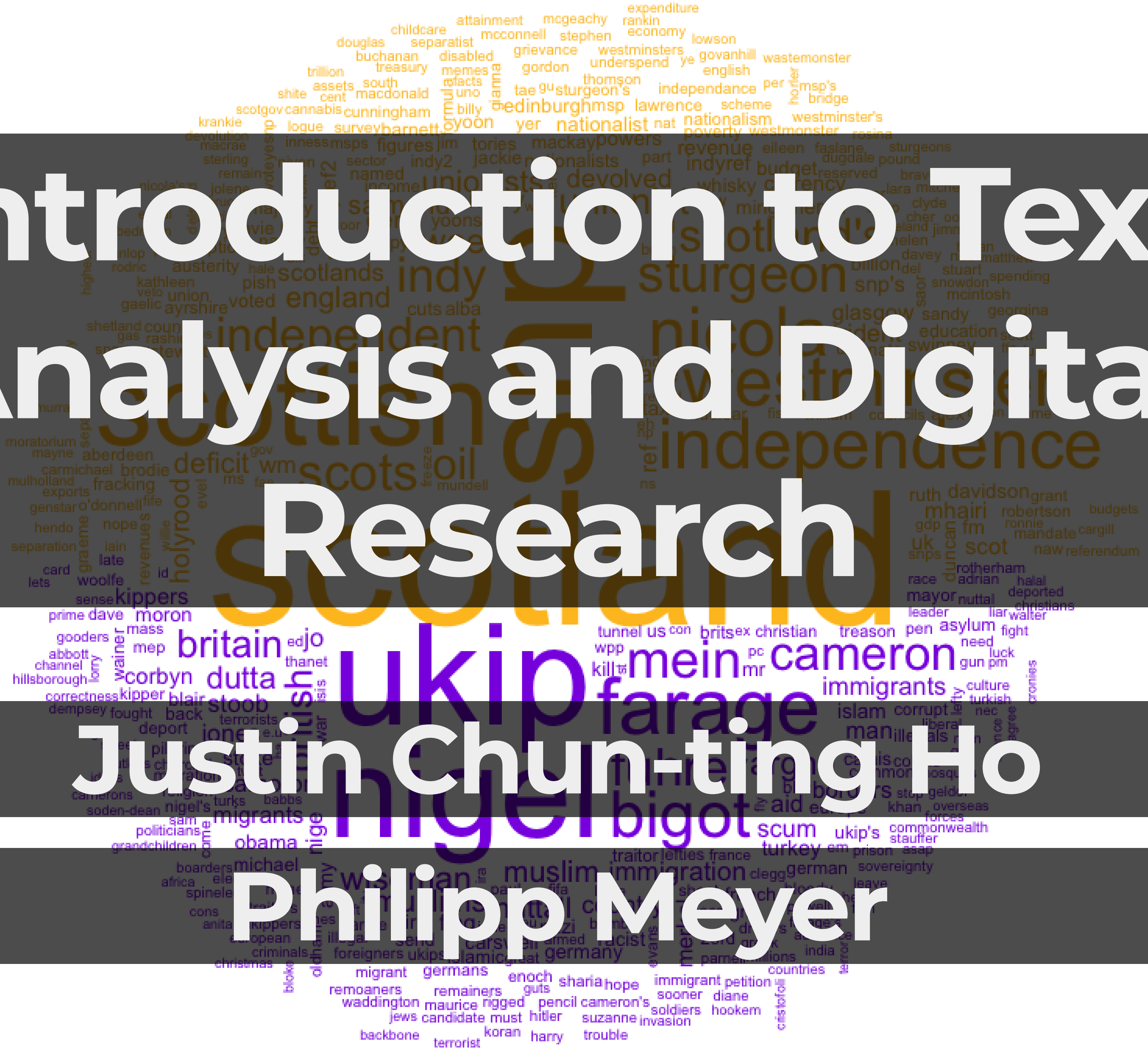


Introduction to Text Analysis and Digital Research

Justin Chun-ting Ho

Philipp Meyer

Overview of the Course

Typical Schedule

Time	Task
First Hour	Lecture & Examples
Second Hour	In-class Exercises
Third Hour	Your Dataset

Preparations

Required:

- R
- R Studio

Recommended:

- Firefox
- LibreOffice
- ATOM

Web and Social Media Scraping

App:

- Netvizz
- Facepager

R:

- Twitter
- Guardian
- Imdb
- Press Release

Time for R

Tips for Web Scraping

Tip 1: Download Once, Load Many Times

Tips for Web Scraping

Tip 2: Selector Gadget

Example

Tips for Web Scraping

Tip 3: Split the HTML and process each section

Tips for Web Scraping

Tip 4: Always start at Page 2

Tips for Web Scraping

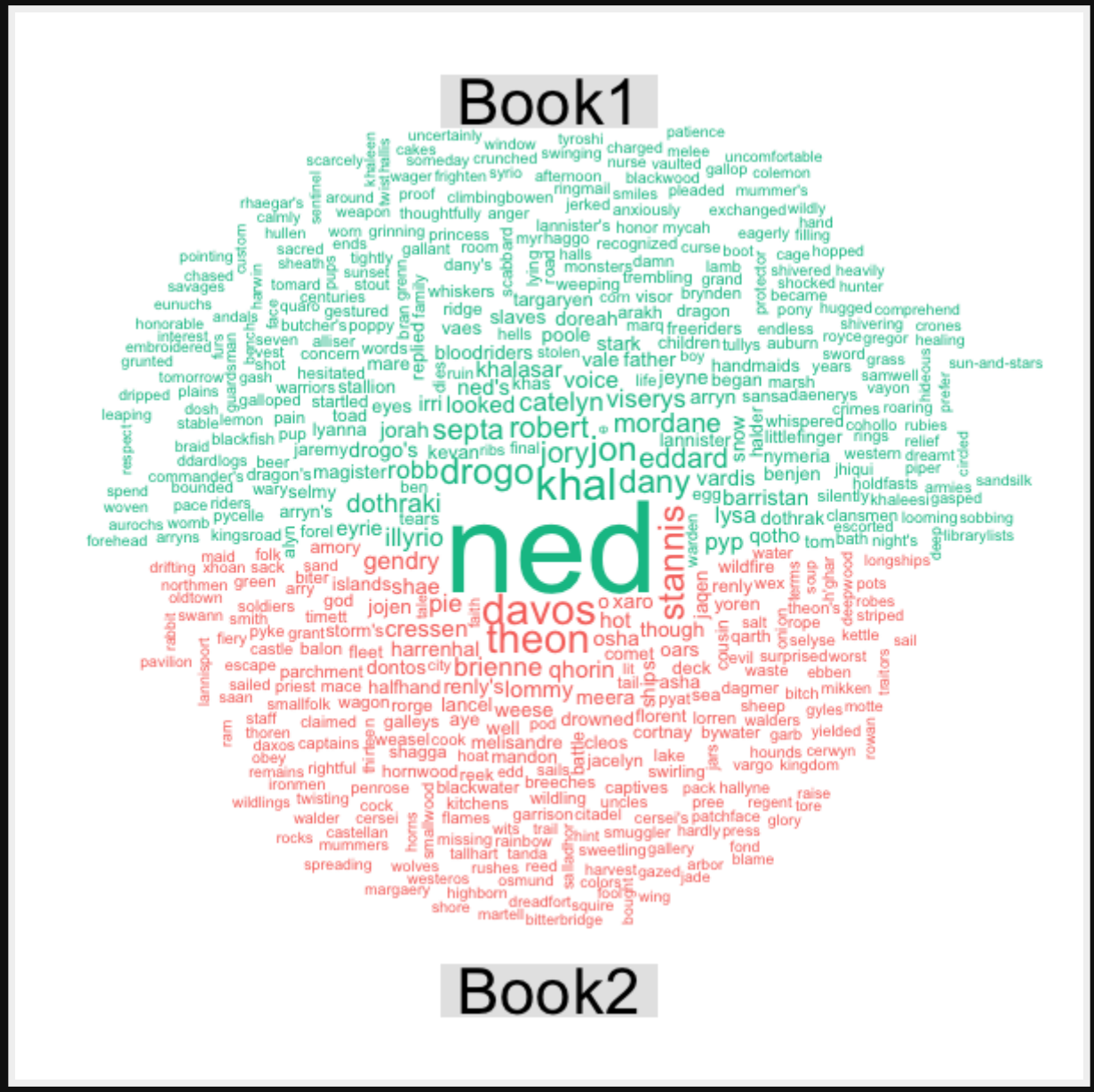
Tip 5: Sometimes, you can "hack" the website

Tips for Web Scraping

Tip 6: Respect the rules and limits of the websites

Text Analysis: The Basics

What is Text Analysis?



