

Tag 1: Der Daten-Lifecycle und: Woher kommen Daten?

Session 1: Data Science - Ein Überblick

Simon Munzert
Hertie School

1. Willkommen!
2. Was ist Data Science?
3. Was kann Data Science?
4. Ziele für dieses Modul

Herzlich Willkommen!

Vorstellungsrunde

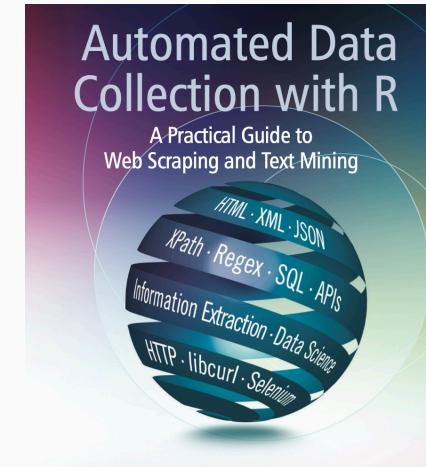
Hertie School

Über mich

👤 Ich bin Simon Munzert [si'mən munsərt], oder Simon [saɪmən].

✉️ munzert@hertie-school.org

🎓 Professor für Data Science und Public Policy | Director, Data Science Lab



Simon Munzert | Christian Rubba | Peter Meißner | Dominic Nyhuis

American Political Science Review, Page 1 of 17
doi:10.1017/S000305418000424

© American Political Science Association 2018

Examining a Most Likely Case for Strong Campaign Effects: Hitler's Speeches and the Rise of the Nazi Party, 1927–1933

PETER SELB¹ University of Konstanz
SIMON MUNZERT² Hertie School of Governance

Hitler's rise to power amidst an unprecedented propaganda campaign initiated scholarly interest in campaign effects. To the surprise of many, empirical studies often find minimal effects. The predominant view is that early work on campaign effects through comparative analysis of the archetypal German wars most likely for strong effects has rarely been studied. We collect extensive data about Hitler's speeches and gauge their impact on voter support at five national elections preceding the dictatorship. We use a semi-parametric difference-in-differences approach to estimate effects in the face of potential confounding due to the deliberate scheduling of events. Our findings suggest that Hitler's speeches, while rationally targeted, had a negligible impact on the Nazis' electoral fortunes. Only the 1932 presidential runoff, an election preceded by an extraordinarily short, intense, and one-sided campaign, yielded positive effects. This study questions the importance of charismatic leaders for the success of populist movements.

Zweitstimme.org. Ein strukturell-dynamisches Vorhersagemodell für Bundestagswahlen

Simon Munzert, Lukas Stötzer, Thomas Gschwend, Marcel Neunhoeffer und Sebastian Sternberg

Zweitstimme.org. A structural-dynamic forecasting model for German federal elections Abstract: We present results of an ex-ante forecast of party-specific vote shares at the German Federal Election 2017. To that end, we combine data from published trial heat polls with structural information. The model takes care of the multi-party nature of the setting and allows making statements about the probability of certain events, such as the plurality of votes for a party or the majority for coalition options in parliament. The forecasts of our model are continuously being updated on the platform zweitstimme.org. The value of our approach goes beyond the realms of academia: We equip journalists, political pundits, and ordinary citizens with information that can help make sense of the parties' latent support and ultimately make voting decisions better informed.

nature
human behaviour
<https://doi.org/10.1038/s41562-020-01044-x>
ARTICLES
Check for updates

Tracking and promoting the usage of a COVID-19 contact tracing app

Simon Munzert^{1,2}, Peter Selb², Anita Gohdes³, Lukas F. Stötzer³ and Will Lowe³

Digital contact tracing apps have been introduced globally as an instrument to contain the COVID-19 pandemic. Yet, privacy by design impedes both the evaluation of these tools and the deployment of evidence-based interventions to stimulate uptake. We combine an online survey with mobile tracking data to examine actual usage of Germany's official contact tracing app and other similar releases. We find that with increased risk aversion, but lower relative risk aversion, individuals take a heightened risk of exposure to COVID-19. Using a randomized intervention, we show that information and motivation via video messages have very limited effect on uptake. However, findings from a second intervention suggest that even small monetary incentives can strongly increase uptake and help make digital contact tracing a more effective tool.

PNAS
NEXUS

PNAS Nexus, 2025, 4, pga032
<https://doi.org/10.1093/pnasnexus/pga032>
Advance access publication 12 February 2025
Research Report

Citizen preferences for online hate speech regulation

Simon Munzert^{1,*}, Richard Traumüller², Pablo Barberá³, Andrew Guess⁴ and JungHwan Yang⁵

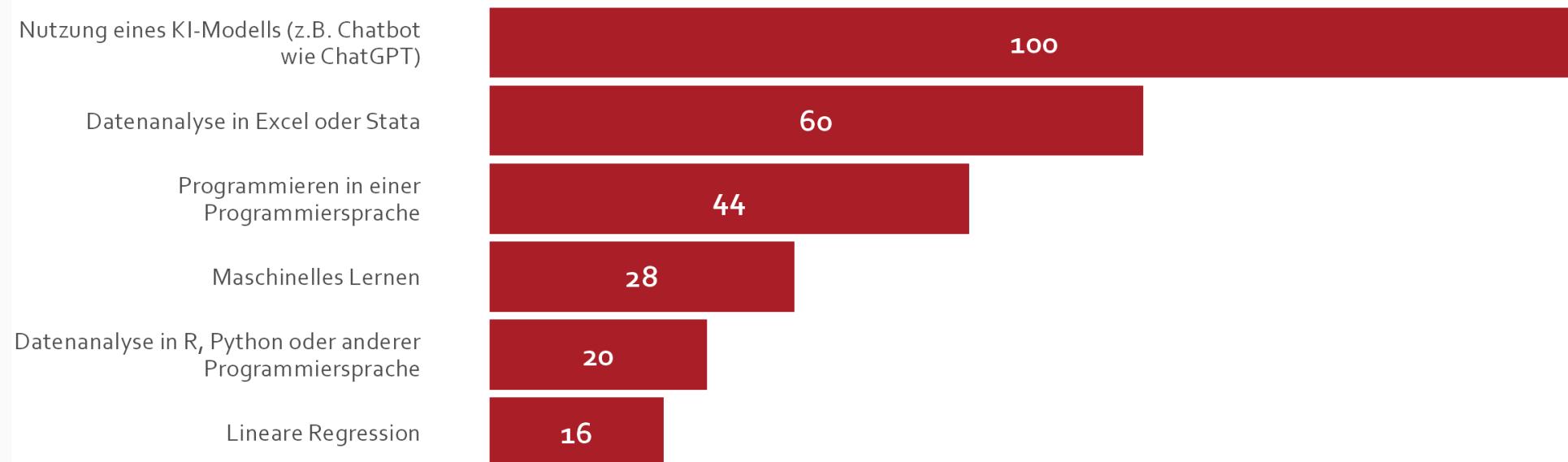
¹Data Science Lab, Hertie School, 1017 Berlin, BE, Germany
²School of Social Sciences, University of Münster, 48159 Münster, NW, Germany
³Department of Political Science, University of California, San Diego, La Jolla, CA 92093, USA
⁴Department of Politics and School of Public and International Affairs, Princeton University, Princeton, NJ 08544, USA
⁵Department of Communication, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
^{*}To whom correspondence should be addressed. Email: munzert@hertie-school.org

Abstract

The shift of public discourse to online platforms has intensified the debate over content moderation by platforms and the regulation of online speech. Designing rules that are met with wide acceptance requires learning about public preferences. We present a visual vignette study using a sample ($n = 2,622$) of German and US citizens that were exposed to 20 synthetic social media vignettes mimicking actual cases of hateful speech. We find people's evaluations to be primarily shaped by message type and severity, and less by contextual factors. When messages are more severe, they are more popular; more extreme sanctions like job loss find little support even in cases of extreme hate. Further evidence suggests in-group favoritism among political partisans. Experimental evidence shows that exposure to hateful speech reduces tolerance of unpopular opinions.

