

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

---

Supervisors : Clémentine NEMO, Luc CHAMPALLE

Tutor : Lionel FILLATRE

09-2017

1

# Introduction

- ▶ Enterprise: Atos
  - Informatics services consultancy
- ▶ Client: CNAF
  - 50 Atos collaborators working for them
- ▶ 1<sup>st</sup> semester:
  - ½ time in a team → functional work
  - ½ time autonomous work for Data Science integration
- ▶ 2<sup>nd</sup> 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

Introduction

Subject

Technical  
realization

Conclusion

- ▶ 1<sup>st</sup> semester: Open data CNAF ([data.caf.fr](http://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:

- « Analysis of CNAF ticketing system »

The screenshot displays the 'TRAITEMENT SOLLICITATION' (Ticket Processing) interface of the CNAF ticketing system. The interface is organized into several sections:

- Informations:** Contains fields for 'N° sollicitation', 'Priorité' (set to PO), 'Date engage.' (04/09/2017 08:19:51), 'Statut' (Ouvert), 'Equipe' (Equipe CDR Contentieux), and 'Equipe précédente'.
- Description:** Includes 'Demandeur' (ZERROUKI, Said, 06E), 'Equipe utilisateur' (DNE\_PPS\_Qualification\_Authentication), 'Type' (Demande métier), 'Nature' (Fonctionnel), 'Service', 'Catégorie', 'Desc. courte' (Evolution technique), and 'Macro-processus'. It also shows 'Région' (06E), 'Date ouverture' (01/09/2017 16:49:51), 'Organisme' (06E), 'Environnement' (Production), and 'Urgence'.
- Diagnostic:** Features 'Impact' (set to Non), 'Sécurité' (set to Non), 'Cell. Crise' (set to Non), 'Version', 'Composant', and 'Aide au diag.'.
- Résolution:** Includes a 'Réponse interne' field.
- Disponibilité du service:** A section with multiple dropdown menus for service availability, including 'Prod. prestations familiales', 'Accueil Physique', 'Vérification comptable', 'Plate-forme téléphonique', 'Action Sociale (SIAS)', 'Gestion CORALI', 'Gestion GRH', and 'Gestion MAGIC'. It also has fields for 'Date de début', 'Date de fin', 'Localisation', and 'Origine'.
- Site(s) impacté(s):** Includes 'Siège' (set to Non) and 'Toutes les antennes' (set to Non).
- Pièces jointes:** A section for attachments.
- Impact:** A section with 'Financier' and 'Volume' fields.

The interface also features a sidebar on the left with 'OUTILS' and 'DONNÉES' sections, and a top navigation bar with links like 'Service Mgt.', 'Référentiel', 'Actifs', 'Niveaux de service', 'Sollicitation', 'Problème', 'FAQ / Plan d'action', and 'Changement'.

- ▶ 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

Introduction

Subject

Technical realization

Conclusion

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





# From ideas to realizations

Introduction

Subject

Technical  
realization

Conclusion

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

Team  
prediction

Text

Deep  
learning

Team

Similarity  
prediction

Text 1

Text 2

Deep  
learning

Similar ?

