



**HAL**  
open science

# Introduction to Text Mining: from basics to applications

Luca Foppiano

► **To cite this version:**

| Luca Foppiano. Introduction to Text Mining: from basics to applications. 2019. hal-02045341

**HAL Id: hal-02045341**

**<https://inria.hal.science/hal-02045341>**

Submitted on 21 Feb 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Introduction to Text  
Mining:  
from basics to applications**

Luca Foppiano

# About me

- Started as software engineer, currently working in Text and Data mining.
- 7 years at the European Patent Office (NL)
- 3 years at Inria (National Institute for Computer science and Mathematics) (FR)
- Currently at NIMS (National Institute for Material Science) (JP)
- Main topics: Machine Learning and TDM

# Text and Data Mining

- Text Mining or TDM (Text and Data Mining) is the process of deriving high quality information from other sources
  - Information retrieval
  - Information extraction (mining)
  - Knowledge management



- The approach for TDM require an understanding (shallow or more deep) of the text
- NLP (Natural Language Processing) is a very wide subfield, studying ways to program computers to interact with natural language data (text for example)

# Ideal conditions

- Scalability
- Repeateability
- Genericity (\*)
- Automatic

# Artificial Intelligence

- Artificial Intelligence (shorten for AI) is the branch of Computer Science that study methods and techniques aiming to mimic the functioning of the human brain
- Origin dating back to the 50' (Alan Turing, Turing test)
- Continuously evolving, more than 50 years of investments
- Most known applications: OCR, Voice Recognition, Self Driving Cars, Spam recognition, etc...