Erfahrung im Umgang mit Datenanalyse und KI

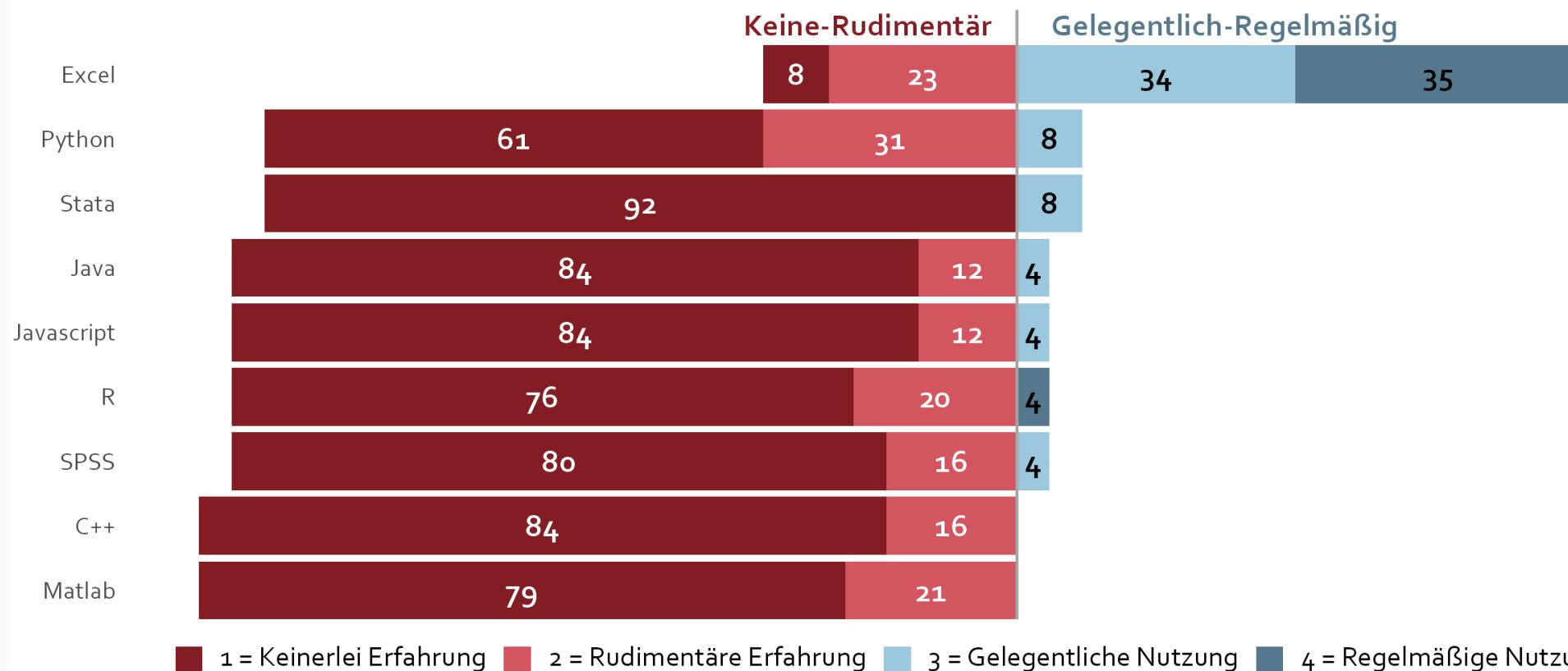
% der Teilnehmer*innen (n = 25)



Erfahrungen mit Programmiersprachen und Datenanalyseprogrammen

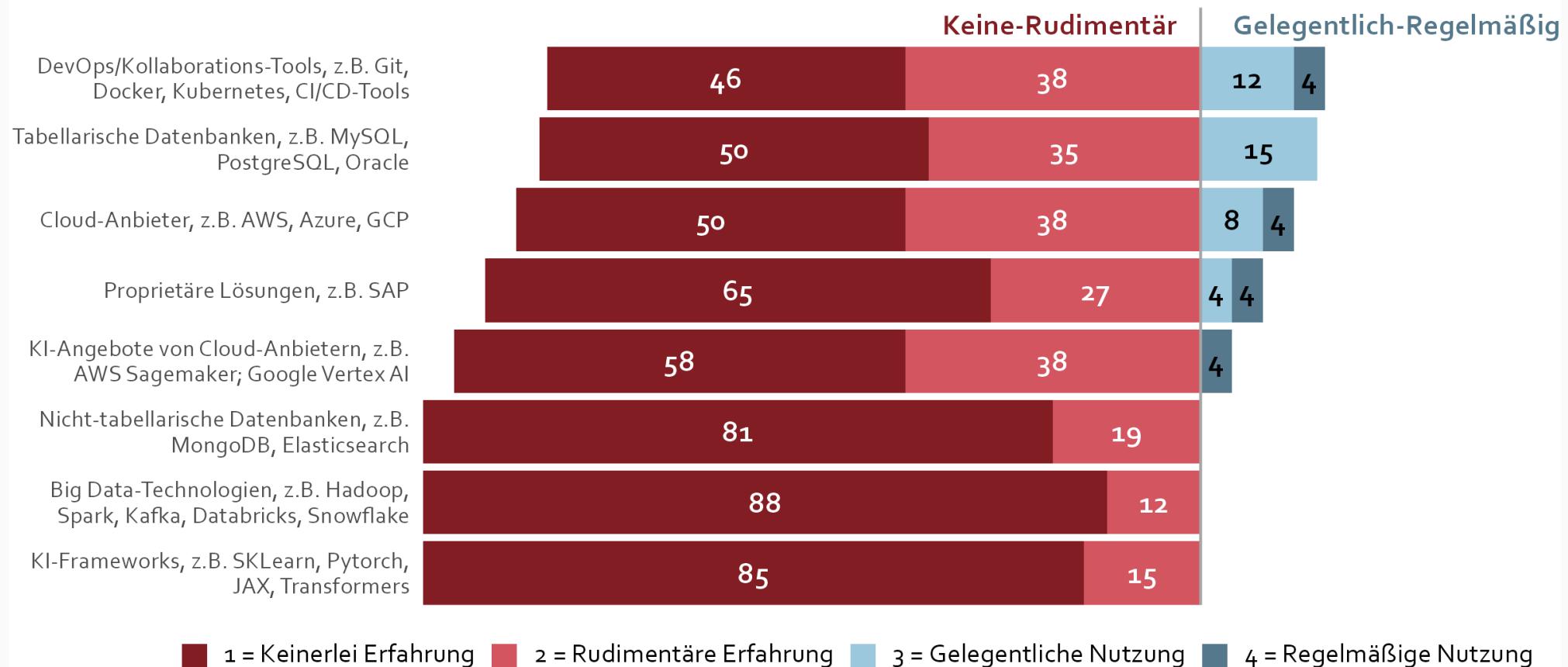
% der Teilnehmer*innen ($n = 25$).

Außerdem genannt: Cobol, Natural, Basic, Orange, SQL, VBA, Ada



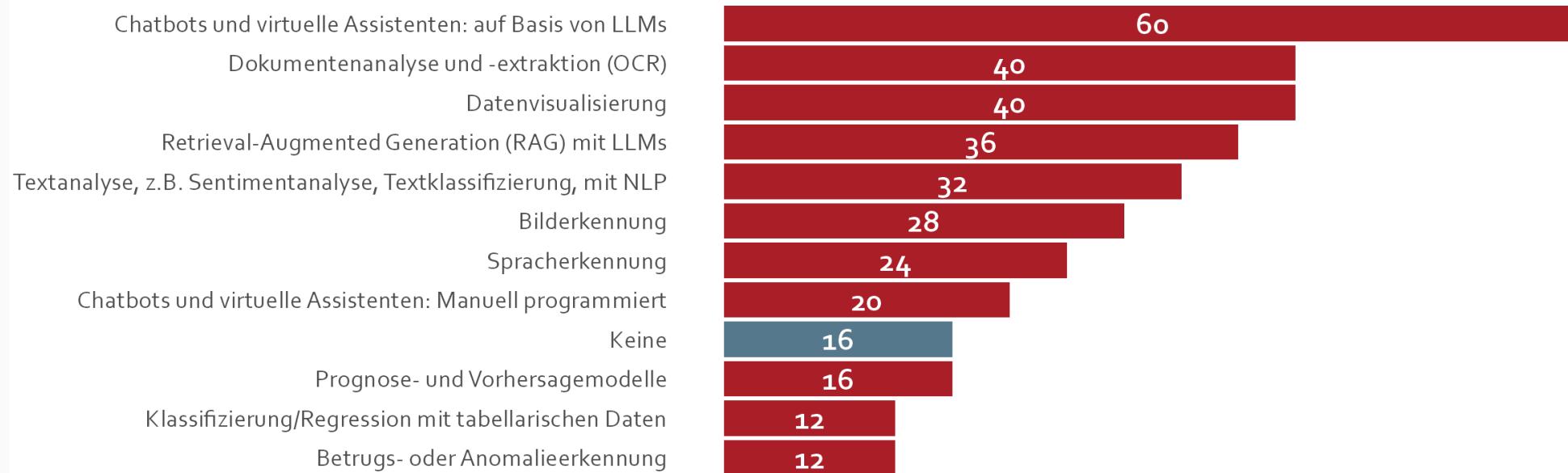
Erfahrungen mit weiteren Technologien

% der Teilnehmer*innen (n = 25)

