Apprentissage et Reconnaisance de Formes

Nicolas Baskiotis

nicolas.baskiotis@lip6.fr

équipe MLIA, Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie (UPMC)

S2 (2015-2016)

Plan

1. Organisation de l'UE
2. Introduction
3. Les problèmatiques générales
4. Premier exemple : classification de films et arbres de décision

Informations administratives

Crénaux

- Cours : Mercredi 16h-18h
- TD/TME :
  - groupe 1 : Mercredi 8h30-12h45 (Nicolas Baskiotis)
  - groupe 2 : Jeudi 8h30-12h45 (Sylvain Lamprier)

Supports et références

- Site du master : `http://dac.lip6.fr/master`
- Beaucoup de références on-line, beaucoup de ebooks et de livres, cf site
- en cas de questions/problèmes
  → email : `nicolas.baskiotis@lip6.fr` (et Piazza ?)

Évaluation

- CC :
  - travail en TME
  - projet et/ou partiel
- Examen

Pré-requis

- Mathématiques :
  - Probabiltiés et Statistiques
    $\rightarrow$ Cours MAPSI de M1-S1, Cours 3I005 de L3
  - Notions d'algèbre linéaire : matrice, vecteur, norme, produit scalaire, ...
  - Notions d'analyse : dérivée, dérivée partielle, gradient, intégrale, ...

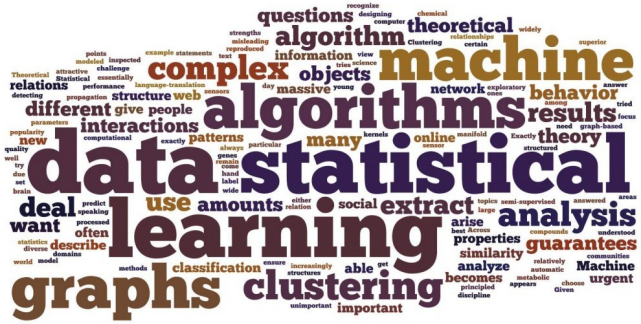
- Du formalisme logique :
  - savoir lire une formule mathématique :
    $$\forall x \in \mathcal{E} \bigcup \mathcal{F}, \exists Y \subset \mathcal{Y}, p(x) \in Y \Rightarrow \neg q(Y)$$

- Programmation (pour les TMEs et projet) : Python
  - savoir faire des scripts (MAPSI)
  - modules : `matplotlib, numpy`

Plan

1. Organisation de l'UE
2. **Introduction**
3. Les problèmatiques générales
4. Premier exemple : classification de films et arbres de décision

Apprentissage artificielle ? (Machine Learning)

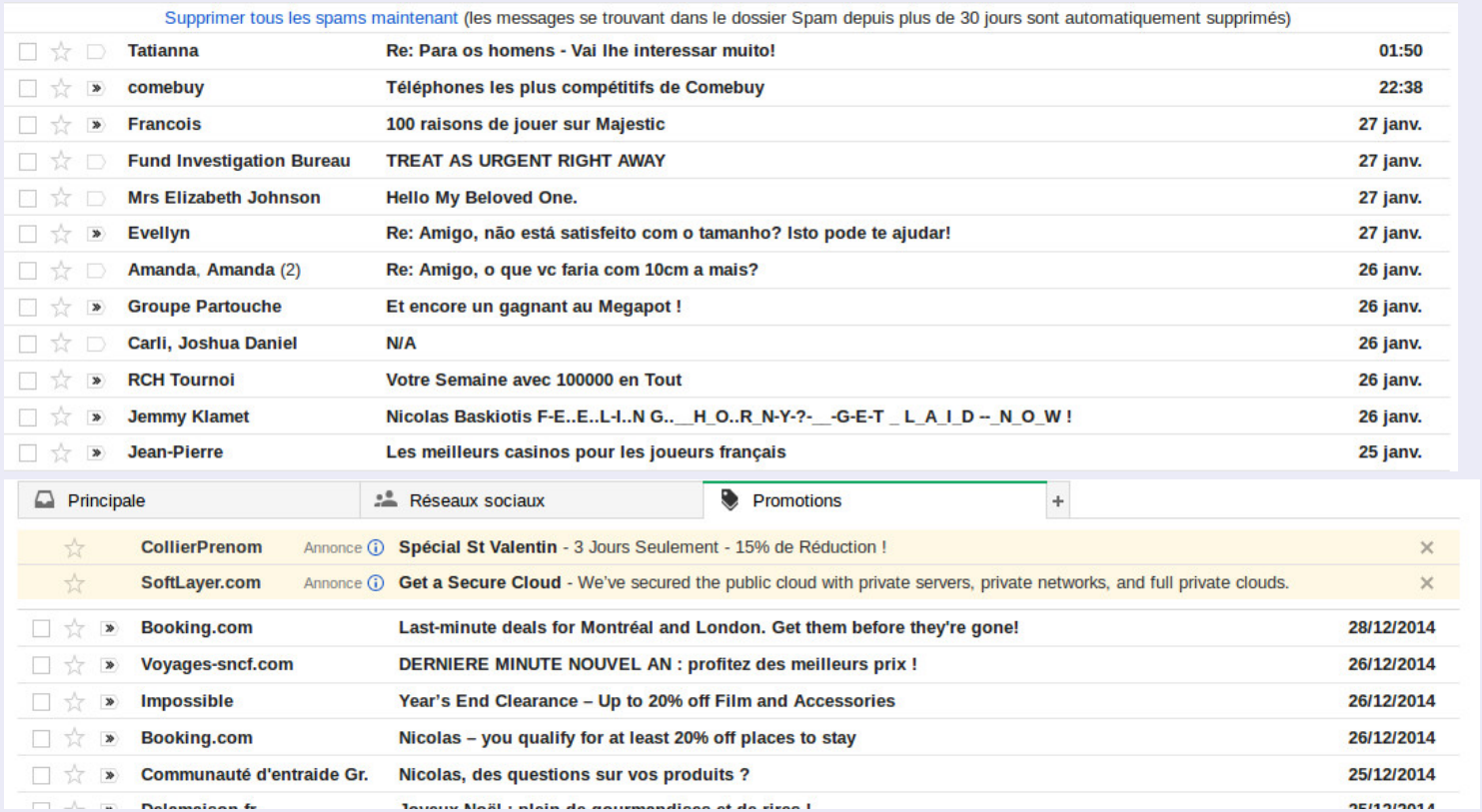


A votre avis, ca regroupe quoi ?

En texte

Classification de documents

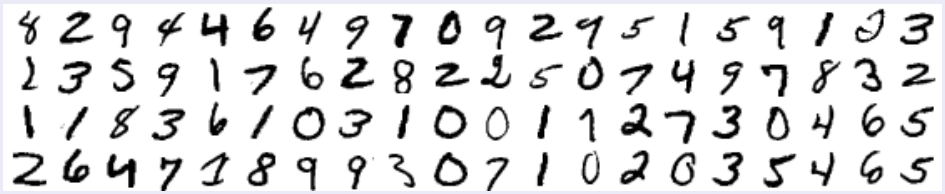
E-mails en spam, shopping, travail, . . .