“Systematic, objective, quantitative analysis of message characteristics"

–Neuendorf, The Content Analysis Guidebook

Types of Text Analysis

Degree of human involvement:

- Human coding (100%)
- Supervised
- Unsupervised (0%)

Type of output:

- Scaling
- Classification

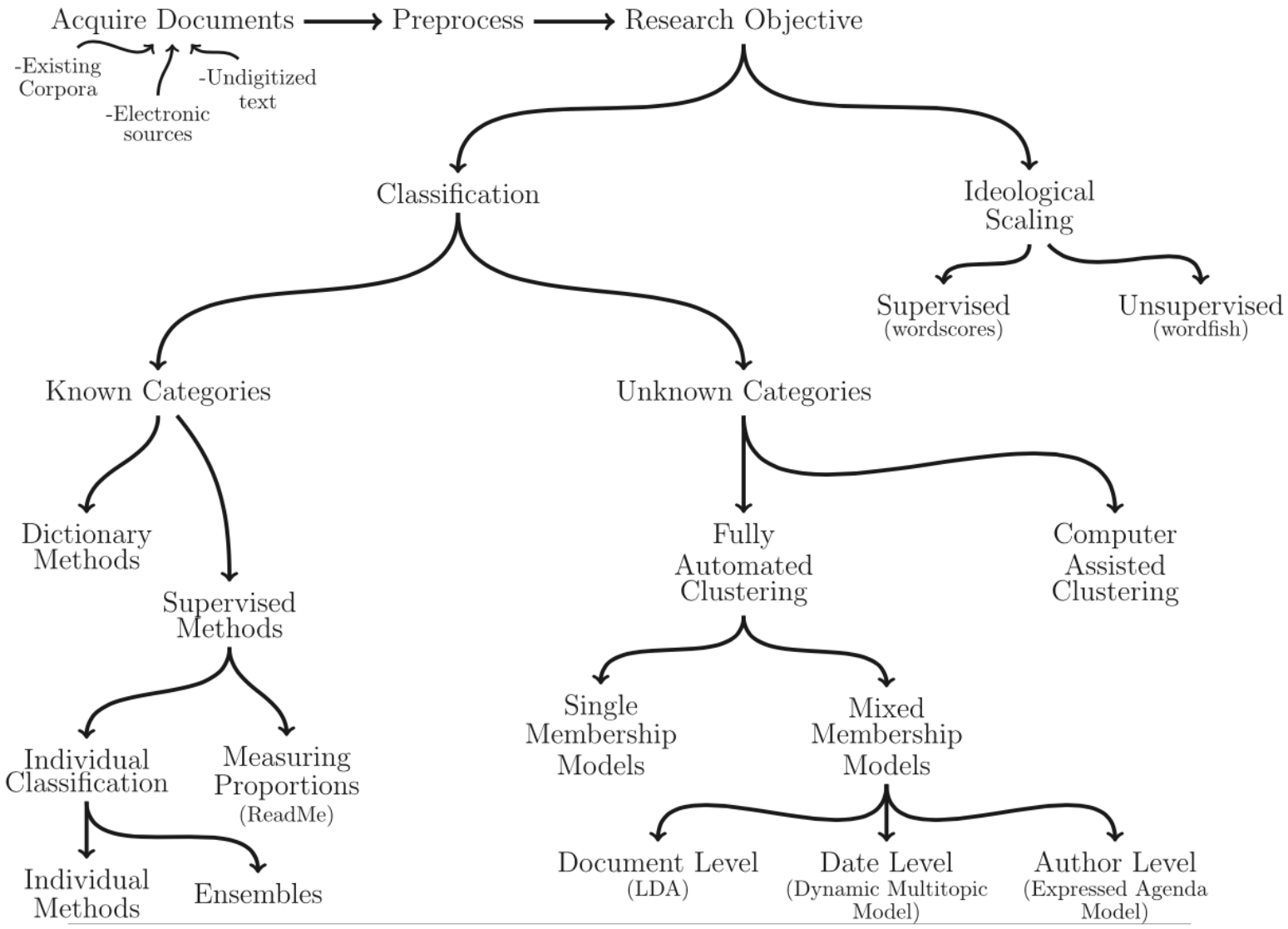


Fig. 1 An overview of text as data methods.

4 Principles of Text Analysis

1. All Quantitative Models of Language Are Wrong—But Some Are Useful
2. Quantitative Methods Augment Humans, Not Replace Them
3. There Is No Globally Best Method for Automated Text Analysis
4. Validate, Validate, Validate

Read: Grimmer, J. and Steward, B. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267-297. doi:10.1093/pan/mps028

1. All Quantitative Models of Language Are Wrong—But Some Are Useful

- Data generation process for any text is a mystery
- All methods necessarily fail to provide an accurate account of the data-generating process
- Meanings change drastically: “Time flies like an arrow. Fruit flies like a banana.”

2. Quantitative Methods Augment Humans, Not Replace Them

- Text Analysis will not eliminate the need for careful thought nor remove the necessity of reading
- Rather than replace humans, computers amplify human abilities

3. There Is No Globally Best Method for Automated Text Analysis

- Different research questions and designs need different models
- The same model will perform well on some data sets, but poorly when applied to other

4. Validate, Validate, Validate

- The output of the models may be misleading or simply wrong
- Supervised methods: able to reliably replicate human coding
- Unsupervised methods: the measures are as conceptually valid

Bag of Words Assumption

- **Word order doesn’t matter**
- **The followings mean exactly the same:**

I enjoy eating food and being with my family
I enjoy eating my family and being with food
and being eating enjoy family food I my with

The Output

End Product

Document-Feature Matrix

	word1	word2	word3	...	wordN
doc1	0	2	0		3
doc2	1	0	0		0
doc3	0	0	2		1
...					
docN	0	1	0		0

End Product

Meta-data Martix

	var1	var2	var3	...	varN
doc1	Y	1	0		3
doc2	Y	0	0		2
doc3	N	0	0		1
...					
docN	Y	0	0		3

The Process

Original Text

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 181
##
## Article 1. All human beings are born free and equal in dignity
```

Remove Punctuation

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 178
##
## Article 1 All human beings are born free and equal in dignity
```

To Lower Case

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 178
##
## article 1 all human beings are born free and equal in dignity
```

Remove Numbers

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 177
##
## article  all human beings are born free and equal in dignity a
```

Remove Stopwords

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 135
##
## article   human beings  born free  equal  dignity  rights   en
```

Stemming

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 108
##
## articl human be born free equal digniti right endow reason con
```

