le cnam

Dans quelle mesure le Big Data renouvelle t'il les méthodes d'analyse de données ?

Gilbert Saporta
CEDRIC- CNAM,
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Plan

1. Le phénomène Big Data
2. Too big ?
3. Vous avez dit « modèles »?
4. Agrégation de modèles
5. Comment valider
6. La fin de la science?
7. Compétences et formations

1. Le phénomène Big Data

Le Monde

Big data

Les données, nouvelle devise

Les grands volumes d'information finiront-ils par devenir la matière première des entreprises ? Ils constituent déjà, grâce au « big data », une ressource essentielle pour créer des produits plus ciblés et augmenter la productivité

- Une révolution

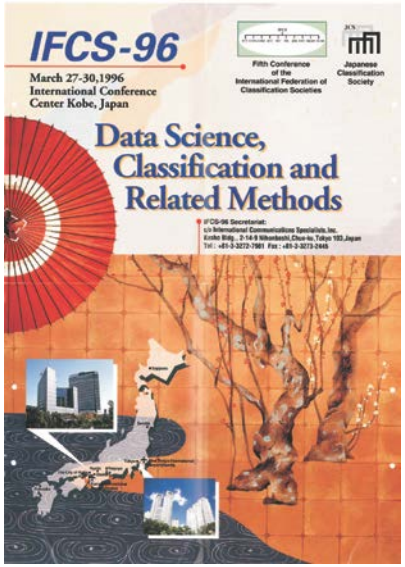
DEFINING THE DATA REVOLUTION

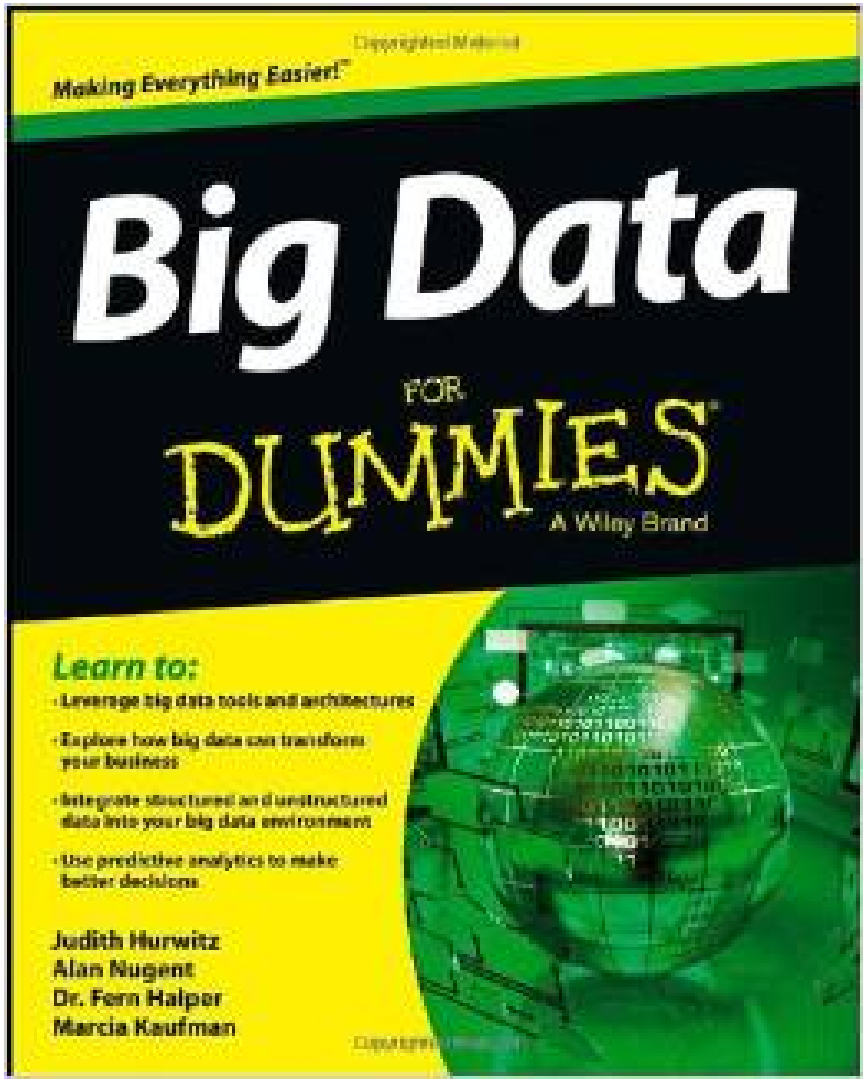
'The data revolution is: an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the 'Internet of Things,' and from other sources, such as qualitative data, citizen-generated data and perceptions data; A growing demand for data from all parts of society.'

