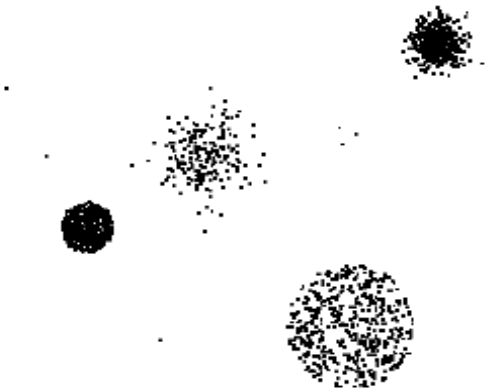
Thème 5: Méthodes avancées – partie 1

Détection des anomalies

S. Wang
– traduit du livre et des diaspos de Tan, Steinbach, Karpatne et Kumar

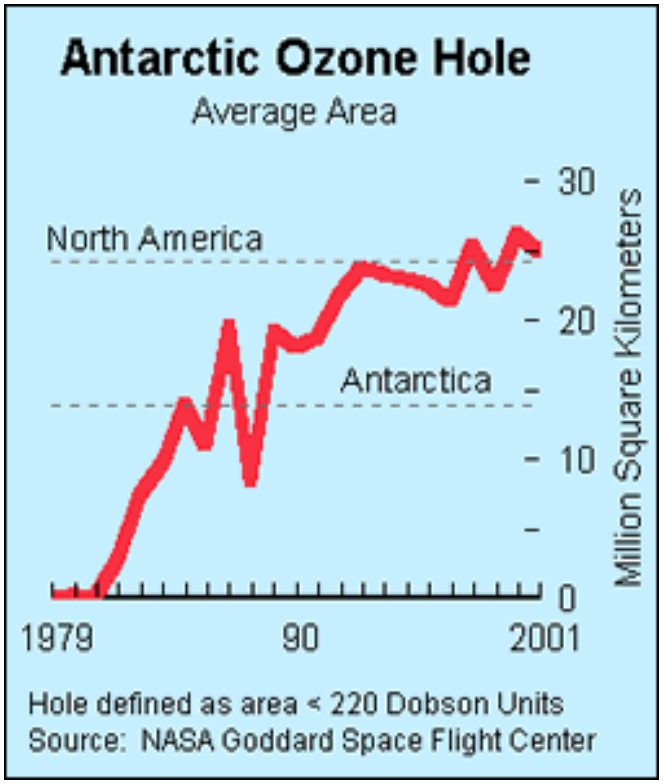
Détection d'anomalies/valeurs aberrantes (outliers)

- Que sont les anomalies/valeurs aberrantes ?
  - L'ensemble (le sous-ensemble) de points de données qui sont considérablement différents du reste des données
- Les anomalies sont relativement rares
  - Par exemple, un sur mille : se produit souvent en présence d'un grand volume de données
  - Le contexte est important, par exemple, les températures glaciales en juillet
- Une information très importante, ex.:
  - Une intrusion dans un syst. d'informatique
  - Tension artérielle anormalement élevée



Importance of Anomaly Detection

- Historique de l'appauvrissement de la couche d'ozone
  - En 1985, trois chercheurs (Farman, Gardinar et Shanklin) ont été intrigués par les données recueillies par le British Antarctic Survey montrant que les niveaux d'ozone pour l'Antarctique avaient chuté de 10 % en dessous des niveaux normaux.
  - Pourquoi le satellite Nimbus 7, qui avait à bord des instruments pour enregistrer les niveaux d'ozone, n'a-t-il pas enregistré des concentrations d'ozone aussi faibles ?
  - Les concentrations d'ozone enregistrées par le satellite étaient si faibles qu'elles ont été traitées comme des valeurs aberrantes par le programme informatique et rejetées !



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Causes des anomalies

- Données en provenance de différentes classes
  - Par ex. on veut mesurer le poids des oranges, mais y trouve quelques pamplemousses dans le panier
- Variation naturelle
  - Des personnes anormalement grandes
- Erreurs de données
  - 200 livres pour un enfant de 2 ans

Distinction entre le bruit et les anomalies

- Les bruits et les anomalies sont des concepts liés mais distincts
  - Les bruits sont des « erreurs », peut-être aléatoires, des valeurs ou des objets contaminants
  - Les bruits sont causés par le procédé/les instruments d'acquisition
- Les bruits ne produisent pas nécessairement des valeurs ou des objets inhabituels
- Les bruits ne sont pas intéressants, alors que les anomalies peuvent être intéressantes si elles ne sont pas dues au bruit

Problèmes généraux : nombre d'attributs

- Anomalies sont souvent définies en termes des attributs individuels
  - Hauteur
  - Forme
  - Couleur
- Peut être difficile de trouver une anomalie en utilisant tous les attributs
  - Attributs bruyants ou non pertinents
  - L'objet n'est anormal que par rapport à certains attributs
- Cependant, un objet anormal peut ne pas être anormal dans aucun attribut. Ex. une personne de 1m et 60kg

Problèmes généraux : cotation (scoring) des anomalies

- Beaucoup de techniques de détection d'anomalies ne fournissent qu'une catégorisation binaire
  - Un objet est une anomalie ou ne l'est pas
  - Surtout le cas pour des approches basées sur la classification
- D'autres approches attribuent un score à tous les points
  - Ce score mesure le degré auquel un objet est une anomalie
  - Cela permet de classer les objets
- Au final, il faut souvent une décision binaire
  - Cette transaction par carte de crédit doit-elle être signalée ?
  - Toutefois, il est toujours utile pour avoir un score
- Combien y a-t-il d'anomalies ?

