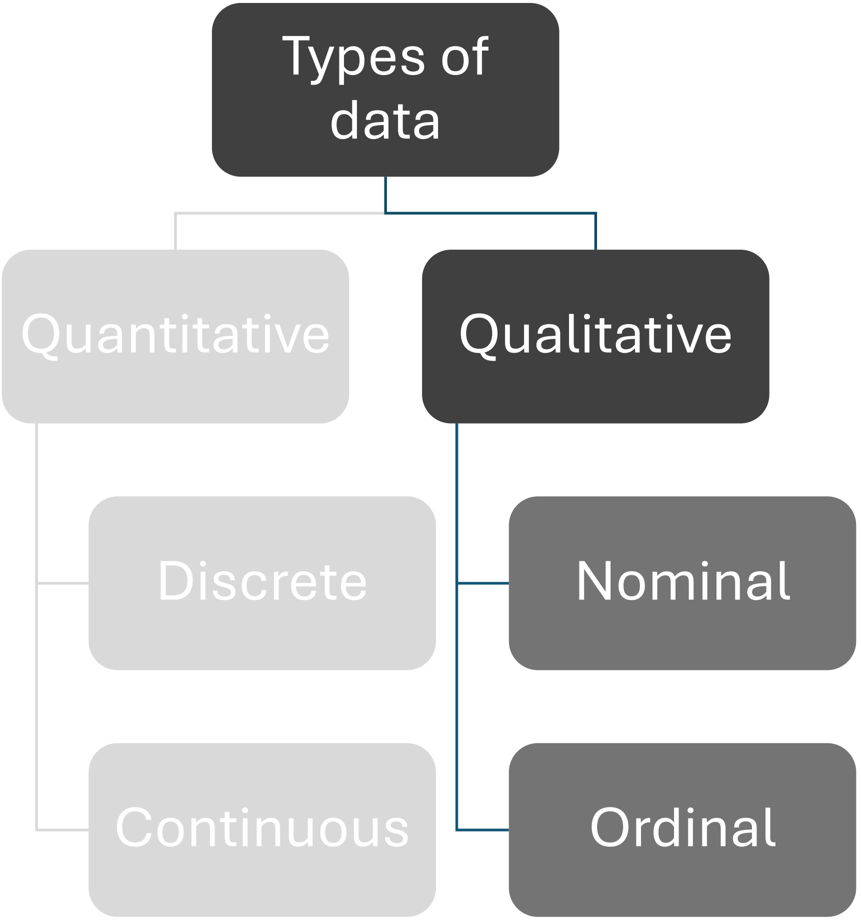
Types of data

- Quantitative data can be numerically measured, are inherently ordered and arithmetical (+, -, ÷, …, valid)
  - Discrete data – essentially integers, they are countable.
    - E.g., number of students in a class, number of days since hospital discharge
  - Continuous data – real numbers, uncountable.
    - E.g., market price, weight in kg, temperature



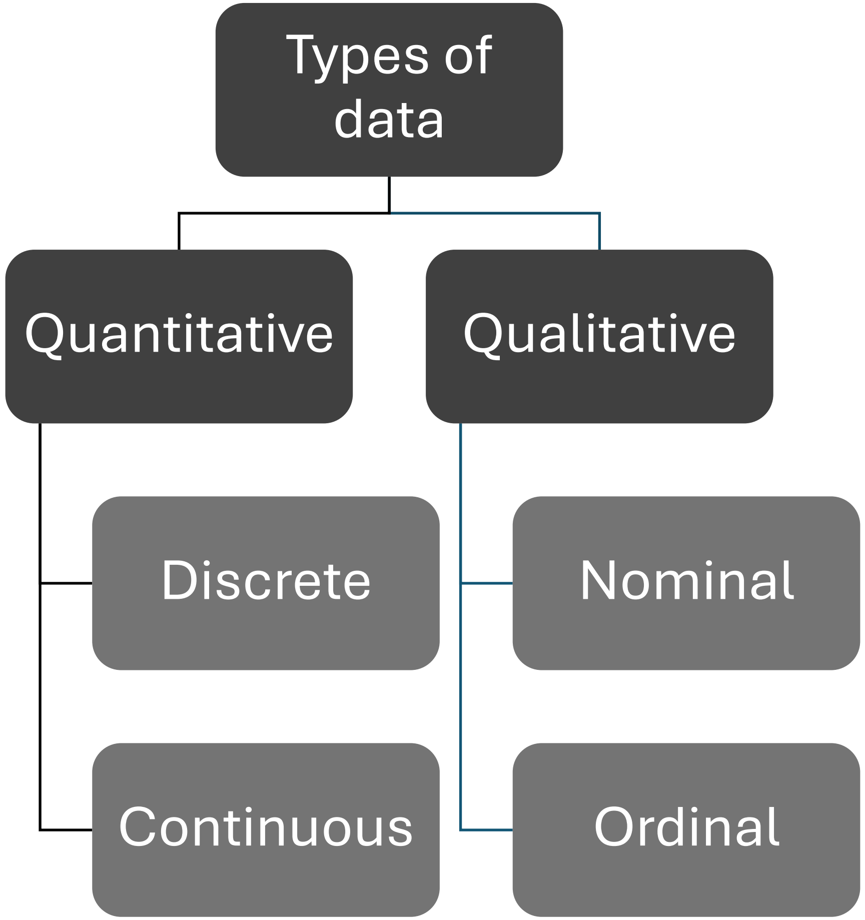
Types of data

- Qualitative (categorical) data cannot be numerically measured
  - Nominal data – nameable labels without a prescribed order
    - E.g., citizenship, diagnosis, species
  - Ordinal data – labels with a prescribed order
    - Comparison operators apply
    - E.g., star ratings, letter grades, economic status, education level
- Arithmetic operations cannot be applied to qualitative data (+, -, ×, ÷)



