1. Introduction

MAT8594

UQAM

1.1 Sciences des données

Définition

- On attribue à J. Wills la définition suivante :
  Data scientist (n) : Person who is better at statistics than any software engineer and better at software than any statistician.
- Il s'agit d'un terme qui regroupe des résultats théoriques et des compétences issus des mathématiques, des statistiques, de l'informatique, etc.

Bref historique

- **Des origines au début des années 70 :** statistique inférentielle, calculs ≪ à la main ≫, $n = \pm 50$ observations et $p = \pm 10$ variables, modèle de régression linéaire.
- **1970 :** premiers outils informatiques, les balbutiements de l'analyse de données, trouver des alternatives au modèle de régression linéaire et à la loi Normale.
- **1980 :** utilisation des premiers réseaux neurones, modèles non-paramétriques.
- **1990 :** début du *data mining*, développement de l'apprentissage statistique (machine), développement de l'intelligence artificielle.
  $\rightarrow 1^{\text{er}}$ **changement de paradigme :** les données ne sont plus ≪ planifiées ≫.

Bref historique (suite)

- **2000 :** poursuite du développement de l'apprentissage statistique (compromis biais/variance, erreur d'approximation, erreur d'estimation, etc.), sélection des modèles, sélection des variables.
  $\rightarrow$ 2^{e} **changement de paradigme :** explosion de la valeur de p.
- **2010 à aujourd'hui :** grande variété de données (images, sons, nombres, textes, etc.) à analyser, délocalisation des données (nuage), explosion de la puissance informatique.
  $\rightarrow$ 3^{e} **changement de paradigme :** explosion de la valeur de n.

Logiciels et interfaces

- SAS (*Enterprise Miner*)
- SPSS (*Clementine*)
- R (*R Studio*) $\rightarrow$ modélisation et interprétation
- Python $\rightarrow$ modélisation et prédiction

1.2 Apprentissage statistique (ou machine)

Problématique

$$\text{Nombre de sinistre(s)} = \beta_0 + \beta_1(\text{km parcourus}) + \beta_2(\text{âge}) + \epsilon$$
$$\epsilon \sim \text{Normale}(0, \sigma^2).$$

- Un modèle combinant une part d'explication (variables explicatives) et une part d'aléatoire (le terme d'erreur).
- On cherche à estimer les paramètres du modèle $(\beta_0, \beta_1, \ldots)$ à partir d'observations en contrôlant la portion stochastique. En termes informatiques, on parle d'*apprentissage*.

Quel est l'objectif ?

- Développer une nouvelle approche ?
- Gagner un concours (type *Kaggle*) ?
- Utilisation en industrie (compagnie d'assurance) ?
- Utilisation personnelle ?

Apprentissage supervisé et non-supervisé

- Supervisé : on cherche à expliquer une variable Y à l'aide d'un ensemble de variables $\mathbf{X}$:

$$Y = f(\mathbf{X}) + \epsilon.$$

  On cherche alors une fonction $\widehat{f}$ qui permettra de reproduire au mieux (à définir selon un certain critère) les valeurs y observées.
  Exemples : régression, analyse discriminante, classement, etc.
- Non-supervisé : il n'y a pas de variable réponse à expliquer ou prédire. On cherche à regrouper, classifier, diviser, comprendre des observations.
  Exemples : analyses en composantes principales, analyse factorielle, classification, etc.

1.3 Apprentissage non supervisé

Objectifs principaux

- **Réduire les dimensions d'un jeu de données** : compression d'images, analyse de données de télématique, analyse économique et/ou financière, etc.
- **Classifier des données** : trouver des groupes dans une population, trouver des classes de risque dans un portefeuille, regrouper des actifs, etc.
- Estimer la fonction de densité d'une variable aléatoire.

Réduction de dimensionnalité

- On dispose d'un jeu de données dans un espace de dimension élevé $\mathbb{R}^k$ avec k élevés : $\mathbf{X}_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ik} \end{bmatrix}$, $i = 1, \dots, n$.
- On cherche une transformation $g : \mathbf{X}_i \to \mathbf{X}_i^*$ avec $\mathbf{X}^* \in \mathbb{R}^q$ où $q <<< k$ en perdant le moins d'information possible.

