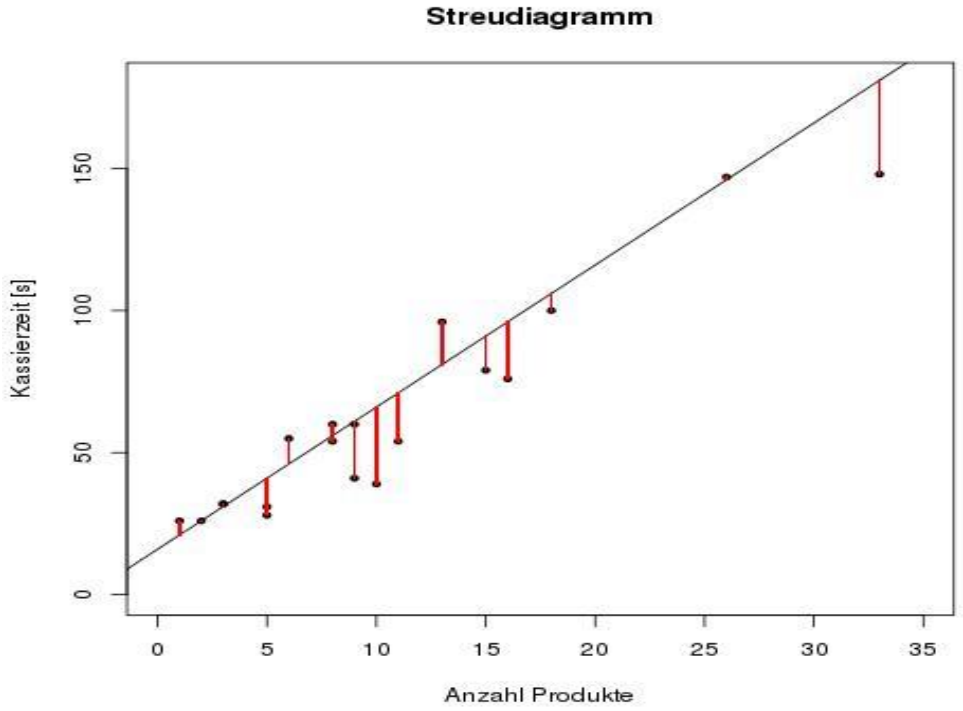
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Multiple Lineare Regression

Statistik (Biol./Pharm./HST) – Herbst 2013

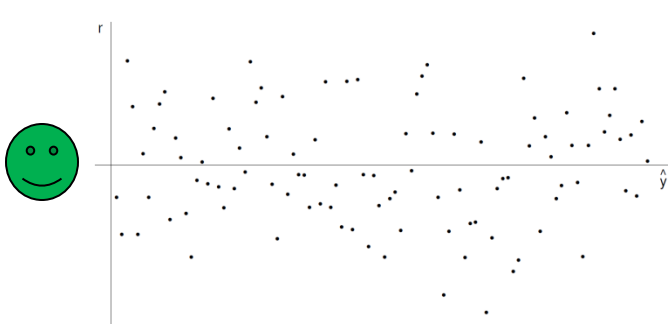
Wdh: Einfache lineare Regression

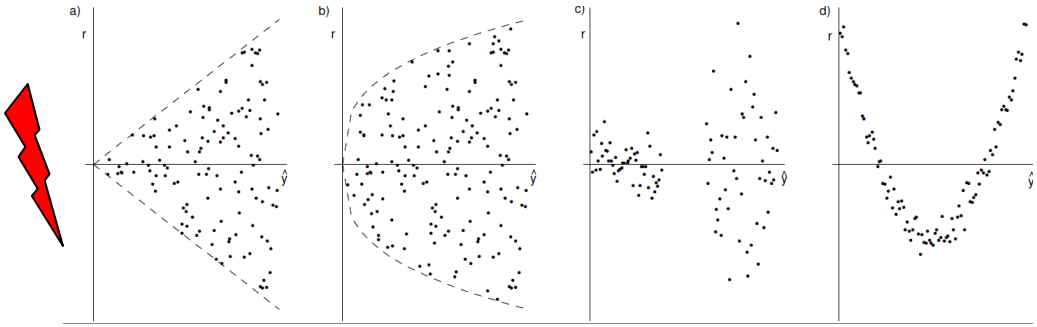
- Modell: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \;\; i.i.d$
- Finde $\widehat{\beta_0}, \widehat{\beta_1}$: Methode der kleinsten Quadrate
  $\widehat{\sigma^2}$ ist geschätzte Varianz der Residuen
- $\frac{\hat{\beta}_k - \beta_k}{\widehat{s.e.}(\hat{\beta}_k)} \sim t_{n-2}$ → t-Test: $H_0: \beta_k = 0, \;\; H_A: \beta_k \neq 0$
- R: Funktion ‘lm’



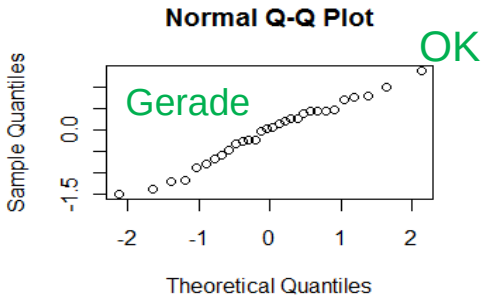
Wdh: Residuenanalyse
Sind Modellannahmen erfüllt?

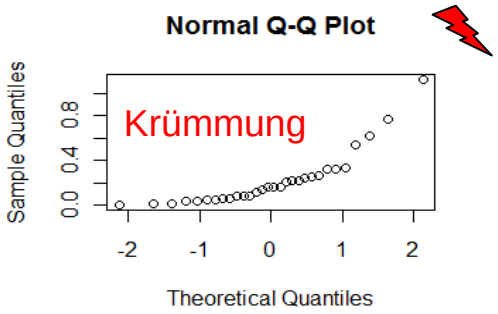
- Tukey-Anscombe Plot: Modellwert vs. Residuen
  (Fehlervarianz konstant, systematische Fehler)





- QQ-Plot: Empirische Quantile vs. theoretische Quantile
  (Residuen normalverteilt)





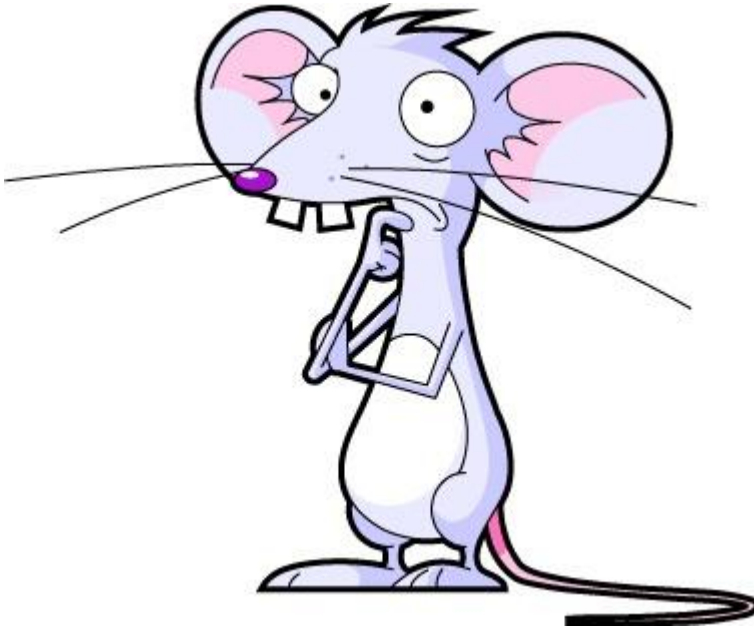
Falls Residuenanalyse schlecht: Transformationen

Zusammenhang:

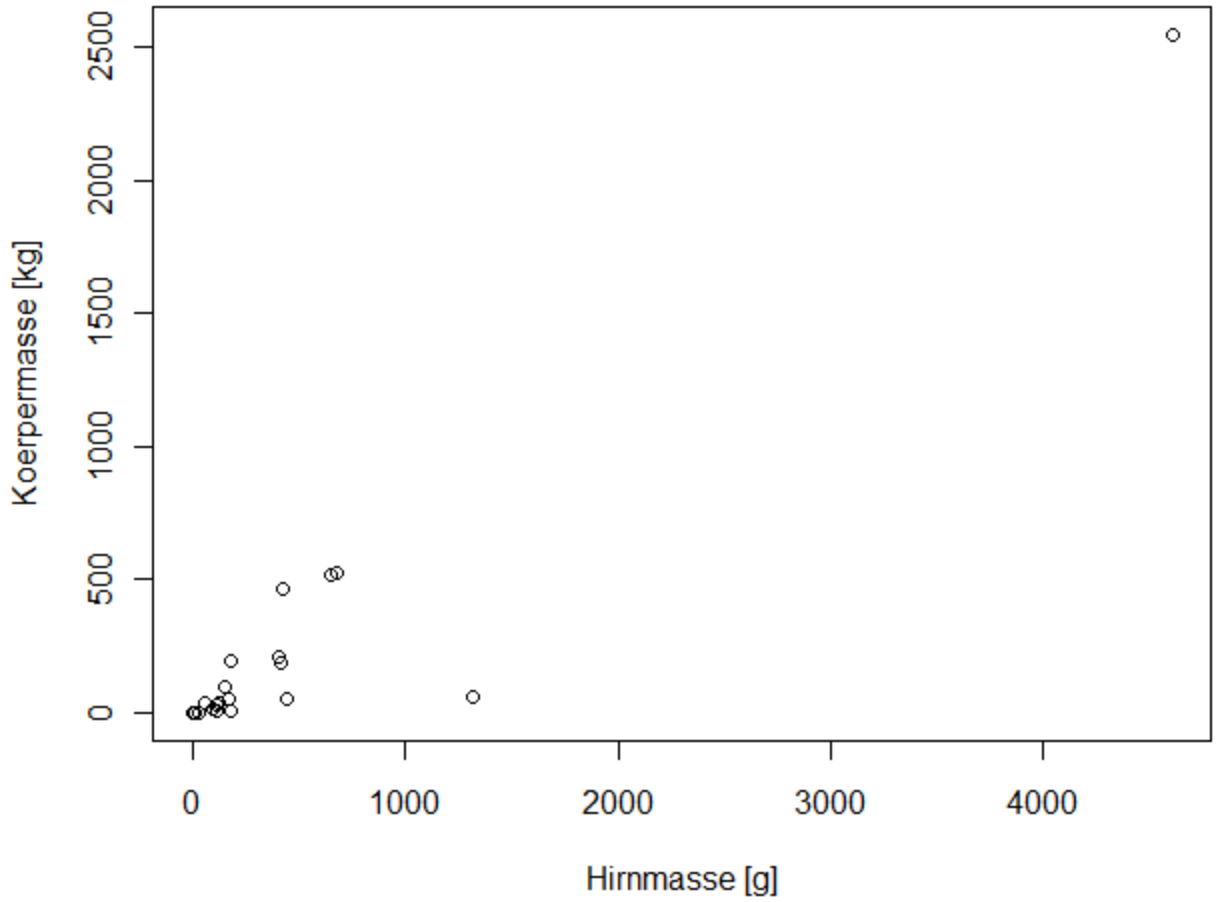
Hirnmasse

und

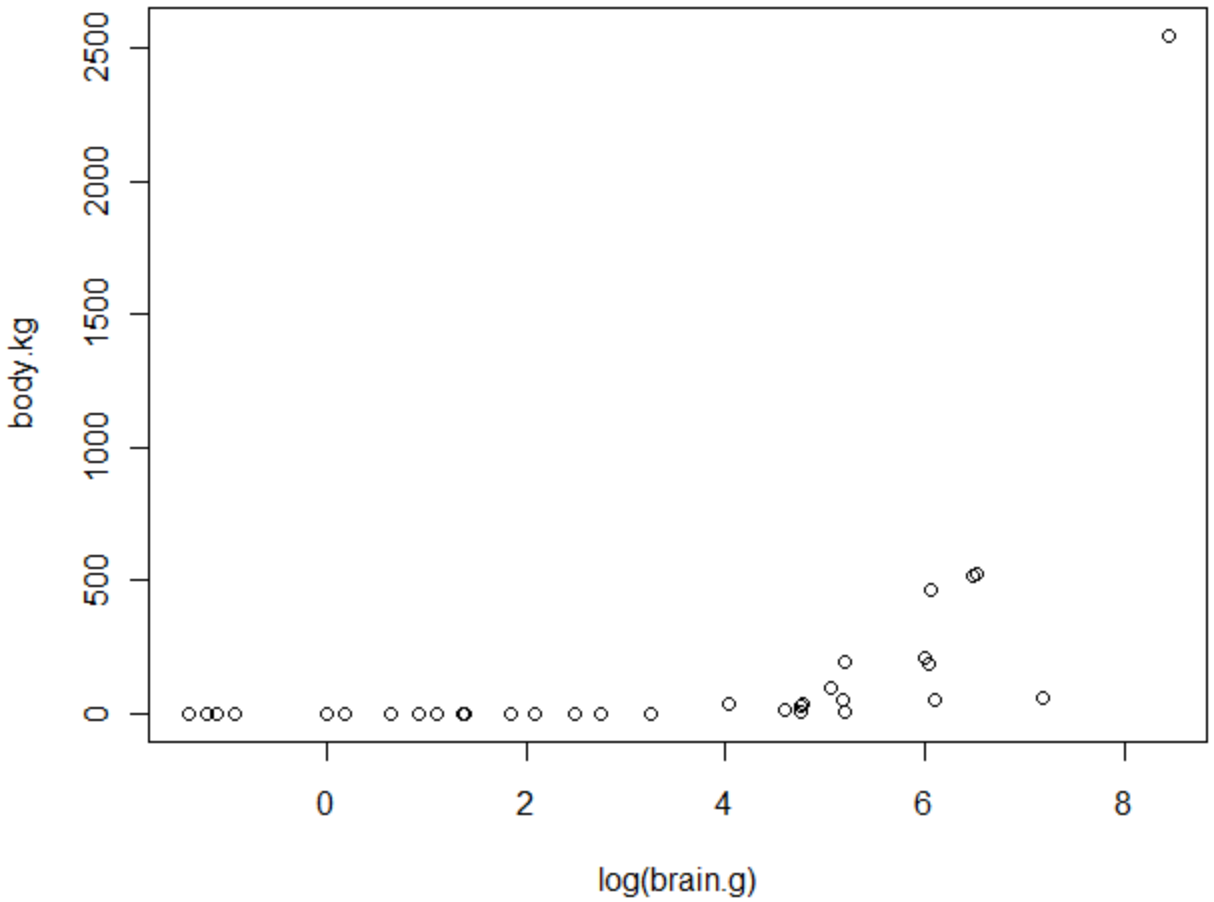
Körpermasse



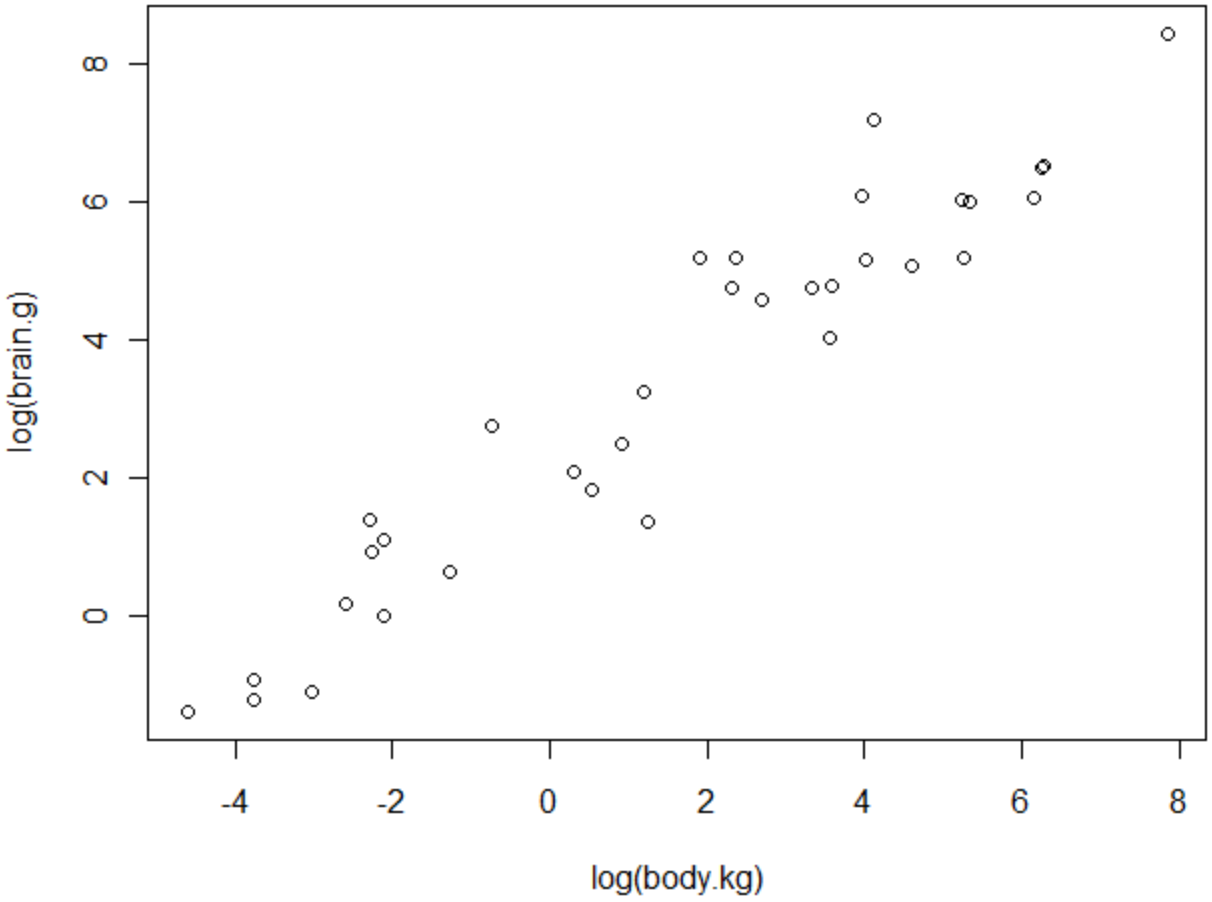
Bsp: Hirnmasse vs. Körpermasse



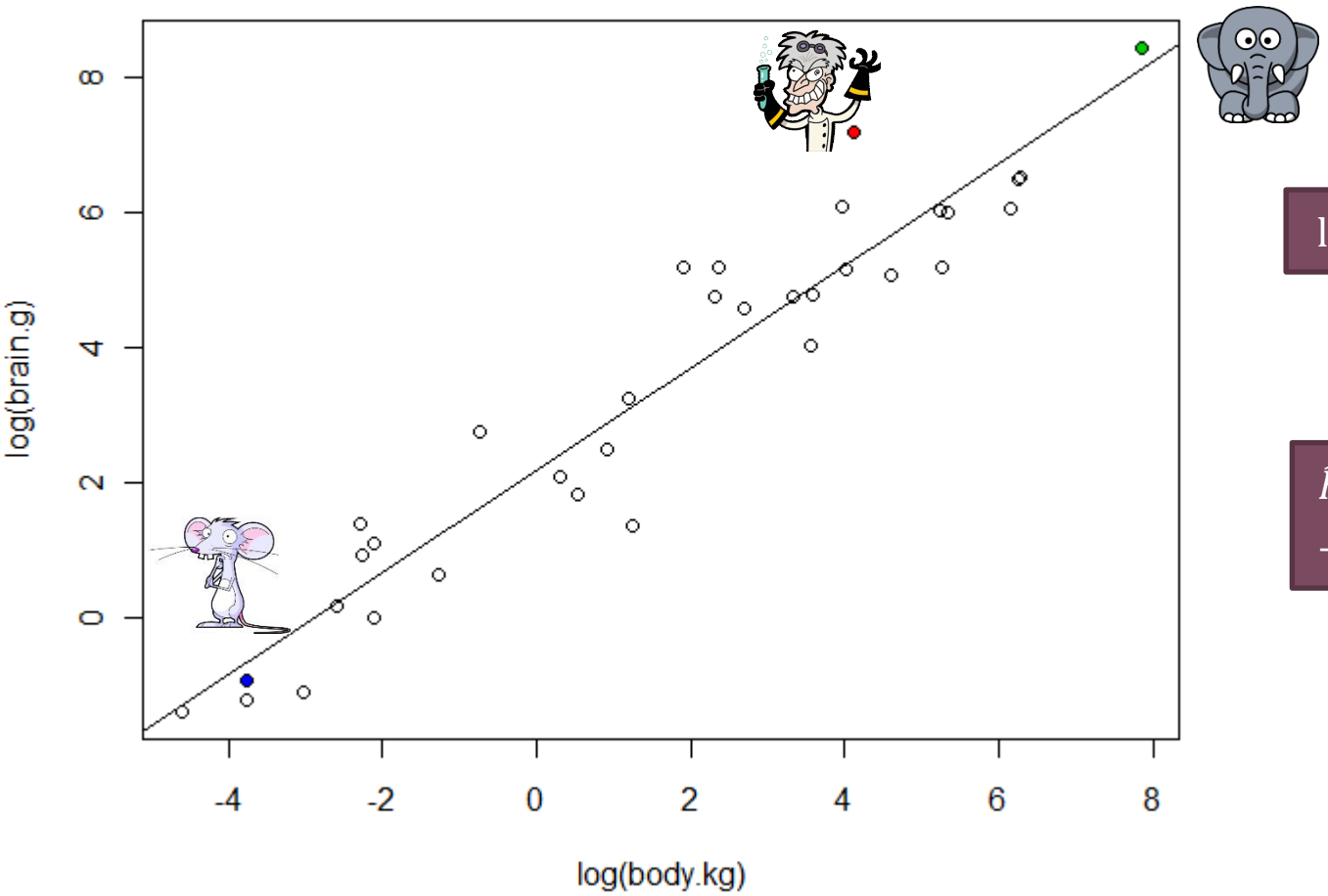
Bsp: log(Hirnmasse) vs. Körpermasse



Bsp: log(Hirnmasse) vs. log(Körpermasse)



