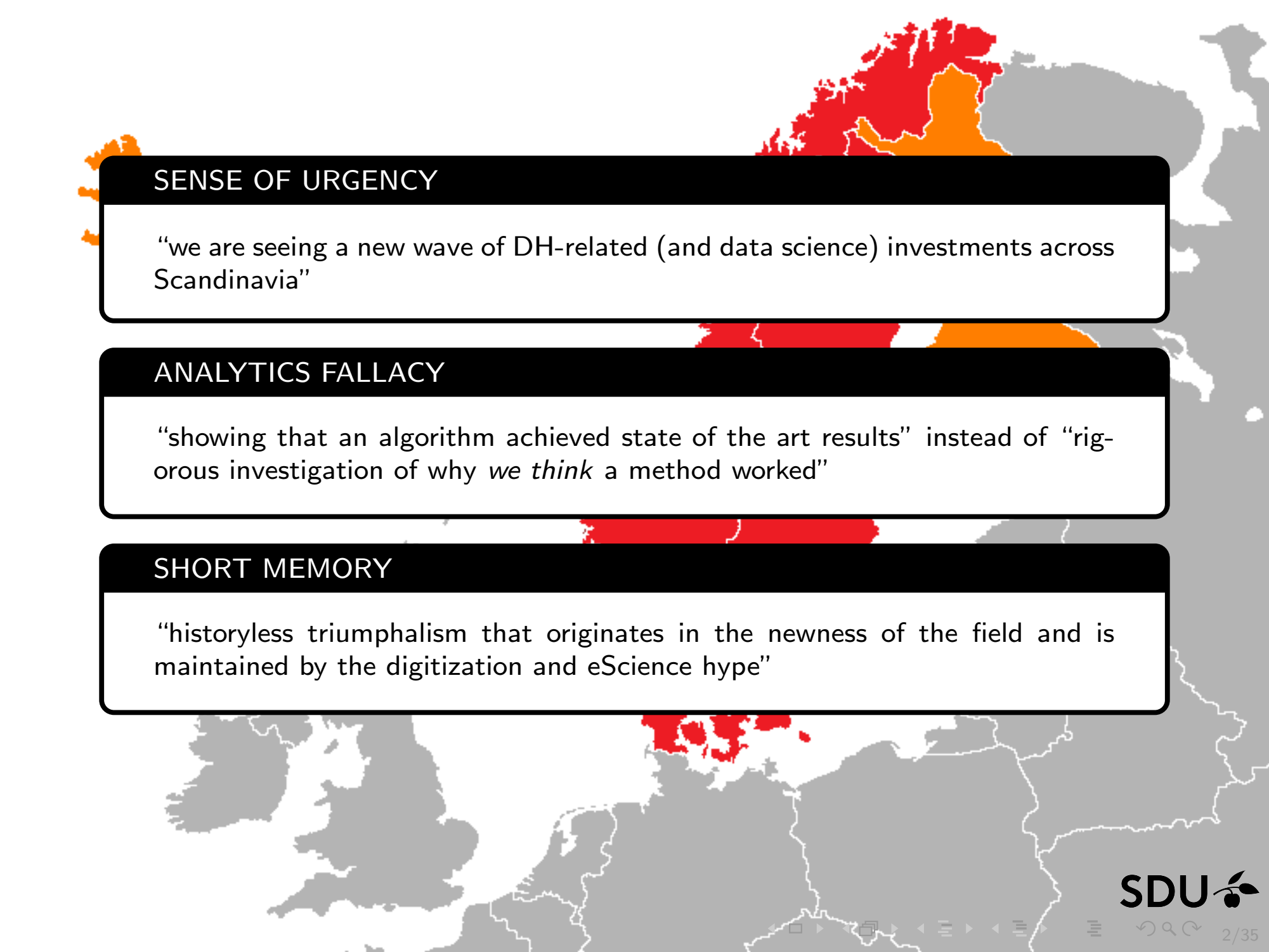
When Humanities Scale

On the Emergence of Analytics in Culture Research

Kristoffer L Nielbo
knielbo@sdu.dk
knielbo.github.io/

March 16, 2018



SENSE OF URGENCY

"we are seeing a new wave of DH-related (and data science) investments across Scandinavia"

ANALYTICS FALLACY

"showing that an algorithm achieved state of the art results" instead of "rigorous investigation of why *we think* a method worked"

SHORT MEMORY

"historyless triumphalism that originates in the newness of the field and is maintained by the digitization and eScience hype"

PROGRAM

0.00	**HUMANITIES & DATA**	a few data points are not enough
0.20	**CULTURE ANALYTICS**	an emerging field
0.30	**APPLICATIONS***	humanities data and computing
0.55	**SUMMARY**	...

