

# **Web Data Collection with R**

## **Course Taster Teaser**

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

**Introduction**

**Applications**

**Conclusion**

# Introduction

# Course Taster

find a course taster at:

[http://pmeissner.com/downloads/user2015\\_meissner\\_webscraping.pdf](http://pmeissner.com/downloads/user2015_meissner_webscraping.pdf)

## Course Teaser

... back to the course teaser

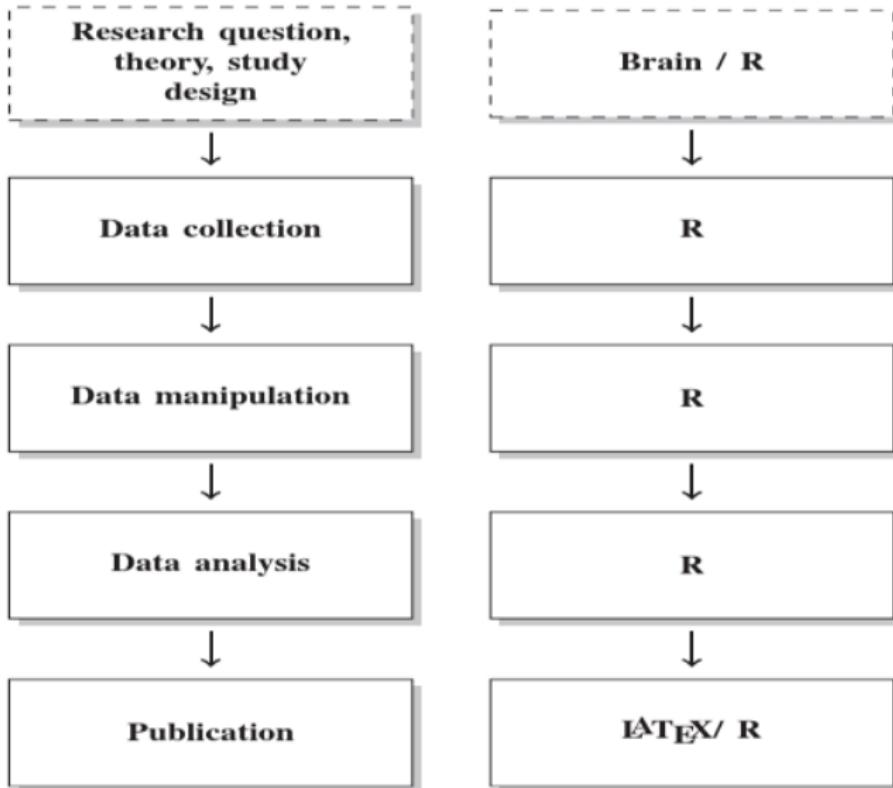
# THE WEB

- ▶ **web pages** (e.g. <http://example.com>, <http://ecpr.eu/>)
- ▶ **web formats** (XML, HTML, JSON, ...)
- ▶ **web frameworks** (HTTP, URL, APIs, ...)
- ▶ **social media** (Twitter, Facebook, LinkedIn, Snapchat, Tumbler, ...)
- ▶ **data in the web** (politician's biography, laws, policy reports, news, ... )
- ▶ **web data** (page views, page ranks, IP-addresses, ... )

# THE PROBLEMS

phase	problems	examples
<b>download</b>	protocols	HTTP, HTTPS, POST, GET, ...
	procedures	cookies, authentication, forms, ...
<b>extraction</b>	parsing	translating HTML (XML, JSON, ...) into R
	extraction	getting the relevant parts
	cleansing	cleaning up, restructure, combine

# THE SOLUTION



# **Applications**

# MP Biographies

[www.bundestag.de/bundestag/abgeordnete18/biografien/C/caesar\\_cajus/258254](http://www.bundestag.de/bundestag/abgeordnete18/biografien/C/caesar_cajus/258254)

