PADERBORN UNIVERSITY

data analytics group

ANALYSIS OF UNSTRUCTURED TEXT DATA WITH TOPIC MODELS

WK RECH 2019, Frankfurt School of Finance & Management

PROF. DR. OLIVER MÜLLER

LS Wirtschaftsinformatik, insb. Data Analytics

- Since October 2018: Professor at the Department of Management Information Systems, Paderborn University
- 2016-2018: Associate Professor at the Business IT Department, IT University of Copenhagen
- 2011-2016: Assistant Professor at the Institute of Information Systems, University of Liechtenstein
- 2007-2011: PhD at European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster

Research Interests

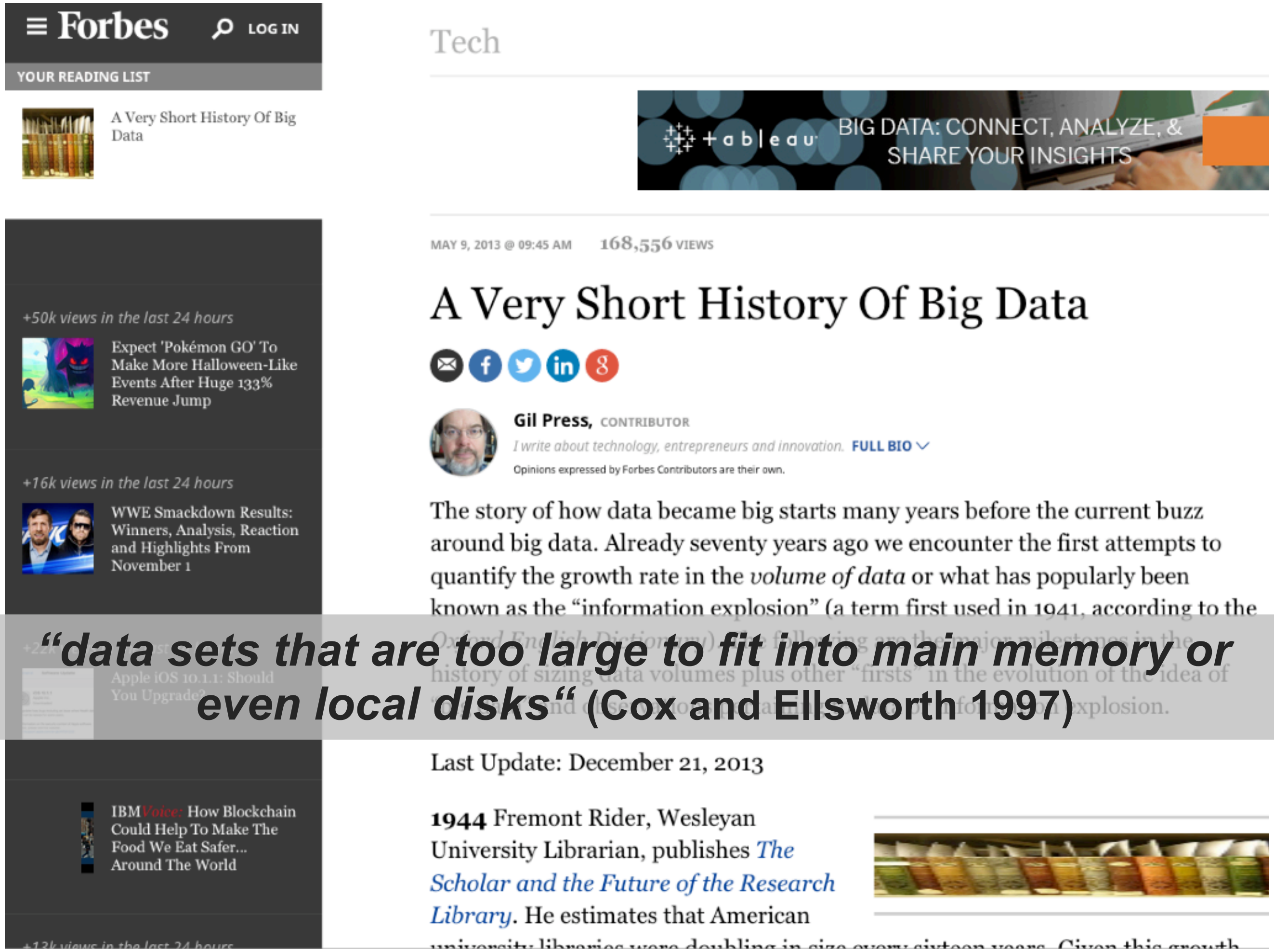
1. Using big data and machine learning to solve relevant business and societal problems
2. Analysis of unstructured data (e.g., text, images)
3. Acceptance and value of big data analytics

I. Big Text Data
II. Fundamentals of Topic Modeling
III. Topic Modeling Walkthrough

Paderborn University

BIG TEXT DATA

BIG TEXT DATA



Forbes LOG IN

YOUR READING LIST

A Very Short History Of Big Data

+50k views in the last 24 hours

Expect 'Pokémon GO' To Make More Halloween-Like Events After Huge 133% Revenue Jump

+16k views in the last 24 hours

WWE Smackdown Results: Winners, Analysis, Reaction and Highlights From November 1

Apple iOS 10.1.1: Should You Upgrade?

IBMVoice: How Blockchain Could Help To Make The Food We Eat Safer... Around The World

Tech

MAY 9, 2013 @ 09:45 AM 168,556 VIEWS

A Very Short History Of Big Data

Gil Press, CONTRIBUTOR

I write about technology, entrepreneurs and innovation. FULL BIO

Opinions expressed by Forbes Contributors are their own.

The story of how data became big starts many years before the current buzz around big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the *volume of data* or what has popularly been known as the "information explosion" (a term first used in 1941, according to the *Oxford English Dictionary*). The following are the major milestones in the history of sizing data volumes plus other "firsts" in the evolution of the idea of "big data" and observations pertaining to data or information explosion.

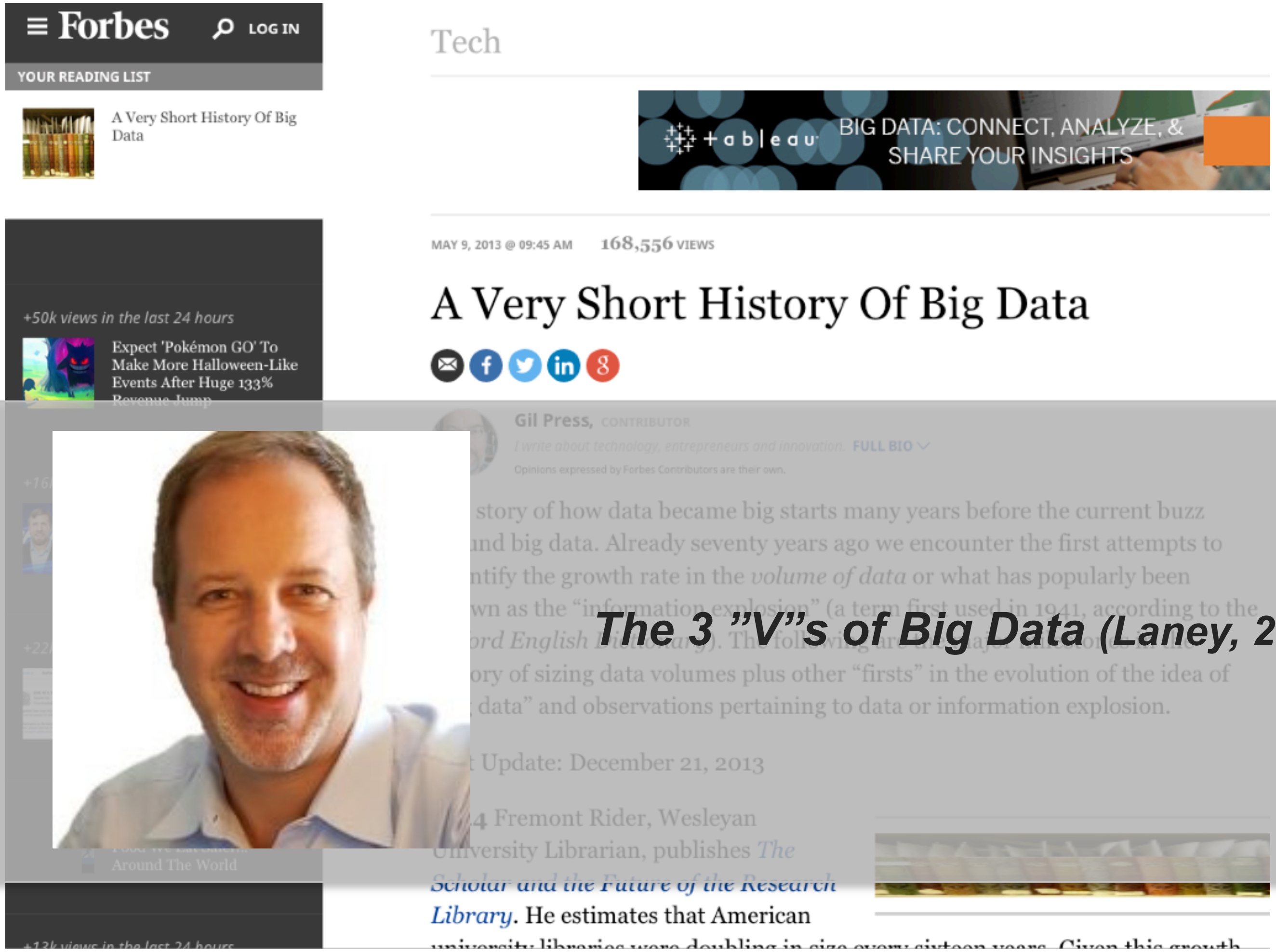
Last Update: December 21, 2013

1944 Fremont Rider, Wesleyan University Librarian, publishes *The Scholar and the Future of the Research Library*. He estimates that American

"data sets that are too large to fit into main memory or even local disks" **(Cox and Ellsworth 1997)**

Source: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#41b6f62055da

BIG TEXT DATA



≡ Forbes LOG IN

YOUR READING LIST

A Very Short History Of Big Data

+50k views in the last 24 hours

Expect 'Pokémon GO' To Make More Halloween-Like Events After Huge 133% Revenue Jump

Tech

BIG DATA: CONNECT, ANALYZE, & SHARE YOUR INSIGHTS

MAY 9, 2013 @ 09:45 AM 168,556 VIEWS

A Very Short History Of Big Data

Gil Press, CONTRIBUTOR

I write about technology, entrepreneurs and innovation. FULL BIO

Opinions expressed by Forbes Contributors are their own.

…story of how data became big starts many years before the current buzz …und big data. Already seventy years ago we encounter the first attempts to …ntify the growth rate in the *volume of data* or what has popularly been …wn as the "information explosion" (a term first used in 1941, according to the …ord English Dictionary). The following … …ory of sizing data volumes plus other "firsts" in the evolution of the idea of … data" and observations pertaining to data or information explosion.

