Sciences Economiques & Sociales de la Santé
& Traitement de l'Information Médicale

www.sesstim-orspaca.org

Laura RICHERT

MD, PhD
INSERM U1219 - INRIA SISTIM - ISPED, Université de Bordeaux

Big data en épidémiologie

mai 2016

Cliquez ici pour voir l'intégralité des ressources associées à ce document

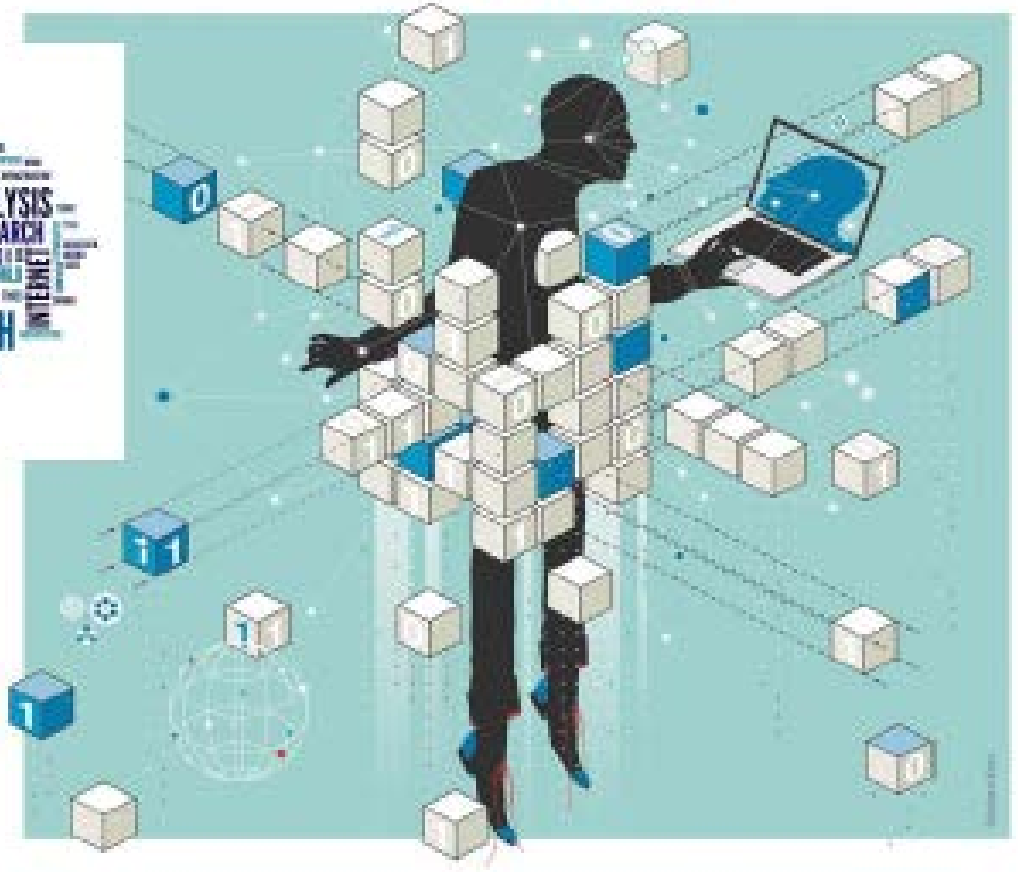
Webinar QuanTIM
20 mai 2016

Big data en épidémiologie

Laura Richert

INSERM U1219/INRIA SISTM
ISPED, Université de Bordeaux

BIGDATA



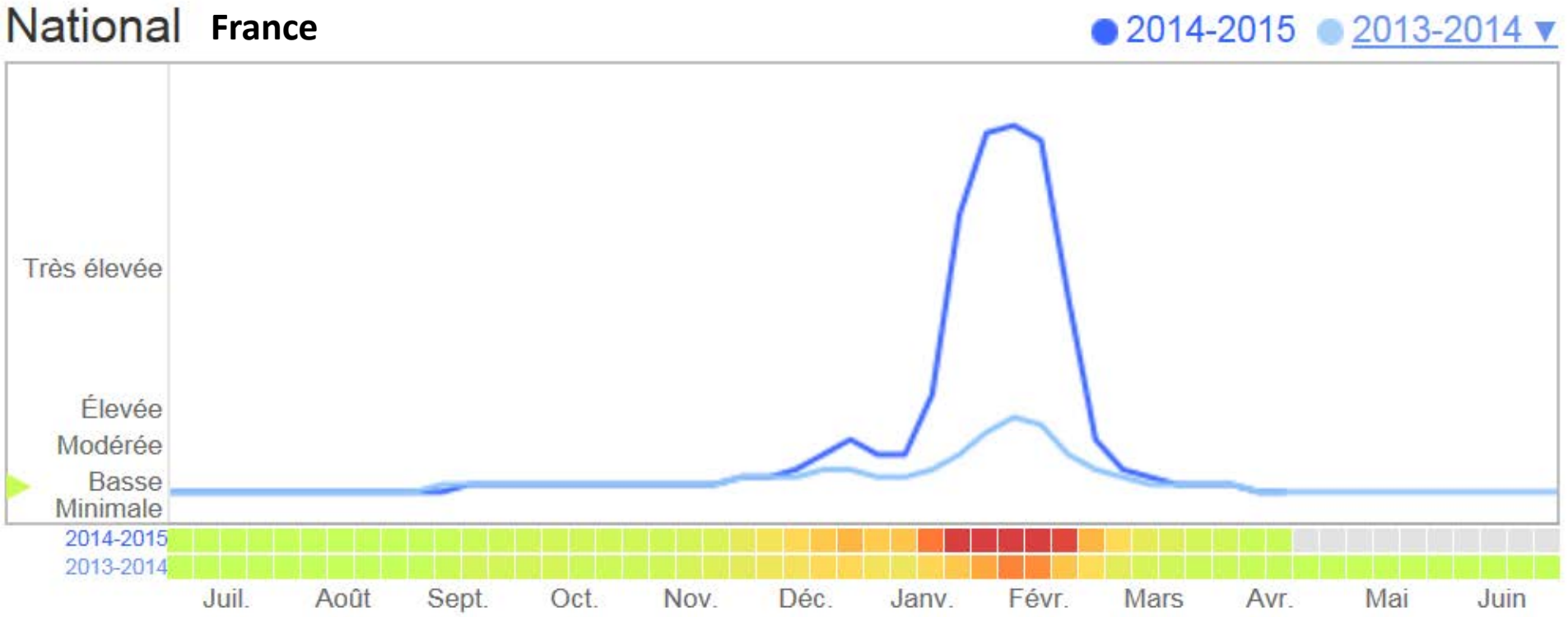
Le nouvel eldorado des données

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

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

Big Data

https://www.google.org/flutrends/fr/#FR



OPEN ACCESS Freely available online

PLoS one

Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic

Samantha Cook[1], Corrie Conrad[2]*, Ashley L. Fowlkes[3], Matthew H. Mohebbi[1]

