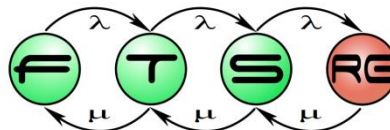


Visuelle Datenanalyse

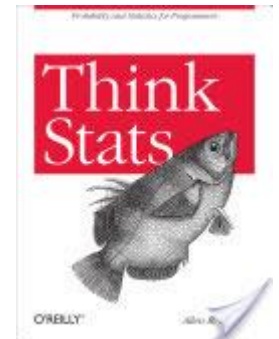
Dr. Pataricza András

**Budapest University of Technology and Economics
Fault Tolerant Systems Research Group**



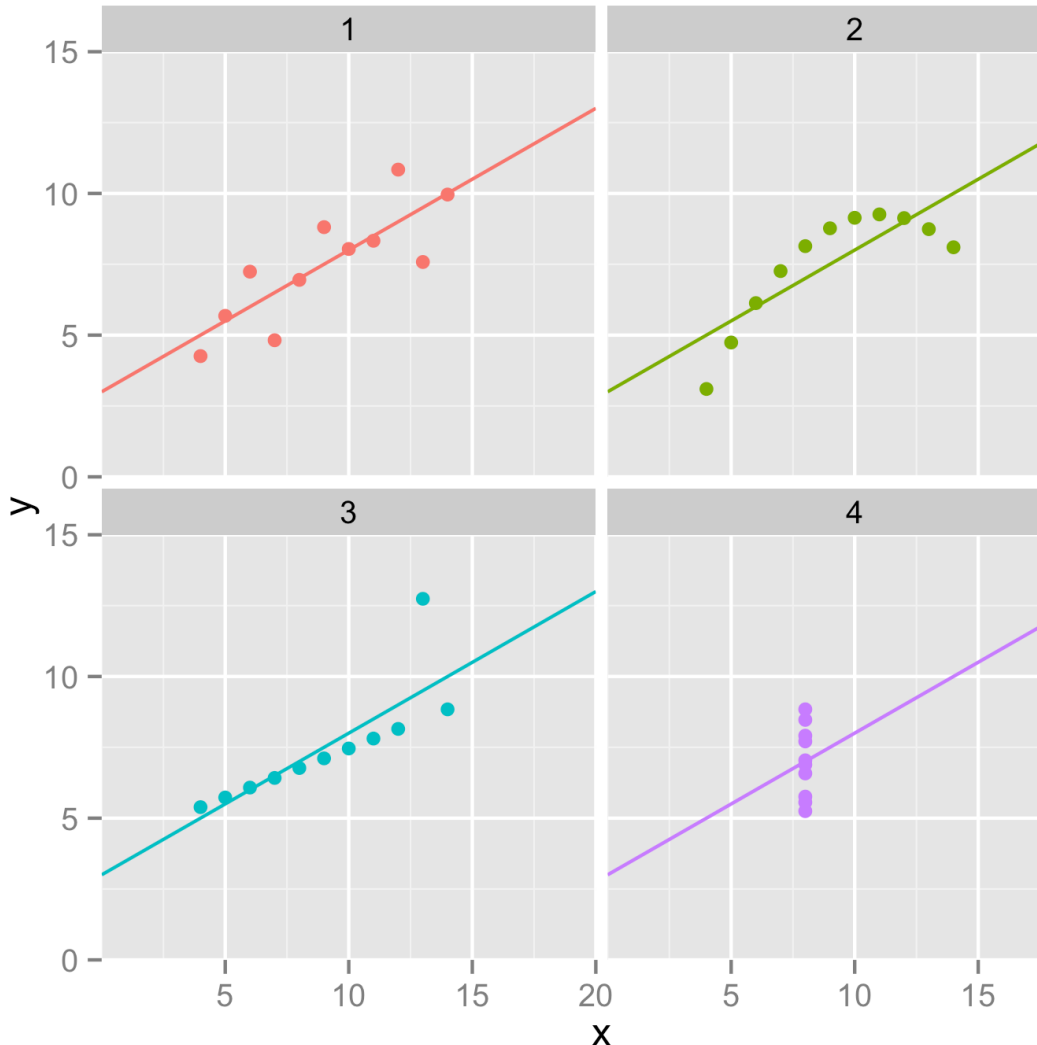
Wiederholung: Statistik (Grundlagen)

- Min: der kleinste Wert
- Max: der größte Wert
- Der Durchschnitt: $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
- Die Varianz (Streuung): $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$
- Think Stats: Probability and Statistics for Programmers



Überprüfung der Berechnungen

Anscombe's Quartet

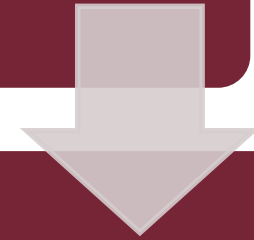


- 4 Datensätze mit gleichen:
- Durchschnitt von x (9)
 - Streuung von x (11)
 - Durchschnitt von y (7,5)
 - Streuung von y (4,12)
 - Korrelation der beiden (0,816)
 - Regressionsgerade

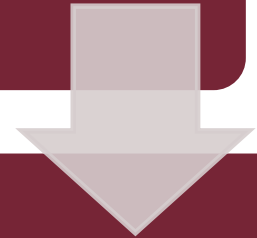
Vermeidung der
fehlerhaften Annahmen...
und Intuition

Inhalt

Visualisieren – Warum?

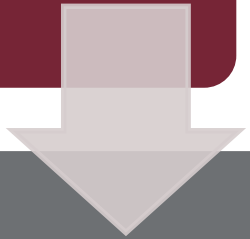


Visualisieren – Was?



Visualisieren – Wie?

Visualisieren – Warum?



Visualisieren – Was?



Visualisieren – Wie?

Visualisieren im Alltag

Analoge Anzeige



Digitale Anzeige



Analog + Koordinat

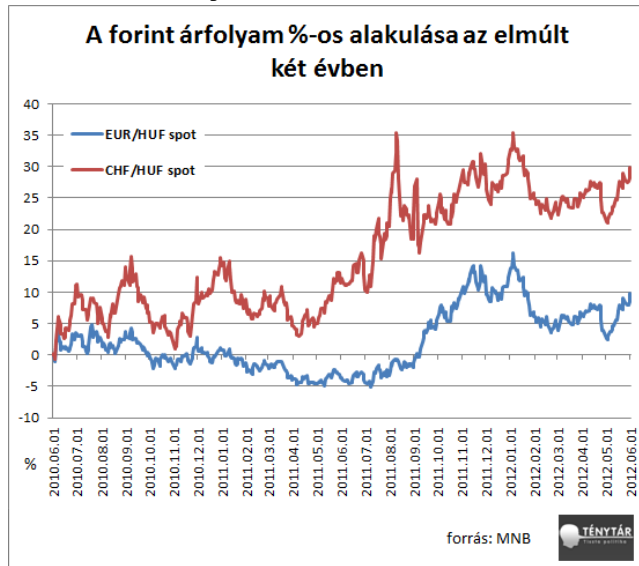


Hybride Anzeige



Visualisieren im Alltag

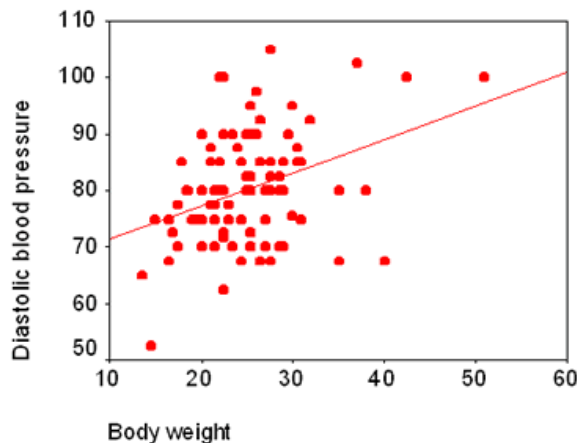
Trendanalyse und Vorhersage



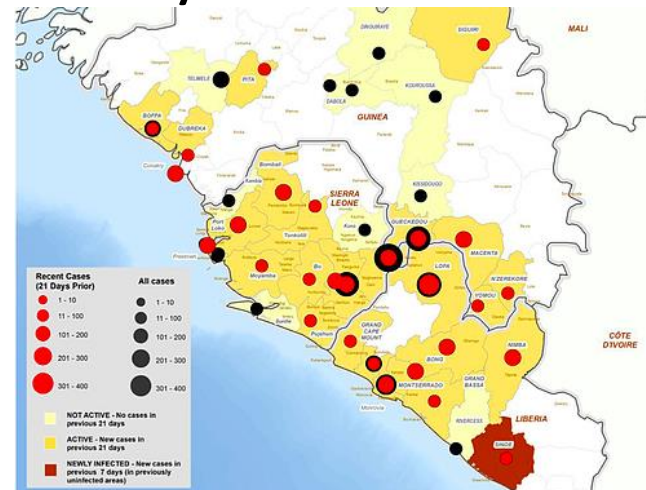
Zeitreihenanalyse



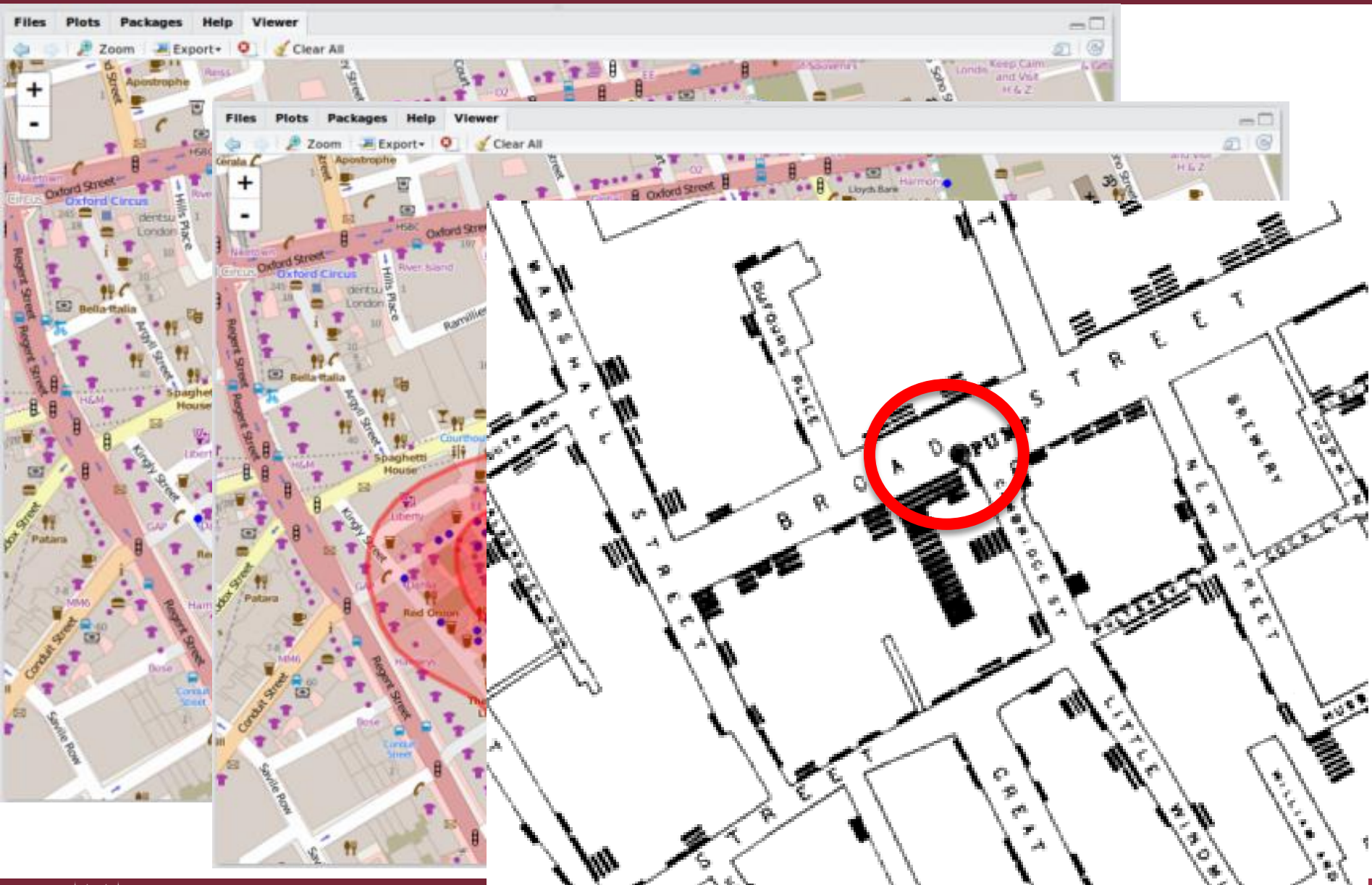
Korrelationsanalyse



Analyse der räumlichen Vert.



Erschließen von Zusammenhängen



Alle Augen auf die Daten!

„Massiv-Parallele“ Bearbeitung

- 120.000.000 Sensoren

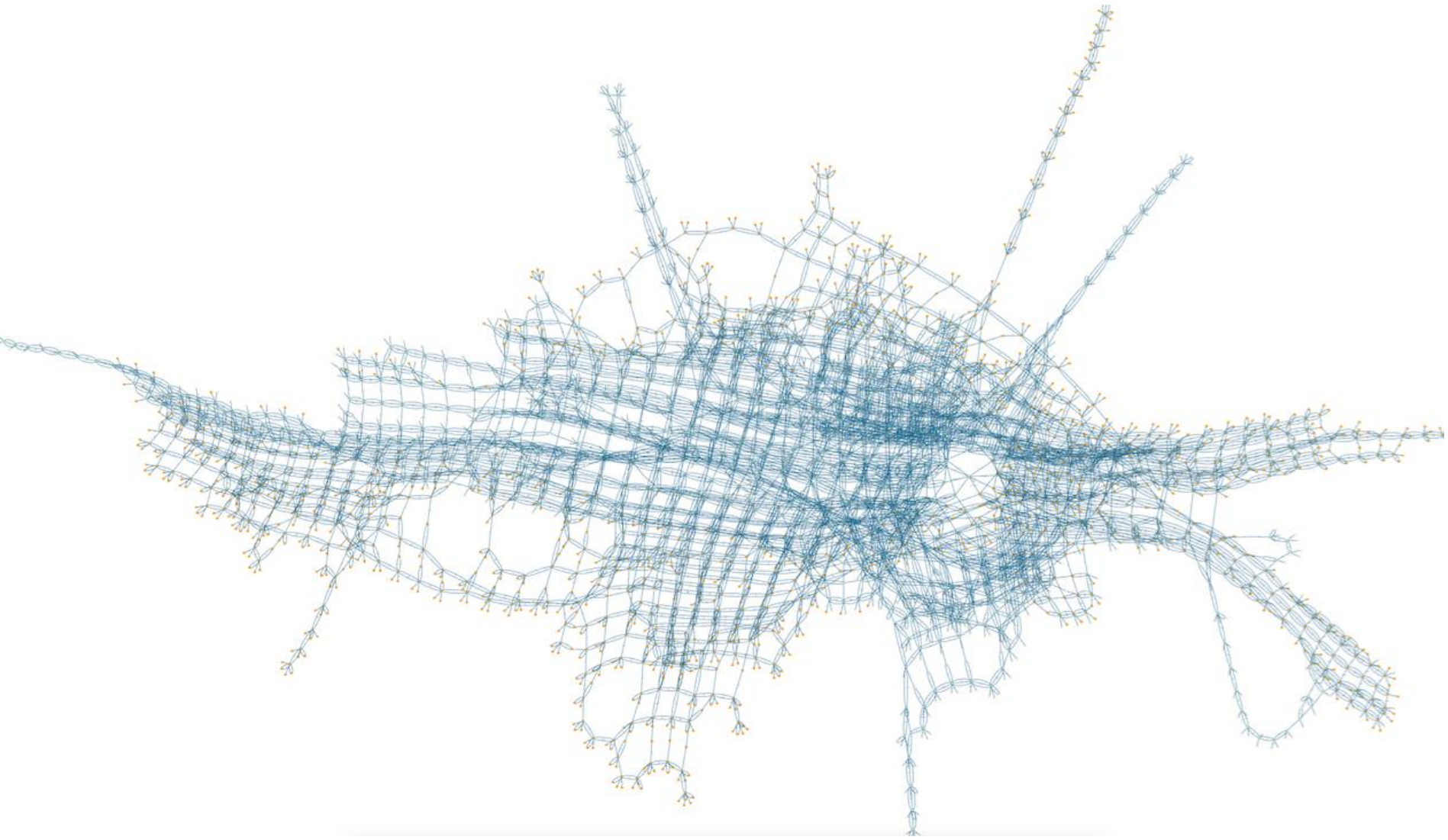
10010



3. Visuelle Auswahl und Manipulation

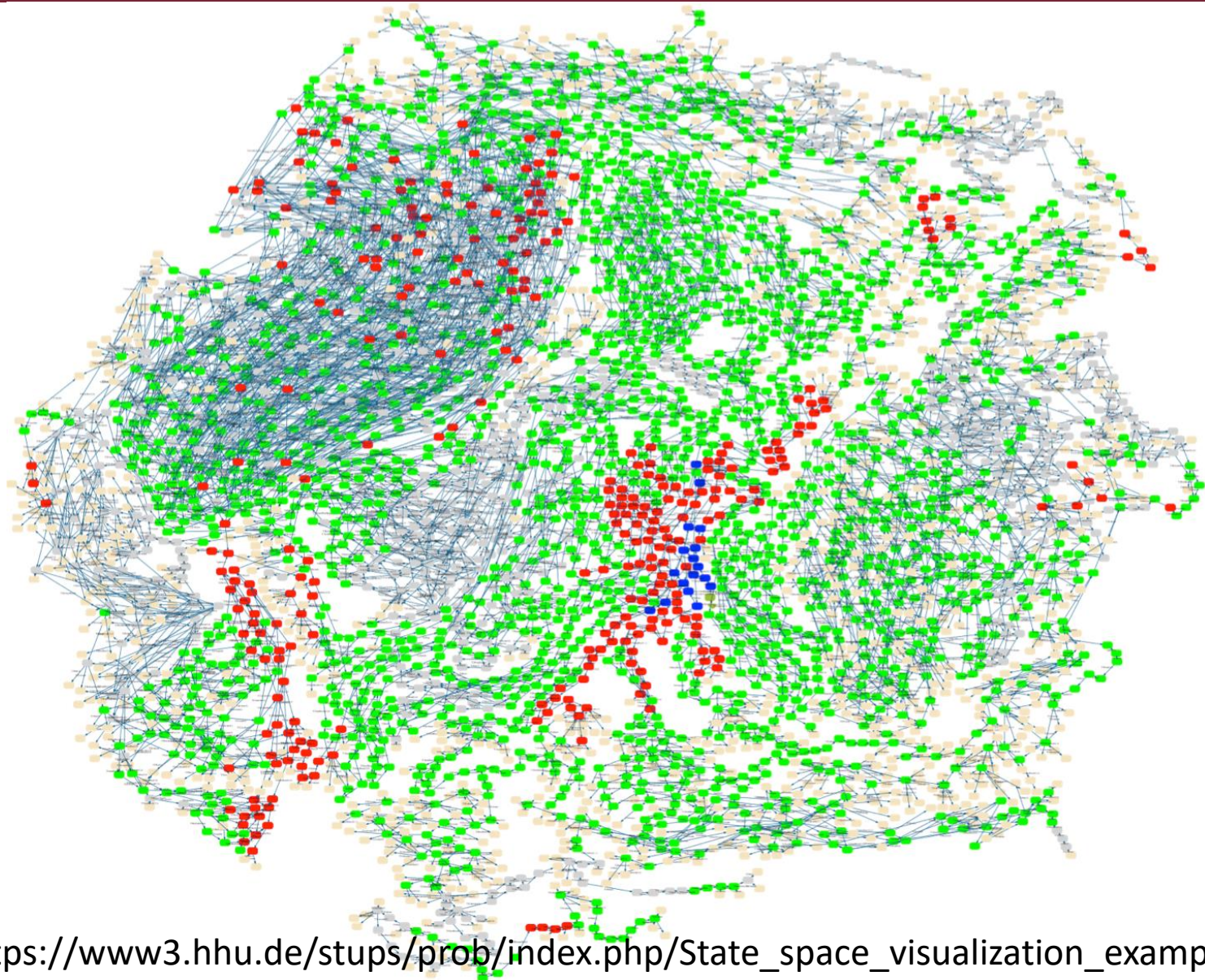
4. Interpretation, Korrelation mit anderen Modellen, Auswertung