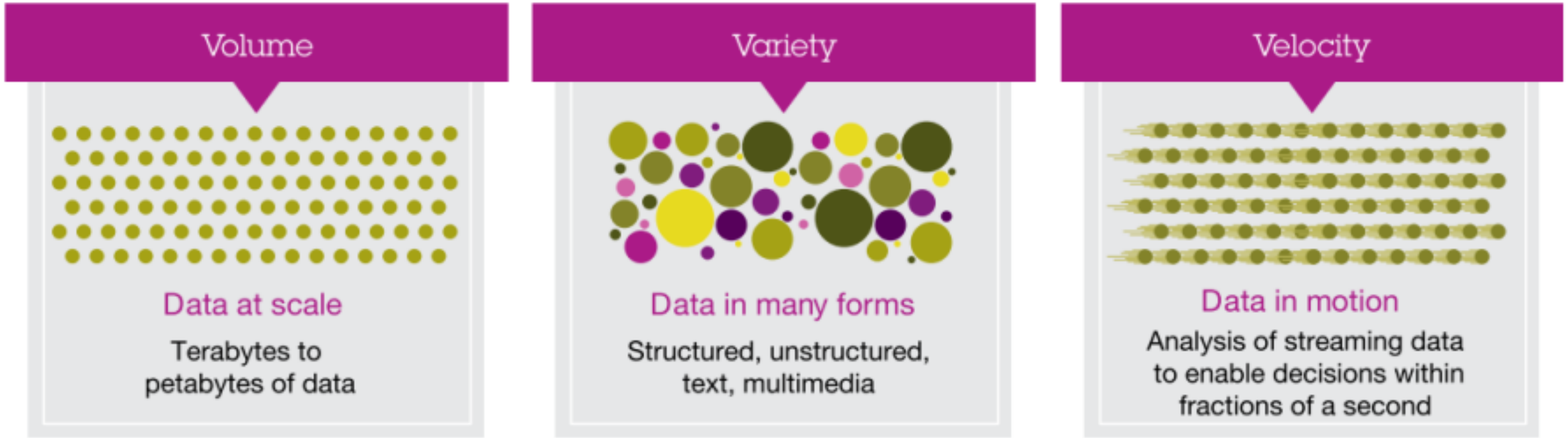
…t Update: December 21, 2013

…4 Fremont Rider, Wesleyan University Librarian, publishes *The Scholar and the Future of the Research Library*. He estimates that American …

The 3 "V"s of Big Data (Laney, 2001)

Source: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#41b6f62055da

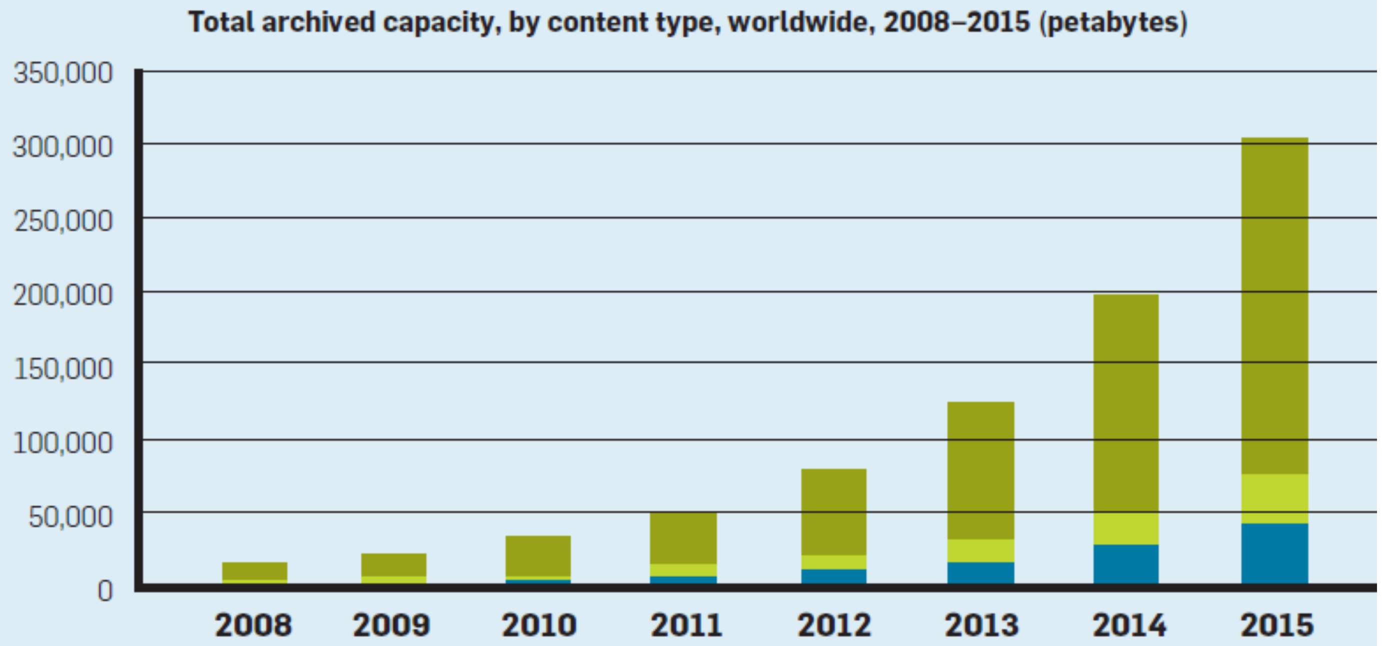
BIG TEXT DATA



Source: Laney (2001), IBM (2012)

PADERBORN UNIVERSITY

Volume and Variety

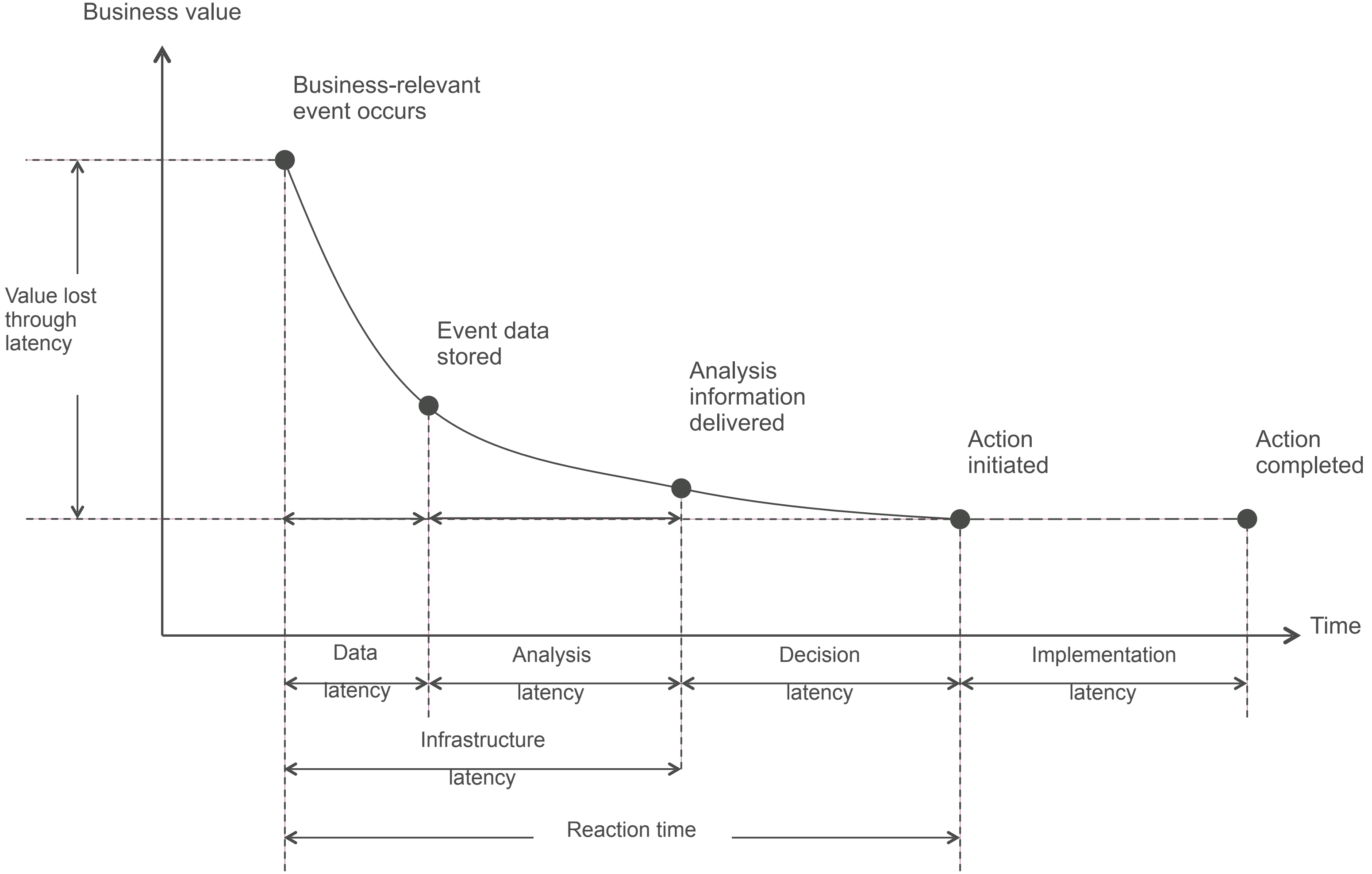


	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

Source: Dhar (2013)

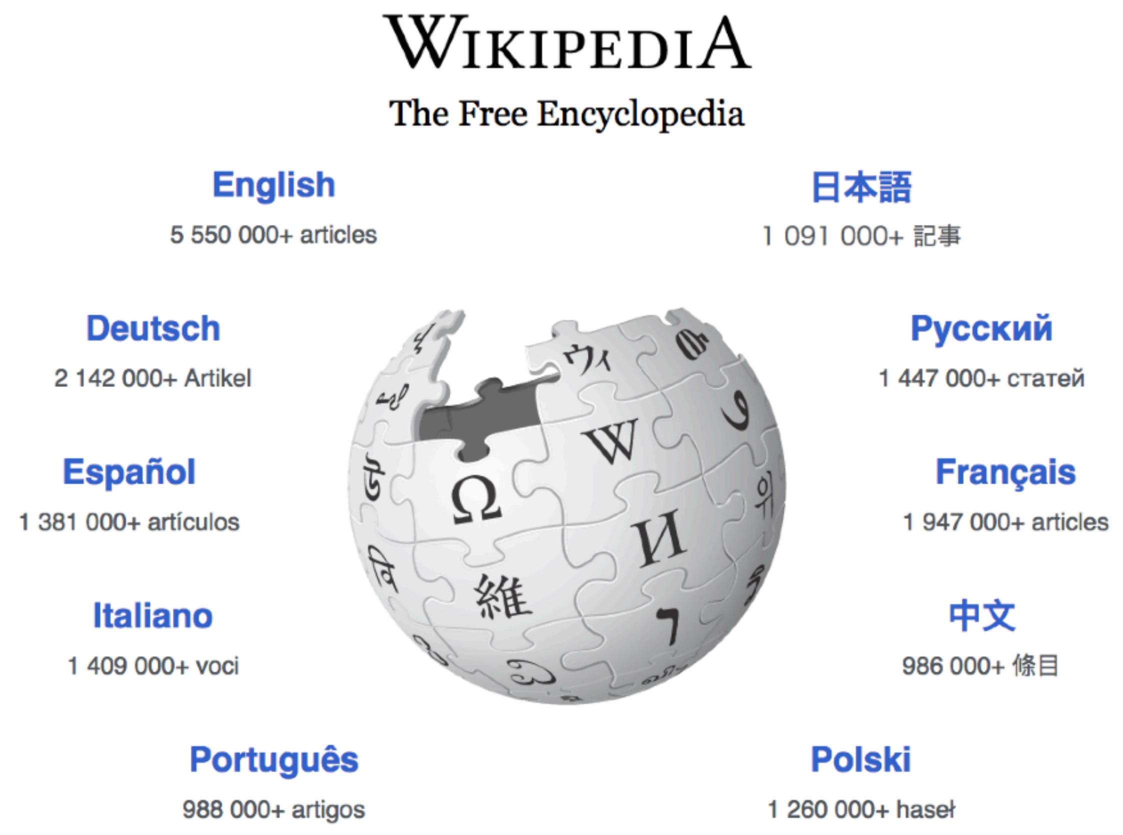
PADERBORN UNIVERSITY

Velocity



Source: Zur Mühlen & Shapiro (2010)

Big or not?


WIKIPEDIA
The Free Encyclopedia
English
5 550 000+ articles
日本語
1 091 000+ 記事
Deutsch
2 142 000+ Artikel
Русский
1 447 000+ статей
Español
1 381 000+ artículos
Français
1 947 000+ articles
Italiano
1 409 000+ voci
中文
986 000+ 條目
Português
988 000+ artigos
Polski
1 260 000+ haseł

PADERBORN UNIVERSITY

Big or not?

PADERBORN UNIVERSITY

Why is Text Analytics Difficult?

