

# Data mining et statistiques

Hugues Bersini

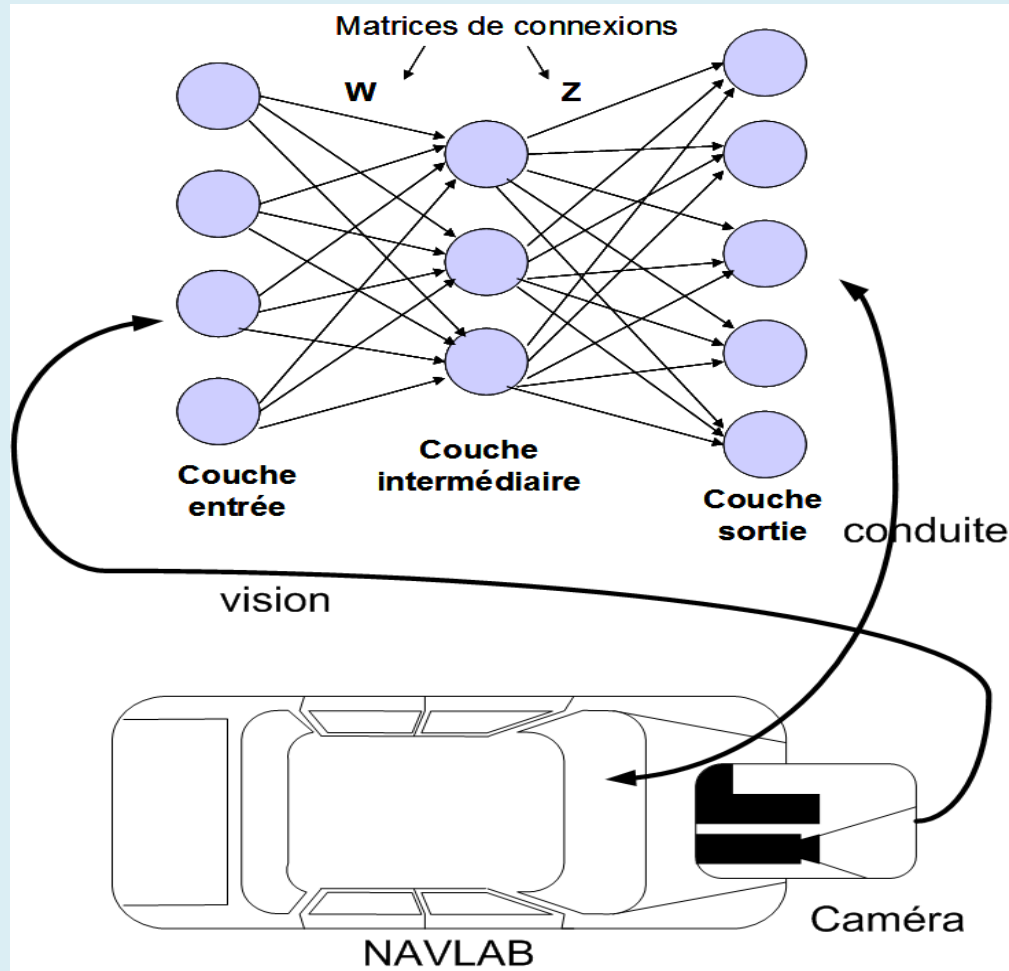
IRIDIA / ULB

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



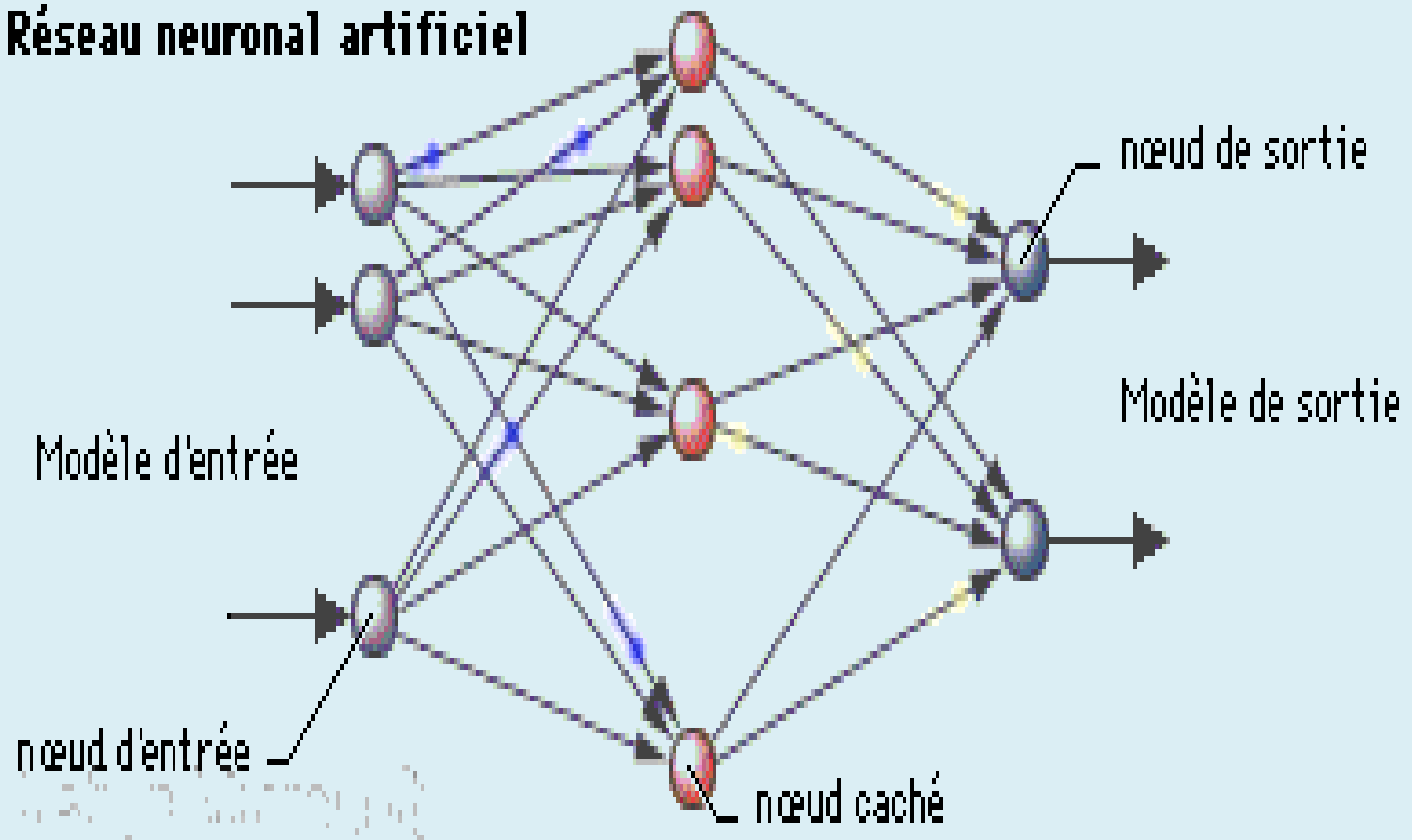
Théorie des croyances vs Probabilités

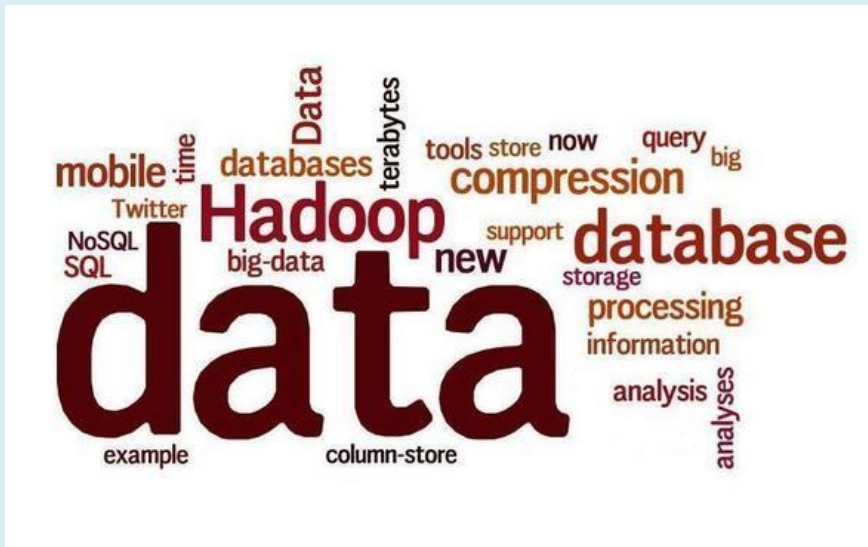
# IA, IRIDIA et Data Mining



# Les réseaux de neurones

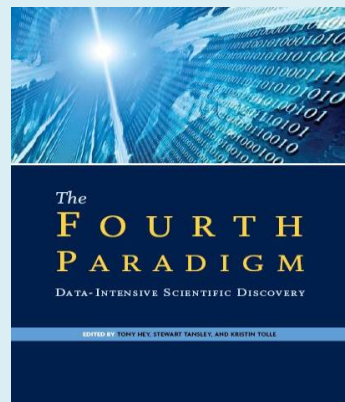
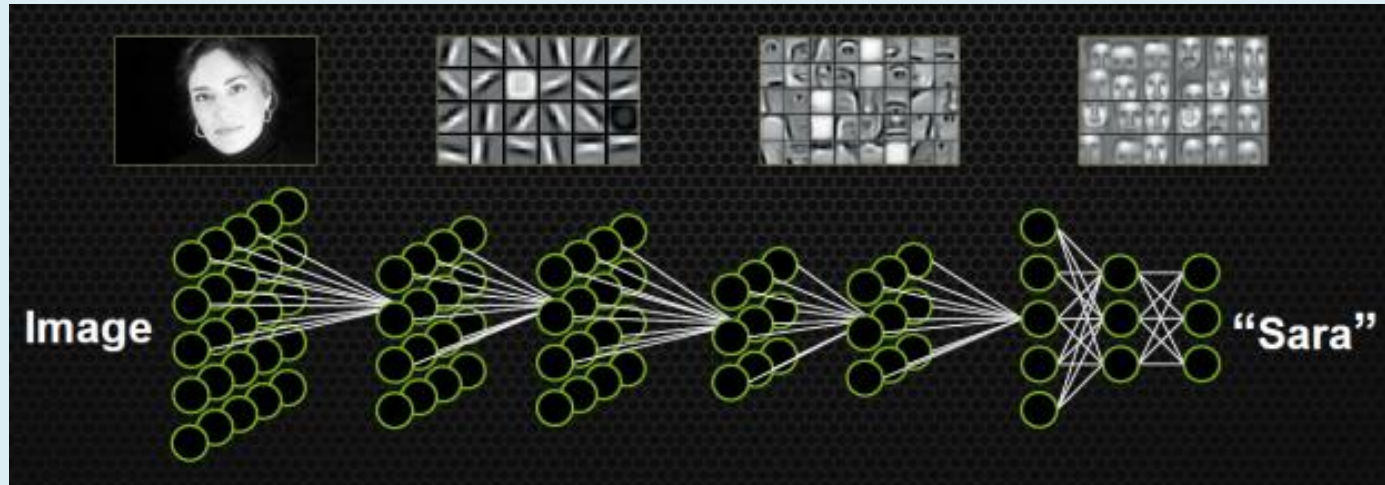
## Réseau neuronal artificiel



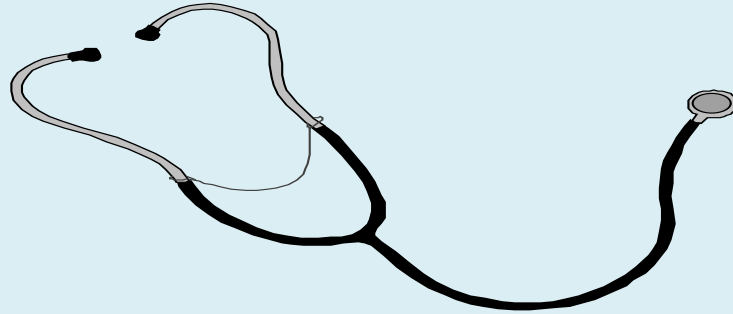


IA à la Google

