

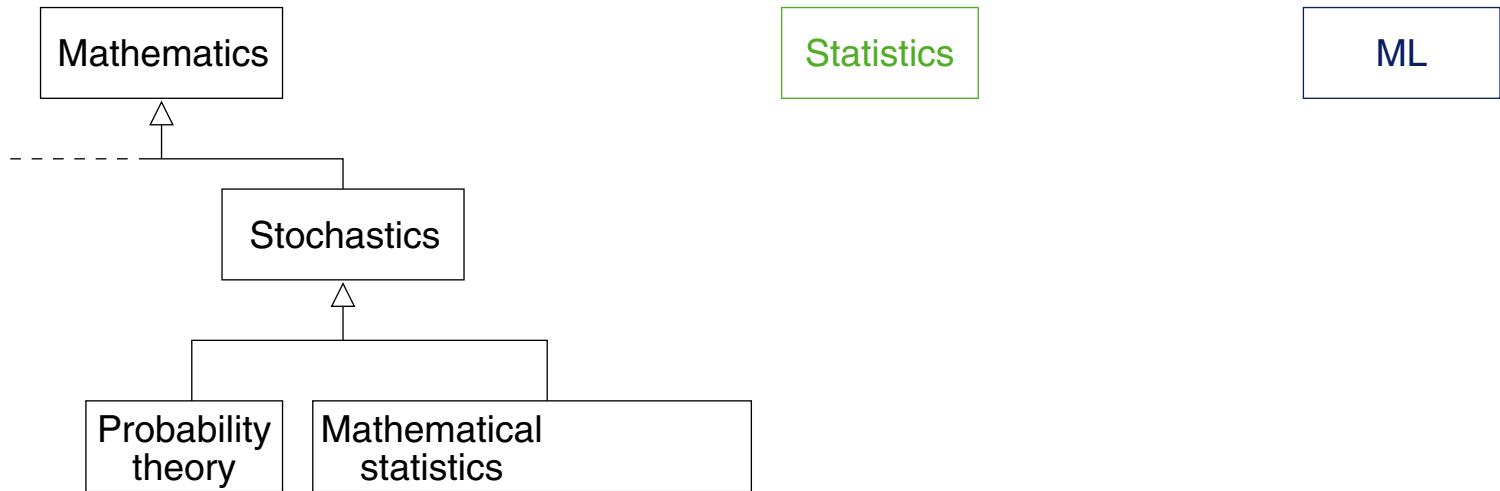
# Kapitel SF:I

## I. Einführung

- Begriffserklärung und Einordnung
- Ausgewählte Anwendungsgebiete
- Geschichte der Statistik
- Philosophie der Statistik
- Missbrauch der Statistik
- Vergleichende Syntax-Übersicht
- Übersetzungen statistischer Fachbegriffe

# Begriffserklärung und Einordnung

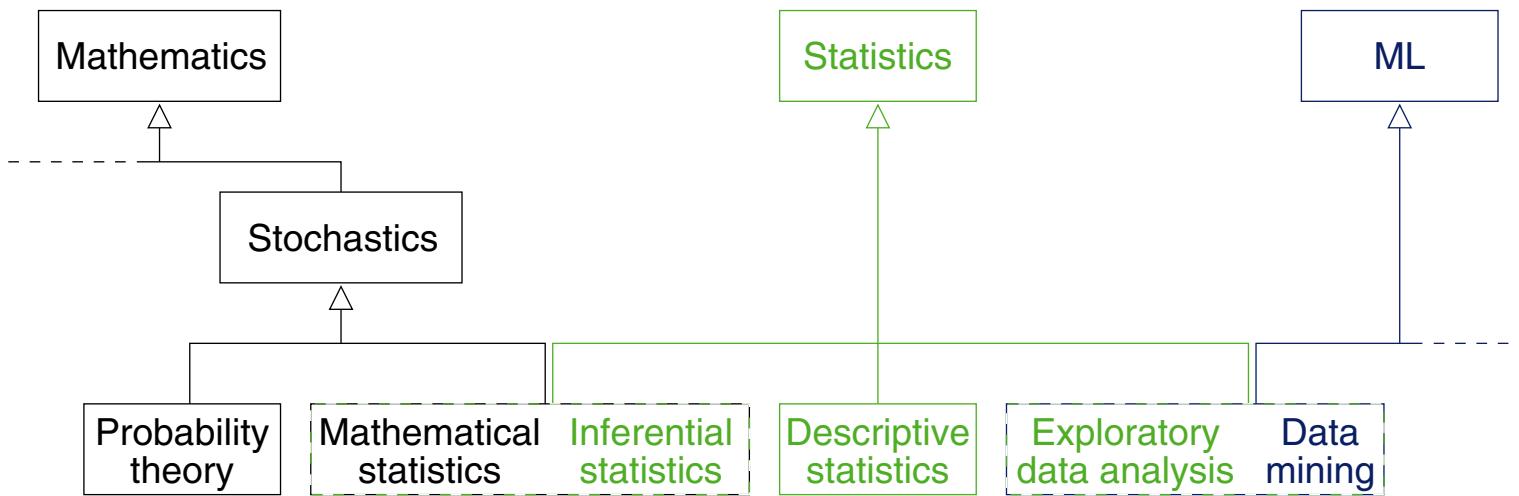
## Überblick



- Wahrscheinlichkeitstheorie: Vorhersage möglicher Ereignisse aus Regeln; Regeln  $\Rightarrow$  Daten.
- Mathematische Statistik: Rückschluss auf Regeln aus Ereignisdaten; Regeln  $\leftarrow$  Daten.

