Süddeutsche Zeitung

Who are we?

Two data journalists working at Süddeutsche Zeitung in the data and digital investigations team.

- Martina Schories, Twitter: @MSchories
- Katharina Brunner, Twitter: @cutterkom
- other team members: Vanessa, Hannes, Christian, Benedict, Moritz, Felix

What is Süddeutsche Zeitung (SZ)?

- One of the most important quality newspapers in Germany
- Focus on politics, business, sports
- International fame because of Panama Papers and Paradise Papers

What is data journalism #ddj?

In contrast to traditional journalism which is based on qualitative reporting, data journalism is based on quantitative reporting.

Our protagonists are means or filtered columns or standard deviations. But at first there is always the question: Which data could answer your question? How can we get it? How to extract the information?

Be aware: Data only tells that something IS, but never WHY.

R at SZ

We use R in different parts of data journalism

- How to analyse crowd-sourced answers
- Show more uncertainty to be more precise
- Using text mining in political reporting
- Automated text generation for election coverage
- When all you got is a hammer, everything looks like nails
- Getting an idea of the blackbox Facebook

How to analyse crowd-sourced answers

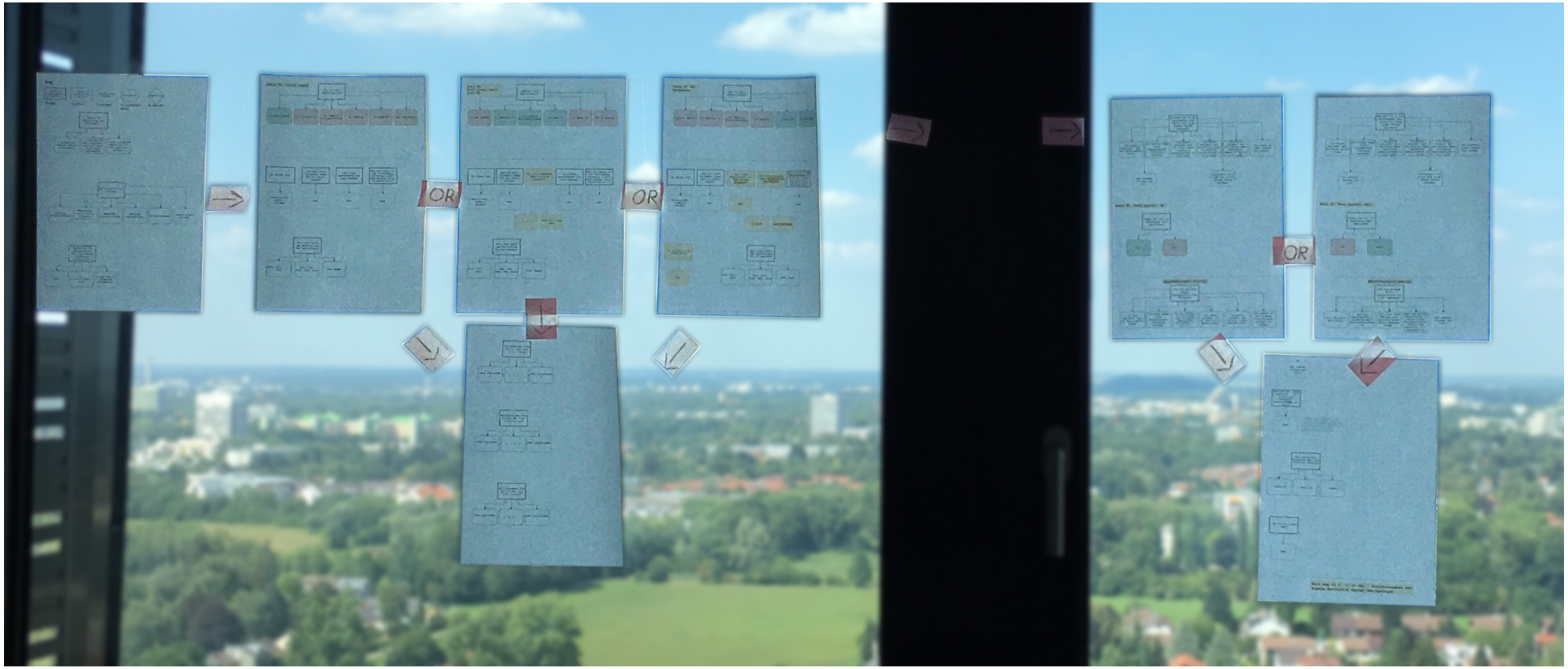
"Der Mietmarkt ist kaputt" (German, English)

The housing market is one of the biggest social problems these days - especially in Munich.

