# Natural Language and Text Analytics

Andrea Esuli

# Natural Language

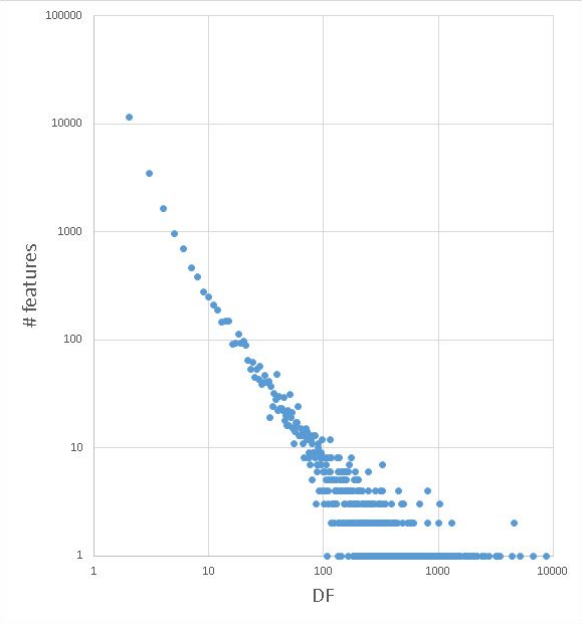
The use of a complex language is a **distinctive trait of humans**.

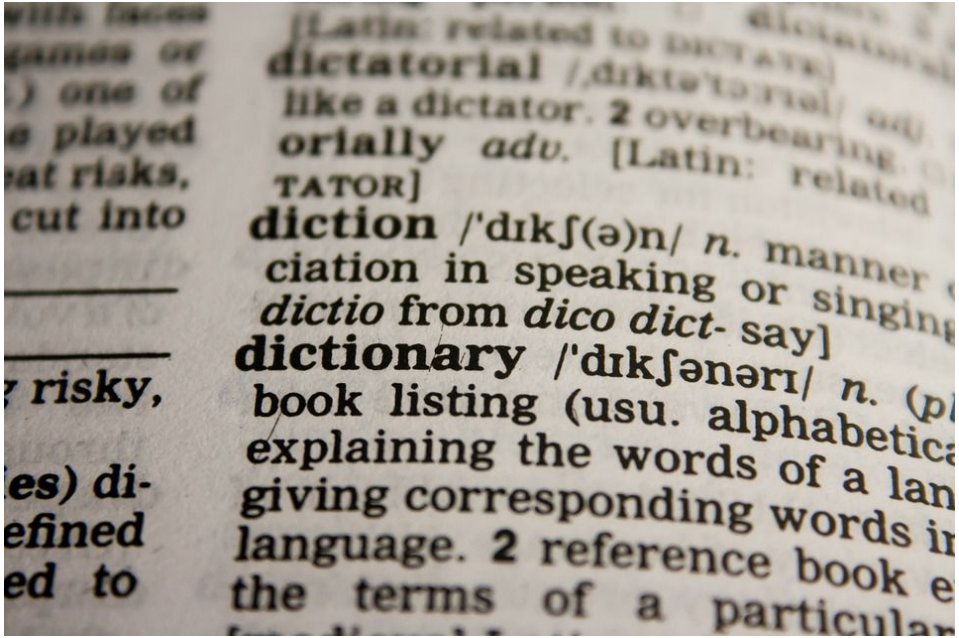
- Thousands of symbols.
- Complex syntax.
- Semantic is *mostly* compositional.
- Potentially **ambiguous**.
- Relies on **shared world knowledge**.
- Learned from scratch **along life**.
- It **evolves** along our lives.

James R Hurford, *"Human uniqueness, learned symbols and recursive though"*
*Thomas C. Scott-Phillips, Richard A. Blythe, "Why is combinatorial communication rare in the natural world, and why is language an exception to this trend?"*

# Natural Language

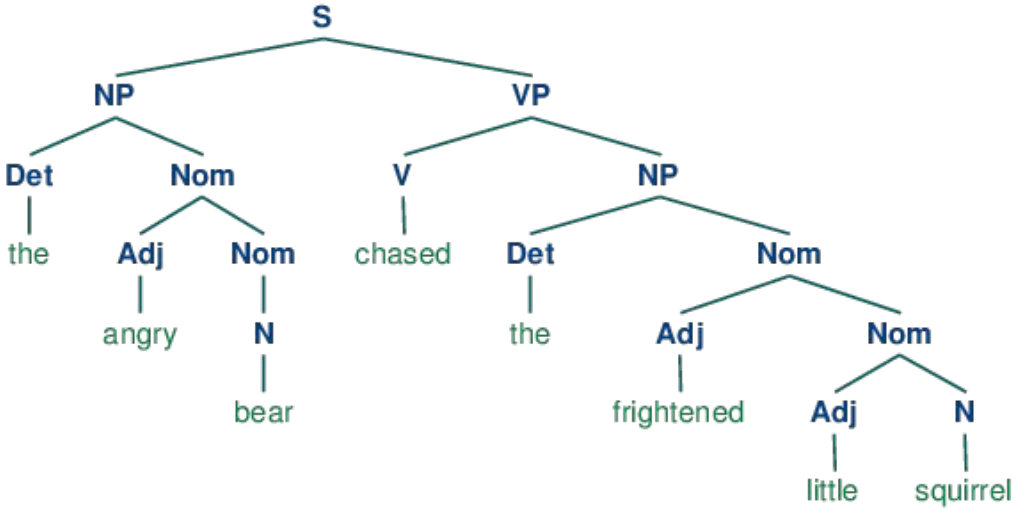
- Thousands of symbols.

