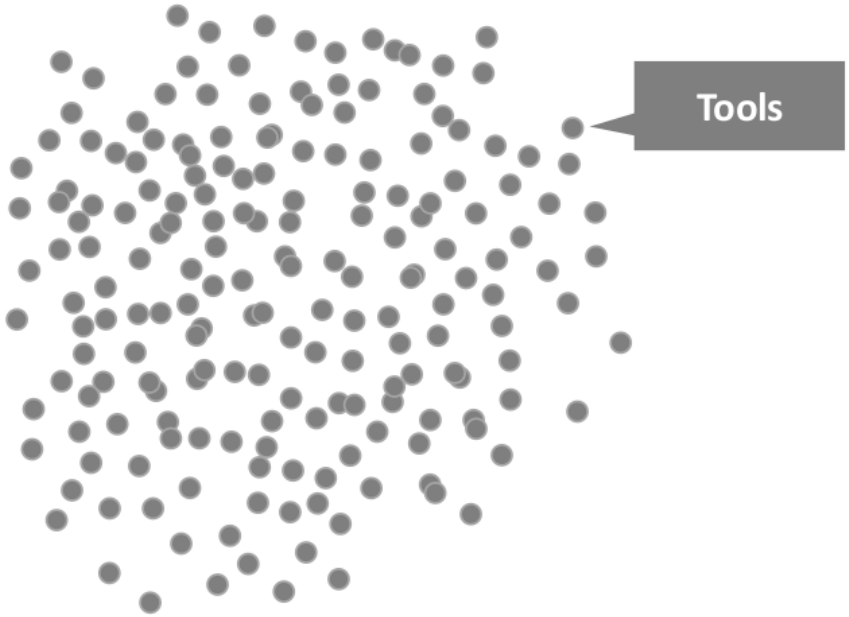
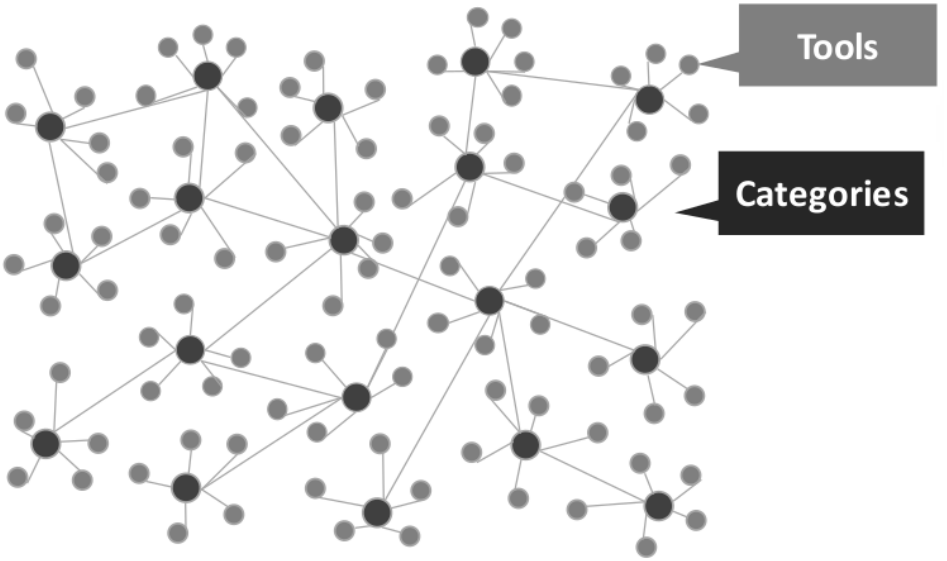
Intelligent analysis of scientific documents

