Data science for “Caisse Nationale des Allocations Familiales” (CNAF)

Supervisors : Clémentine NEMO, Luc CHAMPALLE

Tutor : Lionel FILLATRE

09-2017

Atos

Introduction

Context

- Enterprise: Atos
  - Informatics services consultancy

- Client: CNAF
  - 50 Atos collaborators working for them

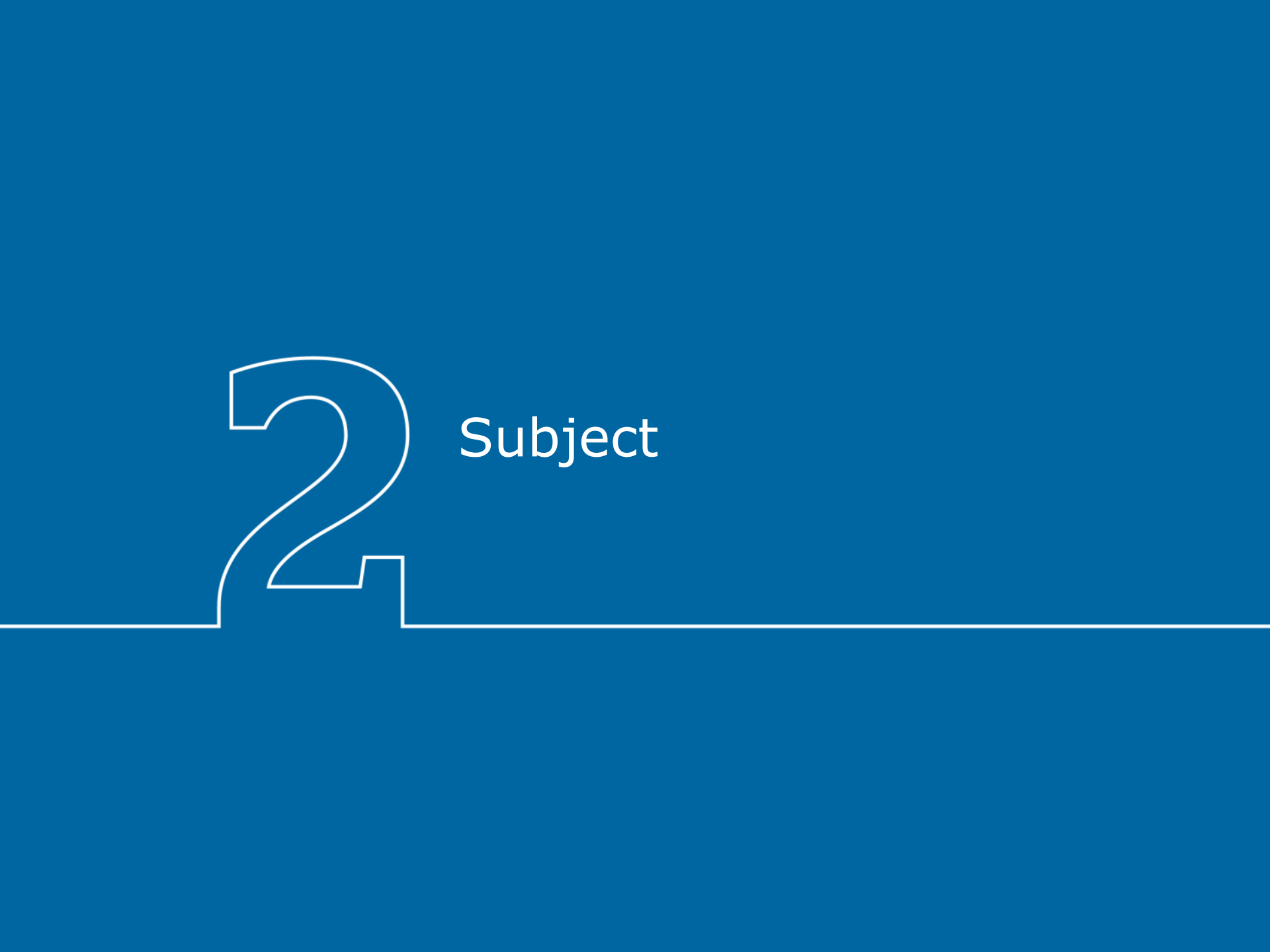
- 1st semester:
  - ½ time in a team → functional work
  - ½ time autonomous work for Data Science integration