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



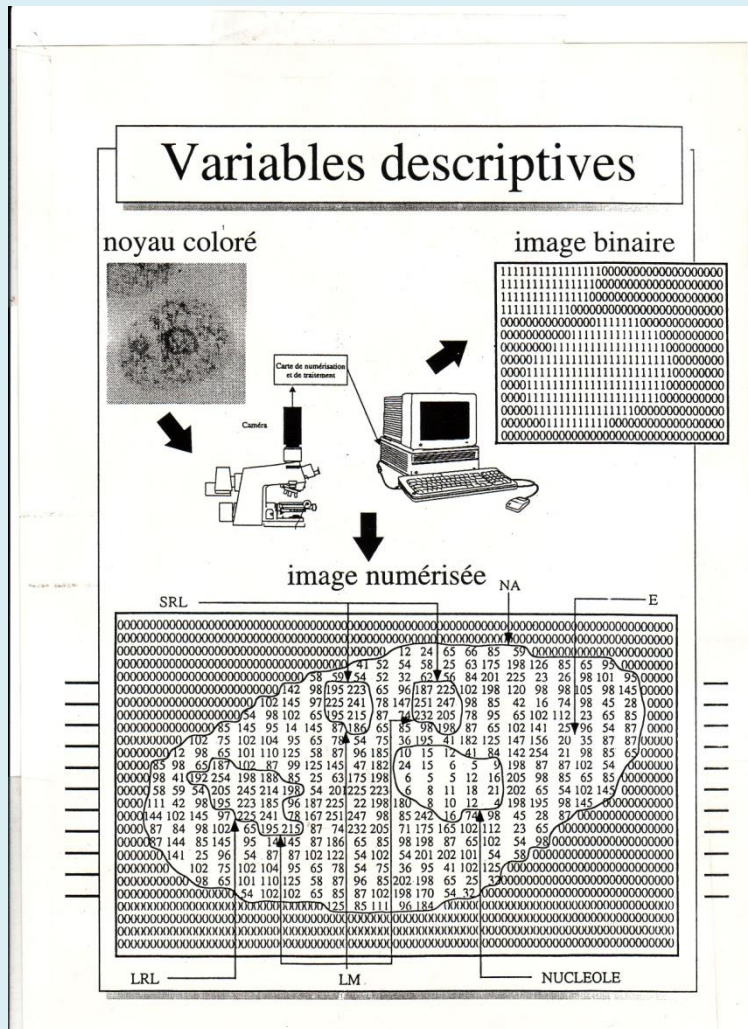
## IRIDIA dans le domaine médical



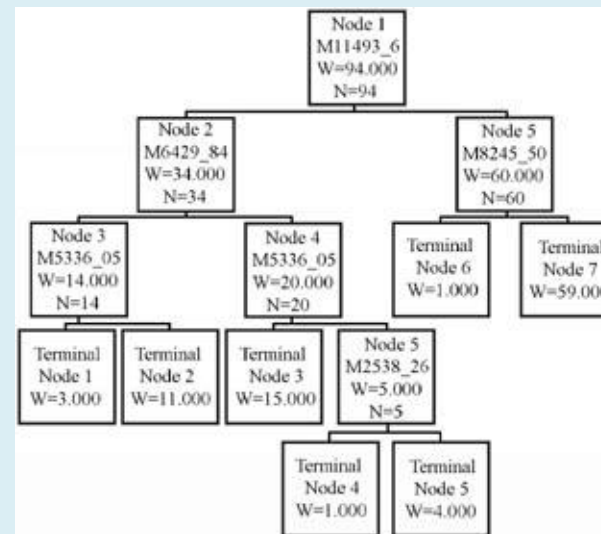
- Diagnostique automatique de cancer: Decaestecker, Van Ham, Kiss
- Détection de problèmes respiratoires: Mathys, Degroot, Kahn
- Analyse d'électrocardiogrammes: Bableyantz
- Aide à la marche assistée de paraplégiques: Preumont
- Aide à l'empowerment des diabétiques
- Classification des cancers par puce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