Emeric DYNOMANT, Data scientist & PhD candidate

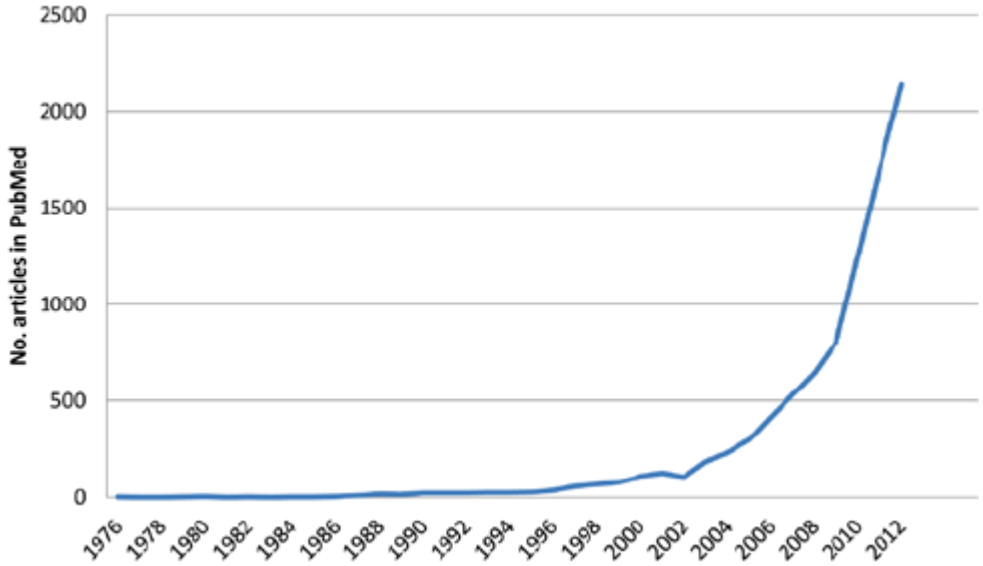
OMIC TOOLS

OMICTools, the meeting point for biology and informatics





Our goal: ultrafast document parsing and classification



Source: researchgate.com

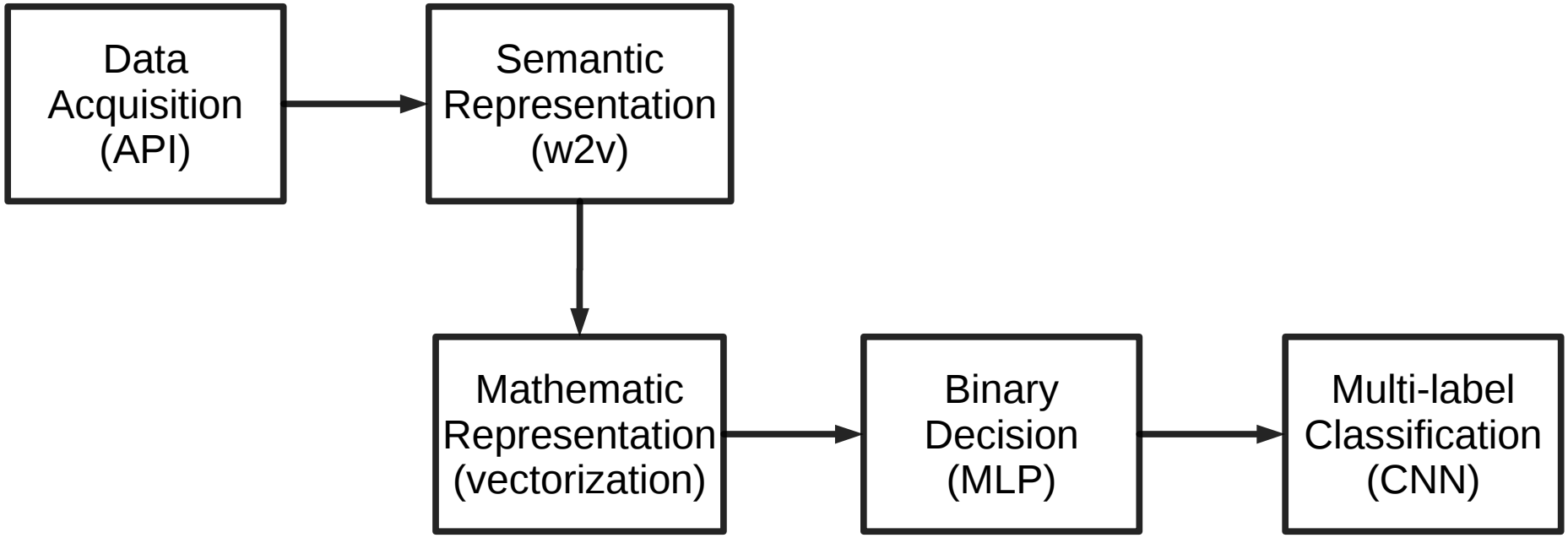
Text analysis steps:

