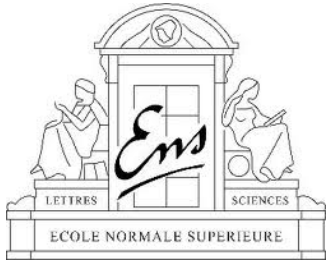
Intro: Apprentissage Automatique et Big Data

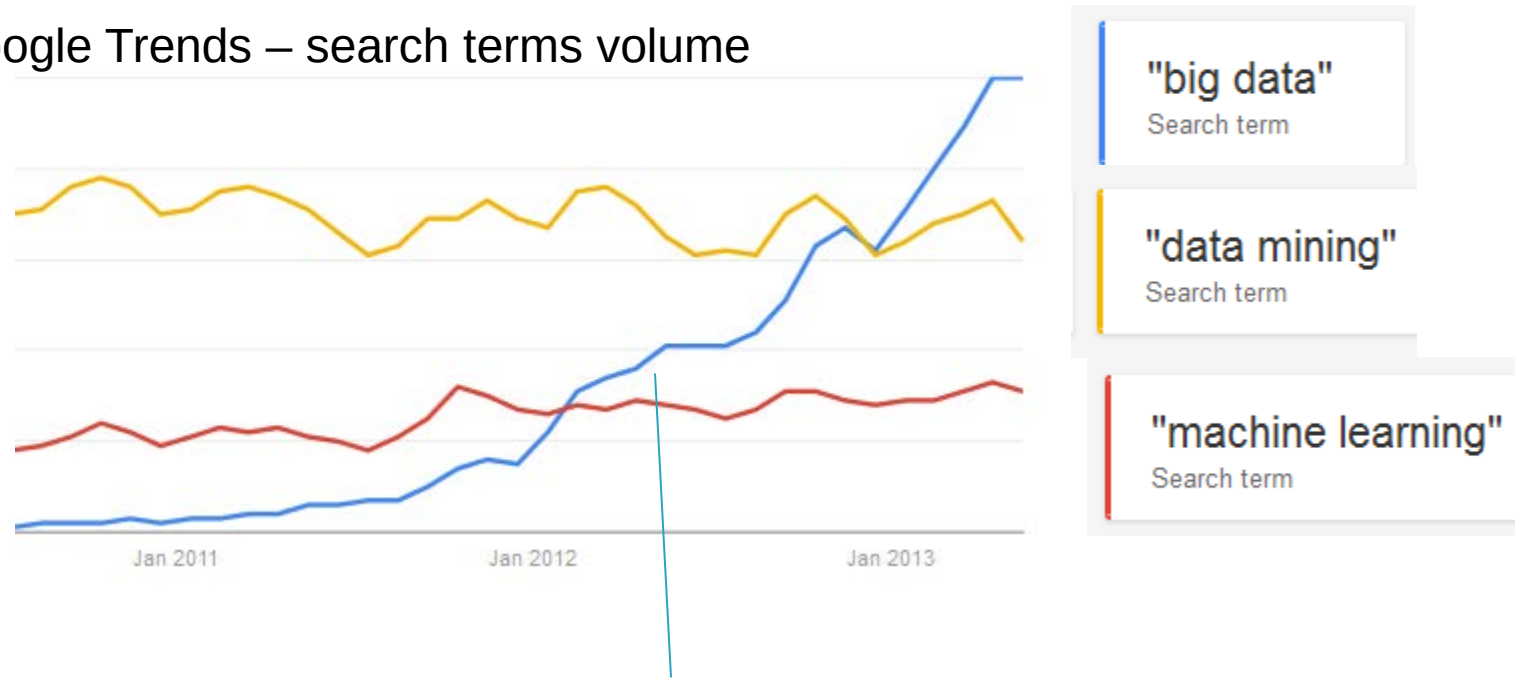
Simon Lacoste-Julien
Chercheur CR
Équipe-Projet SIERRA, INRIA - École Normale Supérieure



Qu'est-ce que le Big Data?

- Mot tendance pour décrire *beaucoup* de données!
- Buzz:

Google Trends – search terms volume



Qu'est-ce que le Big Data?

- Mot tendance pour décrire *beaucoup* de données!
- nous vivons à l'**ère de l'information**
  - accumulation de données dans tous les domaines:
    - internet
    - biologie: génome humain, séquençage d'ADN
    - physique: Large Hadron Collider, 10^{20} octets/jour par senseurs
  - appareils d'enregistrement:
    - senseurs, portables, interactions sur internet, ...

- défis en informatique:
  - stockage, recouvrement, calcul distribué...
  - 3V's: volume, vélocité, variété
- donner un sens aux données: **apprentissage automatique**

Donner du sens au (Big) Data

- Nous voulons utiliser les données pour:
  - faire des prédictions, détecter des failles, résoudre des problèmes...
- Science derrière tout cela:
  - apprentissage automatique / statistiques computationnelles
- Autres termes en pratique:
  - data mining, business analytics, pattern recognition, ...

Qu'est-ce que l'apprentissage automatique?

- Question centrale selon Tom Mitchell:

  "Comment **construire des systèmes informatiques** qui **s'améliorent avec l'expérience**, et quelles sont les **lois** fondamentales qui gouvernent **tous les processus d'apprentissage automatique**?"

- Mélange **d'informatique** et de **statistiques**

  CS: "Comment construire des machines qui résolvent des problèmes, et quels problèmes sont intrinsèquement faisables / infaisables?"

  Statistiques: "Que peut-il être déduit à partir de données et un ensemble d'hypothèses de modélisation?

  -> comment un ordinateur peut-il ***apprendre*** à partir de données?

