Text Analytics (in the Digital Humanities)

aword of the conjunction of Digital and Humanities

HUMANITIES

redrawing boundary lines among the humanities, the social sciences, the arts, and the natural sciences

expanding the audience and social impact of scholarship in the humanties

developing new forms of inquiry and knowledge production

training future generations of humanists through hands-on, project-based learning

increase visability of humanitistic research

DIGITAL

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5*†] Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17*†]

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

The New Science of the Birth and Death of Words

Have physicists discovered the evolutionary laws of language in Google's library?

HUMANITIES 2.0

Analyzing Literature by Words and Numbers

Supercomputer predicts revolution

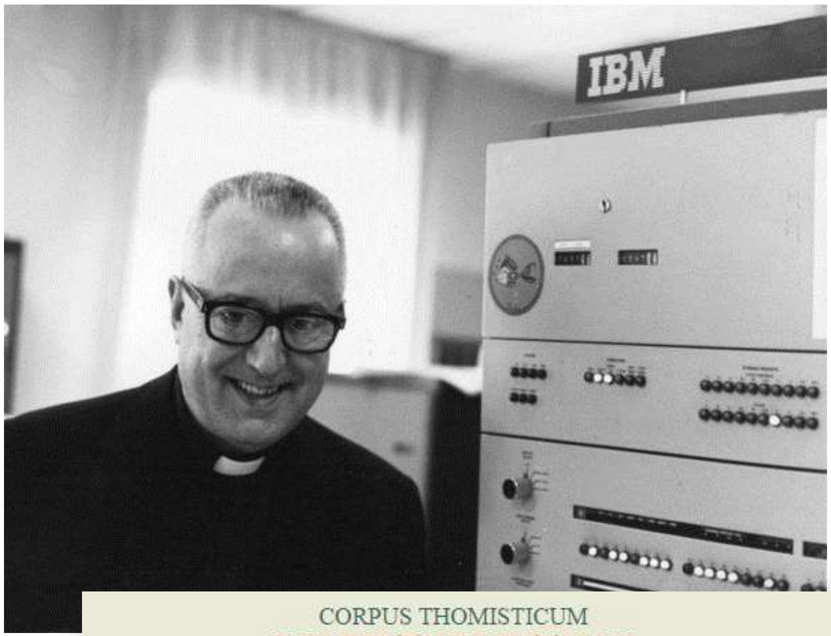
GRAY MATTER

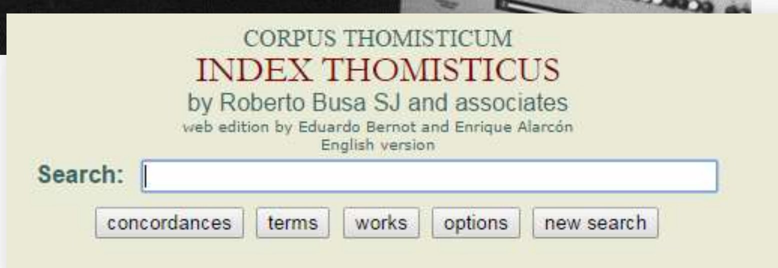
Twitterology: A New Science?

By BEN ZIMMER

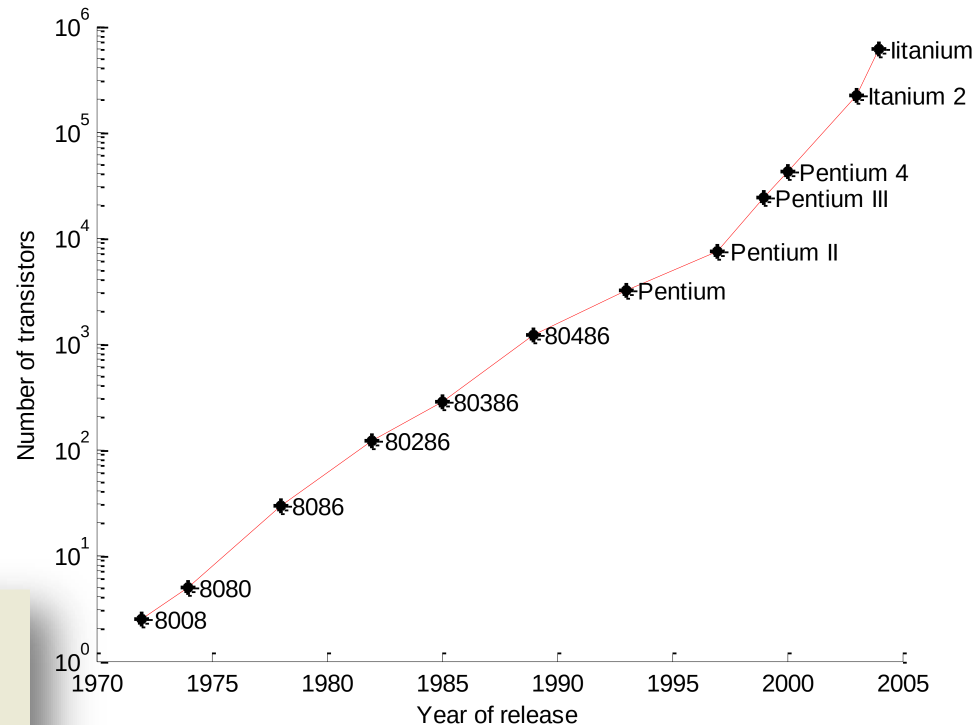
Published: October 29, 2011

DENIZENS of the Twitter-verse, please be advised: Whether you are a Libyan celebrating the demise of Col. Muammar el-Qaddafi, a New Zealand office worker sleepily starting your day or a California teenager trying out the latest slang, your words are being analyzed.





CORPUS THOMISTICUM
INDEX THOMISTICUS
by Roberto Busa SJ and associates
web edition by Eduardo Bernot and Enrique Alarcón
English version
Search:
concordances
terms
works
options
new search



Number of transistors
Year of release
10^0
10^1
10^2
10^3
10^4
10^5
10^6
1970
1975
1980
1985
1990
1995
2000
2005
8008
8080
8086
80286
80386
80486
Pentium
Pentium II
Pentium III
Pentium 4
Itanium 2
Iitanium 2

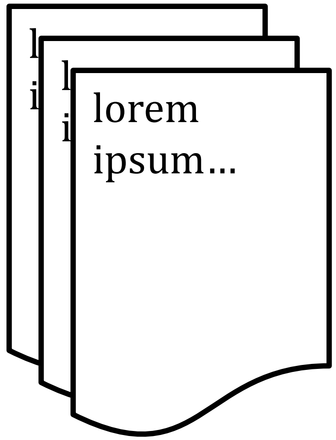
TEXT MINING

The Devil is known by many names

- text analytics
- predictive analytics
- automated text analysis
- computer assisted text analysis
- …

'is extracting high quality information from text through machine learning' (~ Miner et al. 2012)

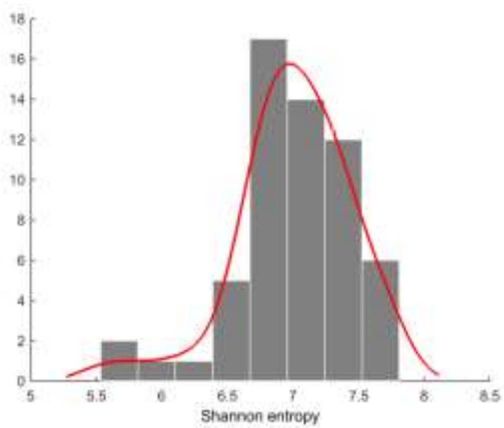
'is a tool for discovery and measurement in textual data of prevalent attitudes, concepts, or events.' (~ O'Connor, Bamman & Smith 2011)



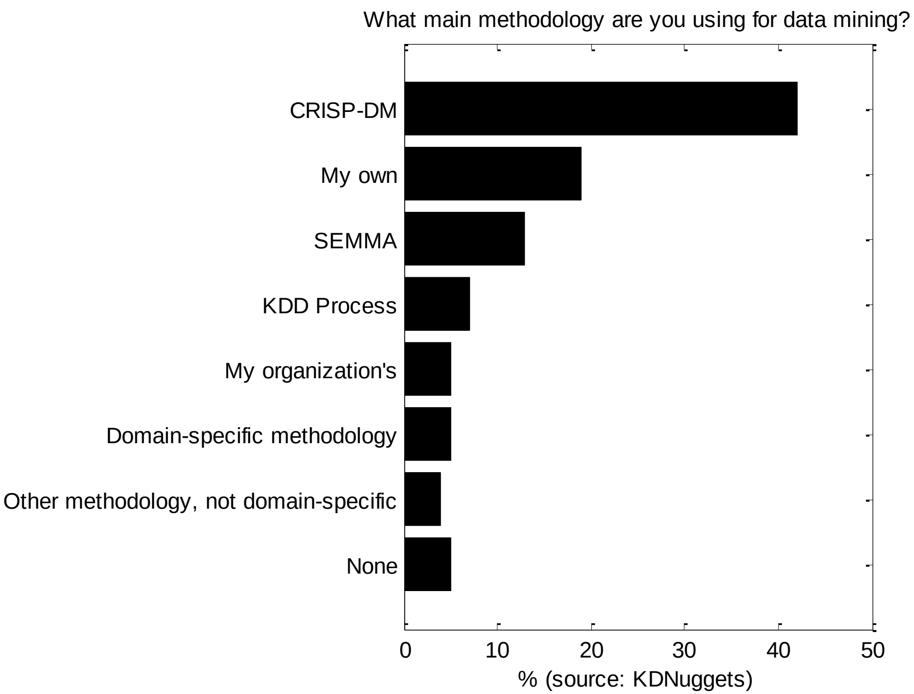
$$
\begin{array}{c}
\\
D_1 \\
D_2 \\
\vdots \\
\vdots \\
D_n
\end{array}
\begin{array}{cccc}
T_1 & T_2 & \cdots & T_t \\
\end{array}
$$

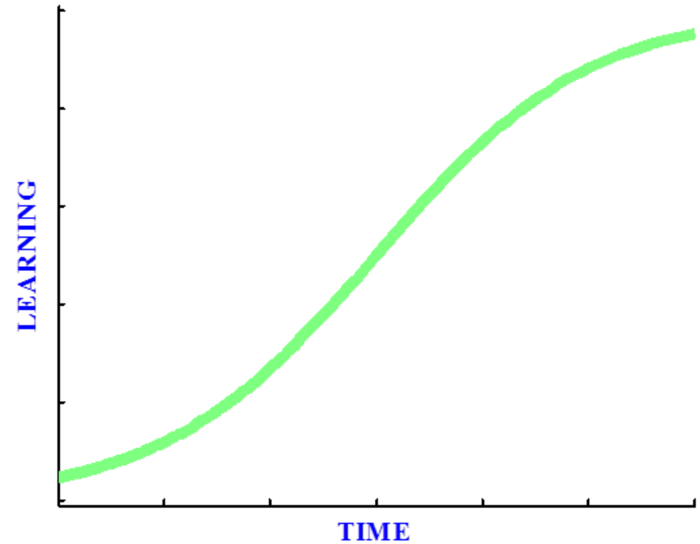
$$
\begin{array}{c|cccc}
 & T_1 & T_2 & \cdots & T_t \\
\hline
D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\
D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\
\vdots & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \vdots & & \vdots \\
D_n & w_{1n} & w_{2n} & \cdots & w_{tn}
\end{array}
$$

