INTRODUCTION TO DIGITAL HUMANITIES

APPLICATIONS OF DH: LANGUAGE AND LITERATURE TOPIC MODELLING

Dr. Thomas Köntges,
Digital Humanities, University of Leipzig

10 Novemver 2015

WHAT IS TOPIC-MODELLING

- Collective knowledge continues to grow; it becomes more difficult to find what one is looking for
- Topic Models more than search&link-approach
- Zooming in and out is possible with TMs
- Topic Model algorithms do not require prior annotation or labelling of the texts
- Topic Models discover the hidden thematic structure in large archives of documents

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

WHAT IS TOPIC-MODELLING

- A method to find clusters of words in large bodies of text
- Those clusters are called topics
- A topic is recurring pattern co-occurring words
- Topic models are probabilistic models that are often based on the number of topics in the corpus being assumed and fixed

WHAT IS TOPIC-MODELLING

http://thomask81.github.io/INNZ_tm_vis/index.html#topic=0&lambda=1&term=

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

MORE (BETTER) DATA AND SIMPLE ALGORITHMS OFTEN BEAT **COMPLEX ANALYTICS MODELS**

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

"We don't have better algorithms. We just have more data " - Peter Norvig (Google)

[Banko and Brill, 2001]

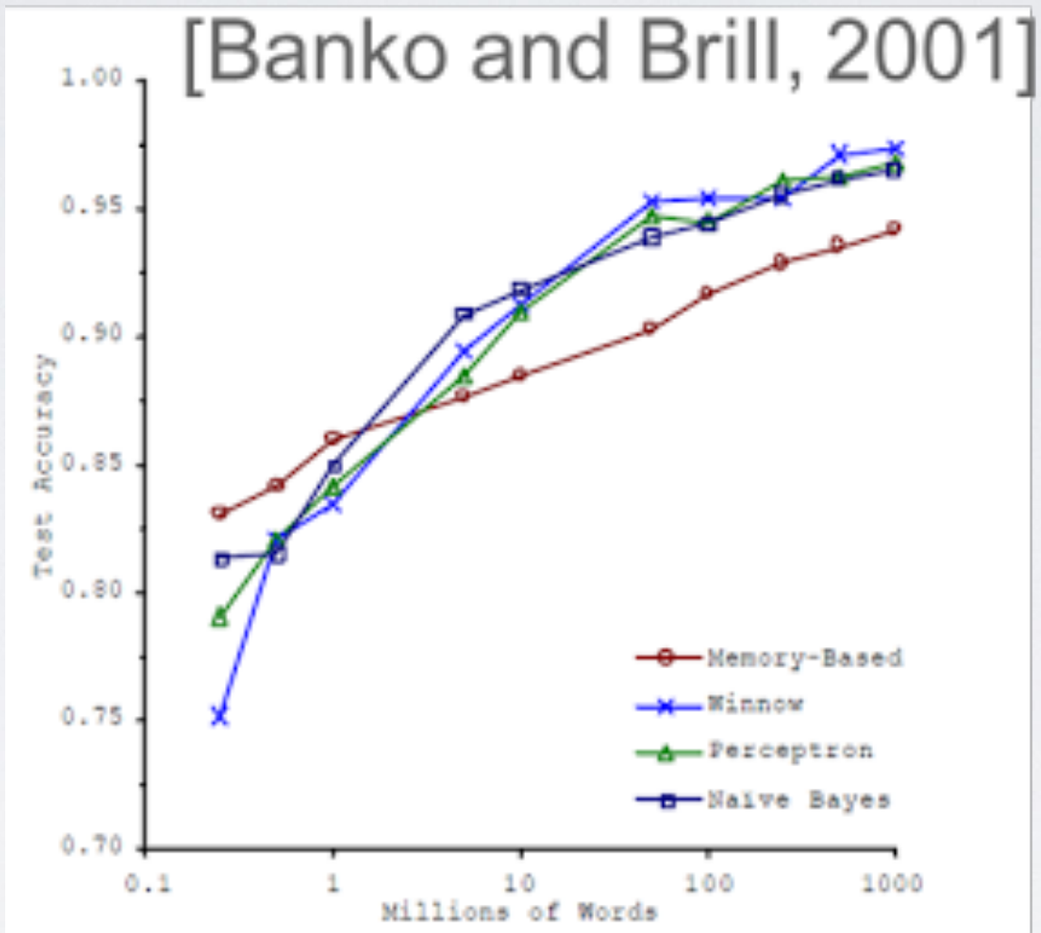


Figure 1. Learning Curves for Confusion Set Disambiguation

SO LET'S START WITH A SIMPLE ALGORITHM

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

LATENT DIRICHLET ALLOCATION

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

STATISTICAL INFERENCE

Deducing properties of an underlying distribution by analysis of data.

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

LATENT DIRICHLET ALLOCATION (LDA)

Assumes that one can find a Dirichlet Distribution by analysing the words in a corpus.

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

DIRICHLET DISTRIBUTION

Distribution on a probability distribution

Dirichlet dist.

