IKON

Prof. Dr. Frank Steinicke
Human-Computer Interaction
Fachbereich Informatik
Universität Hamburg

Mensch-Computer-Interaktion

Evaluierung

Prof. Dr. Frank Steinicke
Human-Computer Interaction, Universität Hamburg

Evaluierung

- **Evaluierung** sammelt Informationen über Performanz, Problemen und Erfahrungen der Benutzer mit interaktiven Systemen
- **Evaluierung** ist wesentlicher Bestandteil der Mensch-Computer-Interaktion

Agenda

- DECIDE Framework
- Arten der Evaluierung
- Feld- & Laborstudien
- Usability-Tests und Benutzerstudien
- Empirische Methoden
- Datenanalyse

Mensch-Computer-Interaktion

Evaluierung

DECIDE-Framework

Evaluierung

- Vor Evaluierung müssen Reihe von Fragen beantwortet werden
  - Wer und Warum?
  - Was und Wie?
  - Wann und Wo?

DECIDE

- **DECIDE** bietet Framework, welches hilft Benutzerstudien zu planen und durchzuführen
- **DECIDE** dient als Checkliste für Evaluierungen

DECIDE

Determine Goals

- Was sind Ziele der Evaluation?
  - Wer will Evaluierung?
  - Warum soll Evaluierung durchgeführt werden?
  - Was soll herausgefunden werden?

DECIDE

Explore the Question

- Wie lautet Frage, die Evaluierung beantworten soll?
  - große Fragen müssen häufig in kleinere Fragen und damit kleinere Evaluierungen zerlegt werden

DECIDE

Choose Evaluation Method

- Welche Methode soll gewählt werden?
  - Auswahl hängt von Zielen und Fragestellung ab
  - häufig Kombination mehrere Methoden (vgl. Triangulation)

DECIDE

Identify Practical Issues

- Welche Aspekte beeinflussen reibungslosen Ablauf der Evaluierung?
  - Laborausstattung, Zeit- und Budgeteinschränkungen, Expertise
- Pilot-Studien sind essentiell für korrekten Ablauf von Studien

DECIDE

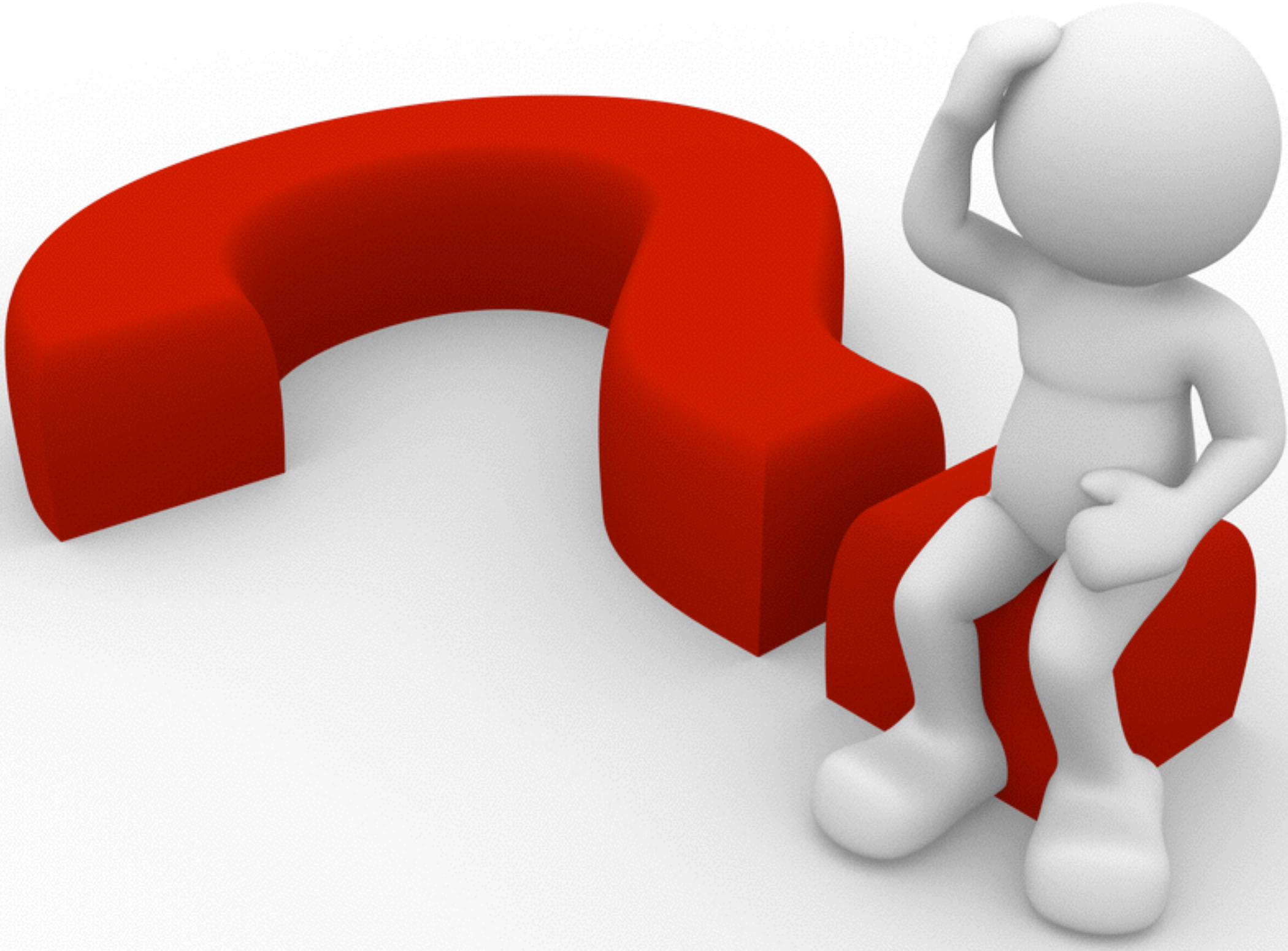
Decide on Ethical Issues

- Wie wird mit ethischen Fragestellungen umgegangen?
  - Zertifizierung von Ethikkommission notwendig, z.B. *Institutional Review Board (IRB)*?

DECIDE

Evaluate

- Wie wird Evaluierung durchgeführt und gesammelte Daten analysiert, interpretiert und präsentiert?



Mensch-Computer-Interaktion

Evaluierung

Arten der Evaluierung

Evaluierung

Warum?

- **Exploration**
  - (frühe) qualitative und informelle Erkundung von Benutzeranforderungen
- **Beurteilung**
  - Einschätzung des Stands der UI-Entwicklung
  - Beurteilung von Designs

Evaluierung

Warum?

- **Vergleich**
  - Gegenüberstellung von Alternativen, z.B. alt gegen neu, unser gegen deren etc.
- **Validierung**
  - Überprüfung gegen Ende der Entwicklungen, um Behauptungen/ Hypothesen zu überprüfen

Evaluierung

Wann?

- Vor Entwicklung
  - Anforderungsanalyse
- Während Entwicklung
  - verfeinern der Anforderungsanalyse
  - Testen verschiedener Konzepte
- Nach Entwicklung
  - Validierung oder Vergleich

formativ

summativ

Evaluierung

Wo?

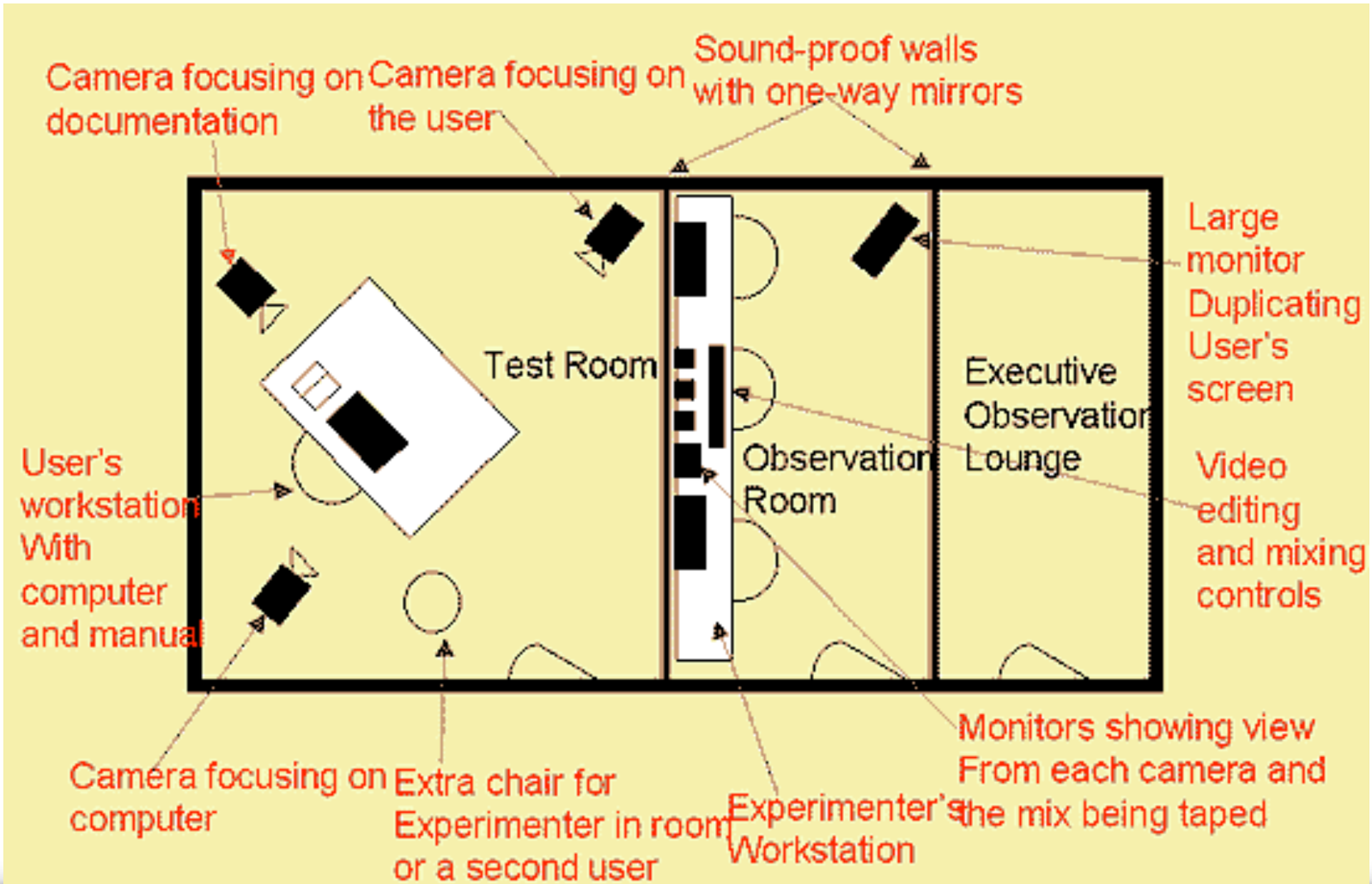
- **Feldstudie** ist systematische wissenschaftliche Beobachtung unter natürlichen Bedingungen
- **Laborstudie** ist wissenschaftliche Methode, um mit Hilfe von Laborexperimenten bestimmte Arbeitshypothese zu testen

Feldstudie

Beispiel: Fahrkartenautomaten

Laborstudie

Bsp: Sun Microsystem Lab



Laborstudie

Beispiel: Usability-Lab

Drive-in Usability-Lab, 2012

Usability-Lab

Bsp.: Usability Living Lab

Evaluierung

Was?

