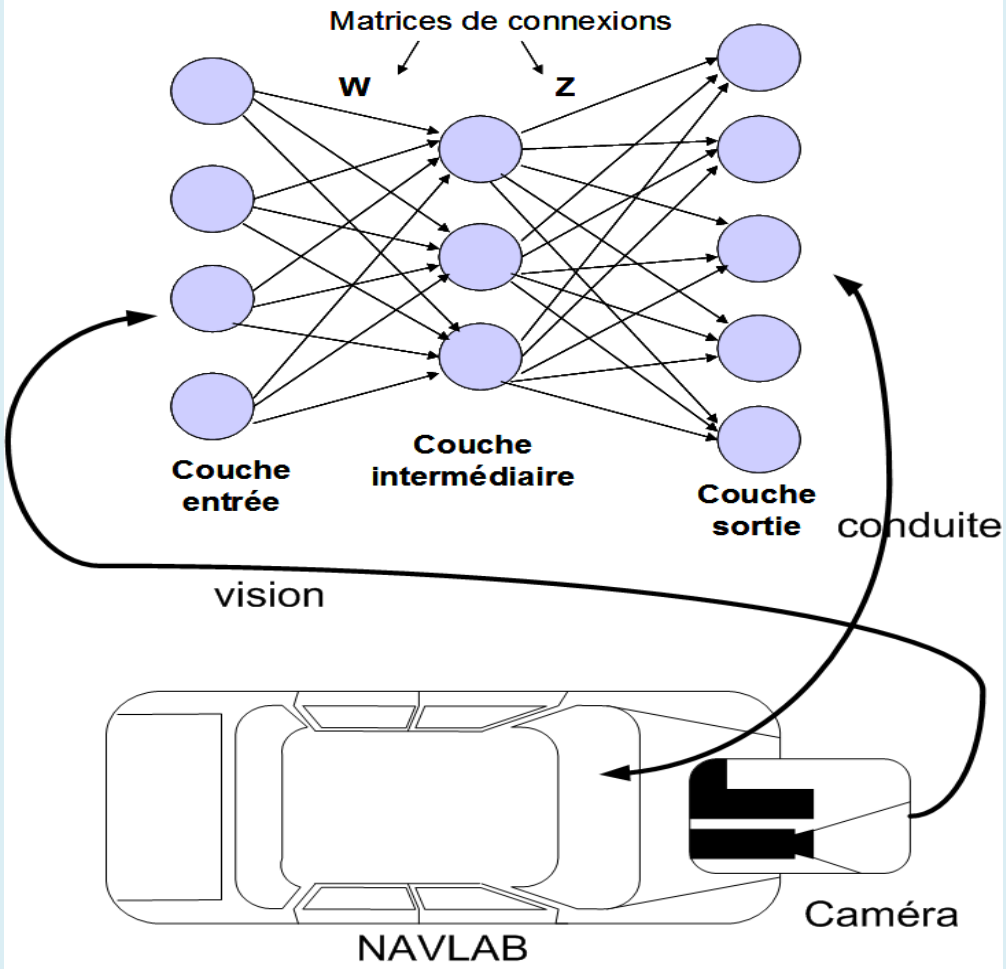
Data mining et statistiques

Hugues Bersini
IRIDIA / ULB

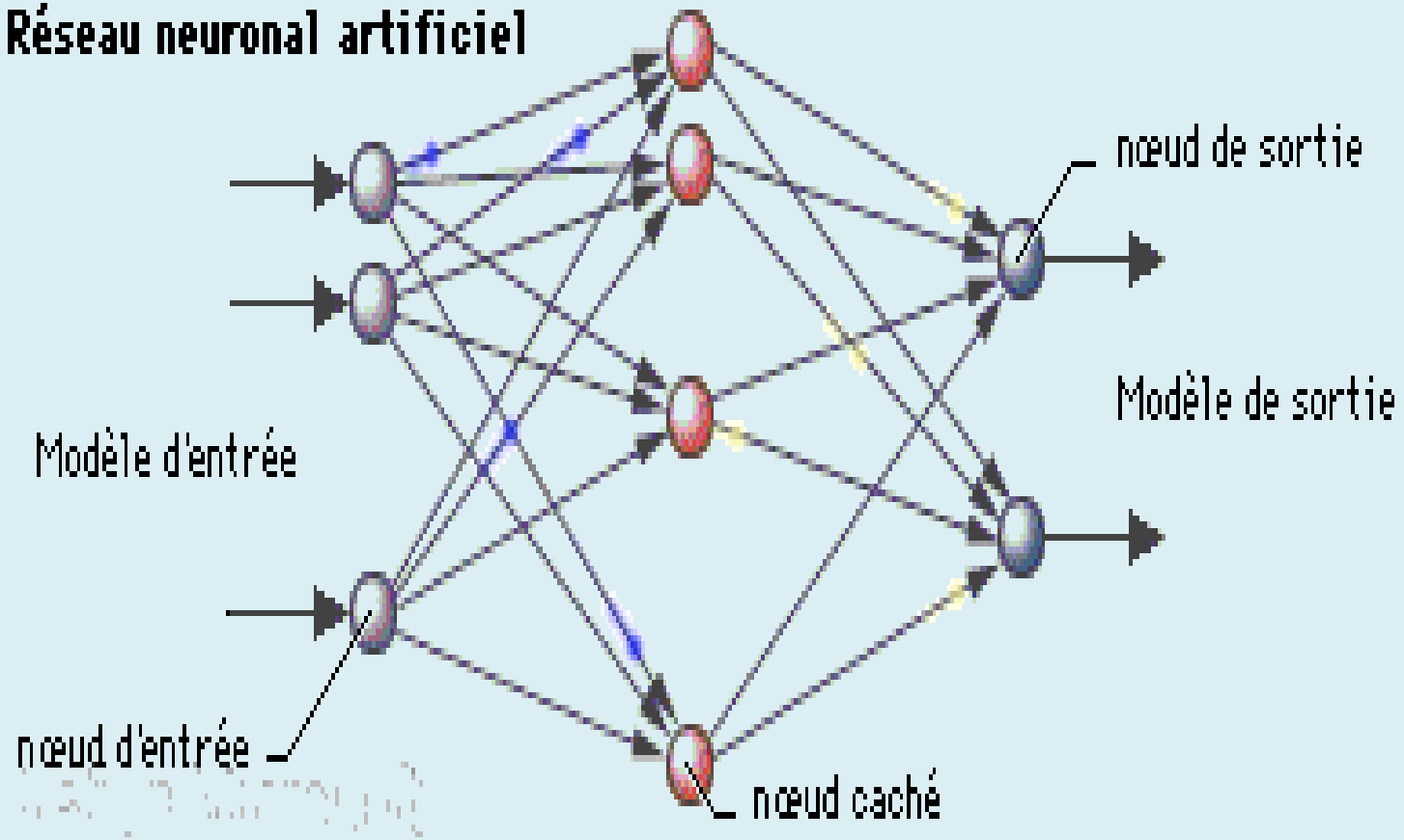
Philippe Smets (1938-2005), créateur d'IRIDIA