Startseite > Der Bundestag > Abgeordnete > Biografien > C

› Aufgaben

▼ Abgeordnete

› Biografien

› Nach Fraktionen

› Nach Bundesländern

› Wahlkreise

› Gesamtliste

› Ausgeschiedene

› Nebentätigkeiten

› Entschädigung

› Abgeordnete in Zahlen

› Plenum

› Ausschüsse

› Weitere Gremien

› Präsidium

› Ältestenrat

› Fraktionen

› Wahlen



© Cagus Caesar / Matthias Herbst

## Cagus Caesar, CDU/CSU

### Diplom-Forstingenieur

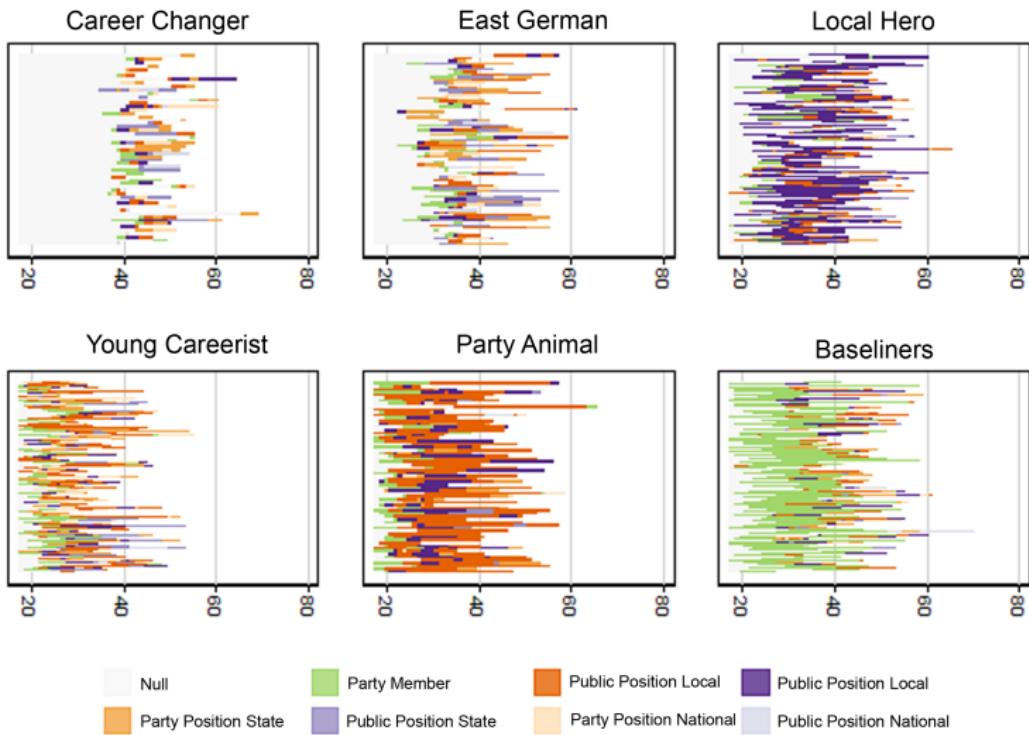
Geboren am 22. Januar 1951 in Rinteln; evangelisch; verheiratet mit Gudrun Caesar; zwei Söhne und eine Tochter; wohnhaft in Kalletal-Westorf Kreis Lippe.

Mittlere Reife; 1969 bis 1971 Forstlehre Landesforstschule Arnsberg Fachhochschulreife; 1971 bis 1974 Revierförsteranwärter; Diplom-Forstingenieur; 1974 bis 1978 Revierleiter Forstrevier Lage; 1978 bis 1980 Forsteinrichtung Landesverband Lippe; 1980 bis 1998 Revierleiter Forstrevier Kirchberg; 2006 Projektleiter Forstmanagement Landesverband Lippe; 1998 bis 2005 und ab 2007 Bundestagsabgeordneter der CDU

Seit 1969 Mitglied der CDU; 1980 bis 1984 Ortsvorsitzender der JU Kalletal; 1985 bis 1990 Gemeindeverbandsvorsitzender der CDU Kalletal; seit 1990 Kreisvorsitzender der CDU Lippe; seit 1992 Bezirksvorstandsmittelglied der CDU Ostwestfalen-Lippe; 1980 bis 1999 Mitglied der Gemeinderatsfraktion Kalletal; 1984 bis 1999 Mitglied des Lippischen Kreistages; u. a. Vorsitzender des Umweltausschusses und stellvertretender Fraktionsvorsitzender; 1986 bis 1999 Mitglied der Landschaftsversammlung beim Landschaftsverband Westfalen-Lippe in Münster; dort u. a. Mitglied im Personal- und Kulturausschuss; umweltpolitischer Sprecher und stellvertretender Fraktionsvorsitzender der CDU; 2000 Verleihung der Freiherr-vom-Stein-Medaille; 1999 bis 2005 Abgeordneter der Verbandsversammlung des Landesverbandes Lippe; dort Fraktionsvorsitzender der CDU; 1998 bis 2005, 2007 bis 2009 und seit 2011 Bundestagsabgeordneter der CDU für den Wahlkreis Lippe I (135).

