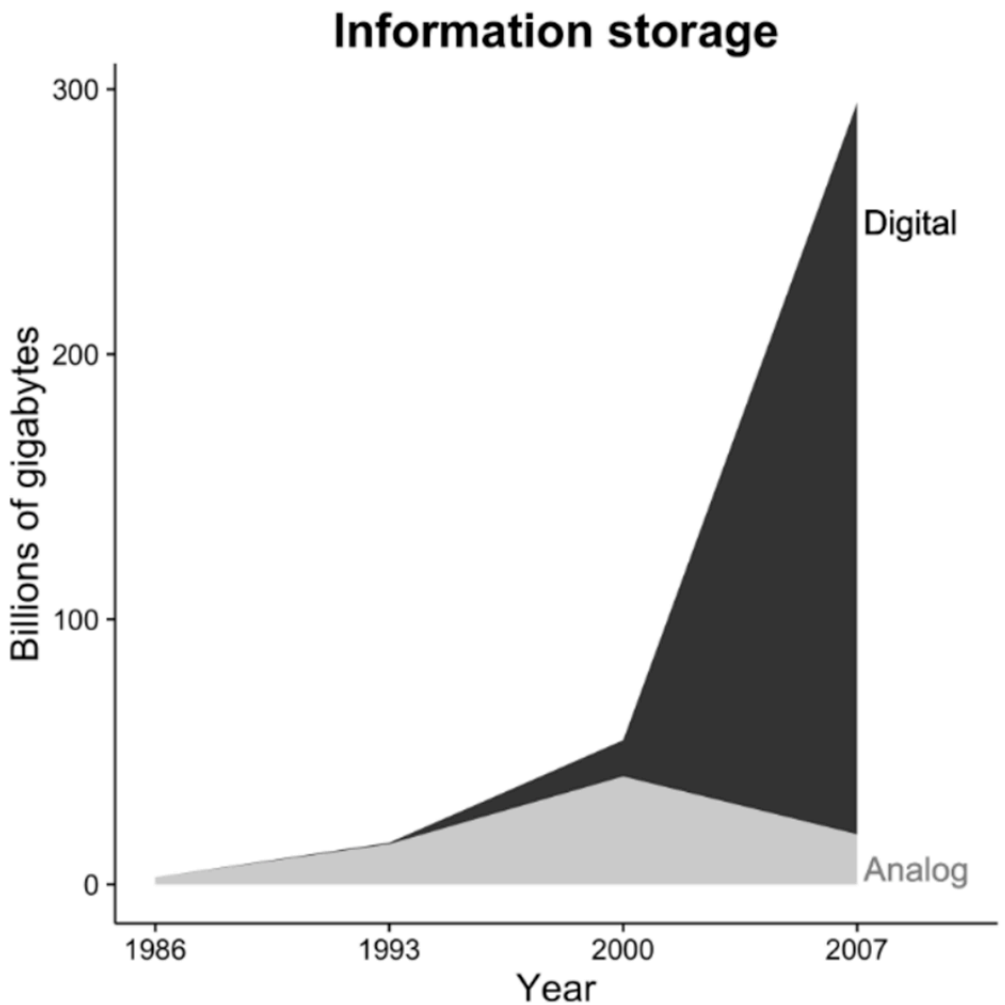
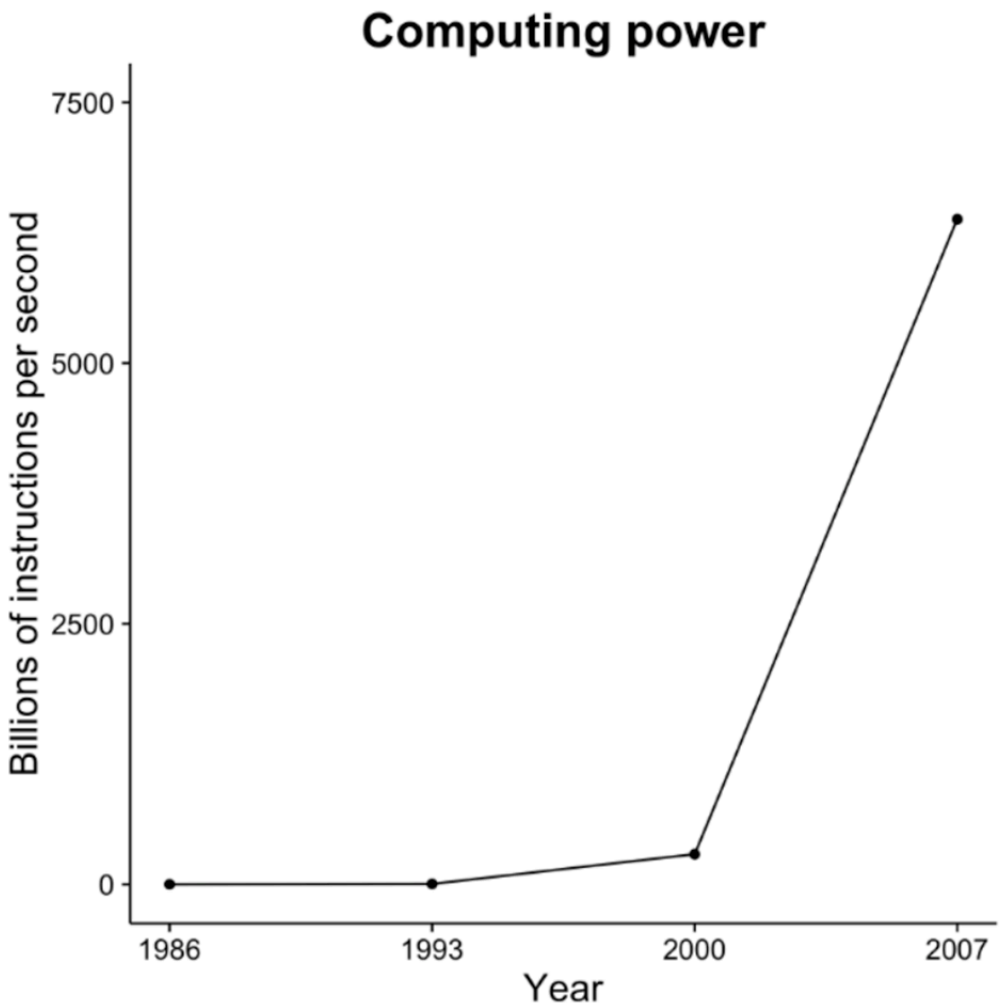
Computational Analysis of Digital Communication

Week 1: Introduction to Computational Methods in Communication Science

Prof. Dr. Wouter van Atteveldt

INCREASING AMOUNT OF DATA AVAILABLE ONLINE





Hilbert & Lopez, 2011

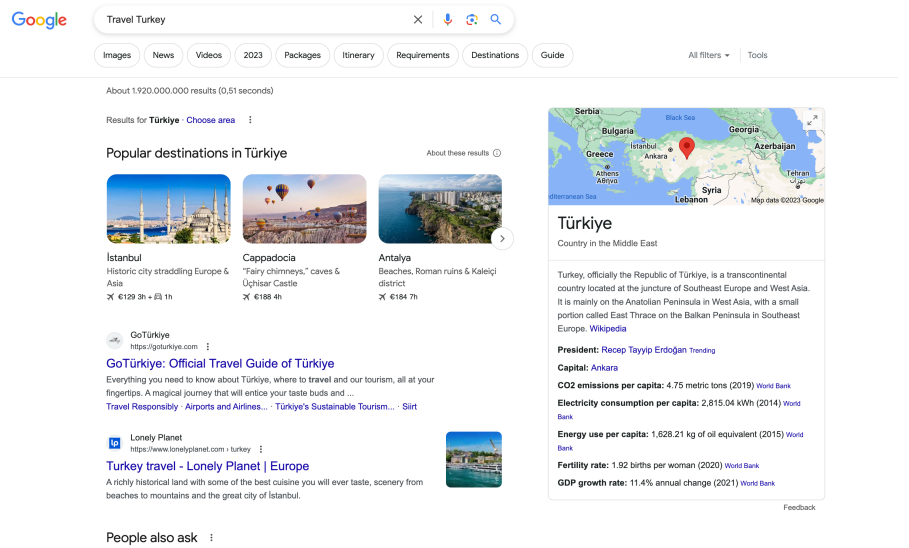
MUCH OF WHAT WE KNOW ABOUT HUMAN BEHAVIOR…

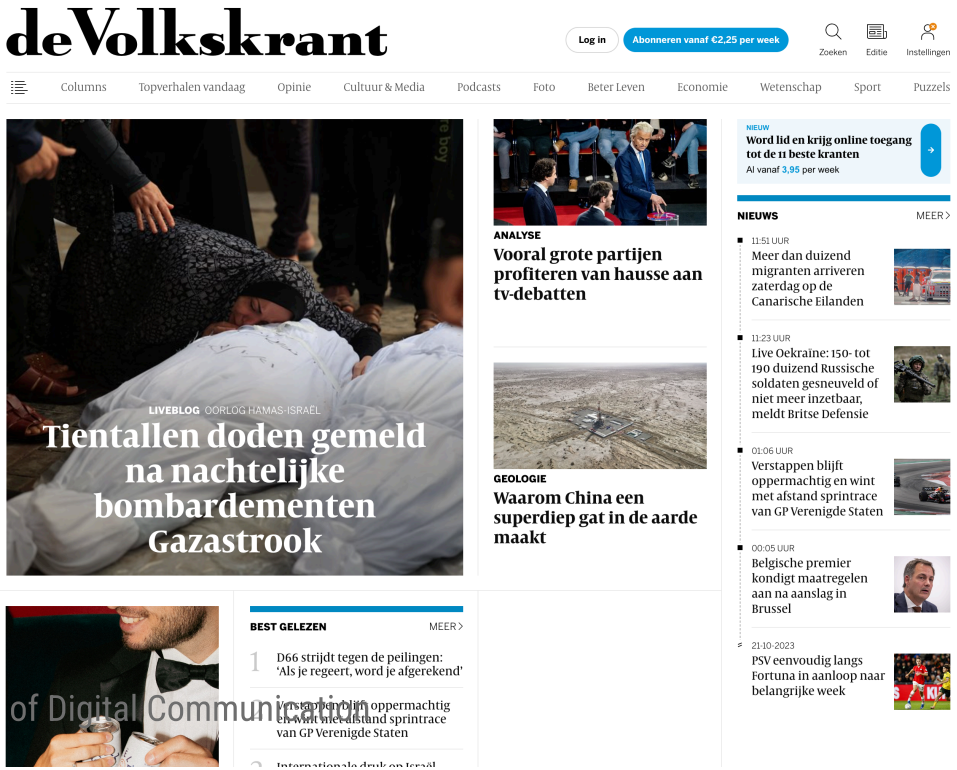
…is based on what people tell us:

- in self-report measures in surveys
- in responses in experimental research
- in qualitative interviews

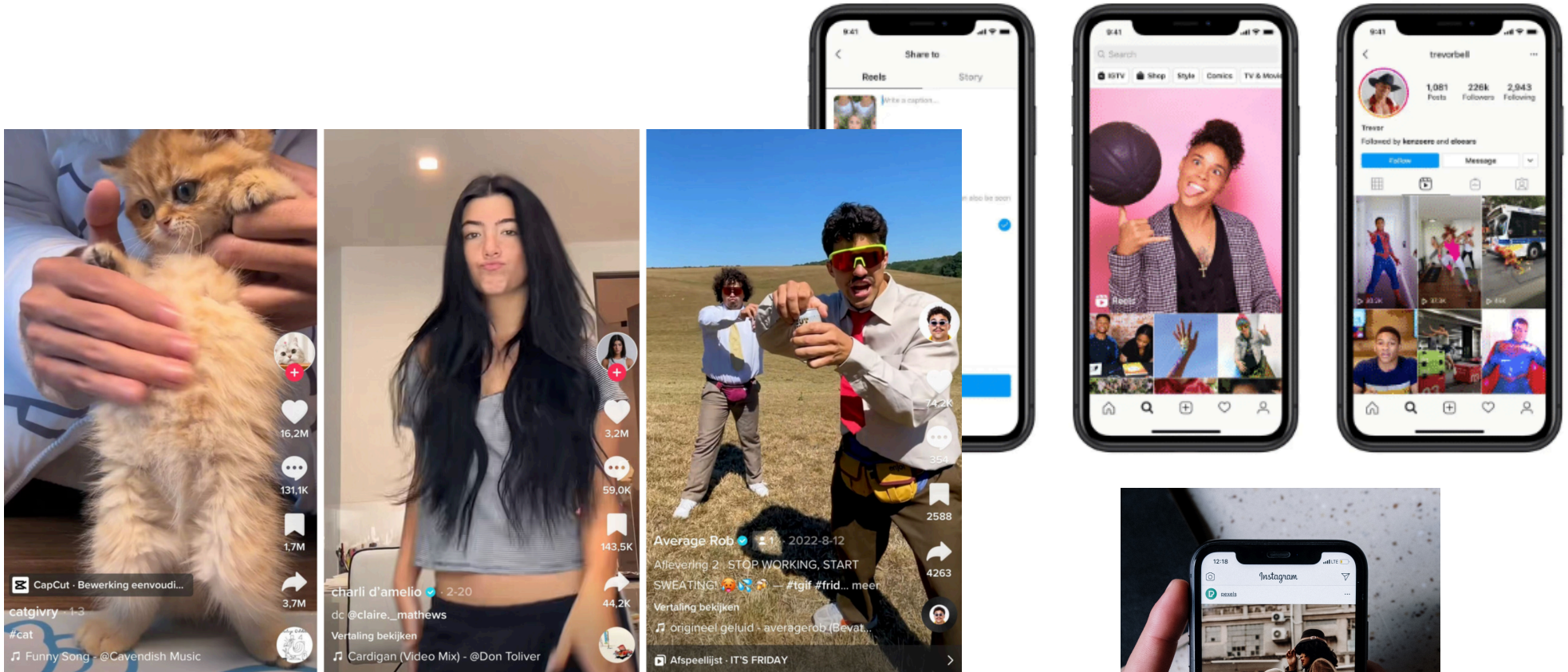
Note: Although valuable, such measurements can be biased (Scharkow, 2013; Parry et al., 2021)!

