le cnam

Science des données, données massives : défis et nouveaux métiers

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. Quelques applications
3. Des défis pour la statistique
4. Valider et interpréter
5. La fin de la théorie?
6. Les Data Scientists
7. Conclusion

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

06/11/2014

What steam was to the 19th century, and oil has been to the 20th, data is to the 21st. It's the driver of prosperity, the revolutionary resource that is transforming the nature of economic activity, the capability that differentiates successful from unsuccessful societies.

The Data Manifesto, Royal Statistical Society, 2014

- 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)

Le nouvel eldorado des données

La trace laissée par les hommes dans l'univers numérique est devenue un véritable marché que se disputent les géants informatiques. Et sur lequel plane la question de la protection de la vie privée

Monsanto n'a pas pour habitude de planter ses graines au hasard. Le géant américain, symbole honni de l'industrie agricole et du maïs transgénique, vient de s'offrir une petite entreprise californienne, peu connue, pour 930 millions de dollars. Une paille.

Au fil des calculs, la centaine d'ingénieurs et de mathématiciens de Climate Corporation (la société rachetée) sont devenus des experts des rendements agricoles. Mais foin de bon sens paysan. Eux moissonnent et tirent parti d'une masse de données météorologiques afin de déterminer le meilleur jour de l'année pour semer, traiter ou récolter. En étudiant le ciel, les nouveaux meilleurs amis des agriculteurs sont capables de doubler, voire de tripler, le rendement d'un champ de maïs. Sans OGM.

C'est une des applications concrètes du « Big Data » – à traduire par « grosses données », auquel est consacré le salon Big Data Expo, qui ouvre ses portes mercredi 16 octobre à Paris. En français, cela ne signifie pas grand-chose. Et pourtant, ce n'est rien de moins que la rencontre de l'infiniment grand avec le tout petit. La récupération, le traitement et l'analyse d'une masse démesurée d'informations au profit d'un objectif très ciblé. Une start-up de Boston, Recorded Future, propose ainsi d'analyser les commentaires sur l'avenir jetés sur le Web pour prédire le futur. Elle commercialise déjà ses services à des gouvernements et des grands groupes. Encore un peu et la « psychohistoire », la science-fiction décrite par le célèbre écrivain américain Isaac Asimov dans *Le Cycle de Fondation*, deviendrait une réalité.

Tout cela est rendu possible grâce aux données que laissent derrière eux la nature, les hommes, les objets. C'est surtout la dernière étape en date des progrès de l'informatique. Le Big Data est synonyme de la numérisation du monde. La transcription de la réalité en nombres, le langage informatique par excellence. On peut désormais stocker et faire « mouliner » d'immenses bases de données dans de gigantesques usines de serveurs, les *data centers*. Le créneau est devenu un filon en or pour les mastodontes de l'octet, qui se battent à coups de milliards de dollars pour se positionner sur ce nouvel eldorado.

Car les paysans sont loin d'être les seuls concernés par l'analyse des données. Les publicitaires, les chercheurs, les fonctionnaires, les assureurs, les logisticiens, les espions

> Le Big Data est synonyme de la numérisation du monde. La transcription de la réalité en nombres, le langage informatique par excellence

ou les commerçants ne le savent peut-être pas encore, mais leur monde est en train de changer. Autant de clients à séduire pour les petits nouveaux et les grands anciens de l'informatique.

Comme pour toute nouveauté, il s'agit de tirer au clair les avantages apportés et les inconvénients associés. Côté pile, le potentiel est imposant. L'analyse des gros volumes de données peut devenir d'une efficacité sans faille, ou presque. Elle permettra par exemple de proposer à un internaute une réclame ciblée ou à un patient un traitement sur mesure.

Mais le côté face est bien chargé aussi. La préservation de la vie privée de chacun est la grande question qui plane au-dessus de l'analyse des données. Quelle intimité pour le particulier, si sa manière de manger, de dormir, de dépenser ou de se déplacer devient accessible à n'importe quel statisticien ? Le problème est de taille, puisque, en matière de numérique, la législation évolue toujours avec un train de retard sur les pratiques des entreprises. C'est le cas de la fiscalité. Et le respect des libertés individuelles ne fait pas exception.

Pour faire leur trou, les « grosses données » devront éviter de se brûler les ailes sur ce point crucial. Les méthodes d'anonymisation et de protection des informations personnelles, voilà un autre champ très lucratif à défricher. ■

JULIEN DUPONT-CALBO

Les trois V

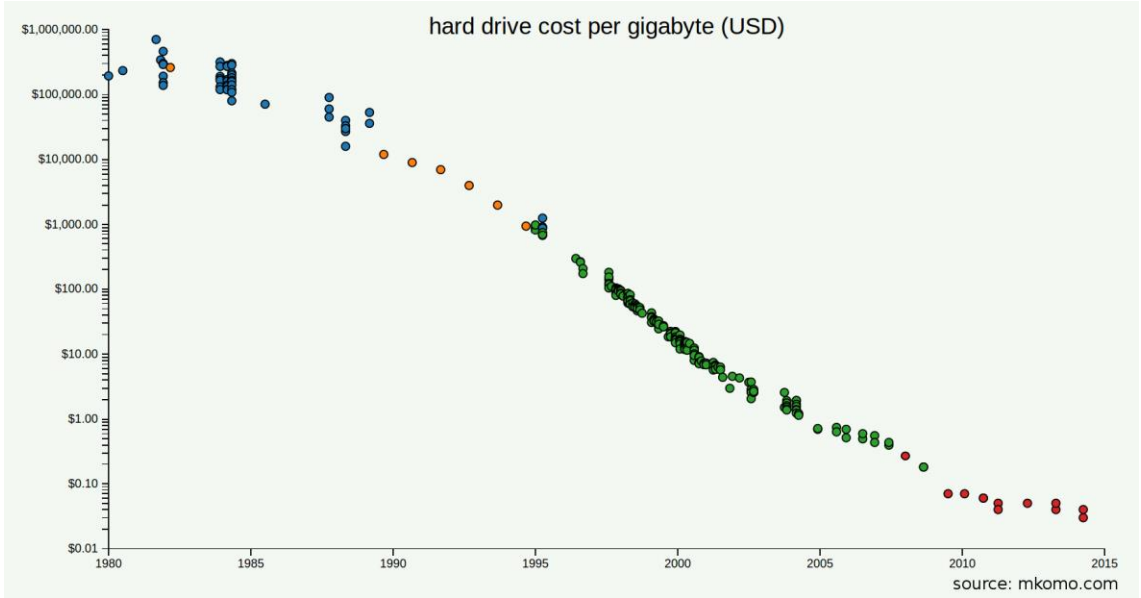
- **Volume** : 90 % de la masse d'information créée depuis le début de l'humanité l'a été ces deux dernières années. Intel prédit qu'il y aura 200 milliards d'objets connectés dans le monde d'ici à 2020. Google traite plus de 24 petabytes par jour
- **Vélocité** :La masse des données créées par les entreprises, les machines et les particuliers augmente chaque année de 40 %. Chaque minute, 200 millions de mails sont envoyés, 400 millions de tweets sont postés chaque jour. 10 millions de photos sont chargées chaque heure sur Facebook.
- **Variété**: données numériques, catégorielles, graphes (réseaux sociaux), textes, videos, etc.

- Big Data: collecte automatique, le plus souvent passive

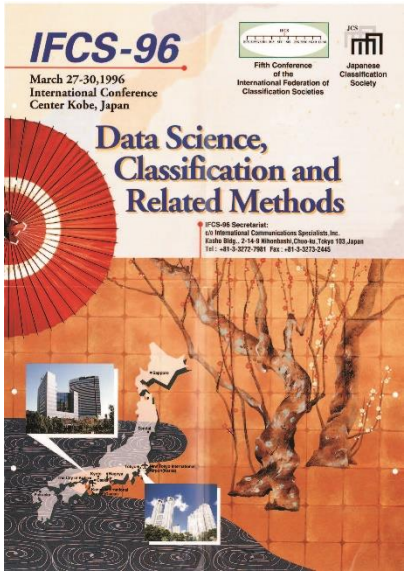
- Par opposition aux méthodes traditionnelles:
  - Enquêtes par sondage
  - Recensements
  - Plans d'expériences

Technologie et Big Data

- Bases de données immenses pour l'apprentissage
- Couts de stockage décroissants
- Processeurs graphiques GPU



- Big Data apparait pour la première fois en 1997:
  - Cox & Ellsworth (NASA, pas NSA!) « Managing Big Data for Visualisation » *ACM SIGGRAPH '97*
- Data Science: Peter Naur 1960
  - Journal of Data Science 2003



2. Des applications

2.1 Business

- Marketing ciblé
- Relances
- Analyse prédictive. « churn »

