

# **Tag 2: Big Data, Good Data? Die kritische Nutzung von Daten**

## Session 7: Informierter Konsum von Daten und Statistiken

---

Simon Munzert  
Hertie School

1. Deskriptive Statistik verstehen
2. Wahrscheinlichkeit verstehen
3. Statistische Signifikanz verstehen
4. Klassifikationsmetriken verstehen

# **Deskriptive Statistik verstehen**

---

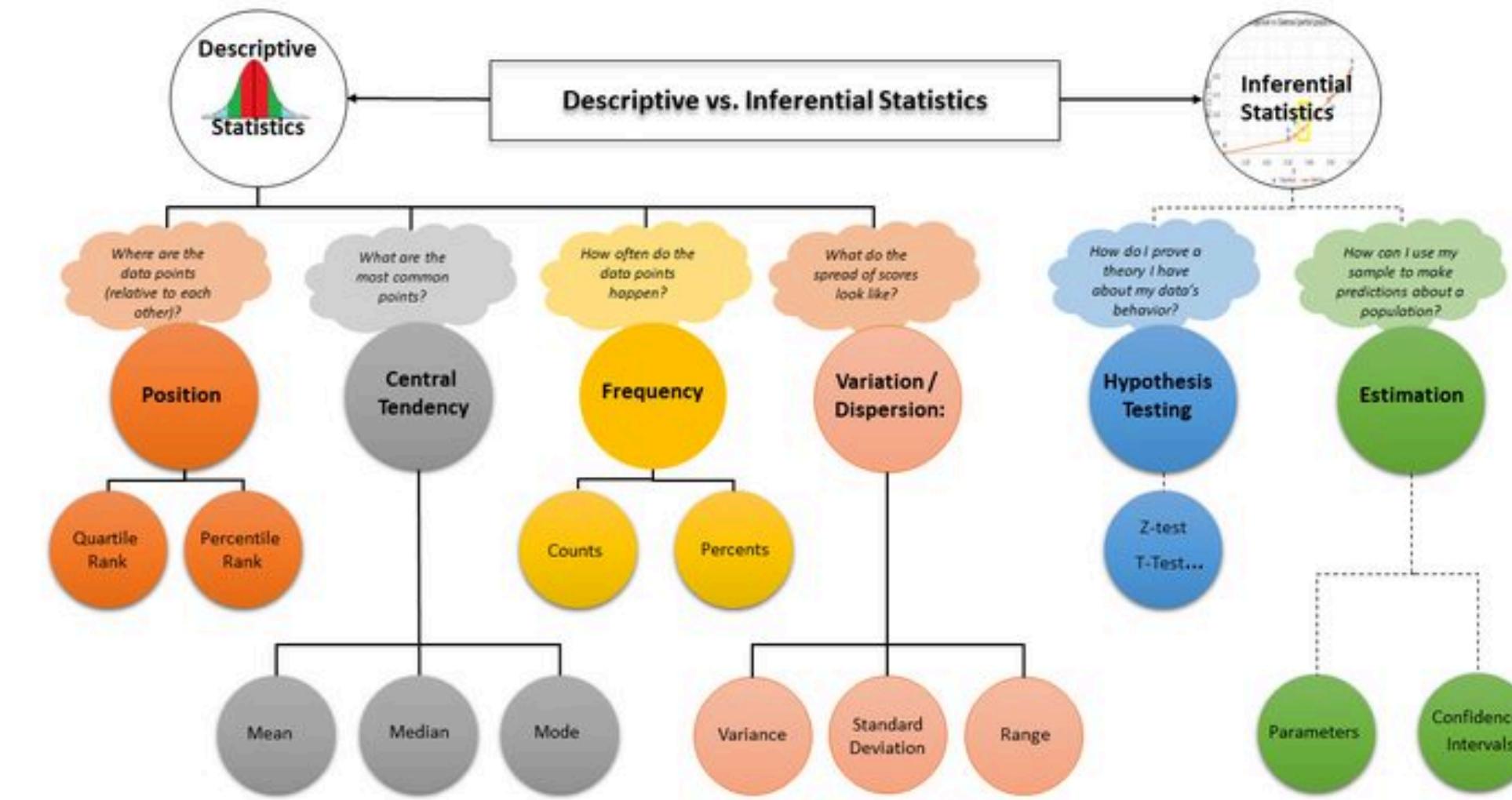
## Deskriptive Statistik

- Zusammenfassen und Beschreiben von Merkmalen einer Stichprobe oder Population
- Kann numerisch und visuell kommuniziert werden
- Unterschiedliche Skalen (Messniveaus) erfordern unterschiedliche deskriptive Statistiken
- Eine gute Beschreibung kann schwierig sein, wenn die Datenerfassung oder Messung komplex ist.

## Inferenzstatistik

- Schlussfolgerungen über eine Grundgesamtheit auf der Grundlage einer Stichprobe
- Es können Rückschlüsse auf Mittelwerte, Proportionen, Beziehungen usw. gezogen werden
- Kann numerisch und visuell kommuniziert werden
- Gute Beschreibung ist die Grundlage für gute Schlussfolgerungen

# Deskriptive vs. Inferenzstatistik



## Drei populäre Maße für zentrale Tendenz

- (Arithmetisches) Mittelwert: Der Durchschnitt aller Werte in einem Datensatz
- Median: Der mittlere Wert eines Datensatzes
- Modus: Der häufigste Wert in einem Datensatz

Warum „zentrale Tendenz“? Beschreibt die Tendenz von quantitativen Daten, sich um einen zentralen Wert zu gruppieren.

## Probieren Sie es aus

Ermitteln Sie den Modus, den Median und den Mittelwert der folgenden Werte:

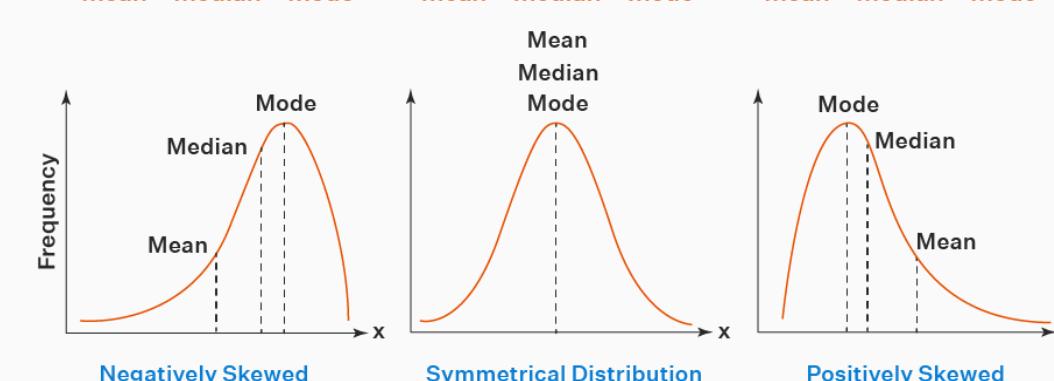
8, 2, 4, 2, 18, 6, 2

## Welches Maß ist zu verwenden?

- **Mittelwert:** Empfindlich gegenüber Ausreißern, aber intuitiv
- **Median:** Robust gegenüber Ausreißern, etwas weniger intuitiv
- **Modus:** Nützlich für kategoriale Daten, kann aber bei kontinuierlichen Daten irreführend sein

## Verzerrte Verteilungen

mean < median < mode



# Maße für zentrale Tendenz: Beispiele

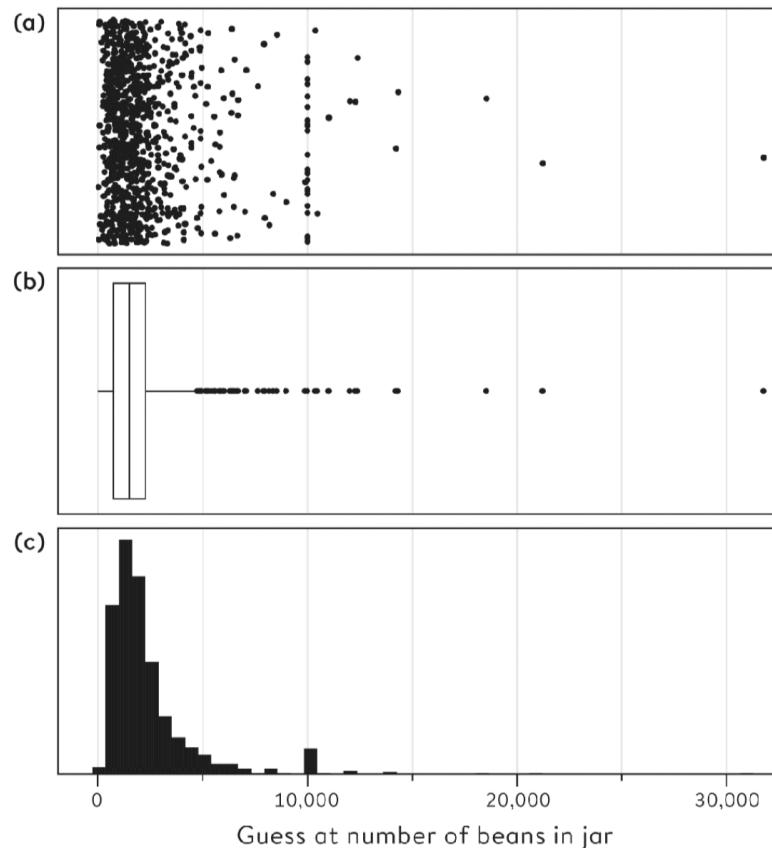


Figure 2.2

Different ways of showing the pattern of 915 guesses of the number of jelly beans in the jar. (a) A strip-chart or dot-diagram, with a jitter to prevent points lying on top of each other; (b) a box-and-whisker plot; (c) a histogram

## Average vs median income

Median and mean income between 2012 and 2014 in selected OECD countries in USD; weighted by the currencies' respective [purchasing power \(PPP\)](#).

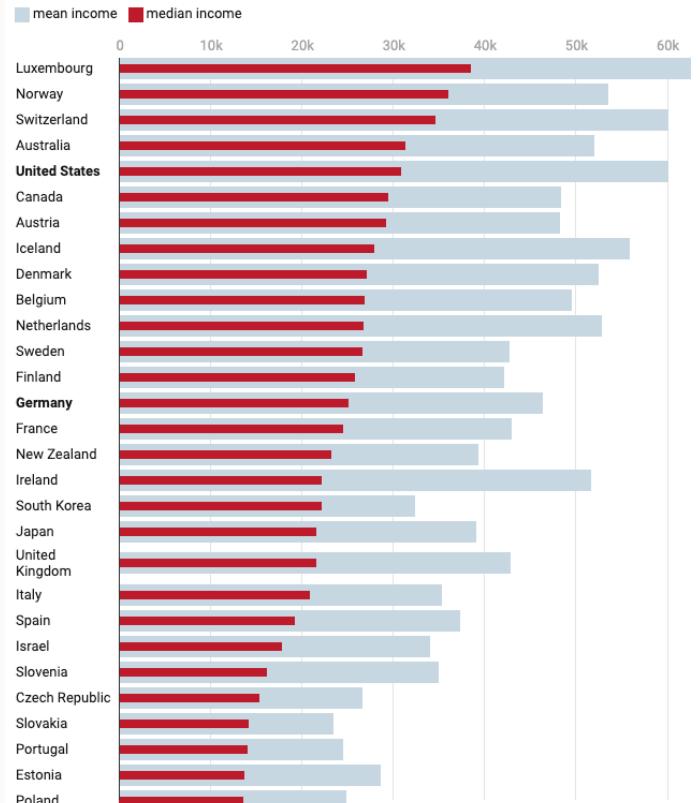


Chart: Lisa Charlotte Rost, Datawrapper • Source: [OECD](#) • [Get the data](#) • Created with Datawrapper

## Warum brauchen wir Streuungsmaße?

- Die zentrale Tendenz allein ist nicht aussagekräftig für die Verteilung
- „Wie weit streuen unsere Daten?“

