Extracting latent topics in large text corpora by taking metadata into account:

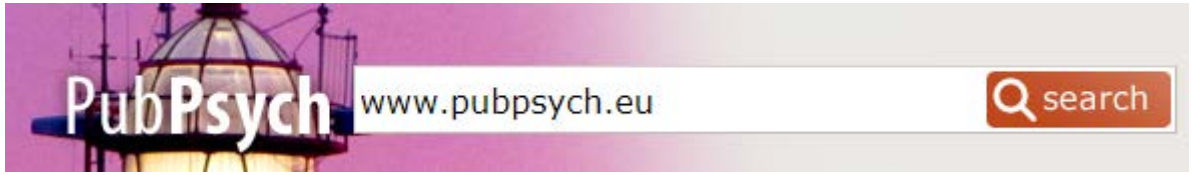
Do empirical studies address psychological research topics differently than nonempirical studies do?

André Bittermann
abi@leibniz-psychology.org

How to extract latent topics within large text corpora?

e. g. **PSYNDEX** database of psychological literature:

more than 340,000 documented publications



Topic Modeling

Idealization!

document			
1	2	3	4
love	happiness	disgust	amazement
hate	joy	anger	surprise
fear	serenity	rage	joy
disgust	love	hate	happiness
intervention	therapy	psychoanalysis	psychotherapy
therapist	therapist	transference	counseling
client	client	client	disorder
disorder	treatment	disorder	treatment
mother	intervention	treatment	outcome
brother	disorder	intervention	exposition
sister	parents	mother	client
father	siblings	father	therapist
school	learning	parents	parents
learning	teacher	child	mother
grades	class	grades	college
class	college	achievement	university

Topic content

topic			
1	2	3	4
love	client	parents	class
joy	disorder	mother	college
happiness	therapist	father	grades
disgust	intervention	child	learning
amazement	treatment	brother	teacher

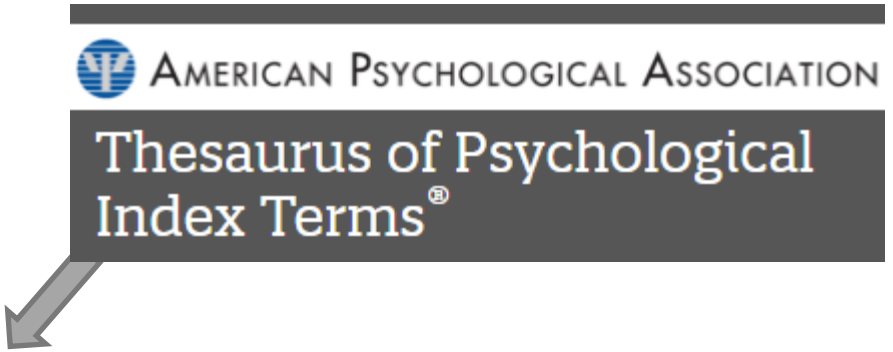
Topic prevalence

	topic				
document	1	2	3	4	sum
1	0.250	0.250	0.250	0.250	1
2	0.250	0.375	0.125	0.250	1
3	0.250	0.375	0.250	0.125	1
4	0.250	0.500	0.125	0.125	1
mean	0.250	0.375	0.188	0.188	1

Do empirical studies address psychological research topics differently than nonempirical studies do?