gmail.com

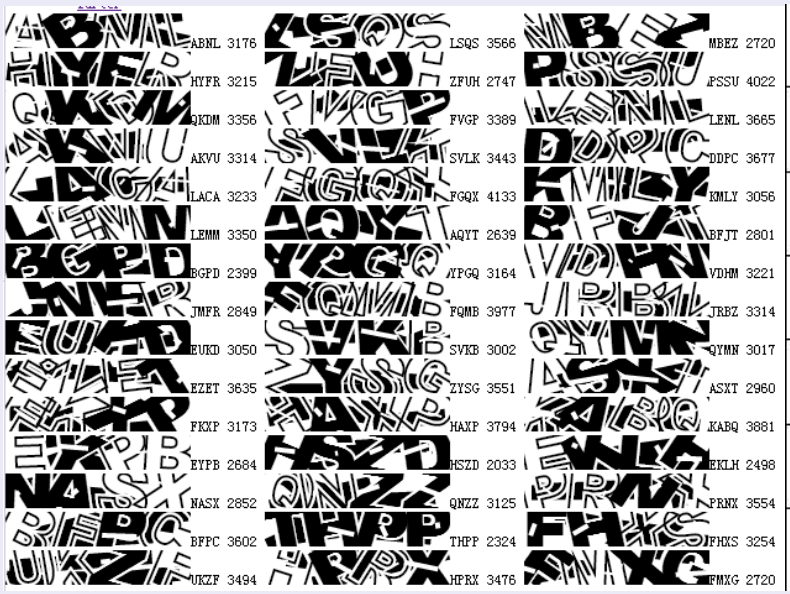
En texte toujours

Reconnaissance de chiffres



Ou de captcha

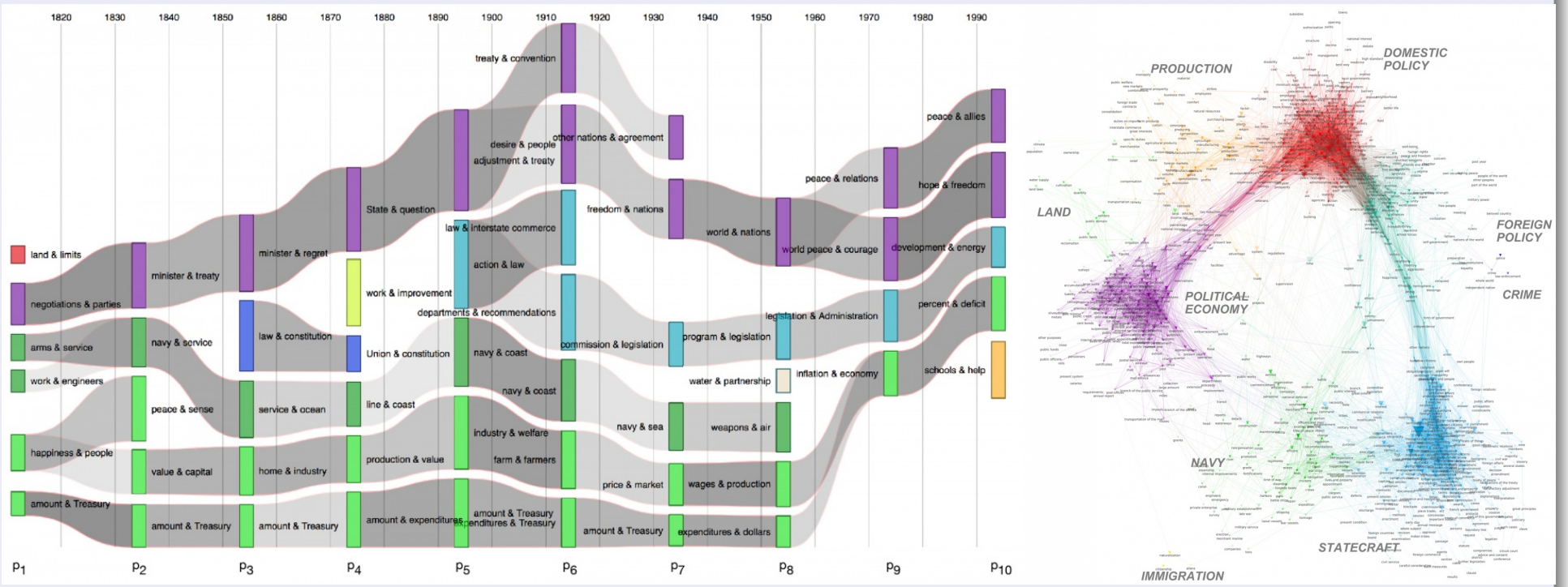
[Yann et al. 08], Newcastle University



Characters under typical distortions	Recognition rate
	~100%
	96+%
	100%
	98%
	~100%
	95+%

Sur des documents

Détection de thèmes (topic detection)



Analyse de 255 discours de l'état de l'union, États-Unis [Rule et al, 2014]

Et plein d'autres applications : traduction, détection de plagiat, résumé automatique, . . .

$\Rightarrow$ U.E. Traitement Automatique du language

En image

Détection de visages (opencv)

Mais aussi ... (betafaceapi.com)

	Score: 0.42 X: 398.67 Y: 29.66 Width: 26.79 Height: 26.79 Angle: -5.45	age : 37 (16%), gender : male, race : white, chin size : average, color background : 4c5042 (15%), color clothes middle : 3295eb (48%), color clothes sides : 38a9f5 (96%), color eyes : ac8066, color hair : fbf2ea (80%), color mustache : a56855 (65%), color skin : dbb5a1, eyebrows corners : extra low, eyebrows position : average, eyebrows size : extra thin, eyes corners : low, eyes distance : average, eyes position : average, eyes shape : extra round, glasses rim : no, hair beard : none, hair color type : blond (80%), hair forehead : yes, hair length : none, hair mustache : thick, hair sides : very thin, hair top : short, head shape : average, head width : extra narrow, mouth corners : low, mouth height : extra thin, mouth width : extra small, nose shape : extra straight, nose width : wide, teeth visible : no [collapse]
	Score: 0.57 X: 216.66 Y: 155.08 Width: 28.34 Height: 28.34 Angle: 0.95	age : 46 (23%), gender : male, race : white, chin size : extra small, color background : 0c0c0d (36%), color beard : 4a2617 (50%), color clothes middle : a22e55 (82%), color clothes sides : a54031 (74%), color eyes : 966a58, color hair : 655348 (77%) , color skin : b98f78, eyebrows corners : average, eyebrows position : extra high, eyebrows size : extra thin, eyes corners : average, eyes distance : close, eyes position : extra low, eyes shape : extra thin, glasses rim : no, hair beard : short, hair color type : brown light (77%), hair forehead : no, hair length : short, hair mustache : none, hair sides : thin, hair top : short, head shape : rect, head width : extra wide, mouth corners : average, mouth height : extra thin, mouth width : average, nose shape : average, nose width : extra narrow, teeth visible : no [collapse]