1- Metadata extraction

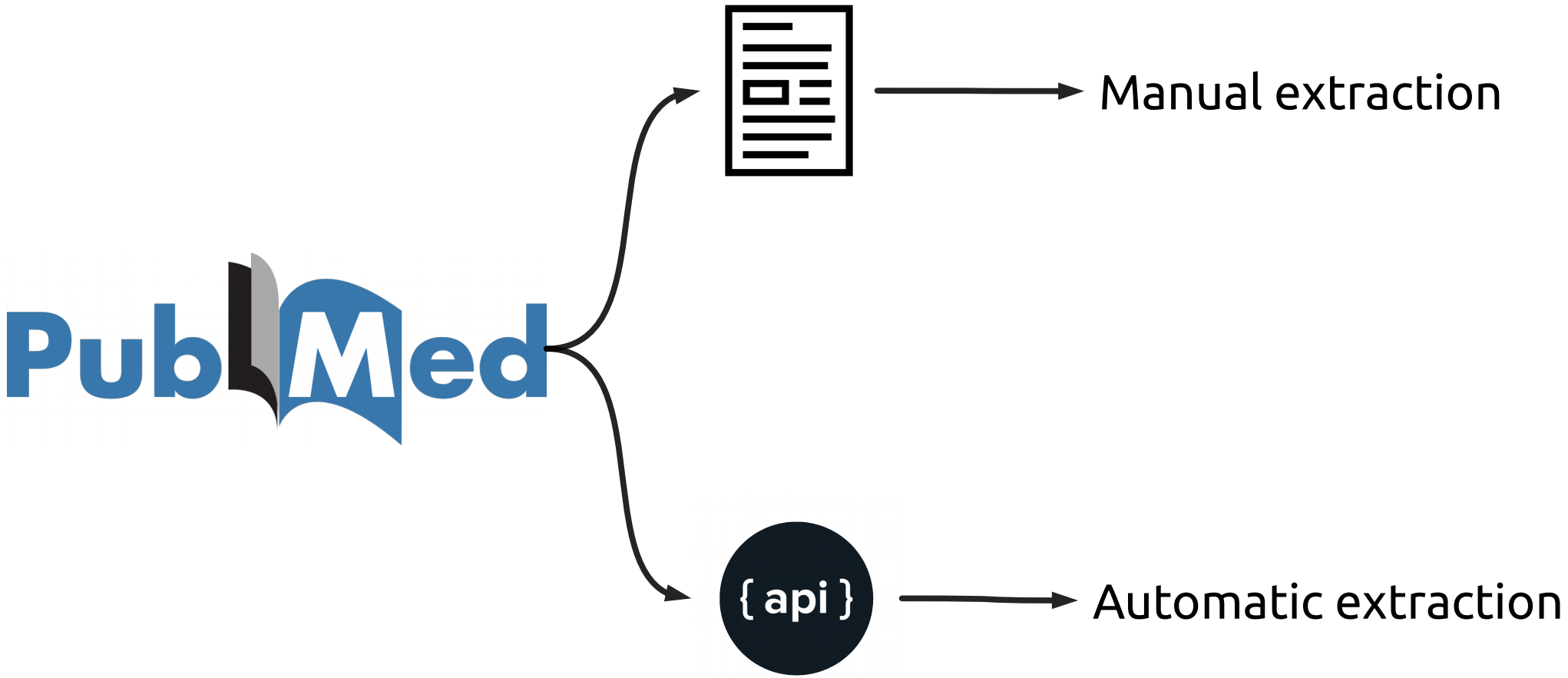
2- Semantic representation

3- Classification

Workflow proposed



Step 1: data mining



Step 1: data mining

Raw metadata

- PMID / DOI / ISSN
- Volume / issue
- Title
- Abstract
- MESH
- Chemicals
- Databanks
- Grants

