Introduction to Text Mining and its 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

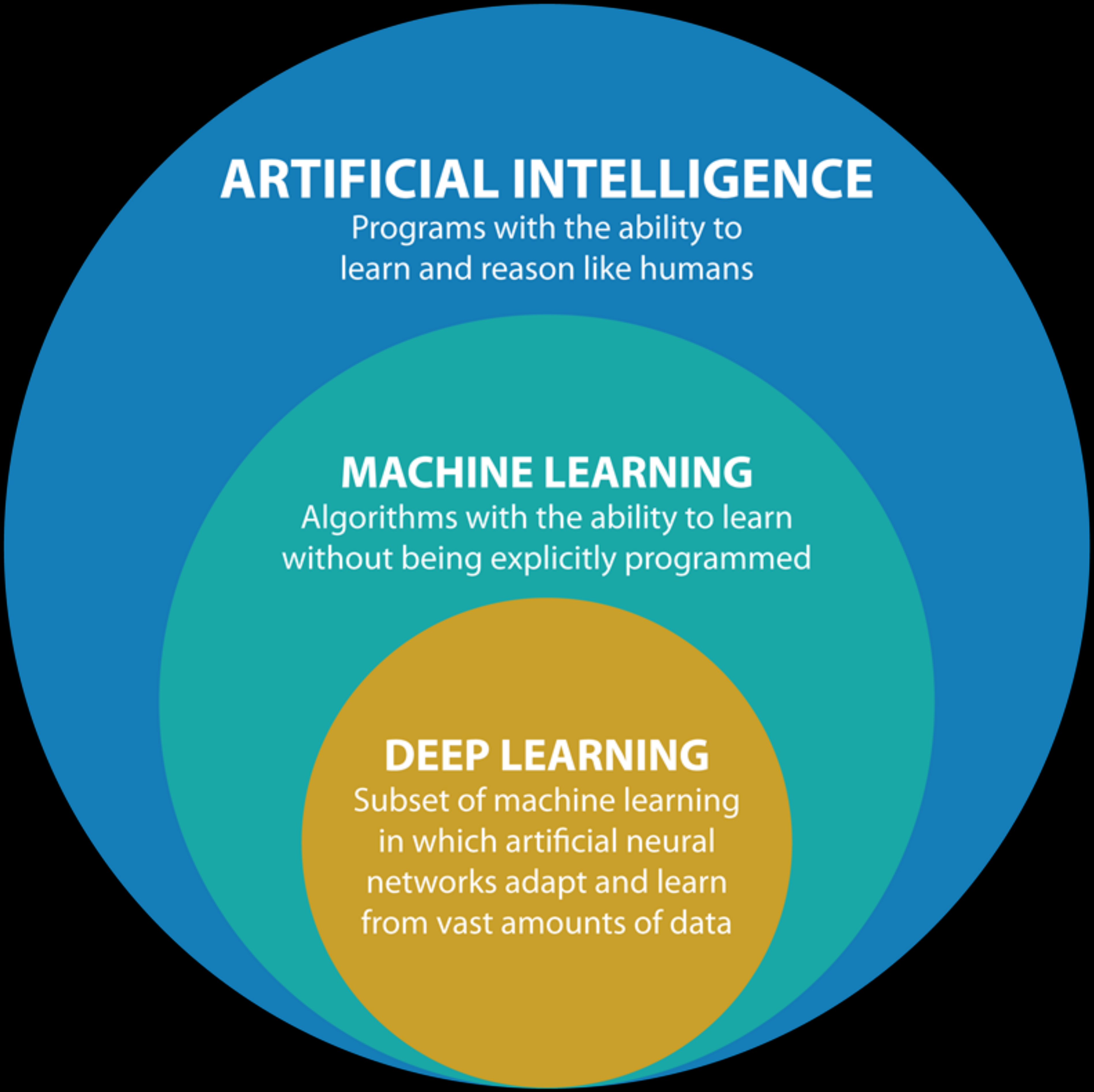
Artificial Intelligence

- Artificial Intelligence (shorten for AI) is the branch of Computer Science that study methods and tecniques 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