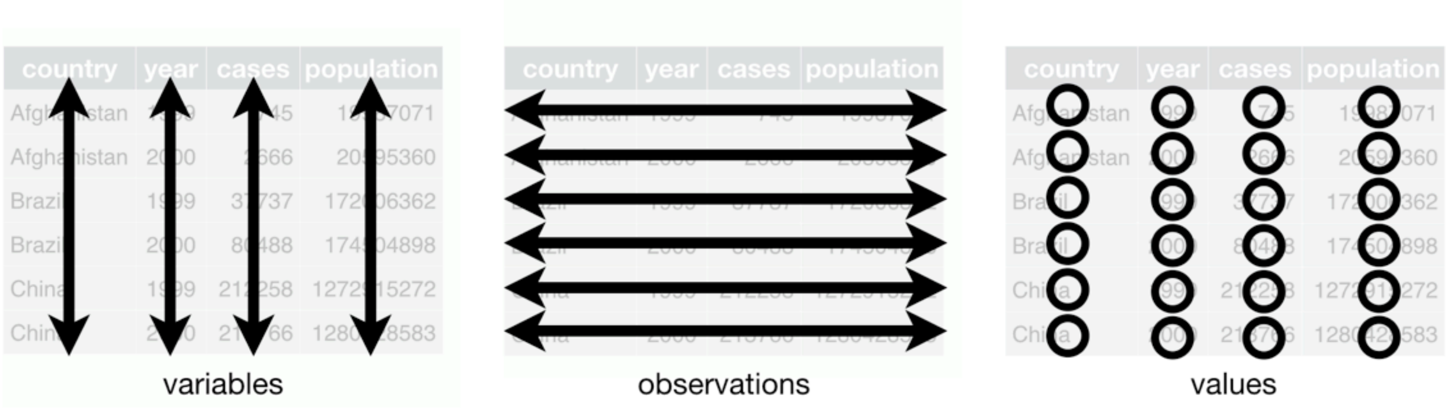
- **Text is Messy**
  - Cannot easily be represented in rows and columns of tables
  - Has complex linguistics structures that differ across languages

- **Text is Dirty**
  - Lots of words that are in no dictionary (e.g., spelling mistakes, slang, abbreviations, technical terms)

- **Text is Ambiguous**
  - Meaning of words depends on context

PADERBORN UNIVERSITY

Text is Messy



Lorem Ipsum

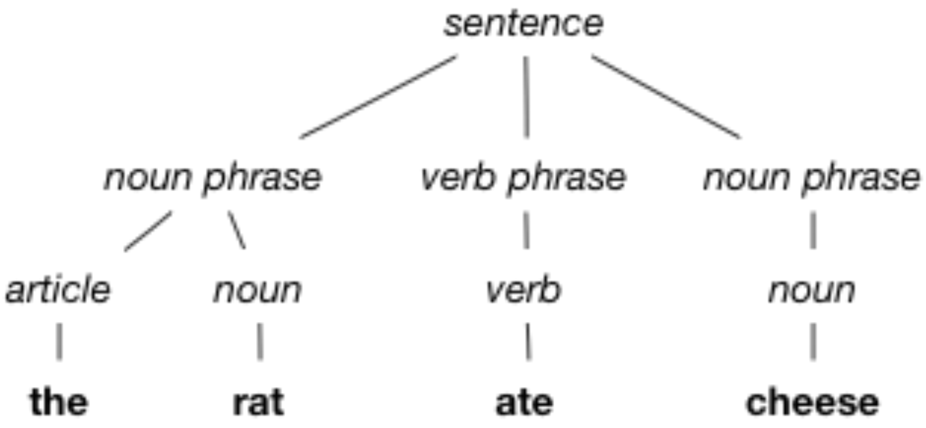
"Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit..."
"There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain..."

What is Lorem Ipsum?

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

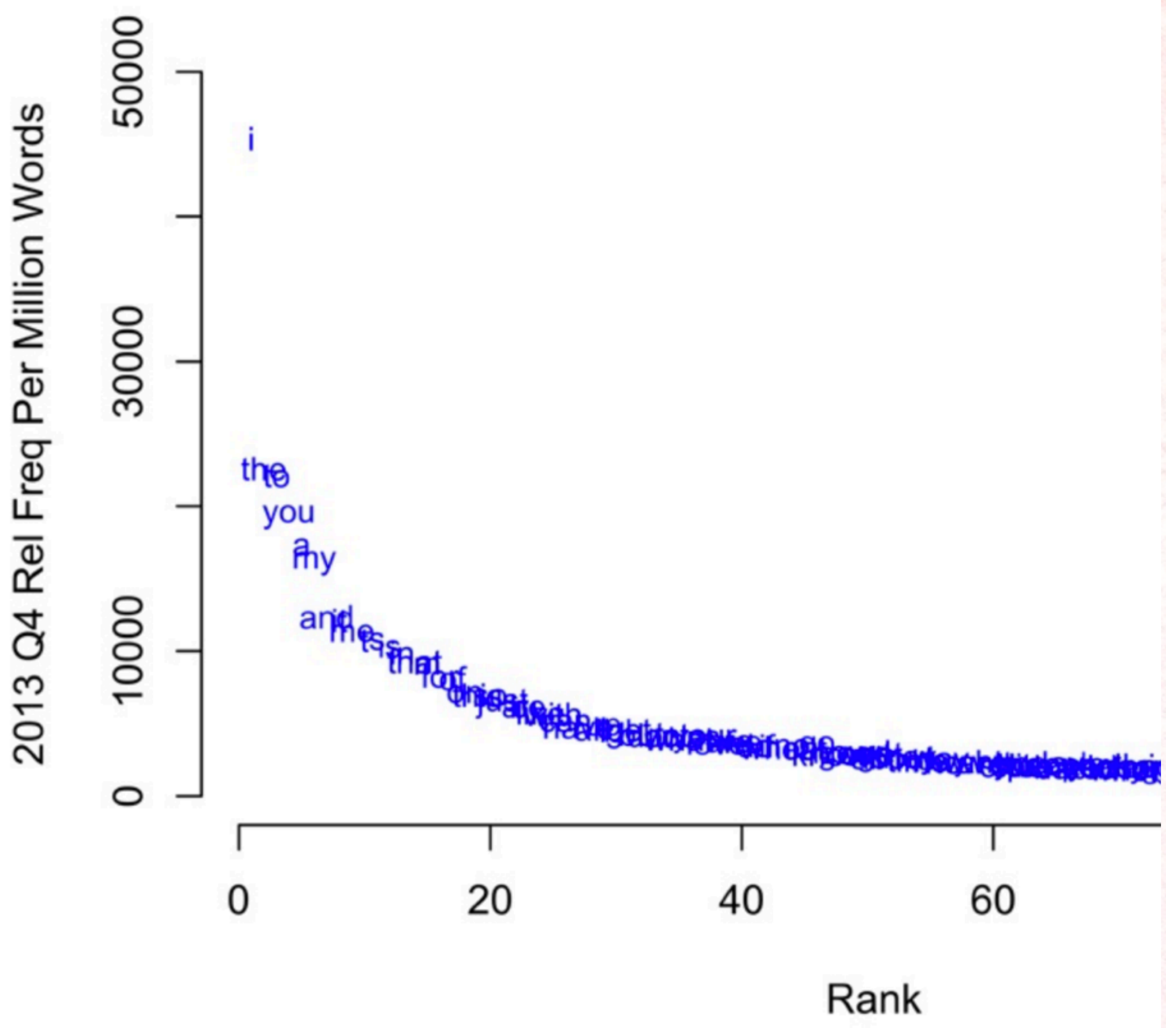
Why do we use it?

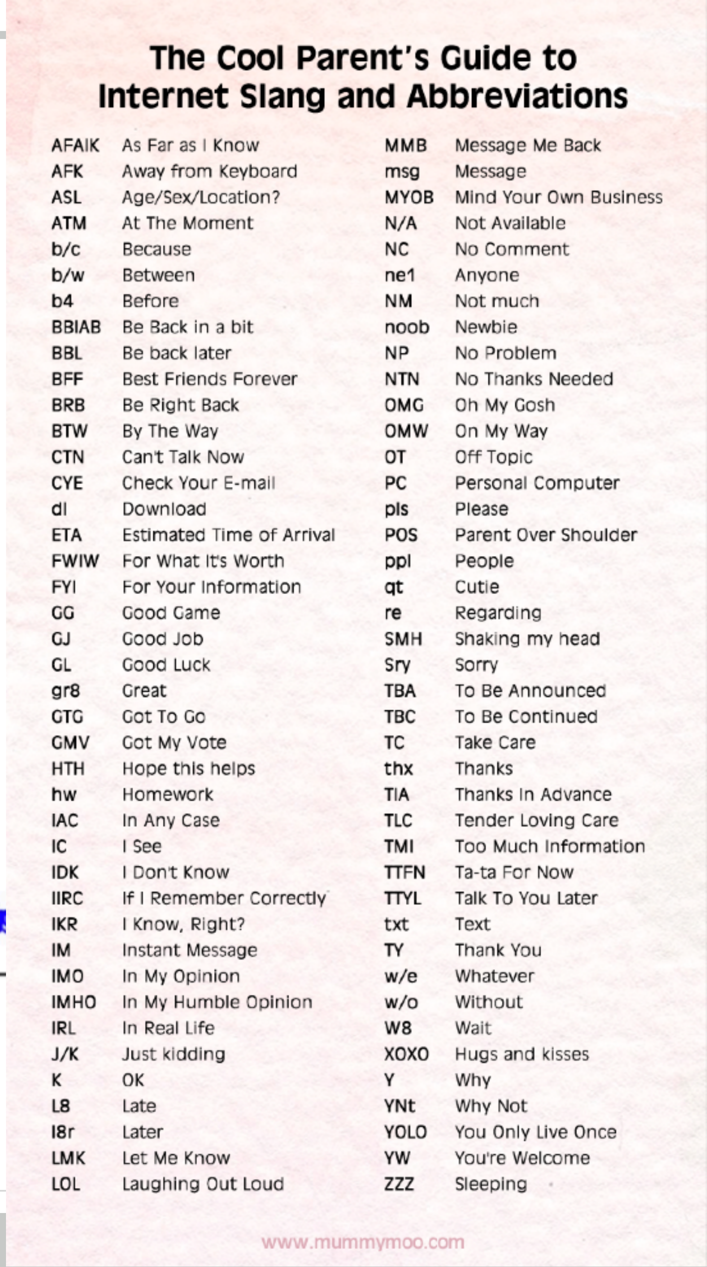
It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a more-or-less normal distribution of letters, as opposed to using 'Content here, content here', making it look like readable English. Many desktop publishing packages and web page editors now use Lorem Ipsum as their default model text, and a search for 'lorem ipsum' will uncover many web sites still in their infancy. Various versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like).