Beispiel: Visualisierung des Zustandsraumes



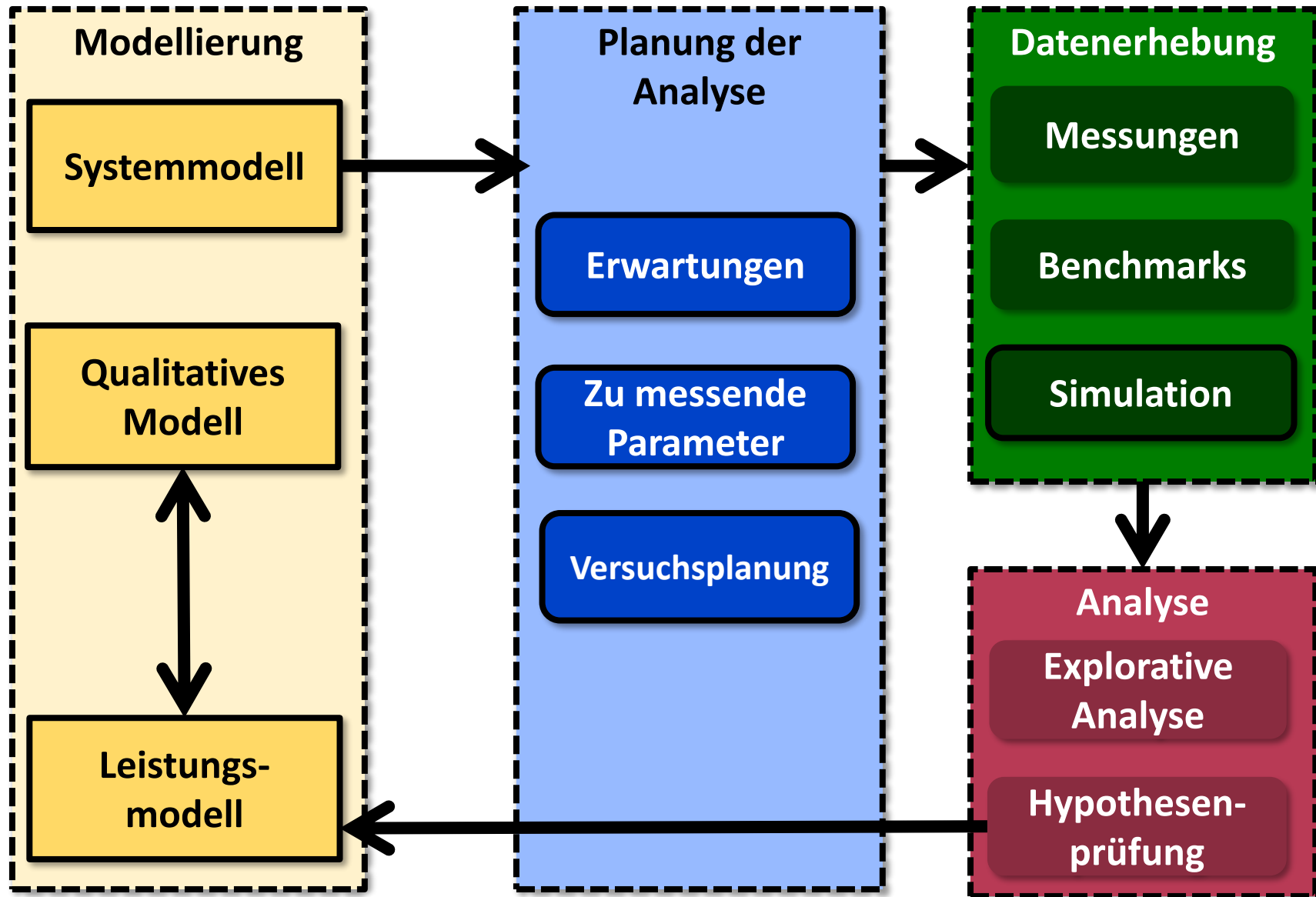
https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples

Beispiel: Zustandsraum des CAN-Bus

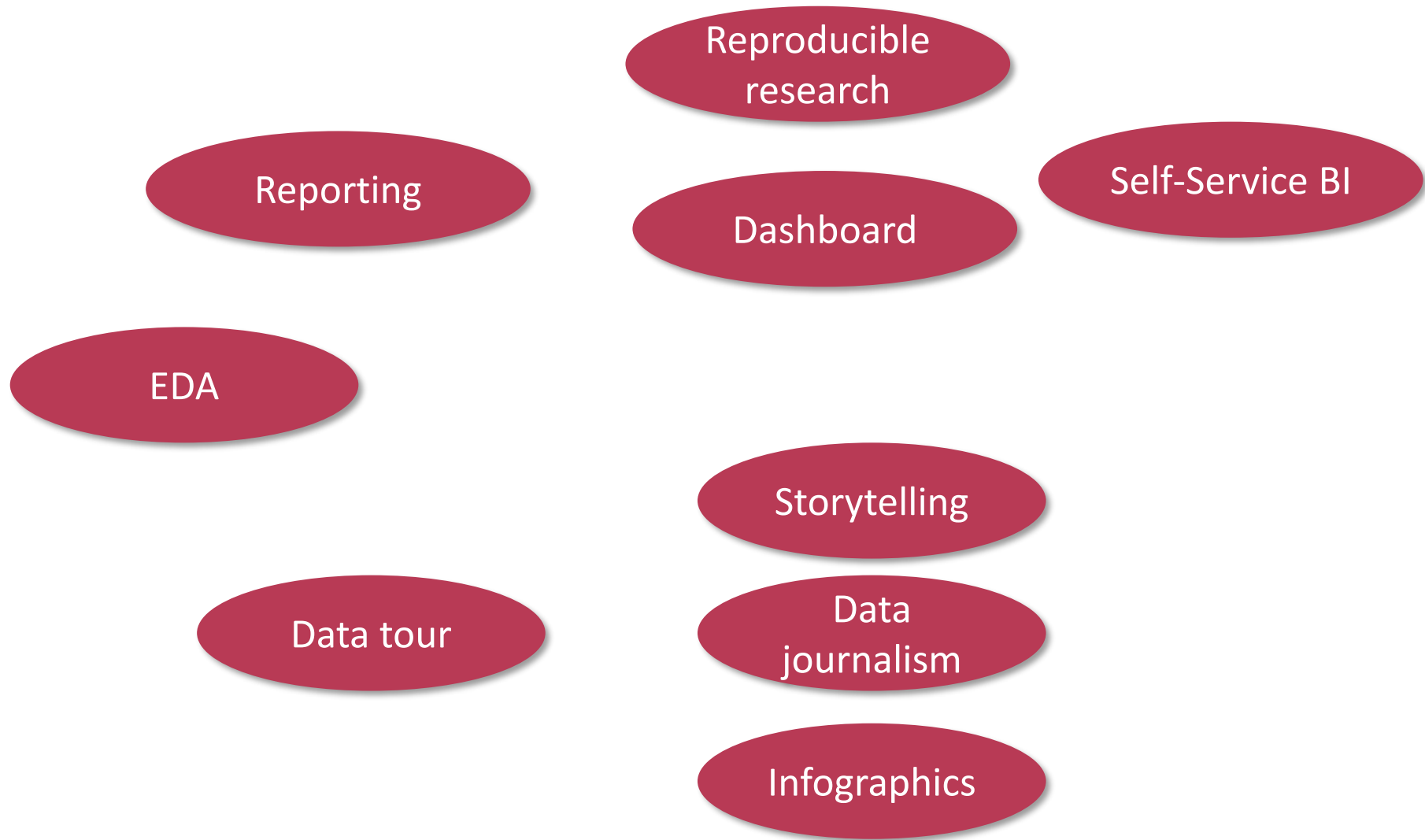


https://www3.hhu.de/stups/prob/index.php/State_space_visualization_examples

Beispiel: Systemmodell → Leistungsmodell

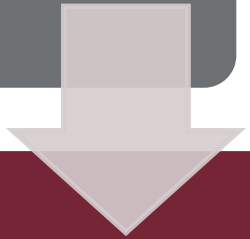


Ziele der Datenvisualisierung

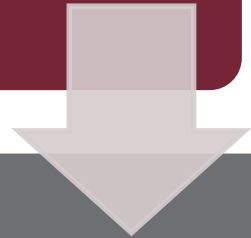


Inhalt

Visualisieren – Warum?



Visualisieren – Was?



Visualisieren – Wie?

Zur Erinnerung: Tabellarische Darstellung

- **Zeilen der Tabelle** = Modellelemente
- **Spalten der Tabelle** = Eigenschaften

Name ▼	Typ ▼	Grösse (kB) ▼	Letze Änderung ▼
Dokumente	Verzeichnis		2016.02.02
Verträge.pdf	Datei	569	2015.11.09
Bilder	Verzeichnis		2016.02.02
Logo.png	Datei	92	2015.03.06
Grundriss.jpg	Datei	1226	2016.02.02

- Datenanalyse Sprachen (z.B. R, Python): **dataframe**
 - Eine Zeile eine Messung
 - Die Spalten haben ihre eigenen **Typen**

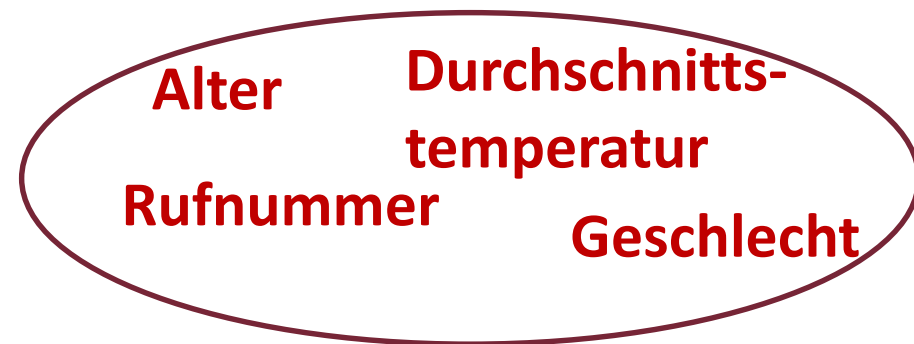
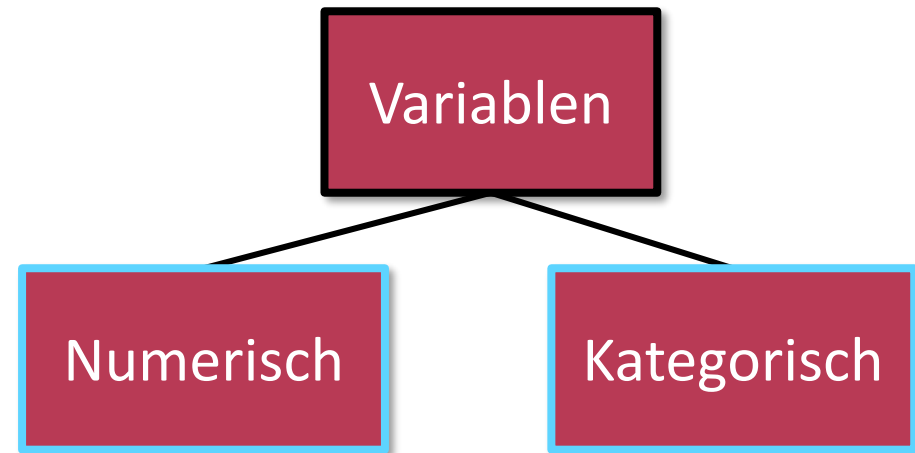
Numerische und kategoriale Variablen

- Numerisch (numerical)

- Die arithmetische Grundoperationen sind sinnvoll definierbar

- Kategorisch (categorical)

- Die Kategorien dienen der Unterscheidung, die Operationen sind sinnlos



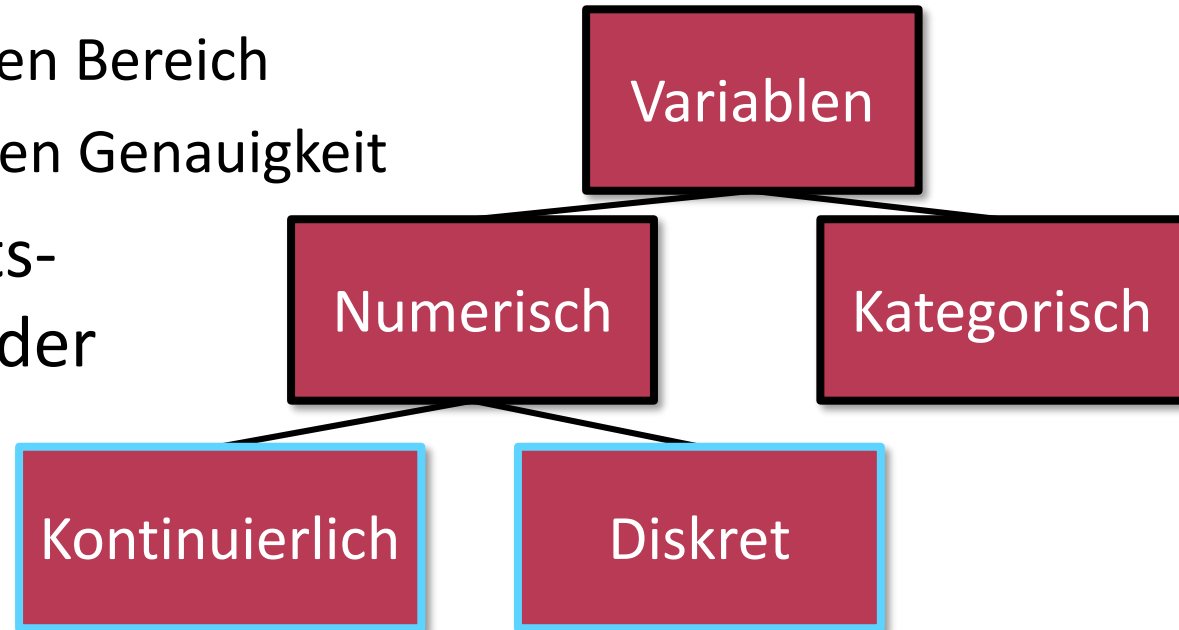
Numerische Variablen

■ Kontinuierlich

- Gemessen – kann beliebige Werte aufnehmen
 - in dem gegebenen Bereich
 - bei der gegebenen Genauigkeit
- z.B. Durchschnitts-Klausurergebnis der Anwesenden

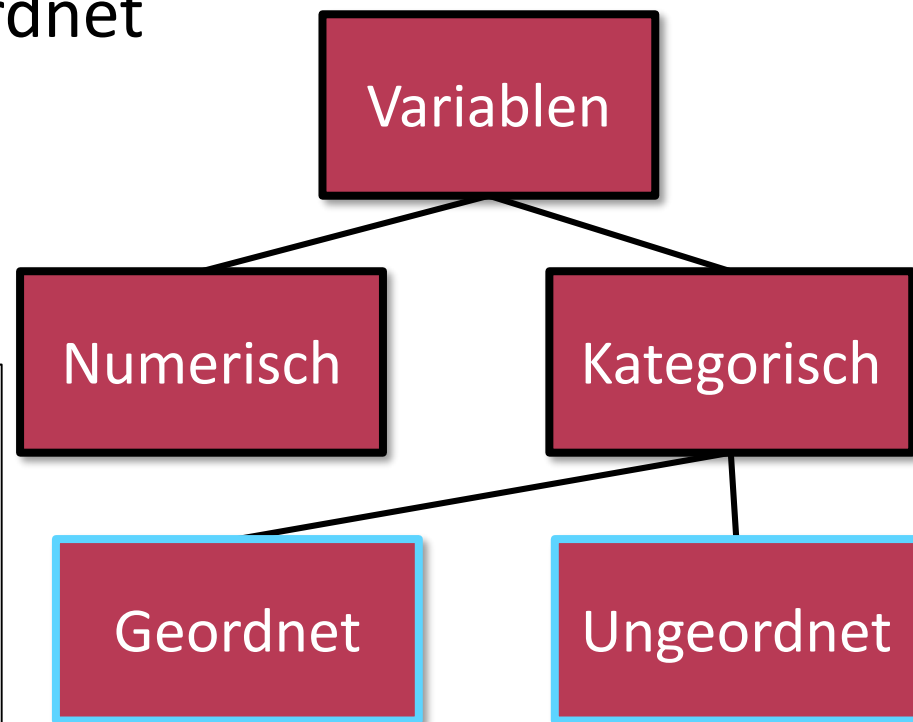
■ Diskret

- Gezählt – kann nur endlich viele verschieden Werte aufnehmen
- z.B. Anzahl der Anwesenden



Kategorische Variablen

- Geordnet (Kategorisch → Ordinal)
 - Der Wertebereich ist geordnet
 - Vollständig geordnet?
- Ungeordnet (regulär)



9. Ajánlanád-e a tárgyat másoknak?

- ☐ Mindenkit rábeszelnék
- ☐ Nyugodtan ajánlanám
- ☐ Esetleg ajánlanám
- ☐ Inkább lebeszelném róla
- ☐ Feltétlenül lebeszelném
- ☐ Nem kívánok válaszolni

Inhalt

Visualisieren – Warum?



Visualisieren – Was?



Visualisieren – Wie?

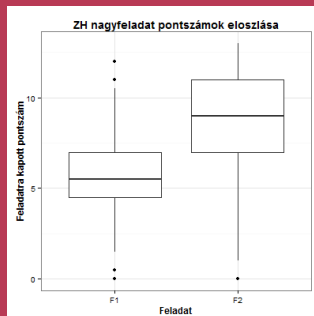
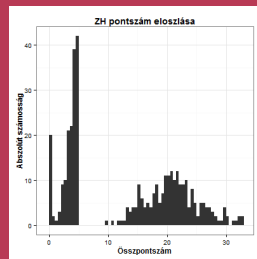
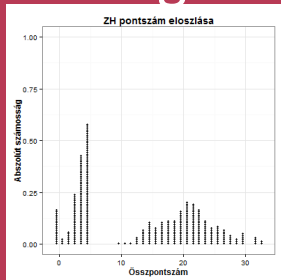
Verteilung einer Variable

Variablen

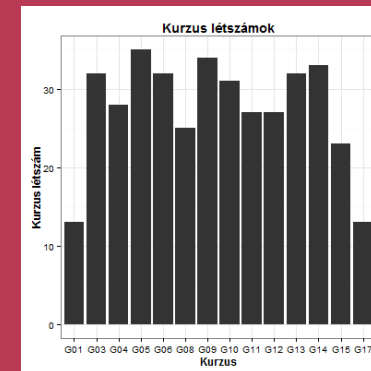
Numerisch

Kategorisch

KI-Ergebnis: [13, 15, 2, ...]



Kurse: [G01, G03, G15, G17, ...]

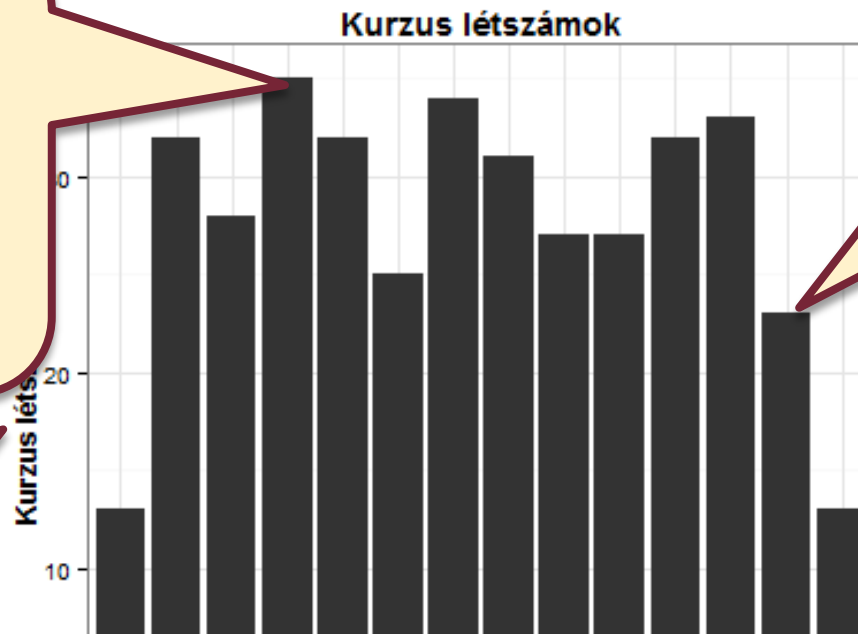


Säulendiagramm/Balkendiagramm

- Eingabevariable: Kurs-Kode
- Frage: wie viele haben sich angemeldet?

Gibt es populäre Zeitpunkte/Übungsleiter?

absolute Häufigkeit!



Säulenhöhe:
Häufigkeit des
gegebenen
Wertes

Entwurfsentscheidung: Aufteilung des Definitionsbereiches
z.B.: Dienstag-Donnerstag-Freitag?

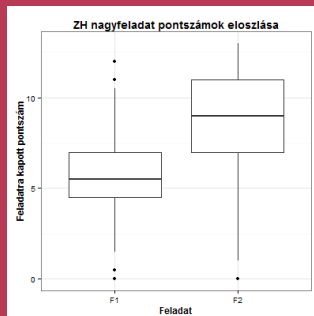
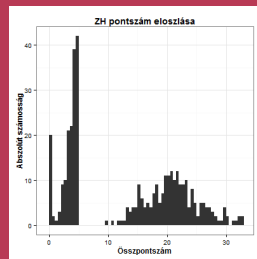
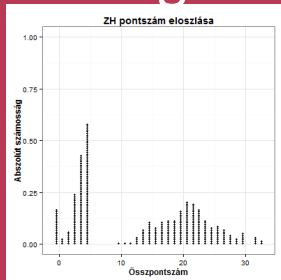
Verteilung einer Variable

Variablen

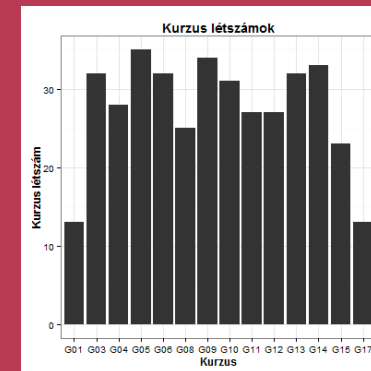
Numerisch

Kategorisch

KI-Ergebnis: [13, 15, 2, ...]

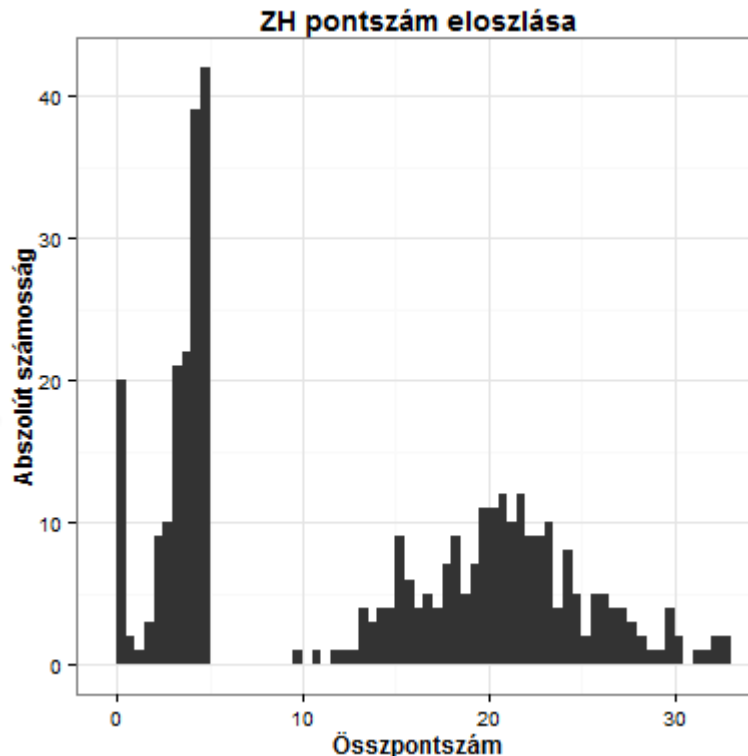


Kurse: [G01, G03, G15, G17, ...]



