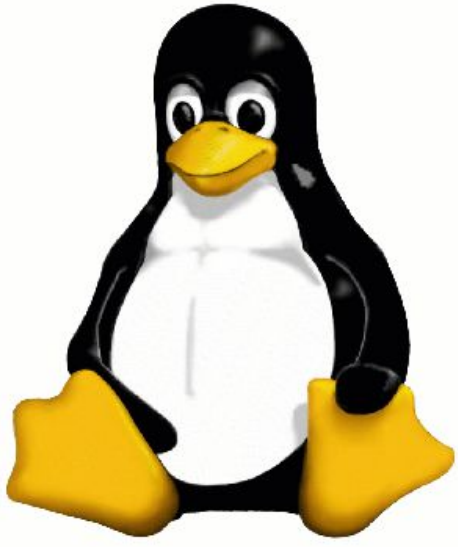
Datenbanken und -analyse

Nils Glück
&
Manuel Blechschmidt
CdE Pfingstakademie 2014
Kirchheim

Bei unklaren Begriffen bitte sofort melden

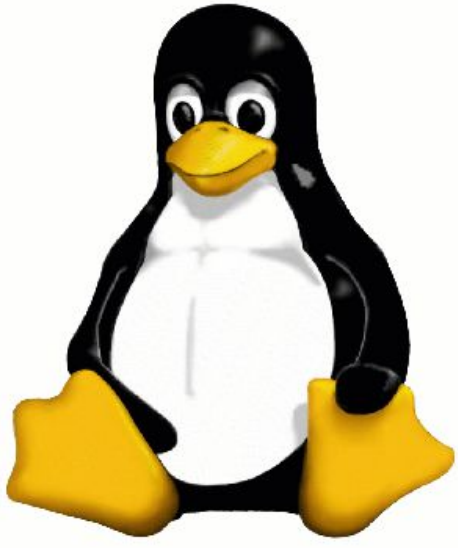
Datenanalyse

- Buzzwords:
  - BigData
  - Data Mining
  - Data Visualization
  - Business Intelligence
  - Advanced Analytics
  - Actionable Insights
  - ...



Chabos wissen wer der Babo ist

- Unser Datensatz
  - 2 Tabellen
    - adressbuch
    - telefonate
- Ganster die miteinander telefonieren

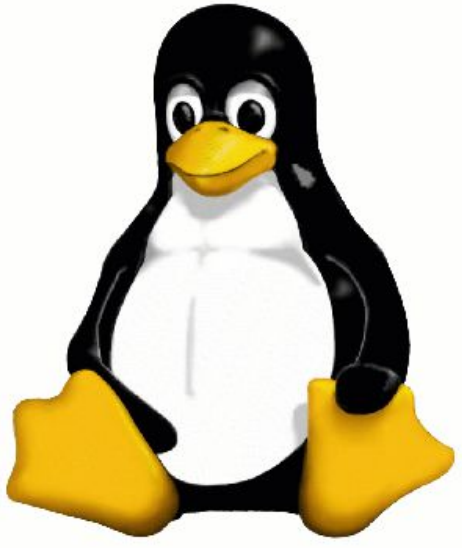


Unsere Aufgabe

- Finde heraus, ob Frauen länger telefonieren als Männer
- Finde heraus, ob die Stadt einen Einfluss auf die Länge der Telefonate hat
- Finde den ober Ganster
  - Ein Gangster ist wichtig, wenn er von vielen Leuten angerufen wird
  - Ein Gangster ist wichtig, wenn er von vielen wichtigen Gangstern angerufen wird

