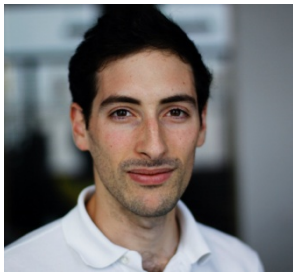
Cours MASH: Projets informatiques

enseignant:

Simon Lacoste-Julien

TD / suivi par:

Fajwel Fogel

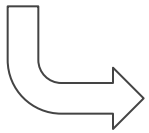


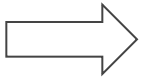
Équipe-Projet SIERRA, INRIA / ENS

Je me présente...

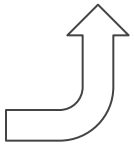
Simon Lacoste-Julien
Chercheur CR
Équipe-Projet SIERRA, INRIA - École Normale Supérieure

McGill





UNIVERSITY OF CAMBRIDGE

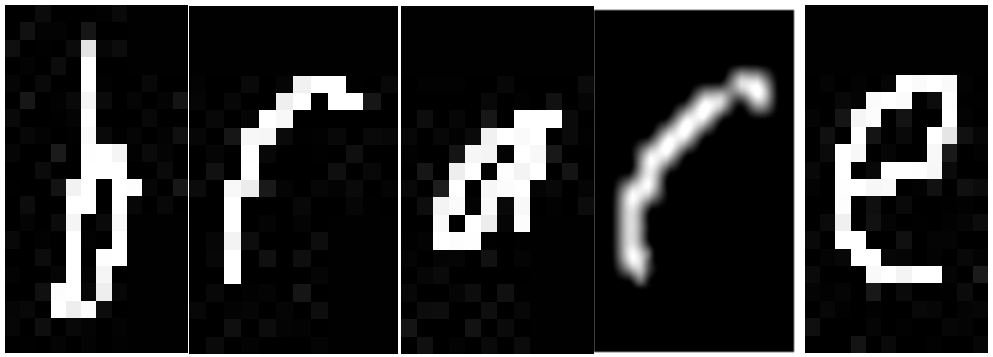


- thèmes de recherche:
  - prédiction structurée
  - optimisation
  - applications: vision, NLP, information retrieval, computational biology

parenthèse:
prédiction structurée

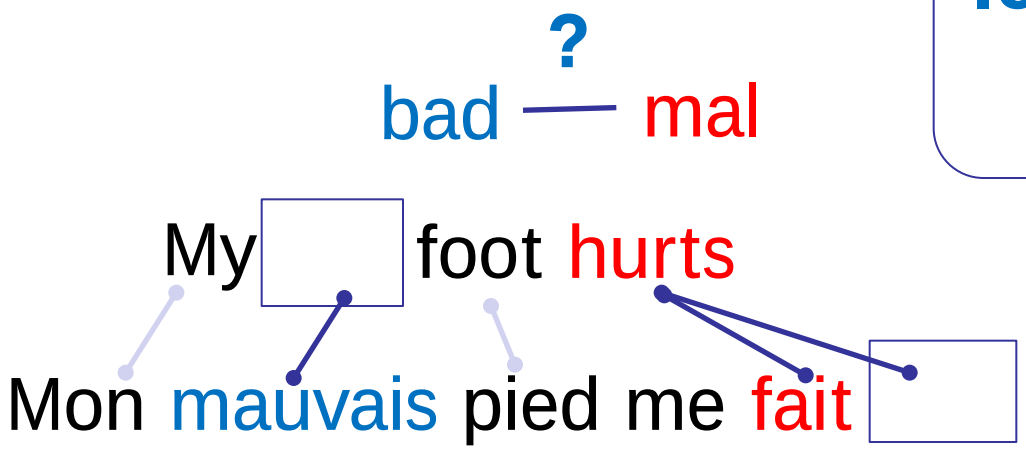
1) Exploiter structure: motivation

- reconnaissance de mots



brace

- alignement de mots



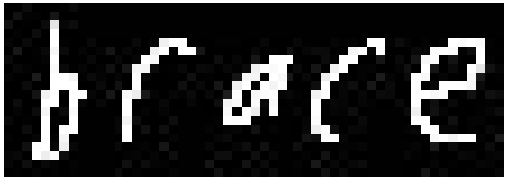
le contexte aide!

