BIST Founding Conference

Growing together, advancing science

BIST Barcelona Institute of Science and Technology

BIG DATA IN BIOINFORMATICS

MARCH 31, 2017

CCCB | Barcelona

BIST Founding Conference

Growing together, advancing science

BIST Barcelona Institute of Science and Technology

BIG DATA IS BIOINFORMATICS

MARCH 31, 2017

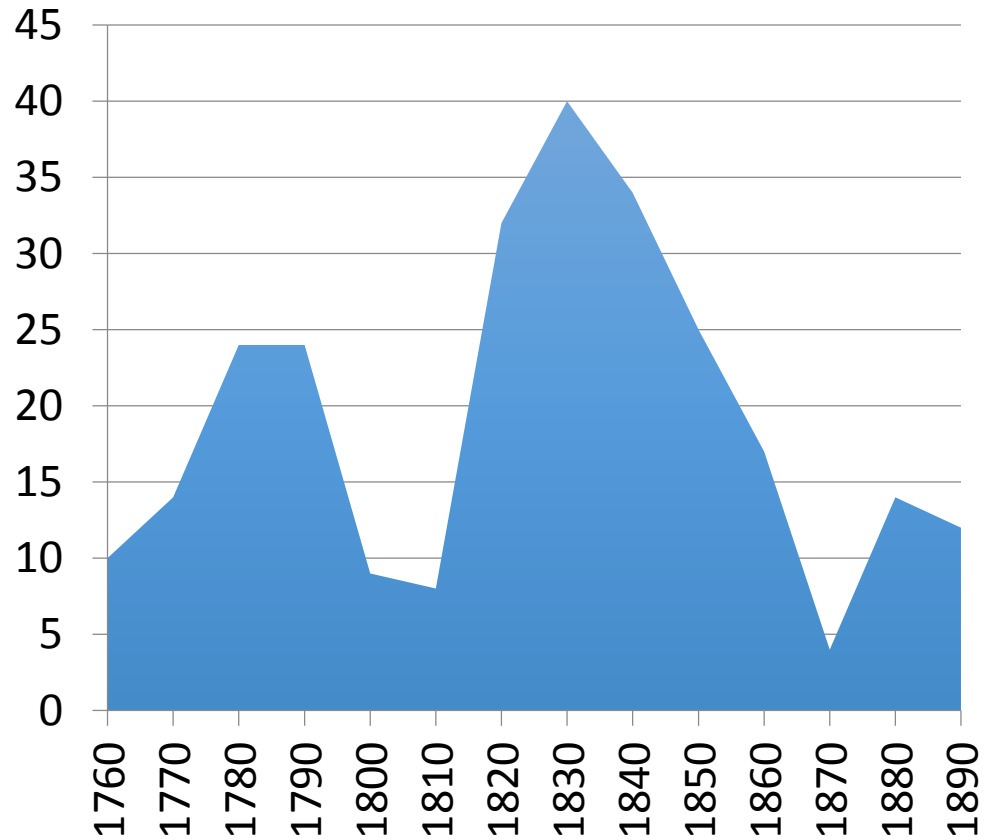
CCCB | Barcelona

BiG data is bioinformatics

- Heterogeneous data
  - numerical
  - non-numerical
    - Structures at different levels (from molecules to organisms)—images
    - Sequences
    - Longitudinal/dynamic—movies
  - Multi-dimensional
- Collected at multiple sites
  - Produced by indivual small labs to large international consortiums
- Shared through the internet
  - Real time acces
- Need of integrative analysis

Google search: X-informatics (june 4,2015)

