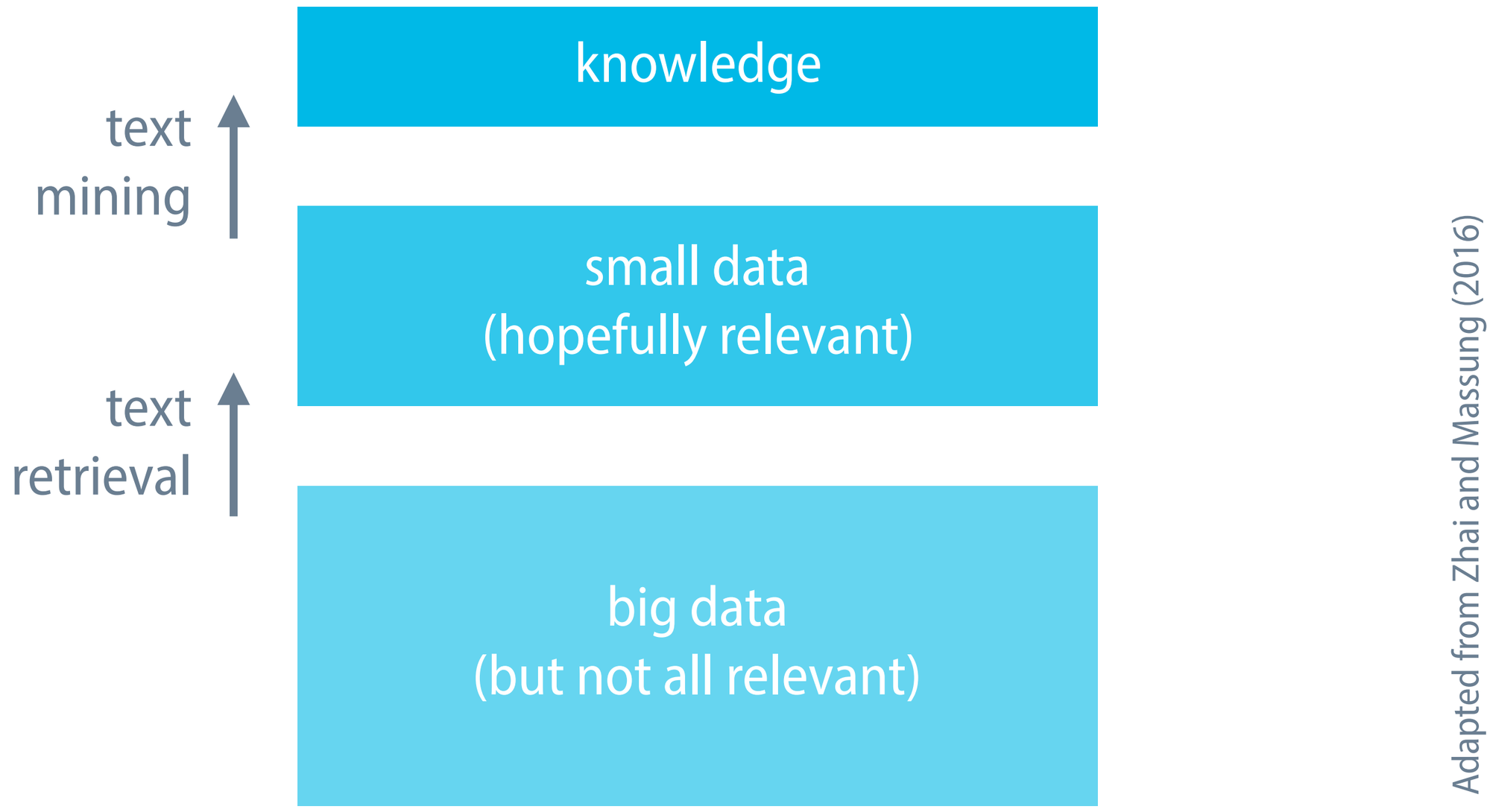
TDDE16/732A92 Text Mining (2018)

