

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

Session 2: Datentypen und ihre Verwendung

Simon Munzert
Hertie School

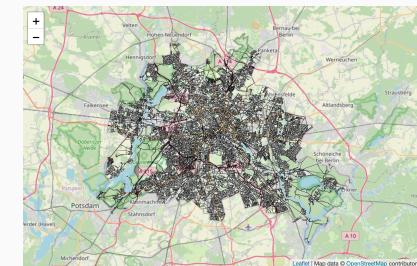
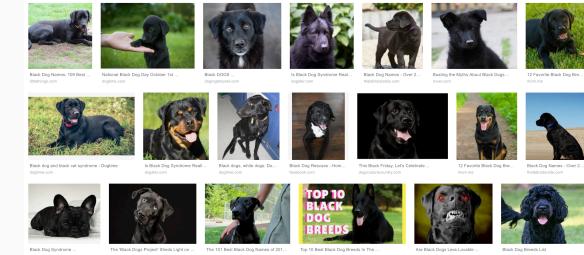
1. Datentypen - ein Überblick
2. Datenformat: JSON
3. Datentypen und Anwendungen
4. Ihre Daten

Datentypen - ein Überblick

Grundgedanken

1. Alles ist Daten.

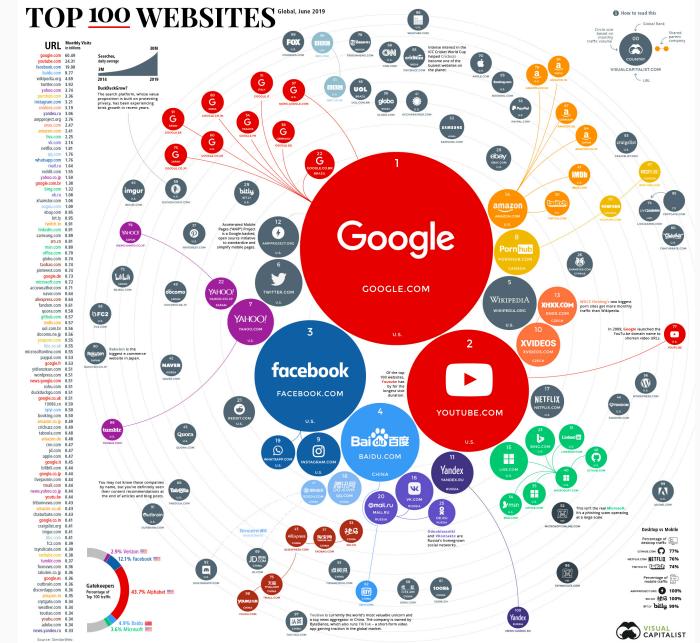
	A	B	C	D	E	F	G
1	OrderDate	Region	Rep	Item	Unit	UnitCost	Total
2	1/6/21	East	Jones	Pencil	95	1.99	189.05
3	1/23/21	Central	Kivell	Binder	50	19.99	999.50
4	2/9/21	Central	Jardine	Pencil	36	4.99	179.64
5	2/26/21	Central	Gill	Pen	27	19.99	539.73
6	3/15/21	West	Sorvino	Pencil	56	2.99	167.44
7	4/1/21	East	Jones	Binder	60	4.99	299.40
8	4/18/21	Central	Andrews	Pencil	75	1.99	149.25
9	5/5/21	Central	Jardine	Pencil	90	4.99	449.10
10	5/22/21	West	Thompson	Pencil	32	1.99	63.68
11	6/8/21	East	Jones	Binder	60	8.99	539.40
12	6/25/21	Central	Morgan	Pencil	90	4.99	449.10
13	7/12/21	East	Howard	Binder	29	1.99	57.71
14	7/29/21	East	Parent	Binder	81	19.99	1,619.19
15	8/15/21	East	Jones	Pencil	35	4.99	174.65
16	9/1/21	Central	Smith	Desk	2	125.00	250.00
17	9/18/21	East	Jones	Pen	16	15.00	240.00



Grundgedanken

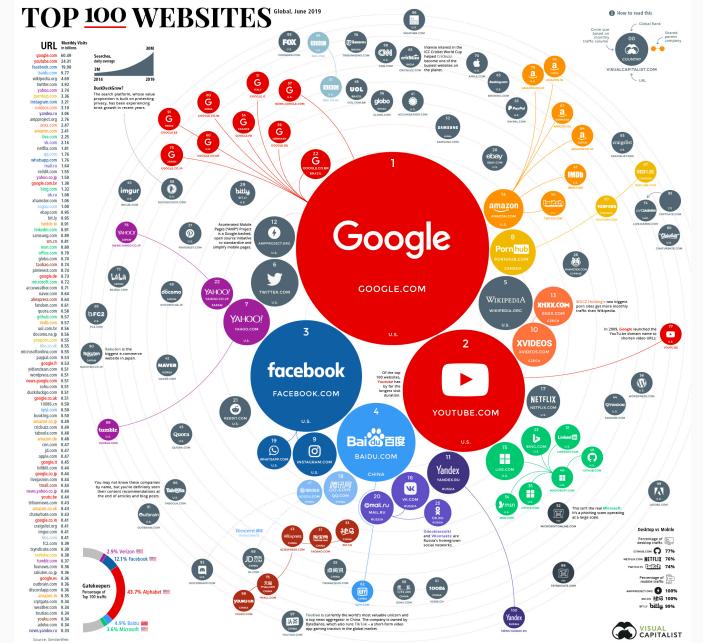
1. Alles ist Daten.