Types of data

- Some concepts are represented with arbitrary types
  - Gender may be nominal (i.e., ‘female’, ‘male’, ‘nonbinary’, …) but is often represented as numbers (i.e., 0, 1, 2).
  - Grades may be ordinal (i.e., A+, A, A-, …) but are often represented nominally (‘A+’, ‘A’, ‘A-’, …)
  - What are the consequences of this?
- Some operations change the data type
  - E.g., average([4, 3, 2, 1]) = 2.5



Data representation

- In the end everything ends up with a numerical representation
- How to deal with nominal data?
    - One-hot encoding – a technique used to convert categorical variables into a binary (0 and 1) format, making it suitable for use in machine learning algorithms.
    - Each unique category value becomes a new column. Each row has a binary value (0 or 1) indicating the presence of the category in that row.

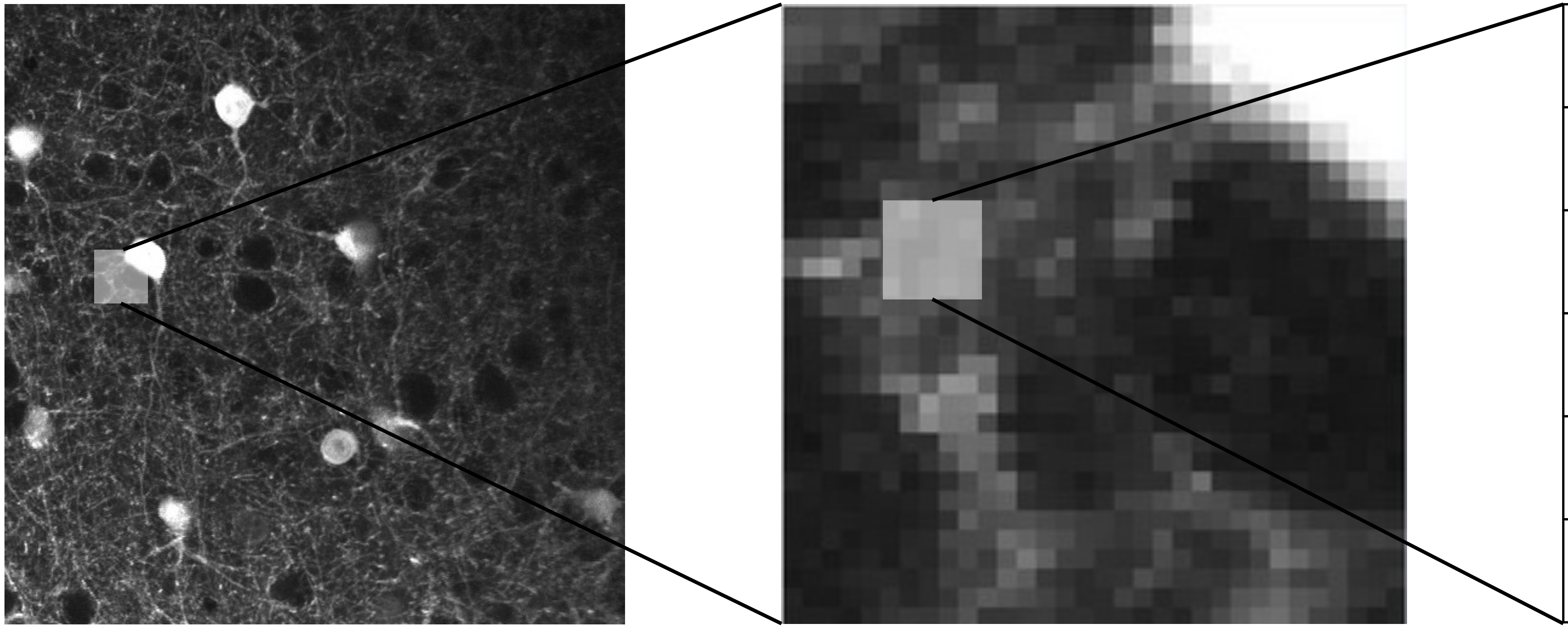
Color	Color_Red	Color_Blue	Color_Green
Red	1	0	0
Blue	0	1	0
Green	0	0	1
Red	1	0	0

Data representation

- Bag of words model – A text representation technique that converts text data into numerical features based on word frequency.
- Example:
    - Document 1: "Machine learning is fun”
    - Document 2: "Learning machine learning”
    - ['is', 'learning', 'machine', 'fun’]

Document	is	learning	machine	fun
Document 1	1	1	1	1
Document 2		2	1	

Data representation



78	64	71	57	43	44
86	93	69	60	48	46
75	64	57	59	55	62
70	65	69	85	56	54
54	68	88	63	46	43
51	51	74	66	43	35

Structure of data

1. Structured data – often tabular or relational. There are many fields and variables with defined meanings and relations.
   - Patient bloodwork
2. Unstructured data – raw and acquired in a variety of contexts.
   - Images, videos, audio, and text
3. Semi-structured data – combination of the above, often with relatively rigid metadata
   - List of scientific publications with their text and associated metadata

Data quality

1. Accuracy – correctly represents the target
2. Precision – the resolution is sufficient for analysis
3. Completeness – available with no missing values
4. Consistency – at an equivalent state of processing and combinable
5. Timeliness – not waiting for additional data
6. Believability – has a subjective effect of trustworthiness
7. Interpretability – understandable