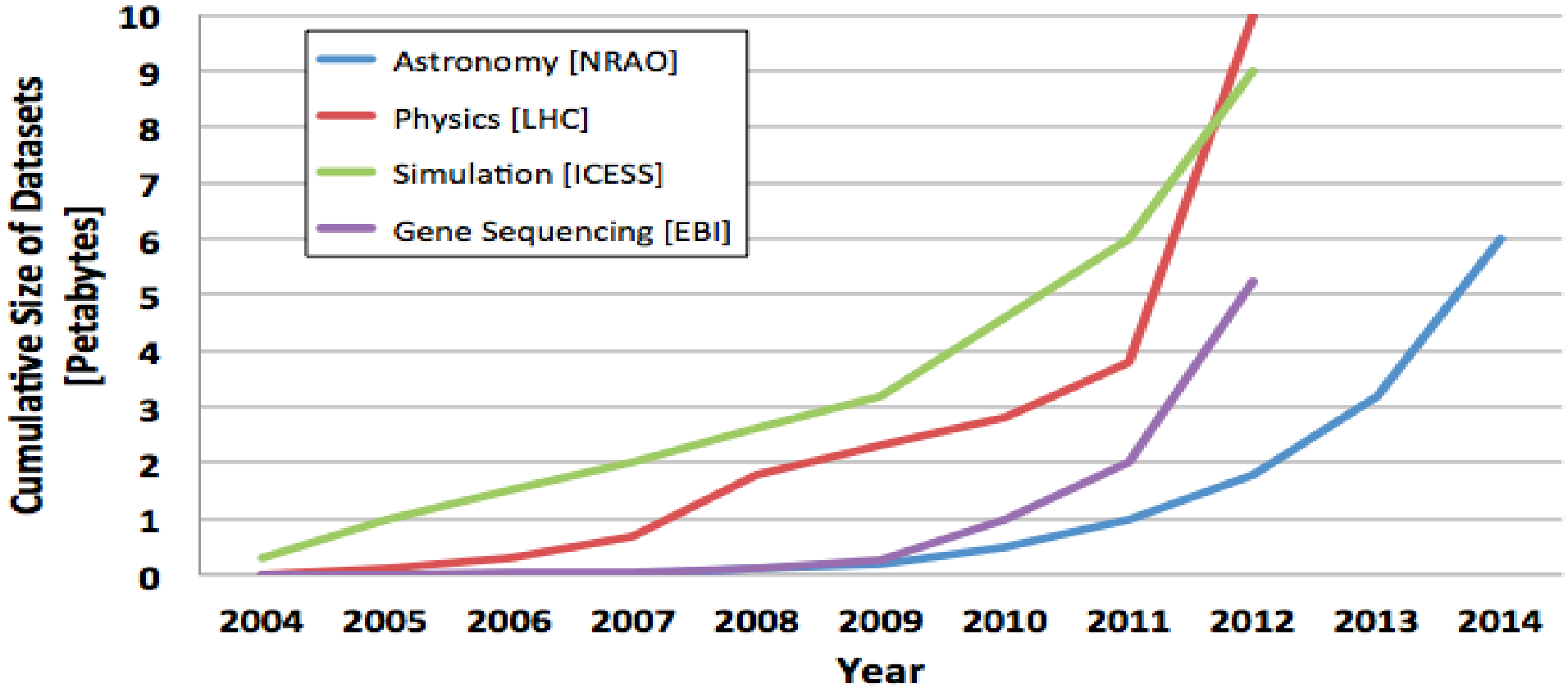
bioinformatics	24,500,000
chemoinformatics	275,000
astroinformatics	27,800
neuroinformatics	331,000
socioinformatics	14,100
geoinformatics	548,000
meteoinformatics	146
econoinformatics	2,010
ecoinformatics	92,800
physicoinformatics	5,390