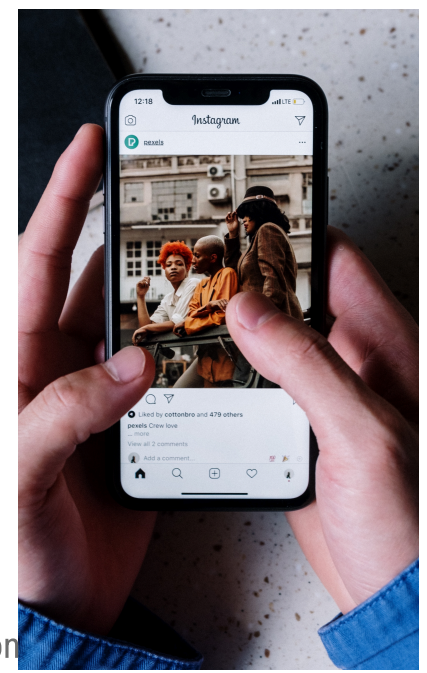
BUT A LOT OF (MASS) COMMUNICATION LOOKS LIKE THIS…





…OR IS BASED ON USER-GENERATED CONTENT





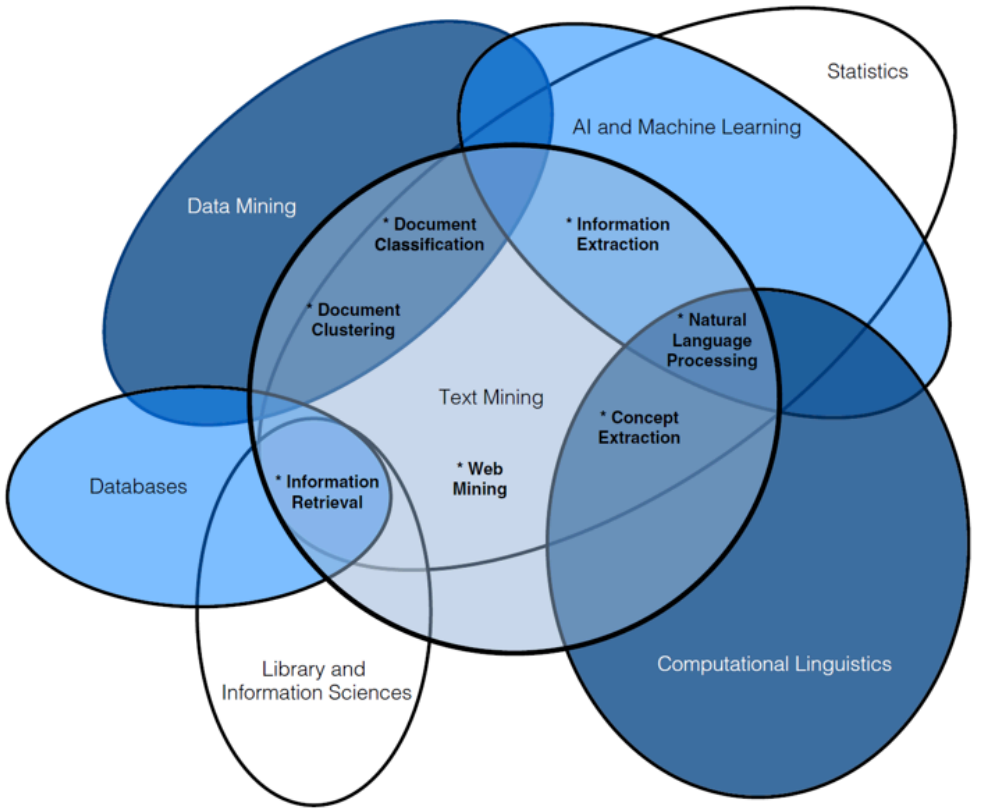
How can we analyze large amount of texts?

This is what we will discuss in this course!

OBJECTIVES AND LEARNING GOALS

After completion of the course, you will…

1. be able to identify data analytic problems, analyze them critically, and find appropriate solutions
2. have a good understanding of the general text classification pipeline
3. have practical knowledge about different approaches of text classification (incl. dictionary approaches, machine learning, large language models…)



SKILLS AND METHODS

With regard to the specific methods being taught in R, you will be able to…

- gather, scrape, and import data from different file types, APIs, and websites
- link data from different sources to create new insights
- clean and transform messy data into a tidy data format ready for text classification and analysis
- use different approaches (e.g., dictionary, classic machine learning, transformer, LLMs) to extract information from textual data
- perform statistical analyses on the substantive data

```
  dfm %>%
  dfm_trim(min_docfreq = 10)
dfm_train
```

```
## Document-feature matrix of: 51,765 documents, 16,463 features (99.28% sparse) and 6 docvars.
##        features
## docs    awesom song a mellow groov that has great weird video
##   text1      1    2 3      1     1    1   1     1     1     1
##   text2      0    3 4      0     0    0   0     0     0     0
##   text3      2    1 5      0     0    1   0     0     0     0
##   text4      0    3 0      0     0    1   0     1     0     0
##   text5      0    0 4      0     0    3   0     0     0     0
##   text6      0    0 0      0     0    1   0     0     0     0
## [ reached max_ndoc ... 51,759 more documents, reached max_nfeat ... 16,453 more features ]
```

Machine Learning

Training the algorithm

Now, we can train a text model with e.g., the naive bayes algorithm:

```
library(quanteda.textmodels) ## install first!
nbmodel <- textmodel_nb(dfm_train, dfm_train$fivestar)
summary(nbmodel)
```

```
##
## Call:
## textmodel_nb.dfm(x = dfm_train, y = dfm_train$fivestar)
##
## Class Priors:
## (showing first 2 elements)
## FALSE  TRUE
##   0.5   0.5
##
## Estimated Feature Scores:
##          awesom     song       a    mellow     groov    that      has    great
## FALSE 0.0001261 0.008263 0.02145 0.0001228 0.0001774 0.01078 0.002952 0.002307
## TRUE  0.0002994 0.008533 0.02115 0.0001399 0.0002143 0.01001 0.002895 0.003351
##           weird     video        .       an     eleph       goe      to
## FALSE 8.496e-05 0.0002017 0.04698 0.002594 3.475e-06 0.0002446 0.01714
## TRUE  7.766e-05 0.0002448 0.04741 0.002864 3.442e-06 0.0002228 0.01676
##             find       him       in   paradis       if      you      want
## FALSE 0.0006062 0.0006350 0.009124 2.606e-05 0.002430 0.005672 0.0008946
## TRUE  0.0005668 0.0005968 0.010349 3.442e-05 0.002229 0.006521 0.0007907
##            cool       ,       get    this      one        !     best       cd
## FALSE 0.0003110 0.04215 0.002298 0.01157 0.004107 0.003561 0.001975 0.002317
## TRUE  0.0002644 0.04141 0.002116 0.01357 0.005136 0.006476 0.002731 0.002722
```

WHO AM I?

- Professor of Computational Communication Science & Political Communication
- Research Interest
  - Effects of (social) media consumption
  - Changing patterns of media consumption & production
- Methodological Interests
  - Computational Methods and Machine Learning
  - Text (and multimedia) analysis
  - Data Donation
- More info: https://vanatteveldt.com

CONTENT OF THIS LECTURE

1. What is Computational Social Science?
   1.1. Definition and Examples
   1.2. Relevance of Computational Methods
   1.3. Characteristics of Big Data
   1.4. Opportunities and Pitfalls
2. What is Computational Communication Science?
   2.1. Definition
   2.2. Typical Research Areas
   2.3. Examples of Computational Communication Research
3. Introduction to Automated Text Analysis
   3.1. Text as Data
   3.2. The General Text Classification Pipeline
   3.3. Timeline of Natural Language Processing
   3.3. A Small Example
4. Ethics of Computational Communication Research
   4.1. A Controversional Study (Kramer et al., 2014)
   4.2. Ethical Challenges
   4.3. Practical Guidelines
5. Course Formalities
   5.1. Introduction of Teachers
   5.2. Course Material and Schedule
   5.3. Attendance and Assignments

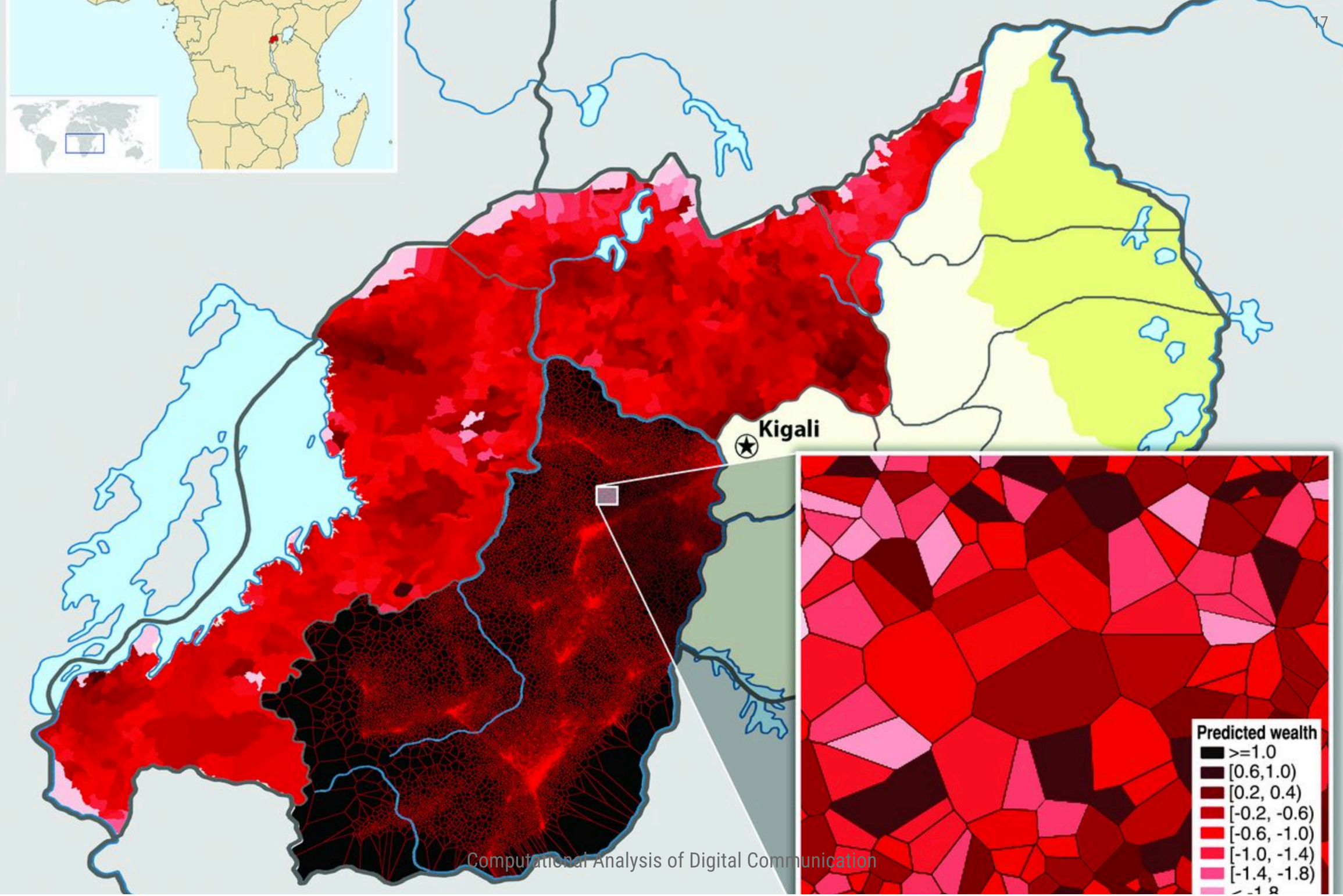
What is Computational Social Science?

…and why should we care?