2. Daten sind nicht gleich Information.



Grundgedanken

1. Alles ist Daten.
2. Daten sind nicht gleich Information.
3. Wir können Daten klassifizieren danach, ...

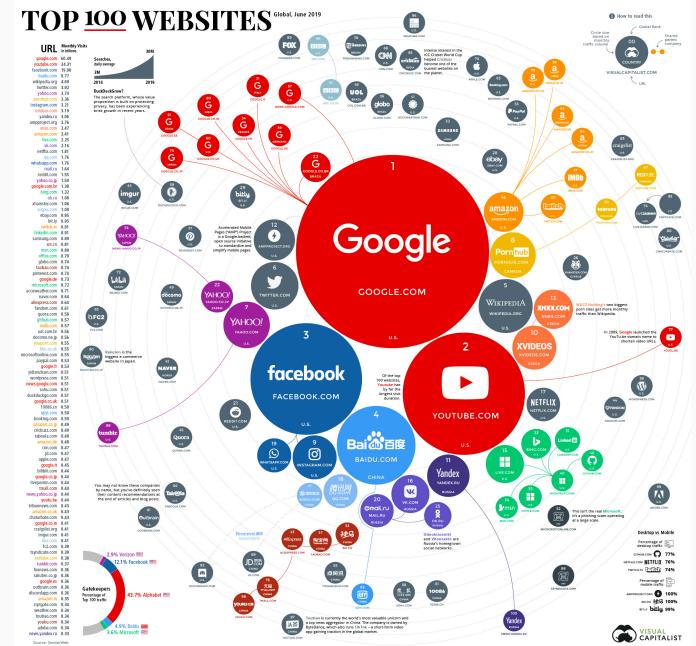


1. Alles ist Daten.

2. Daten sind nicht gleich Information.

3. Wir können Daten klassifizieren danach, ...

- wer sie generiert (Mensch vs. Maschine) bzw. woher sie stammen (z.B. *Verwaltung, Geschichte, Medizin usw.*)

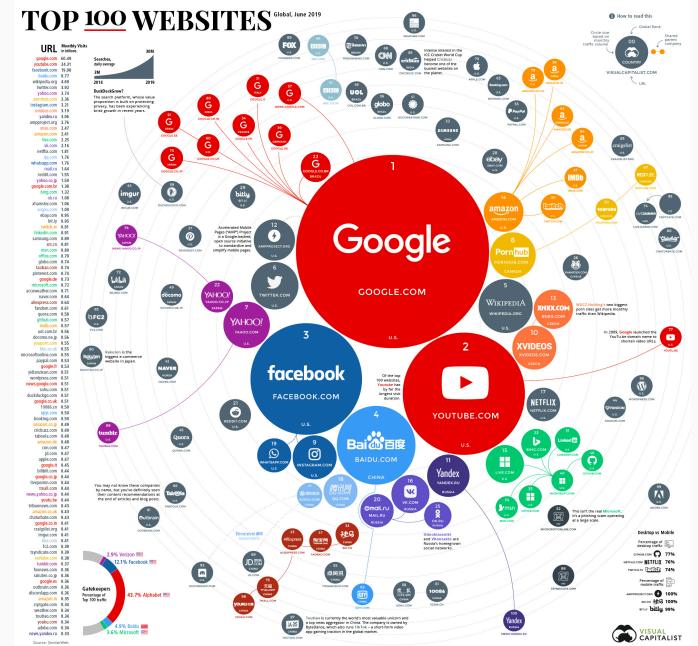


1. Alles ist Daten.

2. Daten sind nicht gleich Information.

3. Wir können Daten klassifizieren danach, ...

- wer sie generiert (Mensch vs. Maschine) bzw. woher sie stammen (z.B. *Verwaltung, Geschichte, Medizin usw.*)
 - wie bewusst sie erhoben werden (prozessgeneriert vs. aktiv erhoben)