Authors

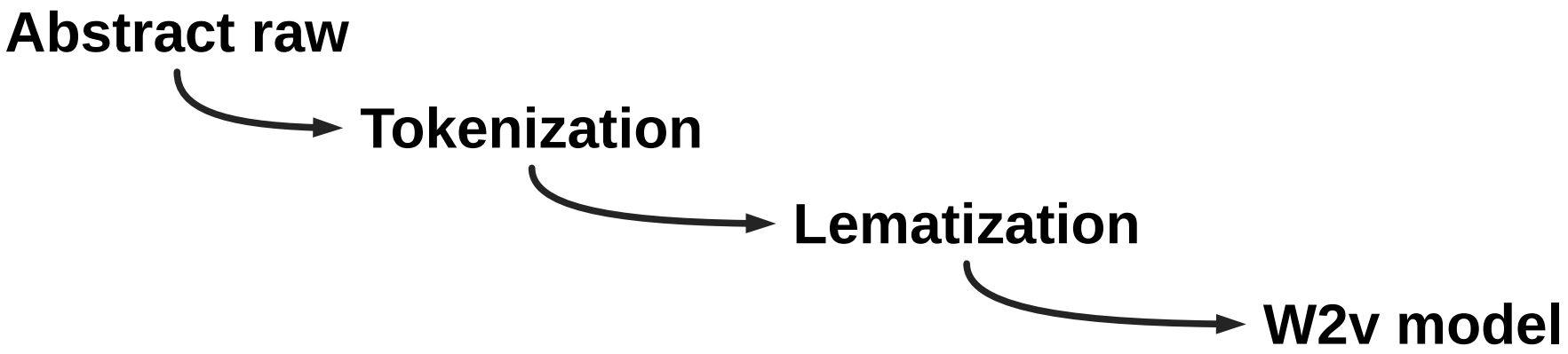
- Specific API designed
- Clean affiliations (21 levels)

Step 1: data mining


```
"affiliations": [
  {
    "country": "UK",
    "level_1": {
      "center": [
        "Li Ka Shing Centre"
      ],
      "institut": [
        "gridCRUK Cambridge Institute and Department of Oncology"
      ],
      "university": [
        "University of Cambridge"
      ]
    },
    "level_2": {
      "department": [
        "gridCRUK Cambridge Institute and Department of Oncology"
      ]
    },
    "not_matched": [
      "CB RE UK"
    ],
    "other": {
      "adress": [
        "Robinson Way"
      ]
    },
    "raw": "gridCRUK Cambridge Institute and Department of Oncology, University of Camb
  }
],
"corresponding": "True",
"email": [
  "carlos.caldas@cruk.cam.ac.uk"
],
"fore_name": "Carlos",
"last_name": "Caldas"
}
```


PMID: 29304755

Step 2: Indexation



Important features are “mapped”:

- EDAM ontology
- OmicX’s termnology

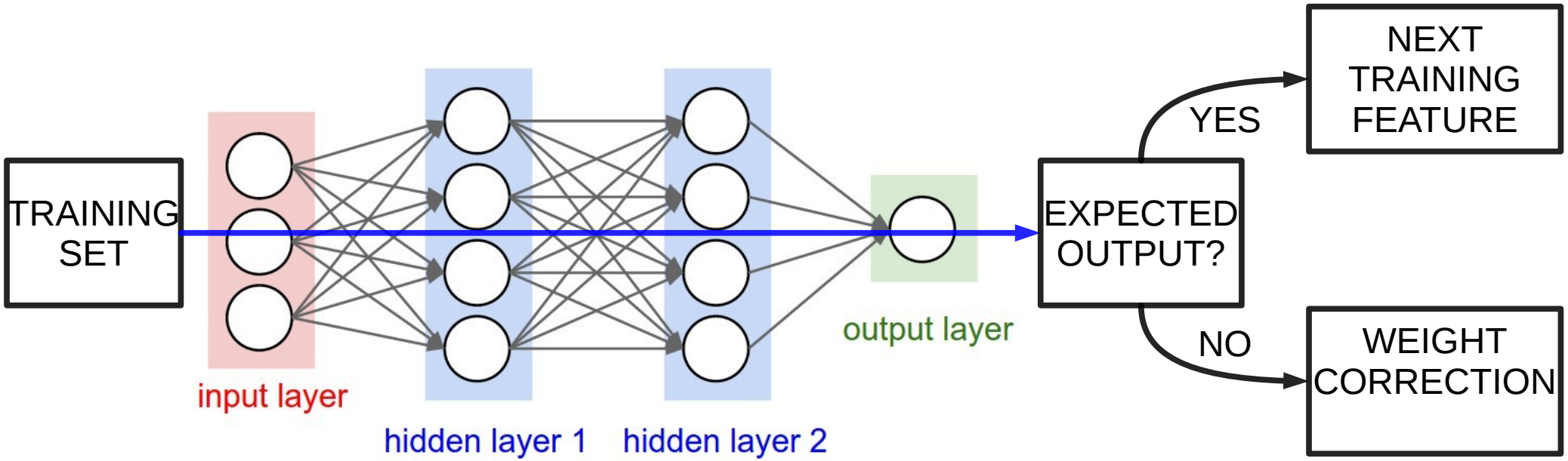
- Other futures (ICD-10, GO, Uniprot ...)

Step 3: Vectorization

PMID	VECTOR	PREDICTION
15369363	5,-10,0,0,0,0,0,0,0,0,0,-5,0,0	0
23303508	5,10,6,0,15,0,0,0,0,15,20,5,0,0	1
23303335	5,-10,0,0,0,0,0,0,0,0,0,-5,0,0	0
18503082	5,10,6,0,0,3,30,20,10,0,10,5,0,0	1
18503281	5,-10,0,0,0,0,0,0,0,0,10,-5,0,5	0
23892897	5,10,0,0,0,3,10,0,0,0,10,5,0,0	1
23892781	-5,0,0,0,0,0,0,0,0,0,0,5,0,0	0
19875695	-5,0,0,0,3,0,0,0,0,0,0,5,0,5	1
19876280	-5,0,0,0,0,0,0,0,0,0,0,-5,0,0	0

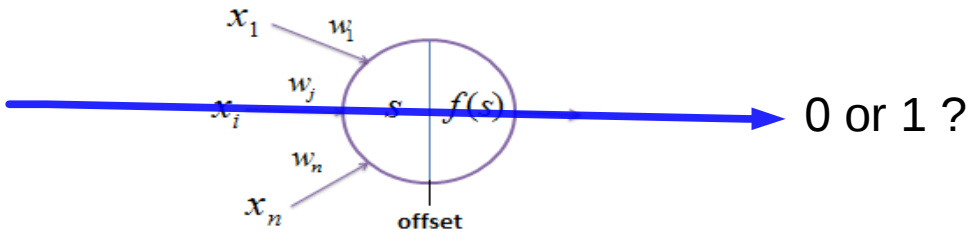
ITEM 1: score_tool coming from the ":" in the article's title
ITEM 2: if there's less than 3 words before the ":"
ITEM 3/4: finding a programming language into the abstract (3) or into the title (4)
ITEM 5/6: programming vocabulary in the abstract (5) or in the title (6)
ITEM 7/8: word in capital letters into the title (7) and/or abstract (8)
ITEM 9/10: finding a link (9) / code repo (10) in the abstract
ITEM 11: words contained into the title/abstract
ITEM 12: is journal name in OMICTools'top 100 ?
ITEM 13: common DB names
ITEM 14: institute match good roots

Break: what is suppervised deep learning?