UN Secretary-General's Independent Expert Advisory Group on a Data Revolution (A World That Counts report, page 6)

- Big Data apparait pour la première fois en 1997:
  - Cox & Ellsworth (NASA, pas NSA!)
    « Managing Big Data for Visualisation »
    ACM SIGGRAPH '97

- Data Science bien plus ancien
  - P.Naur 1960
  - IFCS (Kobe, 1996) "Data Science, classification, and related methods“
  - Journal of Data Science depuis 2003




Making Everything Easier!™
Big Data
FOR
DUMMIES
A Wiley Brand
Learn to:
· Leverage big data tools and architectures
· Explore how big data can transform your business
· Integrate structured and unstructured data into your big data environment
· Use predictive analytics to make better decisions
Judith Hurwitz
Alan Nugent
Dr. Fern Halper
Marcia Kaufman


Tout savoir sur...
Big Data
Le cinéma avait déjà tout imaginé !
UN CRIME DANS LA TÊTE
BACK TO THE FUTURE
SCHWARZENEGGER
TERMINATOR 3
MATRIX
TOM CRUISE
THE MANCHURIAN CANDIDATE
REQUIEM
DEAD ZONE
MINORITY REPORT
Un livre de Xavier Perret & Guy Jacquemelle
Préface de Stéphane Richard
Une collection dirigée par Henri Kaufman
EDITIONS kawa
www.editions-kawa.com

- Cet exposé :
  - V de volume

2. Too big ?

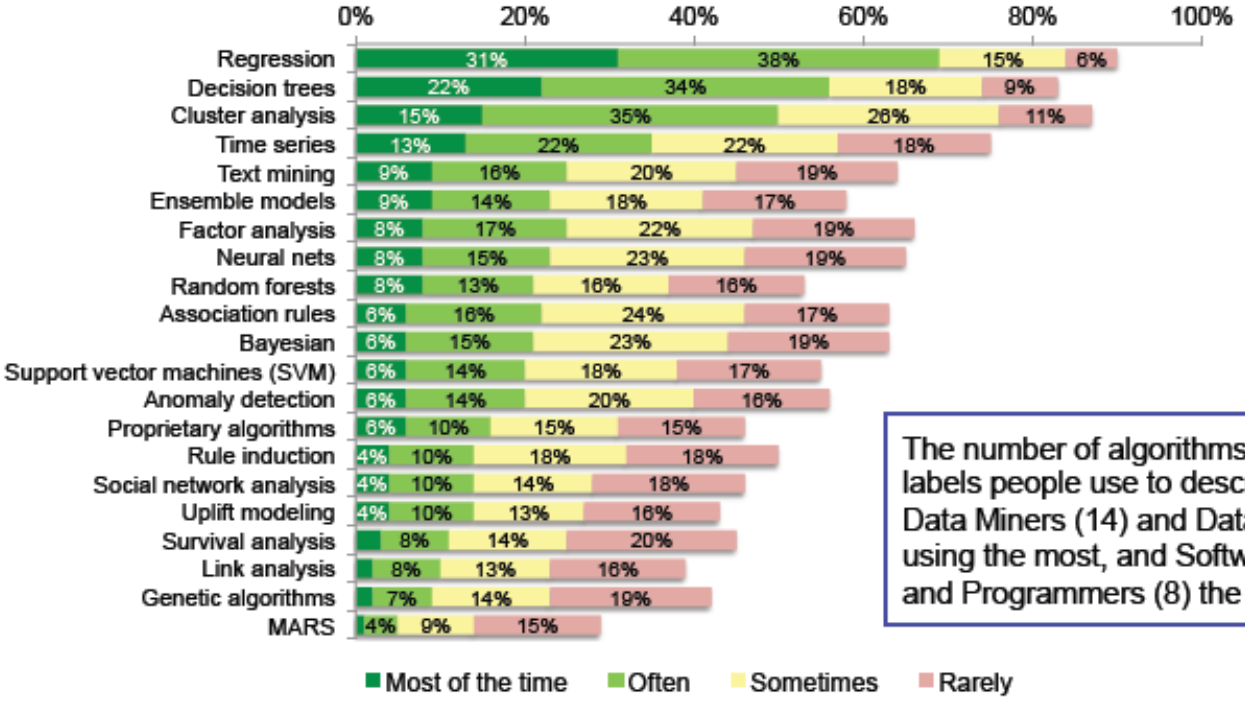
- Estimation, tests, modèles classiques inadaptés
  - Tout est significatif!
    - si $n=10^6$ un coefficient de corrélation égal à 0,002 est significativement différent de 0 mais bien inutile
    - Un modèle de régression pourri aura un R^2 « significatif » mais la plupart des modèles classiques sont rejetés puisque le moindre écart devient significatif
    - Intervalles de confiance réduits à néant