# **ARTIFICIAL INTELLIGENCE**

Programs with the ability to learn and reason like humans

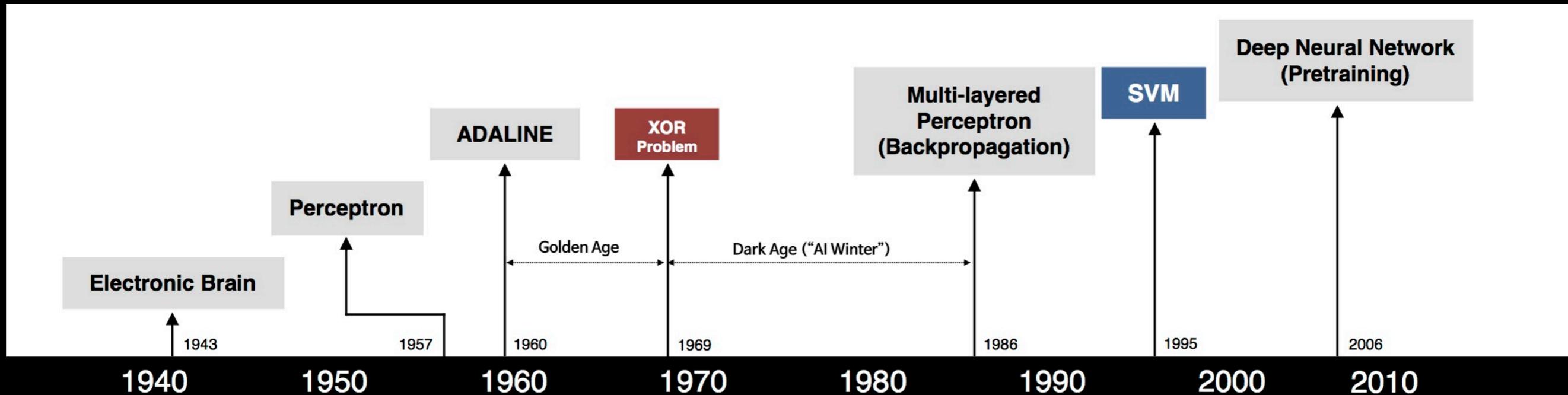
## **MACHINE LEARNING**

Algorithms with the ability to learn without being explicitly programmed

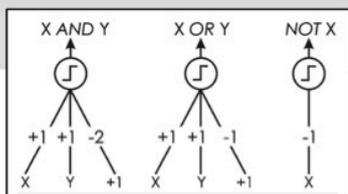
## **DEEP LEARNING**

Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

# AI Evolution



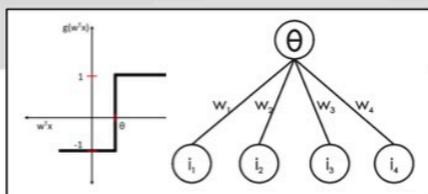
S. McCulloch - W. Pitts



- Adjustable Weights
- Weights are not Learned



F. Rosenblatt



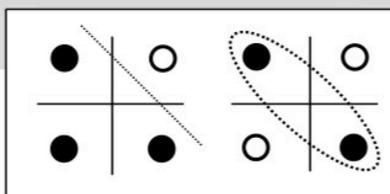
- Learnable Weights and Threshold



B. Widrow - M. Hoff



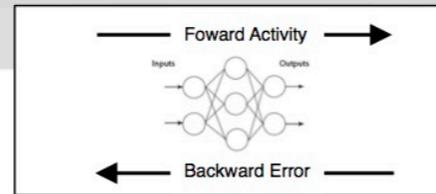
M. Minsky - S. Papert



- XOR Problem



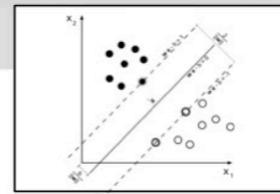
D. Rumelhart - G. Hinton - R. Williams



- Solution to nonlinearly separable problems
- Big computation, local optima and overfitting



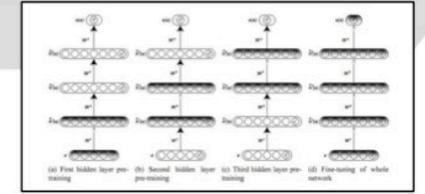
V. Vapnik - C. Cortes



- Limitations of learning prior knowledge
- Kernel function: Human Intervention



G. Hinton - S. Ruslan



- Hierarchical feature Learning

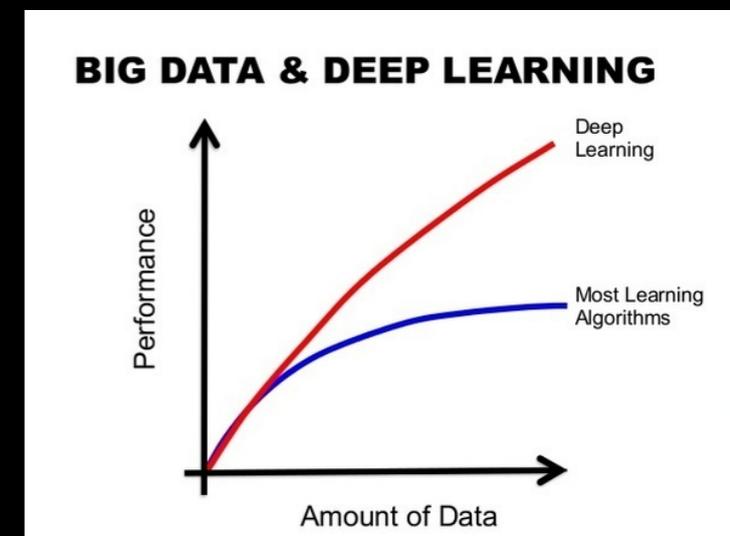
# Deep Learning

The state-of-the-art based on the brain representation concept defined in the 50 by Turing and researched in the 70 with the neural networks.

Computers weren't powerful enough so for deep networks had only 2-3 layers (overtaken by statistical models, like SVM, CRF)

In the last decade with the serveral improvements:

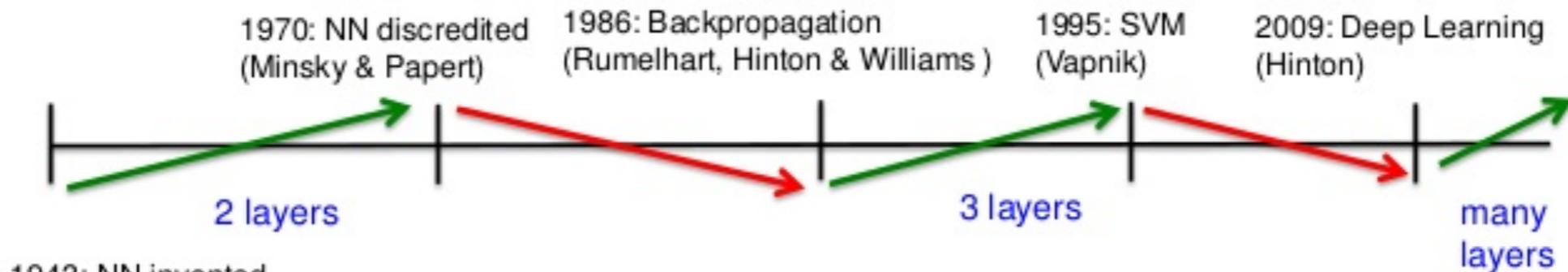
- computer power (thanks gamers!!)
- algorithms
- data availability



