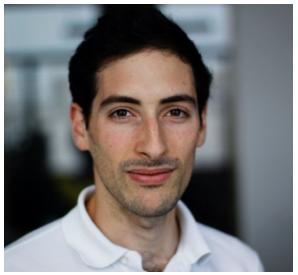
Cours MASH: Projets informatiques

enseignant:

Simon Lacoste-Julien

TD / suivi par:

Fajwel Fogel

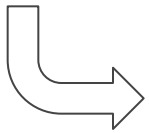


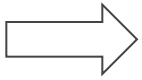
Équipe-Projet SIERRA, INRIA / ENS

Je me présente...

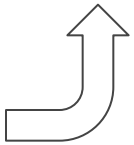
Simon Lacoste-Julien
Chercheur CR
Équipe-Projet SIERRA, INRIA - École Normale Supérieure

McGill





UNIVERSITY OF CAMBRIDGE

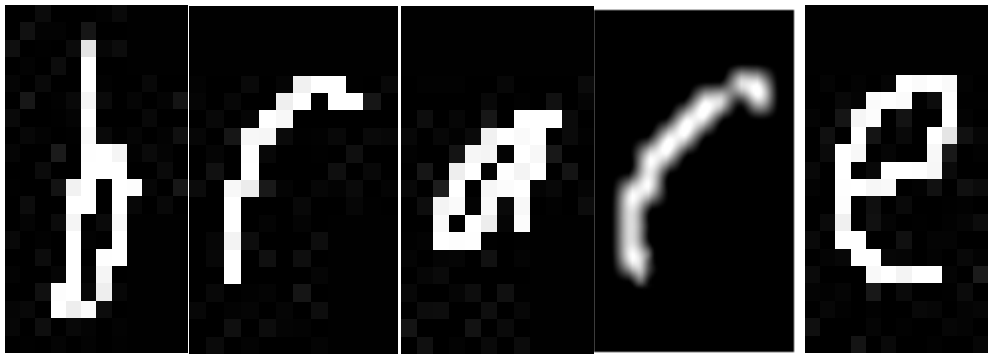


- thèmes de recherche:
  - prédiction structurée
  - optimisation
  - applications: vision, NLP, information retrieval, computational biology

parenthèse:
prédiction structurée

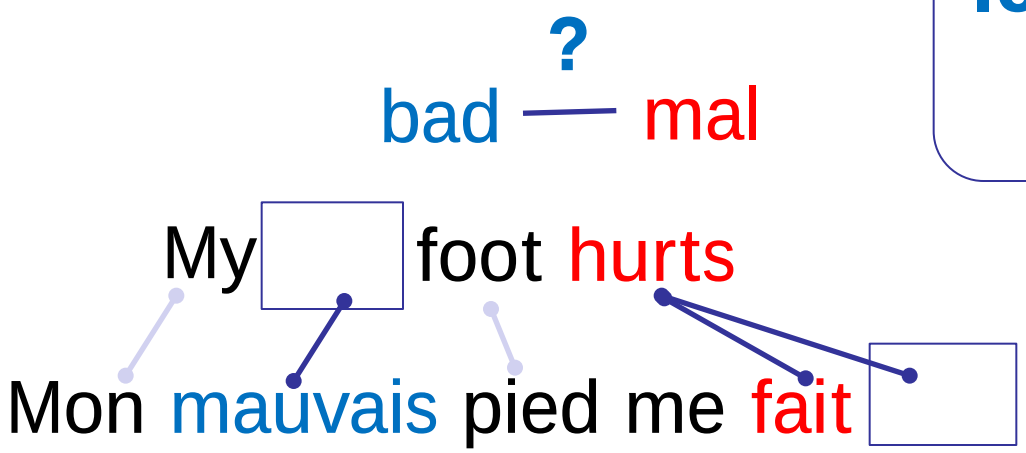
1) Exploiter structure: motivation

- reconnaissance de mots



brace

- alignement de mots



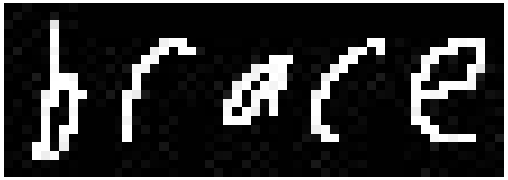
le contexte aide!

