A leading institution at the heart of the digital society

EURECOM
Sophia Antipolis

Highly durable and dense data storage through synthetic DNA

Raja Appuswamy

08-08-2023

@CERN Openlab

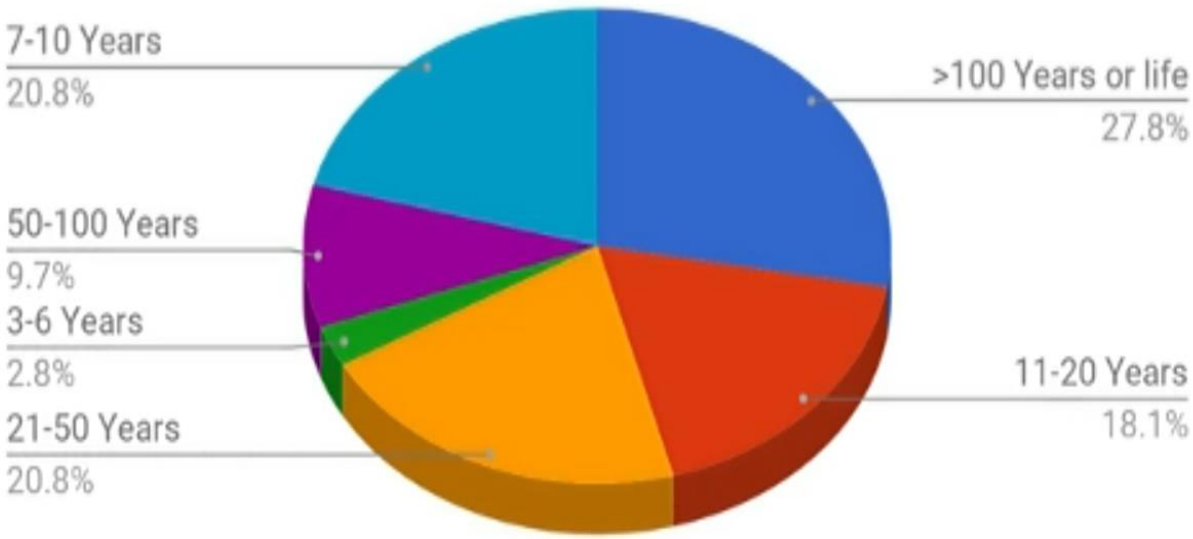
Cti
Commission des Titres d'Ingénieur

Growth of archival data

“50% of 175ZB global datasphere will be enterprise data in 2025” ***[IDC]***

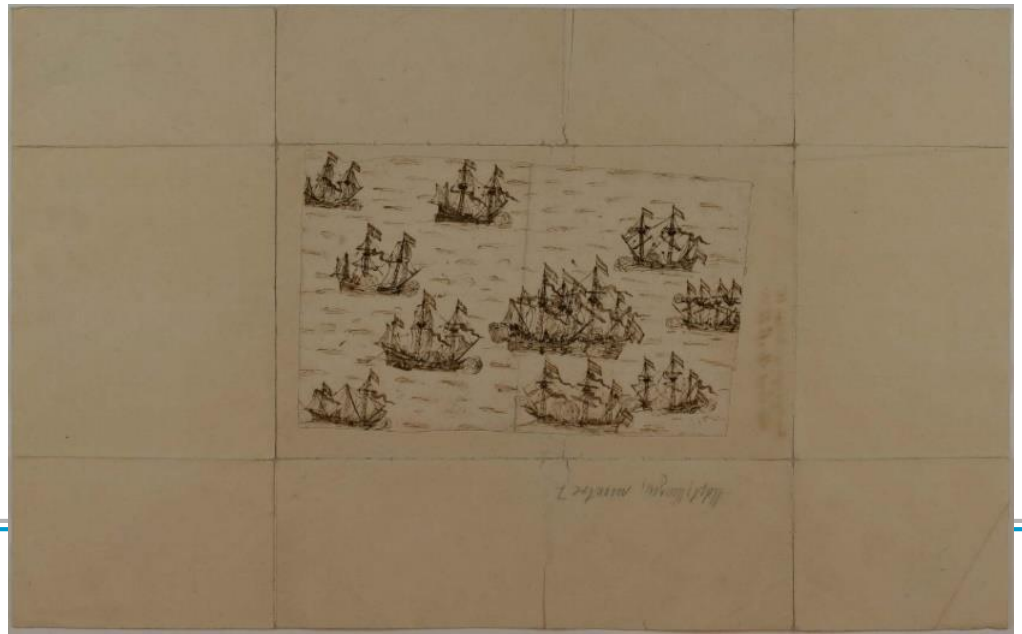
“80% data is cold, and increasing at 60% CAGR” **[Horison]**