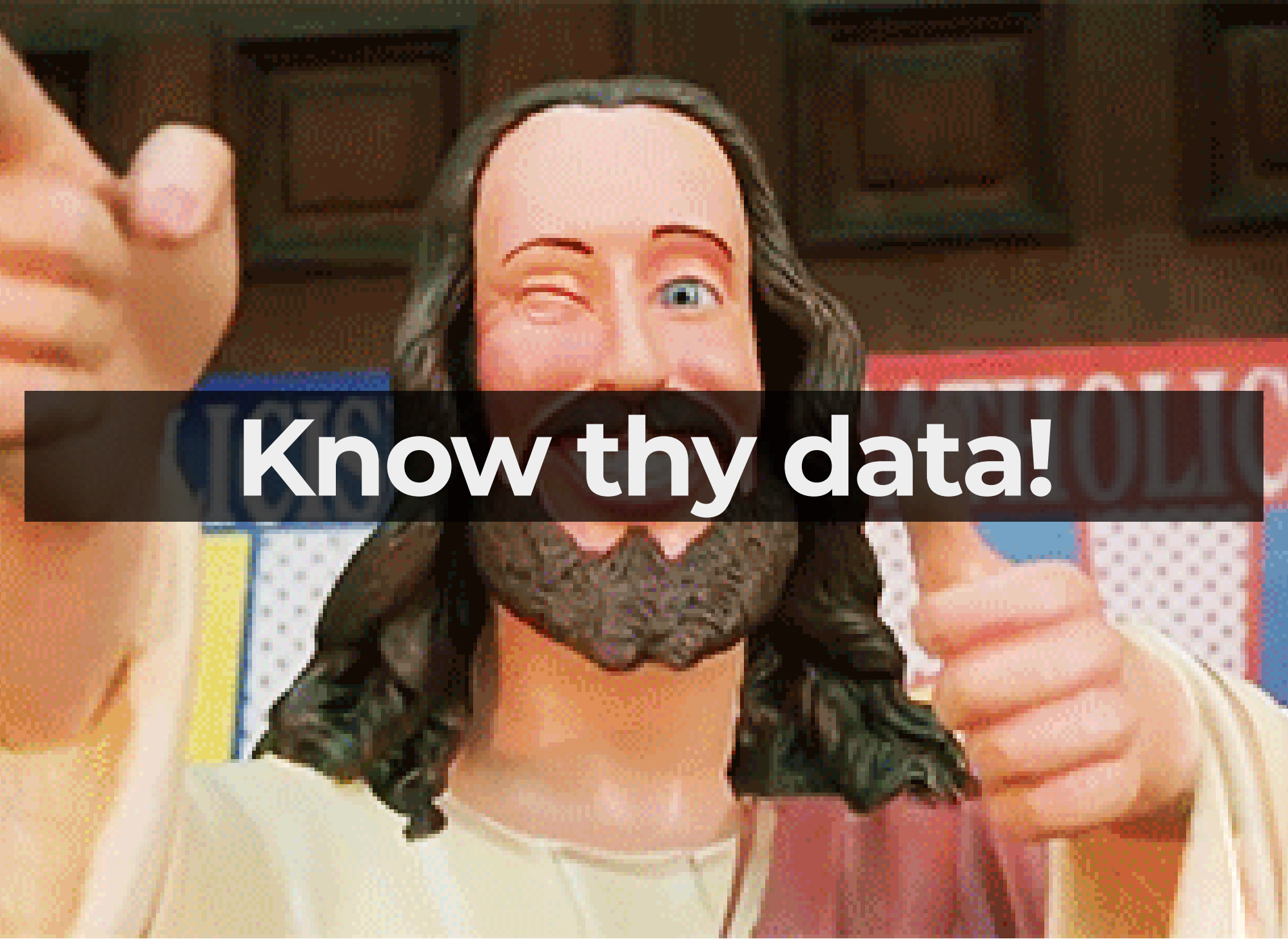
Create DFM

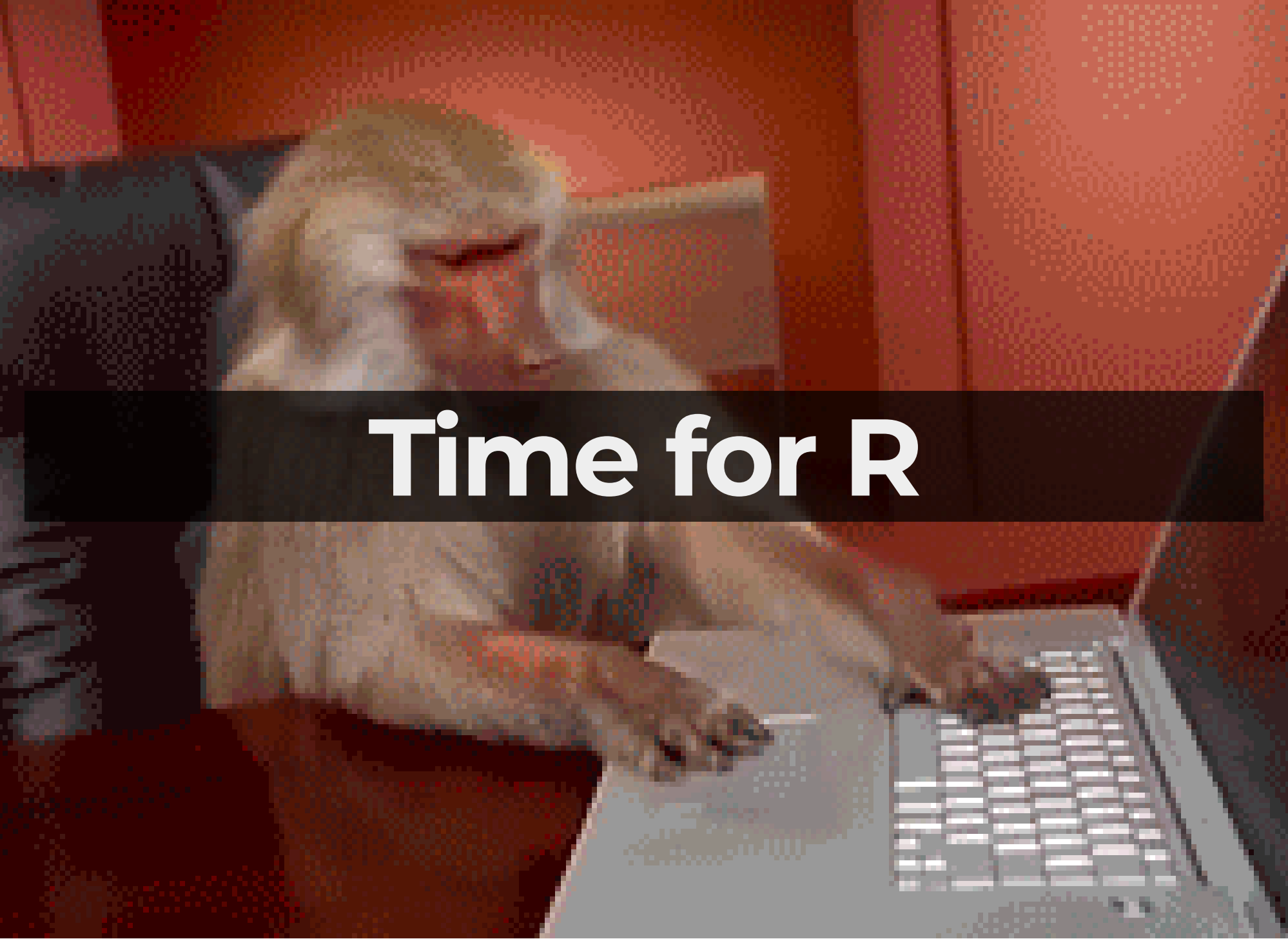
```
##      Terms
## Docs act anoth articl born brotherhood conscienc digniti endow
##    1   1     1      1    1           1         1       1     1
```

Optional: Create N-Gram

```
##
## article_1 1_all all_human human_beings being_are are_born born
```

Take home message


Know thy data!


Time for R

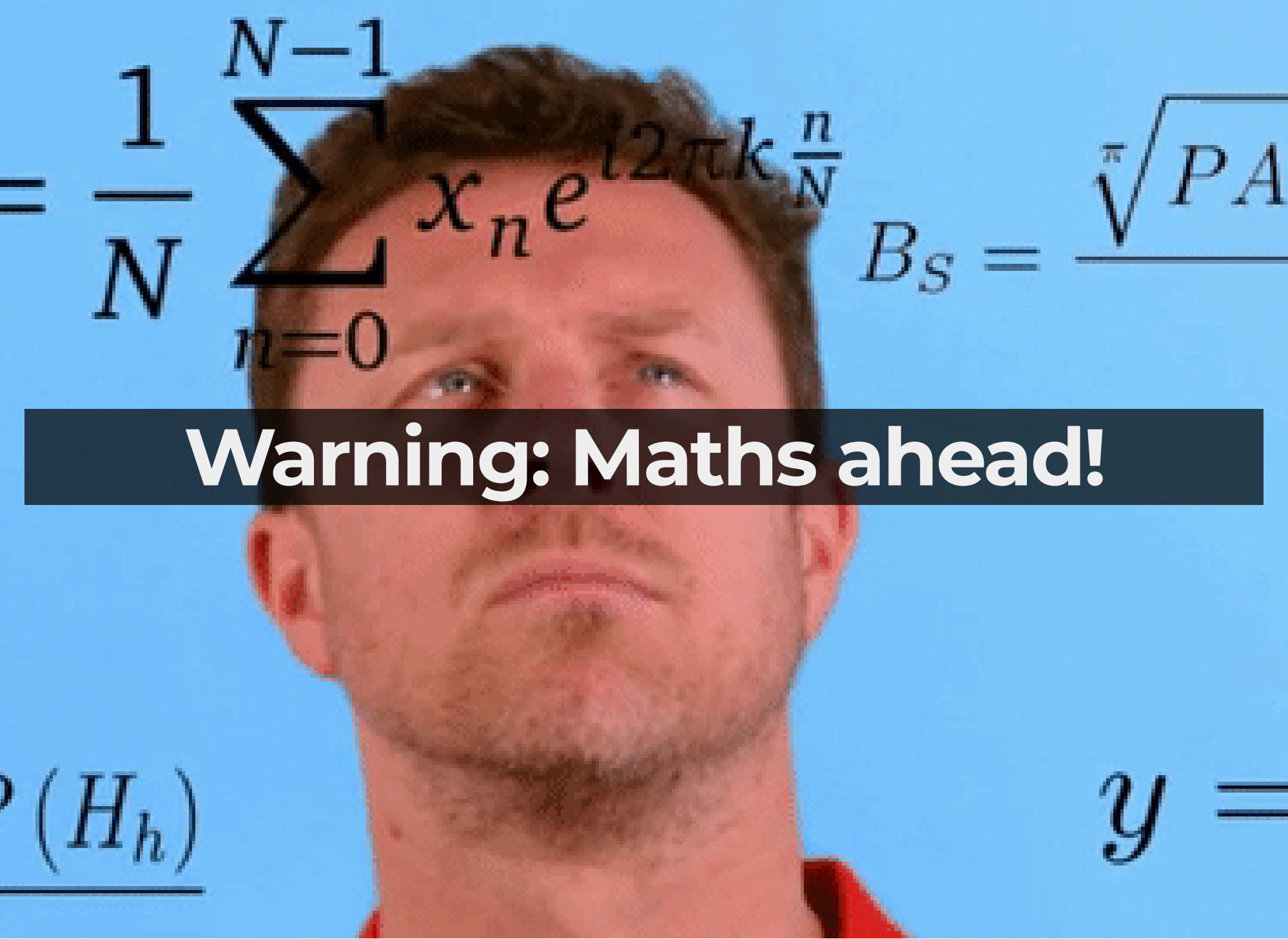
Keyword Analysis

Keyword Analysis

What is a Keyword?

“A keyword may be defined as a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind”

(Scott, M. (1997). PC analysis of key words - and key key words. System, 25(2), 233-45.)


Warning: Maths ahead!

Keyword Analysis

What is Keyness?

“The keyness of a keyword represents the value of log-likelihood or Chi-square statistics; in other words it provides an indicator of a keyword’s importance as a content descriptor for the appeal.”

(Scott, M. (1997). PC analysis of key words - and key key words. System, 25(2), 233-45.)

Keyword Analysis

What is a Chi-squared test?

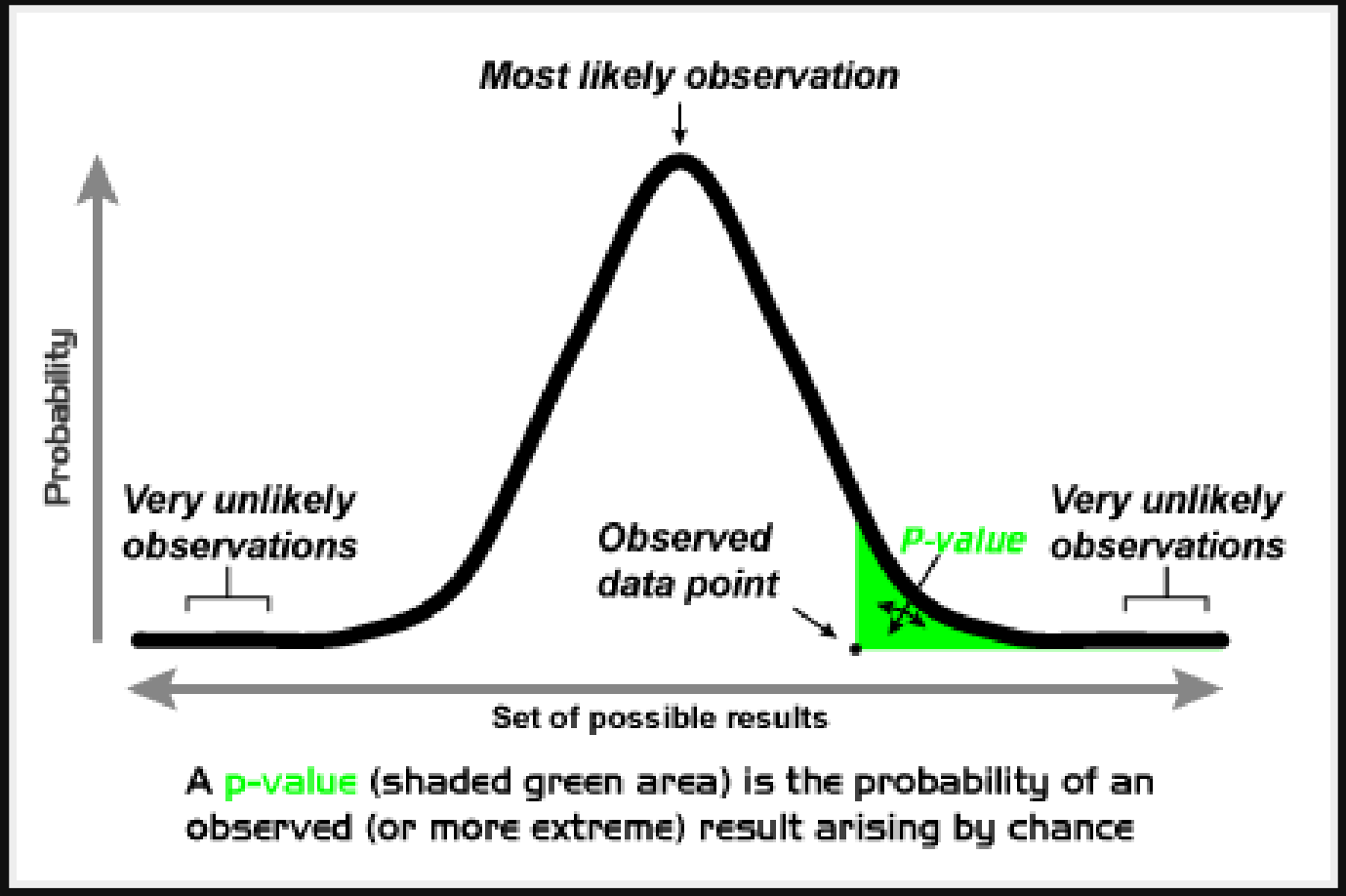
Comparison between the observed frequency and expected frequency.

	love	All other words	Total
Male	414	1714029	1714443
Female	1214	2592238	2593452
Total	1628	4306267	4307895

- Expected frequency: **Row total** times **Column total** divided by the **total number of words** in the corpus.
- Plug into this equation: $$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$
- Comparison between the observed frequency and expected frequency of the word in question and all other words.