Séquençage génétique

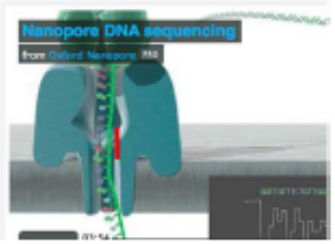
2005 — 2010 — 2012

13 Mb/hour
$10/Mb

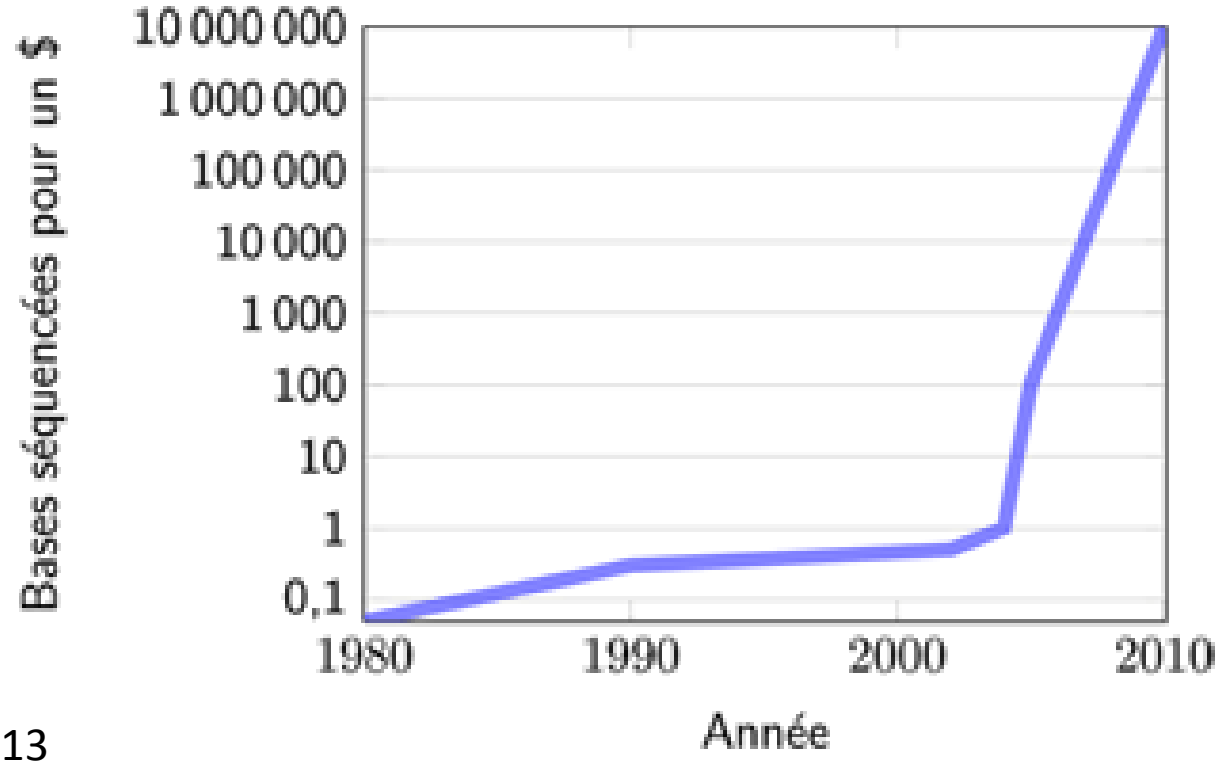
200 Mb/hour
$2/Mb

0,5 Gb/hour
$1/Mb

2,5 Gb/hour
$0.1/Mb



nanopore



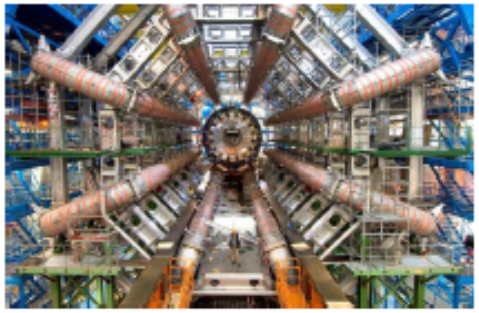
Alain Viari 04/2013

WHAT DO WE MEAN BY « BIG DATA »?

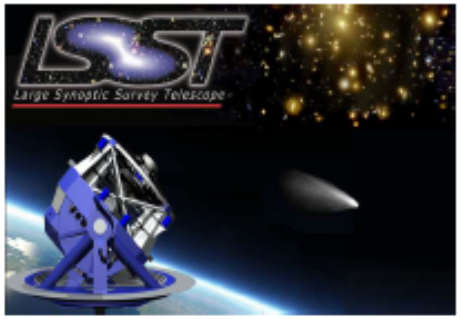
QUE VEUT-ON DIRE PAR « DONNEES MASSIVES »?

- **Edition 2016 du *Robert***
  - Big data **n. m.** (mots anglais « données massives») anglic. Le big data : l'ensemble des données générées par les nouvelles technologies, caractérisées par leur volume colossal. – recomm. offic. *mégadonnées* **n. f. pl.**

Un film de 1h30 (qualité dvd): 2.5 Go



LHC
~ 1 To/h (raw: 300 Go/h)
15 Po/year



LSST
~ 1 To/h (raw: 40 Tb/h)
100 Po/year

facebook

4 Po / year

YouTube™

1 h vidéo / sec
(1 année = 36 siècles de TV classique)

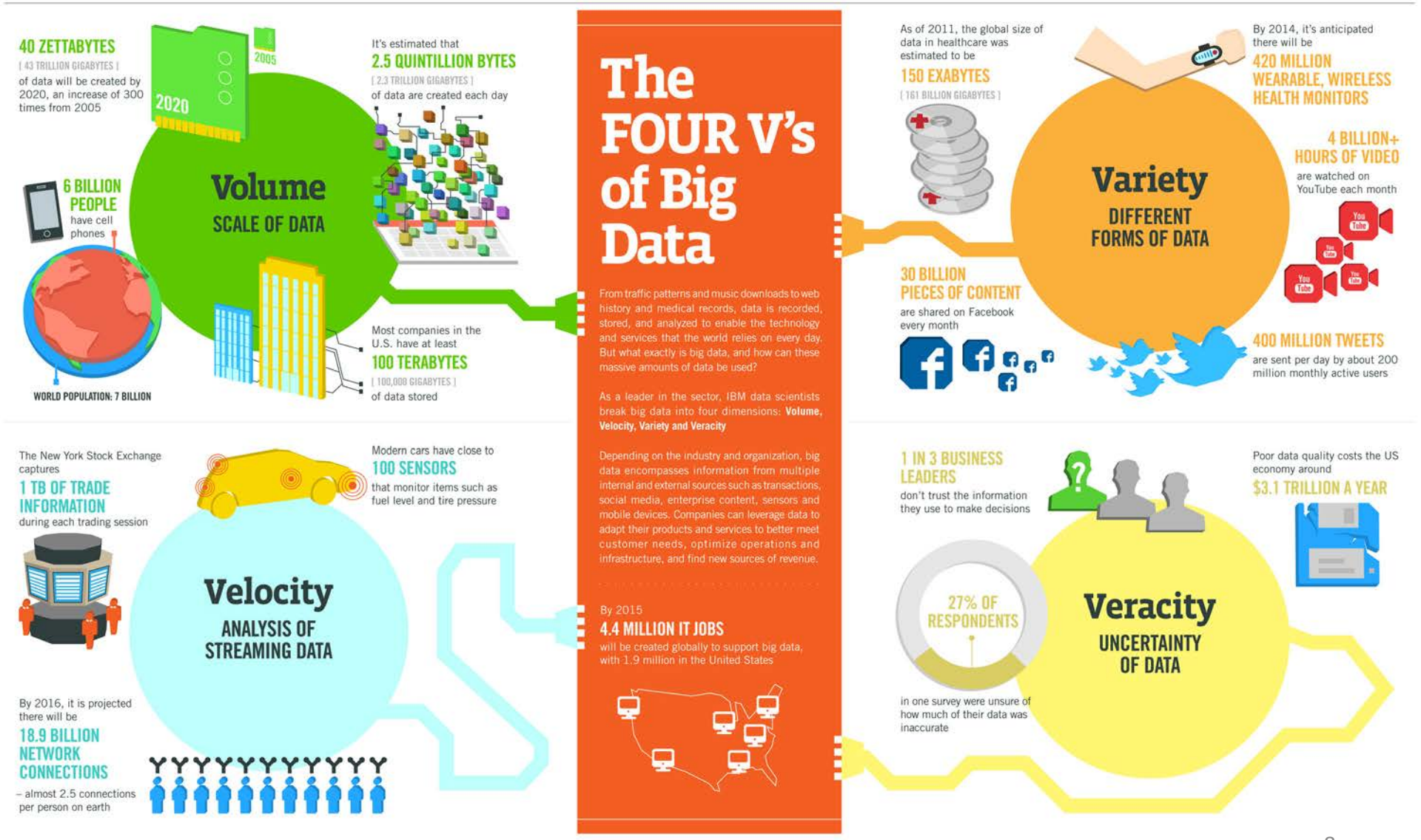
1 bit : 0/1
1 octet = 8 bits
1 Mo = 10^6 octets
1 Go = 10^9 octets
1 To = 10^{12} octets
1 Po = 10^{15} octets

600 Gb/run
~ 5 Go/h

BGI : 20 Po (sur 8 sites)

FranceGenomique: 5 Po
IFB: 1 Po

Les 4 dimensions des big data



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Et en épidémiologie?

