BigData accelerated computing in R: an application in Metagenomics

Ndeye Aram GAYE
ndeye-aram.gaye@jouy.inra.fr

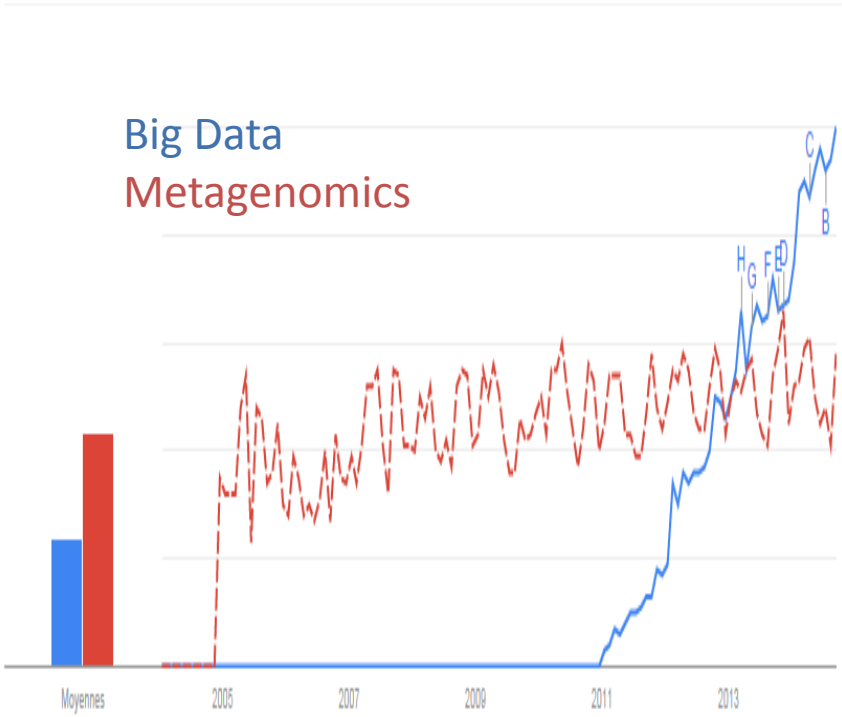
What is BIG DATA ?

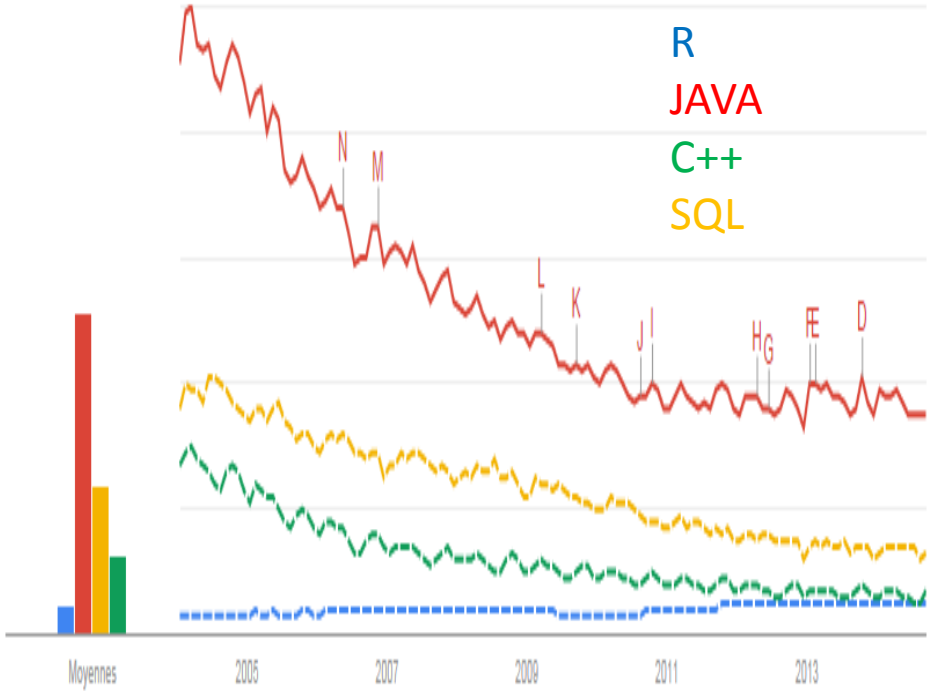
➢ Ambiguous and vague

- **Volume :** data intensity
- **Velocity :** data fast produced and the speed
- **Variety :** the degree of diversity
- **Veracity :** quality and provenance of data
- **Complexity :** data management become complex

➢ What are applications domains ?

