## Mit Umweltdaten Gutes tun!

Vom Daten finden & säubern









#### Open Knowledge Foundation Deutschland e.V.

#### **Unsere Themen**

Offene Daten

Offene Daten sind Daten, die von jedermann frei verwendet, nachgenutzt und verbreitet werden können.

**Civic Tech** 

Zvilgesellschaftliche Interessen mit neuen und partizipativen Möglichkeiten besser umzusetzen.

**Open Government** 

Durch offenes und transparentes Regierungshandeln wird politische Rechenschaft gestärkt und Beteiligung gefördert.

Offene Bildung

Schulische und außerschulische Initiativen, die Bildung partizipativ, offen, zeitgemäß und digital gestalten.



#### **Civic Tech**

"Digital tools and technology that improve public processes and systems, promote engagement between communities and organizations, and improve the lives of individual citizens."

gov20.govfresh.com/defining-civic-innovation-definition-open-government

#### **Public Interest Tech**

"Public interest technology is exactly what it sounds like – technology used to serve the public good."

https://www.publicinteresttech.net/

#### **Datenschule / School of Data**

000

...ermutigt Organisationen durch Workshops, bewusst **Daten und Technologien bei ihrer Arbeit einzusetzen.** 

... befähigt Organisationen, mit datengestützten Projekten öffentliche Debatten evidenz-basiert mitzugestalten.

...beteiligt sich aktiv an der **digitalen Alphabetisierung** im gesellschaftspolitischen Bereich.





## Offene Daten









#### **Offene Daten**





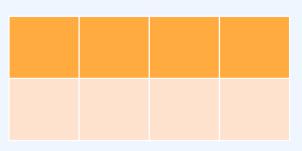
### **Technische Aspekte**

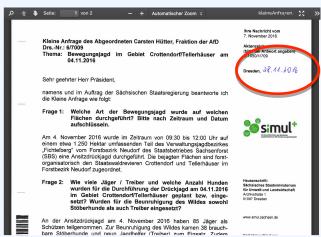
gut



schlecht







Grafik: <u>kleineAnfragen.de</u>

#### Rechtliche Aspekte / Offene Lizenzen



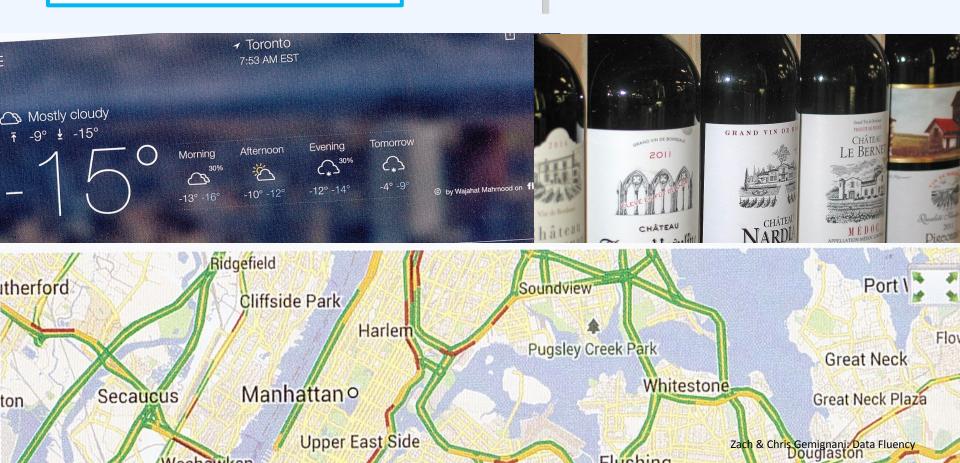
Gemeinfreiheit/ Public Domain



Public Domain dedicated

## **Creative Commons** Namensnennung frei nutzbar Namensnennung-Weitergabe unter gleichen Bedingungen Namensnennung-KeineBearbeitung Namensnennung-NichtKommerziell Namensnennung-NichtKommerziell-Weitergabe unter gleichen Bedingungen Namensnennung-NichtKommerziell, KeineBearbeitung

#### **Daten im Alltag**



Wer seid ihr?

Was interessiert euch an Daten?

Was erwartet ihr vom Workshop?