Bailer, Mei  ner, Ohmura, Selb (2013): Seiteneinstieger im Deutschen Bundestag. Springer VS

# MP Biographies



Bailer, Mei<sup>ß</sup>nner, Ohmura, Selb (2013): Seiteneinstieger im Deutschen Bundestag. Springer VS

# Legislative Process

dipbt.bundestag.de/dip21.web/searchProcedures/simple\_search\_list.do?selId=68572&method=select&offset=0&anzahl=100&sort=3

- ▼ Beratungsabläufe
  - ▶ Einfache Suche
  - ▶ Erweiterte Suche
- ▼ Aktivitäten
- ▼ Dokumente
  - ▶ Einführung
  - ▶ Dokumente ab 1949
  - ▶ DIP 8.-15. Wahlperiode
  - ▶ Kontakt
  - ▶ Kontakt
  - ▶ Parlamentsdokumentation
  - ▶ Fachinformationen und Analysen
  - ▼ Parlamentsarchiv
  - ▼ Datenhandbuch
  - ▼ Bibliothek
  - ▼ Web- und Textarchiv
  - ▶ Registrierte Verbände

[zurück](#) Datensatz 100/429 [weiter >](#)

**Für Dateiausgabe merken:**  [Drucken](#)

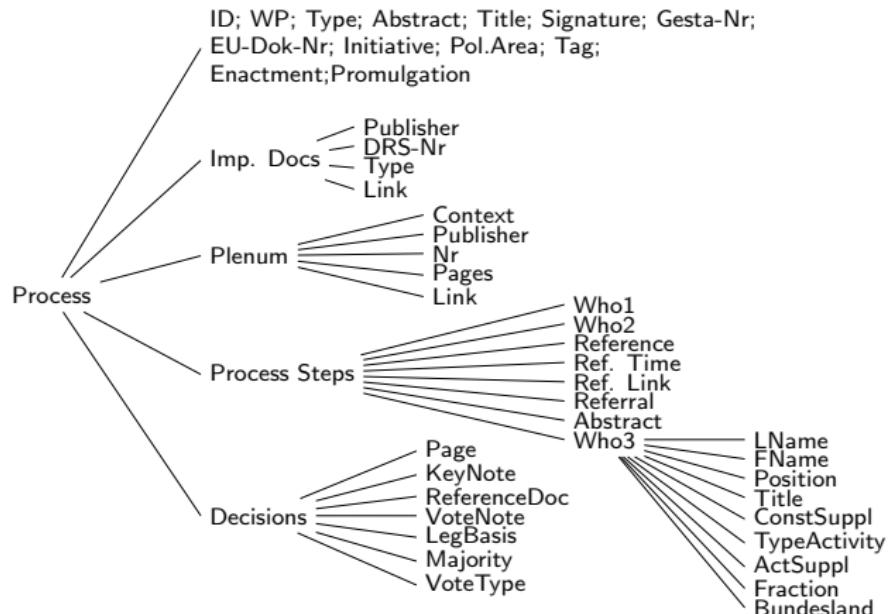
**Basisinformationen über den Vorgang**

[ID: 18-68572] [Version für Lesezeichen / zum Verlinken](#)

