

# Data Science for Social Good

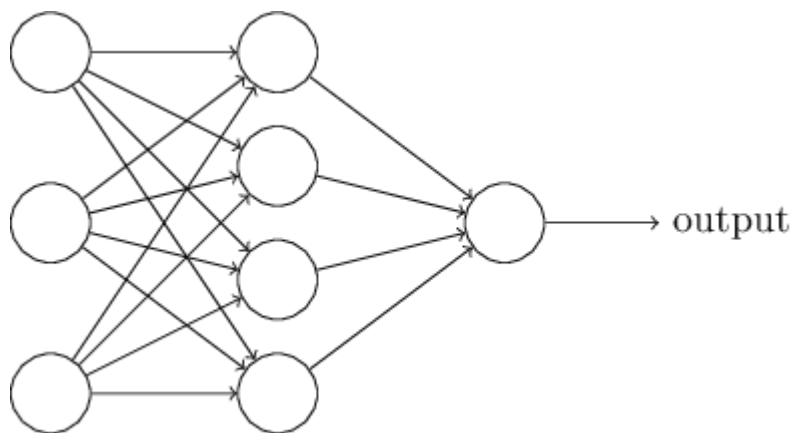
**Johannes J. Müller – CorrelAid e.V.**

**23.06.2018**

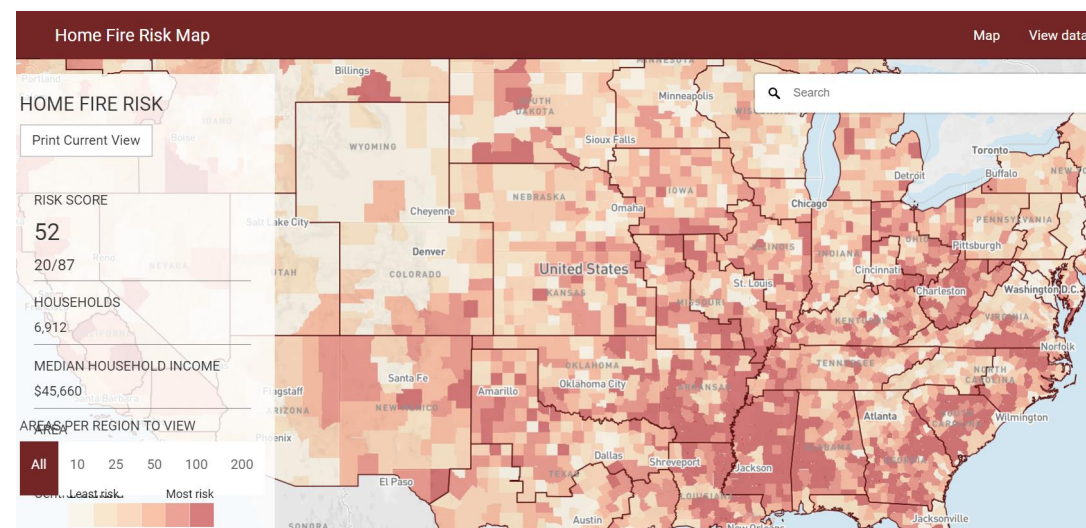
Twitter: @jj\_mllr

# Agenda

## Part 1: Machine Learning



## Part 2: Use Cases



# Teil 1: Was ist Machine Learning?





**The Cambridge  
Analytica Files**

# **'I made Steve Bannon's psychological warfare tool': meet the data war whistleblower**



CORRELAID

# Welche Werbung funktioniert bei wem besonders gut?

## Version 1:

### Neurotisches Profil



COMPLICIT

38.214 Aufrufe

763 137 TEILEN ...



Donald J. Trump for President  
Am 22.01.2018 veröffentlicht

ABONNIEREN 111 TSD.

## Version 2:

### Rationaleres Profil



Donald Trump - Crooked Hillary Ad

5.709 Aufrufe

49 4 TEILEN ...



Kapitäl  
Am 27.07.2016 veröffentlicht

ABONNIEREN 1,6 TSD.



CORRELAID

# Welche Werbung für wen?



# Was brauchen wir für die Klassifikation

1. Wie können wir messen ob jemand neurotisch ist oder nicht?
2. Woher bekommen wir einen Datensatz mit dem wir unser Modell „trainieren“ können?
3. Wie können wir klassifizieren ob jemand zur ängstlichen Gruppe gehört - nur mit Facebook-Daten?




# Persönlichkeitsprofile







# Persönlichkeitsprofile

	INACCURATE		NEUTRAL		ACCURATE
I have a kind word for everyone.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am always prepared.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I feel comfortable around people.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I often feel blue.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I believe in the importance of art.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**facebook** 




[My Personality Profile](#) [Compare to Friends](#) [More Tests](#) 0

Latest news: [New Schwartz's V](#)

**Your Friends' Personalities**

Compare Feature: Big Five Personality Questionnaire

**Most Like Me**  
*Your Personality Soulmate*



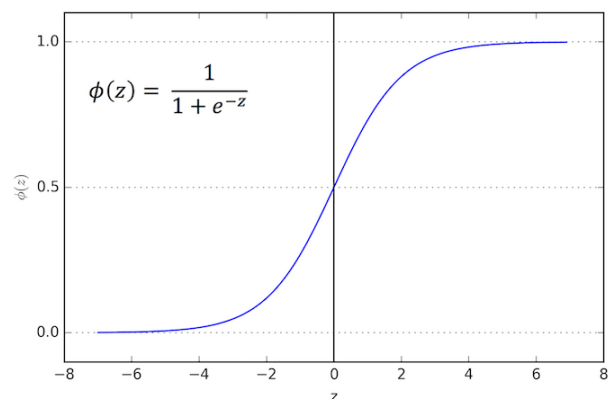
**Cristina Stoian**  
Similarity Score: **93.29%**  
(How was this calculated?)

Trait	0	50	100	% (diff.)
O	<div></div>			88% (-)
C	<div></div>			81% (+10%)
E	<div></div>			81% (-2%)
A	<div></div>			75% (-)
N	<div></div>			31% (-11%)

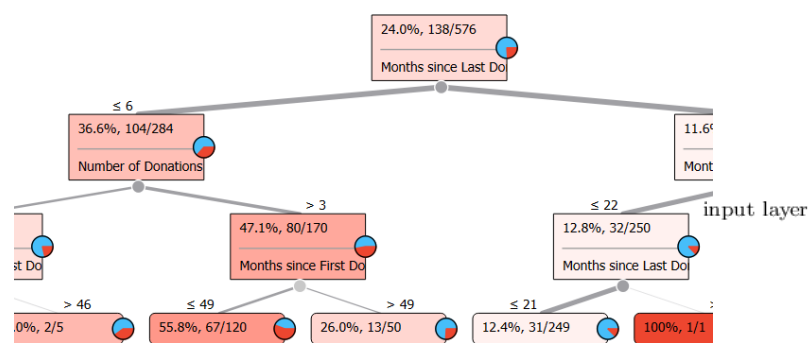
# Datensatz

Name	Neurotisch	BMW_Like	Adele_Like	Trump_Like
Alan	JA	JA	JA	JA
Betty	NEIN	NEIN	JA	NEIN
Jean	JA	NEIN	NEIN	JA
Satoshi	NEIN	JA	JA	NEIN
...	...	...	...	...

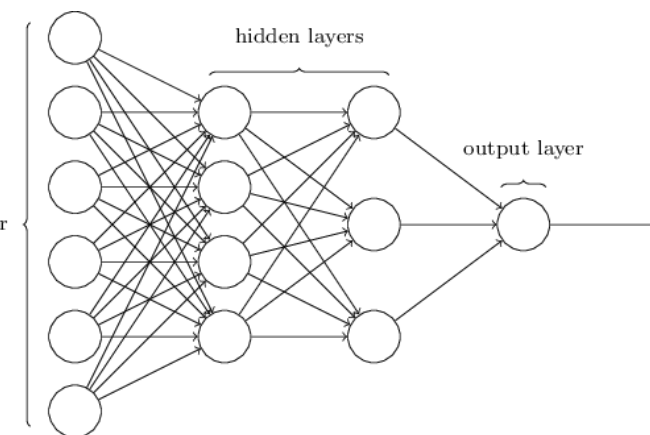
# Supervised Learning



**Logistic Regression**

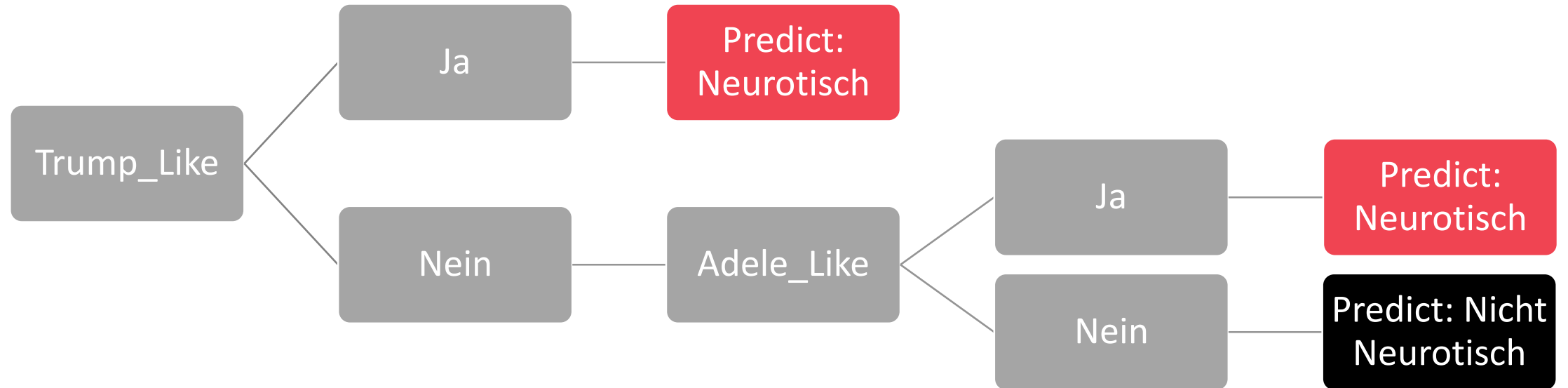


**Decision Trees**



**Neural Networks**

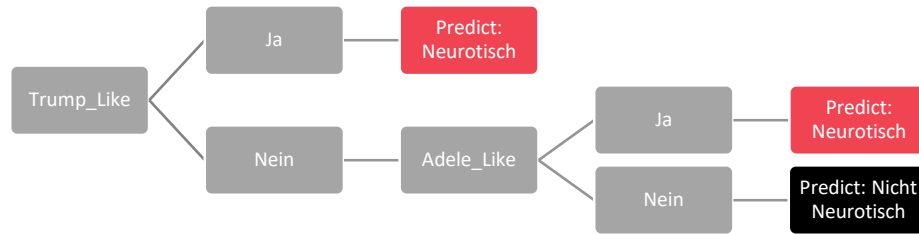
# Klassifikationsmethode: Decision Tree



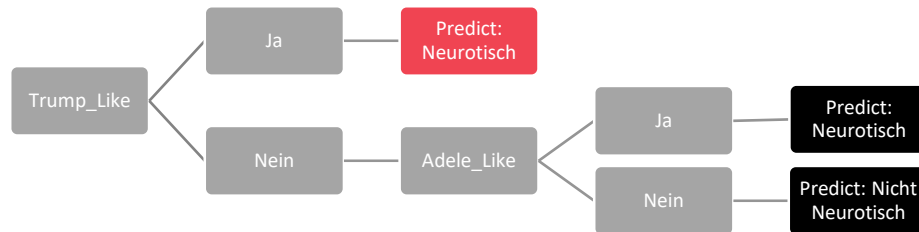
# Datensatz

ID	Neurotisch	BMW_Like	Adele_Like	Trump_Like	Vorhersage
Alan	JA	JA	JA	JA	RICHTIG
Betty	NEIN	NEIN	JA	NEIN	RICHTIG
Jean	JA	NEIN	NEIN	JA	FALSCH
Satoshi	NEIN	JA	JA	NEIN	RICHTIG
...	...	...	...	...	

