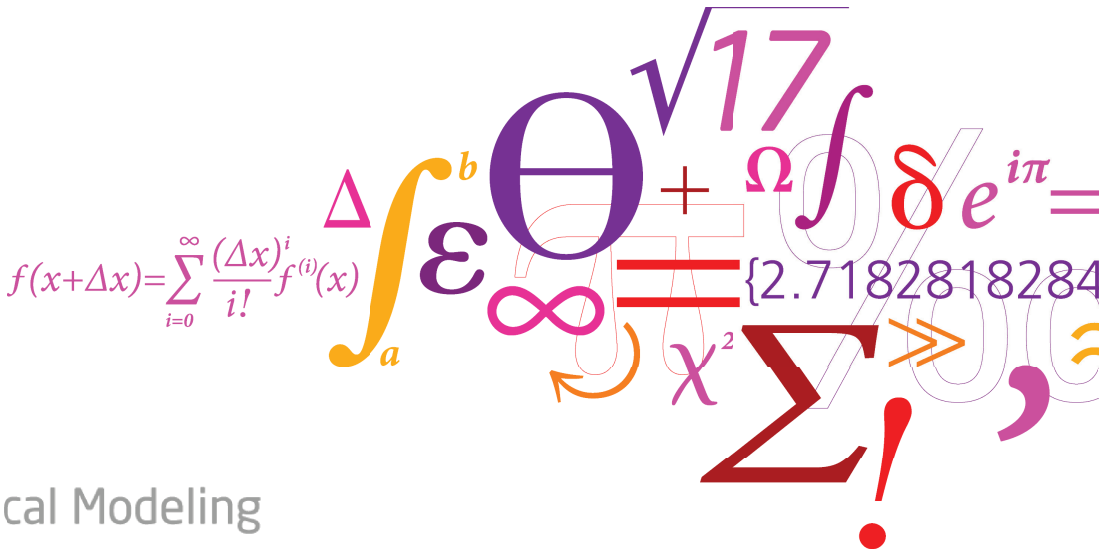
02450 Introduction to machine learning and data modeling

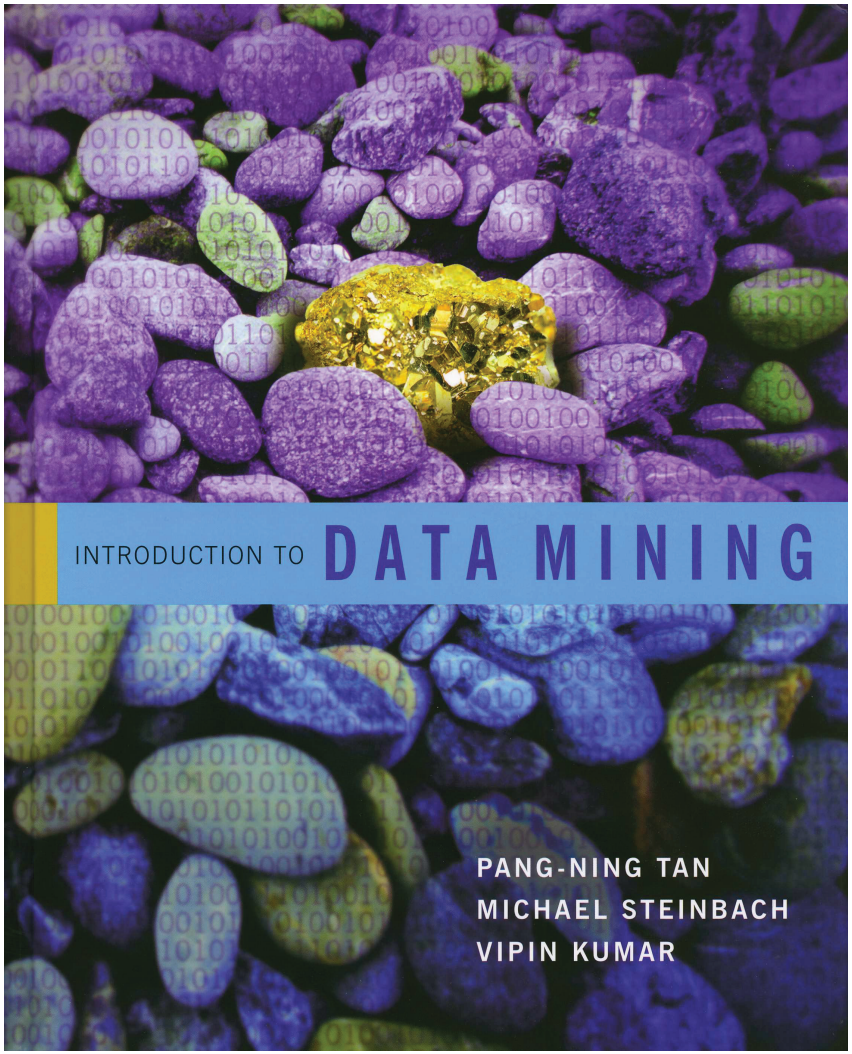


DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"

Section 2.4 + 3.1-3.2 + C1-C2



Group(s) of the day:

Jan Selliah
Carsten Nilsson
Mette Vestergaard Lauridsen
Anders Vinther Olsen
Mikkel Liisborg hansen
Nanna Thorning-Schmidt

Lecture schedule

1. Introduction
   (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
   (Tan 2.1-2.3 + B1 (+ A))
3. **Measures of similarity and summary statistics**
   (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
   (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
   (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
   (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
   (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers
   (Tan 5.6-5.8)

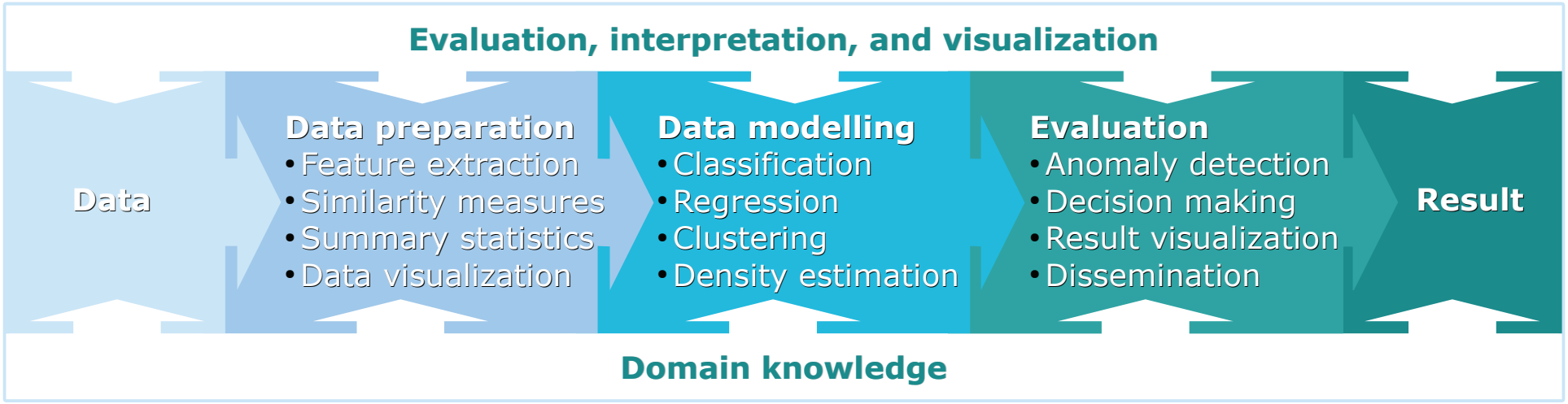
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
   (Tan 8.1-8.3+8.5.7)
10. Mixture models and association mining
    (Tan 9.2.2 + 6.1-6.3)
11. Density estimation and anomaly detection
    (Tan 10.1-10.4)

Machine learning and data modelling in practice

12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



Todays learning objectives:

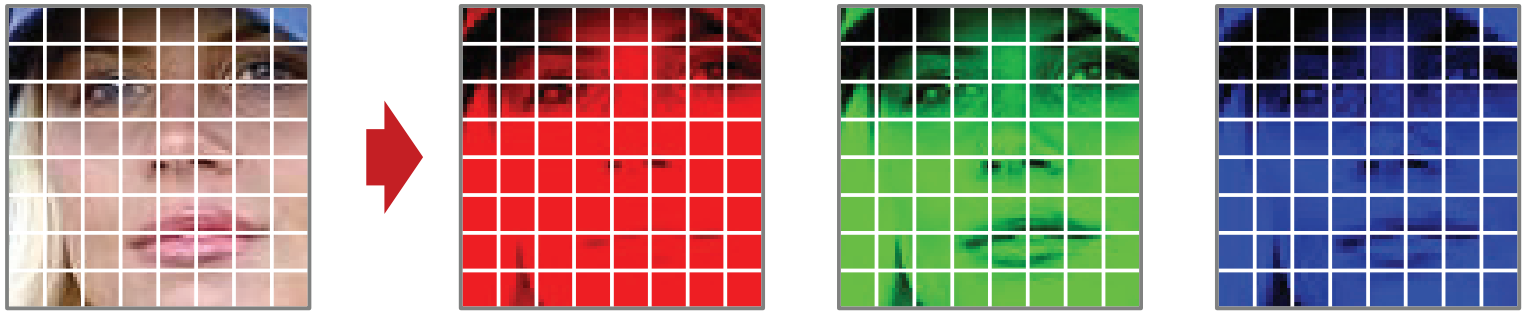
Be able to calculate various measures of similarity and dissimilarity.
Understand how various summmary statistics are calculated and can be interpreted
Explain and apply Bayes theorem
Understand the normal and multi-variate normal distribution and the role of the covariance matrix

Example: Principal component analysis of images

- 1000 images, 86 x 86 pixels, 3 RGB intensities

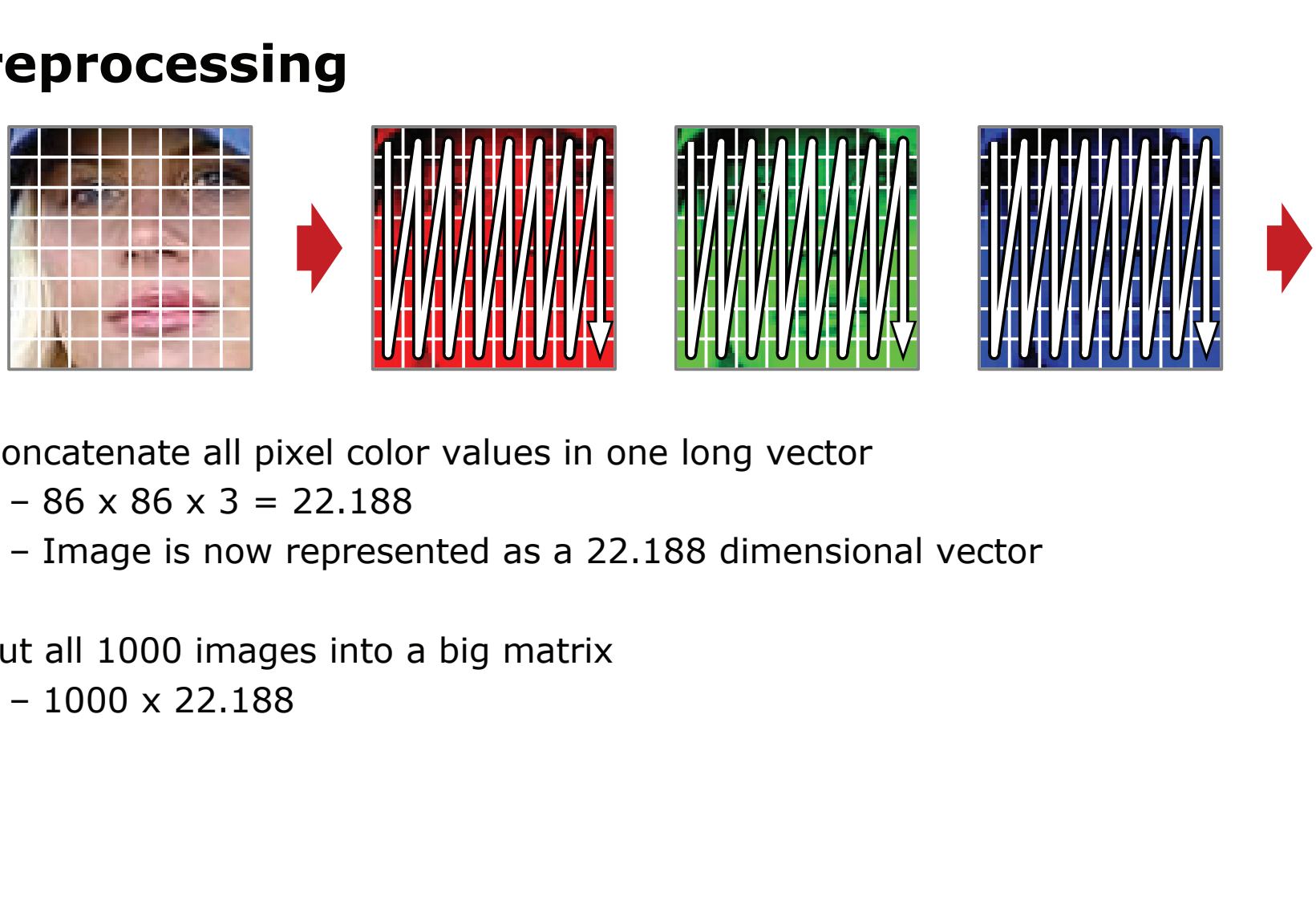
Tamara Berg "Faces in the wild"

