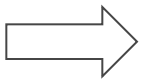
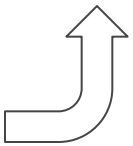
Apprentissage Automatique

Francis Bach
Chercheur INRIA
Équipe-Projet SIERRA, INRIA - École Normale Supérieure



Berkeley
University of California

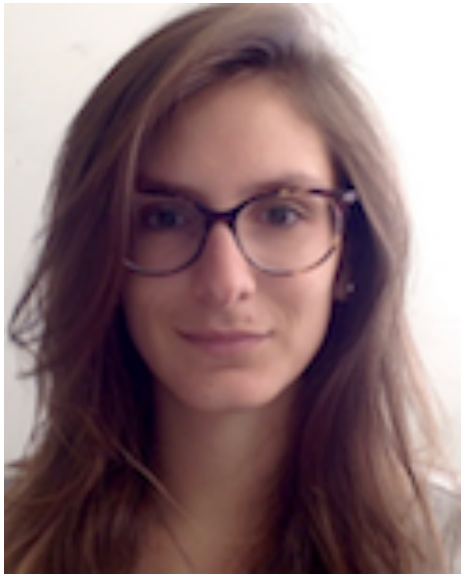


Apprentissage Automatique

Francis Bach
DI

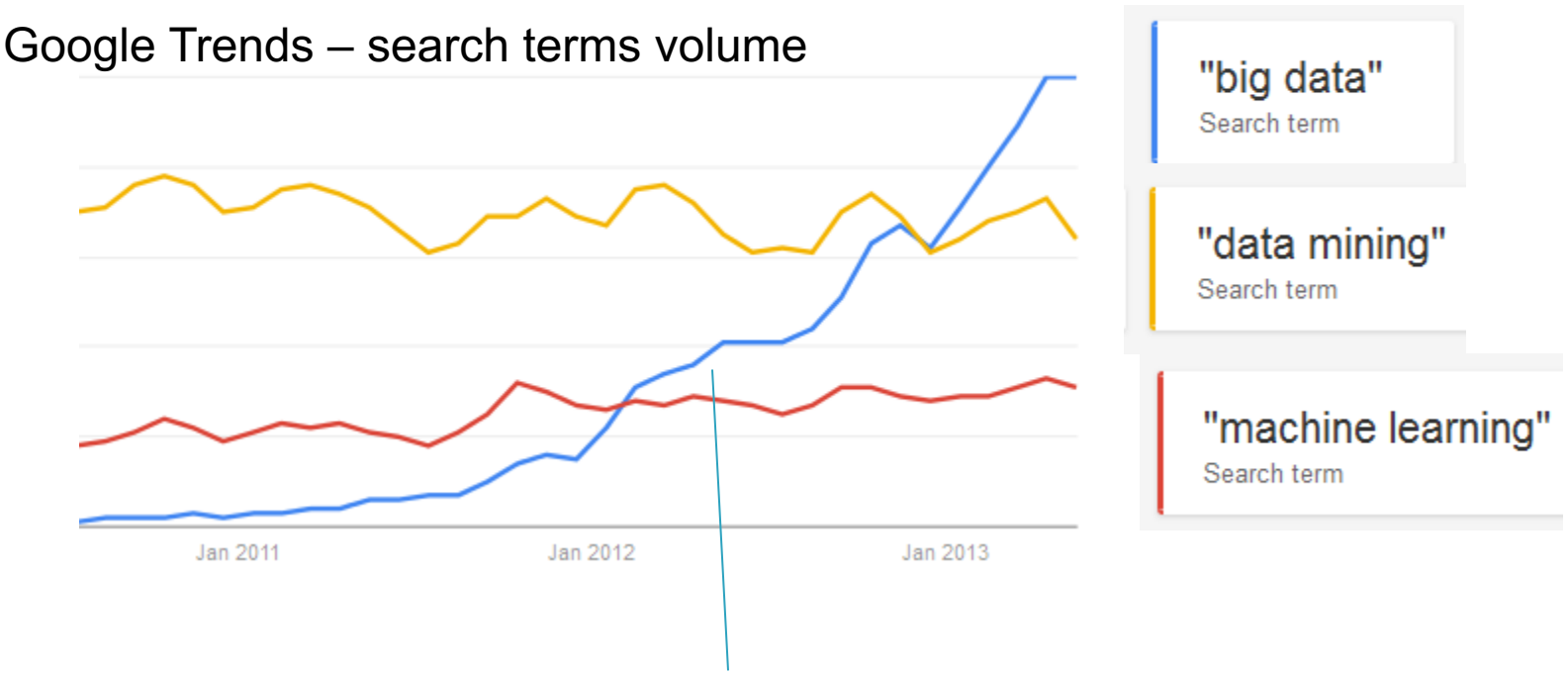
Pierre Gaillard
DI

Aude Genevay
DMA



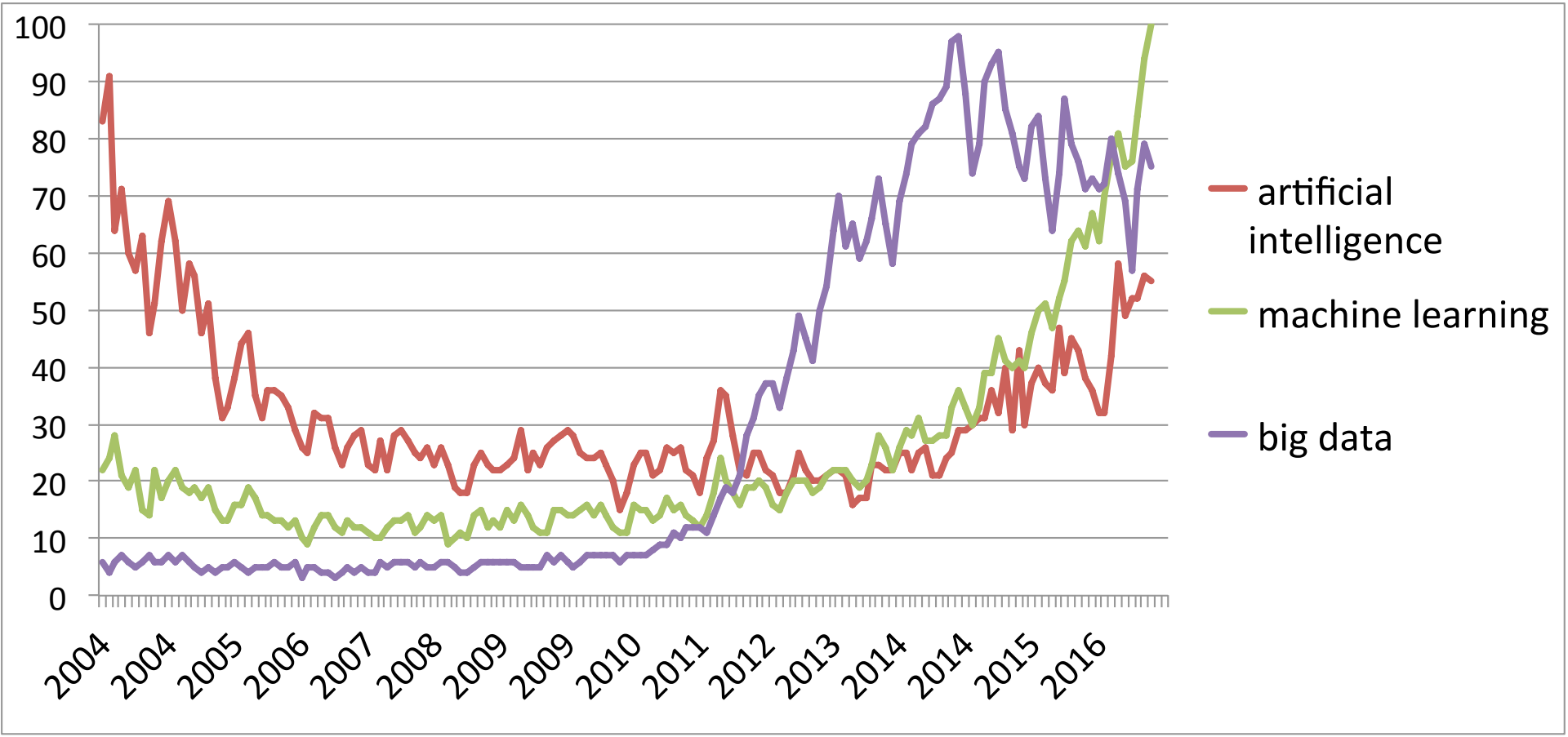
Qu'est-ce que le Big Data?

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



Maintenant on dit IA…

- Renouveau de l'intelligence artificielle



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, artificial intelligence ...