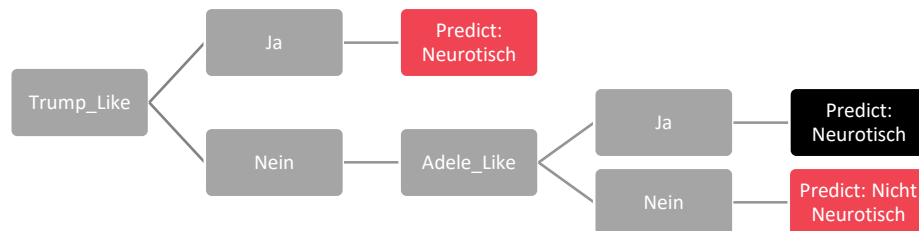
# Den Decision Tree „trainieren“



→ Richtige Vorhersagen: 85%



→ Richtige Vorhersagen: 70%



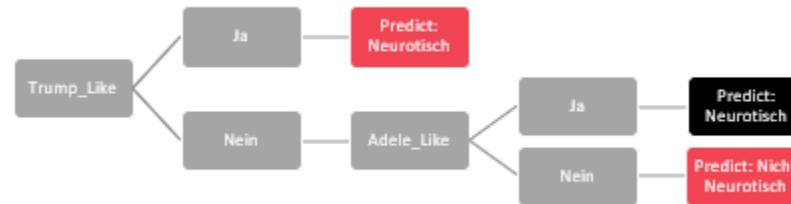
→ Richtige Vorhersagen: 90%



# Klassifikation



Klassifikation



CORRELAID

# Steps of Machine Learning

1. Daten sammeln
2. Daten labeln
3. Daten aufbereiten
4. Modell trainieren, evaluieren, anpassen, ....
5. Vorhersagen treffen







Wie können wir diese Technik für etwas Gutes nutzen?



# CASE: Blutspenden vorhersagen

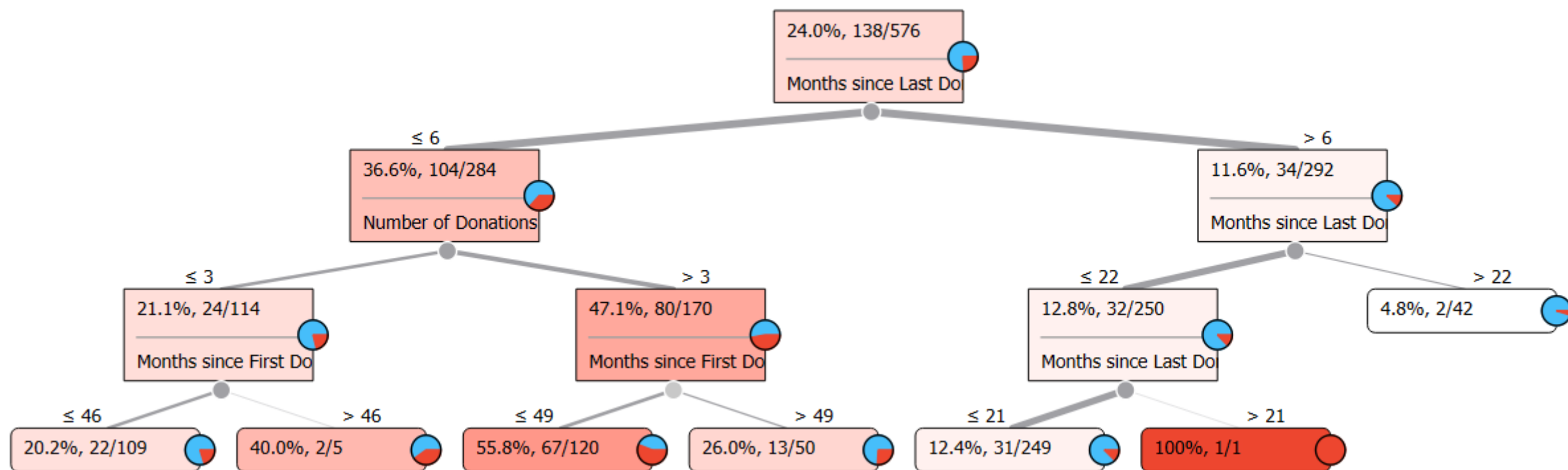


**Mobile Blood Donation Vehicle in Taiwan:** The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes to campus.

# Vorgehen

	Months since Last Donation	Number of Donations	Total Volume Donated (c.c.)	Months since First Donation
619	2	50	12500	98
664	0	13	3250	28
441	1	16	4000	35
160	2	20	5000	45
358	1	24	6000	77

- Daten über BlutspenderInnen werden anonymisiert
- Für jedeN SpenderIn wird berechnet wie wahrscheinlich er oder sie wieder spenden wird
- Daraus lässt sich ableiten ob der Bedarf durch die Blutspenden gedeckt werden kann