Preprocessing



- Each image
  - 86 x 86 pixels
  - 3 RGB intensities
- Split image into red, green, and blue color channels

Preprocessing



- Concatenate all pixel color values in one long vector
  - 86 x 86 x 3 = 22.188
  - Image is now represented as a 22.188 dimensional vector

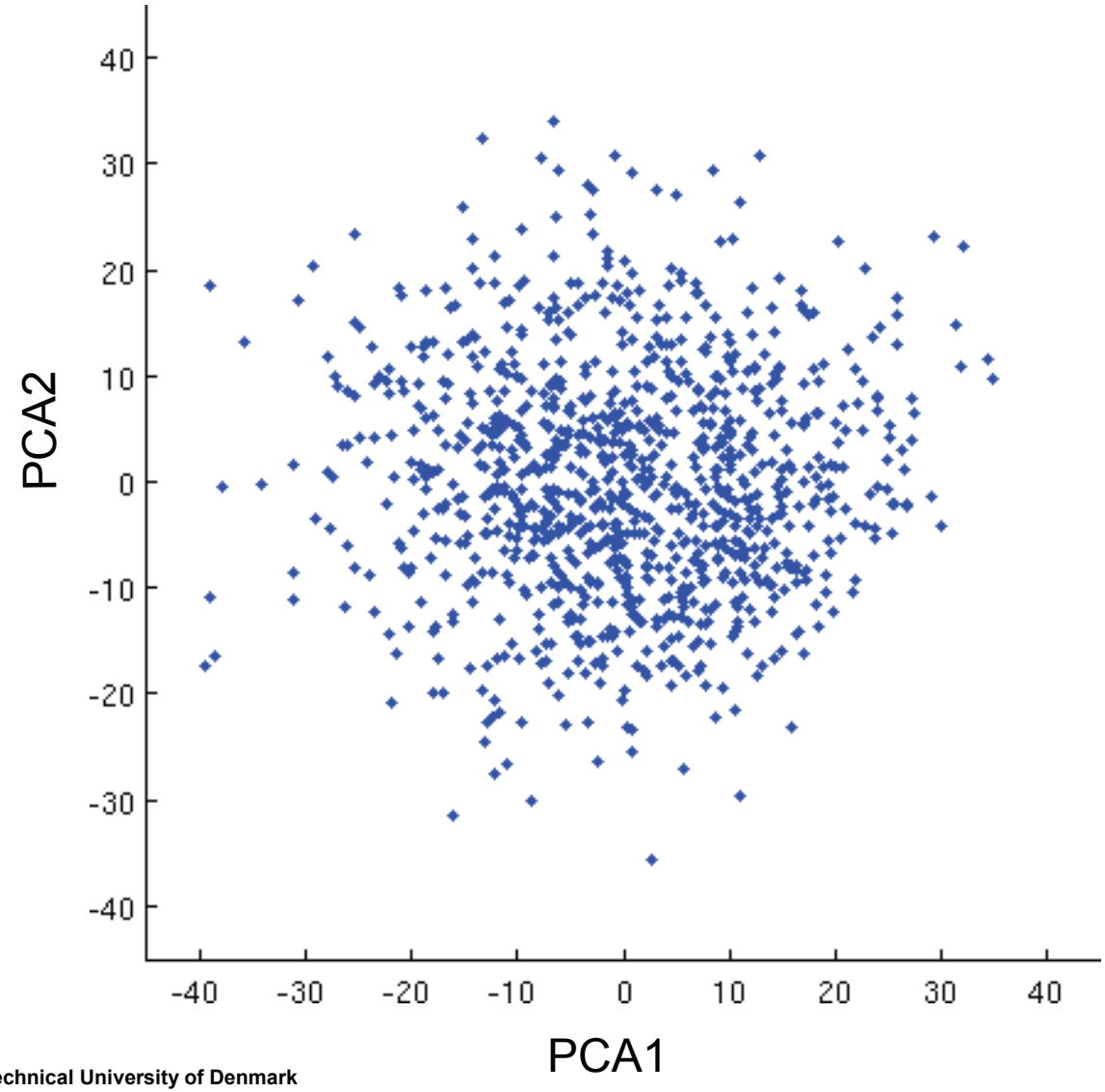
- Put all 1000 images into a big matrix
  - 1000 x 22.188

Principal component analysis (PCA)

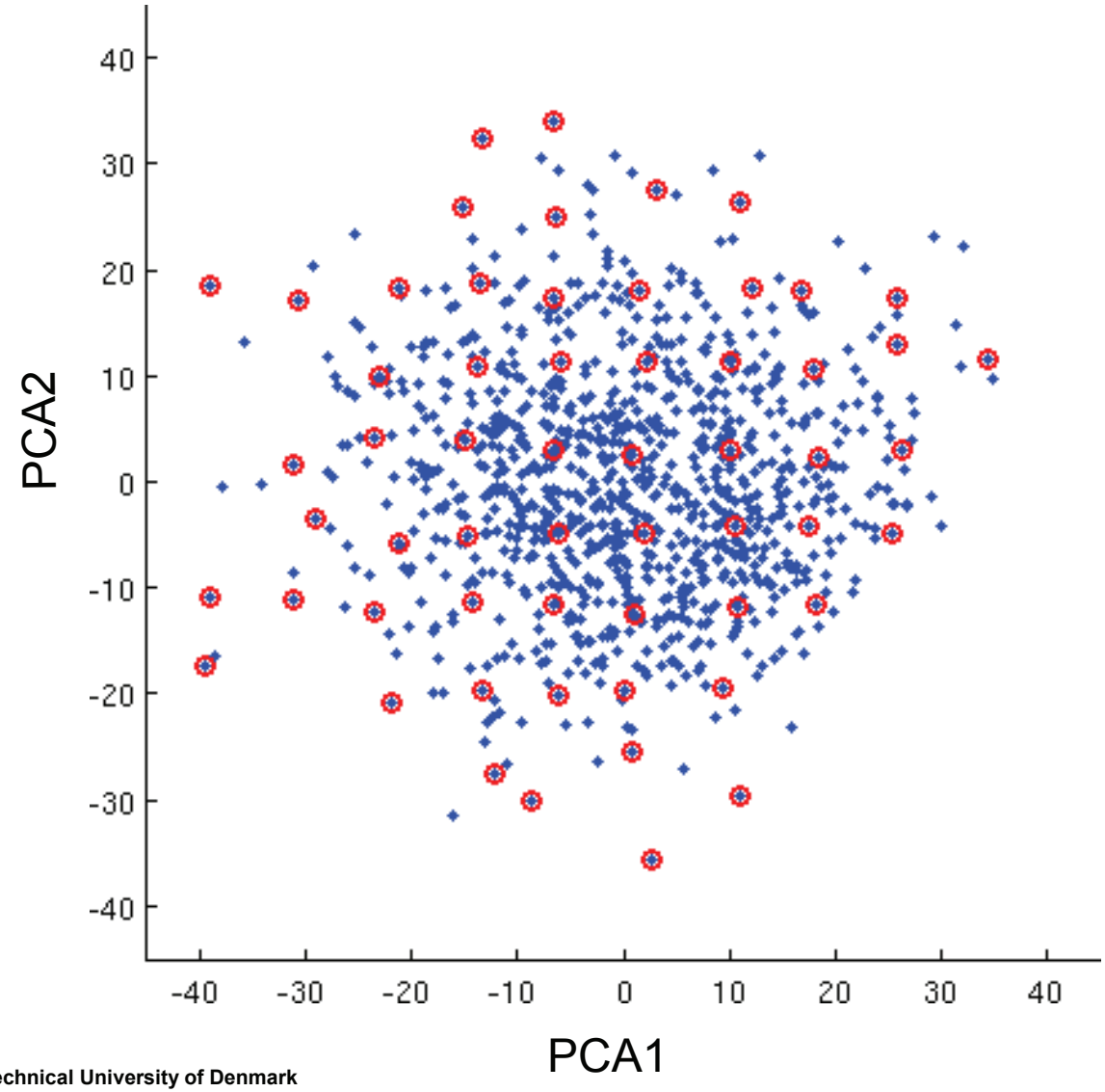
1. **Subtract the mean**
2. **Compute the singular value decomposition (SVD)**
   - Orthogonal linear transformation
   - Transforms data to a new coordinate system
     - Greatest variance along the first axis
     - Second greatest variance along the second axis
     - Etc.

- **Plot data in the transformed coordinate system**
  - Corresponds to looking at data from an angle where it is most spread out

