text analytics

tm-bootcamp

kristoffer l nielbo
`kln@cas.au.dk`
`github.com/kln-courses/tm_bootcamp`
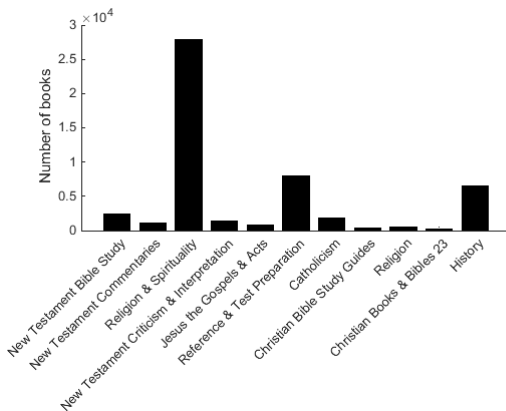`tmbootcamp.slack.com`

IMC|AARHUS UNIVERSITY

AARHUS UNIVERSITET

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

# Gospel of Marc (KJV) $\sim$ 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'
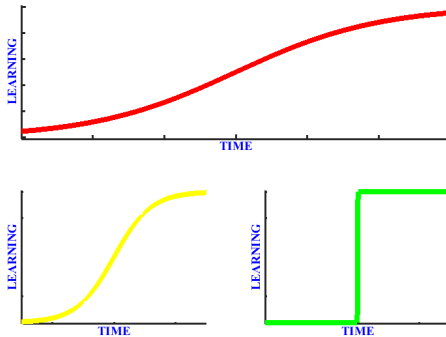
Eric Smith (Google)

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

Jim Gray (Fourth Paradigm)
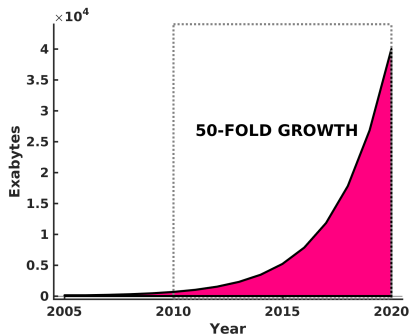
computational sciences are entering the exa-scale era
+
digital technologies are disruptive on a new scale

every knowledge-intensive industry have to "break" the learning curve

INTERVENTION|from the console

———

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

AARHUS UNIVERSITET

Pannang Curry 1-2-3!

1. Open your right refrigerator door and remove ingredients from the following locations: Door shelf 2, Spot 1; Crisper drawer 1, Spot 3; Crisper drawer 1, Spot 5.

2. Open your third kitchen drawer and remove the utensils labeled "1", "3", "4", "9", and "12".

3. Use your arms to apply utensil 1 to ingredients 1-3. Place ingredients inside utensil 3.

*Note: This recipe uses ShaKL the Shared Kitchen Layout. To use ShaKL, you'll need to have installed ShaKL shelving, cabinetry, and utensils throughout your kitchen and pantry and have basic understanding of ShaKL managers. To learn more, read *Up and Running with ShaKL* (O'Billy Press, 2015). Want to improve ShaKL? Consider contributing to our team.