BIG DATA: evolution



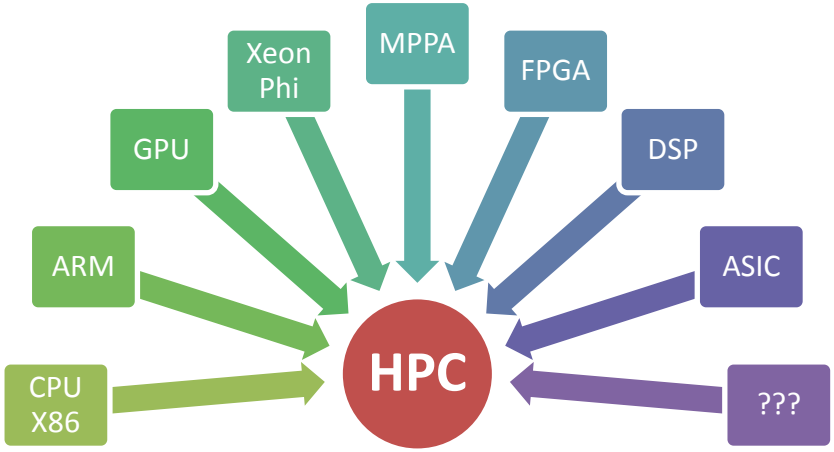


Source: Google Trends

Mckinsey prediction : big data the next frontier for innovation , May 2011

How to process Big Data ?

Hardware solutions



Software solutions

- **What is the complexity of these solutions ?**
- **How introduce such solutions to finals users ?**

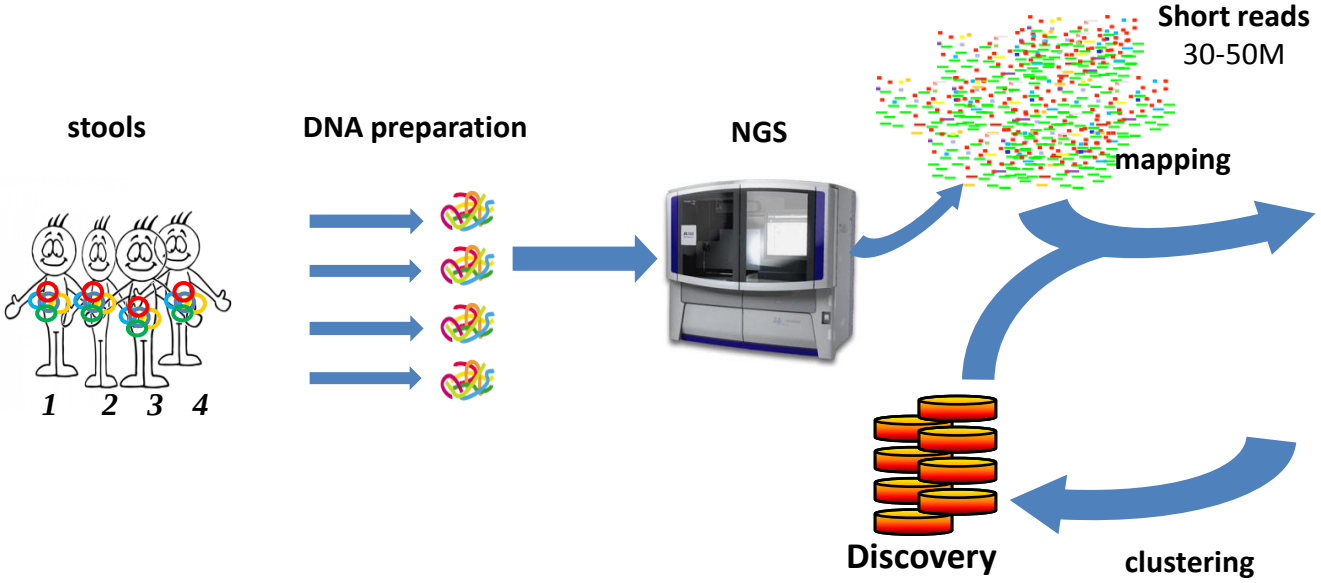
- Language and scientific computing environment
- Free and open source software
- Large community of users

