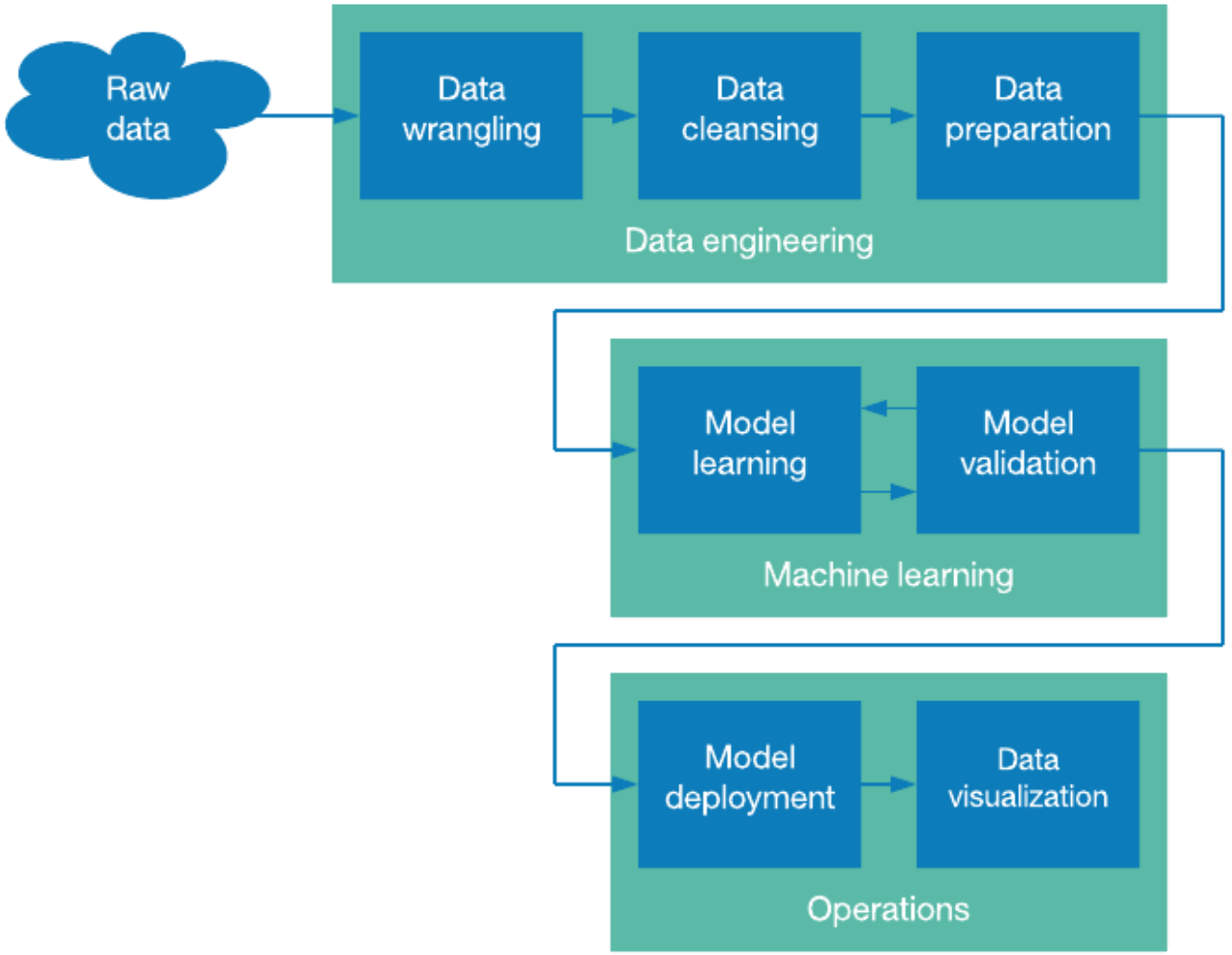
Introduction to Data

20_KIN2 – Artificial Intelligence and Machine Learning

Lecture Contents

- The Data Science pipeline
- Taxonomy of data
- Data representation
- Graphical exploration tools
- Summarization
- Metrics for data

The Data Science Pipeline



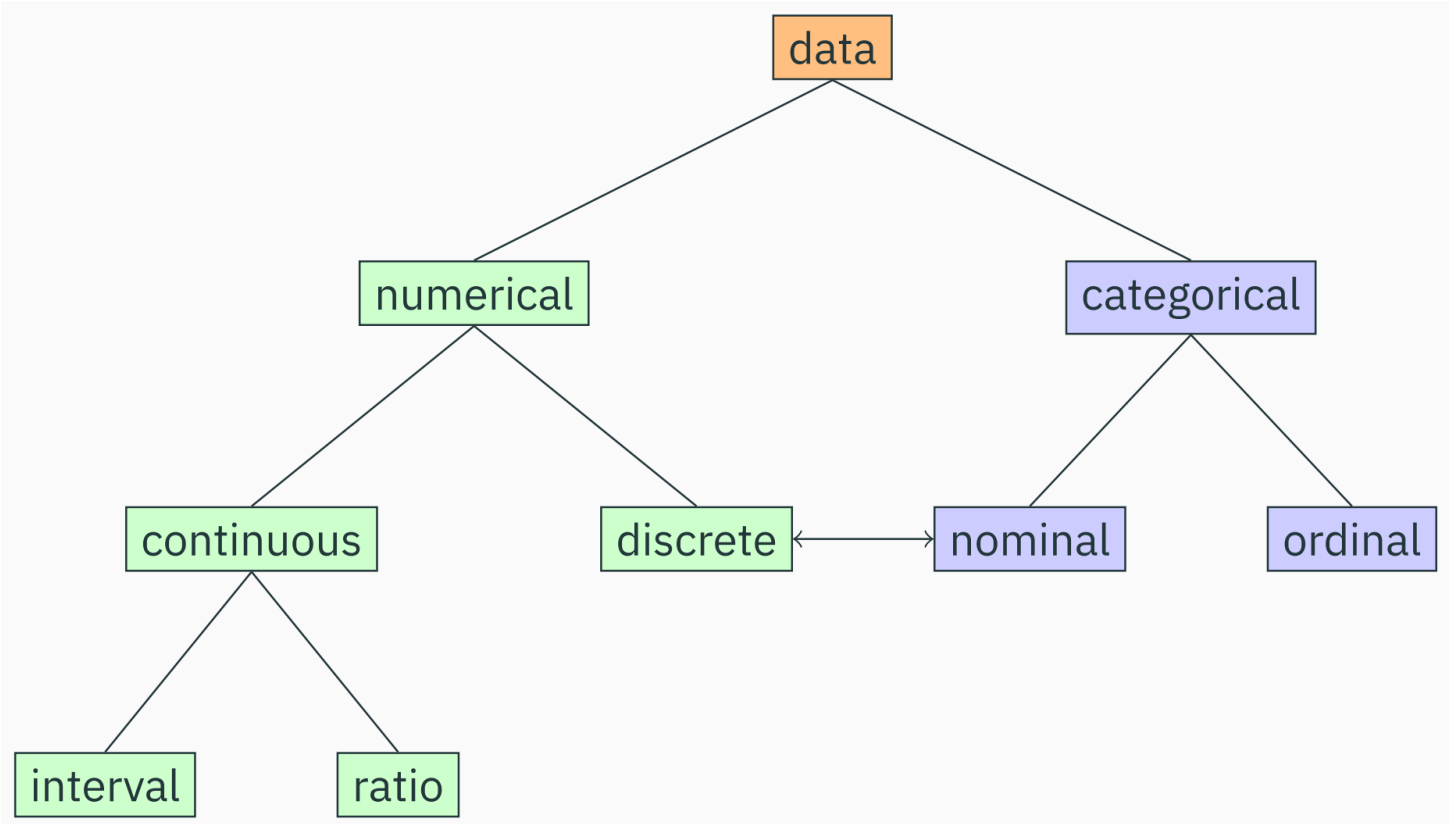
Types of Data

Data is more important than you think

A dumb algorithm with lots and lots of data beats a clever one with modest amounts of it.

Taxonomy of Data

Two fundamental data types: numerical and categorical



Taxonomy of Data

Discrete data

- Can’t be measured but can be counted
- Can be categorized into a classification (nominal)

Continuous data

- Can be measured but can’t be counted
- Interval data: ordered units (equidistant), does not have “true” zero
- Ratio data: same as interval but has “true” zero

Taxonomy of Data

Nominal data

- Discrete labels with no quantitative value
- Has no order

Ordinal data

- Same as nominal data except that ordering matters
- Ordinal scales are used to measure features like education level, customer satisfaction, happiness, etc.

Applicable transformations

Data type	Transformation	Comments
Nominal	Permutation of values	Variables can be put into categories.
Ordinal	Order-preserving mapping, ie. $x' = f(x)$ where f is monotonic.	Variables can be ordered (ranked) but not described by a degree of difference.
Interval	Linear scaling, ie., $x' = a \cdot x + b$ where a and b are constants.	Meaningful difference between two values. Arbitrary zero point. Negative values possible.
Ratio	$x' = a \cdot x$ where a is a constant.	Clearly-defined zero point.

https://en.wikipedia.org/wiki/Statistical_data_type

Applicable statistics

Statistic	Nominal	Ordinal	Interval
ratio, coefficient of variation	x	x	x
mean, stddev, standard error of the mean	x	x	✓
add or subtract	x	x	✓
median and percentiles	x	✓	✓
frequency distribution	✓	✓	✓

https://en.wikipedia.org/wiki/Statistical_data_type

Alternative typologies

Mosteller and Tukey's (1977)

1. Names
2. Grades (ordered labels)
3. Ranks (integer orders)
4. Counted fractions (within [0, 1])
5. Counts (integer ≥ 0)
6. Amount (real ≥ 0)
7. Balance (real)

Chrisman (1998)

1. Nominal
2. Gradation of membership
3. Ordinal
4. Interval
5. Log-interval
6. Extensive ratio
7. Cyclical ratio
8. Derived ratio
9. Counts
10. Absolute

Example

Nominal

- Comfortable
- Uncomfortable

Ordinal

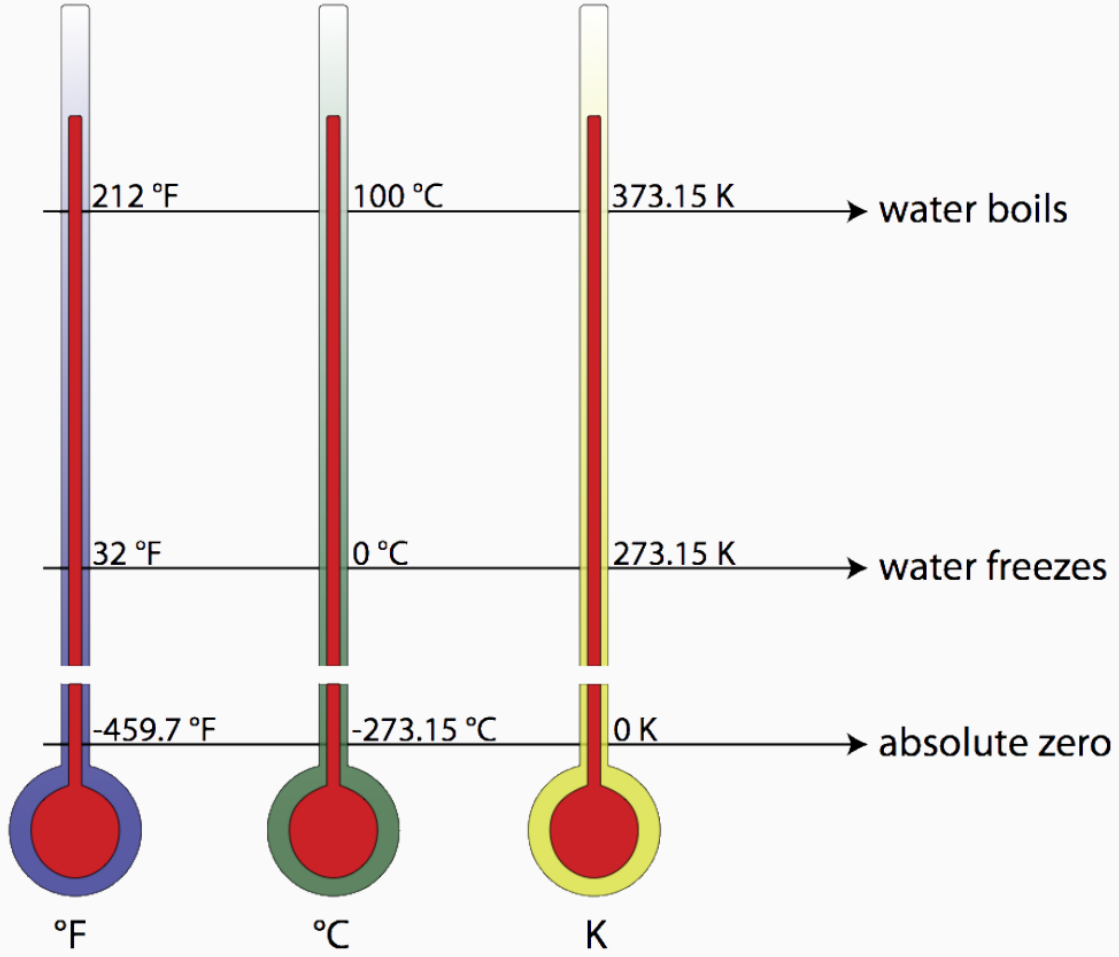
- Cold
- Warm
- Hot

Interval

- Celsius
- Fahrenheit

Ratio

- Kelvin (absolute zero)



Data representation

- We are mostly concerned with data in numerical form
- A quantitative variable takes numerical values about an object
- A dataset is a collection of data, where
  - Each column is a variable or feature
  - Each row is an instance
- A feature is an individual measurable property or characteristic

Data representation

What if data is not numerical?