AARHUS UNIVERSITET

IMC
INTERACTING MINDS CENTRE
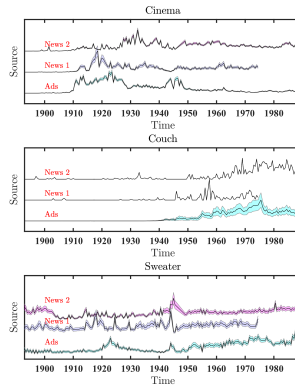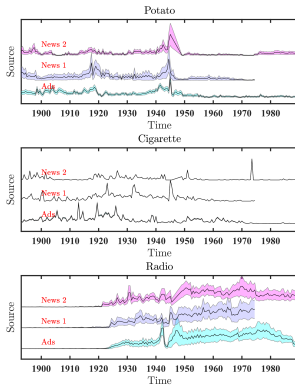
Digital history and media studies
– prerequisite: humanistic domain experts that use content analysis
– source digitization (newspapers) og super computing change resolution and scale
– technologies create new standards for the domains involved
– share technology, but not data!
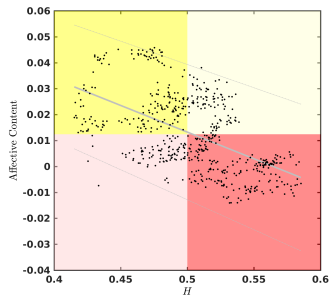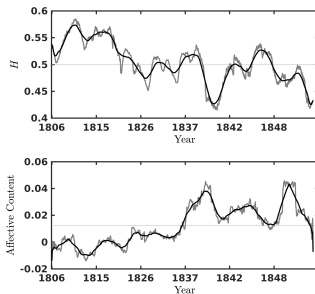
Computational literary history
– prerequisite: humanistic domain experts that study writers and literary periods
– high quality digitization of writers, annotation and NLP changes perspective and scale
– technologies that are creating new standards
– sharing of tehcnology and data

Data   Information   Presentation   Knowledge

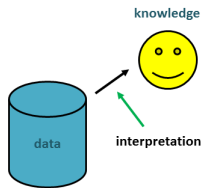**Data**   **Information**   **Presentation**   **Knowledge**
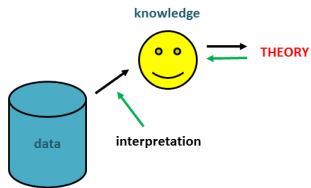
Data

Information

Presentation

Knowledge

knowledge

THEORY

knowledge

patterns

interpretation
& evalutation

DATA MINING

processed
data

target
data

preprocessing
& cleaning

data

selection

data selection → target data → preprocessing & cleaning → processed data → TEXT MINING → patterns → interpretation & evalutation → knowledge
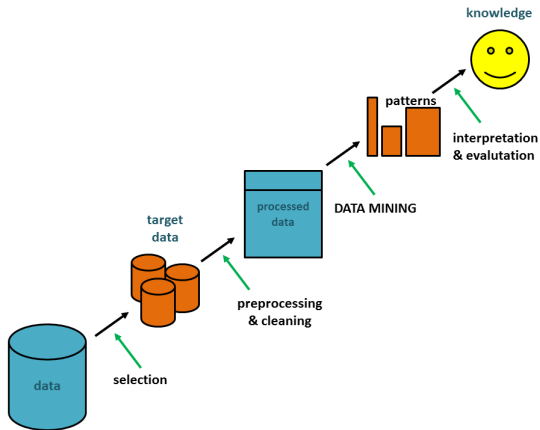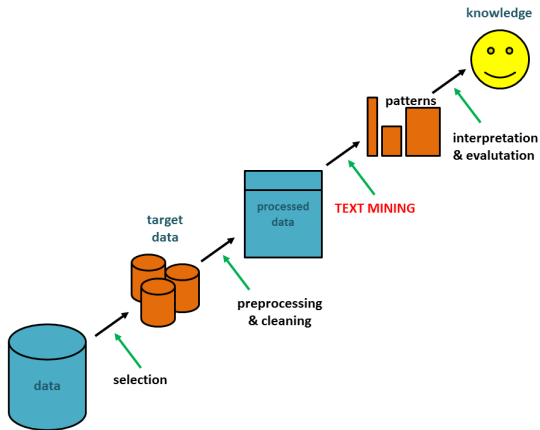
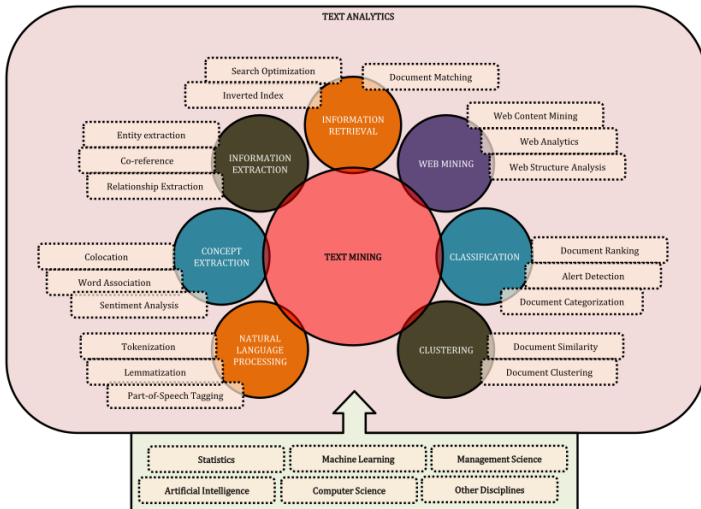`text analytics` $\sim$ text mining $\sim$ automated text analysis