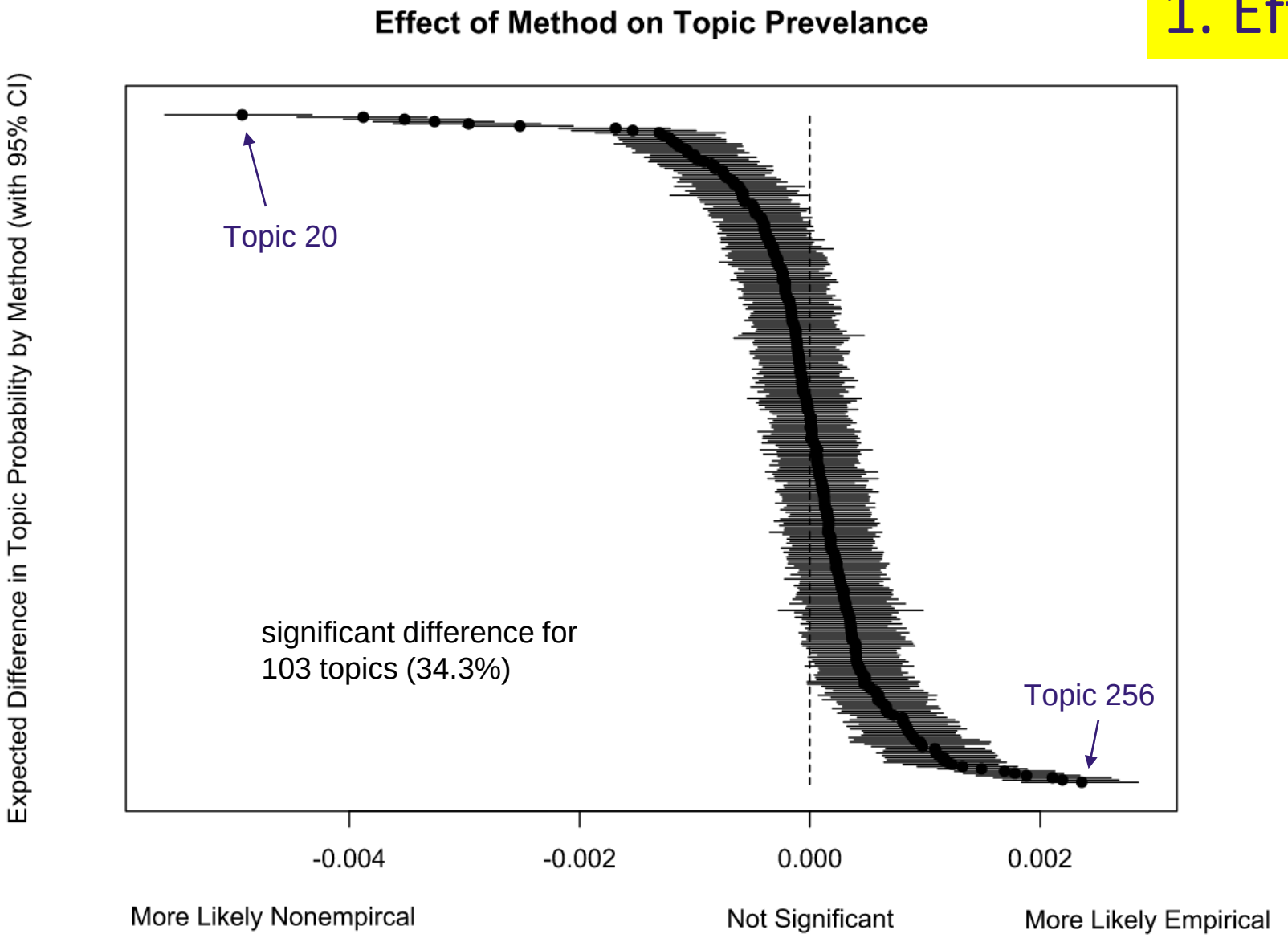
effects on topic
1. prevalence
2. content

Method



Corpus	controlled terms of *N* = 237,625 publications (PSYNDEX)
Metadata (Covariates)	- study methodology (empirical vs. nonempirical) - publication year (1980–2016)
Method	Structural Topic Modeling (Roberts et al., 2014)
Preprocessing	no stemming, no stopwords, no thresholds for sparse terms
Optimal *k*	via heldout likelihood
Software	stm (1.3.3) package for R

