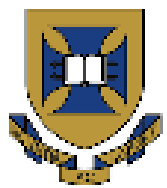
Best Practices in and a Practical Introduction to Text Analysis

Martin Schweinberger
The University of Queensland
m.schweinberger@uq.edu.au



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Today's topics

Theoretical background

- Why Text Analysis (TA)?
- What is TA?
- Why do we need best practices?
- Best practices in TA
- LADAL (Language Technology and Data Analysis Lab.)

Hands-on session

- Creating concordances (KWIC displays)
- Visualizing word frequencies
- Collocations
- Sentiment Analysis

Before we begin...

- Open R Studio (if R Studio is not yet installed, go to your favorite search engine, type "`r studio download`" in the search window, open the first link, and follow the installation instructions).
- Open a new script in R Studio (File → New File → R Script)
- Go to `https://slcladal.github.io/textanalysis.html`
- Copy and paste the code in the section "Preparation" into the R script in R Studio and run it by executing "`Run`" above the script window.

Best practices in and a practical introduction to TA

This session introduces and discusses the concept of "best practices" in Text Analysis and exemplifies how best practices can be implemented when working with textual data.

The first part of the session showcases best practice procedures and principles in text analysis that aim to guarantee transparency, replicability, and reproducibility of research practices. As Text Analysis is still a relatively new approach to analysing textual data, we are in a position to establish communal best practices while this field is still emerging.

The second half of the session serves as a practical introduction to basic examples and methods of Text Analysis. We will introduce methods such as; concordancing and visualizations of word frequencies, and explore collocations using the open source and freely available environment R.

The practical part of this session introduces and makes use of online materials provided by the Language Technology and Data Analysis Laboratory (LADAL).

The course is designed for PhD candidates or researchers in the early stages of their research and anyone interested in working with text based data sets.

Why Text Analysis?

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

A Computational Economy in a Digital Era

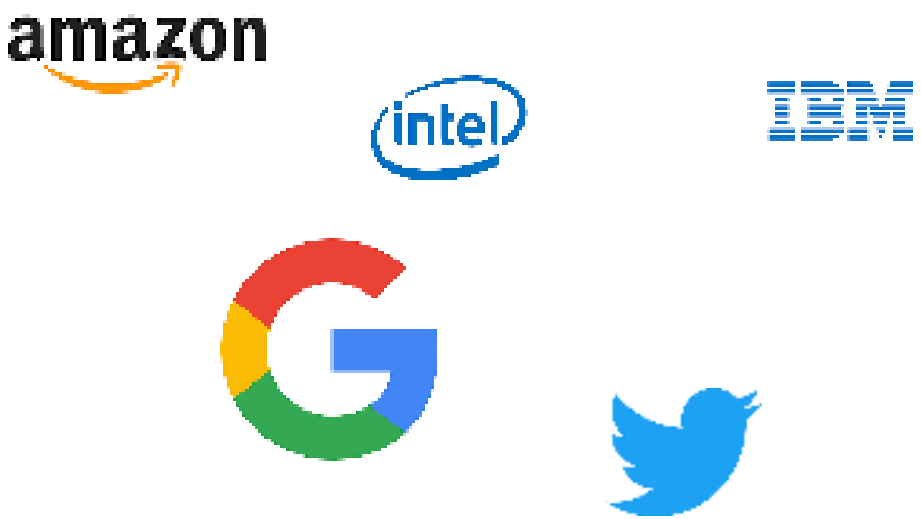
Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

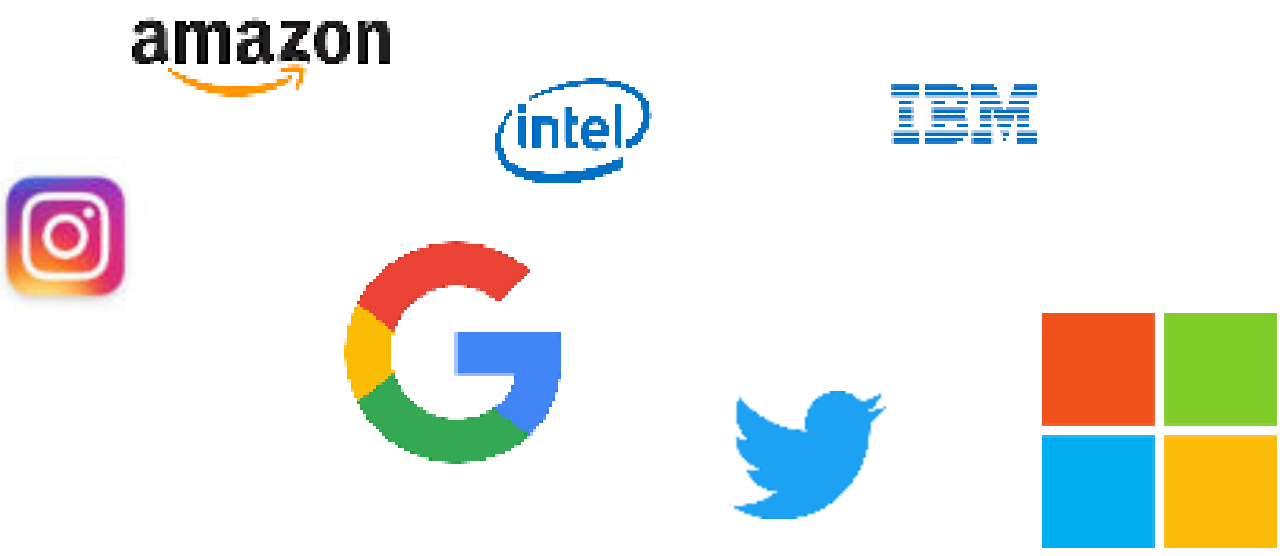


A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

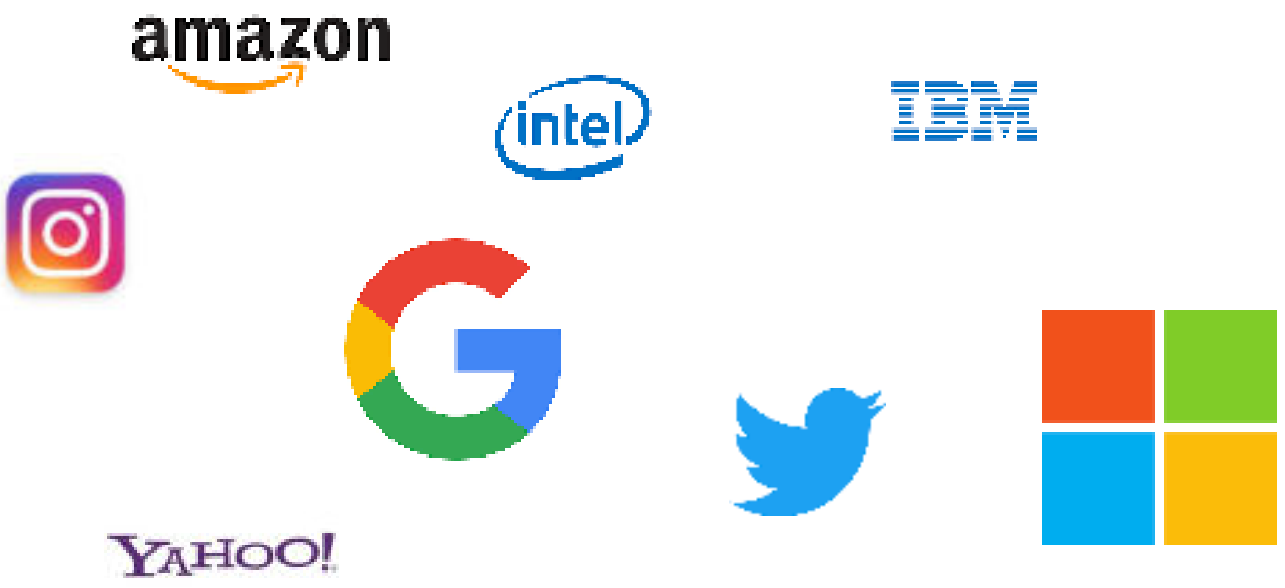
A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.



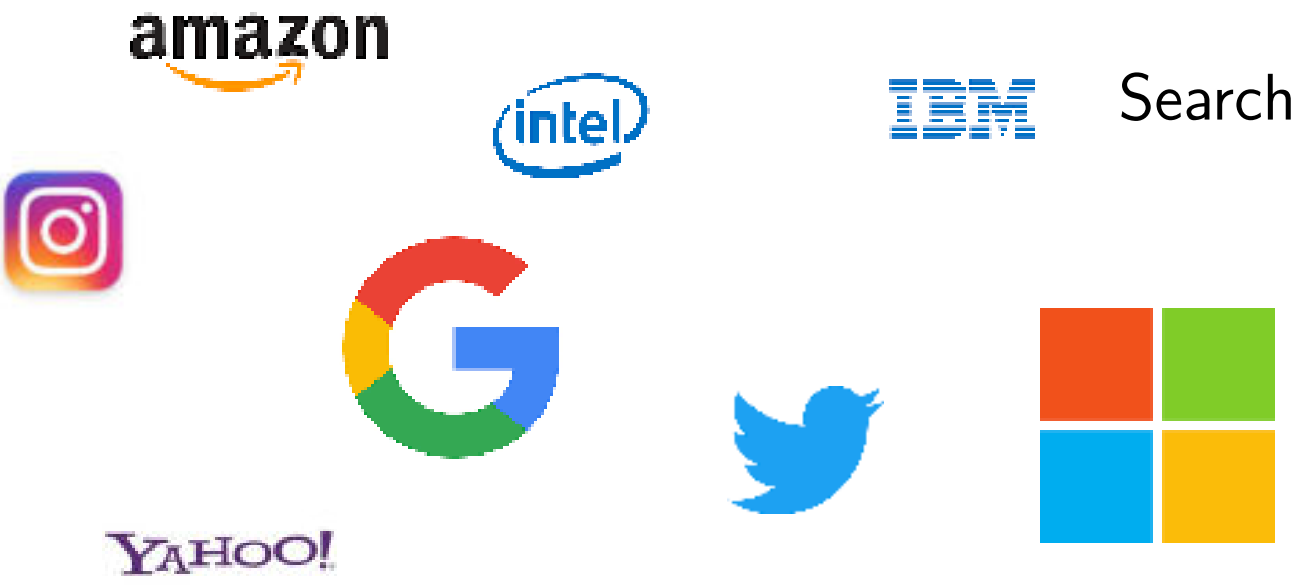
A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.