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

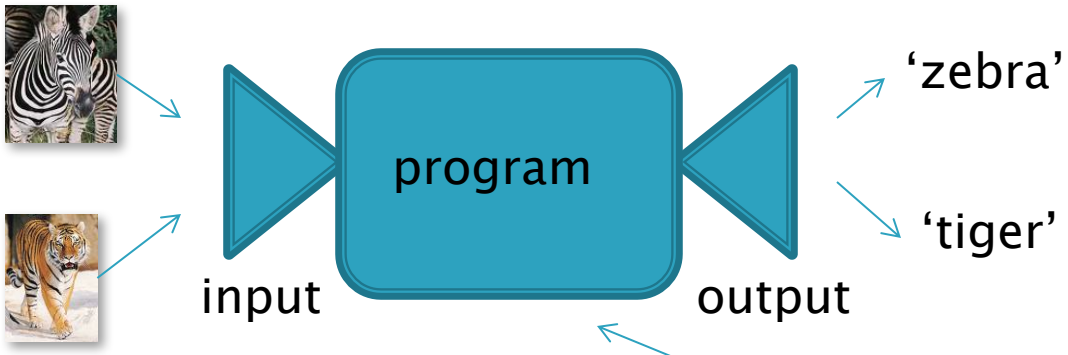
- informatique + statistique / math. appliquées

vs statistiques traditionnelles:

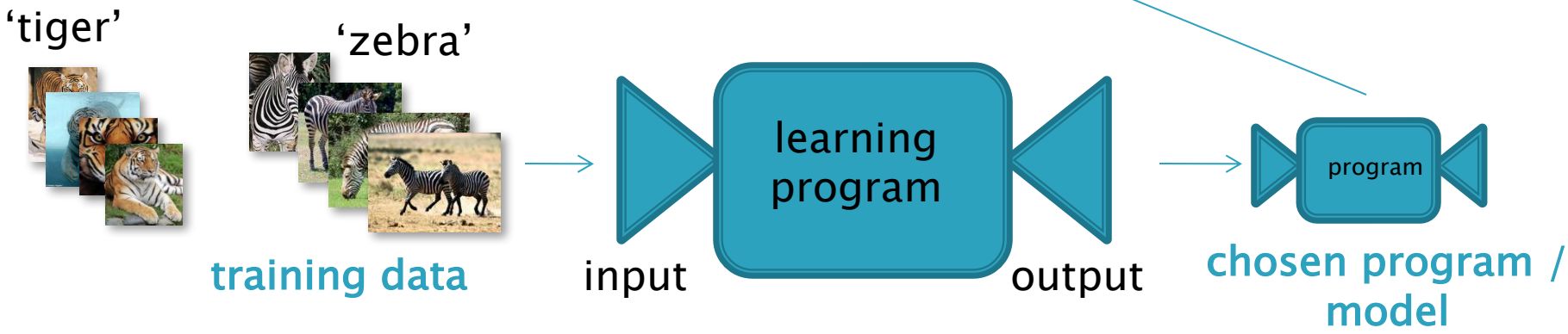
- analyse de données en **grande dimension**
  - modèles **complexes** / **structurées**
- sensible aussi à l'efficacité des algorithmes (aspect computationnel)

Intro à l'apprentissage automatique

- Traditional programming:



- Machine learning:

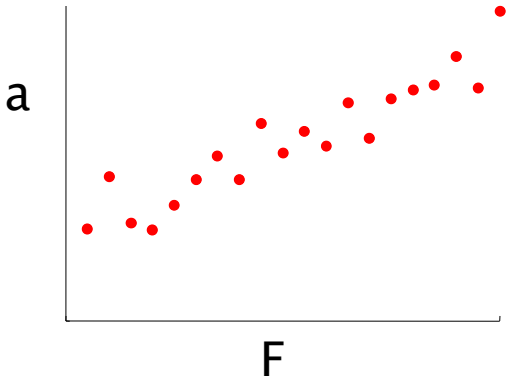


Exemple simple: régression linéaire

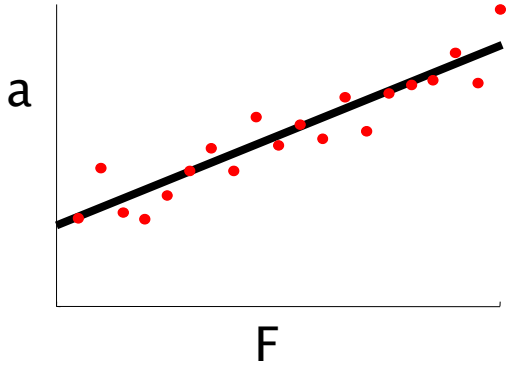
- learn a predictive model

possible models

$a = w_0 + w_1 * F$



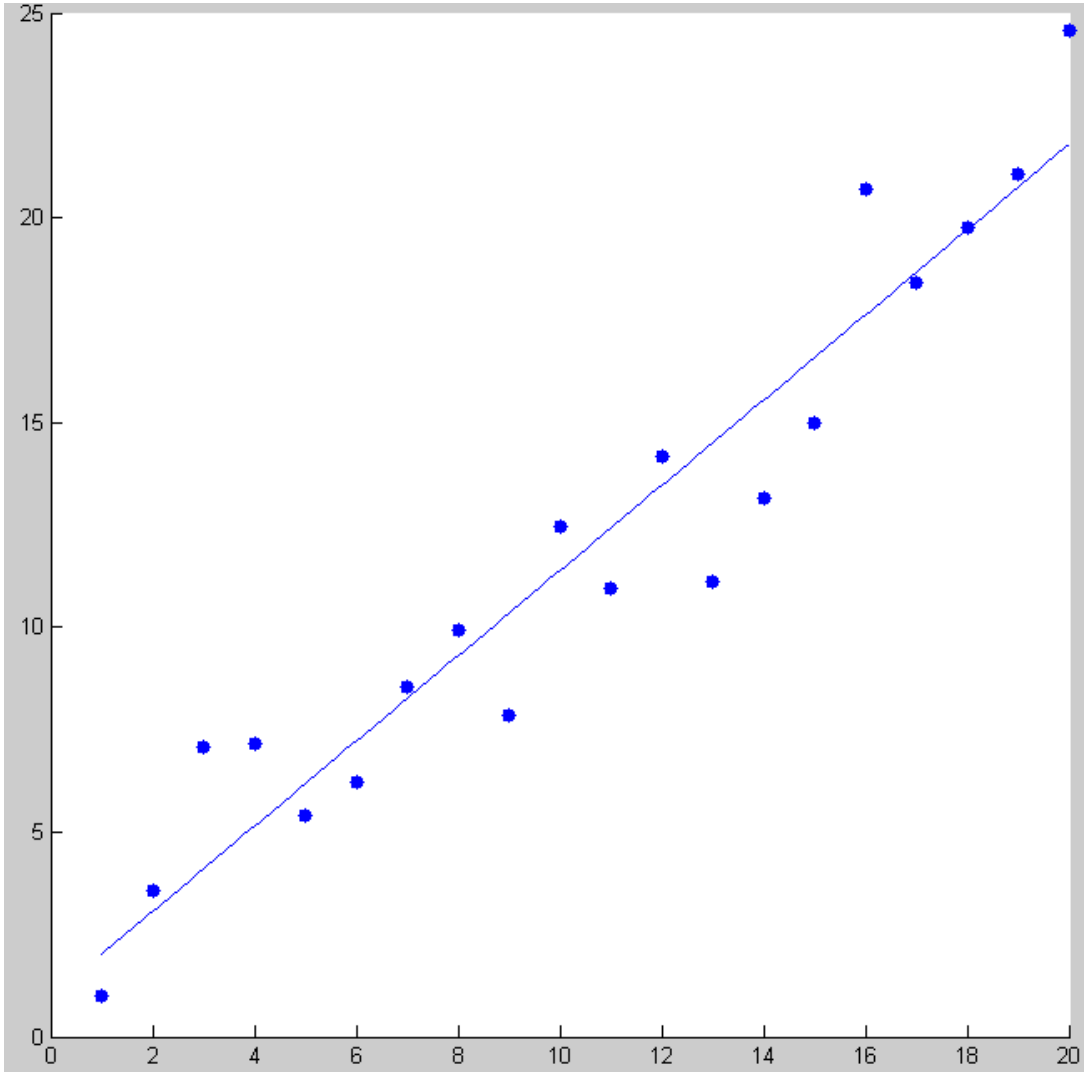
training data



Choose w_0, w_1 to minimize sum of squared errors

Learning law #1:
Occam's razor and overfitting

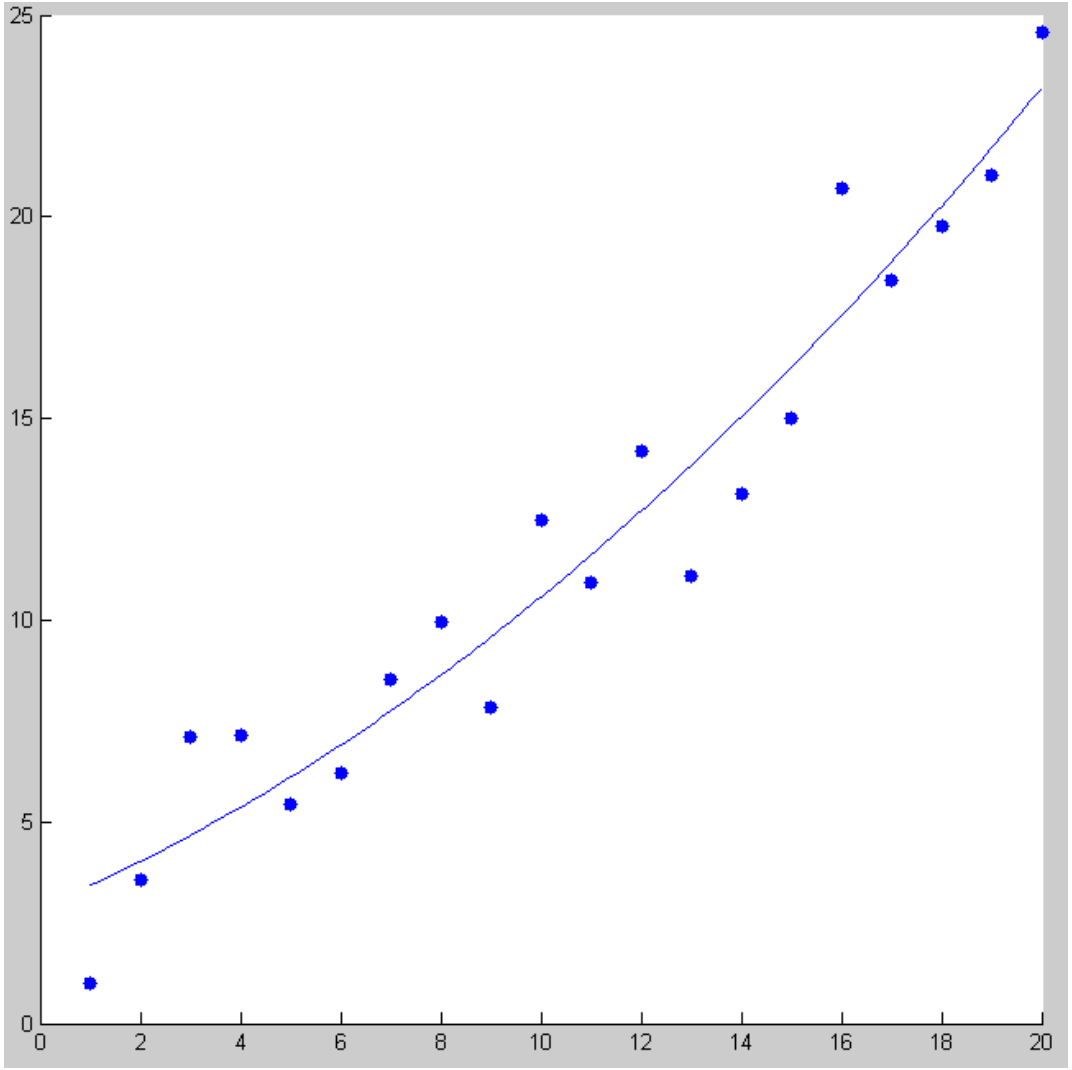
Overfitting in regression...