Les systèmes de recommandation

- Collecte
  - Passive
  - Active : note de * à *****
- Recommandation
  - Personnalisée: passé du client
  - Objet: profils similaires (mots-clés, genre, …)
  - Sociale: comportements d'utilisateurs semblables
  - Hybride : Amazon utilise les 3 précédentes

2.2 Assurance

- Voiture ou conducteur connecté et « pay as you drive » . Au-delà de l'assurance au kilomètre
- Assurance santé et « quantified self » (mesure de soi)
- Individualisation des primes versus mutualisation
- Risque de discrimination

2.3 Santé

Le Monde

L'ouverture des données de santé, nouvelle frontière de l'innovation

▸ Mardi 7 avril, l'Assemblée nationale va débattre de l'article 47 du projet de loi de santé qui touche à l'ouverture publique des données de la « Sécu »

▸ Feuilles de soins, dossiers d'hospitalisation, informations médico-sociales... Dix-sept milliards de données sont concernées

▸ La mise à disposition publique de ce « big data » médical fait consensus après avoir longtemps divisé les professionnels du secteur de la santé

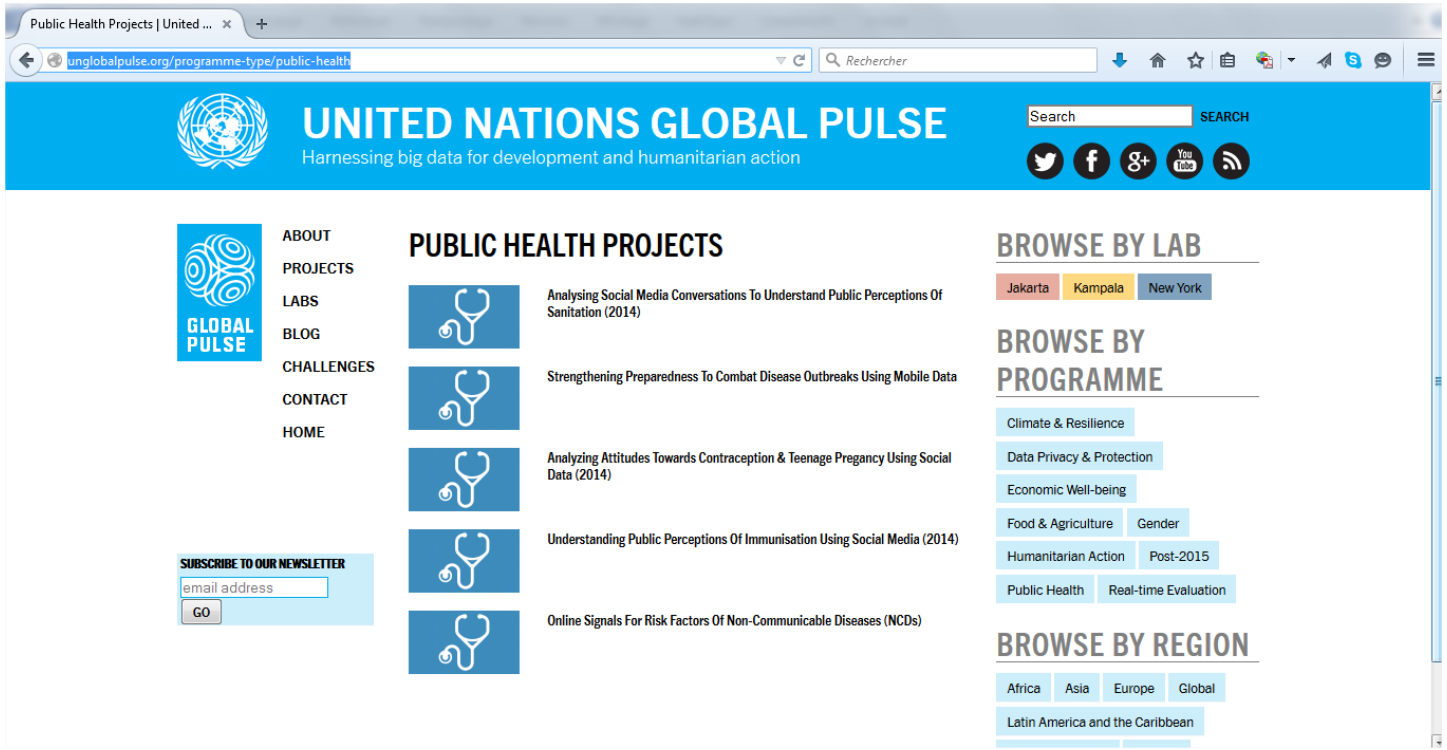
▸ A terme, le croisement des données publiques et privées pourrait améliorer les services rendus aux malades

→ LIRE PAGES 6 ET 7

Médecine personnalisée, détection de risques, …

La France, pays centralisé de 65 millions d'habitants dispose aujourd'hui avec l'ensemble SNIIRAM-PMSI d'une des plus grandes, voire de la plus grande base médico-administrative au monde : elle regroupe par an 1,2 milliard de feuilles de soins , 500 millions d'actes médicaux et 11 millions de séjours hospitaliers (en médecine, chirurgie et obstétrique), avec une profondeur historique de près de 20 ans

Le SNIIRAM, le PMSI et le BCMD (causes de décès) sont regroupés dans le **SNDS** Système National des Données de Santé



UN Global Pulse is a United Nations innovation initiative working to discover and mainstream applications of **big data** and **artificial intelligence** for sustainable development, humanitarian action, and peace.

http://unglobalpulse.org/programme-type/public-health

2.4 Internet des objets

- Maintenance prédictive grâce aux capteurs connectés
- Optimisation de la gestion du réseau d'électricité avec les compteurs connectés
- Smart cities
- Agriculture connectée
- ….

2.5 Statistique officielle et Big Data

- Quelques exemples
  - Données de téléphonie mobile (tourisme, mobilité, pauvreté, crime)
  - Collecte de prix sur le web et aux caisses des magasins
  - Offres d'emploi et taux de chomage
  - Compteurs électriques et occupation des logements
- Avantages:
  - Rapidité
  - Économies

- Inconvénients
  - Absence de contrôle sur la production des données
  - Manque de vérité de terrain
  - qualité et précision variables
    - Capteurs, caméras, téléphonie
    - Réseaux sociaux, e-commerce
  - Pérennité; public-privé
  - Risque de dégradation de l'image des INS
- Nécessité de protéger la confidentialité
  - Risque de réidentification
  - éthique

3. Des défis pour la statistique

4 | Le Monde Mercredi 29 janvier 2014 | SCIENCE & MÉDECINE | ÉVÉNEMENT

« Big data »

Trois défis pour les maths

INFORMATIQUE

L'analyse de grandes masses de données pour en tirer des informations pertinentes est un domaine en pleine expansion. Les « data scientists » doivent imaginer de nouveaux algorithmes pour maîtriser les volumes, la vitesse et la variabilité de ce déluge numérique

- 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
    - Modéles usuels rejetés
    - Intervalles de confiance de longueur nulle

Les deux cultures

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

- The generative modelling culture
  - seeks to develop stochastic models which **fits** the data, and then make inferences about the data-generating mechanism based on the structure of those models. Implicit (…) is the notion that there is a **true model** generating the data, and often a truly `best' way to analyze the data.
- The predictive modelling culture
  - is silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets. **Machine Learning** is identified by Breiman as the epicenter of the Predictive Modeling culture.

From Donoho, 2015

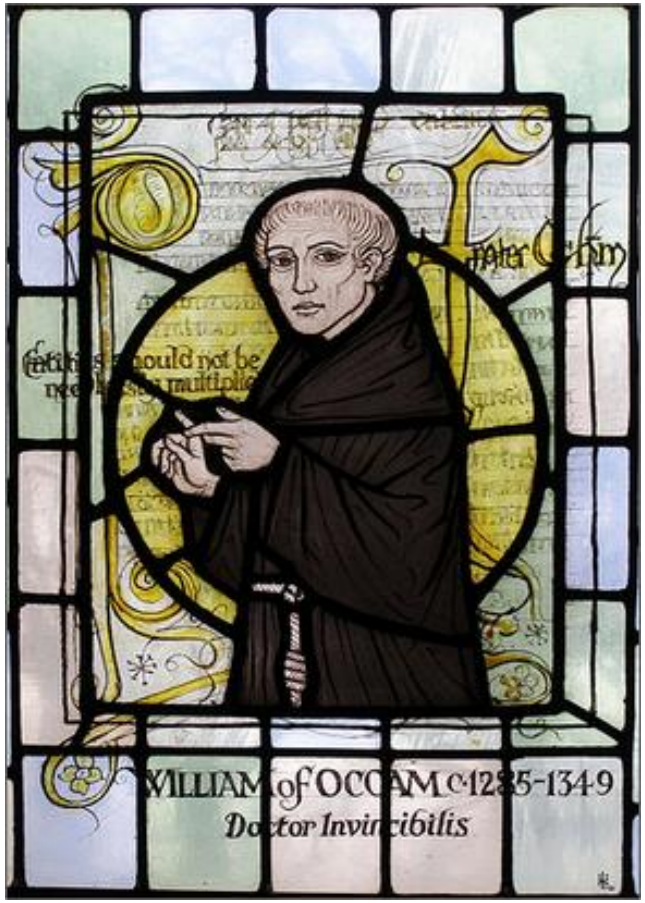
- Conception standard (modèles pour comprendre)
  - Fournir une certaine compréhension des données et de leur mécanisme générateur à travers une représentation parcimonieuse .
  - Un modèle doit être simple et ses paramètres interprétables pour le spécialiste : élasticité, odds-ratio, etc.
- En « Big Data Analytics » (modèles pour prédire)
  - Pour de nouvelles observations: généralisation
  - Les modèles ne sont que des algorithmes

