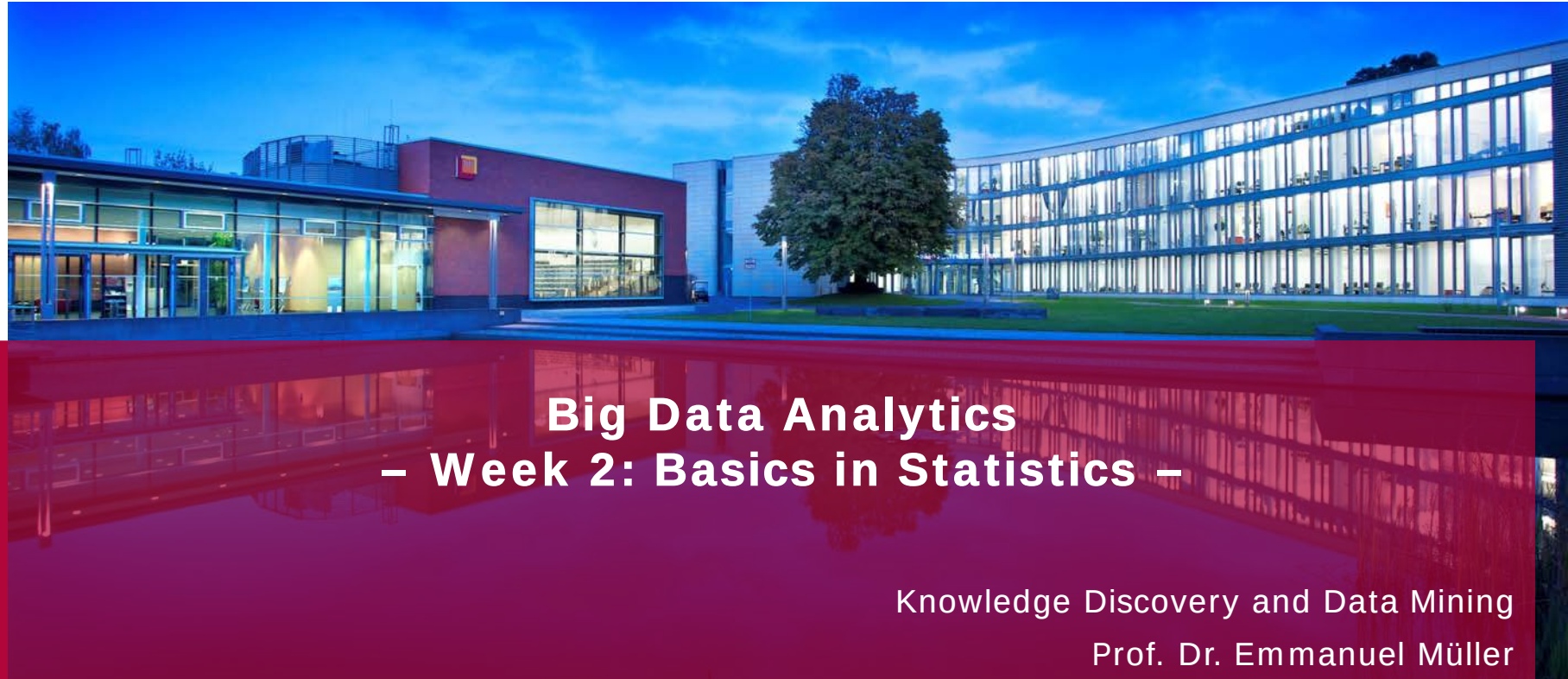


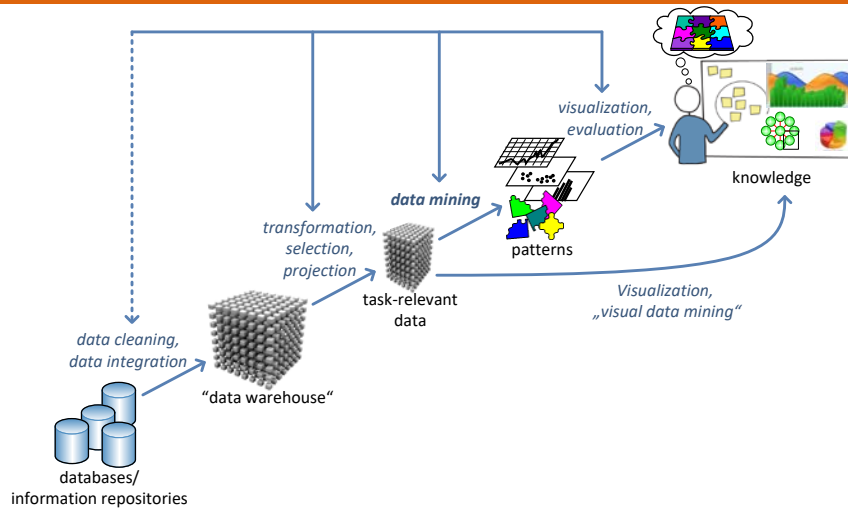


Big Data Analytics

– Week 2: Basics in Statistics –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

Understanding data



- Motivation
 - To better understand the data: **central tendency, variation and spread**
- Data dispersion characteristics
 - median, max, min, quantiles, outliers, variance, etc.

