Apprentissage automatique et méga données

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu

Séminiare ICube 2015, Strasbourg

June 11, 2015

Plan

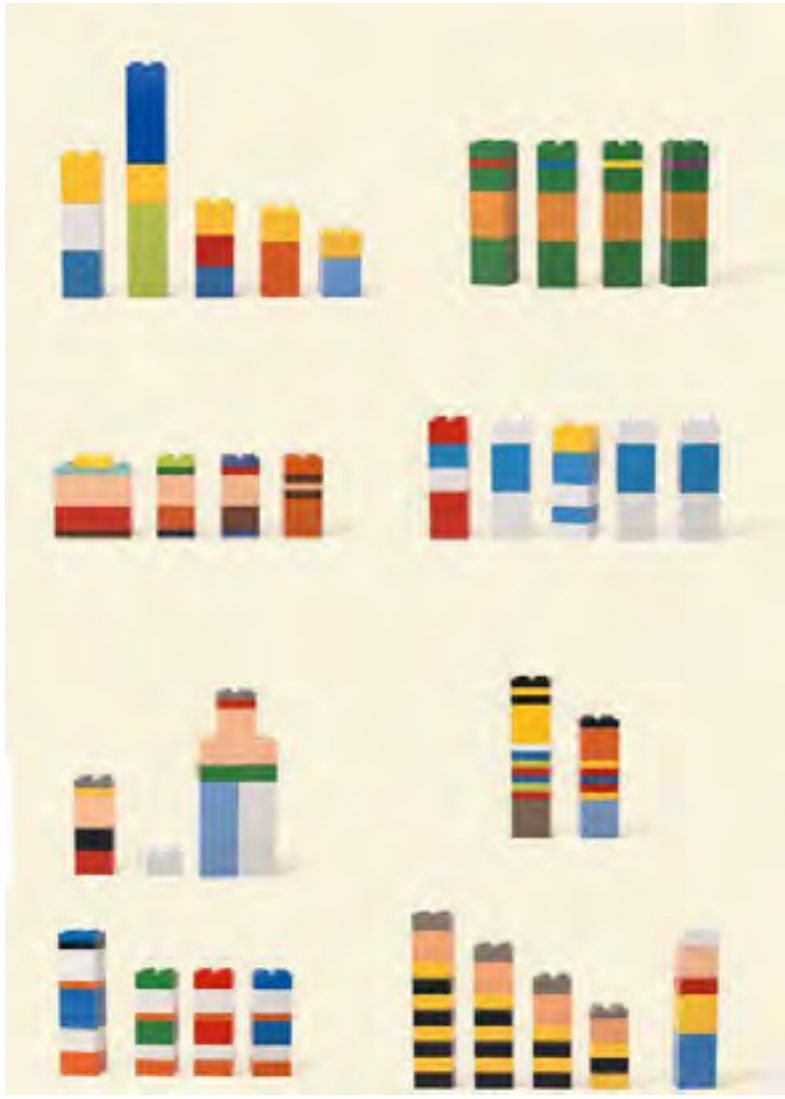
Plan

Apprentissage : humain vs. machine

Les apprentissages d'un enfant

- marcher : un an
- parler : deux ans
- raisonner : le reste

Apprendre à raisonner



Deux définitions de l'apprentissage

Arthur Samuel (1959)

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (The Discipline of Machine Learning, 2006)

"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"

A computer program CP is said to learn from experience E with respect to some class of tasks T and performance measure P, if its **performance at tasks in T, as measured by P, improves** with experience E

Une tâche T « apprenable »

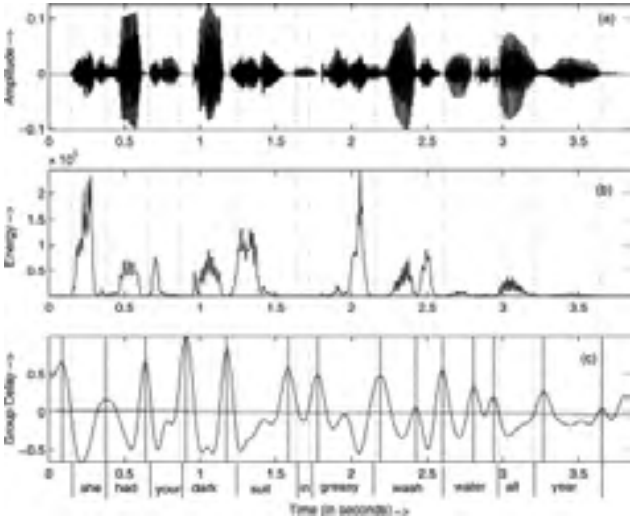
Machine Learning

A computer program CP learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E

Construire un programme CP

- experience E : statistiques
- performance measure P : optimisation
- tasks T : utilité
  - traduction automatique
  - jouer aux échec
  - conduire, ... et faire ce que l'on fait

Les tâches T


Amplitude -->
Energy -->
Group Delay -->
she
had
your
dark
suit
in
greasy
wash
water
all
year
Time (in seconds) -->


A laser sensor scans 360 degrees around the vehicle for objects.
A processor reads the data and regulates vehicle behavior.
Radar measures the speed of vehicles ahead.
An orientation sensor tracks the car's motion and balance.
A wheel-hub sensor detects the number of rotations to help determine the car's location.

L'experience E : les données

capteurs →	variables qualitatives / ordinales / quantitatives
texte →	chaine de caractères
parole →	temps - série temporelle
images/vidéos →	dépendances 2/3 d
réseaux →	graphes
jeux →	séquences d'interactions
flots →	tickets de caisse, web logs, traffic. . .
étiquettes →	information d'évaluation

Les mégadonnées (volume, vitesse, variété, véracité et valeur)

Des données qui arrivent sans qu'on ait décidé de les collecter
→ l'importance des pré-traitements (nettoyage, normalisation, codage. . .)
→ et de la représentation : de la donnée au vecteur

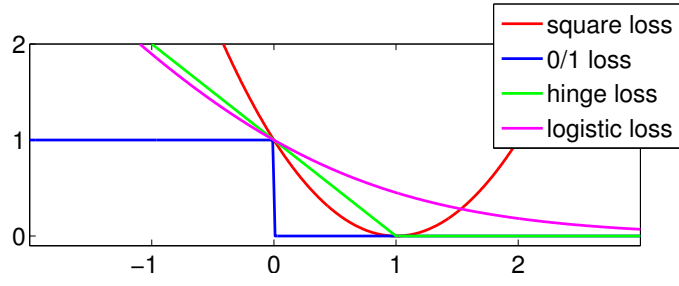
Objectifs, tâches et mesures de performance P

non supervisé	▸ clustering	homogénéité des groupes
supervisé	▸ classification	
	★ bi classe	nombre d'erreur (0/1)
	★ multi classe	
	★ détection	nombre de non détection
	▸ ordonnancement	erreur de rang
	▸ regression	erreur quadratique
renforcement		probabilité de gagner
expliquer		qualité du résumé

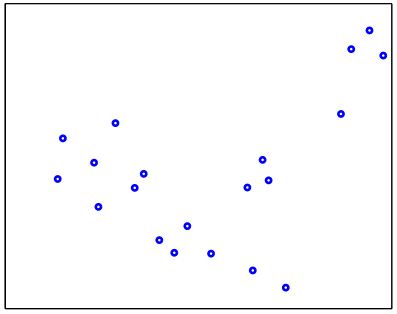
. . . et semi supervisé, transductif, séquentiel, actif. . .

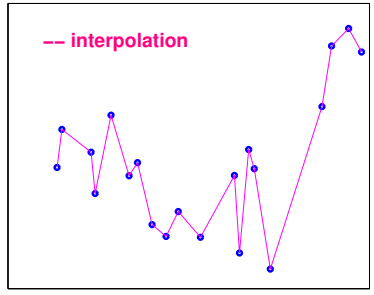
Généraliser

$\rightarrow$ bien faire (minimiser P) sur les futures données qu'on ne connait pas



L'apprentissage comme un problème d'optimisation bi critère





data, interpolation

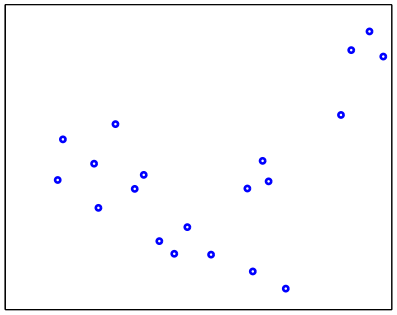
.

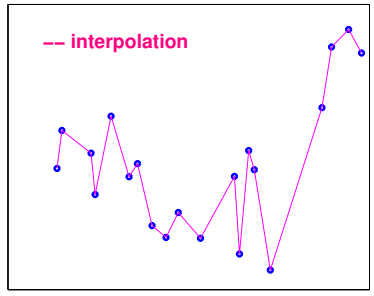
$$\begin{cases} \min\limits_{f \in H} & \text{s'ajuster aux données} \end{cases}$$

Apprendre c'est sélectionner une bonne hypothèse

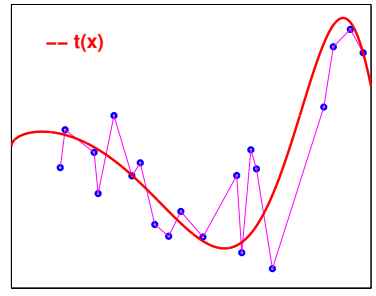
- un bon model H (ensemble d'hypothèses universelles)
  - cout d'ajustement aux données

L'apprentissage comme un problème d'optimisation bi critère





data, interpolation



data, interpolation and the target function

$$\begin{cases} \min\limits_{f \in H} & \text{s'ajuster aux données} \\ \min\limits_{f \in H} & \text{bien généraliser} \end{cases}$$

Apprendre c'est sélectionner une bonne hypothèse

- un bon model H (ensemble d'hypothèses universelles)
- P = deux critères contradictoires
  - cout d'ajustement aux données
  - pénalité favorisant une certaine régularité
- une bonne gestion de cette contradiction

Exemples de model H (ensemble d'hypothèses universelles)

- dictionnaire fixe $f(x) = \sum_{j=1}^{p} \alpha_j \phi_j(x)$
- dictionnaire adapté aux observations $f(x) = \sum_{j=1}^{p} \alpha_j \phi_j(x, x_i)$
  - les machines à noyaux $f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$
- dictionnaire adapté aux observations et aux étiquettes

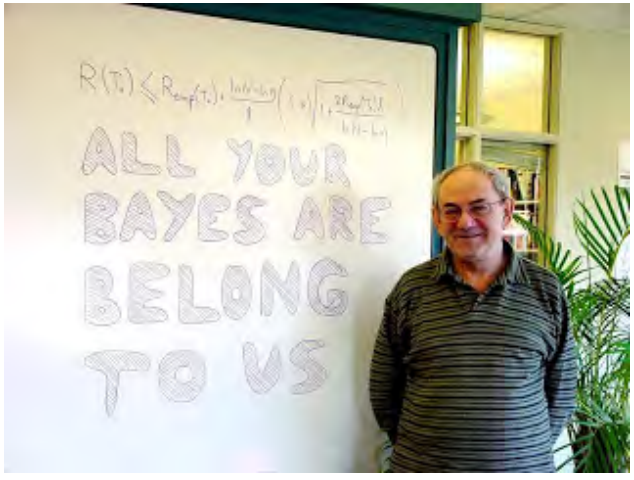
$$f(x) = \sum_{i=1}^{p} \alpha_j \phi_j(x, x_i, y_i)$$

  - les réseaux de neurones de type perceptron multicouche $f(x) = \sum_{k=1}^{p} \alpha_k \Phi \left(\sum_{j=1}^{p} \beta_j \mathbf{x} \right)$
  - méthodes d'ensemble

Les objectifs de l'apprentissage automatique

- Algorithmes d'apprentissage
  - bonne généralisation
  - généricité
  - passage à l'échelle

- Questions théoriques (Vapnik's Book, 1982 - Valiant, 1984)
  - apprenabilité, sous quelles conditions ?
  - complexité (en temps, en échantillon)



Une brève histoire de l'apprentissage artificiel

1940 Etude de la logique en temps qu'objet mathématique (Shannon, Godel et al).

1950 Test de Turing,
1956 Dartmouth conférence : *artificial intelligence*.

1960 *"Within a generation ... the problem of creating 'artificial intelligence' will substantially be solved."*

1960 *Le perceptron*

1970 *AI symbolique : les systèmes experts*

1980 *AI Winter & Réseaux de neurones (perceptron non linéaire)*

1990 *SVM & COLT*

2000 *les applications : Google, Amazon (recommendation), spam filtering, OCR, parole, traduction...*

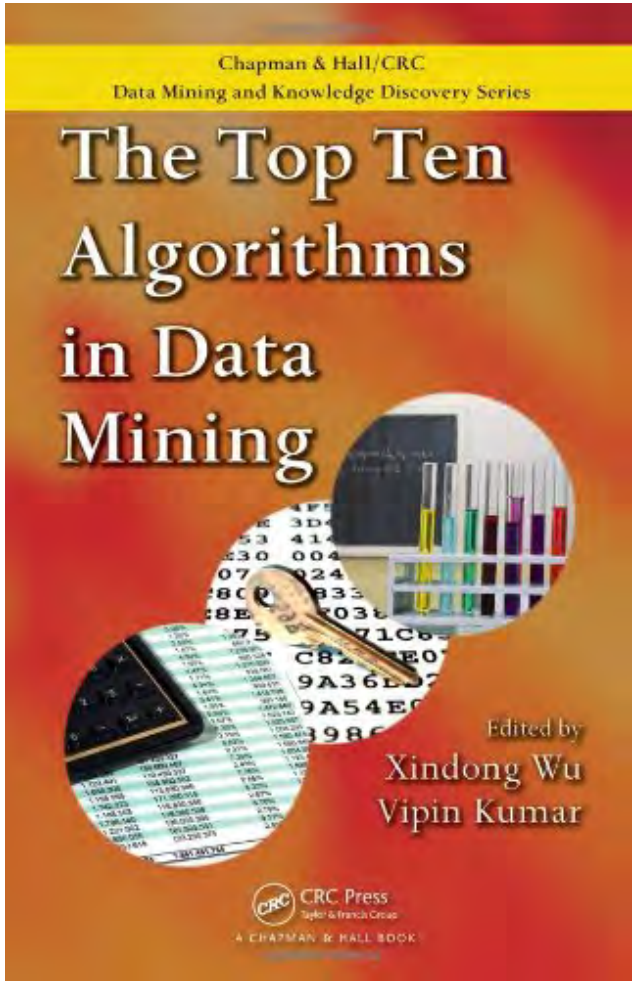
2010 *Big data, Go player*

http://www.cs.manchester.ac.uk/ugt/COMP24111

Une brève histoire des modes en apprentissage

Top 10 algorithmes en fouille de données

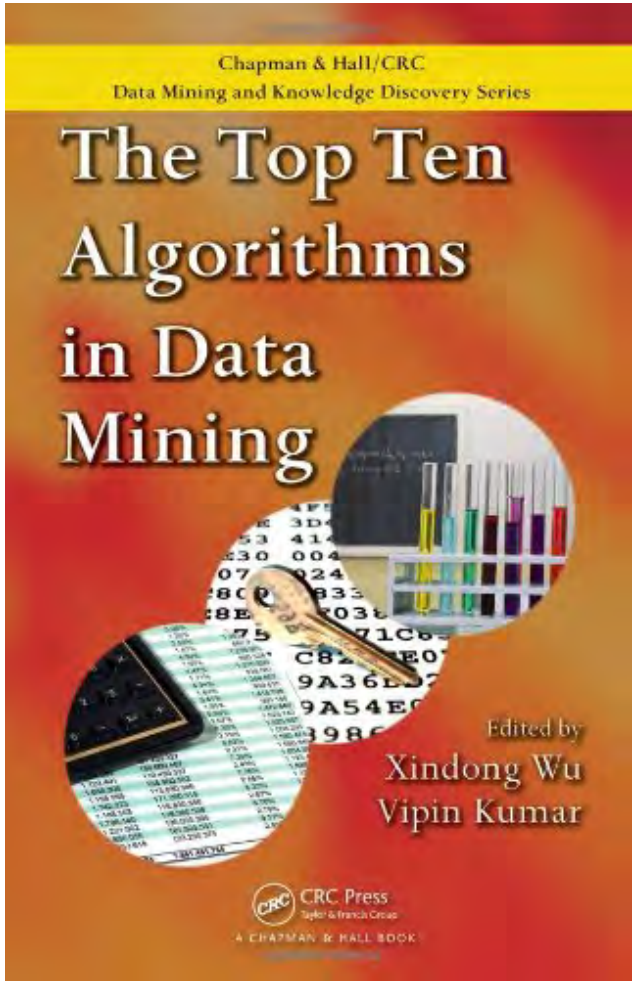
- Arbres de décision (et les forêts)
  - C4.5,
  - CART,
- SVM,
- AdaBoost,
- kNN,
- Bayesien
  - k-Means,
  - Apriori,
  - EM,
  - Naive Bayes,
- PageRank,
-



identified by the IEEE International Conference on Data Mining (ICDM) in December 2006

Top 10 algorithmes en fouille de données

- Arbres de décision (et les forêts)
  - C4.5,
  - CART,
- SVM,
- AdaBoost,
- kNN,
- Bayesien
  - k-Means,
  - Apriori,
  - EM,
  - Naive Bayes,
- PageRank,



- en 2014 un 11 ème : deep networks

identified by the IEEE International Conference on Data Mining (ICDM) in December 2006

Meilleure compréhension du monde à partir d'observations en vue de prédiction

C'est beau mais

$\rightarrow$ Comment ça marche ?