“60% of archival data stored longer than 20 years”
[SNIA]



Digital Preservation Example (1): Danish National Archive (DNA)

- **Danish National Archive**
  - Preservation of digitally created/retro-digitized data since 1970
- **Digitized hand drawings of King Christian IV**
  - Actual drawings date back to 1583-1591
  - Material ranked as having unique national significance



Digital Preservation Example (2): ODEUROPA Project

- **ODEUROPA**
  - Award winning EU project on preserving olfactory heritage
- **Preserving frankincense**
  - Artwork, odour descriptors, knowledge graphs, articles, etc
  - Information assembled by 10 members of ODEUROPA consortium

Digital Preservation Example (3): HORROR FILM Pilot

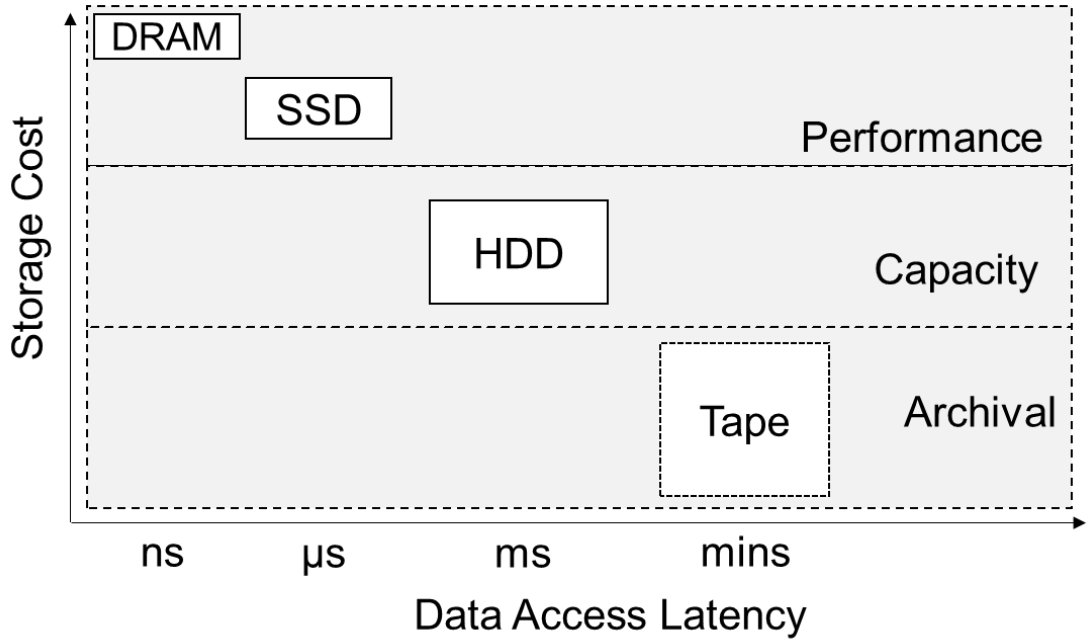
- **Horror Film by Malcolm Legrice**
  - First of its kind expanded cinema involving moving image and a performer
  - Made by the leading artist of London Film Makers' Co-op
  - Movie is an active visual aid to performers & not a passive film
- **Preserving Horror Film through time**
  - Instructions to perform the film, video explainer, performances, etc.
  - Collab. with Louise Curham & Lucas Ihlein from Canberra, Legrice family

Growth of archival data

“50% of 175ZB global datasphere will be enterprise data in 2025” *[IDC]*

“80% data is cold, and increasing at 60% CAGR” [Horison]

“60% of archival data stored longer than 20 years” *[SNIA]*



Current tape-based archival suffers from fundamental limitations

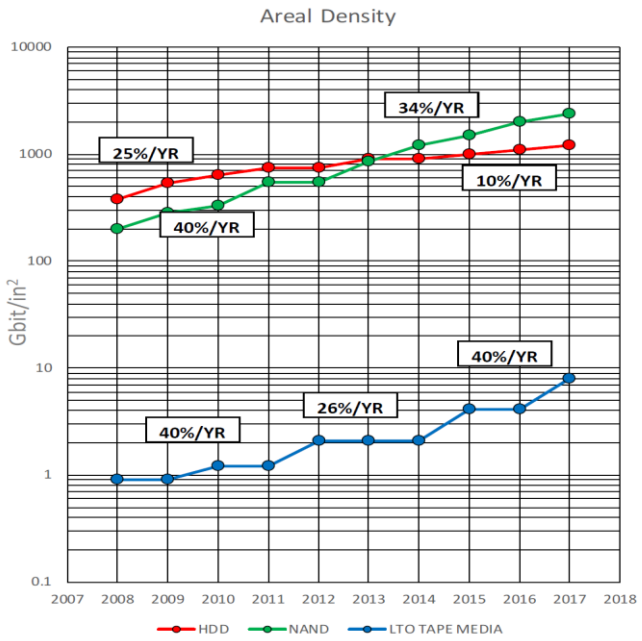
Long-term Archival Challenge: Media Obsolescence