Categorization of Data

Values			
Categorical (qualitative)		Numerical (quantitative)	
Nominal	Ordinal	Discrete	Continuous
No intrinsic order	Exhibits order	Counting	Measurement
Colors (blue, red, yellow), Names (Alice, Bob, Eve), ...	Sizes (XS<S<M<L<XL<XXL), Grades (1>2>3>4>5>6), ...	Integers in equally-spaced Interval ([-20 .. 99]), No. of participants, ...	Time (0:00, 0:10, 0:17, 0:04), Blood pressure (121/68, 139/91 mmHg), ...

- One-Dimensional (18, 20, 43, 32, ...)
- Multi-dimensional ((1,4), (5,3), (6.2, 10.4), ...)
- High-dimensional (e.g. Time Series Data)
- No-dimensional (e.g. Protein Folding Structures)

Measuring the Central Tendency

Mean (weighted)		Midrange	Median	Mode
$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$	$(\max + \min) / 2$	The middle value	most frequent value(s)
Well-known measure for central tendency (<i>average</i>)		Average of the largest and the smallest value in a data set	Middle value if odd number of values, or average of the middle two values otherwise	Unimodal data has one most frequent value; data can be multimodal (bimodal, trimodal, ...)
Only applicable for numerical data		Applicable for numerical and metric data	Applicable for numerical and ordinal data, but not for nominal data	Can be problematic for continuous data (<i>no mode</i> if each data value occurs only once)
{1, 2, 3, 4, 5, 6} → 3.5 {2, 2, 2, 3, 3, 102} → 19		{1, 2, 3, 4, 5, 6} → $(6 + 1) / 2 = 3.5$ {2, 2, 2, 3, 3, 102} → 52	{1, 2, 3, 4, 5, 6} → 3.5 {2, 2, 2, 3, 3, 102} → 2.5	{1, 2, 3, 4, 5, 6} → No mode! {a, b, b, c, d, d} → Bimodal: b, d

Measures: Distributive

Distributive

Algebraic

Holistic

Distributive: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning.

Examples:

$\text{count}(D_1 \cup D_2)$

$= \text{count}(D_1) + \text{count}(D_2)$

$\text{sum}(D_1 \cup D_2)$

$= \text{sum}(D_1) + \text{sum}(D_2)$

$\text{min}(D_1 \cup D_2)$

$= \text{min}(\text{min}(D_1), \text{min}(D_2))$

$\text{max}(D_1 \cup D_2)$

$= \text{max}(\text{max}(D_1), \text{max}(D_2))$

Measures: Algebraic

Distributive

Algebraic

Holistic

Algebraic: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function.

■ Examples:

- $\text{avg}() = \text{sum}() / \text{count}()$
- $\text{standard_deviation}()$

$$\begin{aligned}\text{avg} (D1 \cup D2) &= \text{sum} (D1 \cup D2) / \text{count} (D1 \cup D2) \\ &= (\text{sum} (D1) + \text{sum} (D2)) / (\text{count} (D1) + \text{count}(D2)) \\ &\neq (\text{avg}(\text{avg}(D1), \text{avg}(D2))\end{aligned}$$

Measures: Holistic

Distributive

Algebraic

Holistic

Holistic: if there is no constant bound on the storage size which is needed to determine / describe a sub-aggregate

Examples:

- median: value in the middle of a sorted series of values
(= 50% quantile)
- $\text{median}(D1 \cup D2) \neq \text{simple_function}(\text{median}(D1) + \text{median}(D2))$
- mode: value that appears most often in a set of values
- rank: k-smallest / k-largest value (cf. quantiles, percentiles)

Measuring the Dispersion of Data

Quartiles, outliers and boxplots

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, Q_1 , median, Q_3 , max
- **Boxplot**: ends of the box are the quartiles, median is marked, whiskers; plot outlier individually
- **Outlier**: usually, values that are more than $1.5 \times IQR$ below Q_1 or above Q_3

Variance and standard deviation

- **Variance** (algebraic, scalable computation):
- **Standard deviation** σ : square root of variance

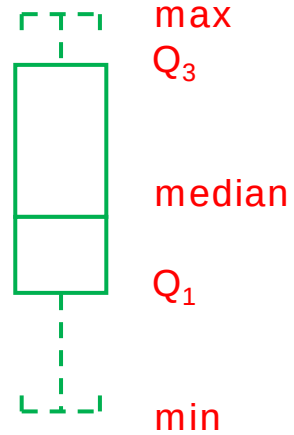
Boxplot Analysis

Five-number summary of a distribution

- Minimum, Q1, Median, Q3, Maximum
- = 0%, 25%, 50%, 75%, 100%-quantiles (25-percentile, etc.)