- Interpreted language:

  Pros: easy to use, high level language,
  Cons: slow, no suitable for big data analysis

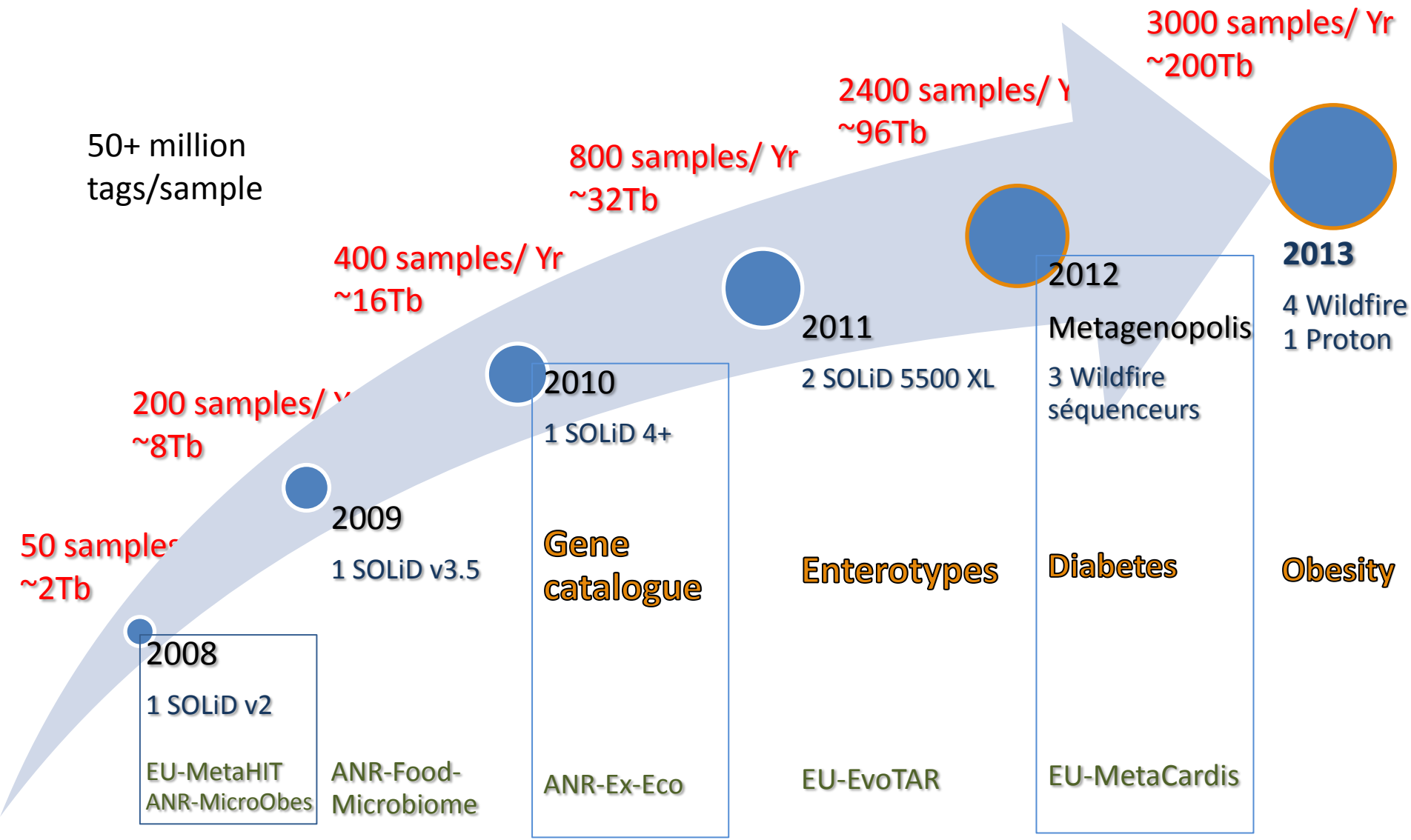
An application domain: Metagenomics

Biology → DNA preparation → Get Sequences → Compare to reference → Count



BiG dAtA

BiG dAtA


50+ million tags/sample
50 samples
~2Tb
200 samples/ Y
~8Tb
400 samples/ Yr
~16Tb
800 samples/ Yr
~32Tb
2400 samples/ Y
~96Tb
3000 samples/ Yr
~200Tb
2008
1 SOLiD v2
EU-MetaHIT
ANR-MicroObes
2009
1 SOLiD v3.5
ANR-Food-Microbiome
2010
1 SOLiD 4+
Gene catalogue
ANR-Ex-Eco
2011
2 SOLiD 5500 XL
Enterotypes
