Document Scraping and Text Analysis in R

Juraj Medzihorsky

2014-11-03


IRWIN
QUICK GRIP

Why R?

- Free
- Open source
- Large and diverse user community
- Flexible
- Multi-platform
- Advanced: add-on packages, graphics

Community

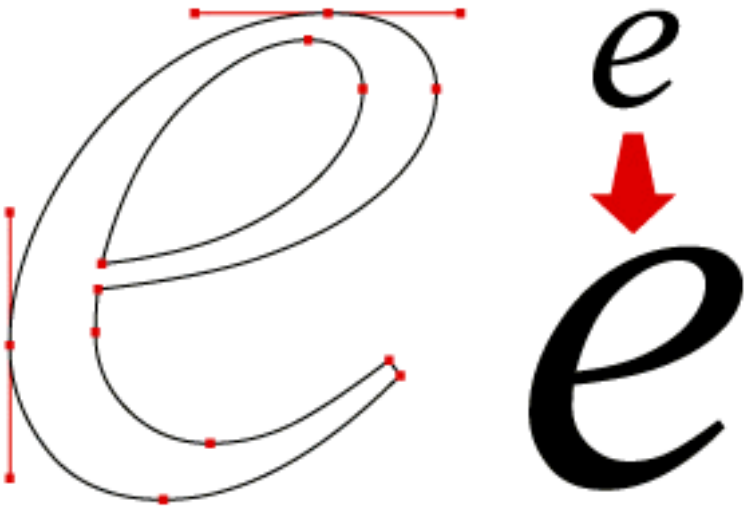
- Googling often more efficient than the manuals.
- Many specialized mailing lists.
- stackowerflow

Graphics

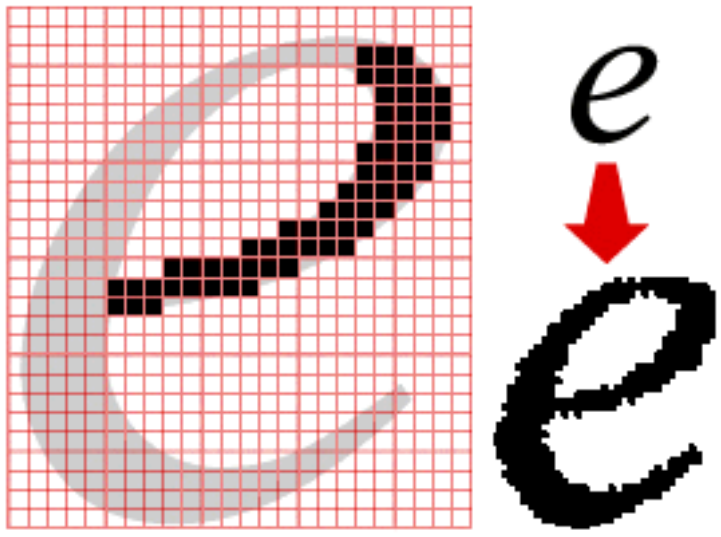
- Many options.
- Many formats: vector vs. raster graphics.

Vector vs. Raster Formats

VECTOR GRAPHICS



BITMAPPED (RASTER) GRAPHICS



Vector vs. Raster Formats

- Raster: .jpg, .tiff, .png, .bmp ...
- Vector: .pdf, .svg, .eps, .ppt ...

How to use R?

- **Directly:** command line, console.
- **GUI:** default, RStudio, JGR, Deducer etc.
- **Text editor + plugin:** Emacs, vim, Sublime Text etc.