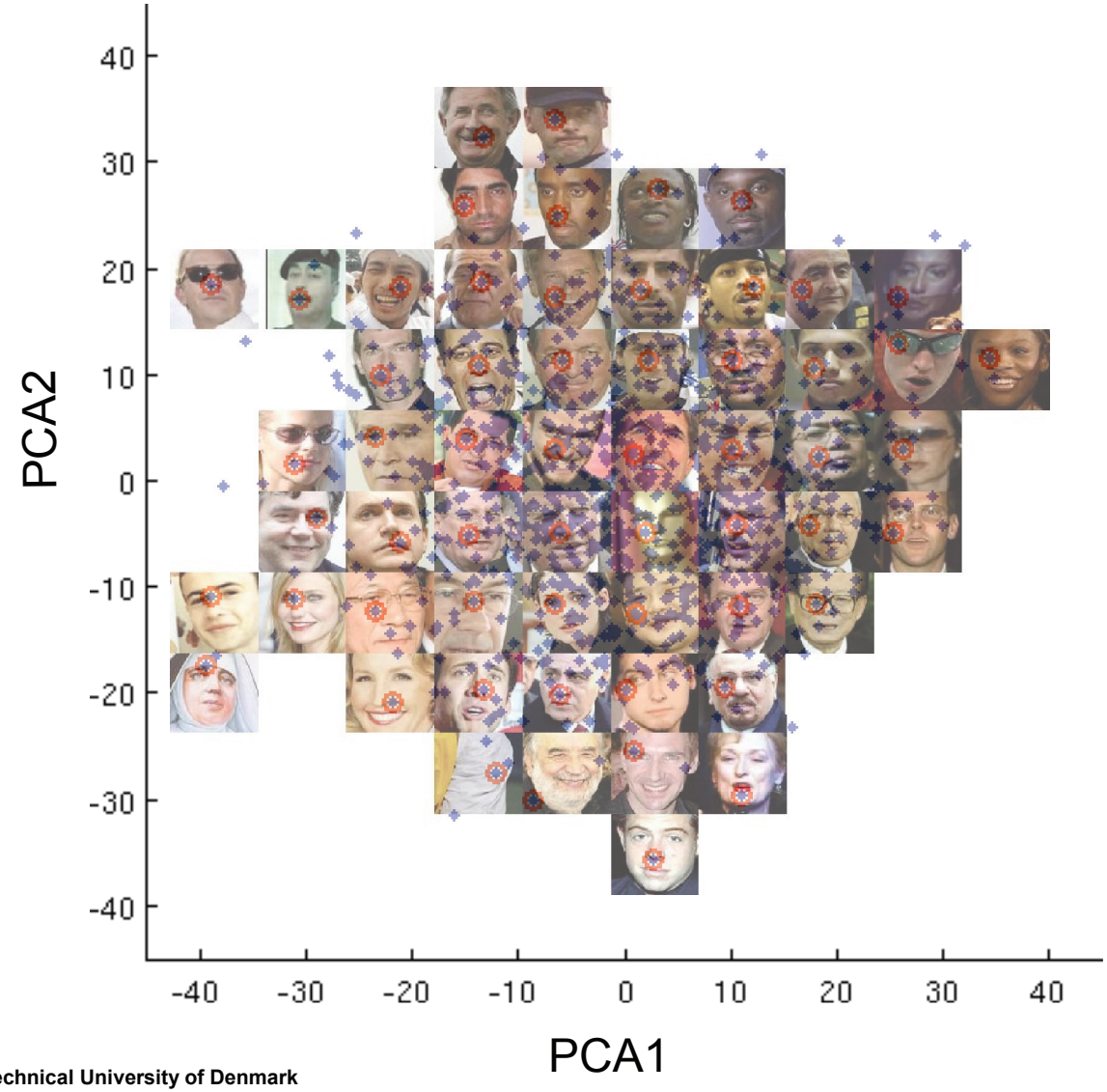
PCA of face images

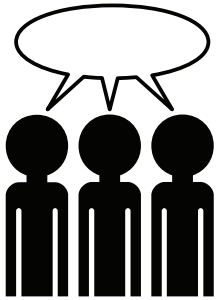


PCA of face images

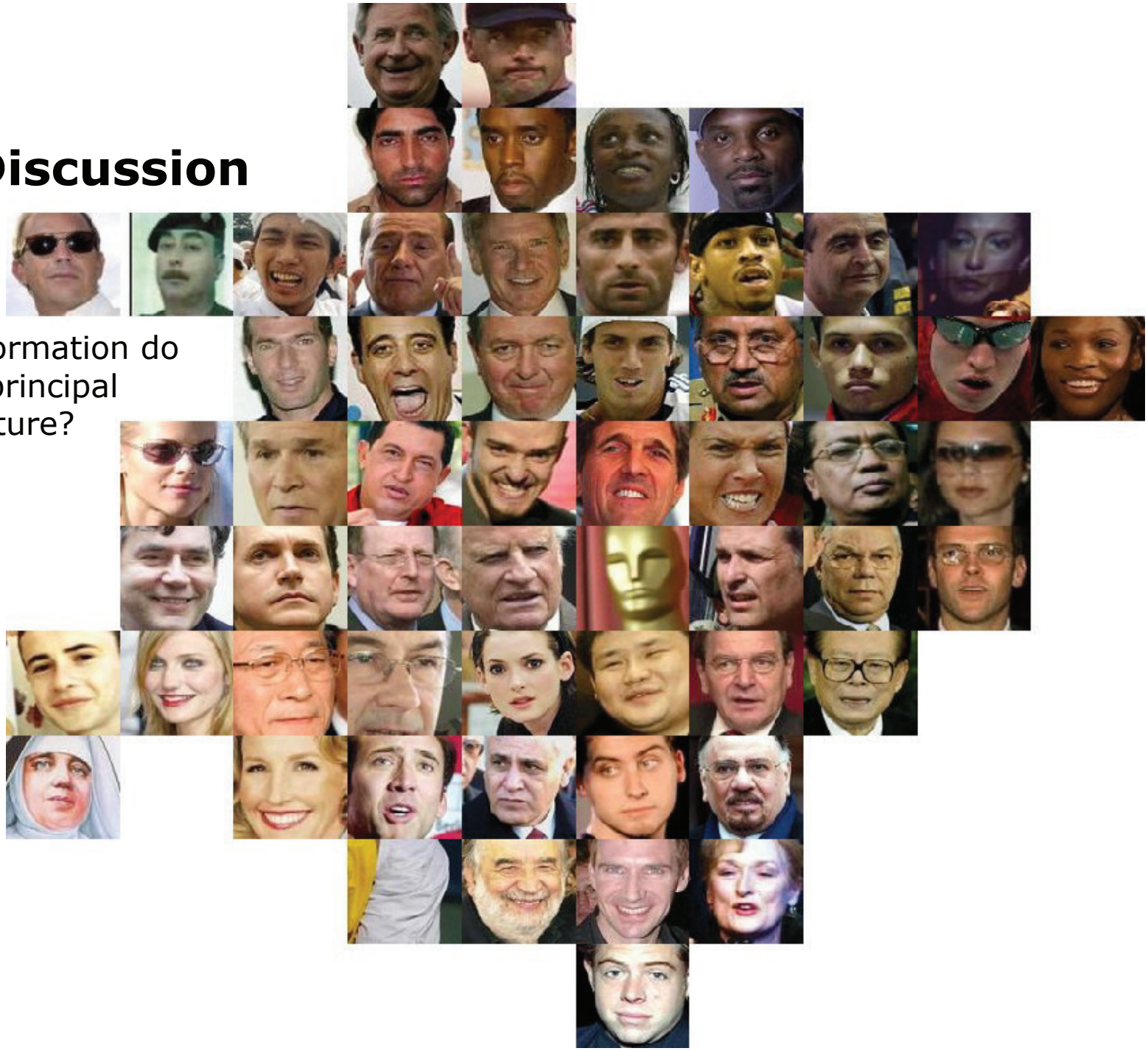


PCA of face images

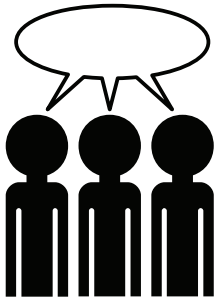




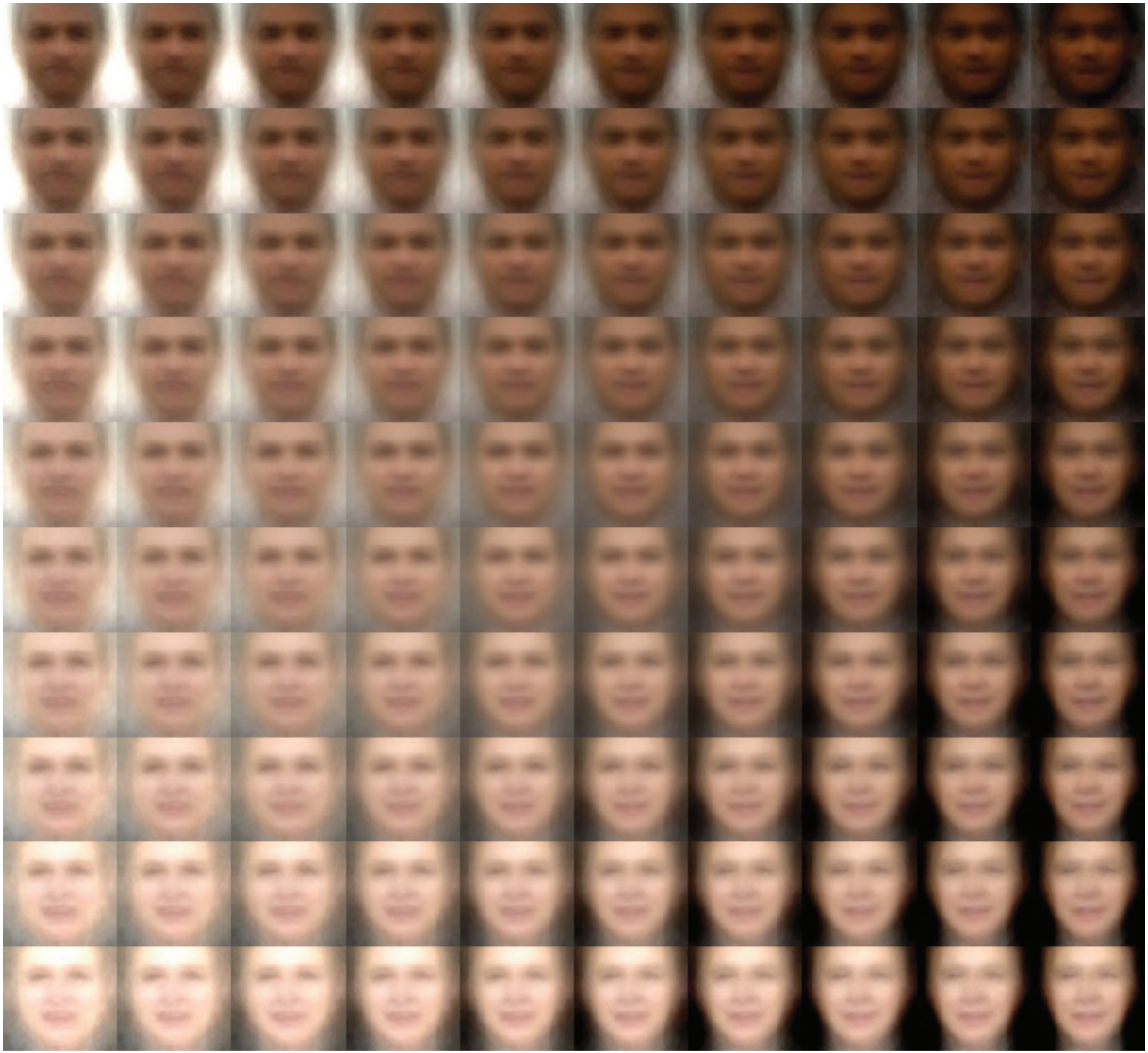
Discussion

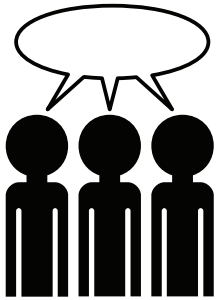


- What information do the two principal axes capture?