- 2nd semester
  - Data Science project for CNAF

- No Data Science at Atos Sophia → **Technically autonomous**

NOT DETAILED

NOT DETAILED


2
Subject

From Open Data to ticketing system

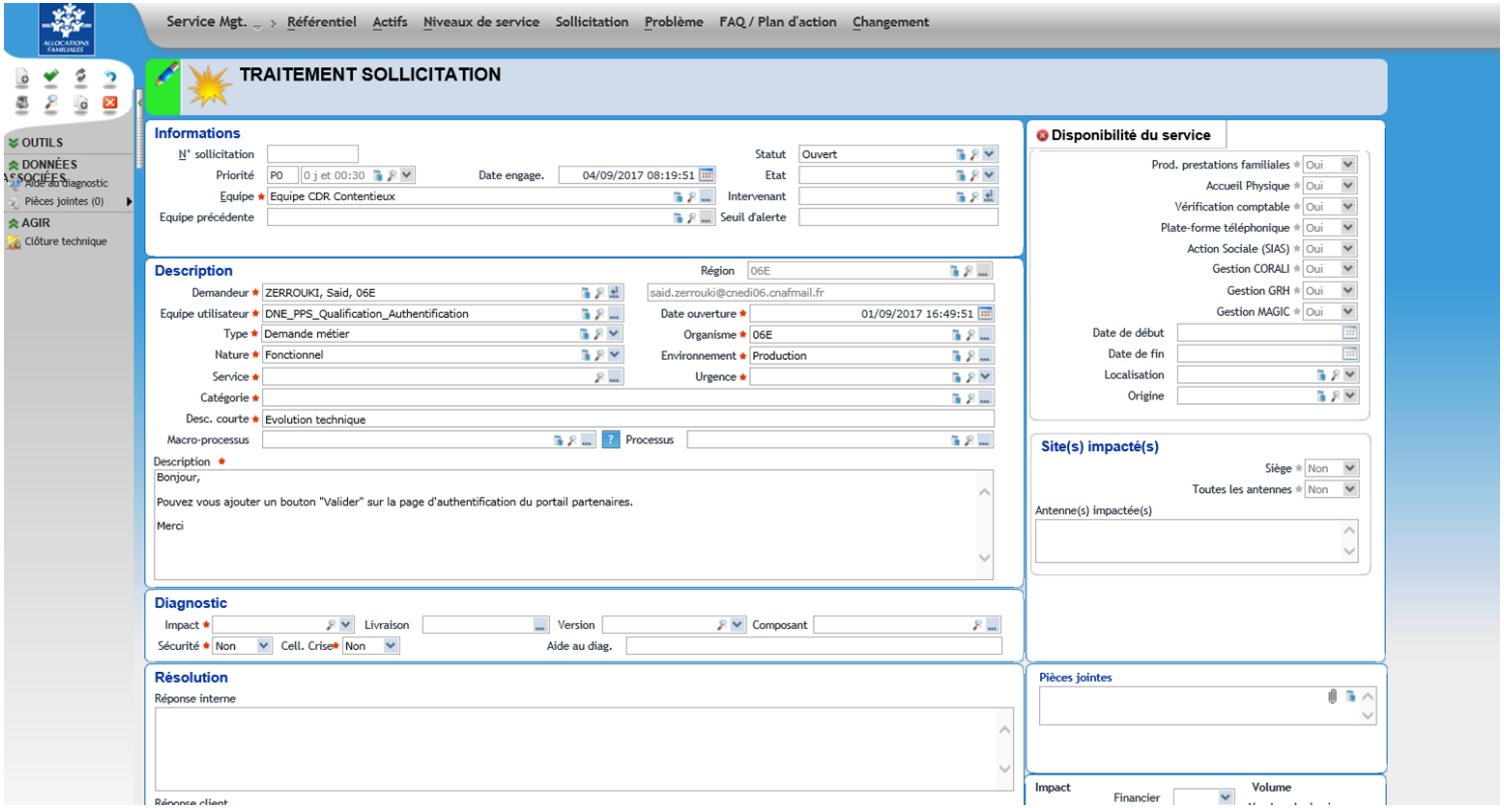
- 1st semester: Open data CNAF (data.caf.fr)

- Proposition open data:
  - Data visualizations
  - Improve open data with semantic web
  - Emotion analysis on tweets

- During meeting presenting solutions:
  - Visualizations & improving Open Data → Team already work on it
  - Tweets → Text analysis → data with text → Ticketing system

Subject

- Subject:
  - « Analysis of CNAF ticketing system»



Issue

- 300 000 tickets on 5 years

- Issue: «Overwork on the system. Improvement possible thanks to Data Science ?»

- **Problems identified thanks to visualizations**
- Multiple duplicates and not always identified or sometimes late in the process