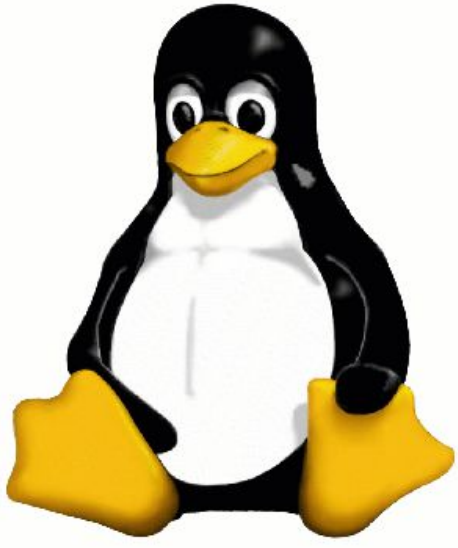
Deskriptive Statistik

- Hypothesentests
  - Sagen aus, ob eine Hypothese mit einem Signifikanzniveau (α) angenommen oder abgelehnt werden
  - Typische Niveaus: 0.05, 0.01, 0.001
- Beispiel:
  - Telefonieren Frauen länger als Männer



T-Test

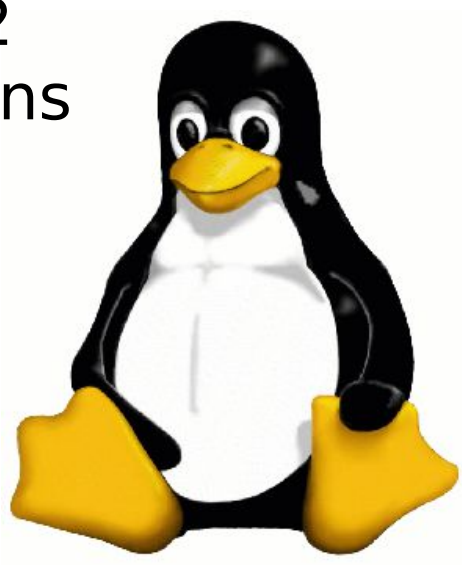
- Signifikanztest für Normalverteilungen mit einem Merkmal was nur zwei Ausprägungen haben kann
  - Mann/Frau
  - jung/alt
  - arm/reich



Beispiel

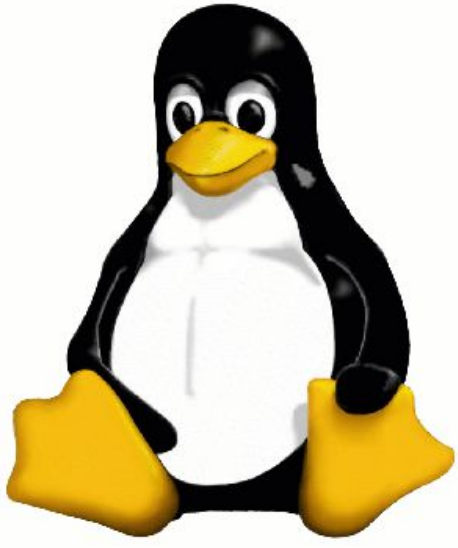
Hypothese: Männer und Frauen telefonieren nicht gleich lang.

```
t.test(telefonateMaenner, telefonateFrauen)

    Welch Two Sample t-test

data:  telefonateMaenner and telefonateFrauen
t = -2.0496, df = 213.332, p-value = 0.04162
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
 -47.50058  -0.92712
sample estimates:
mean of x mean of y
 92.58824 116.80208
```



ANOVA

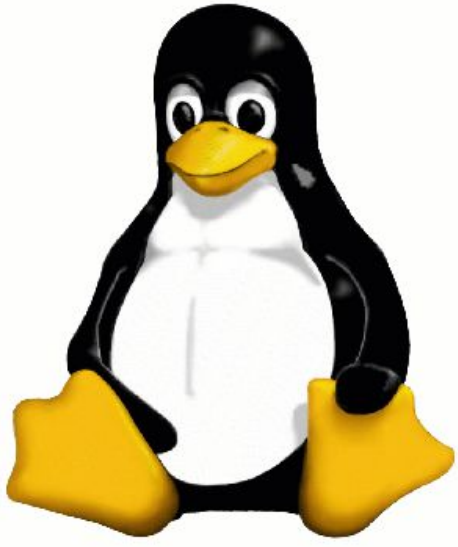
- Analysis of Variance
- Möglichkeit um festzustellen, ob ein Merkmal mit mehrer Ausprägungen Einfluss auf einer Größe hat z.B.
  - Land
  - Gewählte Partei
  - Hautfarbe
  - Schuhmarke
  - etc.