Plan

Etude de cas

- web : google, facebook...
- Marketing : Walmart & big data
  - volume
  - variété
  - vitesse
- Quelques challenges +
  image +
  la traduction automatique :
  deep learning strikes back

Plan

Une brève histoire des moteurs de recherche

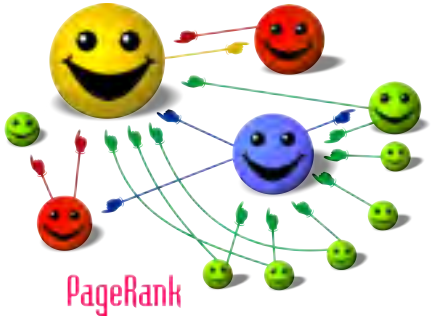
1994

solution manuelle

1995

solution logique

1996



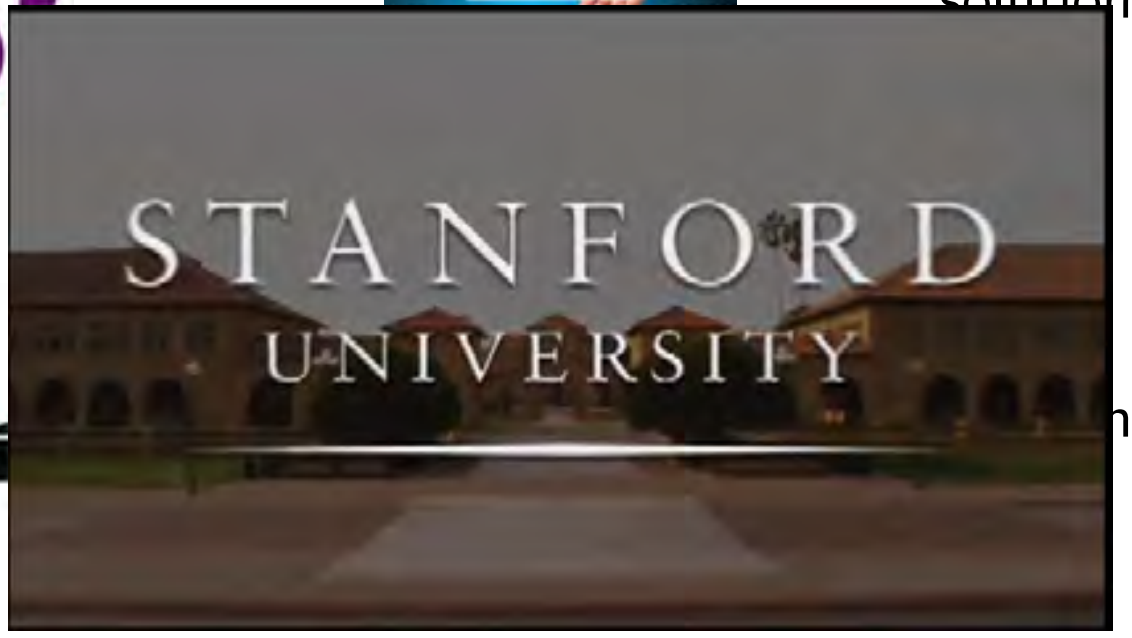
modèle statistique
+ valeur propre

Une brève histoire des moteurs de recherche

1994

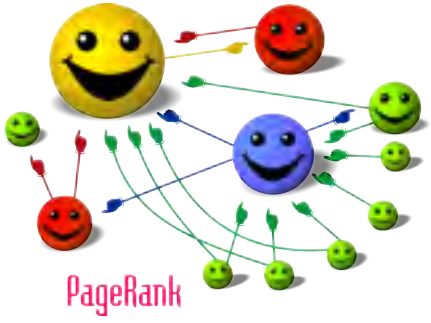
solution manuelle

1995



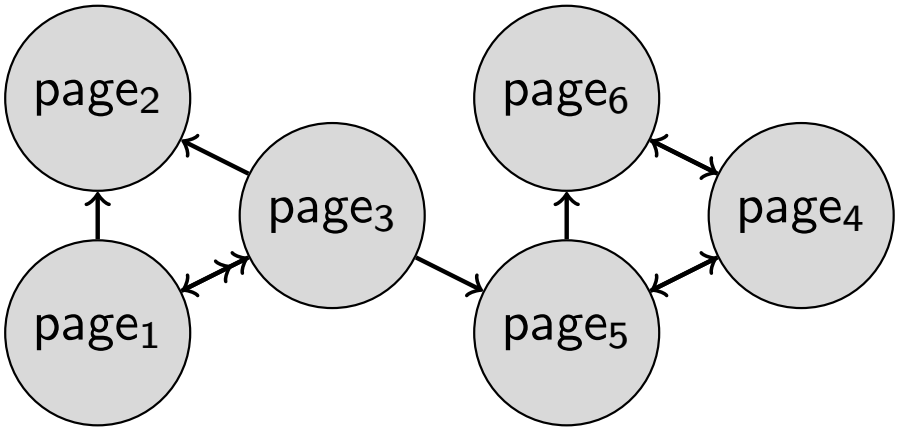
n logique

1996



modèle statistique
+ valeur propre

La recherche sur le web comme un problème de matrice

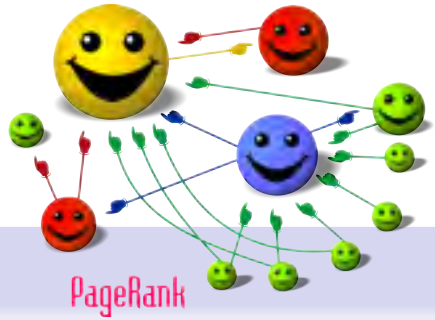


$$Y = \begin{pmatrix} 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 1/2 & 0 & 0 & 0 \\ 2/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/3 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

$$\mathbb{P}(\text{visiter la page}) \quad = \quad \text{une mesure de l'intérêt de la page}$$

$$Y : \mathbb{P}(j \to i) \quad = \quad \text{matrice des transitions}$$

Formalisation du problème

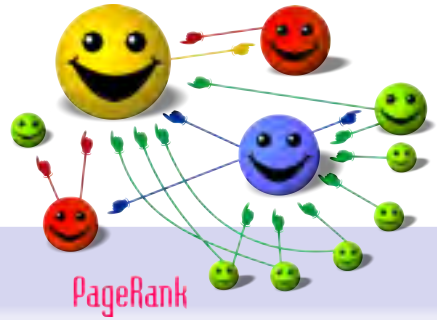


Modèle

$$\begin{array}{rcccc} \mathbb{P}(\text{visiter } p_i) & = & \sum_j \mathbb{P}(\text{visiter } p_j)\mathbb{P}(j \to i) & + & \mathbb{P}(\text{hasard}) \\ \mathbb{P} & = & (1-\delta) Y \mathbb{P} & + & \delta \mathbb{1}\mathbb{P} \\ \textit{observation} & = & \textit{information} & + & \textit{bruit} \end{array}$$

avec Y la matrice des transitions
et $\mathbb{P}$ le vecteur des probabilités de visiter une page

Formalisation du problème



Modèle

$$\mathbb{P}(\text{visiter } p_i) = \sum_j \mathbb{P}(\text{visiter } p_j)\mathbb{P}(j \to i) + \mathbb{P}(\text{hasard})$$

$$\mathbb{P} = (1-\delta) Y \mathbb{P} + \delta \mathbb{1} \mathbb{P}$$

$$\textit{observation} = \textit{information} + \textit{bruit}$$

avec Y la matrice des transitions
et $\mathbb{P}$ le vecteur des probabilités de visiter une page

Problème : retrouver $\mathbb{P}(\text{visiter } p_i)$

$$M\mathbb{P} = \mathbb{P} \qquad \text{avec} \qquad M = (1-\delta) Y + \delta \mathbb{1}$$

- C'est un problème de **valeur propre** de très grande taille
- avec $\delta = 0.15$

math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html

Les leçons de l'histoire des moteurs de recherche

- l'arrivée des big data

mégadonnées

- innover avec la recherche fondamentale & académique

créer des partenariats

- poser les problèmes : faire des maths et des statistiques

programmation à base d'exemples

- poser les vrai usages : interroger un moteur de recherche

organiser l'information au niveau mondial

- business model : la valeur c'est la requête

vers la recommandation

Plan

La recommandation de tous les jours

Systèmes existants

Libre: Myrrix (facto, Web Mahout), Easyrec (item-based, Web), Duine (multi approches), Propriétaires: KXpro...

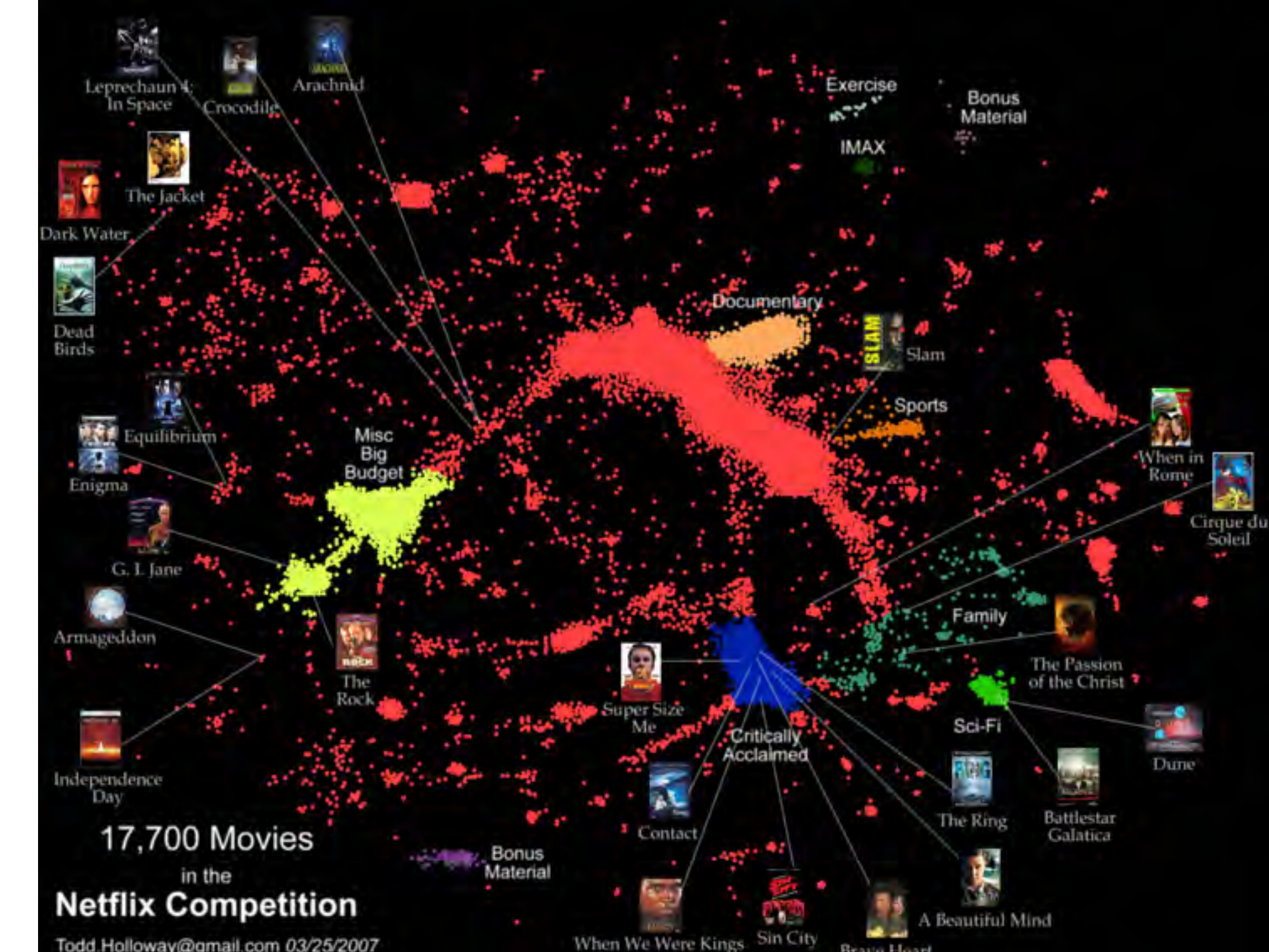
Quelle méthode choisir ?

Netflix Challenge (2007-2009)

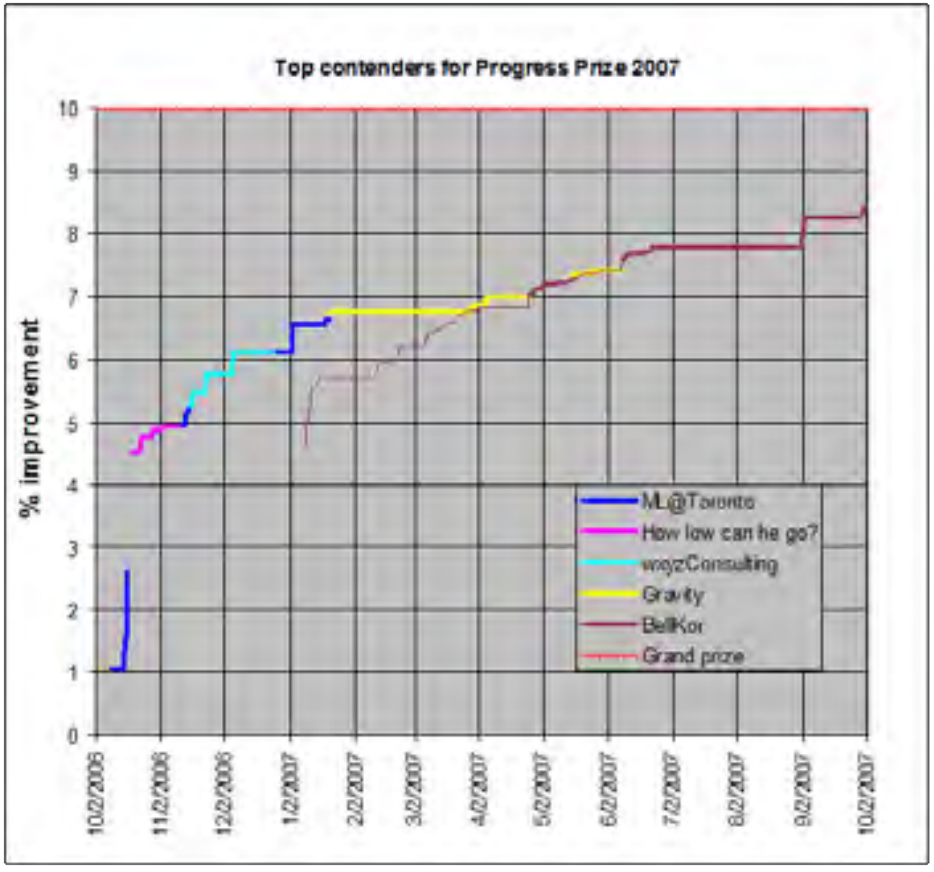
- Recommandation de films
- Moteur maison : CineMatch
- Améliorer les performances de 10%
- critère : erreur quadratique

