ACADÉMIE DES TECHNOLOGIES
POUR UN PROGRÈS RAISONNÉ CHOISI PARTAGÉ

NATIONAL ACADEMY OF TECHNOLOGIES OF FRANCE
SHARING A REASONED, CHOSEN PROGRESS

Archiving Big Data on DNA

François KÉPÈS

French Academy of Agriculture

National Academy of Technologies of France (NATF)

NATF Workgroup "DNA: reading, writing, storing information"

The current model: DataCenters

The Global DataSphere:
2018 ⇔ 33,000 times this center
2040 ⇔ ≥ 15,000,000 times this center

In the 2010's :

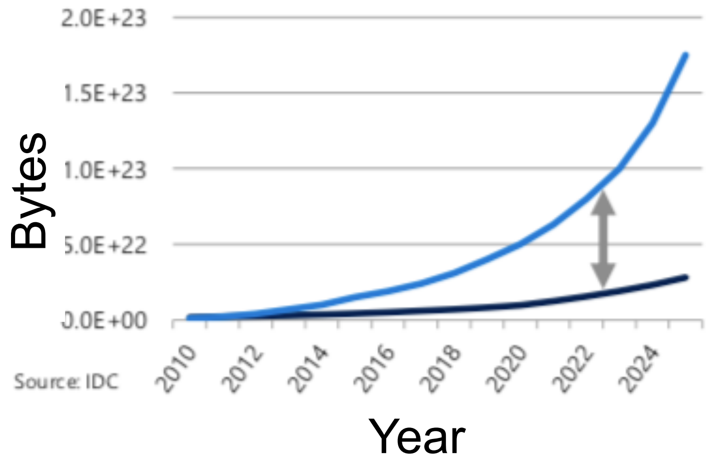
8,6 million datacenters
170 million m^2 (1/1.000.000 ^{eme})

Annual Consumption:
1,000 TWh electricity
1,000 megatons eq-CO_2 emitted
450 billion € invested.

Shortages for DataCenters

Electronic-grade Silicium :
24,000 tons produced (1%)
2,400,000 tons required within 20 years

Data storage capacity:
- overrun by quantity



Data Center of 1 Eb (10^{18} bytes)

Global DataSphere:

2018 ⇔ 66 grams
2040 ⇔ 10,000 grams

vs.

DNA

the dot here in the circle

Archiving on DNA

- **Advantages**
- **Principle**
- **History**

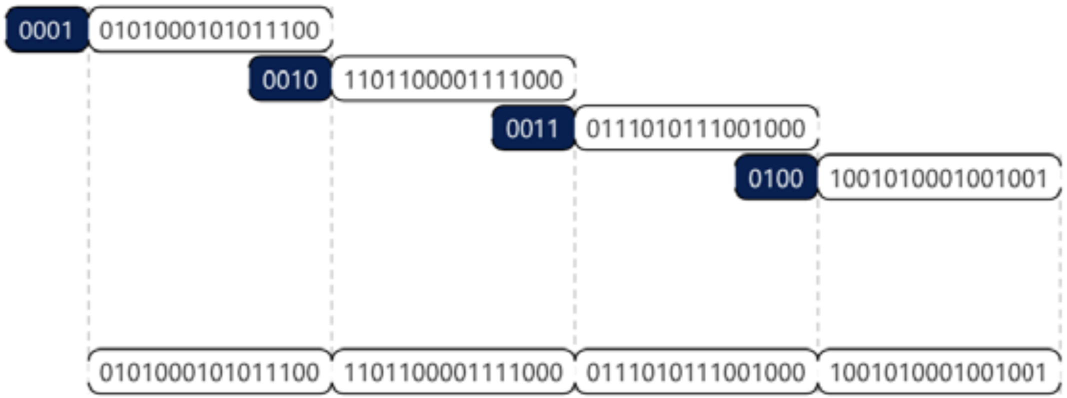
Archiving on DNA: DENSITY

Global DataSphere:

2018 ⇔ 66 grams
2040 ⇔ 10,000 grams

In practice, much more DNA:

- millions of identical copies
- signals for addressing, indexation, quality control
- macroscopic container



Archiving on DNA: DENSITY

Global DataSphere:

2018 ⇔ a van
2040 ⇔ a truck

In practice, much more DNA:

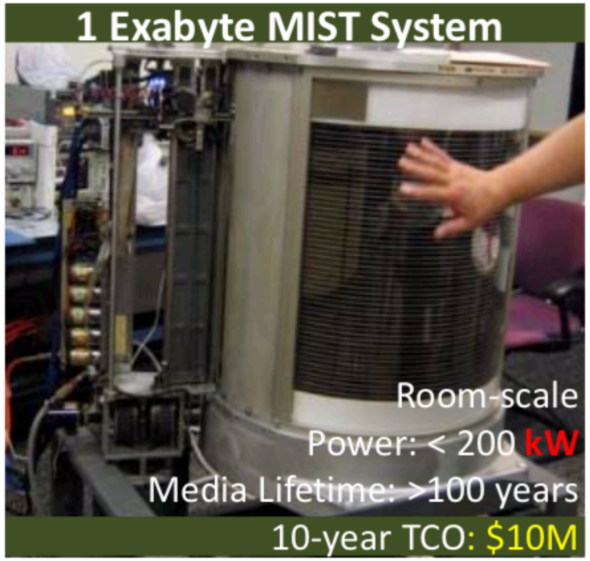
- millions of identical copies
- signals for addressing, indexation, quality control
- macroscopic container

