

- Statistique multivariée
- Analyse des données
- Apprentissage statistique à partir des données
  - Supervisé
  - Non supervisé
- Reconnaissance des formes statistique
- Fouille de données ou *Data mining*

- Explosion de la quantité de données disponibles
  - Informatique de gestion
  - Appareils de mesure : capteurs de pollution, images satellitaires, ...
  - Fichiers de logs
  - Le Web
- Objectif : extraire des informations à partir de ces données
- Moyens :
  - Constitution d'entrepôts de données (*datawarehouse*)
  - Outils d'analyse : Data mining

- Exemple de méthodes (modèles )
  - Visualiser et interpréter : ACP, MDS
  - Prévoir : classement, régression
  - Découvrir des structures
  - Rechercher des règles
- Intersection de plusieurs disciplines
  - Base de données
  - Statistique et analyse de données
  - Apprentissage (*Machine learning*)
  - Intelligence artificielle

# Étapes du processus d'extraction d'information

- Nettoyage des données (60 % du processus)
  - Données manquantes, données atypiques (*outliers*), mise au format,...
- Sélection d'un jeu de données pertinent en fonction de l'objectif fixé
  - Sélection des variables, des individus,...
- Data mining
  - Choix des objectifs : résumé, classification, régression
  - Choix des méthodes
  - Application des méthodes
- Analyse des résultats
  - Visualisation
  - Interprétation
  - Retour aux étapes précédentes

# Exemple des Moucherons (1)

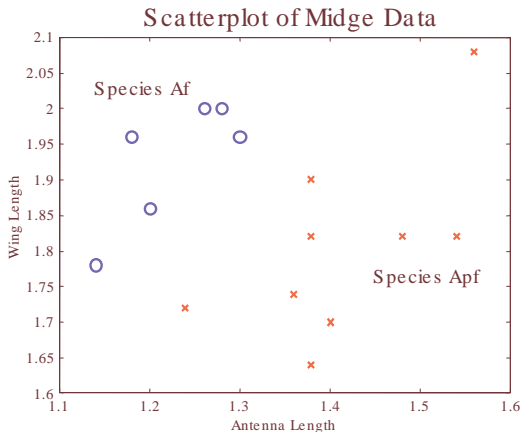
- Découverte de 2 espèces de moucheron (1981)
- Difficile de les distinguer
- On cherche à le faire avec des caractéristiques externes simples
- Données :
  - 9 moucherons Af et 6 moucherons Apf
  - Longueurs de l'aile et de l'antenne en mm

# Exemple des Mouchérons (2)

0	1.38	1.64
0	1.40	1.70
0	1.24	1.72
0	1.36	1.74
0	1.38	1.82
0	1.48	1.82
0	1.54	1.82
0	1.38	1.90
0	1.56	2.08
1	1.14	1.78
1	1.20	1.86
1	1.18	1.96
1	1.30	1.96
1	1.26	2.00
1	1.28	2.00

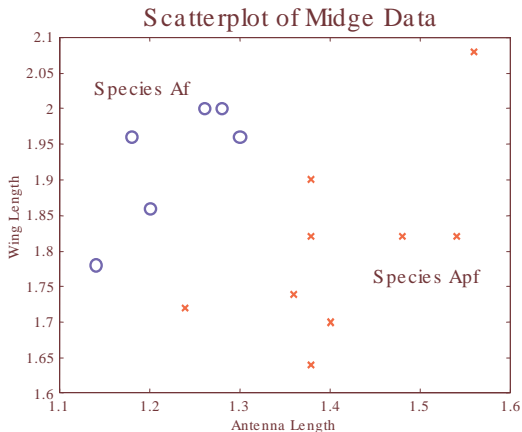
# Exemple des Moucheron (2)

0	1.38	1.64
0	1.40	1.70
0	1.24	1.72
0	1.36	1.74
0	1.38	1.82
0	1.48	1.82
0	1.54	1.82
0	1.38	1.90
0	1.56	2.08
1	1.14	1.78
1	1.20	1.86
1	1.18	1.96
1	1.30	1.96
1	1.26	2.00
1	1.28	2.00