linear model:

$a = w_0 + w_1 * F$

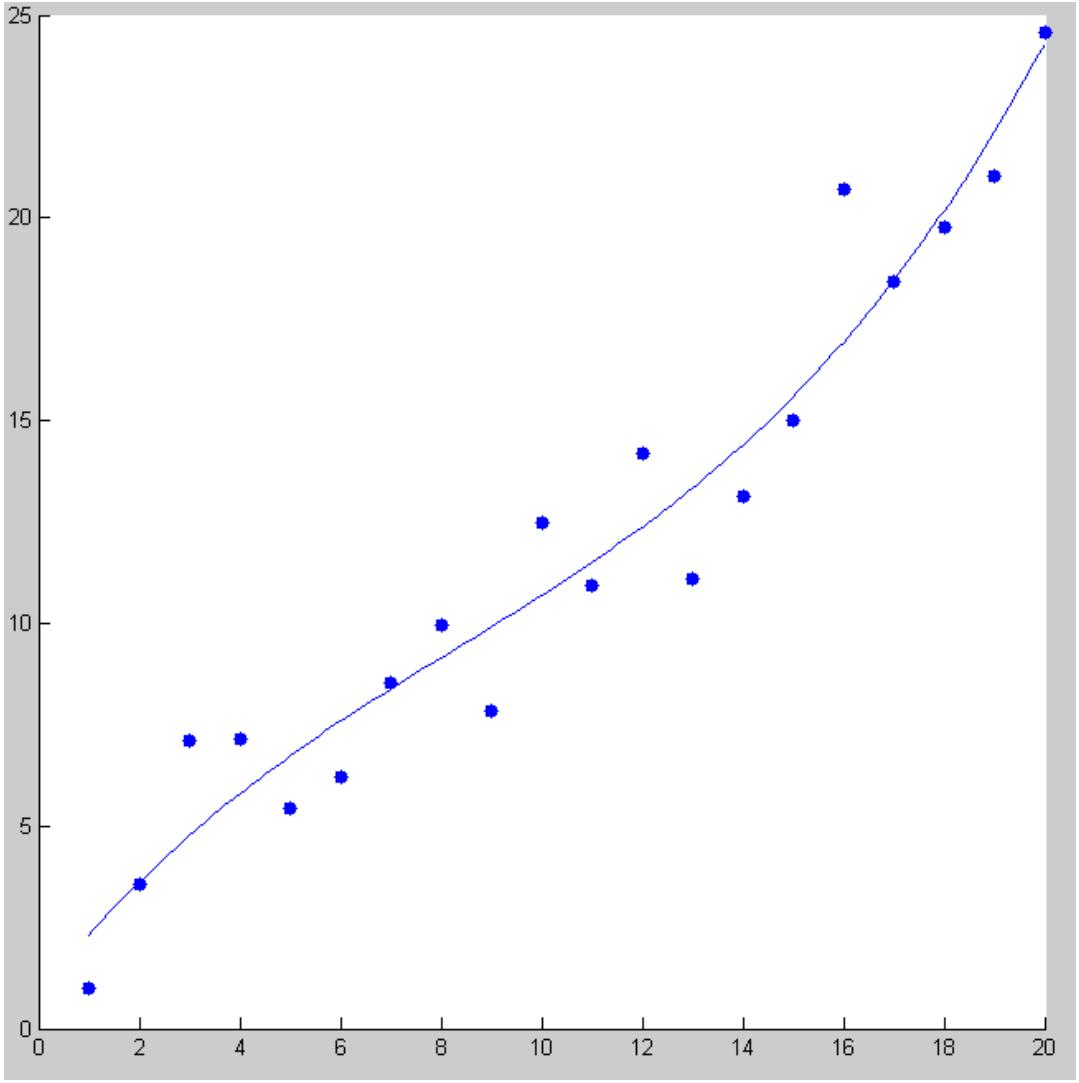
Overfitting in regression...



quadratic model:

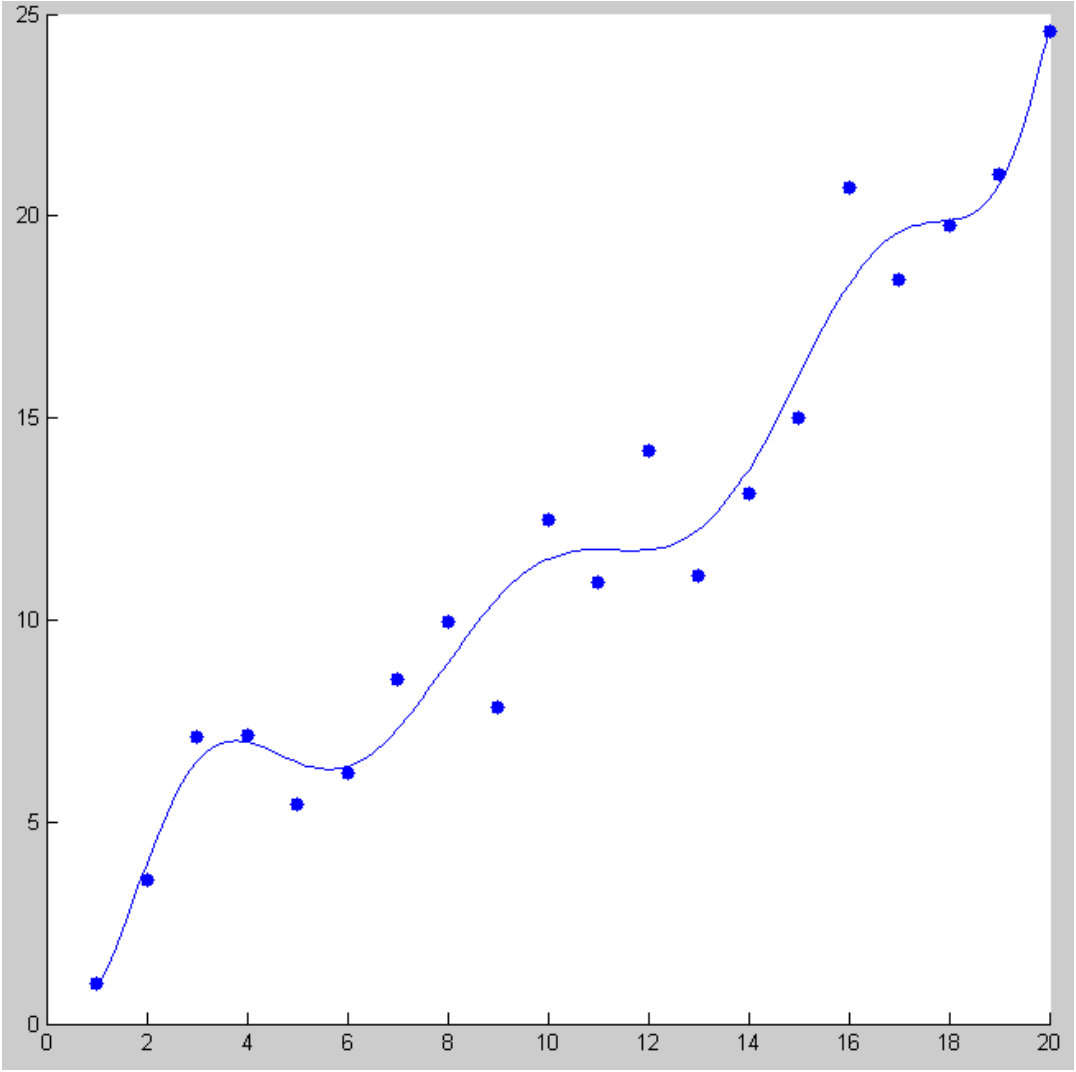
$a = w_0 + w_1 * F + w_2 * F^2$

Overfitting in regression...