Gain by a factor of **10 million**

Archiving on DNA: DENSITY

1 Eb (10^{18} bytes)

→



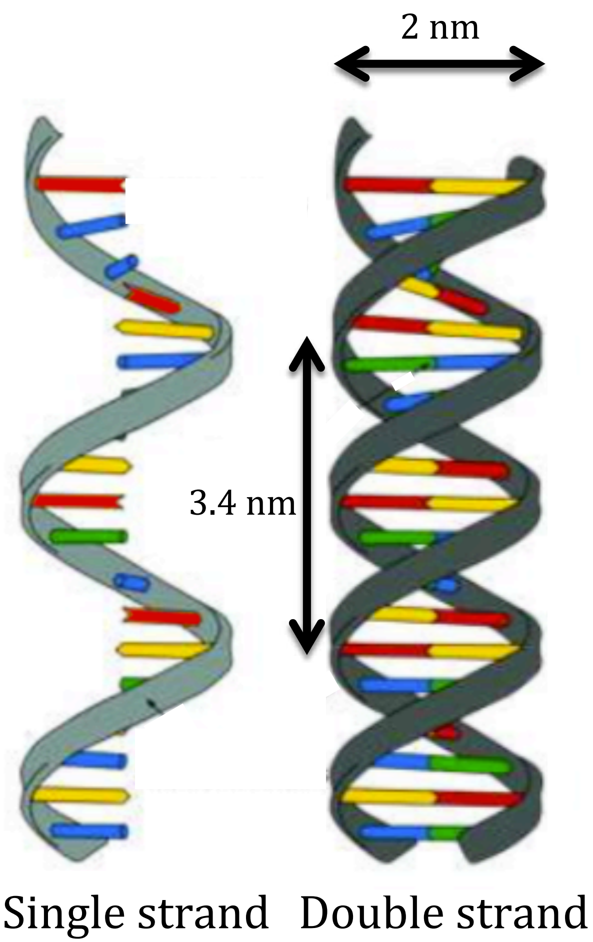
TCO = Total Cost of Ownership

1 bit ⇔ 50 atoms

DNA: a quick reminder

1 bit ⇔ 50 atoms

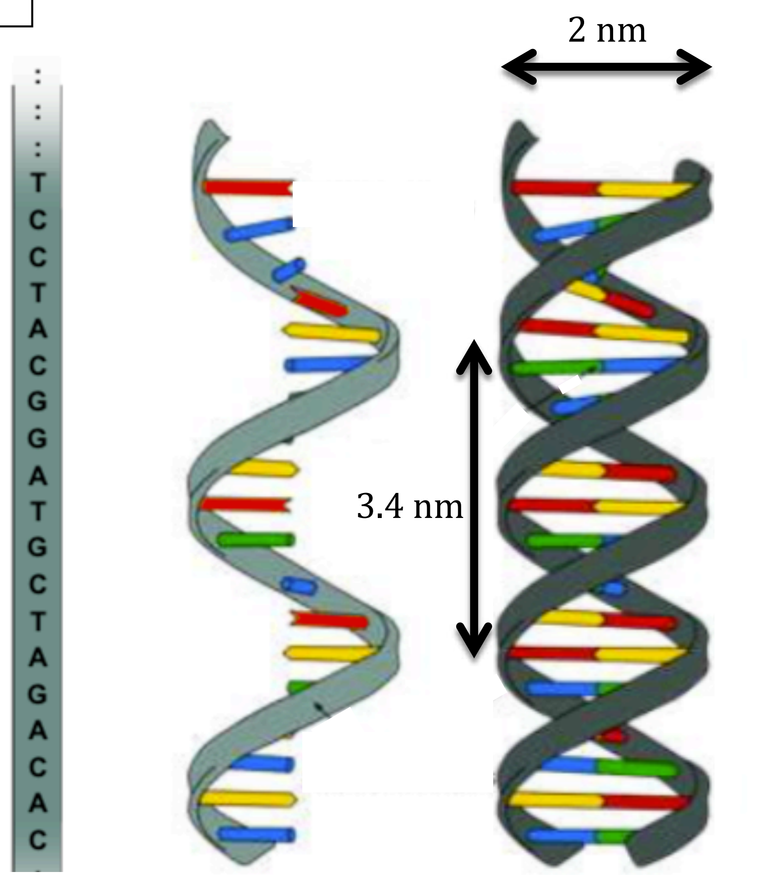
DNA



DNA: a quick reminder

DNA Sequence: '…TCCTACGGAT …'