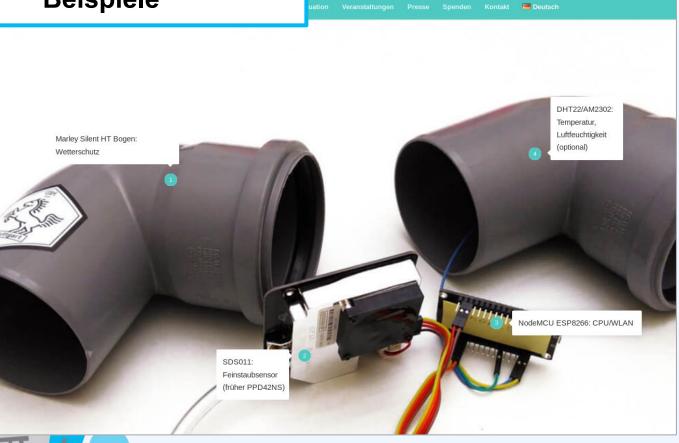
# Nutzung von Umweltdaten

**Gute Beispiele** 

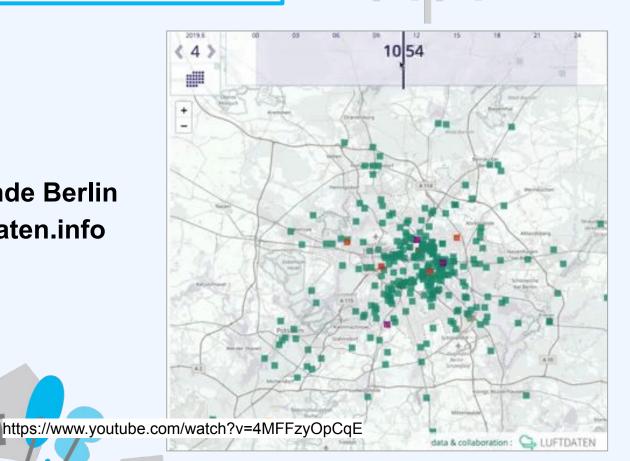




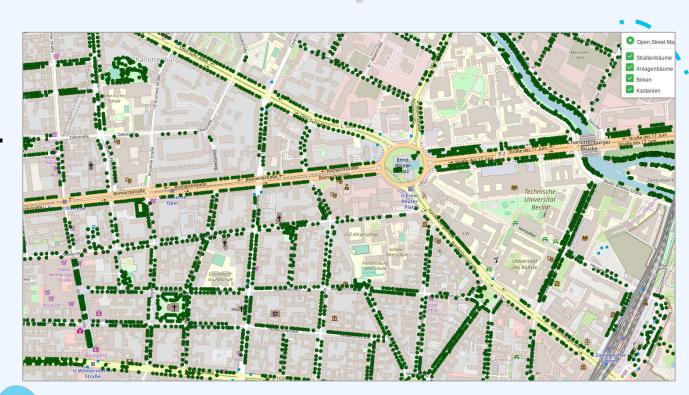








Baumkatasterdaten zugänglich machen



# Wasserqualität in Potsdam





# Data Pipeline

Methodologie









## **Data Pipeline**







## **Data Pipeline**



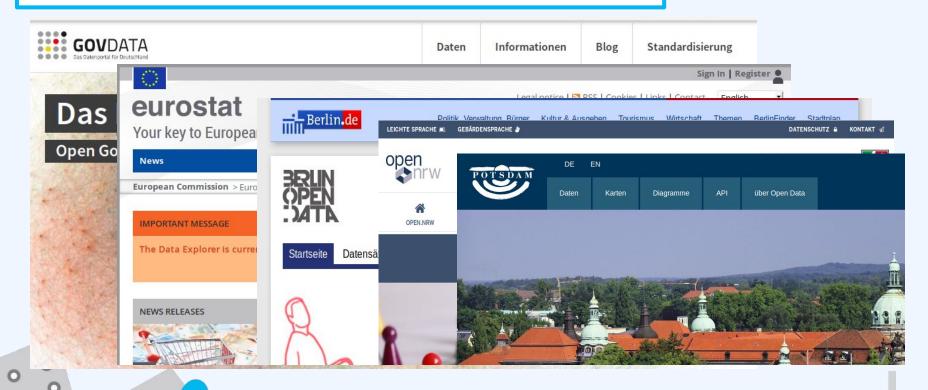






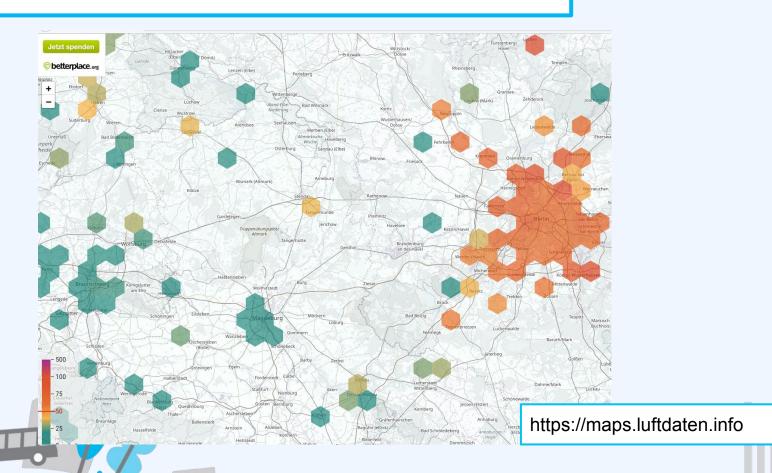
- Verwendung von Suchmaschinen
  - erweiterte Suchoperationen nutzen (<u>siehe unser</u>
     <u>Lernmaterial</u>)
  - außerhalb der eigenen Filterblase suchen (verschiedene Suchmaschienen, Tor Netzwerk)
- wenn möglich, verschiedene Quellen nutzen
- Datenportale

