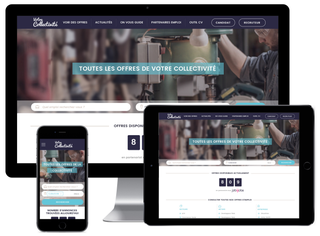
Des réseaux de neurones pour rapprocher CV et offres d'emploi

Justine BEL-LETOILE, Data scientist à Jobijoba

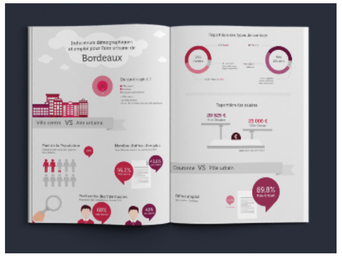
2ème Journée Aquitaine IA, RO et Data Science
7 Février 2019 - Bordeaux - Kedge Business School

Les activités de Jobijoba

Un site grand public
www.jobijoba.com



Une offre pour les collectivités

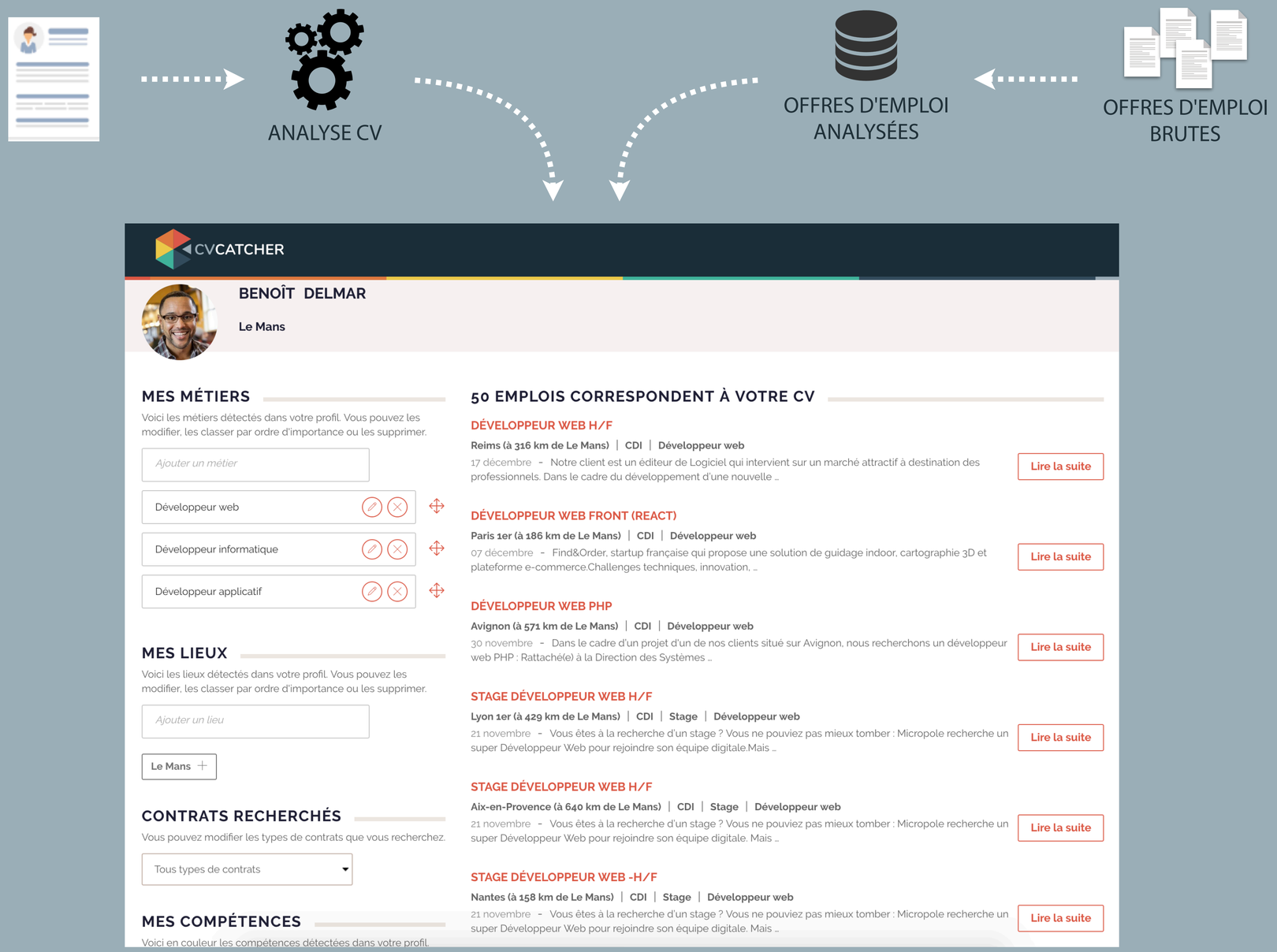


Des observatoires et baromètres

Une offre pour les recruteurs
www.cvcatcher.io

CV CATCHER

Qu'est-ce que le CV Catcher ?