DNA

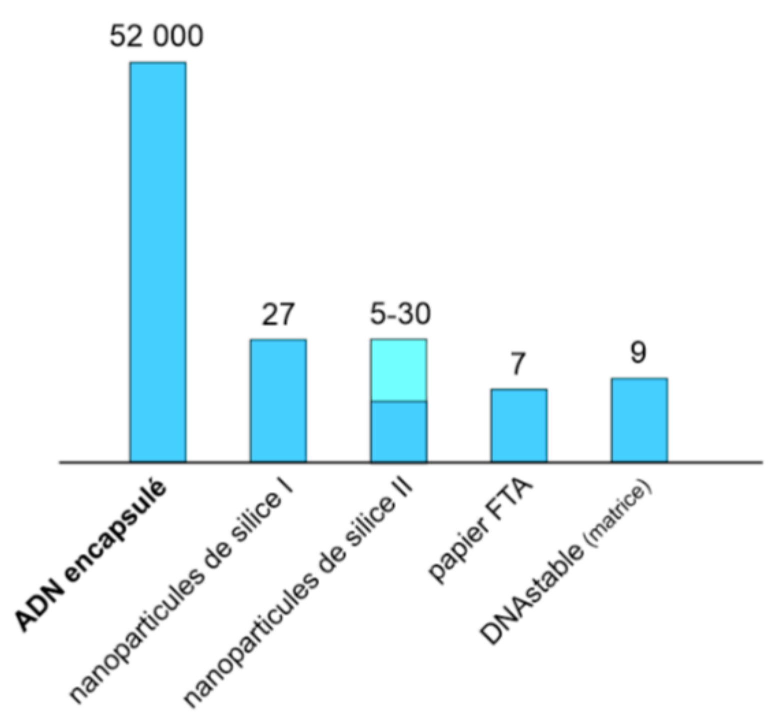


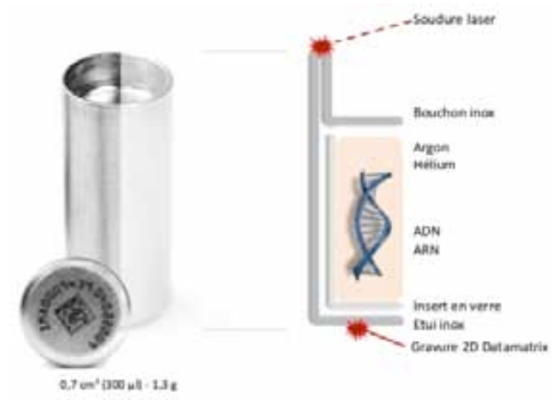
Archiving on DNA: LIFESPAN

2,000 to 2,000,000 years

Archiving on DNA: LIFESPAN

In the laboratory:
Half-lives (years) at 25°C





DataCenters:

- frequent and recurrent controls
- renewal every 5-7 years

Archiving on DNA: CONSUMPTION

<u>Storage</u> *per* se at room temperature with no consumption of energy, water etc.

<u>Operations</u> on DNA consume various resources

Electricity (according to IARPA, USA): > 1,000 times less

DataCenters:
- 2 - 4% of electricity

Archiving on DNA: DURABILITY

DNA technology will not lapse

ADN :
Physical support of our heredity
➔ with no obsolescence.

DataCenters:
- rapid obsolescence of formats and devices

Archiving on DNA: AMPLIFICATION

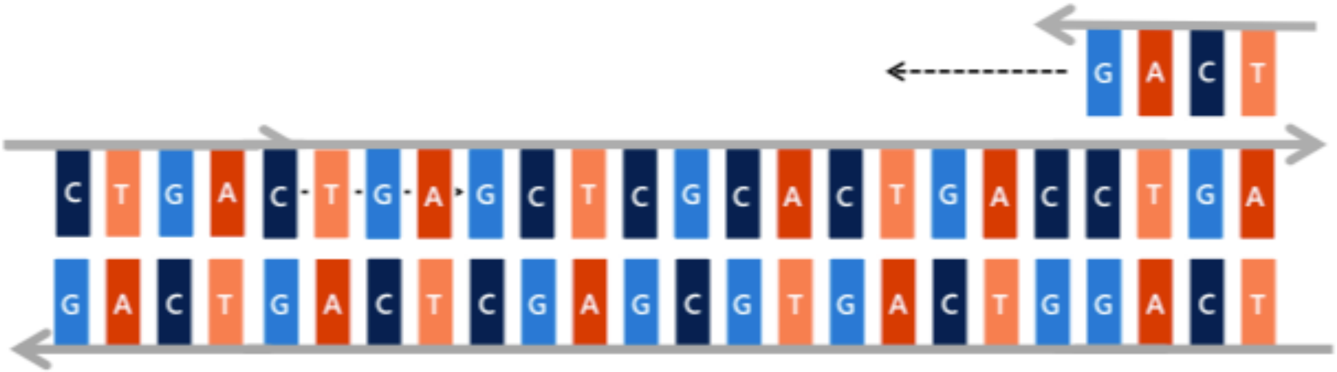
Multiply and dispatch digital data

DNA:

- 1 billion identical copies in 3 hours
- a few euro cents
- by PCR (polymerase chain reaction)

DataCenters:

- costly duplication



Archiving on DNA: DESTRUCTION

Destroy digital data at will

DNA:

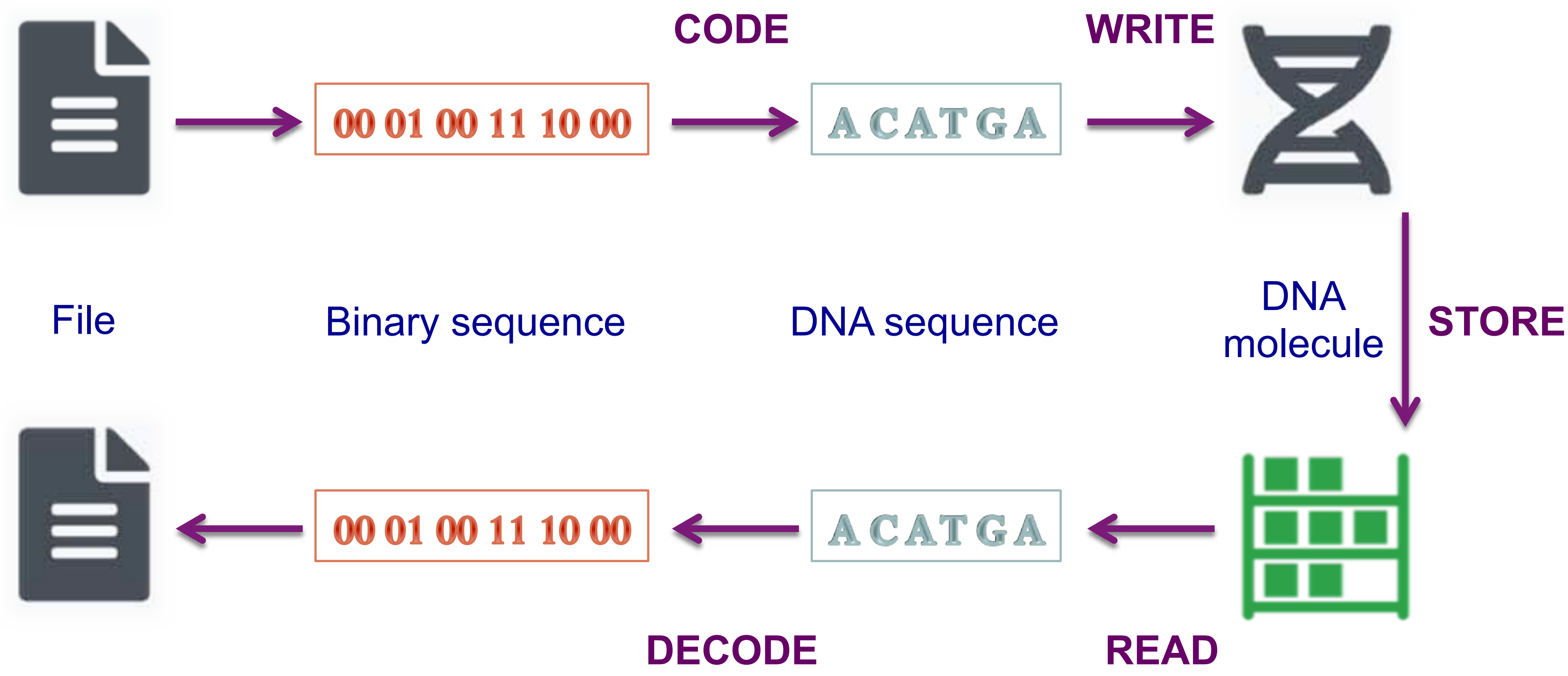
DNAase
quasi-instantaneous
a few euro cents

or pH, temperature etc.