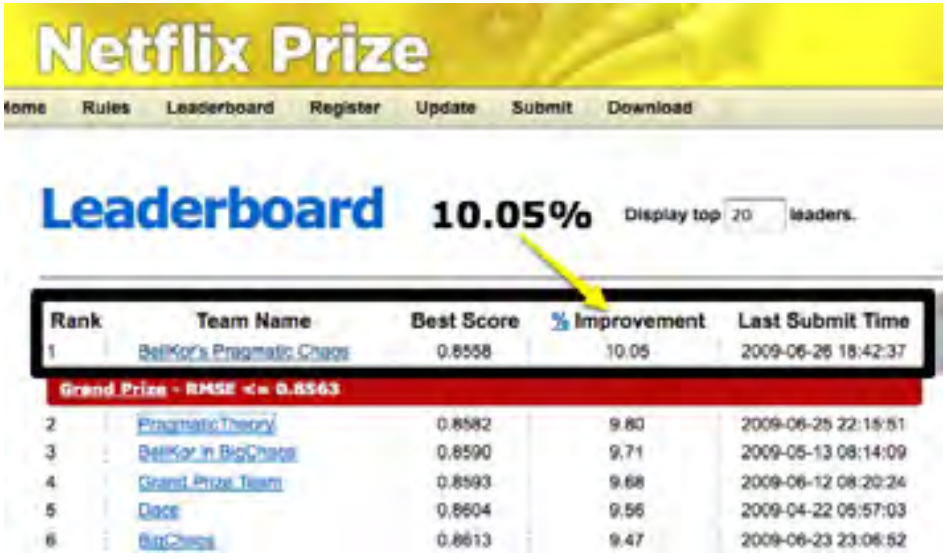
Les données

- 450 000 spectateurs
- 17 000 films
- 100 millons d'évaluations (de 1 à 5)
- taux de remplissage : 1,3 %
- test sur les 3 millions les plus récents
- 48 000 téléchargements des données

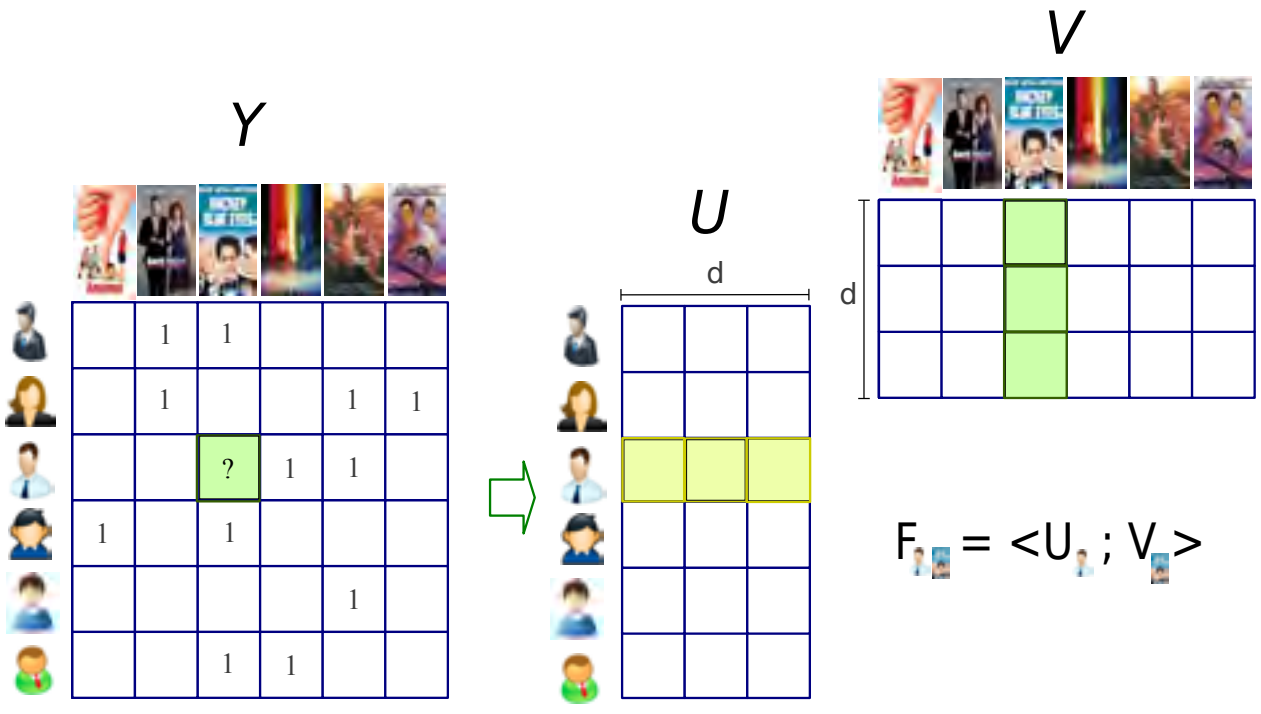

Leprechaun 4: In Space
Crocodile
Arachnid
Exercise
Bonus Material
IMAX
The Jacket
Dark Water
Dead Birds
Documentary
Slam
Sports
Equilibrium
Misc Big Budget
Enigma
When in Rome
Cirque du Soleil
G. I. Jane
Armageddon
The Rock
Family
The Passion of the Christ
Super Size Me
Critically Acclaimed
Sci-Fi
Dune
Independence Day
Contact
The Ring
Battlestar Galatica
17,700 Movies
in the
Netflix Competition
Bonus Material
A Beautiful Mind
When We Were Kings
Sin City
Todd Holloway@gmail.com 03/25/2007

Faire de la science un sport !





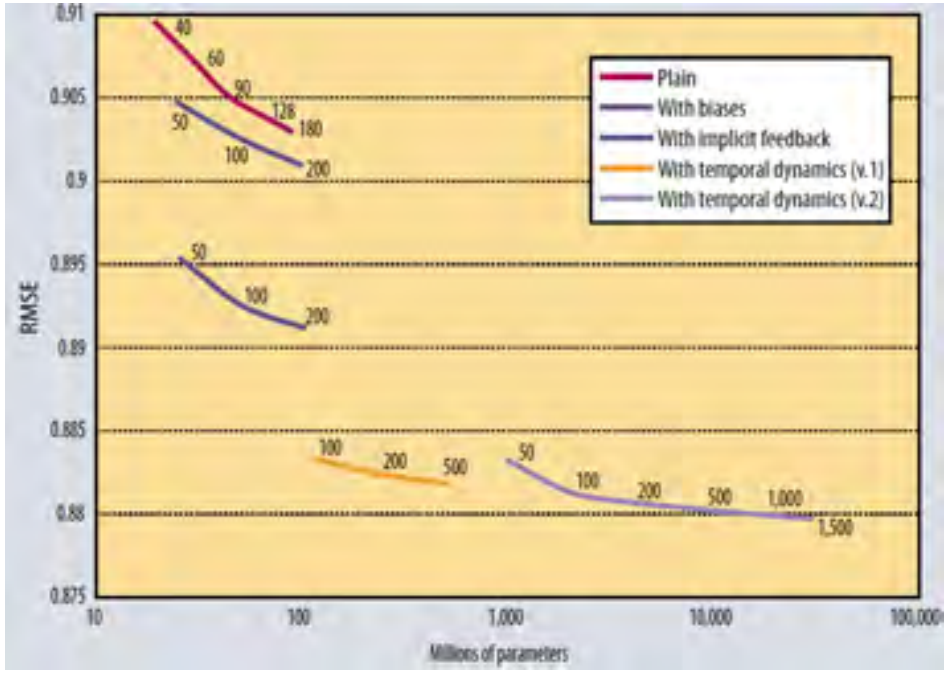
La recommandation comme un problème de matrice



$\langle U_i, V_j \rangle$ une mesure d'appétence de l'utilisateur i pour le produit j

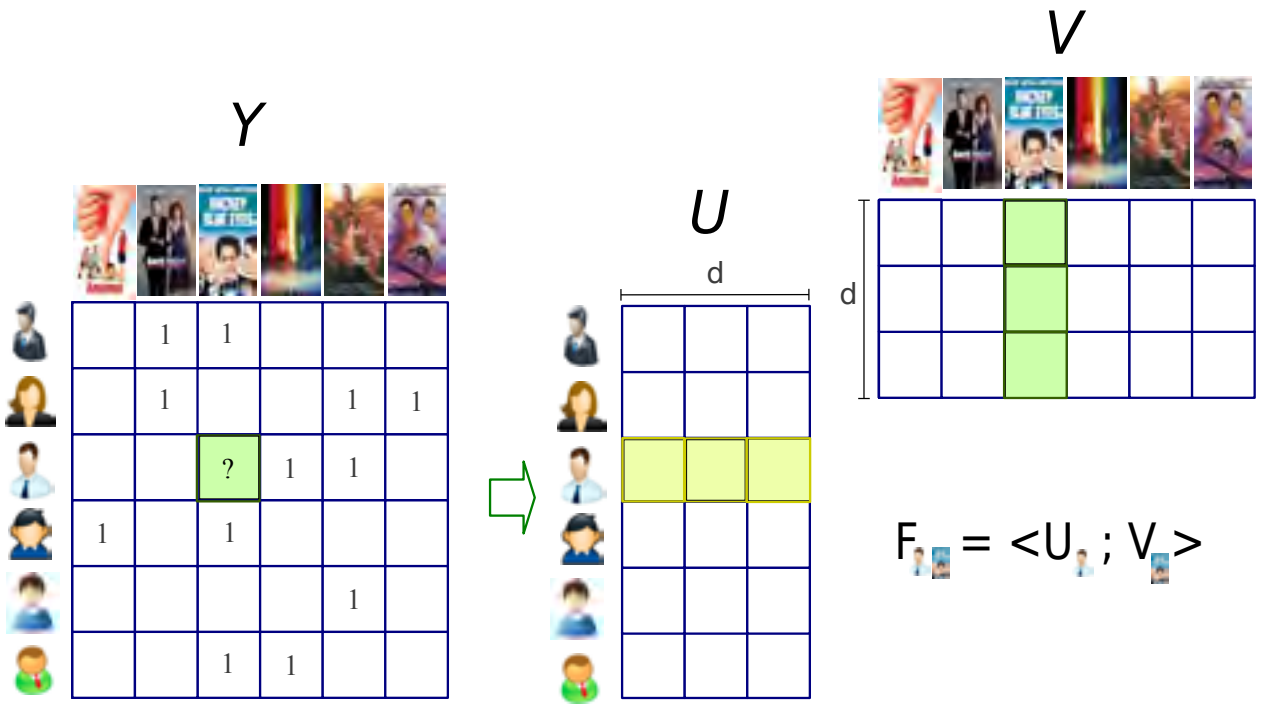
Les résultats sur Netflix

- erreur initiale : 0.9514
- factorisation - 4 %
- nombres de facteurs - .5 %
- améliorations - 2.5 %
  - normalisation
  - poids
  - temps
- bagging : - 3 % = 0.8563
  mélange 100 méthodes



Koren, Y., Bell, R.M., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Computer 42(8), 30–37 (2009)

La recommandation comme un problème de matrice



$\langle U_i, V_j \rangle$ une mesure d'appétence de l'utilisateur i pour le produit j

Formalisation du problème

Modèle

$$\begin{array}{ccccc} Y & = & F & + & R \\ observation & = & information & + & bruit \end{array}$$

avec F régulière et R une **matrice** de bruit.

Formalisation du problème

Modèle

$$\begin{array}{ccccc} Y & = & F & + & R \\ observation & = & information & + & bruit \end{array}$$

avec F régulière et R une **matrice** de bruit.

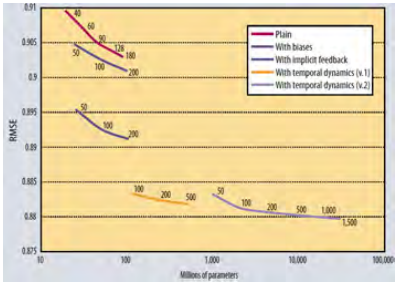
Problème : construire $\widehat{F}$ un estimateur de F

$$\min_{\widehat{F}} \underbrace{\mathcal{L}(Y, \widehat{F})}_{\text{attache aux données}} + \lambda \underbrace{\Omega(\widehat{F})}_{\text{pénalisation}}$$

- optimiser un cout bien choisi
- avec $\widehat{F} = UV$ factorisation (...à l'aide de valeurs propres)

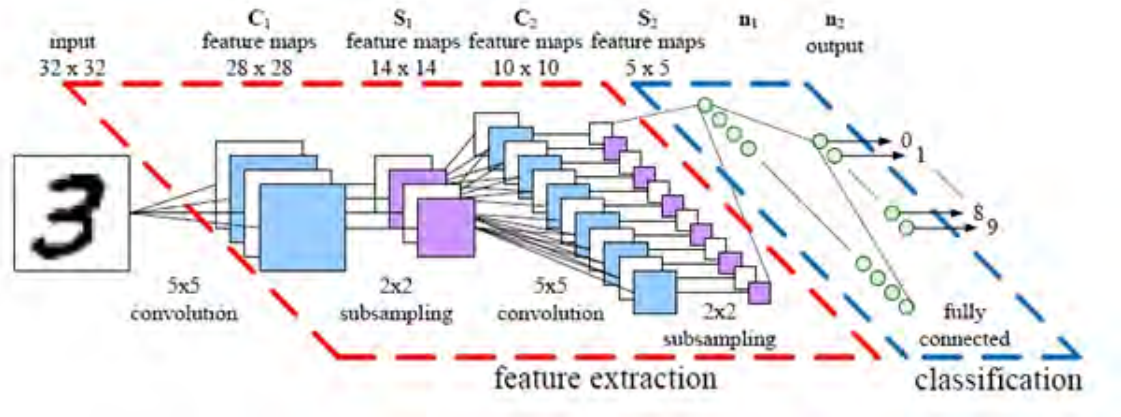
Les leçons de Netflix

- trop d'efforts spécifiques pour être généralisée
- Modèle statistique : factorisation *is the solution*
- prendre en compte les spécificités (ici l'effet temps et les feedback implicites)

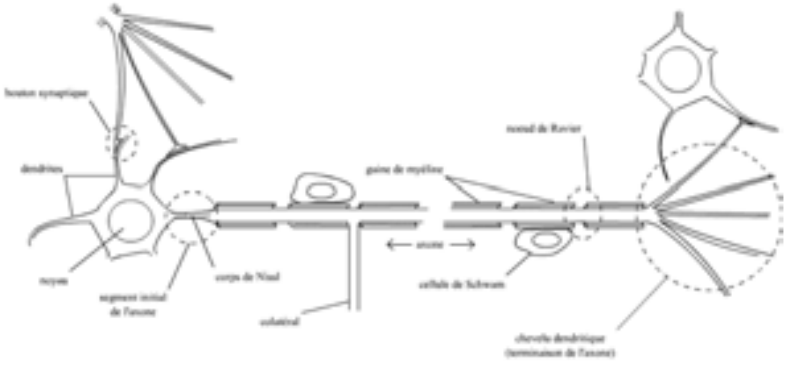


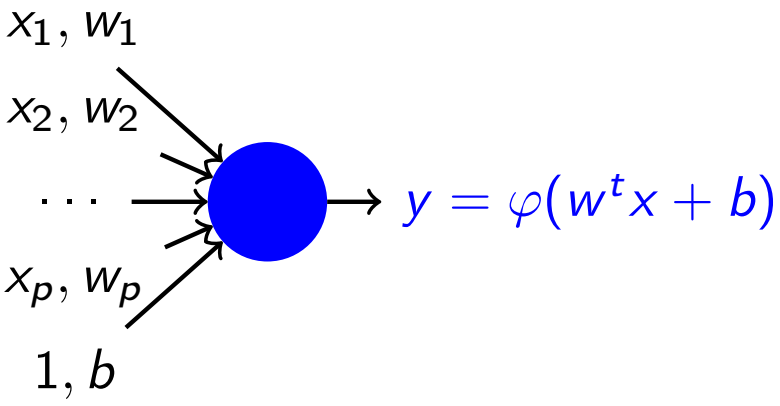
- une optimisation de type gradient stochastique
  - flexible
  - qui passe à l'échelle
- quels usages pour la recommandation ?