“60% of archival data stored longer than 20 years”

[SNIA]

	Capacity	Durability
Flash	TBs	~5 yrs
HDD	100s TBs	~5 yrs
Tape	PBs	~10s yrs

“Kryder’s rate of tape: 31%/YR average”



Limited backwards compatibility

	Tape Drives				
Version	LTO-6	LTO-5	LTO-4	LTO-3	LTO-2
LTO6	Read/Write				
LTO6 WORM	Read/Write				
LTO5	Read/Write	Read/Write			
LTO5 WORM	Read/Write	Read/Write			
LTO4	Read	Read/Write	Read/Write		
LTO4 WORM	Read	Read/Write	Read/Write		
LTO3		Read	Read/Write	Read/Write	
LTO3 WORM		Read	Read/Write	Read/Write	
LTO2			Read	Read/Write	Read/Write
LTO1				Read	Read/Write
Cleaning Tape	Supported	Supported	Supported	Supported	Supported

Net Effect of Obsolescence

28 Apr 2017 | 15:00 GMT

The Lost Picture Show: Hollywood Archivists Can't Outpace Obsolescence

Studios invested heavily in magnetic-tape storage for film archiving but now struggle to keep up with the technology

By **Marty Perlmutter**

"There's going to be a large dead period," he told me, "from the late '90s through 2020, where most media will be lost."

Using Analog Media for Archiving Digital Data

ISO 9706
Permanent paper

- **Analog media (Paper, microform, film)**
  - Preserving documents/art work in museums and archives

- **Can also preserve digital data**
  - Convert data into barcodes and shoot to film
  - UNICEF child rights, health data (PIQL)
  - Micr'Olonys in CIDR 2021

- **"Solving" media decay & obsolescence**
  - ISO 9706 paper, LE-500 microfilm, PIQL film lasts >500 years
  - Paper, film requires basic scanning technology
  - No need to migrate data