#### **Datenportale**



https://opendata.potsdam.de

#### Citizen Science: Luftdaten selber messen



#### Institutionen



WETTER

KLIMA UND UMWELT

FORSCHUNG

LEISTUNGEN

DER DWD

Startseite > Leistungen





Suche Stichwort eingeben

ок

Home

Grundlagen

Steckbriefe

Datenrecherche

Ergebnisse

Service

Glossar

Inhaltsübersicht | English

Sie sind hier: Home

Wir untersuchen die Belastung des Menschen und seiner Umwelt durch Chemikalien in verschiedenen Lebensräumen. Seit 1985 sammeln wir Jahr für Jahr Tausende von Proben und lagern sie dauerhaft.

Gehen Sie auf Spurensuche im Archiv der ökologischen und toxikologischen Beweissicherung!



Zeitbezüge



alyte



#### Ausgewählte Ergebnisse



Benzo[b]naphtho[2,1-d]thiophen in Miesmuscheln aus dem Jadebusen

Deutlich abnehmende Belastung seit Ende der 1980er Jahre

>> WEITERE ERGEBNISSE

#### Im Fokus: Der Boden

Boden

Am 5. Dezember ist Weltbodentag. Er geht zurück auf eine Initiative der Internationalen Bodenkundlichen Union (IUSS) im Jahr 2002. Im selben Jahr wurde Boden in das Untersuchungsprogramm der

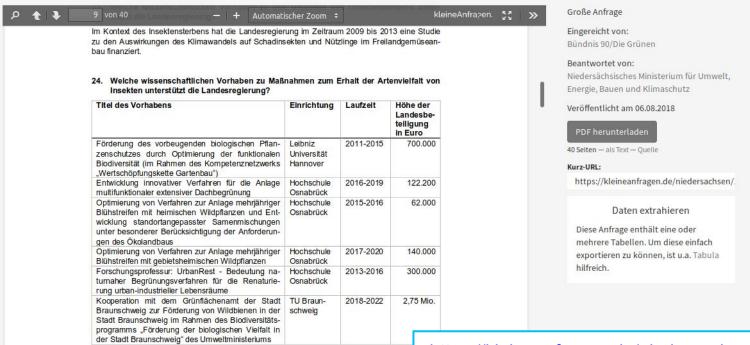
https://www.umweltprobenbank.de/de

#### Civic Tech: kleineAnfragen.de

Was tut die Landesregierung, um das Insektensterben zu stoppen, und wann werden Bienengifte wie Neonicotinoide verboten?

defere ejekt die Landesmalerung Urseehen des Artensehuunde in der Intenside

18/1347



https://kleineanfragen.de/niedersachsen/18/1347

#### IFG & FragDenStaat

FragDenStaat.de

Angefragt

Blog

FAQ

Stellen Sie eine Anfrage!

Suchen





## Was ist FragDenStaat?

Jede Person hat das Recht auf Informationen.

FragDenStaat hilft Ihnen, Ihr Recht wahrzunehmen.

Fragen Sie über diese Plattform Behörden in

Deutschland nach Informationen und Dokumenten!

Suchen Sie in 12428 Anfragen und 11756 Behörden:

z.B. Schule oder NSA

Q Suchen!

Informationsfreiheit
FragDerStaat.de



fragdenstaat.de

### **Datensammlung**

SOIT\_OIGH = DE OCATO] = HEIG\_HCERISE\_WDGGH700A1303

#### Umweltdatensätze, Stories & Projekte

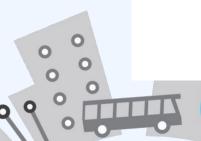
Offene Umweltdaten

Allgemeine Daten que Umweltdatensätze, St

- Thrude //thru.de, Umweltdaten von Industriebetrieben: Emissionen, Schadstoffe, Abfälle, etc.
- //Luftdaten.info, Citizen Science-Projekt, sammelt selbst erhobene Daten Feinstaubbelastung in Deutschland und anderen Ländern, inkl. Anleitung eigenen Feinstaubmessers