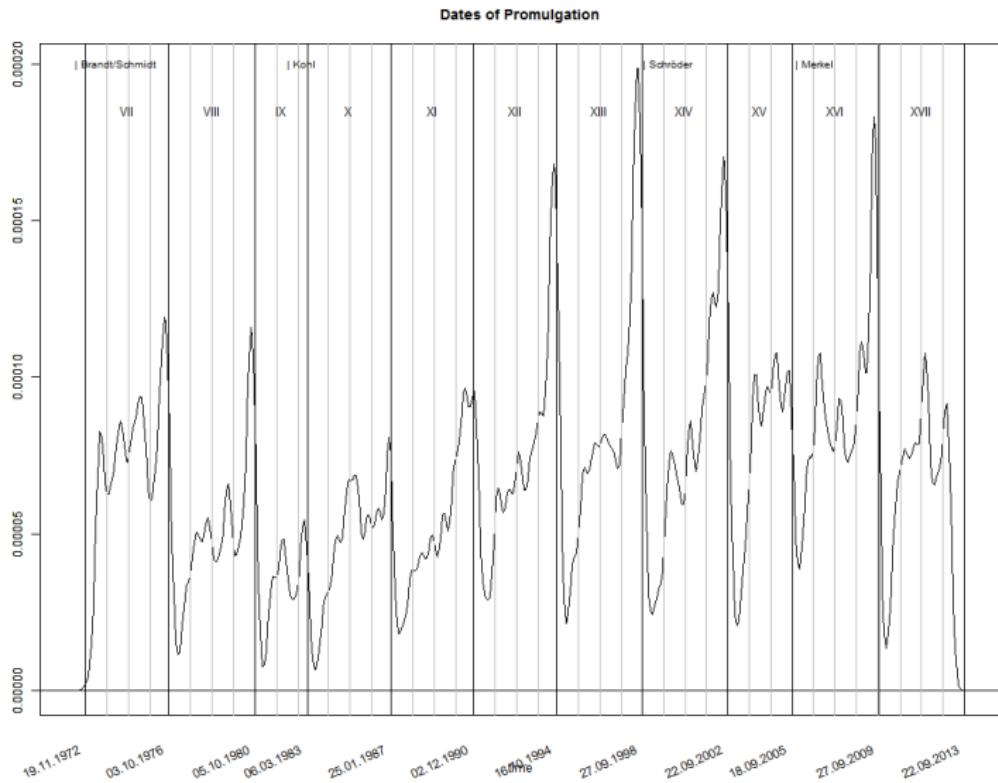
<b>18. Wahlperiode</b>	Gesetzgebung
	<b>Gesetz zur Durchführung der Verordnung (EU) Nr. 1007/2011 und zur Ablösung des Textilkennzeichnungsgesetzes</b>
<b>Vorgangstyp:</b>	Bundesregierung
<b>Initiative:</b>	Bundesregierung
<b>Aktueller Stand:</b>	Verkündet
<b>GESTA-Ordnungsnummer:</b>	E017
<b>Zustimmungsbedürftigkeit:</b>	Nein , laut Gesetzentwurf (Drs 362/15) Ja , laut Bundesrat (Drs 362/15(Beschluss)) Nein , laut Bundesregierung (Drs 18/6488) Nein , laut Verkündung (BGBl I)
<b>Wichtige Drucksachen:</b>	<a href="#">BR-Drs 362/15 (Gesetzentwurf)</a> <a href="#">BT-Drs 18/6488 (Gesetzentwurf)</a> <a href="#">BT-Drs 18/6662 (Beschlussempfehlung und Bericht)</a>
<b>Plenum:</b>	<a href="#">1. Durchgang: BR-PIPr 936 , S. 332A - 332B</a> <a href="#">1. Beratung: BT-PIPr 18/133 , S. 12935D - 12936A</a> <a href="#">2. Beratung: BT-PIPr 18/136 , S. 13282A - 13282B</a> <a href="#">3. Beratung: BT-PIPr 18/136 , S. 13282B</a> <a href="#">2. Durchgang: BR-PIPr 940 , S. 513B - 513C</a>
<b>Verkündung:</b>	<a href="#">Gesetz vom 15.02.2016 - Bundesgesetzblatt Teil I 2016 Nr. 8 23.02.2016 S. 198</a>
<b>Sachgebiete:</b>	Wirtschaft