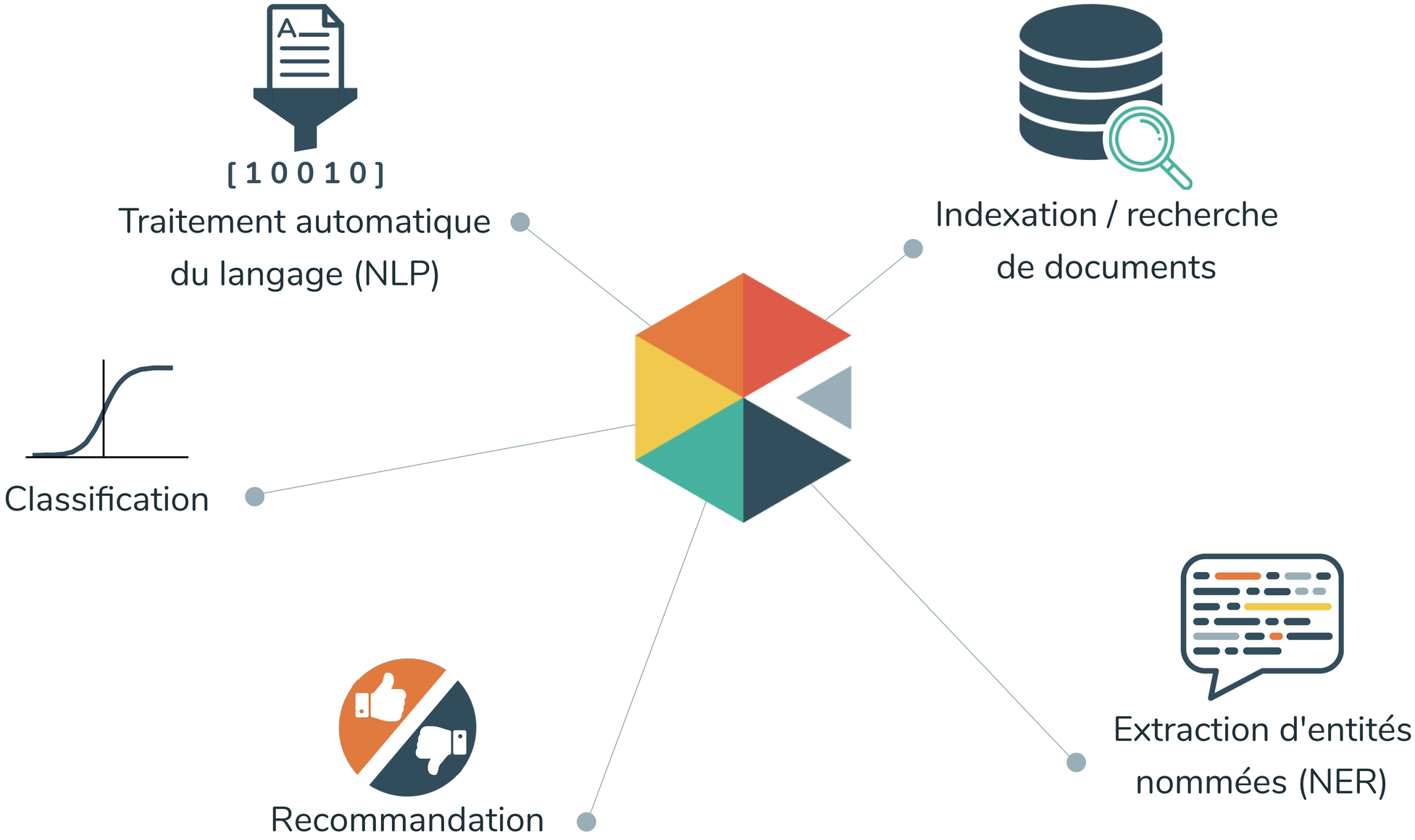
Job matching

Sourcing

Mobilité

Expérience candidat

La Data science pour le CV Catcher



Word embeddings

Word embeddings

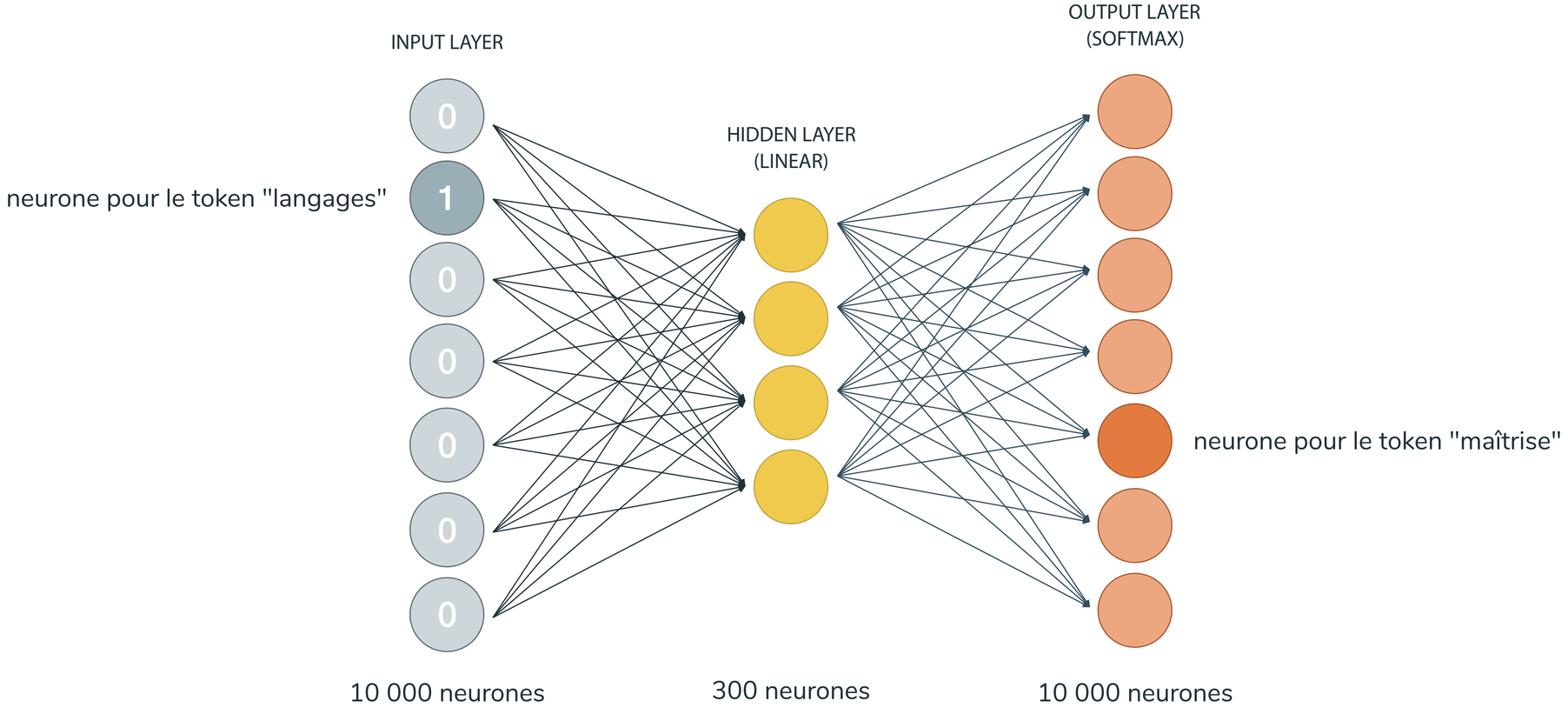
- Représenter chaque mot par un vecteur de $\mathbb{R}^n$
- Concept qui n'est pas nouveau (VSM, LSA)
- Avancées récentes
  - → Word2vec, T. Mikolov (Google), 2013
  - → GloVe, J. Pennington (Stanford), 2014
  - → FastText, Facebook AI Research, 2016

Word2vec : le principe de la "fake task"