From "Big Epidemiology" to "Colossal Epidemiology"

When All Eggs Are in One Basket

Miguel A. Hernán[a] *and David A. Savitz*[b]

Is Size the Next Big Thing in Epidemiology?

Sengwee Toh and Richard Platt

Epidemiology • Volume 24, Number 3, May 2013

Epidémiologie classique

Population

Modèle de régression

$$Logit\ P(Y=1) = X^T \beta$$

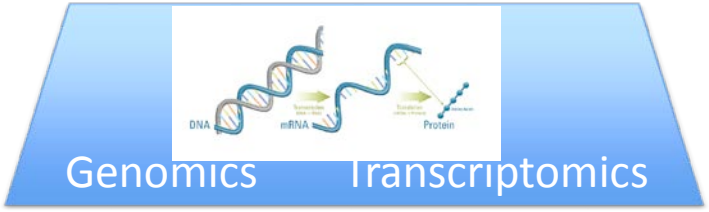
Genome Wise Association Study

Population

Modèle de régression

$$Logit\ P(Y=1) = X^T \beta$$

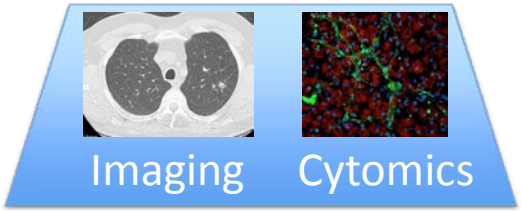
Gènes



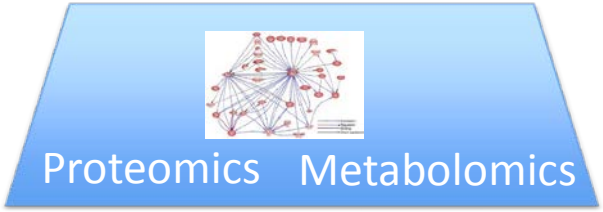
Epidémiologie systémique

Population

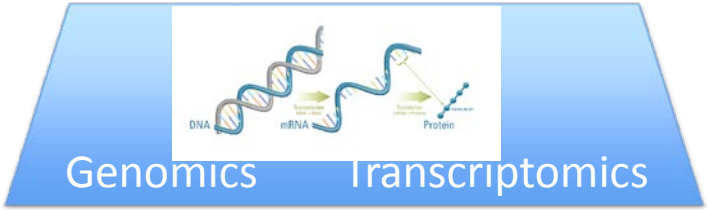
Tissus
Cellules



Molécules



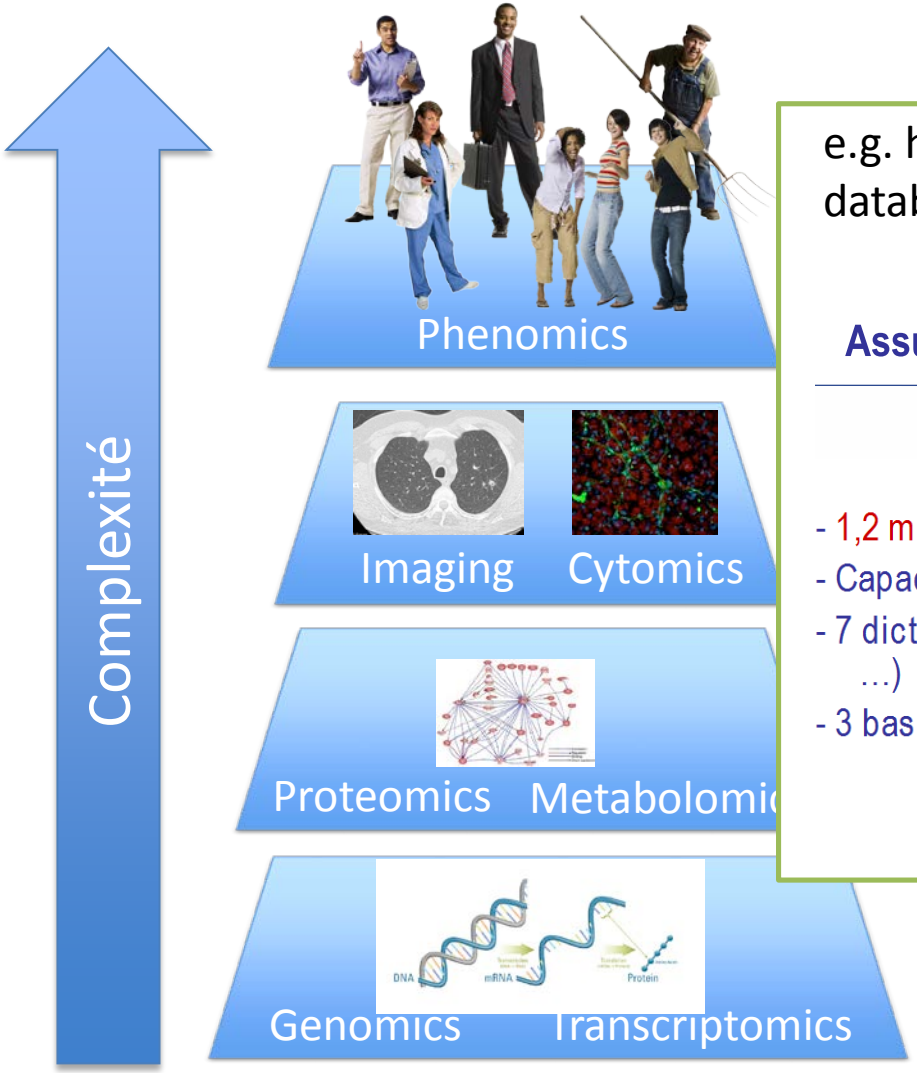
Gènes



Analyse integrative

Qui?
Comment?

Big data



e.g. health national insurance database for 57 M people

Assurance Maladie

- 1,2 milliards de feuilles de soins gérées chaque année
- Capacité de stockage égale à 450 Téraoctets
- 7 dictionnaires avec 785 000 objets (tables, index, synonymes …)
- 3 bases de données d'une volumétrie supérieure à 25 téraoctets

Taille

PB N°1: LA RÉPÉTITION DES TESTS (NOMBREUSES VARIABLES)

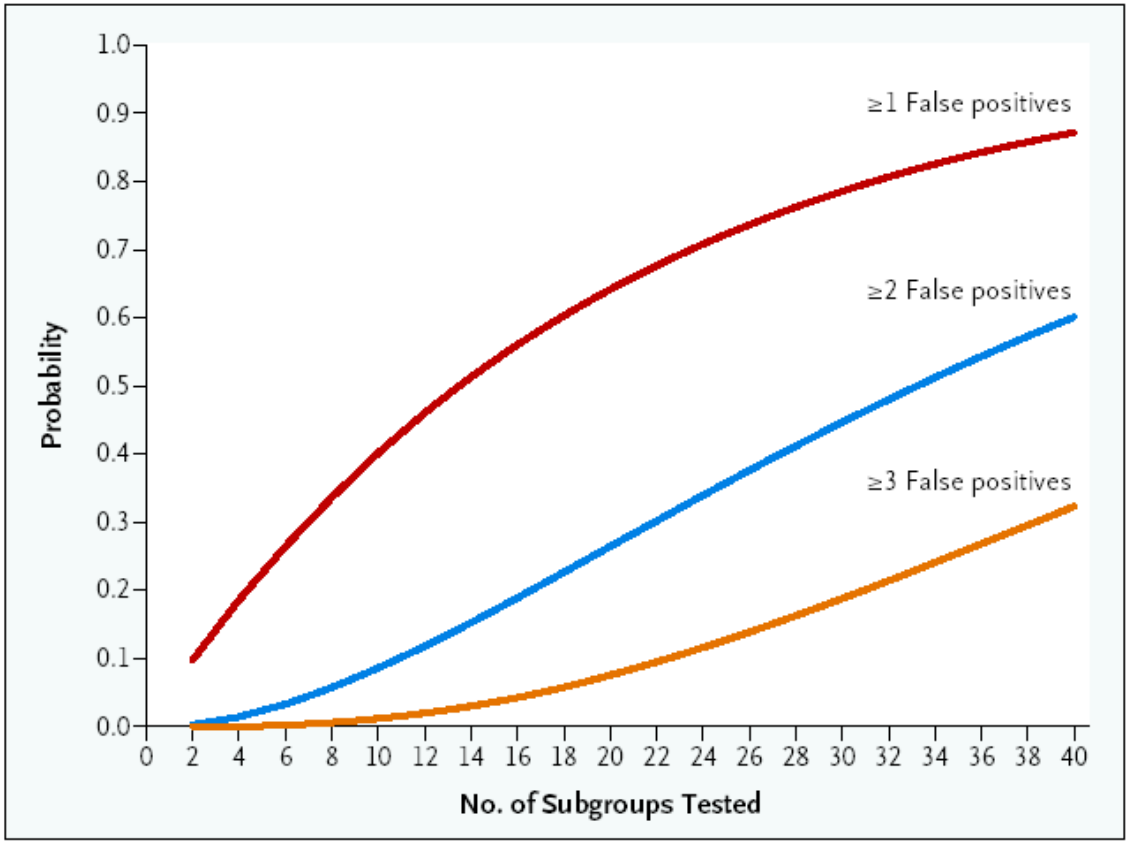
Multiplicité des tests statistiques

STATISTICS AND MEDICINE

The Challenge of Subgroup Analyses — Reporting without Distorting

Stephen W. Lagakos, Ph.D.