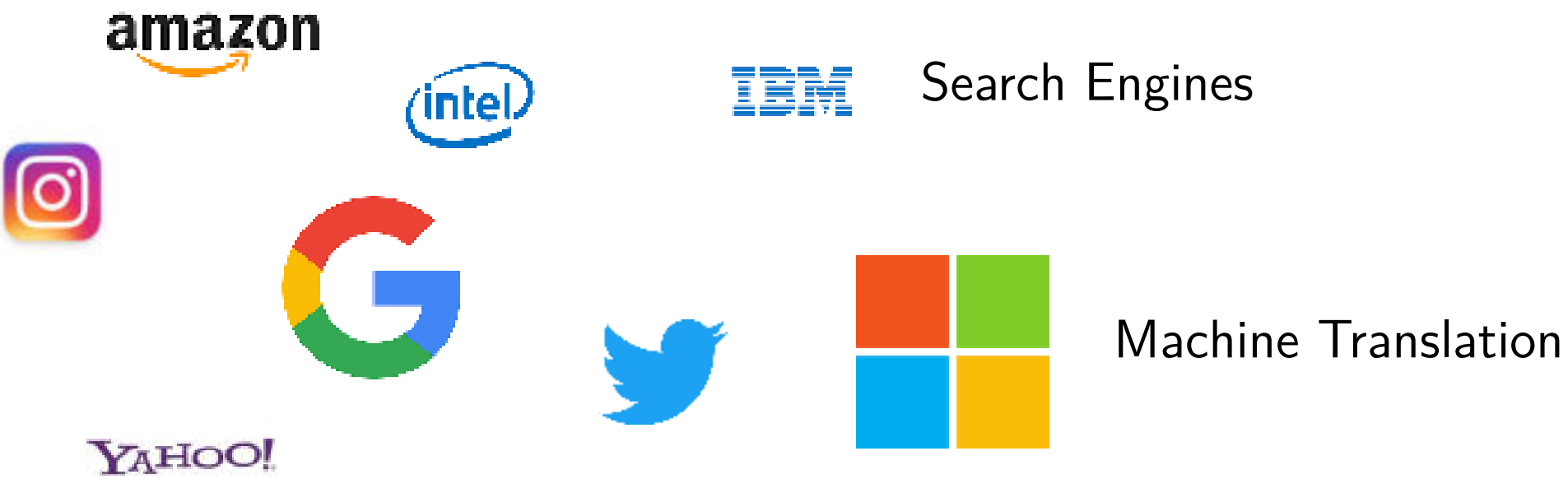
A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.



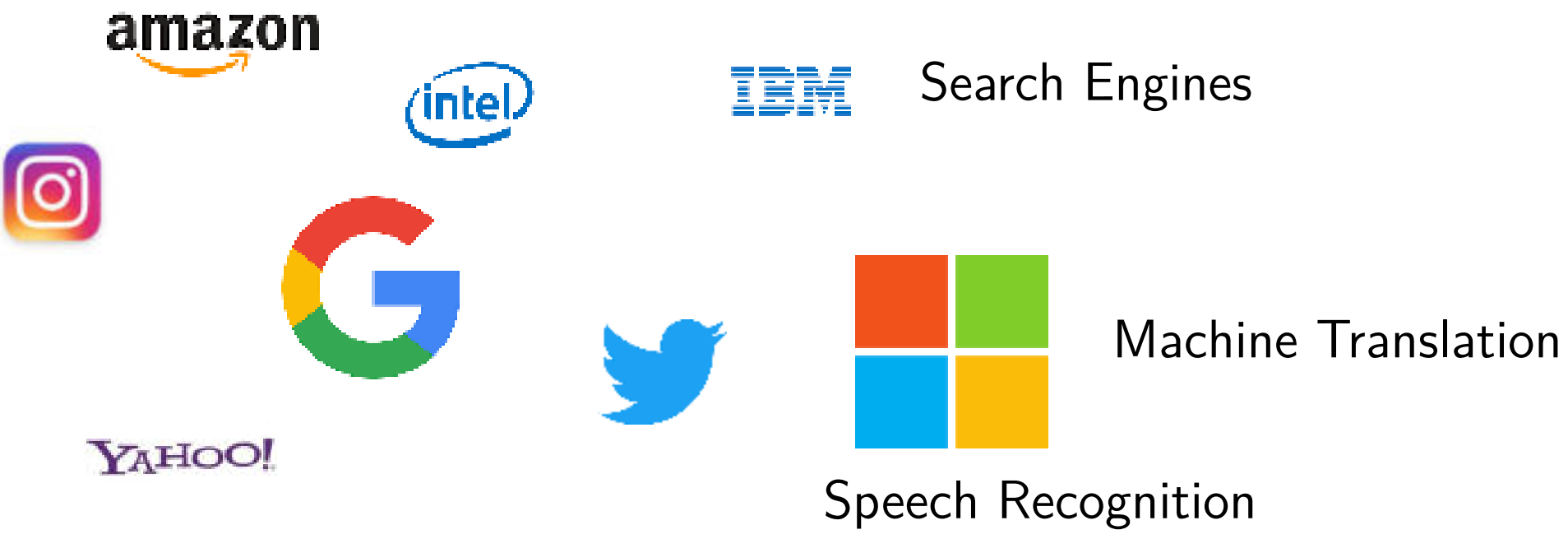
A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.



A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

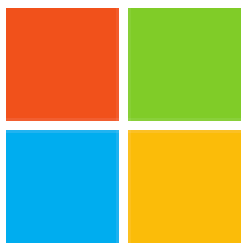


A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

Named Entity Recognition

Search Engines



Machine Translation

Speech Recognition

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

Named Entity Recognition

Search Engines

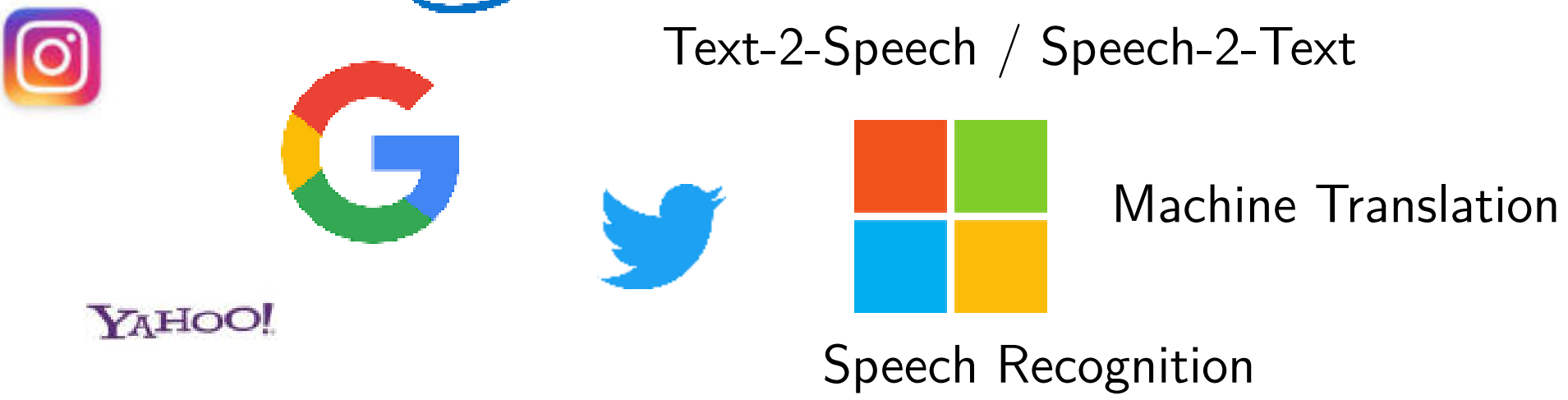
Text-2-Speech / Speech-2-Text

Machine Translation

Speech Recognition

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.



A Computational Economy in a Digital Era

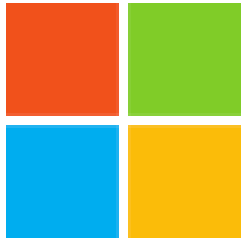
Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

Topic Modeling

Named Entity Recognition

Search Engines

Text-2-Speech / Speech-2-Text



Machine Translation

Spelling Correction

Speech Recognition

A Computational Economy in a Digital Era

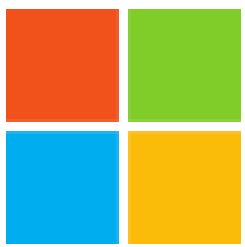
Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.

Topic Modeling

Named Entity Recognition

Search Engines

Text-2-Speech / Speech-2-Text



Sentiment Analysis

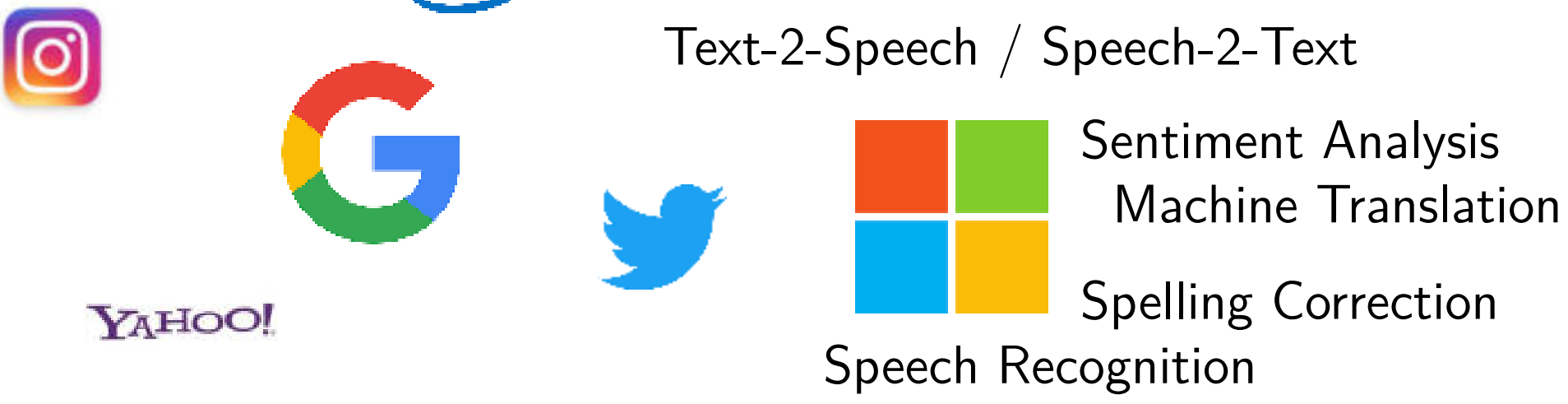
Machine Translation

Spelling Correction

Speech Recognition

A Computational Economy in a Digital Era

Computational approaches to processing and transforming, analyzing and visualizing data are becoming increasingly prevalent.



