




Tutorial: NLP and Topic Modeling

John Aoga

BIOGRAPHY

AI Seminar
19th December 2022







John Aoga, PhD

Who am I ?
 Doctor & Engineer in Science and Technology
 Specialist in Data science & AI
Online Content author and Teacher
 Co Founder of MIFY SARL company

Goals and Aspirations
 Promote and develop AI 4 Africa In Africa
 Promote and develop Education tools

Domains & Interests
 Algorithms and Optimization
 Data/Pattern Mining Approches and applications
 Deep Learning & NLP for local languages
 Social Data Analysis


Scientific References

✉    

Aoga John () AI Seminar 2

MIFY-SARL (MACHINE INTELLIGENT FOR YOU)

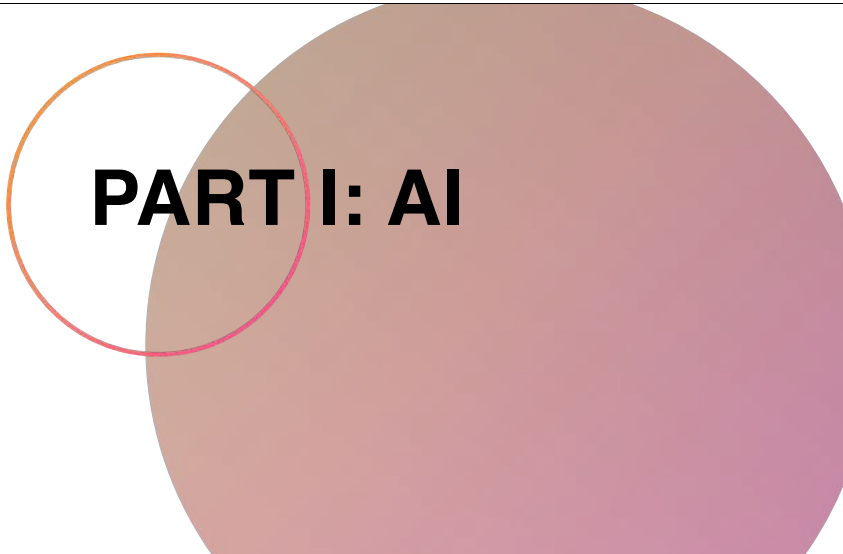
AI Seminar
19th December 2022



- IOT**
Build IA objects to help people in their daily workflow
- AI/OPTIMIZATION**
 - Covid assistant
 - NLP products
 - MAIC
- DATA SCIENCE & IS**
 - Help Entreprixe to build IS
 - Data Analysis and BI
- OTHER SERVICES**
 - Training
 - App Dev.

Aoga John () AI Seminar 3

PART I: AI





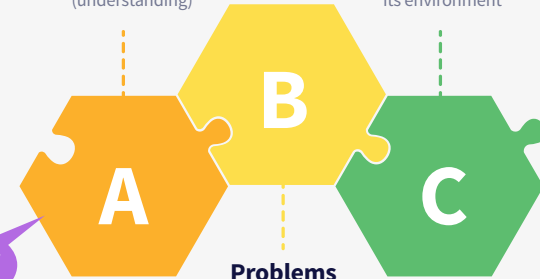
What's AI?



- PROGRAMM
- MACHINE
- OBJECT

Reasoning
Simulate reasoning
(understanding)

Adaptation
Can understand & live in
its environment



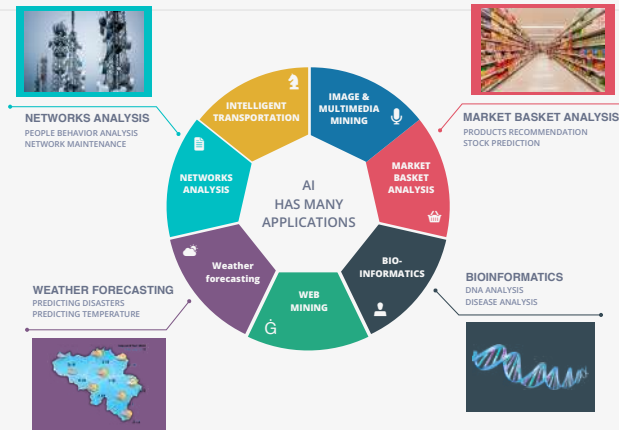
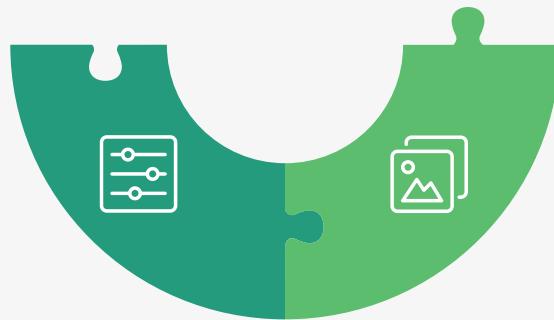
Not necessary
from Human

Problems
Solve Problems (Find
feasible solutions)



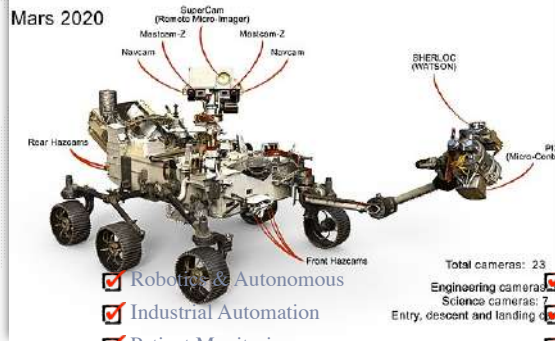
DATA

ALGORITHMS





Mars 2020



- Robotics & Autonomous
- Industrial Automation
- Patient Monitoring

Total cameras: 23
Engineering cameras: 12
Science cameras: 7
Entry, descent and landing camera



- Electricity Use Forecasting
- Self Driving Cars
- Predictive Maintenance



Predictive

Use analytics to predict machine failures

Reactive

Fix when the equipment is down

Proactive

Eliminate defects at an early stage

Periodic

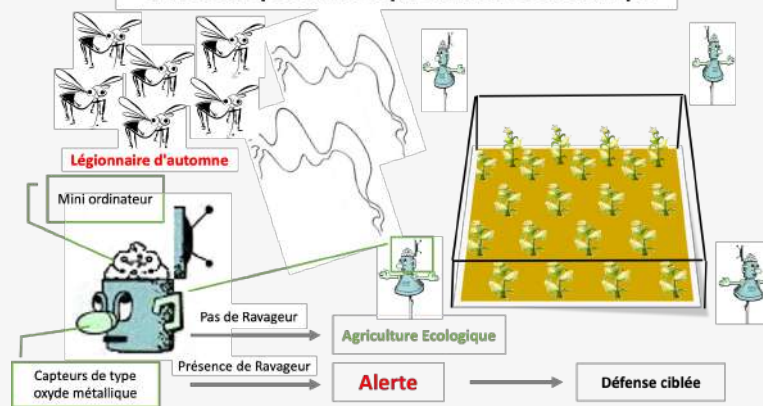
Scheduled Maintenance



AI Applications (Closer to us!)

