

AUGMENTATION DE DONNÉES POUR L'EXPLORATION DE DOSSIERS MÉDICAUX

Joris Falip, Amine Aït Younes, Frédéric Blanchard, Michel Herbin
joris.falip@univ-reims.fr

24 Janvier 2017

CReSTIC, Université Reims Champagne-Ardenne

CONTEXTE

"I've a feeling we're not in low dimensional anymore." -
Wizard of Oz

Les données médicales sont...

- Abondantes
- Sous-exploitées
- Non structurées

Des difficultés inhérentes

- Haute dimensionnalité
- Données non-renseignées
- Informations hétérogènes

Solution attendue

- Visualisation
- Exploration
- Recommandation

Reposant sur la comparaison et structuration automatique

Solution attendue

- Visualisation
- Exploration
- Recommandation

Reposant sur la comparaison et structuration automatique

Objectifs multiples

- Faciliter les stratégies médicales personnalisées
- Capitaliser les connaissances d'experts médicaux
- Assister les études épidémiologiques

Spécificités

- Éviter la généralisation
- Ne pas noyer les situations atypiques
- Extraire des individus représentatifs

MÉTHODOLOGIE

"I'm going to give him a vote he can't refuse." - The Godfather

Émergence d'individus typiques via un mécanisme d'élection.

Étapes

- Processus de vote dans chaque dimension
- Agrégation des résultats en un indice de *représentativité*
- Détermination des individus emblématiques

Soit n individus définis en dimension p .

Sur chacune des p dimensions

- Calcul de la matrice de dissimilarité

Soit n individus définis en dimension p .

Sur chacune des p dimensions

- Calcul de la matrice de dissimilarité
- Transformation en matrice de rangs

Soit n individus définis en dimension p .

Sur chacune des p dimensions

- Calcul de la matrice de dissimilarité
- Transformation en matrice de rangs
- Transformation des rangs en scores de rang

Soit n individus définis en dimension p .

Sur chacune des p dimensions

- Calcul de la matrice de dissimilarité
- Transformation en matrice de rangs
- Transformation des rangs en scores de rang

Globalement

- Agrégation et normalisation, pour chaque individu, des scores obtenus en un indice de *représentativité*.

Soit n individus définis en dimension p .

Sur chacune des p dimensions

- Calcul de la matrice de dissimilarité
- Transformation en matrice de rangs
- Transformation des rangs en scores de rang

Globalement

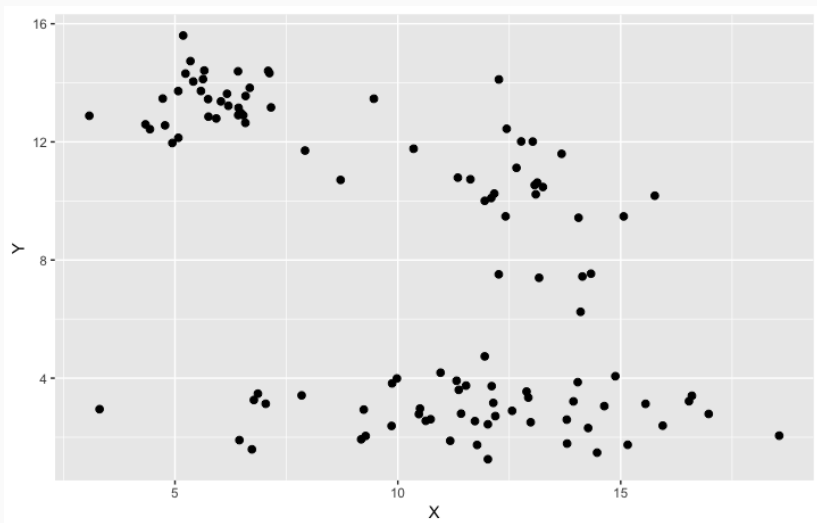
- Agrégation et normalisation, pour chaque individu, des scores obtenus en un indice de *représentativité*.
- Chaque individu est ensuite connecté, parmi ses k plus proches voisins, à celui dont la *représentativité* est la plus élevée. k constitue un paramètre de granularité.

- Représentativité
 - L'individu est-il un bon représentant ?
- Singularité
 - L'individu est-il un cas atypique ?
- Généricité
 - L'individu illustre-t-il une grande partie de la population ?
- Structuration
 - Comment les individus sont-ils associés ?