## Drei gängige Streuungsmaße

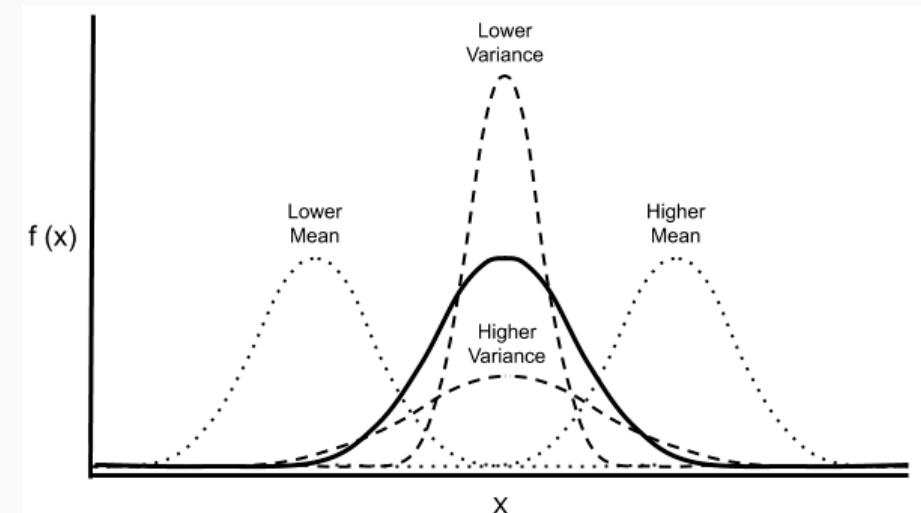
- **Spannweite:** Die Differenz zwischen dem höchsten und dem niedrigsten Wert in einem Datensatz
- **Varianz:** Der Durchschnitt der quadrierten Differenzen vom Mittelwert
- **Standardabweichung:** Die Quadratwurzel aus der Varianz

Formel zur Berechnung der Varianz:  $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

Siehe [hier](#) für interaktive Intuition.

## Warum ist das wichtig?

- Die meisten Menschen sind in der Tat nicht „durchschnittlich“. Variation kann eine Quelle für Einblicke in die zugrunde liegenden Prozesse sein
- Schlüsselmaß für nachgelagerte Statistiken, z.B. der Standardfehler als Schätzung der Unsicherheit einer Schätzung



## Wie können alle der folgenden Punkte wahr sein?<sup>1</sup>

1. 80% der 100 prominentesten deutschen TikToker sind männlich.
2. Weibliche TikToker haben im Durchschnitt 500 Follower, männliche nur 300.
3. Es gibt ungefähr gleich viele männliche und weibliche deutsche TikToker.



<sup>1</sup> „Wahr“ im Sinne von ‚theoretisch wahr‘. Die Zahlen sind alle erfunden.

## Wie können alle der folgenden Punkte wahr sein?<sup>1</sup>

1. An einer Universität ist die Zulassungsquote in jedem der vier Fachbereiche für Frauen höher als für Männer.
2. Über alle Fachbereiche hinweg ist die Zulassungsquote bei den Männern höher.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	<b>825</b>	62%	108	82%
B	585	63%	<b>560</b>	63%	25	68%
C	792	34%	417	33%	375	35%
D	714	6%	<b>373</b>	6%	341	7%
Total	<b>3024</b>	<b>39%</b>	<b>2175</b>	<b>47%</b>	<b>849</b>	<b>31%</b>

Legend:

 greater percentage of successful applicants than the other gender

 greater number of applicants than the other gender

**bold** - the two 'most applied for' departments for each gender

## Paradoxon erklärt

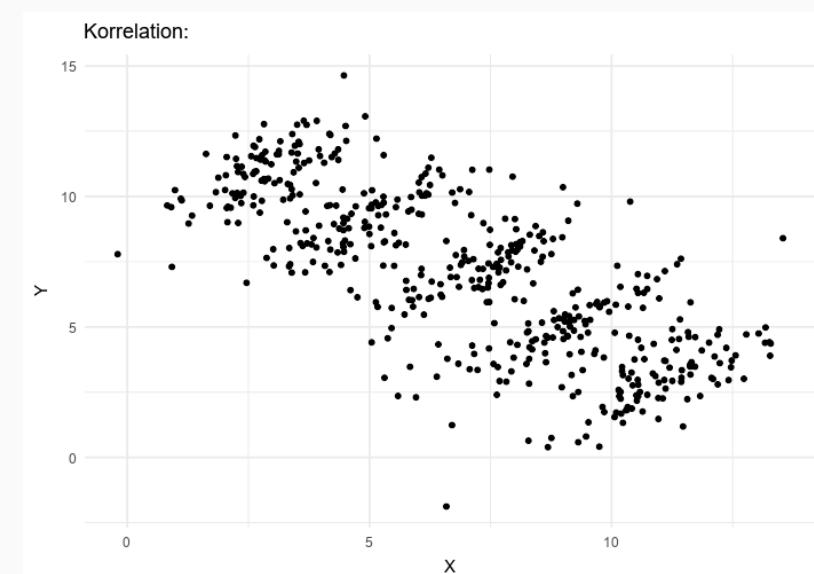
- Einige Fachbereiche (C+D) waren umkämpfter als andere, und dort bewarben sich mehr Frauen.
- Die allgemeinen Zulassungsquoten und die Zulassungsquoten innerhalb eines Fachbereichs unterscheiden sich!
- Wenn wir die gruppierende Variable anpassen/kontrollieren, ändert sich die Beziehung zwischen den Variablen.

## Das Phänomen, verallgemeinert

- Ein Trend erscheint in verschiedenen Datengruppen, verschwindet aber oder kehrt sich um, wenn diese Gruppen kombiniert werden.
- Auch bei Korrelationen möglich (positive vs. negative Korrelation innerhalb vs. zwischen Gruppen).

## Relevanz im Policy-Kontext

- Analyse von Mustern auf verschiedenen Ebenen (z.B. Region vs. Bund, Schulen vs. Schulbezirke)
- Wenn gruppeninterne Muster nicht berücksichtigt werden, könnten die politischen Schlussfolgerungen irreführend sein.



Quelle

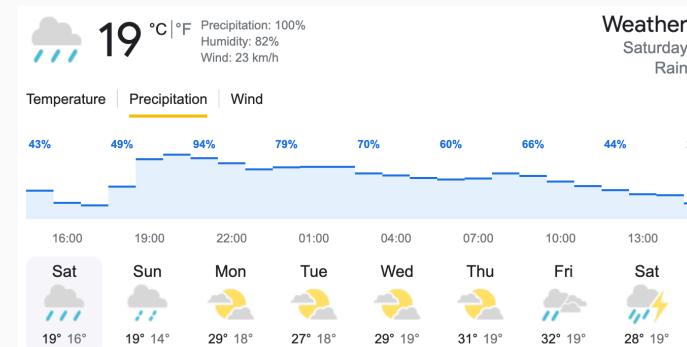
Wikipedia, "Simpson's paradox"

# **Wahrscheinlichkeit verstehen**

---

## Was sind Wahrscheinlichkeiten?<sup>1</sup>

- Wahrscheinlichkeiten quantifizieren die Möglichkeit eines Ereignisses
- Wahrscheinlichkeiten liegen zwischen 0 und 1 (oder 0 und 100%)
- Wahrscheinlichkeiten können verbal oder numerisch kommuniziert werden

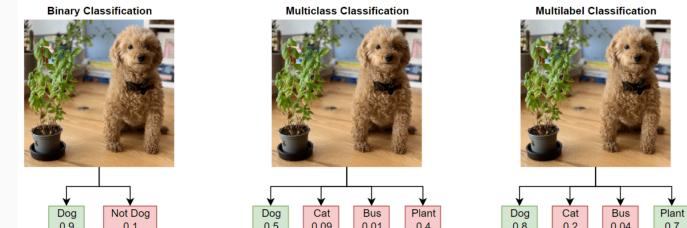


## Relevanz von Wahrscheinlichkeiten für politische Entscheidungen

Wahrscheinlichkeiten sind ...

- ... das Kernstück der Risikobewertung und Entscheidungsfindung
- ... werden zur Quantifizierung von Unsicherheit verwendet
- ... zur Bewertung der Wirksamkeit von Maßnahmen
- ... eine Kernmetrik für AI-basierte Entscheidungssysteme

2024 PRESIDENTIAL ELECTION ODDS & BETTING			
US Presidential Election Futures   Winner			
Donald Trump	1.86	Joe Biden	2.71
Michelle Obama	37	Robert Kennedy Jr.	43
Gavin Newsom	64	Kamala Harris	88



<sup>1</sup>Hier eine gute Einführung in die Wahrscheinlichkeitsrechnung und Simulation.

## Randwahrscheinlichkeit

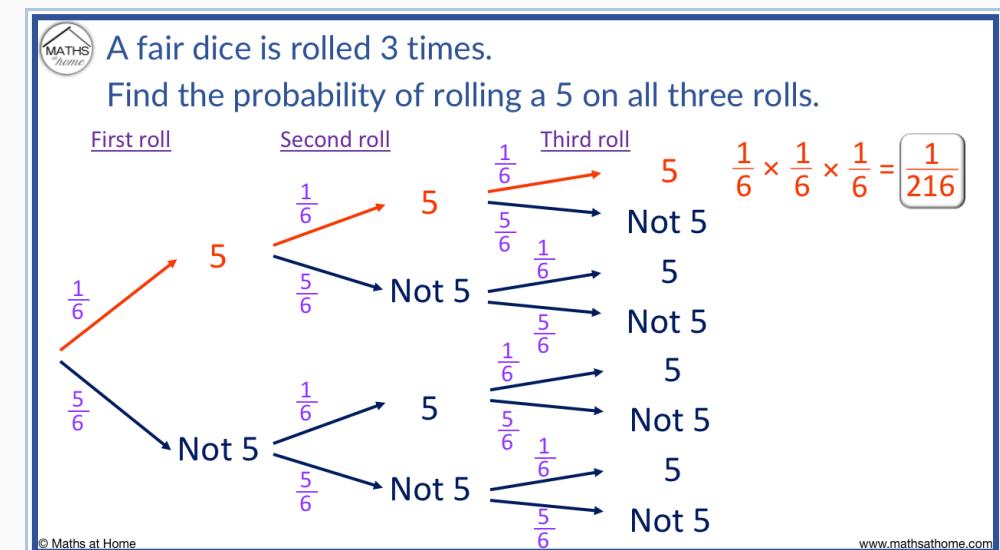
- Die Wahrscheinlichkeit des Eintretens eines Ereignisses:  $p(A)$
- Unbedingte Wahrscheinlichkeit ist nicht von einem anderen Ereignis abhängig
- Beispiel:  $p(\text{Würfeln einer } 5) = 1/6$

## Bedingte Wahrscheinlichkeit

- Die Wahrscheinlichkeit, dass das Ereignis A eintritt, wenn das Ereignis B eintritt:  $p(A|B)$
- Wichtig: Die Randwahrscheinlichkeit von B spielt hier keine Rolle!
- Beispiel:  
 $p(\text{Würfeln einer } 5 | \text{Würfeln ungerader Zahl}) = 1/3$

## Gemeinsame Wahrscheinlichkeit

- Die Wahrscheinlichkeit, dass Ereignis A und Ereignis B eintreten:  $p(A \text{ und } B) = p(A \cap B)$
- Beispiel:  
 $p(\text{Würfeln einer } 5 \text{ und einer geraden Zahl}) = 0$



# Bedingte Wahrscheinlichkeiten

$$P(A) = 0.500 \text{ or } 50.0\%$$

$$P(B) = 0.300 \text{ or } 30.0\%$$

$$P(A \cap B) = 0.150 \text{ or } 15.0\%$$

$$P(B|A) = 0.300 \text{ or } 30.0\%$$

If we have a ball and we know it hit the red shelf, there's a 30.0% chance it also hit the blue shelf.

$$P(A|B) = 0.500 \text{ or } 50.0\%$$

If we have a ball and we know it hit the blue shelf, there's a 50.0% chance it also hit the red shelf.