# Gradation automatique des cancers



Christine Decaestecker



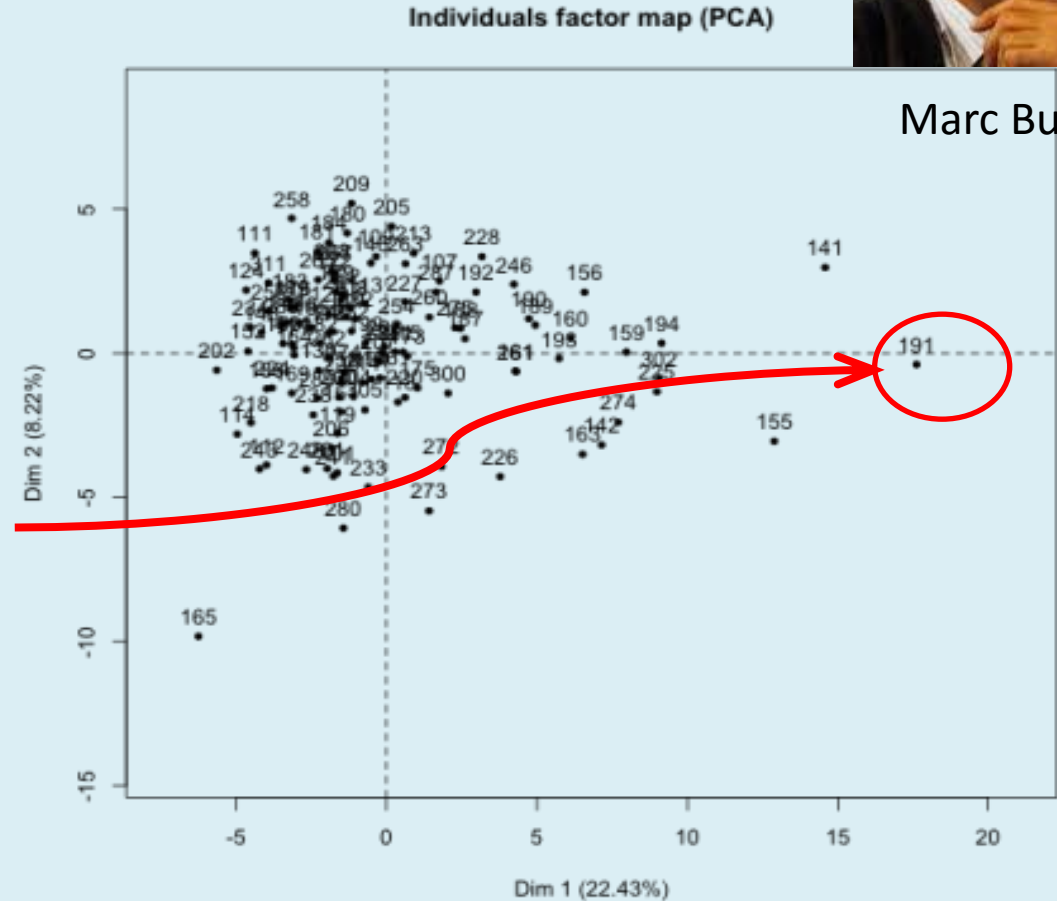


## Cluepoints : detection of outlier clinical site



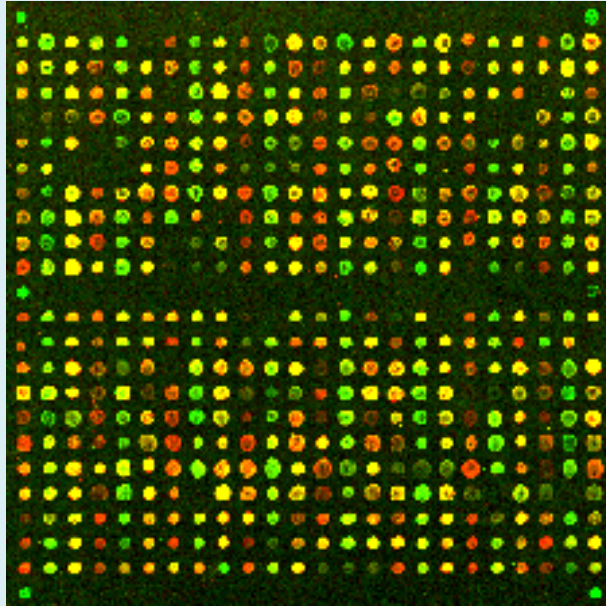
Marc Buyse

- 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



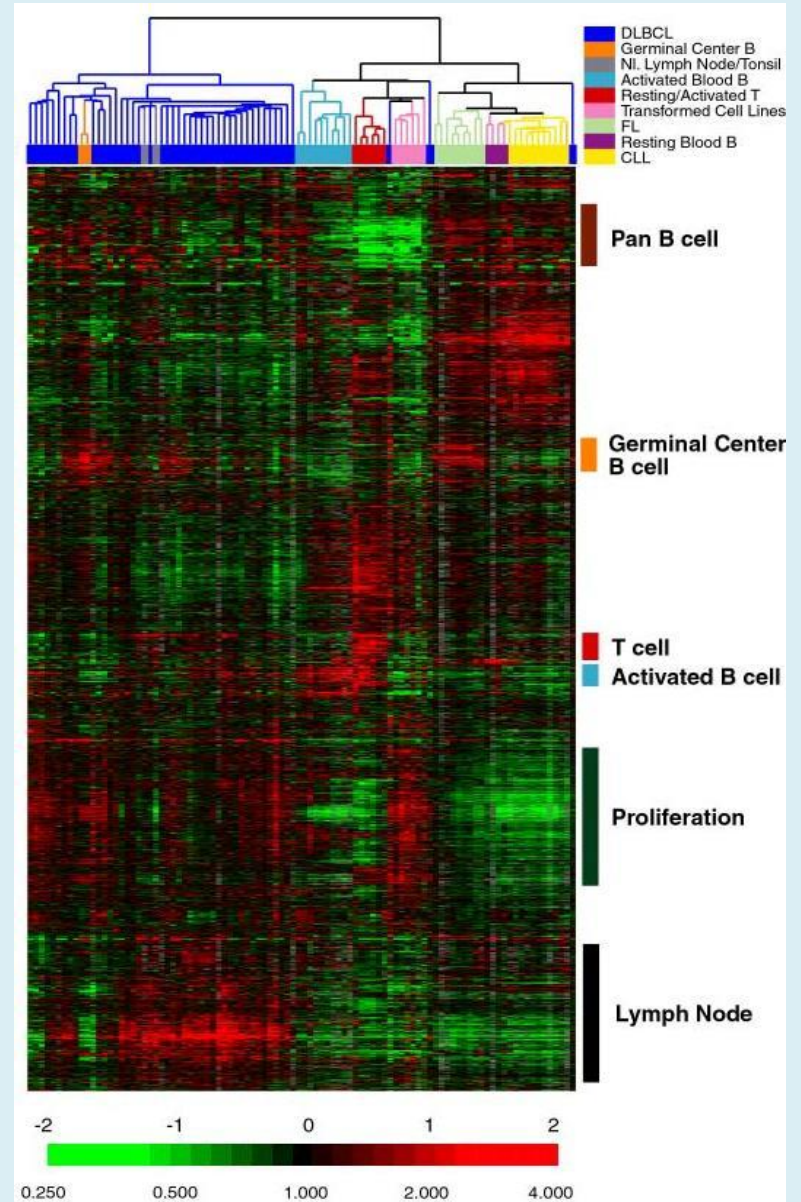
Microarray chip



Vincent Detours



Gianluca Bontempi



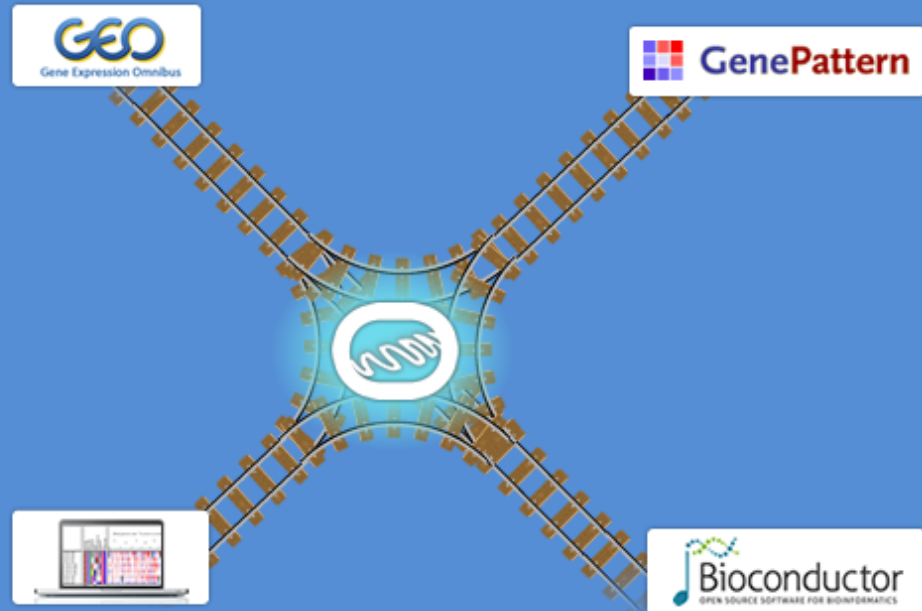
# LA SPINOFF INSILICO

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

David Weiss et Alain Coletta



**YOUR GENOMICS DATA HUB**  
AN EFFICIENT STARTING POINT  
FOR GENOMICS ANALYSIS



InSilico DB = Data + Tools

# Website for genomics datasets management and query

The screenshot displays the INSILICO website interface for browsing genomic datasets. The main search results table is as follows:

Title	#Samples	Public
<a href="#">Anti-tumor Activity of Histone Deacetylase Inhibitors in Non-Small Cell Lung Cancer Cells</a>	23	<input checked="" type="checkbox"/>
<a href="#">Gene expression differences between adenocarcinoma and squamous cell carcinoma in human NSCLC</a>	58	<input checked="" type="checkbox"/>
<a href="#">MERLION LUNG CANCER STUDY</a>	72	<input checked="" type="checkbox"/>
<a href="#">Gene expression profile of lung tumors</a>	19	<input checked="" type="checkbox"/>
<a href="#">H522 lung cancer cells' TGFbeta response was restored by putting back BRG1 and TGFbRII</a>	6	<input checked="" type="checkbox"/>
	30	<input checked="" type="checkbox"/>
	13	<input checked="" type="checkbox"/>
<a href="#">Metastatic cancers</a>	187	<input checked="" type="checkbox"/>
<a href="#">Lung cancer in smokers with airway dysplasia</a>	30	<input checked="" type="checkbox"/>
<a href="#">Lung cancer in smokers with airway dysplasia</a>	4	<input checked="" type="checkbox"/>
<a href="#">Lung Cancer Cells</a>	24	<input checked="" type="checkbox"/>