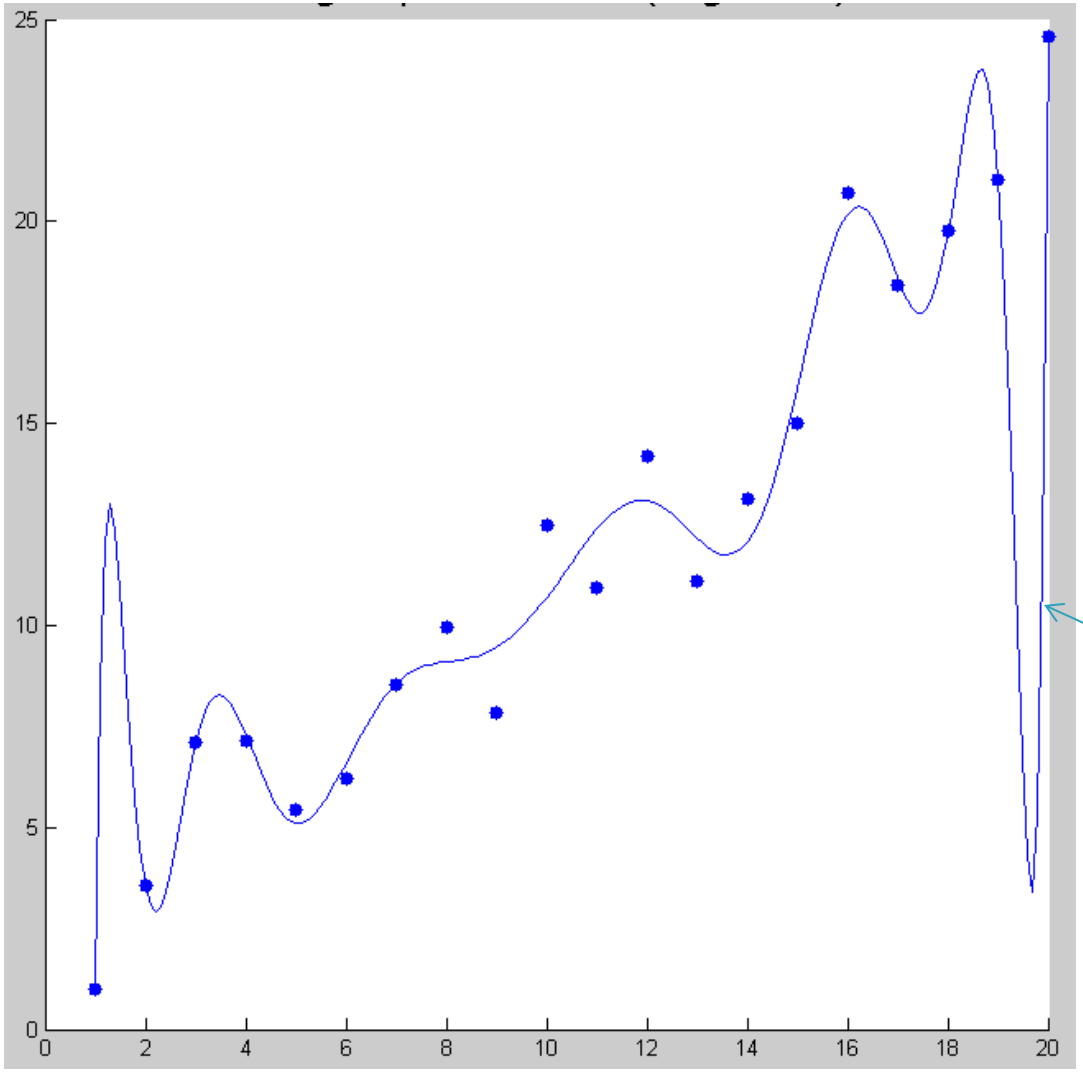
cubic model
(degree 3)

Overfitting in regression...



degree 10

Overfitting in regression…



degree 15

overfitting!

Occam's razor principle:

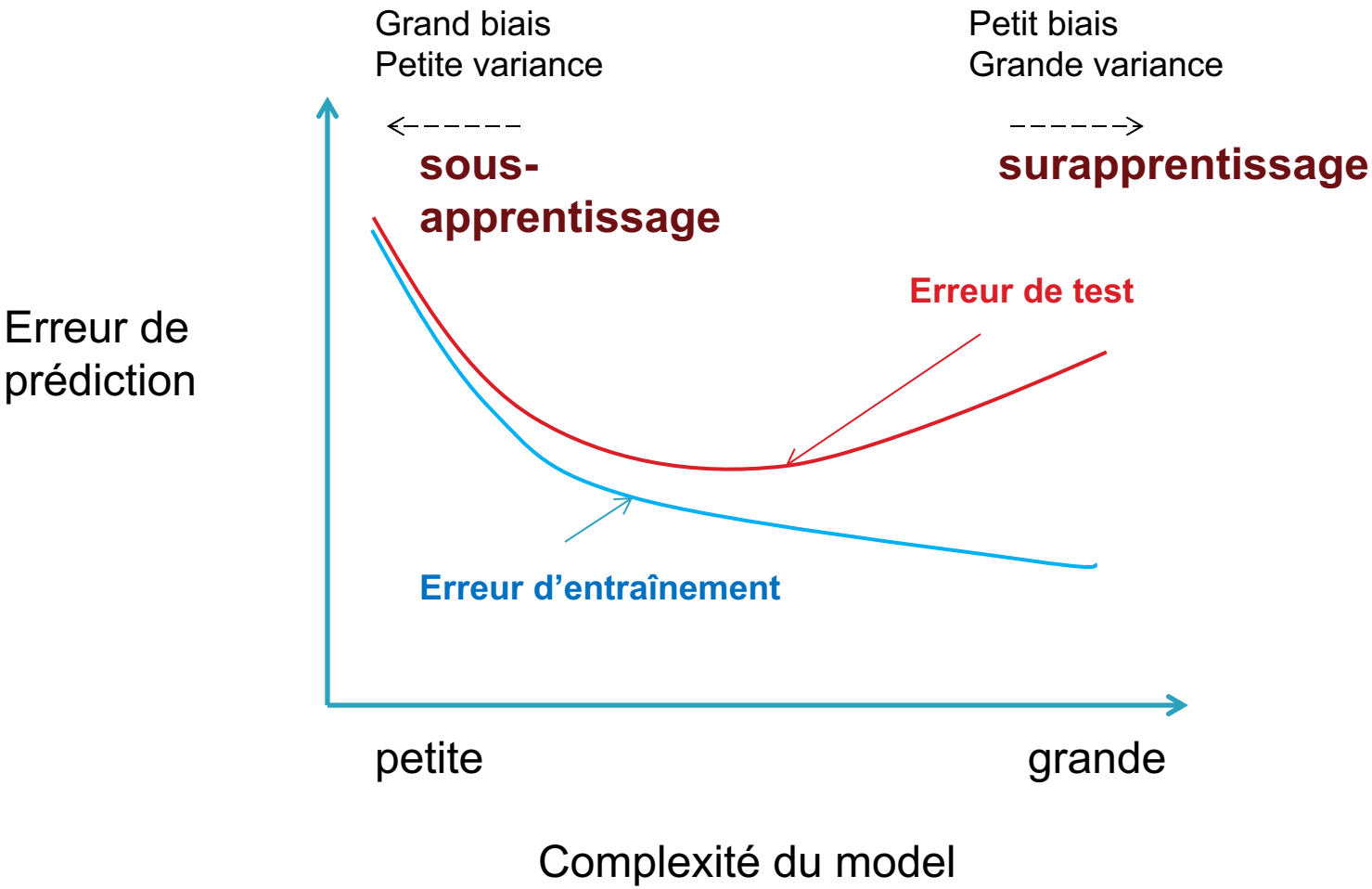
- Between two models / hypotheses which explain as well the data, choose the **simplest** one