Cf GS, compstat 2008

Même formule: $y = f(x;\theta)+\varepsilon$

- **Modélisation explicative**
  - Théorie sous-jacente
  - Ensemble restreint de modèles
  - Qualité d'ajustement: prédire le passé
  - Erreur: bruit blanc
- **Modélisation prédictive**
  - Modèles issus des données
  - Modèles algorithmiques
  - Prédire de nouvelles données: prédire l'avenir
  - Erreur: à minimiser

Parcimonie et complexité



- Le rasoir d'Ockham
  - *pluralitas non est ponenda sine necessitate*
  - Un principe scientifique : éviter des hypothèses inutiles

- AIC, BIC et autres vraisemblances pénalisées sont souvent considérées comme des versions modernes du rasoir d'Ockham

  $$AIC = -2\ln(L) + 2K$$

  $$BIC = -2\ln(L) + K\ln(n)$$

- Une similarité trompeuse : **AIC et BIC issus de théories différentes**
  - AIC : approximation de la divergence de Kullback-Leibler entre la vraie distribution f et le meilleur choix dans une famille paramétrée
  - BIC : choix bayesien parmi des modèles paramétriques a priori équiprobables
- **Illogique de les utiliser simultanément**

Comparaison AIC BIC

- L'AIC est un critère prédictif tandis que le BIC est un critère explicatif.
- Si *n* tend vers l'infini la probabilité que le BIC choisisse le **vrai** modèle tend vers 1, ce qui est faux pour l'AIC.
- Pour *n* fini: résultats contradictoires. BIC ne choisit pas toujours le vrai modèle: il a tendance à choisir des modèles trop simples en raison de sa plus forte pénalisation

AIC BIC réalistes?

- Vraisemblance pas toujours calculable.
- Nombre de paramétres non plus: ridge, PLS, arbres, etc.
- « **Vrai** » modèle?

  “Essentially, all models are wrong, but some are useful ”
  (G.Box,1987)

* Box, G.E.P. and Draper, N.R.:
Empirical Model-Building and Response Surfaces, p. 424, Wiley, 1987

- « Occam’s Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately in prediction, accuracy and simplicity (interpretability) are in conflict » Breiman, 2011

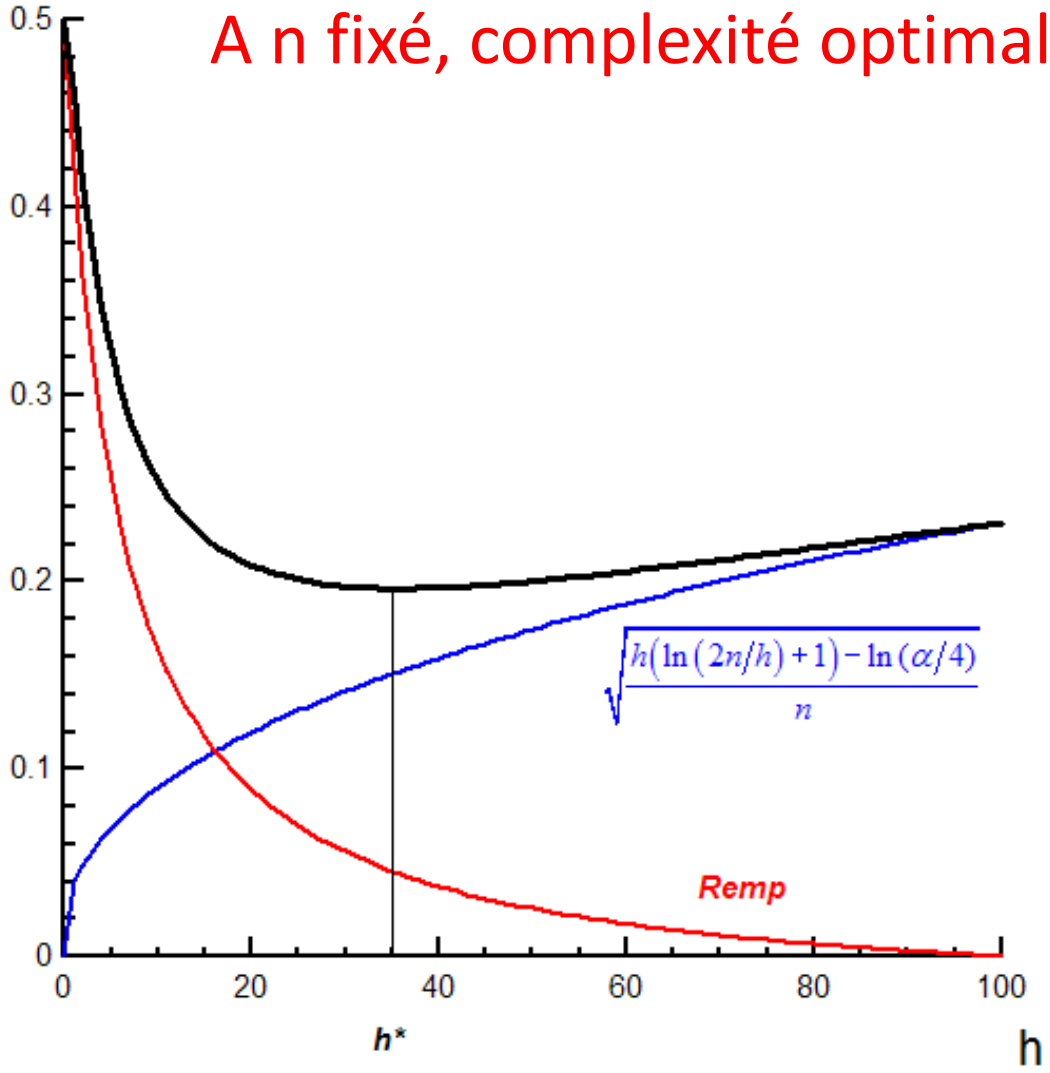
- Pénalisation en contradiction avec la théorie de l'apprentissage
- L'inégalité de Vapnik

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

dépend de n/h d'où ces conséquences :

- On améliore la capacité prédictive si la complexité h croit mais moins vite que n
- On peut augmenter la complexité du modèle avec n

A n fixé, complexité optimale h*



- Les meta-modèles ou méthodes d'ensemble
  - Bagging, boosting, **random forests** améliorent les algorithmes élémentaires
  - Idem pour le **stacking** (Wolpert, Breiman) qui combine linéairement les prévisions de modèles de toutes sortes: linéaires, arbres, ppv, réseaux de neurones etc.