Prédiction structurée:

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

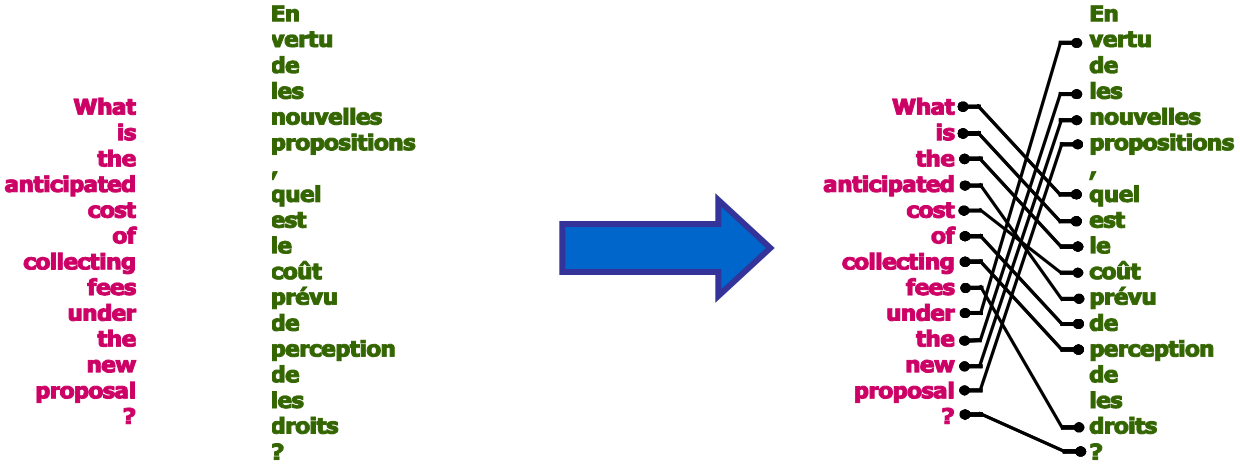
nombre exponentiel!

Reconnaissance de mots écrits



brace

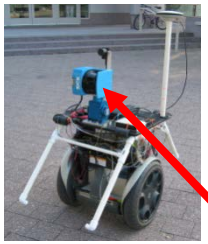
Alignement de mots



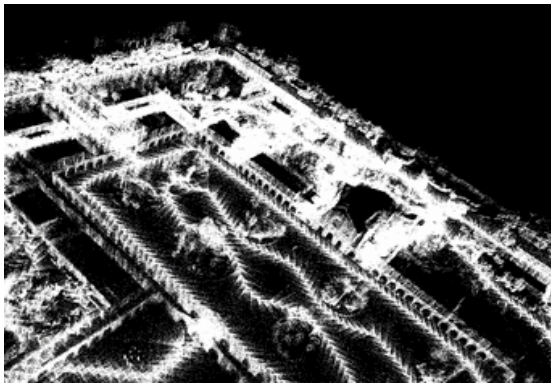
d'autres exemples...

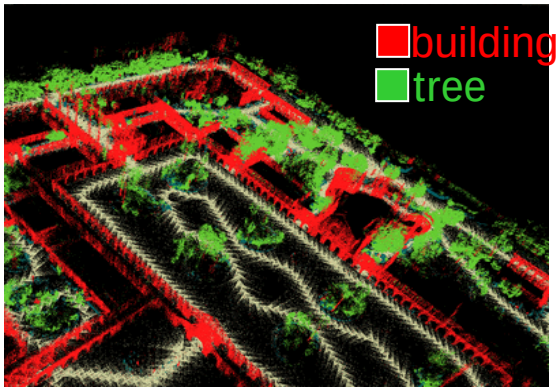
fonction d'erreur structurée: $\ell(\mathbf{y}, \mathbf{y}')$	Entrée $\mathbf{x} \in \mathcal{X}$	Sortie $\mathbf{y} \in \mathcal{Y}$
Traduction automatique	'Ce n'est pas un autre problème de classification.'	'This is not just another classification problem.'

