Data-driven Intelligent Systems

Lecture 2
Properties of Data

KNOWLEDGE
TECHNOLOGY

http://www.informatik.uni-hamburg.de/WTM/

Overview

- Types of Data
- Representing Data
  - Relational Table
  - Statistical Descriptions
- Curse of Dimensionality

Important Characteristics of structured Data

- Dimensionality
  - Curse of dimensionality
- Resolution
  - Patterns depend on the scale
- Sparsity
  - Few values are present
- Distribution
  - Centrality and dispersion
- Similarities
  - Find outliers

Types of Data

- Structured Records
  - Tables
  - Transaction data
  - Relational records
- Sequential and semi-structured
  - Documents with text data
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential data: transaction sequences
  - Genetic sequence data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Data Objects

- ***Data sets*** are made up of data objects. Examples:
  - sales dataset: customers, store items, sales
  - medical dataset: patients, treatments
  - university dataset: students, professors, courses
- A ***data object*** represents an entity
  - Also called *sample, example, instance, tuple,* ***data point***
- Data objects are described by ***attributes***
- A data set as a ***matrix***:
  - rows -> data objects; columns ->attributes
- A ***database*** is an organised collection of data (sets)

Attributes

- ***Attribute*** (or ***dimensions***, ***features***, ***variables***):
  - a data field, representing a characteristic of a data object
  - **E.g.**, customer_ID, name, address
- Types:
  - Nominal
  - Binary
  - Ordinal
  - Numeric, quantitative:
    - Interval
    - Ratio

Attribute Types

- ***Nominal*:** categories, states, or “names of things”
  - *Hair_color* = {*black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
  - However, there is no meaningful order
- ***Binary*:** nominal attribute with only 2 states (0 and 1)
  - **Symmetric** binary: both outcomes equally important
    - e.g., gender
  - **Asymmetric** binary: outcomes not equally important
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., cancer positive)
- ***Ordinal***
  - Values have a meaningful order (ranking)
  - Magnitude between successive values *not* known
  - *Size* = {*small, medium, large*}, army rankings, grades

Numeric Attribute Types

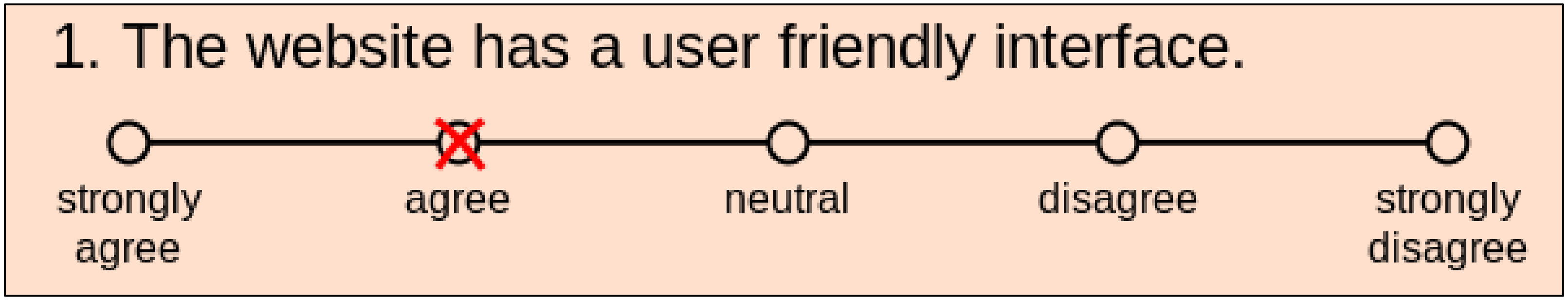
- ***Interval***
  - Measured on a scale of ***equal-sized units***
  - Values have ***order***
    - **Examples:** *temperature in C˚or F˚, calendar dates*
  - Differences between units can be ***quantified***
  - However, no true zero-point
- ***Ratio***
  - Inherent ***zero-point***
  - We can distinguish values by order of magnitude
    - "100 is 3 orders of magnitude larger than 0.1"
    - **Examples:** *temperature in Kelvin, length, durations of events, monetary quantities*

Attribute Types Overview

Type	Description	Examples	Operations
Nominal	Uses a label or name to distinguish one object from another	ZIP-Code, ID, Gender	= or !=
Ordinal	Uses values to provide the ordering of objects	Opinion, grades	< or >
Interval	Uses units of measurements, but the origin is arbitrary	Celsius, Fahrenheit, calendar dates	+ or -
Ratio	Uses units of measurement with fixed origin	Kelvin, length, counts, age, income	+, -, *, /

Likert Scale

- **Example:**



- Of which type are the attributes of a Likert scale?
  - Nominal
  - Ordinal
  - Interval (not well-defined intervals)
  - Ratio

Defining the Center of Multiple Data Points

- Each data type has its own natural way to characterize one “typical” value among multiple data points
  - Nominal — mode (most frequent value)
  - Ordinal — median (value in the middle)
  - Interval — mean (average)
  - Ratio — geometric mean

→ “Central Tendency“ (later in this lecture)

Discrete vs. Continuous Attributes
(Another Dimension of Data Classification I)

- ***Discrete Attribute***
  - Has only a finite or countable infinite set of values
    - **E.g.**, zip codes, profession, or set of words in collection of documents
  - Can all be mapped to integer values
  - Special case: Binary attributes
- ***Continuous Attribute***
  - Has continuous values
    - **E.g.**, temperature, height, or weight
  - One cannot list all possible values
  - Typically, real numbers represented as floating-point variables
    - Practically, represented using a finite number of digits

Static vs. Temporal Attributes
(Another Dimension of Data Classification II)

- Some data do not change with time:
  - ***static data***
- Some attribute values do change with time:
  - ***dynamic*** or ***temporal data***
- The majority of methods, software and commercial tools for data analysis and mining are more suitable for static data!

Experimental vs. Observational Data
(Another Dimension of Data Classification III)

- **Experimental Data** (Primary, Prospective)
  - Hypothesis H
  - Design an experiment to test H
  - Collect data, infer how likely it is that H is true
  - **E.g.**, *clinical trials in medicine*

- **Observational Data** (Secondary, Retrospective)
  - Massive non-experimental data sets
    - **E.g.**, human genome, atmospheric data, retail data, web logs for Amazon, Google, etc.
  - Not constrained by experimental design
  - Cheap compared to experimental data

Overview

- Types of Data
- Representing Data
  - Relational Table
  - Statistical Descriptions
- Curse of Dimensionality

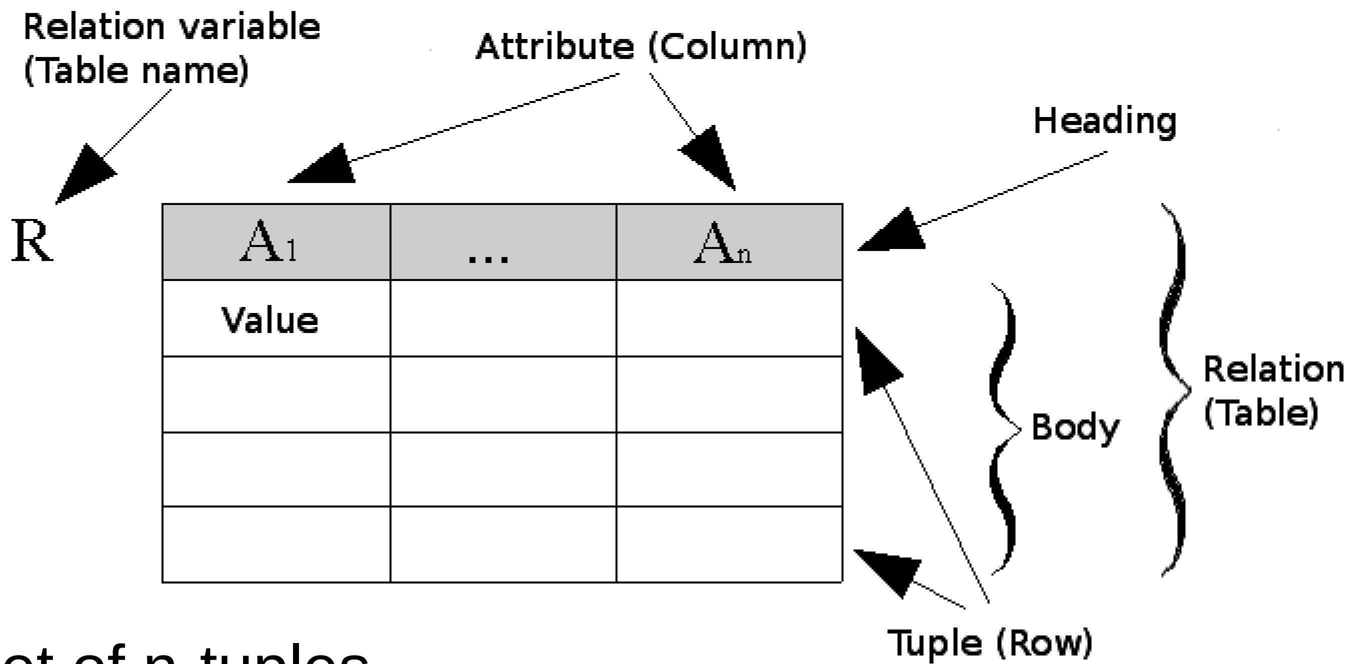
Preparing the Data

Two central tasks for the preparation of data:

- To organize data into a standard form, typically, a ***relational table*** (or tables)
- To prepare data sets by ***preprocessing***, such as dimensionality reduction

… for best performance of knowledge extraction algorithms

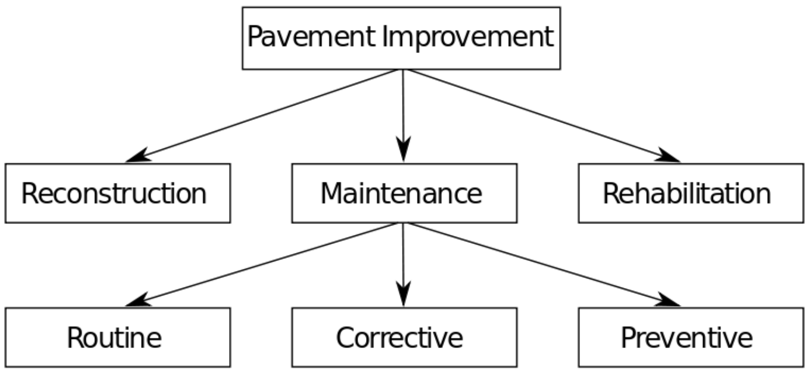
Relational Database Model



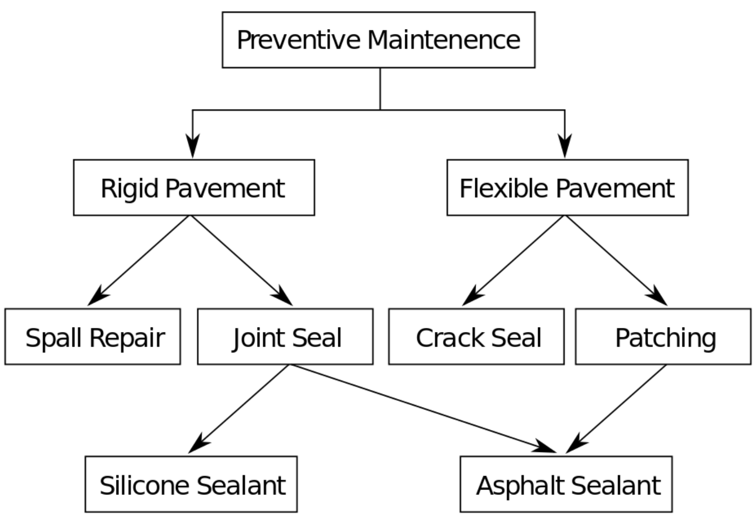
- Relation ~ set of n-tuples
- Tuples have no order
  - attribute names are used instead
- An attribute may serve as a key to link to other tables
- Mostly SQL data definition and query lang.