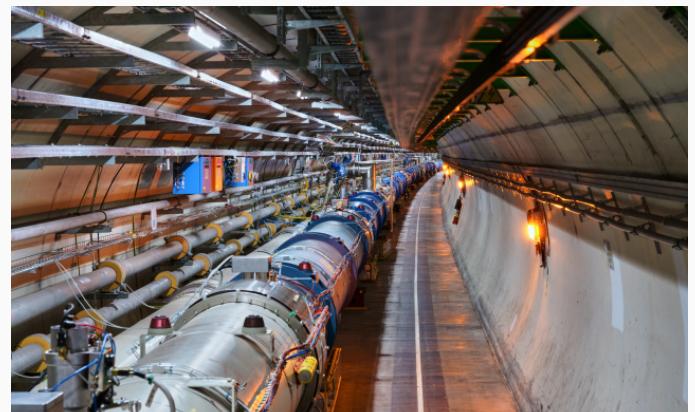
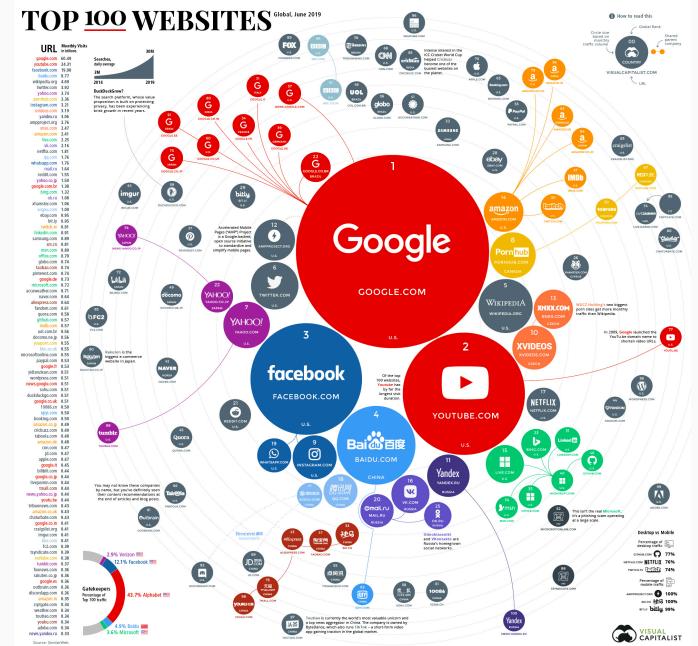


1. Alles ist Daten.

2. Daten sind nicht gleich Information.

3. Wir können Daten klassifizieren danach, ...

- wer sie generiert (Mensch vs. Maschine) bzw. woher sie stammen (z.B. *Verwaltung, Geschichte, Medizin usw.*)
 - wie bewusst sie erhoben werden (prozessgeneriert vs. aktiv erhoben)
 - wie strukturiert sie sind (maschinenlesbar vs. nicht maschinenlesbar)

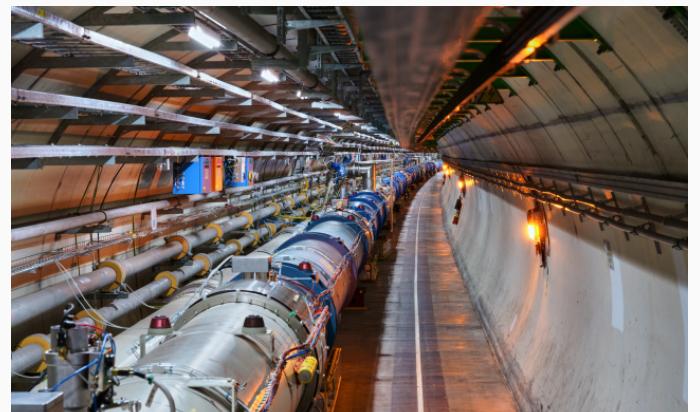
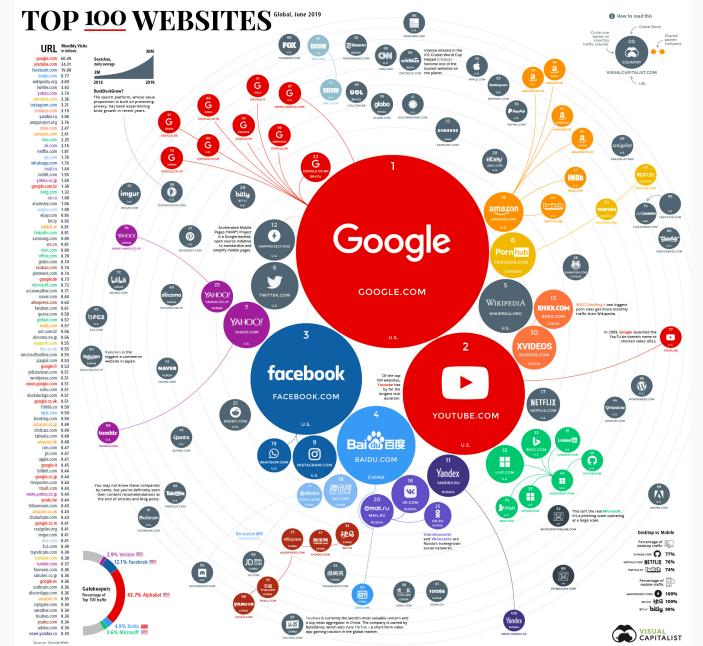