Introduction

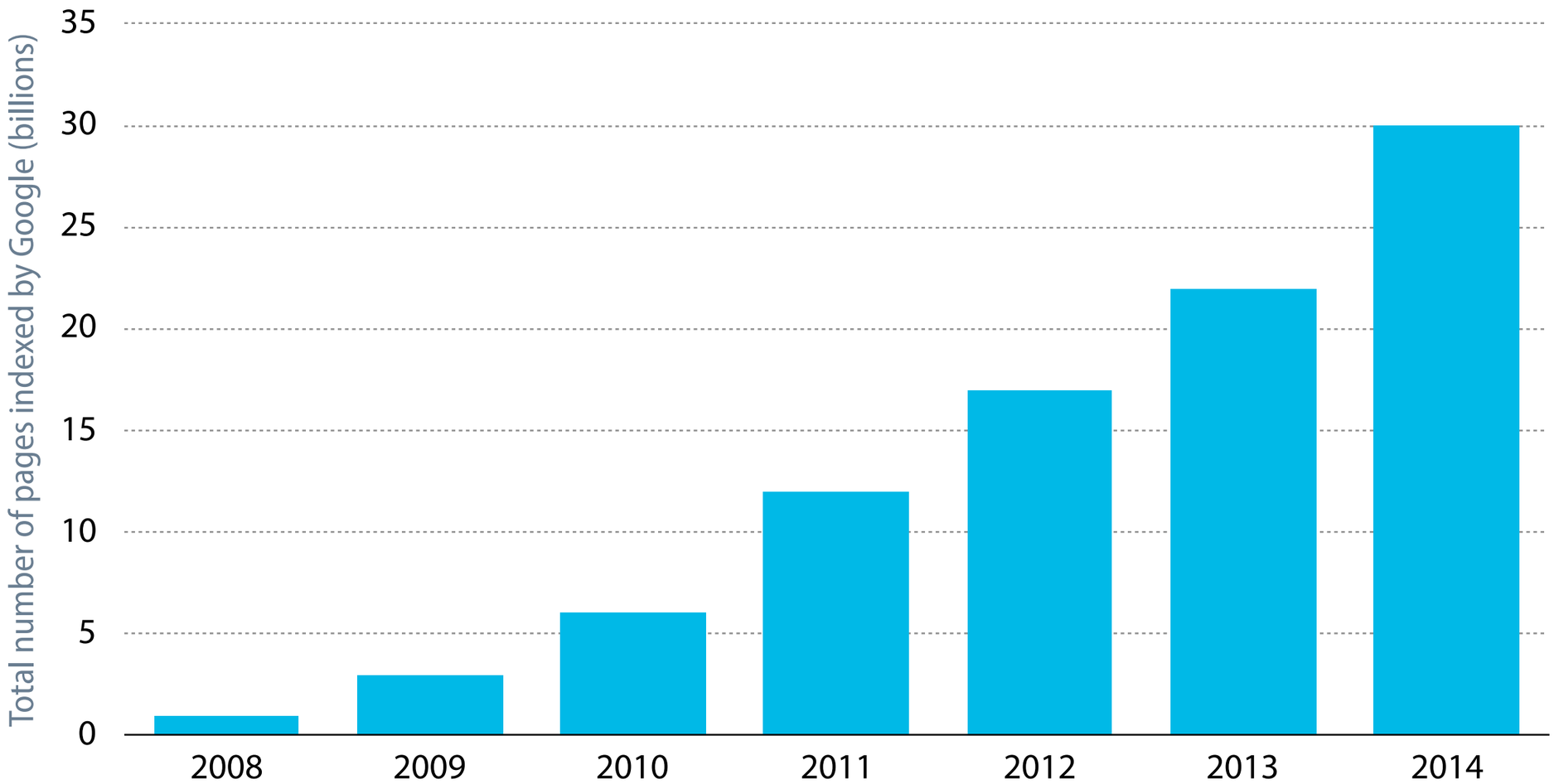
Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

Text retrieval and text mining



‘We are drowning in information.’