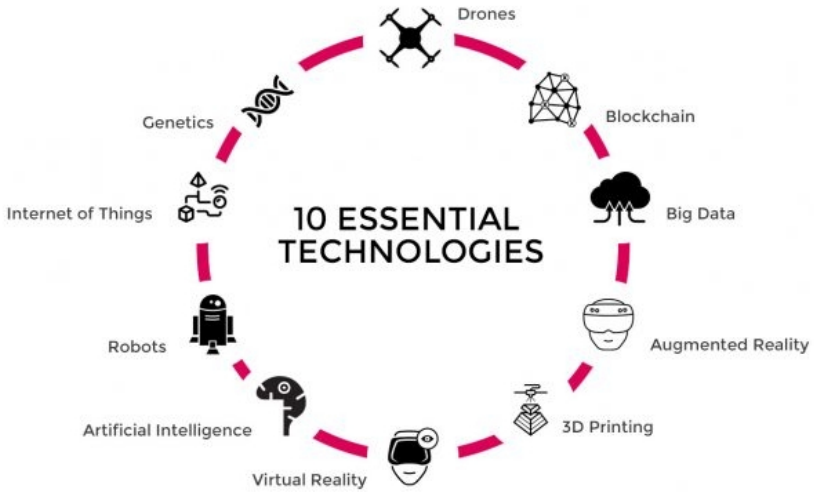
* interrupted by short `digressions`

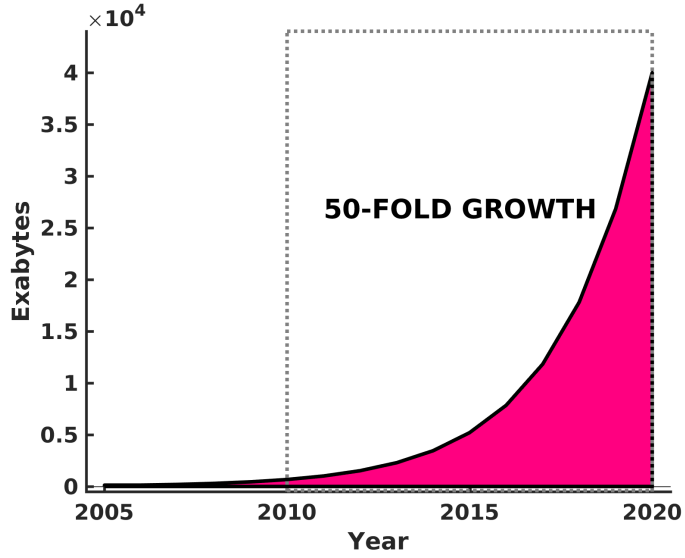
HUMANITIES & DATA

– domain knowledge in history, language, literature &c combined with microscopic and (predominantly) qualitative analysis of human cultural manifestations

anti-thesis to data-intensive research

– research that solely relies on very few data points, a "myopic" perspective and human computation



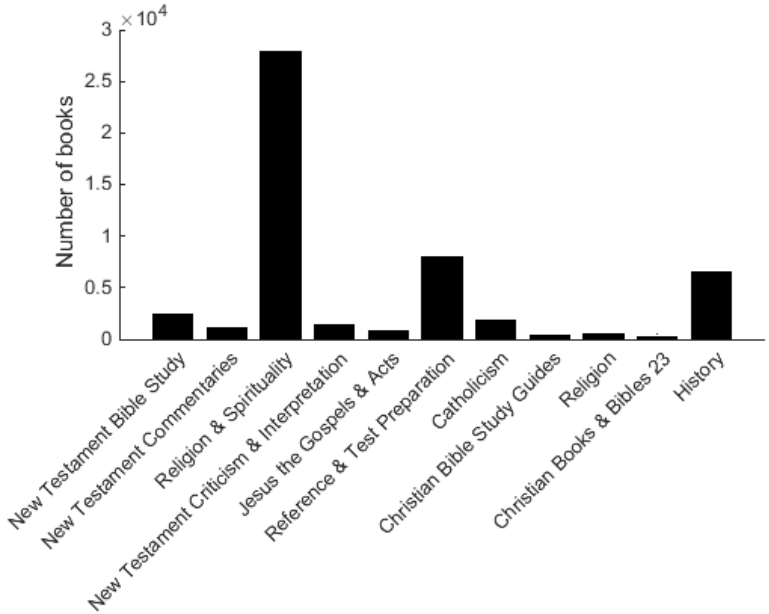


– the `data deluge` is transforming knowledge discovery and understanding in every domain of human inquiry

– knowledge discovery depends critically on advanced computing capabilities

a large part of these data are unstructured and fundamentally cultural

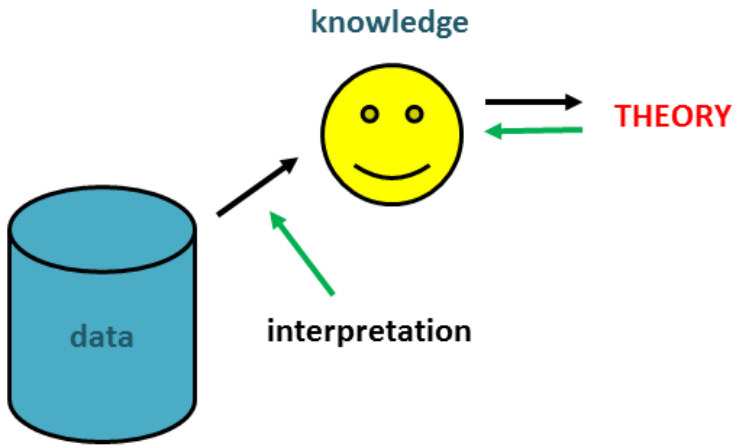
– to get additional value from these data, faculties of humanities must become computationally and data literate