There are many data sources on prices and rents out there, but one thing is missing: **What is the rent burden?**

$$rent_burden = rent/income$$

Survey with 32 questions:



Lessons from working with other departments: `.Rmd` files are the way to connect data analysis with first drafts of the storyline

Lessons from asking people: provide a free textfield for getting useful information you never thought of

Analyse München

Show more uncertainty to be more precise

"Wie wir über Umfragen berichten" (German, English)

Traditionally, media outlets are reporting about a new poll in the following style:

```
If an election would be held today, party x would get y
percent of the votes. This is a decline of z percent
compared to the previous week.
```

Covering polls like this is oversimplifying and even dangerous. Polls have real impact on decisions of politicians and voters, e.g. due to feedback loops.

Common misconceptions about polls

- Polls **aren't predictions**.
- Polls are **always biased**.
- Polls methodology **is unknown**.
- Polls are **never exact**.

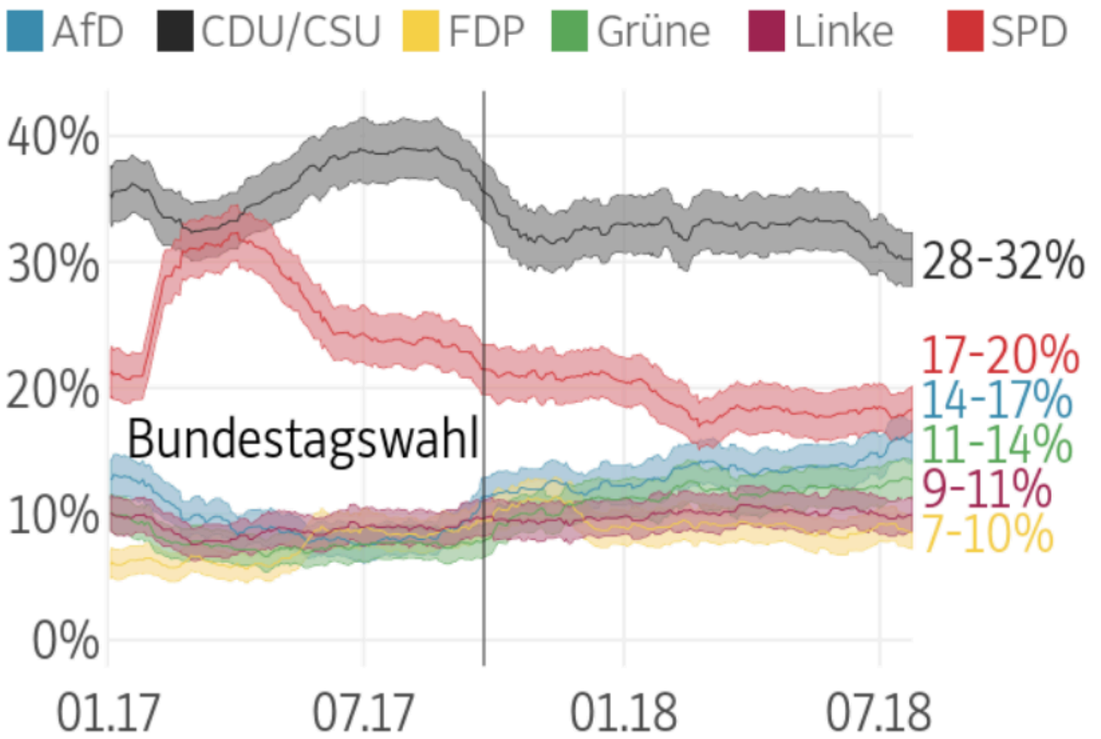
Pollsters want to mirror the views of a whole electorate by asking 1000 to 2000 people. Of course, there is uncertainty!

Our approach: Making the visualization more complex, but more precise by **showing the uncertainty**.

1. Scrape data from wahlrecht.de
2. Calculate standard deviations and the weighted mean of seven pollsters
3. Generate the plots with `ggplot` and an `sz-theme`
4. Automated this process: all code is written in `R` and deployed on Jenkins

Wen würden Sie wählen, wenn am Sonntag Bundestagswahl wäre?

Umfrageergebnisse liefern keine exakten Werte, sondern geben eine Spanne an, innerhalb der die Ergebnisse für eine Partei wahrscheinlich liegen. Die Institute setzen verschiedene Methoden ein, die zu unterschiedlichen Ergebnissen führen. Die Linie zeigt den gewichteten Mittelwert der jeweils neuesten Umfrage von sieben Instituten.

