Analyser les tweets de chefs politiques canadiens grâce à l'Analyse Topologique de Données

Maxime Turgeon

18 Mars 2021

Université du Manitoba
Départements de Statistique et d'Informatique

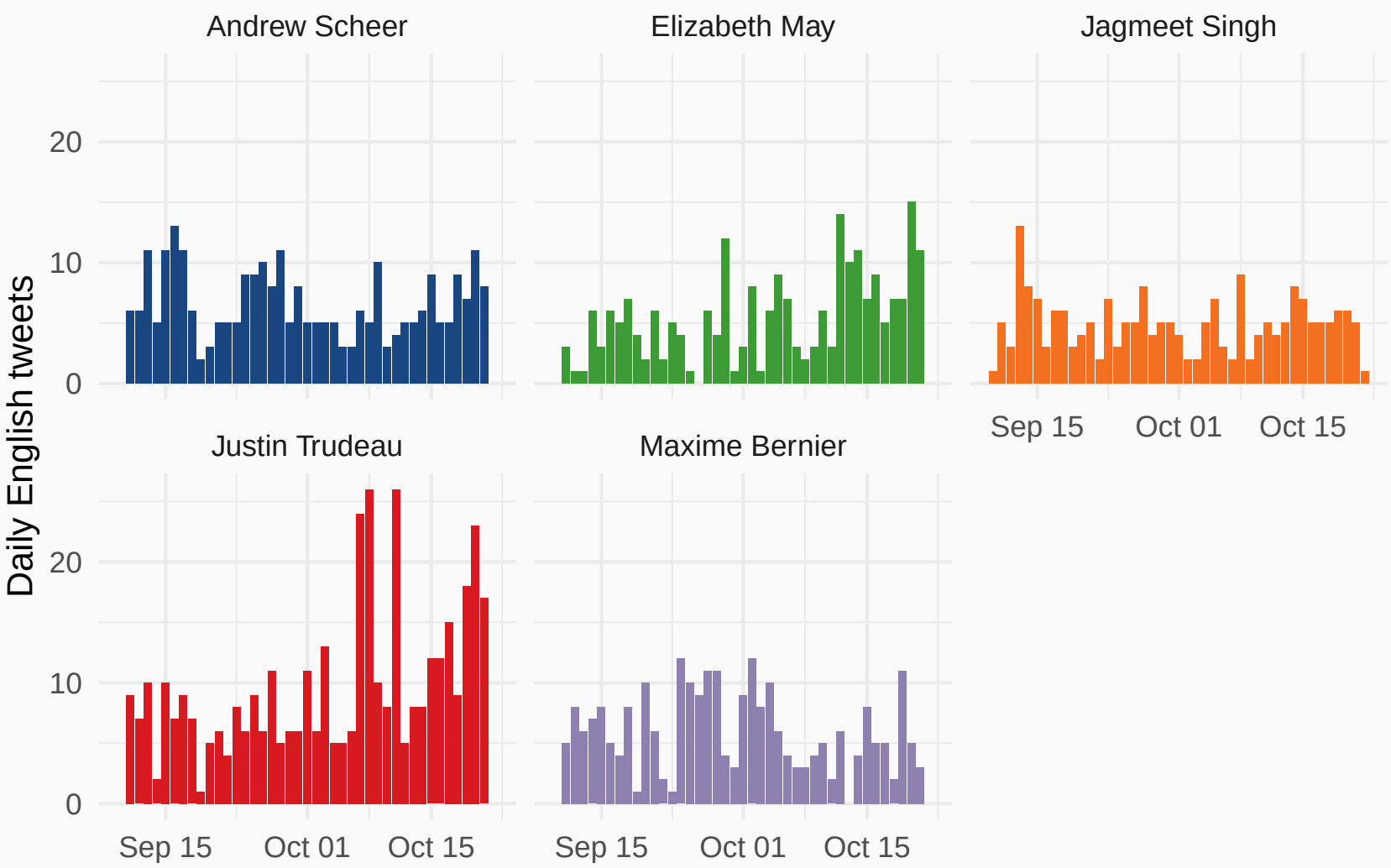
Motivation

- Pour mon cours d'analyse multivariée appliquée, je demande aux étudiants d'analyser un jeu de données de leur choix.
- L'an dernier, un étudiant (J. Hamilton) a choisi les tweets des chefs politiques
- **Objectif principal**: Peut-on recouvrir les thèmes des débats seulement à partir du texte?
    - Conclusion: oui pour certains thèmes, mais le signal est faible.
- Une question persiste: *Quelle est la meilleure façon d'analyser de telles données?*

