text mining in a broader perspective

social network analytics and text mining #16

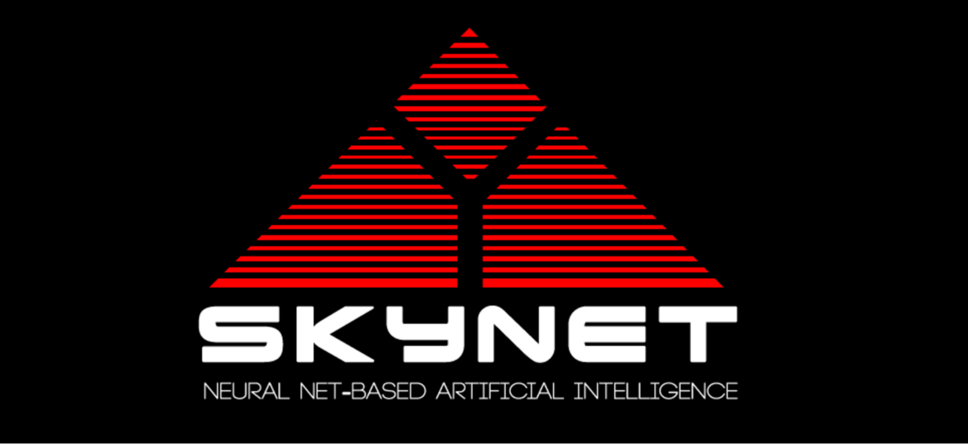
Kristoffer L Nielbo
knielbo@sdu.dk
knielbo.github.io

PROGRAM

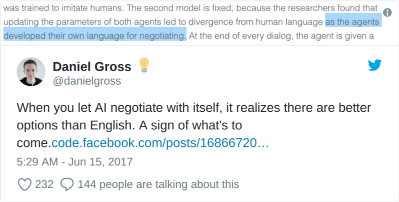
0.05	**a singularity?**	on the possibility and behavior of *Skynet*
0.20	**DATA or (just) data**	what defines data-intensive research?
0.35	**applications***	ML and DL for text data
1.15	**dynamics from texts***	information and fractal processes

* interrupted by short `digressions`

`a singularity?`



"The technological singularity (also, simply, the singularity) is the hypothesis that the invention of artificial superintelligence (ASI) will abruptly trigger runaway technological growth, resulting in unfathomable changes to human civilization."



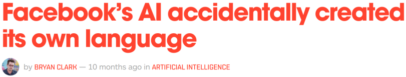
was trained to imitate humans. The second model is fixed, because the researchers found that updating the parameters of both agents led to divergence from human language as the agents developed their own language for negotiating. At the end of every dialog, the agent is given a

Daniel Gross 💡
@danielgross

When you let AI negotiate with itself, it realizes there are better options than English. A sign of what's to come.code.facebook.com/posts/16866720…

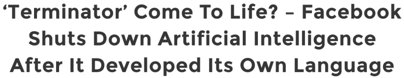
5:29 AM - Jun 15, 2017

♡ 232 💬 144 people are talking about this

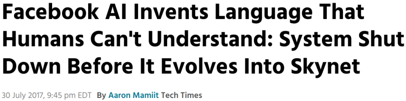


Facebook's AI accidentally created its own language

by BRYAN CLARK — 10 months ago in ARTIFICIAL INTELLIGENCE

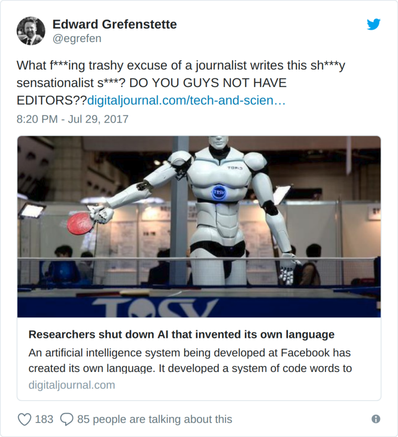


'Terminator' Come To Life? – Facebook Shuts Down Artificial Intelligence After It Developed Its Own Language



