


**EGC 2020..** Or, How I Learned to Stop  
Worrying and Love Clustering.

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

# OUTLINE

- Mise en situation
- Le jeu des données
- Préparation des données
- Topic modeling
- Clustering -KMeans-
- Distribution à travers le temps
- Conclusion

# Mise en situation

- ❑ L'association EGC propose un nouveau défi EGC pour la 20ème édition d'EGC qui aura lieu en janvier 2020.
- ❑ Le but : prédire les sujets de discussion des années à venir

# Les Données

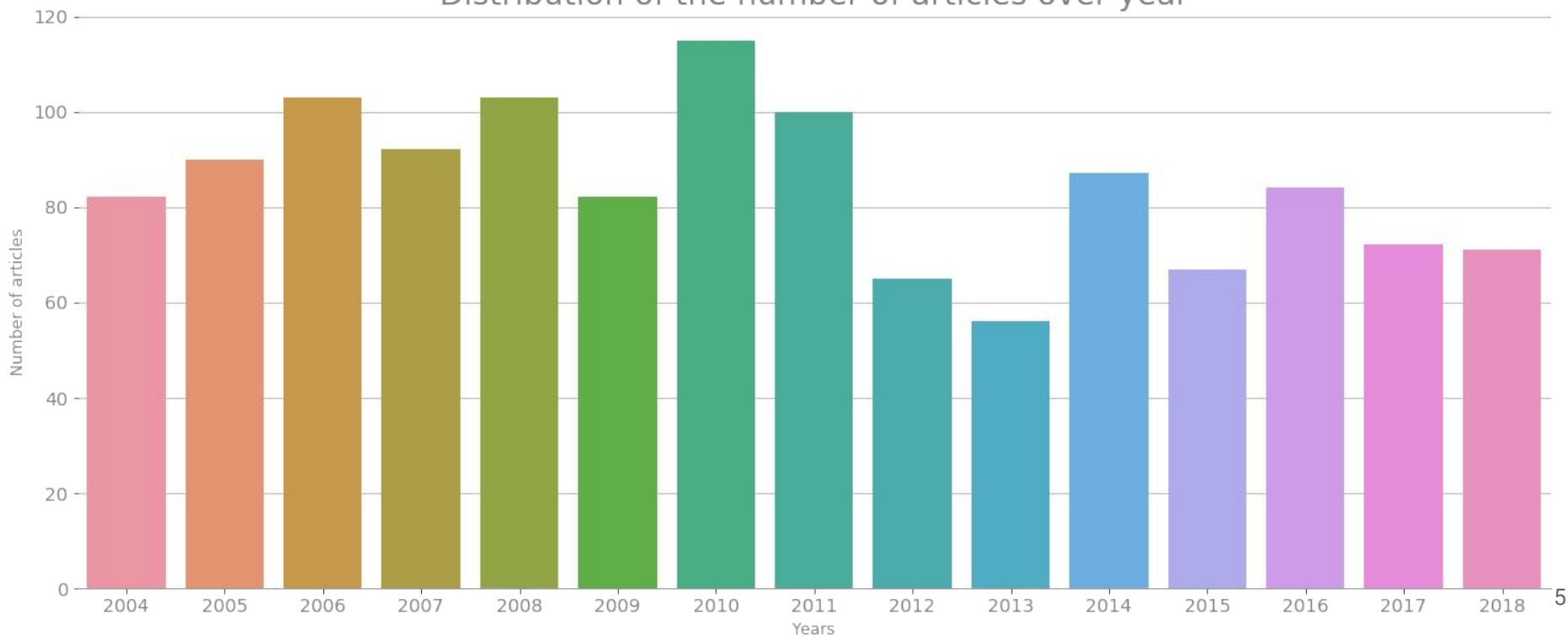
- ❑ 1269 articles scientifiques
- ❑ Sur la période de 2004 - 2018