https://ogy.de/uds

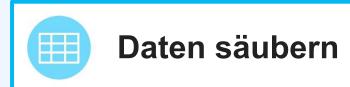
- //farmsubsidy.org, Agrarsubventionen auf Europäischer Ebene, gesammelt von verschiednen Stellen
- Climate Data Center vom Deutschen Wetter Dienst (DWD) //opendata.dwd.de
- CORINE Land Cover //land.copernicus.eu, Datensatz der Oberflächennutzung durch Satellitenbilderkennung. Original Satellitenbilder //code-de.org
- Umweltverträglichkeitsprüfungen (UVP) //uvp-portal.de, hauptsächlich PDFs
- Elektromagnetische Felder (EMF) Datenbank der Bundesnetzagentur //emf2.bundesnetzagentur.de
- CodeCheck //corporate.codecheck.info/Produktdatenbank, zu Inhaltsstoffen von Kostmetika etc (Privatunternehmen)
- //nanopartikel.info, Produktinformationen zu Nanopartikeln
- ROSYS Rohstoffinformationssystem //rosys.dera.bgr.de
- Observatory of Economic Complexity (OEC), Internationale Handelsströme //atlas.media.mit.edu
- Service you Correctly legarice correctly ended by place EOO Macon united by staigandon



#### Daten bekommen

- Download von maschinenlesbaren Dateien
  - z.B. CSV, TSV, JSON, XML
- Scraping von PDF-Dokumenten und Webseiten
  - Programmierung (z.B. Python)
  - Freie Tools (<u>Web Scraper</u>)





- sind die Daten für den Menschen oder die Maschine aufbereitet?
- werden alle Daten benötigt oder nur eine gefilterte Version?
- immer Arbeitskopie erstellen!



#### Daten säubern

- Korrekter Zeichensatz eingestellt
- Korrespondierende Werte sind gleich geschrieben (U.S. ≠ USA)
- Keine Leerzeichen am Anfang und Ende (aka. trailing Spaces)
- Alle Spalten sind eindeutig benannt (keine Dopplungen)

- Keine versteckten Zeilen/Spalten
- Einheitliche Formatierung (z.B. Punkt oder Komma zwischen Zahlenwerten)
- Einheitliche Kennzeichnung der Abwesenheit eines Wertes

#### **Tools: Tabula**



#### Tabula



Tabula is a tool for liberating data tables locked inside PDF files.

View the Project on GitHub



Current Version: 1.2.1

Other Versions: pre-releases & archives

Need help? Open an issue on Github.

**Donate:** Help support this project by backing us on OpenCollective.

We'd love to hear from you! Say hi on Twitter at @TabulaPDF Latest Version: Tabula 1.2.1

June 4, 2018

https://tabula.technology

Tabula 1.2.1 fixes several bugs in the user interface and processing backend. (You can read about all the changes in the release notes.)

Download Tabula below, or on the release notes page

Special thanks to our OpenCollective back if you find Tabula useful in your work, pleadonation.

Lernmaterial zu Tabula https://tinyurl.com/ds-tabula

#### How Can Tabula Help Me?

If you've ever tried to do anything with data provided to you in PDFs, you know

Beispiel: Jährliche Auswertung Feinstaub 2018 https://tinyurl.com/uba-feinstaub

Willo Oses Tabala.

Tabula is used to power investigative reporting at news organizations of all sizes, including ProPublica, The Times of London, Foreign Policy, La Nación (Argentina), The New York Times and the St. Paul (MN) Pioneer Press.

### **Tools: Open Refine**





#### OpenRefine

A free, open source, powerful tool for working with messy data



Home

Community

Documentation

Download

Contact Us

Blog

Enhanced with Java profiler

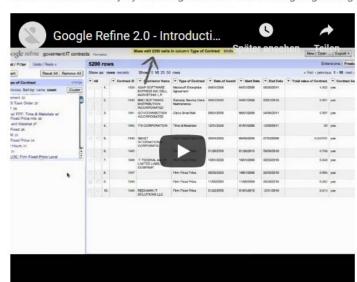


**JPROFILER** 

#### Introduction to OpenRefine

#### 1. Explore Data

OpenRefine can help you explore large data sets with ease. You can find out more about this functionality by watching the video below and going through these articles