	29	<input checked="" type="checkbox"/>
	36	<input checked="" type="checkbox"/>

The search interface (inset) shows the INSILICO logo with the tagline "GENOMICS AT YOUR FINGERTIPS" and a search bar containing "lung cancer". It also features a "GenePattern" logo and the text "Access genomic datasets in SECONDS".

# Integration with visualisation and analysis tools

**GenePattern | Job Results**  
<http://genepattern.broadinstitute.org/gp/pages/jobResults.jsf>

**GenePattern**  
 Modules & Pipelines Suites Job Results

**Job Results** show: My job results

Status	Job	delete	Module Name
✓	229454	<input type="checkbox"/>	GetDatasetInSilico
✓	229455	<input type="checkbox"/>	1. DownloadURL
✓	229456	<input type="checkbox"/>	2. untar

**Curated biological samples information**

**Smoker**

	No	Yes	Description
1	Red	Yellow	angiogenin, ribonuclease, RNase A
2	Red	Yellow	melanoma antigen family D, 1
3	Red	Yellow	NAD(P)H dehydrogenase
4	Red	Yellow	CAP, adenylate cyclase
5	Red	Yellow	WAP four-disulfide core domain
6	Red	Yellow	histone cluster H1
7	Red	Yellow	homogentisate 1,2-dioxygenase
8	Red	Yellow	pirin (iron-binding)
9	Red	Yellow	family with sequence similarity to claudin 10
10	Red	Yellow	transcobalamin II
11	Red	Yellow	carbonyl reductase
12	Red	Yellow	abhydrolase domain containing 10
13	Red	Yellow	lin-related protein 1
14	Red	Yellow	-alpha (glycosylated)
15	Red	Yellow	ly with sequence similarity to claudin 10
16	Red	Yellow	ribonuclease H1

**Excel**

NUMBER	SUB-ARRAY	GENE	log ratio	mean (log)	S.D.	T-TEST	P-VALUE
1	1	pooled ma	44	0.13685	1.196416	-0.2964	
2	2	pooled ma	35	-0.02827	0.753543	0.0326	
3	3	salmon sp	1	na	na	na	na
4	4	luciferase	removed	removed	removed	removed	removed
5	5	salmon sp	removed	removed	removed	removed	removed

**IGV**  
 File View Tracks Help

**Integr. Gen. Viewer**

Whole genome view. To jump to...