1. Alles ist Daten.

2. Daten sind nicht gleich Information.

3. Wir können Daten klassifizieren danach, ...

- wer sie generiert (Mensch vs. Maschine) bzw. woher sie stammen (z.B. *Verwaltung, Geschichte, Medizin usw.*)
 - wie bewusst sie erhoben werden (prozessgeneriert vs. aktiv erhoben)
 - wie strukturiert sie sind (maschinenlesbar vs. nicht maschinenlesbar)
 - was sie abbilden (welche Art von Information, z.B. Text, Bild, Geodaten, Netzwerkdaten)



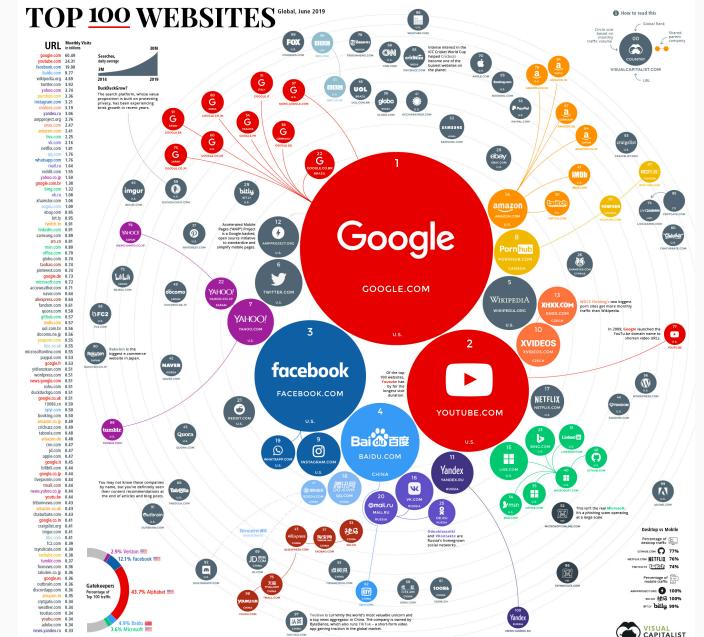
1. Alles ist Daten.

2. Daten sind nicht gleich Information.

3. Wir können Daten klassifizieren danach, ...

- wer sie generiert (Mensch vs. Maschine) bzw. woher sie stammen (z.B. Verwaltung, Geschichte, Medizin usw.)
- wie bewusst sie erhoben werden (prozessgeneriert vs. aktiv erhoben)
- wie strukturiert sie sind (maschinenlesbar vs. nicht maschinenlesbar)
- was sie abbilden (welche Art von Information, z.B. Text, Bild, Geodaten, Netzwerkdaten)

Und nach vielem mehr (Beständigkeit, Digitalisierung, etc.). Für uns eher relevant: Welche **Datentypen** sind im Data-Science- und AI-Kontext besonders relevant und wie machen wir sie nutzbar?



Datenformate

- **Datenformate** sind die Art und Weise, wie Daten gespeichert und übertragen werden.
- Sie sind **nicht** dasselbe wie Datentypen.
- Datenformate legen fest, wie Daten strukturiert und abgelegt werden, und wie sie bei ihrer Verarbeitung zu interpretieren sind.
- Es gibt eine nahezu unerschöpfliche Masse an Datenformaten, die sich zum Beispiel in der Vielfalt von **Dateinamenserweiterungen** widerspiegelt.



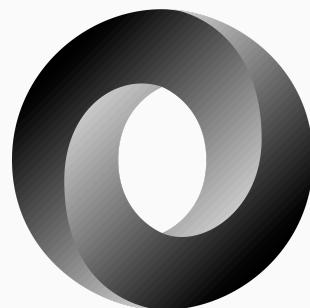
Wahl von Datenformaten

- Datenwissenschaftler haben oft mehr oder weniger ausgeprägte Präferenzen für bestimmte Datenformate und -strukturen.
- **Beispiel 1** (für einfach strukturierte Daten): Excel Binary File Format (.XLS) < Office Open XML (.XLSX) < **CSV**
- **Beispiel 2** (für semistrukturierte Daten): **JSON** vs. **XML**
- **Beispiel 3** (für vektorielle Geodaten): **Shapefile** vs. **GeoJSON**

Beispiel: Datenformat JSON

Zusammengefasst

- JavaScript Object Notation
- Beliebtes Datenaustauschformat für Webservices / APIs
- "die fettfreie Alternative zu XML"
- JSON \neq Java, sondern ein Teilmengen-Format von JavaScript
- Sehr flexibel; nicht von einer bestimmten Programmiersprache abhängig
- Import von JSON-Daten z.B. in R ist mit dem `jsonlite`-Paket unkompliziert