http://openrefine.org/

#### **Excel, LibreOffice Calc, Google sheets**

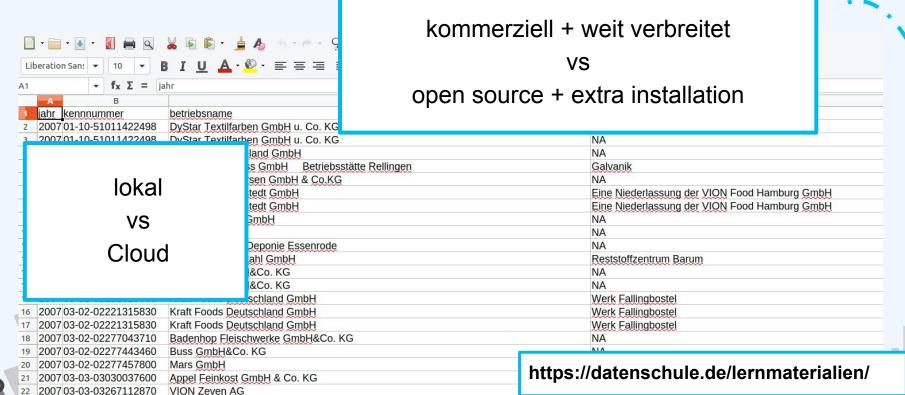
23 2007 03-03-03267112870

2007 03-03-03267112870

VION Zeven AG

VION Zeven AG





NΔ

#### Säubern: Funktionen



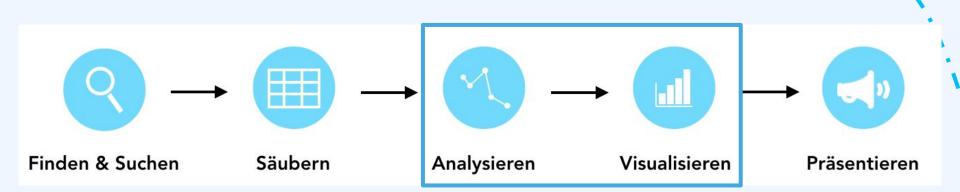
Helfen bei der Vereinheitlichung der Formatierung

- SÄUBERN()
- GLÄTTEN()
- SPLIT()
- GROSS()
- KLEIN()
- GROSS2()

Auf eine Zelle angewendet oder auf Spalten/ Reihen (**ArrayFormula()**).



# **Data Pipeline**







# Be a statistician

**Grundlagen & interaktive Übung** 









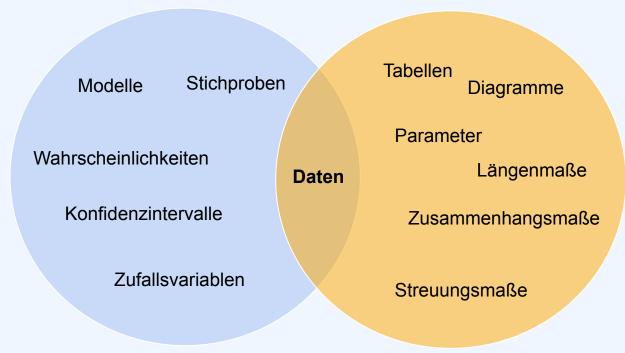
# **Statistik**

= die Wissenschaft des Organisierens,Analysierens, Sammelns und Darstellens von Daten





#### **Statistik**



**Mathematische/ induktive Statistik** 

**Deskriptive Statistik** 

#### Maße der zentralen Tendenz

= zusammenfassende Statistiken, die darauf abzielen, eine Menge von Daten in einer bestimmten Anzahl zu beschreiben.

- Mittelwert der Durchschnitt: Alle Datenpunkte dividiert durch die Anzahl der Beobachtungen
- Modus der häufigste Wert: Datenpunkt, der im Datensatz am häufigsten vorkommt
- Median der Wert der Mitte: der Wert, der den Datensatz in zwei gleiche Hälften teilt (50/50)

### Streuungsmaße

= zusammenfassende Statistiken, die die Streuungsbreite der Werte beschreiben.

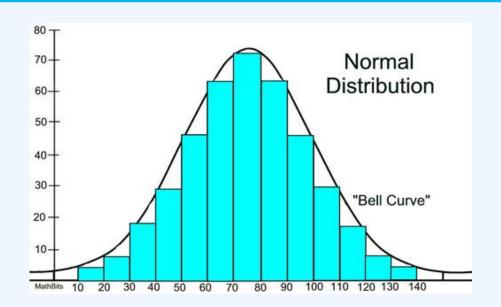
- Spannweite:
  - Minimum: Niedrigster Wert im Datensatz
  - Maximum: Höchster Wert im Datensatz
- Standardabweichung: misst, was "normal" oder erwartet wird; durchschnittliche Entfernung zum Mittelwert
- Varianz: Quadrat der Standardabweichung

### Normalverteilung

Zentraler Grenzwertsatz: Die Durchschnittswerte der Zufallsvariablen konvergieren zur Normalverteilung, wenn die Anzahl der Beobachtungen hoch genug ist

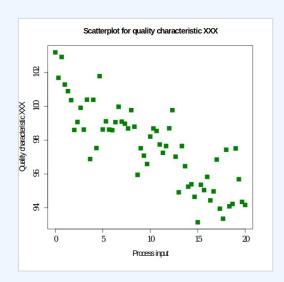
x-Achse: Wert

y-Achse: Häufigkeit





#### **Korrelation**



= statistische Beziehung zwischen zwei Variablen

- y = abhängige Variable (zu erklärende Variable),
   z. B. Schuhgröße
- x = unabhängige Variable (erklärende Variable),
   z. B. Länge des Fußes