Other Database Models (Examples)

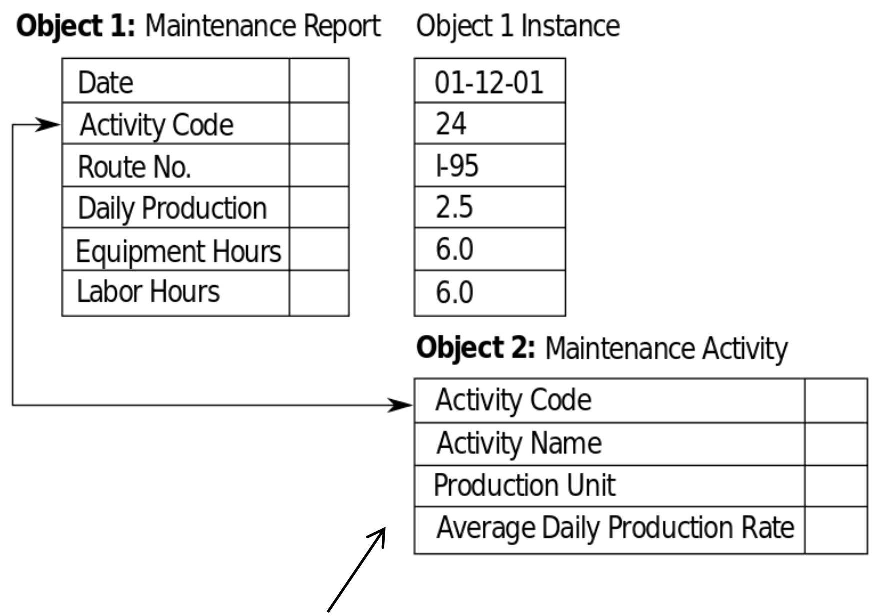
Hierarchical Model



Network Model

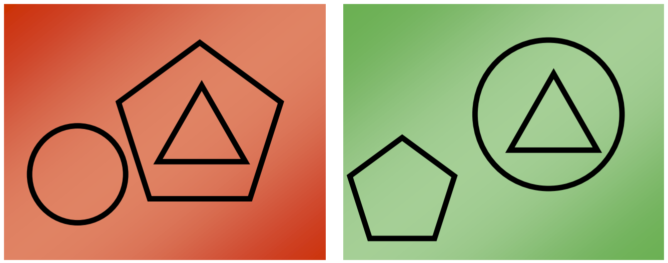


Object-Oriented Model



Information represented as objects in object oriented programming

Representing Data with Tables



Scene S1 Scene S2

Relational Representation

SCENE		
SceneID	ObjectID	Shape
S1	O1	Triangle
S1	O2	Circle
S1	O3	Pentagon
S2	O1	Triangle
S2	O2	Circle
S2	O3	Pentagon

INSIDE		
SceneID	ObjectID	ObjectID
S1	O1	O3
S2	O1	O2

Single Table Representation

SCENE				
SceneID	Triangle	Square	Circle	Pentagon
S1	+	-	+	+
S2	+	-	+	+

Representing Data with Tables: Market Baskets

Each basket represents one sample

TID: 100

TID: 200

TID:300

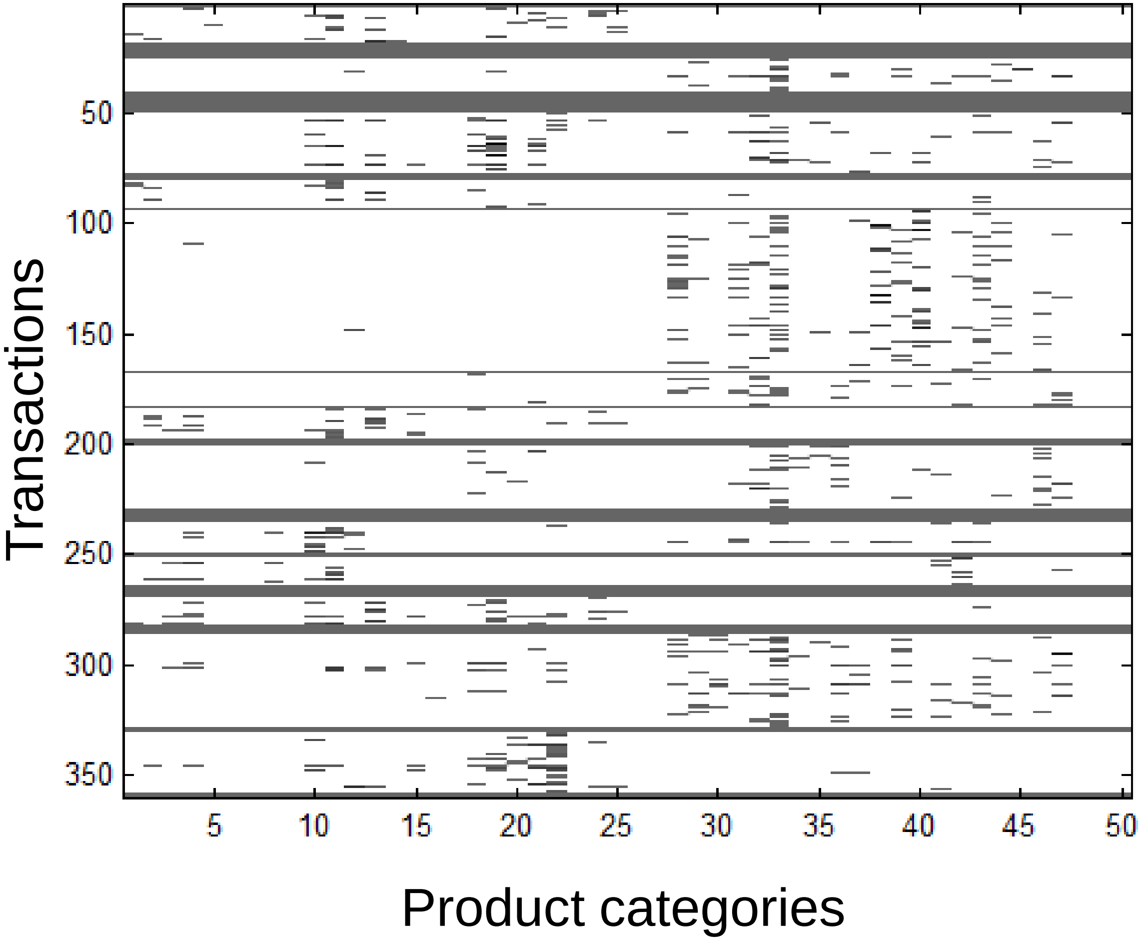
TID: 400

TID	Garlic	Milk	Detergent	Ketchup	Wine
100	Yes	No	Yes	Yes	No
200	No	Yes	Yes	No	Yes
300	Yes	Yes	Yes	No	Yes
400	No	Yes	No	No	Yes

Sparsity: eliminate “No‘s”

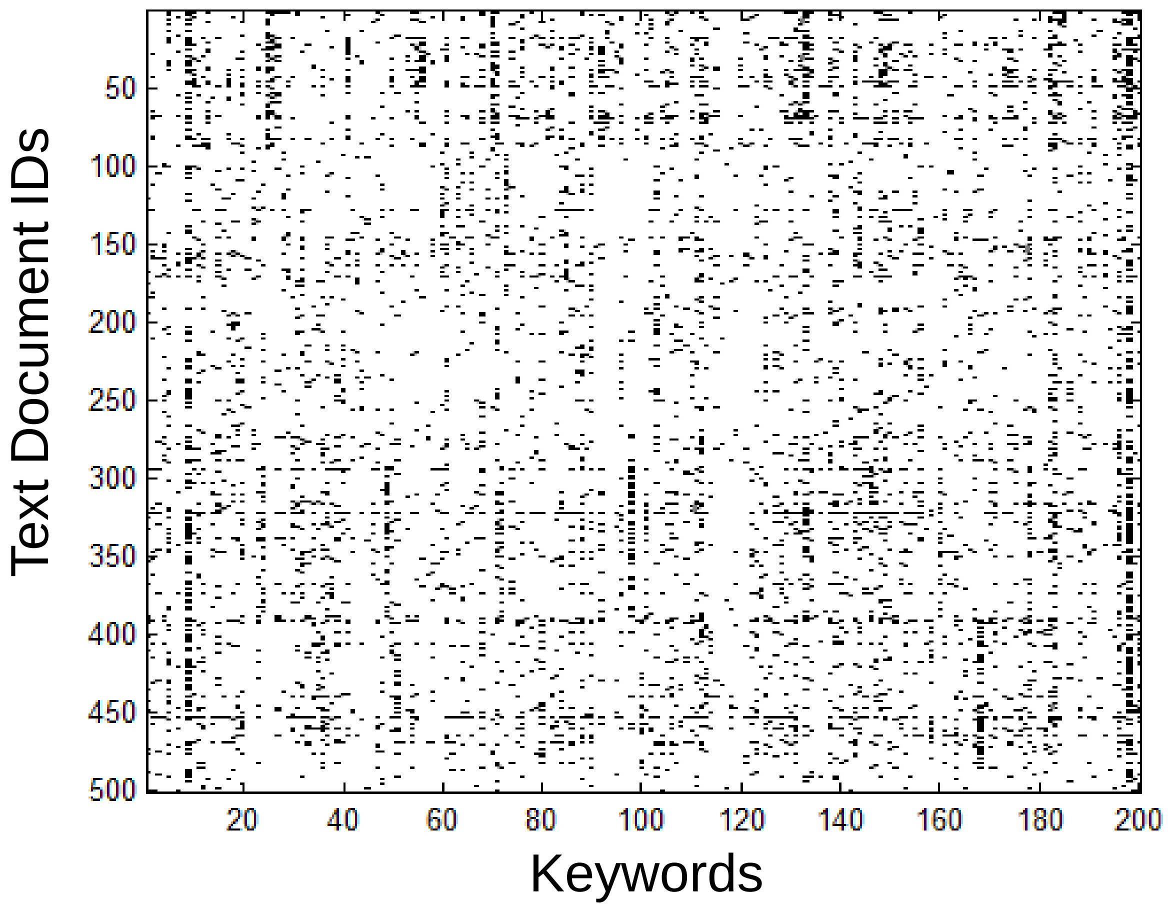
TID	Items
100	{Garlic, Detergent, Ketchup}
200	{Milk, Detergent, Wine}
300	{Garlic, Milk, Detergent, Wine}
400	{Milk, Wine}

Market Basket Data



TID	Items
01	01, 03, 44, 76
02	22, 37, 76
...	...