# The brain model

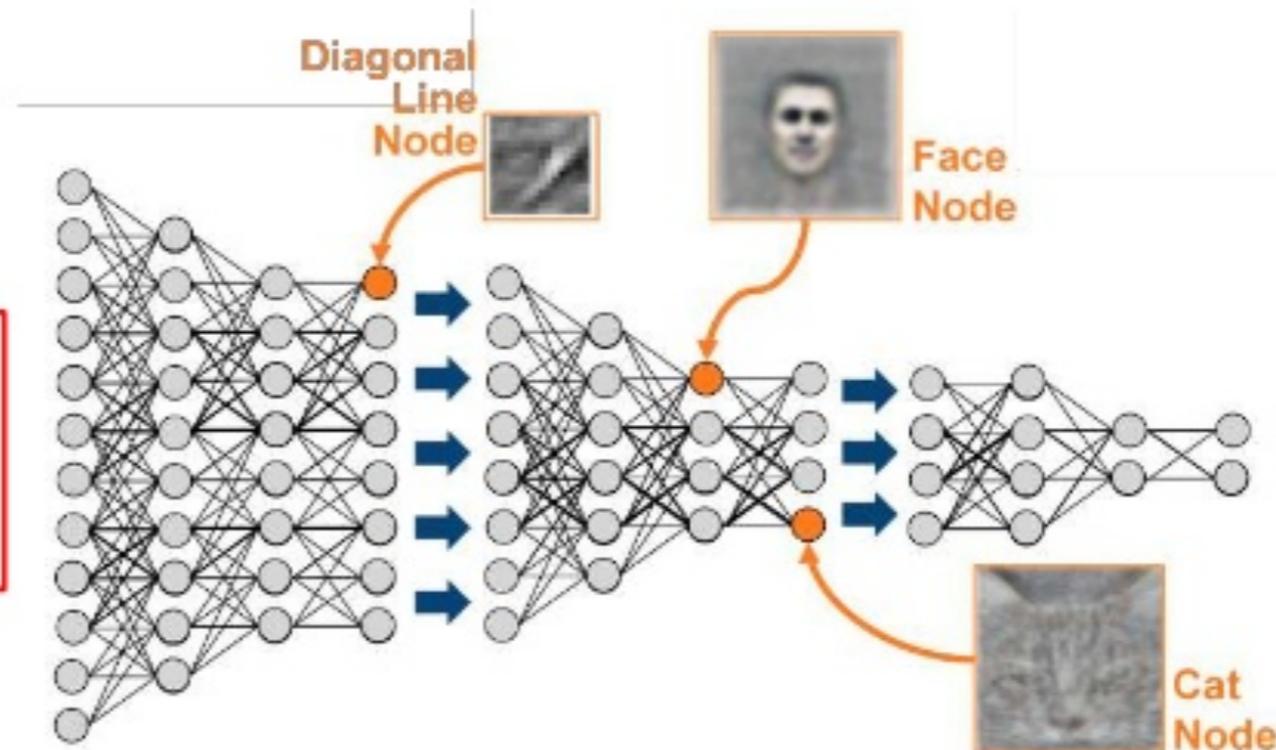
D S  
R C

## Deep Learning: Neural Nets Strike Back(again)



1943: NN invented  
(McCulloch & Pitts)

-Model Size: 10B parameters  
-Used by: Yahoo!, Google,  
Microsoft, Baidu,  
IBM, Scyfer ☺





# DL and TDM

- Images or Audio are easily representable as vectors or matrices...
- Text is shallow, require an alternative representation of their basic component: words
- How to represent words?

