Rencontres 2019

Journée Dép. Bios. Montpellier, Juillet 2019

cirad

https://intranet.cirad.fr/

Plan

1. Introduction
2. Expériences au Cirad
    - Marie-Claude Deboin (Dgdrs-Dist)
    - Manuel Ruiz (UMR AGAP, Plateforme SouthGreen)
3. Session débat

Introduction

Qu'est-ce que c'est ?

> *umbrella term for non-traditional **strategies** and **technologies** needed to gather, organize, process, and gather insights from **large** or **complex** datasets*

- **large** : impossible de stocker et/ou processer dans un seul ordinateur
- **complex** : issus de plusieurs sources et fortement structurés
- Trois Vs : **V**olume, **V**elocity, **V**ariety

D'autres gros mots associés

- **Data science** : Les statistiques adaptés au Big Data (+ informatique)
- **Machine Learning / Intelligence Artificielle** : Méthodes **prédictifs** qui necessitent (souvent) beaucoup des données d'entrainement

Ça bouge !

Mars - INRA P. Ezanno et al. *L'Intelligence Artificielle & les Recherches en Santé Animale à l'Inra*.

Rapport du groupe de travail mandaté par le DSA

Juin - IRD C. Stephens (UNAM). *Using spatial data to unravel the complexity of emerging diseases*.

Analyse de co-occurrence d'espèces à partir de données du GBIF (*Global Biodiversity Information Facility*)

Juin - IRD F. Masseglia. (Inria). *BIG DATA : Analyser de GRANDS corpus de données scientifiques*.

Enjeux particuliers du domaine scientifique.

Mars 2018. Mission confiée par le Premier Ministre Édouard Philippe.

> *"Dopée par les progrès de l'intelligence artificielle, la révolution du **big data** contribue à rendre le monde plus transparent, plus quantifiable, mesurable à l'infini."*

Expériences au Cirad

-  **Marie-Claude DEBOIN** (Dgdrs-Dist)

  IST et big data : points de rencontre

-  **Manuel Ruiz** (UMR AGAP, Plateforme SouthGreen)

  Défis liés aux Big Genomics Data

Débat

The end of the theory ?

> *"This is a world where **massive amounts of data and applied mathematics replace every other tool** that might be brought to bear. (...) Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. **With enough data, the numbers speak for themselves.** (...) **We can stop looking for models.** We can analyze the data **without hypotheses** about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms **find patterns where science cannot.**"*

Chris Anderson, 2008, Wired[1]

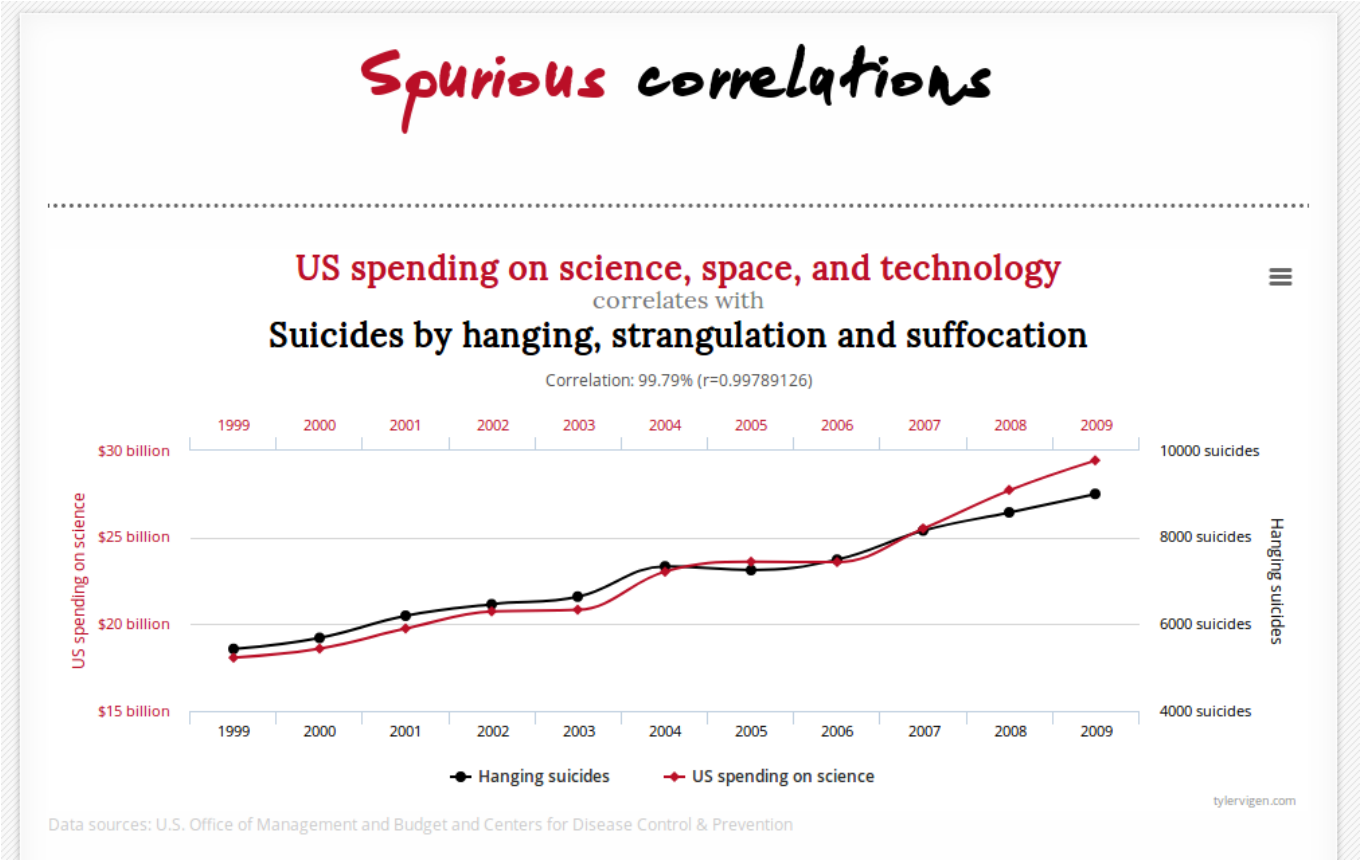
[1]The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired https://www.wired.com/2008/06/pb-theory/