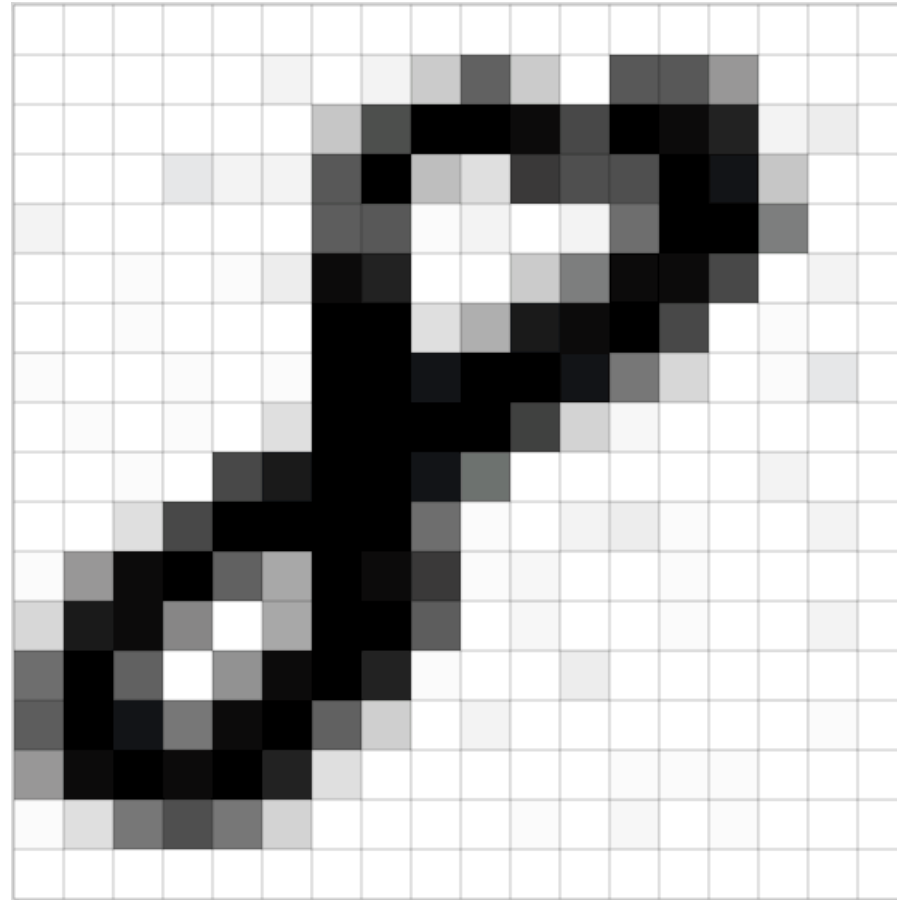


# What's up with Deep Neural Networks?



CORRELAID

# Was der Computer sieht



<https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4>





# Was der Computer sieht

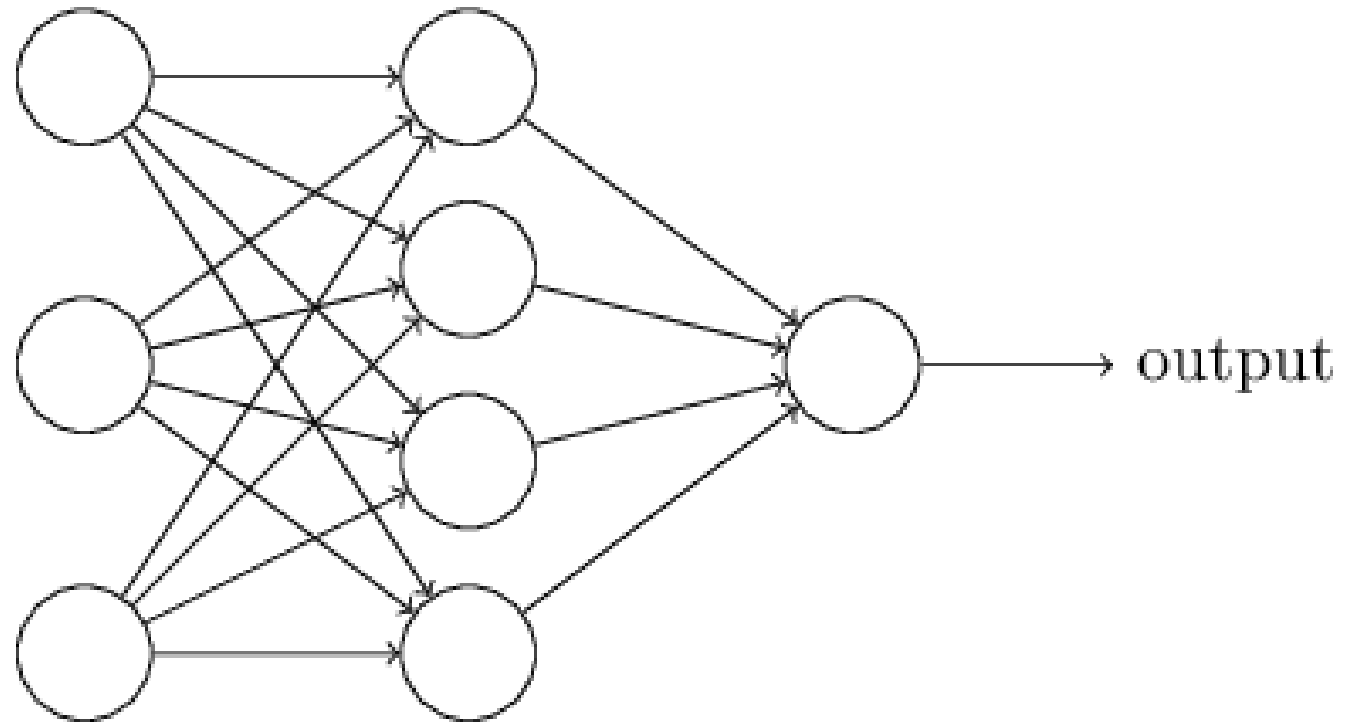


=

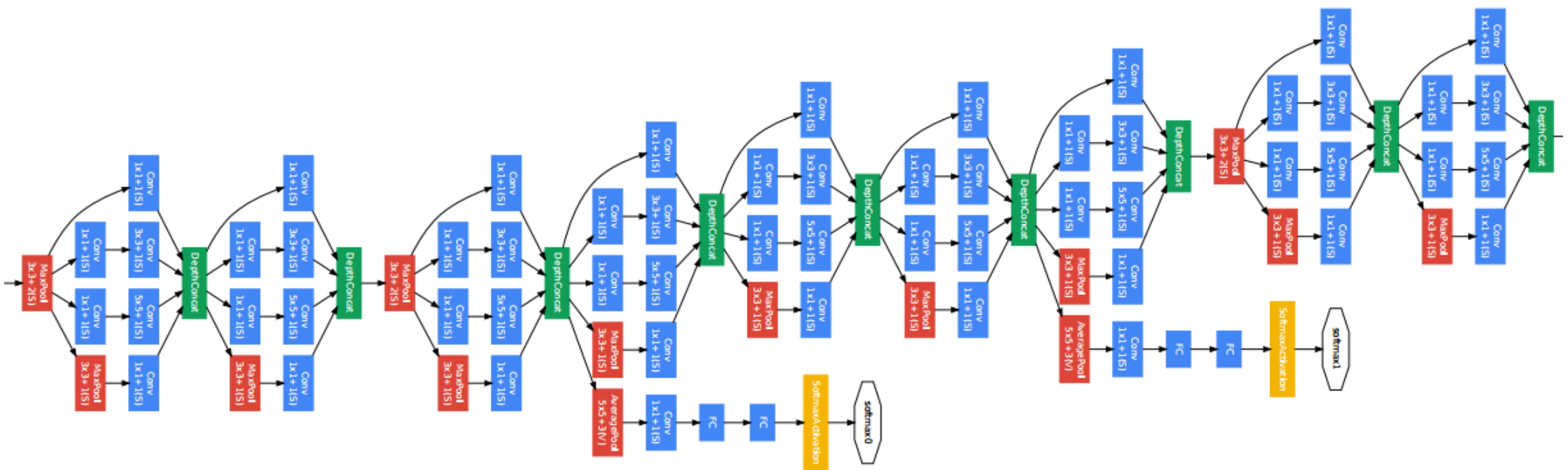
```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 12, 0, 11, 39, 137, 37, 0, 152, 147, 84, 0, 0, 0, 0,
0, 1, 0, 0, 0, 41, 160, 250, 255, 235, 162, 255, 238, 206, 11, 13, 0, 0, 0, 0, 16, 9, 9, 150, 251, 45, 21, 184, 159, 154, 2
55, 233, 40, 0, 0, 10, 0, 0, 0, 0, 0, 145, 146, 3, 10, 0, 11, 124, 253, 255, 107, 0, 0, 0, 0, 3, 0, 4, 15, 236, 216, 0, 0,
38, 109, 247, 240, 169, 0, 11, 0, 1, 0, 2, 0, 0, 0, 253, 253, 23, 62, 224, 241, 255, 164, 0, 5, 0, 0, 6, 0, 0, 4, 0, 3, 252
, 250, 228, 255, 255, 234, 112, 28, 0, 2, 17, 0, 0, 2, 1, 4, 0, 21, 255, 253, 251, 255, 172, 31, 8, 0, 1, 0, 0, 0, 0, 0, 4,
0, 163, 225, 251, 255, 229, 120, 0, 0, 0, 0, 0, 11, 0, 0, 0, 0, 21, 162, 255, 255, 254, 255, 126, 6, 0, 10, 14, 6, 0, 0, 9
, 0, 3, 79, 242, 255, 141, 66, 255, 245, 189, 7, 8, 0, 0, 5, 0, 0, 0, 0, 26, 221, 237, 98, 0, 67, 251, 255, 144, 0, 8, 0, 0
, 7, 0, 0, 11, 0, 125, 255, 141, 0, 87, 244, 255, 208, 3, 0, 0, 13, 0, 1, 0, 1, 0, 0, 145, 248, 228, 116, 235, 255, 141, 34
, 0, 11, 0, 1, 0, 0, 0, 1, 3, 0, 85, 237, 253, 246, 255, 210, 21, 1, 0, 1, 0, 0, 6, 2, 4, 0, 0, 0, 6, 23, 112, 157, 114, 32
, 0, 0, 0, 0, 0, 2, 0, 8, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```



CORRELAID







# CASE: Give Directly

**Dymo**


User: brian  
Image: [KE2013071461-iron.png](#)  
Number Left: 1452

**Instructions:**

- Identify **thatch** roofs by clicking on them.
- Identify **iron** roofs by shift+clicking on them.
- If you need to restart, press 'Clear'.
- When you're done with an image, press 'Submit'

**Labels:**

- **iron** x: 133, y: 227
- **thatch** x: 68, y: 230
- **thatch** x: 33, y: 118
- **thatch** x: 63, y: 95
- **thatch** x: 299, y: 137
- **thatch** x: 360, y: 171
- **iron** x: 374, y: 353
- **thatch** x: 139, y: 376
- **thatch** x: 180, y: 217
- **thatch** x: 203, y: 234