Related article, page 1706



Probability That Multiple Subgroup Analyses Will Yield at Least One (Red), Two (Blue), or Three (Yellow) False Positive Results.

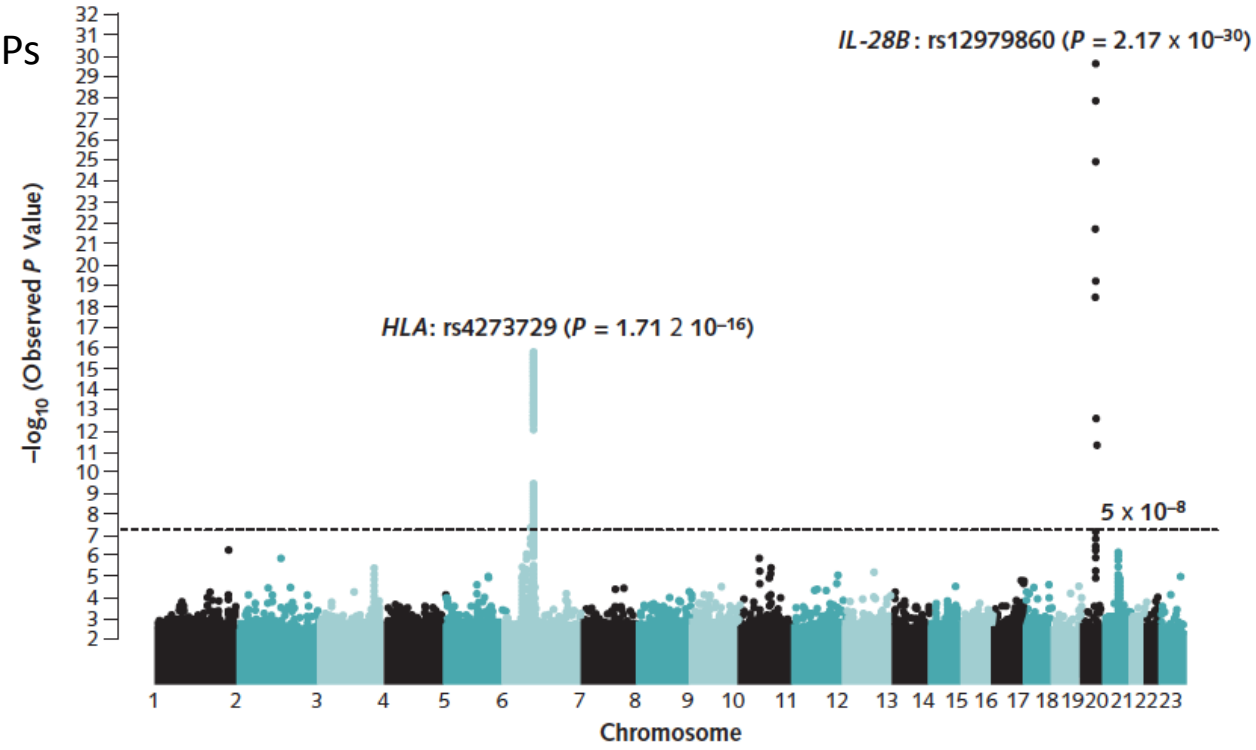
NEJM 2006;354:1706

Etude d'associations génétiques pangénomiques **GWAS (genome-wide association study)**

« Manhattan plot »

Figure 1. **Manhattan plot summarizing the genome-wide association results in 919 persons with spontaneous resolution of hepatitis C virus infection and 1482 persons with chronic hepatitis C virus infection.**

792 721 SNPs



Each point corresponds to a *P* value from a test of association for a single nucleotide polymorphism. The $-\log_{10}$ *P* values are plotted by location of the person's single nucleotide polymorphism across the genome. The dashed line represents an accepted level of genome-wide significance ($P = 5 \times 10^{-8}$). Single nucleotide polymorphisms in the *HLA* and *IL-28B* region on chromosomes 6 and 19, respectively, exceed this threshold. *IL-28B* = *interleukin-28B*.

Ann Intern Med. 2013;158(4):235-245. doi:10.7326/0003-4819-158-4-201302190-00003

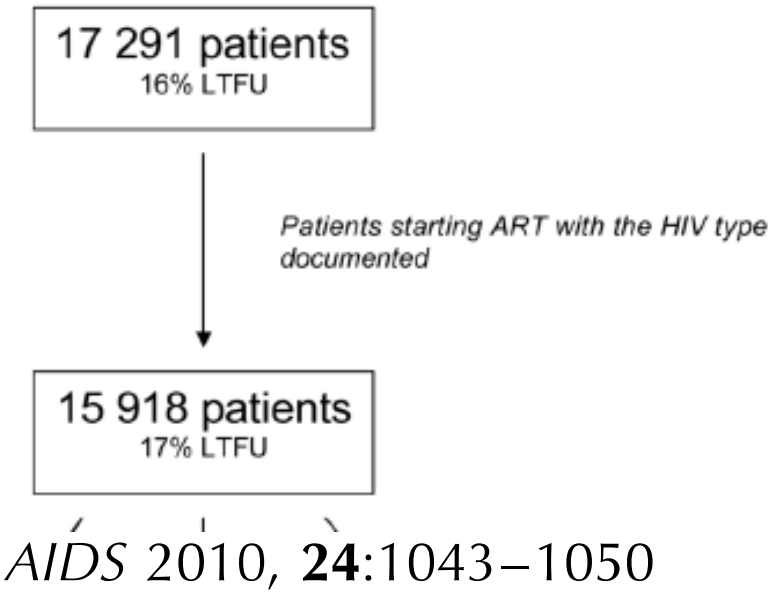
PB N°2: LES DONNÉES MANQUANTES

- Facteurs de confusion connus non mesurés
- Sorties d'étude, perdus de vue

IeDEA

International Epidemiologic Databases to Evaluate AIDS

GLOBAL COHORTS COMPOSED OF
>1,000,000
PERSONS LIVING WITH HIV/AIDS



AIDS 2010, **24**:1043–1050

JAMA®

Online article and related content current as of August 25, 2008.

Sampling-Based Approach to Determining Outcomes of Patients Lost to Follow-Up in Antiretroviral Therapy Scale-Up Programs in Africa

Elvin H. Geng; Nneka Emenyonu; Mwebesa Bosco Bwana; et al.