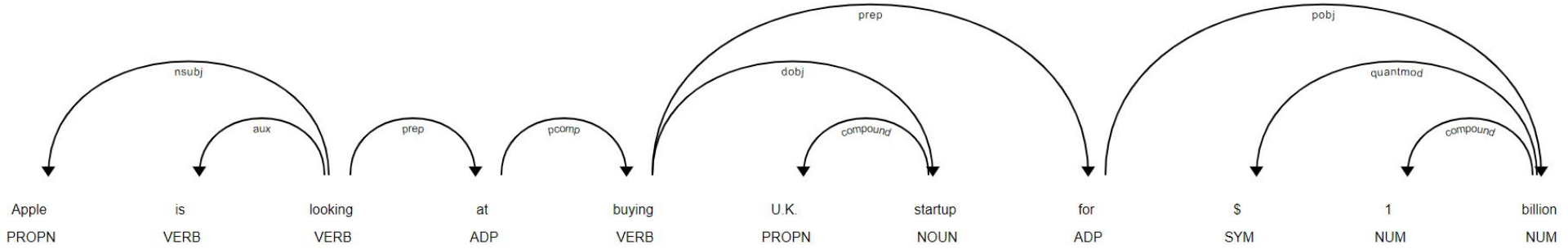


# Natural Language

- Complex syntax.



# Natural Language

- Semantic is *mostly* compositional.

**compositional:**

to buy a car
leggere un libro

**collocation:**

make the bed
saltare la lezione

**idiomatic expression:**

break a leg
dalla padella nella brace

# Natural Language

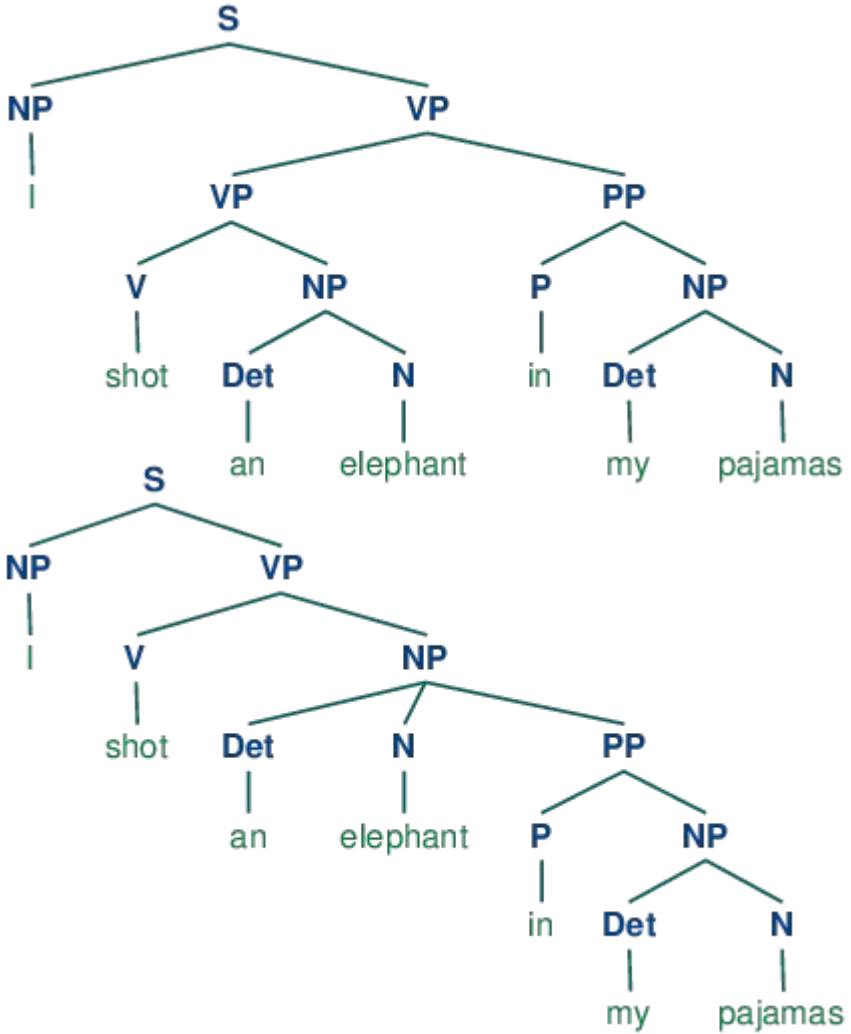
- Potentially ambiguous.
  - Part of speech ambiguity

    *beat*: Verb, Noun?
    *Listen to this cool beat.*
    *I'll beat you at checkers!*

  - Word sense ambiguity

    *interest*: attention or money?
    *There is interest on this topic.*
    *The bank raised the interest rate.*

  - Syntactic ambiguity →



# Natural Language

- Relies on shared world knowledge.

| | |
|---|---|
| "I'm faster than Lewis Hamilton" | Faster doing what? |
| "Alexis and Kate are sisters" | Relation. |
| "Alexis and Kate are mothers" | Same structure, no relation. |
| "My mother is younger than me." | Not possible. |
| "I was on a bike with a helmet" | The helmet is on my head. |
| "I was on a bike with an electric engine" | The engine is part of the bike. |

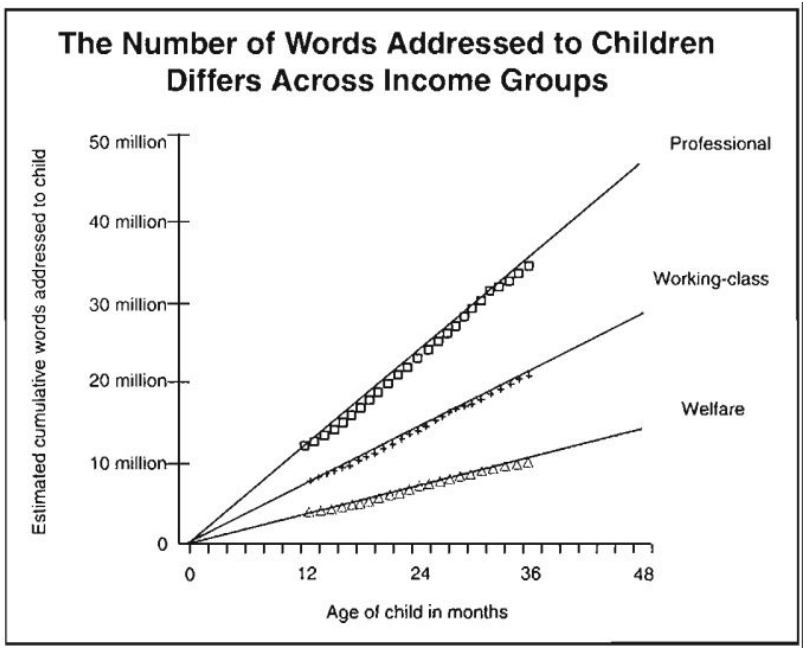
# Natural Language

- Learned from scratch along life.