Histogramm

- Eingabevariable: Klausurergebnis
- Frage: was für Ergebnisse sind geboren?



absolute
Häufigkeit!

Säulenhöhe:
Häufigkeit des
gegebenen
Intervalls

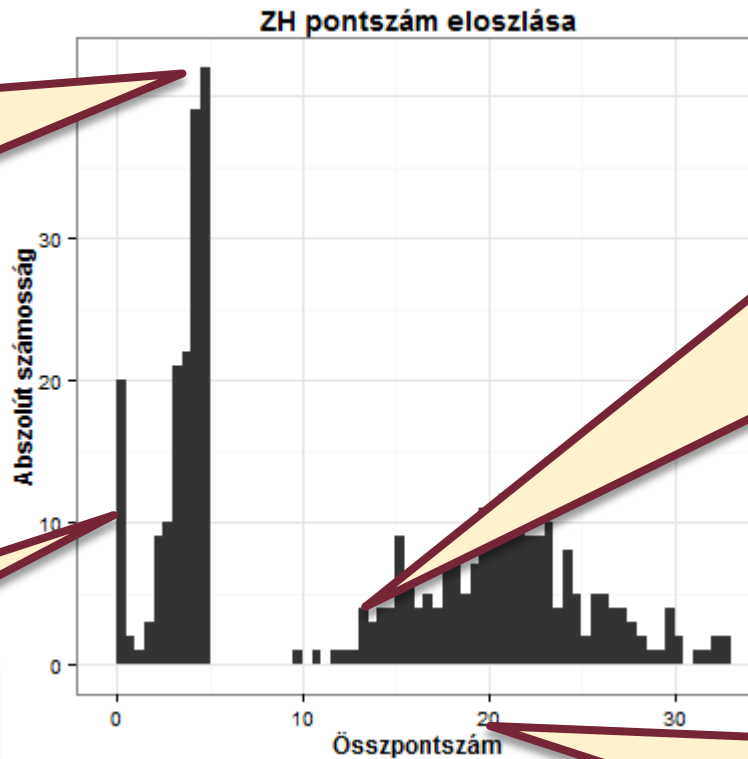
Entwurfsentscheidung: Bestimmung der Intervalllänge
Z.B.: 1-Punkte-Auflösung vs. 0,5-Punkte-Auflösung?

Histogramm

- Eingabevariable: Klausurergebnis
- Frage: was für Ergebnisse sind geboren?

Viele haben das Test nur fast geschafft

Die gar nicht erschienen

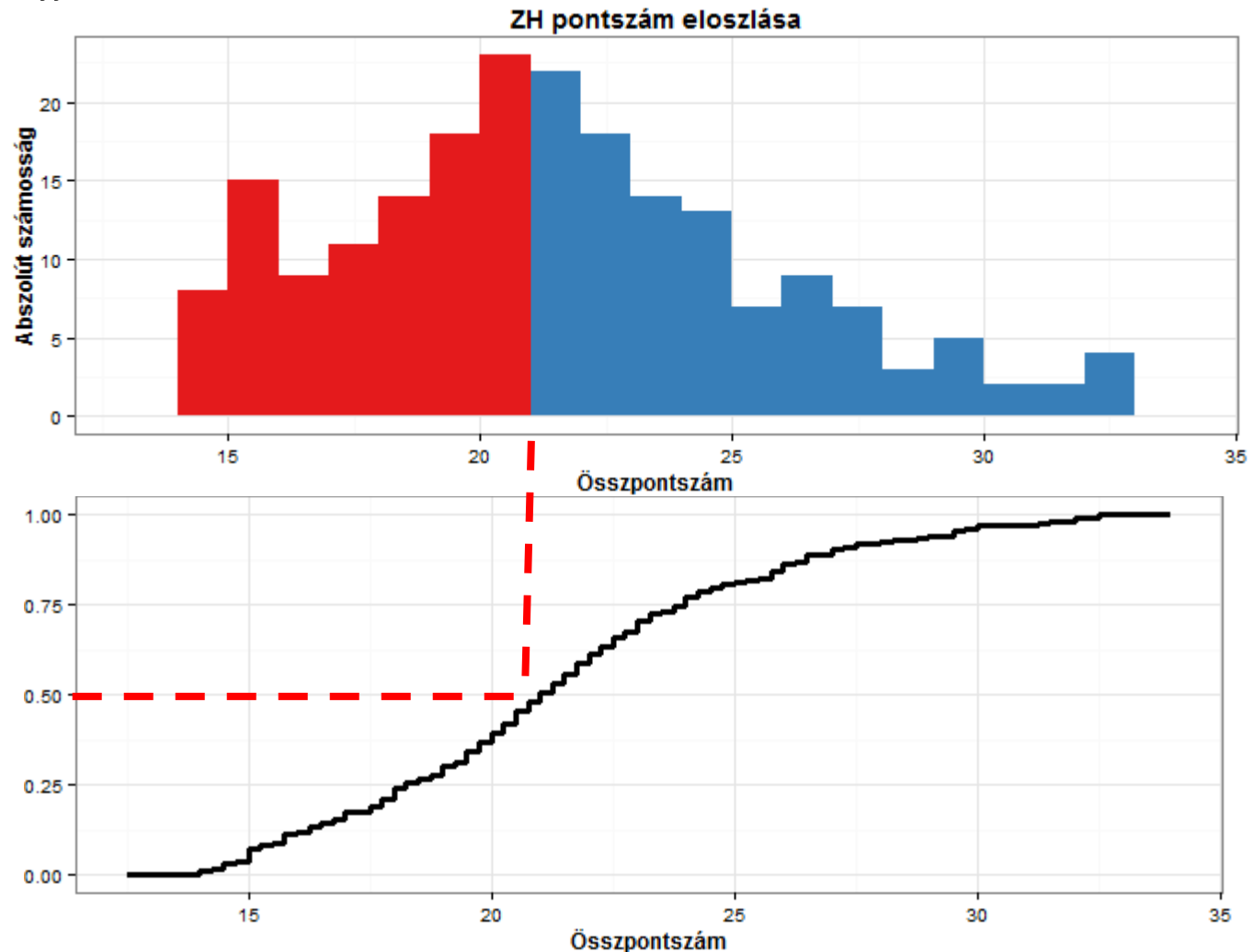


Die meisten von denen, die das Test schon bestanden haben, haben die Klausur geschafft.

Der Durchschnitt/Median war so um 20 Punkte

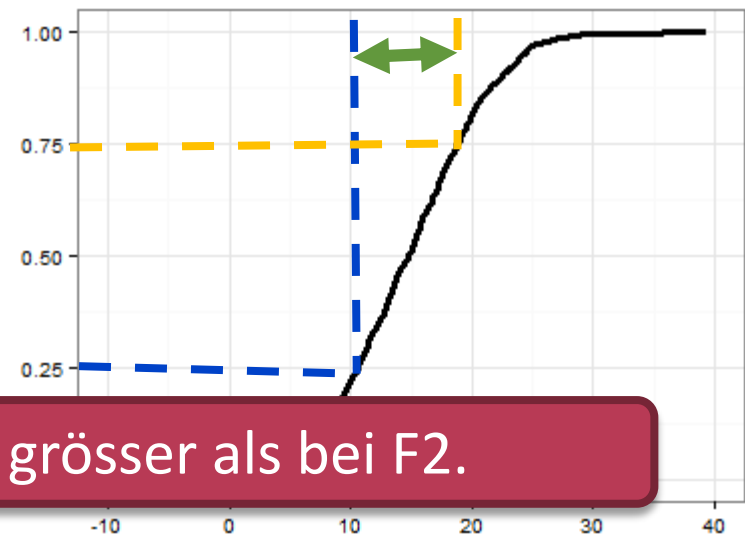
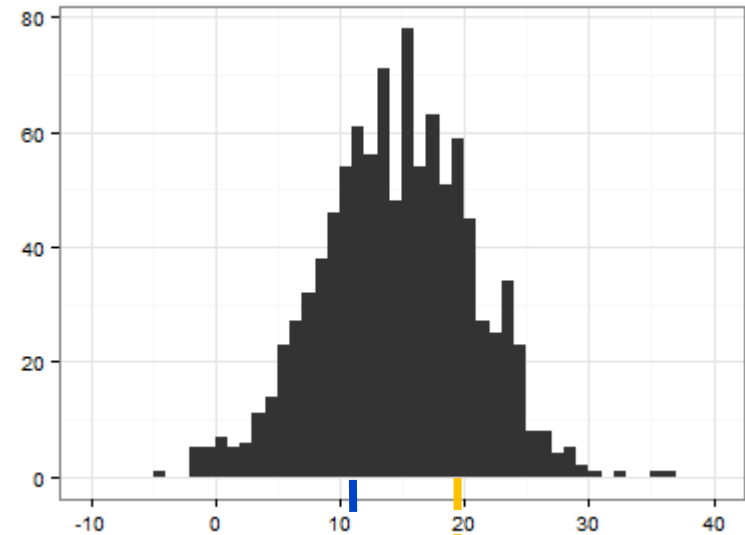
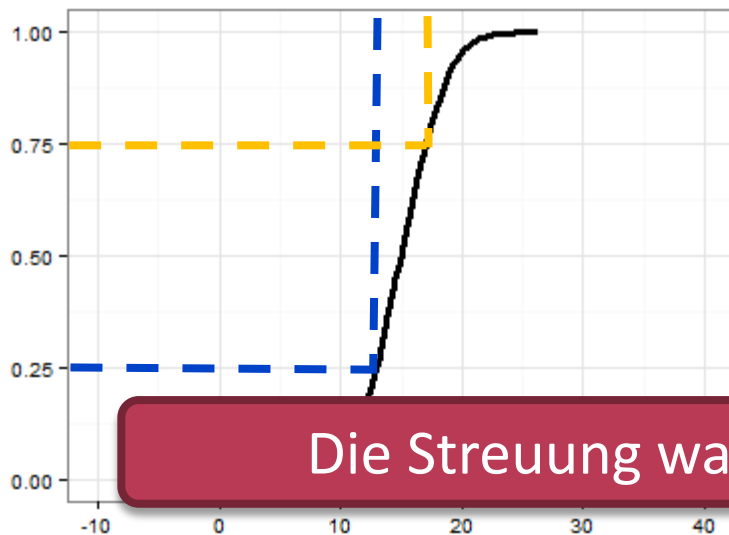
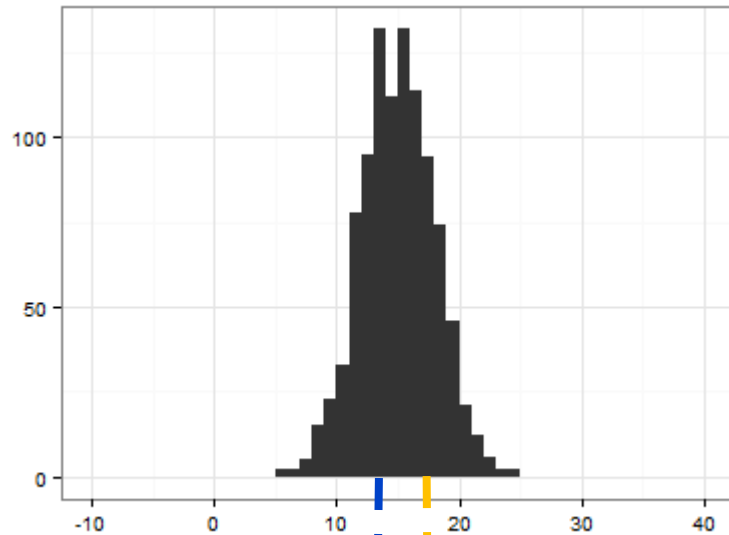
Einfache statistische Beschreibung

- Wo ist „die Mitte“ der Werte?



Einfache statistische Beschreibung

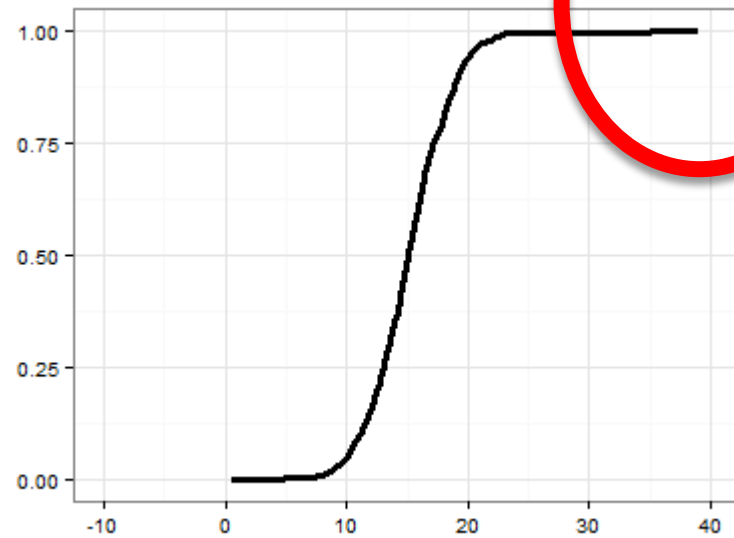
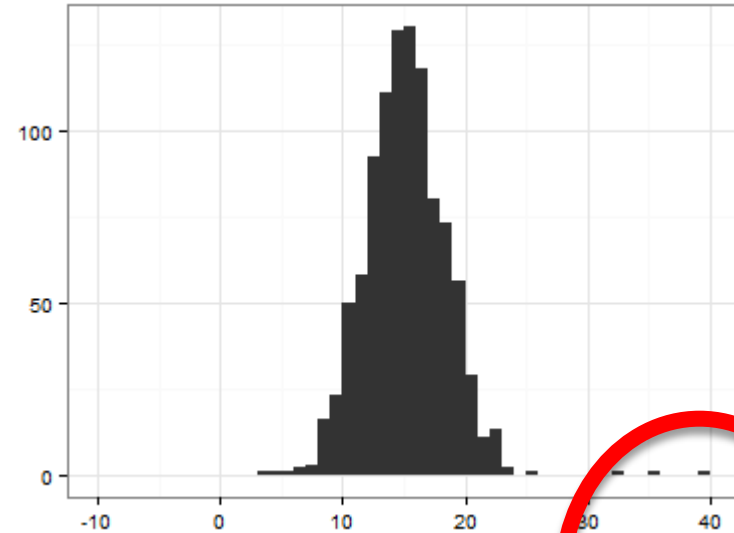
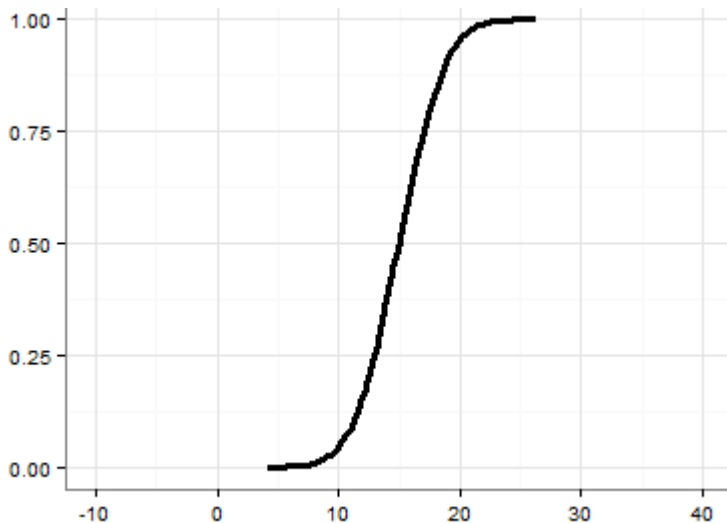
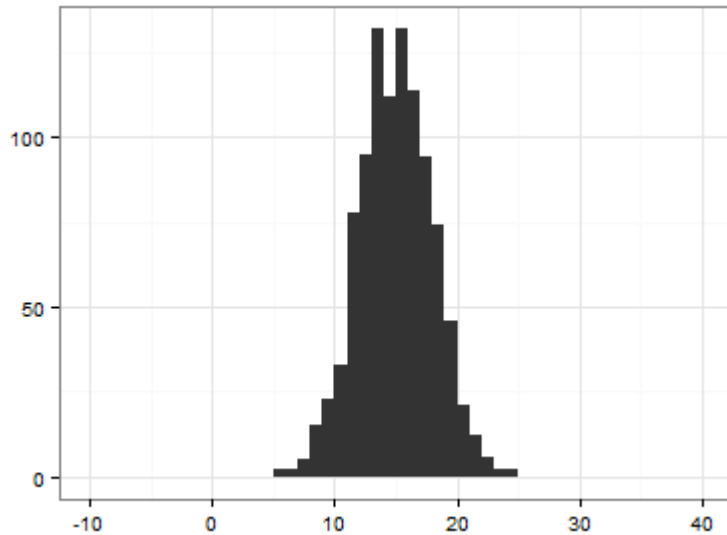
- Wie weit sind die Werte „gestreut“?



Die Streuung war bei F1 grösser als bei F2.

Einfache statistische Beschreibung

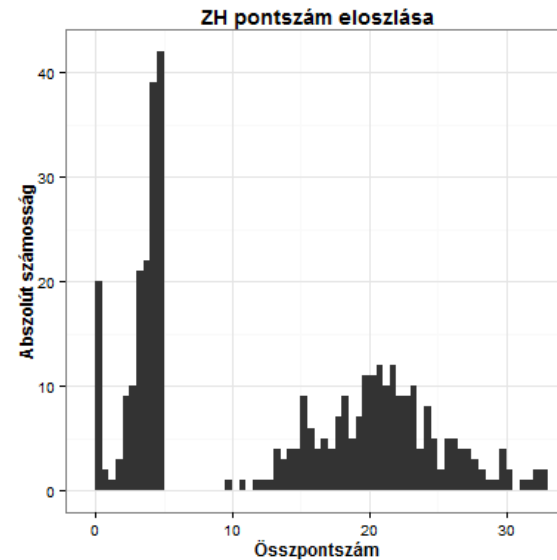
■ Gibt es Ausreisser?



Kastengraphik/Boxplot

- Eingabevariable: Klausurergebnis
- Frage: was für Ergebnisse sind geboren
so ungefähr?

Eine Art Abstraktion:
nehmen wir Intervalle,
die einzelnen Werte sind
uninteressant



Beschreibung (kontinuierlicher) Beobachtungen

- Beschreibung der „Mitte“
 1. Durchschnitt (Mittelwert) – arithmetisches Mittel
 2. **Median** (Zentralwert) – mittleres Element (geordnet)
 3. Modalwert – häufigstes Element
 - Beispiel: {3, 4, 4, 5, 5, 6, 10, 20}
 - Durchschnitt: ~ 7.125
 - Median: 5
 - Modalwert: 4 und 5
- Beschreibung der „Ausdehnung“
 - Perzentile (Frequenzen für kategoriale Typen)

Zentralwert anstatt Mittelwert – Warum?

- Grundmenge: 1000 Punkte $\sim U(1, 5)$ mit gleichmässiger Verteilung

- *Mittelwert = Zentralwert = 3 ms*



3ms $\pm$ 2 ms



Neuer Zentralwert: `sort(resp. times)[501]` = 3.004 ms

Zentralwert



Mittelwert



Neuer Mittelwert: $(2 * 10^4 + 3 * 10^3) / 1001 = 23 \text{ ms!}$

Beschreibung (kontinuierl.) Beobachtungen

Wenn die Elemente eines Datensatzes steigend geordnet sind, ist das mittlere Element **der Median** des Datensatzes. Falls es kein mittleres Element gibt (gerade Elementenanzahl), ist **der Median** der Durchschnitt der zwei mittleren Elemente.

Das Modalwert ist das häufigste Element (die häufigste Elemente) des Datensatzes. Wenn das häufigste Element nicht eindeutig ist, hat der Datensatz mehrere Modalwerte.