Daten in JSON-Format

```
[  
  {  
    "name" : "van Pelt, Lucy",  
    "sex" : "weiblich",  
    "age" : 32  
  },  
  {  
    "name" : "Peppermint, Patty",  
    "sex" : "weiblich",  
    "age" : null  
  },  
  {  
    "name" : "Brown, Charlie",  
    "sex" : "männlich",  
    "age" : 27  
  }  
]
```

JSON Daten in R-Datentabelle überführt

```
R> library(jsonlite)  
R> json_dat ←  
+ '[{ "name" : "van Pelt, Lucy",  
+       "sex" : "weiblich",  
+       "age" : 32},  
+     { "name" : "Peppermint, Patty",  
+       "sex" : "weiblich",  
+       "age" : null},  
+     { "name" : "Brown, Charlie",  
+       "sex" : "männlich",  
+       "age" : 27}]'  
R> fromJSON(json_dat, flatten = TRUE)  
##           name      sex age  
## 1   van Pelt, Lucy weiblich  32  
## 2 Peppermint, Patty weiblich  NA  
## 3   Brown, Charlie männlich  27
```

Arten von Klammern

1. Geschweifte Klammern, { und }, umfassen Objekte. Objekte funktionieren ähnlich wie Elemente in XML/HTML und können andere Objekte, Schlüssel-Wert-Paare oder Arrays enthalten.
2. Eckige Klammern, [und], umfassen Arrays. Ein Array ist eine geordnete Sequenz von Objekten oder Werten.

Datenstruktur

JSON kann komplexe Datenstrukturen abbilden (verschachtelte Objekte etc.), was die Umwandlung in flache Datenstrukturen (z. B. Tabellen) erschweren kann.

Zum Glück sind JSON-Dateien aus Webservices meist nicht sehr komplex.

Key-Value-Paare

Keys stehen in Anführungszeichen; Values nur, wenn es sich um Strings handelt.

```
"name" : "van Pelt, Lucy"  
"age" : 32
```

Keys und Values werden durch : getrennt.

```
"age" : 32
```

Key-Value-Paare werden durch , getrennt.

```
{"name" : "van Pelt, Lucy", "age" : 32}
```

Werte in Arrays werden durch , getrennt.

```
["van Pelt, Lucy", "Peppermint, Patty"]
```

Beispiel: JSON in R einlesen

- Es gibt verschiedene Pakete zum Einlesen von JSON in R.
- `jsonlite` von Jeroen Ooms: Gut gepflegt und mit überzeugenden Abbildungsregeln.

Zentrale Funktionen

Es gibt zwei zentrale Funktionen in `jsonlite`:

- `fromJSON()`: konvertiert JSON-Daten in R-Objekte nach bestimmten **Konventionen**.
- `toJSON()`: konvertiert R-Objekte in JSON-Daten

Zum Ausprobieren: [Code-Vignette](#).

Konvertierungsregeln von `jsonlite`

```
R> library(jsonlite)
R> x <- '[1, 2, true, false]'
R> fromJSON(x)

## [1] 1 2 1 0
```

```
R> x <- '["foo", true, false]'
R> fromJSON(x)
```

```
## [1] "foo"    "TRUE"   "FALSE"

R> x <- '[1, "foo", null, false]'
R> fromJSON(x)
```

```
## [1] "1"      "foo"    NA       "FALSE"
```

Datentypen und Anwendungen

"Your AI Pair Programmer"

- GitHub Copilot ist ein KI-gestütztes Tool, das Entwickler/innen beim Schreiben von Code unterstützt.
- Es wurde von GitHub und OpenAI entwickelt und basierte ursprünglich auf GPT-3.
- Copilot kann Codevorschläge machen, die auf dem Kontext des Codes basieren, den Sie gerade schreiben.
- Es kann auch ganze Funktionen schreiben, wenn man eine Beschreibung gibt (💻 Demonstration).
- Copilot ist in der Lage, in vielen verschiedenen Programmiersprachen zu arbeiten.



Datengrundlage

- Massen an Code, vermutlich von Plattformen wie GitHub, StackOverflow, etc.
- Substanzeller Impact auf Developer-Community, aber auch Kritik bzgl. Urheberrechten und Datenschutz.