Autres problèmes liés à la détection d'anomalies

- Défis liés à la détection : toutes les anomalies à la fois ou une à la fois
  - Inondation : phénomène consistant à étiqueter les objets normaux comme des anomalies.
  - Masquage : les anomalies ne sont pas détectées
  - Situations supervisées ou non supervisées
- Évaluation
  - Comment mesurer les performances ?
- Efficacité
- Contexte

Variantes des problèmes de détection d'anomalies

- Étant donné un ensemble de données D, trouver tous les points de données $\mathbf{x} \in D$ avec des scores d'anomalie supérieurs à un certain seuil *t*
- Étant donné un ensemble de données D, trouvez tous les points de données $\mathbf{x} \in D$ ayant les n plus grands (*top-n*) scores d'anomalie
- Étant donné un ensemble de données D contenant principalement des points de données normaux (mais non étiquetés), et un point de test **x**, calculez le score d'anomalie de **x** par rapport à D

Détection d'anomalies basée sur un modèle

- Construire un modèle pour les données et tester
  - Non-supervisé
    - Les anomalies sont des points qui ne correspondent (*fitre*) pas bien au modèle
    - Les anomalies sont des points qui déforment le modèle
    - Exemples des modèles:
      - Distribution statistique
      - Clustering
      - Régression
      - Géométrique
      - Graphique
  - Supervisé
    - Les anomalies sont considérées comme une classe rare
    - Besoin d'avoir des données d'entraînement

Autres techniques de détection d'anomalies

- Basé sur la proximité
  - Les anomalies sont des points éloignés des autres points
  - Peut détecter cela graphiquement dans certains cas
- Basé sur la densité
  - Les points de faible densité sont des valeurs aberrantes
- Correspondance (appariement) de motifs
  - Créer des profils ou des modèles d'événements ou d'objets atypiques mais importants
  - Les algorithmes pour détecter ces modèles sont généralement simples et efficaces
- Basé sur le désordre

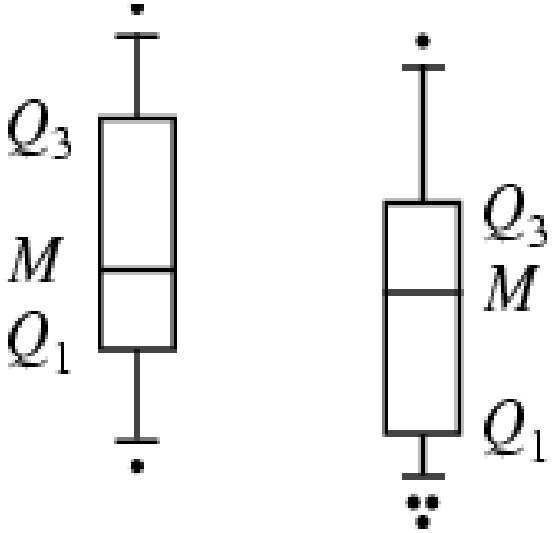
Voici quelques approches/méthodes

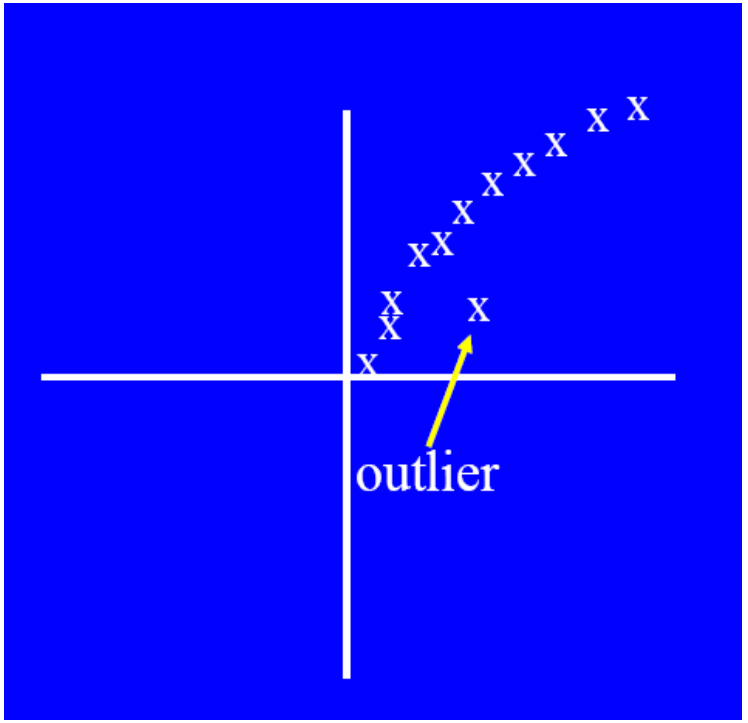
- Je vais introduire chacune des approches suivantes. Cependant, certaines approches sont moins « abordables » à cause des bases mathématiques nécessaires.
- Les diapositives suivantes présentant des méthodes moins « abordables » et ne faisant pas partie de la matière de l'examen final seront marquées des ***.

Approches visuelles

- "Boxplots" ou "scatter plots"