Solution → SIMILARITY PREDICTION

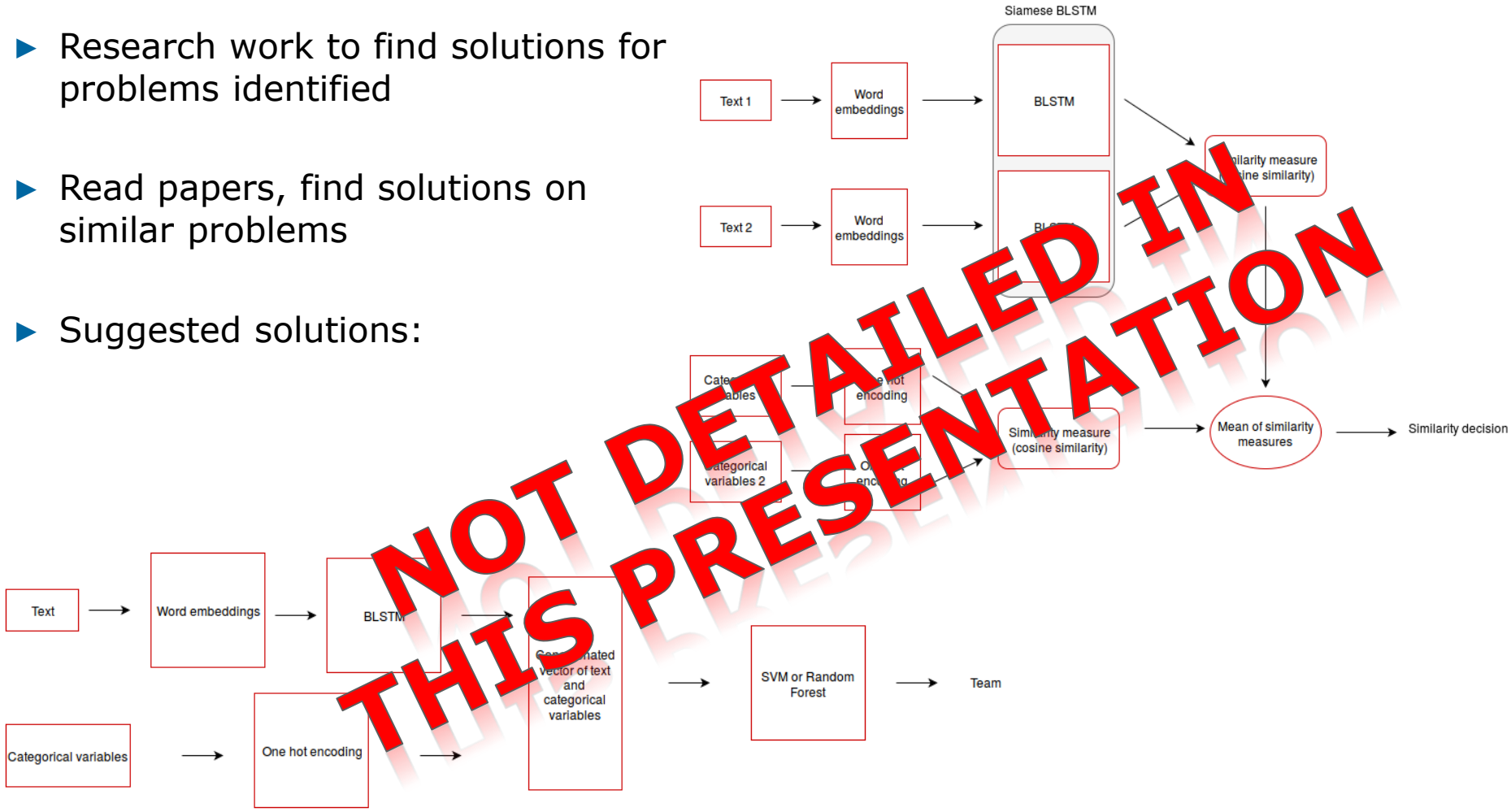
- Process to redirect ticket to team able to solve the ticket.
- Redirection is not always easy and some tickets are long to solve because of it.

Solution → TEAM PREDICTION

The purpose of every solution suggested is to help decision

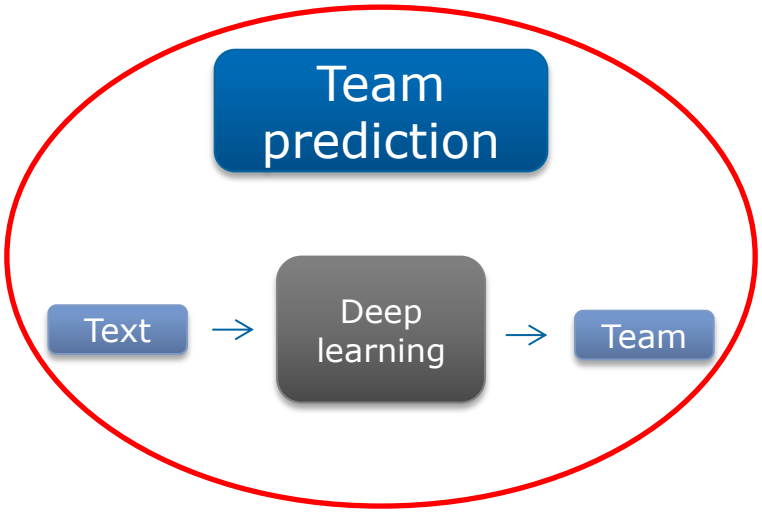
From ideas to realizations

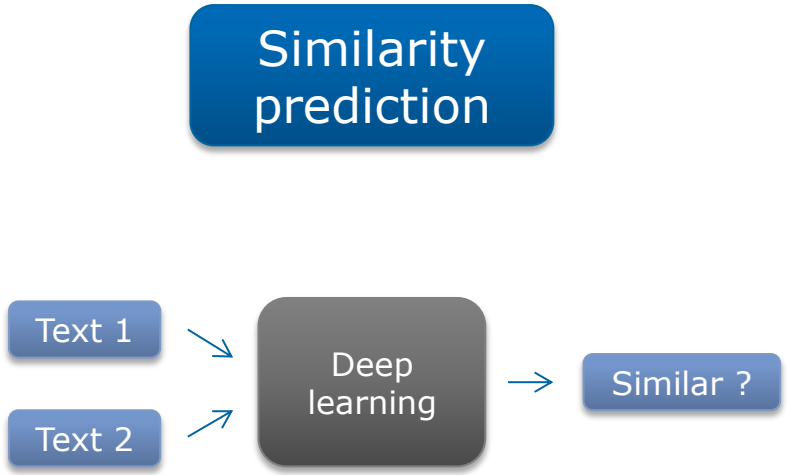
- Research work to find solutions for problems identified
- Read papers, find solutions on similar problems
- Suggested solutions:



From ideas to realizations

- Meeting with CNAF statisticians → Focus on text only to reuse it
- Suggested solutions:





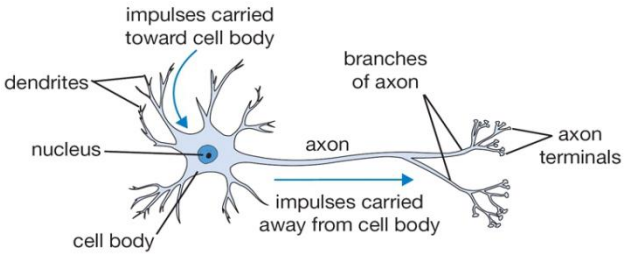
Scientific approach

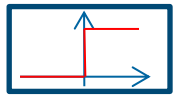
- Why deep learning?
  - Better results
  - Known algorithm thanks to previous internship
- CNAF statistician → No knowledge on deep learning
- Comparison between classic approach and deep learning algorithms

3 Technical realization

Deep learning basics

Biological neuron

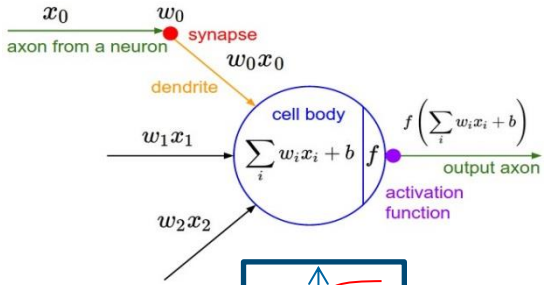


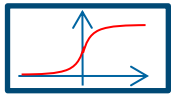


cell body activation

≠

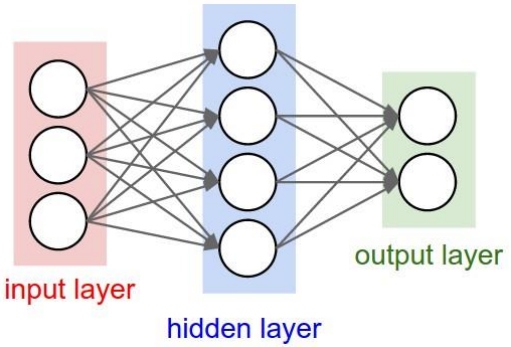
Artificial neuron