$\theta = (\theta_1, \dots, \theta_n)$ $\quad \alpha = (\alpha_1, \dots, \alpha_n), \alpha_i > 0$

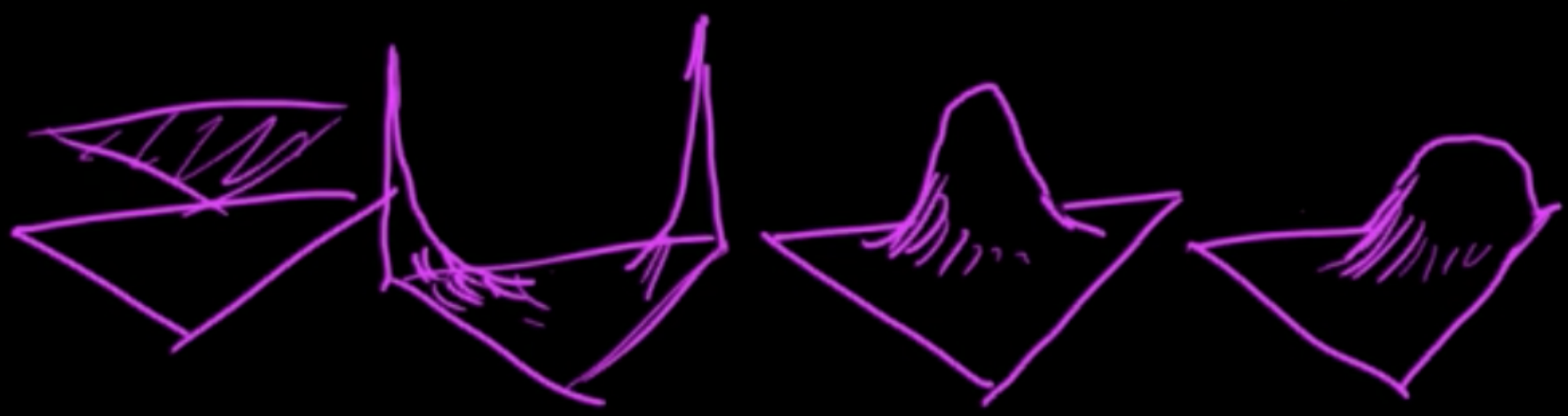
$\alpha_0 = \sum_{i=1}^{n} \alpha_i$.

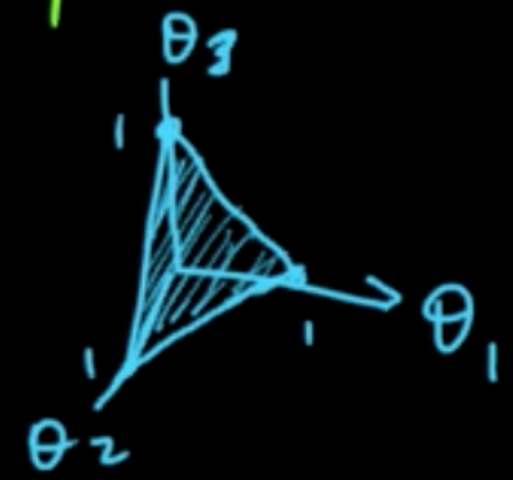
$$\theta \sim \mathrm{Dir}(\alpha) \qquad p(\theta) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} \theta_i^{\alpha_i - 1} I(\theta \in S)$$

$$\frac{1}{B(\alpha)} = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)}.$$

$$S = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^{n} x_i = 1\}.$$

↳ prob. simplex.





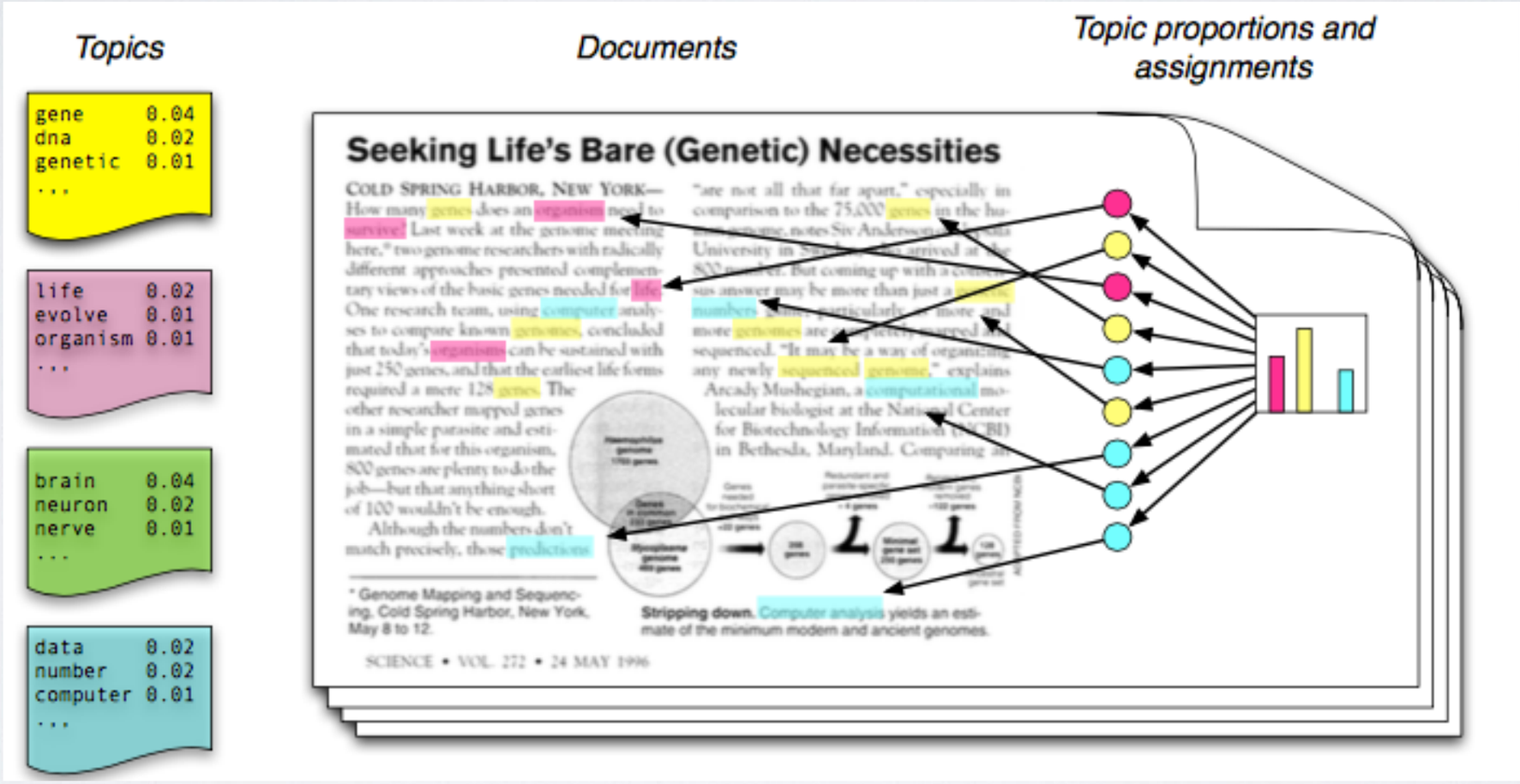
DIRICHLET DISTRIBUTION OF WORDS IN CORPUS