Perzentile

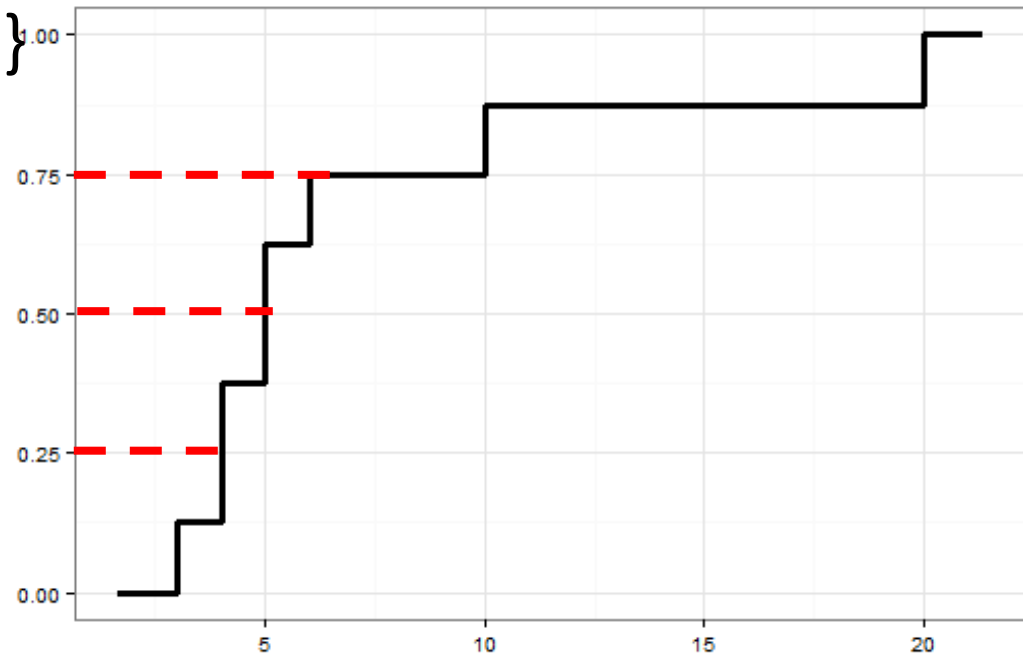
■ Perzentil

- $n\%$ der Werte sind kleiner als das **n -te Perzentil**

Frequenz: $n\%$ der Werte fallen in die gegebene Kategorie(n)

- {3, 4, 4, 5, 5, 6, 10, 20}

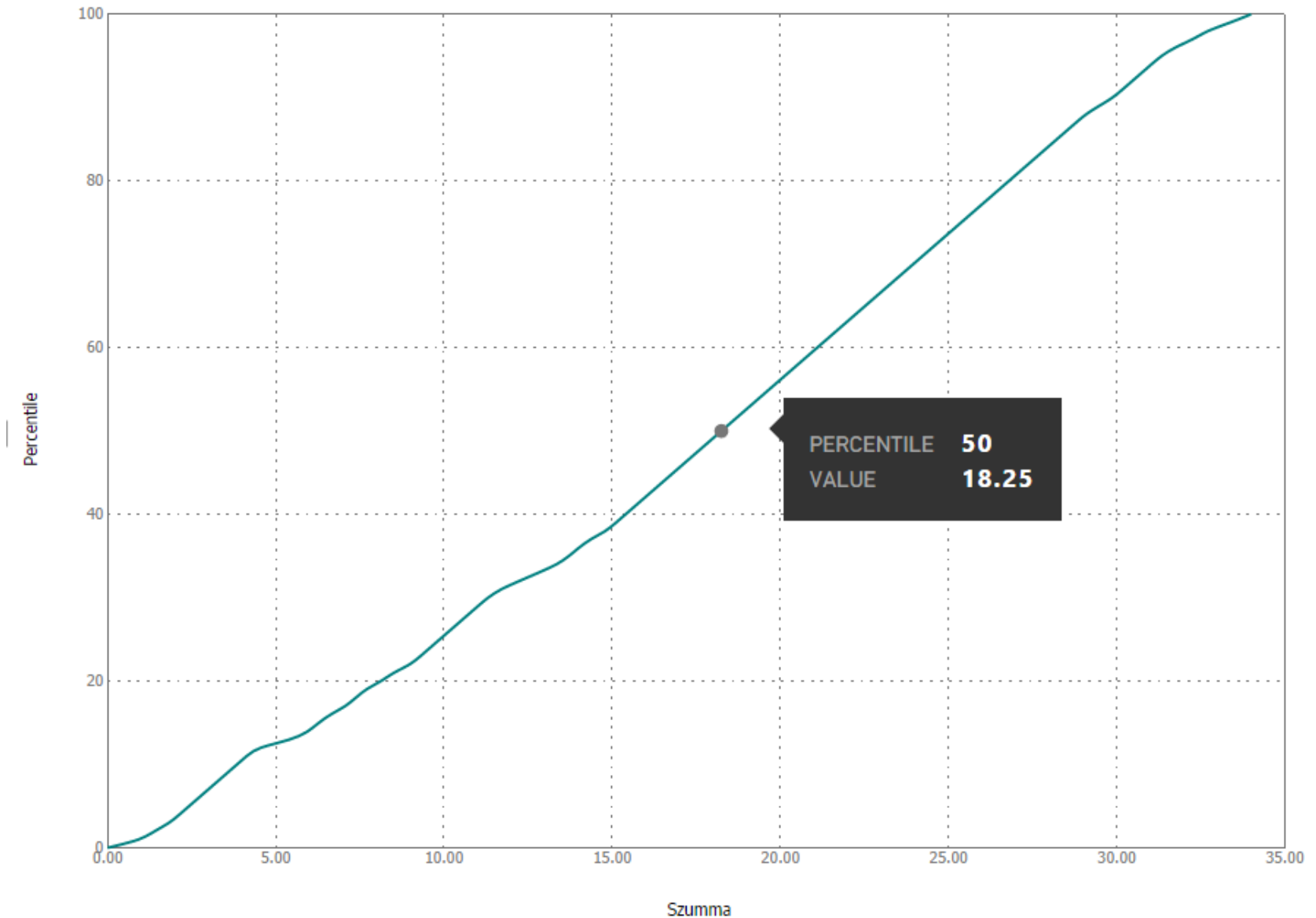
- 50. Perzentil: 5
- 25. Perzentil: 4
- 75. Perzentil: 10



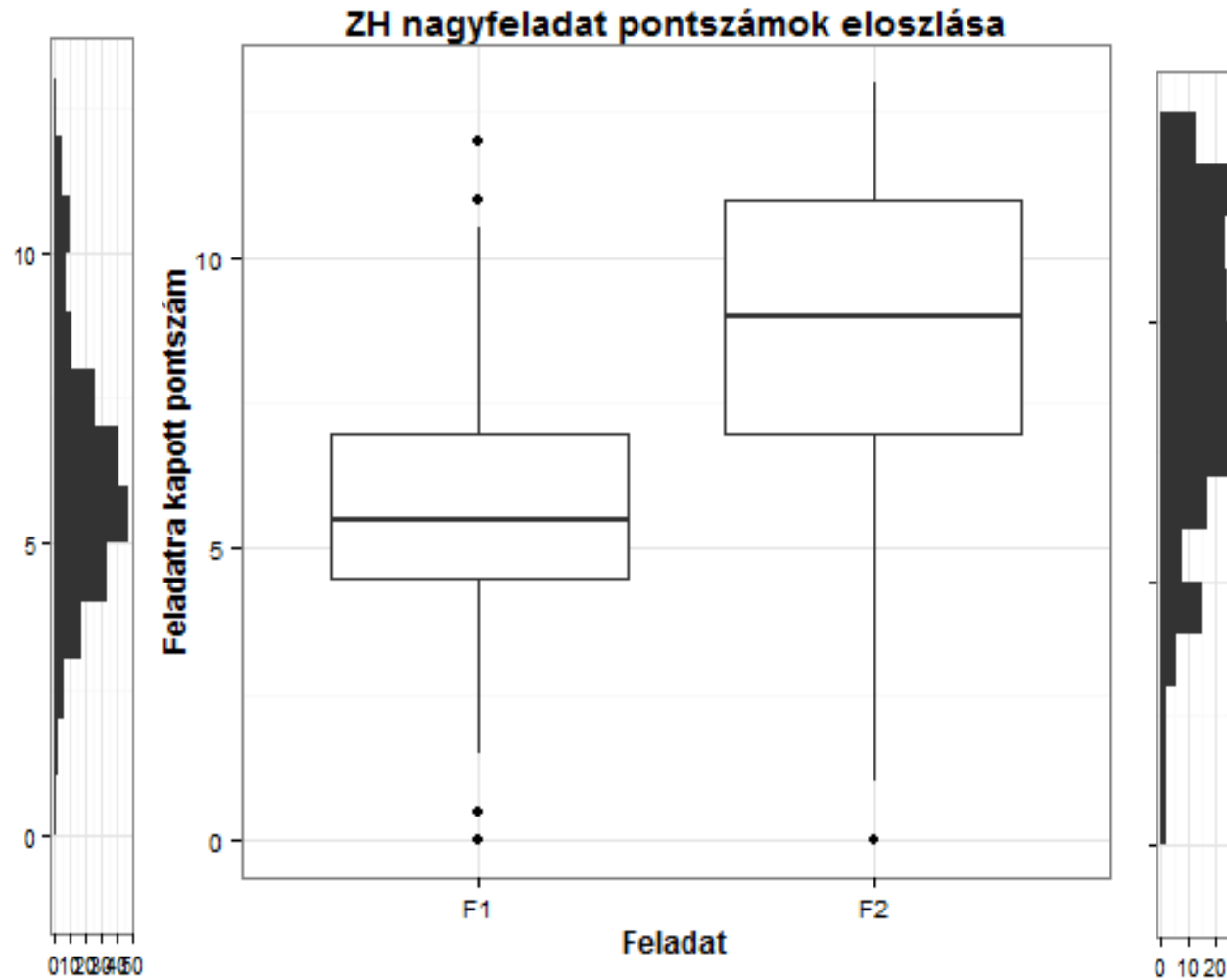
■ Quartil

- Q1: das 25. Perzentil
- Q3: das 75. Perzentil
- **Q2: der Median**

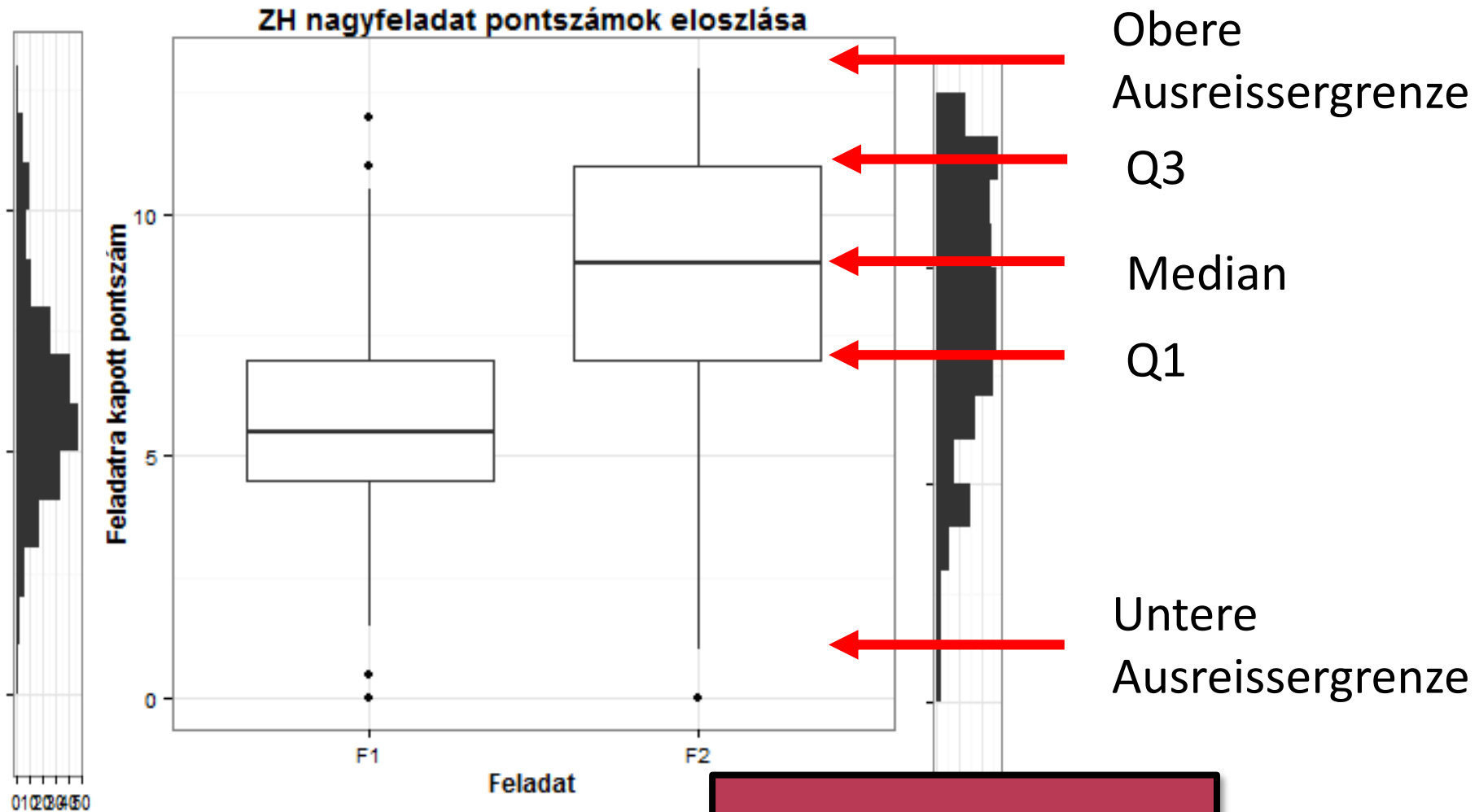
Beispiel: Darstellung der Perzentile



Kastengraphik (Box and whisker plot)



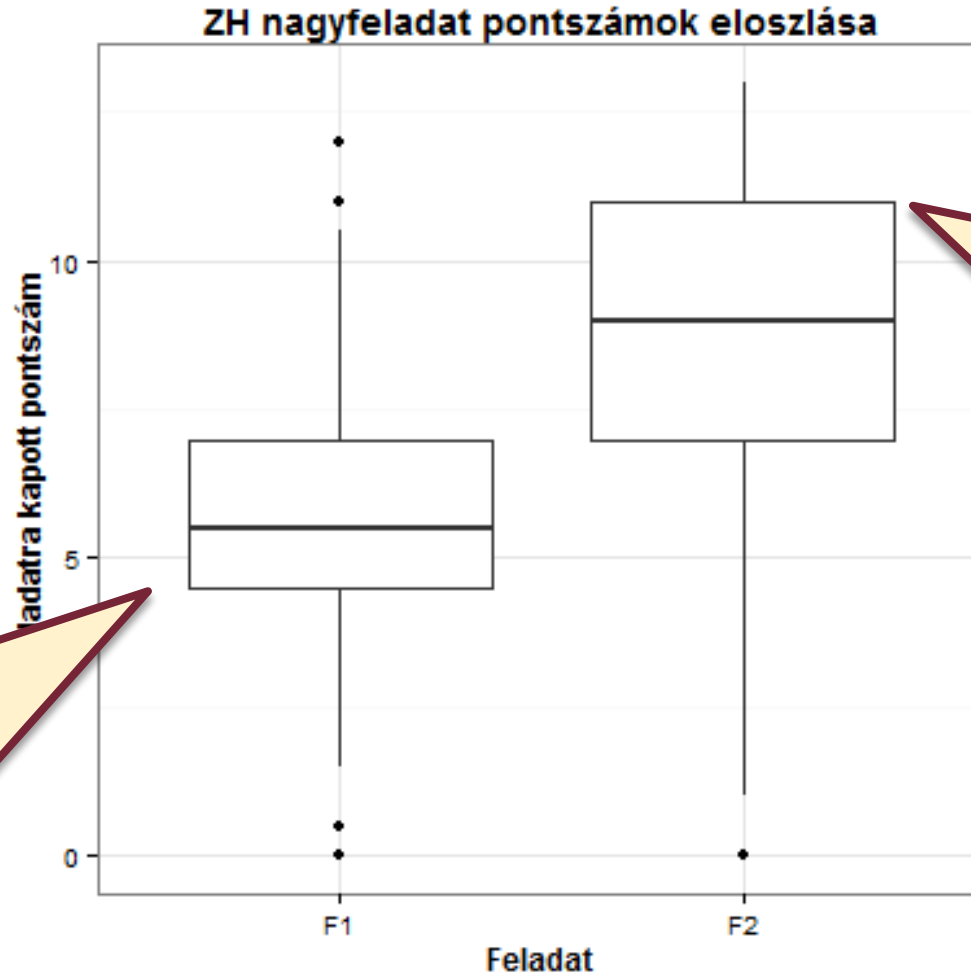
Kastengraphik (Box and whisker plot)



Wie geht es in Excel?

(www.youtube.com/watch?v=ucWmfmXb1kk)

Kastengraphik (Box and whisker plot)

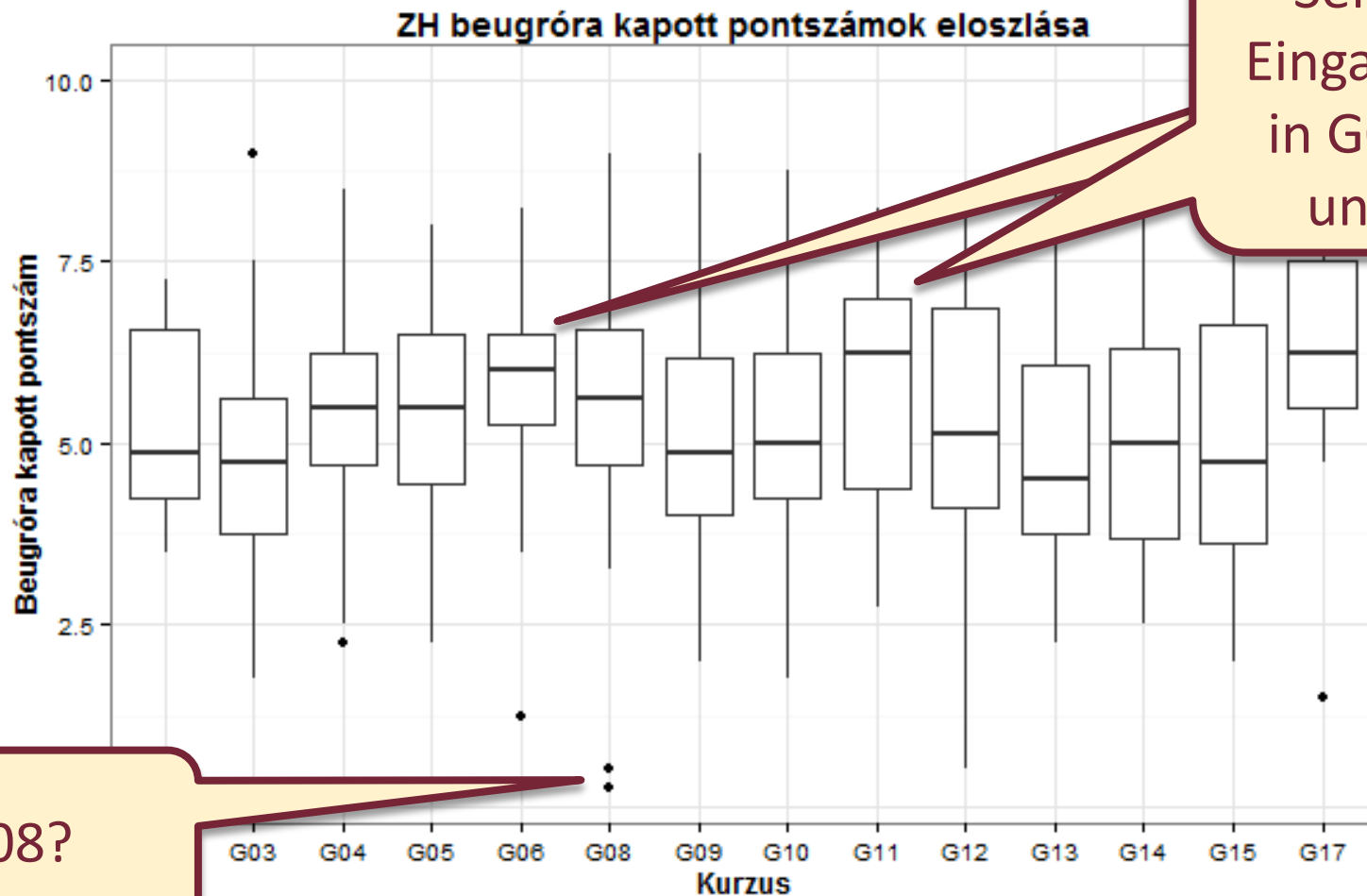


50% der Ergebnisse der Aufgabe 1 liegen zwischen 4.5 und 7.5.

Aufgabe 2 lieferte (im Allg.) bessere Ergebnisse als Aufgabe 1.

Kastengraphik (Box and whisker plot)

- Was waren die Ergebnisse per Übungsgruppe?

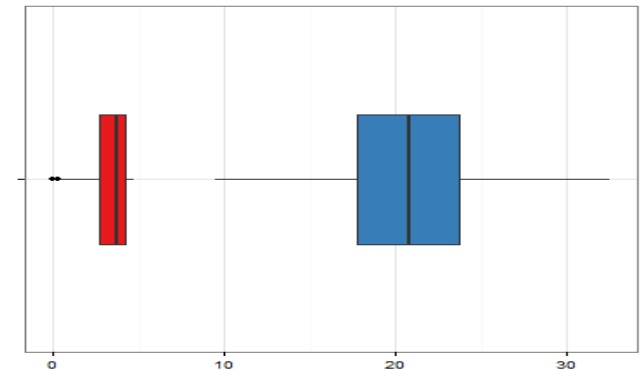
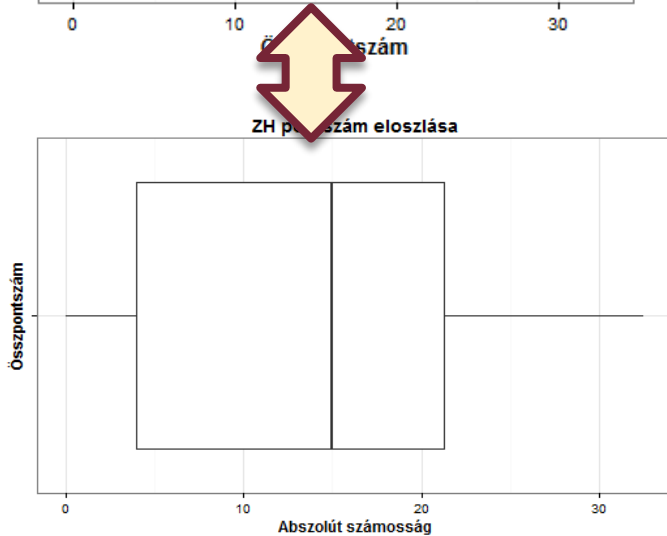
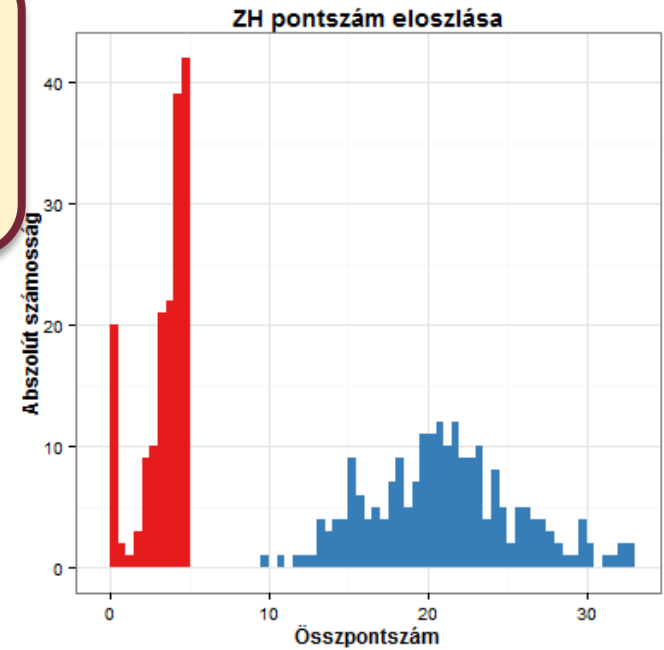
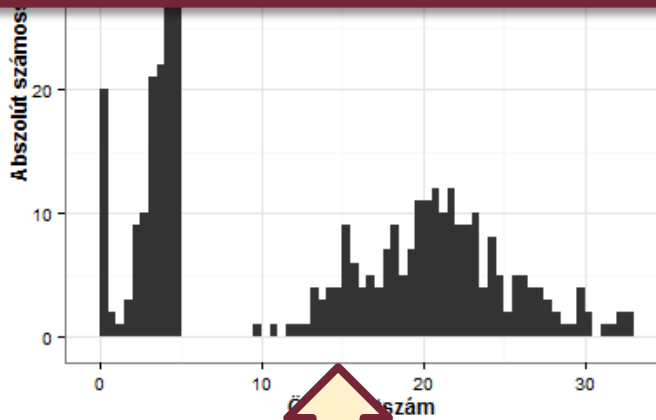


Sehr gute
Eingangstests
in G06, G11
und G17

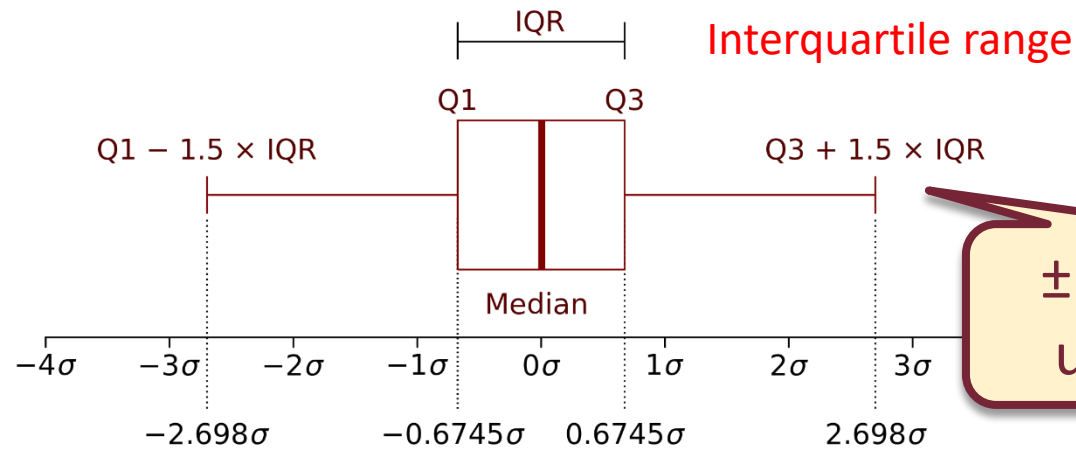
G08?