Data engineering

Concert non-numerical data and give it a quantitative meaning

- Word embeddings for text
- Distance matrices for graph data
- Many other feature engineering techniques

Concepts and terminology

Population

Includes all the elements from a set of data (e.g. all heights of humans)

- Typically, it is impossible to survey/measure the entire population
- If it is possible, it is often costly to do so (and would take lots time)

Sample

Consists of one or more observations drawn from the population

Parameter

A measurable characteristic from a population (μ, σ, etc)

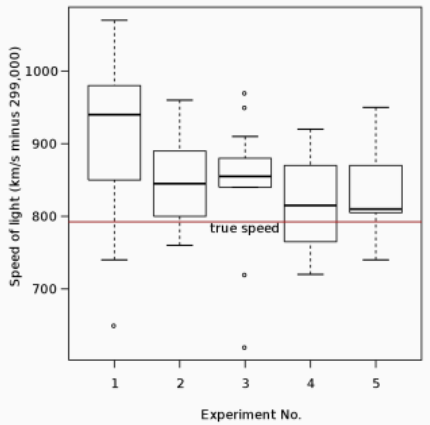
Statistic

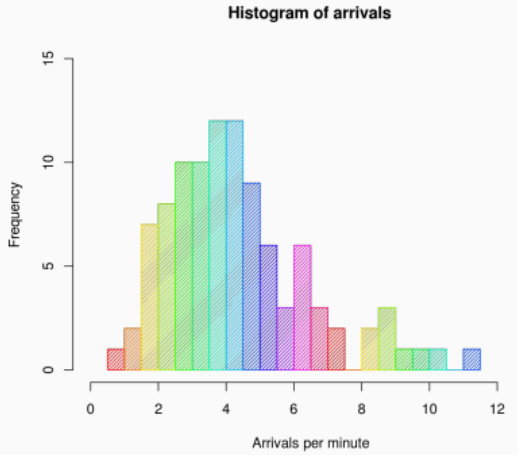
A measurable characteristic of a sample ($\bar{x}, s$, etc)

Graphical exploration

Continuous data

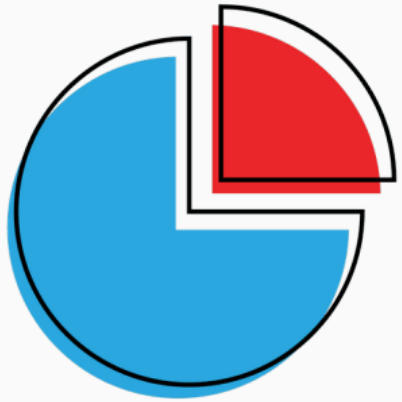
- Boxplots
- Histogram

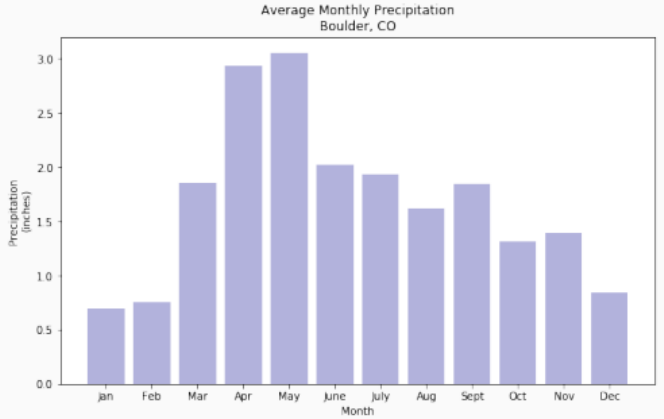




Nominal data

- Frequencies
- Proportion
- Percentage





Exploratory Analysis

- Interactive is best
- Make plots and tables (plot as much of the data as you can)
- For large data, subsample before plotting
- Identify problems: missing data, inconsistencies, outliers
- Calculate summary statistics

Descriptive data summarization

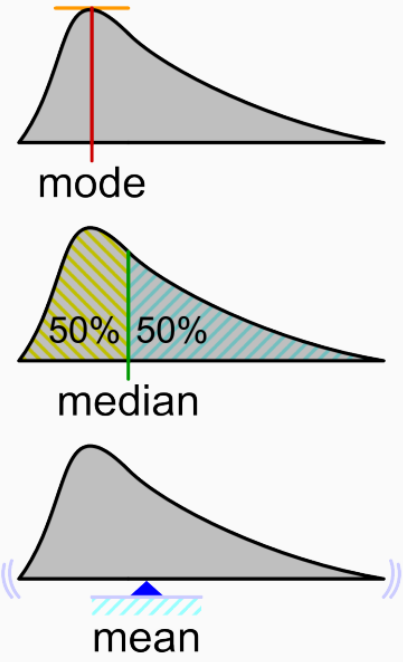
Basic statistics

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(mean)}$$

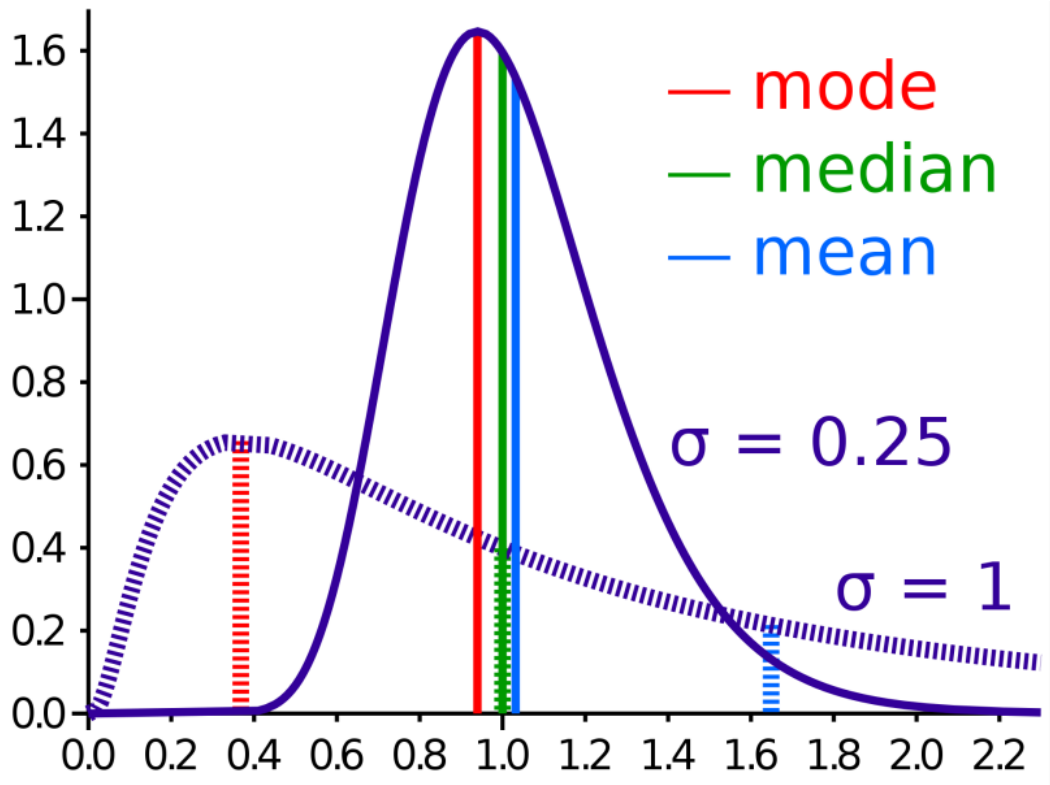
$$\tilde{x} = \frac{\texttt{sorted}(x)_{\lfloor \frac{n+1}{2} \rfloor} + \texttt{sorted}(x)_{\lceil \frac{n+1}{2} \rceil}}{2} \qquad \text{(median)}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \qquad \text{(variance)}$$



Descriptive data summarization

Know the whole story



Descriptive data summarization

Variables can be described by various statistics

Variables describing objects/systems/processes/phenomena are usually not independent.

It’s difficult to find “true” relationship – what ML is about.

The Boxplot

Quartiles

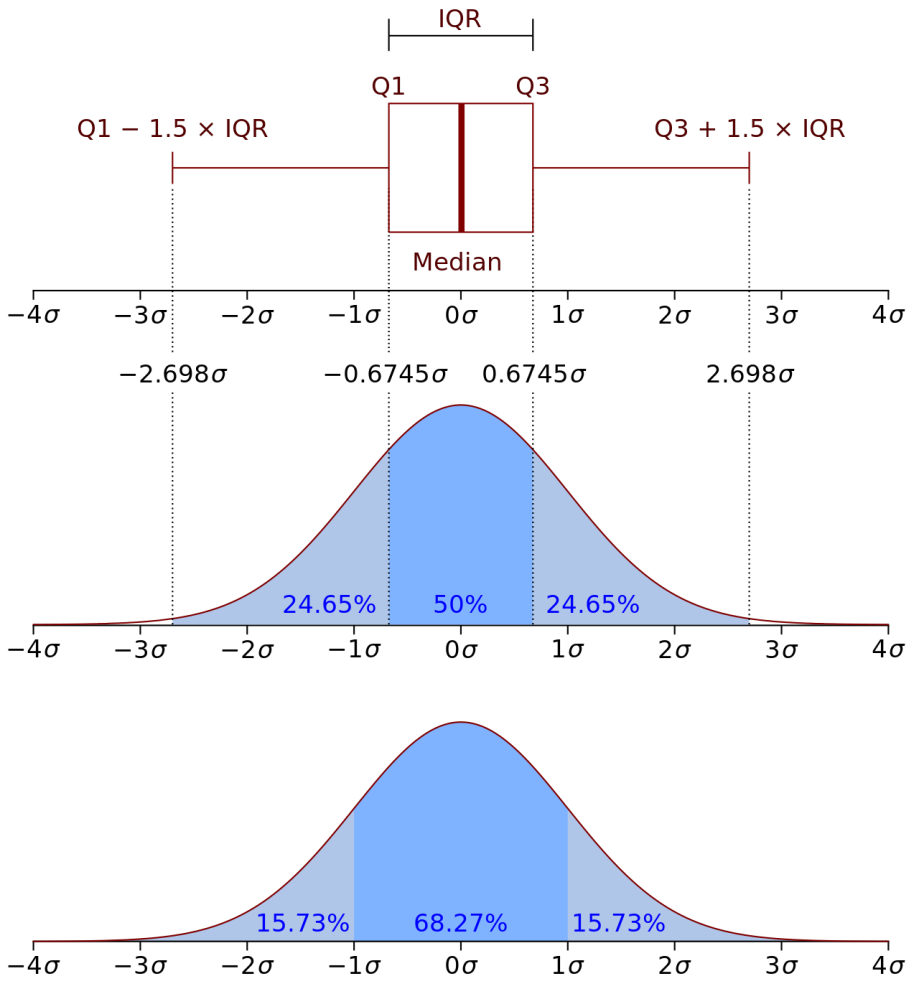
- Q1: 25% of values smaller than Q1
- Q2: 50% of values smaller than Q2 (is the median)
- Q3: 75% of values smaller than Q3

Interquartile Range (IQR)

- IQR = Q3 – Q1 (width of the box)

Whiskers (“minimum” and “maximum”)

- At distance of 1.5 x IQR from Q1, Q3



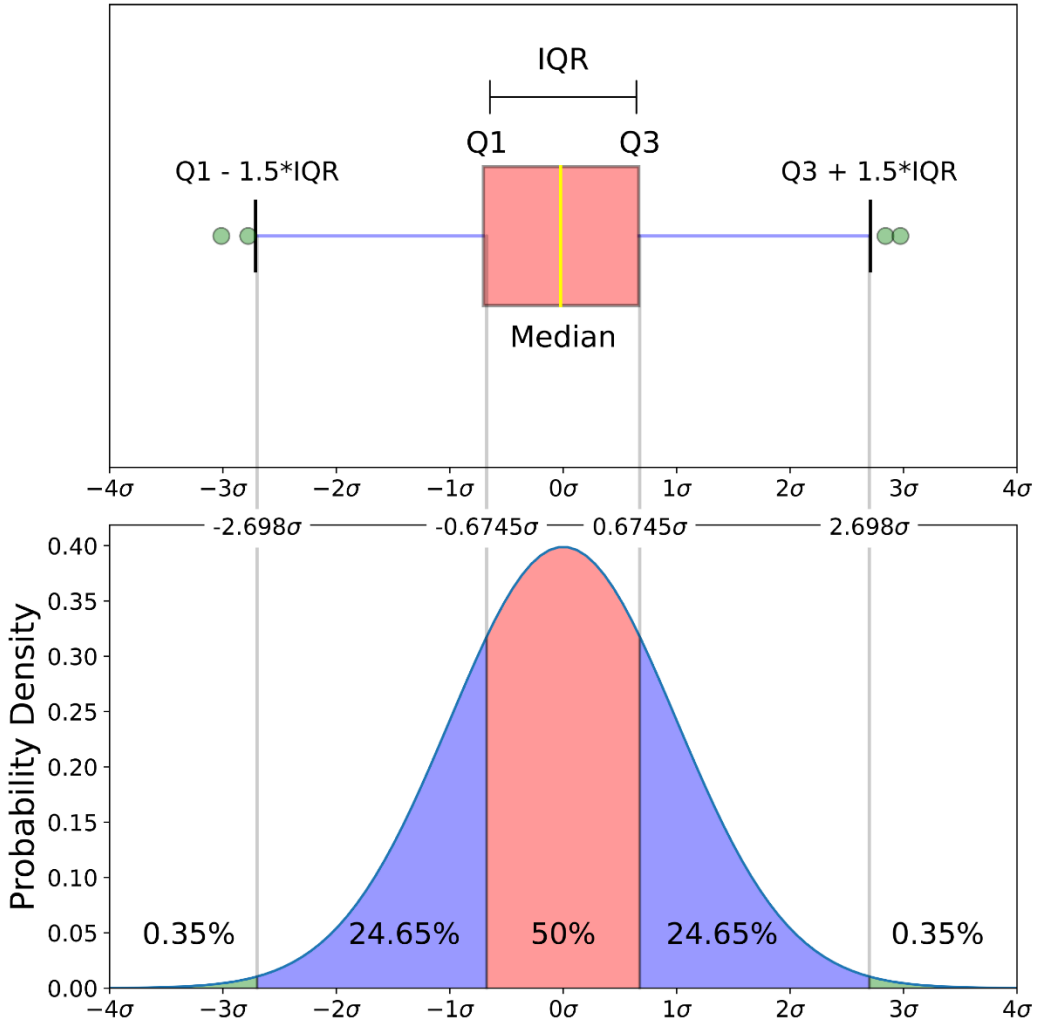
The Boxplot

Useful graphical summary of the data

Can help identify outliers

Outlier

A datapoint that does not belong (possibly noise or an error)



Trivia: why is the normal distribution so common?