1. For each document draw a topic distribution

2. For each word in the document.
   a. Draw a specific topic
   b. Draw a word

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

DIRICHLET DISTRIBUTION OF WORDS IN CORPUS



Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

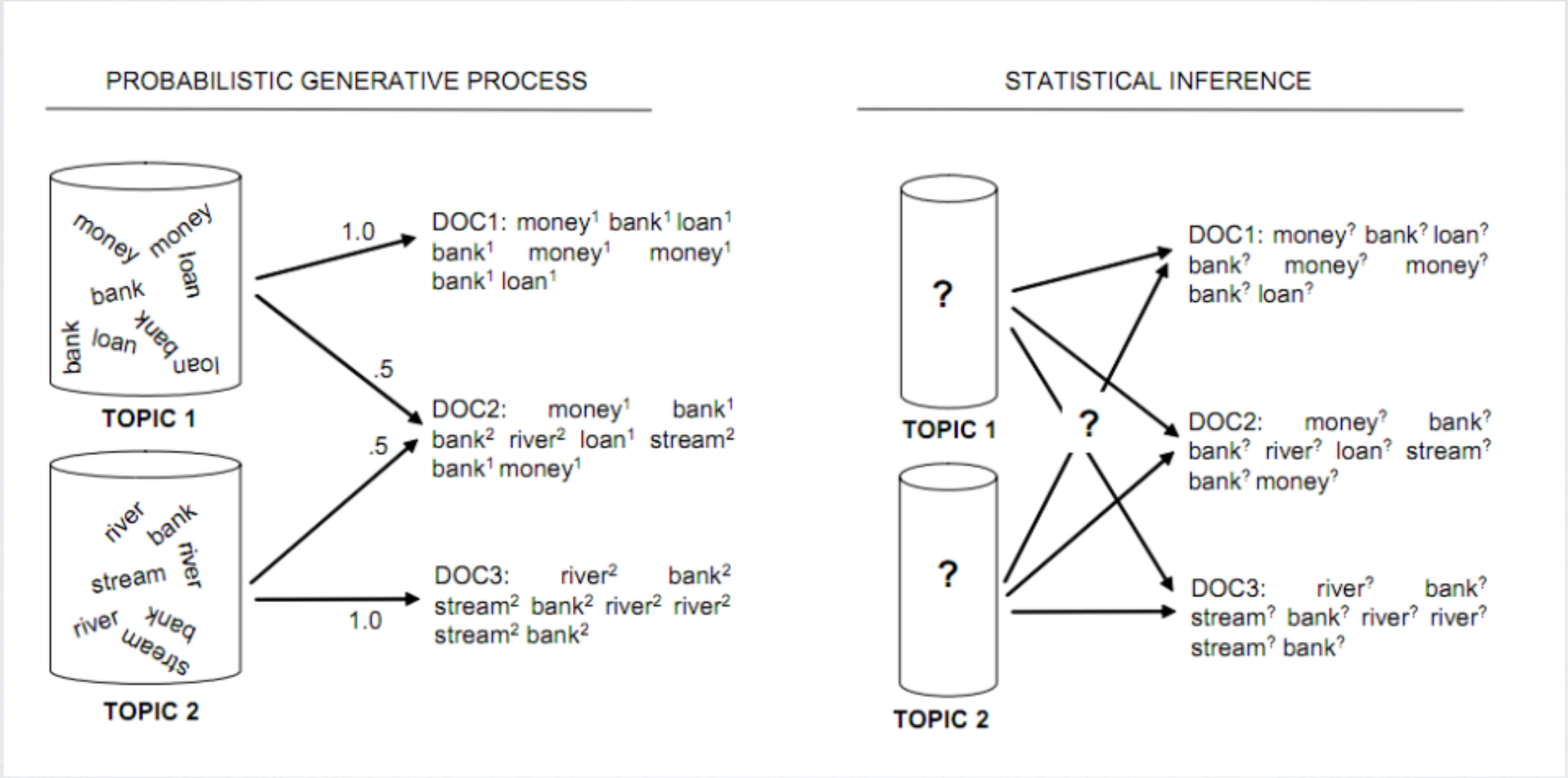
Digital Humanities
UNIVERSITÄT LEIPZIG

LATENT DIRICHLET ALLOCATION

- Simplification of how the documents in a dataset were created using Dirichlet Distribution
- Documents are Bag-of-words (order of the words in each document is irrelevant)
- Each corpus has words from a number of known topics
- We do not know which words belong to which topic
- “panel” in document A and “panel” in document B are not the same “word”

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

LATENT DIRICHLET ALLOCATION



Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

LATENT DIRICHLET ALLOCATION

- The central goal of topic modelling is to automatically discover the topics from a collection of documents.
- The central inferential problem for LDA is determining the posterior distribution of the latent variables given the document

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

SUCCESS AND RESULTS OF LDA RELY ON APRIORI-SET VARIABLES

- Number of topics
- Number of iterations
- Normalisation of the data
- Stopwords

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

A PRACTICAL EXAMPLE

- https://github.com/ThomasK81/TopicModellingR

WHAT DOES EACH VARIABLE DO?

