Advanced text analysis methods

Or: (How) are new methods better?

Theresa Gessler

University of Zurich | http://theresagessler.eu/ | @th_ges

2022-05-12

Program

- general trends
- Methods that we have not covered
- word embeddings
- transformer models
- making sure your model is good

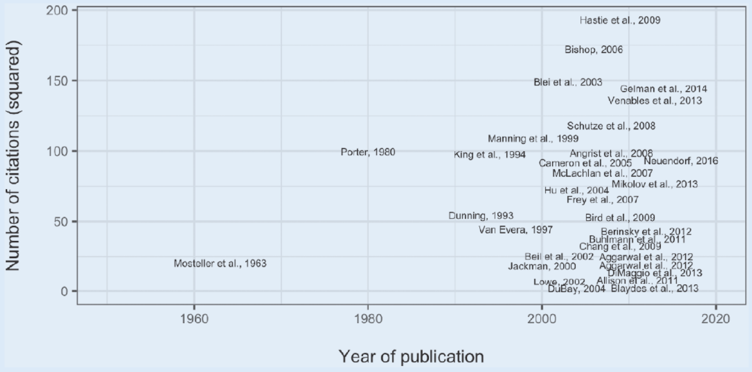
General Trends

- Text as Data is a rapidly growing field
  - recent development of the field
  - rapid innovations
  - transfer of methods from other fields

→ **importance of understanding principles, rather than specific methods**

Figure 3

Publication Year of the Most-Cited Sources



Fréchet, Nadjim, Justin Savoie, und Yannick Dufresne. "Analysis of Text-Analysis Syllabi: Building a Text-Analysis Syllabus Using Scaling". PS: Political Science & Politics 53, Nr. 2 (2020): 338–43. https://doi.org/10.1017/S1049096519001732.

General Trends

- new ways to implement methods (e.g. new algorithms)
  - focus on change
  - mostly building on bag of words and feature frequencies
- recent innovations that draw on the availability of large quantities of text
  - learn from big corpora and transfer lessons to small(er) amounts of data
    - word embeddings
    - transformers

Other bag-of-word methods

Other bag-of-word methods

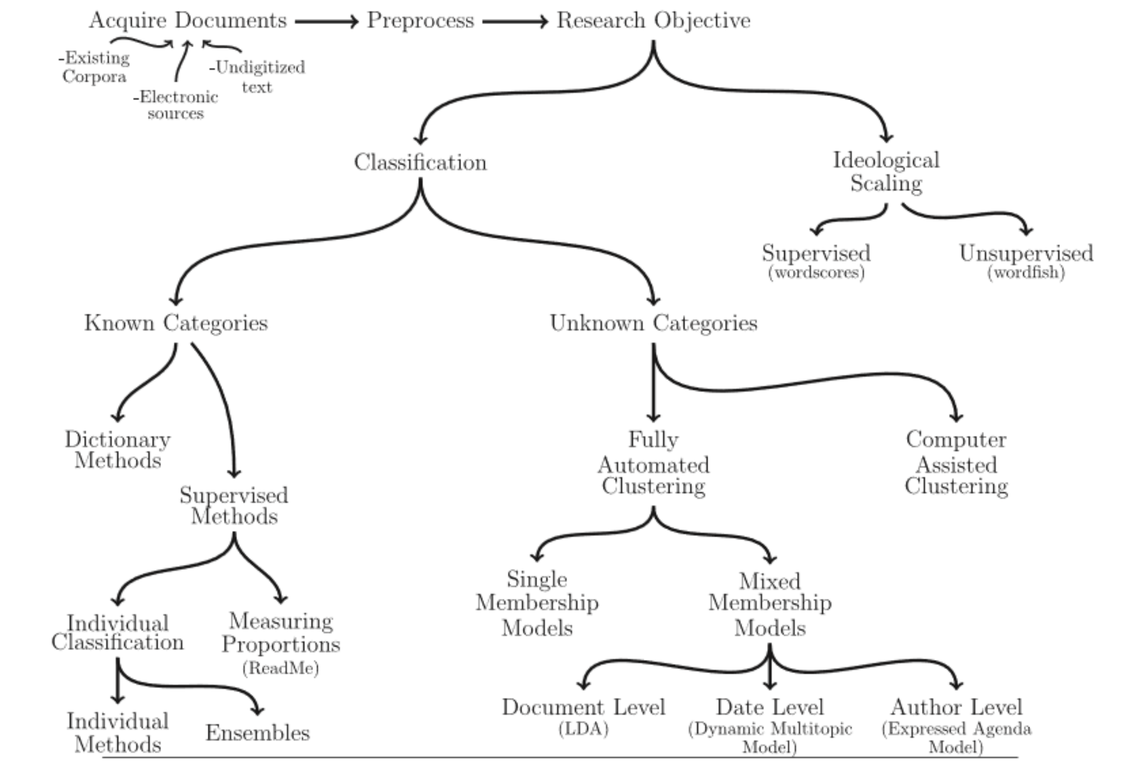


Fig. 1 An overview of text as data methods.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21, 267-297.