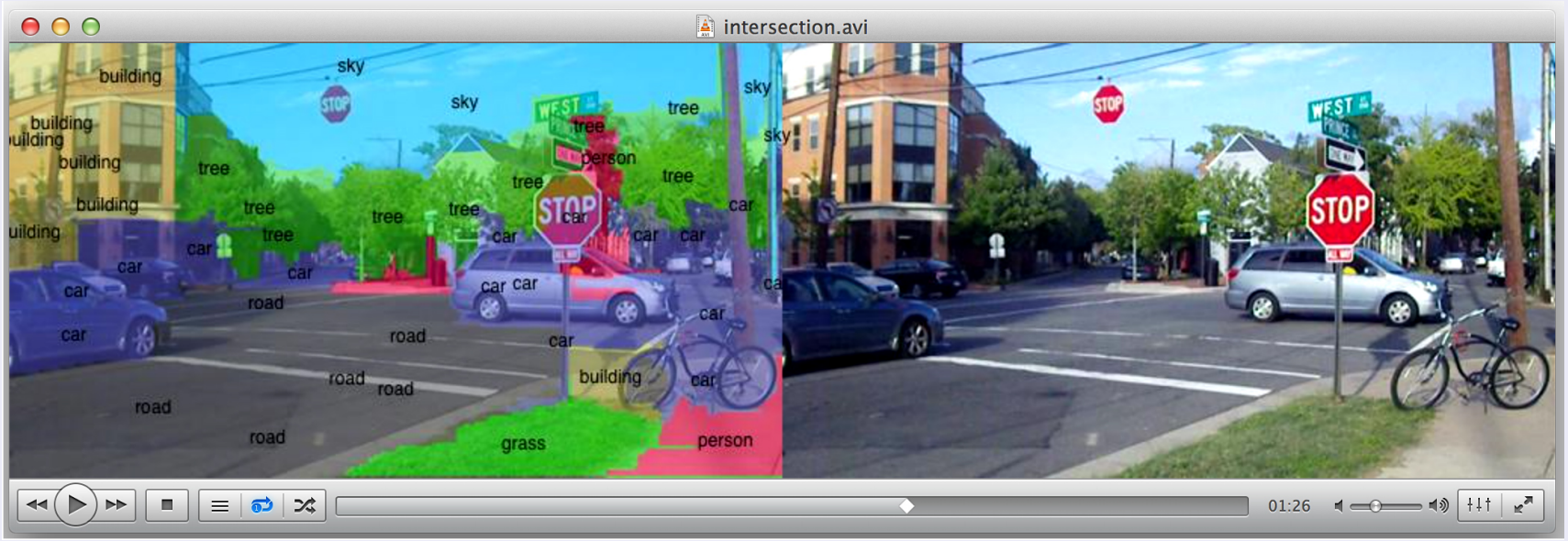
En image

Classification et organisation automatique

En image

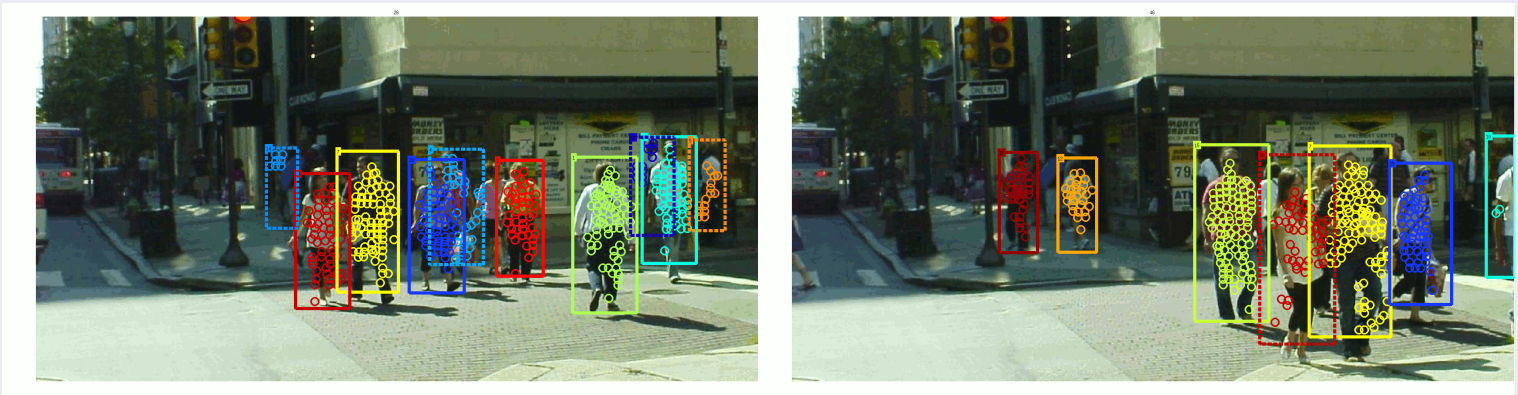
Détection d'objets

teradeep.com, Purdue University



Tracking

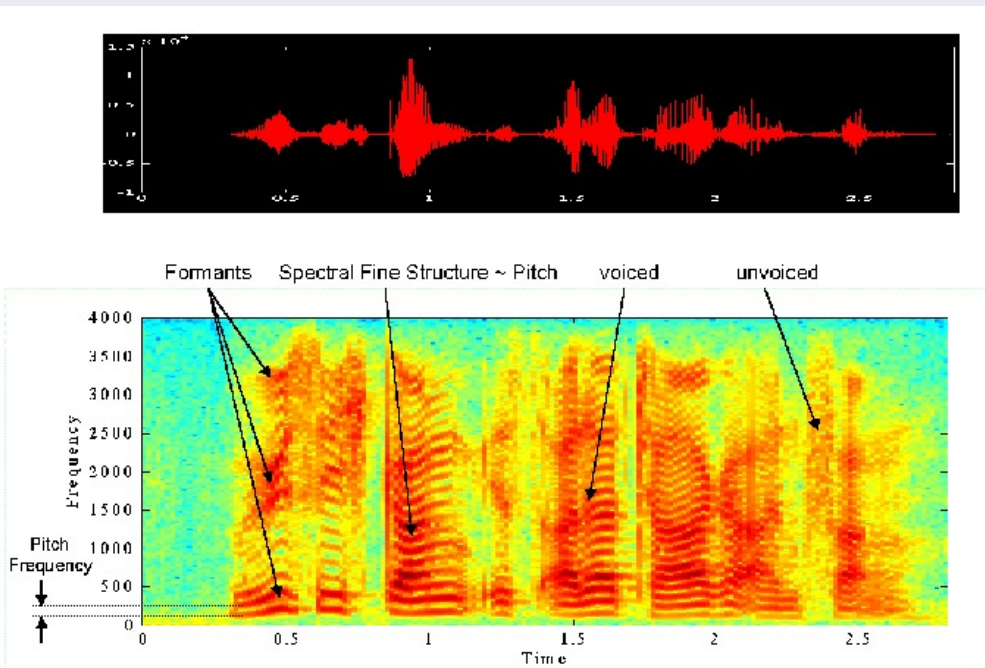
[Fragkiadaki et al. 12], Pennsylvania University

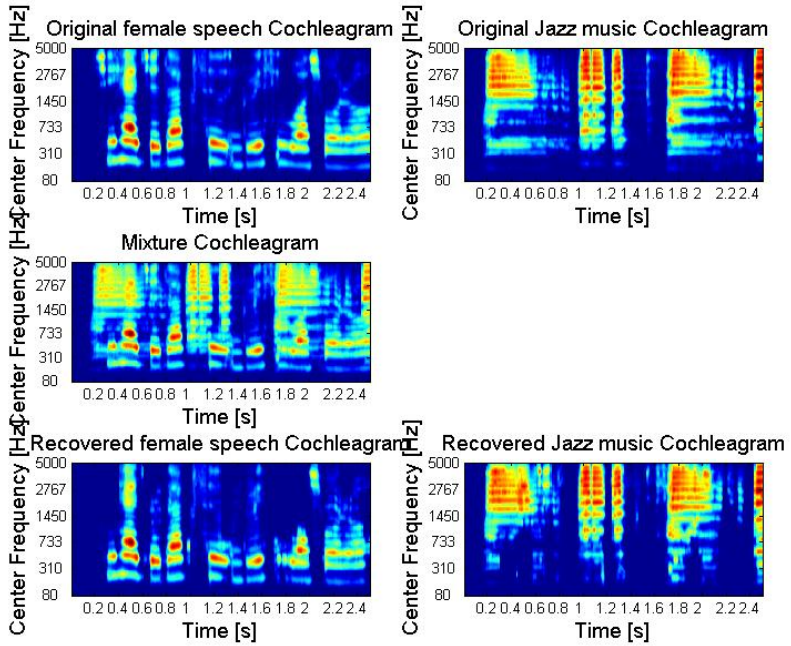


Et l'audio . . .

Reconnaissance de la parole, séparation de sources

http://markus-hauenstein.de

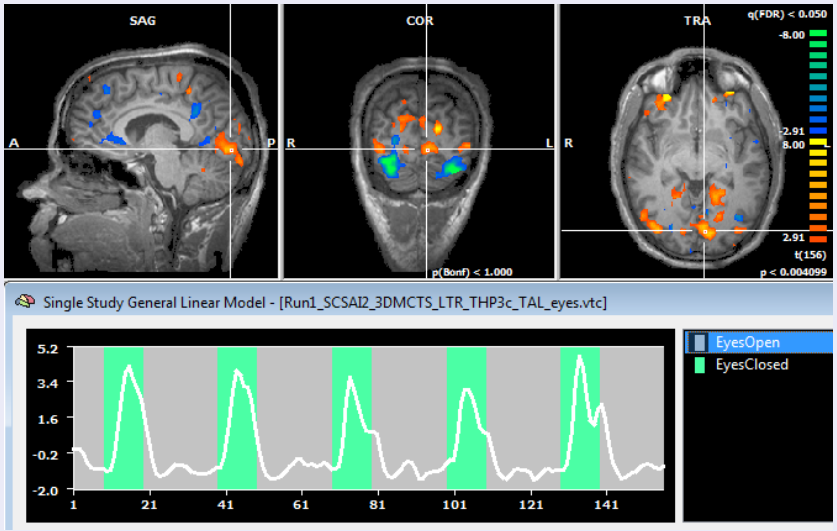




Mais aussi débruitage, transcription musicale, reconnaissance du locuteur, classification/identification de musiques. . .