Keyword Analysis

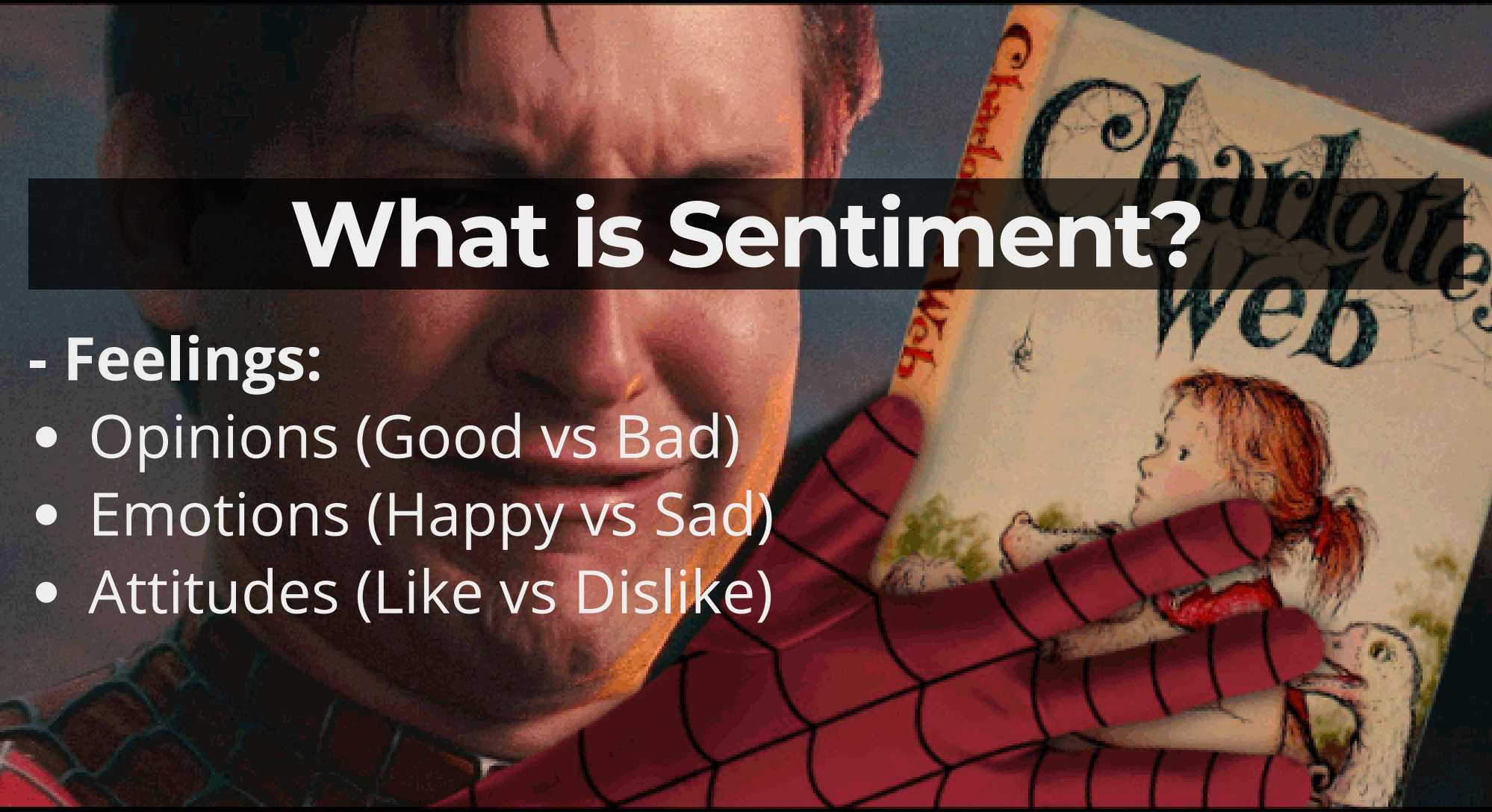
What is p-value?



Sentiment Analysis

What is Sentiment Analysis?

Sentiment analysis, also called *opinion mining*, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.



What is Sentiment?

- Feelings:

- Opinions (Good vs Bad)
- Emotions (Happy vs Sad)
- Attitudes (Like vs Dislike)

Levels of Analysis

- Document level
- Sentence level
- Entity and Aspect level

Examples of Usage

- Product Review
- Public Opinion
- Voters Support
- Wellbeing/Mental Health

Lexicons

- "Dictionary" for sentiment

- Popular Lexicons:

- LIWC
- SentiwordNet (postive to negative from +1 to -1, neutrality)
- AFINN (postive to negative from +5 to -5)
- Bing (postive, netgative)
- NRC (postive, netgative, anger, anticipation, disgust, fear, joy, sadness, surprise, trust)

Challenges

1. Opposite orientations in different applications.

“This camera sucks.” vs
“This vacuum cleaner really sucks.”

Challenges

2. Sentence containing sentiment words may not express any sentiment.

“If I can find a good camera in the shop, I will buy it.”

Challenges

3. Sarcastism

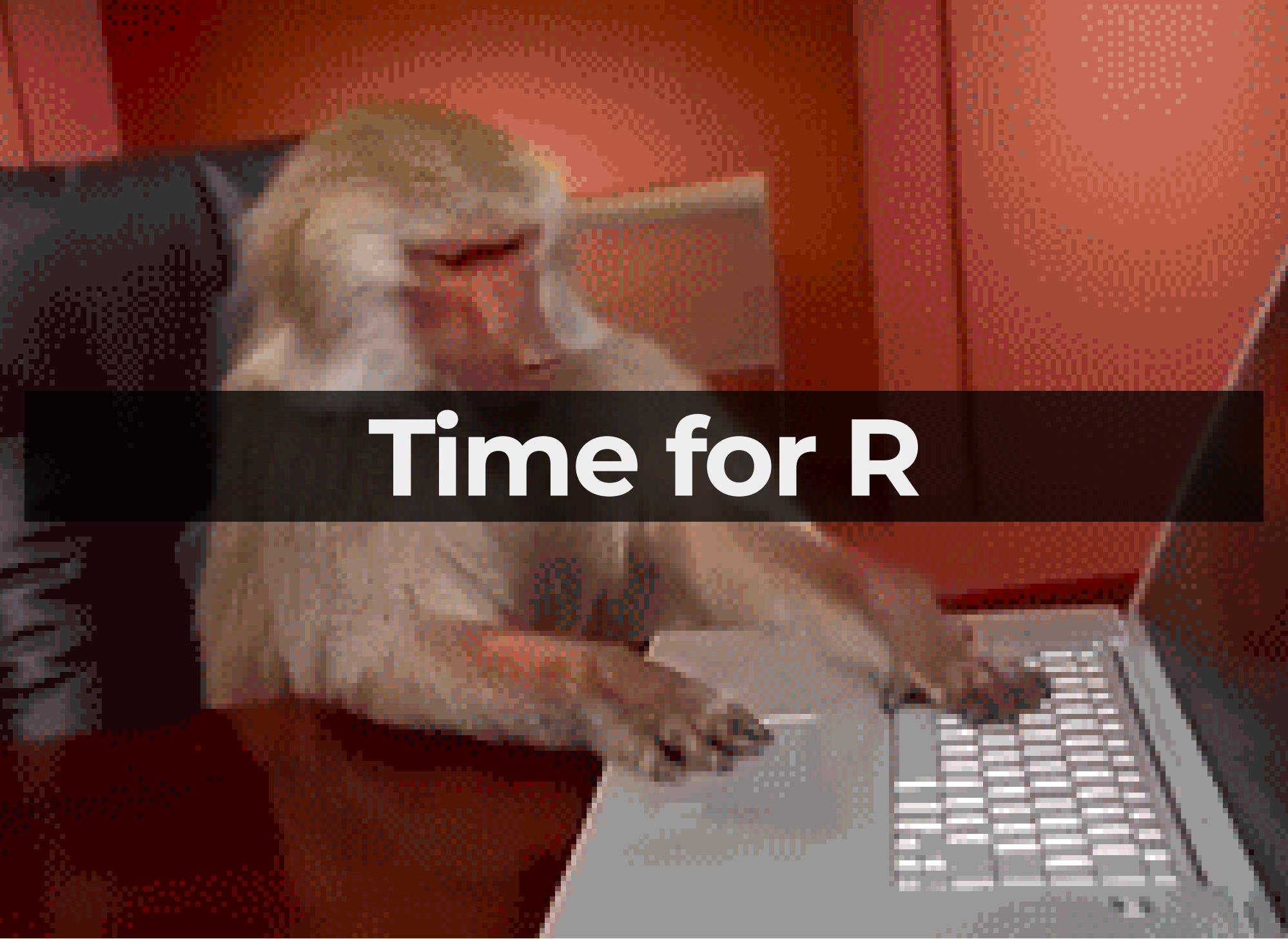
“What a Genius! You uploaded your passwords to Github!”