Quelle Victor Powell, setosa.io (Siehe zur interaktiven Simulation)

## Wie können alle der folgenden Punkte wahr sein?

1. Ein Impfstoff ist hochwirksam beim Schutz gegen eine Krankheit.
2. Die meisten Menschen, die die Krankheit bekommen, sind geimpft worden.



Quelle [Hakan Nural, Unsplash](#)

## Wie können alle der folgenden Punkte wahr sein?

1. Ein Impfstoff ist hochwirksam beim Schutz gegen eine Krankheit.
2. Die meisten Menschen, die die Krankheit bekommen, sind geimpft worden.

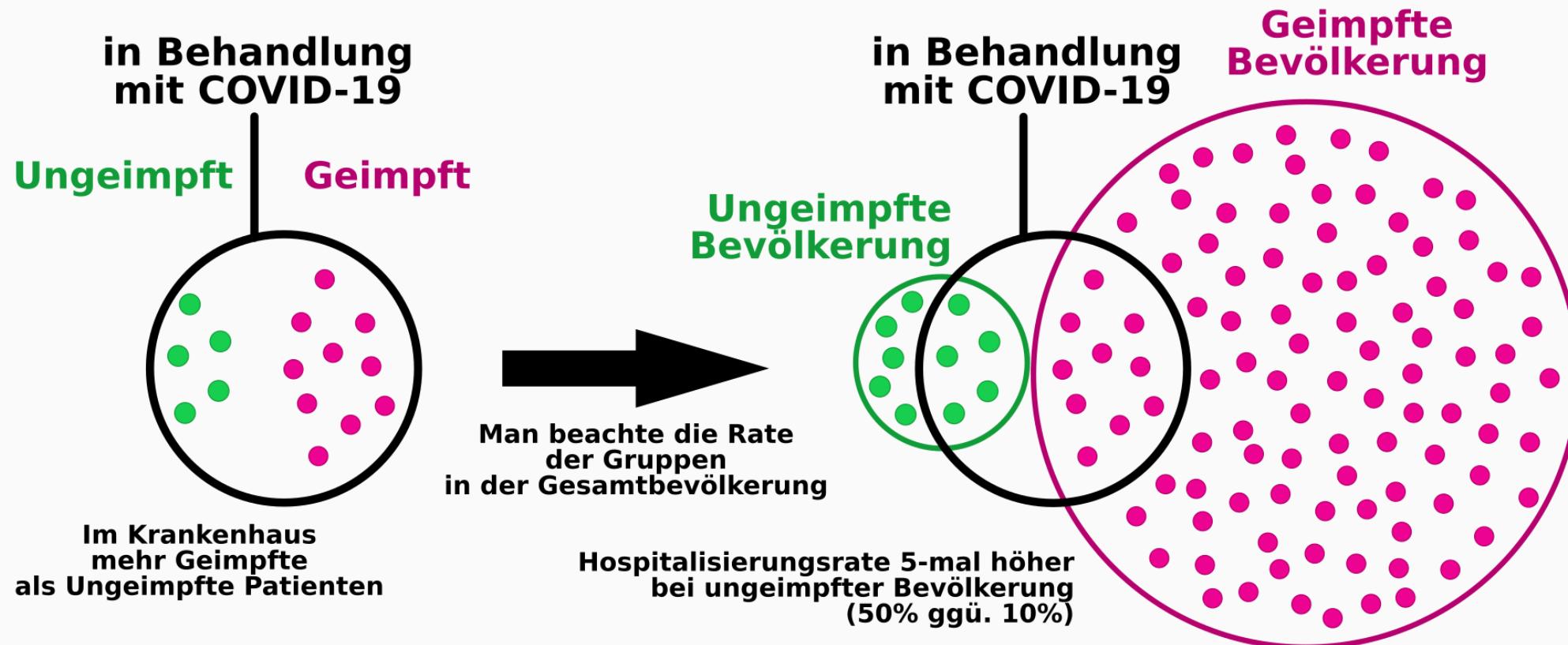
## Base Rate Fallacy ("Prävalenzfehler")

- Dies ist ein klassischer Fall des „Base Rate Fallacy“ oder „Prosecutor's Fallacy“.
- Wenn die Impfrate  $P(\text{geimpft})$  in der Bevölkerung hoch ist, besteht für geimpfte Personen eine viel größere Chance, ins Krankenhaus zu kommen, als für ungeimpfte Personen.



Quelle [Hakan Nural, Unsplash](#)

# "Base rate fallacy", illustriert



Quelle Marc Rumilly

## Politik für seltene Ereignisse

- Diverse Politikmaßnahmen sind darauf ausgerichtet, seltene und gleichzeitig extrem kostspielige Ereignisse zu verhindern.
- Beispiele: Terroranschläge, Krieg, Naturkatastrophen
- Die Vorhersage solcher Ereignisse ist von Natur aus schwierig.
- KI-gestützte Erkennungssysteme versprechen hohe Erkennungsraten. Aber selbst bei einer sehr hohen Genauigkeit kann die Anzahl der falsch-positiven Meldungen inakzeptabel hoch sein.

## Leitfaden

- Berücksichtigen Sie bei der Evaluation von solchen Tools immer die **Basiswahrscheinlichkeit**.
- Wählen Sie geeignete Performanzmetriken (FPR, FNR, etc. - mehr dazu später).

## Beispiel: Terrorismuserkennung

- In einer Stadt mit 1 Mio. Einwohnern gibt es 100 Terroristen und 999.900 Nicht-Terroristen:  
 $p(\text{Terrorist}) = 0,0001$
- Überwachung mit Gesichtserkennungssoftware mit zwei Fehlerquoten von 1%:
  1. Falsch-negativ-Rate (FNR):  
 $p(\text{kein Alarm}|\text{Terrorist}) = 0,01$
  2. Falsch-positiv-Rate (FPR):  
 $p(\text{Alarm}|\text{kein Terrorist}) = 0.01$

Was bedeutet das, wenn wir einen Alarm erhalten?<sup>1</sup>

$$p(\text{Terrorist}|\text{Alarm}) = \frac{p(\text{Alarm}|\text{Terrorist})p(\text{Terrorist})}{p(\text{Alarm})}$$
$$= \frac{0.99 * 0.0001}{0.01} = 0.01$$

<sup>1</sup> $p(\text{Alarm}) = p(\text{Alarm}|\text{Terrorist}) * p(\text{Terrorist}) + p(\text{Alarm}|\text{Nicht-Terrorist}) * p(\text{Nicht-Terrorist})$   
 $= 0.99 * 0.0001 + 0.01 * 0.9999 = 0.01$

# Kommunikation von Wahrscheinlichkeiten mit verbalen Ausdrücken

UCL Institute School



## Variability in the interpretation of probability phrases used in Dutch news articles — a risk for miscommunication

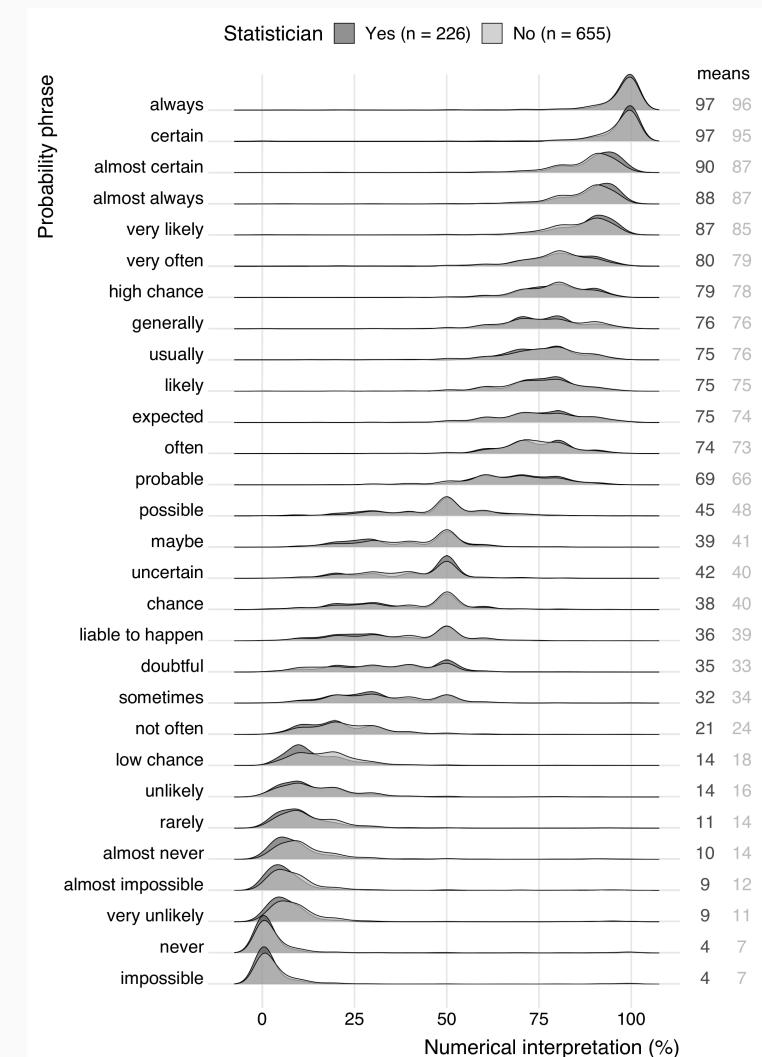
---

**Sanne Willems, Casper Albers and Ionica Smeets**

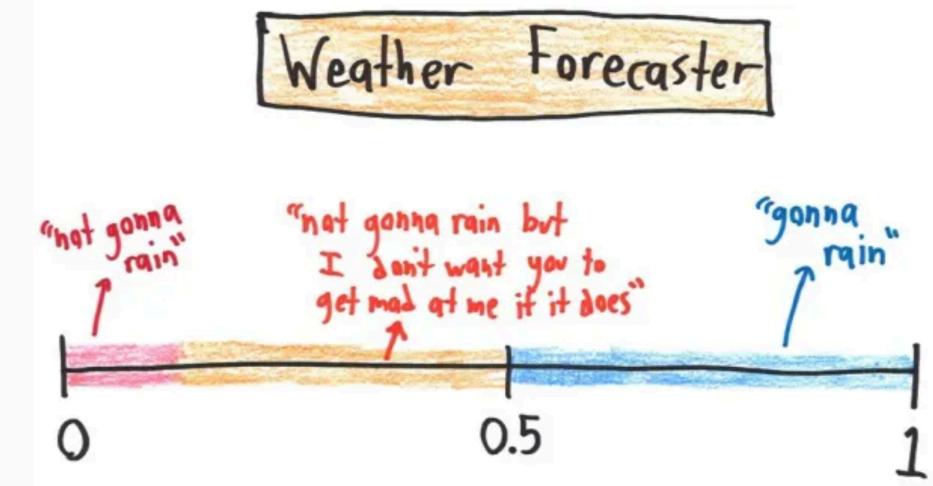
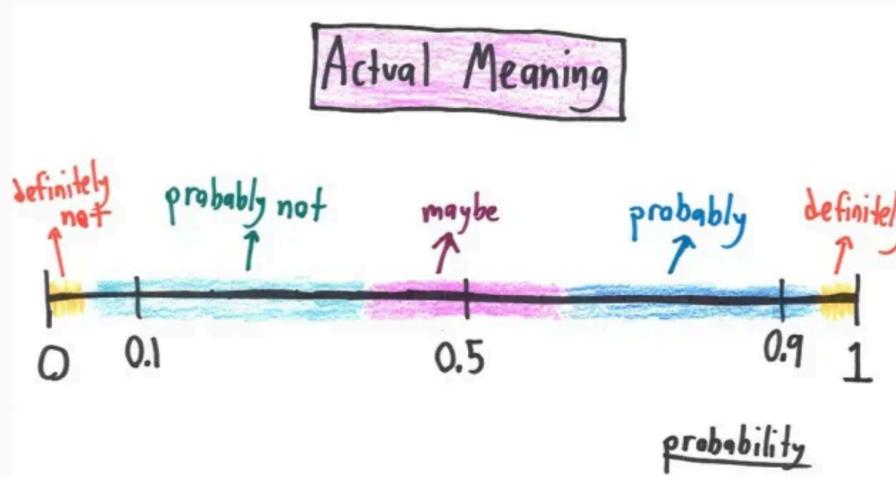
**Abstract** Verbal probability phrases are often used in science communication to express estimated risks in words instead of numbers. In this study we look at how laypeople and statisticians interpret Dutch probability phrases that are regularly used in news articles. We found that there is a large variability in interpretations, even if the phrases are given in a neutral context. Also, statisticians do not agree on the interpretation of the phrases. We conclude that science communicators should be careful in using verbal probability expressions.