Interface cerveau-machine (BCI)

Classification d'actions, de pensées

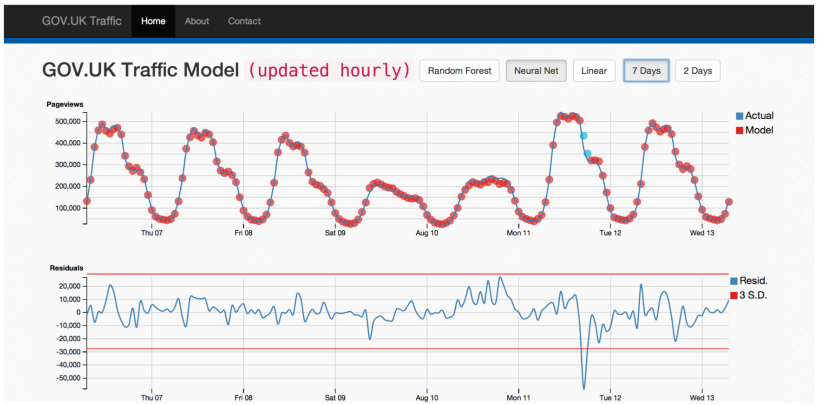


Contrôle

Objets connectés

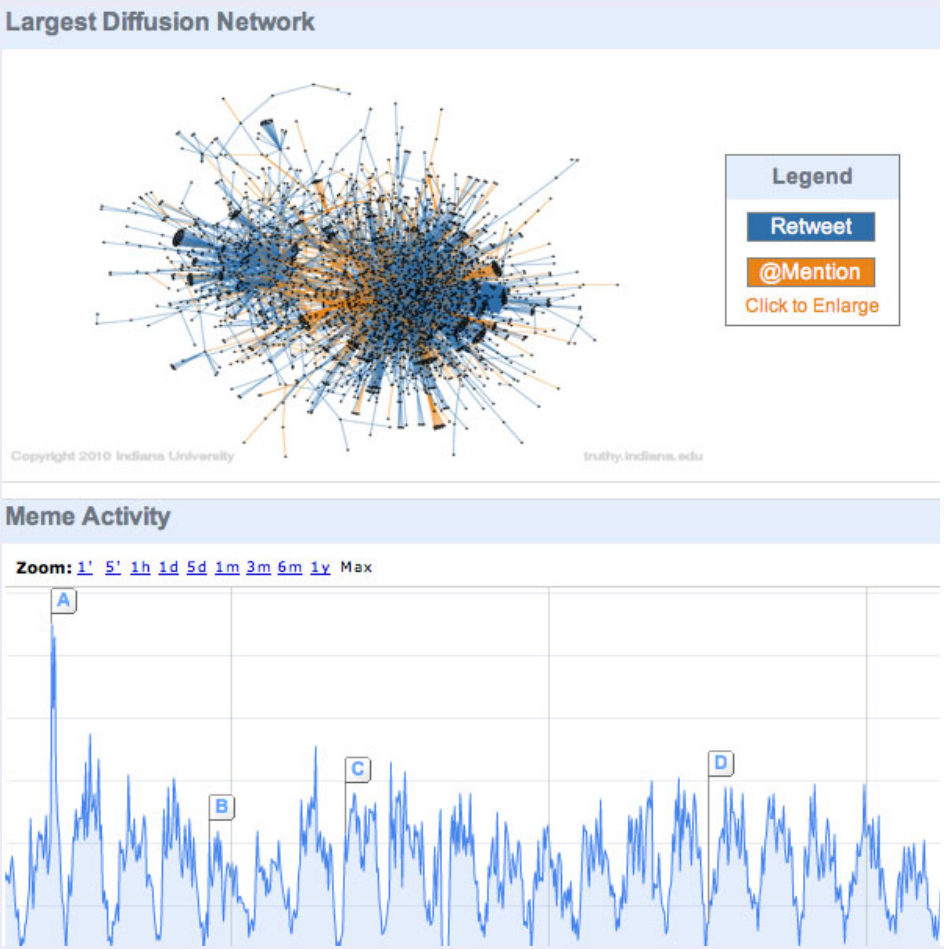
Traqueurs d'activité

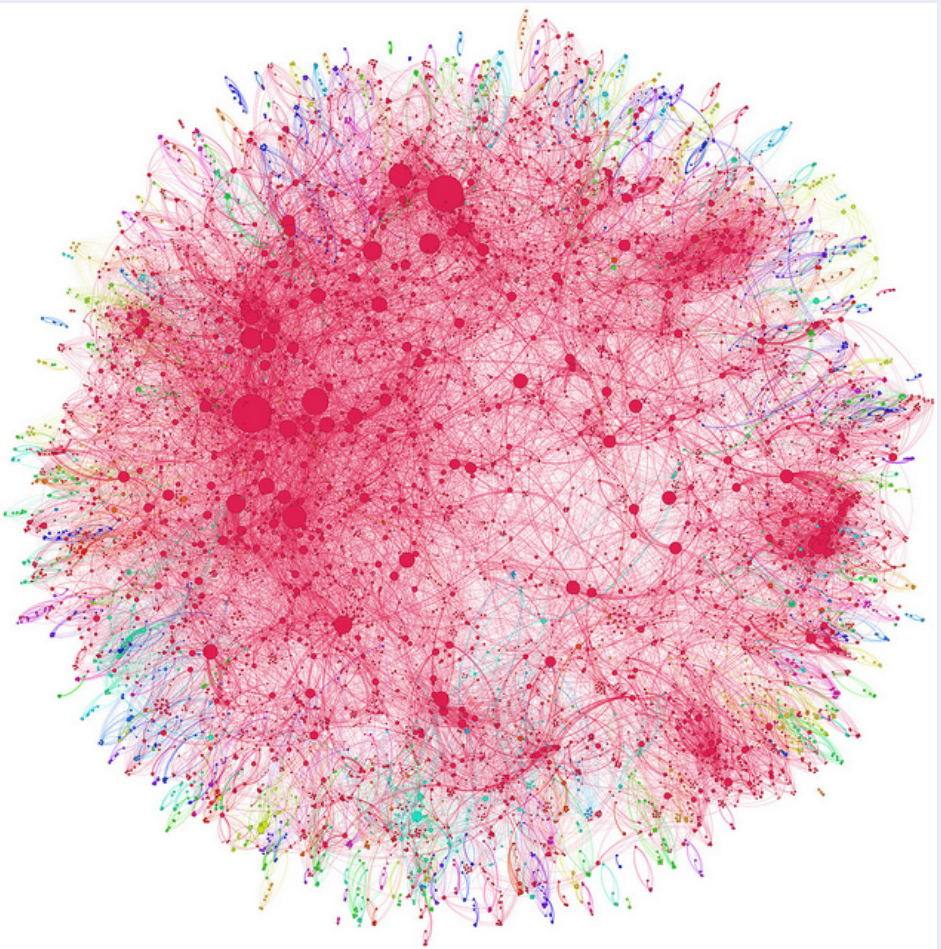
Surveillance vidéo, monitoring consommation électrique, sécurité réseau



Réseaux sociaux

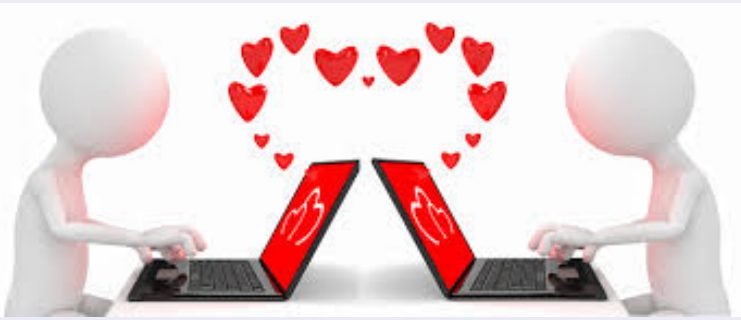
Détection de communauté, phénomènes de diffusions, classification