- Limitations
  - Pas automatique
  - Subjectif

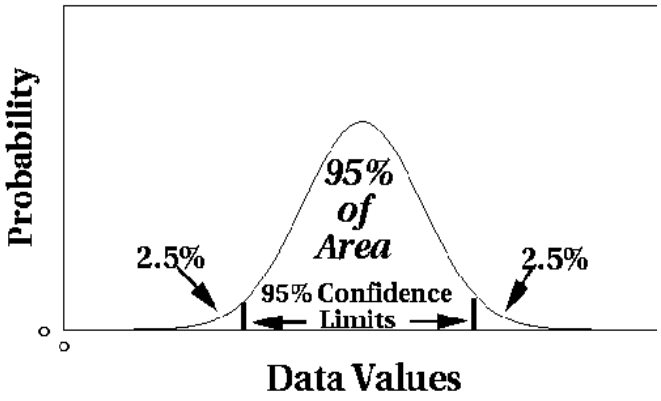




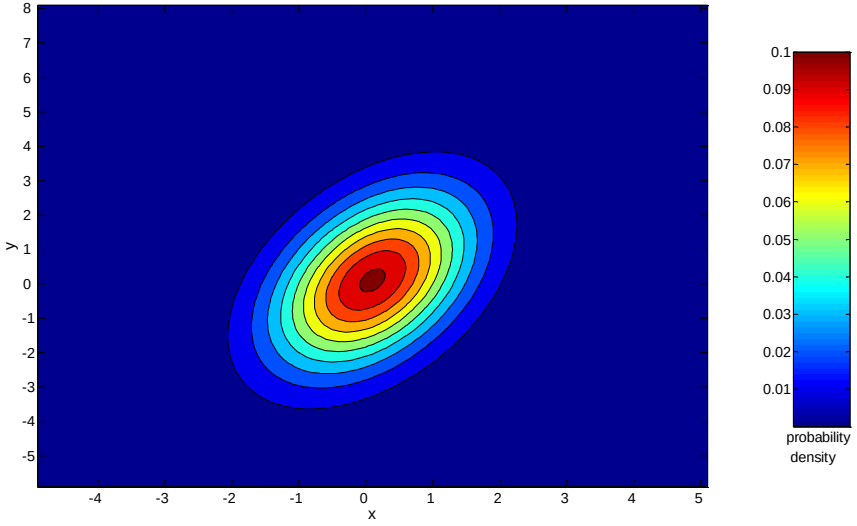
Approches statistiques

- Définition probabiliste d'un objet aberrant :
  - Un objet aberrant est un objet qui a une faible probabilité par rapport à un modèle de distribution de probabilité des données.
  - Généralement, on suppose un modèle paramétrique décrivant la distribution des données (par exemple, une distribution normale)
  - Appliquer un test statistique qui dépend de
    - Distribution des données
    - Paramètres de distribution (par exemple, moyenne, variance)
    - Nombre de valeurs aberrantes attendues (limite de confiance)
  - Problèmes
    - Identifier la distribution d'un ensemble de données
      - Distribution à queue lourde (Heavy tailed distribution)
    - Nombre d'attributs
    - Les données sont-elles un mélange de distributions?

Distributions normales



Gaussienne unidimensionnelle



Gaussienne bidimensionnelle

*** Test de Grubbs

- Détecter les objets aberrants dans les données univariées
  - Supposons que les données proviennent d'une distribution normale
  - Détecter un objet aberrant à la fois, supprime l'objet aberrant et répète
    - H_0 : il n'y a pas d'objet aberrant dans les données
    - H_A : Il y a au moins un objet aberrant
  - Statistique du test de Grubbs : $$G = \frac{\max\left|X - \overline{X}\right|}{s}$$
  - Rejeter H0 si : $$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

Approches statistiques – basée sur la vraisemblance

- Supposons que l'ensemble de données D contienne des échantillons d'un mélange de deux distributions de probabilité :
  - M (distribution majoritaire)
  - A (distribution anormale)
- Distribution de données D = (1 – λ) M + λ A
- M est une distribution de probabilité estimée à partir des données
  - Peut être basé sur n'importe quelle méthode de modélisation (Bayes naïf, entropie maximale, etc.)
- A est (initialement) supposée d'être une distribution uniforme

Approches statistiques – basée sur la vraisemblance

- Vraissemblance à l'instant t :

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

- L'idée de l'algorithme :
  - Initialement, supposons que tous les points appartiennent à M
  - Soit $LL_t(D)$ le log de vraisemblance de D au temps t
  - Pour chaque point x_t qui appartient à M, déplacez-le vers A
    - Soit $LL_{t+1}(D)$ le nouveau log de vraisemblance.
    - Calculer la différence, $\Delta = LL_{t+1}(D) - LL_t(D)$
    - Si $\Delta >$ c (un seuil), alors x_t est déclaré comme une anomalie et déplacé de façon permanente de M vers A