- 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

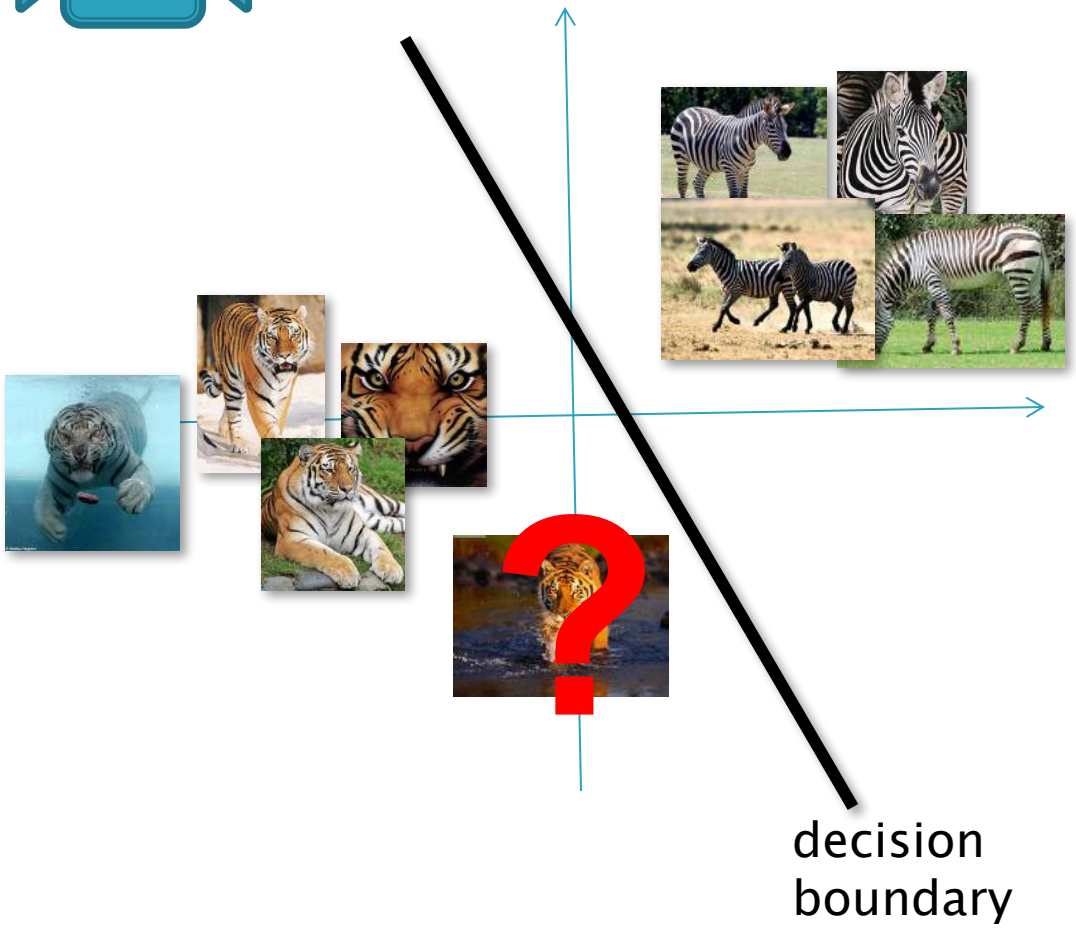


Classification

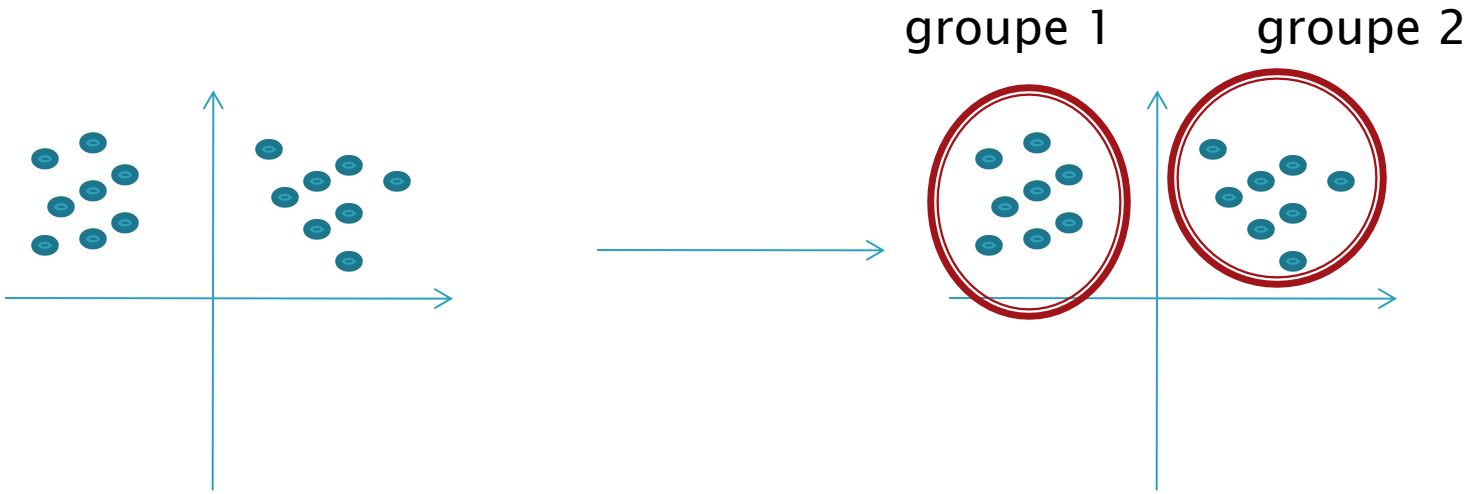
‘tiger’

‘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
- **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

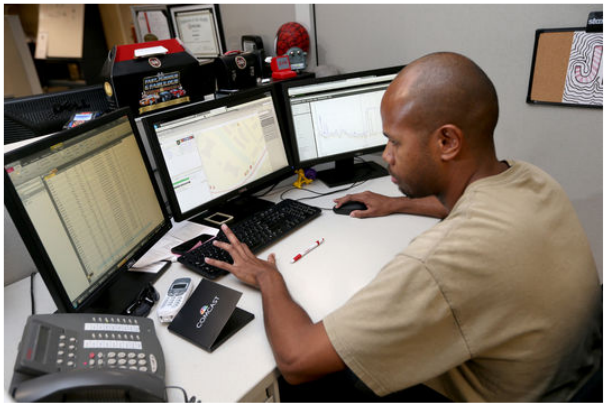
- Buy PDF

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 ».

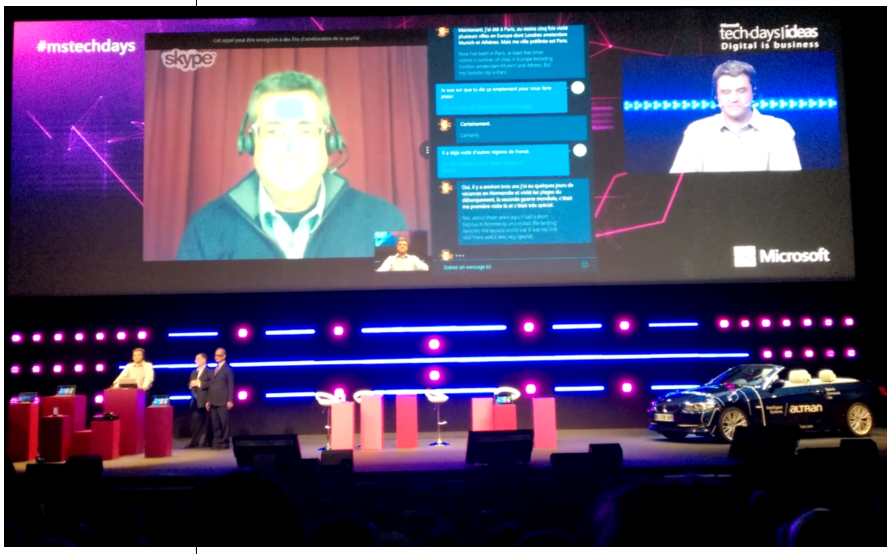
http://abonnes.lemonde.fr/economie/article/2014/04/08/un-metier-sexy-datascientifique_4397951_3234.html?xtmc=data_scientist&xtcr=1

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)

- Démonstration de Skype Translator à Microsoft Techdays



https://www.youtube.com/watch?v=QuwfcZ23fAI

TechDays : Microsoft passe à l'heure du machine learning

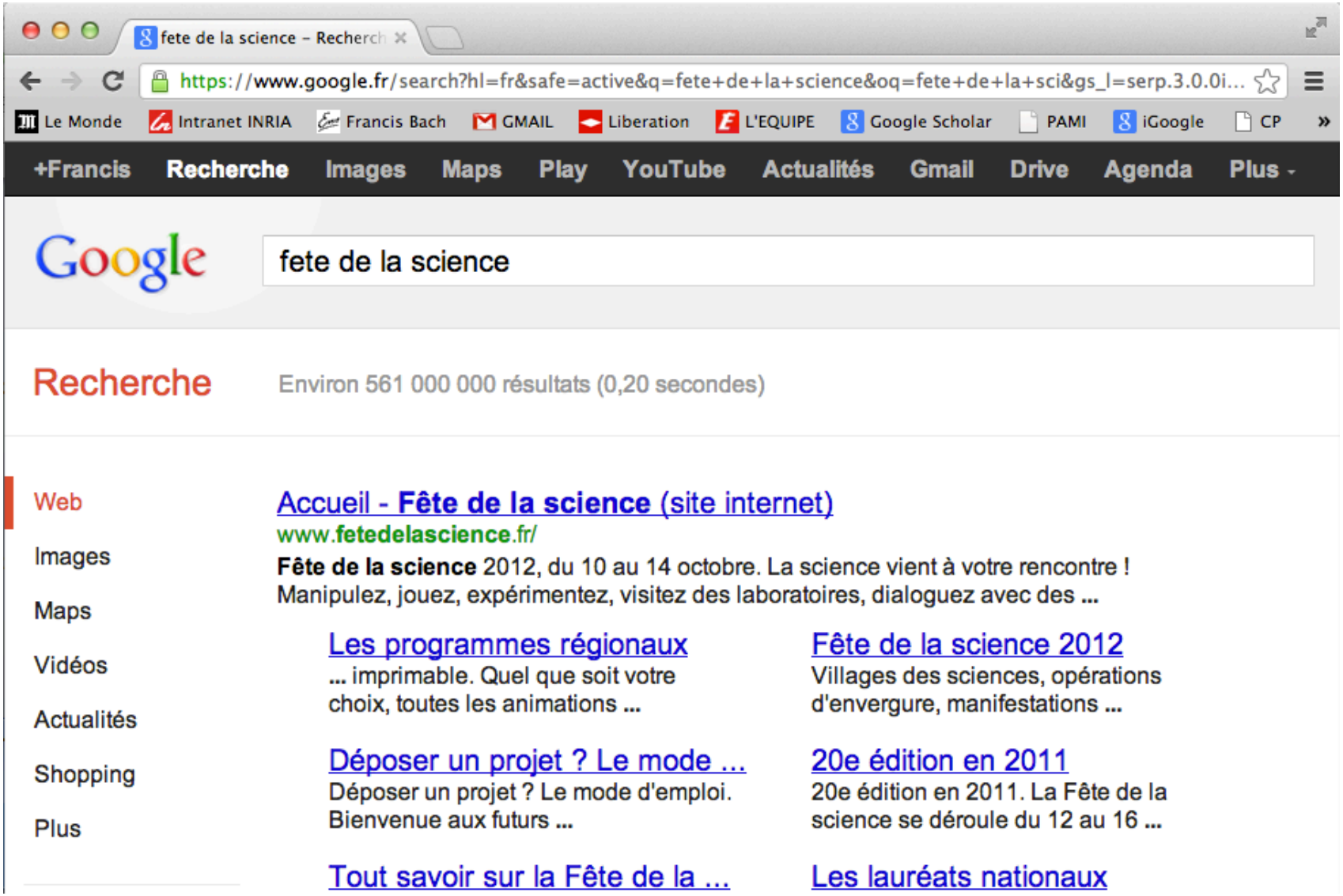
Virgile Juhan
JDN
12/02/15 09:46

Twitter LinkedIn Facebook Email

Le géant a profité de son événement pour montrer tout le potentiel du machine learning, avec des cas concrets... et son cloud Azure.

voir aussi: http://www.journaldunet.com/solutions/analytics/techdays-2015-microsoft-passe-a-l-heure-du-machine-learning-hdinsight-et-azure-machine-learning.shtml

Moteurs de recherche



Moteurs de recherche – 2

Google photo (3).JPG chateau lagrange

Search About 21 results (0.75 seconds)

Web
Images
Maps
Videos
News
Shopping
More

Search by image
Visually similar
More sizes

Any time
Past hour
Past 24 hours
Past week
Past month

Image size:
1536 × 2048

No other sizes of this image found.