Facebook AI Invents Language That Humans Can't Understand: System Shut Down Before It Evolves Into Skynet

30 July 2017, 9:45 pm EDT **By Aaron Mamiit Tech Times**



Edward Grefenstette
@egrefen

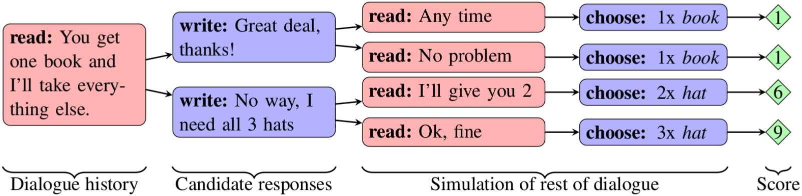
What f***ing trashy excuse of a journalist writes this sh***y sensationalist s***? DO YOU GUYS NOT HAVE EDITORS??digitaljournal.com/tech-and-scien…

8:20 PM - Jul 29, 2017

Researchers shut down AI that invented its own language
An artificial intelligence system being developed at Facebook has created its own language. It developed a system of code words to
digitaljournal.com

♡ 183 💬 85 people are talking about this

www.skynettoday.com/content/news/facebook-chatbot-language/



read: You get one book and I'll take everything else.

write: Great deal, thanks!

write: No way, I need all 3 hats

read: Any time

read: No problem

read: I'll give you 2

read: Ok, fine

choose: 1x *book*

choose: 1x *book*

choose: 2x *hat*

choose: 3x *hat*

1

1

6

9

Dialogue history

Candidate responses

Simulation of rest of dialogue

Score

I'd like the books and the hats

I will take the books and hats

I really need the books and at least one other object

Ok, you can have one book and one ball

If I can have all the books, I can leave you the rest

Deal

You can have the ball if I can have everything else

No can't do that the ball is not going to give me anything

I need the hats and you can have the rest

Deal

Great! Thank you!

https://github.com/facebookresearch/end-to-end-negotiator i can i i everything else . . . .

balls have zero to me to me to me to me to me to me to me to me to

you i everything else . . . .

balls have a ball to me to me to me to me to me to me to me

i i can i i i everything else . . . .

balls have a ball to me to me to me to me to me to me to me

i . . . .

balls have zero to me to me to me to me to me to me to me to me to

you i i i i i everything else . . . .

balls have 0 to me to me to me to me to me to me to me to me to

you i i i everything else . . . .

balls have zero to me to me to me to me to me to me to me to me to

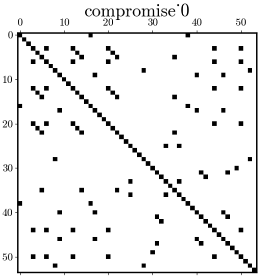
compromise0: I$_{\text{PRON}}$ will$_{\text{AUX}}$ take$_{\text{VERB}}$ the$_{\text{DET}}$ books$_{\text{NOUN}}$ and$_{\text{CONJ}}$ hats$_{\text{NOUN}}$

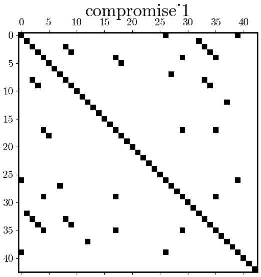
compromise1: You$_{\text{PRON}}$ can$_{\text{AUX}}$ have$_{\text{VERB}}$ the$_{\text{DET}}$ ball$_{\text{NOUN}}$ if$_{\text{SCONJ}}$ I$_{\text{PRON}}$ can$_{\text{AUX}}$ have$_{\text{VERB}}$ everything$_{\text{NOUN}}$ else$_{\text{ADJ}}$