Workspace

- Contains **objects**
- List contents with `ls()`

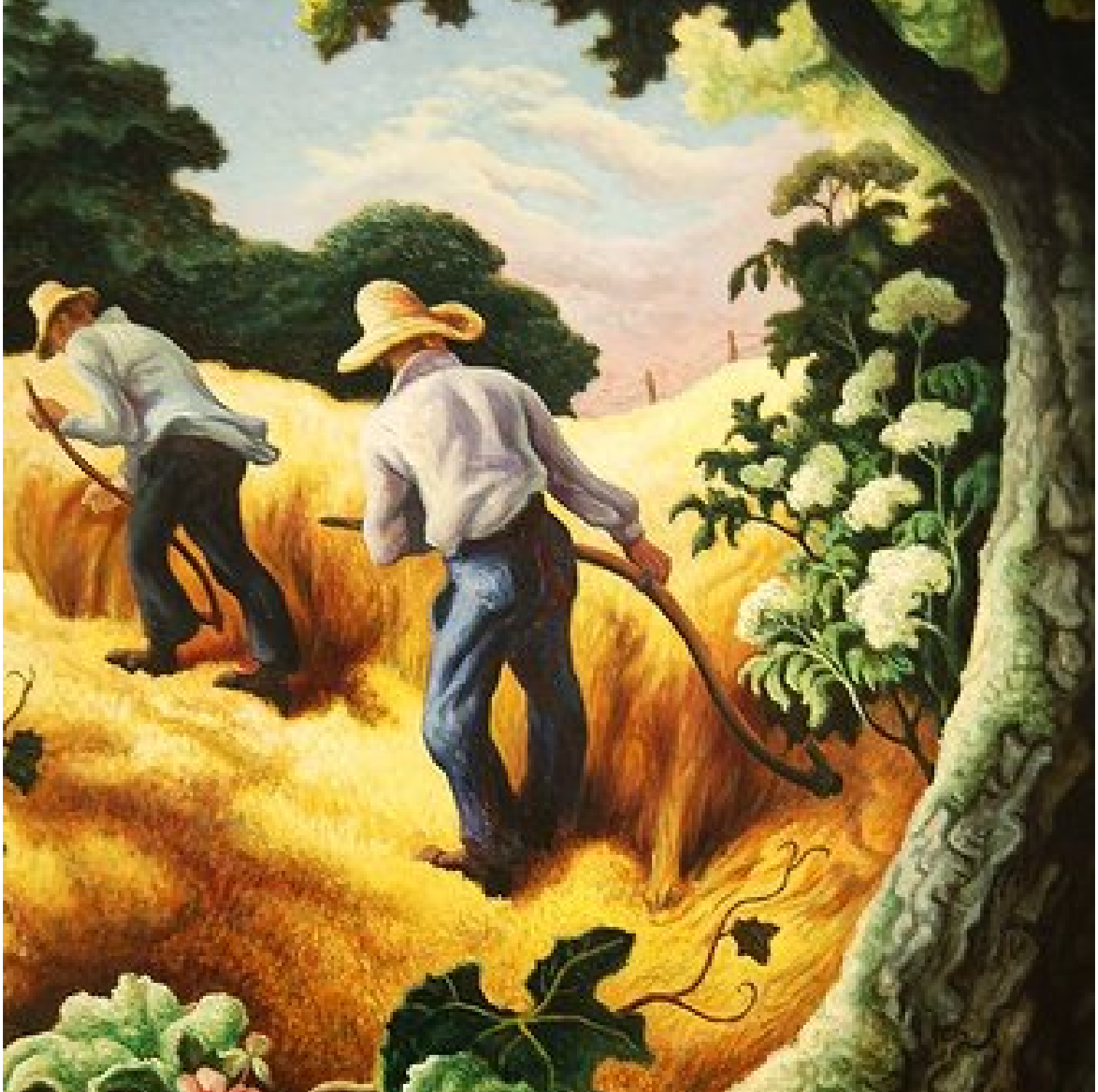
Objects

- Data
- Functions

Document Scraping

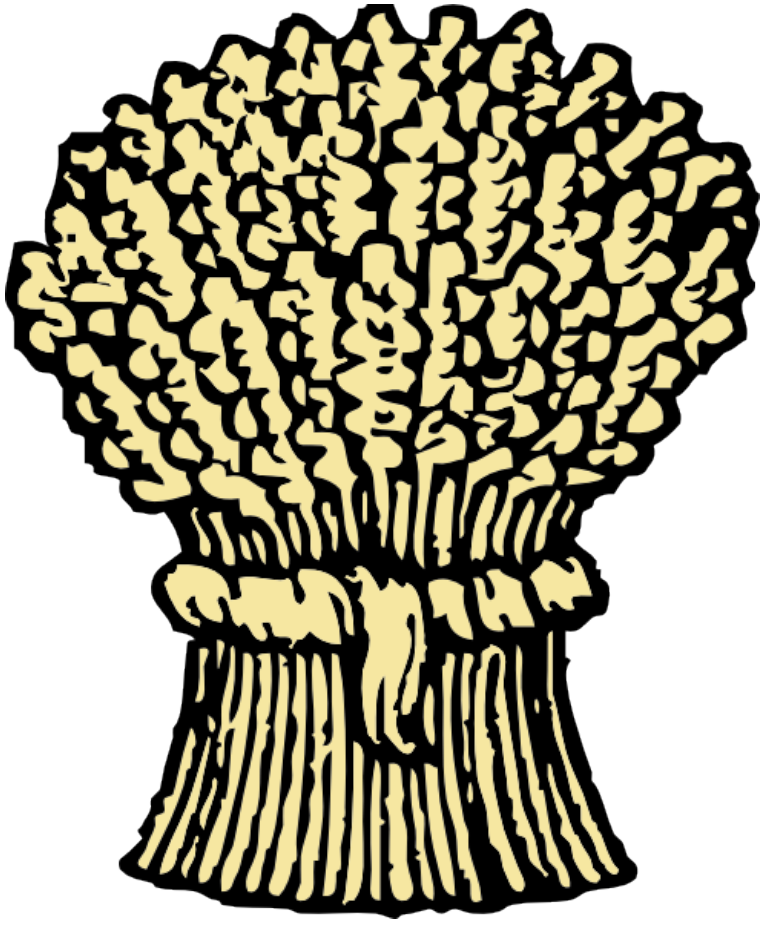
Document Scraping

- Numbers and text in files
- Local
- Web



Web Scraping

- HyperText Markup Language (HTML) familiarity



Text Analysis

Text Analysis

- Content Analysis
- 'Manual' vs Computer Assisted Text Analysis (CATA)

‘Manual’ Text Analysis

- Humans do most of the work
- Expensive
- Slow
- Reliability issues

Computer Assisted Text Analysis

- Computers do most of the work
- Boom
  - Huge amount of digitized text available
  - Cheap computing power
  - New methods – CS & PS

‘Bag of Words’

- Common assumption in CATA
- Order of words (n-tuplets of words) does not matter
- Text as vector of word counts