Apprentissage statistique

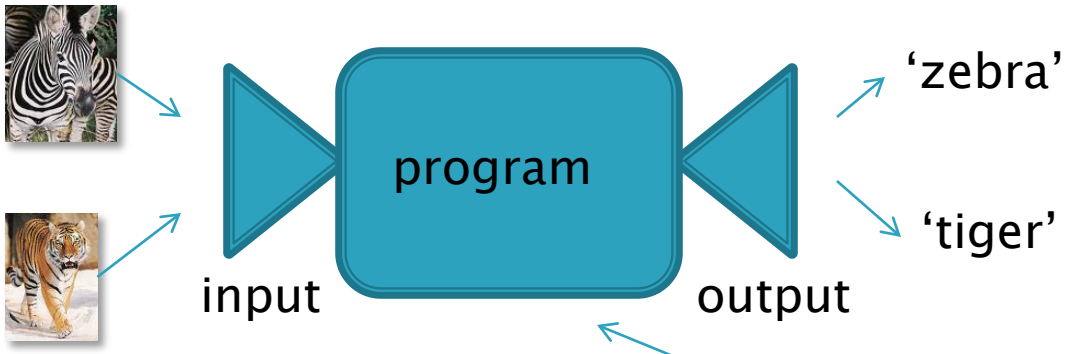
- informatique + statistique / math. appliquées

vs statistiques traditionnelles:

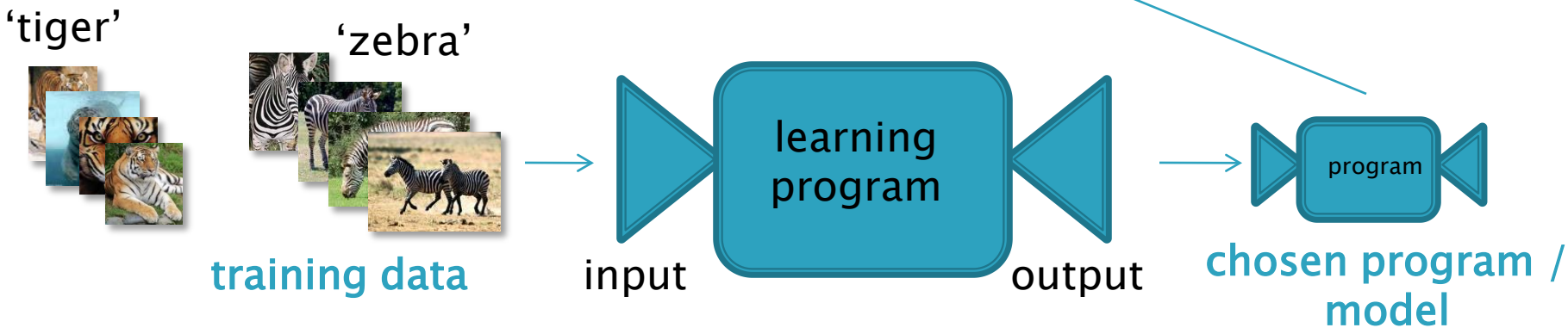
- analyse de données en **grande dimension**
  - modèles **complexes** / **structurées**
- sensible aussi à l'efficacité des algorithmes (aspect computationnel)

Intro à l'apprentissage automatique

- Traditional programming:



- Machine learning:

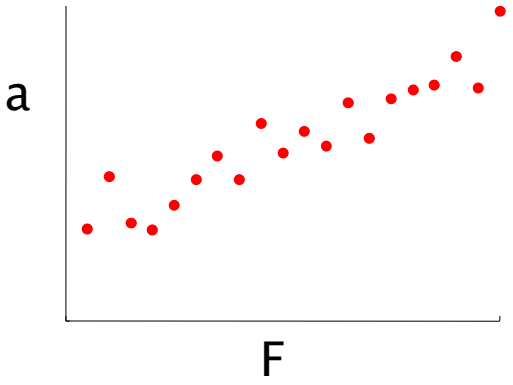


Exemple simple: régression linéaire

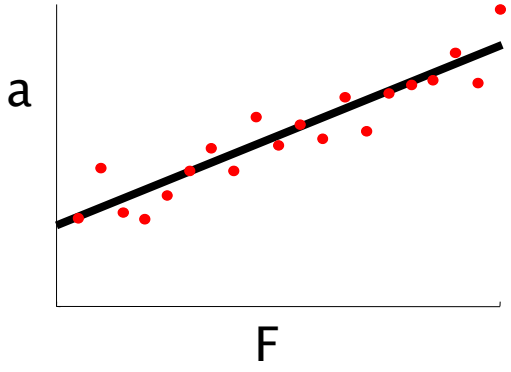
- learn a predictive model

possible models

$a = w_0 + w_1 * F$



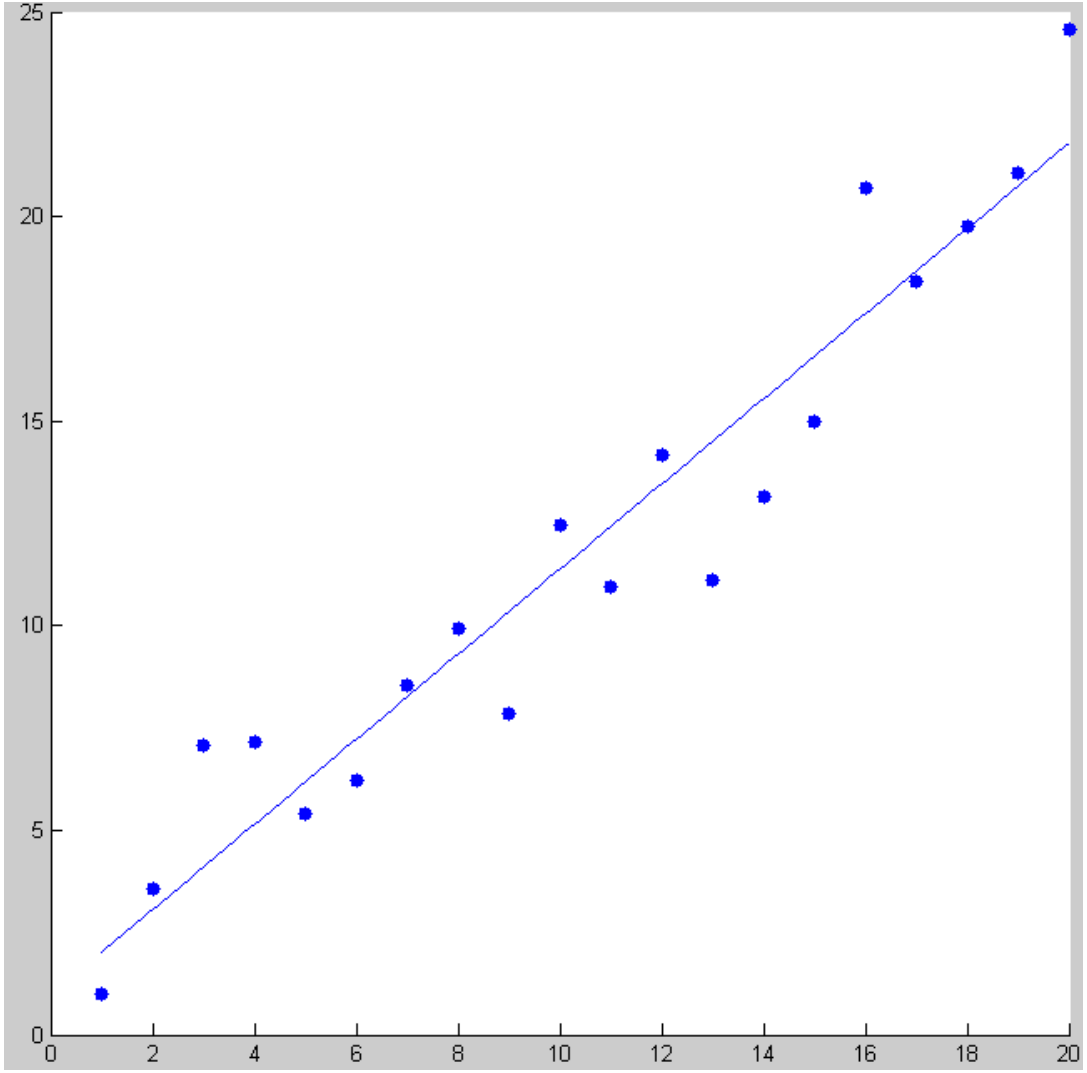
training data



Choose w_0, w_1 to minimize sum of squared errors

Learning law #1:
Occam's razor and overfitting

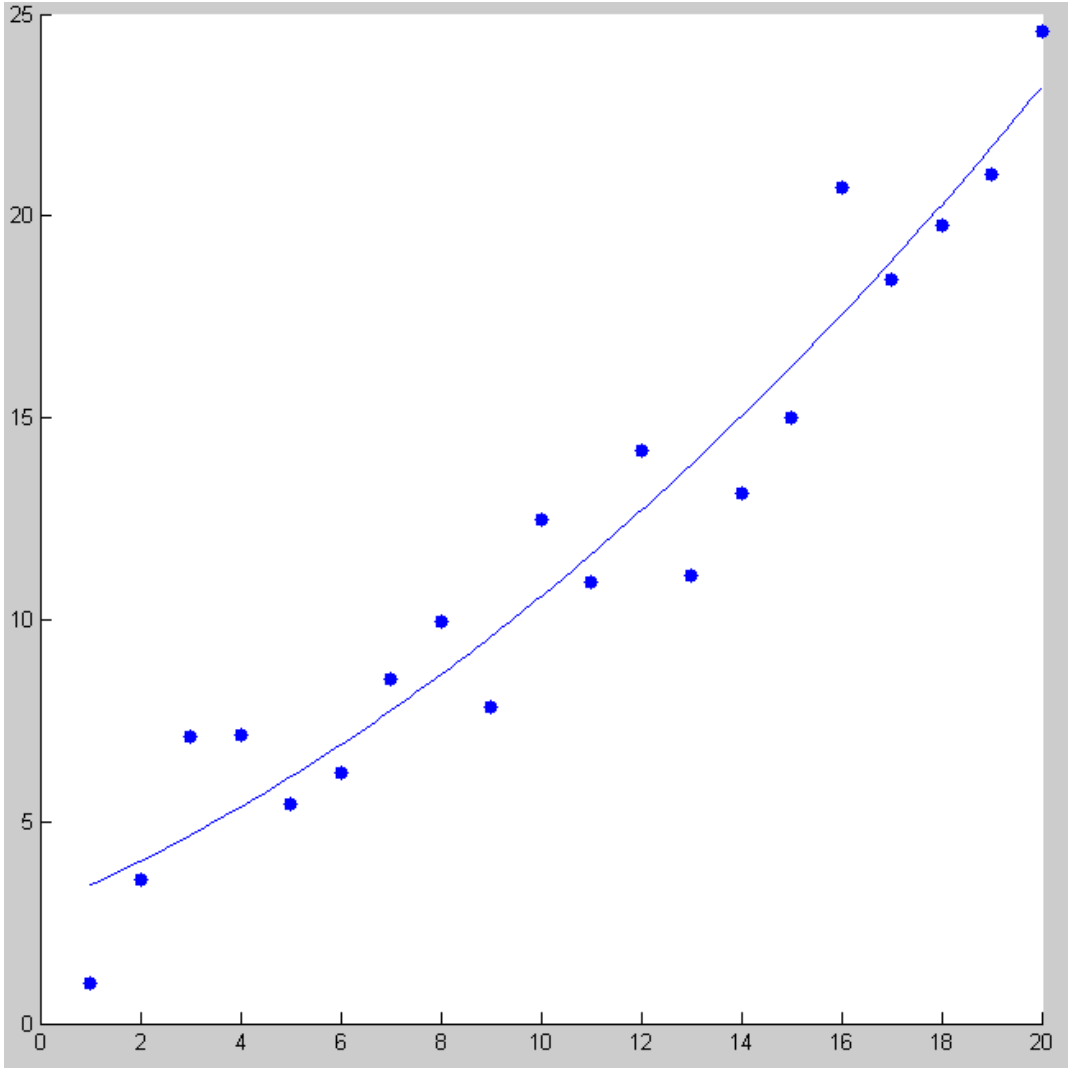
Overfitting in regression...



linear model:

$a = w_0 + w_1 * F$

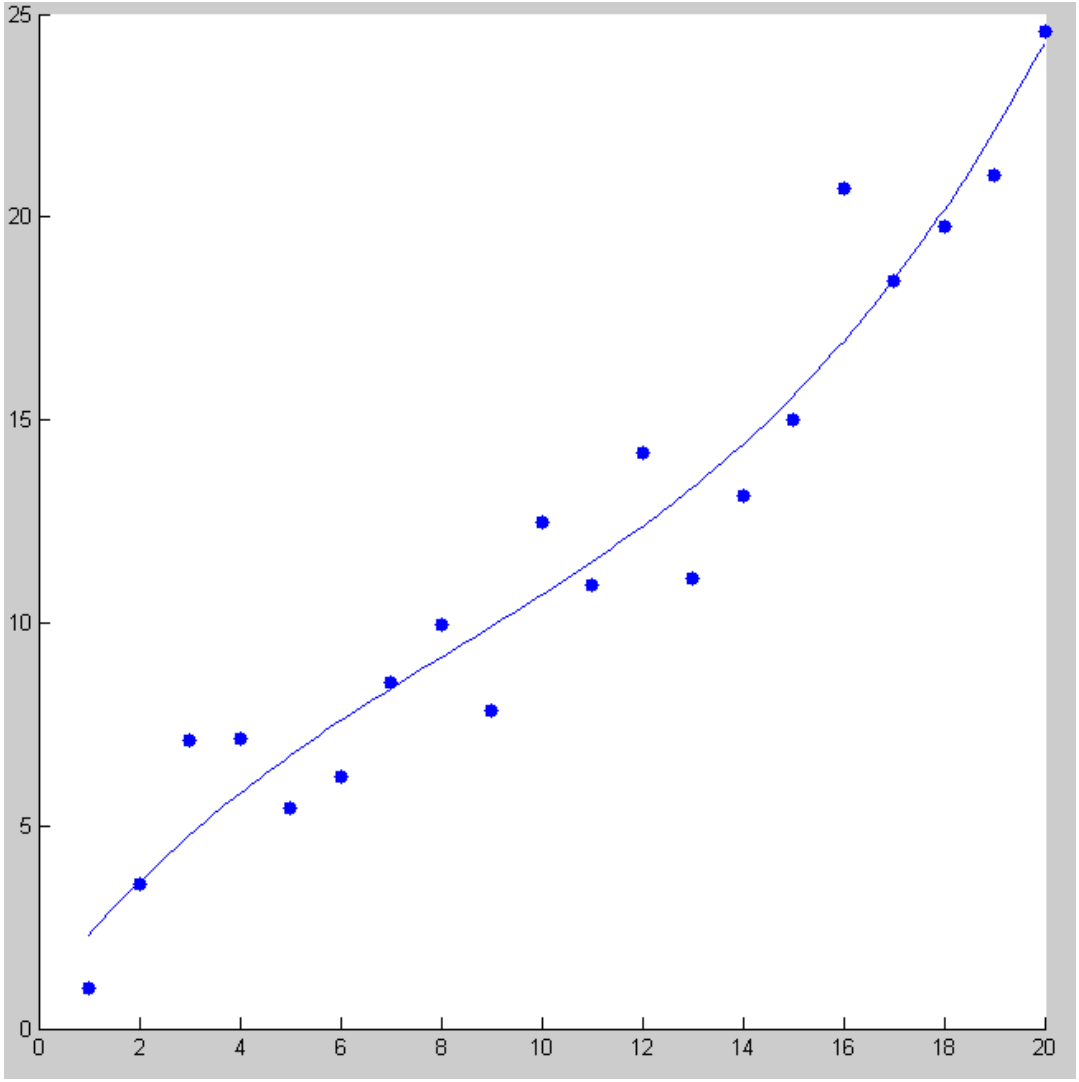
Overfitting in regression...



quadratic model:

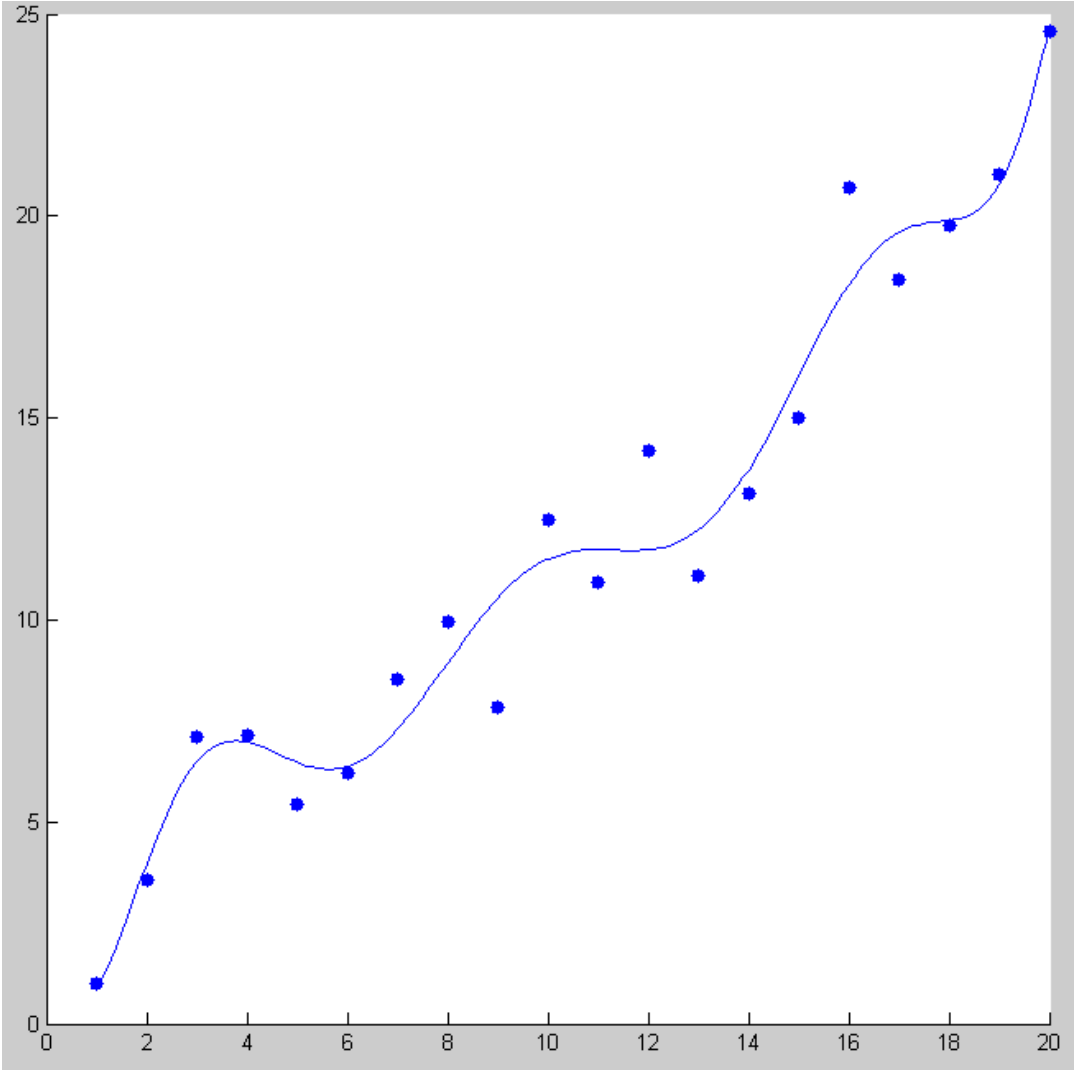
$a = w_0 + w_1 * F + w_2 * F^2$

Overfitting in regression...



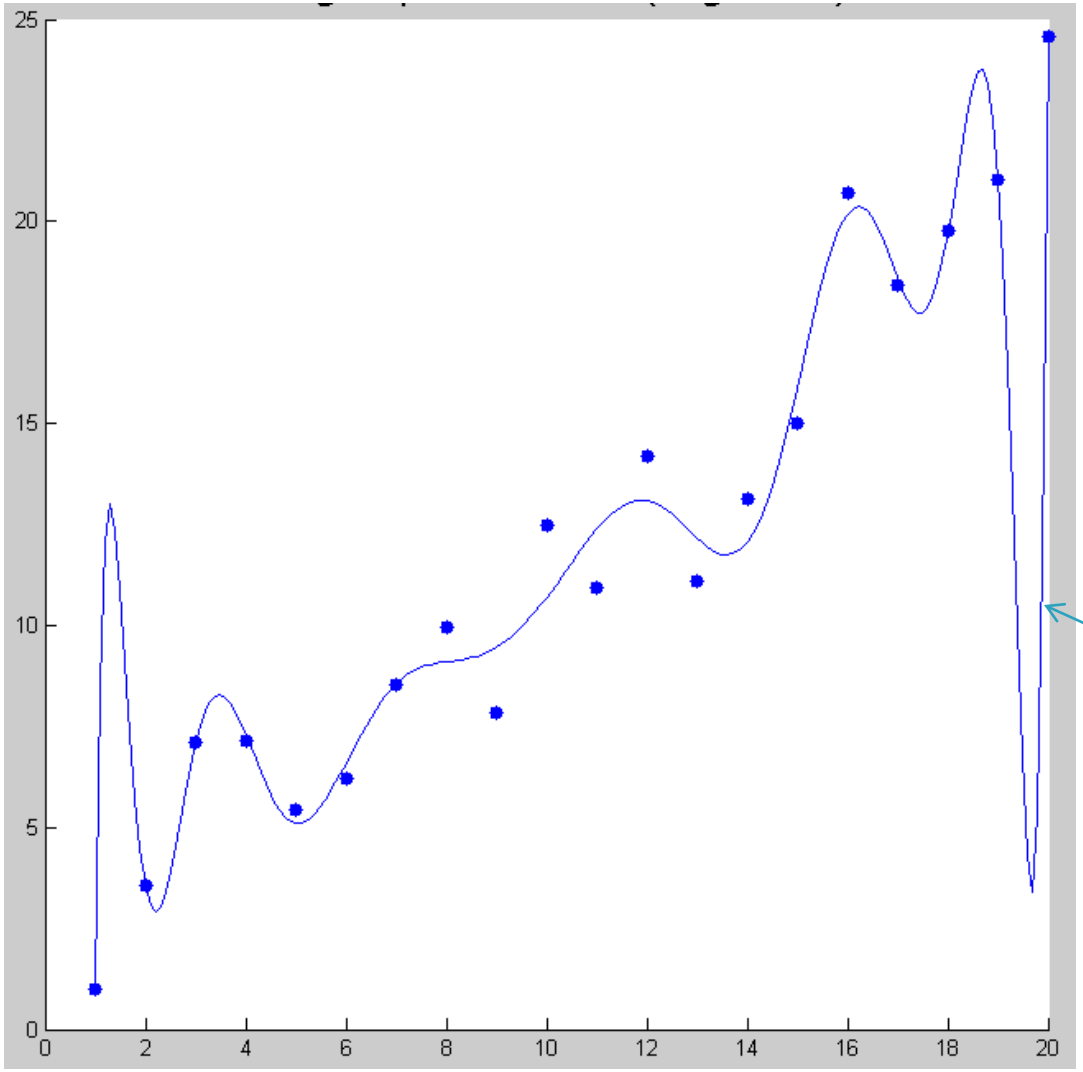
cubic model
(degree 3)

Overfitting in regression...



degree 10

Overfitting in regression...



degree 15

overfitting!