Plan

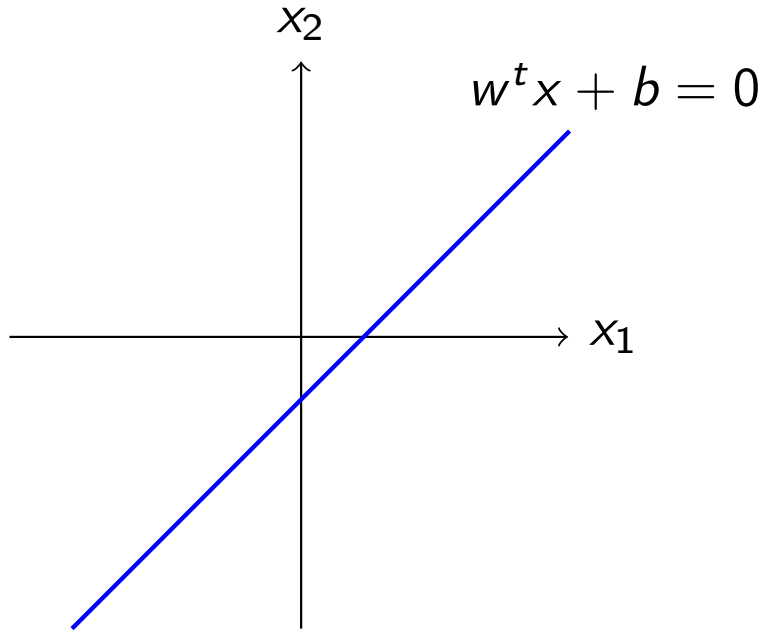


The formal neuron



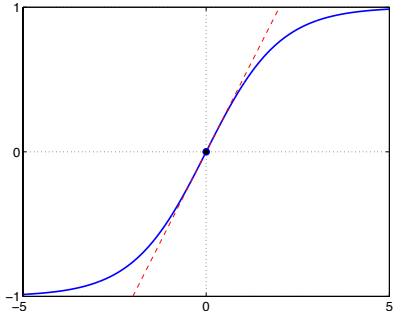


Defines a hyperplane



- x input
- w weight, b bias
- φ activation function
- y output

$\varphi(t) = \tanh(t)$



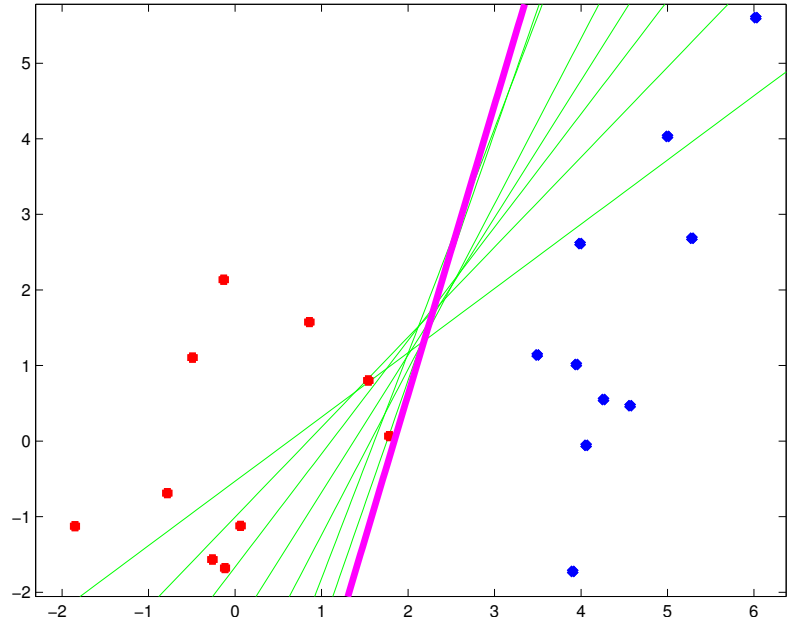
The formal neuron as a learning machine

Fitting the data for the 2 classes classification problem

given n pairs of input–output data $\mathbf{x}_i, y_i, i = 1, n$
find w such that $\underbrace{\varphi(\mathrm{w}^t \mathbf{x}_i)}_{\text{prediction of the model}} = \underbrace{y_i}_{\text{ground truth}}$

make some initial guess for w

- pick some data
- extract information
- improve w



what about new data – generalization?

Fitting the data with an energy-based model

Fitting the data

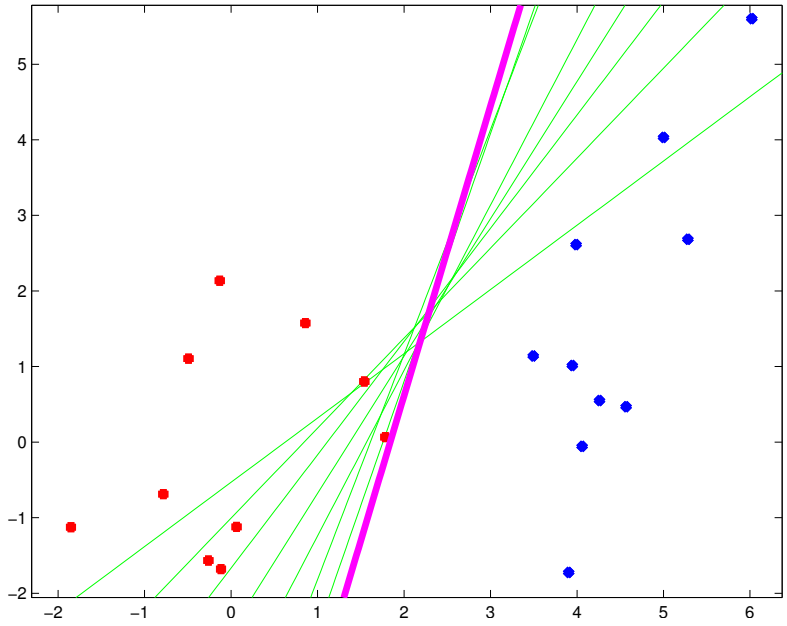
$$\min_{\mathrm{w}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \Big(\underbrace{\varphi(\mathrm{w}^t \mathbf{x}_i)}_{\text{prediction}} - \underbrace{y_i}_{\text{truth}} \Big)^2$$

Improve w

$$\mathrm{w}^{\text{new}} = \mathrm{w}^{\text{old}} - \rho \underbrace{\big(\varphi(w^t \mathbf{x}_i) - y_i\big) \nabla_{\mathrm{w}} \varphi(w^t \mathbf{x}_i)}_{\text{extracted information}}$$

Algorithm 1 Adaline

Data: w initialization, ρ fixed stepsize
Result: w
while *not converged* **do**
  $\mathbf{x}_i, y_i \leftarrow$ pick a point i at random
  $\mathbf{d} \leftarrow (w^t \mathbf{x}_i - y_i)\mathbf{x}_i$
  $\mathrm{w} \leftarrow \mathrm{w} - \rho \mathbf{d}$
  check that the cost decease
end



Fitting the data with an energy-based model (part of P)

Coding the output (truth?)

$$y_i = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \text{class 1} \\ -1 & \text{else} \end{cases}$$

Fitting the data

$$\min_{\mathrm{w}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} (\underbrace{\varphi(\mathrm{w}^t\mathbf{x}_i)}_{\text{prediction p}} - \underbrace{y_i}_{\text{truth t}})^2$$

$$(\varphi(\mathrm{w}^t\mathbf{x}_i) - y_i)^2 = y_i^2 (\underbrace{y_i\varphi(\mathrm{w}^t\mathbf{x}_i)}_{\text{score } z_i} - 1)^2$$

As a loss

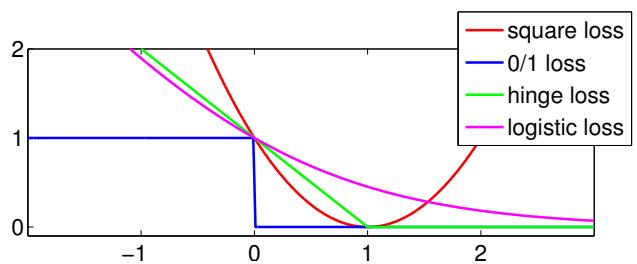
$$\min_{\mathrm{w}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \mathit{cost}(z_i) \qquad z_i = \varphi(\mathrm{w}^t\mathbf{x}_i)y_i$$

0/1 loss

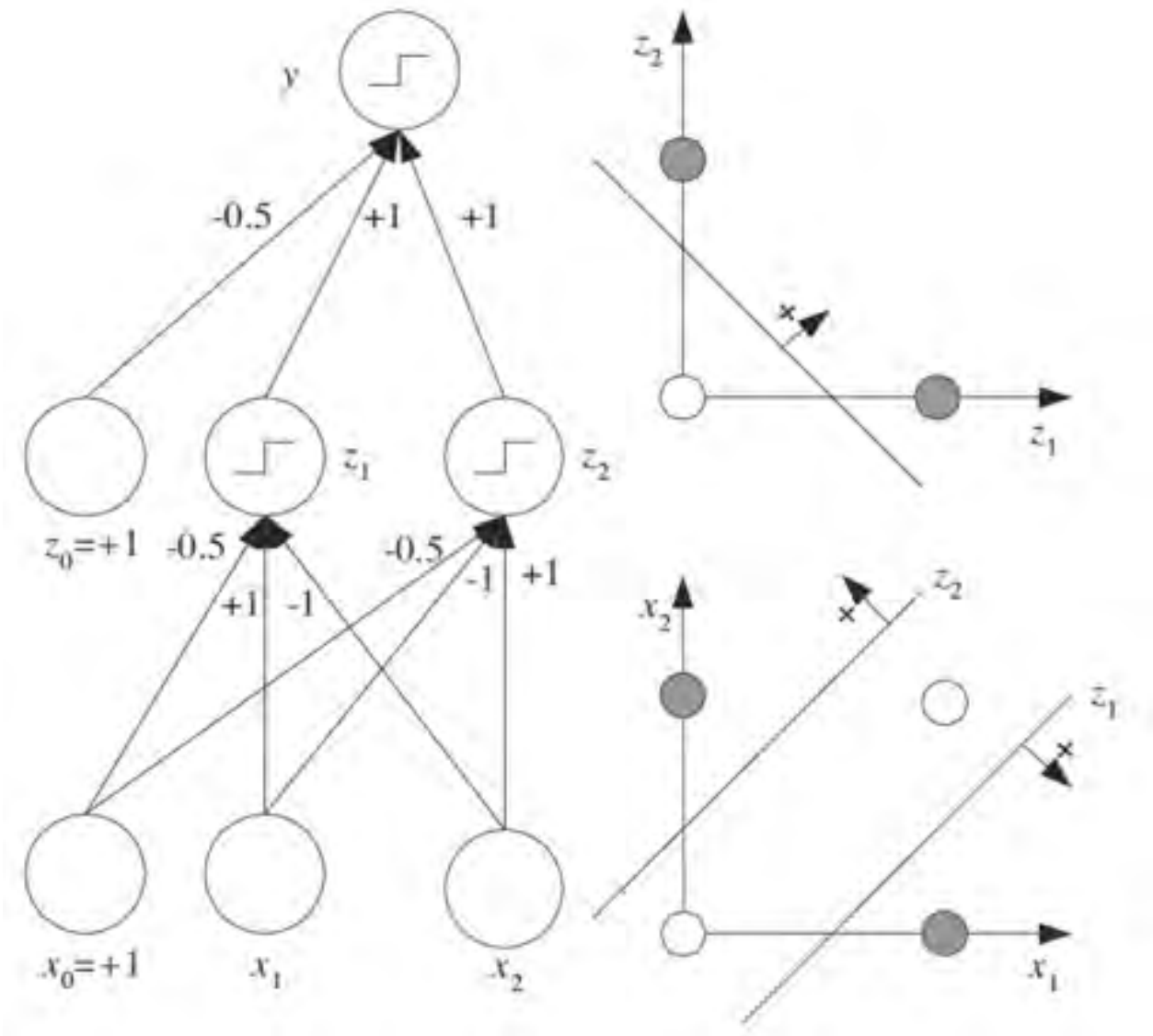
$$\mathit{cost}(z_i) = \begin{cases} 0 & \text{if } \mathit{sign}(\mathrm{w}^t\mathbf{x}_i) = \mathit{sign}(y_i) \\ 1 & \text{else} \end{cases}$$

