



# Projet

## Grand Débat National

Chloé Clavel, Stephan Clémençon,
Florence d'Alché, James Eagan,
Alexandre Garcia, Marc Jeanmougin,
Adrien Nouvellet, Julien Romero

# Le contexte

## Cabinet de M. Sébastien Lecornu

- CNRS, INRIA, Télécom ParisTech
- 7/2/19 - Première réunion, proposition d'analyses
- Télécom ParisTech :
  - Recherche : INFRES, IDS, SES
  - Etudiants : MS Big Data, Filière SD, CES Data Scientist
- MS BGD :
  - Hackathon en P3
  - en P4, 'Données du Web' et 'Visualisation'

# Les Données du Grand Débat

- Divisé en 4 catégories (DEMOCRATIE ET CITOYENNETE, LA FISCALITE ET LES DEPENSES PUBLIQUES, LA TRANSITION ECOLOGIQUE, ORGANISATION DE L'ETAT ET DES SERVICES PUBLICS)
- 36 questions sur la démocratie, 8 questions sur la fiscalité, 16 questions sur l'écologie et 32 questions sur les services publiques
- 190.000 réponses de 94.000 utilisateurs
- Quelques données des utilisateurs : Id, code postal (autodéclaré), date de création, de modification,...

# Disponibilité des données

- Les données sont publiques et arriveront au fur et à mesure. Il faut donc développer et partager des pipelines pour l'analyse des données.
- Les données ainsi que des datasets complémentaires sont disponible sur `https://nextcloud.r2.enst.fr/nextcloud/index.php/s/cxeBz4WPoFLEM27`
- Les données sont disponible sur ElasticSearch sur `http://lame23.enst.fr:9200/` et sur Kibana `http://lame23.enst.fr:5601` (depuis Télécom)

# Données complémentaires

- Les données principales peuvent bénéficier d'autres sources
- Réseaux sociaux (Twitter, Reddit, forums de discussions...)
- Données structurées de Wikidata comme des informations démographiques
- Données sur les précédentes élections présidentielles
- (bientôt) Données du «  Vrai Débat »
- ... proposez d'autres sources potentielles :)

