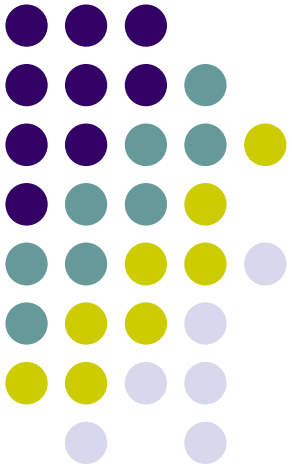
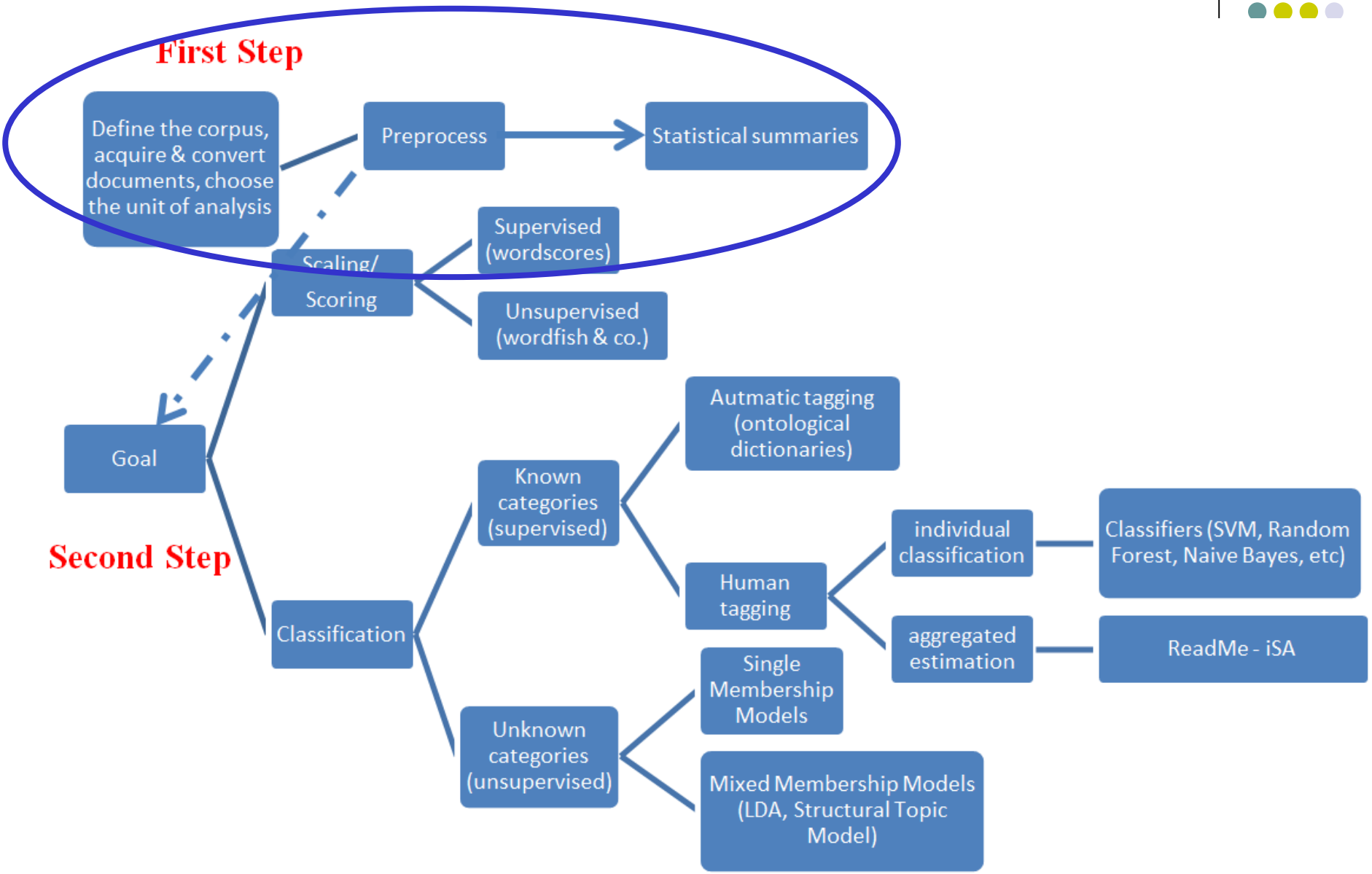
Big Data Analytics

