Clouds in biomedical sciences Part IV – entering a new world

Vincent Breton

July 28th 2014

Enrico Fermi school of physics

Session IV: the future

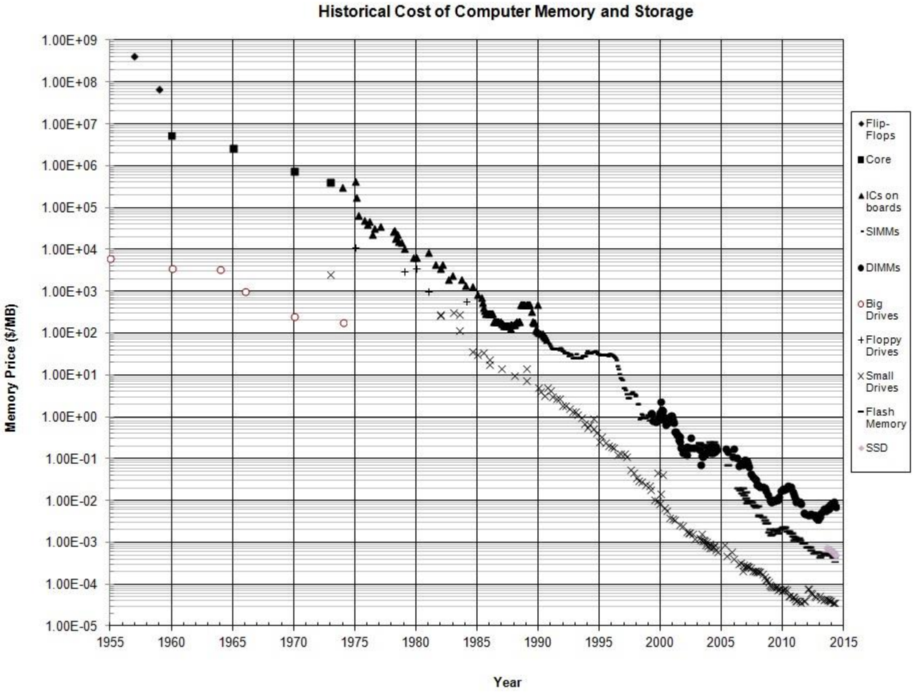
- Welcome to a new world
- Learn from history to prepare future: an introduction to Big Data
- What I do of my spare time…

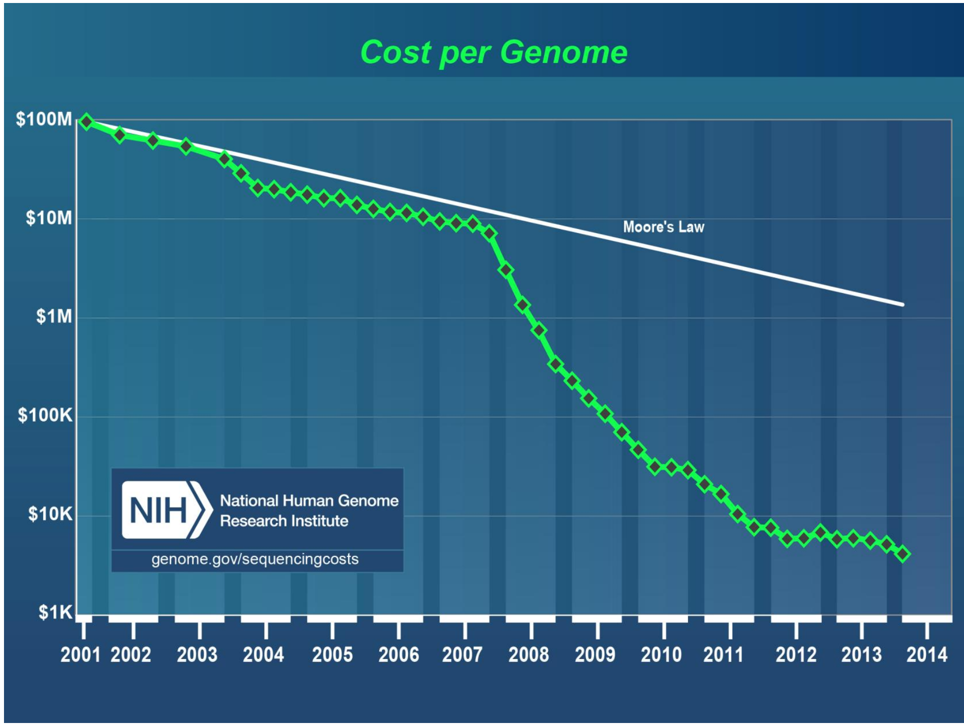
A new world beyond the standard model

- For more than 30 years, validation of the standard model
  - Electroweak physics at LEP
  - Top quark discovery at TEVATRON
  - Higgs Boson discovery at LHC
- New exploratory phase beyond the standard model
  - Where is the new physics?