Types of dirty data

- Inconsistent data types
  - E.g., {M, ‘male’, 1}
- Inconsistent data values
  - E.g., 0.1 dL ⇔ 10 CC
- Missing data
- Data outliers
- Recording errors and noise
  - E.g., incorrect labels

Inconsistent types & values

- Checking for inconsistent types can usually be automated
  - Check if a value is an integer number, text, or real number and flag/fix inconsistencies
- Checking for inconsistent values is harder automate
  - Can use clustering to check for large differences in values (dL vs CC)
  - However, if values are too close then it’s difficult (€ EUR, $ USD)
- An understanding of the data set is often needed
  - Domain experts

Missing data

- Usually obvious and finding it can be automated
  - Missing fields:

June	45	NaN	100	119	282.0

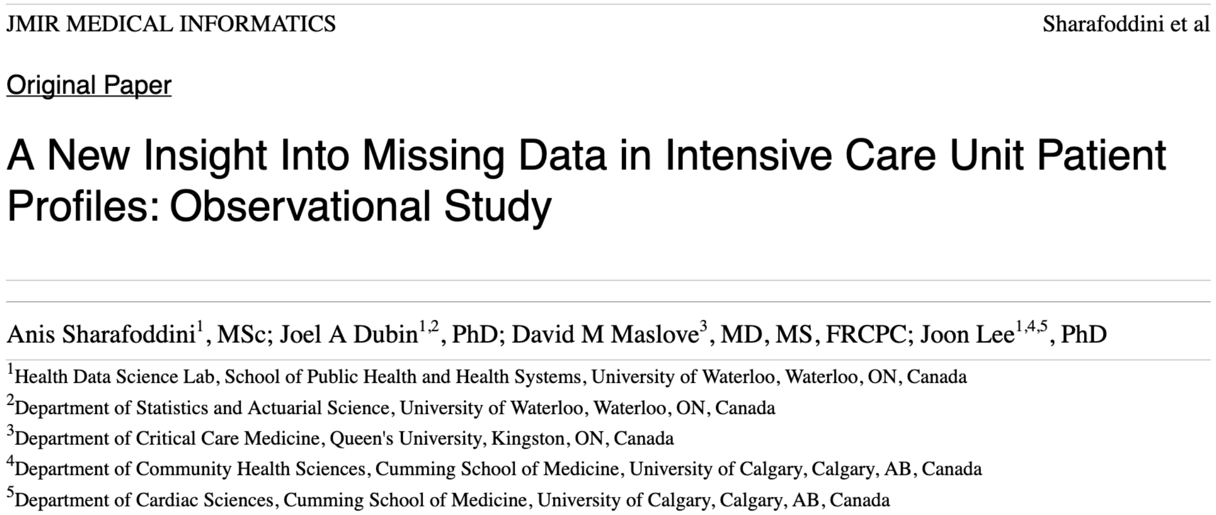
  - Missing entries:

June	45	86	100	119	282.0
August	47	82	100	118	294.8

- Sometimes recorded as an invalid value
  - E.g., ‘age’: ‘-1’
  - Treated as missing data

Missing data

- Some types of missing data are impossible to track and recover
  - E.g., patient deleted from database
- Missingness can be useful!
  - If a patient is discharged without intubation they were not intubated
  - This data point can be used as part of the negative class



JMIR MEDICAL INFORMATICS Sharafoddini et al

Original Paper

A New Insight Into Missing Data in Intensive Care Unit Patient Profiles: Observational Study

Anis Sharafoddini[1], MSc; Joel A Dubin[1,2], PhD; David M Maslove[3], MD, MS, FRCPC; Joon Lee[1,4,5], PhD

[1] Health Data Science Lab, School of Public Health and Health Systems, University of Waterloo, Waterloo, ON, Canada
[2] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada
[3] Department of Critical Care Medicine, Queen's University, Kingston, ON, Canada
[4] Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[5] Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

Resolving missing data

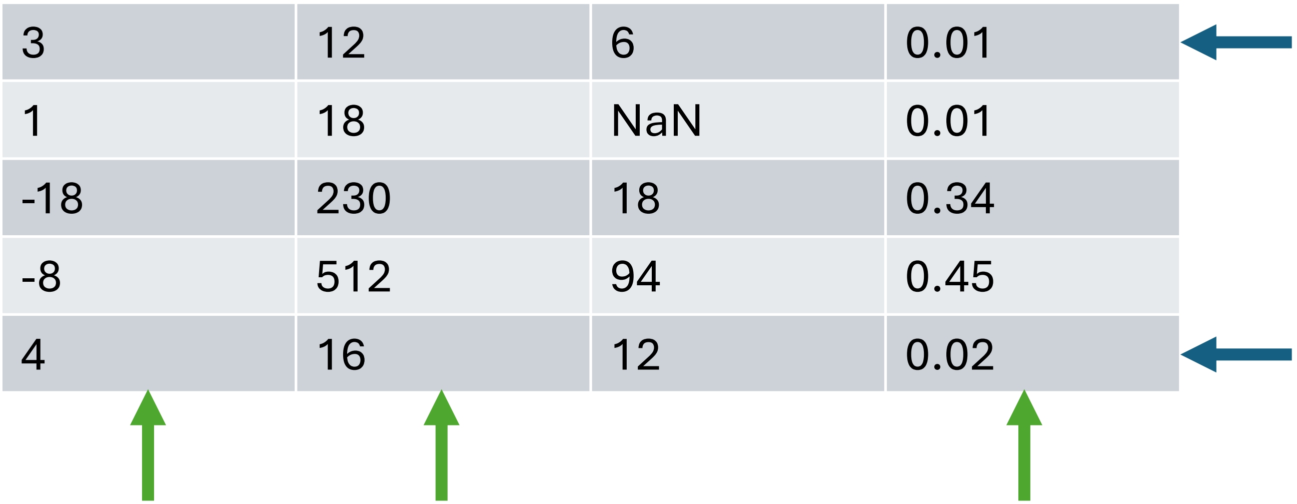
- Ignore the entry entirely
  - For classification this is often done if class label is missing
  - This loses data
- Regenerate data manually
  - Can be very time consuming
  - Infeasible for large data set