Langage	Title	Abstract
Fr	1134	991
En	120	105
Autres	15	0

**= 1096**

# Les Données

Distribution of the number of articles over year



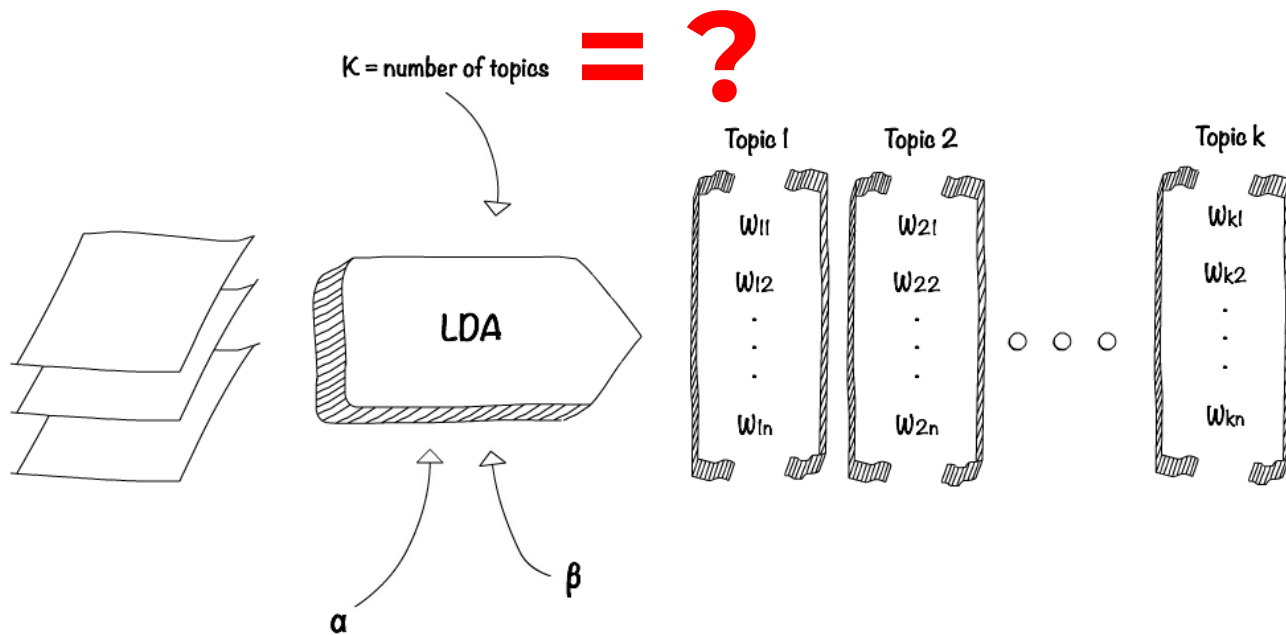
# Préparation des Données

- ❑ Traduction des 'Titles' et 'Abstracts' vers la langue française.
- ❑ Suppression des stop-words et de la ponctuation.
- ❑ Tokenisation
- ❑ POS tagging -- **StanfordPostTagger**
- ❑ Lemmatisation -- **French\_Lefff\_Lemmatizer**

# Préparation des Données

- ❑ Tri sur les tokens :
  - ❑ Présent dans moins de 90% des documents
  - ❑ On garde les 1000 premiers
- ❑ Tf.idf