In some situations, when independent random variables are added, their (properly normalized) sum tends towards a normal distribution, *even if the variables themselves are not normal-distributed*

Consequence

Probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions

Metrics for data

Are data “similar” or “different”?

- Can we group them together?
- Do they belong in the same category?
- Can we quantify this relationship?

Distance and similarity are fundamental concepts

They are building blocks for many supervised and unsupervised learning algorithms

Metrics for data

Kinds of distances

- Between data points
- Between a point and a distribution
- Between distributions

In most cases, we take the mathematical notion of distance (a metric)

Similarity is taken to be (in some sense) the inverse of the distance

Properties of a distance

1. Non-negativity
$$d(p,q) \geq 0, \forall p,q$$
2. Identity of indiscernibles
$$d(p,q) = 0 \Rightarrow\ p = q$$
3. Symmetry
$$d(p,q) = d(q,p), \forall p,q$$
4. Triangle inequality (subadditivity)
$$d(p,r) \leq d(p,q) + d(q,r), \forall p,q,r$$

Properties of a similarity

1. Boundedness

$$s(p,q) \in [0,1], \forall p,q$$

2. Identity of indiscernibles

$$s(p,q) = 0 \Rightarrow p \text{ and } q \text{ have nothing in common}$$
$$s(p,q) = 1 \Rightarrow p \text{ and } q \text{ are the same}$$

3. Symmetry

$$s(p,q) = s(q,p), \forall p,q$$

In general, similarity can be seen as the opposite of (normalized) distance

Scalar, vector, multi-vector distance

Scalar data

Each instance is a scalar

Vector data

Each instance is a d-dimensional vector

Multi-vector data

Each instance is represented by several d-dimensional feature vectors

Usually summarized by Gaussian distributions

Scalar distance

Co-occurrence count between two data items

- Two gene sequences in a DNA helix
- Two words in a set of documents
- Two pixel values in a set of images

Usually, a feature co-occurrence matrix is built (in a given context)

Useful in the task of collocation extraction

Vector distance

Vector data

Each instance is a d-dimensional vector

Example:

- Euclidean distance
- Manhattan distance
- Correlation
- Cosine distance (similarity)
- Mahalanobis distance

Manhattan distance

Taxicab distance

- Also known as: L_1 distance, L^1 distance, l_1 norm, snake distance, city block distance, Manhattan distance
- The distance between two points is the sum of the absolute differences of their Cartesian coordinates.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

- Useful when Euclidean distance is not enough (high dimensions)
- The L_1 distance metric (Manhattan distance metric) is the most preferable for high dimensional applications, followed by the Euclidean metric $(L_2)^{11}$.

Euclidean distance

Euclidean distance

Very useful because it is versatile, adaptable and has “nice” mathematical properties.

Straight line between two points in Euclidean space.

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Applications

- Clustering
- KNN classification and regression
- Dimensionality reduction (convert high dimensional data to representation where Euclidean distance works)

Euclidean distance

Why is Euclidean distance not a good metric in high dimensions?

In high dimensions… intuition fails

- Most of the volume of a orange is in the skin, not the pulp
- Most of the mass of a multivariate Gaussian distribution is not near the mean, but in an increasingly distant “shell” around it
- If we approximate a hypersphere by inscribing it in a hypercube, almost all the volume of the hypercube is outside the hypersphere

Pedro Domingo, *A Few Useful Things to Know about Machine Learning*

http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Correlation coefficient

For a population

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

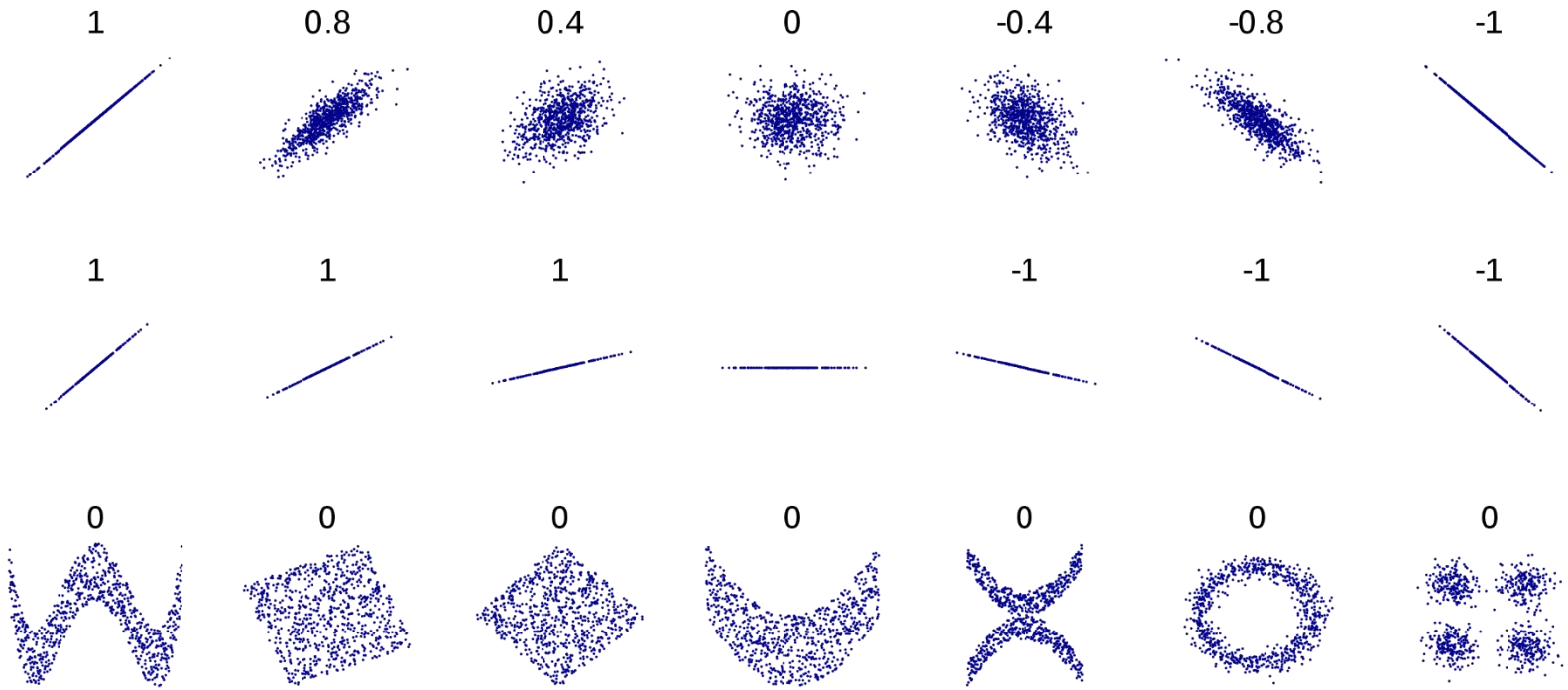
For a sample

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- represents a measure of linear correlation
- it has a value between -1 (total negative linear correlation) and $+1$ (total positive linear correlation)
- zero means no correlation
- do not confuse with R^2 (coefficient of determination)

Correlation coefficient

- Reflects the strength and direction of a linear relationship
- Does not reflect nonlinear relationships



Cosine similarity

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\text{similarity} = \cos\theta = \frac{\mathbf{p}\cdot\mathbf{q}}{\|\mathbf{p}\|\cdot\|\mathbf{q}\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$

-1 Exactly opposite

0 Orthogonality (decorrelation)

1 Exactly the same

No magnitude information

Mahalanobis distance

If our data D can be described by a covariance matrix Σ, then the distance between two vectors x and y from D is given by

$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

If the covariance matrix Σ is the identity matrix, then the Mahalanobis distance reduces to the Euclidean distance.

Distance between distributions

Distance between distributions can be defined using the concept of informational entropy

Entropy (from Greek τροπή [trop¯e] – transformation) is related to the conservation of energy. Coined by Rudolf Clausius (founder of thermodynamics)

Boltzmann's Entropy (1877) – Statistical Thermodynamics

Measure of statistical "mixedupness" or disorder

$$S = k_B \ln \Omega$$

Where k_B is Boltzmann's constant and Ω is the number of microstates consistent with the given equilibrium macrostate.

Distance between distributions

Gibbs Entropy

Gibbs coined the term statistical mechanics, explaining the laws of thermodynamics as consequences of the statistical properties of ensembles of the possible states of a physical system composed of many particles.

$$S = -k_B \sum p_i \ln p(i)$$

Mathematically equivalent to Boltzmann's entropy (but not always physical)

Distance between distributions

Informational Entropy

Developed by **Claude Shannon (1948)** to mathematically quantify the statistical nature of “lost information” in phone-line signals.

Unaware of thermodynamics, Shannon wanted to call it uncertainty.

Renamed it to entropy at the advice of John von Neumann.

The average amount of information produced by a stochastic source of data:

$$H = -K \sum_{i=1}^{k} p(i) \log p(i)$$

where K is a positive constant and $p(i)$ is the probability of state i.

Distance between distributions

Informational entropy is present whenever there are unknown quantities that can be described only by a probability distribution.

Example: coin toss

Entropy is maximized if the coin is fair ($p_{heads} = p_{tails}$). Let X be a random variable with two possible outcomes x_1, x_2.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) = -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2} = -\sum_{i=1}^{2} \frac{1}{2} \times (-1) = 1$$

Each coin toss delivers one full bit of information.

If the coin is biased (probabilities are not equal), then it's easier to predict the outcome → less entropy (uncertainty).

Distance between distributions

Kullback-Leibler Divergence

For two probability distributions P and Q, $D_{KL}(P \parallel Q)$ can be seen as the *relative entropy* of P with respect to Q, or the *amount of information lost* when Q is used to approximate P.

$$D_{KL}(P \parallel Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)} = \sum_i P(i) \log \frac{P(i)}{Q(i)} \qquad \text{(discrete)}$$

$$D_{KL}(P \parallel Q) = \int p(x) \log \frac{p(x)}{q(x)} dx \qquad \text{(continuous)}$$

Distance between distributions

Kullback-Leibler Divergence

For two probability distributions P and Q, $D_{KL}(P \parallel Q)$ can be seen as the *relative entropy* of P with respect to Q, or the *amount of information lost* when Q is used to approximate P.

$$D_{KL}(P \parallel Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)} = \sum_i P(i) \log \frac{P(i)}{Q(i)} \qquad \text{(discrete)}$$

$$D_{KL}(P \parallel Q) = \int p(x) \log \frac{p(x)}{q(x)} dx \qquad \text{(continuous)}$$

Why the logarithm?

Statistical interpretation

If p_i are the probabilities of independent events, then the total probability is:

$$p = \prod_i p_i$$

It is easier and more intuitive to work with the sum of log-probabilities:

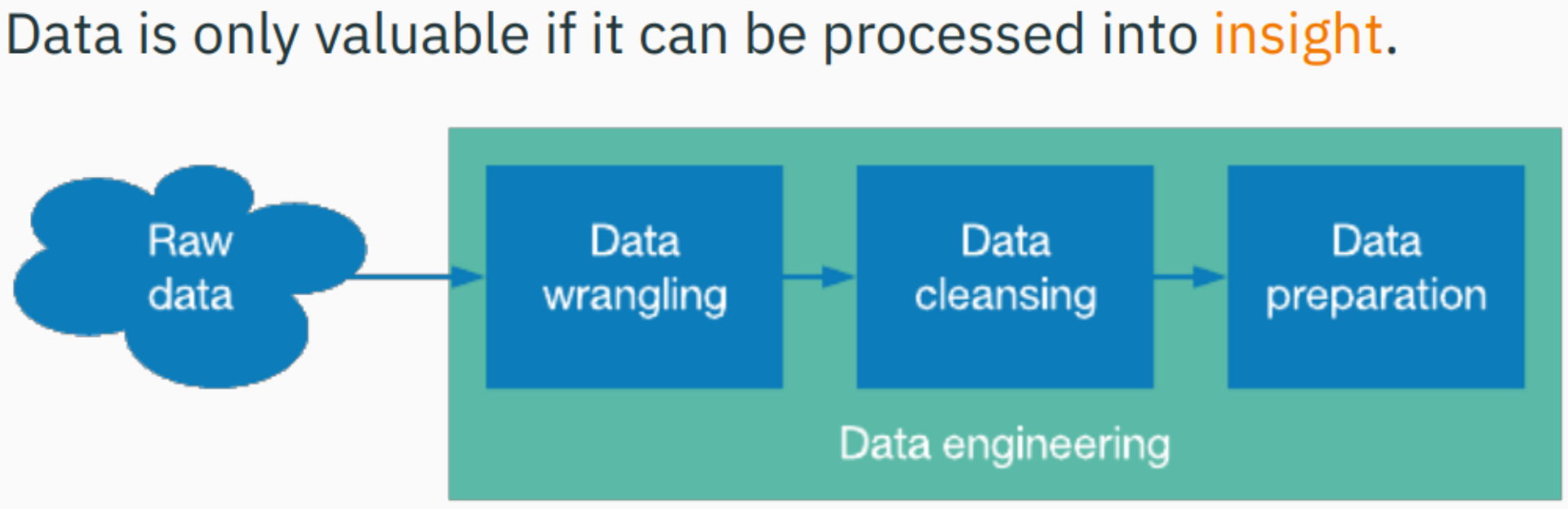
$$\ln p = \sum_i \ln p_i$$

Physical interpretation

We want to define entropy as an extensive property. Extensive properties of systems are additive for subsystems (like *mass* or *volume*).

One (fair) coin → one bit, two (fair) coins → two bits of information.

Data engineering



Data scientists spend around 80% of their time collecting, cleaning, preparing data for use in machine learning.