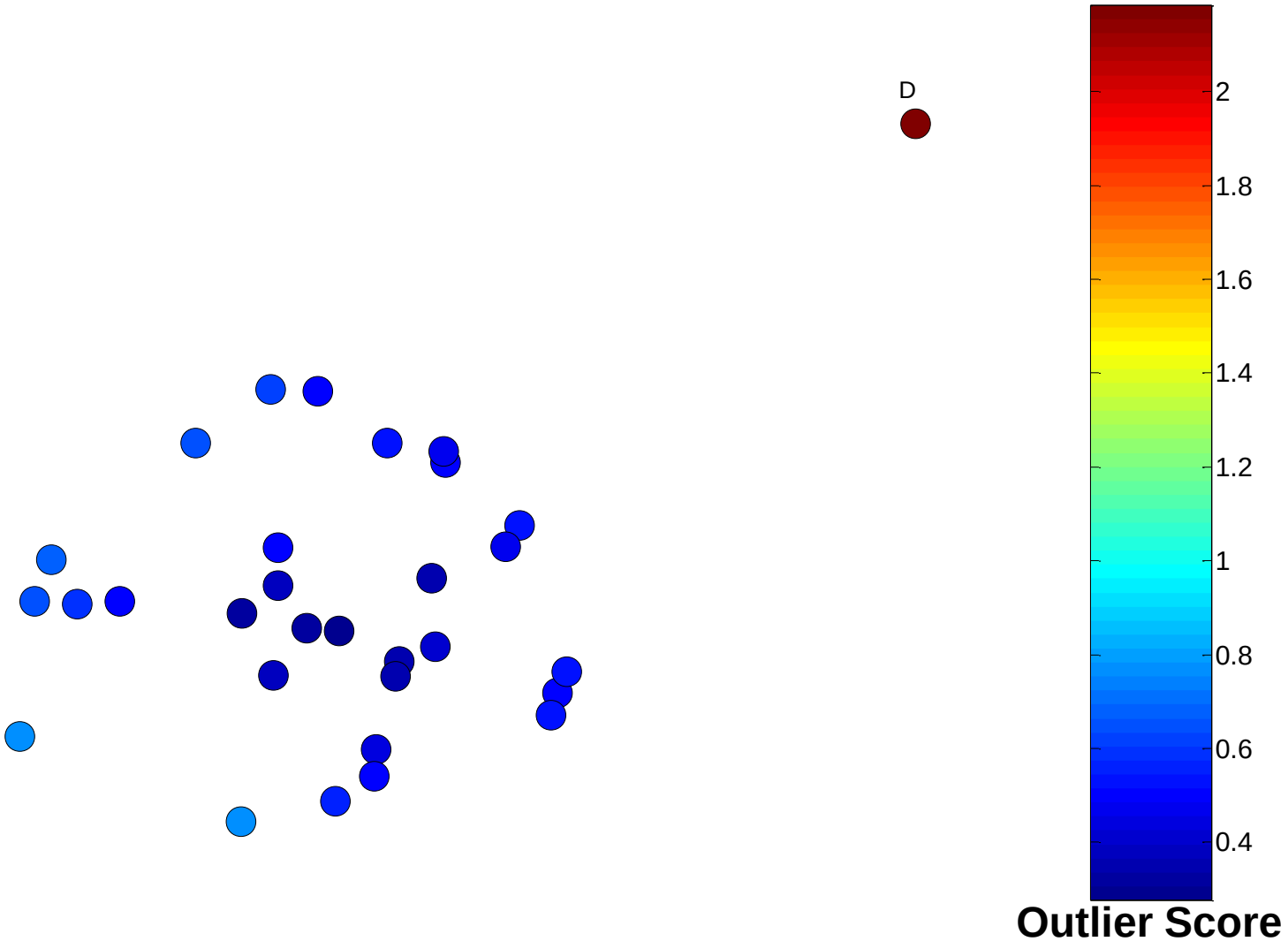
Forces/faiblesses des approches statistiques

- Base mathématique solide
- Peut être très efficace
- Bons résultats si la distribution est connue
- Dans de nombreux cas, la distribution des données peut ne pas être connue
- Pour les données de grande dimension, il peut être difficile d'estimer la vraie distribution
- Les anomalies peuvent fausser les paramètres de la distribution

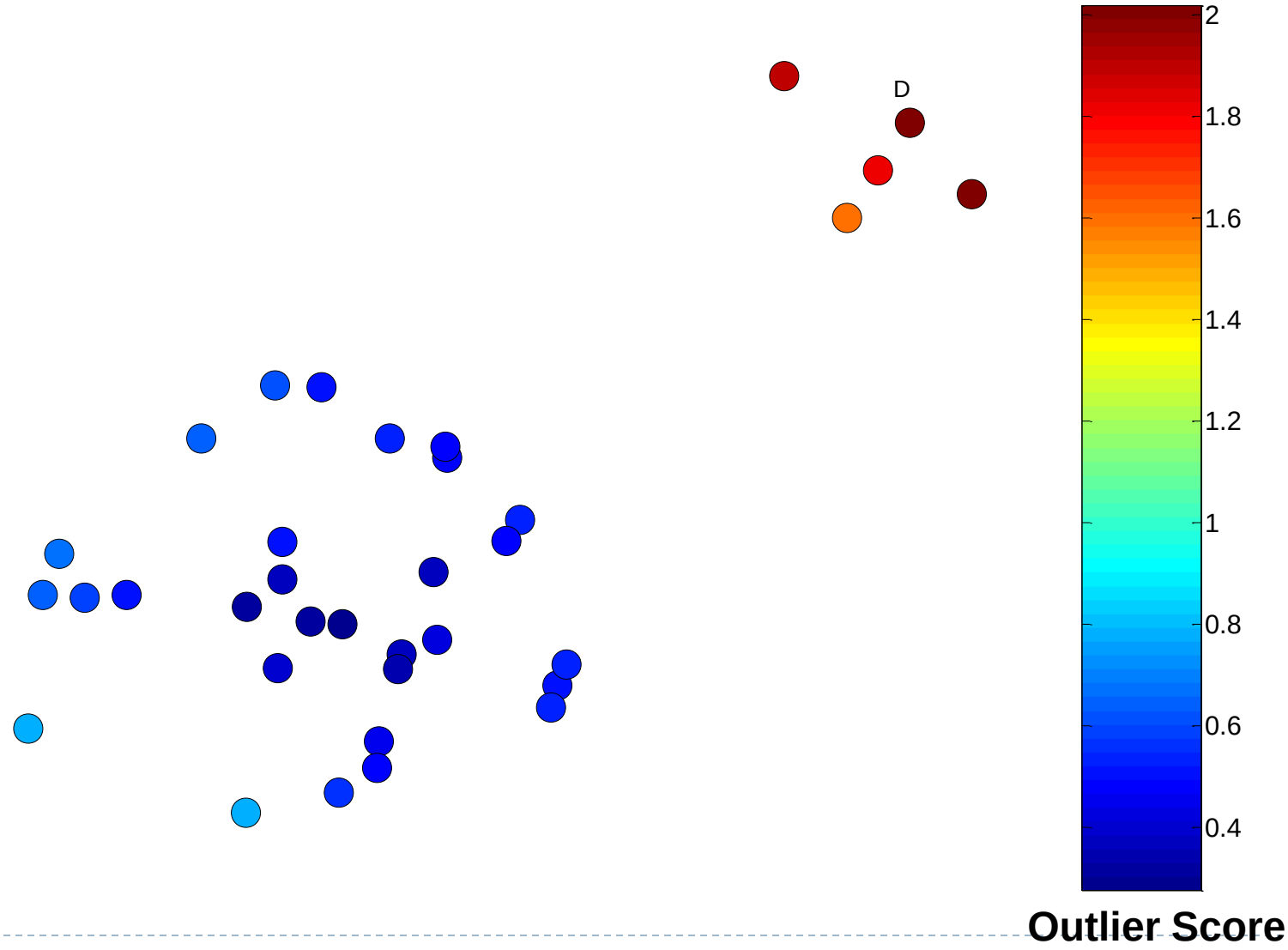
Approches basées sur la distance

- Plusieurs techniques différentes
- Un objet est aberrant si une fraction spécifiée des objets est à plus d'une distance spécifiée (Knorr, Ng 1998)
  - Certaines définitions statistiques sont des cas particuliers de celle-ci
- Le score aberrant (*outlier score*) d'un objet peut être défini comme la distance à son kième voisin le plus proche.