Resolving missing data

- Generate the missing data automatically with
  - A constant
  - Attribute mean
  - Attribute mean for all samples belonging to the class
  - Most likely value based on other inference methods
  - Interpolation and regression
  - Imputation

Resolving missing data

- Imputation – Use the *k* points of existing data most similar to the missing data, replace the missing data with the average (or other combination) of those points.

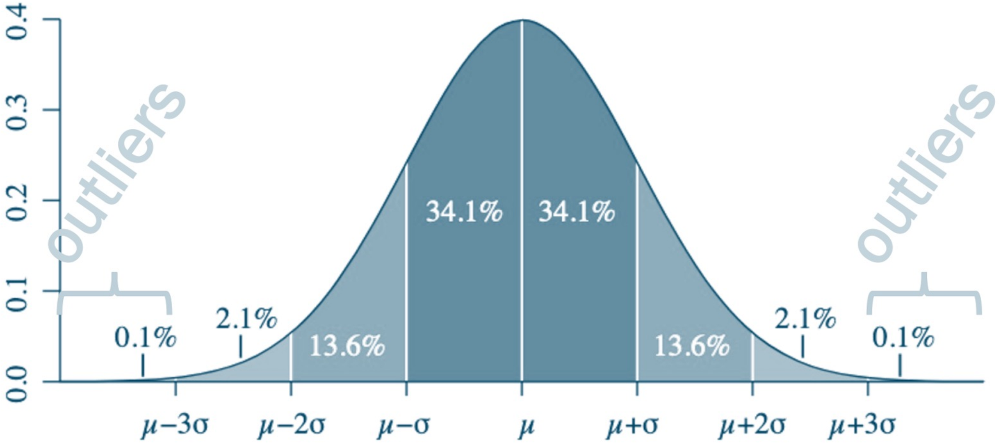


3	12	6	0.01
1	18	NaN	0.01
-18	230	18	0.34
-8	512	94	0.45
4	16	12	0.02

$$\frac{6+12}{2}=9$$

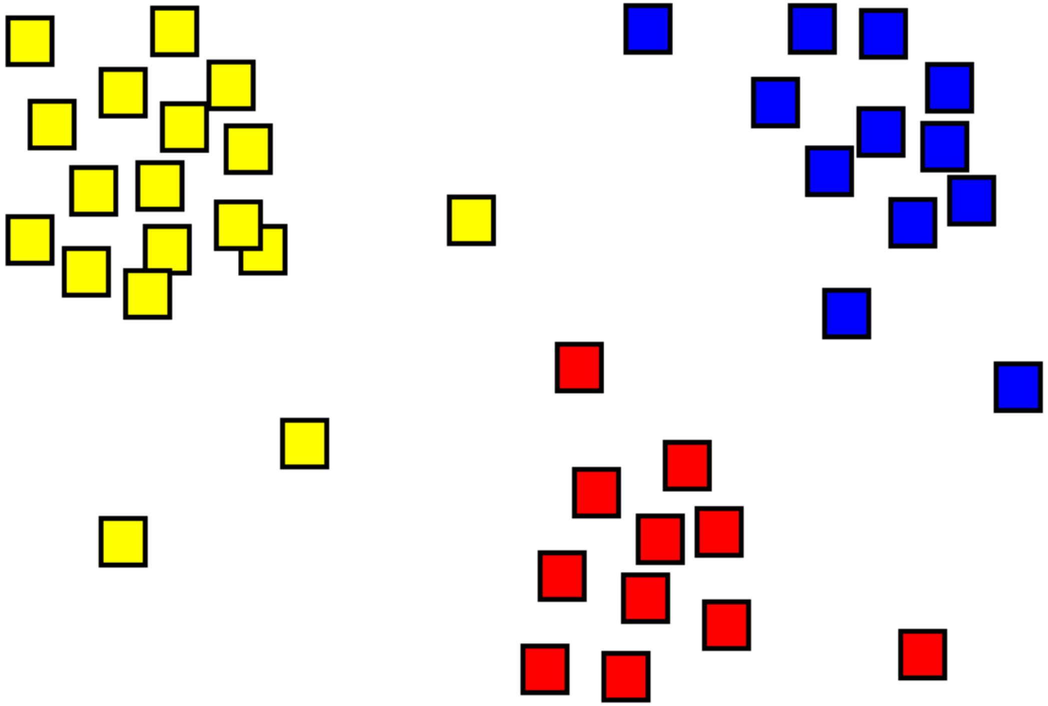
Outliers

- Datapoint that differs *significantly* from the others
  - Often caused by variability, novel sources of data, or simple error
  - No single mathematical definition, they are often defined by their relationship to a distribution mean
  - If we know the distribution *a priori* we can often assume outliers occur more than 2 standard deviations away from the mean.