# 30 million words gap

A study started in 1982 (currently debated) by Hart and Risley measured the number of words listened from their parent by children during the first three years of their life, when the main source of experience about the world is the communication with the parents.

They observed a gap of 30 million words (a factor of four) between children living in high income families and those living in low income families.



# 30 million words gap

The study than found that the measure of number of word listened from parent in the early years is a good predictor of the child future language skills, even at distance of years.

*"... We were awestruck at how well our measure of accomplishments at age 3 predicted measures of language skill at age 9-10 ... Vocabulary use at age 3 was equally predictive of measures of language skill at age 9-10 ..."*

Successive studies measured different gaps, and discovered significant biases in the original study.

Yet, many studies confirmed the positive effect of experiencing more use of language in the early, **pre-scholar** phases of life.

# Natural Language

- It evolves along our lives.

"airplane"

"quark"

"automobile"

"genocide"

"Tesla" (books, web)

"a ship"

"I googled it"

"it's in the cloud"

"crowdfunded"

"to navigate"

"super spreader"

"wear a mask"

# Natural Language

The use of a complex language is a **distinctive trait of humans**.

- Thousands of symbols.
- Complex syntax.
- Semantic is *mostly* compositional.
- Potentially ambiguous.
- Relies on shared world knowledge.
- Learned from scratch along life.
- It evolves along our lives.

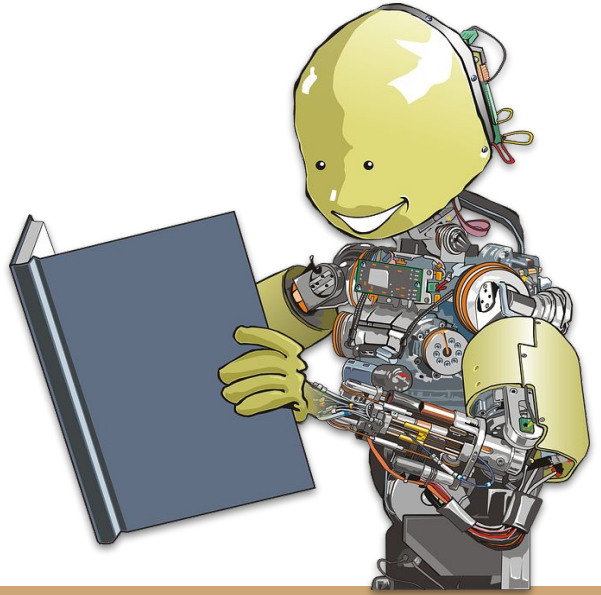
James R Hurford, *"Human uniqueness, learned symbols and recursive though"*
*Thomas C. Scott-Phillips, Richard A. Blythe, "Why is combinatorial communication rare in the natural world, and why is language an exception to this trend?"*

# Natural Language Understanding (by machines)

Natural Language Understanding aims at building machines that are able to receive and give information using natural language, like humans do.

From a computational point of view, natural language understanding is considered to be an **AI-complete** problem.

Practical text analytics tasks are simplifications of NLU that make the problem easier to solve.



# Text Mining

*"Text mining is the **analysis of data contained in natural language text**. Text mining works by transposing words and phrases in **unstructured data** into numerical values which can then be linked with **structured data** in a database and analyzed with **traditional data mining techniques**."*

Source

# Text Mining

Text mining supports data analysis problems by **recognizing** and **extracting structured knowledge** that is otherwise locked in the **unstructured** form of **natural language**, and thus **out of reach** from traditional **data mining** methods.

A complete, integrated text mining and data mining pipeline can be faster than humans at taking [critical decisions.]

| A | Symbol | Open | High | Low | Close | Net Chg | %Chg | Vol |
|---|---|---|---|---|---|---|---|---|
| A10 Networks | ATEN | 7.72 | 7.88 | 7.59 | 7.85 | 0.14 | 1.82 | 462,746 |
| AAC Holdings | AAC | 24.62 | 24.82 | 21.14 | 24.72 | 0.18 | 0.73 | 149,108 |
| AAR Corp. | AIR | 24.61 | 24.72 | 24.42 | 24.65 | 0.09 | 0.37 | 114,148 |
| Aaron's Inc. | AAN | 24.42 | 24.75 | 24.27 | 24.50 | 0.23 | 0.95 | 808,727 |
| ABB ADR | ABB | 18.74 | 18.78 | 18.48 | 18.60 | -0.22 | -1.17 | 2,599,799 |
| Abbott Laboratories | ABT | 45.22 | 45.52 | 44.94 | 45.46 | 0.54 | 1.20 | 5,522,257 |
| AbbVie | ABBV | 58.23 | 59.09 | 57.16 | 59.02 | 0.87 | 1.50 | 11,025,268 |
| Abercrombie&Fitch | ANF | 25.90 | 26.32 | 25.70 | 25.75 | 0.18 | 0.70 | 4,881,104 |
| ABM Industries | ABM | 29.75 | 30.13 | 29.59 | 30.10 | 0.43 | 1.45 | 168,706 |
| Acadia Realty Trust | AKR | 33.41 | 33.94 | 33.01 | 33.55 | 0.01 | 0.03 | 302,163 |
| Accenture Cl A | ACN | 107.29 | 108.25 | 107.22 | 108.13 | 0.91 | 0.85 | 2,405,340 |
| ACCO Brands | ACCO | 7.76 | 7.88 | 7.70 | 7.88 | 0.16 | 2.07 | 358,227 |
| Accuride | ACW | 2.57 | 2.57 | 2.45 | 2.50 | -0.08 | -3.10 | 84,018 |
| ACE | ACE | 114.63 | 116.22 | 114.54 | 116.13 | 1.28 | 1.11 | 1,811,121 |
| Acorn International ADR | ATV | 2.73 | 20.85 | 2.62 | 19.20 | 16.73 | 677.49 | 5,070,208 |
| Actuant Cl A | ATU | 24.79 | 24.96 | 23.93 | 24.36 | -0.40 | -1.62 | 623,034 |
| Acuity Brands | AYI | 232.17 | 233.74 | 231.60 | 233.22 | 2.34 | 1.01 | 214,090 |
| Adecoagro | AGRO | 11.28 | 11.28 | 10.84 | 10.98 | -0.27 | -2.40 | 307,478 |