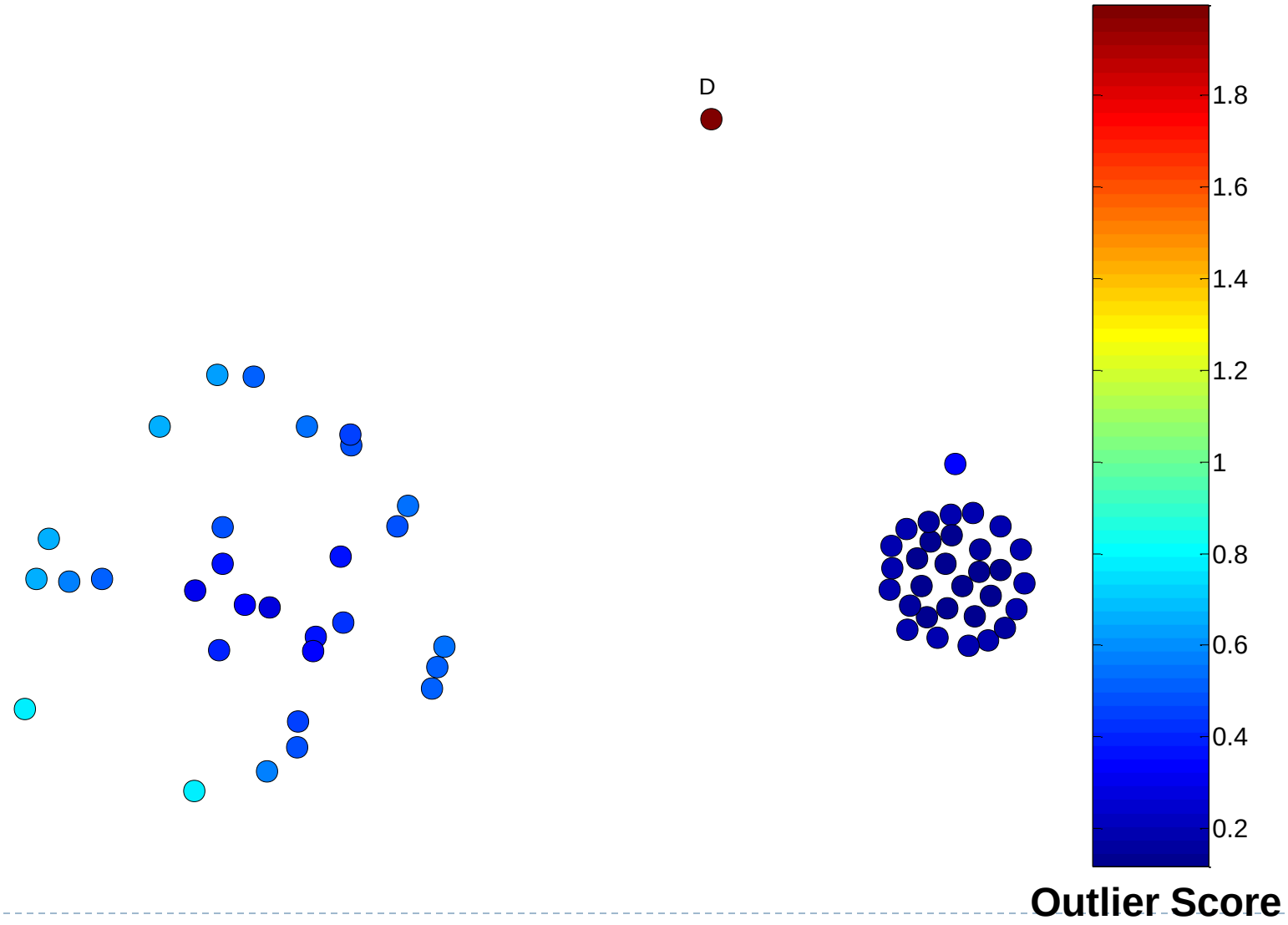
Le voisin le plus proche – Un objet aberrant



Cinq voisins les plus proches - Petit cluster



Cinq voisins les plus proches - Densité différente



Forces/faiblesses des approches basées sur la distance

- Simple
- Exigente en termes of calcul – $O(n^2)$
- Sensible aux paramètres
- Sensible aux variations de densité
- La distance devient moins significative dans l'espace de grande dimension

Approches basées sur la densité

- Objet aberrant basé sur la densité : le score d'un objet aberrant est l'inverse de la densité autour de l'objet.
  - Peut être défini en fonction des k voisins les plus proches
  - Par exemple : inverse de la distance au $k^{ième}$ voisin
  - Ou encore : Inverse de la distance moyenne aux k voisins
  - Ou selon définition dans DBSCAN

- S'il existe des régions de densité différente, cette approche peut poser des problèmes