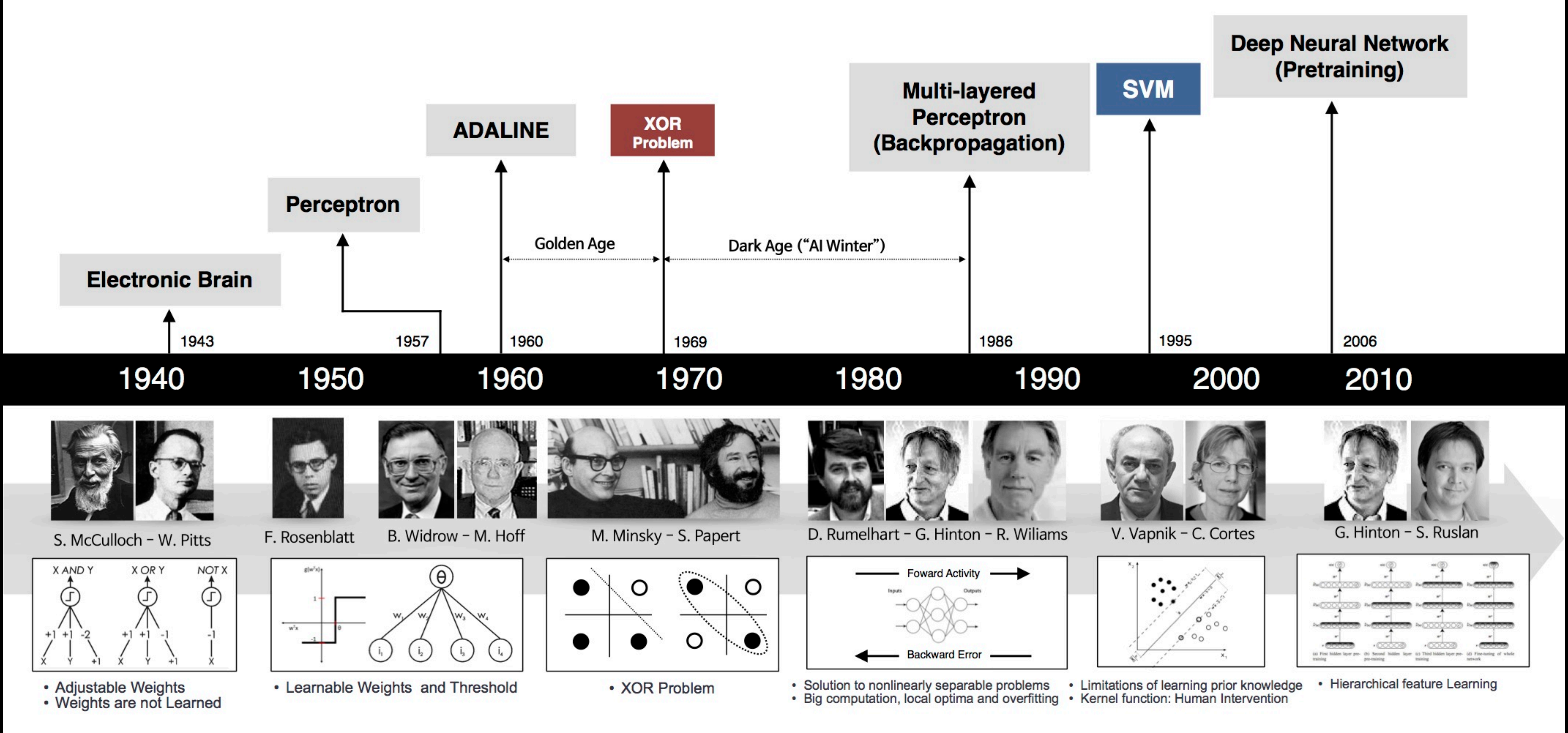
MACHINE LEARNING

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



Electronic Brain

1943

Perceptron

1957

ADALINE

1960

XOR Problem

1969

Multi-layered Perceptron (Backpropagation)

1986

SVM

1995

Deep Neural Network (Pretraining)