EXAMPLE: SURPRISING SOURCES OF INFORMATION

- In 2009, researchers wanted to study wealth and poverty in Rwanda
- They conducted a survey with a random sample of 1,000 customers of the largest mobile phone provider
- They collected demographics, social, and economic characteristics (incl. wealth)
- So far, traditional social science, right?
- The authors also had access to complete call records from 1.5 million people
- Combining both data sources, they used the survey data to “train” a machine learning model to predict a person’s wealth based on their call records
- They also estimated the places of residence based on the geographic information embedded in call records

Blumenstock, Cadamura, & On, 2015


Kigali
Predicted wealth
>=1.0
[0.6,1.0)
[0.2, 0.4)
[-0.2, -0.6)
[-0.6, -1.0)
[-1.0, -1.4)
[-1.4, -1.8)

COMPUTATIONAL SOCIAL SCIENCE

- Field of social science that uses algorithmic tools and large/unstructured data to understand human and social behavior
- Complements rather than replaces traditional methodologies: Methods are not the goal, but contribute to data generation
- Includes methods such as, e.g.,:
  - Data mining (e.g., scraping and gathering of large data sets)
  - Software development for social science experiments
  - Automated text analysis (e.g., sentiment analysis, keyword extraction, dictionary approaches)
  - Image classification (e.g., face recognition, visual topic modeling)
  - Machine learning approaches (e.g., for classification, prediction, topic modeling)
  - Actor-based modeling (e.g., simulation of social behavior, spreading of information)
  - …

WHY IS THIS IMPORTANT NOW?

- Vast **amounts of digitally available data**, ranging from social media messages and other digital traces to web archives and newly digitized newspaper and other historical archives
- Large-scale records (**big data**) of persons or businesses are created constantly
- Powerful and comparatively cheap **processing power**, and easy to use **computing infrastructure** for processing these data
- Improved **tools to analyze** this data, including network analysis methods and automatic text analysis methods such as supervised text classification, topic modeling, word embeddings, as well as large language models

10 CHARACTERISTICS OF BIG DATA

#	Characteristic	Description
1	Big	The scale or volume of some current data sets is often impressive. However, big data sets are not an end in themselves, but they can enable certain kinds of research including the study of rare events, the estimation of heterogeneity, and the detection of small differences
2	Always-on	Many big data systems are constantly collecting data and thus enable to study unexpected events and allow for real-time measurement
3	Nonreactive	Participants are generally not aware that their data are being captured or they have become so accustomed to this data collection that it no longer changes their behavior.
4	Incomplete	Most big data sources are incomplete, in the sense that they don't have the information that you will want for your research. This is a common feature of data that were created for purposes other than research.
5	Inaccessible	Data held by companies and governments are difficult for researchers to access.

10 CHARACTERISTICS OF BIG DATA

#	Characteristic	Description
6	Nonrepresentative	Most big datasets are nonetheless not representative of certain populations. Out-of-sample generalizations are hence difficult or impossible.
7	Drifting	Many big data systems are changing constantly, thus making it difficult to study long-term trends
8	Algorithmically confounded	Behavior in big data systems is not natural; it is driven by the engineering goals of the systems.
9	Dirty	Big data often includes a lot of noise (e.g., junk, spam, spurious data points…)
10	Sensitive	Some of the information that companies and governments have is sensitive.

Salganik, 2017, chap. 2.3

PRO'S AND CON'S OF COMPUTATIONAL METHODS

Opportunities

- We can study actual behavior instead of simply self-reports
- We can study human beings in their social context instead of in an artificial lab setting
- We can increase our N (higher power)
- Potential to uncover patterns and insights that we couldn't investigate before

Pitfalls

- Techniques often (rather) complicated
- Data is often proprietary (not shared openly)
- Samples are often biased
- Often, data have only insufficient metadata
- Risks of no longer understanding the models we use (black box)

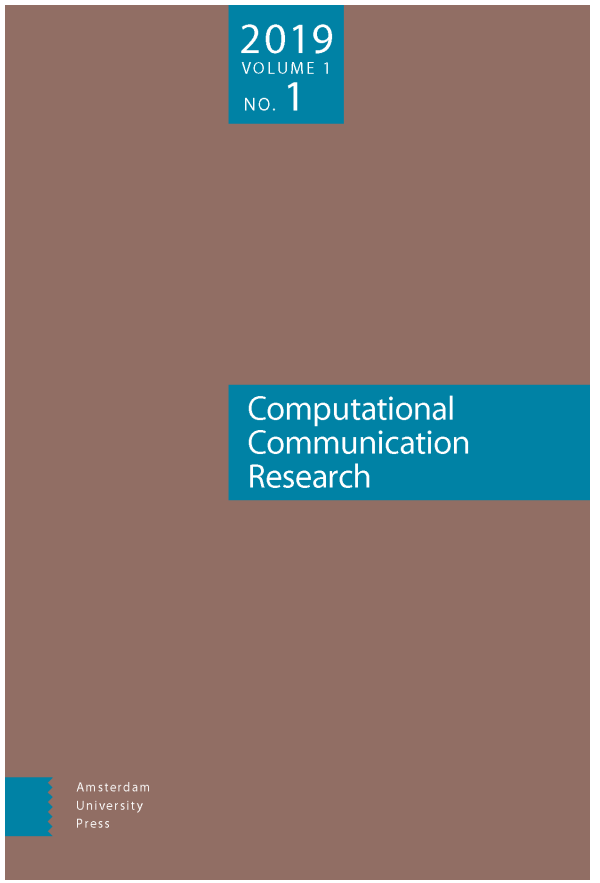
Computational Communication Science

Why computational methods are important for communication research…

DEFINITION

> "Computational Communication Science (CCS) is the label applied to the emerging subfield that investigates the use of computational algorithms to gather and analyze big and often semi- or unstructured data sets to develop and test communication science theories"

Van Atteveldt & Peng, 2018



TYPICAL RESEARCH AREAS

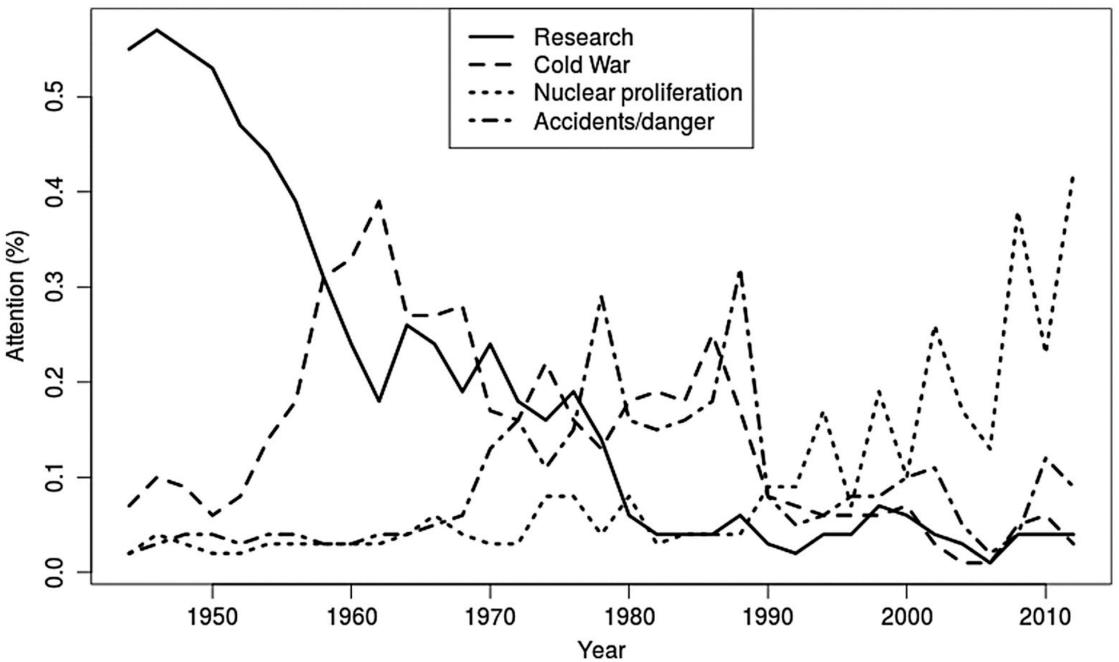
Computational communication science studies thus usually involve:

1. large and complex data set
2. consisting of digital traces and other "naturally occurring" data
3. requiring algorithmic solutions to analyze (e.g., machine learning, LLMs)
4. allowing the study of human communication by applying and testing communication theory

- Political Communication
  - Democratization and Polarization
  - Hate Speech
- Social Media Use
  - Tracking of actual social media use
  - Spreading of behavior, information, or emotions
- Health Communication
  - Prevalence of health information online
- (Online) Journalism
  - News coverage across decades
  - Gender equality

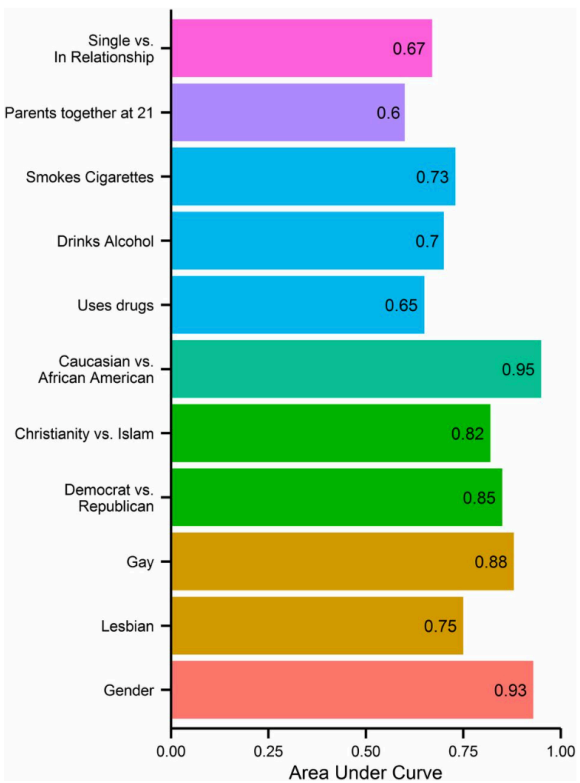
EXAMPLE 1: ANALYZING NEWS COVERAGE

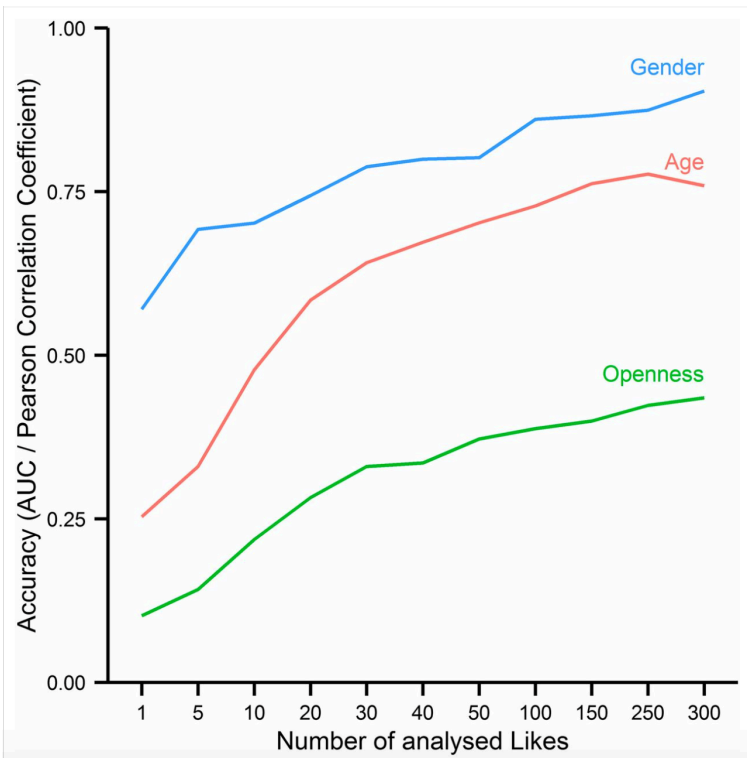
- Jacobi and colleagues (2016) analyzed the coverage of nuclear technology from 1945 to 2014 in the New York Times
- Analysis of 51,528 news stories (headline and lead): Way too much for human coding!
- Used “LDA topic modeling” to extract latent topics and analyzed their occurrence over time