As a constraint

$$\mathit{sign}(\mathrm{w}^t\mathbf{x}_i)\mathit{sign}(y_i) \geq 0$$



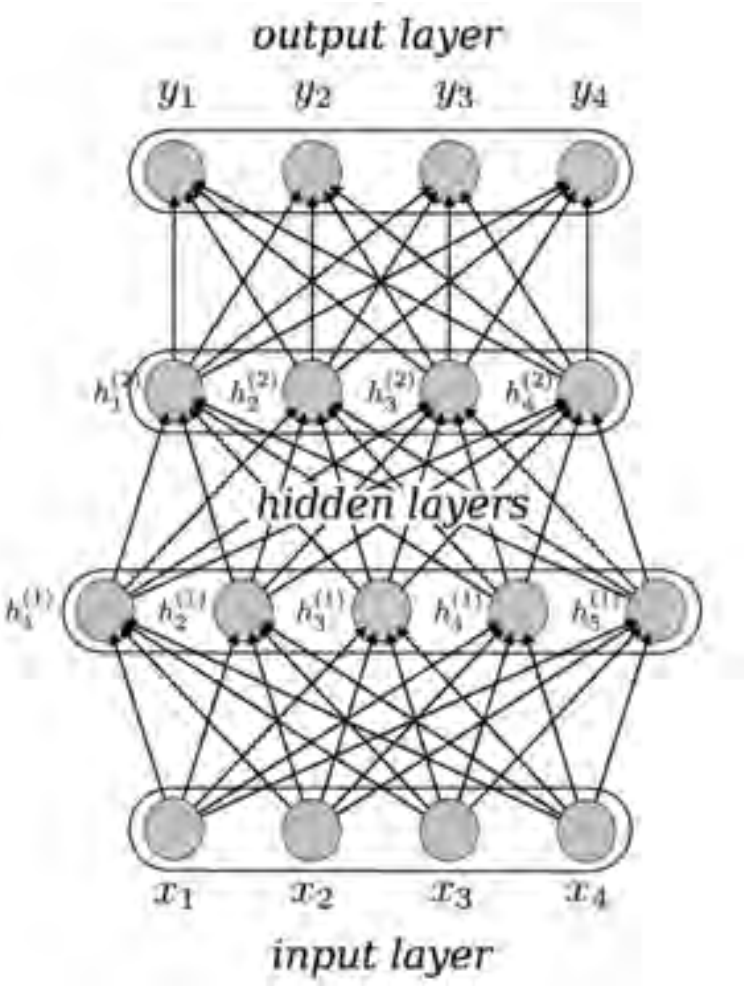
Non linearity combining linear neurons



Alpaydın, Introduction to Machine Learning, 2010

Neural networks as universal approximator

Running several neurons at the same time



$$y = \varphi(W_3 h^{(2)})$$

$$h^{(2)} = \varphi(W_2 h^{(1)})$$

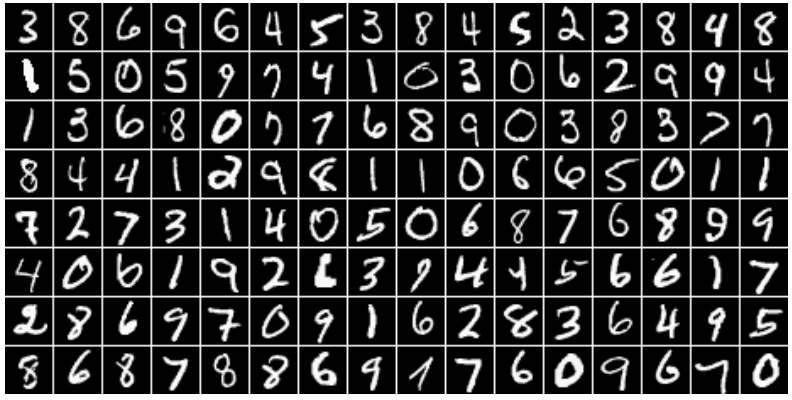
$$h^{(1)} = \varphi(W_1 \mathbf{x})$$

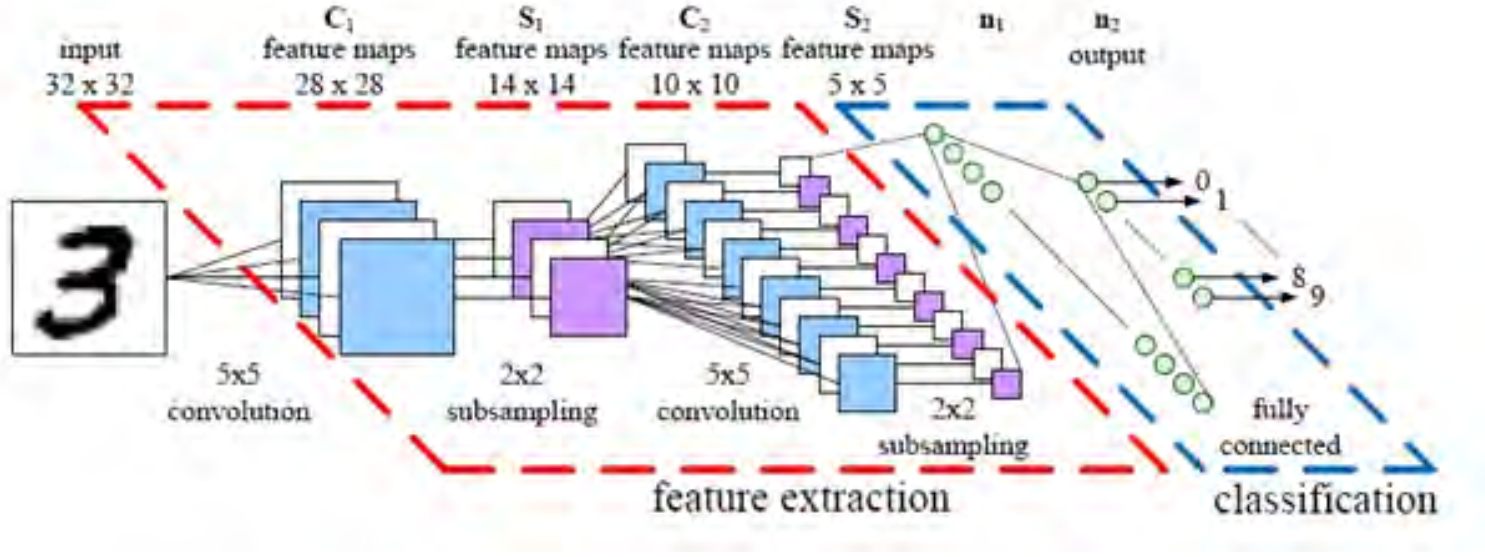
$$\mathbf{x}$$

Multilayered neural network: learning internal representation W_1, W_2, W_3

from L. Arnold PhD

OCR: the MNIST database (Y. LeCun, 1989)

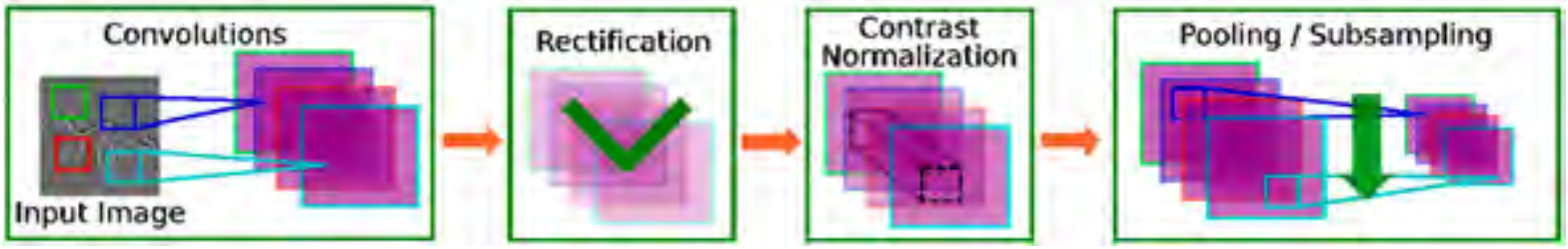




The caltech 101 database (2004)

- 101 classes
- 30 training images per category
- ...and the winner is NOT a deep network
  - dataset is too small

Deep architecture lessons from caltech 101 (2009)



Single Stage System: $[64.F^{9\times9}_{CSG} - R/N/P^{5\times5}]$ - log_reg					
	$R_{abs} - N - P_A$	$R_{abs} - P_A$	$N - P_M$	$N - P_A$	P_A
U^+	54.2%	50.0%	44.3%	18.5%	14.5%
R^+	54.8%	47.0%	38.0%	16.3%	14.3%
U	52.2%	43.3%(±1.6)	44.0%	17.2%	13.4%
R	53.3%	31.7%	32.1%	15.3%	12.1%(±2.2)
G	52.3%				
Two Stage System: $[64.F^{9\times9}_{CSG} - R/N/P^{5\times5}] - [256.F^{9\times9}_{CSG} - R/N/P^{4\times4}]$ - log_reg					
	$R_{abs} - N - P_A$	$R_{abs} - P_A$	$N - P_M$	$N - P_A$	P_A
U^+U^+	65.5%	60.5%	61.0%	34.0%	32.0%
R^+R^+	64.7%	59.5%	60.0%	31.0%	29.7%
UU	63.7%	46.7%	56.0%	23.1%	9.1%
RR	62.9%	33.7%(±1.5)	37.6%(±1.9)	19.6%	8.8%
GT	55.8%				
Single Stage: $[64.F^{9\times9}_{CSG} - R_{abs}/N/P^{5\times5}_A]$ - PMK-SVM					
U	64.0%				
Two Stages: $[64.F^{9\times9}_{CSG} - R_{abs}/N/P^{5\times5}_A] - [256.F^{9\times9}_{CSG} - R_{abs}/N]$ - PMK-SVM					
UU	52.8%				

R_{abs} Rectification Layer

N Local Contrast Normalization Layer

P_A Average Pooling and Subsampling Layer

in What is the Best Multi-Stage Architecture for Object Recognition? Jarrett et al, 2009

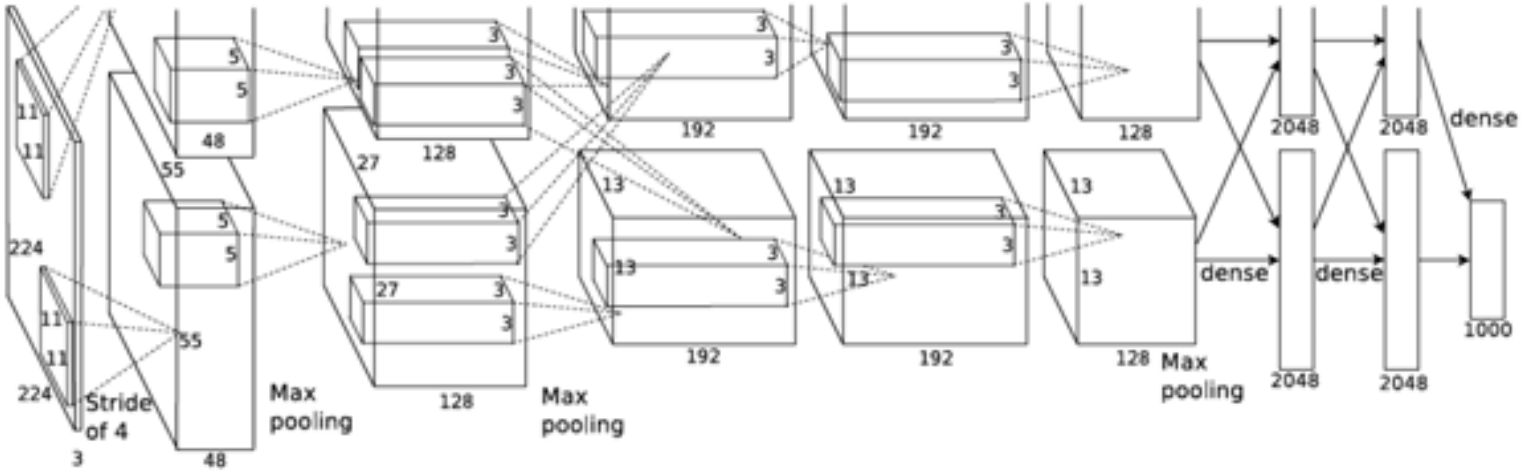
The image net database (Deng et al., 2012)

ImageNet = 15 million labeled high-resolution images of 22,000 categories.
Large-Scale Visual Recognition Challenge (a subset of ImageNet)

- 1000 categories.
- 1.2 million training images,
- 50,000 validation images,
- 150,000 testing images.

www.image-net.org/challenges/LSVRC/2012/

Deep architecture and the image net



The architecture of the CNN [Krizhevsky, Sutskever, Hinton, 2012]

- 60 million parameters
- using 2 GPU
- regularization
  - data augmentation
  - dropout
  - weight decay

A new fashion in image processing

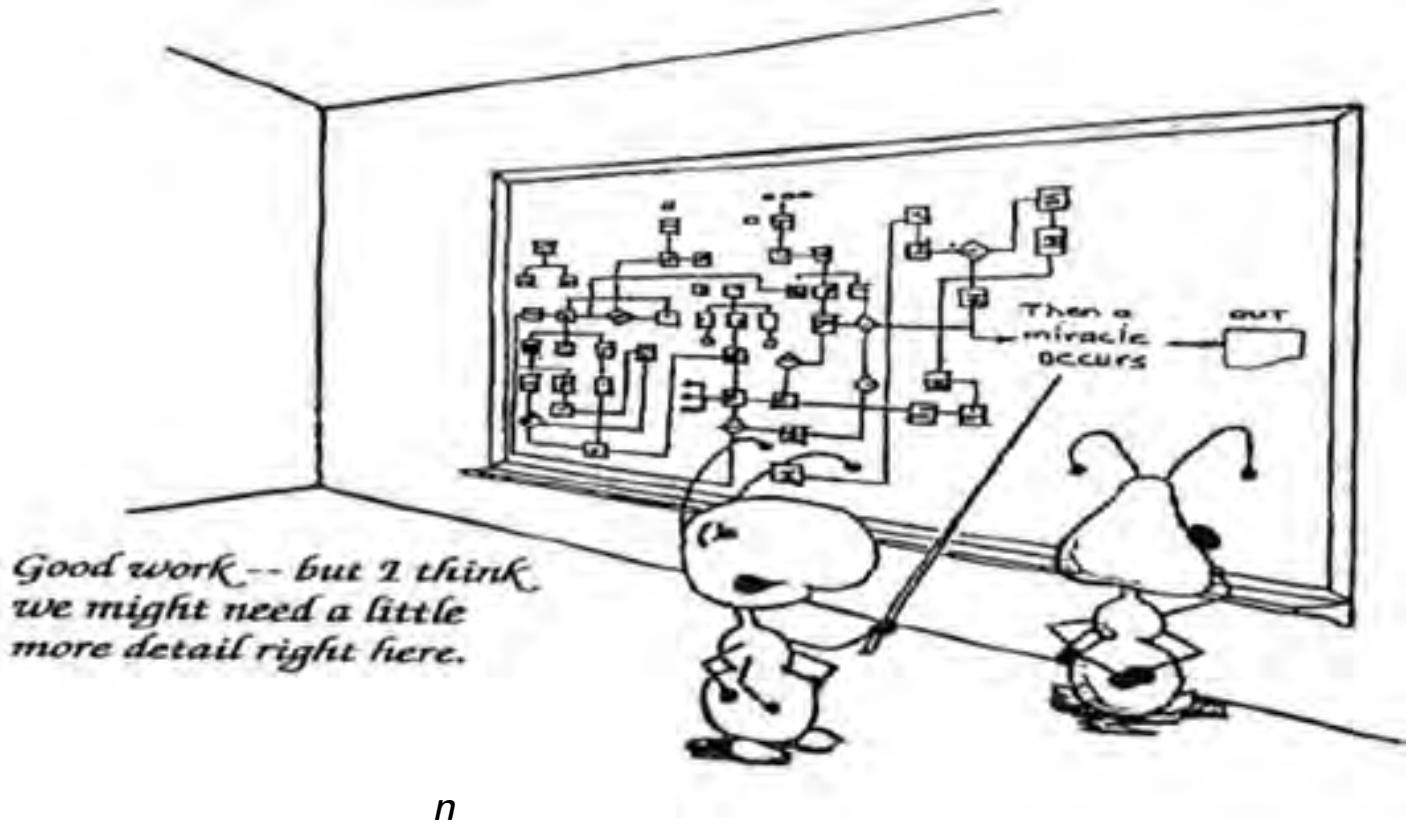
2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

Baidu: 5%

Learning Deep architecture

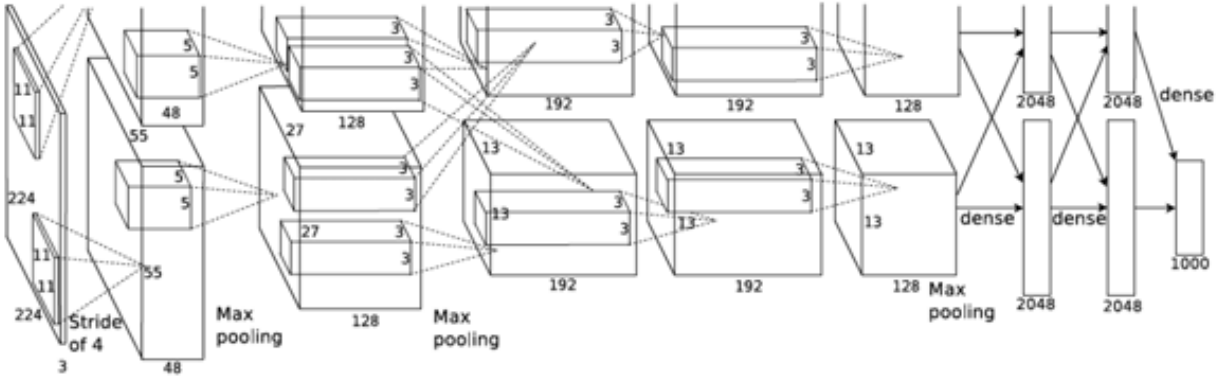


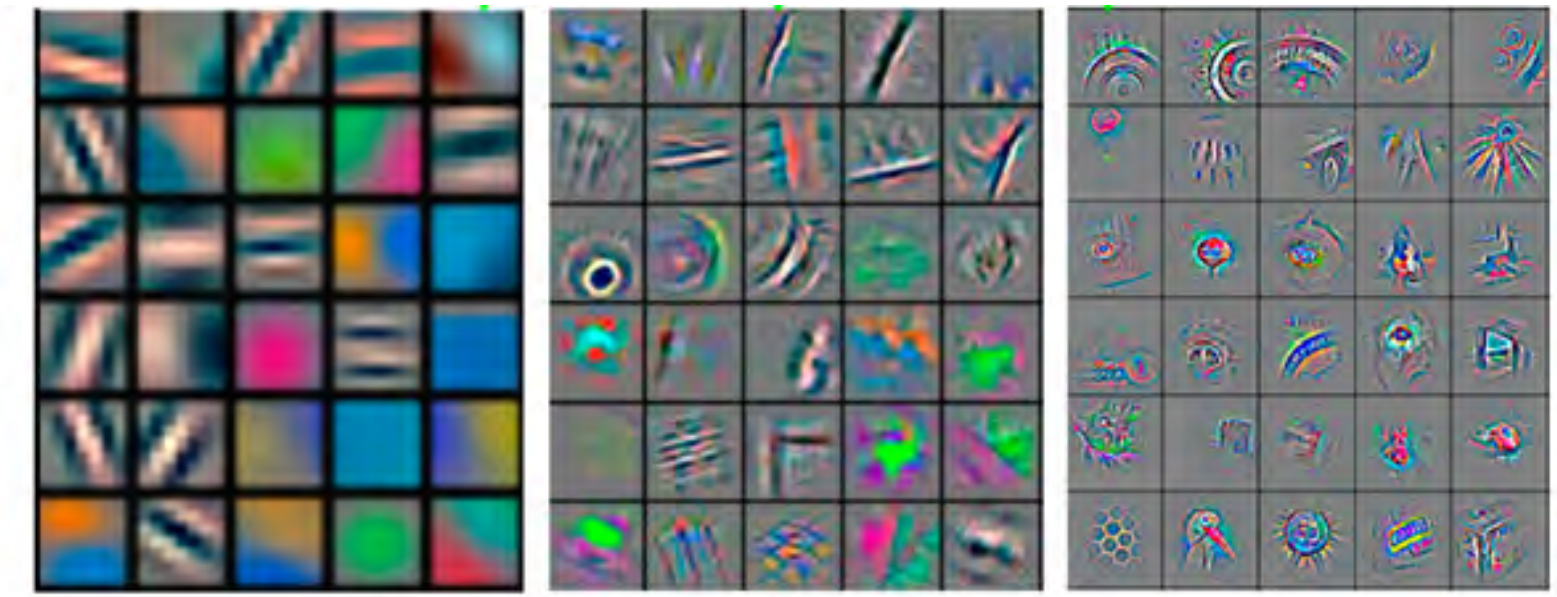
$$\min_{W \in \mathbb{R}^d} \sum_{i=1}^{n} \|f(x_i, W) - y_i\|^2 + \lambda \|w\|^2$$

- $d = 60 \times 10^6$
- $n = 1,200\,000$ ++
- $\lambda = 0.0005$

f is a deep NN

Then a m. occurs: learning internal representation





Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Deep architecture: more contests and benchmark records

- speech (phoneme) recognition,
- automatic translation,
- Optical Character Recognition (OCR),
- ICDAR Chinese handwriting recognition benchmark,
- Grand Challenge on Mitosis Detection, Road sign recognition
- Higgs boson challenge

Machine Learning in High Energy Physics

The winning model is "brute force"

- a bag of 70 dropout neural networks
- three hidden layers of 600 neurons
- produced by 2-fold stratified cross-validations

https://www.kaggle.com/c/higgs-boson

Comment ça marche ?