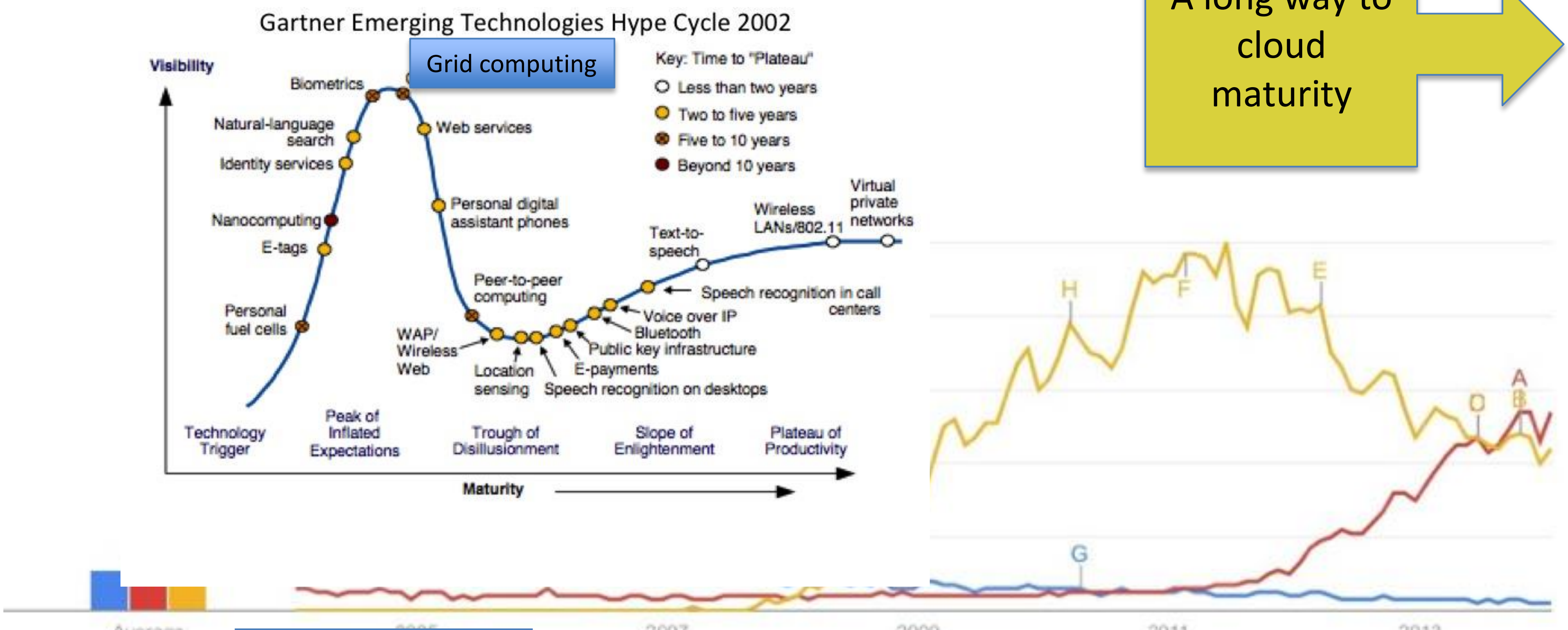
A new world without Moore’s law

- Moore’s law does not apply any more to storage capacities… nor to sequencing data production