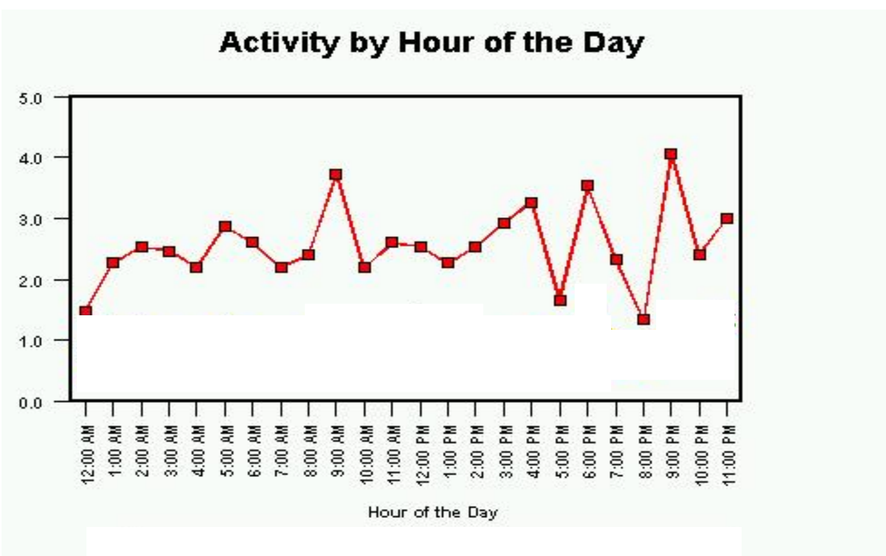
Representing Text as Tables

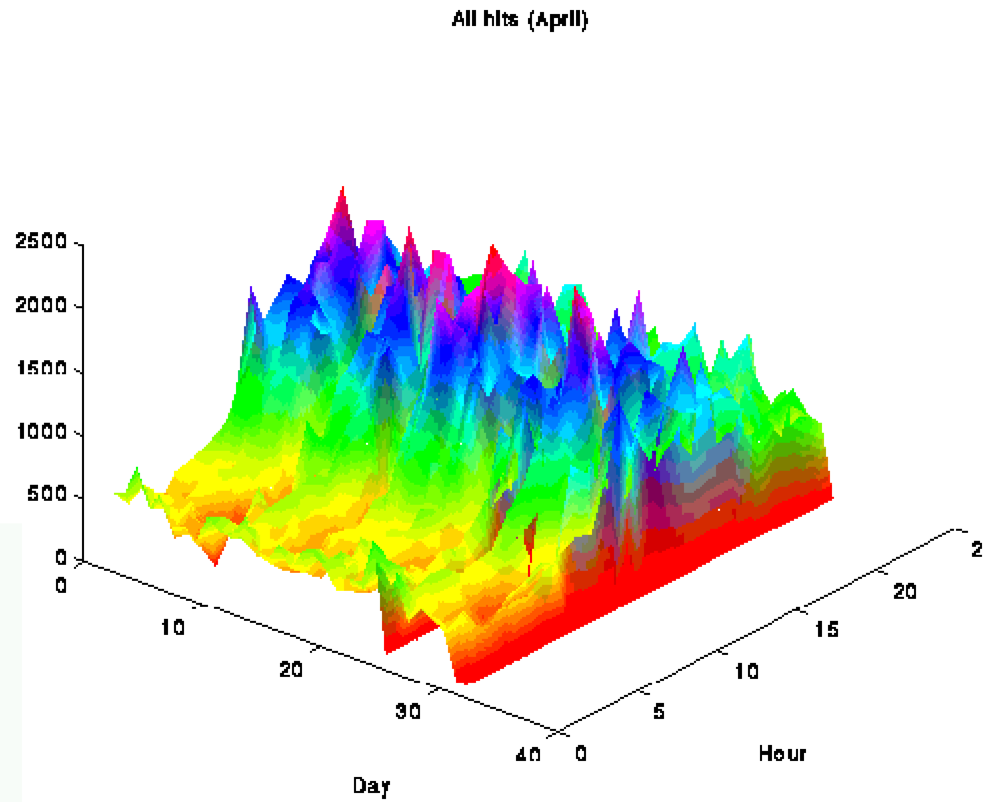


Text ID	Keywords
001	56, 34, 79
002	07, 122, 189
…	…

Web Log Data over Time as a Table

Day	Hour	# of hits
06/06/13	5 a.m.	58
06/07/13	6 a.m.	83
…	…	…





Time Series Data as a Table

Time	TS1	TS2		TSn
1	86	74	…	140
2	99	133	…	91
…	…	…	…	…

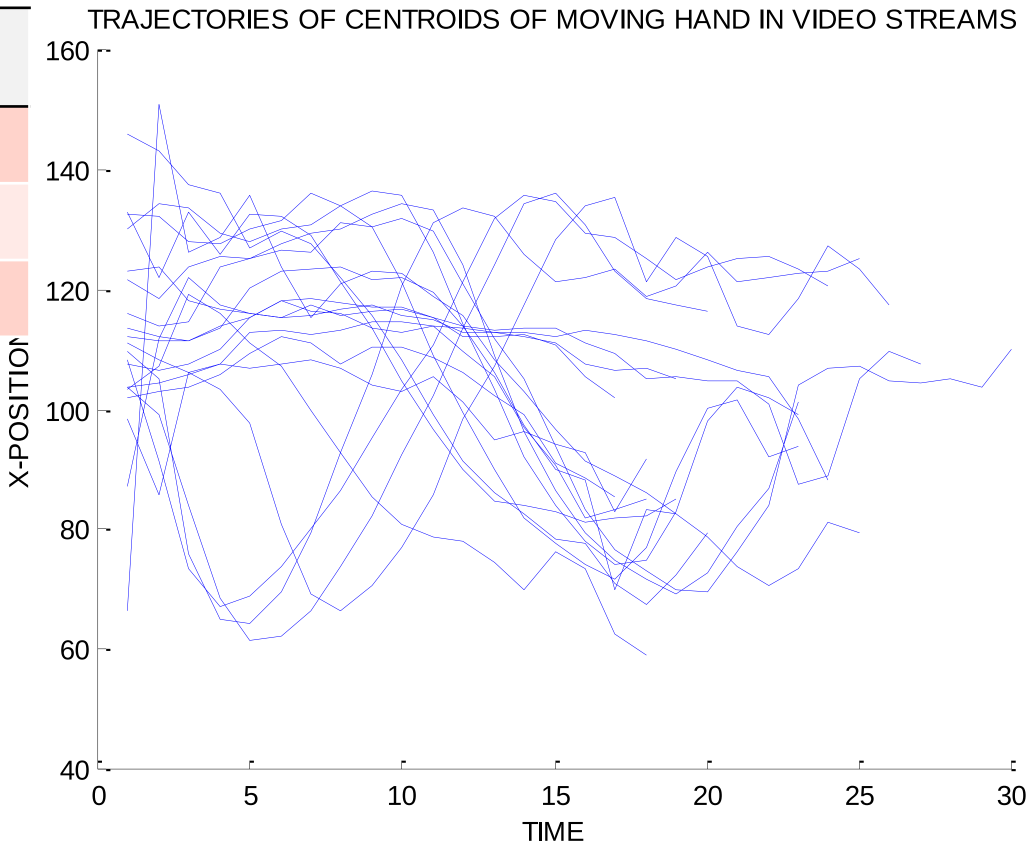
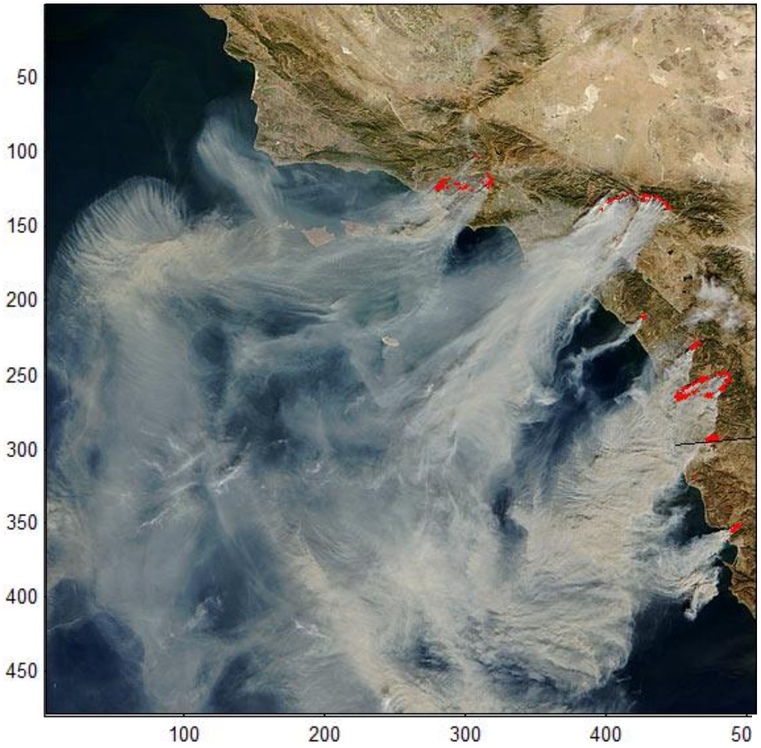