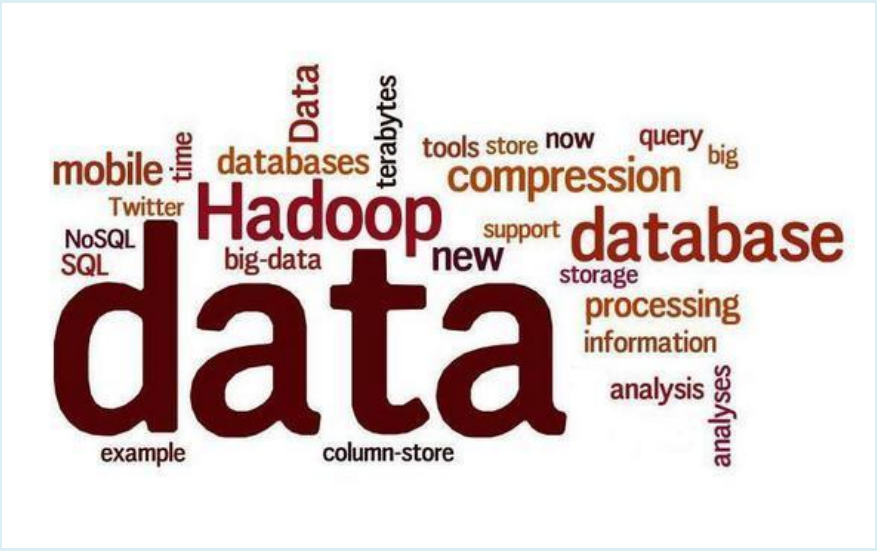
Théorie des croyances vs Probabilités

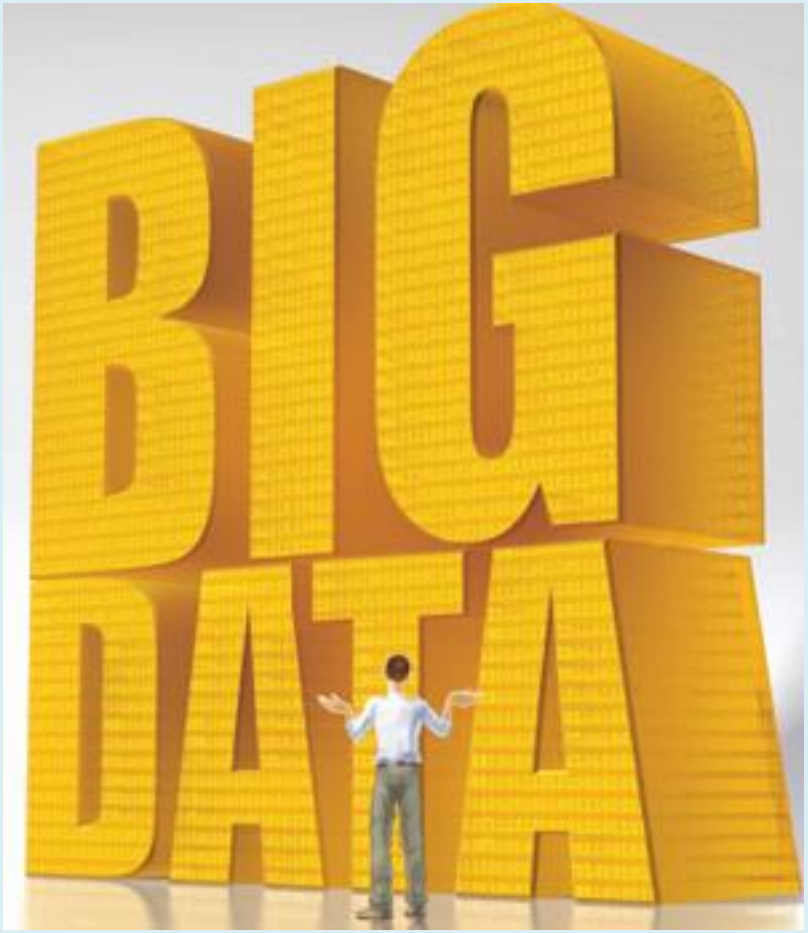
IA, IRIDIA et Data Mining

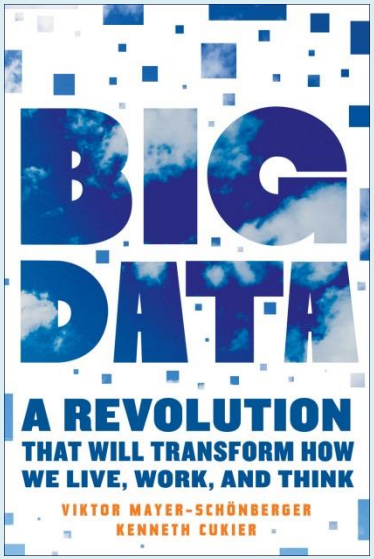


Les réseaux de neurones



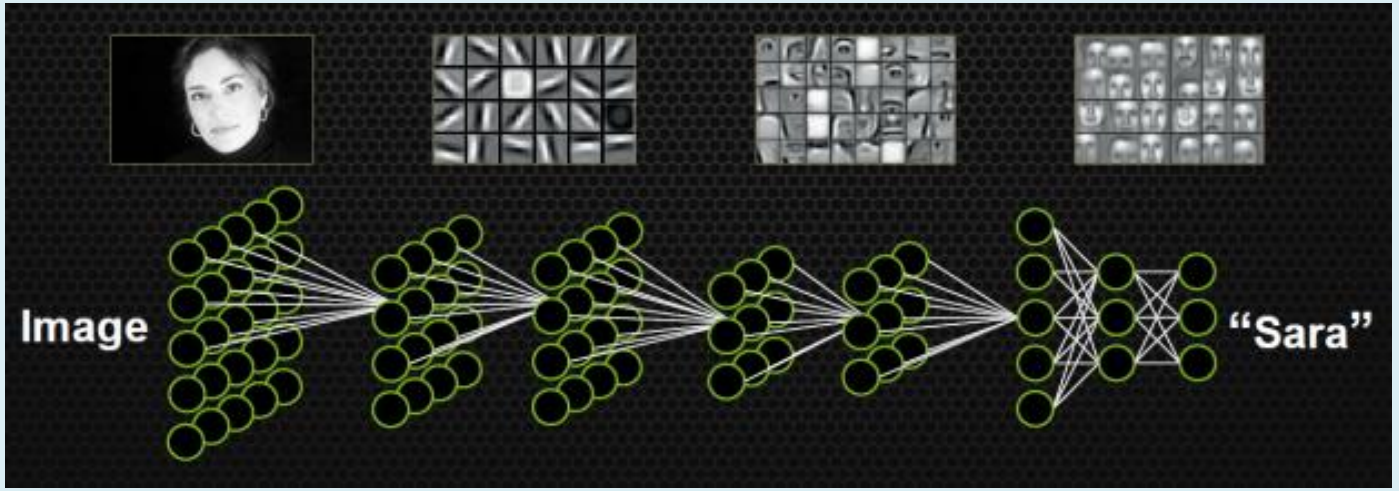


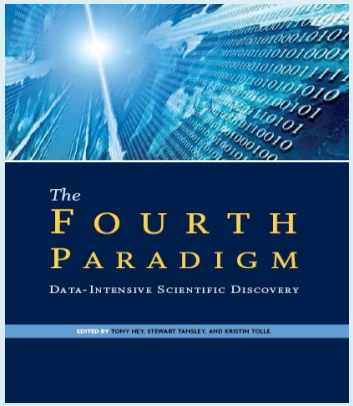




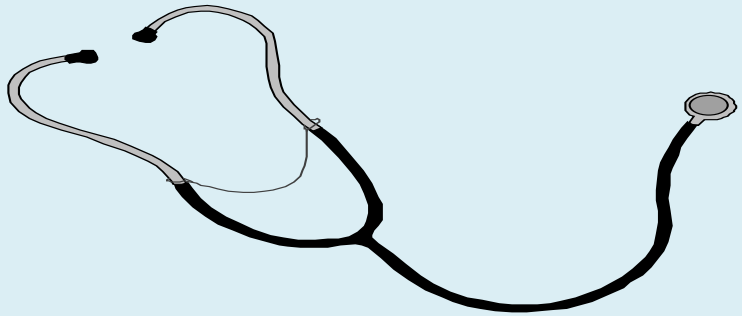
IA à la Google

Deep Learning: Reconnaissance d'images, traduction, conduite automatique,… (déterministe)