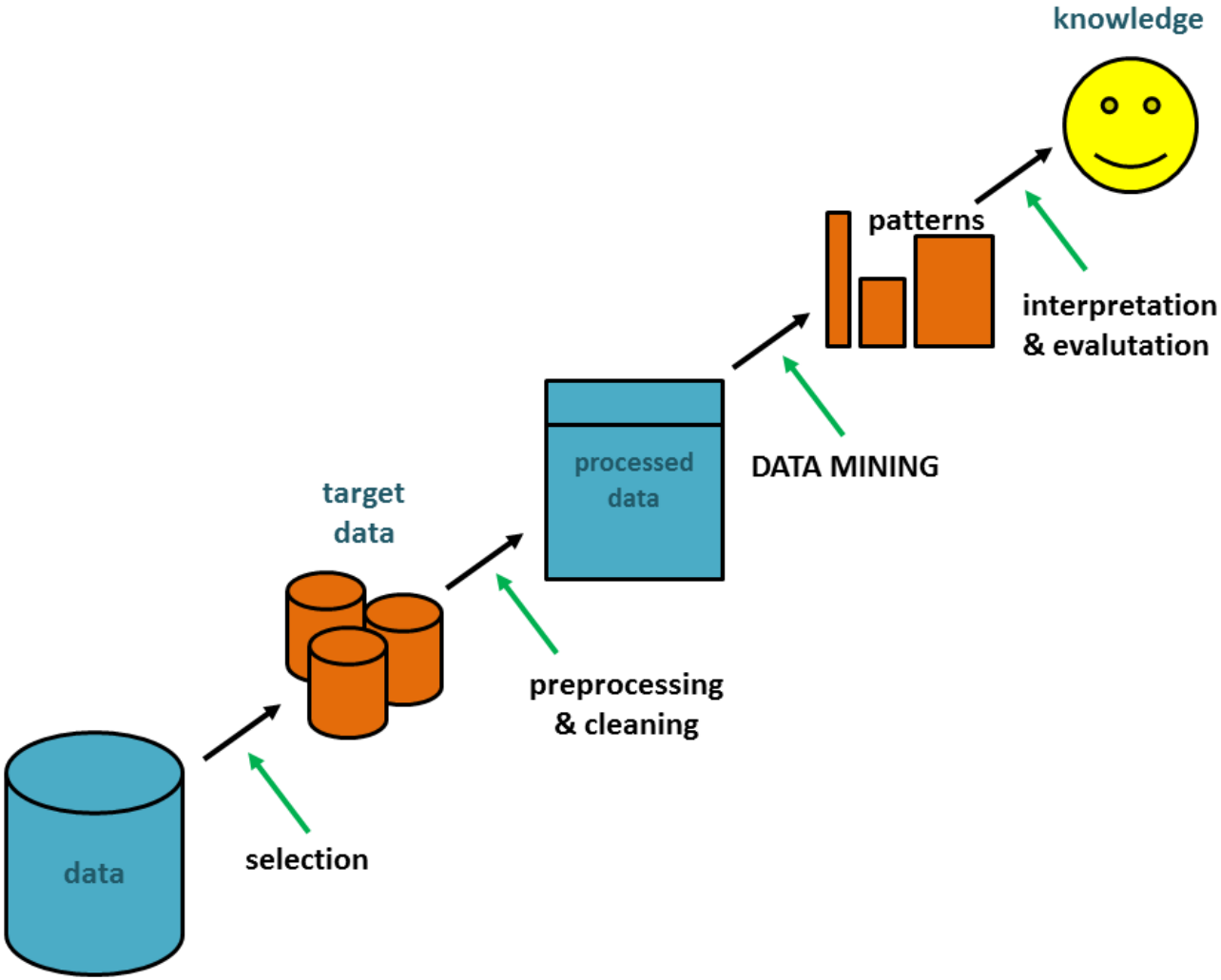


– number of research publications alone makes computational literacy a necessity for the humanities scholar

  – publications related to Gospel of Marc (KJV) $>$ 50K, $\sim$ 16,500 words in 16 chp. on 11 p.

– plus a massive increase in digitized cultural heritage databases (libraries, archieves, museums)


knowledge
THEORY
data
interpretation


knowledge
patterns
interpretation
& evalutation
processed
data
DATA MINING
target
data
preprocessing
& cleaning
data
selection

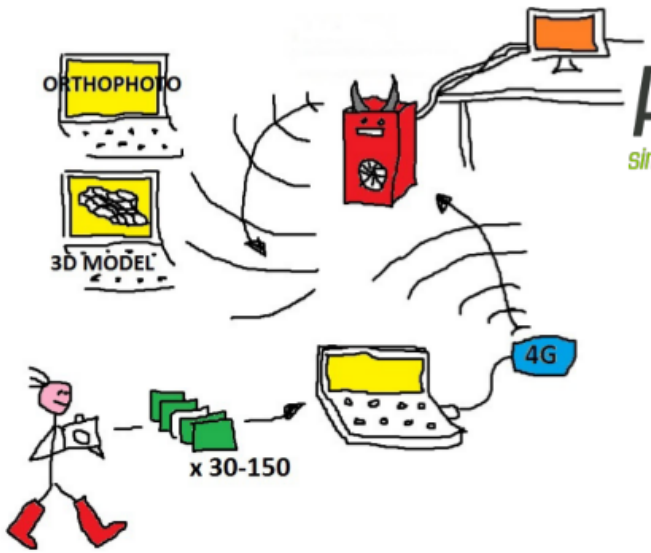
Big Data or (just) data

– Depending on definition, most humanities data are not *Big Data*
– they are however "big enough for us"

> "Instead of focusing on a 'big data revolution,' perhaps it is time we were focused on an 'all data revolution,' where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world."

(Lazer,Kennedy, King & Vespignani 2014)

Archaeology|3D modeling



PHOTOMODELER

PIX4D
simply powerful

sfm

AUTODESK 123D

PhotoScan
3D Modeling and Mapping

– humanistic domain experts (archaeologist) that use research technique (excavation)
– digital technologies have increased the scale and changed the research area

- archaeology and interaction studies currently use Big Data/HPC when the computational needs are present

  – scale alone does not necessarily change methods or perspective
  – reduce ++data points to a few by relying on our myopic perspective for analysis
  – we essentially lack a culture *of* analytics in the humanities