- **thatch** x: 269, y: 223
- **iron** x: 276, y: 150
- **thatch** x: 269, y: 44
- **thatch** x: 17, y: 326



- Projekt mit GiveDirectly in Uganda & Kenya
- Wo werden Micro-Spenden am dringendsten gebraucht?
- Dichte der Metall-Dächer als Proxy für finanzielle Lage eines Dorfes
- Machine Learning um Dächer zu klassifizieren

# ml5.js



Friendly  
Machine  
Learning for  
the Web.



CORRELAID

## **TASK 1:**

1. Überlegt euch einen Use Case für eine Bildklassifikations-App
2. Bringt euren ersten Prototypen zum laufen
3. Testet die App



# Das Potential von Datenanalyse demokratisieren



CORRELAID

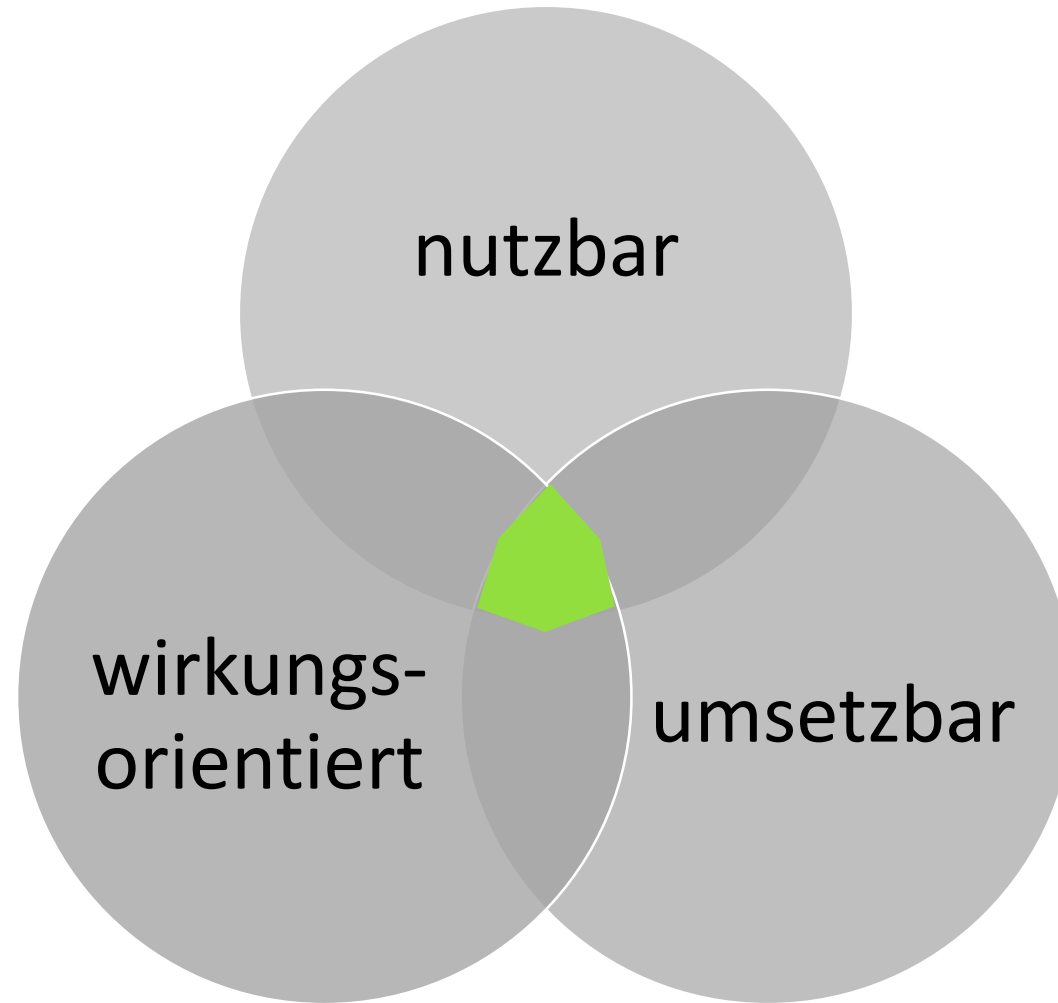


## Teil 2: Use Cases

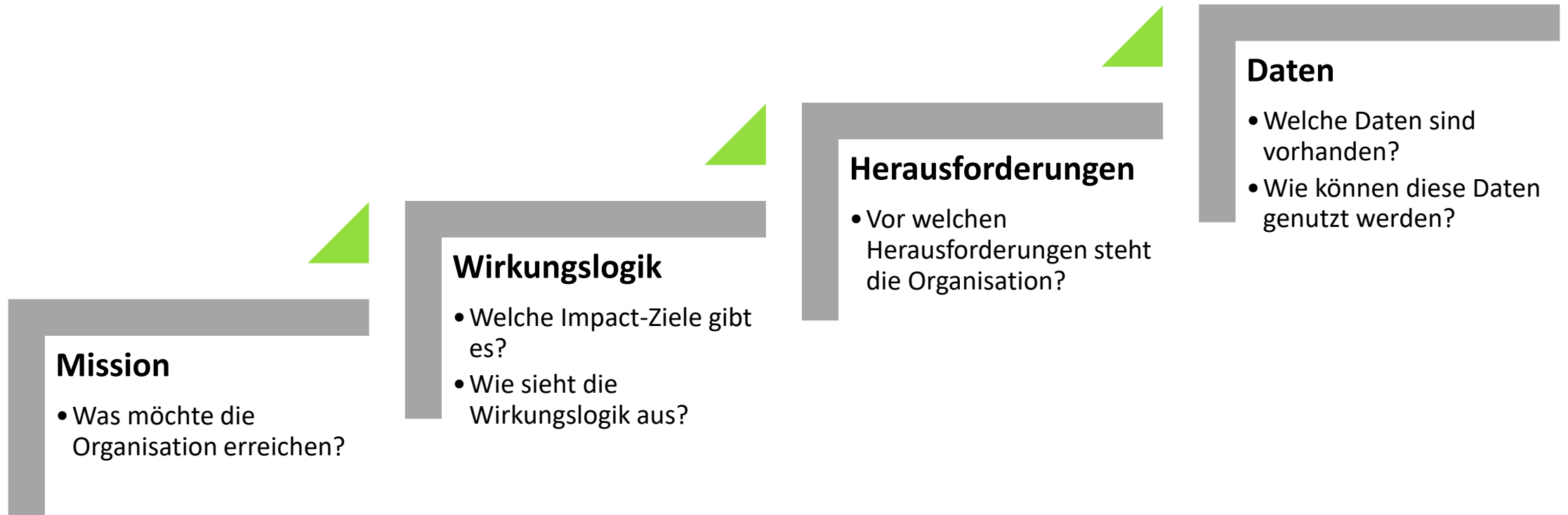
Wir entwickeln Data-for-Good Use Cases



# Ideation @CorrelAid



# Ideation @CorrelAid





## **TASK 1:**

Diskutiert 10 Minuten was Herausforderungen in gemeinnützigen Organisationen sind? Am besten am Beispiel einer Organisation in der Ihr euch engagiert.



# **Worüber wir reden wenn wir von Daten reden**



File Home Insert Page Layout Formulas Data Review View **WIP** Design Tell me what you want to do

Insert Toggle Actions Pane Reset WIP Report Data Report Completed Contracts Import Export Validate Help Add-In Activation About WIP Add-In Altova on the Web WIP Add-In

WIP Report XBRL

A4 200

	E	F	G	H	I	J	K	L	M	N	O	P	Q
1													
2		Total Contract			From Inception to June 13, 2016			At June 13, 2016			For the Period Ended June 13, 2016		
3	Estimated Revenue	Estimated Costs	Estimated Gross Profit	Earned Contract Revenue	Contract Costs	Gross Profit	Contract Billings	Estimated Costs to Complete	Percent Complete	Under (Over) Billings	Earned Contract Revenue	Contract Costs	Gross Profit (Loss)
4	29,831,262	22,771,956	7,059,306	12,113,470	9,246,924	2,866,546	11,987,630	13,525,032	41%	125,840	3,740,588	2,855,269	885,319
5	4,765,875	3,915,859	850,016	4,761,592	3,912,340	849,252	4,748,777	3,519	100%	12,815	319,663	185,925	133,738
6	3,165,949	2,635,676	530,273	3,073,180	2,558,445	514,735	3,092,332	77,231	97%	(19,152)	1,212,380	1,019,868	192,512
