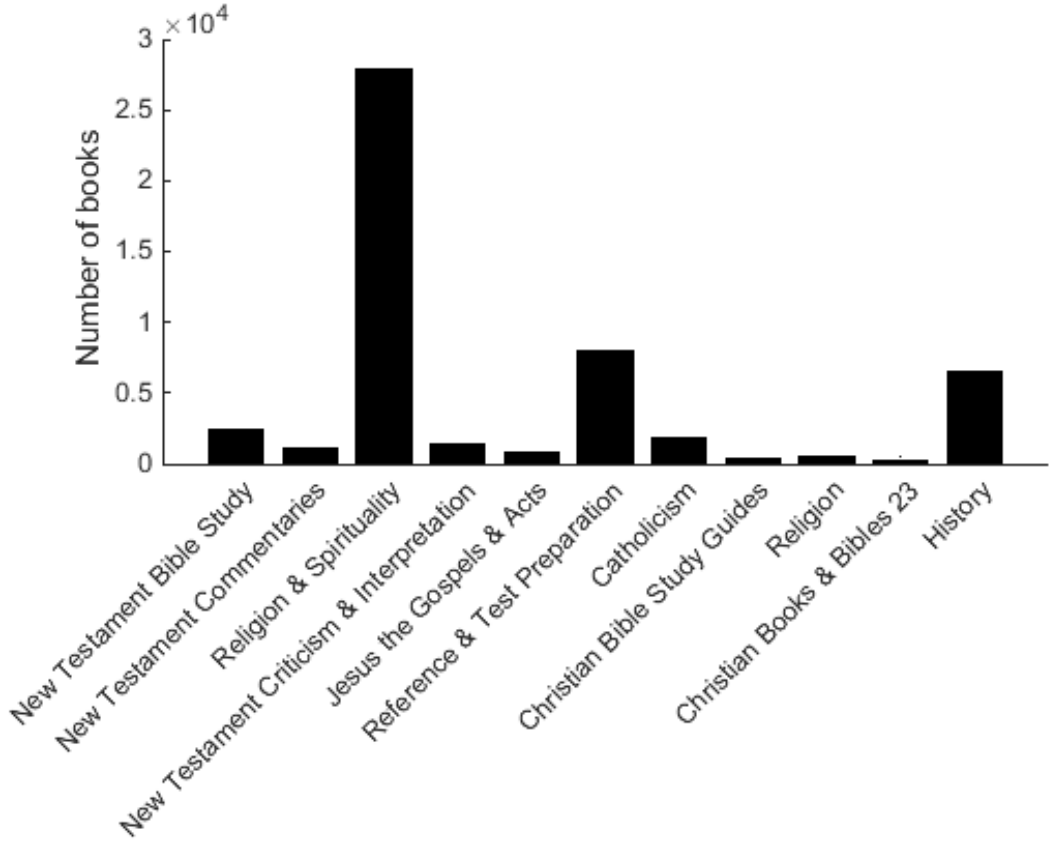
text analytics and generic tools

Kristoffer L. Nielbo
kln@cas.au.dk

DTL|Digital Arts Initiative
Interacting Minds Centre|Aarhus University

Gospel of Marc (KJV) $\sim$ 16500 words in 16 chp. on 11 p.