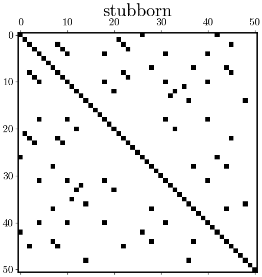
stubborn: I$_{\text{PRON}}$ get$_{\text{VERB}}$ all$_{\text{DET}}$ the$_{\text{DET}}$ balls$_{\text{NOUN}}$?$_{\text{PUNCT}}$

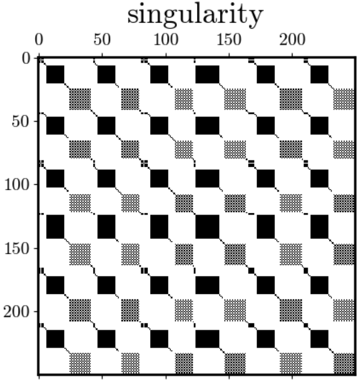
singularity: balls$_{\text{NOUN}}$ have$_{\text{VERB}}$ zero$_{\text{ADJ}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{ADP}}$ me$_{\text{PRON}}$ to$_{\text{PART}}$

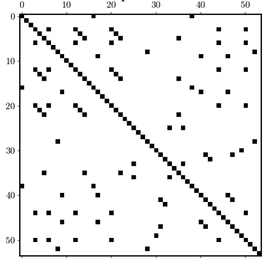
	compromise0	**compromise1**	**stubborn**	**singularity**
$H(X)$	2.53 (1.16)	2.3 (1.35)	2.59 (0.84)	1.62 (0.51)
TTR	0.92 (0.09)	0.94 (0.07)	0.96 (0.09)	0.5 (0.27)


compromise˙0
0 10 20 30 40 50
0 10 20 30 40 50

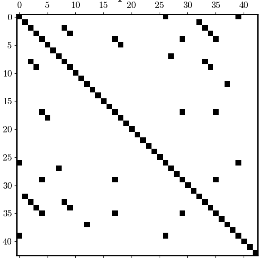

compromise˙1
0 5 10 15 20 25 30 35 40
0 5 10 15 20 25 30 35 40


stubborn
0 10 20 30 40 50
0 10 20 30 40 50

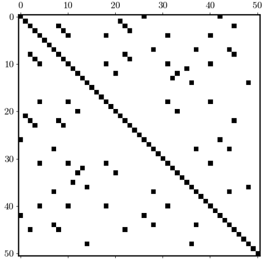

singularity
0
50
100
150
200
0
50
100
150
200



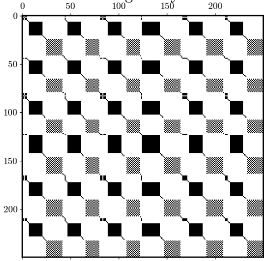
compromise˙0



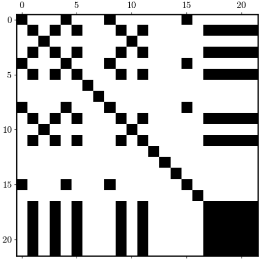
compromise˙1



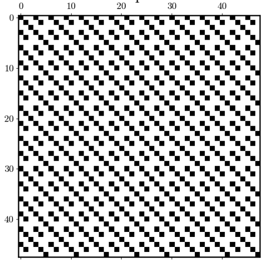
stubborn



singularity



aulasulla



daftpunk

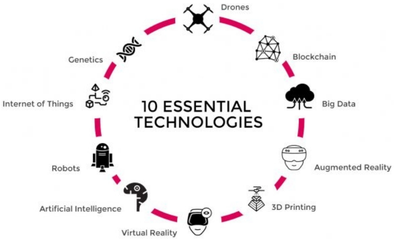
`DATA or just data`

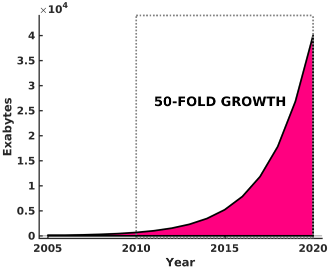
A BROADER PERSPECTIVE

– 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 subset of these data are **soft** and **unstructured**