1. Effects on Topic Prevalence



Step 1: **linear regression**
topic prevalence ~ method

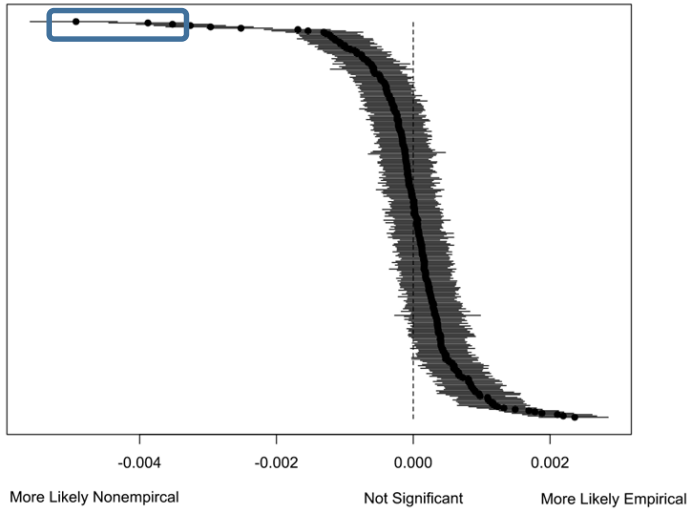
Step 2: for each topic
calculate the **difference**
between
- topic probability for empirical studies
- topic probability for nonempirical studies

Step 3: **plot** the differences
(and 95% CI)

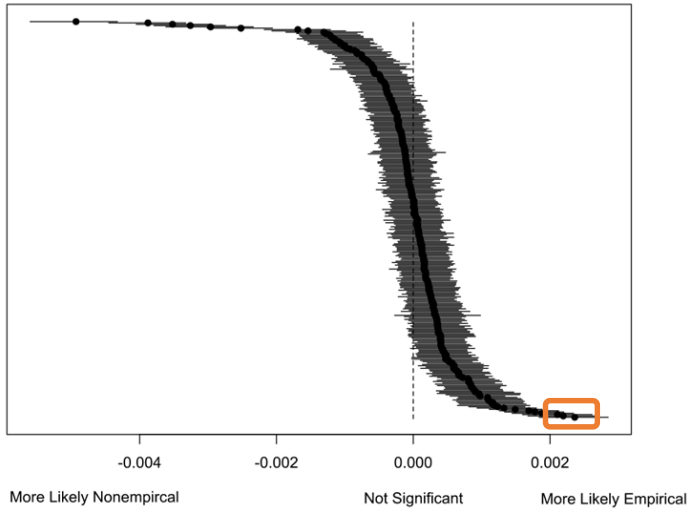
Topic 20:
Individual Psychotherapy
Eye Movement Desensitization Therapy
Brief Psychotherapy
Directed Reverie Therapy
Schema

Topic 56:
Activity Theory
Psychologists
Applied Psychology
Functionalism
Adlerian Psychotherapy

Topic 57:
Rationalization (Defense Mechanism)
Castration Anxiety
Reaction Formation
Omnipotence
Seduction