Other bag-of-word methods

- we cannot cover everything...
- ...but many 'traditional' text-as-data methods are fairly straight-forward extensions of what we have covered
    - e.g. different variants of clustering
    - e.g. different types of topic models
    - ...even scaling!

The typical steps:

- convert documents into a document-feature-matrix
- find a method to connect features and concepts

Other bag-of-word methods

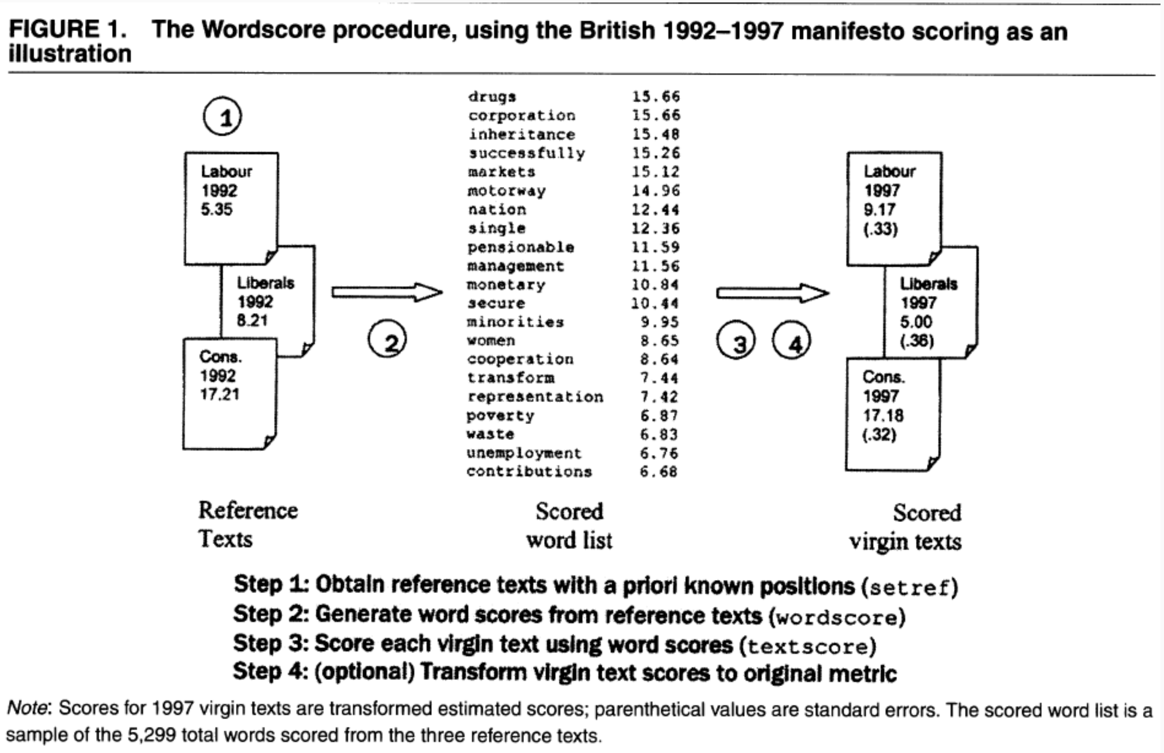


FIGURE 1. The Wordscore procedure, using the British 1992–1997 manifesto scoring as an illustration

drugs	15.66
corporation	15.66
inheritance	15.48
successfully	15.26
markets	15.12
motorway	14.96
nation	12.44
single	12.36
pensionable	11.59
management	11.56
monetary	10.84
secure	10.44
minorities	9.95
women	8.65
cooperation	8.64
transform	7.44
representation	7.42
poverty	6.87
waste	6.83
unemployment	6.76
contributions	6.68

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.