- ▶ 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

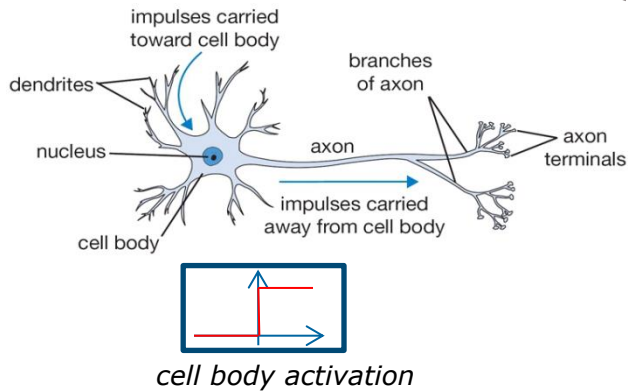
Introduction

Subject

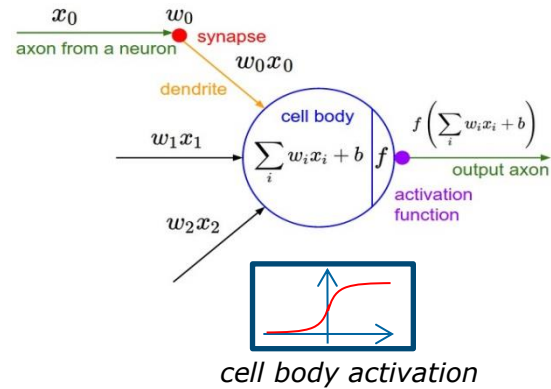
Technical realization

Conclusion

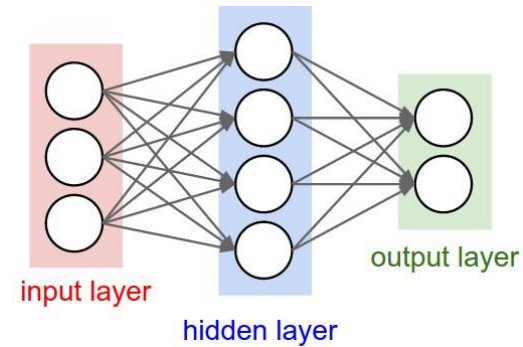
## Biological neuron



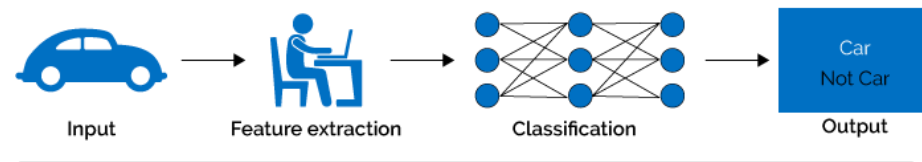