Humanities and Computing

The digital revolution has led to an ever-increasing availability of textual data in large quantities

- Google indexes over 130 trillion pages (Aug 2010) just over half of which are in English
- Project Gutenberg offers over 59,000 free eBooks.

Vast potential for the humanities by extending computational methods (TA/NLP) to humanties research

- Being able to use Big Data (e.g. mega corpora)
- Finding and visualizing patterns that cannot be detected by traditional means
- Enables multivariate analyses $\rightarrow$ which factors are relevant, and what is their relative impact?

What is Text Analysis?

Definitions and Related Concepts

Text Analysis (TA)

- TA (also Text Analytics) is the process of deriving meaningful information from text data using computation.

Natural Language Processing (NLP)

- The goal of TA is to turn text into data for analysis via NLP which is that part of Computer Science, Machine Learning, and Artificial Intelligence which deals with human language

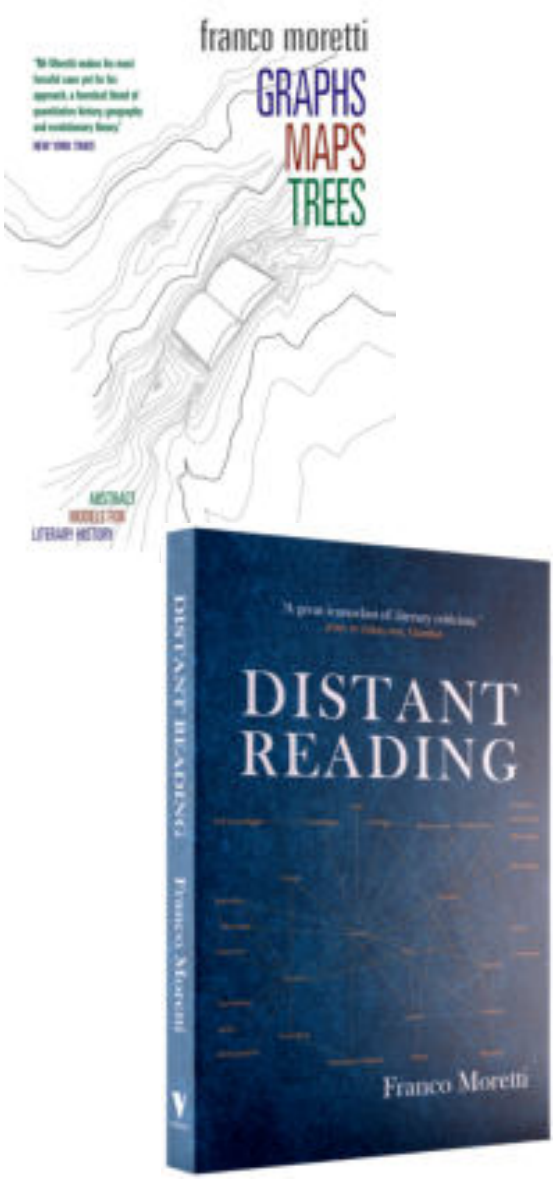
Distant Reading (DR)

- DR is applying TA methods to find patterns in a large set of texts that would remain undetected if only few texts are read closely.

Origin of TA and DR

TA and DR was popularized by Franco Moretti, an Italian literary historian, who applied quantitative methods in literary studies.

In 2005, Moretti published *Graphs, Maps, Trees: Abstract Models for Literary History* and in 2013 *Distant Reading* which introduced text analytic methods into the classical humanities.



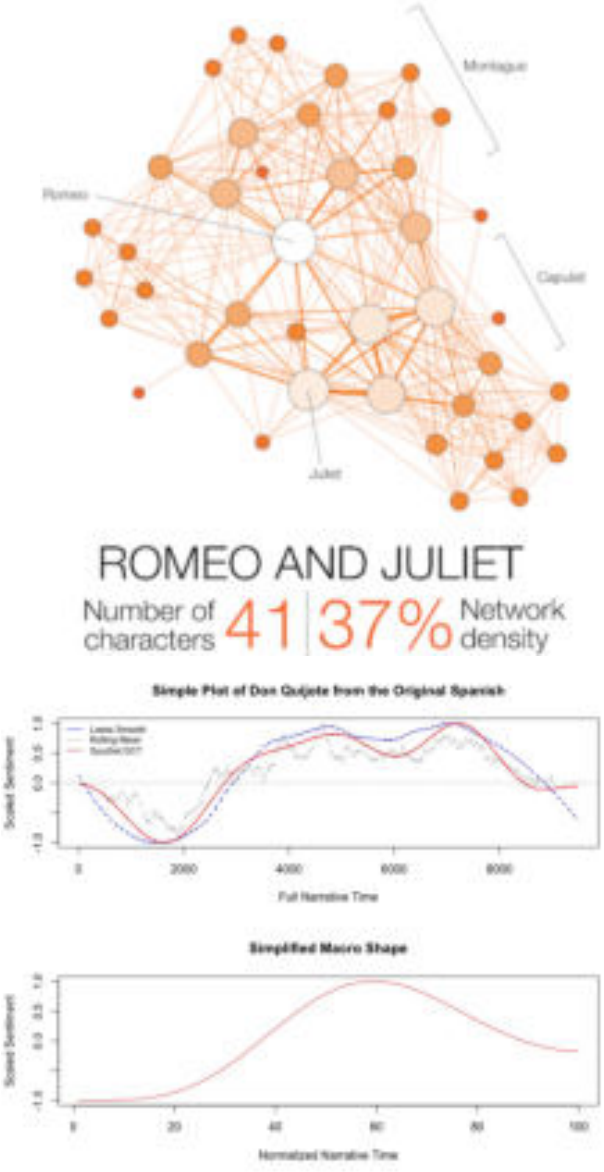
Applications of TA

Since Moretti, literary scholars have adopted his approach to investigate e.g. networks of characters in Shakespearean plays

(Martin Grandjean http://www.martingrandjean.ch/)

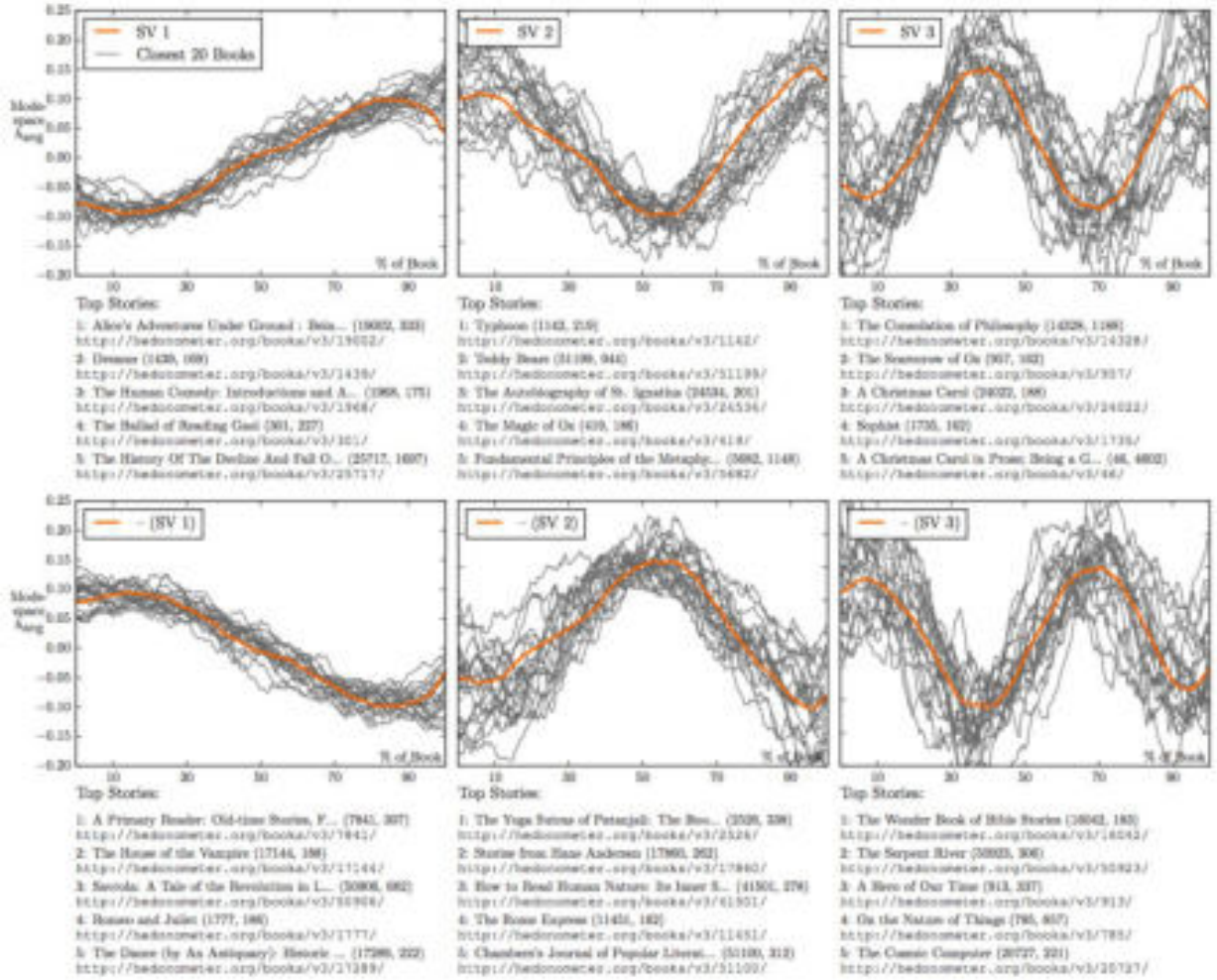
or analyzed typical emotional development in literary narratives

(Matthew L. Jockers http://www.matthewjockers.net/)