## Nokia Launches $16.6 Billion Offer for Alcatel

Nokia CEO Rajeev Suri says merger will mean 'we are just ahead of the curve'

By JENS HANSEGARD
Nov. 18, 2015 6:19 a.m. ET

HELSINKI—Finnish wireless-equipment specialist Nokia Corp. on Wednesday commenced its share-exchange offer for Alcatel-Lucent SA shareholders in Paris and New York, betting its proposed €15.6 billion ($16.6 billion) acquisition will allow a combined company to better compete in a global race for scale in the business of making telecommunications and Internet gear.

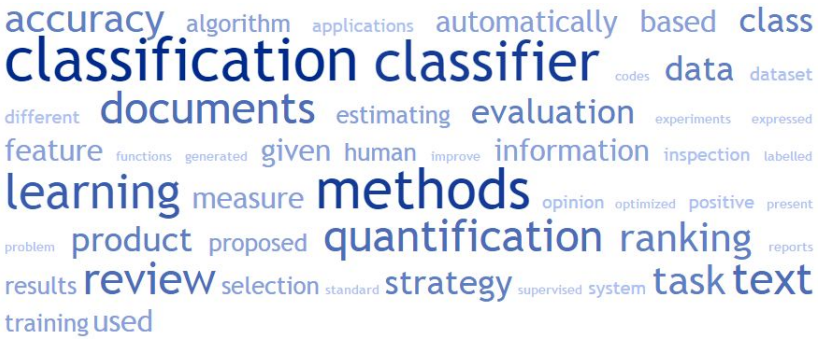
Nokia's attempt to create a one-stop shop for telecom companies and Internet service providers comes as the company faces heightened competition from new...

# Text Mining

TM methods may be aimed at **directly extracting** the answer to an information need from text,

or

give an **explicit structure** to the information contained in text in order to enable the application of data mining methods



# Text Mining

85 percent of business-relevant information originates in unstructured form, primarily text, according to Anant Jhingran of IBM Research.

Information is mostly communicated by reading or writing e-mails, reports, or articles and the like, in conversations, or by listening/watching media

Attempts to turn text into (semi-)structured forms (HTML) or microformats (atom, card, calendar, geo) only scratch the surface, but:

- Requires universal agreed ontologies
- Additional effort to make structure explicit

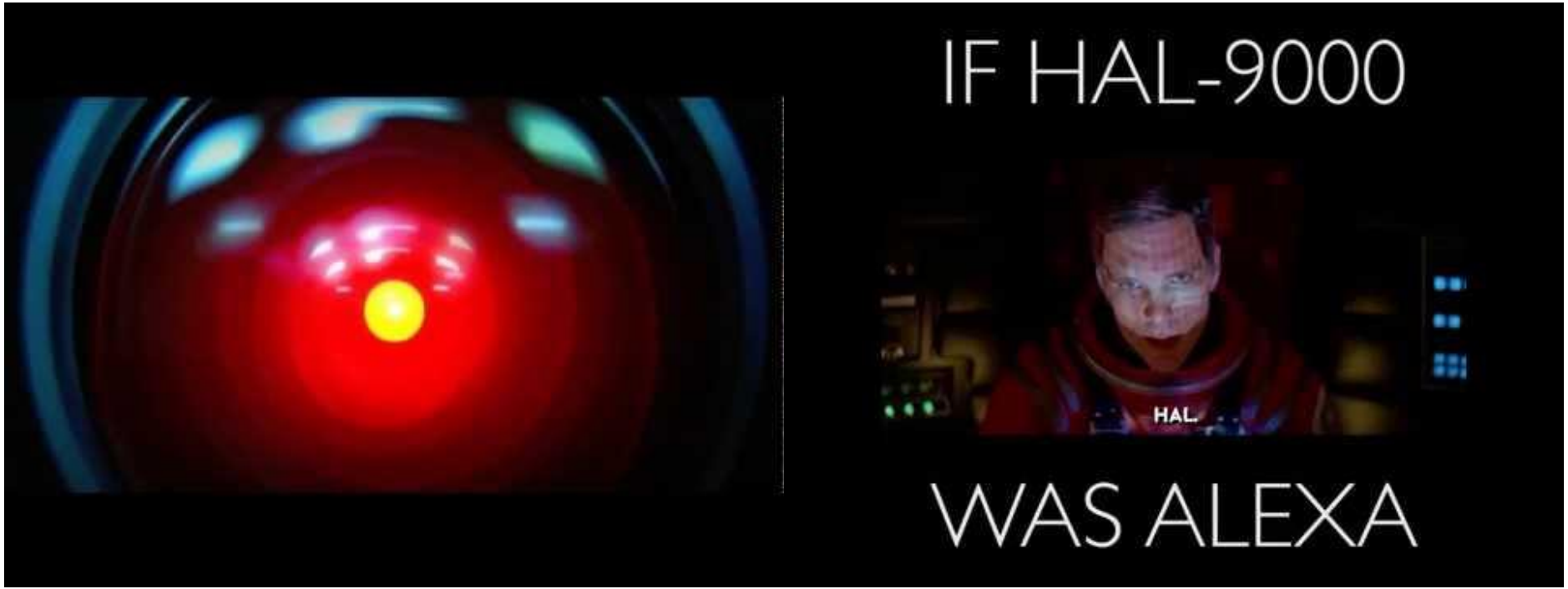
Entity linking attempts to provide a bridge.