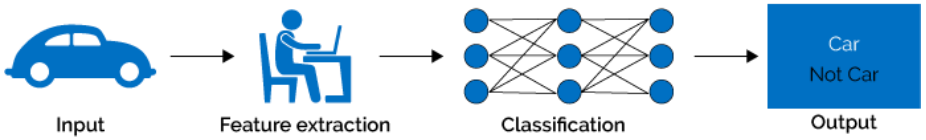


cell body activation

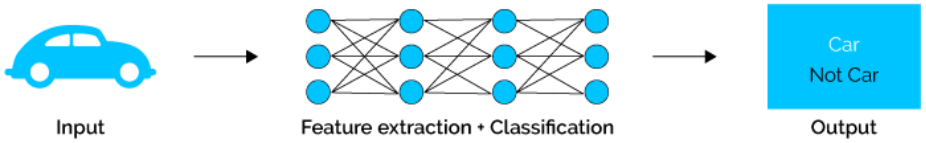
Neural network

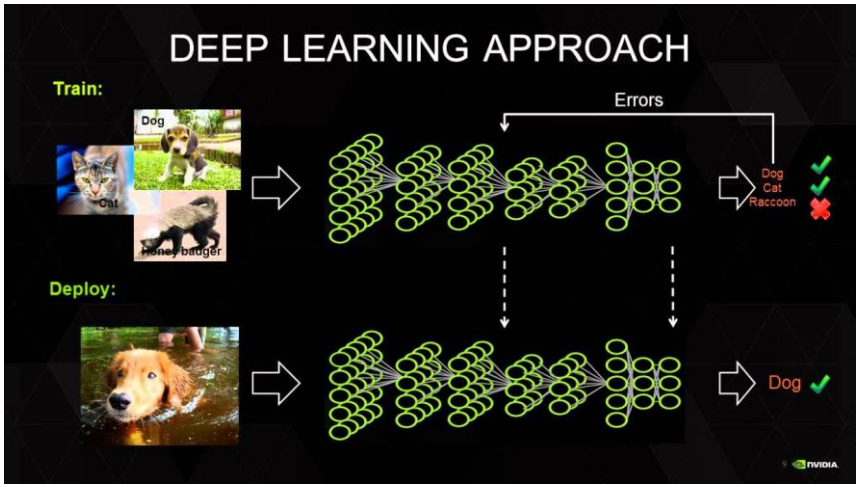


Machine Learning

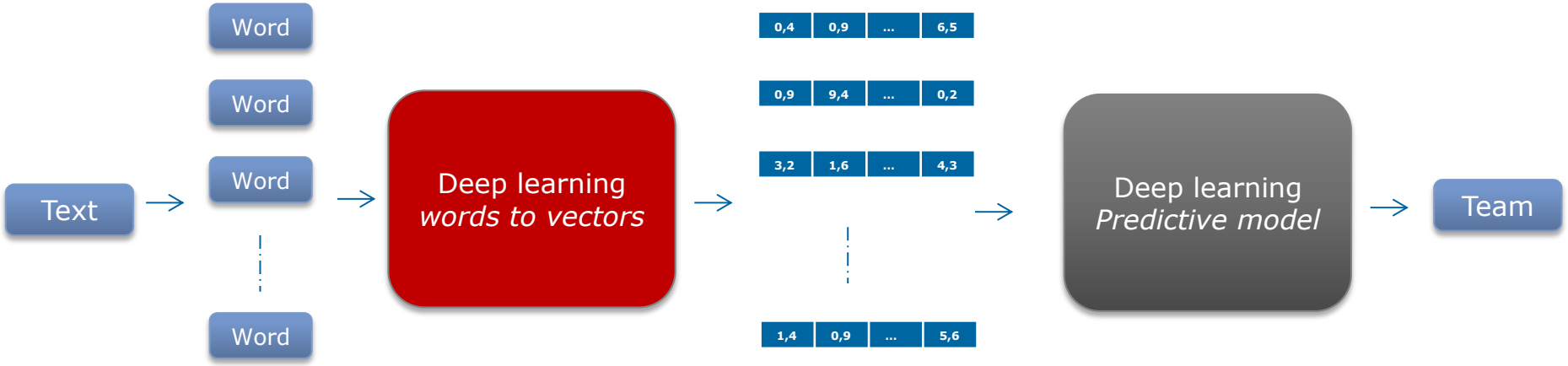


Deep Learning



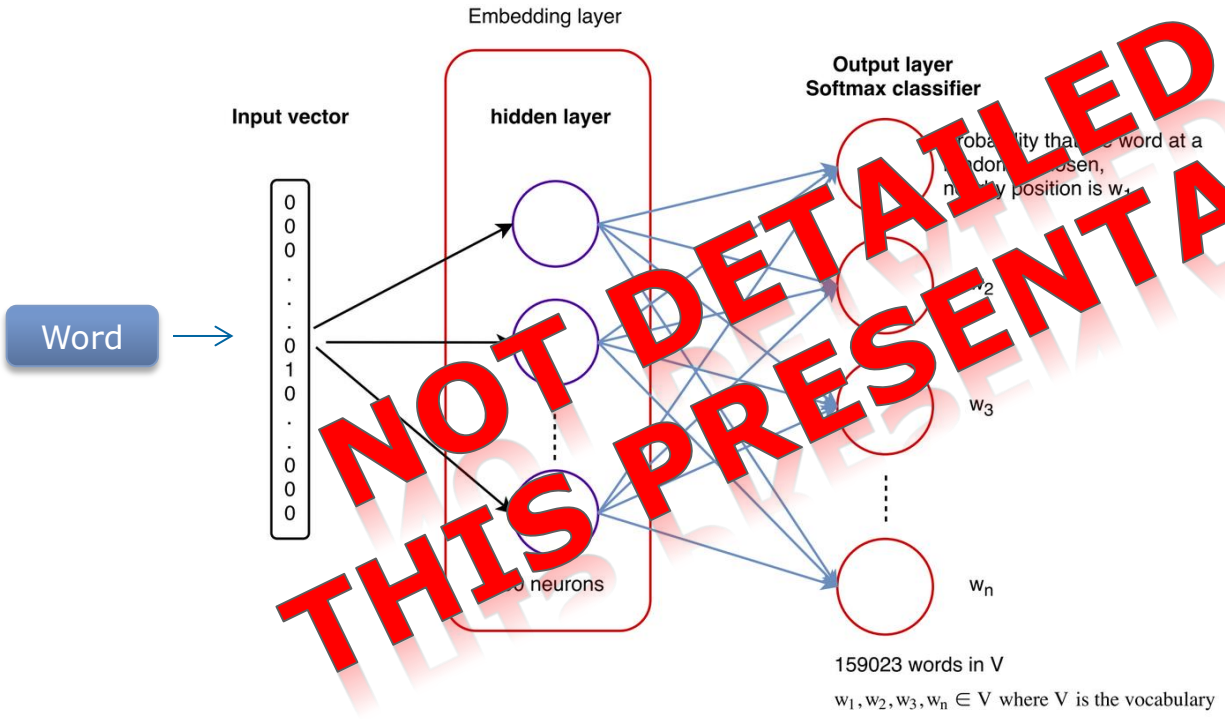


Model


Text
Word
Word
Word
Word
Deep learning
words to vectors
0,4 0,9 ... 6,5
0,9 9,4 ... 0,2
3,2 1,6 ... 4,3
1,4 0,9 ... 5,6
Deep learning
Predictive model
Team

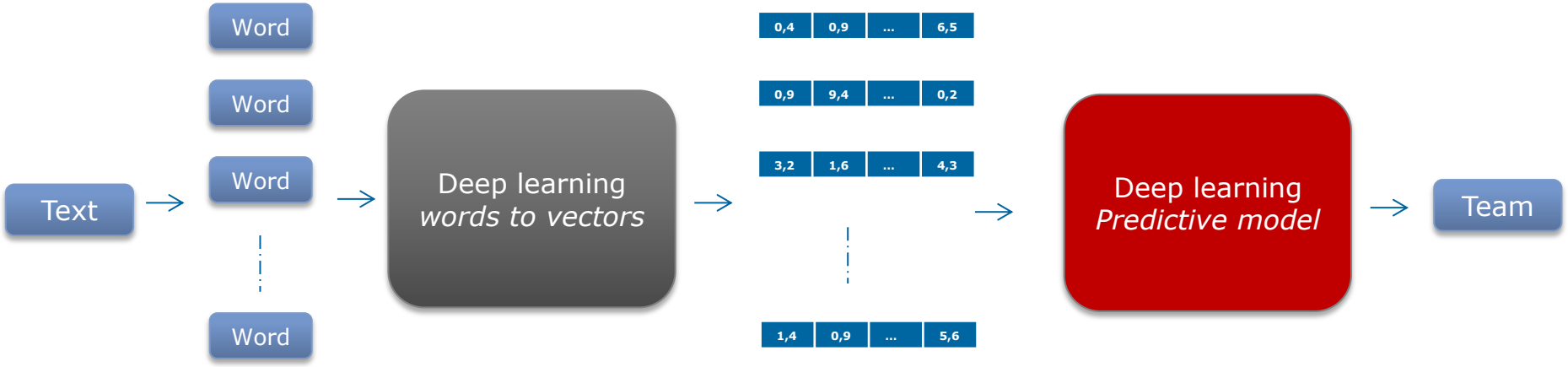
From words to vectors

Embedding model (Skip gram)