Image Data as a Table

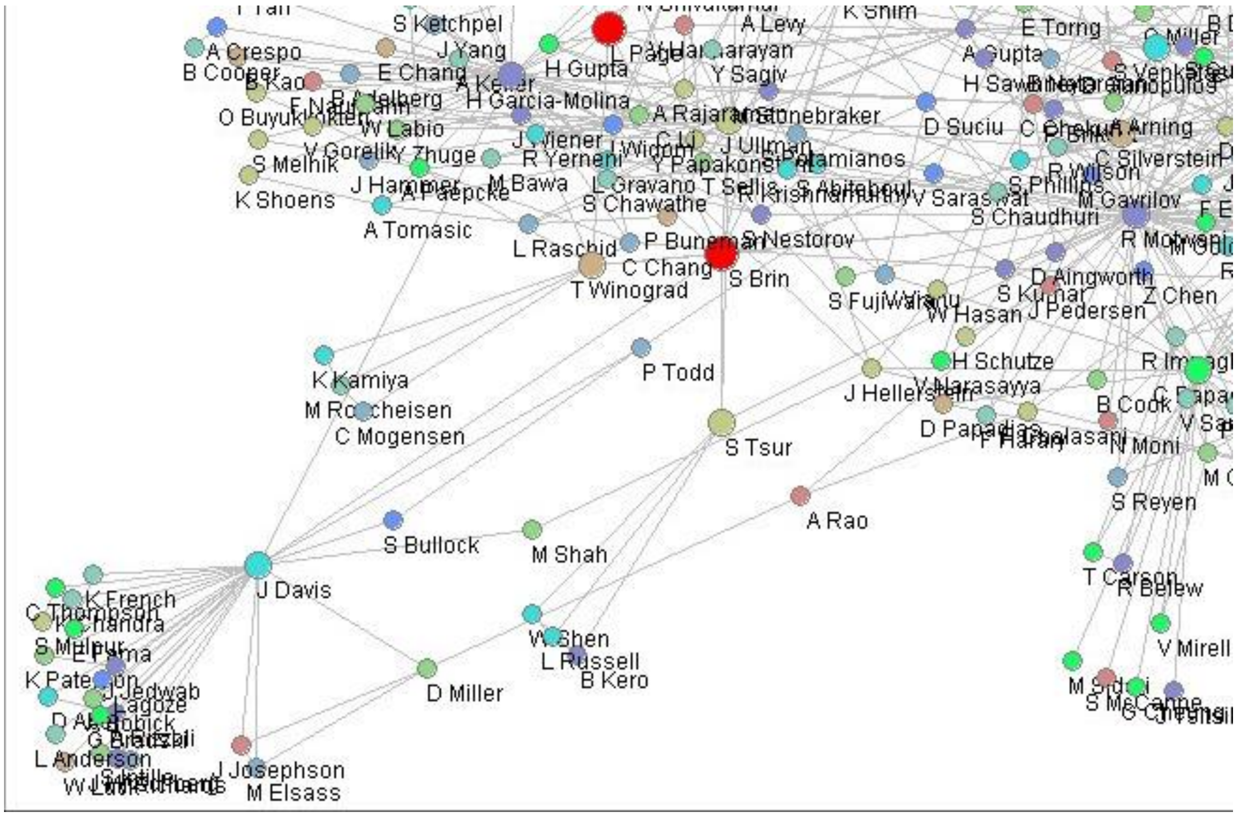
X coor.	Y coor.	red	green	blue
100	250	87	107	43
100	251	85	104	39
…	…	…		



Relational Data (=*Graph*) as a Table

Beginning node	Ending node	Distance
Bullock	Todd	134
Miller	Davis	87
…	…	…

Each row contains the beginning and ending node in one connection, and weight factor (here distance) connected with this link.



Overview

- Types of Data
- Representing Data
  - Relational Table
  - Statistical Descriptions
- Curse of Dimensionality

Basic Statistical Descriptions of Data

- Motivation
  - To better ***understand*** the data: central tendency, variation and spread
- Data ***dispersion*** characteristics
  - analyzed with multiple granularities of precision
  - median, max, min, quantiles, outliers, variance, etc.
  - ***Boxplot or quantile analysis*** on sorted intervals

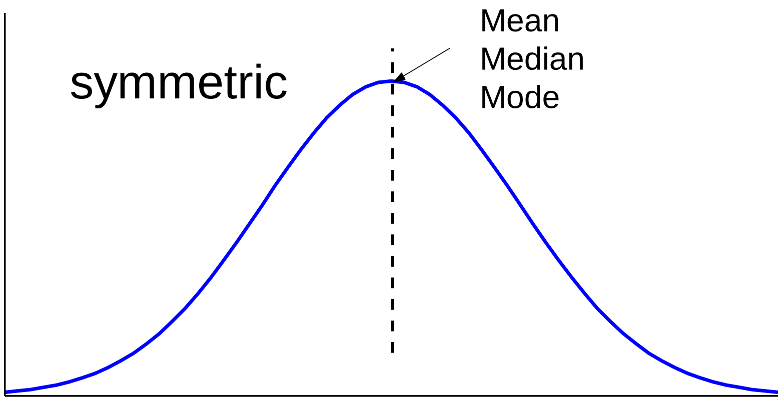
Measures of the Central Tendency

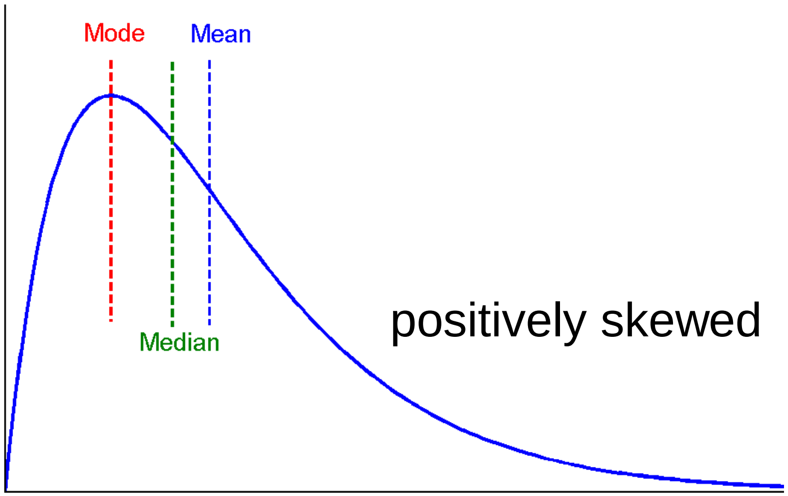
- ***Mode*** *good for **nominal** data*
  - Value that occurs most often in the data
  - Unimodal, bimodal, trimodal are data sets with 1, 2, 3 modes
- ***Median*** *good for **ordinal** data*
  - Middle value if odd number of values,
    or average of the middle two values otherwise
- ***Mean*** *good for **interval** data*
  - Population mean (*N* = population size): $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Mean estimated from samples (*n* = sample size): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Mostly *n* << *N*
- ***Geometric Mean*** *good for **ratio** data type*

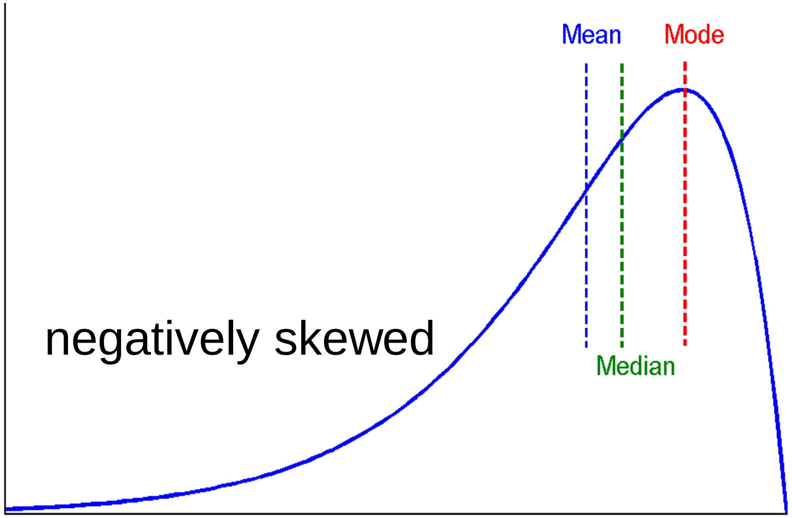
$$\bar{X}_{geom} = \sqrt[n]{x_1.x_1.......x_n}$$

Symmetric vs. Skewed Data

- Symmetric data:
  - Median = mean = mode
- Skewed data:
  - Median ≠ mean ≠ mode







Empirical formula for moderately asymmetrical curves:

$$mean - mode = 3 \times (mean - median)$$

Dispersion of Data: Standard Deviation

- Variance and standard deviation
  - ***Variance***: Variance estimated from sample:

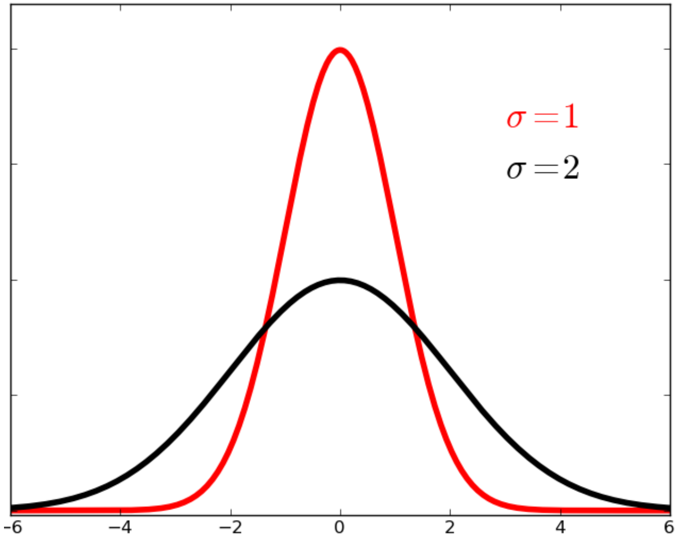
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  - ***Standard deviation*** σ is the square root of variance σ^2
  - Outliers contribute over-proportionally to the variance, due to the square

Symmetric Example: Normal Distribution

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

normalizer (not exact on discrete space!)