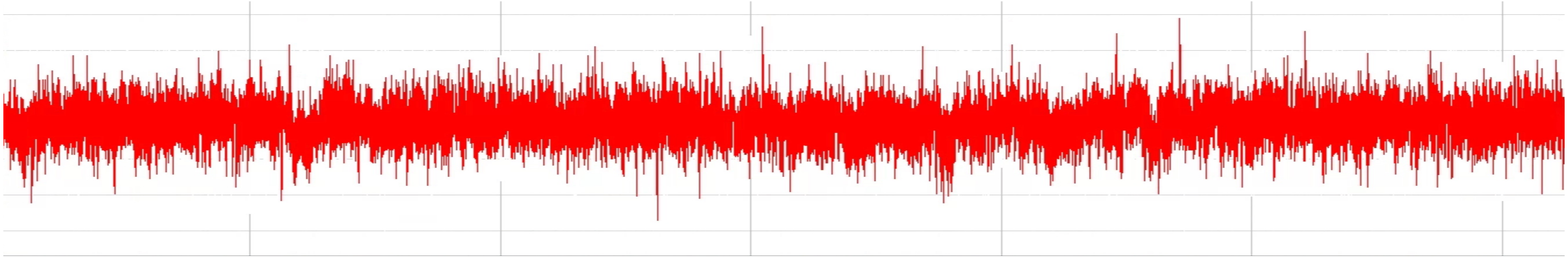
Outliers

- Can also use clustering to detect outliers
    - Anything that's not part of a cluster may be an outlier
- Note that not all clustering algorithms allow for data points to remain unclustered



Noise

- Channel noise
  - Recording equipment can have different sensitivities to noise (e.g., EEG)
  - Can filter out using various filtering or smoothing methods
- Human noise
  - Manual labelling will have some inherent error and bias that can’t be avoided
  - Can quantify this error by using multiple annotators and then calculating their inter-annotator agreement



Cleaning text data

- Remove links, symbols, emojis, anything that’s not directly relevant to the analysis
- Translate all text into the language you’re working in
- Lowercase all text (optional)
- Perform spelling correction (optional)
- Remove unnecessary blank text between words (optional)

NHANES dataset

- National Health and Nutrition Examination Survey
- Done by the Centers for Disease Control and Prevention in the USA
- Combination of interview and physical examination
- Samples approximately 5,000 people each year
- We’ll be using data from the 2017-2018 edition

Demo

Data transformation

- Normalisation – scales the values of an attribute to fall within a specified range
- Min-max normalisation – linearly scale the attribute to a new range from new min to new max

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)}(newMax - newMin) + newMin$$

  - Suppose that the income range is \$10,000 to \$96,000 and we want to normalise such that the new min is -1 and new max is 1. Then \$43,000 is mapped to:

$$\frac{43{,}000 - 10{,}000}{96{,}000 - 10{,}000}(1 - (-1)) + (-1) = -0.2326$$

Data transformation

- Z score normalisation – scales the values such that the data follows the Standard Normal distribution