It takes many years from hype to production quality



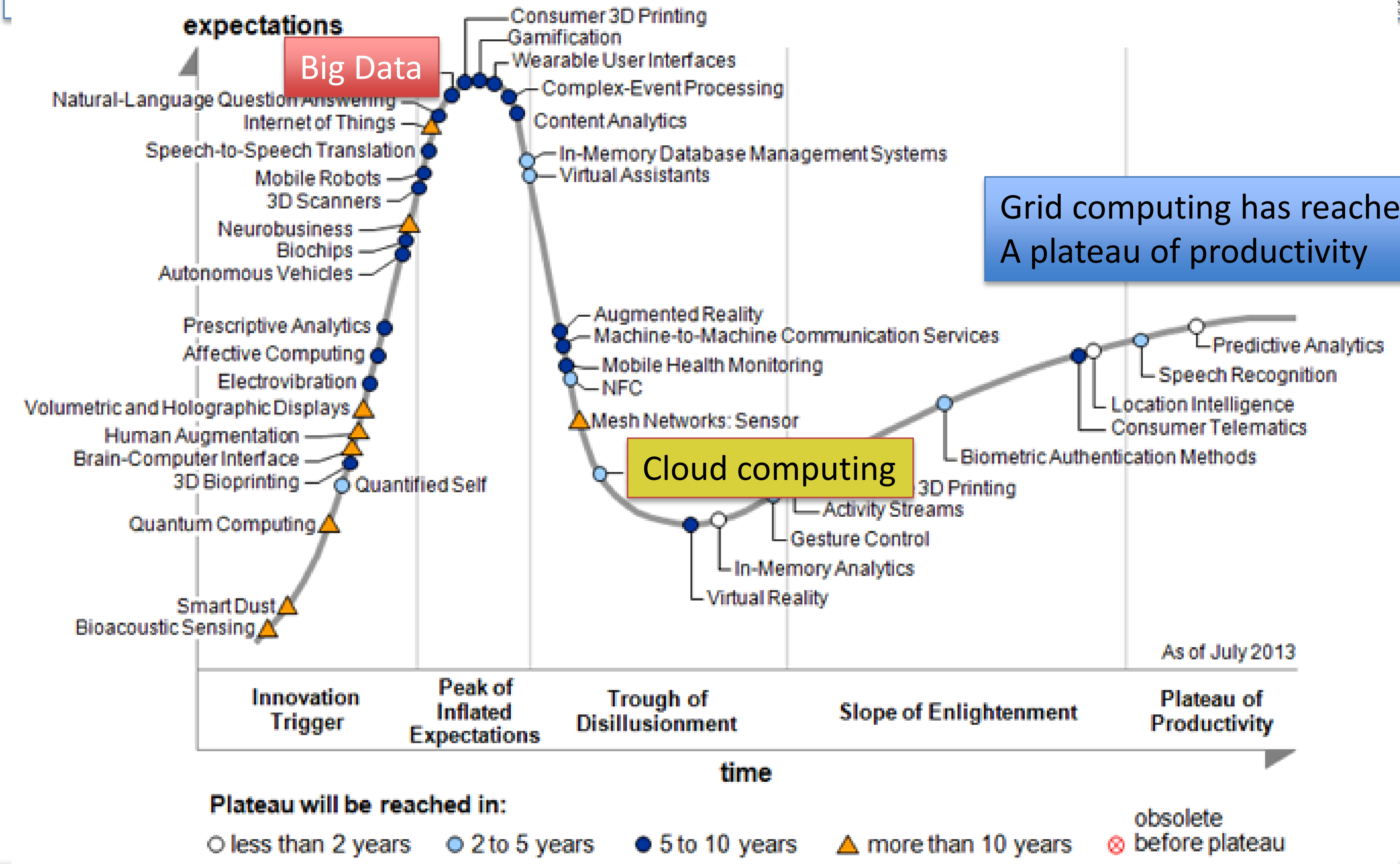
Grid computing

A long way to cloud maturity

Grid peak of expectations back in 2002

Grid maturity

Gardner hype curve for 2013



Big Data

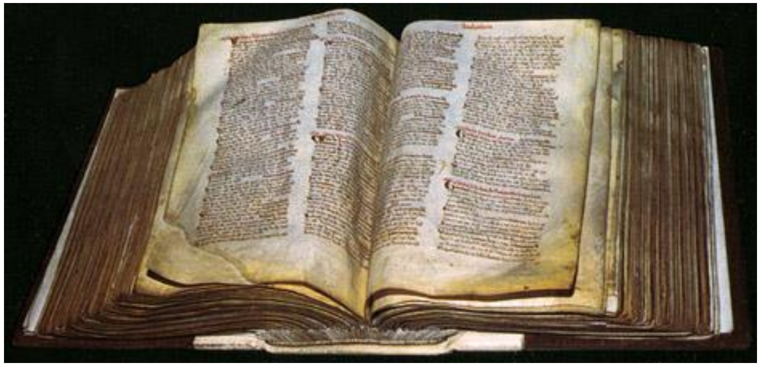
Grid computing has reached
A plateau of productivity

Cloud computing

Learning from history to build the future

- The greatest achievement of grids is not the capacity it has built
  - Obsolescence in three years for the hardware
  - Obsolescence of the grid middleware
- The greatest achievement are the human networks it has created
  - Fantastic human adventure

Learning from history: the Domesday Book (1087)



- Manuscript record of the great survey, completed in 1086 on orders of William the Conqueror

«While spending the Christmas time of 1085 in Gloucester, William had deep speech with his counsellors and sent men all over England to each shire to find out what or how much each landholder had in land and livestock, and what it was worth» *Anglo-Saxon chronicle*

- Absolute authority to define property rights since Middle Age

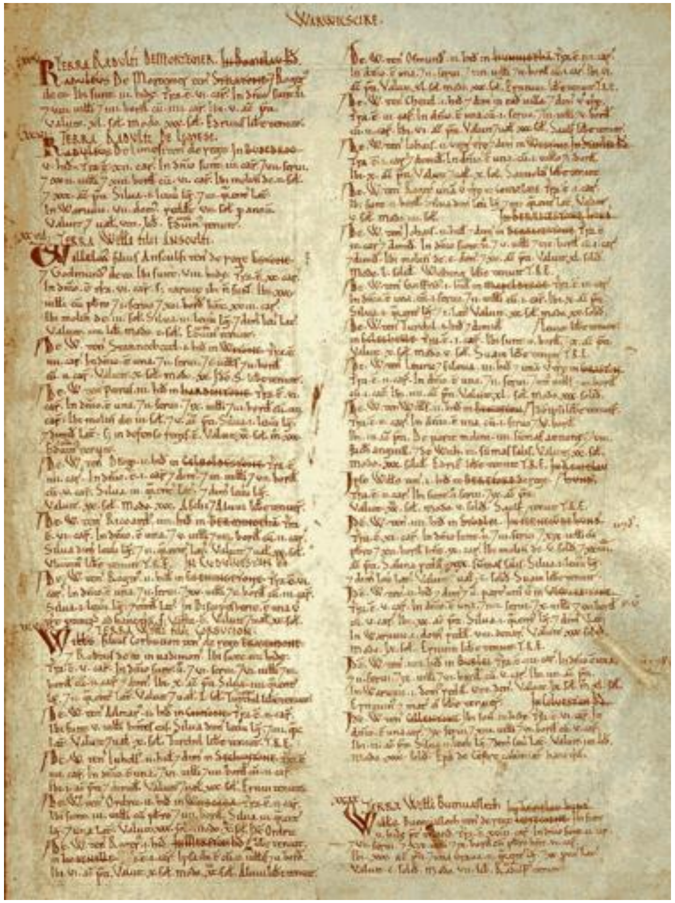
for as the sentence of that strict and terrible last account cannot be evaded by any skilful subterfuge, so when this book is appealed to ... its sentence cannot be quashed or set aside with impunity. That is why we have called the book 'the Book of Judgement' ... because its decisions, like those of the Last Judgement, are unalterable. *Richard Fitzneal, Dialogus de Scaccario, 1179*

09:21:00

Credit: Genevieve Bell, keynote talk @ SuperComputing 2013

Big data issues (I/II)