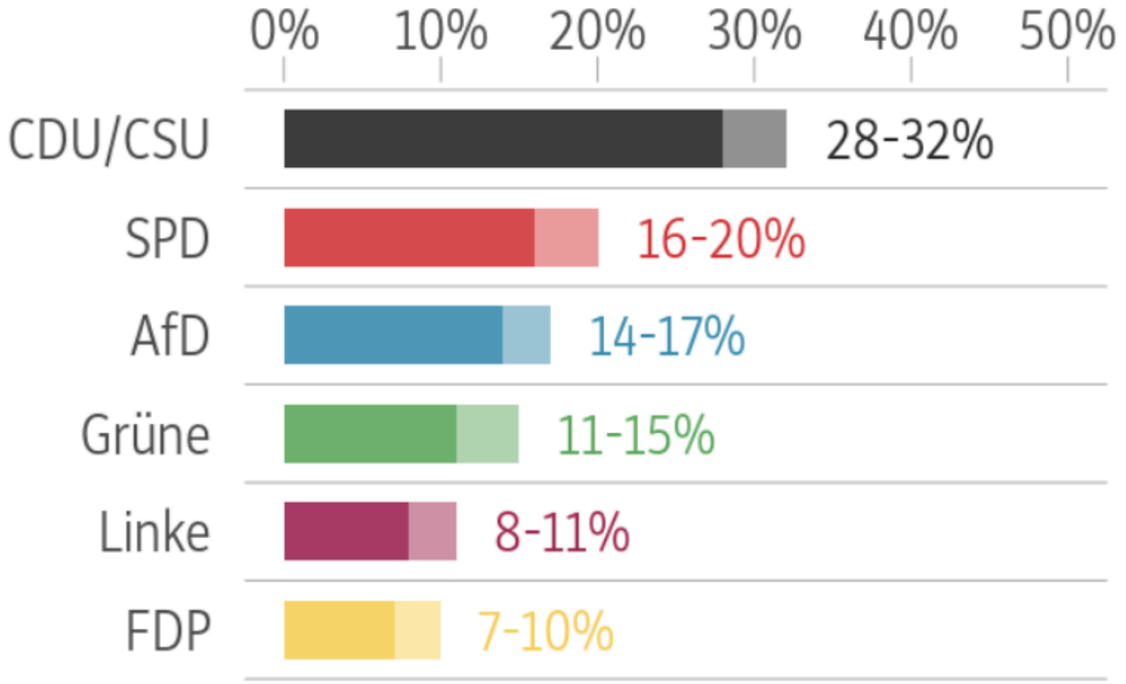


Katharina Brunner, Martina Schories

Stand: 25.07.2018

Wen würden Sie wählen, wenn am Sonntag Bundestagswahl wäre?

Umfrageergebnisse liefern keine exakten Werte, sondern geben eine Spanne an, innerhalb der die Ergebnisse für eine Partei wahrscheinlich liegen. Die Institute setzen verschiedene Methoden ein, die zu unterschiedlichen Ergebnissen führen. Die Balken zeigen den gewichteten Mittelwert der jeweils neuesten Umfrage von sieben Instituten.



Stand: 30.07.2018

sztheme:

```
sztheme_lines <- theme(
  strip.background = element_blank(),
  strip.text.y = element_blank(),
  strip.text.x = element_blank(),
  axis.text = element_text(family = "SZoSansCond-Light", size = 18),
  axis.text.x = element_text(margin = margin(0.1,0,0,0,"in")),
  axis.line.y = element_blank(),
  axis.ticks = element_blank(),
  axis.ticks.length = unit(0,"lines"),
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.background = element_blank(),
  panel.border = element_blank(),
  panel.grid.major.y = element_line(colour = "#eeeeee", size = 0.3),
  panel.grid.major.x = element_line(colour = "#eeeeee", size = 0.3),
  panel.spacing = unit(c(0,0,0,0), "lines"),
  plot.background = element_blank(),
  plot.margin = unit(c(0, 0.8, 0.1, 0), "in"),
  legend.margin = margin(0.1, 0, 0.1, -3.5, "in"),
  legend.background = element_blank(),
  legend.title = element_blank(),
  legend.position = "top",
  legend.direction = "vertical",
  legend.key = element_blank(),
  legend.key.size = unit(0.15, "in"),
  legend.text = element_text(family = "SZoSans-Light", size = 16, colour = "#666666"),
  text = element_text(size = 18, family = "SZoSansCond-Light", colour = "#666666")
)
```

Not everything is fun with R

- Labels like `28-32%` are not part of standard labeling.
- When two parties have the same values, only the last plotted value is visible. To solve this problem, we pasted non-printing characters to the label, e.g.

```
hidden_chars <- c("\U200C","\u200D","\u200E","\u200F",
"\U200C","\u200D")
```

- Why on earth are there different units in `ggplot` font sizes?

More about our approach:

- Code on Github
- Methodology: Wie wir über Umfragen berichten - in English: How we report on polls
- Umfragen sind nur ein Schnappschuss der Gegenwart - in English: Surveys are just a snapshot of the present
- Süddeutsche Zeitung is improving the way media reports on political polls: @GENinnovate about the project

Using text mining in political reporting

"Das gespaltene Parlament" (German, English)

Research question: How does a right-wing, populist party change the atmosphere and the debates in the German parliament?

Inofficial research question: Who applauds the AfD (by clapping)?

Answers can be found in the official protocols of the Bundestag.