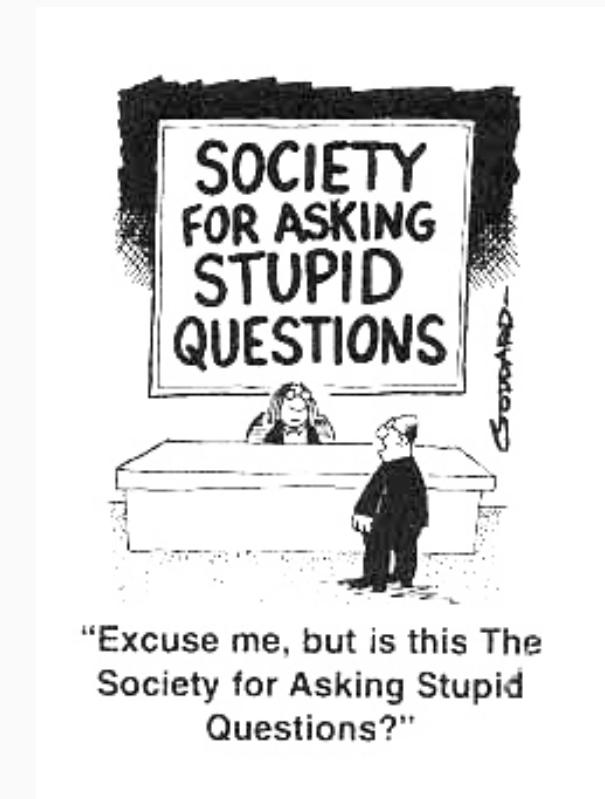


Erfahrung mit KI-Lösungen und Datenanalyseverfahren

% der Teilnehmer*innen, die Technologie kennen bzw. in Projekt umgesetzt haben (n = 23)



- Wir haben viele Themen vor uns. Ich habe Material vorbereitet, aber entscheidend ist: **Signalisieren Sie Ihre Interessen und Bedarfe** – ich vertiefe gerne, schwenke um oder greife andere Beispiele auf, soweit möglich.
- Ich bin kein Experte für Ihre Behörde oder Ihre spezifischen Themen. **Bringen Sie Ihr Fachwissen und Ihre Erfahrung ein.** Sie bilden die Grundlage für eine fundierte, evidenzbasierte Diskussion über Data Science im öffentlichen Sektor.
- Bitte nutzen Sie die Gelegenheit, **jederzeit Fragen zu stellen.** Es gibt keine schlechten Fragen – jede trägt zur gemeinsamen Diskussion bei.



"Excuse me, but is this The Society for Asking Stupid Questions?"

Diskussion

Was möchten Sie über Daten und Data Science lernen?

Was sind Ihre Erwartungen an dieses Modul?

Excalidraw - excalidraw.com

Was ist Data Science?

Was ist Data Science?

Was ist Data Science?

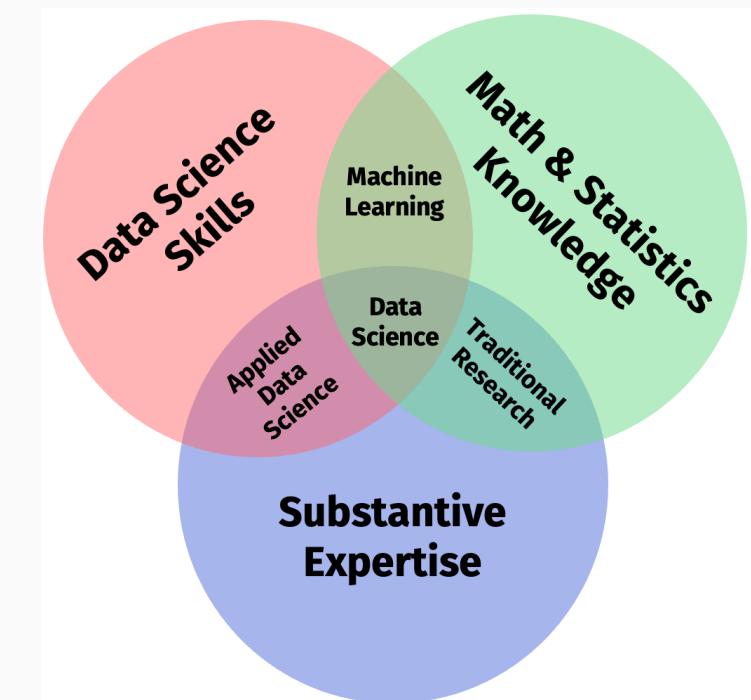
"Der interdisziplinäre Forschungszweig Data Science befasst sich mit dem Management und der Analyse von Daten." - [Daniel Keim, Kai-Uwe Sattler](#)

"Data Science ist ein interdisziplinäres Wissenschaftsfeld, welches wissenschaftlich fundierte Methoden, Prozesse, Algorithmen und Systeme zur Extraktion von Erkenntnissen, Mustern und Schlüssen sowohl aus strukturierten als auch unstrukturierten Daten ermöglicht." - [Wikipedia](#)

"Data Science ist ein Konzept, das Statistik, Datenanalyse, Informatik und die damit verbundenen Methoden vereint, um aktuelle Phänomene anhand von Daten zu verstehen und zu analysieren." - [Chikio Hayashi](#)

☞ Zwischen Disziplinen angesiedelt, anwendungsorientiert

Eine Arbeitsdefinition



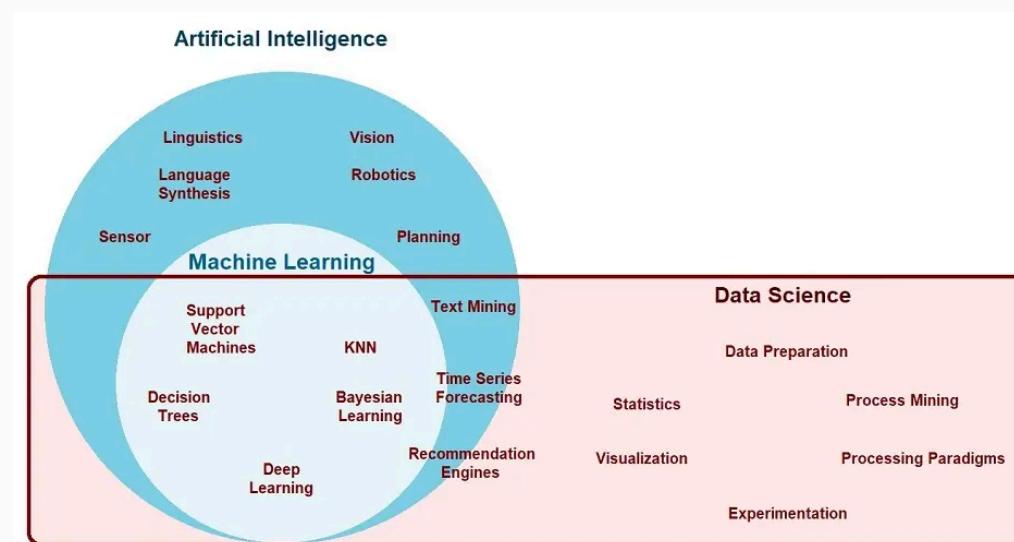
Quelle [Drew Conway, 2010](#) (angepasst)

Künstliche Intelligenz (KI)

- Teilbereich von (u.a.) Data Science, der sich mit der Automatisierung von intelligentem Verhalten befasst.
- Praktisch alle KI-Modelle sind ohne Daten nutzlos.
- Das "I" in KI bezieht sich in der Regel darauf, dass die Modelle **lernen** können, d.h. auf variablen Input angemessen reagieren.

Data Science

- Überbegriff für Methoden und Technologien, die auch KI mit einschließen.
- Aber auch Methoden zur Datenerhebung, -verarbeitung, -modellierung, -visualisierung, -kommunikation, -archivierung.
- Damit bereitet Data Science die essentielle Grundlage für KI-Anwendungen.



1. Beschreibung

- Wie ist der Zustand der Welt?
- Wie entwickeln sich Trends?
- Wie gestalten sich Gruppenunterschiede?

2. Erklärung

- Welche Wirkung hat eine Maßnahme?
- Ist sie über Gruppen heterogen?
- Was sind Wirkungsmechanismen?

3. Vorhersage

- Wie wird Person X sich verhalten?
- (Wann) wird Ereignis X eintreten?
- Zu welchem Typ gehört eine Beobachtung?