DATA FILE	DATA TYPE	LINKING_ID	PARTICIPANT_ID	SAMPLE_ID	TUMOR TYPE
550001402...07011.A05					
550001402...07011.A05					
550001402...07011.A07					
550001402...07011.A07					
5500014026...607011.B05					
5500014026...607011.B05					
5500014026...607011.B07					
5500014026...607011.B07					
550001402...07011.C05					
550001402...07011.C05					
550001402...07011.C07					

**R/Bioconductor**

**Bioconductor**  
 OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



---

## Website Stats. Jan-Sep 2011

---

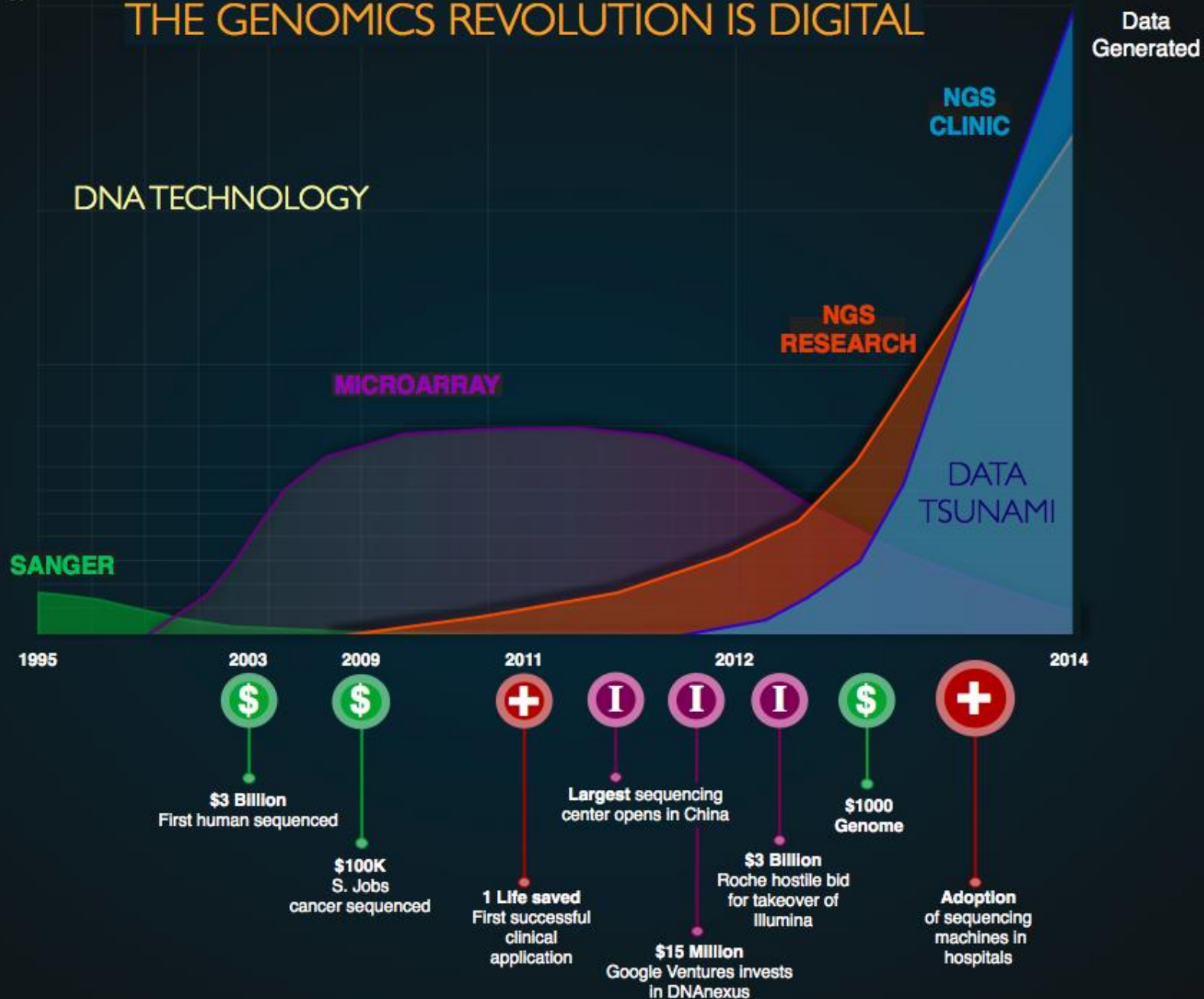
Visits	3,328
Unique visitors	996
Average time on site	9.5 minutes
Registered users	180
Visits with downloads	102

