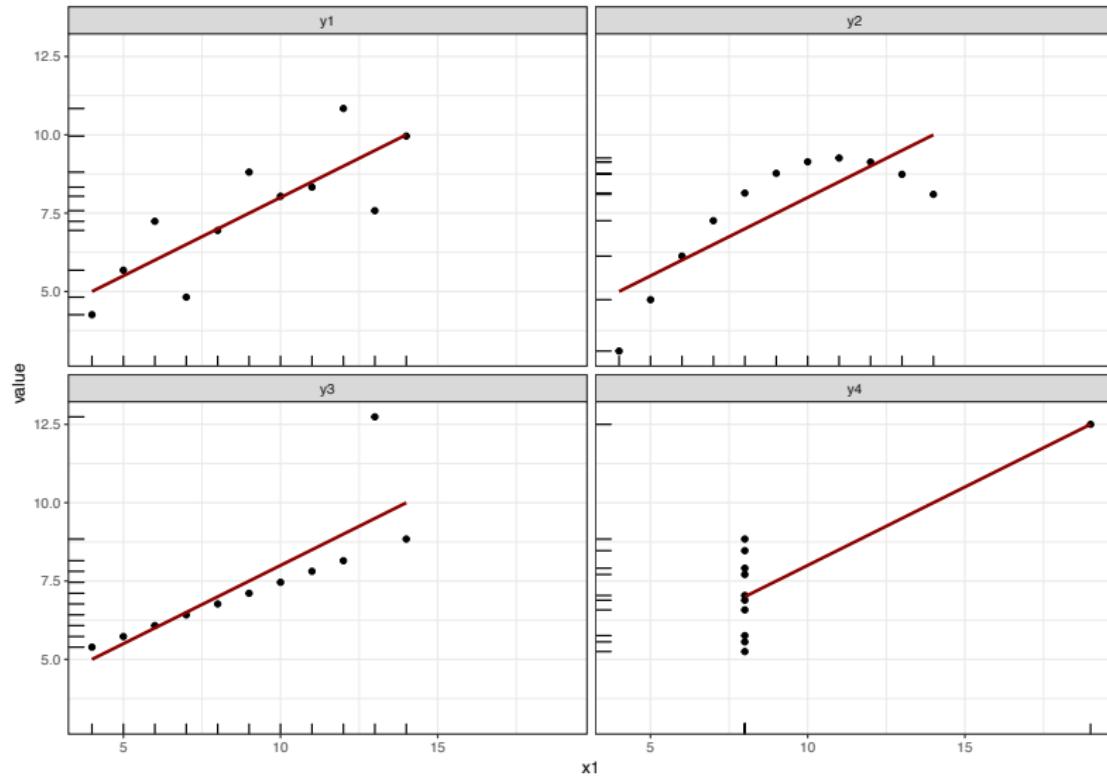


Multidimensionale Daten visualisieren

Prof. Dr. Gerrit Hirschfeld

26 Januar 2016

Anscombe's Quartett (Anscombe, 1973)



Neue Hardware erlaubt es, dass jeden Tag neuen Analysemethoden entwickelt werden, um Daten zu analysieren.

Graphiken ermöglichen es, Methoden zu verwenden, mit denen wir jeden Tag seit unserer Geburt arbeiten.

Warum funktioniert Anscombe's Quartett?

- Intuitiv verständlich: **Mappings**
 - Variable x horizontale Position
 - Variable y vertikale Position
- Kombiniert Daten und Modelle: **Layers**
 - Punkte
 - Regressionsgerade
- Wiederholt sich: Small Multiples aka **Facets**
 - Im Prinzip viermal dieselbe Graphik mit anderen Daten/Variablen.

1 Intro

2 GGplot

- Mappings
- Layers
- Facets

3 Multidimensionale Daten

- ggplot
- spezielle Pakete

4 Fazit

Das R-Paket `ggplot2` implementiert eine Grammar of Graphics (Wickham, 2010).

Graphiken bestehen aus:

- Daten und Mappings der Daten zu ästhetischen Eigenschaften
- Layers (Geometrische Objekte)
- Optional (da es hierfür jeweils gute default-Werte gibt)
 - Skalen: Die die default-Mappings verändern,
 - Koordinatensysteme, z.B. logarithmische Skalen
 - Facets, die Subplots erstellen.

Beispiel: Gewalt gegen Flüchtlinge

- Datensatz der Amadeu Antonio Stiftung (<http://www.amadeu-antonio-stiftung.de>).
- N = 3311 Vorfälle (Stand: 7.1.2017)
- Aktuelle Daten unter: <https://raw.githubusercontent.com/ax31/chronik-vorfaelle/data/vorfaelle.csv>

lon	lat	typ	monat
12.109271	49.32619	Brandanschlag	25
12.109271	49.32619	Brandanschlag	25
9.935205	51.53276	Sonstige Angriffe auf Unterkünfte	25
12.466217	50.84994	Tälicher Übergriff/Körperverletzung	25
7.849400	47.99609	Sonstige Angriffe auf Unterkünfte	25
10.792532	51.50516	Sonstige Angriffe auf Unterkünfte	25

Mapping

Wie werden Eigenschaften der Daten auf ästhetische Eigenschaften der Graphik gemappt?

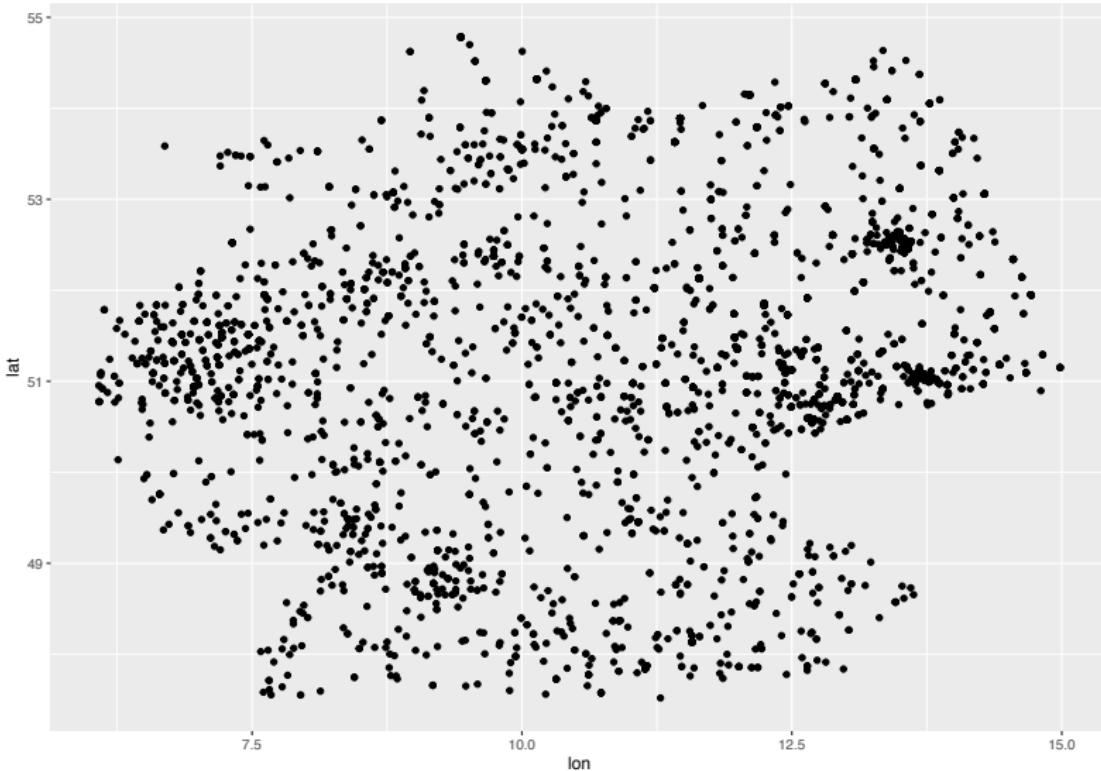
Häufigstes Mapping: 2D-Position

- Variable X: horizontale Position
 - Variable Y: vertikale Position

weitere Mappings:

- Farbe
 - Form

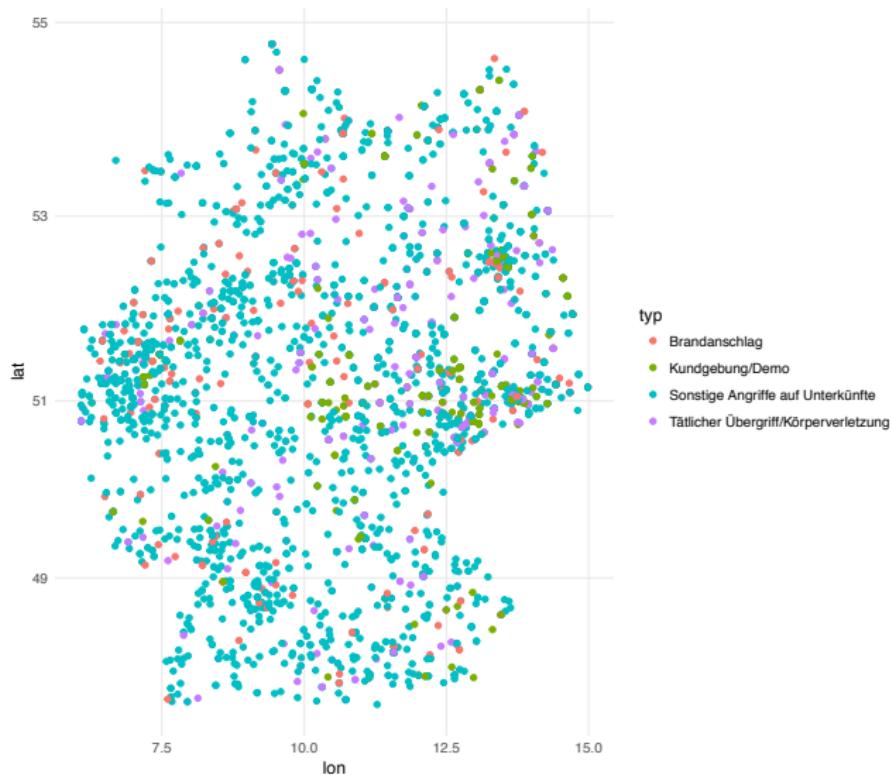
```
ggplot(brand, aes(x=lon, y=lat)) +  
  geom_point()
```



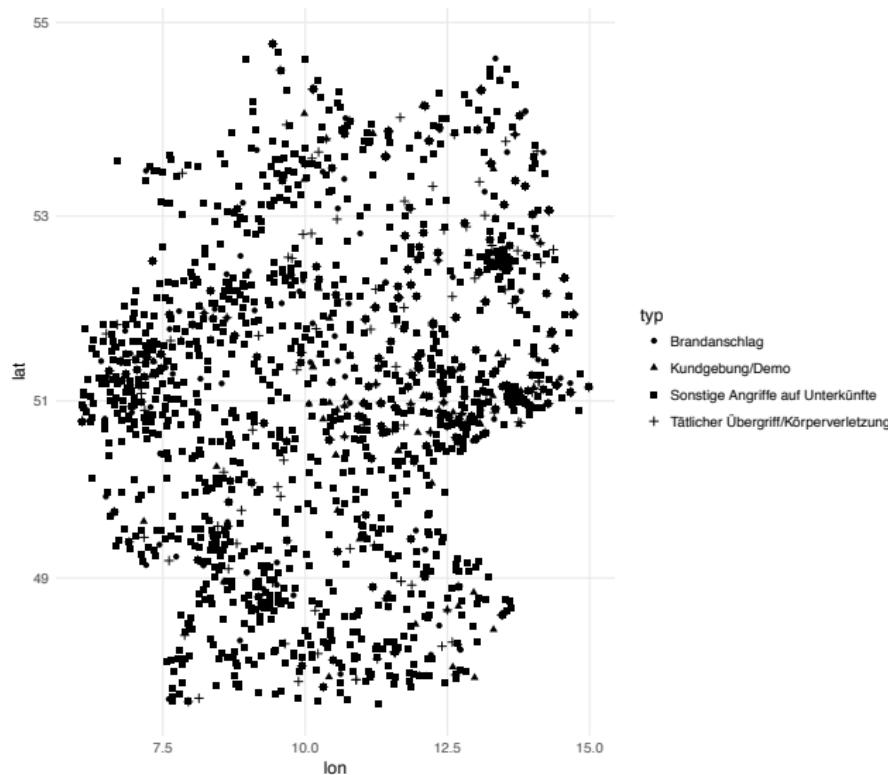
```
# Richtiges Projektion und Hintergrund entfernen  
ggplot(brand, aes(x=lon, y=lat)) +  
  geom_point() +  
  coord_map() + theme_minimal()
```



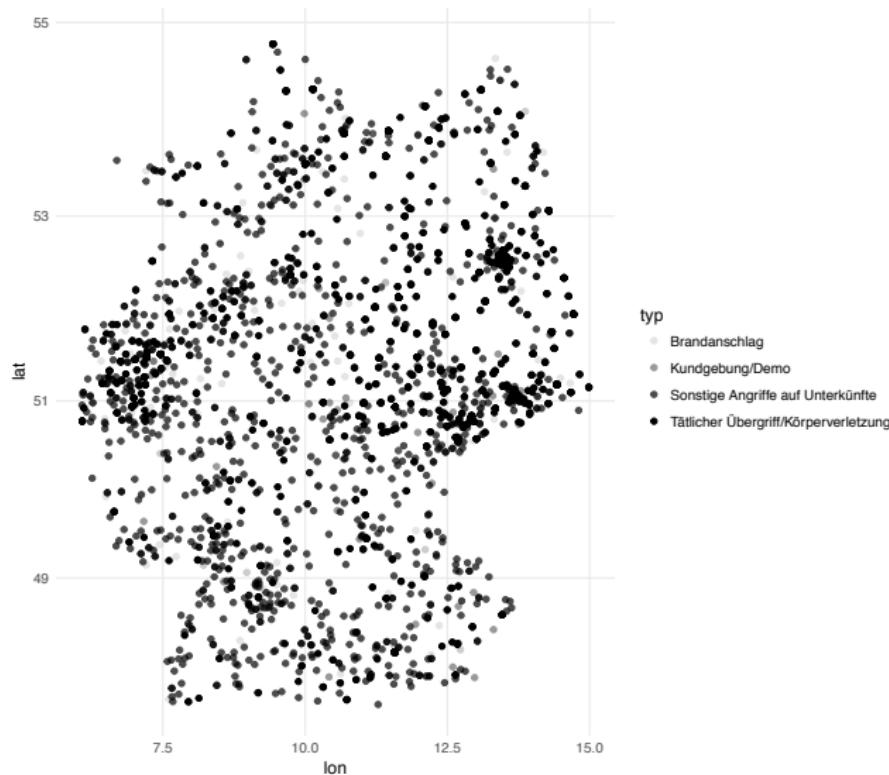
```
# Typ des Vorfalls auf Farbe gemappt
ggplot(brand, aes(x=lon, y=lat, col = typ)) +
  geom_point() +
  coord_map() + theme_minimal()
```