Densité relative

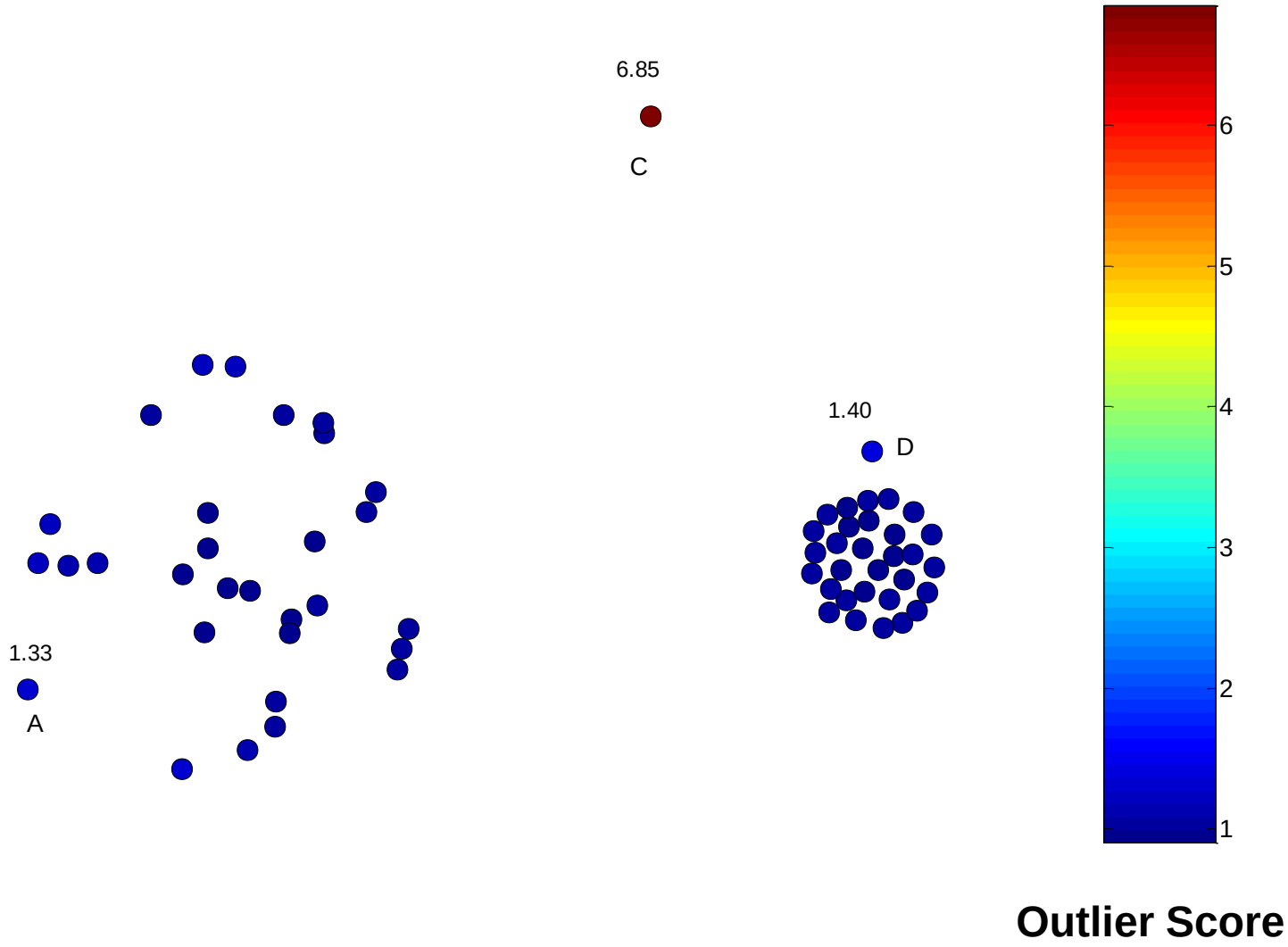
- Considérons la densité d'un point par rapport à celle de ses k plus proches voisins

$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(\mathbf{y}, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

Algorithm 10.2 Relative density outlier score algorithm.

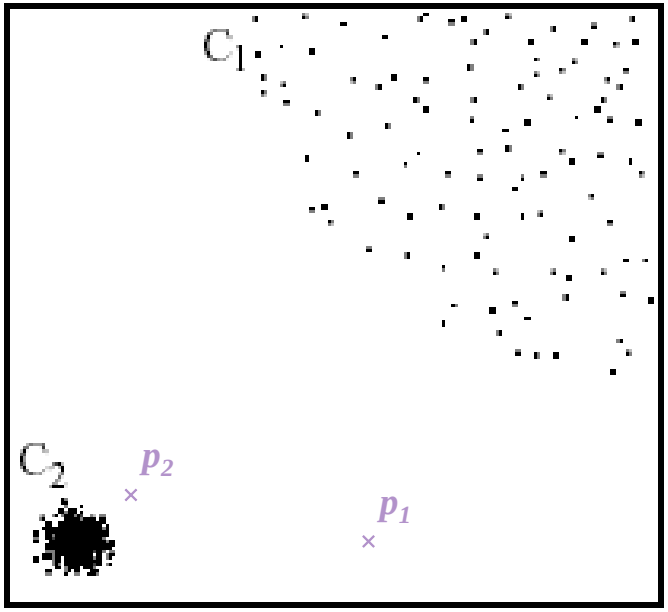
1: {k is the number of nearest neighbors}
2: **for all** objects $\mathbf{x}$ **do**
3: Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
4: Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
5: **end for**
6: **for all** objects $\mathbf{x}$ **do**
7: Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

Scores des valeurs aberrantes de la densité relative



Basée sur la densité : approche LOF (local outlier factor)

- Pour chaque point, calculez la densité de son voisinage local
- Calculer le facteur local aberrant (LOF) d'un échantillon p comme la moyenne des rapports de la densité de l'échantillon p et de la densité de ses voisins les plus proches
- Les valeurs aberrantes sont des points avec la plus grande valeur LOF