EXAMPLE 2: FACEBOOK DATA TO PREDICT PERSONALITY

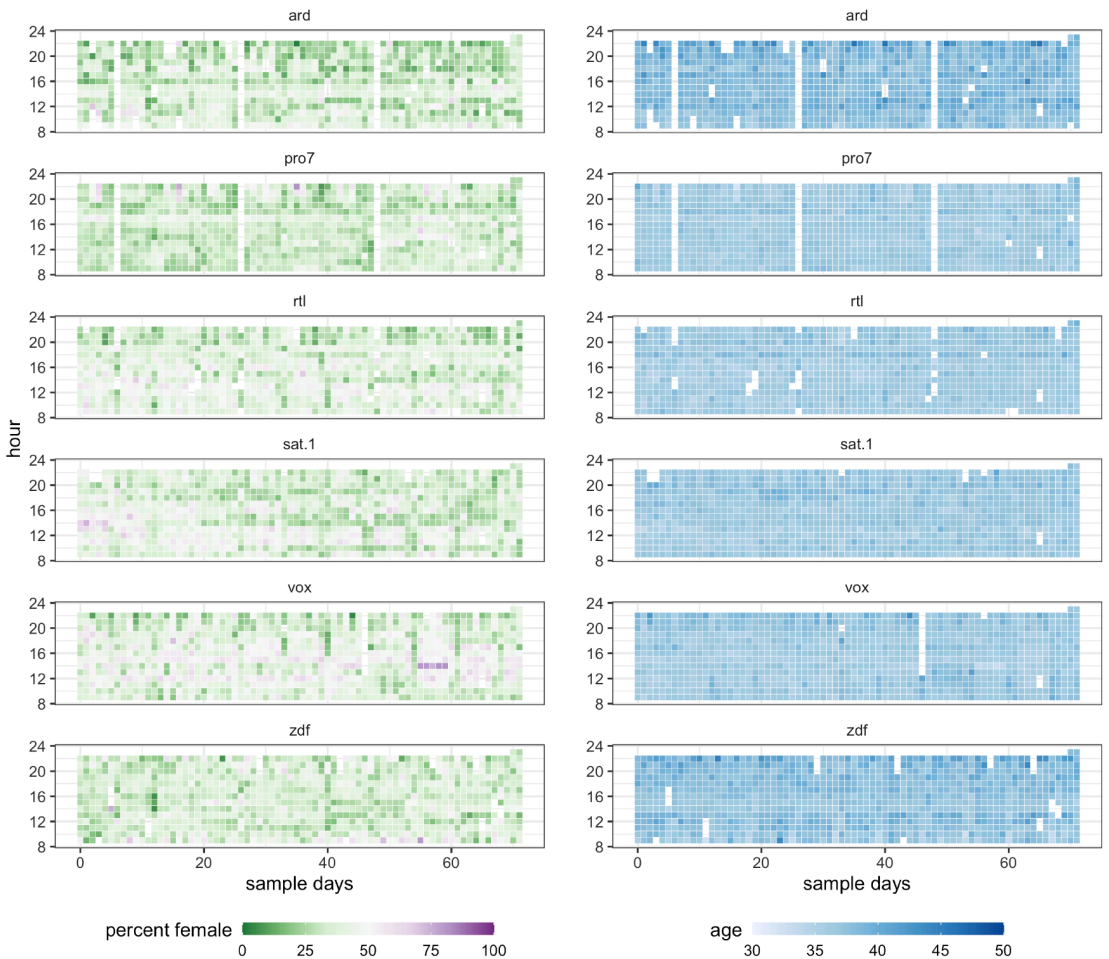
- Kosinski and colleagues (2013) used a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric test
- Were able to show that one can predict a variety of personal characteristics and personality traits from simple Facebook likes





EXAMPLE 3: GENDER REPRESENTATION IN TV

- Women on average remained underrepresented on TV, with 6.3 million female faces out of 16 million total (estimated proportion .39, 95% CI: .37-.42)
- This strong overall bias was mirrored across specific subsamples (news, sports, advertising…)



Introduction to Automated Text Analysis

The core topic of this course!

A “NEW” KIND OF DATA

- A lot of communication is encoded in texts
- But text does not look like data we can easily analyze…

Experimental data

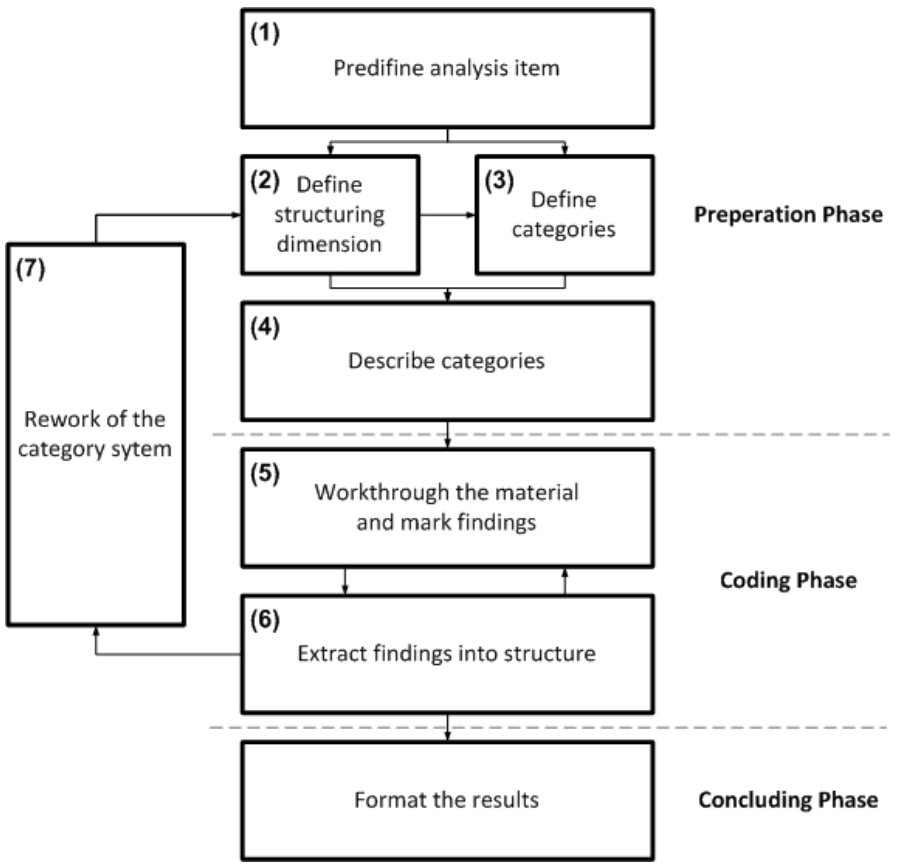
```
# A tibble: 6 × 5
     id condition sns_use well_being pers1
  <int> <chr>       <dbl>      <dbl> <dbl>
1     1 A            1.42     -0.986  2.22
2     2 B            4.29      4.44   6.47
3     3 A            4.78      4.66   3.73
4     4 B            3.84      1.83   1.47
5     5 A            4.16      8.18   2.44
6     6 B            1.60     -1.38   4.66
```

Text data

```
# A tibble: 6 × 1
  text
  <chr>
1 "North Korea launched a ballistic mis…
2 "Tributes poured in for former Republ…
3 "An Indian couple have arrived for th…
4 "Unique events that led to civilisati…
5 "Billionaire Peter Thiel is facing op…
6 "In some ways, accounts of “human ori…
```

TRADITIONAL *TEXT* ANALYSIS

- Choosing the texts that contain the content and one wants to analyze (1)
- Define the units and categories of analysis (2) & (3)
- Describe categories and develop a set of rules for the manual coding process (4)
- Coding the text according to the rules (5), which usually requires a lot of manual work
- Make sense of codes (6) and rework the codes and rules (7) and redo the analysis
- Analyze frequencies, relationships, differences, similarities between units/codes



Problem: Requires a lot of work and there are always more texts than humans can possibly code manually!

DEFINITION

> Text analysis is “a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use”

Krippendorff, 2004

WHAT IS TEXT?

tion and analysis that require external software modules or that go beyond the bag-of-words assumption, using word positions and syntactic relations. The purpose of this section is to provide a glimpse of alternatives that are possible in R, but might be more difficult to use.

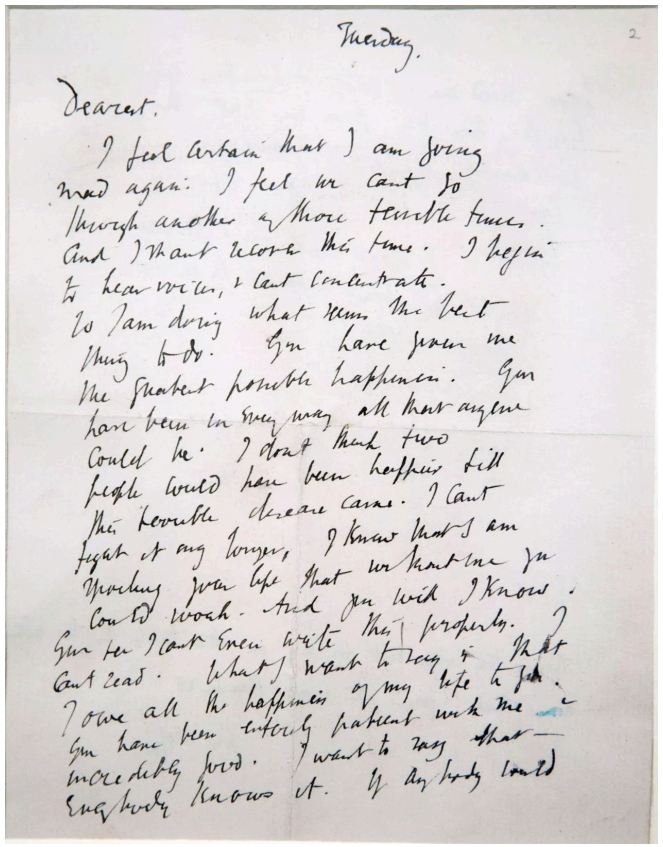
Within each category we distinguish several groups of operations, and for each operation we demonstrate how they can be implemented in R. To provide parsimonious and easy to replicate examples, we have chosen a specific selection of packages that are easy to use and broadly applicable. However, there are many alternative packages in R that can perform the same or similar operations. Due to the open-source nature of R, different people from often different disciplines have worked on similar problems, creating some duplication in functionality across different packages. This also offers a range of choice, however, providing alternatives to suit a user's needs and tastes. Depending on the research project, as well as personal preference, other packages might be better suited to different readers. While a fully comprehensive review and comparison of text analysis packages for R is beyond our scope here—especially given that existing and new packages are constantly being developed—we have tried to cover, or at least mention, a variety of alternative packages for each text analysis operation.[6] In general, these packages often use the same standards for data formats, and thus are easy to substitute or combine with the other packages discussed in this teacher's corner.

Data preparation

Data preparation is the starting point for any data analysis. Not only is computational text analysis no different in this regard, but also frequently presents special challenges for data preparation that can be daunting for novice and advanced practitioners alike. Furthermore, preparing texts for analysis requires making choices that can affect the accuracy, validity, and findings of a text analysis study as much as the techniques used for the analysis (Crone, Lessmann, & Stahlbock, 2006; Günther & Quandt, 2016; Leopold & Kindermann, 2002). Here we distinguish five general steps: importing text, string operations, preprocessing, creating a document-term matrix (DTM), and filtering and weighting the DTM.

Importing text

Getting text into R is the first step in any R-based text analytic project. Textual data can be stored in a wide variety of file formats. R natively supports reading regular flat text files such as CSV and TXT, but additional packages are required for processing formatted text files such as JSON (Ooms, 2014), HTML, and XML (Lang & the CRAN Team, 2017), and for reading complex file formats such as Word (Ooms, 2017a), Excel (Wickham & Bryan, 2017) and PDF (Ooms, 2017b). Working with these different packages and their different interfaces and output can be challenging, especially if different file formats are used together in the same project. A convenient solution for this problem is the `readtext` package(Benoit & Obeng, 2017), that wraps various import packages together to offer a single catch-all function for importing many types of data in a uniform format. The following lines of code illustrate how to read a CSV file with the `readtext` function, by providing the path to the file as the main argument (the path can also be an URL, as used in our example). An



Tuesday.

2.

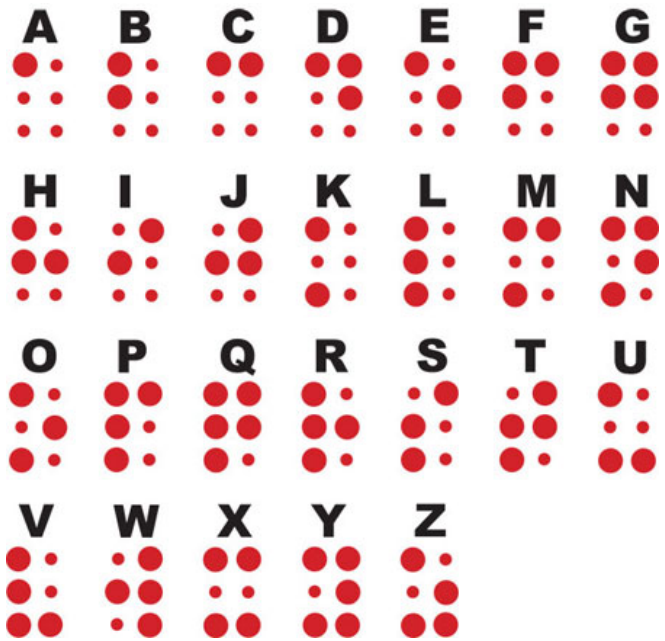
Dearest.