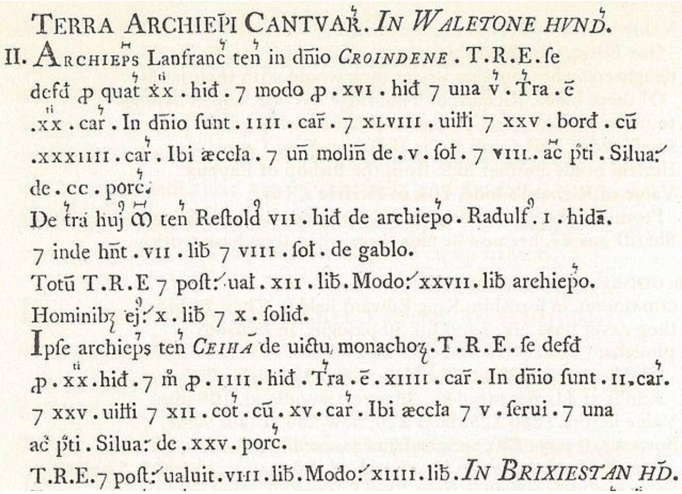
- Data collection
  - Every shire visited by a group of royal officers (1085-1086)
  - The unit of inquiry was the Hundred (a subdivision of the county)
- Data veracity
  - return for each Hundred was sworn to by twelve local jurors, half of them English and half of them Normans.
- Data analysis
  - names of the new holders of lands and assessments on which their tax was to be paid
  - national valuation list, estimating the annual value of all the land in the country



Crédit: Wikipedia

Big Data issues (II/II)

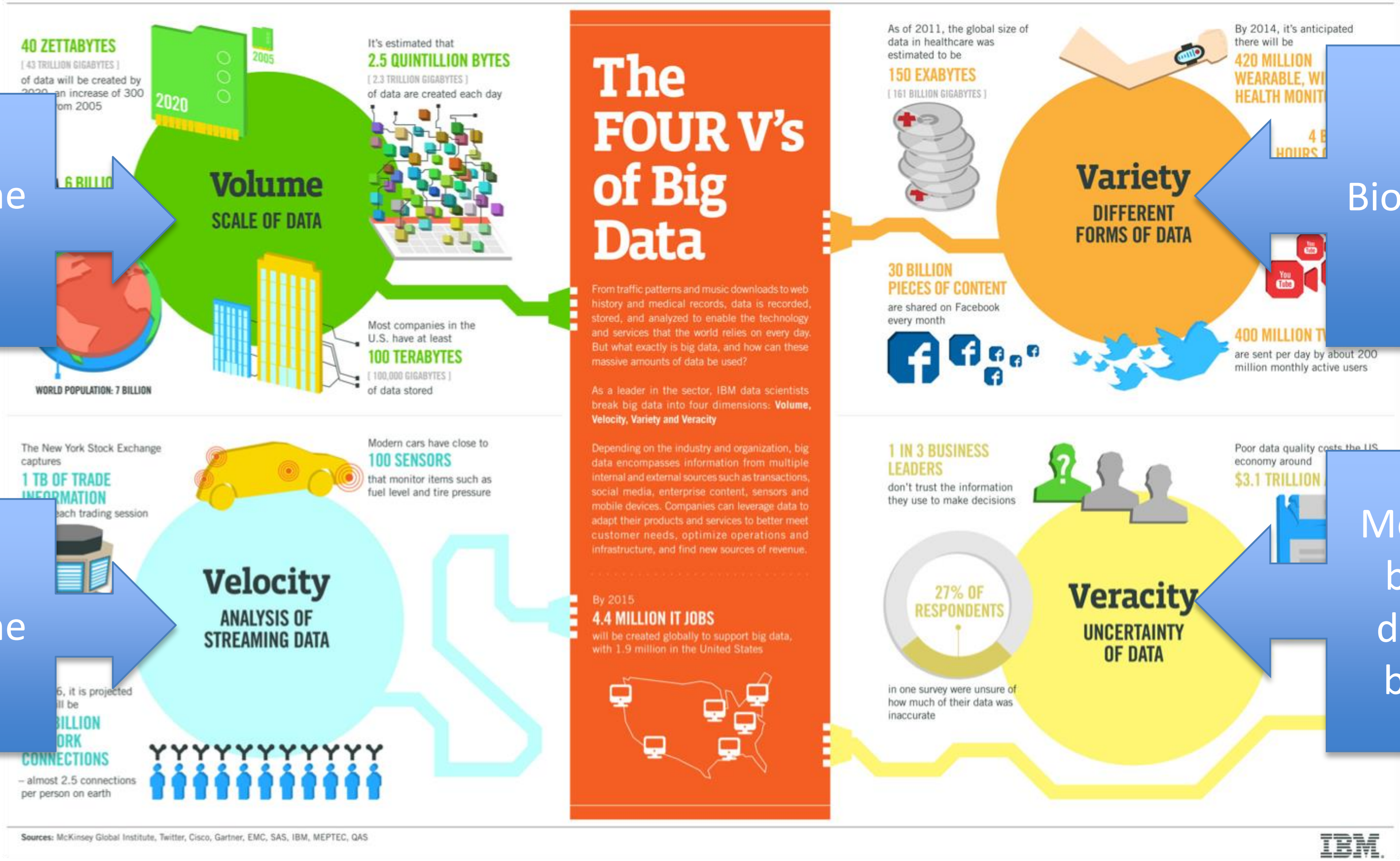
- Data presentation
  - Properties listed by fiefs
  - Properties listed by owner categories
    - king's holdings
    - holdings of churchmen and religious houses
    - Aristocrats
    - Lay men
- Data preservation
  - Preservation in the Royal Treasury in Westminster till 19th century
  - Stored at UK National Archives in Kew
  - 1986: digital version
  - 2002: access problem to digital version



