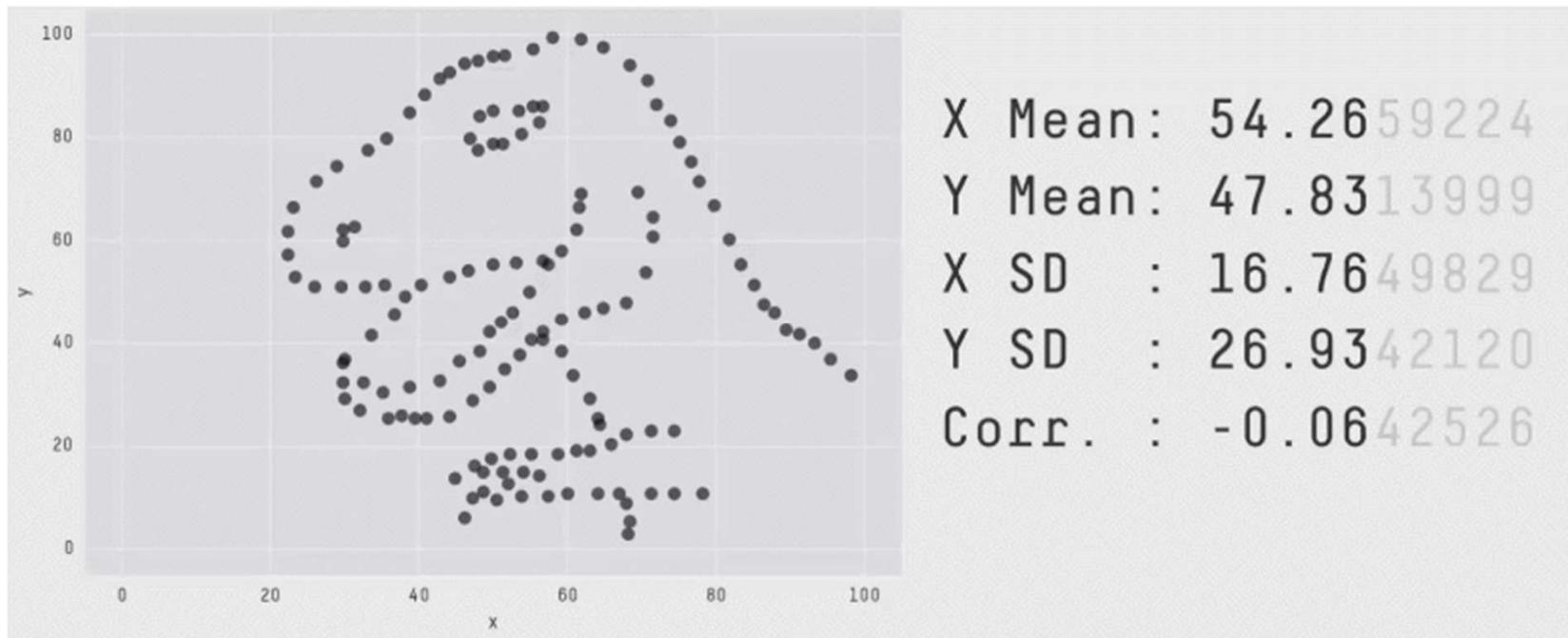

Seminarbaustein: Data Science mit

R ist das am häufigsten genutzte Werkzeug zur statistischen Datenanalyse und zur grafischen Darstellung der Ergebnisse. Es ist kostenfrei, flexibel einsetzbar und wird von einer großen Community ständig weiter entwickelt. Im Seminar erlernen Sie den Umgang damit und erhalten hilfreiche Unterstützung zur Anwendung.

Hier eine kurze Einführung.