# Exemples de Visualisation

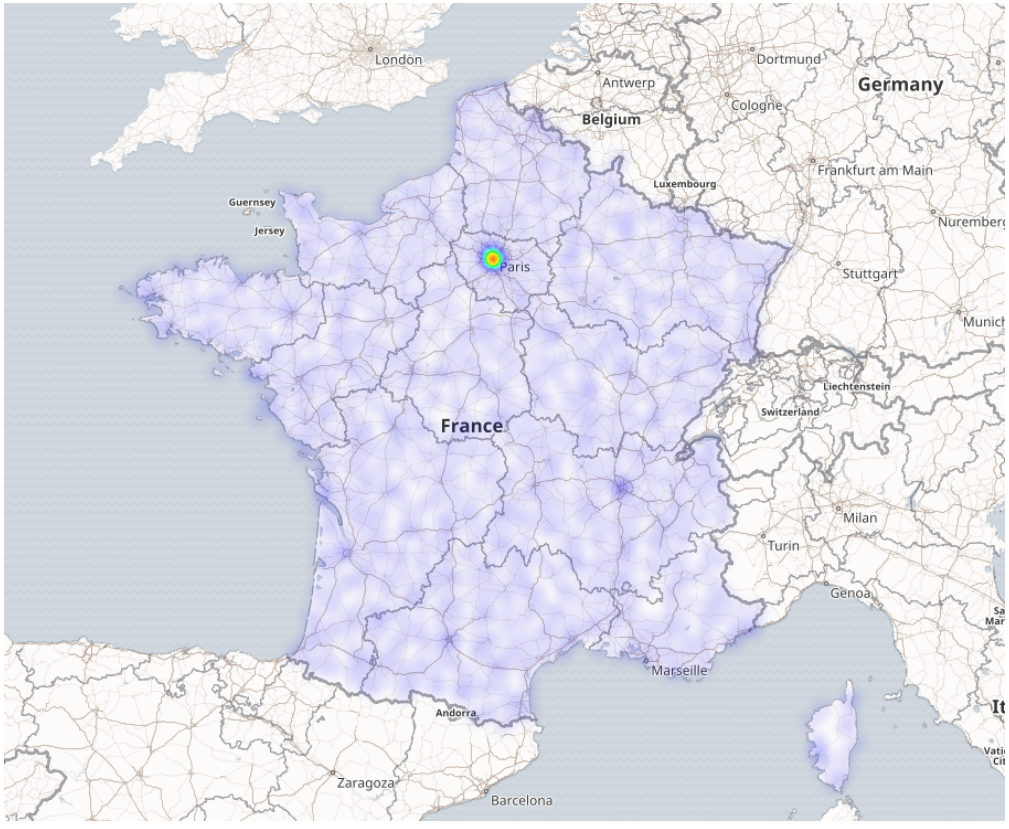



FIGURE – Heatmap

# Exemples de Visualisation

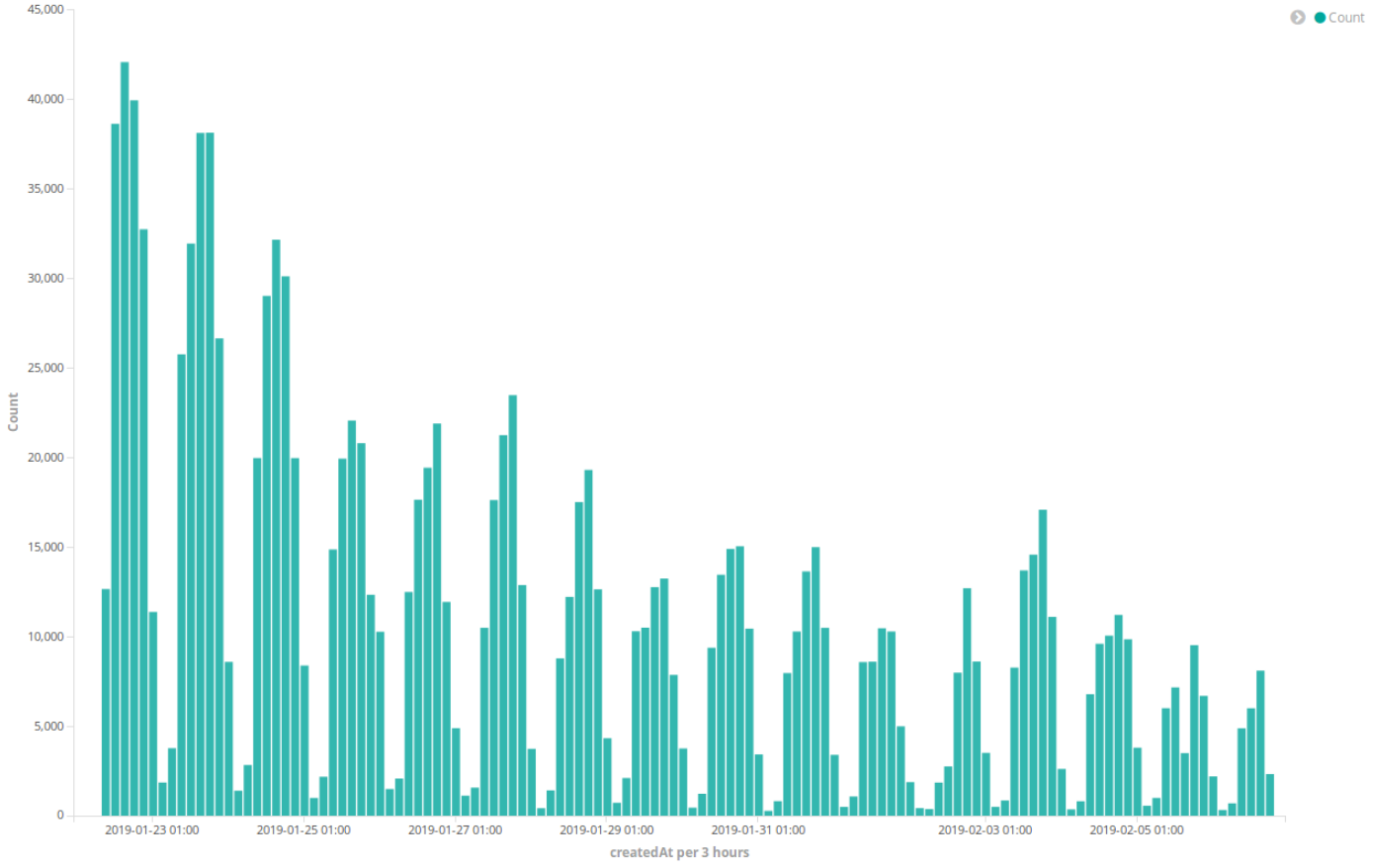



FIGURE – Dates des soumissions

# Exemples de Visualisation


FIGURE – Nuage de mots des titres

# Exemples de Visualisation


FIGURE – Distribution des contributions par code postal

# Soumettre nouvelles dimensions

- Par exemple, position GPS, analyses de sentiments,...
- Nous envoyer une description et le code pour les génèner, nous les ajouterons à Kibana pour que tout le monde en profite