# Legislative Process

## What Kind of Information Can We Get?

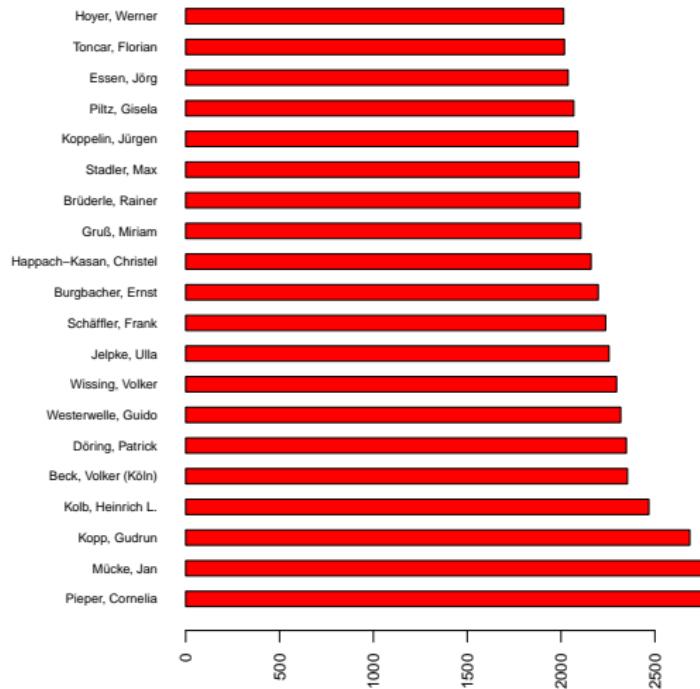


# Legislative Process



# Legislative Process

**Distribution of Most Active Persons in BT**  
- Top 20 -

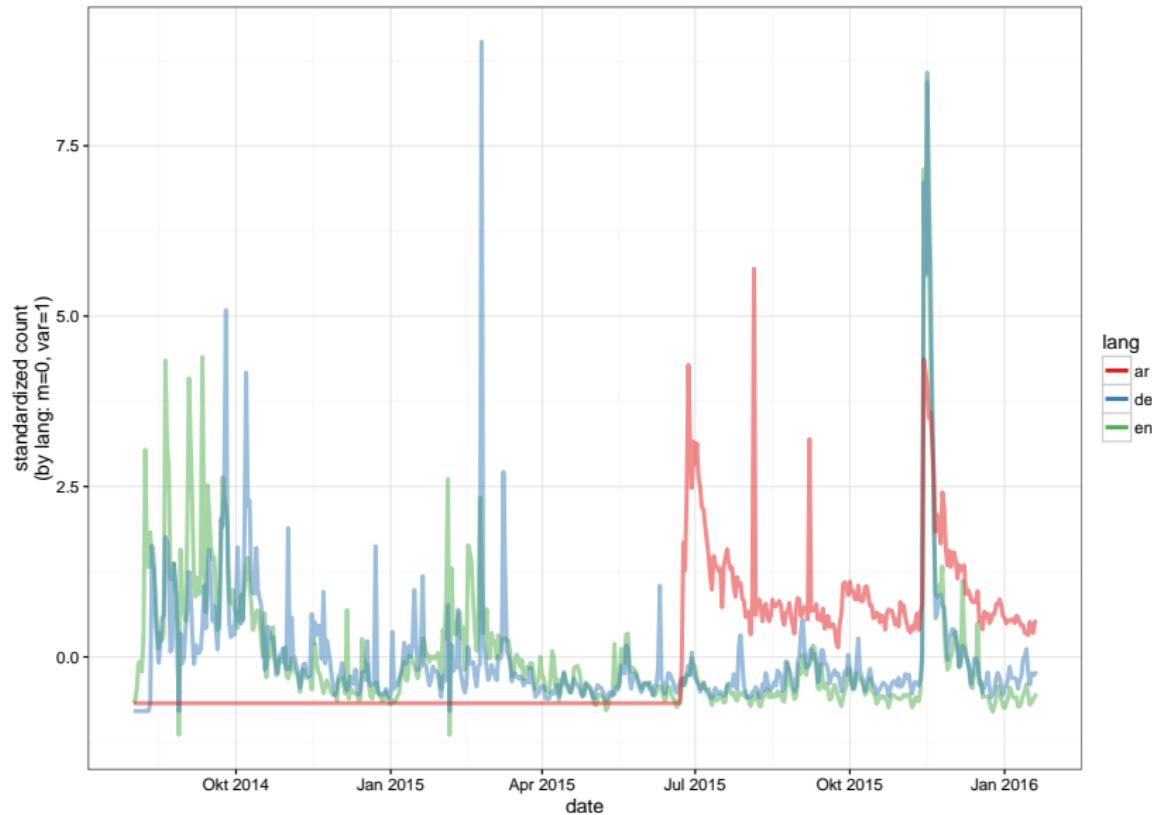


# Legislative Process

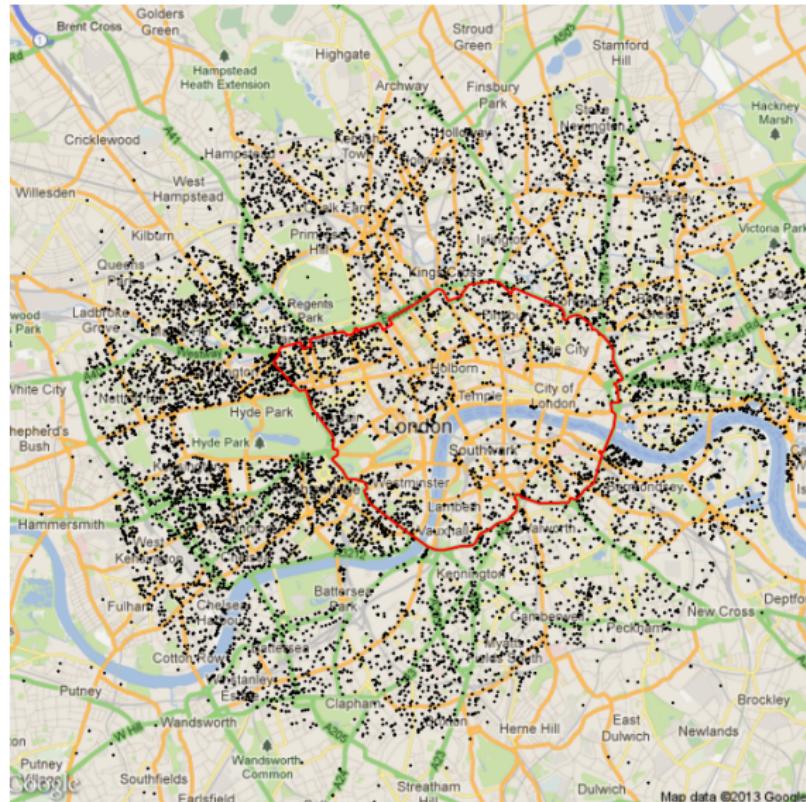
Distribution of Referral  
to Leading Committee  
– Top 20 –