Prédiction structurée:

Entrée	Sortie
$\mathbf{x} \in \mathcal{X}$	$\mathbf{y} \in \mathcal{Y}$

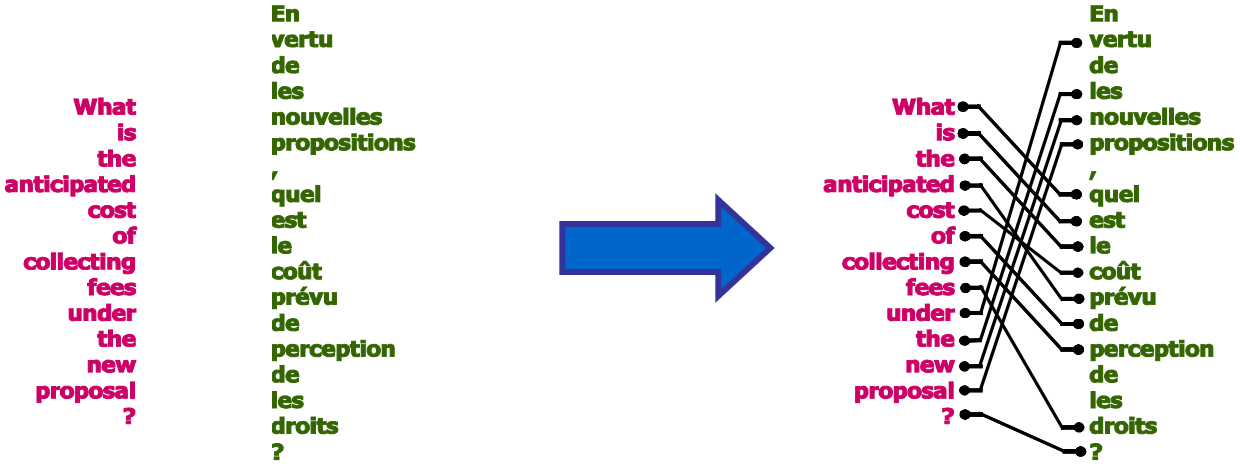
nombre exponentiel!

Reconnaissance de mots écrits



brace

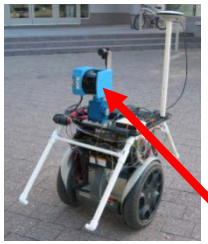
Alignement de mots



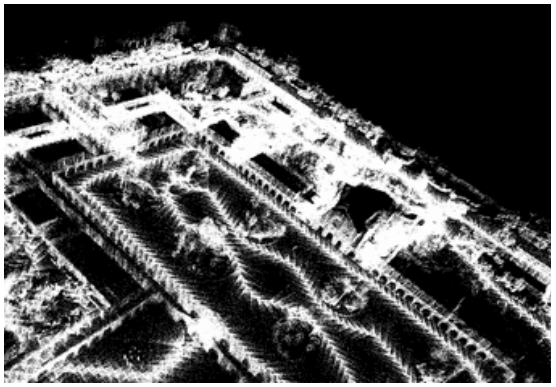
d'autres exemples...

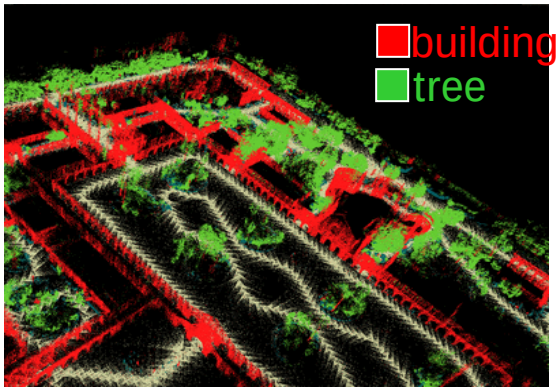
fonction d'erreur structurée: $\ell(\mathbf{y}, \mathbf{y}')$	Entrée $\mathbf{x} \in \mathcal{X}$	Sortie $\mathbf{y} \in \mathcal{Y}$
Traduction automatique	'Ce n'est pas un autre problème de classification.'	'This is not just another classification problem.'

Reconnaissance objet 3D



télémètre laser



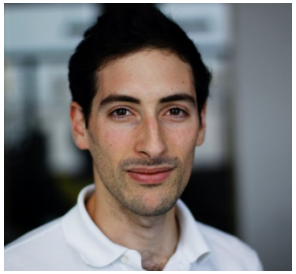


(fin de la parenthèse!)

Sujet du cours

- cours pratique!
  - but: mettre en pratique les techniques d'apprentissage automatique sur des **vraies données**
  - se familiariser avec les outils informatiques (python, scikit-learn)
  - ** PROJET **

Structure du cours



- projet en équipe de 2 (max)
- « office hour » (OH) par Fajwel Vogel
  - en général (+ exceptions): les lundis de 10h-11h -> avec système de réservation
- remise du projet lundi 30 mars
- toutes les infos sur site web à venir...