TERRA ARCHIEPI CANTVAR. *IN WALETONE HVND.*

II. ARCHIEPS Lanfranc ten in dn̄io *CROINDENE* . T.R.E.ſe
defd ꝑ quat xx . hid . 7 modo ꝑ . xvi . hid 7 una v . Tra . ē
.xx . car . In dn̄io ſunt . IIII . car̄ . 7 XLVIII . uilli 7 xxv . bord . cū
.xxxIIII . car . Ibi æccla . 7 un̄ molin̄ de . v . ſol . 7 vIII . ac pti . Silua:
de . cc . porc.
De tra huj ꝯ ten Reſtold vII . hid de archiepo . Radulf . I . hidā.
7 inde hn̄t . vII . lib 7 vIII . ſol . de gablo.
Totū T.R.E 7 poſt: ual . xII . lib . Modo: xxvII . lib archiepo.
Hominibꝫ ej: x . lib 7 x . ſolid.
Ipſe archieps ten *CEIHĀ* de uictu monachoꝝ . T.R.E. ſe defd
ꝑ . xx . hid . 7 m̊ ꝑ . IIII . hid . Tra . ē . xIIII . car̄ . In dn̄io ſunt . II . car.
7 xxv . uilli 7 xII . cot . cū . xv . car . Ibi æccla 7 v . ſerui . 7 una
ac pti . Silua: de . xxv . porc.
T.R.E. 7 poſt: ualuit . vIII . lib . Modo: xIIII . lib . *IN BRIXIESTAN HD.*

Big Data 4 Vs



40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION

WORLD POPULATION: 7 BILLION

Volume
SCALE OF DATA

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
each trading session

Velocity
ANALYSIS OF STREAMING DATA

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

6, it is projected
ll be
BILLION
ORK
CONNECTIONS
– almost 2.5 connections per person on earth

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

Variety
DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WI
HEALTH MONIT

4 B
HOURS

400 MILLION T
are sent per day by about 200 million monthly active users

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Veracity
UNCERTAINTY OF DATA

Poor data quality costs the US economy around
$3.1 TRILLION

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

IBM.

LHC experiments

Biodiversity

LHC experiments

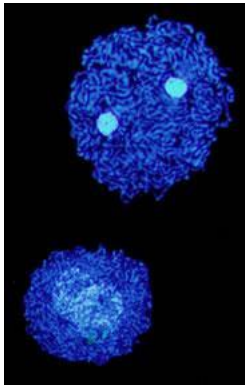
Molecular biology data are bugged

Data volumes: the example of metagenomics

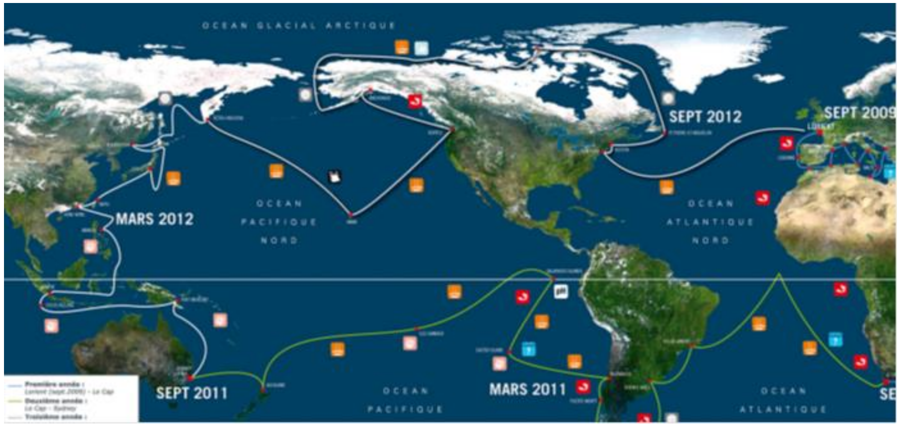
Metagenomics is the study of genetic material recovered directly from environmental samples.

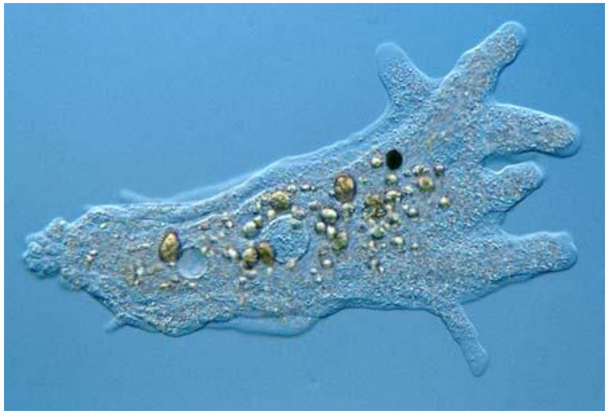
Evolution of sequencing techniques

Sanger technology	500 base pairs (bp)
454 technology	10^5 400-600 bp reads
Illumina Technology	10^6 100 bp reads
TARA project	10^7 100-400 bp reads



Smallest non viral genome: *Carsonella ruddii* (0,16Mbp)