Note: Very common in political discussion, especially on social media.

Challenges

4. Sentences without sentiment words can also imply opinions.

“This car burns a lot of fuel.”


Time for R

Topic Modeling

What is Topic Modeling?

"Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents."

Common Usage

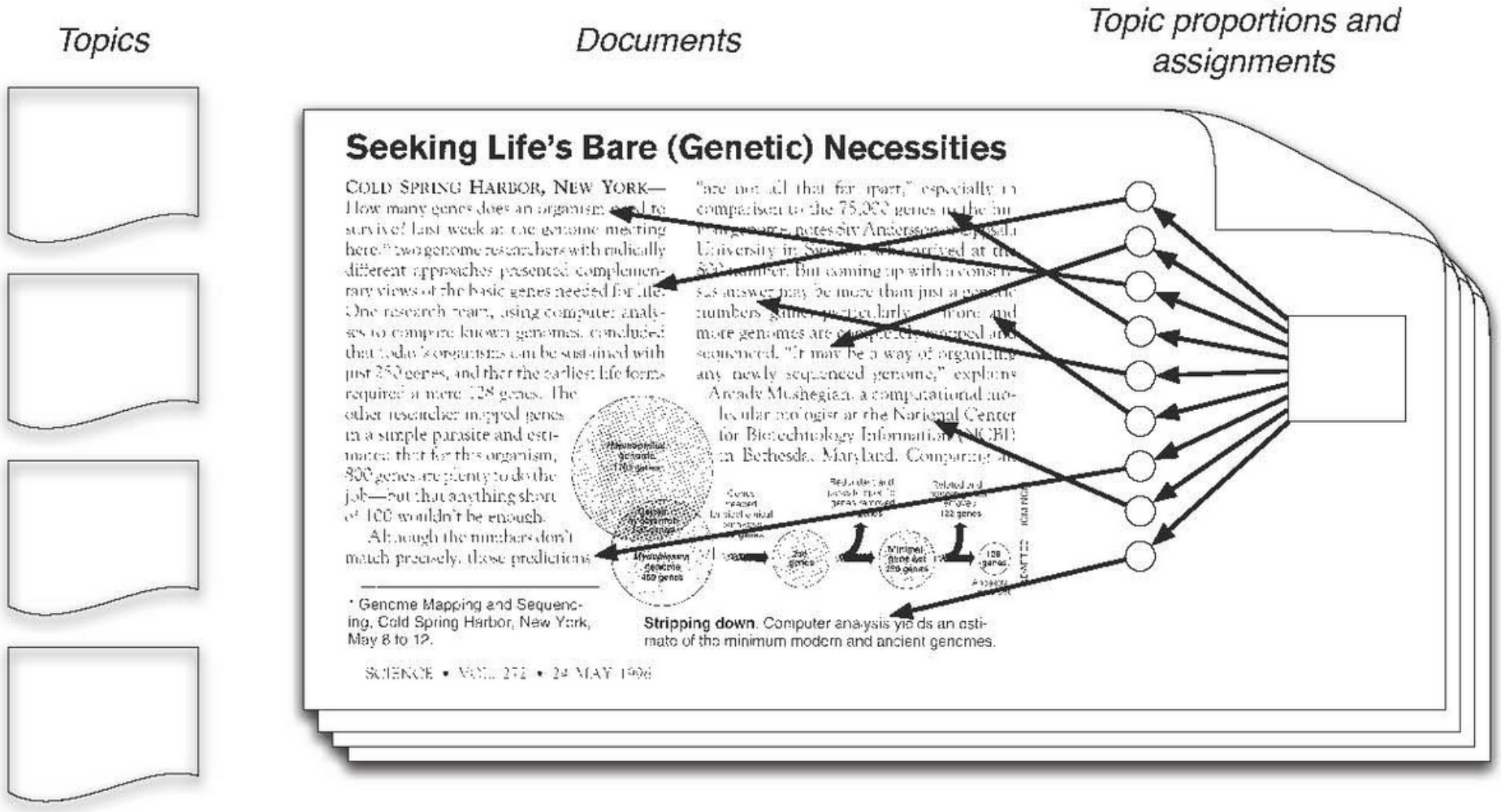
- Information Retrieval
- Recommender Systems
- Exploration of text collections
- Answering Research questions

Data

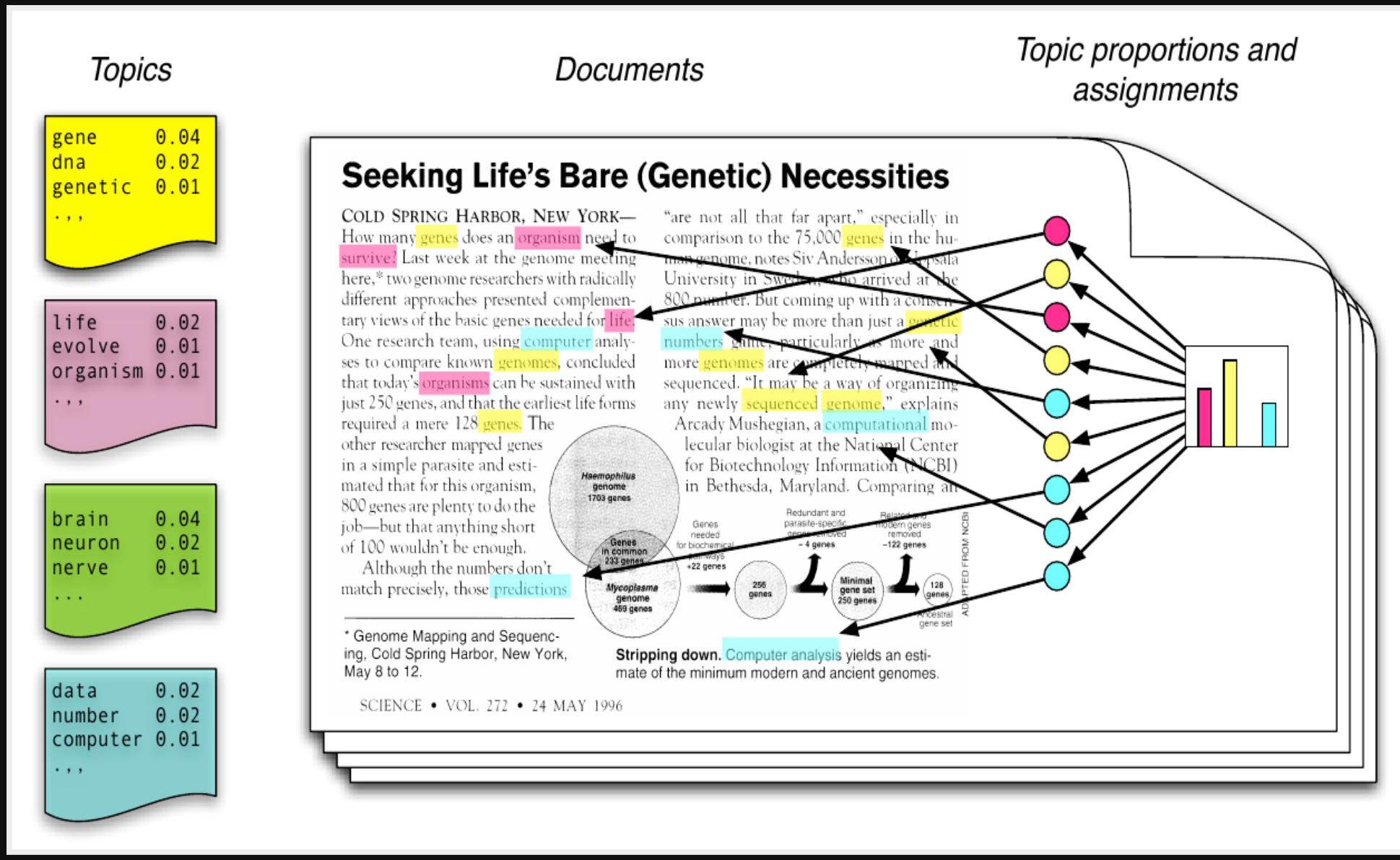
- (Large)Text Corpus
- Genetic Data
- Images

Example

How is it done?



(David Blei, 2012)



(David Blei, 2012)

Two Things you Need to Know

- Each topic is made up of words of varying importance and relevance to the topic

- Each document is made up of several topics in various proportions

Technical Things you might want to Know

- A topic model is an abstract representation of all topics and documents in a collection

- A topic is a probability distribution over words

- A document is a probability distribution over topics

- The Dirichlet distribution (in LDA) describes the topic mixture distribution of the model

Latent Dirichlet Allocation (LDA):

"A generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar."

Almost there...