- Art der Ergebnisse unterscheiden sich in
  - schwer quantifizierbare Beschreibungen, z.B. Lösungsansätze
  - Benutzeraussagen, z.B. Präferenzen, Selbsteinschätzungen
  - quantitativ messbare Ergebnisse
  - technisch aufzeichen- & auswertbar

subjektiv

objektiv

"When, among a set of observations, ***any single observation*** *is a* ***number*** *that represents an amount or a* ***count****, then the data is* ***quantitative****."*

Witte & Witte: Statistics, 2009

Evaluierung
Quantitative

- **Quantitative Evaluierung** liefert *quantitive*, d.h. sinnvoll in Zahlen ausdrückbare, Ergebnisse
  - Beispiele: Messwerte wie Fehler, Anzahl Tastendrücke, Ausführungszeiten …

"When, among a set of observations, ***any single observation*** *is a* ***word****, or a* ***sentence****, or a* ***description****, or a* ***code*** *that* ***represents*** *a* ***category*** *then the data is* ***qualitative****."*

Witte & Witte: Statistics, 2009

Evaluierung
Qualitative

- **Qualitative Evaluierung** liefert *qualitative* Ergebnisse, die sich nicht sinnvoll numerisch fassen lassen
  - Beispiele: Notizen, Interview, Videos …
- **Qualitative Ergebnisse** beinhalten häufig interessante Details, die nicht aus numerischen Daten entnommen werden können

Evaluierung
Quantitativ & Qualitativ

- **Qualitative** und **quantitive** Ergebnisse ergänzen sich häufig sinnvoll
  - **quantitive** Daten erlauben statistische Auswertung und liefern belastbare Ergebnisse
  - **qualitative** Daten runden Ergebnisse einer Studie ab

Analytische vs. Empirische Untersuchung

Evaluierung

Analytisch

- **Analytische Methoden** untersuchen System durch reine Analyse
- **Analytische Methoden** liefern Erklärungen von Arbeitsweisen, Bestandteilen oder Eigenschaften
- **Analytische Methoden** beziehen i.d.R. keine Testpersonen ein

Evaluierung

Empirisch

- **Empirische Methoden** befassen sich mit Ergebnissen bei Bedienung des interaktiven Systems/Produkts
- **Empirische Methoden** beziehen i.d.R. Testpersonen ein

Evaluierung

Empirie vs. Analyse

- **Analytische** und **Empirische Methoden** lassen sich häufig kombinieren
  - **Empirische Methoden** liefern belastbare Ergebnisse über Benutzung interaktiver Systeme
  - **Analytische Methoden** liefern mögliche Erklärungen

Analytische Methoden
Beispiele

- formal analytische Verfahren
  - KLM-GOMS
- Inspektionsverfahren
  - Heuristische Evaluation auf Basis von Nielsen oder Shneiderman
  - Cognitive Walkthrough
  - …

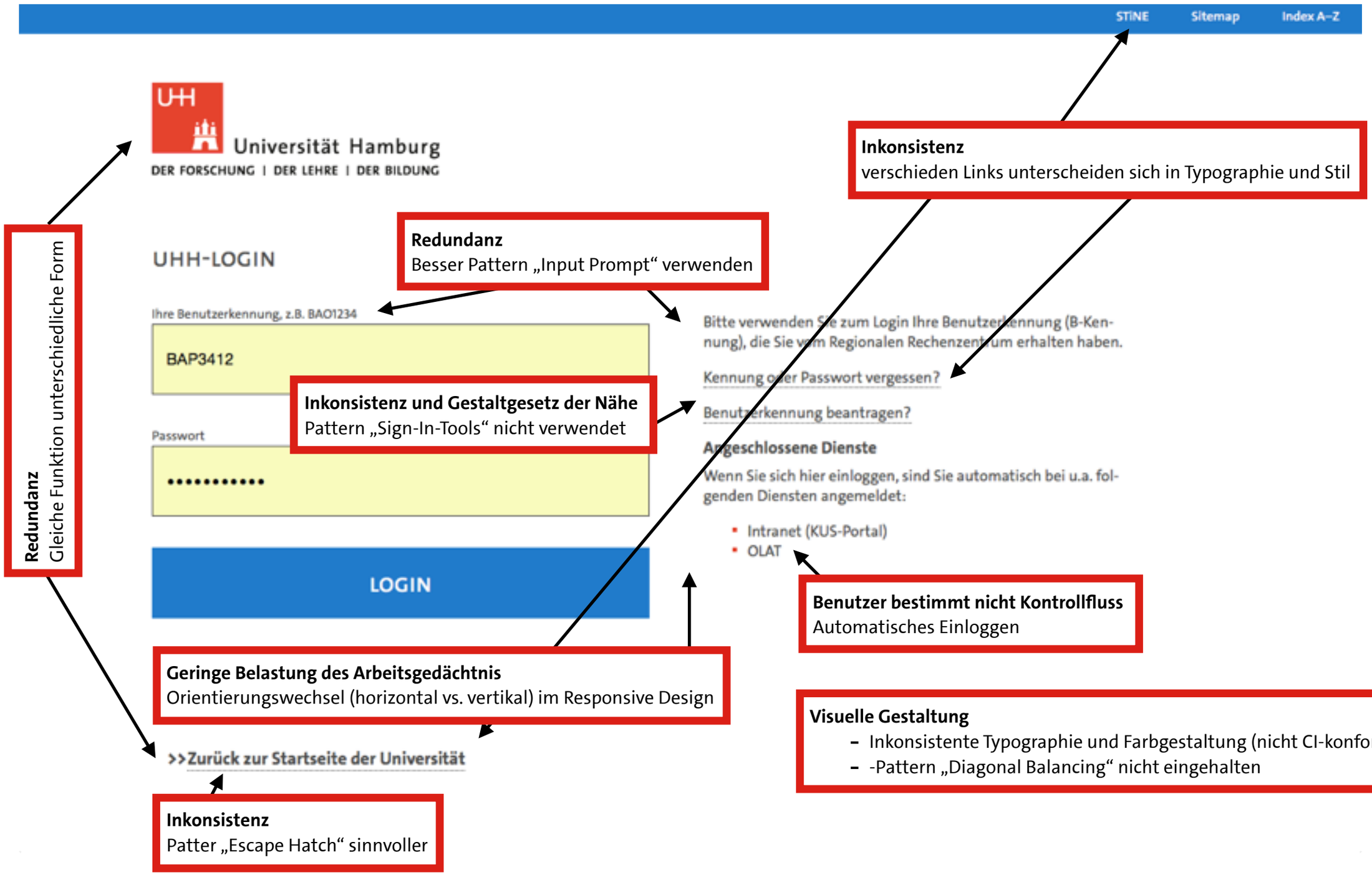
Heuristische Evaluation

Vorgehen

- 3-5 Experten begutachten Prototypen gemäß von Heuristiken
- Orientierung an vorgegebener Aufgabe und vermuteter Benutzereigenschaften
- Zusätzlich sollten Styleguides und abgeleitete Anforderungen berücksichtigt werden

Heuristische Evaluation

Beispiel: KUS-Portal



Fokus IxD

Bsp: Bewertungsskalen

- **Bewertungsskalen** (engl. ***Rating Scales***) erlauben es Einschätzung zu machen, z.B. wie einfach, schnell, sinnvoll …

- Bsp. **Likert-Skala**

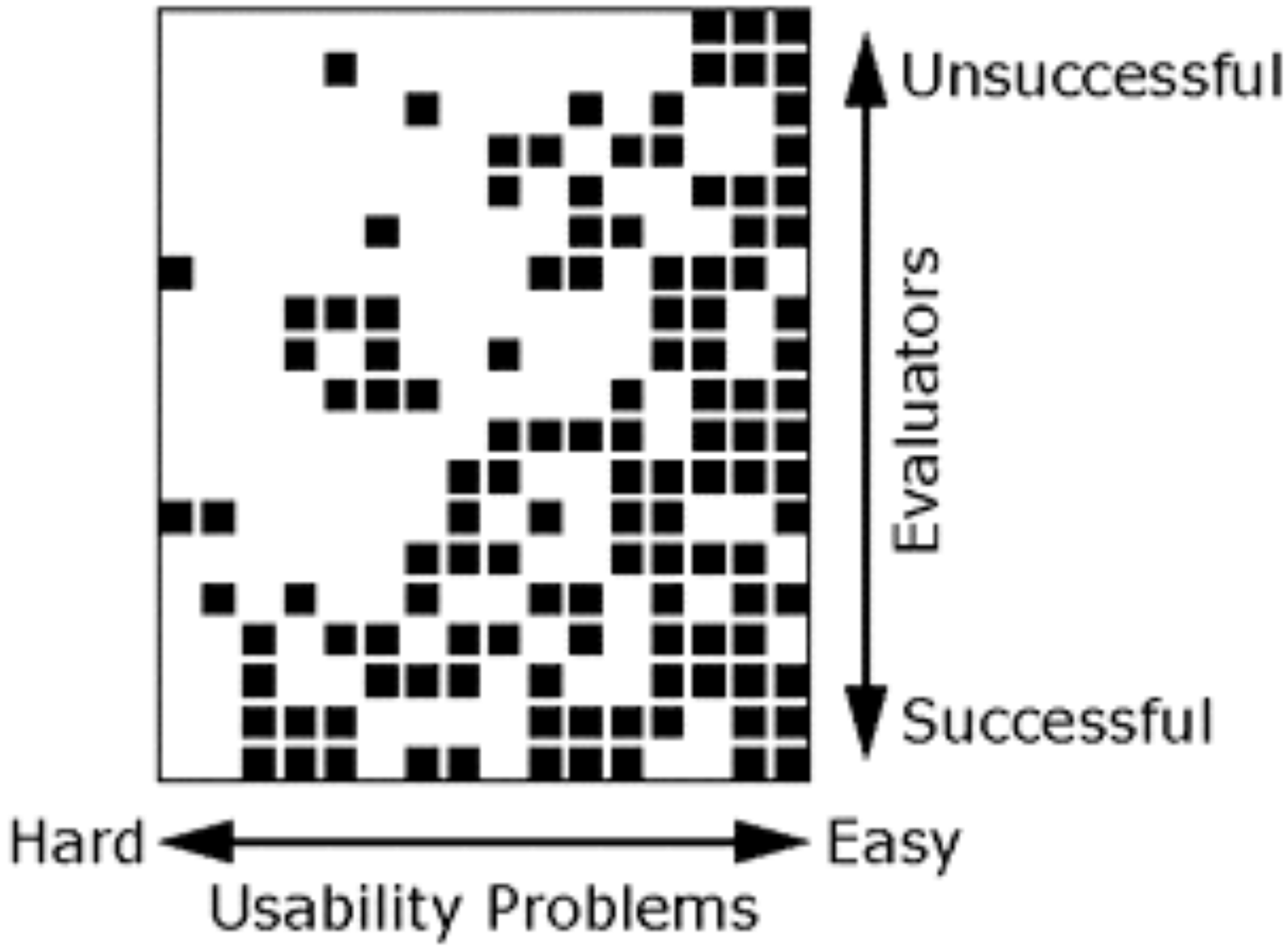
Use of color is important (where 1 presents *strongly agree* and 5 represents *strongly disagree*):