Applications of TA

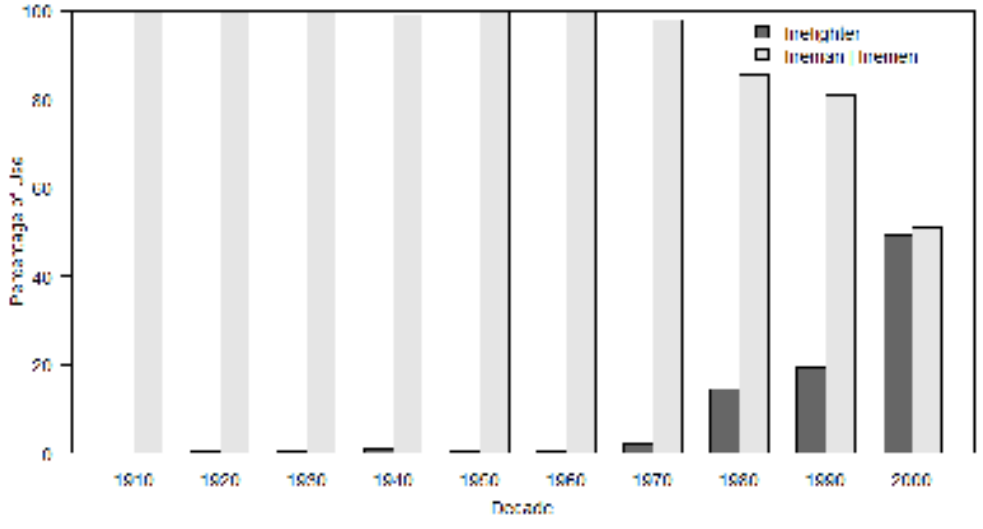
(Scharping 2016)

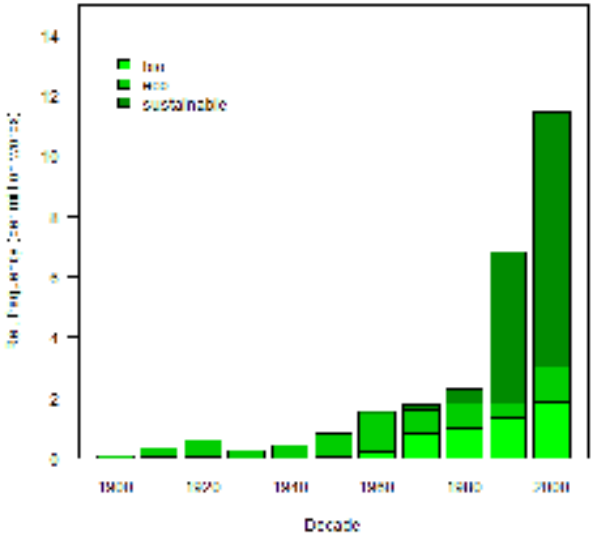


Applications of TA

Investigating Changes in Culture

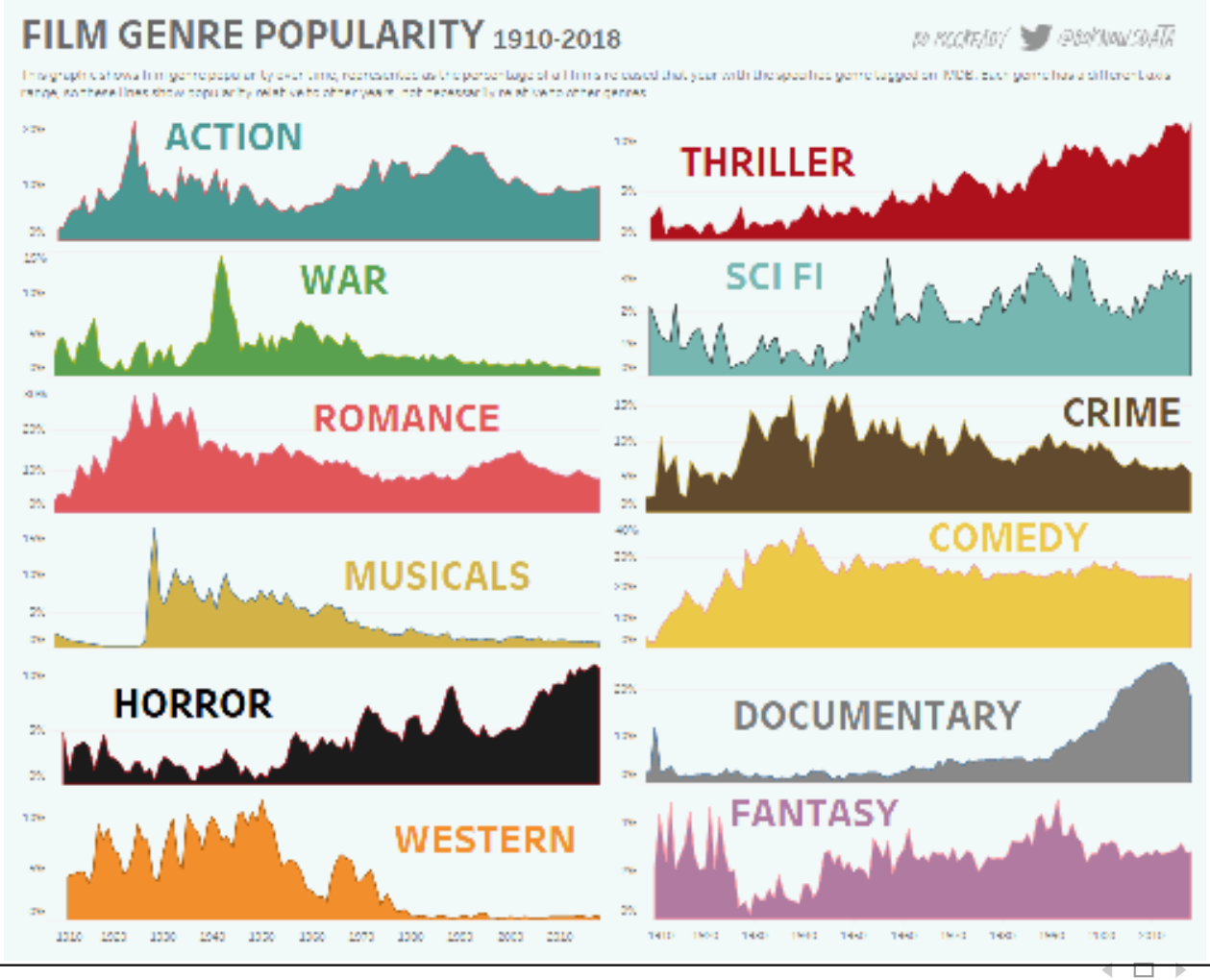
- We can use online corpora like the *Corpus of Historical American English* (COHA) to investigate cultural changes and their reflections in language use.





Applications of TA

(Girdwood 2019)

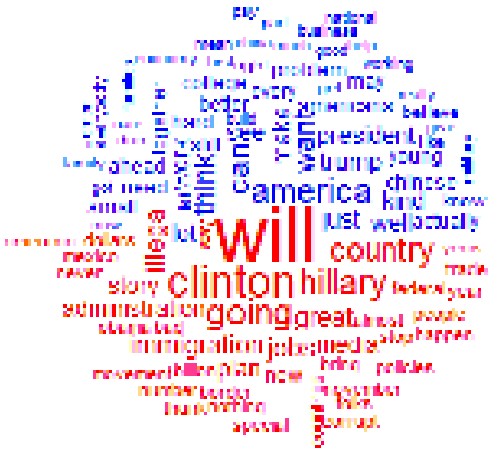


Applications of TA

Comparative Word Clouds display frequency differences of word usage and they can be utilized to investigate, e.g., the political discourse of the 2016 US presidential campaigns of Donald Trump and Hillary Clinton

- Which words were particularly distinctive for Clinton and Trump in the 2016 US presidential campaign?
- How did the rally speeches of the candidates differ?
- Do word frequencies reflect the images of the candidates?

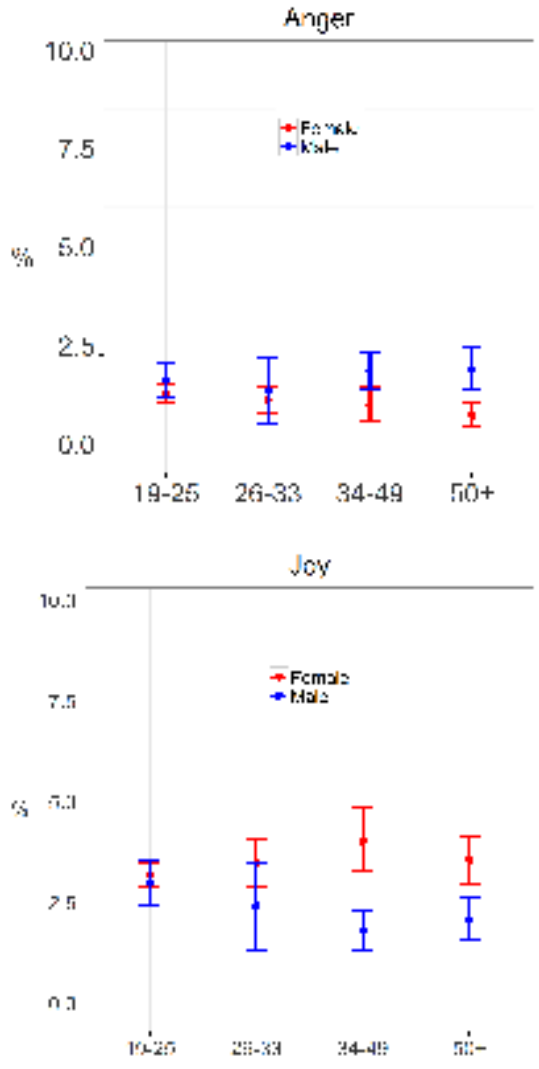
CLINTON



TRUMP

Applications of TA

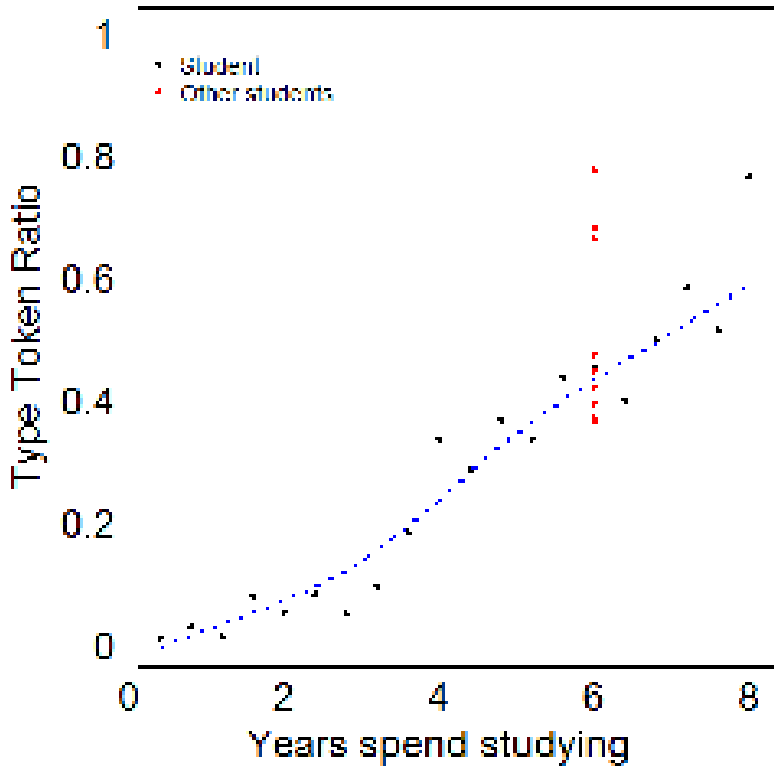
TA can be used to investigate cultural norms and social stereotypes, e.g., by using a sentiment analysis to investigate whether men and women differ in their expression of emotions.