Assumption of how Documents are Created

- Topics are specified
- Randomly choose a distribution over topics
- For each word:

1. Randomly choose a topic
2. Randomly choose a word from the topics

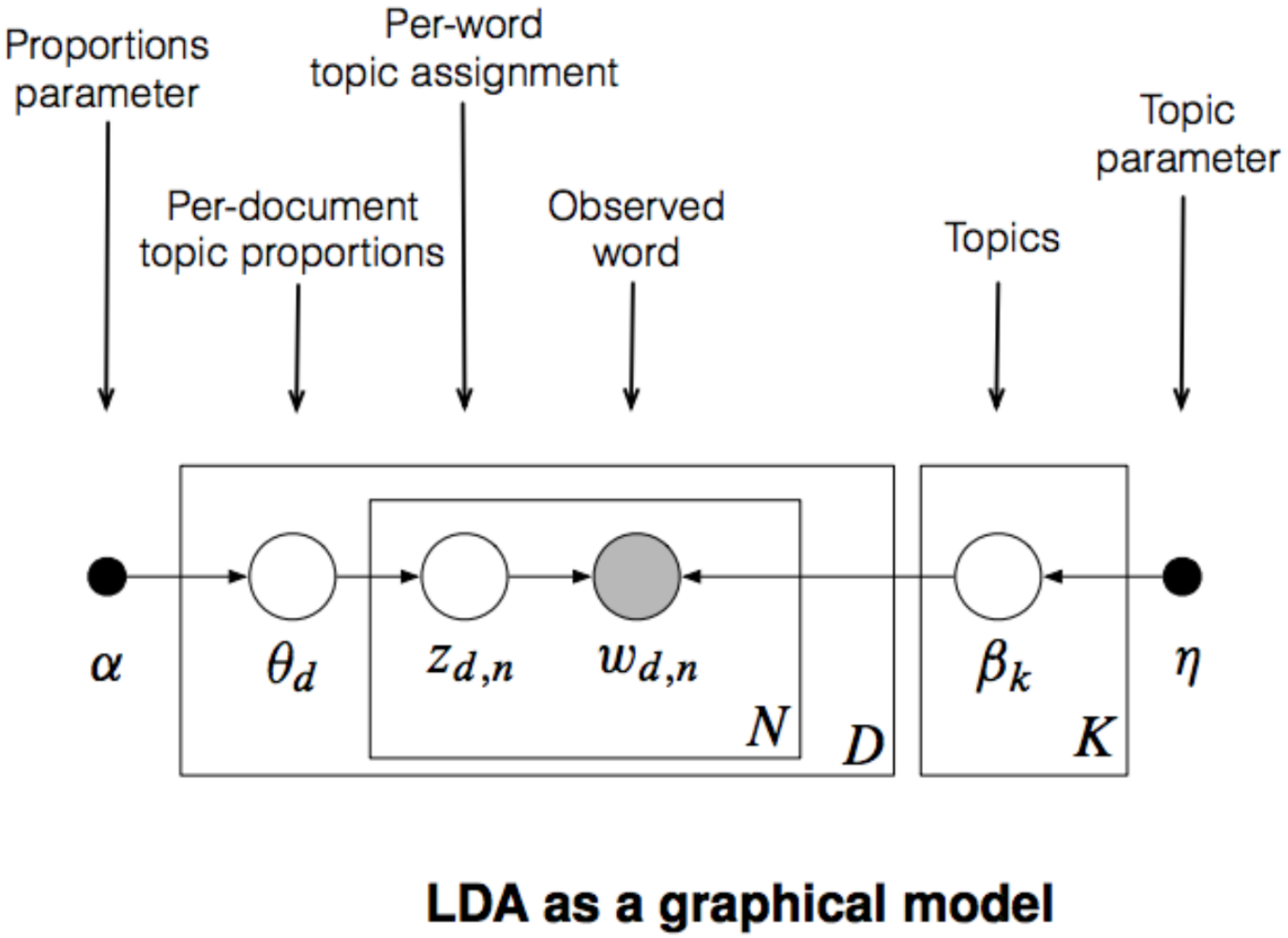
Bayesian Statistics

"The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, the conditional distribution of the hidden variables, given the documents."

(David Blei, 2012)

$p(Z, \varphi, \theta \mid w, \alpha, \beta)$

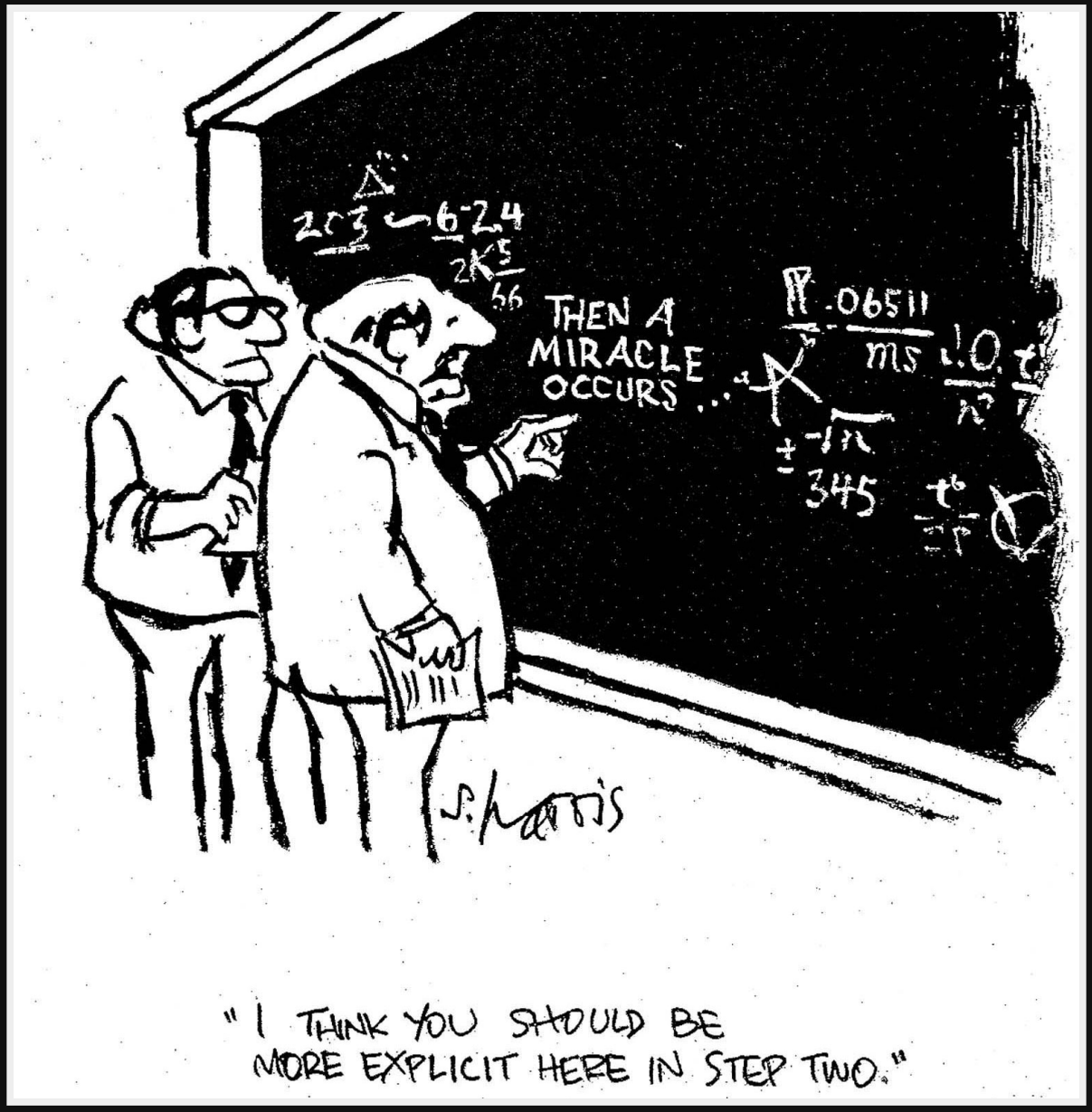
Compute the probability of:
Z = assignments of each word in each document to a topic
φ (phi) = distribution over words (for each topic)
θ (theta) = distribution over topics (for each document)
Given...
w = the data (observed words)
α = parameter of the Dirichlet prior for topics per document
β = parameter of the Dirichlet prior for words per topic



LDA as a graphical model

(David Blei, 2012)

The Process Behind...



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

The Task

What we have: Documents and the words within

What we want: Word distributions per topic, Topic distributions per document, and Topic assignment of each word

Both distributions are dependent on each other. All of them need to fit with the original documents

The Problem

For each word in each document: 1. We sample a topic from the topic distribution of that document; 2. We sample a word from the word distribution of that topic

This works only if we have the distributions. BUT WE DON'T.

The Solution: Random Initialization

For each document, we generate a random distribution over topics

For each topic, we generate a random distribution over words

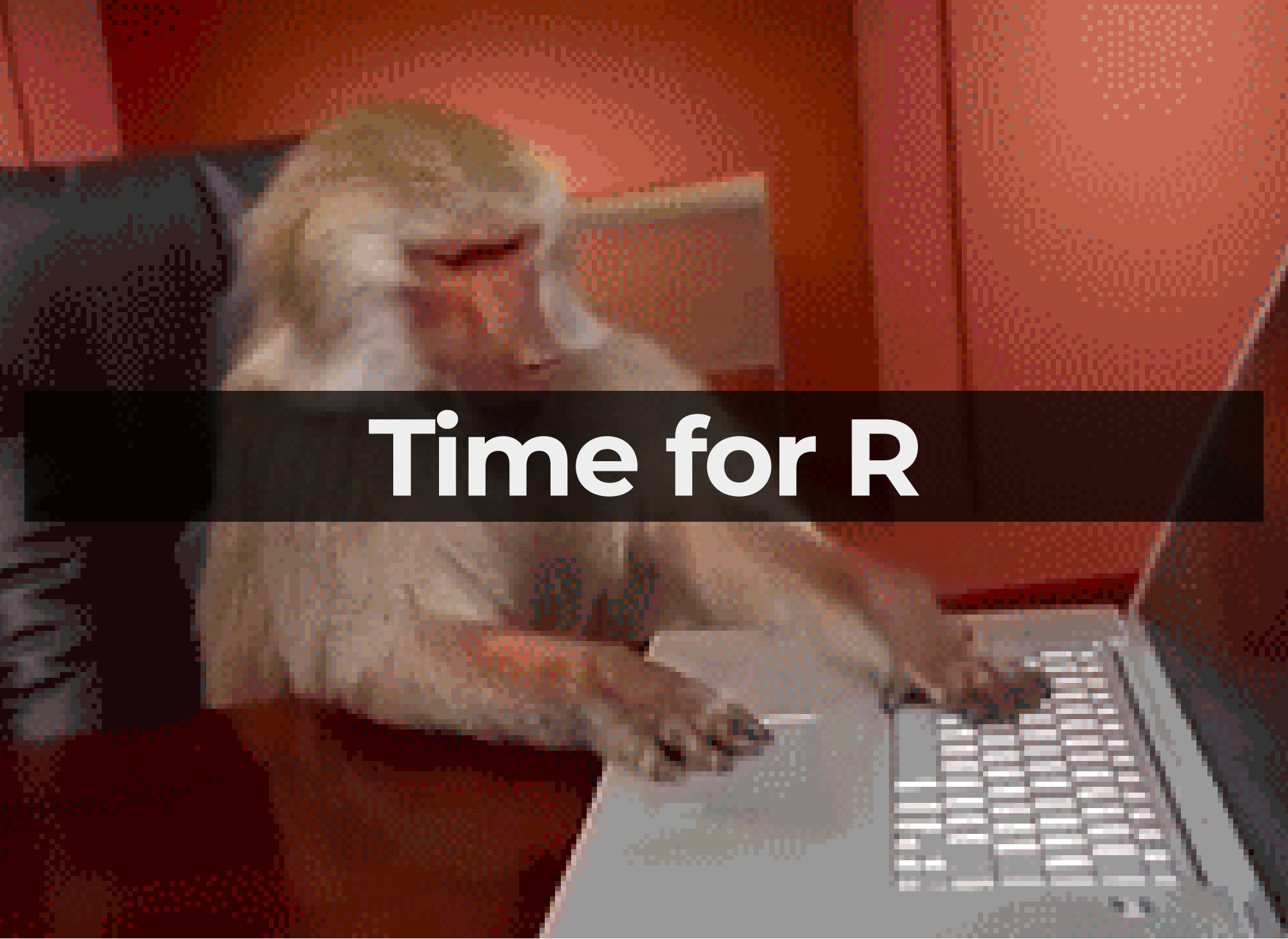
For each word in each document:
- Sample a topic from the topic distribution
- Sample a word from the word distribution of topic