La boîte à outils

- Exploratoire ou non supervisé
  - Analyses factorielles, k-means
  - Règles d'association
- Prédictif ou supervisé
  - Modèles explicites de type régression, avec régularisation, arbres ..
  - Modèles de type boîte noire (neurones, SVM, ..)

Algorithms

- Regression, decision trees, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent since the first Data Miner Survey in 2007.
- The average respondent reports typically using 12 algorithms. People with more years of experience use more algorithms, and consultants use more algorithms (13) than people working in other settings (11).



The number of algorithms used varies by the labels people use to describe themselves, with Data Miners (14) and Data Scientists (14) using the most, and Software Developers (9) and Programmers (8) the fewest.

Question: What algorithms / analytic methods do you TYPICALLY use? (Select all that apply)

3. Vous avez dit « modèles »

- Vision classique: modèles pour comprendre
  - Fournir une certaine compréhension des données et du mécanisme qui les a engendrées à travers une représentation parcimonieuse d'un phénomène aléatoire. Nécessite en général la collaboration d'un statisticien et d'un expert du domaine. **Modèle génératif**
  - un modèle doit être simple, et ses paramètres interprétables en termes du domaine d'application : élasticité, odds-ratio, etc.

- Paradoxe n° 1
  - Un « bon » modèle statistique ne donne pas nécessairement des prédictions précises au niveau individuel. eg: facteurs de risque en épidémiologie
- Paradoxe n°2
  - On peut prévoir sans comprendre:
    - pas besoin d'une théorie du consommateur pour faire du ciblage

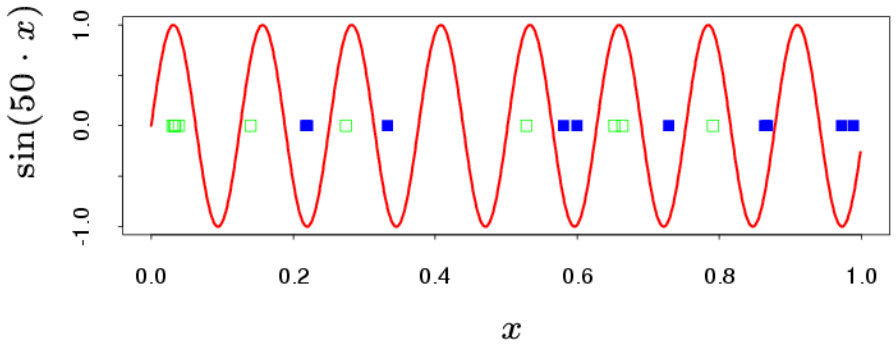
- Vision « Big Data Analytics »: **modèle prédictif**
  - capacité prédictive sur de nouvelles observations «généralisation »
  - différent de l'ajustement aux données (prédire le passé)
    - Un modèle trop précis sur les données se comporte de manière instable sur de nouvelles données : phénomène de surapprentissage
    - Un modèle trop robuste (rigide) ne donnera pas un bon ajustement sur les données
  - modéles issus des données (« data driven »); en Data Mining et Machine learning un modèle n'est qu'un algorithme