# Analyses - Envisagées(1/2)

Liste d'analyses envisagées (énumération non-exhaustive)

- Modélisation de l'évolution des contribution dans le temps et / ou dans l'espace (modélisation des tendances [**?**], rééchantillonage) : Application à la détection de spams / publication coalisée.
- Mise en place d'une interface de requête intelligente (c.f. gensim LsiModel [**?**])
- Inférence du département/ville des réponses, visualisation d'une carte des contributions.

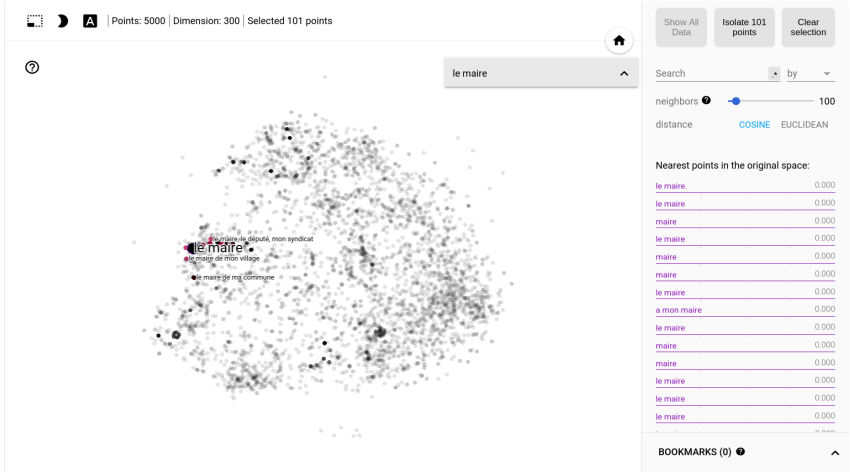
# Analyses - Envisagées (2/2)

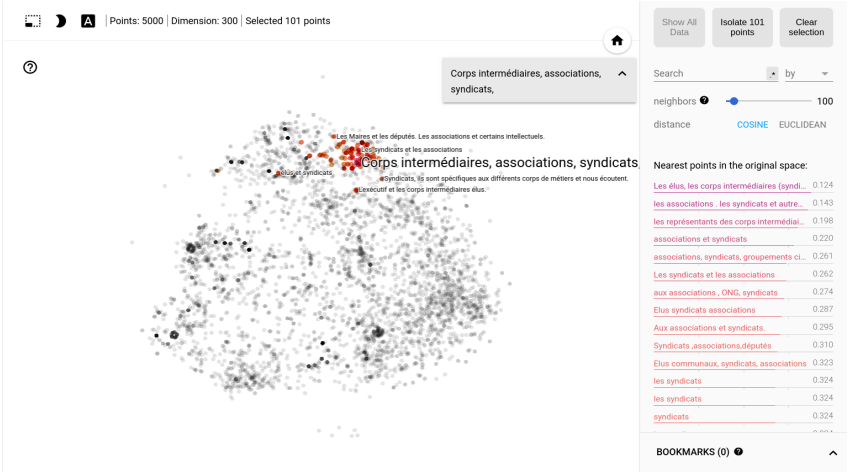
Liste d'analyses envisagées (énumération non-exhaustive)