The protocols are published as XML files. Hooray! (At first sight)

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="dbtplenarprotokoll.css" type="text/css" charset=
"UTF-8"?>
<!DOCTYPE dbtplenarprotokoll SYSTEM "dbtplenarprotokoll.dtd">
<dbtplenarprotokoll herstellung="Satz: Satzweiss.com Print, Web,
Software GmbH, Mainzer Straße 116, 66121 Saarbrücken,
www.satzweiss.com, Druck: Printsystem GmbH, Schafwäsche 1-3, 71296
Heimsheim, www.printsystem.de" sitzung-datum="21.03.2018"
sitzung-ende-uhrzeit="22:22" sitzung-naechste-datum="22.03.2018"
sitzung-nr="22" sitzung-start-uhrzeit="11:31" start-seitennr="1795"
vertrieb="Bundesanzeiger Verlagsgesellschaft mbH, Postfach 1 0 05 34,
50445 Köln, Telefon (02 21) 97 66 83 40, Fax (02 21) 97 66 83 44,
www.betrifft-gesetze.de" wahlperiode="19">
  <vorspann>
    <kopfdaten>
      <plenarprotokoll-nummer>Plenarprotokoll <wahlperiode>19</
      wahlperiode>/<sitzungsnr>22</sitzungsnr></plenarprotokoll-nummer>
      <herausgeber>Deutscher Bundestag</herausgeber>
      <berichtart>Stenografischer Bericht</berichtart>
      <sitzungstitel><sitzungsnr>22</sitzungsnr>. Sitzung</
      sitzungstitel>
      <veranstaltungsdaten><ort>Berlin</ort>, <datum date="21.03.2018">
      Mittwoch, den 21. März 2018</datum></veranstaltungsdaten>
    </kopfdaten>
```

Unfortunately, it's a bit more complicated. We want to analyze the <kommentar> tags, the meta data of a single speech: clapping, interruptions, laughing ...

```
<p klasse="0">Die AfD fordert, das rückgängig zu machen. Staatsvolk ist Wahlvolk,
und Wahlvolk kann Staat, Grundgesetz und Demokratie aus den Angeln heben. Ein zur
Regel entarteter Doppelpass untergräbt Staat und Demokratie.</p>
<kommentar>(Claudia Roth [Augsburg] [BÜNDNIS 90/DIE GRÜNEN]: Jetzt reden wir hier
schon von „Entartung“! – Dr. Anton Hofreiter [BÜNDNIS 90/DIE GRÜNEN]:
Entschuldigen Sie mal! Schämen Sie sich für Ihre Sprache! Überhaupt kein Anstand!)
</kommentar>
<p klasse="0">Das wollen wir nicht.</p>
<kommentar>(Beifall bei der AfD – Dr. Anton Hofreiter [BÜNDNIS 90/DIE GRÜNEN]:
Schämen Sie sich! Hier wird NS-Sprache benutzt, und Sie klatschen für das! Sie
haben ja wohl überhaupt keinen Anstand! – Gegenruf des Abg. Dr. Alexander Gauland
[AfD]: Halten Sie den Mund! Wir sind hier im Parlament und nicht auf einer
Maikundgebung!)</kommentar>
```

The challenge: Extract all `<kommentar>` tags and transform these unstructured texts in structured data aka a `dataframe`.

The procedure:

- Cut all strings into the smallest possible parts: `str_extract` <3
- Find rules for text mining. We wrote **many** regexes.
- Extract the party, action, ... to create a longform dataframe

sitzung_id	type	party	speaker_name	speaker_party	zurufer_name	zuruf_text
12	beifall	afd	Gottfried Curio	afd	*NA*	*NA*
12	zuruf	gruene	Gottfried Curio	afd	Dr. Anton Hofreiter	Schämen Sie sich! Hier wird NS-Sprache benutzt, und …
12	zuruf	afd	Gottfried Curio	afd	Dr. Alexander Gauland	Halten Sie den Mund! Wir sind hier im Parlament und …

We did this for every protocol which resulted in a dataframe with

- 24 396 rows for 24 meetings
- more than 1500 speakers.

Which stories can we find inside the data?

How to get from a dataframe to a story?

We interview our dataset.

A lot of `filter()`, `select()`, `group_by`, `mutate()`, `ggplot()`.

And serveral `.Rmd` files later... The final questions and answers(!) our story is based on, you can find on Github.

More on that story

Main story: Das gespaltene Parlament in German,

Methodology

Github Repository

Automated text generation for election coverage

"Wie hat Ihr Wahlkreis gewählt" (German, English)

Context: After an election there are two tasks for journalism:

1. **report** the results instantly (which is easy)

2. **interpret** the results (which is not that easy)

We wanted to do these two things in an automated way and as fast as possible for all 299 election districts on the morning after the election.

User-centered approach:

How did **my** election district vote?
Compared to the national level and other districts?

What we wanted: Offer useful interpretation, not just copying and publishing plain data you can find anywhere on the Internet.

Our solution:

- auto-generated texts based on the results of every single district
- bar charts for every district

Wie hat mein Wahlkreis gewählt?

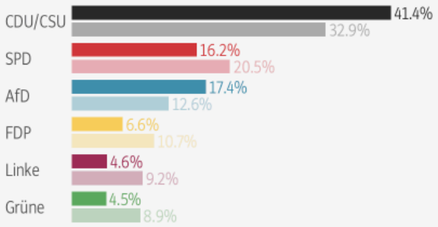
Geben Sie hier Ihren Wohnort ein und wählen Sie den entsprechenden Bezirk oder Wahlkreis aus.

Teublitz

Zufälliger Wahlkreis

Das sind die Ergebnisse für den **Wahlkreis Schwandorf**.

Für jede Partei zeigt der obere dunkle Balken das Ergebnis des Wahlkreises an, der untere hellere den Stimmenanteil bundesweit:



Dieser Wahlkreis hat **außergewöhnlich** gewählt. Stärkste Kraft ist die **CDU/CSU** mit 41.4 Prozent. Damit liegt sie **24 Prozentpunkte** vor der zweitplatzierten **AfD** mit 17.4 Prozent.
Die **CDU/CSU** schneidet besser ab als im Bundesdurchschnitt. Die **Grünen** schneiden schlechter ab als im Bundesdurchschnitt.
Für die **Linken** und die **FDP** lief es **nicht gut**: Von allen 299 Wahlkreisen liegen sie auf dem **viert-** und **neuntletzten Platz**.
Übrigens, eine Kleinstpartei fällt auf: Bundesweit erzielen die **Freien Wähler** hier ihr **zweitbestes Resultat**. Wäre nur dieser Wahlkreis entscheidend, kämen die **Freien Wähler** mit 5.1 Prozent sogar in den Bundestag.

Im nächsten Bundestag vertritt **Karl Holmeier** von der **CSU** den Wahlkreis.

Die Wahlbeteiligung liegt bei **75.3 Prozent** und liegt damit um 10 Prozentpunkte **über** der Wahlbeteiligung zur Bundestagswahl 2013 in diesem Wahlkreis.

Vorläufiges Ergebnis: 25.09.2017 04:03

Did a district vote extraordinarily? Similar to the national level? Just slightly different?

Our method for finding differences between the districts was calculating jenks buckets for every party.

The big advantage of Jenks is that they refer to the data and cluster them by the range: *"seeks to minimize the variance within categories, while maximizing the variance between categories"* (R documentation `BAMMtools`)

The hardest part: auto-generated German sentences that are good enough to be published at SZ, a newspaper that is proud of its high text quality (German grammar!)

Because hardly none of the election districts are similar to the others, the possiblities of combining the sentences are “endless”.

