Big Data and Data Science

Mark Asch - IMU/VLP/CSU

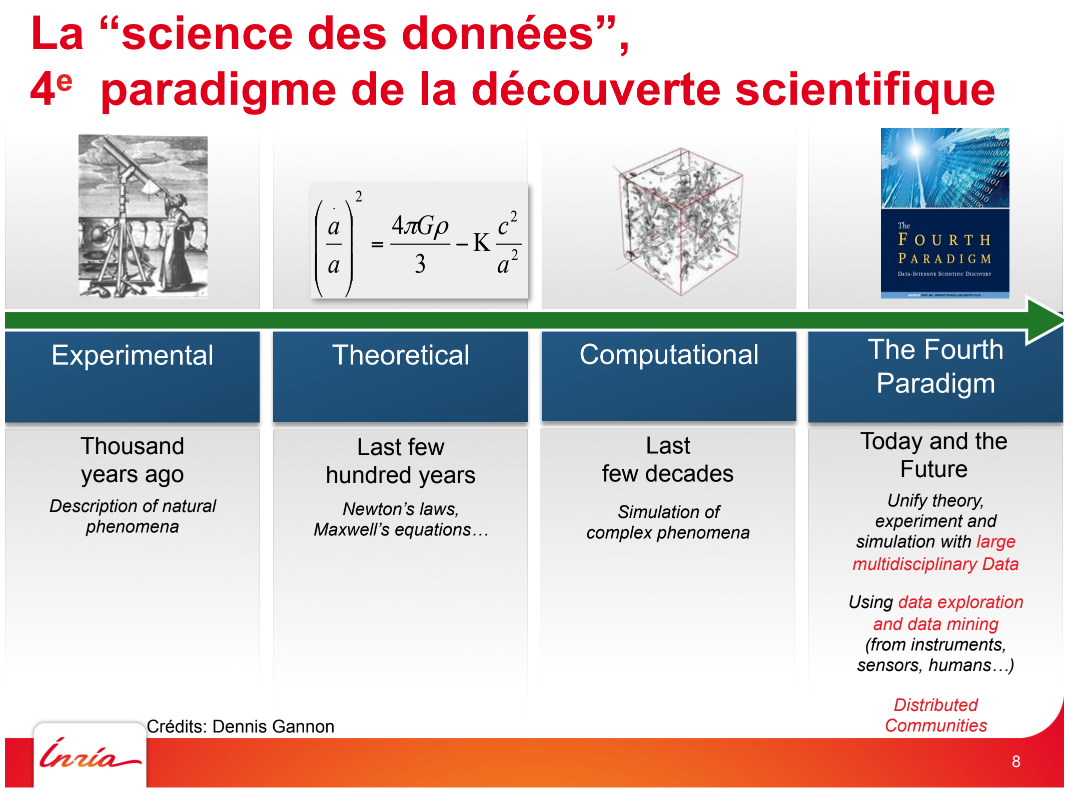
2023

The 4th paradigm

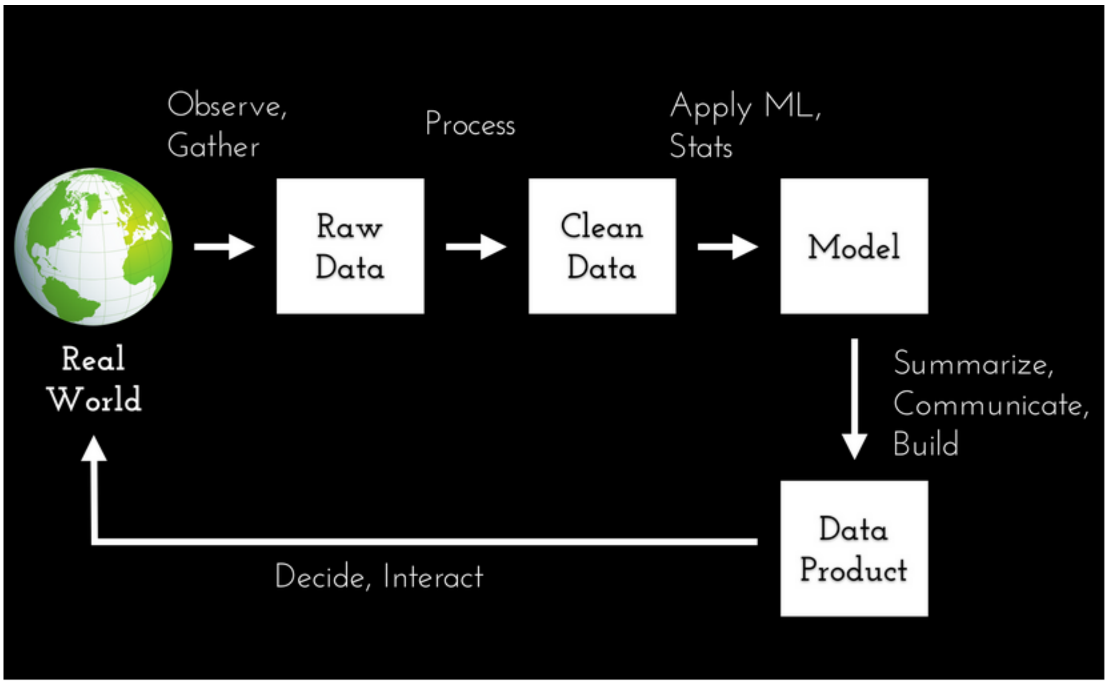
1. The scientific methodology: (three pillars)
   (a) Experiments
   (b) Theory
   (c) Simulation

2. The 4th paradigm: science from the data
   (a) the data deluge - instruments, connected objetcs, internet, etc. (exponential growth)
   (b) data science... "let the data speak "