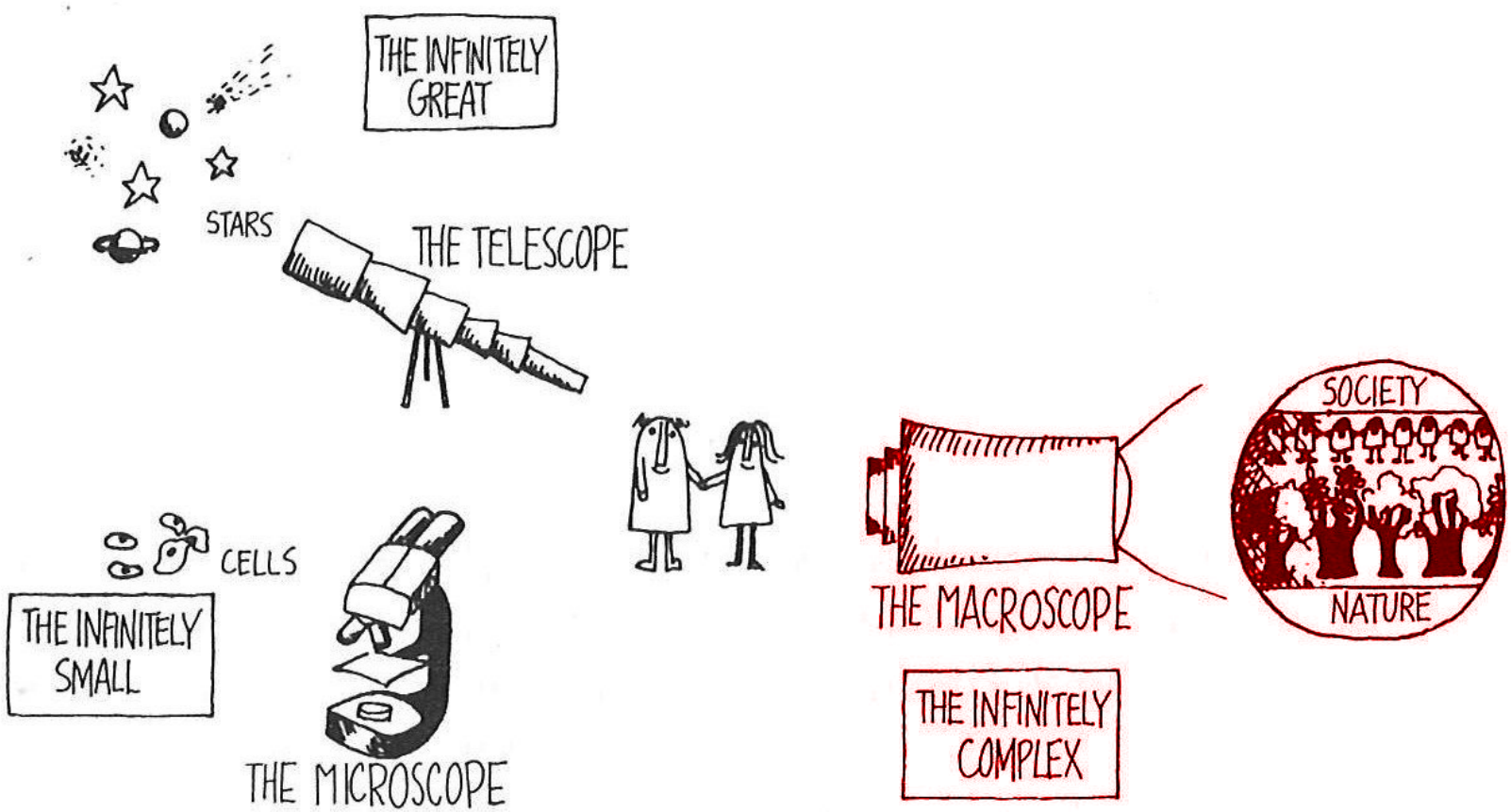
CULTURE ANALYTICS

In the humanities a culture of analytics is an analytics of culture

We have to study "*the dynamics of culturally informed interactions between people, and the cultural expressive forms that result from these interaction ... at scales hitherto unimaginable*"

So we need to develop an "*intellectually and ethically sound approach to the study of cultures across time and across space, leveraging the enormous gains made in the past decade in computation and machine readable cultural archives, from libraries and museum collections to the born digital cultural expressions of billions of people on the internet*"[1]

[1] From the Culture Analytics White Papers' *Introduction*



– Culture Analytics seeks to understand cultural phenomena as inherently multi-scale and multi-resolution

– preference for micro to macro-movement ("scale from one object")

CULTURE ANALYTICS

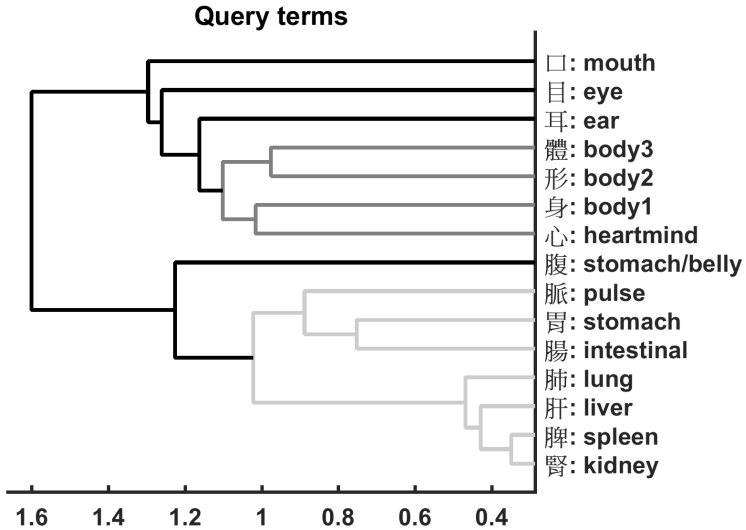
In comparison to analytics proper

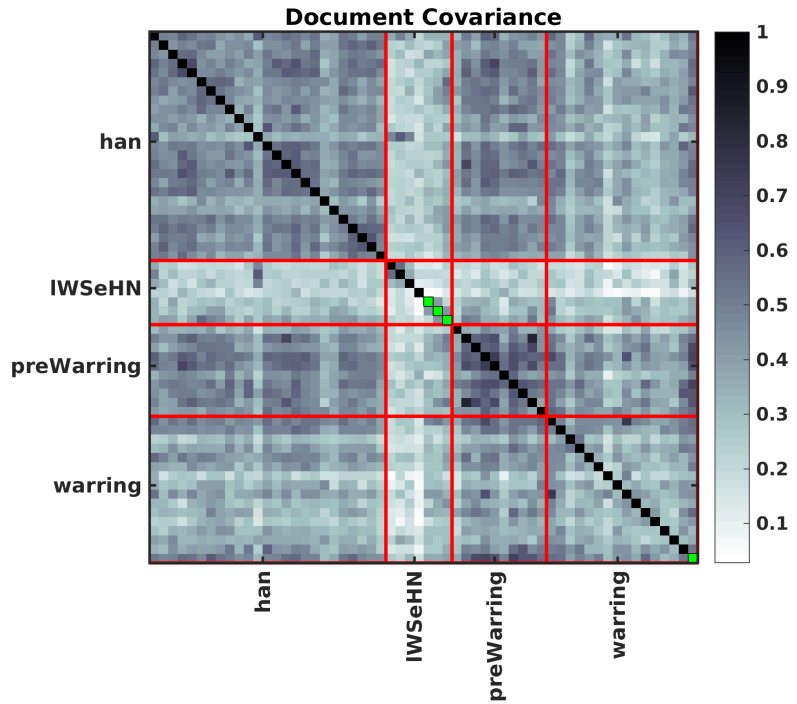
- descriptive not predictive
- neither side of the interdisciplinary divide is conceptualized as service
- preference for micro-scale analysis
- predominantly unstructured data
- low-resource varieties/historical perspective (cultural heritage data)
- reliance on qualitative assesment (e.g., hyper-parameters and validation procedures)