Der Wert für die Politikgestaltung

- Im Zentrum des **Monitorings**
- „Wie viele Menschen konsumieren Fehlinformationen im Internet?“
- „Wie entwickelt sich die Arbeitslosigkeit in Bezirken?“
- „Wie viele Radunfälle gab es pro Straßenabschnitt?“

Der Wert für die Politikgestaltung

- Im Mittelpunkt der **Evaluation**
- „Hat die Digitalisierungsmaßnahme zu Zeitersparnis geführt?“
- „Bei wem wirkte die Anti-Fake-News-Kampagne?“
- „Warum hat die Intervention nicht zu den erwarteten Ergebnissen geführt?“

Der Wert für die Politikgestaltung

- Steht im Mittelpunkt der **Vorhersage**, aber auch der **Zielsetzung** und **Messung**.
- „Wird die Person rückfällig werden?“
- „Enthält dieser Post Hate Speech?“
- „Wie stark wird die Wasserverschmutzung an einem bestimmten Tag sein?“

Die Data-Science-Pipeline



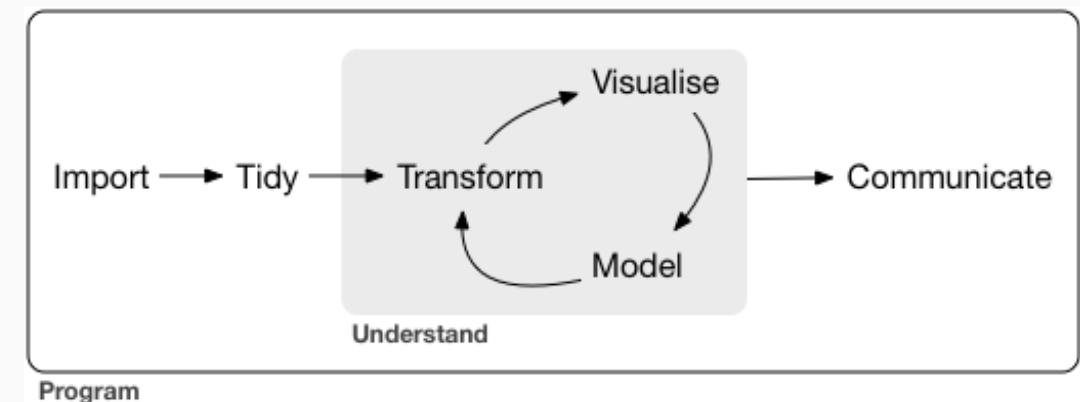
Vorbereitung

- **Problemstellung** Vorhersage, Inferenz, Beschreibung
- **Design** Konzept erstellen, Datenerfassung aufsetzen
- **Datenerhebung** rekrutieren, sammeln, überwachen

Vorbereitung

- **Problemstellung** Vorhersage, Inferenz, Beschreibung
- **Design** Konzept erstellen, Datenerfassung aufsetzen
- **Datenerhebung** rekrutieren, sammeln, überwachen

Datenverarbeitung



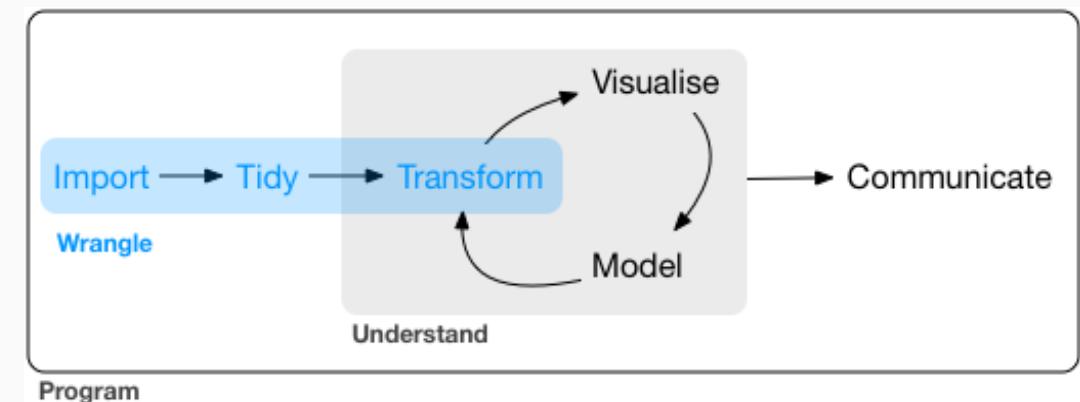
Quelle H. Wickham, M. Çetinkaya-Rundel, G. Gronemund:
R for Data Science

Vorbereitung

- **Problemstellung** Vorhersage, Inferenz, Beschreibung
- **Design** Konzept erstellen, Datenerfassung aufsetzen
- **Datenerhebung** rekrutieren, sammeln, überwachen

Datenverarbeitung

- **Aufbereiten:** importieren, bereinigen, anreichern



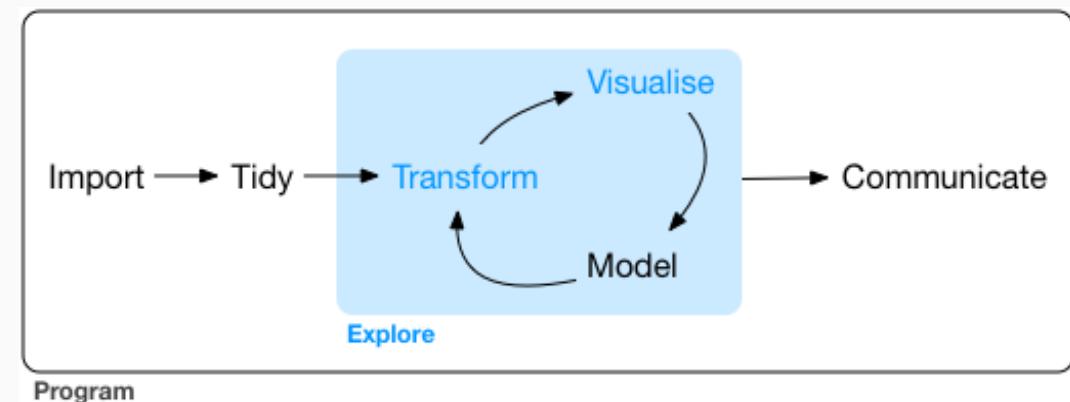
Quelle H. Wickham, M. Çetinkaya-Rundel, G. Gronemund:
R for Data Science

Vorbereitung

- **Problemstellung** Vorhersage, Inferenz, Beschreibung
- **Design** Konzept erstellen, Datenerfassung aufsetzen
- **Datenerhebung** rekrutieren, sammeln, überwachen

Datenverarbeitung

- **Aufbereiten:** importieren, bereinigen, anreichern
- **Explorieren:** visualisieren, beschreiben, entdecken



Quelle H. Wickham, M. Çetinkaya-Rundel, G. Gronemund:
R for Data Science

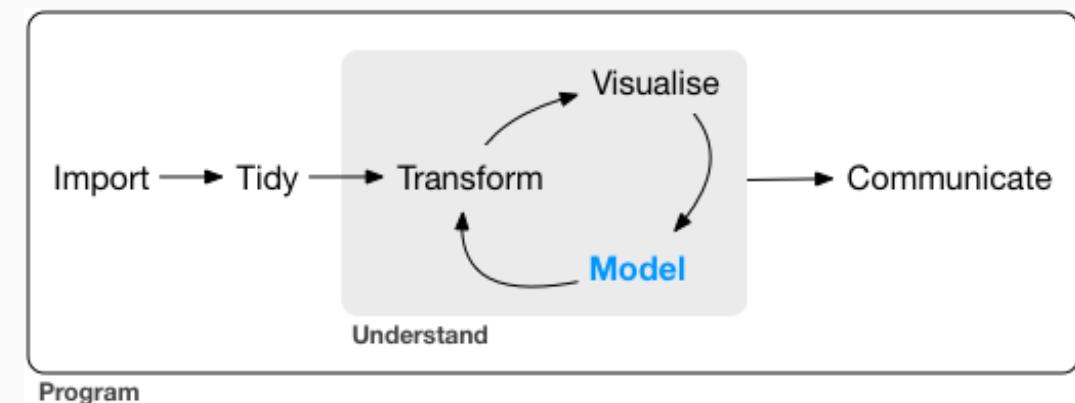
Die Data-Science-Pipeline

Vorbereitung

- **Problemstellung** Vorhersage, Inferenz, Beschreibung
- **Design** Konzept erstellen, Datenerfassung aufsetzen
- **Datenerhebung** rekrutieren, sammeln, überwachen

Datenverarbeitung

- **Aufbereiten:** importieren, bereinigen, anreichern
- **Explorieren:** visualisieren, beschreiben, entdecken
- **Modellieren:** testen, inferieren, vorhersagen



Quelle H. Wickham, M. Çetinkaya-Rundel, G. Gronemund:
R for Data Science

Vorbereitung

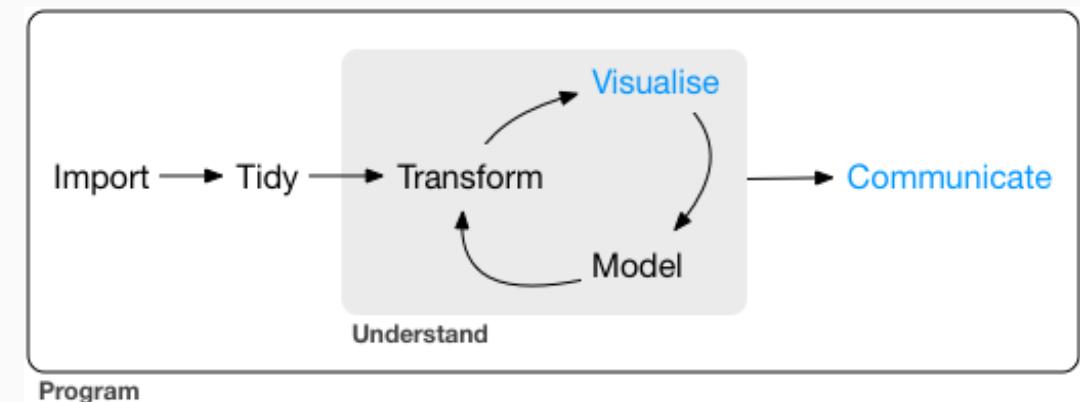
- **Problemstellung** Vorhersage, Inferenz, Beschreibung
- **Design** Konzept erstellen, Datenerfassung aufsetzen
- **Datenerhebung** rekrutieren, sammeln, überwachen

