De la statistique à la Data Science

Eric Matzner-Løber

ENSAE-ENSAI formation continue

http://www.lecepe.fr

ENSAE-ENSAI
FORMATION CONTINUE

Un contexte nouveau

- Des données
  - multiples et dans de nombreux domaines
  - une variété forte
  - une volumétrie importante

Un contexte nouveau

- Des données
  - multiples et dans de nombreux domaines
  - une variété forte
  - une volumétrie importante
- Des unités de calcul abordables
  - ordinateurs, serveurs
  - cloud computing

Un contexte nouveau

- Des données
  - multiples et dans de nombreux domaines
  - une variété forte
  - une volumétrie importante
- Des unités de calcul abordables
  - ordinateurs, serveurs
  - cloud computing
- Un intérêt académique et industriel en forte croissance.

Data Science

Four major influences act today :

- The formal theories of statistics
- Accelerating developments in computers and display devices
- The challenge, in many fields, of more and ever larger bodies of data
- The emphasis on quantification in an ever wider variety of disciplines

Data Science

Four major influences act today :

- The formal theories of statistics
- Accelerating developments in computers and display devices
- The challenge, in many fields, of more and ever larger bodies of data
- The emphasis on quantification in an ever wider variety of disciplines

Tukey 1962, he was talking of Data Analysis !

Evolution

- ...
- Avant 1985, Statistical Data Analysis
- 1985 - 2005, Data Mining, Business Intelligence
- 2005 —-, Big Analytics

Nous sommes donc dans l'ère du **Big** avec des domaines d'applications variés.

Evolution

- ...
- Avant 1985, Statistical Data Analysis
- 1985 - 2005, Data Mining, Business Intelligence
- 2005 —-, Big Analytics

Nous sommes donc dans l'ère du **Big** avec des domaines d'applications variés.

Les moyens (humains et techniques), les méthodes, les outils ?

Les méthodes ?

- Régression biaisée (1970 ridge) puis différentes contraintes de norme (2000...)

Les méthodes ?

- Régression biaisée (1970 ridge) puis différentes contraintes de norme (2000...)
- Perceptron de Rosenblatt (1957), réseau de neurones

Les méthodes ?

- Régression biaisée (1970 ridge) puis différentes contraintes de norme (2000...)
- Perceptron de Rosenblatt (1957), réseau de neurones
- Méthodes nonparamétriques (ppv, noyau, splines) 1960

Les méthodes ?

- Régression biaisée (1970 ridge) puis différentes contraintes de norme (2000...)
- Perceptron de Rosenblatt (1957), réseau de neurones
- Méthodes nonparamétriques (ppv, noyau, splines) 1960
- Arbres (1980) et forêts (fin des années 1990)

Les méthodes ?

- Régression biaisée (1970 ridge) puis différentes contraintes de norme (2000...)
- Perceptron de Rosenblatt (1957), réseau de neurones
- Méthodes nonparamétriques (ppv, noyau, splines) 1960
- Arbres (1980) et forêts (fin des années 1990)
- Support vector machine (1990)

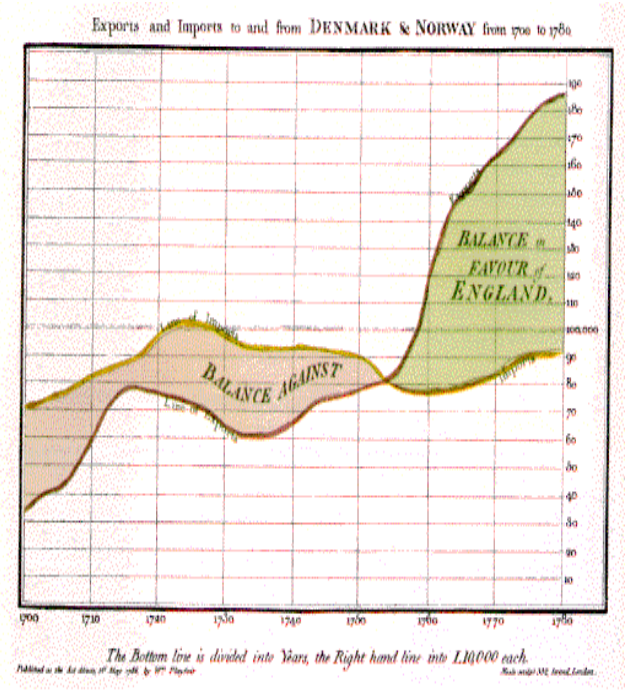
Les méthodes ?