Iterative Approximation

Using the observed data to improve the model

One of the methods: Gibbs sampling

- Remove the existing topic assignment of one word
- Based on the model and the other words in the document, assign a new topic to the word
- Update the overall model according to this assignment
- Repeat until convergence (you can't get better result)


Time for R

The Problem of Party Positions

"Despite the importance of party positions to the study of comparative politics, locating parties in a political space over time is a difficult task."

(Slapin and Proksch, 2008)

Expert Survey

- Expert's judgement based on various sources, various sources, including manifestos, speeches, voting patterns, and media reports

- Difficult and expensive to repeat over time and across countries
- Difficult to know whether different experts understand and answer the questions in a similar manner

Comparative Manifestos Project

In operation since 1979

Coded 2,347 party manifestos issued by 632 different parties in 52 countries over the postwar era

Link

Comparative Manifestos Project

Sampling unit: Party manifestos for election
Recording unit: quasi-sentence
Coding: Allocating to substantial category
Coding scheme: 56 cat., 7 policy domains, 13 L & R
Summarizing: Category frequencies, Scale

CMP Coding Scheme Example

305 Political Authority
408 Economic Goals
410 Productivity
605 Law and Order
606 Social Harmony

CMP Coding Scheme Example

1. Manufacturing output from Britain is back to the level of nearly twenty years ago.

2. Unemployment is still rising and there are now generations of school-leavers who no longer even hope for work.

3. Mrs. Thatcher's government stands idly by, hoping that the blind forces of the marketplace will restore the jobs and factories that its indifference has destroyed.

Therefore...

Wordscore

Goal

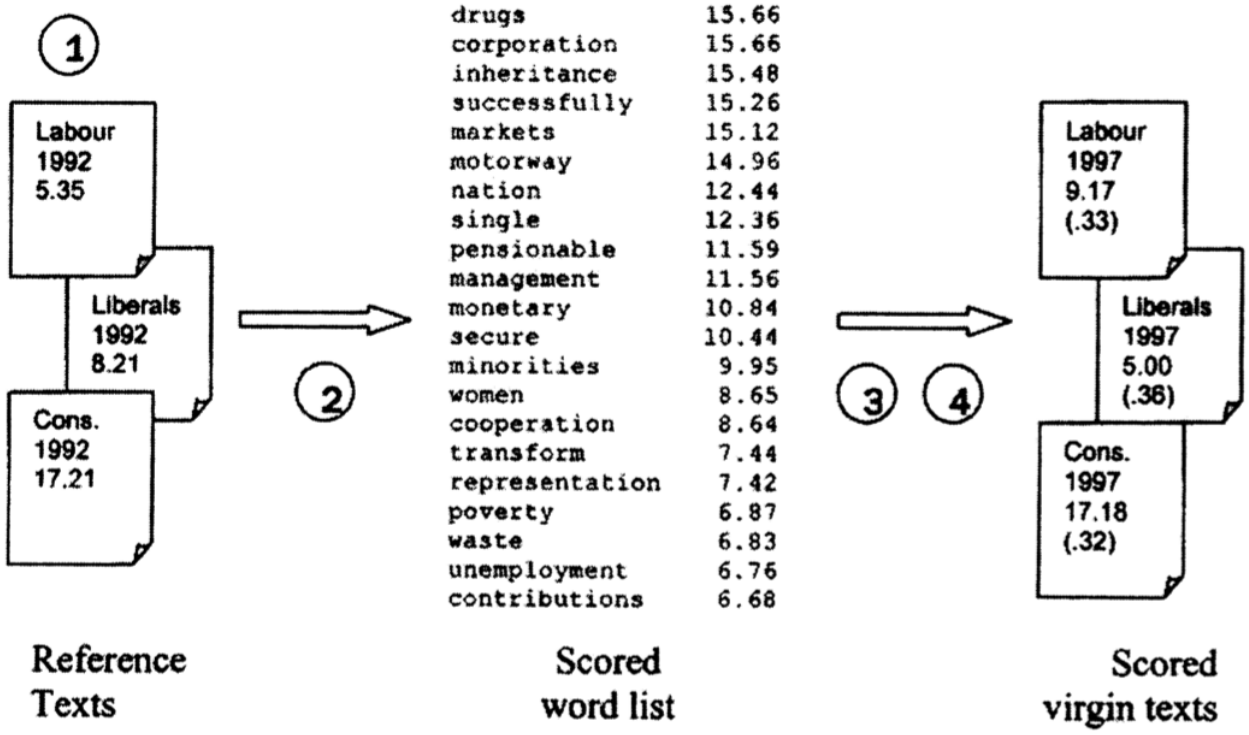
Overcoming the reliability problems of hand-coding and enabling researchers to quickly estimate policy positions for a vast number of texts

Basic Idea

Estimate policy positions by comparing two sets of texts:

- Reference texts: Documents with known policy positions (eg. from expert surveys)
- Virgin texts: Documents which we know nothing

FIGURE 1. The Wordscore procedure, using the British 1992–1997 manifesto scoring as an illustration



Step 1: Obtain reference texts with a priori known positions (`setref`)
Step 2: Generate word scores from reference texts (`wordscore`)
Step 3: Score each virgin text using word scores (`textscore`)
Step 4: (optional) Transform virgin text scores to original metric

Note: Scores for 1997 virgin texts are transformed estimated scores; parenthetical values are standard errors. The scored word list is a sample of the 5,299 total words scored from the three reference texts.

Selecting Reference Texts

1. the reference texts should use the same lexicon, in the same context, as the virgin texts being analyzed

2. Policy positions of the reference texts should span across the whole dimension

3. Set of reference texts should contain as many different words as possible

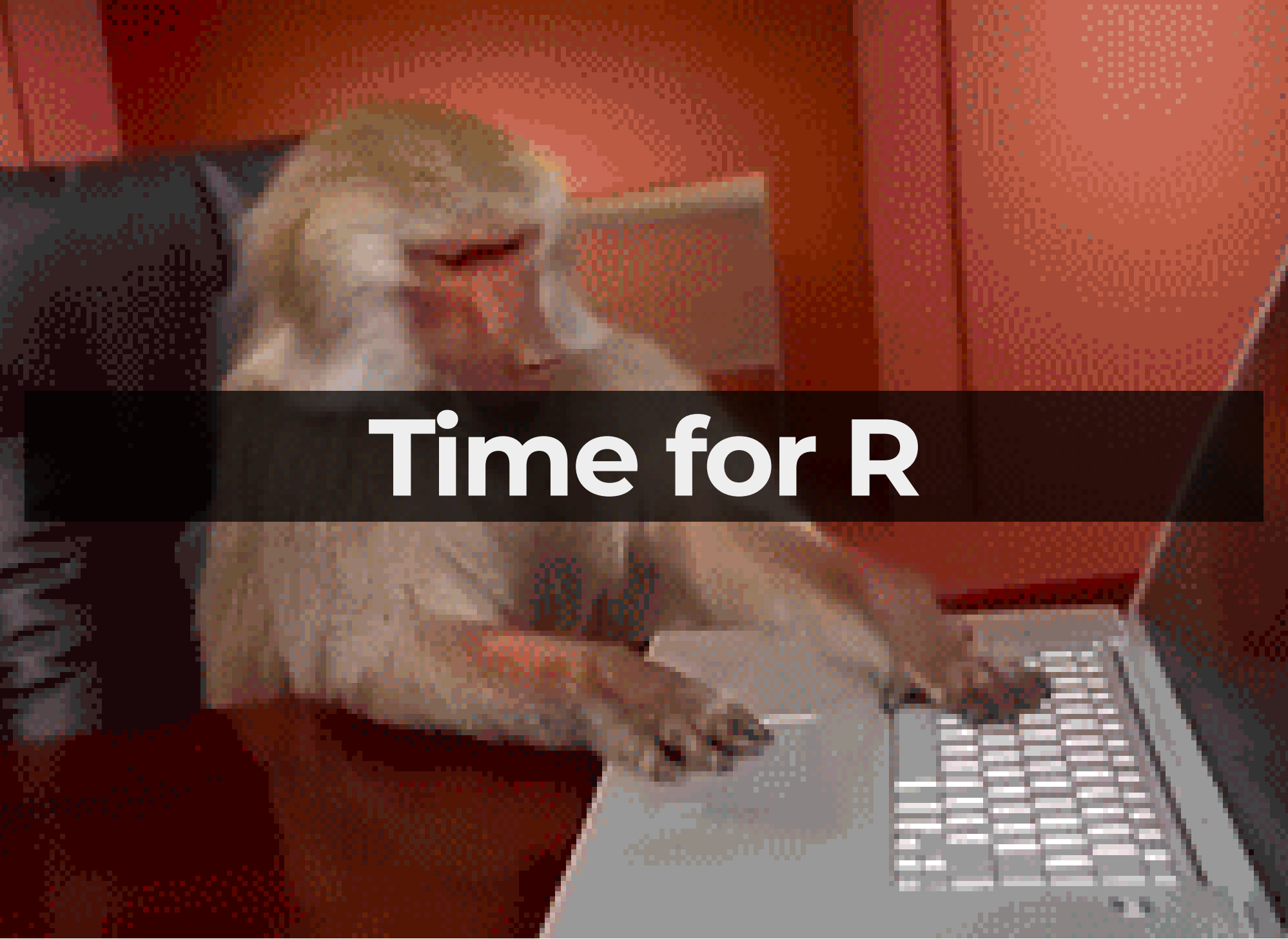
Output

Text score: Mean score of all of the scored words that a virgin text contains weighted by the frequency of the scored words