# Begriffserklärung und Einordnung

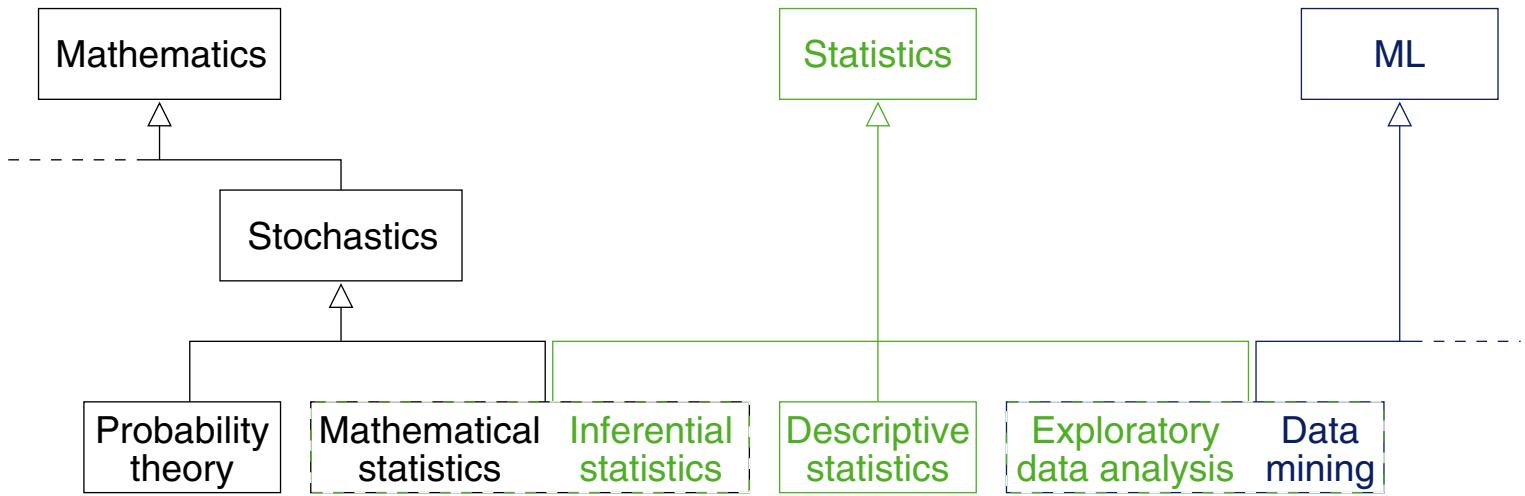
## Überblick



- **Wahrscheinlichkeitstheorie:** Vorhersage möglicher Ereignisse aus Regeln; Regeln  $\Rightarrow$  Daten.
- **Mathematische Statistik:** Rückschluss auf Regeln aus Ereignisdaten; Regeln  $\leftarrow$  Daten.
- **Schließende Statistik:** ebd.
- **Deskriptive Statistik:** Erhebung, Darstellung und Charakterisierung von Daten.
- **Explorative Datenanalyse:** Gewinnung von Einsichten über mögliche Regeln aus Daten.
- **Data Mining:** Algorithmische Erkennung von Mustern in Daten.

# Begriffserklärung und Einordnung

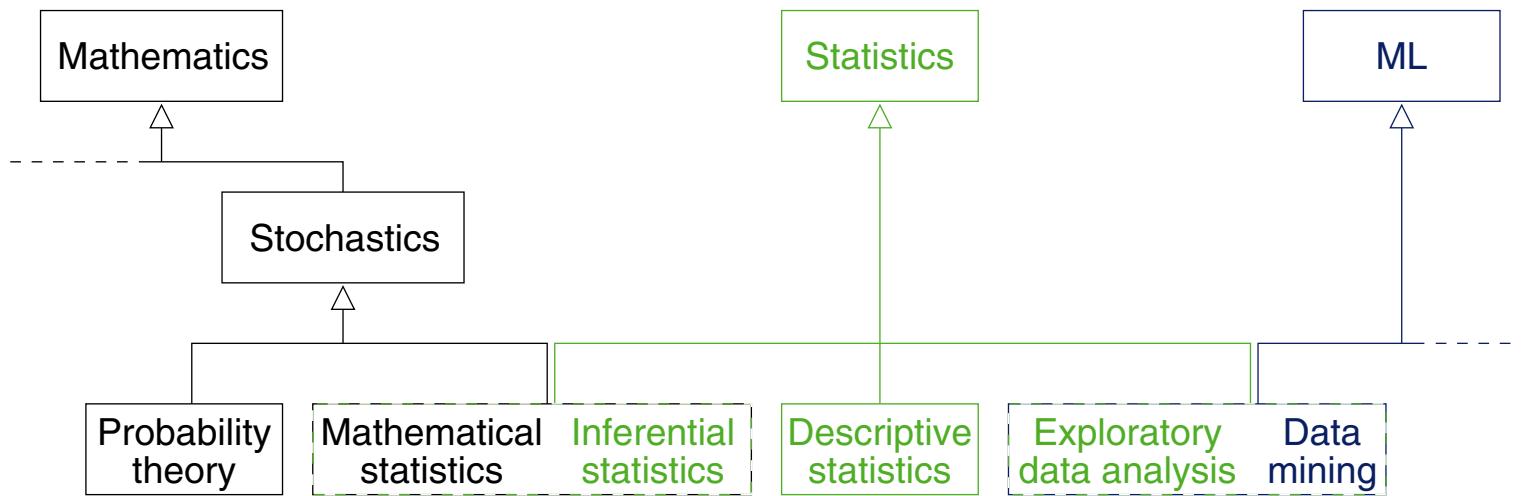
## Überblick



- **Wahrscheinlichkeitstheorie:** zufällige Ereignisse, Zufallsvariablen, stochastische Prozesse
- **Mathematische Statistik:** Schätztheorie, Schätzverfahren, Schätzfunktionen
- **Schließende Statistik:** Konfidenzintervalle, statistische Tests
- **Deskriptive Statistik:** Tabellen, Kennzahlen (Varianz, Erwartungswert)
- **Explorative Datenanalyse:** Box-Plot, Histogramm, Streudiagramm
- **Data Mining:** Anomalieerkennung, Cluster-Analyse

# Begriffserklärung und Einordnung

## Überblick



- Stochastik = Wahrscheinlichkeitstheorie + mathematische Statistik
- Statistik: „Lehre von Methoden zum Umgang mit quantitativen Informationen.“  
[Rinne 2008]
  
- Wahrscheinlichkeitstheorie ist die Grundlage für die Statistik.
- Beides sind die Grundlagen für das Maschinelle Lernen.