- Régression biaisée (1970 ridge) puis différentes contraintes de norme (2000...)
- Perceptron de Rosenblatt (1957), réseau de neurones
- Méthodes nonparamétriques (ppv, noyau, splines) 1960
- Arbres (1980) et forêts (fin des années 1990)
- Support vector machine (1990)
- Méthodes d'aggrégation (années 2000)

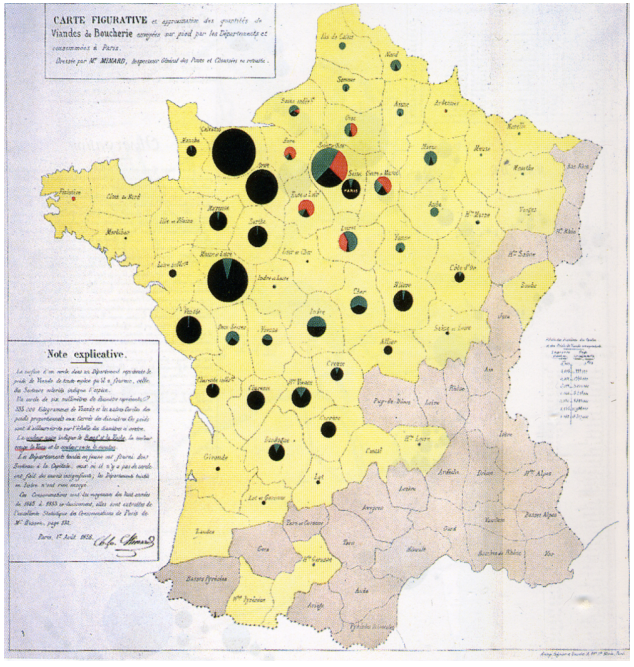
Les méthodes ?

- Régression biaisée (1970 ridge) puis différentes contraintes de norme (2000...)
- Perceptron de Rosenblatt (1957), réseau de neurones
- Méthodes nonparamétriques (ppv, noyau, splines) 1960
- Arbres (1980) et forêts (fin des années 1990)
- Support vector machine (1990)
- Méthodes d'aggrégation (années 2000)
- Analyses textuelles, analyse des réseaux

Visualisation



Playfair 1759-1823



Minard 1781-1870

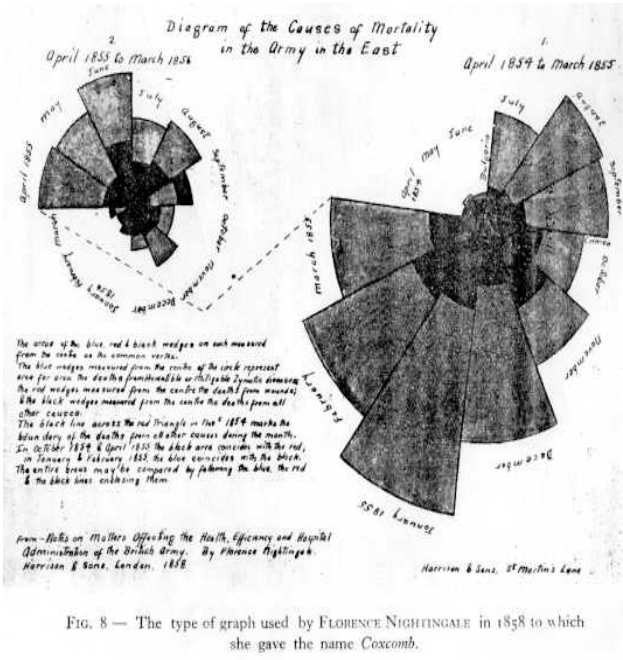


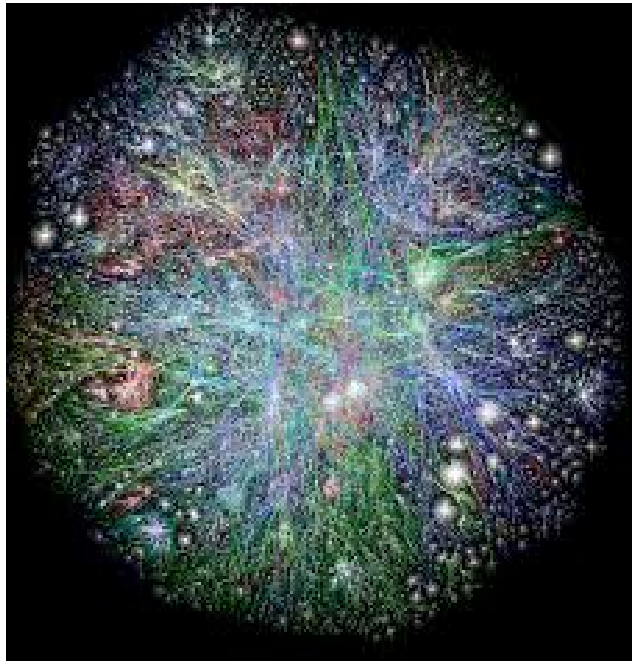
FIG. 8 — The type of graph used by FLORENCE NIGHTINGALE in 1858 to which she gave the name *Coxcomb*.