# Exemple des Moucheron (2)

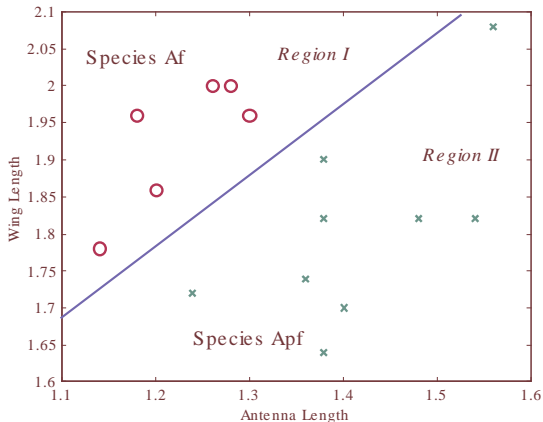
0	1.38	1.64
0	1.40	1.70
0	1.24	1.72
0	1.36	1.74
0	1.38	1.82
0	1.48	1.82
0	1.54	1.82
0	1.38	1.90
0	1.56	2.08
1	1.14	1.78
1	1.20	1.86
1	1.18	1.96
1	1.30	1.96
1	1.26	2.00
1	1.28	2.00



Comment distinguer les 2 groupes ?

# Exemple des Moucheron (3)

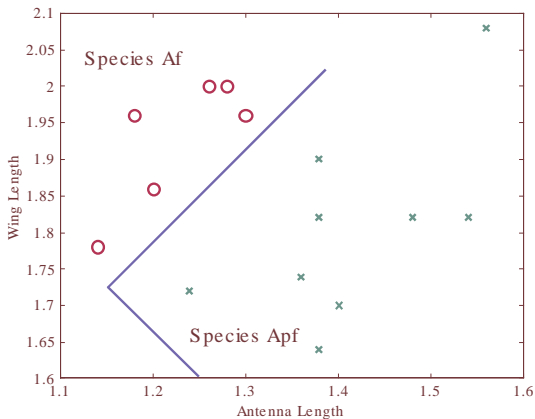
0	1.38	1.64
0	1.40	1.70
0	1.24	1.72
0	1.36	1.74
0	1.38	1.82
0	1.48	1.82
0	1.54	1.82
0	1.38	1.90
0	1.56	2.08
1	1.14	1.78
1	1.20	1.86
1	1.18	1.96
1	1.30	1.96
1	1.26	2.00
1	1.28	2.00



Il est facile visuellement de tracer une ligne



# Exemple des Moucheron (5)



On peut voir cela comment un changement de variables

# Exemple des Mouchérons (6)

- **Variable intéressante** :  $d = \text{aile-antenne}$
- Mais aussi : aile/antenne
- Intérêt du **graphe de dispersion** (scatter plot)
- Mais que faire si on a plus de 2 variables ?
- Problème de **discrimination** et ensemble d'**apprentissage**
- **Validité** des résultats sur la population totale ?
  - Nécessité de la Statistique :
    - Vecteur aléatoire : (aile, antenne)
    - Loi jointe, lois marginales, lois conditionnelles, ...

- Une personne cherche à filtrer ses emails : email et spam
- Données : 3601 emails, classés en email et spam pour lesquels on connaît la fréquence de 57 mots souvent utilisés

	Georg	you	your	hp	free	hpl	!	Our	re	edu	remove
Spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.23
Email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Problème de discrimination

- **Données** : 97 patients
  - Niveau de gravité (log) :  $lcavol$  (difficile à calculer)
  - Poids de la prostate (log) :  $weight$
  - âge
  - Quantité d'hyperplasie prostatique (log) :  $lbph$
  - Invasion de la vésicule séminale :  $svi$
  - Pénétration capsulaire (log) :  $lcp$
  - Score de Gleason
  - Pourcentage de score de Gleason 4 ou 5 :  $pgg45$
- **Prédire  $lcavol$**  (pour décider d'une opération ou non)
- Problème de **régression**

- Introduction
- Méthodes exploratoires élémentaires
- Analyse en Composantes Principales
- Classification Automatique
- Vecteur aléatoire et statistique multidimensionnelle
- Théorie de la décision
- Discrimination dans le cas gaussien
- Régression linéaire
- Méthodes non paramétriques (kppv,...)
- Sélection de variables
- Evaluation des méthodes

- Probabilités, analyse de données et statistique, Saporta, G., Technip, Paris (2006)
- A First Course in Multivariate Statistics, Flury, Springer (1997)
- The elements of statistical learning, Hastie, Tibshirani, Friedman, Springer (2001)
- Pattern recognition, Duda, Hart et Stork, Wiley (2000)