## Artificial neuron



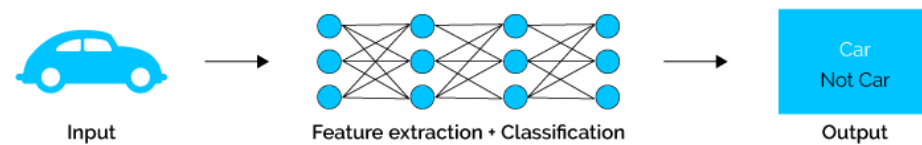
## Neural network



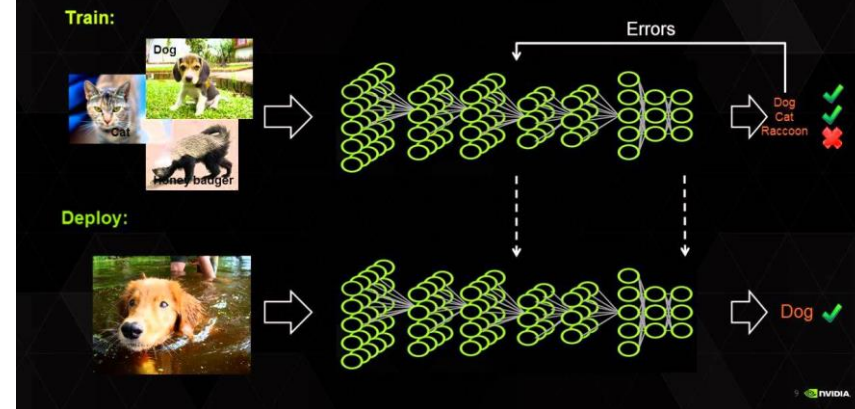
## Machine Learning

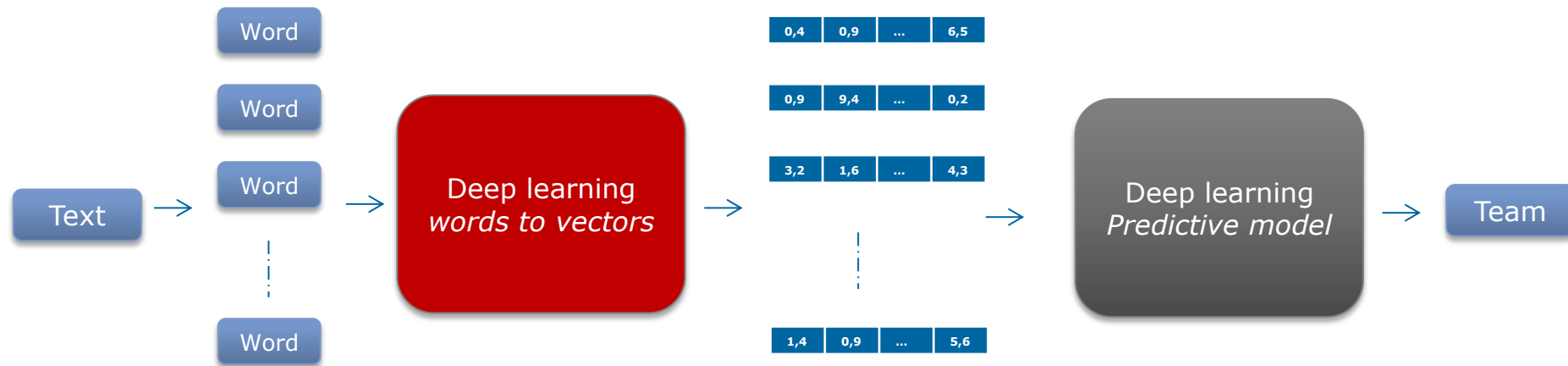


## Deep Learning



## DEEP LEARNING APPROACH





# From words to vectors

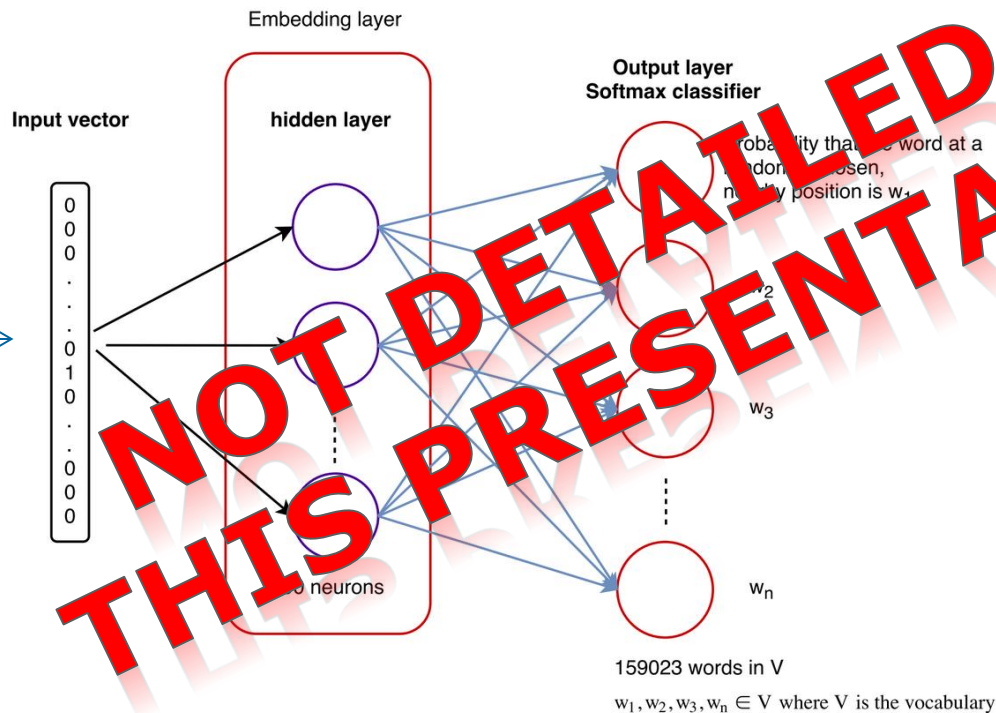
## Embedding model (Skip gram)