BIG TEXT DATA

Text is Dirty



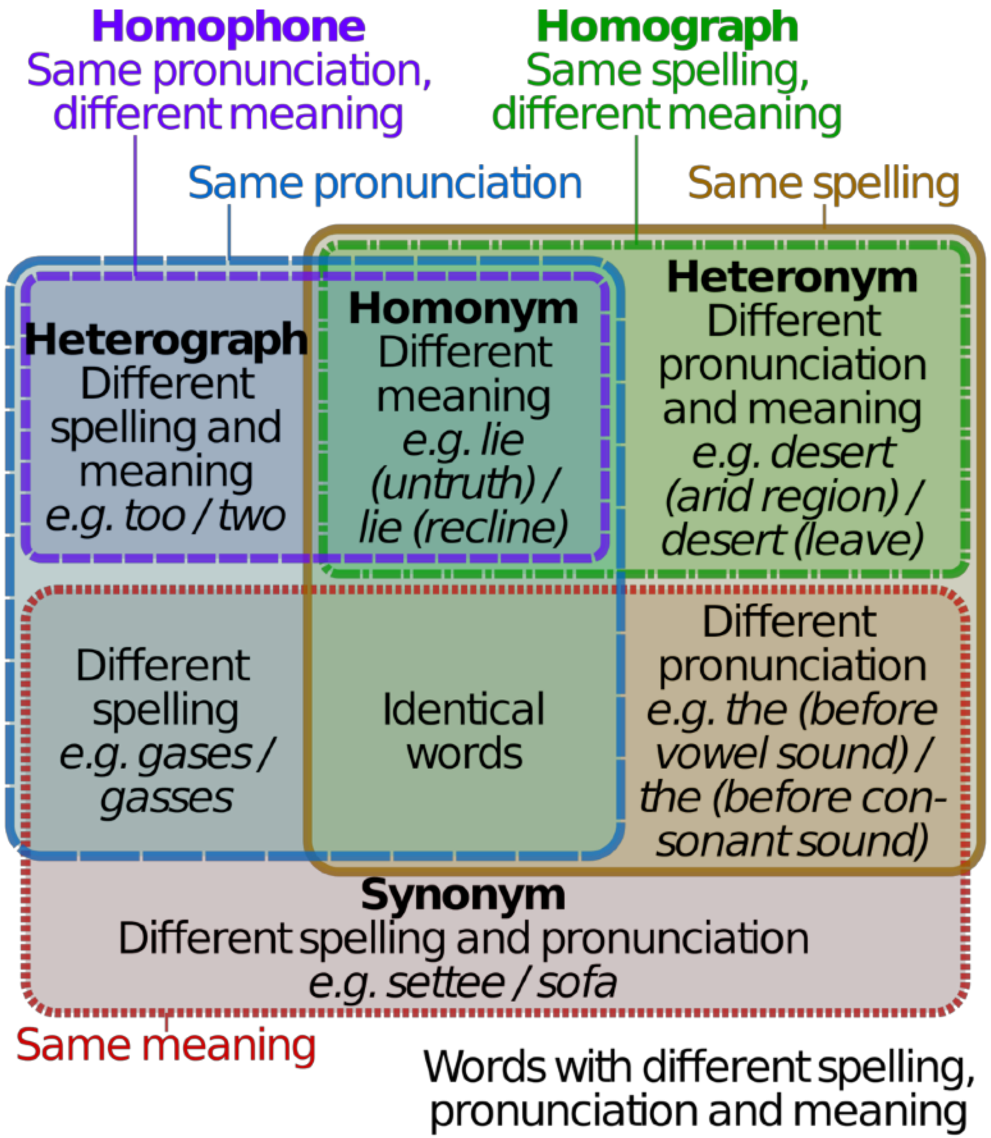


The Cool Parent's Guide to Internet Slang and Abbreviations

AFAIK	As Far as I Know	MMB	Message Me Back
AFK	Away from Keyboard	msg	Message
ASL	Age/Sex/Location?	MYOB	Mind Your Own Business
ATM	At The Moment	N/A	Not Available
b/c	Because	NC	No Comment
b/w	Between	ne1	Anyone
b4	Before	NM	Not much
BBIAB	Be Back in a bit	noob	Newbie
BBL	Be back later	NP	No Problem
BFF	Best Friends Forever	NTN	No Thanks Needed
BRB	Be Right Back	OMG	Oh My Gosh
BTW	By The Way	OMW	On My Way
CTN	Can't Talk Now	OT	Off Topic
CYE	Check Your E-mail	PC	Personal Computer
dl	Download	pls	Please
ETA	Estimated Time of Arrival	POS	Parent Over Shoulder
FWIW	For What It's Worth	ppl	People
FYI	For Your Information	qt	Cutie
GG	Good Game	re	Regarding
GJ	Good Job	SMH	Shaking my head
GL	Good Luck	Sry	Sorry
gr8	Great	TBA	To Be Announced
GTG	Got To Go	TBC	To Be Continued
GMV	Got My Vote	TC	Take Care
HTH	Hope this helps	thx	Thanks
hw	Homework	TIA	Thanks In Advance
IAC	In Any Case	TLC	Tender Loving Care
IC	I See	TMI	Too Much Information
IDK	I Don't Know	TTFN	Ta-ta For Now
IIRC	If I Remember Correctly	TTYL	Talk To You Later
IKR	I Know, Right?	txt	Text
IM	Instant Message	TY	Thank You
IMO	In My Opinion	w/e	Whatever
IMHO	In My Humble Opinion	w/o	Without
IRL	In Real Life	W8	Wait
J/K	Just kidding	XOXO	Hugs and kisses
K	OK	Y	Why
L8	Late	YNt	Why Not
l8r	Later	YOLO	You Only Live Once
LMK	Let Me Know	YW	You're Welcome
LOL	Laughing Out Loud	ZZZ	Sleeping

www.mummymoo.com

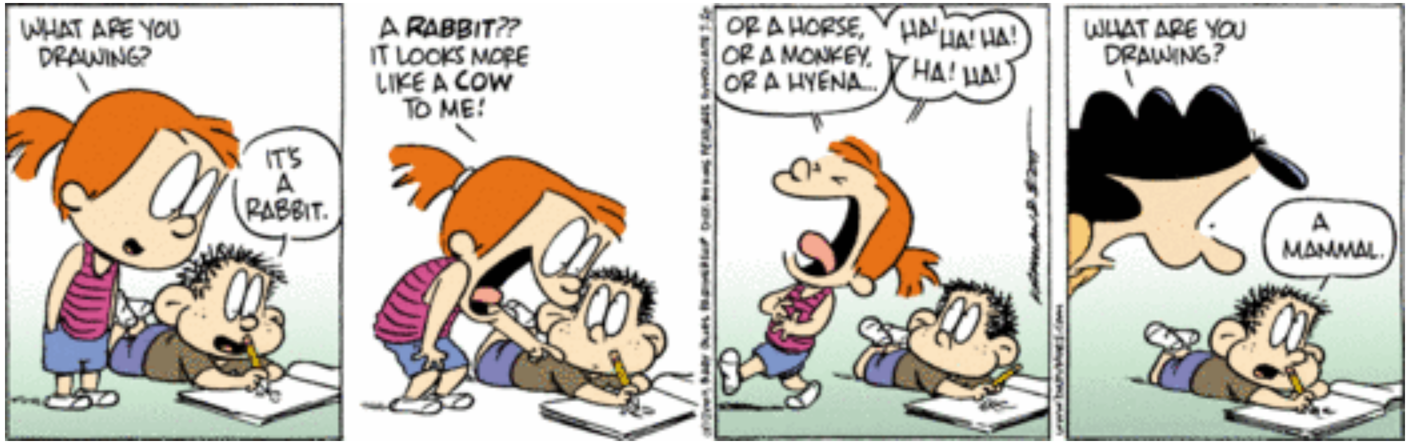
Text is Ambiguous



Source: Wikipedia

FUNDAMENTALS OF TOPIC MODELING

From Counting to Categorizing



http://www.cse.buffalo.edu/~rapaport/575/categories.html

PADERBORN UNIVERSITY

Text Categorization

	Manual coding (bottom up)	Manual coding (top down)	Dictionaries	Supervised machine learning	Unsupervised machine learning
Assumptions					
Categories are predefined	No	Yes	Yes	Yes	No
Relevant text features are known	Yes	Yes	Yes	Yes	Yes
Mapping between text features and categories is known	No	No	Yes	No	No
Costs					
Pre-analysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Low	High	High	High	Low
Analysis costs					
Person-hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate	Moderate	Low	Low	Low
Post-analysis costs					
Person hours spent interpreting	Moderate	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Source: Debortoli et al. (2016)

PADERBORN UNIVERSITY

Text Categorization

	Manual coding (bottom up)	Manual coding (top down)	Dictionaries	Supervised machine learning	Unsupervised machine learning
Assumptions					
Categories are predefined	No	Yes	Yes	Yes	No
Relevant text features are known	Yes	Yes	Yes	Yes	Yes
Mapping between text features and categories is known	No	No	Yes	No	No
Costs					
Pre-analysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Low	High	High	High	Low
Analysis costs					
Person-hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate	Moderate	Low	Low	Low
Post-analysis costs					
Person hours spent interpreting	Moderate	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Source: Debortoli et al. (2016)

FUNDAMENTALS OF TOPIC MODELING

Text Categorization

	Manual coding (bottom up)	Manual coding (top down)	Dictionaries	Supervised machine learning	Unsupervised machine learning
Assumptions					
Categories are predefined	No	Yes	Yes	Yes	No
Relevant text features are known	Yes	Yes	Yes	Yes	Yes
Mapping between text features and categories is known	No	No	Yes	No	No
Costs					
Pre-analysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Low	High	High	High	Low
Analysis costs					
Person-hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate	Moderate	Low	Low	Low
Post-analysis costs					
Person hours spent interpreting	Moderate	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Source: Debortoli et al. (2016)

PADERBORN UNIVERSITY

Text Categorization

	Manual coding (bottom up)	Manual coding (top down)	Dictionaries	Supervised machine learning	Unsupervised machine learning
Assumptions					
Categories are predefined	No	Yes	Yes	Yes	No
Relevant text features are known	Yes	Yes	Yes	Yes	Yes
Mapping between text features and categories is known	No	No	Yes	No	No
Costs					
Pre-analysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Low	High	High	High	Low
Analysis costs					
Person-hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate	Moderate	Low	Low	Low
Post-analysis costs					
Person hours spent interpreting	Moderate	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Source: Debortoli et al. (2016)

PADERBORN UNIVERSITY

Text Categorization

	Manual coding (bottom up)	Manual coding (top down)	Dictionaries	**Supervised machine learning**	Unsupervised machine learning
Assumptions					
Categories are predefined	No	Yes	Yes	Yes	No
Relevant text features are known	Yes	Yes	Yes	Yes	Yes
Mapping between text features and categories is known	No	No	Yes	No	No
Costs					
Pre-analysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Low	High	High	High	Low
Analysis costs					
Person-hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate	Moderate	Low	Low	Low
Post-analysis costs					
Person hours spent interpreting	Moderate	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Source: Debortoli et al. (2016)

FUNDAMENTALS OF TOPIC MODELING

PADERBORN UNIVERSITY

Text Categorization

	Manual coding (bottom up)	Manual coding (top down)	Dictionaries	Supervised machine learning	**Unsupervised machine learning**
Assumptions					
Categories are predefined	No	Yes	Yes	Yes	No
Relevant text features are known	Yes	Yes	Yes	Yes	Yes
Mapping between text features and categories is known	No	No	Yes	No	No
Costs					
Pre-analysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Low	High	High	High	Low
Analysis costs					
Person-hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate	Moderate	Low	Low	Low
Post-analysis costs					
Person hours spent interpreting	Moderate	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Source: Debortoli et al. (2016)

FUNDAMENTALS OF TOPIC MODELING

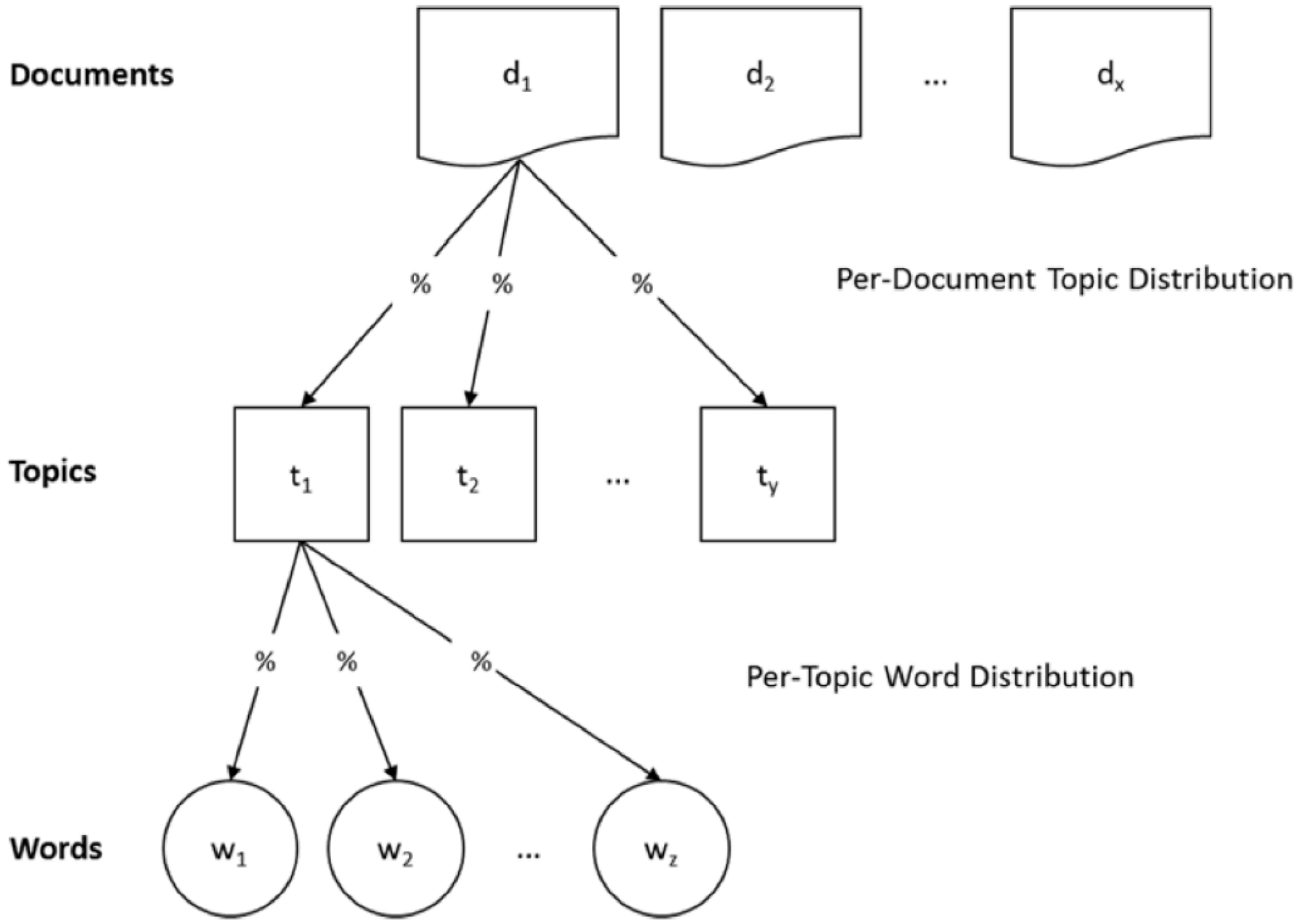
What are Topic Models?

- **Unsupervised** machine learning methods for text mining (e.g., Latent Semantic Analysis, Latent Dirichlet Allocation)

- Theoretical grounding: **Distributional hypothesis** of linguistics
  - Words that co-occur together in similar contexts (e.g., ball, goal, offside) tend to have similar meanings
  - Co-occurrence patterns can be interpreted as topics (e.g., football) and used to cluster documents

FUNDAMENTALS OF TOPIC MODELING

Schematic Overview of Probabilistic Topic Modeling with LDA



Source: Debortoli et al. (2016)

Illustrative Example of Probabilistic Topic Modeling with LDA

Exemplary Customer Review about a **Fitbit Flex**

I bought this for my 14 year old daughter as a gift. She received it in July. It works great - she lost 6 pounds in 2 weeks. The Fitbit makes staying in shape easy. The iPhone app works fine.