Datenverarbeitung

- **Aufbereiten:** importieren, bereinigen, anreichern
- **Explorieren:** visualisieren, beschreiben, entdecken
- **Modellieren:** testen, inferieren, vorhersagen

Verbreitung

- **Kommunikation:** Öffentlichkeit, Entscheidungsträger
- **Veröffentlichen:** Zeitschriften, Software, Berichte
- **Produktivieren:** nutzbar, robust, skalierbar machen



Quelle H. Wickham, M. Çetinkaya-Rundel, G. Gronemund:
R for Data Science

Vorbereitung

- **Problemstellung** Vorhersage, Inferenz, Beschreibung
- **Design** Konzept erstellen, Datenerfassung aufsetzen
- **Datenerhebung** rekrutieren, sammeln, überwachen

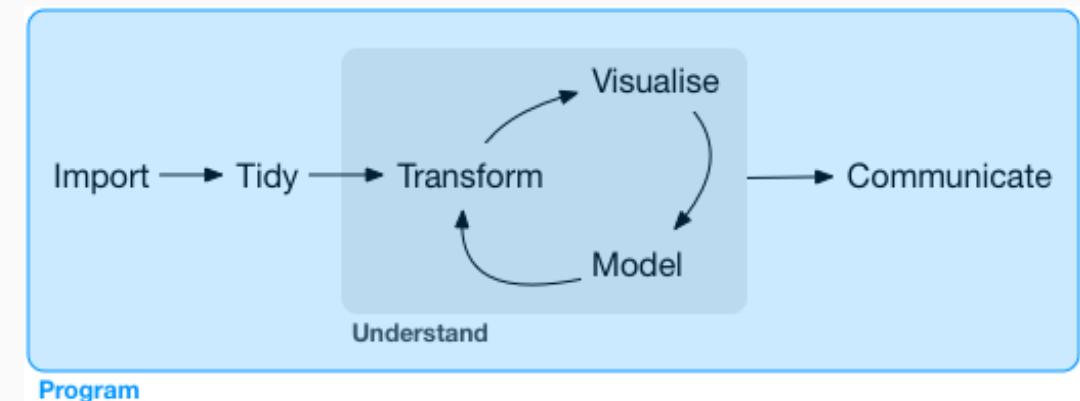
Datenverarbeitung

- **Aufbereiten:** importieren, bereinigen, anreichern
- **Explorieren:** visualisieren, beschreiben, entdecken
- **Modellieren:** testen, inferieren, vorhersagen

Verbreitung

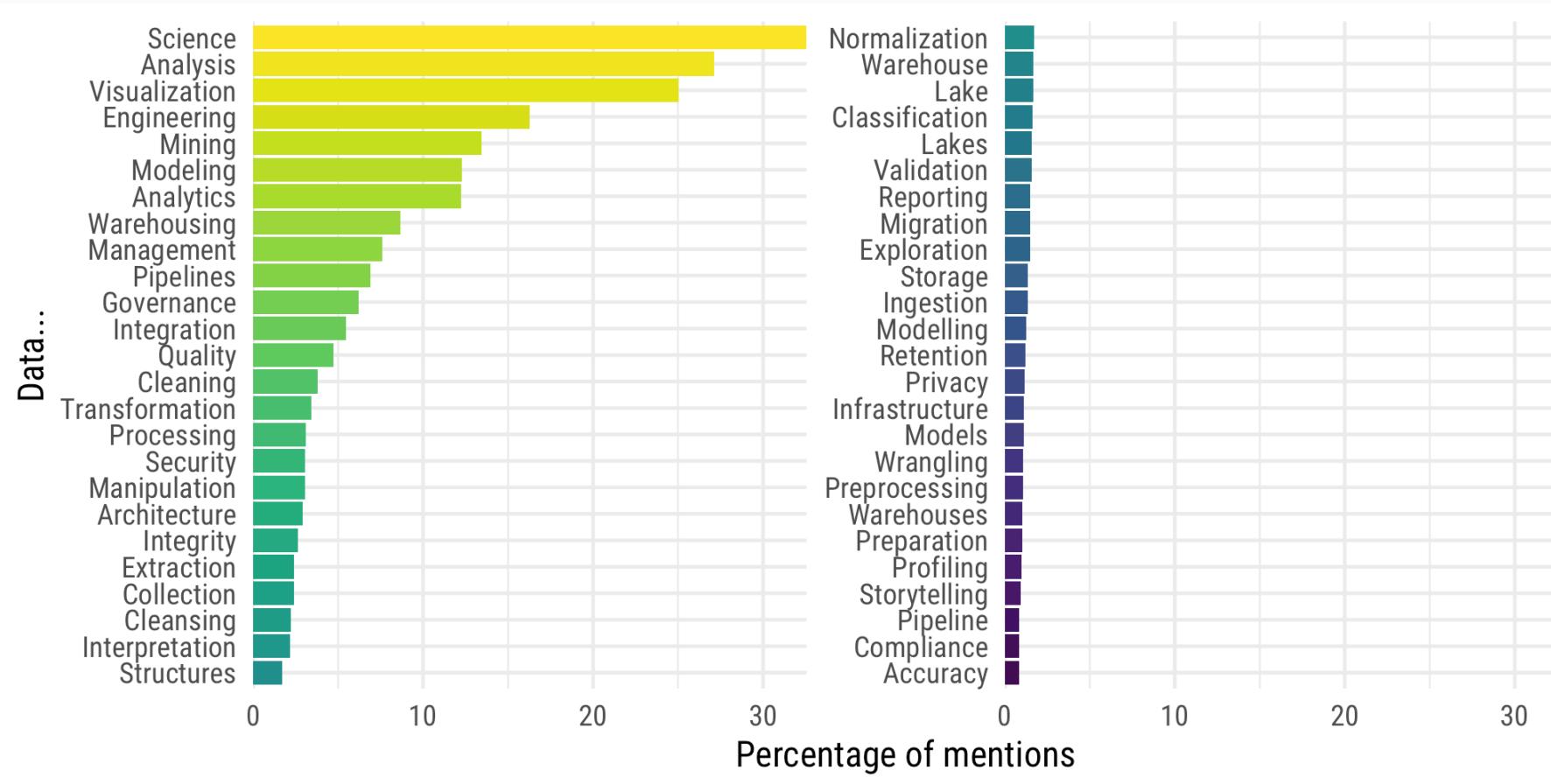
- **Kommunikation:** Öffentlichkeit, Entscheidungsträger
- **Veröffentlichen:** Zeitschriften, Software, Berichte
- **Produktivieren:** nutzbar, robust, skalierbar machen