$[X_1, X_2, X_3, \ldots X_n]$

$[W_1, W_2, W_3, \ldots W_n]$

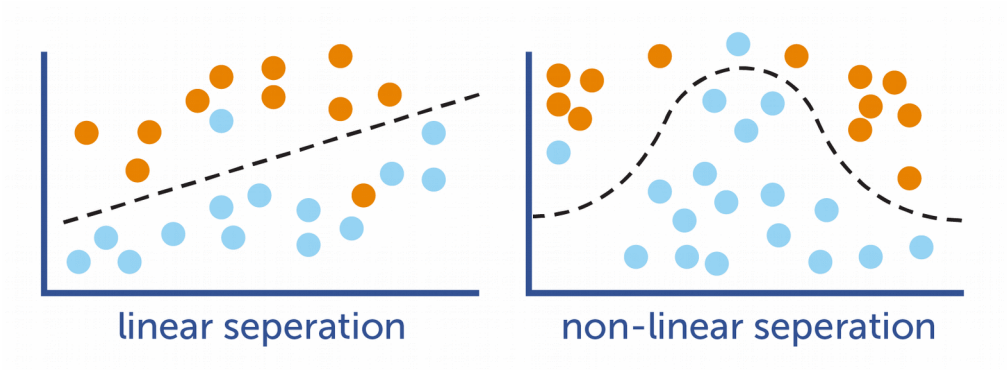


0 or 1 ?

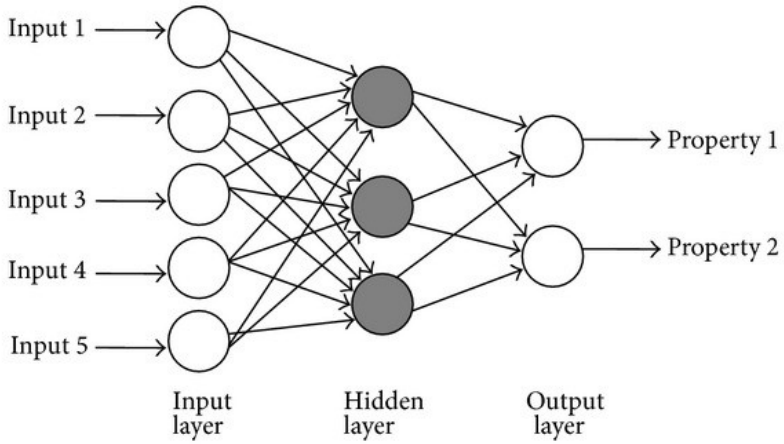
Summation $s = \sum w \cdot x$

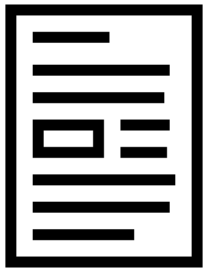
Transformation $f(s) = \frac{1}{1+e^{-s}}$