Topics

Birthday present

Loosing weight

Mobile app

FUNDAMENTALS OF TOPIC MODELING

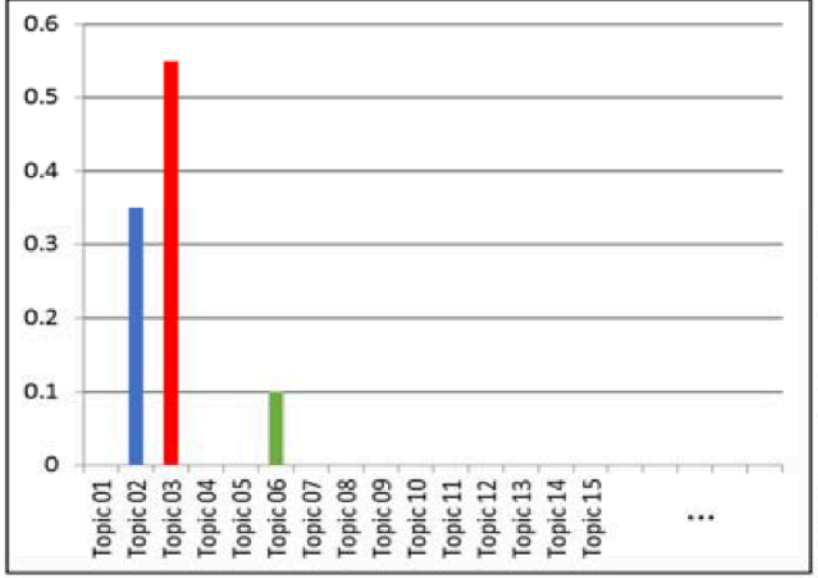
PADERBORN UNIVERSITY

Illustrative Example of Probabilistic Topic Modeling with LDA

Document

«I bought this for my 14 year old daughter as a gift. She received it in July. She lost 6 pounds in 2 weeks. The Fitbit makes staying in shape easy. The iPhone app works fine.»

Topic Distribution



Topic 01		Topic 02		Topic 03	
love	0,13	gift	0,10	weight	0,08
recommend	0,08	love	0,07	loss	0,05
thing	0,07	christmas	0,07	pounds	0,04
color	0,07	bought	0,06	lose	0,04
purchased	0,06	husband	0,05	week	0,02
band	0,06	daughter	0,04	lb	0,02
buy	0,03	received	0,03	month	0,02
mine	0,03	birthday	0,03	helped	0,01
amazing	0,03	present	0,02	eat	0,01
friend	0,02	son	0,02	goal	0,01

Topic 04		Topic 05		Topic 06	
battery	0,15	heart	0,11	app	0,12
day	0,09	rate	0,10	iphone	0,08
charge	0,07	monitor	0,07	sync	0,03
life	0,05	blood	0,02	ipad	0,02
week	0,03	pressure	0,02	work	0,01
time	0,02	pedometer	0,02	io	0,01
hour	0,02	measure	0,02	apple	0,01
low	0,01	tracking	0,01	android	0,01
recharge	0,01	device	0,01	computer	0,01
dead	0,01	glucose	0,01	update	0,01

... ...

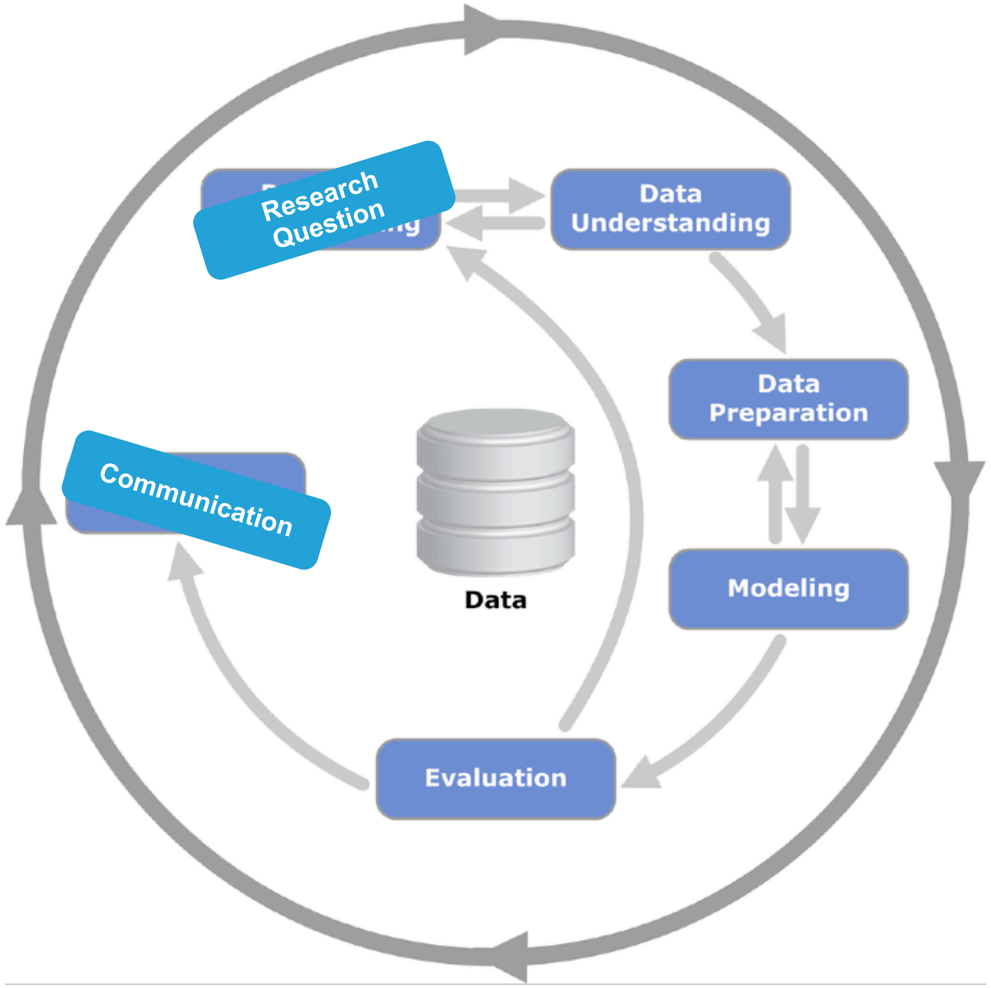
Source: Debortoli et al. (2016)

TOPIC MODELING WALKTHROUGH

tidyverse, quanteda, stm

TOPIC MODELING WALKTHROUGH

Cross Industry Standard Process for Data Mining



Source: Shearer et al. (2000)

PADERBORN UNIVERSITY

Research Question

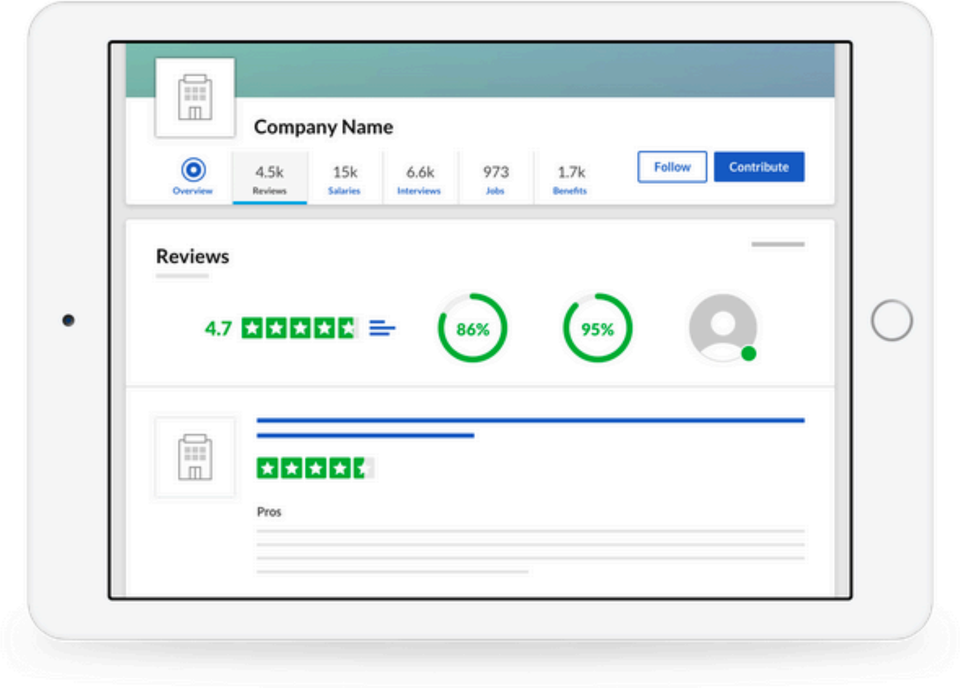
"What factors drive employees' company ratings?"

glassdoor

Company reviews and ratings. Get the whole story.

Search ratings and reviews of over 600,000 companies worldwide. Get the inside scoop and find out what it's really like from people who've actually worked there.

Write a Review



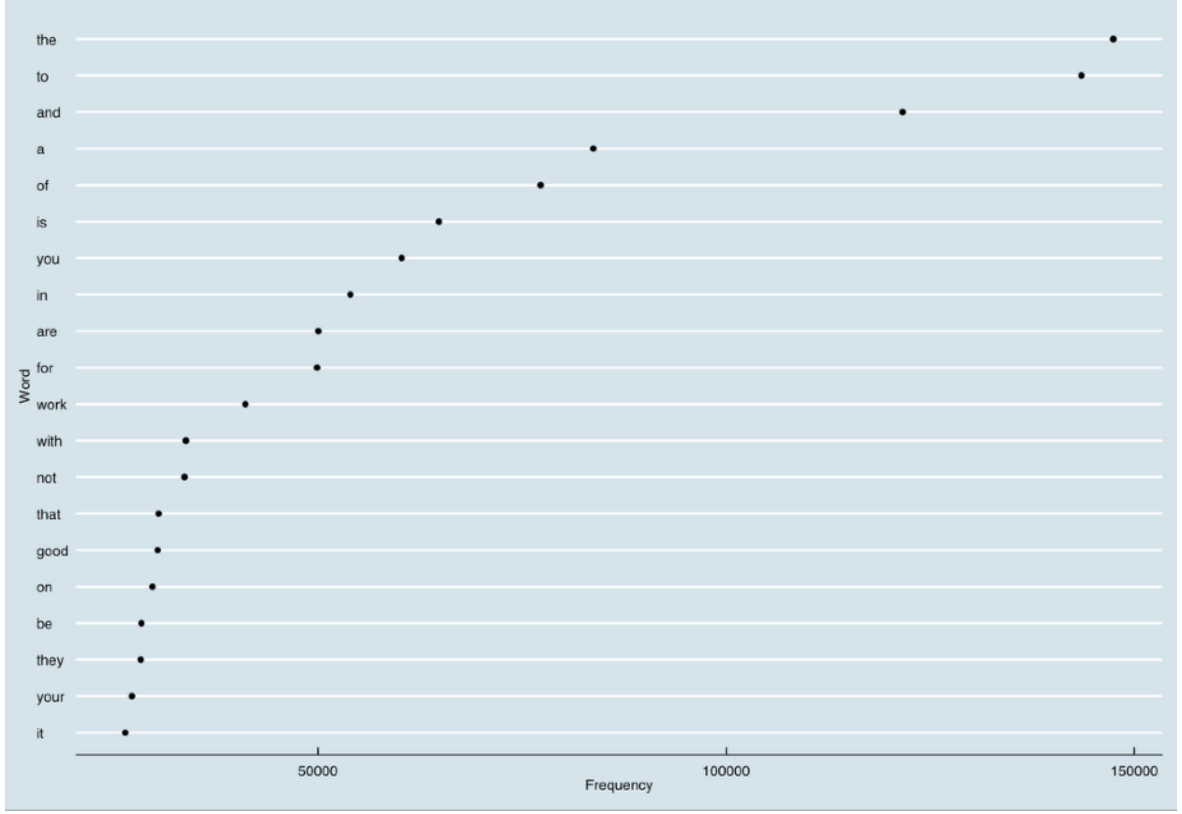
TOPIC MODELING WALKTHROUGH

Data Understanding


```
{
    "shortName": "Bank of America",
    "date": "2010-10-03",
    "stars": 3,
    "text": "Great benefits for associates, Paid maternity/paternity leave, most associates receive 3
        weeks of vacation leave per year (SSS, PB, AM and four weeks for BCM). Micro-management,
        poor leadership, lack of recognition, extremely under staffed. Do not forget the human aspect.
        Micro-management is not the answer to every situation. Put more people in the branches."
}
```