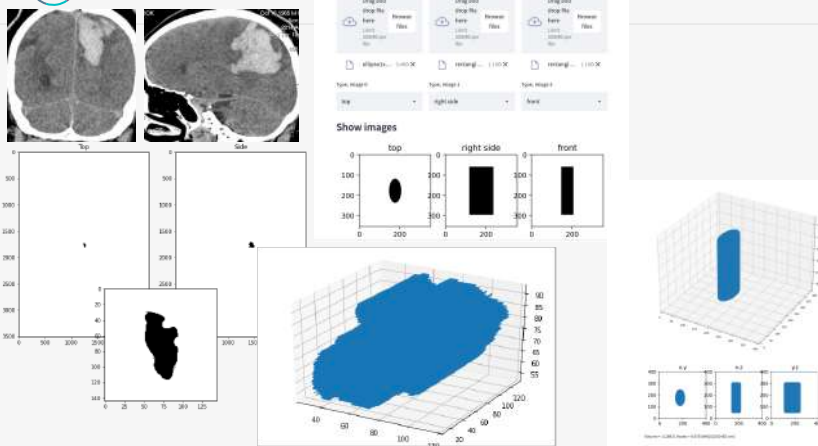
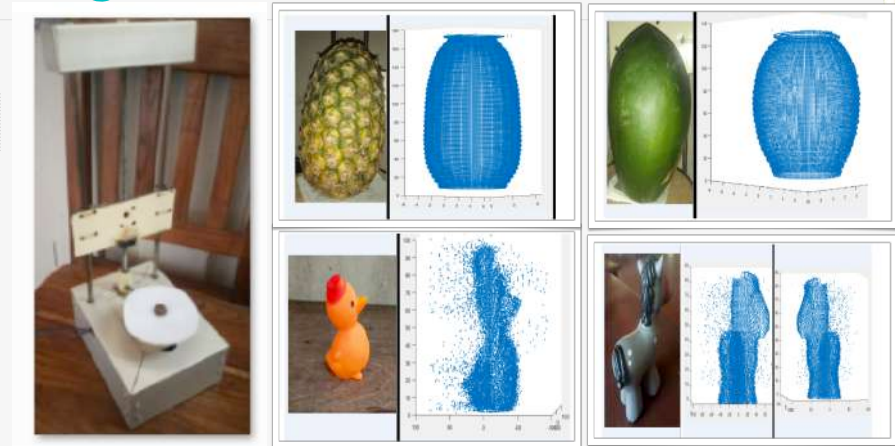
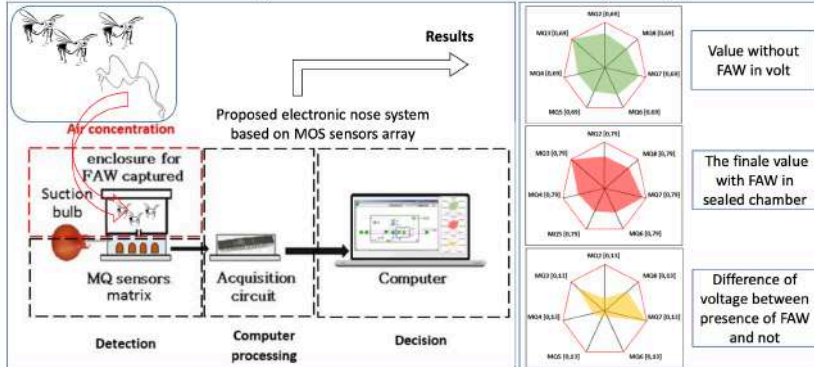


Un robot Epouvantail qui surveille nos champs





An Affordable E-Nose Using Metal Oxide Semiconductor Sensor Array for early detection of fall armyworm (FAW) pest



Prévision du débit d'une rivière par les Réseaux de Neurons Artificiels dans le bassin versant de l'Ouémé à l'exutoire de Savé (Bénin, Afrique de l'Ouest)

Zohou Pierre Jérôme^{1,2,3} Biaou Iboukoun Elizezer^{4,5} Aoga John^{1,2} Houessou Oscar⁵ Alimou Adéchinu Eric^{4,6} Ezin C. Eugène^{1,2}
 Université d'Abomey-Calavi (UAC), Bénin¹ Centre International en Physique Mathématique et Applications (CI-PMA-CIAP) UAC, Bénin² Laboratoire du Recherche en Sciences Économiques et Applications (LRSEA) UAC, Bénin³ Université Nationale des Sciences, Technologies, Ingénierie et Mathématiques (UNSTIM) Abomey, Bénin⁴ Laboratoire d'Hydrologie Appliquée (LHA) UAC, Bénin⁵

Introduction

Les précipitations sont des phénomènes naturels et constituent généralement les plus grands apports de débit hydrologique d'un bassin versant. Prédire et gérer les débits des cours d'eau est donc une nécessité pour la maîtrise des crues.



Figure 1: Localisation géographique du bassin versant de l'Ouémé à Savé.

Le schéma d'un bassin versant d'après [14] Q: Débit; P: Précipitation; E: Évapotranspiration. Les modèles hydrologiques, autres que conceptuels et physiques (n'arrivent pas à bien simuler les hautes eaux. Au vu de l'existence des nouvelles méthodes basées sur l'apprentissage automatique, il est intéressant de tester de nouvelles approches de modélisation plus précises afin d'obtenir au mieux les débits.