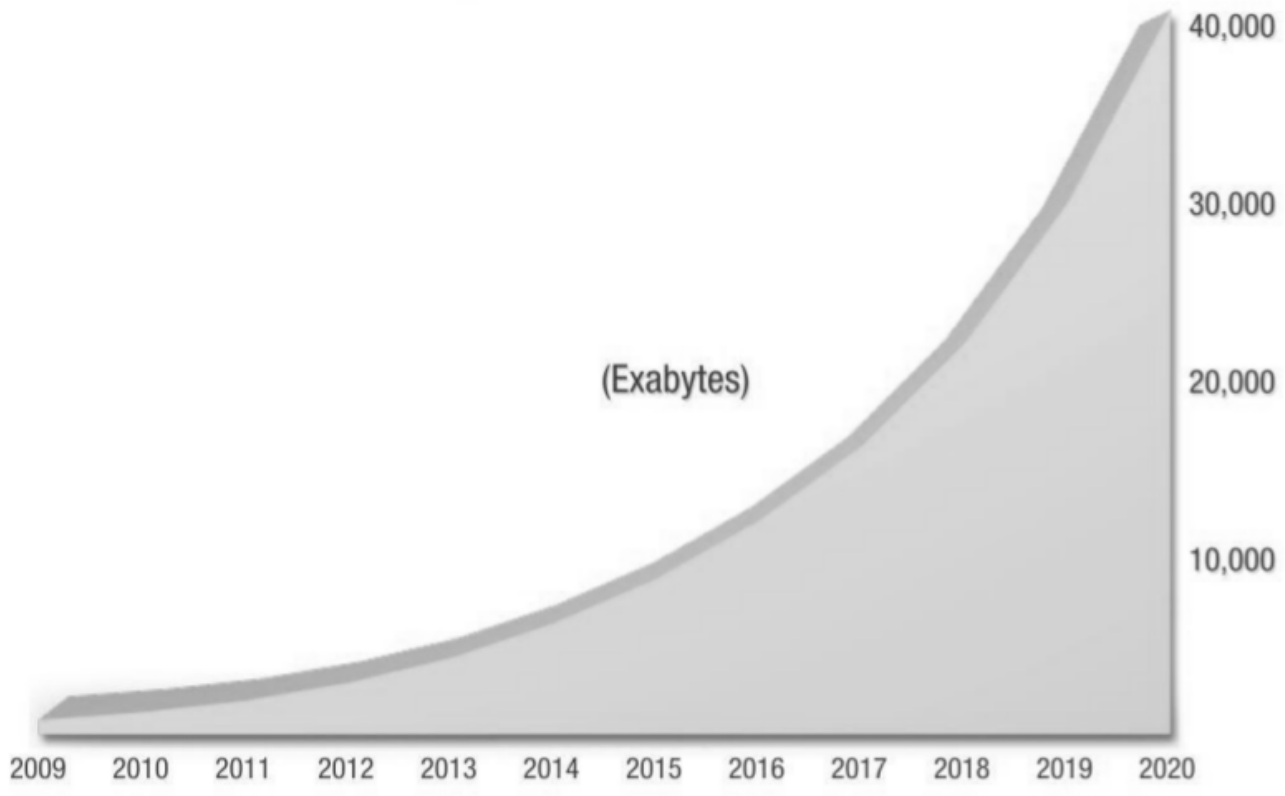


'from the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days ... and the pace is accelerating'

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)


40,000
30,000
20,000
10,000
(Exabytes)
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Data	Information	Presentation	Knowledge
			