2006

Golden Age

Dark Age ("AI Winter")

1940 1950 1960 1970 1980 1990 2000 2010

S. McCulloch – W. Pitts

- Adjustable Weights
- Weights are not Learned

F. Rosenblatt

B. Widrow – M. Hoff

- Learnable Weights and Threshold

M. Minsky – S. Papert

- XOR Problem

D. Rumelhart – G. Hinton – R. Wiliams

- 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

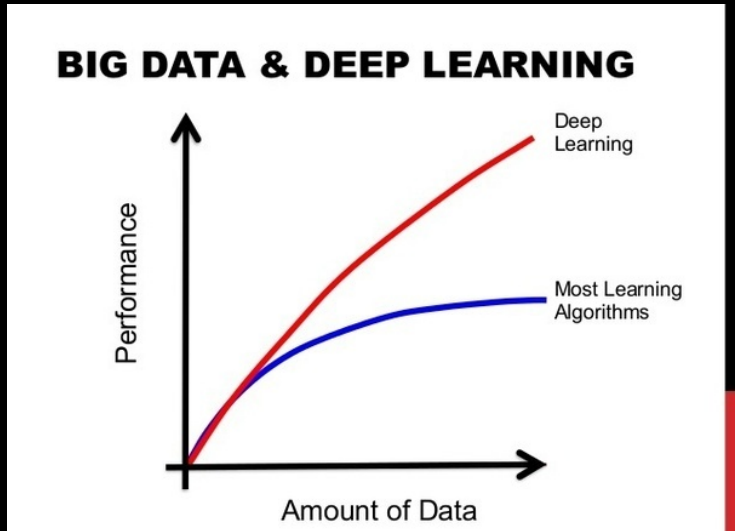
Deep Learning

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.

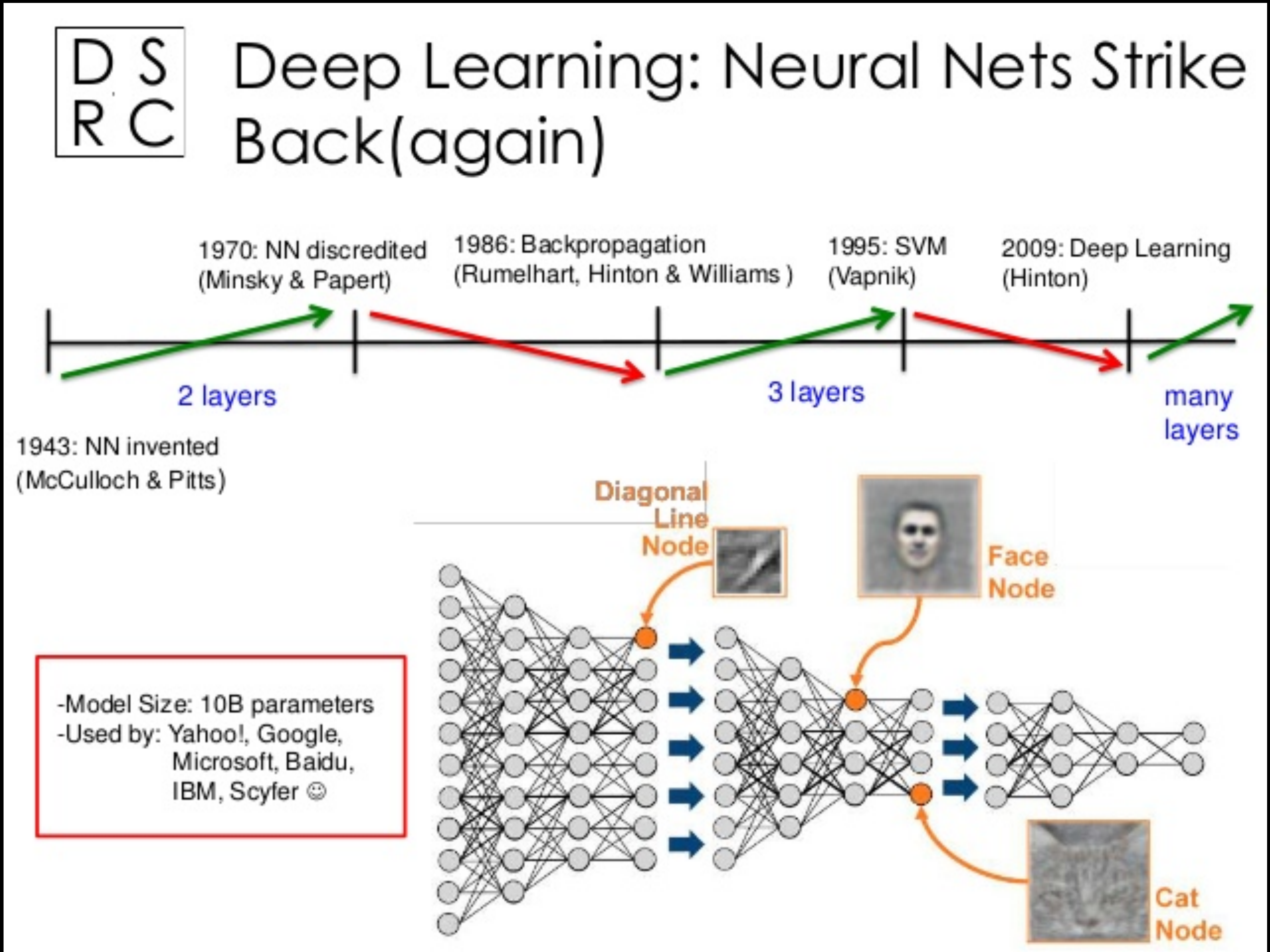
Computers weren’t powerful enough so for deep networks had only 2-3 layers (overtaken by statistical models, like SVM, CRF)