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

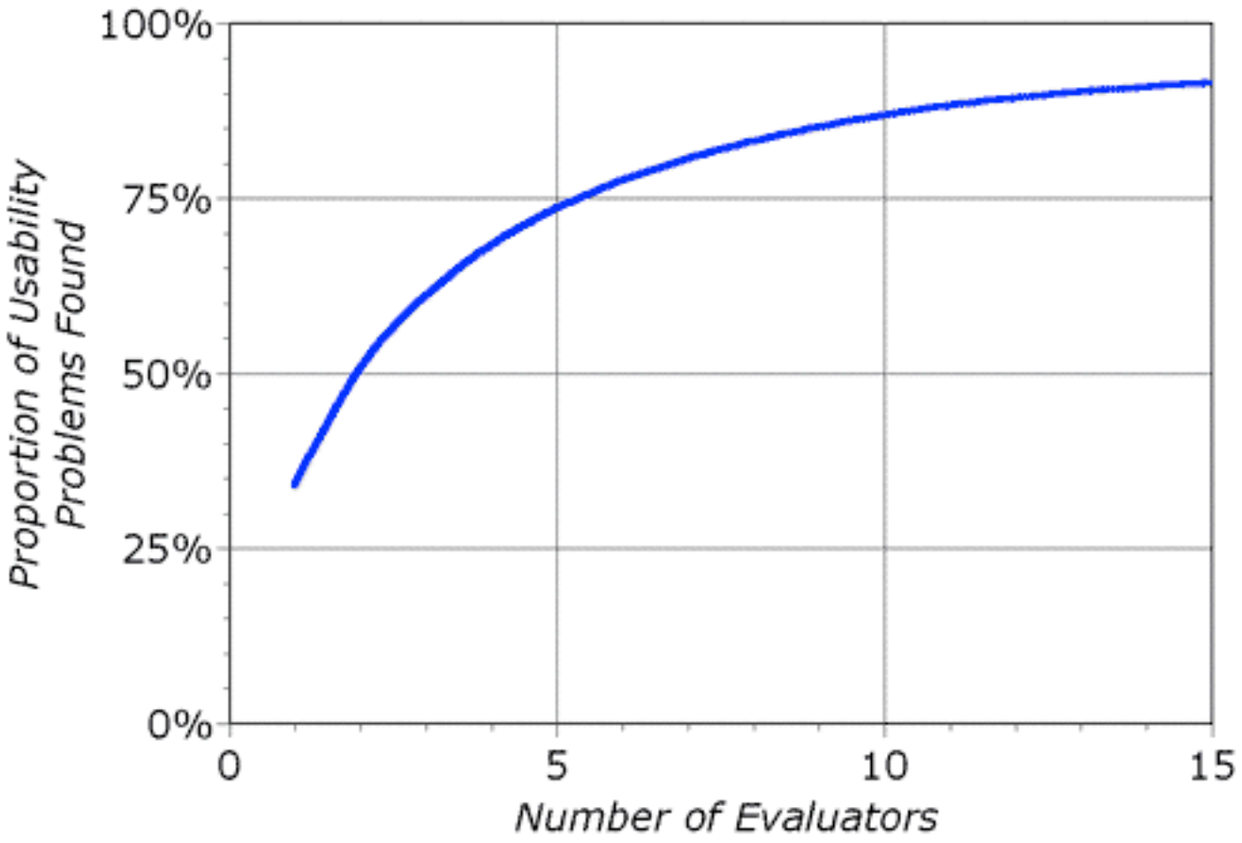
Use of color is important:

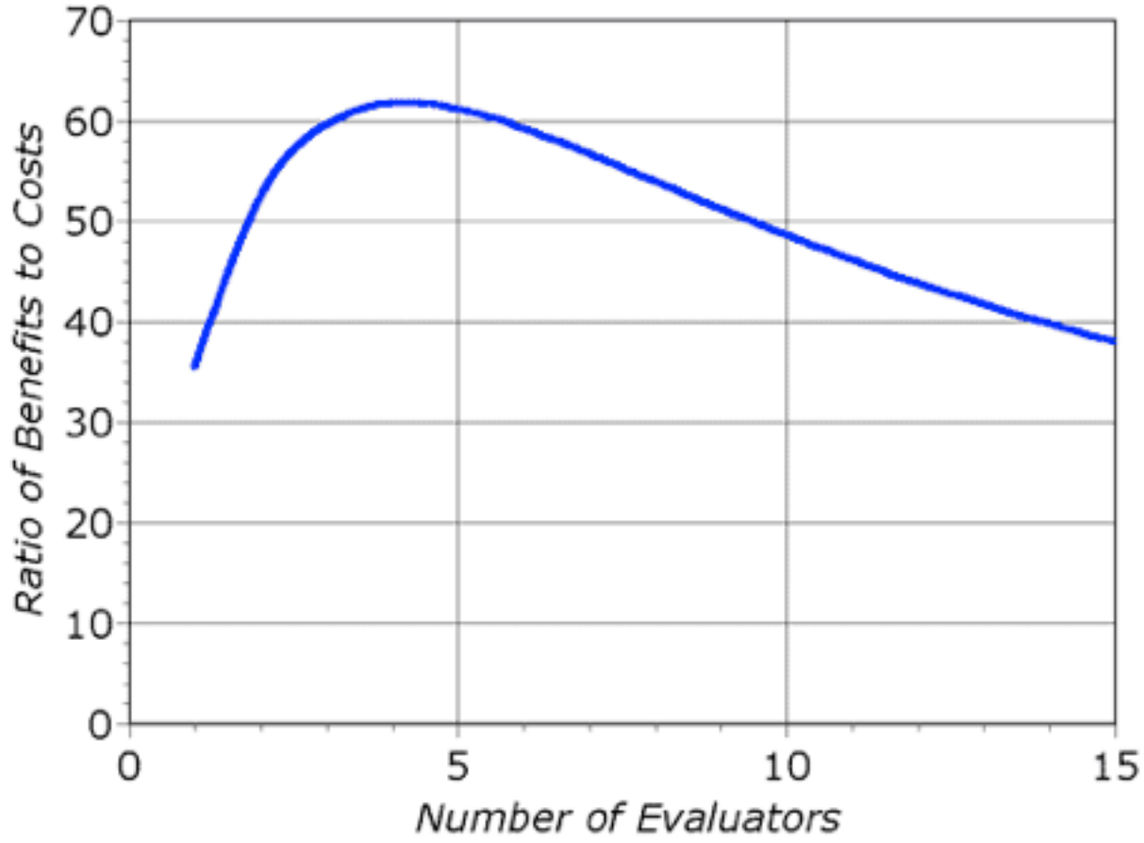
strongly agree	agree	neutral	disagree	strongly disagree
☐	☐	☐	☐	☐

R. Likert: A technique for the measurement of attitudes. Likert, R. Archives of Psychology, Vol 22 140, 1932.

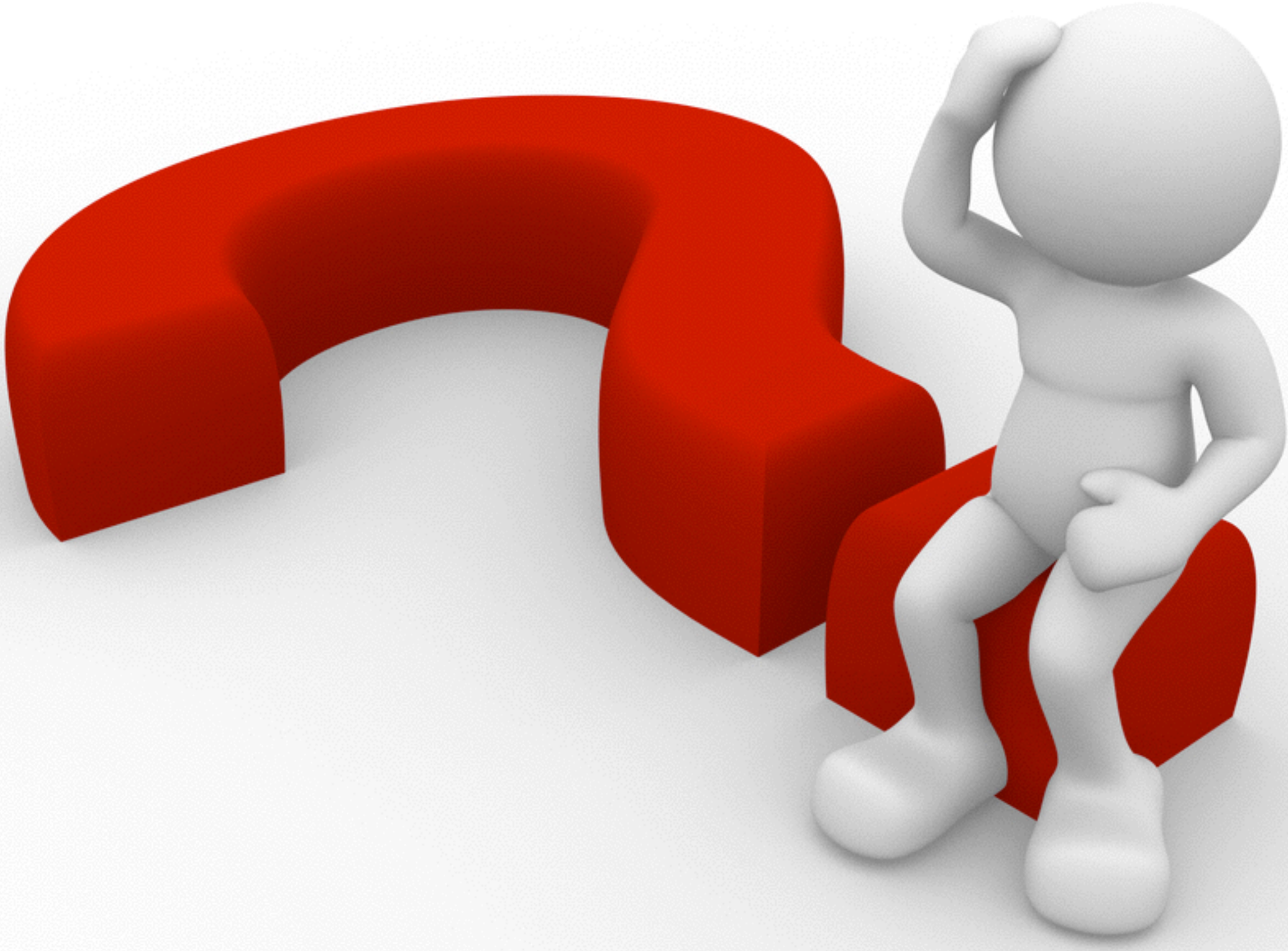


J. Nielsen: How to Conduct a Heuristic Evaluation, 1995





J. Nielsen: How to Conduct a Heuristic Evaluation, 1995



Mensch-Computer-Interaktion

Evaluierung

Empirische Methoden

Empirische Methoden

- **Empirische Methoden** bezeichnen alle Formen von Evaluation, die durch **Messung** oder **anderweitige Sammlung** in Experimenten, Beobachtungen oder Befragungen **Daten** erheben, auf deren Basis **wissenschaftliche Aussagen** gemacht werden können