Occam's razor principle:

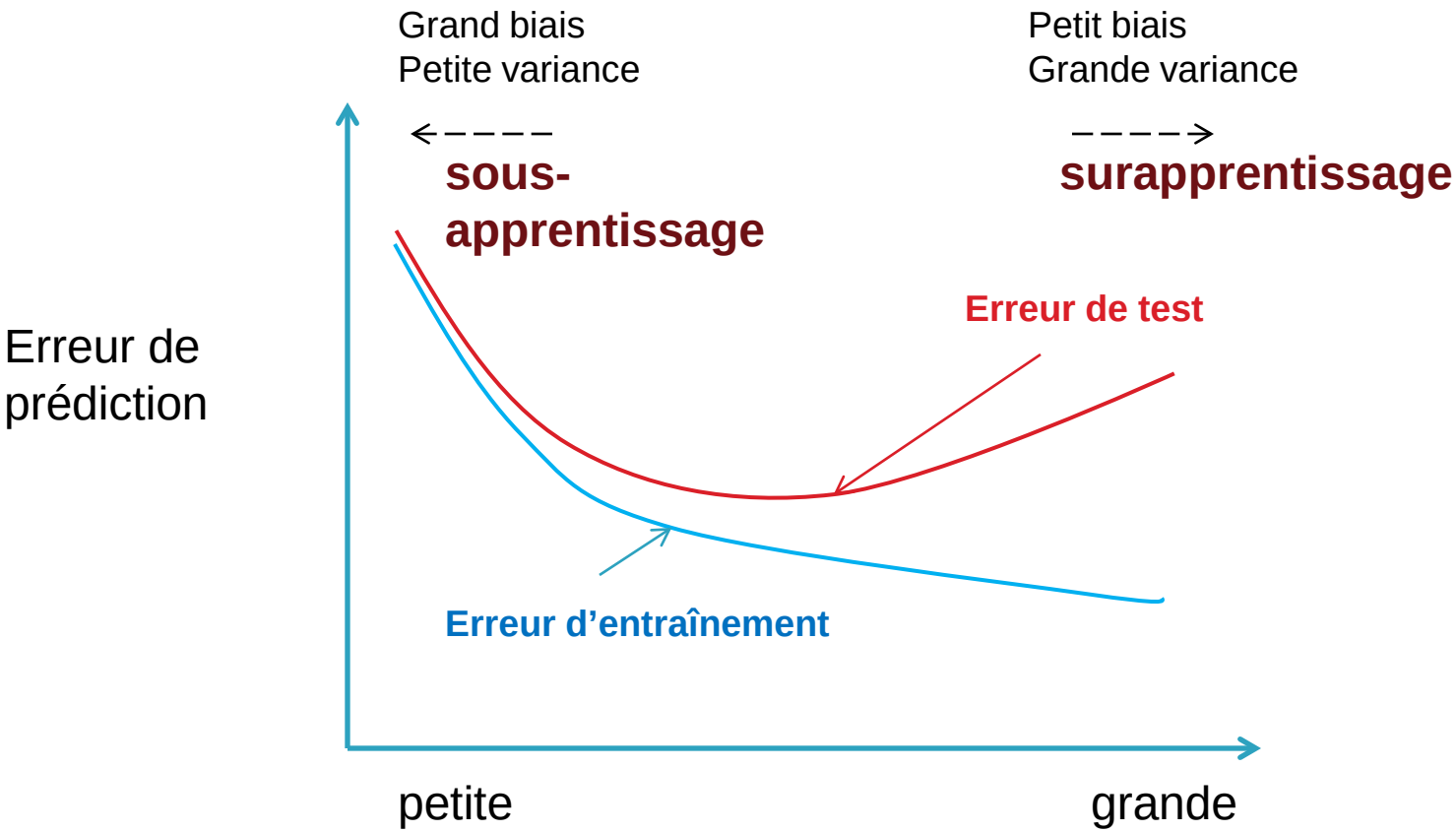
- Between two models / hypotheses which explain as well the data, choose the **simplest** one

- In Machine Learning:
  - we usually need to tradeoff between
    - training error
    - model complexity

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \underbrace{\hat{\mathbb{E}}\left[\ell(\mathbf{y}, h(\mathbf{x}))\right]}_{\text{empirical error}} + \underbrace{\Omega(h)}_{\text{regularizer}}$$

  - can be formalized precisely in statistics (bias-variance tradeoff, etc.)