Données

- 1356 tweets en anglais des 5 chefs principaux:
  - Andrew Scheer (CPC), Elizabeth May (GPC), Jagmeet Singh (NDP), Justin Trudeau (LPC) et Maxime Bernier (PPC)
  - Yves-François Blanchet (BQ) est omis; un seul tweet en anglais
- Tweets entre 10 Septembre et 22 Octobre 2019.
- Trudeau tweete le plus souvent (401), Singh tweete le moins souvent (210)

Tweets par jour



Préparation des données

- Chaque tweet est divisé en un ensemble de mots (modèle "bag-of-words")
- Hashtags, mentions, stop-words, emojis, et chiffres sont omis.
- Tweets avec moins de 4 mots sont omis.
- *Résultat*: une matrice document-mot 1256 par 3932.

Mots les plus communs

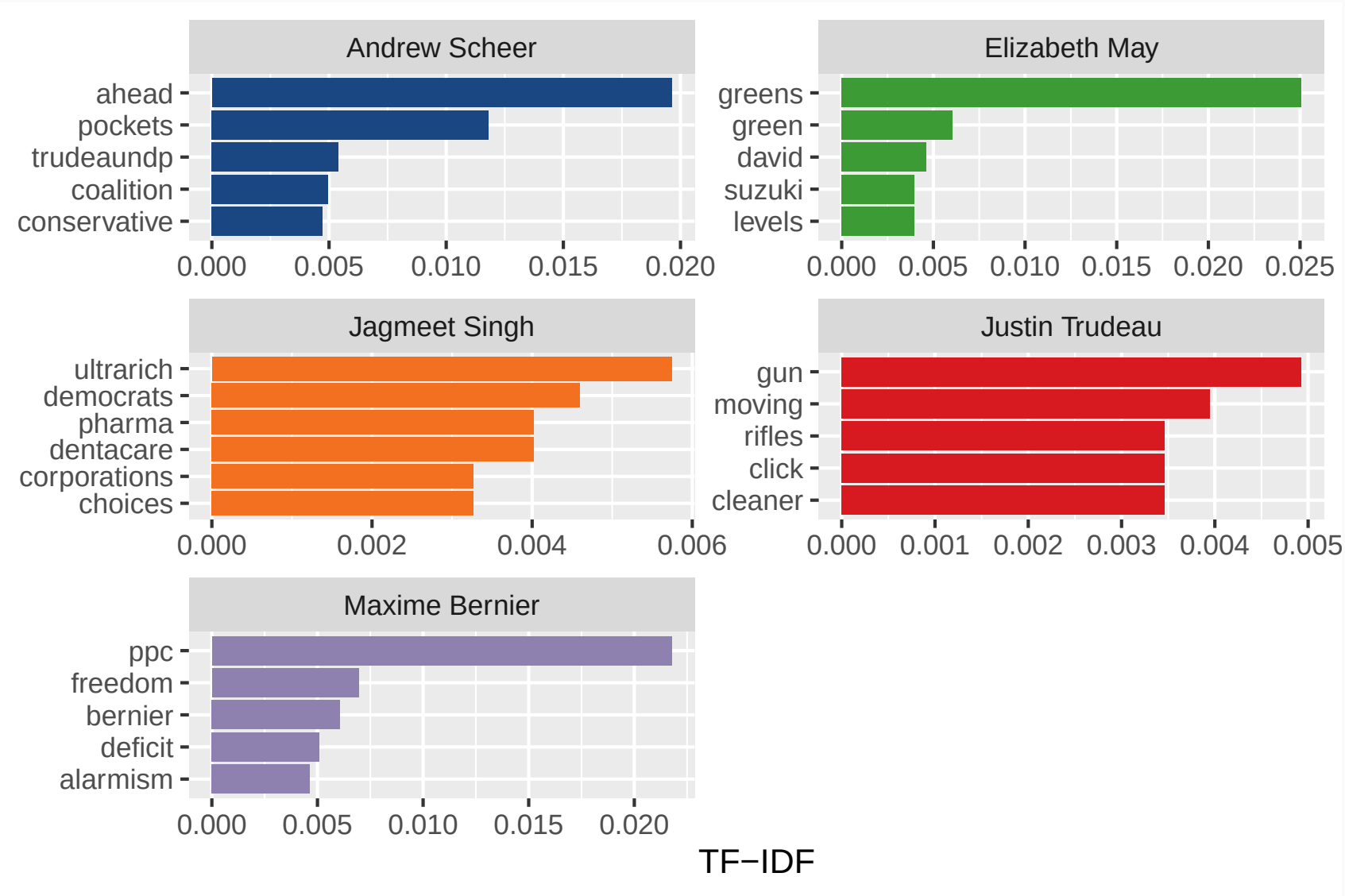
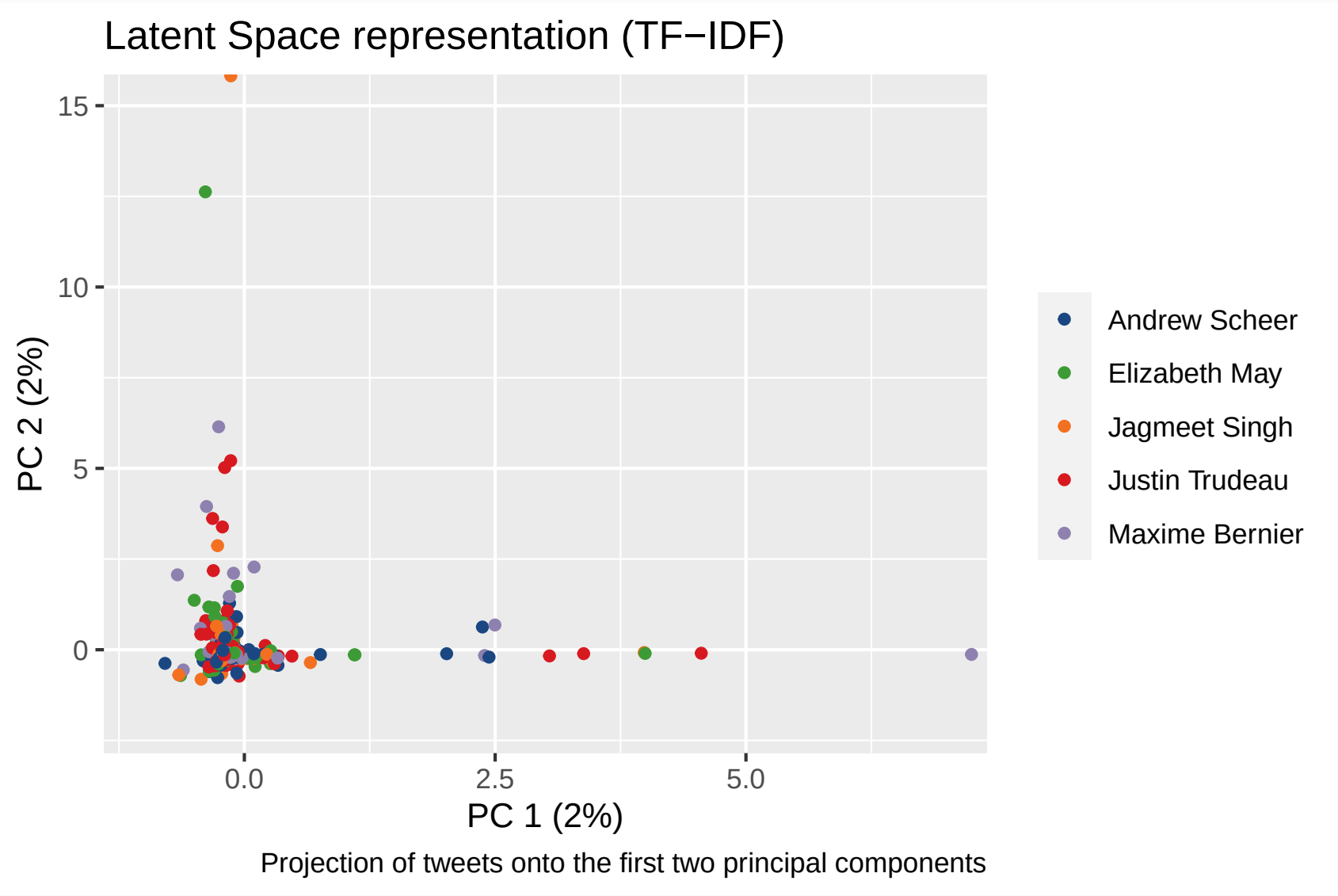


Figure 1: TF-IDF: Term frequency-Inverse Document Frequency

Analyse en composantes principales



Projection of tweets onto the first two principal components

Observations

- PCA ne fonctionne pas très bien...
  - Les rayons sont typiques de données non-gaussiennes (e.g. Rohe & Zeng, 2020).
- La matrice est très éparse (plus de 99%!!!)
- Les données sont de haute dimension.