Empirische Methoden
Wissenschaftliche Aussagekraft

- **Objektivität** bedeutet, dass erhobene Daten unabhängig von Messmethode, Erwartung und Hypothesen des Experimentators sind
- **Reproduzierbarkeit** bedeutet, dass Experiment hinreichend genau beschrieben ist, so dass es wiederholbar wird

Empirische Methoden
Wissenschaftliche Aussagekraft

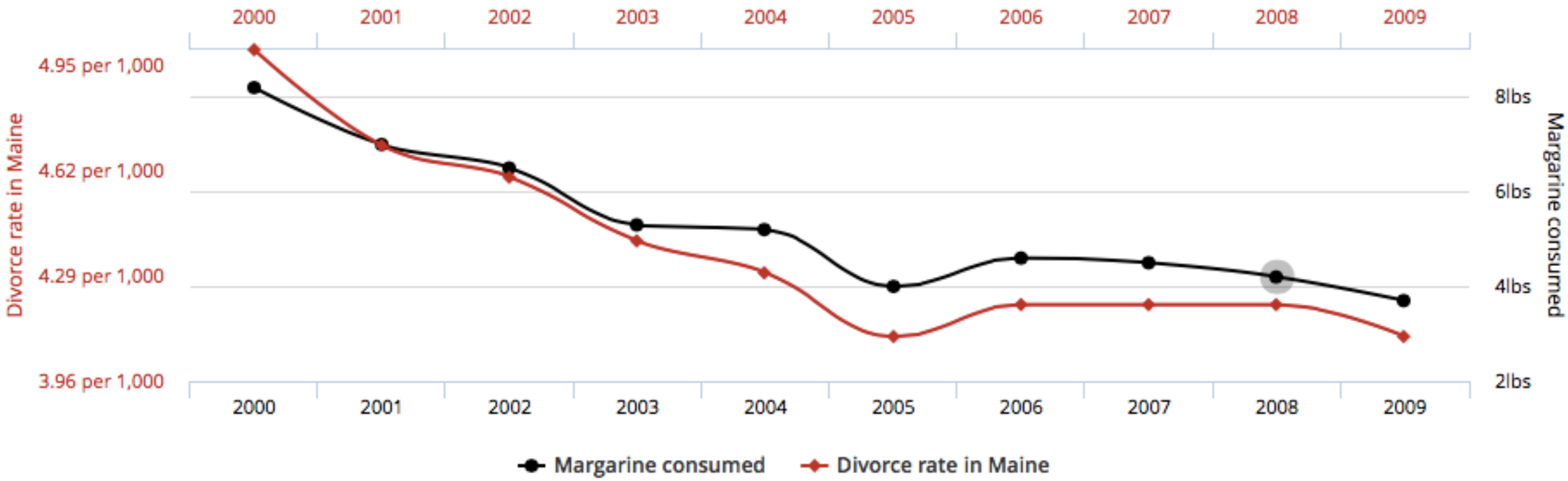
- **Validität**, d.h. Ergebnisse messen nur das, was sie messen sollen (**interne Validität**) und sind repräsentativ für Allgemeinheit (**externe Validität**)
- **Relevanz** bedeutet, dass Ergebnisse tatsächlich neue Erkenntnisse liefern

Divorce rate in Maine

correlates with

Per capita consumption of margarine

Correlation: 99.26% (r=0.992558)



http://www.tylervigen.com/spurious-correlations

Kontrolliertes Experiment

- **Kontrolliertes Experiment** sind Form von empirischen Studien, bei denen alle relevanten Einflussfaktoren kontrolliert werden
- **Kontrolliertes Experiment** sollen mit möglichst wenigen Versuchen **Kausalität** zwischen **Einflussfaktoren** und **Zielgrößen** möglichst genau ermitteln

Experimentdesign

Variablen

- Soll Zusammenhang zwischen mehreren Variablen untersucht werden, sind **unabhängige Variablen** solche Variablen mit deren Ausprägungen einer oder mehrerer anderer Variablen (**abhängige Variablen**) erklärt werden sollen

Experimentdesign

Beispiel: Variablen

- **unabhängige Variable**
  - Eingabegerät (Maus, Tastatur, Touchscreen)
  - Interface-Design
  - …
- **abhängige Variable**
  - Genauigkeit, Zeit, Anzahl der Fehler
  - …

Experimentdesign

Weitere Variablen

- **Kontrollvariablen** könnten Einfluss auf abhängige Variablen haben und sollten daher konstant gehalten & erhoben werden
  - Beleuchtung, Temperatur, Lautstärke …
- **Zufallsvariablen** könnten Einfluss auf abhängige Variable haben, bleiben aber „Zufall überlassen“ (**Generalisierbarkeit**)
  - Größe / Gewicht von Testpersonen …

Experimentdesign
Weitere Variablen

- **Störfaktoren** sind Variablen, die sich mit unabhängiger Variable verändern
- **Störfaktoren** sind problematisch, da unklar ist, ob Kausalität zwischen unabhängiger Variable oder Störfaktor und abhängiger Variable besteht
  - Beispiel: unterschiedliche **Konditionen** (Maus vs. Tastatur) haben Einfluss auf Formfaktoren, Latenz, Auflösung …

Variablentypen

Nominal

- Merkmalsausprägungen *ohne* natürliche Ordnung
- Beispiele: Geschlecht, Berufsstatus, dichotome Antwort vom Typ "ja/nein"

Variablentypen

Ordinal

- Merkmalsausprägungen *mit* natürlicher Ordnung
- Beispiel: Nutzung ("jeden Tag", "einmal in der Woche", "einmal im Monat", "einmal im Jahr", "noch nie genutzt")

Variablentypen

Intervall (metrisch)

- Merkmalsausprägungen *mit* natürlicher Ordnung, *gleiche* Abstände zwischen Werten, *ohne* absoluten Nullpunkt
- Beispiele: Likert-Skalen mit als gleich ansehbaren Abständen
  - "strongly agree", "agree", "neutral", "disagree", "strongly disagree")
  - 1="very high", ..., 7="very low"

Variablentypen

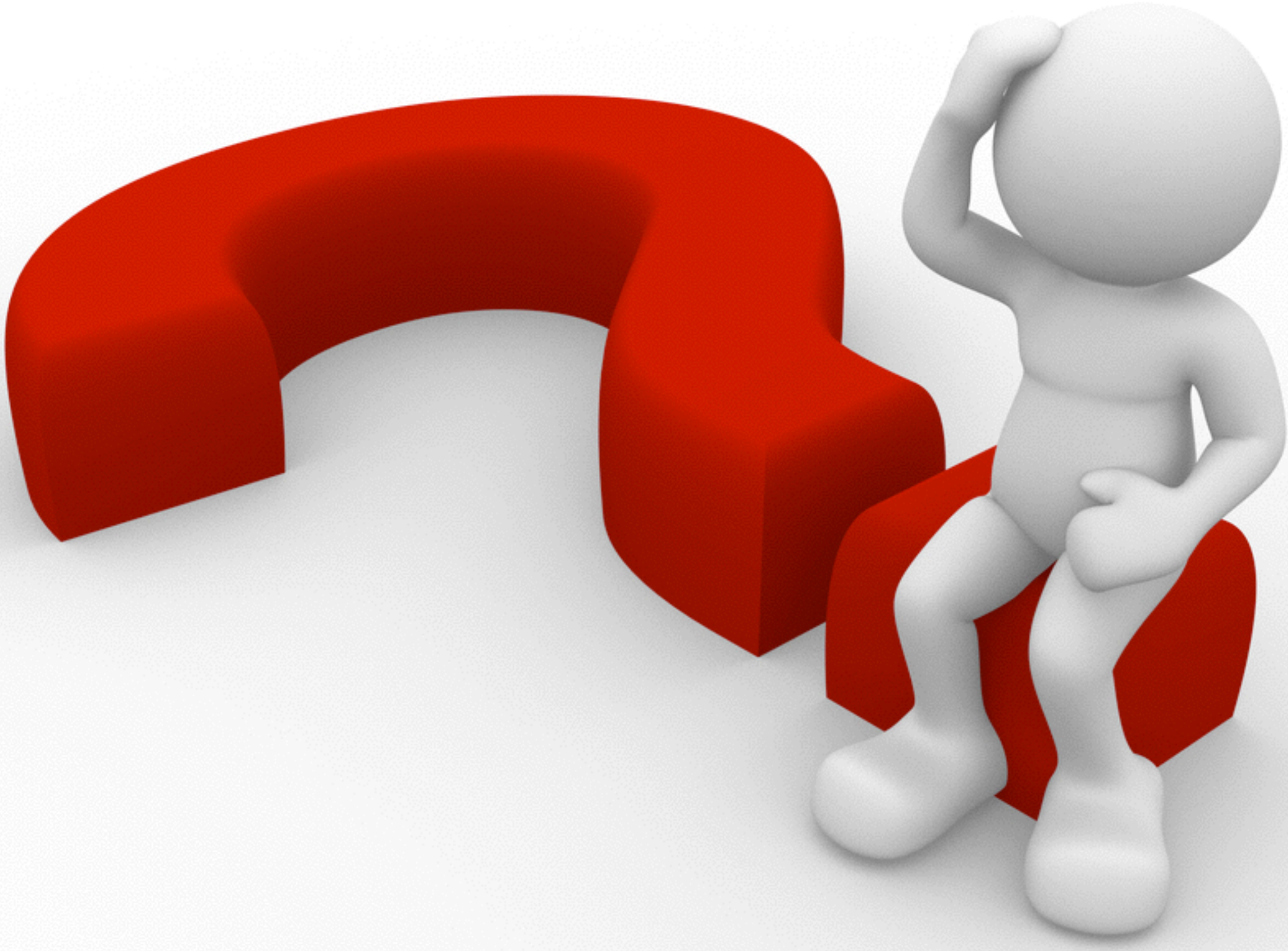
Ratio (metrisch)

- Merkmalsausprägungen *mit* natürlicher Ordnung, *gleiche* Abstände zwischen Werten, *mit* absolutem Nullpunkt
- Beispiele: Einkommen (in Euro), Alter (in Jahren), Leistung (in Stück pro Stunde, in km/h), Gewicht, Größe, Länge, Zeit und Fehlerrate

Experiment

Ziel

- **Ableitung** von **Kausalität** aus **Korrelation** zwischen **unabhängigen** und **abhängigen Variablen**
- Achtung: Lern-/Ermüdungseffekte verhindern durch Permutation
  - Beispiele: (Pseudo-)Randomisierung, Lateinische Quadrate …



Mensch-Computer-Interaktion

Evaluierung

Datenanalyse

Datenanalyse

- **Datenanalyse** ist der Prozess erhobene Daten beschreibend aufzubereiten, um Schlussfolgerungen ziehen zu können
- Analyse basiert auf Datenformat
  - quantitativ (i.d.R. numerisch)
  - qualitativ (i.d.R. nicht-numerisch, lassen sich aber teilweise transformieren)