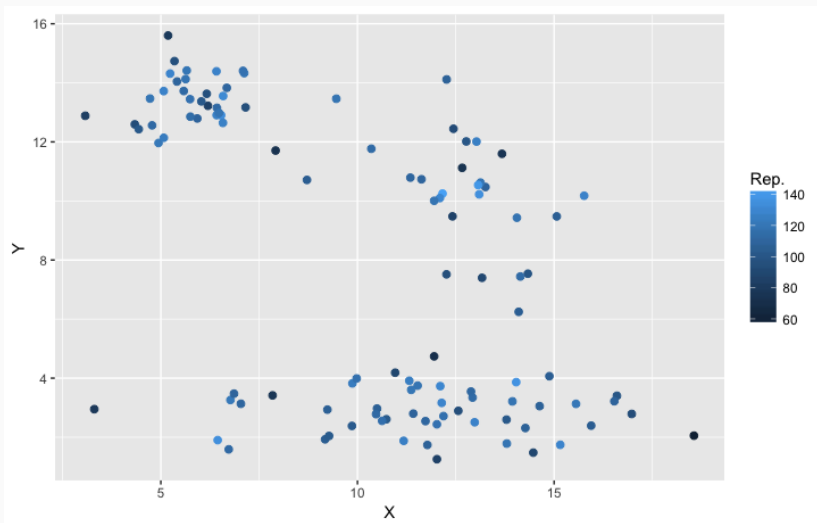
ILLUSTRATIONS

"We're gonna need a bigger dataset." - Jaws

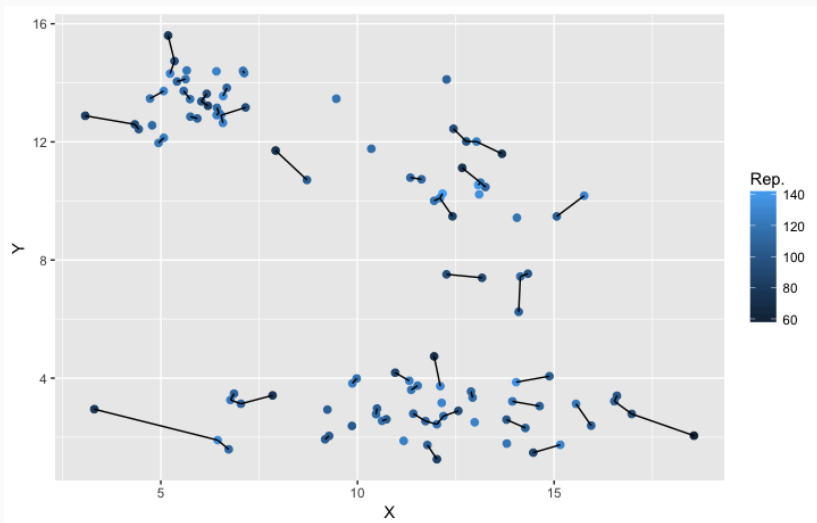
DATASET SYNTHÉTIQUE



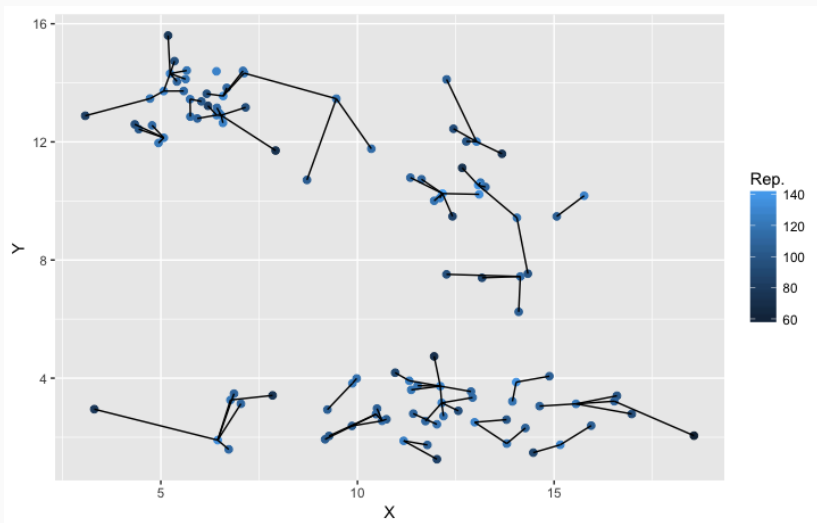
DATASET SYNTHÉTIQUE



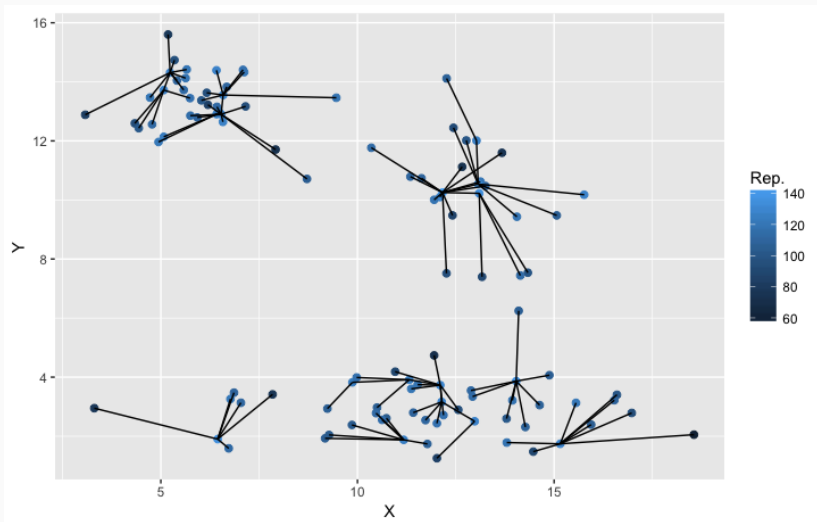
DATASET SYNTHÉTIQUE | $K = 2$



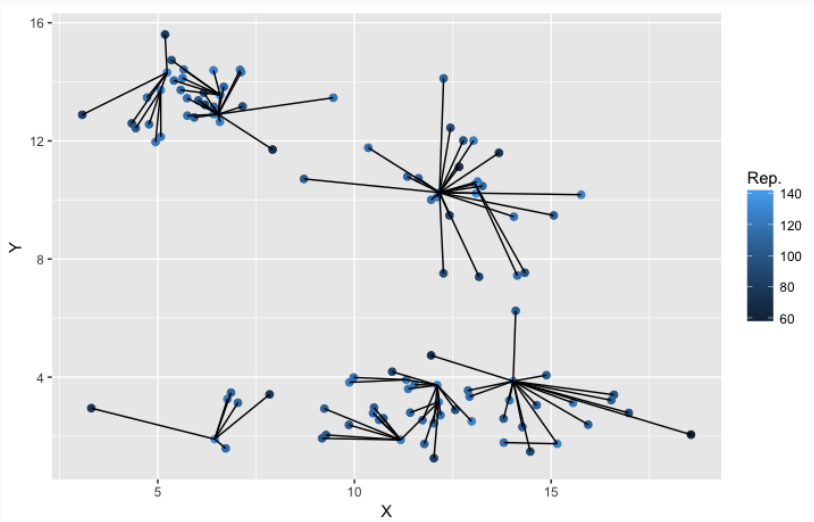
DATASET SYNTHÉTIQUE | $K = 5$



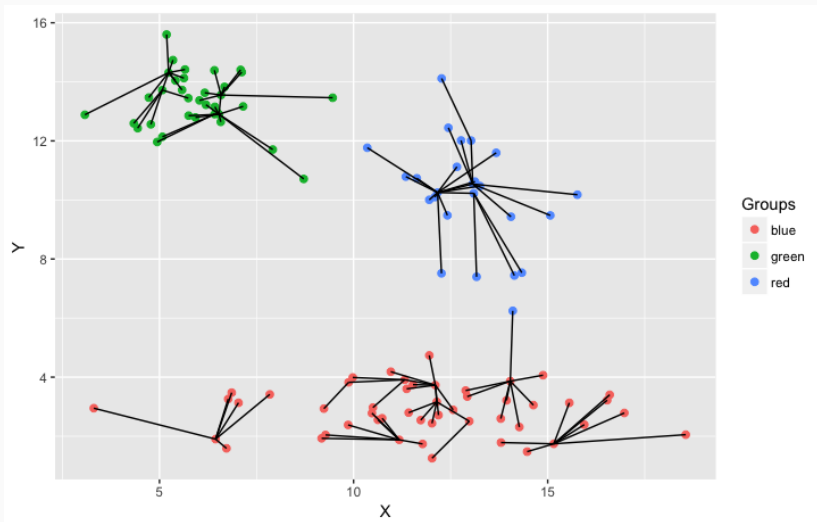