# Topic modeling



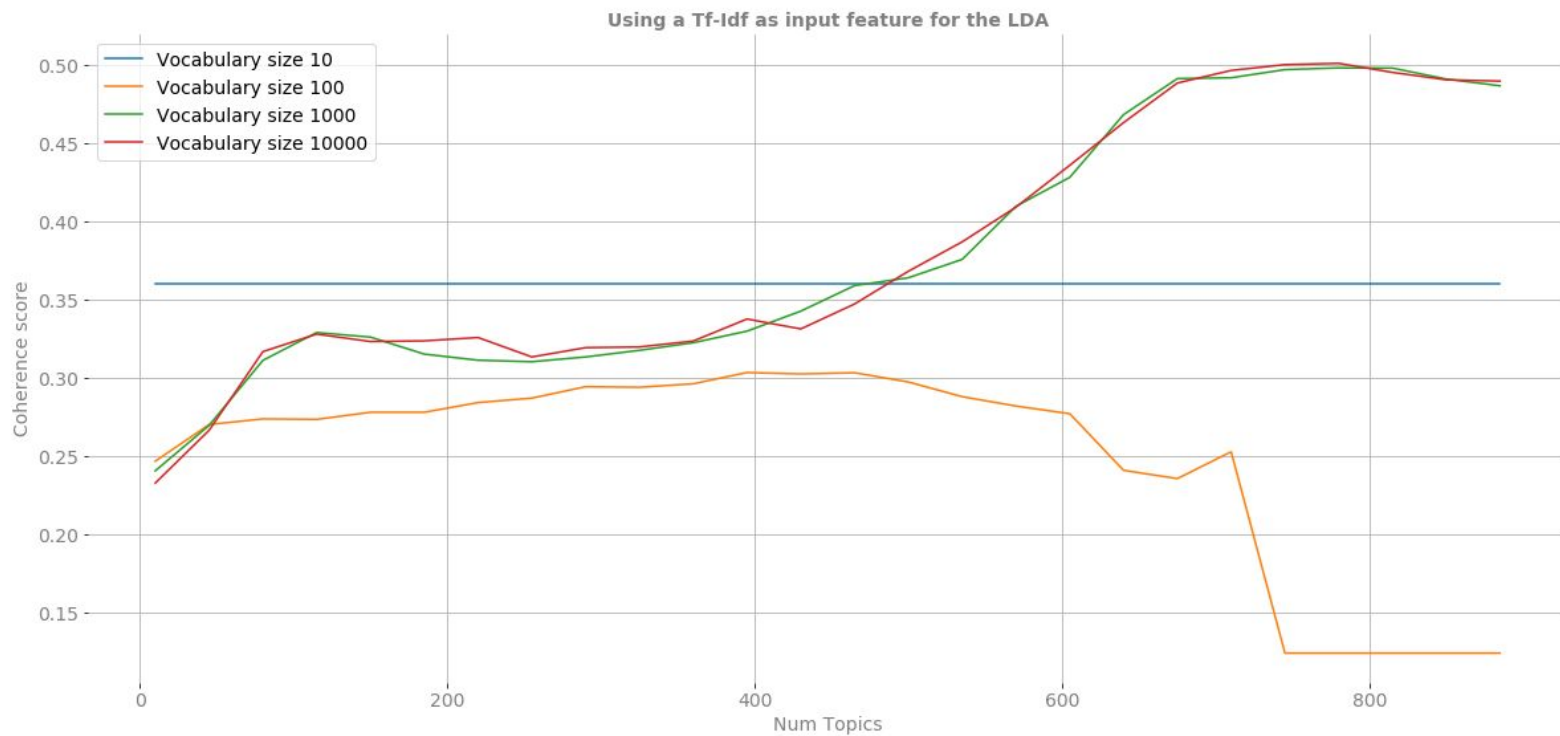


# Topic modeling



$$\text{Coherence} = \sum_{i < j} \text{score}(w_i, w_j)$$

Coherence score / Num of Topics in LDA



# Topic modeling



$$\text{Coherence} = \sum_{i < j} \text{score}(w_i, w_j)$$

Coherence score / Num of Topics in LDA

Using a Bag of words as input feature for the LDA



# Topic modeling



$$\text{Coherence} = \sum_{i < j} \text{score}(w_i, w_j)$$

## ❏ Evaluation

**#Topics = 50**



<TF-IDF>

**Coherence Score: 0.326**

# Clustering K-Means

## ❑ Pourquoi !?

- ❑ **K-means** et **LDA** sont des algorithmes d'apprentissage **non supervisés**.
- ❑ LDA assigne un document à un **plusieurs** de Topics.
- ❑ K-means va partitionner les documents en groupes **disjoints**.
  - ❑ Marche mal, **dur à évaluer**.