Laver, Michael, Kenneth Benoit, und John Garry. 2003. „Extracting Policy Positions from Political Texts Using Words as Data“. The American Political Science Review 97 (2): 311–31.

generate dfms → calculate probability for finding words in left vs. right texts as word scores → calculate left-right positions of new texts from word scores

Other bag-of-word methods

→ Think of text analysis methods as a family tree where everyone is highly related!

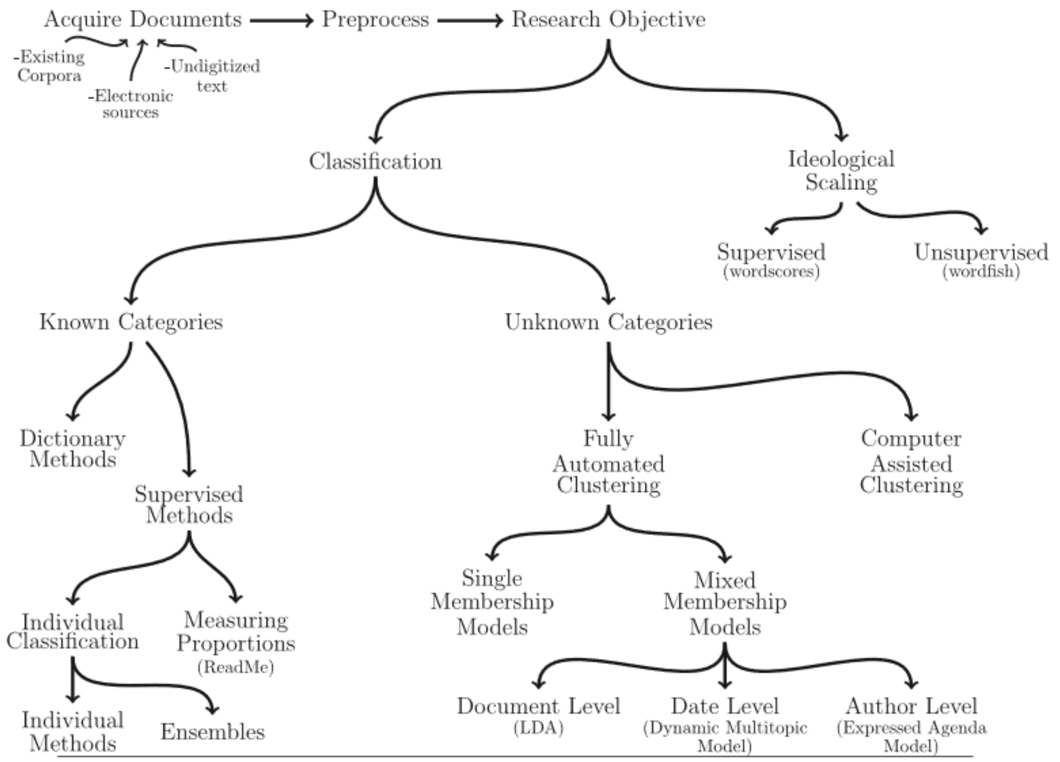


Fig. 1 An overview of text as data methods.

→ try to understand new methods by comparison to methods you know

Word Embeddings

Word Embeddings

Some newer methods mark a (partial) departure from this - for example word embeddings

The idea

- most recent innovations draw on the availability of large quantities of text
  - typically, they use this to enhance feature definitions
  - most of them use this to move beyond the 'bag of words'

Word Embeddings

The idea

- *self supervision*: supervised tasks with labels provided by data itself
  - this can be an external data source (e.g. Wikipedia, news articles, webpages, ...)
- *understand words by nearby words* ('context window') → context provides information to define word
- *reduce dimensionality* (~50-500 dimensions) to create dense representations of words