Objectif général: Modéliser la relation pluie-débit grâce aux réseaux de neurones récurrents de type LSTM et GRU.

Matériels

Zone d'étude et données
 • Situation géographique: 09°12' latitude Nord et 02°16' longitude Est.
 • Climat tropical savanien Saison sèche de mai octobre à avril Saison pluvieuse de mai à octobre
 • Précipitations moyennes interannuelles: 1275 mm
 • Néohaut: 1161 à 2010
 • Sources: Agence Nationale de la Météorologie du Bénin (Météo Bénin) et de la Direction Générale de l'ETA.

Technologies utilisées
 Python TensorFlow Keras

Méthodes

Architecture LSTM et GRU
 Pour simuler efficacement ces débits nous avons utilisé les réseaux de neurones récurrents à mémoire court terme et long terme (Long Short Term Memory en anglais [22]) et le réseau de neurones récurrents à ports Gated Recurrent Unit en anglais[13].



Processus d'apprentissage
 Le processus d'apprentissage des réseaux de neurones artificiels se présente comme suit:



Figure 2: Processus d'apprentissage.

Résultats

Optimisation des hyperparamètres
 Pour améliorer la qualité de nos modèles nous avons optimisé certains hyperparamètres à savoir: le taux d'apprentissage, la taille du lot, le nombre d'époques, le nombre d'unités.

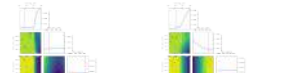


Tableau 1: Valeurs des hyperparamètres du modèle LSTM. Tableau 2: Valeurs des hyperparamètres du modèle GRU.

Ces valeurs optimales des hyperparamètres sont présentées dans le tableau 1.

Modèle	Taux d'apprentissage	Nombre d'époques	Époques de patience	Taille du lot
LSTM	0.0025	144	913	154
GRU	0.02	222	254	144

Simulation avec les modèles

Après l'entraînement, la validation des deux modèles a été effectuée avec les données d'essai qui servent à cette première phase. L'ensemble des données de validation constitue 20% de l'ensemble des données de départ.

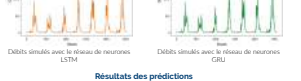


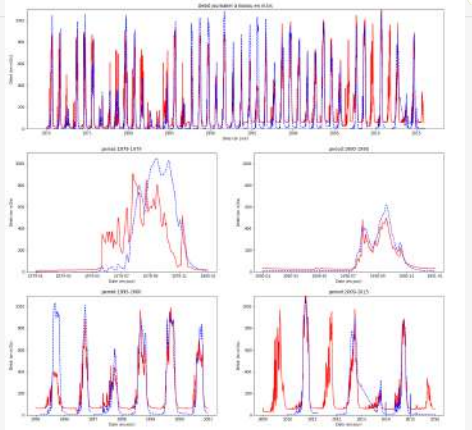
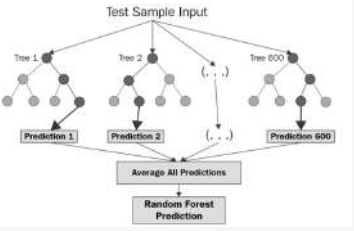
Tableau 2: Performances en entraînement et validation.

Cratères de performances	Entraînement	Validation
Coeficient de détermination (R ²)	0.917	0.810
Erreur Moyenne Absolue (MAE)	0.024	0.026
Erreur Moyenne Quadratique (RMSE)	0.044	0.044

Ces performances en entraînement et en validation prouvent que nos modèles dépassent bien le seul acceptable en modélisation pluie-débit à l'échelle journalière proposée par [1] qui est de 0,5.

- Références
 [1] Wengchen Guo, Ben Van Meirvenne, Capel Gannon, David Vitarina, Felix Baigun, Roger Kibria, and Markus Becht. Learning from experimental data to model stream flow discharge by artificial neural networks. *Water*, 2016.
 [2] How to calculate the coefficient of determination.
 [3] David Hertzog, Robert Zechman, and Anand Chaturvedi. *Hydrology and water quality models: An introduction and applications*. Springer, 2015.
 [4] Hisham Khamis Anwar. *Artificial neural networks: A practical approach*. CRC Press, 2015.

FLOOD FORECASTING USING MACHINE LEARNING



AI for Development

AI AND INTERNATIONAL DEVELOPMENT

- 01 **SDG principle**
Sustainable Development Goals
- 02 **No Poverty**
Food sustainability, Outcome optimization, reduce inequalities
- 03 **Quality Education**
Online courses (+ create educational content)
- 04 **Clean resources**
Clean water and sanitation, Affordable and clean energy, Responsible consumption

- 05 **Suitable Cities**
Well design roads, transportation flows and optimization
- 06 **Suitable Services**
Digitalization of administration (Quick and efficient services)
- 07 **Production**
Optimization of production, Use of suitable Tools
- 08 **Health**
New ways to efficiently tackles healthcare problems