Applications of TA

TA and Language Technology (LT) can enhance language learning and teaching and it can serve to help in grading and provide objective feedback to students

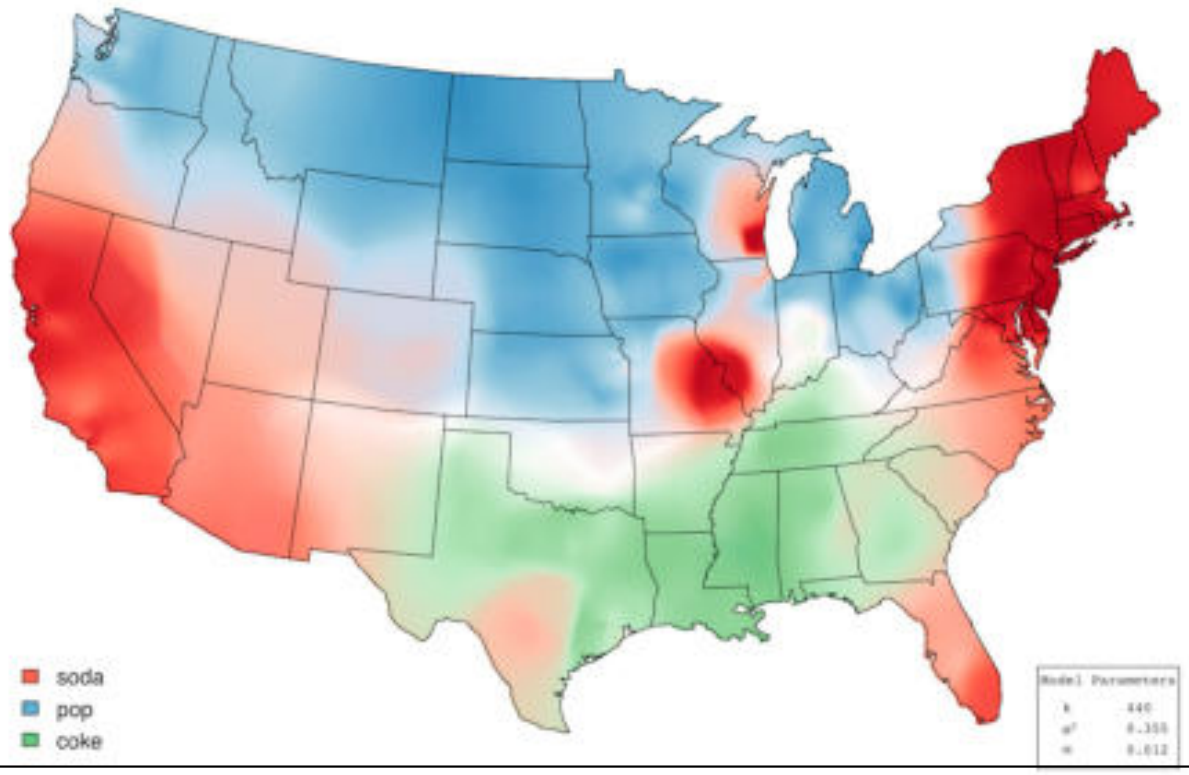
- Type-Token Ratios (TTRs) of language learners over time and across learners
- Do the TTRs of learners increase over time?
- Do the TTRs increase after a stay abroad?



Type = different words in a text.
Token = overall words in a text.
Example = 10 Types/100 Token = 0.1 TTR

Applications of TA

Computers provide us with means to visualize language data in ways that have not been available to linguistics before and can thereby offer new ways of seeing the world.



Why do we need best practices?

Best practice

Best practice is a method or technique that is superior to alternatives because it produces superior results or assures to maintain quality and standards.

- Avoidance of ethical or legal issues;
- Assures high quality (quality control)
- Assists in efficiency (time/effort saving)

Replication crisis (RC)

. . . ongoing methodological crisis primarily affecting parts of the social and life sciences beginning in the early 2010s.

- growing awareness of the problem that results of many scientific studies are difficult or impossible to replicate/reproduce.
- reproducibility is an essential part of the scientific method,
- inability to replicate the studies of others has potentially grave consequences for many fields of science in which significant theories are grounded on unreproducible work.

Adds to loss of public confidence in the social and life sciences!

SCIENTIFIC AMERICAN

HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS

Observations

(Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018

SCIENTIFIC AMERICAN

HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS

More social science studies just failed to replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

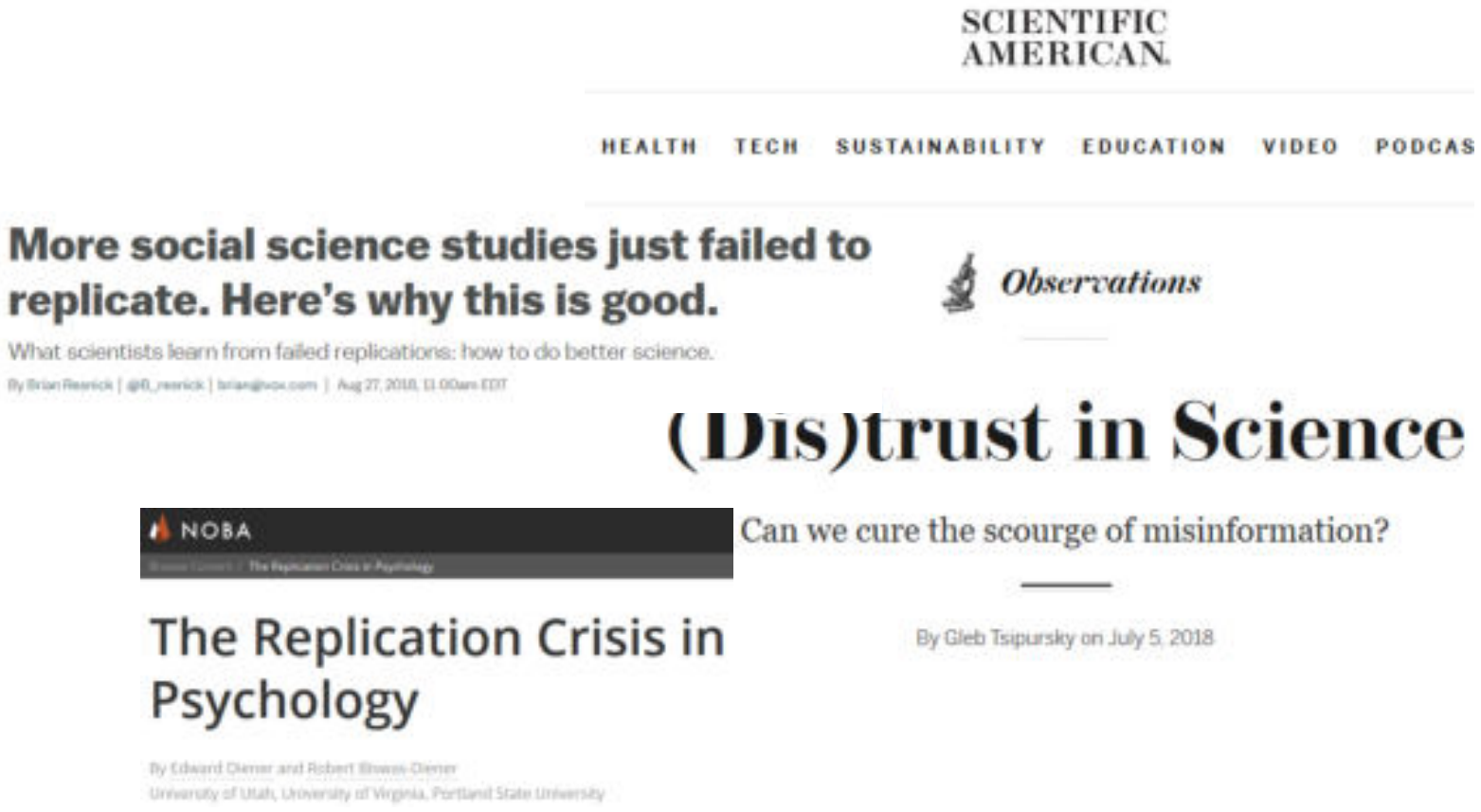
By Brian Resnick | @B_resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