```
if (abs(diff) >= 9){
#brutale abweichung
    rv_first <- "Dieser Wahlkreis hat <b>außergewöhnlich</b> gewählt. "
    rv_pp <- get_stories(df_pp_diff, tell_story_pparty_extreme)
    rv_kleine <- get_kleine_stories(df_kleine_diff)
} else {
# check concrete jenk_diffs for telling the story
## wenn eine volkspartei um 2 jenks schwankt, allgemeiner satz schwankt sehr
if(check_pps_extreme(df_pp_diff)){
  rv_first <- "Dieser Wahlkreis <b>weicht deutlich ab </b> vom Wahlergebnis in ganz Deutschland. "
  rv_pp <- get_stories(df_pp_diff, tell_story_pparty_extreme)
}
## wenn eine der volksparteien um ein jenks schwankt, weicht er bisschen ab.
if(check_pps_simple(df_pp_diff)){
  if(rv_first == ""){
    rv_first <- paste0("Der Wahlkreis hat alles in allem <b>recht ähnlich zu ganz Deutschland</b> gewählt. ")
  }
}
## wenn eine der kleinen um 2 jenks schwankt, allgemeiner satz, weicht ab
if(check_kleine(df_kleine_diff)){
  if(rv_first == ""){
    rv_first = "Das Ergebnis in diesem Wahlkreis unterscheidet sich nur ein bisschen. Die kleine Variation liegt überwiegend an den kleinen Parteien. "
    rv_kleine = get_kleine_stories(df_kleine_diff)
  }
}
if( !check_pps_extreme(df_pp_diff) && !check_pps_simple(df_pp_diff) && !check_kleine(df_kleine_diff)){
  rv_first <- paste0("Die Menschen in diesem Wahlkreis haben recht ähnlich zu allen deutschen Wähler abgestimmt. ")
}
```

From where we started:

- a `xml` from the Bundeswahlleiter with election results from 2013 as training set
- we knew the structure of the `xml` file that will be updated all night long after the election in 2017

Meaning: we were prepared!