Meta-Fähigkeit: Programmierung



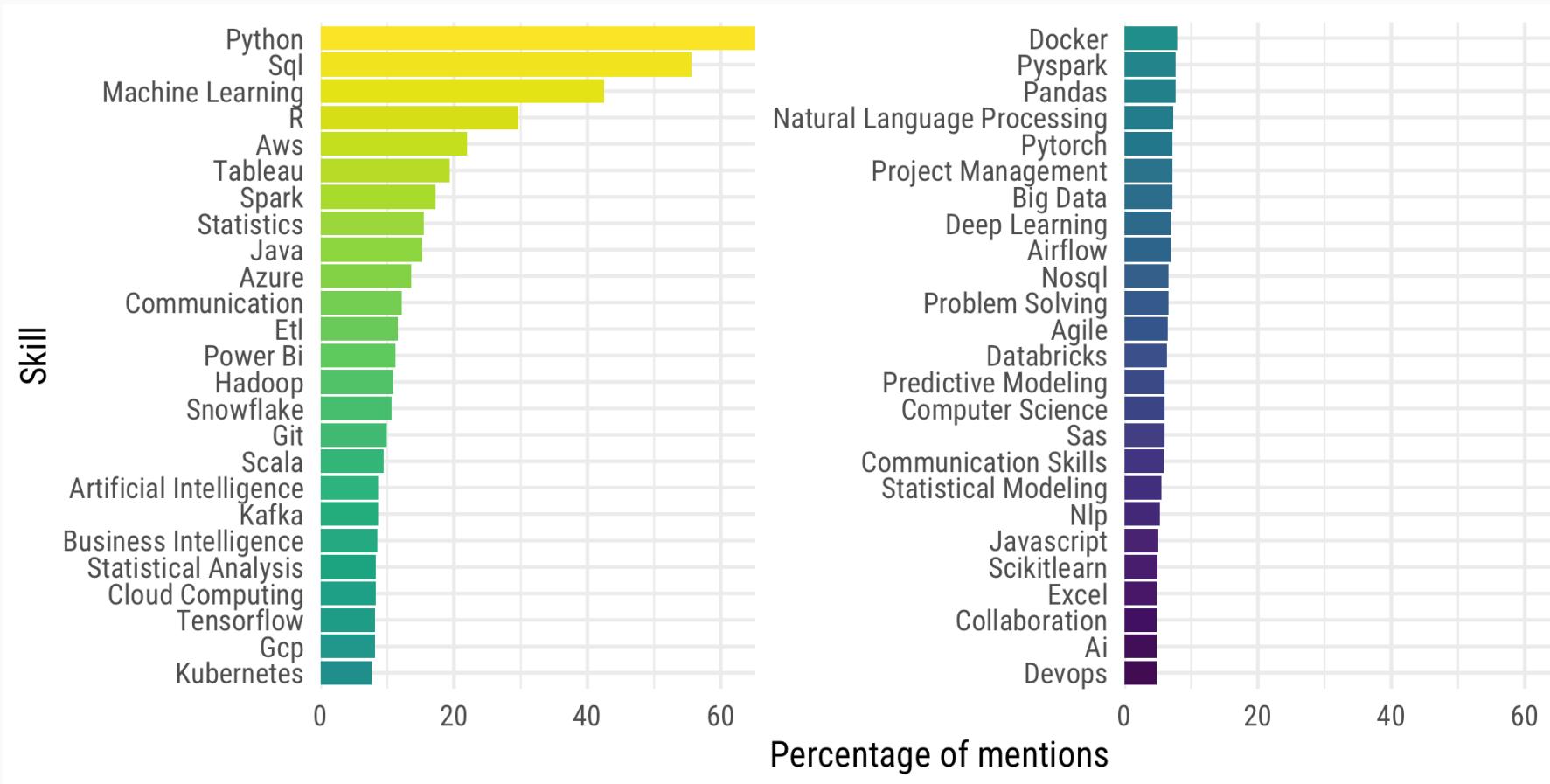
Quelle H. Wickham, M. Çetinkaya-Rundel, G. Gronemund:
R for Data Science

Anforderungen in der Industrie



Quelle Eigene Darstellung; Daten: doi.org/10.34740/kaggle/dsv/8217982; scraped „Data Scientist“ job postings from LinkedIn; n=4,342; countries: US, UK, CA, AU

Anforderungen in der Industrie II



Quelle Eigene Darstellung; Daten: doi.org/10.34740/kaggle/dsv/8217982; scraped „Data Scientist“ job postings from LinkedIn; n=4,342; countries: US, UK, CA, AU

Was kann Data Science?

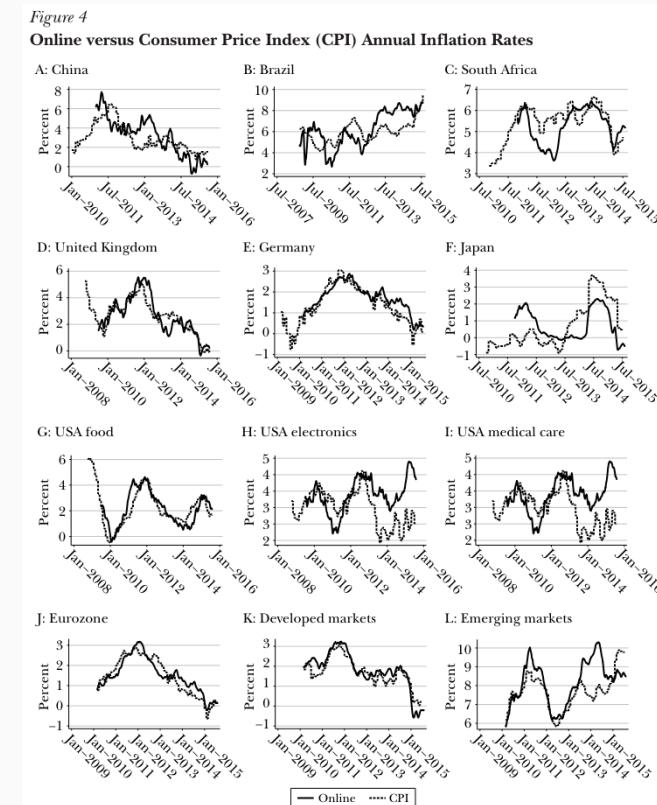
Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178

The Billion Prices Project: Using Online Prices for Measurement and Research

Alberto Cavallo and Roberto Rigobon

New data-gathering techniques, often referred to as “Big Data,” have the potential to improve statistics and empirical research in economics. This paper presents one example of how this can be achieved by using the vast number of online prices displayed on the web. We describe our work with the Billion Prices Project at MIT, and emphasize key lessons that can be used for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices.

The basic procedure used in most countries to collect inflation data has remained roughly the same for decades. A large number of people working for national statistical offices visit hundreds of stores on a monthly or bimonthly basis to collect prices for a preselected basket of goods and services. The micro data are then processed and used to construct consumer price indexes and other related indicators. This process is expensive, complex, and often too slow for some users of the data. Infrequent sampling and slow updates to the baskets can complicate



Source: Authors using online price indexes computed by PriceStats and consumer price indexes sourced from the national statistical office in each country.

Notes: Figure 4 compares inflation as measured by online prices and by the offline prices in the official consumer price index for a selection of countries, sectors, and regions. Annual inflation rates for daily online price indexes are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. The series are nonseasonally adjusted. Indexes are “all-items” with the exception of China, where an online supermarket index is shown next to the official food index. Global aggregates in the last row are computed using 2010 consumption weights in each country and CPIs from official sources.

Das Billion-Prices-Projekt (MIT)

Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178

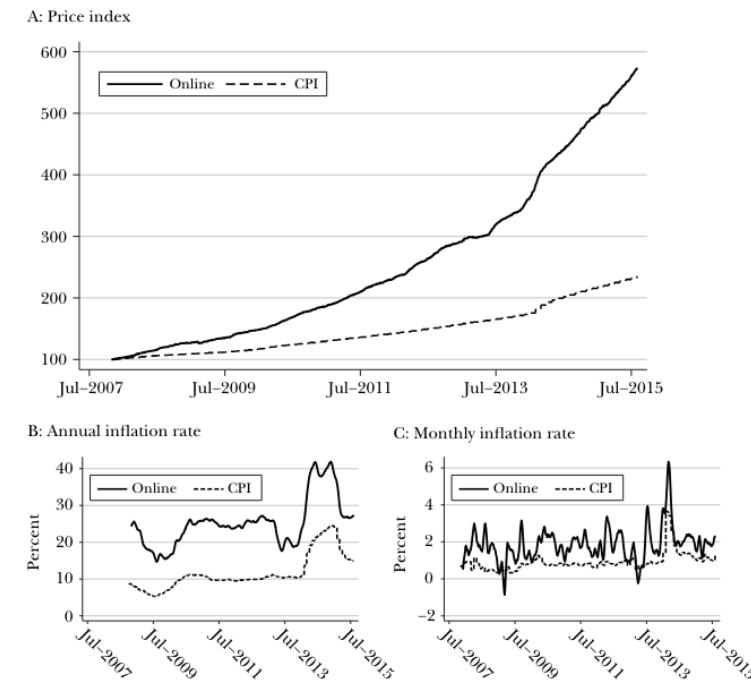
The Billion Prices Project: Using Online Prices for Measurement and Research

Alberto Cavallo and Roberto Rigobon

New data-gathering techniques, often referred to as “Big Data,” have the potential to improve statistics and empirical research in economics. This paper presents one example of how this can be achieved by using the vast number of online prices displayed on the web. We describe our work with the Billion Prices Project at MIT, and emphasize key lessons that can be used for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices.