No quality data → no quality results

Data engineering pipeline

Wrangling

Identify, collect, merge and preprocess one or more datasets in preparation for data cleansing

Cleaning

Ensure that the data is syntactically and semantically correct

- Syntax: missing values, incorrect delimiters, inconsistencies
- Semantics: filtering outliers and noise

Preprocessing

- Scaling, normalization, standardization
- Reduction, discretization
- Encoding categorical data into numerical values

Data engineering pipeline

Main factors in data quality

Accuracy. Inaccurate data may result from:

- Human/computer errors
- Obfuscated or incorrect values
- Incorrect formats
- Duplication of records

Data engineering pipeline

Main factors in data quality

Completeness. The data should include all relevant attributes/features

- Data (un-)availability
- Features removed due to being considered irrelevant

Consistency. Depends on data sources and how they are aggregated

Dealing with missing values

Approaches

- Remove the training example
- Remove a feature if more than % of values are missing
- Try to fill in missing values (manually or automatically)

Fill in missing data

- Use a standard value (i.e., 0, nan, 'N/A')
- Use central tendency (mean, median, mode) (of the entire feature or separated i.e., by class label)
- Use prediction (i.e., regression, decision trees)
- Use statistical methods

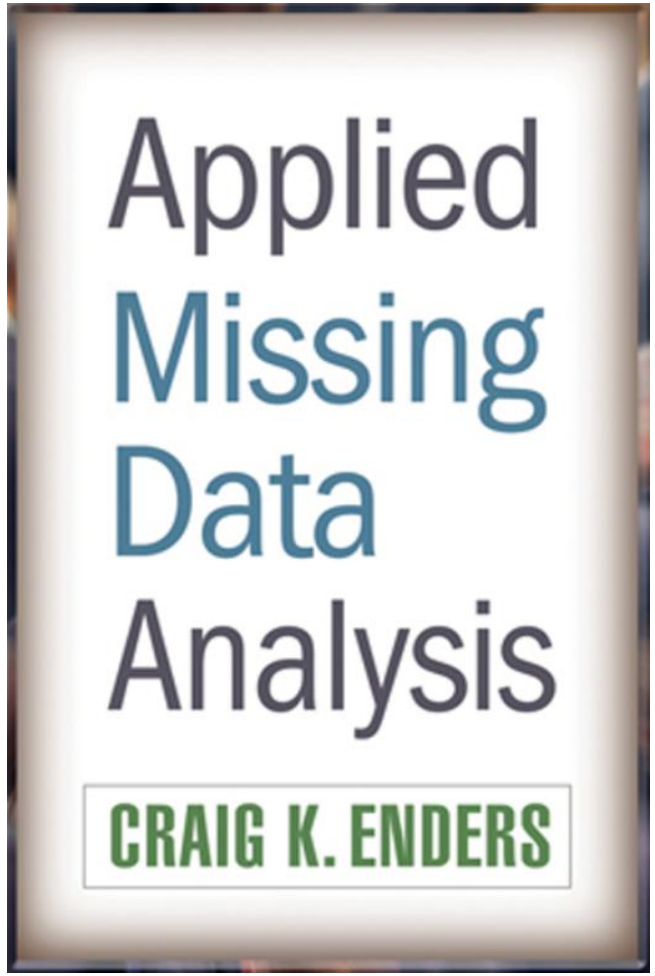
Missing data

Missing data analysis

- Data is missing everywhere
- Many ad-hoc techniques exist
- Simple methods rely on assumptions about the cause of missing data
- Wrong assumptions → introducing bias
- Need to find the pattern behind missing data

Current state of the art

- Maximum likelihood estimation
- Multiple imputation



Missing data

Typology of missing data

Roderick and Rubin, *Statistical Analysis with Missing Data*

Missing completely at random (MCAR)

Missingness on x unrelated to observed values of other variables or unobserved values of x

Missing at random (MAR)

Missingness on x uncorrelated with the unobserved value of x, after adjusting for observed variables

Missing not at random (MNAR)

Missingness on x is correlated with the unobserved value of x

Some methods handle missing data, but most don't

Dealing with noise

Noise defined as random variance in a measured variable.

Causes

- Faulty data collection (e.g., sensors)
- Data entry problems
- Transmission problems or other technological limitations

Symptoms

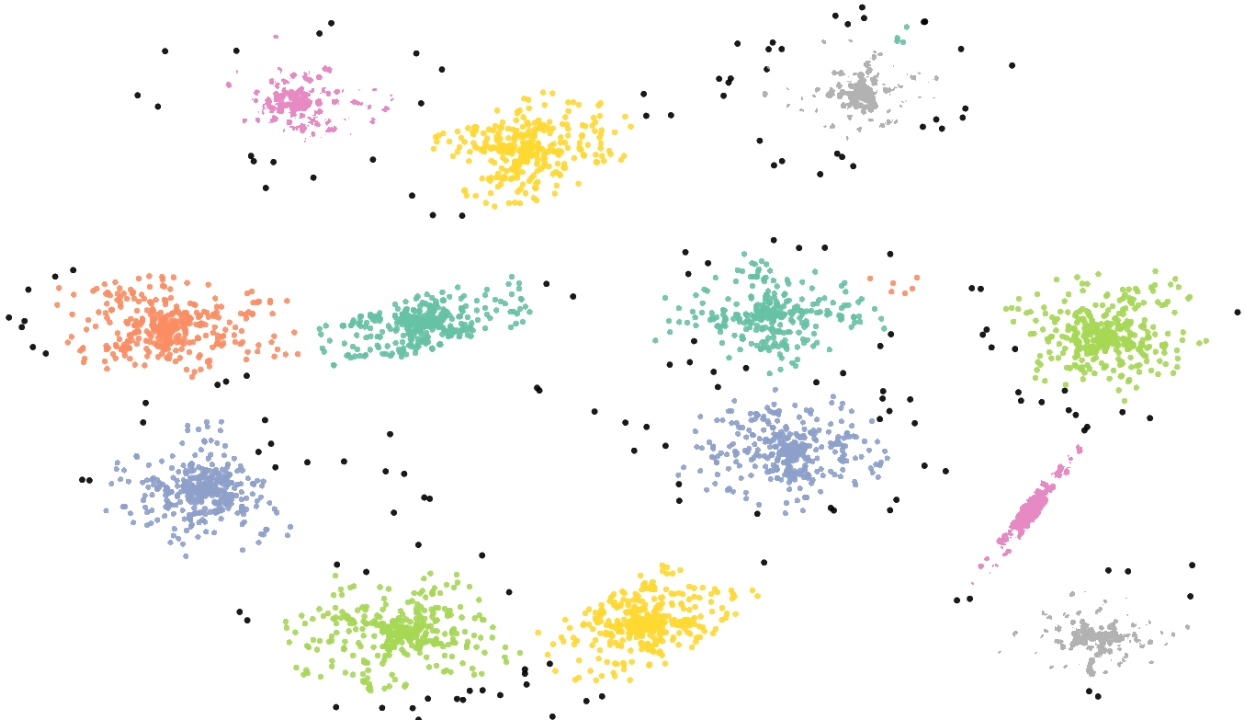
- Duplicate records
- Incomplete data
- Inconsistent data

Dealing with noise

Boxplots and scatter plots can be used to identify outliers.

Smoothing

- Binning
- Regression
- Outlier analysis (i.e., clustering)



Descriptive data summarization

Before everything else, we must get an overview

- Central tendency
- Variation
- Dispersion

Statistics

- Median, mean
- Min, max
- Quantiles, variance, outliers

Use different levels of granularity

Descriptive data summarization

Measuring the central tendency

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Weighted mean

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Trimmed mean: chopping extreme values

Descriptive data summarization

Measuring the central tendency

Median – a holistic measure

$$median(x) = f(x) = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{(n+1)/2}}{2}, & \text{if } n \text{ is even} \end{cases}$$

Middle value if n is odd, average of two middle values otherwise.

Half the observations are smaller than it and half are larger.

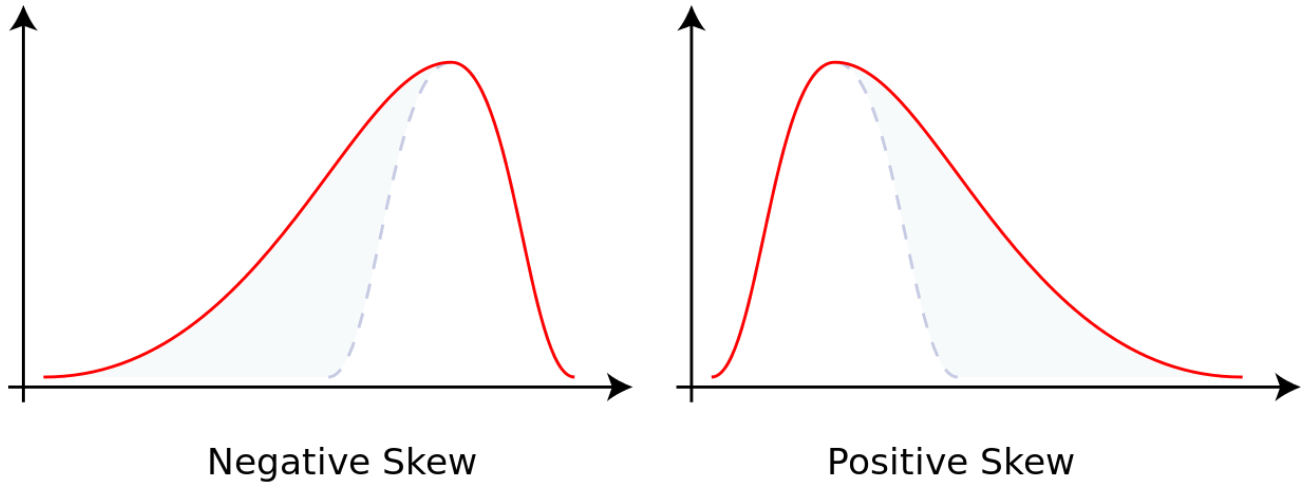
(this is why the median is also known as the balance point of the data)

Skewness

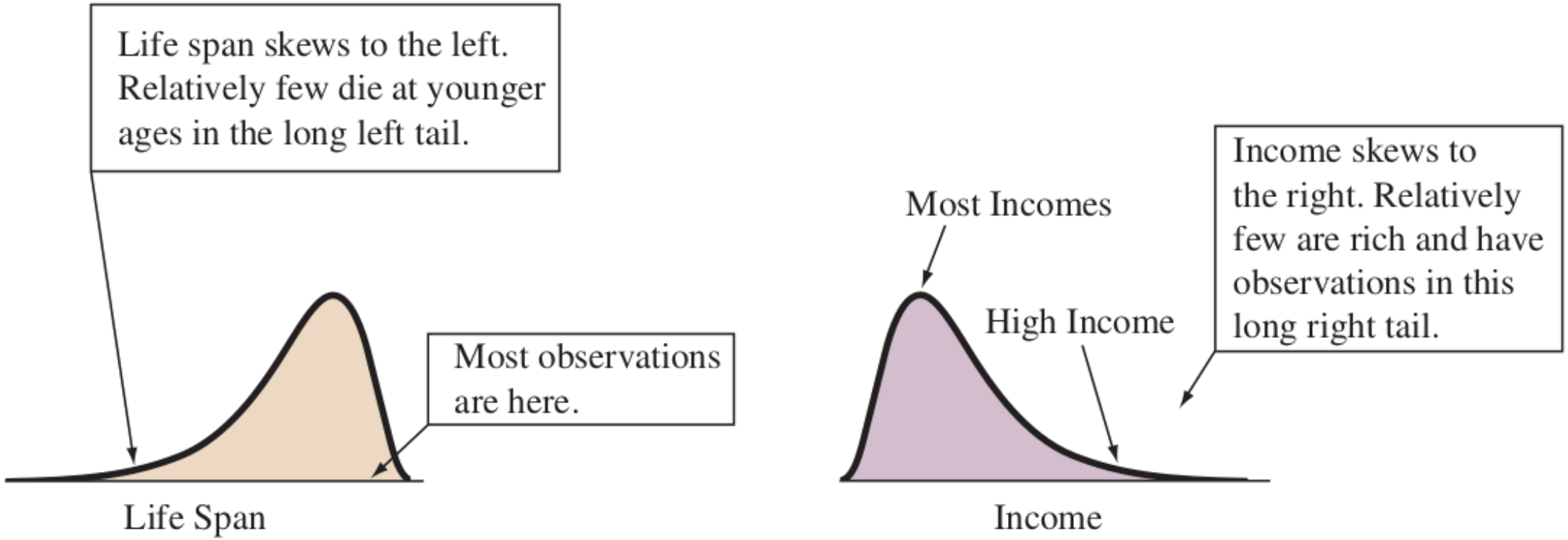
Skewness is a measure of asymmetry in the data.

For a unimodal distribution, the skew indicates which side the tail of the distribution is:

- Positive skew: tail to the right
- Negative skew: tail to the left

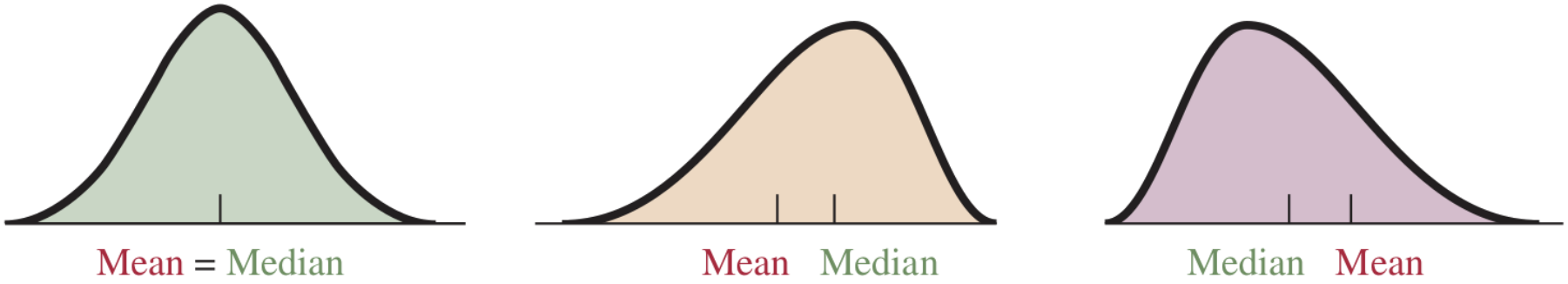


Skewness


Life span skews to the left. Relatively few die at younger ages in the long left tail.
Most observations are here.
Life Span
Most Incomes
High Income
Income skews to the right. Relatively few are rich and have observations in this long right tail.
Income

Skewness

Relationship between mean and median (rule-of-thumb)



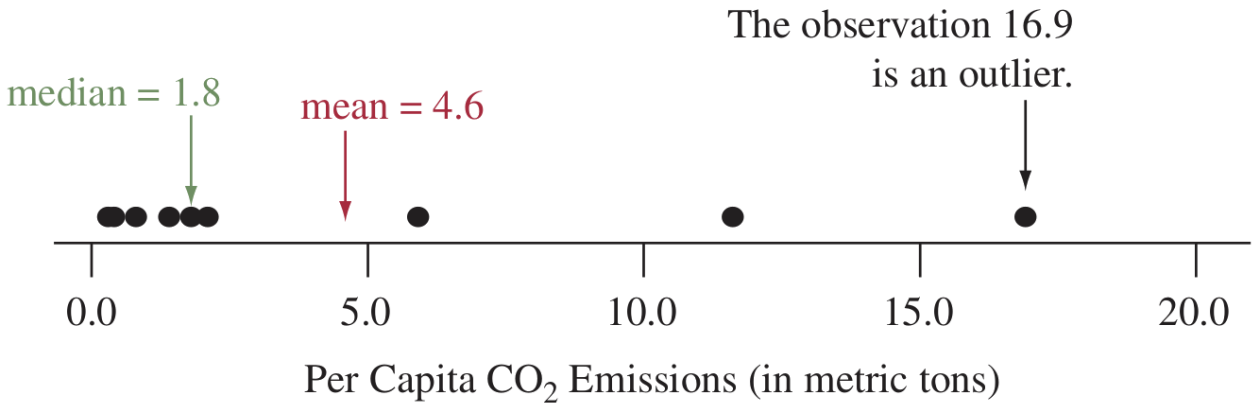
This rule may fail in multimodal distributions or when one tail is *long* and the other one is *heavy*.