Kastengraphik (Box and whisker plot)

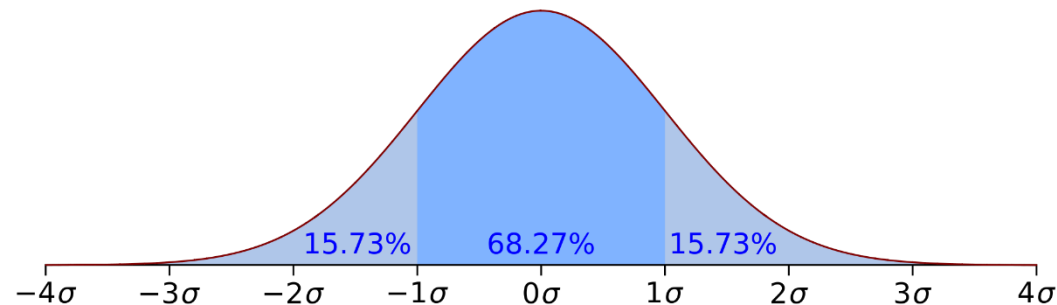
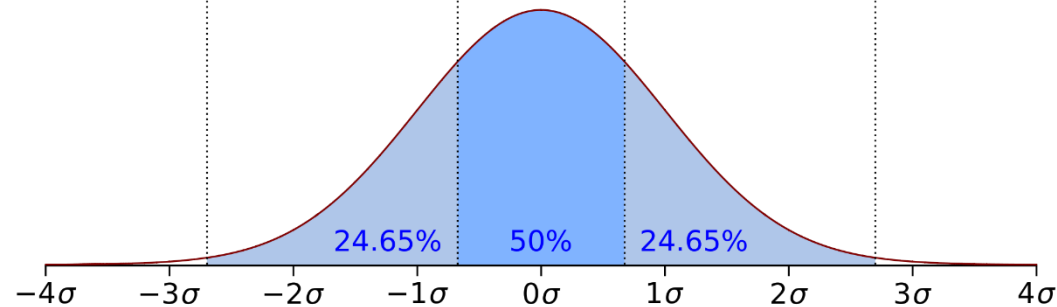
Abstraktion: Mit Kastengraphik können wir wichtige Informationen verlieren.



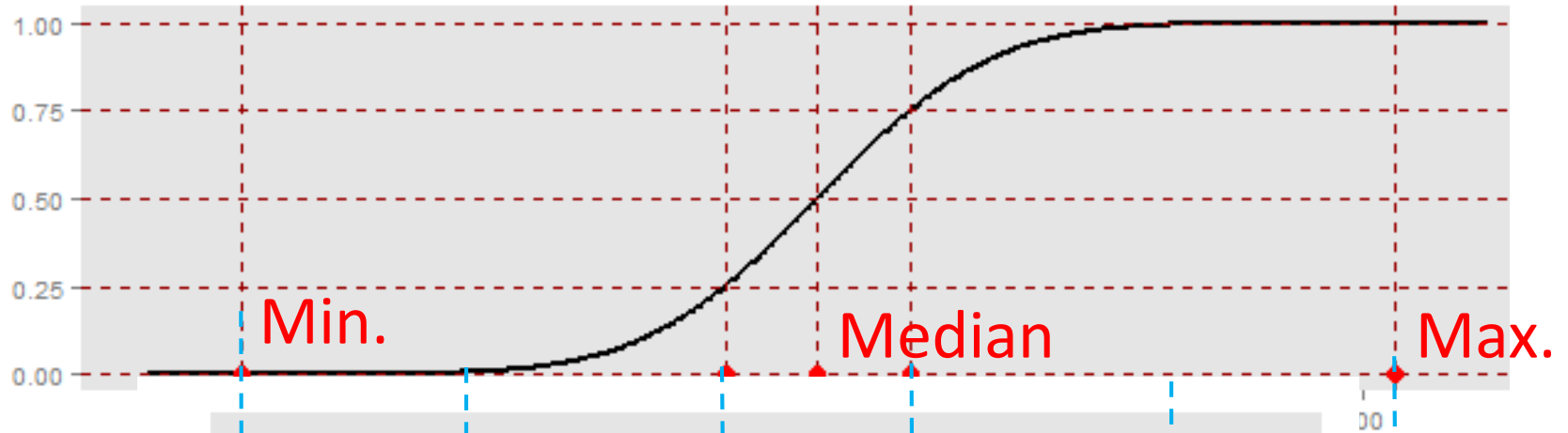
Boxplot (Box and whisker plot)



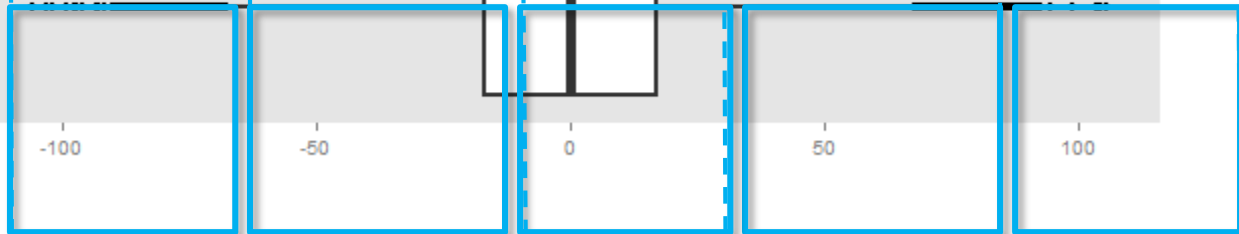
$\pm 1.5 \times IQR$ ist
ungefähr 3σ



Boxplot: Qualitative Beschreibung



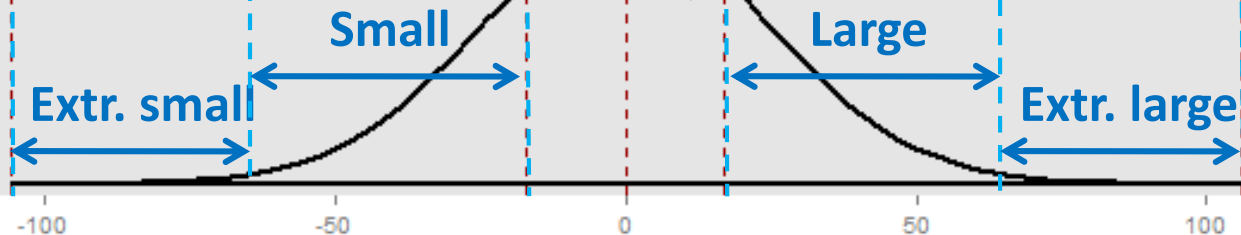
Extr. small Small Normal Large Extr. large



Q1 – 1.5 IQR

Normal

Quantitativer
Bereich



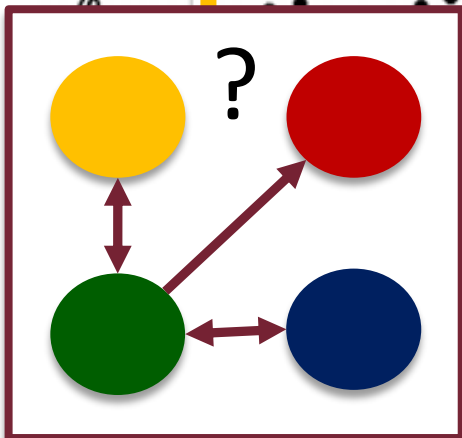
Beispiel: Last vs. Auslastung

**Ausreisserwerte/
Hintergrundoperationen?**

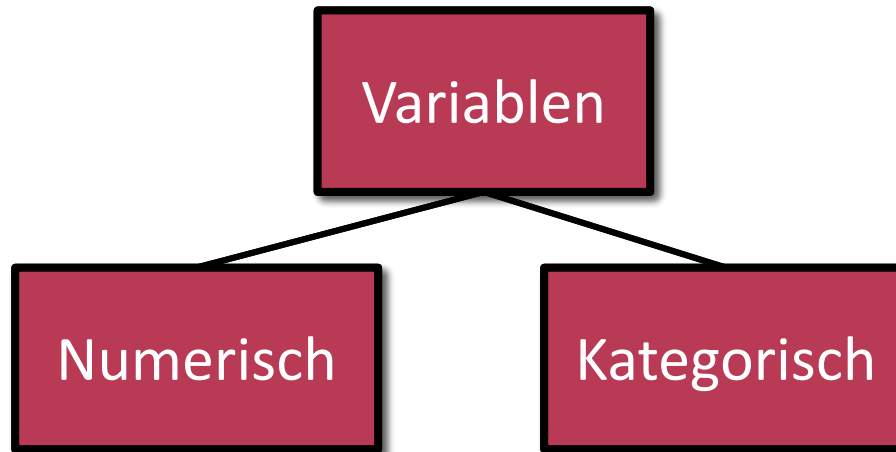
Abweichungen?