But analog media is severely limited in density

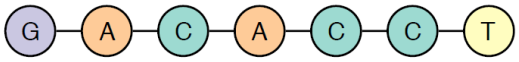
DNA as a digital storage media

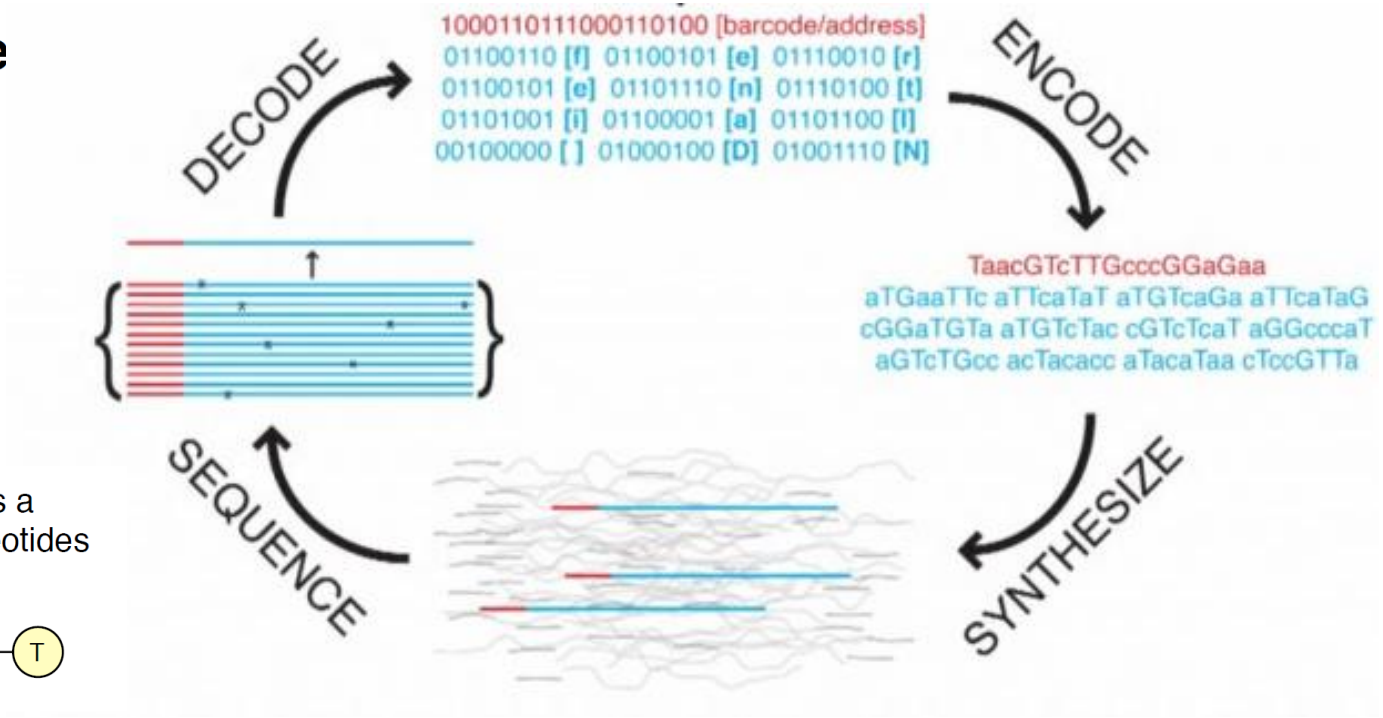
DNA molecule

Four nucleotides:

- A Adenine
- C Cytosine
- G Guanine
- T Thymine

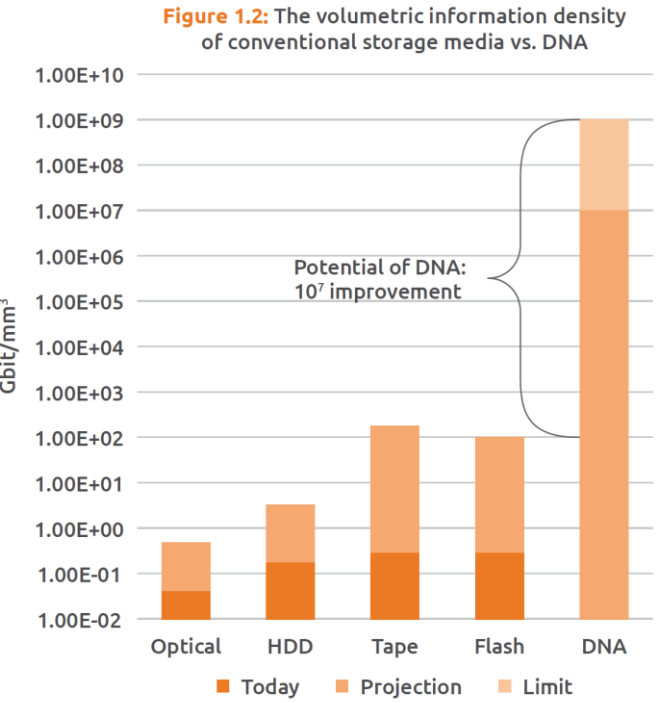
DNA strand (oligonucleotide) is a linear sequence of these nucleotides





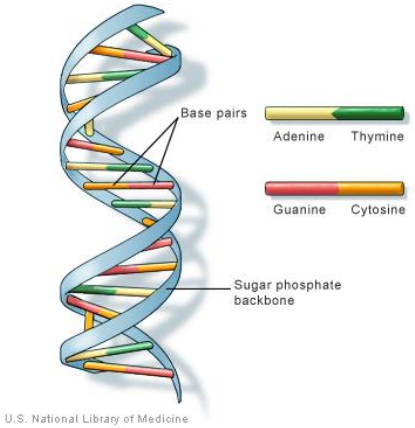
Why DNA?

Dense



Durable

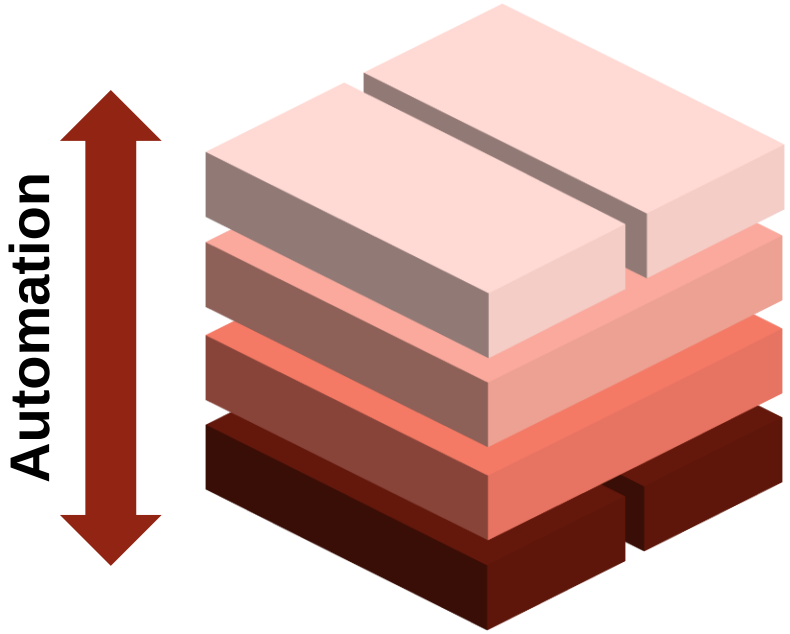
Kryder's rate: 0



How do we use DNA as an archival media?