Lecture 1 (Lab part)



Our Course Map

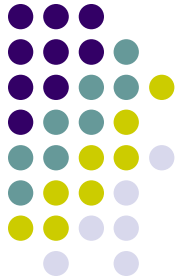


Statistical summaries

Once you have your DfM, you can start by running some statistical summaries

Statistical summary methods are essentially quantitative summaries of texts to describe their characteristics on some indicator, and may use (or not) statistical methods based on sampling theory for comparison

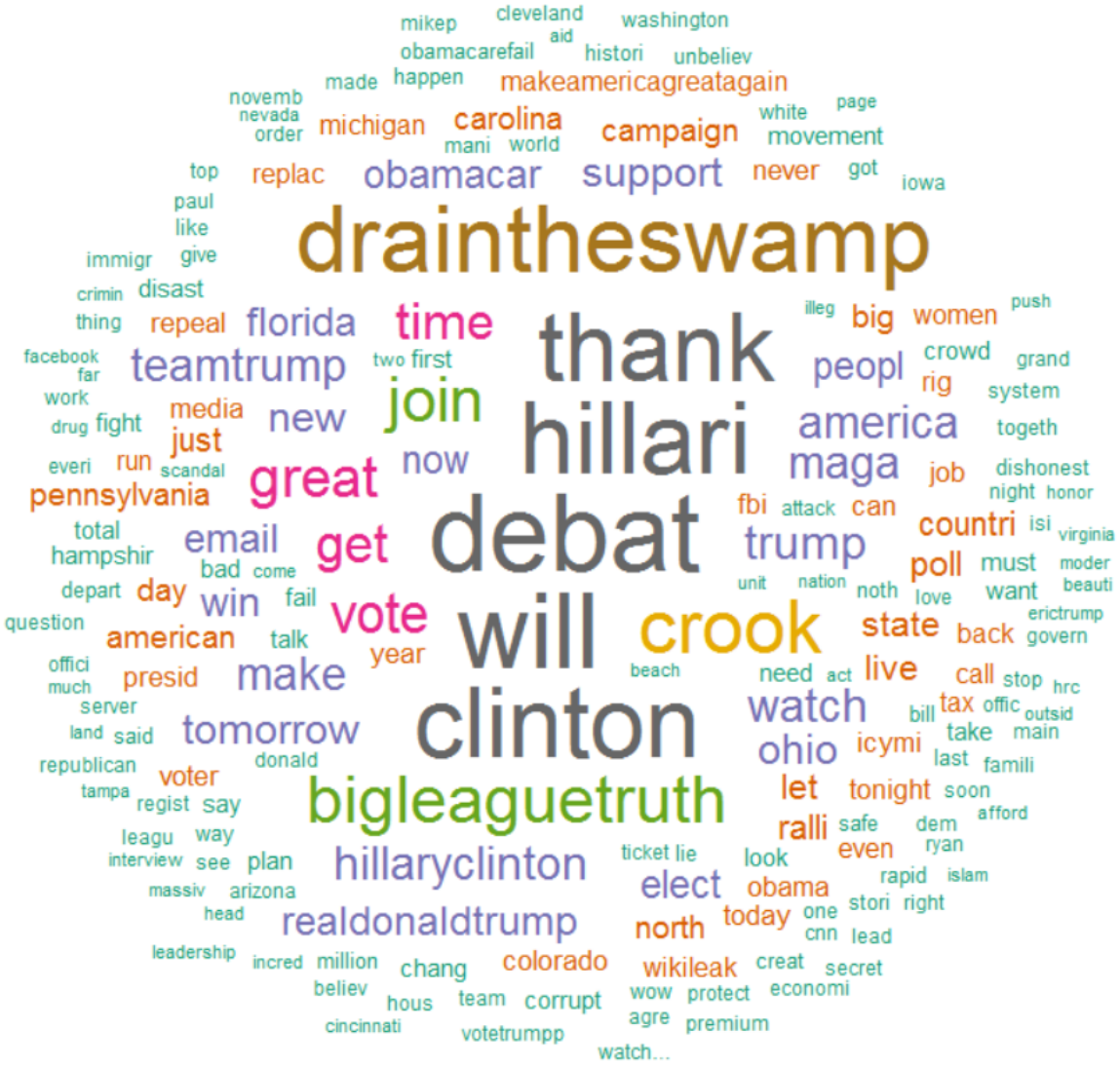
Statistical summaries (1)