DataCenters, or paper:

- not fullproof

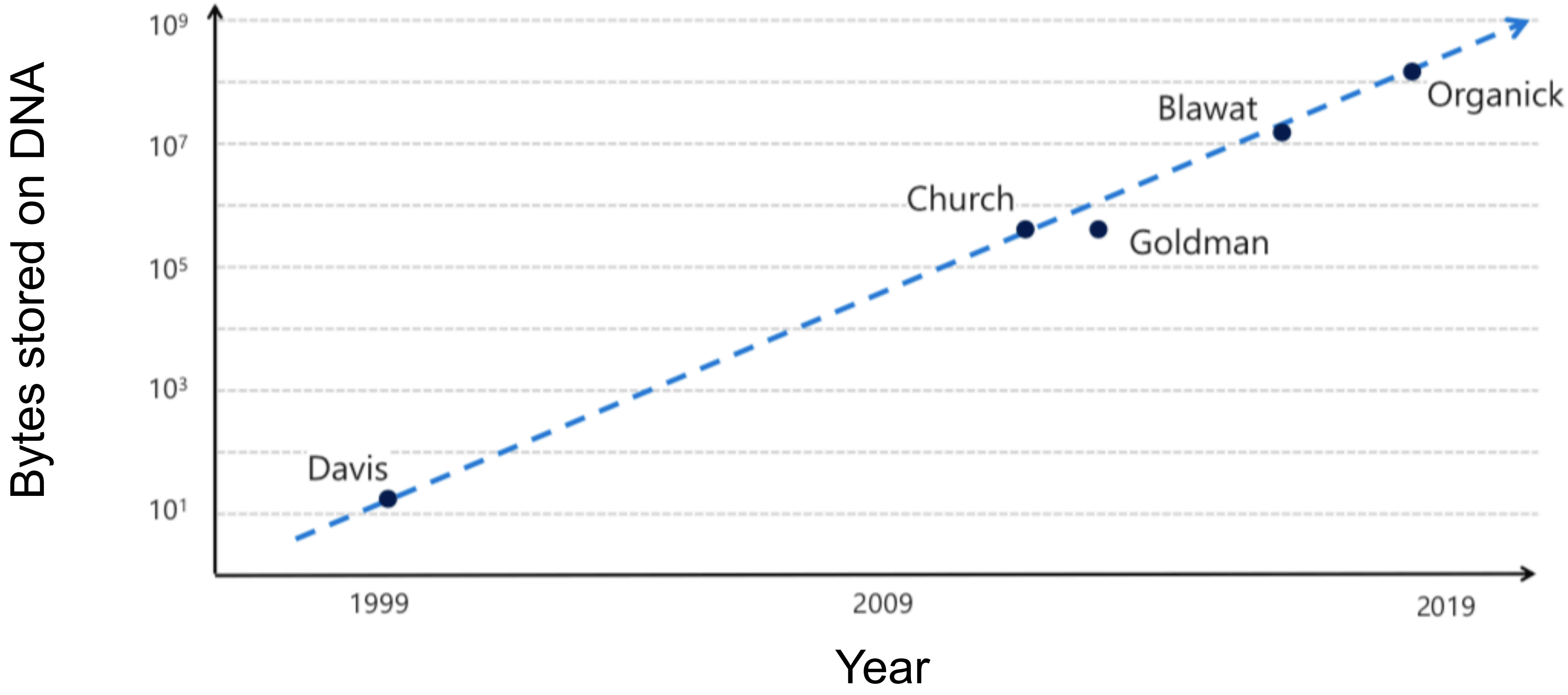
Archiving on DNA: PRINCIPLE



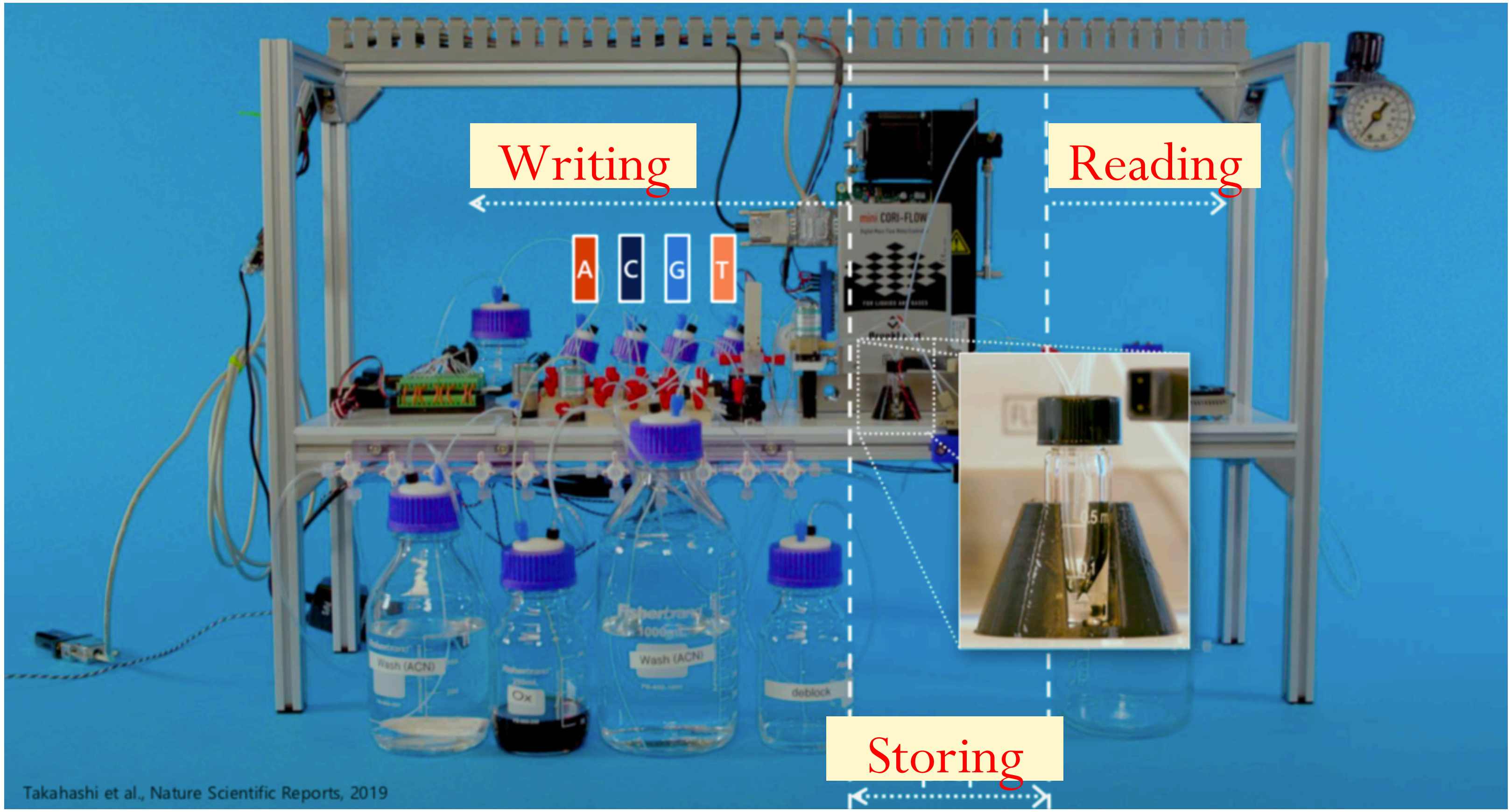
Archiving on DNA: A SHORT HISTORY

Record in 2018: 1 Gb (Microsoft Corp. & Univ. Washington, USA)

Expected in 2024: 1 Tb (IARPA, USA)



Automated Prototype