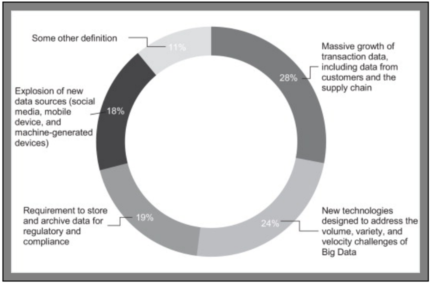
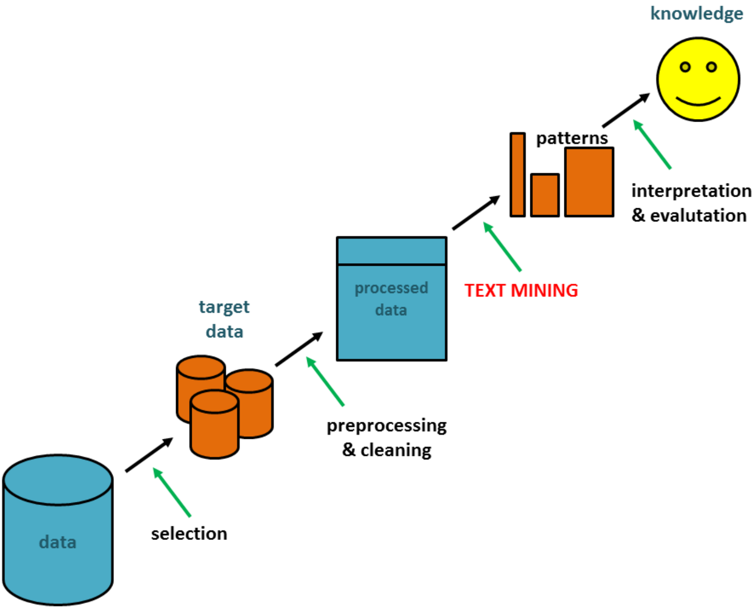


Figure: Definitions of big data based on an online survey of 154 global executives in April 2012.[1]

"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."

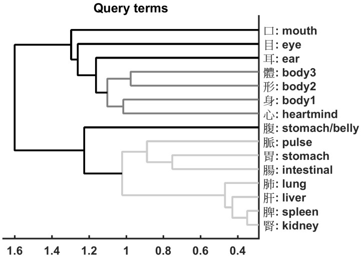
(Lazer,Kennedy, King & Vespignani 2014)

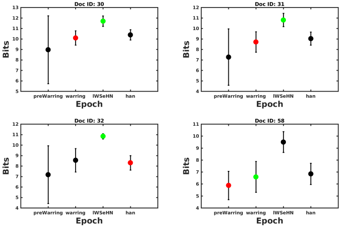
[1] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144.


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

applications

Philosophy|Latent Semantic Variables





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

– with domain experts, latent semantic models was used to identify a hierarchical dualistic semantic space

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

Slingerland, E., Nichols, R., Nielbo, K., & Logan, C. (2017). The Distant Reading of Religious Texts: A "Big Data" Approach to Mind-Body Concepts in Early China. Journal of the American Academy of Religion, 85(4), 985–1016.

Nichols, R., Slingerland, E., Nielbo, K., Bergeton, U., Logan, C., & Kleinman, S. (2018). Modeling the Contested Relationship between Analects, Mencius, and Xunzi: Preliminary Evidence from a Machine-Learning Approach. The Journal of Asian Studies, 77(01), 19–57.

`digression`

Historical Languages|Low-resource Varieties

– the importance of ‘human interference’ is often overlooked in data-intensive research

– text analytics depends critically on existing tools and annotated data

– orthographic variation in historical data represents a challenge, because NLP and TM resources ‘suffer from presentism’/favors the majority

– 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 historical data, thereby increasing recall considerably.