Bsp: log(Hirnmasse) vs. log(Körpermasse)



$$\log(\widehat{H}) = \widehat{\beta_0} + \widehat{\beta_1} * \log(K)$$

$$\widehat{H} = \exp\left(\widehat{\beta_0} + \widehat{\beta_1} * \log(K)\right)$$
$$\rightarrow H = \hat{a} * K^{\hat{b}}$$

$\widehat{\beta_0} = 2.19$ (95%-VI: [1.89; 2.49]); $\widehat{\beta_1} = 0.75$ (95%-VI: [0.67; 0.83])

$\hat{a} = \exp\left(\widehat{\beta_0}\right) = 8.94$ (95%-VI: $[\exp(1.89)\,;\exp(2.49)]$ = [6.60; 12.02]
$\hat{b} = \widehat{\beta_1}$ (95%-VI: [0.67; 0.83])

Übersicht über nützliche Transformationen

- Linearer Zusammenahng:
  $y = a + bx$ (keine Transofrmation nötig)

- Exponentieller Zusammenhang:
  $$\log(y) = a + bx \rightarrow y = \exp(a) * \exp(bx)$$

- Polynomieller Zusammenhang:
  $$\log(y) = a + b \cdot \log(x) \rightarrow \mathrm{y} = \exp(\mathrm{a} + \mathrm{b} \cdot \log(x))$$
  $$\rightarrow y = \exp(a) \cdot x^b$$

Multiple Lineare Regression:
Wie hängt Energie von Eiweiss, Kohlehydraten und Fett ab?

100 ml enthalten ca. / contiennent env. / contengono ca.:	
Energie / énergie / energia	270 kJ (63 kcal)
Eiweiss / protéines / proteine	3.5 g
Kohlenhydrate / glucides / carboidrati	10 g
Fett / lipides / grassi	1.0 g
Calcium / calcium / calcio	120 mg
Vitamin B2	0.24 mg
Vitamin B12	0.18 µg

Multiple Lineare Regression: Interpretation

- Energie (E), Eiweiss (EW), Kohlehydrate (K), Fett (F)
- Modell:

$$E[kcal] = \beta_0 + \beta_1 EW[g] + \beta_2 K[g] + \beta_3 F[g] + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Was bedeutet es, wenn in diesem Modell $\beta_3 = 8$?

A: Wenn ein Nahrungsmittel ein Gramm mehr Fett als ein anderes hat, enthält es im Schnitt 8 kcal mehr Energie.

B: Wenn ein Nahrungsmittel ein Gramm mehr Fett als ein anderes hat und gleich viel Eiweiss und Kohlehydrate enthält, enthält es im Schnitt 8 kcal mehr Energie.

Einfache oder Multiple Regression

(Gilt für alle GLMs; hier am Bsp der linearen Regression)

- Einfache Regression:
  "Totaler Effekt"
  y ~ x → "Wenn sich x um eine Einheit erhöht, erhöht sich y um β_1"
- Multiple Regression
  "Bereinigter Effekt"
  y ~ x1 + x2 → "Wenn sich x1 um eine Einheit erhöht und x2 konstant bleibt, erhöht sich y um β_1.

- Kein "richtig" oder "falsch"; eher zwei verschiedene Sichtweisen auf das gleiche Problem

Vorteil von Multipler Regression

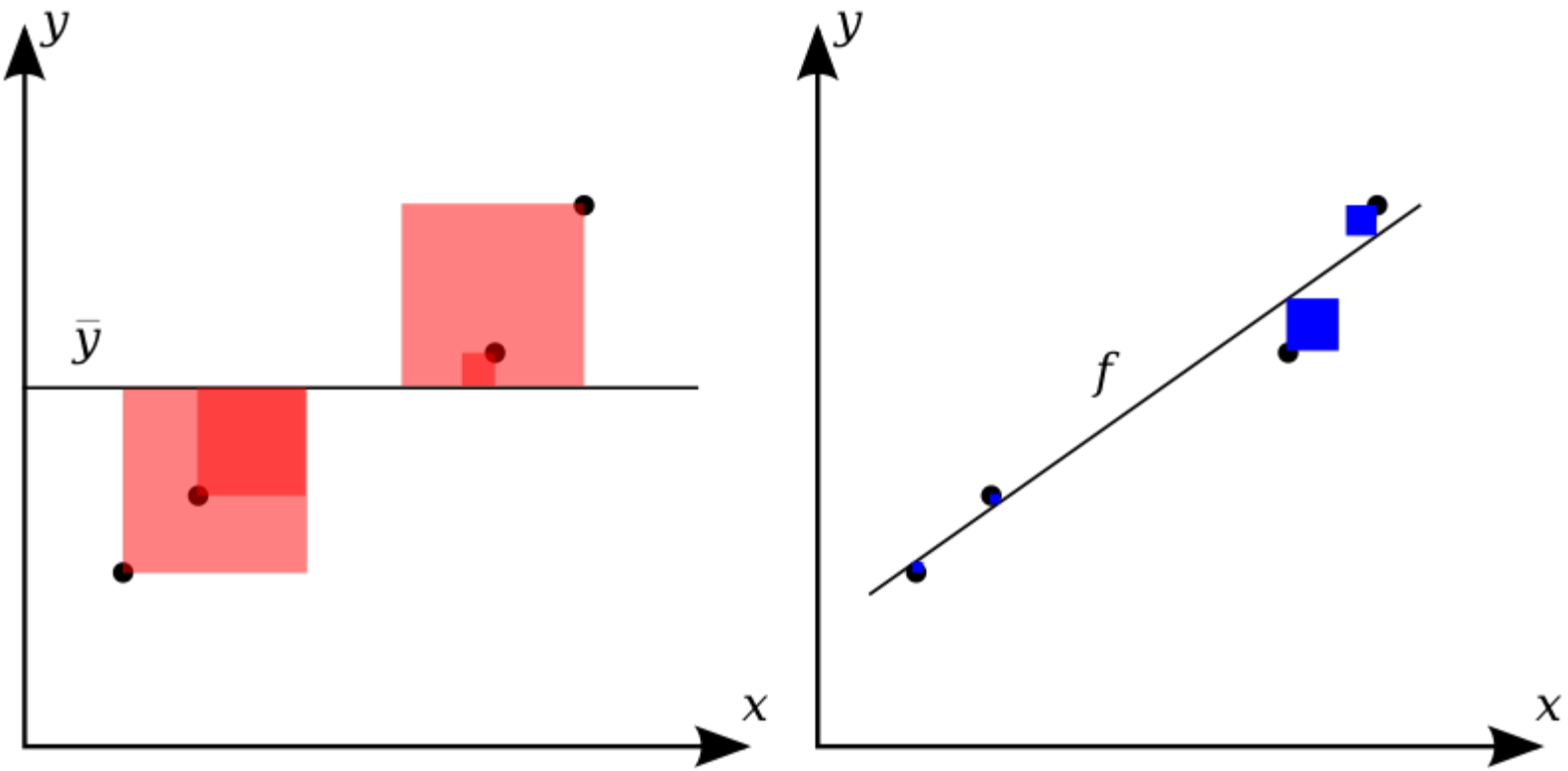
- Andere Einflüsse werden ausgeschaltet

Bsp: Diskriminierung

- Einfache Regression:
  Zulassung ~ Geschlecht
- Multiple Regression:
  Zulassung ~ Geschlecht + Job

  Berühmtes Beispiel: Simpson’s Paradox

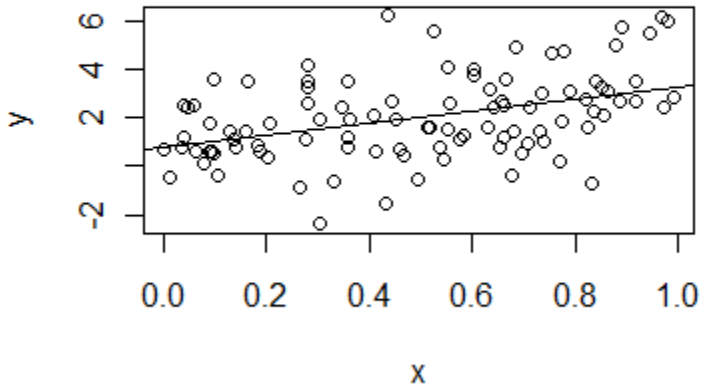
Bestimmtheitsmass R^2



$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

R^2: “Wie nahe liegen Punkte auf der Geraden?”
(im Vergleich zur ursprünglichen Streuung der y-Werte)

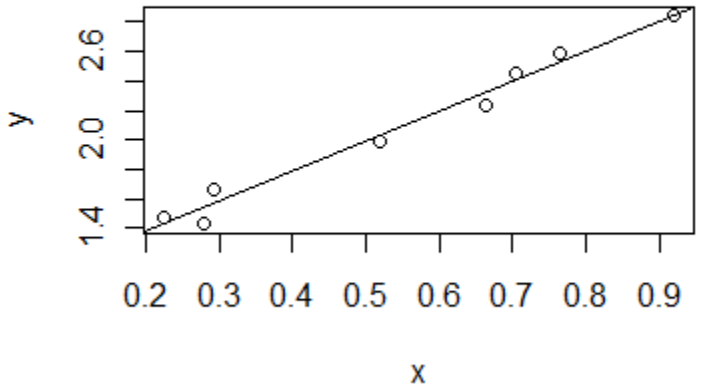
Signifikanz vs. Relevanz



Signifikant, aber evtl. nicht relevant

$H_0: \beta_1 = 0 \rightarrow p = 0.00008$

$R^2 = 0.15$ *oder* $|\widehat{\beta_1}|$ *sehr "klein"*



Signifikant und wohl auch relevant (?)