EU-EvoTAR
2012
Metagenopolis
3 Wildfire
séquenceurs
Diabetes
EU-MetaCardis
2013
4 Wildfire
1 Proton
Obesity

✓data analysis of quantitative metagenomics

Preprocessing

Downsizing

Normalization

MetaOMineR

Processing

Gene Counting

Sample Clustering

Biomarker Identification

.
.
.

Wish of Users

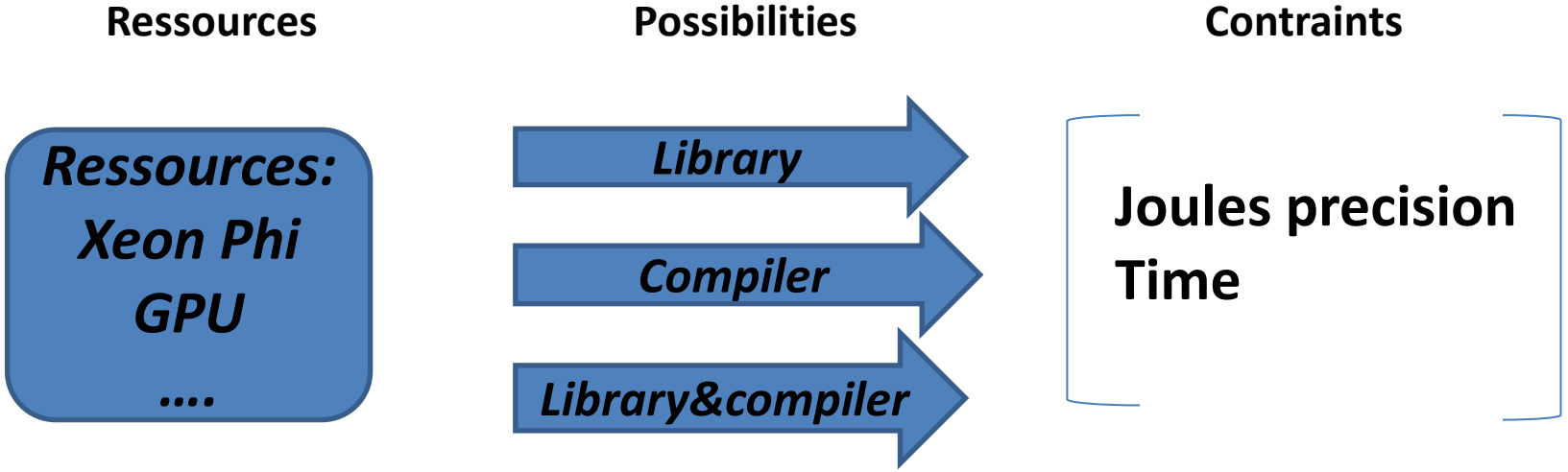
« I want to process and analyze Big Data on R without changing my programming habits … »

- How to satisfy such requirements ?
- What is the best way to present my solution users ?
- Give them a lesson on their HPC or hide the complexity of using HPC tools ?