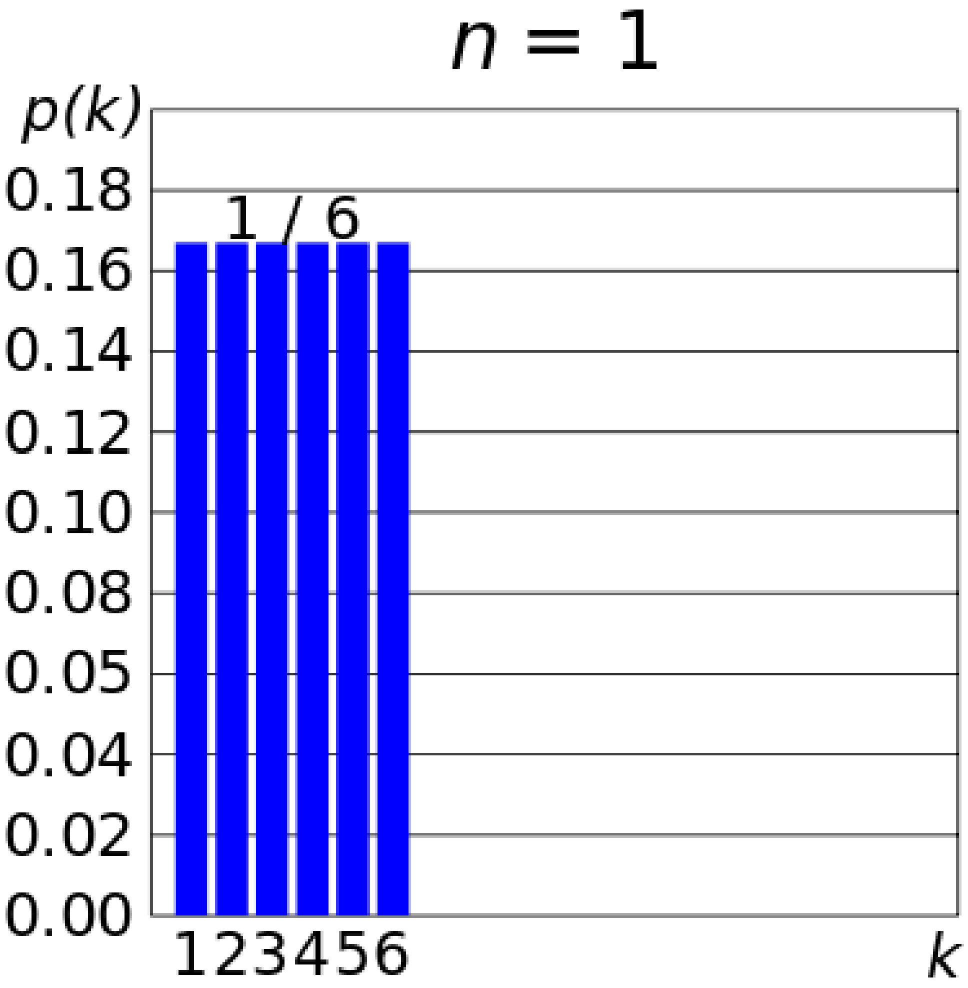


- Central Limit Theorem: (under certain conditions …) the sum of many random variables converges to a Gaussian