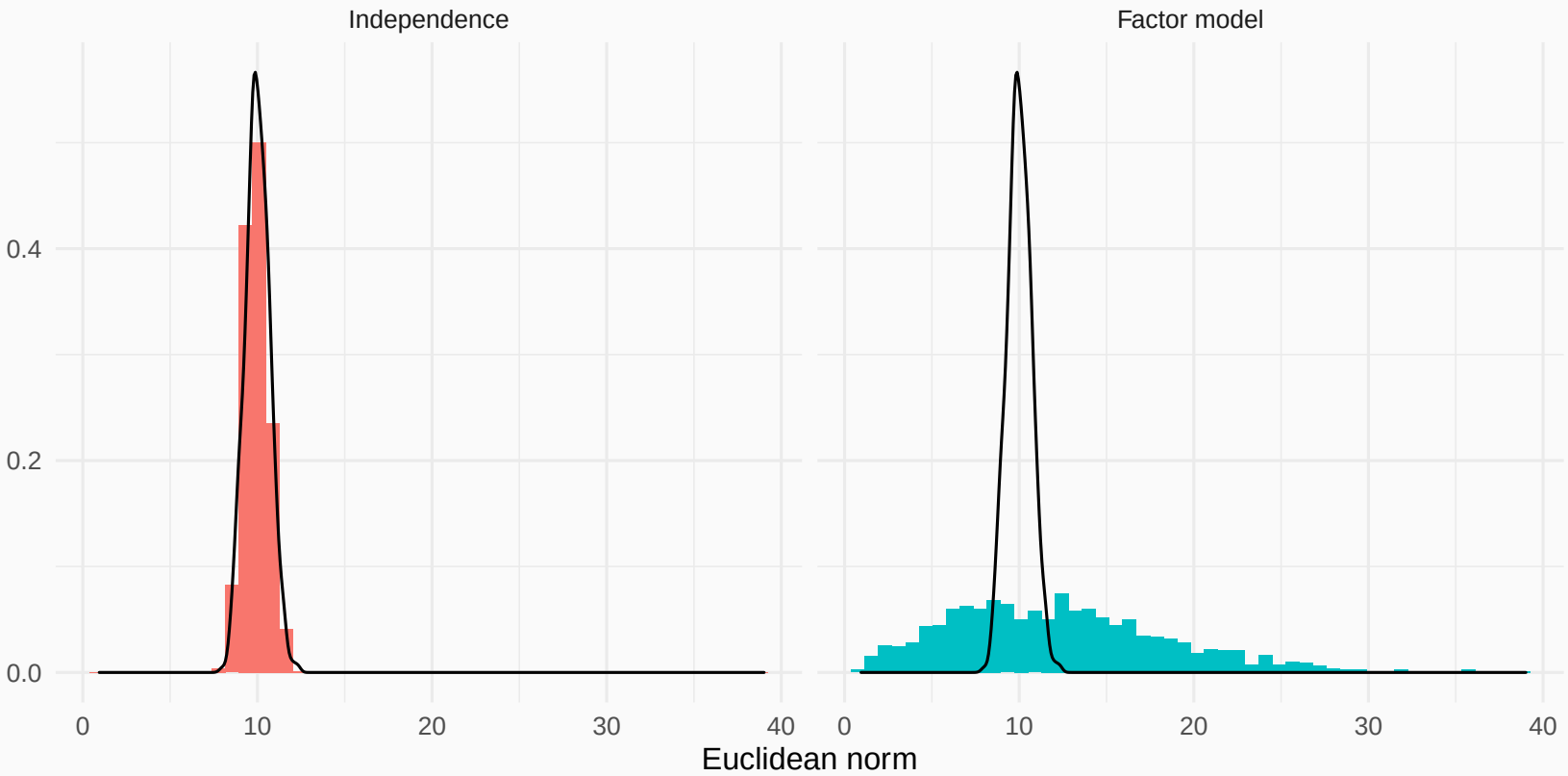
Malédiction de la Dimensionalité

Les données de haute dimension souffre de la **malédiction de la dimensionalité**:

- Beaucoup d'espace vide, les voisins sont loins...
- Les observations sont majoritairement loin de l'espérance.
- Défis computationels

La malédiction est-elle si terrible?

Exemple—Norme en haute dimension



“Big data matrices are approximately low rank”

- Empiriquement, ce n’est pas ce qu’on observe...
- **Pourquoi?** Ces résultat maudits requièrent l’indépendance des variables.
- Ce qui est presque toujours faux.
- Les grandes matrices de données ont généralement de petits rangs (Udell and Townsend, 2019)

Mitiger la malédiction

Plait-il? Mitiger en exploitant la *structure* de nos données.