Best guess for this image: ***chateau lagrange***

Château Lagrange Grand Cru Classé Saint-Julien :. Accueil
www.chateau-lagrange.com/ - Translate this page
Présentation du **château**. Historique, vignoble et fiche technique de ce saint-julien. 1 image

Château Lagrange - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Château_Lagrange
Château Lagrange is a winery in the Saint-Julien appellation of the Bordeaux region of France. **Château Lagrange** is also the name of the red wine produced by ... 1 image

Visually similar images - Report images



Recherche Google

https://www.google.fr/search?tbs=sbi:AMhZZitd0-7DEs0WcTd-pl0AmmnQtDV4WpkWaFYduZBcqyA5dbmxlo4o...

Le Monde | Intranet INRIA | Francis Bach | GMAIL | Liberation | L'EQUIPE | Google Scholar | PAMI | iGoogle | CP

+Francis Recherche **Images** Maps Play YouTube Actualités Gmail Francis Bach 0 Partager...

Google essai.jpg × Décrivez l'image ici

Recherche

- Web
- Images
- Maps
- Vidéos
- Actualités
- Shopping
- Plus

Recherche par image

Apparence similaire

Autres tailles

Date indifférente

Conseil : Essayez d'entrer un mot descriptif dans le champ de recherche.

Taille de l'image :
1355 × 804

Aucune autre taille d'image trouvée.

Images similaires - Signaler des images inappropriées




Recherche Google

https://www.google.fr/search?tbs=sbi:AMhZZivztYKFKxgPg_1gJfmUlczu6q9yVQiu6x_1ZDVer7UNnZW-_1R79p...

Le Monde | Intranet INRIA | Francis Bach | GMAIL | Liberation | L'EQUIPE | Google Scholar | PAMI | iGoogle | CP

+Francis | Recherche | **Images** | Maps | Play | YouTube | Actualités | Gmail | Francis Bach | 0 | Partager...

Google

photo...cation.jpg ×

Décrivez l'image ici

Recherche

Environ 5 résultats (0,64 secondes)

- Web
- Images
- Maps
- Vidéos
- Actualités
- Shopping
- Plus

Recherche par image

- Apparence similaire
- Autres tailles

Taille de l'image :
1773 × 1182

Trouver d'autres tailles de l'image :
Toutes les tailles - Petite

Pages contenant des images identiques

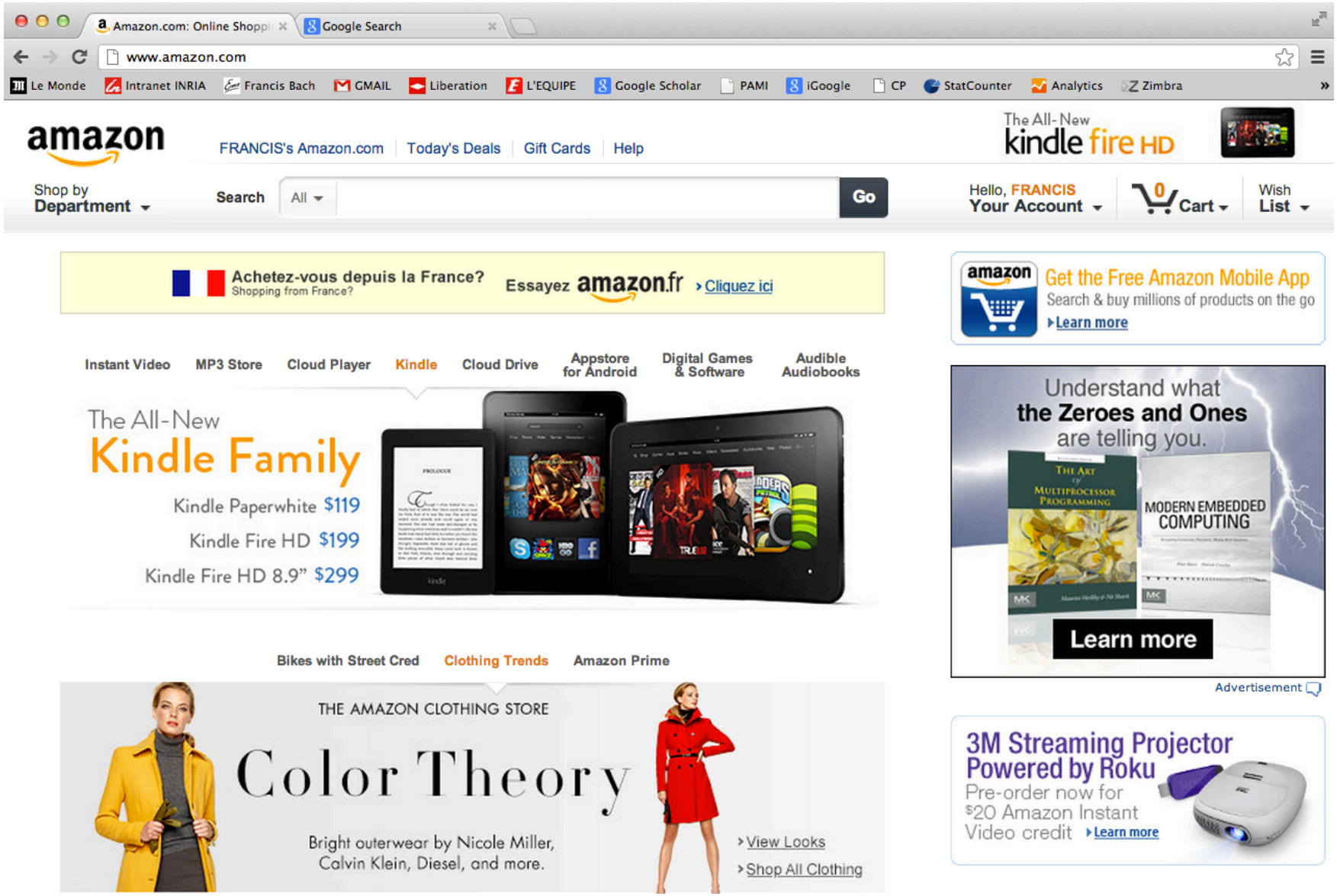
300 × 200

[Page 7 - Russian Online Dating - Russian Love Match]
onlinedating.russianlovematch.com/?page=7 - Traduire cette page
Are you finding that the messages that you and your Russian bride send each other have the same style and rhythm? This can be a sign that the two of you are a ...

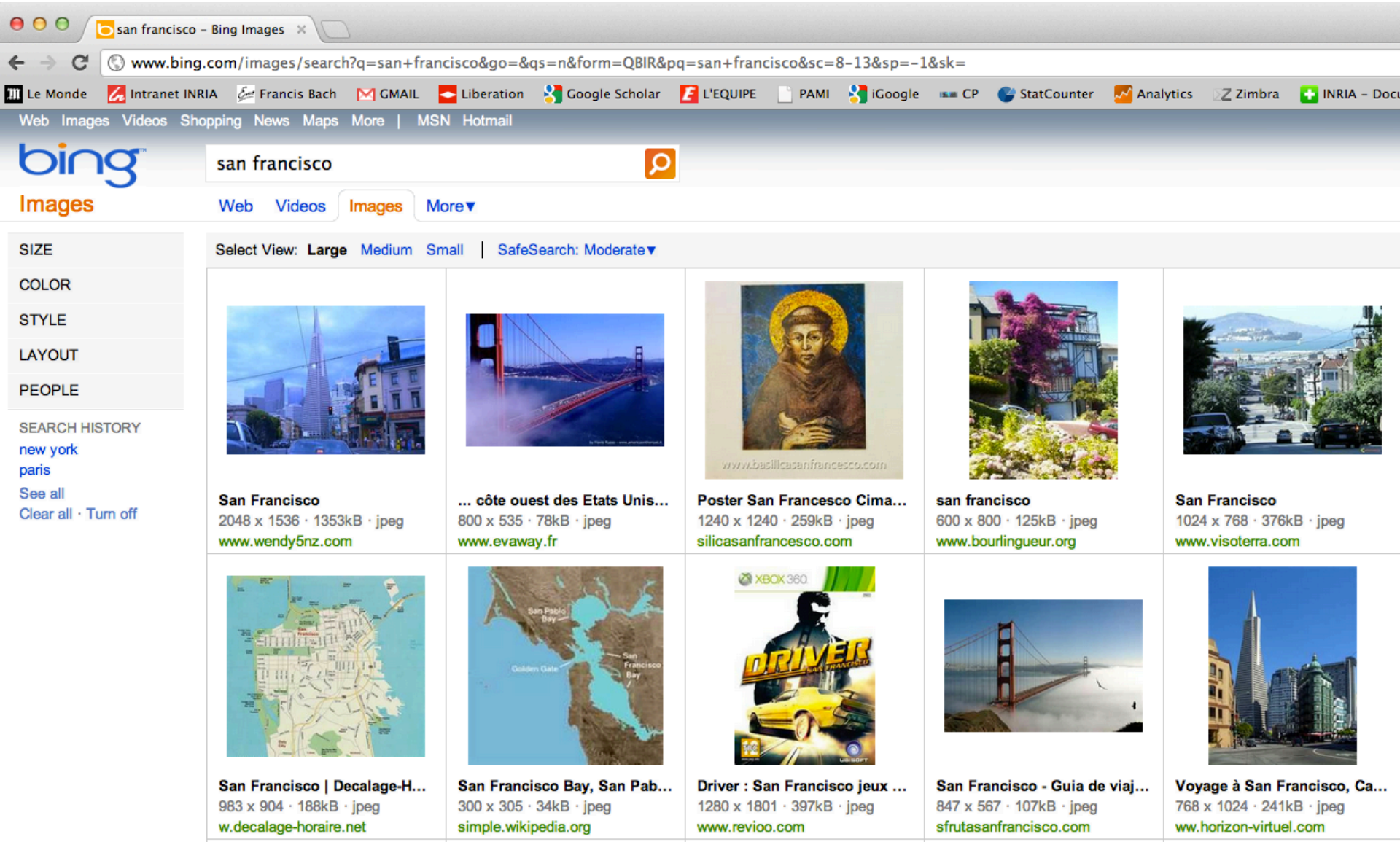
160 × 107

[Bezirksleiterin Daniela Brandes - Ihr Ansprechpartner vor Ort]
www.wuestenrot.de/de/.../adphomepage.php?vnr... - Traduire cette page
Ihre Ansprechpartnerin vor Ort. Bezirksleiterin Daniela Brandes. Bei uns sind Sie in guten Händen! Sie wollen. sich den Wunsch vom eigenen Haus oder der ...