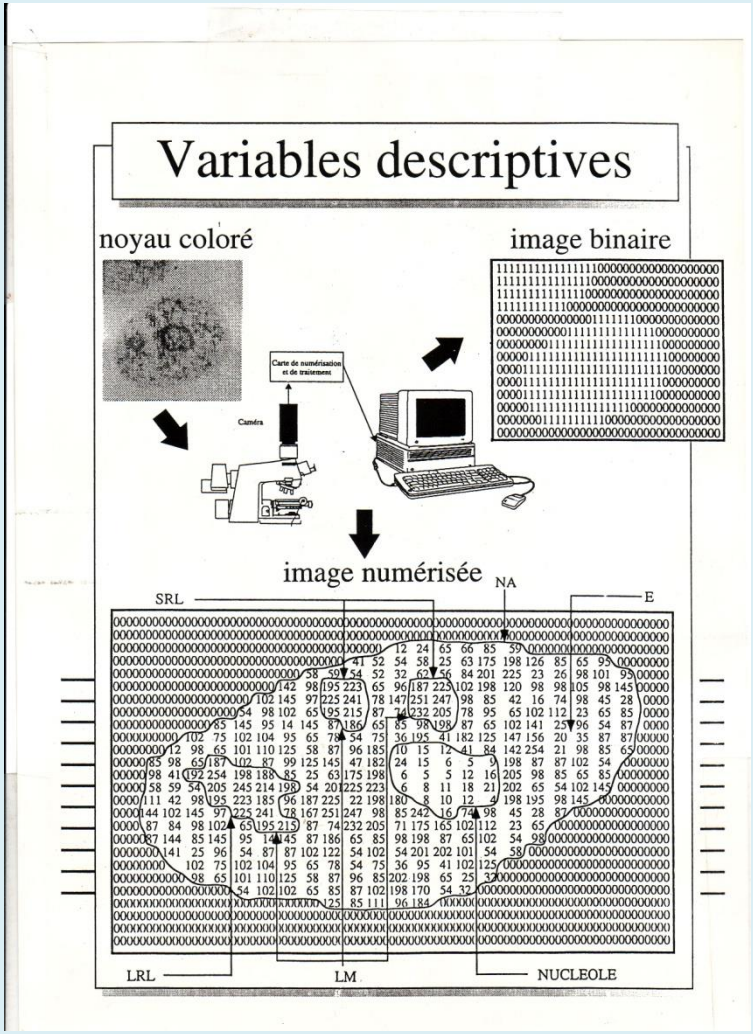


IRIDIA dans le domaine médical

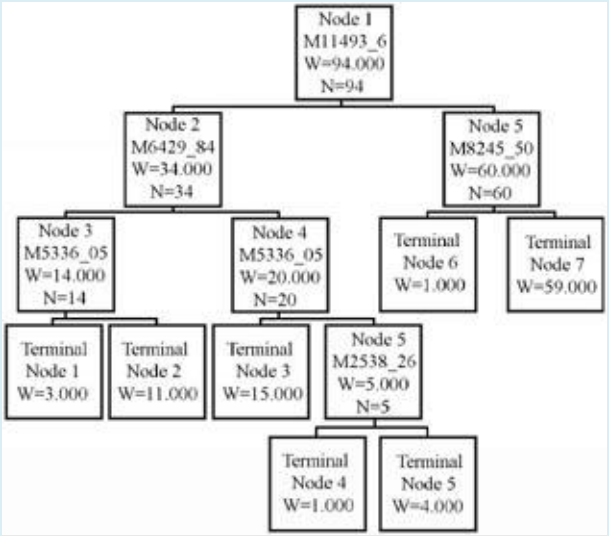


- Diagnostique automatique de cancer: Decaestecker, Van Ham, Kiss
- Détection de problèmes respiratoires: Mathys, Degroot, Kahn
- Analyse d'électrocardiogrammes: Bableyantz
- Aide à la marche assistée de paraplégiques: Preumont
- Aide à l'empowerment des diabétiques
- Classification des cancers par puce ADN: IRIBHM et Bordet
- Création de l'institut IB^2, projet BridgeIris: Bontempi

Gradation automatique des cancers



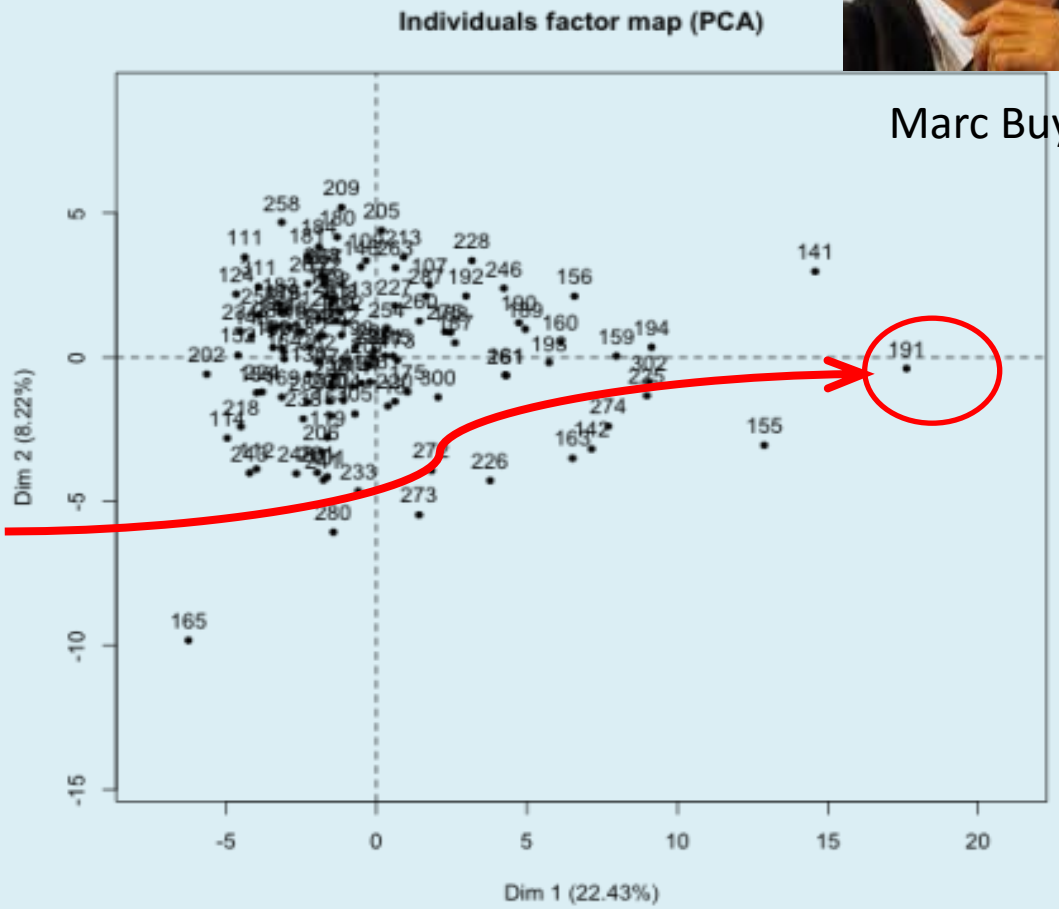
Christine Decaestecker



Cluepoints : detection of outlier clinical site

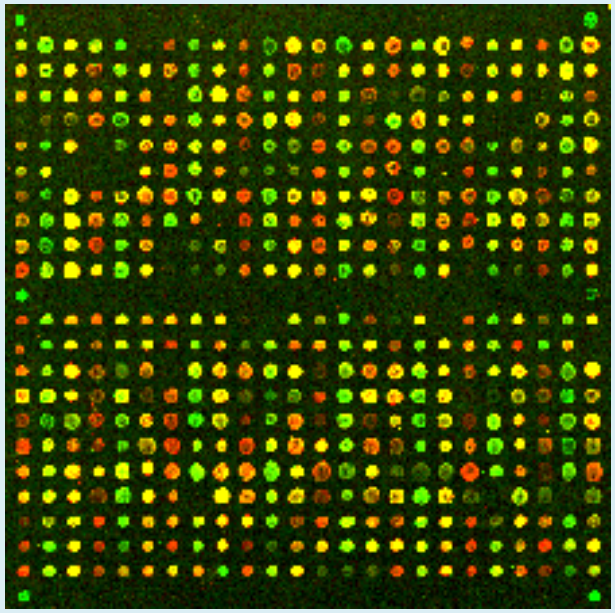
Marc Buyse

- Real example
  - Known fraud in center 191
- SMART analysis
  - 191 is an outlier
- Other centers?
  - 141, 155, 165?
  - Most frauds are undetected by current methods