$H_0: \beta_1 = 0 \rightarrow p = 0.00002$

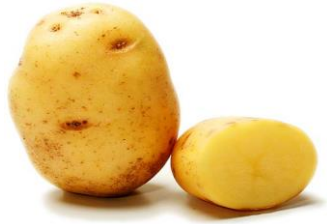
$R^2 = 0.98$ *oder* $|\widehat{\beta_1}|$ *"gross"*

Statistik: Entscheidet **Signifikanz**

Wissenschaft: Entscheidet **Relevanz**

(je nach Fach: Unterschiedliche Werte von R^2 gefordert)

Energiegehalt von 20 Lebensmitteln



Daten (pro 100 g)

Name	kcal	gE	gK	gF
Butter	729	0.5	0.5	82.0
Laetta	370	0.0	4.0	39.0
Mozzarella	257	19.0	1.0	20.0
Cantadou	323	7.0	3.0	32.0
Lc1	105	3.5	15.5	3.0
Emmi	130	4.0	16.0	5.5
Quark	65	12.0	2.5	0.1
LightKaese	249	29.0	2.0	14.0
Banane	93	1.0	22.0	0.0
Zucchini	19	1.6	3.3	0.4
Tomate	17	1.0	2.6	0.2
Kartoffel	86	2.0	19.0	0.1
Brot	282	11.0	53.0	1.5
CremeSchnitte	311	4.5	48.0	11.0
Pizza	227	13.0	31.0	5.0
Schoko	569	7.0	46.0	40.0
Chips	517	7.0	51.0	32.0
Spaghetti	350	12.0	72.2	1.5
Reis	358	5.0	83.0	0.5
Stocki	320	9.0	70.0	1.0

Multiple Lineare Regression

```
lm(formula = kcal ~ gE + gK + gF, data = dat)
```

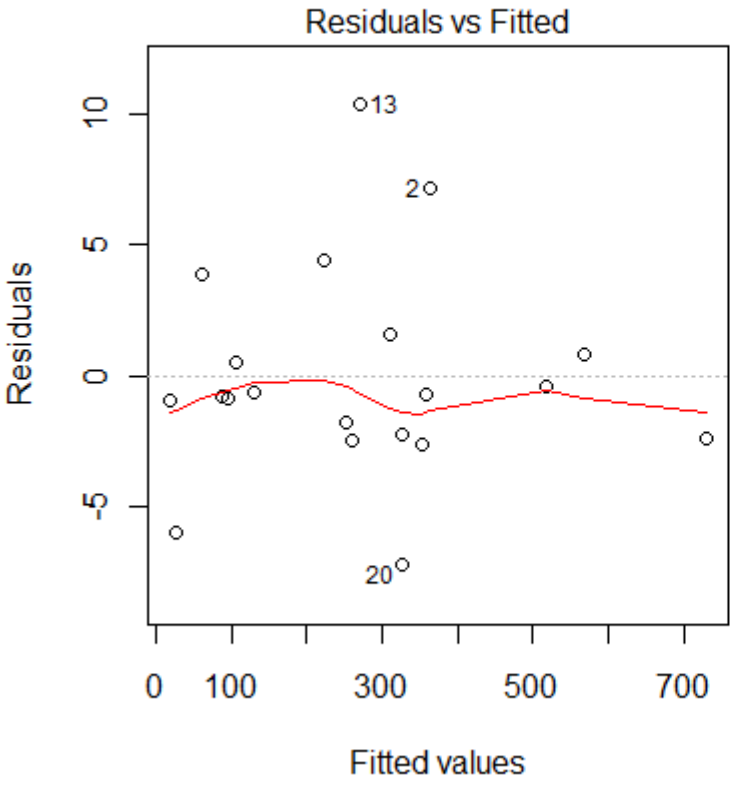
Ein Lebensmittel, das ein Gramm mehr Fett
aber gleich viel Eiweis und Kohlenhydrate enthält,
enthält im Schnitt 8.8 kcal (95%-VI: [7.8; 9.8]) mehr Energie.

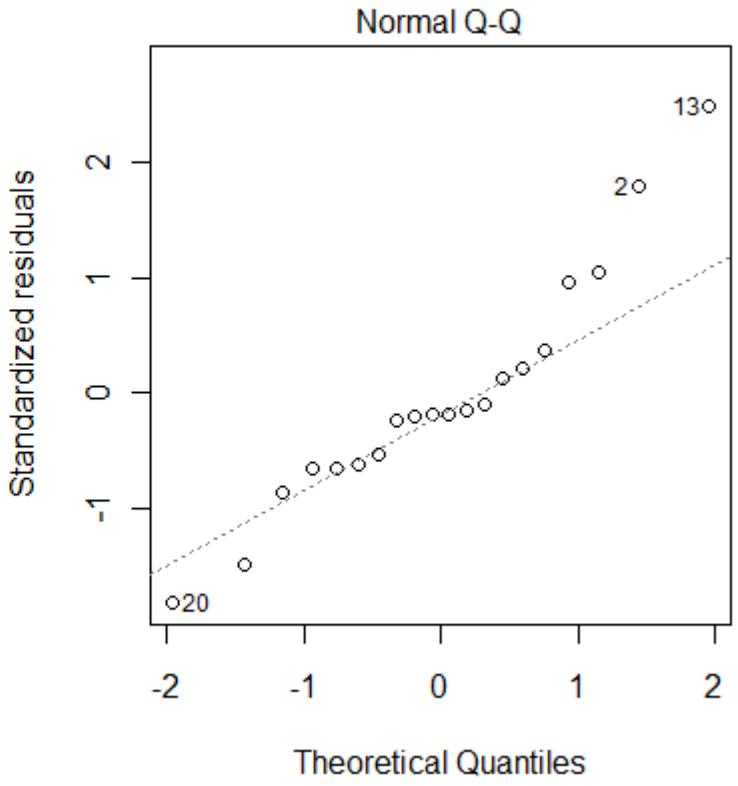
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.70736    2.10299   0.812    0.429
gE           4.04087    0.14280  28.298  4.3e-15
gK           4.00415    0.03838 104.330  < 2e-16
gF           8.84937    0.05025 176.115  < 2e-16