# Word 2 Vec

- First draft published by Google in 2013 (word2vec)
- Google provided a pre-trained models on millions of sentences
- Each word can be transformed into a vector
- “Similar” words output similar vectors
- called Embeddings

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

# TDM in action

- Search
- Recommendation systems (Netflix, Amazon...)
- Analytics
- Disambiguation / Entity Linking
- Mining (extraction of specific information)

# Search

- Ambiguities in wording
- Correctly infer the context (short query)
- How to prioritise the results? Nobody goes to page 2
- Lot of data, but not clean

# Recommendation systems

Because you watched Star Trek: Discovery >



- Your Daily Mix
- Recently Played
- Songs
- Albums
- Artists
- Stations
- Local Files
- Videos
- Podcasts

+ New Playlist



## The Beatles

PLAY

FOLLOW

...

MONTHLY LISTENERS  
10,262,687

OVERVIEW RELATED ARTISTS ABOUT

### Popular

	1	+	Here Comes The Sun - Remastered	1,156,436,448
	2	+	Come Together - Remastered	777,487,854
	3	+	Let It Be - Remastered	659,320,647
	4	+	Hey Jude - Remastered 2015	653,969,858
	5	+	Twist And Shout - Remastered 2009	642,976

SHOW 5 MORE

Merch

### Related Artists

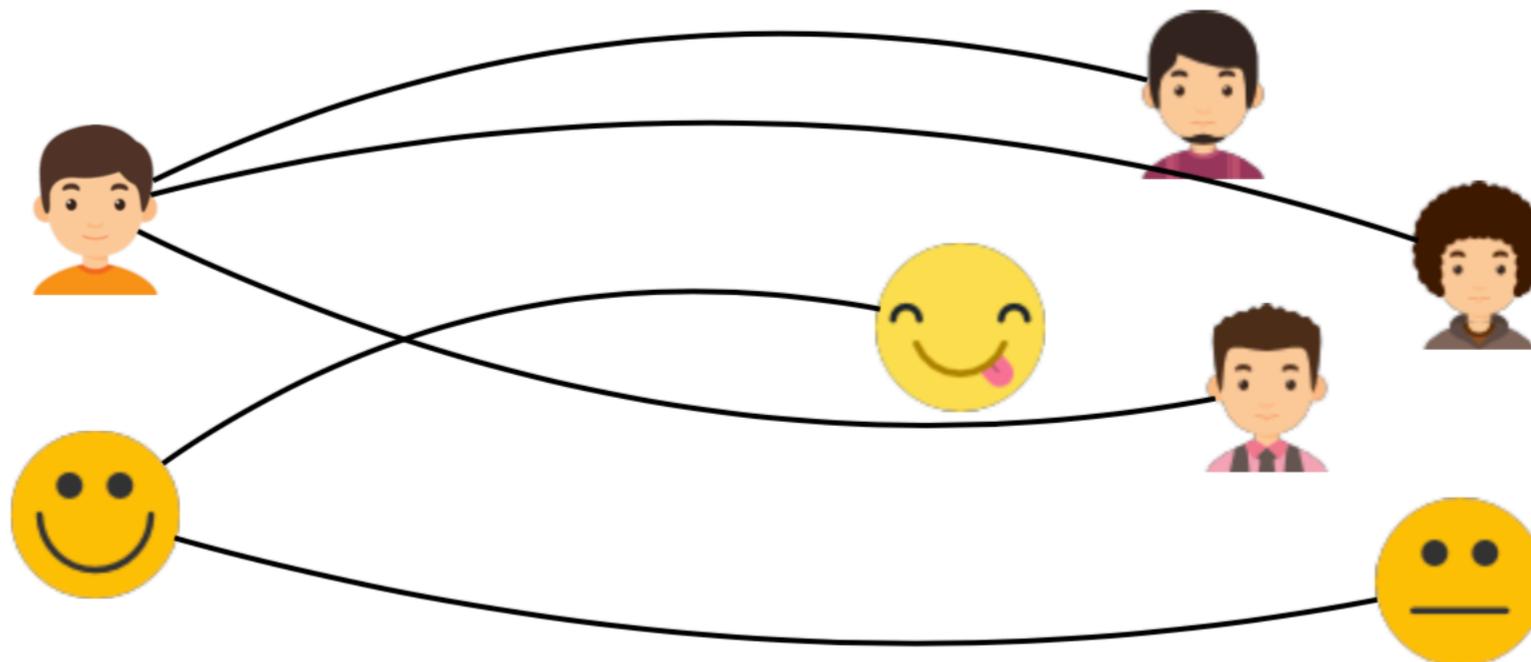
- George Harrison
- John Lennon
- Paul McCartney
- The Beach Boys