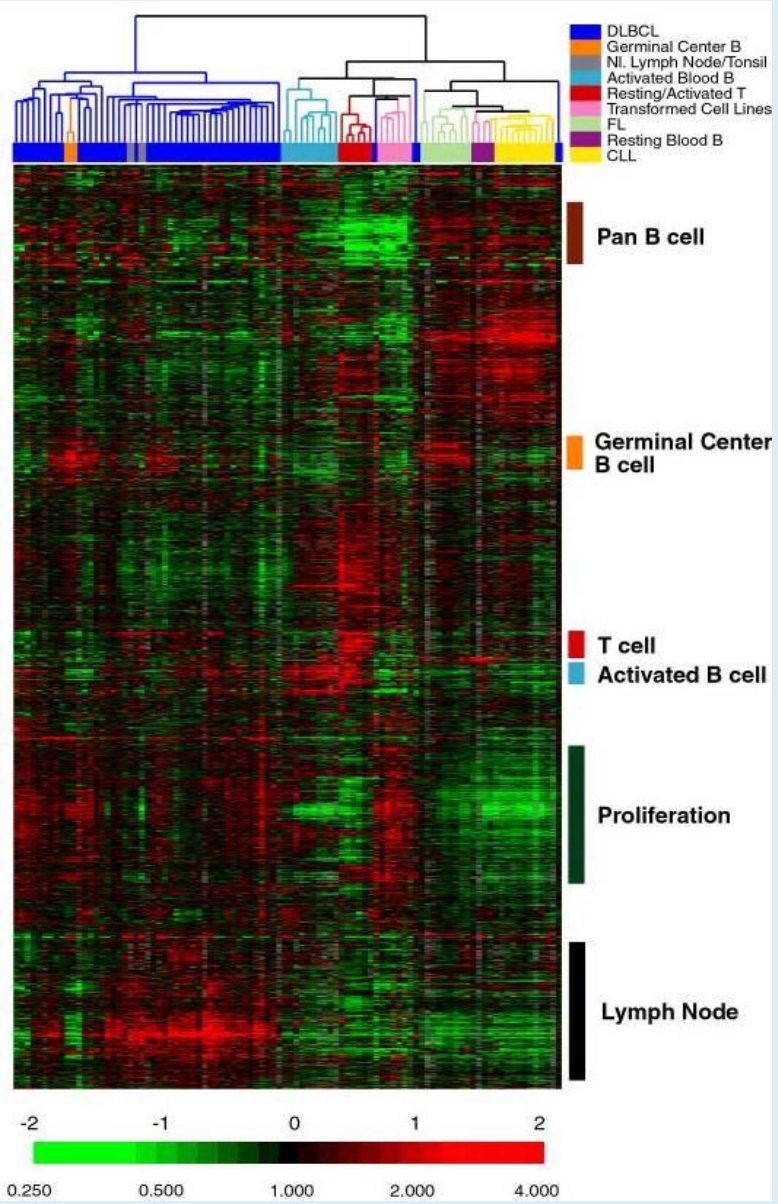


Summary through PCA of a SMART analysis

PUCE ADN



Microarray chip



Vincent Detours

Gianluca Bontempi

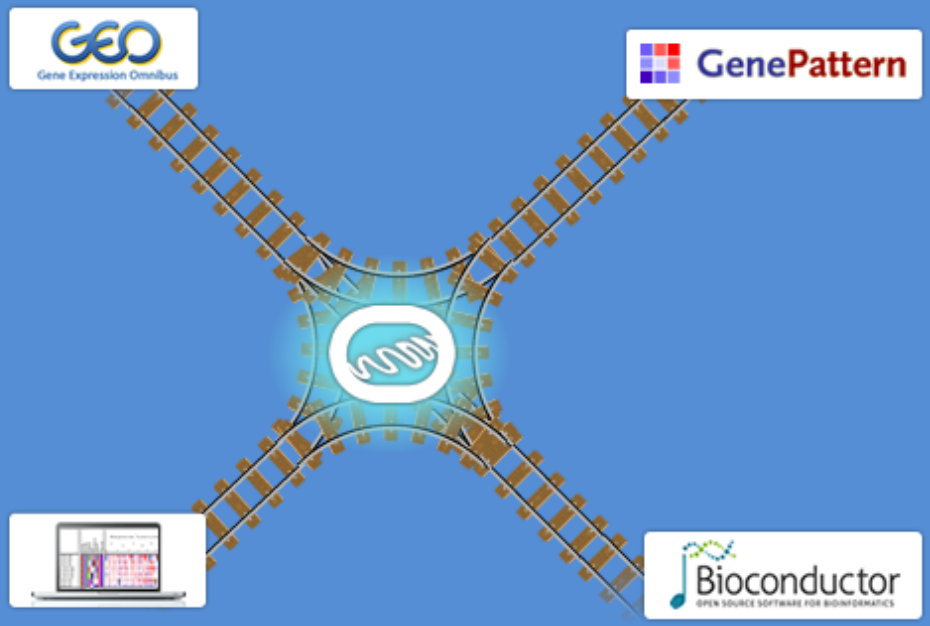
LA SPINOFF INSILICO

The concept: InSilico DB is a central data hub for genomics-based biomedical research