## Bemerkungen:

- „Stochastik“, aus dem altgriechischen „στοχαστικὴ τέχνη“ („stochastikē technē“), bedeutet so viel wie „Kunst des Vermutens“. Vermutungen sind mit Unsicherheit verbunden. Die Wahrscheinlichkeitstheorie dient dem Bestimmen des Grades der Unsicherheit. Die mathematische Statistik ist die Anwendung der Wahrscheinlichkeitstheorie.
- Die Wahrscheinlichkeitstheorie „schaut voraus“; man macht Vorhersagen über die Zukunft.
- Die (mathematische, schließende) Statistik „schaut zurück“; man lernt etwas über die Vergangenheit aus aufgezeichneten Daten.
- Die Wahrscheinlichkeitstheorie wird oft mit deduktivem Schließen (Folgerung aus Annahmen), die Statistik mit induktivem Schließen (Verallgemeinerung aus Beispielen) assoziiert.
- „Statistik“ ist vom neulateinischen „statisticum collegium“ („Staatsrat“) sowie dem italienischen „statista“ („Staatsmann“, „Politiker“) abgeleitet und bezeichnet ursprünglich die beschreibende Statistik zum Herausgreifen des Wesentlichen zur Beschreibung der gesellschaftlichen, politischen und wirtschaftlichen Eigenschaften eines Staates.
- Synonyme Begriffe
  - Wahrscheinlichkeitstheorie: Wahrscheinlichkeitsrechnung, Probabilistik
  - Mathematische Statistik: schließende Statistik, Inferenzstatistik, induktive Statistik, beurteilende Statistik
  - Deskriptive Statistik: beschreibende Statistik
  - Explorative Datenanalyse: explorative Statistik

# Ausgewählte Anwendungsgebiete

## Informatik

Probabilistische und statistische Methoden sind fester Bestandteil einer Vielzahl von Bereichen der Informatik.

ACM Computing Classification System (some sub-topics):

- Applied computing: Physical sciences and engineering, Life sciences, Arts and humanities, ...
- Computer systems organization: Distributed architectures, Neural networks, Robotics, ...
- Computing methodologies: Artificial intelligence, Machine learning, Simulation...
- General and Reference: Reliability, Empirical studies, Measurement, Metrics, Evaluation, ...
- Hardware: Power and energy, Simulation and emulation, Reliability, Quantum computation, ...
- Human-centered computing: User models, User studies, Empirical studies, Visualization, ...
- Information systems: Data structures, Query optimization, Data mining, Information retrieval ...
- Mathematics of computing: Combinatorics, Probability and statistics, Optimization, ...
- Networks: Algorithms, Performance evaluation, Monitoring, Peer-to-peer networks, ...
- Security and privacy: Cryptography, Authentication, Anonymity, Intrusion detection, ...
- Social and professional topics: Sustainability, Computer crime, User characteristics, ...
- Software and its engineering: Memory management, Scheduling, Genetic programming, ...
- Theory of computation: Probabilistic computation, Analysis of algorithms, Randomness, ...

# Ausgewählte Anwendungsgebiete

## Informatik

Probabilistische und statistische Methoden sind fester Bestandteil einer Vielzahl von Bereichen der Informatik.

ACM Computing Classification System (some sub-topics):

- Applied computing: Physical sciences and engineering, Life sciences, Arts and humanities, ...
- Computer systems organization: Distributed architectures, Neural networks, Robotics, ...
- Con...
- Gen...
- Hard...
- Hum...
- Info...
- Mat...
- Netw...
- Sec...
- Social and professional topics: Sustainability, Computer crime, User characteristics, ...
- Software and its engineering: Memory management, Scheduling, Genetic programming, ...
- Theory of computation: Probabilistic computation, Analysis of algorithms, Randomness, ...



Informatiker mit Statistikverständnis

...  
ence, Machine learning, Simulation...  
al studies, Measurement, Metrics, Evaluation, ...  
nd emulation, Reliability, Quantum computation, ...  
User studies, Empirical studies, Visualization, ...  
y optimization, Data mining, Information retrieval ...  
Probability and statistics, Optimization, ...  
ion, Monitoring, Peer-to-peer networks, ...  
tication, Anonymity, Intrusion detection, ...

# Ausgewählte Anwendungsgebiete

# Informatik

Probabilistische und statistische Methoden sind fester Bestandteil einer Vielzahl von Bereichen der Informatik.

## ACM Computing Classification System (some sub-topics):

- Applied computing: Physical sciences and engineering, Life sciences, Arts and humanities, ...

- Computer systems organization: Distributed architectures. Neural networks. Robotics. ...



## Informatiker mit Statistikverständnis

ence,   
al stud  
nd em  
User s  
y optim  
Probab  
ion, M  
tication



... und ohne

- ❑ Social and professional topics: Sustainability, Computer crime, User characteristics, ...
  - ❑ Software and its engineering: Memory management, Scheduling, Genetic programming, ...
  - ❑ Theory of computation: Probabilistic computation, Analysis of algorithms, Randomness, ...