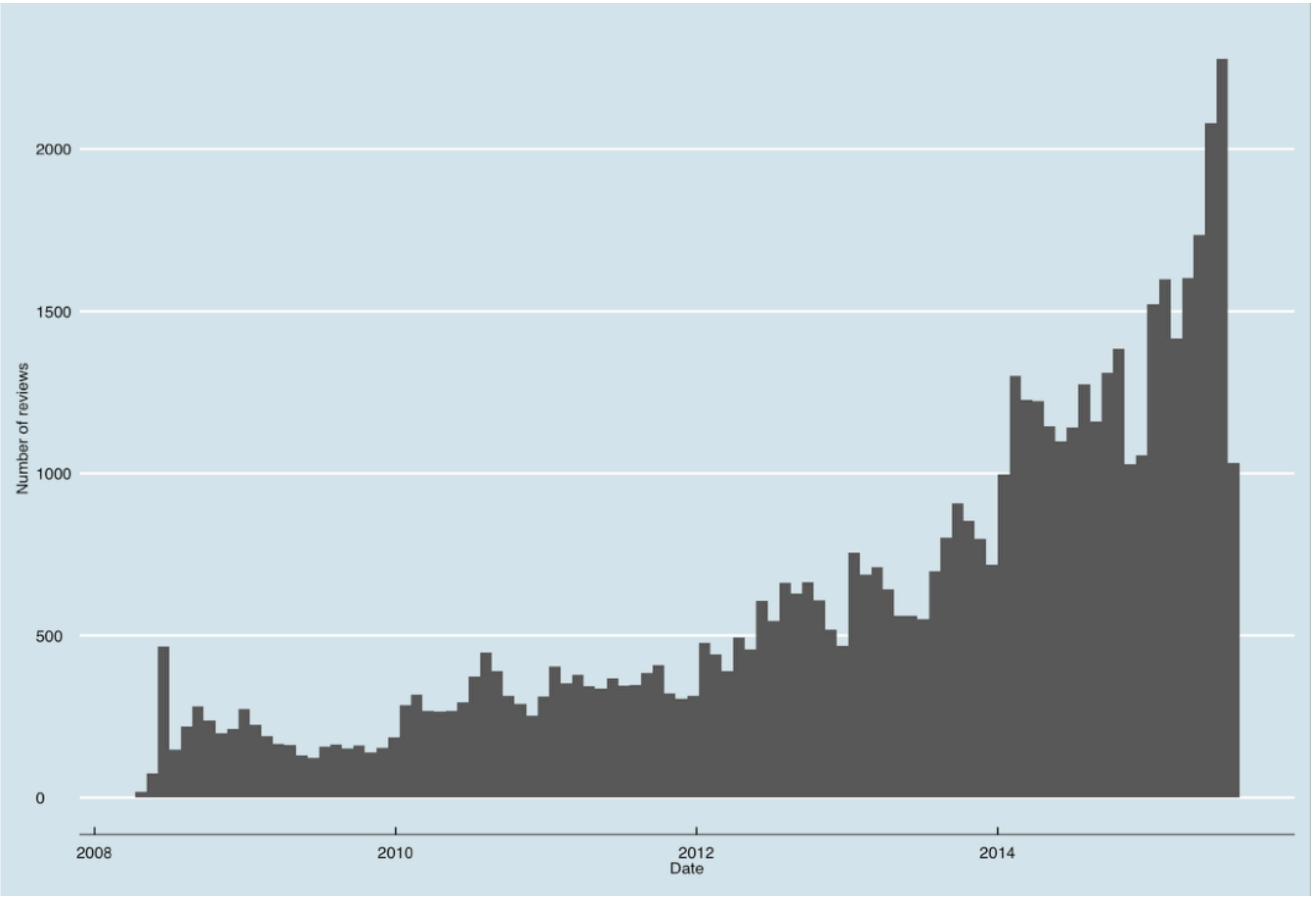
Data Understanding

- Sub-sample
  - Finance industry only
- Number of documents
  - 57,765
- Number or words (tokens)
  - 1,608,259
- Number of unique words
  - 1,740