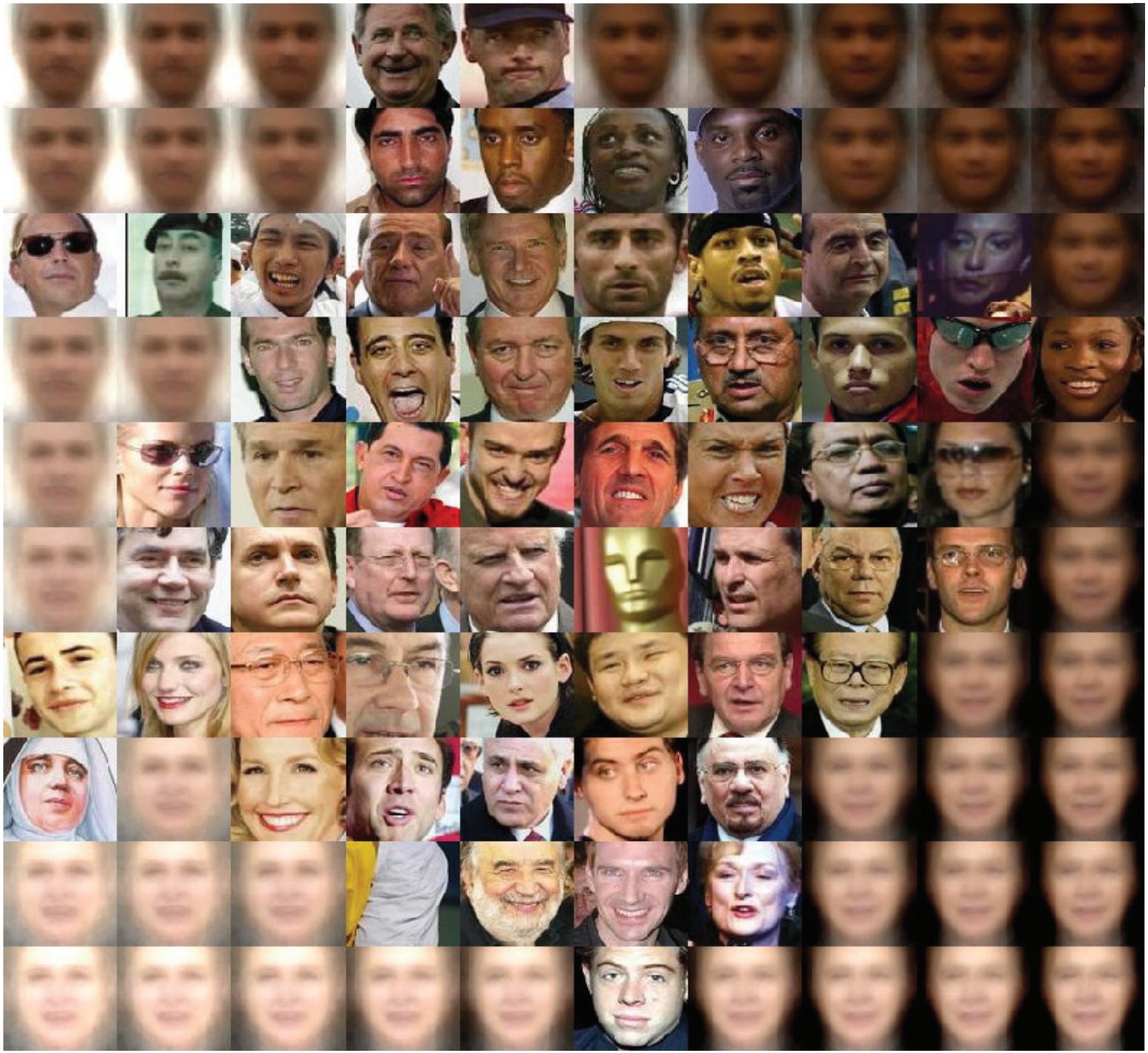


- What information do the two principal axes capture?





- What information do the two principal axes capture?

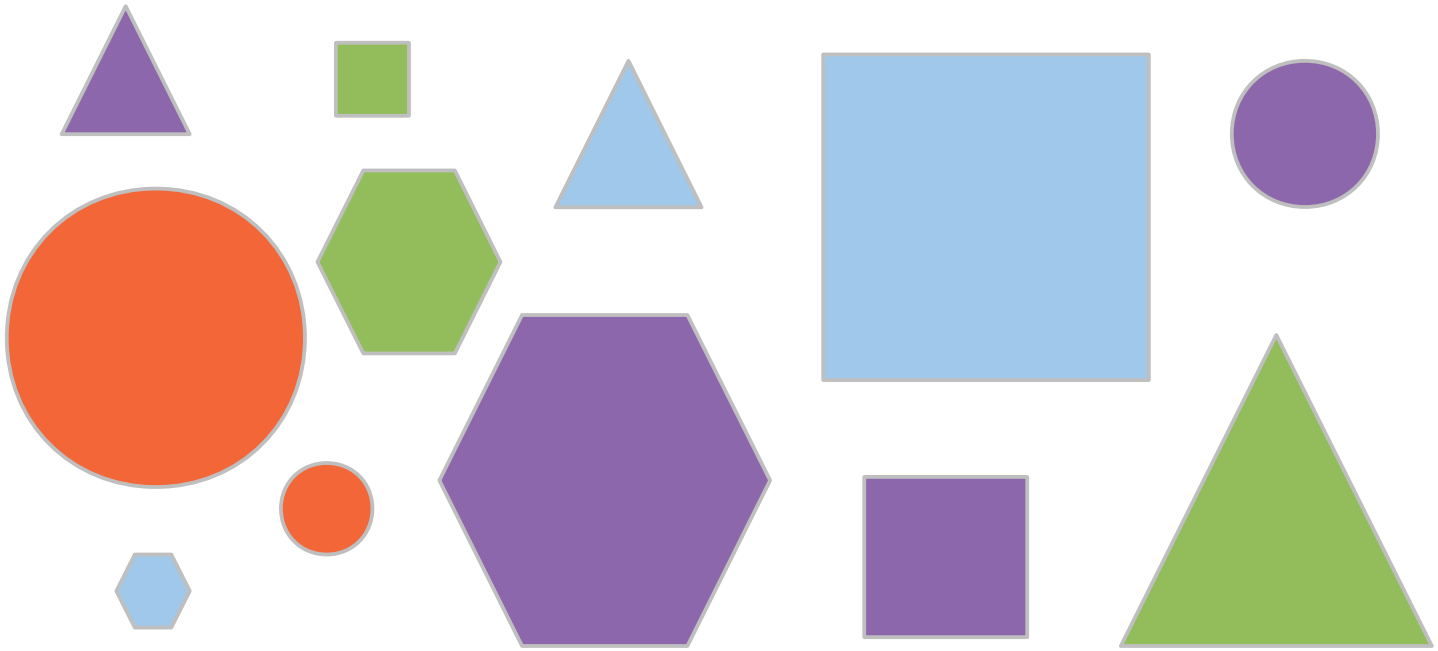


Similarity / Dissimilarity

- **Definition**
  - A numerical measure of *how alike/different* two data objects are
  - Often defined on the interval [0,1]

- **Similarity / dissimilarity between two data objects**

$$s(\boldsymbol{x}, \boldsymbol{y}), \quad d(\boldsymbol{x}, \boldsymbol{y})$$



Dissimilarity measures

- **Euclidean distance (2-norm)**

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

- **Minkowski distance (p-norm)**

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{n=1}^{N} |x_n - y_n|^p \right)^{1/p}$$

Properties of dissimilarity measures

- **Positivity**

$$d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$$
$$d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$$

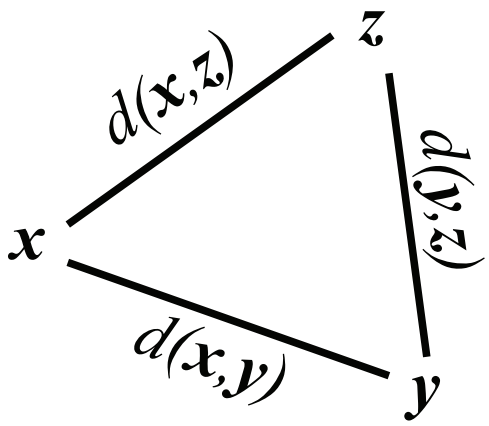
- **Symmetry**

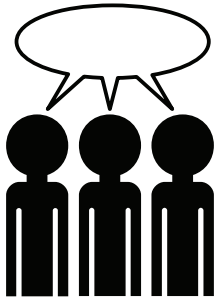
$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$$

- **Triangle inequality**

$$d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$$

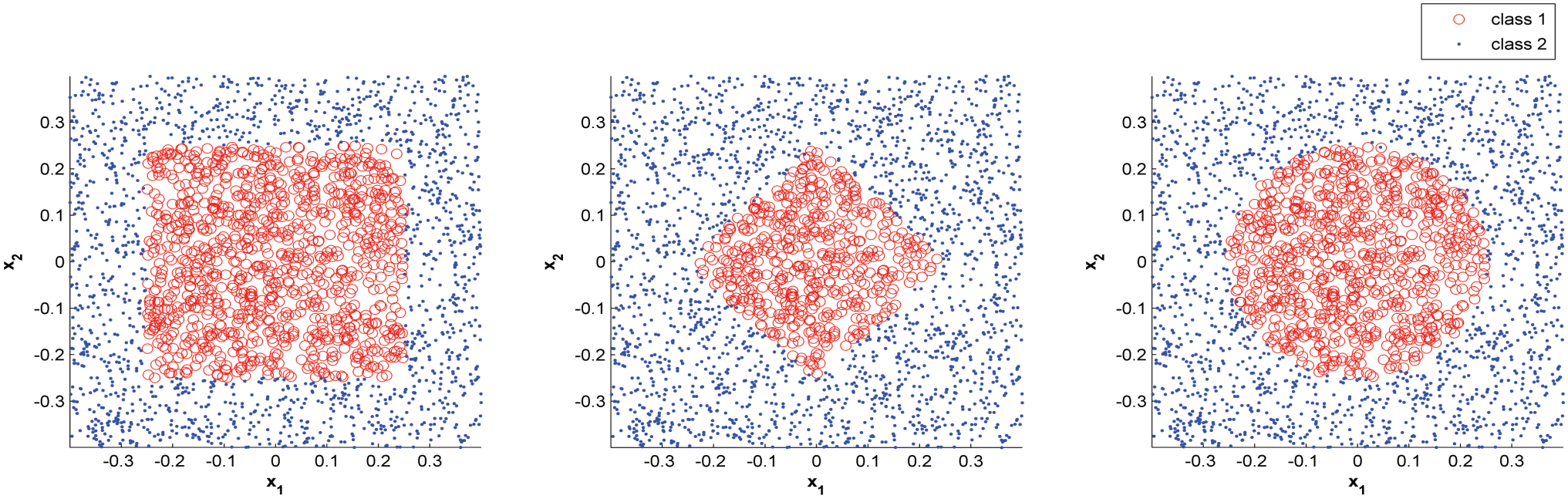
- Measures that satisfy all these properties: **Metrics**





Minkowski distance

Which Minkowski distance (p-norm) can be used to separate the two classes, by measuring the distance to the origo (0,0)?



$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{n=1}^{N} |x_n - y_n|^p \right)^{1/p}$$

Binary similarity measures

- **Simple matching coefficient (SMC)**
  - Symmetric: Counts present and absent attributes equally

$$\mathrm{SMC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{00} + f_{11}}{K}$$

- **Jaccard coefficient**
  - Asymmetric: Counts only present attributes

$$\mathrm{J}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{11}}{K - f_{00}}$$

K : Total number of attributes

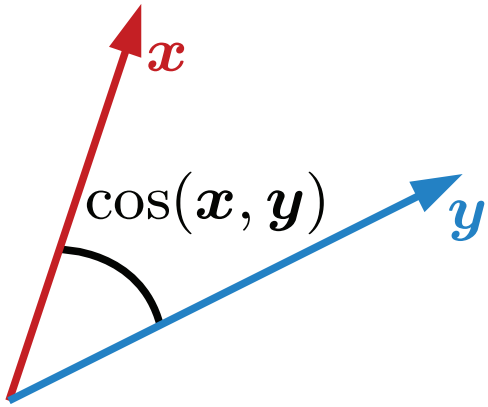
f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

Continuous similarity measures

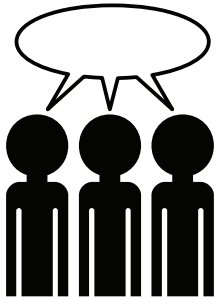
- **Cosine similarity**

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}$$



- **Extended Jaccard coefficient**

$$EJ(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - \boldsymbol{x}^\top \boldsymbol{y}}$$



Calculate the SMC, Jaccard, Cosine and Extended Jaccard similarity between customer 1 and customer 2 in the market basket data below.