# Ausgewählte Anwendungsgebiete

## Data Science

Data  
Consumption  
Layer

Data  
Analytics  
Layer

Data  
Management  
Layer

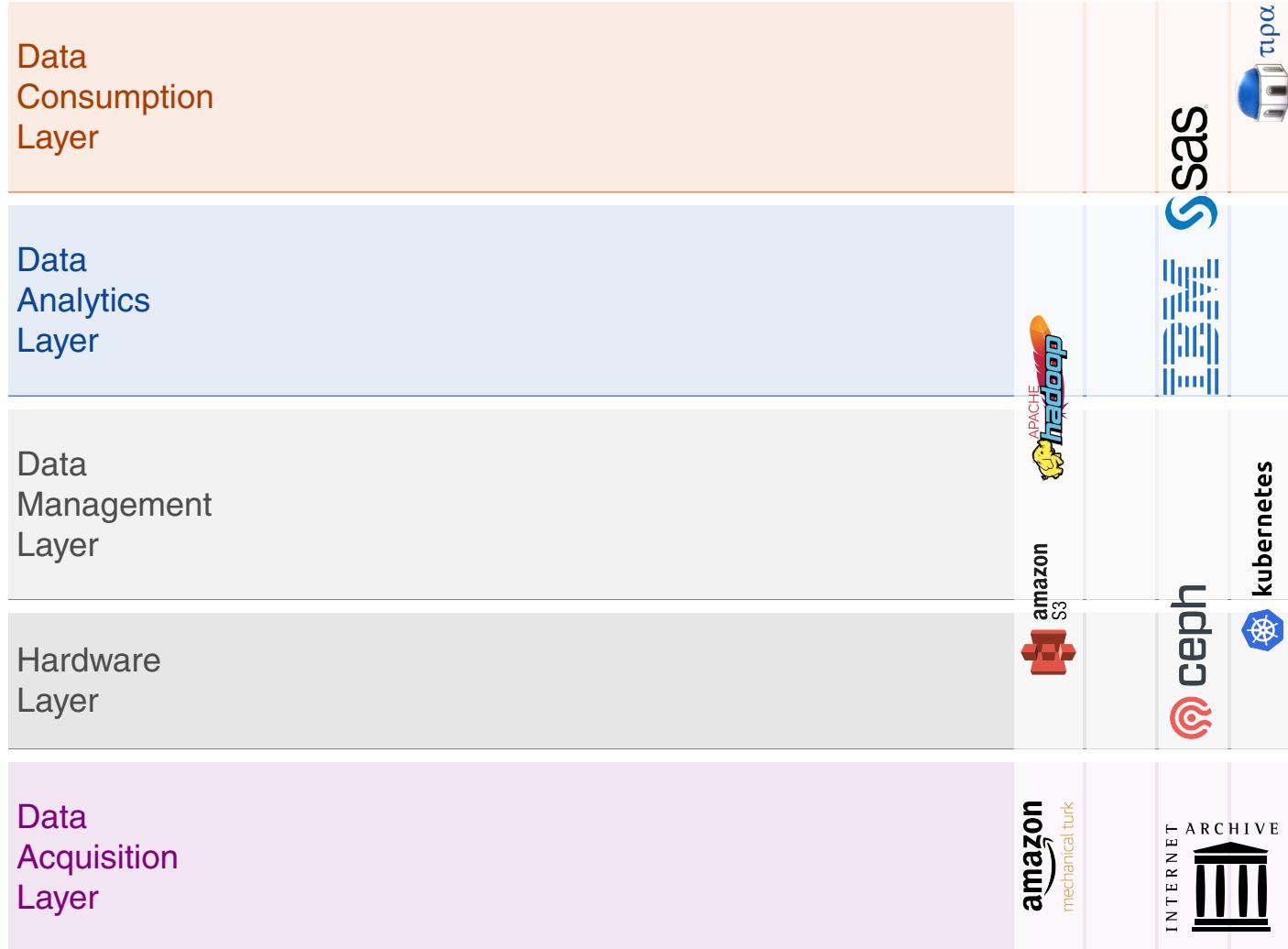
Hardware  
Layer

Data  
Acquisition  
Layer

# Ausgewählte Anwendungsgebiete

## Data Science

Vendor stack



# Ausgewählte Anwendungsgebiete

## Data Science

	Technology stack	Vendor stack
Data Consumption Layer	<ul style="list-style-type: none"><li>- Visual analytics</li><li>- Immersive technologies</li><li>- Intelligent agents</li></ul>	
Data Analytics Layer	<ul style="list-style-type: none"><li>- Distributed learning</li><li>- State-space search</li><li>- Symbolic inference</li></ul>	
Data Management Layer	<ul style="list-style-type: none"><li>- Key-value store</li><li>- RDF triple store</li><li>- Object store</li></ul>	 
Hardware Layer	<ul style="list-style-type: none"><li>- Replication</li><li>- Parallelization</li><li>- Cloud versus inhome</li></ul>	
Data Acquisition Layer	<ul style="list-style-type: none"><li>- Distant supervision</li><li>- Crowdsourcing</li><li>- Crawling and archiving</li></ul>	 

# Ausgewählte Anwendungsgebiete

## Data Science

	Task Stack	Technology stack	Vendor stack	
Data Consumption Layer	<ul style="list-style-type: none"><li>- Query and explore</li><li>- Visualize and interact</li><li>- Explain and justify</li></ul>	<ul style="list-style-type: none"><li>- Visual analytics</li><li>- Immersive technologies</li><li>- Intelligent agents</li></ul>		 
Data Analytics Layer	<ul style="list-style-type: none"><li>- Diagnose and reason</li><li>- Structure identification</li><li>- Structure verification</li></ul>	<ul style="list-style-type: none"><li>- Distributed learning</li><li>- State-space search</li><li>- Symbolic inference</li></ul>		 
Data Management Layer	<ul style="list-style-type: none"><li>- Provenance tracking</li><li>- Normalize</li><li>- Cleansing</li></ul>	<ul style="list-style-type: none"><li>- Key-value store</li><li>- RDF triple store</li><li>- Object store</li></ul>		  
Hardware Layer	<ul style="list-style-type: none"><li>- Virtualization</li><li>- Orchestration</li></ul>	<ul style="list-style-type: none"><li>- Replication</li><li>- Parallelization</li><li>- Cloud versus inhome</li></ul>	 	
Data Acquisition Layer	<ul style="list-style-type: none"><li>- Replay</li><li>- Collect</li><li>- Log</li></ul>	<ul style="list-style-type: none"><li>- Distant supervision</li><li>- Crowdsourcing</li><li>- Crawling and archiving</li></ul>	 	