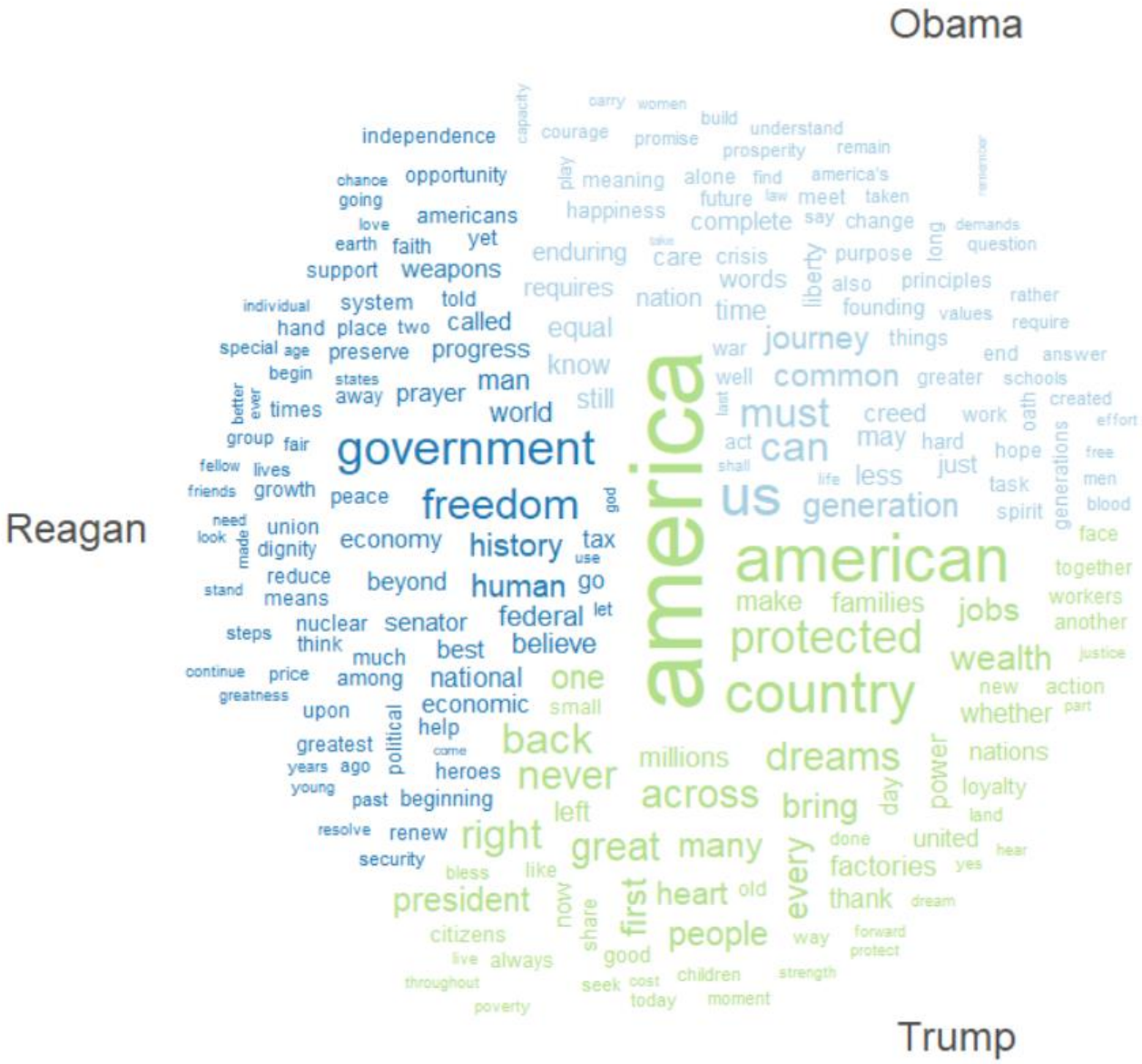
The simplest such measures identify the most commonly occurring words, and summarize these as frequency distributions