Author of this slide: Giuseppe Attardi

# Text Mining

The vision...

...the reality



# Text Mining Tasks

The are many different practical information processing tasks in which text is the main media:

- Tagging
- Parsing
- Search
- Question Answering
- Summarization
- Translation
- Classification
- Clustering
- Regression
- Quantification
- Information Extraction
- Named Entity Recognition
- Coreference Resolution
- Entity linking
- Paraphrase
- Dialog

# Text Mining Tasks

The performance of the available methods of the tasks varies:

Very good:

- Tagging
- Parsing
- Named Entity Recognition
- Search

Progressing:

- Translation
- Classification
- Clustering
- Coreference Resolution
- Regression
- Information Extraction
- Entity linking

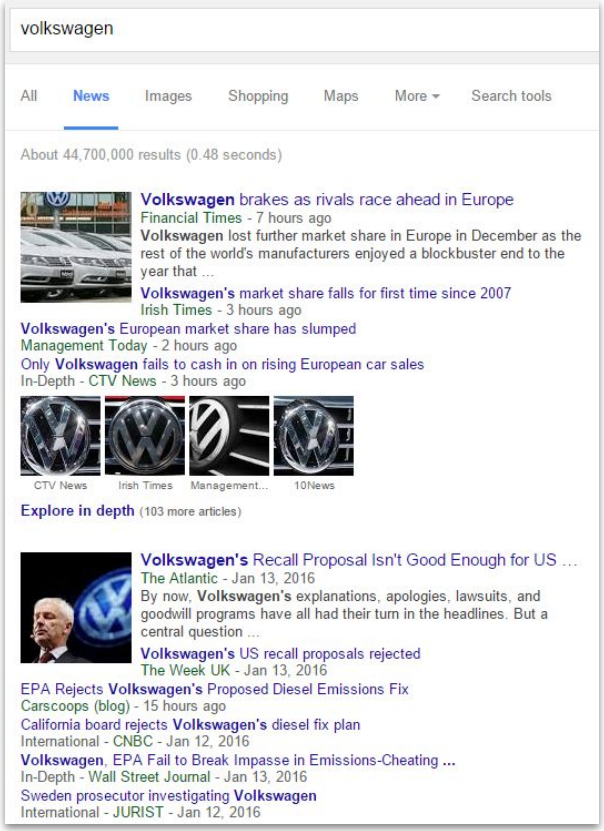
Not yet ready:

- Question Answering
- Summarization
- Quantification
- Paraphrase
- Dialog

# Practical applications

Text mining methods find application in many everyday situations.

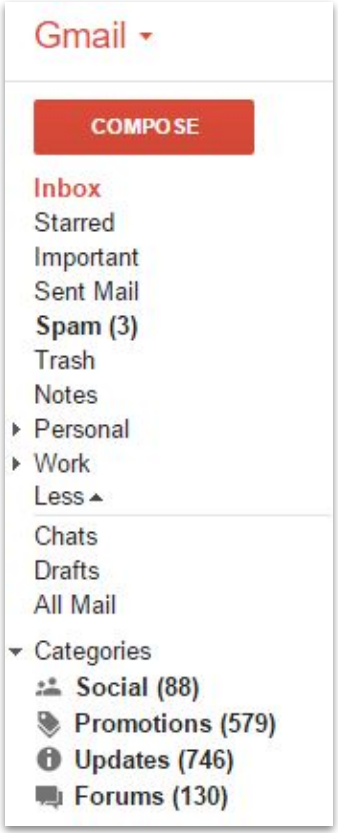
# Practical applications



# Practical applications



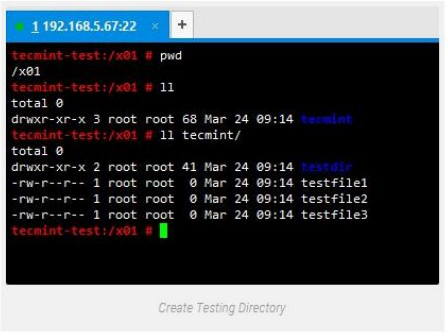
# Practical applications



# Practical applications

## Step 2: Create a Test Directory and Enable SimpleHTTPServer

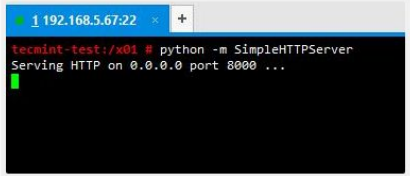
3. Create a test directory where you don't mess with system files. In my case I have a partition called `/x01` and I have created a directory called `tecmint` in there and also I have added some test files for testing.



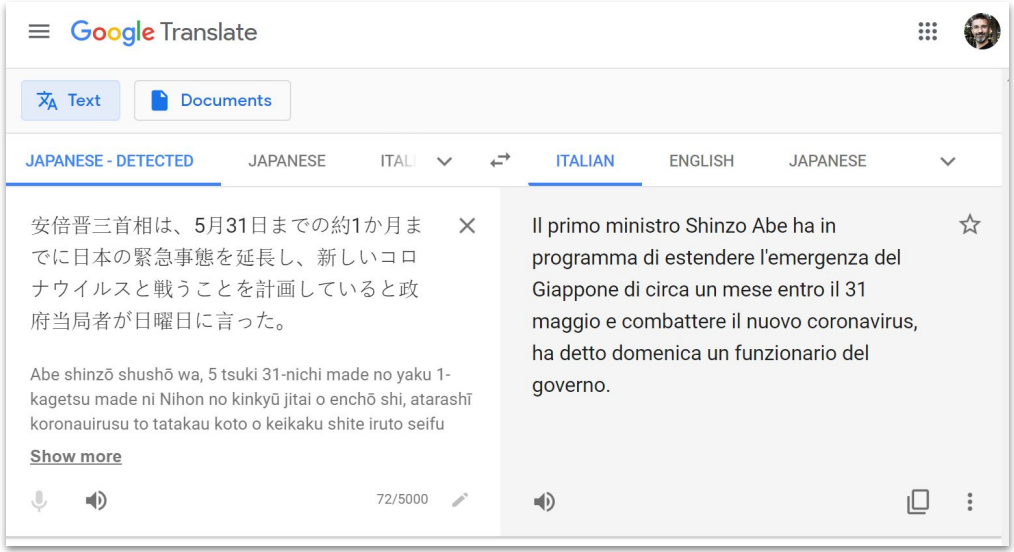

*Create Testing Directory*

4. Your prerequisites are ready now. All you have to do is try python's **SimpleHTTPServer** module by issuing below command within your test directory (In my case, **/x01//**).