**Lineare
Abhängigkeit**

**Bei grosser Last kann
die Reaktion nicht
vorhergesagt werden**



Zusammenhänge zweier Variablen

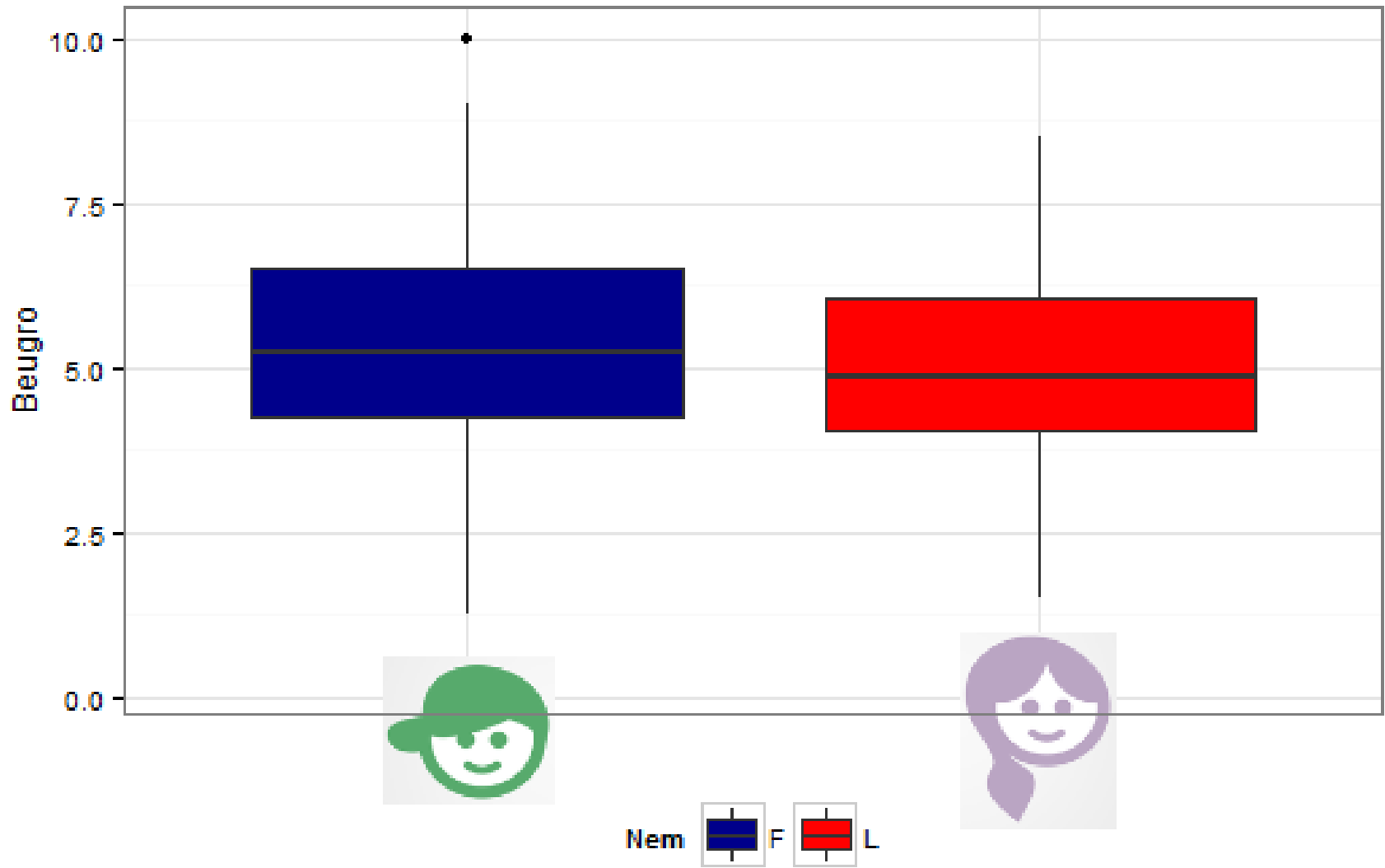


2 numerische

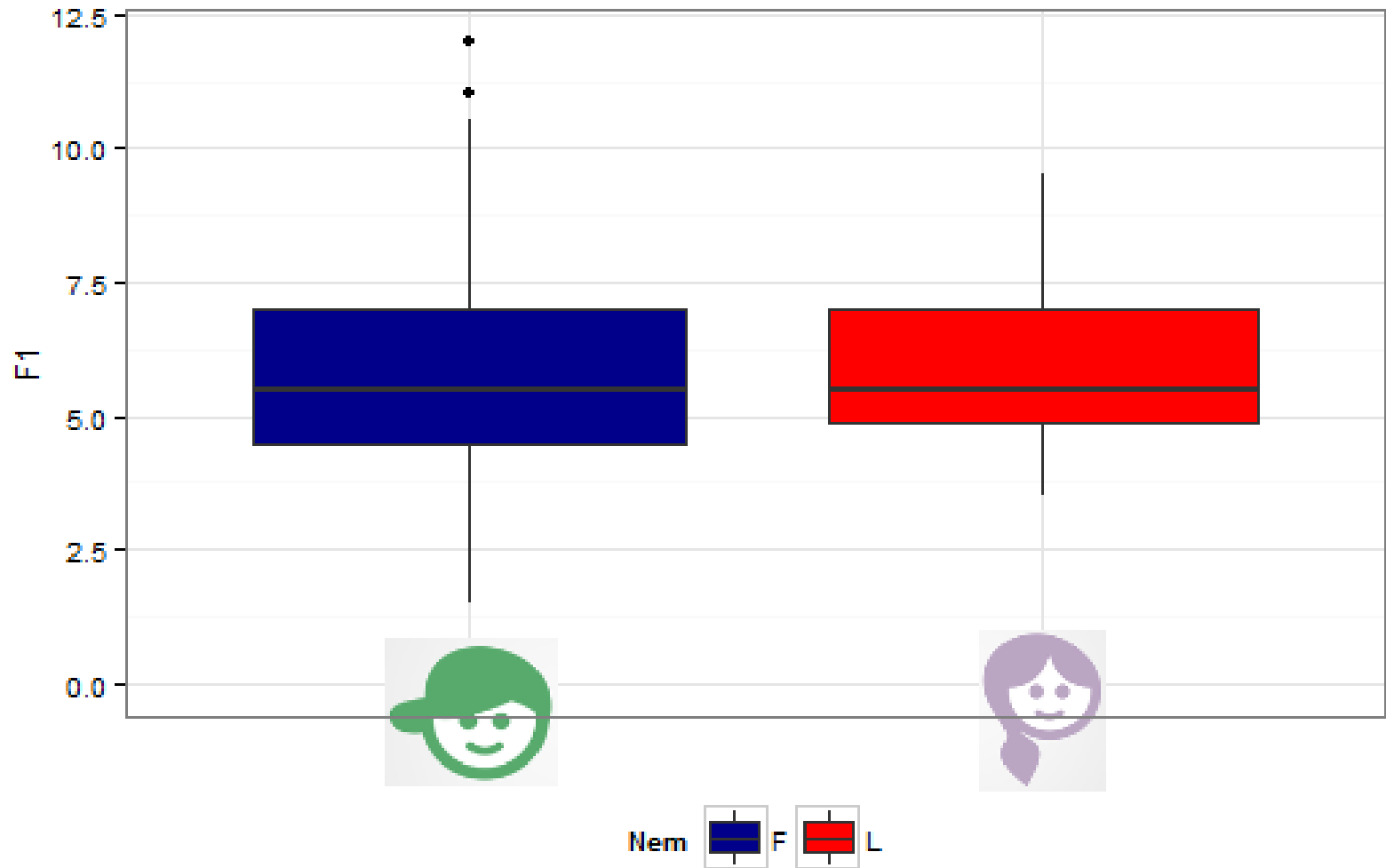
1 numerisch,
1 kategorisch

2 kategorische

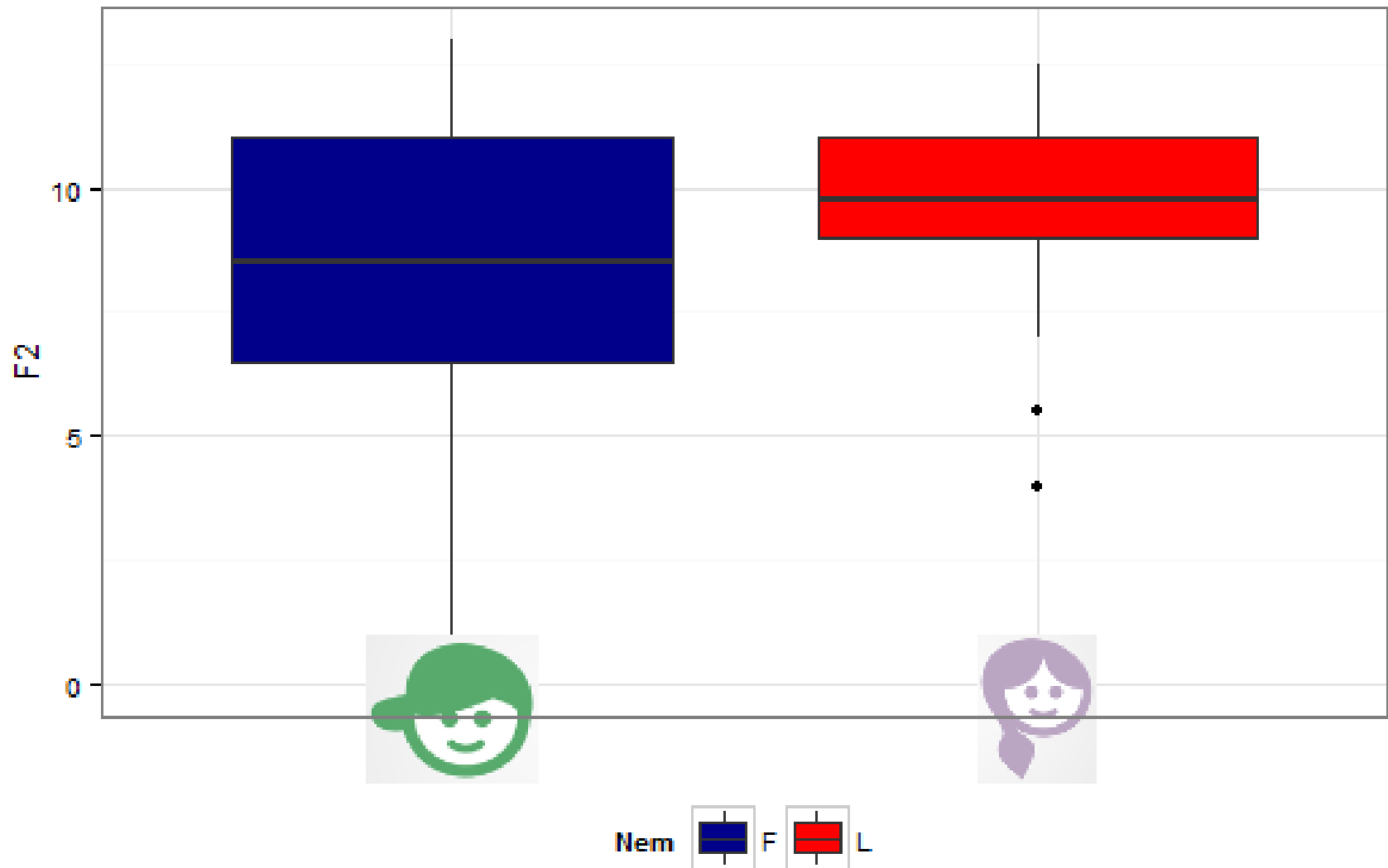
Numerisch, per Kategorie



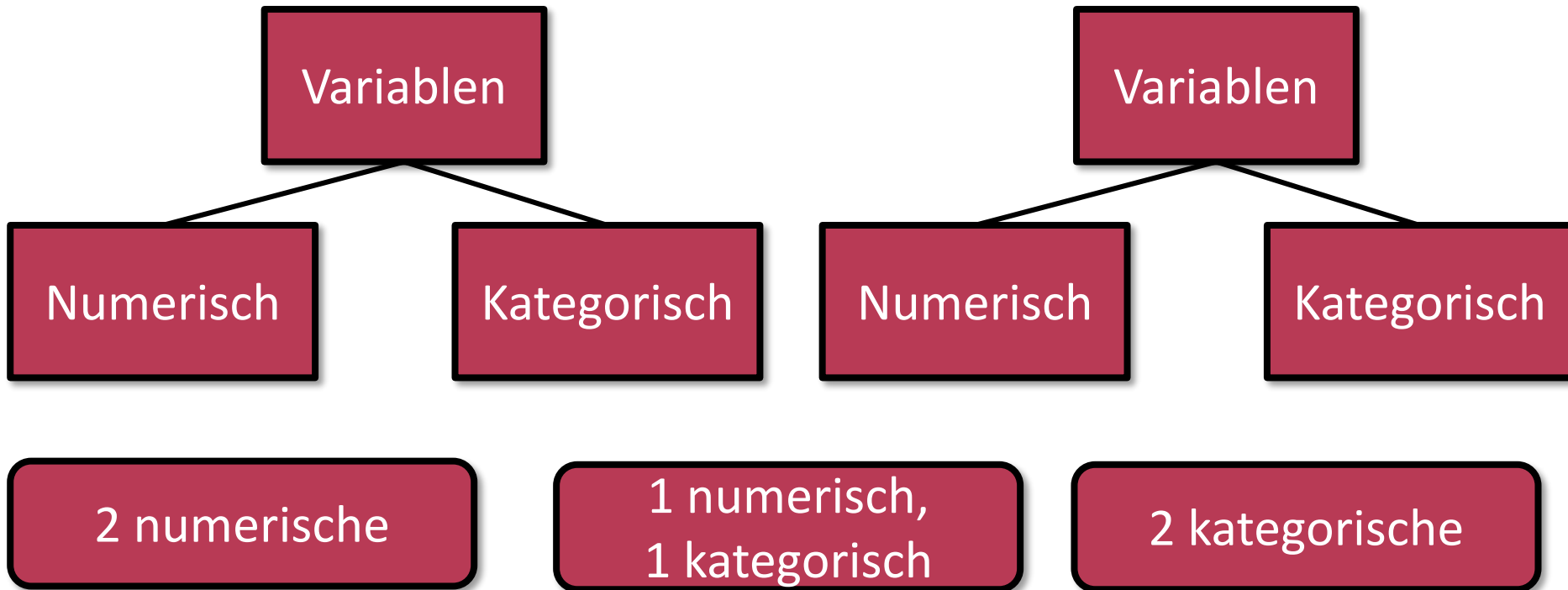
Numerisch, per Kategorie



Numerisch, per Kategorie



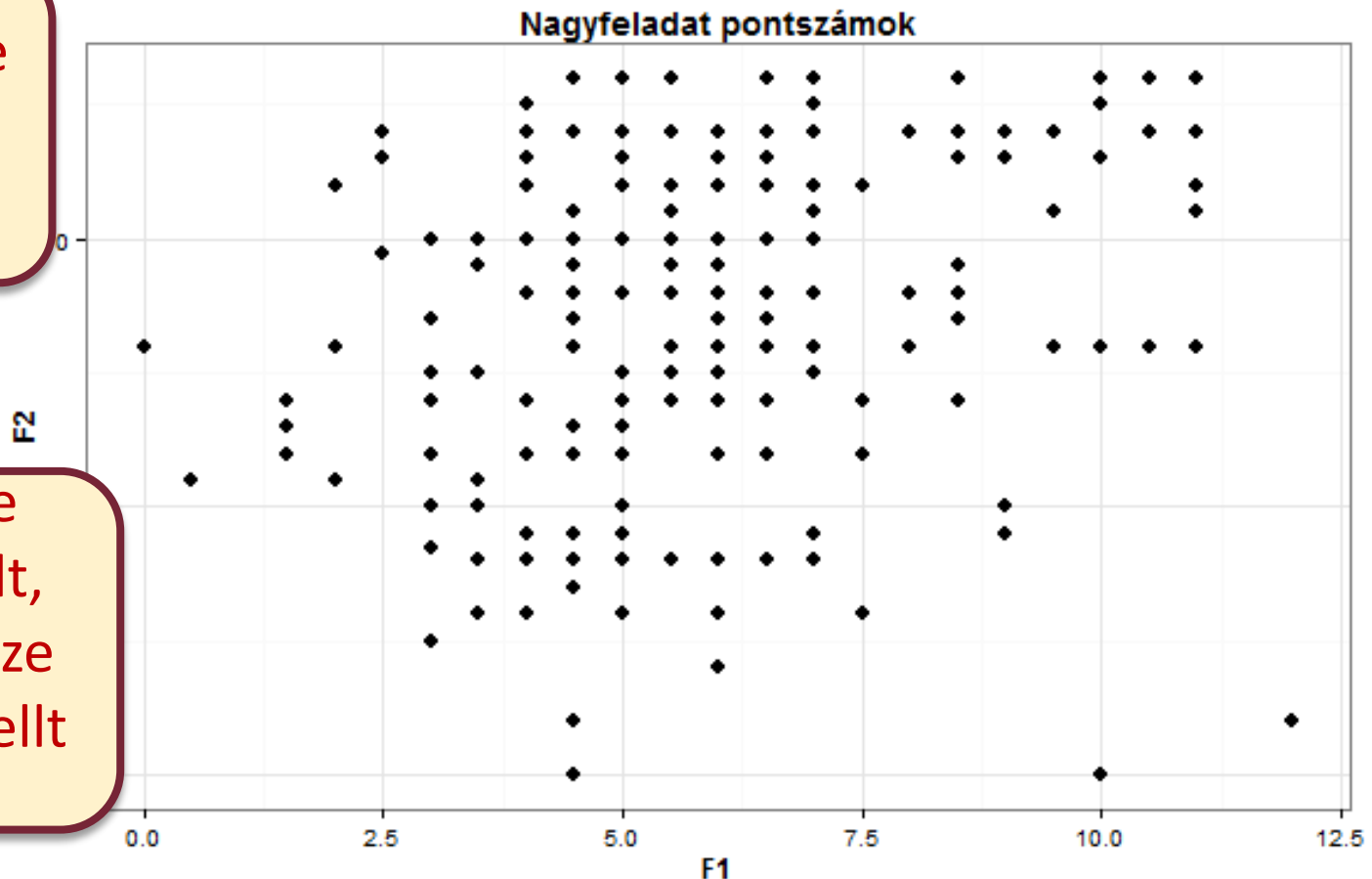
Zusammenhänge zweier Variablen



Streudiagramme (scatterplot)

- Eingabevariablen: Ergebnisse der Großaufgaben
- Frage: Was ist ihr Verhältnis?

Ergebnispaare
sind
visualisiert

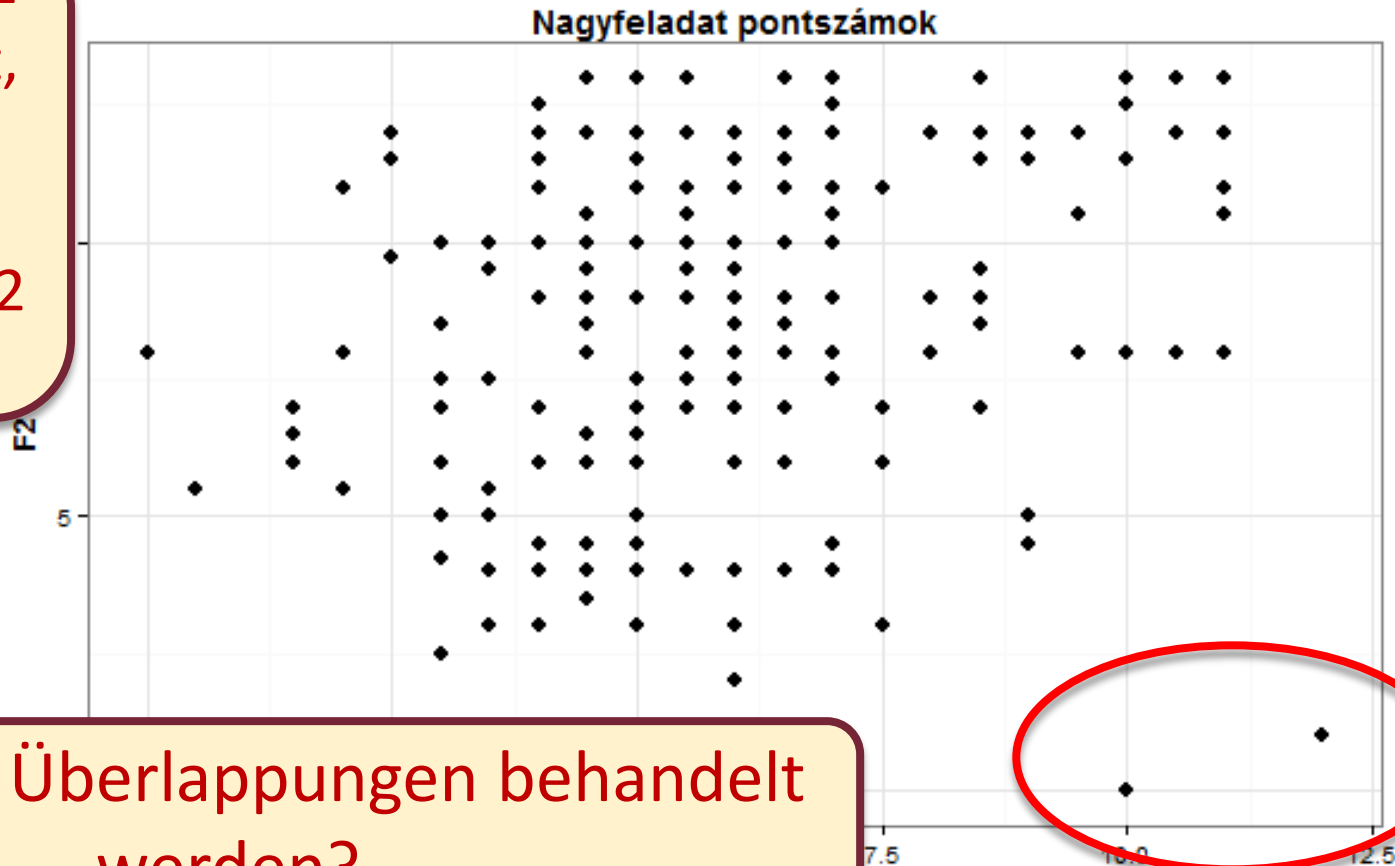


Wo das eine
Ergebnis fehlt,
kann das ganze
nicht dargestellt
werden. ☹️

Streudiagramme (scatterplot)

- Eingabevariablen: Ergebnisse der Großaufgaben
- Frage: Was ist ihr Verhältnis?

Wer Aufgabe1
gut gelöst hat,
hat nicht
unbedingt
auch Aufgabe2
gut gelöst.



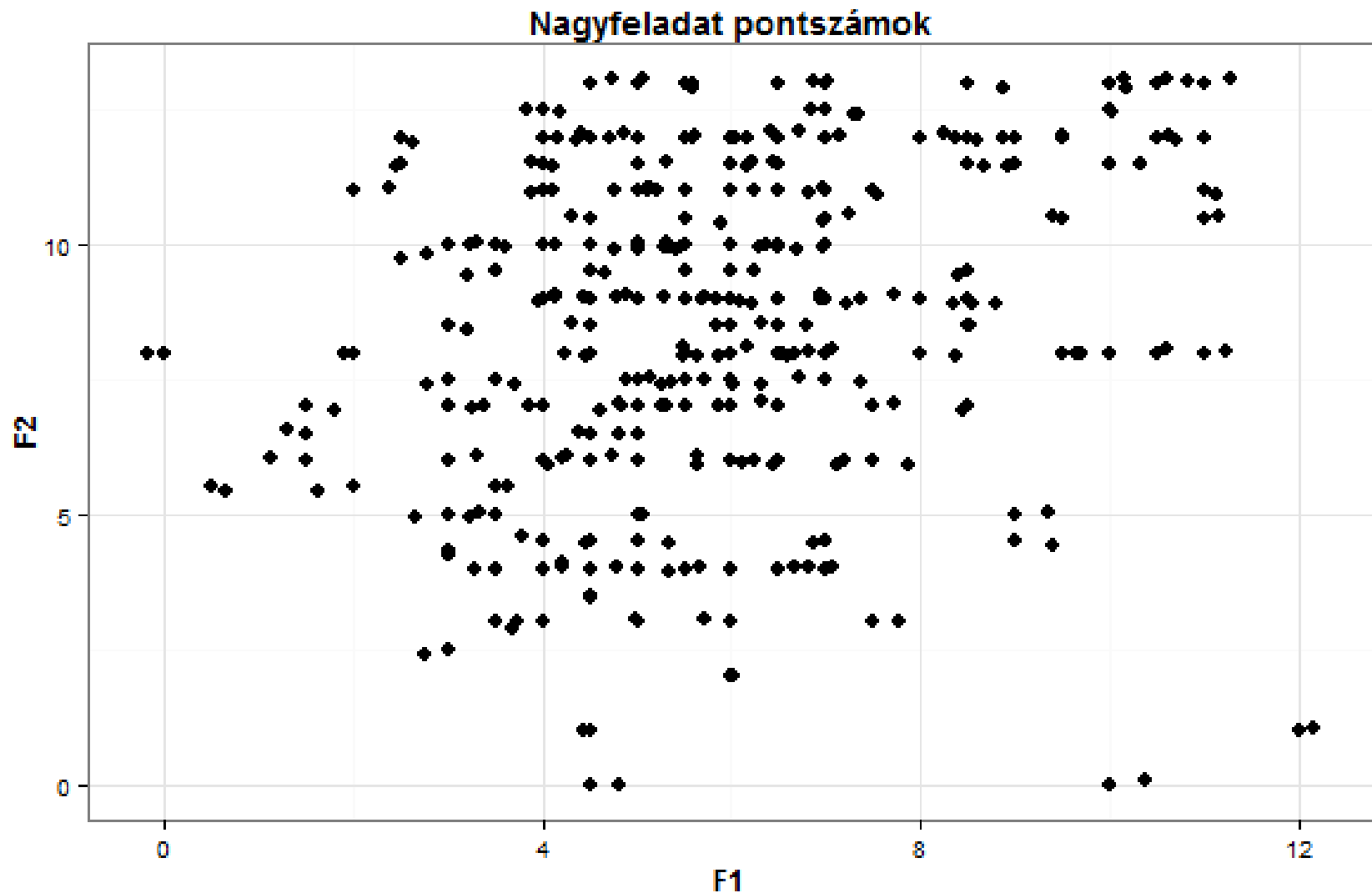
Wie sollten Überlappungen behandelt werden?