Multiple R-squared: 0.9995
```

Die Punkte liegen äusserst genau
auf der geschätzten Geraden.

(verglichen mit der ursprünglichen Streuung der Energiewerte)

Residuenanalyse





Im Allgemeinen sind die Modellannahmen erfüllt. Allerdings fallen Beobachtungen 2 (Lätta) und 13 (Brot) etwas aus dem Rahmen (5-10 kcal mehr als vorhergesagt).

Ursache und Wirkung (Kausalität)

Randomisiertes, kontrolliertes Experiment

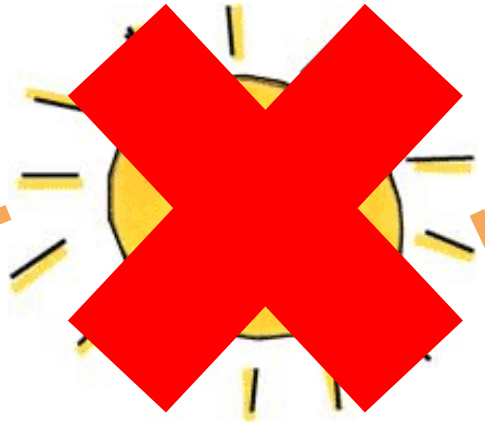
Ursache und Wirkung

Opfer durch Ertrinken

?

Eisverkauf

Ursache und Wirkung



Opfer durch Ertrinken

Eisverkauf

Kausaler Zusammenhang

≠

Korrelation

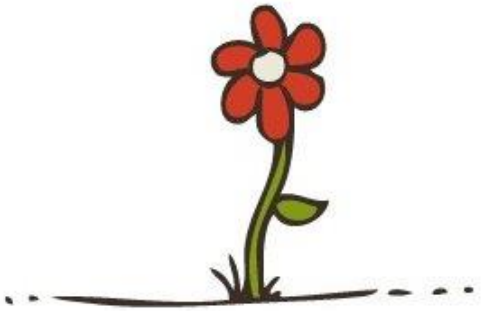
Wie findet man Kausalzusammenhänge?

Randomisiertes, kontrolliertes Experiment

Kausaleffekt finden

Experiment

?

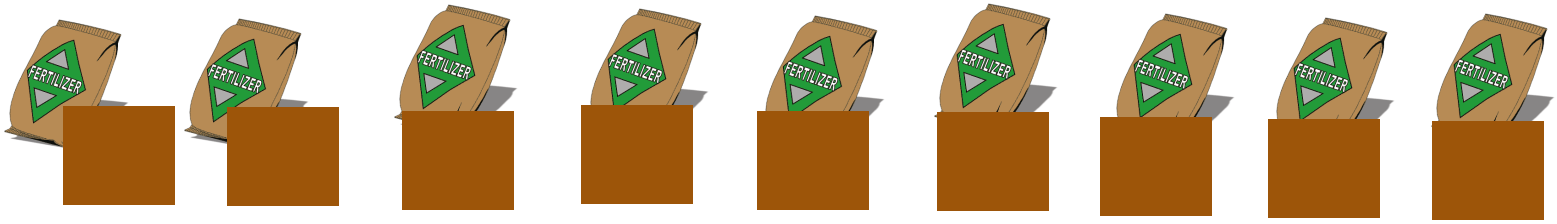


Kausaleffekt finden

Experiment

Kausaleffekt finden

Experiment



Kausaleffekt finden

Experiment

Kausaleffekt finden

Experiment

Dünger besser als kein Dünger?

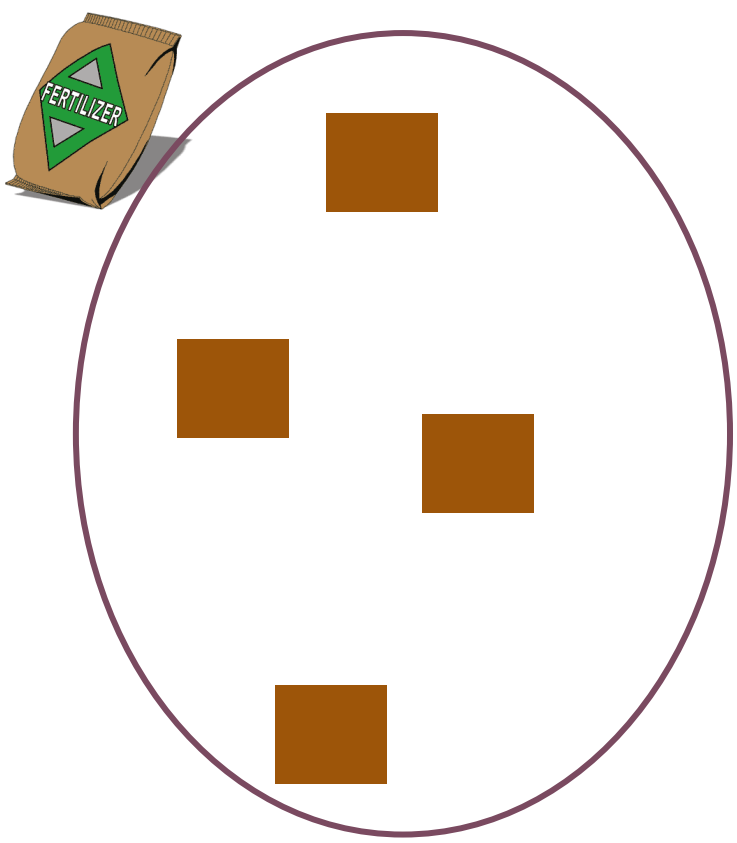
Keine Ahnung!

Wie viele rote Blumen hätte es ohne Dünger gegeben?

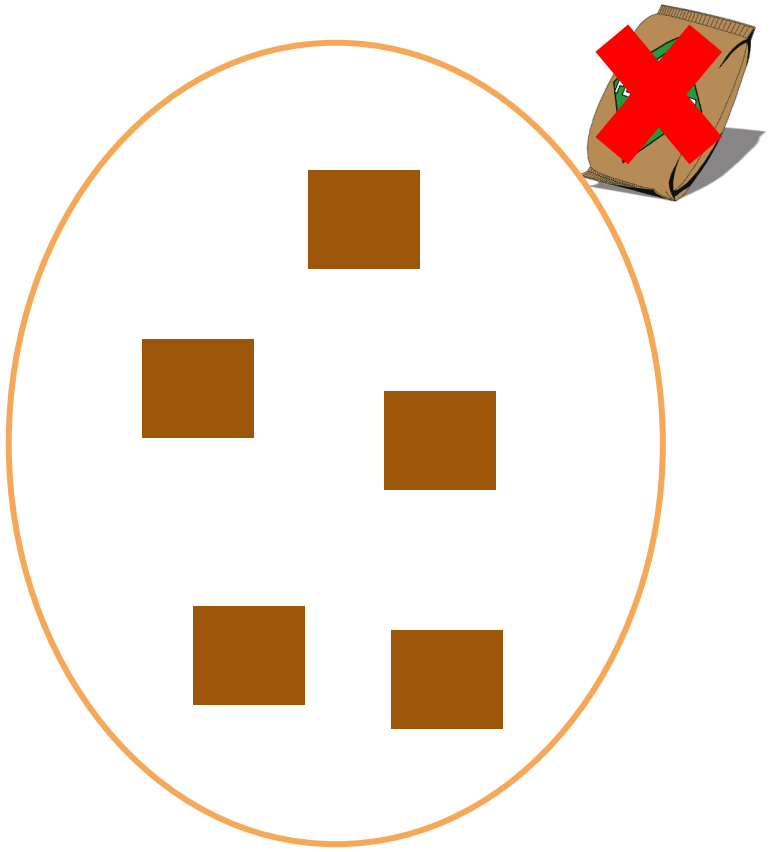
Brauchen eine Kontrollgruppe

Kausaleffekt finden

Experiment



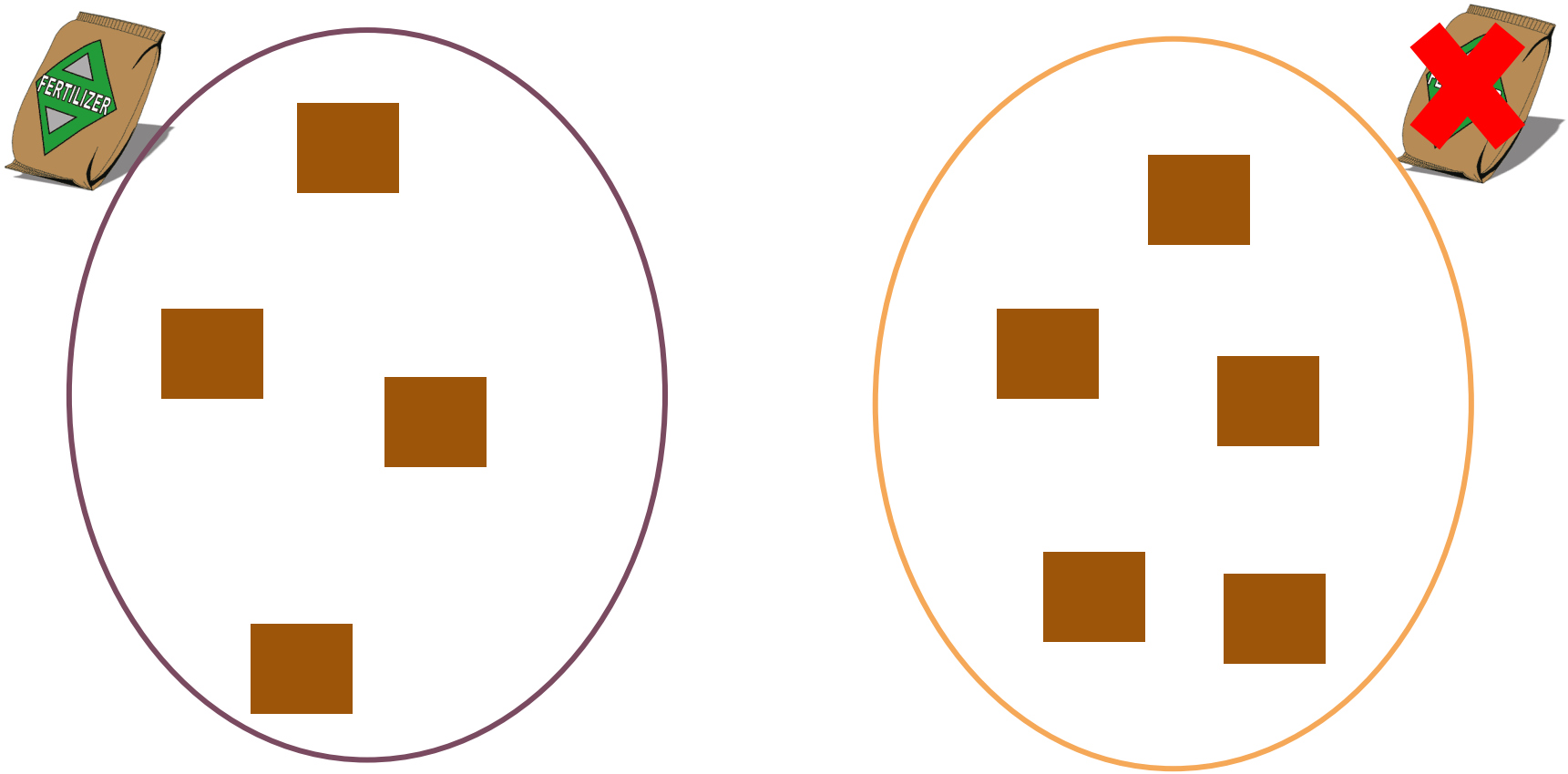
Behandlungsgruppe



Kontrollgruppe

Kausaleffekt finden

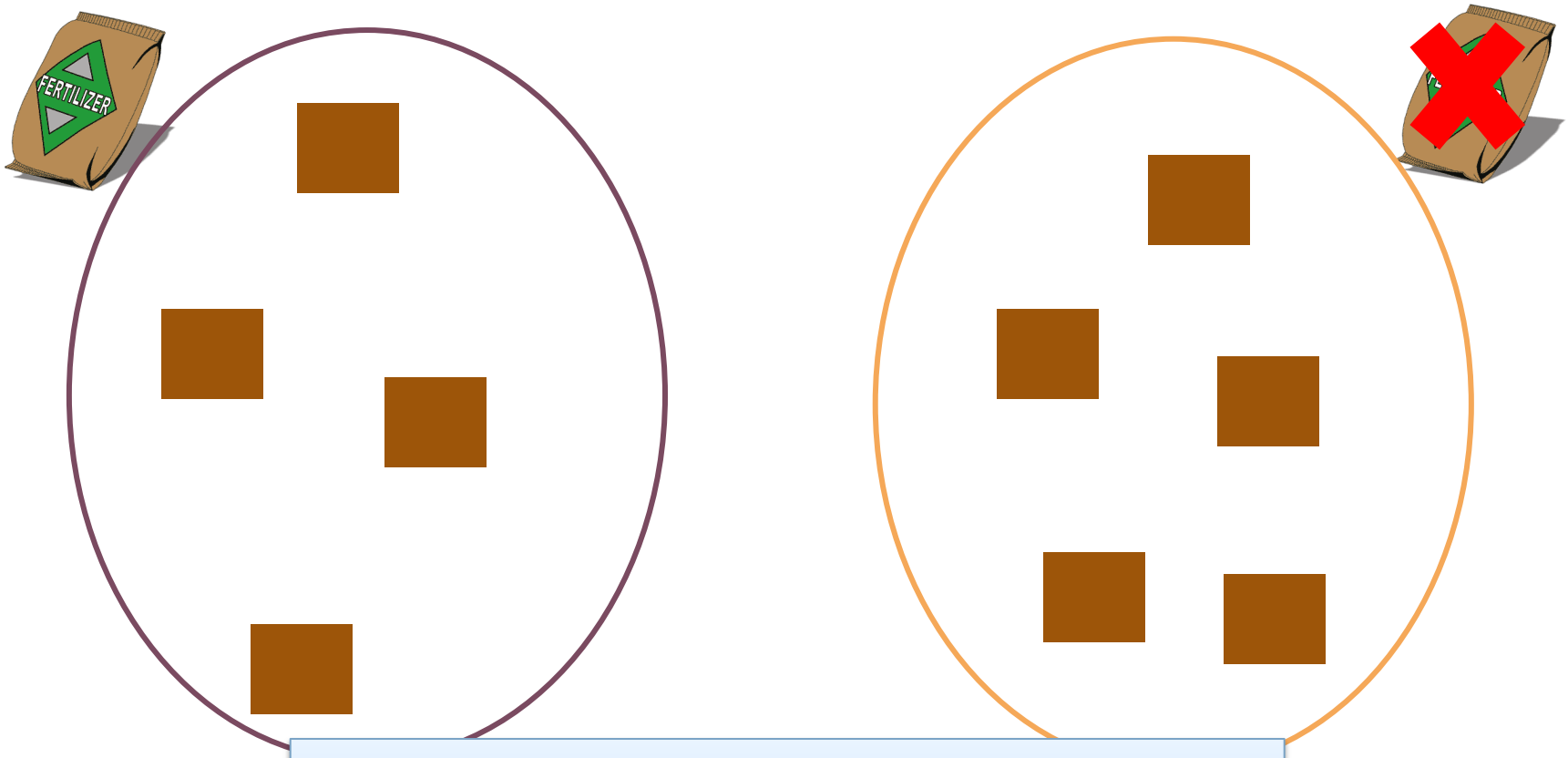
Experiment