Data

Information

Presentation

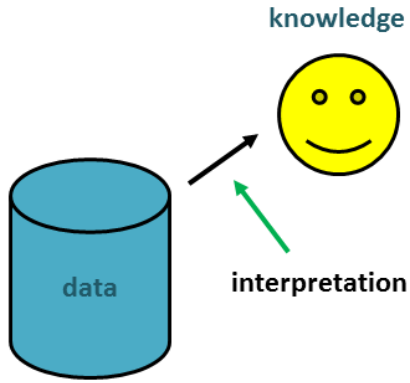
Knowledge

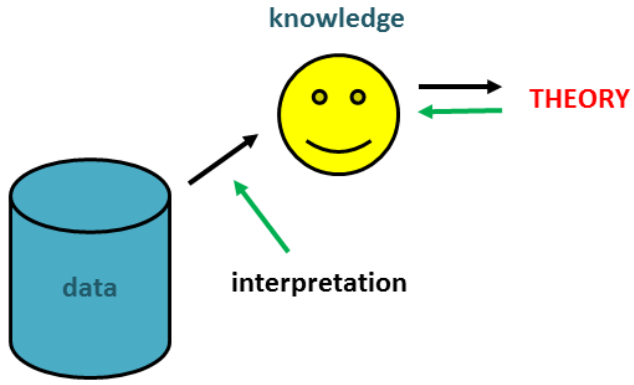
Data

Information

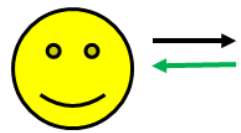
Presentation

Knowledge

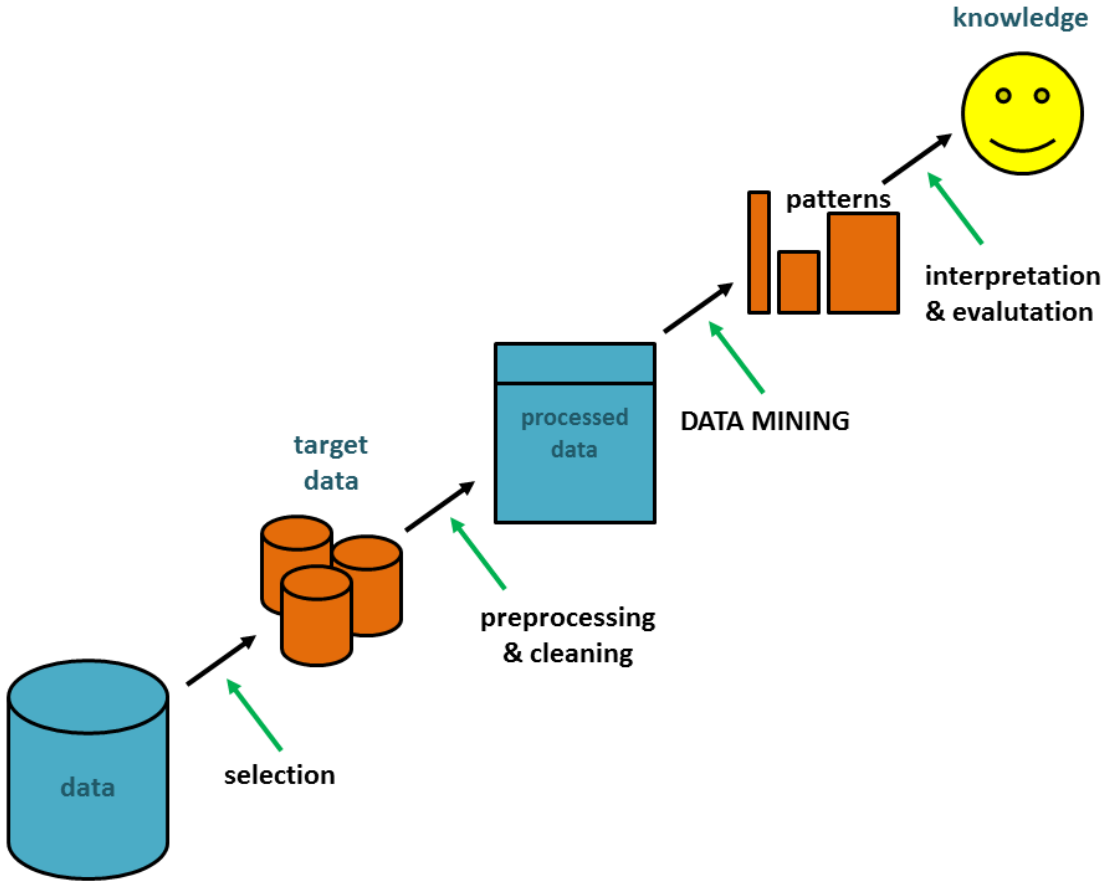

knowledge
data
interpretation

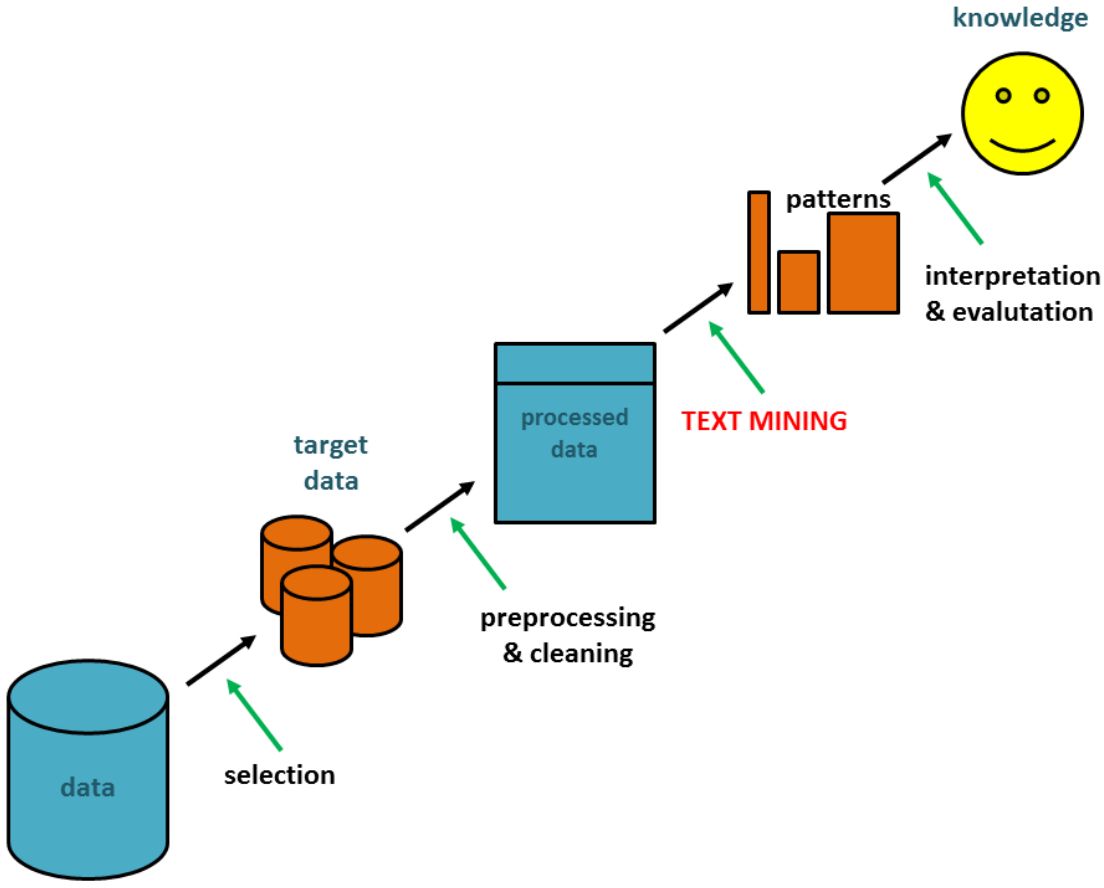

knowledge
THEORY
data
interpretation