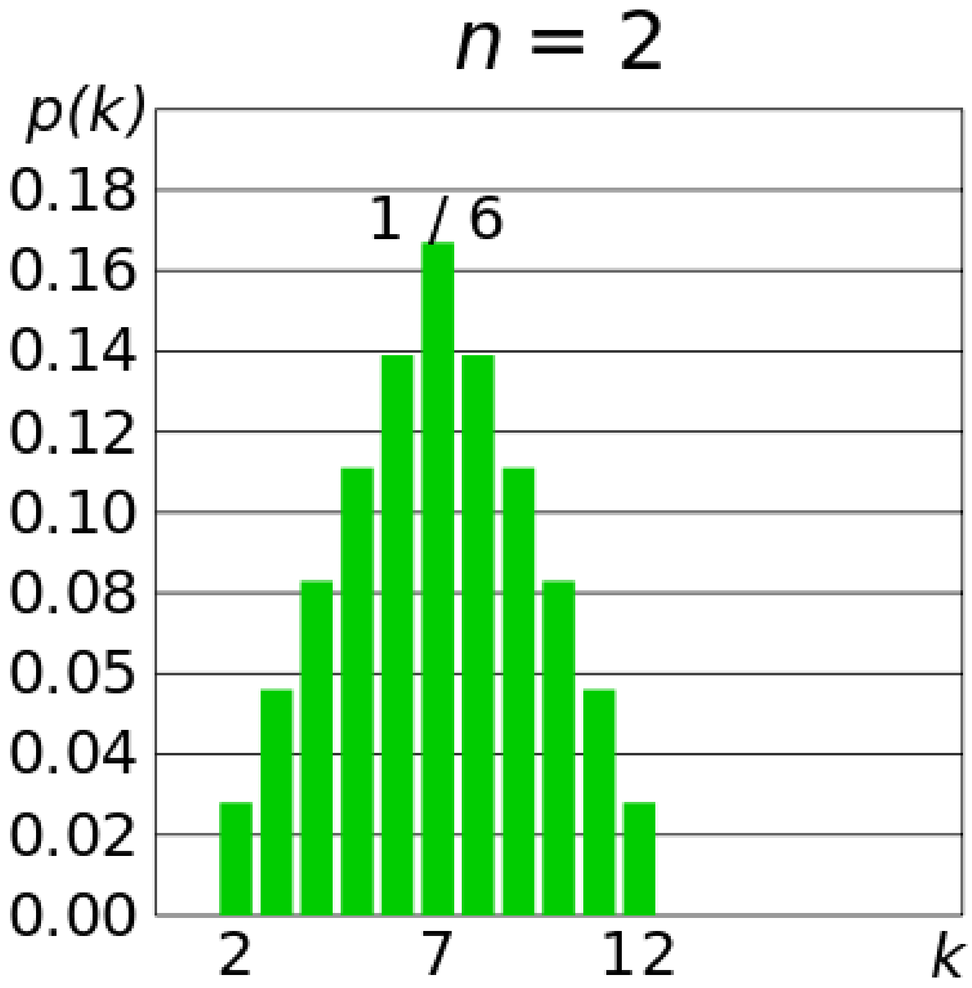
aka. Gaussian Distribution

- **Example:** sum of n fair 6-sided dice

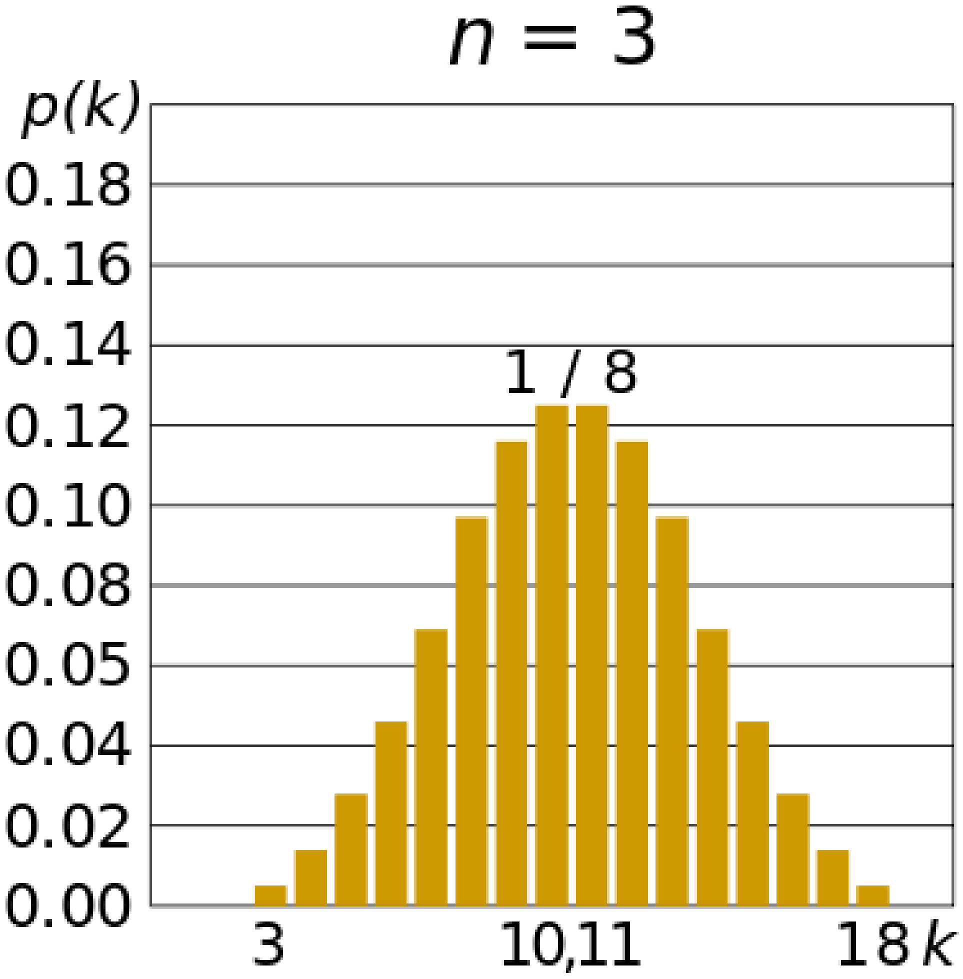
Sum of random variables: 6-sided dice



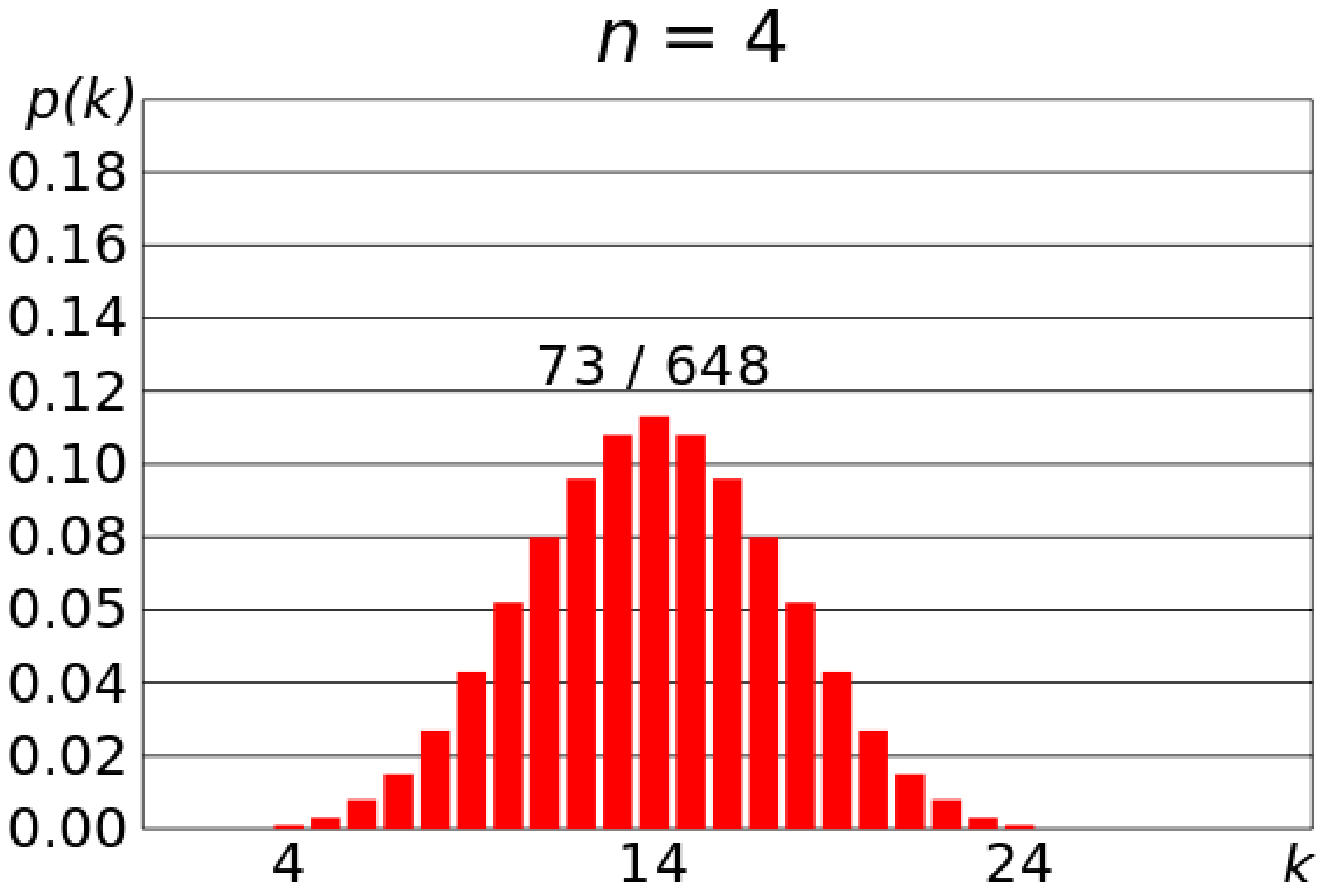
Sum of random variables: 6-sided dice



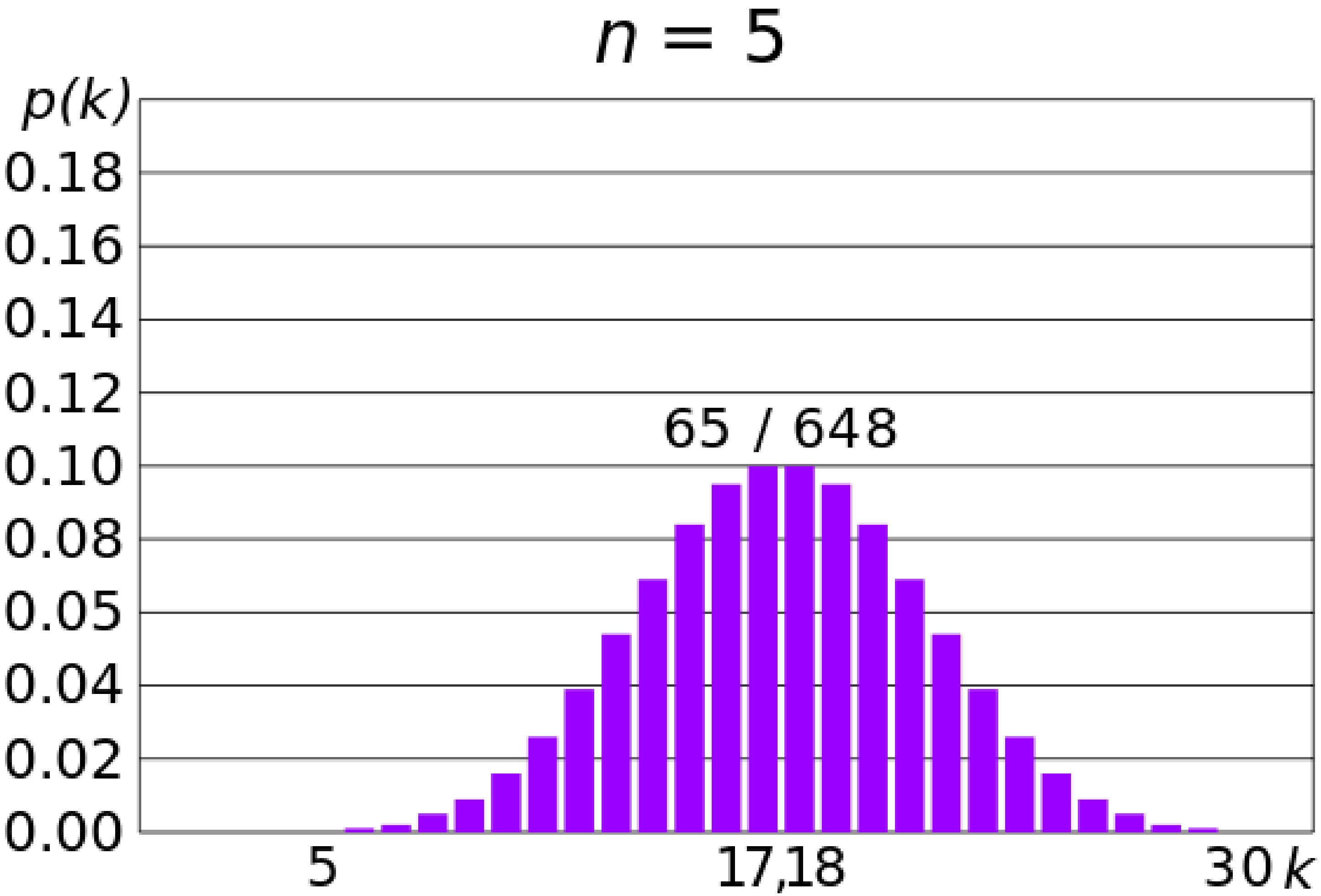
Sum of random variables: 6-sided dice



Sum of random variables: 6-sided dice



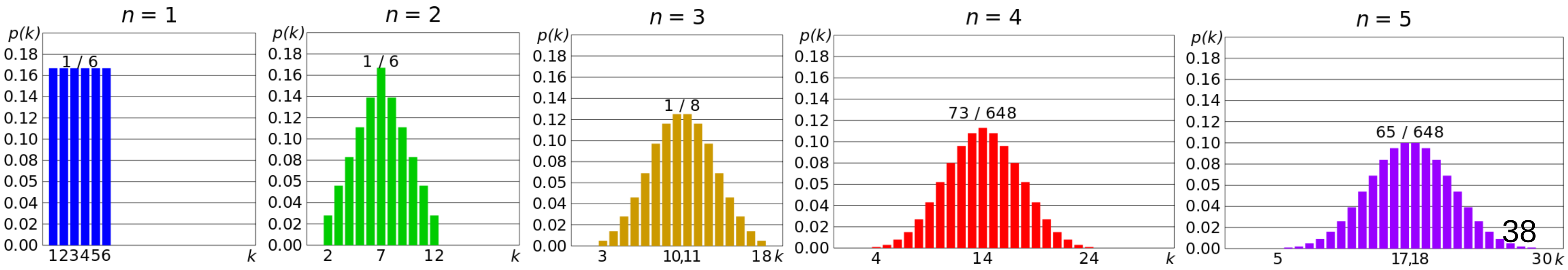
Sum of random variables: 6-sided dice



Normal (Gaussian) Distribution

- Central Limit Theorem: the sum of many random variables converges to a Gaussian

- **Example:** sum of n fair 6-sided dice

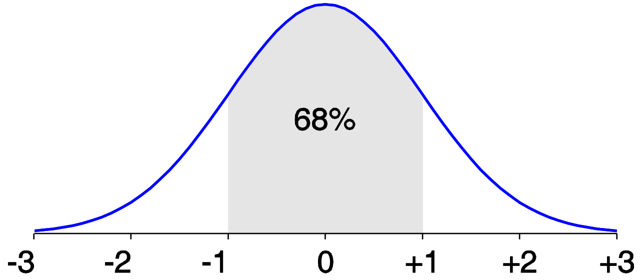


Normal (Gaussian) Distribution

- From $\mu - \sigma$ to $\mu + \sigma$:

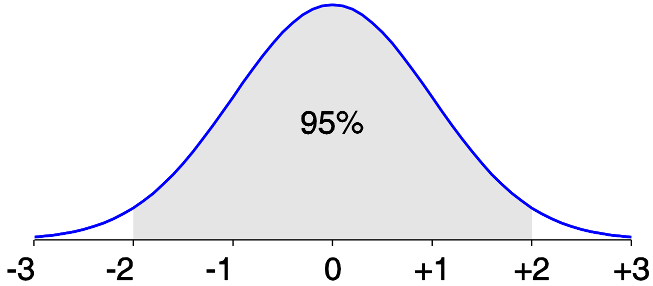
  contains ~ 68%
  of the measurements

  (μ: mean, σ: standard deviation)



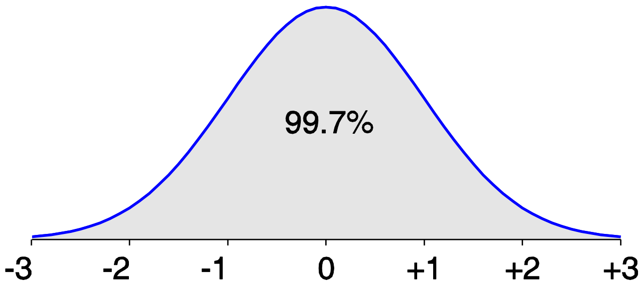
- From $\mu - 2\sigma$ to $\mu + 2\sigma$:

  contains ~ 95%



- From $\mu - 3\sigma$ to $\mu + 3\sigma$:

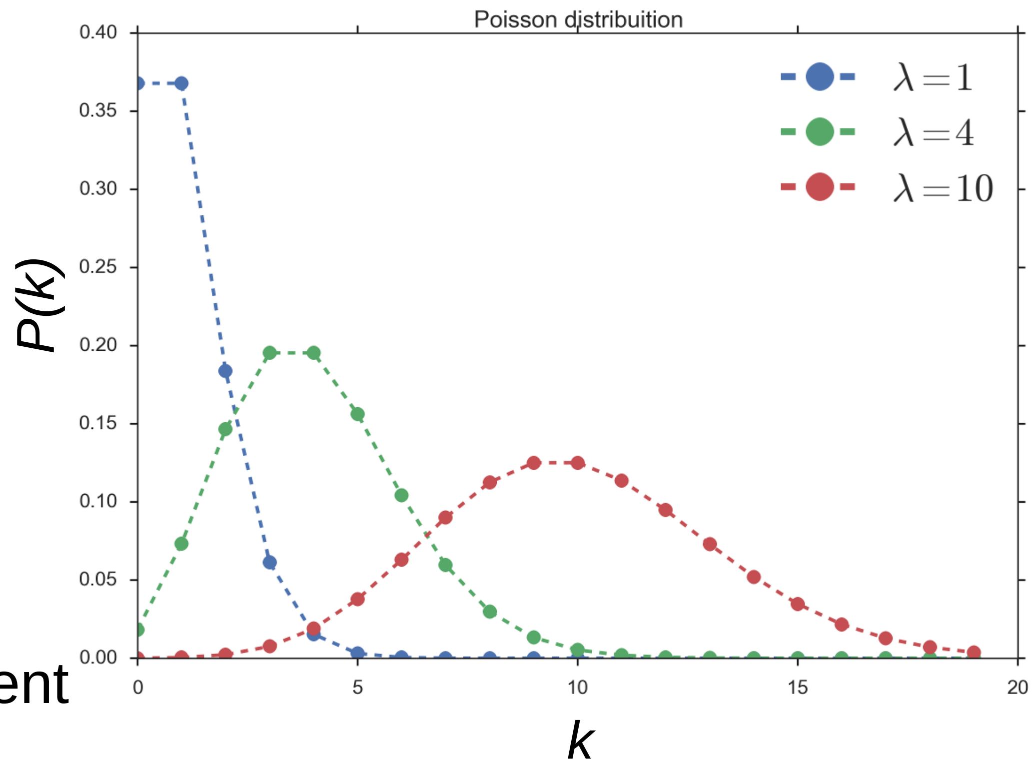
  contains ~ 99.7%



- Of all distributions with given mean and variance, the Gaussian maximises the entropy

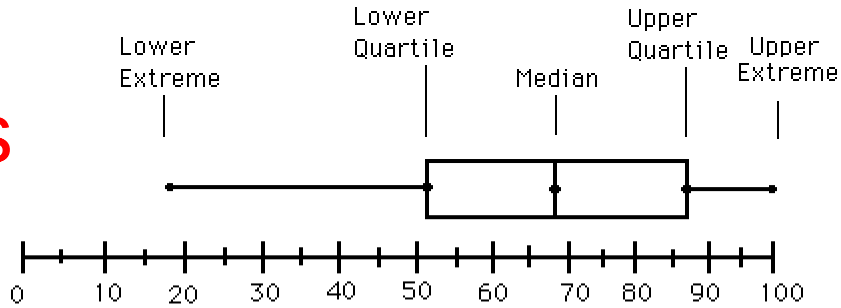
Skewed Example: Poisson Distribution

$$P(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

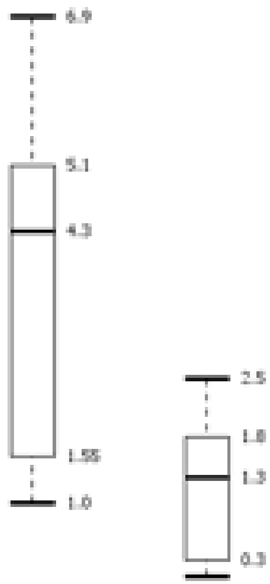


- Probablity that *k* events happen in a given interval
- *λ* = average number of events in an interval
- Events must be independent
- Large *λ* → ≈ Gaussian-like
- **E.g.:**
  - # meteors that hit earth per year
  - # patients arriving at an emergency room at a given hour
  - # neural spikes per second (model)

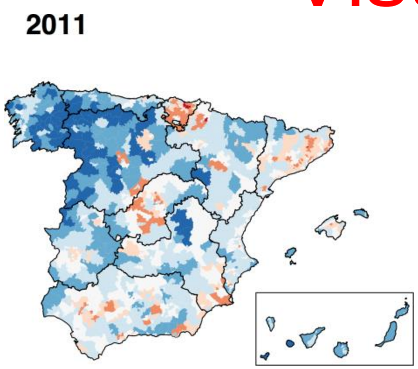
Box (-and-Whisker) Plots

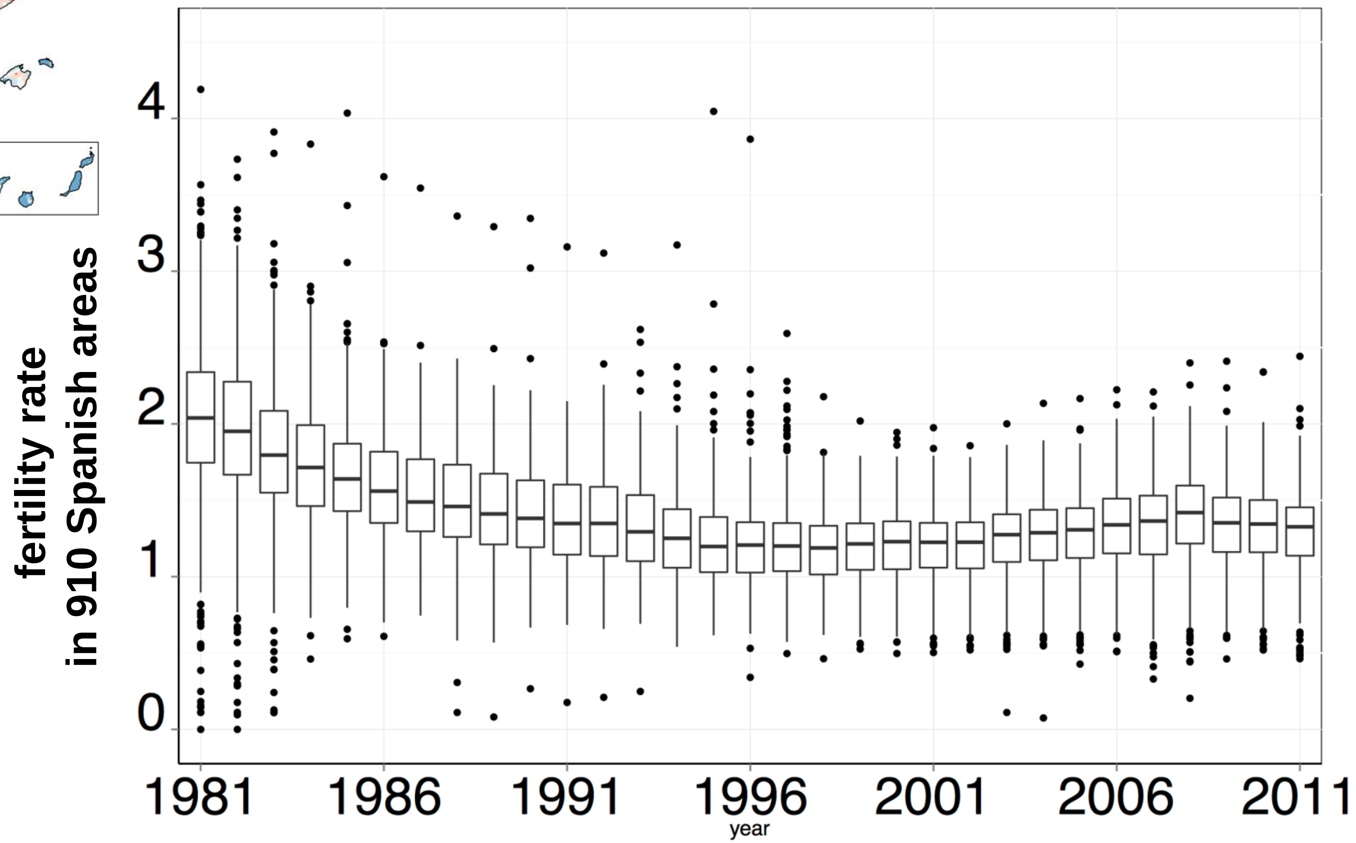


- ***Five-number summary*** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- ***Boxplot***
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is the interquartile range (IQR)
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to minimum and maximum
  - If outliers: points beyond specified thresholds, plotted individually, e.g. value lower than Q1 - 1.5·IQR or higher than Q3 + 1.5·IQR. Whiskers extend only to the non-outlier data.



Visualization of Data Dispersion: Boxplot Time Series





Here:

- Lines in the boxes show national *average* value (instead of *median*)

Overview

- Types of Data
- Representing Data
  - Relational Table
  - Statistical Descriptions
- Curse of Dimensionality

Curse of Dimensionality
(Geometric Approach I)

The "*curse of dimensionality*" is due to the geometry of high-dimensional spaces.

- The properties of high-dimensional spaces often appear *counterintuitive* because our experience with the physical world is in low, 2- or 3-dimensional space.
- Conceptually, objects *in high-dimensional spaces* have a *larger amount of surface* area for a given volume than objects in low-dimensional spaces.

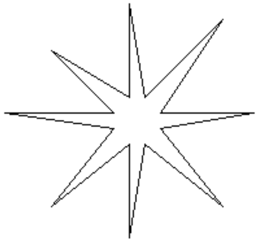
Curse of Dimensionality
(Geometric Approach II)

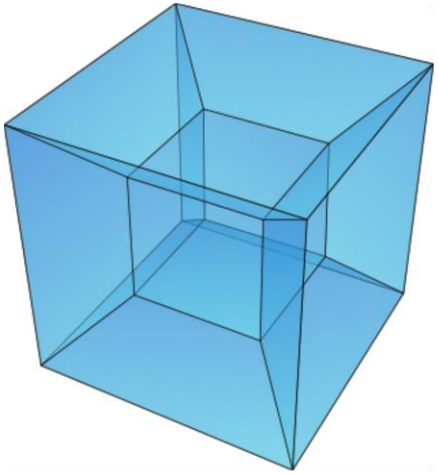
For example:

- A high-dimensional hypercube may be visualized as a porcupine (or even a hedgehog, as small 3D things have more surface per volume:

  surface~length^2
  volume~length^3)

- As the dimensionality grows, the surface grows relative to the central part of the hypercube.