7	6,845,696	5,348,200	1,497,496	5,935,890	4,637,414	1,298,476	5,727,306	710,786	87%	208,584	2,985,189	2,344,782	640,407
8	3,202,917	2,139,767	1,063,150	3,197,769	2,136,328	1,061,441	3,199,414	3,439	100%	(1,645)	386,839	241,974	144,865
9	3,267,627	2,402,206	865,421	3,122,086	2,295,211	826,875	3,143,402	106,995	96%	(21,316)	254,751	101,060	153,691
10	3,513,815	2,260,925	1,252,890	2,839,759	1,827,211	1,012,548	2,573,819	433,714	81%	265,940	1,823,265	1,173,159	650,106
11	3,913,079	3,104,573	808,506	3,591,755	2,849,640	742,115	3,503,374	254,933	92%	88,381	2,651,445	2,039,028	612,417
12	12,187,491	13,500,000	(1,312,509)	2,193,165	3,505,674	(1,312,509)	2,476,537	9,994,326	26%	(283,372)	2,193,165	3,505,674	(1,312,509)
13	3,274,077	2,798,357	475,720	35,779	30,580	5,199	0	2,767,777	1%	35,779	35,779	30,580	5,199
14	3,835,139	4,296,527	(461,388)	2,578,713	3,040,101	(461,388)	2,386,461	1,256,426	71%	192,252	2,578,713	3,040,101	(461,388)
15	13,500,000	10,227,273	3,272,727	8,553,041	6,479,577	2,073,464	8,321,142	3,747,696	63%	231,899	8,553,041	6,479,577	2,073,464
16	3,849,262	3,137,190	712,072	274,615	223,814	50,801	1,741,936	2,913,376	7%	(1,467,321)	274,615	223,814	50,801
17	74,614,943	64,402,779	10,212,164	46,921,464	41,803,708	5,117,756	43,715,328	22,599,071			29,854,173	27,271,295	2,582,878
18													
19	169,767,132	142,941,288	26,825,844	99,192,278	84,546,967	14,645,311	96,617,458	58,394,321		(631,316)	56,863,606	50,512,106	6,351,500
20													
21				Costs and estimated gross profit in excess of billings on contracts in progress							4,841,687		
22				Billings in excess of costs and estimated gross profit on contracts in progress							(2,266,867)		
23													
24										2,574,820			
25													
26													
27													
28													
29													

**WIP Report Pane**

WIP Report Properties

- Data**
  - Accuracy
  - Currency
- Document Information**
  - Period Start Date
  - Period End Date
  - Fiscal Year Focus
  - Fiscal Period Focus**
- Entity Information**
  - Registrant Name
  - Current Fiscal Year End Date
  - Tax Identification Number
  - Data Universal Numbering System (DUNS)