---





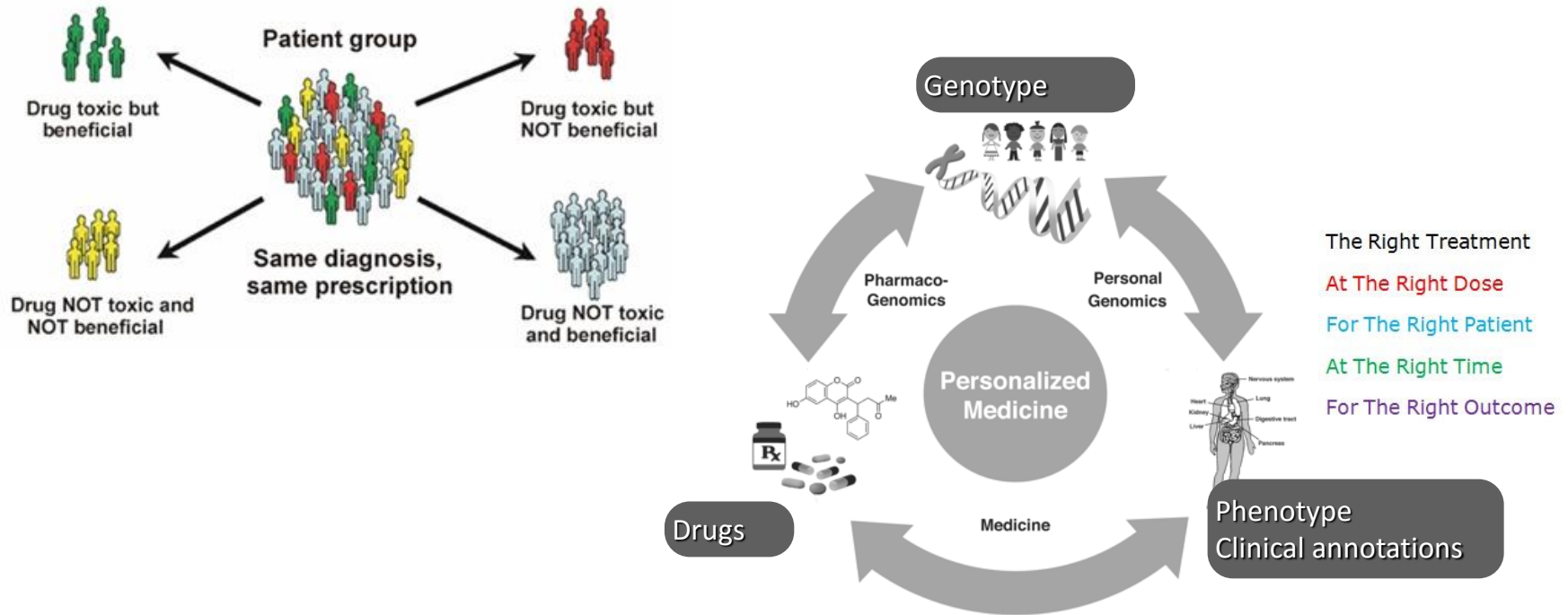
# THE GENOMICS REVOLUTION IS DIGITAL



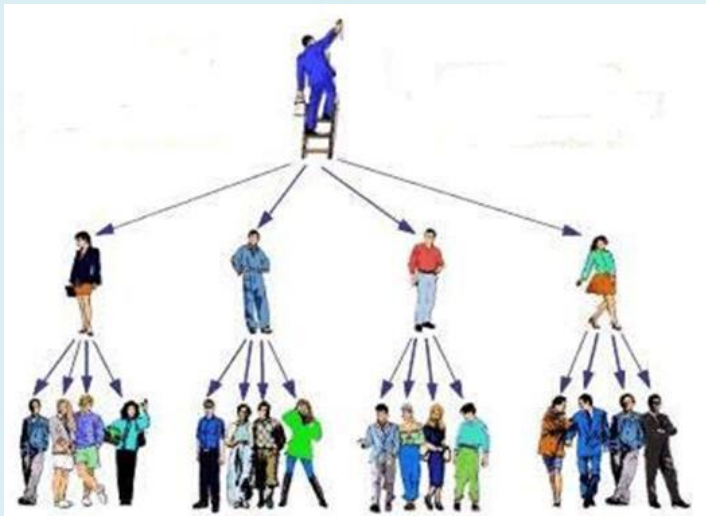
Exposé Bordet 14/11/2015



# Personalized Medicine



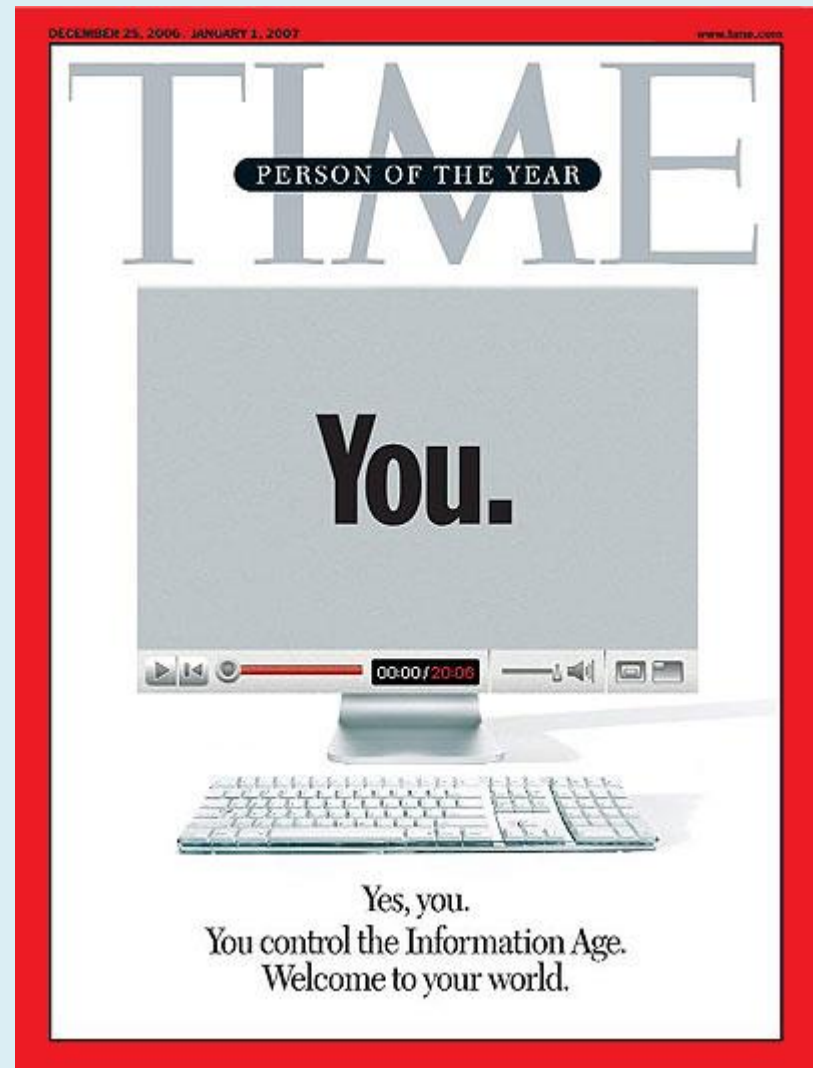
- Identifying robust drugs–genotype–phenotype relationships is a key challenge in the process of making this vision a reality.
- This requires very large sample sets for discovery and validation.



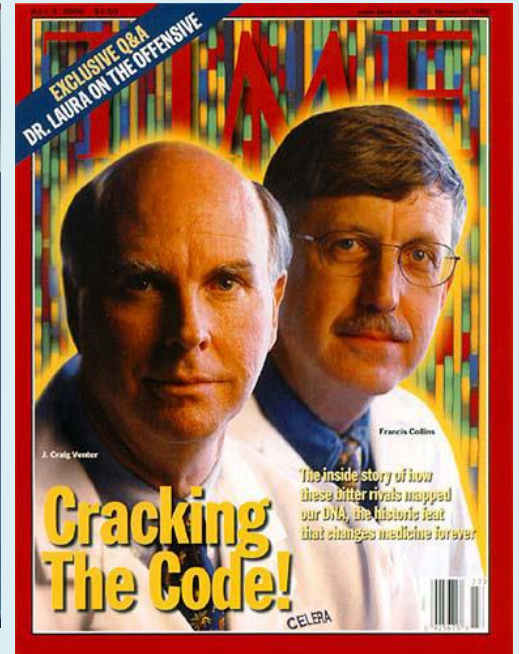
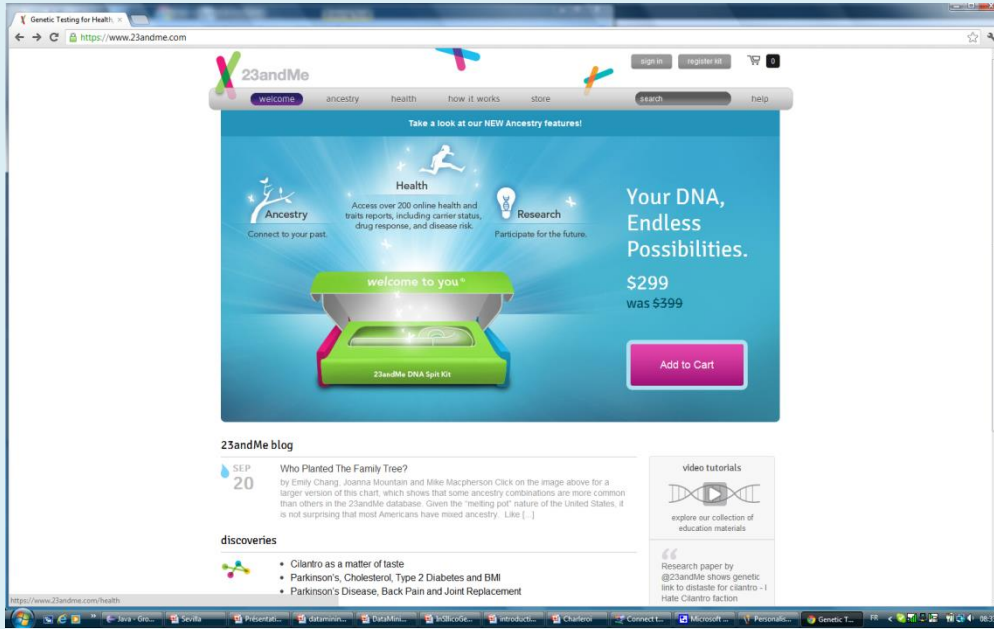
↑  
**Profilage**  
↑