# Clustering K-Means

## ❏ Evaluation

- ❏ Silhouette Coefficient & Silhouette scores for each sample
- ❏ Intuitions

Pour 10 clusters :

**Silhouette Coefficient = 0.20**

0.71 - 1.0	A strong structure has been found
0.51 - 0.70	A reasonable structure has been found
0.26 - 0.50	The structure is weak and could be artificial
< 0.25	No substantial structure has been found

[Legend](#) on stackoverflow

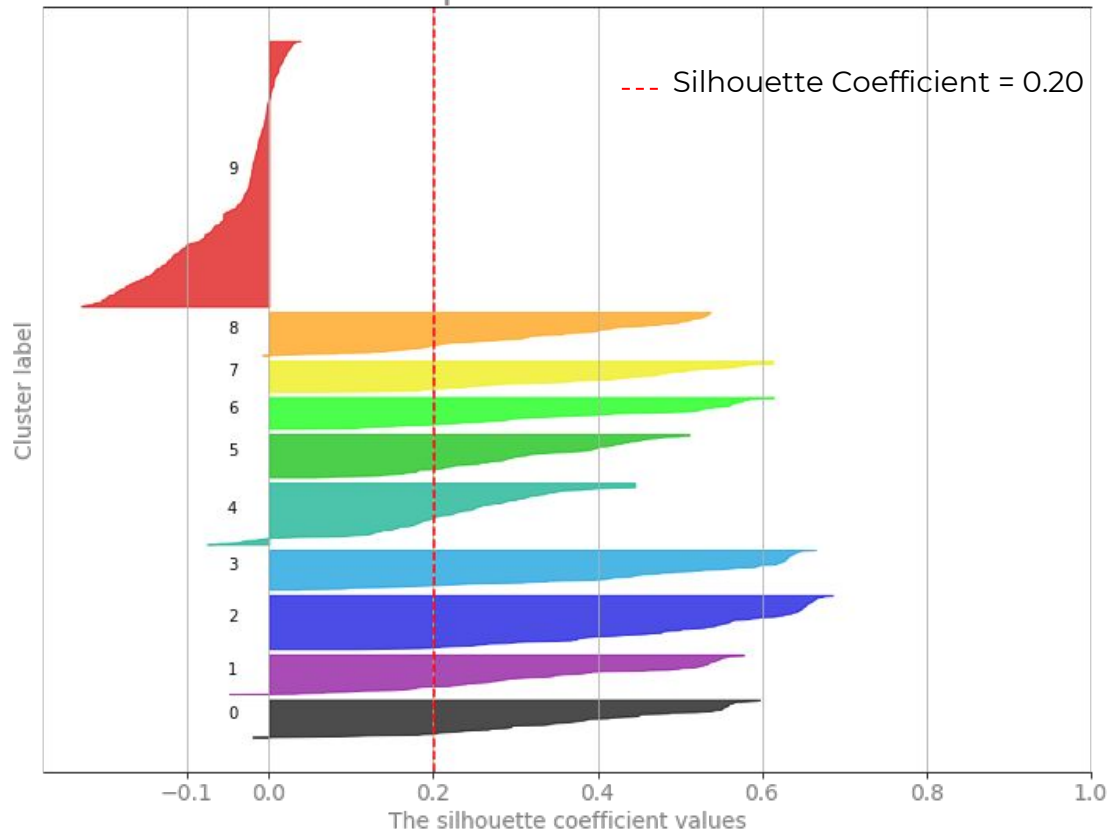
# Clustering K-Means

Garbage Cluster = Cluster 9