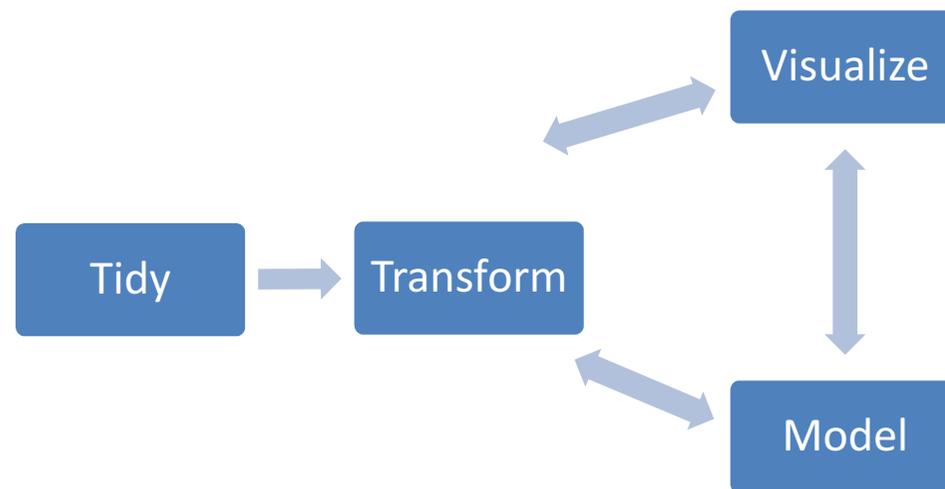
Dozent: Dr. Andreas Busjahn

Was versteckt sich in den Daten? (?Dinosaurier?)



- Skriptsprache für statistische Analysen und Grafik
- Multiversum universeller und hochspezialisierter Lösungen
- Freie Software mit 13.741 Erweiterungspaketen
- Unterstützt durch IBM, Microsoft, Google, Oracle...
- Bioconductor: Teilprojekt für Bioinformatik, ...omics
- Grammar of Graphics: Flexibles Grafiksystem

- Import (1 bis 50 Zeilen)
- Datenvorbereitung (50 bis 1000 Zeilen)
- Visualisierungen (10 bis 100 Zeilen)
- Tests / Modelle (ca. 100 Zeilen, Tests: 1 bis 5)



Graphische Benutzeroberfläche RStudio

5

The screenshot displays the RStudio environment with the following components:

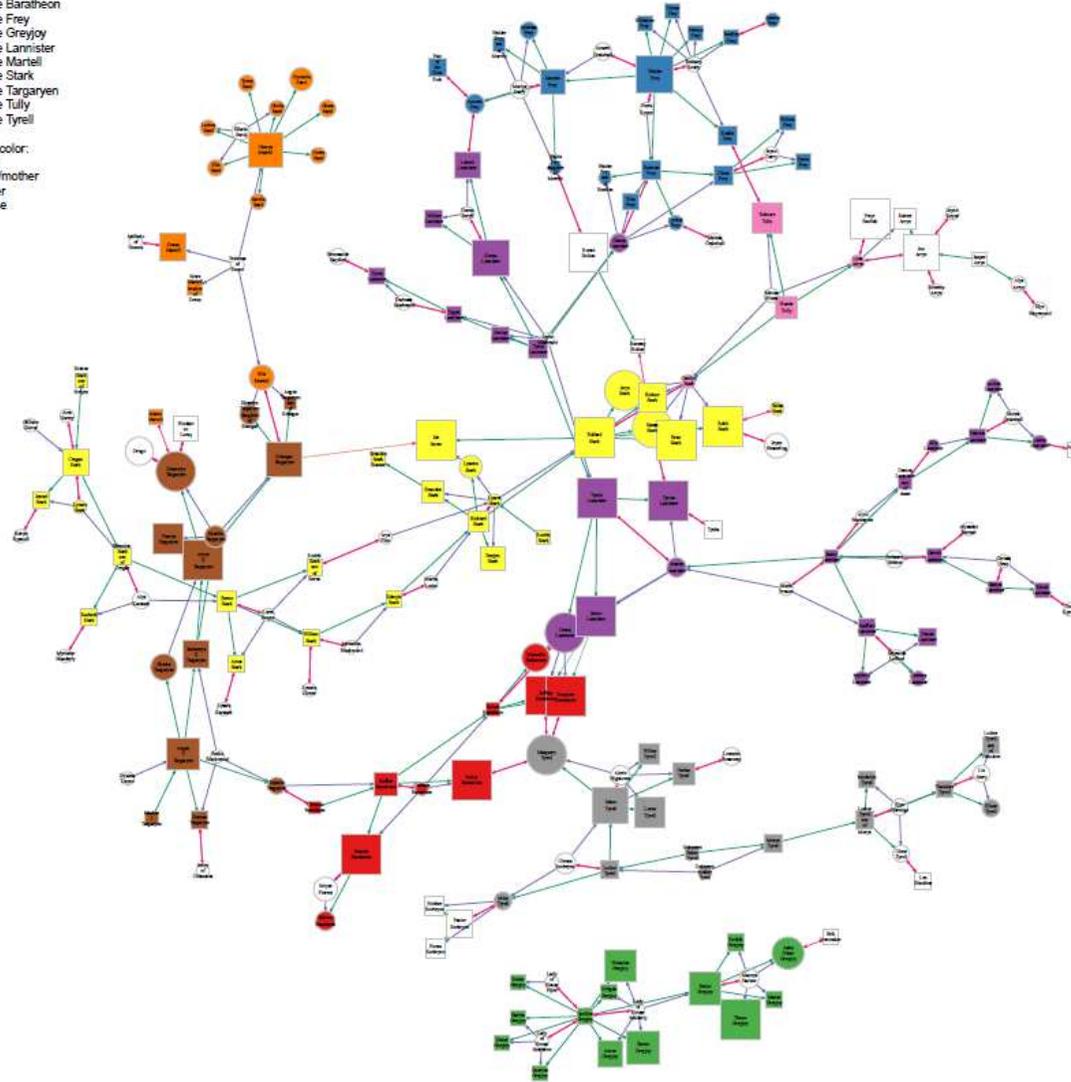
- Source Editor:** Contains R code for creating a scatter plot with a grid. The code uses `ggplot2` to plot red points of varying sizes on a grid. The grid lines are black, with major lines every 4 units and minor lines every 1 unit. The plot area is divided into a 3x3 grid of quadrants.
- Console:** Shows the execution of the R code, resulting in a scatter plot.
- Environment:** Lists the objects in the global environment, including `Kammertiefe_m...0.1`, `Konzentration...1e+06`, `Volumen_Gross...1e-04`, `Volumen_Gross...0.1`, `Volumen_Probe...10`, and `Zellen_in_Kam...900`.
- Plots Panel:** Displays the resulting scatter plot. The x-axis is labeled 'x' and the y-axis is labeled 'y', both ranging from 0 to 12. The plot shows a dense distribution of red points of varying sizes, overlaid on a black grid.