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

$$\hat{y} = \sum_{j=1}^{m} w_j \hat{f}_j(\mathbf{x})$$

    - Avec des poids positifs de somme 1: version fréquentiste du Bayesian Model Averaging

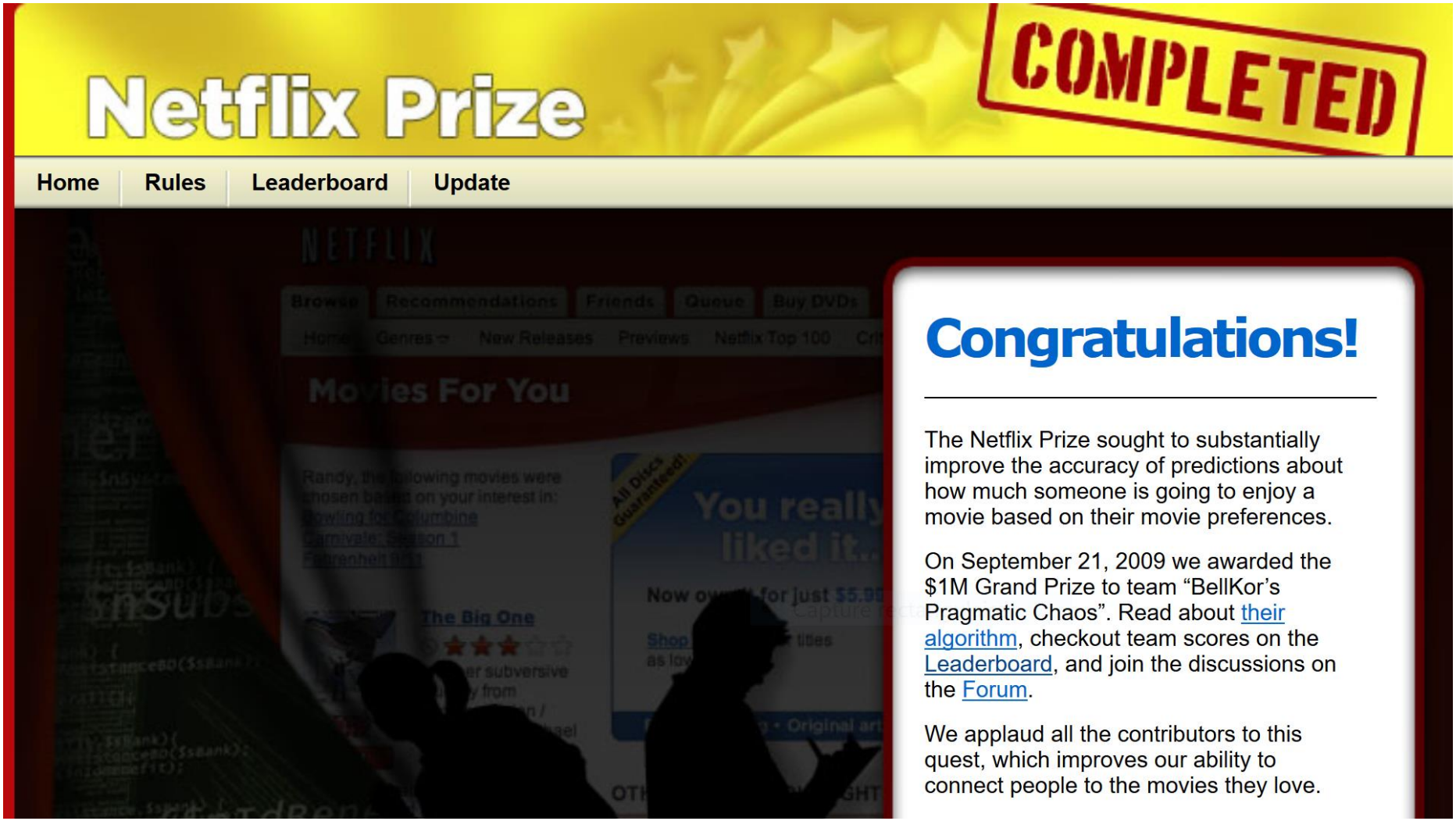
- Première idée: poids obtenus par mco:

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

  - favorise les modèles les plus complexes: surapprentissage
- Solution: utiliser les valeurs prédites par leave one out

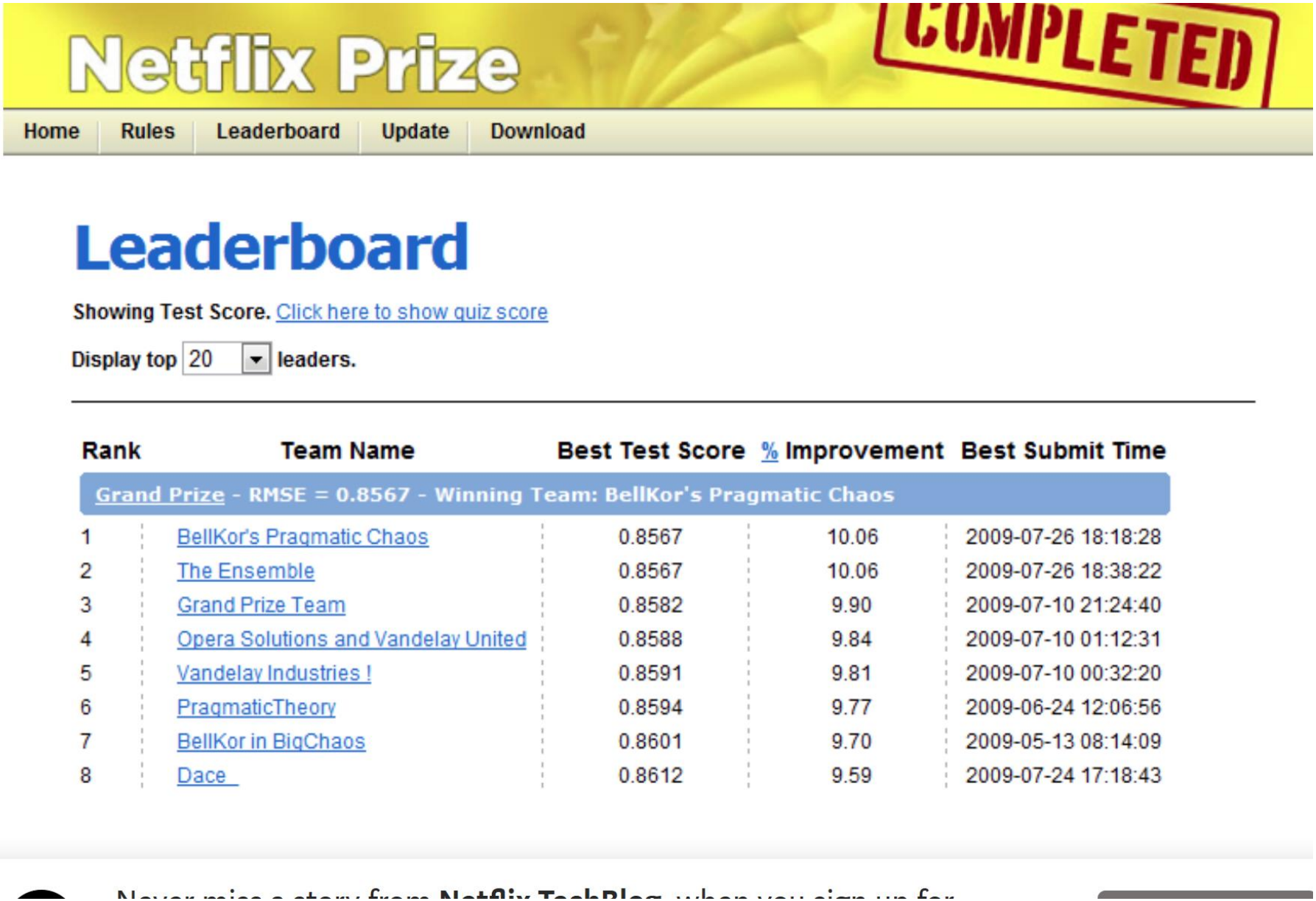
$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j^{-i}(\mathbf{x}) \right)^2$$

- Améliorations : (Noçairi et al., 2016)
  - Régression PLS ou autre méthode régularisée car les m prévisions sont très corrélées
- Empiriquement le stacking surpasse sur de nombreux cas le BMA avec des calculs bien plus simples (Clarke, 2003)



We offered $1 million to whoever improved the accuracy of our existing system called *Cinematch* by 10%. The race was on to beat our RMSE of 0.9525 with the finish line of reducing it to 0.8572 or less

- The Netflix dataset contains more than 100 million 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.
- We offered $1 million to whoever improved the accuracy of our existing system called *Cinematch* by 10%. The race was on to beat our RMSE of 0.9525 with the finish line of reducing it to 0.8572 or less


Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Download
Leaderboard
Showing Test Score. Click here to show quiz score
Display top 20 leaders.
Rank
Team Name
Best Test Score
% Improvement
Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos
1 BellKor's Pragmatic Chaos 0.8567 10.06 2009-07-26 18:18:28
2 The Ensemble 0.8567 10.06 2009-07-26 18:38:22
3 Grand Prize Team 0.8582 9.90 2009-07-10 21:24:40
4 Opera Solutions and Vandelay United 0.8588 9.84 2009-07-10 01:12:31
5 Vandelay Industries ! 0.8591 9.81 2009-07-10 00:32:20
6 PragmaticTheory 0.8594 9.77 2009-06-24 12:06:56
7 BellKor in BigChaos 0.8601 9.70 2009-05-13 08:14:09
8 Dace 0.8612 9.59 2009-07-24 17:18:43

- 44 014 valid submissions from 5 169 different teams
- *BellKor's Pragmatic Chaos team*. A blend of hundreds of different models
- *The Ensemble Team* . Blend of 24 predictions