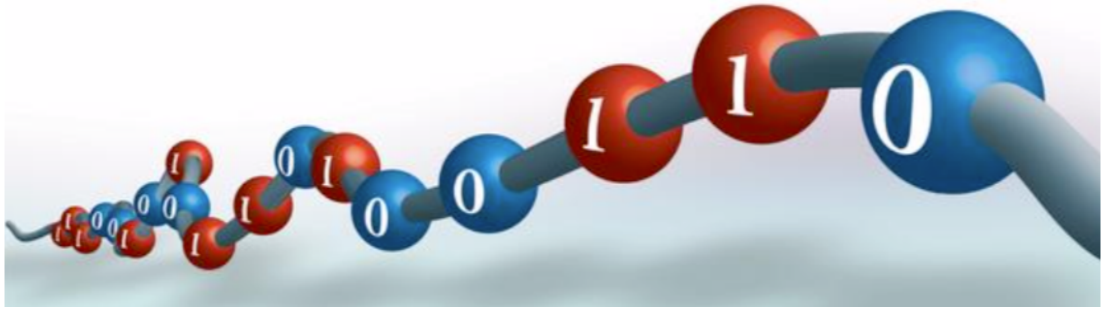


This first prototype carries out all the operations (Microsoft Corp., 2019).

DNA or other polymers?

Any hetero-polymer or co-polymer whose synthesis can be controlled step by step

"Digital" polymers:
- *Reading*: mass spectrometry
  following controlled fractionation
- *Writing*: multi-step elongation