- Utiliser des méthodes sparses (e.g. régression avec penalité, lasso graphique)
- Utiliser des estimateurs de covariance structurés (e.g. Turgeon, Oualkacha, *et al*, 2018)

Analyse Topologique des Données

- L'**Analyse Topologique des Données** (TDA) permet d'étudier la géométrie du support et ses contraintes.
- Mélange de géométrie et d'algèbre linéaire computationelles, etc.
- Nous nous concentrerons sur une technique en particulier: l'**homologie persistante**.

Rappel de topologie algébrique

- La **topologie**: étudier des objets géométriques indépendemment d'un système de coordonnées ou de fonctions de distance.
  - Géometrie "primitive"
- La **topologie algébrique** utilise des outils d'algèbre linéaire et abstraite pour étudier et classifier ces objets.
- L'**homologie** attache une séquence d'espaces vectoriels à chaque objet.
  - On appelle **nombres de Betti** $(\beta_0, \beta_1, \ldots)$ la dimension de chaque espace.
  - Préservés sous homéomorphisme.
- Les nombres de Betti dénombrent des caractéristiques topologiques importantes:
  - Composantes connexes, trous, cavités, etc.

Et les données?

- Supposons que le support de nos données de haute dimension p est une *variété* de dimension $k \ll p$.
- Un ensemble fini n'est pas très intéressant...
- **Idée**: À partir de nos données, construire un *complexe simplicial*, ce qui est beaucoup plus intéressant!

Complexe simplicial

- Un **complexe simplicial** est un espace topologique construit à partir de l'enveloppe convexe de sous-ensembles de points (appelé un simplexe).
  - Plus quelques règles pour assurer une certaine cohérence
- Un point est un simplexe de dimension 0
- Un segment est un simplexe de dimension 1
- Un triangle est un simplexe de dimension 2
- etc.

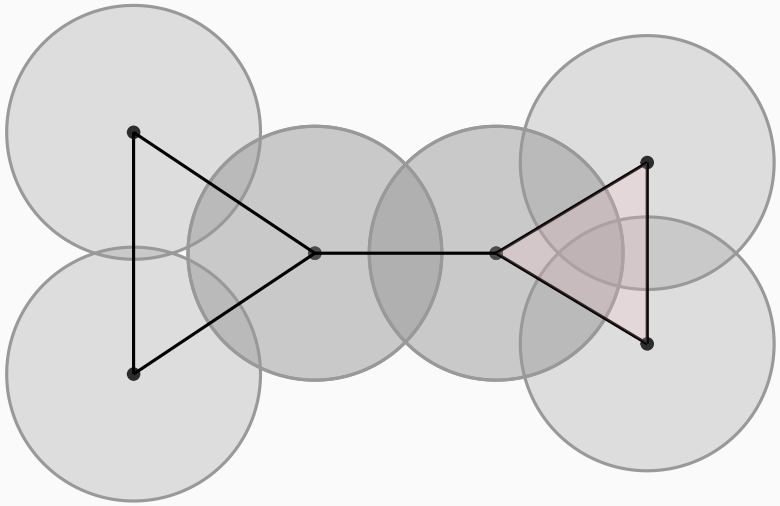
Complexes de Cech et Vietori-Rips

Considérons deux constructions possibles (soit $r > 0$):

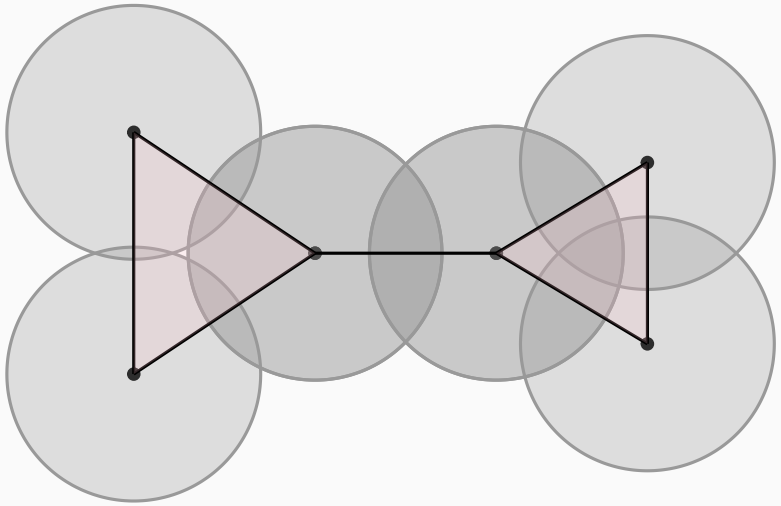
- **Complexe de Cech**: Chaque observation est un simplexe; un sous-ensemble de k observations est un simplexe si l'intersection de boules ouvertes de rayon r autour d'eux est non-vide.
- **Complexe de Vietori-Rips**: Chaque observation est un simplexe; un sous-ensemble de k observations est un simplexe si la distance entre chaque paire est au plus $2r$.