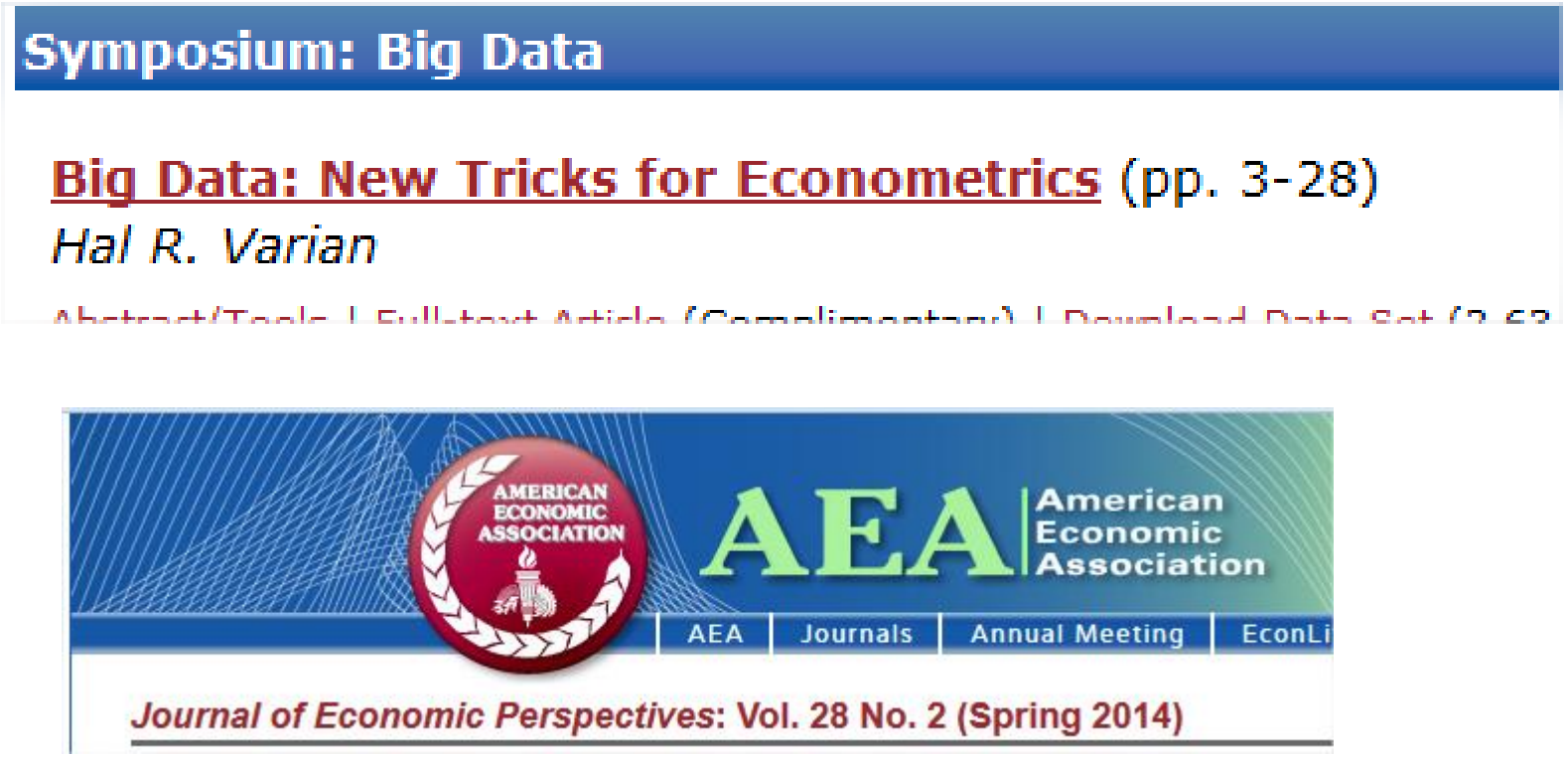
  Same Test RMSE : 0.8567 (10.06%)
- Bellkor's Pragmatic Chaos defeated The Ensemble by submitting just 20 minutes earlier!

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

- Et pour les économétres…



- "Data manipulation tools and techniques developed for small datasets will become increasingly inadequate to deal with new problems. Researchers in machine learning have developed ways to deal with large datasets and economists interested in dealing with such data would be well advised to invest in learning these techniques."

4. Valider et interpréter

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

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

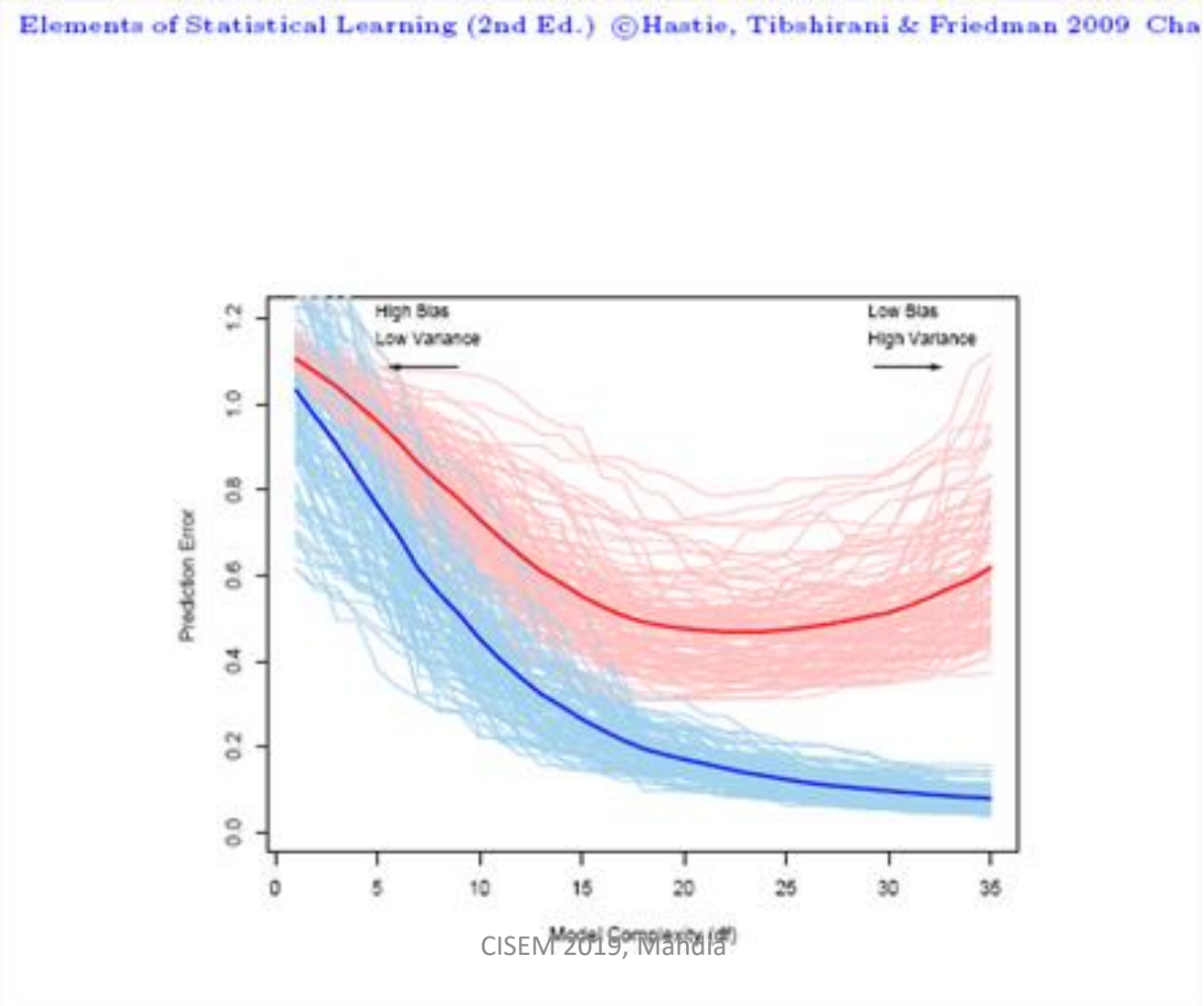
- Apprentissage: pour estimer les paramètres des modèles
- Test : pour choisir le meilleur modèle
- Validation : pour estimer la performance sur des données futures
  - **Estimer les paramètres ≠ estimer la performance**
  - Réestimation du modèle final**:** **avec toutes les données disponibles**

- Précurseurs:
  - Paul Horst (1903-1999)

    « *the usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established*» 1941
  - Leave one out : Lachenbruch et Mickey, 1968
  - Validation croisée: Stone, 1974

- **Elémentaire?**
  - Pas si sur…
  - Voir publications en économétrie, épidémiologie, .. prédictions rarement validées sur des données « hold-out » (sauf en prévision de séries temporelles)

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



Paradigmes et paradoxes

- Comprendre sans prédire
  - Un « bon » modèle qui s'ajuste bien peut fournir des prévisions médiocres au niveau individuel (eg épidémiologie)
- Prédire sans comprendre
  - Des modèles ininterprétables (eg *deep learning*) peuvent donner de bonnes prévisions (ciblage marketing, …)

Le paradigme de la boîte noire

- modèle génératif y=f(x)+ ε . On cherche une fonction qui approxime en un certain sens le comportement de la boîte noire.
- Deux conceptions très différentes :
  - soit on cherche à approximer la vraie fonction f,
  - soit on cherche à obtenir des prévisions de y aussi précises que possible.

– Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data (Breiman, 2001).

– Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms (Vapnik, 2006).

- Garder (ou pas) des variables significatives ou non?
  - A researcher might choose to retain a causal covariate which has a strong theoretical justification *even if is statistically insignificant*.
  - statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, *even if they are statistically significant*, results in improved prediction accuracy

(Shmueli, 2010)

- Le modèle « vrai » ne prédit pas toujours mieux
  - the underspecified linear regression model that leaves out q predictors has a lower EPE when the following inequality holds:

$$q\sigma^2 > \beta_2' X_2' (I - H_1) X_2 \beta_2.$$

- when the data are very noisy (large σ);
- when the true absolute values of the left-out parameters (in our example β_2) are small;
- when the predictors are highly correlated; and
- when the sample size is small or the range of left-out variables is small.

(Shmueli, 2010)

Interpréter les modèles

- On croit souvent que les modèles simples (régression linéaire, logistique) s'interprètent aisément
- C'est loin d'être le cas!
- Sauf dans des dispositifs orthogonaux, la valeur des paramètres ne reflète que rarement l'importance des variables

- Plus de 11 méthodes pour quantifier l'importance des variables en régression linéaire! (Grömping, 2015, Wallard, 2015)
  - Eg Shapley value: un sous-ensemble de prédicteurs vu comme une coalition en théorie des jeux
- Encore des idées de Breiman:
  - « A variable might be considered important if deleting it seriously affects prediction accuracy »
  - permutation aléatoire des valeurs « shuffling »
  - Utilisées dans les random forests mais applicable pour tout modèle: ***« model-agnostic methods»***

- Mais :
  - « toutes choses égales par ailleurs» souvent impossible
  - Faire varier un prédicteur (intervention) peut impliquer des variations des autres d'où un effet complexe
  - Nécessité de schémas causaux

6. La fin de la théorie?

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/ »

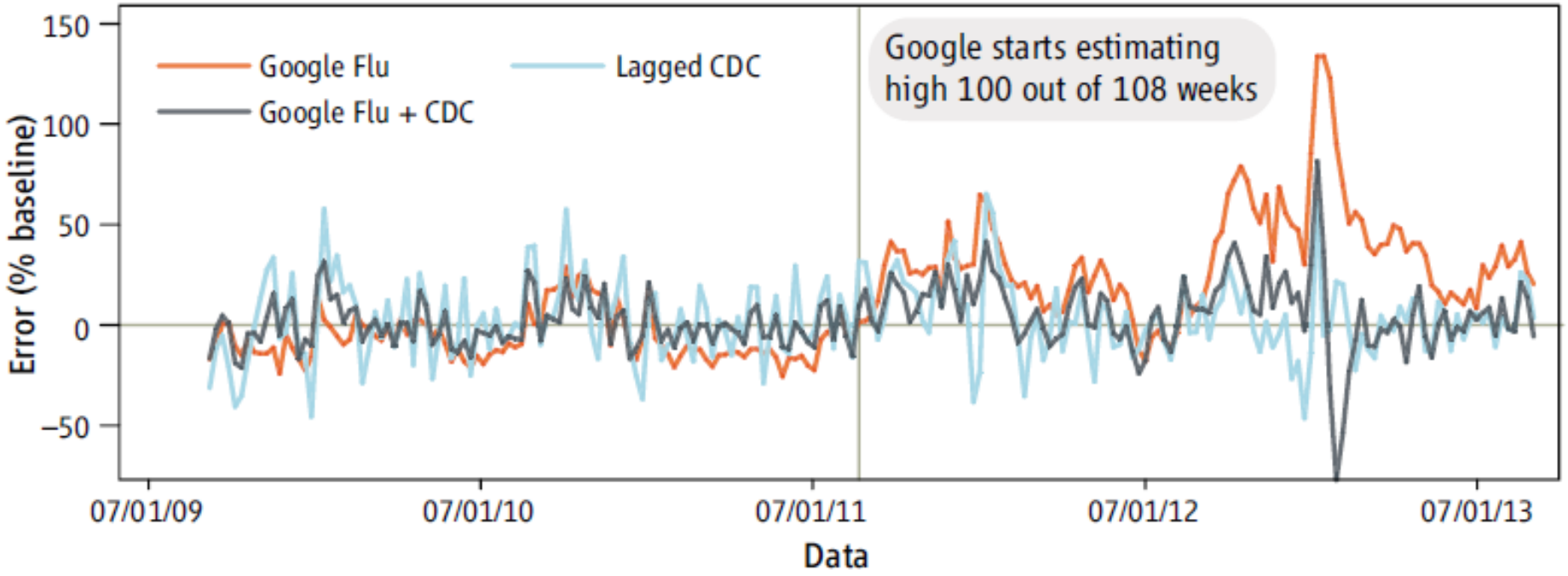
http://esante.gouv.fr/le-mag-numero-10/decryptage-le-big-data-sante

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014



Comprendre pour mieux prédire

- Confusion entre corrélation et causalité
- Difficile d'inférer la causalité à partir de données d'observations.
  - Effet d'un traitement: certains individus ont eu X=1, d'autres non

Inférence causale et raisonnement contrefactuel

- L'identité de base de décomposition du résultat d'un traitement:

$$
\begin{aligned}
&\text{Outcome for treated} - \text{Outcome for untreated} \\
&= [\text{Outcome for treated} - \text{Outcome for treated if not treated}] \\
&\quad + [\text{Outcome for treated if not treated} \\
&\qquad - \text{Outcome for untreated}] \\
&= \text{Impact of treatment on treated} + \text{selection bias.}
\end{aligned}
$$

- Partie contrefactuelle : « Outcome for treated if not treated"

- Problème: comment savoir ce qui serait arrivé aux individus traités s'ils ne l'avaient pas été!
- Estimer la partie contrefactuelle
  - Inférence causale de Rubin 1974
  - Propensity score matching de Rosenbaum et Rubin 1983
  - Pearl 2000

Le retour de l'expérimentation

- As Box et al. put it, "To find out what happens when you change something, it is necessary to change it." … the best way to answer causal questions is usually to run an experiment. (Varian, 2016)
- L'identité de base montre l'intérêt des essais randomisés: le biais de sélection est alors d'espérance nulle, d'où la possibilité d'estimer l'impact causal .
  - A/B testing, Marketing, publicité sur le web (Bottou,2013)

Arthur M. Sackler COLLOQUIA

Drawing Causal Inference from Big Data

This meeting was held March 26-27, 2015 at the National Academy of Sciences 2101Constitution Ave. NW in Washington, D.C.