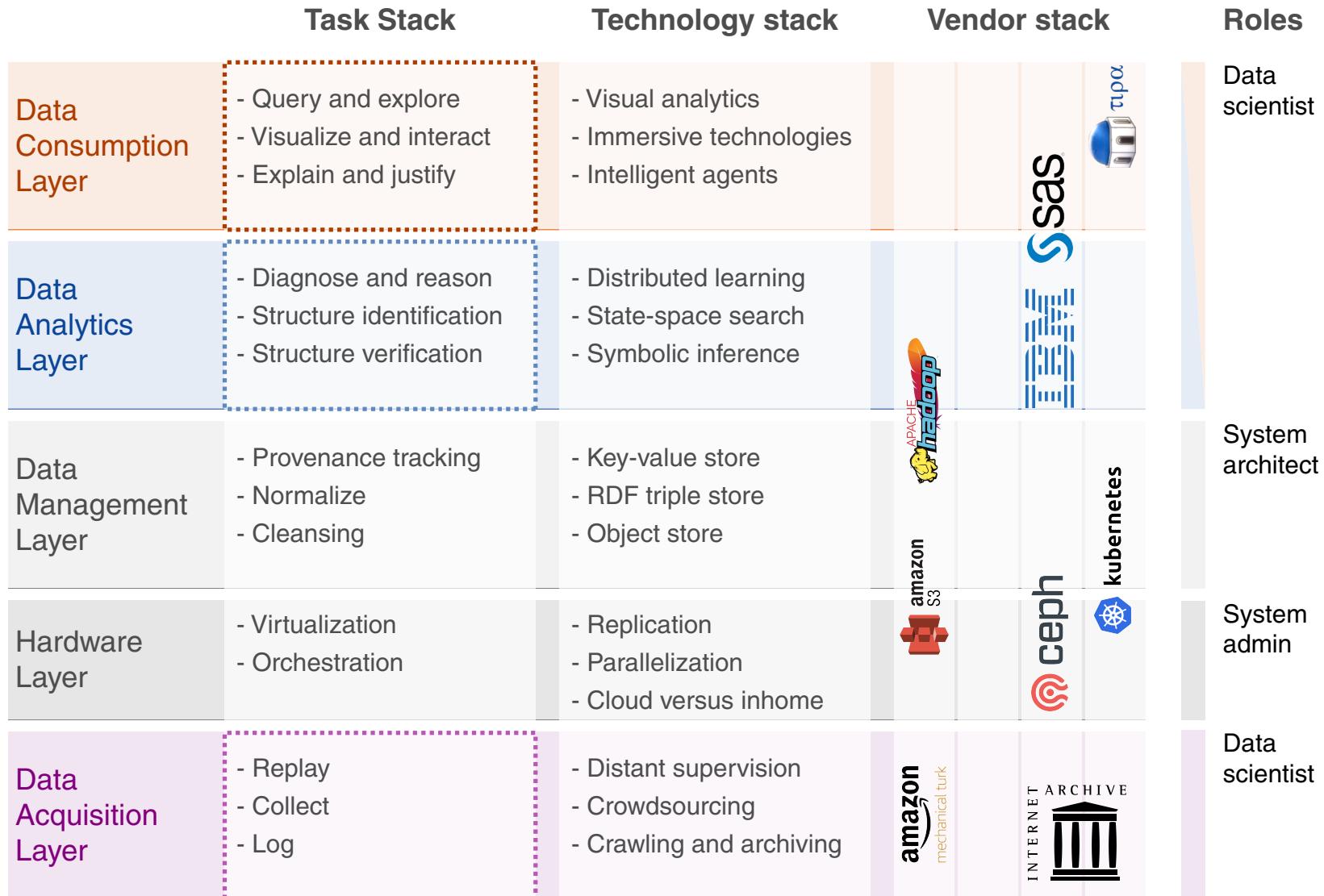
# Ausgewählte Anwendungsgebiete

## Data Science

	Task Stack	Technology stack	Vendor stack	Roles
Data Consumption Layer	<ul style="list-style-type: none"> <li>- Query and explore</li> <li>- Visualize and interact</li> <li>- Explain and justify</li> </ul>	<ul style="list-style-type: none"> <li>- Visual analytics</li> <li>- Immersive technologies</li> <li>- Intelligent agents</li> </ul>		
Data Analytics Layer	<ul style="list-style-type: none"> <li>- Diagnose and reason</li> <li>- Structure identification</li> <li>- Structure verification</li> </ul>	<ul style="list-style-type: none"> <li>- Distributed learning</li> <li>- State-space search</li> <li>- Symbolic inference</li> </ul>		
Data Management Layer	<ul style="list-style-type: none"> <li>- Provenance tracking</li> <li>- Normalize</li> <li>- Cleansing</li> </ul>	<ul style="list-style-type: none"> <li>- Key-value store</li> <li>- RDF triple store</li> <li>- Object store</li> </ul>		
Hardware Layer	<ul style="list-style-type: none"> <li>- Virtualization</li> <li>- Orchestration</li> </ul>	<ul style="list-style-type: none"> <li>- Replication</li> <li>- Parallelization</li> <li>- Cloud versus inhome</li> </ul>		
Data Acquisition Layer	<ul style="list-style-type: none"> <li>- Replay</li> <li>- Collect</li> <li>- Log</li> </ul>	<ul style="list-style-type: none"> <li>- Distant supervision</li> <li>- Crowdsourcing</li> <li>- Crawling and archiving</li> </ul>		