Réduction de dimensionnalité

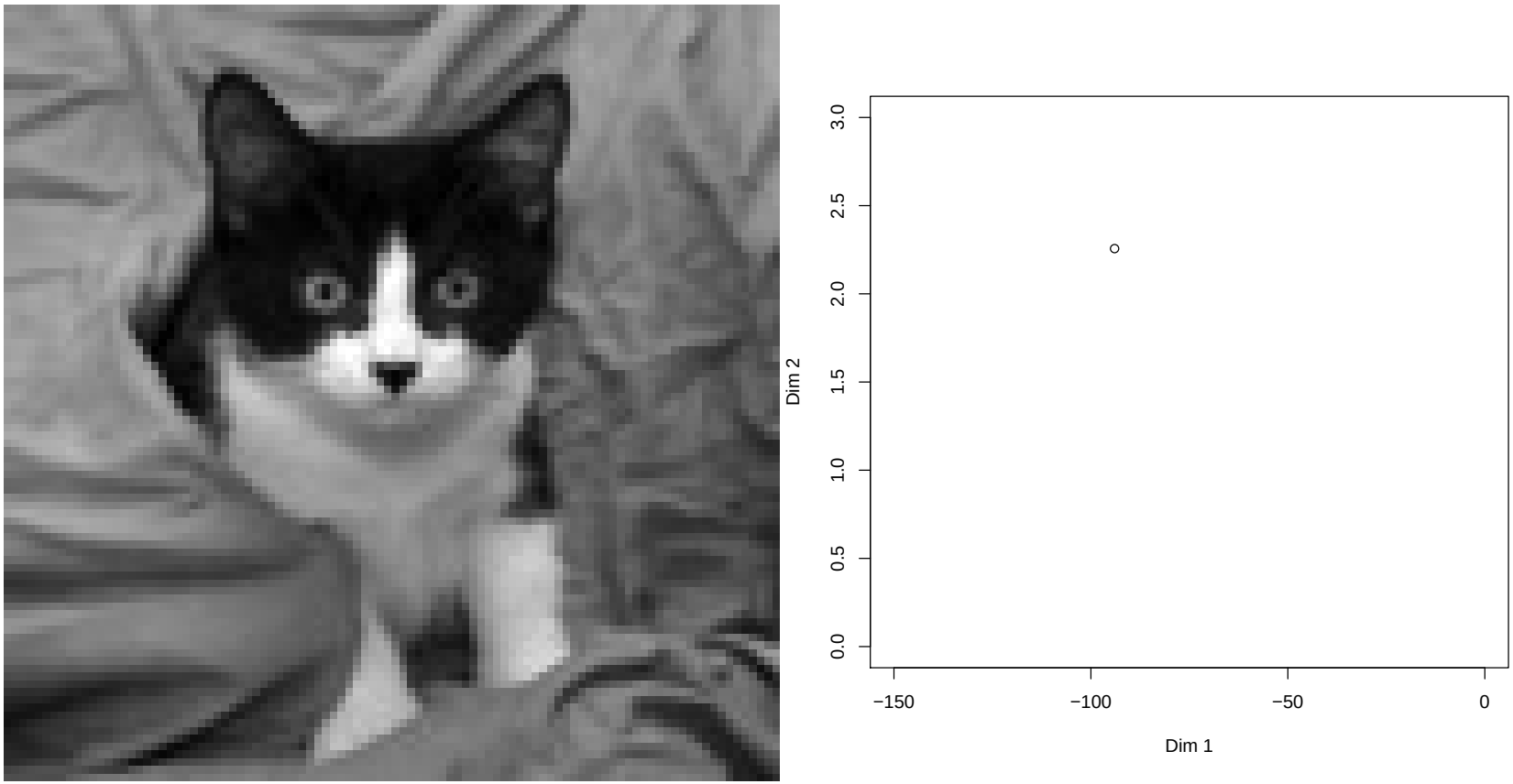

FIGURE: Un chat en dimension $200 \times 200 \to \mathbb{R}^{40\,000}$ (gauche) et en dimension $\mathbb{R}^2$ (droite).

Réduction de dimentionnalité

- **Analyse en composantes principales :** k variables quantitatives
- Analyse factorielle discriminante : k variables quantitatives et 1 variable catégorielle
- Analyse factorielle des correspondances : 2 variables catégorielles (simple) ou k variables catégorielles (multiple)
- etc.