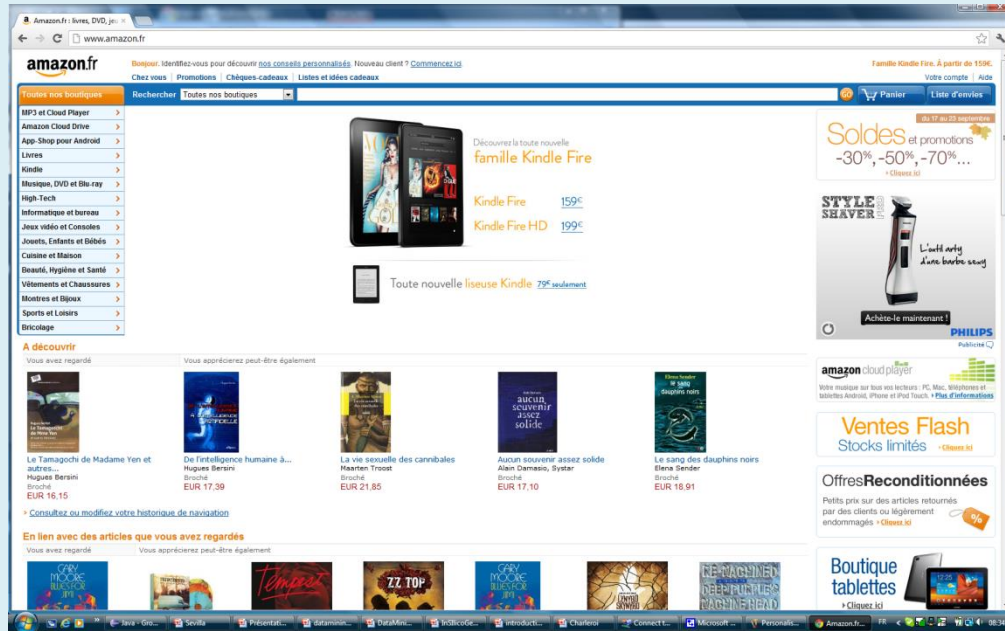
Qui se ressemble s'assemble



23AndMe

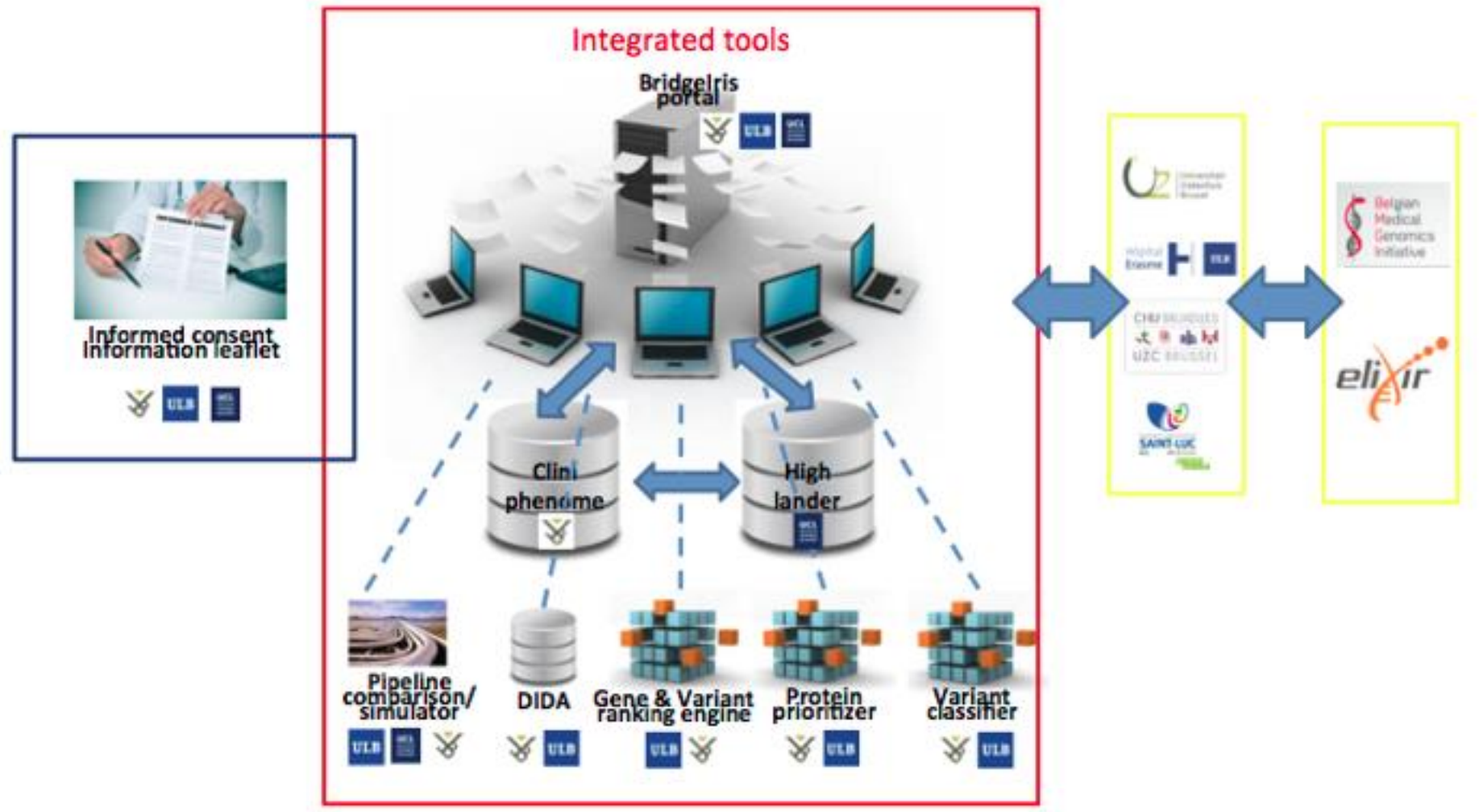


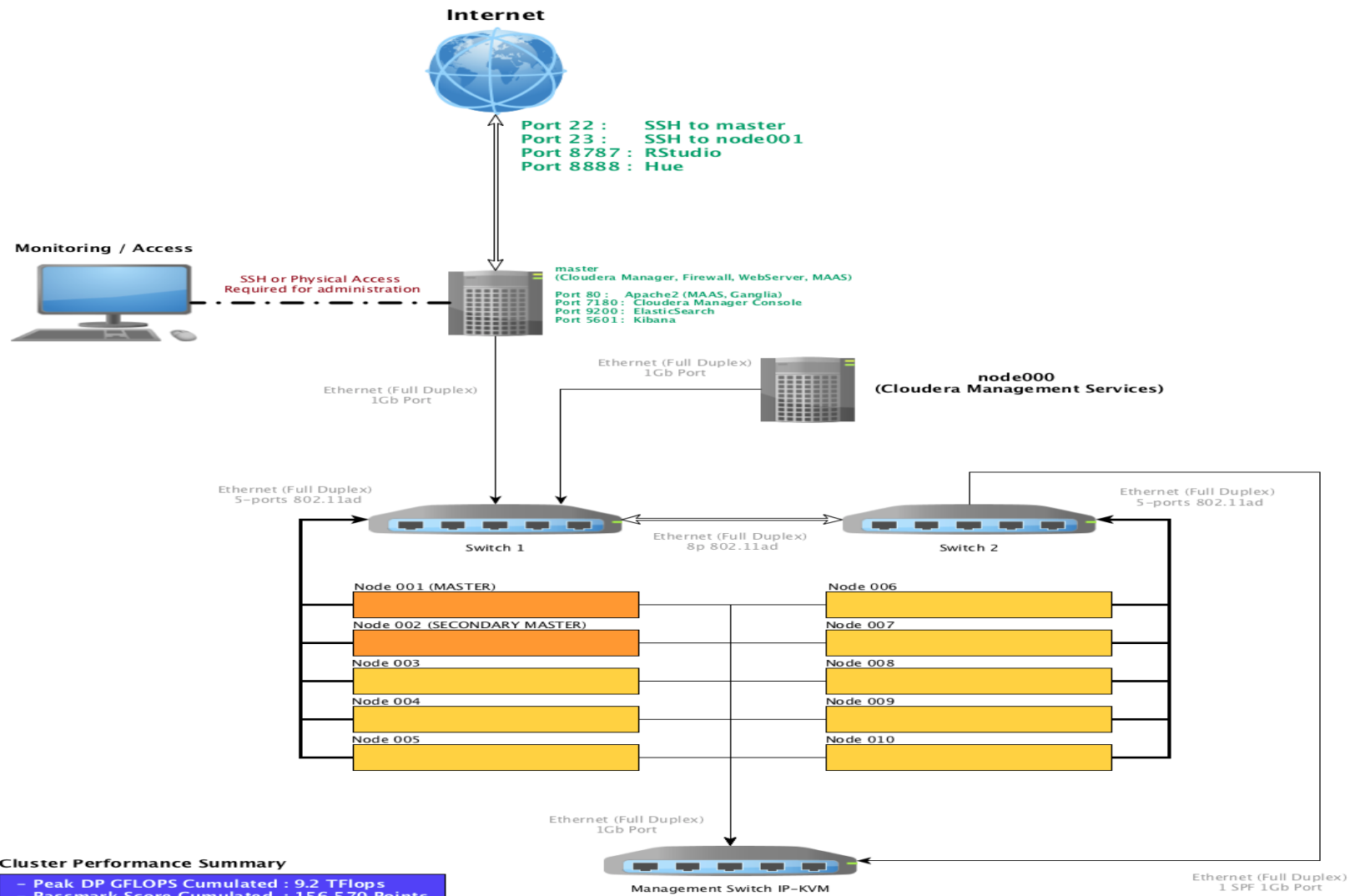
Amazon





# Dernier en date: Bridgelris





#### Cluster Performance Summary

- Peak DP GFLOPS Cumulated : 9.2 TFlops
- Passmark Score Cumulated : 156570 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’opérationnalité, le déterminisme du monde et l’absence de risque (envoi de mails, sms ...) le rendent rarement nécessaire ... mais en médecine ??
4. Le BigData renforce le bricolage opérationnel et diminue le recours à l’évaluation scientifique des résultats.



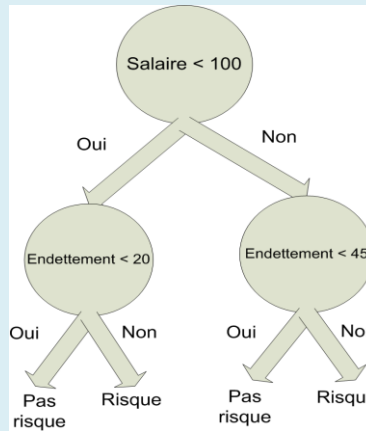
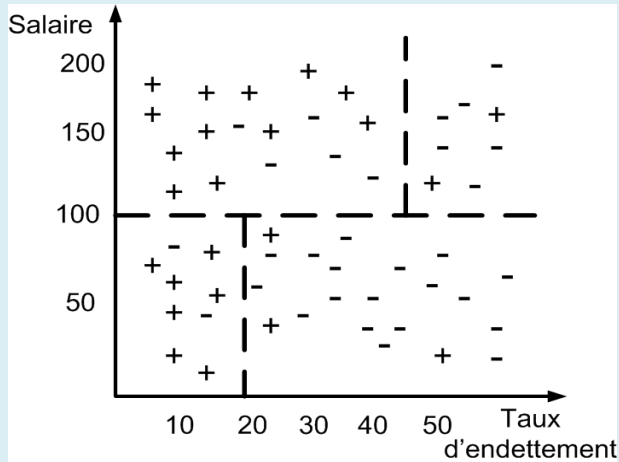
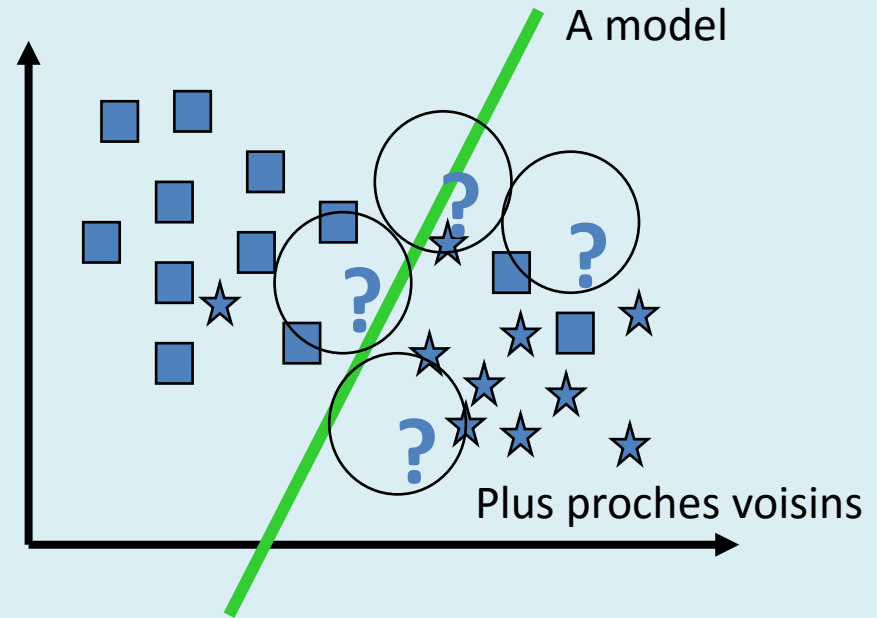
# Statistiques « inconscientes » pour la classification: Trois exemples

$$P(\textit{spam}|\textit{penis}, \textit{viagra})$$