  - State Registration Number

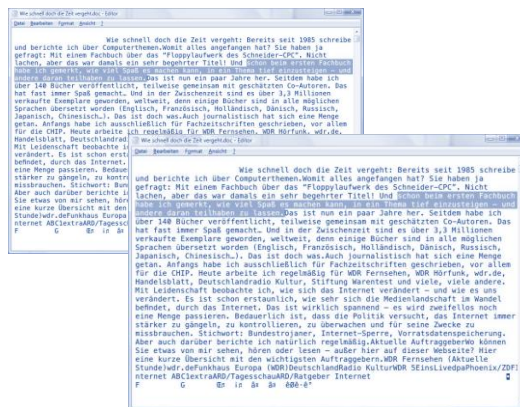
**Fiscal Period Focus**  
Specifies the fiscal period attributed with the report.

**Cell Documentation**  
Select a table data cell to display a short description.

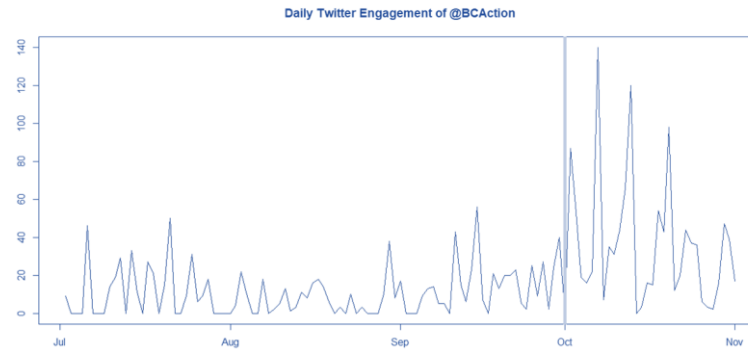


# Aufgezeichnete Daten

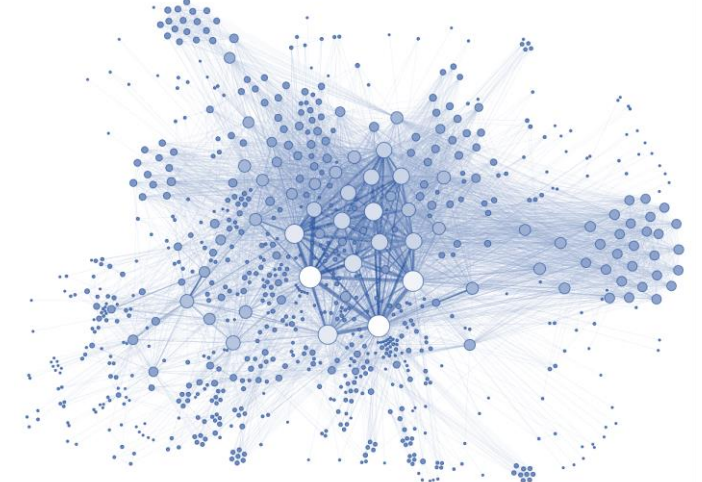
	A	B	C	D	E
1	Datum	Kunde	Artikel	Menge	Wert
2	15.09.1996	Mayer	Ball	12	300.00
3	15.09.1996	Mayer	Schläger	3	3.750.00
4	15.09.1996	Mayer	Schuhe	5	3.900.00
5	17.09.1996	Berger	Dress	1	290.00
6	17.09.1996	Berger	Ball	6	150.00
7	17.09.1996	Huber	Ball	24	600.00
8	22.09.1996	Mayer	Dress	4	1.160.00
9	24.09.1996	Huber	Schläger	6	7.500.00
10	25.09.1996	Mayer	Dress	2	580.00
11	25.09.1996	Mayer	Schläger	8	10.000.00
12	26.09.1996	Huber	Schläger	2	2.500.00
13	26.09.1996	Huber	Ball	12	300.00
14	26.09.1996	Berger	Schläger	7	8.750.00
15	26.09.1996	Berger	Ball	9	225.00
16	02.10.1996	Mayer	Schläger	8	10.000.00
17	05.10.1996	Berger	Ball	36	900.00
18	07.10.1996	Huber	Ball	24	600.00
19	07.10.1996	Huber	Dress	1	290.00
20	07.10.1996	Huber	Schläger	3	3.750.00



# Geordnete Daten

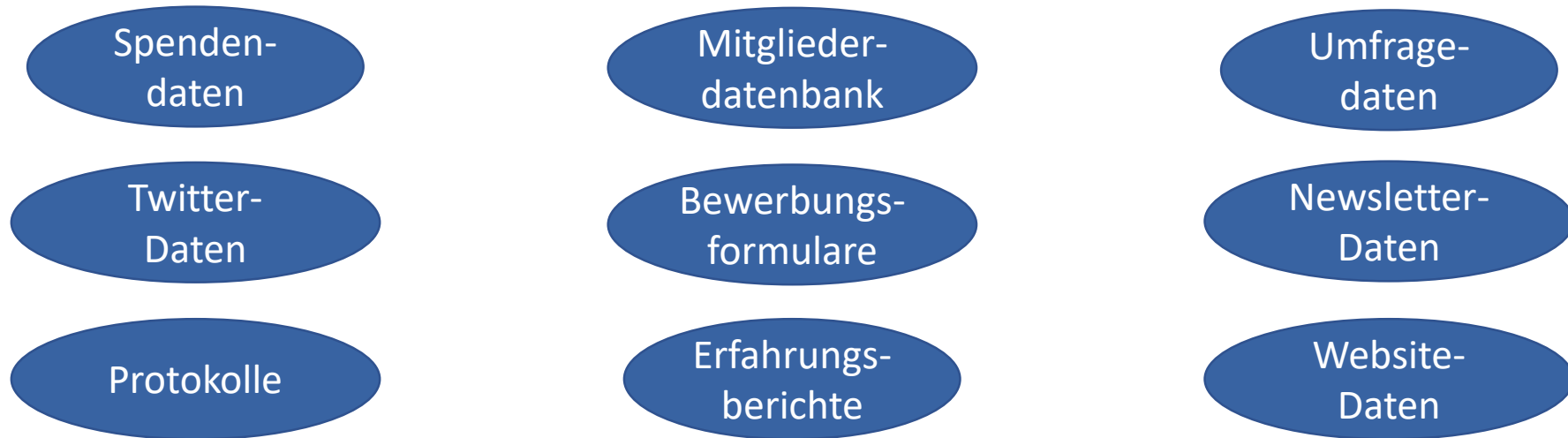


# Netzwerk-Daten



CORRELAID

# Wo kommen die Daten her?



**Strukturiert**

Spenden-  
daten

Mitglieder-  
datenbank

Umfrage-  
daten

Twitter-  
Daten

Newsletter-  
Daten

Website-  
Daten

Bewerbungs-  
formulare

**Unstrukturiert**

Protokolle

Erfahrungs-  
berichte

**„Gefundene  
Daten“**

**„bewusste  
Datenerhebung“**



CORRELAID



**Strukturiert**

**Unstrukturiert**

Datenqualität

**„Gefundene  
Daten“**

**„bewusste  
Datenerhebung“**

## **TASK 2:**

Überlegt, welche Datenquellen es in eurer Organisation geben könnte. Sind diese zugänglich? Was für Probleme könnte es mit den Daten geben?



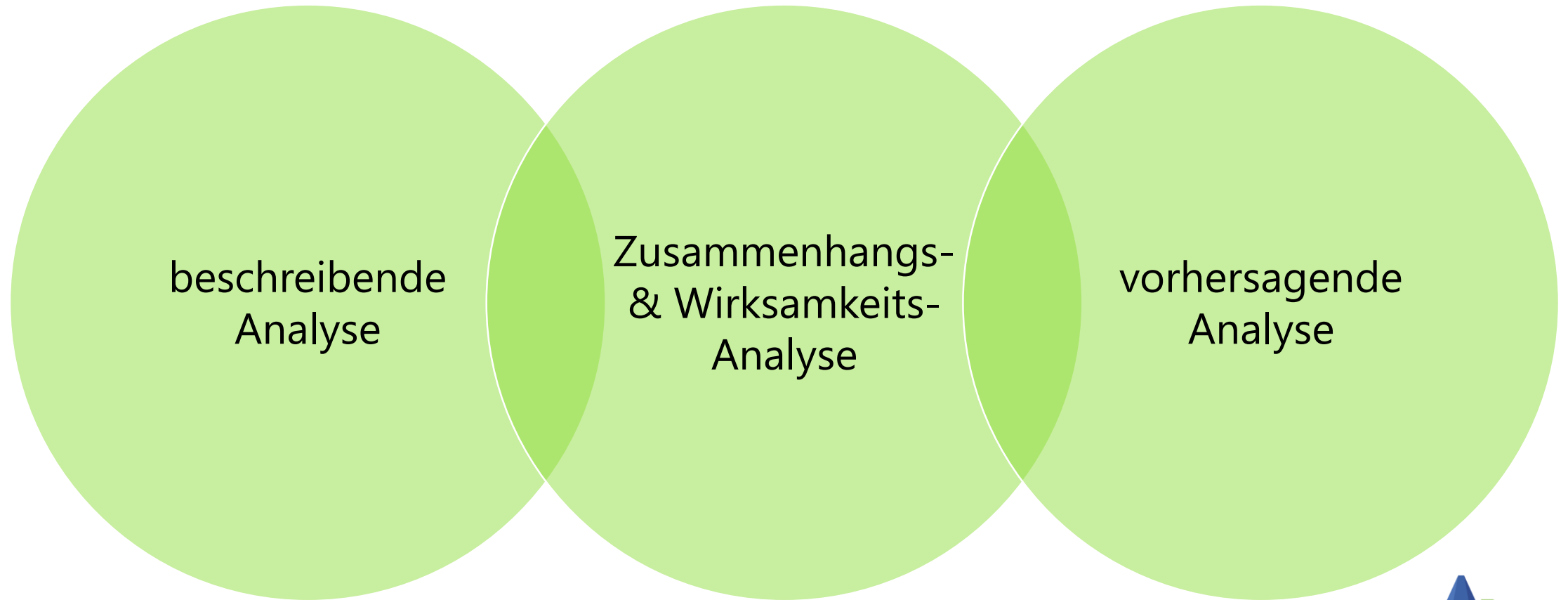


# Mögliche Ziele

- Vorhersagen
- Klassifikation/ Mustererkennung
- Vorschläge
- Wirkungsanalyse
- Informierte Entscheidungen
- ...



# Analyseverfahren



# Use Cases: Teil 1

## Klassifikation

- Muster entdecken
- Gruppen zuordnen

## Vorhersage

- Bedarf vorhersagen