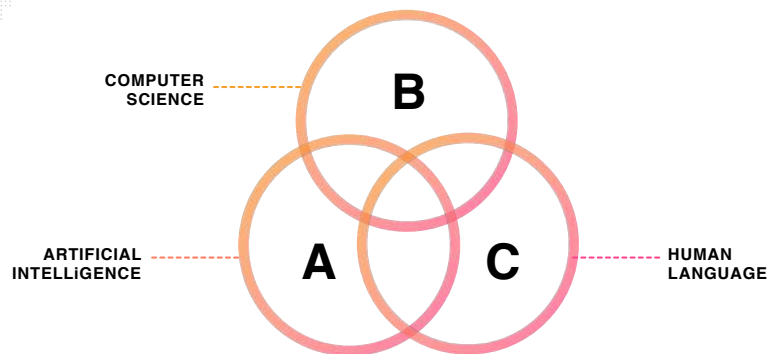
PART II: Discovery

AI & Natural Language Processing

What's NLP

Make computers «understand» and parse Natural(Human) Language

23



NLP Components

Make a computer to « understand » and parse Natural Language

24

@ NL Understanding

Mapping input to useful representation and Analyzing different of languages

1 Question and Answering

2 Sentiment analysis

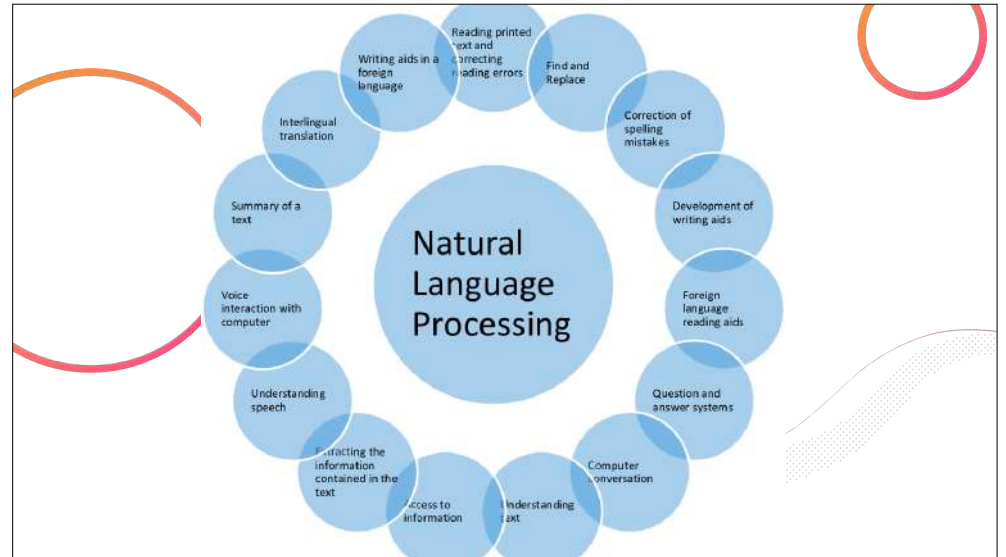
@ NL Generation

Produce meaningful phrases following a structure of a languages

1 Text Summarization

2 Text To Speech

NLP Applications



Human & Computer communication

Before: Programming Language

```

1 def play_audio_file(fname):
2     """Simple callback function to play a wave file.
3
4     :param str fname: wave file name
5     :return: None
6     """
7     ding_wav = wave.open(fname, 'rb')
8     ding_data = ding_wav.readframes(ding_wav.getnframes())
9     audio = pyaudio.PyAudio()
10    stream_out = audio.open(
11        format=audio.get_format_from_width(ding_wav.getsampwidth()),
12        channels=ding_wav.getnchannels(),
13        rate=ding_wav.getframerate(), input=False, output=True)
14    stream_out.start_stream()
15    stream_out.write(ding_data)
16    time.sleep(0.2)
17    stream_out.stop_stream()
18    stream_out.close()
19    audio.terminate()