Ways to do this…

- DSL: a programming language whose specifications are dedicated to a specific application domain
- Old fashion optimization : smart library
- New fashion : DSL -> smart compiler
- Smart library and smart compiler

To sum up



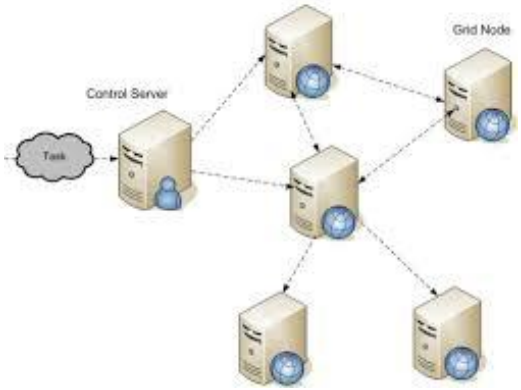
- A graph problem

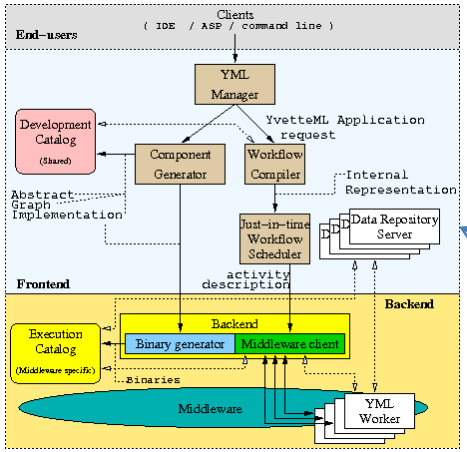
Mach Project (Massive calculations on heteregeous systems)