- Search Text
- Collection
- Year
- Additional Stopwords
- Number of Topics
- Numbers of Terms Shown
- Iterations

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

ZOOMING IN AND OUT?

Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

THE RISK OF NORMALISATION

- http://thomask81.github.io/Greek_vis/#topic=4&lambda=1&term=
- ArabicMorph
- ArabicTranslated
- ArabicTM

TOPIC MODELLING IS NOT AN END, BUT A MEANS TO AN END

- https://www.google.com/fusiontables/DataSource?docid=1HuEta4_g-S01X9LEow8F6llidt8s5u5_GKkur_so&pli=1#card:id=2

- https://www.google.com/fusiontables/DataSource?docid=1fppC5YlmhtxbtDWiEwbgq_N2FCQQKQKTtN8rPJV9#card:id=2

USE CASE: CARTOONS VERSUS ARTICLES

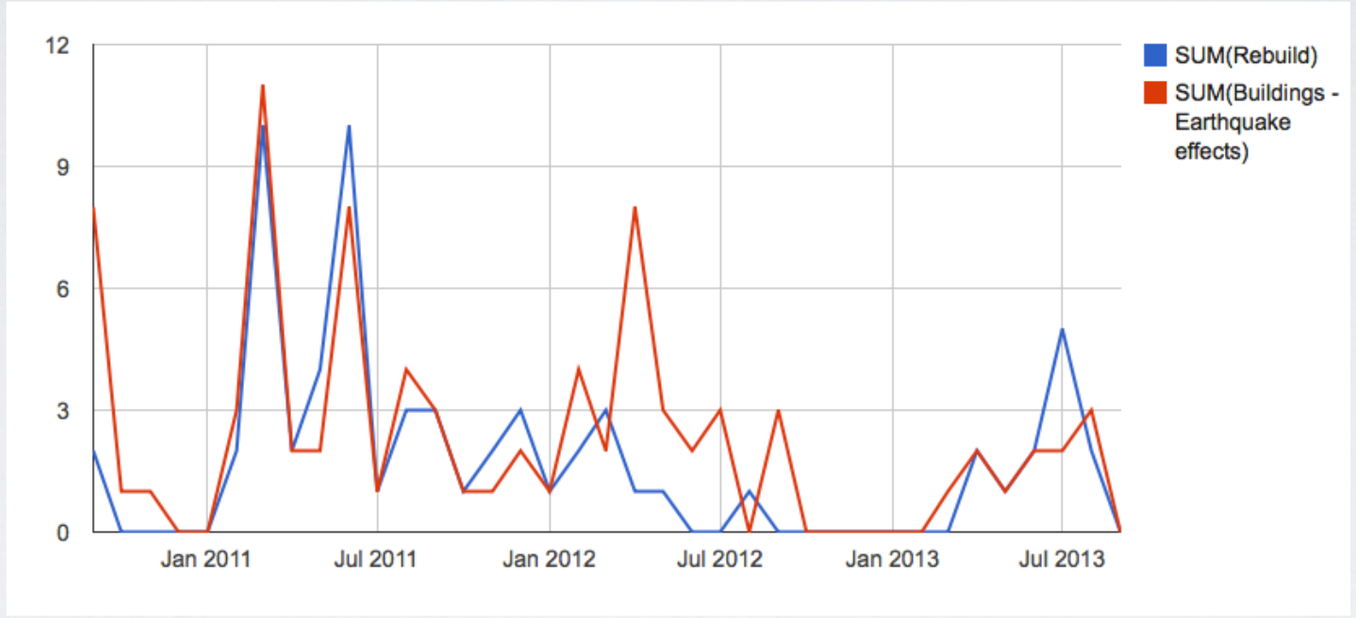
- Looking at the perception of the Canterbury EQ in published articles and published editorial cartoons

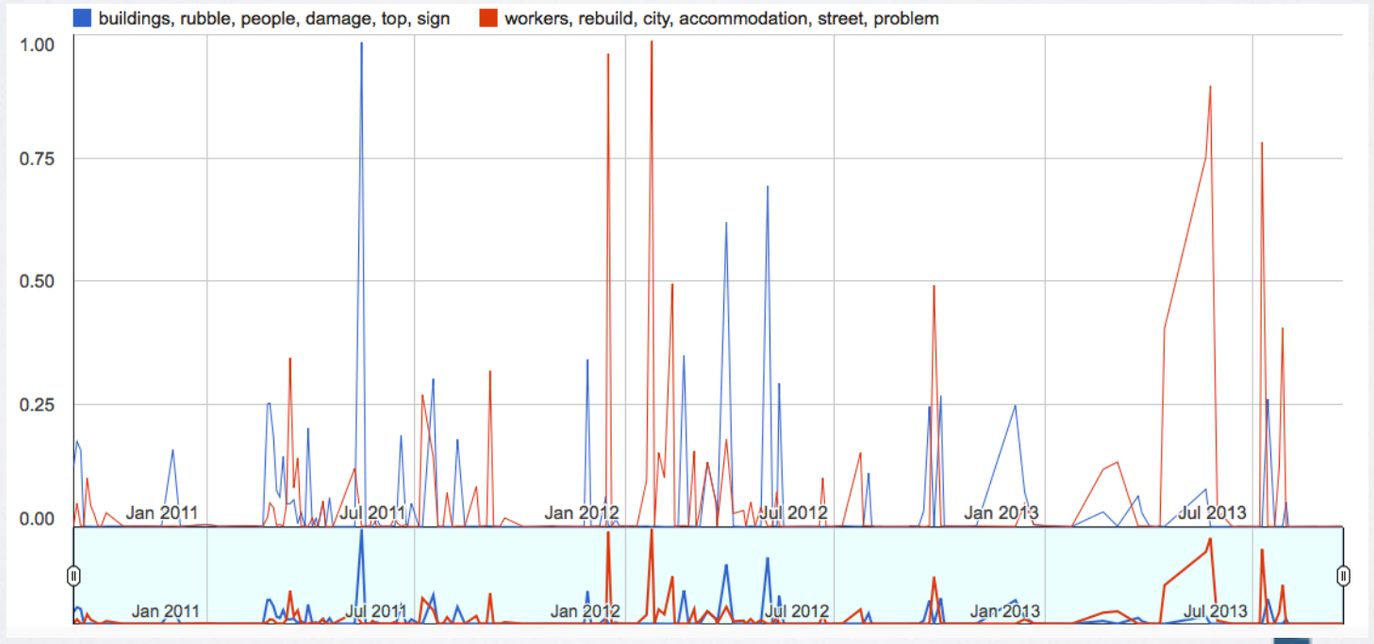
- Topic-Modelling is just one part of the overall research and is part of the quantitative analysis of the research data