In the last decade with the serveral improvements:

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



The brain model



Deep Learning: Neural Nets Strike Back(again)

1943: NN invented (McCulloch & Pitts)

1970: NN discredited (Minsky & Papert)

1986: Backpropagation (Rumelhart, Hinton & Williams)

1995: SVM (Vapnik)

2009: Deep Learning (Hinton)

2 layers

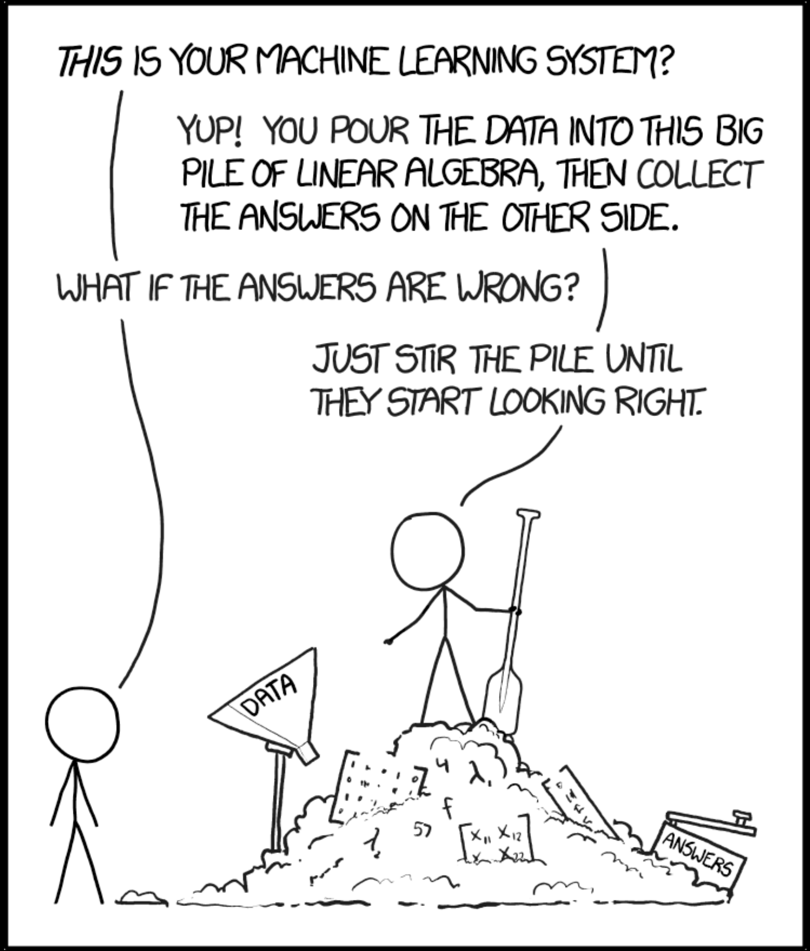
3 layers

many layers

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

Deep Learning

- There are still grey areas
- The knowledge required to get decent results has decresed - “if in doubt, add a layer”
- Less needs to tune
- Common approach is to use as much data as possible and let the network figure out what to do with it