- Programmation par l'exemple (et beaucoup - des méga données)
- un bon critère :
  - intégre des connaissances a priori
  - adjustment aux données (énergie)
  - un ou des mécanismes de régularisation
- une bonne procédure d'optimisation
- des ressources informatiques suffisantes

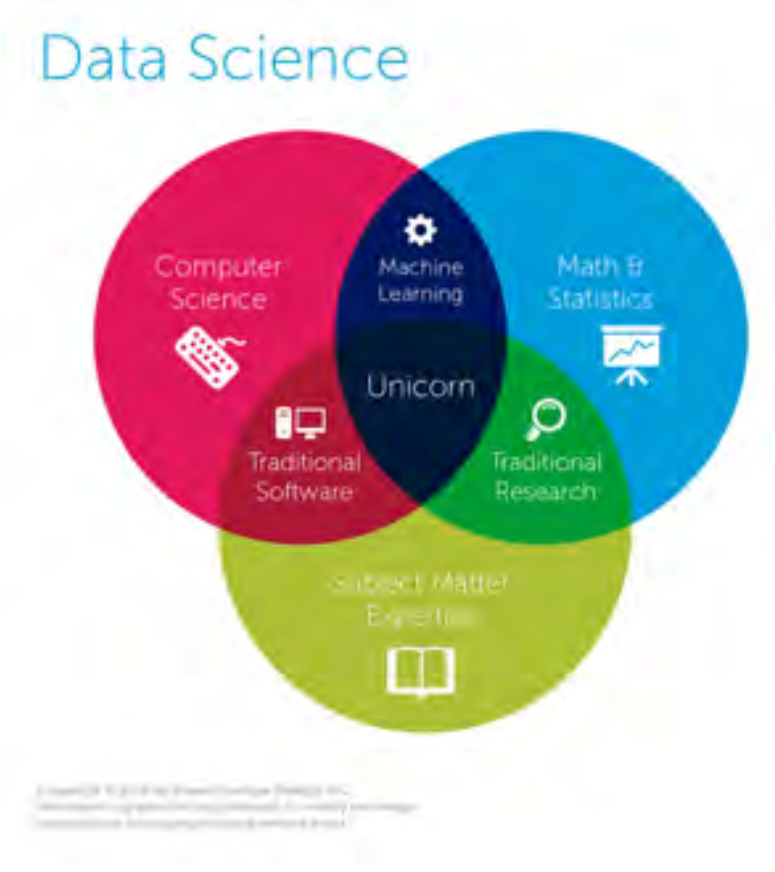
Ce que l'on ne comprend pas

- gérer et protéger les **données** et les contextes ?
- qu'est-ce qu'apprendre ?
  - pourquoi les architectures profondes fonctionnent ?
  - apprendre à apprendre (transfert, transport) ?
  - quelles **garanties** peut on avoir ?
  - comment gérer les **dynamiques** ?
- comment **optimiser** efficacement des fonctionnelles non convexes et non régulières ?
- comment **automatiser** la chaine de traitement nécessaire à l'apprentissage ?

Statistiques + optimisation + informatique

vers la science des données

Plan



New trends in ML fashion: big data

ICML 2014 workshops

- Designing Machine Learning Platforms for Big Data

  Xiangxiang Meng, Wayne Thompson, Xiaodong Lin

- New Learning Frameworks and Models for Big Data

  Massih-Reza Amini, Eric Gaussier, James Kwok, Yiming Yang

- Deep Learning Models for Emerging Big Data Applications

  Shan Suthaharan, Jinzhu Jia

- Unsupervised Learning for Bioacoustic Big Data

  Hervé Glotin, P. Dugan, F. Chamroukhi, C. Clark, Yann LeCun

- Knowledge-Powered Deep Learning for Text Mining

  Bin Gao, Scott Yih, Richard Socher, Jiang Bian

- Optimizing Customer Lifetime Value in Online Marketing

  Georgios Theocharous, Mohammad Ghavamzadeh, Shie Mannor

Comment traiter ces *big data :* data science

this joke makes perfect sense

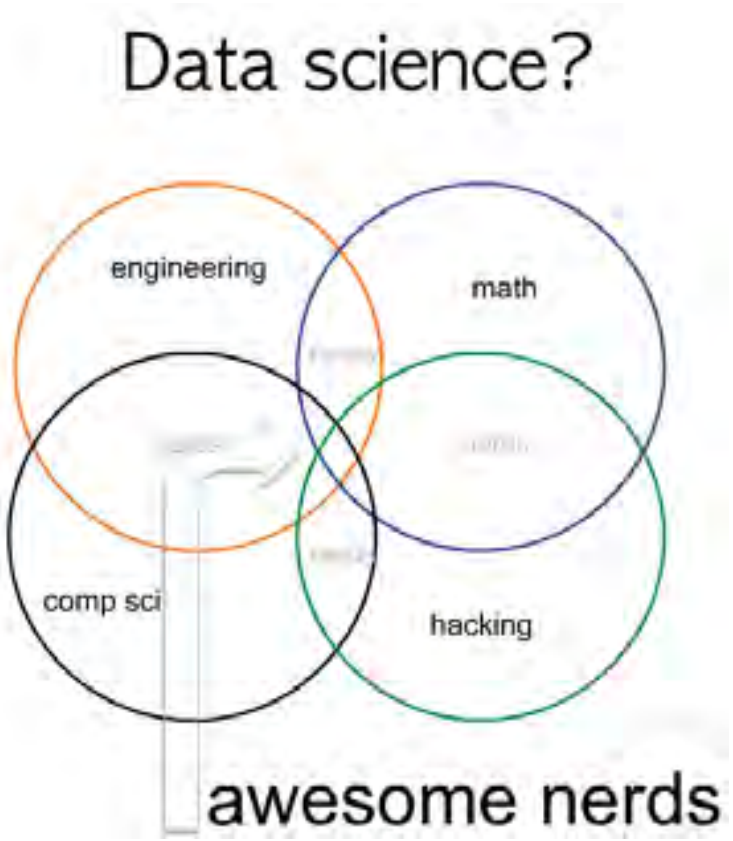
4 January 2013 12:28

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it

@[redacted]

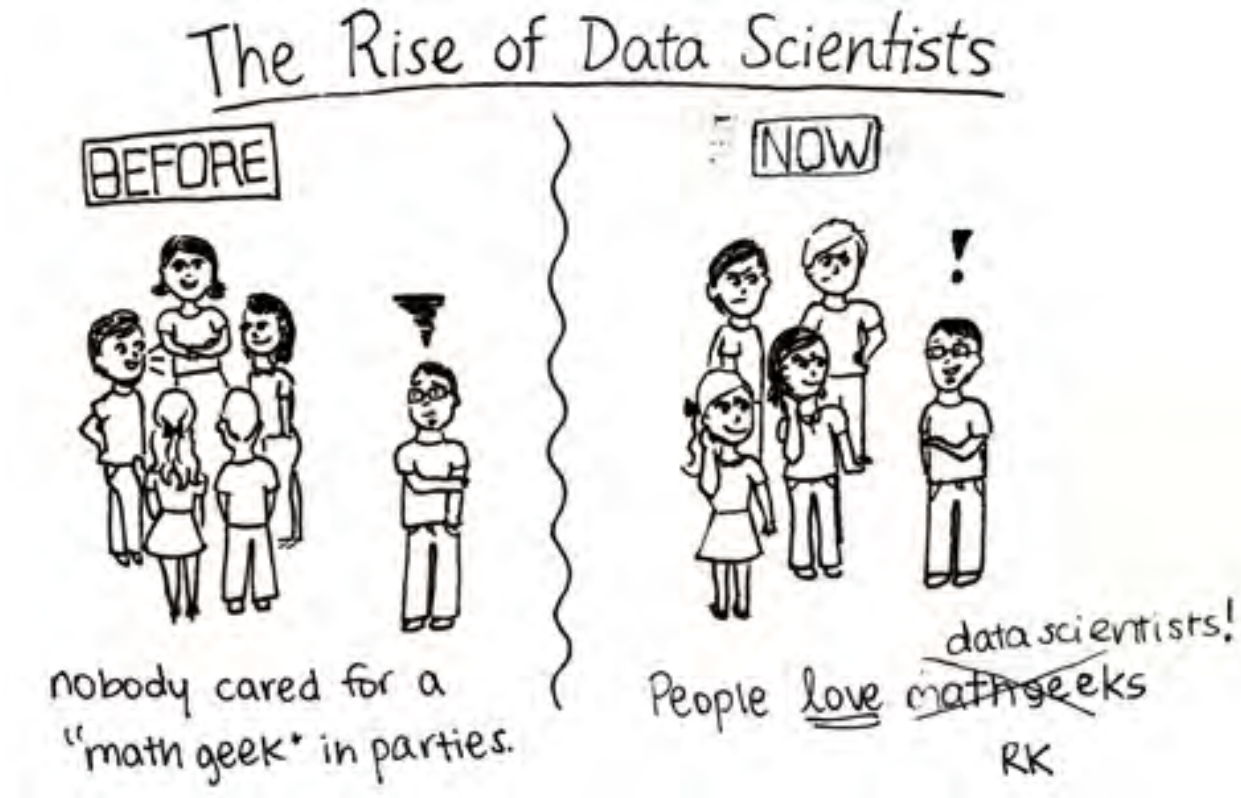
weibo.com/bohaishiyi

New trends in ML fashion: data science



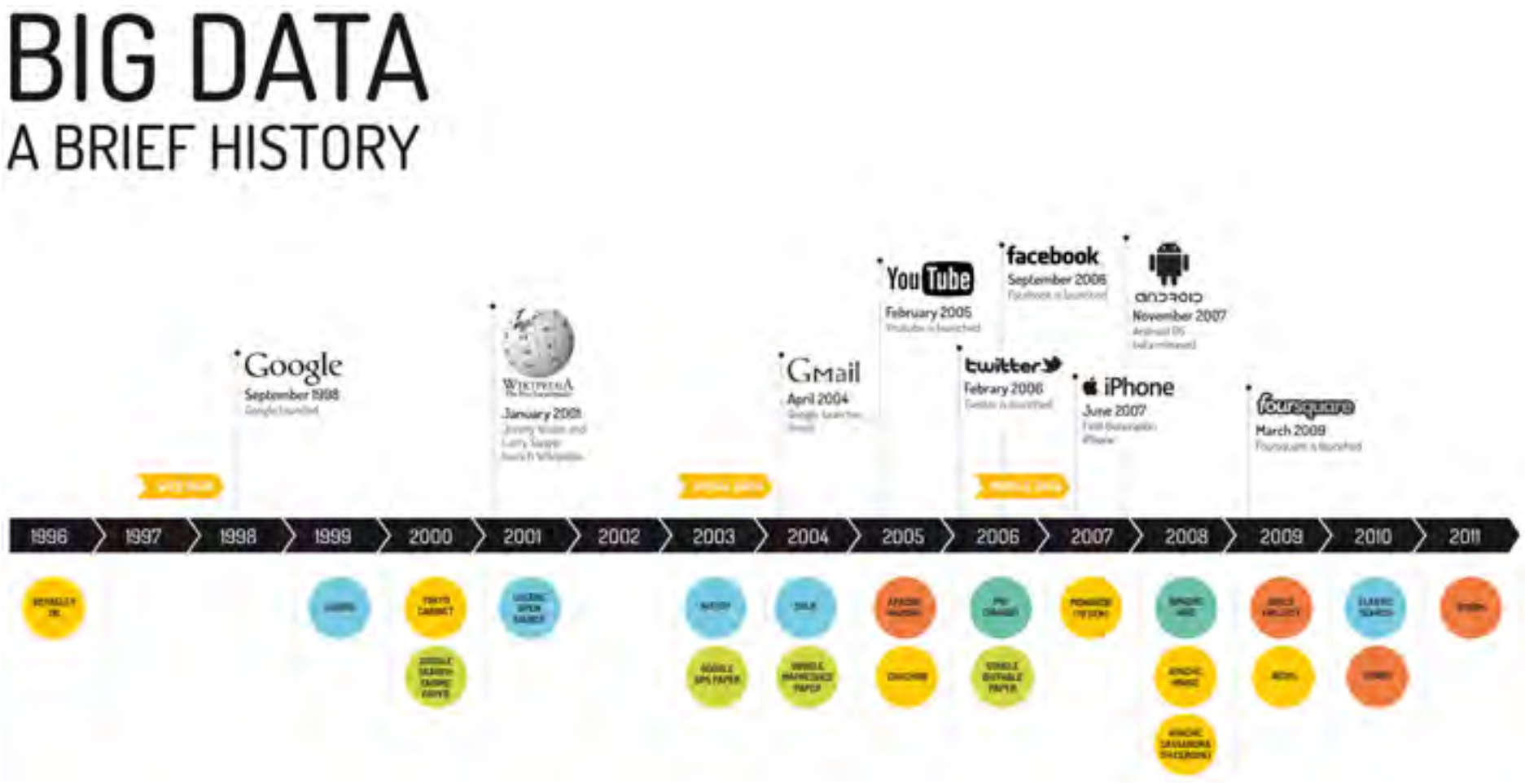
Hilary Mason, `http://www.infoq.com/presentations/Machine-Learning`