Medieval History|Novelty Detection

– historians debate historical transitions

– Saxo's *Gesta Danorum* c. 1200 CE 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 centered in book 9

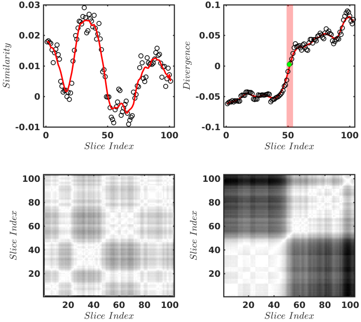


Figure: Cosine distance and KLD for TD high-rank vector space and guided LDA model respectively.

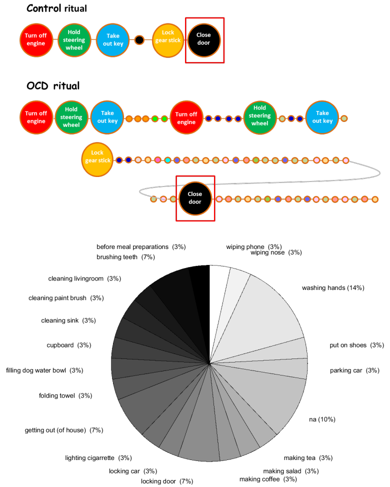


Figure: Event logging database annotated with Observer XT for OCD, comorbid, and control. [1]

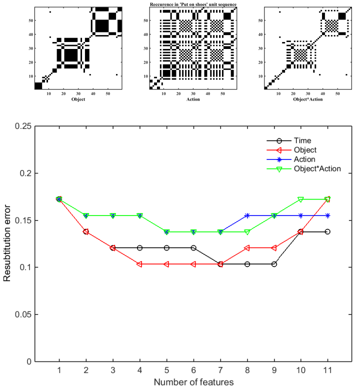


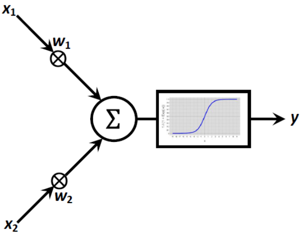
Figure: Binomial classifier (OCD vs. control) on unseen data.

[1] Zor, R., Hermesh, H., Szechtman, H., & Eilam, D. (2009). Turning order into chaos through repetition and addition of elementary acts in obsessive-compulsive disorder (OCD). World Journal of Biological Psychiatry, 10(4.2), 480–487.

Nielbo, K. L., Fux, M., Mort, J., Zamir, R., & Eilam, D. (2017). Structural differences among individuals, genders and generations as the key for ritual transmission, stereotypy and flexibility. Behaviour, 154(1), 93–114.

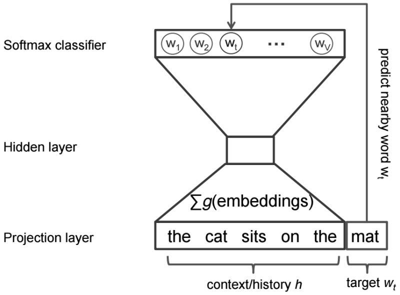
for text data, deep learning* has become an increasingly popular technology for feature engineering

embeddings are trained for all levels of text representation
– word, sentence, document with topic, sentiments, POS &c
– distributed representations with semantic properties

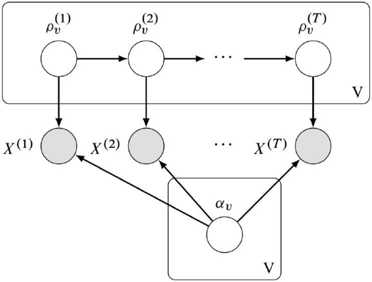


$(Copenhagen - Denmark) + Norway \approx Oslo$
$(summer - warm) + winter \approx cold$
$(dogs - dog) + cat \approx cats$

– words are similar if they appear in similar contexts
– embedding encode the **distribution of word contexts** under certain conditions (e.g., window, auxiliary task, topics &c)



www.tensorflow.org



COMPUTER (Senate)