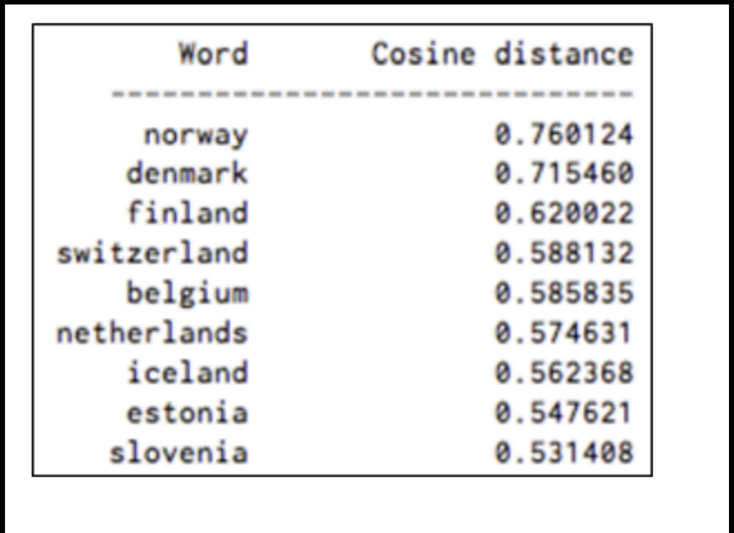

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

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
```

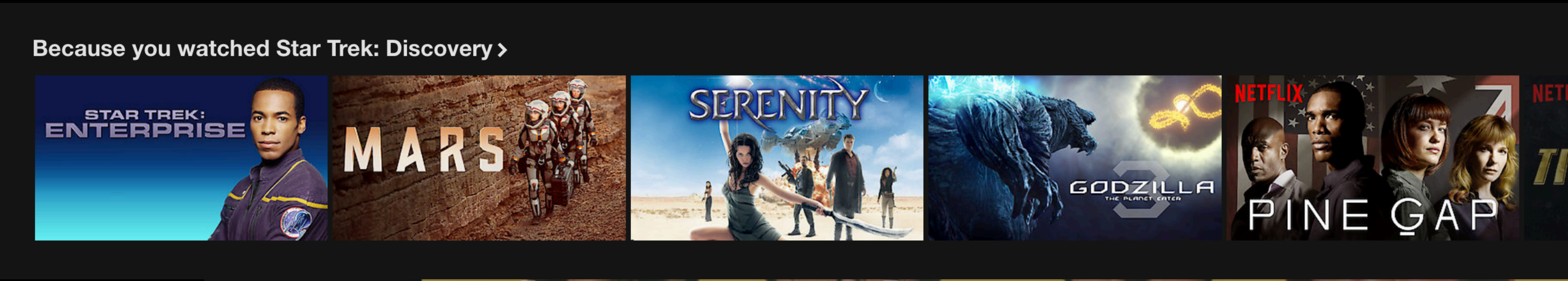
TDM in action

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

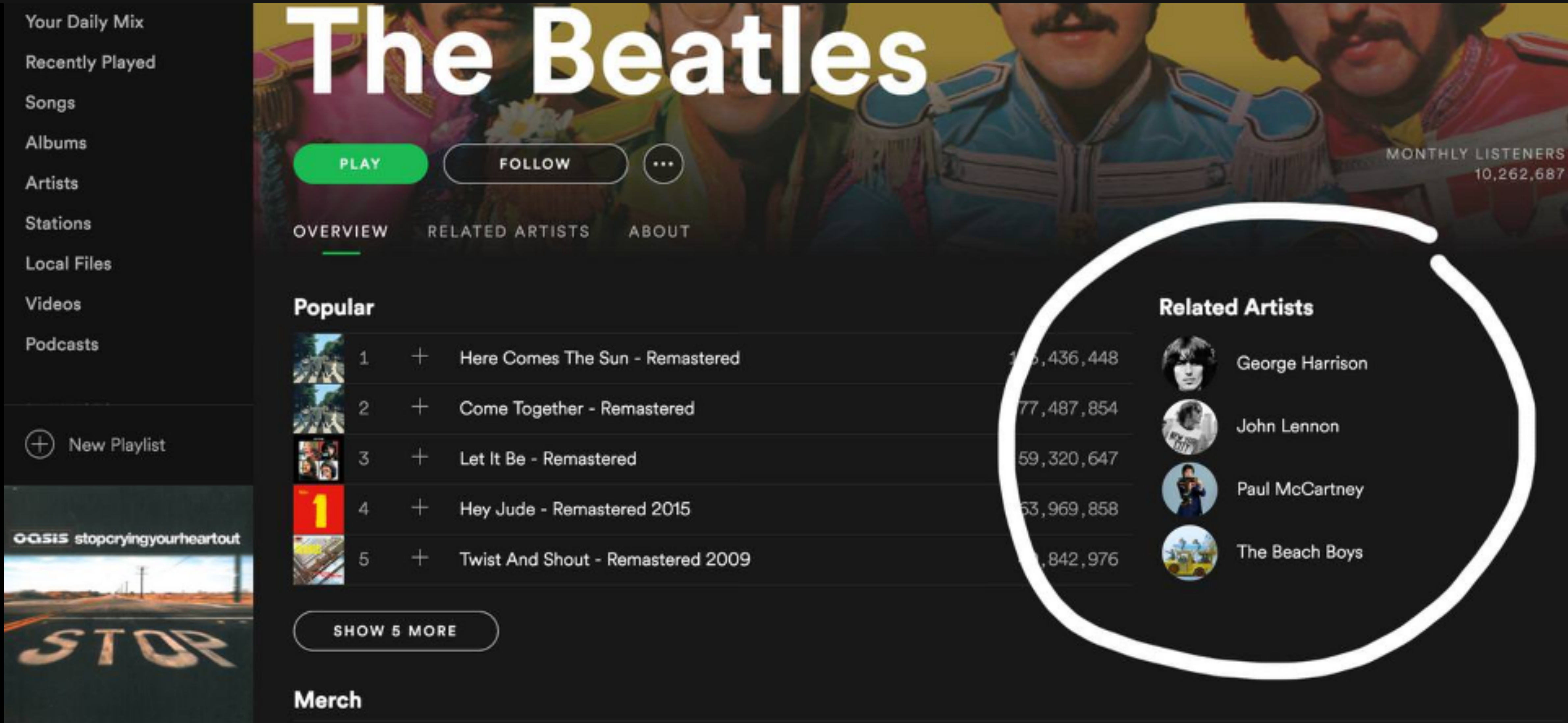