Matchmaking

de profils, sites de rencontre



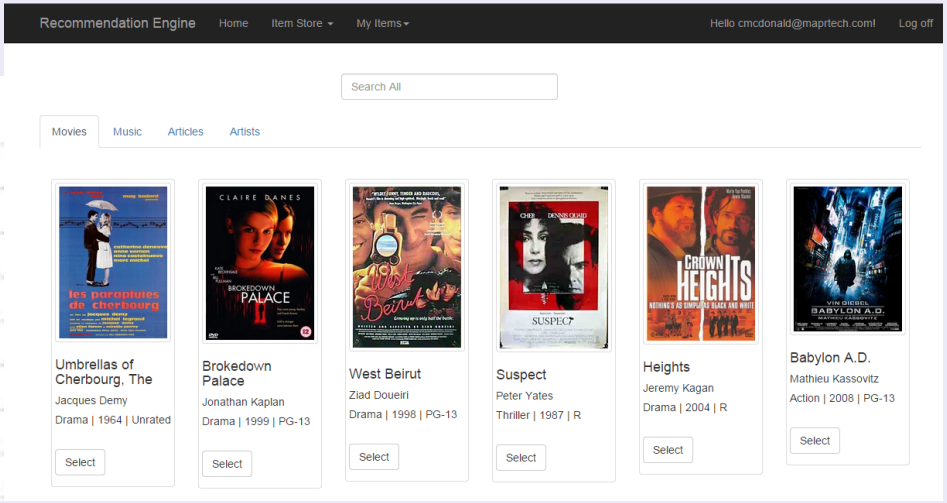
badoo
tinder meetic
adopteunmec.com

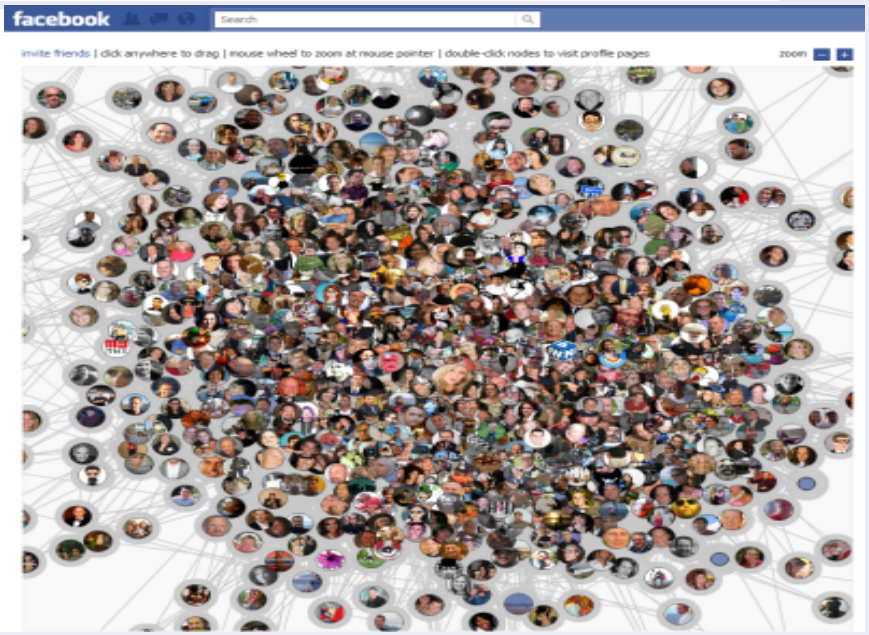
Experts, CV - Emplois, Jeux

LinkedIn
viadeo

Systèmes de recommendation

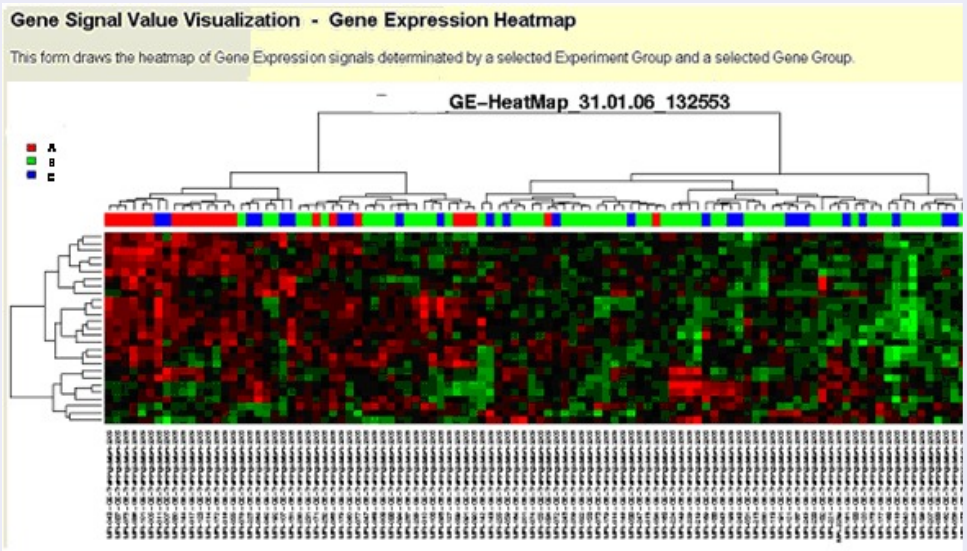
De musiques, de films, de produits, d'amis





Et dans d'autres sciences

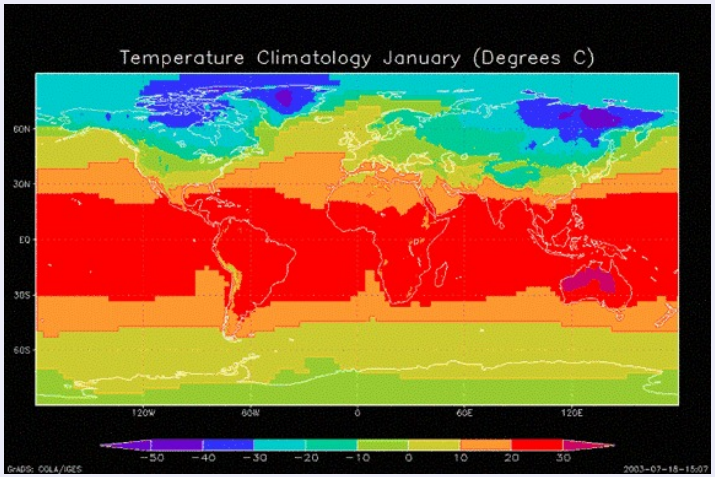
Biologie IZBI, Leipzig University



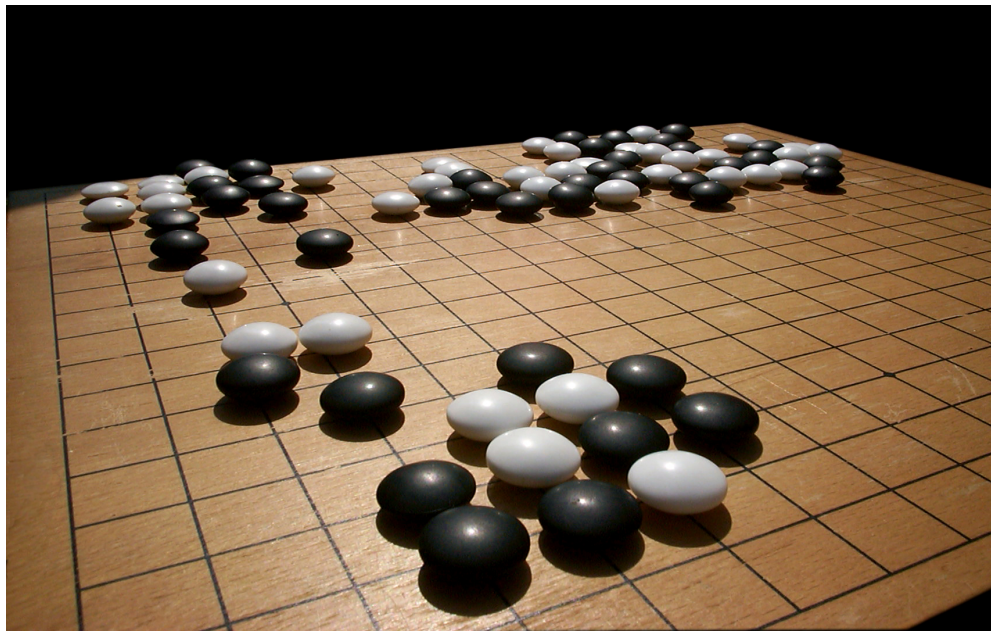
Économie

Astronomie

Climatologie (complétion données)

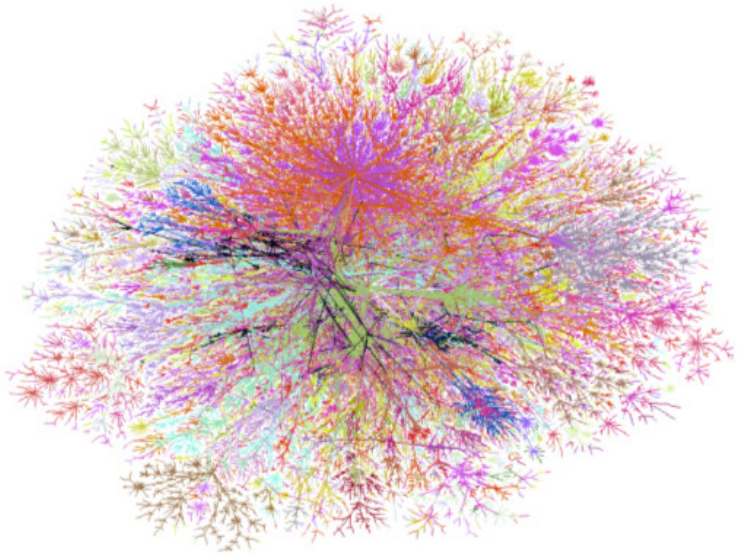


Dans les jeux et la robotique



L'apprentissage aujourd'hui : Big Data

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



>10B useful webpages

Carnegie Mellon University

extrait du cours d'A. Smola

Entreprises concernées :

Yahoo, Google, Amazon, Netflix, Microsoft, Apple, Xerox, Samsung, Critéo, Facebook, Twitter, Flicker, Instagram, Reddit, Valve, Steam, Deezer, Dailymotion, Youtube, STIF, SNCF, AXA, EDF, GDF-Suez, Veolia, Safran, Thalès, les médias, . . .