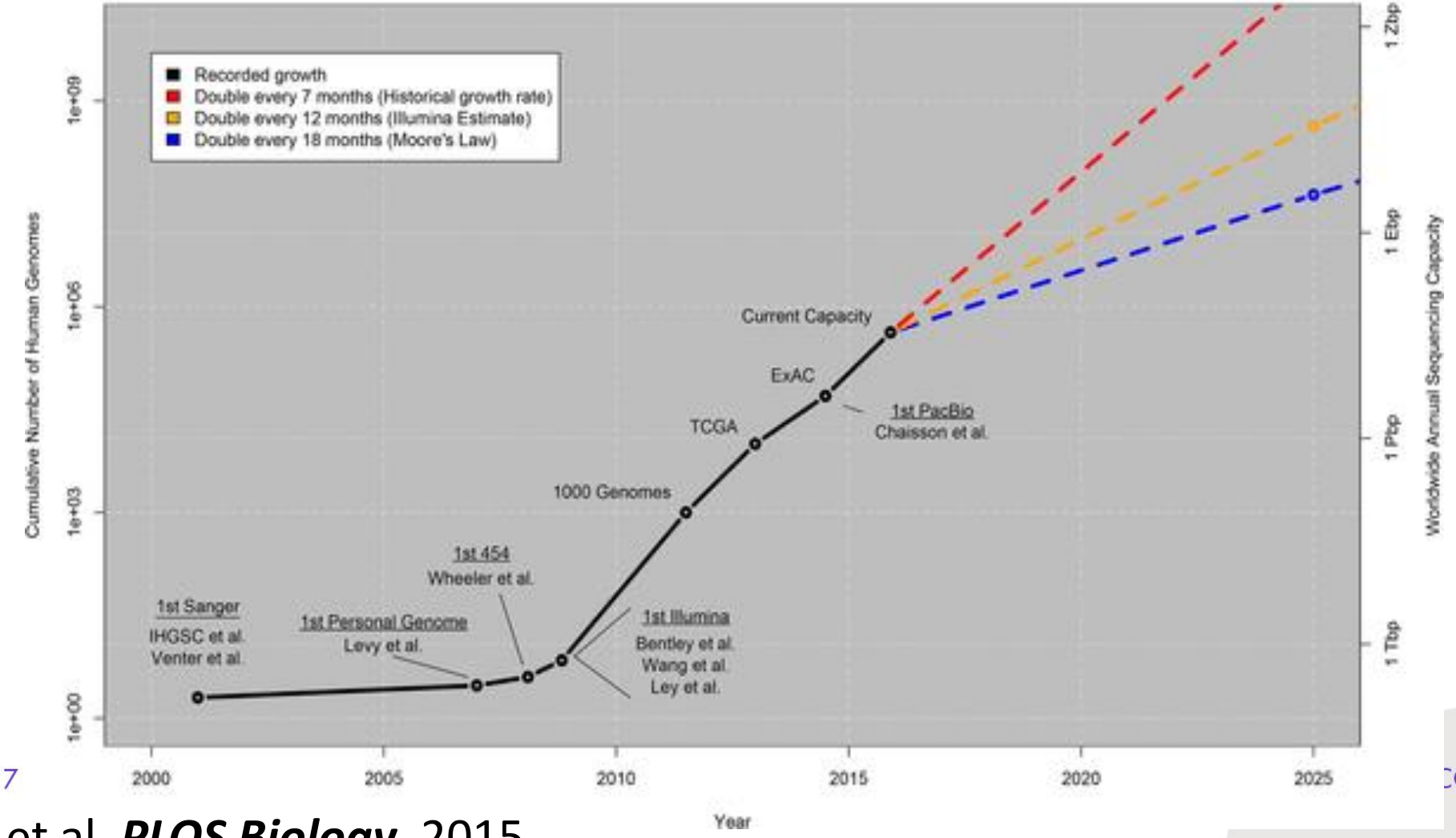


Number of scientific expeditions

Cedric Notredame, CRG

Cedric Notredame, CRG


Cumulative Size of Datasets [Petabytes]
10
9
8
7
6
5
4
3
2
1
0
Astronomy [NRAO]
Physics [LHC]
Simulation [ICESS]
Gene Sequencing [EBI]
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year



Stephens ZD et al. ***PLOS Biology***, 2015

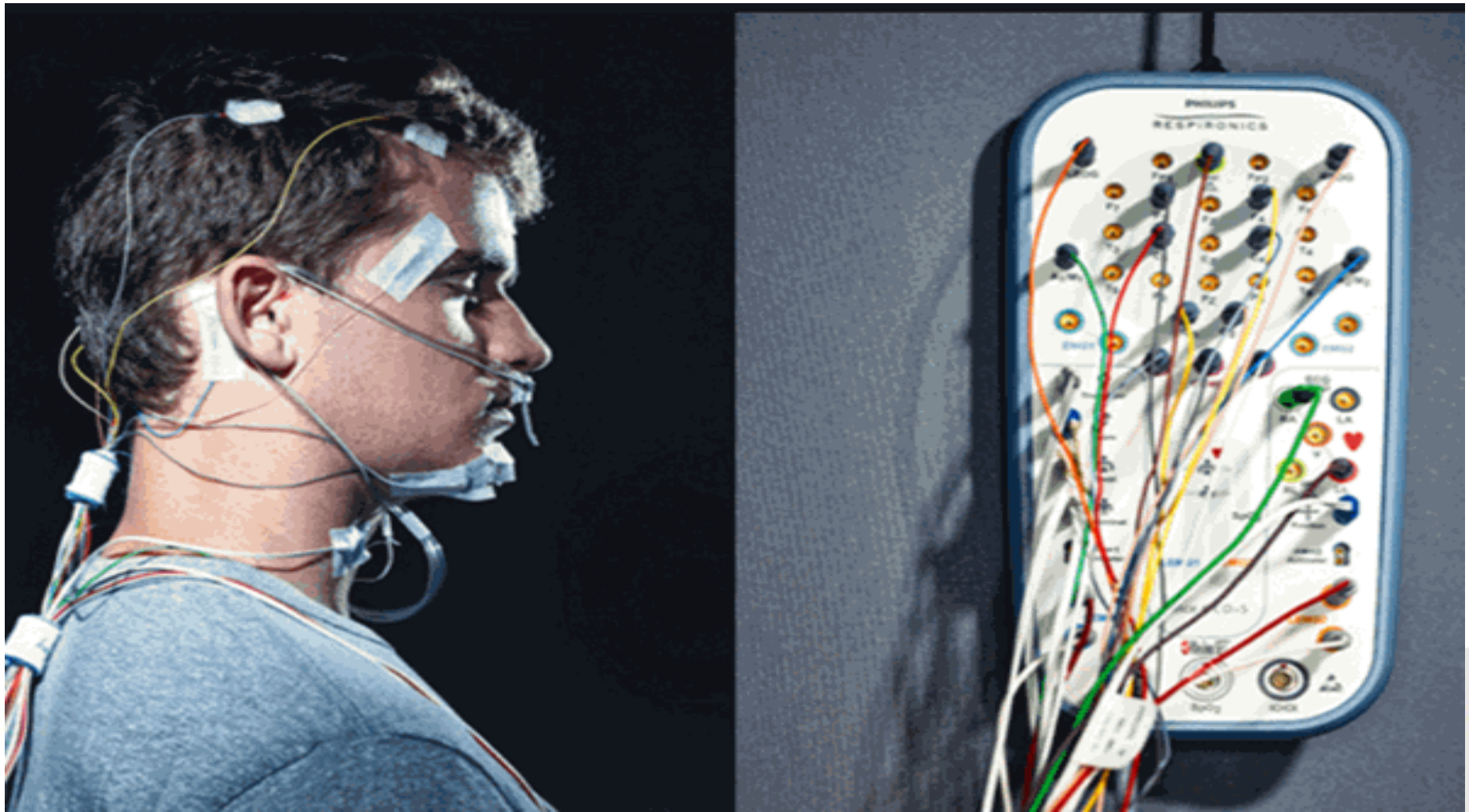
Big Data: Astronomical or Genomical?