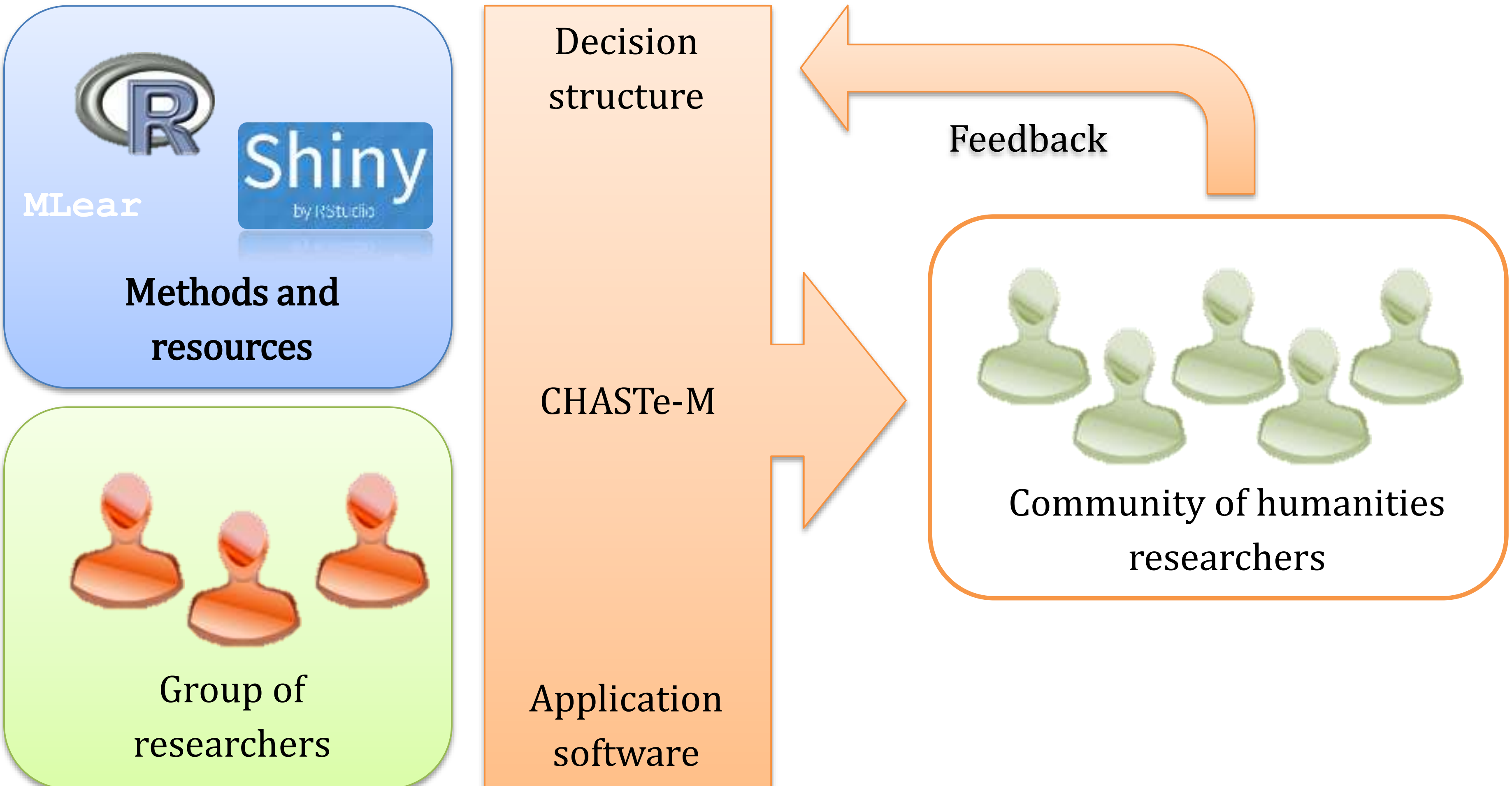
```
alltokensNum = cell(size(tokens));
tic
for t = 1:size(tokens,1)
    types = unique(tokens{t});
    typesNum = 1:length(types);
    tokensNum = zeros(size(tokens{t}));
        for i = 1:length(types)
            ii = strcmp(types(i),tokens{t});
            tokensNum(ii) = typesNum(i);
        end
    alltokensNum{t} = tokensNum;
    disp(t)
end
toc
```

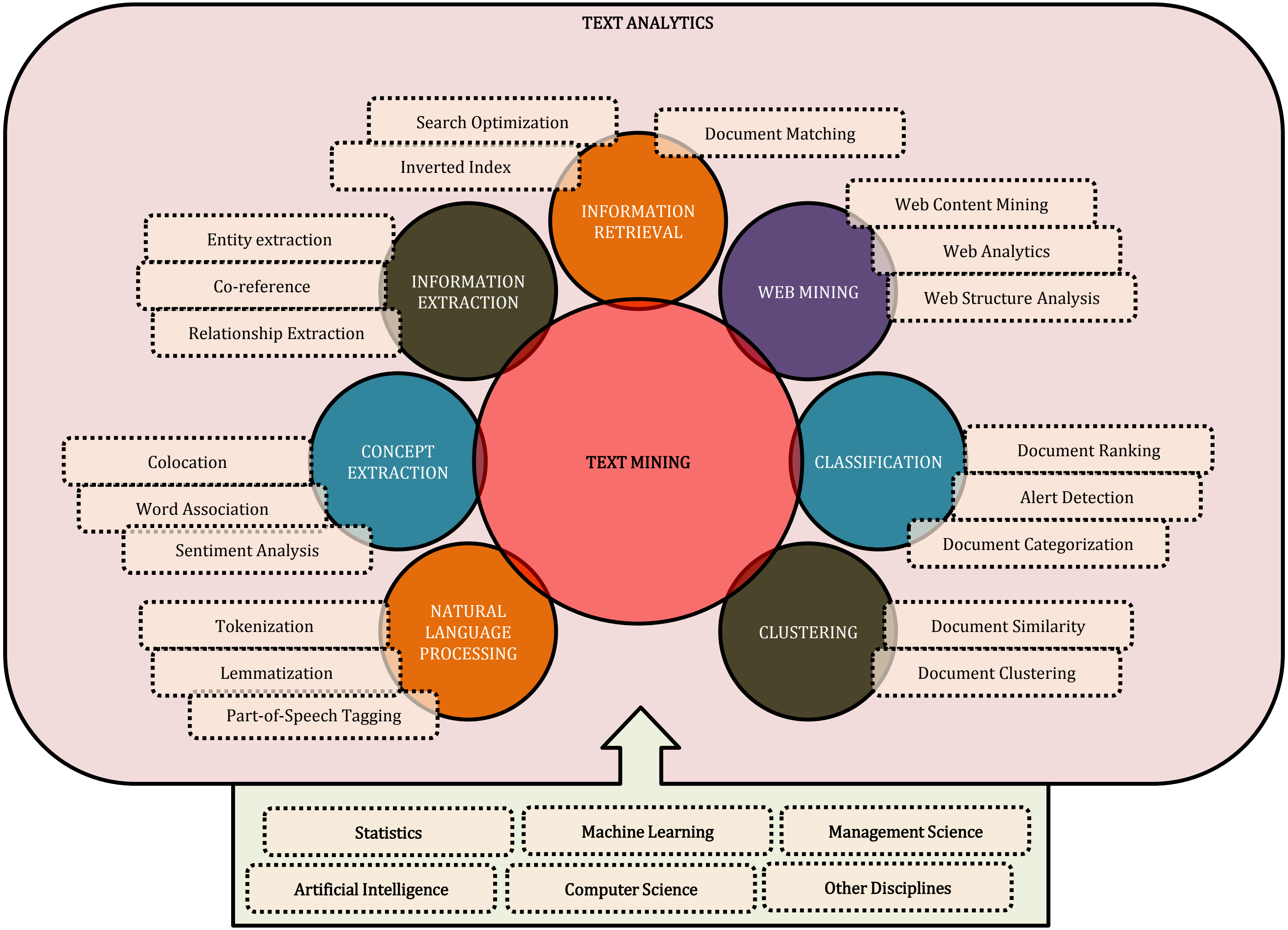


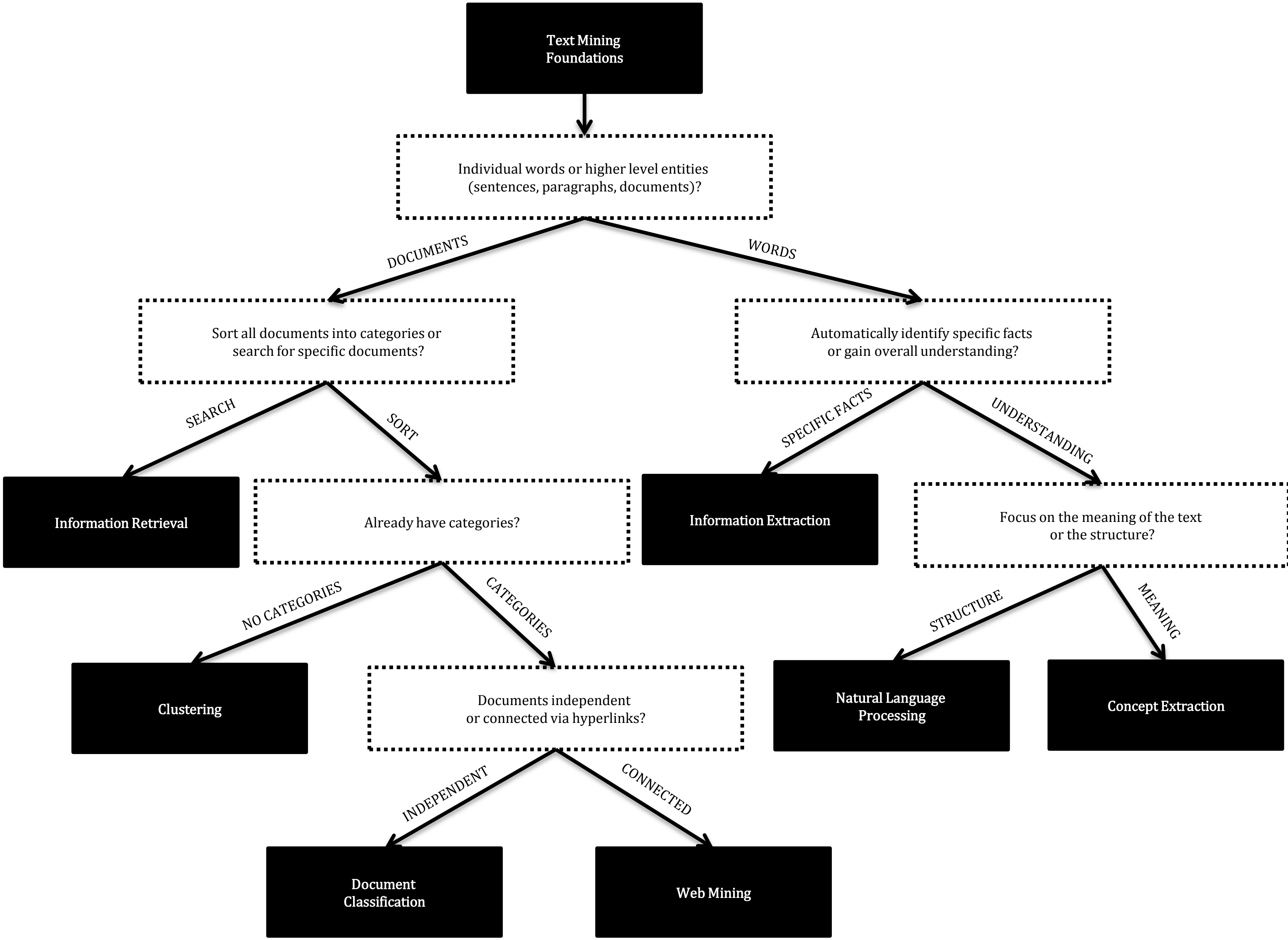
Text mining issues specific to the humanities



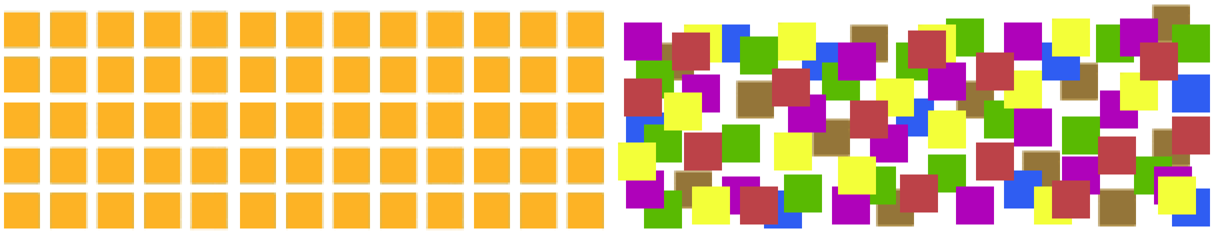



MLear
Shiny
by RStudio
Methods and resources
Group of researchers
Decision structure
CHASTe-M
Application software
Feedback
Community of humanities researchers


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

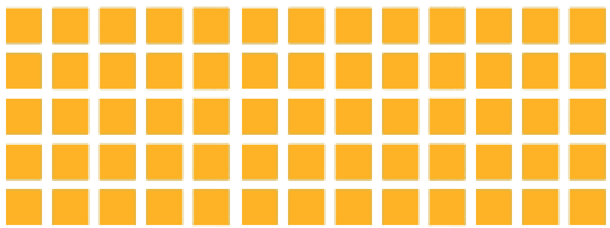

Text Mining Foundations
Individual words or higher level entities (sentences, paragraphs, documents)?
DOCUMENTS
WORDS
Sort all documents into categories or search for specific documents?
Automatically identify specific facts or gain overall understanding?
SEARCH
SORT
SPECIFIC FACTS
UNDERSTANDING
Information 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

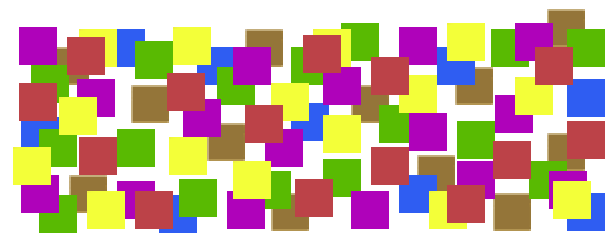
structured data
~ well-defined variables (quantitative)



unstructured data
~ text-heavy data (qualitative)

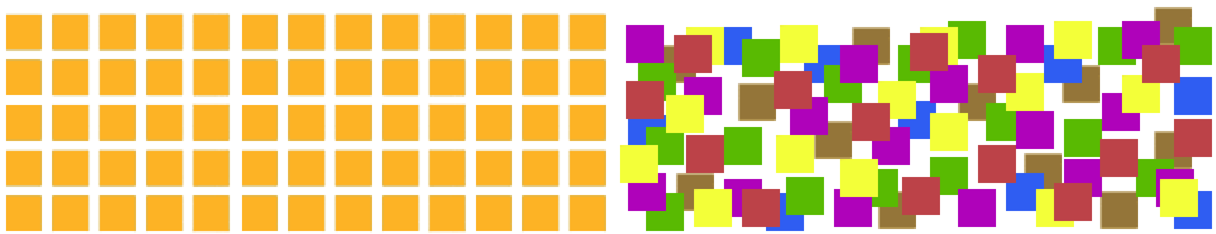
if you got this





why would you ever ... ?

if you got this



why would you ever ... ?