Step 4: Decision



MLP allow to separate data non linearly



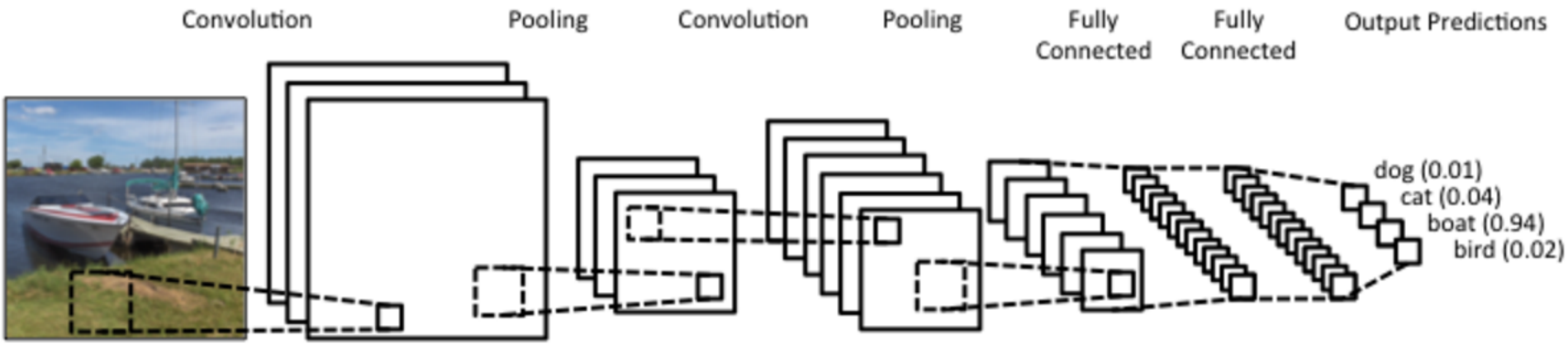


Training:
- 23 275 articles from group 1
- Idem for group 0 (ID created)

MLP

Group 0: not bioinformatics
> Next article
Group 1: is bioinformatics
> Next step

Break: What is convolution?



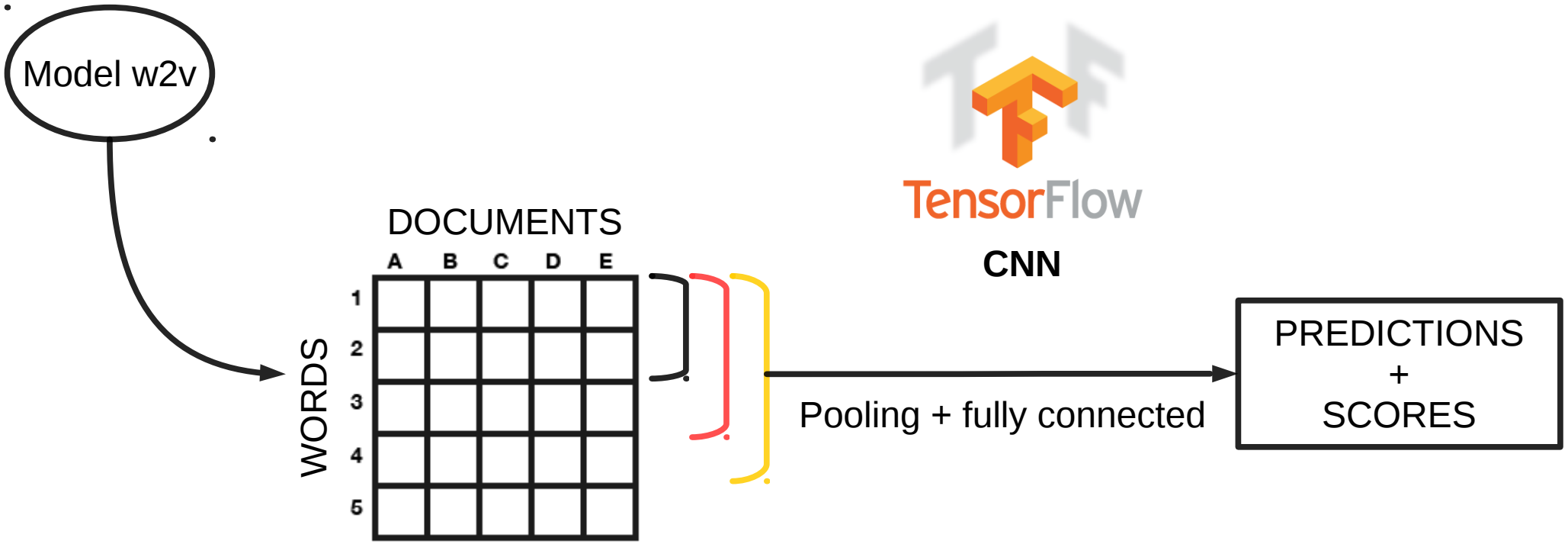
- Biologically inspired (visual cortex)