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	0	1	1	0	1

$$\mathrm{SMC}(\boldsymbol{x},\boldsymbol{y}) = \frac{f_{00}+f_{11}}{K}$$

$$\mathrm{J}(\boldsymbol{x},\boldsymbol{y}) = \frac{f_{11}}{K-f_{00}}$$

$$\cos(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}$$

$$EJ(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|_2^2+\|\boldsymbol{y}\|_2^2-\boldsymbol{x}^\top \boldsymbol{y}}$$

K : Total number of attributes

f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

Empirical statistics

- **Empirical mean**

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- **Empirical covariance**

$$\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

- **Empirical variance**

$$\mathrm{var}(\boldsymbol{x}) = \mathrm{cov}(\boldsymbol{x}, \boldsymbol{x}) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

- **Empirical standard deviation**

$$\mathrm{std}(\boldsymbol{x}) = \sqrt{\mathrm{var}(\boldsymbol{x})}$$

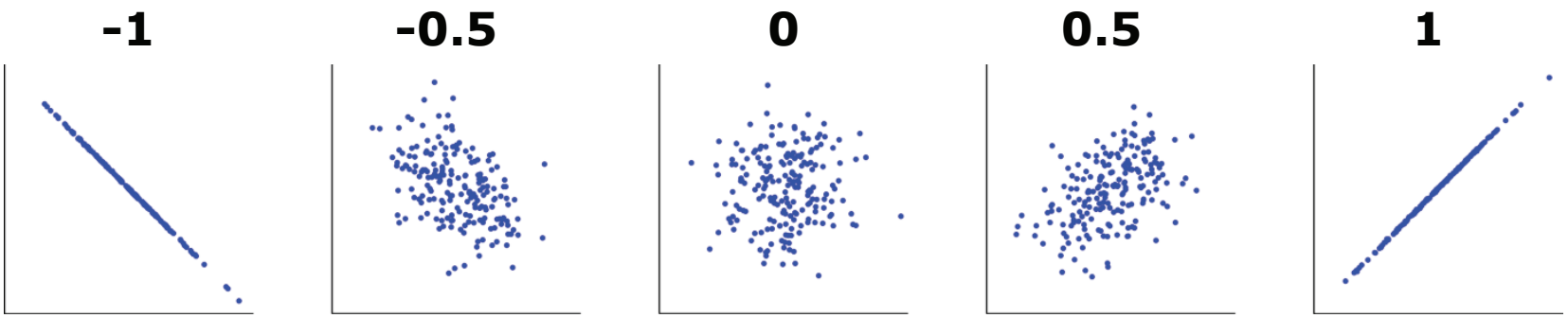
Correlation

- **Measure of linear relation**

$$\mathrm{corr}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{std}(\boldsymbol{x})\mathrm{std}(\boldsymbol{y})}$$

  – A correlation of **1** or **-1** means there is a perfect linear relation

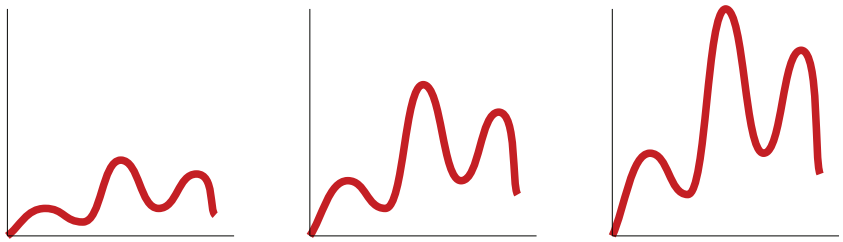
$$x_k = ay_k + b$$



Invariance

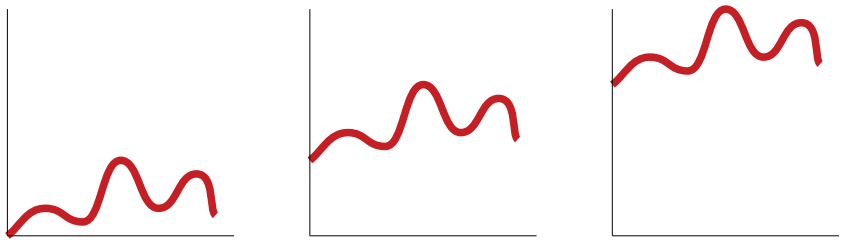
- **Scale**

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\alpha \boldsymbol{x}, \boldsymbol{y})$$



- **Translation**

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\beta + \boldsymbol{x}, \boldsymbol{y})$$

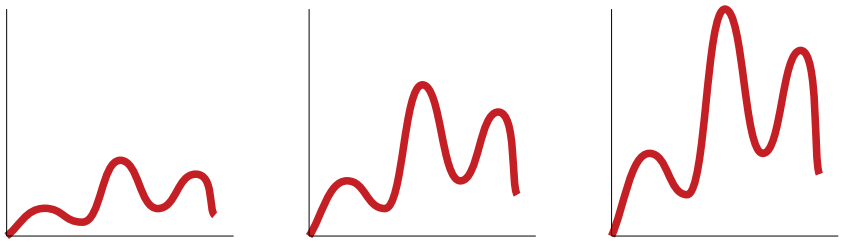


Discussion

- When would a **scale invariant** similarity measure be useful
  - Give an example
- When would a **translation invariant** similarity measure be useful
  - Give an example

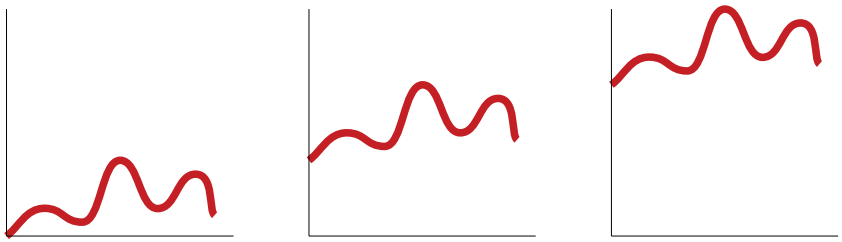
Scale invariance

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\alpha\boldsymbol{x}, \boldsymbol{y})$$

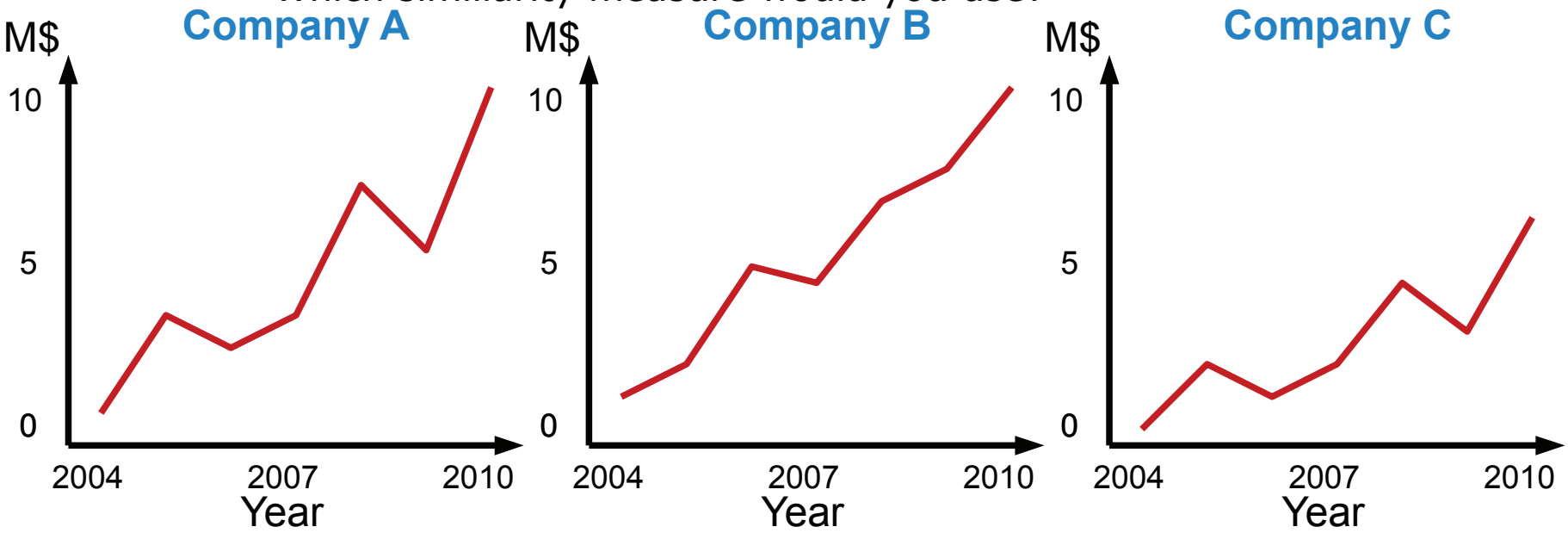


Translation invariance

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\beta + \boldsymbol{x}, \boldsymbol{y})$$



- Which of these three companies are similar and in what way?
  - Which similiarity measure would you use?



Issues in proximity calculation

- **What to do when attributes have different**
  - **Scale?**
    - Standardize attributes
    - Use scale invariant similarity measure
  - **Type?**
    - Compute similarities for each attribute and combine
  - **Importance?**
    - Compute a weighted similarity measure

Standardization

- **Attributes have different scales.**
  - Example:
    - **Number of children** ~ 0-5
    - **Age** ~ 0-100 years
    - **Annual income** ~ 0-50.000 €

- Unless we do something, **Annual income** will dominate
  - **Standardization**: Subtract mean and divide by standard deviation

$$x_k^* = \frac{x_k - \bar{x}_k}{\mathrm{std}(x_k)}$$

Combining heterogeneous attributes

- **Attributes have different type**
  - Example:
    - **Age:** Continuous
    - **Education:** Binary
      - Primary (yes/no)
      - Secondary (yes/no)
      - Tertiary (yes/no)

- Similarity measure must handle **continuous** and **binary** features
  - **Compute similarities for each attribute and combine**

$$s_{\text{Age}} = d_1(x_{\text{Age}}, y_{\text{Age}}) \qquad s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}(s_{\text{Age}} + s_{\text{Edu.}})$$

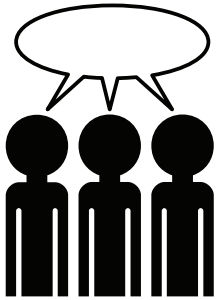
Weighting attributes by importance

- **Attributes have different importance**
  - Example:
    - **Age:** Very important
    - **Education:** Less important

- Similarity measure must take **importance** into account
  - Introduce **importance weights** for each attribute

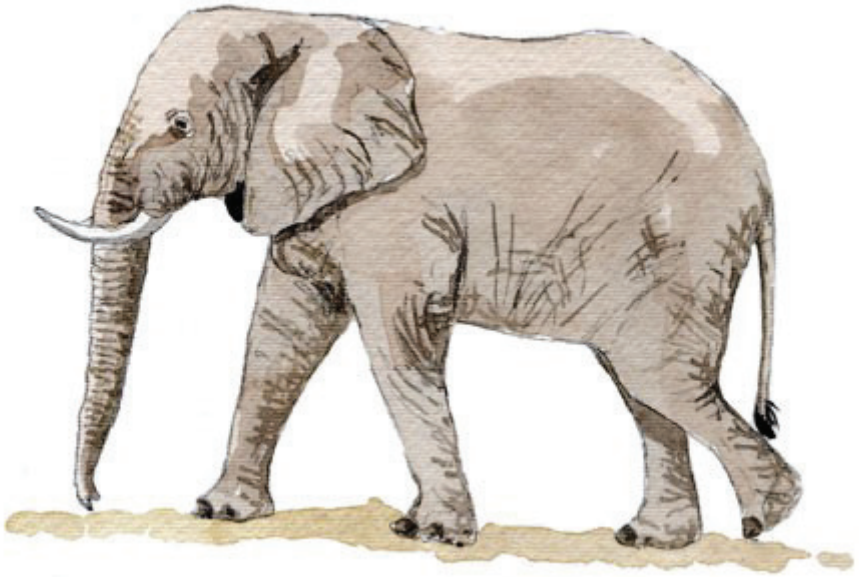
$$s_{\text{Age}} = d_1(x_{\text{Age}}, y_{\text{Age}}) \qquad s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

$$s(\boldsymbol{x}, \boldsymbol{y}) = 0.99 \cdot s_{\text{Age}} + 0.01 \cdot s_{\text{Edu.}}$$