Introduction

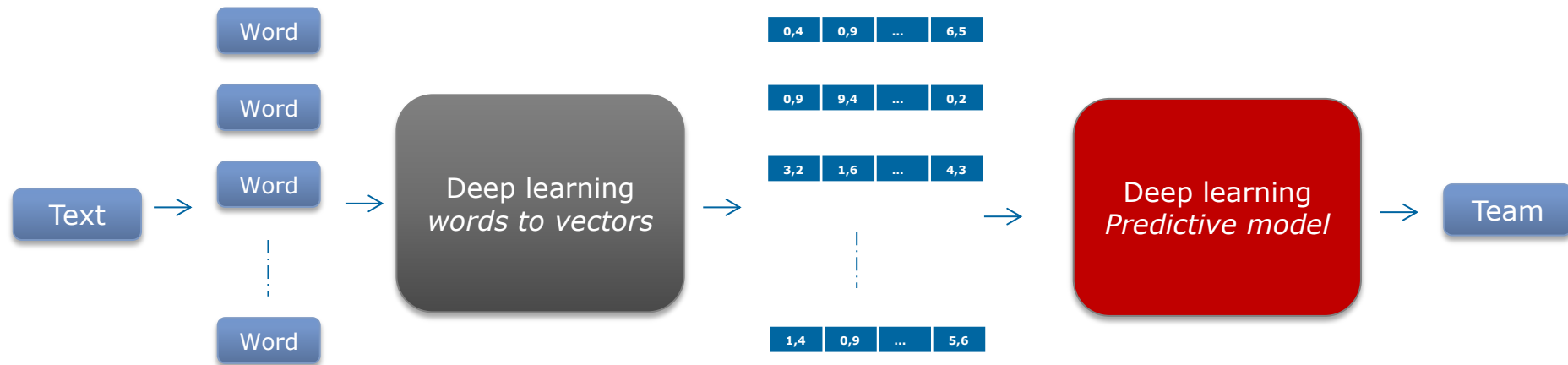
Subject

Technical  
realization

Conclusion

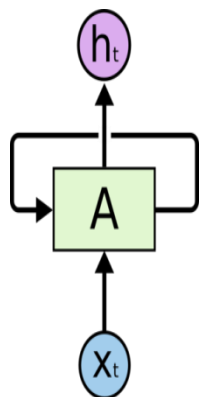


- Fake task: predict surrounding words of the input
- hidden layer is the representation by a vector of reals of the input word