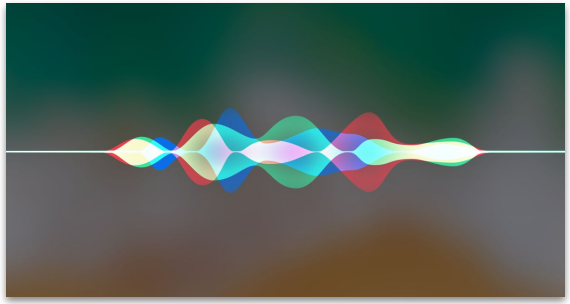
```
# python -m SimpleHTTPServer
```

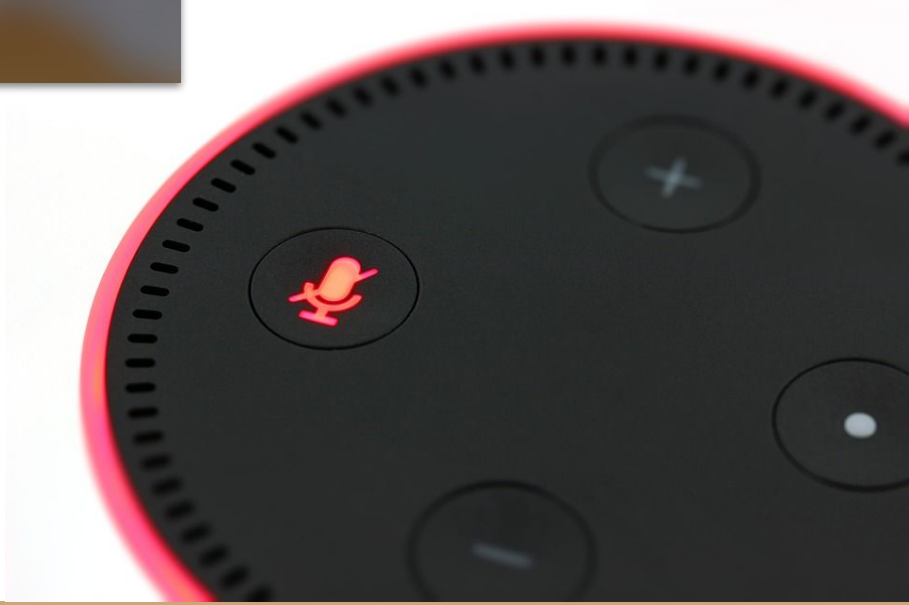


# Practical applications

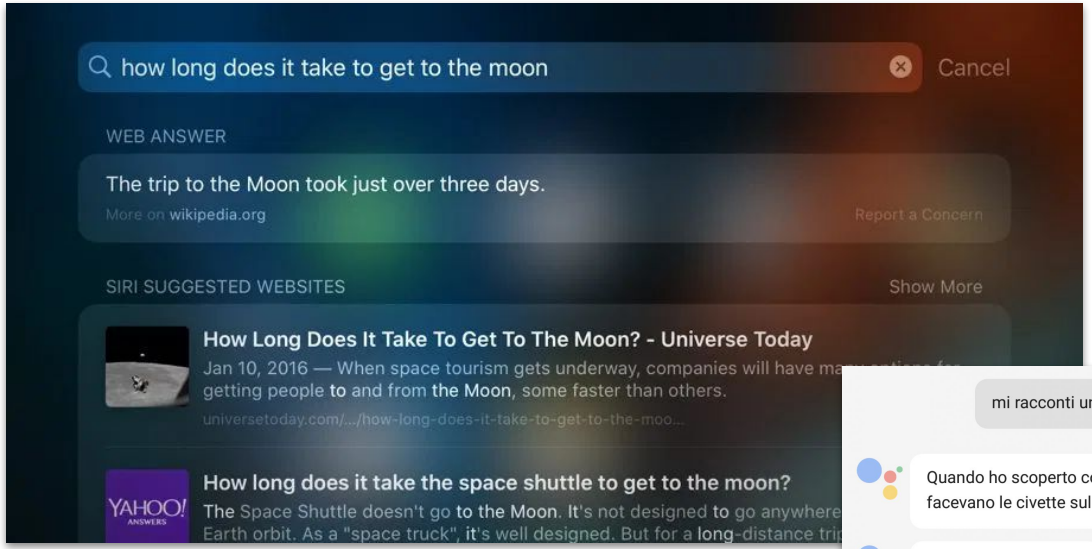



# Practical applications

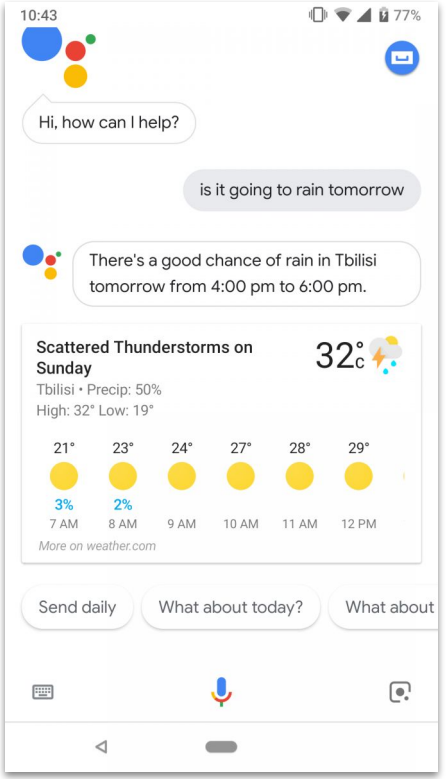




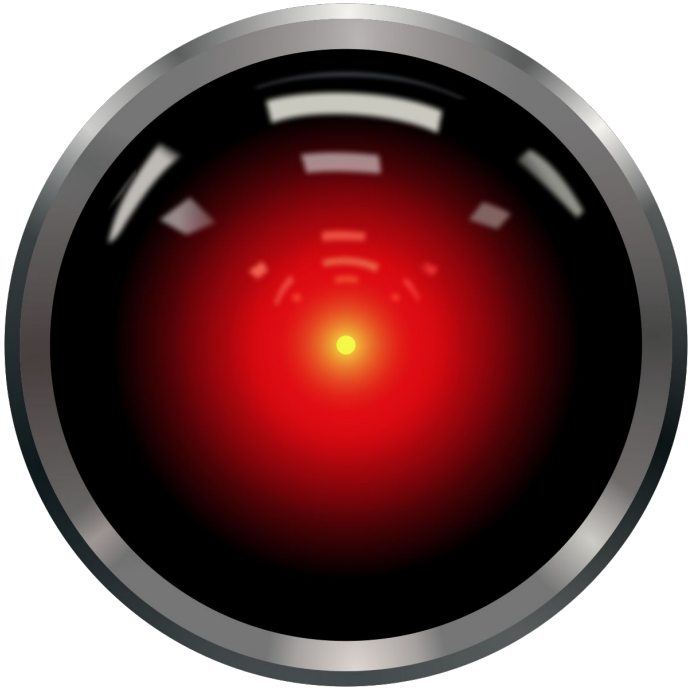
# Practical applications







# Practical applications



# Relations with other data types: tabular data

Tabular data usually contain numerical, categorical, or symbolic values that have explicit and fixed structure, constraints and relations.

Extracting structured information from historical text sources may enable data mining on information not originally designed for such processes.

Using natural language questions to query databases:

**Table**

| Rank | Name | No. of reigns | Combined days |
|---|---|---|---|
| 1 | Lou Thesz | 3 | 3,749 |
| 2 | Ric Flair | 8 | 3,103 |
| 3 | Harley Race | 7 | 1,799 |
| 4 | Dory Funk Jr. | 1 | 1,563 |
| 5 | Dan Severn | 2 | 1,559 |
| 6 | Gene Kiniski | 1 | 1,131 |

**Example questions**

| # | Question | Answer |
|---|---|---|
| 1 | *Which wrestler had the most number of reigns?* | Ric Flair |
| 2 | *Average time as champion for top 2 wrestlers?* | AVG(3749,3103)=3426 |
| 3 | *How many world champions are there with only one reign?* | COUNT(Dory Funk Jr., Gene Kiniski)=2 |
| 4 | *What is the number of reigns for* Harley Race? | 7 |