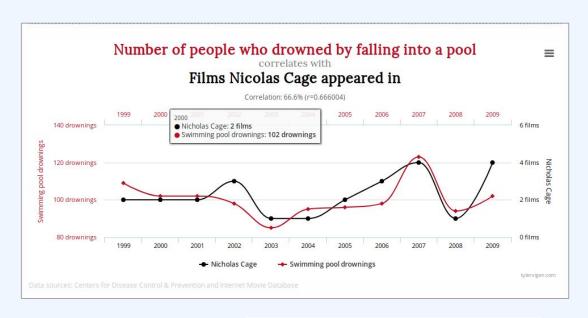
#### Korrelation vs. Kausalität

#### Aber: Korrelation bedeutet

nicht Kausalität!

#### Beispiel:

Anzahl der Leute die in Pools ertrunken sind und Anzahl der Filme mit Nicolas Cage



http://tylervigen.com/spurious-correlations

### Analyse in Excel et al



#### Zentrale Tendenz

- Modus: MODUS.EINF() MODE()
- Median: **MEDIAN()**
- Average: MITTELWERT() AVERAGE()

#### Streuungsmaße

- MIN() MAX()
- Spannweite
- Standardabweichung STABW() STDEV()



#### **Mehr Tabellen! Pivot!**



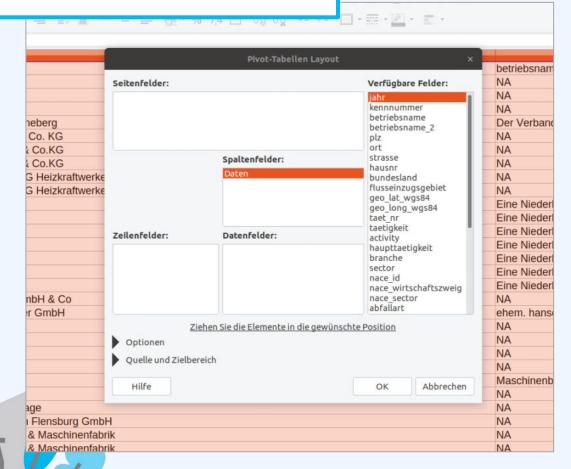
- Alle Daten markieren
- Daten > Pivot Tabelle > Erstellen
- Im Dialogfenster Aktuelle Auswahl > OK





#### Mehr Tabellen! Pivot!





# Daten visualisieren







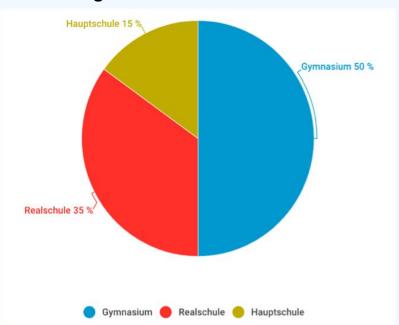




# **Pie Chart**



#### Verteilung der Schularten in Deutschland



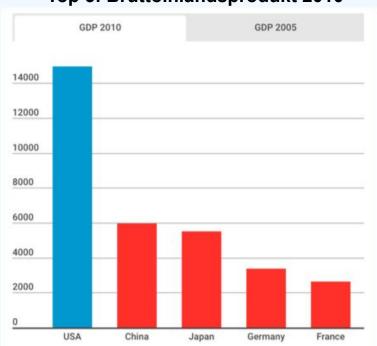
# **Bar Chart**





BIP (Mrd.)

**Top 5: Bruttoinlandsprodukt 2010** 



Länder (top 5 Industrienationen)