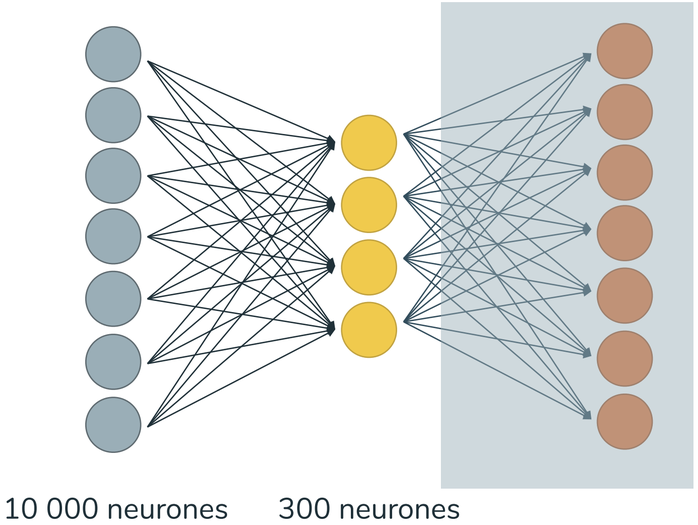
Maîtrise	des	langages	R	et	Python

TRAINING DATA

(langages, maîtrise)
(langages, des)
(langages, R)
(langages, et)
(langages, python)



Récupération des vecteurs



Apprentissage sur la donnée emploi

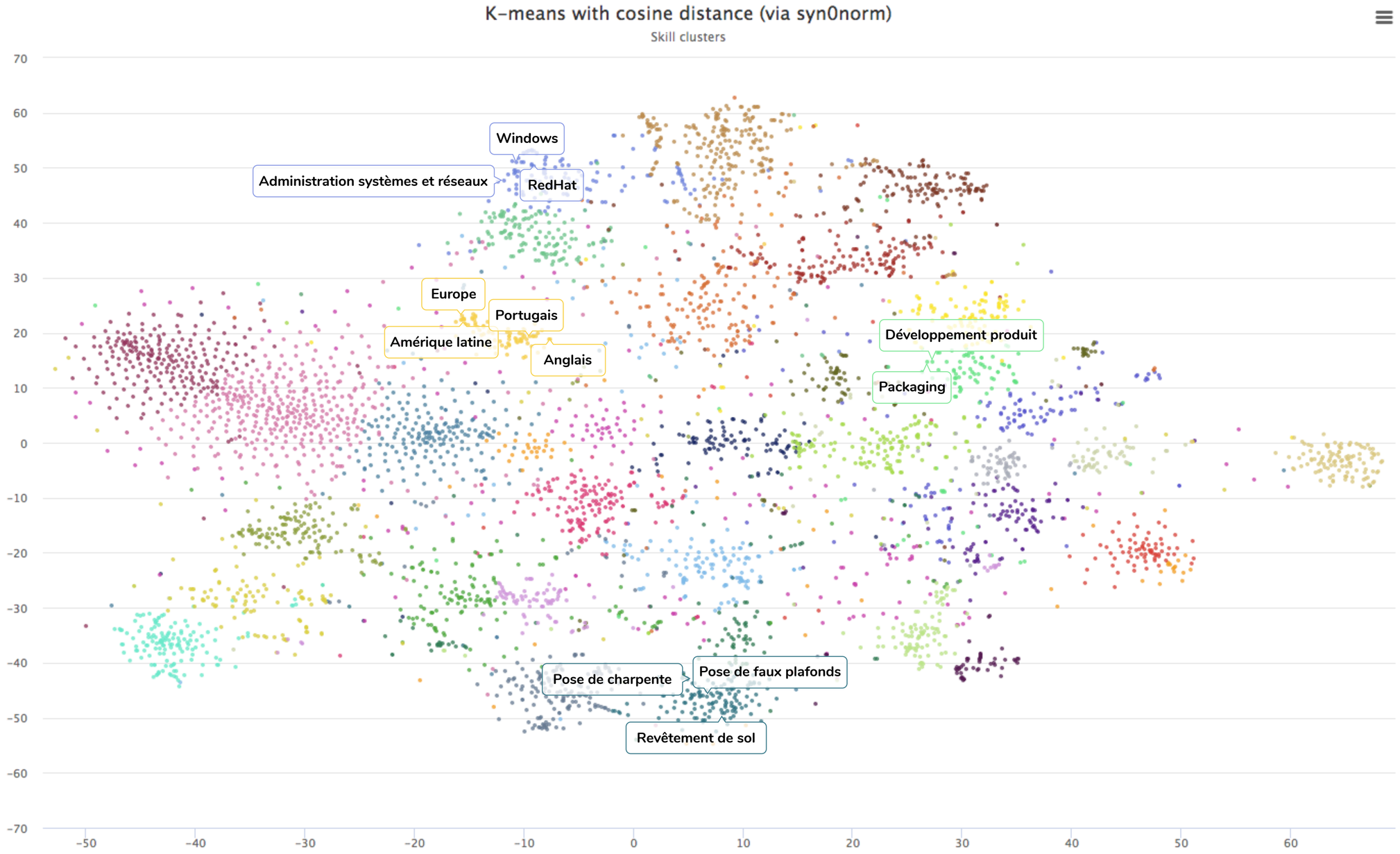
- Textes issus de 1 000 000 d'offres d'emploi et 1 500 000 CVs
- Librairie gensim
- Pré-traitement sur les textes

A ce titre, vos missions principales seront :
- Contrôle qualité
- Suivi des flux et des stocks matières
- Participation aux inventaires

Vous serez également en charge d'évaluer les fournisseurs, et de créer et suivre les tableaux de bord.

a	ce	titre	vos	mission	principal	seront
Gestion de la qualité						
Gestion des flux	et	stock	matière			
participe	aux	Gestion des inventaires				
vous	serez	egalement	en	charge	d'	Evaluation des fournisseurs
et	de	Suivi et analyse des indicateurs				

Intérêt des word embeddings



Les word embeddings pour enrichir nos thesaurus

- Explorer les plus proches voisins des vecteurs pour les compétences connues

- Enrichissement semi-automatisé

```
kakemono                  => synonyme de "Outils de communication"
artificial_intelligent    => synonyme de "Intelligence artificielle"
Manager_une_équipes       => synonyme de "Gestion d'équipe"
communication_digitale    => nouvelle entrée "Communication digitale"
Spanning_Tree             => nouvelle entrée "STP" (Spanning Tree Protocol)
```