- Erfolgsfaktoren identifizieren

## Exploration

- Zielgruppe besser verstehen
- Mitgliederstrukturen verstehen
- Hypothesen testen



# Use Cases: Teil 2

## Netzwerkanalyse

- Einflussreiche Personen im Netzwerk identifizieren
- Beziehungen verstehen
- Informationsflüsse nachvollziehen

## Evaluation

- Wirksamkeit messen
- Marketingmaßnahmen evaluieren

## Visualisierung

- Storytelling
- Tracking von KPIs



# Wie sozial inklusiv ist meine Organisation?

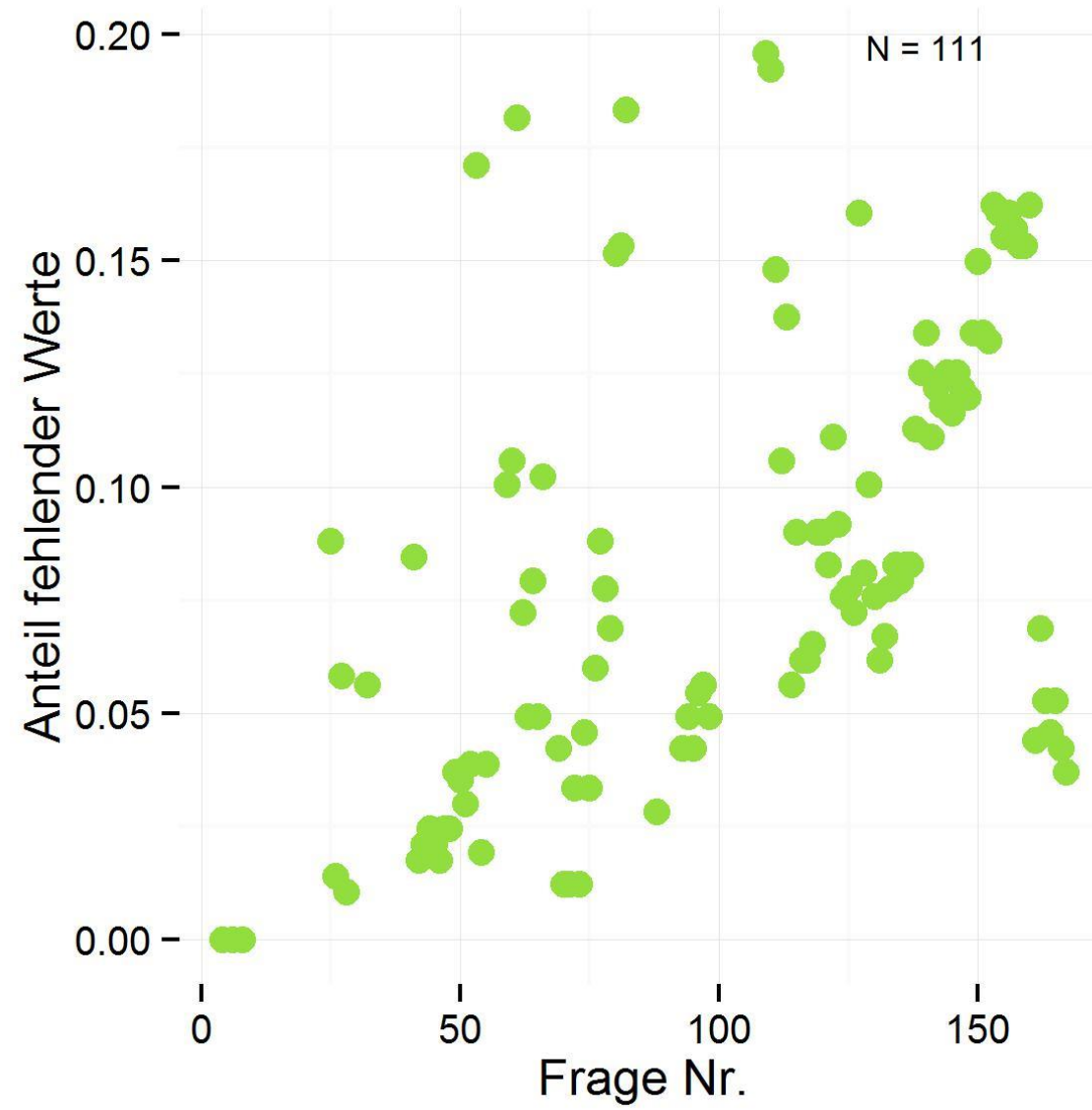
Beschreibende Analyse



# CASE: PFADFINDER- UND PFADFINDERINNENBUND NORD

- Jugendorganisation mit 1000 Mitgliedern aus Norddeutschland, die das Ziel hat die Entwicklung junger Menschen zu fördern.
- Befragung von 567 PfadfinderInnen
- Fragen:
  - Wie sind Minderheiten repräsentiert im Vergleich zur allg. Bevölkerung?
  - Wer übernimmt Verantwortung?
  - Wie wurden die Mitglieder geworben?
- Impact: Tätigkeiten besser auf Mitglieder abstimmen und potentiell interessierte Jugendliche gezielter ansprechen





# Wo ist die Gefahr am größten? Wo werden unsere Ressourcen am dringendsten gebraucht?

Vorhersagende Analyse



CORRELAID



# CASE: Brandvorhersage in NYC



**American  
Red Cross**

**DataKind**

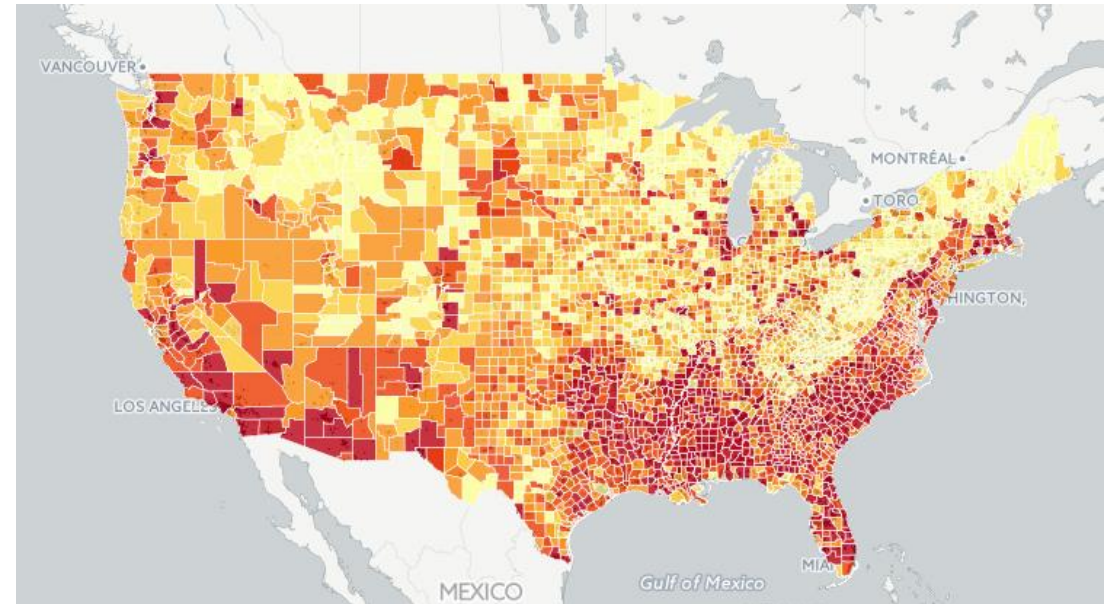
“The **American Red Cross Home Fire Preparedness Campaign** aims to reduce the number of home-fire deaths and injuries by 25 percent over the next five years, working with community partners and stakeholders to install smoke alarms and provide fire- and disaster-safety education in communities at risk for home fires.”

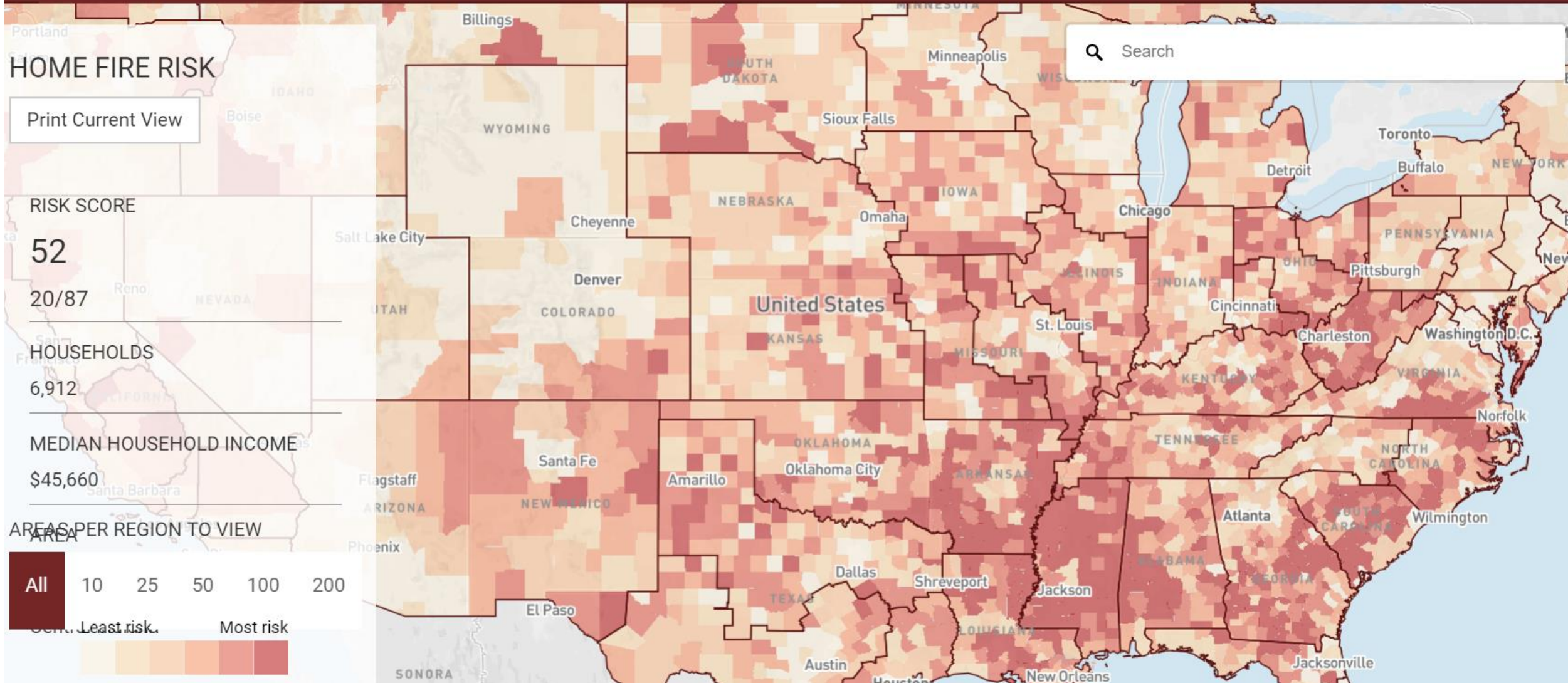


CORRELAID

# Vorgehen

- Daten über häusliche Katastrophen von American Red Cross werden anonymisiert
- Übersicht erstellen von Mustern in den USA
- Erstellen eines Vorhersagemodells um die risikoreichsten Regionen in den USA zu identifizieren
- Impact: American Red Cross kann entscheiden wo sie ihre nächsten Kampagnen starten sollen





# Hat meine Marketing- oder Awareness Kampagne funktioniert?

Wirksamkeitsanalyse

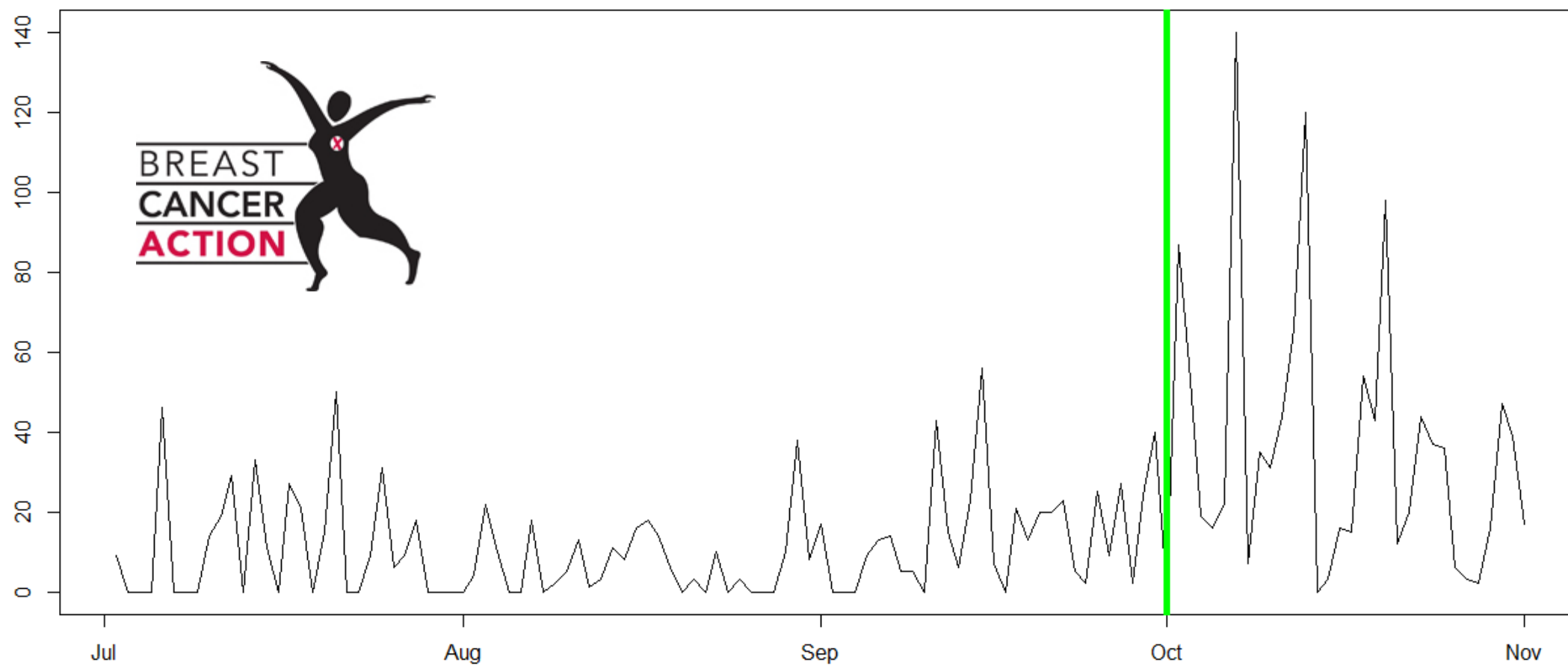