Word Embeddings

The idea

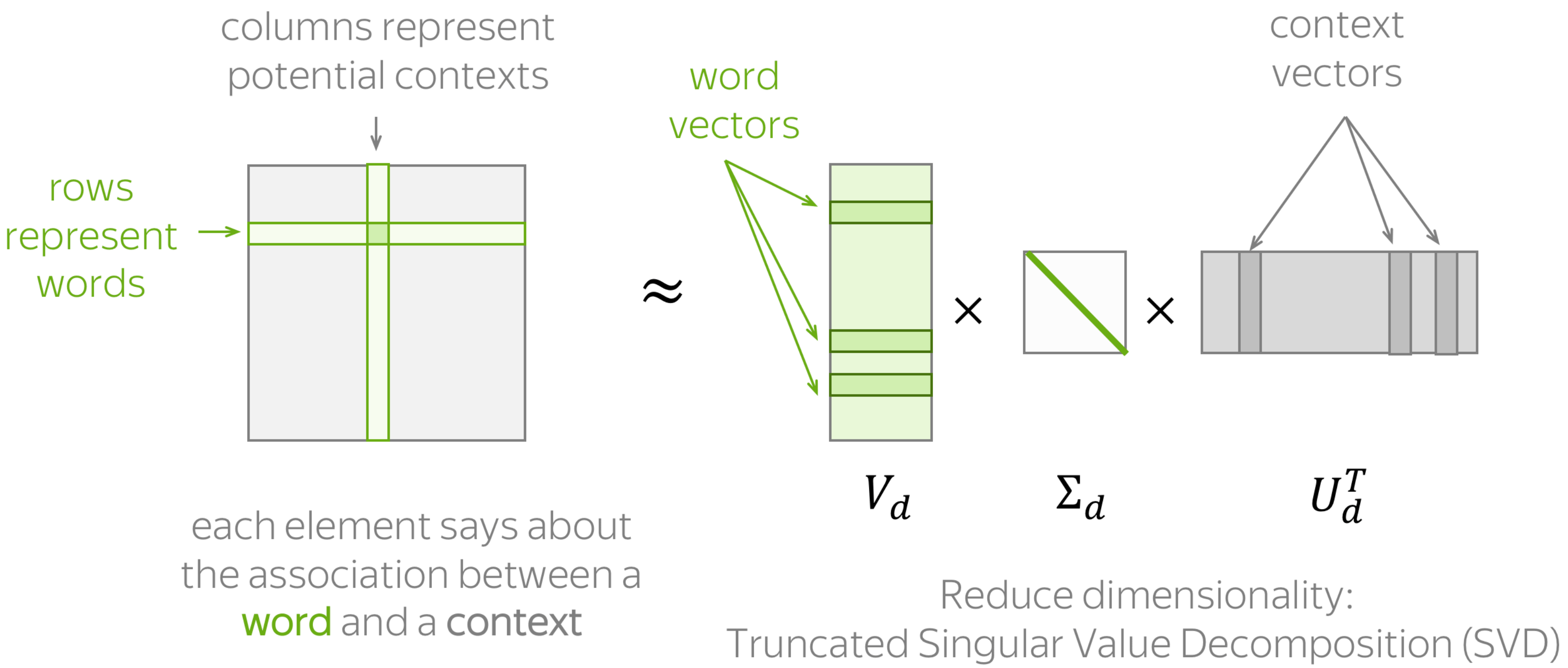

columns represent potential contexts
rows represent words
each element says about the association between a word and a context
≈
word vectors
×
×
context vectors
V_d
Σ_d
U_d^T
Reduce dimensionality:
Truncated Singular Value Decomposition (SVD)

Image source & more details: Lena Voita's NLP Course

→ methods differ in their context window, their metric etc.

Word Embeddings

Advantages of Word Embeddings (Grimmer et al 2022, 79):

- **encode similarity**: in bag-of-words, we cannot see whether words are similar
- **automatic generalization**: learning about one word helps us learn about related words (e.g. their positivity) → sharing information across words
- **measure of meaning**: measure whether words are used in similar contexts

Word Embeddings

Use

- word embeddings are used to solve many tasks that are not directly relevant to many political science tasks
    - e.g. named entity recognition
- for many politicial science uses, we have to aggregate them to the document level (weighting, doc2vec, ...)
- or use them as feature-definitions within traditional methods
- but there's a rapidly growing literature on applying word embeddings within political science

Word Embeddings

Use

- Enhance machine learning performance for sentiment analysis: Rudkowsky et al 2018
- dictionary-like methods for measuring stereotype associations Kroon et al
- expand corpus-specific dictionaries: Rheault et al 2016
- broader guidelines
    - e.g. general guidelines for using word embeddings: Spirling & Rodríguez 2021
    - e.g. overtime embeddings: Rodman 2020
    - e.g. 'embedding regressions' Rodríguez, Spirling & Stewart

Transformer Models

Transformer Models

The idea

- word embeddings were a nice idea
- BUT: they still don't fully account for word order, differences in context etc.
- ...with the growth of text data, we may be able to account for this