CATA in Political Science

- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

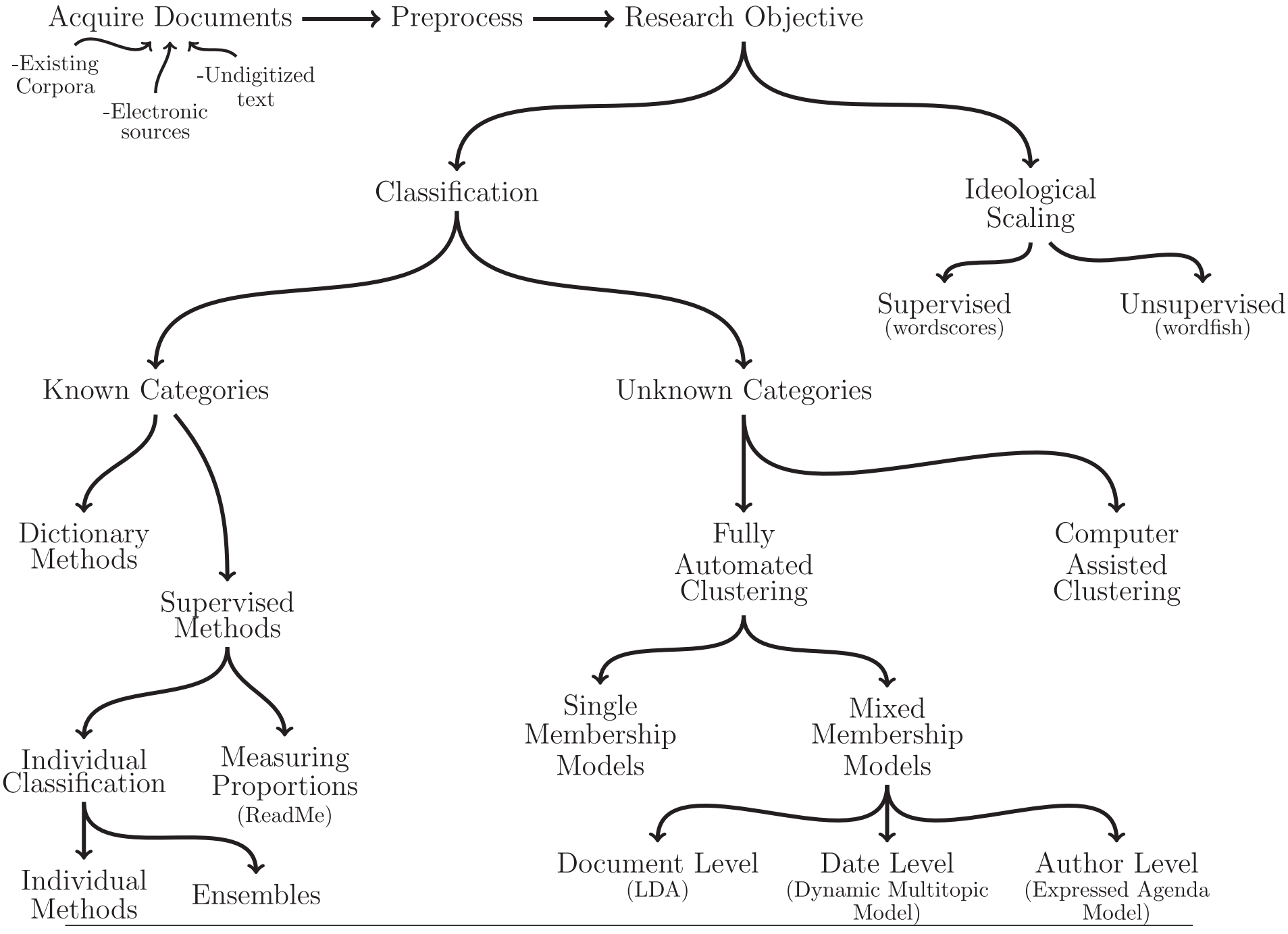


Fig. 1 An overview of text as data methods.

Scaling Goals

- 1 or more dimensions