```

After: Natural Language

(Chat)bots

Voice Assistant

ASSISTANT DOMOTIQUE XWEFA
PERMET DE GERER LES INSTALLATIONS ELECTRIQUES DE CHEZ VOUS

ASSISTANT VOCAL CICA
PERMET DE LANCER DES COMMANDES SUR SON PC OU TELEPHONE

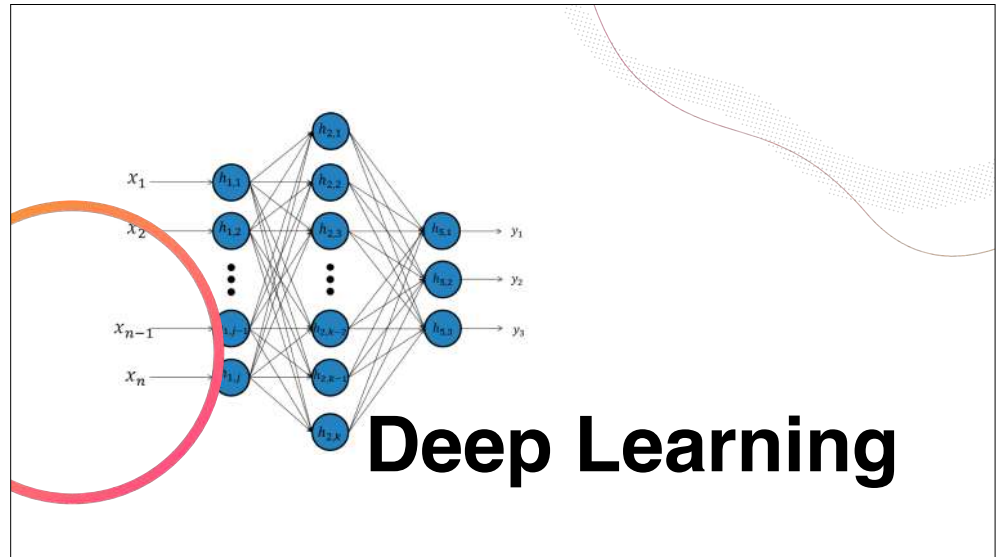
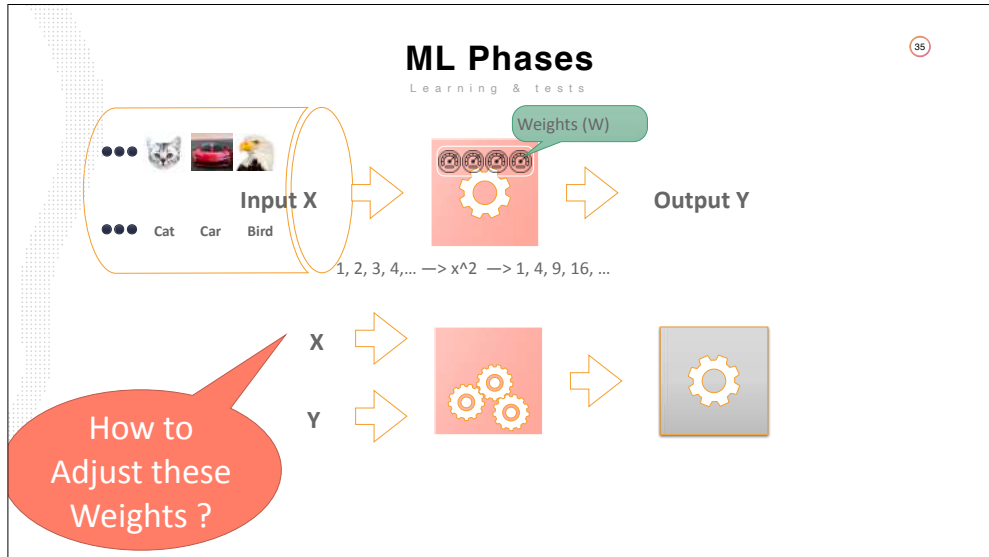
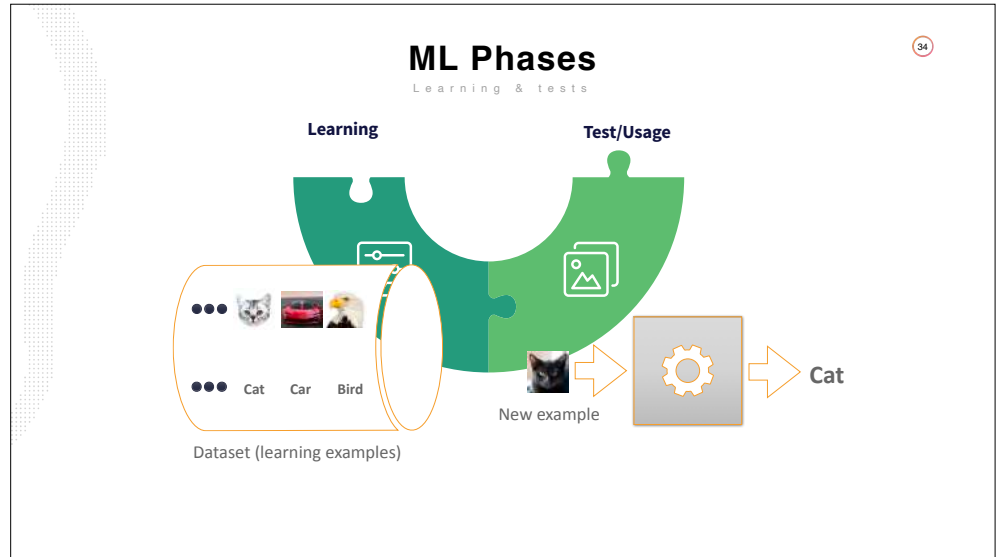
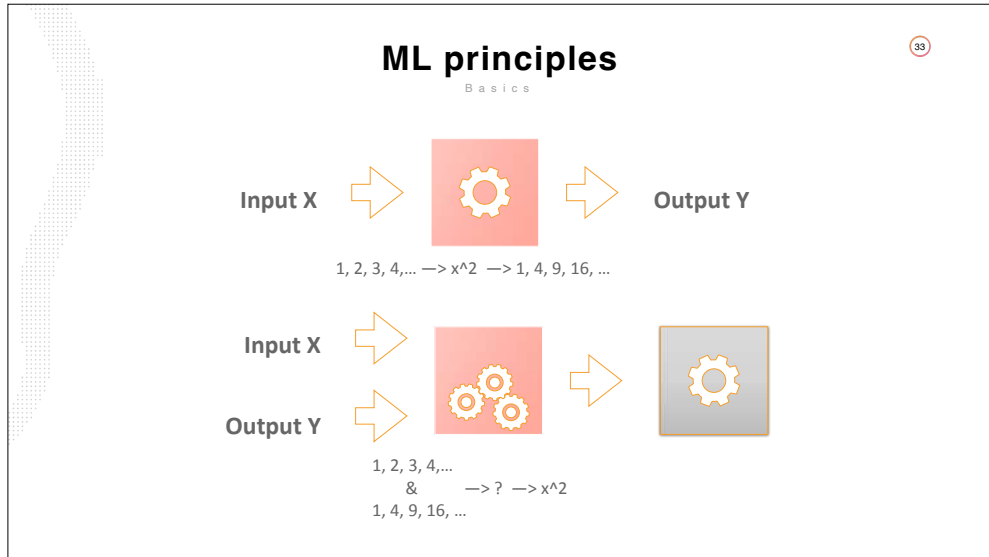
MIFY SARL dans TALN

- Application Cica
- Application Xwefa
- Application Dofongbé
- Application Diagnostic COVID19

Do Computers (Really) Understand?

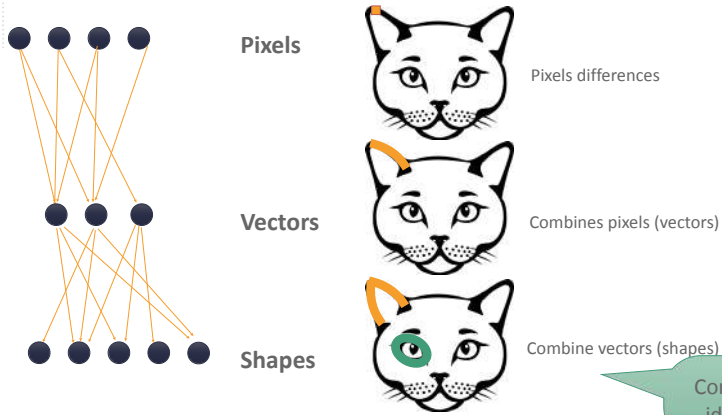
HOW?