Quelle [Scalablepath.com](https://scalablepath.com)

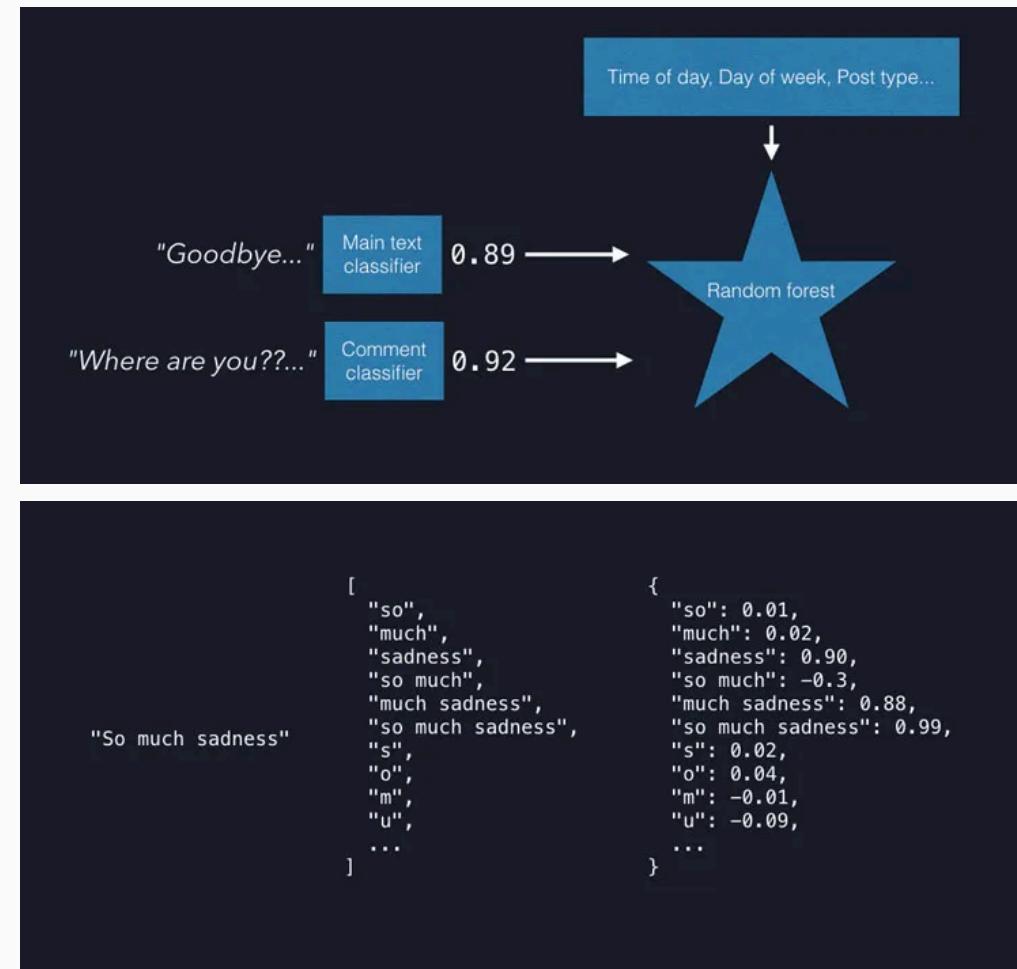
Textdaten: Facebook Suicide Prediction

Absicht

- Suizidrisiko vorhersagen mit Nutzerdaten und Interaktionen auf Facebook
- **Behauptung:** Vorhersage von Suizidgefährdung zur Prävention

Hintergrund

- **Belege:** keine konkreten Genauigkeitszahlen oder Details veröffentlicht ([s. Facebook Blogpost](#))
- **Anfechtbarkeit:** keine Zustimmung durch Nutzer/innen, keine Einsicht welche Daten verwendet wurden
- **Training der Modelle:** insb. Facebook-Kommentare und Meldungen, potenzielle Verzerrungen durch unterschiedliche Nutzergruppen und Sprachbarrieren



Quelle [Murellio et al., 2018](#)

Das Potential von Satellitenbildern und Luftaufnahmen

- Mit massiv gestiegener Auflösung sowie räumlicher und zeitlicher Abdeckung werden Satellitenbilder mittlerweile für viele Forschungs- und kommerzielle Zwecke eingesetzt
- Rohdaten alleine sind wenig nützlich: Machine-Learning-Techniken, um Muster zu erkennen bzw. inferieren

Beispiele

- Vorhersage von Armut ([Neal Jean et al., Science](#))
- Vorhersage von ökonomischer Entwicklung ([Burke et al.](#))
- Erkennung von Photovoltaik-Anlagen ([de Hoog et al.](#))

RESEARCH ARTICLES

ECONOMICS

Combining satellite imagery and machine learning to predict poverty

Neal Jean,^{1,2*} Marshall Burke,^{3,4,5*} Michael Xie,¹ W. Matthew Davis,⁴ David B. Lobell,^{3,4} Stefano Ermon¹

Reliable data on economic livelihoods remain scarce in the developing world, hampering efforts to study these outcomes and to design policies that improve them. Here we demonstrate an accurate, inexpensive, and scalable method for estimating consumption expenditure and asset wealth from high-resolution satellite imagery. Using survey and satellite data from five African countries—Nigeria, Tanzania, Uganda, Malawi, and Rwanda—we show how a convolutional neural network can be trained to identify image features that can explain up to 75% of the variation in local-level economic outcomes. Our method, which requires only publicly available data, could transform efforts to track and target poverty in developing countries. It also demonstrates how powerful machine learning techniques can be applied in a setting with limited training data, suggesting broad potential application across many scientific domains.