- De « nouveaux » modèles issus du Machine Learning
  - Réseaux de neurones et deep learning
  - SVM
  - Règles d'association
  - Systèmes de recommandation (eg Amazon)
- Enrichissement des données avec les réseaux sociaux (ex. fraude)

L'apport de la théorie de l'apprentissage

- La complexité ne se limite pas au nombre de paramètres

V. Vapnik en 1990



Hastie et al. 2001

$$f(x,w) = sign\ (sin\ (w.x)\ ) \quad c < x < 1, c>0$$

Un seul paramètre, VC dimension infinie

- Quelques conséquences surprenantes de la théorie de Vapnik
  - Si la complexité augmente moins vite que n, on améliore la généralisation
  - Possibilité d'utiliser des modèles de plus en plus complexes quand n est très grand!
  - Pas forcément une bonne idée si p est très grand

4. Agrégation de modèles

- Pourquoi choisir entre plusieurs modèles?
- Combinaison linéaire des prévisions de m modèles

- Stacking
  - Combinaison de m prédictions obtenues par des modèles différents
  - Première idée : régression linéaire

$$\hat{f}_1(\mathbf{x}),\ \hat{f}_2(\mathbf{x}),\ldots,\ \hat{f}_m(\mathbf{x})$$

$$\min \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} w_j\, \hat{f}_j(\mathbf{x}) \right)^2$$

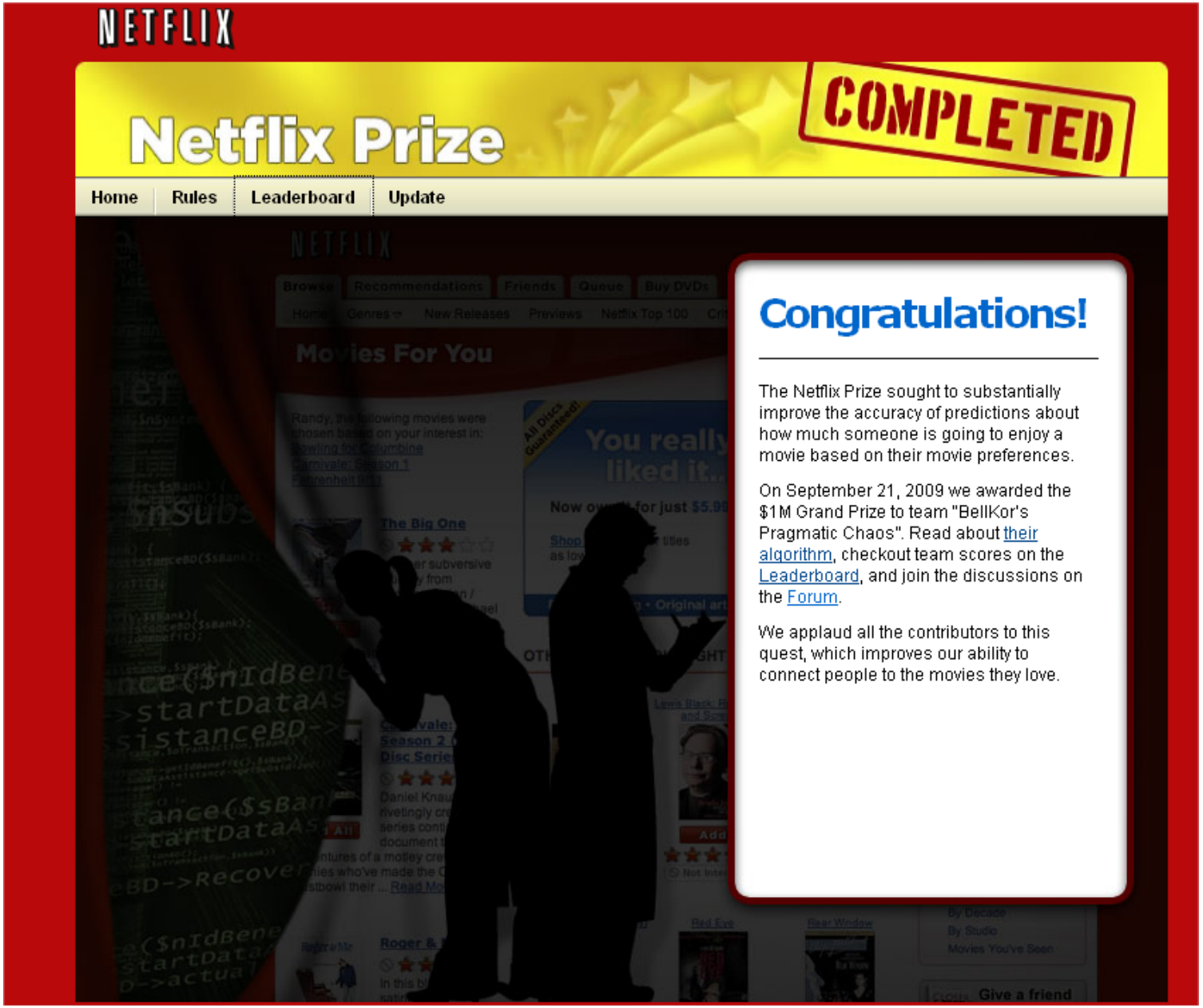
    - Favorise les modèles les plus complexes: surapprentissage