# Relations with other data types: graphs

Social networks are an important source of always up-to-date information about social groups, events and debated topics.

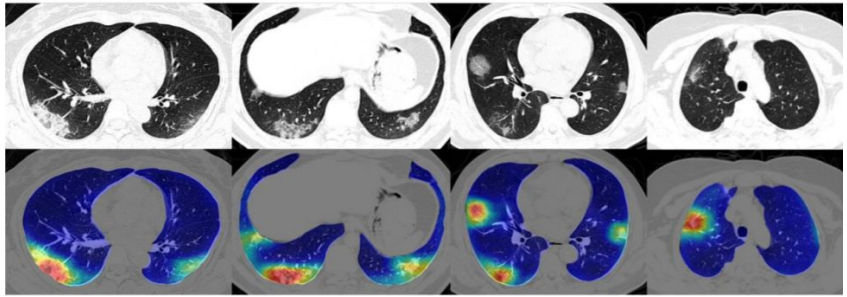
Methods based on **graph analysis** enables discovering the **structure and the dynamics** of a network at various levels: local relations, communities, large scale behaviors.

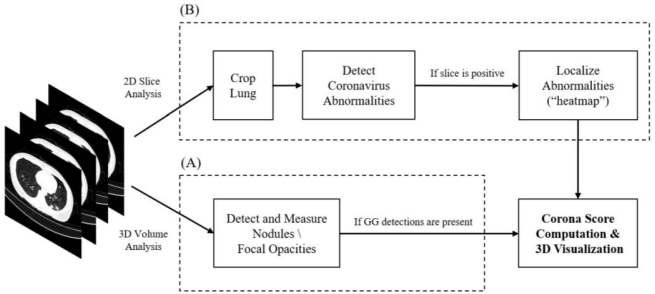
Enriching these methods with the **analysis of the actual content** that is shared on a network enables to better characterize the elements of the network and to give more detailed meaning to the interactions, enabling to investigate complex processes, e.g., abnormal behaviors, manipulation attempts...

# Relations with other data types: images and video

- Images and video represent a signal that is directly linked with a low-level perception of the world, i.e., vision.
- The entities involved in typical image processing tasks are usually bound to real world physical objects and actions (e.g., object tracking in autonomous driving).

This properties make vision-based problem relatively easier than language ones, allowing NN-based methods to achieve human-like performances.





# Relations with other data types: images and video

Cross-media processing:

- Image captioning is the task of describing the content of an image.

It has application in retrieval processes.

It is of help for people with vision problems.