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

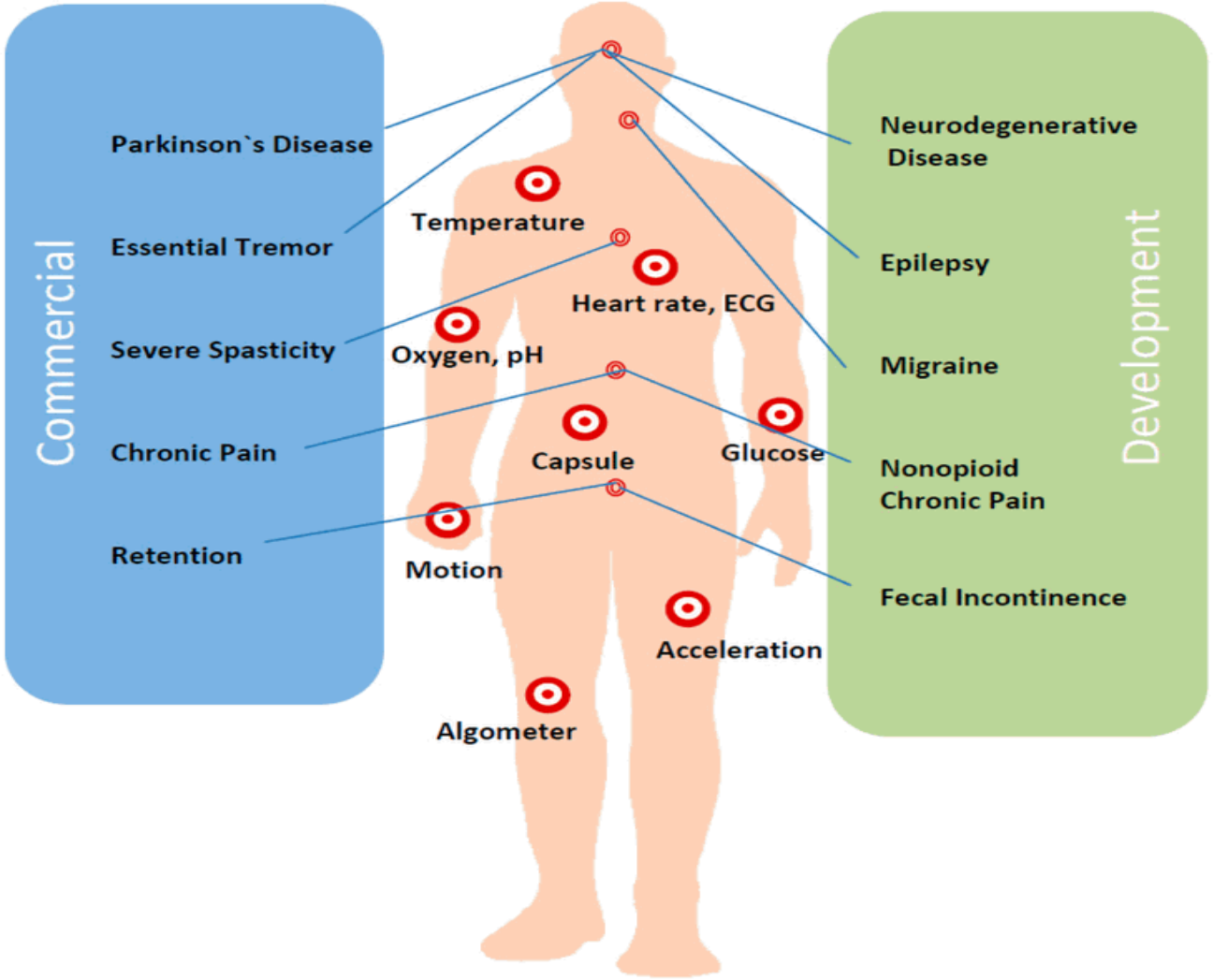
Table 1. Four domains of Big Data in 2025