**Keywords** Risk communication; Science and media; Science writing

Quelle [Willems et al. 2020](#)



# Was bedeutet die Wahrscheinlichkeit für verschiedene Berufe? Hertie School



# **Statistische Signifikanz verstehen**

---

## Attractive names sustain increased vegetable intake in schools

Brian Wansink <sup>a,\*</sup>, David R. Just <sup>b</sup>, Collin R. Payne <sup>c</sup>, Matthew Z. Klinger <sup>d</sup>

<sup>a</sup> Department of Applied Economics and Management at Cornell University, 15 Warren Hall, Ithaca, NY 14853-7801, USA

<sup>b</sup> Department of Applied Economics and Management at Cornell University, 16 Warren Hall, Ithaca, NY 14853-7801, USA

<sup>c</sup> New Mexico State University, College of Business, MSC 5280, PO Box 30001, Las Cruces, NM 88003-8001, USA

<sup>d</sup> Half Hollow Hills High School East, 50 Vanderbilt Parkway, Dix Hills, NY 11746, USA

### ABSTRACT

*Objective:* This study will determine if the selective use of attractive names can be a sustainable, scalable means to increase the selection of vegetables in school lunchrooms.

*Methods:* Study 1 paired an attractive name with carrots in five elementary schools ( $n=147$ ) and measured selection and consumption over a week compared to controls. Study 2 tracked food sales of vegetables in two elementary schools ( $n=1017$ ) that were systematically attractively named or not named over a two-month period. Both studies were conducted in New York in 2011.

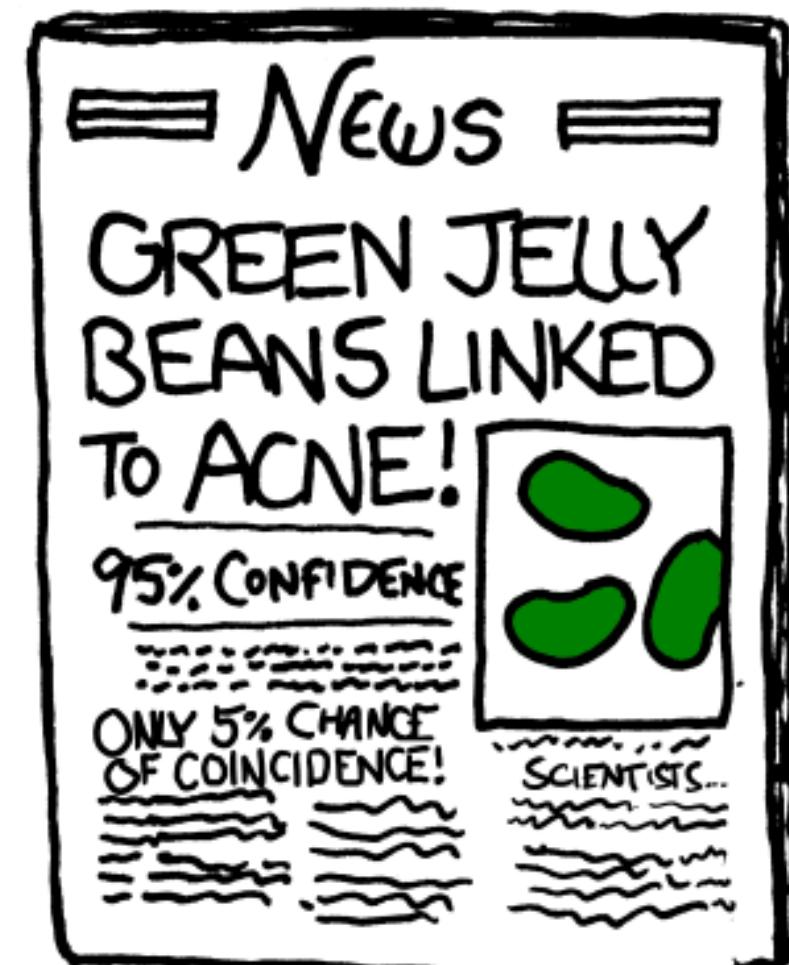
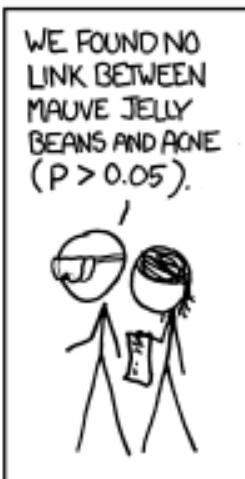
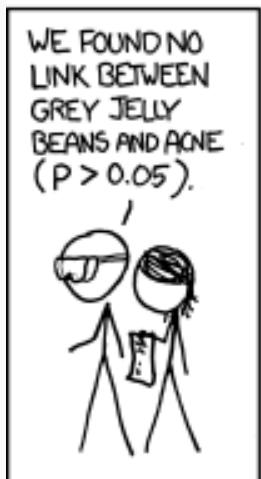
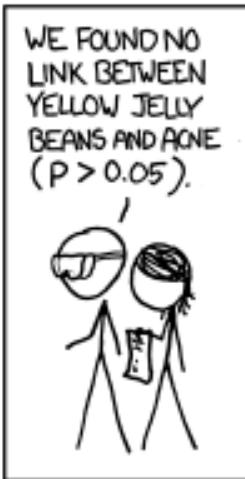
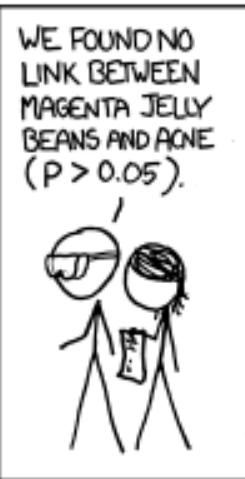
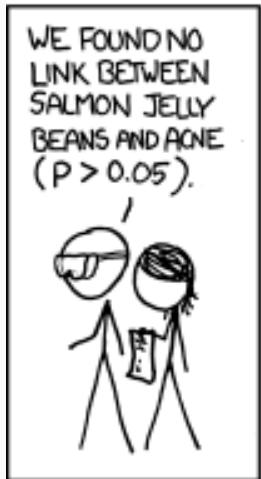
*Results:* Study 1 found that elementary students ate twice the percentage of their carrots if attractively named as "X-ray Vision Carrots," than if un-named or generically named as the "Food of the Day." Study 2 found that elementary school students were 16% more likely to persistently choose more hot vegetable dishes ( $p<0.001$ ) when they were given fun or attractive names.

*Discussion:* Attractive names effectively and persistently increased healthy food consumption in elementary schools. The scalability of this is underscored by the success of Study 2, which was implemented and executed for negligible cost by a high school student volunteer.



Quelle Wansink et al.,  
Retraction Watch

# "Statistische Signifikanz" überall

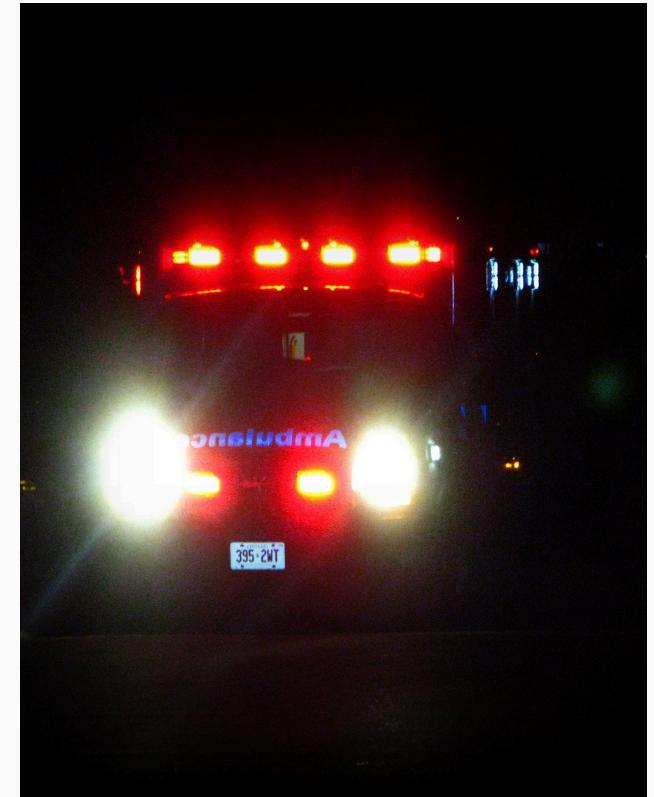


## Beispiel

Sie sind Rettungssanitäter und nähern sich dem Ort eines Autounfalls. Ein Opfer liegt regungslos auf der Straße und Sie müssen einschätzen, ob das Opfer tot oder lebendig ist, und das Opfer entsprechend behandeln.  
Ausgehend von diesen Informationen, **welcher Fehler ist schwerwiegender?**

## Hypothesen

- Nullhypothese: Das Opfer ist am Leben.
- Alternative Hypothese: Das Opfer ist nicht am Leben.



Quelle [jeffalltogether, StackExchange.com](#)

## Hypothesen

- Nullhypothese: Das Opfer ist am Leben.
- Alternative Hypothese: Das Opfer ist nicht am Leben.

## Fehlertypen

- Typ-I-Fehler: Sie verwerfen die Null(hypothese), obwohl sie tatsächlich wahr ist („falsch positiv“).
- Typ II-Fehler: Sie verwerfen die Null nicht, obwohl sie tatsächlich falsch ist. („falsch negativ“)

## Kosten

- **Typ I-Fehler:** Sie erklären das Opfer für tot, obwohl es in Wirklichkeit noch lebt. Das Opfer wird nicht zur möglicherweise lebensrettenden medizinischen Behandlung ins Krankenhaus gebracht → **Extrem kostspieliger Fehler**
- **Typ II-Fehler:** Sie erklären das Opfer für lebendig, obwohl es in Wirklichkeit tot ist. Sie schicken eine tote Person fälschlicherweise mit einem Krankenwagen ins Krankenhaus → **Nicht so kostspieliger Fehler**

# Fehlertypen im Hypothesentesten

		The Truth (Based on Entire Population)	
		Nothing Is There ( $H_0$ Is True)	Something Is There ( $H_0$ Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

## Statistische Signifikanz vs. praktische Signifikanz

- Sie sind nicht dasselbe.
- Bei der **statistischen Signifikanz** geht es um die Wahrscheinlichkeit der Beobachtung der Daten unter Berücksichtigung der Nullhypothese.
- Bei der **praktischen Signifikanz** geht es um die Bedeutung des Ergebnisses in der realen Welt.

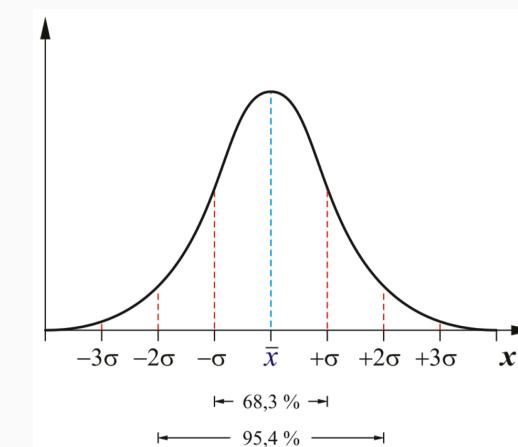
## Vom Hypothesentest zur statistischen Signifikanz

Der dreistufige Ansatz:

1. Formulierung von Null- und Alternativhypotesen.
2. Berechnen Sie eine Teststatistik, zum Beispiel die Effektgröße in einer Regression geteilt durch den Standardfehler.
3. Vergleichen Sie die Teststatistik mit einem kritischen Wert; berechnen Sie einen p-Wert.