The basic procedure used in most countries to collect inflation data has remained roughly the same for decades. A large number of people working for national statistical offices visit hundreds of stores on a monthly or bimonthly basis to collect prices for a preselected basket of goods and services. The micro data are then processed and used to construct consumer price indexes and other related indicators. This process is expensive, complex, and often too slow for some users of the data. Infrequent sampling and slow updates to the baskets can complicate

Figure 1
Argentina



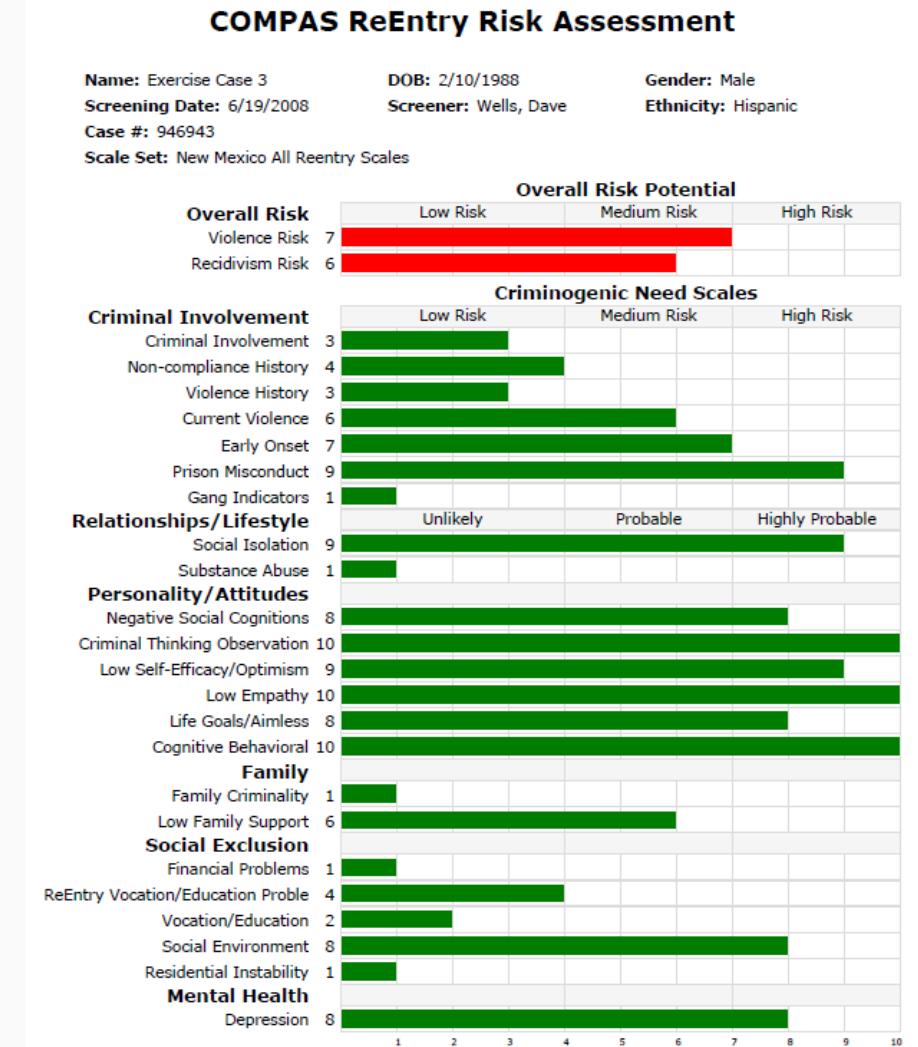
Source: Authors using online price index computed by PriceStats and the consumer price index from the national statistical office in Argentina (INDEC).

Notes: The figure compares a price index produced with online data to a comparable official consumer price index (CPI) for the case of Argentina from 2007 to 2015. It also looks at annual and monthly inflation rates using each source of data. Monthly inflation rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average a month before. Annual inflation rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. All price indexes are nonseasonally adjusted.

COMPAS: Vorhersage der Rückfälligkeit von Straftätern

Hintergrund

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) ist ein von Northpointe (jetzt Equivant) entwickeltes Entscheidungshilfeinstrument, das von US-Gerichten zur **Bewertung der Rückfallwahrscheinlichkeit** eingesetzt wird.
- Erstellt mehrere Skalen (Risiko der vorzeitigen Entlassung, allgemeine Rückfälligkeit, gewalttätige Rückfälligkeit) auf der Grundlage von Faktoren wie Alter, Vorstrafen und Drogenmissbrauch
- Der Algorithmus ist urheberrechtlich geschützt und seine inneren Abläufe sind nicht öffentlich.



Data scientists have the potential to help save the world

By Leo Borrett May 17, 2017

With an untold number of crises emerging every year, big data is becoming increasingly important for helping aid organisations respond quickly to chaotic and evolving situations.

HOW DATA SCIENCE IS SAVING LIVES



AVINASH N Sep 29 · 2 min read



For all the people first priority is about their life. Life is one of the most precious thing in the world. Can Data Science techniques save life, is it possible? Yes, using Data Science techniques to analyze large data sets today has a huge impact on saving lives.

Health

Artificial intelligence and covid-19: Can the machines save us?

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

How AI Will Save Thousands of Lives

Sepsis is the problem; data are the cure



Drew Smith, PhD Jan 10, 2020 · 5 min read ★



STUDENTS

Data Science: Why It Matters and How It Can Make You Rich

The Cambridge Analytica case: What's a data scientist to do?

The Cambridge Analytica controversy has highlighted data ethics issues especially dear to early career stage data scientists

Researchers just released profile data on 70,000 OkCupid users without permission

By Brian Resnick | @B_resnick | brian@vox.com | May 12, 2016, 6:00pm EDT

An Algorithm That ‘Predicts’ Criminality Based on a Face Sparks a Furor

Its creators said they could use facial analysis to determine if someone would become a criminal. Critics said the work recalled debunked “race science.”

Data Failed the Election, But There's Still Hope for Business Everyone is blaming data for failing to predict Trump's win. But it's the data handlers who need the real reexamination. ↗

Übung

Was kann Data Science und AI tun, was nicht?

Whiteboard | AI-Anwendungen | Performanz | Schädlichkeit

Was kann Data Science und AI tun, was nicht?

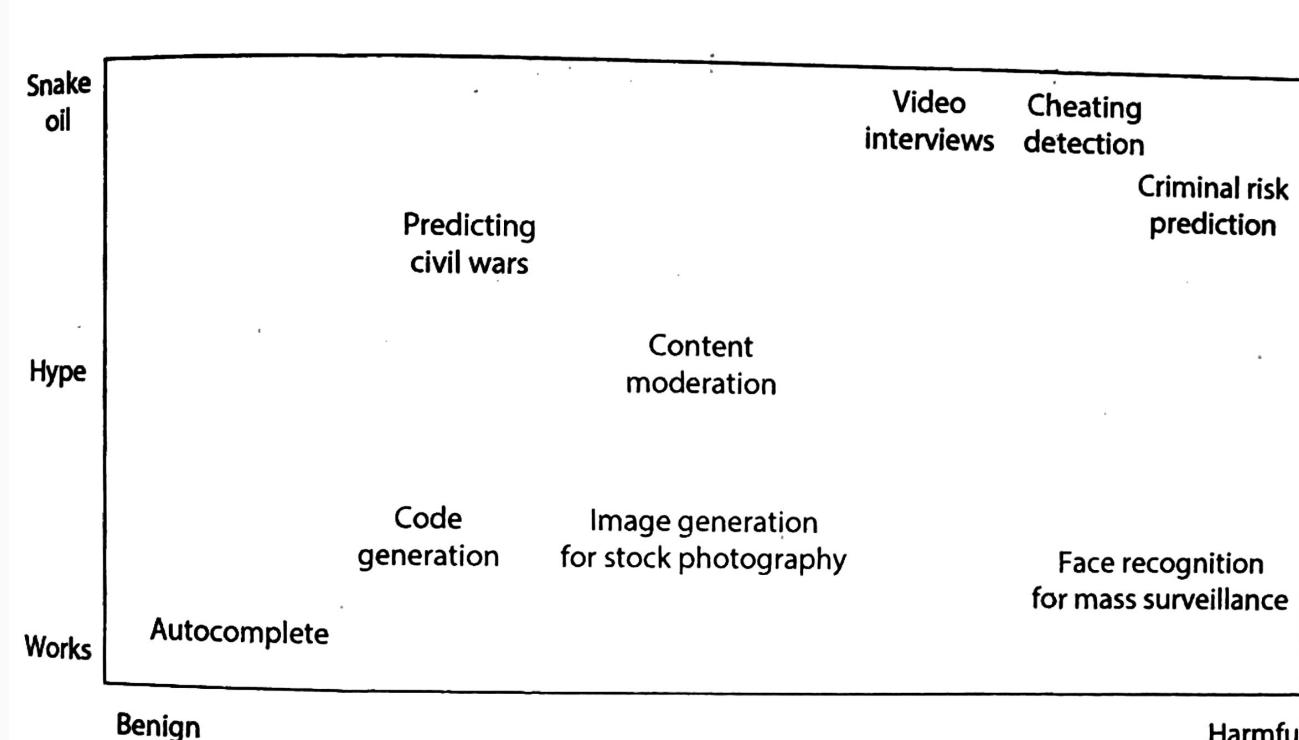


FIGURE 1.2. The landscape of AI snake oil, hype, and harms, showing a few illustrative applications.