JAMA. 2008;300(5):506-507 (doi:10.1001/jama.300.5.506)

Figure 1. Derivation of Probability Weights

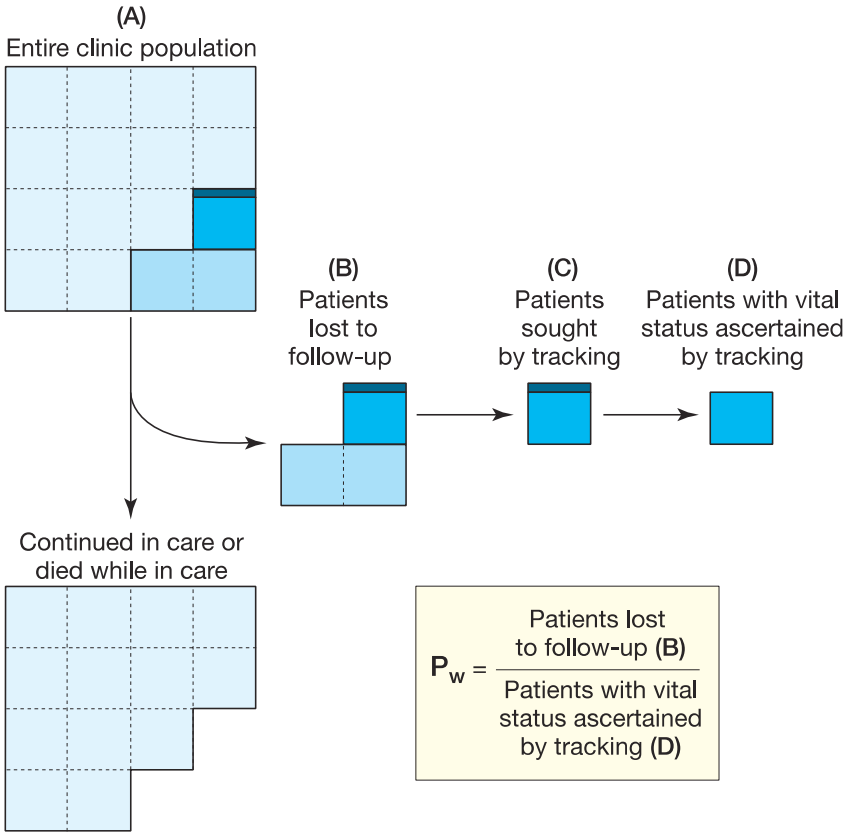
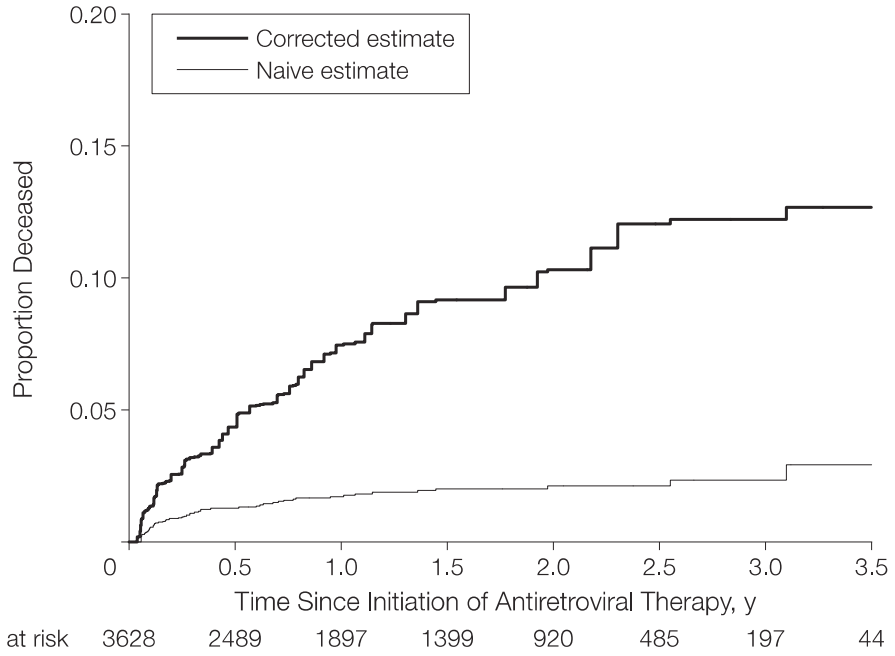


Figure 2. Naive and Corrected Mortality Estimates



PB N°3: LA CAUSALITÉ

Causalité

Précautions

- Questions/hypothèses posées en amont
- Critères de causalité (Bradford Hill)
  - Critères internes à l'étude (force de l'association, relation temporelle, relation dose-réponse et spécificité de la relation)
  - Externes (consistance, cohérence, plausibilité biologique, analogie, expérimentation)
- Données –omiques
  - Connaissances des mécanismes biologiques
  - Validation des résultats par des nouvelles expérimentations spécifiques

EN RECHERCHE CLINIQUE

Disponible en ligne sur
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

Revue d'Épidémiologie et de Santé Publique
Epidemiology and Public Health

Revue d'Épidémiologie et de Santé Publique 62 (2014) 1–4

Éditorial

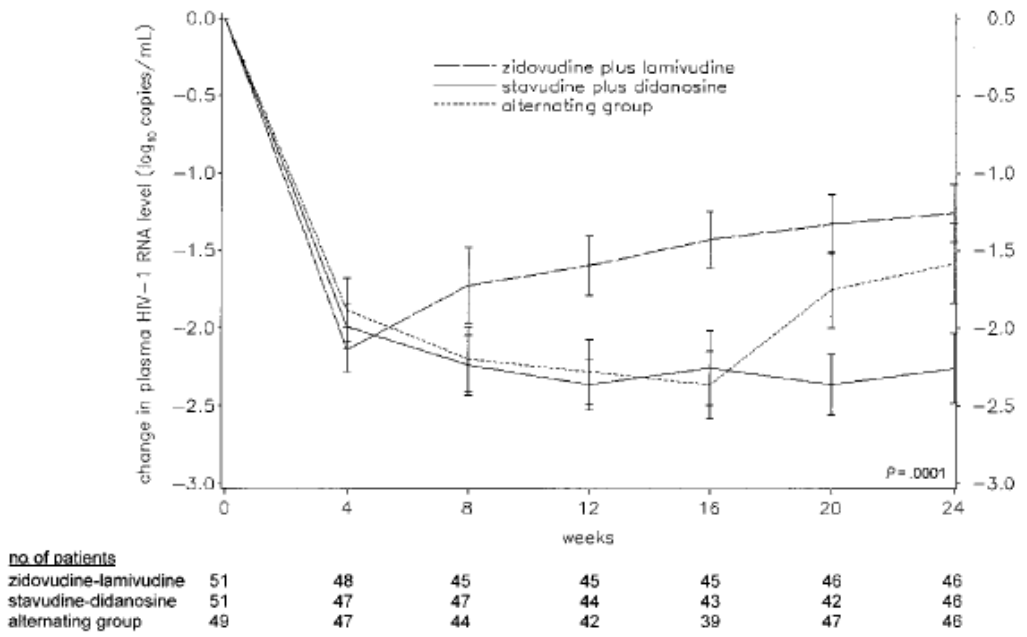
L'analyse des « Big Data » en recherche clinique

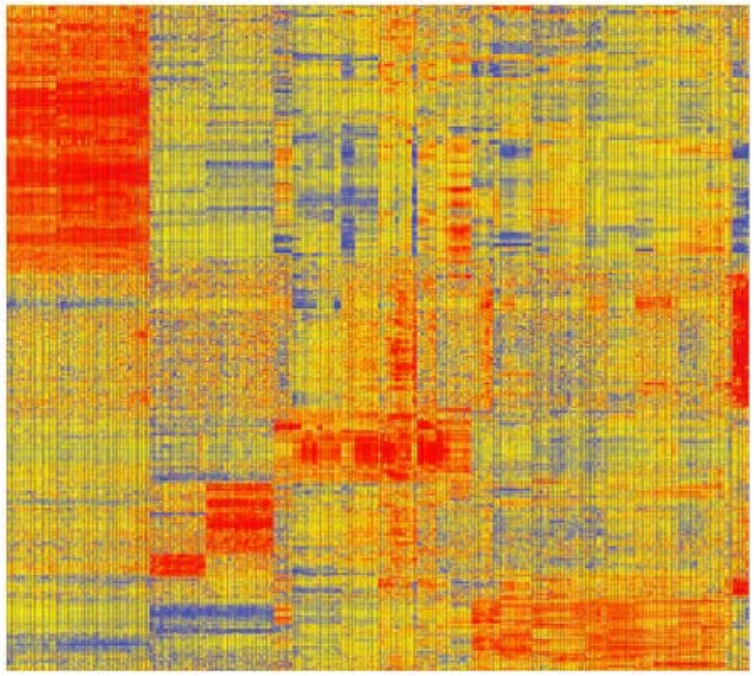
The analysis of ''Big Data'' in clinical research