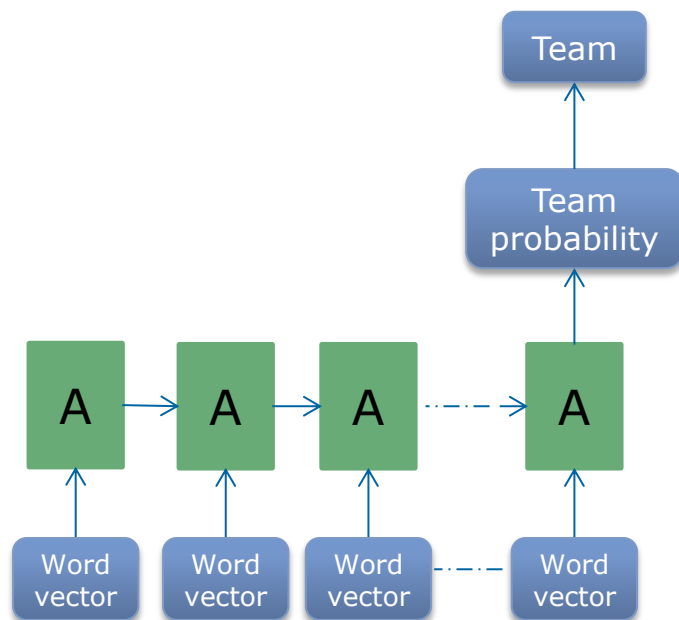


## Recurrent Neural Network (RNN) and bidirectional RNN

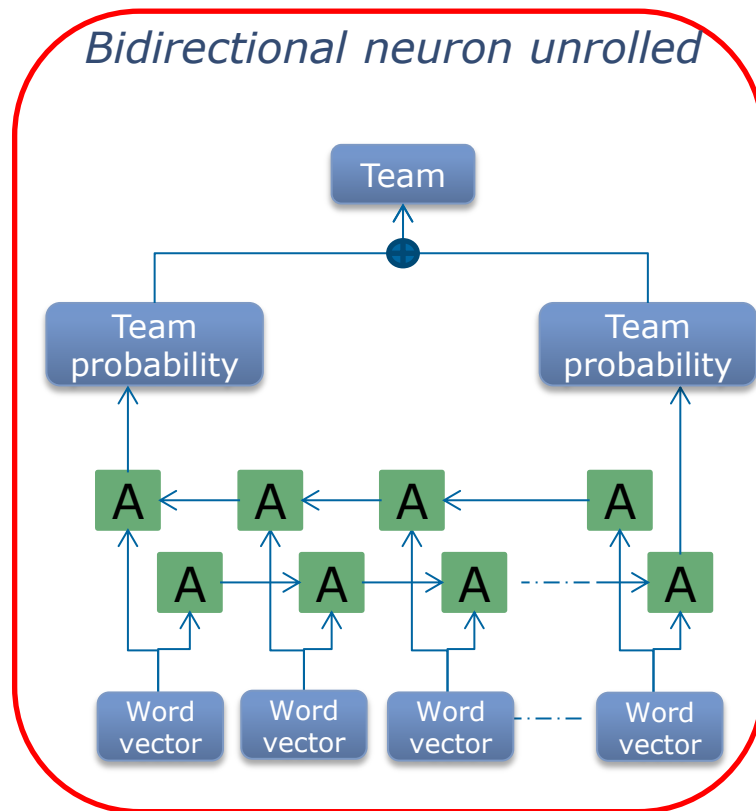
*Recurrent neuron*



*Recurrent neuron unrolled*



*Bidirectional neuron unrolled*





# Team prediction

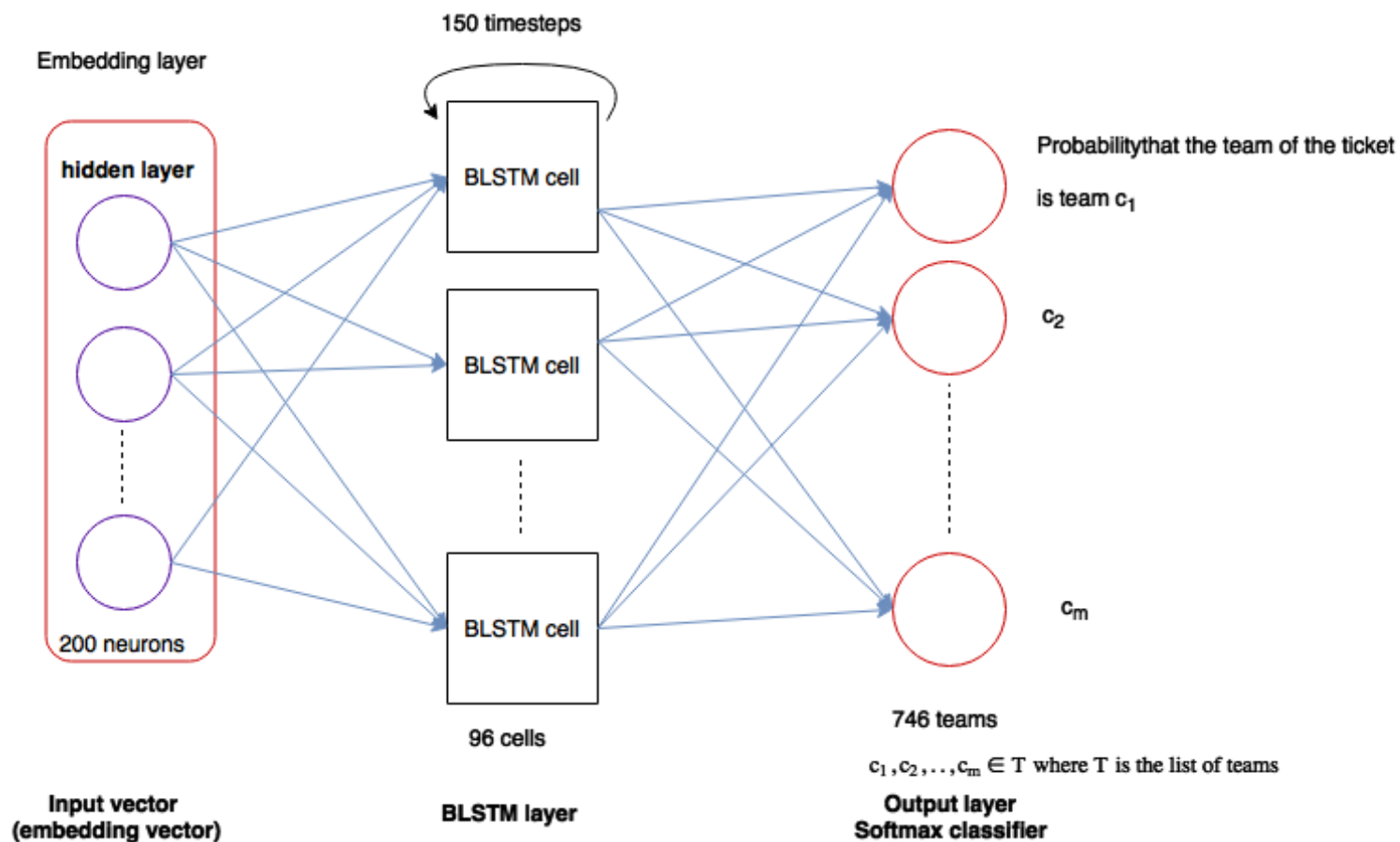
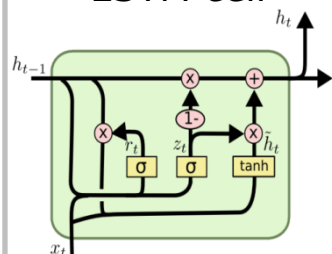
Introduction