```
### the main script:
# source("config-testdaten.R")
source("config-livedaten.R")

### das datum und uhrzeit wird nicht mitgeliefert und muss hier händisch eingetragen werden
stand_der_daten <- "22.09.2017 16:30"


### zuerst die erstimmengewinner aus der xml-datei extrahieren
source("1_transform/bwl_get_winners.R")

### nach den erstimmengewinnern, werden die Daten aus der csv-datei aufbereitet und mit den erstimmengewinnern gemerged
### hier werden csv-dateien gespeichert, für die deutschlandweiten Ergebnisse, mit Wahlkreisdaten und für die Wahlkreiskarten
### Ort: 0-daten/output/2017/
source("1_transform/bwl_transform.R")

### aus der dpa-historie heraus läuft das script mit zahlenangaben dafür, ob ein wahlkreis ausgezählt ist oder nicht.
### da wir eh erst veröffentlichen, wenn alles ausgezählt ist, setzen wir die historische variabel auf ausgezählt: 4
### wenn aber noch keine daten da sind, soll live der satz stehen, dieser wahlkreis ist noch nicht ausgezählt. das wird
### dann in die daten geschrieben, wenn die varible 0 ist.
### alle ausgezählt: 4
### noch keine daten: 0

ALL_WK_COUNTED <- 4

### hier wird die analyse zu den Wahlkreisergebnissen gemacht. Standardabweichung, Jenks-Buckets und -diffs und Hitlisten errechnet
### speichert eine csv, die die Analysewerte enthält
### Ort: 0-daten/output/2017/
source("2_analyse/bwl_analyse_wk_data.R")

### hier wird der wahlkreisvergleich zusammengebaut
source("3_wk_vergleich/write_wk_data_jenks.R")

# Upload to Server
system(UPLOAD_SCRIPT)
```

When all you got is a hammer, everything looks like nails

"Hire and Fire" unter Trump (German, English)

Problem: In many newsrooms the software for publishing is not up-to-date. For example, the Content Management Systems are not flexible enough to cope with modern formats in journalism.

Solution:

- We created a system that uses Google Docs as input, by using ArchieML created by the NYTimes.
- we import that data into R and build `html` websites with `.Rmd` files.

Content with key-value-pairs in a Google Doc:

[section]

Name: Scott Pruitt
Art: Rücktritt
Rolle: Umweltminister
Tage: 504
Datum: 5. Juli 2018
Bild: pruitt.png
[.+text]
Niemand sonst in der Trump-Regierung hat derart viele Fragezeichen und hochgezogene Augenbrauen verursacht wie Pruitt. Mehr als ein Dutzend Untersuchungen sind in seinem Fall anhängig. Pruitt, 50, hat in seinem früheren Leben als Justizminister von Oklahoma die Umweltbehörde EPA über ein Dutzend Mal verklagt. Dann wurde er Chef der Behörde und nutzte sein Amt, um so viele Umweltregeln wie möglich außer Kraft zu setzen. Wirtschaftsförderung sollte klar vor Umweltschutz stehen. Die konsequente Umsetzung dieser Haltung hat ihn lange vor dem Rücktritt bewahrt. Aber Pruitts Extravaganzen waren wohl auch für Trump zuletzt zu sehr zur Belastung geworden. Hier eine Auswahl: Er soll gerne und häufig erster Klasse in seine Heimat Oklahoma geflogen sein, obwohl die Richtlinien der Regierung zu Sparsamkeit mahnen. Als EPA-Direktor hat er einige Monate in dem Washingtoner Appartement der Frau eines befreundeten Energie-Lobbyisten gewohnt. Beste Lage für 50 Dollar die Nacht. Er hat sich einen Rund-um-die-Uhr-Personenschutz zuweisen lassen. Er ließ sich für mehrere zehntausend Dollar eine schalldichte Telefonbox in seinem Büro installieren - um vertrauliche Gespräche führen zu können. Unter anderem, um einen Mitarbeiter nach einer einer gebrauchten Matratze aus einem Trump-Hotel zu schicken.
[]

Simple example for an `.Rmd` file:

````
---
name: ""
author: ""
output:
  html_document:
    theme: null
    css: style.css
    self_contained: no
    highlight: null
    mathjax: null
---
```{r load data, results='asis', echo=FALSE, warning=F, message=F}
load("data/archie_list.RData")
```

```{r create html, results='asis', echo=FALSE, warning=F, message=F}

for (section in names(archie_list)[1]) {

  for (i in 1:nrow(archie_list[[section]])) {

    name <- archie_list[[section]][["Name"]][[i]]

    cat(paste0("<div><b>Name: </b> ", name, "</div>"))

    # write more code for other key-value-pairs
  }
}