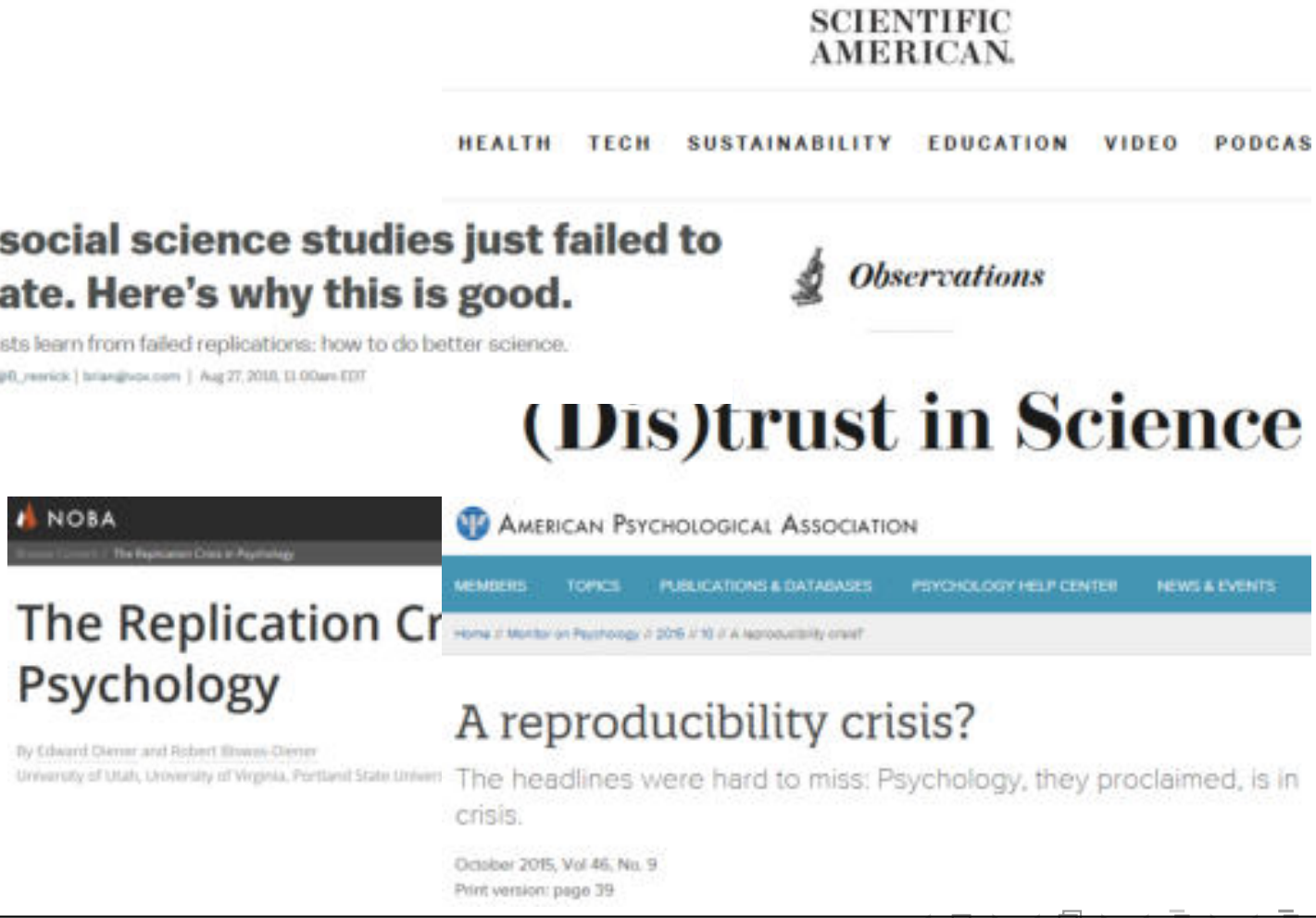
Observations

(Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018

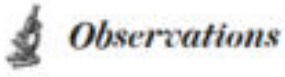

SCIENTIFIC AMERICAN
HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS
More social science studies just failed to replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.
Observations
(Dis)trust in Science
Can we cure the scourge of misinformation?
By Gleb Tsipursky on July 5, 2018
NOBA
The Replication Crisis in Psychology

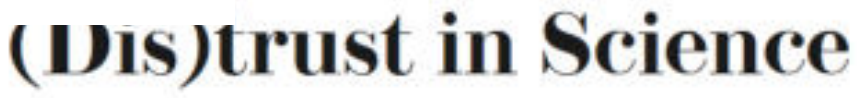

SCIENTIFIC AMERICAN
HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS
More social science studies just failed to replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.
Observations
(Dis)trust in Science
NOBA
The Replication Cr
Psychology
AMERICAN PSYCHOLOGICAL ASSOCIATION
MEMBERS TOPICS PUBLICATIONS & DATABASES PSYCHOLOGY HELP CENTER NEWS & EVENTS
A reproducibility crisis?
The headlines were hard to miss: Psychology, they proclaimed, is in crisis.
October 2015, Vol 46, No. 9
Print version: page 39


PHYS.ORG
Researcher discusses the the science replication crisis


More social science studies just failed to replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.