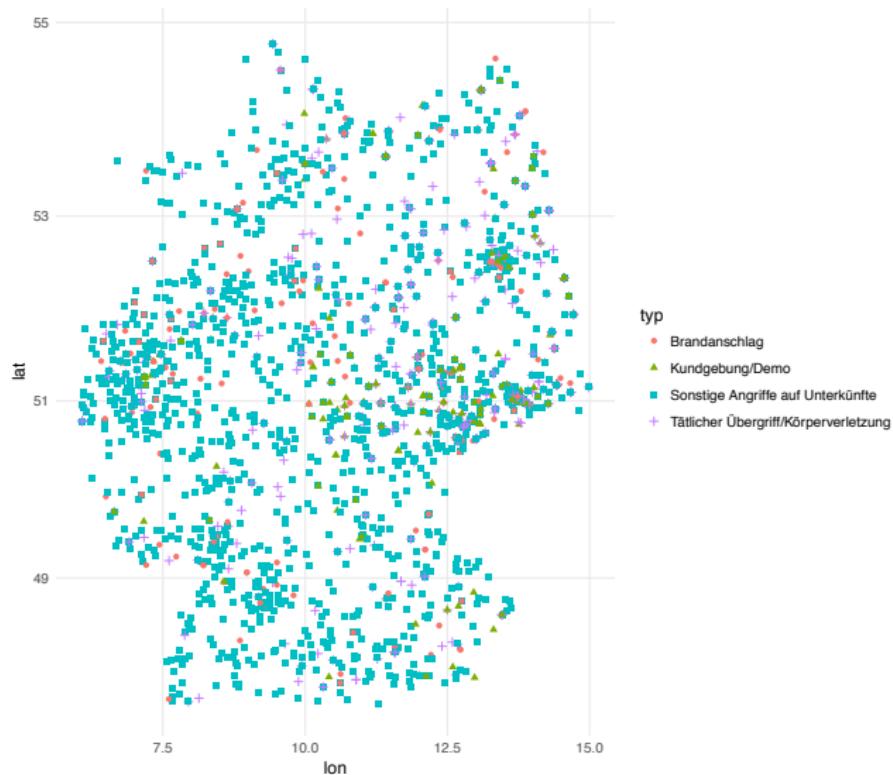
```
# Typ des Vorfalls auf Form gemappt  
ggplot(brand, aes(x=lon, y=lat, shape = typ)) +  
  geom_point() +  
  coord_map() + theme_minimal()
```



```
# Typ des Vorfalls auf Transparenz gemappt  
ggplot(brand, aes(x=lon, y=lat, alpha = typ)) +  
  geom_point() +  
  coord_map() + theme_minimal()
```

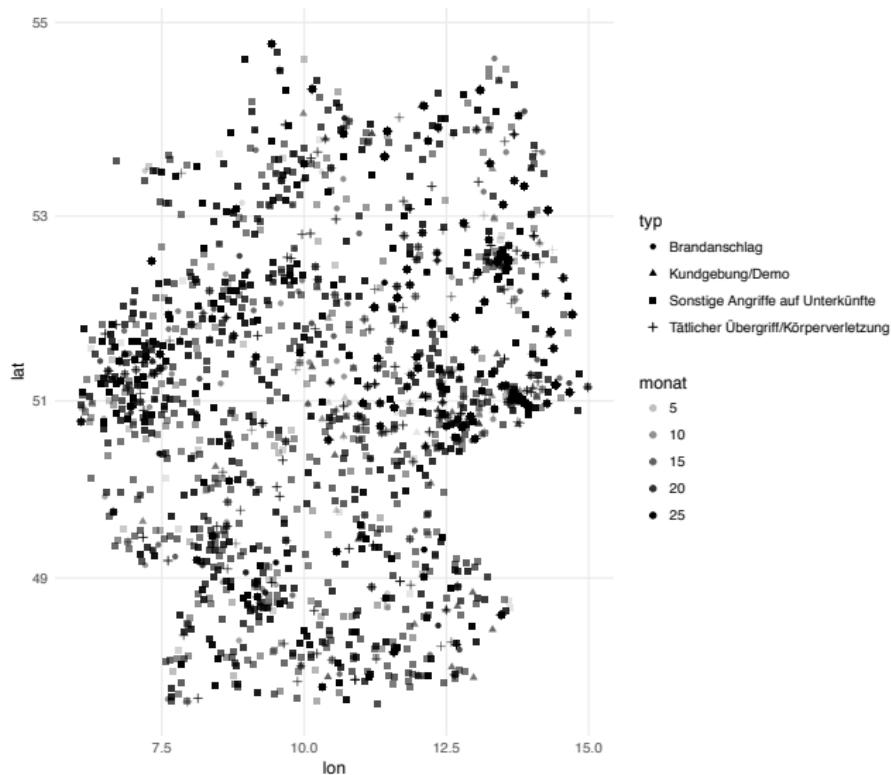


```
# Typ des Vorfalls auf Farbe und Form gemappt
ggplot(brand, aes(x=lon, y=lat, col = typ, shape = typ)) +
  geom_point() +
  coord_map() + theme_minimal()
```