Heavy: heavier than the exponential distribution

Long: large number of occurrences far from the “head” or central part of the distribution

Outlier detection

Descriptive statistics may help detecting outliers



China 5.9	Indonesia 1.8	Nigeria 0.3
India 1.4	Brazil 2.1	Bangladesh 0.4
United States 16.9	Pakistan 0.8	Russian Federation 11.6

An outlier is a point that is far away from the rest (bulk) of the data.

Measuring the dispersion of data

Quartiles

Q_1 (25th percentile, 0.25 quantile)

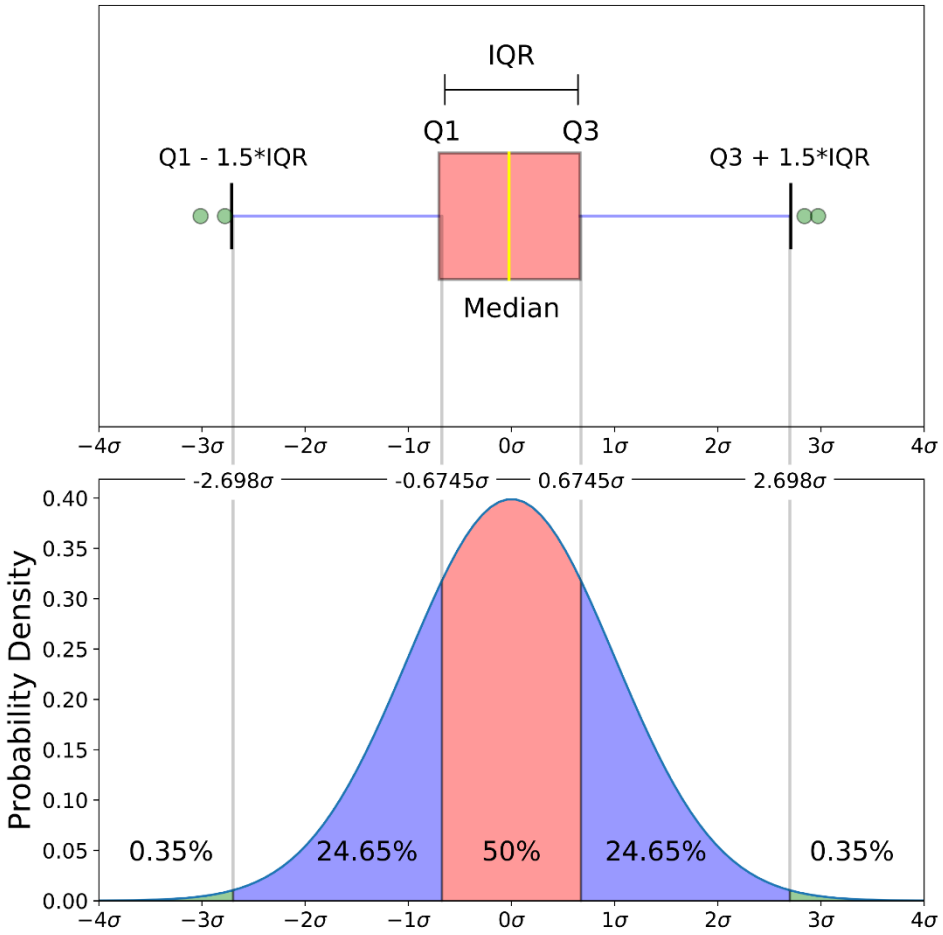
Q_3 (75th percentile, 0.75 quantile)

Inter-quartile range

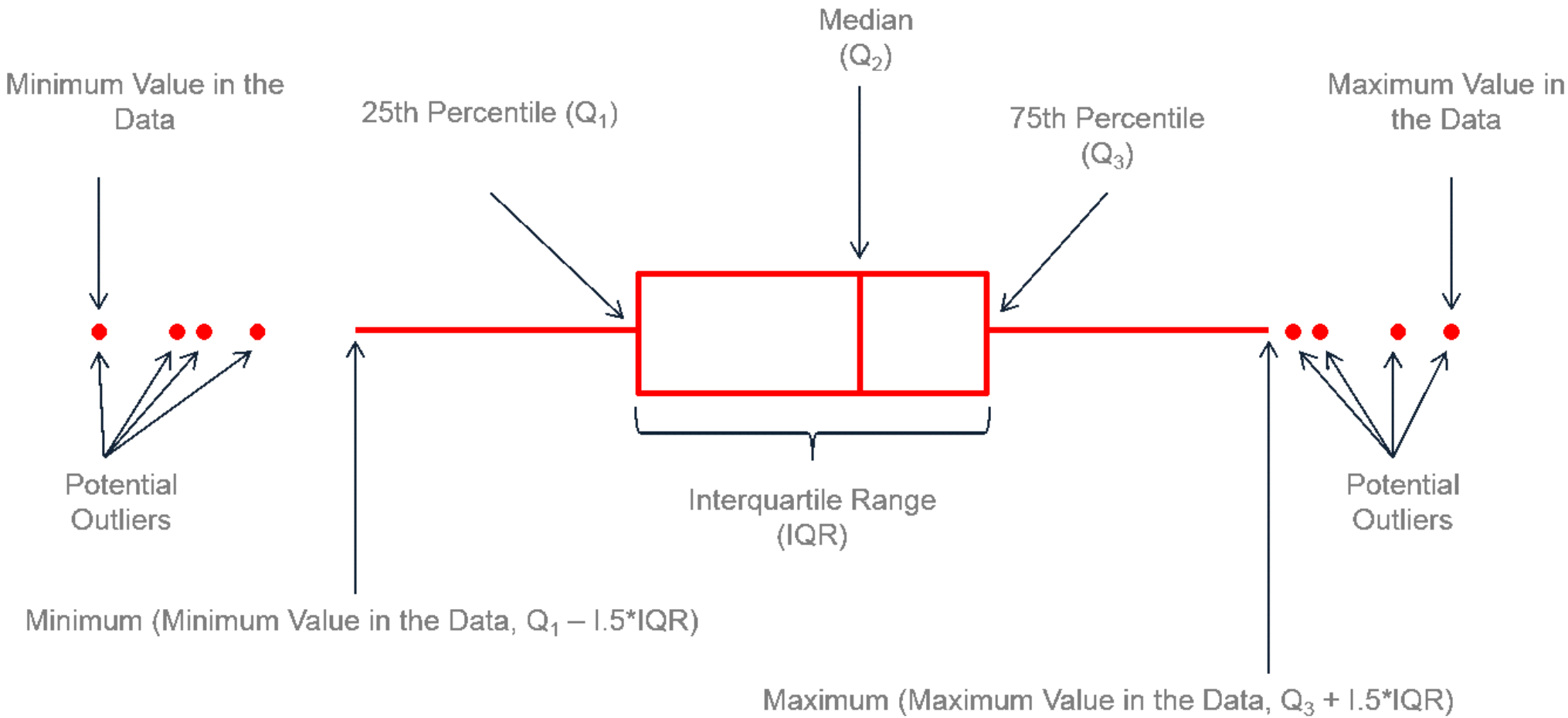
$IQR = Q_3 - Q_1$

Five number summary: min, Q_1, M, Q_3, max

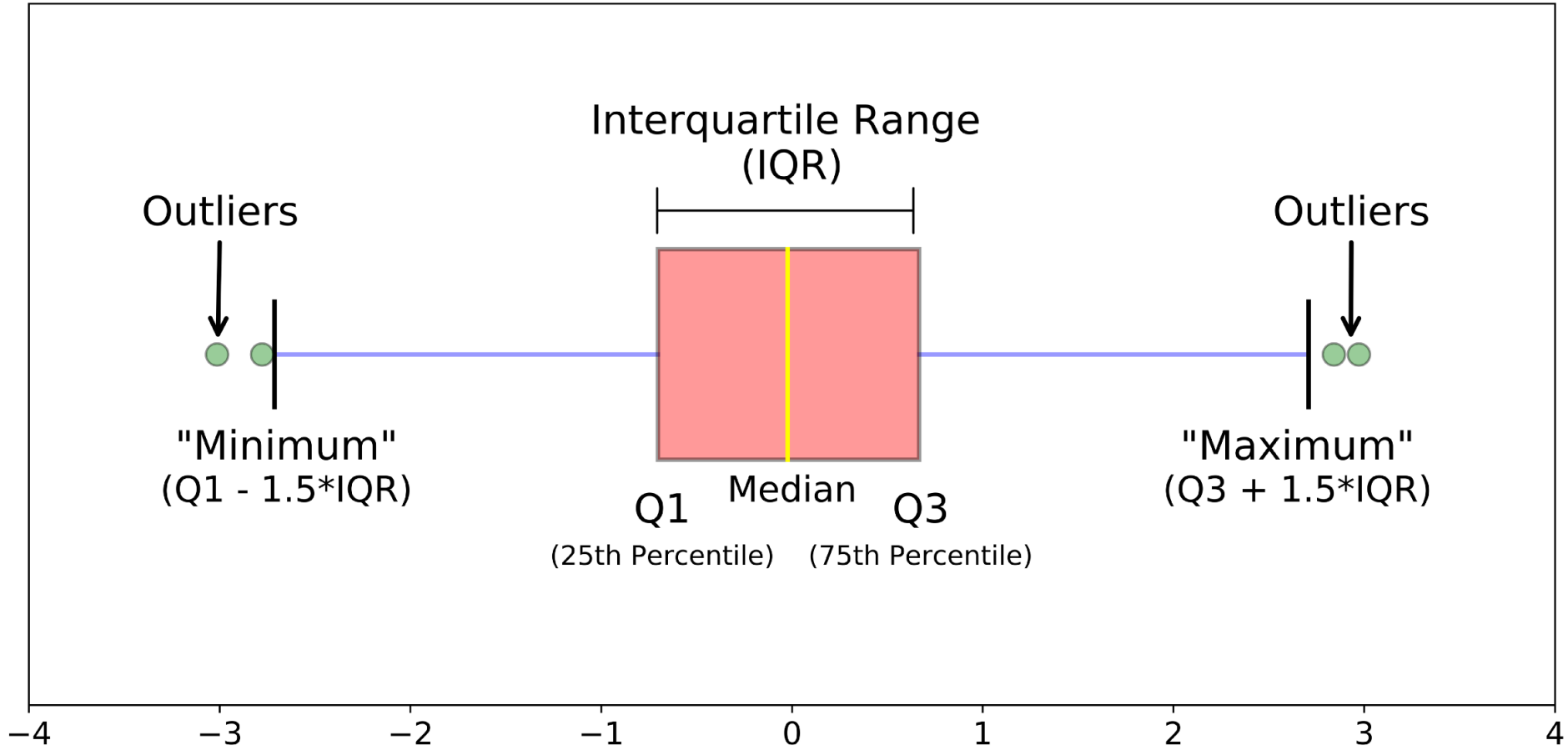
Outlier: usually a value outside $1.5 \times IQR$



The box plot

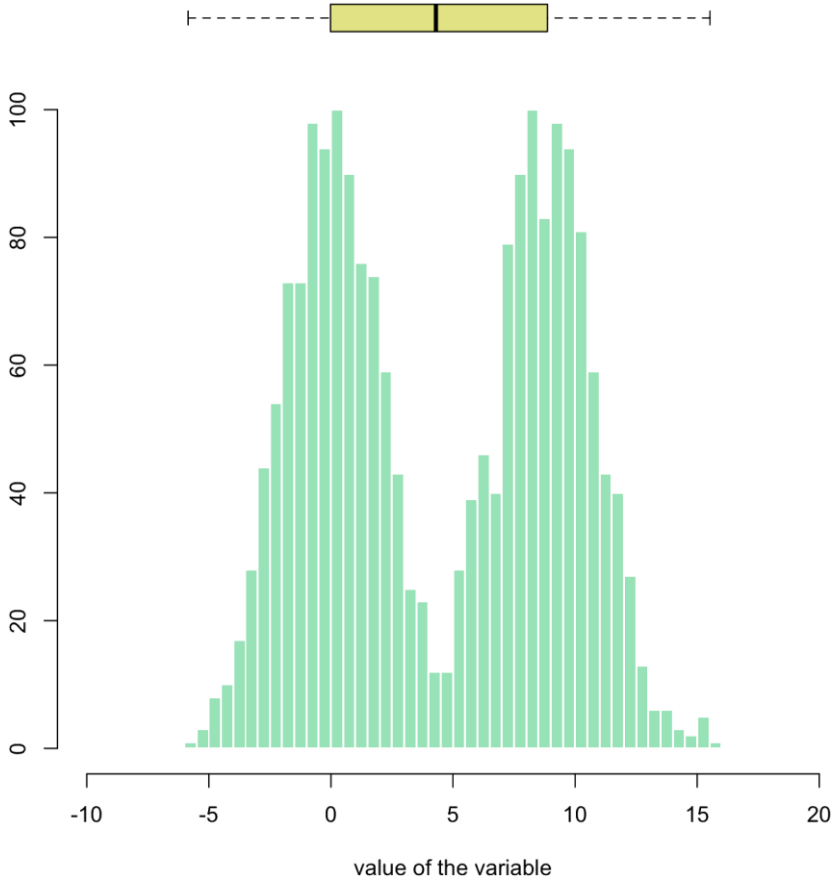

Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile
(Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range
(IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