I feel certain that I am going
mad again. I feel we can't go
through another more terrible times.
And I want record this time. I begin
to hear voices, & can't concentrate.
So I am doing what seems the best
thing to do. You have given me
the greatest possible happiness. You
have been in every way all that anyone
could be. I don't think two
people could have been happier till
this terrible disease came. I can't
fight it any longer, I know that I am
spoiling your life that we have you
could work. And you will I know.
You see I can't even write this properly. I
can't read. What I want to say is that
I owe all the happiness of my life to you.
You have been entirely patient with me &
incredibly good. I want to say that —
everybody knows it. If any body could

BUT TEXT CAN ALSO LOOK VERY DIFFERENT

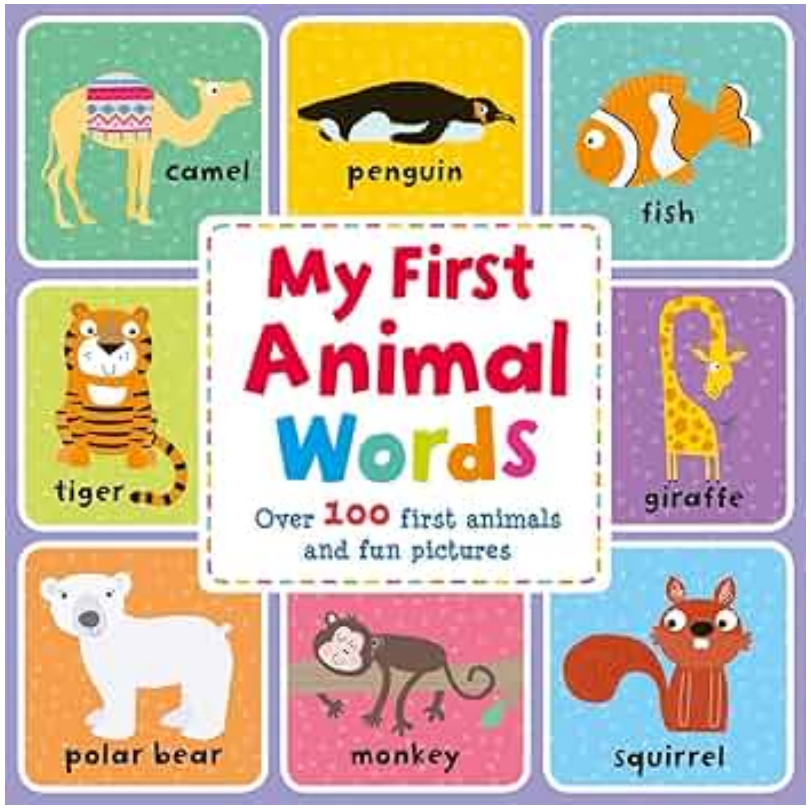
```python
import os, time
x = "Welcome to The Mirror."
y = 0

while y <= len(x):
    os.system("clear")
    print(x[:y])
    time.sleep(0.2)
    y = y+1
time.sleep(2)
x = "You will stay here for as long as you can, you may forfeit when you would like to."
y = 0

while y <= len(x):
    os.system("clear")
    print(x[:y])
    time.sleep(0.2)
    y = y+1
time.sleep(2)
x = "Please stand here, and stare into the mirror."
y = 0

while y <= len(x):
    os.system("clear")
    print(x[:y])
    time.sleep(0.2)
    y = y+1
time.sleep(2)
x = "This is you in the mirror:"
y = 0

while y <= len(x):
    os.system("clear")
    print(x[:y])
```

```html
<div class="navbar navbar-inverse  navbar-fixed-top" role="navigation">
  <div class="container">
    <div class="navbar-header">
      <button type="button" class="navbar-toggle collapsed" data-toggle="collapse" data-bs-toggle="collapse" data-target="#nav
        <span class="icon-bar"></span>
        <span class="icon-bar"></span>
        <span class="icon-bar"></span>
      </button>
      <a class="navbar-brand" href="index.html">Home</a>
    </div>
    <div id="navbar" class="navbar-collapse collapse">
      <ul class="nav navbar-nav">

      </ul>
      <ul class="nav navbar-nav navbar-right">
        <li>
  <a href="content.html">Content &amp;amp; Learning Goals</a>
</li>
<li>
  <a href="overview.html">Overview</a>
</li>
<li>
  <a href="material.html">Slides &amp;amp; Material</a>
</li>
<li>
  <a href="contact.html">Contact</a>
</li>
      </ul>
    </div><!--/.nav-collapse -->
  </div><!--/.container -->
</div><!--/.navbar -->

<div id="header">



<h1 class="title toc-ignore">Computational Analysis of Digital
Communication</h1>

</div>


<p>This page contains all information and materials for the course
"Computational Analysis of Digital Communication (CADC)" developed at
the Vrije Universiteit Amsterdam. In the course, students will learn
about common computational methods in Communication Science. They will
learn how to use the <a href="https://www.r-project.org/">statistical
programming environment R</a> to</p>
<ul>
<li>import various data formats as well as gather data from online
sources,</li>
<li>transform and wrangle data to get it ready for analysis,</li>
<li>perform text classification and text analysis (including dictionary,
classic machine learning approaches, and using large language models),
and</li>
<li>perform advanced statistical analysis.</li>
</ul>
<p><br></p>
<p><img
src="https://socialsciences.nature.com/cdn-cgi/image/quality=90/https://images.zapnito.com/uploads/d39ae7f9b96dbb3ee350deac50
<p><em><a
href="https://socialsciences.nature.com/posts/54262-computational-social-science-heralds-the-age-of-interdisciplinary-science"
source: Sutherland, 2019</a></em></p>
<hr>
<p style="margin-left: auto; margin-right: 0;">
This course is published under the following <a
href="https://github.com/masurp/VU_CADC/blob/main/LICENSE.md">license</a>.
</p>



</div>

<script>

// add bootstrap table styles to pandoc tables
function bootstrapStylePandocTables() {
  $('tr.odd').parent('tbody').parent('table').addClass('table table-condensed');
```



SYMBOLS AND MEANING

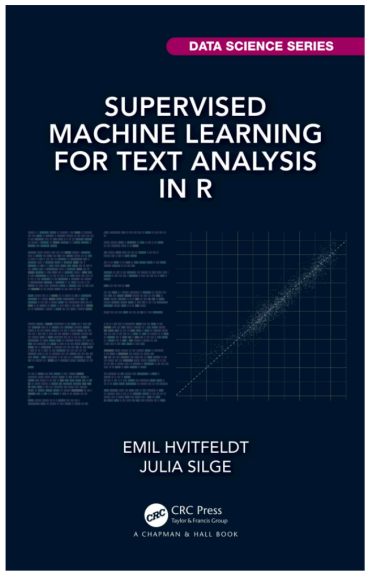
- Text consists of *symbols*
- Symbols by themselves do not have meaning
- A symbol itself is a mark, sign, or word that indicates, signifies, or is understood as representing an idea, object, or relationship
- Symbols thereby allow people to go beyond what is known or seen by creating linkages between otherwise very different concepts and experiences
- Text (a collection of symbols) only attains meaning when interpreted (in its context)
- Main challenge in Automatic Text Analysis: Bridge the gap from symbols to meaningful interpretation



UNDERSTANDING LANGUAGE

> "As natural language processing (NLP) practitioners, we bring our assumptions about what language is and how language works into the task of creating modeling features from natural language and using those features as inputs to statistical models. This is true even when we don't think about how language works very deeply or when our understanding is unsophisticated or inaccurate […] We can improve our machine learning models for text by heightening that knowledge."

Hvitfeldt & Silge, 2021



A SHORT OVERVIEW OF LINGUISTICS

- Each field studies a *different level* at which language exhibits organization
- When we engage in text analysis, we use these levels of organization to create language features (e.g., tokens, n-grams,…)
- In classic machine learning, they often depend on the morphological characteristics of language, such as when text is broken into sequences of characters, words, sentences
- In modern approaches (e.g. using LLMs), it also involves syntactical and pragmatic characteristics. In case of audio-text transformation also phonetical or phonological characteristics!

Subfield	What does it focus on?
Phonetics	Sounds that people use in language
Phonology	Systems of sounds in particular languages
Morphology	How words are formed
Syntax	How sentences are formed from words
Semantics	What sentences mean
Pragmatics	How language is used in context

Hvitfeldt & Silge, 2021 (chap. 1)

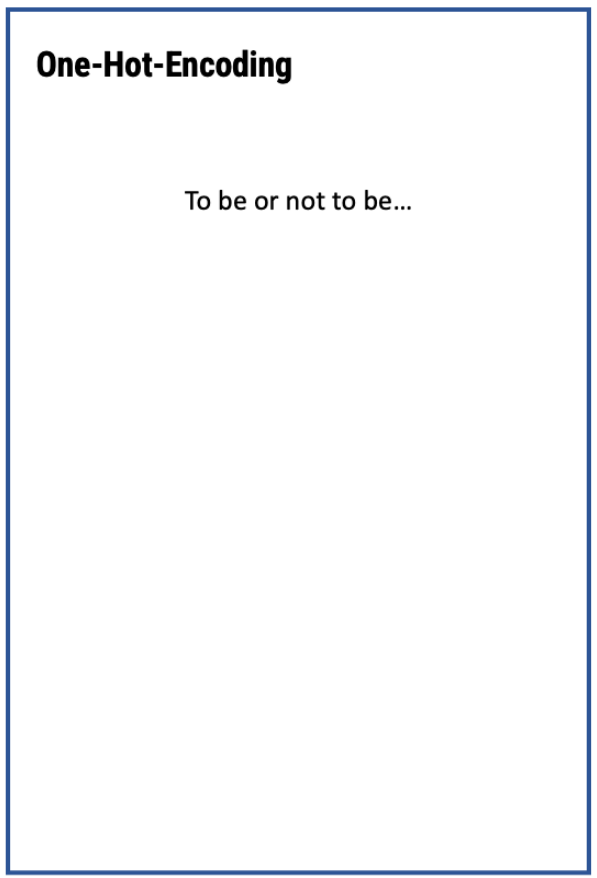
MORPHOLOGY

- When we build text classification models for text data, we use these levels of organization to create natural language features (i.e. predictors for our models)
- What features we extract often depend on the morphological characteristics of language, such as when text is broken into sequences of characters
- Yet, because the organization and the rules of language can be ambiguous, our ability to create text features for machine learning is constrained by the very nature of language

Type of Feature	Example
Sentence	“Include Your Children When Baking Cookies”
Word	“Include”, “Your”, “Children”, “When”, “Baking”, “Cookies”
Bigrams	“Include”, “Your Children”, “When”, “Baking Cookies”
n-grams	“Include Your Children”, “When Baking Cookies”

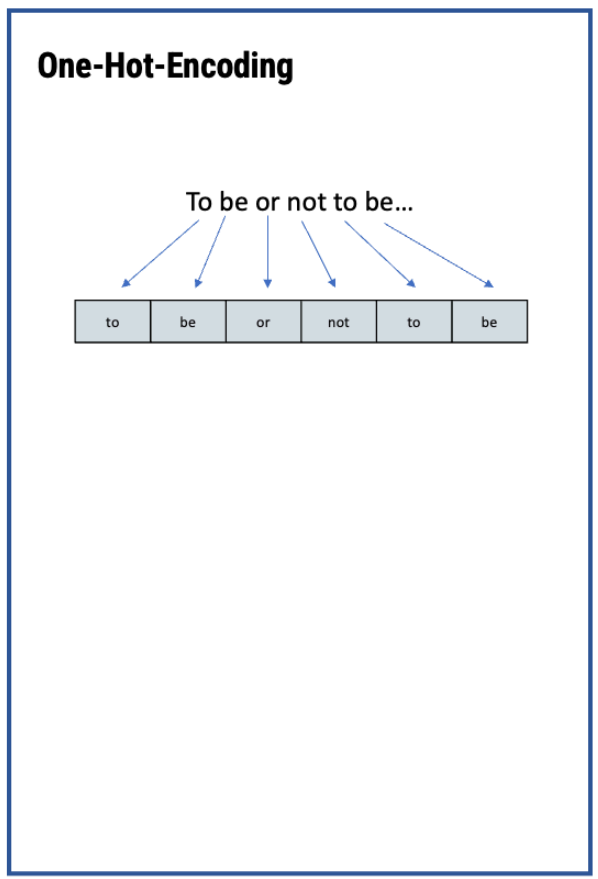
From text to numbers

In automated text classification, we use understandings of morphology (and other fields of linguistics) to break text into tokens and then represent these tokens as numbers, so that a computer can read them:



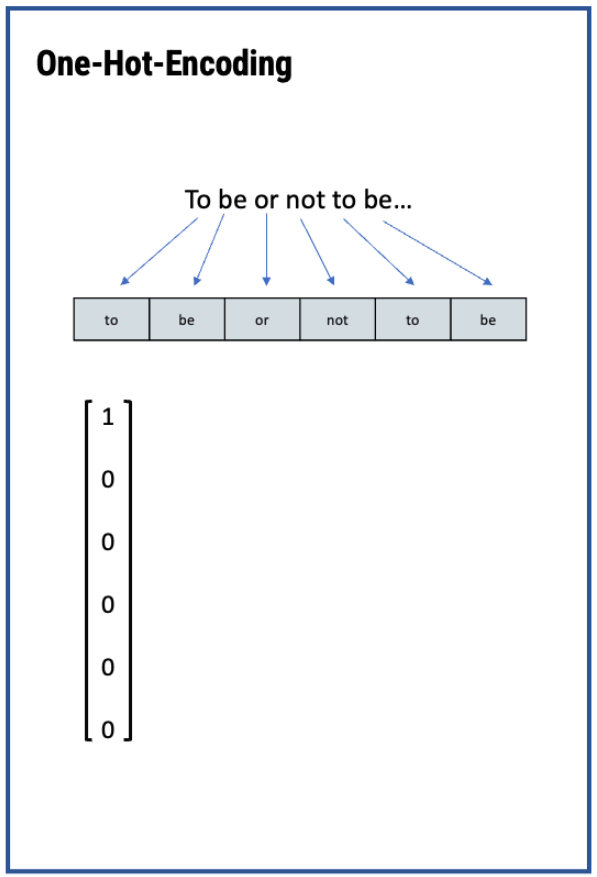
From text to numbers

In automated text classification, we use understandings of morphology (and other fields of linguistics) to break text into tokens and then represent these tokens as numbers, so that a computer can read them:



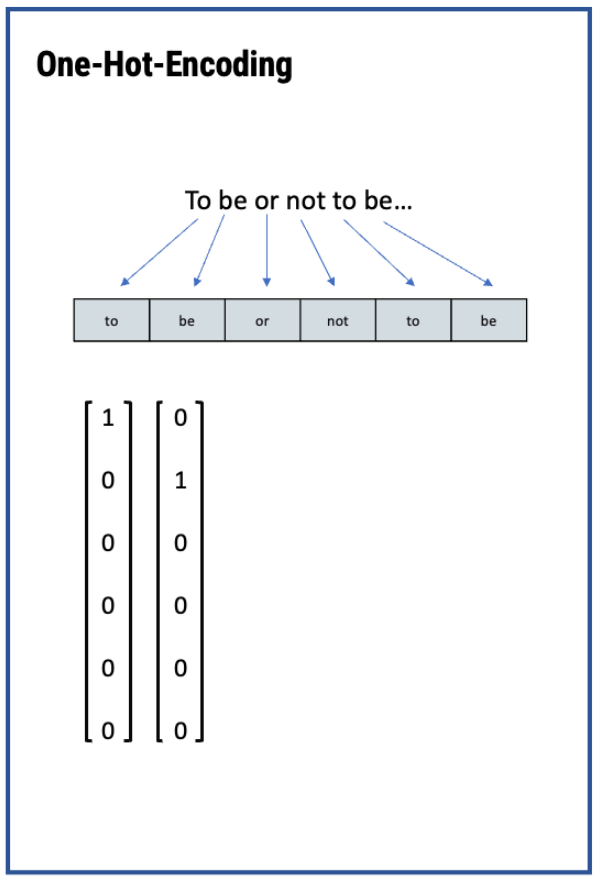
From text to numbers

In automated text classification, we use understandings of morphology (and other fields of linguistics) to break text into tokens and then represent these tokens as numbers, so that a computer can read them:



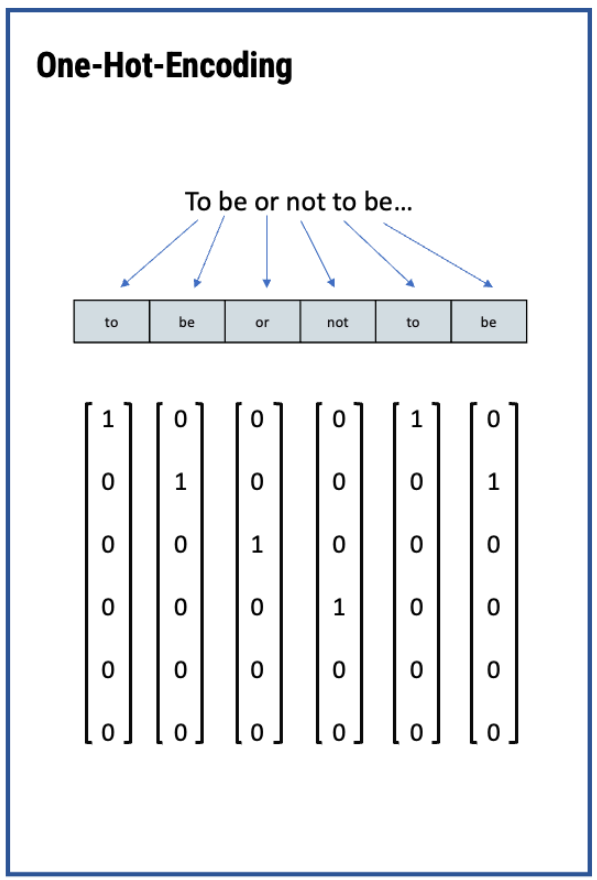
From text to numbers

In automated text classification, we use understandings of morphology (and other fields of linguistics) to break text into tokens and then represent these tokens as numbers, so that a computer can read them:



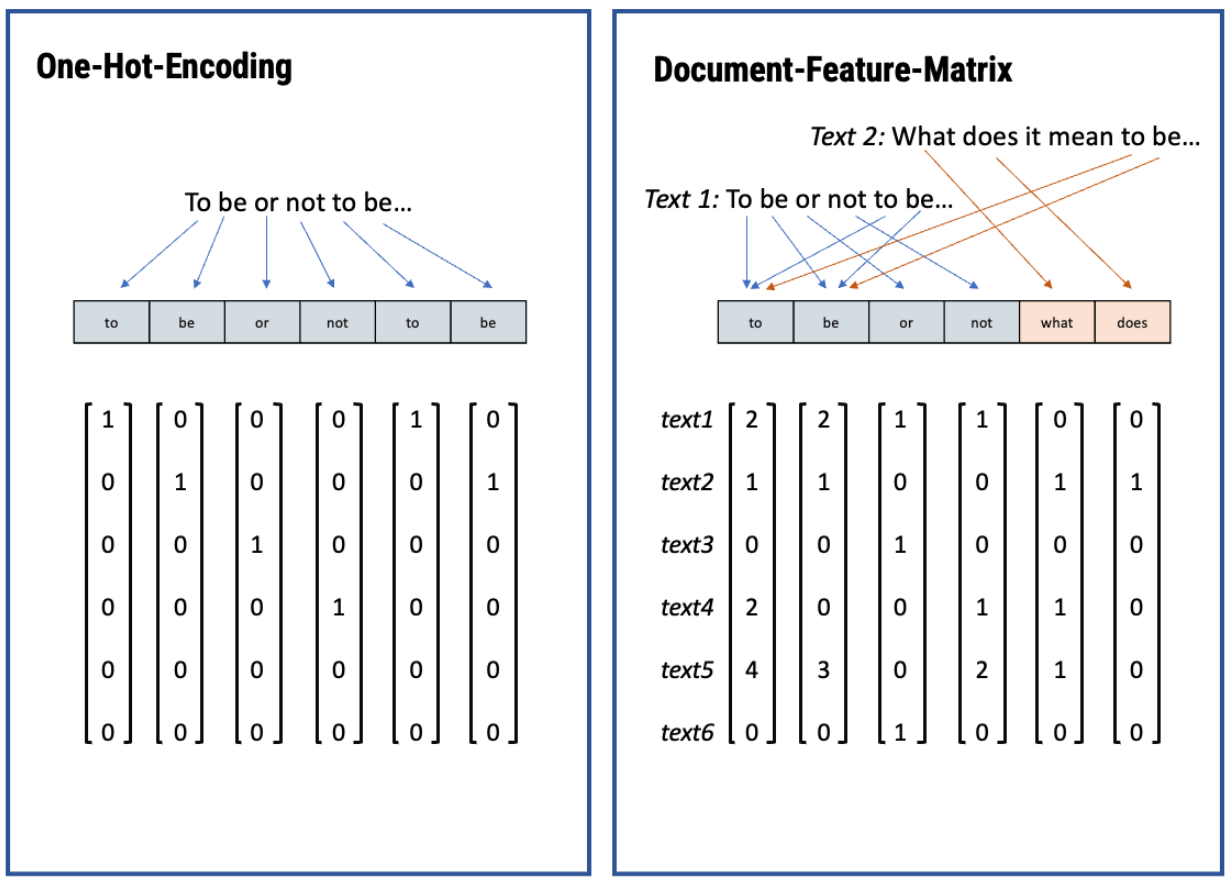
From text to numbers

In automated text classification, we use understandings of morphology (and other fields of linguistics) to break text into tokens and then represent these tokens as numbers, so that a computer can read them:



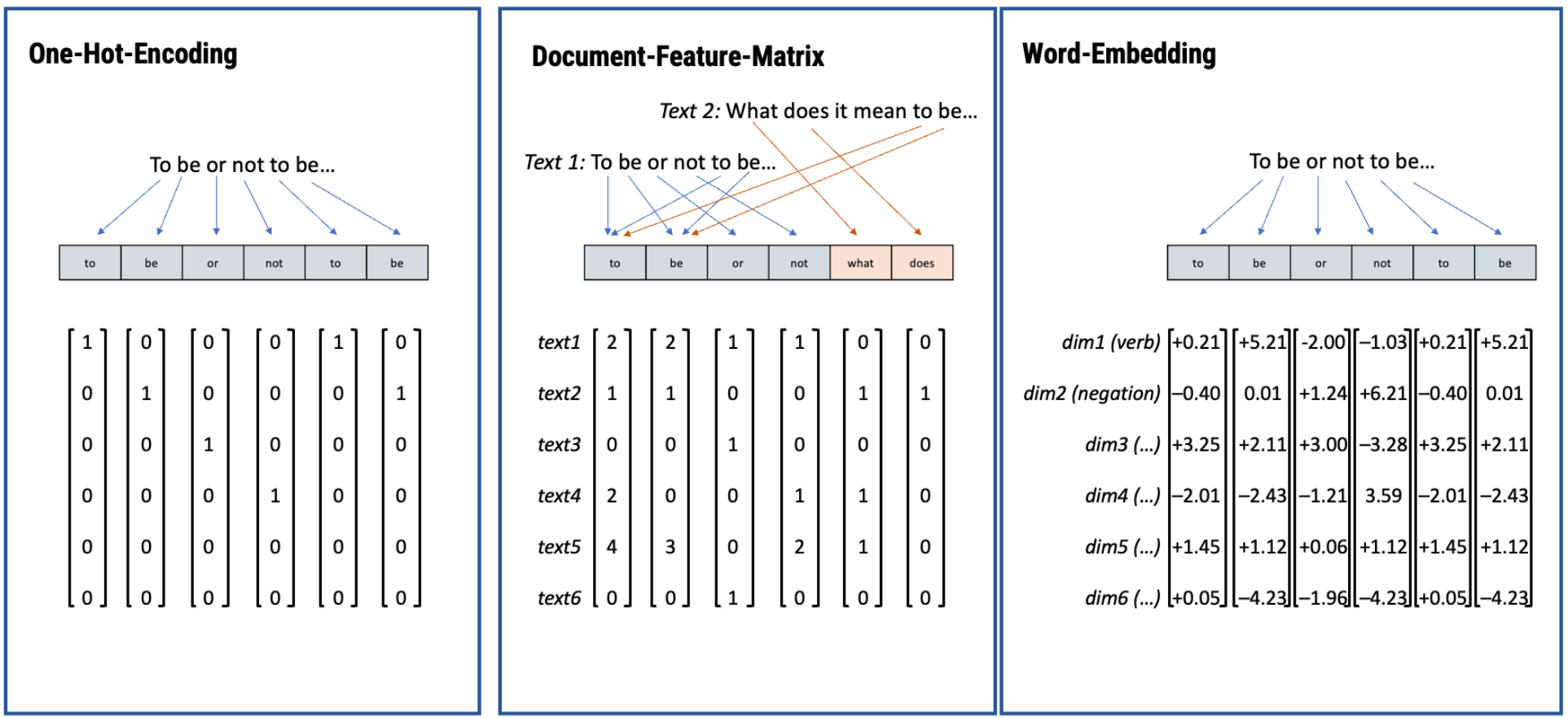
From text to numbers

In automated text classification, we use understandings of morphology (and other fields of linguistics) to break text into tokens and then represent these tokens as numbers, so that a computer can read them:



From text to numbers

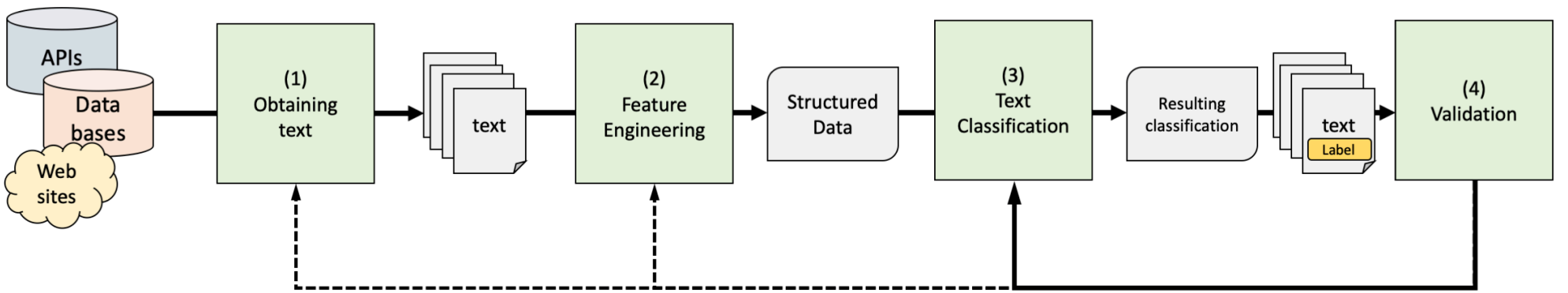
In automated text classification, we use understandings of morphology (and other fields of linguistics) to break text into tokens and then represent these tokens as numbers, so that a computer can read them:



General Text Classification Workflow

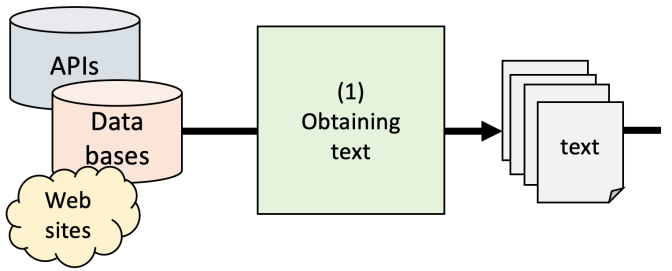
A Framework for this Course

GENERAL TEXT CLASSIFICATION PIPELINE



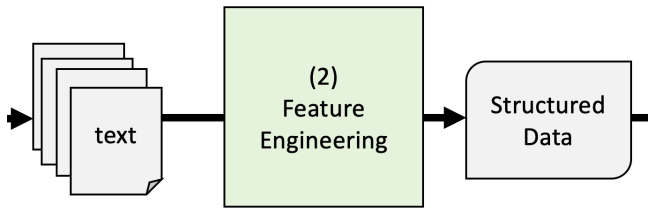
- **General Goal:** to label (or annotate) previously unlabeled text
  - Entails labeling a sentence, paragraph, or entire text (e.g., with the topic, the sentiment,…)
  - Specific methods may differ, but the necessary steps (1-4) usually remain the same
- We will always come back to this general pipeline to make sure we understand the core principles and goals

(1) OBTAINING TEXT



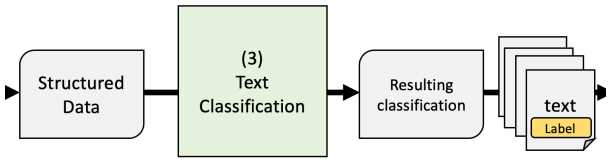
- From publicly available data sets
  - e.g.: Political texts, news from publisher / library
  - Great if you can find it, often not available
- By scraping primary sources
  - e.g.: Press releases from party website, existing archives
  - Writing scrapers can be trivial or very complex depending on web site
  - Make sure to check legal issues
- Via proprietary texts from third parties
  - e.g.: digital archives (LexisNexis, factiva etc.), social media APIs
  - Often custom format, API restrictions, API changes
  - Terms of use not conducive to research, sharing

(2) FEATURE ENGINEERING



- Feature engineering is the process of selecting, manipulating, and transforming raw data into so-called features
- What type of feature engineering is necessary or useful strongly depends on the method used, but may include
  - **feature creation:** breaking down text into the features that we want to analyze (so-called tokens such as words, sentence, bi-grams…)
  - **feature transformation:** involves text cleaning such as stopword removal, stemming, normalization
  - **feature selection:** frequency trimming
  - **creation of structured data:** translating tokens into vectors or numbers (e.g., a document-feature matrix for classic ML approaches or dense vector-matrices for deep learning
- Modern approaches such require less and less manual feature engineering and at times, make it entirely obsolete (because these models automatically “do it”)

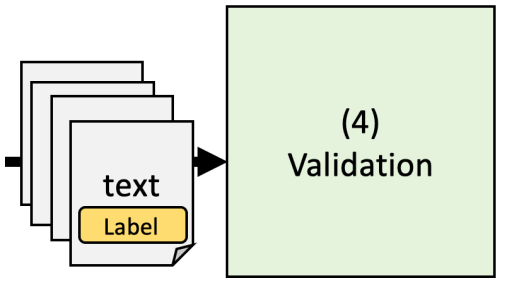
(3) TEXT CLASSIFICATION

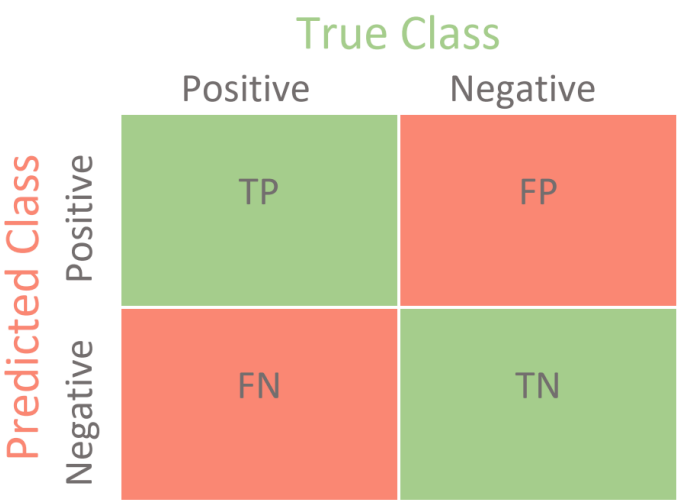


	Rule-Based Analyses	Unsupervised Machine Learning	Supervised Machine Learning	Semi-Supervised Training and Transfer Learning
Description	Code/annotate unlabeled texts based on keywords, expressions or concepts based on pre-defined word lists	Finding clusters or patterns of words that co-occur in unlabeled texts	Training a model on already labeled texts (*training data*), validate it (on labeled *test data*) to then annotate unlabeled text (unlabeled *new data*)	Training a model on a large corpus in a semi-supervised fashion (e.g., word prediction in random word masking tasks) to "learn the language". Resulting model then completes a completely new task (e.g., labeling a text) without further training
Logic	*"if word X occurs, the text means Y"*	*"these words form a pattern, which I think means X"*	*"text X is like other texts in the training data that were labeled negative, so X is probably also negative"*	*"Based on the relationships between words, phrases, and contexts learned from vast amounts of text data, the model understands that text X shares similarities with texts that tend to be labeled as negative, so X is likely negative."*
Meaning assignment	Meaning of text-word associations is assigned by researcher a priori	Meaning is assigned a posteriori by the researcher by interpreting the resulting clusters of words	Meaning of text-word-associations is generalized from human coding of the training material	Meaning is automatically provided by the model
Examples	• Dictionary approaches	• Unsupervised topic modeling	• Training a text classifier using algorithms such as Naïve Bayes, SVM, neural networks	• Text classification using pre-trained large language models (e.g., GPT, Llama, Claude…)

(4) VALIDATION

- Many text analysis processes are 'black boxes'
  - even manual coding
  - dictionaries are ultimately opaque
  - complex algorithms cannot be deciphered
- Computer does not 'understand' natural language
  - It just predicts labels based on features
  - False-positive and false-negatives occur
- We need to prove that the analysis is valid
  - Validate by comparing text analyis output to a known good
  - Reference: often manual annotation of a 'gold standard'





TIMELINE OF NATURAL LANGUAGE PROCESSING

1763 — BAYES' THEOREM

- Reverend Thomas Bayes's work is published two years after his death
- It presents major underpinnings of Bayes theorem that is still used in Machine Learning (and Bayesian Statistics) today

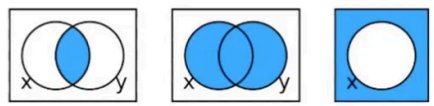
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

1842 — FIRST ALGORITHM?

- Ada Lovelace develops the first algorithm "Note G"
- She is widely seen as the first programmer
- She suggested that a machine could take advantage of computation in order to deal with the realms of music, graphics, and language

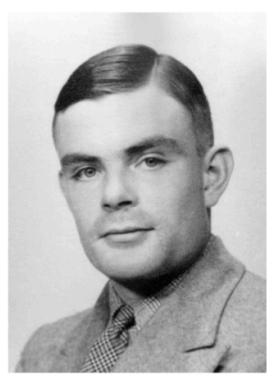
1854 — BOOLEAN LOGIC

- George Boole invents the Boolean Algebra
- It has been fundamental in the development of digital electronics, and is provided for in all modern programming languages



1934 — MACHINES THAT LEARN?

- Alan Turing proposes a machine that can learn
- The Universal Turing Machine is an abstract computing machine consisting of a limitless memory and a scanner that moves back and forth through the memory, symbol by symbol, reading what it finds and writing further symbols



TIMELINE OF NATURAL LANGUAGE PROCESSING

1949

WEAVER'S MEMORANDUM

- most influential publication in the early days of machine translation
- formulated goals and methods before most people had any idea of what computers might be capable of
- direct stimulus for the beginnings of research into natural language processing

1957

GRAMMAR THEORIES

- early work in MT was very simple: dictionary-lookup of appropriate words for translation
- Noam Chomsky developed a rule based-system of syntactic structures
- brought insight into how mainstream linguistics could help machine translation

1960s

ELIZA, ALPAC REPORT & AI WINTER

- The most famous prototype is perhaps ELIZA, designed to replicate the conversation between a psychologist and a patient
- The report by the Automatic Language Processing Advisory Committee (ALPAC), a committee established in 1964 by the United States, led to significant cuts in funding

1970s-80s

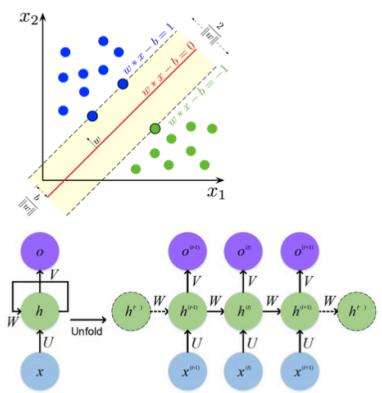
CONCEPTUAL ONTOLOGIES

- The 1970s brought new ideas into NLP
- Development of conceptual ontologies, i.e. sets of concepts and categories that shows their properties and the relations between them
- structured real-world information into computer-understandable data

1990

STATISTICAL MODELS

- statistical models came as a revolution in NLP (SVM, regression-based approaches)
- replacing most natural language processing systems based on complex sets of hand-written rules
- Recurrent neural networks and long short-term memory networks

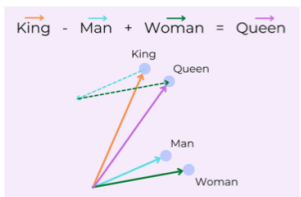


TIMELINE OF NATURAL LANGUAGE PROCESSING

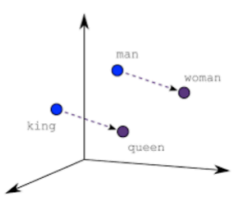
2000s

LANGUAGE MODELING, WORD EMBEDDINGS

- neural networks are used for learning the distribution of language words and predicting the next words in a text given the previous ones
- Start of representing words with dense vectors of numbers
- 2012: First unsupervised neural network by Google



2013



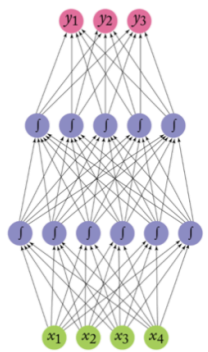
WORD2VEC

- first algorithm capable of learning word embedding efficiently
- possible to do mathematical operations with them that represent semantic operations
- using pre-trained embeddings as features for ML models improves performance in downstream tasks

2014-2016

LSTM RNNs, CNNs, ENCODER-DECODER

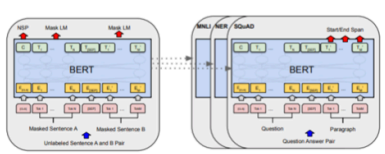
- Most use models are long short-term memory recurrent neural networks, and convolutional neural networks
- sequence-to-sequence learning, a general end-to-end approach for mapping one sequence to another using a neural network
- the first encodes the entire input into a vector, the second decodes the output from the vector



2017

ATTENTION AND TRANSFORMERS

- Introduction of the principle of attention
- Rise of Transformers, which are still used heavily today
- First generative pre-trained transformer by OpenAI in 2018!



2020s

PRE-TRAINED MODELS AND LARGE LANGUAGE MODELS

- Transformers were trained in a self-supervised way on large data sets to become "language models"
- These models can then be reused for different task after only small adaption traning (fine-tuning)

2022

GPT, BART and Co

- Increasingly larger language models that are trained on the "Internet"
- November 2022: Launch of chatGPT and wide-spread use of LLMs
- AI revolution?