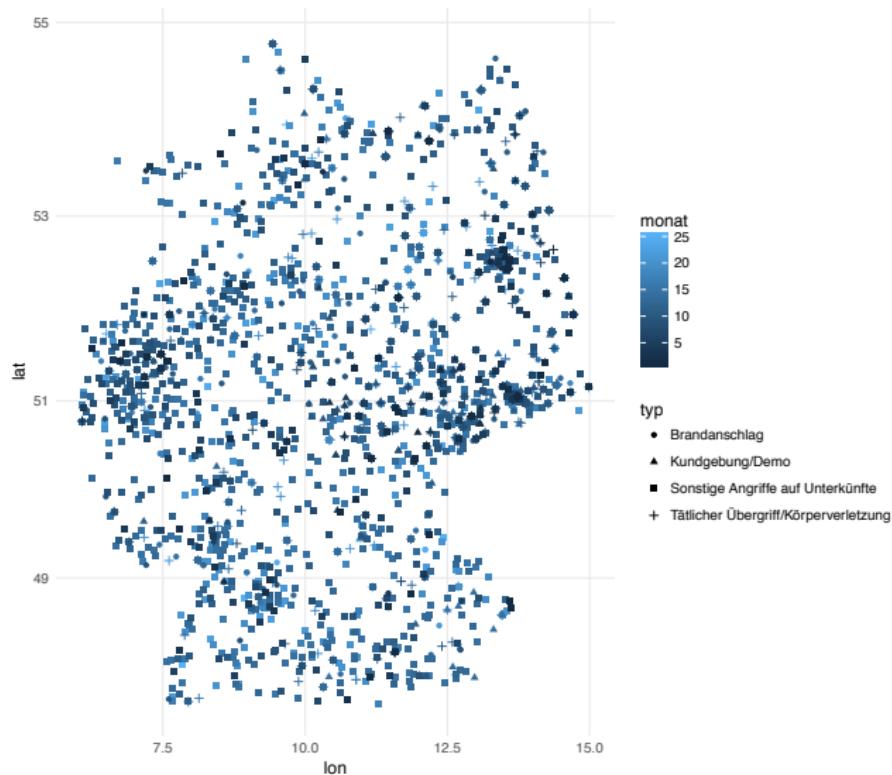


Typ des Vorfalls auf Form, Datum auf Transparenz gemapppt

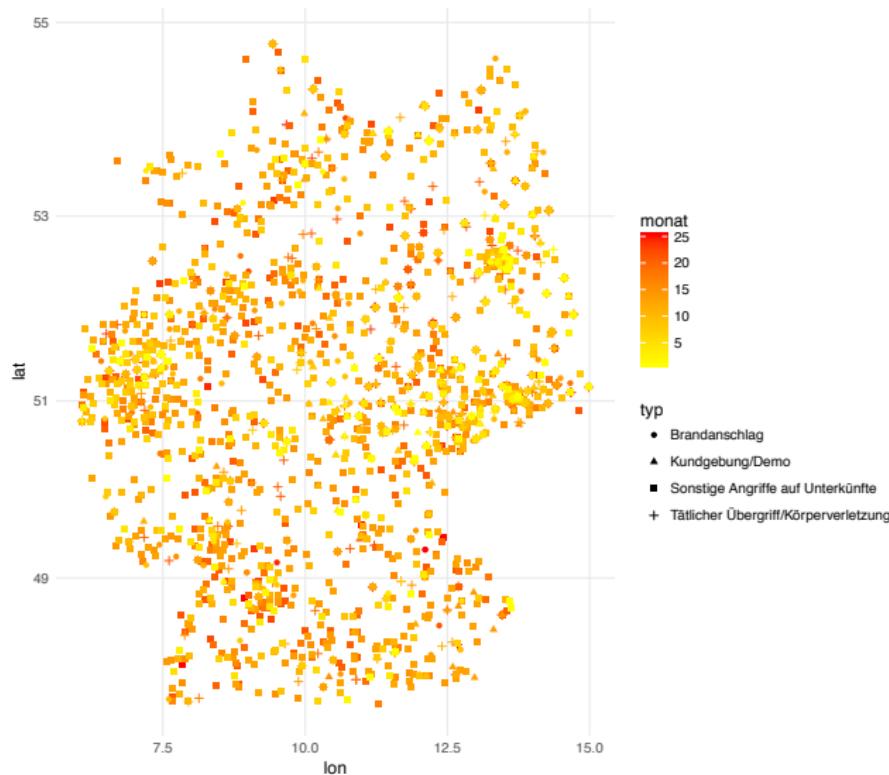
```
ggplot(brand, aes(x=lon, y=lat, shape = typ, alpha = monat)) +  
  geom_point() +  
  coord_map() + theme_minimal()
```



```
# Typ des Vorfalls auf Form, Datum auf Farbe gemappt
ggplot(brand, aes(x=lon, y=lat, shape = typ, col = monat)) +
  geom_point() +
  coord_map() + theme_minimal()
```



```
# Alle Mappings lassen sich durch scale_x anpassen!
ggplot(brand, aes(x=lon, y=lat, shape = typ, col = monat)) +
  geom_point() +
  scale_colour_gradient(low="yellow", high = "red") +
  coord_map() + theme_minimal()
```



Übersicht Mappings:

Art des Mappings	Qual.	Ord.	Quant.	Obacht
Position	X	X	X	Karten, Log.-Skalen
Helligkeit	X	X	-	Medienspezifisch
Transparenz	X	X	-	Medienspezifisch
Größe	X	X	-	Medienspezifisch
Farbe:	X	X	-	Farbfehlsichtigkeit
Form:	X	-	-	

— Weiterlesen: <http://www.cookbook-r.com/Graphs/> —

- Wie man die Mappings durch scale verändert.
- Wie man die Legende gestaltet.

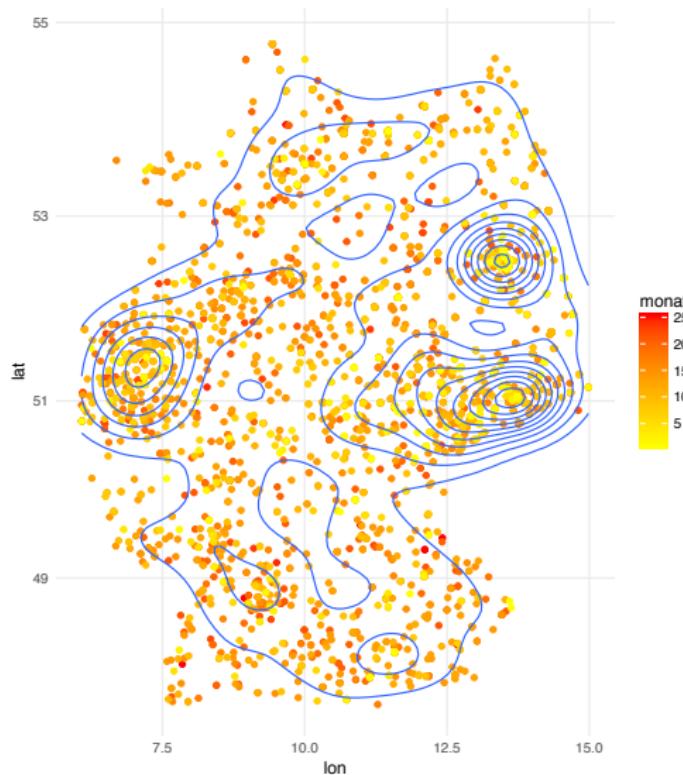
Layers

Wie kombiniert man verschiedene graphische Objekte (Punkte, Linien etc.) innerhalb desselben Koordinatensystems?