PLOS BIOLOGY

We are the Big Data

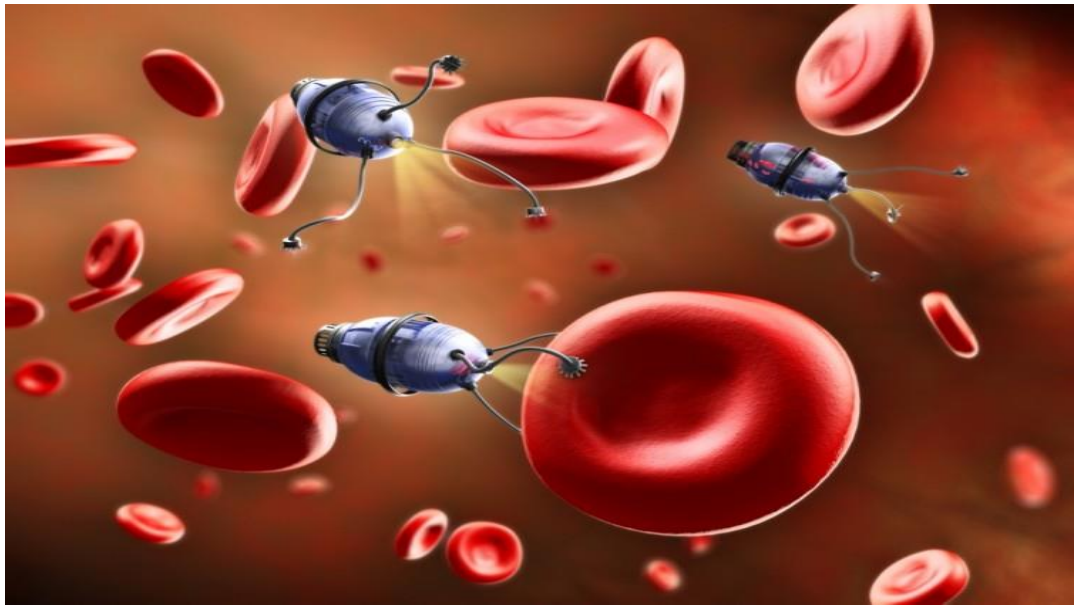


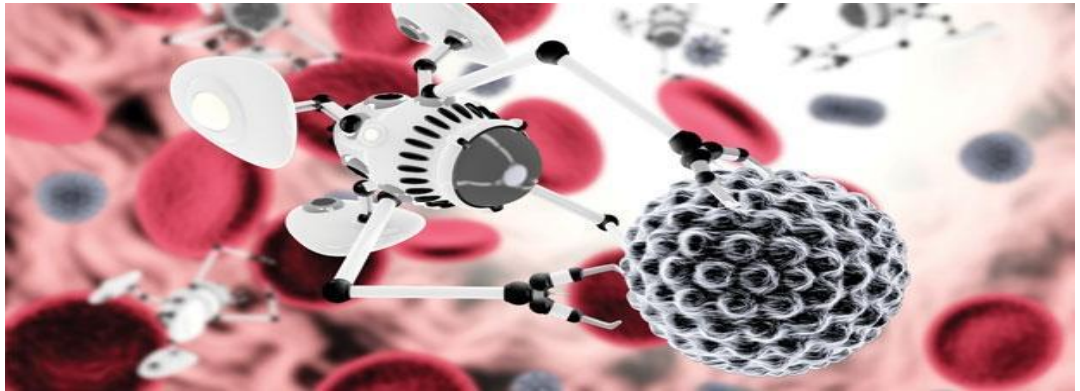
Wearable medical devices

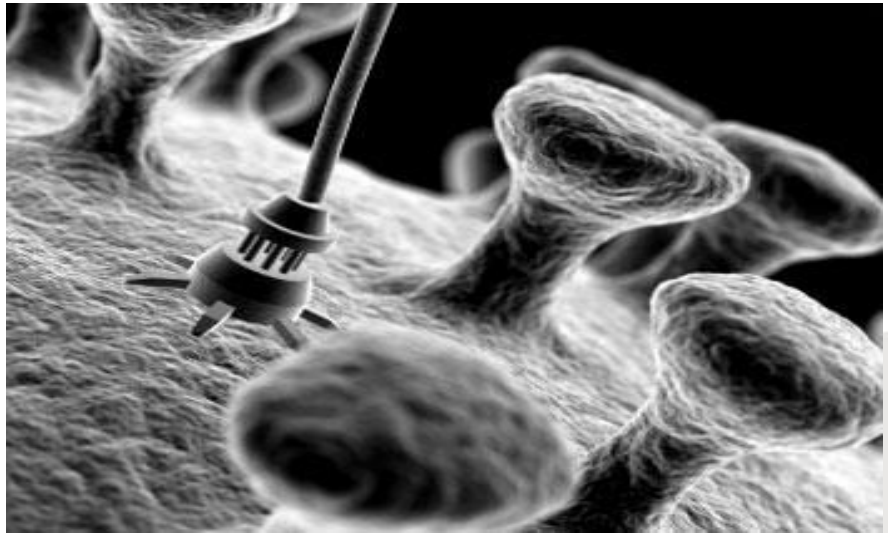


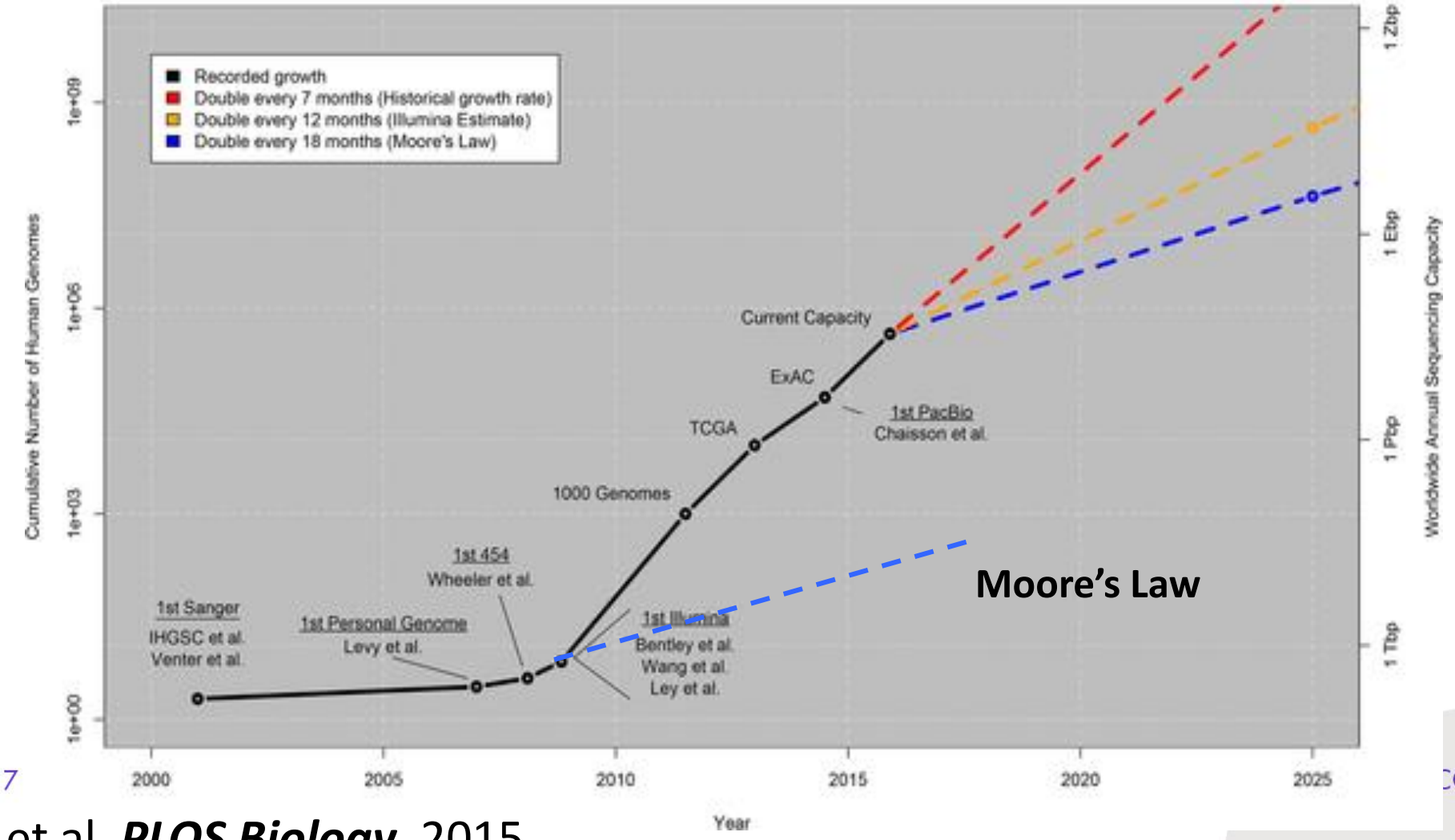
Implantable wearable devices

Nanowearables









Stephens ZD et al. ***PLOS Biology***, 2015

Molecular Computation of Solutions to Combinatorial Problems

Leonard M. Adleman