The box plot

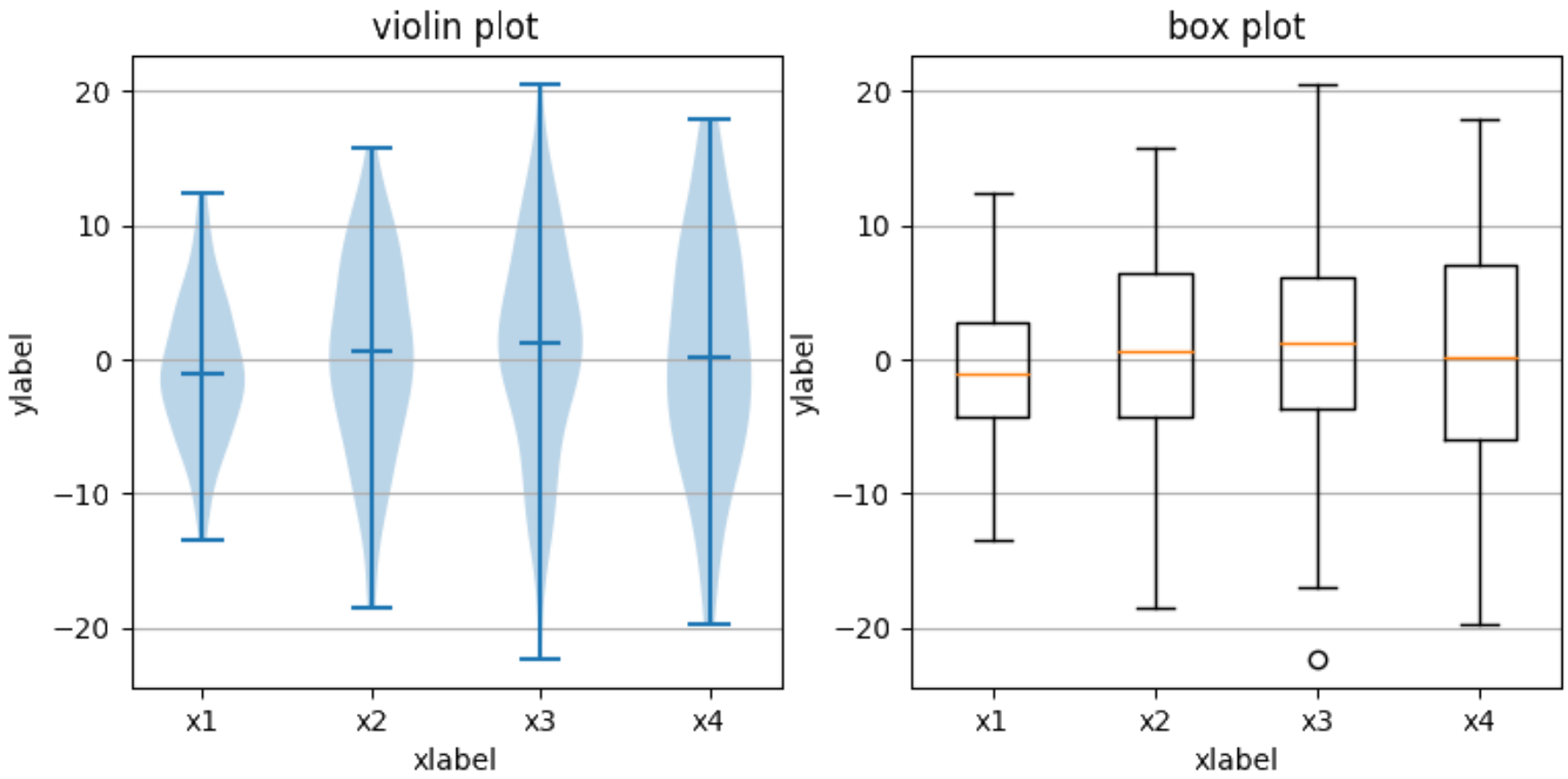

Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

Other representations

The box plot is a powerful visualization but not always sufficient



Other representations



Binning

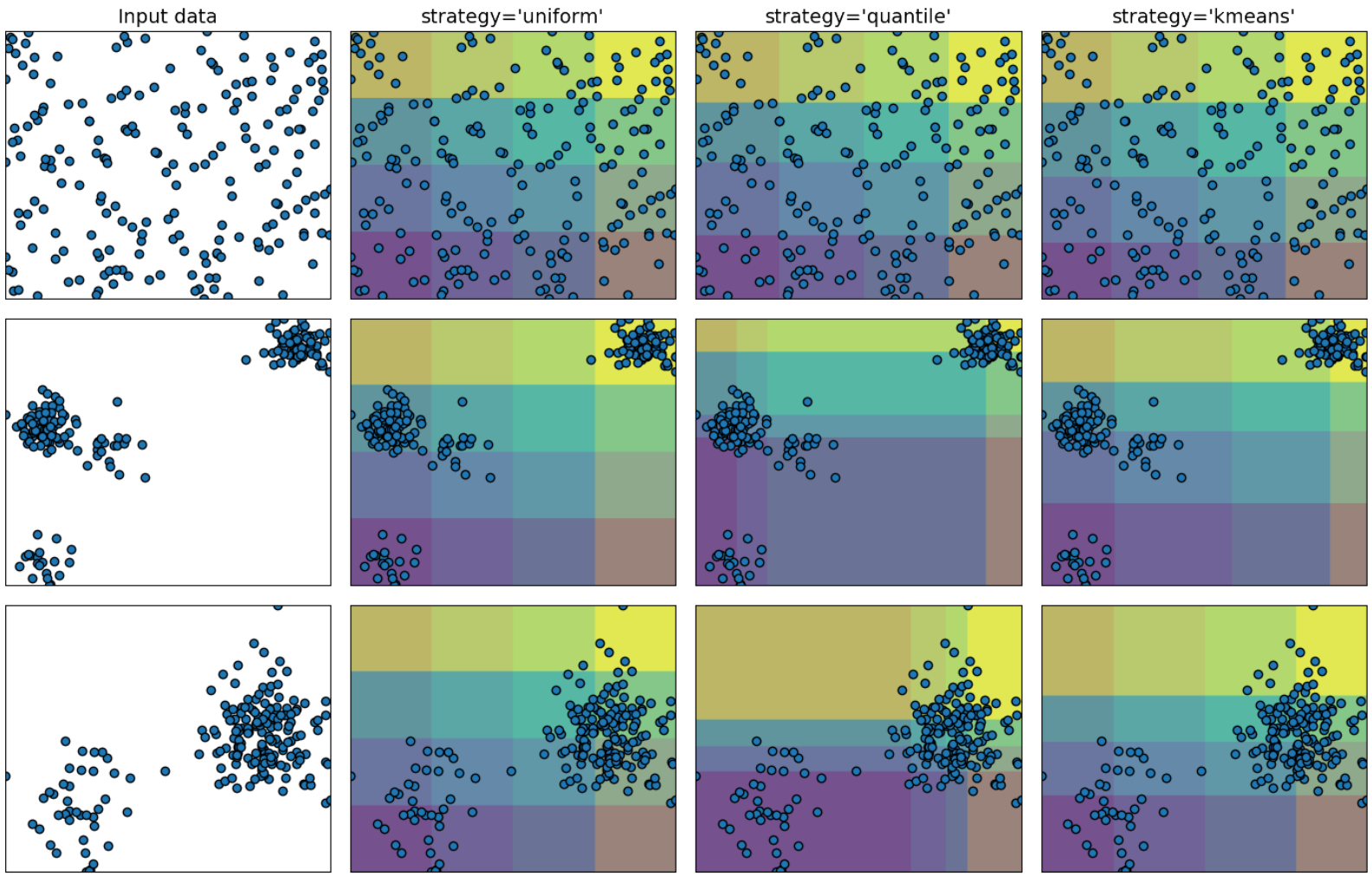
Used to discretize data and deal with noise

Uniform: bin widths are constant in each dimension

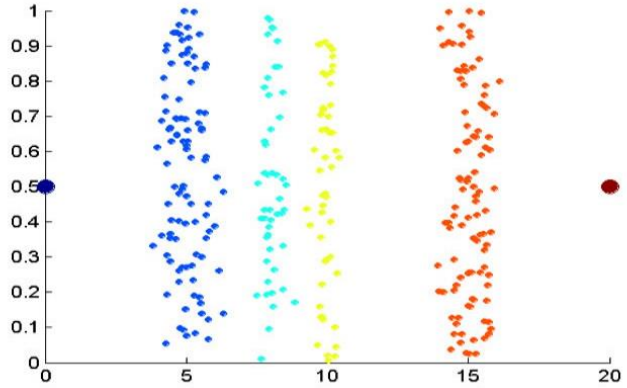
Quantile: bin widths according to quantiles, ~same number of samples

K-means: discretization based on k-means clustering

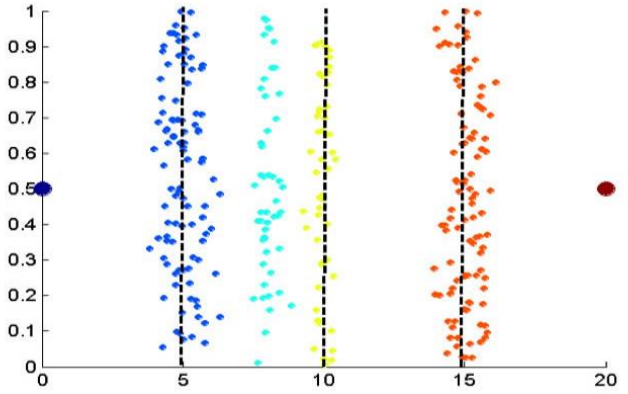
Binning


Input data
strategy='uniform'
strategy='quantile'
strategy='kmeans'

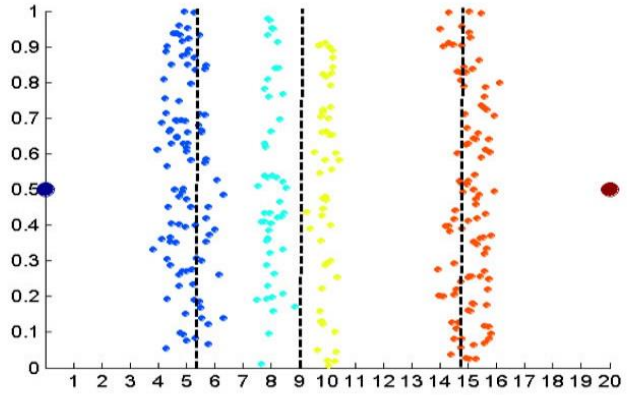
Binning



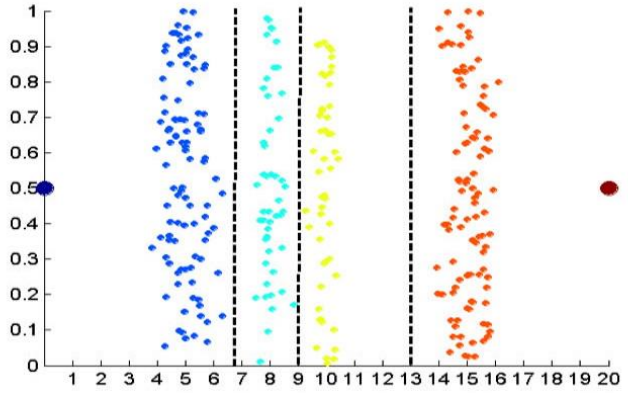
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.



Equal interval width approach used to obtain 4 values.

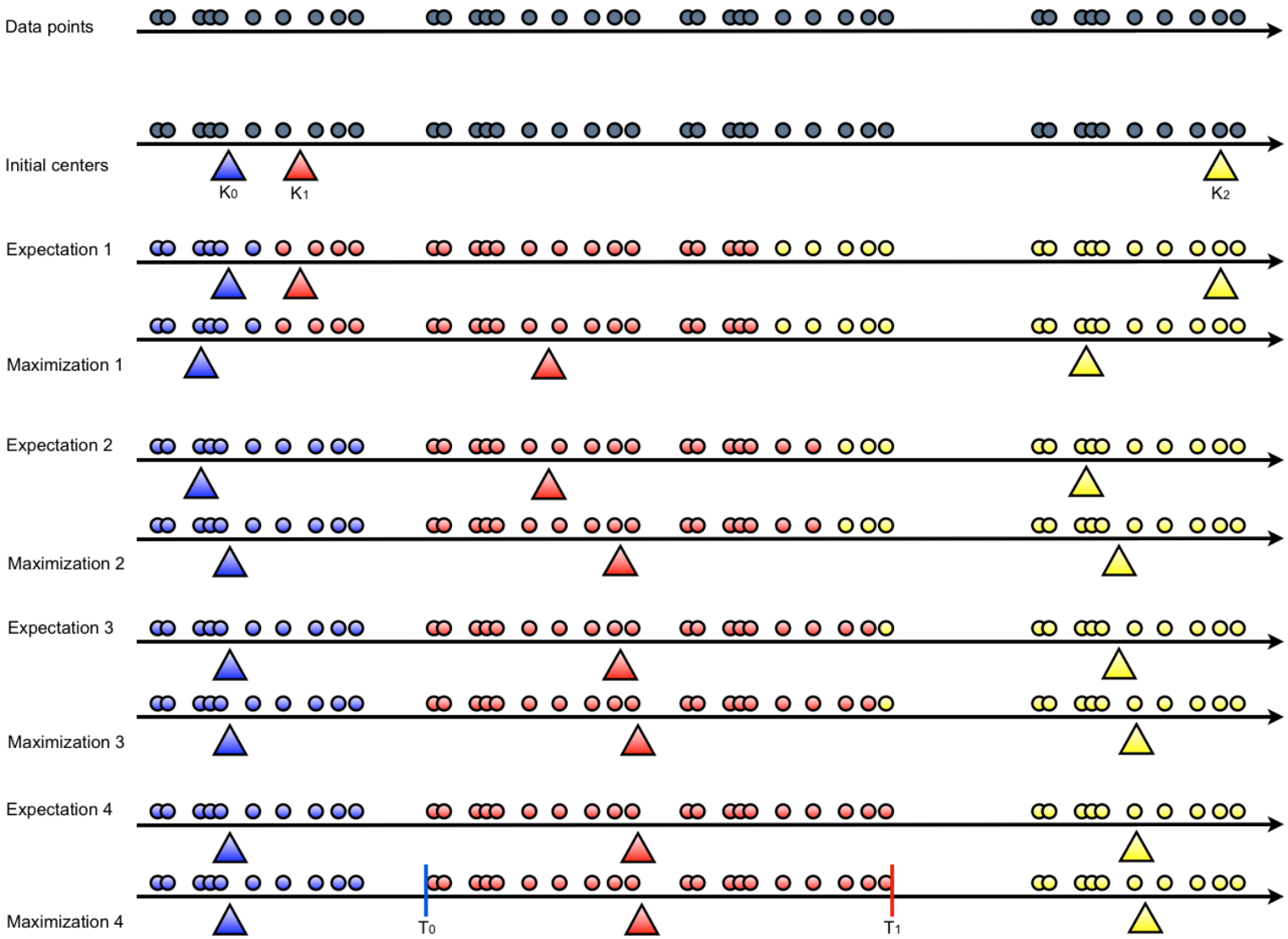


Equal frequency approach used to obtain 4 values.