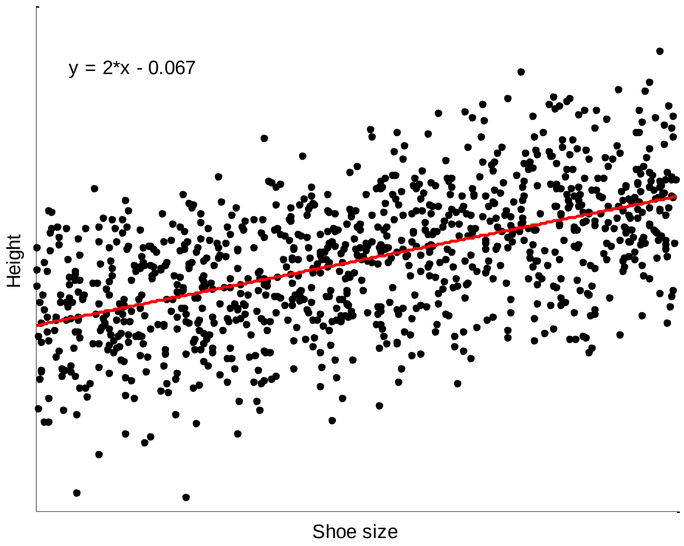
because

1: sheer mass

2: knowledge domains (human communication)

3: domain-specific specializations ~ humanities (language & culture)

a word on MODELING



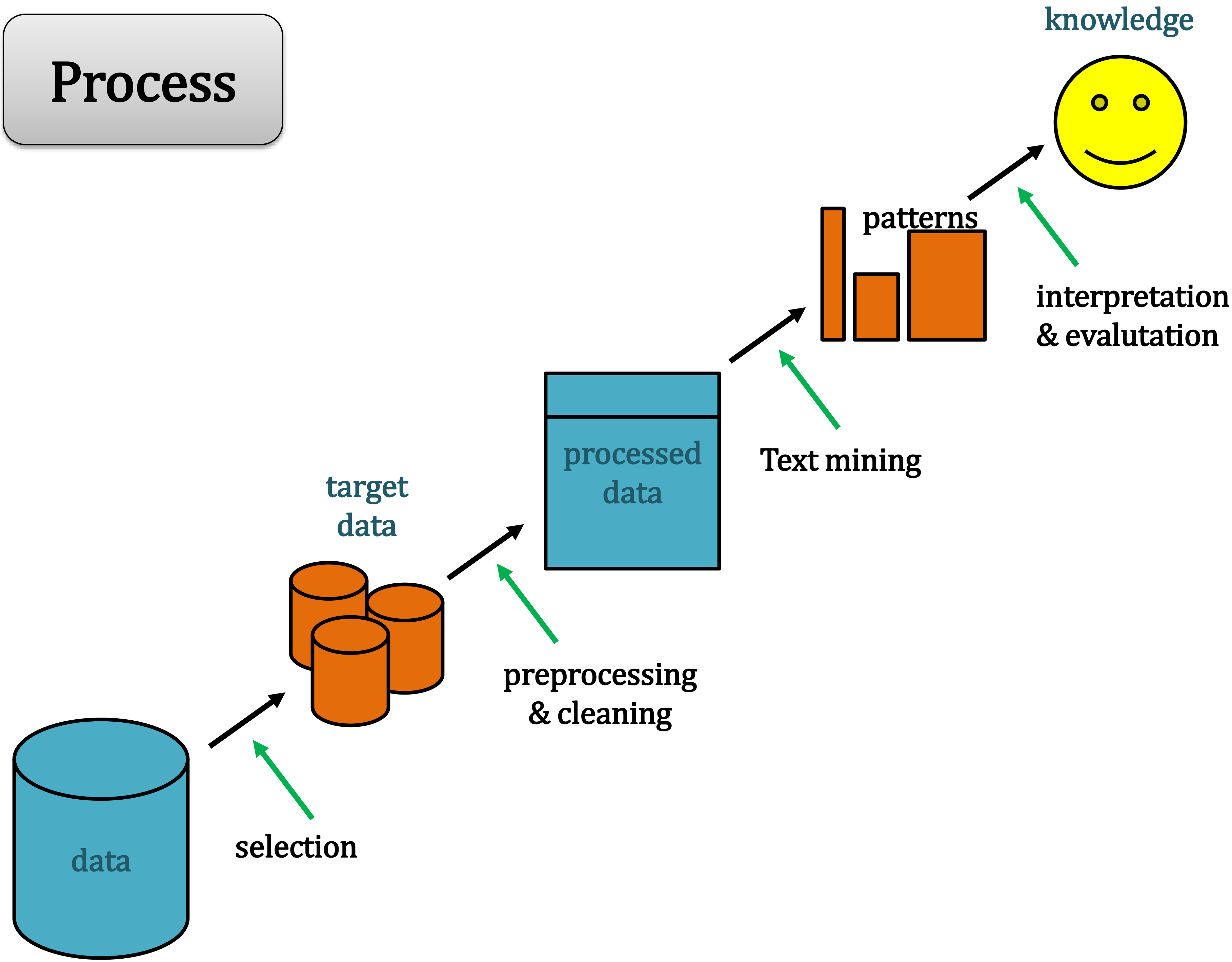
Models are mathematical simplifications

- they are abstractions
- they distinguish elements and make explicit the relations between them
- they make a range of explicit assumptions
- they are (by necessity) WRONG, but useful

Modeling of language

- Based on mathematical models of language
- Probabilistic or geometric
- Do not explain in themselves, but need to be interpreted
- Evaluated according to their ability to support inferences, insights and generate new interpretations

Process

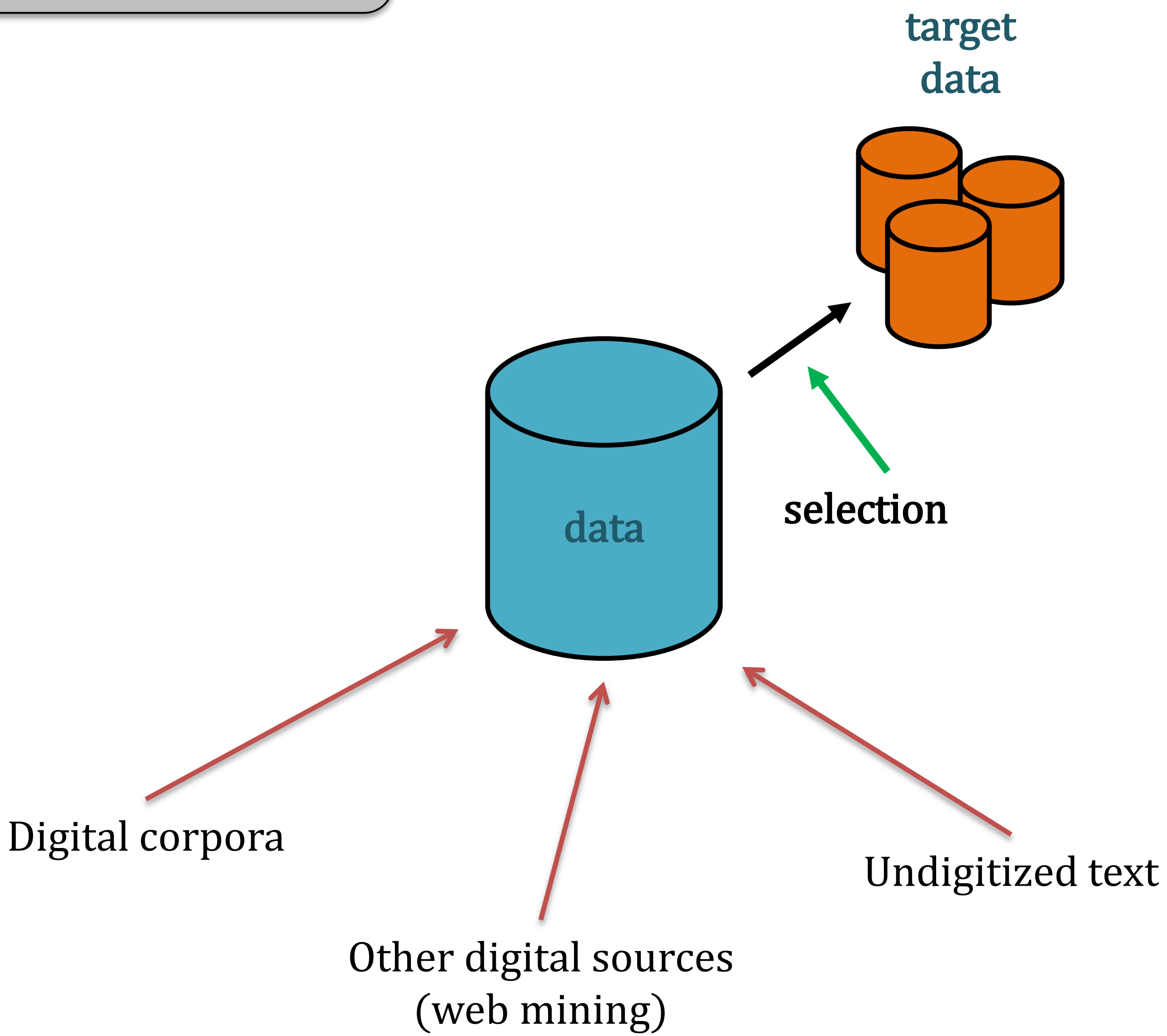


knowledge
patterns
interpretation
& evalutation
processed
data
Text mining
target
data
preprocessing
& cleaning
data
selection


OUR CONSULTANT HAS BEEN MINING DATA ALL DAY.
THE RESULTS ARE QUITE SHOCKING.
www.dilbert.com
ACCORDING TO THE DATA, SALES ARE ALWAYS HIGHEST WHEN I DO THIS...

1. Collecting data
2. Preprocessing
3. Analysis
4. Evaluation

Collecting data



Digital corpora

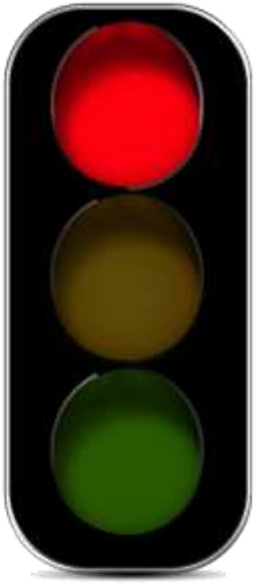
- Ideally available in XML
- High quality of text and metadata
    - Text Encoding Initiative
- TXT is fine (copy-paste)
- Beware of licensing agreements!
- e.g., Project Gutenberg, Internet Sacred Text Archive,

Other digital source

- Some texts can be obtained through an API
- Others can be scraped
- Might need custom programming
- Little or no metadata
- Beware of website restrictions!

Undigitized texts

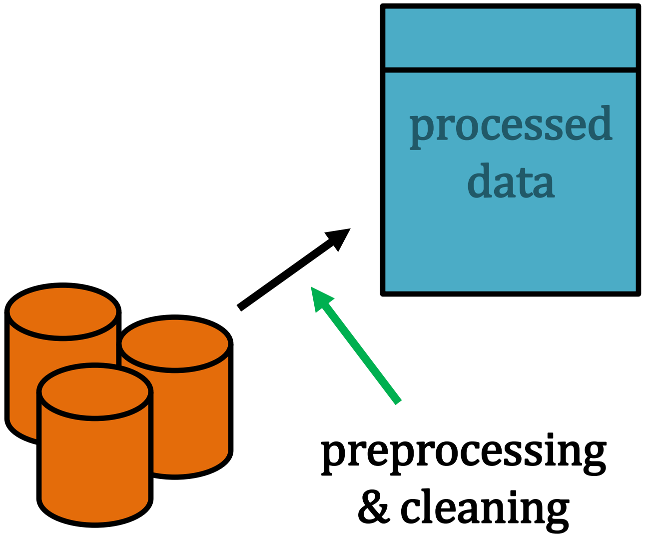
- Scanned and subjected to Optical Character Recognition (OCR)
- Costly
- Error prone (Dirty OCR)
- Add metadata
- Sometimes the only option → team up