Nightingale 1820-1910

J. Snow (1813–1858), R. Fisher (1890-1962), J. Tukey (1915-2000), J. Bertin (1918-2010), E. Tufte (1942–)...

Navigation dans les données, support à des pistes de réflexion...

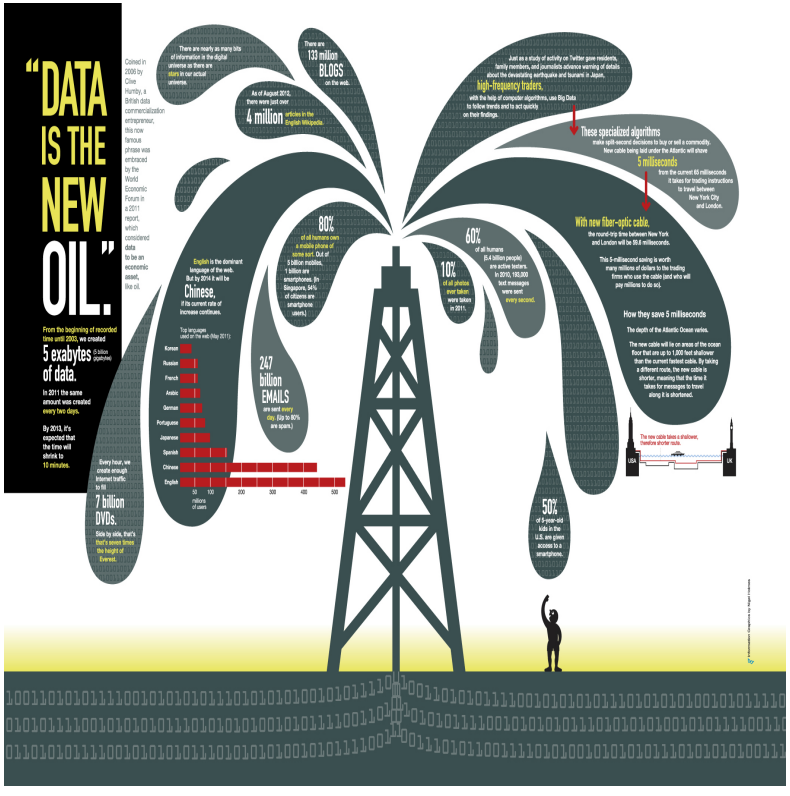
Analyse des graphes



Konigsberg bridge problem, Euler 1736 !

La perception des autres

Articles dans la presse, rapport d'*experts*,



Bits

OCTOBER 24, 2012, 9:00 AM | 4 Comments

Big Data in More Hands

By QUENTIN HARDY

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT

Business people, Big Data is coming for you.

Software that captures lots of data and uses it to make predictions has mostly been the province of engineers skilled in arcane databases and statisticians capable of developing complex algorithms. As the business gets bigger, however, software makers are domesticating their products in the hope they will prove attractive to a broader population.

Cloudera, which offers a popular version of the open source database called Hadoop, released software on Wednesday that makes it possible to run queries from a more mainstream SQL programming language interface. SQL, thanks to its adoption by Oracle, Microsoft and others, is known to millions of business analysts.

"This enables us to talk to a whole other class of customer," said Mike Olson, the chief executive of Cloudera. "The knock against Hadoop was that it is too complex."

Volonté politique d'avoir son **lab**, ses **POC**, ses équipes de data-scientistes.