... to similar trends (e.g., culturomics, cliodynamics)

- multi-scale/multi-resolution
- data-intensive ethos (scalability matters)

APPLICATIONS



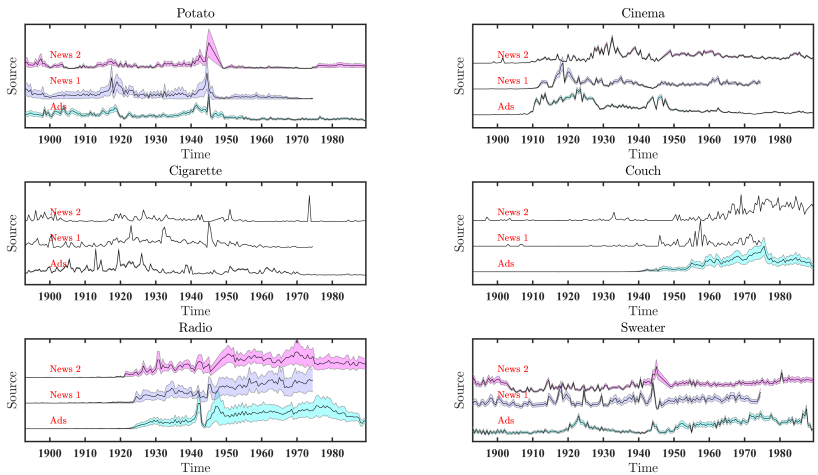


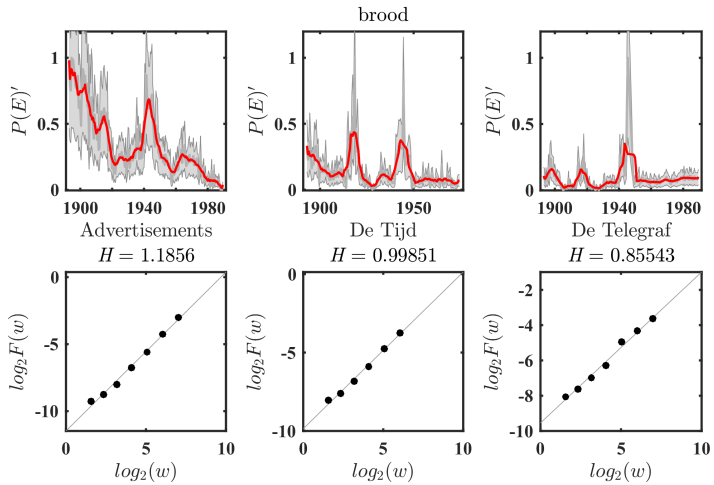
– philosphers and sinologists have been debating the existence of mind-body dualism in classical Chinese philosophy

– with domain experts, unsupervised learning was used to identify a multi-level dualistic semantic space

– one model (LDA) was further utilized to predict class of origin for controversial texts slices

History|Predictive Causality & Slow Decay





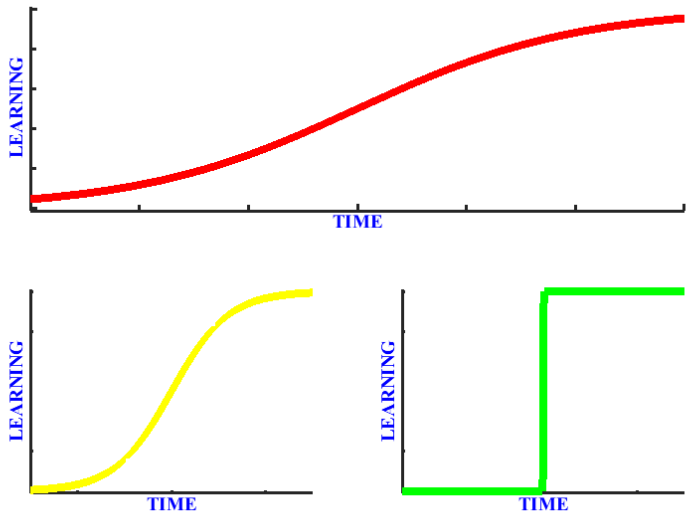
– historians and media researchers theorize about the causal dependencies between public discourse and advertisement

– time series analysis of keyword frequencies (from seedlists) indicated that for some categories ‘ads shape society’, while other categories merely ‘reflect’

– advertisements show a faster decay (on-off intermittant behavior) than public discourse (long-range dependencies)

digression #1.1

Computational Literacy|Programming & Stats



software carpentry

DATA CARPENTRY

– every knowledge intensive organization has to break the learning curve, but certain sectors are more challenged

– we out-sourced the task to an international non-profit organization w. years of experience in scientific computing

– promote a common language and import best practice from software development

  – unix shell, python and version control

`digression #1.2`

Computational Literacy|Programming & Stats

GUI → CLI

- novice-friendly visual approach to computer interaction w. a fast learning curve ERROR
- expert-friendly text-based approach to computer interaction w. ++freedom VALID
- CONFLICT break the learning curve through training intensive, non-intuitive, and specialized tools

- locally, we try to solve this conflict with a mix of science and guerrilla warfare by establishing small, semi-autonomous eScience units that intervene in humanities research

