

text analytics

tm-bootcamp

kristoffer l nielbo

kln@cas.au.dk

github.com/kln-courses/tm_bootcamp

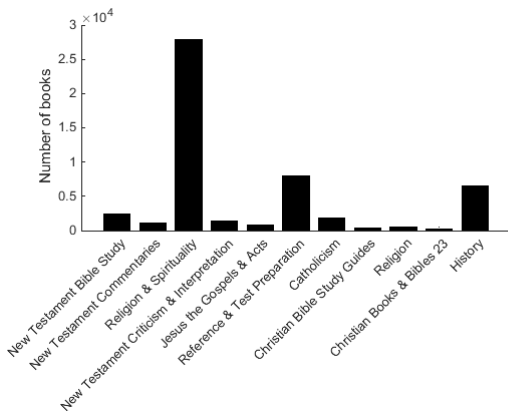
tmbootcamp.slack.com

IMC | AARHUS UNIVERSITY



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

Gospel of Marc (KJV) ~ 16500 words in 16 chp. on 11 p.

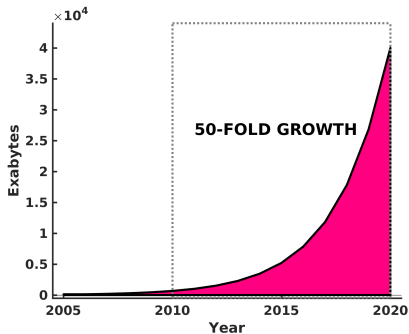


'from the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days ... and the pace is accelerating'

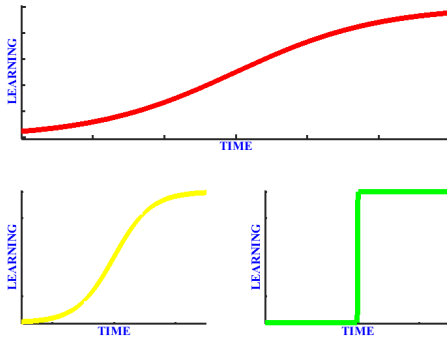
Eric Smith (Google)

'increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets '

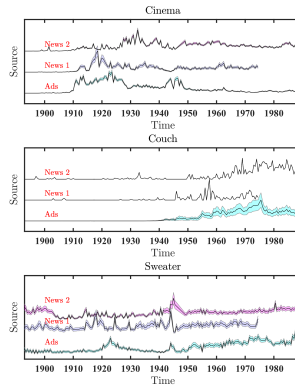
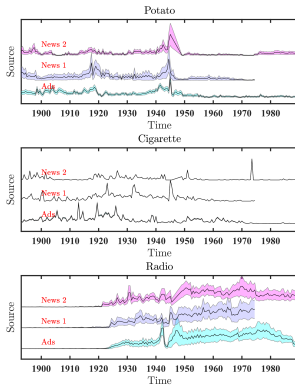
Jim Gray (Fourth Paradigm)



computational sciences are entering the exa-scale era
+
digital technologies are disruptive on a new scale

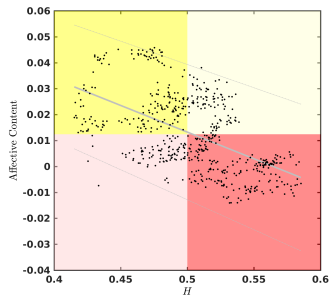
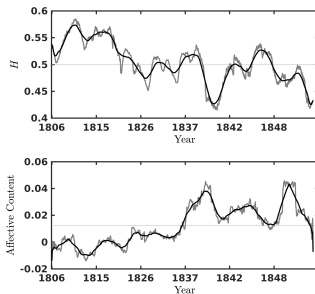


every knowledge-intensive industry have to “break” the learning curve



Digital history and media studies

- prerequisite: humanistic domain experts that use content analysis
- source digitization (newspapers) og super computing change resolution and scale
- technologies create new standards for the domains involved
- share technology, but not data!



Computational literary history

- prerequisite: humanistic domain experts that study writers and literary periods
- high quality digitization of writers, annotation and NLP changes perspective and scale
- technologies that are creating new standards
- sharing of technology and data