Machine Learning



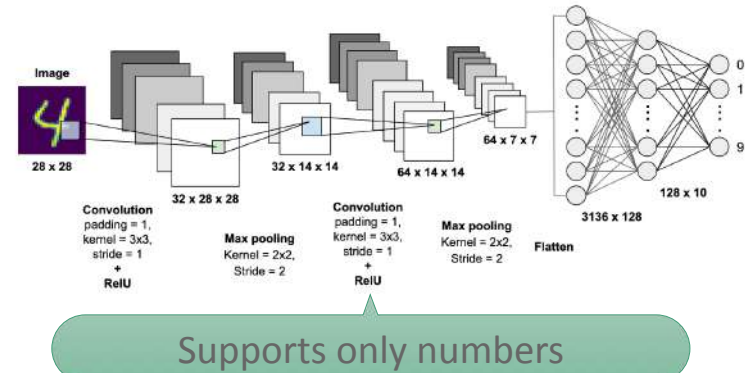
Convolutional Neural Network

37



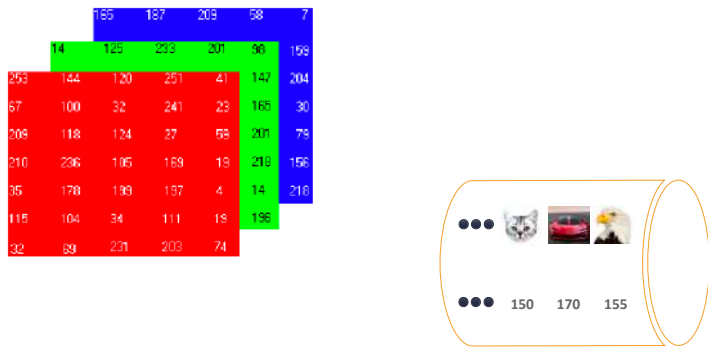
Convolutional Neural Network

38



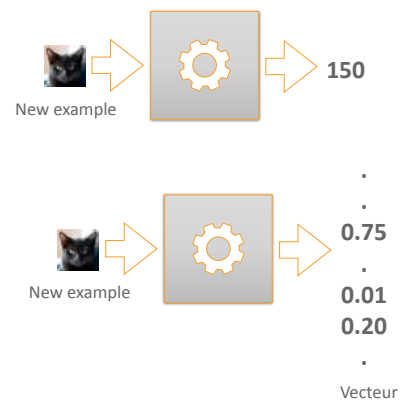
Solve number's Problems

39



Solve number's Problems

40

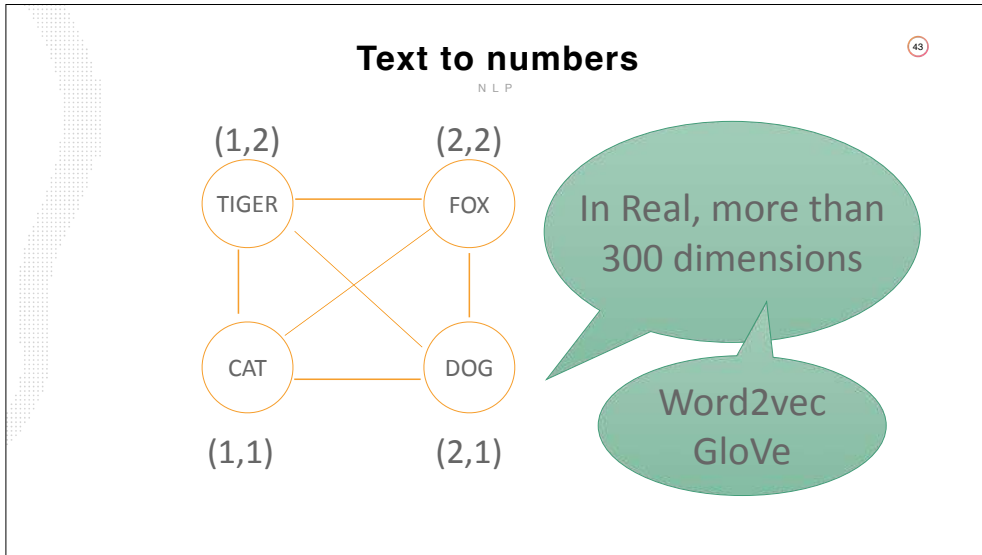


Deep Learning With Language

Language Processing Pain Points

04 pain points highlights here

- 1 Text -> numbers**
- 2 Text with different sizes**
Image with fix size at the begin, if not resize
- 3 Order is really important**
Difficult to split the sentence without loosing information
- 4 Meaning**
Emotion detection in a sentence