Search

- 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 >



Popular

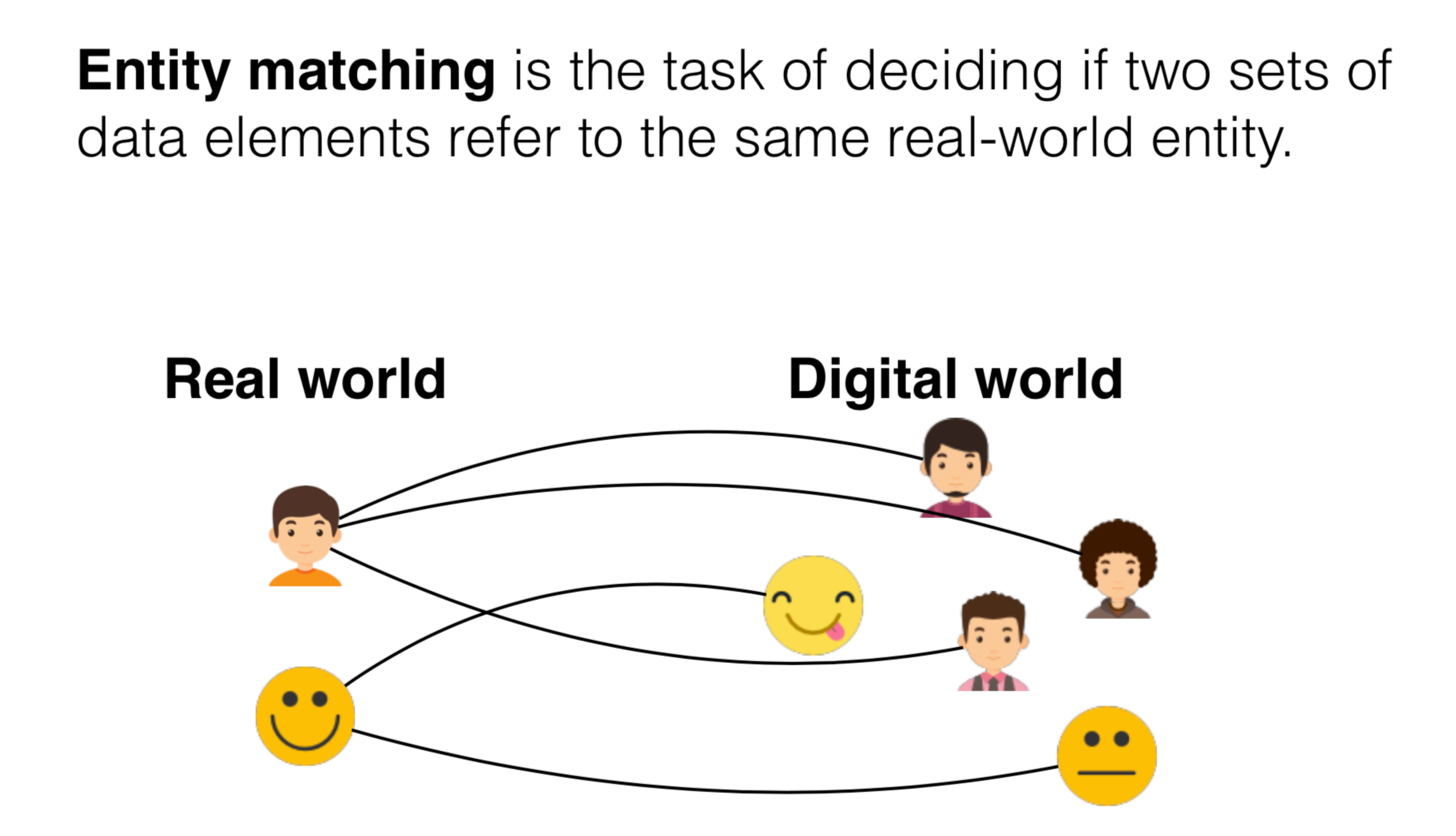
1	+	Here Comes The Sun - Remastered	1 ,436,448
2	+	Come Together - Remastered	77,487,854
3	+	Let It Be - Remastered	59,320,647
4	+	Hey Jude - Remastered 2015	63,969,858