## Der p-Wert

- Der p-Wert ist die Wahrscheinlichkeit, dass ein Ergebnis beobachtet wird, das mindestens so extrem ist wie das beobachtete Ergebnis, wenn die Nullhypothese wahr wäre.
- Der p-Wert wird mit einem Schwellenwert (z.B. 0.05) verglichen, um zu entscheiden, ob die Nullhypothese verworfen werden soll.
- Wichtig: Der p-Wert gibt nicht die Wahrscheinlichkeit an, dass die Nullhypothese wahr oder falsch ist!



# Statistische Signifikanz im Auge behalten

	Prominence	Influence
Senate	0.906*** (0.060)	1.483*** (0.067)
Sessions served	0.163*** (0.016)	0.292*** (0.017)
Party (Independent)	0.701* (0.368)	1.059** (0.412)
Party (Republican)	0.035 (0.047)	-0.080 (0.052)
Office: Governor	0.266* (0.158)	0.450** (0.177)
Office: Lt. Governor	-0.031 (0.257)	0.089 (0.288)
Office: US Secretary	0.551** (0.262)	0.372 (0.294)
Position: House Speaker	1.896*** (0.385)	2.670*** (0.431)
Position: Majority/Minority Leader	0.185 (0.308)	0.711** (0.345)
Position: Whip	0.231 (0.233)	0.848*** (0.261)
Position: Deputy Whip	0.698*** (0.234)	0.462* (0.262)
Position: Party Chairman	-0.115 (0.215)	-0.255 (0.241)
(Intercept)	1.648*** (0.050)	1.527*** (0.057)
N	492	492
R-squared	0.493	0.694
Adj. R-squared	0.481	0.687
Residual Std. Error (df = 479)	0.505	0.565
F Statistic (df = 12; 479)	38.890***	90.715***

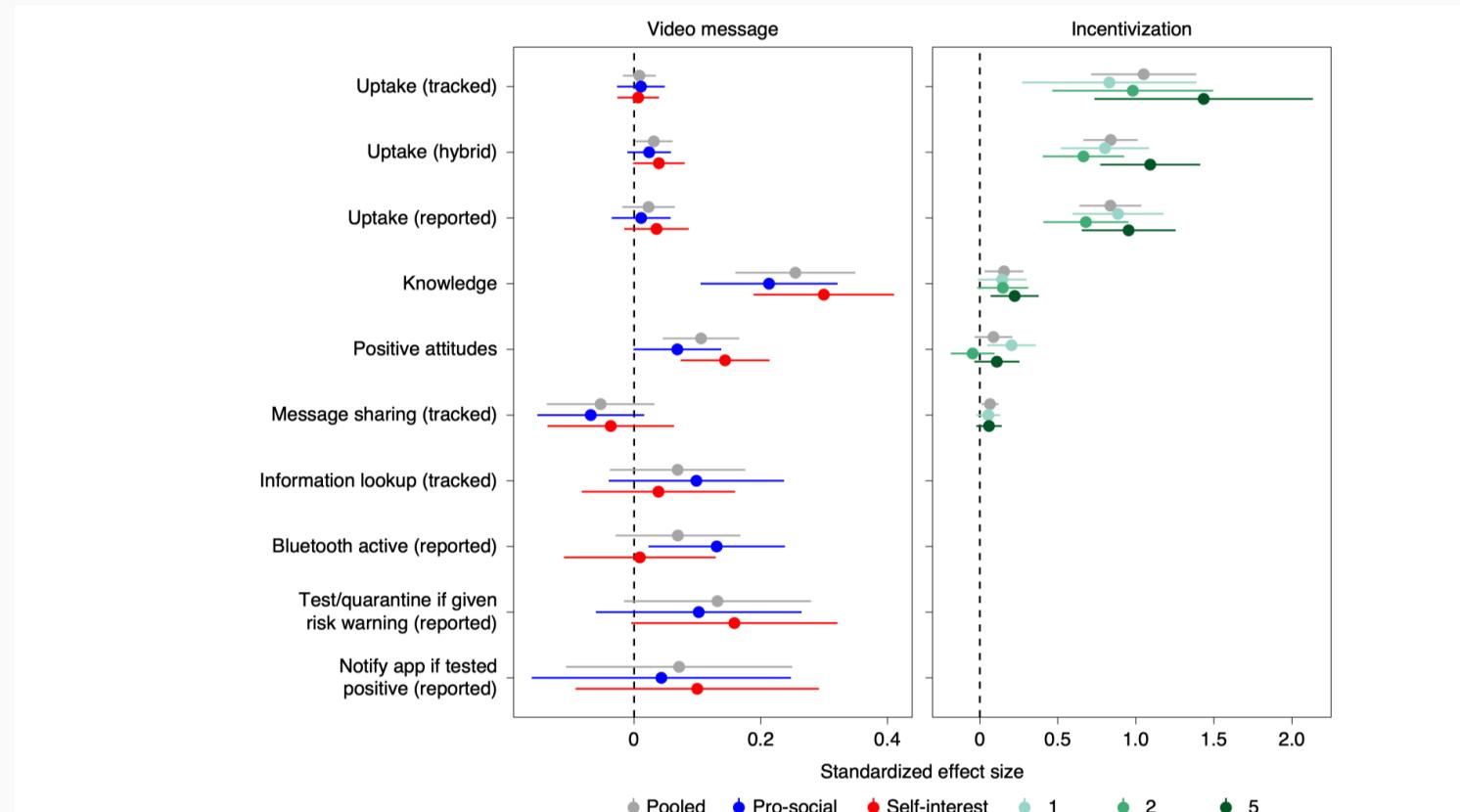
\*\*\* p < .01; \*\* p < .05; \* p < .1

Table 1. Summary statistics

Variable	South (1)	North (2)	Difference in means (3)	Adjusted difference in means (4)	P value (5)
<b>Panel 1: Air pollution exposure at China's Disease Surveillance Points</b>					
TSPs, $\mu\text{g}/\text{m}^3$	354.7	551.6	196.8***	199.5***	<0.001/0.002
SO <sub>2</sub> , $\mu\text{g}/\text{m}^3$	91.2	94.5	3.4	-3.1	0.812/0.903
NO <sub>x</sub> , $\mu\text{g}/\text{m}^3$	37.9	50.2	12.3***	-4.3	<0.001/0.468
<b>Panel 2: Climate at the Disease Surveillance Points</b>					
Heating degree days	2,876	6,220	3,344***	482	<0.001/0.262
Cooling degree days	2,050	1,141	-910***	-183	<0.001/0.371
<b>Panel 3: Demographic features of China's Disease Surveillance Points</b>					
Years of education	7.23	7.57	0.34	-0.65	0.187/0.171
Share in manufacturing	0.14	0.11	-0.03	-0.15***	0.202/0.002
Share minority	0.11	0.05	-0.05	0.04	0.132/0.443
Share urban	0.42	0.42	0.00	-0.20*	0.999/0.088
Share tap water	0.50	0.51	0.02	-0.32**	0.821/0.035
Rural, poor	0.21	0.23	0.01	-0.33*	0.879/0.09
Rural, average income	0.34	0.33	0.00	0.24	0.979/0.308
Rural, high income	0.21	0.19	-0.02	0.27	0.772/0.141
Urban site	0.24	0.25	0.01	-0.19	0.859/0.241
Predicted life expectancy	74.0	75.5	1.54***	-0.24	<0.001/0.811
Actual life expectancy	74.0	75.5	1.55	-5.04**	0.158/0.044

The sample ( $n = 125$ ) is restricted to DSP locations within 150 km of an air quality monitoring station. TSP ( $\mu\text{g}/\text{m}^3$ ) in the years 1981–2000 before the DSP period is used to calculate city-specific averages. Degree days are the deviation of each day's average temperature from 65°F, averaged over the years 1981–2000 before the DSP period. The results in column (4) are adjusted for a cubic of degrees of latitude north of the Huai River boundary. Predicted life expectancy is calculated by OLS using all of the demographic and meteorological covariates shown. All results are weighted by the population at the DSP location. One DSP location is excluded due to invalid mortality data. \*Significant at 10%, \*\*significant at 5%, \*\*\*significant at 1%. Sources: China Disease Surveillance Points (1991–2000), *China Environment Yearbook* (1981–2000), and World Meteorological Association (1980–2000).

# Statistische Signifikanz im Auge behalten



**Fig. 3 | Effect of message and incentive treatments on uptake, knowledge, attitudes and behaviour.** Each plot shows standardized ITT estimates with 95% CIs from fully saturated ordinary least squares regression models fit using the pre-registered LASSO covariate selection procedure. The video message sample comprises  $n=2,044, 1,356$  and  $1,337$  respondents for estimation of the pooled, pro-social and self-interest treatment effects, respectively. The incentive sample comprises  $n=1,015, 513, 516$  and  $494$  respondents for estimation of the pooled, €1, €2 and €5 treatment effects, respectively.

## Einige Probleme

- Nur weil ein Effekt signifikant ist, heißt das nicht, dass er substanzial bedeutsam (groß) ist.
- Es gibt einen Anreiz für Forschende (und Al-Verkäufer), statistisch signifikante Ergebnisse zu produzieren.
- Statistische Signifikanz ist (auch) eine Funktion der Stichprobengröße. Es ist **trivial, mit großen Daten signifikante Ergebnisse zu erzielen**.
- Leider ist es auch oft **trivial, mit kleinen Daten signifikante Ergebnisse zu erzielen**, wenn man in Bezug auf seine Hypothesen flexibel ist.

## Sprachakrobatik mit Signifikanz

"The following list is culled from peer-reviewed journal articles in which (a) the authors set themselves the threshold of 0.05 for significance, (b) failed to achieve that threshold value for p and (c) described it in such a way as to make it seem more interesting." - **Matthew Hankins, Probable Error**

(barely) not statistically significant ( $p=0.052$ ), ..., a certain trend toward significance ( $p=0.08$ ), a clear tendency to significance ( $p=0.052$ ), a clear trend ( $p<0.09$ ), a clear, strong trend ( $p=0.09$ ), very closely brushed the limit of statistical significance ( $p=0.051$ ), very narrowly missed significance ( $p<0.06$ ), very nearly significant ( $p=0.0656$ ), weakly non-significant ( $p=0.07$ ), weakly significant ( $p=0.11$ ), weakly statistically significant ( $p=0.0557$ ), well-nigh significant ( $p=0.11$ )

## Die Krankheit

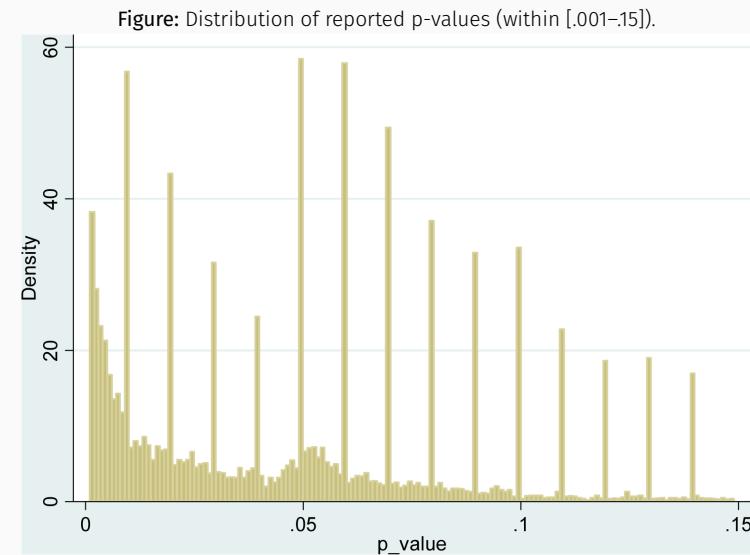
Die Dominanz der statistischen Signifikanz als Entscheidungskriterium bei wissenschaftlichen Veröffentlichungen macht die p-Werte zu einem wichtigen Zielkriterium bei der statistischen Analyse. Kleine p-Werte werden häufiger berichtet, als man erwarten würde!