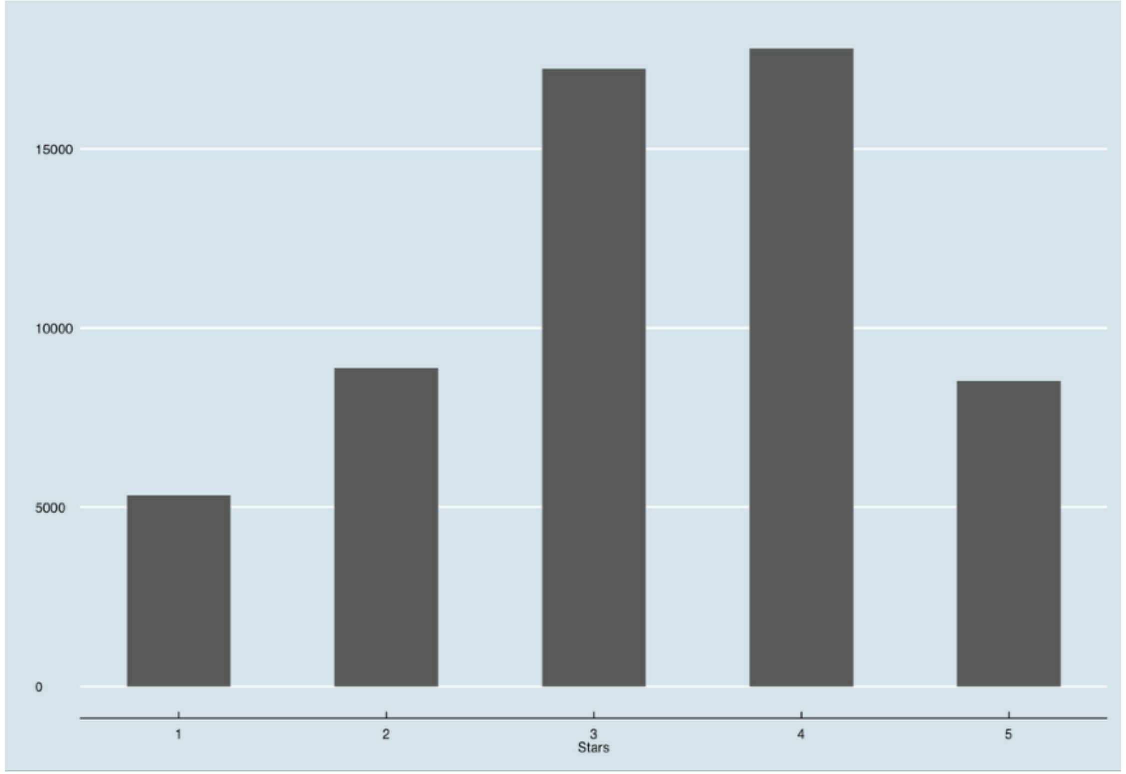
TOPIC MODELING WALKTHROUGH

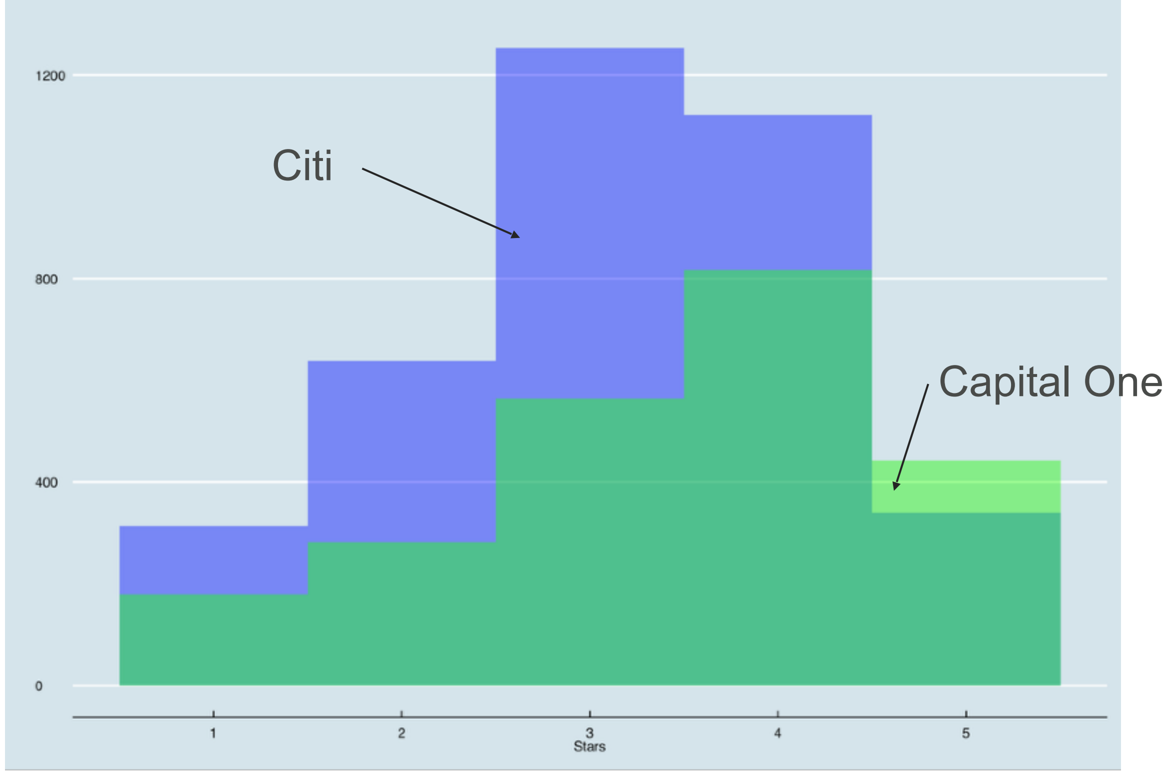
Data Understanding



TOPIC MODELING WALKTHROUGH

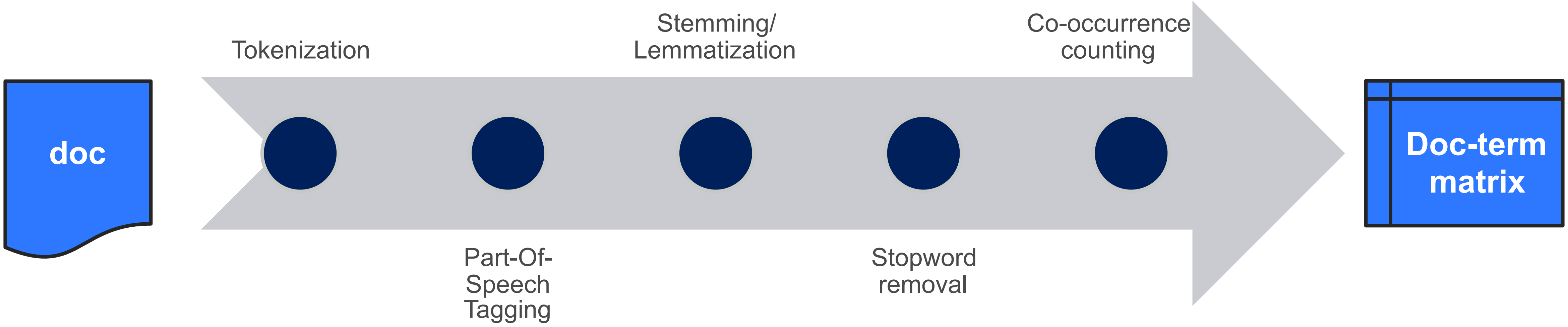
Data Understanding





PADERBORN UNIVERSITY

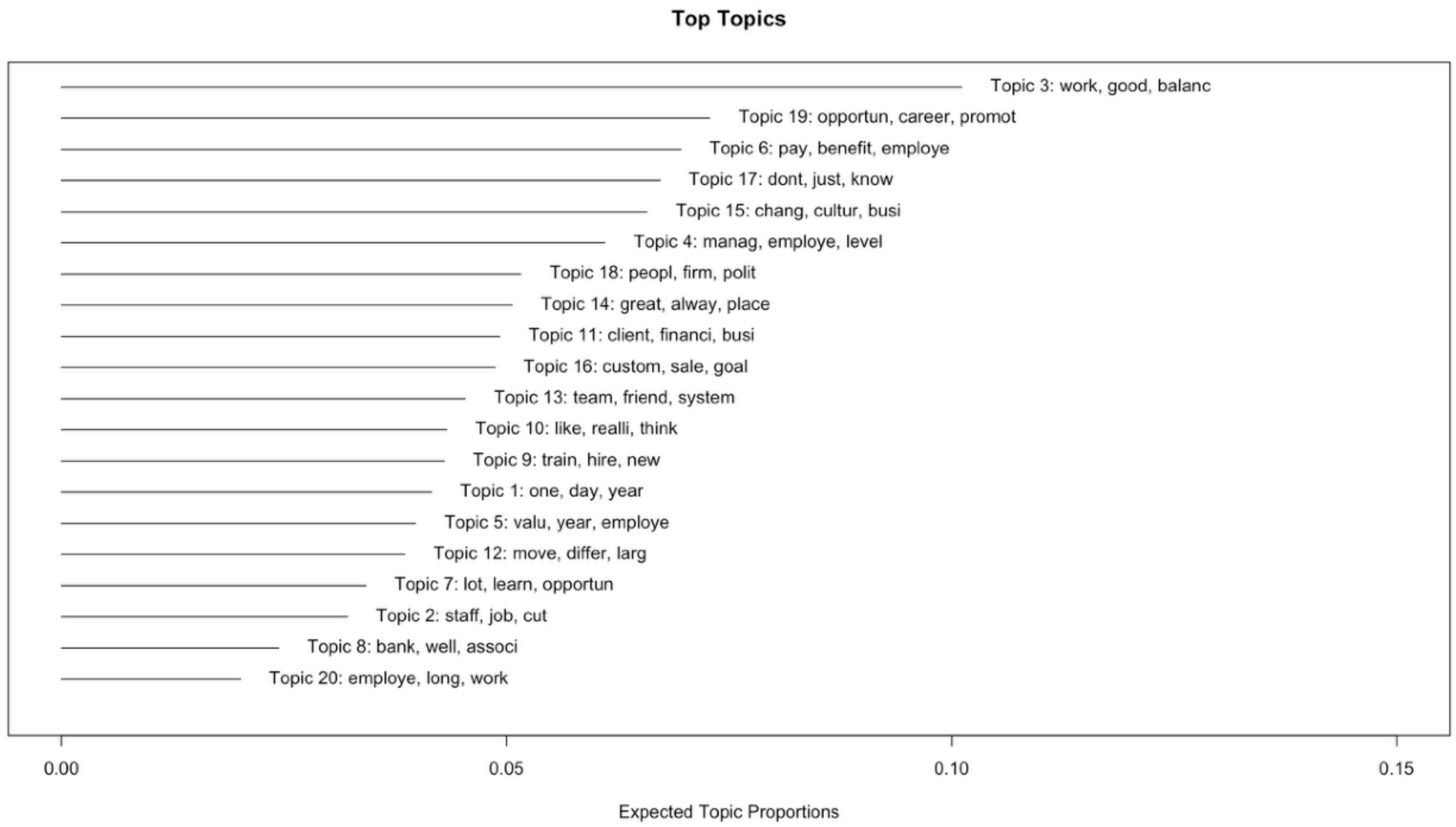
Data Preparation



TOPIC MODELING WALKTHROUGH

Modeling and Evaluation – Iteration #1

- Estimate 30 most prevalent topics
  - Takes approx. 10 minutes on a MacBook Pro
  - Wait for the “aha” effect



PADERBORN UNIVERSITY

Modeling and Evaluation – Iteration #1

- Estimate 30 most prevalent topics
  - Takes approx. 10 minutes on a MacBook Pro
  - Wait for the "aha" effect

Top-3 Documents for Topic 6

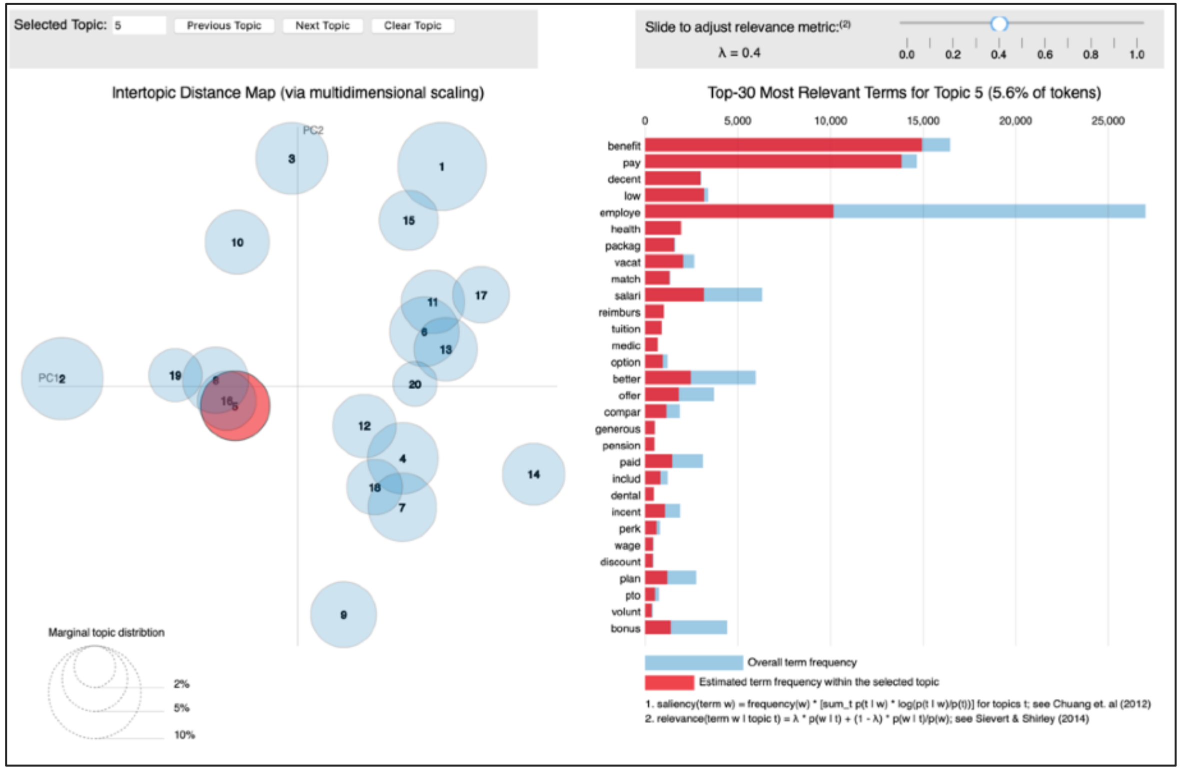
Health insurance, Vacation, sick pay, paid maternity leave 12 weeks. Every year perks decrease and are eliminated. Uneducated people with their nose in the air.
Decent benefits, decent bonuses, decent vacation/off time. Inadequate pay and inadequate coverage.
Huge Annual Bonus amount, Paid Overtime amount can be earned, 1 Time free meal & free transport facility. Less On paper CTC offered. Should include the approximate Annual Bonus & Gratuity in the offered CTC on-paper.

PADERBORN UNIVERSITY

Modeling and Evaluation – Iteration #2

- Find the right number of topics
  - **Manual** investigation
    - Are topics coherent? No duplicate topics? No fused topics?
    - Increase or reduce number of topics

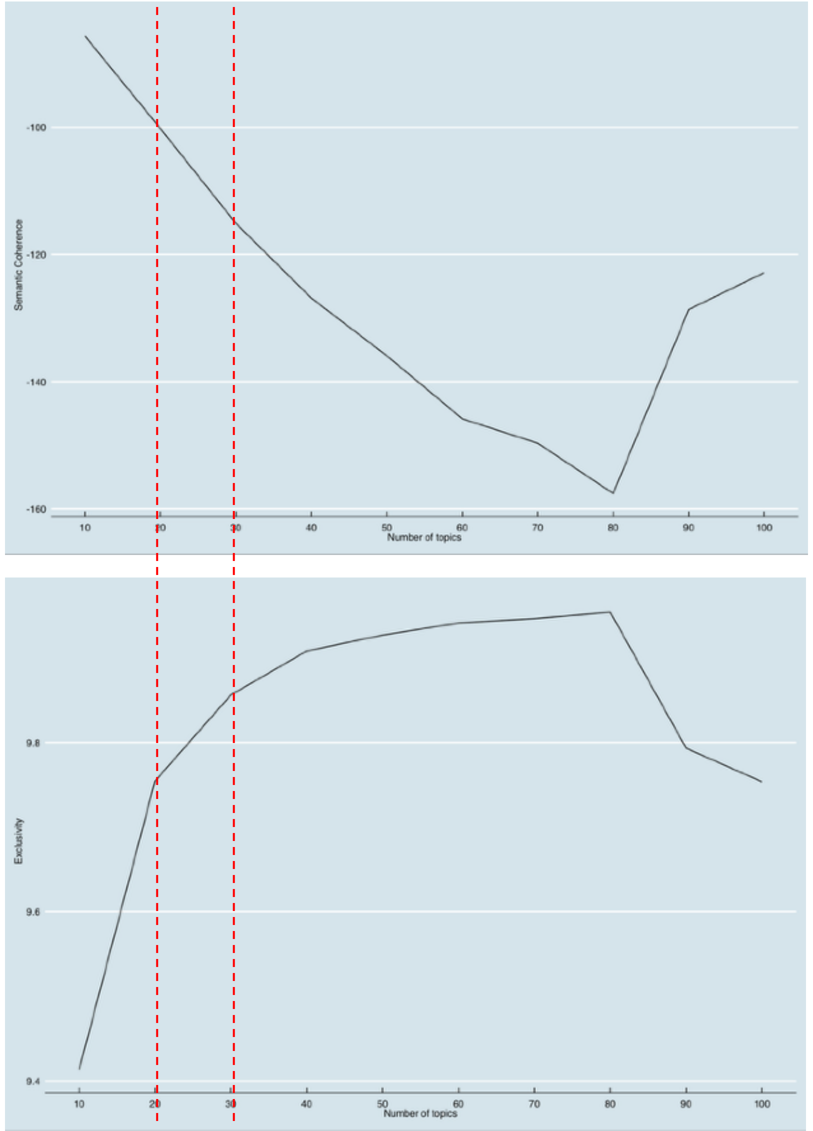
LDAvis: A R package for interactive topic model visualization.



TOPIC MODELING WALKTHROUGH

Modeling and Evaluation – Iteration #2

- Find the right number of topics
  - **Automated** search
    - e.g.: From 10 to 100 topics, in steps of 10
    - Takes several hours on a MacBook Pro
    - Evaluate models with regards to **Semantic Coherence** and **Exclusivity**



TOPIC MODELING WALKTHROUGH

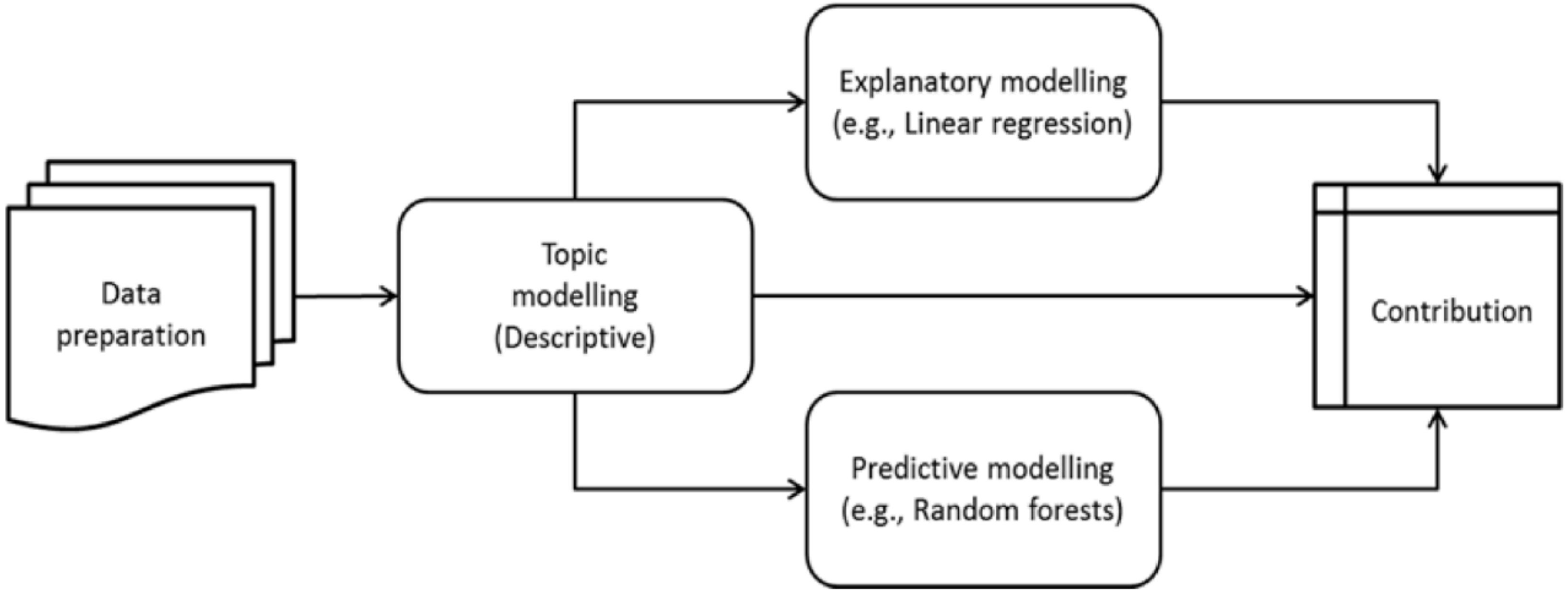
Modeling and Evaluation – Iteration #3

- Final experimental evaluation through human coders
  - Word intrusion task
  - Topic intrusion task