OligoArchive (https://oligoarchive.eu)
MOSS (https://moss-eu.com/)

Goal: implement a custom storage stack for data archival on DNA



Application Layer
Encoding structured (database) and unstructured (imaging) data

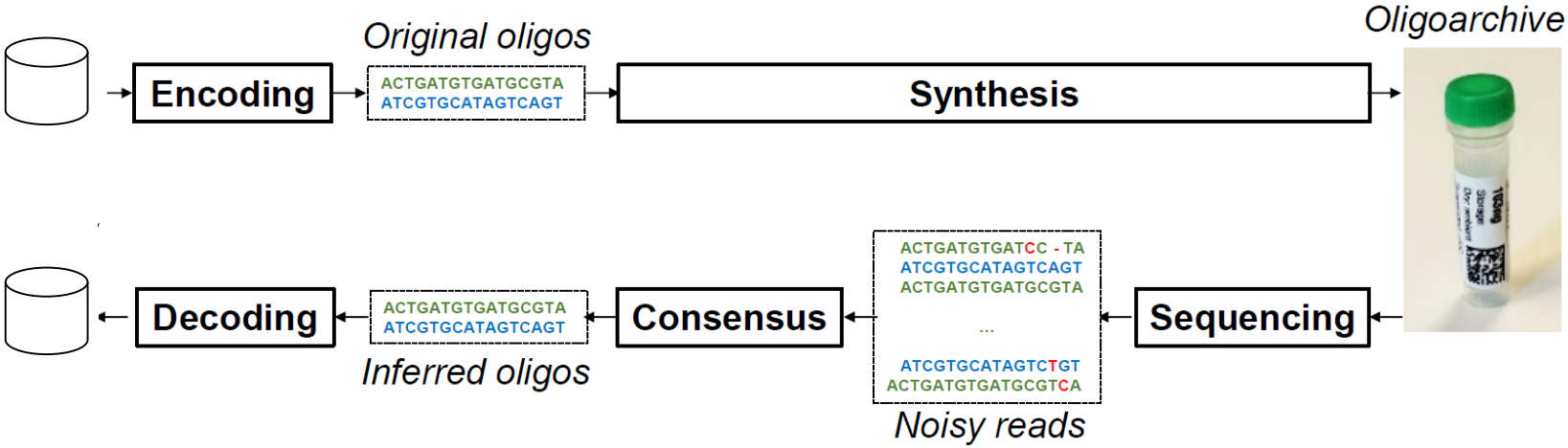
OS Layer
File system abstraction

Controller Layer
Data processing capabilities

Media Layer
Synthesis and Sequencing

DNA Archival & Restoration: Challenges

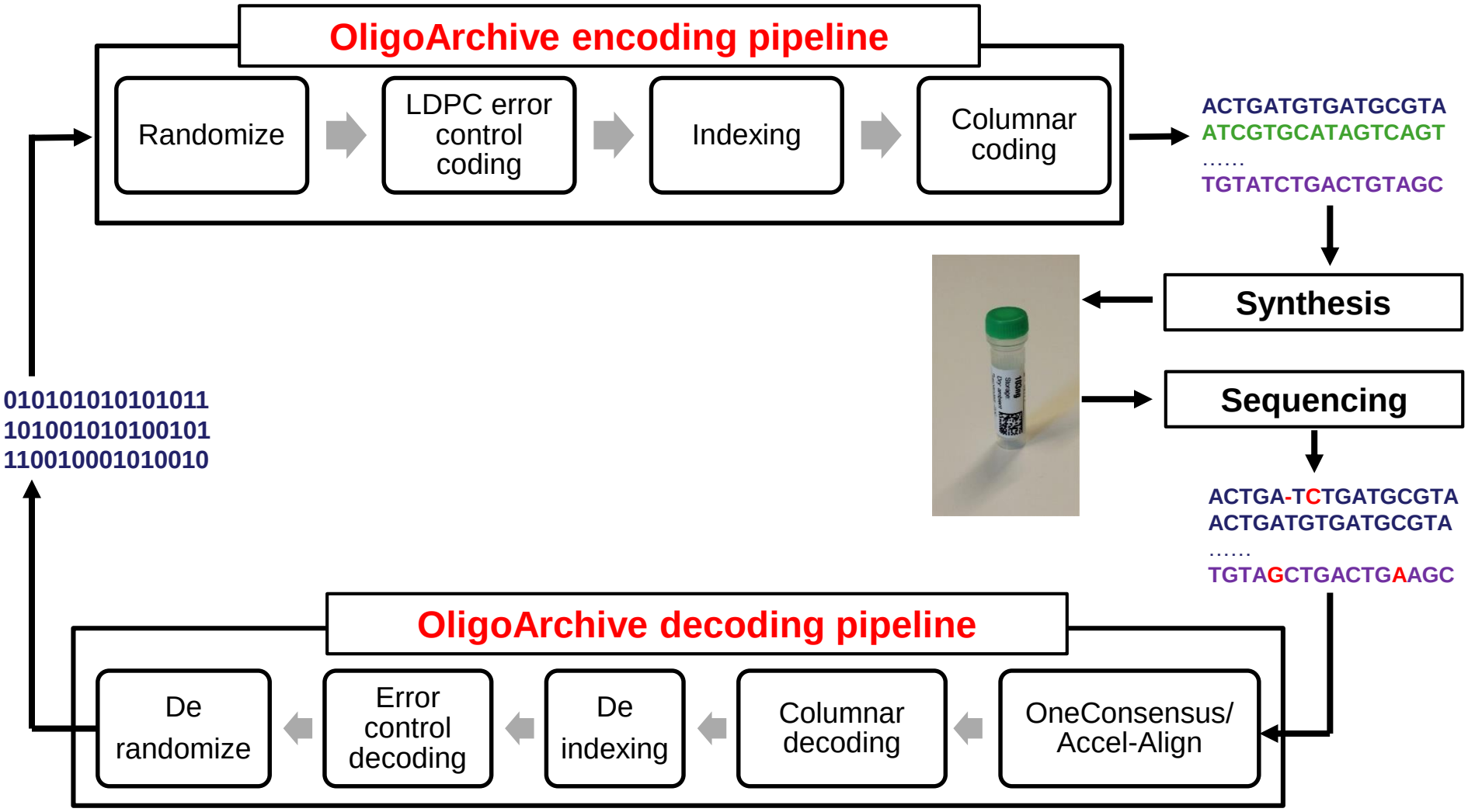
- **Each DNA is limited to a few hundred nucleotides**
  - Data spread out across millions of DNA
- **Not all DNA are created equal**
  - G-C content limitations, homopolymers
- **DNA has no addressing**