Complexes de Cech et Vietori-Rips

Cech



Rips



Théorème du nerf

- Un théorème important en topologie algébrique (le *théorème du nerf*) peut être utilisé pour prouver des résultats de cohérence pour la topologie du complexe de Cech et pour le support des données.
- **Et Vietori-Rips?** Puisque

$$\mathrm{Rips}_r(\mathcal{X}) \subseteq \mathrm{Cech}_r(\mathcal{X}) \subseteq \mathrm{Rips}_{2r}(\mathcal{X}),$$

  on obtient directement des résultats de cohérence pour la topologie du complexe de Vietori-Rips.

Avec une taille d'échantillon suffisante, un support "lisse", et un rayon r approprié, il est possible de calculer les nombres de Betti du support en utilisant le complexe de Cech (ou Vietori-Rips).

Homologie persistante

- **Comment choisir** r**?** Pas besoin de choisir!
- **L'homologie persistante** permet de calculer l'homologie d'une séquence d'espaces topologiques:

$$\mathrm{Cech}_{r_1}(\mathcal{X}) \subseteq \mathrm{Cech}_{r_2}(\mathcal{X}) \subseteq \mathrm{Cech}_{r_3}(\mathcal{X}) \subseteq \cdots$$

- Pour de petits r: $\beta_0 = n; \beta_k = 0$ for $k \geq 1$
- Pour de grands r: $\beta_0 = 1; \beta_k = 0$ for $k \geq 1$
- On cherche les caractéristiques topologiques qui *persistent* sur de longs intervalles de r.

Codes à barre et diagrammes

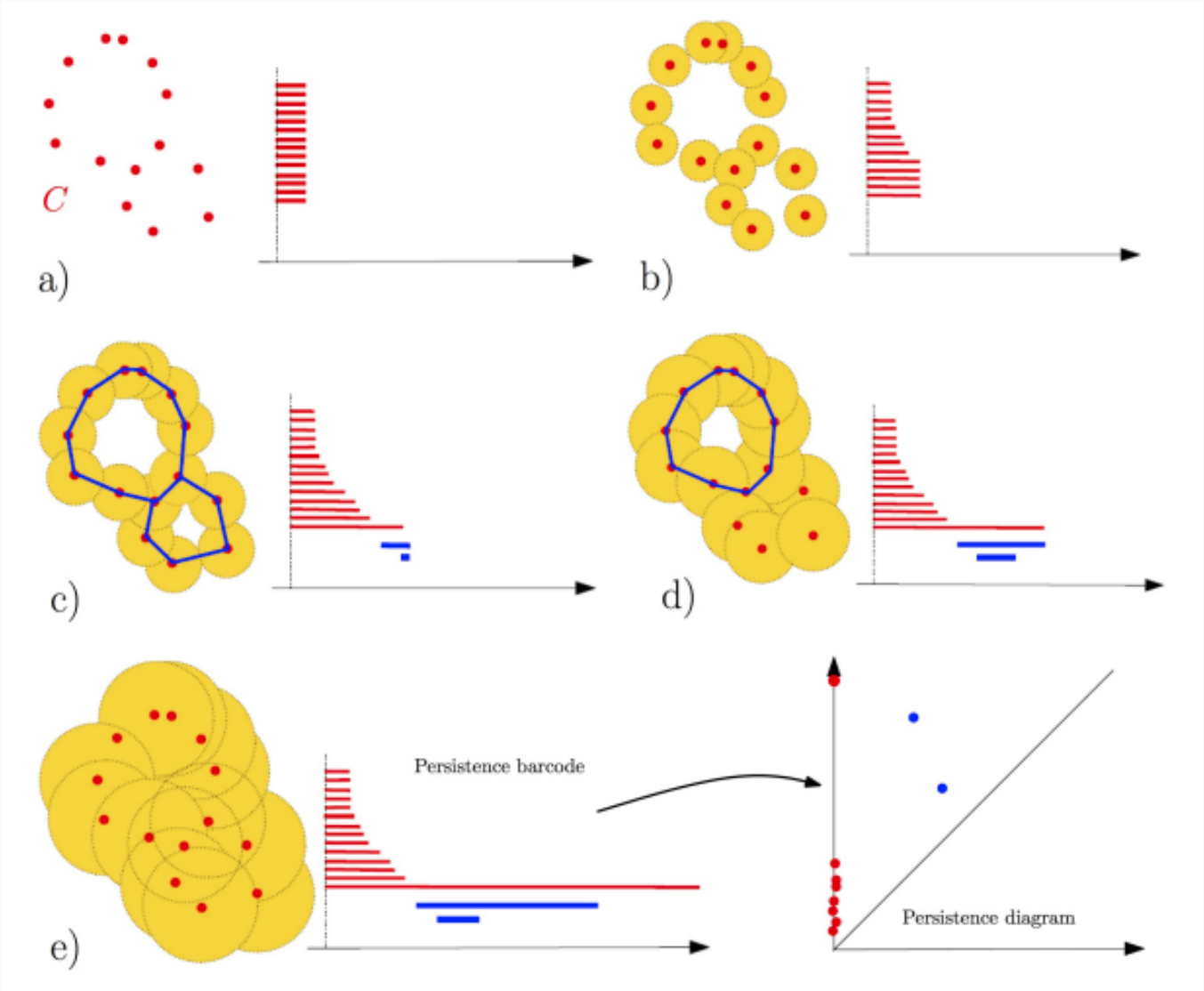
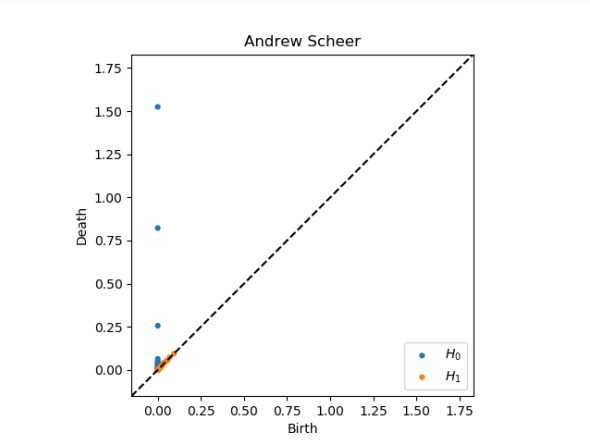