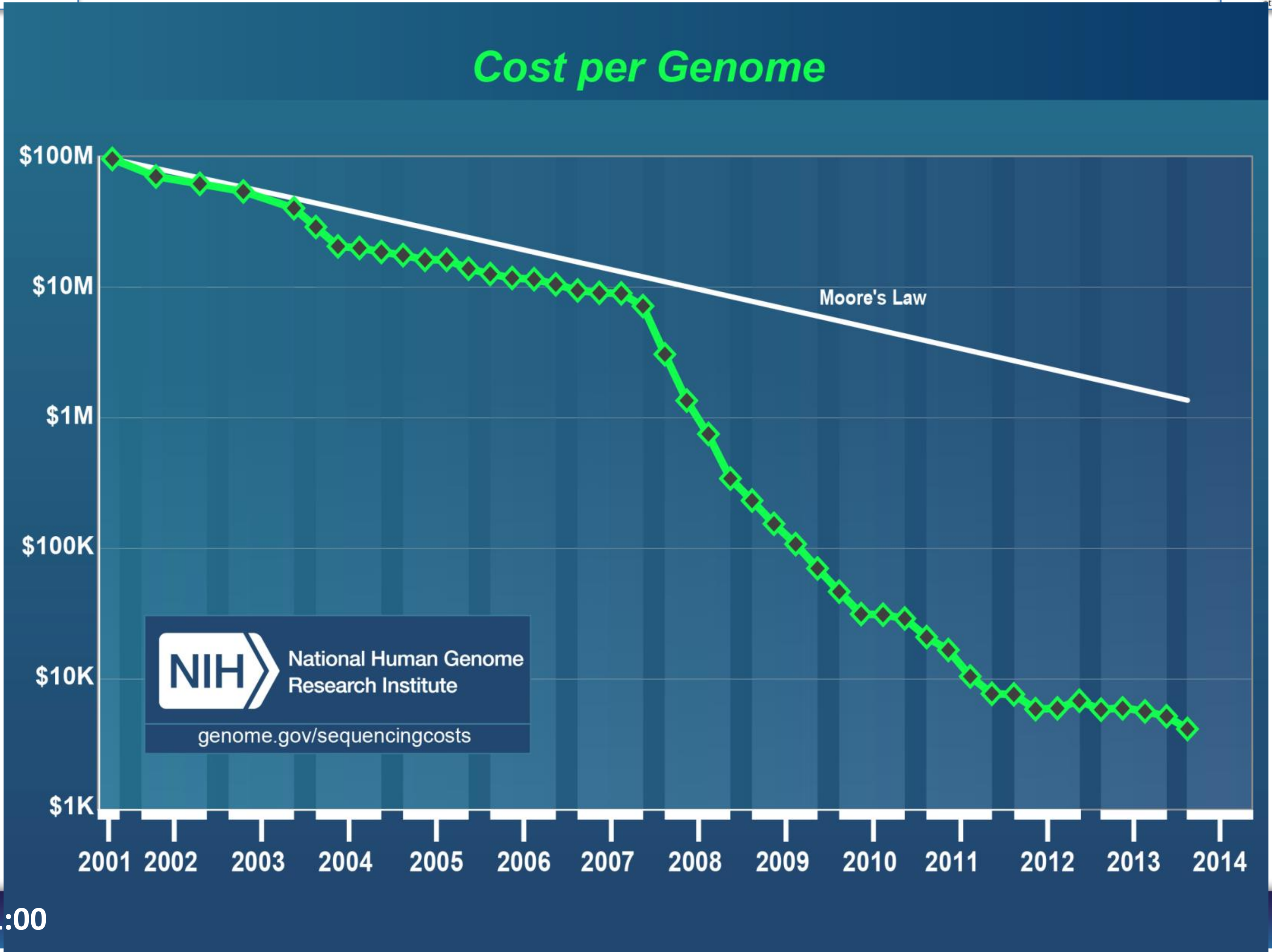




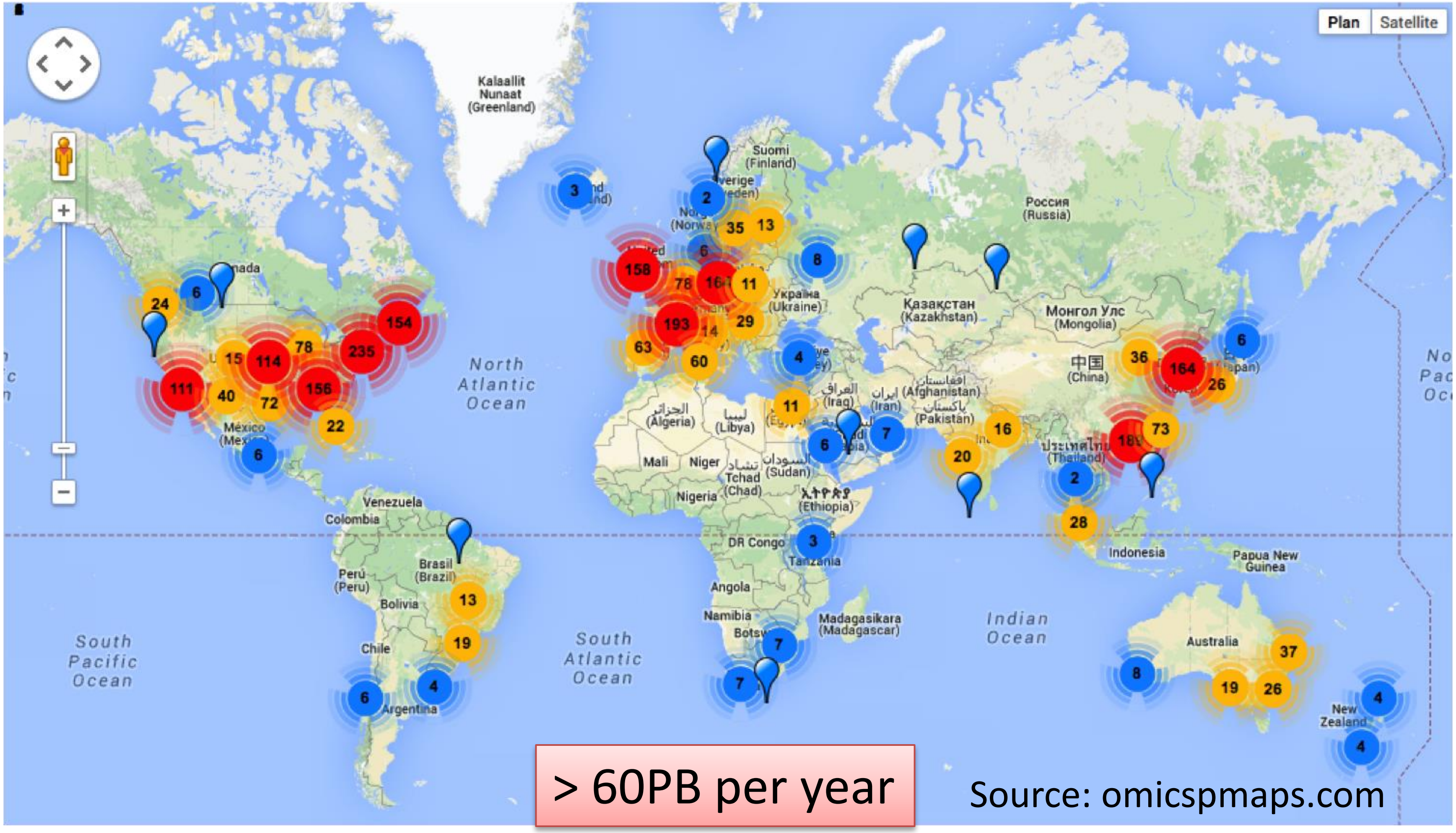
Largest genome: *Polychaos dubium* (670Gbp)