The tools of molecular biology were used to solve an instance of the directed Hamiltonian path problem. A small graph was encoded in molecules of DNA, and the ''operations'' of the computation were performed with standard protocols and enzymes. This experiment demonstrates the feasibility of carrying out computations at the molecular level.

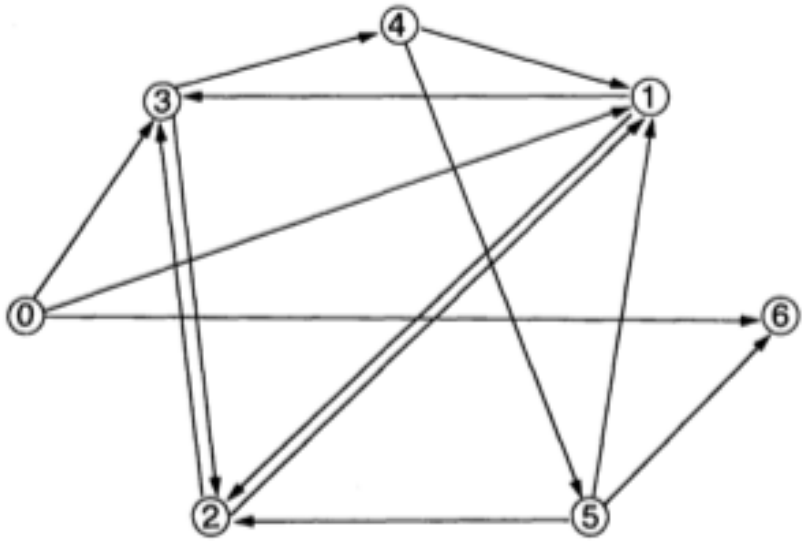
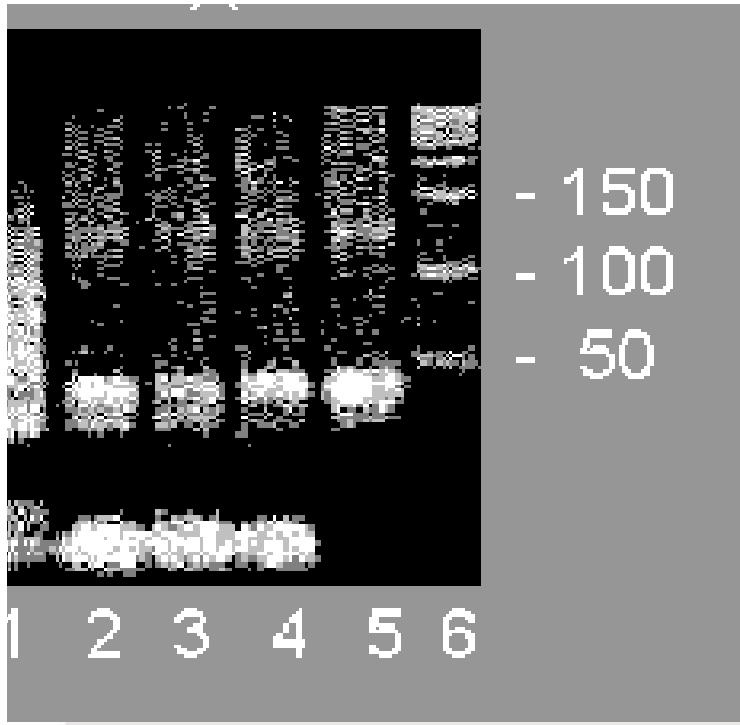