Now,
let's analyze some
text


Welcome to the
purgatory of
preprocessing

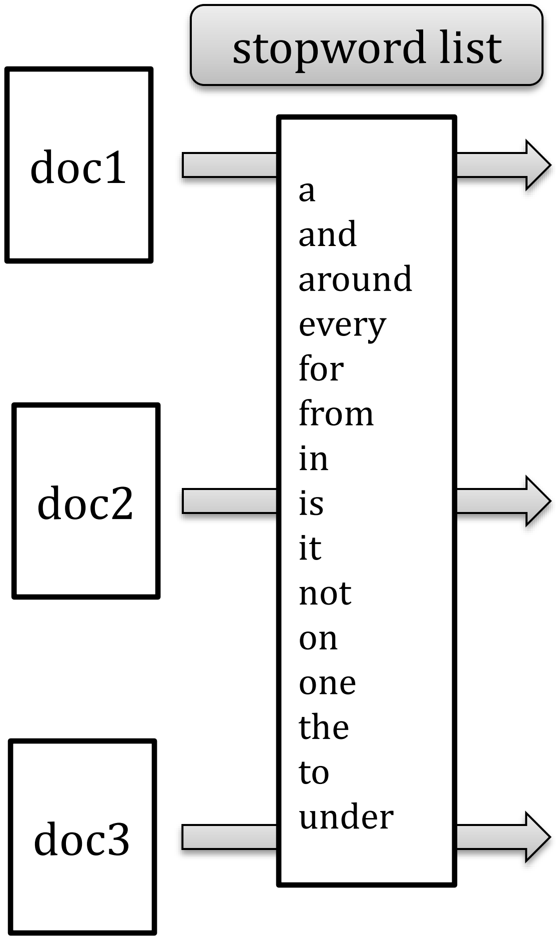
Preprocessing



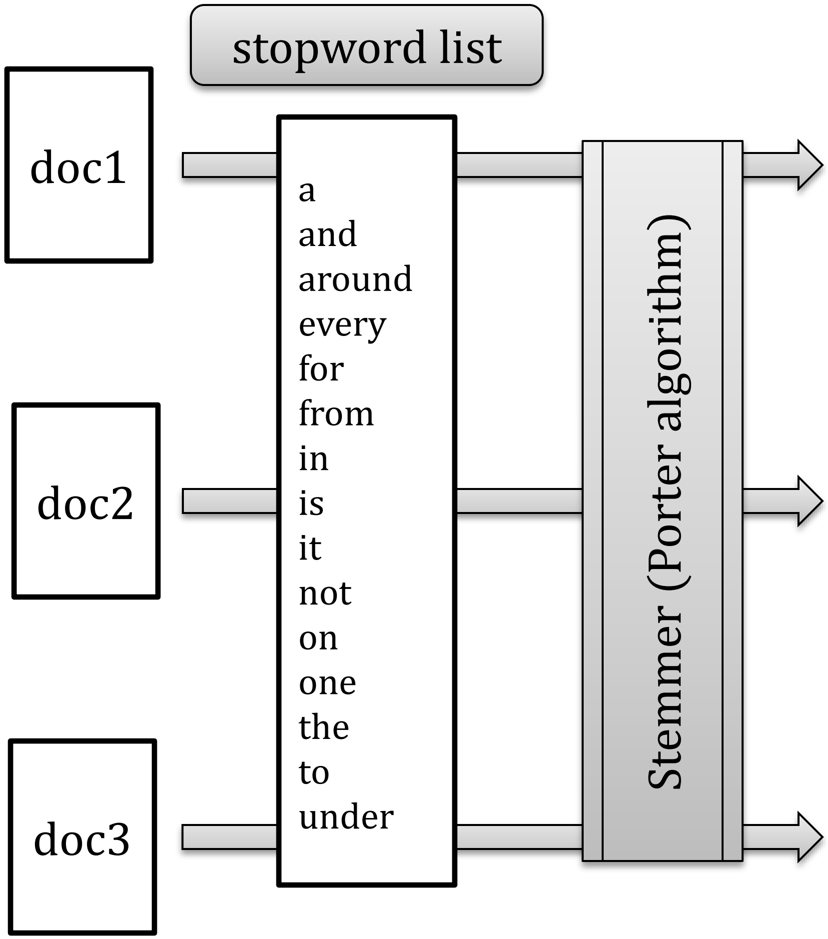
- You will spend most of your text mining career preprocessing your texts
- OCR errors
- Words broken over across lines
- Running headers and footers
- Breaking into paragraphs, sentences &c
- Tokenization
- Filtering
- Tagging
- ....

"1:1 The book of the generation of Jesus Christ, the son of David,the son of Abraham."
1:2 Abraham begat Isaac; and Isaac begat Jacob; and Jacob begat Judas and his brethren;
1:3 And Judas begat Phares and Zara of Thamar; and Phares begat Esrom; and Esrom begat Aram;
1:4 And Aram begat Aminadab; and Aminadab begat Naasson; and Naasson begat Salmon;
1:5 And Salmon begat Booz of Rachab; and Booz begat Obed of Ruth; and Obed begat Jesse;
1:6 And Jesse begat David the king; and David the king begat Solomon of her [that had been the wife] of Urias;
1:7 And Solomon begat Roboam; and Roboam begat Abia; and Abia begat Asa;
1:8 And Asa begat Josaphat; and Josaphat begat Joram; and Joram begat Ozias;
1:9 And Ozias begat Joatham; and Joatham begat Achaz; and Achaz begat Ezekias;
1:10 And Ezekias begat Manasses; and Manasses begat Amon; and Amon begat Josias;
"1:11 And Josias begat Jechonias and his brethren, about the time they were carried away to Babylon:"
"1:12 And after they were brought to Babylon, Jechonias begat Salathiel; and Salathiel begat Zorobabel;"
1:13 And Zorobabel begat Abiud; and Abiud begat Eliakim; and Eliakim begat Azor;
1:14 And Azor begat Sadoc; and Sadoc begat Achim; and Achim begat Eliud;
1:15 And Eliud begat Eleazar; and Eleazar begat than; and than begat Jacob;
"1:16 And Jacob begat Joseph the husband of Mary, of whom was born Jesus, who is called Christ."
1:17 So all the generations from Abraham to David are fourteen generations; and from David until the carrying away into Babylon are fourteen generations; and from the carrying away into Babylon unto Christ are fourteen generations.
"1:18 Now the birth of Jesus Christ was on this wise: When as his mother Mary was espoused to Joseph, before they came together, she was found with child of the Holy Ghost."
"1:19 Then Joseph her husband, being a just man, and not willing to make her a publick example, was minded to put her away privily."
"1:20 But while he thought on these things, behold, the angel of the LORD appeared unto him in a dream, saying, Joseph, thou son of David, fear not to take unto thee Mary thy wife: for that which is conceived in her is of the Holy Ghost."
"1:21 And she shall bring forth a son, and thou shalt call his name JESUS: for he shall save his people from their sins."
"1:22 Now all this was done, that it might be fulfilled which was spoken of the Lord by the prophet, saying,"
"1:23 Behold, a virgin shall be with child, and shall bring forth a son, and they shall call his name Emmanuel, which being interpreted is, God with us."
"1:24 Then Joseph being raised from sleep did as the angel of the Lord had bidden him, and took unto him his wife:"
1:25 And knew her not till she had brought forth her firstborn son: and he called his name JESUS.

mysteries	1.0
naked	3.0
name	21.0
named	2.0
names	1.0
narrow	1.0
nation	3.0
nations	4.0
nay	1.0
near	2.0
neck	1.0
need	7.0
needle	1.0
needs	1.0
neglect	2.0
neighbour	3.0
neither	26.0
nests	1.0
net	2.0
nets	2.0
never	6.0
nevertheless	3.0



mysteries	1.0
naked	3.0
name	21.0
named	2.0
names	1.0
narrow	1.0
nation	3.0
nations	4.0
nay	1.0
~~near~~	2.0
neck	1.0
~~need~~	7.0
needle	1.0
~~needs~~	1.0
neglect	2.0
neighbour	3.0
~~neither~~	26.0
nests	1.0
net	2.0
nets	2.0
~~never~~	6.0
~~nevertheless~~	3.0



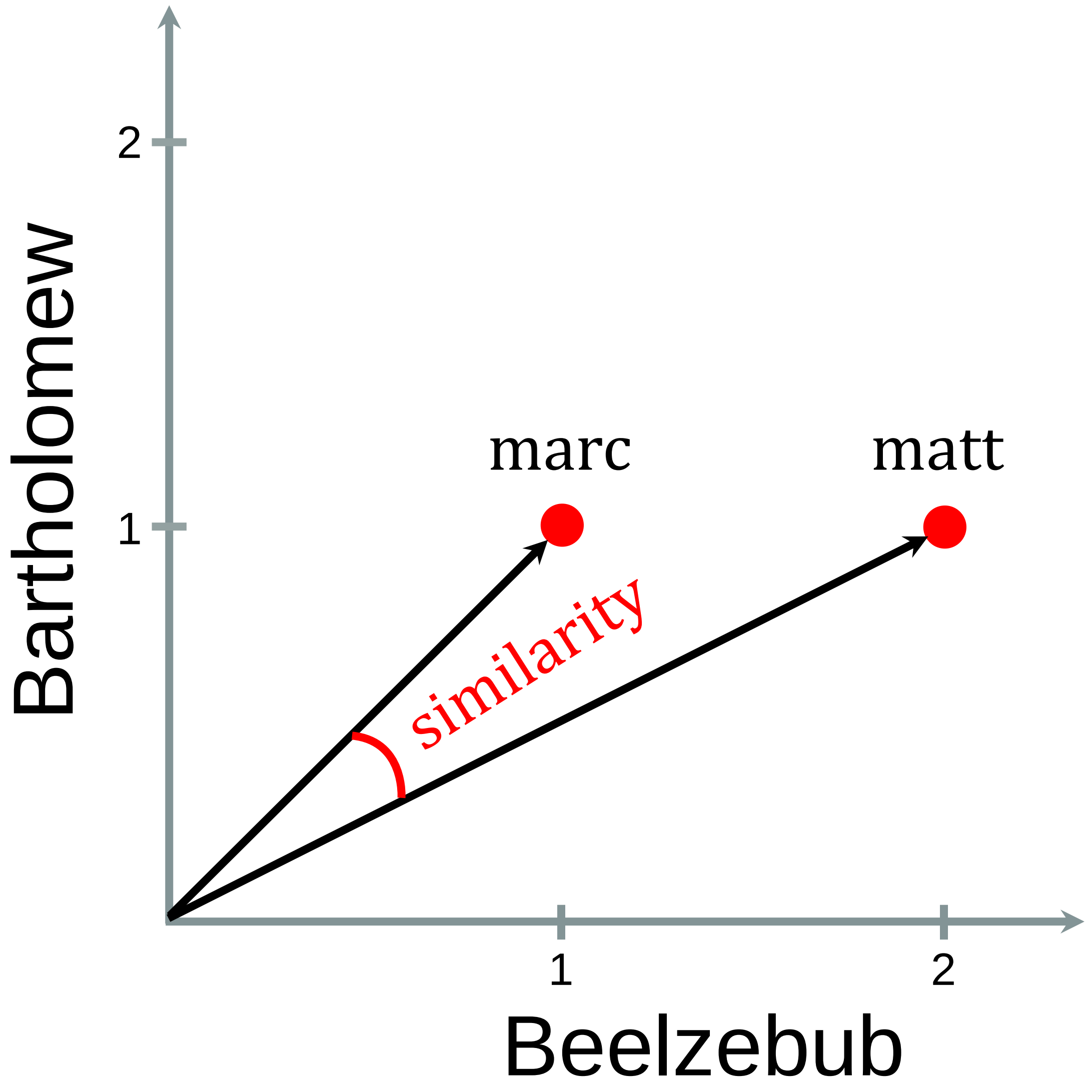
mysteri	1.0
naasson	2.0
nai	3.0
nake	4.0
name	24.0
narrow	1.0
nation	7.0
nazaren	1.0
nazareth	4.0
neck	1.0
needl	1.0
neglect	2.0
neighbour	3.0
nephthalim	2.0
nest	1.0
net	4.0

doc	Augustus	Avenge	Azor	Babylon	Baptist	Barabbas	Barachias	Barjona	Bartholomew	Bartimaeus	Beelzebub	Behold	Believe
john.txt	.0	.0	.0	.0	.0	2.0	.0	.0	.0	.0	.0	10.0	1.0
luke.txt	1.0	1.0	.0	.0	4.0	1.0	.0	.0	1.0	.0	3.0	14.0	.0
marc.txt	.0	.0	.0	.0	4.0	3.0	.0	.0	1.0	1.0	1.0	7.0	.0
matt.txt	.0	.0	2.0	4.0	7.0	5.0	1.0	1.0	1.0	.0	2.0	18.0	1.0

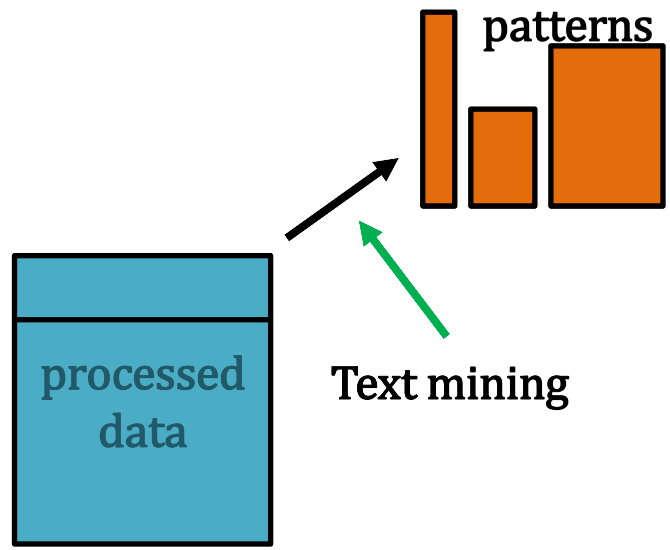
Behold: a vector representation of text

How similar are Matt and Marc?

doc	Augustus	Avenge	Azor	Babylon	Baptist	Barabbas	Barachias	Barjona	Bartholomew	Bartimaeus	Beelzebub	Behold	Believe
john.txt	.0	.0	.0	.0	.0	2.0	.0	.0	.0	.0	.0	10.0	1.0
luke.txt	1.0	1.0	.0	.0	4.0	1.0	.0	.0	1.0	.0	3.0	14.0	.0
marc.txt	.0	.0	.0	.0	4.0	3.0	.0	.0	1.0	1.0	1.0	7.0	.0
matt.txt	.0	.0	2.0	4.0	7.0	5.0	1.0	1.0	1.0	.0	2.0	18.0	1.0