- European Project: private and public partners
- Duration: 3 years

Architectures Targets

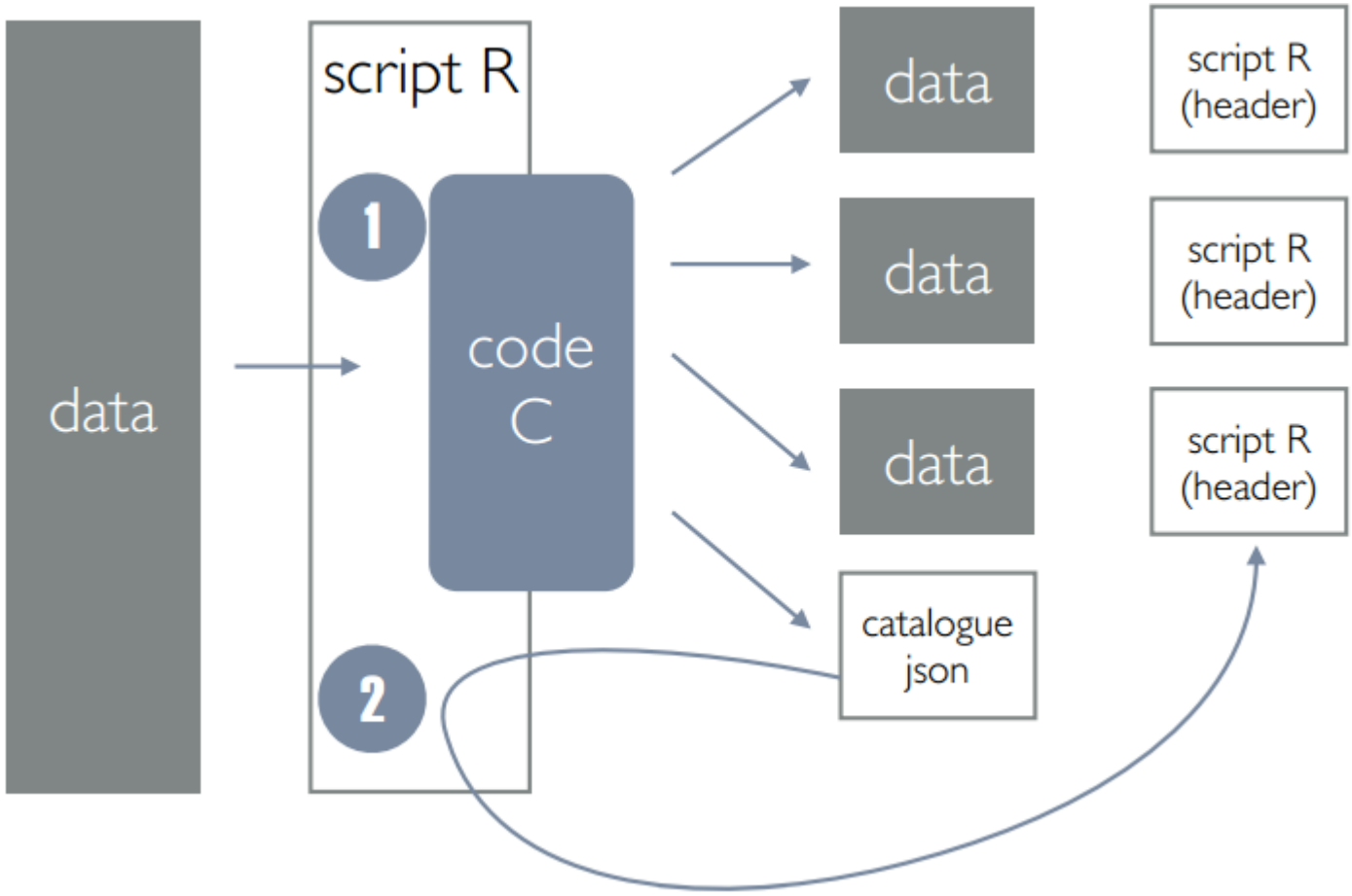
Scheduler





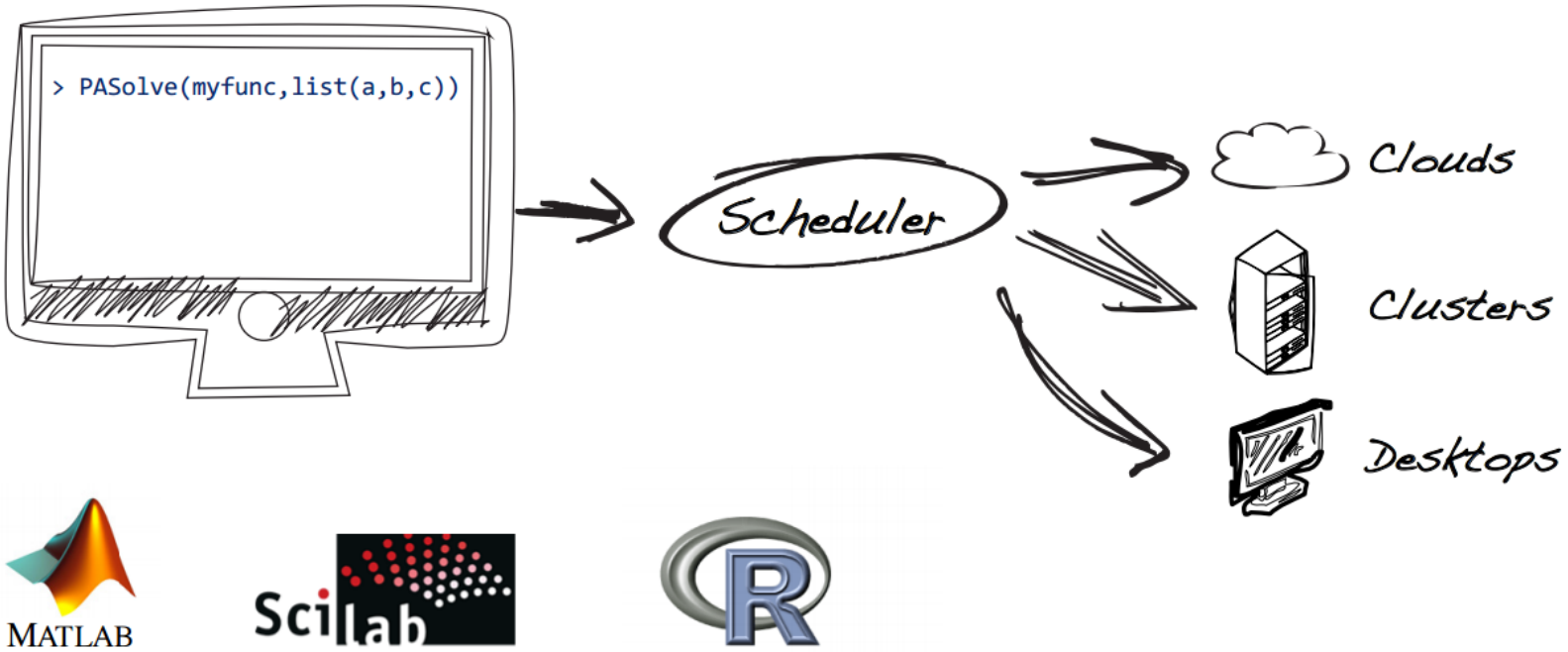
At Current Day

MapReduce/ 'R' Megapack



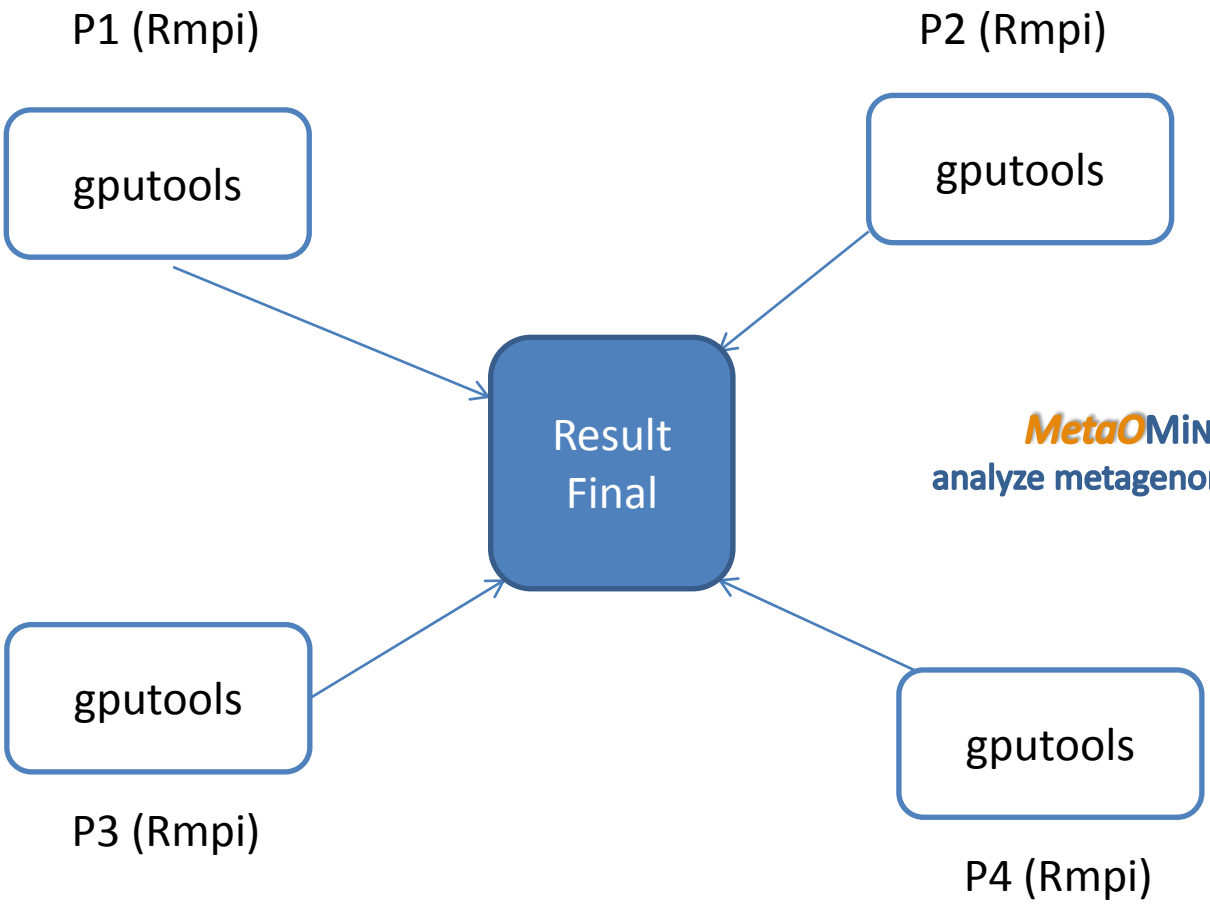
ACTIVEEON
SCALE BEYOND LIMITS

Jobs soumission/ 'R' PARConnector