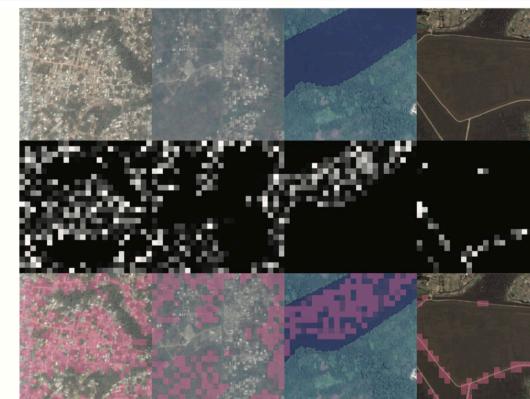


Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter “highlights” the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

Luftaufnahmen: Autobahnparkplatzerkennung

Remote Sensing of Car and Truck Parking Lots at German Highway Rest Areas using Semantic Segmentation*

Master Thesis

Benedikt N. Korbach[†]

b.korbach@students.hertie-school.org

April 29, 2024

Abstract

The widespread adoption of electric vehicles (EVs) necessitates the expansion of public charging infrastructure, particularly along highways to facilitate long-distance, low-carbon transportation. For optimal charging site planning, accurate information on parking lot size and type is crucial. Public parking inventories and community-based databases such as OpenStreetMap (OSM), however, often provide unreliable and inconsistent parking information. This paper addresses these challenges by leveraging aerial imagery to accurately estimate parking lots for cars and trucks at German highway rest areas. As a main contribution, a dataset of 246 high-resolution images of rest areas is created from publicly accessible sources, with manually verified and corrected ground-truth annotations gathered from OSM. Various semantic segmentation algorithms (U-Net, LinkNet, Feature Pyramid Network) trained on this dataset are assessed for their ability to detect specific parking lots. The best-performing model demonstrates robust performance in parking area extraction, achieving a mean Intersection over Union (mIoU) score of 0.70 on full-scale test cases. This method offers a validation technique to enhance the accuracy of parking inventory data along highways, thereby informing the strategic development of charging infrastructure.

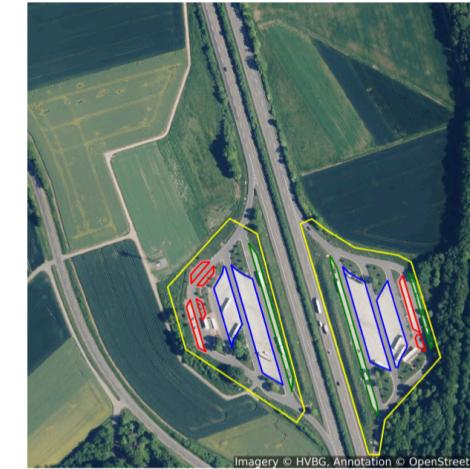


Figure 1: Example of a Correctly Labelled Rest Area. Yellow line polygons are used to indicate service station areas, blue polygons for truck parking areas, red polygons for car parking areas, and green polygons for pull-off zones.

Absicht

- Image Content Moderation mit KI: [Google's efforts to combat Child Sexual Abuse Material \(CSAM\)](#) über Google Photos; Sperren von Accounts mutmaßlicher Täter
- Launch 2018: [Google Blogpost](#); letzter Stand [hier](#)



Bildquelle

Hintergrund

- **Belege:** keine Angaben zu Accuracy des Tools; siehe auch [Narayanan/Kapoor, 2024: S. 181, 194](#)
- **Anfechtbarkeit:** Google Nutzer/innen kennen Kriterien nicht oder können Missklassifikation anfechten
- **NYT Recherche:** Familievater fotografierte geschwollenen Intimbereich seines Babys für den Hausarzt, Missklassifikation als Kindesmissbrauch und Zugang zu allen Google Services, Google SIM Karte und Google beruflicher Account gesperrt
- Zahllose weitere [Berichte](#) von gesperrten Eltern

Digitale Spurendaten: COVID-19-App-Nutzung



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

Simon Munzert¹✉, Peter Selb¹, Anita Gohdes¹, Lukas F. Stoetzer³ and Will Lowe¹

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 panel survey with mobile tracking data to measure the actual usage of Germany's official contact tracing app and reveal higher uptake rates among respondents with an increased risk of severe illness, but lower rates among those with a heightened risk of exposure to COVID-19. Using a randomized intervention, we show that informative and motivational 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.