knowledge



THEORY


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


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

text analytics $\sim$ text mining $\sim$ automated text analysis

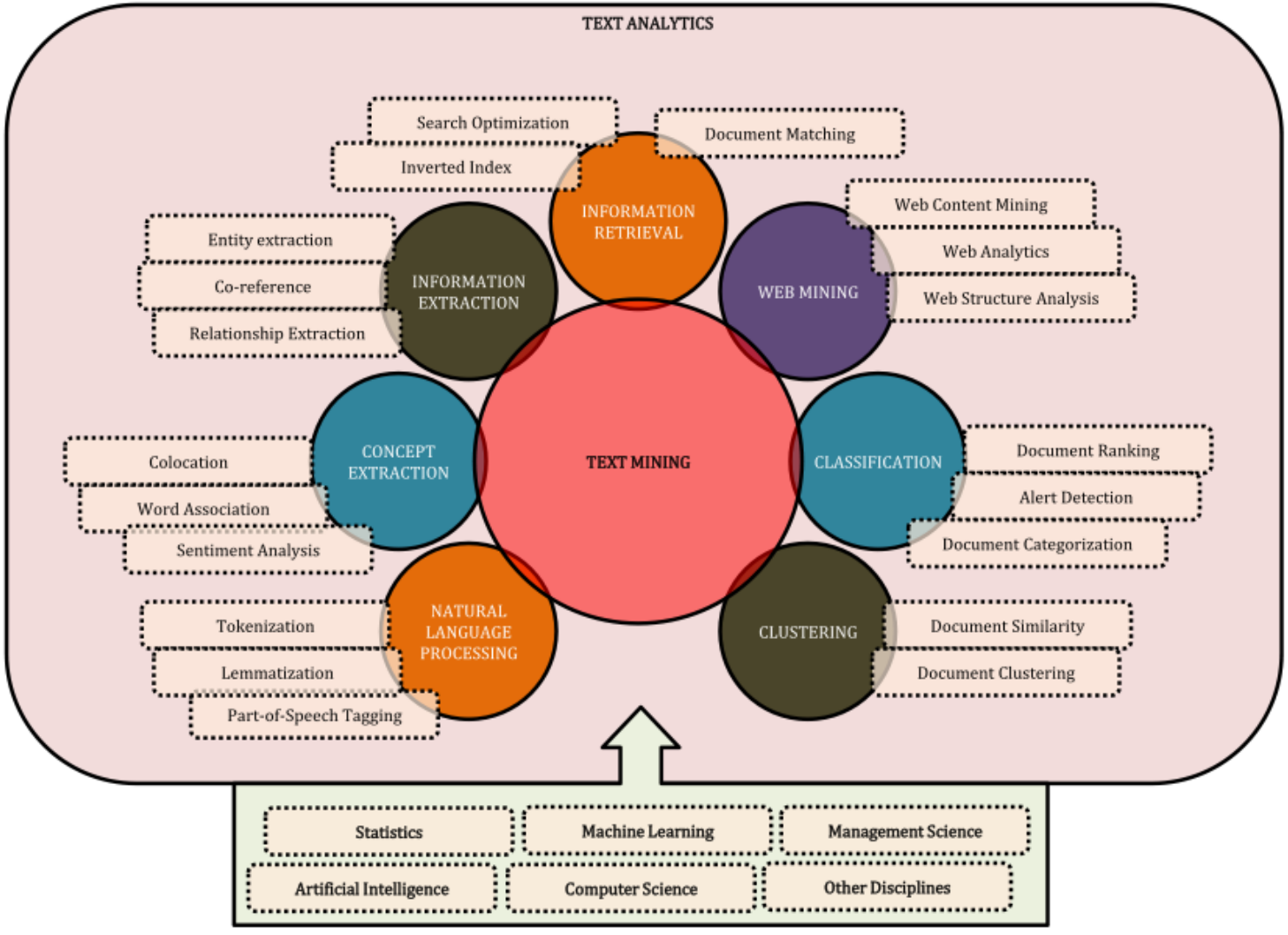
set of data mining[1] techniques for extracting high quality information from large scale text-heavy (unstructured) data sets