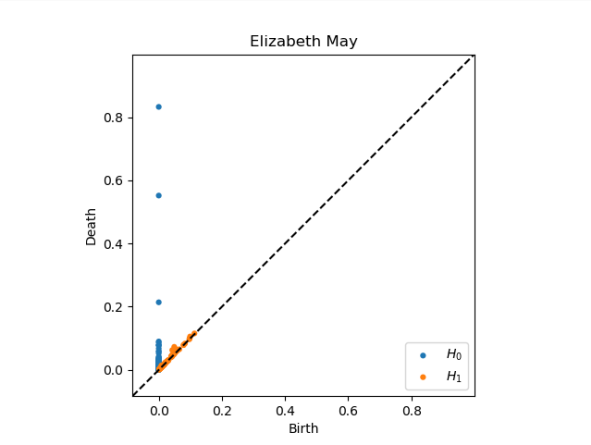
Figure 2: Chazal & Michel, Arxiv 2017

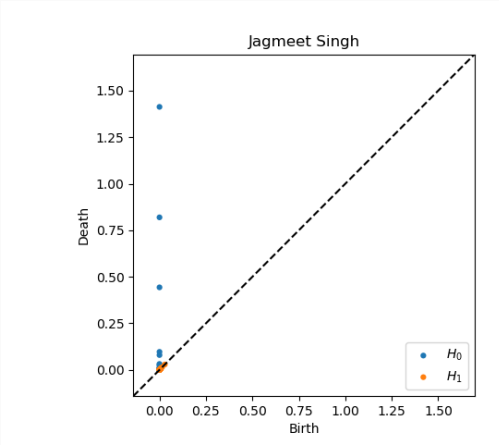
Analyse de données

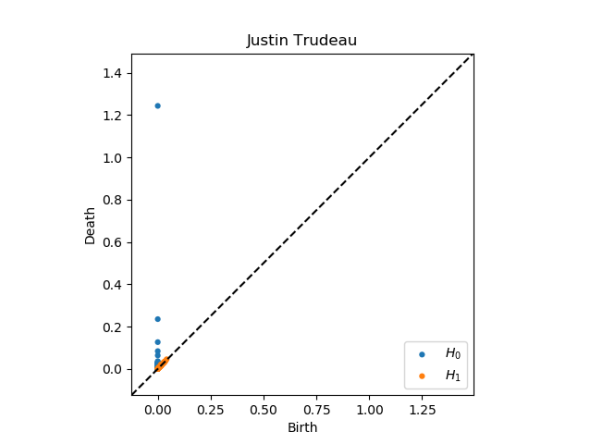
- Diviser la matrice document-mot en 5 sous-matrices
  - Une pour chaque chef
- Réduire la dimension avec PCA
- Calculer l'homologie persistante pour les données réduites
- Module `scikit-tda` en Python