# Wikipedia Page Views - IS



# Policy Effects

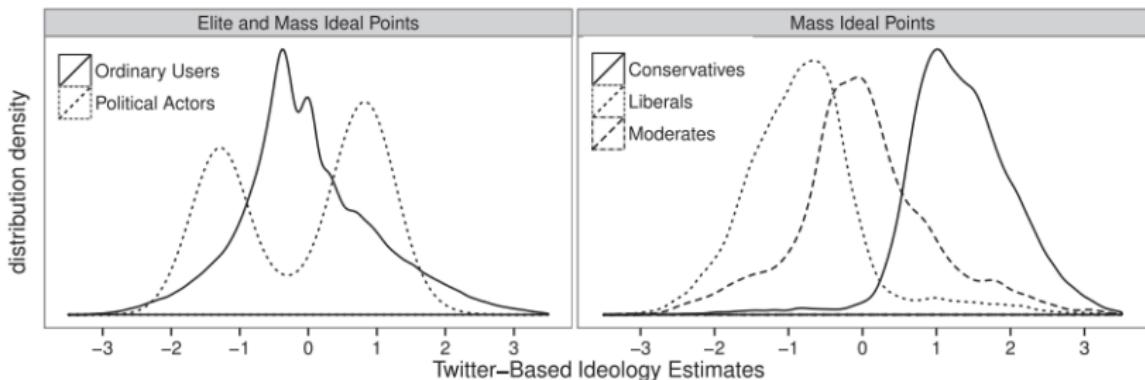


Map data ©2013 Google

# Mass Idealpoint Estimation

10

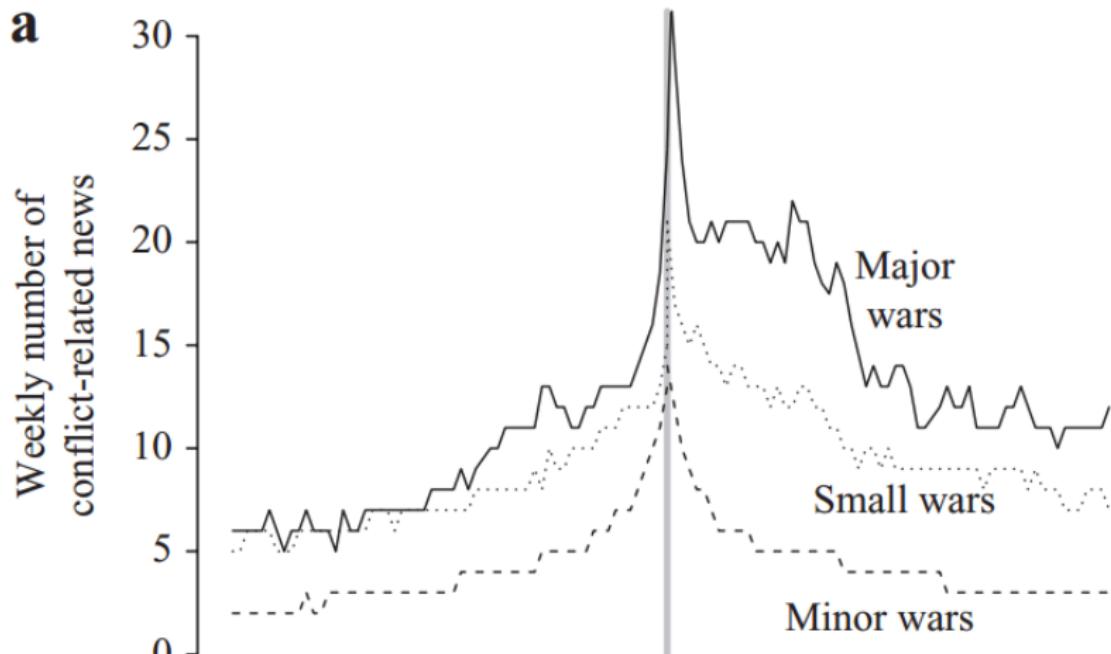
Pablo Barberá



**Fig. 4** Distribution of political actors and ordinary Twitter users' ideal points.

Barberá (2014)