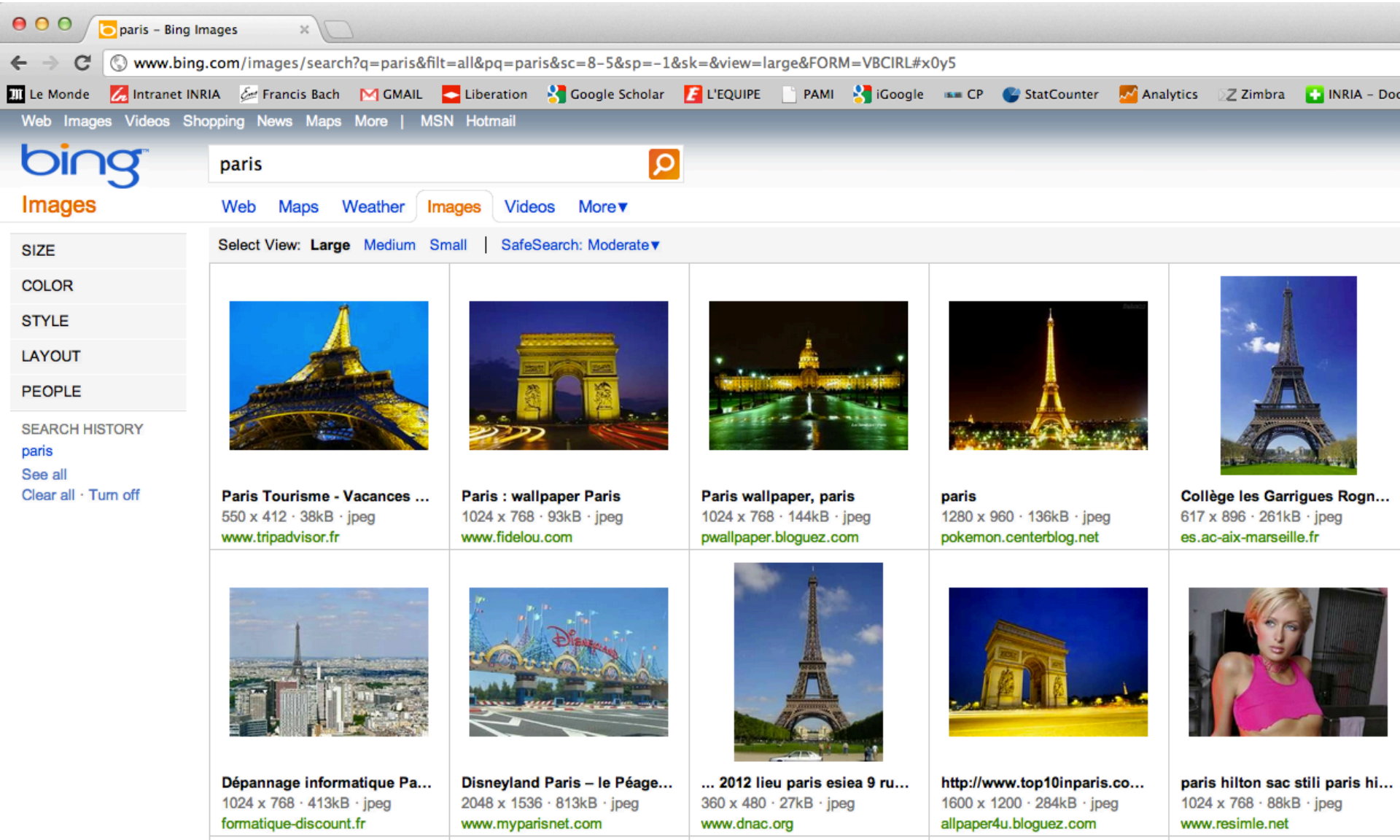
Publicite


Amazon.com: Online Shoppi
Google Search
www.amazon.com
Le Monde
Intranet INRIA
Francis Bach
GMAIL
Liberation
L'EQUIPE
Google Scholar
PAMI
iGoogle
CP
StatCounter
Analytics
Zimbra
amazon
FRANCIS's Amazon.com
Today's Deals
Gift Cards
Help
The All-New kindle fire HD
Shop by Department
Search
All
Go
Hello, FRANCIS Your Account
0 Cart
Wish List
Achetez-vous depuis la France?
Shopping from France?
Essayez amazon.fr
Cliquez ici
Instant Video
MP3 Store
Cloud Player
Kindle
Cloud Drive
Appstore for Android
Digital Games & Software
Audible Audiobooks
The All-New
Kindle Family
Kindle Paperwhite $119
Kindle Fire HD $199
Kindle Fire HD 8.9" $299
Bikes with Street Cred
Clothing Trends
Amazon Prime
THE AMAZON CLOTHING STORE
Color Theory
Bright outerwear by Nicole Miller, Calvin Klein, Diesel, and more.
View Looks
Shop All Clothing
amazon
Get the Free Amazon Mobile App
Search & buy millions of products on the go
Learn more
Understand what
the Zeroes and Ones
are telling you.
THE ART OF MULTIPROCESSOR PROGRAMMING
MODERN EMBEDDED COMPUTING
Learn more
Advertisement
3M Streaming Projector
Powered by Roku
Pre-order now for $20 Amazon Instant Video credit
Learn more