- Place documents (texts, speeches) in space
- Place words in the same space

Scaling Methods

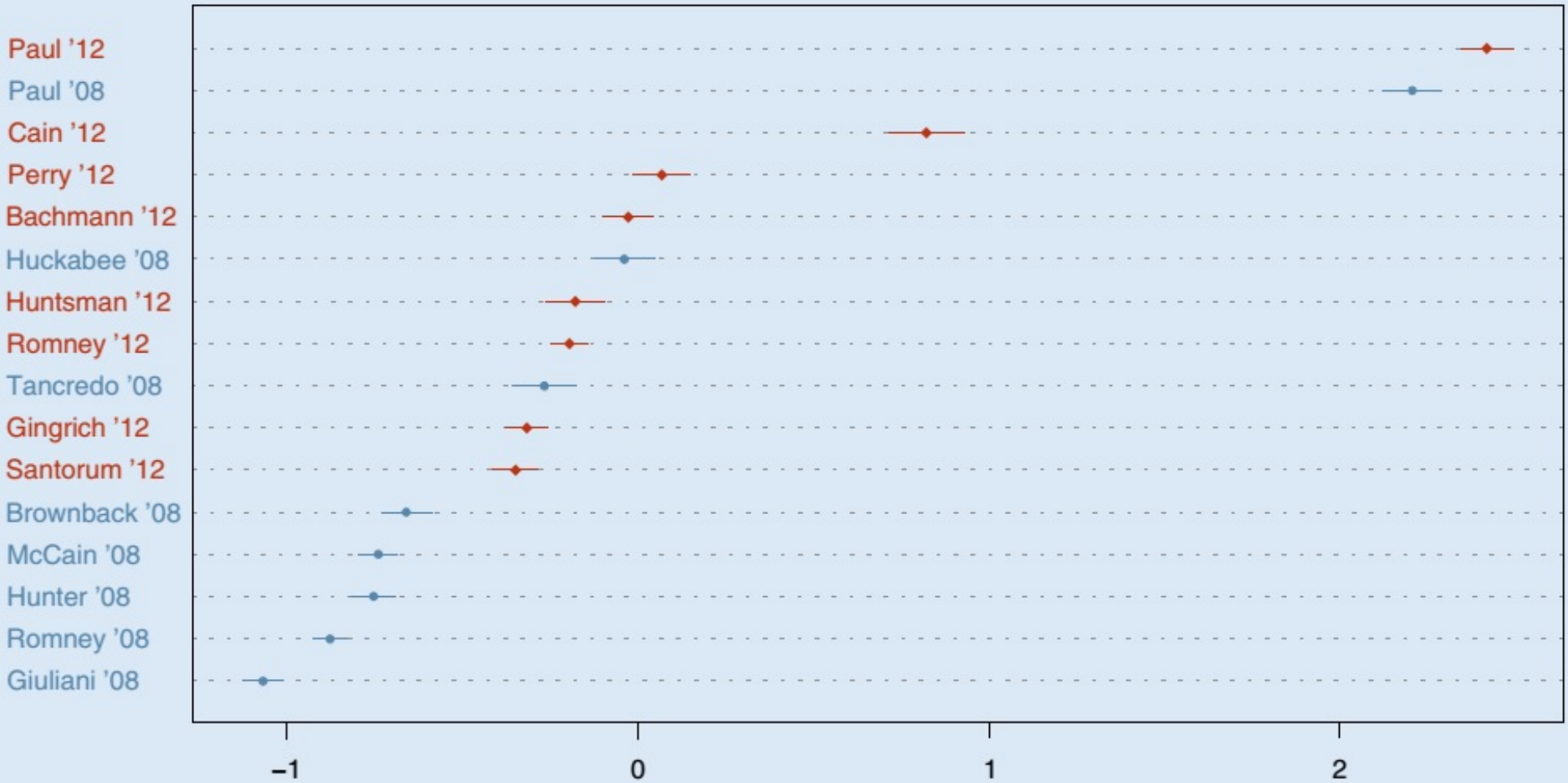
- Supervised: Wordscores
- Unsuperwised: Wordfish (1-factor Poisson IRT)
- Unsupervised: Correspondence Analysis