# News Based War Prediction



Chadefaux (2014)

# Collective Action and Organization Formation

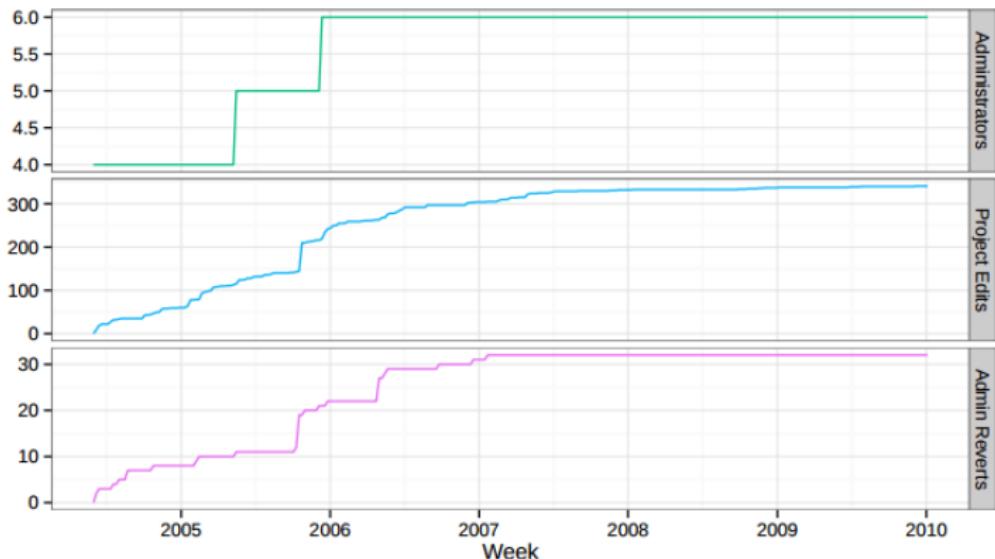
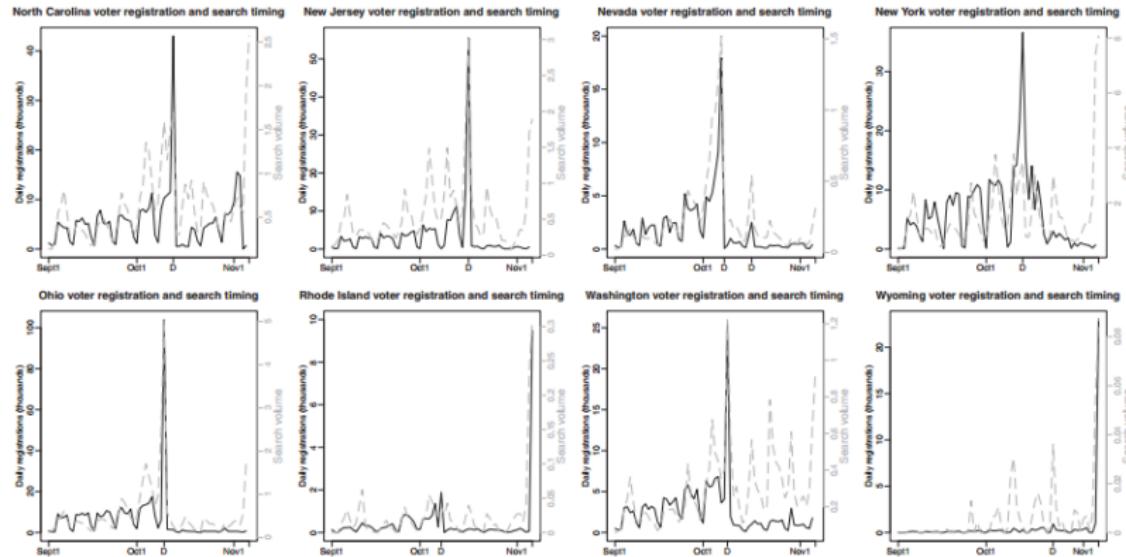


Figure 1: Cumulative plots of covariates for *Seattle Wiki*, a collaborative website for information about Seattle and one of the online communities in our dataset.