# NBC2 VIEWER VOTE

NBC-2.COM

# HOW CONCERNED ARE YOU ABOUT THE ZIKA VIRUS?





# **Line Chart**

# Bruttoinlandsprodukt 2010 - 2015 (USA, China, Japan)

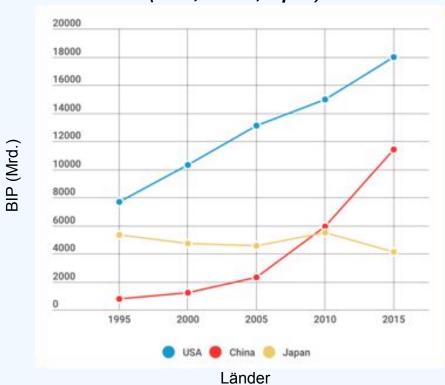
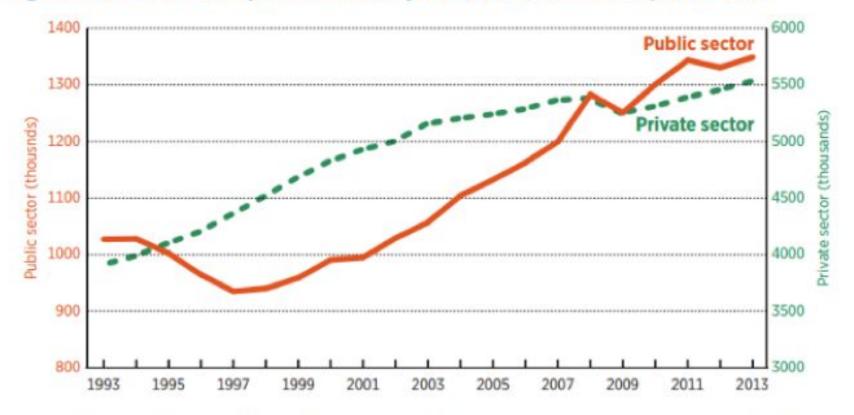






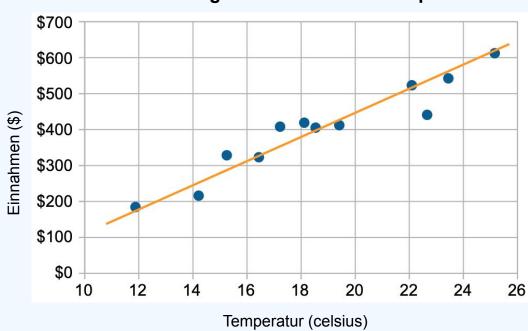
Figure 10: Public- and private-sector jobs (000s) in Ontario, 1993-2013



Source: Statistics Canada, CANSIM Table 282-0089: employment by class of worker and sex, seasonally adjusted and unadjusted; Ontario; Public sector and private sector employees; Both sexes; Seasonally adjusted (x 1,000).

# **Scatterplott**

#### Zusammenhang: Eisverkäufe und Temperatur





#### Daten kommunizieren



- Usability over Design
  - lieber simpel und klar als komplex und kompliziert
- Kernaussagen klar herausstellen





### Allgemeine Regeln

Konzentriere dich darauf, welche Information du vermitteln möchtest

Eine Aussage pro Grafik

Was ist relevant?