- Solution: utiliser les valeurs prédites en otant à chaque fois l'unité i

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j^{-i}(\mathbf{x}) \right)^2$$

- Améliorations:
  - Combinaisons linéaires à coefficients positifs (et de somme 1)
  - Régression PLS ou autre méthode régularisée car les m prévisions sont très corrélées

- Avantages
  - Prévision meilleures qu'avec le meilleur modèle
  - Possibilité de mélanger des modéles de toutes natures: arbres , ppv, réseaux de neurones etc.


NETFLIX
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Congratulations!
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

- The Netflix dataset contains more than 100 million datestamped movie ratings performed by anonymous Netflix customers between Dec 31, 1999 and Dec 31, 2005. This dataset gives ratings about m = 480 189 users and n = 17 770 movies
- The contest was designed in a training-test set format. A hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user (or fewer if a user had not rated at least 18 movies over the entire period).The remaining data made up the training set.

- *BellKor's Pragmatic Chaos team*. A blend of hundreds of different models
- *The Ensemble Team* . Blend of 24 predictions
- Same Test RMSE : 0.8567 (10.06%)

- Bellkor's Pragmatic Chaos defeated The Ensemble by submitting just 20 minutes earlier!

Mais Netflix n'a pas implémenté la méthode victorieuse:

We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment.