Fig. 1. Directed graph. When $v_{in} = 0$ and $v_{out} = 6$, a unique Hamiltonian path exists: 0→1, 1→2, 2→3, 3→4, 4→5, 5→6.

Fig. 2. Encoding a graph in DNA. For each vertex *i* in the graph, a random 20-mer oligonucleotide O_i is

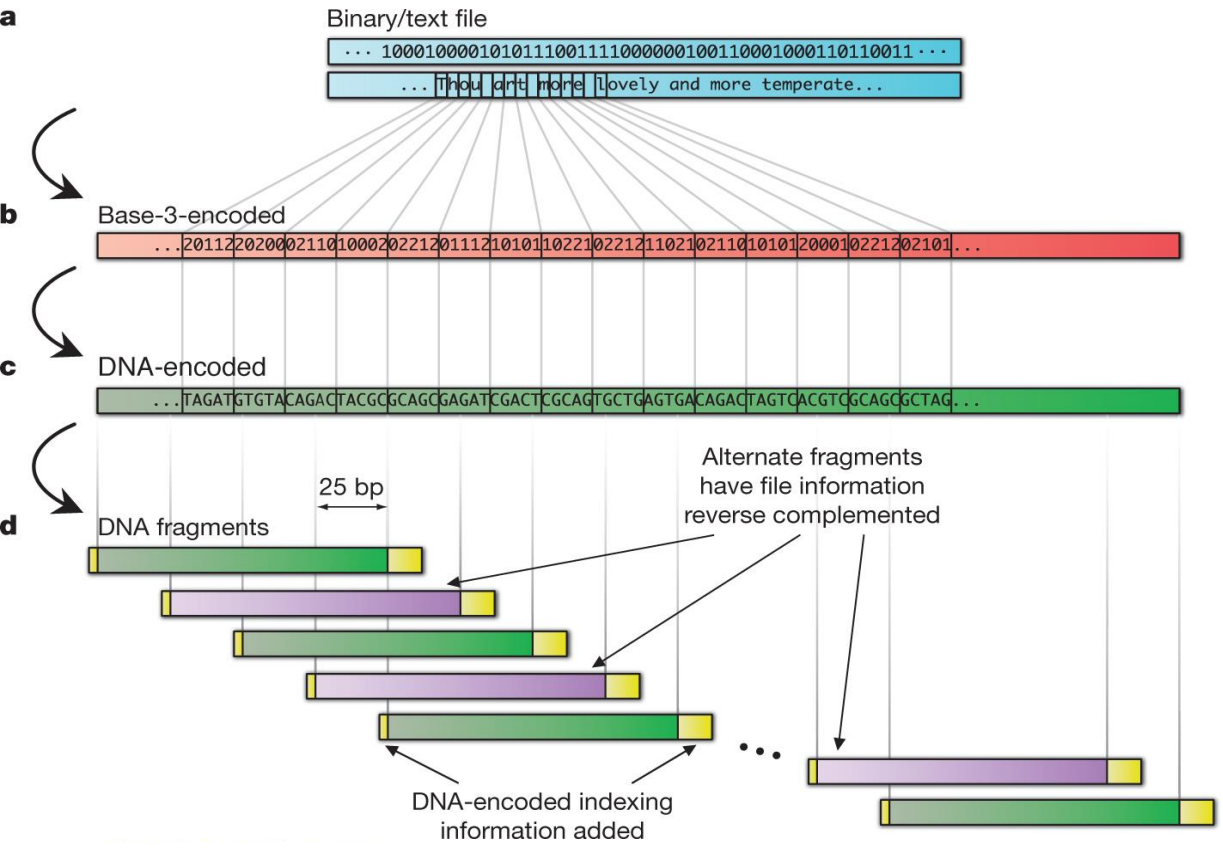


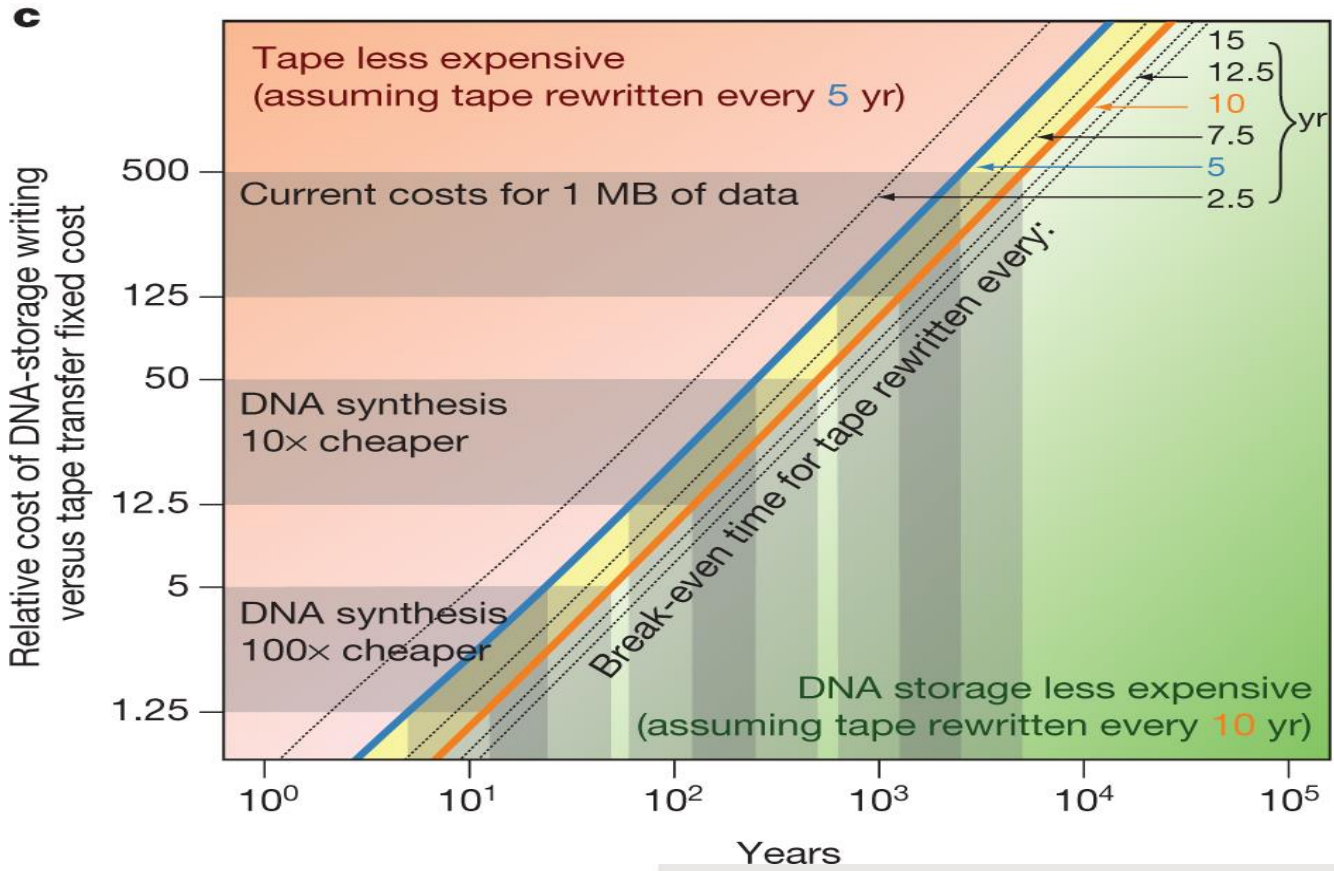
doi:10.1038/nature11875

Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

Nick Goldman[1], Paul Bertone[1], Siyuan Chen[2], Christophe Dessimoz[1], Emily M. LeProust[2], Botond Sipos[1] & Ewan Birney[1]

2 PB per 1 g DNA





“Goldman prediction”

- 2PB per 1g DNA (2 x 10^{15} bytes)
- Total world info (2013): 3ZBytes (3 x 10^{21} bytes)
  - Aproximately 1.5 x 10^{6} g (1,5 tonnes of DNA) to store all information
- Information doubling time: 2 years
- Mass of earth: 6 x 10^{27} g (google)
- 1.5 x 10^{6} x $2^{x/2}$ ≈ 6 x 10^{27} ➔ x ≈ 140 years
- the mass of total info in the world stored in DNA exceeds the mass of the Earth in year 2157

REPORT

DNA STORAGE

DNA Fountain enables a robust and efficient storage architecture

215 PB per 1 g DNA

Yaniv Erlich[1,2,3]* and Dina Zielinski[1]

DNA is an attractive medium to store digital information. Here we report a storage strategy, called DNA Fountain, that is highly robust and approaches the information capacity per nucleotide. Using our approach, we stored a full computer operating system, movie, and other files with a total of 2.14×10^6 bytes in DNA oligonucleotides and perfectly retrieved the information from a sequencing coverage equivalent to a single tile of Illumina sequencing. We also tested a process that can allow 2.18×10^{15} retrievals using the original DNA sample and were able to perfectly decode the data. Finally, we explored the limit of our architecture in terms of bytes per molecule and obtained a perfect retrieval from a density of 215 petabytes per gram of DNA, orders of magnitude higher than previous reports.