David Weiss et Alain Coletta

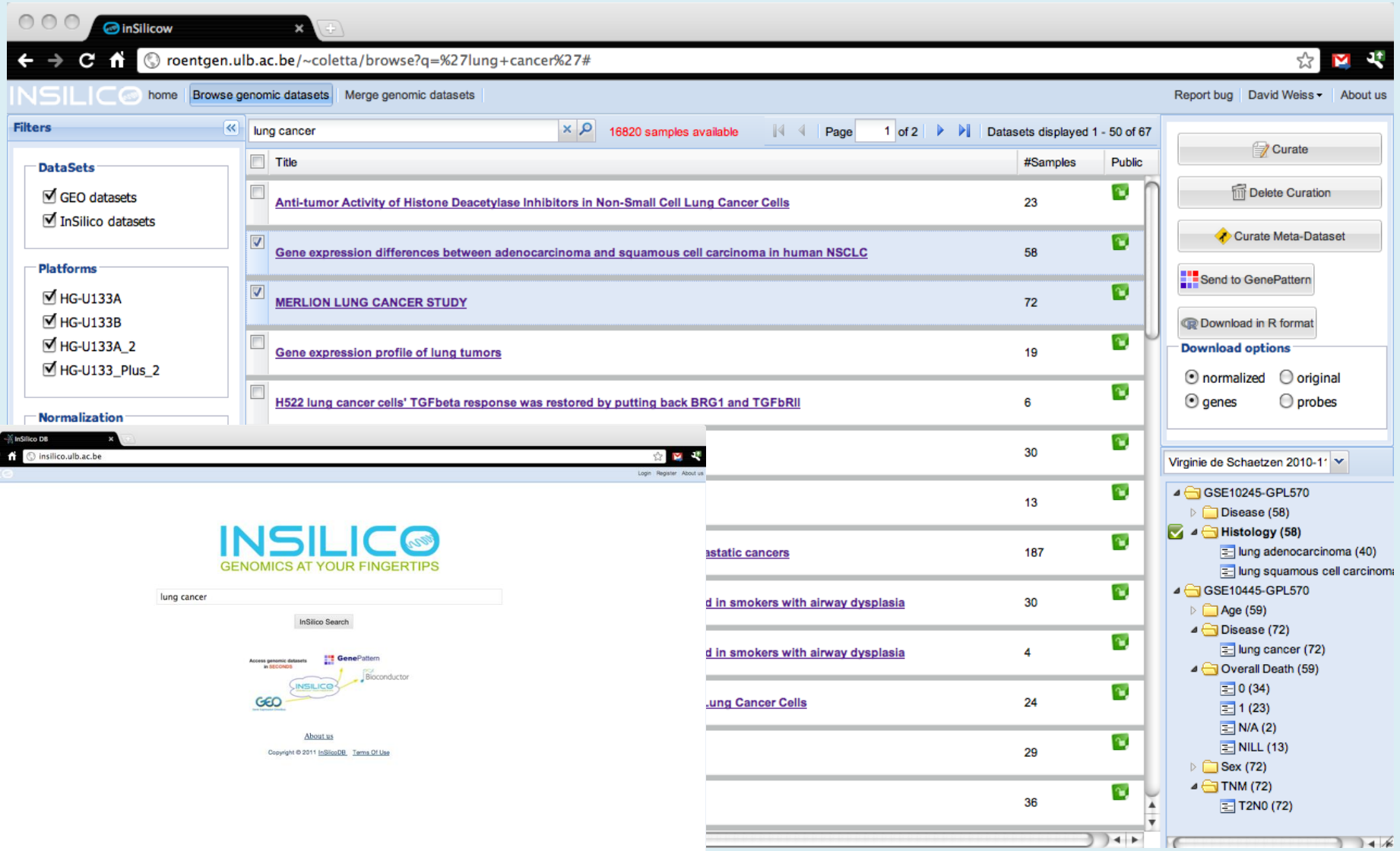
INSILICO
GENOMICS AT YOUR FINGERTIPS

YOUR GENOMICS DATA HUB
AN EFFICIENT STARTING POINT FOR GENOMICS ANALYSIS

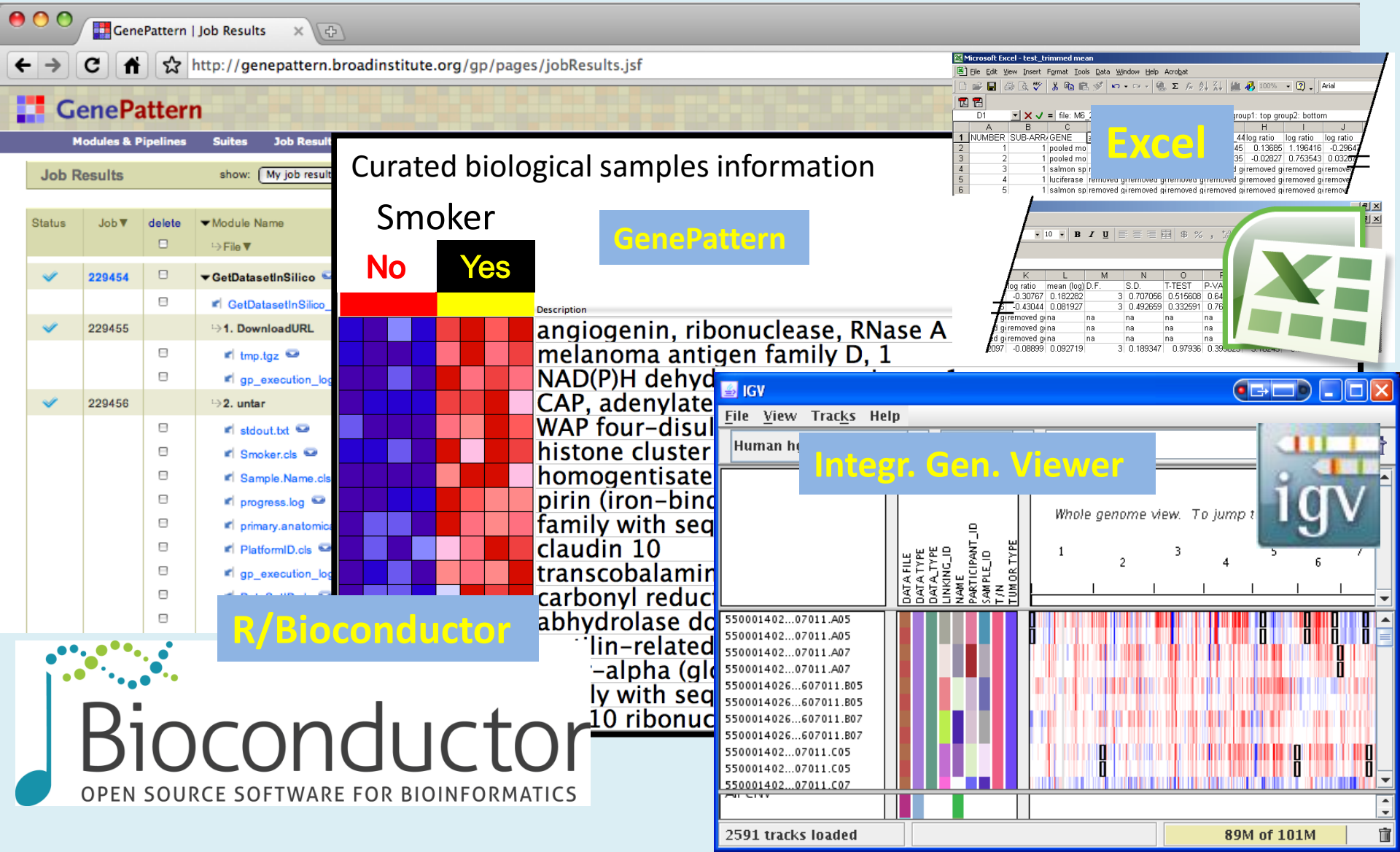


InSilico DB = Data + Tools

Website for genomics datasets management and query

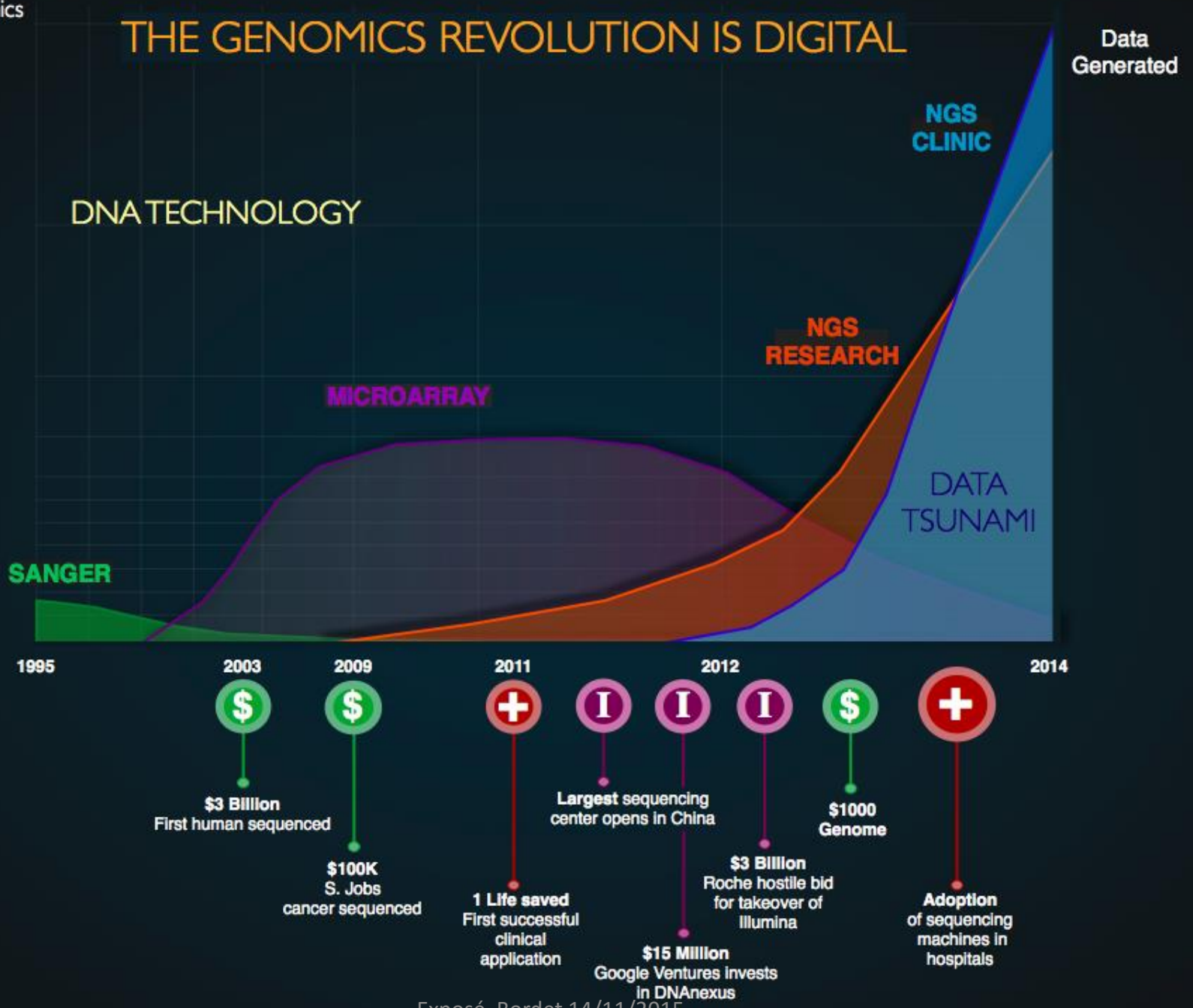


Integration with visualisation and analysis tools

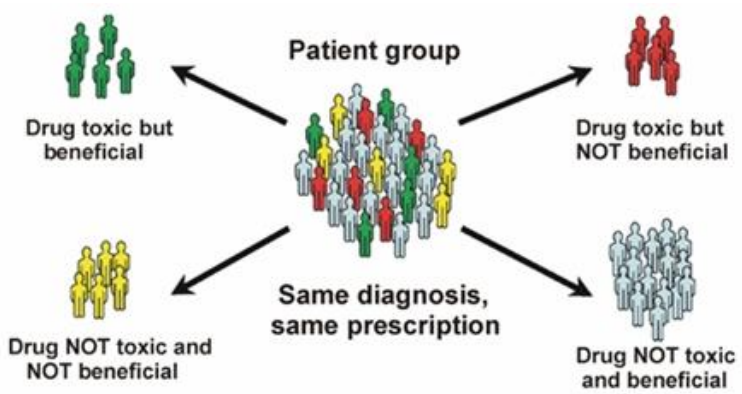


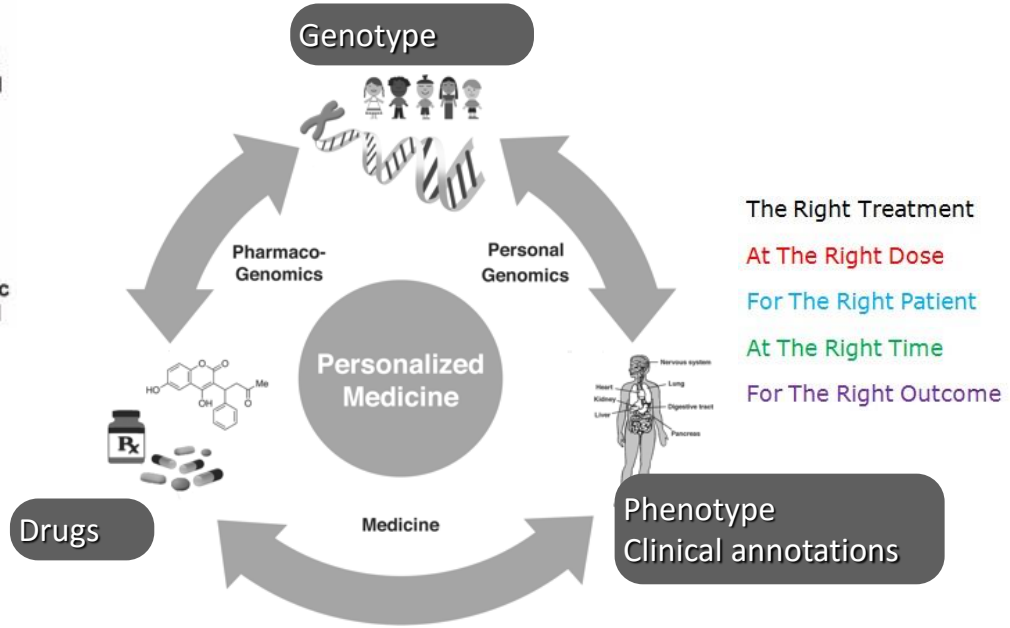
Website Stats. Jan–Sep 2011	
Visits	3,328
Unique visitors	996
Average time on site	9.5 minutes
Registered users	180
Visits with downloads	102

INSILICO
GENOMICS AT YOUR FINGERTIPS

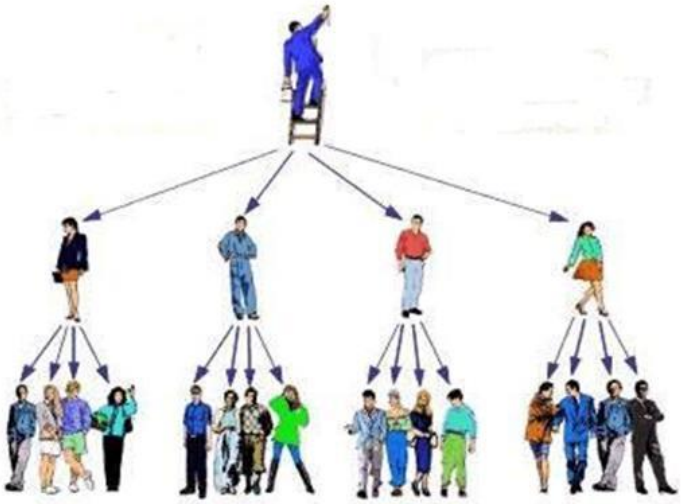