Subject

Technical realization

Conclusion

*LSTM cell*



# Tricks to improve learning

## Finetuning

Introduction

Subject

Technical realization

Conclusion

80%

Training data

x20

60%

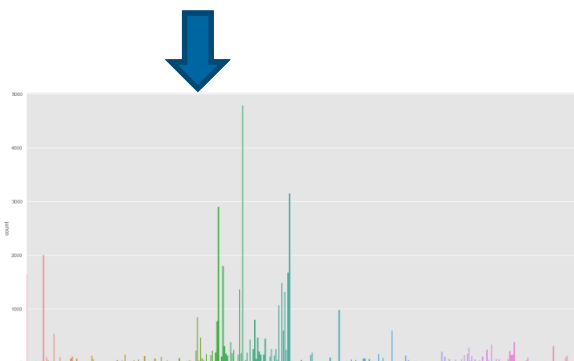
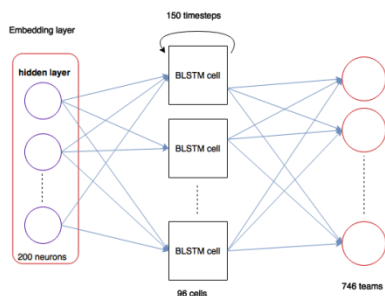
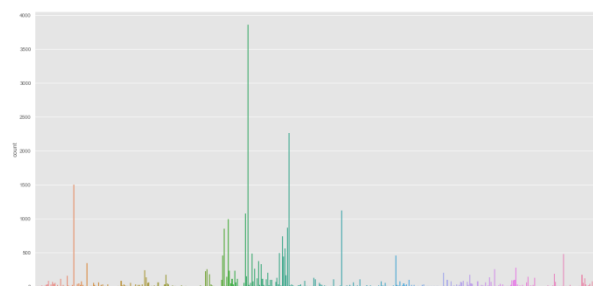
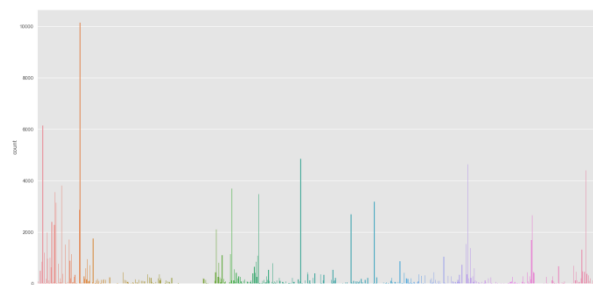
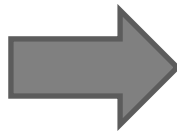
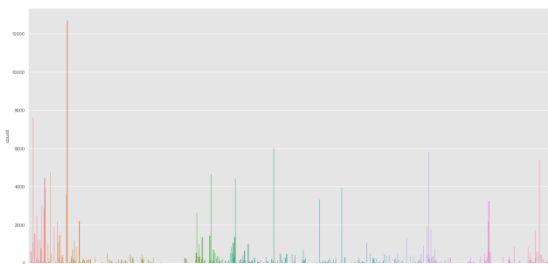
Training data