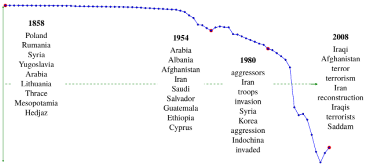
1858	1986
computer	computer
draftsman	software
draftsmen	computers
copyist	copyright
photographer	technological
computers	innovation
copyists	mechanical
janitor	hardware
accountant	technologies
bookkeeper	vehicles

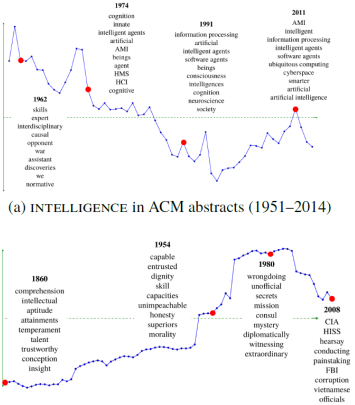
BUSH (Senate)

1858	1990
bush	bush
barberry	cheney
rust	nonsense
bushes	nixon
borer	reagan
eradication	george
grasshoppers	headed
cancer	criticized
tick	clinton
eradicate	blindness

DATA (ACM)

1961	1969	1991	2011	2014
data	data	data	data	data
directories	repositories	voluminous	raw data	data streams
files	voluminous	raw data	voluminous	voluminous
bibliographic	lineage	repositories	data sources	raw data
formatted	metadata	data streams	data streams	warehouses
retrieval	snapshots	data sources	dws	dws
publishing	data streams	volumes	repositories	repositories
archival	raw data	dws	warehouses	data sources
archives	cleansing	dsms	marts	data mining
manuscripts	data mining	data access	volumes	marts



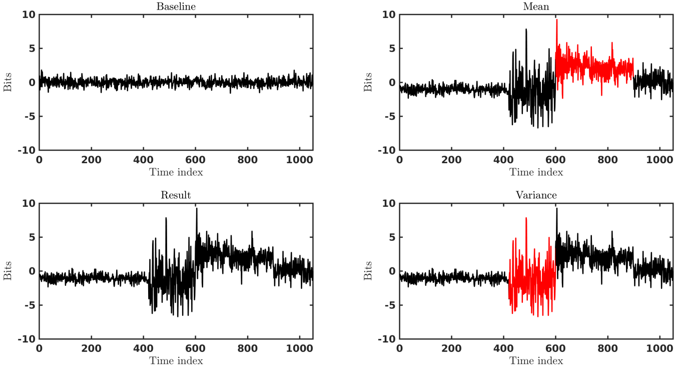


(a) INTELLIGENCE in ACM abstracts (1951–2014)

(b) INTELLIGENCE in U.S. Senate speeches (1858–2009)

Rudolph, M., Blei, D. (2017). Dynamic Bernoulli Embeddings for Language Evolution, arXiv 1703.08052

Media Studies and Cliometrics|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