2
1
Bartholomew
marc
matt
similarity
1
2
Beelzebub

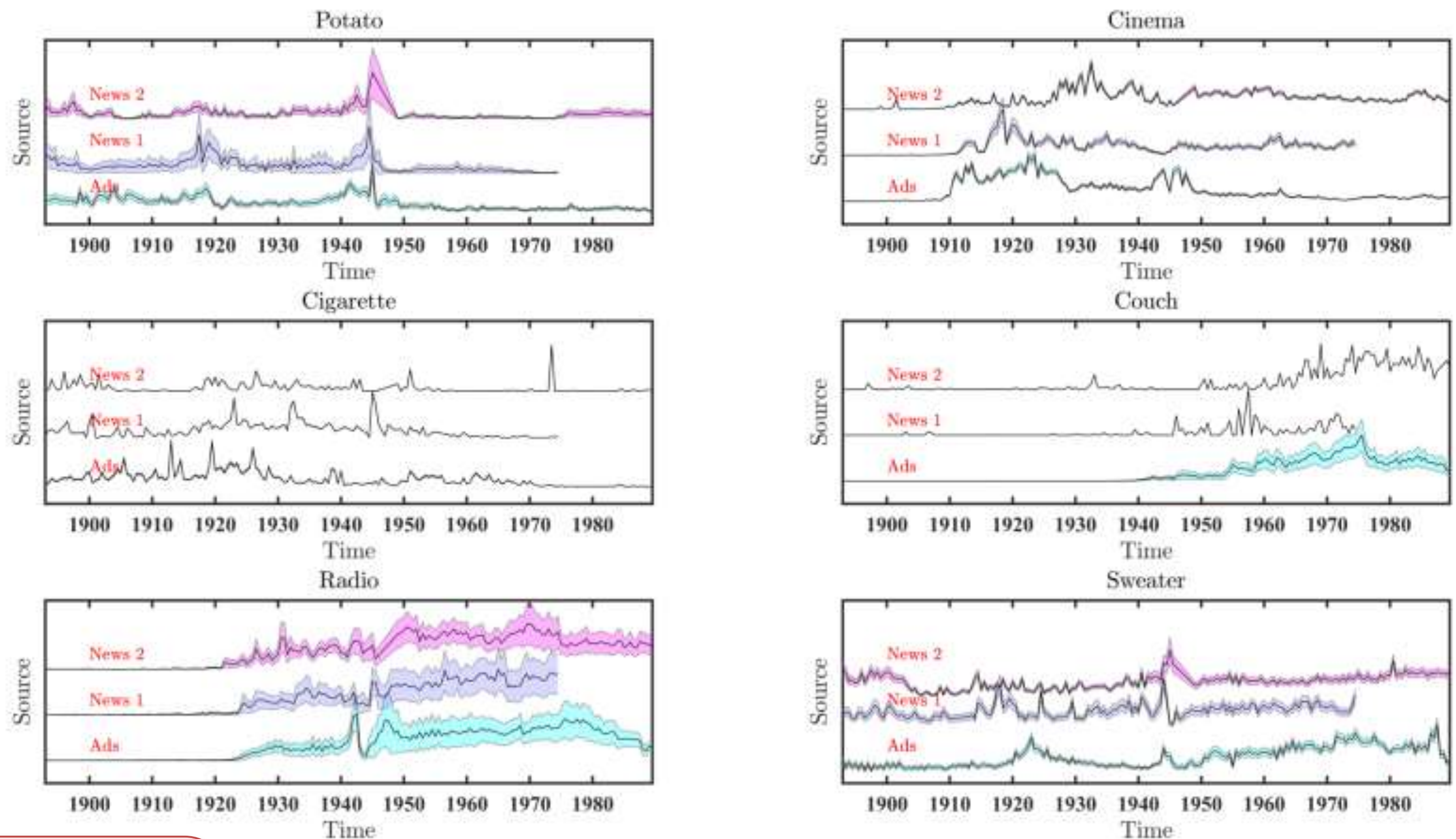
Analysis



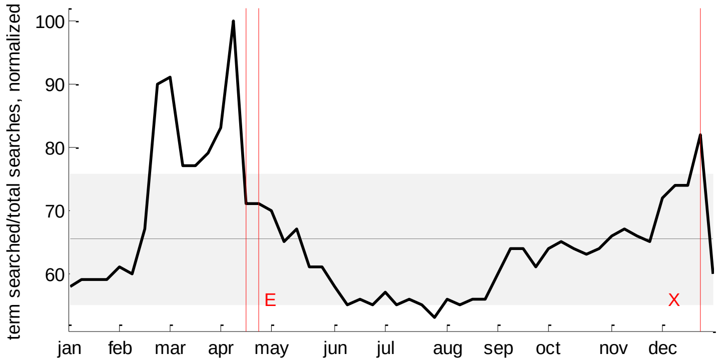
- Reading
- Counting
- Human coding
- Dictionary
- Unsupervised learning
- Supervised learning

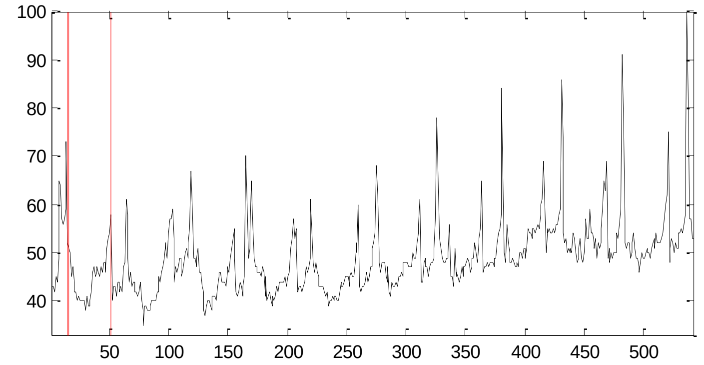
- Reading
- **Counting**
- Human coding
- Dictionary
- **Unsupervised learning**
- **Supervised learning**

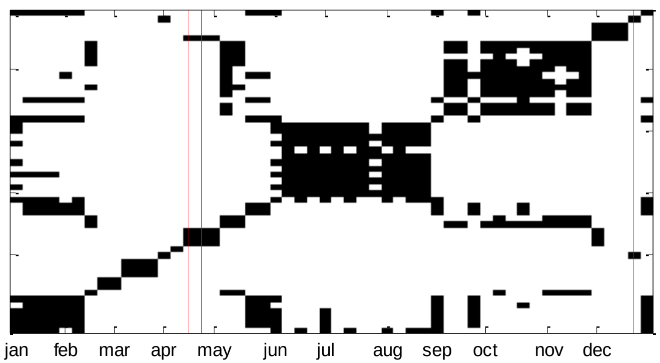
Counting

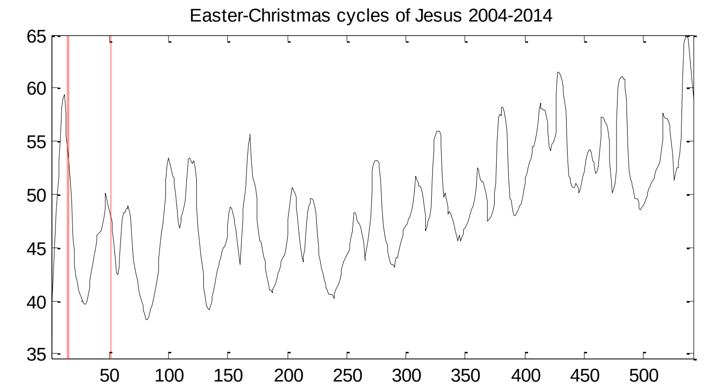


Easy to
- compute
- replicate



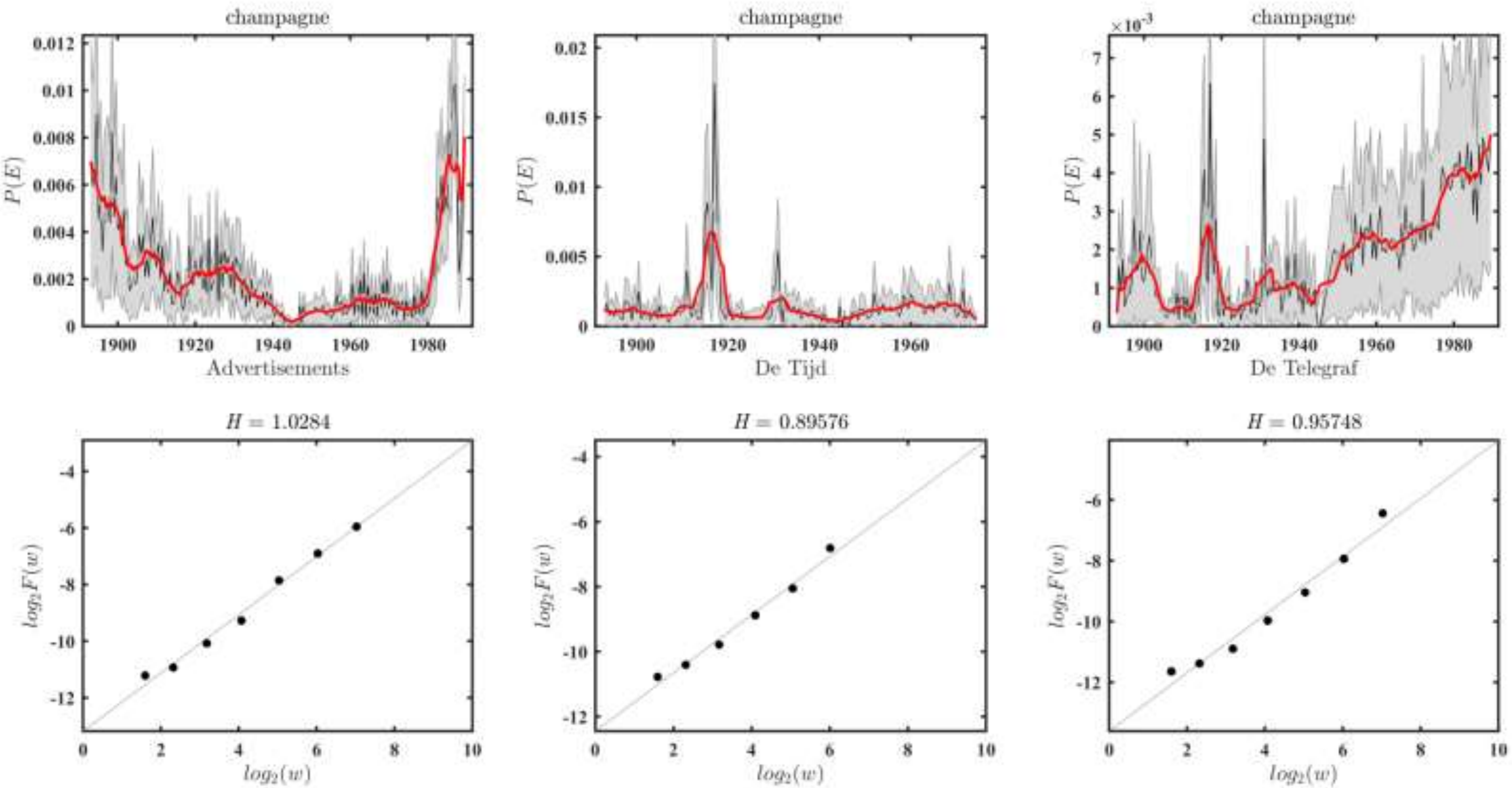






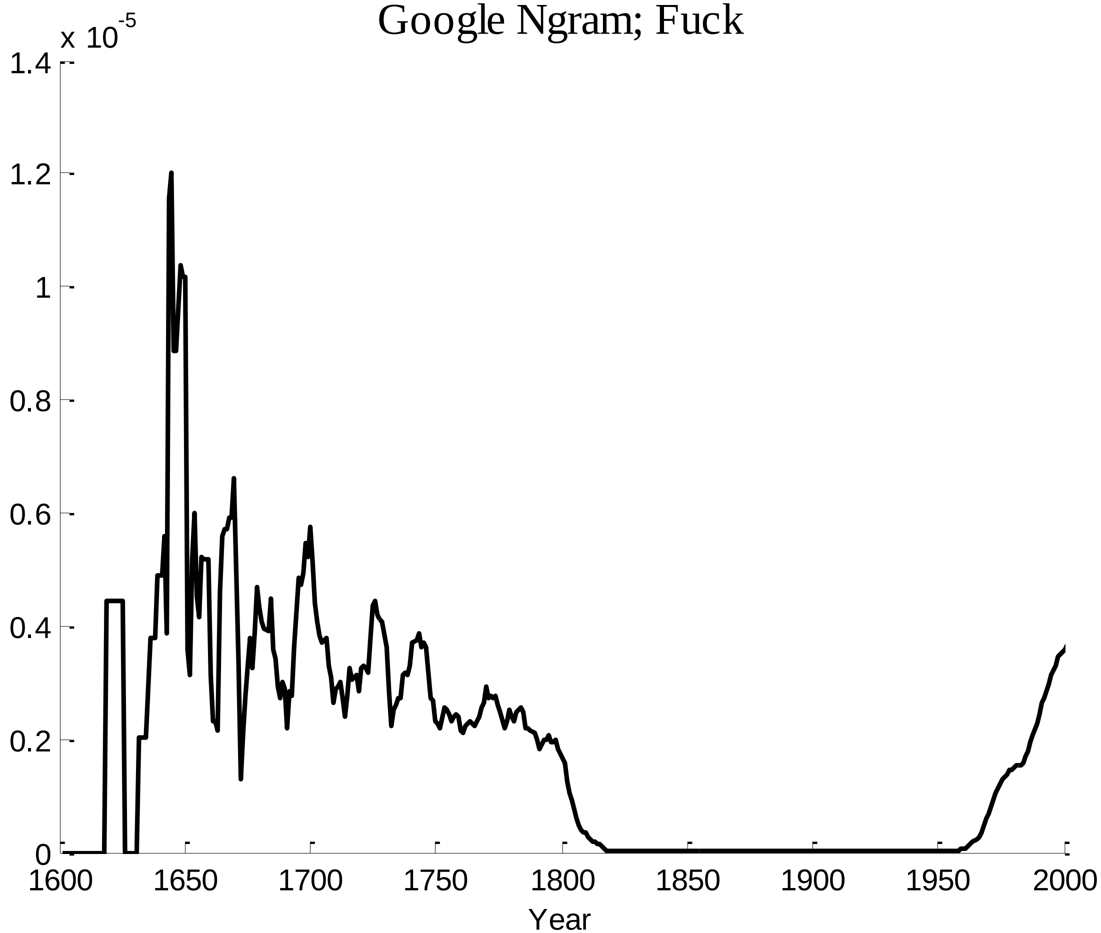
Comparison requires metadata

- language
- time
- location
- ...