---

Medieval History|Novelty Detection

– historians debate historical transitions

– Saxo's *Gesta Danorum* c. 1200 AD. history of the Danish royal dynasty

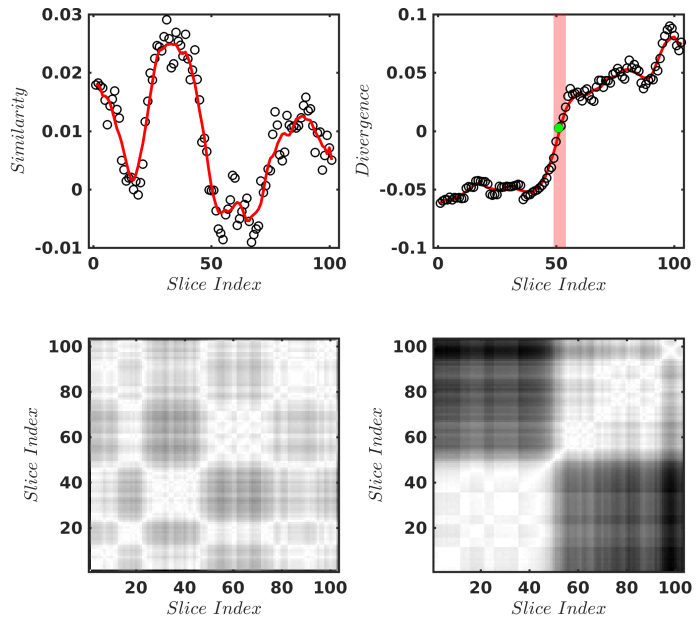
– transition between book 8 or 9?

– transition point or gradual?

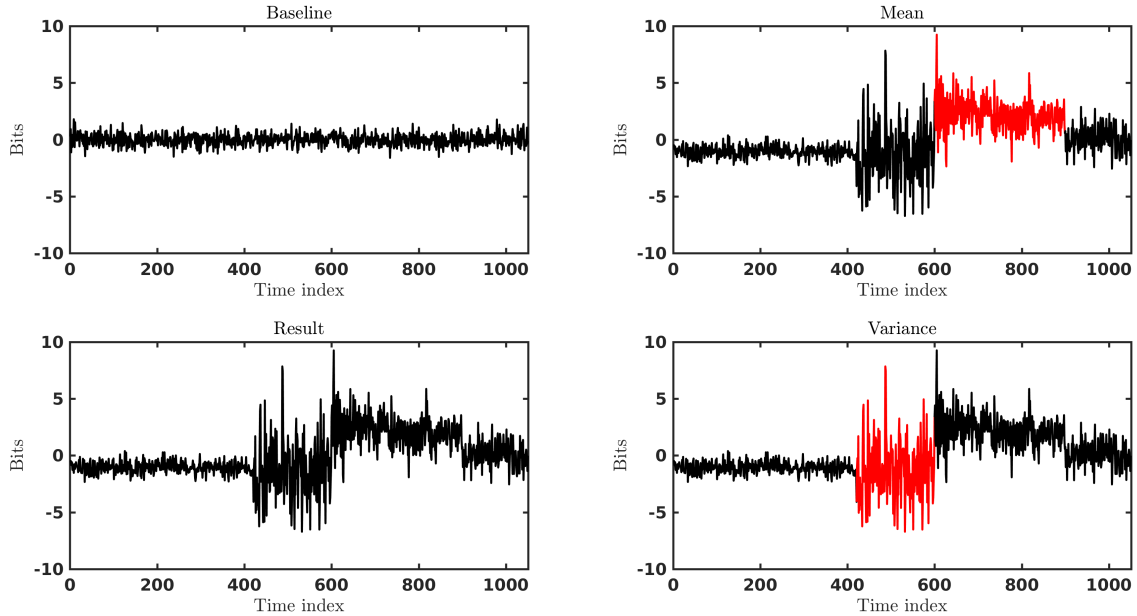
– traditional word-level representation is ambivalent

– latent semantic model was trained over sentence windows

– change detection and recurrence plot used to identify phase transition focusd in book 9



Media Studies|Novelty Detection



– change point detection in topicality space applies to "a change in the media tone"

– train model on 200 years of newspapers in a comparative study between DK and NL

– collaboration between historians, media studies and information science with a predictive scope

`digression #2`

Copyright & Privacy|Data Access and Mobility

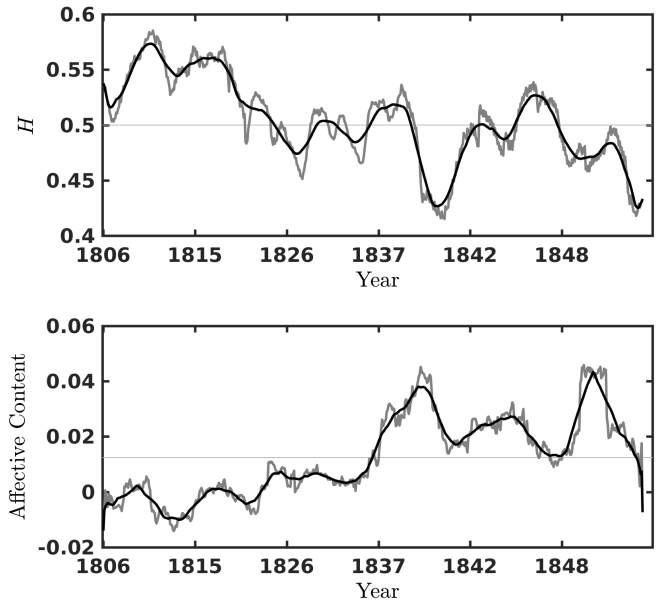
Challenges to computationally empowering humanities:

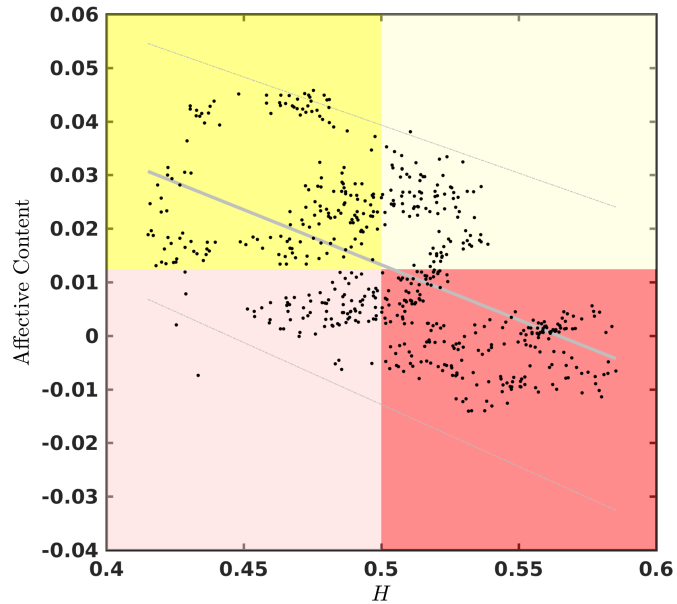
- technical competencies
- interdisciplinary respect and understanding
- epistemology differences
- data access and mobility