THE GENOMICS REVOLUTION IS DIGITAL
Data Generated
NGS CLINIC
DNA TECHNOLOGY
NGS RESEARCH
MICROARRAY
DATA TSUNAMI
SANGER
1995
2003
2009
2011
2012
2014
$3 Billion
First human sequenced
$100K
S. Jobs
cancer sequenced
1 Life saved
First successful clinical application
Largest sequencing center opens in China
$15 Million
Google Ventures invests in DNAnexus
$3 Billion
Roche hostile bid for takeover of Illumina
$1000 Genome
Adoption
of sequencing machines in hospitals

Personalized Medicine



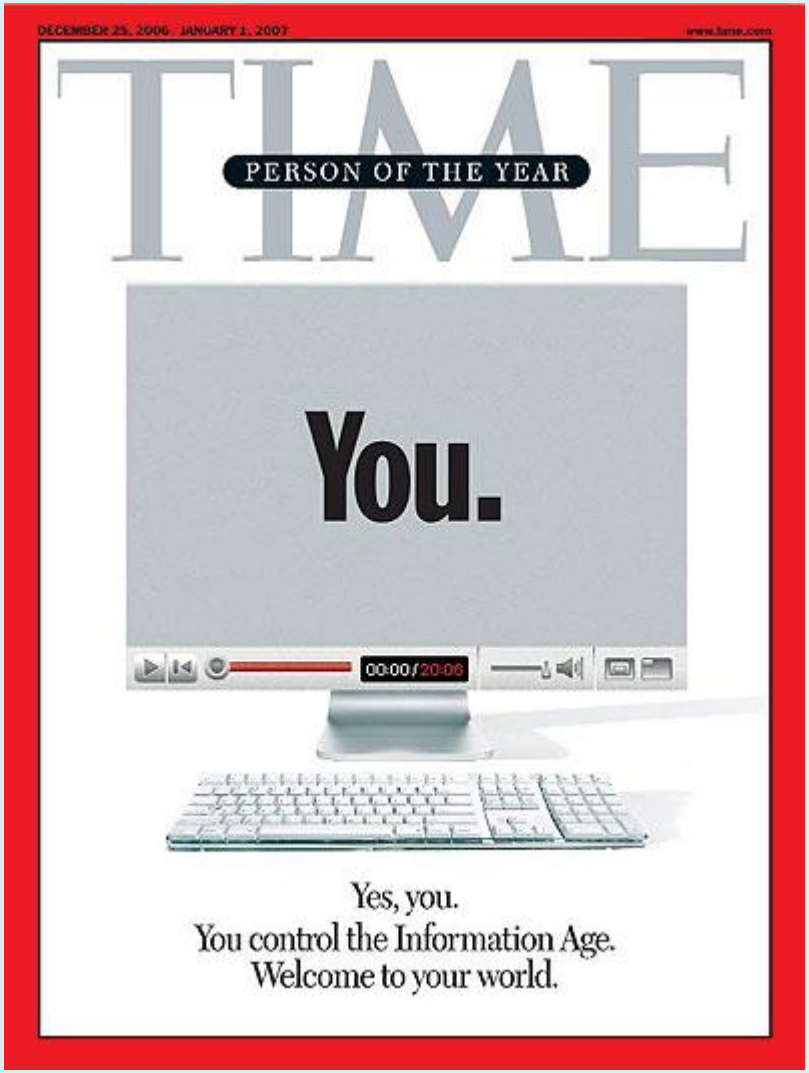


- Identifying robust drugs–genotype–phenotype relationships is a key challenge in the process of making this vision a reality.
- This requires very large sample sets for discovery and validation.