Scaling Example 1

- Republican presidential primaries for 2010 and 2014
- Transcripts of speeches from a UCSB website
- Expected move towards Tea Party positions
- Wordfish

Figure 1

Candidate Positions



Candidate positions extracted from their pre-Iowa debate speeches with bootstrapped 95% confidence intervals (1,000 replications). 2008 candidates denoted by circles (blue) and 2012 candidates by diamonds (red).

Table 4

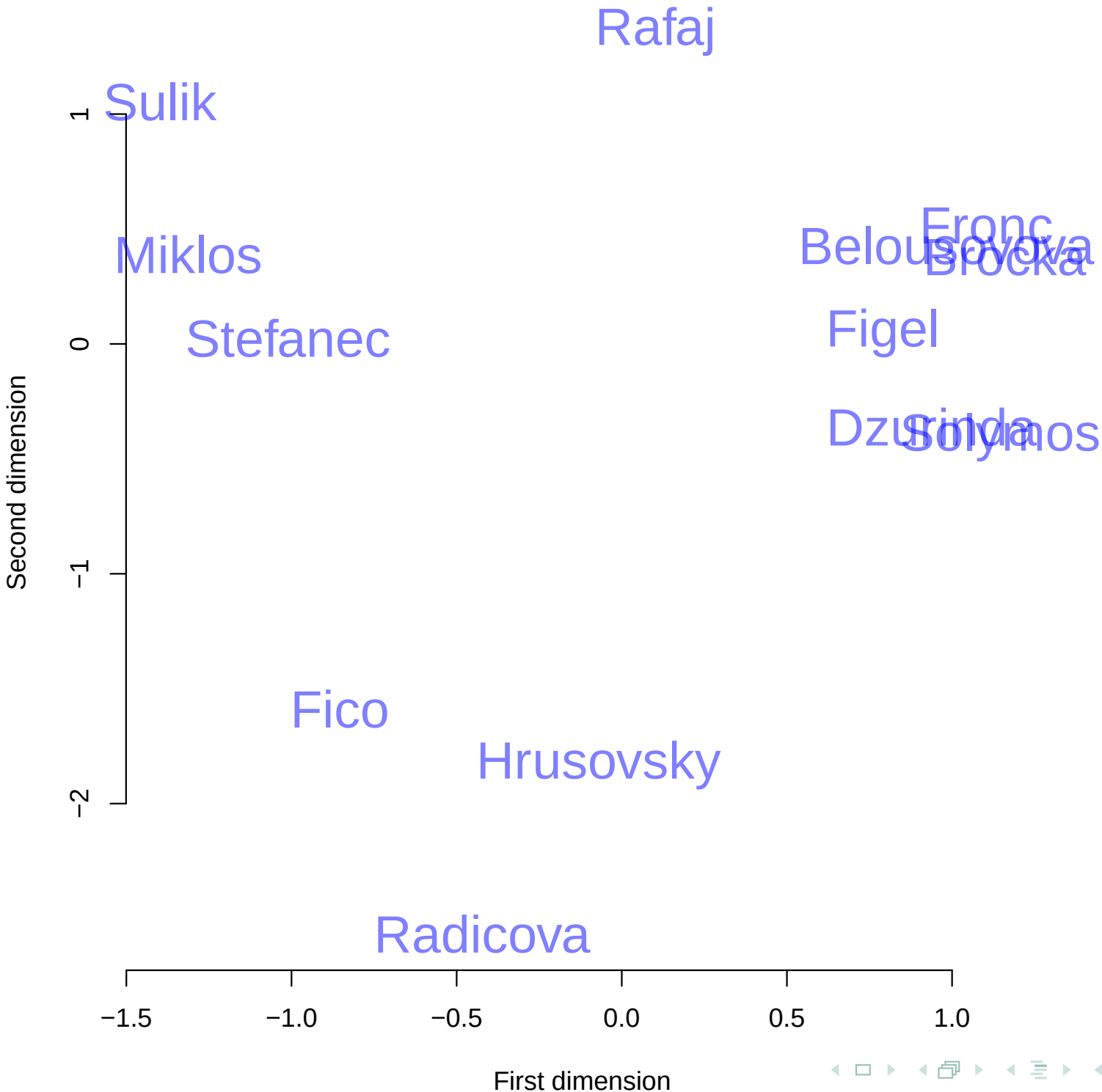
Selected Five-Sentence Sequences Spoken by a Single Candidate in a Single Debate