4th paradigm



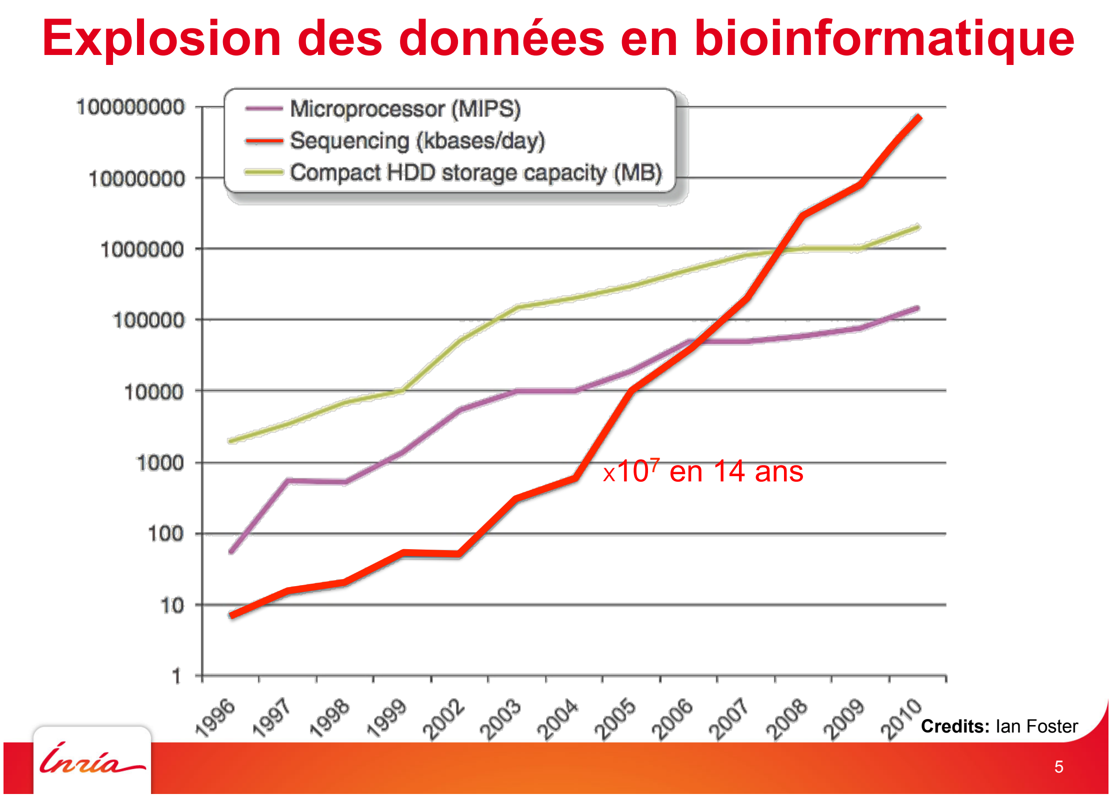
Data Science



The 4 "V's" of Big Data

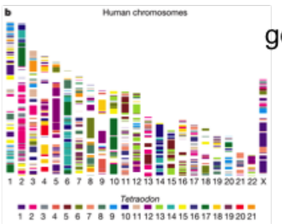
1. Volume
2. Variety
3. Velocity
4. Veracity

Volume



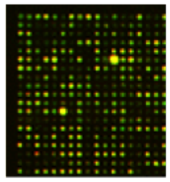
Explosion des données en bioinformatique
Microprocessor (MIPS)
Sequencing (kbases/day)
Compact HDD storage capacity (MB)
100000000
10000000
1000000
100000
10000
1000
100
10
1
x10^7 en 14 ans
1996
1997
1998
1999
2002
2003
2004
2005
2006
2007
2008
2009
2010
Credits: Ian Foster
5