Alles, was deine Aussage unterstützt

Alles, was Kontext herstellt

Überschriften

Erkläre was der Leser sieht

#### Halte es Einfach!

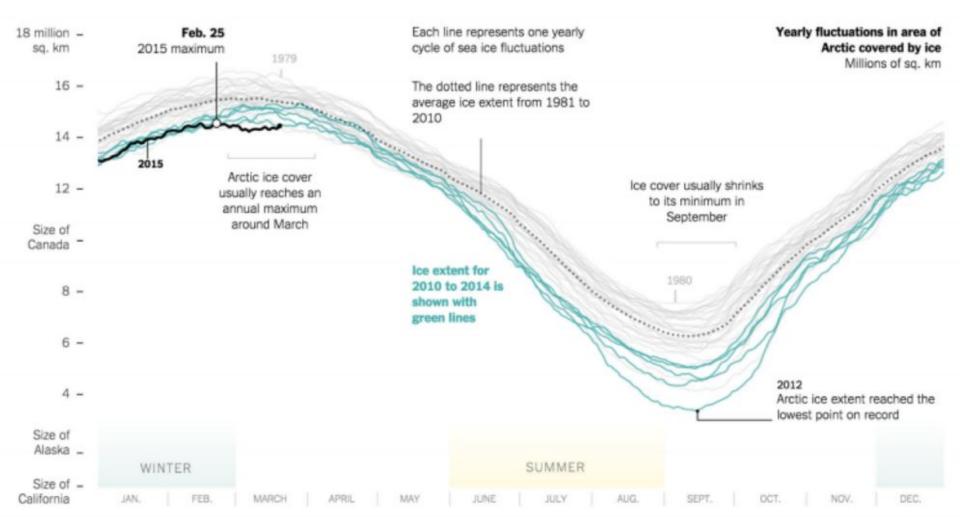
Reduziere die Informationen

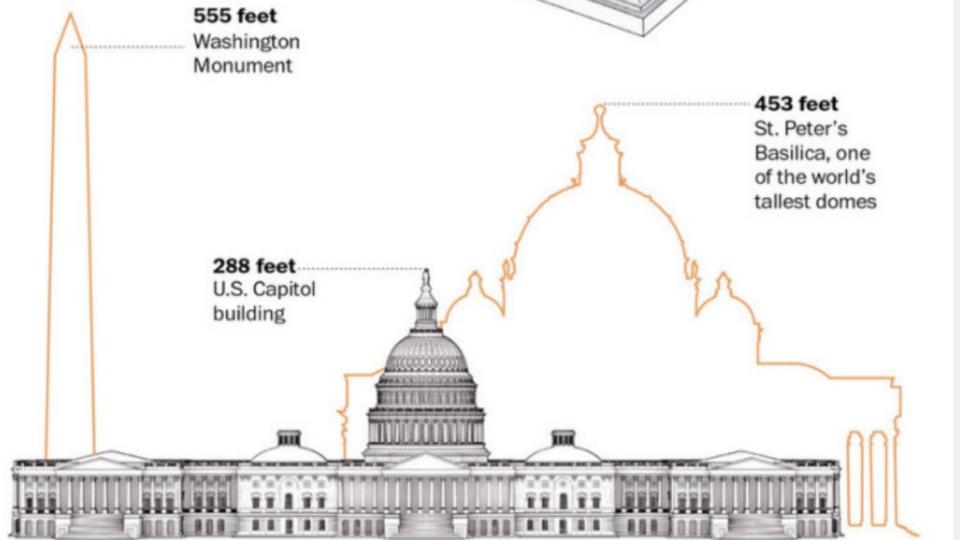
Weniger Farbe (max 7),

Graue/Bedeckte Hintergrund

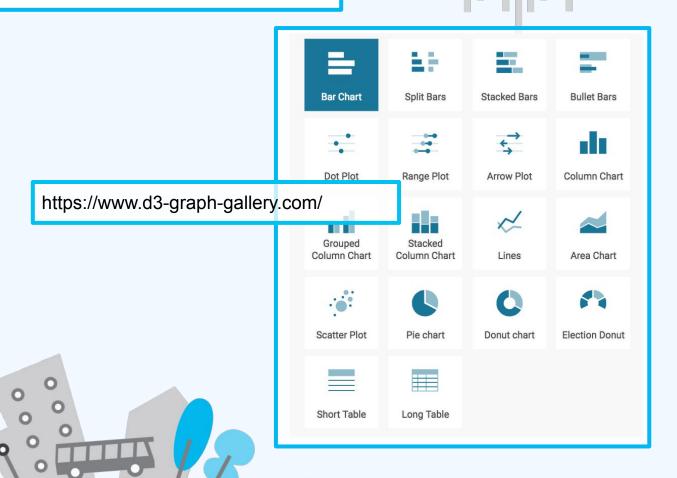
Verwende logarithmische Skalen nur, wenn dies sinnvoll Erscheint!

Check Mehrdeutigkeit



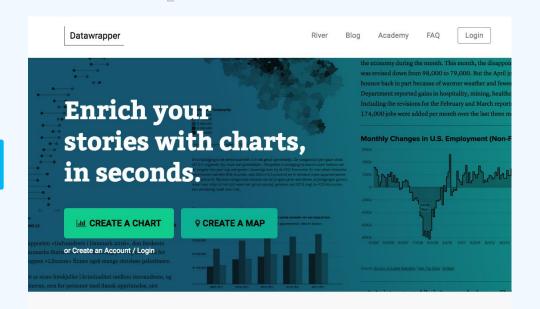


# **Arten von Visualisierung**



# **Data Wrapper**

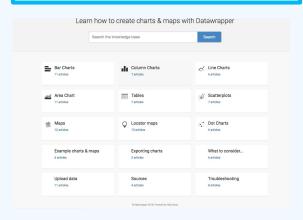
https://www.datawrapper.de/

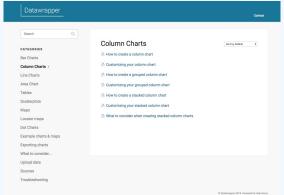


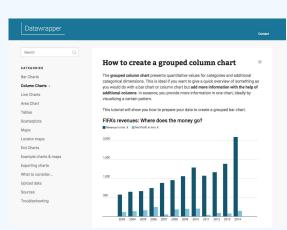


# **Data Wrapper**

#### Academy







https://academy.datawrapper.de/article/21-how-to-create-a-grouped-column-chart



# **Vorbereitung: Daten**

Header

Column - Kategorien

Column - Numerische Werte

Pivot Tabelle

Paste Values Only

Keine Gesamtsumme

	female	male
CDU	442000	14252500
CSU	1067250	14148250
DIE GRÜNEN	51500	273250
DIE LINKE	189000	183000
fraktionslos	0	0
SPD	1385250	3391500



#### **Analysiere und visualisiere selbst!**

- Suche dir einen Datensatz aus, den du analysieren willst
- Überlege dir eine Fragestellung, die du beantworten willst
- Visualisiere das Ergebnis deiner Analyse
- Stelle deine Visualisierung den anderen vor

# Präsentiere deine Visualisierung!

000

- Was war deine Fragestellung?
- Woher kommen die Daten?
- Was ist das Ergebnis?