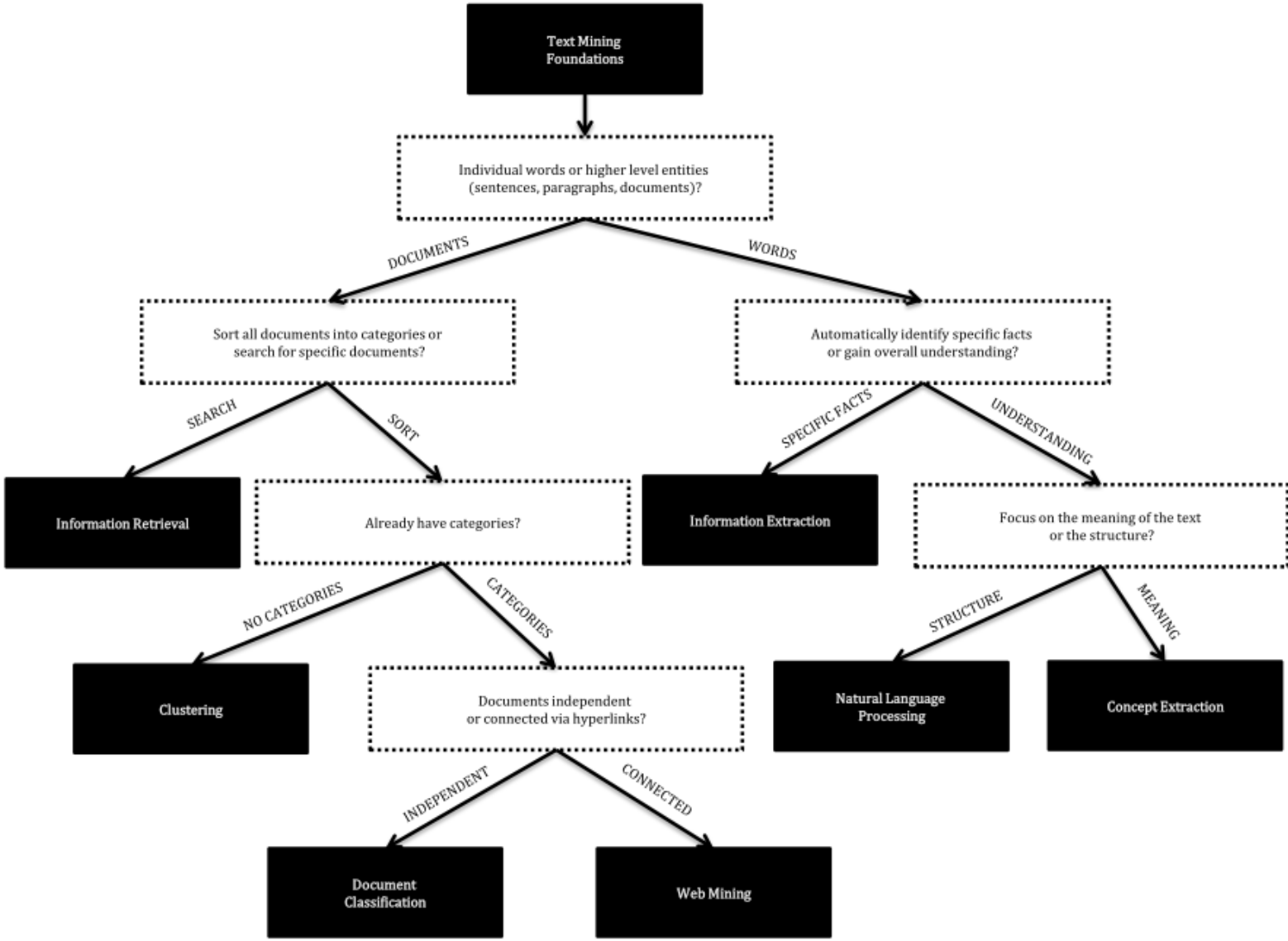
($\sim$ Miner et al 2012)

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

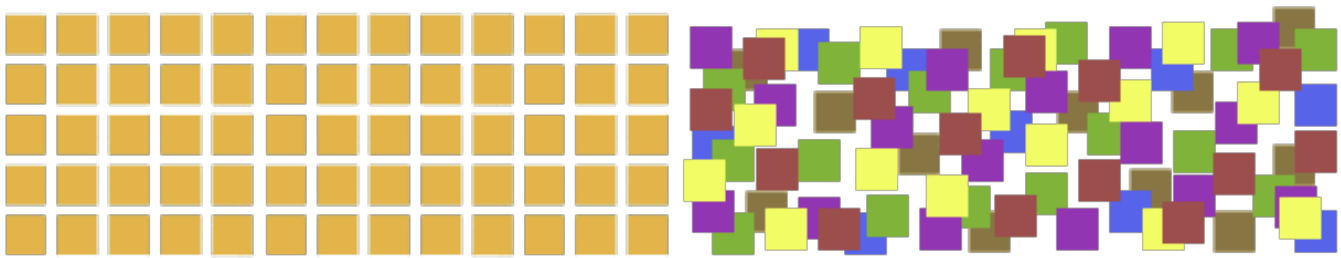
($\sim$ O'Connor, Bamman & Smith 2011)