Overplotting

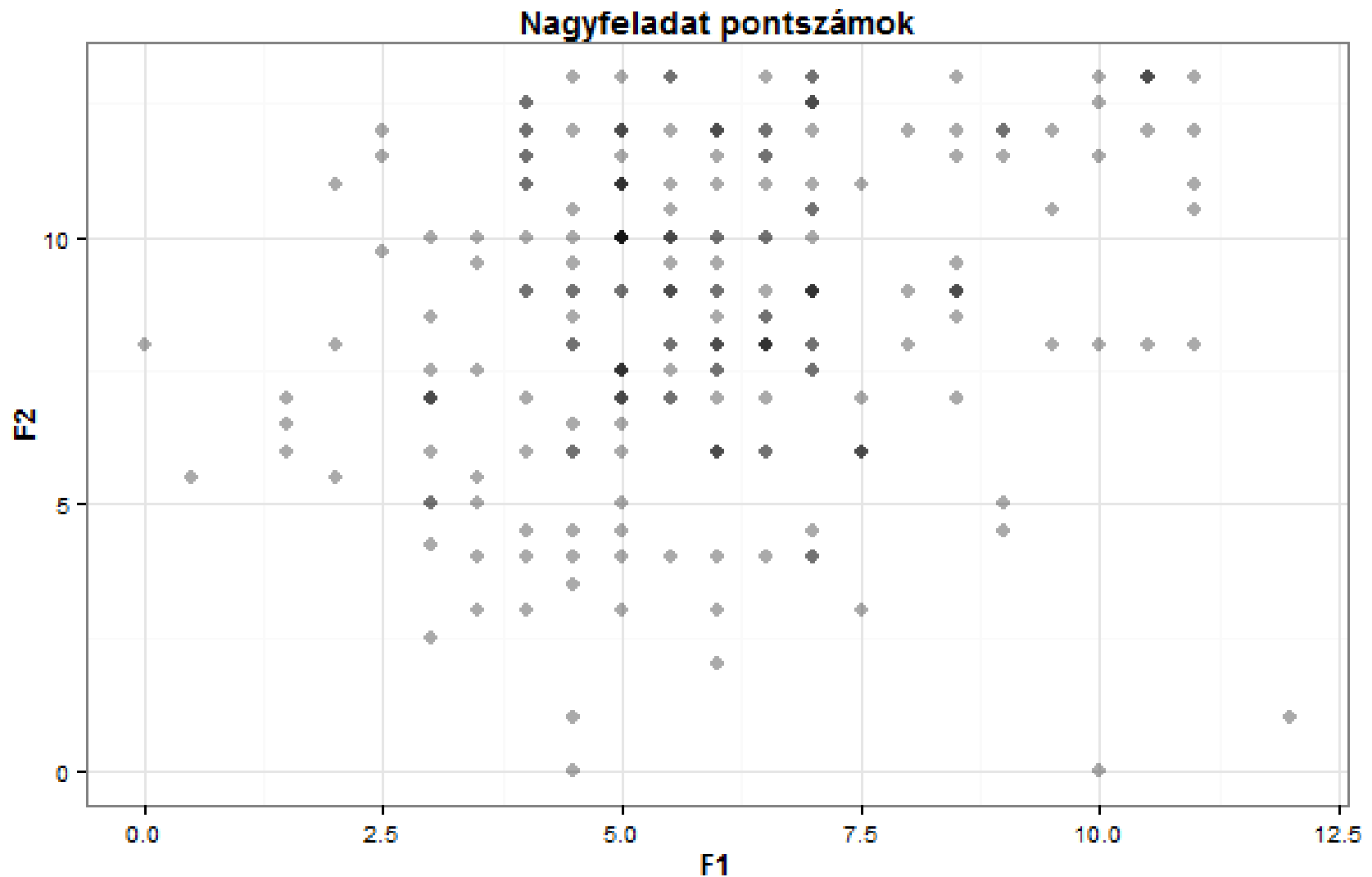


<http://www.holigyunkma.hu/>

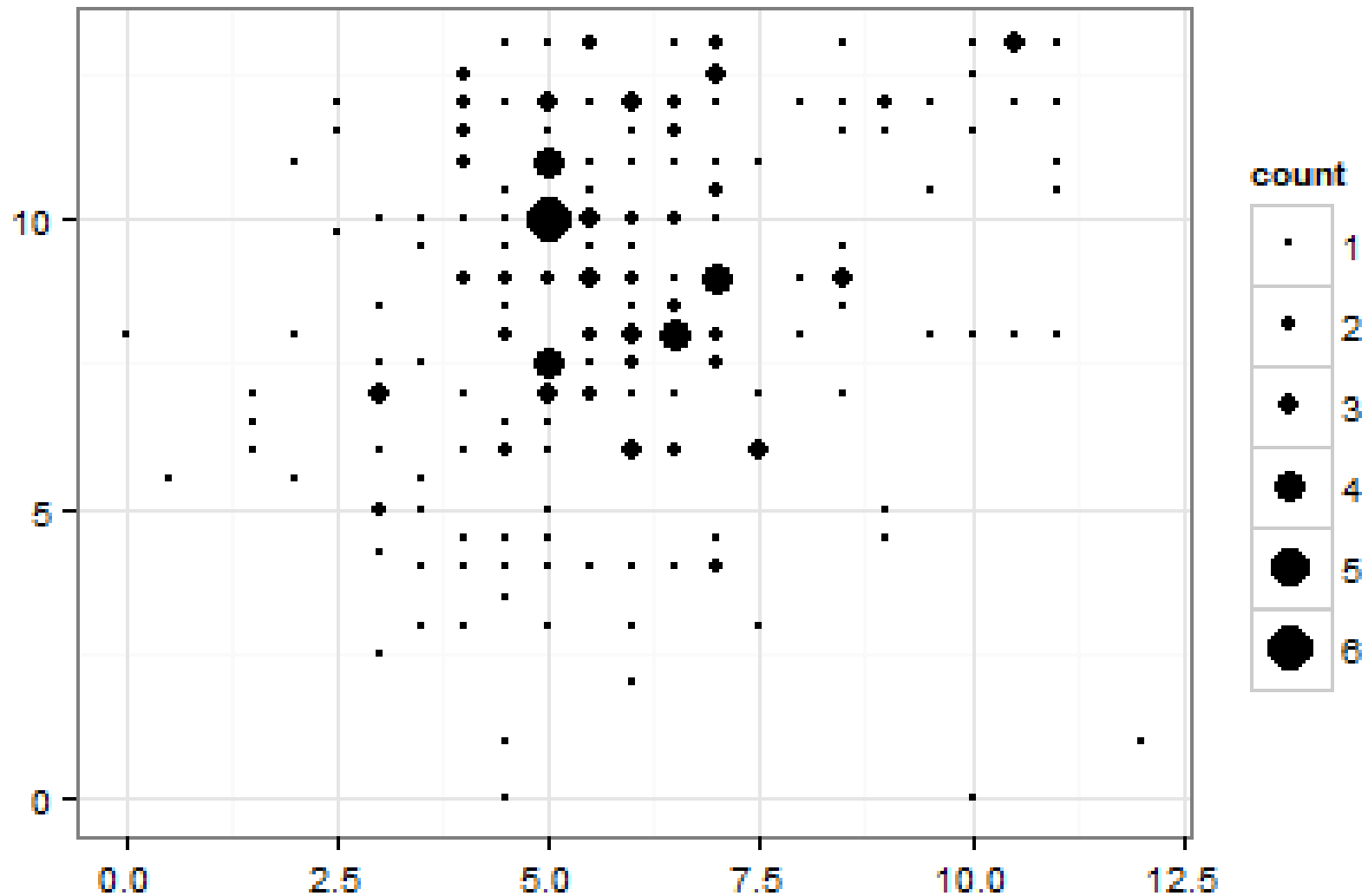
Overplotting – Lösung 1: jitter



Overplotting – Lösung 2: Durchsichtigkeit



Overplotting – Lösung 3: Grösse



Zusammenhänge zweier Variablen

Variablen

Numerisch

Kategorisch

Variablen

Numerisch

Kategorisch

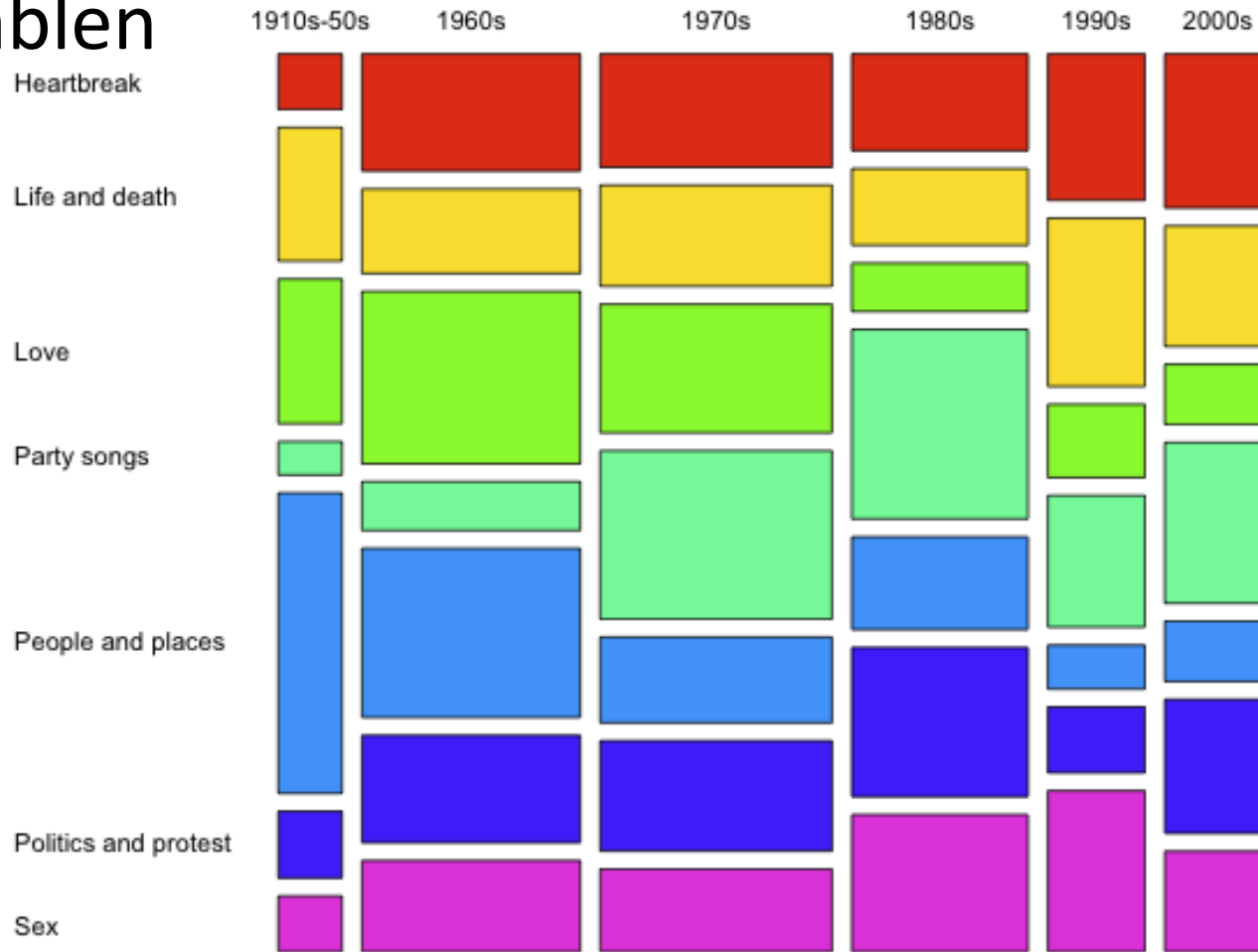
2 numerische

1 numerisch,
1 kategorisch

2 kategorische

Mosaikplot

- Zusammenhänge zweier oder mehr kategorischen Variablen



stubbornmule.net

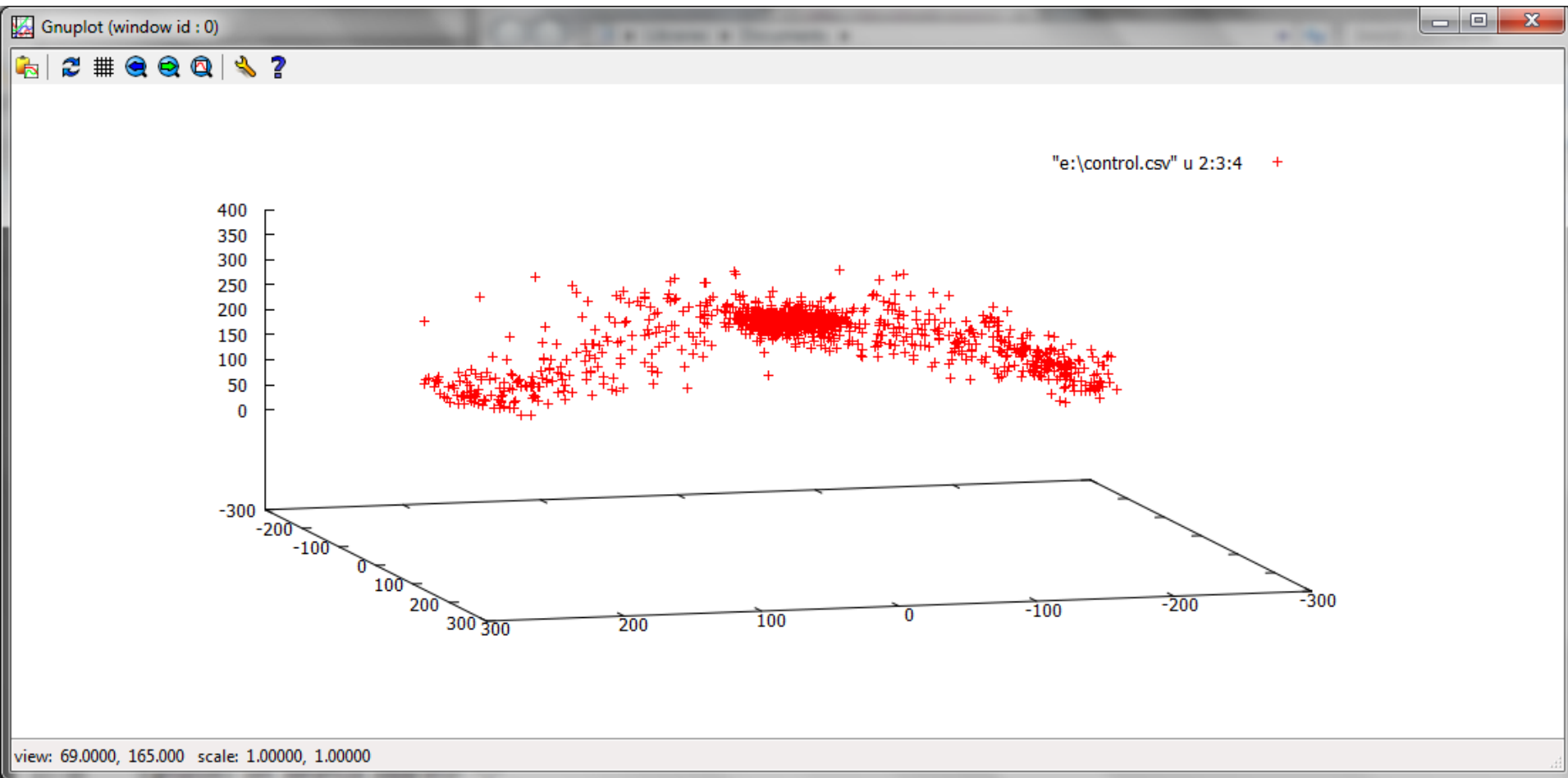
Guardian's list of "1000 songs to hear before you die"

MEHRERE VARIABLEN

≥ 3 Variablen

- Wir ändern die Attribute der graphischen Objekte
 - Farbe
 - Grösse
 - Textur
 - Stelle – es kann auch trivial sein, aber bei „treemaps“ hat Stelle eine konkrete Bedeutung
- Z.B. bubble chart, heatmap, treemap

3D Plot



Bubble chart: Durchschnittsalter nach Regionen

GAPMINDER WORLD

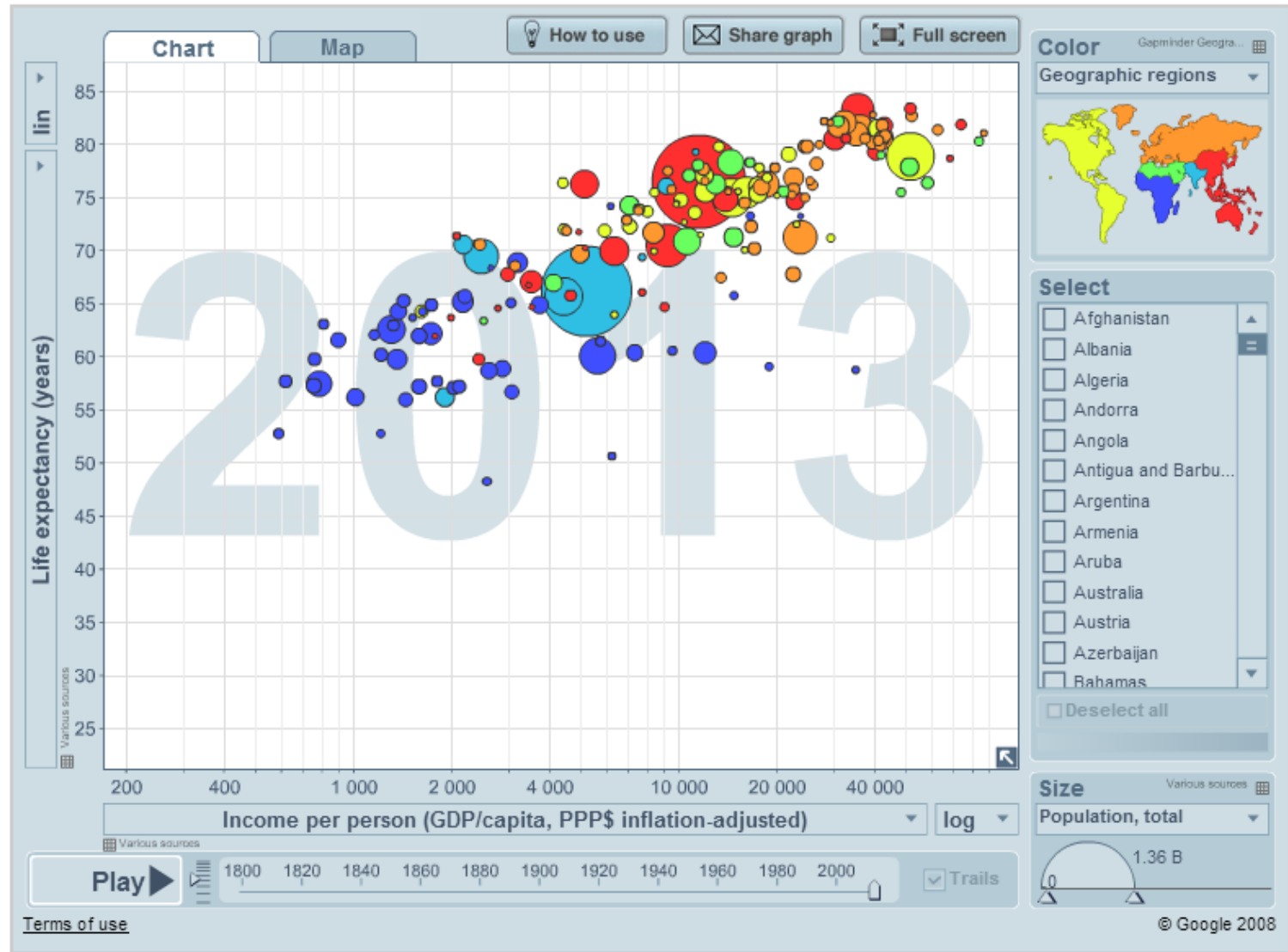
VIDEOS

DOWNLOADS

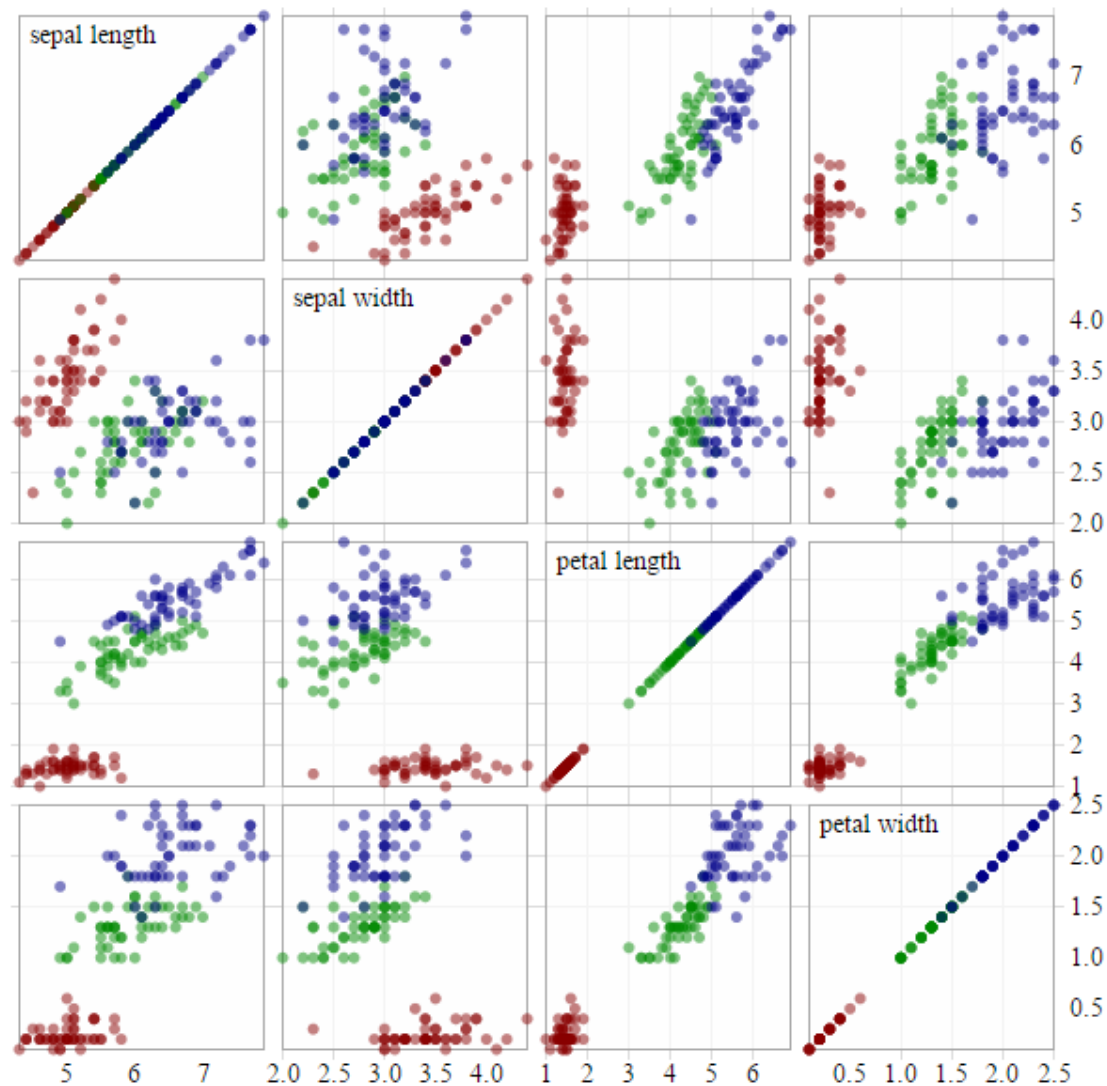
TEACH

IGNORANCE

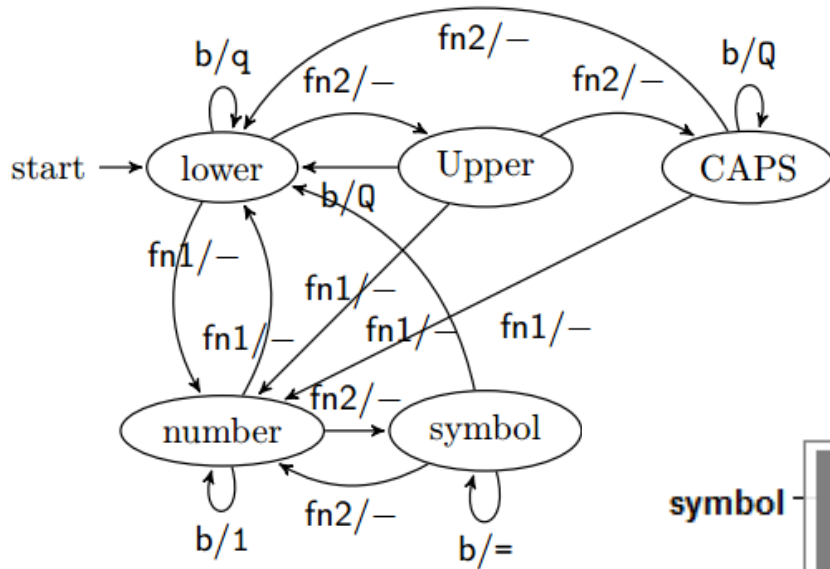
DATA



Scatterplot Matrix

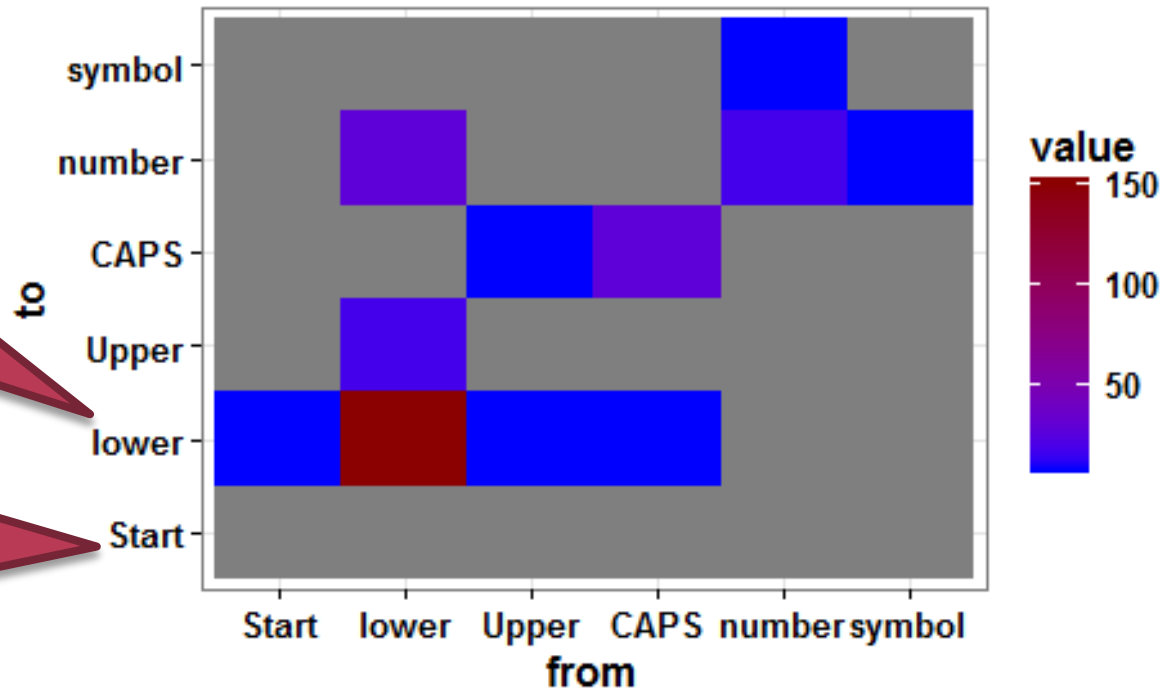


Heatmap: Ablaufstatistiken

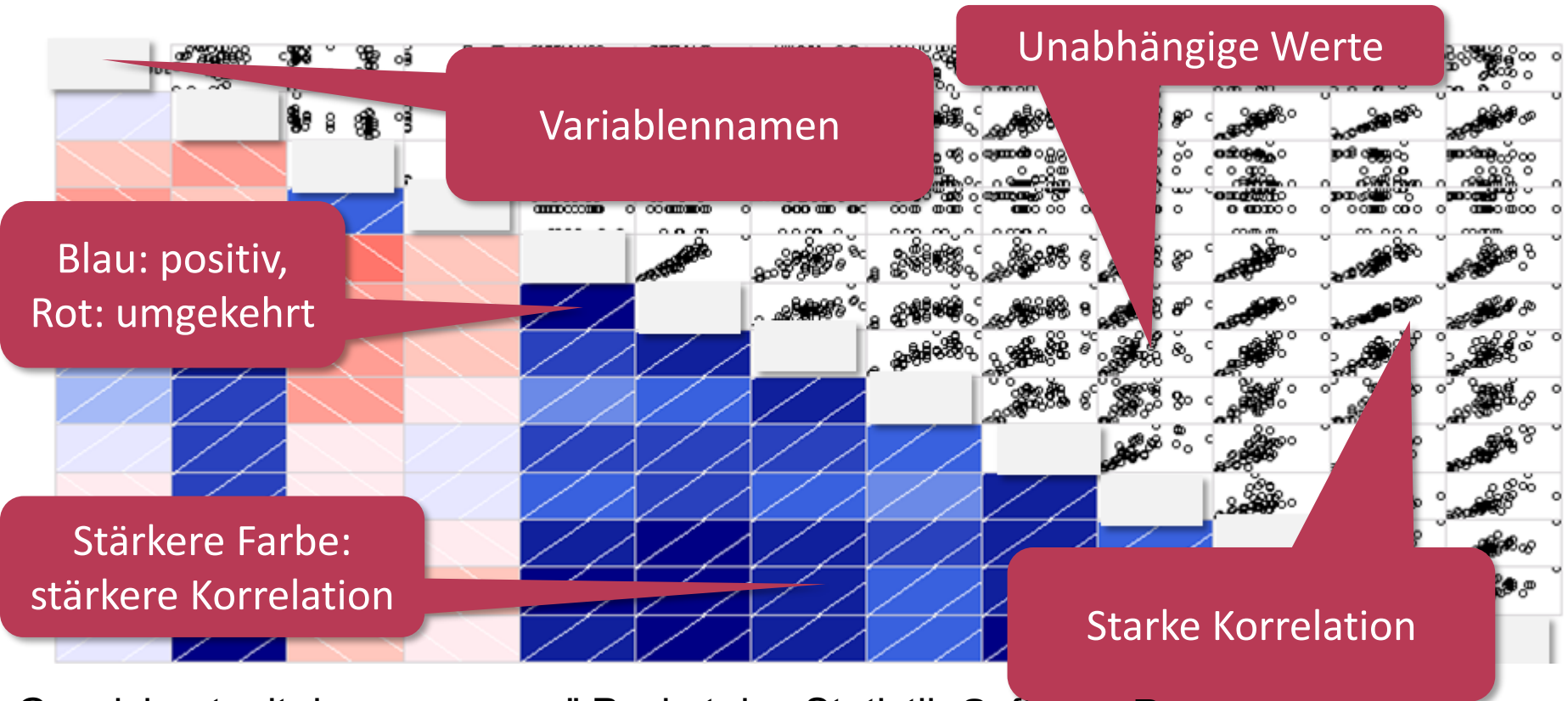


Meistens schreiben wir einfaches Text

In Start fangen wir nur an, wir kehren nie zurück



Ausblick: Paarweise Korrelation mehrerer Werte



Gezeichnet mit dem „corrgram” Packet des Statistik-Software R.