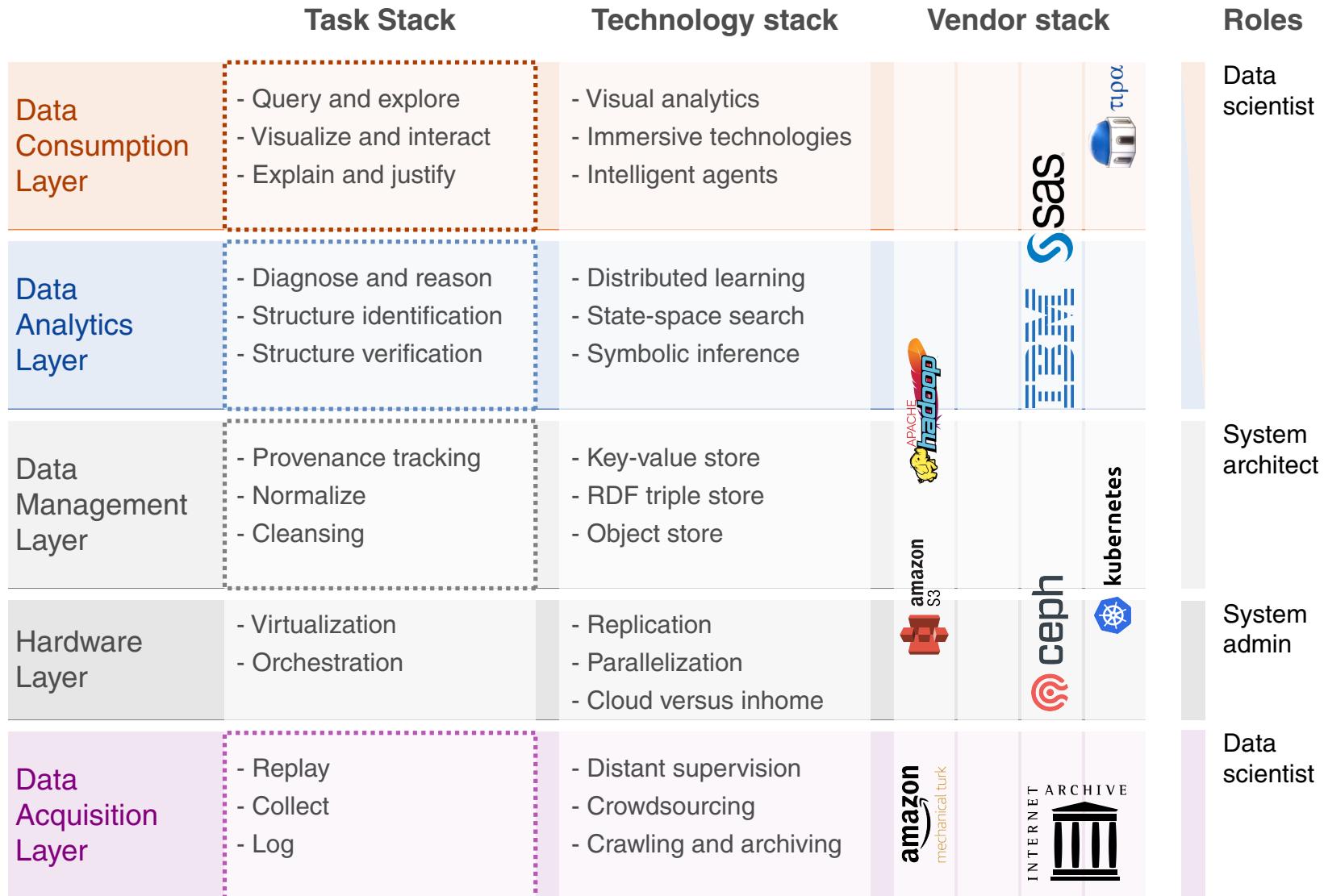
# Ausgewählte Anwendungsgebiete

## Data Science



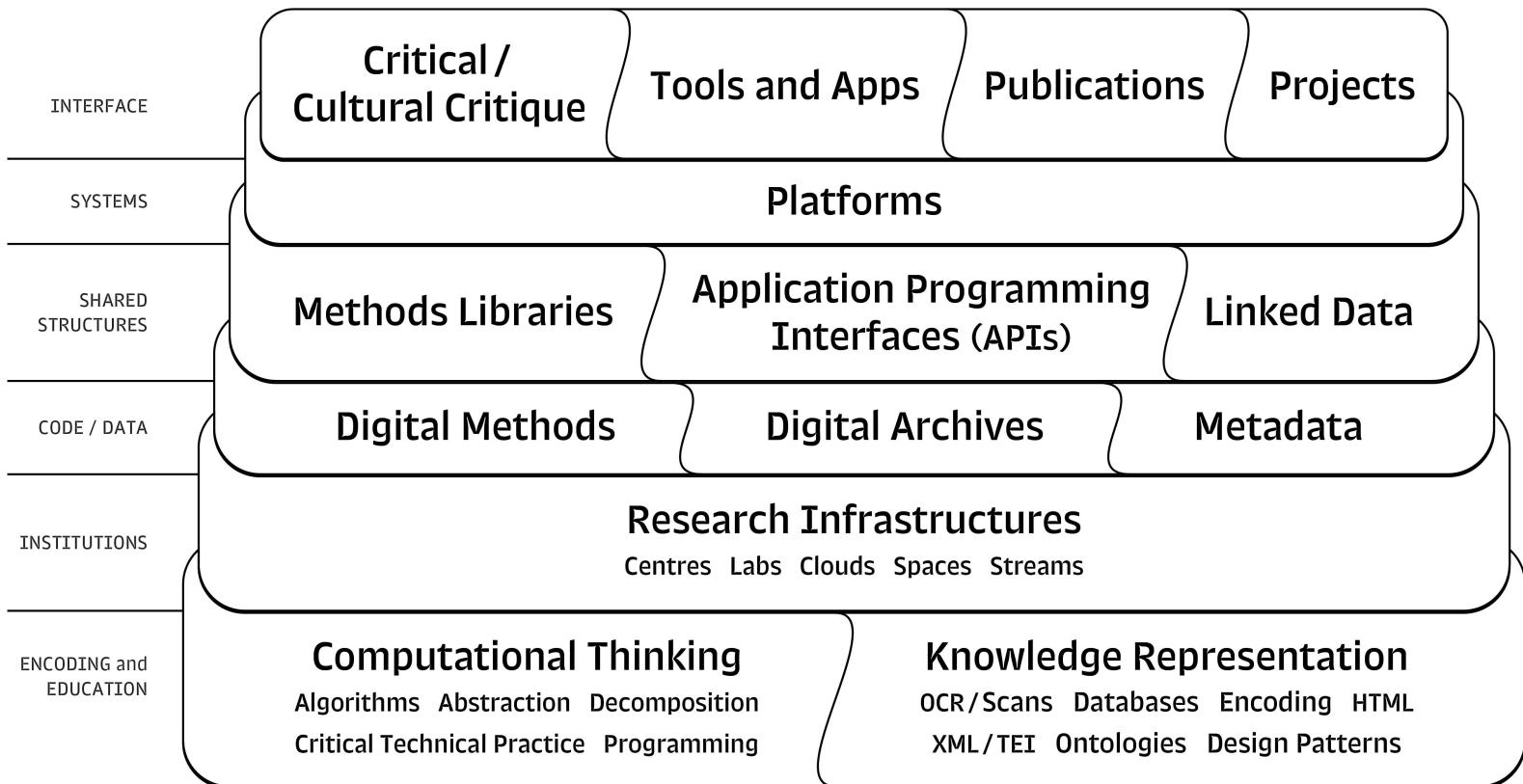
# Ausgewählte Anwendungsgebiete

## Data Science



# Ausgewählte Anwendungsgebiete

## Digital Humanities



[Berry and Fagerjord, 2017]

# Geschichte der Statistik

[Fienberg 1992]

- ~1570 Gerolamo Cardano schreibt zwischen 1525 und 1570 über Glücksspiele und die Wahrscheinlichkeitsrechnung (1663 posthum veröffentlicht).
- 1654 Blaise Pascal korrespondiert mit Pierre de Fermat über Glücksspiele und findet das „arithmetische Dreieck“ und den Binomialkoeffizient.
- 1662 John Graunt (Mitbegründer der Demographie) erfindet die Sterbetafel auf Basis der „Bills of Mortality“ und schätzt die Einwohnerzahl Londons.
- 1669 Christiaan Huygens, inspiriert durch die Arbeiten von Pascal und Fermat, berechnet mit Graunts Daten erwartete Lebenszeiten.
- 1693 Edmond Halley schätzt anhand von Breslauer Sterbetafeln kostendeckende staatliche Rentenbeiträge abhängig vom Eintrittsalter für England.

# Geschichte der Statistik

[Fienberg 1992]

- 1710 John Arbuthnot untersucht Graunts Beobachtung einer unausgeglichenen Geschlechterverteilung (1,06:1) und führt den ersten Signifikanztest durch.
- 1713 Acht Jahre nach dem Tod Jakob Bernoullis erscheint seine Abhandlung „Ars Conjectandi“, die u.a. sein „Gesetz der großen Zahlen“ und die Konzepte subjektiver Wahrcheinlichkeit und moralischer Gewissheit behandelt.
- 1738 Abraham de Moivre veröffentlicht eine Anwendung des Grenzwertsatzes und was heute als Normal-Approximation an die Binomialverteilung bekannt ist. De Moivre erkannte die Wichtigkeit des  $\sqrt{n}$ -Gesetzes.
- 1738 Daniel Bernoulli (Jakobs Neffe) diskutiert subjektiven „Nutzen“ als Maß für unsichere Entscheidungen im Gegensatz zu Gewinnmaximierung.