Variety



génomique

NGS

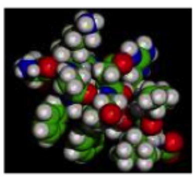


transcript-
omiques

marrays
RNASeq

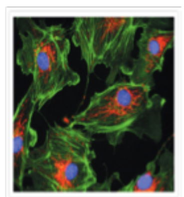
protéomiques
métabolomiques

spectro.
masse



structurales

cristallo.
RMN



images
biologiques

microscopie

images
médicales

IRM

cliniques

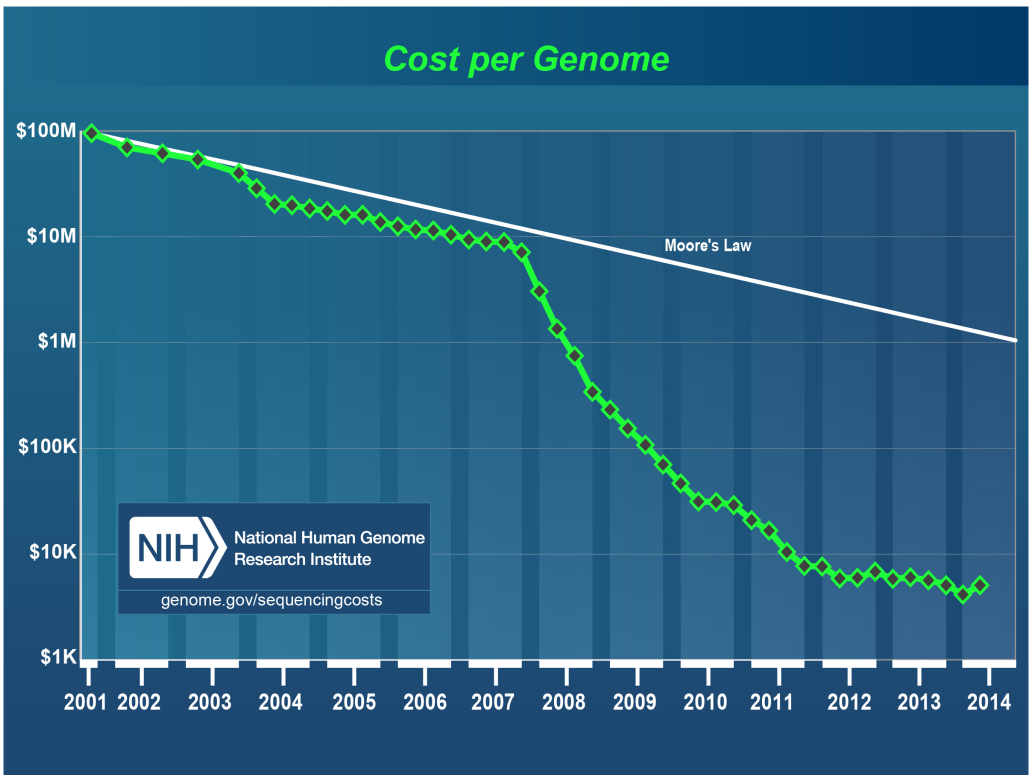
cohortes
biobanques

biblio-
graphiques

environ-
nementales

Velocity

600 Gb/run
~ 5 Go/h



Veracity

- trust, integrity, quality, transparency
- security
- confidentiality
- protection of private life
- intellectual propriety

Big Data and Statistics

- For both, we want to learn from the data...
- Statistiques analyze primary data (experimental, samples) and try to verify hypotheses.
- Big data attempts to use secondary data (observations) to deduce the hypotheses and hence to create new insights.
- Conclusion : the 2 approaches are complementary and should proceed hand-in-hand... to facilitate and provide tools for decision-making based on the data. The 2 fields should be united to draw reliable onclusions from available data.
- A lack of expertise in statistics can (and has - eg. using omic data...) led to fundamental errors!

  $\Rightarrow$ clinical trials cancelled for an anti-cancer treatment;
  $\Rightarrow$ effect of GMO on cancer;
  $\Rightarrow$ Google flu... (correlation vs causation!)

Statistics for Big Data

✔ General skills of a "data scientis":

- statistics
- linear algebra
- programming

✔ Complementary skills:

- data preparation ("data wrangling, munging, scraping")
- modeling
- visualization
- communication

Statistical Inference

✔ The data represent traces of real-world processes that depend on the way we collect the data.

✔ Two sources of uncertainty:

- random character of the process itself,
- uncertainty due to the data collection method

✔ The process that leads us from the world to the data, then from the data to the world, is called statistical inference

- procedures, methods, theorems that allow us to extract information from data that come from a random/stochastic process

Populations and Samples

Definition 1. *A population is the set of all the objects (observations) being studied. Their number is denoted by N.*

Definition 2. *A sample is a subset, of size n, $n \leq N$, drawn from the population. We examine these observations to draw conclusions and infer things about the population.*

For Big Data:

✘ N =ALL ???

✘ Correlation $\Longrightarrow$ Causation ???

Models

Definition 3. *A model is an attempt to understand and represent the nature of reality. It is a construction from which all superfluous details have been eliminated.*

CAVEAT :

- "Models are models, not reality!"
- "All models are wrong, but some are useful."

Statistical Models

We seek the underlying process...

✔ What comes first?

✔ What influences what?

✔ What is the cause of what?

✔ What would be a test of his?