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

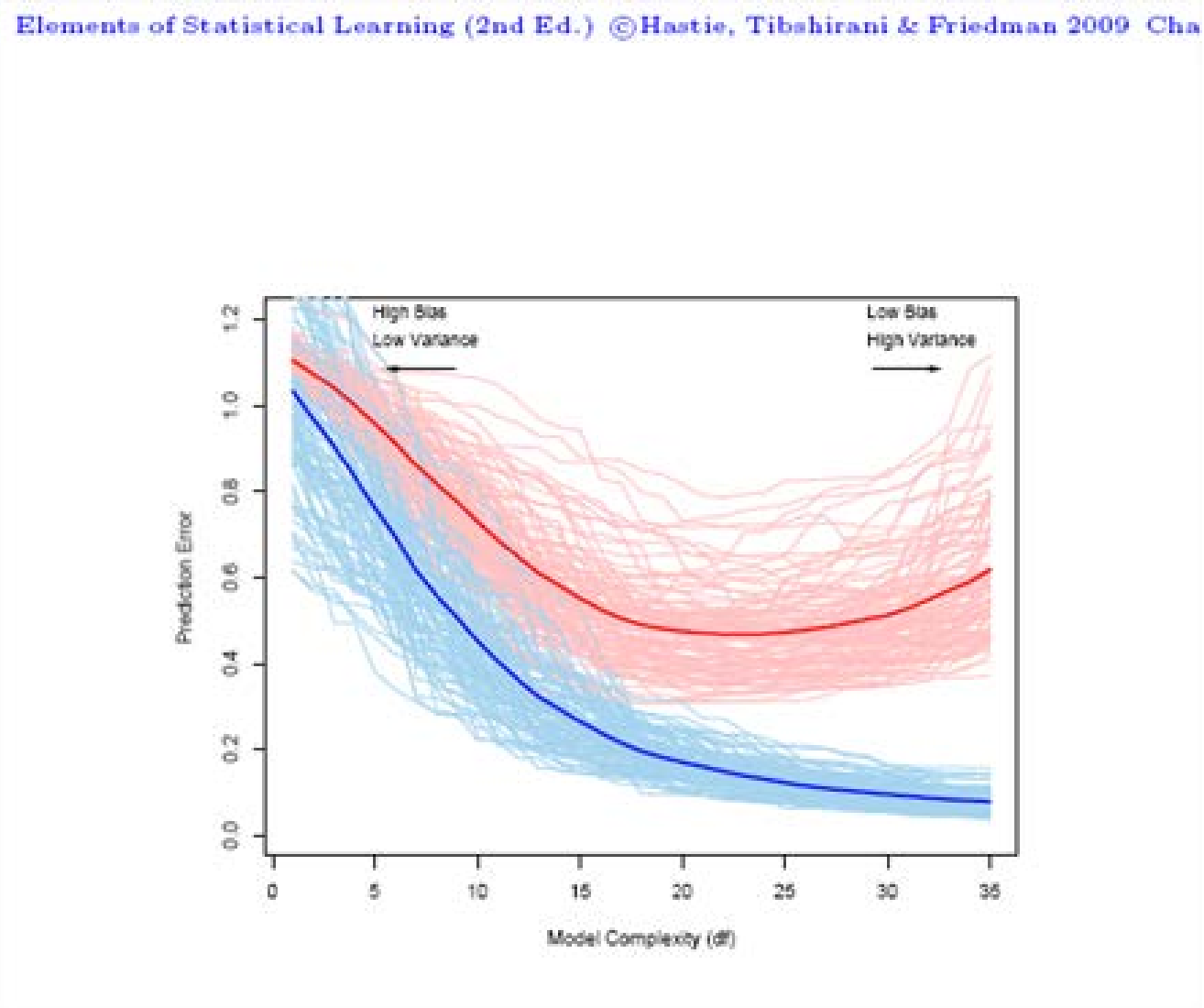
5. Comment valider

- Nécessité de marier Machine Learning et statistique
  - Un bon modèle est celui qui prédit bien
  - Différence entre ajustement et prévision
  - Ensembles d'apprentissage et de validation

Une démarche avec 3 échantillons pour choisir entre plusieurs familles de modèles:

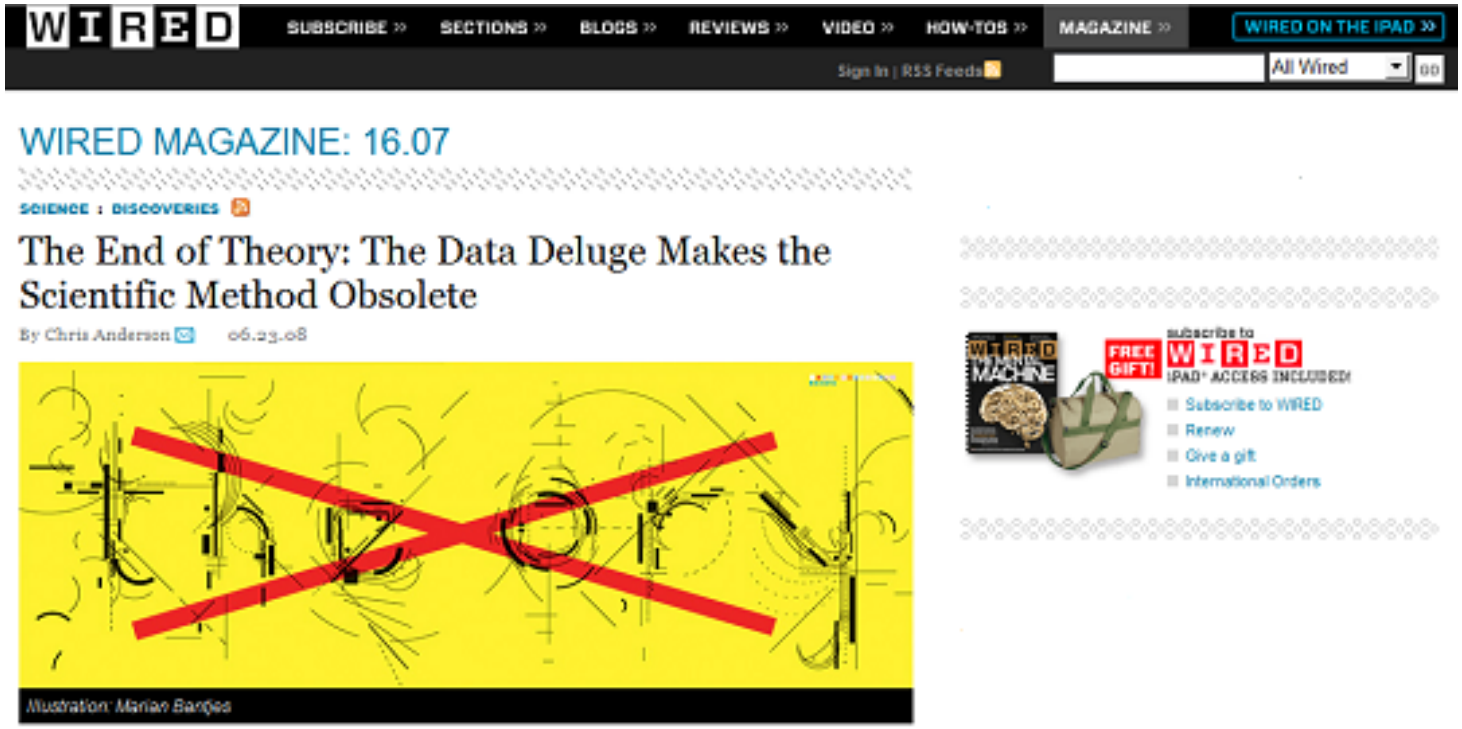
- Apprentissage: pour estimer les paramètres des modèles
- Test : pour choisir le meilleur modèle
  - Réestimation du modèle final**: avec toutes les données disponibles**
- Validation : pour estimer la performance sur des données futures
  - Estimer les paramètres ≠ estimer la performance

- Séparer (une fois) les données en apprentissage, test et validation ne suffit pas



- **Elémentaire?**
  - Pas si sur…
  - Voir publications en économètrie, actuariat, épidémiologie

6. La fin de la science?



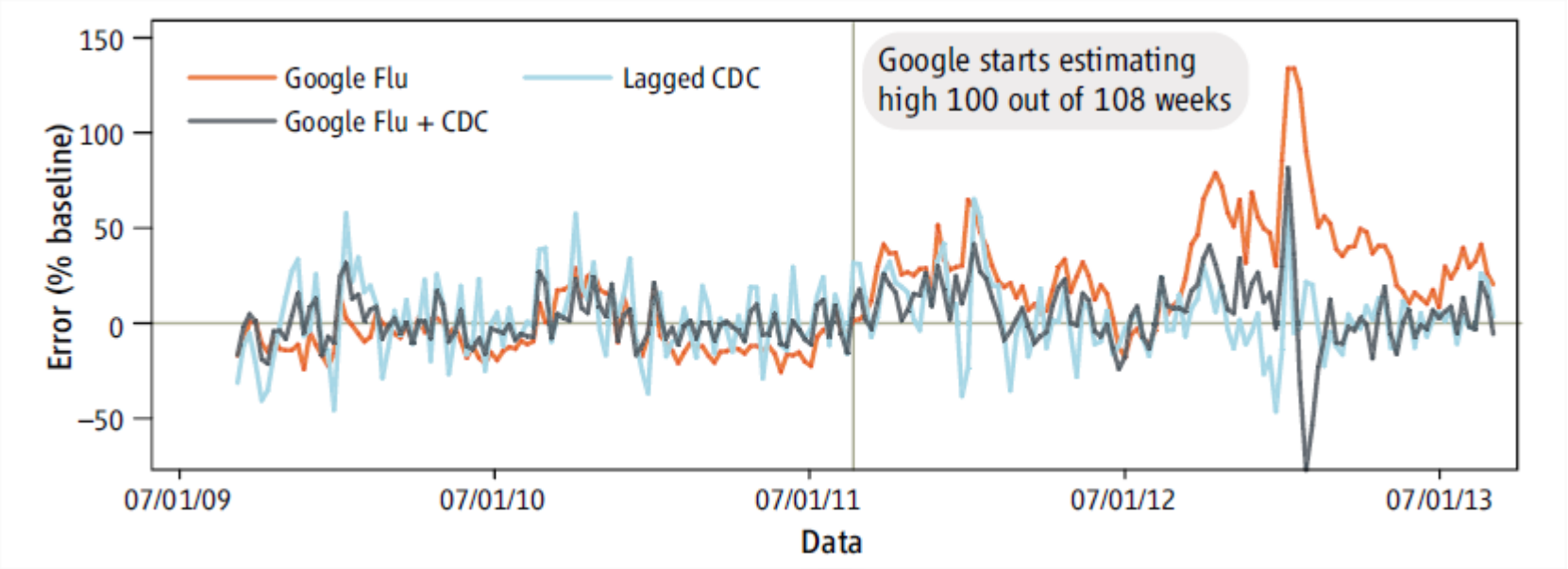
Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