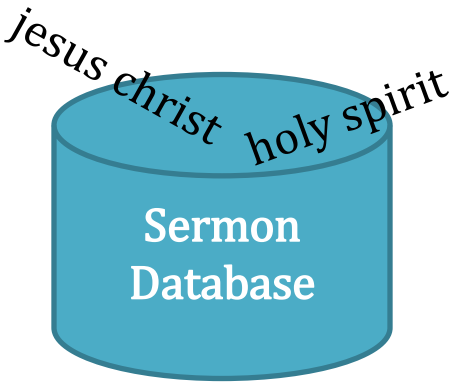
Word use is
- ambiguous
- spelling can vary

Google Ngram; Fuck



guilt for which God and Man, yea and themſelves alſo ſhall equally accuſe them, and to keep their expences within ſuch limits, that as Bees ſuck, but do not violate or deface the flowers, ſo they as joint proprietaries with the Husbands, may enjoy, but not devour and deſtroy his fortune.

Association rules



1 of 5 documents *Holy Spirit* occurs

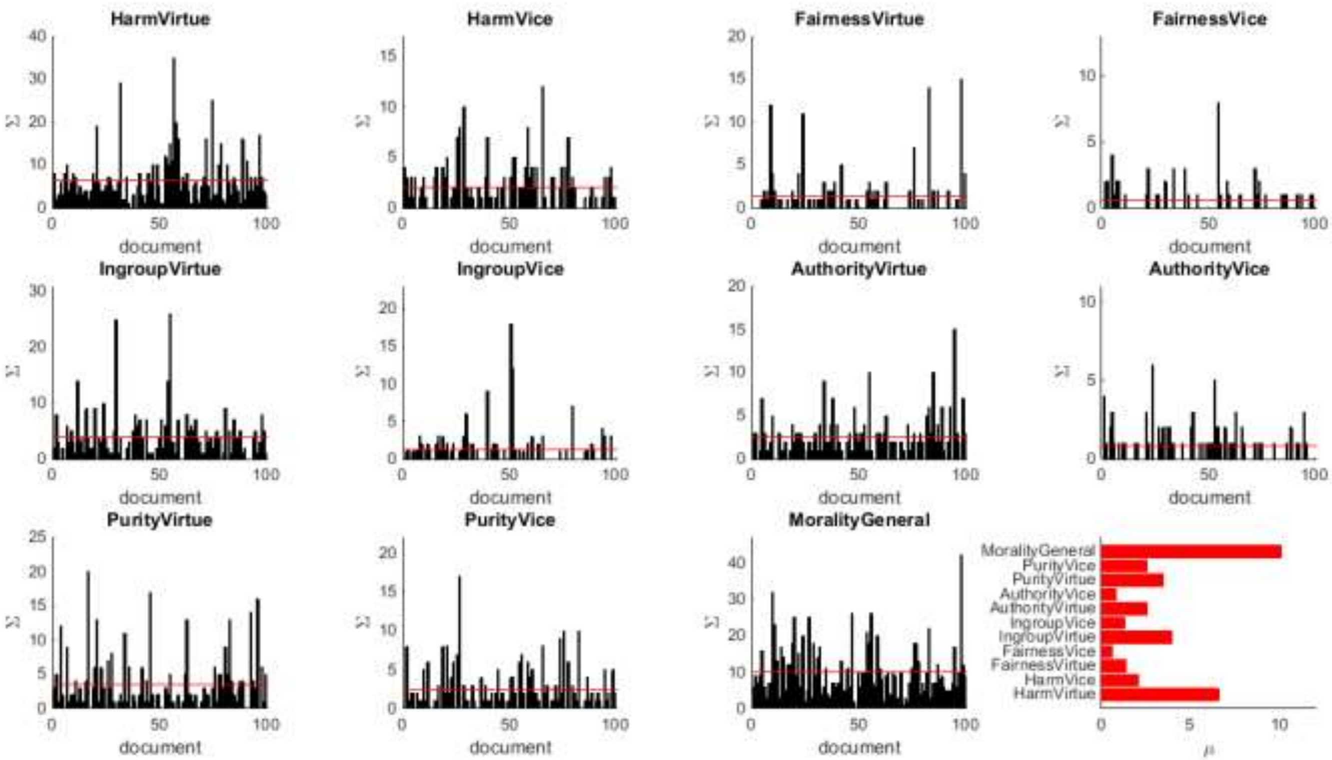
1 of 4 documents *Jesus Christ* occurs

no association
- 1 of 20 documents have occurrences of *Holly Spirit* & *Jesus Christ*
- 1/5 x 1/4 = 1/20

reality
- *Holly Spirit* & *Jesus Christ* > 5%
- correlated

association mining
- looking of associations that occur above chance level
- Association: attribute/value pair
- e.g., *Holly Spirit* = true & *Jesus Christ* = true
- counting the association rules

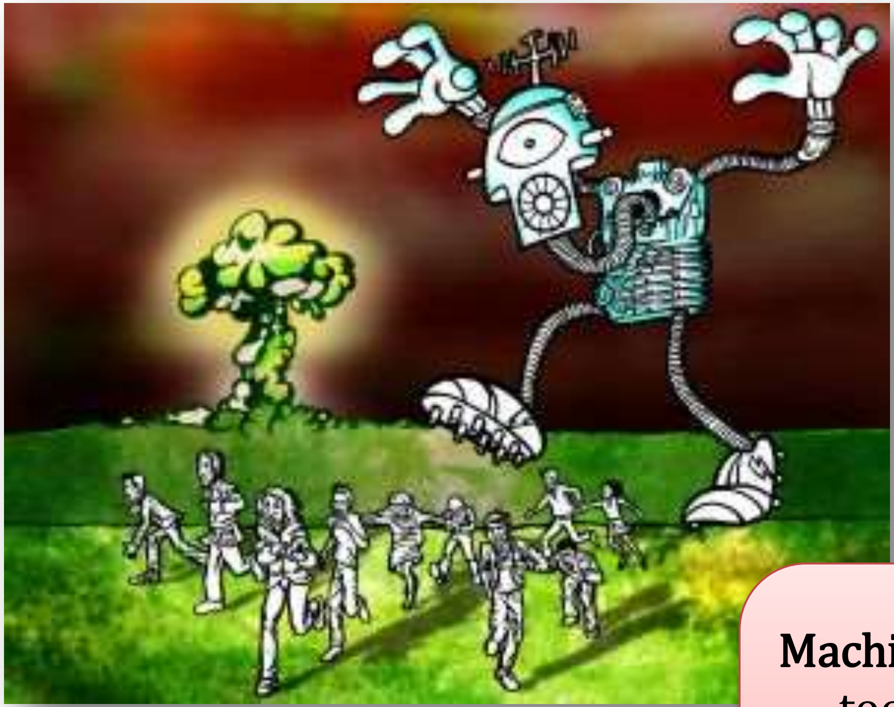
Sentiment analysis



Sentiment analysis

- Dictionaries: list of words that are compiled for specific categories
  - e.g., positive and negative affective terms
  - Custom-built or reused
- machine learning

machine learning



Machine Learning/Mlear/ML

- tools that use computers to transform data into actionable knowledge
- making sense of complex data

'a machine is able to learn if it can take experience and utilize it such that its performance improves up on similar experiences in the future'

3-step process

- **Data input** utilizes observation, memory storage, and recall to provide a factual basis for further reasoning
- **Abstraction** involves translation of data into broader representations
- **Generalization** uses abstracted data to for a basis for action

Unsupervised (machine) learning

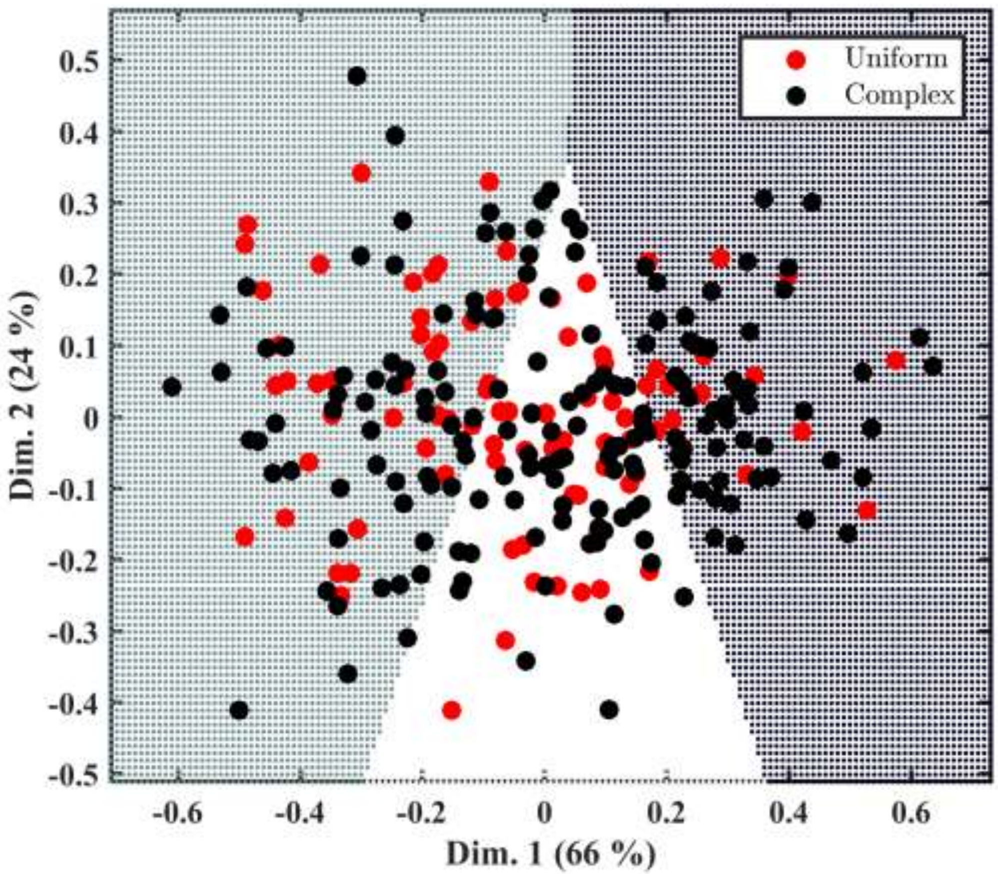
When?
- Don’t know the categories
- Want to discover new categories
- Exploratory

Unsupervised learning
- Let the machine explore and find possible categorizations for you
- Clustering, cluster analysis

Can the robot do the thinking?
- Yes, no manual coding pre-analysis
- But, evaluation of suggested categories is needed