Plan

1 Organisation de l'UE

2 Introduction

3 **Les problèmatiques générales**

4 Premier exemple : classification de films et arbres de décision

L'apprentissage (selon M. Sahani, UCL)

En quelques mots

- Trouver des structures, des régularités dans des observations.
- Prédire de nouvelles observations.

Touche à beaucoup de domaine, interdisciplinarité très forte

- Statistiques : théorie de l'apprentissage, fouille de données, inférence
- Informatique : IA, vision, RI
- Ingénierie : signal, controle, robotique
- Science cognitive, psychologie, neuroscience, épistémologie
- Économie : théorie de la décision, théorie des jeux

L'apprentissage artificiel

- étudie les algorithmes qui améliorent leur performance sur une tache donnée en fonction de leur expérience.
- fondements mathématiques, informatiques et applications concrètes des systèmes qui apprennent, raisonnent et agissent.

Quand appliquer l'apprentissage ?

Lorsque :

- l'expertise humaine est absente
- impossible d'expliquer cette expertise
- les solutions sont dynamiques
- les solutions doivent être adaptées à beaucoup de cas spécifiques
- la taille du problème est trop grand pour que l'humain puisse le résoudre

Les grandes familles

Apprentissage supervisé

- Classification
- Regression
- Forecasting
- Complétion de données
- Ranking
- Recommendation

Apprentissage non supervisé

- Clustering
- Apprentissage de représentation, de dictionnaire
- Analyse de séquences
- Représentation hiérarchique
- Détection d'anomalies

Apprentissage par renforcement

- Apprendre à jouer
- Apprendre à interagir avec l'environement

Apprentissage supervisé

Données du problème

- Une représentation X des objets de l'étude
- Une sortie d'intérêt y qui peut être numérique, catégorielle, structurée, complexe (label, réponse, étiquette, ...)
- Un ensemble d'exemples, d'échantillons, sous leur représentation X et avec leur sortie connue $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

Objectifs

- Prédire de manière précise la sortie y pour un nouvel exemple x non vu
- Comprendre quels facteurs influencent la sortie
- Évaluer la qualité de nos prédictions

Apprentissage non supervisé

Données du problème

- Une représentation X des objets de l'étude
- Un ensemble d'exemples, d'échantillons, sous leur représentation X, $\{x_1, \ldots, x_n\}$
- Pas de variable de sortie !

Objectifs

- Trouver des groupes d'objets "semblables"
- Organiser les données d'une manière "logique"
- Trouver les "similarités" des objets
- Trouver des "représentations" des objets

$\Rightarrow$ on ne sait pas bien ce que l'on cherche

$\Rightarrow$ tout un art !

Apprentissage par renforcement

Apprentissage continu en fonction du retour d'expérience

Données du problème

- Un état décrit l'environement courant
- Un ensemble d'actions sont possibles
- Une politique permet de choisir en fonction de l'état l'action à effectuer
- A l'issue de chaque action, une récompense est observée

Objectifs

- S'améliorer ! (améliorer la politique de choix de l'action)
- Éviter les situations d'échecs
- Comprendre la dynamique du problème

Ce cours

C'est une introduction !

- Problématiques générales (biais, variance, évaluation, sur-apprentissage, représentation des données)
- Algorithmes supervisées (k-nn, bayésien, perceptron, réseaux de neuronnes, svm, . . .)
- Algorithmes non supervisées (hiérarchique, k-means, . . .)
- Apprentissage par renforcement (introduction)
- Pas de méthodes avancées $\Rightarrow$ en M2 !

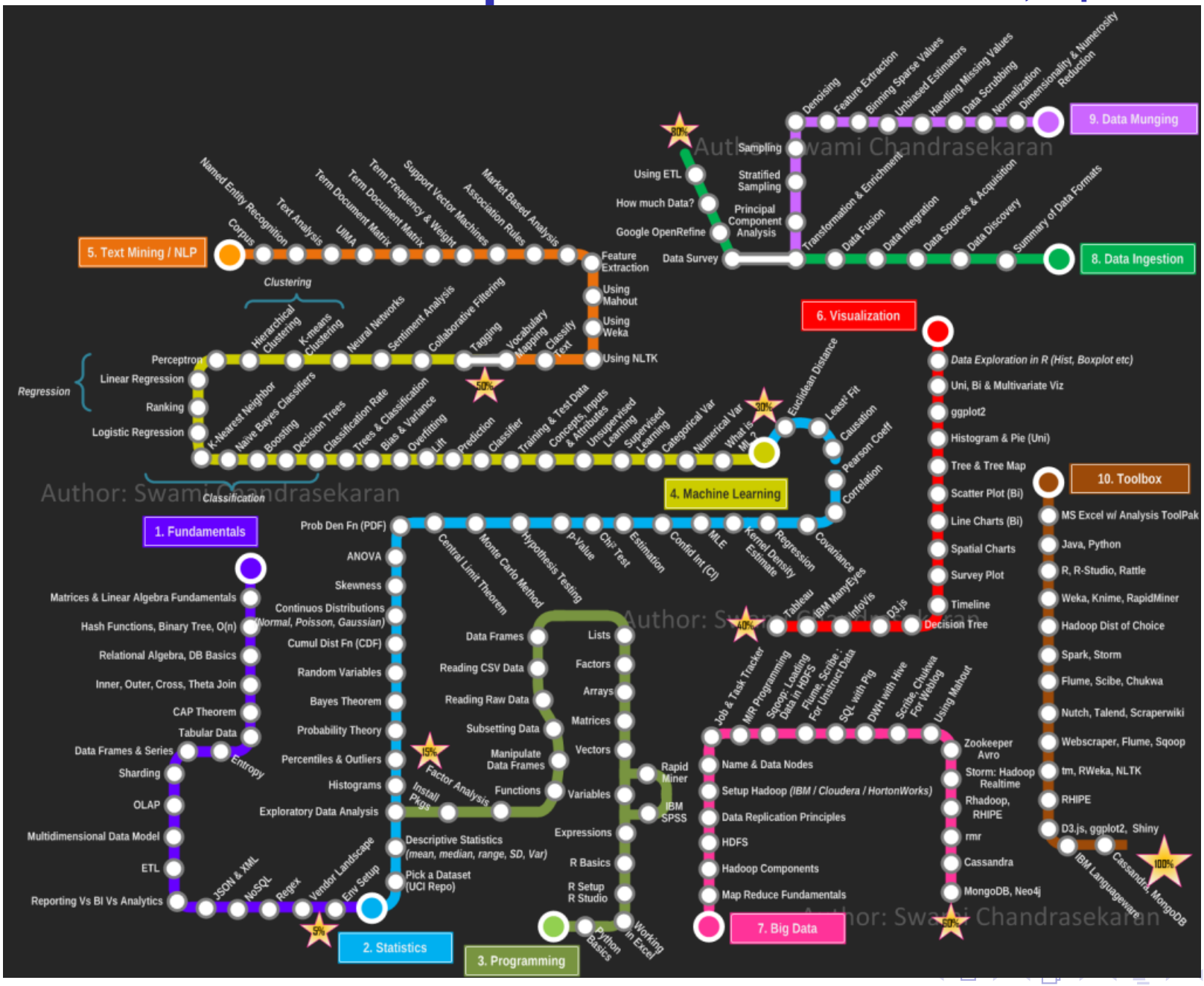
Objectifs

- Comprendre les différentes techniques en profondeur, algorithmiquement et théoriquement
- Savoir évaluer une approche
- Savoir sélectionner des méthodes
- Savoir traiter des données de différentes natures