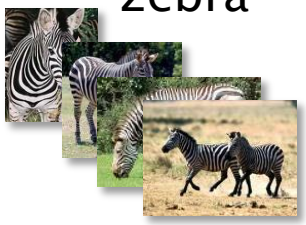
Rasoir d'Occam



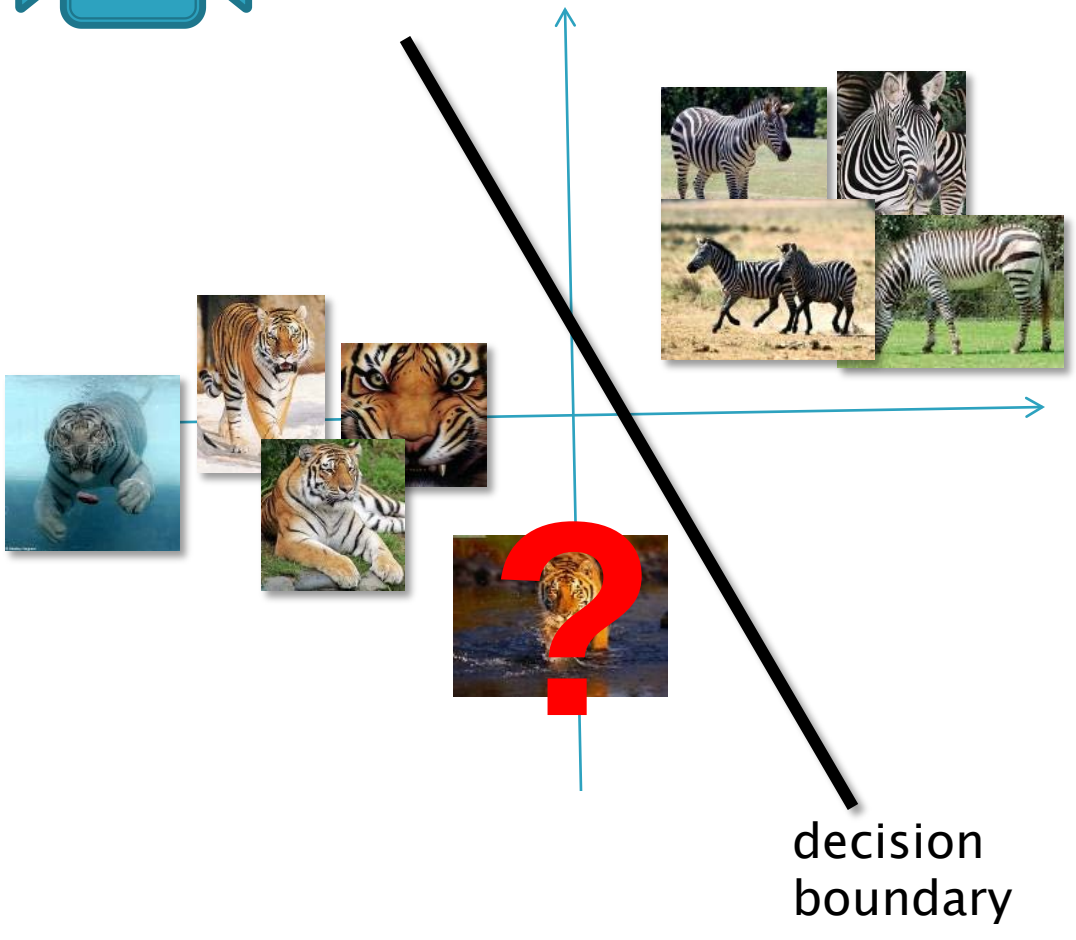
Classification

‘tiger’

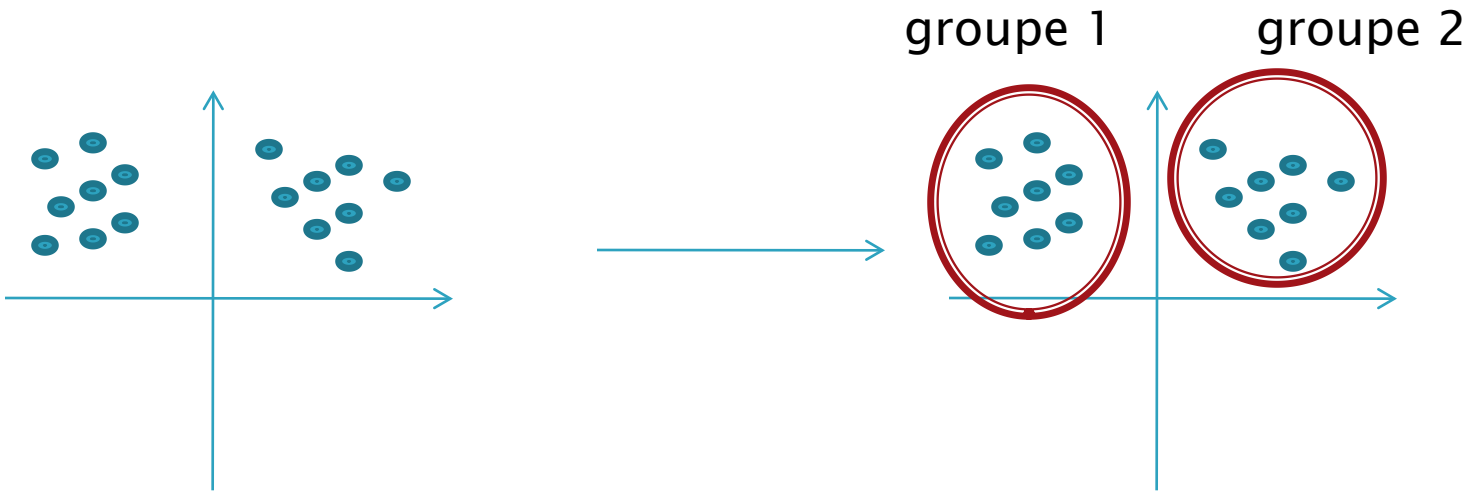
‘zebra’



training data



Regroupement (clustering)



some warnings...

Pitfalls for Big Data hype

- Hype: With enough data, we can solve "everything" with "no assumptions"!

- Theory: **No Free Lunch Theorem!**
  - If we do not make assumptions about the data, **all** learning methods do **as bad *"on average"*** on unseen data as a **random prediction**!

- consequence: need **some assumptions**
  - for example, that time series vary 'smoothly'

Fléau de la dimension

- Problème avec données en grandes dimensions: explosion combinatoire de possibilités (**exponentiel** en d)
- Exemple: classification d'images
  - entrées: 16x16 pixels binaires ($d=16^2=256$)
  - sortie: {-1,1} [2 classes]
  - nombre d'entrées possible: $2^{256} \sim 10^{77}$

    vs. nombre d'images sur Facebook: $\sim 10^{12}$
- => impossible d'apprendre la fonction de classification sans supposition!

Pitfall 2 – mining random patterns

- We can 'discover' meaningless random patterns if we look through too many possibilities
  - "Bonferroni's principle"; exemplified by Birthday Paradox
- **NSA example**: say we consider **suspicious** when a pair of (unrelated) people **stayed at least twice in the same hotel on the same day**
  - suppose 10^9 people tracked during 1000 days
  - each person stays in a hotel 1% of the time (1 day out of 100)
  - each hotel holds 100 people (so need 10^5 hotels)

  -> if everyone behaves **randomly** (i.e. no terrorist), can we still detect something suspicious?

  Probability that a **specific** pair of people visit same hotel on same day is 10^{-9}; probability this happens twice is thus 10^{-18} (tiny),

  ... but there are many possible pairs

  => **Expected number of "suspicious" pairs is actually about 250,000!**

example taken from Rajamaran et al., Mining of Massive Datasets

Morale de l'histoire...