Classification

- On dispose d'un jeu de données dans un espace $\mathbb{R}^k$:
  $\mathbf{X}_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ik} \end{bmatrix}$, $i = 1, \dots, n$.
- On cherche à associer à chaque $\mathbf{X}_i$ une catégorie $y \in \{1, \dots, q\}$ de manière à regrouper les observations similaires (à définir) et à ne pas regrouper les observations différentes (à définir également). Les catégories **ne** sont **pas** connues à l'avance.

Classification

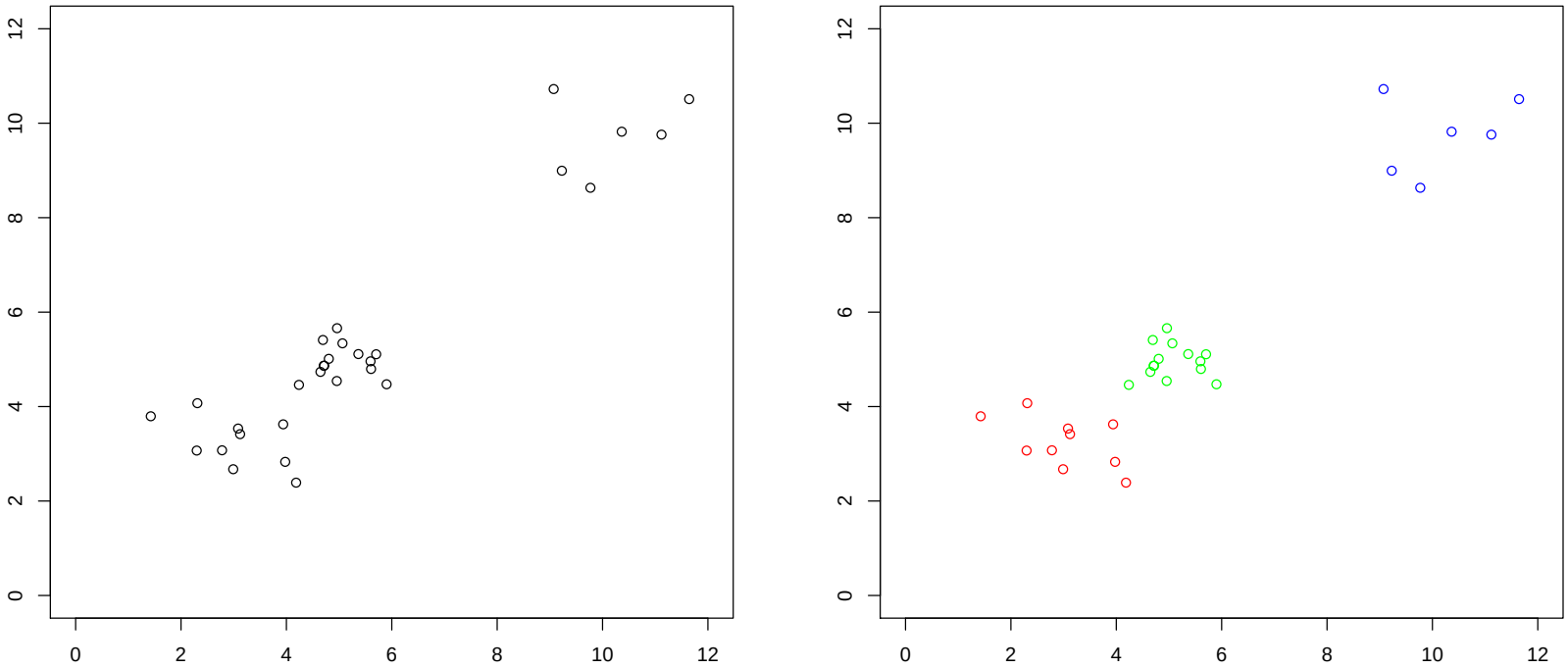

FIGURE: Données brutes (gauche) et classifiées (droite).

Classification

- **Partitionnement** (k-moyennes)
- **Hiérarchique** (divisif, agglomératif, etc.)
- Graphique
- Densité
- Conceptuel
- etc.

Quelques références

- G. James, D. Witten, T. Hastie et R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Springer.
- T. Hastie, R. Tibshirani et J. Friedman. *The Elements of Statistical Learning*, Springer.
- F. Husson, S. Lê et J. Pagès. *Analyse de données avec R*, Presses universitaires de Rennes.