Source: statisticbrain.com

Text data is special

- Text data is generally produced by humans, rather than by computers or sensors.

  contrast with e.g. image data

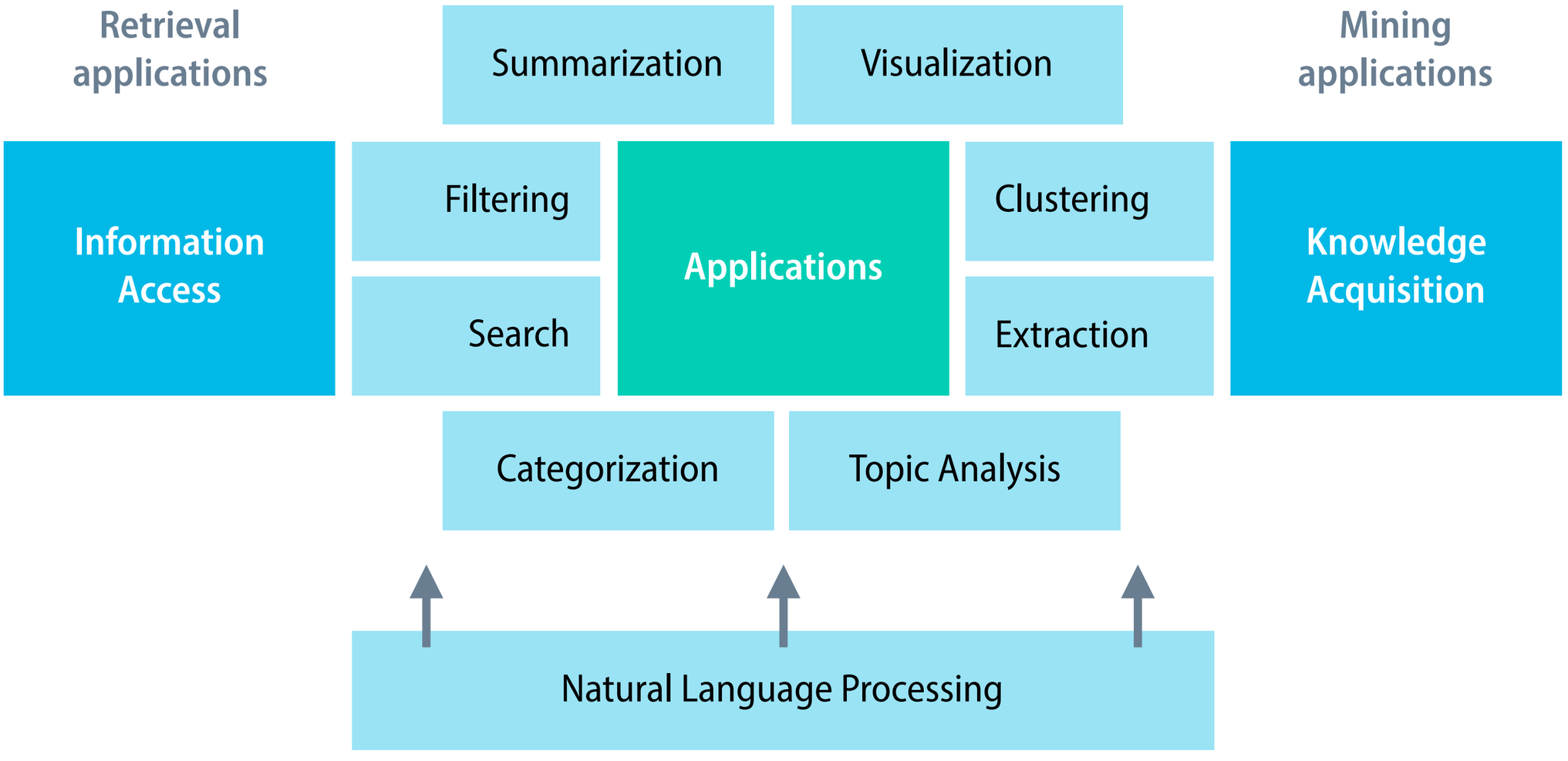
- Text data is generally meant to be consumed by humans, rather than by computers.

  so-called unstructured data

Typical applications

- **Search.** Take a user's query and return relevant documents.
- **Filtering.** Filter a stream of incoming documents.
- **Categorization.** Sort documents into predefined categories.
- **Summarization.** Generate a summary of a document collection.
- **Topic Analysis.** Identify topics in a document collection.
- **Information Extraction.** Extract entities and relations between them.
- **Clustering.** Discover groups of similar text documents.
- **Visualization.** Visually display patterns in text data.