→ **transformer models**

Transformer Models

The idea

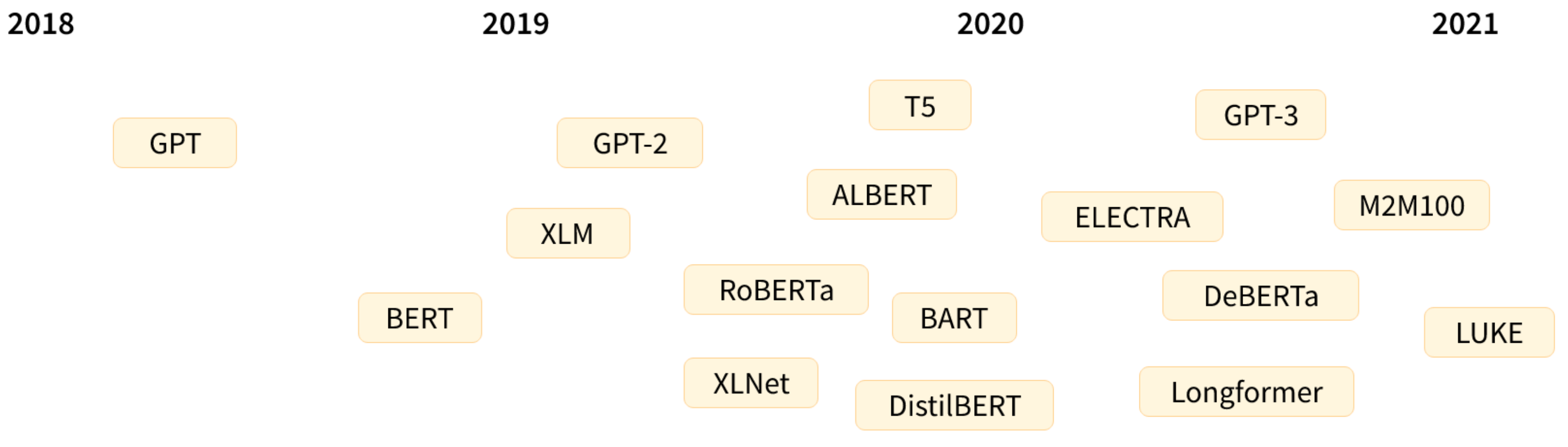


Image source: HuggingFace

→ Training on a huge amount of data (e.g.: Wikipedia), often for several weeks

- multi-layered models that try to predict elements of sentences while some parts are masked → **attention**