Nouveauté

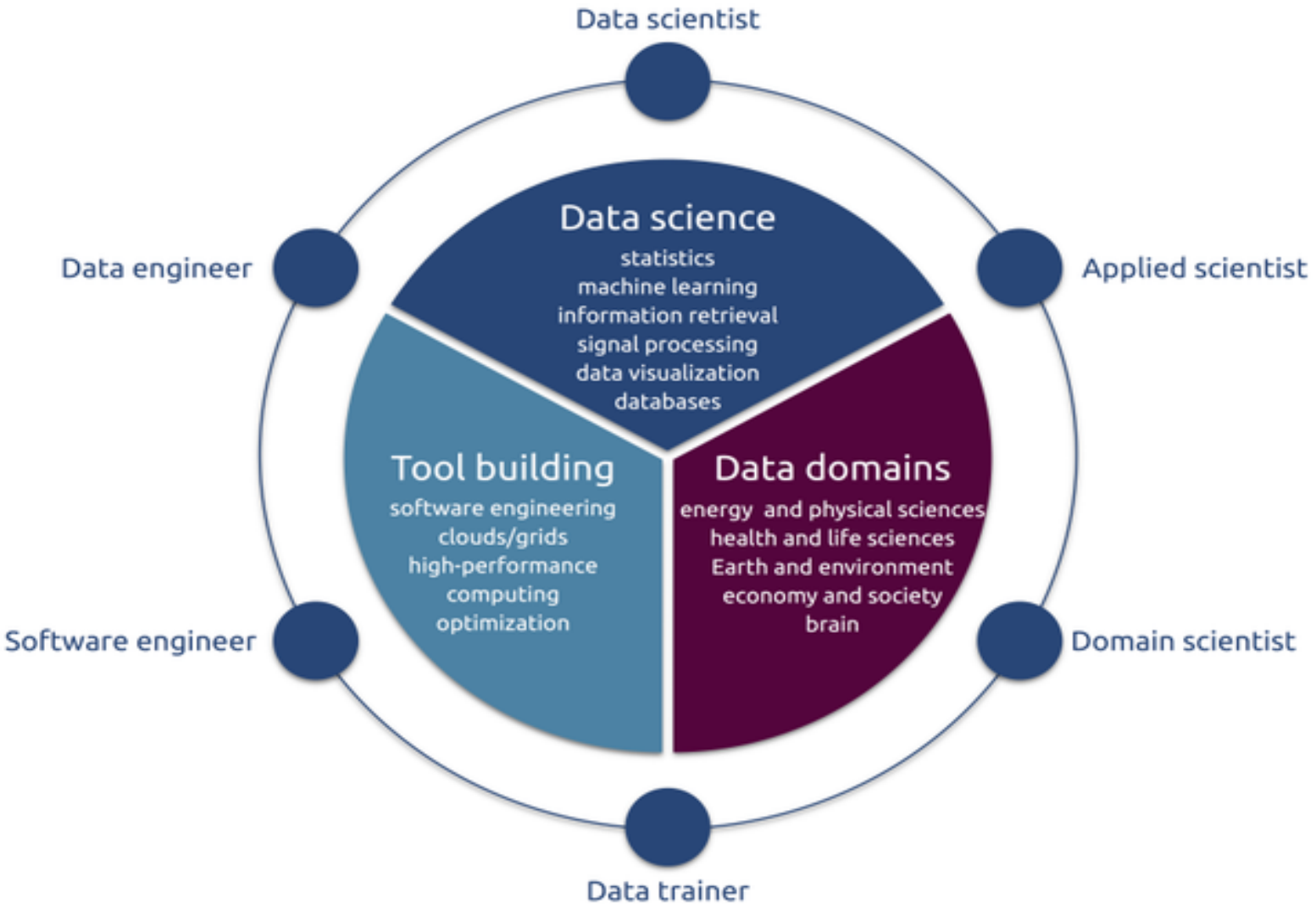
- Les données sont de plus en plus prises en compte en interne (prix de la mesure et du stockage en baisse)
- des données externes accessibles (open-data)
- des logiciels de traitements libres (R, Python, Julia)
- des moyens de calcul performants abordables
- une demande forte avec des programmes de formation en léger décalage

Des équipes pluridisciplinaires...

Le data scientiste

- Métier : business analyste, connaissance marché...
- Data Management : collecter, stocker, nettoyer
- Statistique et Machine learning : modélisation, prédiction, optimisation...
- Programmation : développement, passage à l'échelle...
- Interface : visualisation, navigation...

Data Scientiste



Compétences suite

- En Statistique : modélisation, échantillonnage, prévision...
- En Mathématiques : optimisation..
- En Data engineering : base de données, architecture...
- En computer science : algorithmes, optimisation
- En connaissances métiers
- En code
- ....

Comment devenir un Data Scientist ?

- Des filières spécifiques à l'université
- Certaines grandes écoles d'ingénieurs
- Via la formation continue :
  - master spécifique ou mastère spécialisé
  - certificats
  - cours au catalogue
  - auto formation
  - ...

Conclusion

Le temps de passage de la recherche académique à l'industrialisation s'est drastiquement réduit. Il faut donc des data scientistes

- curieux
- appétents aux nouveautés
- prêts à expérimenter/tester.

Il faut donc leur laisser du temps pour se (auto)former, pour traiter des projets nouveaux, la possibilité de travailler avec des chercheurs académiques...

ENSAE-ENSAI
FORMATION CONTINUE