  - Need to add ordering information in DNA



Biochemical errors

- substitution, insertions, deletions,
- Bias & duplication

OligoArchive DNA Storage Pipeline



OligoArchive enables high-density digital archival on DNA

Digital Preservation & DNA storage: Danish National Archive (DNA)

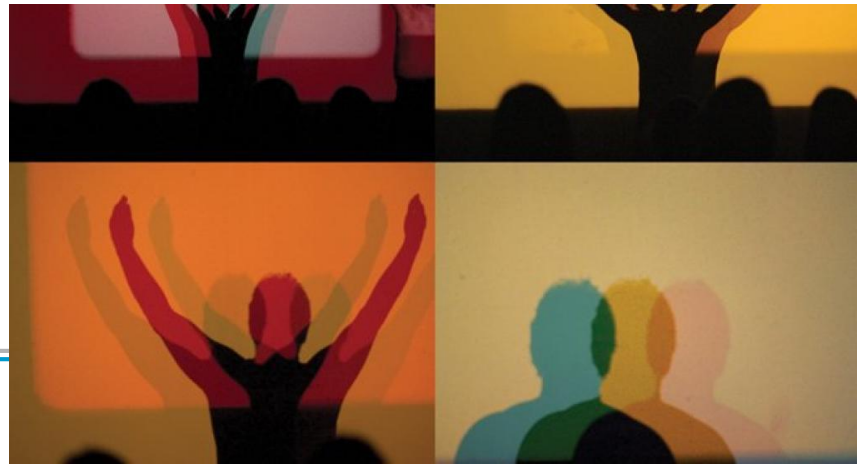
- **Danish National Archive**
  - Preservation of digitally created/retro-digitized data since 1970
- **Digitized hand drawings of King Christian IV**
  - Actual drawings date back to 1583-1591
  - Material ranked as having unique national significance
- **DNA4DNA Pilot**
  - 14MB SIARD archive
  - Encoded into 261,336 oligos (280nt)
  - 100% recovery verified