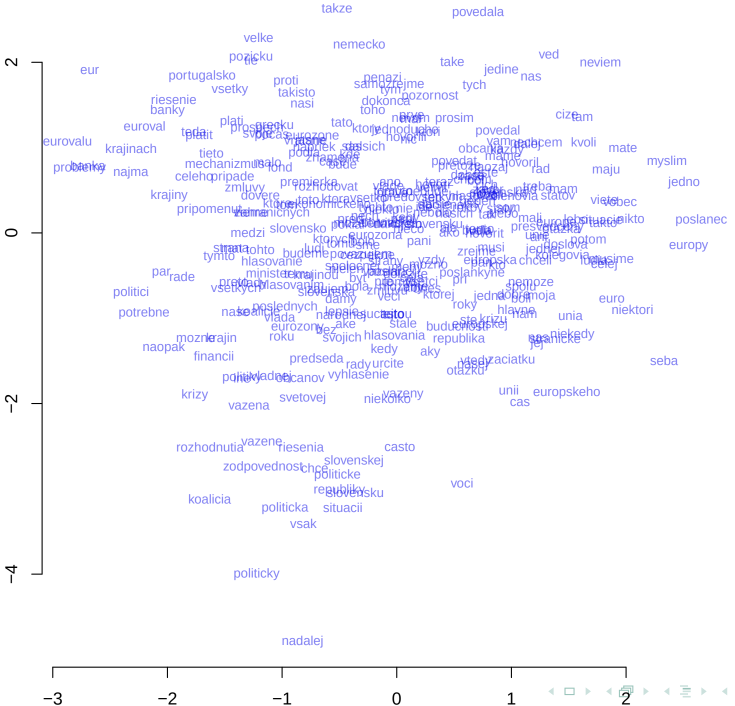
NEGATIVE, −1±0.1	POSITIVE, +1±0.1
"I have joined together across the aisle on a number of pieces of legislation, many of them very important. I'm proud of my legislative record of conserving my ideals and my conservative principles and getting things done in Washington. And I am proud of that, and I will continue to hold to those ideals. But I will reach across the aisle to the Democrats who I have worked with, who know me, and we know we can work together for the good of this country. Let's raise the level of dialogue and discussion and debate in this campaign." (McCain on December 12, 2007; score: −1.1)	"Repeal Dodd-Frank, repeal Obamacare. It really isn't that tough if you try. It is easy to turn around this economy, just have the backbone to do it. Well, as president of the United States, I would not be reappointing Ben Bernanke, but I want to say this. During the bailout, the $700 billion bailout, I worked behind the scenes against the bailout, because one of the things that I saw from the Federal Reserve, the enabling act legislation is written so broadly that, quite literally, Congress has given the Federal Reserve almost unlimited power over the economy." (Bachmann on September 12, 2011; score: 1.9)
"It's the one place I found to agree with President Obama. If every parent in America had a choice of the school their child went to, if that school had to report its scores, if there was a real opportunity, you'd have a dramatic improvement. I visited schools where, three years earlier, there were fights, there were dropouts, there was no hope. They were taken over by a charter school in downtown Philadelphia, and all of a sudden the kids didn't fight anymore, because they were disciplined. They were all asked every day, what college are you going to? Not are you going to go to college, what college are you going." (Gingrich on September 7, 2011; score: −1)	"If we look for it, you'll realize that our national sovereignty is under threat. Yes, and I would like to state that, to the statement earlier made that we all went to Washington to change Washington and Washington changed us, I don't think that applies to me; Washington did not change me. I would like to change Washington, and we could by cutting three programs, such as the Department of Education–Ronald Reagan used to talk about that–Department of Energy, Department of Homeland Security is the biggest bureaucracy we ever had. And besides, what we can do is we can have a stronger national defense by changing our foreign policy. Our foreign policy is costing us a trillion dollars, and we can spend most of that or a lot of that money home if we would bring our troops home." (Paul on November 28, 2007; score: 2)
"I can tell you a good union, the Steel Workers Union. When last year, Chris, we had a strike in a Kansas plant that made the tires for our humvees, I called up the president of the Steelworkers and the president of Goodyear, and within a very short period of time, they were working together, they got that thing done for the good of the country. A union is a receptacle of power, just like management. But those folks love this country, they love their family, and they helped to build a middle class, which has been important for America and for our party. We need to work with unions to win this presidency." (Hunter on October 9, 2007; score: −0.9)	"There's a responsible way for the federal government to do the things that it should do. Running organizations like the TSA, I would agree with Representative Paul, no. Having the federal government responsible for trying to micromanage Medicare, no, trying to micromanage education, no. The federal government is not good at micromanaging anything. This is why I believe in empowering the states to do more and limit what the federal government does with regard to those kinds of program." (Cain on August 11, 2011; score: 2.1)

Sequences were selected from all such sequences longer than 500 characters and within ±0.1 to two points on the dimension: −1 and +2. Sequences in the first column characterize the negative end of the extracted dimension and sequences in the second column the positive end.

Scaling Example 2

- Fall of Radicova's gov't
- Transcripts of speeches from NRSR website
- Expected 2 dimensions: gov't support vs ESF support
- Correspondence analysis


Rafaj
Sulik
Miklos
Stefanec
Beloussovova
Fronc
Brocka
Figel
Dzurinda
Sulyanos
Fico
Hrusovsky
Radicova
Second dimension
1
0
−1
−2
−1.5
−1.0
−0.5
0.0
0.5
1.0
First dimension



takze
povedala
velke
nemecko
2
eur
portugalsko
vsetky
proti
takisto
nasi
take
ved
neviem
jedine
nas
tych
pozornost
dokonca
toho
riesenie
banky
euroval
eurovalu
krajinach
prosim
povedal
kvoli
mate
myslim
jedno
maju
rad
tieto
najma
krajiny
pripomenut
medzi
slovensko
0
poslanec
europy
potom
pani
hlasovanie
politici
potrebne
euro
niektori
unia
naopak
financii
predseda
hlasovania
republika
seba
kedy
aky
urcite
vyhlasenie
−2
krizy
svetovej
vazena
unii
europskeho
cas
rozhodnutia
casto
slovenskej
zodpovednost
politicke
voci
koalicia
politicka
situacii
vsak
−4
politicky
nadalej
−3
−2
−1
0
1
2