Zwei Gruppen von Feldern **in allem gleich**
(Bodenqualität, Wasser, Sonnenlicht, …)

Kausaleffekt finden

Experiment



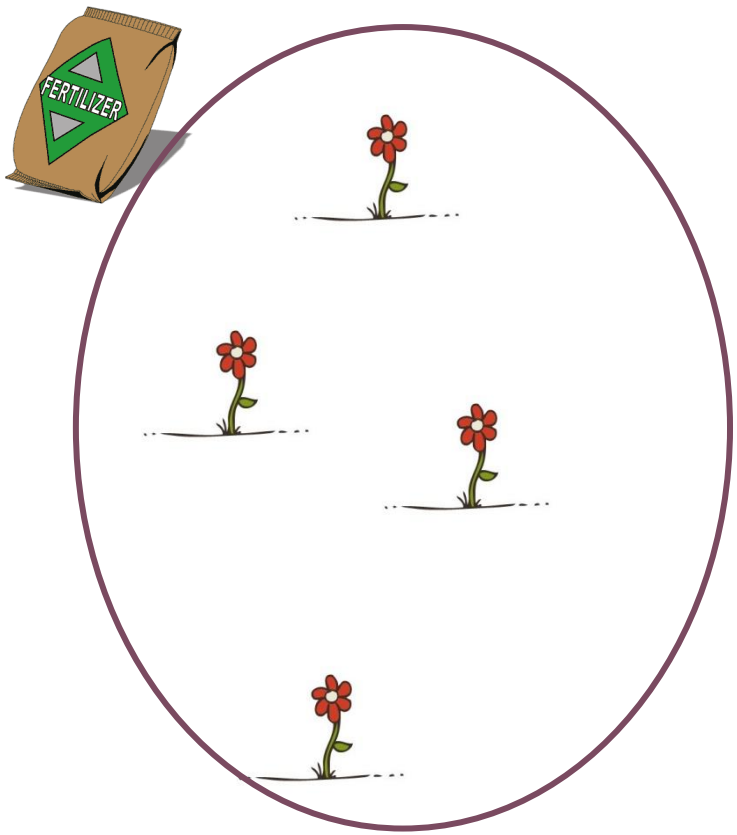
Zwei Gruppen von Feldern **in allem gleich**:

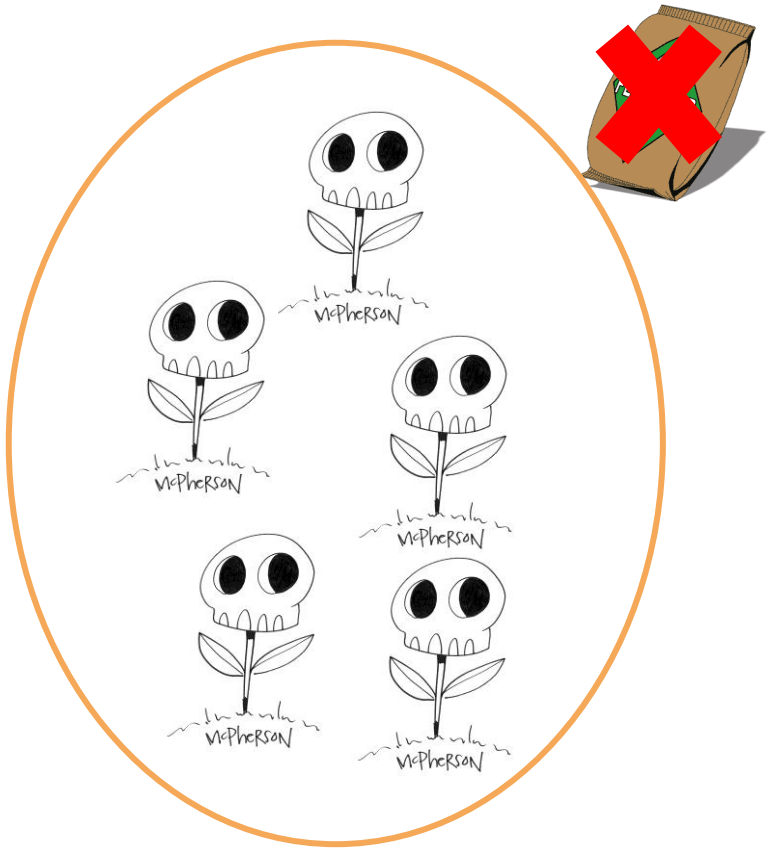
(Bodenqualität, Wasser, Sonnenlicht, …)

Praxis: Zufällige Zuordnung der Felder

Kausaleffekt finden

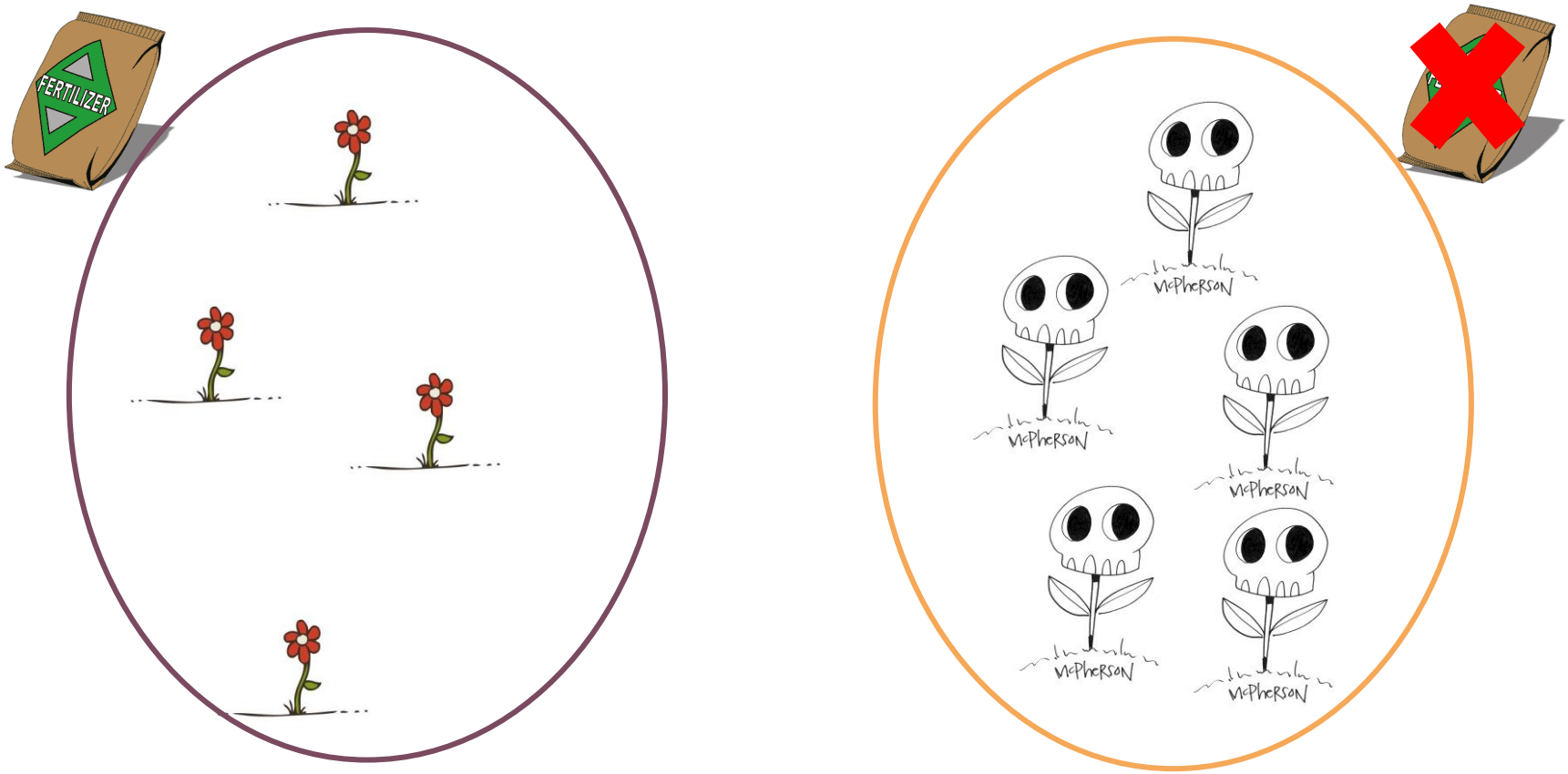
Experiment





Kausaleffekt finden

Experiment



Ergebnis ist wegen Dünger,
weil alles andere gleich war

Manchmal sind randomisierte, kontrollierte Experimente nicht machbar

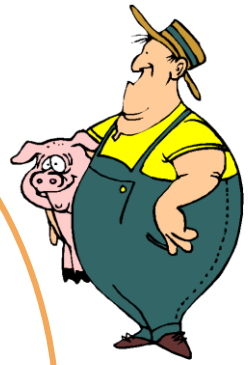
- zu teuer, zu zeitaufwändig (Genexpressionsdaten)
- unethisch, nicht machbar (HIV Behandlung, Rauchen)

Falls Experiment nicht machbar…

Beobachtungsstudie

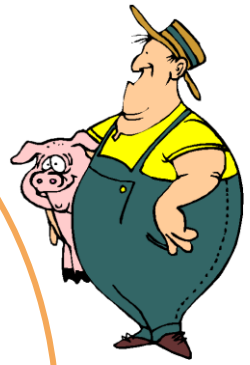
… mache Beobachtungen.

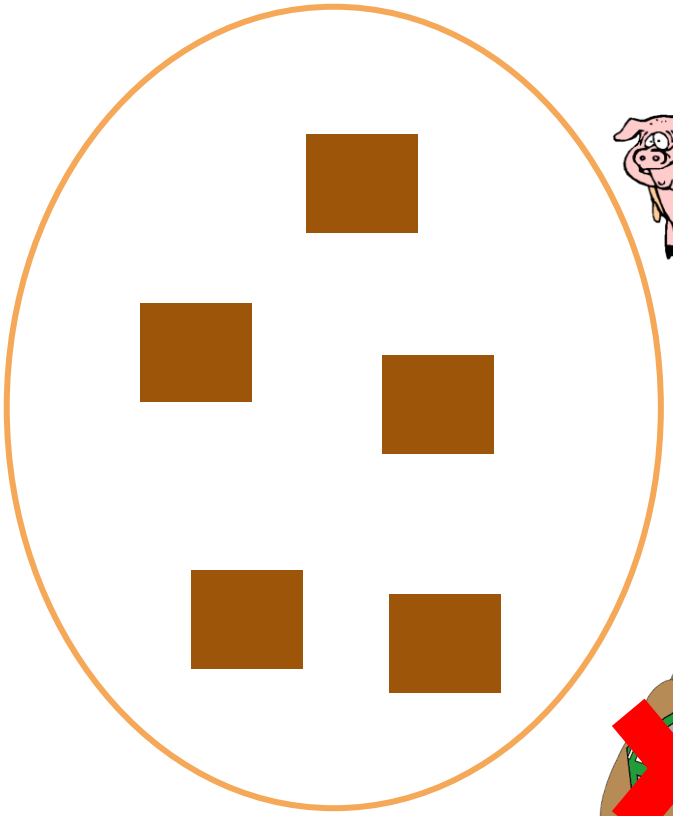
Beobachtungsstudie



… mache Beobachtungen.

Beobachtungsstudie

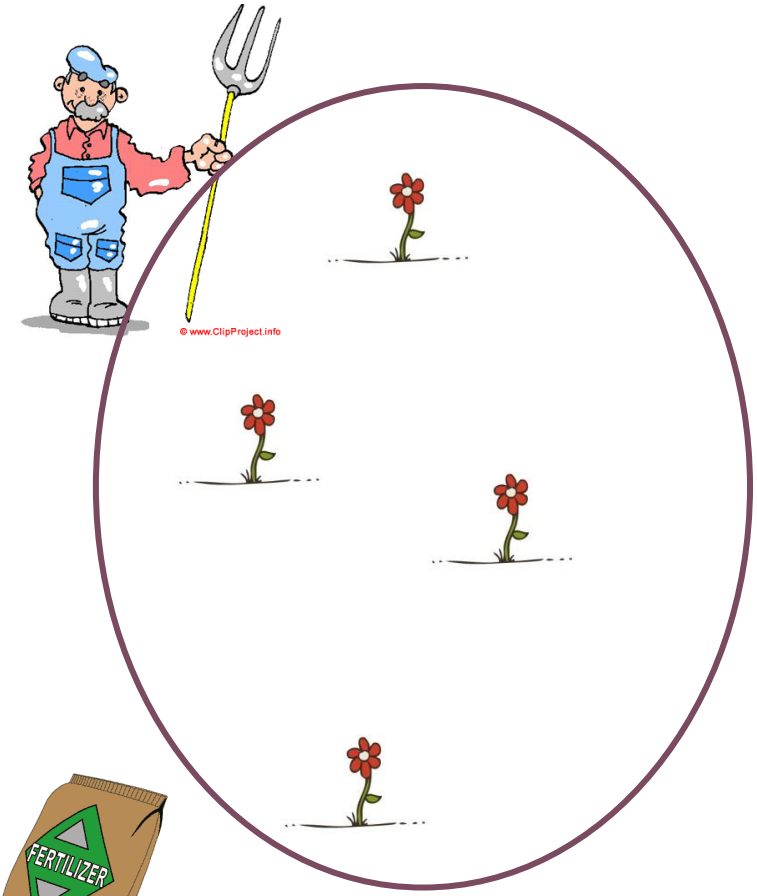


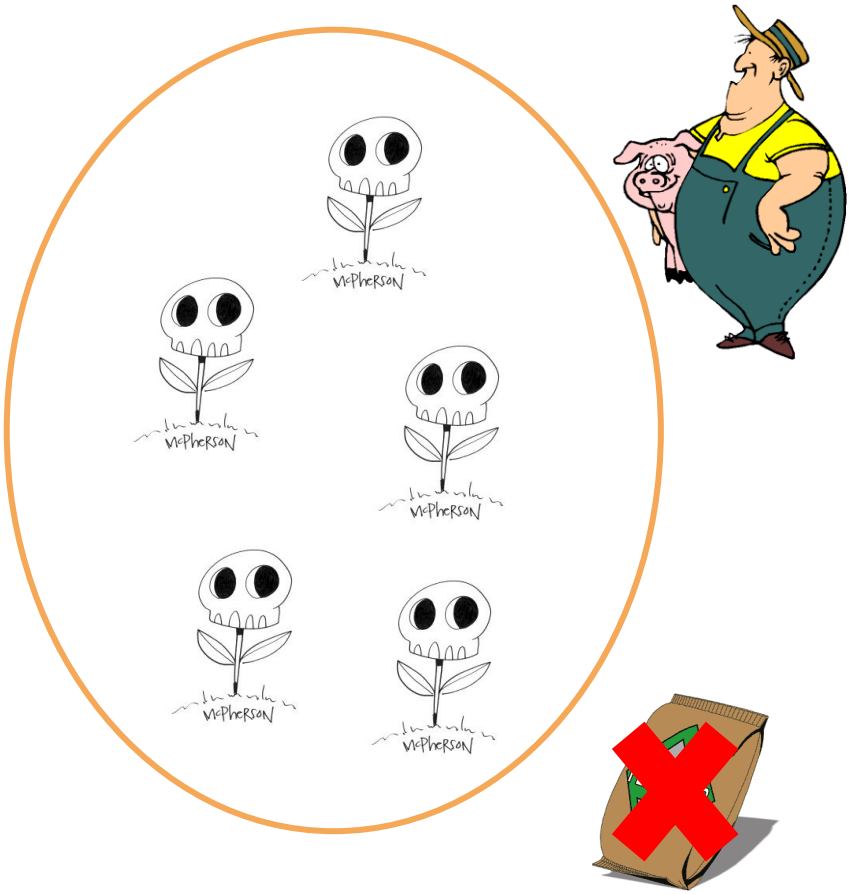


Es ist nicht garantiert, dass beide Gruppen in allen Aspekten gleich sind

… mache Beobachtungen.

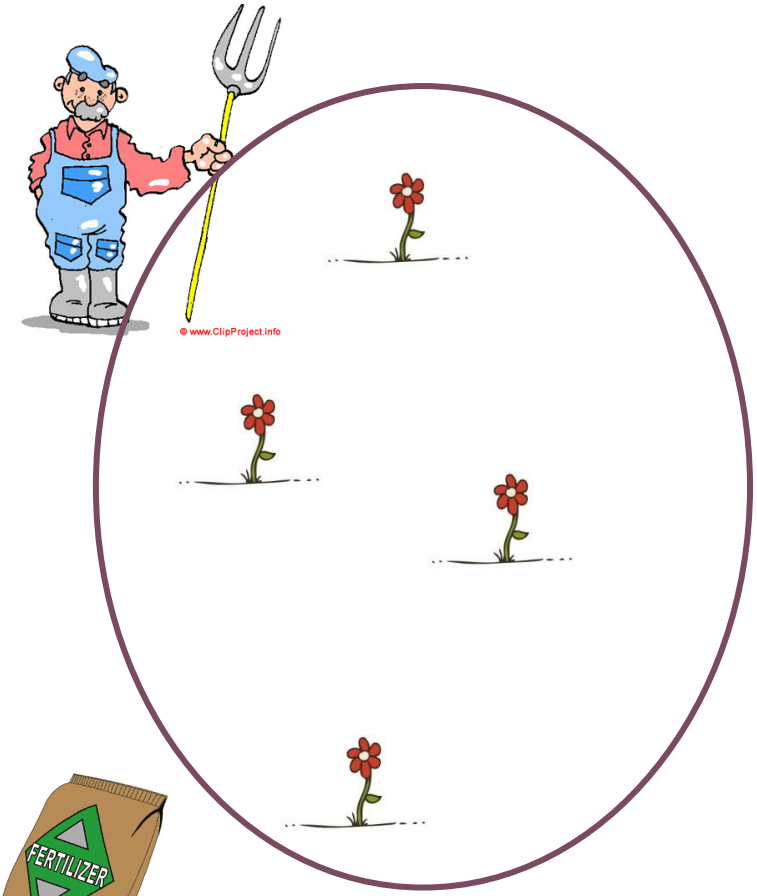
Beobachtungsstudie

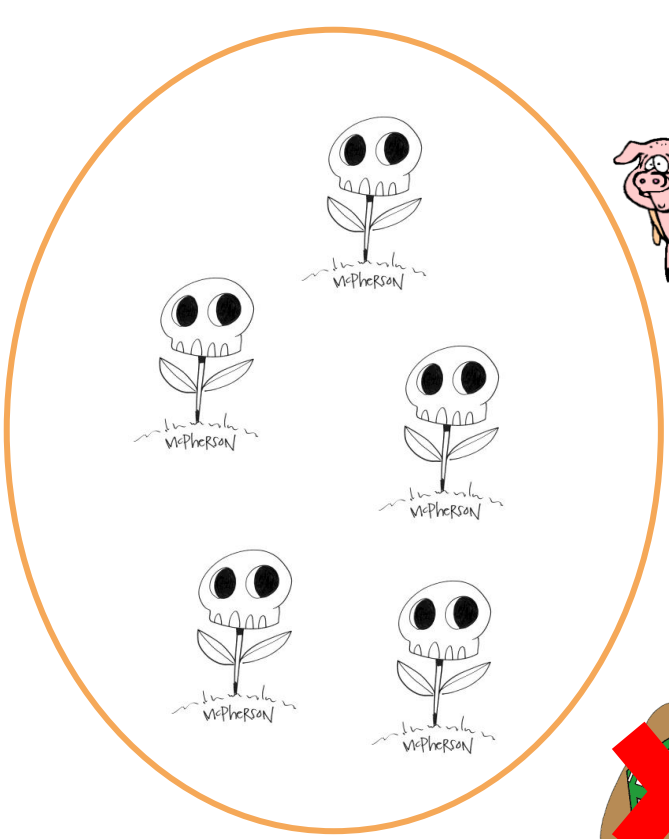




… mache Beobachtungen.

Beobachtungsstudie



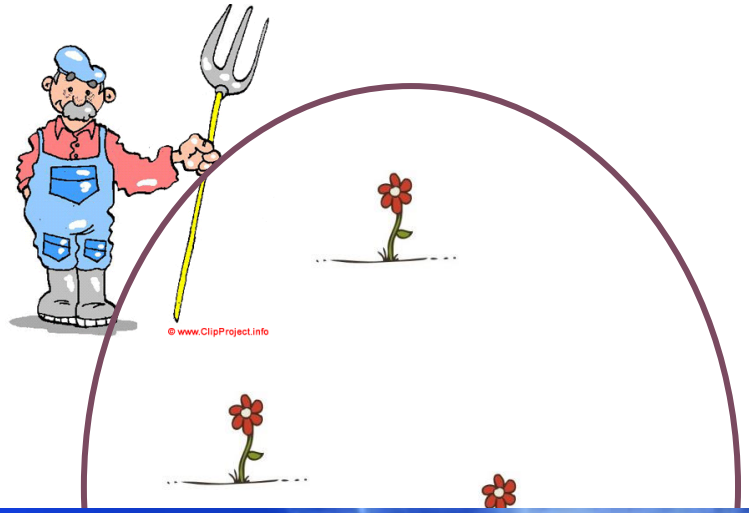