R. Thiébaut [a,b,c,d,e,*], B. Hejblum [a,b,c,d,e], L. Richert [a,b,c,d,e]

[a] Inserm U897[1] épidémiologie et biostatistique, 33000 Bordeaux, France
[b] Inria SISTM[2], 33000 Bordeaux, France
[c] Institut de santé publique d'épidémiologie et de développement (ISPED[3]), université Bordeaux, 33000 Bordeaux, France
[d] Vaccine Research Institute, 94010 Créteil, France
[e] Unité de soutien méthodologique à la recherche clinique et épidémiologique (USMR), CHU de Bordeaux, 33000 Bordeaux, France

Exemple de deux essais cliniques





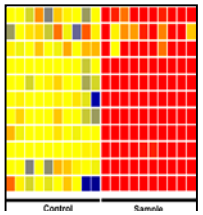
Essai	ALBI ANRS 070	DALIA-1
Nb de participants	151	19
Taille	67 Ko	200 Go

VACCINE RESEARCH INSTITUTE

DALIA TRIAL
(Dendritic cells And Lipo5 Immunization against Aids)

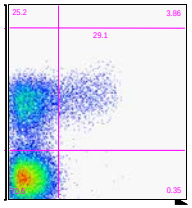
Les (big) data

- 846 000 probes (18 temps x 47 000 sondes) 26 Mo
- 18 612 000 beads (22 billes/sonde) 6 Go

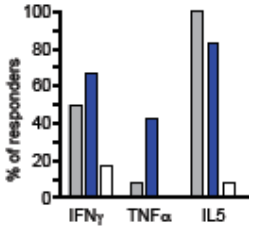


Gene profiling

- 30 populations cellulaires 0.05 Mo
- 2160 anticorps (18 temps x 15 tubes x 8 anticorps) pour 2.6 Go