- Quantitative analysis made it possible to reduce the sample from over 100,000 articles to around 2000 and reduce the number of cartoons from over 30,000 to under 1000.

- Combined with qualitative analysis (on the smaller sample)

SUBJECT HEADINGS VS TM

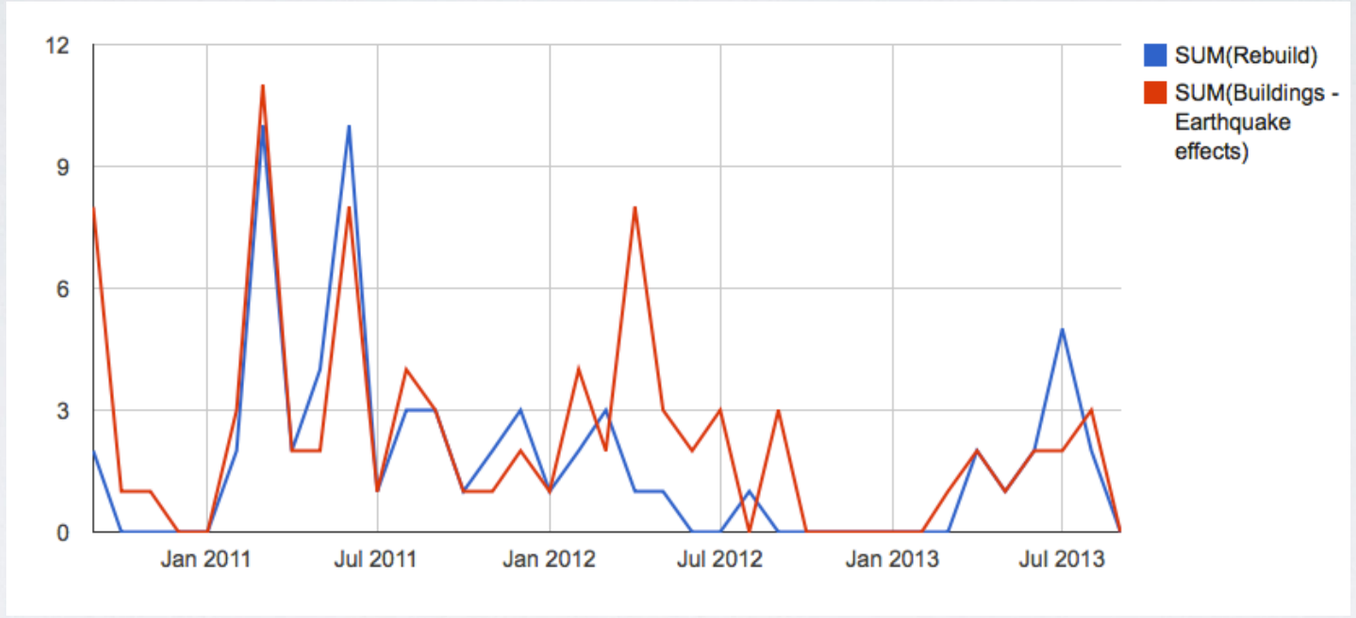


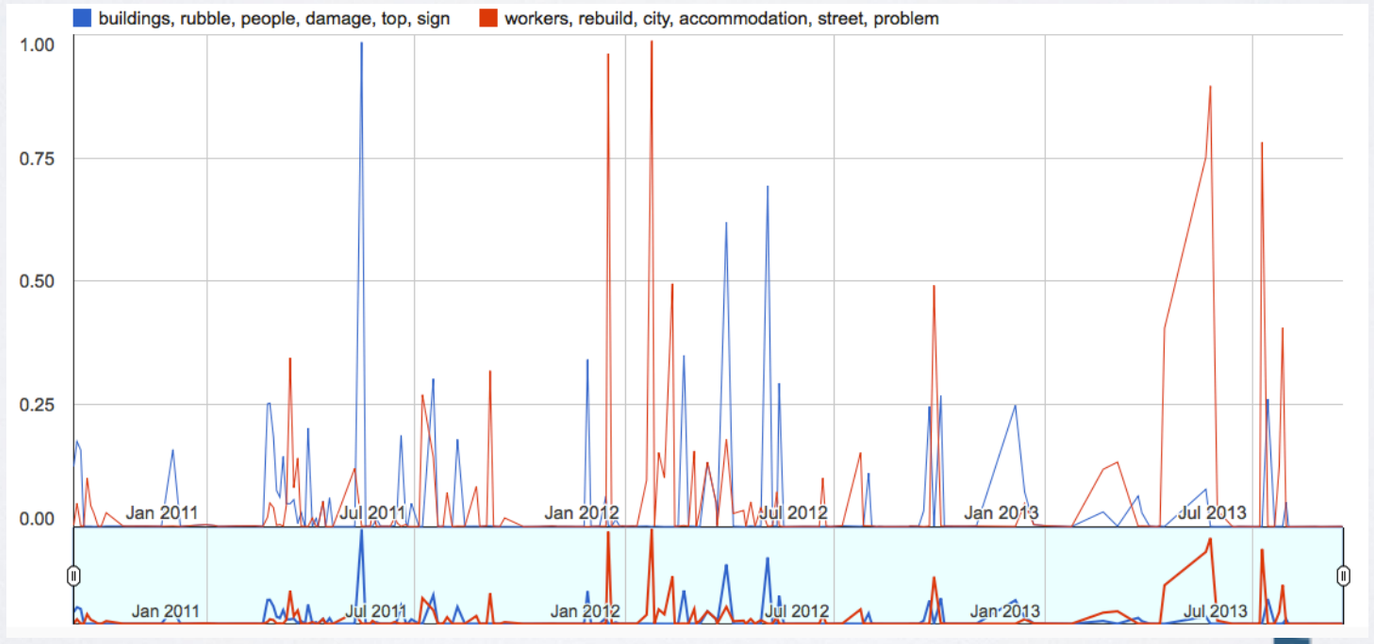


Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

SUBJECT HEADINGS VS TM

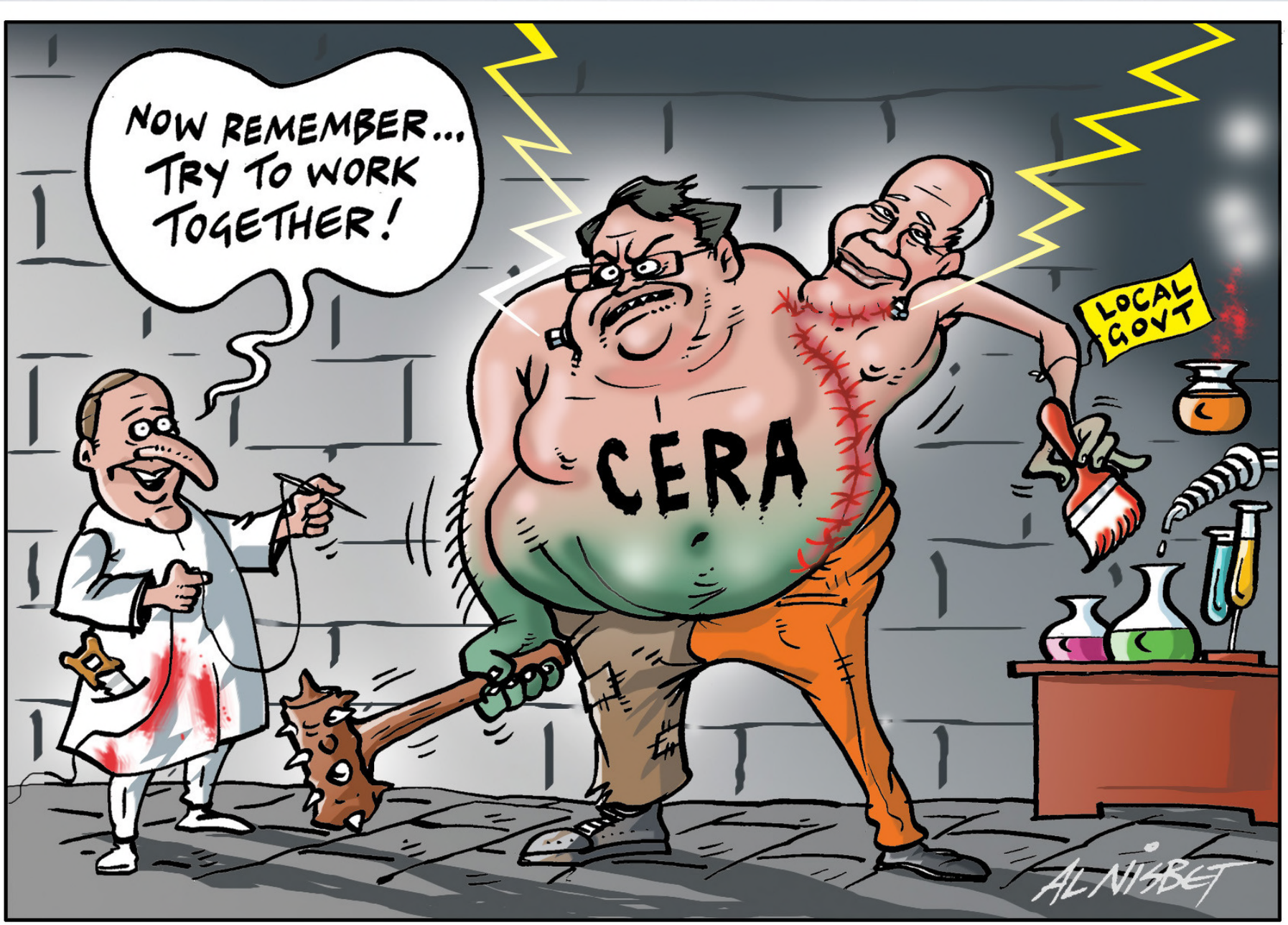




Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