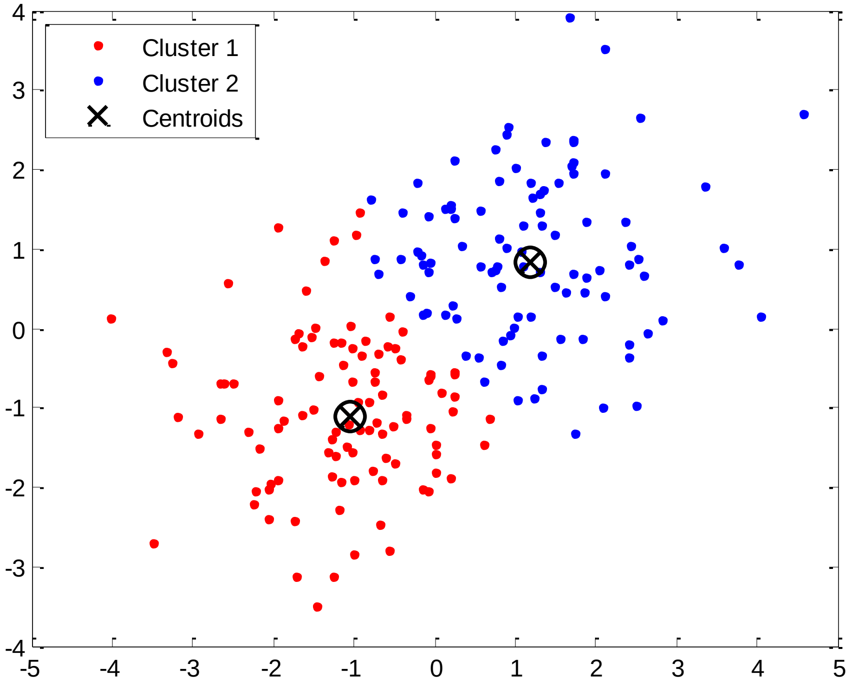
Single membership clustering

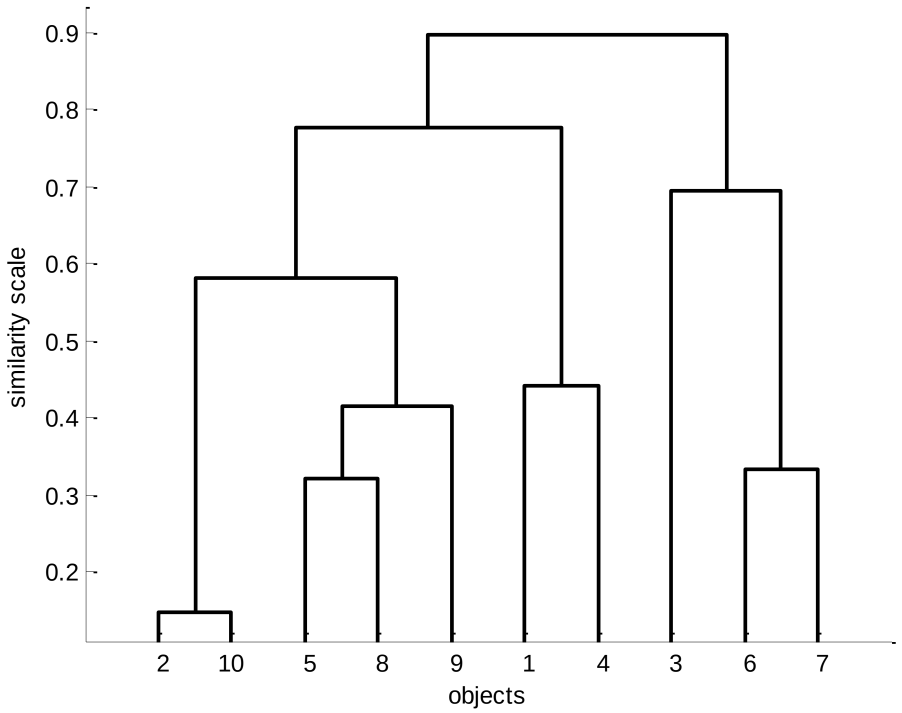
- Define similarity measure
- Define measure of how good a cluster is
- Define a process for optimization of overall goodness

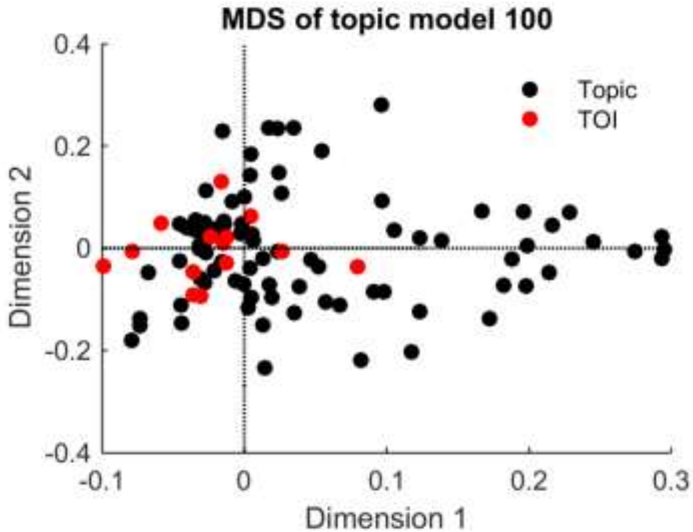


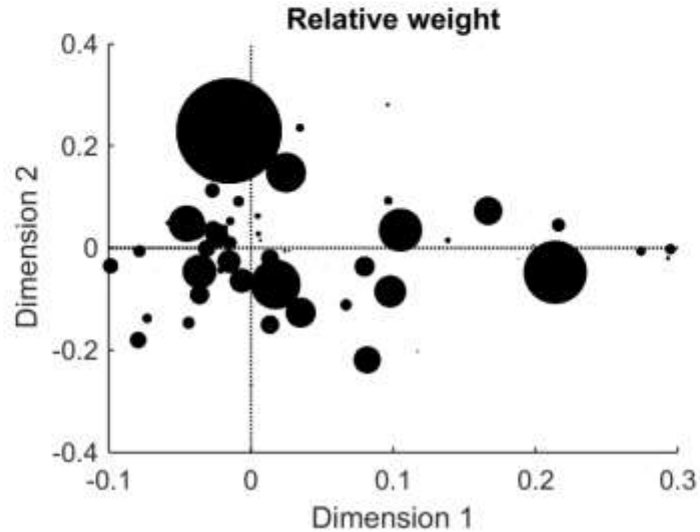
exclusive clusters

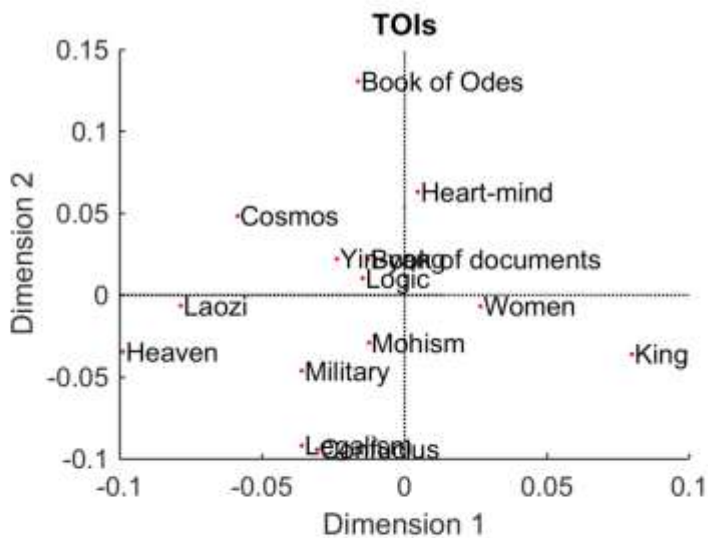
embedded clusters











Mixed membership clustering

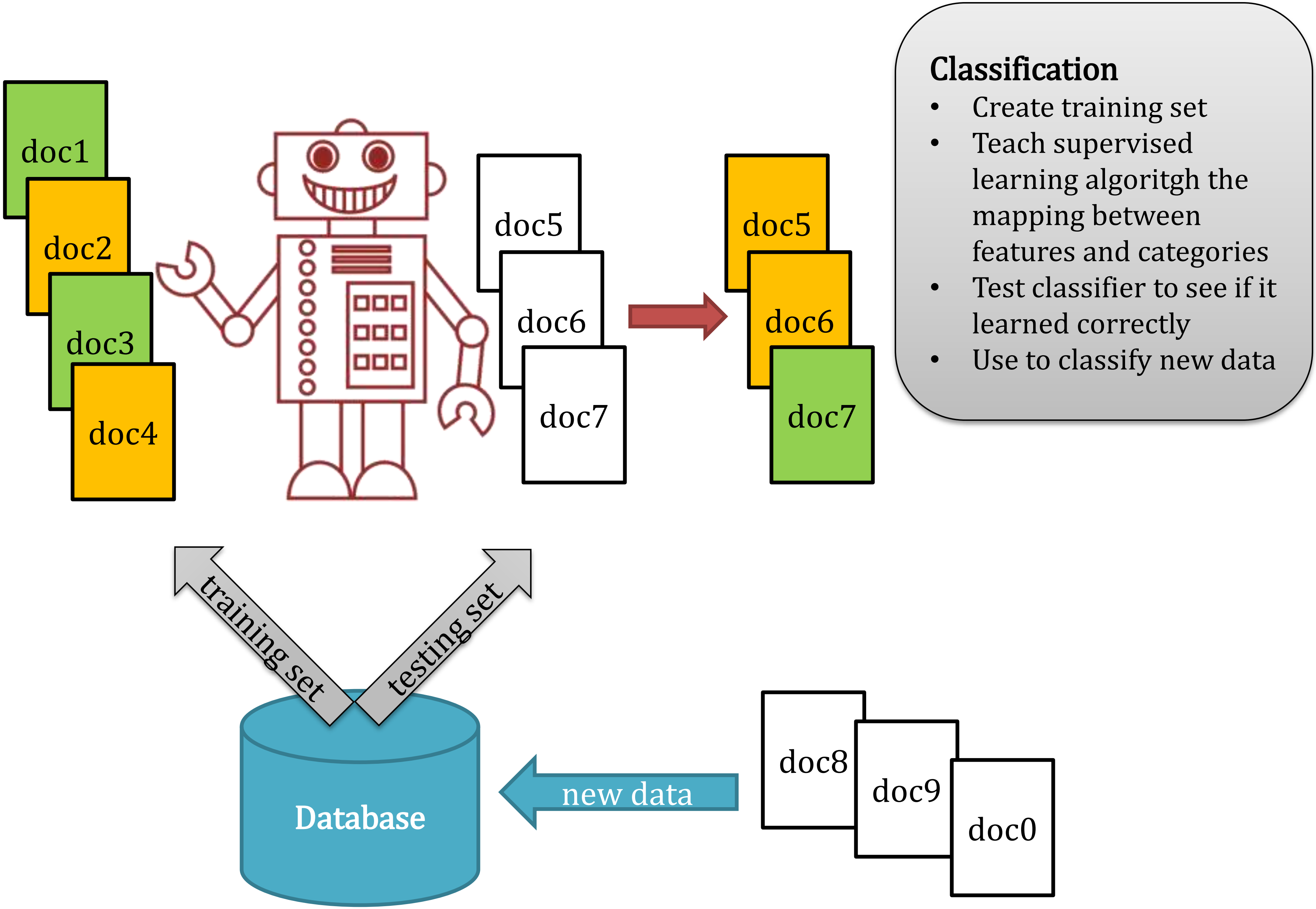
- Topic modeling
- Each document is a mixture of topics (or categories)
- A document is a probability distribution over topics
- A topic is a probability distribution over words

Supervised (machine) learning

When?
- Know the categories
- Human coding doesn't scale
- Closer to hypothesis testing

Supervised learning
- Let the machine train on and test your categories
- Classification

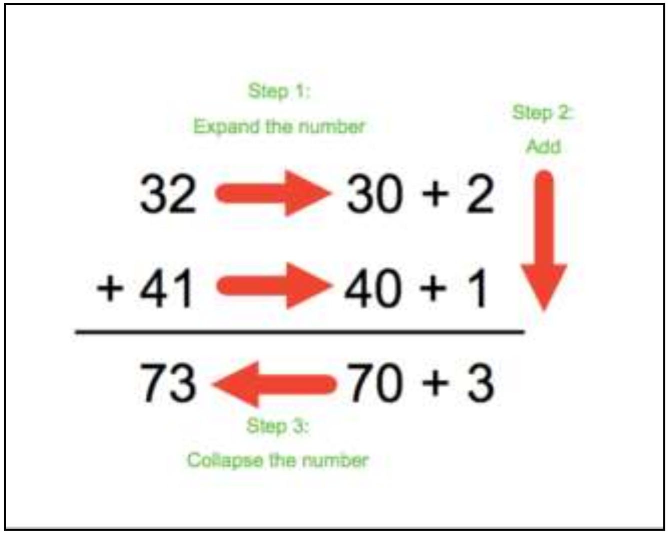
Can the robot do the thinking?
- Yes, if you do not have too many categories
- But, takes time to figure out what drives the classification



Classification
• Create training set
• Teach supervised learning algoritgh the mapping between features and categories
• Test classifier to see if it learned correctly
• Use to classify new data
doc1
doc2
doc3
doc4
doc5
doc6
doc7
doc5
doc6
doc7
training set
testing set
Database
new data
doc8
doc9
doc0

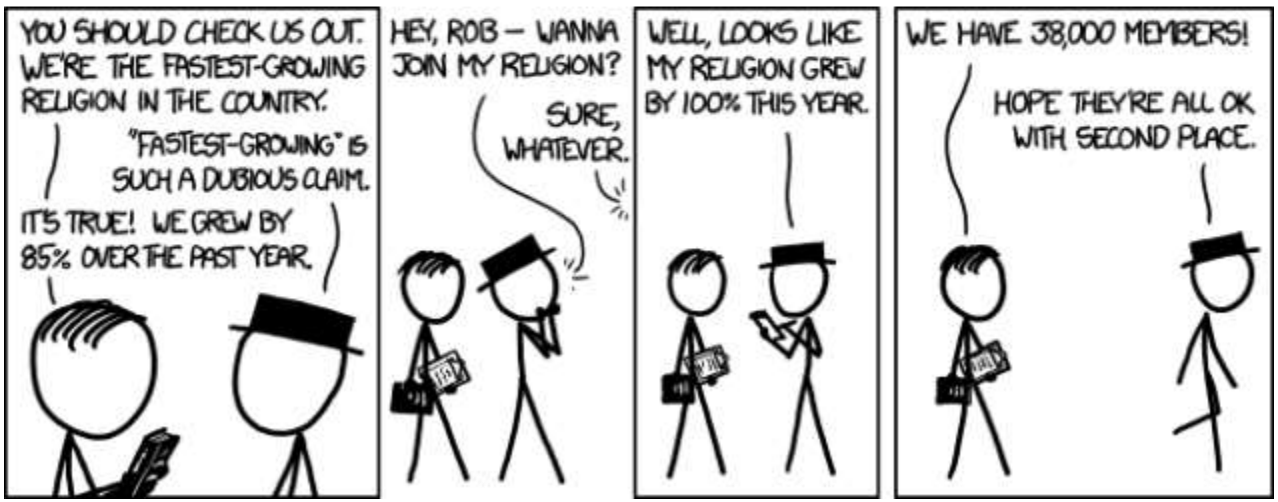
Supervised learning algorithms

- Multitude of SL algorithms
  - Naïve Bayes
  - Decision trees
  - Support vector machines
  - Neural networks
- Performance is domain and dataset specific
- Ensembles of different algorithms outperform single algorithms

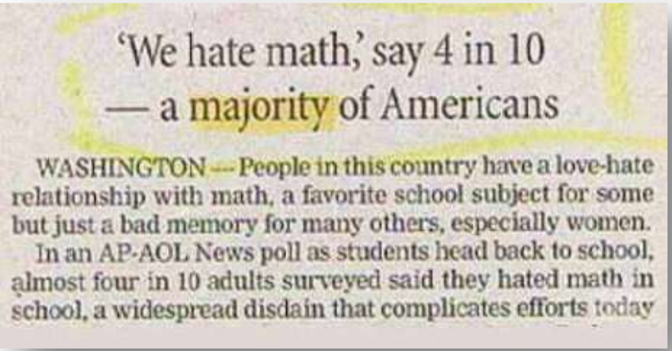


Algorithm?