## Bemerkungen:

- Die Arbeiten vieler Akteure dieser Zeitspanne wurde getrieben von ihren theologischen Studien zum Gottesbeweis. Man glaubte an eine deterministische Welt („Gott würfelt nicht nicht“), griff jedoch auf die Wahrscheinlichkeitstheorie zurück, um Glücksspiele und Probleme der Moral und der Theologie zu erklären.
- Die deskriptive Statistik entwickelte sich zunächst weitgehend unabhängig von der Wahrscheinlichkeitsrechnung, ohne Rückgriff auf die aus ihr hervorgehenden Theorien.

# Geschichte der Statistik

[Fienberg 1992]

- 1750 Tobias Mayer schätzt die Mondposition unter Berücksichtigung seiner Libration indem er 27 Beobachtungen durch 3 Zentroiden ersetzt und so die größtmögliche Information aus den zur Verfügung stehenden Daten zieht.
- 1755 Thomas Simpson untersucht Verteilungen von (Mess-)Fehlern in Daten und empfiehlt ihre Minimierung durch Durchschnittsbildung vieler Beobachtungen.
- 1764 Thomas Bayes führt die bedingte Wahrscheinlichkeit ein und sucht erstmals einen Rückschluss von Beobachtungen auf Verteilungsparameter zu ziehen; er formuliert den speziellen Satz von Bayes („inverse Wahrscheinlichkeit“).
- 1774 Pierre-Simon Laplace findet den Satz von Bayes erneut und popularisiert ihn.
- 1778 Daniel Bernoulli schließt auf Verteilungsparameter aus Beobachtungen durch Maximierung ihrer Wahrscheinlichkeit, wird jedoch von Euler harsch kritisiert.

# Geschichte der Statistik [Fienberg 1992]

- 1805 Adrien-Marie Legendre veröffentlicht die Methode der kleinsten Quadrate.
- 1809 Johann Carl Friedrich Gauß veröffentlicht seine unabhängig gefundene Methode der kleinsten Quadrate und stellt die Verbindung zur ebenfalls neuen Normalverteilung her, die als Verteilung für Fehler angenommen wird; sowie das Gauß-Eliminationsverfahren und das Gauß-Newton-Verfahren.
- 1810 Laplace generalisiert den ersten zentralen Grenzwertsatz von de Moivre und zeigt, dass die Summe vieler Zufallsvariablen (z.B. viele Durchschnittswerte von Beobachtungsreihen) annähernd normalverteilt ist, was die Besonderheit der Normalverteilung erklärt.
- 1835 Adolphe Quetelet beschreibt den Durchschnittsmenschen als normalverteilte Variablen, u.a. den immer noch gebräuchlichen Body-Mass-Index.
- 1843 Antoine Cournot führt den frequentistischen Wahrscheinlichkeitsbegriff ein.

## Bemerkungen:

- Die vorgenannten Entwicklungen können in zwei Linien unterteilt werden: Die Entwicklung der schließenden Statistik sowie die der Methode der kleinsten Quadrate um Beobachtungen zu kombinieren.
- Beide Linien hatten zum Ziel, mit unsicheren, fehlerbehafteten Daten umzugehen, einerseits durch Suche nach der bestpassendsten Wahrscheinlichkeitsverteilung und andererseits durch eine optimale Annäherung einer Linearfunktion.
- Treibende Kraft dieser Entwicklungen waren vor allem Probleme der Astronomie, wie die Vorhersage von Planetenbahnen auf Basis weniger, fehlerbehafteter Beobachtungen, aber auch Probleme der Biologie und Geodäsie.

# Geschichte der Statistik [Fienberg 1992]

- 1869 Francis Galton veröffentlicht eine Studie zur Erblichkeit von Intelligenz.
- 1885 Galton beobachtet in Vererbungsexperimenten die Regression zur Mitte und prägt den Begriff der Regression.
- 1888 Galton entwickelt zusammen mit Karl Pearson den Korrelationskoeffizient.
- 1897 Udny Yule beschreibt die multiple und partielle Korrelation und legt die Grundlage für eine Synthese von Korrelation und Regression mit der Methode der kleinsten Quadrate und der Theorie von Fehlern.
- 1900 Pearson findet den chi-square test; der erste Hypothesentest.
- 1907 Yule beschreibt die multiple Regression, die bis heute genutzt wird.

## Bemerkungen:

- Galton prägte auf Basis seiner Forschung zur Vererbung auch den Begriff der Eugenik, „die Wissenschaft, die sich mit allen Einflüssen befaßt, welche die angeborenen Eigenschaften einer Rasse verbessern“, welche Grundlage der nationalsozialistischen „Rassenhygiene“ wurde. Auch sein Schüler Pearson verfocht diese zur damaligen Zeit weit verbreitete Lehre.

# Geschichte der Statistik

[Fienberg 1992]

- 1908 William Sealy “Student” Gosset veröffentlicht die t-Verteilung, die die Durchschnitte kleiner Stichproben normalverteilter Populationen beschreibt.
- 1913 Ronald Aylmer Fisher beginnt seine Karriere als Statistiker und Biologe und leistet in den folgenden Jahrzehnten zahlreiche grundlegende Beiträge:
- 1922 Schätztheorie zu Suffizienz, Konsistenz, Effizienz und Maximum-Likelihood.
- 1923 Theorie der Varianz als Maß für Streuung einer Zufallsvariable sowie die Varianzanalyse („Analysis of Variance“, ANOVA).  
Jerzy Neymann schlägt erstmals randomisierte Experimente vor.
- 1925 Signifikanzanalyse auf Basis des  $p$ -Werts, der Wahrscheinlichkeit, einen Wert zu erhalten der mindestens so extrem ist wie der betrachtete.

# Geschichte der Statistik

[Fienberg 1992]

- 1926 Frank Plumpton Ramseys Arbeit zu subjektiver Wahrscheinlichkeit und Nutzen begründet die moderne Entscheidungstheorie.
- 1930 Bruno de Finetti entwickelt unabhängig von Ramsey eine Theorie subjektiver Wahrscheinlichkeit.
- 1933 Andrei Nikolajewitsch Kolmogorov veröffentlicht eine Axiomatisierung der Wahrscheinlichkeitsrechnung.  
Neyman und Pearson erweitern Fishers Signifikanztests zu Hypothesentests und führen den Fehler 2. Art und die Trennschärfe eines Tests ein.
- 1934 Neyman beschreibt repräsentative Stichproben sowie Konfidenzintervalle.
- 1939 Harold Jeffreys systematisiert die Baysessche Statistik insbesondere im Hinblick auf a-priori Verteilungen.
- 1945 Abraham Wald beschleunigt mittels Sequenzanalyse Experimente.
- 1954 Leonard Savage, inspiriert von de Finetti, begründet in „Foundations of Statistics“ die moderne Bayessche Statistik.