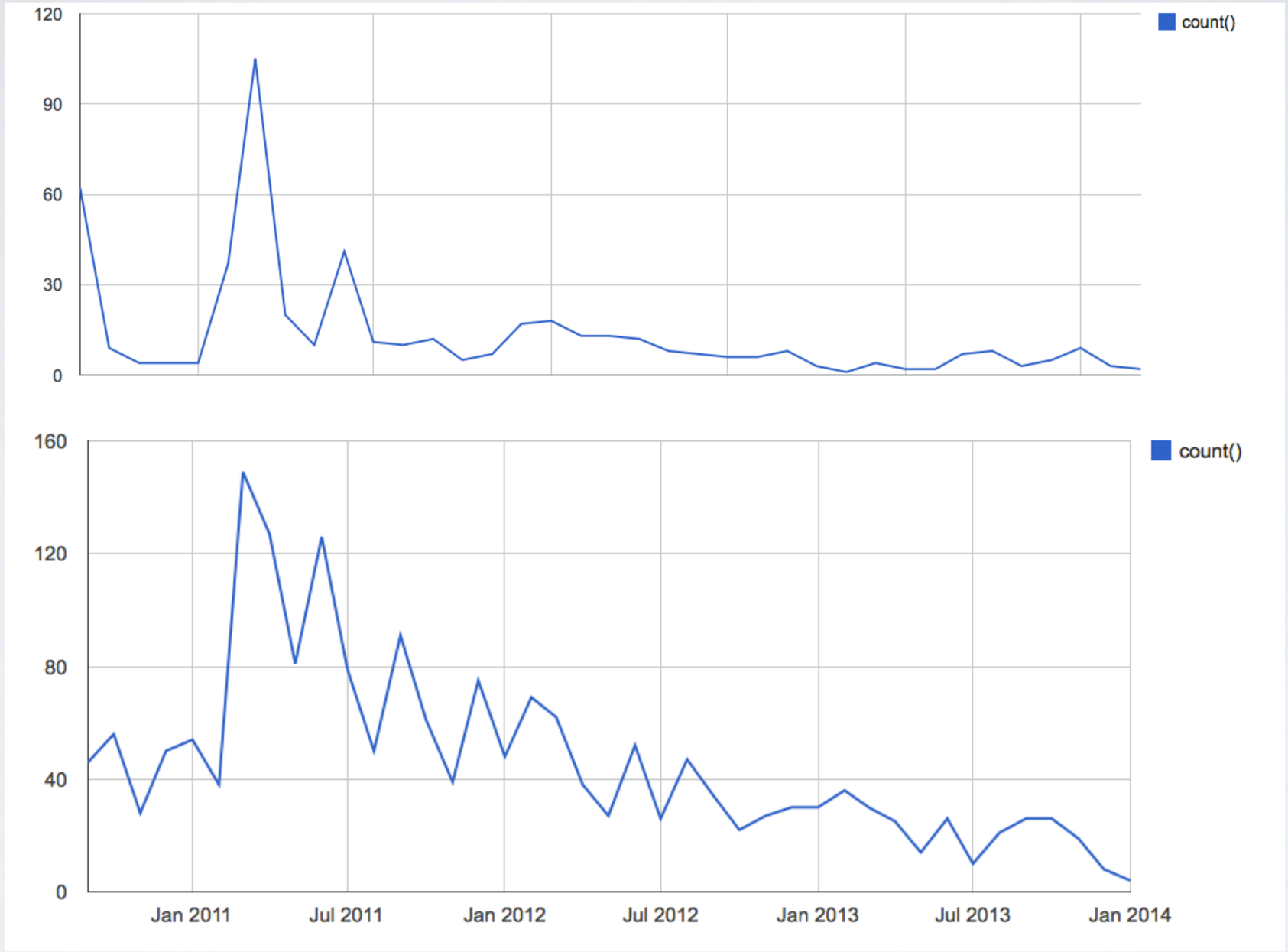
COMBINING DATA-SETS



Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

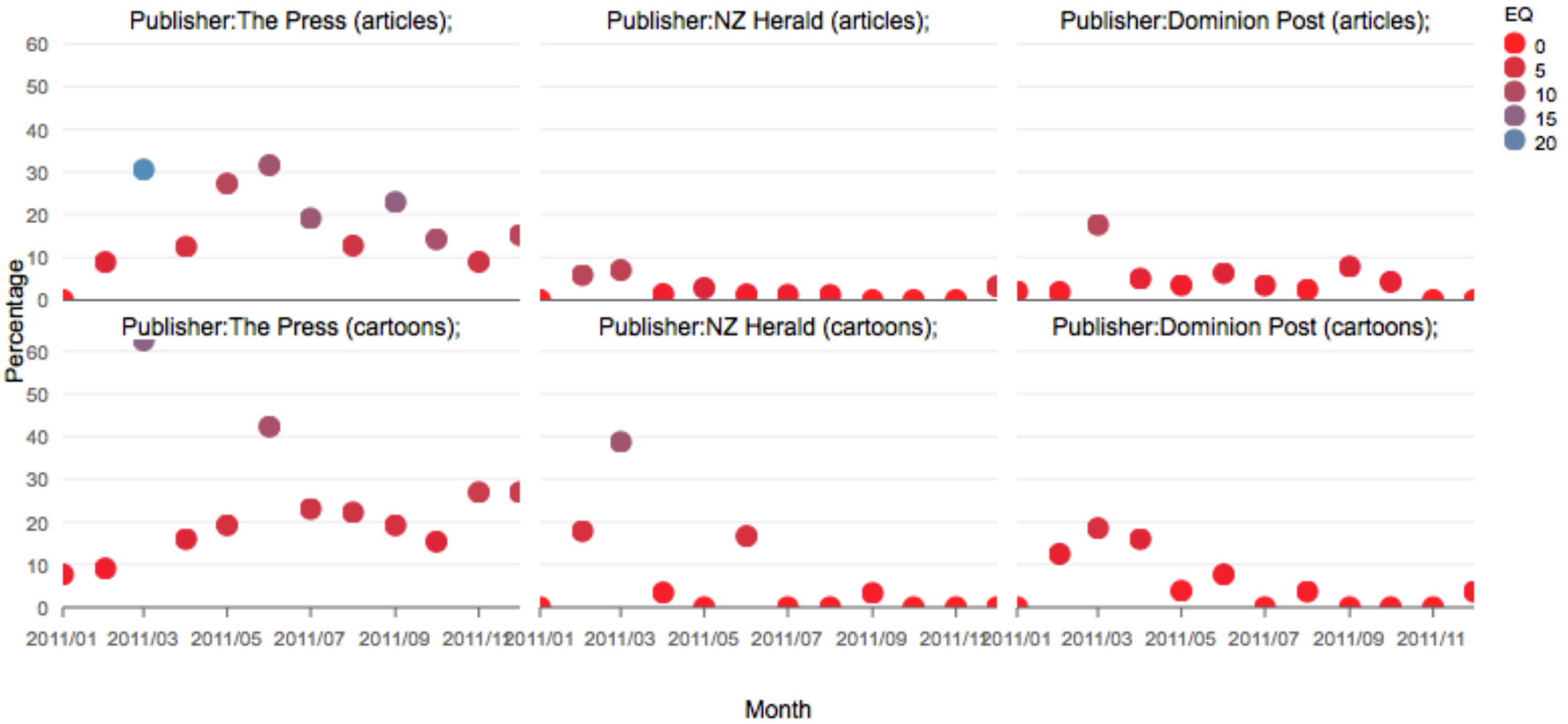
COMBINING DATA-SETS



Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