Quelle Narayanan/Kapoor, AI Snake Oil, S.29

Ziele für dieses Modul

	Tag 1: Der Daten-Lifecycle und: Woher kommen Daten?	Tag 2: Big Data, Good Data? Die kritische Nutzung von Daten	Tag 3: Auf dem Weg zur datenbewussten Organisation
9-11 h		<i>Session 5</i> Möglichkeiten und Probleme mit Big Data	<i>Session 9</i> Daten-Workflows und Open Government Data in der Behörde
11-13 h	<i>Session 1</i> Data Science - Ein Überblick	<i>Session 6</i> Praxis-Session II: Daten weiterverarbeiten und strukturieren	<i>Session 10</i> Wie gewinnt man Data Scientists nachhaltig für Behörden?
13-14h	<i>Mittagspause</i>		
14-16 h	<i>Session 2</i> Datengenerierung und was dabei schief gehen kann	<i>Session 7</i> Informierter Konsum von Daten und Statistiken	<i>Session 11</i> Wrap-up und Ausblick
16-18 h	<i>Session 3</i> Datentypen und ihre Verwendung	<i>Session 8</i> Kommunikation von und mit Daten	
18-19 h	<i>Session 4</i> Praxis-Session I: Daten sammeln und einlesen: Web Scraping und OCR		
19h	<i>Gemeinsames Abendessen L'Osteria am Tacheles</i>		

Was wir nicht behandeln werden

Programmierung

- Kenntnisse in Python, R, SQL usw. sind für die Datenwissenschaft unerlässlich
- Die Lernkurve ist steil und erfordert viel Übung
- Wir werfen einen (kleinen) Blick hinter die Kulissen



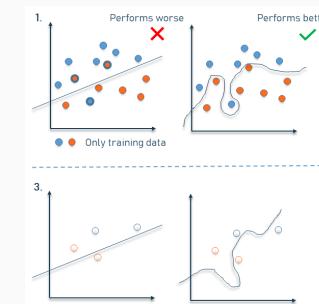
Aktive Modellierung

- Seriöse AI-Modellierung erfordert mehr theoretisches und praktisches Wissen, als wir in diesem Workshop abdecken können.
- Die Konzentration auf die Grundsätze statistischer Argumentation sollte ausreichen, um Daten und Modelle kritisch zu beurteilen.

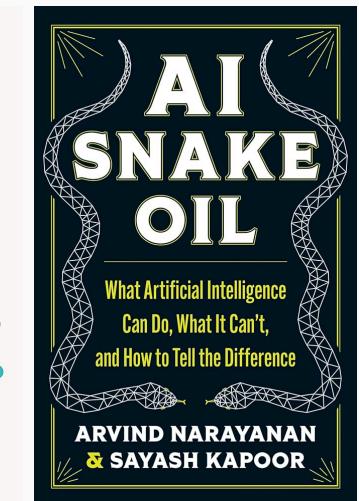
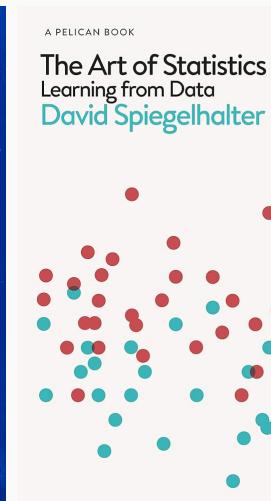
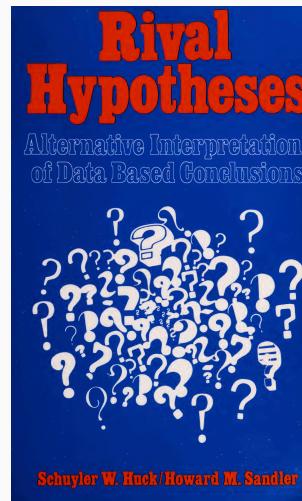
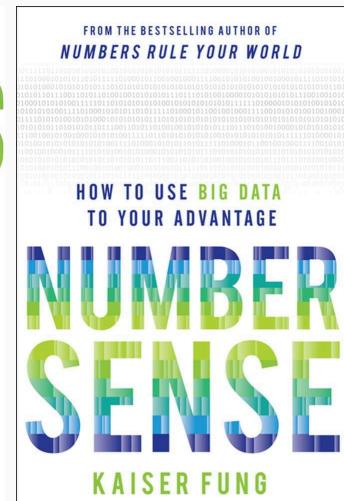
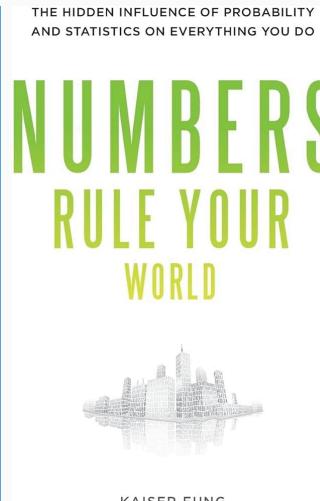
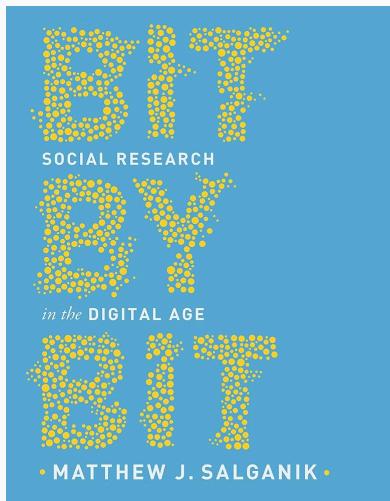
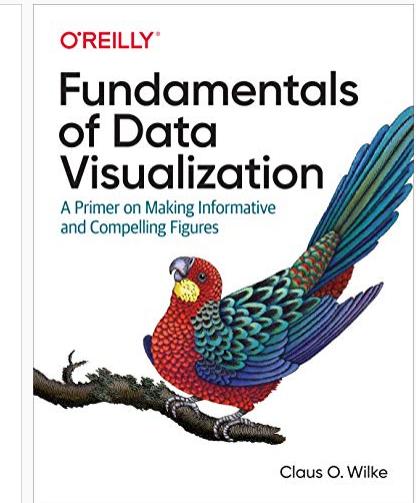
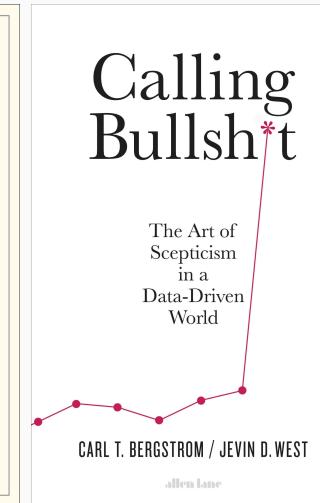
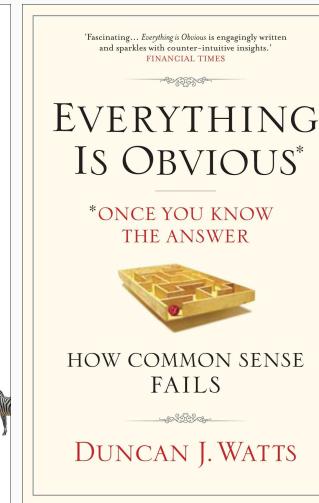
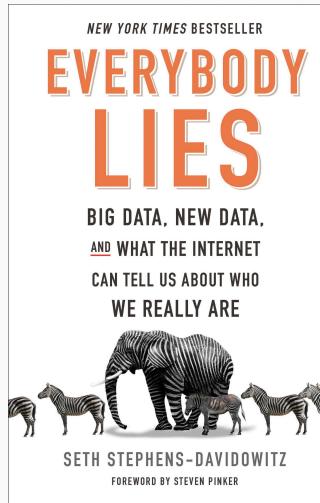
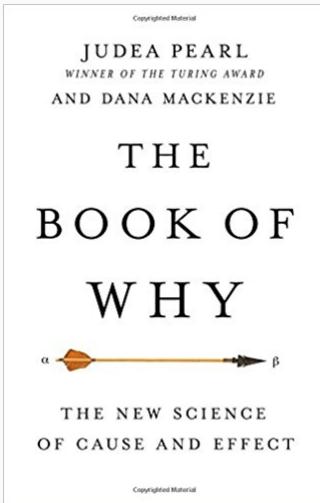
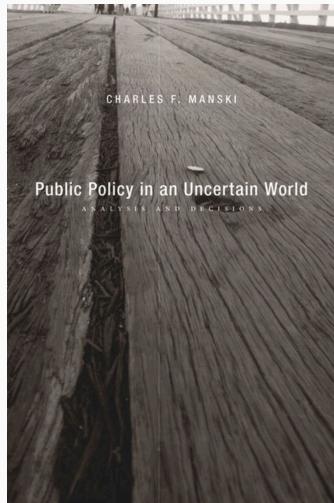


Fortgeschrittenes Maschinelles Lernen, NLP

- ML, DL, NLP sind Technologien, die viele der spannendsten Anwendungen der Datenwissenschaft vorantreiben.
- Modul 3 wird hier anschließen; wir werden uns auf die Datengrundlagen konzentrieren.



Weiterführende Literatur



Weiterführende Podcasts