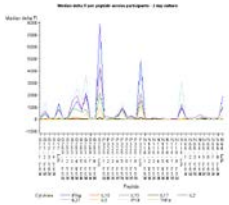


Cell responses

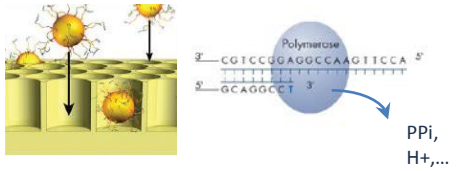


Cytokines

- 800 mesures/temps
- 0.35 Mo



Epitope mapping



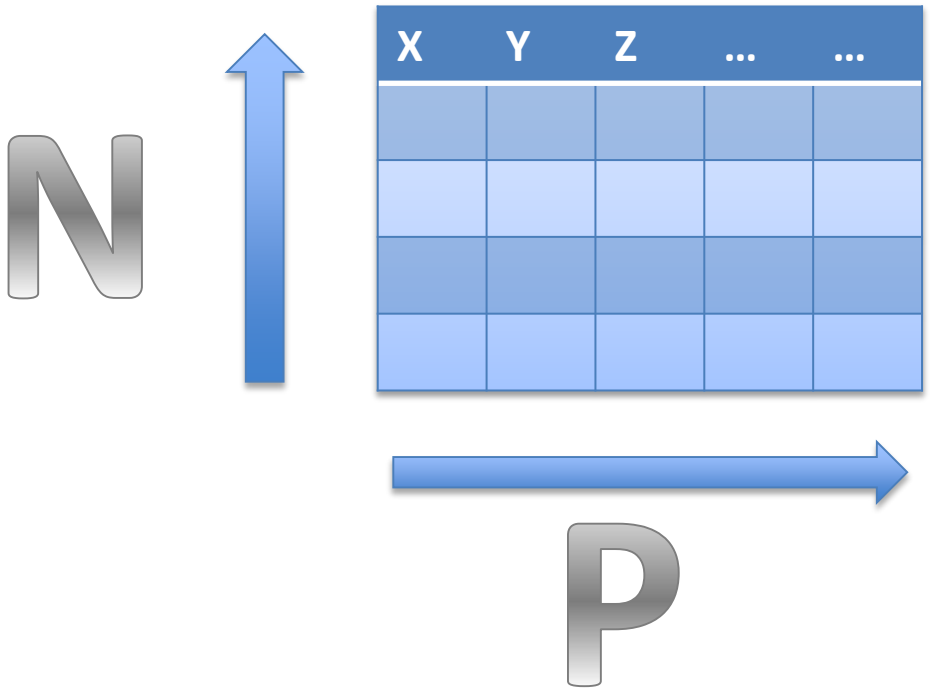
Viral changes/ adaptation

- 200 séquences
- 20 Mo

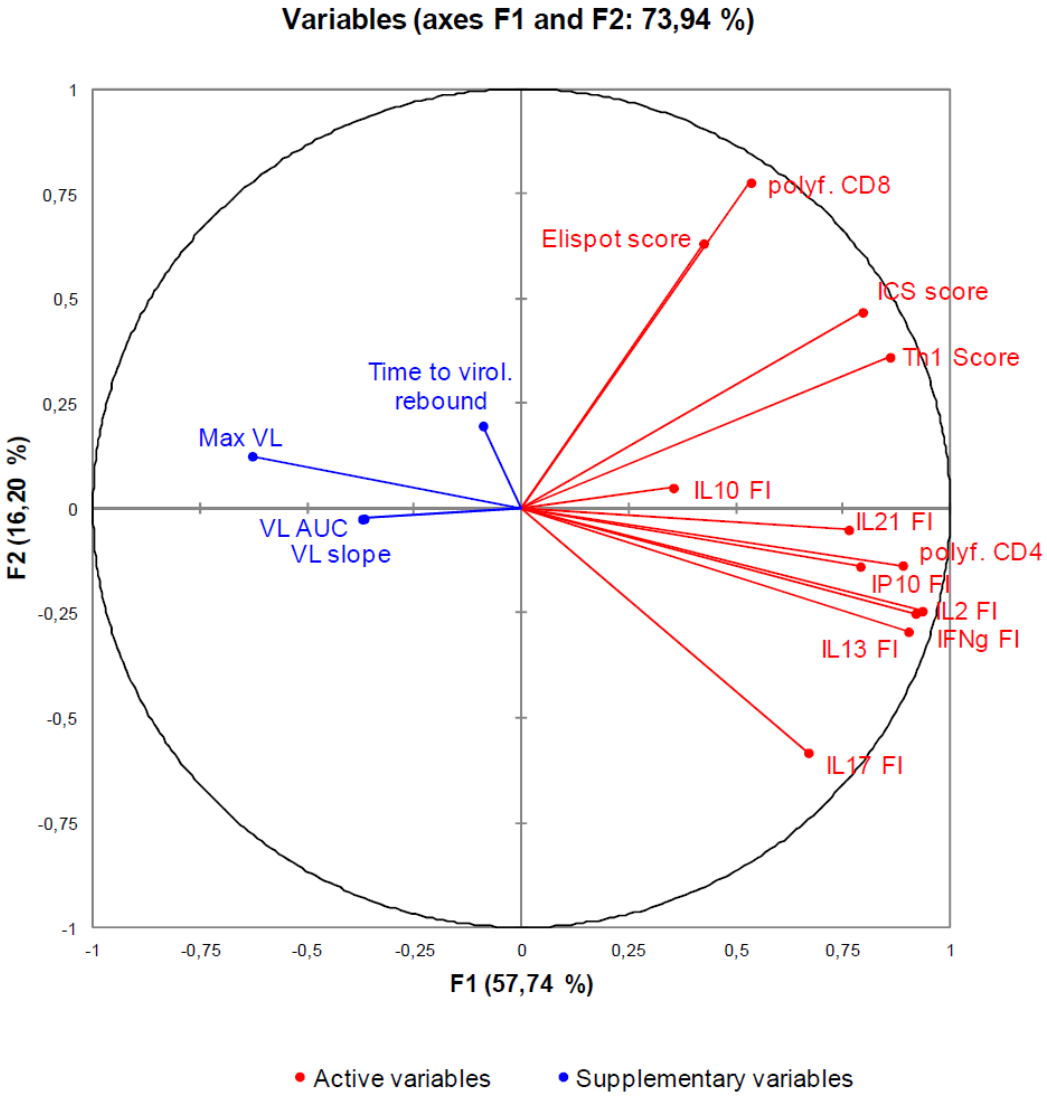
Point de vue statistique

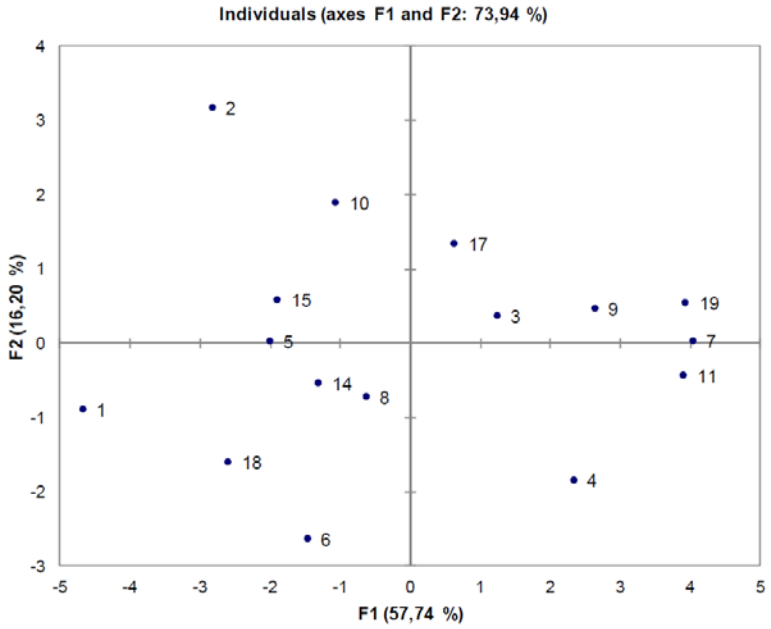
Données multidimensionnelles

$$N << P$$



Réponses immunologiques et virales



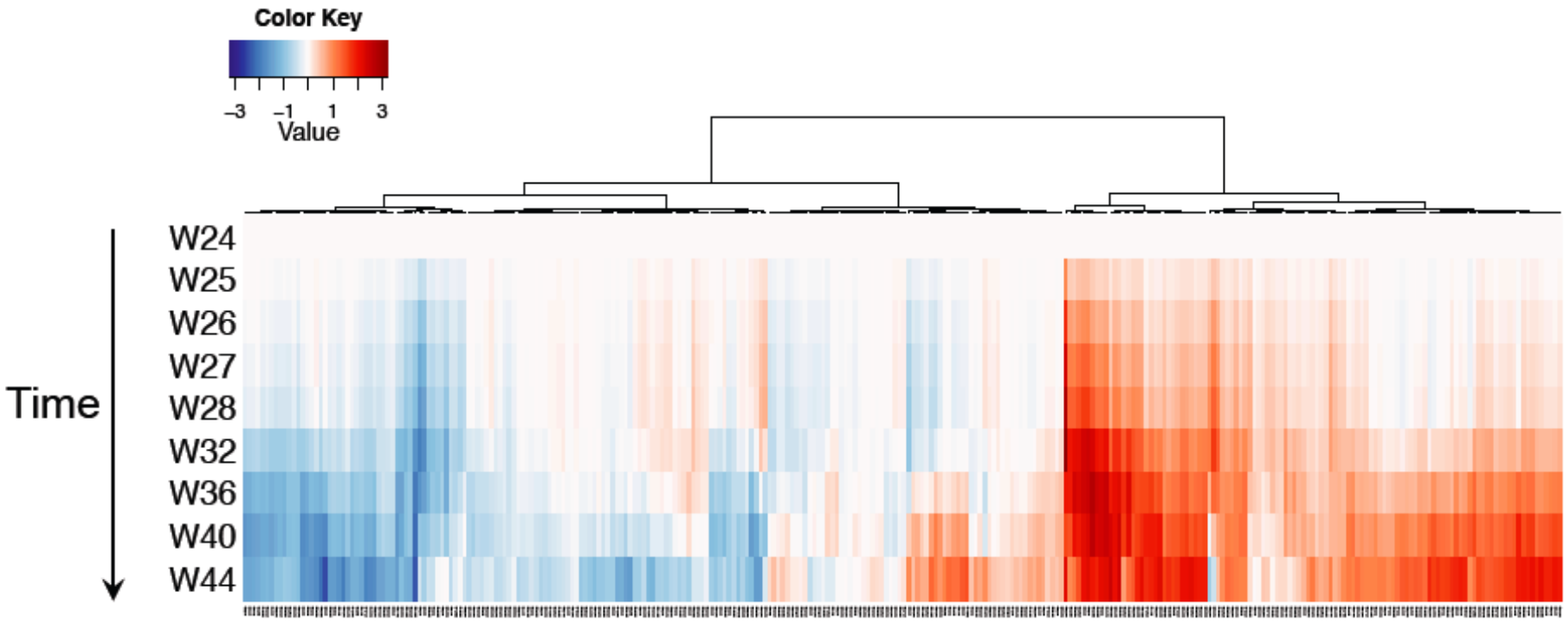


Levy et al. EJI 2014

26

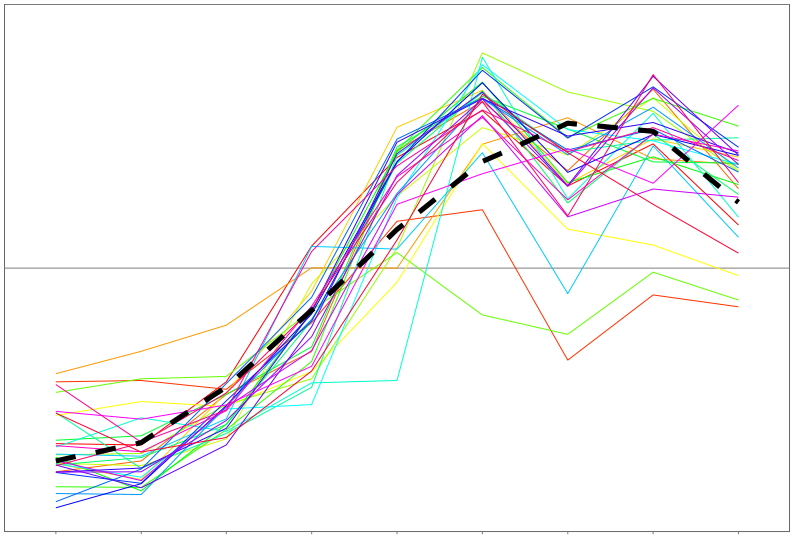
VACCINE RESEARCH INSTITUTE

Expression génique

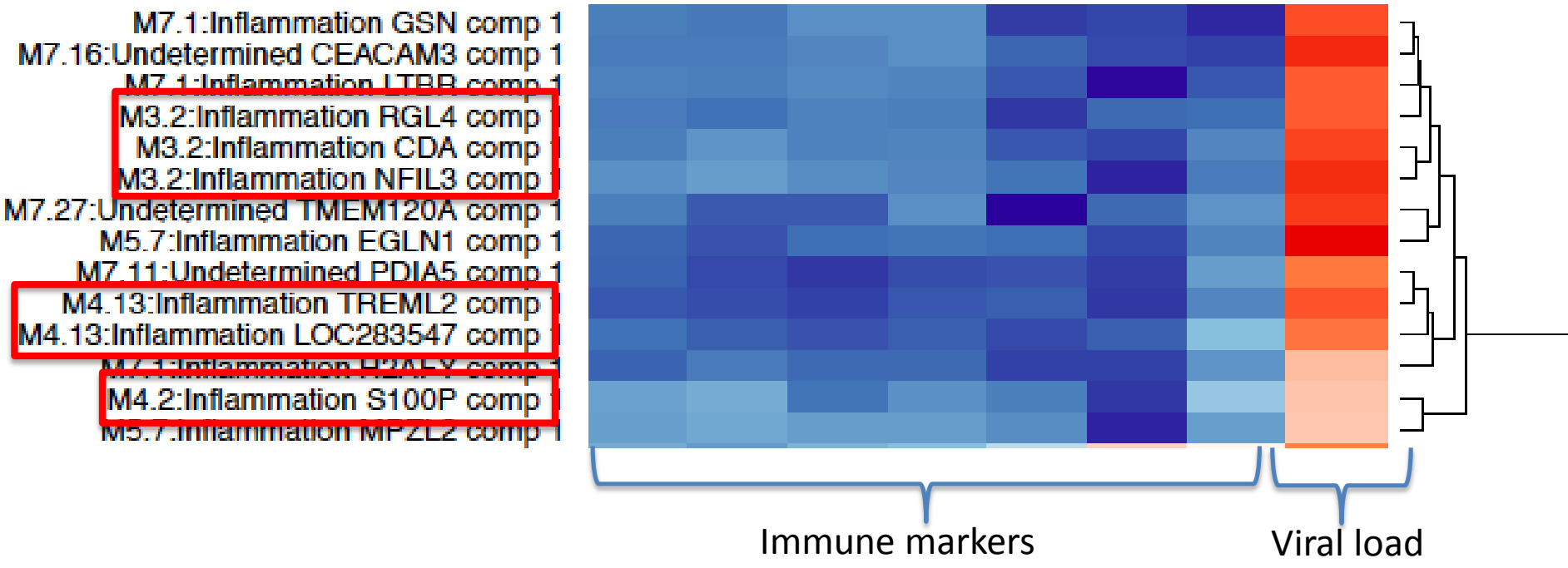


VACCINE RESEARCH INSTITUTE

Expression génique:
Time-course Gene Set Analysis



Correlations entre groupes d'expression génique, réponses immunes (W16) et pic de charge virale