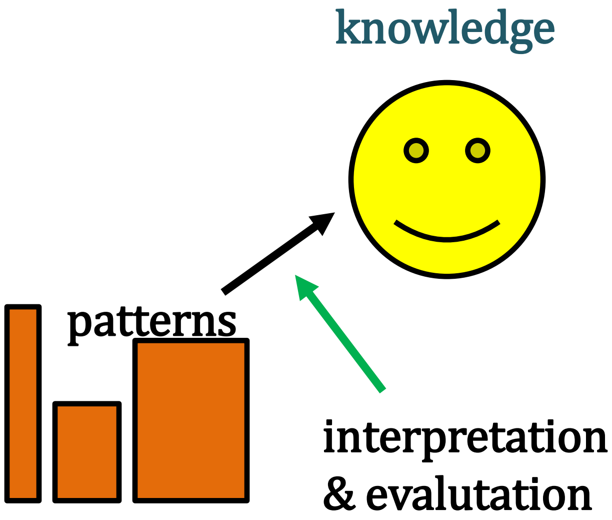
- Stepwise procedure for conducting a computation
- 'recipe' for solving a problem


YOU SHOULD CHECK US OUT. WE'RE THE FASTEST-GROWING RELIGION IN THE COUNTRY.
"FASTEST-GROWING" IS SUCH A DUBIOUS CLAIM.
IT'S TRUE! WE GREW BY 85% OVER THE PAST YEAR.
HEY, ROB — WANNA JOIN MY RELIGION?
SURE, WHATEVER.
WELL, LOOKS LIKE MY RELIGION GREW BY 100% THIS YEAR.
WE HAVE 38,000 MEMBERS!
HOPE THEY'RE ALL OK WITH SECOND PLACE.

Evaluation


'We hate math,' say 4 in 10 — a majority of Americans

WASHINGTON — People in this country have a love-hate relationship with math, a favorite school subject for some but just a bad memory for many others, especially women.
In an AP-AOL News poll as students head back to school, almost four in 10 adults surveyed said they hated math in school, a widespread disdain that complicates efforts today



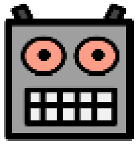
How to validate results?

- Easily lead astray by the facticity of numbers
- However, it always depends on your design
- Use common sense (+ some validation techniques)

Counting

- Text data often have errors (e.g., problems with OCR)
- Errors in metadata
- Multiple instantiations of text (copy-paste, automatic methods, & web-mining).
- Collections can be very biased samples (e.g., Google Books*)

Supervised learning



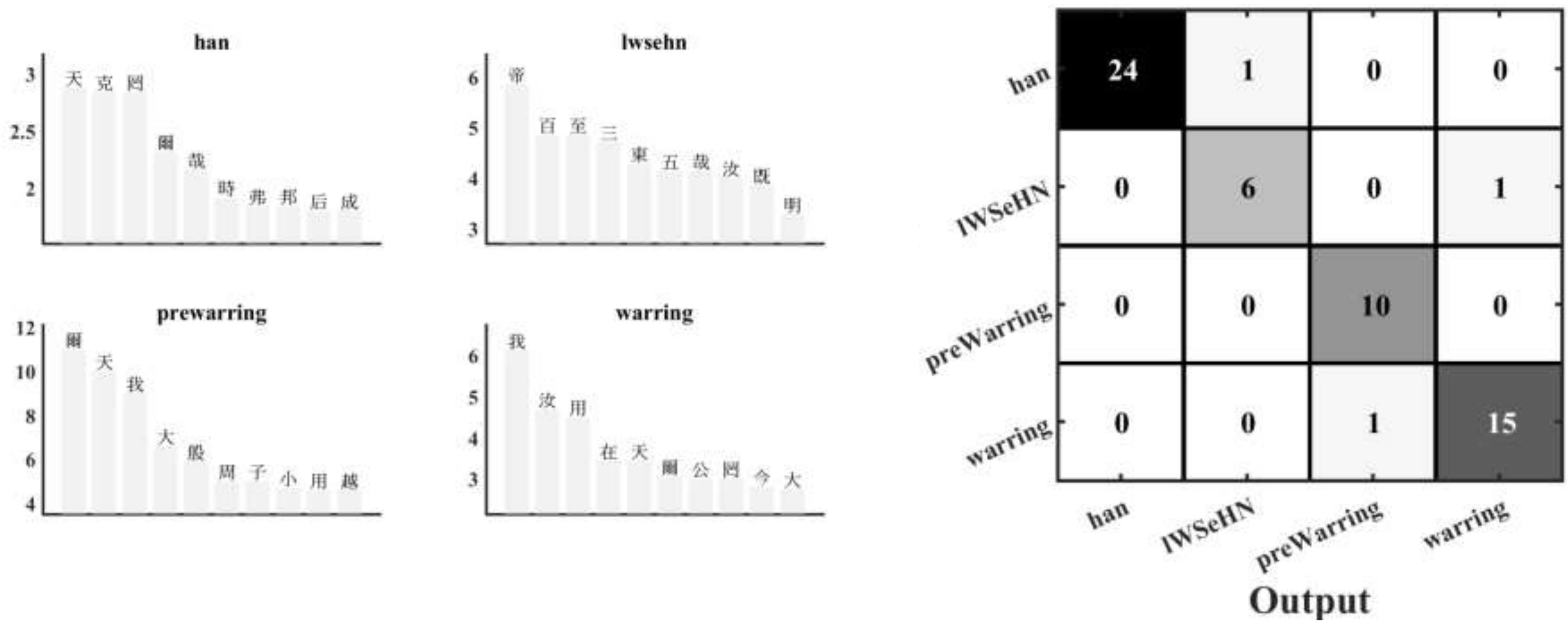
	class1	class2
class1	644	89
class2	106	661

Documents in total: $644 + 106 + 661 + 89 = 1500$

Accuracy: $\frac{(644+661)}{1500} = 0.87$

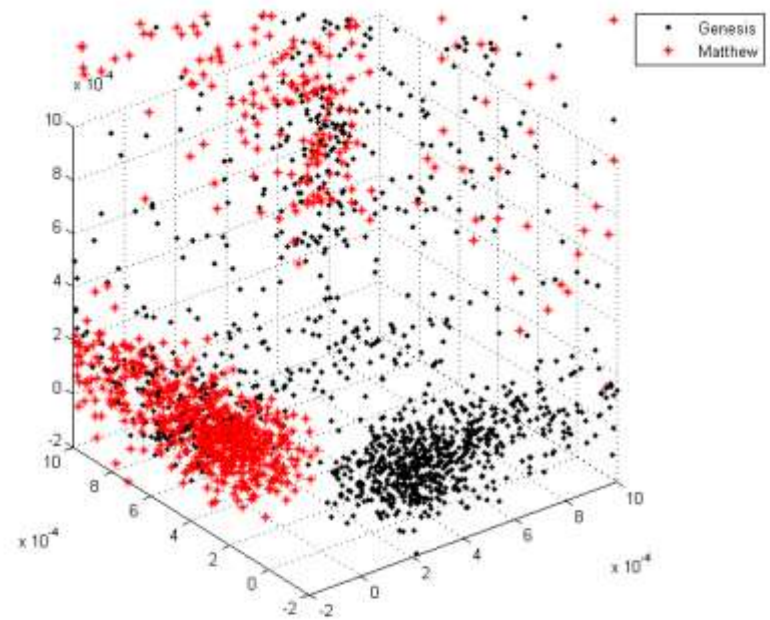
Accuracy in percent: $0.87 \times 100 = 87\%$

How many times were the model right given the population?
Proportion of correctly classified verses

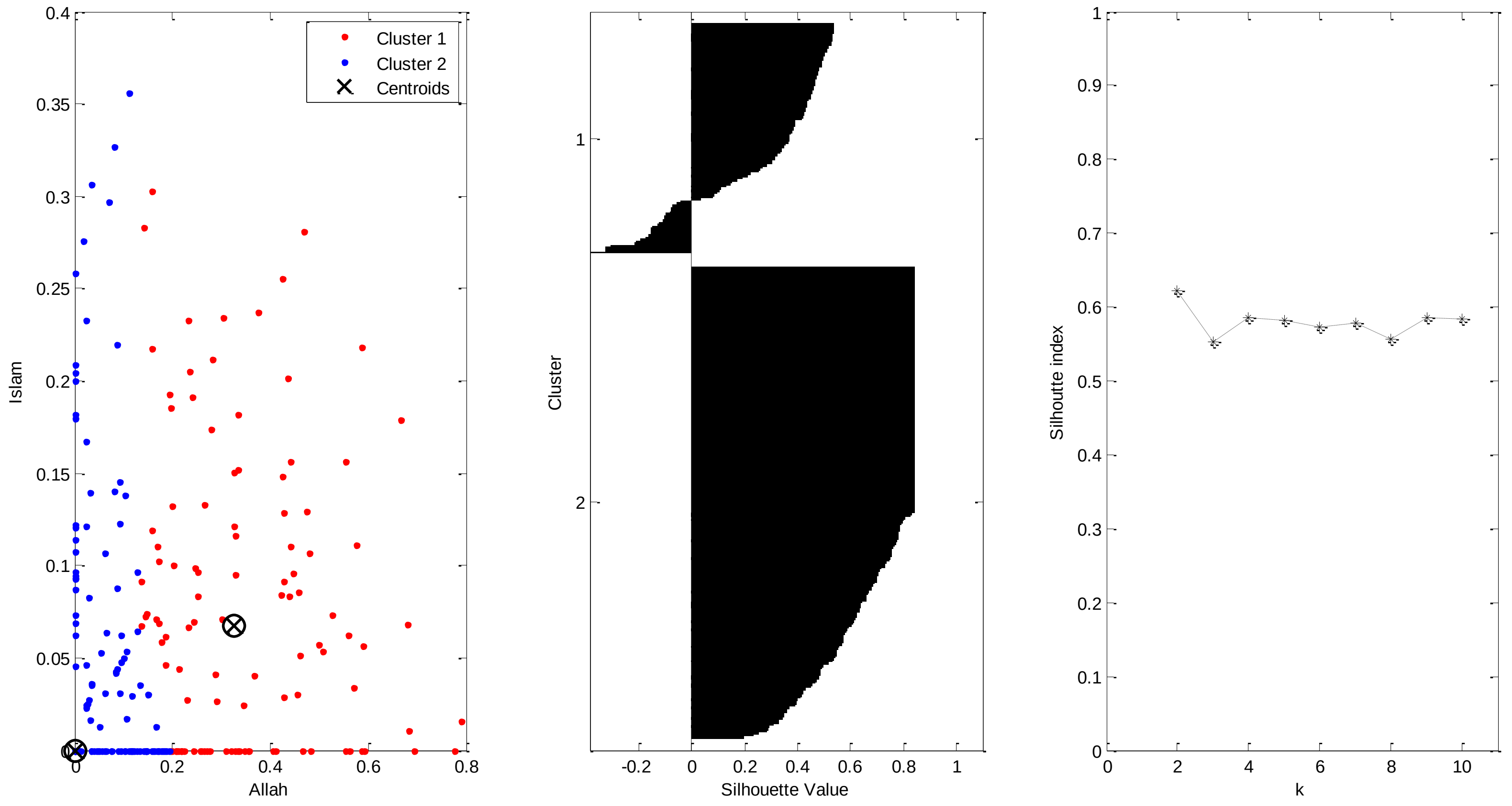

han
天 克 罔 爾 哉 時 弗 邦 后 成
lwsehn
帝 百 至 三 東 五 哉 汝 既 明
prewarring
爾 天 我 大 殷 周 子 小 用 越
warring
我 汝 用 在 天 爾 公 罔 今 大
han
lWSeHN
preWarring
warring
24 1 0 0
0 6 0 1
0 0 10 0
0 0 1 15
Output

Unsupervised learning

- Compare categorizations to existing categorization schemes (natural or manual)
- Match categories to text-external factors (e.g., author data or context)
- Test through supervised learning – convert clusters to coding sheme



validation of clusters



THANK YOU

Kristoffer L. Nielbo
kln@cas.au.dk