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


TEXT ANALYTICS
Search Optimization
Inverted Index
Document Matching
INFORMATION RETRIEVAL
Entity extraction
Co-reference
Relationship Extraction
INFORMATION EXTRACTION
Web Content Mining
Web Analytics
Web Structure Analysis
WEB MINING
Colocation
Word Association
Sentiment Analysis
CONCEPT EXTRACTION
TEXT MINING
Document Ranking
Alert Detection
Document Categorization
CLASSIFICATION
Tokenization
Lemmatization
Part-of-Speech Tagging
NATURAL LANGUAGE PROCESSING
Document Similarity
Document Clustering
CLUSTERING
Statistics
Machine Learning
Management Science
Artificial Intelligence
Computer Science
Other Disciplines


Text Mining Foundations
Individual words or higher level entities (sentences, paragraphs, documents)?
DOCUMENTS
WORDS
Sort all documents into categories or search for specific documents?
Automatically identify specific facts or gain overall understanding?
SEARCH
SORT
SPECIFIC FACTS
UNDERSTANDING
Information Retrieval
Already have categories?
Information Extraction
Focus on the meaning of the text or the structure?
NO CATEGORIES
CATEGORIES
STRUCTURE
MEANING
Clustering
Documents independent or connected via hyperlinks?
Natural Language Processing
Concept Extraction
INDEPENDENT
CONNECTED
Document Classification
Web Mining

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

the goal of **statistical learning** is to build a machine that can learn from data and automatically make the right decisions

supervised learning infer mapping between data & class-information $\rightarrow$ 'ground truth'

unsupervised learning identify latent classes in the data $\rightarrow$ lack 'ground truth'

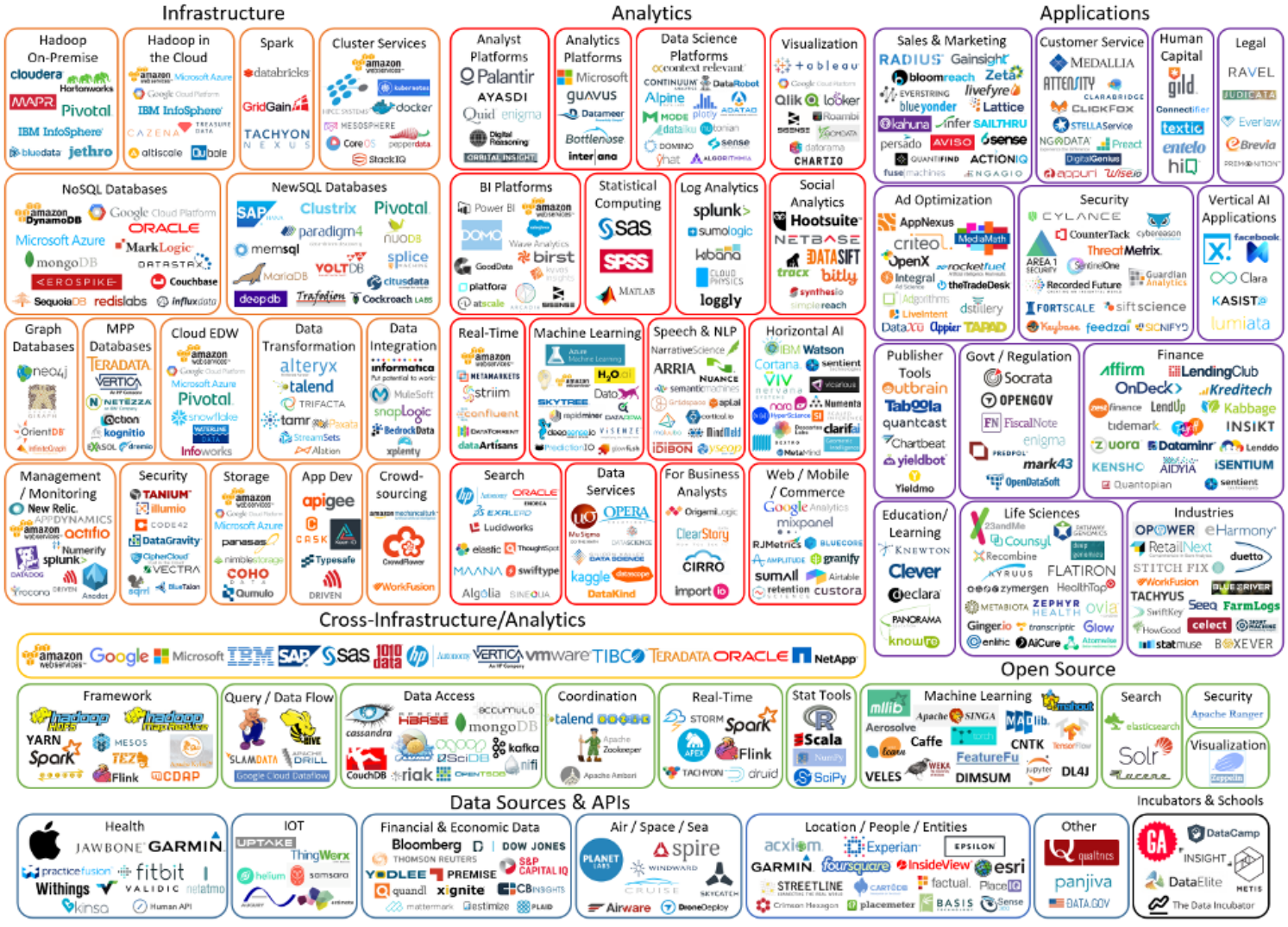
adequate problem solution requires that we test a range of approaches (algorithms, (hyper-)parameter estimation) - the validation of an approach is an **experiment**