## Die Symptome

- **Fishing:** Testen vieler Hypothesen bis zur Signifikanz
- **P-Hacking:** Optimieren der Analyse (z. B. Hinzufügen/Entfernen von Kontrollen, Transformieren von Variablen, Ändern von Modellen) bis zur Signifikanz
- **HARKing:** Hypothesizing after the results are known

## Die Kur?

Special issue in *The American Statistician*, 2019: "Statistical Inference in the 21st Century: A World Beyond  $p < 0.05$ "



"The data set consists of over 135'000 records. The data have been harvested by means of computer-based search from (...) five Journals of Experimental Psychology in the period January 1996–March 2008."

THE AMERICAN STATISTICIAN  
2016, VOL. 70, NO. 2, 129–133  
<http://dx.doi.org/10.1080/00031305.2016.1154108>

## The ASA's Statement on *p*-Values: Context, Process, and Purpose

### Sechs Prinzipien

1. *p*-Werte können angeben, wie unvereinbar die Daten mit einem bestimmten statistischen Modell sind.
2. *p*-Werte messen nicht die Wahrscheinlichkeit, dass die untersuchte Hypothese wahr ist, oder die Wahrscheinlichkeit, dass die Daten allein durch Zufall entstanden sind.
3. **Wissenschaftliche Schlussfolgerungen und geschäftliche oder politische Entscheidungen sollten nicht nur darauf beruhen, ob ein *p*-Wert einen bestimmten Schwellenwert überschreitet.**
4. Eine korrekte Schlussfolgerung erfordert eine vollständige Berichterstattung und Transparenz.
5. Ein *p*-Wert sagt nichts über die Größe eines Effekts oder die Bedeutung eines Ergebnisses aus.
6. Ein *p*-Wert an sich ist kein guter Maßstab für die Evidenz eines Modells oder einer Hypothese.

## Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland<sup>1</sup> · Stephen J. Senn<sup>2</sup> · Kenneth J. Rothman<sup>3</sup> · John B. Carlin<sup>4</sup> · Charles Poole<sup>5</sup> · Steven N. Goodman<sup>6</sup> · Douglas G. Altman<sup>7</sup>

Received: 9 April 2016/Accepted: 9 April 2016/Published online: 21 May 2016  
© The Author(s) 2016. This article is published with open access at Springerlink.com

**Abstract** Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific

---

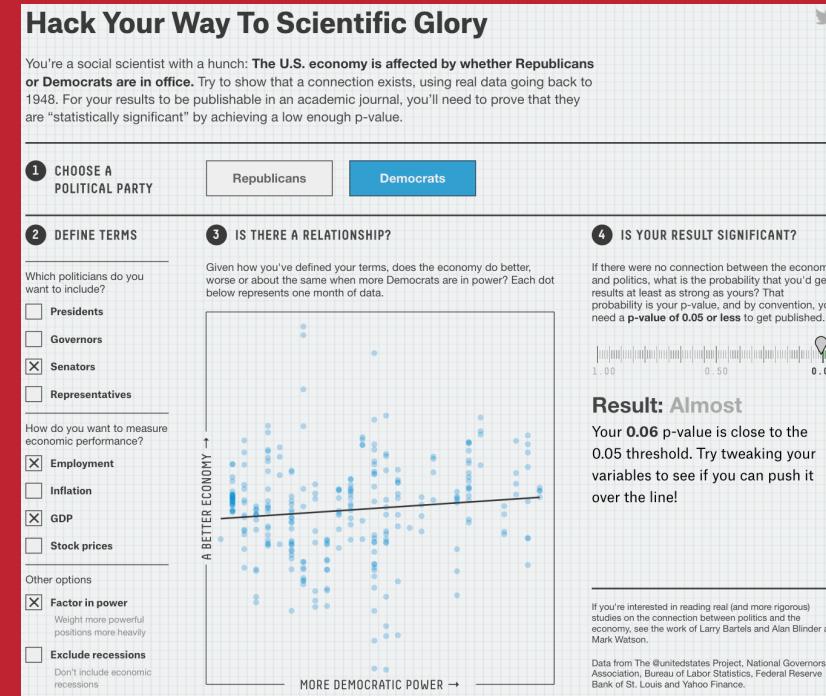
**Editor's note** This article has been published online as supplementary material with an article of Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process and purpose. *The American Statistician* 2016.

---

literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

# Übung

Erlangen Sie wissenschaftlichen Ruhm durch 5 Minuten P-Hacking bei [https://projects.fivethirtyeight.com/p-hacking/!](https://projects.fivethirtyeight.com/p-hacking/)  
(Mehr Hintergrund [hier.](#))



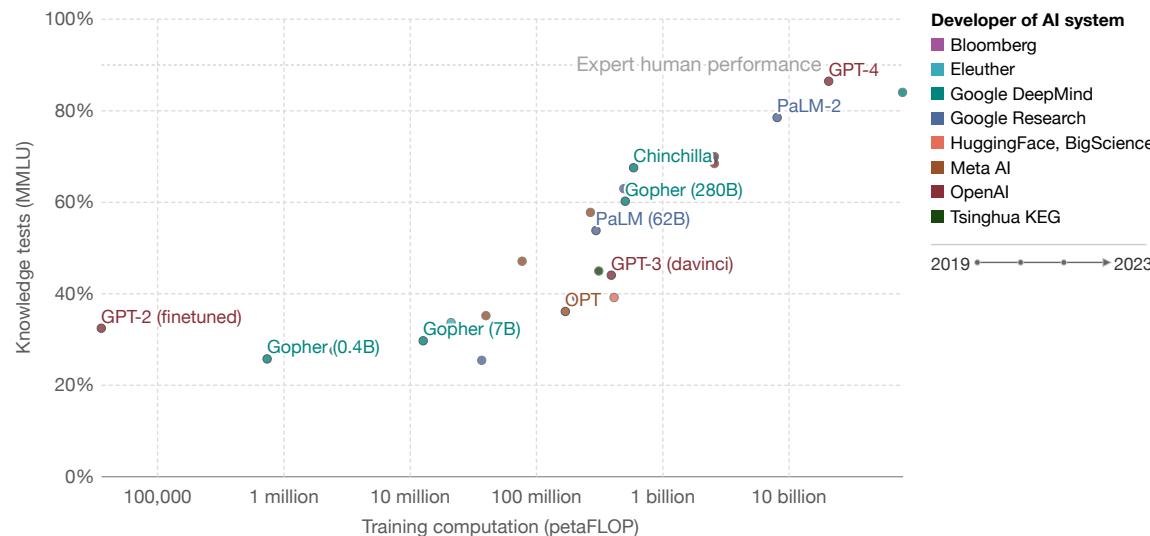
# Klassifikationsmetriken verstehen

---

## Artificial intelligence: Performance on knowledge tests vs. training computation

Our World in Data

Performance on knowledge tests is measured with the MMLU benchmark<sup>1</sup>, here with 5-shot learning, which gauges a model's accuracy after receiving only five examples for each task. Training computation is measured in total petaFLOP, which is  $10^{15}$  floating-point operations<sup>2</sup>.



Data source: Epoch (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: The values for training computation are estimates and come with some uncertainty, especially for models for which only minimal information has been disclosed, such as GPT-4.

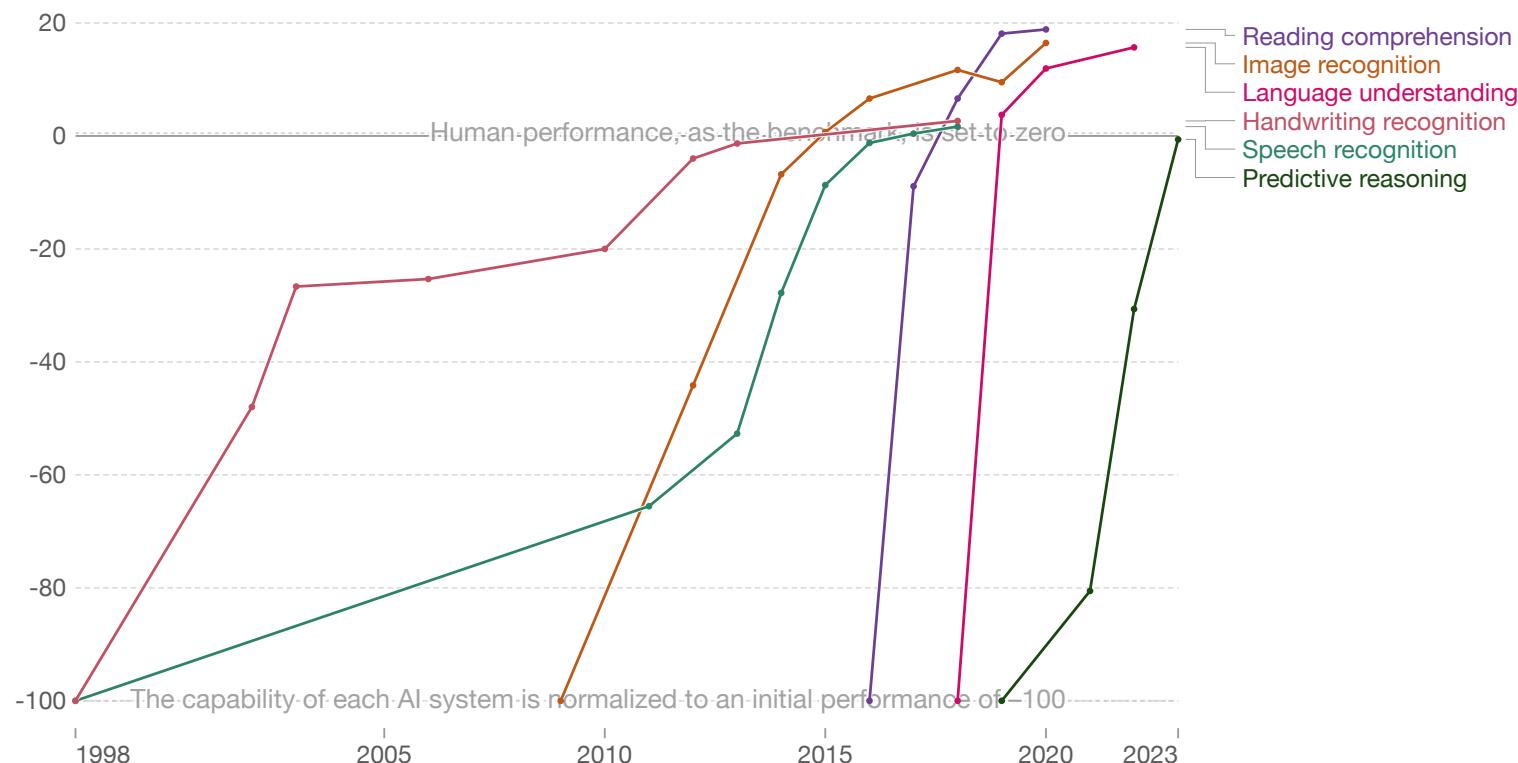
**1. MMLU benchmark:** The Massive Multitask Language Understanding (MMLU) benchmark mimics a multiple-choice knowledge quiz designed to gauge how proficiently AI systems can comprehend various topics like history, science, or psychology. It has 57 different sections, each one looking at a particular subject. The MMLU test has 15,908 questions in total, which are split up into smaller sets. There are at least 100 questions about each subject. The questions in the test come from many places, like practice tests for big exams or questions from university courses. The difficulty of the questions varies, some are as easy as elementary school level, while others are as hard as what professionals in a field might know. The scores achieved by humans on this test are largely dependent on their level of expertise in the subject matter. Individuals who are not specialists in a given area typically achieve a correctness rate of around 34.5%. However, those with a deep understanding and proficiency in their field, such as doctors sitting for a medical examination, can attain a high score of up to 89.8% on the test.