# Disambiguation

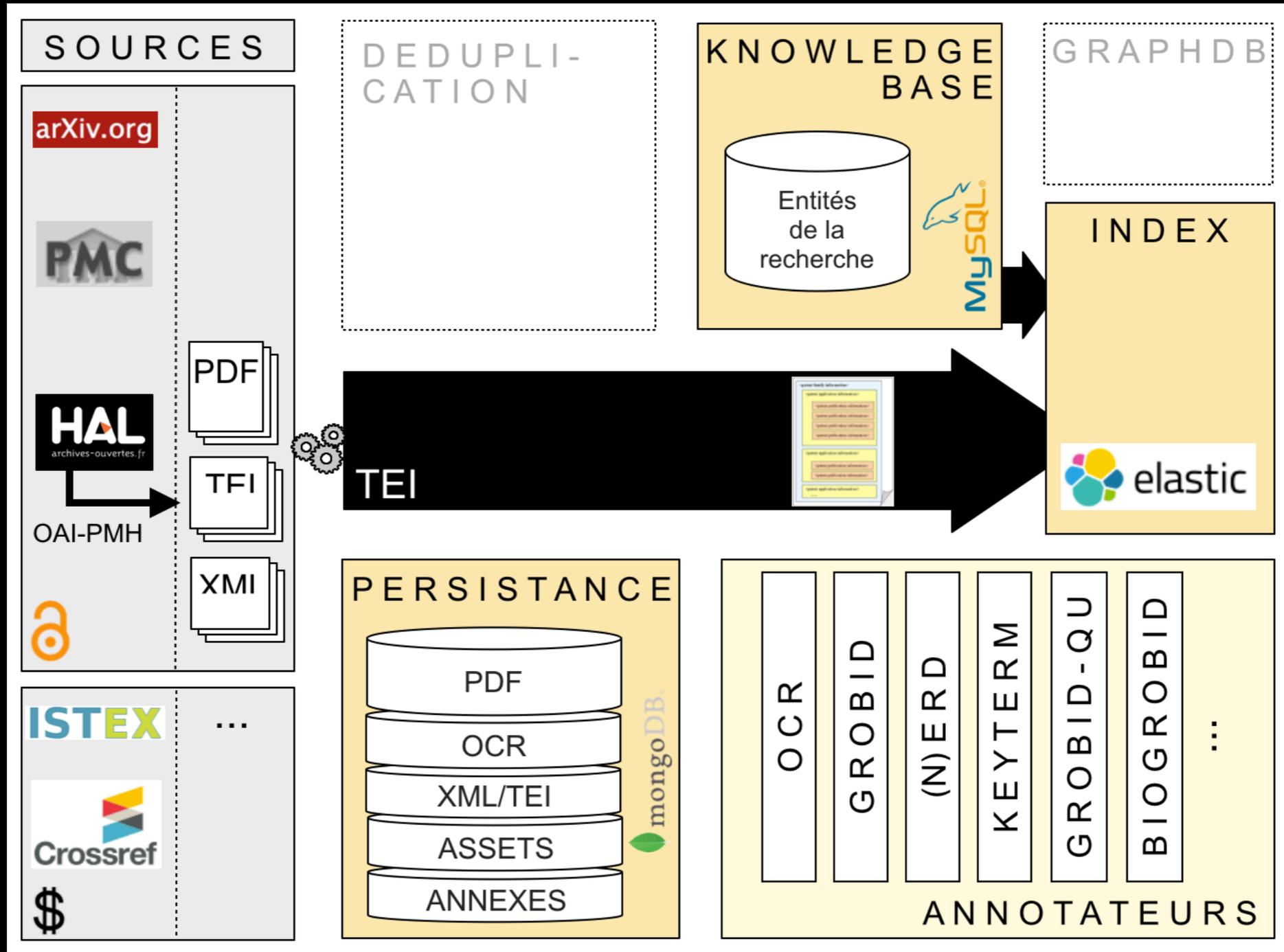
**Entity matching** is the task of deciding if two sets of data elements refer to the same real-world entity.

**Real world**

**Digital world**



# Inria anHALytics



Demo

**Thank you**