## Bemerkungen:

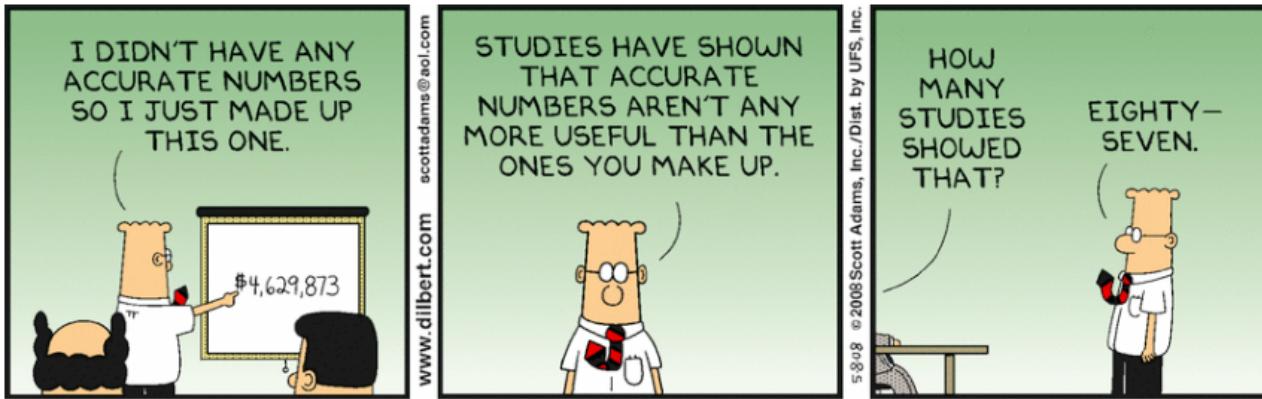
- Statistische Methoden und Gesetze sind durchsetzt mit dem was heute als Stigler's Gesetz bekannt worden ist: Keine wissenschaftliche Erkenntnis ist nach ihrem ursprünglichen Erfinder benannt. Gerade in der frühen Geschichte der Statistik gab es zahlreiche Prioritätskonflikte.
- Die moderne Statistik ist dagegen seit Jahrzehnten tief gespalten in der Frage, wie induktive Schlussfolgerungen aus Daten gezogen werden müssen:

*It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis. [Savage 1954]*

# Missbrauch der Statistik

„Lies, damned lies, and statistics“

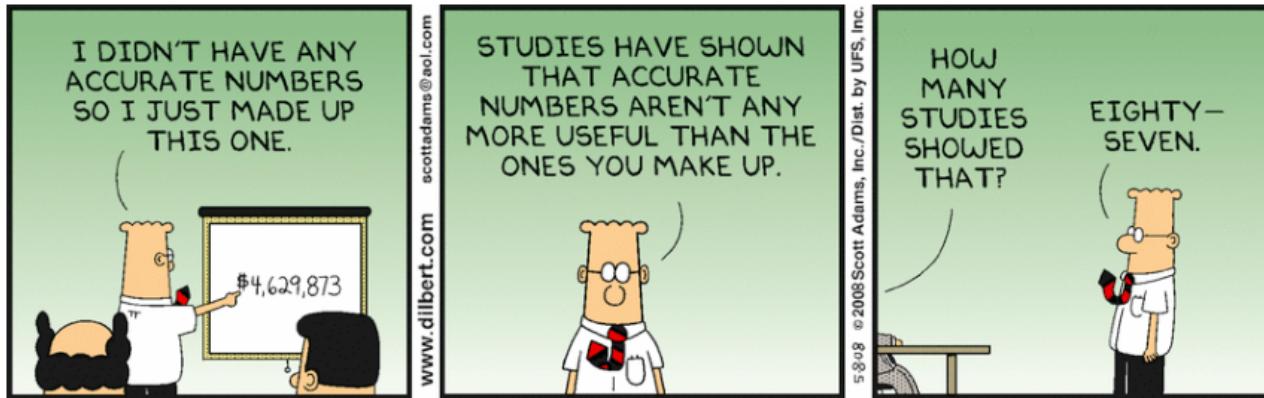
Erfundene / manipulierte Daten:



# Missbrauch der Statistik

„Lies, damned lies, and statistics“

Erfundene / manipulierte Daten:



“Geschicktes” Auswerten [Lopez-Real 1988]:

- Statistik 1: Fliegen ist sicherer als Bahnhfahren
  - Bahn 9 Verkehrstote pro 10 Milliarden Passagierkilometer
  - Flugzeug 3 Verkehrstote pro 10 Milliarden Passagierkilometer
- Statistik 2: Bahnhfahren ist sicherer als Fliegen
  - Bahn 7 Verkehrstote pro 100 Millionen Passagierstunden
  - Flugzeug 24 Verkehrstote pro 100 Millionen Passagierstunden

## Bemerkungen:

- Bahn vs. Flugzeug ist ein sogenanntes Adäquationsproblem, nämlich die Schwierigkeit dasjenige Datenmaterial zur Untersuchung heranzuziehen, das die Realität hinreichend repräsentiert.
- Im Beispiel besteht der Unterschied in der höheren Durchschnittsgeschwindigkeit eines Flugzeugs gegenüber einer Bahn (Annahme: 800 km/h vs. 80 km/h).
- Interpretation der Verkehrssicherheitsdaten (Anzahl Verkehrstote) ist abhängig von der Fragestellung:
  - Eine gegebene Strecke ist sicherer mit dem Flugzeug zurückzulegen.
  - Vor einem achtstündigen Flug sollte man trotzdem mehr Todesangst als vor einer achtstündigen Bahnfahrt verspüren.

# Missbrauch der Statistik

„How to Lie with Statistics“ [Huff 1954]

Häufige Ursachen: [\[Wikipedia\]](#)

- Domänenexpertise ohne Statistik-Expertise
- Statistik-Expertise ohne Domänenexpertise
- Das zugrundeliegende Problem/Thema/Ziel ist nicht wohldefiniert
- Mangelnde Datenqualität
- Übertreibungen durch die Presse
- „Politicians use statistics in the same way that a drunk uses lamp-posts—for support rather than illumination.“ [\[Andrew Lang 1910\]](#)