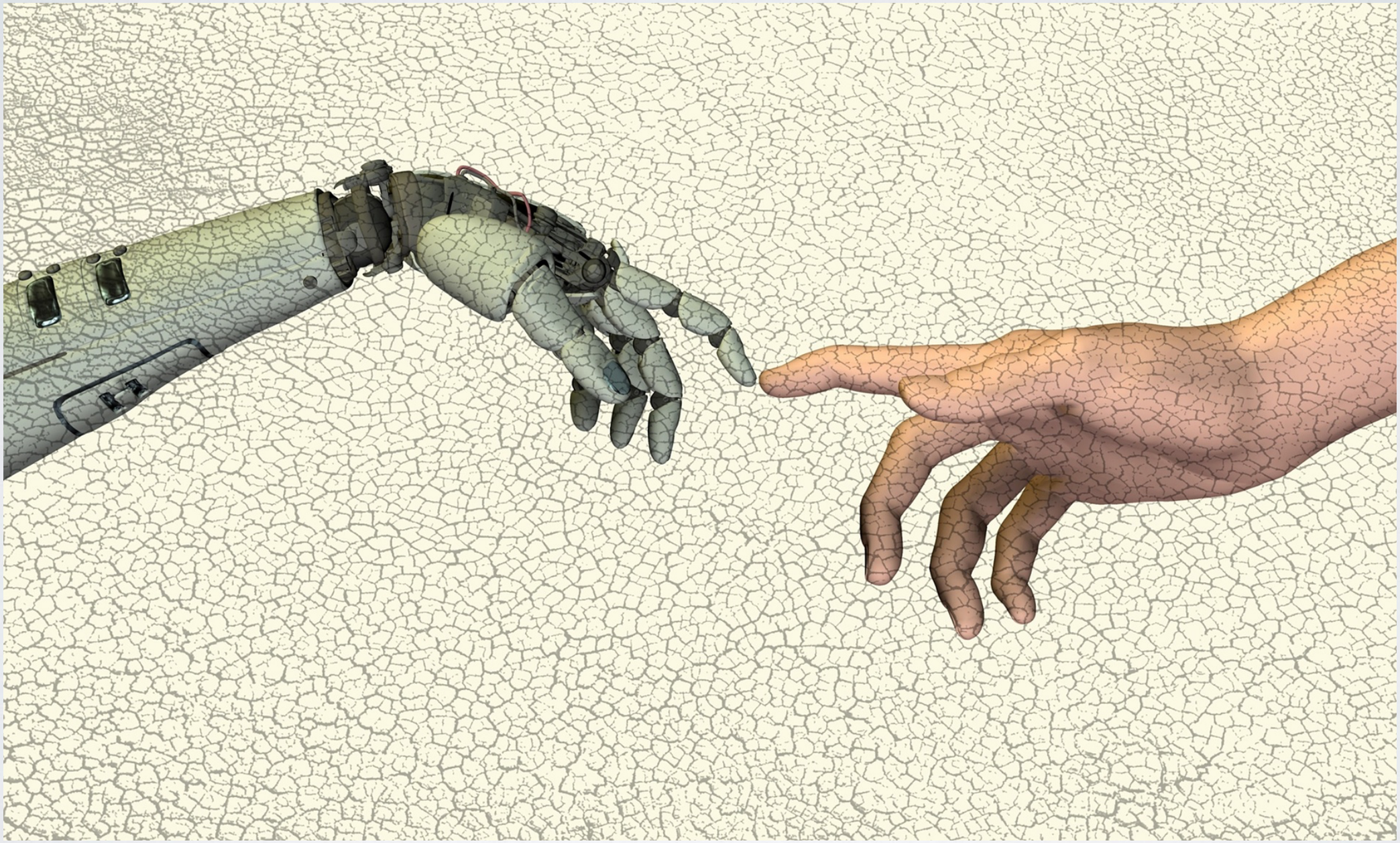
COMBINING DATA-SETS



Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG

QUESTIONS?



Thomas Köntges
Introduction to Digital Humanities
University of Leipzig

Digital Humanities
UNIVERSITÄT LEIPZIG