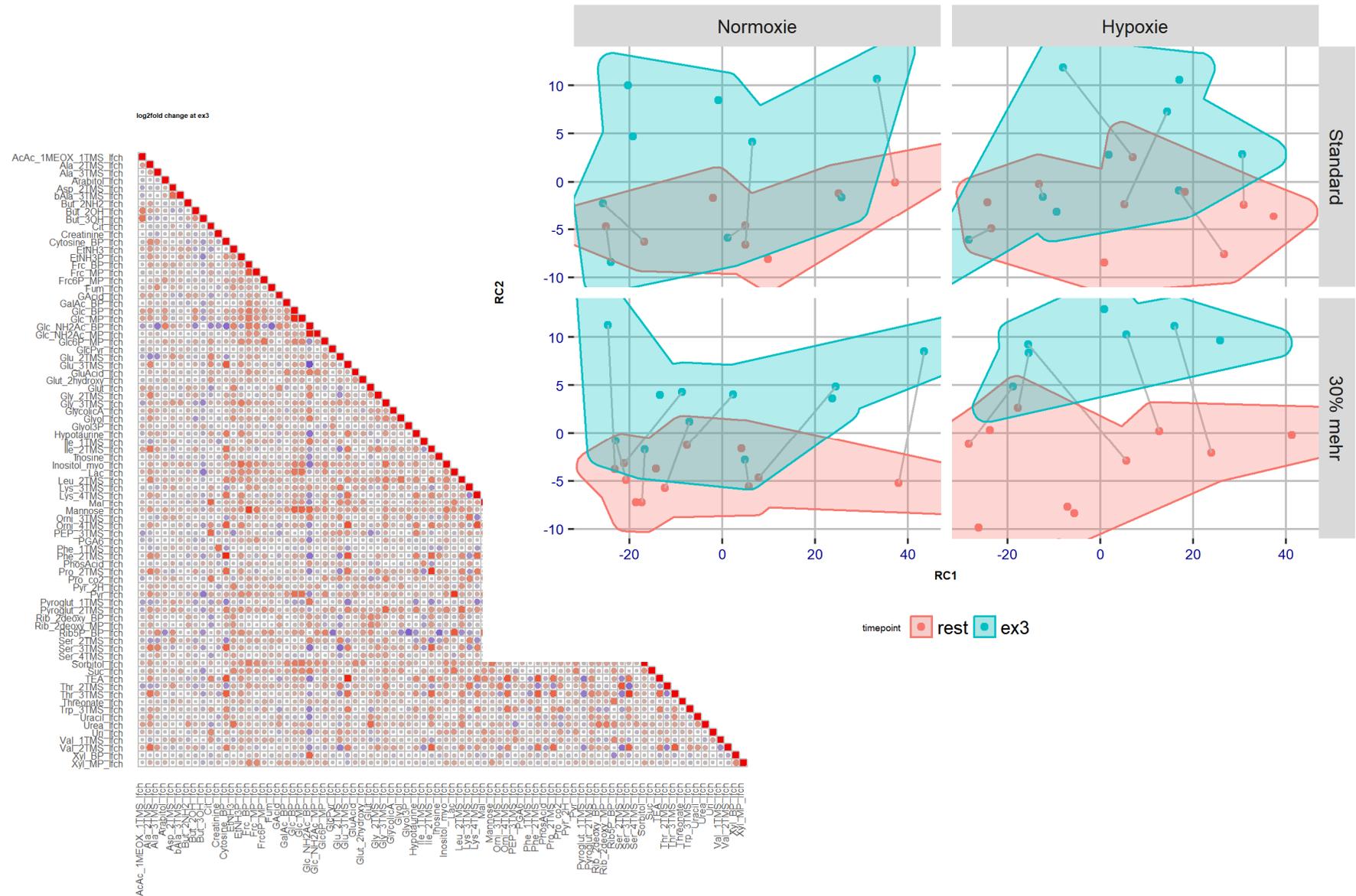
```
18 ggplot(Zelldaten, aes(x=x, y=y, size=size))+
19   geom_point(color='brown', alpha=.7)+
20   scale_x_continuous(breaks=c(0, 4, 8, 12),
21                     minor_breaks=c(0:3, seq(4
22 , 8, .25), 9:12))+
23   scale_y_continuous(breaks=c(0, 4, 8, 12),
24                     minor_breaks=c(0:3, seq(4
25 , 8, .25), 9:12))+
26   scale_size(range=c(1, 4), guide=F)+
27   theme(panel.grid.major=element_line(color
28 = 'black',
29                                               size=1.1)
30         ,
31         panel.grid.minor=element_line(color
32 = 'black', size=.8))
33
34 Zelldaten$Quadrant_x <- floor(Zelldaten$x)/4 + 1
35 Zelldaten$Quadrant_y <- floor(Zelldaten$y)/4 + 1
36 Zelldaten$Grossquadrant_x <- floor((Zelldaten$x
37 - Zelldaten$x%4)/4) + 1
38 Zelldaten$Grossquadrant_y <- floor((Zelldaten$y
```