How to construct a model

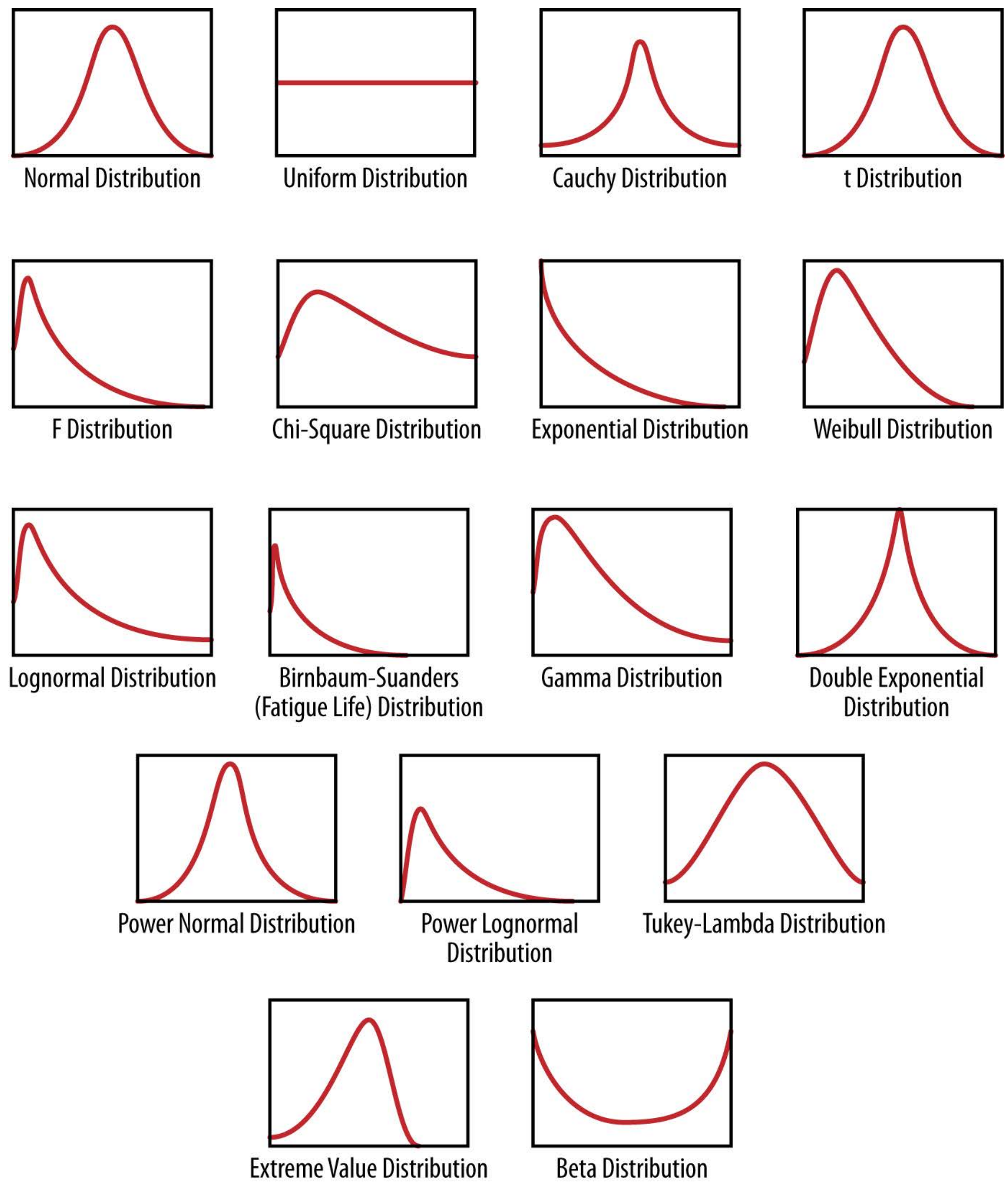
✔ Exploratory data analysis (see below)

- calculate basic statistics
- draw graphics
- obtain intuition

✔ Probability distributions

- the foundations of classical statistical models
- not all processes generate data that ressemble known distributions (Gauss, Poisson, Weibull, etc.), but many do
- these laws attribute a probability to a subset of possible outcomes by means of a corresponding function

Law/distributions of Probability



References

1. M. DeGroot, M. Schervish, *Probability and Statistics*, Addison Wesley, 2002.

2. Spiegel, Murray et Larry Stephens, *Statistics: Cours et problèmes,* 3ème édition, Série Schaum/McGraw Hill. 2000.

3. V. Mayer-Schönberger et K. Cukier. *Big Data: La révolution des données est en marche.* Robert Laffont. 2014.

4. H. Laude. *Data Scientist et langage R - Guide d'autoformation à l'exploitation des Big Data*. Editions ENI. 2016.

5. M. Lutz. *Data Science : fondamentaux et études de cas: Machine Learning avec Python et R*. Eyrolles. 2015.

6. G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer. 2013. Freely availaible for download.

7. Rachel Schutt and Cathy O'Neil. *Doing Data Science.* O'Reilly. 2014.

8. I. Goodfellow, Y. Bengio, A. Courville. *Deep Learning.* MIT Press. 2016.

http://www.deeplearningbook.org