Glove - Google

N L P

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

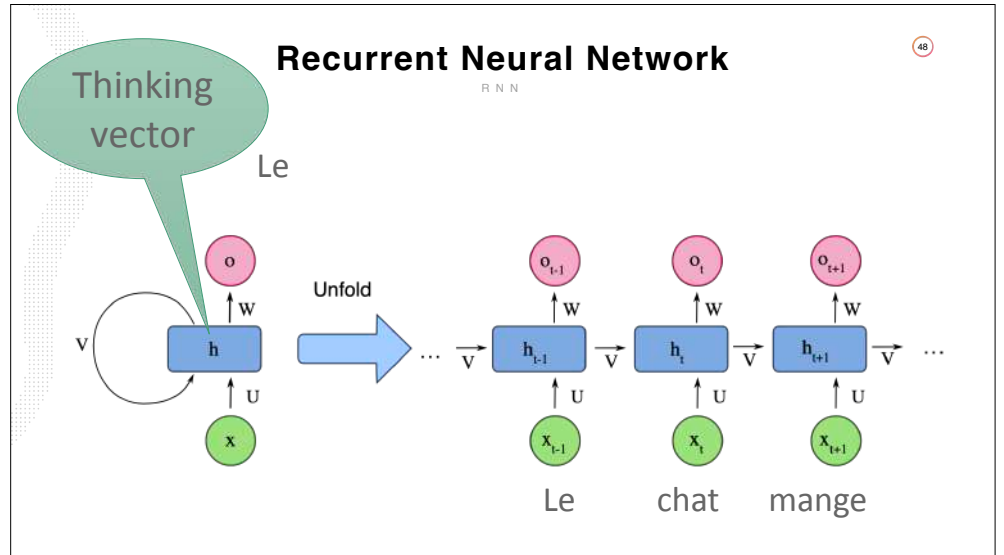
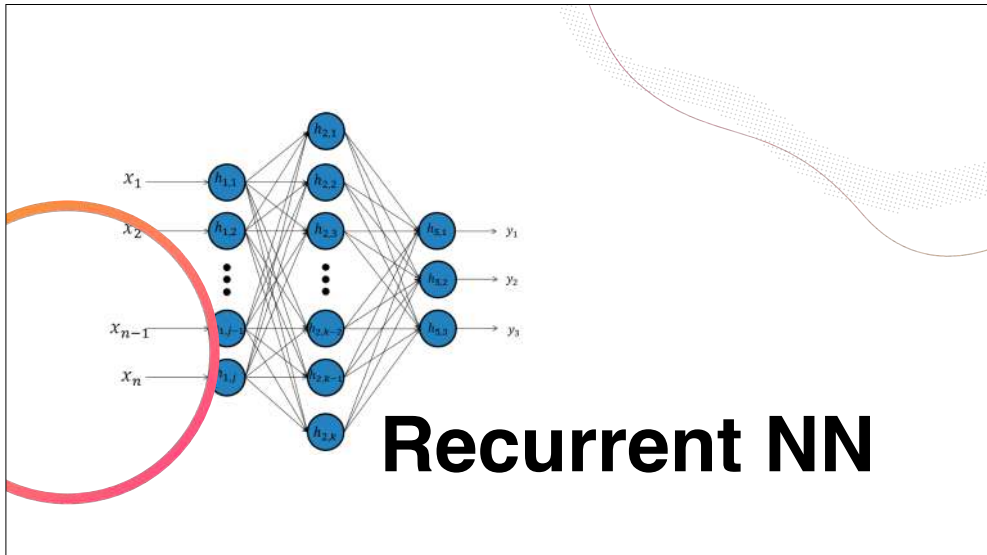
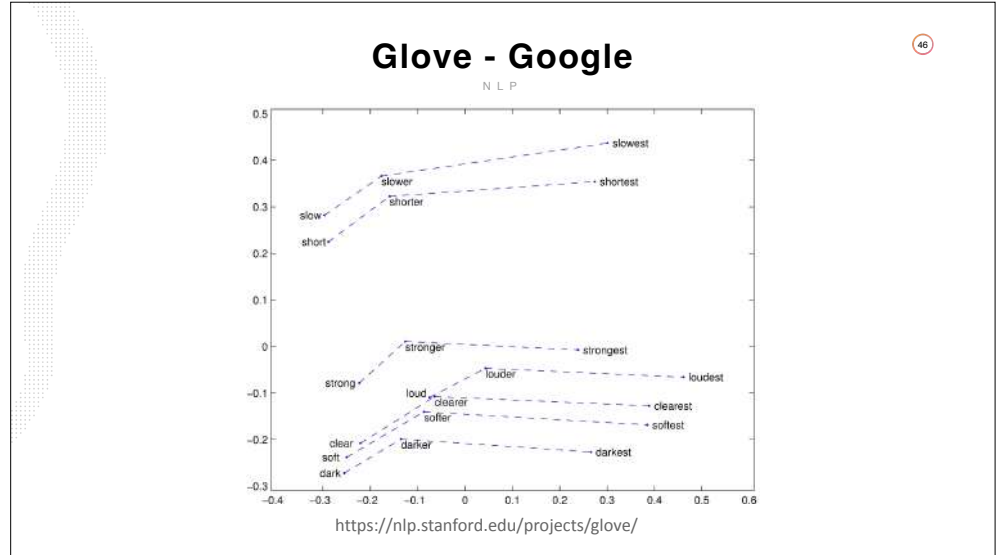
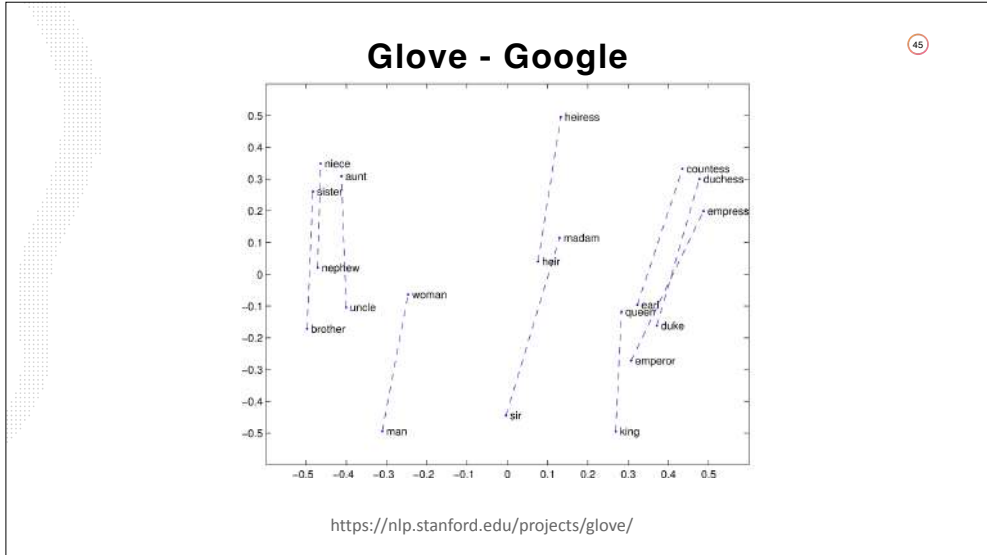
Getting started (Code download)

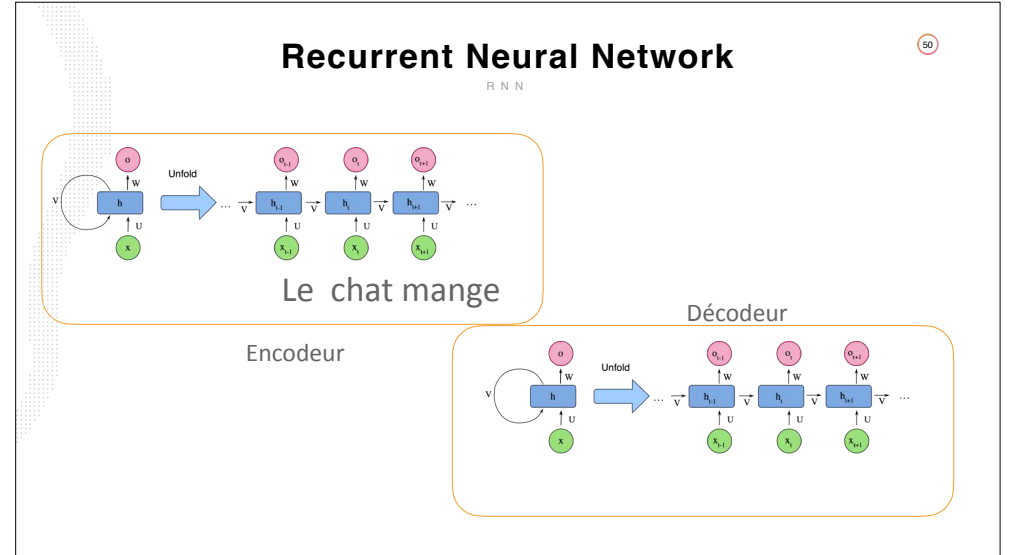
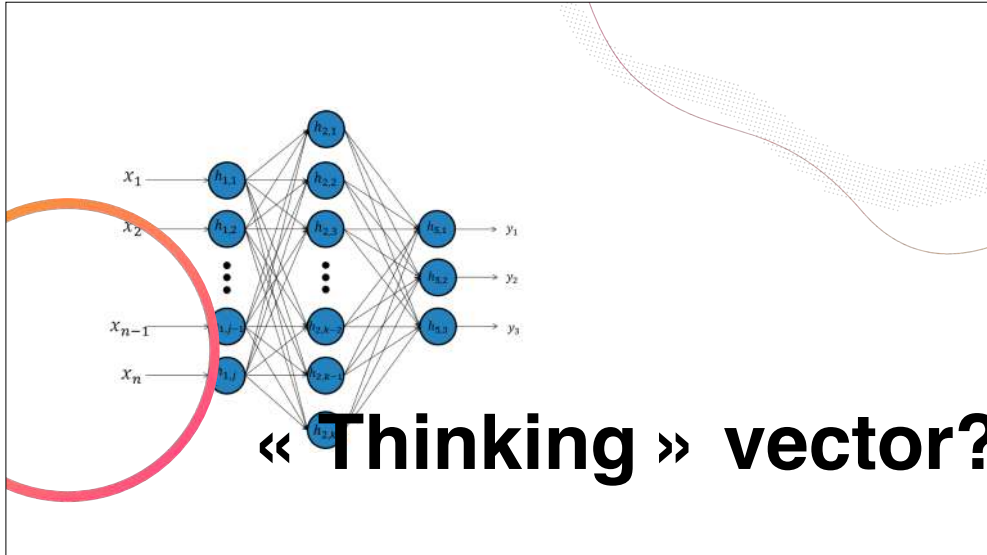
- Download the latest [source code](#) (licensed under the [Apache License, Version 2.0](#))
- Look for "Clone or download"
- Unpack the files, using `mkdir`
- Compile the library, using GloVe's `make` rule
- Run the demo script, `demo.sh`
- Consult the included `README` for further usage details, or ask a [question](#)