Data silos (the true punishment for the fall of man) often originate in "cultural differences", not technical or legislative issues

copyright is a bigger challenge than data protection laws

Literary History|Lexical Density





Literary scholars and creativity researchers argue for the "tortured artist"

- "writers' creative state is inversely related to their emotional state"
- "writers' creative state *depends* on their emotional state"

– look for dependencies in lexical density and sentiment scores for highly profilic writers to identify state incongruences

`digression #3`

Historical Languages|Low-resource Varieties

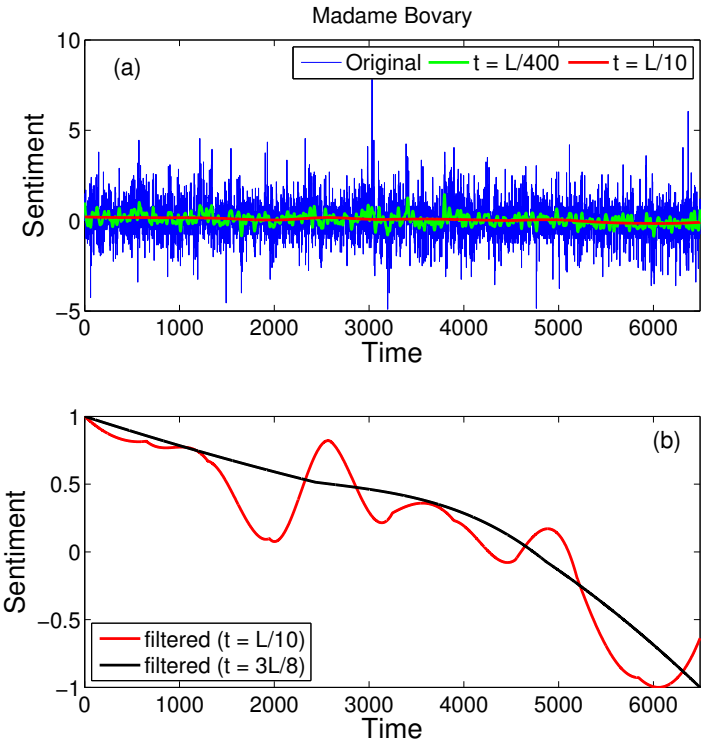
– text analytics depends critically on existing tools and data (ex. sentiment dictionaries)

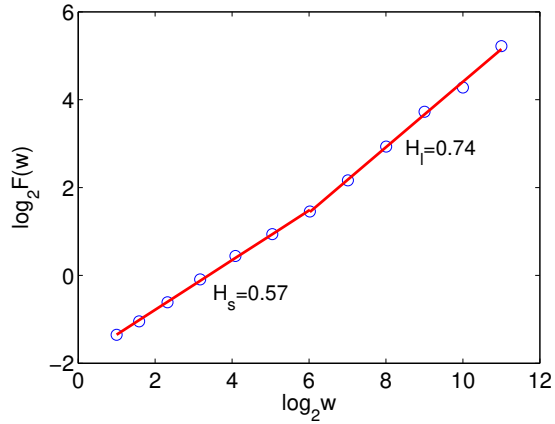
– orthographic variation in historical data represents a challenge, because NLP and TM resources "suffer from presentism"

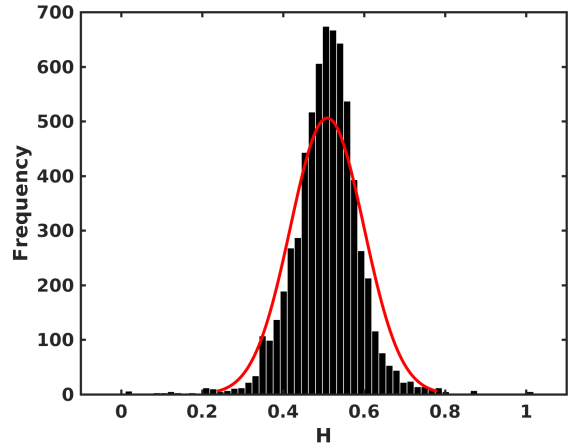
– projects often try to adapt the tool (ex. modify dictionary to historical data set)

– this solution scales badly due to lack of standardization

For Scandinavian languages we use spelling correction (rule-based and probabilistic) to normalize (or modernize) historical data increasing recall considerably