Shaw & Hill (2014)

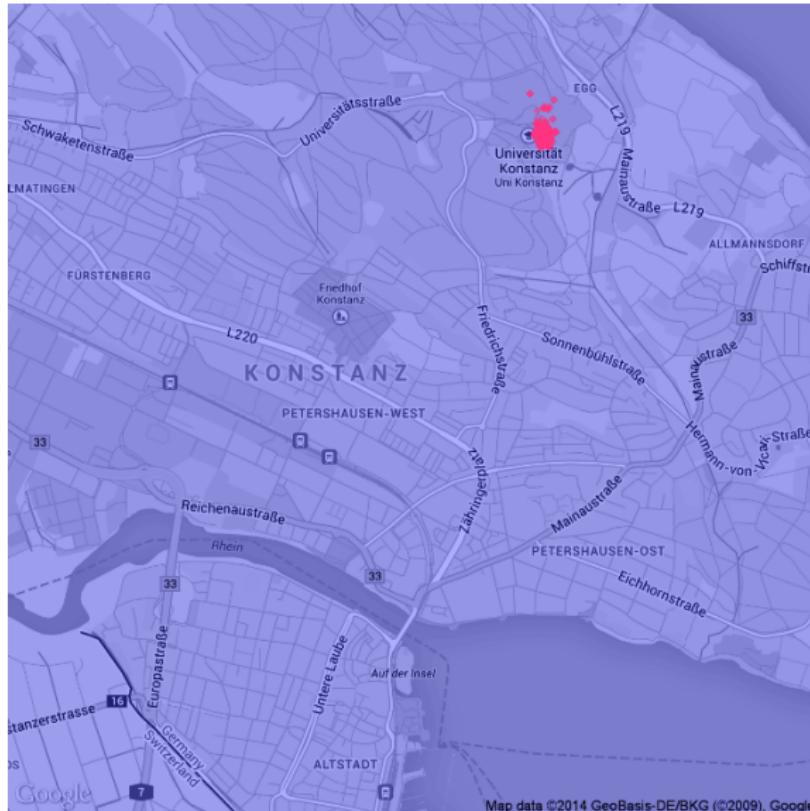
# Electoral Rule Effects



**Fig. 1** Web searches for “voter registration” and observed registration numbers, September to November 2012. Black lines and left axes show daily registrations, in thousands. Dashed gray lines and right axes show standardized search volume. Horizontal axes show dates; D marks the mail and in-person registration deadlines (the same day in most states).

Street et al. (2015)

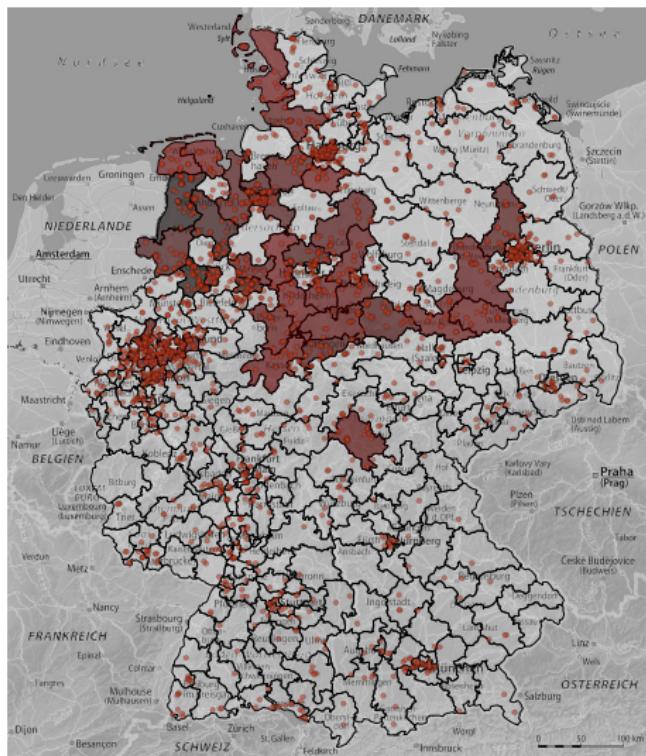
# Mobil Phone Meta Data



# Mobil Phone Meta Data



# Name Distribution



## **Conclusion**

# Conclusion

- ▶ applications are diverse and many fold
- ▶ the web is everywhere
- ▶ web data formats are not only in the web (e.g. EPub, Docx, KML are XML)
- ▶ data extraction skills (e.g. RegEx) are swiss army knives