K-means approach to obtain 4 values.

Binning


Data points
Initial centers
K0
K1
K2
Expectation 1
Maximization 1
Expectation 2
Maximization 2
Expectation 3
Maximization 3
Expectation 4
Maximization 4
T0
T1

Handling redundancy

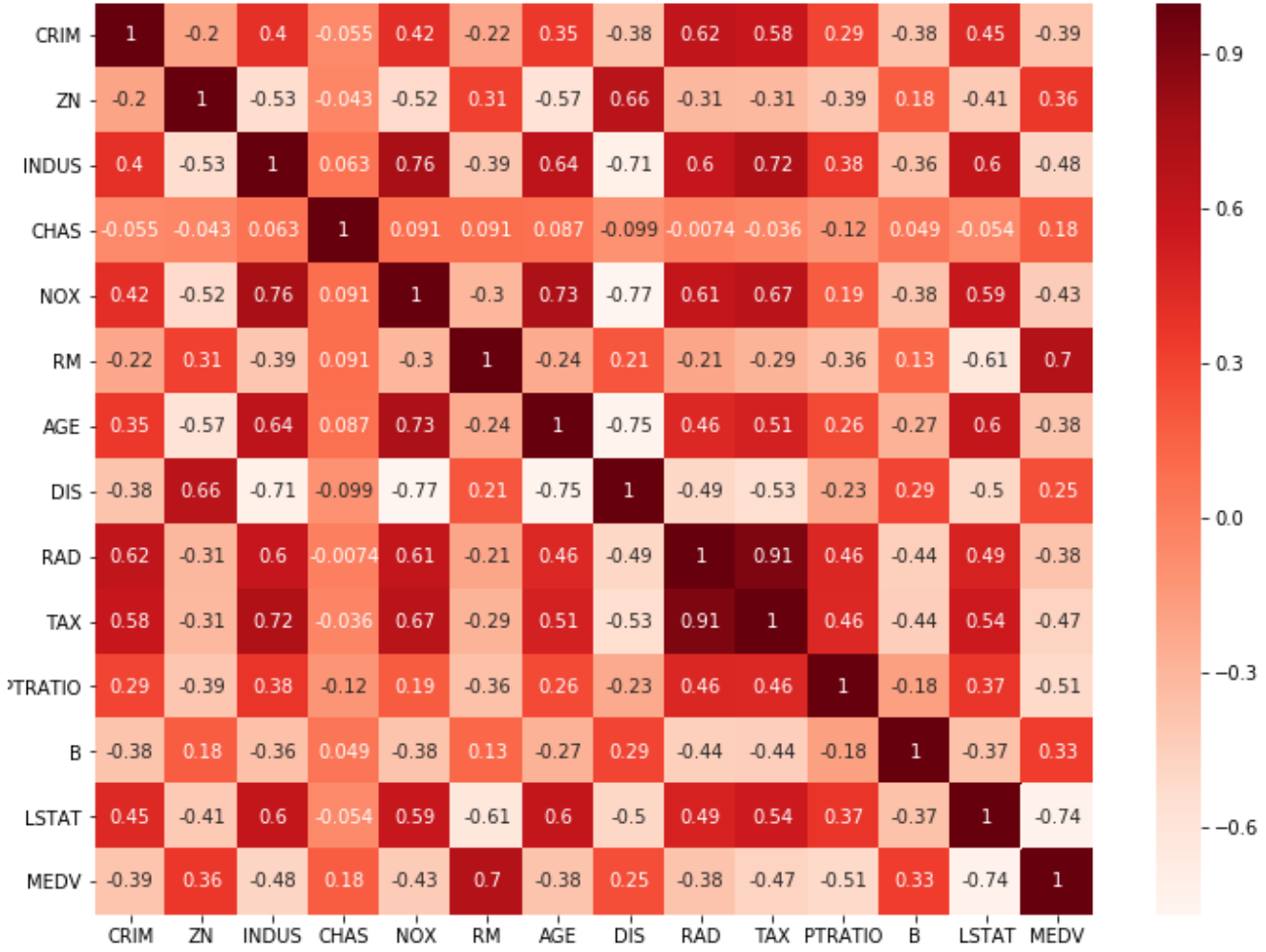
Redundant data may occur when integrating multiple data sources.

- The same attribute may have different names in different databases
- One attribute may be a “derived” attribute in another table (i.e., electric power derived from voltage and current)

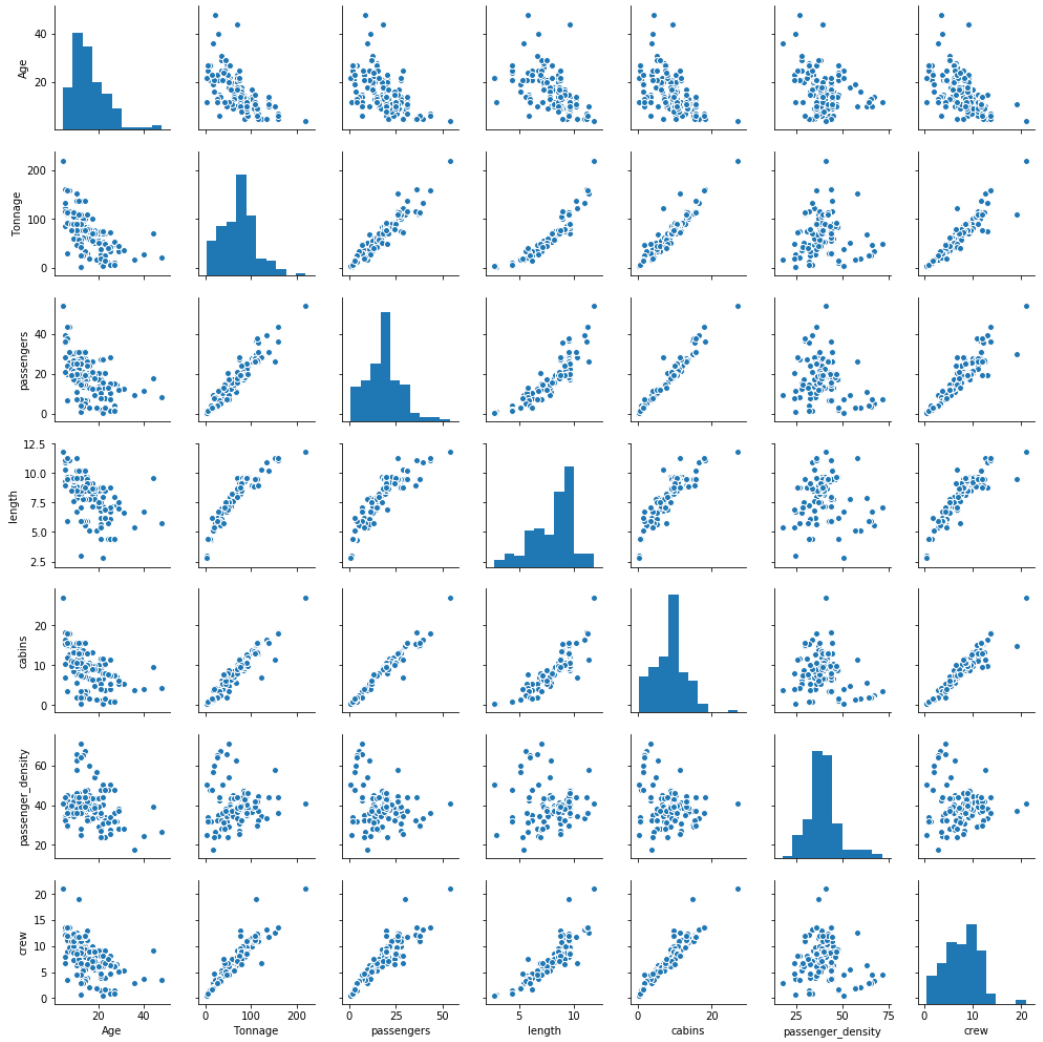
Correlation analysis may help in detecting redundancy.

In the context of ML, we look at the feature correlation matrix.

Handling redundancy



Handling redundancy



Data scaling

Data scaling refers to techniques for transforming the range of the data.

- Min-max normalization
- Z-score normalization
- Normalization by decimal scaling

Objective functions used by ML algorithm usually do not work properly without normalization.

Data scaling

Min-max normalization

Scaling a range of values to [0, 1] or [−1, 1]

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Scaling to an arbitrary range [a, b]

$$x' = a + \frac{(x - x_{min})(b - a)}{x_{max} - x_{min}}$$

Data scaling

Normalization by decimal scaling

Scaling a range of values to [0, 1] or [−1, 1]

$$x' = \frac{x}{10^i}$$

where i is the smallest integer such that $|x'| < 1$

Data scaling

Mean normalization

$$x' = \frac{x - \bar{x}}{x_{max} - x_{min}}$$

Standardization

$$x' = \frac{x - \bar{x}}{s}$$

Ensures that each feature in the data has zero-mean and unit-variance.

Data reduction

Motivation

- Storage space (some datasets consist of terabytes of data)
- Algorithm efficiency and improved modeling results

Reduction concept

- Obtain reduced representation while preserving information
- Preserve ability to learn from data

Strategies

Compression, numerosity reduction, dimensionality reduction, discretization

Principal component analysis

Unsupervised method - ignores class labels, invented in 1901 by **Karl Pearson**

Can reveal the structure of the data in a way that best explains the variance

- Used as a tool in exploratory data analysis
- Used as a preprocessing step before finding predictive models
- Often used to visualize distance and relatedness

Project a feature space onto a smaller subspace that represents the data well, by means of a linear transformation

Principal component analysis

Operating principle

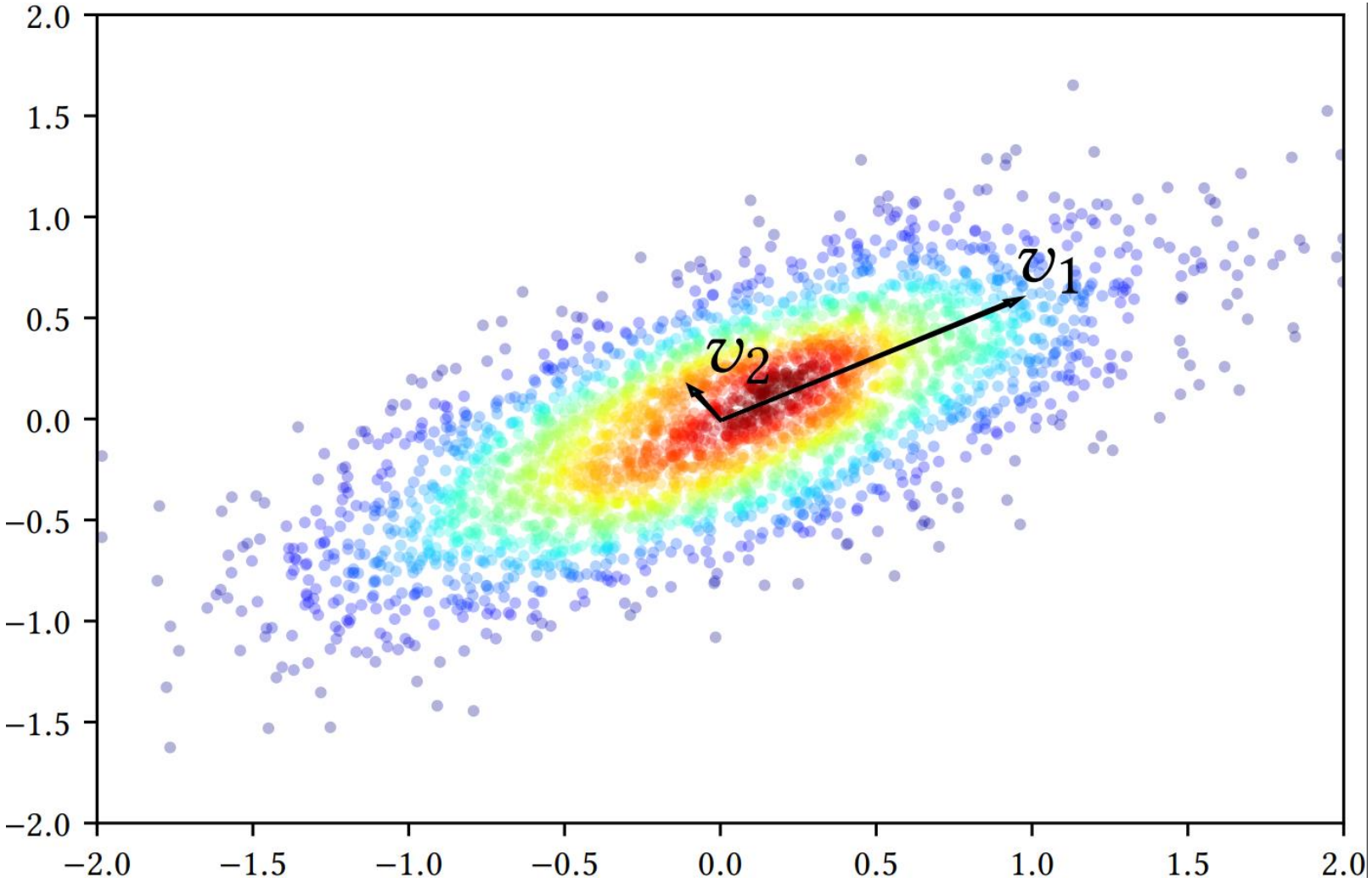
Eigenvalue decomposition of the data covariance or correlation matrix