MAIS: 9 autres cluster dans le vert

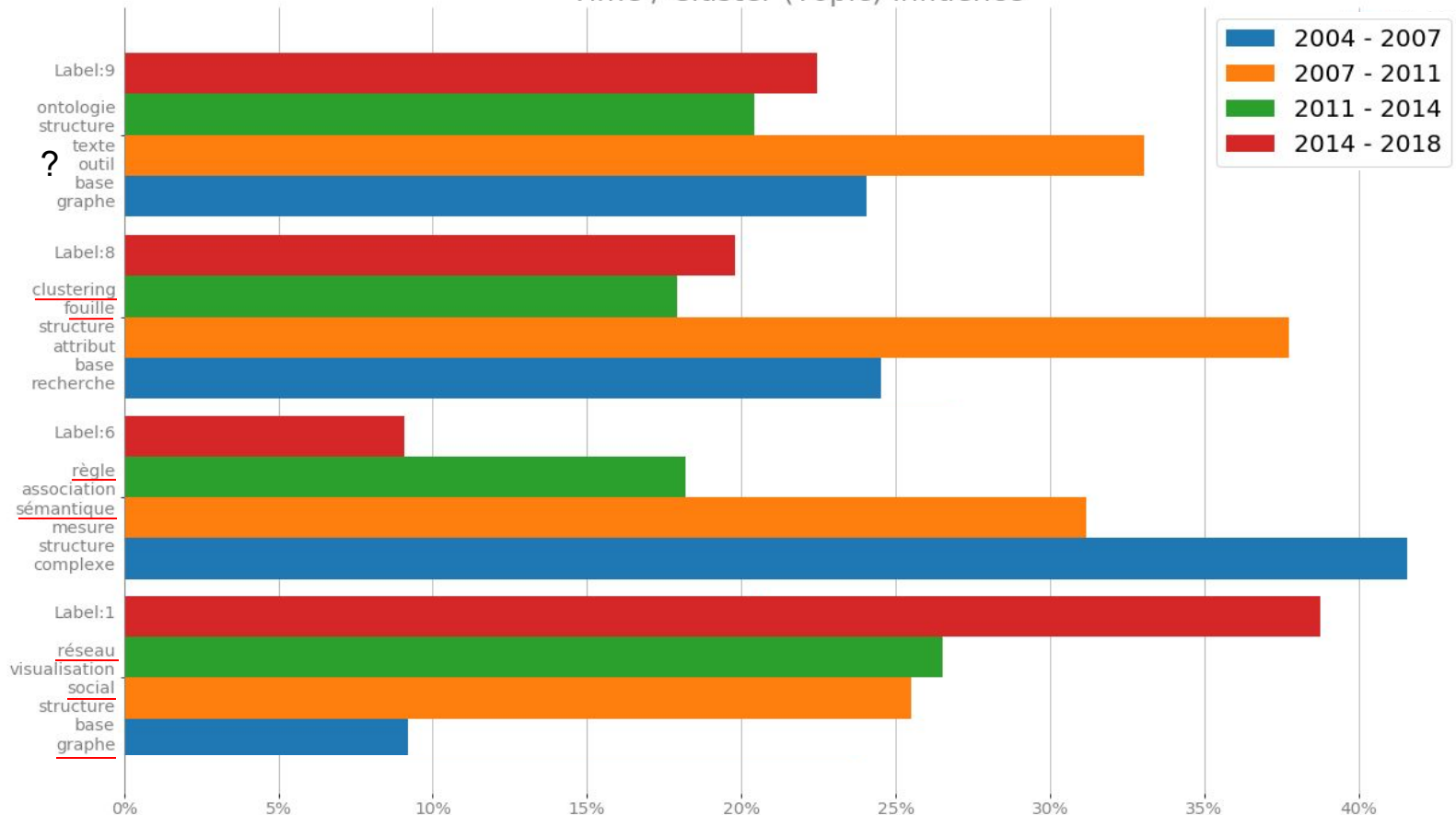
Silhouette analysis for KMeans clustering on sample data with  $n\_clusters = 10$

The silhouette plot for the various clusters.



# Distribution à travers le temps

Time / Cluster (Topic) influence



# Perspective

- ❑ Fuzzy K-Means
- ❑ Analyse diachronique avec le Raisonnement bayésien
- ❑ Doc2Vec



# Conclusion

- ❑ Peu de tendances, mais celles qu'on distingue sont clairement vérifiables
- ❑ Peut-être trop peu de document ?