Uncertainty: Variance of all of the scored words that a virgin text contains weighted by the frequency of the scored words

Assumptions

- Policy positions are reflected in the relative frequency of word usage
- Word meaning remains stable over time
- All words carry the same weight
- All words of interest are contained in the reference texts


Time for R

Wordfish

The model:

$$y^{ijt} \approx POISSON(\lambda^{ijt})$$

y^{ijt} is the count of word j in document i's at time t

The Lambda parameter λ:

$$\lambda^{ijt} = exp(\alpha^{it} + \psi^{j} + \beta^{j} * \omega^{it})$$

with α as a set of document fixed effects at time t, Ψ as a set of word fixed effects, β as estimates of word specific weights capturing the importance of word j in discriminating between documents, and ω as the estimate ofauthor i's position at time t

In English

α: author-document fixed effect

→ document length

ψ: word fixed effect

→ word frequency in dataset

β: word's importance

→ Score of word in dimension

ω: author's position in document

→ Score of document in dimension

Assumptions

1. Documents have one underlying-latent-dimension
2. Distribution of words follows Poisson distribution

Expectation Maximization algorithm

Step 1: Calculate starting values

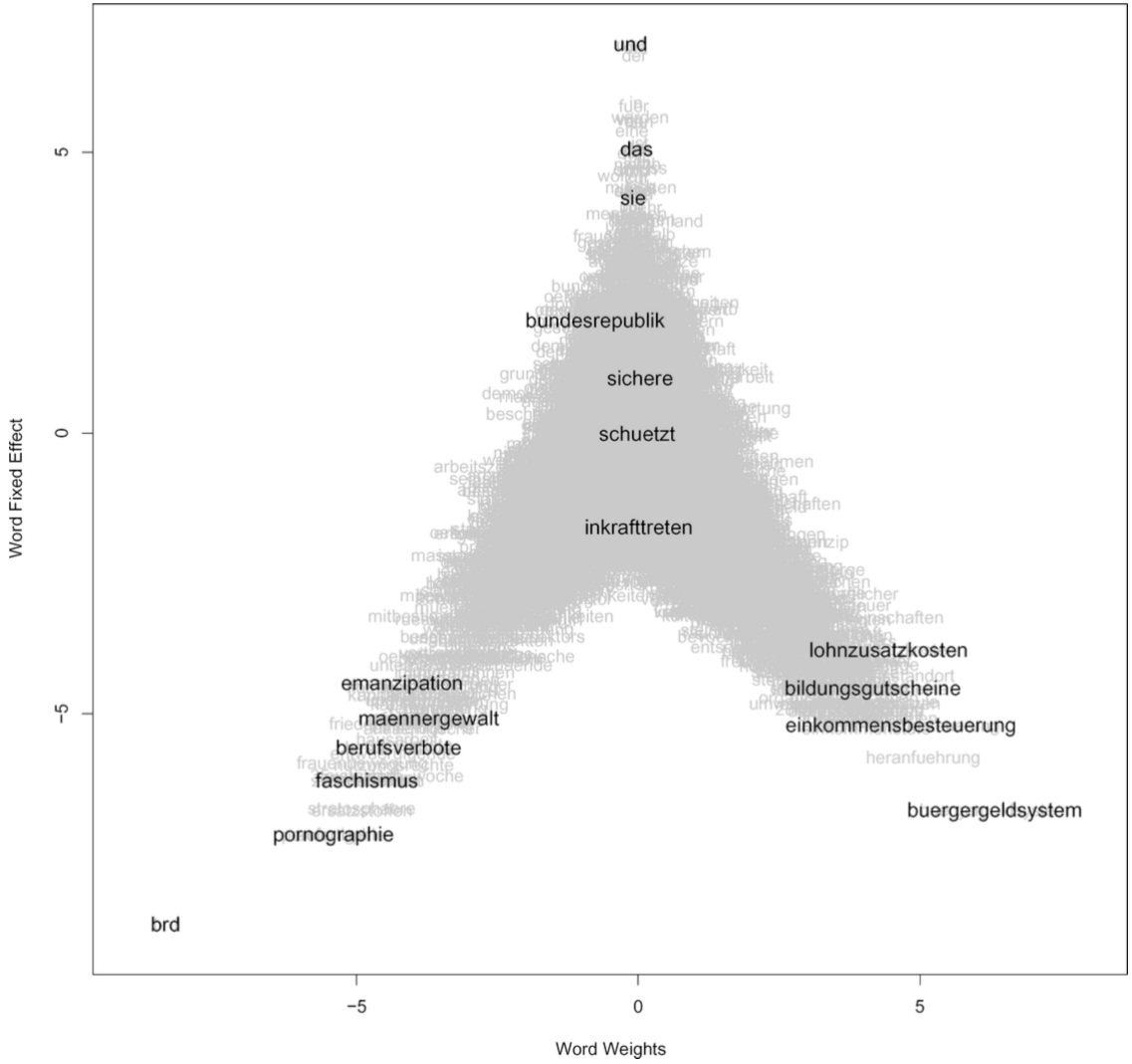
Step 2: Estimate party parameters by length and score of documents

Step 3: Estimate word parameters by expectation for the party parameters

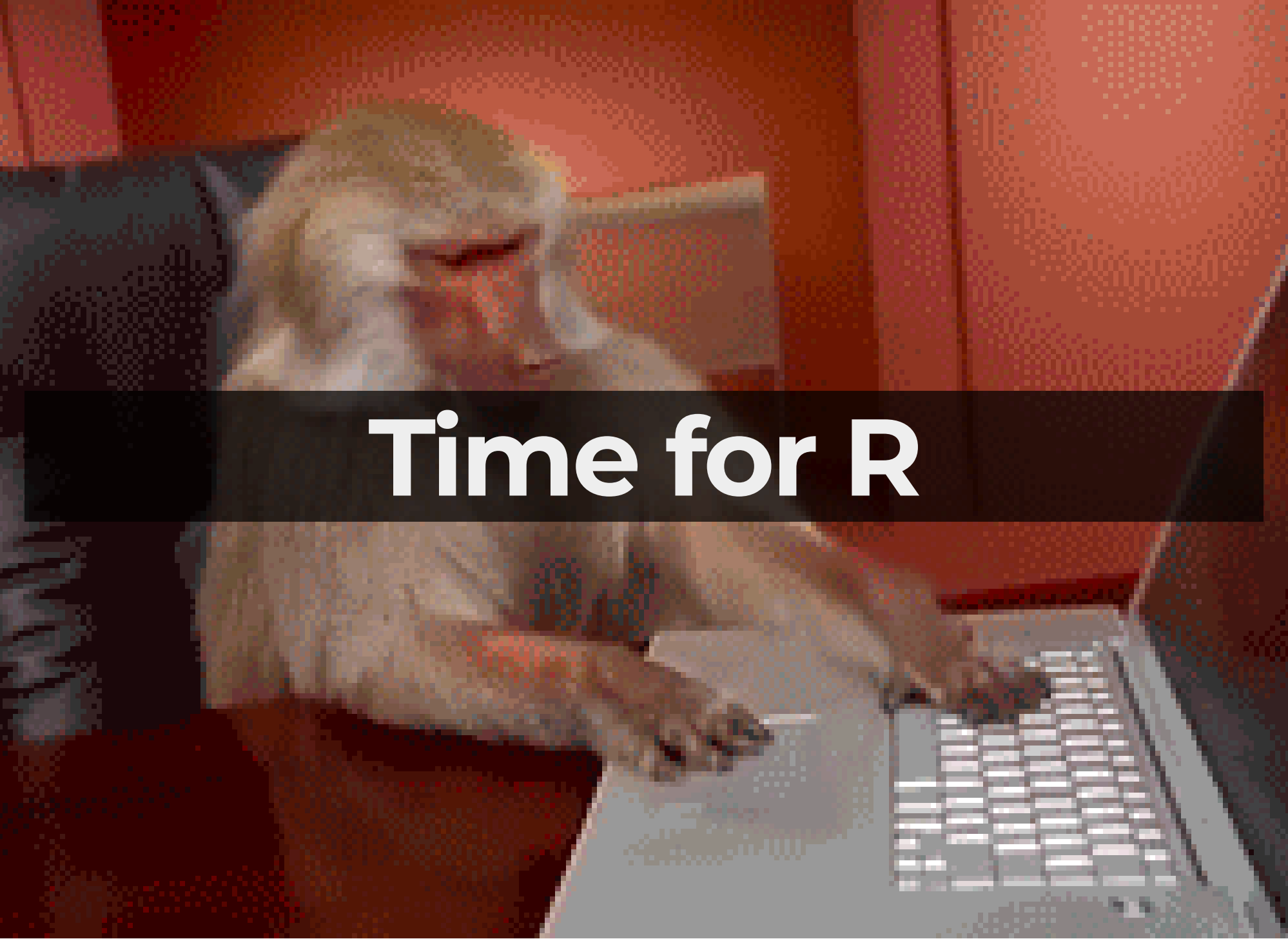
Step 4: Calculatelog-likelihood

Step 5: Repeat steps 2–4 until convergence

FIGURE 2 **Word Weights vs. Word Fixed Effects. Left-Right Dimension, Germany 1990–2005 (Translations given in text)**



Issue of validity


Time for R