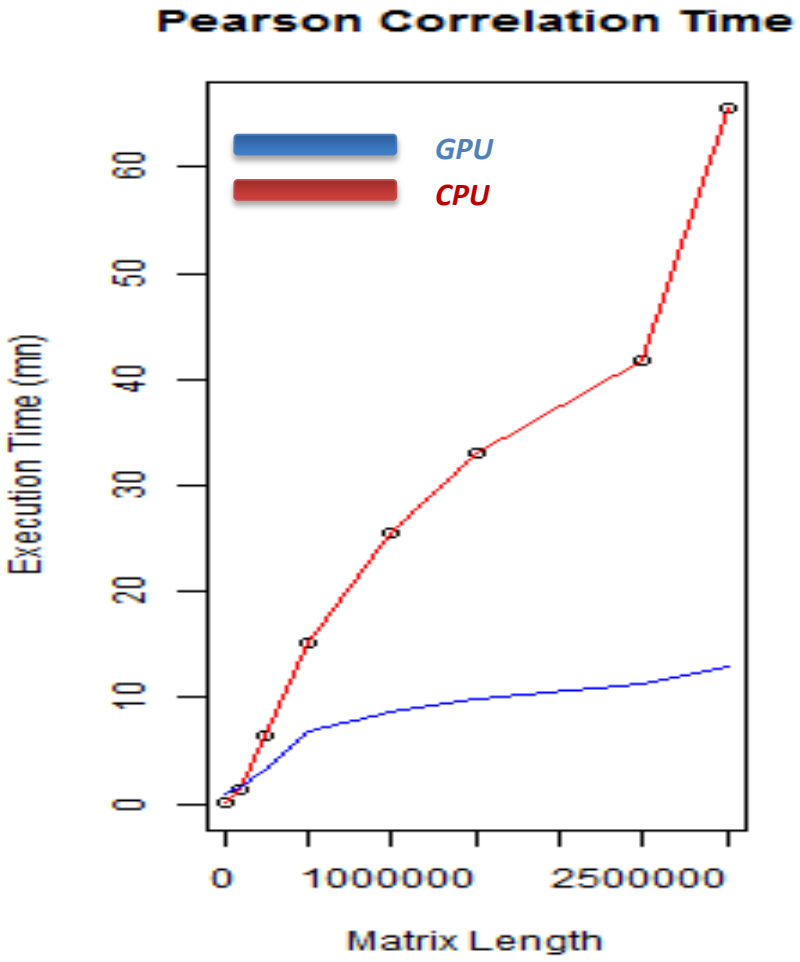
Intégration transparente dans votre environnement scientifique

'R'gpuStat



MetaOMiNeR:
analyze metagenomic datasets.

Benchmark of Pearson correlation in GPU



Conclusion

- Approach the problem from a graph is it the best solution?

- Big Data is -there becoming a barrier to innovation?

Efficient Maximum Flow Algorithms, By Andrew V. GOLBERG && Robert E. TARJAN, 2014 ACM 0001-0782/14/08

Thanks

Jad Abou Ghantous

Anne-Sophie Alvarez

Jean-Michel Batto

Magali Berland

Nahid Emad

Amine Ghozlane

Vincent Heuschling

Emmanuelle Le Chatelier

Pierre Léonard

Nicolas Pons

Florian Plaza

Edi Prifti