5	+	Twist And Shout - Remastered 2009	,842,976

Merch

Related Artists

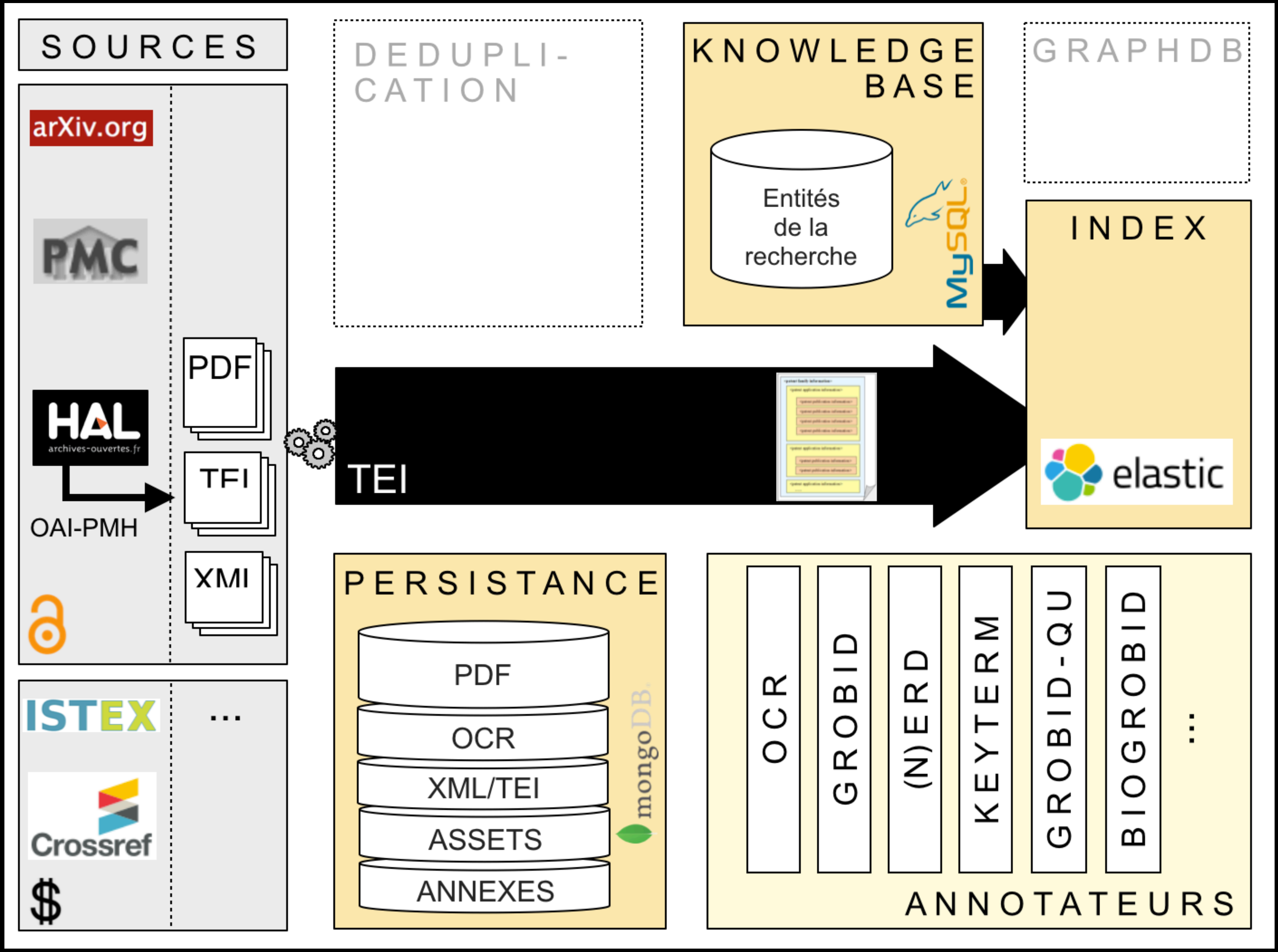
- George Harrison
- John Lennon
- Paul McCartney
- The Beach Boys

Disambiguation



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

Inria anHALytics



SOURCES

arXiv.org

PMC

HAL archives-ouvertes.fr

OAI-PMH

PDF

TFI

XMI

ISTEX

Crossref

$

...

DEDUPLICATION

KNOWLEDGE BASE

Entités de la recherche

MySQL

GRAPHDB

INDEX

elastic

TEI

PERSISTANCE

PDF

OCR

XML/TEI

ASSETS

ANNEXES

mongoDB

OCR

GROBID

(N)ERD

KEYTERM

GROBID-QU

BIOGROBID

...

ANNOTATEURS

Demo

Thank you