http://etc.ch/bRsu

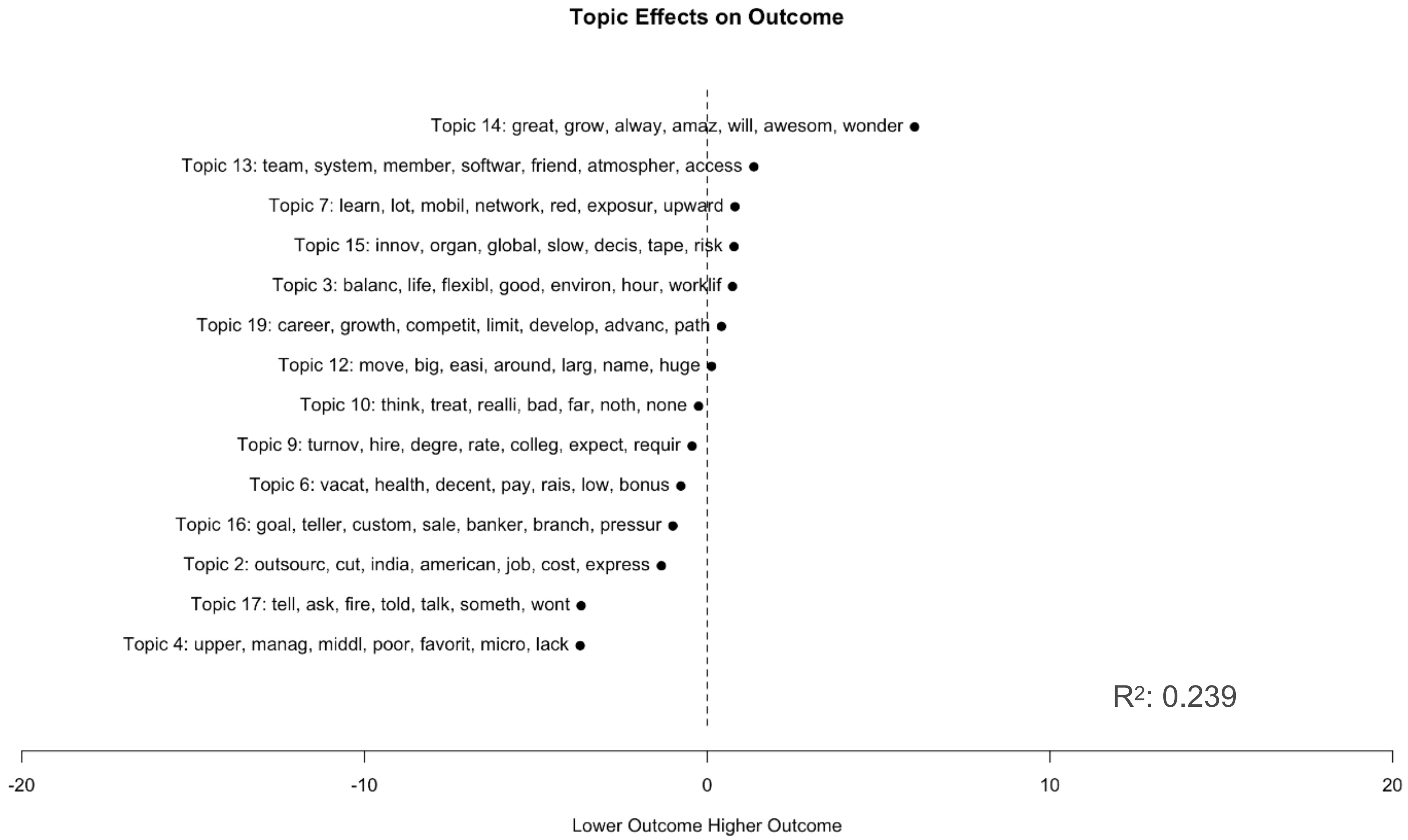
Modeling and Evaluation – Iteration #4

- Modeling the relationship between topics and stars



TOPIC MODELING WALKTHROUGH

Modeling and Evaluation – Iteration #4



PADERBORN UNIVERSITY

Communication

Communications of the Association for Information Systems

Tutorial ISSN: 1529-3181

Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial

Stefan Debortoli
University of Liechtenstein
Institute of Information Systems
Vaduz, Liechtenstein
stefan.debortoli@uni.li

Oliver Müller
IT University of Copenhagen
Information Management Section
Copenhagen, Denmark

Iris Junglas
Florida State University
College of Business
Tallahassee, FL, USA

Jan vom Brocke
University of Liechtenstein
Institute of Information Systems
Vaduz, Liechtenstein

Abstract:

Analysts have estimated that more than 80 percent of today's data is stored in unstructured form (e.g., text, audio, image, video)—much of it expressed in rich and ambiguous natural language. Traditionally, to analyze natural language, one has used qualitative data-analysis approaches, such as manual coding. Yet, the size of text data sets obtained from the Internet makes manual analysis virtually impossible. In this tutorial, we discuss the challenges encountered when applying automated text mining techniques in information systems research. In particular, we showcase how to use probabilistic topic modeling via Latent Dirichlet allocation, an unsupervised text mining technique, with a LASSO multinomial logistic regression to explain user satisfaction with an IT artifact by automatically analyzing more than 12,000 online customer reviews. For fellow information systems researchers, this tutorial provides guidance for conducting text-mining studies on their own and for evaluating the quality of others.

Keywords: Text Mining, Topic Modeling, Latent Dirichlet Allocation, Online Customer Reviews, User Satisfaction.

This manuscript underwent editorial review. It was received 05/26/2015 and was with the authors for 8 months for 3 revisions. The Associate Editor chose to remain anonymous.

Volume 39 Paper 7 pp. 110 – 135 July 2016

Check for updates

Article

Organizational Research Methods
1-28
© The Author(s) 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1094428118773858
journals.sagepub.com/home/orm
SAGE

Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture

Theresa Schmiedel[1], Oliver Müller[2], and Jan vom Brocke[1]

Abstract

Research has emphasized the limitations of qualitative and quantitative approaches to studying organizational phenomena. For example, in-depth interviews are resource-intensive, while questionnaires with closed-ended questions can only measure predefined constructs. With the recent availability of large textual data sets and increased computational power, text mining has become an attractive method that has the potential to mitigate some of these limitations. Thus, we suggest applying topic modeling, a specific text mining technique, as a new and complementary strategy of inquiry to study organizational phenomena. In particular, we outline the potentials of structural topic modeling for organizational research and provide a step-by-step tutorial on how to apply it. Our application example builds on 428,492 reviews of *Fortune* 500 companies from the online platform Glassdoor, on which employees can evaluate organizations. We demonstrate how structural topic models allow to inductively identify topics that matter to employees and quantify their relationship with employees' perception of organizational culture. We discuss the advantages and limitations of topic modeling as a research method and outline how future research can apply the technique to study organizational phenomena.

Keywords
topic modeling, structural topic model, tutorial, organizational culture, online reviews

Introduction

Organizational research follows both quantitative and qualitative research paradigms and applies various empirical methods for data collection and analysis (Currall, Hammer, Baggett, & Doniger,

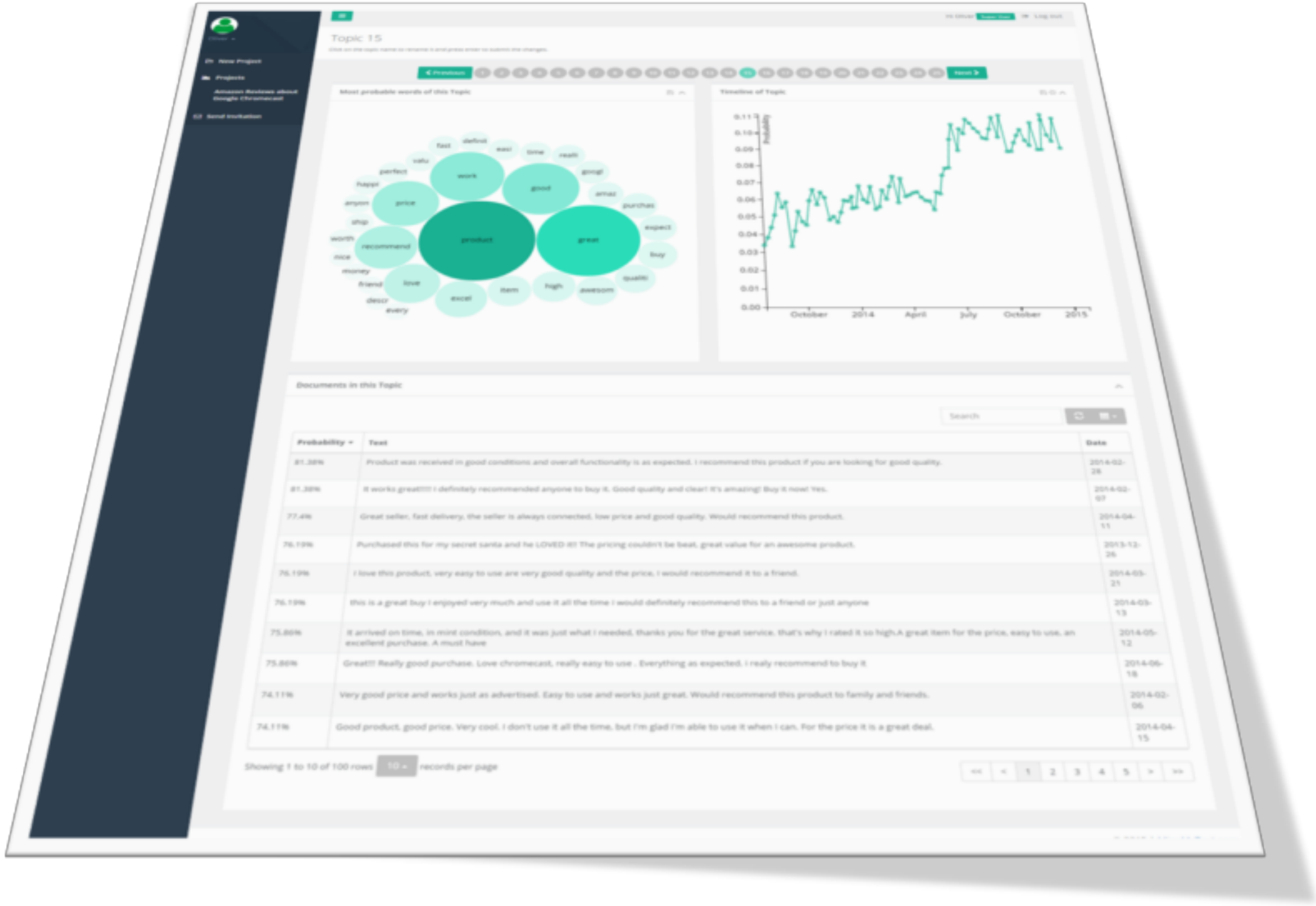
[1]University of Liechtenstein, Liechtenstein
[2]IT University of Copenhagen, Copenhagen, Denmark

Corresponding Author:
Theresa Schmiedel, University of Liechtenstein, Fuerst-Franz-Josef-Strasse, 9490 Vaduz, Principality of Liechtenstein.
Email: theresa.schmiedel@uni.li

TOPIC MODELING WALKTHROUGH

PADERBORN UNIVERSITY

www.MineMyText.com



THE END

Prof. Dr. Oliver Müller
Lehrstuhl für Wirtschaftsinformatik, insb. Data Analytics
Universität Paderborn
Warburger Str. 100, 33098 Paderborn
R: Q2.457
E: oliver.mueller@uni-paderborn.de
T: +49-5251-605245
W: https://wiwi.uni-paderborn.de/dep3/mueller/

- Cox M, Ellsworth D (1997) Application-controlled demand paging for out-of-core visualization. Proceedings of the 8th Conference on Visualization, 235–244.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70).
- IBM. (2012). Analytics: The real-world use of big data. Retrieved from http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html
- Dhar, V. (2013). Data Science and Prediction. Communications of the ACM, 56 (12), 64–73.
- Zur Mühlen, M., & Shapiro, R. (2010). Business process analytics. In J. Vom Brocke & M. Rosemann (Eds.), Handbook on Business Process Management (Vol. 2). Springer.
- Mitchell, T. M. (1997). Machine learning. McGraw Hill.
- Shearer C. (2000). The CRISP-DM model: the new blueprint for data mining, Journal of Data Warehousing, 5(4), 13-22.
- Debortoli, S., Müller, O., Junglas, I. A., & vom Brocke, J. (2016). Text mining for information systems researchers: an annotated topic modeling tutorial. Communication of the AIS, 39, 7.
- Schmiedel, T., Müller, O., & vom Brocke, J. (2018). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. Organizational Research Methods, 1094428118773858.

- Icons: questions by Depb Dew, Jemis Mali from the Noun Project