- Il faut bien connaître ses **statistiques** en plus de l'informatique pour faire du sens du Big Data!
  - -> métier de « Data-Scientifique »

Harvard Business Review

THE MAGAZINE | BLOGS | VIDEO | BOOKS | CASES | WEBINARS | CO

Guest

Register today and save 20% off your first order! Details

THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article: Sign in or Register for free. HBR Subscribers activate your free archive access »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (87)

RELATED

Executive Summary

ALSO AVAILABLE

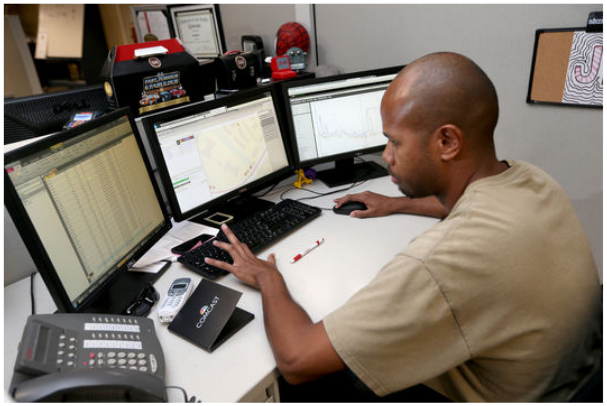
- Buy PDF

http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1

Un métier « sexy » ? Datascientifique !

LE MONDE | 08.04.2014 à 21h10 • Mis à jour le 09.04.2014 à 11h03 |

Par Maryline Baumard



Vous connaissez le métier le plus « sexy » du moment ? La très sérieuse *Harvard Business Review* ose ce qualificatif pour les *« data scientists »*, ces « scientifiques des données ». Si l'article paru fin 2012 a fait grand bruit, la revue n'a rien inventé, juste donné un bel écho à l'idée lancée par Hal Varian. Le chef économiste de Google, professeur à Berkeley, en Californie, avait déclaré que *« le métier le plus sexy du moment* [était celui de] *statisticien »*. Il ne parlait évidemment pas du statisticien lambda qui se bagarre avec deux colonnes de chiffres, mais du « datascientifique ».

http://abonnes.lemonde.fr/economie/article/2014/04/08/un-metier-sexy-datascientifique_4397951_3234.html?xtmc=data_scientist&xtcr=1

opportunities...

Some success stories using machine learning

- spam classification (Google)
- machine translation (not pretty, but ‘functional’)
- speech recognition (used in your smart phone)
- self-driving cars (again Google)

- Demo from Microsoft Chief Research Officer at Microsoft Research Asia meeting in Oct 2012

Microsoft_demo_cut.mp4

http://phys.org/news/2012-11-microsoft-applause-tone-preserving-video.html

Résumé

- ‘révolution’ du Big Data:

  disponibilités de données

  \+

  avancées dans les outils computationnels et statistiques

  = opportunités pour résoudre de nouveaux problèmes!

- apprentissage automatique – domaine en pleine croissance...
  - par contre domaine extrêmement multidisciplinaire: combine informatique, maths appliquées, statistiques
- `success stories’ dans les domaines des sciences et technologies

Statistics vs. Machine Learning

- from Larry Wasserman's blog:

http://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/

Statistics	Machine Learning
Estimation	Learning
Classifier	Hypothesis
Data point	Example/Instance
Regression	Supervised Learning
Classification	Supervised Learning
Covariate	Feature
Response	Label

and of course:
Statisticians use R.
Machine Learners use Matlab.

Cours M1: Apprentissage Statistique

http://www.di.ens.fr/~slacoste/teaching/apprentissage-fall2014/

Vendredi 8h30-12h30 - Salle Henri Cartan
premier cours: 26 sept.

co-enseigné par:

Simon Lacoste-Julien

Francis Bach

chargé de TD:

Rémi Lajugie

Équipe-Projet SIERRA, INRIA / ENS

Liens avec d'autres disciplines

Math:

- Statistiques et théorie de l'information
- Optimisation et analyse convexe
- Mais aussi:
  - Théorie spectrale des opérateurs
  - Transformée de Fourier (traitement du signal)
  - Géométrie différentielle et riemannienne

Info:

- Algorithmique (e.g. programmation dynamique)
- Programmation

Domaines appliqués:

- Vision par ordinateur
- Biologie Computationnelle
- Traitement du Langage Naturel
- Robotique
- Fouille de données

Pourquoi prendre ce cours?

- comme porte d'entrée pour le master MVA de l'ENS Cachan!
- pour comprendre la base de l'analyse de données de grande dimension
  - soit pour continuer en recherche en statistiques, traitement du signal, apprentissage, etc.
  - soit pour avoir la base théorique pour poursuivre en industrie (croissance des rôles de data scientists)
  - soit par curiosité! Concepts utilisés dans plusieurs domaines où traitent des données...

Logistique:

- 6 ECTS
- Note déterminée à 60% par l'examen et 40% par un TP à rendre
- Normalement:
  - cours magistral de 8h30 à 10h20
  - une pause d'environ 20 minutes
  - TD de 10h40 à 12h30 -> apportez votre portable!
- Nos mails de contact se trouvent sur nos sites webs!

Curriculum (prévisionnel)

26/09	Simon	2h	Introduction
	Simon	2h	Apprentissage supervisé
03/10	Simon	2h	Méthodes par moyennage local
	Rémi	2h	(TD) Apprentissage supervisé
10/10	Simon	2h	Validation croisée / sélection de modèles
	Rémi	2h	(TD) Méthodes par moyennage local
17/10	Francis	2h	Analyse convexe
	Rémi	2h	(TD) Analyse convexe
24/10	Francis	2h	Optimisation convexe
	Rémi	2h	(TD) Optimisation convexe
31/10	Simon	2h	Théorie, concentration et borne PAC-Bayes
	Rémi	2h	(TD) Théorie, concentration et borne PAC-Bayes
07/11	Simon	2h	Méthodes probabilitistes (maximum de vraisemblance)
	Rémi	2h	(TD) Méthodes probabilistes (maximum de vraisemblance)
14/11	Simon	2h	Régression linéaire / logistique
	Rémi	2h	(TD) Régression linéaire / logistique

21/11	Francis	2h	Méthode à noyaux (I)
	Francis	2h	Méthode à noyaux (II)
28/11	Simon	2h	Régularisation (Stein, analyse biais/variance)
	Rémi	2h	(TD) Régularisation (Stein, analyse biais/variance)
05/12	Francis	2h	Classification linéaire par pertes convexes
	Rémi	2h	(TD) Méthode à noyaux
09/01	Simon	2h	Prédiction structurée et applications
	Simon	2h	Résumé et questions / réponses
16/01		3h	EXAMEN