SCIENTIFIC AMERICAN
TAINABILITY EDUCATION VIDEO PODCAS


Observations

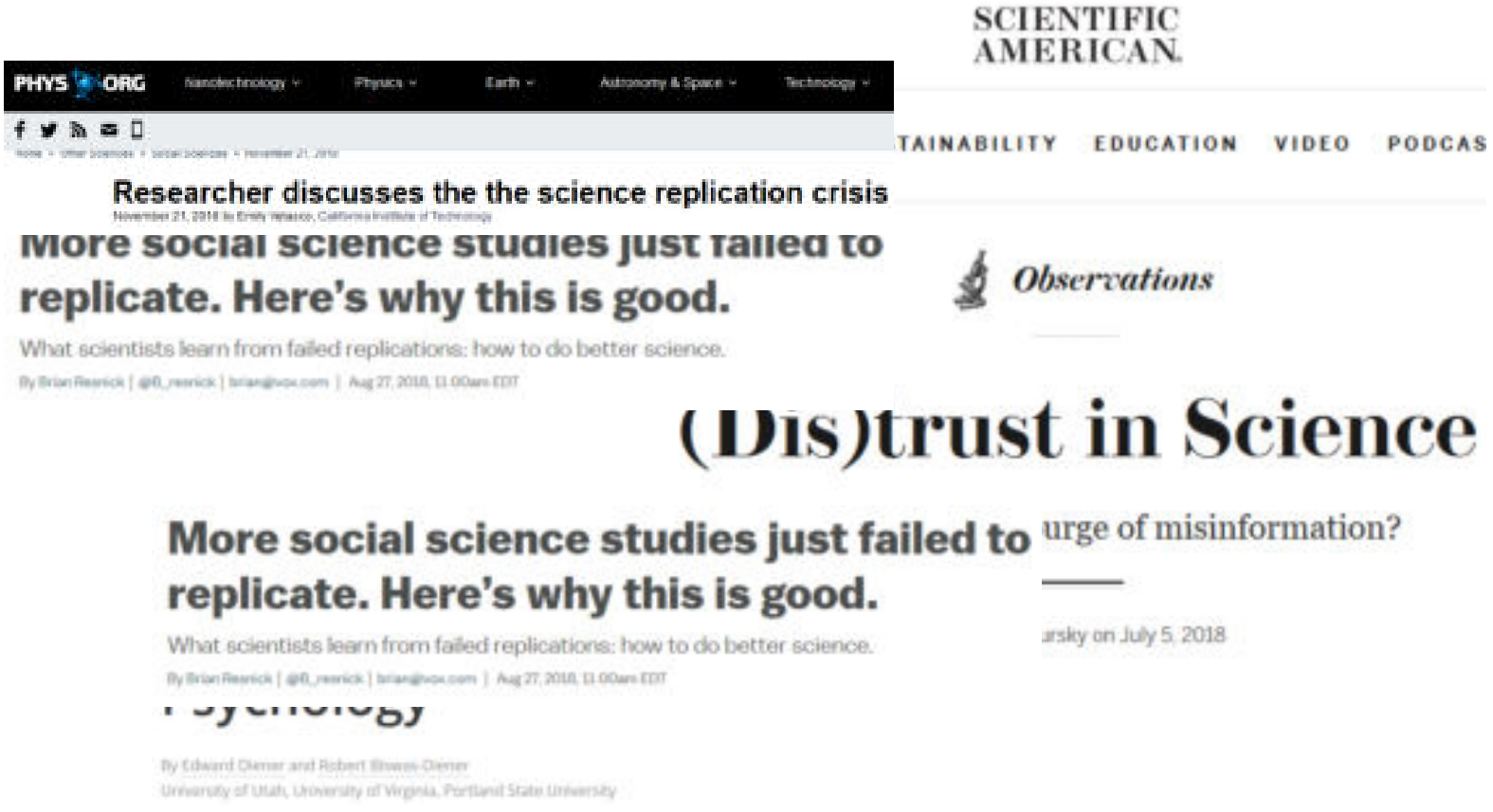

(Dis)trust in Science

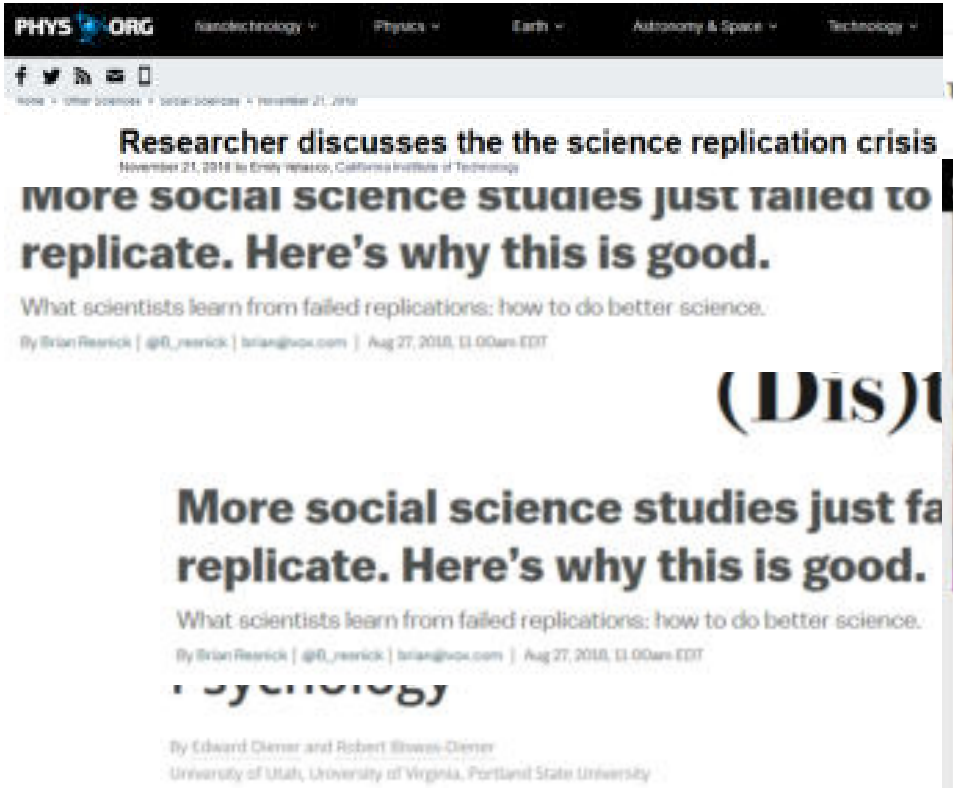

NOBA

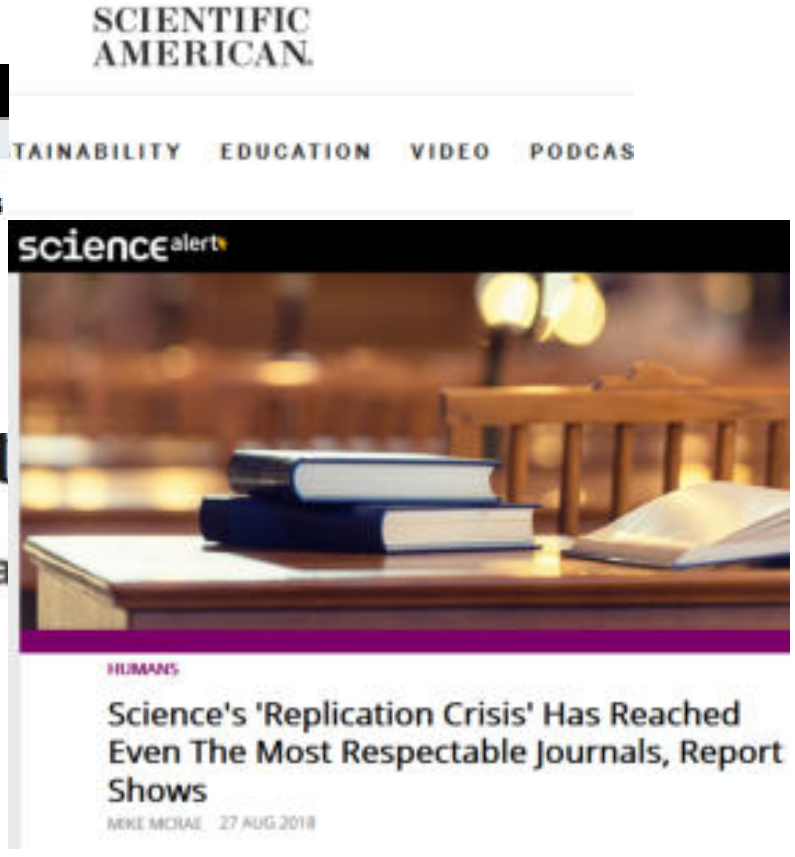

The Replication Crisis in Psychology

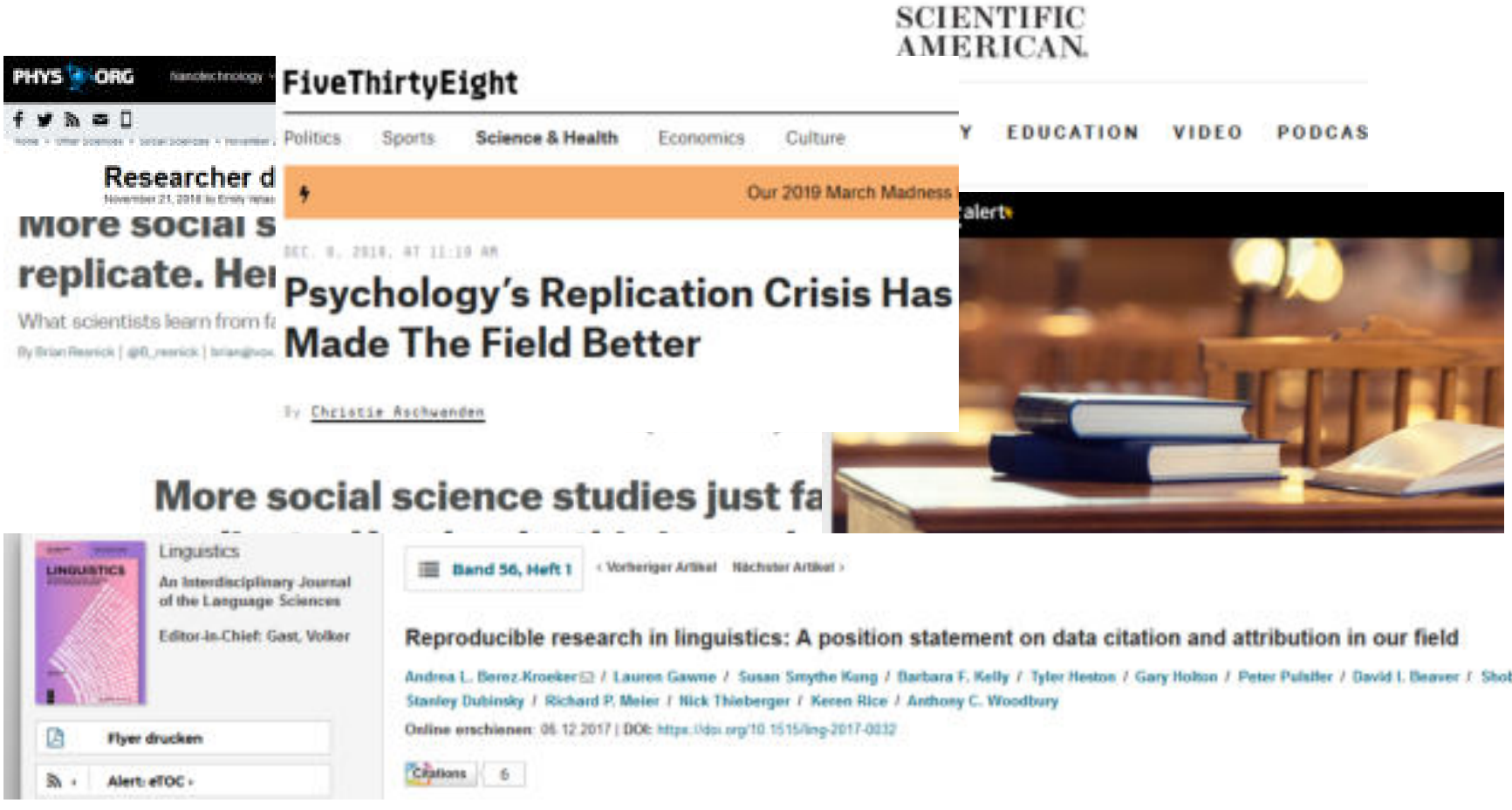

Can we cure the scourge of misinformation?


By Gleb Tsipursky on July 5, 2018


PHYS.ORG
Researcher discusses the the science replication crisis
More social science studies just failed to replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.
SCIENTIFIC AMERICAN
TAINABILITY EDUCATION VIDEO PODCAS
Observations
(Dis)trust in Science
More social science studies just failed to replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.
urge of misinformation?
arsky on July 5, 2018


PHYS.ORG
Researcher discusses the the science replication crisis
More social science studies just failed to replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.
More social science studies just fa
replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.


SCIENTIFIC AMERICAN
TAINABILITY EDUCATION VIDEO PODCAS
sciencealert
HUMANS
Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows
MIKE MCRAE 27 AUG 2018


PHYS.ORG
FiveThirtyEight
Politics Sports Science & Health Economics Culture
Our 2019 March Madness
Psychology's Replication Crisis Has Made The Field Better
By Christie Aschwanden
SCIENTIFIC AMERICAN
EDUCATION VIDEO PODCAS
More social science studies just fa
Linguistics
An Interdisciplinary Journal of the Language Sciences
Editor-in-Chief: Gast, Volker
Band 56, Heft 1
Reproducible research in linguistics: A position statement on data citation and attribution in our field
Andrea L. Berez-Kroeker / Lauren Gawne / Susan Smythe Kung / Barbara F. Kelly / Tyler Heston / Gary Holton / Peter Pulsifer / David I. Beaver / Shobhana Chelliah / Stanley Dubinsky / Richard P. Meier / Nick Thieberger / Keren Rice / Anthony C. Woodbury
Flyer drucken
Alert: eTOC

Replication crisis (RC)

Nature 2016 poll of 1,500 scientists

- 70% had failed to reproduce at least one other scientist's experiment
- 50% had failed to reproduce one of their own experiments (cf. Fanelli 2009)

2009 meta-analysis of surveys on science fraud (Fanelli 2009)

- 2% admitted to falsifying studies at least once
- 14% admitted to personally knowing someone who did

More importantly: data analysis is often too lengthy/complex to describe in detail...

Problem

We just do not know how bad our science is...
(outright forgery, data manipulation, p-hacking, etc.)
because we do not (or only rarely)
reproduce and replicate...

Best Practices in TA

RC in HASS

Good

- blind peer-review (for problems cf., e.g., Tancock (2018))
- we are open and share if we are asked (sometimes)
- discussion has begun (cf. e.g. Berez-Kroeker et al. 2018)

Bad

- analyses are not reproducible/replicated
- reliance on tools not scripts
- reproduction is discouraged
  (if successful: journals are not interested in publishing the same analysis twice/several times;
  if unsuccessful: researchers do not want to threaten the face of other researchers)

Solutions

Open Access Data

Access to data sets to enable reproduction

Code

Scripts rather than tools

Publication

Only studies that provide data and code and thereby enable reproduction/replication should be published

Solutions

Open Access Data
Access to data sets to enable reproduction

Code
Scripts rather than tools

Publication
Only studies that provide data and code and thereby enable reproduction/replication should be published

Code

- Scripts allow exact replication (total transparency)
- Only practical solutions for true replication
  (too time consuming to replicate a tool-based analysis)
- Data analysis is too fine-grained to be described in papers
  (including all steps the researcher has undertaken)
- Training programs for basic programming at
  universities/schools (obligatory for grad programs)