- Fake task: predict surrounding words of the input
- hidden layer is the representation by a vector of reals of the input word

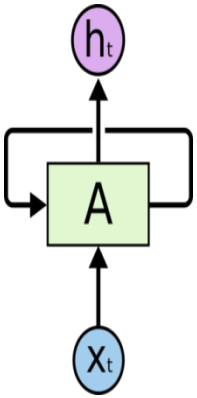
Model


Text
Word
Word
Word
Word
Deep learning
words to vectors
0,4 0,9 ... 6,5
0,9 9,4 ... 0,2
3,2 1,6 ... 4,3
1,4 0,9 ... 5,6
Deep learning
Predictive model
Team

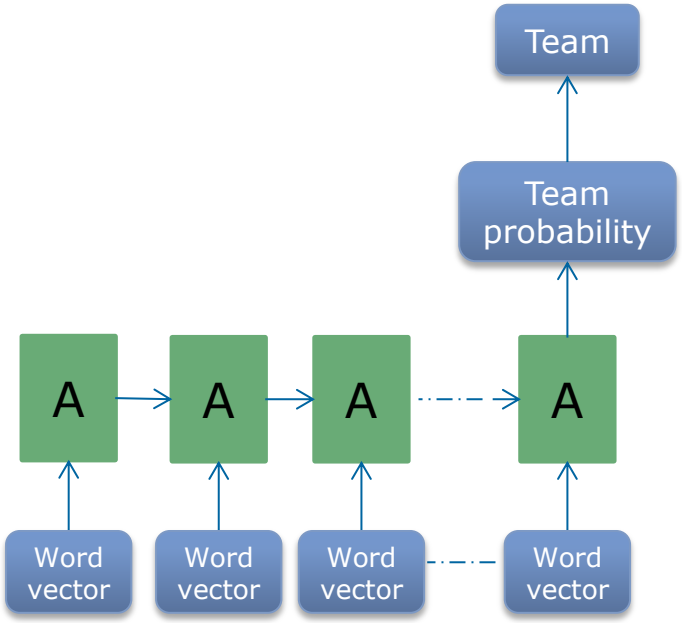
Predictive model

Recurrent Neural Network (RNN) and bidirectional RNN

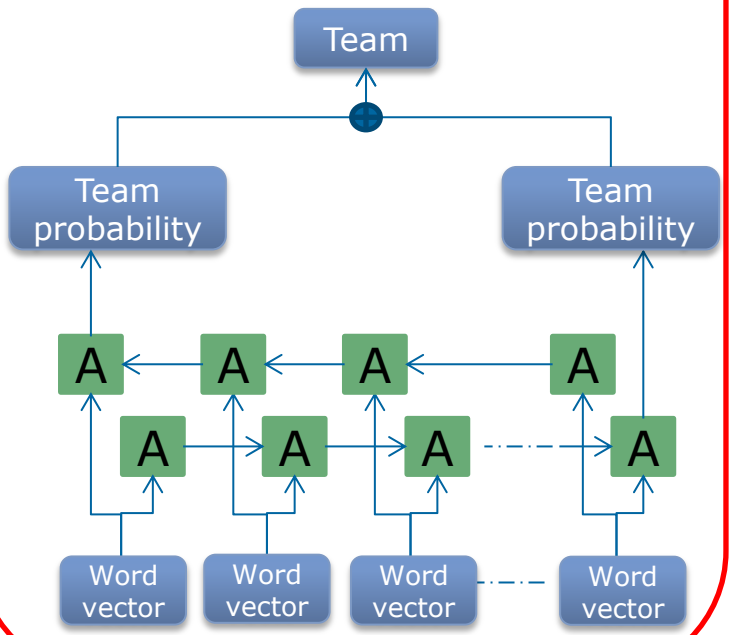
Recurrent neuron



Recurrent neuron unrolled

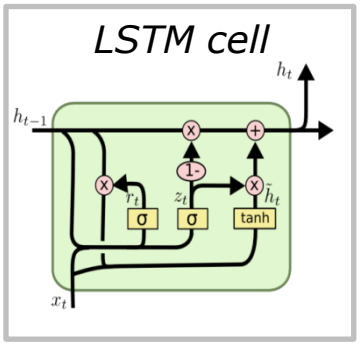


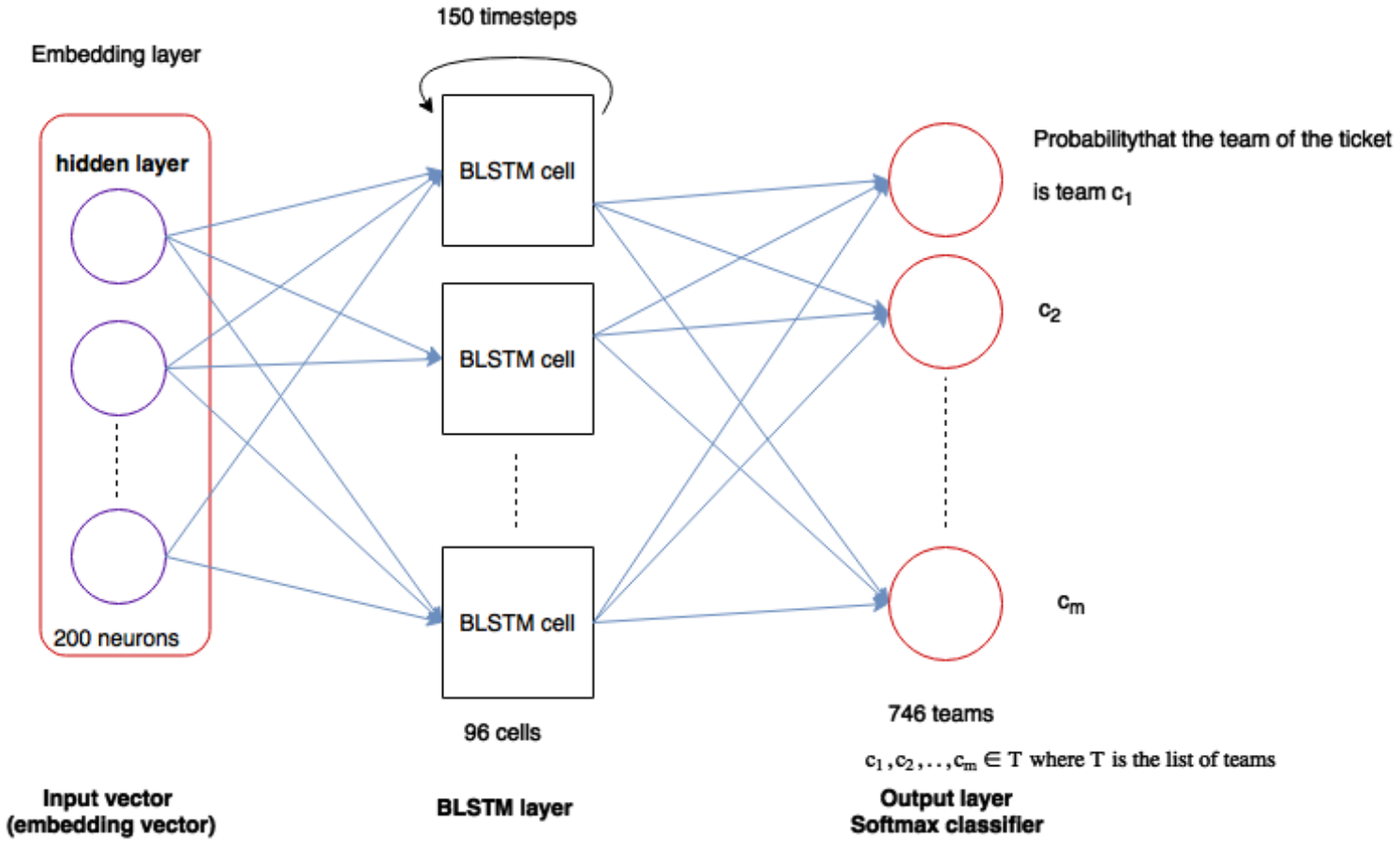
Bidirectional neuron unrolled



Atos

Team prediction



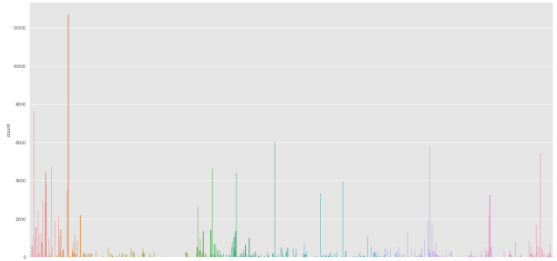


Tricks to improve learning

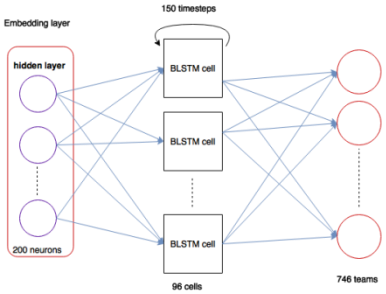
Finetuning

80%

Training data



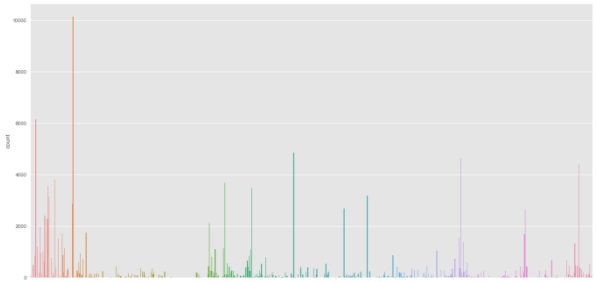
x20



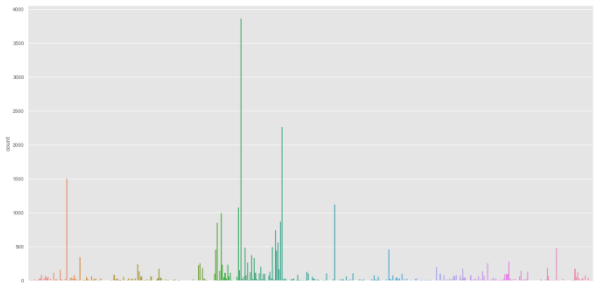
x5

60%

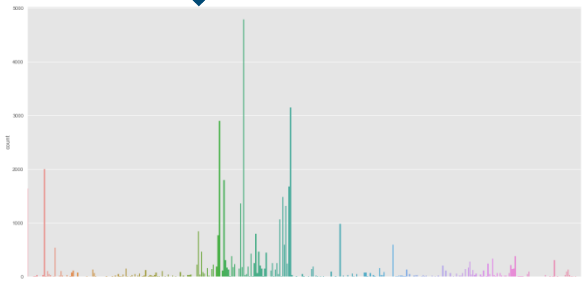
Training data



20% training data



20% Testing data



Atos

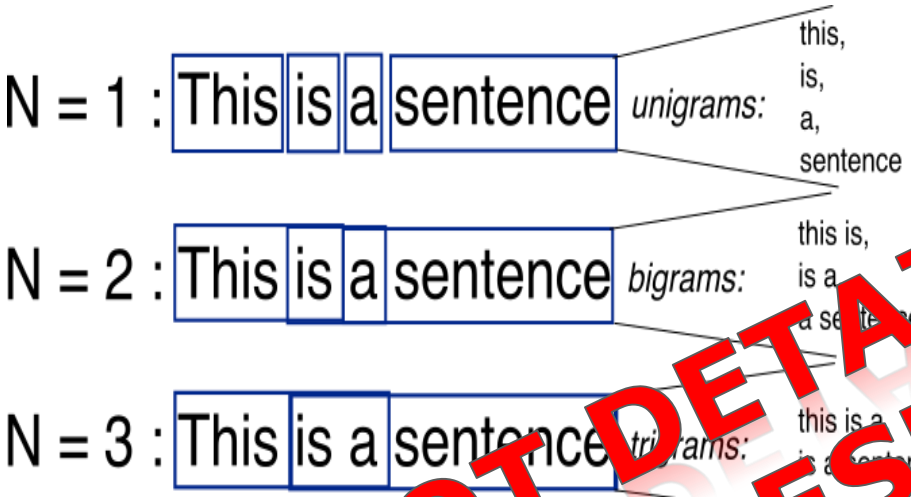
Results

- +700 teams

Model	Train accuracy (%)	Test accuracy (%)