Deskriptive Statistik

- Beschreibung der Experimentdaten
  - **Minimum**, **Maximum**
  - **Mittelwert** (Summe dividiert durch # Werte)
  - **Standardabweichung** (durchschnittliche Entfernung aller gemessenen Werte vom Mittelwert)
  - **Median** (Mittlerer Wert in geordneter Liste)
  - **Modus** (am häufigsten auftretender Wert)

Deskriptive Statistik

Beispiel

Datenwerte: $0, 1, 1, 2, 3, 4, 4, 4, 5$

- **Minimum:** 0; **Maximum:** 5
- **Mittelwert**:

$$\mu = \frac{0+1+1+2+3+4+4+4+5}{9} \approx 2.67$$

- **Standardabweichung:**

$$\sigma = \sqrt{\frac{(0-2.67)^2 + ... + (5-2.67)^2)}{9}} \approx 1.63$$

- **Median:** 3; **Modus**: 4

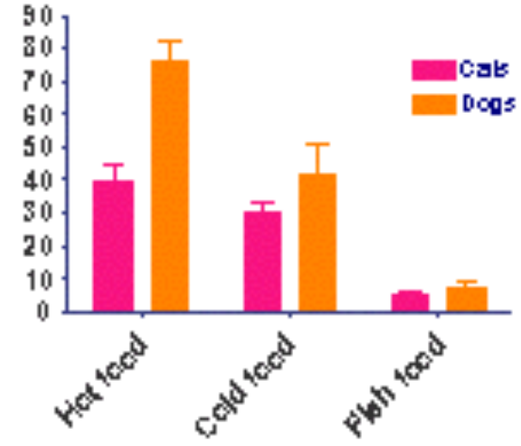
Deskriptive Statistik

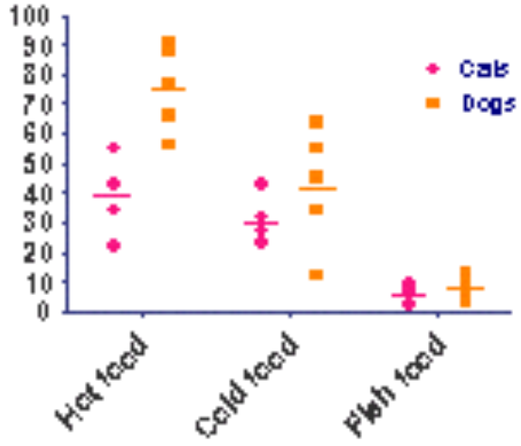
Beispiele

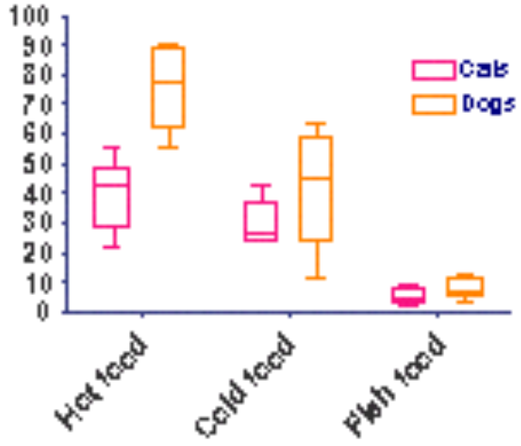
- *”Zehn Teilnehmer haben beide Experimentteile durchgeführt (5 Männer, 5 Frauen; durchschnittliches Alter 22,4 Jahre, zwischen 18-37 Jahren).”*
- *”Die durchschnittsliche Bewegungszeit im Fitts’ Law Experiment in der Gruppe mit Maussteuerung war 34,5 Sekunden (SD=5.4 Sek., min=19.2 Sek., max=305.1 Sek.).”*

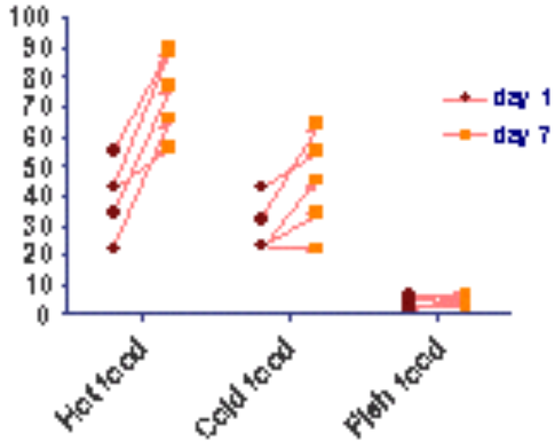
Deskriptive Statistik

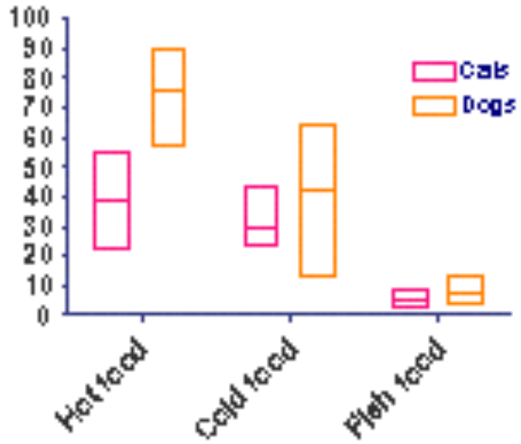
- **Grafische Darstellung** mit Diagrammen gibt Überblick über Daten