Korrelation (siehe Wahrscheinlichkeitsrechnung):

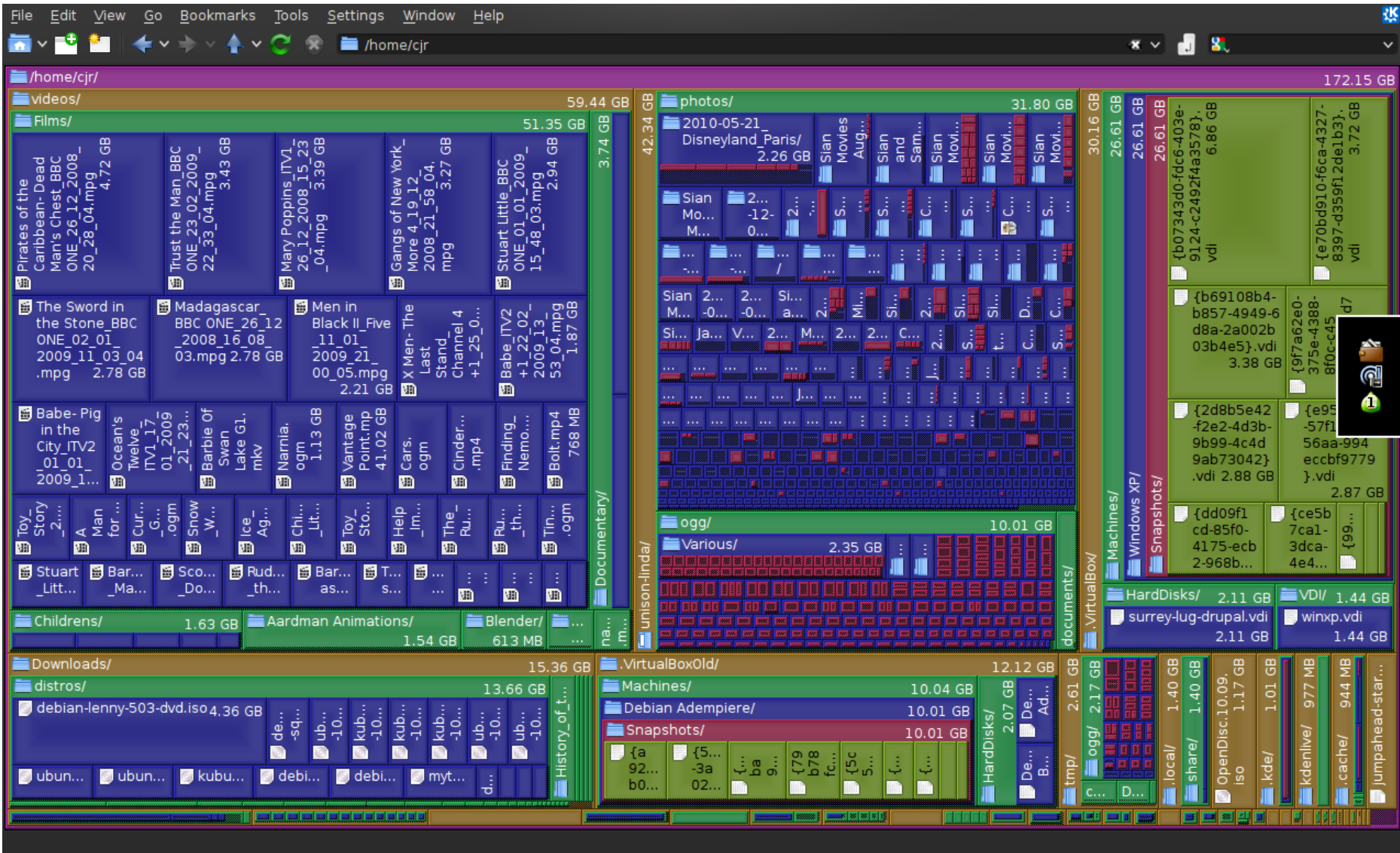
Stärke und Richtung des linearen Zusammenhanges zwischen zwei Variablen

Über der Diagonale: **scatterplot matrix**

Ziel: Herausfiltern der zusammenhängenden Variablen, Identifikation der **Ausreisser**.

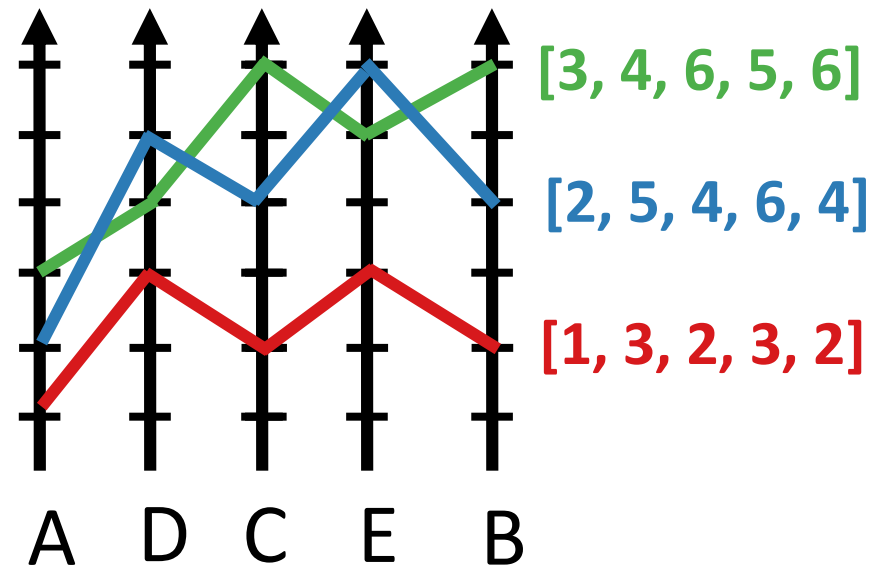
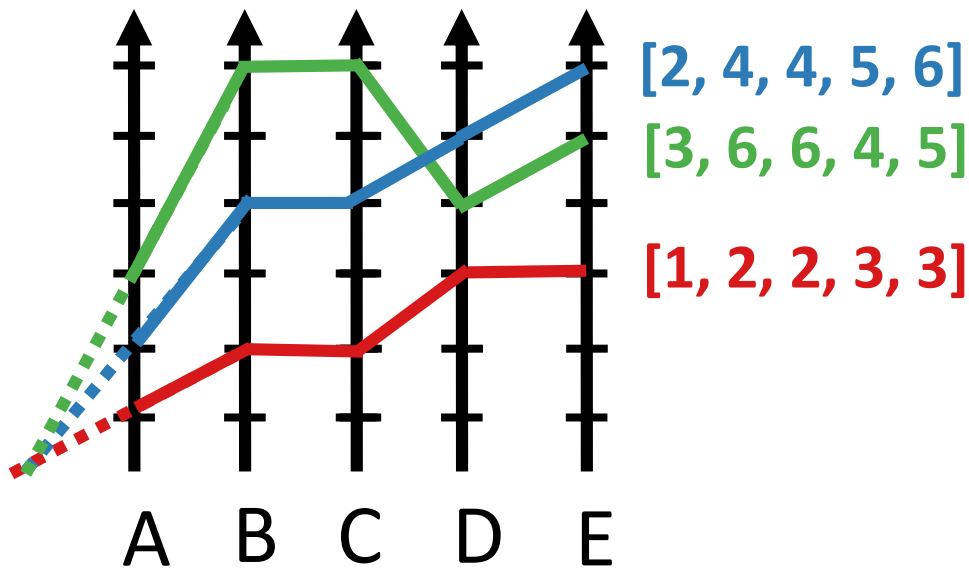
→ Welche Variablen sind für eine Hervorsage der Belastung wichtig?

Treemap: Dateisystem



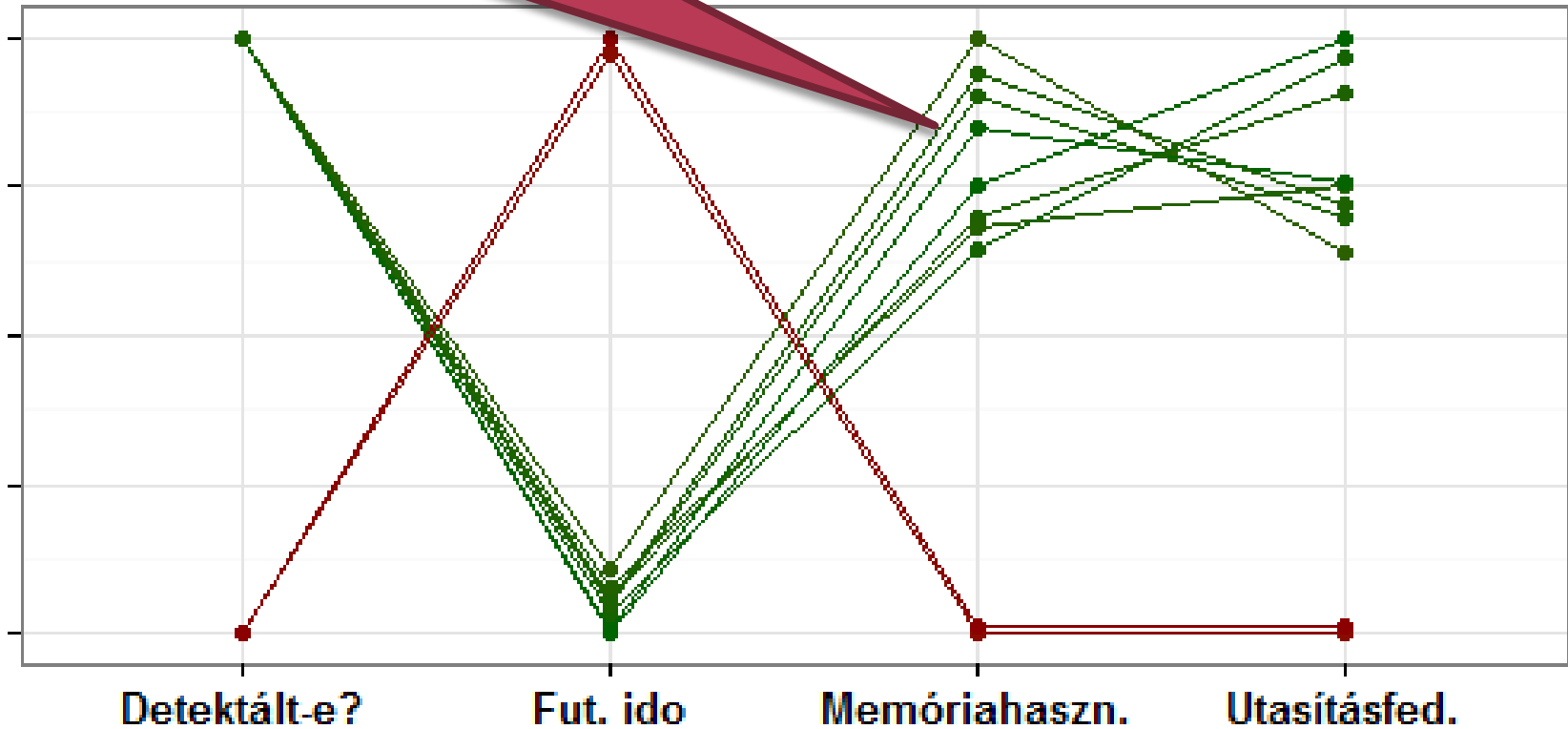
Parallele Koordinaten

- Multi-dimensionale Visualisation
- Kompakt, skalierbar
- Reihenfolge der Achsen?



Parallele Koordinaten: Analyse der Testfälle

1 Testfall: 1 gebrochene Linie

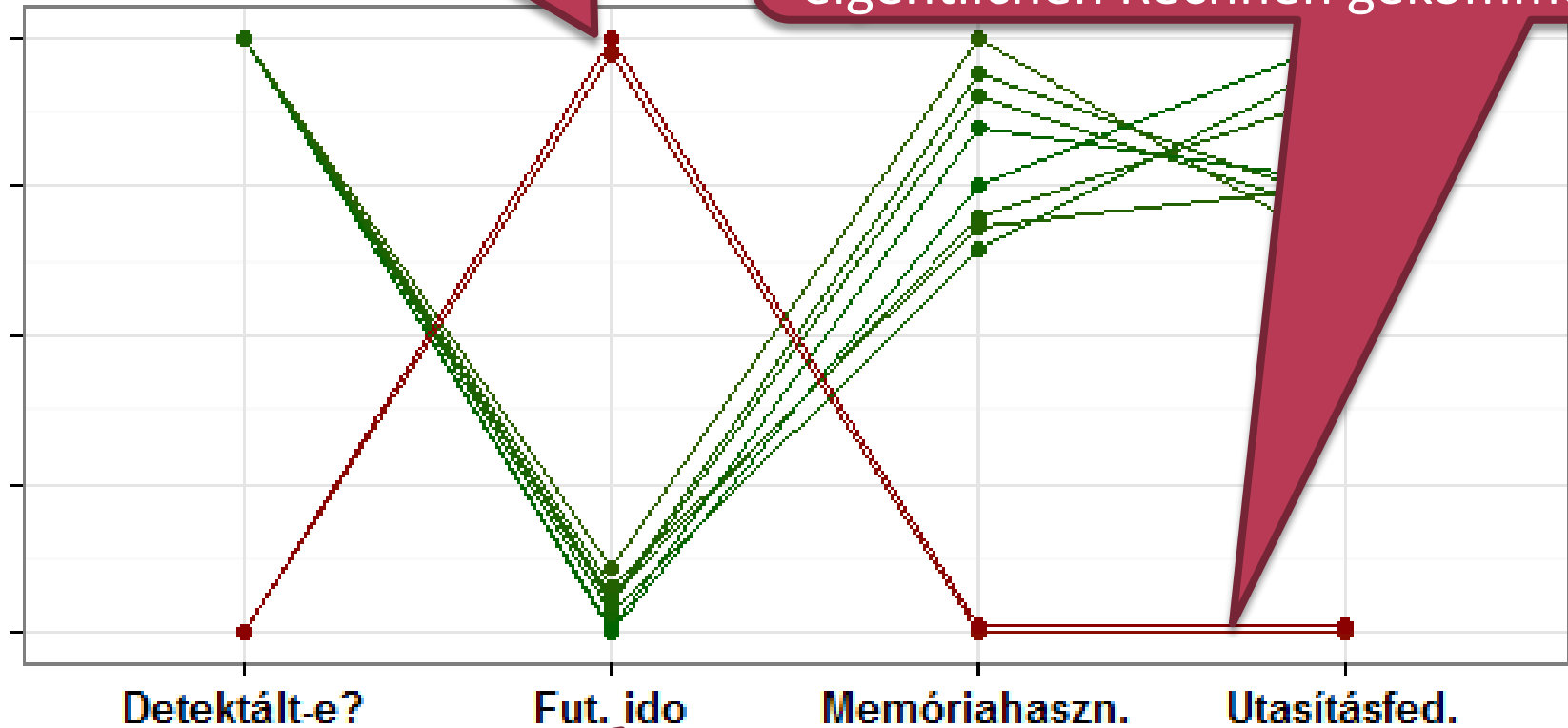


Die Variablen erscheinen
an der x -Achse

Parallele Koordinaten: Analyse der Testfälle

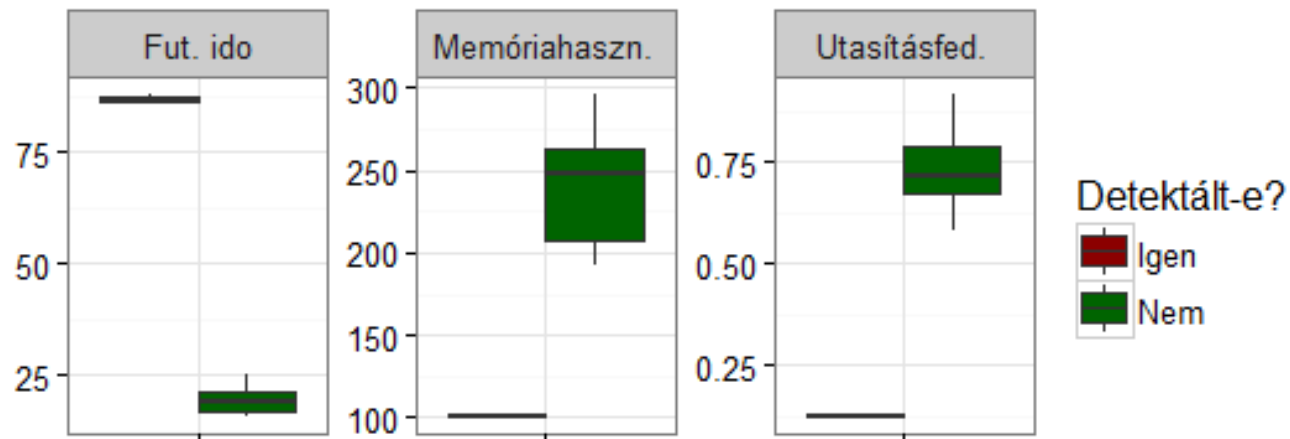
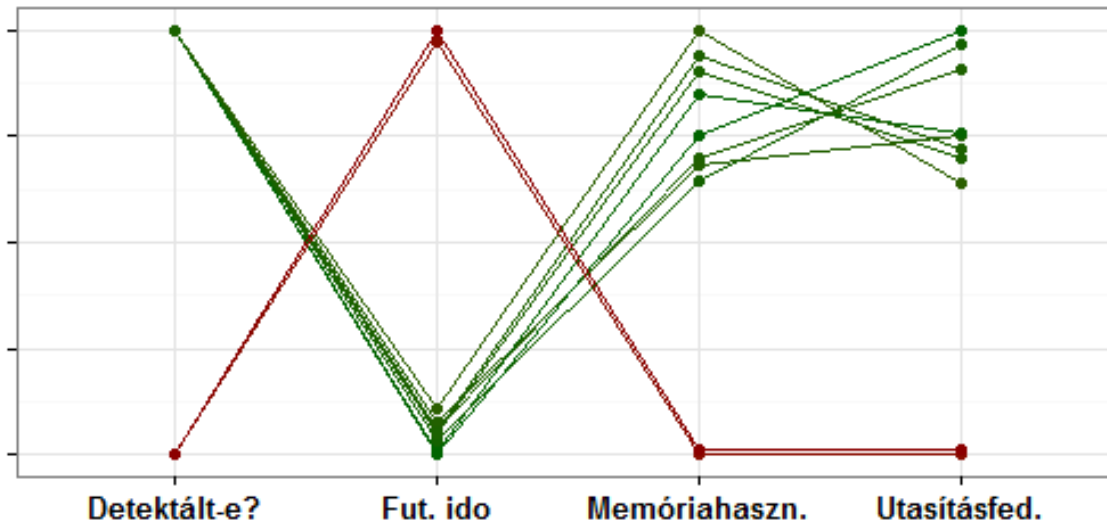
Timeout?

Die, die einen Fehler detektiert haben, sind gar nicht zum eigentlichen Rechnen gekommen.



Laufzeit und Speicherverbrauch scheinen in positivem Zusammenhang zu stehen (bei erfolgreichem Test)

Parallela Koordinaten: die Alternativen



Radar chart: Erweiterung der parallelen Koord.