Download pre-trained word vectors

- Pre-trained word vectors. This data is made available under the [Public Domain Dedication and License](#) v1.0 whose full text can be found at <http://www.gnu.org/licenses/old-licenses/pddl-1.0.txt>
 - [Word2vec 2014](#) - Gigaword 5 (6B tokens, 400K words, uncsd), 300d, 100d, 300d & 300d vectors, 821 MB download, [glove.6B.zip](#)
 - Common Crawl (65B tokens, 18M words, uncsd), 300d vectors, 1.75 GB download, [glove.6B.commoncrawl.zip](#)
 - Common Crawl (65B tokens, 2M words, csd), 300d vectors, 2.03 GB download, [glove.6B.commoncrawl.csd.zip](#)
 - Twitter (2B tweets, 2B tokens, 12M words, uncsd), 300d, 100d & 300d vectors, 1.42 GB download, [glove.twitter.27B.zip](#)
- Run `utils.py` for preprocessing Twitter data

<https://nlp.stanford.edu/projects/glove/>

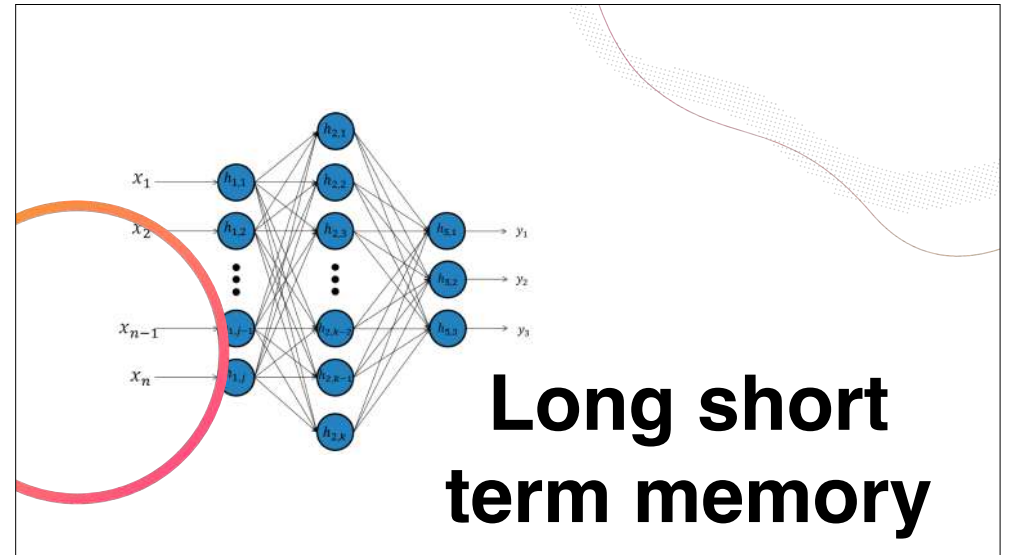




Recurrent Neural Network

Issue

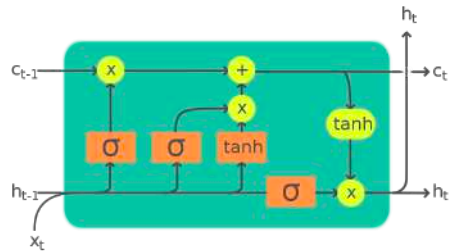
Bohicon est une tr s belle ville du Benin, j'est v cu cinq et c'est l  j'ai appris   parler ...



Long Short Term Memory

L s t m

55



Legend:

Layer	ComponentwiseCopy	Concatenate



TAKEAWAYS
AI for all discipline

AI Seminar
26th May 2022

Can Machine Replace Human One Day?

1 AI is everywhere and is growing faster

2 The future is no longer a time-frame. The future is a mind-set!

4 Promote Interdisciplinary strategy

Aoga John () AI Seminar 56



**ANY
QUESTIONS?**



**What can you
solve in your env
with AI?**



Tuto Time



Practical sessions

Named Entity Recognition (NER)

Identify entities in a text

61

1 Content Classification for news providers

automatically scan entire articles and reveal which are the major people, organizations, and places discussed in them

2 Efficient Search algorithm

Speed up the search by identifying entities (tags) first

3 Powering Content Recommendation

Recommend articles with similar/common entities

4 Customer Support

Use entities to categorize complaints and assign them to the right department

With Spacy

<https://tinyurl.com/2p9ydzd9>

62