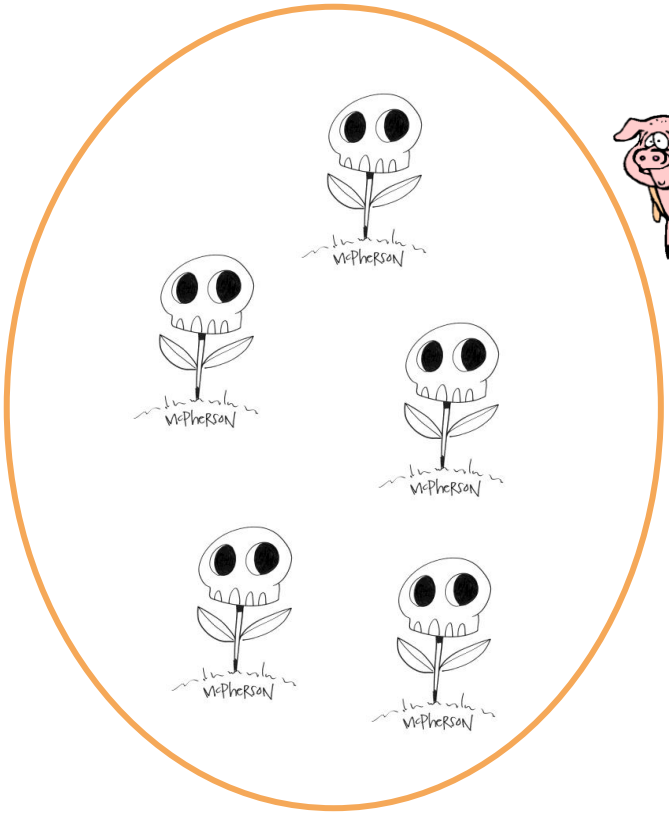
Ist das Ergebnis wegen Dünger?

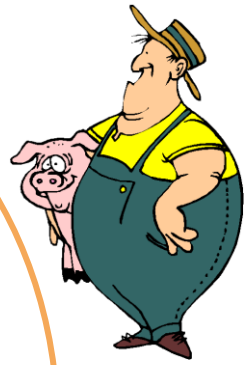
Keine Ahnung!

… mache Beobachtungen.

Beobachtungsstudie

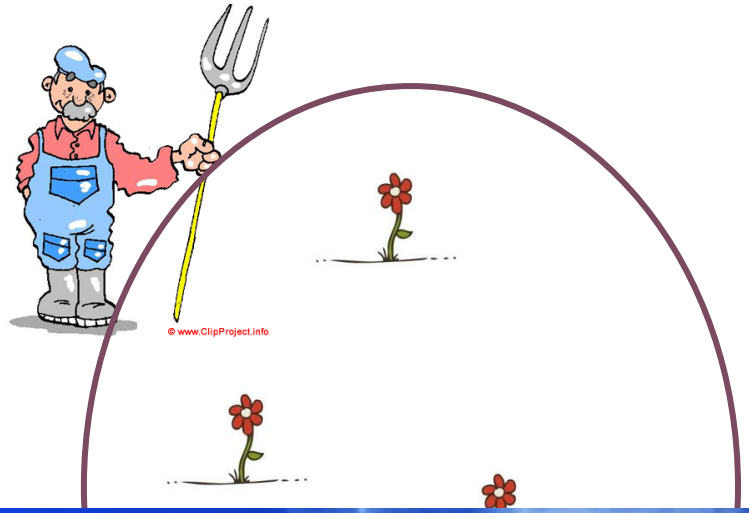


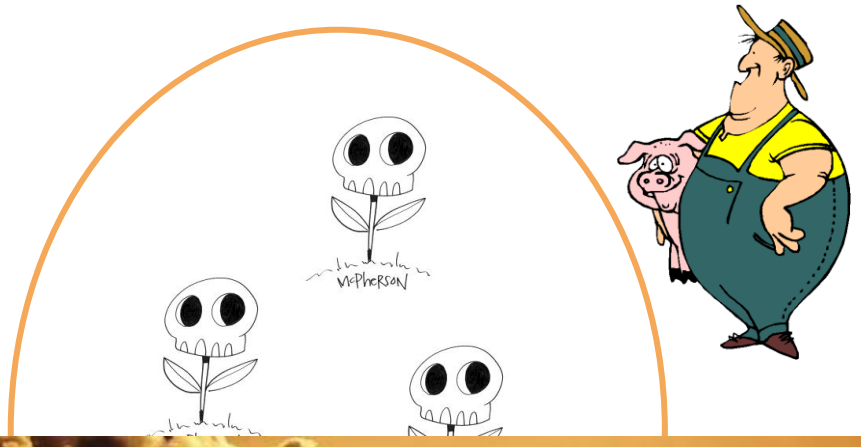




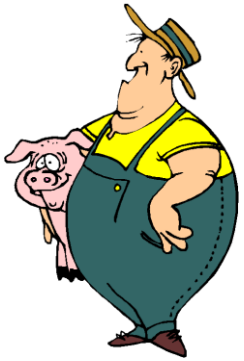
… mache Beobachtungen.

Beobachtungsstudie

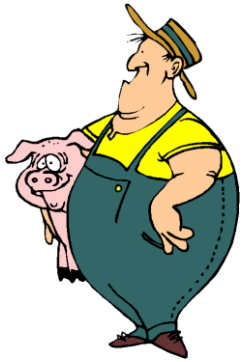




Besser: Vergleiche Bauern, die in möglichst vielen Punkten übereinstimmen.

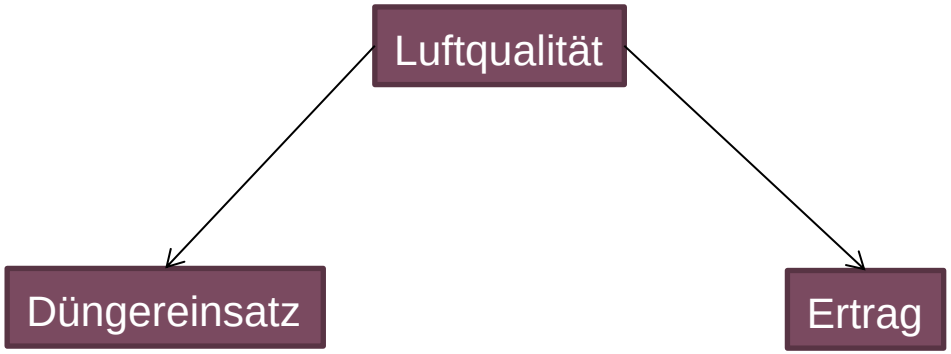


Aber: Wir können nie sicher sein, dass es nicht doch noch irgendwelche relevanten Unterschiede zwischen den Gruppen gibt.



Zusammenfassung

- **Randomisierte, kontrollierte Experiment:** Beste Möglichkeit, Daten zu sammeln (“Goldstandard”)
- **Beobachtungsstudie:** Man muss skeptisch sein – kam der Effekt (viele schöne Blumen) durch die Behandlung (Dünger), oder durch einen Umstand, der in beiden Gruppen unterschiedlich war (Luftqualität)?



Beispiel für rand. kontr. Exp.: Das grösste medizinische Experiment aller Zeiten

- 1954: Potenzieller Impfstoff gegen Polio
- Randomisiertes, kontrolliertes Experiment bei Kindern in 1. bis 3. Schulklasse
  [http://www.stat.luc.edu/StatisticsfortheSciences/MeierPolio.htm]

	Anzahl Kinder	Polio bekommen (pro 100.000 Kinder)
Behandlung	200.000	28
Kontrolle	200.000	71
Verweigert	350.000	46