L'apprentissage statistique est au cœur de la formation d'un *data scientist*

Data scientist map

Chandrasekaran, 2013]



Plan

1. Organisation de l'UE
2. Introduction
3. Les problèmatiques générales
4. **Premier exemple : classification de films et arbres de décision**

Formalisation de l'apprentissage supervisé

On dispose :

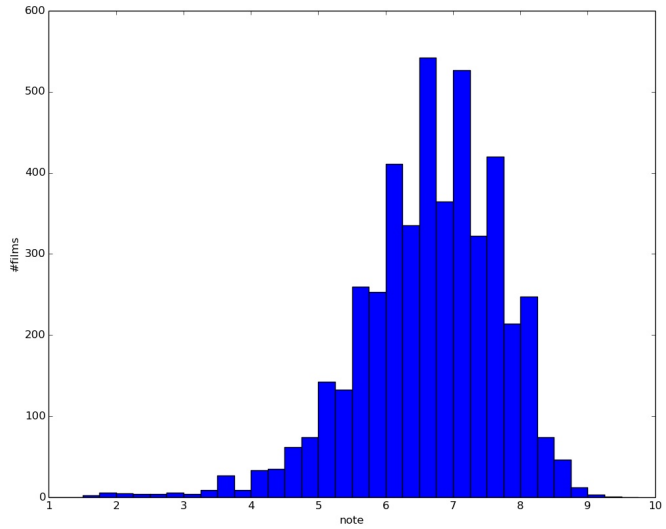
- d'un espace de représentation $\mathcal{X}$, usuellement $\mathbb{R}^n$
  n est la dimension de l'espace de représentation,
  chaque dimension $=$ un attribut
- d'un ensemble d'exemples X décrit dans cette espace :
  $x \in X, x = (x_1, x_2, x_3, \ldots, x_n)$
- d'un ensemble d'étiquettes/labels Y décrivant les classes d'intérêt
  quand Y contient deux classes $\rightarrow$ classification binaire, usuellement
  $Y = \{0, 1\}$ ou $Y = \{-1, 1\}$
- pour chaque exemple x^i de X, son étiquette y^i
  $\Rightarrow$ ensemble d'apprentissage $E = \{(x^i, y^i)\}$

On veut :

Trouver une fonction $f : \mathcal{X} \rightarrow Y$ telle que la prédiction sur de futurs exemples soit la plus précise possible.

Première étape : choix de la représentation

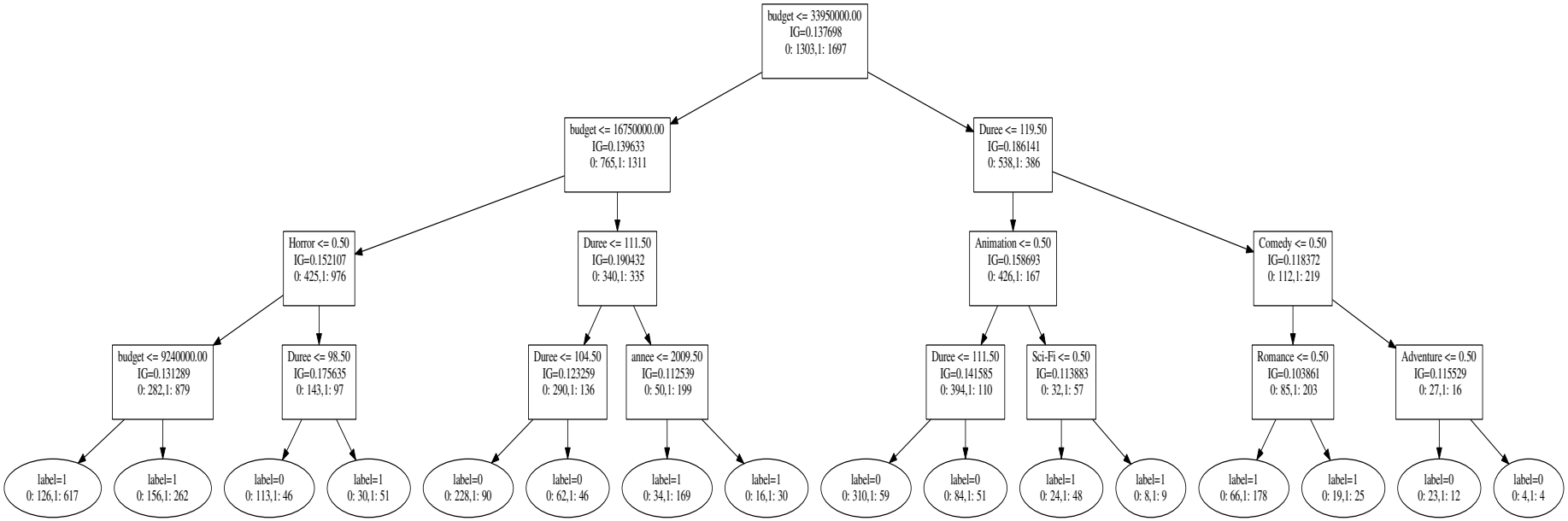
Collection de films, des notes à chaque film



Questions

- Comment représenter un film ?
- Comment classifier ?

Arbres de décision

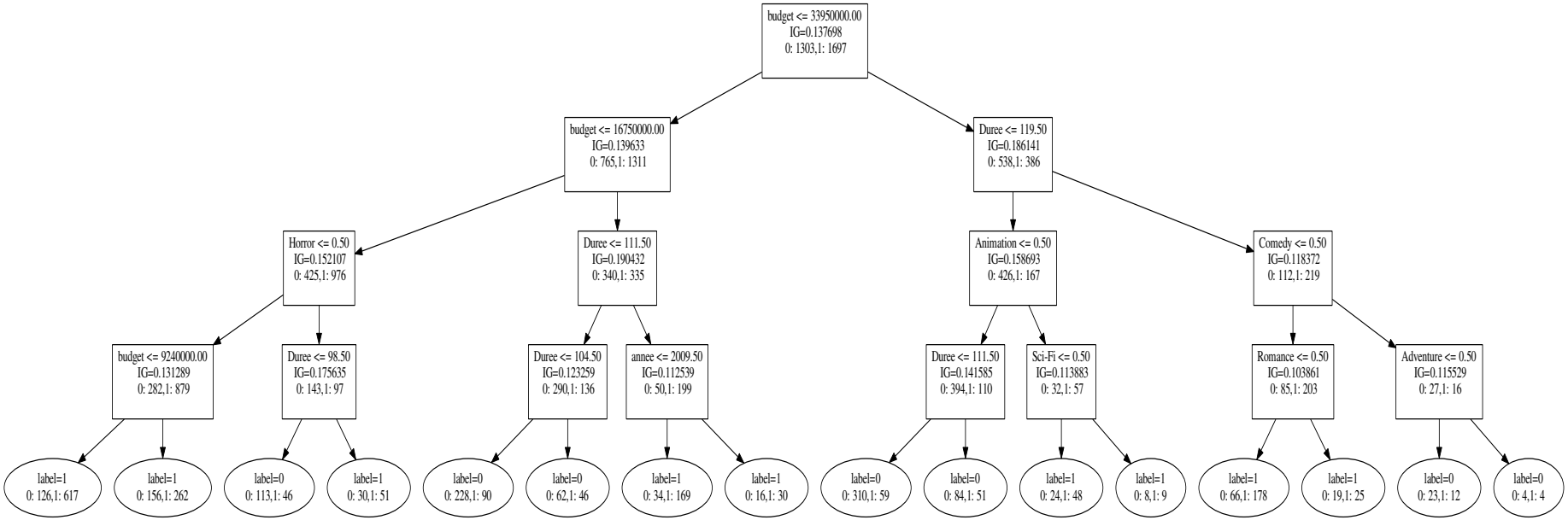


Principe

- Chaque nœud interne : un test sur une des dimensions de $\mathcal{X}$
- Chaque branche : un résultat du test
- Chaque feuille : un label de Y

$\Rightarrow$ classification en parcourant un chemin de la racine à une feuille.