For example: **tag clouds**! A tag cloud is a visual representation of text data, in which tags are single words whose frequency is shown with different font size (and/or color)

Tag-cloud of the tweets posted in @realDonaldTrump during the last 2 months of the electoral campaign



Comparing tag-clouds! The US Presidential inaugural speeches example



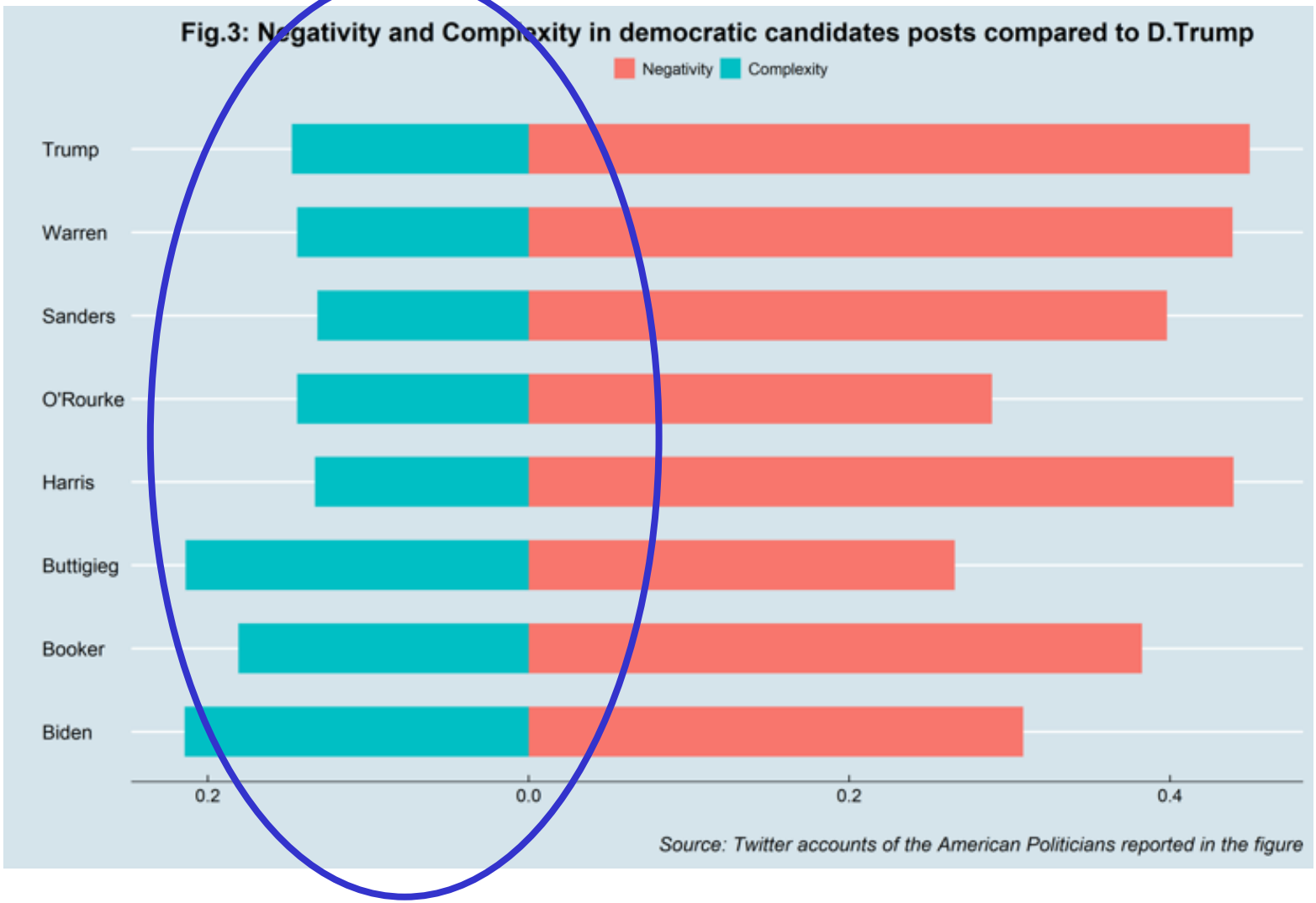
Statistical summaries (2)