```
````

Example code as HTML output:

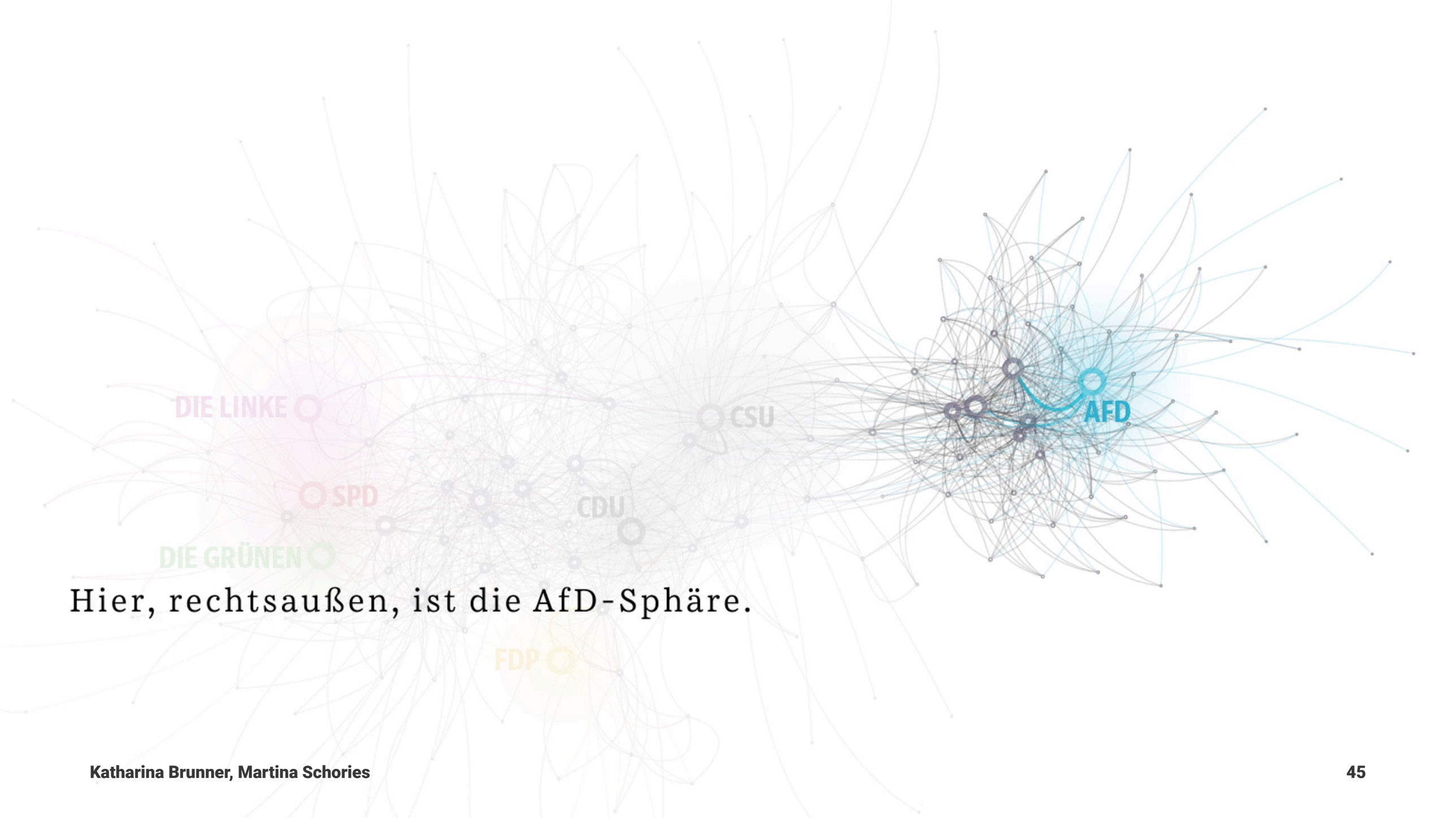
Name: Scott Pruitt
Name: Thomas Bossert
Name: H.R. McMaster
Name: John McEntee
Name: Rex Tillerson
Name: Gary Cohn
Name: Hope Hicks
Name: Rob Porter
Name: Andrew McCabe
Name: Brenda Fitzgerald
Name: Taylor Weyeneth
Name: David Sorensen
Name: Rachel Brand
Name: Omarosa Manigault Newman
Name: Tom Price
Name: Keith Schiller
Name: Sebastian Gorka
Name: Carl Icahn
Name: Stephen Bannon
Name: Kenneth C. Frazier
Name: Anthony Scaramucci
Name: Reince Priebus
Name: Sean Spicer
Name: Walter Shaub
Name: Robert Iger
Name: Elon Musk
Name: Mike Dubke
Name: Kathleen T. McFarland
Name: James Comey
Name: Angella Reid
Name: Katie Walsh
Name: Craig Deare
Name: Mike Flynn
Name: Travis Kalanick
Name: Sally Yates

Getting an idea of the blackbox Facebook

Der Facebook-Faktor - Wie das soziale Netzwerk die Wahl beeinflusst (German, English)

Goal: Investigating the political sphere on Facebook by crawling the sites of political parties and active users.

- We evaluated more than one million public Facebook likes from a little less than 5000 politically interested Facebook users.