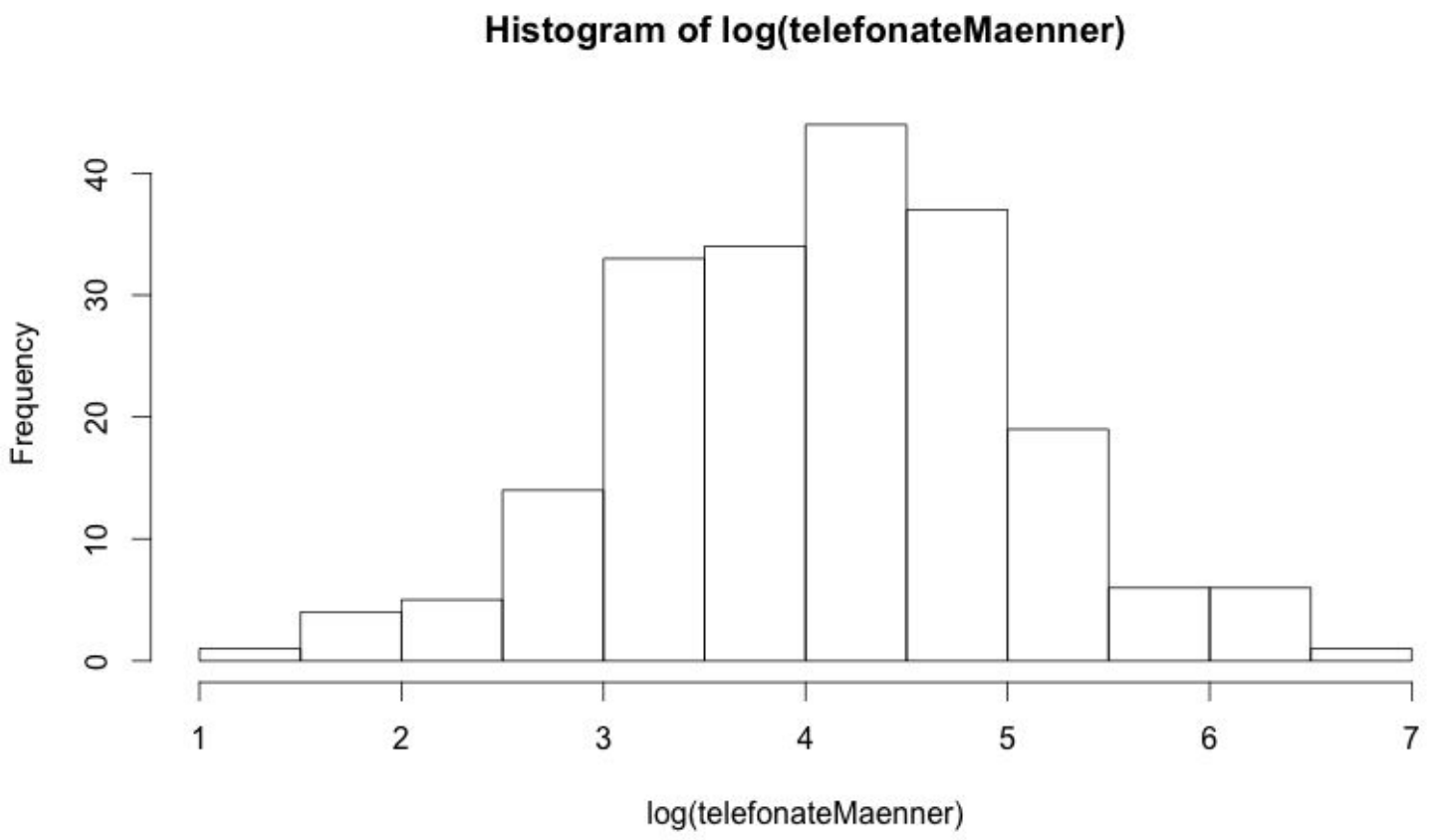
Beispiel

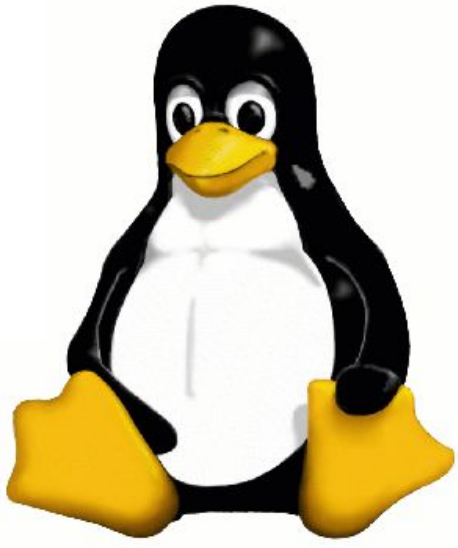
Hypothese: Das Land hat Einfluss auf die Telefonlänge

```
> model <- aov(laenge ~ land, telefonateLand)
> summary(model)
             Df  Sum Sq Mean Sq F value Pr(>F)
land         11  143424   13039   1.288   0.23
Residuals   288 2914679   10120
```



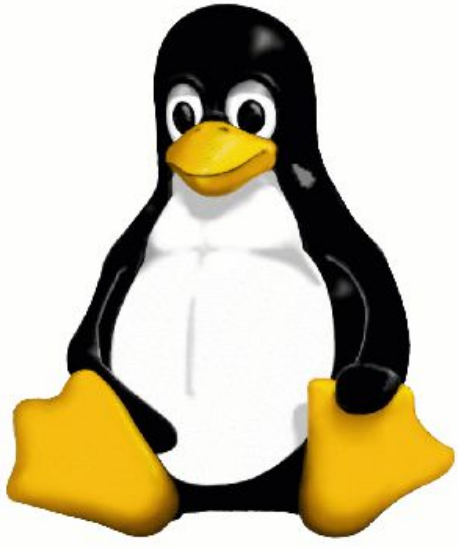
Plotten


Histogram of log(telefonateMaenner)
Frequency
0
10
20
30
40
1
2
3
4
5
6
7
log(telefonateMaenner)