4968 compétences

5532 compétences

Les word embeddings pour suggérer des compétences

```
["Word", "Excel"]
    ===>
    ("Libre Office", 0.753753125668)
    ("Tableur", 0.727019906044)
    ("Open Office", 0.722301483154)
    ("PowerPoint", 0.72156560421)
    ("Photofiltre", 0.683253884315)
```

```
"vous aimez le machine learning et la data science"
    ===>
    ("Apprentissage supervisé", 0.875446438789)
    ("Visualisation de données", 0.768740534782)
    ("Big-data", 0.753282666206)
    ("Algorithmes", 0.729285657406)
    ("Traitement de données", 0.722404718399)
    ("Text-mining", 0.712874412537)
    ("Modélisation", 0.691904425621)
```

Suggestion de compétences sur le CV Catcher

MES COMPÉTENCES

Ajouter une compétence

- Machine learning
- Recherche opérationnelle
- Algorithmes
- Elasticsearch
- Java
- Python
- Sémantique
- Gestion de la logistique
- Assurer le reporting
- Gestion de stock
- Mise en place d'indicateurs
- Amélioration continue
- Mathématiques

- Apprentissage supervisé +
- Traitement d'image +
- Calcul numérique +
- Traitement du signal +
- Méthodes de régression +
- Big-data +
- Suivi et analyse des indicateurs +
- Gestion des résultats +
- KPI +
- Visualisation de données +
- Text-mining +
- Gestion des flux +
- Gestion des achats +
- Mongodb +
- Scala +
- Mathématiques appliquées +
- Cours particuliers +

DATA SCIENTIST

Nanterre (à 493 km de Bruges) | CDI | Data Scientist

06 janvier - Equipe datalab Pas de POC, industrialisation des process Machine learning au coeur des projets Expertise Python demandée

Lire la suite

DATA SCIENTIST F/H

Paris (à 495 km de Bruges) | CDI | Alternance | Data Scientist

26 décembre - Cabinet de recrutement de dimension internationale, S&you est spécialisé dans le recrutement d'Experts et Cadres. Nos 30 consultants expérimentés mettent ...

Lire la suite

DATA SCIENTIST MANAGER H/F

Le Plessis-Robinson (à 484 km de Bruges) | CDI | Data Scientist

07 janvier - Notre entreprise Sodicam² - Groupe Renault, filiale détenue à 100 % par le Groupe Renault depuis 1999, est en charge de la commercialisation et de l'animation ...

Lire la suite

ASSISTANT DATA SCIENTIST H/F

Poissy (à 491 km de Bruges) | Data Scientist

01 janvier - Métier : Système d'informationContrat : StageDescription du poste : Dans le cadre de notre développement, nous recherchons un Assistant Data Scientist. ...

Lire la suite

CONSULTANT DATA ANALYST / DATA SCIENTIST - H/F

Lille (à 693 km de Bruges) | CDI | Data Scientist

21 novembre - Groupe international spécialisé en conseil et technologies innovantes, Micropole recrute des consultants DataAnalyst / DataScientist.Acteurs incontournables du ...

Lire la suite

Offres précédentes | Offres suivantes

Mesurer la qualité d'un word embedding

- Analogies "a est pour b ce que c est pour..."
  - analogies syntaxiques et sémantiques
  - $d = arg\ \underset{i}{max} \frac{(x_b - x_a + x_c)^T x_i}{\|x_b - x_a + x_c\|}$

```
: capital-common-countries
Athens Greece Baghdad Iraq
Athens Greece Bangkok Thailand
: gram2-opposite
acceptable unacceptable aware unaware
clear unclear logical illogical
```

source: http://download.tensorflow.org/data/questions-words.txt

- Corrélation avec le jugement humain
  - score de similarité entre 2 mots sur une échelle de 0 à 10
  - corrélation entre les scores et les distances entre vecteurs

Word 1	Word 2	Human (mean)
tiger	cat	7.35
tiger	tiger	10.00
book	paper	7.46
stock	phone	1.62
king	queen	8.58
school	center	3.44
street	avenue	8.88

source: http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

Extraction d'entités nommées

Extraction d'entités nommées

Benoît FIRST NAME DELMAR LAST NAME

12 allées des Tulipes - 72000 Le MANS LOC

06.00.00.00.00

benoit.delmar@yopmail.com

Ingénieur informatique JOB

Compétences : Java SKILL, J2EE SKILL, .Net SKILL, C++ SKILL, HTML SKILL / CSS SKILL, Bootstrap SKILL

EXPERIENCES PROFESSIONNELLES

Développeur web JOB junior / 2012 - 2018 (Le Mans)

Développement, référencement et mise à jour de sites web SKILL et d'applications.

Gestion de projet SKILL autour d'une équipe pluridisciplinaire.

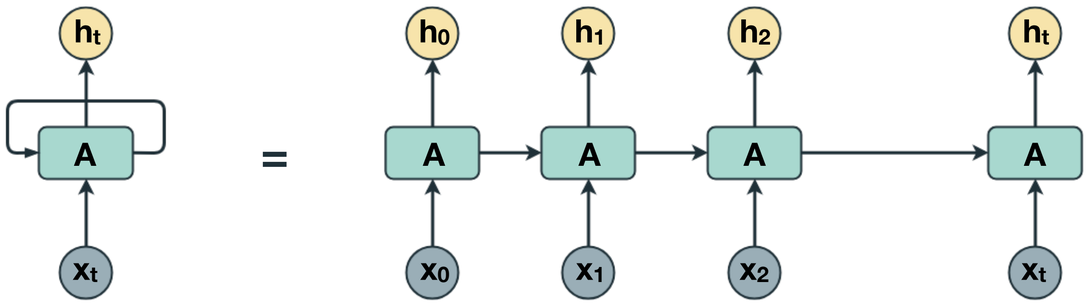
Recueil du besoin SKILL client.

Respect des délais et des spécifications clients.