**2. Floating-point operation:** A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

## Test scores of AI systems on various capabilities relative to human performance

Our World  
in Data

Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

[OurWorldInData.org/artificial-intelligence](https://OurWorldInData.org/artificial-intelligence) | CC BY

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

## Klassifizierungsprobleme in freier Wildbahn

Klassifizierungsprobleme sind weit verbreitet (wahrscheinlich häufiger als Regressionsprobleme). **Beispiele:**

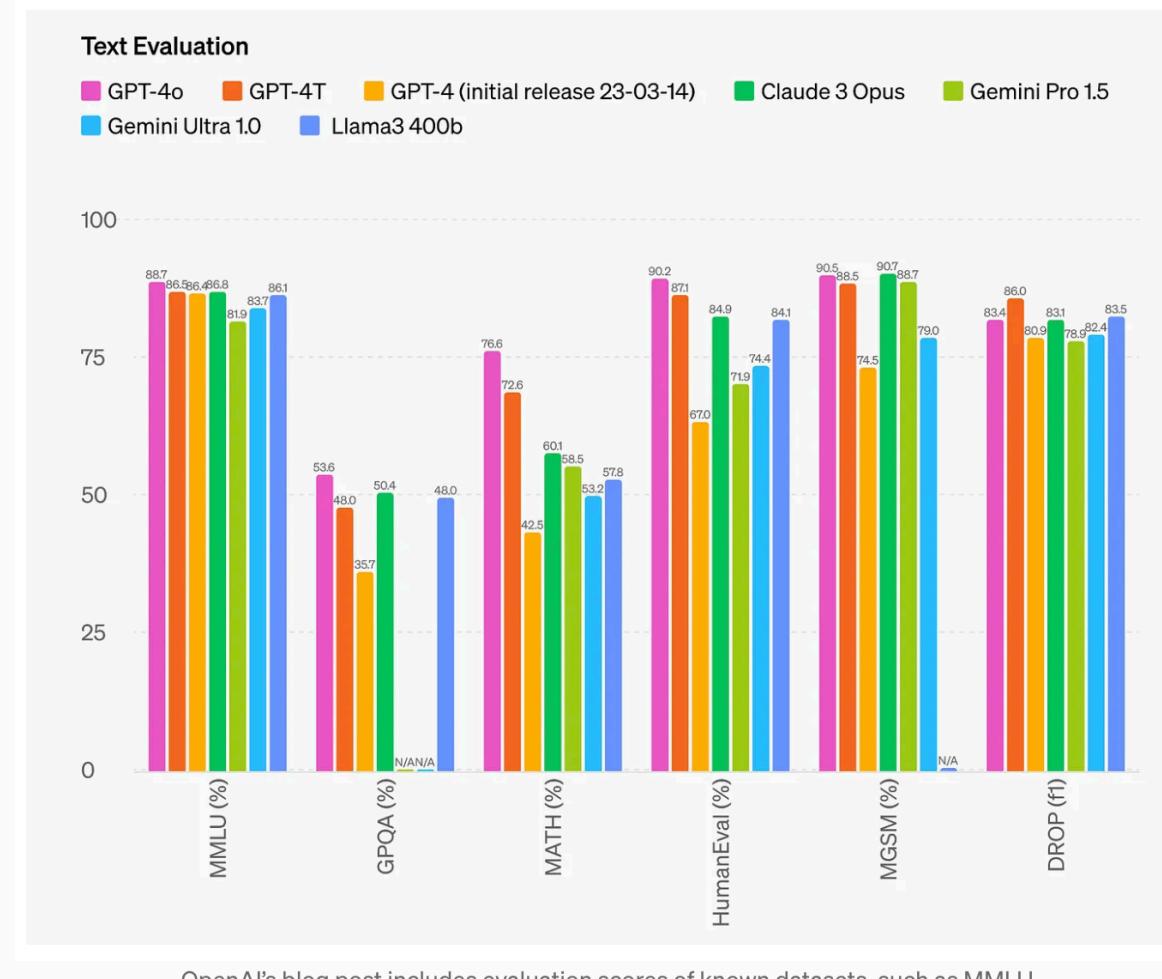
1. Eine Frau kommt mit einer Reihe von Symptomen in die Notaufnahme. Welche Krankheit hat sie?
2. Ein Online-Banking-Dienst muss in der Lage sein, anhand der IP-Adresse des Benutzers, seiner bisherigen Transaktionen usw. festzustellen, ob eine Transaktion betrügerisch ist oder nicht.
3. Eine Biologin möchte anhand der DNA-Sequenzdaten einer Reihe von Patienten mit und ohne eine bestimmte Krankheit herausfinden, welche DNA-Mutationen schädlich (krankheitsverursachend) sind und welche nicht.
4. Ein Moderator einer Social-Media-Plattform muss entscheiden, ob ein von einem Nutzer erstellter Beitrag als unangemessen gekennzeichnet werden soll oder nicht.

## Automatisierte Entscheidungsfindung

Entscheidungsfindungsprobleme sind oft Klassifizierungsprobleme!

Wenn automatisierte Systeme zur Entscheidungsfindung eingesetzt werden, beruhen sie häufig auf Klassifizierungsmodellen. Ob sie als rechtmäßig angesehen werden oder nicht, hängt oft entscheidend von der Leistung dieser Modelle ab.

# ML-Performanz-Benchmarking: Beispiel



# ML-Performanz-Benchmarking: Beispiel 2

Dataset	Moderation Service	ROC AUC	F1	FPR	FNR
ToxiGen	Amazon	70.4%	68.9%	7.2%	52.0%
	Google	62.7%	62.7%	39.1%	35.5%
	OpenAI	70.3%	68.1%	33.2%	56.0%
	Microsoft	59.8%	57.4%	16.4%	64.0%
Jigsaw	Amazon	92.2%	92.2%	7.5%	8.1%
	Google	69.9%	67.2%	58.4%	1.8%
	OpenAI	78.6%	78.6%	17.1%	25.6%
	Microsoft	75.8%	75.7%	20.4%	28.1%
MegaSpeech	Amazon	72.8 %	72.0 %	10.4 %	43.9 %
	Google	73.3 %	72.3 %	41.3 %	12.0 %
	OpenAI	77.1 %	76.7 %	8.4 %	37.3 %
	Microsoft	70.6 %	70.1 %	16.9 %	41.9 %

**TABLE 2:** Performance metrics by moderation service and dataset. The names of moderation services were abbreviated for better readability. ROC AUC is threshold-invariant, while F1, False Positive Rate and False Negative Rate are threshold-variant. Good performance maximises ROC AUC and F1, and minimises False Positive and False Negative Rates. Per dataset, blue shading signals the best performance, while red shading indicates the worst performance. ToxiGen includes 7,800 observations, while Jigsaw and MegaSpeech each contain 50,000. All datasets are balanced on toxic and non-toxic phrases.

## Accuracy ("Korrektklassifizierungsrate")

- Accuracy =  
$$\frac{\text{Anzahl der richtigen Vorhersagen}}{\text{Gesamtzahl der Vorhersagen}} = \frac{TP+TN}{TP+TN+FP+FN}$$
- Falschklassifikationsrate:  $1 - \text{Accuracy}$

## Nützlichkeit

- Die Accuracy ist eine einfache und intuitive Metrik.
- Sie kann jedoch irreführend sein, insbesondere in unausgewogenen Datensätzen, in denen die Klassen nicht gleichmäßig vertreten sind.
- Beispiel: In einem Datensatz mit 90% der Klasse A und 10% der Klasse B hat ein Modell, das alle Fälle als Klasse A vorhersagt, eine Genauigkeit von 90%, ist aber für die Vorhersage von Fällen der Klasse B nicht nützlich.

## Beispiel

		Outcome recidivism: 1 = recidivate, 0 = not recidivate	
		Predicted values	
		0	1
True/actual values		Total	
0		540	224
1		255	424
Total		795	648
			1443

Wie hoch ist die **Accuracy** unseres Rückfall-Klassifizierers?

# Bestimmung der Performanz eines binären Klassifizierers

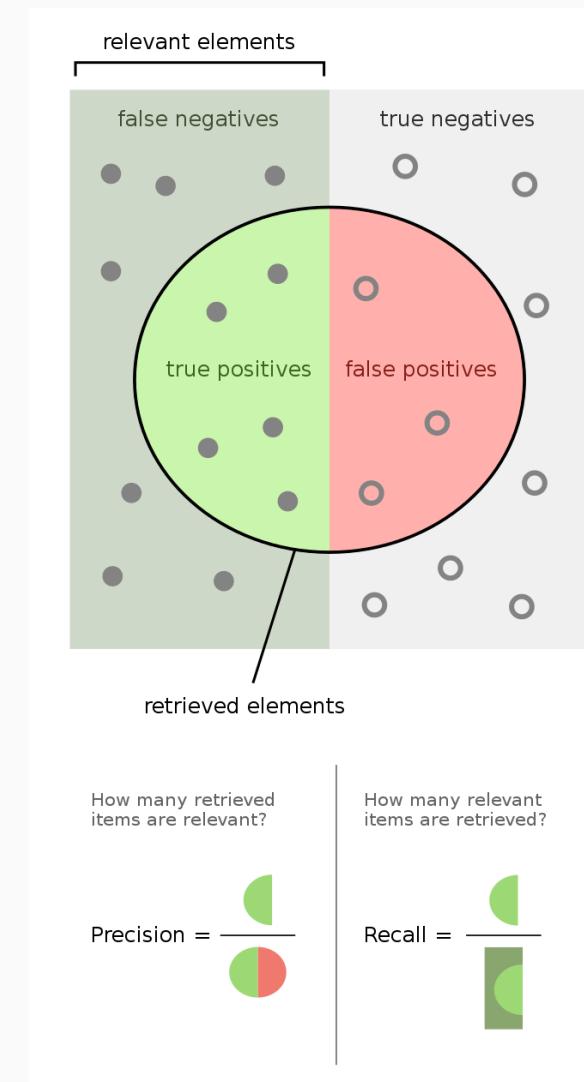
## Präzision

- Präzision =

$$\frac{\text{Anzahl der echten positiven Vorhersagen}}{\text{Anzahl der positiven Vorhersagen}} = \frac{TP}{TP+FP}$$

## Nützlichkeit

- Die Präzision konzentriert sich auf die Genauigkeit der positiven Vorhersagen und ist nützlich, wenn die Kosten für falsch positive Vorhersagen hoch sind.



## Präzision

- Präzision =

$$\frac{\text{Anzahl der echten positiven Vorhersagen}}{\text{Anzahl der positiven Vorhersagen}} = \frac{TP}{TP+FP}$$

## Nützlichkeit

- Die Präzision konzentriert sich auf die Genauigkeit der positiven Vorhersagen und ist nützlich, wenn die Kosten für falsch positive Vorhersagen hoch sind.

## Example

Outcome recidivism: 1 = recidivate, 0 = not recidivate		
Predicted values		Total
True/actual values	0	1
0	540	224
1	255	424
Total	795	648
		1443

Wie hoch ist die **Präzision** unseres Rückfall-Klassifizierers?