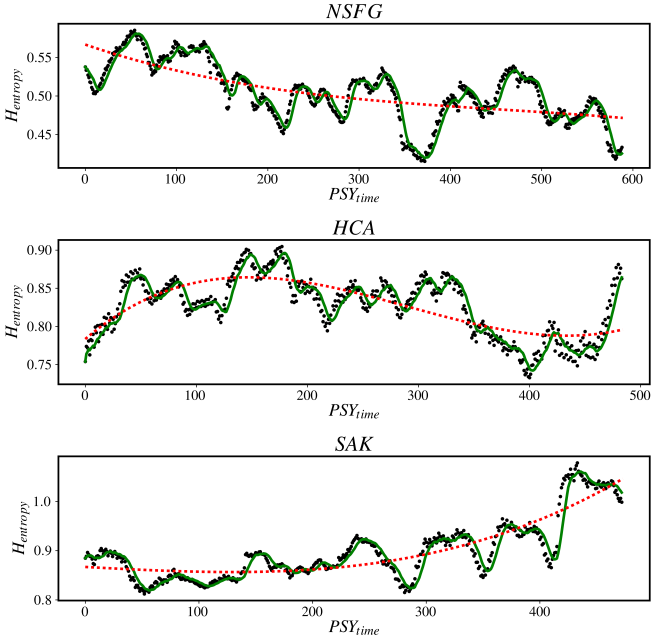


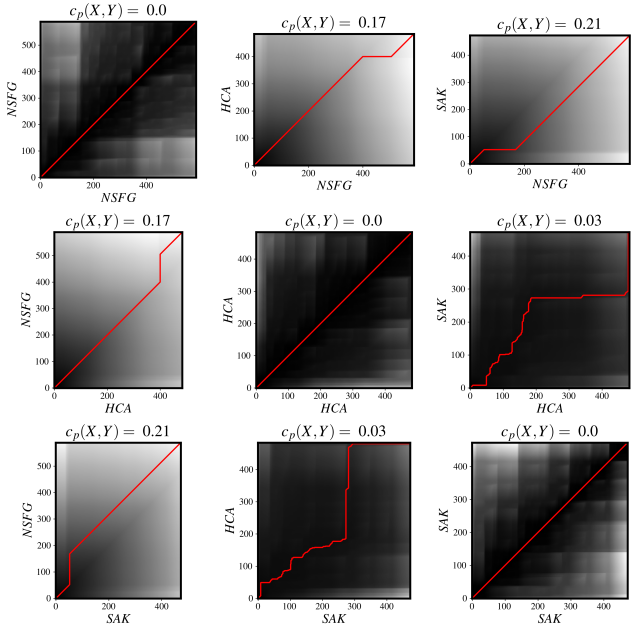
– dictionary-based sentiment analysis can reconstruct narrative/plot vectors that reflect human reading

– basic insights from structural linguistics and narratology can be captured by this approach

– a particular scaling-range, $0.6 < H \leq 0.8$, seems to indicate literary optimality

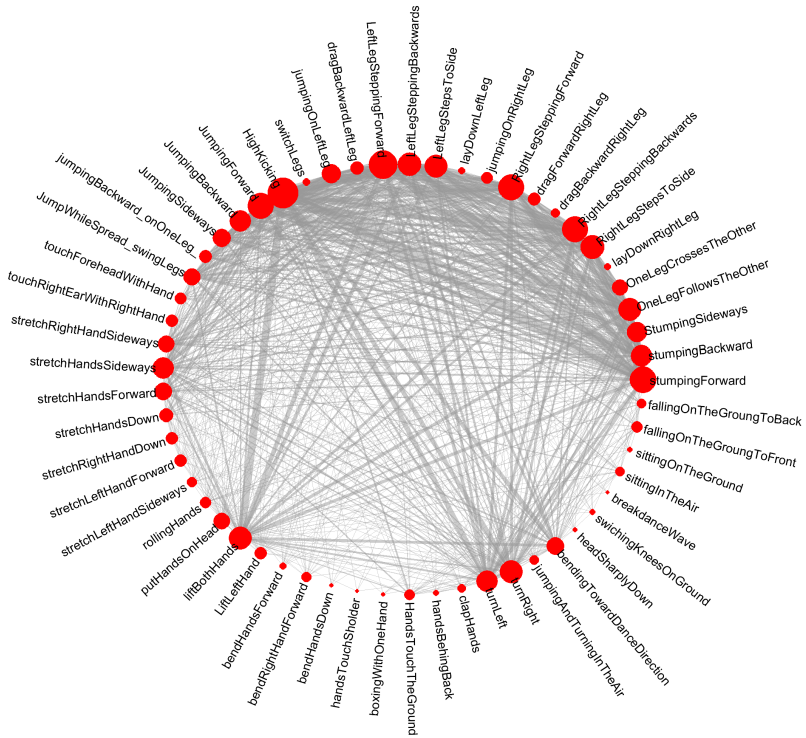
Literary History|Sequence Alignment

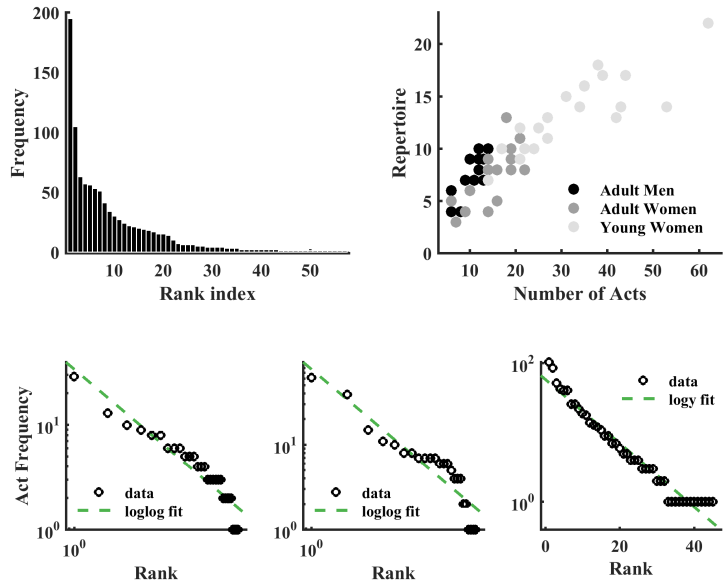




– there is a new biographical trend is In literary history

– using lexical density and sequence alignment, we can compare creative trajectories of authors





– anthropologists discuss why rituals appear rigid, while they seem to maintain behavioral variability

– manual annotation of ritual dance applied to ethnographic video archieves from multiple generations

– very few behavioral units are transmitted between generations (compulsory), allowing for both flexibility and rigidity

SUMMARY

Summary

All knowledge-intensive organization are experiencing the data deluge

- demands new forms of expertise and (strange) bed-fellows
- unique situation where compute and data can *empower humanities domain experts* and change our scale and perspective
- humanities are part of the solution

BUT,

- scaling (Big Data) alone is not enough (e.g., archaeology, interaction studies)
- we need a culture of analytics

CULTURE ANALYTICS

- cultural behavior and products at scale
- descriptive, transdisciplinary, historical, qualitative
- challenged by lack of training, data access, low-resource varieties

THANK YOU

knielbo@sdu.dk
knielbo.github.io

& credits to

Max R. Echardt and Katrine F. Baunvig, datakube, University of Southern Denmark, DK
Jianbo Gao and Bin Liu, Institute of Complexity Science and Big Data, Guangxi University, CHN
Culture Analytics @ Institute of Pure and Applied Mathematics, UCLA, US