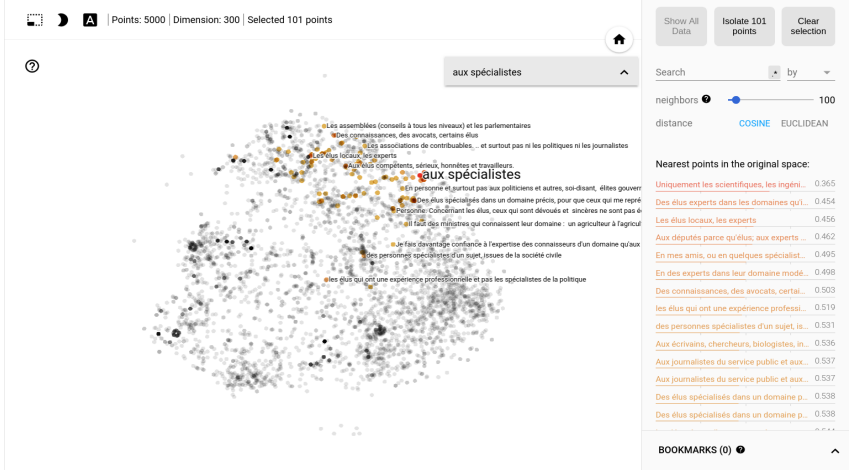
- Topic modeling (LDA [**?**])
- Analyse de sentiments (Ressources en français : Stanford Core NLP pour l'extraction de caractéristiques grammaticales et possibilité de coupler avec des modèles appris sur des datasets d'analyse de sentiment (twitter) ou des lexiques de polarité.
- Clustering des réponses (bag-of-words, word embeddings [**?**, **?**], sentence embeddings [**?**])

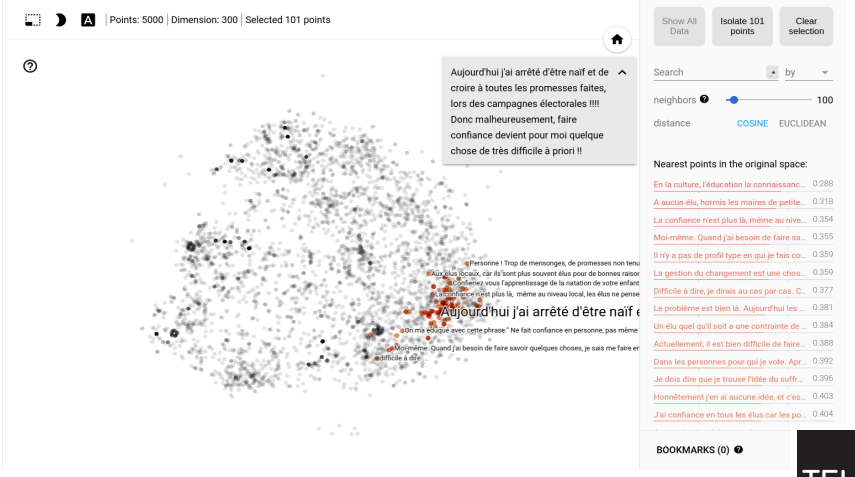
# Exemple : Visualisation Embeddings

"En qui faites-vous le plus confiance pour vous faire représenter dans la société et pourquoi ?"



# Modalités de Travail

- Formation de groupes de 3-4 personnes.
- Mise à disposition de serveurs par TeraLab
  - Accès à JupyterLab (compte à partager entre les membres de chaque groupe)
  - Environnement pré-installé : python3, panda, scikit-learn, matplotlib, numpy, nltk, scapy, tensorflow, keras...
  - Possibilité d'installer de nouveaux modules par le biais de pip
- Espace d'échange collaboratif : Discourse : `https://discourse.iscpif.fr` section « Les grands débats »et moodle `https://moodle.r2.enst.fr/moodle/course/view.php?id=60`.

# Encadrement

- Marc Jeanmougin (Infres)
  marc.jeanmougin@telecom-paristech.fr
- Julien Romero (Infres)
  julien.romero@telecom-paristech.fr
- Adrien Nouvellet (ex IDS, IT4PME)
  adrien.nouvellet@gmail.com
- SD-TSIA Alexandre Garcia (IDS)
  garcia@telecom-paristech.fr
- SD-TSIA Florence d'Alché (IDS)
  florence.dalche@telecom-paristech.fr
- SD-TSIA Alex Lambert (IDS)
  alex.lambert@telecom-paristech.fr

# Projects for TSIA-SD210

1. T1 : Dimension reduction, representation learning, metrics for clustering
2. T2 : Clustering and cluster interpretability with decision trees
3. T3 : Supervised learning for location prediction from the text content
4. T4 (more advanced) : topic modeling
5. T5 (more advanced) : distribution modeling of text contents with GANNs
6. TN : N >5, any topics related to the lectures you propose by email.

# Projects for TSIA-SD210

- Final Deadline : April 12
- Midterm deadline : March 15 (notebook, first analysis)
- Office hours (1H30/ 2 persons/week), starting March 4
- Project choice : February 20