Data Scientist Voted Sexiest Job of 21st Century



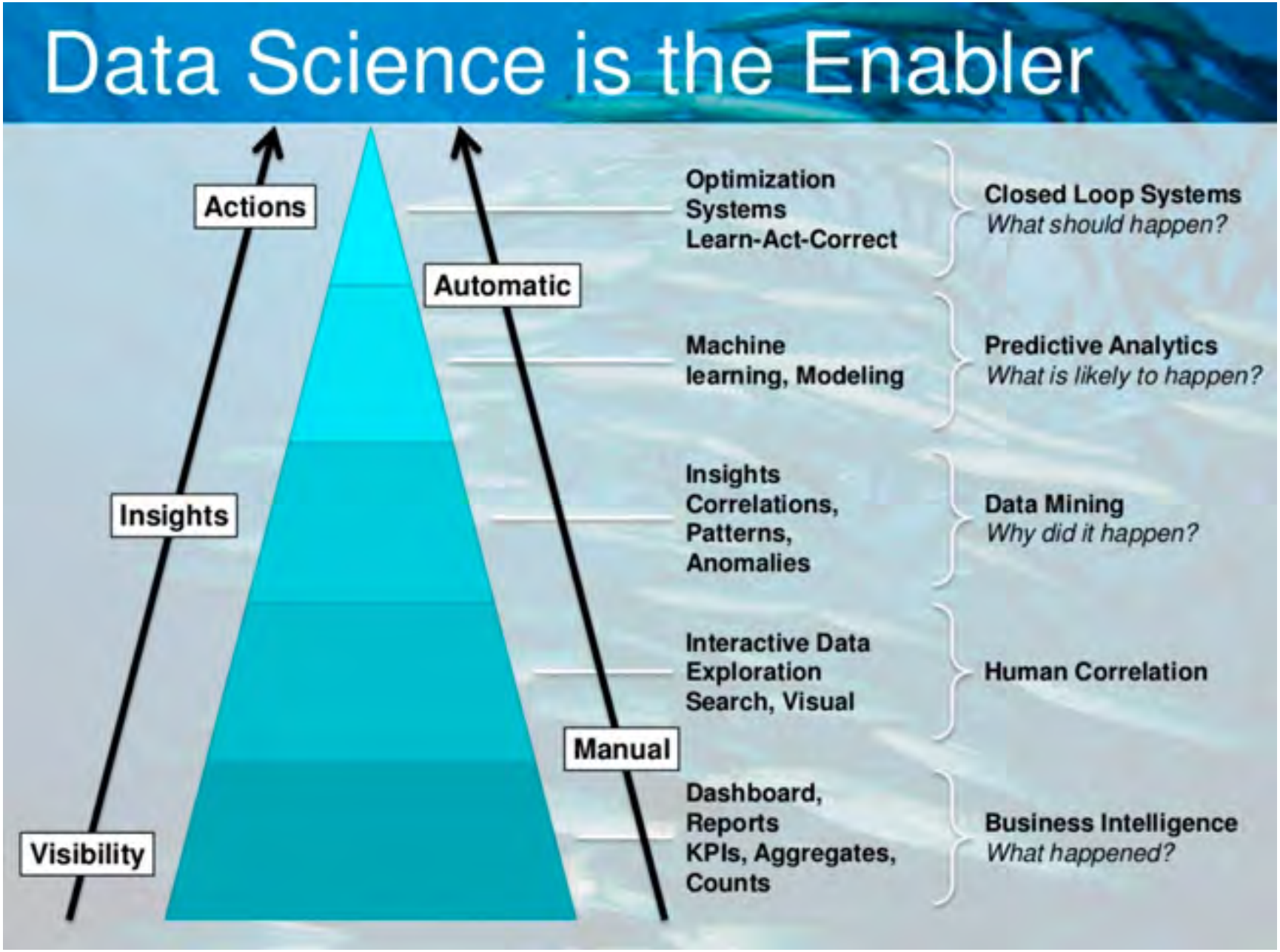
New trends in ML fashion: big data hiring boom

Big data brief history



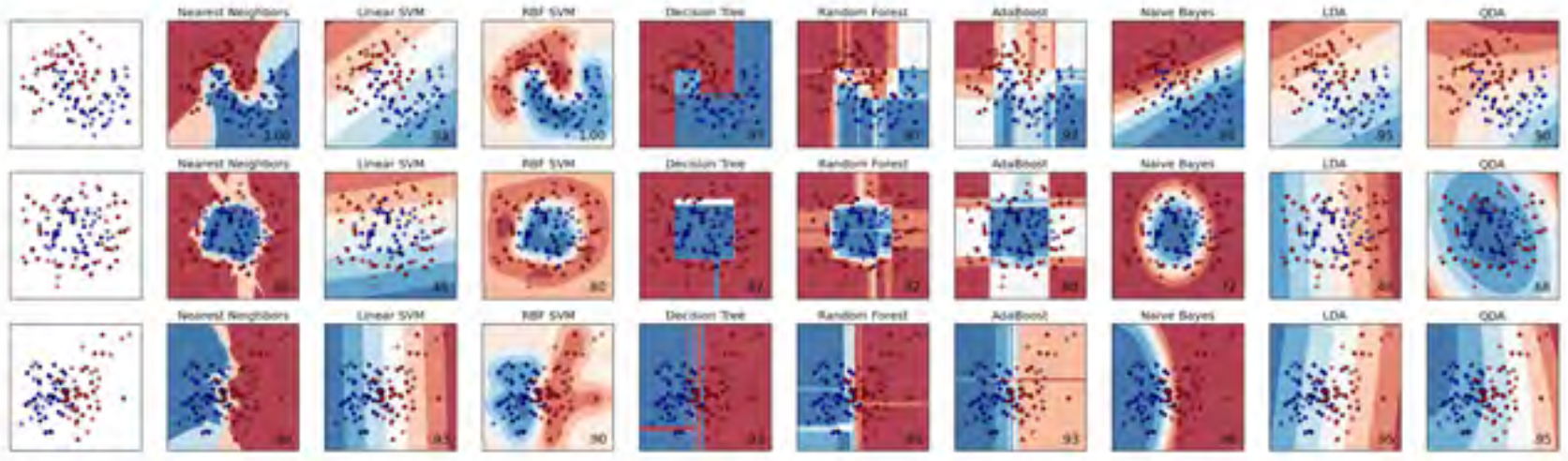
Original source at http://invent.ge/Uah9YO

Les tâches T

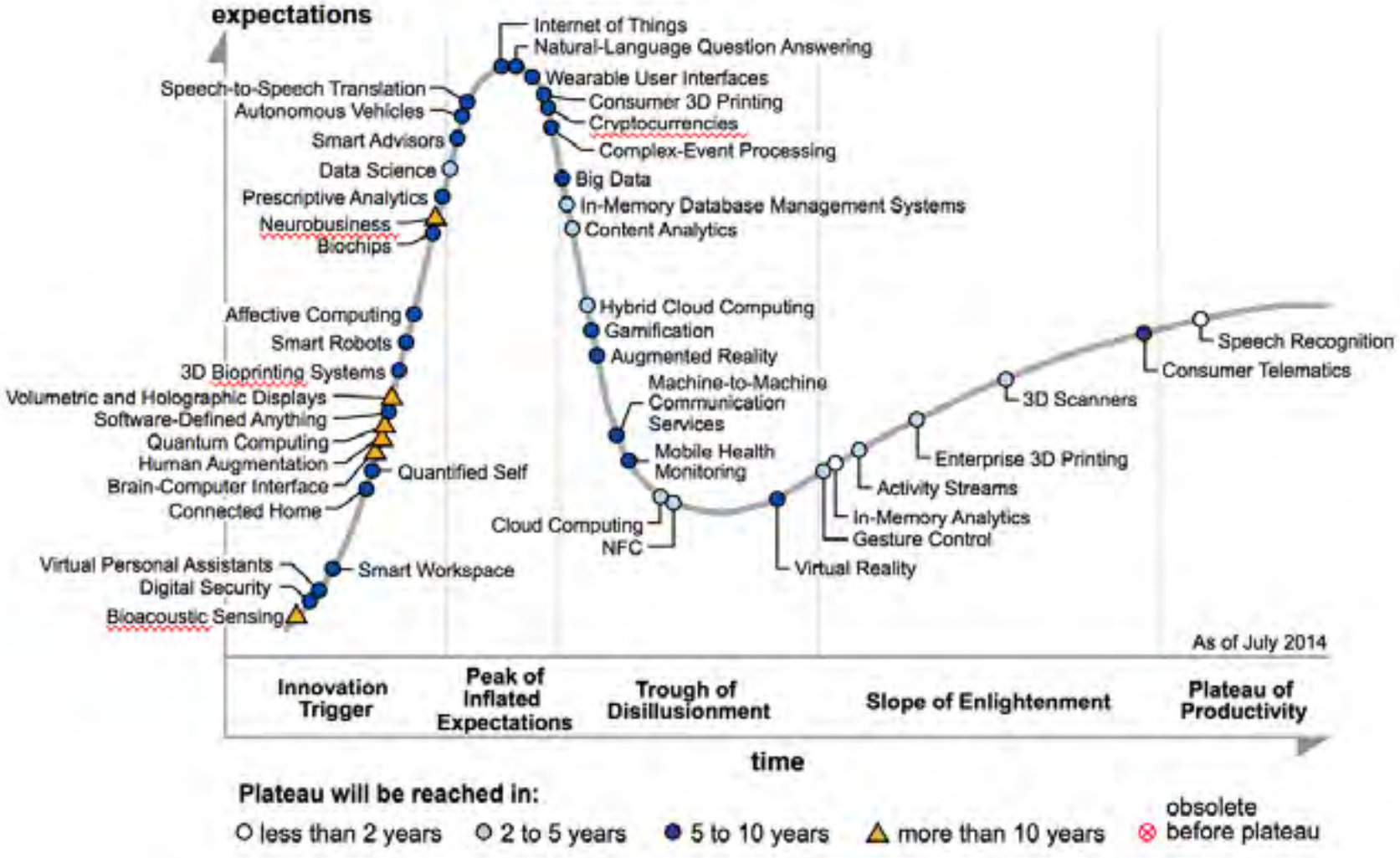


hivedata.com/

M.L. fashion as classifier comparison (from scikit-learn)

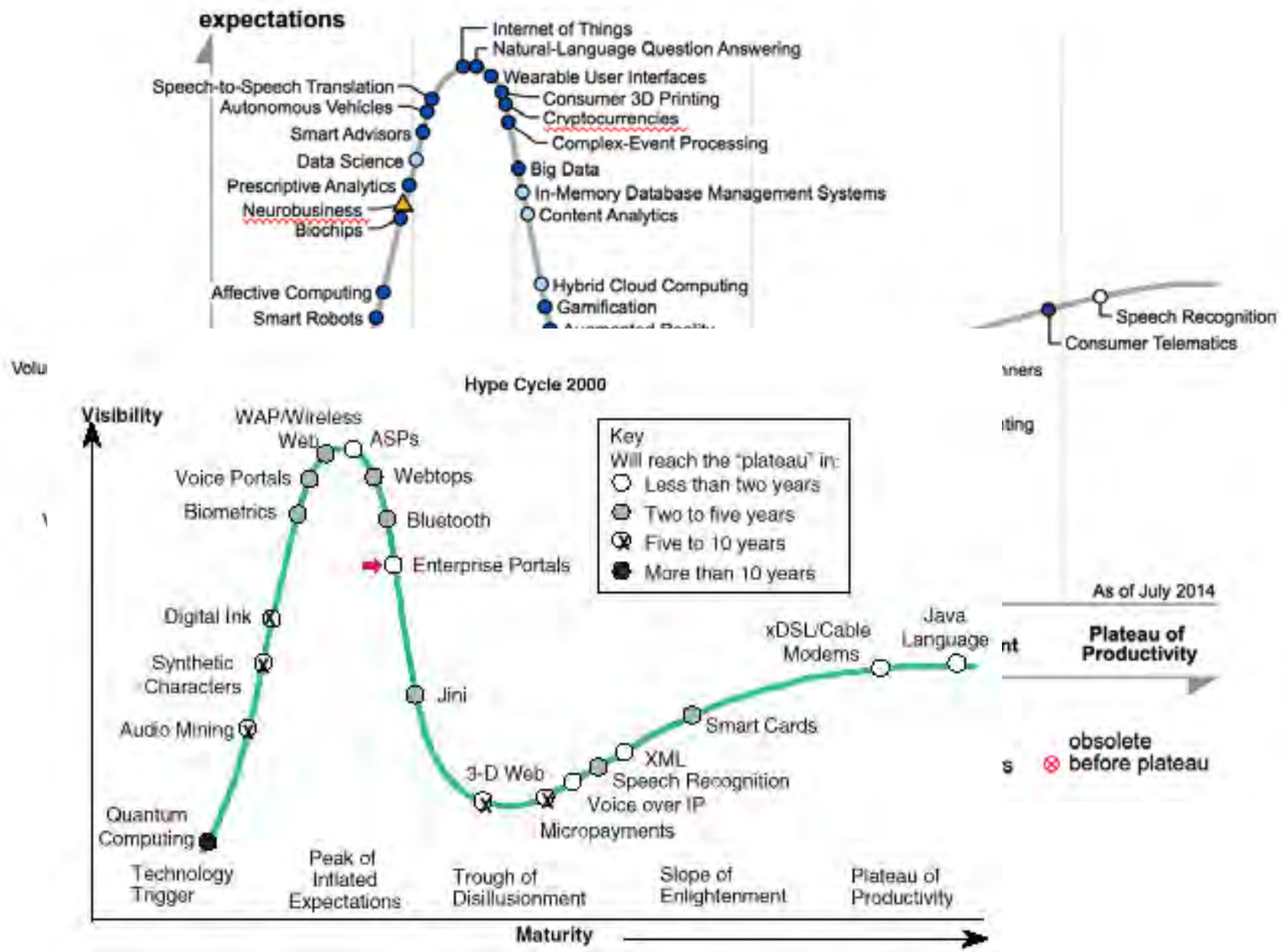


Une brève histoire de la mode en technology (by Gartner)



Gartner's 2014 Hype Cycle for Emerging Technologies Maps the Journey to Digital Business
http://www.gartner.com/newsroom/id/2819918

Une brève histoire de la mode en technology (by Gartner)



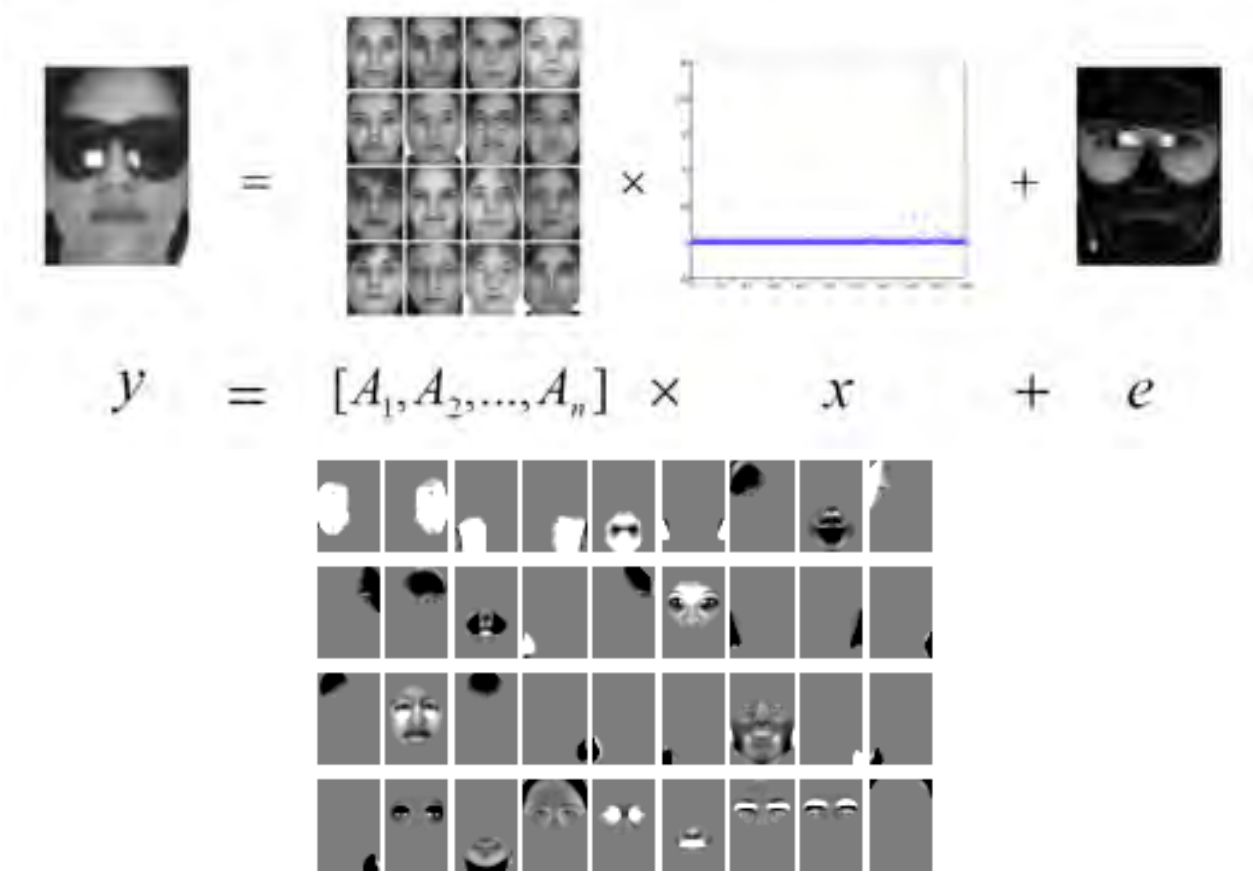
Gartner's 2000 Hype Cycle for Emerging Technologies Maps the Journey to Digital Business