Most likely addressed by nonempirical publications



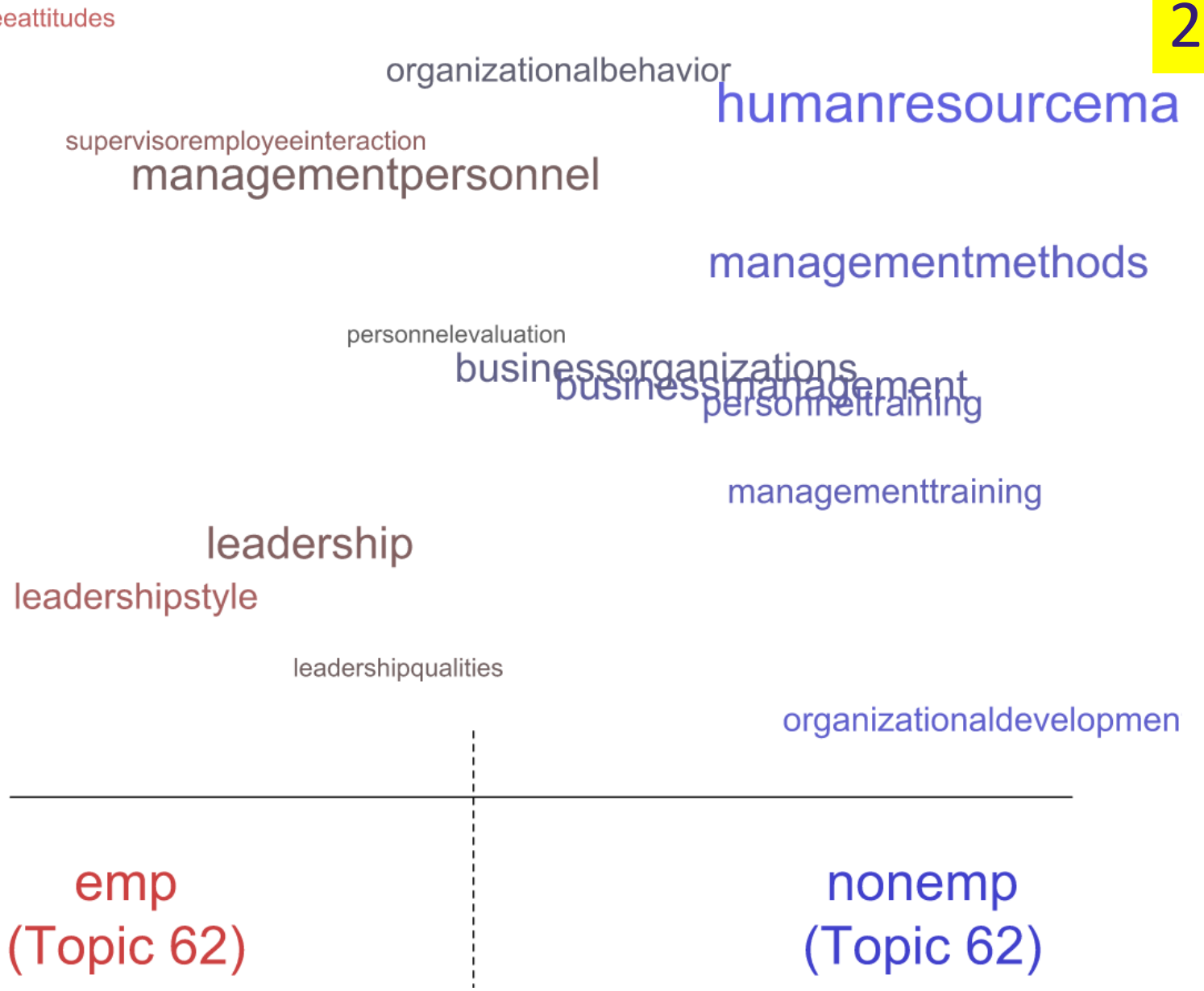
Topic 132:
Digital Span Testing
Dual Task Performance
Visual Memory
Visuospatial Memory
Cognitive Processing Speed

Topic 246:
Alpha Rhythm
Electroretinography
Gamma Rhythm
Afferent Stimulation
Delta Rhythm

Topic 256:
Academic Self-Concept
Ability Grouping
Academic Overachievement
Academic Achievement Motivation
Academic Achievement Prediction