- Reduce complexity and analysis time

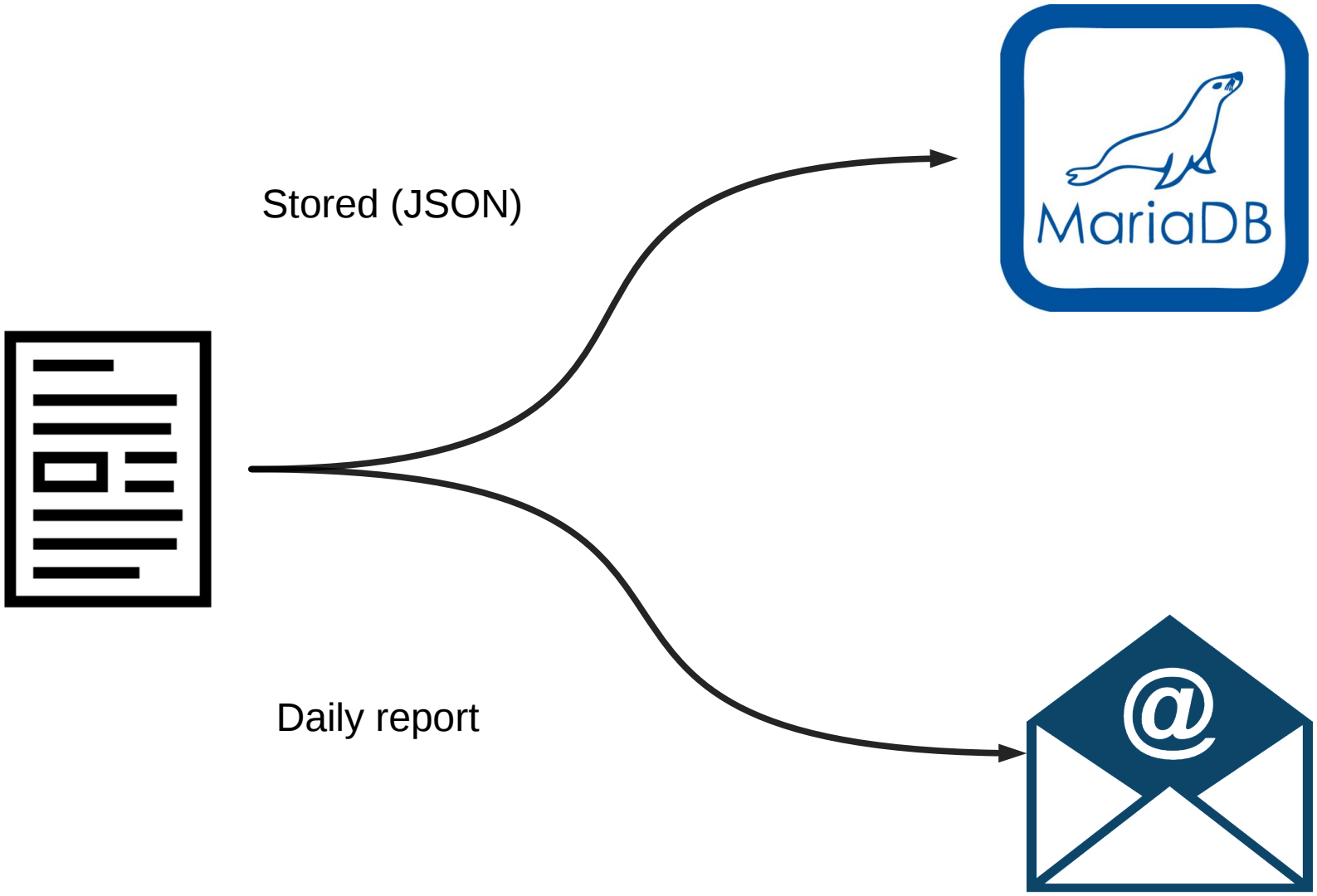
- Very deep neural networks

Step 5: Classification

Suppervised (23 275 PMID already classified)



Step 6: DB storage and report



Next Steps

- **SVM** instead of MLP
- **Deep learning** for mapping + generic block
- **Pre-prints** analysis
- Study on **other corpus** (medical..)

→ Stronger decision

→ Comparison vs. ECMT

→ BiorXiv, arXiv ...

→ Binary decision + classification

Thanks!

Emeric DYNOMANT

emeric.dynomant@omictools.com

http://omictools.com

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

litis

D2IM Département d'Informatique et d'Information Médicales