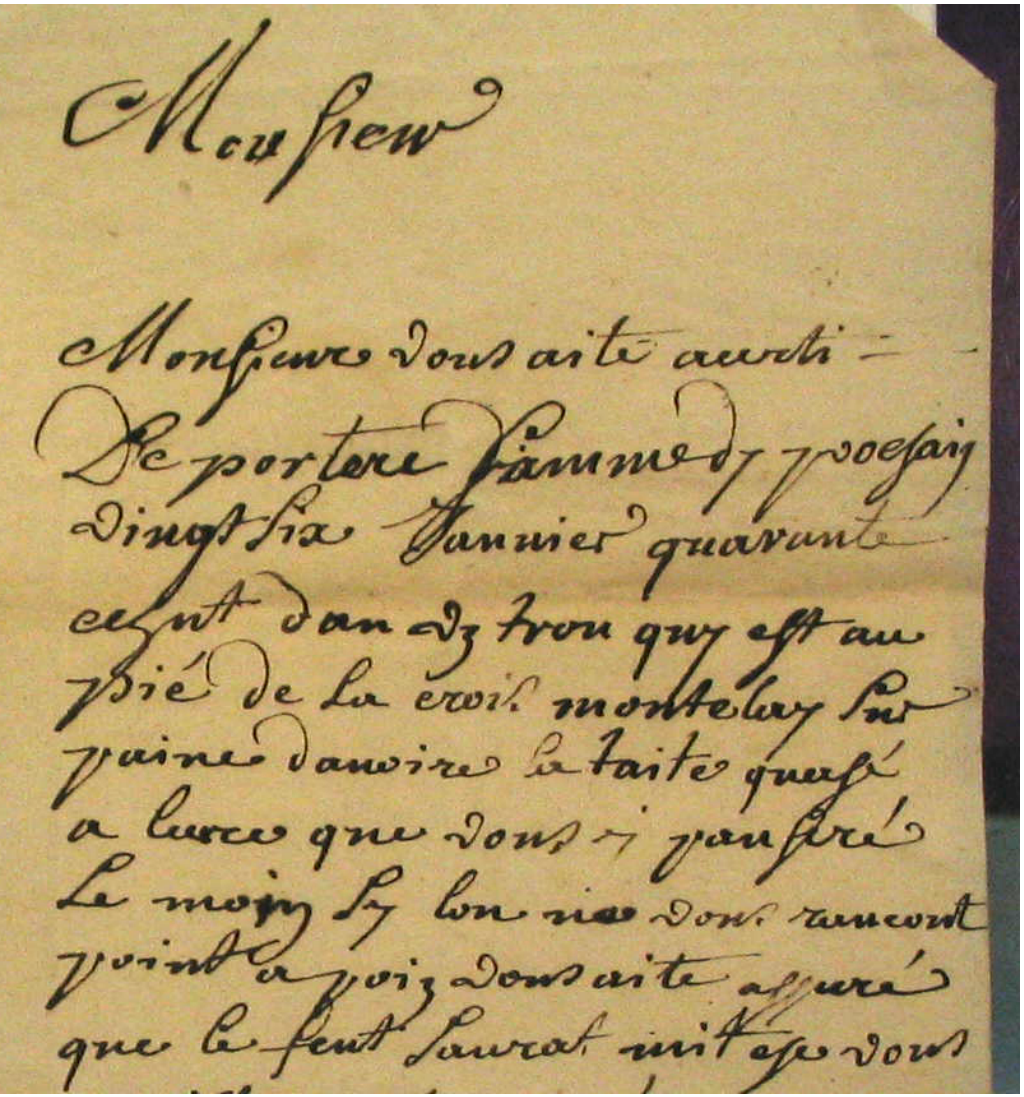
Apprentissage pour donnees multimedia



Apprentissage pour donnees multimedia



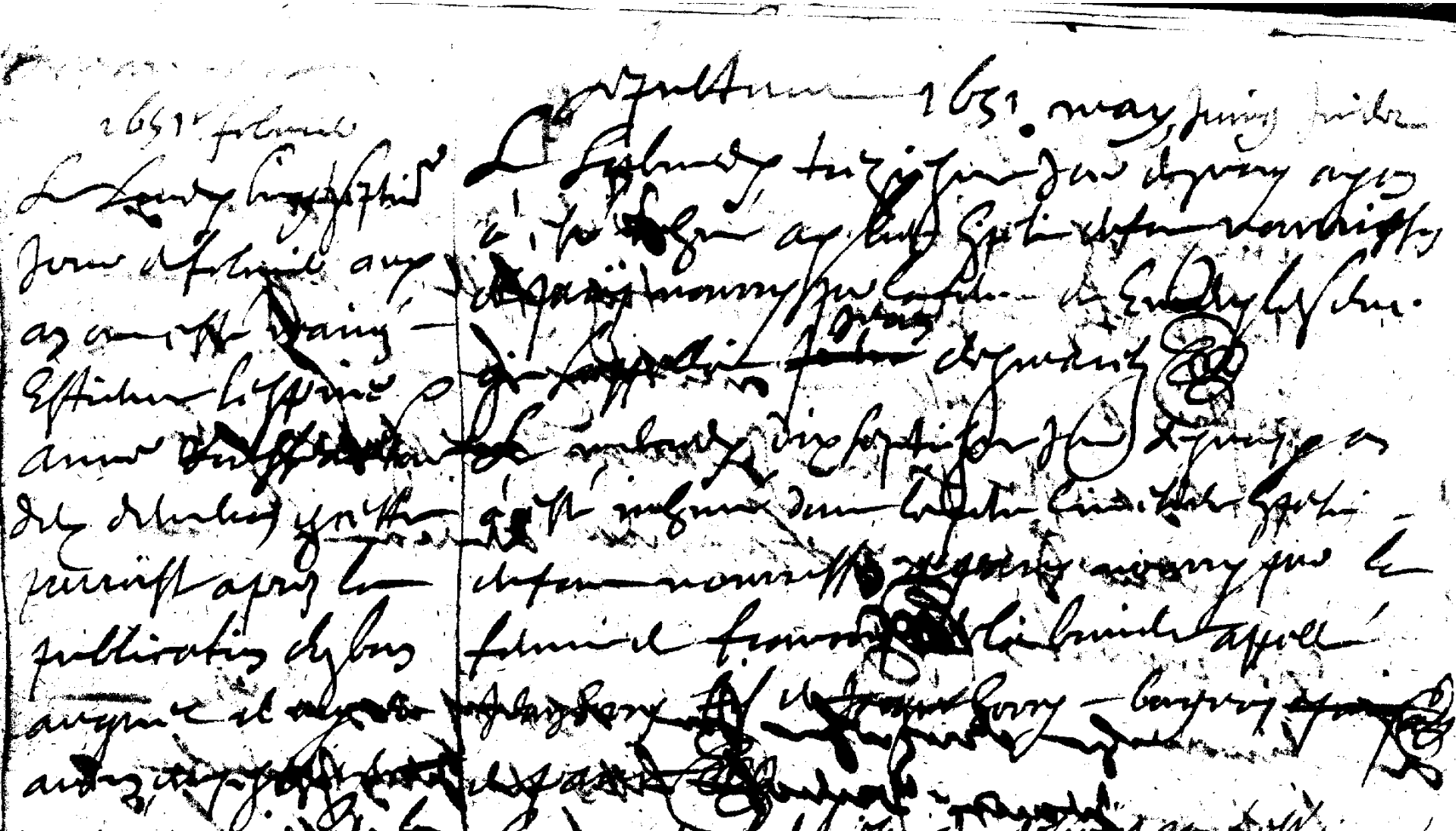
Apprentissage pour donnees multimedia



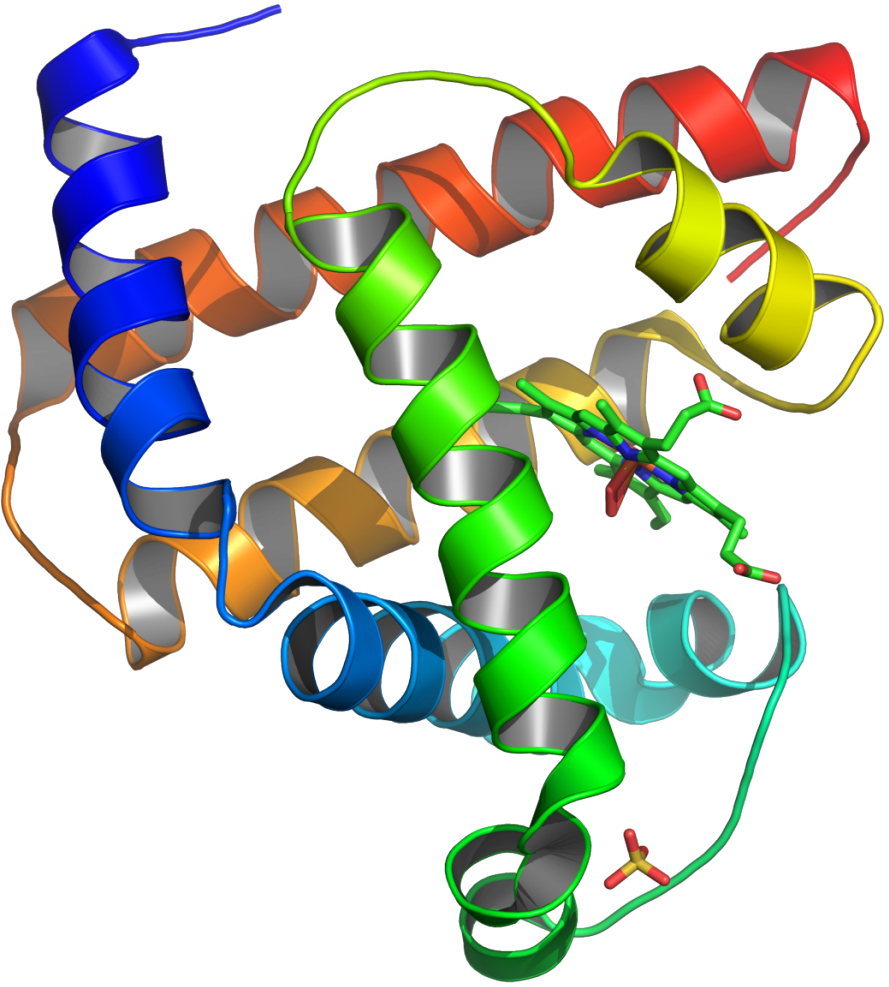
Monsieur,
Vous êtes averti de porter samedi prochain 26 janvier quarante écus dans un trou qui est au pied de la croix Montelay sous peine d'avoir la tête cassée à l'heure que vous y penserez le moins. Si l'on ne vous rencontre point vous êtes assuré que le feu sera mis chez vous. S'il en est parlé à qui que ce soit la tête cassée vous aurez.

Archives du Val d'Oise - 1737

Apprentissage pour donnees multimedia



Apprentissage pour la bioinformatique (proteines)



- Eléments essentiels de la vie de la cellule
- Prédiction des multiples fonctions et interactions des protéines
- Données massives
  - 2 millions pour l'homme!
- Données complexes
  - Chaîne d'acides aminés
  - Lien avec l'ADN
  - Molécule tri-dimensionnelle

Apprentissage pour la bioinformatique (puce a ADN)

- **Mesures du niveau d'expression des gènes**
- **Beaucoup de gènes / peu de patients**

Résumé

- ‘révolution’ du Big Data et IA:

  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/appstat

Vendredi 8h30-12h30 - Salle R
premier cours: 15 sept.

co-enseigné par:

Pierre Gaillard

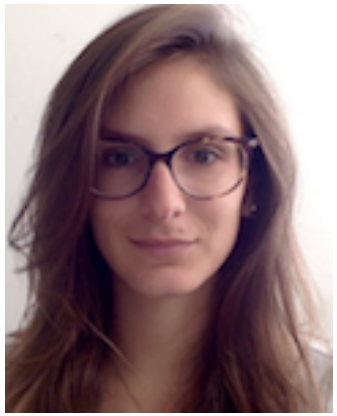
DI, ENS

Francis Bach

DI, ENS

chargé de TD:

Aude Genevay



DMA, 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 rendre plus concret des outils des math appliquées (statistiques, algèbre linéaire, analyse, etc.)
- 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 les données sont analysées

Logistique:

- 9 ECTS
- Note: 40% par l'examen, 40% par un TP à rendre, et 20% par les TDs a finir a la maison
- 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!
- **Premier DM pour le prochain cours:**
  - inscription sur la mailing list:

    http://tinyurl.com/hm4t8fc

    faire le TP d'intro de Matlab (voir site web)

Curriculum (prévisionnel)

15/09	Francis	2h	Introduction
	Francis	2h	Apprentissage supervisé
22/09	Pierre	2h	Régression linéaire / logistique (+regularisation)
	Aude	2h	(TP/TD) Régression linéaire / logistique
29/09	Pierre	2h	Plus proches voisins - arbres de decision - forets aleatoires
	Aude	2h	(TD) K-plus proche voisins
06/10	Francis	2h	Analyse convexe
	Aude	2h	(TD) Analyse convexe
13/10	Francis	2h	Optimisation convexe
	Aude	2h	(TD) Optimisation convexe
20/10	Pierre	2h	Théorie, concentration et borne PAC
	Aude	2h	(TD) Théorie, concentration et borne PAC
27/10	Pierre	2h	Méthodes probabilitistes (maximum de vraisemblance)
	Aude	2h	(TD) Méthodes probabilistes (maximum de vraisemblance)
3/11			Pas de classe

Curriculum (prévisionnel)

10/11	Pierre Aude	2h 2h	Méthode à noyaux (I) (TD) Méthode à noyaux (I)
17/11	Pierre Aude	2h 2h	Méthode à noyaux (II) (TD) Méthodes à noyaux (II)
24/11	Pierre Aude	2h 2h	Selection de variable (Lasso) (TP/TD) Selection de variable
1/12	Pierre Aude	2h 2h	Apprentissage Sequentiel - Renforcement
8/12			Pas de classe
15/12	Francis Aude	2h 2h	Apprentissage non supervise (TD) Kmeans et PCA
22/12			Pas de classe
29/12			Vacances
5/1			Vacance
12/1	Francis Aude	2h 2h	Résumé et questions / réponses Exercises d'entrainement
19/1		4h	Exam