- Eigenvectors are vectors which are fixed in direction under a given linear transformation
- The scaling factors of these eigenvectors are called the eigenvalues

Suppose we have a set of data with a certain distribution

- Eigenvectors v_i tell us the orientation of the distribution
- Eigenvalues λ_i tell the amount of variance in each dimension

Principal component analysis



Principal component analysis

Step-by-step guide

1. Standardize the data (zero-mean, unit-variance)
2. Calculate *eigenvectors* and *eigenvalues*
3. Choose *k* principal components based on the *k* largest *eigenvalues*
4. Construct projection matrix W from the selected *k eigenvectors*
5. Project original feature space X using W to obtain *k*-dimensional feature subspace Y

$$\overbrace{\mathbf{Y}}^{\text{Projected data}} = \underbrace{\mathbf{X}}_{\text{Original data}} \cdot \overbrace{\mathbf{W}}^{\text{Projection matrix}}$$

PCA Example

R. A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, 1936

- Data describing morphologic variation of Iris flowers
- Three related species (class labels): setosa, versicolor, virginica
- Four features: sepal length and width, petal length and width (in cm)

setosa

versicolor

virginica

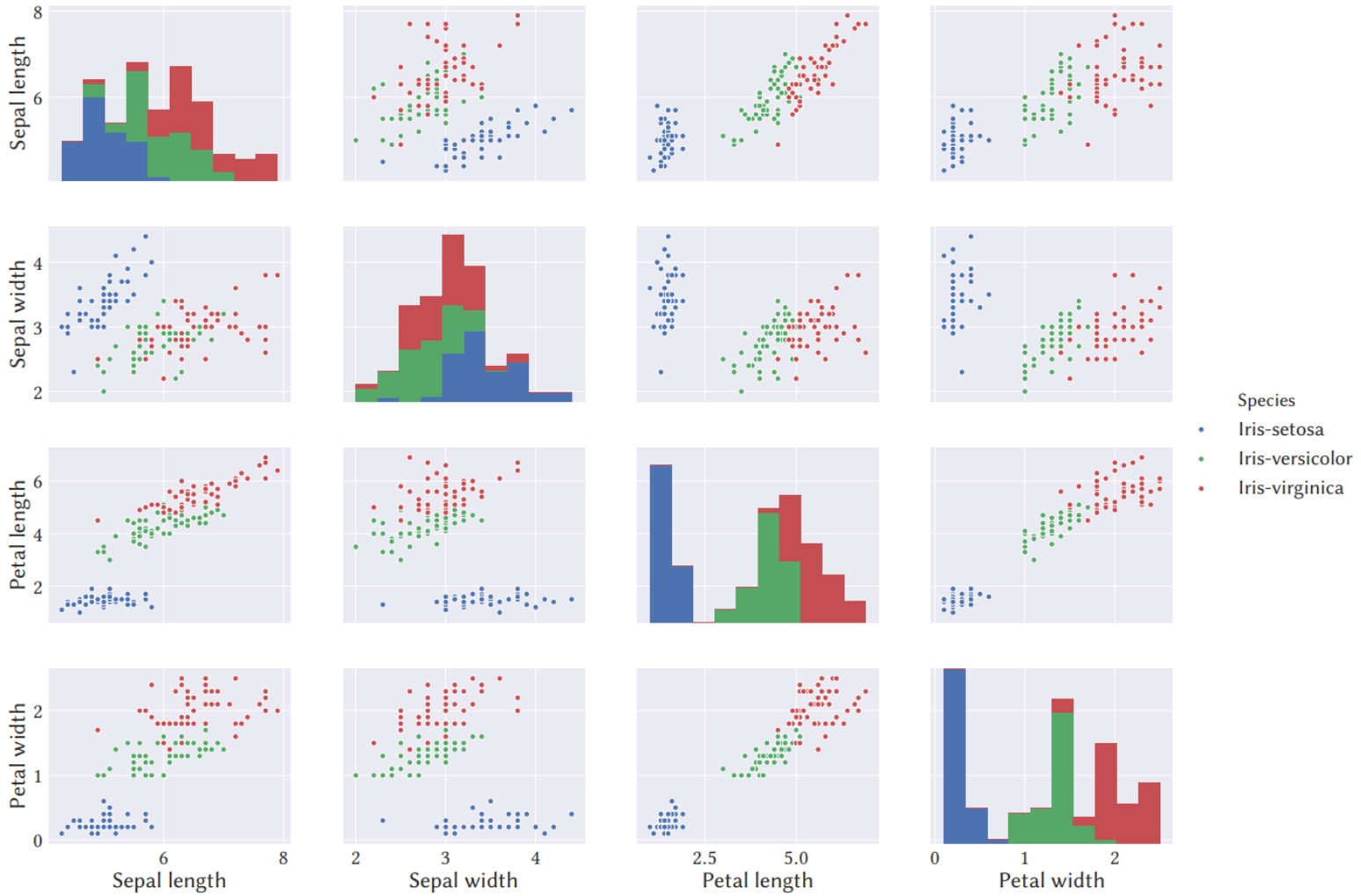
PCA Example

Input data X

$$\mathbf{x}^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \text{sepal length} \\ \text{sepal width} \\ \text{petal length} \\ \text{petal width} \end{bmatrix}$$

- 150 instances (50 for each class), no missing values
- Predicted attribute: species

PCA Example


Sepal length
Sepal width
Petal length
Petal width
Species
Iris-setosa
Iris-versicolor
Iris-virginica

PCA Example

1. **Standardization (Z-score normalization)**

$$x' = \frac{x - \bar{x}}{s}$$

- important step for many ML algorithms (e.g., k-means, KNN, SVM, LDA)
- when in doubt, standardize

PCA Example

2. Eigenvectors and eigenvalues

- Compute covariance matrix Σ from data
- Eigendecomposition of covariance matrix Σ
  (also possible to use correlation matrix)
- In practice, we prefer singular value decomposition (more efficient)

Four features (four dimensions) → four eigen vector/value pairs

Values represent the amount of variance explained by each principal component

PCA Example

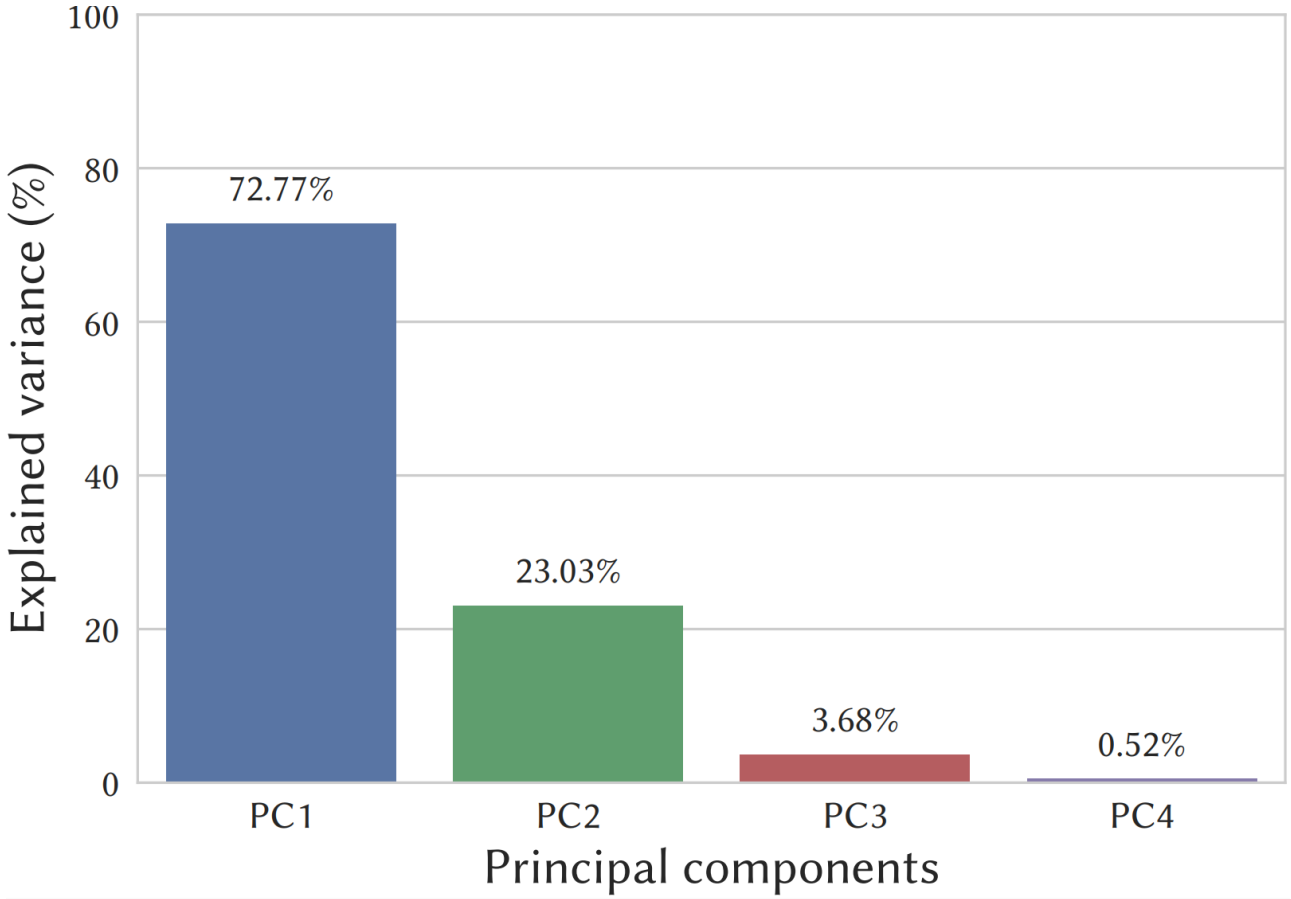
2. Eigenvectors and eigenvalues

$$\Sigma = \begin{bmatrix} 1.0067 & -0.1184 & 0.8776 & 0.8234 \\ -0.1184 & 1.0067 & -0.4313 & -0.3686 \\ 0.8776 & -0.4313 & 1.0067 & 0.9693 \\ 0.8234 & -0.3686 & 0.9693 & 1.0067 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.5211 & -0.3774 & -0.7196 & 0.2613 \\ -0.2693 & -0.9233 & 0.2444 & -0.1235 \\ 0.5804 & -0.0245 & 0.1421 & -0.8014 \\ 0.5649 & -0.0669 & 0.6343 & 0.5236 \end{bmatrix}, \lambda = \begin{bmatrix} 2.9381 \\ 0.9202 \\ 0.1477 \\ 0.0209 \end{bmatrix}$$

PCA Example

3. Choose principal components



PCA Example

4. Construct the projection matrix

We keep the first two *eigenvectors* corresponding to the two principal components, and we obtain

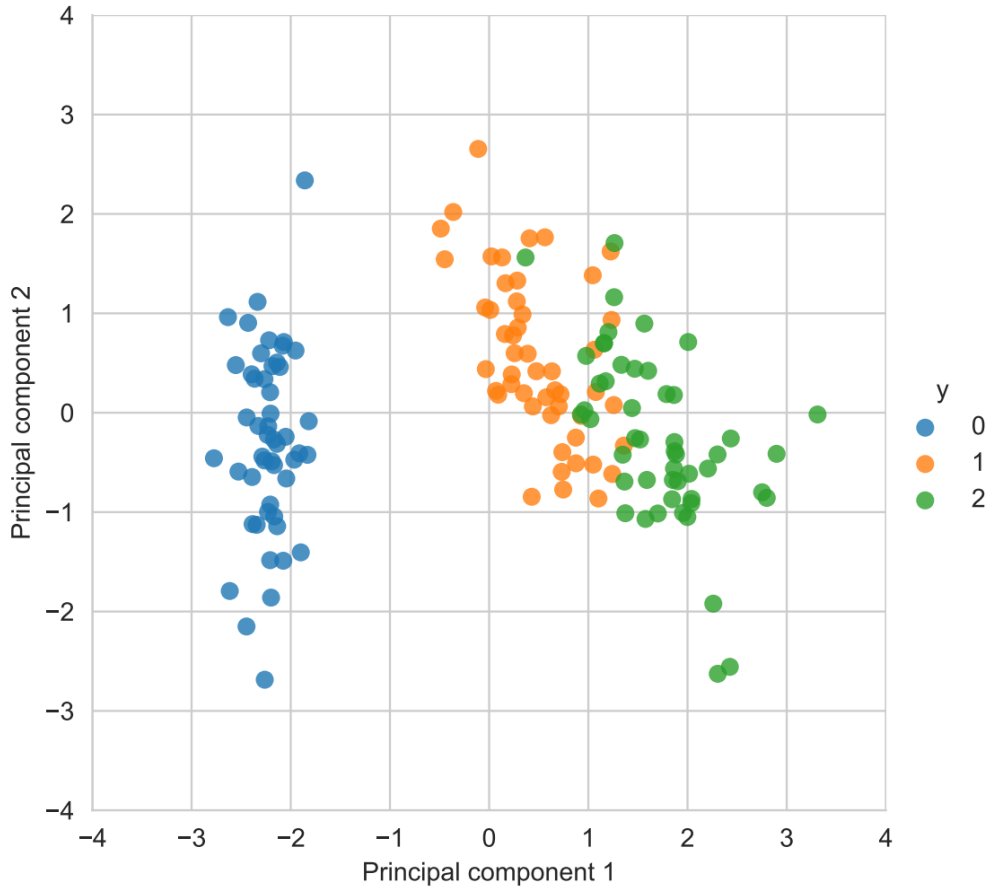
$$\boldsymbol{W} = \begin{bmatrix} 0.5211 & -0.3774 \\ -0.2693 & -0.9233 \\ 0.5804 & -0.0245 \\ 0.5649 & -0.0669 \end{bmatrix}$$

5. Project the feature space

$$\boldsymbol{Y} = \boldsymbol{XW}$$

PCA Example

Samples scatterplot in the lower-dimensional (projected) feature space



PCA Example

How to reconstruct the original features?

Use $\boldsymbol{W^T}$ to map the data back to the original dimensions

$$\widehat{\boldsymbol{X}} = \boldsymbol{Y}\boldsymbol{W^T}$$

Due to this feature, PCA can also be used for compression

PCA Example

Image compression

Original image

Reconstruction from 50 PC