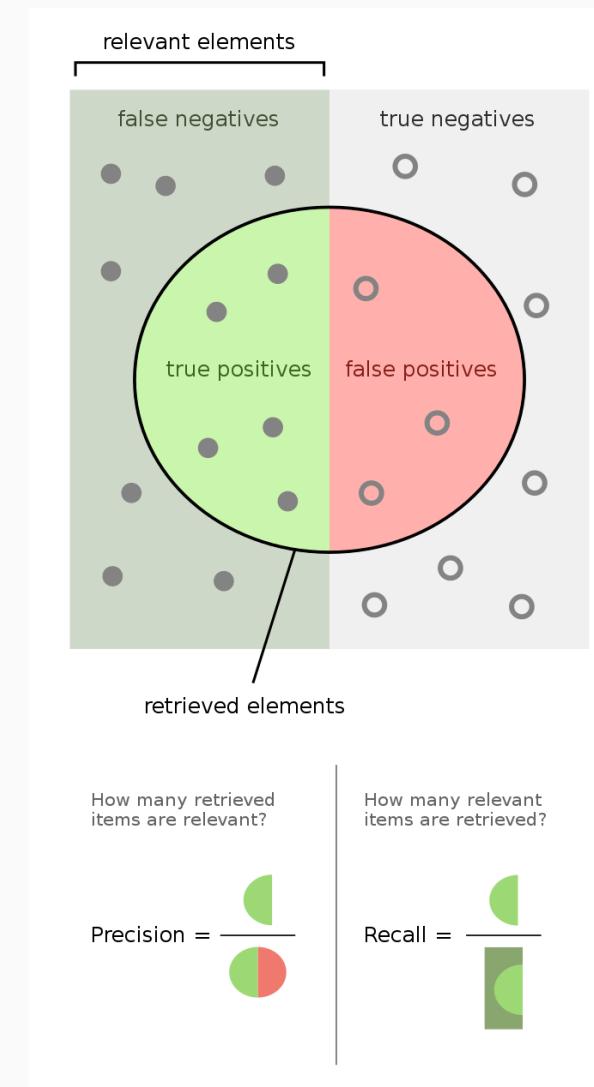
# Bestimmung der Performanz eines binären Klassifizierers

## Recall (Sensitivität)

- Recall =  
$$\frac{\text{Anzahl der wahren positiven Vorhersagen}}{\text{Anzahl der wahren positiven Werte}} = \frac{TP}{TP+FN}$$
- „True-Positive-Rate“

## Nützlichkeit

- Der Recall konzentriert sich auf die Erfassung aller positiven Werte und ist wichtig, wenn die Kosten für Falsch-negativ-Vorhersagen hoch sind.
- **Beispiel:** Bei einer medizinischen Diagnose ist der Recall wichtig, um sicherzustellen, dass alle Patienten mit einer Krankheit korrekt identifiziert werden.
- Das ergänzende Maß ist die Spezifität (Rate der echten Negativbefunde; z. B. wie viele gesunde Personen als nicht erkrankt identifiziert werden).



## Recall (Sensitivität)

- Recall =  
$$\frac{\text{Anzahl der wahren positiven Vorhersagen}}{\text{Anzahl der wahren positiven Werte}} = \frac{TP}{TP+FN}$$
- „True-Positive-Rate“

## Nützlichkeit

- Der Recall konzentriert sich auf die Erfassung aller positiven Werte und ist wichtig, wenn die Kosten für Falsch-negativ-Vorhersagen hoch sind.
- **Beispiel:** Bei einer medizinischen Diagnose ist der Recall wichtig, um sicherzustellen, dass alle Patienten mit einer Krankheit korrekt identifiziert werden.
- Das ergänzende Maß ist die Spezifität (Rate der echten Negativbefunde; z. B. wie viele gesunde Personen als nicht erkrankt identifiziert werden).

## Beispiel

		Outcome recidivism: 1 = recidivate, 0 = not recidivate	
		Predicted values	
		0	1
True/actual values		Total	
0		540	224
1		255	424
Total		795	648
			1443

Wie hoch ist der Recall unseres Rückfall-Klassifizierers?

## F1-Score

- F1 score =  $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$
- Der F1-Score ist das harmonische Mittel aus Präzision und Recall.

## Nützlichkeit

- Der F1-Score berichtet eine Balance zwischen Präzision und Recall, was insbesondere dann hilfreich ist, wenn ein Ungleichgewicht zwischen den Klassen besteht.
- Der F1-Score reicht von 0 bis 1, wobei 1 für perfekte Präzision und Recall und 0 für eine schlechte Leistung steht.

## Illustration

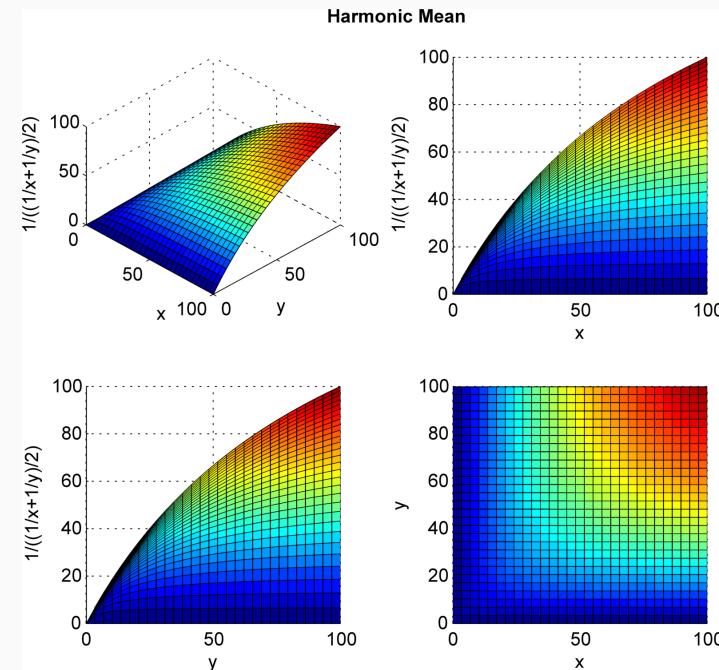


Abbildung des normalisierten harmonischen Mittels, wobei x die Präzision, y den Recall und die vertikale Achse den F1-Score in %-Punkten darstellt.

# Bestimmung der Performanz eines binären Klassifizierers

## F1-Score

- F1 score =  $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$
- Der F1-Score ist das harmonische Mittel aus Präzision und Recall.

## Nützlichkeit

- Der F1-Score berichtet eine Balance zwischen Präzision und Recall, was insbesondere dann hilfreich ist, wenn ein Ungleichgewicht zwischen den Klassen besteht.
- Der F1-Score reicht von 0 bis 1, wobei 1 für perfekte Präzision und Recall und 0 für eine schlechte Leistung steht.

## Beispiel

Outcome recidivism: 1 = recidivate, 0 = not recidivate			
		Predicted values	
		0	1
True/actual values		Total	
0		540	224
1		255	424
Total		795	648

Wie hoch ist der **F1-Score** unseres Rückfall-Klassifizierers?

## Szenario

- **Outcome:** Rückfälligkeit, wenn die Person rückfällig wird (1) oder nicht (0)
- **Falsch positiv (FP):** Das Modell sagt voraus, dass eine Person rückfällig wird, wenn sie es tatsächlich nicht wird.
- **Falsch Negativ (FN):** Das Modell sagt voraus, dass eine Person nicht rückfällig wird, obwohl sie es tatsächlich wird.

## Kosten

- Was sind die nachgelagerten Kosten von FP und FN?
- FP könnte dazu führen, dass einer Person Maßnahmen auferlegt werden, die nicht notwendig wären
- FN könnte dazu führen, dass gefährdete Personen ohne angemessene Intervention entlassen werden, was möglicherweise zu erneuten Straftaten führt.
- Auf welcher Ebene fallen die Kosten an - individuell, gesellschaftlich, ...?

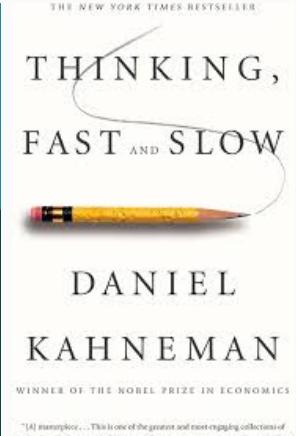
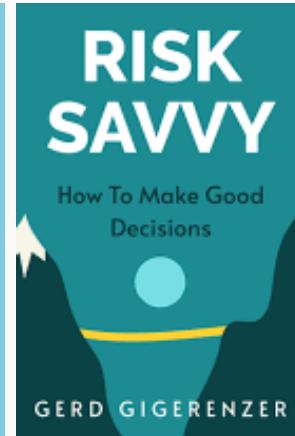
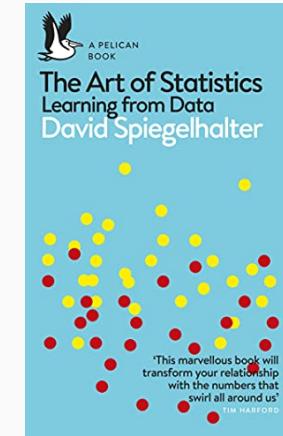
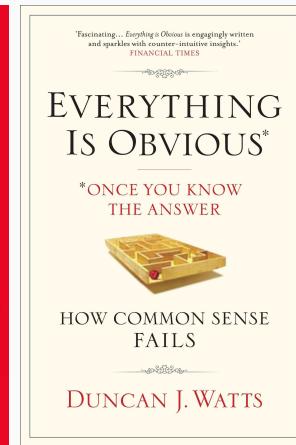
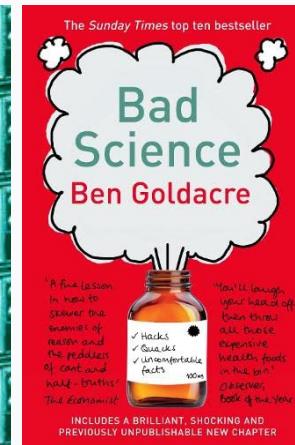
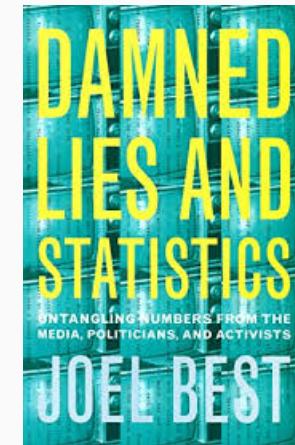
## Ethische und wirtschaftliche Überlegungen

- Wie sollten wir die Kosten von FP und FN abwägen?
- Was sollte in unserem Modell Vorrang haben - die Verringerung von FP, FN oder ein Kompromiss zwischen beiden?

# Statistiken verwenden: lessons learned

## Lies, damned lies and statistics

- Politische Debatten drehen sich fast zwangsläufig auch um Statistiken
- Strategischer Anreiz, Beweise zu seinen Gunsten zu beeinflussen
- Statistische Fallstricke: Nicht alles, was logisch klingt, ist statistisch fundiert
- Ein grundlegendes Verständnis der grundlegenden Konzepte der Statistik ist der Schlüssel zu kritischem Konsum von Daten und Statistiken
- Einige populäre Irrtümer und Fehler treten immer wieder auf - man kann trainieren, sie zu erkennen!



## "OptiClassify: Effiziente Aktenkategorisierung – Präzision, die Sie brauchen!"

Mit OptiClassify bieten wir eine KI-basierte Lösung, die Verwaltungsakten automatisch in 5 spezifische Kategorien einteilt. Unsere Technologie erreicht eine statistisch signifikante Trefferquote (Accuracy) von 95%, unterstützt durch modernste Machine-Learning-Algorithmen.

Die Falsch-Negative-Rate liegt bei 5%, was bedeutet, dass nur wenige Dokumente in die falsche Kategorie eingeordnet werden. Dank der hohen Signifikanz von  $p < 0.01$  können Sie sich auf die Zuverlässigkeit unserer Lösung verlassen, insbesondere bei großen Datenmengen.

In Tests mit über 100.000 realen Verwaltungsakten haben wir bewiesen, dass OptiClassify selbst in komplexen Szenarien konsistent arbeitet und den manuell erzeugten Kategorisierungsaufwand um 75% reduziert.

Entdecken Sie die Zukunft der Verwaltung mit OptiClassify – sparen Sie Zeit, Kosten und Ressourcen, indem Sie unsere praxiserprobte Lösung einsetzen!



## Mögliche Fragen

1. Auf Basis welcher Daten wurde das Tool trainiert?

☞ Die Qualität und Passung der Trainingsdaten ist entscheidend für die Leistung des Tools im konkreten Kontext.

2. Wie wird die Accuracy von 95% berechnet, und wie verteilt sich diese auf die einzelnen Kategorien?

☞ Die Accuracy könnte in einigen Kategorien viel niedriger sein.

3. Wie wirkt sich die False-Negative-Rate von 5% auf kritische Akten aus?

☞ Es ist wichtig zu wissen, ob Dokumente, die zwingend in eine bestimmte Kategorie gehören, zuverlässig erkannt werden.

☞ Bei einer großen Anzahl von Dokumenten könnten niedrige Fehlerraten zu vielen manuellen Nacharbeiten führen.

4. Wie wurde die statistische Signifikanz ( $p < 0.01$ ) berechnet?

☞ Welche Hypothese liegt hier zugrunde?