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-12h -> avec système de réservation
- remise du projet mardi 29 mars
- toutes les infos sur site web:

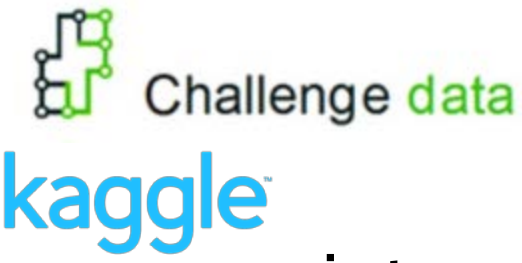
http://www.di.ens.fr/~slacoste/teaching/projet-MASH-2015/

Plan de cours + jalons

- cours: lundi 9 & 23 nov. 10-12h: TD python / scikit-learn par Fajwel
- lundi 30 novembre:
  - jalon: avoir formé son équipe + choisi son projet + envoie par mail 1 page résumé de projet
  - OH obligatoire: rendez-vous avec Fajwel
- [OH optionnelle commence lundi 14 décembre et continue le 4 janvier et continue chaque semaine (avec exceptions)]
- lundi 18 janvier & 15 février:
  - JALONS: présentation de 10 minutes devant la classe pour chaque équipe sur l'état du projet (problème, approche, résultats)
- **mardi 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