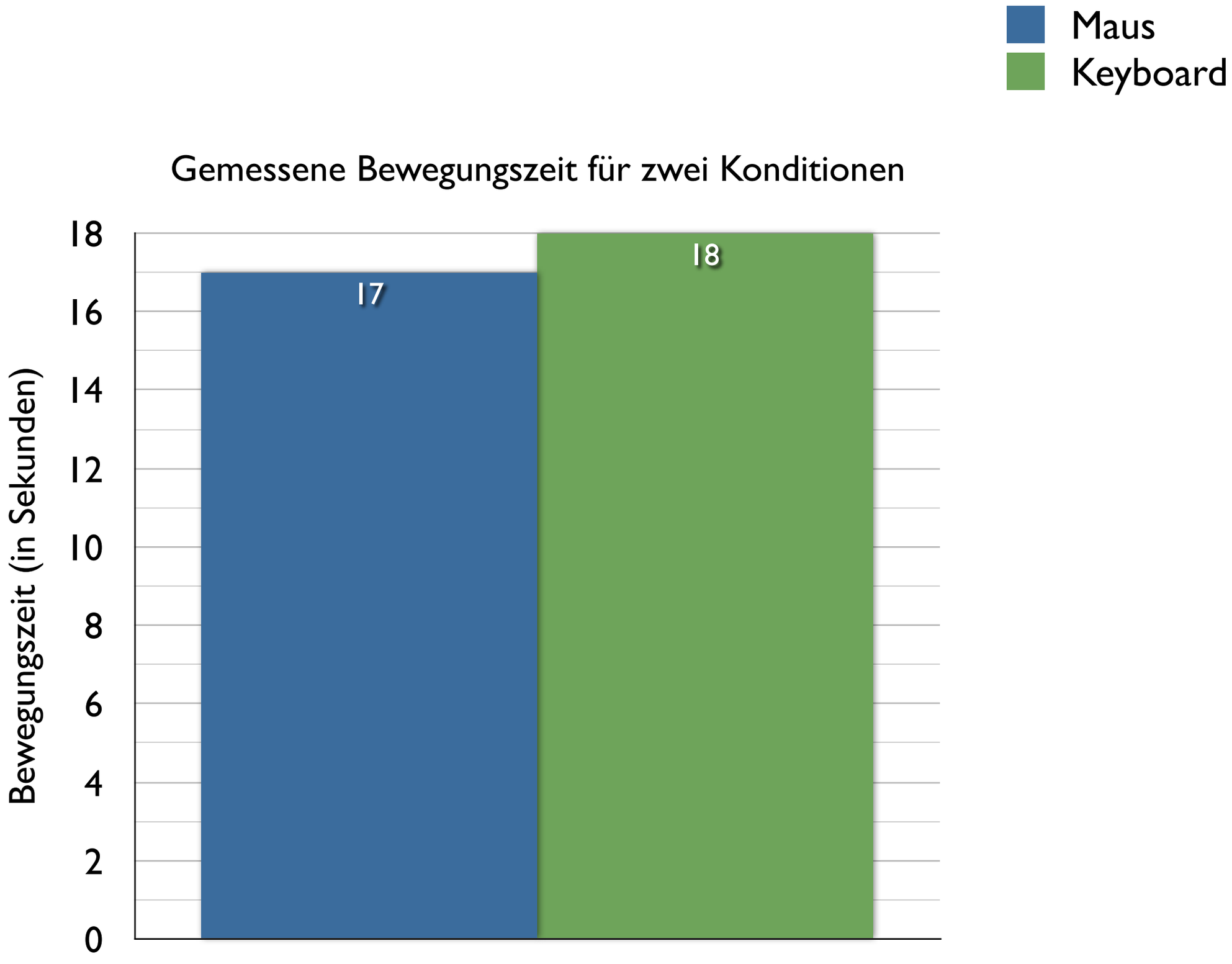


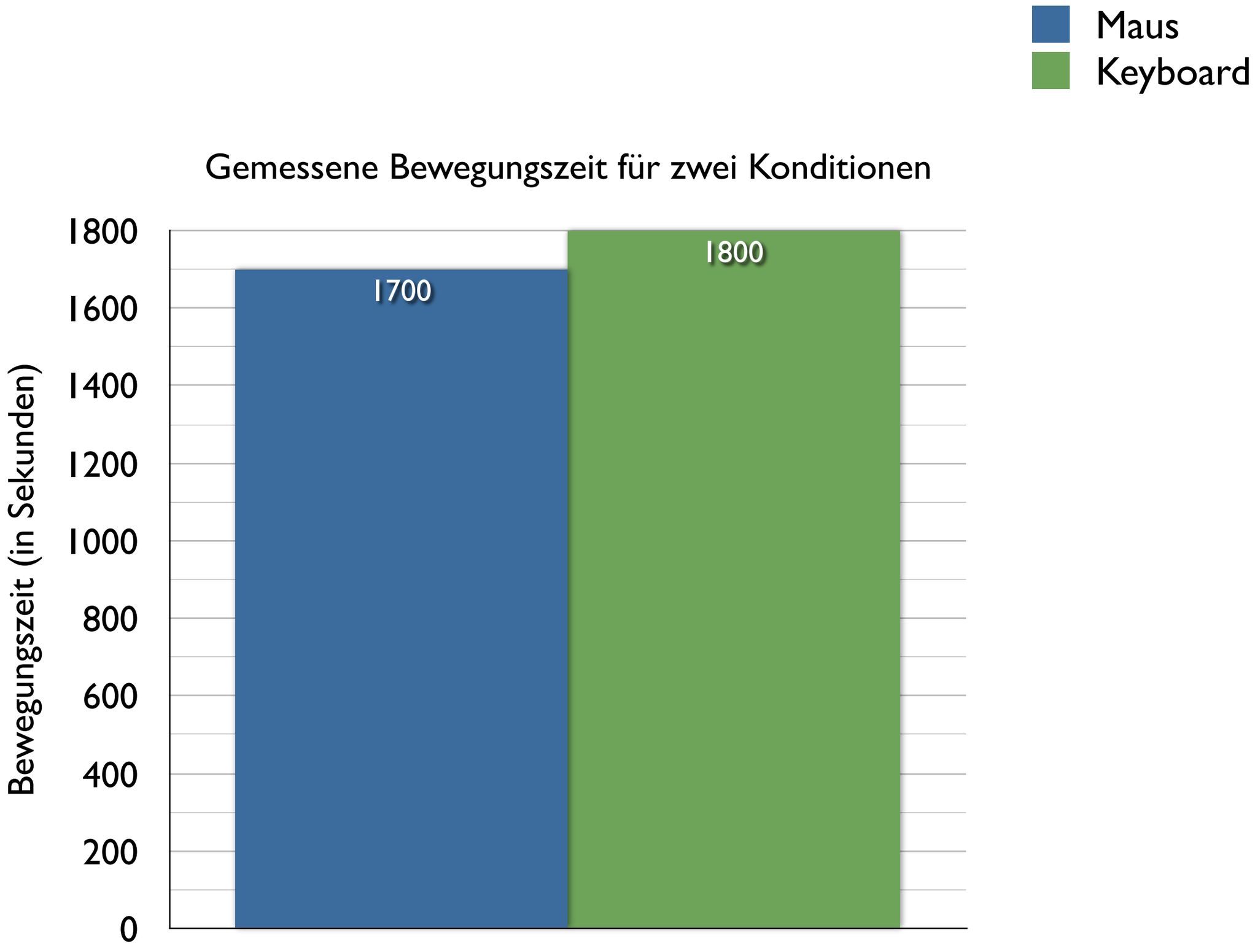


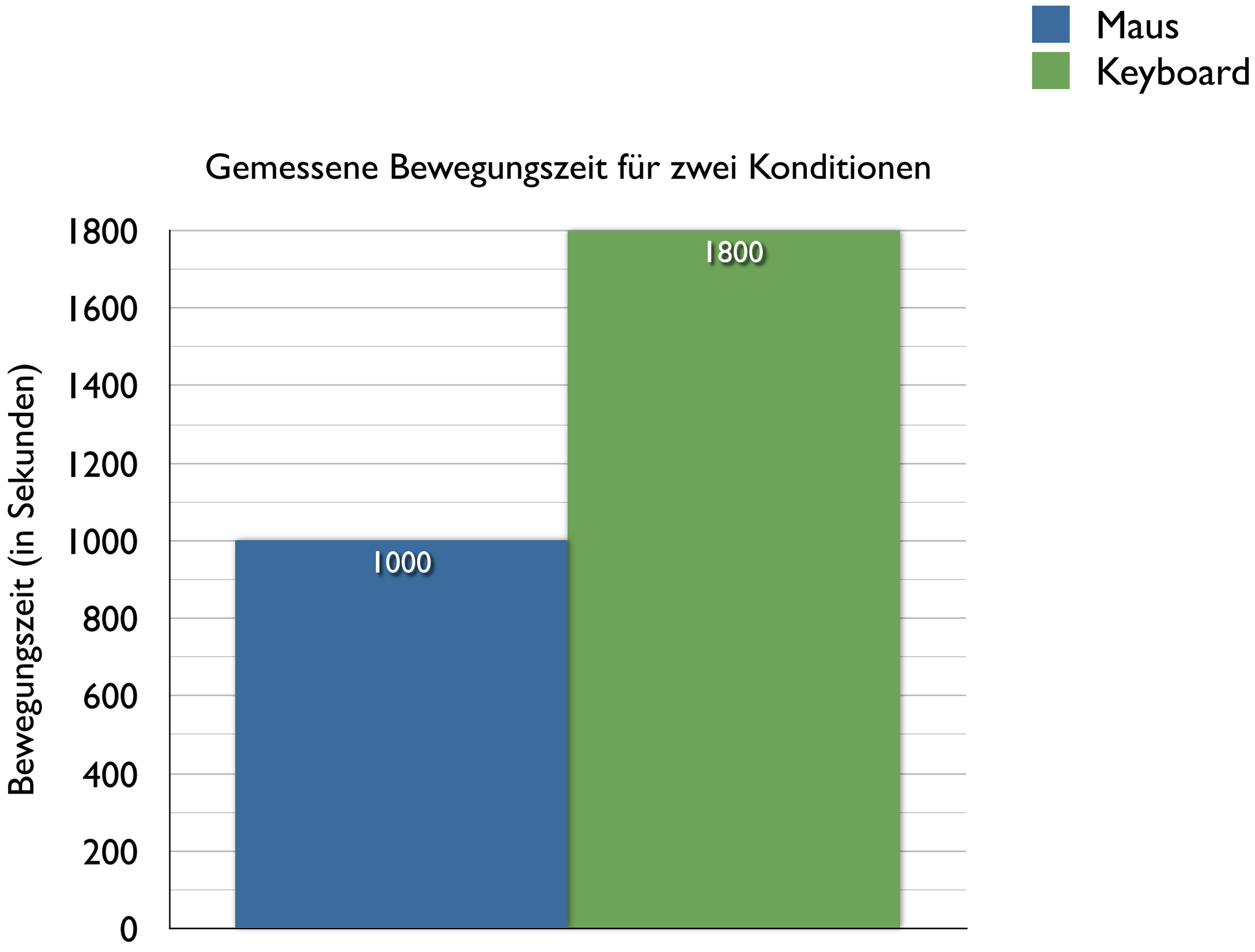
Grafische Darstellung

- **Grafische Darstellungen** geben Eindruck von Verteilungen, Verhältnissen, Größen
- Diagramme sagen nicht ob **signifikante**, d.h. unter gleichen Bedingungen höchstwahrscheinlich reproduzierbare, **Unterschiede** existieren

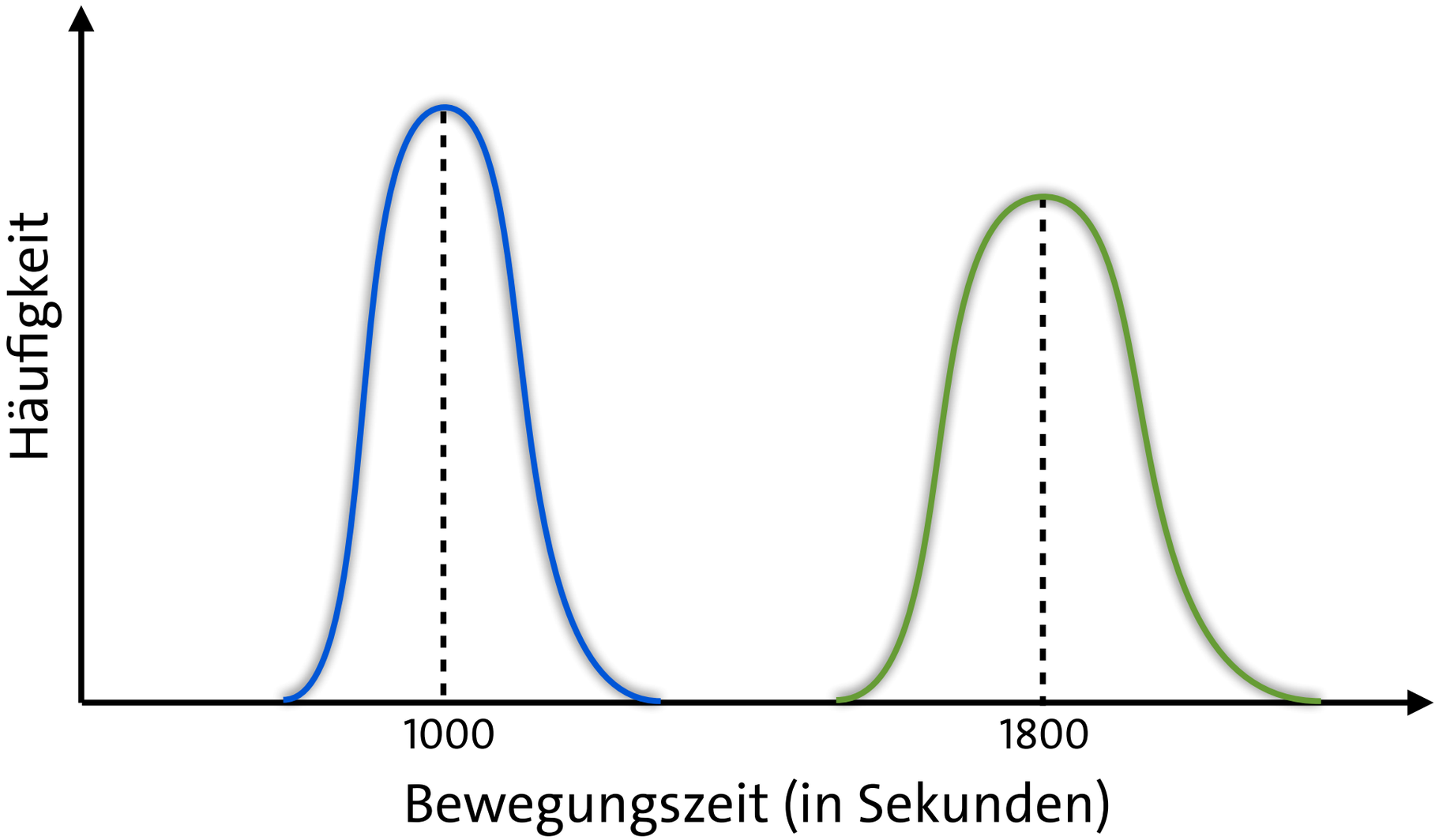
➡ Einsatz statistischer Signifikanztests


Maus
Keyboard
Gemessene Bewegungszeit für zwei Konditionen
Bewegungszeit (in Sekunden)
18
16
14
12
10
8
6
4
2
0
17
18