Plan de cours + jalons

- cours: jeudi 27 nov. 10h-12h: TD python / scikit-learn par Fajwel
- lundi 1er décembre:
  - jalon: avoir formé son équipe + quelques pistes de projet
  - OH obligatoire: rendez-vous avec Fajwel
- mercredi 17 décembre:
  - jalon: envoie par mail 1 page avec résumé du projet
  - OH obligatoire: rendez-vous avec Fajwel (avoir feedback)
- [OH optionnelle commence lundi 5 janvier et continue chaque semaine (avec exceptions)]
- lundi 16 février:
  - JALON: présentation de 10 minutes devant la classe pour chaque équipe sur l'état du projet (problème, approche, résultats)
- **lundi 30 mars – évaluation finale**
  - remettre un rapport écrit sur le projet d'environ 5 pages
  - session poster (8 pages A4); 10 minutes de présentation évaluée

Choix de projet

- suggestions de projet sera disponibles sur le site web pour le 24 novembre
  - principalement kaggle
  - quelques choix de 1000mercis (contact: Anne Guérin) – [option: projet spécial avec éram]
- vous pouvez suggérer votre propre projet:
  - quelles sont les données?
  - quelle est la tâche? les sorties désirées?
  - quel est la *métrique d'évaluation*?
- critères pour projet:
  - défi / votre intérêt / vous faire apprendre!
  - pour évaluation: feedback individualisé pour savoir ce que vous devriez accomplir...

Quelques étape en analyse de données (apprentissage automatique appliqué)

1) Définition du problème
2) Télécharger données
3) Exploration données: résumer, visualisation
4) Data processing: sous-échantillonner, nettoyage, standardisation, transformation, définir « features »
5) Choix modèle / algorithme
6) Évaluer les résultats [répéter 3-5!]
7) Présenter solution / résultats

Kind of features [statistics terminology]

a) **Nominal** qty.: distinct symbols with no ordering

e.g. color $\in$ {red, blue, green}

b) **Ordinal** qty.: values can be ordered

e.g. for temperature: cool < mild < hot

c) **Interval** qty: fixed units, but no **scale** [no absolute 0]

e.g. p = position in space of an object; 2p doesn't make sense

but $p_2 - p_1$ does

d) **Ratio** qty: absolute zero gives meaningful **scale**

e.g. mass of object; frequency of words in document, ...

scikit-learn...

(cours le jeudi 27 novembre)

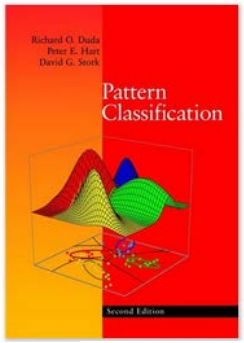
Suggestion approche pour projet

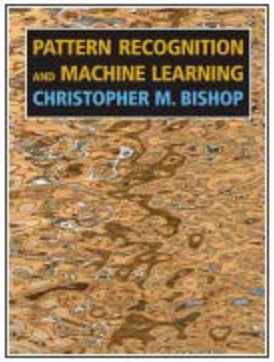
- Commencer avec méthodes / modèles simples
- Étudier où ça brise!
- Modifier features / méthode / modèle en conséquence
- Répéter!

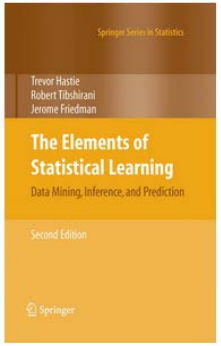
exemples kaggle...

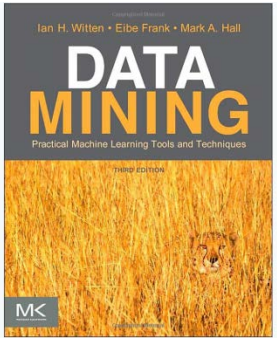
Quelques ressources

- Logiciels:
  - SciKit Learn (Python): http://scikit-learn.org
  - Weka (Java): http://www.cs.waikato.ac.nz/ml/weka/
  - RapidMiner (nicer GUI?): http://rapid-i.com/
- Livres:
  - Pattern Classification (Duda, Hart & Stork)
  - Pattern Recognition and Machine Learning (Bishop)
  - **Data Mining** (Witten, Frank & Hall)
  - The Elements of Statistical Learning (Hastie, Tibshirani & Friedman)
- Cours en python:
  - cours cs188 de Dan Klein à Berkeley: http://inst.eecs.berkeley.edu/~cs188/fa10/lectures.html









Action points!

- Former vos équipes (pour 1er décembre) + commencer à réfléchir à projets (voir site web)

- Avant cours du 27 novembre:
  - installer Anaconda Python
  - faire tutoriel Python / scikit-learn (info par mail)