Suggestions d'évolutions via la méthodologie agile SKILL

Développement d'applications mobiles SKILL android SKILL

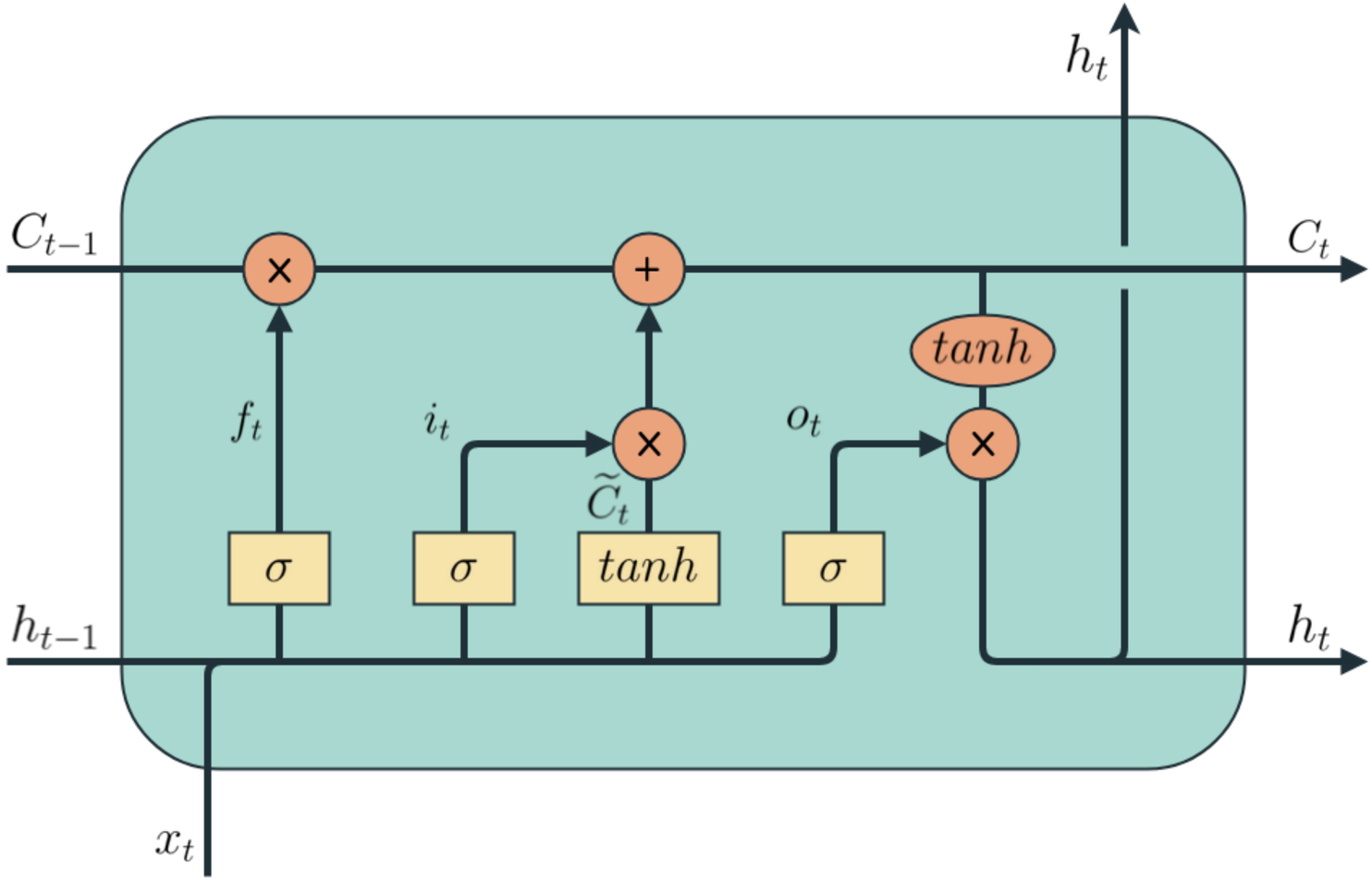
Les réseaux LSTM pour traiter des séquences



- Réseau de neurones récurrent
  - → garder en mémoire les informations passées
  - → disparition du gradient

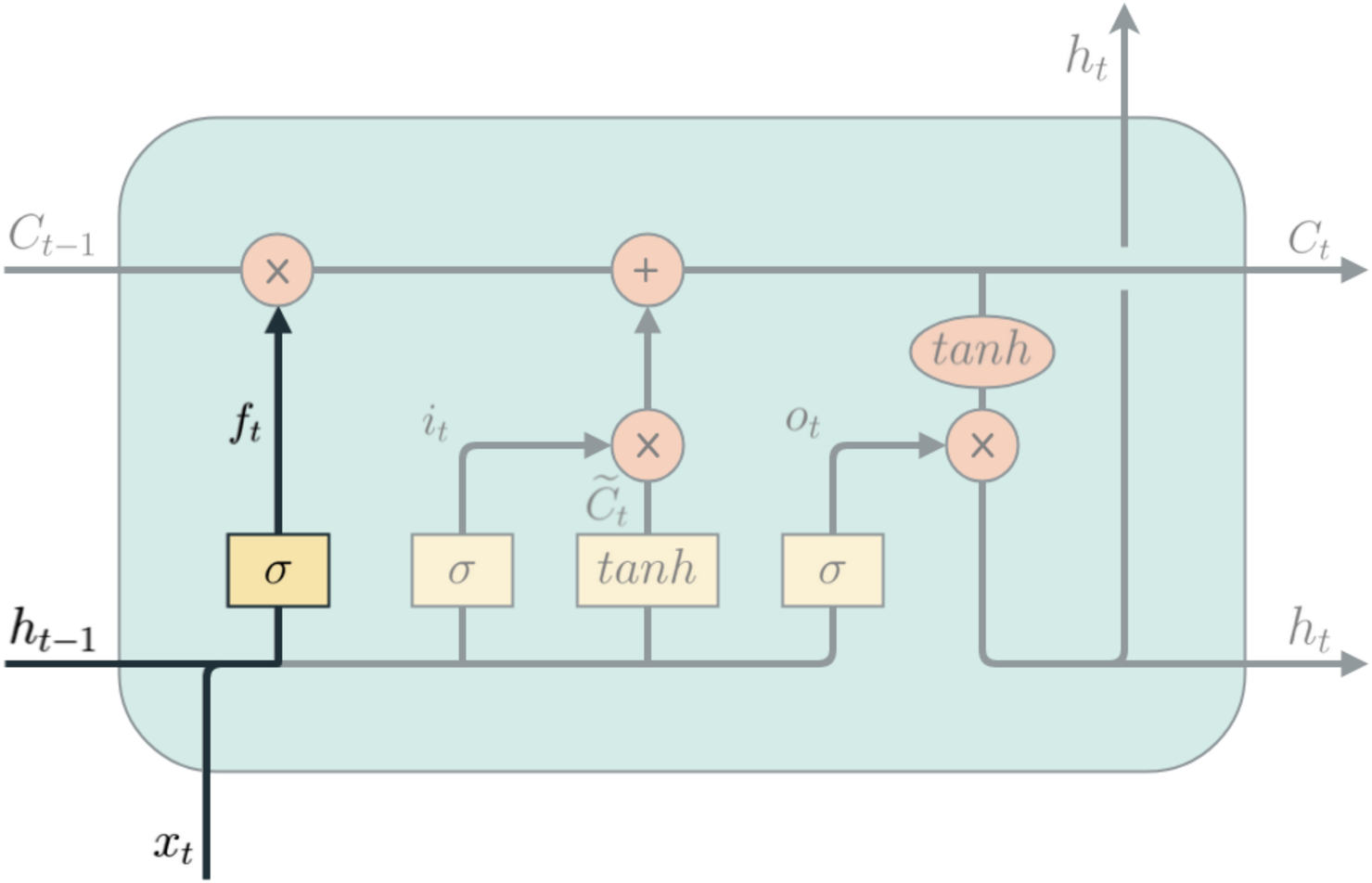
- Long Short-Term Memory
  - → oublier, garder en mémoire ou mettre à jour l'info qui est transmise