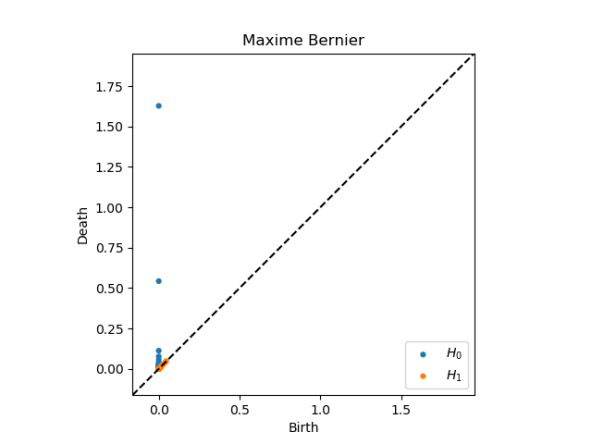
Diagrammes de persistance







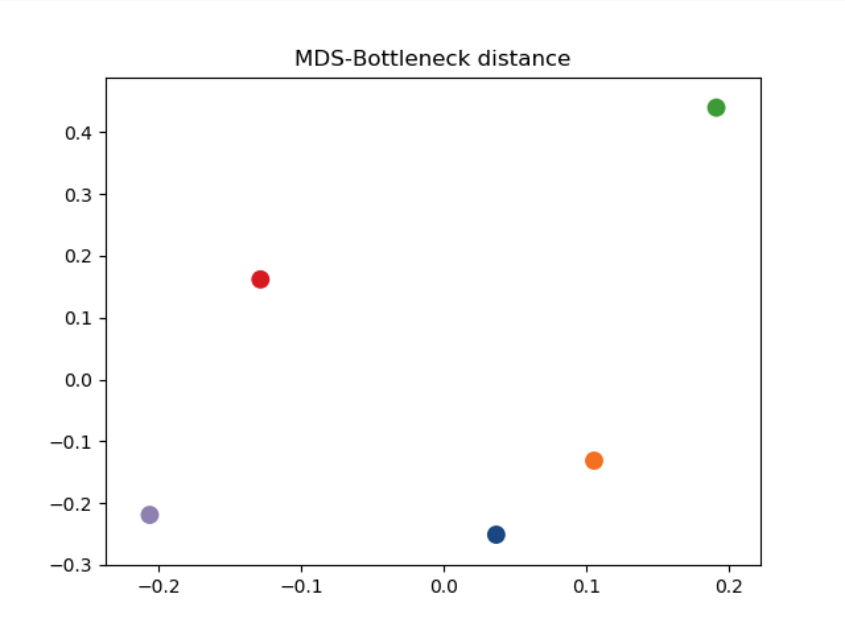




Visualisation de la mesure goulot

Mesure goulot ("Bottleneck distance") pour comparer l'homologie de degré zéro + Analyse Multidimensionelle.

	AS	EM	JS	JT	MB
AS		0.67	0.20	0.42	0.30
EM	0.67		0.56	0.39	0.78
JS	0.20	0.56		0.41	0.28
JT	0.42	0.39	0.41		0.39
MB	0.30	0.78	0.28	0.39	



Discussion

- Les chefs les plus distants dirigent les plus petits partis (EM et MB)
- AS et JS ont un but semblable: remplacer JT. Explication possible pour leur proximité?
- JS est plus près de EM que MB, et c'est le contraire pour AS.

Discussion

- L'homologie persistante peut être utilisée pour extraire des variables latentes des données textuelles (Gholizadeh *et al*, 2020), pour grouper et classifier des documents (Guan *et al*, 2016).
- Les diagrammes de persistance admettent une immersion dans un espace d'Hilbert, et ils peuvent donc être combinés avec des méthodes à noyaux pour construire des modèles de prd́iction
- Autres applications: analyse de rèseaux sociaux (Almgren *et al*, 2017), analyse de point de changement (Islambekov *et al*, 2019), étude des réseaux neuronaux profonds (Gebhart *et al*, 2019)

Comment améliorer

- La sparsité est une caractéristique des données textuelles, mais la lemmatisation peut aider à réduire son impact.
- Bootstrap et diagrammes de persistance.
- Mesurer l'importance de certains mots à partir de leur impact sur l'homologie?

Projets futurs

- Quel est l'impact de la réduction dimensionnelle sur les calculs topologiques?
- Jusqu'à quelle dimension peut-on pousser ces algorithmes?
- Peut-on reconstruire la variété? Où sont les trous?

Questions ou commentaires?

Pour plus d'informations, visitez

`maxturgeon.ca`.