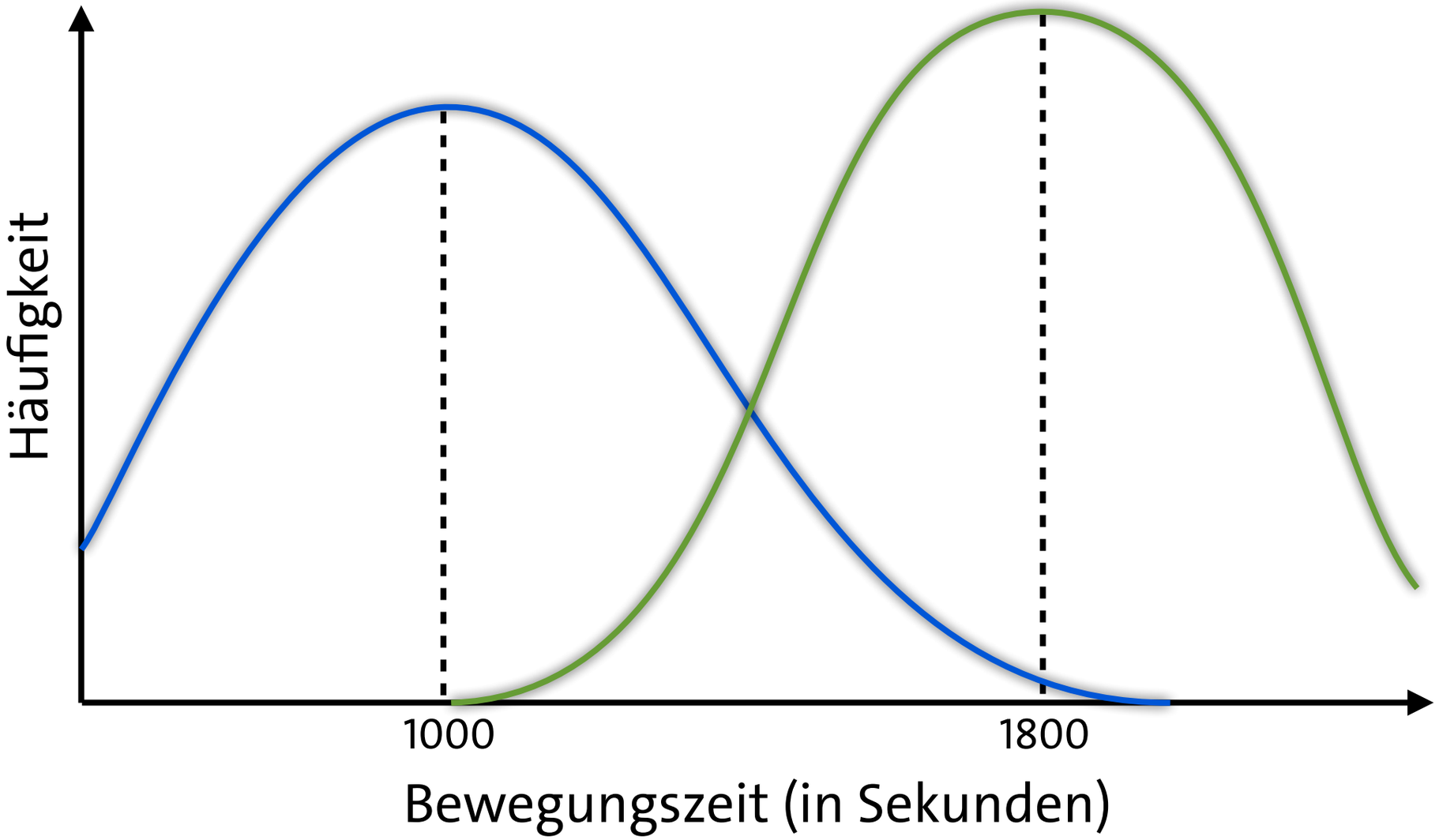

Maus
Keyboard
Gemessene Bewegungszeit für zwei Konditionen
Bewegungszeit (in Sekunden)
1800
1600
1400
1200
1000
800
600
400
200
0
1700
1800


Maus
Keyboard
Gemessene Bewegungszeit für zwei Konditionen
Bewegungszeit (in Sekunden)
1800
1600
1400
1200
1000
800
600
400
200
0
1800
1000

Verteilungskurven der Messwerte

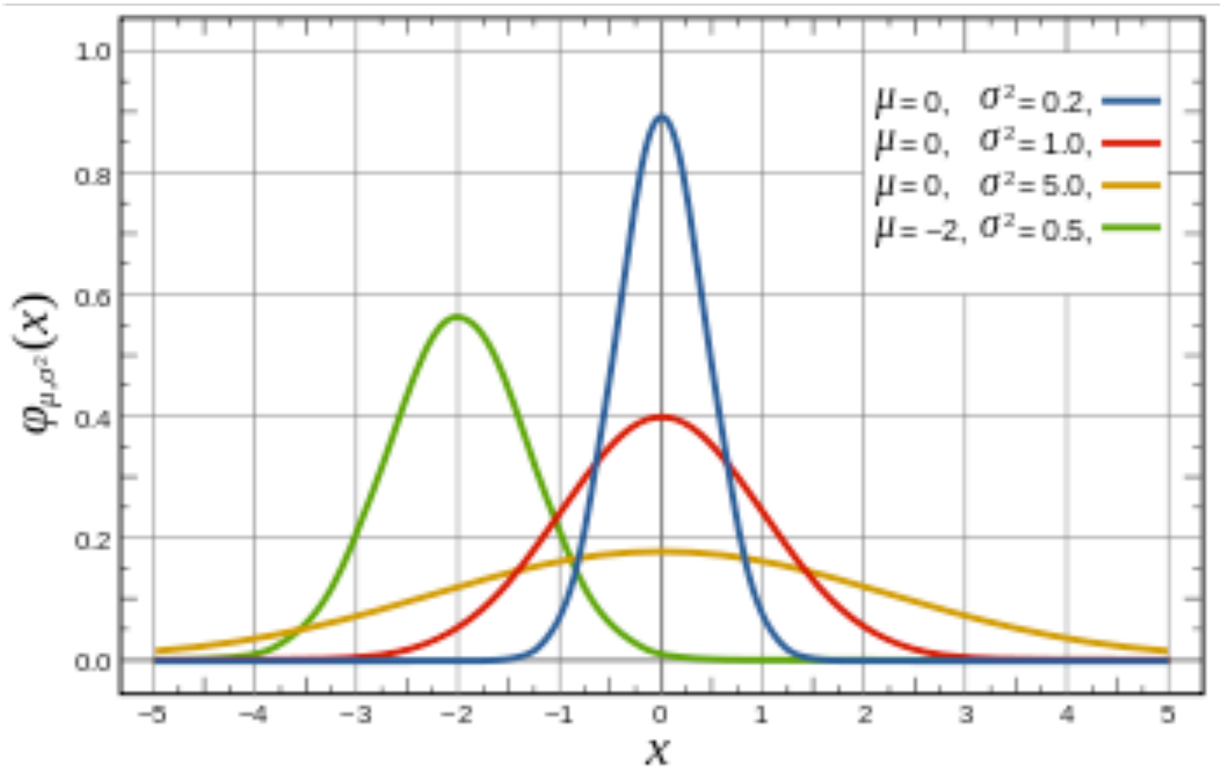


Verteilungskurven der Messwerte



Normalverteilung

Verteilung

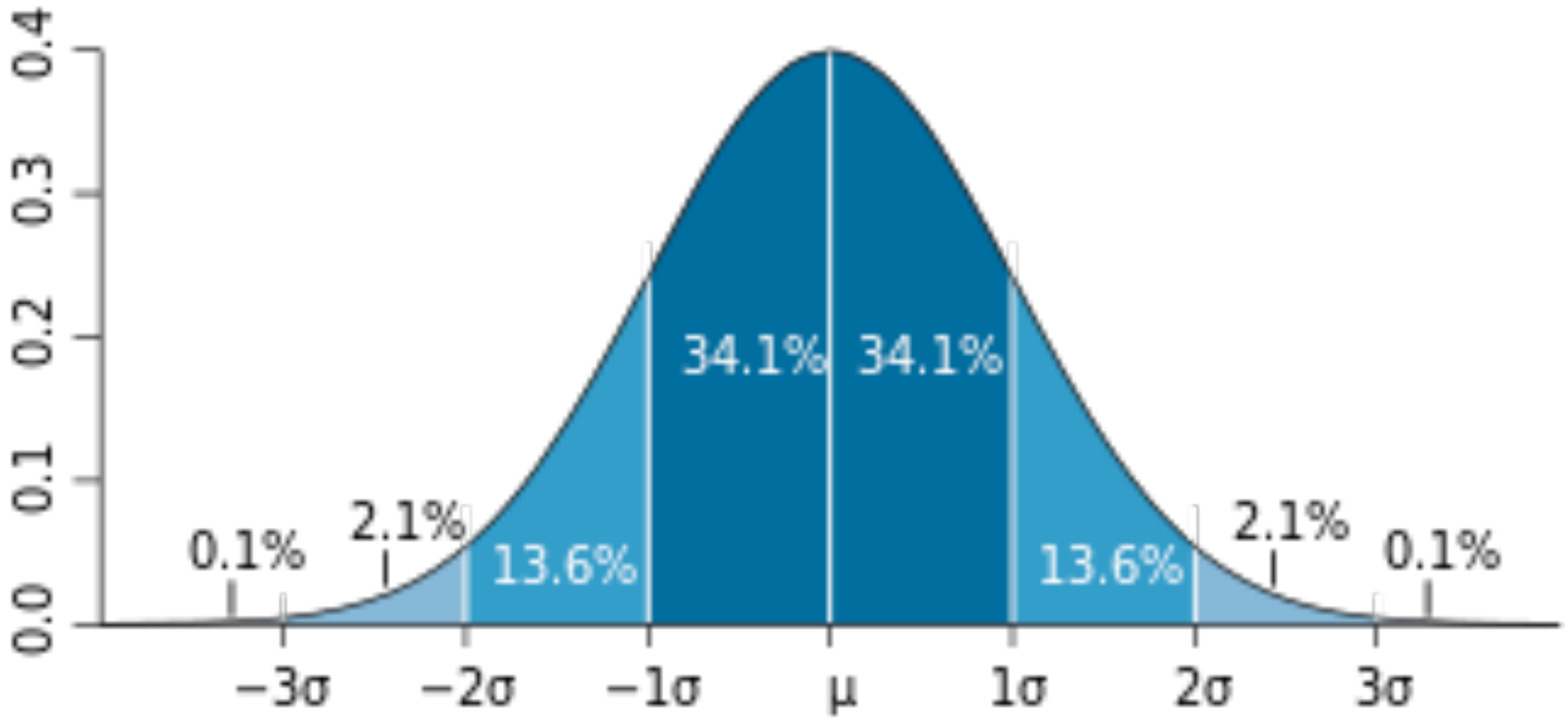


- Gebräuchlichste Verteilung, symmetrisch um **Mittelwert** μ, Breite gegeben durch **Standardabweichung** σ