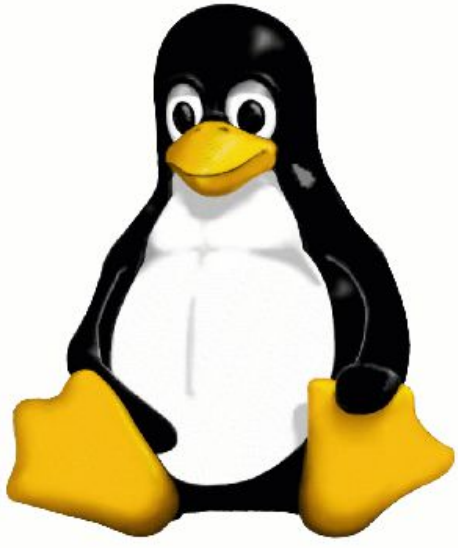
Graph Analysis mit igraph

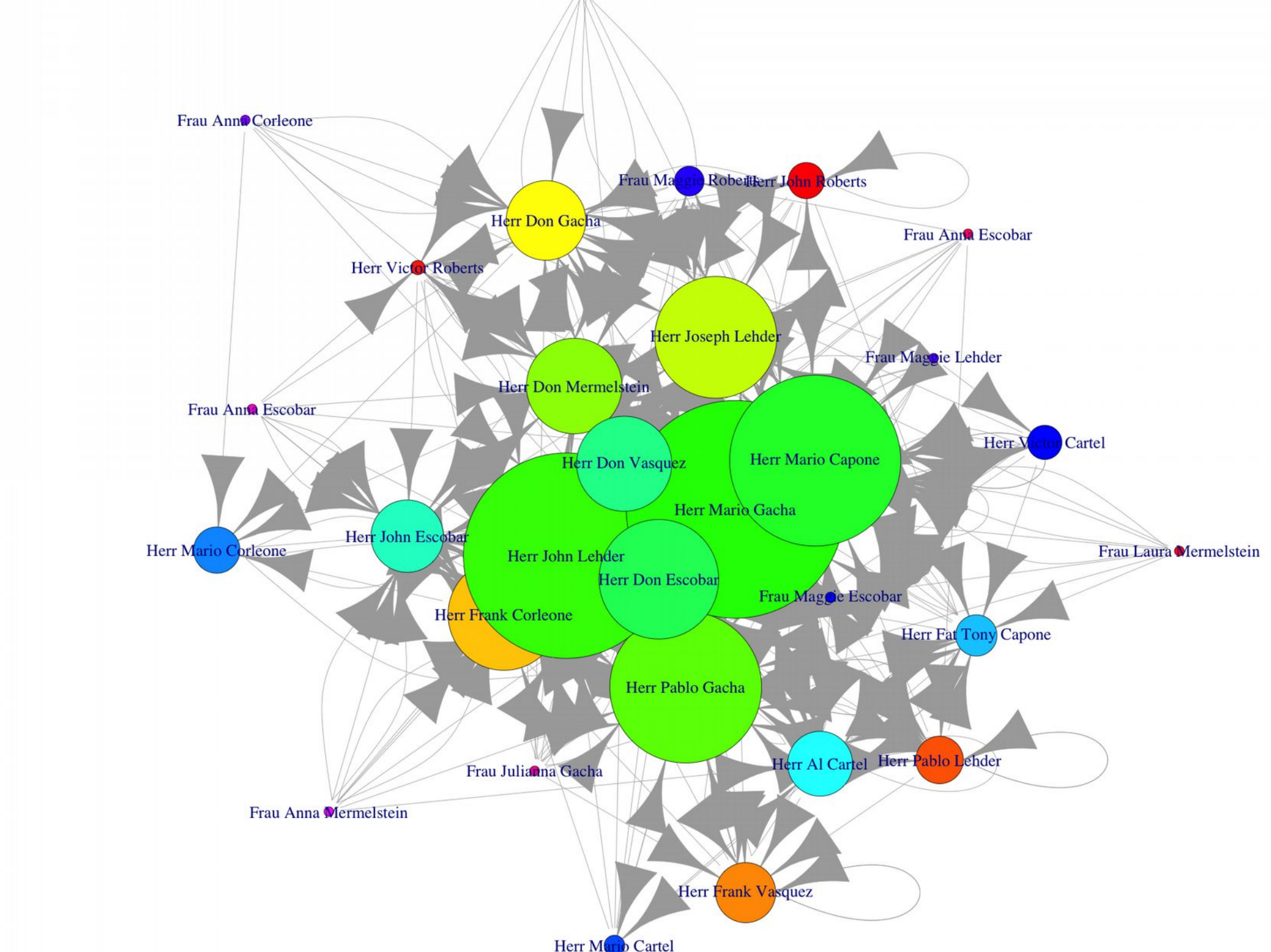
- igraph ist eine Library für verschiedene Programmiersprachen um Graphen zu analysieren und visualisieren
- Sprachen
  - R
  - Python
  - C



Aufgabe

- Daten laden
  - ETL - Extract Transform Load
- Graph bauen
- Graph analysieren
- Graph plotten




Frau Anna Corleone
Frau Maggie Roberts
Herr John Roberts
Herr Don Gacha
Frau Anna Escobar
Herr Victor Roberts
Herr Joseph Lehder
Frau Maggie Lehder
Herr Don Mermelstein
Frau Anna Escobar
Herr Victor Cartel
Herr Don Vasquez
Herr Mario Capone
Herr Mario Gacha
Herr John Escobar
Herr Mario Corleone
Herr John Lehder
Frau Laura Mermelstein
Herr Don Escobar
Frau Maggie Escobar
Herr Frank Corleone
Herr Fat Tony Capone
Herr Pablo Gacha
Herr Al Cartel
Herr Pablo Lehder
Frau Julianna Gacha
Frau Anna Mermelstein
Herr Frank Vasquez
Herr Mario Cartel

Ende