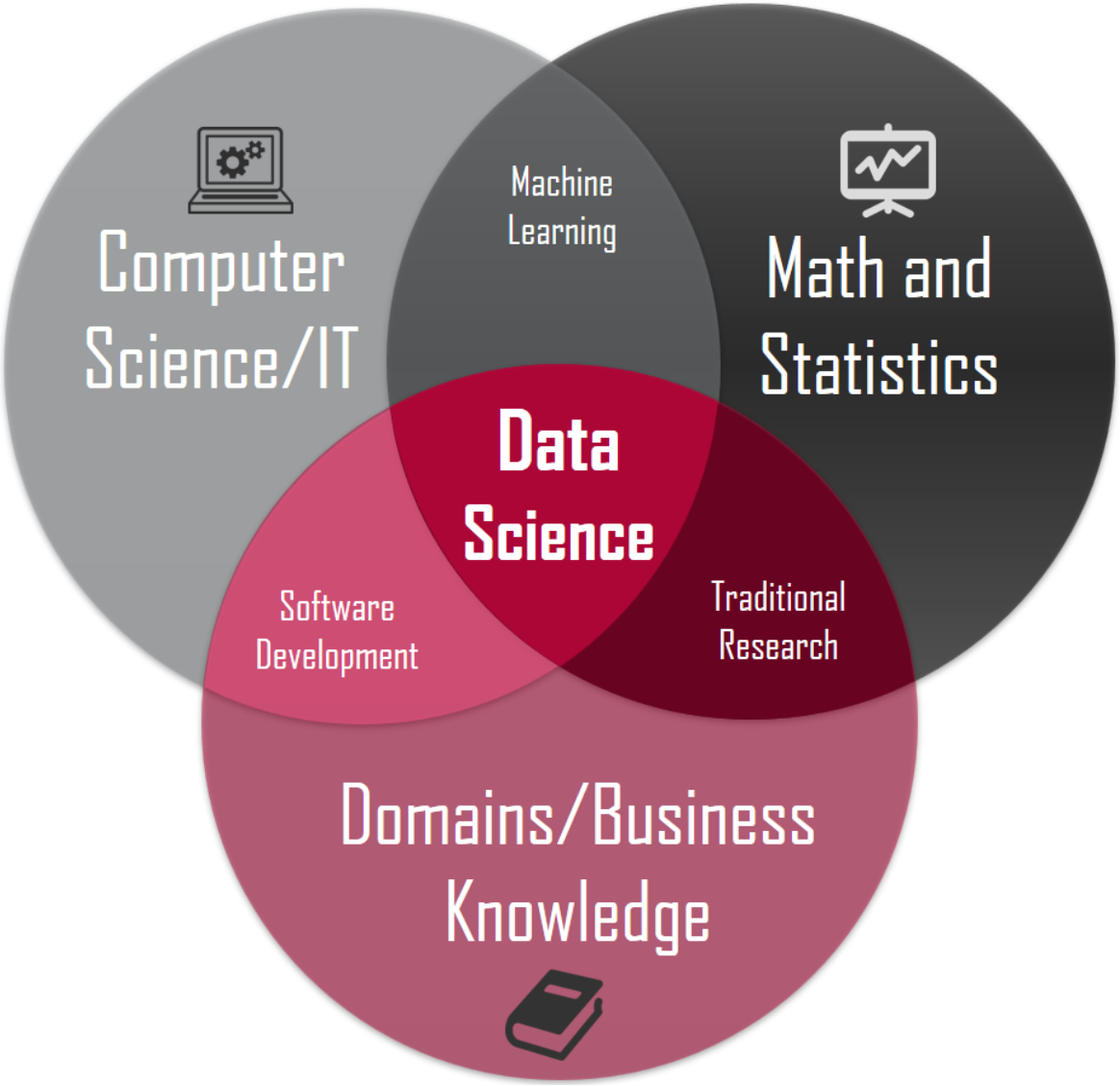
Organized by Richard M. Shiffrin (Indiana University), Susan Dumais (Microsoft Corporation), Mike Hawrylycz (Allen Institute), Jennifer Hill (New York University), Michael Jordan (University of California, Berkeley), Bernhard Schölkopf (Max Planck Institute) and Jasjeet Sekhon (University of California, Berkeley)

Graduate Student / Postdoctoral Researcher travel awards sponsored by the National Science Foundation and the Ford Foundation.

July 5, 2016
vol. 113 no. 27

6. Les Data Scientists

- Les spécialistes du Big Data Analytics
- Compétences:
  - Mathématique, Statistique, Machine Learning
  - Informatique: bases de données réparties, calcul parallèle, algorithmique et codage
  - Compréhension du métier et sens (feeling) des données



https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

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."

- **Data Scientist est le job numéro 1 du classement Glassdoor 2017**
- Tous les ans, le site Glassdoor établit un classement des meilleurs jobs du monde en fonction du salaire moyen, du nombre d'offres d'emplois et du taux de satisfaction des personnes qui exercent ces professions.
- En 2017, le job numéro 1 est celui de Data Scientist avec un taux de satisfaction de 4,4/5 et un salaire médian de 110 000 dollars

DATA SCIENTIST

Le *data scientist* est expert en mathématiques, statistique et informatique. Ce métier d'explorateur de la donnée relève, selon les profils – plutôt orientés mathématique ou informatique –, de la recherche, du développement d'algorithmes, voire de l'industrialisation de solutions.

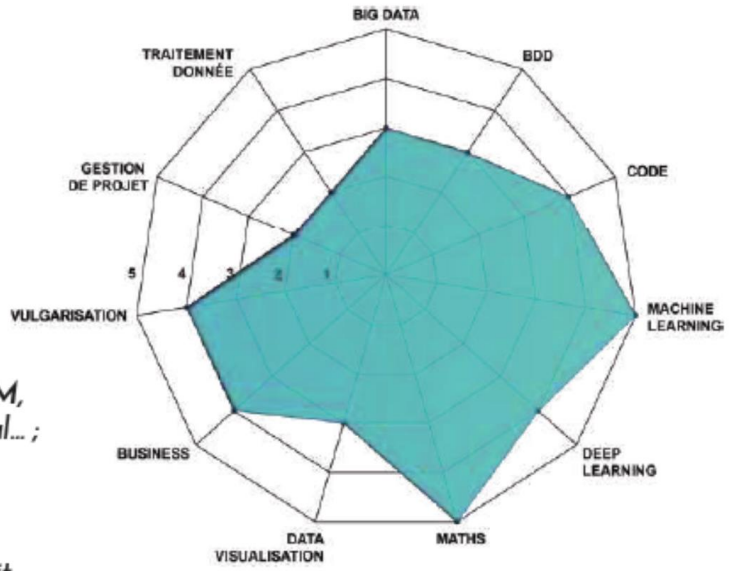
L'EXPLORATEUR DE DONNÉES COMPLEXES

Au quotidien :

- *Explorer de nouvelles pistes dans les ensembles de données et tester des hypothèses*
- *Créer de nouveaux modèles de données*
- *Concevoir des algorithmes à base de machine learning et les implémenter*

Compétences :

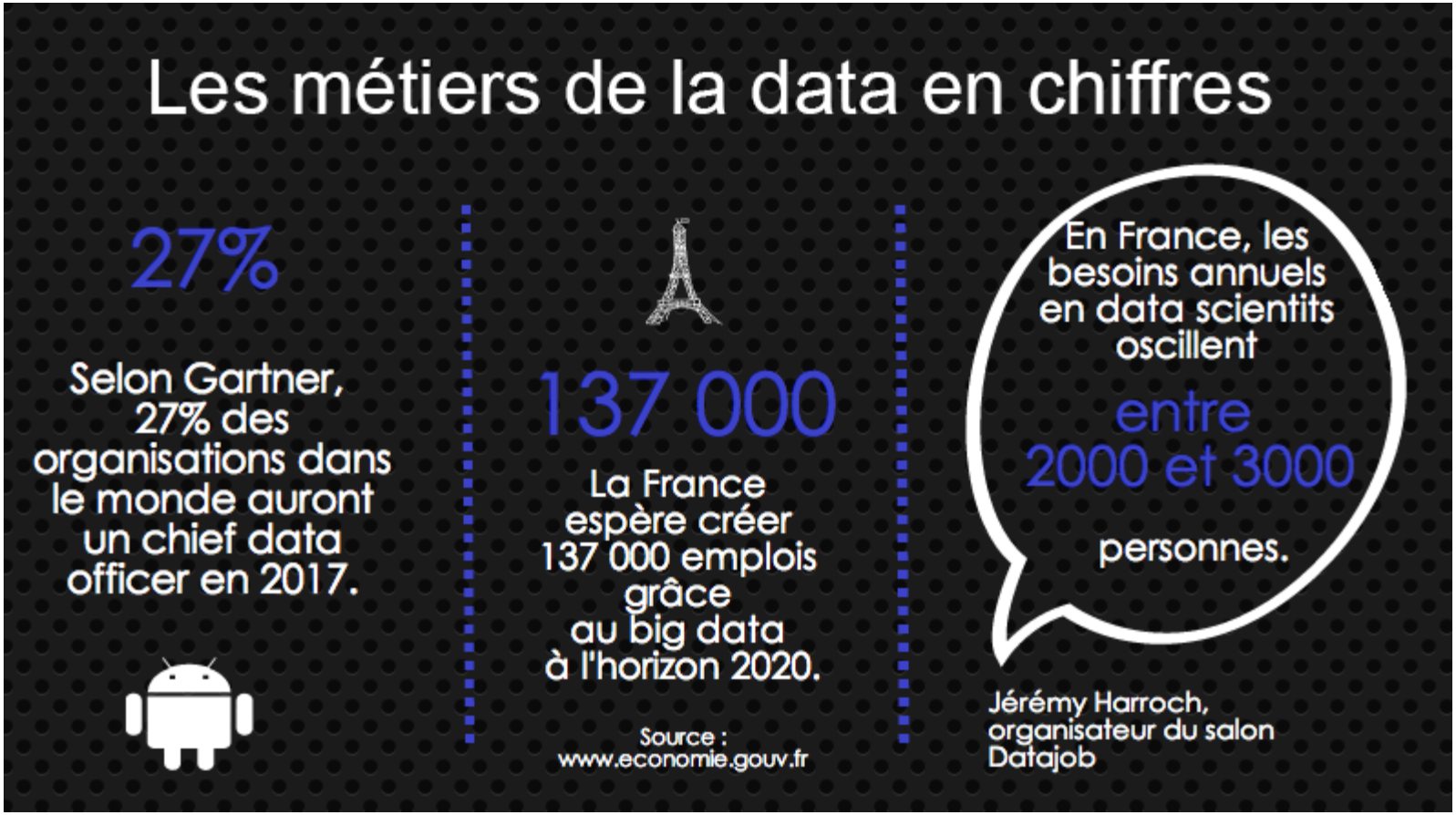
- *Maîtrise du machine learning, deep Learning, Random Forests, modèles de Markov cachés, SVM, régression, séries temporelles, traitement du signal… ;*
- *Maîtrise des outils Big data ;*
- *Maîtrise de Python, R , Java, C, C++, Matlab, écosystème Hadoop, Spark… ;*
- *Esprit d'équipe, communication, ouverture d'esprit, créativité, curiosité, transdisciplinarité ;*
- *Sensibilité aux enjeux business, notamment dans les secteurs comme le marketing, le web, la publicité…*