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

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

Naïve Bayes



Arbres de décision

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

**Classification incertaine**  
Diagnostic médical

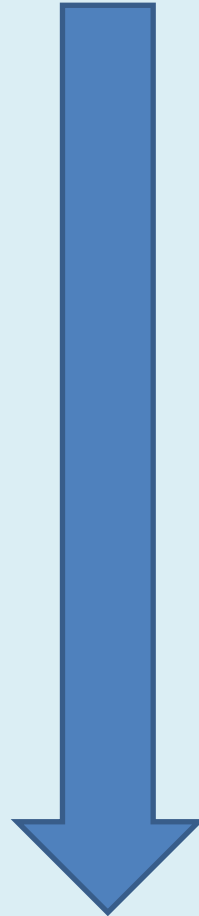
**Erreurs conséquentes**

Stats plus formelles

**Classification certaine**  
Reconnaissance  
de caractère, de scène

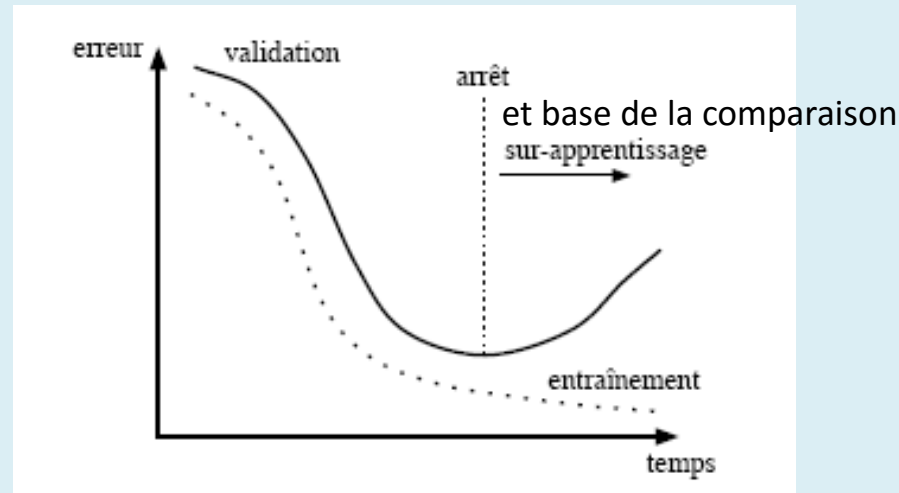
**Erreurs inconséquentes**

Approches plus informelles



# Evaluation de modèles plus “formelle”

- Validation croisée



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

# Conclusions

- Stats inconscientes (le minimum qu'il faut savoir) dans l'élaboration 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'accroît avec l'incertitude inhérente au monde et les conséquences des erreurs.
- Mais la culture statistique est toujours un grand plus!