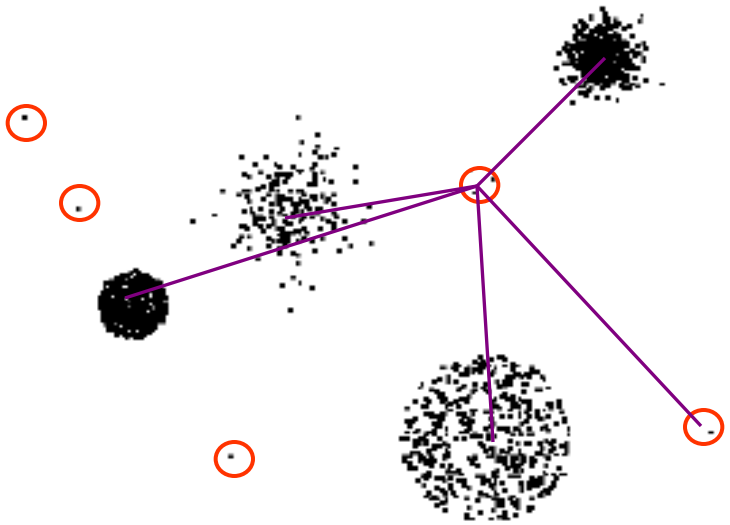
Dans l'approche la plus proche voisin, p2 n'est pas considéré comme un objet aberrant, mais l'approche LOF trouve à la fois p1 et p2 comme objets aberrants

Forces/faiblesses des approches basées sur la densité

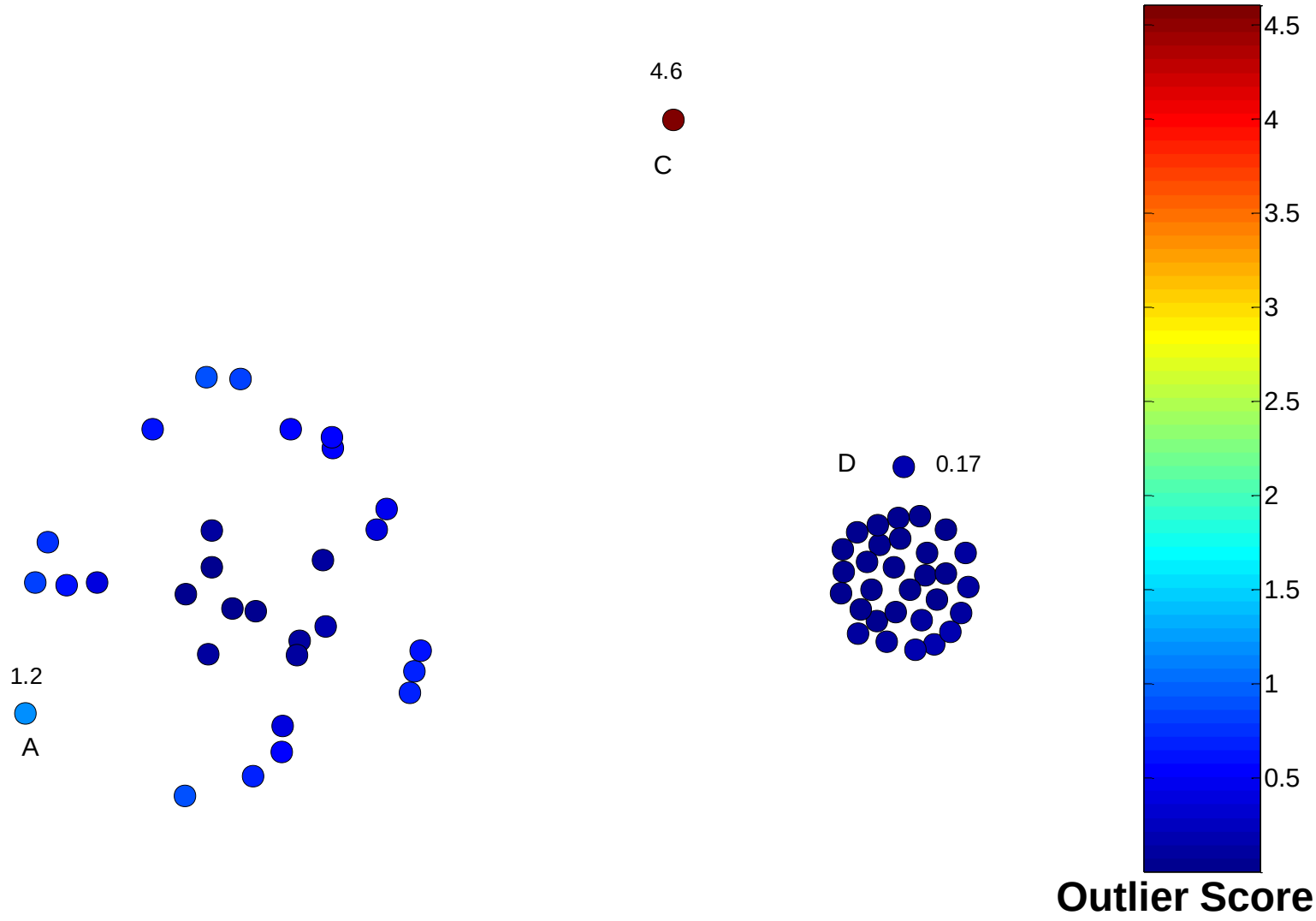
- Simple
- Exigente en termes of calcul – $O(n^2)$
- Sensible aux paramètres
- La densité devient moins significative dans l'espace de grande dimension