Arbres de décision

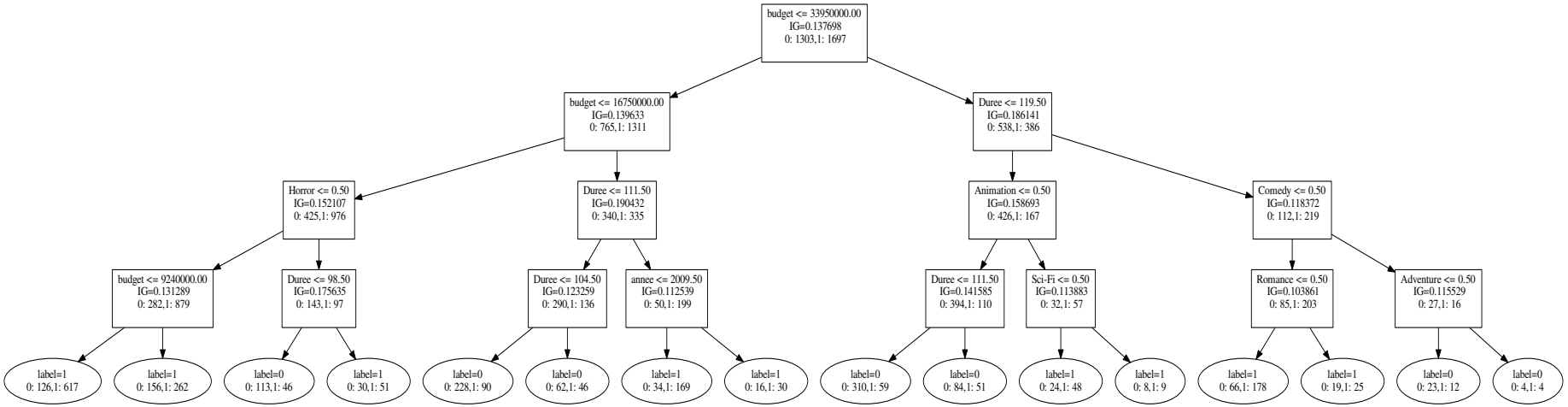


Exercice

Comment représenter :

- $x_2 = 1 \wedge x_3 = 1$?
- $x_1 = 1 \vee x_5 = 1$?
- $x_2 = 1 \wedge x_3 = 1 \quad \vee \quad x_1 = 1 \wedge x_2 = 0$?

Apprentissage d'un arbre de décision

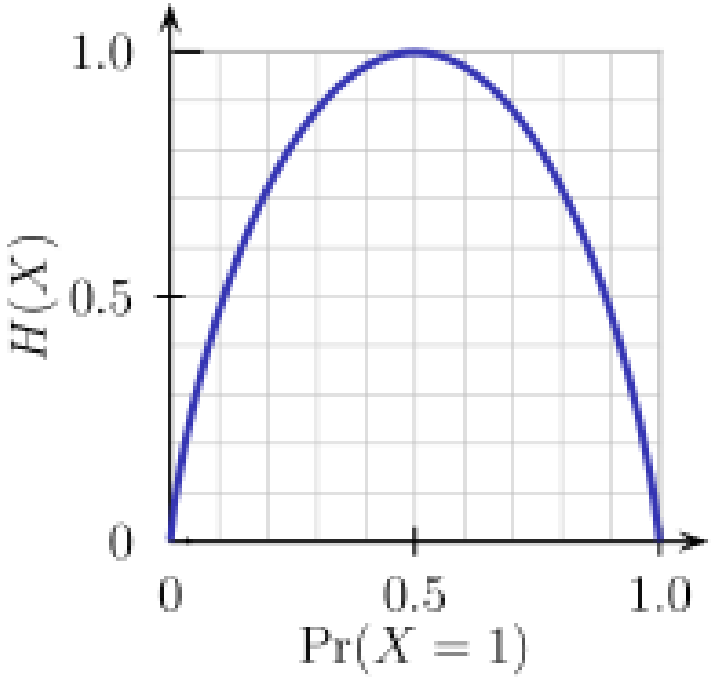


Algorithme glouton, top-down

Initialisation à la racine, considérer tous les exemples

- Si le nœud n'est pas pur, alors
  - Trouver x_i le "meilleur" attribut pour ce nœud et les cas tests associés
  - Pour chaque test, créer un fils au nœud courant
  - Faire "tomber" les exemples du nœud courant à leur fils correspondant
- sinon transformer le nœud en feuille.

Sélectionner le meilleur attribut



Entropie d'une variable aléatoire

Soit X une variable aléatoire pouvant prendre n valeurs :

$$H(X) = -\sum_{i=1}^{n} P(X = u) log(P(X = i))$$

Plus l'entropie est grande, plus le désordre est grand.
Entropie nulle $\rightarrow$ pas d'aléa.

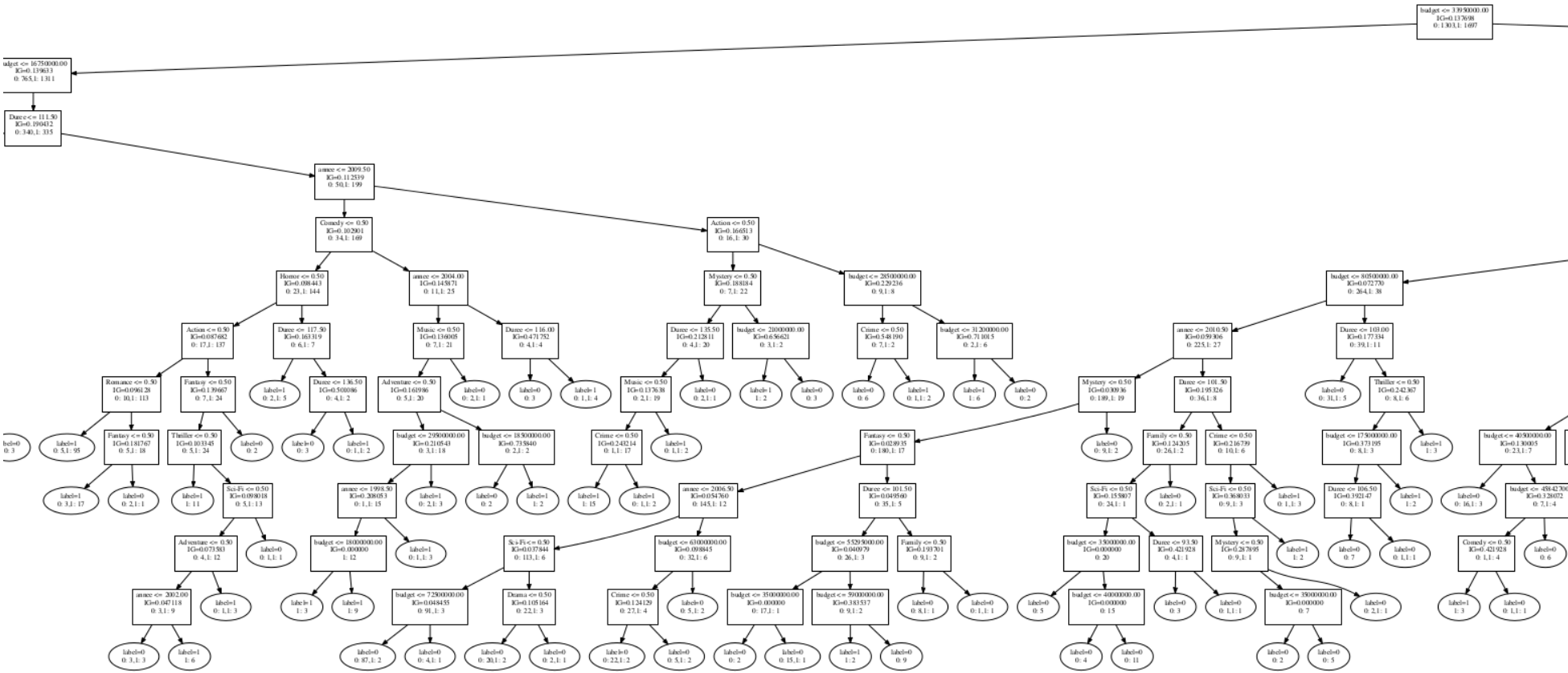
Sélectionner le meilleur attribut

Entropie d'un échantillon : cas binaire

- X un ensemble de données
- p_+ la proportion d'exemples positifs
- p_- la proportion d'exemples négatifs
- $H(X) = -p_+ log(p_+) - p_- log(p_-)$

Cas général : entropie conditionelle

- $H(X|Y = y) = -\sum_{i=1}^{n} P(X = i|Y = y) log P(X = i|Y = y)$
- $H(X|Y) = \sum_{y \in Y} P(Y = y) H(X|Y = y)$

$\Rightarrow$ Gain d'information : $I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Où s'arrêter ? Est ce un bon modèle ?