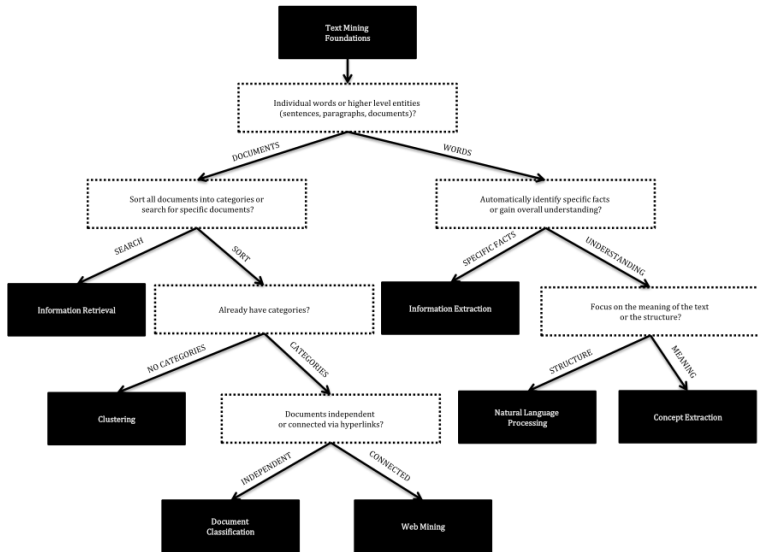set of data mining[1] techniques for extracting high quality information from
large scale text-heavy (unstructured) data sets

($\sim$ 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)

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

TEXT ANALYTICS

INFORMATION RETRIEVAL
- Search Optimization
- Inverted Index
- Document Matching

INFORMATION EXTRACTION
- Entity extraction
- Co-reference
- Relationship Extraction

WEB MINING
- Web Content Mining
- Web Analytics
- Web Structure Analysis

CONCEPT EXTRACTION
- Colocation
- Word Association
- Sentiment Analysis

TEXT MINING

CLASSIFICATION
- Document Ranking
- Alert Detection
- Document Categorization

NATURAL LANGUAGE PROCESSING
- Tokenization
- Lemmatization
- Part-of-Speech Tagging

CLUSTERING
- Document Similarity
- Document Clustering

- Statistics
- Artificial Intelligence
- Machine Learning
- Computer Science
- Management Science
- Other Disciplines

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

<div align="center">

`a complex and error-prone process`

</div>

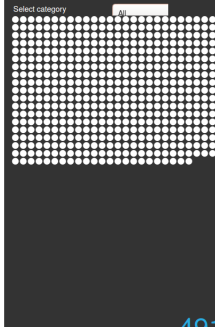$\Rightarrow$ systematically comment your work and process and use version control and source code management

AARHUS UNIVERSITET

IMC
INTERACTING MINDS CENTRE

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