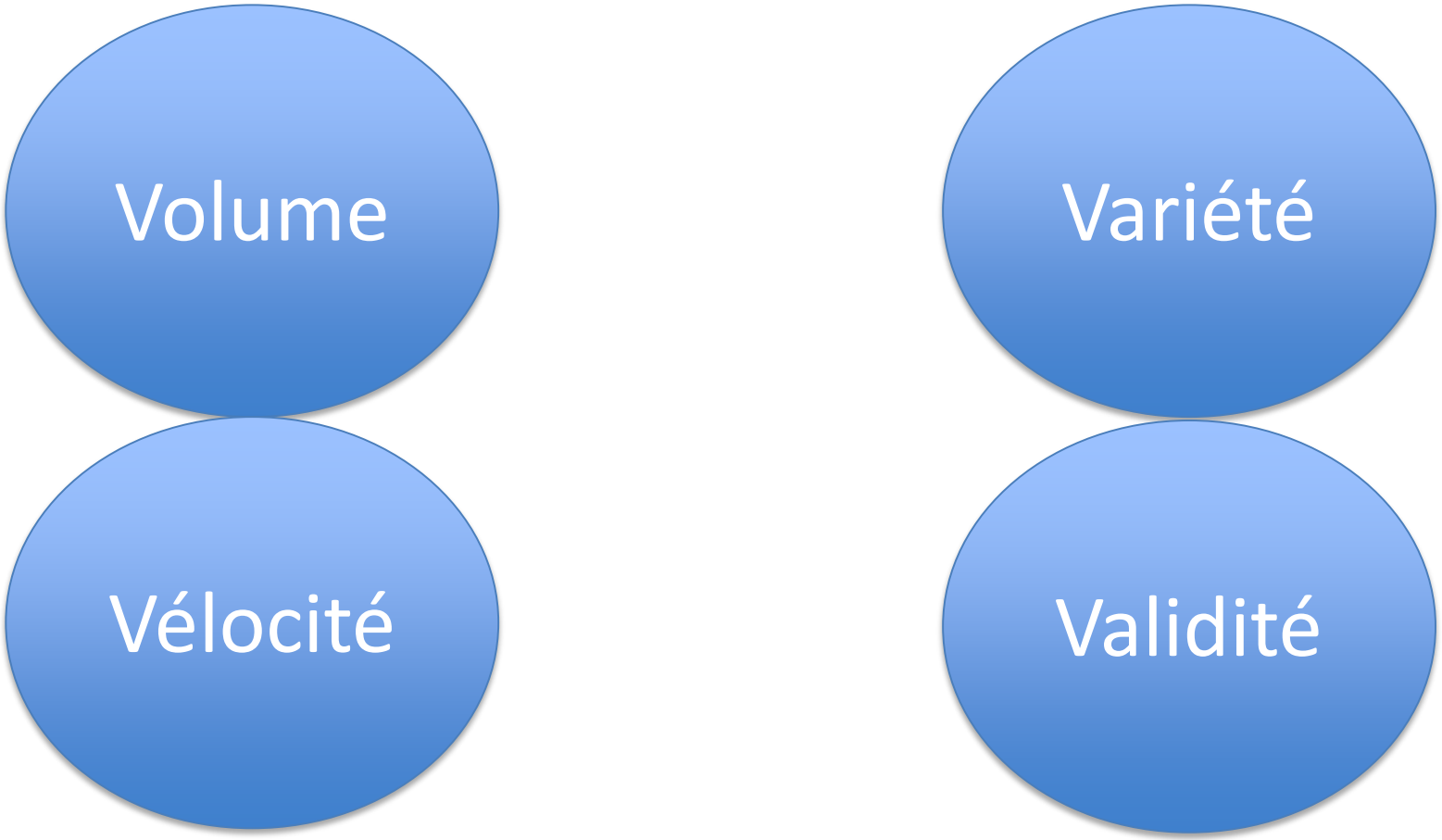
VACCINE RESEARCH INSTITUTE

Association négative des groupes de gènes retrouvée aussi avec les réponses à la vaccination anti-pneumococciques (Obermoser et al., Immunity 2013)

Conclusion: Big data en épidémiologie
Deux situations

- Les données volumineuses issues
  - Des bases de données médico-administratives
  - Des collaborations multicohortes
  - Des consortium (GWAS)
  - ➢ ! facteurs non mesurés, puissance statistique
- Les données « riches » issues de nouvelles technologies
  - Données omiques
  - Données d'imagerie…
  - ➢ ! Fléau de la dimensionalité

Conclusion: les 4 V



- Stockage, sécurité
- Débit
- Puissance de calcul

- **Comprendre les données**
- **Visualiser les données**
- **Analyser les données**

BMC Bioinformatics

BioMed Central

Research article

Open Access

Sparse canonical methods for biological data integration: application to a cross-platform study

Kim-Anh Lê Cao*[1,2], Pascal GP Martin[3], Christèle Robert-Gr
Philippe Besse[2]

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 147–154 • DOI: 10.1155/JBB.2005.147

RESEARCH ARTICLE

Classification and Selection of Biomarkers in Genomic Data Using LASSO

Debashis Ghosh[1] and Arul M. Chinnaiyan[2]

[1]*Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA*
[2]*Departments of Pathology and Urology, University of Michigan, 1300 Catherine Road, Ann Arbor, MI 48109-1063, USA*

ELSEVIER MASSON

Disponible en ligne sur ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

Revue d'Épidémiologie et de Santé Publique

Epidemiology and Public Health

Revue d'Épidémiologie et de Santé Publique 62 (2014) 1–4

Éditorial

L'analyse des « Big Data » en recherche clinique

The analysis of ''Big Data'' in clinical research

R. Thiébaut [a,b,c,d,e,*], B. Hejblum [a,b,c,d,e], L. Richert [a,b,c,d,e]

[a]Inserm U897[1] épidémiologie et biostatistique, 33000 Bordeaux, France
[b]Inria SISTM[2], 33000 Bordeaux, France
[c]Institut de santé publique d'épidémiologie et de développement (ISPED[3]), université Bordeaux, 33000 Bordeaux, France
[d]Vaccine Research Institute, 94010 Créteil, France
[e]Unité de soutien méthodologique à la recherche clinique et épidémiologique (USMR), CHU de Bordeaux, 33000 Bordeaux, France

Systems biology approaches to epidemiological studies of complex diseases

Hongzhe Li*

WIREs Syst Biol Med 2013, 5:677–686.

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context

Benoît Liquet [1,2,3,4,5,*], Pierre Lafaye de Micheaux [6], Boris P. Hejblum [2,3,4,5] and Rodolphe Thiébaut [2,3,4,5]

[1]School of Mathematics and Physics, The University of Queensland, Brisbane 4066, Australia
[2]Inria, SISTM, Talence, France
[3]Inserm, U897, Bordeaux, France
[4]Bordeaux University, Bordeaux, France
[5]Vaccine Research Institute, Creteil, France
[6] Department of Mathematics and Statistics, Université de Montreal, Canada

Qualité des données

- Garbage in, garbage out?

Conclusion

- Big data in epidémiology = Big opportunity + Big challenge

- Besoin:
  - De multidisciplinarité
  - De méthodes d'analyse adaptées
  - Et du savoir faire épidémiologique et de l'intégrité scientifique

Remerciements

Rodolphe Thiébaut

Linda Wittkop

Boris Hejblum