Module LSTM



source: Understanding LSTM networks, C. Olah

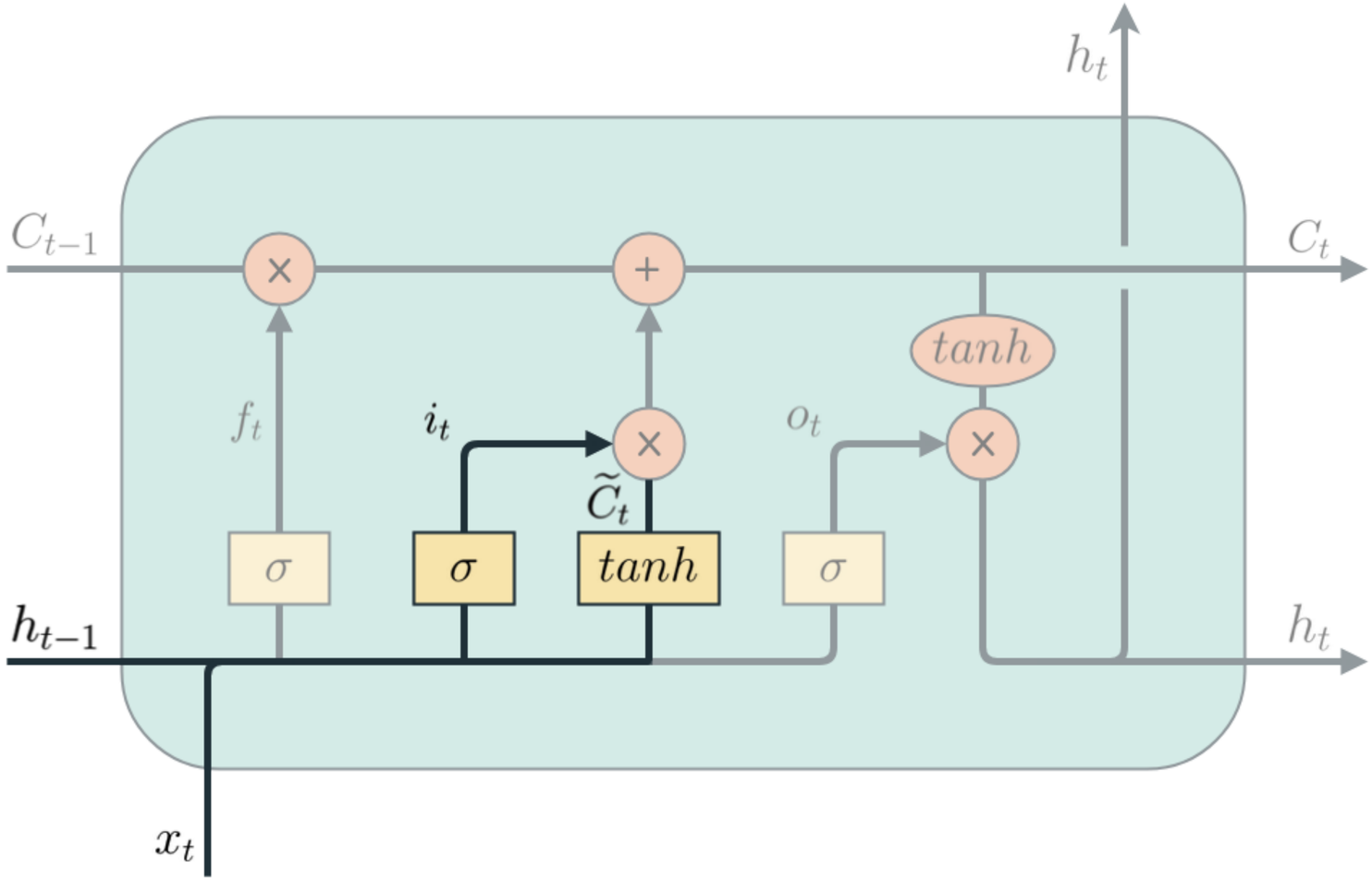
Module LSTM



$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

source: Understanding LSTM networks, C. Olah

Module LSTM

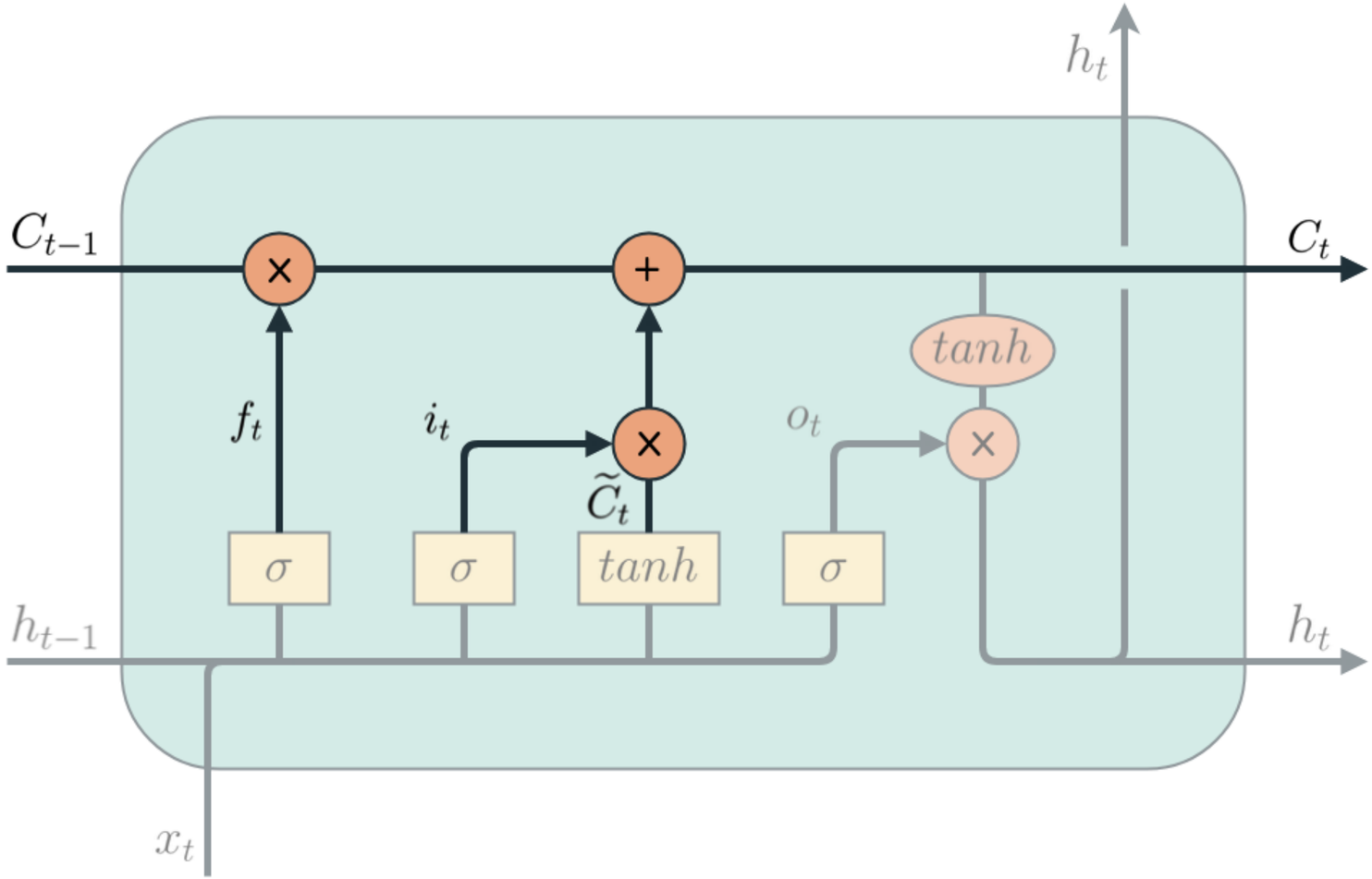


$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

source: Understanding LSTM networks, C. Olah

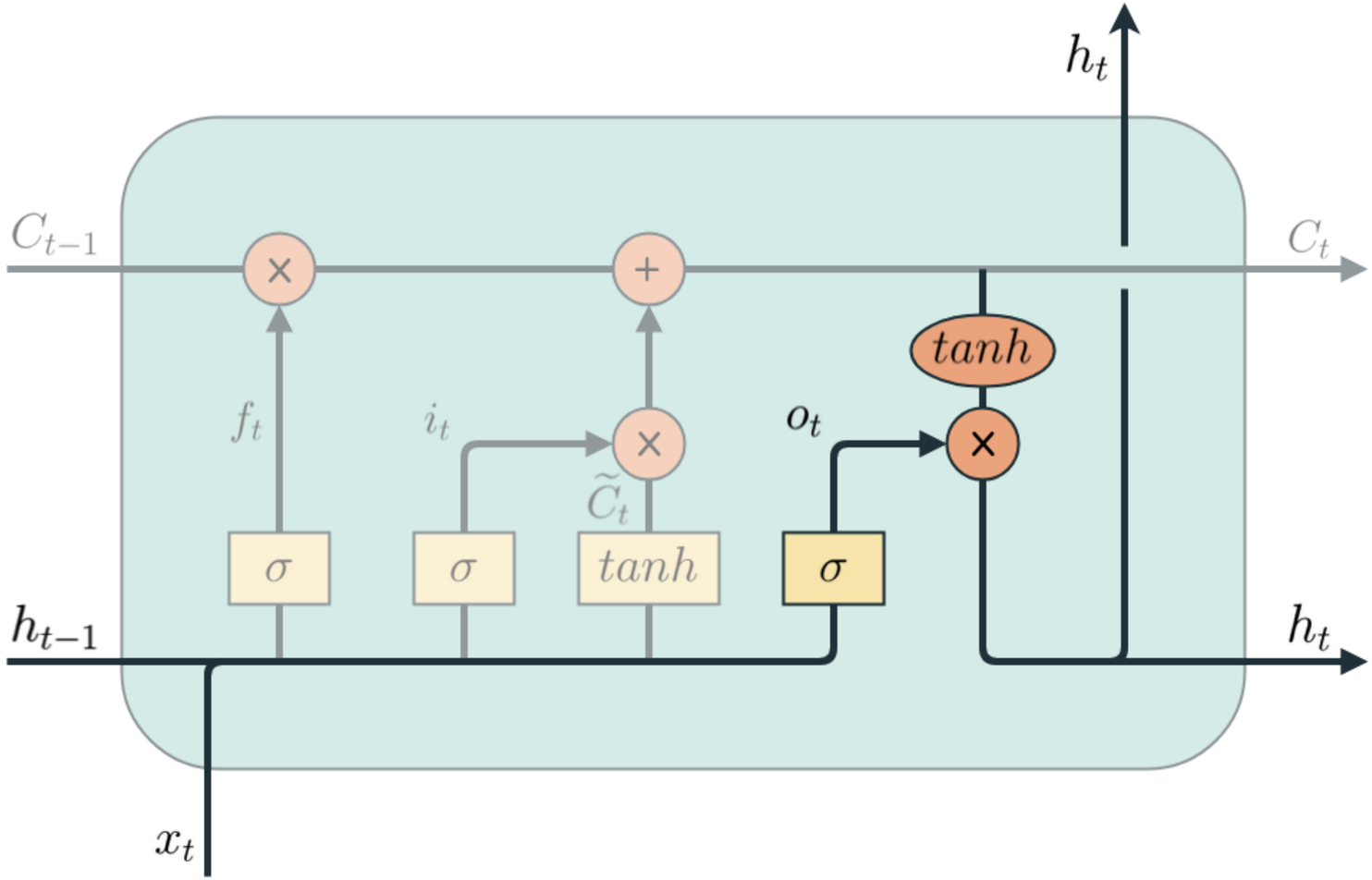
Module LSTM



$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t$$

source: Understanding LSTM networks, C. Olah

Module LSTM



$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * tanh(C_t)$$

source: Understanding LSTM networks, C. Olah

Les Conditional Random Fields pour traiter des séquences

- modèle probabiliste discriminant
- utiliser les informations de la séquence de labels à prédire

feature	$b(\boldsymbol{x},\ i) = \begin{cases} 1 \text{ si le mot } x_i \text{ est tout en majuscule} \\ 0 \text{ sinon} \end{cases}$
transition feature function	$t_j(\boldsymbol{x},\ i,\ l_i,\ l_{i\text{-}1}) = \begin{cases} b(\boldsymbol{x},\ i) \text{ si } l_i = \text{LAST NAME et } l_{i\text{-}1} = \text{FIRST NAME} \\ 0 \text{ sinon} \end{cases}$
state feature function	$s_j(\boldsymbol{x},\ i,\ l_i) = \begin{cases} 1 \text{ si } i = 1 \text{ et si } l_i = \text{FIRST NAME} \\ 0 \text{ sinon} \end{cases}$