Beispiel

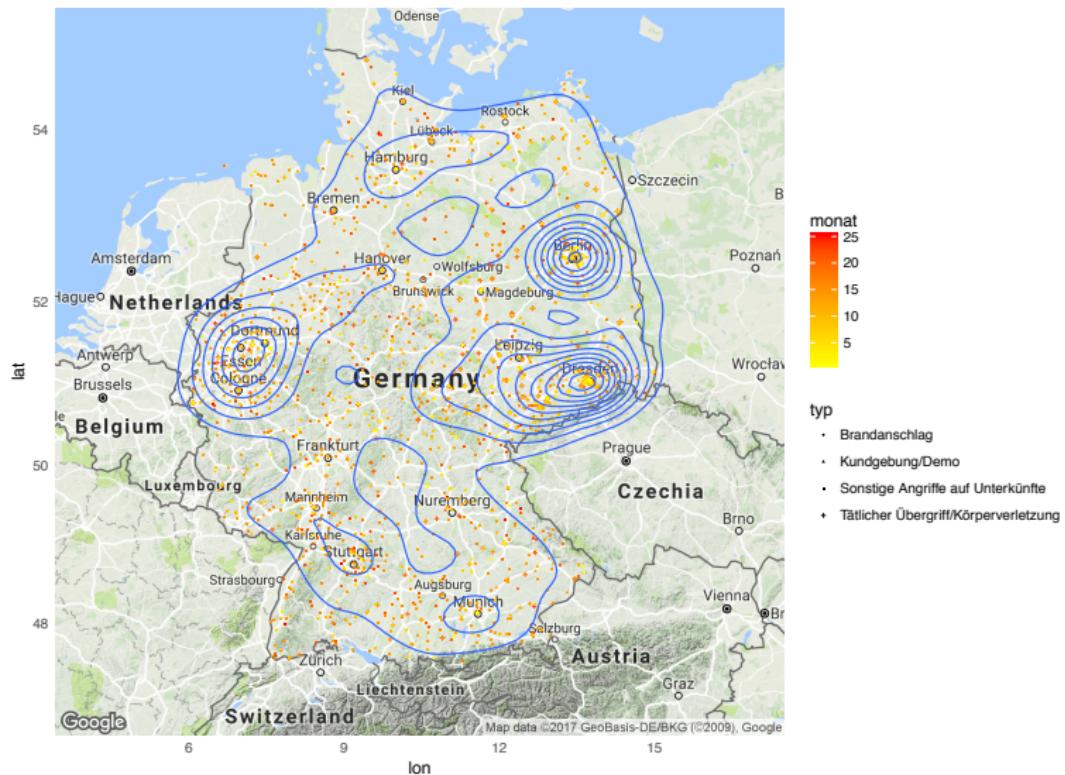
- Layer 1: Rohdaten
- Layer 2: Statistiken
 - Mittelwerte
 - Standardabweichungen / Konfidenzintervalle
 - Boxplots
 - Regressionsgeraden
- Layer 3: Ländergrenzen

```
# ggplot kann einige Statistiken berechnen und einzeichnen!
ggplot(brand, aes(x=lon, y=lat, col = monat)) +
  geom_point() +
  geom_density2d() +
  scale_colour_gradient(low="yellow", high = "red") +
  coord_map() + theme_minimal()
```

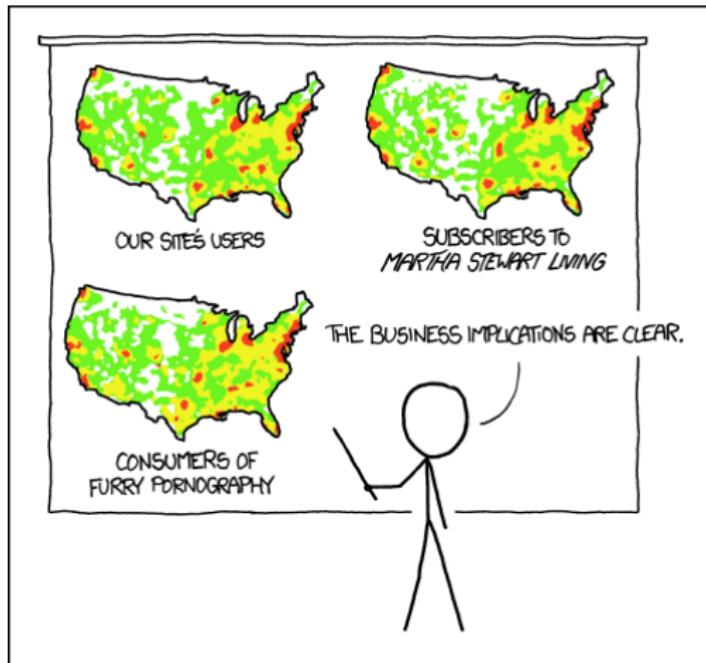


```
library(ggmap)
google.map <- get_map(location="Germany", zoom =6, maptype = "te

ggmap(google.map) +
  geom_point(data = brand, aes(x=lon, y=lat, shape = typ, col =
  geom_density2d(data = brand, aes(x=lon, y=lat)) +
  scale_colour_gradient(low="yellow", high = "red") +
  coord_map() + theme_minimal()
```