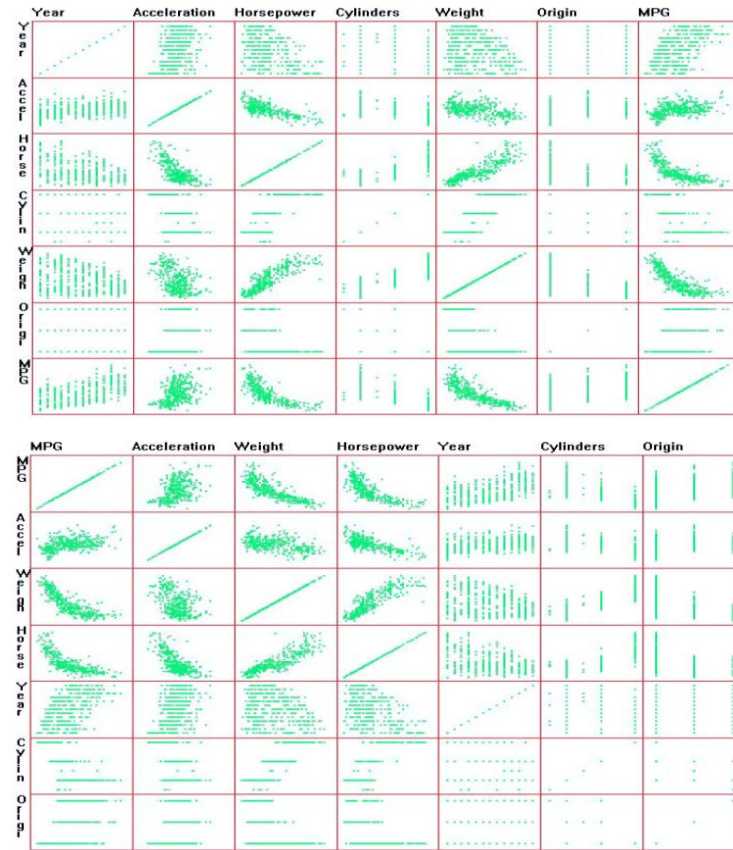
Boxplot

- 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 IQR
- The median is marked by a line within the box
- Whiskers: two lines outside the box extend to Minimum and Maximum



Scatterplot Matrix

- Matrix of scatterplots (x-y diagrams) of d -dimensional data
- Ordering of dimensions is important
- Dimension reordering
 - The interestingness of different orderings can be evaluated with quality metrics (e.g. Peng et al.)
 - Reduces clutter
 - Better visualization and understanding of data



Figures from Peng et al., Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering, IEEE Symp. on Inf. Vis., 2004.

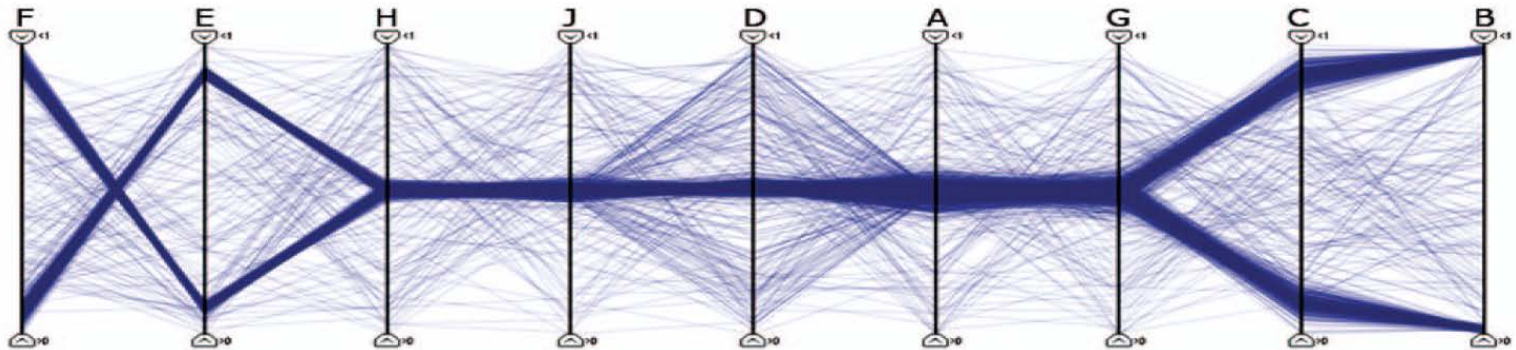
Parallel Coordinates

Inselberg, A.: The Plane with Parallel Coordinates, Special Issue on Computational Geometry. The Visual Computer, Vol. 1, pp. 69-97, 1985.

A d -dimensional data space is visualized by d parallel axes

Each axis is scaled to the min-max range in the corresponding dimension

Every data object is visualized as a polygonal line which intersects each of the axes at the point that corresponds to the value of the object in the respective dimension



Slide by Daniel Keim, Visual Techniques for Exploring Databases, Tutorial Slides, KDD 1997.

Figure from Bertini et al., Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, Trans. on Vis. and Comp. Graph., 2011.

Parallel Coordinates

Again, the ordering of the dimensions is important

Interestingness of an ordering can be measured with a quality metric

Quality or interestingness of orderings depends on what you want to visualize

Example:

- The first ordering is well-suited to visualize clusters in the data
- The second ordering is well-suited to visualize correlation between the dimensions