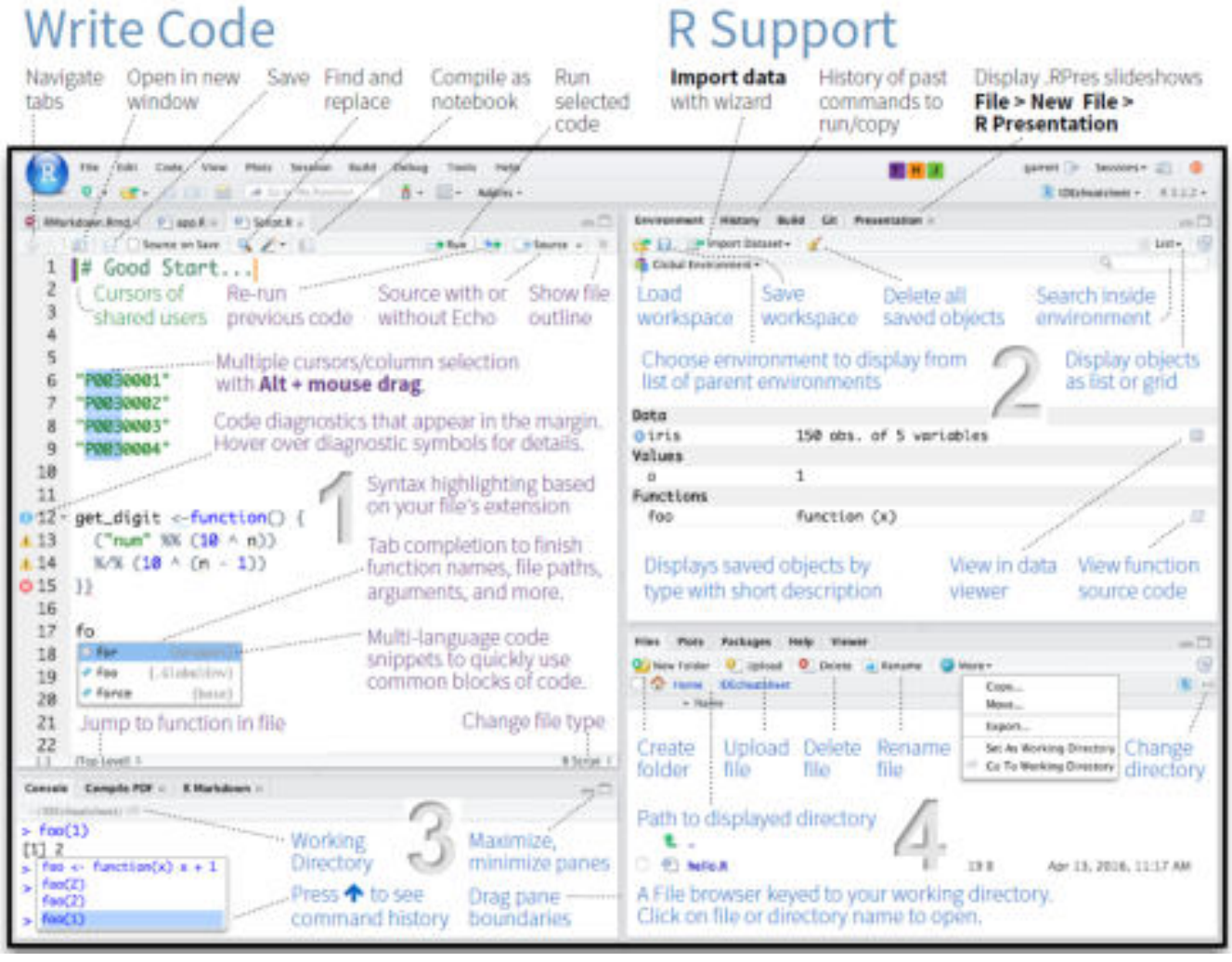
```
#install.packages(Rling)       # install Rling library (remove # to activate)
library(Rling)                 # activate Rling library
library(partykit)              # activate partykit library
library(dplyr)                 # activate dplyr library
options(stringsAsFactors = T)  # set options: do not convert strings
options(scipen = 999)          # set options: supress math. notation
options(max.prAmplified=10000) # set options
# load data
citdata <- read.delim("data/treedata.txt", header = T, sep = "\t")
head(citdata)
```

A Brief Note on R

Why R?

- Open source free-ware
- Full transparency: all steps of the analysis are documented and scripts can be shared easily
- One of the fastest growing world's top 10 programming environments and is a fully fledged human/user-centered programming language
- Allows complex text analysis/data analysis (statz)/data viz (including geo mapping), speech analysis, creating websites, slides, apps, or notebooks, etc.
- Compatible with other software apps common in HASS (Praat, MAUS, Excel, etc.)

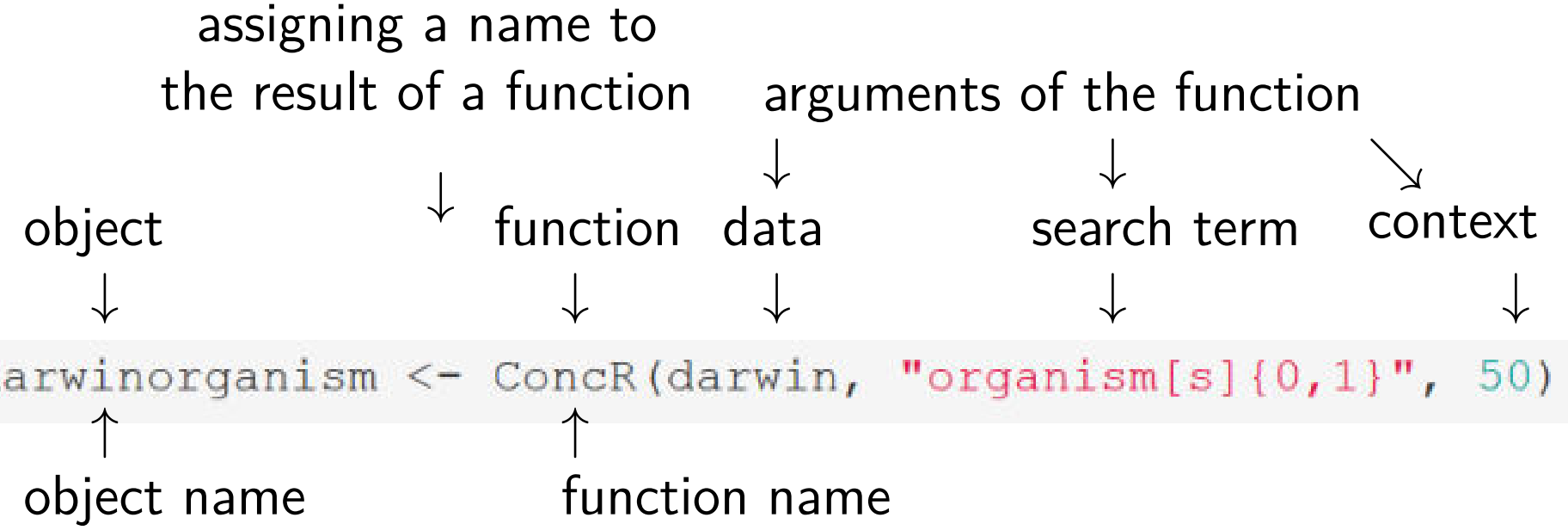
R Basics: R Studio



1. File editor
2. Environment variables
3. R console
4. Management panes (File browser, plots, help display, R packages).

R Basics: Code Structure

```r
# load function for concordancing
source("https://slcladal.github.io/rscripts/ConcR_2.5_LoadedFiles.R")
# start concordancing
darwinorganism <- ConcR(darwin, "organism[s]{0,1}", 50)
# inspect data
darwinorganism[1:5, 2:ncol(darwinorganism)]
```



Solutions at UQ

- Training program: workshops on R √/X
  (for all levels of expertise *Center for Digital Scholarship*/*School of Languages and Cultures*)
- Materials √/X
  Language Technology and Data Analysis Laboratory (LADAL) website (data and text analysis with R: `https://slcladal.github.io/index.html`)
- Study program √/X (beginning to plan a program)
  Digital HASS (BA/MA program including modules on data and text analysis with R)

Language Technology and Data Analysis Lab.

(LADAL)

LADAL

- "Best Practices" in HASS eResearch
- Sustainable research data infrastructure
- Replication oriented research practices
- Provision of specific, hands-on training in
  - Data extraction / transformation / processing
  - NLP applications (Text Analysis)
  - Visualisation techniques
  - Statistics/Classification/etc.

NOTE: LADAL is work-in-progress!

Today's topics

Theoretical background

- Why Text Analysis (TA)?
- What is TA?
- Why do we need best practices?
- Best practices in TA
- LADAL (Language Technology and Data Analysis Lab.)

Hands-on session

- Creating concordances (KWIC displays)
- Visualizing word frequencies
- Collocations
- Sentiment Analysis

Hands-on session

https://slcladal.github.io/textanalysis.html

So, what do you think???

Comments? Feedback? Suggestions?

Aschwanden, C. (2018). Psychology's replication crisis has made the field better.

Berez-Kroeker, A. L., L. Gawne, S. S. Kung, B. F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. I. Beaver, S. Chelliah, S. Dubinsky, et al. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics 56*(1), 1–18.

Diener, E. and R. Biswas-Diener (2019). The replication crisis in psychology.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS One 4*, e5738.

Girdwood, A. (2019). The rise and fall of film genres. https://www.geeknative.com/64459/the-rise-and-fall-of-film-genres/.

Grandjean, M. (2015). Network visualization: mapping shakespeare's tragedies. http://www.martingrandjean.ch/network-visualization-shakespeare/.

Jockers, M. L. (2017). Syuzhet 1.0.4 now on cran. http://www.matthewjockers.net/2017/12/16/syuzhet-1-0-4/.

McRae, M. (2018). Science's 'replication crisis' has reached even the most respectable journals, report shows.

Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*. Verso.

Moretti, F. (2013). *Distant reading*. Verso.

Resnick, B. (2018). More social science studies just failed to replicate. here's why this is good.what scientists learn from failed replications: how to do better science.

Scharping, N. (2016). 6 story arcs define western literature, data-mining study reveals. http://blogs.discovermagazine.com/d-brief/2016/07/06/the-6-story-arcs-that-define-western-literature/.

Tancock, C. (2018). When reviewing goes wrong: the ugly side of peer review. https://www.elsevier.com/connect/editors-update/when-reviewing-goes-wrong-the-ugly-side-of-peer-review.

Velasco, E. (2019). Researcher discusses the the science replication crisis.

Weir, K. (2015). A reproducibility crisis? the headlines were hard to miss: Psychology, they proclaimed, is in crisis. *Monitor on Psychology 46*, 39.

Yong, E. (2018). Psychology's replication crisis is running out of excuses. another big project has found that only half of studies can be repeated. and this time, the usual explanations fall flat.