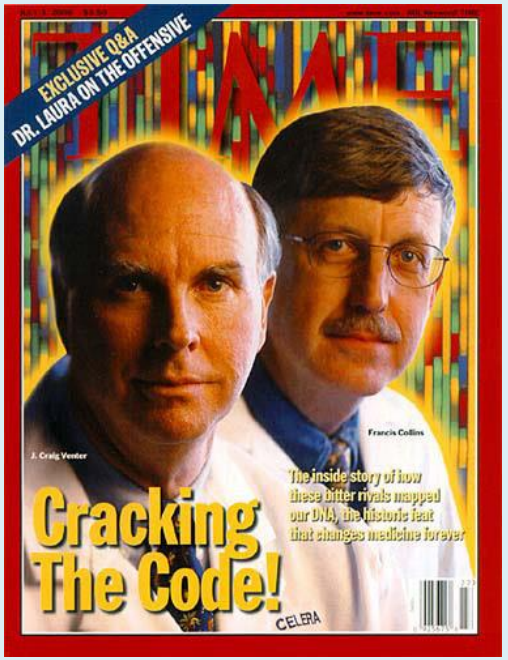


Profilage

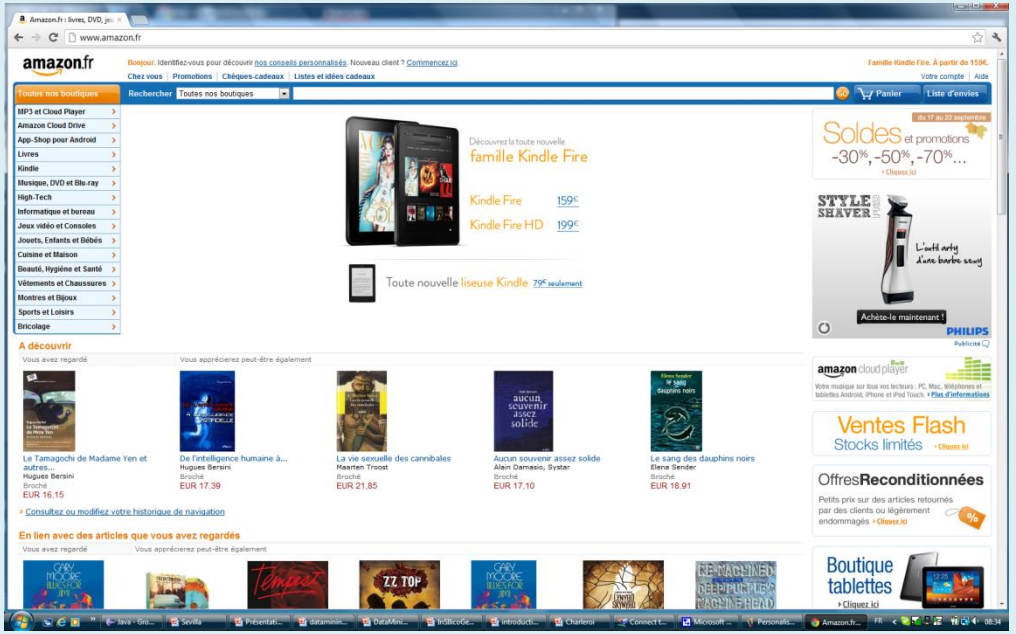
Qui se ressemble s’assemble



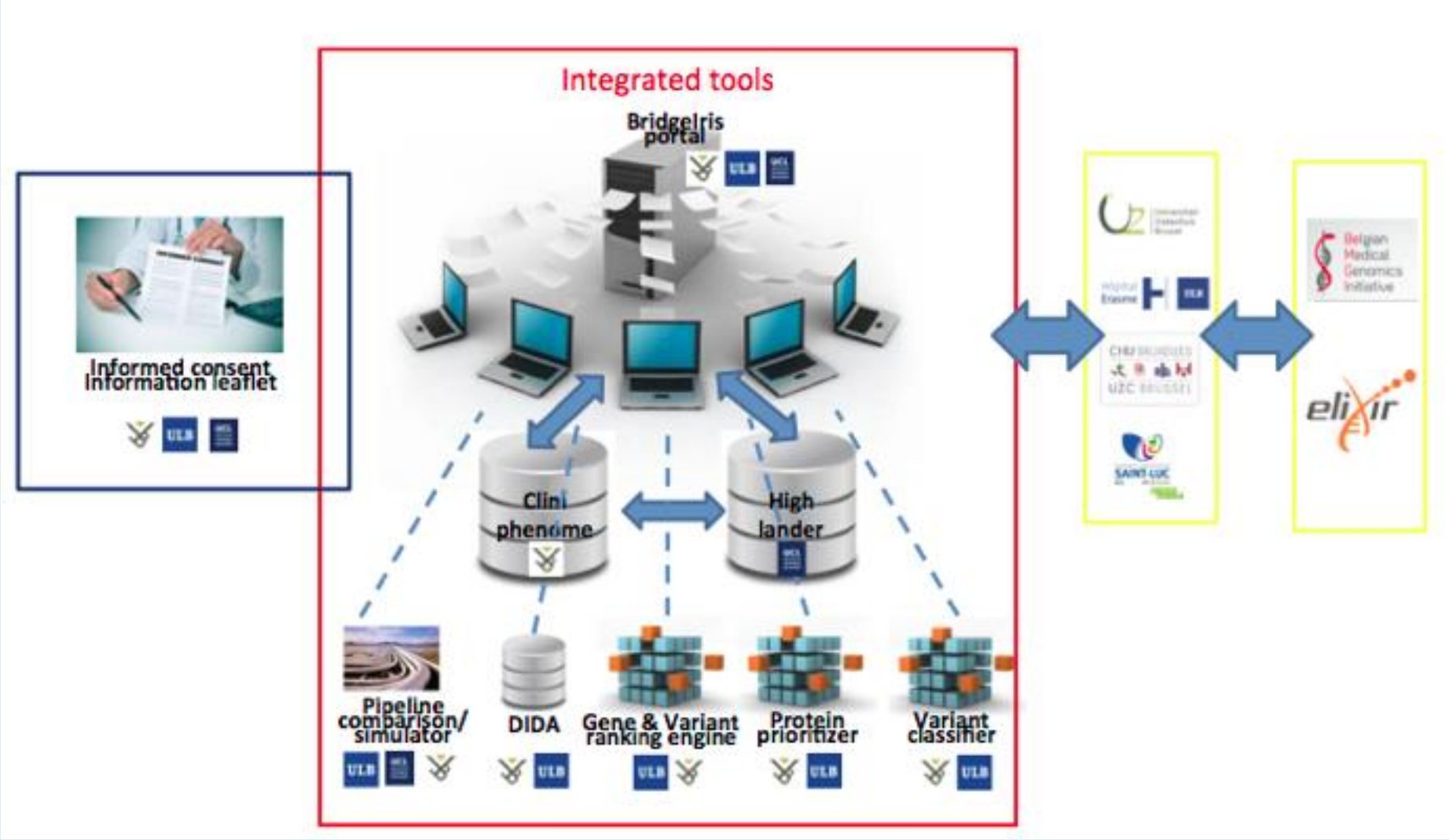
23AndME

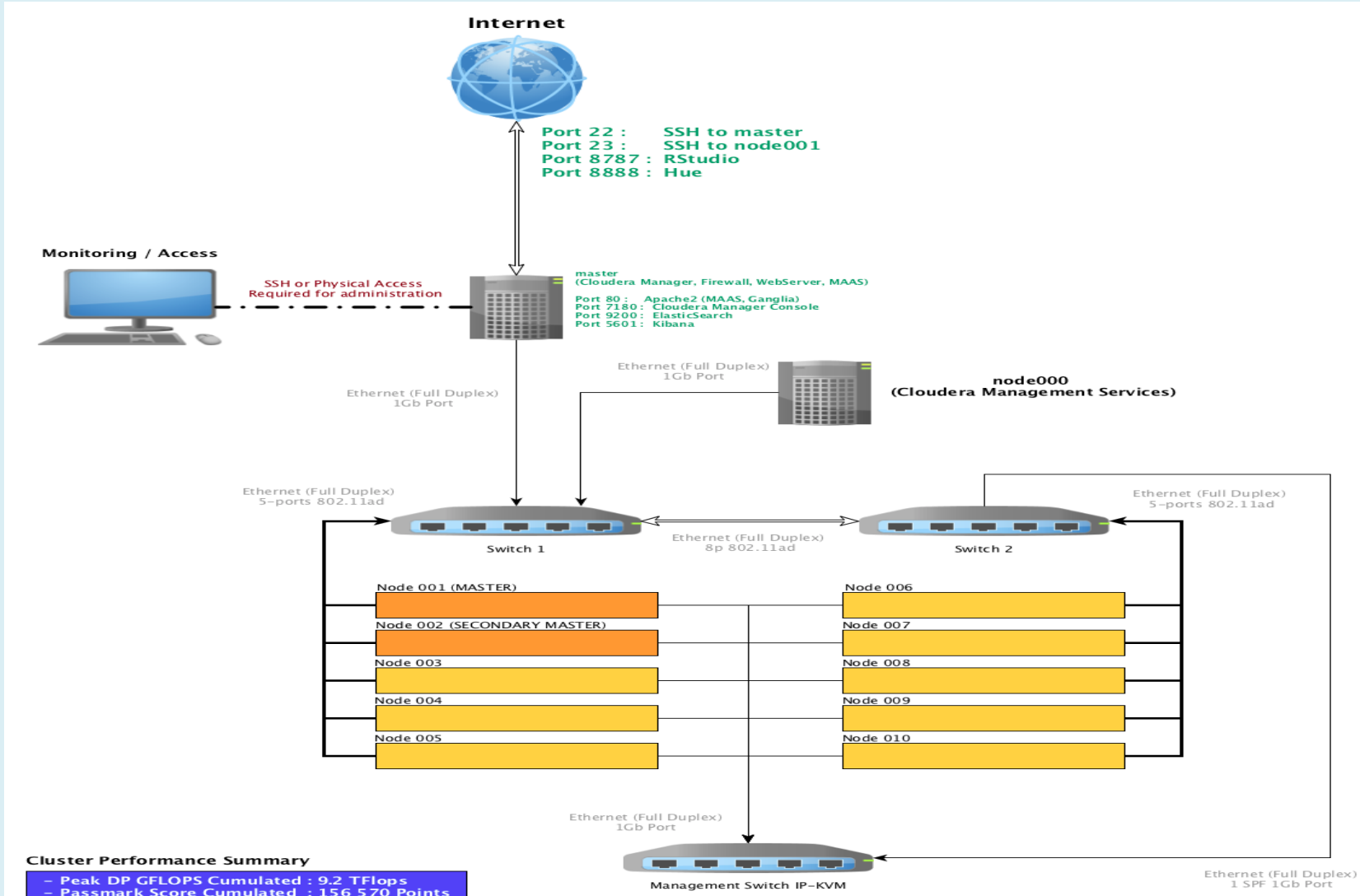


Amazon



Dernier en date: BridgeIris





Cluster Performance Summary

- Peak DP GFLOPS Cumulated : 9.2 TFlops
- Passmark Score Cumulated : 156 570 Points
- Storage HDD JBOD available : 80 Tb
- Memory RAM available : 1.28 Tb