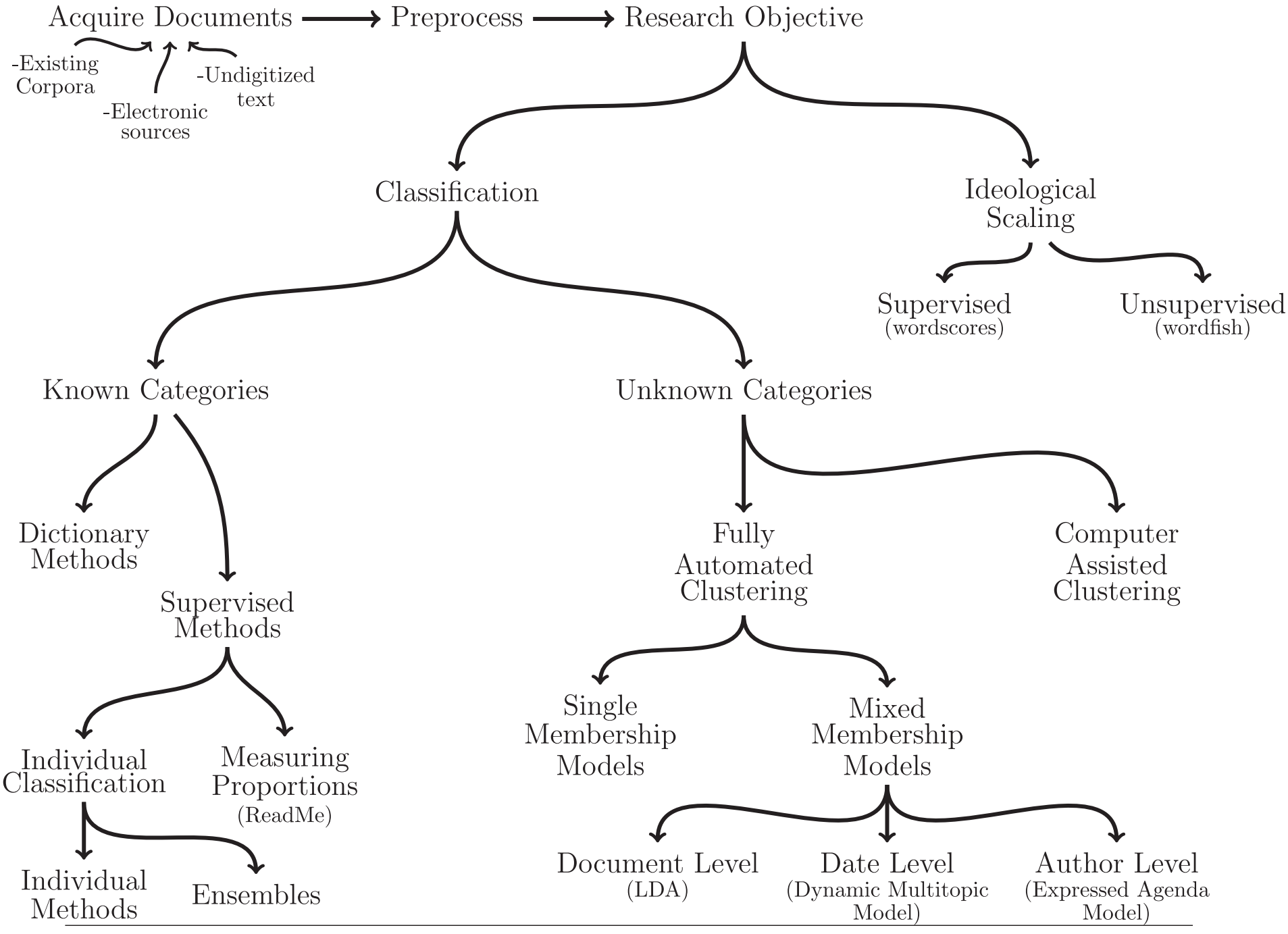


Fig. 1 An overview of text as data methods.


IRWIN
QUICK GRIP