Approches basées sur le clustering

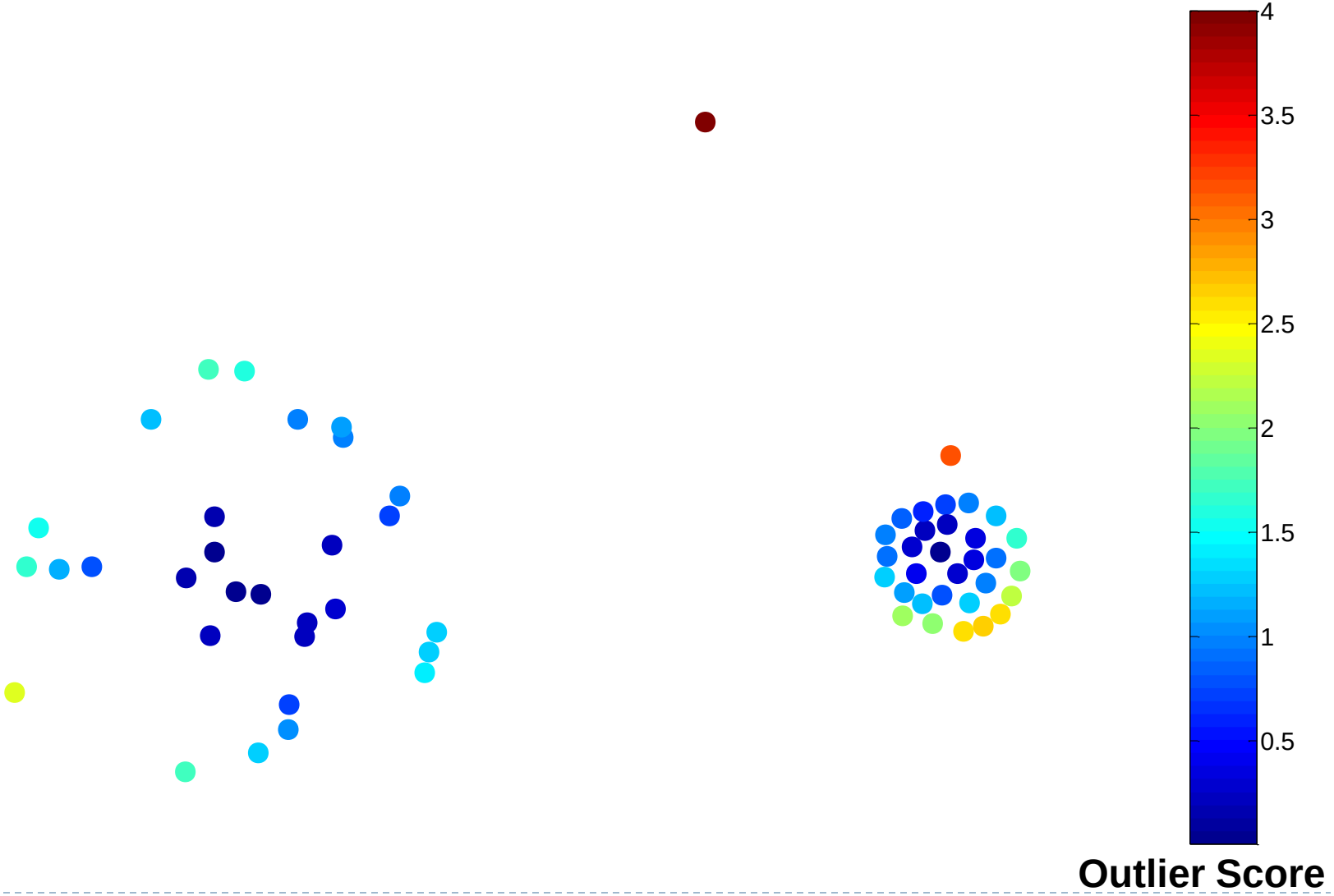
- **Objet aberrant basé sur le cluster** : un objet est une valeur aberrante basée sur un cluster s'il n'appartient pas fortement à un cluster
  - Pour les clusters basés sur des prototypes, un objet est une valeur aberrante s'il n'est pas assez proche d'un centre de cluster
  - Pour les clusters basés sur la densité, un objet est une valeur aberrante si sa densité est trop faible
  - Pour les clusters basés sur des graphes, un objet est une valeur aberrante s'il n'est pas bien connecté
- D'autres problèmes incluent l'impact des objets aberrants sur les clusters et le nombre de clusters



Distance des points aux centroïdes les plus proches



Distance relative des points par rapport au centre de gravité le plus proche



Forces/faiblesses des approches basées sur le clustering

- Simple
- De nombreuses techniques de clustering peuvent être utilisées
- Peut être difficile de décider d'une technique de clustering
- Peut être difficile de décider du nombre de clusters
- Les valeurs aberrantes peuvent fausser les clusters

***Autres approches pour détecter des objets aberrants

- Basée sur la reconstruction : ex. par PCA ou par apprentressage profond
  - Classe normale réside dans un espace de dimension inférieure
  - Trouver cet espace et projeter chaque objet dans l'espace
  - Reconstruire l'objet projeté => objet aberrant genère plus grande erreur de reconstruction
- Basée sur le classement à une classe :
  - SVM (Support Vector Machine) peut être utilisé pour générer la frontière de la classe normale
  - Objets aberrants se trouvent dans l'autre côté de la frontière

*** Autres approches pour détecter des objets aberrants

- Basée sur la théorie d'information (théorie de désordre) : mesurer la quantité d'informations dans l'ensemble des données
  - La fonction d'entropie dans le cas des données catégoriques
  - La complexité Kolmogorov (qui mesure la complexité d'un ensemble de données par la taille du plus petit programme informatique permettant de reproduire les données originales) dans d'autres cas
  - Ou bien, en pratique, utiliser une technique standard pour compresser les données et utiliser la taille du fichier compressé comme la quantité d'information dans l'ensemble des données
  - **L'objet x le plus susceptible d'être aberrant est celui qui maximise le gain d'information définie comme suit** :

$$Gain(x) = Info(D) - Info(D\backslash \mathrm{x})$$