Node xxx Configuration

- Dual Xeon e5-2620v3 2.4 Ghz
- 128 Gb RAM DDR4 2133Mhz ECC Reg Quad Rank
- 2x 4 Tb WD RED 7.2k 64Mo SATA3
- n-Ethernet Gigabit Port Full Duplex 802.11ad Link Agregation

Node xxx Upgradability

- Dual Xeon Skt 2011 v3 Family up to 18 cores HT 145W TDP
- Up to 256-512 Gb RAM DDR4 RDIMM, or 1024 LRDIMM 2133Mhz ECC Reg QR
- Up to 6 additional 3.5" HDD or 8 via 2.5" HDD
- Network upgrades include : Multiple 10GbE or Infiniband via PCIe
- Co-Processor : 4x PCIe 3.0 16x + 2x PCIe 3.0 8x Low Profile Available

Statistiques et Data Mining

1. Les data scientists font tout le temps des stats “fréquentielles”, souvent de manière maladroite et inconsciente pour trouver le meilleur modèle: comptage et métrique.
2. Faut-il évaluer les modèles de manière plus “scientifique”: p-value, student-test, …. ?
3. L’opérationalité, le déterminisme du monde et l’absence de risque (envoi de mails, sms …) le rendent rarement necessaire … mais en médecine ??
4. Le BigData renforce le bricolage opérationnel et diminue le recours à l’évaluation scientifique des résultats.

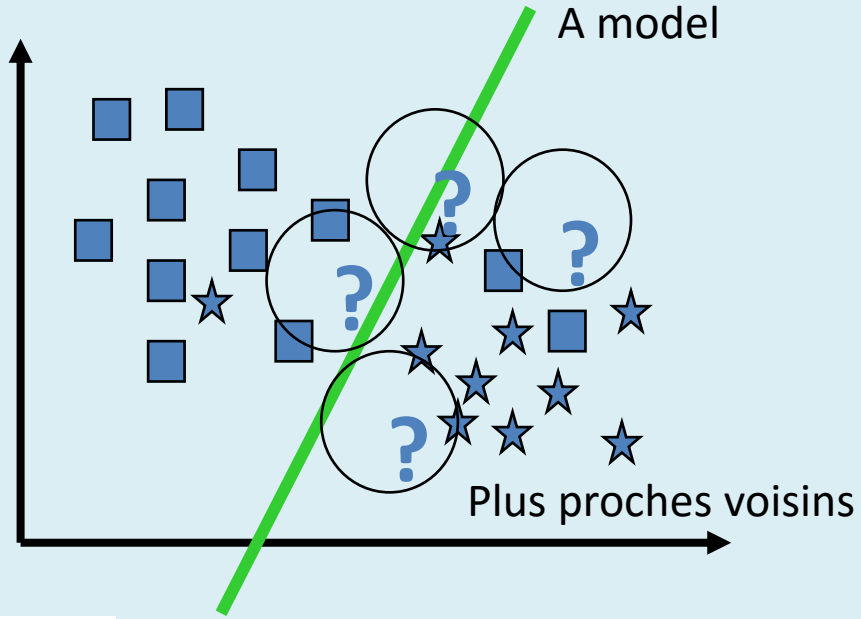
Statistiques « inconscientes » pour la classification: Trois exemples

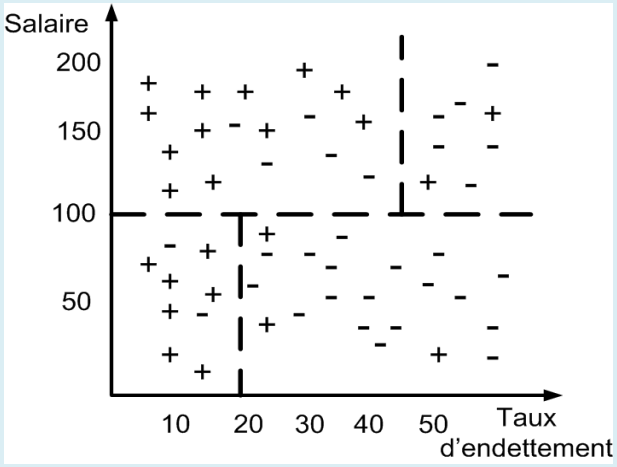
$$P(spam|penis, viagra)$$

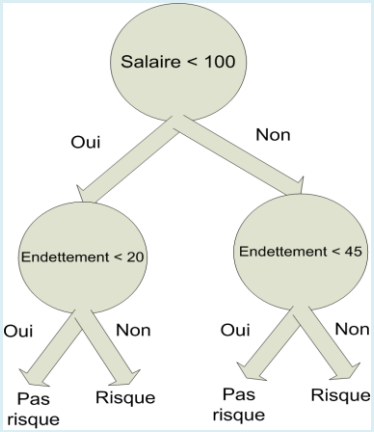
$$= \frac{P(penis|spam) * P(viagra|spam) * P(spam)}{P(penis) * P(viagra)}$$

$$= \frac{\frac{24}{30} * \frac{20}{30} * \frac{30}{74}}{\frac{25}{74} * \frac{51}{74}} = 0.928$$

Naïve Bayes





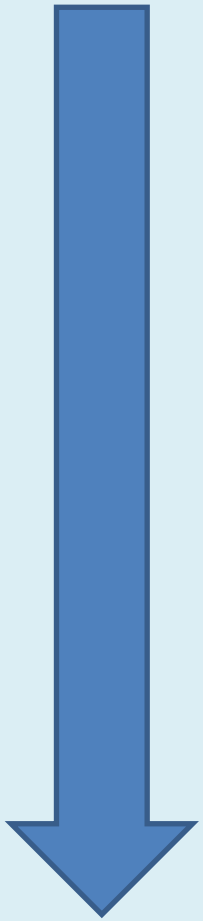


Arbres de décision

$$Gain(S, District) = Entropy(S) - (\frac{5}{14} Entropy(S_{District=Surburban})) + (\frac{5}{14}) Entropy(S_{District=Urban})) + (\frac{4}{14} Entropy(S_{District=Rural})) = 0.246 bits$$

Classification incertaine
Diagnostic médical

Erreurs conséquentes



Stats plus formelles

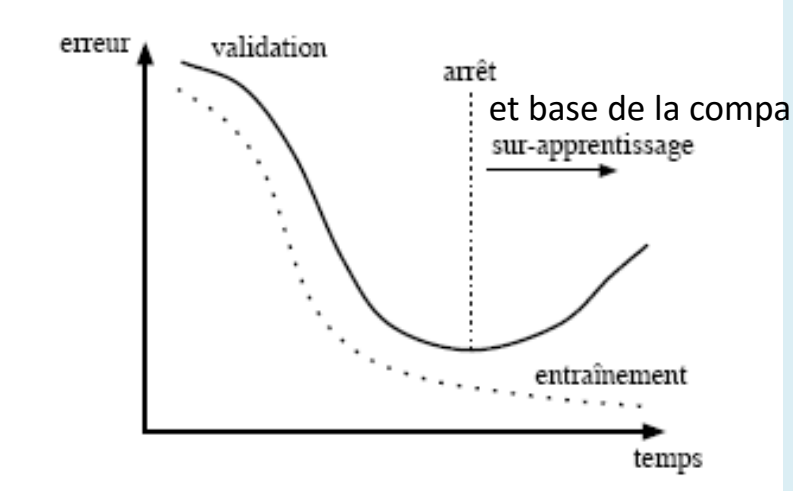
Classification certaine
Reconnaissance
de caractère, de scène

Erreurs inconséquentes

Approches plus informelles

Evaluation de modèles plus “formelle”

- Validation croisée



- Comparaison de modèles: t-test.
- Validation de modèle et intervalle de confiance : Pr[-z < val/N < +z] = c
- Conforte l’idée de l’élargissement de l’ensemble de données

Conclusions

- Stats inconscientes (le minimum qu'il faut savoir) dans l'elaboration des modèles.
- Data Mining et BigData, de plus en plus, privilégient les qualités d'informaticien sur celles de statisticiens.
- Le besoin de rigueur statistique s'accroit avec l'incertitude inhérente au monde et les consequences des erreurs.
- Mais la culture statistique est toujours un grand plus!