http://www.gartner.com/newsroom/id/2819918

Quelles propriétés

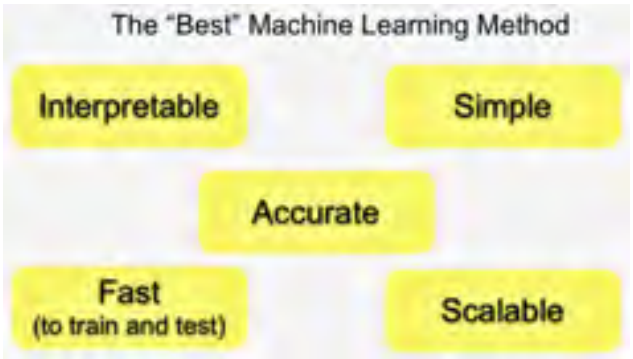
passage à l'échelle $\rightarrow$ parcimonie : sélectionner les exemples et les variables

randomisation $\rightarrow$ traiter la masse par le hasard



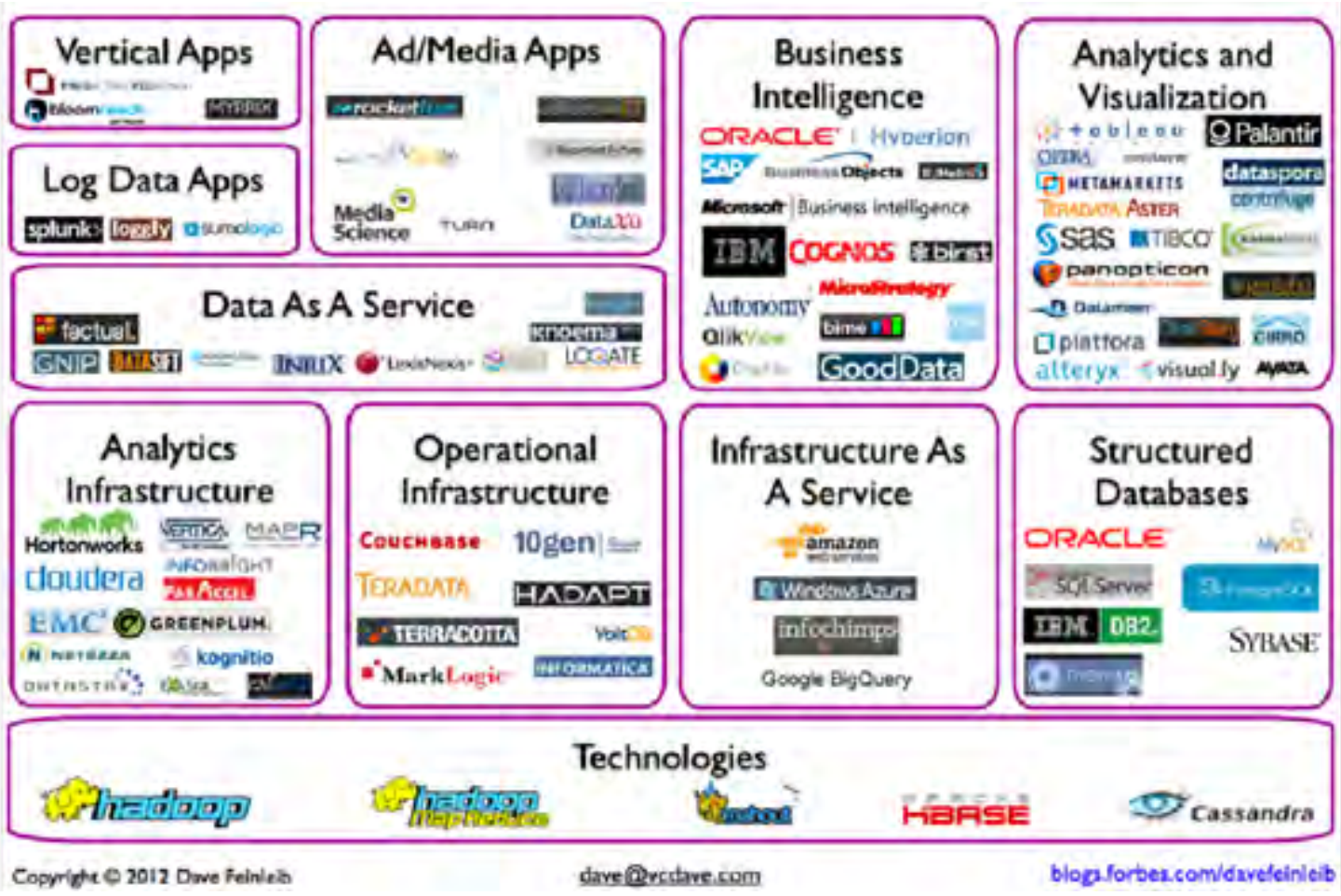
Structured Sparse Principal Component Analysis (R. Jenatton, G. Obozinski and F. Bach)

Plan

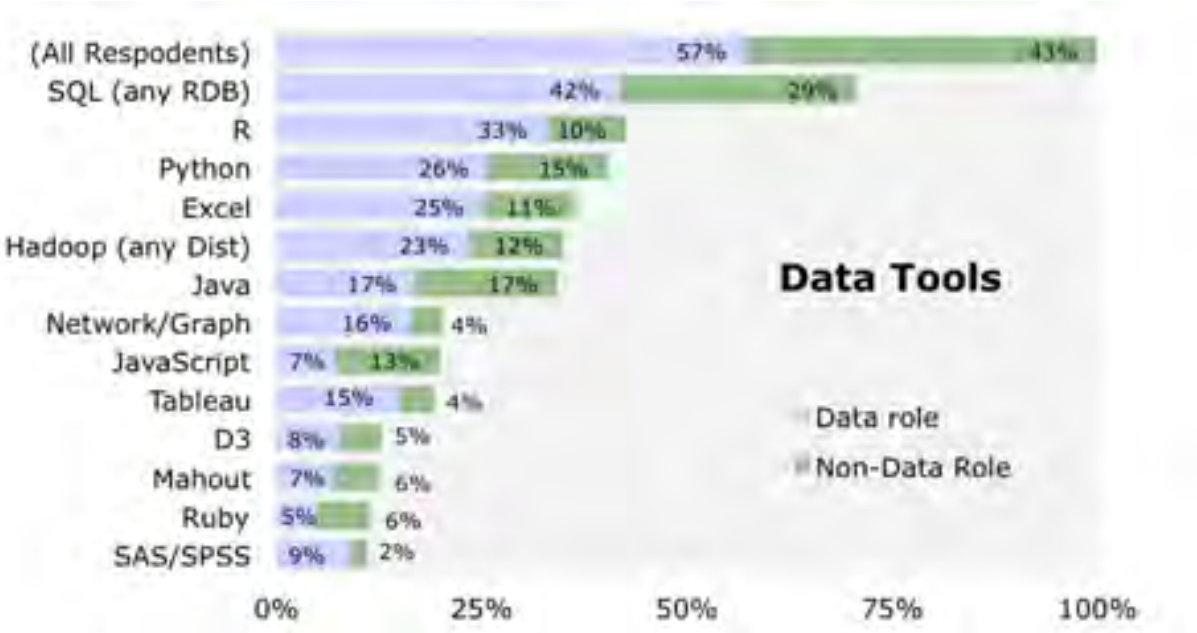


http://radar.oreilly.com/2013/09/gaining-access-to-the-best-machine-learning-methods.html

Les outils de la science des données

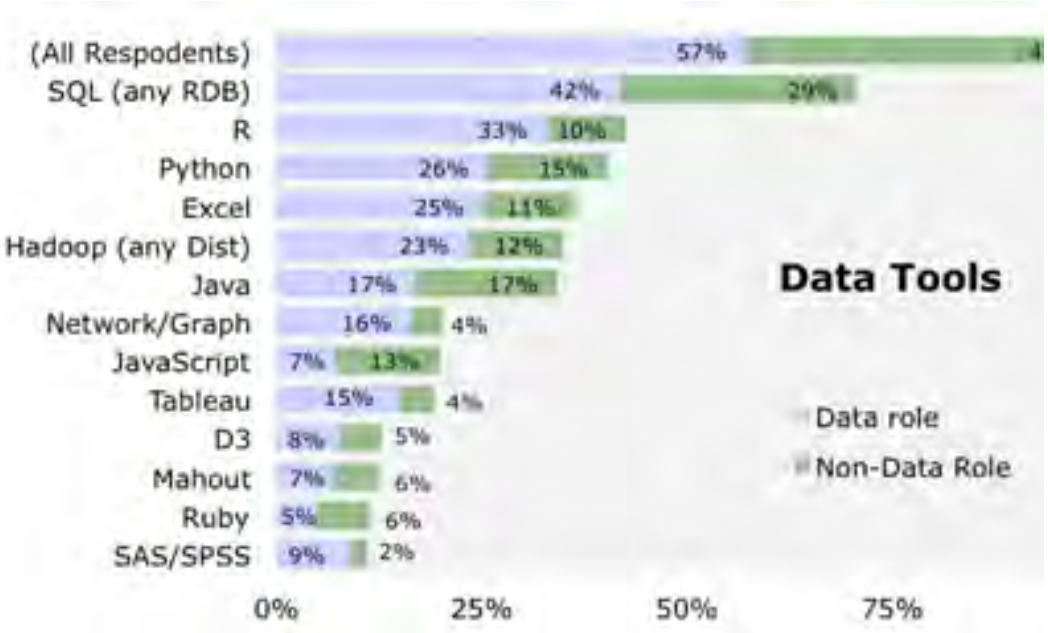


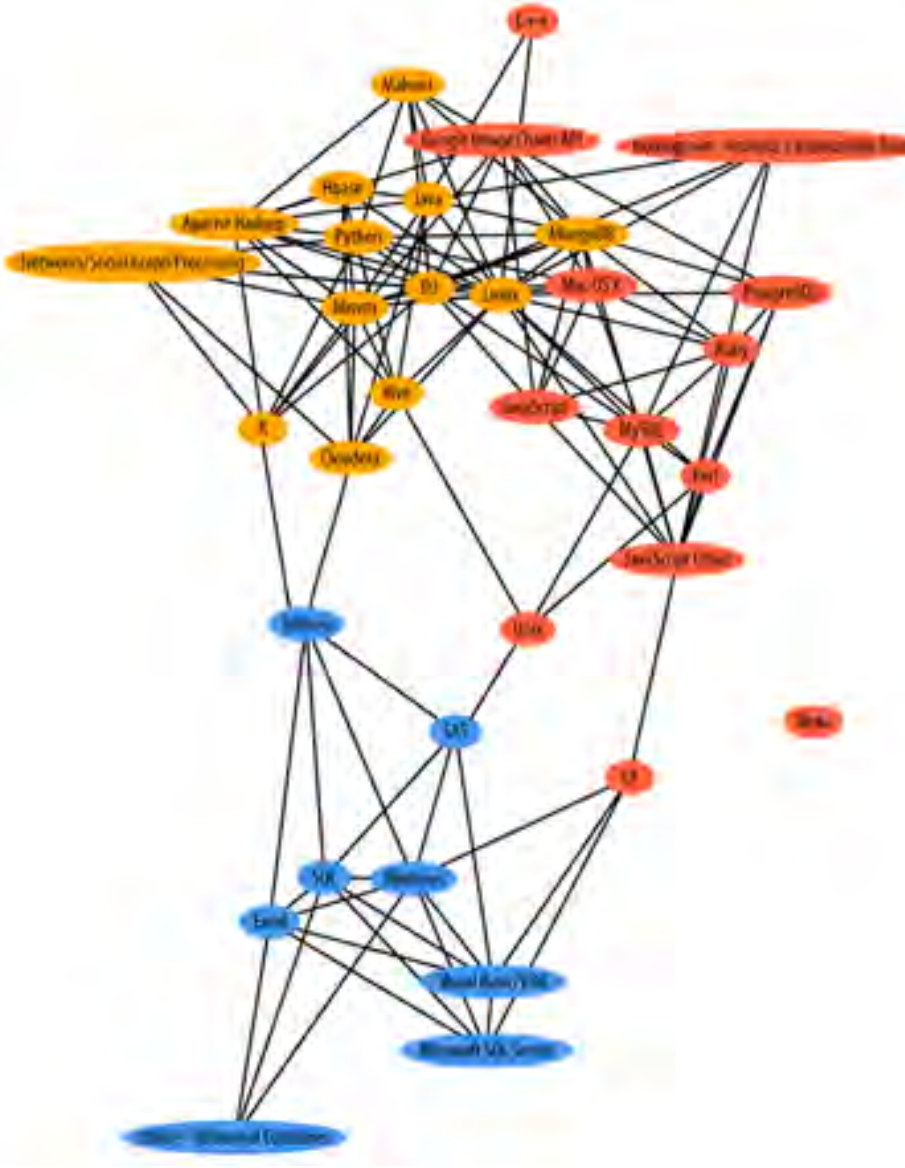
Les outils de la science des données



2013 *Data Science Salary Survey* `oreilly.com/data/free/files/stratasurvey.pdf`

Les outils de la science des données





2013 *Data Science Salary Survey* `oreilly.com/data/free/files/stratasurvey.pdf`

Logiciels (mloss.org)

environ. interactifs
- R (opensource)
- Matlab,

Java
- Weka,
- Rapid miner,
- MLlib (Spark), `spark.apache.org`

Python
- Orange, `http://orange.biolab.si`
- SciPy, `http://www.scipy.org`
- Shogun, MLpy, ...

C, C++
- Vowpal Wabbit, `hunch.net/~vw`
- Torch (Lua), `http://torch.ch/`
- libLinear, libSVM,

ML as a service
- Google prediction API,
- Microsoft Azure Machine Learning API,

Question méthode

1. définir objectifs, les contraintes et les couts

2. étude préliminaire
   - récupérer des données
   - construire un bac à sable
   - affiner les objectifs et les couts en interagissant avec les données
   - proposer un modèle et choisir une solution (un alto)

3. mise en oeuvre sur de véritable flots de données
   - randomisation
   - outils adaptés

Plan

Quelques prédictions... à propos du futur

- Data science - Applications
  - multi compétences
  - chaine de traitements

- Outils pour l'apprentissage
  - l'apprentissage sans paramètres (off-the-shelf)
  - passage à l'échelle (4v - big data - mégadonnées)

- Algorithmes d'apprentissage
  - optimisation (mégadonnées, non convexe)
  - dynamique (interactions)
  - apprendre à apprendre (transfert)

- Théorie de l'apprentissage
  - la nature de l'information

Pour en savoir plus

- livres
  - Bishop, C. M. 1995. Neural Networks for Pattern Recognition. Oxford: Oxford University Press.
  - Duda, R. O., P. E. Hart, and D. G. Stork. 2001. Pattern Classification, 2nd ed. New York: Wiley.
  - Hastie, T., R. Tibshirani, and J. Friedman. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.
  - A. Cornuéjols & L. Miclet, "L'apprentissage artificiel. Concepts et algorithmes". Eyrolles. 2 ème éd. 2010.
- conférences
  - xCML, NIPS, COLT, séminaire SMILE, Paris Machine Learning Applications Group
- revues
  - JMLR, Machine Learning, Foundations and Trends in Machine Learning, machine learning survey `http://www.mlsurveys.com/`