Other quantitative summary measures of documents are designed to characterize specific qualities of texts

Comparing the **rates of types and tokens** forms the foundation for measures of **lexical diversity** (the rate of vocabulary usage), with most common such measure comparing the number of types to the number of tokens (the “type-token ratio”)

For example, it is argued that populist communication means simplified political discourse (lower diversity), in an attempt to reach the public more easily

So different, yet so alike (to Donald Trump?)



Statistical summaries (3)

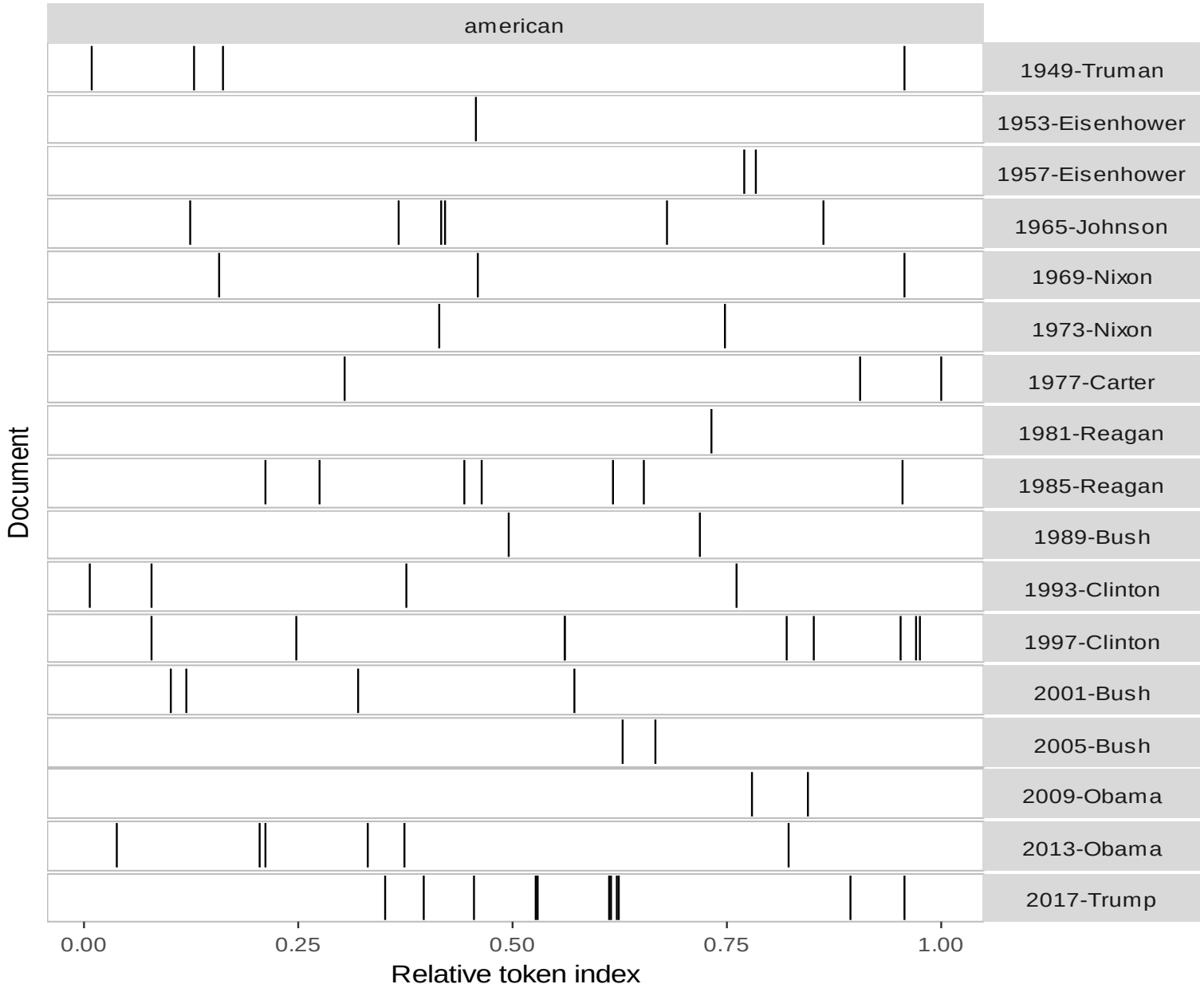
More sophisticated methods **compare the differential occurrences of words across texts** or partitions of a corpus, using statistical association measures, to identify the **words that belong primarily to sub-groups** such as those predominantly associated with male- versus female - authored documents, or Democratic versus Republican speeches