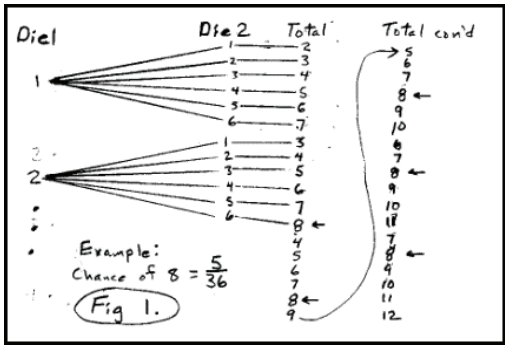
Discussion

- The following **attributes** are measured for a herd of elephants
  - Weight
  - Height
  - Tusk length
  - Trunk length
  - Ear area
  - Gender
- **Based on these measurements**
  - How would you evaluate how similar elephants are?
  - Justify your answer

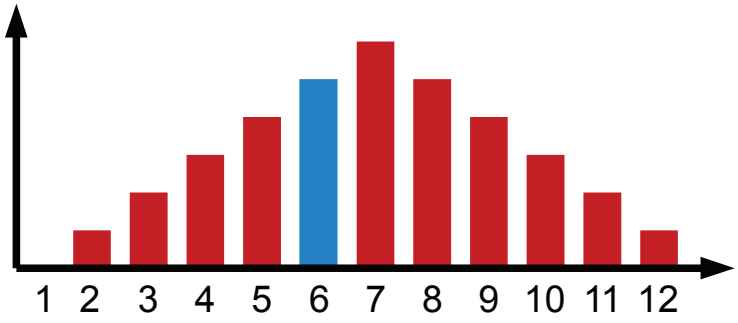


Probabilities

- **Discrete: Probability mass**
  - Example: The sum of two dice

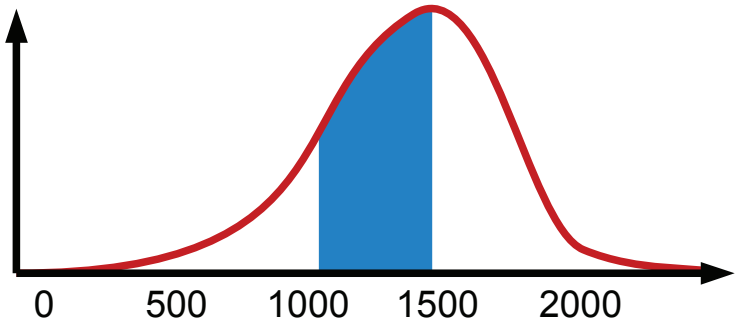


$$P(X = v)$$



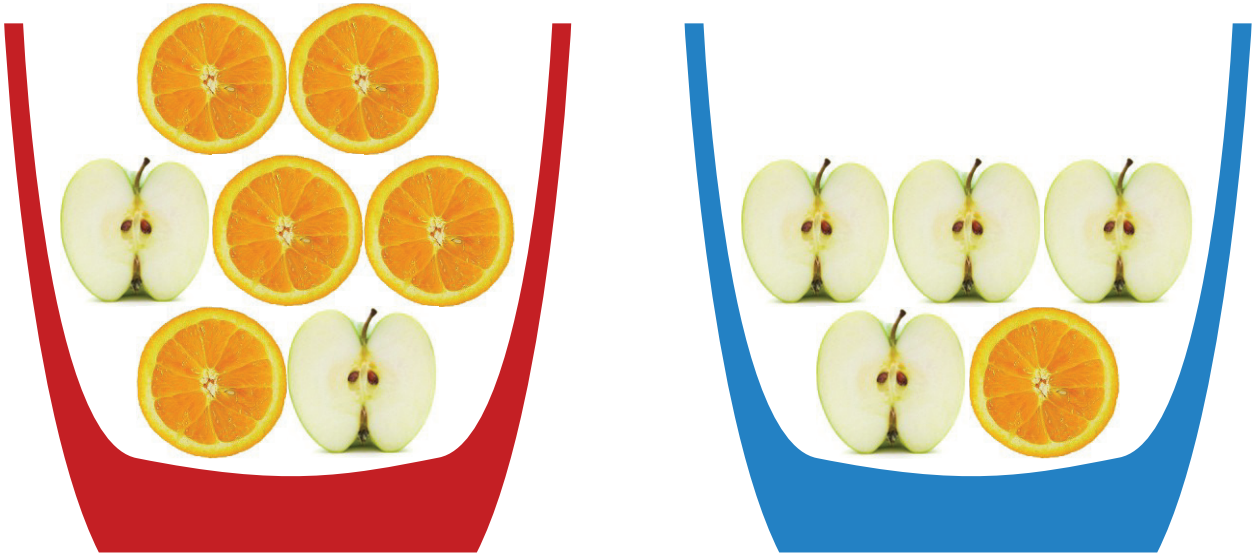
- **Continuous: Probability density**
  - Example: Lifetime of a light bulb

$$P(a \leq X \leq b) = \int_a^b p(X)\mathrm{d}X$$



Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Probabilities

- Basic rules of probability
  - Sum rule

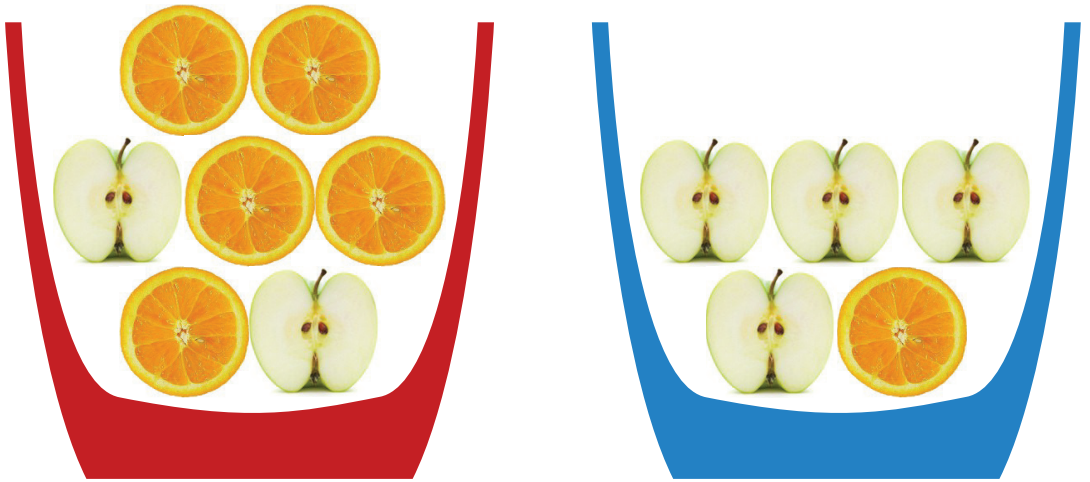
$$p(x) = \sum_y p(x, y)$$

  - Product rule

$$p(x, y) = p(x|y)p(y)$$

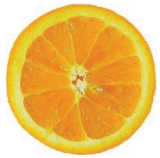
  - Bayes' rule

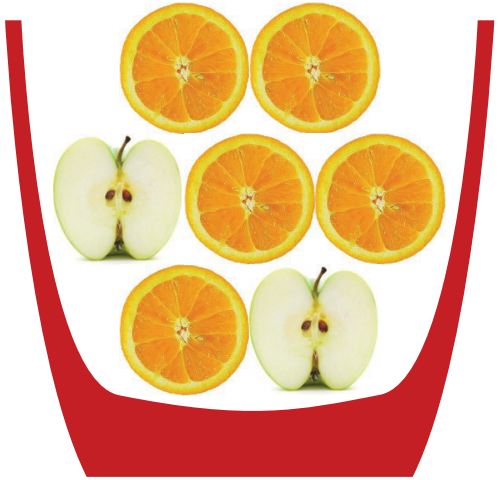
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

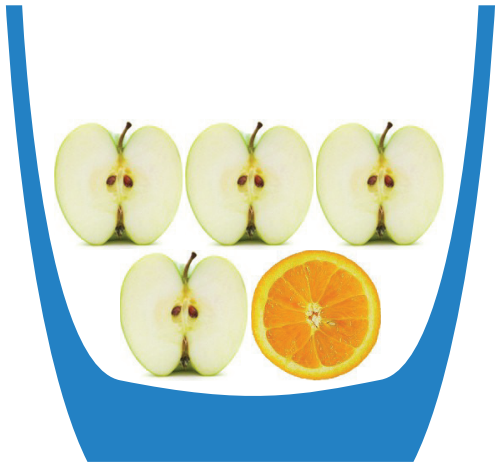


Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

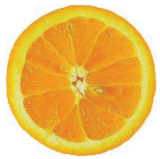
		
	2	5
	4	1

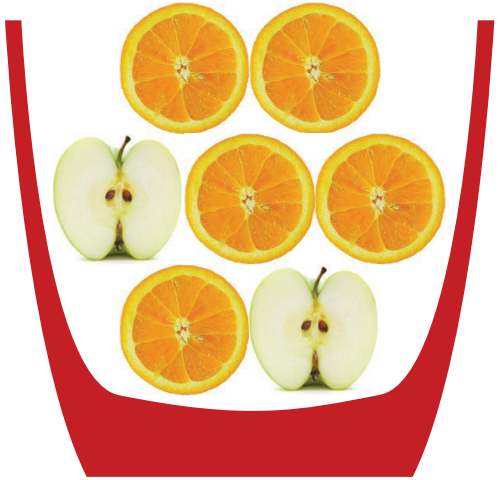


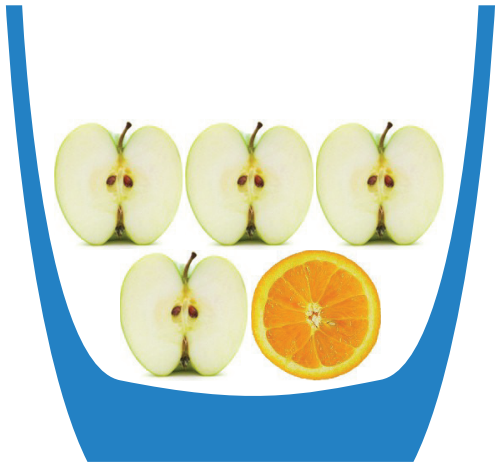


Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

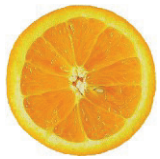
			
	2	5	7
	4	1	5
	6	6	12





Probabilities

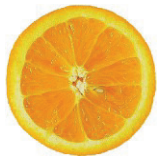
- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

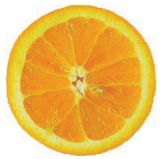
			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

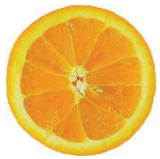
$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

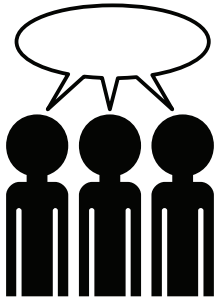
			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

$$= \frac{5/7 \cdot 7/12}{6/12} = 5/6$$



Medical test

A medical test for a given disease

- Correctly identifies the disease 99% of the time (true positives), and
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.

You go to the doctor to get tested, and the test turns out to be positive.

What is the probability you have the disease?