Tara @ http://oceans.taraexpeditions.org/

Cost per Genome is decreasing faster than Moore's law



Consequence: over 2500 Next Generation Sequencing machines in 900+ research centers in the world



> 60PB per year

Source: omicspmaps.com

Welcome to Auvergne, at the heart of France

1,35 Million inhabitants
26013 km^2

AUVERGNE
la région juste et grande

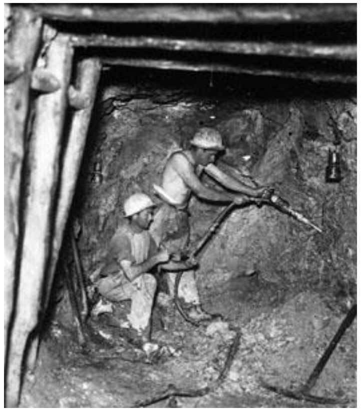
09:21:00

Auvergne at the heart of Uranium production in France

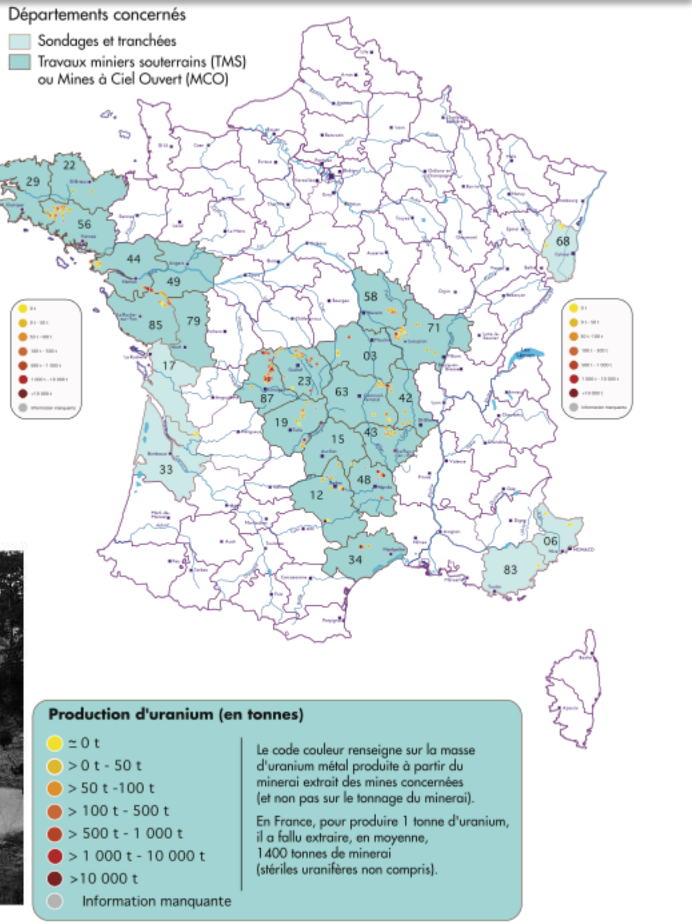
1949: first attempt to extract uranium ore in France in Lachaux (Auvergne)

In 50 years:

- 53 Million tons extracted in France till 2001
- 76000 tons of uranium ore produced in > 200 mines



Map of uranium mines in metropolitan France



Le code couleur renseigne sur la masse d'uranium métal produite à partir du minerai extrait des mines concernées (et non pas sur le tonnage du minerai).

En France, pour produire 1 tonne d'uranium, il a fallu extraire, en moyenne, 1400 tonnes de minerai (stériles uranifères non compris).