# Relations with other data types: images and video

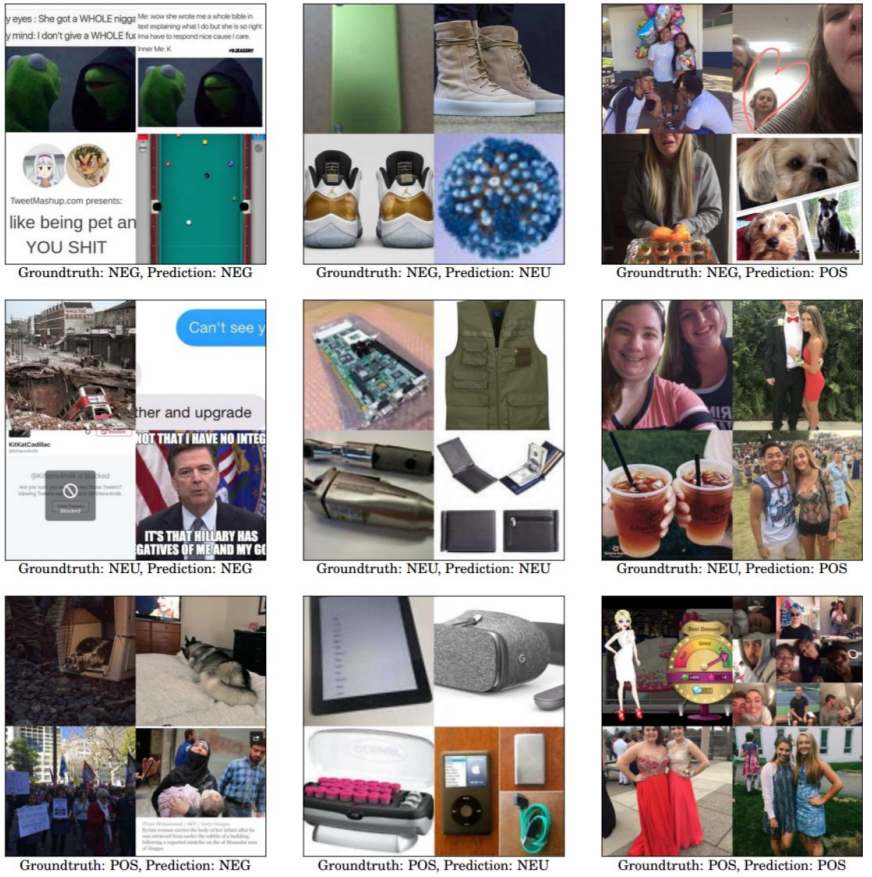
Cross-media processing:

- Abstract visual representation of text enables smarter retrieval methods.

# Relations with other data types: images and video

Cross-media processing:

- The co-occurrence of text and images supports **transfer learning**, e.g., performing sentiment analysis on images.



# Implementing Text Analytics methods

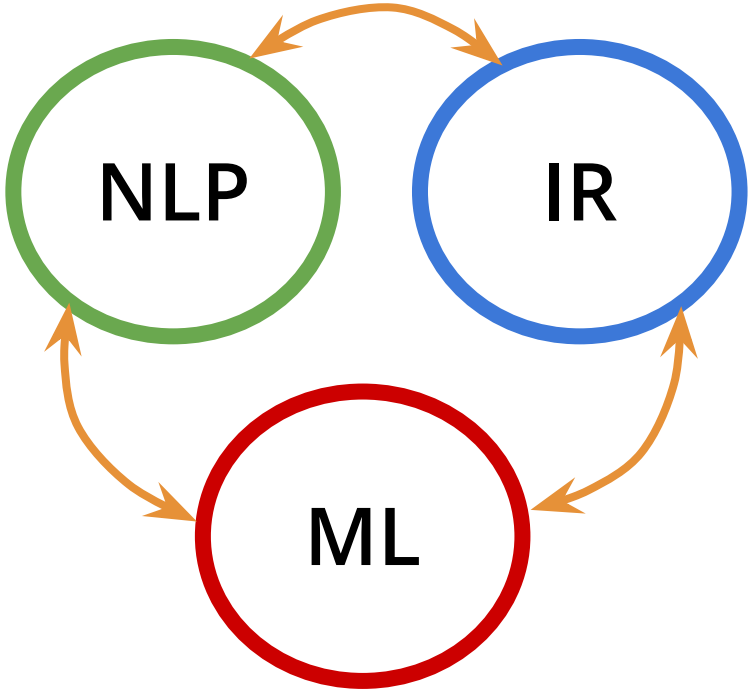
Text Analytics solutions are usually built combining methods from three disciplines:

- Natural Language Processing (NLP): studying how to recognize and process the elements and structures that build language as we humans define them.
- Information Retrieval (IR): studying the statistical properties of language, treating it as a signal on a channel, using Information Theory method.
- Machine Learning (ML): exploiting statistical/neural learning methods on properly processed representations of language to build descriptive/predictive models of some properties of interest.

# Three interlocked disciplines

These disciplines are not one distinct from the other.

- NLP methods are often built themselves on top of IR and ML methods.
- ML adopts IR measures to define the goals of the learned models.
- ML often assumes language to be manipulated by NLP and IR methods to be able to work on it.



# Information through pipelines

The concept of a **processing pipeline** is recurring in text analytics methods, as text is rarely processable *as is* by ML methods and it requires to be transformed into representation that are fit to the specific ML method in use.

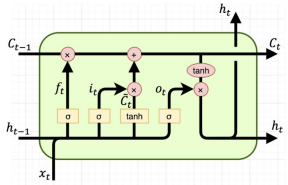
Such transformations are not just simple conversions of format, they may actively filter and modify the information in order to be better exploited by successive stages of the pipeline, so as to improve the final outcome of the process.

# The Neural Network revolution

Neural Networks had a strong comeback starting around 2000.

Three key factors made it possible:

- Lots of data from new models of data production and sharing.
- Powerful parallel computing that is well fit for NN algorithms
  - GPU, TPU...
- Improved and novel NN algorithms and architectures
  - There was no "Deep" in NN before.
  - Few researchers brought NN research through the AI winter...
  - to the AI renaissance: LSTM, CNN, ReLU, Dropout, Residual Networks, Attention, Capsule Networks, Transformers, GAN, RL...

# Neural Networks and Language

The NN revolution had a great impact of some problems:

- Translation
- Summarization
- Question Answering

mainly due to the ability of NN to process huge amounts of text and the large size of their models, i.e., **to learn very large language models.**

Yet, NN for NLP have **limitations**, specially in areas that require system 2 type reasoning: e.g., paraphrase.