Data



Information



Presentation



Knowledge



Data



Information



Presentation



Knowledge



Data



Information

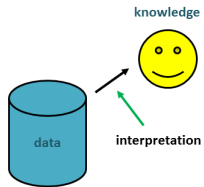


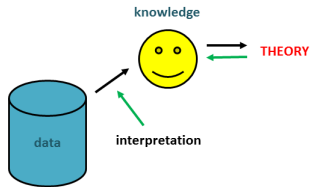
Presentation



Knowledge





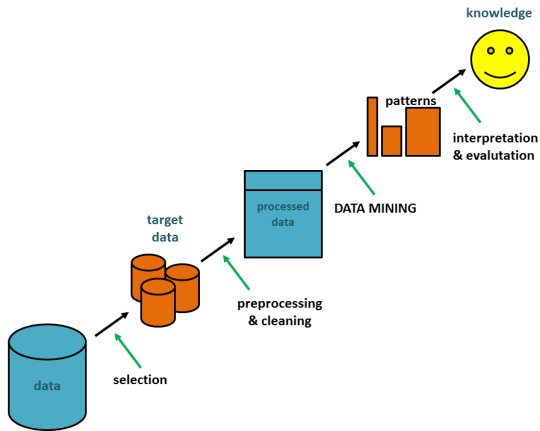


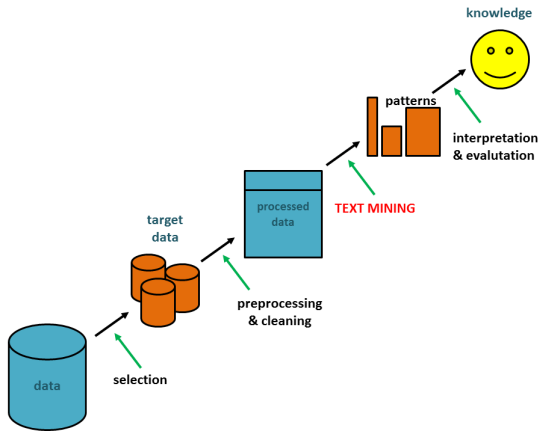
knowledge



THEORY







text analytics \sim text mining \sim automated text analysis

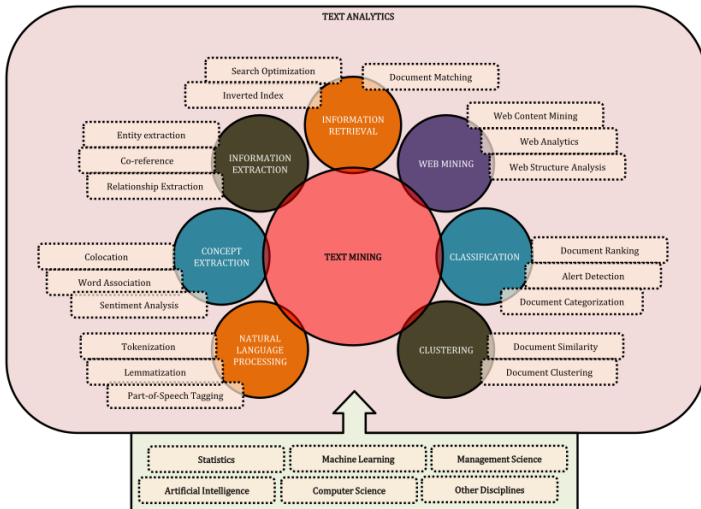
set of data mining¹ techniques for extracting high quality information from
large scale text-heavy (unstructured) data sets

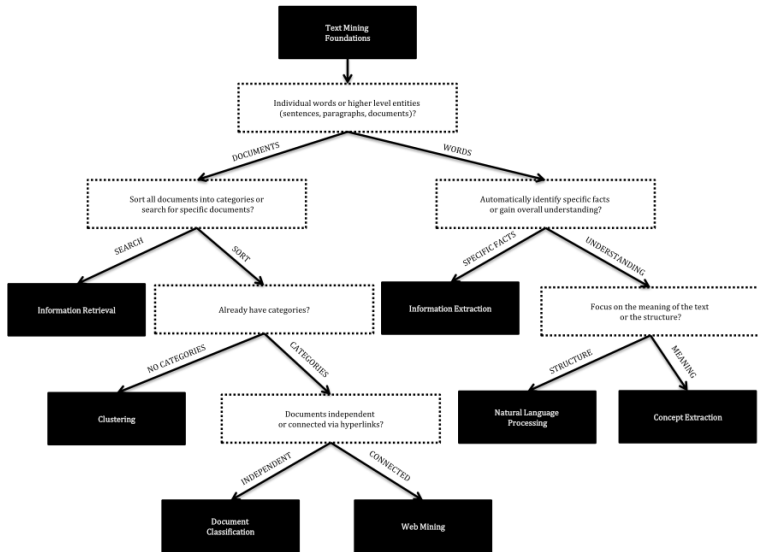
(\sim Miner et al 2012)

a tool for discovery and measurement in textual data of
prevalent attitudes, concepts, or events

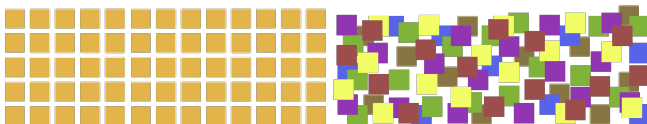
(\sim O'Connor, Bamman & Smith 2011)

¹Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37.





data objects that are described over a set of (qualitative or quantitative) features



fundamental difference between structured data and **unstructured* data**

- word processing files, pdfs, emails, social media posts, digital images, video, and audio
- today > 80% of all data are unstructured
- increased demand for expertise from culture, media and linguistic domains

adequate problem solution requires that we test a range of approaches (algorithms, (hyper-)parameter estimation) - the validation of an approach is an **experiment**

experiment input: code, data sets, hyperparameter values

experiment output: model definition (weights), metric values (experiment comparison), execution logs

a complex and error-prone process

⇒ systematically comment your work and process and use version control and source code management





TAPoR 3

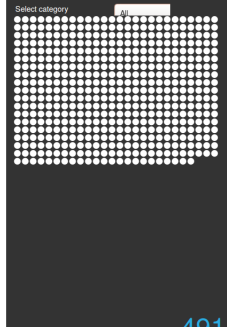
Discover research tools for studying texts.

Search

Voyant 2.0 is a complete rewrite of Voyant. It provides a suite of text analysis tools that will work with most texts you can upload or find on the web. These tools are combined in skins. For documentation see [Documentation for Voyant 2.0](#)



Voyant Tools 2.0 (Corpus View)



“There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea”

Andreas Buja