ChatGPT

Ethics of Computational Communication Research

Ethical challenges in using computational methods

ETHICAL CHALLENGES IN COMPUTATIONAL RESEARCH (1)

1. **Privacy and Data Security:** A key ethical concern is the handling of personal data, including the risk of data breaches and misuse. When using proprietary software/algorithms, should we share personal data with the company behind it (e.g., OpenAI, Google)?
2. **Bias and Fairness:** Computational methods can inherit biases from the data they are trained on. This can result in unfair or discriminatory outcomes, especially when later applied to contexts like hiring, lending, and criminal justice. Addressing these biases remains an ongoing challenge. Validation is of utmost importance!
3. **Transparency and Accountability:** Many computational algorithms are complex and not easily understandable by humans. This lack of transparency can make it difficult to hold individuals or organizations accountable for the decisions made by these algorithms. Lack of explainability can lead to a lack of trust.
4. **Social Manipulation:** Computational methods can be used to manipulate public opinion or behavior, often in unethical ways (see Kramer et al.). This includes spreading misinformation or conducting social experiments without consent and debriefing.

ETHICAL CHALLENGES IN COMPUTATIONAL RESEARCH (2)

5. **Environmental Impact:** The massive computational power required for some methods, such as training deep learning models, can have a significant environmental impact, contributing to concerns about energy consumption and climate change.
6. **Ownership and Intellectual Property:** Questions of intellectual property rights, data and model ownership, and access to computational methods can raise more ethical issues. Some may argue that access to certain methods should be open to all (see also point 3 transparency), while others may seek to protect intellectual property (among other things, to e.g., prevent misuse).
7. **Ethical AI:** In the field of artificial intelligence (AI) and machine learning, ethical concerns often revolve around the development of AI systems that can make autonomous decisions. Ensuring these decisions align with human values is a significant challenge. Ethical dilemmas arise around the degree of control we should and are able to maintain over these systems and who should be responsible for their actions.

see also van Atteveldt & Peng, 2018

GENERAL ETHICAL GUIDELINES

Salganik (2013) proposes the following four ethical guidelines for evaluating the ethical nature of a computational study. They are loosely based on the Belmont Report (1978), which is a fundamental document in prescribing ethical guidelines for the protection of human subjects of research:

- **Respect for persons:** Treating people as autonomous and honoring their wishes
- **Beneficence:** Understanding and improving the risk/benefit profile of a study
- **Justice:** Risks and benefits should be evenly distributed among participants/research subjects
- **Respect for law and public interest**: Respect privacy, copyright, and access rights

Salganik, 2018, chap. 6

ETHICAL PRINCIPLES FOR CCR

#	Principle	Important aspects
1.	**Ensure privacy and data security**	• Prioritize informed consent and data privacy, if possible obtain permission for data use. • Most importantly, ensure the protection of individuals' sensitive information via strict data management and protection rules • When using proprietary software and algorithms, consider risks of sharing subject data with the companies operating them (e.g., OpenAI, Google, etc.)
2.	**Prevent bias and ensure fairness**	• Ensure that text classification models are designed to be fair, with strategies in place to identify and mitigate biases in the data and algorithms. • Regularly assess the performance of text classification models and be committed to refining them to enhance accuracy and fairness. Overall, put a strong emphasis on validation!
3.	**Strive for transparency**	• Use transparent and interpretable text classification models if possible, allowing users to understand how decisions are made and to address concerns regarding model opacity. • Share more complex models for re-use and evaluation.
4.	**Be mindful about conducting social experiments**	• If possible, obtain informed consent and ensure thorough debriefing • If that is not possible or undesired, thoroughly assess potential negative consequences of the study. Only if they are acceptable or not different from everyday use of technology, the study may be ethical to conduct (think about psychological consequences, discrimination, political implications…)
5.	**Reduce environmental impact**	• A lot of task do NOT require the use of computationally extensive approaches (e.g., the fine-tuning of large language models); test performance on smaller subsets before wasting energy on using an overly complex approach for a simple task • Assess the impact of your research: Is it worth using this much energy in light of expected impact?
6.	**Ensure cross-cultural sensitivity**	• Acknowledge and respect cultural differences in language and context, adapting text classification models to be sensitive to various cultural norms and nuances.
7.	**Engage with peers, subjects, community and stakeholders**	• Engage with relevant communities and stakeholders to gather feedback, understand concerns, and involve them in shaping the development and application of text classification methods.

Formalities

How this course is going to work?

TEACHERS

prof. dr. Wouter van Atteveldt	Gianni Quaedvlieg	dr. Kasper Welbers	dr. Sophia Gil-Clavel	dr. Alberto López Ortega
				
Lecturer & Course Coordinator	Teacher & Workgroup Coordinator	Teacher	Teacher	Teacher

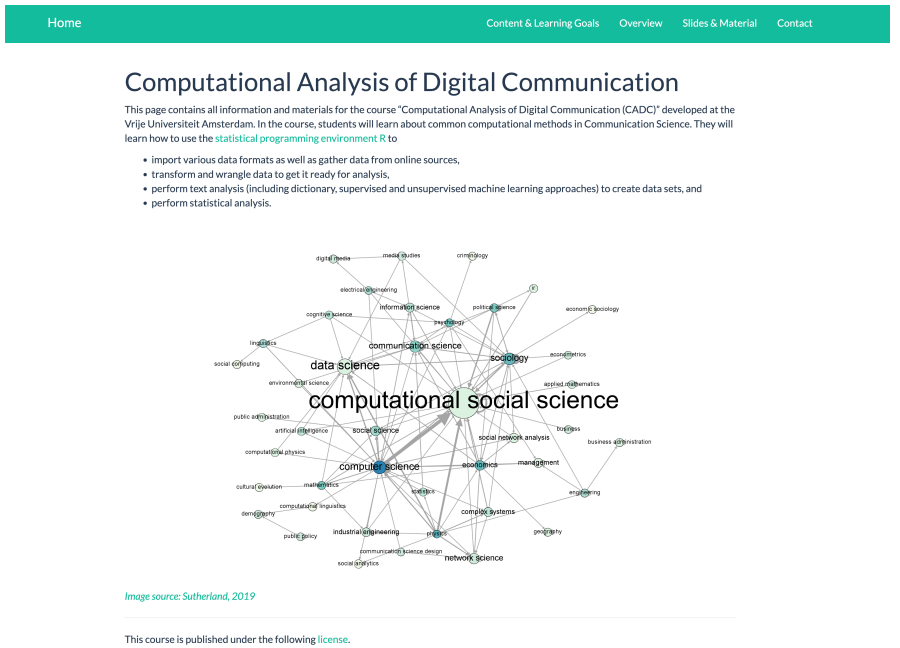
INFORMATION AND MATERIALS

The major hub for this course is the following website:

- https://vanattevelt.com/VU_CADC/

Communication and assignment submission via Canvas:

- S_CADC: Computationele analyse van digitale communicatie



COURSE SCHEDULE

Day	Week	Type	Content	Assignment
Part I: Introduction and Automated Text Analysis				
Monday	1	Lecture	Introduction to Computational Methods in Communication Science	
Tuesday	1	Practical session	Data Wrangling using the tidyverse and tidytext	
Thursday	1	Practical session	Exploratory Data Analysis and Data Visualization	Homework Assignment 1
Monday	2	Lecture	Automated Text Analysis and Dictionary Approaches	
Tuesday	2	Practical session	Basic Text Analysis using tidytext	
Thursday	2	Practical session	Dictionary Approaches using tidytext	Homework Assignment 2
Part II: Text Classification using Machine Learning				
Monday	3	Lecture	Text Classification Using Machine Learning	
Tuesday	3	Practical session	Supervised text classification using Naive Bayes and neural networks	
Thursday	3	Practical session	Supervised text classification with wordembeddings and neural networks	Homework Assignment 3
Monday	4	Lecture	Transformers and Large Language Models	
Tuesday	4	Practical session	Zero-Shot Classification Using Transformers/GPT/llama	
Thursday	4	Practical session	Few-Shot Classification Using GPT/llama	Homework Assignment 4

COURSE SCHEDULE

Exam week				
Friday	5	Exam	Multiple-Choice Exam (Content of Part I & II)	
Part III: Practical Group Projects				
Monday	6	Lecture	Summary and Introduction to Group Projects	
Tuesday	6	Practical session	Meeting with supervisor	
Thursday	6	Practical session	Meeting with supervisor	
Monday	7	No lecture		
Tuesday	7	Practical session	Meeting with supervisor	
Thursday	7	Practical session	Meeting with supervisor	
Tuesday	8	Conference	Presentation of Group Projects	10-min talk

ATTENDANCE

You will realize that this course has a comparatively **steep learning curve**. We will learn about research papers and recreate their analyses in R. It is thus generally recommended to follow all lectures and practical sessions! **BUT:** Despite some initial challenges, you will also experience a lot of self-efficacy: Learning R and computational methods is very rewarding and at the end, you can be proud of what you have achieved!

- Attendance during the regular lectures is highly recommended (this is the content for the exam).
- Attendance of the practical sessions is **mandatory**.
- One absence from one of the workgroup sessions, for serious health, family, or work reasons, can be excused if the instructor is advised **in advance**.

EXAM

After the first two cycles, there will be a written exam (**40% of the final grade**):

- Exam questions will be based on **all material** discussed in the first two cycles, including lecture content, class materials, and required readings

HOMEWORK ASSIGNMENTS

After each week, students are required to hand in a “homework assignment”, which represents a practical application of some of the taught analysis methods (e.g. with a new data set, specific research question) (**30% of the final grade**):

- Assignment 1 and 3 are pass/fail; assignment 2 and 4 are graded
- All assignments are in teams of two students –> Create a team on Canvas before submitting!
- Each week’s assignment requires students to apply the methods they have learned to a new data set.
- Students will receive an RMarkdown template for their code and the respective data set(s), explanations for how to use these templates will be providee in the practical sessions and via Canvas.
- Students are required to hand in the **RMarkdown file** (.rmd) and a compiled **html document** (.html) on Monday the week after. All homework assignments must be submitted to pass!

GROUP PRESENTATION

In the third cycle, students will be assigned to small working groups in which they independently conduct a research project. A final poster presentation in week 8 will be graded per group (**30% of the final grade**).

Students are required to hand in the poster (as PDF) and analyses (again .rmd and .html) beforehand

Questions?

If anything is unclear, do ask now.

REQUIRED READING

- Kramer, A. D. I., Guillory, J. E, & Hancock, J. (2014). Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks. Proceedings of the National Academy of Sciences, 111(24), 8788-8790.
- Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. Communication Methods and Measures, 12(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

(available on Canvas)

REFERENCES

- Andrich, A., Bachl, M., & Domahidi, E. (2023). Goodbye, Gender Stereotypes? Trait Attributions to Politicians in 11 Years of News Coverage. Journalism & Mass Communication Quarterly, 100(3), 473-497. https://doi-org.vu-nl.idm.oclc.org/10.1177/10776990221142248
- Blumenstock, J. E., Cadamuro, G. , and On, R. (2015). Predicting Poverty and Wealth from Mobile Phone Metadata. Science, 350(6264), 1073–6. https://doi.org/10.1126/science.aac4420.
- Haim, M., Scherr, S., & Arendt, F. (2021). How search engines may help reduce drug-related suicides. Drug and Alcohol Dependence, 226(108874). https://dx.doi.org/10.1016/j.drugalcdep.2021.108874
- Jacobi,C., van Atteveldt, W. & Welbers, K. (2016) Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, 4(1), 89-106, DOI: 10.1080/21670811.2015.1093271
- Jürgens, P., Meltzer, C., & Scharkow, M. (2021, in press). Age and Gender Representation on German TV: A Longitudinal Computational Analysis. Computational Communication Research.
- Kramer, A. D. I., Guillory, J. E, & Hancock, J. (2014). Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks. Proceedings of the National Academy of Sciences, 111(24), 8788-8790.
- Parry, D. A., Davidson, B. I., Sewall, C. J., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. Nature Human Behaviour, 5(11), 1535-1547.
- Salganik, M. J. (2018). Bit by Bit: Social Research in the Digital Age. Princeton University Press.
- Scharkow, M. (2016). The accuracy of self-reported internet use—A validation study using client log data. Communication Methods and Measures, 10(1), 13-27.
- Simon, M., Welbers, K., Kroon, A. C. & Trilling, D. (2022). Linked in the dark: A network approach to understanding information flows within the Dutch Telegramsphere. Information, Communication & Society, https://doi.org/10.1080/1369118X.2022.2133549

REFERENCES

- Thompson, T. (2010). Crime software may help police predict violent offences. The Observer. Retrieved from https://www.theguardian.com/uk/2010/jul/25/police-software-crime-prediction
- Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. Communication Methods and Measures, 12(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084
- Wimmer, A. & Lewis, K., (2010). Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook. American Journal of Sociology, 116(2), 583-642.

Thank you for your attention!