Conceptual framework for text mining



Two functions

- **Information Access**
  Enable the user to access relevant information in time.

  search engines (pull), recommender systems (push)

- **Knowledge Acquisition**
  Enable the user to acquire knowledge ‘hidden’ in text.

  information extraction, discover interesting patterns

Adapted from Zhai and Massung (2016)

Two perspectives

- **Natural Language Processing**
  Make limited inferences based on the natural language text.

  information extraction

- **Data Mining**
  Discover and extract interesting patterns in the text data.

  topic modelling

Adapted from Zhai and Massung (2016)

JEOPARDY!®

This Stanford University alumnus co-founded educational technology company Coursera.

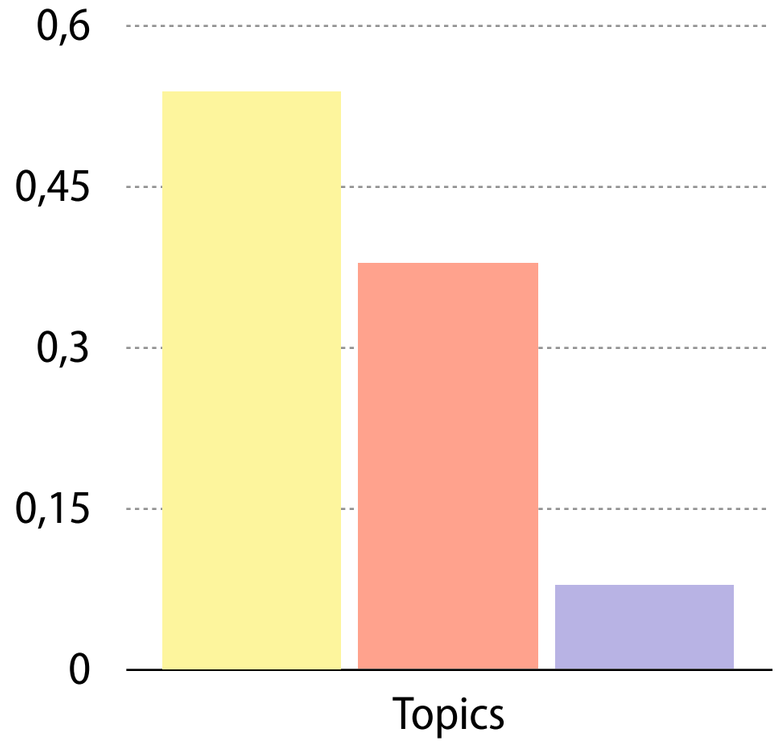
Source: MacArthur Foundation

SPARQL query against DBPedia

```
SELECT DISTINCT ?x WHERE {
 ?x dbpedia-owl:almaMater dbres:Stanford_University.
 dbres:Coursera dbpedia-owl:founder ?x.
}
```

Topic models

How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes.



Source: Blei (2012)

Topic models

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

Three stages

- Retrieving and storing textual data

  Databases and Information Retrieval

- Analysing the linguistic structure of the data

  Natural Language Processing

- Building statistical models of the data

  Statistics and Machine Learning

Three subjects, three teachers

Databases and Information Retrieval

Patrick Lambrix

Natural Language Processing

Marco Kuhlmann

Statistics and Machine Learning

Måns Magnusson

Course outline

- Data Models and Information Retrieval (Lambrix)
- Introduction to Natural Language Processing (Kuhlmann)
- Statistics for Textual Data (Magnusson)
- Text Mining Project (you!)

	Monday	Tuesday	Wednesday	Friday
W45		LEC Course Introduction		
W46	LEC Information Retrieval	LEC LAB Information Retrieval	LAB Information Retrieval	LEC Natural Language Processing
W47	LEC Natural Language Processing	LEC Natural Language Processing	LAB Natural Language Processing	
W48	LEC Statistics for Textual Data	LEC Statistics for Textual Data	LEC LAB Statistics for Textual Data	LEC Introduction to the Project
W49	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W50	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W51	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W52	Christmas Break	Christmas Break	Christmas Break	Christmas Break
W01	Christmas Break	Christmas Break	Christmas Break	Christmas Break
W02	Christmas Break	Christmas Break	Christmas Break	Christmas Break
W03	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision

Examination

	Computer labs	Text Mining Project
ECTS credits	3 credits	3 credits
to be done	in pairs	individually
grading	Pass/Fail	U345, ECTS
form of hand-in	notebooks	written project report

Example projects

- topic classification for cooking recipes
- topic analysis for the TV series *Friends*
- mood classification of songs based on lyrics
- predicting gender and age from blogs
- sentiment classification of Amazon reviews