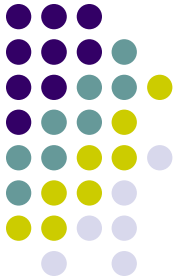
Statistical summaries (4)

Interesting descriptive statistics can also be produced directly by working with the corpus, rather than with the DtM

This allows us to retain the original text sequence, and therefore, for example, to detect both **the relative frequency of an employed word across documents as well as the "timing" of that word** via a *Lexical dispersion plot*

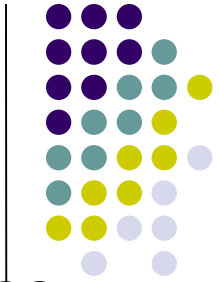
Inaugural Speeches by US Presidents





IMPORTANT!!!

Before using rtweet



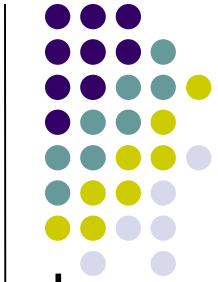
We will use since the next week the `rtweet` package: so start to install it!

install.packages("rtweet", repos='http://cran.us.r-project.org')

install.packages("httpuv", repos='http://cran.us.r-project.org')

install.packages("ggmap", repos='http://cran.us.r-project.org')

Before using rtweet



Then open an R session and type the following commands. Plz let me know if you are able (or not) to download the 10 tweets:

```
library(rtweet)

library(httpuv)

rt <- search_tweets( "#rstats", n = 10,
include_rts = FALSE)

print(rt$text[1:10])
```

Optional

Before we can start geocoding data, we need to obtain an API key from Google. Go to the registration page, and follow the instructions (select all mapping options)

The **geocoding API** is a free service, but you nevertheless need to associate a credit card with the account.

Please note that the Google Maps API is not a free service. There is a free allowance of 40,000 calls to the geocoding API per month, and beyond that calls are $0.005 each

This implies that basically you have a monthly free limit of $200 (more than enough…)

To register you need to have: a) a gmail account; b) a credit card

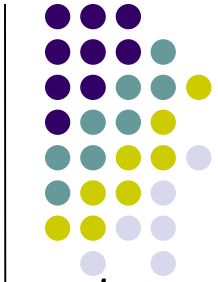
Optional

After you finish the registration (if everything hopefully works fine!) Google gives you back an API number. Save it!

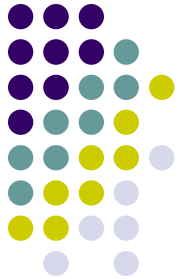
Then type:

```
library(ggmap)
register_google(key = "NUMBER OF YOUR GOOGLE API!")
geocode(c("White House", "Uluru"))
You should get this result back:
# A tibble: 2 x 2
    lon   lat
  <dbl> <dbl>
1 -77.0  38.9
2 131.  -25.3
```

Optional



If you are able to get the Google API, but GGMAP does not get any results back, enable the “geocoding app” in your console developer. Check how to enable GOOGLE API here



Please check that everything is ok with rtweet before next Tuesday!