Hazard following distribution	1,84	1.21
Most represented team	7,34	4.54
Traditional approach (no deep learning)	65,61	23.26
Proposed approach before finetuning	51,02	32,07
Proposed approach after finetuning	67,24	47.12

- Good results and research project to be continued…
- Research report available ask by mail: turpaultn@gmail.com
- Github available soon: https://github.com/turpaultn/Text_classification

Traditional method ?



Term-Frequency Metric	Term(1)	Term(2)	Term(3)	Term(4)	Term(5)	Term(6)	Term(n)
Document(1)	225	300	0	0	0	25	0
Document(2)	78	87			0	175	0
Document(3)		13		0	237	0	21
Document(4)		1	101	0	0	0	0
Document(5)	3	15	0	24	0	48	87
Document(6)	0	0	71	0	0	0	0
Document(n)	109	0	901	221	331	441	551

$$\text{tf-idf}_{(t,d)} = \text{tf}_{(t,d)} \times \text{idf}_{(t)}$$

t = term

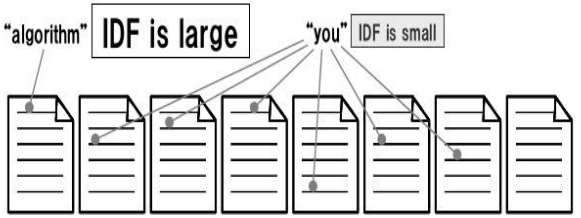
d = document

Inverse Document Frequency (IDF)

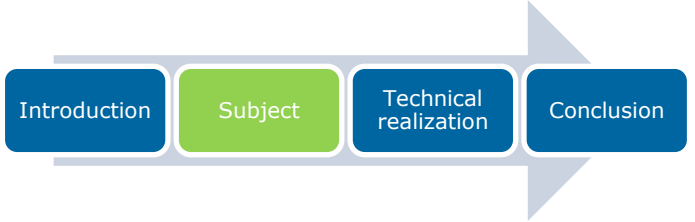
Give more weight to a term occurring in less documents

$$IDF(t) = \log \frac{|D|}{df(t)}$$

t : Term
$df(t)$: Document frequency of t
$|D|$: Number of documents in D



Presentations



- Introduction to Data Science
- Monthly meeting with CNAF (alone from ATOS)

Presentation Deep learning at ATOS with CNAF and ATOS collaborators (~40 pers) *(September, 7th 2017)*

- Introduction to Deep Learning to ATOS and CNAF (video)
- Project presentation to present a concrete use case of Deep Learning (video)
- Feedback on Deep Learning School (video)
- Feed back on Sophia Conf (video)

Conclusion

Experiencing difficulties

- Adapt speech to audience
- Collaborators new to deep learning
- Long to define subject and begin to code the solution
- Evaluate time

Acquired skills

- Technical skills:
  - Developing a concrete solution
  - Research writing
  - Github
  - Define a problem
  - Vulgarization

- Non technical skills:
  - Develop concrete ideas
  - Oral skills
  - adapt speech to audience
  - Vulgarization
  - Consulting work

Conclusion

- Wonderful experience
- I thank ATOS and CNAF for their confidence during this aprenticeship
- Future: PhD at LORIA Nancy:
  - *Deep learning for sound analysis in real environments*

Thanks

For more information please contact:
T+ 33 1 98765432
F+ 33 1 88888888
M+ 33 6 44445678
firstname.lastname@atos.net

Atos