Reading

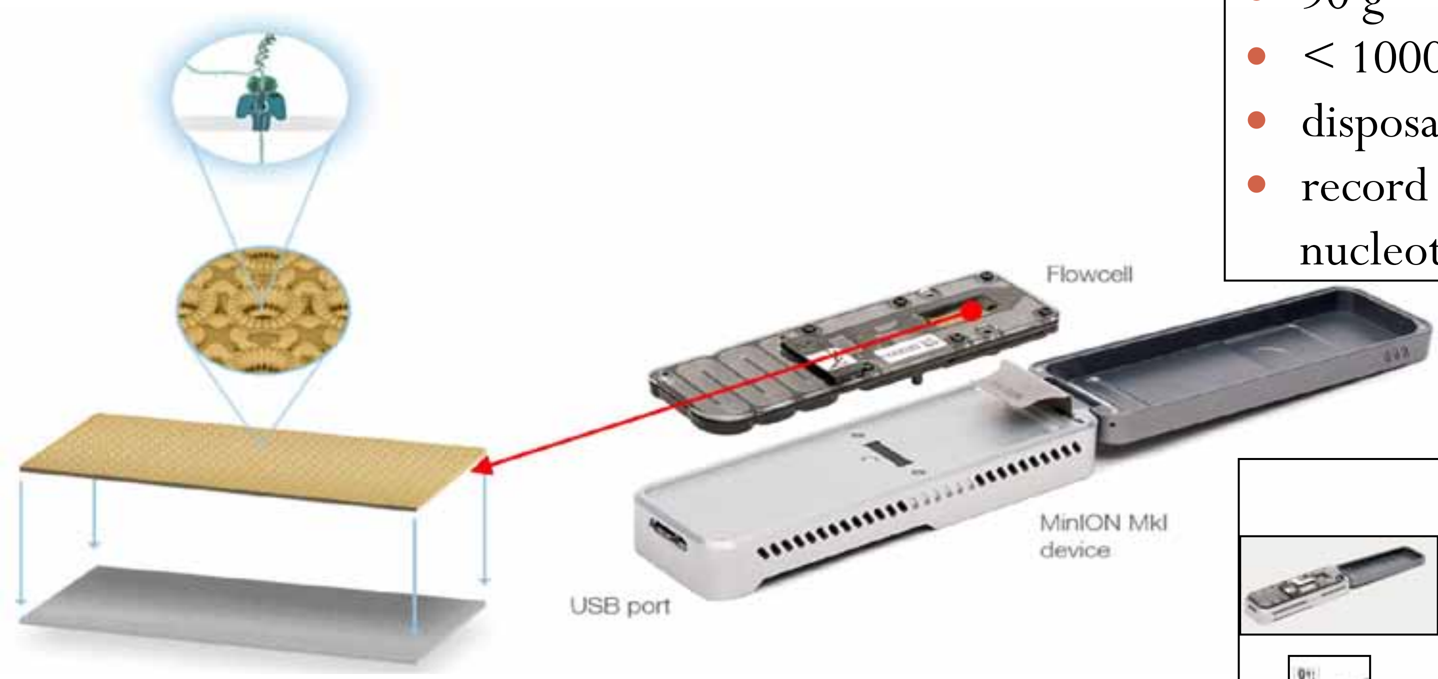
Possible approaches:

• one-stranded synthesis	[1G *or* 2G]
• nanopores	[3G]
• mass spectrometry	[digital polymers]

Reading

Massively parallel through nanopores — *e.g.*, MinION (Oxford Nanopore Technologies)

- > 20 Go of DNA sequence
- 90 g
- < 1000 $
- disposable
- record several million nucleotides in one run



Writing

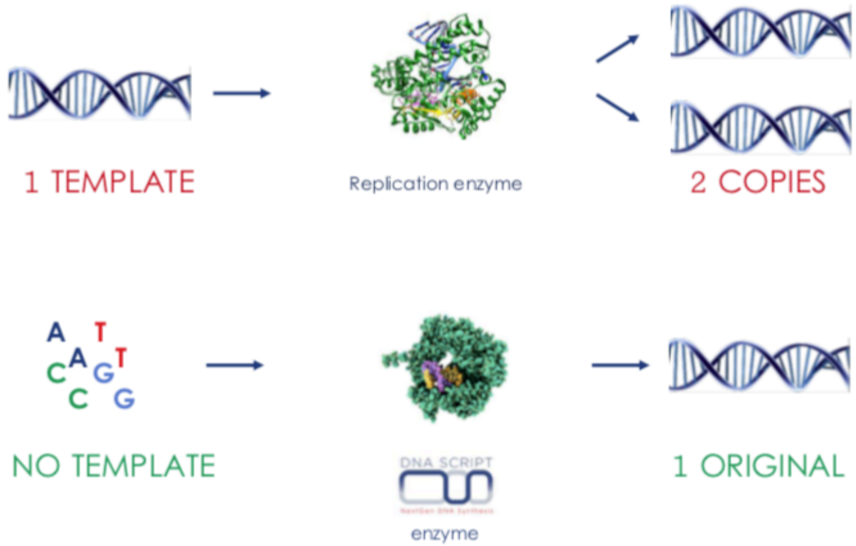
Possible approaches :

- Chemical synthesis of DNA
- Enzymatic synthesis of DNA
- Ligation of pre-fabricated cassettes of DNA
- Synthetic "digital" heteropolymers

Writing

Enzymatic synthesis:
Lowered error rate ➔ potentially longer DNA fragments (>400).

In a live cell, the DNA reading/writing process is faster than in a Flash memory (< 100 µsecond per bit). This gives a notion of the potential of the biological approach.



Economical prospect of DNA archiving

Several orders of magnitude are currently lacking:

- ~ 1,000 for the reading cost
- ~ 100,000,000 for the writing cost

Is it a barrier?

No, DNA technologies gained a 1,000,000 factor in 10 years (2-fold every 6 months ➔ much faster than IT).

No, in some applications, massive parallelization is possible.

In 2024 a single machine will presumably write and read 1 Tb a day.