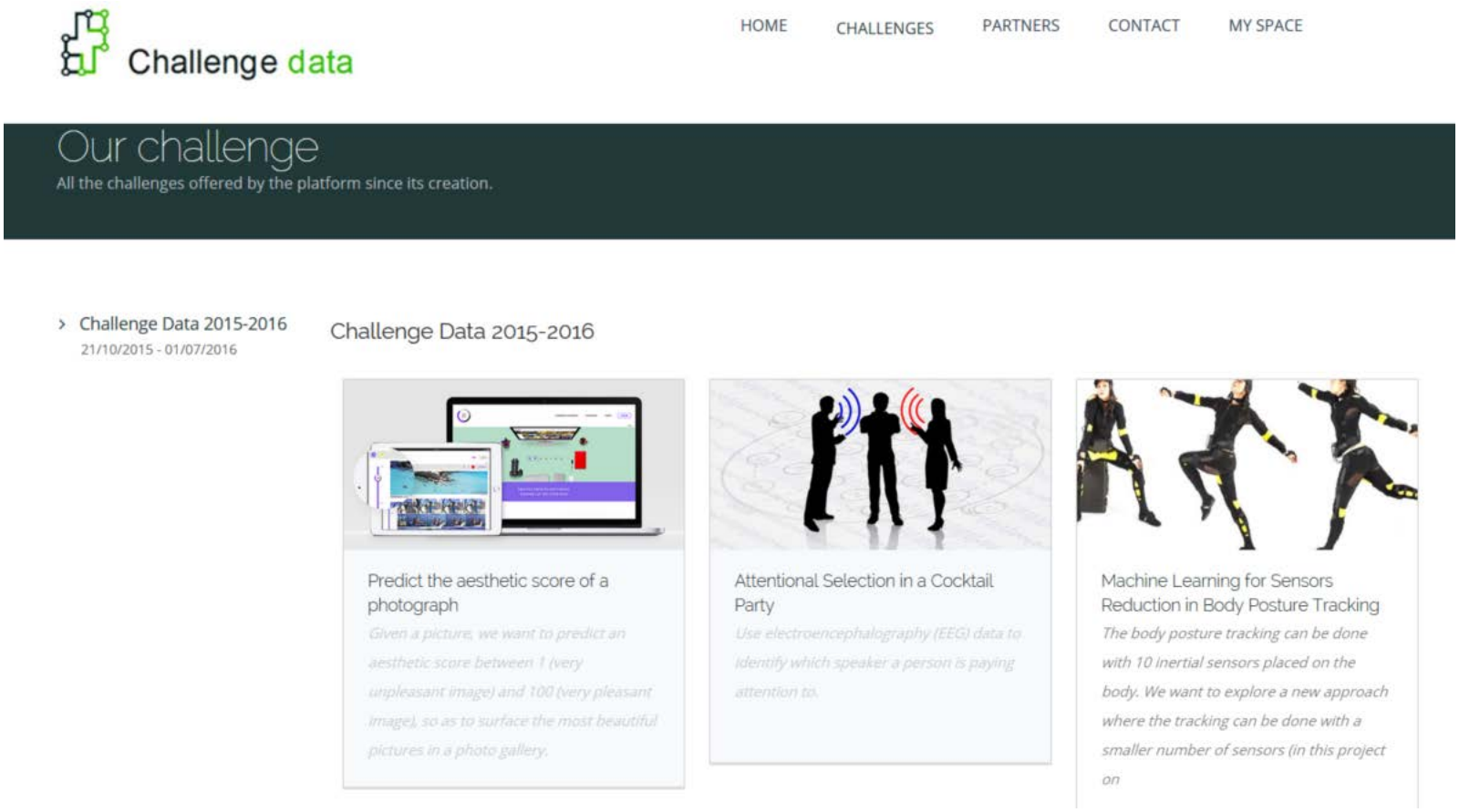
Choix de projet

- suggestions de projet sera disponibles sur le site web pour le 16 novembre
  - principalement Challenge Data
  - aussi possibilité kaggle



- 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...

challengedata.ens.fr



- clôture académique: fin février
- clôture challenge: 30 juin

Exemples de challenges:

Créateur du challenge	Titre du projet	Type de données	Type de problème
Dreem	Classification des états du sommeil	Série temporelle (EEG)	Classification
Sonoscanner	Détéction de bounding box en échographie obstétrique	Images	Regression
Cardiologs	Détection d'inversion d'electrodes	Série temporelle (ECG)	Classification
Shihab Lab	Classification d'EEG du cortex auditif	Série temporelle (EEG)	Classification
Criteo	Prédiction de clicks publicitaires	Données non structurées	Classification
Regaind	Prédiction de la qualité d'une photo	Images	Régression
Quantmetry / SNCF	Prédiction de pannes sur les trains SNCF	Données non structurées	Classification
Plume Labs	Prédiction de la pollution atmosphérique	Série temporelle	Régression
Capital Fund Management	Prediction du volume des transactions financières	Série temporelle	Régression
Université de Toulon	Projet Bird : classification de chants d'oiseau	Audio	Classification
Dassault Systèmes 3DS	Sensor reduction	Série temporelle	Regression
Reminiz	Reconnaissance de visages dans les films	Images	Classification
Oze-Energies	Prédiction de la consommation énergétique dans un immeuble	Série temporelle	Régression

** séances de présentation des challenges les vendredi 13, 20 et 27 novembre en amphi Rataud à l'ENS entre 11h30 et 13h30 (optionnelles)

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...

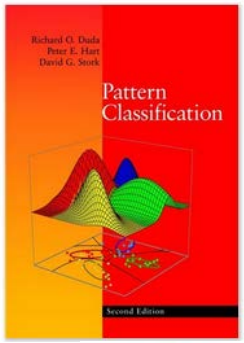
(cours les lundi 9 et 23 novembre)

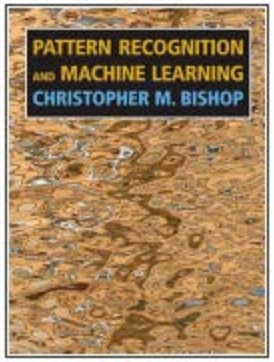
Suggestion approche pour projet

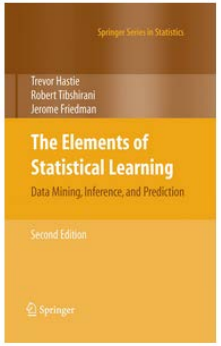
- Commencer avec méthodes / modèles simples
- Étudier où ça brise!
- Modifier features / méthode / modèle en conséquence
- Répéter!

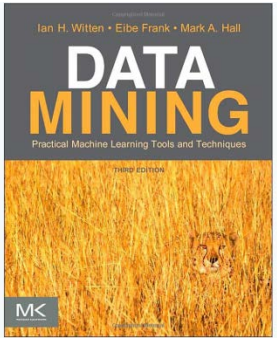
Quelques ressources

- 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 le 30 novembre) + commencer à réfléchir à projets (voir site web)

- Avant cours du 9 novembre:
  - installer Anaconda Python
  - faire tutoriel Python / scikit-learn (info par mail)