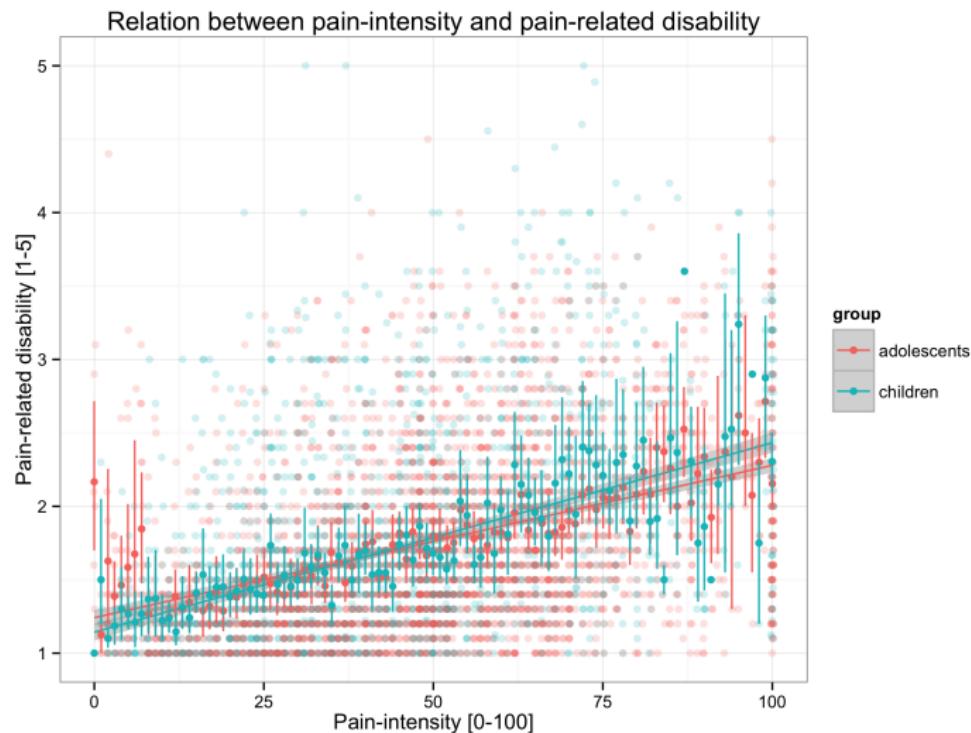
Obacht bei Karten



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

Figure 1: Quelle: <https://xkcd.com/1138/>

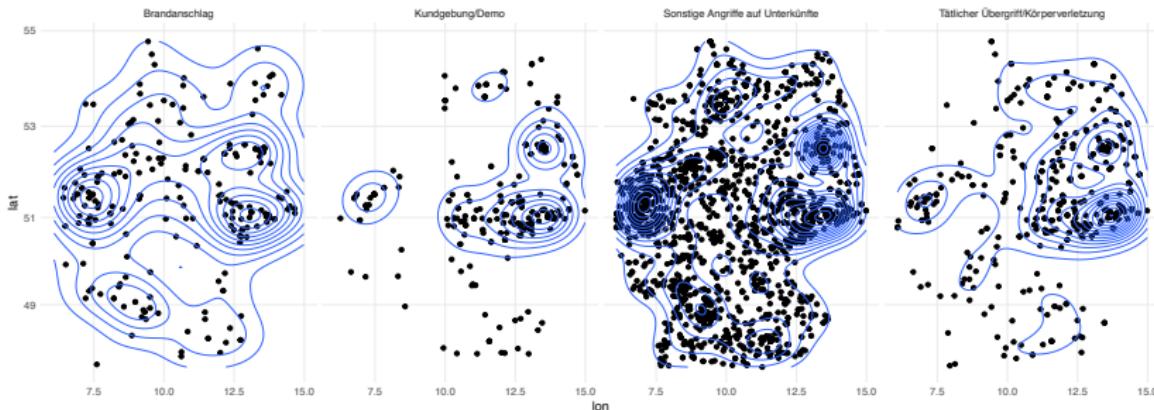
Kombination von Rohdaten, Konfidenzintervallen und Regressionen (Hirschfeld & Zernikow, 2013).



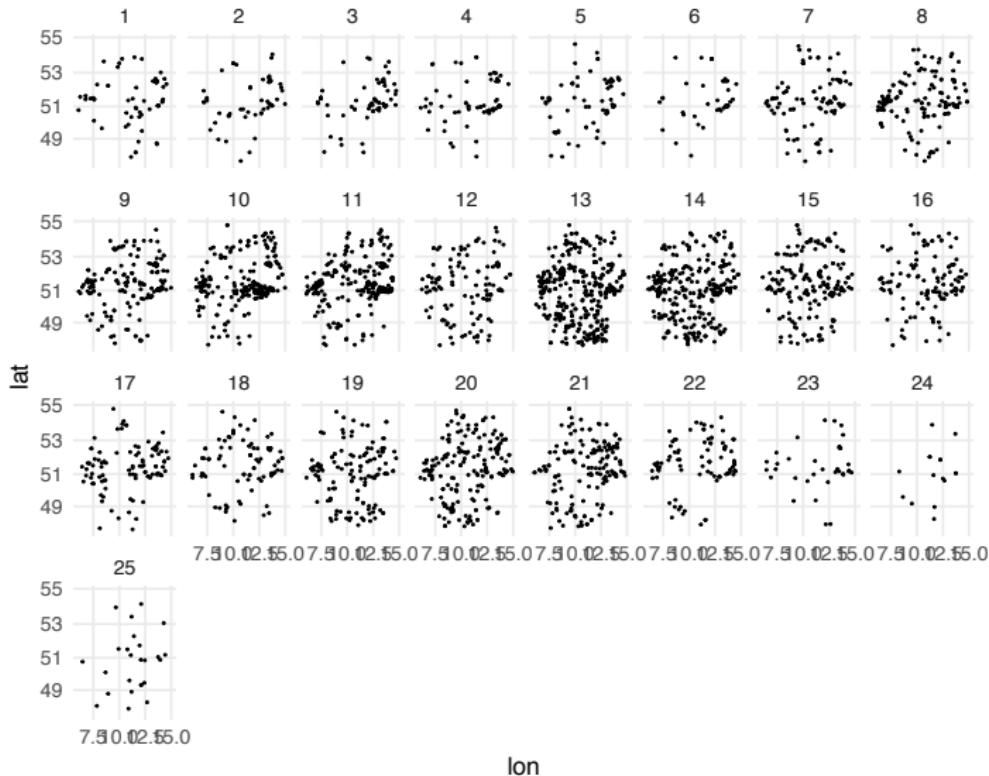
Facets (aka small multiples)

Wie kann man dieselbe Graphik für verschiedene Gruppen zeichnen?

```
ggplot(brand, aes(x=lon, y=lat)) +
  geom_point() +
  geom_density2d() +
  facet_wrap(~typ, ncol = 4) +
  coord_map() + theme_minimal()
```



```
ggplot(brand, aes(x=lon, y=lat)) +
  geom_point(size=.2) +
  facet_wrap(~monat, ncol = 8) +
  coord_map() + theme_minimal()
```



Vielleicht sogar besser als Videos?

```
library(gganimate)

p<-ggmap(google.map) +
  geom_point(data = brand, aes(x=lon, y=lat,
    shape = typ, frame = datum, cumulative = TRUE))

gganimate(p, filename = "output.mp4", ani.width=1440,
  ani.height=900)
```

Grammar of Graphics für multidimensionale Daten

- Daten umstrukturieren!
 - Jede Variable in einem einzelnen Facet.

Spezielle Pakete!

- **Bivariate Korrelationen:** Korrelationstabelle
 - **Netzwerke:** Netzwerkstrukturen
 - **Clusteranalysen:** Dendrogramm
 - **Regressionsmodelle:** Parameterschätzer

Beispiel big 5

Fragebogen zur Persönlichkeit bestehend aus 50 Items z.B. "Ich bin der Mittelpunkt jeder Party", "Ich treffe gern neue Leute.". Antworten von knapp 20.000 Probanden. Rohdaten unter:

http://personality-testing.info/_rawdata/

Wir plotten nur 50.000 der 1 Millionen Datenpunkte...

```

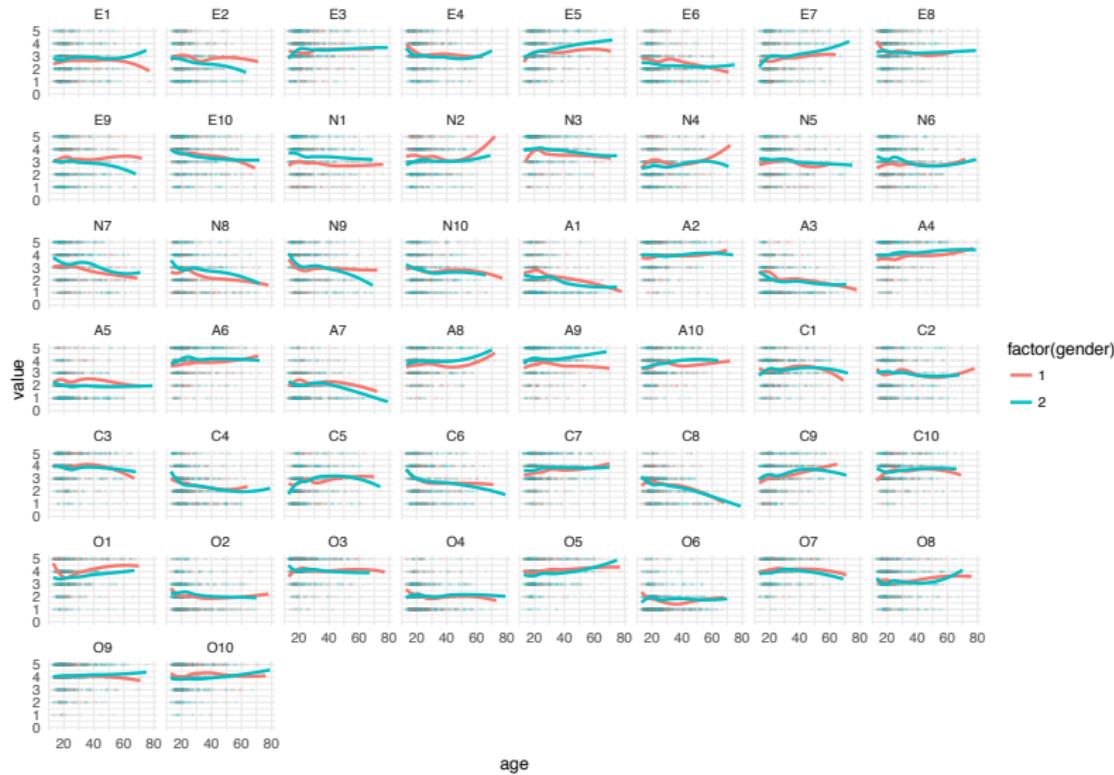
library(reshape2)
big5 <- read.delim("data/big_5.csv")
big5<-subset(big5, age < 90 & gender < 3 & gender >0)

tmp<-melt(big5[,c(2,4,8:57)], id.vars = c("age", "gender"))
tmp2<-tmp[sample(nrow(tmp), 50000),] #Darstellbarkeit!

ggplot(tmp2, aes(x=age, y=value, col = factor(gender))) +
  geom_point(size=.2, alpha = 0.1) +
  stat_smooth( se=F) +
  facet_wrap(~variable) +
  theme_minimal()

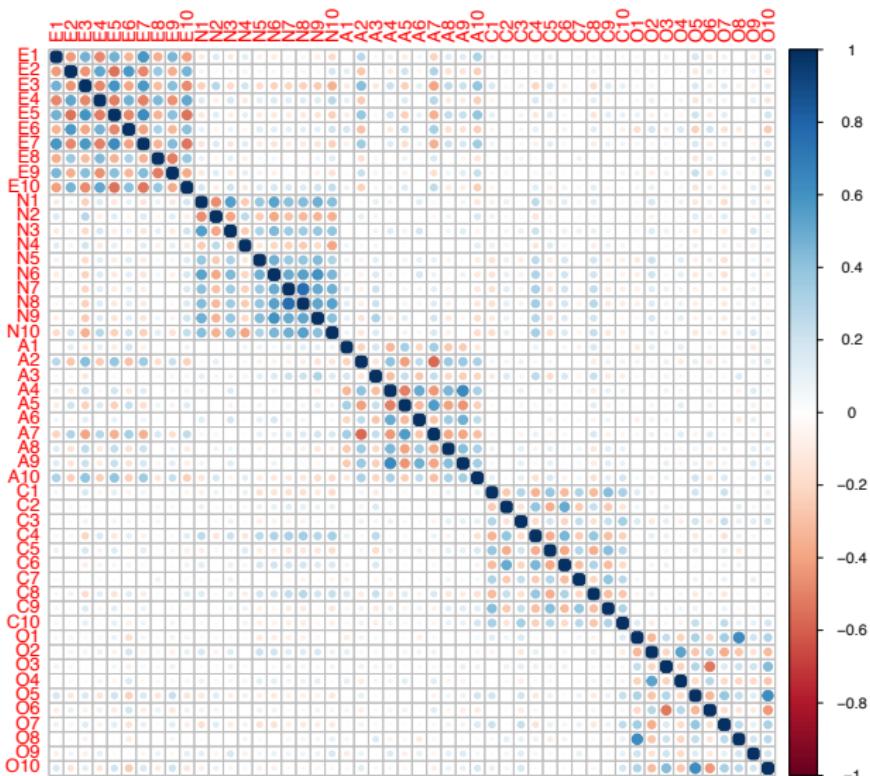
## `geom_smooth()` using method = 'loess'

```



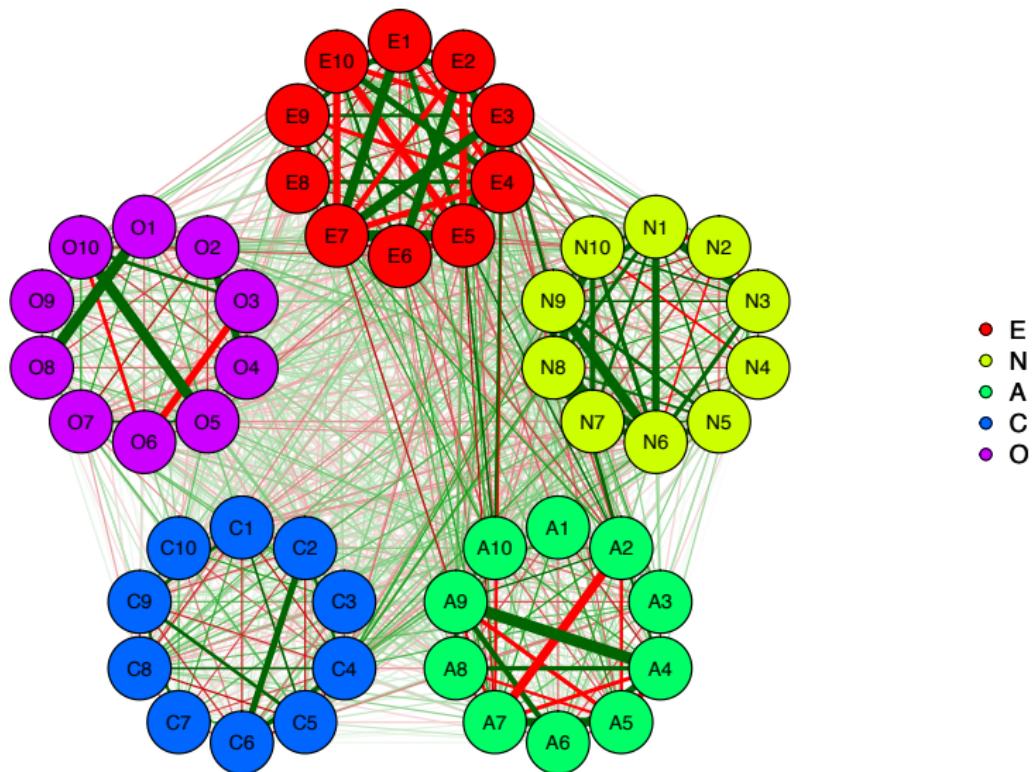
Bivariate Korrelationen

```
library(corrplot)
corrplot(cor(big5[,c(8:57)]))
```