Potential market for DNA archiving

Handicap of DNA or other polymers :
slow and costly writing/reading processes
➔ COLD STORAGE)

- under 3-8 years for niche markets
(cultural, scientific, bank, crypto-keys HERITAGES)
- under 9-18 years for more global markets.

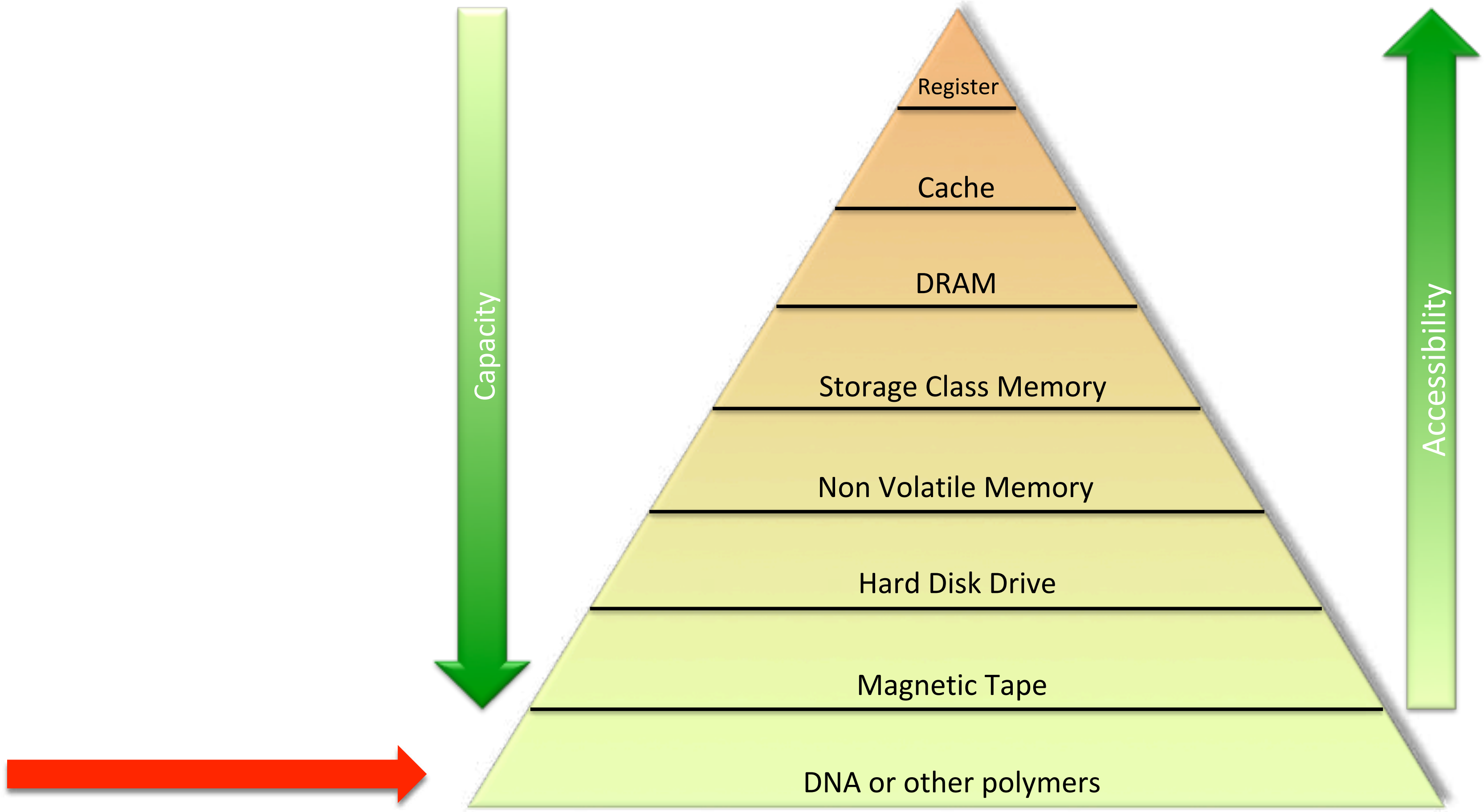
Investments

Public investment as of 2021:
USA: 150 M$ (IARPA, DARPA, NSF)
China: ?? Huawei, BGI Genomics
Europe: 1 lab at EBI (United Kingdom)
Germany: 4,2 M€
Switzerland: 1 lab
France : a few labs in Strasbourg,
Rennes, Nice, Paris

Private investments:
Several companies in
USA, U-K, Ireland, Germany, France

DNA Data Storage Alliance (2020):
Microsoft, Western Digital, Twist Bioscience, Illumina, Quantum …
+ many small actors

Pyramid of memory types in computer systems



ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ CHOISI PARTAGÉ

NATIONAL ACADEMY OF TECHNOLOGIES OF FRANCE
SHARING A REASONED, CHOSEN PROGRESS

Thank you for your attention!