- probabilité conditionnelle : $p(l|x,\lambda) = \dfrac{\exp(\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(x,i,l_i,l_{i-1}))}{\sum_{l'}\exp(\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(x,i,l'_i,l'_{i-1}))}$

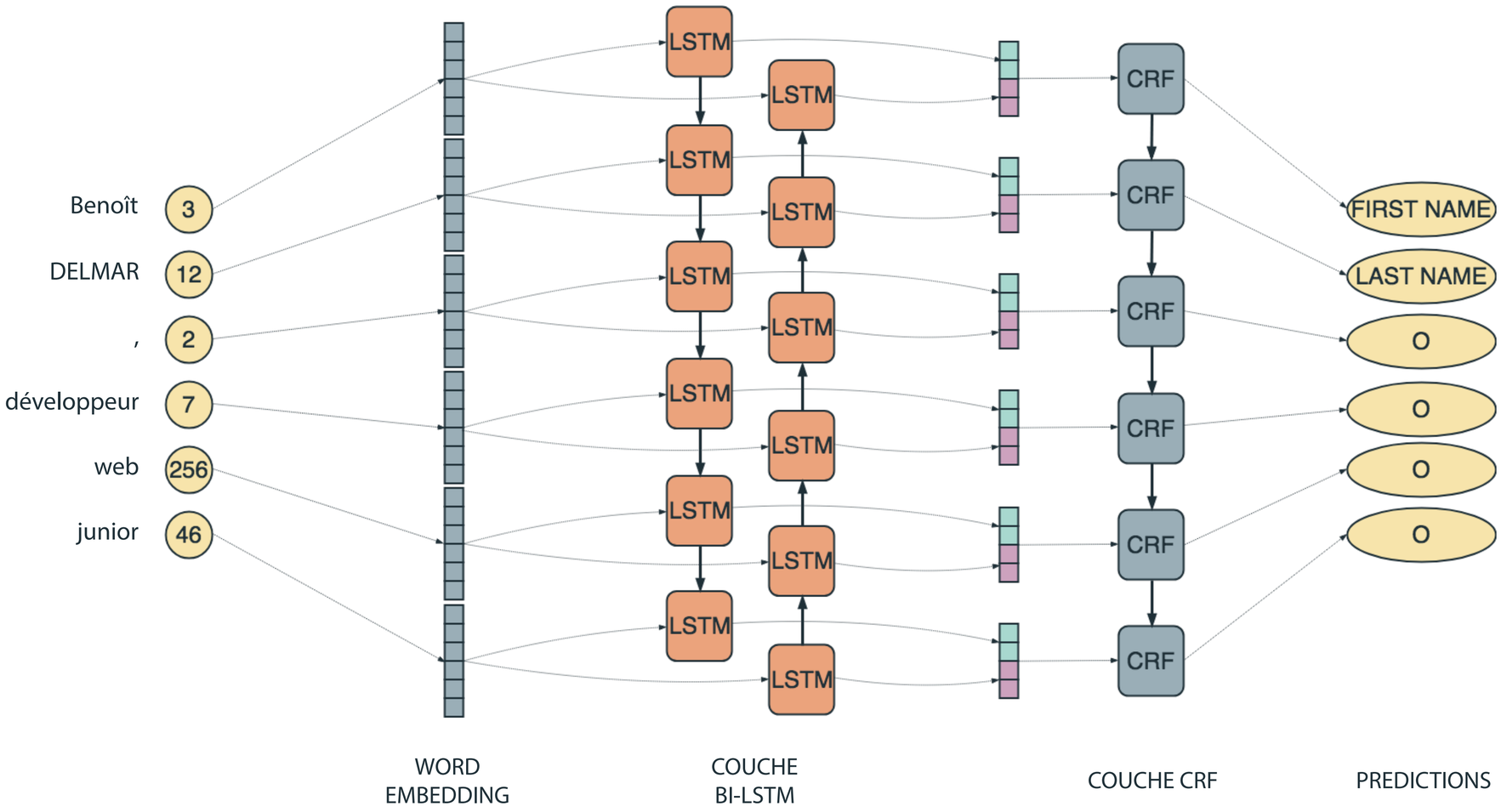
Un réseau Bi-LSTM + CRF

- Etat de l'art par Z. Huang (Baidu), 2015

- Recherche d'entités nommées
  - → couche LSTM bidirectionnel : capturer les features du contexte
  - → couche CRF : exploiter la logique d'enchaînement des labels

Structure du réseau mis en place

"Benoît DELMAR, développeur web junior"



Résultats pour les noms / prénoms

Baseline (régression logistique)

Précision	Rappel	F-score
0.883117	0.538732	0.669218

	Précision	Rappel	F-score
LAST NAME (567)	0.8364	0.3245	0.4676
FIRST NAME (569)	0.9049	0.7522	0.8215

Bi-LSTM + CRF

Précision	Rappel	F-score
0.877378	0.730634	0.797310

	Précision	Rappel	F-score
LAST NAME (567)	0.8678	0.7178	0.7857
FIRST NAME (569)	0.8868	0.7434	0.8088

Ressources pour approfondir le sujet

- Word2vec : Mikolov T., Chen K., Corrado G. & Dean J. (2013) "Efficient estimation of word representations in vector space"
- Glove : Pennington J., Socher R. & Manning C. D. (2014) "Glove: Global Vectors for Word Representation"
- Tutoriel de Chris McCormick sur Word2vec

- Hochreiter S. & Schmidhuber Jü. (1997) "Long short-term memory"
- Tutoriel de Christopher Olah sur les réseaux LSTM
- Lafferty J., McCallum A. & Pereira F. C.N., (2001) "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data"
- Hanna Wallach, Conditional Random Fields: An Introduction (2004)
- Huang Z., Xu W. & Yu K. (2015) "Bidirectional LSTM-CRF Models for Sequence Tagging"
- Tutoriel de Tobias Sterbak sur Sequence Tagging With A LSTM-CRF

Merci

justine@jobijoba.com

www.linkedin.com/in/justine-bel-letoile