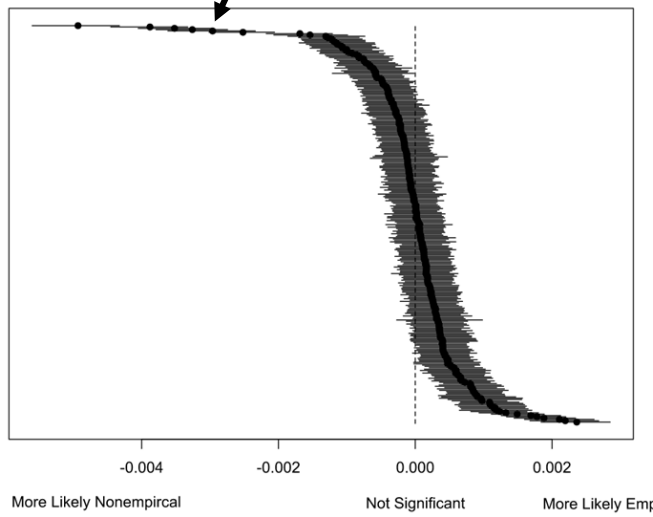
Most likely addressed by empirical publications

2. Effects on Topic Content

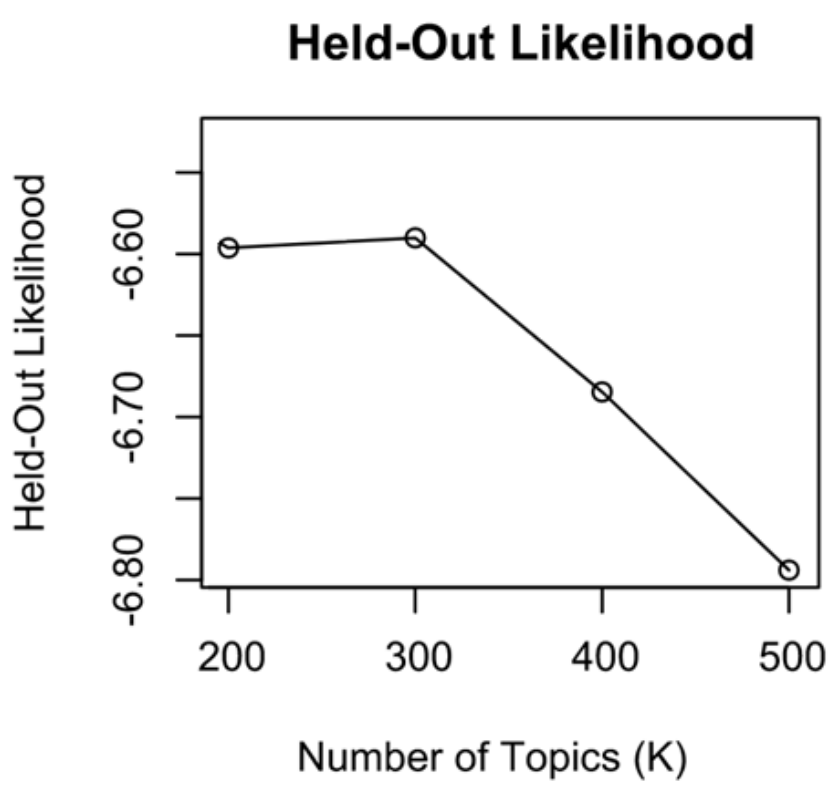


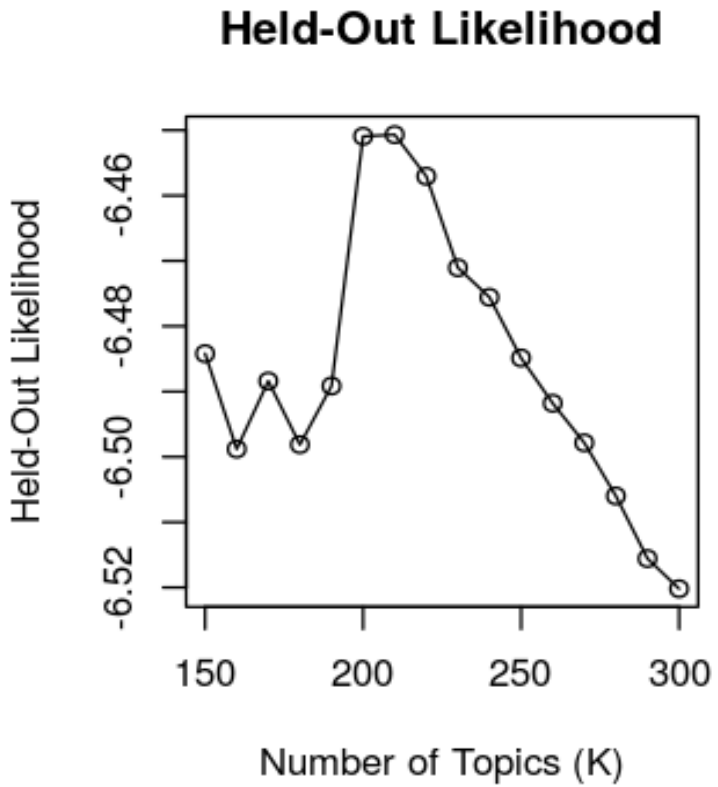
Topic 62:

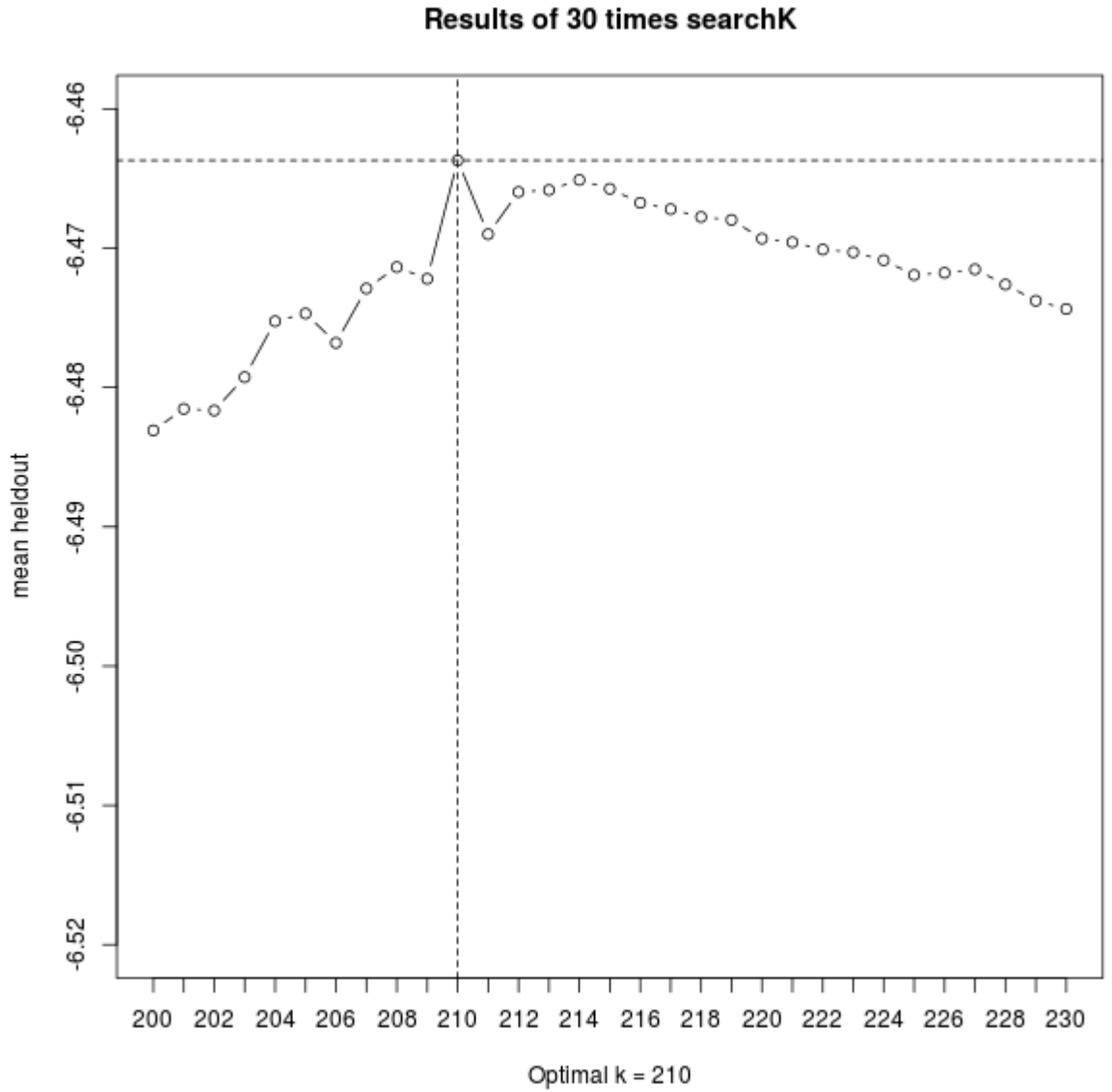
„Leadership & Business Management“



optimal *k*?


Held-Out Likelihood
Held-Out Likelihood
-6.80
-6.70
-6.60
200
300
400
500
Number of Topics (K)


Held-Out Likelihood
Held-Out Likelihood
-6.52
-6.50
-6.48
-6.46
150
200
250
300
Number of Topics (K)



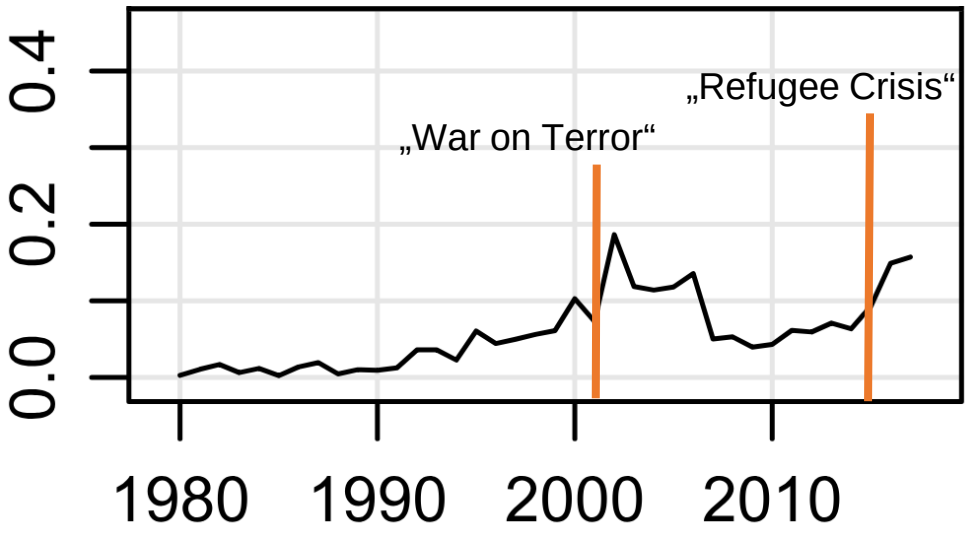
optimizing optimal *k*

30 different sets of training and test data
(by setting different heldout seeds)

Main conclusions

1. blind spots / research desiderata vs. focus areas of empirical research
2. other **covariates**: author information (e. g., gender, profession, …)
3. the issue with **optimal *k*** in large and heterogenous corpora
4. groundwork for manuscript:

   “Does psychological research address current social issues?”



Topic 12:
posttraumatic, refugees traumatized, PTSD, torture_victims

References

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science, 58* (4), 1064–1082. https://doi.org/10.1111/ajps.12103