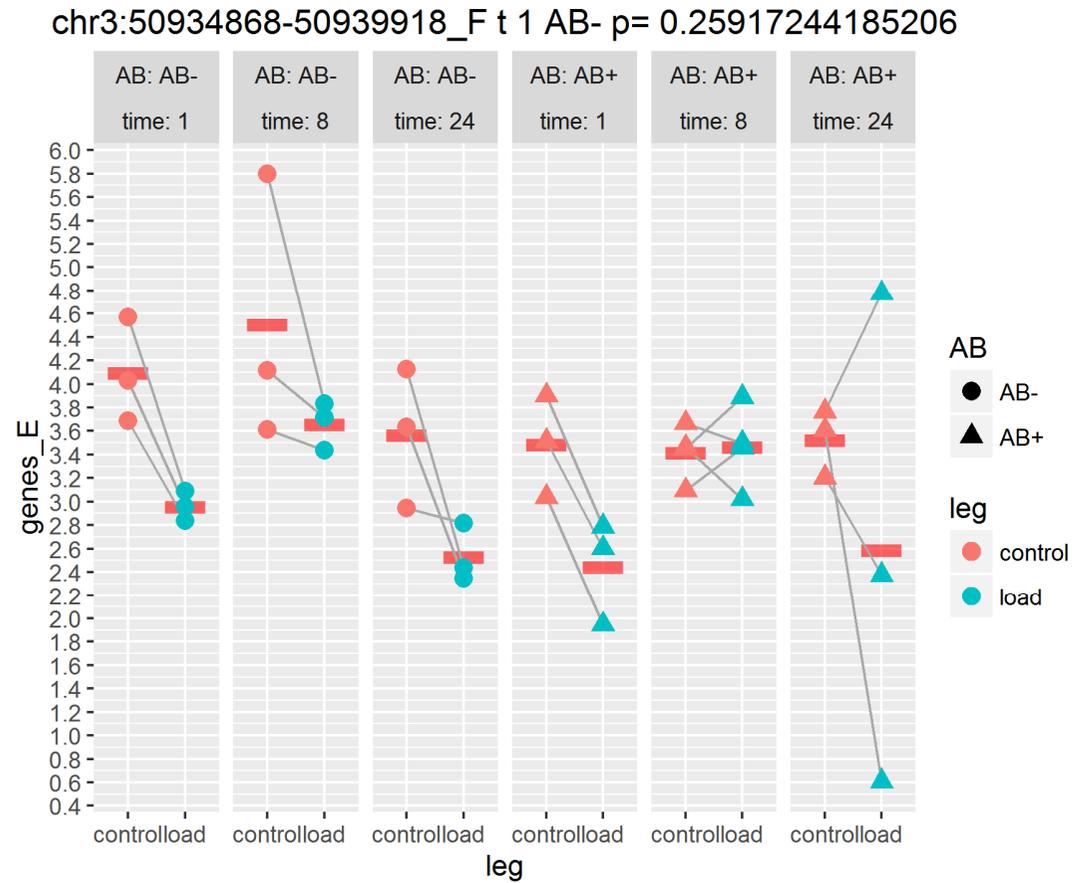
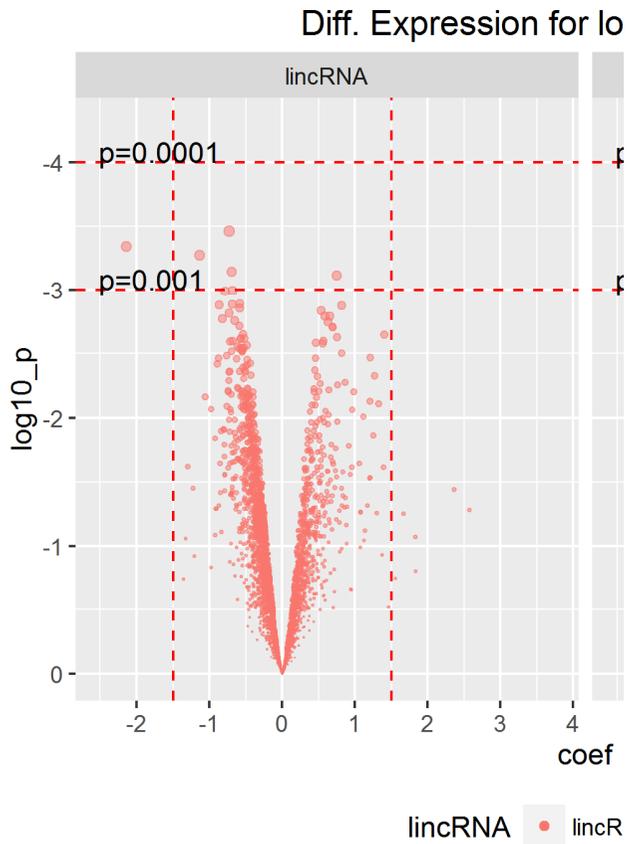
... Verlangen komplexe Analysen

Game of Thrones Family Ties

- Node color:
 - House Baratheon
 - House Frey
 - House Greyjoy
 - House Lannister
 - House Martell
 - House Stark
 - House Targaryen
 - House Tully
 - House Tyrell
- Edge color:
 - father
 - father/mother
 - mother
 - spouse







- Jeder Schritt einer wissenschaftlichen Studie muss nachvollziehbar sein
- Angabe statistischer Methoden (PCA, OPLS...) nicht ausreichend, da hunderte Variationen
- KEIN Schritt der Analyse darf manuell erfolgen (copy/paste in Excel...)

- Lösung:
 - ▶ Scriptbasierte Analysen mit 
 - ▶ R Markdown als Reportgenerator: Text, Tabellen und Abbildungen -> LaTeX / Worddokument