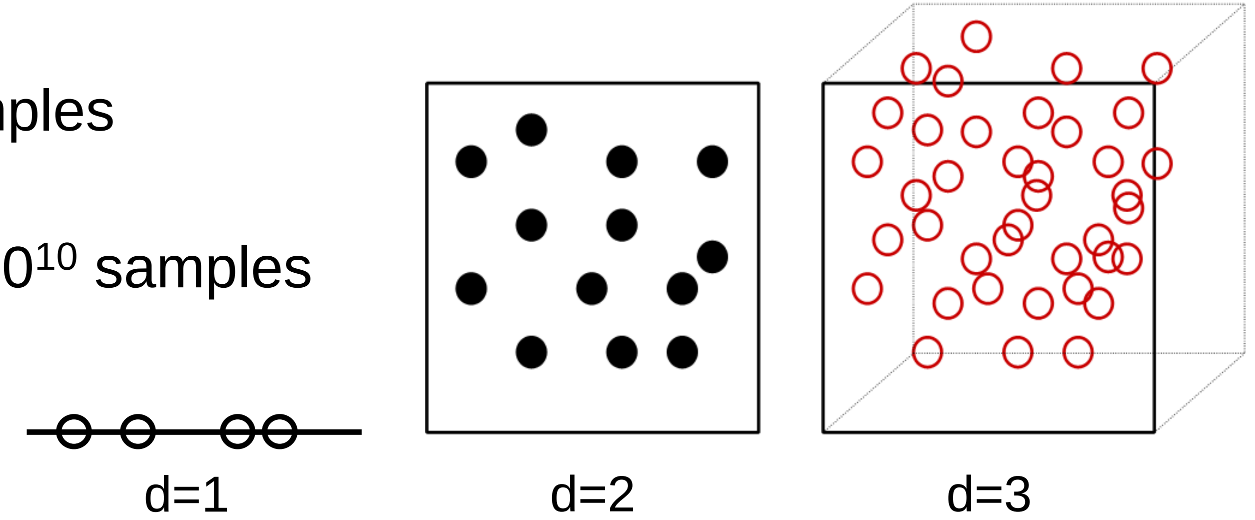
Curse of Dimensionality (1)

- The size of a data set yielding the same density of data points in d-dimensional space, increases ***exponentially*** with dimensions.

  To achieve the same density of points in d dimensions, we need n^d data points.

- **Example**
  - d = 1
    $\rightarrow$ n = 100 samples
  - d = 5
    $\rightarrow$ $n = 100^5 = 10^{10}$ samples

Same density of data



Curve of Dimensionality (2)

- In a high-dimensional space, a ***larger radius*** is needed to enclose the ***same fraction*** of data points.
  The ***edge length e*** of the hypercube scales as:

$$e(p) = p^{1/d}$$

p: pre-specified fraction of samples
d: number of dimensions

- **Example:**

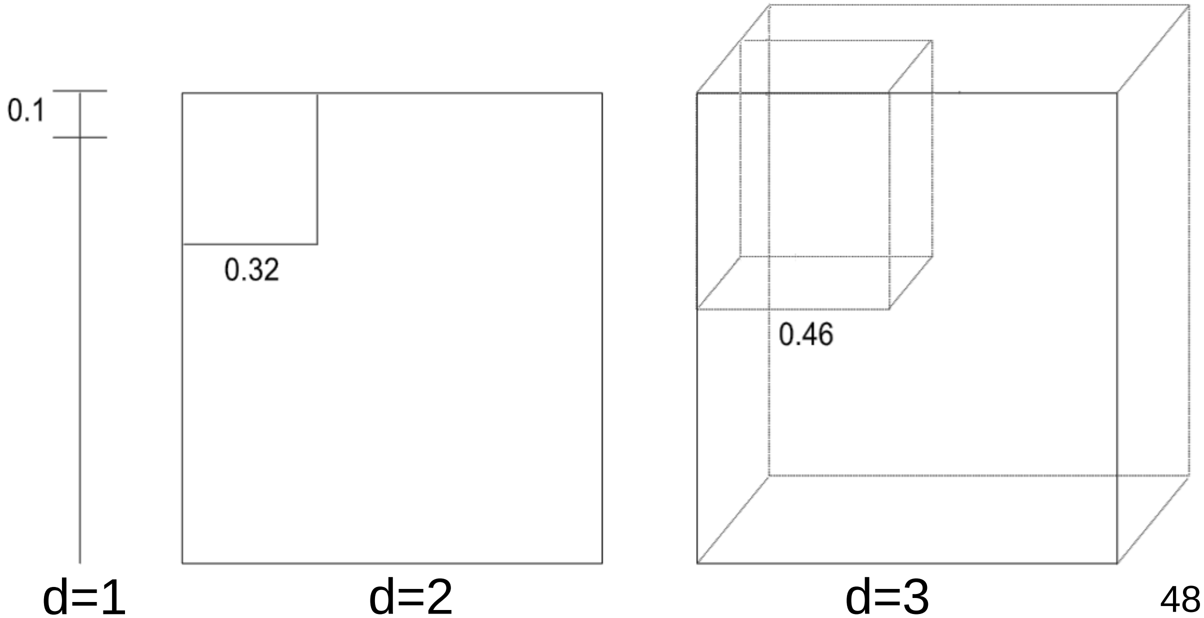
  10% of the samples (p=0.1):

  1-D: $e_1(0.1) = 0.1$

  2-D: $e_2(0.1) = 0.32$

  3-D: $e_3(0.1) = 0.46$

  10-D: $e_{10}(0.1) = 0.8$

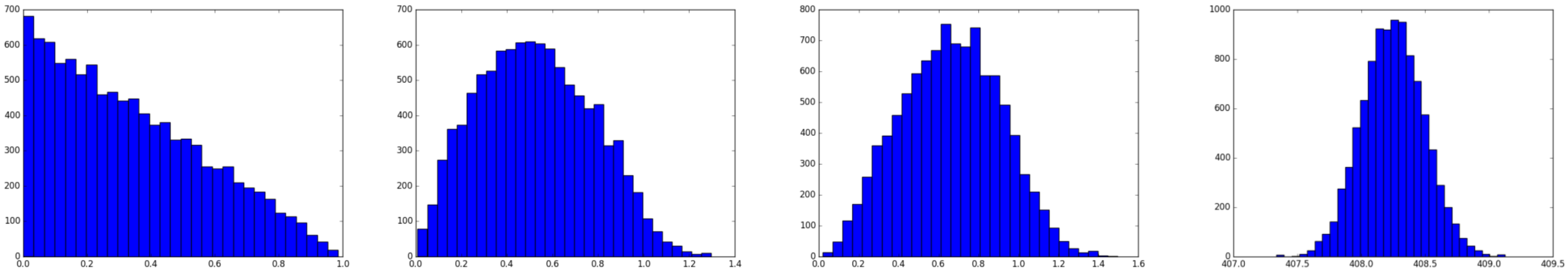


Curse of Dimensionality (3)

- **Average distances increase with higher dimensionality**
  - in high dimensions: no two random points are nearby

- Figures show histograms of Euclidean distances between 10000 pairs of randomly sampled points in a cube of unity length in dimensions d:

d=1 d=2 d=3 d=1000000



Curse of Dimensionality (4)

- In a high-dimensional space
  - The distance to the next sample point gets large:

    For a sample size *n*, the expected distance *D* between normalized data points in *d*-dimensional space is:

$$D(d,n) = \tfrac{1}{2} \cdot \left(\tfrac{1}{n}\right)^{1/d} = \frac{0.5}{\sqrt[d]{n}}$$

- **Example, expected distance between 10000 points:**

  For a 2-D space: $\rightarrow D(2,10000) = 0.005$

  For a 10-D space: $\rightarrow D(10,10000) \approx 0.2$

  different in the Kantardzic book!

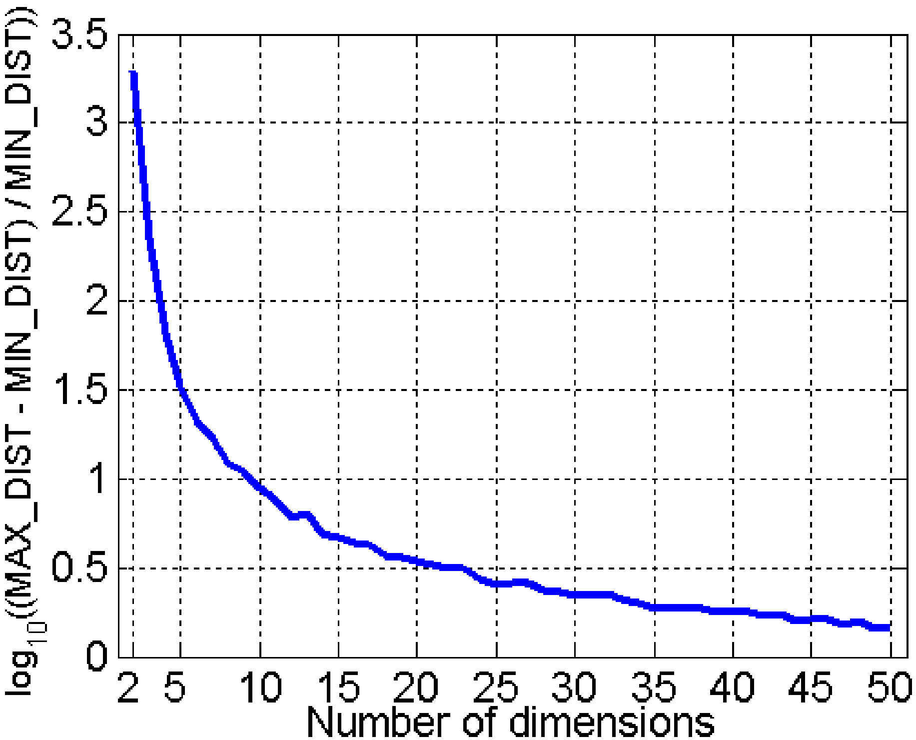
- However, almost every point is close to some edge

Curse of Dimensionality (5)

Experimental Confirmation:

With higher dimensionality:

- distances between data points become more similar
- data becomes increasingly ***sparse***
  → local "density" looses its meaning, if not backed by sufficiently many data
- most are ***outliers***
  → "distance" less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

See also: **Learning in High Dimension Always Amounts to Extrapolation**
https://lauraruis.github.io/2021/11/06/extra.html

Curse of Dimensionality (Summary)

As the dimension increases:

(1) we need exponentially more data for constant density,

(2) a hypercube of larger edge length covers same subspace,

(3) distance between points increases,

(4) distance to an edge decreases,

(5) every point becomes an outlier.

(1),(2) $\rightarrow$ difficult to make local estimates; we need more and more samples to satisfy requirements for analysis.

(3),(4),(5) $\rightarrow$ difficult to predict a response at a given point, since a new point will be far from the training examples.

Summary

- Data attribute types: nominal, ordinal, interval-, ratio-scaled
- Gain insight into the data by:
  - Basic ***statistical*** data ***description***: central tendency, dispersion
  - Normal & Poisson distributions
  - Display as box plots
- If high-dimensional, we need more data for density estimation