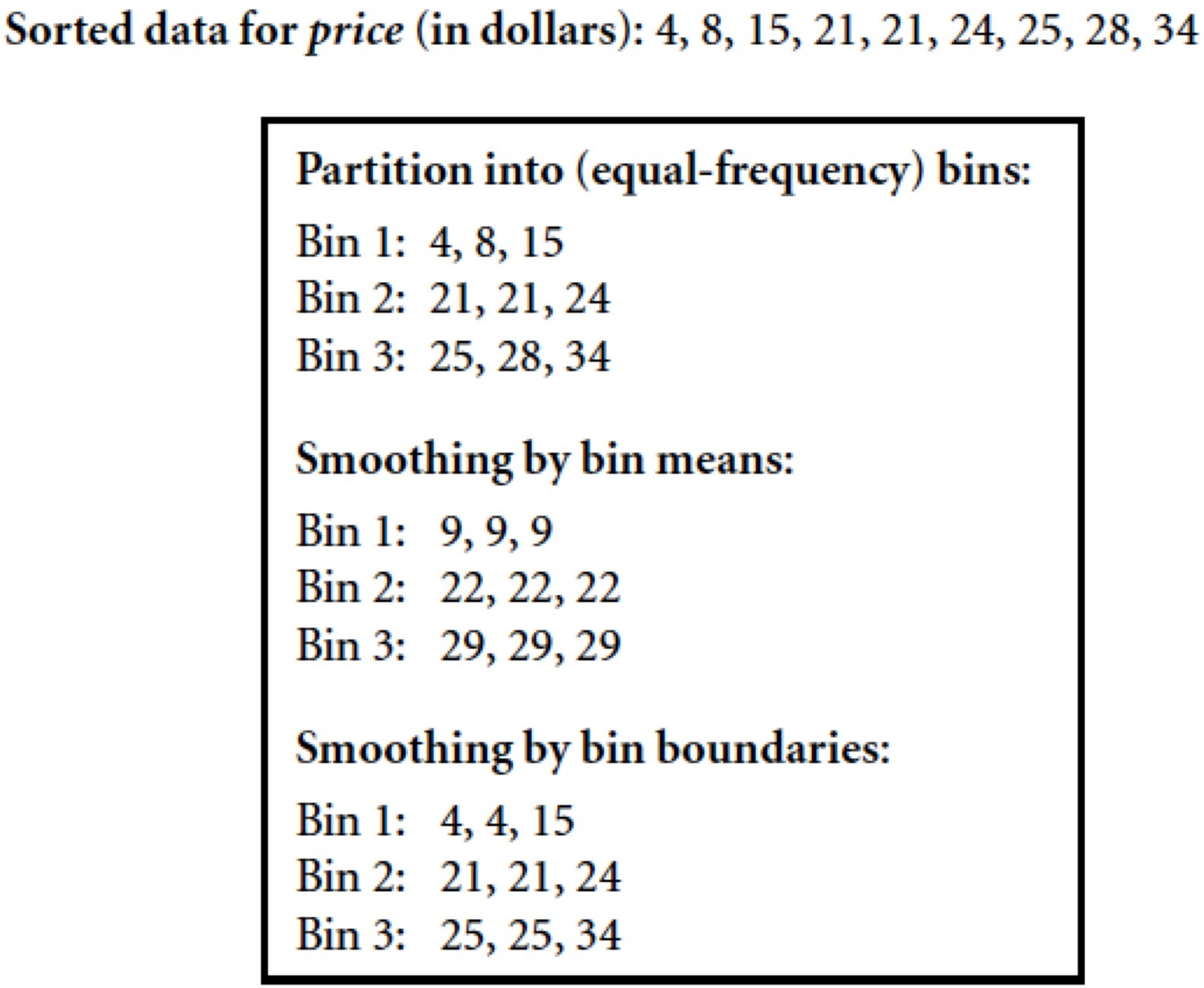
$$x' = \frac{x - \mu}{\sigma}$$

  - E.g., let μ = \$54,000, σ = \$16,000 then \$43,000 is mapped onto:

$$\frac{43{,}000 - 54{,}000}{16{,}000} = -0.6875$$

Discretisation

- Make continuous data discrete – divide the range of a continuous numeric attribute into intervals
- Actual data values can then be replaced using one of the following methods:
    - Bin mean – each value is replaced by the mean value of the bin
    - Bin median – each value is replaced by the median value of the bin
    - Bin boundaries – each value is replaced by either the minimum or maximum of the bin, depending on which is closer



Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

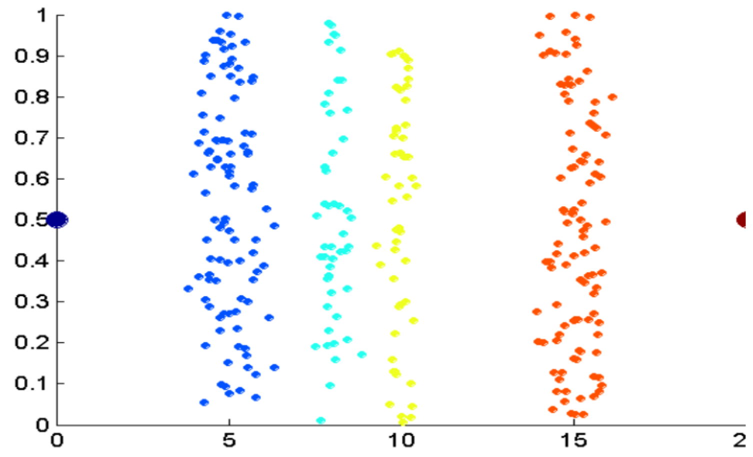
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

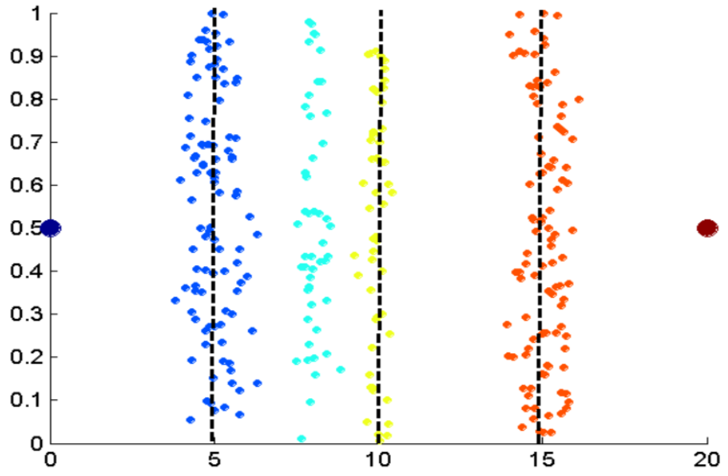
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Discretisation

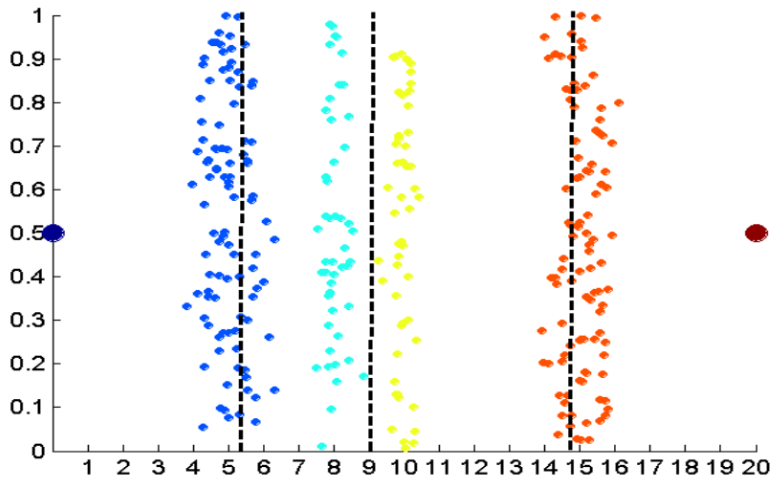
- Equal width – divides the range into *N* intervals of equal size
  - Given min(Data) and max(Data) the width is: W = (max(Data) – min(Data)) / *N*
  - Outliers dominate
  - Some bins may be empty
- Equal depth
  - Divides the range into *N* intervals each containing approximately the same number of samples
- Cluster based
  - Use a clustering algorithm to determine the bins (e.g., K-means clustering)