Bivariate Korrelationen

```
library(qgraph)
b5groups<-list(E=1:10, N=11:20, A=21:30, C=31:40, O=41:50)
qgraph(cor(big5[,c(8:57)]), groups = b5groups)
```



Cluster Analysen

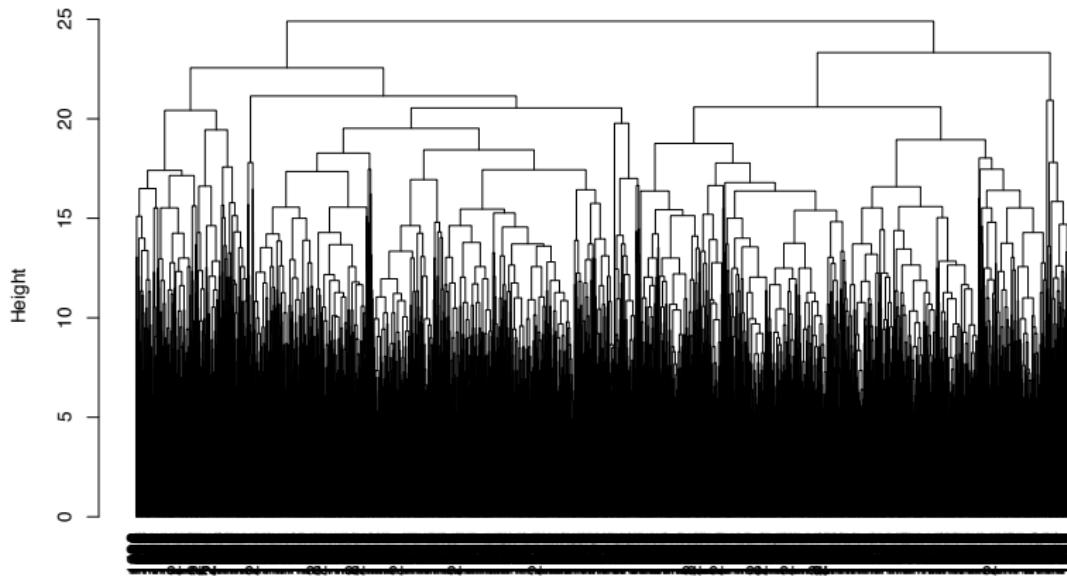
```
library(cluster)
library(flashClust)

##
## Attaching package: 'flashClust'

## The following object is masked from 'package:stats':
## 
##      hclust

big5_dist<-dist(big5[1:2000,c(8:57)])
big5_hc<-flashClust(big5_dist)
plot(big5_hc, h=-1)
```

Cluster Dendrogram



```
    d  
hclust (*, "complete")
```

Regressions Parameter

```
library(sjPlot)
mod<-lm(age ~ E1 + E2 + E3 + N1 + N2 + N3 + A1 + A2, data = big5)
sjp.lm(mod)
```


Obacht!

Nutze bekannte Skalierungen!

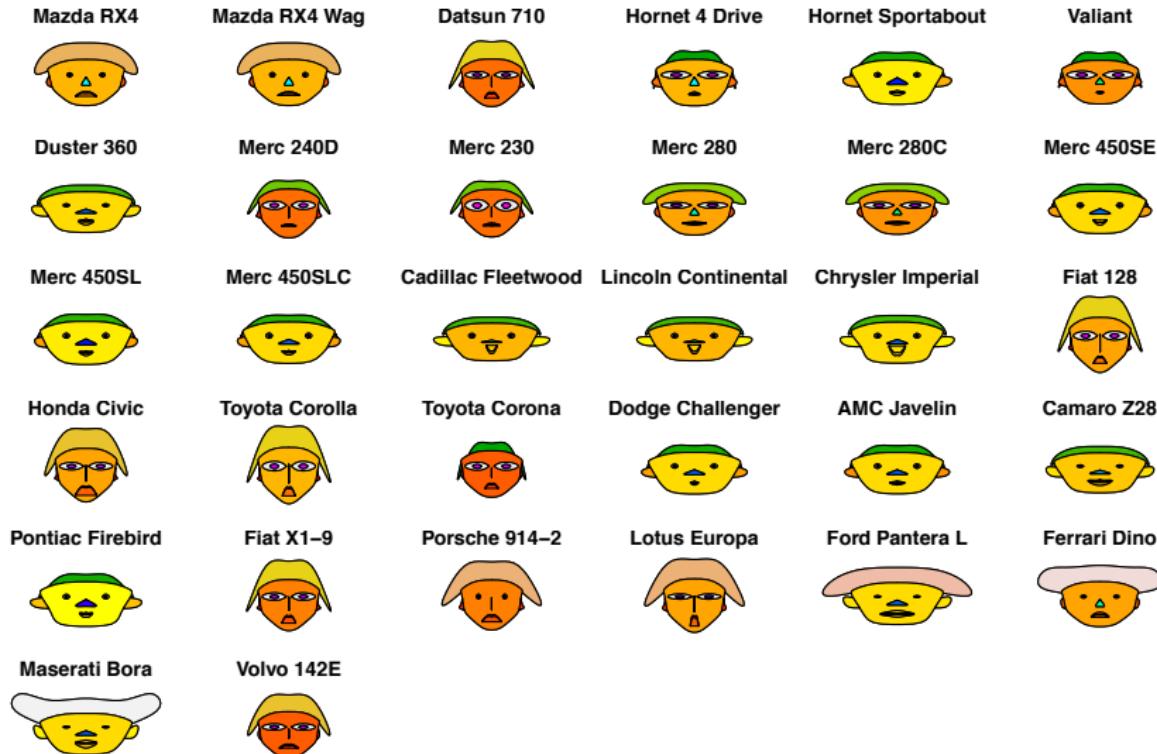
- Bei Karten: Mercator Projektion `coord_map()`
 - Log-skalen nur wenn Standard in dem Feld

Maximiere Data/Ink Ratio!

- Plotte Rohdaten!
 - Kein Hintergrund: *theme_minimal()*

Chernov faces (Chernoff, 1973)

```
library(aplpack)  
faces(mtcars)
```



Fazit

Nutze ggplot / grammar of graphics

- Mappings: Intuitive Mappings!
- Layers: Kombination von Daten und Modellen
- Facets: Wiederhole kleine Abbildungen

Multidimensionale Daten

- Strukturiere die Daten um!
- Verwende spezielle Packages

Referenzen I

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17–21.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342), 361–368.

Hirschfeld, G., & Zernikow, B. (2013). Cut points for mild, moderate, and severe pain on the vas for children and adolescents: What can be learned from 10 million anovas? *PAIN*, 154(12), 2626–2632.

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1), 3–28.