- Google FluTrends

« En 2009, en pleine pandémie de grippe H1N1, le ministère américain de la santé a demandé l'aide de Google. En localisant sur une carte la position la provenance des mots-clés tapés dans le célèbre moteur de recherche, les ingénieurs ont pu dessiner et finalement anticiper l'évolution de l'épidémie: http://www.google.org/flutrends/ »

http://esante.gouv.fr/le-mag-numero-10/decryptage-le-big-data-sante

Surestimation de 50% en 2012-2013



BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014

- Corrélation n’est pas causalité…
- L’influence d’un prédicteur ne se mesure pas par son coefficient de régression (P.Bühlmann)
  - Le « toutes choses égales par ailleurs » est absurde
  - Faire varier un prédicteur entraine des variations des autres prédicteurs (intervention vs corrélation)
  - Nécessité d’un schéma causal

7. Compétences et formations

- Les données massives nécessitent une approche spécifique
- Quels statisticiens pour les Big Data?
  - « Data scientists »
  - Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician (Donoho, 2015)

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Top 10 des meilleurs métiers en 2014 : les notes critère par critère					
	Métier	Cadre de travail (1 = top / 200 = pire)	Stress (1 = cool / 200 = très stressant)	Perspective d'emploi (1 = top / 200 = pires)	Salaire (en $)
1	Mathématicien	2	24	19	101 360
2	Professeur d'université	1	4	47	68 970
3	Statisticien	8	54	12	75 560
4	Actuaire	10	54	8	93 680
5	Audiologiste	31	1	8	69 720
6	Hygiénisite dentaire	51	11	6	70 210
7	Ingénieur logiciel	56	18	24	93 350
8	Ingénieur système	13	58	18	79 680
9	Ergothérapeute	45	28	11	75 400
10	Orthophoniste	36	17	35	69 870

http://etudiant.aujourdhui.fr/etudiant/info/le-top-10-des-meilleurs-metiers-a-envisager-et-les-pires-a-eviter-xxxxx.html

compétences

- Mathématique et Statistique
- Algorithmique
- Informatique: bases de données réparties (cloud, NoSQL) calcul parallèle (Hadoop, MapReduce), visualisation

Quelques formations

- Ecoles et Masteres spécialisés
  - ENSAI, ENSAE, EISTI, Telecom Paris-Tech, ENSIMAG, HEC…
- Masters universitaires
  - dkm² Master course in **Data Mining and Knowledge Management** a european master
  - Master DAC UPMC

- Formation continue: diplômes d'établissement
  - Après Bac+2
    - IUT Paris-Descartes : 20 jours, 3000 €

  - Après Bac+5
    - ~ 300h

      ~ 500 €, soir

Merci pour votre attention