Transformer Models

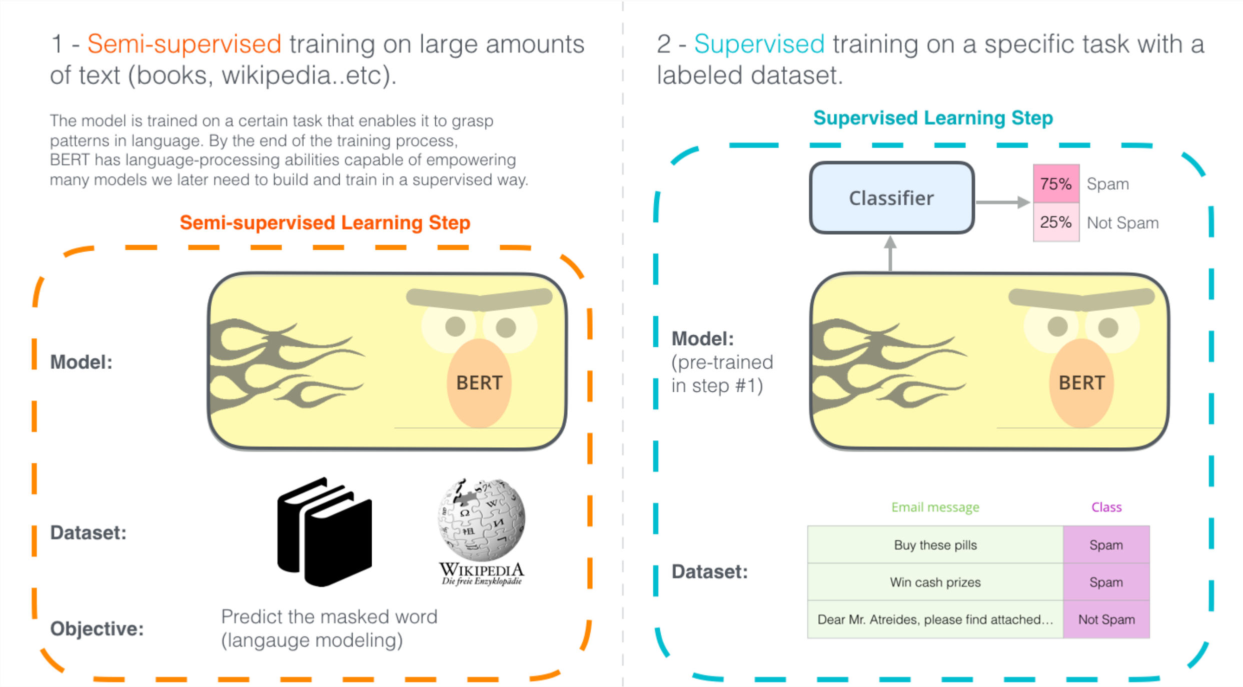
The idea

- proper 'language models' built for aims which are less relevant for political scientists, such as...
    - machine translation
    - text generation, prediction of words and sentences
    - question answering
- ...however, they perform really well on polsci-tasks as well!
    - → fine-tuning for task on a general language model

Transformer Models

The idea

- transfer learning: fine-tuning a model previously trained on large amounts of data



1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

2 - Supervised training on a specific task with a labeled dataset.

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Image source: Jay Alammar: The Illustrated BERT, ELMo and co. (How NLP Cracked Transfer Learning)

Transformer Models

Places to start

HuggingFace Course (*in python!*)

Tutorial by Christopher Klamm, Moritz Laurer and Elliott Ash

Lena Voita: NLP for you

So what is the best model?

So what is the best model?

- "all models of language are wrong but some are useful"

"Once the research question is defined and the researcher is performing measurement, we can assess how accurate a method is at recovering the specific concept of interest to the researcher. Because there were many different things about the document that the researcher could have chosen to measure, we do not assume there is a single model that captures a true data generating process, but hopefully there are options that help us capture what we need for our research question. In other words, we are agnostic about which particular model is used for measurement, as long as the model can accurately and reliably measure the concept of interest." (Grimmer et al 2022, 19)

→ How do you make sure you have a useful model?

So what is the best model?

Know your data

- the better you know your data, the easier it is to judge whether a model works

Know your model

- see what is behind your model predictions
    - e.g. which features does a dictionary actually pick up? Which features are most predictive in a machine learning model?
    - how does your complex model classify example sentences?
    - use quantitative and qualitative validation techniques and present the evidence

Explain your model

- make sure you understand what is happening
- use only as much complexity as needed for addressing a task

→ no inherent benefits of more complex models, though performance statistics for some tasks have become very impressive!

So what is the best model?

Upsides of complex models

- **significantly improved performance for many tasks**

Downsides

- **understandability**
  - Do you understand BERT? Does your reviewer?
- **complexity**
  - hidden biases
- **environmental impact and computational power**
  - change in scale between word embeddings and transformer models
  - "Training a single BERT base model (without hyperparameter tuning) on GPUs was estimated to require as much energy as a trans-American flight." (Bender et al)

Questions?

Where can you go next?

Where can you go next?

Please fill out the evaluation forms (which you'll get via e-mail)

- **to help me improve this course**
- **to help me get a job**

Where can you go next?

Some resources to go further

- Kohei Watanabe, Stefan Müller: quanteda tutorials
- Atteveldt, Wouter van, Damian Trilling, und Carlos Arcíla. Computational analysis of communication: a practical introduction to the analysis of texts, networks, and images with code examples in Python and R. Hoboken, NJ: John Wiley & Sons, 2021. also online
- Grimmer, Justin, Margaret E. Roberts, und Brandon M. Stewart. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton, New Jersey Oxford: Princeton University Press, 2022.
- machine learning with caret
- Julia Silge, David Robinson: Text Mining with R
    - but: they use the `tidytext` package

Where can you go next?

General text books

General theory on text analysis and machine learning

- Daniel Jurafky, James Martin: Speech and Language Processing
    - new edition draft
- Gareth James, Daniela Witten, Hastie, Trevor, Robert Tibshirani. An Introduction to Statistical Learning, Springer
- Hastie, Trevor, Robert Tibshirani, und Jerome Friedman. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY: Springer New York, 2017.
- Christopher Manning, Prabhakar Raghavan, Hinrich Schütze: An Introduction to Information Retrieval

Python

- Text Analysis in Python for Social Scientists. Discovery and Exploration. and Part 2 on Prediction and Classification
- Applied Text Analysis with Python

Thank you!