Hints:

- Identify from the text:
  p(Positive|Disease)
  p(Positive|No Disease)
  p(Disease)
  p(No Disease)
- Use the basic rules of probability given to the right to find:
  p(Disease|Positive)

$$p(x) = \sum_y p(x, y)$$

$$p(x, y) = p(x|y)p(y)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Frequency and mode

- **Frequency:** Percentage of time a value occurs
  - Example: Given the attribute **Gender** and a representative population of people, the value **Female** occurs about 50% of the time
- **Mode:** The most frequent attribute value
  - Example: Given the attribute **Operating System** and a representative population of computers, the value **Microsoft Windows** is the mode

- The notions of frequency and mode are typically used with categorical data

Percentiles

- **Percentiles:** Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the **p**th percentile is a value $\mathbf{x}_p$ of **x** such that **p** percent of the observed values of **x** are less than $\mathbf{x}_p$.
  - Example: The 10th percentile of **x** is the value $\mathbf{x}_{10\%}$ such that 10% of all values are less than $\mathbf{x}_{10\%}$.
- **Median:** The 50th percentile
  - Sort the numbers and take the middle value
    (if there are an even number of values, average the two middle values)

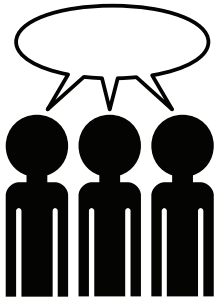
Measures of location

- **Mean:** Average

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- **Median:** Sort the numbers

$$\text{median}(x) = \begin{cases} x_{R+1} & \text{if } N \text{ is odd } (N = 2R+1) \\ \frac{1}{2}(x_R + x_{R+1}) & \text{if } N \text{ is even } (N = 2R) \end{cases}$$



Discussion

- What are the frequencies and the mode of these numbers and what is the mean and median?

0, 1, 1, 3, 5, 590

  - Explain also what to be careful about when using the mean and median

- **Frequency:** Percentage of time a value occurs
- **Mode:** The most frequent attribute value
- **Mean:** Average

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- **Median:** Sort the numbers

$$\text{median}(x) = \begin{cases} x_{R+1} & \text{if } N \text{ is odd } (N = 2R+1) \\ \frac{1}{2}(x_R + x_{R+1}) & \text{if } N \text{ is even } (N = 2R) \end{cases}$$

Measures of spread

- **Range**

$$\text{range}(x) = \max(x) - \min(x)$$

- **Variance**

$$variance(x) = s_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \overline{x})^2$$

- **Absolute average deviation (AAD)**

$$\text{AAD}(x) = \frac{1}{N}\sum_{n=1}^{N}|x_n - \bar{x}|$$

- **Median absolute deviation (MAD)**

$$\text{MAD}(x) = \text{median}\left\{|x_1 - \bar{x}|, \ldots, |x_N - \bar{x}|\right\}$$

- **Interquartile range (IQR)**

$$\text{IQR}(x) = x_{75\%} - x_{25\%}$$

Expected values

- Discrete random variable

$$\mathbb{E}\left[g(X)\right] = \sum_i g(x_i) P(X = x_i)$$

- Continuous random variable

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(X) p(X) \mathrm{d}X$$

Statistics

- **Mean**

$$\bar{x} = \mathbb{E}\left[x\right]$$

- **Covariance**

$$\mathrm{cov}(x, y) = \mathbb{E}\left[(x - \bar{x})(y - \bar{y})\right]$$

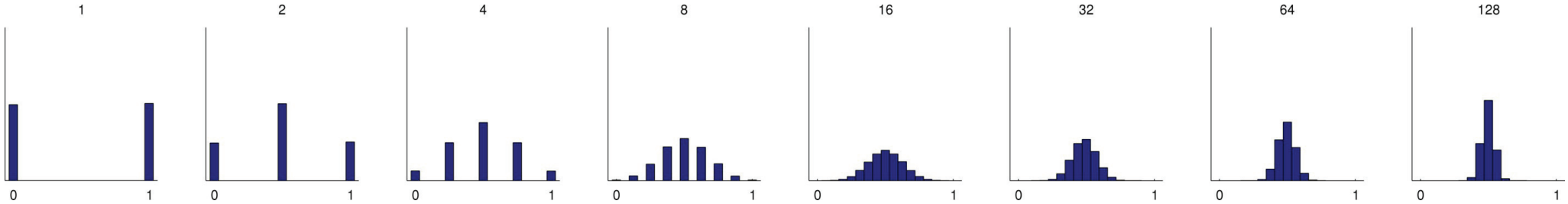
- **Variance**

$$\mathrm{var}(x) = \mathrm{cov}(x, x) = \mathbb{E}\left[(x - \bar{x})^2\right]$$

- **Standard deviation**

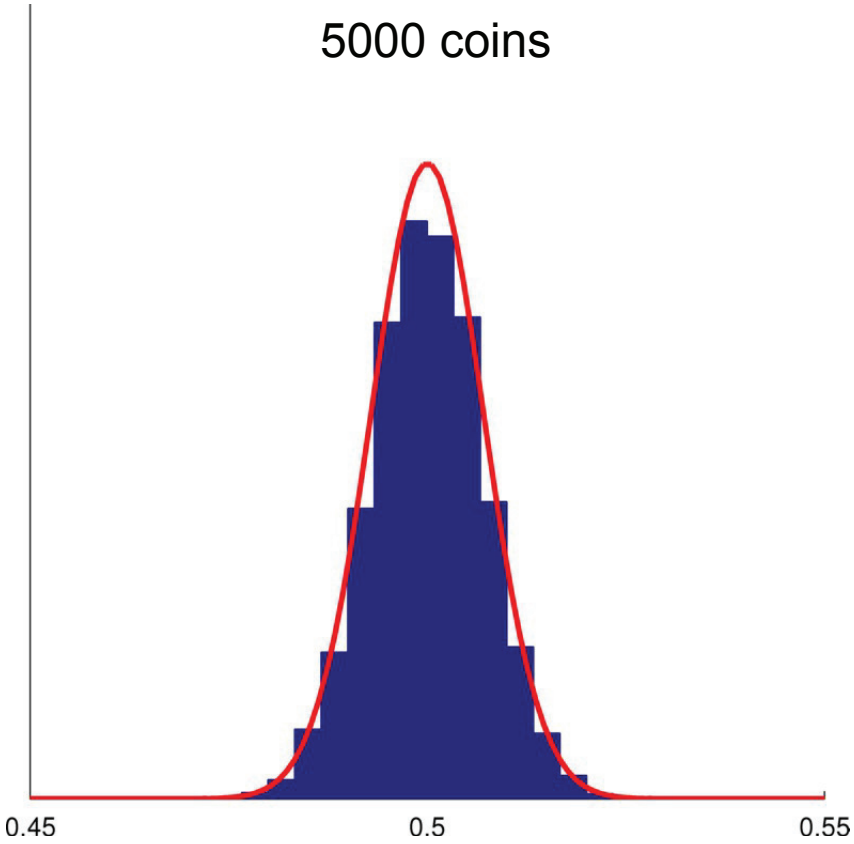
$$\mathrm{std}(\boldsymbol{x}) = \sqrt{\mathrm{var}(\boldsymbol{x})}$$

Normal distribution



- **Central limit theorem**
  - The mean of a large number of random variables will tend to a Normal distribution irrespective of the distribution of the random variables
    (Under certain conditions)

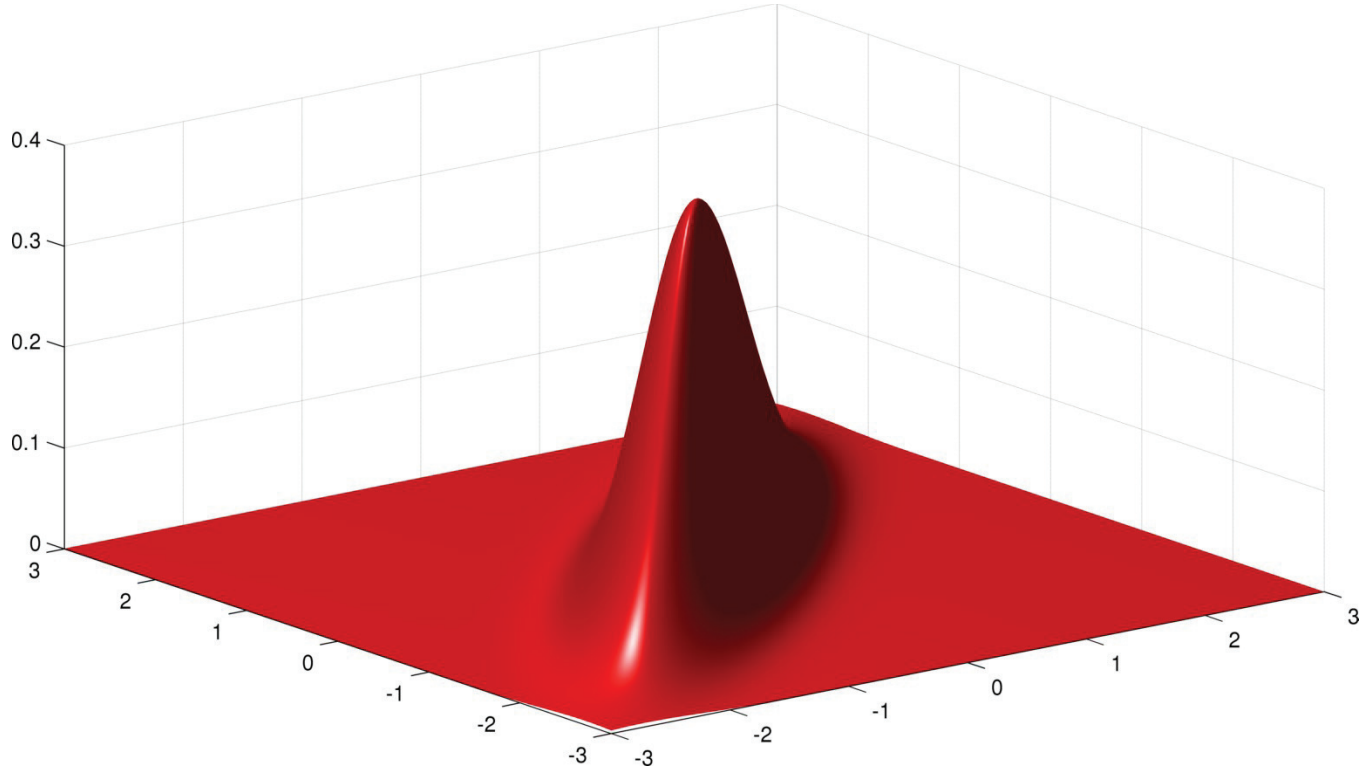
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$



- **Example:** Proportion of heads when flipping
  - 1 coin, 2 coins, 4 coins etc.