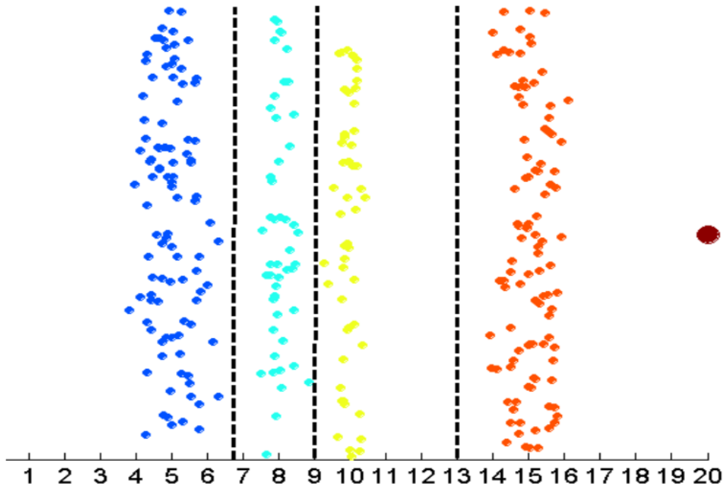
Data



Equal width (binning)



Equal depth (binning)



K-means clustering

Feature selection

- The process of selecting a subset of relevant features (variables, predictors) for use in model construction.
  - Can improve model performance.
  - Can reduce overfitting.
  - Decreases computation time.
  - Can enhances model interpretability.

Feature selection

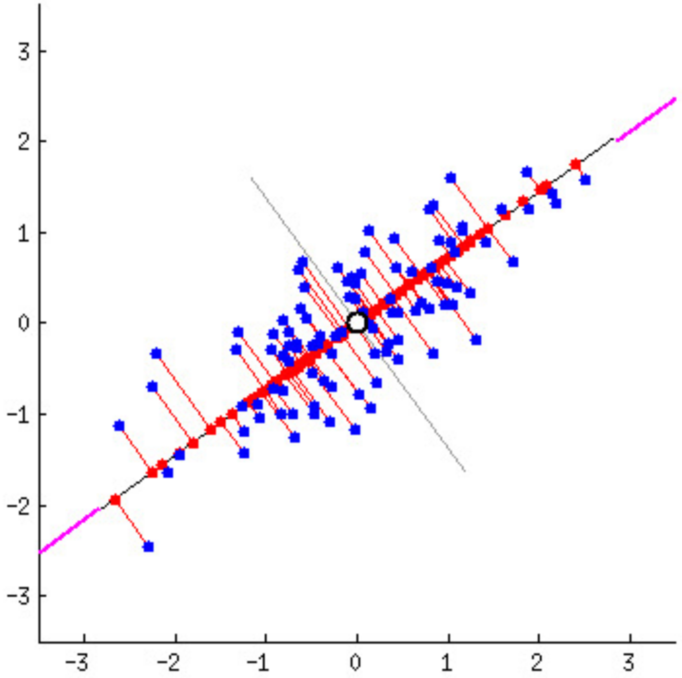
- Filter methods – Use statistical measurements to evaluate importance of features
  - Correlation between continuous features
  - χ^2 test for categorical features
  - Mutual information
- Embedded methods – perform feature selection using model performance
  - Train models based on an individual feature
  - Pick the feature that performs best, and then train models that use this feature in combination with all other features. Repeat until performance is stable or desired number of features reached.

Dimensionality reduction

- Simplifies models and reduces overfitting.
- Decreases computational cost and storage space.
- Helps in visualising high-dimensional data.

Principal component analysis (PCA)

- A statistical procedure that transforms the data into a set of components that capture the maximum variance in the data.
- Steps:
  - Standardise the data
  - Compute statistics to identify the principal component
  - Select the k best components
  - Transform the data into this new space



Demo

Introduction to machine learning

- Using algorithms and statistical models to enable computers to perform tasks without explicit instructions, relying on patterns and inference.
- Key Components:
  - **Data:** The foundation for training models.
  - **Algorithms:** Methods used to learn from data.
  - **Model:** The output of the learning process that can make predictions or decisions.

Supervised vs Unsupervised

- **Supervised Learning:**
  - **Definition:** Learning from labelled data where the output is known.
  - **Examples:** Classification and regression.
- **Unsupervised Learning:**
  - **Definition:** Learning from unlabelled data where the output is unknown.
  - **Examples:** Clustering

Classification vs Regression

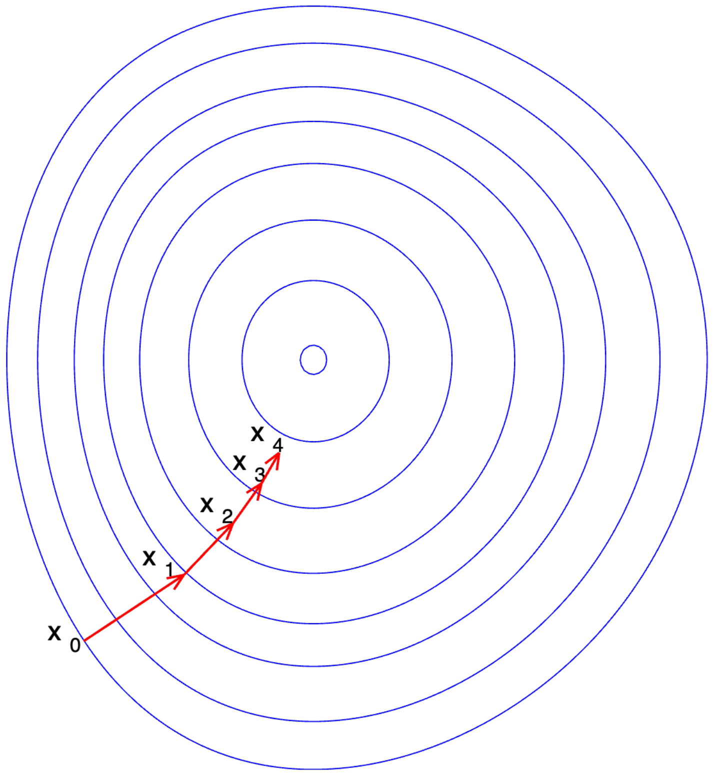
- **Classification:**
  - **Goal:** Predict categorical class labels.
  - **Examples:** image recognition.
- **Regression:**
  - **Goal:** Predict continuous values.
  - **Examples:** recovery time post-surgical intervention.