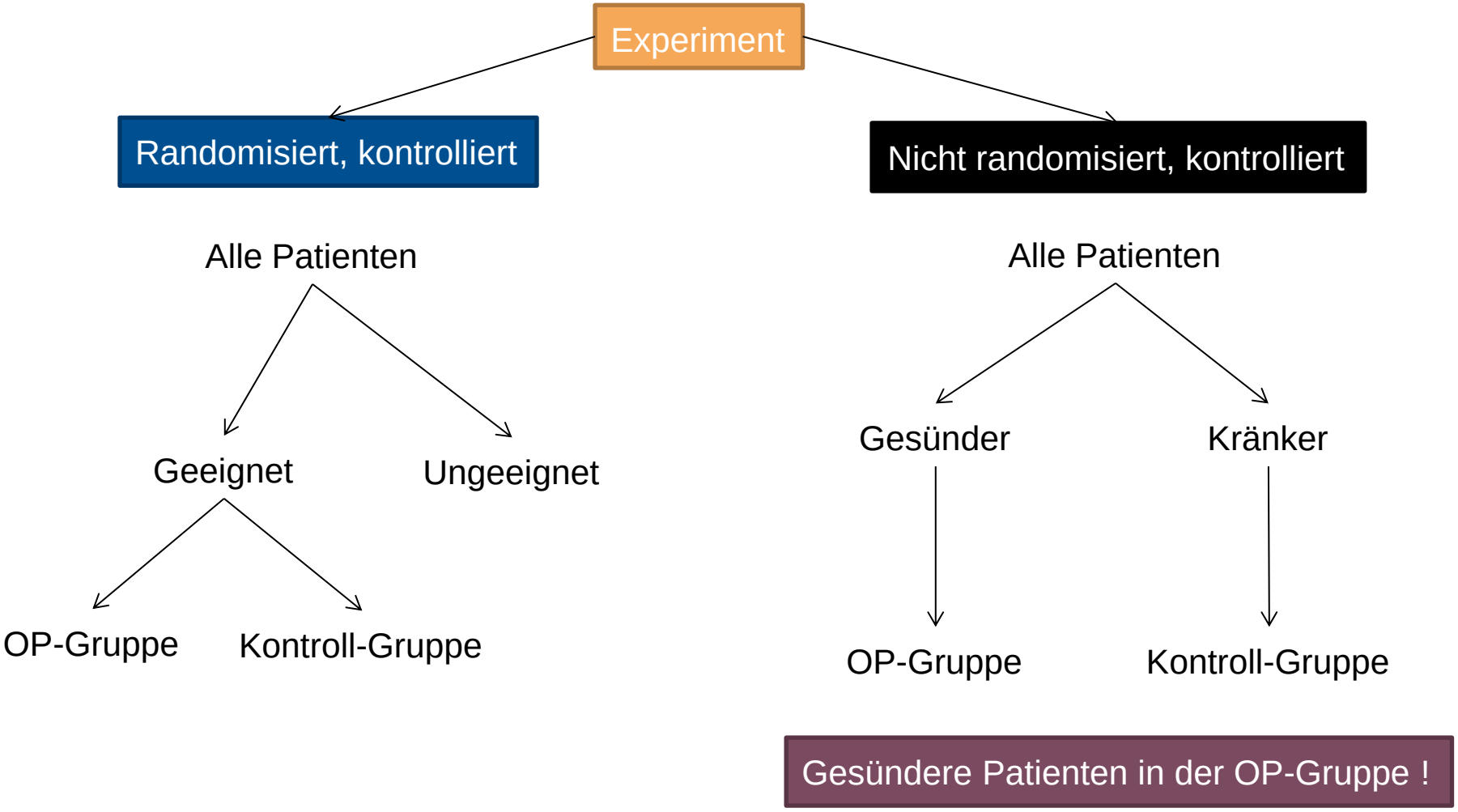
- Kontrollgruppe hatte mehr Polio-Fälle als Behandlungsgruppe: Könnte das Zufall sein?
- KAUM: p-Wert = 0.000000001 (diesen Test haben wir nicht besprochen)

Methodenvergleich: Leberzirrhose – Shunt (Blutumleitung)

- Bringt die riskante Operation einen Vorteil?
- 51 klinische Studien untersucht: “Bringt die Operation einen Vorteil?”

	Ja, sehr	Etwas	Nein
Keine Kontrollgruppe	24	7	1
Kontrollgruppe, nicht randomisiert	10	3	3
Kontrollgruppe, randomisiert	0	1	3

Problem: Gesündere Patienten werden eher operiert



Gesündere Patienten in der OP-Gruppe !

Goldstandard in der Medizin:

Randomisiertes, kontrolliertes, doppelblindes Experiment mit Placebo

- Placebo: Medikamentenattrappe ohne Wirkstoff

  Placebo hat einen starken Effekt!

  (J.A. Turner et.al., “The importance of placebo effects in pain treatment and research”, Journal of the American Medical Association, Vol. 271 (1994), pp. 1609 – 14)

- Doppelblind: Weder Patient noch Arzt weiss, ob er das Placebo oder das

  wirkliche Medikament erhält / verabreicht.
  (Nur Leiter der Studie weiss das.)

Zusammenfassung

- Jede “Behandlung” muss mit einer “Kontrolle” verglichen werden
  (bei Menschen am besten “doppelblind” mit einem Placebo)

- Was nicht kontrolliert werden kann, soll randomisiert werden

- Korrelation ≠ Kausalzusammenhang