20%

training data

x5

20%  
Testing data



- ▶ +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](mailto:turpaultn@gmail.com)
- ▶ Github available soon: <https://github.com/turpaultn/CnafSAXO>

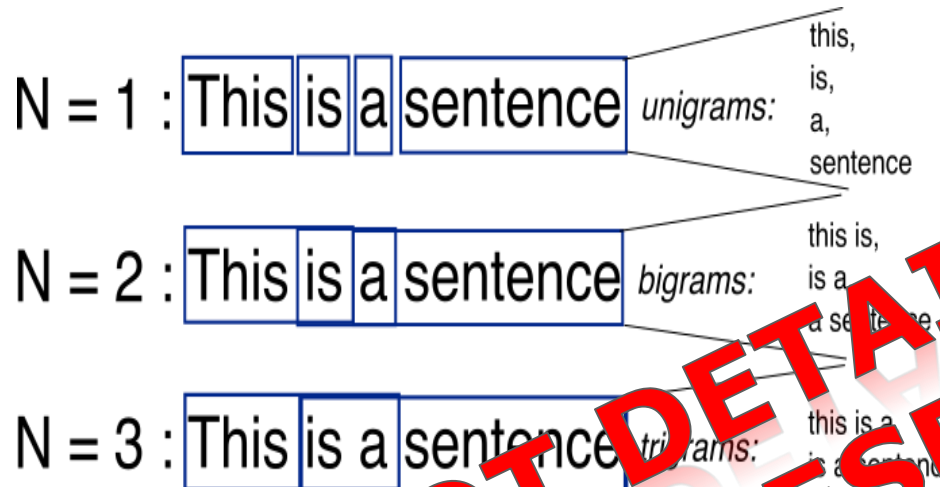
# Traditional method ?

Introduction

Subject

Technical realization

Conclusion



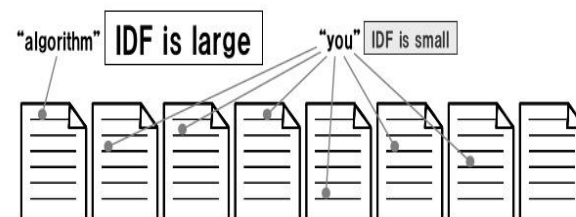
Term-Frequency Metric	Term(1)	Term(2)	Term(3)	Term(4)	Term(5)	Term(6)	Term(n)
Document(1)	225	300	0	0	0	25	0
Document(2)	78	87	0	0	0	175	0
Document(3)	13	0	0	237	0	21	0
Document(4)	1	101	0	0	0	0	0
Document(5)	3	15	0	24	0	48	87
Document(6)	0	0	71	0	0	0	0
Document(n)	109	0	901	221	331	441	551

## 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  $D$



$$tf-idf_{(t,d)} = tf_{(t,d)} \times idf_{(t)}$$

$t$  = term

$d$  = document

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

Presentation Deep learning at ATOS with CNAF and ATOS collaborators (~40 pers)  
(*September, 7<sup>th</sup> 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](#))

4

Conclusion

---

# Experiencing difficulties

Introduction

Subject

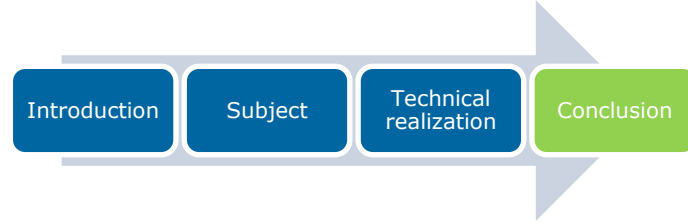
Technical  
realization

Conclusion

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

- ▶ 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





- ▶ Wonderful experience
- ▶ I thank ATOS and CNAF for their confidence during this apprenticeship
- ▶ 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