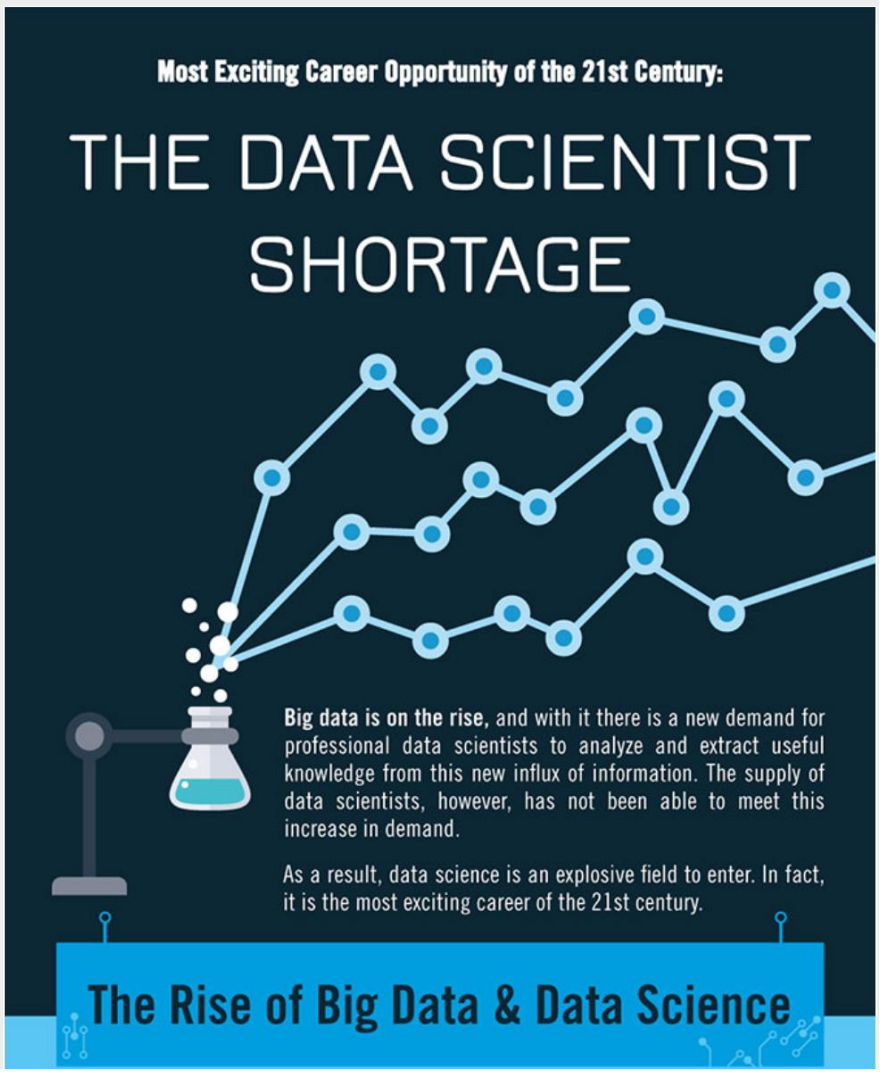


Formation :

- *Masters ou thèses en informatique, sciences cognitives, statistiques, mathématiques… ;*
- *Bac + 4-5 en mathématiques-statistiques ou informatique ;*
- *Formation big data ;*
- *Expérience professionnelle (Data Analyst).*

https://dataanalyticspost.com/wp-content/uploads/2019/03/Guide_metiers_data_DAP.pdf


Les métiers de la data en chiffres
27%
Selon Gartner, 27% des organisations dans le monde auront un chief data officer en 2017.
137 000
La France espère créer 137 000 emplois grâce au big data à l'horizon 2020.
Source : www.economie.gouv.fr
En France, les besoins annuels en data scientits oscillent
entre 2000 et 3000
personnes.
Jérémy Harroch, organisateur du salon Datajob



https://insidebigdata.com/2018/08/19/infographic-data-scientist-shortage

La pénurie de talents aux USA

- In 2015, there was a national surplus of people with data science skills.
- But today, 3 years later, the picture has changed markedly: data science skills shortages are present in almost every large U.S. city. Nationally, we have **a shortage of 151,717 people** with data science skills, with particularly acute shortages in New York City (34,032 people), the San Francisco Bay Area (31,798 people), and Los Angeles (12,251 people).
- As more industries rely on big data to make decisions, data science has become increasingly important across all industries, not just tech and finance.

https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

- **Europe needs 346,000 more data scientists by 2020, but why is the gap so big?**
- Data science is one area of the digital sector that is desperately short of talent. In fact, IBM thinks data science will account for 28% of all digital jobs by 2020
- "Machine learning, big data and data science skills are the most challenging to recruit for, and can potentially create the greatest disruption if not filled," according to IBM's The Quant Crunch report.
- The European Commission thinks that **100,000 new data-related jobs** will be created in the region by 2020, but the fact there are not enough people with the right skills to fill the role is certainly worrying.

https://www.itpro.co.uk/careers/28929/data-scientist-jobs-where-does-the-big-data-talent-gap-lie

Le défi de la formation (Saporta, 2018)

- Capacités de formation initiale insuffisante malgré la création (?) de masters en Data Science
- Des hommes ou des robots?
  - on one hand the crowd- sourced trained data scientists of Kaggle with its 150,000 members, ready to solve problems and on the other hand the machines represented by IBM Watson... in the end, I predict Artificial Intelligence will win. B.Marr, 2016
- Formation tout au long de la vie: perfectionner les ingénieurs, statisticiens, informaticiens déjà en poste (la mission du CNAM).

7. Conclusion

- Big Data: pas seulement un défi technologique mais aussi un défi méthodologique
- Modèles et algorithmes
- Des jobs par milliers pour les « data scientists »
- Nécessite de nouveaux cursus
- Enjeux éthiques: vie privée, confidentialité

Merci pour votre attention

References

- C.Anderson (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, http://www.wired.com/2008/06/pb-theory/
- L.Bottou et al. (2013) Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, *Journal of Machine Learning Research*, 14, 3207–3260,
- Breiman, L., (1996): Stacked Regressions. *Machine Learning*, 24:49-64
- L.Breiman (2001) Statistical Modeling: The Two Cultures, *Statistical Science*, 16, 3, 199–231
- B.Clarke (2003) Comparing Bayes Model Averaging and Stacking When ModelApproximation Error Cannot be Ignored, *Journal of Machine Learning Research,* 4, 683-712
- D.Donoho (2015) 50 years of Data Science, *Tukey Centennial workshop*, https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf

- U.Grömping, (2015). Variable importance in regression models. *WIREs Computational Statistics,* 7, 137-152.
- H. Noçairi , C. Gomes , M. Thomas , G. Saporta (2016) Improving Stacking Methodology for Combining Classifiers; Applications to Cosmetic Industry, *Electronic Journal of Applied Statistical Analysis*, vol. 9(2), 340-361
- G.Saporta (2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
- G. Saporta (2018) Training data scientists: a few challenges , *International Journal of Data Science and Analytics*, vol. 6(3), pp. 201-204
- G. Shmueli (2010) To explain or to predict? *Statistical Science,* 25**,** 289–310
- V.Vapnik (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
- H.Varian (2014) Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28, 2, 3–28
- H.Varian (2016) Causal inference in economics and marketing, *PNAS* , 113, 7310-7315
- H.Wallard (2015) Using Explained Variance Allocation to analyse Importance of Predictors, *16th ASMDA conference proceedings*, 1043-1054