dynamics from text

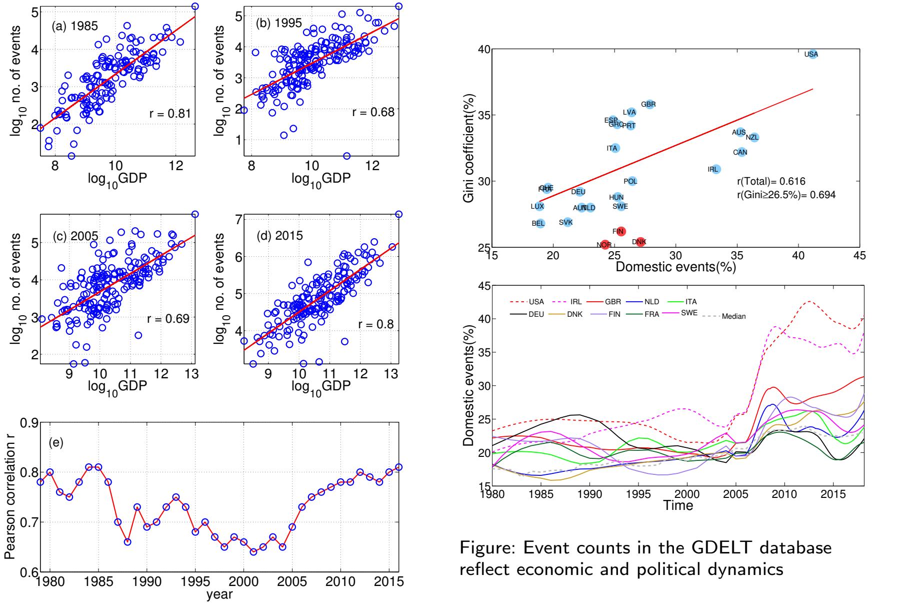


Figure: Event counts in the GDELT database reflect economic and political dynamics

Gao J., Ma M., Liu B., Nielbo K.L., Roepstorff A., Tangherlini T., Roychowdhury V. (in review) Brexit and Trump Presidency: were they black swan events or inevitable?

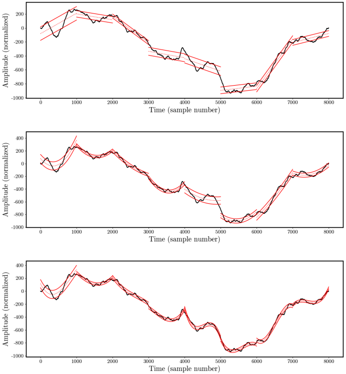


Figure: Computation of local fluctuations, RMS, around linear, quadratic, and cubic trends.

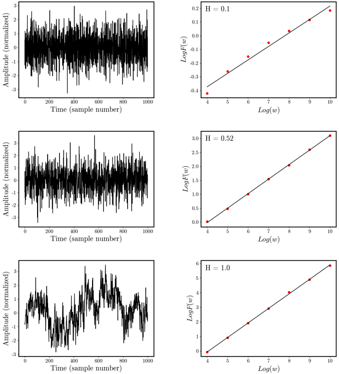
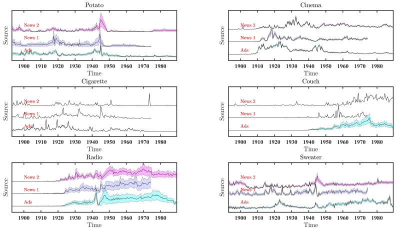


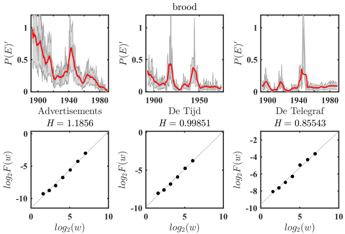
Figure: Computation of Hurst exponent (H) for anti-persistent, memoryless and persistent processes.

for $1/f^{2H+1}$ processes: anti-persistent process: $0 < H < 0.5$, short-range correlations only $H = 0.5$, and $0.5 < H < 1$ **persistent process**

https://github.com/knielbo/saffine

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)

Wevers, M., Nielbo, K. L., & Gao, J. (in review). Tracking the Consumption Junction: Temporal Dependencies in Dutch Newspaper Articles and Advertisements.

digression

Copyright & Privacy|Data Access and Mobility

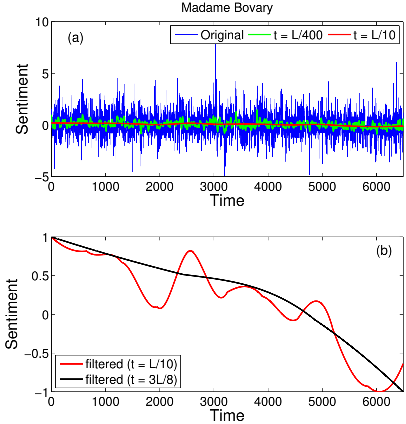
Challenges to computationally empowering humanities:

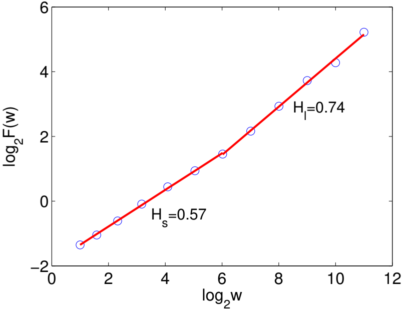
- technical competencies
- interdisciplinary respect and understanding
- epistemology differences
- data access and mobility

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 Studies|Sentiment Analysis





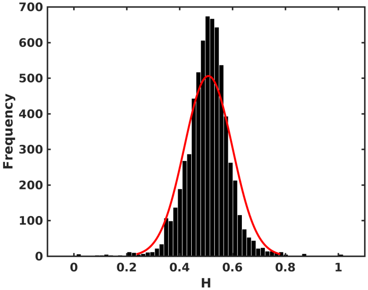


Figure: sentiment analysis and adaptive filtering reconstructs narrative vectors that reflect the reader experience. Particular fractal scaling-range, $0.6 < H \leq 0.8$, indicates literary optimality.

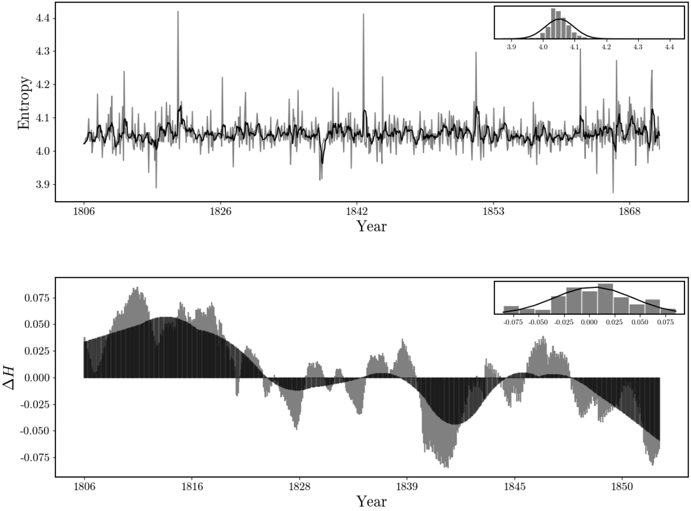
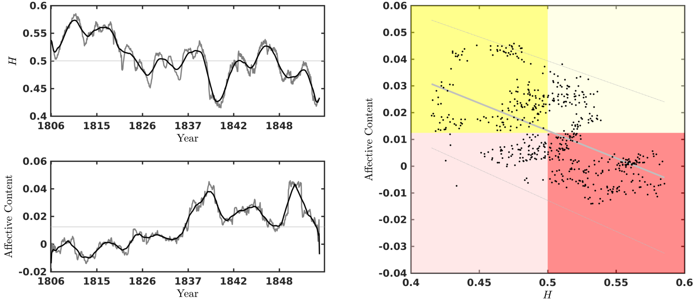


Figure: Author-profiling with time-varying Hurst parameter for lexical variability (entropy). Distinct periods in an author's development (Grundtvig).



EMOTION|Grundtvig

– early phase: negative affective tone

– late phase: positive affective tone

– inverse relation → `state incongruent writer`

– emotional state Granger-causes creative state → `dostoyevskian trope`

THANK YOU

knielbo@sdu.dk
knielbo.github.io

& credits to

Max R. Echardt and Katrine F. Baunvig, datakube, University of Southern Denmark, DK
David Eilam, Department of Zoology, Tel-Aviv University, IL
Jianbo Gao and Bin Liu, Institute of Complexity Science and Big Data, Guangxi University, CHN
Melvin Wevers, DH Lab, KNAW Humanities Cluster, NL
Culture Analytics @ Institute of Pure and Applied Mathematics, UCLA, US