ZATU, a Long Term Ecological Research dedicated to life under natural ionizing radiation

Natural radioactivity

Storage sites of uranium ore extraction residues

- Society in uranium rich territories
  - Social impact of uranium extraction
  - Preserving the long term memory
- Characterization, behavior and transfer of radionucleids
  - long term future of radionucleids in storage sites
- Impact of radiation on living systems
  - Multigenerational effects of chronic exposure to radiation

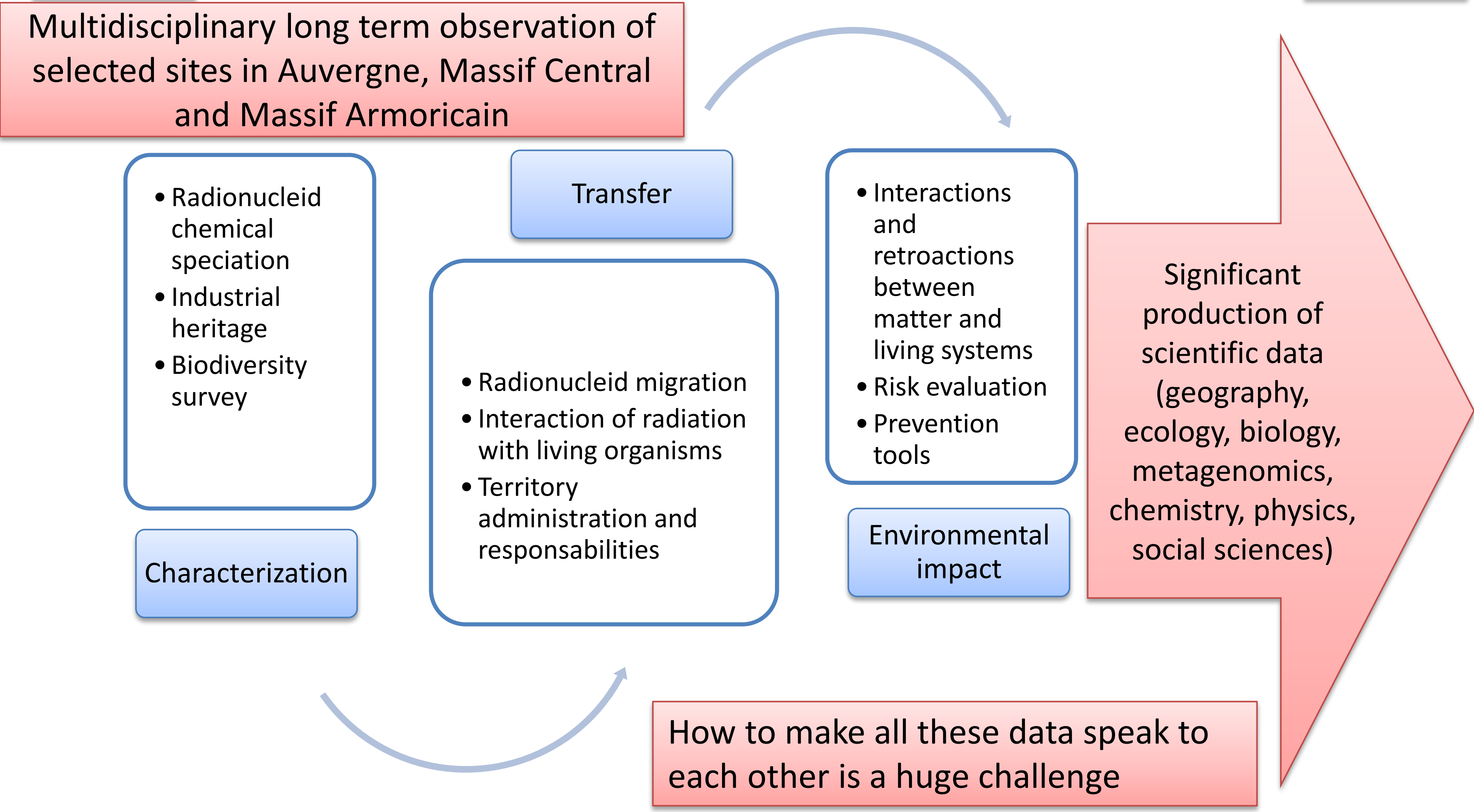
Impact of chronic exposure to low dose ionizing radiation on living organisms

- From the Chernobyl environment, a coherent picture of predictable radiation-induced effects for low-dose-rate exposures has not emerged
  - Contradictory experimental evidences from Chernobyl exclusion zone
- Need to collect more data from Chernobyl exclusion zone but also from other ecosystems under chronic low dose exposure
  - Radioactive water sources
- Point 0: what happens in “total” absence of radioactivity?

Photographs of abnormalities in barn swallows. (a) Normal phenotype. (b–d) Partially albinistic plumage. (e) and (f) Deformed beak. (g) Deformed air sacks. (h) and (i) Bent tail feathers.

Proasellus cavaticus

ZATU strategy



Multidisciplinary long term observation of selected sites in Auvergne, Massif Central and Massif Armoricain

- Radionucleid chemical speciation
- Industrial heritage
- Biodiversity survey

Characterization

Transfer

- Radionucleid migration
- Interaction of radiation with living organisms
- Territory administration and responsabilities

- Interactions and retroactions between matter and living systems
- Risk evaluation
- Prevention tools

Environmental impact

Significant production of scientific data (geography, ecology, biology, metagenomics, chemistry, physics, social sciences)

How to make all these data speak to each other is a huge challenge

Conclusion

- Grid computing has allowed building a truly multidisciplinary distributed IT infrastructure
  - Greatest achievement: human networks
- Cloud computing allows extending the grid functionalities
  - All sciences will benefit even more
  - Still a long way to the plateau of maturity
  - Scientific gateways and pilot agent platforms allow a smooth transition from grids to clouds
- Big Data is the next frontier
  - Volume will not be necessarily the most difficult challenge

Which data produced today will still be used in 900 years?