Multivariate Normal distribution

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



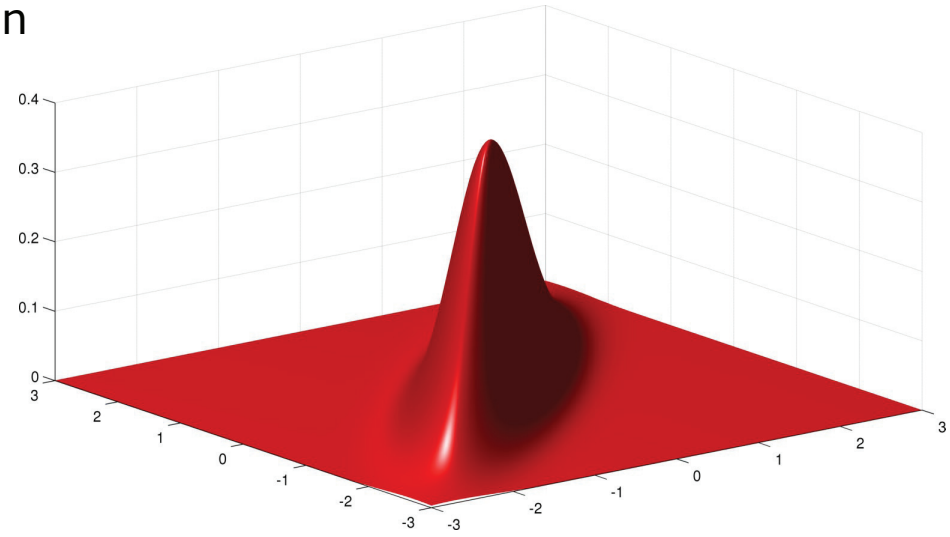
Multivariate Normal distribution

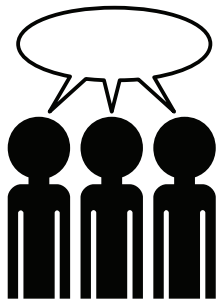
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Example: 2-dimensional Normal distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

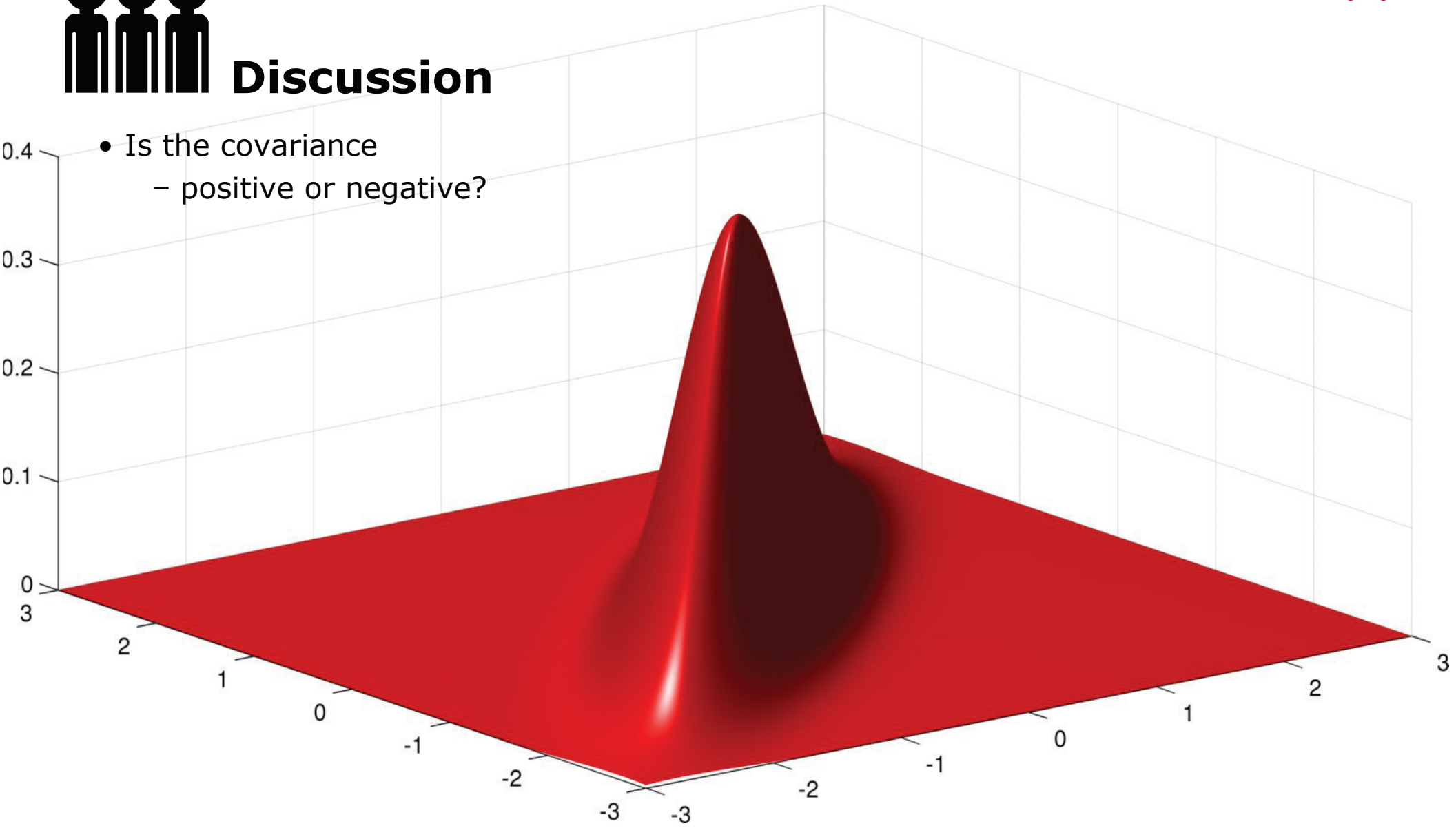
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{bmatrix}$$

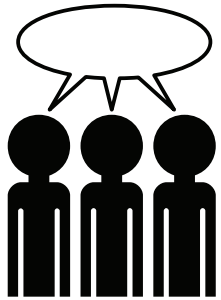




Discussion

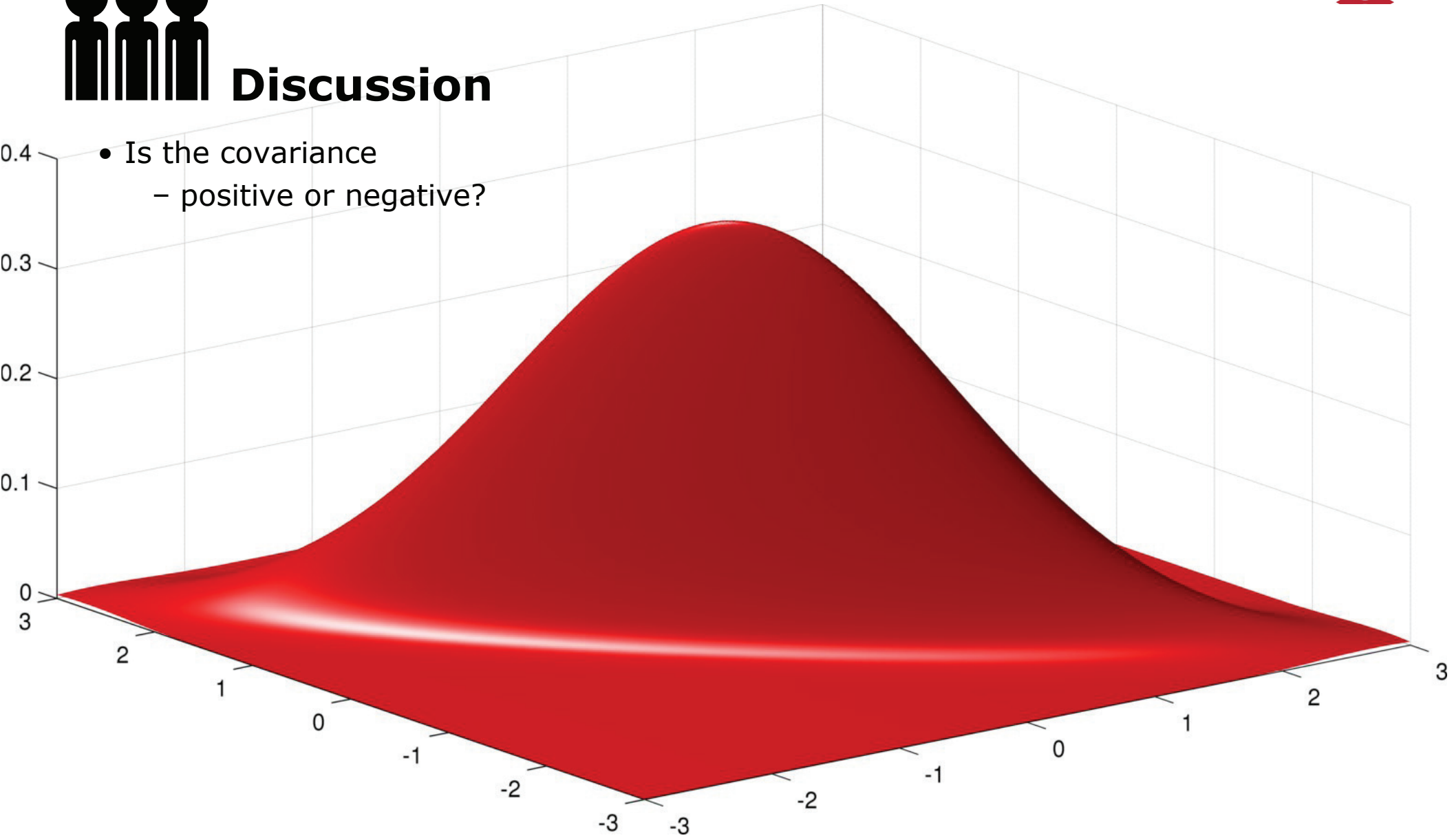
- Is the covariance
  - positive or negative?





Discussion

- Is the covariance
  - positive or negative?



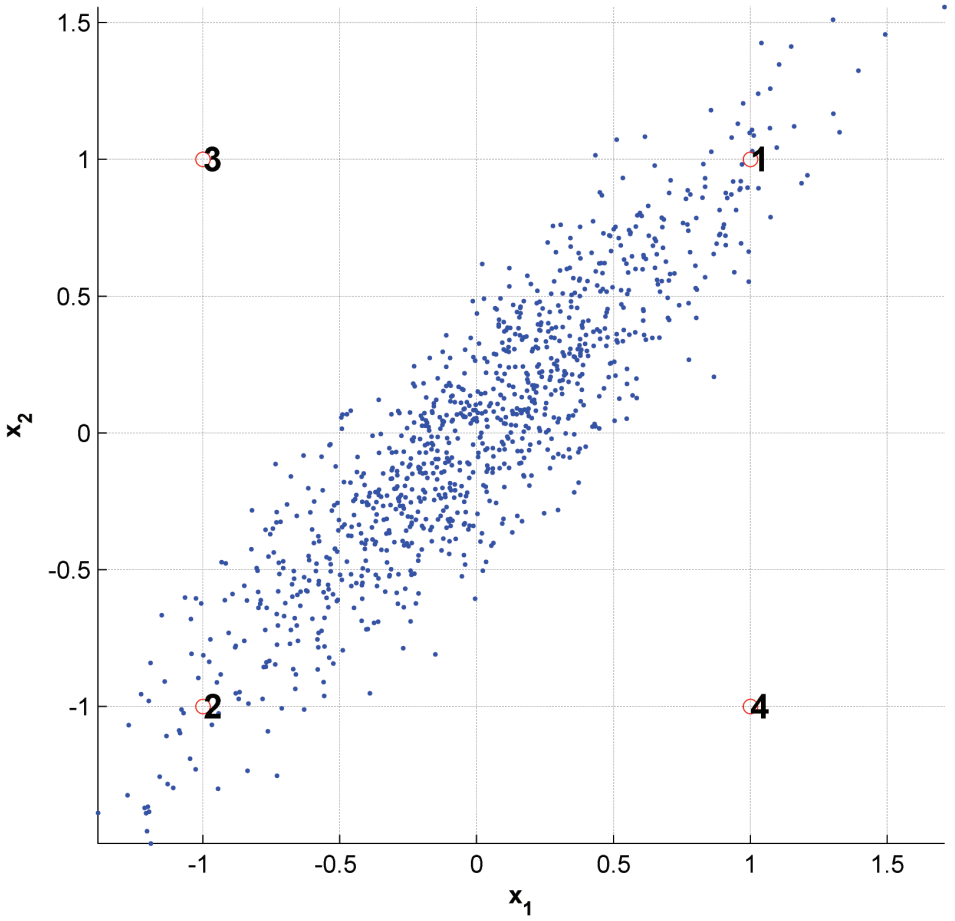
The Mahalanobis distance

How far are x_1 and x_2 apart?

- mahalanobis(x_1,x_2) = 4.2
- $d_{Euclidean}(x_1,x_2)^2$ = 8.0

How far are x_3 and x_4 apart?

- mahalanobis(x_3,x_4) = 80
- $d_{Euclidean}(x_3,x_4)^2$ = 8.0



$$\text{mahalanobis}(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})$$

$$d_{Euclidean}(\boldsymbol{x},\boldsymbol{y})^2 = (\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{I}^{-1}(\boldsymbol{x}-\boldsymbol{y})$$