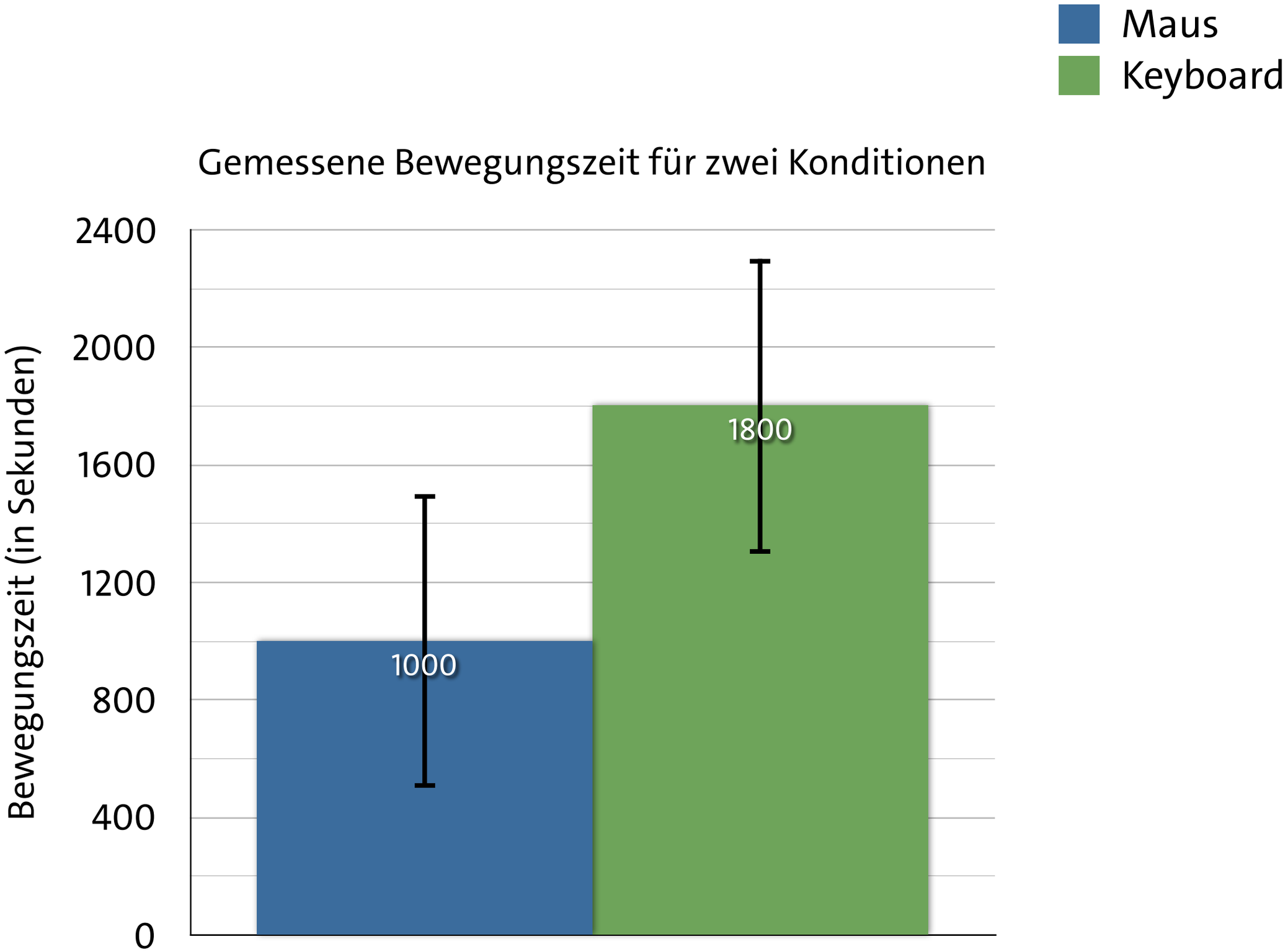
  - *Mittelwert = Median = Mode*

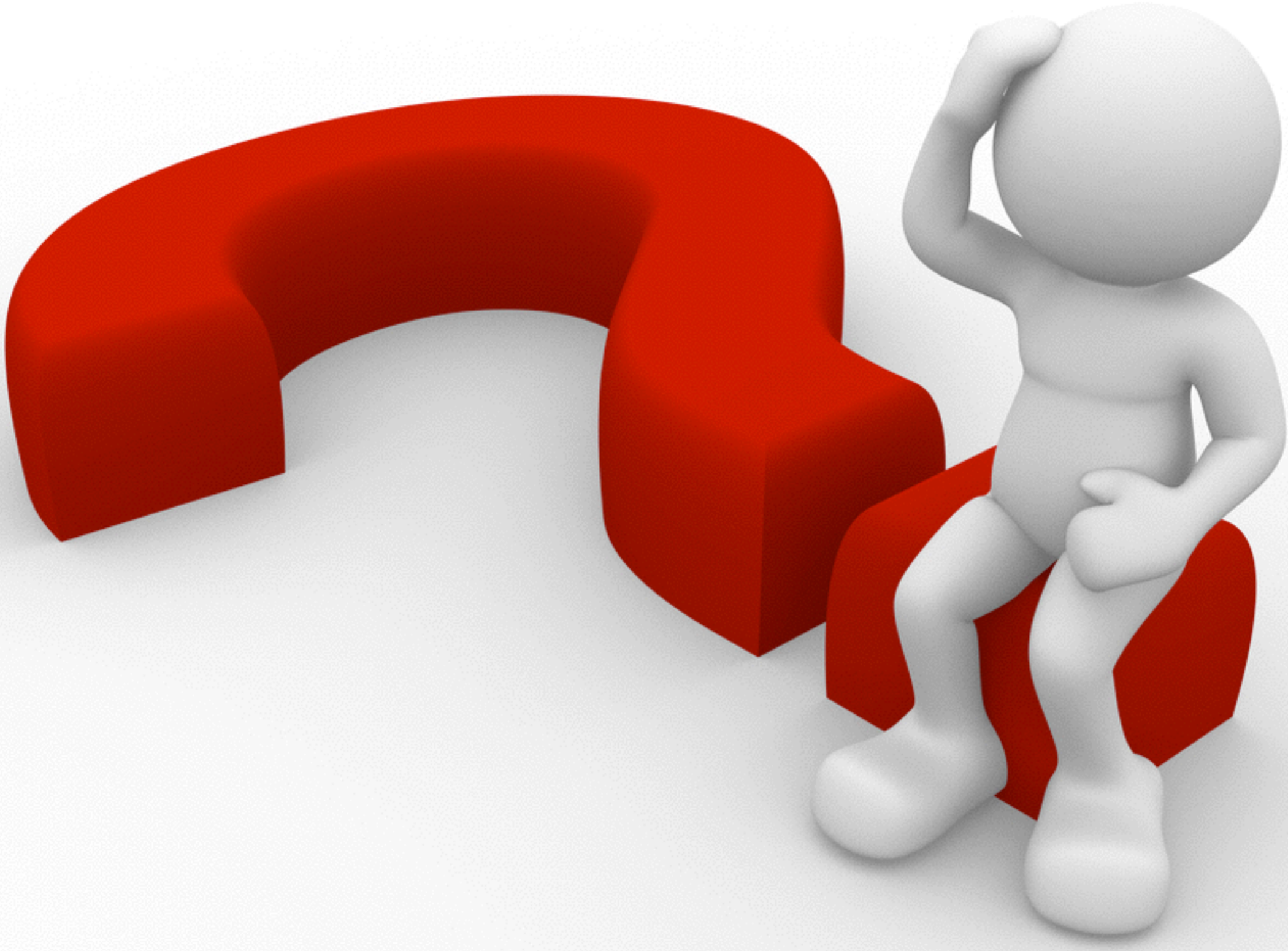
Normalverteilung

Ausreißer



- 99.7% aller Datenpunkte liegen innerhalb von +/- $3 \cdot \sigma$ um μ
- Gängiges Mittel: Datenpunkte außerhalb als Ausreißer entfernen


Maus
Keyboard
Gemessene Bewegungszeit für zwei Konditionen
Bewegungszeit (in Sekunden)
2400
2000
1600
1200
800
400
0
1800
1000



Mensch-Computer-Interaktion

Evaluierung

Statistische Tests

Statistische Tests

- Nach Durchführung eines *quantitativen* Experiments
  - **Auswahl** des *korrekten* statistischen Testverfahrens
  - **Durchführung** des statistischen Tests
  - **Interpretation** & **Präsentation**

Hypothesen

- In jeder Studie gibt es mindestens zwei Hypothesen:
  - **Hypothese**: Vorhergesagter Einfluss der Konditionen auf Messwerte
  - **Nullhypothese**: Konditionen haben keinen Einfluss auf Messwerte
- Ziel: Hypothese statistisch belegen, Nullhypothese verwerfen

Hypothesen
Beispiele

- **Hypothese:** Mittelwerte der Messwerte (z.B. Zeit, Fehler, Genauigkeit ...) unterscheiden sich zwischen den Konditionen
- **Nullhypothese:** Mittelwerte unterscheiden sich nicht

Messwiederholungen

- Teilnehmer im Experiment haben entweder teilgenommen an
  1. allen Konditionen (engl. ***within-subjects design***) oder
  2. nur einen Teil der Konditionen (engl. ***between-subjects design***)

Welchen Test?

		Typ der abhängigen Variable		
	Within-/ between subjects design	**Intervall/Ratio (Normalität angenommen)**	**Intervall/Ratio (Normalität nicht angenommen), Ordinal**	**Dichotomy (Bi-Nomial)**
Mittelwertsvergleich von zwei Gruppen	between	Unpaired t test	Mann-Whitney test	Fisher's test
	within	Paired t test	Wilcoxon test	McNemar's test
Mittelwertsvergleich von mehr als zwei Gruppen	between	ANOVA	Kruskal-Wallis test	Chi-square test
	within	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Zusammenhang zwischen zwei Variablen finden	within/ between	Pearson correlation	Spearman correlation	Cramer's V
Wert vorhersagen mit einer unabhängigen Variable		Linear/Non-linear regression	Non-parametric regression	Logistic regression
Wert vorhersagen mit mehreren unabhängigen Variablen oder binomialen Variablen		Multiple linear/non-linear regression		Multiple logistic regression

Experiment
Beispiel

- **Forschungsfrage**: Interagieren Menschen mit Technik A schneller oder langsamer als mit Technik B
  - **Unabhängige Variable:** Benutzerschnittstelle (Technik A vs. Technik B)
  - **Abhängige Variable:** Zeit um Aufgabe zu erfüllen (in Sekunden)

Experiment
Beispiel

- **Hypothese:** Zeiten zur Erfüllung der Aufgabe unterscheiden sich zwischen Technik A und Technik B
- **Nullhypothese:** Zeiten, die mit Technik A und Technik B benötigt werden, unterscheiden sich nicht

Experiment
Beispiel

- ***Within-subjects Design***: Alle Teilnehmer führen alle Konditionen durch

	Technik A	Technik B
Teilnehmer 1	17 sec	12 sec
Teilnehmer 2	19 sec	15 sec
Teilnehmer 3	13 sec	10 sec
…		

Welchen Test?

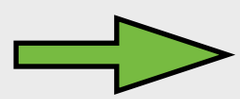
		Typ der abhängigen Variable		
	Within-/ between subjects design	**Intervall/Ratio (Normalität angenommen)**	**Intervall/Ratio (Normalität nicht angenommen), Ordinal**	**Dichotomy (Bi-Nomial)**
Mittelwertsvergleich von zwei Gruppen	between	Unpaired t test	Mann-Whitney test	Fisher's test
	within	Paired t test	Wilcoxon test	McNemar's test
Mittelwertsvergleich von mehr als zwei Gruppen	between	ANOVA	Kruskal-Wallis test	Chi-square test
	within	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Zusammenhang zwischen zwei Variablen finden	within/ between	Pearson correlation	Spearman correlation	Cramer's V
Wert vorhersagen mit einer unabhängigen Variable		Linear/Non-linear regression	Non-parametric regression	Logistic regression
Wert vorhersagen mit mehreren unabhängigen Variablen oder binomialen Variablen		Multiple linear/non-linear regression		Multiple logistic regression

Paired T-Test

Beispiel

```
> t.test(my_data_A,my_data_B,
  paired=TRUE,...)
Paired t-test
data: my_data_A and my_data_B
t = 2.4575, df = 9, p-value = 0.01815
```

Statistik in R



p<.05 ⇒ signifikant

- Reporten z.B. als...

  *"A **paired-samples t-test** was conducted to compare task completion time between conditions with technique A and technique B. We found a **significant** difference in the results for technique A (M=16.1, SD=2.1) and technique B (M=12.2, SD=2.2) at the 5% significance level; **t(9)=2.4575, p=.018**. The results suggest that technique B is **faster** than technique A."*