Cultural preservation & DNA storage: ODEUROPA Project

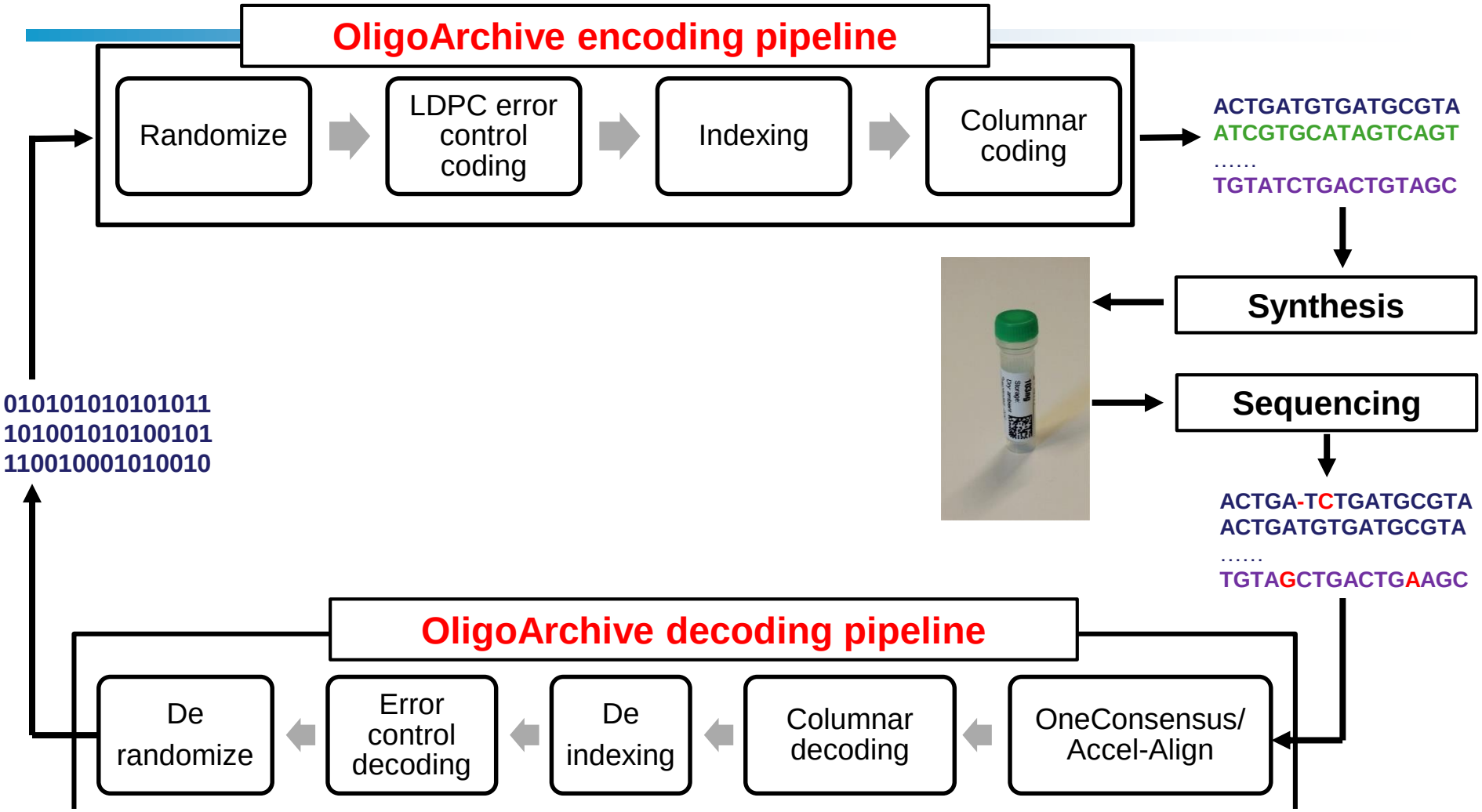
- **ODEUROPA**
  - Award winning HE project on preserving olfactory heritage
- **Preserving frankincense**
  - Artwork, odour descriptors, knowledge graphs, academic articles, etc
  - Information assembled by 10 members of ODEUROPA consortium
- **ODEUROPA Pilot**
  - 10MB SIARD archive
  - Encoded into 177,504 oligos (120nt)
  - 100% recovery verified

Art preservation & DNA storage: HORROR FILM Pilot

- **Horror Film by Malcolm Legrice**
  - First of its kind expanded cinema involving moving image and a performer
  - Made by the leading artist of London Film Makers' Co-op
  - Movie is an active visual aid to performers & not a passive film
- **Preserving Horror Film through time**
  - Instructions to perform the film, video explainer, performances, etc.
  - Collab. with Louise Curham & Lucas Ihlein from Canberra, Legrice family
- **HORROR Film Pilot**
  - 42MB ZIP archive
  - Encoded into 2M oligos! (168nt)
  - Under sequencing now



OligoArchive DNA Storage Pipeline



OligoArchive enables high-density digital archival on DNA

But who archives the OA-DSM decoder?