experiment input: code, data sets, hyperparameter values

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

a complex and error-prone process

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


Infrastructure
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
NoSQL Databases
NewSQL Databases
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Analytics
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
Applications
Sales & Marketing
Customer Service
Human Capital
Legal
Ad Optimization
Security
Vertical AI Applications
Publisher Tools
Govt / Regulation
Finance
Education / Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other
Incubators & Schools


R

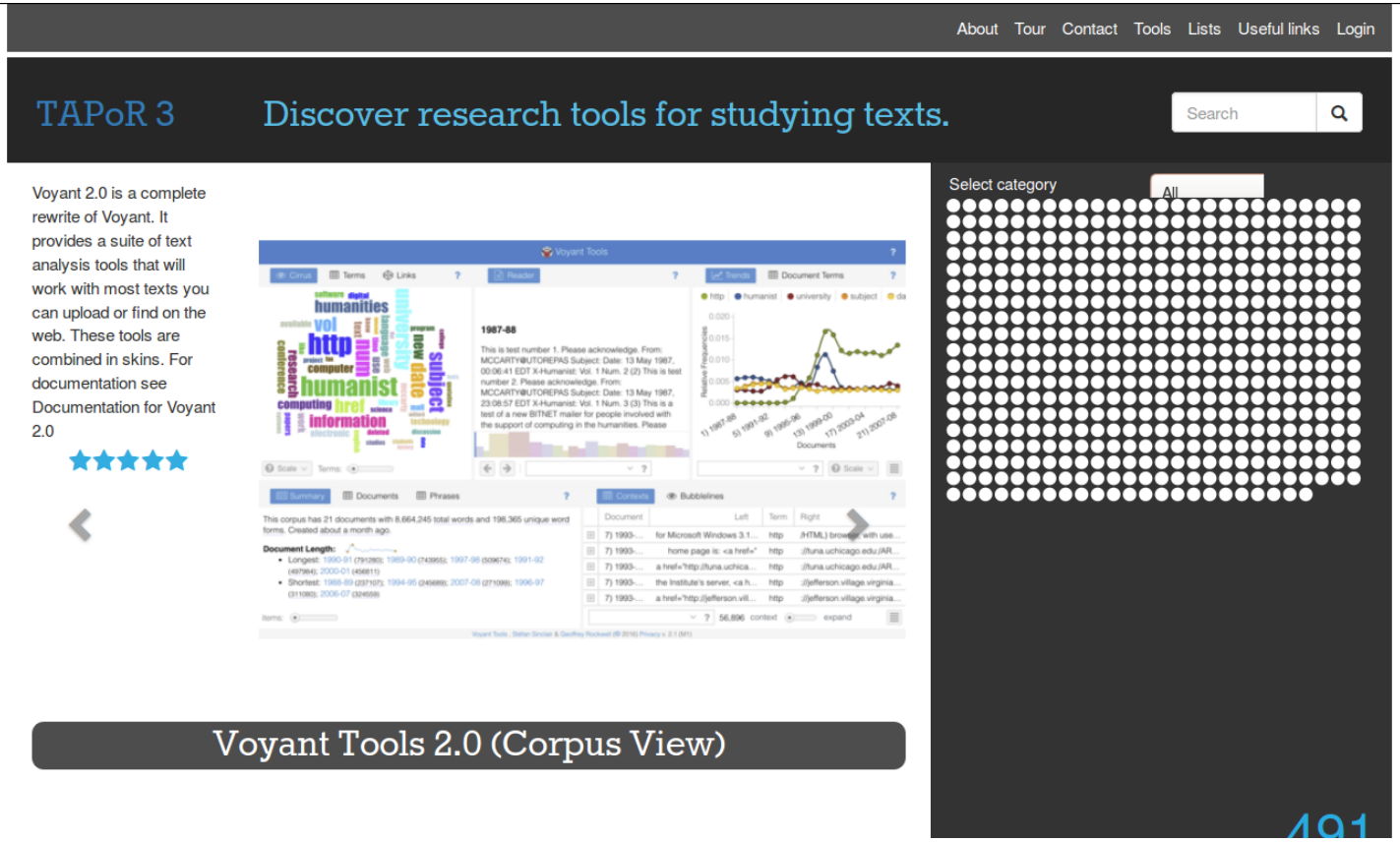

python™

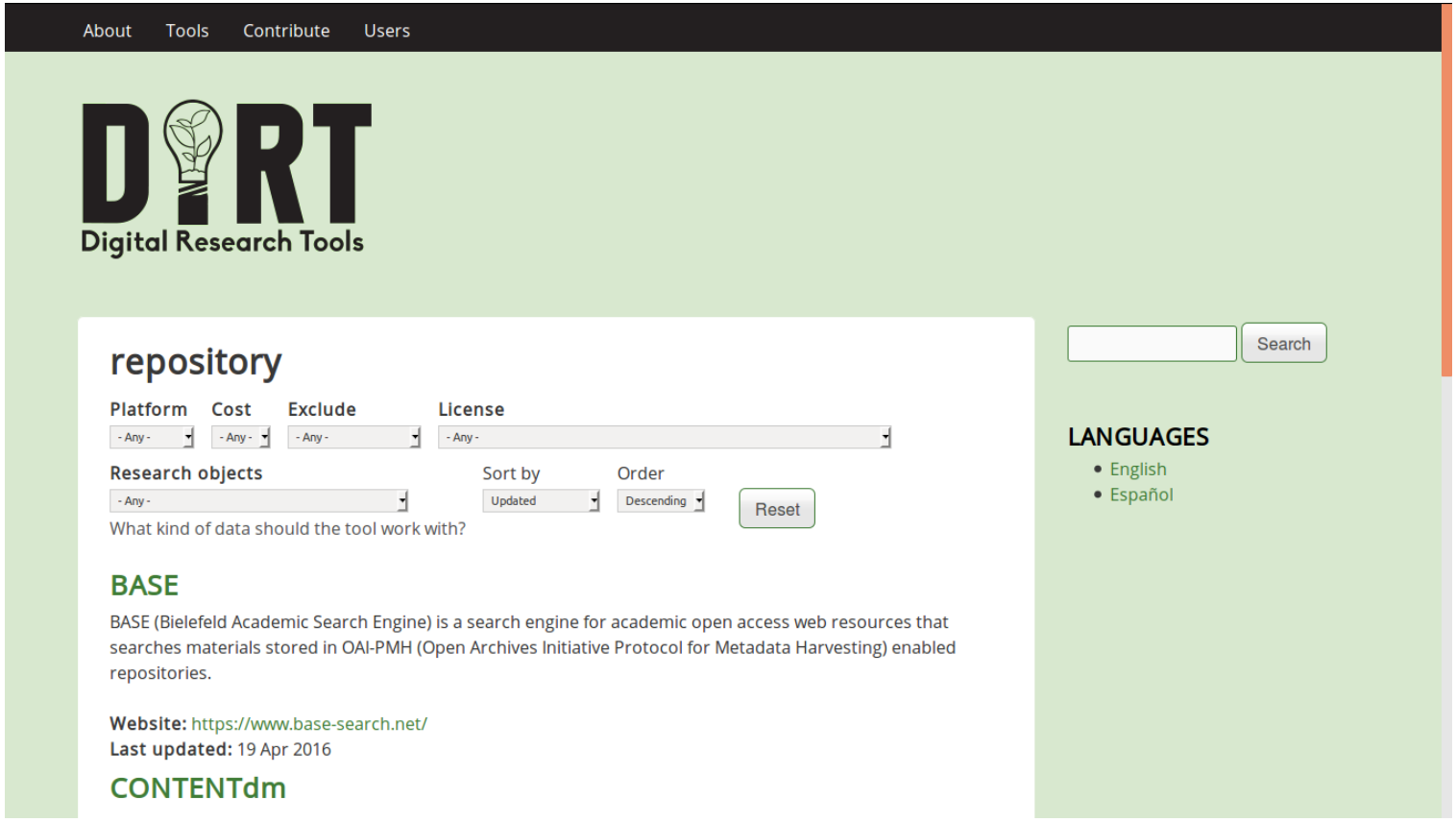

MATLAB


rapidminer


tableau


VOYANT
see through your text


About Tour Contact Tools Lists Useful links Login
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
Select category
All
Voyant Tools 2.0 (Corpus View)


About
Tools
Contribute
Users
DIRT
Digital Research Tools
repository
Platform
Cost
Exclude
License
- Any -
- Any -
- Any -
- Any -
Research objects
- Any -
Sort by
Updated
Order
Descending
Reset
What kind of data should the tool work with?
BASE
BASE (Bielefeld Academic Search Engine) is a search engine for academic open access web resources that searches materials stored in OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) enabled repositories.
Website: https://www.base-search.net/
Last updated: 19 Apr 2016
CONTENTdm
Search
LANGUAGES
English
Español

"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