Big noise ?

> *"The rise of **Big Data** has the potential to help us predict the future, yet **much of it is misleading, useless or distracting**."*

Nate Silver, 2013, *The Signal and the Noise*

Corrélations et causalités



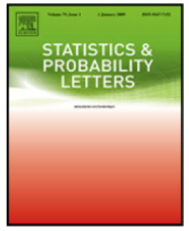
Quantité vs qualité

Statistics and Probability Letters 136 (2018) 142–145

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro



When small data beats big data

Julian J. Faraway, Nicole H. Augustin *

Department of Mathematical Sciences, University of Bath, United Kingdom

ARTICLE INFO

Article history:
Available online 17 February 2018

Keywords:
Big data
Small data

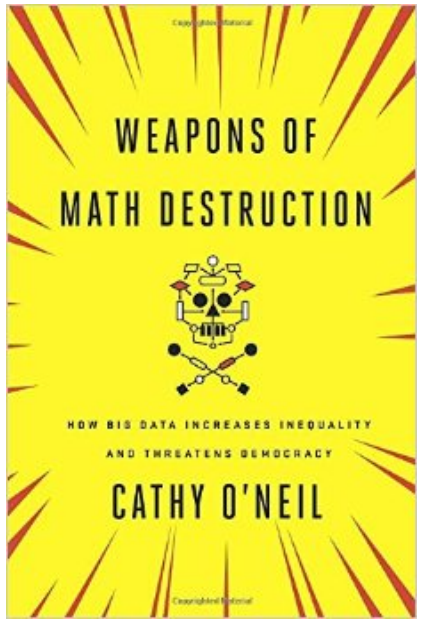
ABSTRACT

Small data is sometimes preferable to big data. A high quality small sample can produce superior inferences to a low quality large sample. Data has acquisition, computation and privacy costs which require costs to be balanced against benefits. Statistical inference works well on small data but not so well on large data. Sometimes aggregation into small datasets is better than large individual-level data. Small data is a better starting point for teaching of Statistics.

Data neutrality

> *"Not everything that counts can be counted, and not everything that can be counted counts."*

William Bruce Cameron (Sociologist)



> *"How Big Data increases inequality and threatens Democracy"*

Somme-nous préparés ?

> *"Don't sleep on the basics. Someone probably solved your problem in the 80s."*

Nick Strayer (2019). **Using AWK and R to parse 25tb**.[1]

> *"You can have a second computer once you've shown you know how to use the first one."*

F. McSherry, M. Isard, D. G. Murray (2015). **Scalability! but at what cost ?** *Usenix conference*[2]

[1] https://livefreeordichotomize.com/2019/06/04/using_awk_and_r_to_parse_25tb/

[2] https://t.co/oV1Qs6zatz

Comment se positionner au Cirad / Bios ?

- On réagit au fur et à mesure ? on se **prépare** ? comment ?
- **Infrastructure** logistique ? formation ? partenariat ? collaborations ?
- On devrait tous nous **former** d'avantage à l'analyse de données ?


BIG
DATA
ANALYTICS
STORAGE
TECHNOLOGIES
INFORMATION
INTERNET
SETS
MANAGEMENT
SOFTWARE
PETABYTES
SYSTEMS
INCLUDE
SIZE
RESEARCH
PARALLEL
SEARCH
DATABASES
NETWORKS
RECORDS
HUNDREDS
CAPACITY
BIOLOGICAL
PROCESSING
COMPLEX
SOCIAL
RELATIONAL
SHARED
DISK
EXAMPLES
EVERY
ALSO
TOOLS
TARGET
APPLIED
SENSOR
MAY
DEFINITION
MOVING
WITHIN
ARCHIVES
MANAGE
CAPTURE
BUSINESS
DISTRIBUTED
GENOMICS
NEEDED
TIME
ONE
LARGE
PROCESS
STORE
ABILITY
PERFORMANCE
LOGS
COST
TERABYTES
CASE
ELAPSED
TOLERABLE
SAN
GARTNER
WORKING
MASSIVELY
GROW
DIFFICULTY
LARGER
CONTINUES
TENS
CITATION
SET
USE
USING

Merci !