- Data is stored in the graph database Neo4J



Hier, rechtsaußen, ist die AfD-Sphäre.

Role of R:

- export data from Neo4J
- aggregate the user data to get a sense of echo chambers
- **a lot** of data transforming

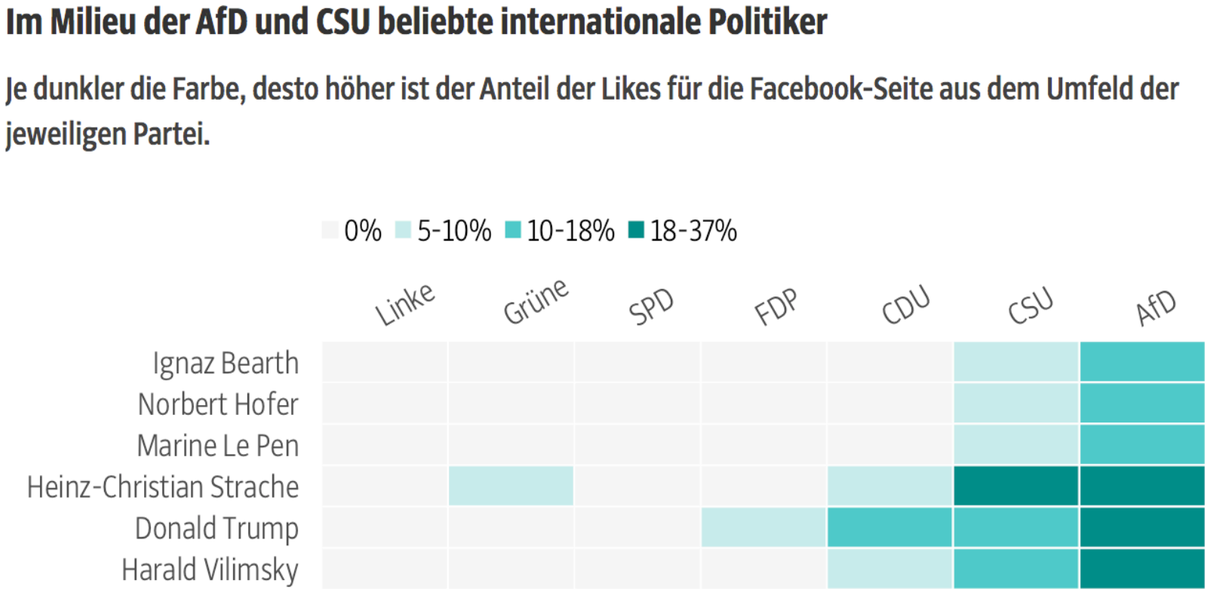
System:

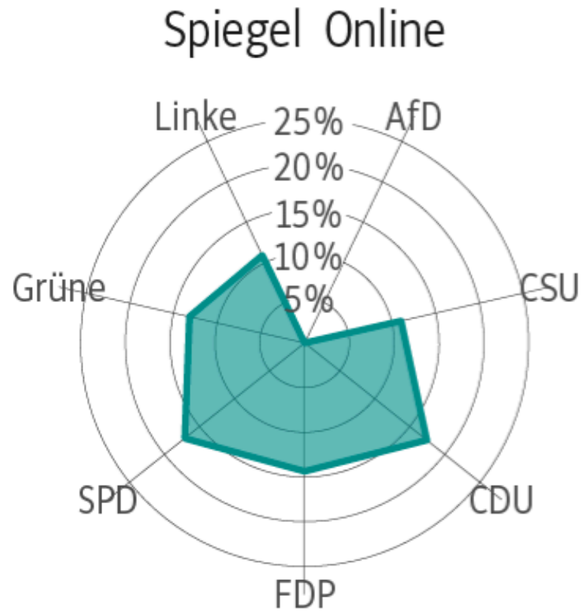
user	affiliated party	like for fb page
User 1	spd	SZ
User 1	spd	Rbloggers
User 2	fdp	SZ
User 2	fdp	R
User 2	fdp	RStudio
User 3	spd	SZ

Calculate overlaps of Facebook spheres:

party	fb page	weight
spd	SZ	2
spd	Rbloggers	1
fdp	SZ	1
fdp	R	1
fdp	RStudio	1

Heavy use of `ggplot` for plotting: We produced many heatmaps and radar plots with R. But refining in Illustrator neccessary. All plots must work on mobile phones, too.





More on the Facebook Factor

- Main story: Der Facebook Faktor
- Methodology: So haben wir die Daten recherchiert
- Von AfD bis Linkspartei - so politisch ist Facebook
- Was links und rechts verbindet - und trennt
- Wie es in Facebooks Echokammern aussieht - von links bis rechts
- All stories

Ideas, Comments, Questions? Feel free to contact us