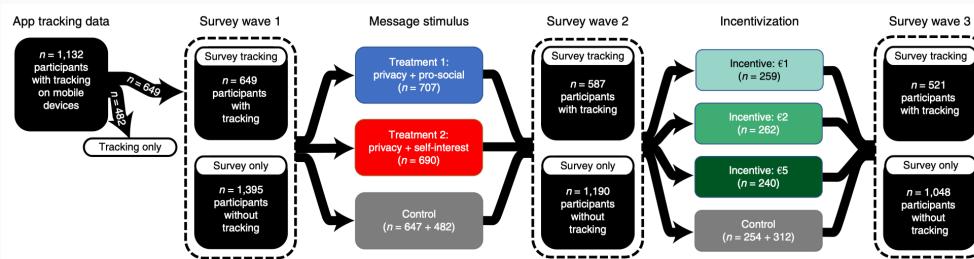


Fig. 1 | Overview of the study design. To collect tracking data (app usage, time stamps, duration and device information) from the app, members of the survey provider's passive tracking panel were incentivized to provide mobile app usage histories via passive metering software (Wakoopa). This was done from 15 June 2020 (1 d before the official app launch) until 21 September 2020 and included only panellists with Android devices. During survey wave 1, participants completed a 20-min survey about sociodemographic, attitudinal and behavioural characteristics. They were then assigned to one of two message interventions (message stimulus) or a control group. For analyses of tracked app usage, the 482 participants with mobile tracking only were used as additional controls. During survey wave 2, an average of 12 d after the initial survey, the participants were surveyed again to reassess their attitudes and behaviours. As part of the follow-up survey, self-reported non-users of the app were randomly assigned to one of three incentivization conditions or to a control group. When analysing tracked app usage, the 312 participants in the tracking-only group who did not have the app installed at the time of the follow-up survey were included as additional controls. Finally, during survey wave 3, an average of 28 d later, the survey wave 2 participants were re-invited to another follow-up survey during which their attitudes and behaviours were re-assessed.

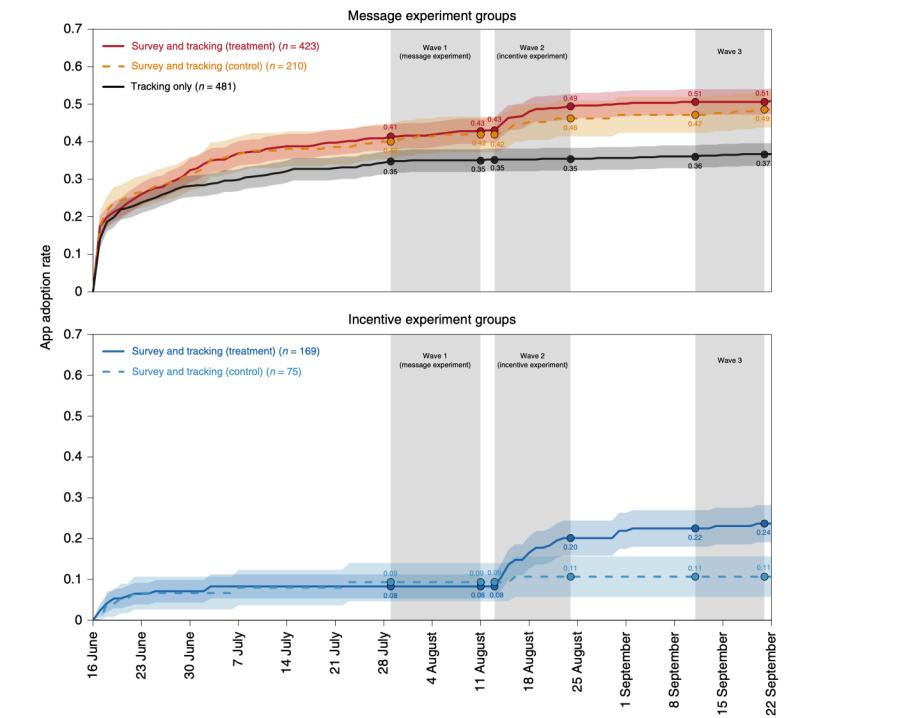


Fig. 4 | App adoption rates over time for message and incentive groups. The top plot shows the estimated rates and 83% CIs (shaded areas) for treated, control and baseline tracking-only respondents. The bottom plot shows the estimated rates and 83% CIs for treated and control respondents in the incentive group. Note that this group was limited to those who never adopted the app or uninstalled it by wave 2. Non-overlapping intervals indicate a significant difference at $P < 0.05$.

Ihre Daten

Übung

Daten-Poster-Karrussell

1. 5 Gruppen à 5 Personen
2. Jede Gruppe hat einen festen Poster-Standort.
3. Jeweils eine Person bleibt beim eigenen Poster stehen und erklärt es.
4. Die 4 anderen hören zu, stellen Fragen.
5. Nach 3–4 Minuten wird gewechselt: Die nächste Person ist dran, bis alle 5 vorgestellt haben.