DATASET SYNTHÉTIQUE | $K = 10$



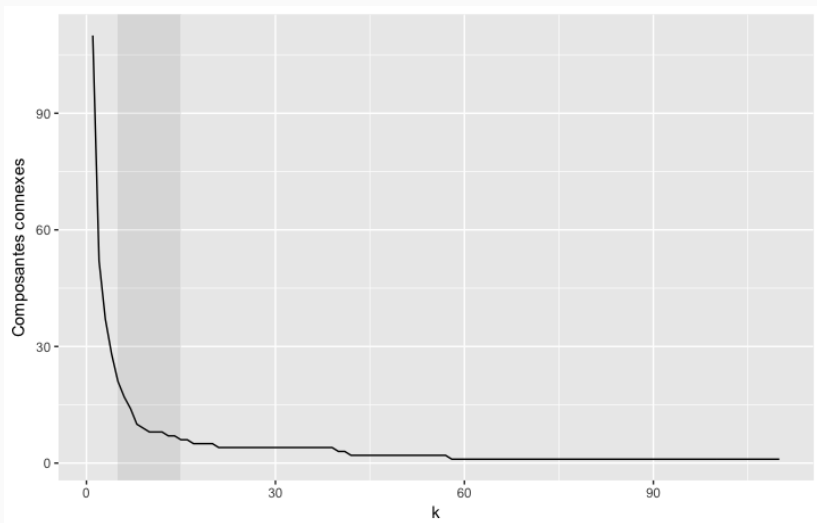
DATASET SYNTHÉTIQUE | $K = 15$



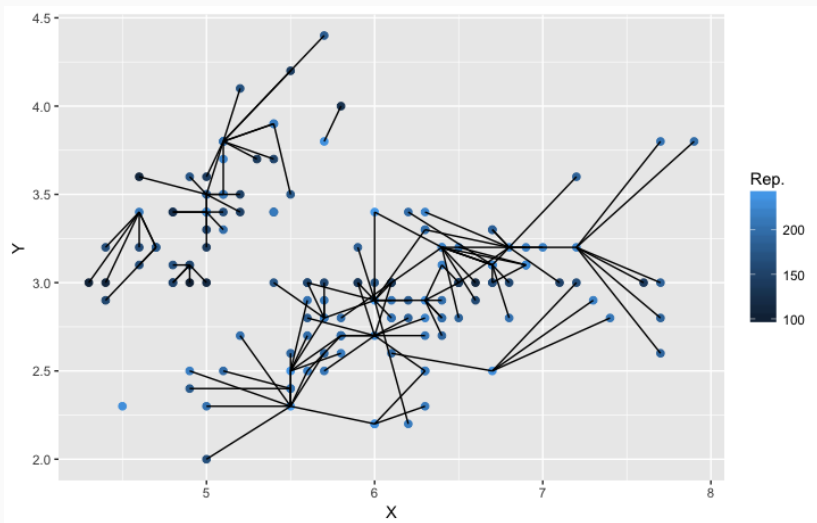
DATASET SYNTHÉTIQUE | K = 10



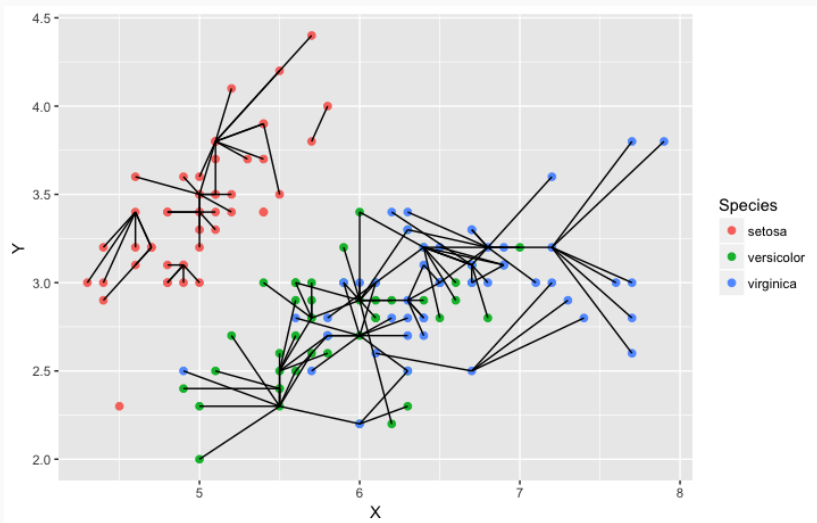
EVOLUTION DES COMPOSANTES CONNEXES



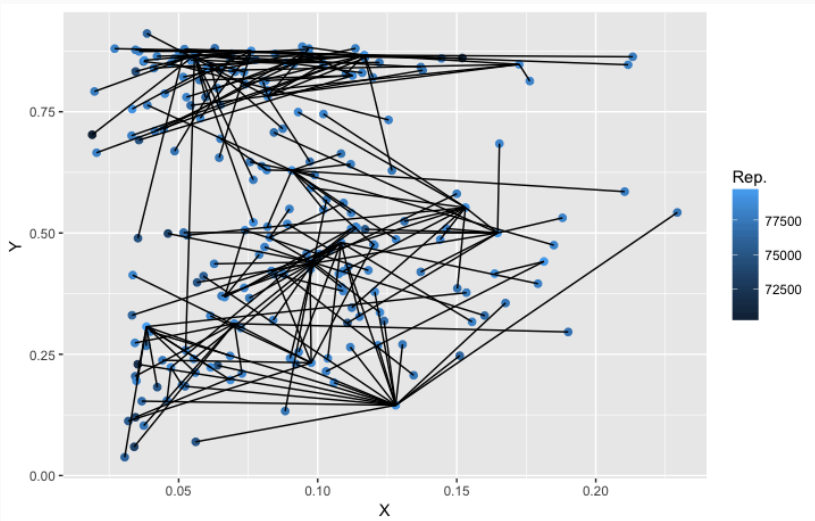
DATASET IRIS | $K = 10$



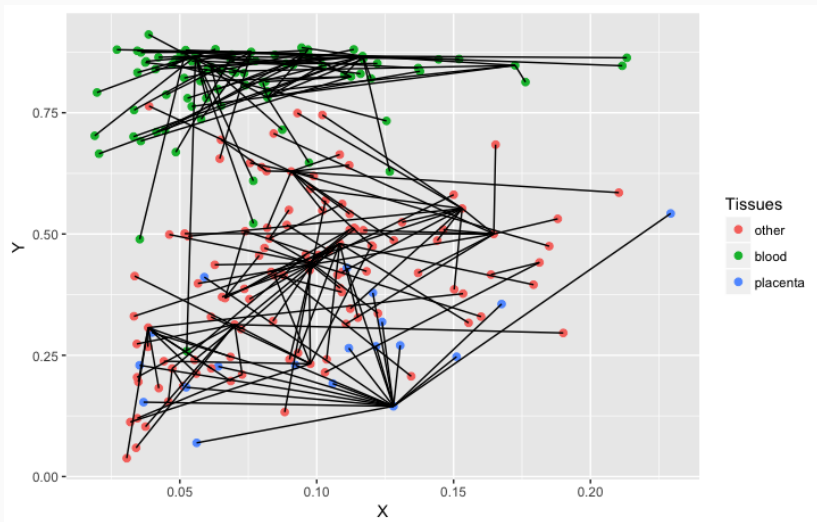
DATASET IRIS | $K = 10$

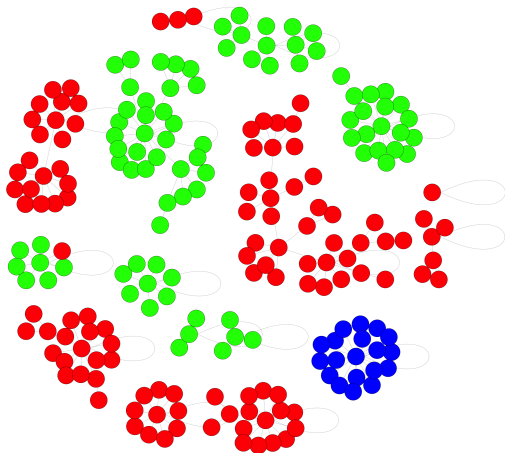


DATASET B. C CHRISTENSEN | $K = 10$



DATASET B. C CHRISTENSEN | $K = 10$





CONCLUSION

"Chewie, we're done." - Star Wars

Résultats

- Méthode d'exploration et visualisation
- Adéquation aux données
- Scalable facilement

Résultats

- Méthode d'exploration et visualisation
- Adéquation aux données
- Scalable facilement

Axes d'améliorations

- Gestion des données manquantes
- Injection d'un individu au dataset

Une palette d'outils large et forte de sens, s'inspirant de la *théorie du choix social*.

- Distribution de "votes négatifs"
- Pondération des dimensions
- Utilisation de systèmes de véto

AUGMENTATION DE DONNÉES POUR L'EXPLORATION DE DOSSIERS MÉDICAUX

Joris Falip, Amine Aït Younes, Frédéric Blanchard, Michel Herbin
joris.falip@univ-reims.fr

24 Janvier 2017

CRéSTIC, Université Reims Champagne-Ardenne