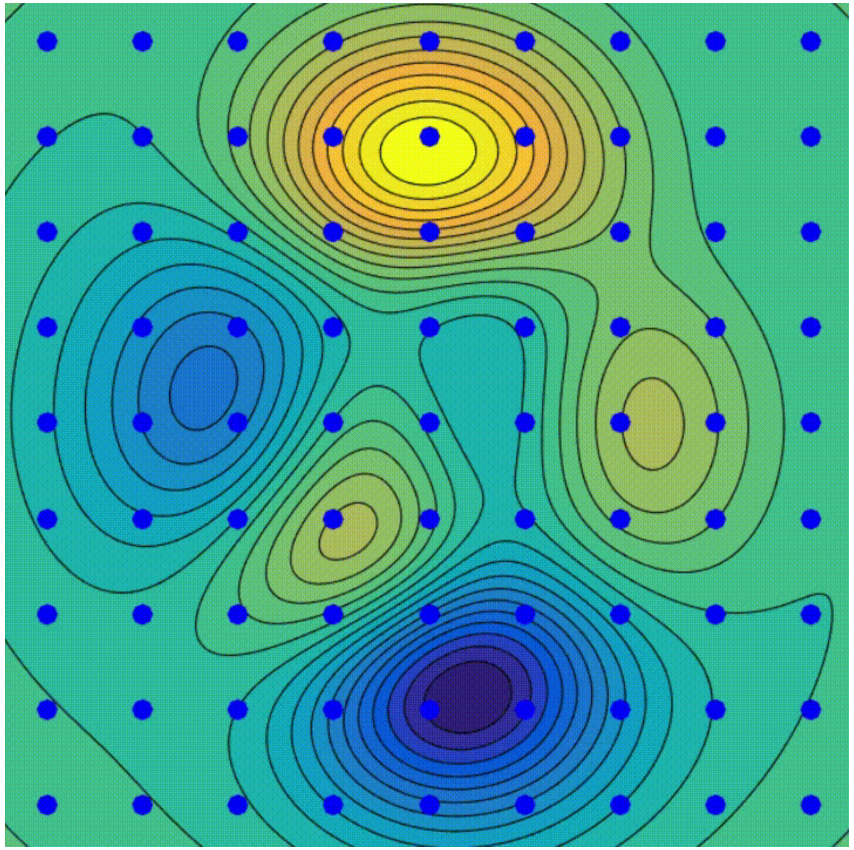
Loss functions

- Measures how well the model's predictions match the actual data.
- Common examples:
    - Mean Squared Error (MSE) for regression
    - Cross-entropy for classification.

Gradient descent

- An optimisation algorithm used to minimise the loss function.





Overall summary

- Data can be quantitative and qualitative
  - Nominal vs ordinal
  - Discrete vs continuous
- Data quality is key for a good model
- Standardisation and normalisation rescale the values to either fit a standard normal distribution, or a new preset range
- PCA can be used as a dimensionality reduction tool
- Machine learning uses data to fit a model to perform a task

Next session!

- Model Selection and Fitting:
  - Choose models for classification, regression, and clustering.
  - Understand overfitting vs. underfitting.
- Evaluation Metrics
  - Important choices for choosing best model in clinical practice.
- Clinical considerations for ML research on health data

Regression

- Linear regression
  - **Definition:** A linear approach to modelling the relationship between a dependent variable and one or more independent variables.
  - **Equation:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_i x_i$
  - **Example:** Predicting house prices based on size.
- Logistic regression
  - **Definition:** A regression model used for binary classification tasks.
  - **Equation:** $\sigma(z) = \frac{1}{1+e^{-z}}$ where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_i x_i$
  - **Example:** Predicting if an email is spam or not.

Clustering algorithms

- **K-Nearest Neighbours (KNN):**
  - **Definition:** Classifies a data point based on the majority class among its k-nearest neighbours.
- **K-Means:**
  - **Definition:** Partitions data into k clusters by minimising the variance within each cluster.
- **DBSCAN:**
  - **Definition:** Density-based clustering that finds core samples of high density and expands clusters from them.
- **Examples**: Patient segmentation, identifying clusters in spatial data

Decision trees

- A tree-like model used for classification that splits the data into subsets based on feature values.
- Multiple decision trees can be combined into a random forest classifier

Support vector machines (SVMs)

- A supervised learning model that finds the optimal hyperplane to separate classes in the feature space.
- Kernel Trick: Allows SVM to create non-linear boundaries by mapping input features into higher-dimensional space.

Neural networks

- Composed of layers of interconnected nodes (neurons) that learn representations of the data through training.
- Common Types:
    - Feedforward Neural Networks
    - Convolutional Neural Networks (CNNs)
    - Recurrent Neural Networks (RNNs)