# Effekt einer Awareness-Kampagne



## Daily Twitter Engagement of @BCAction

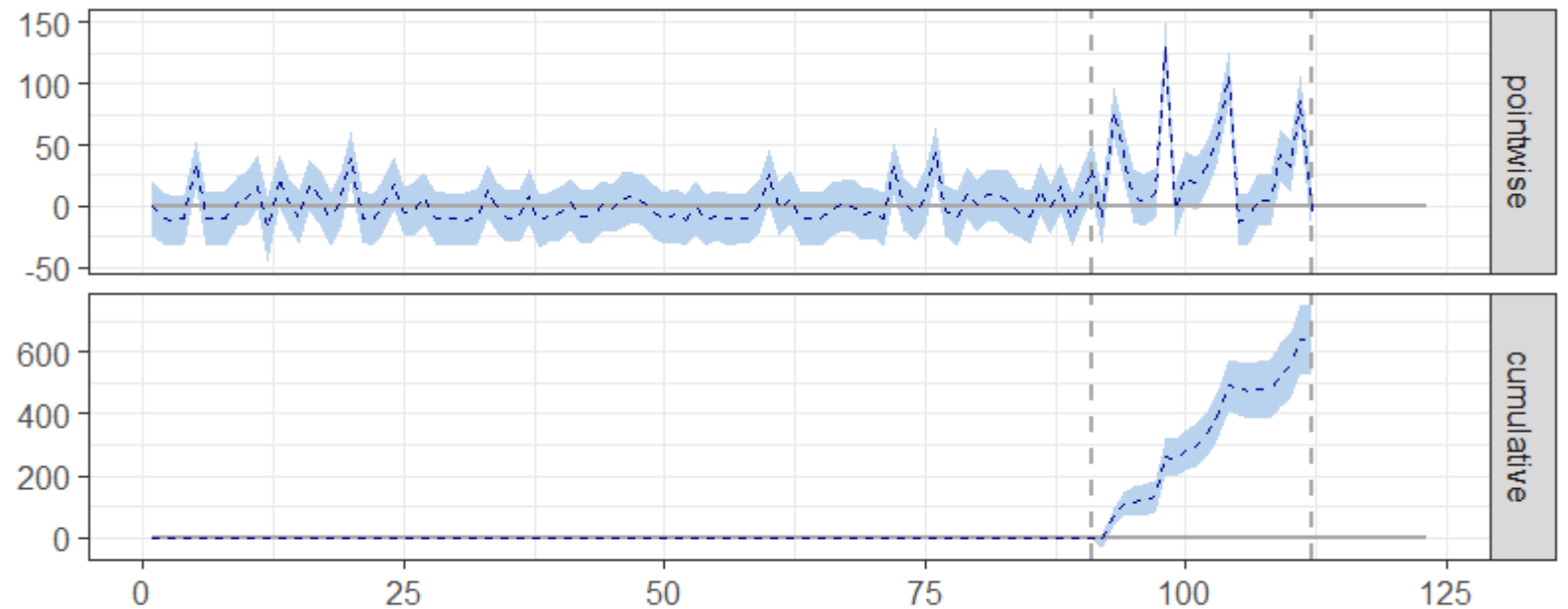


CORRELAID

# Was ist der Effekt in den ersten 20 Tagen?

**Durchschnittlicher  
Täglicher Effekt**  
+ 260 %  
+ 30 Likes/Retweets

**Gesamter Effekt**  
+ 636  
Likes/Retweets



CORRELAID

# Data-driven social media marketing

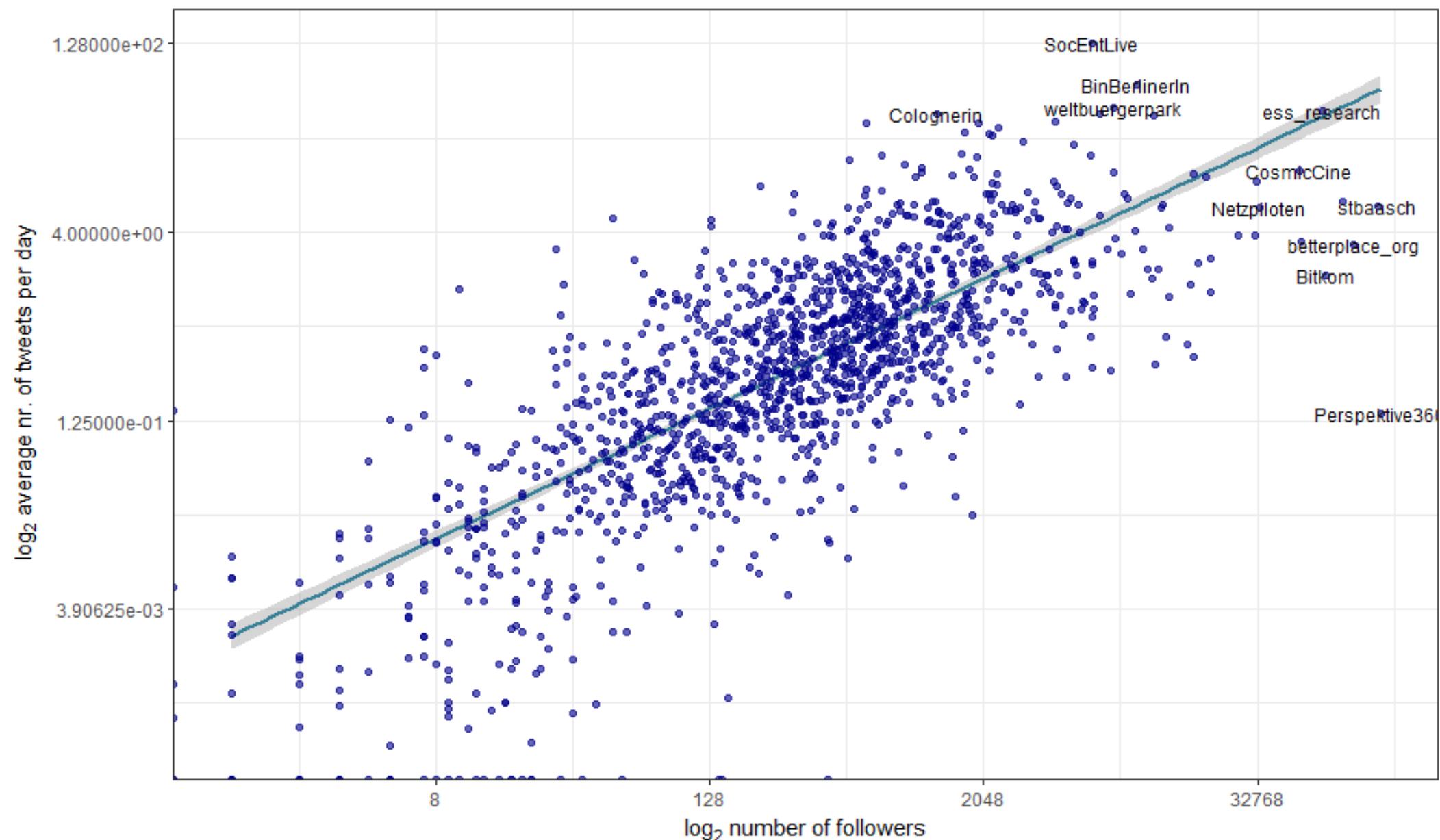
Netzwerkanalyse



CORRELAID

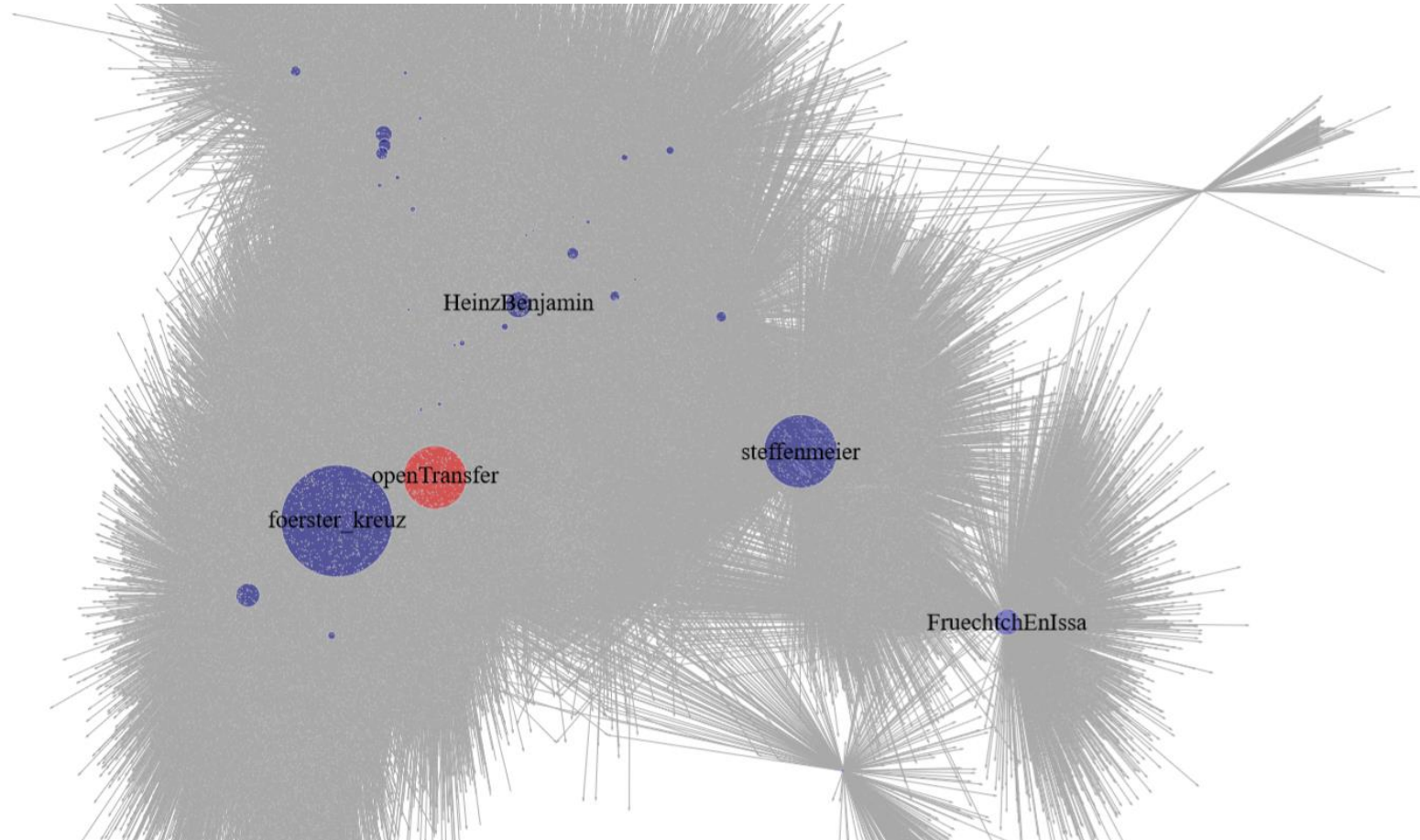


# OpenTransfer's most influential followers



2<sup>nd</sup> degree followers and tweet rate of @openTransfer Twitter followers

# Follower Netzwerke



CORRELAID

# Welche Themen interessiert das Netzwerk?



### **TASK 3:**

Überlegt euch eine oder mehrere Projekte bei denen Datenanalyse diese Organisation helfen könnte.



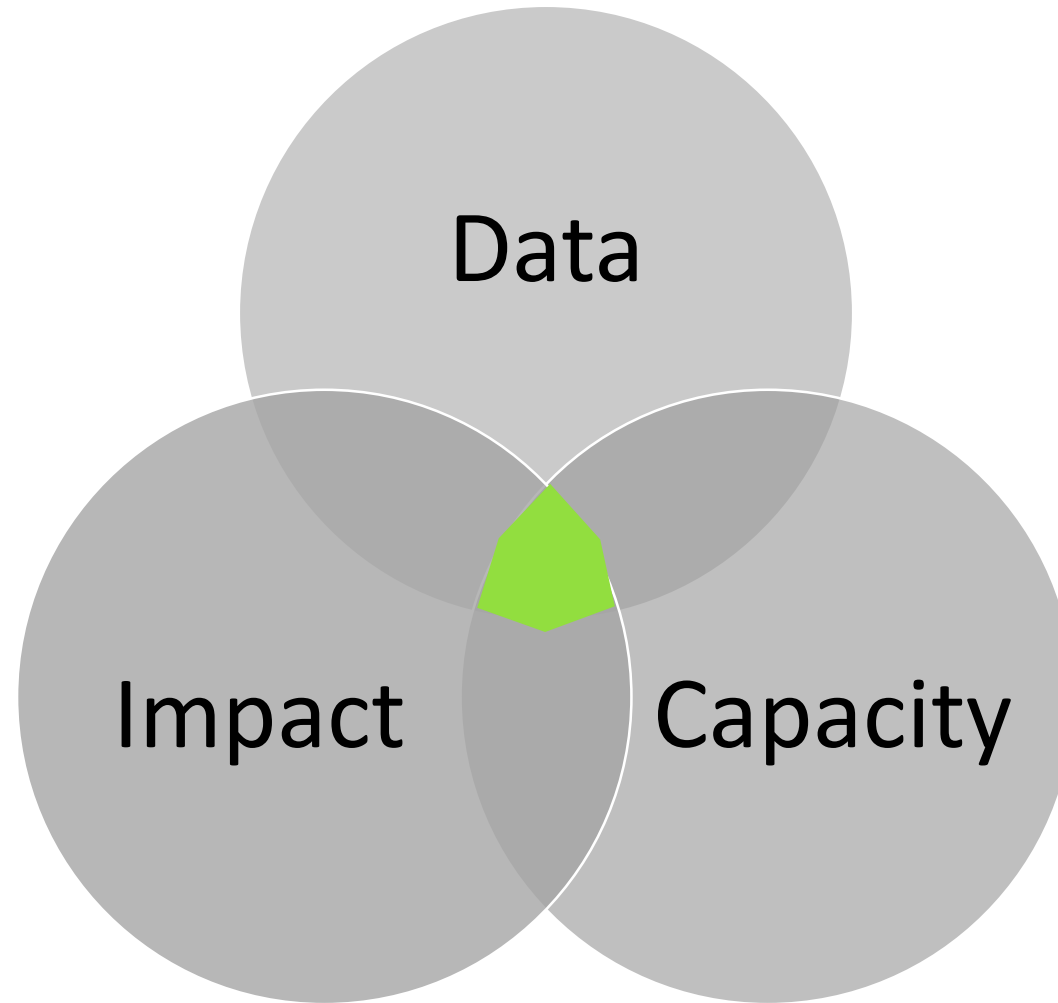


# Das Potential von Datenanalyse demokratisieren

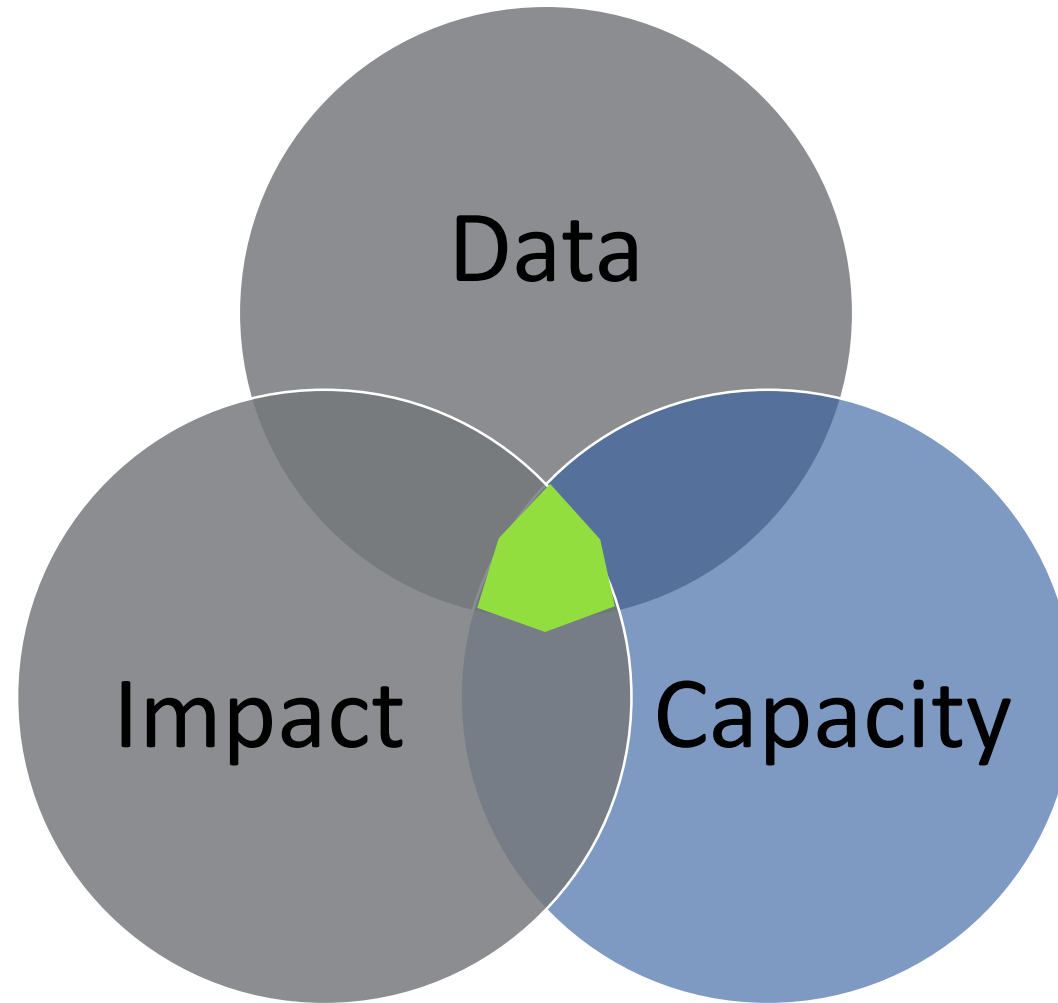


CORRELAID

# Data for Good



# Data for Good



# CorrelAid

Wir sind ein deutschlandweites Netzwerk aus 650 jungen Datenanalytistinnen und -analysten, das mit einem inklusiven, vernetzten und innovativen Ansatz das Potential von Datenanalyse demokratisiert.



CORRELAID



# Unsere drei Säulen

- 1) Wir nehmen in Deutschland eine Vorreiterrolle bei der pro-bono Datenanalyse-Beratung von Organisationen mit sozialem Auftrag ein.
- 2) Wir vernetzen junge und engagierte Data Scientists und bieten ihnen eine Plattform, ihre Kenntnisse anzuwenden und zu erweitern.
- 3) Zudem stoßen wir einen Dialog über den Wert und Nutzen von Daten und Datenanalyse für die Zivilgesellschaft an.



# Kontakt

**Johannes Müller**

*Vorstandsvorsitzender CorrelAid e.V.*

[johannes.m@correlaid.org](mailto:johannes.m@correlaid.org)  
@jj\_mllr

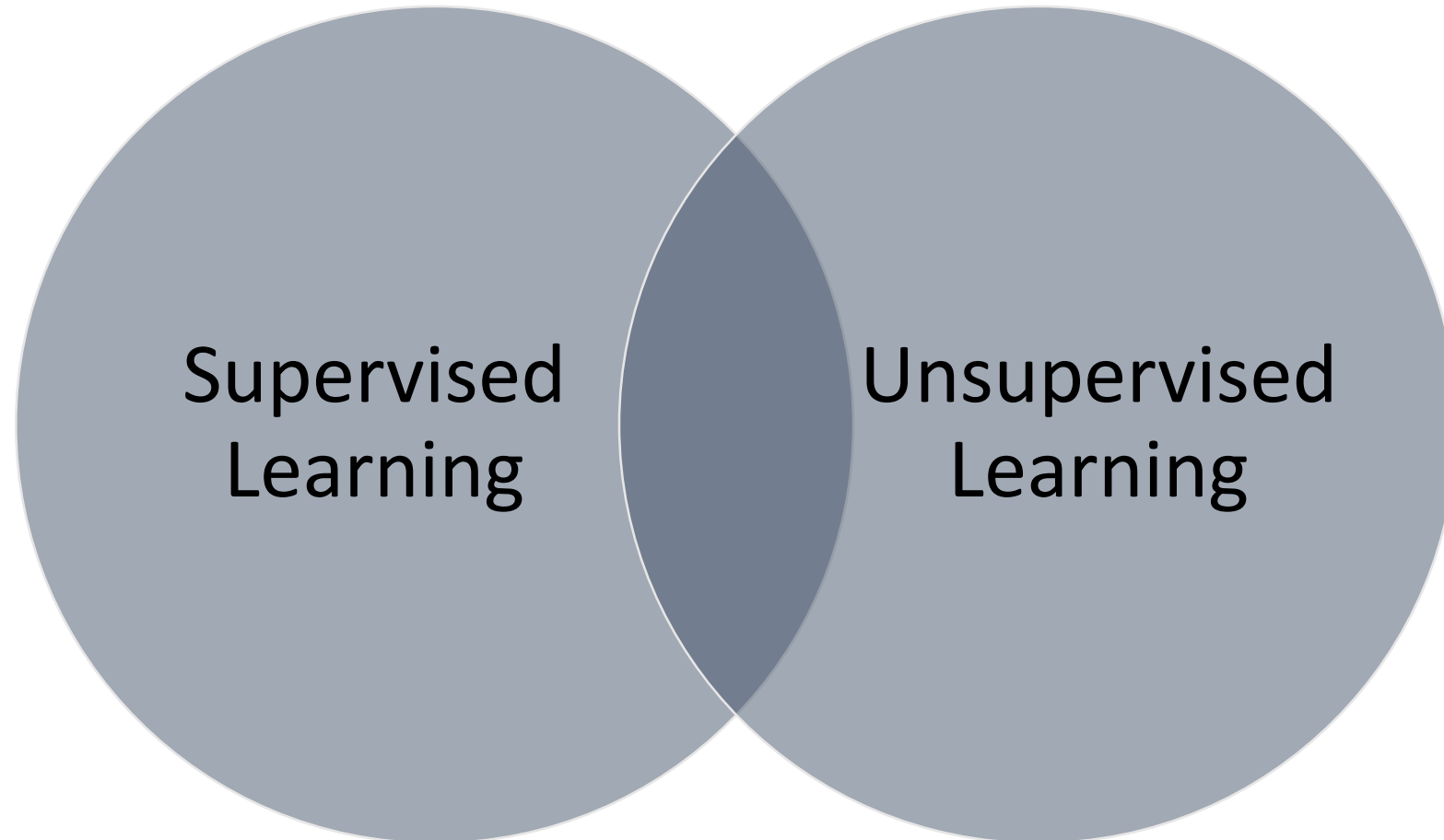
[www.correlaid.org](http://www.correlaid.org)

facebook.com/WeAreCorrelAid  
@CorrelAid



CORRELAID

# Machine Learning



# Schreibt mir gerne...

[johannes.m@correlaid.org](mailto:johannes.m@correlaid.org)

Twitter: @jj\_mlr

[facebook.com/WeAreCorrelAid](https://facebook.com/WeAreCorrelAid)



CORRELAID