Extended Format Obsolescence Issues

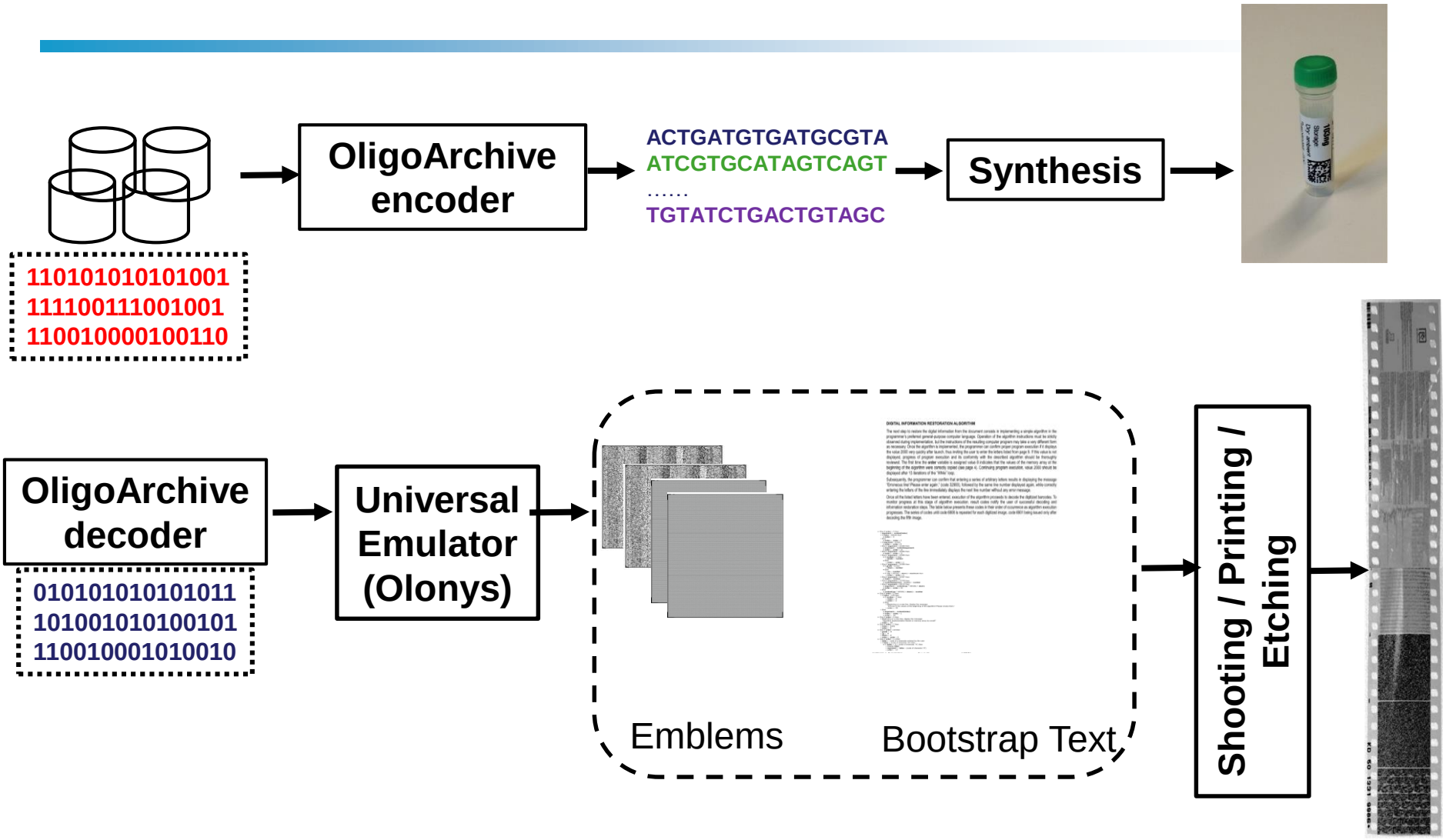
- **New media impose new "media layout"**
  - Storing data on DNA requires encoding data into oligos
  - Getting data from DNA requires converting oligos back into digital data
- **Decoders are complex**
  - Use error-correcting codes that require parity-check matrix and parameters for decoding
- **We want to archive media layout decoders with data**
  - Otherwise, can sequence oligos, but not decode

QN: How can a user run decoders developed today 100 years later on a computing platform that might not exist today?

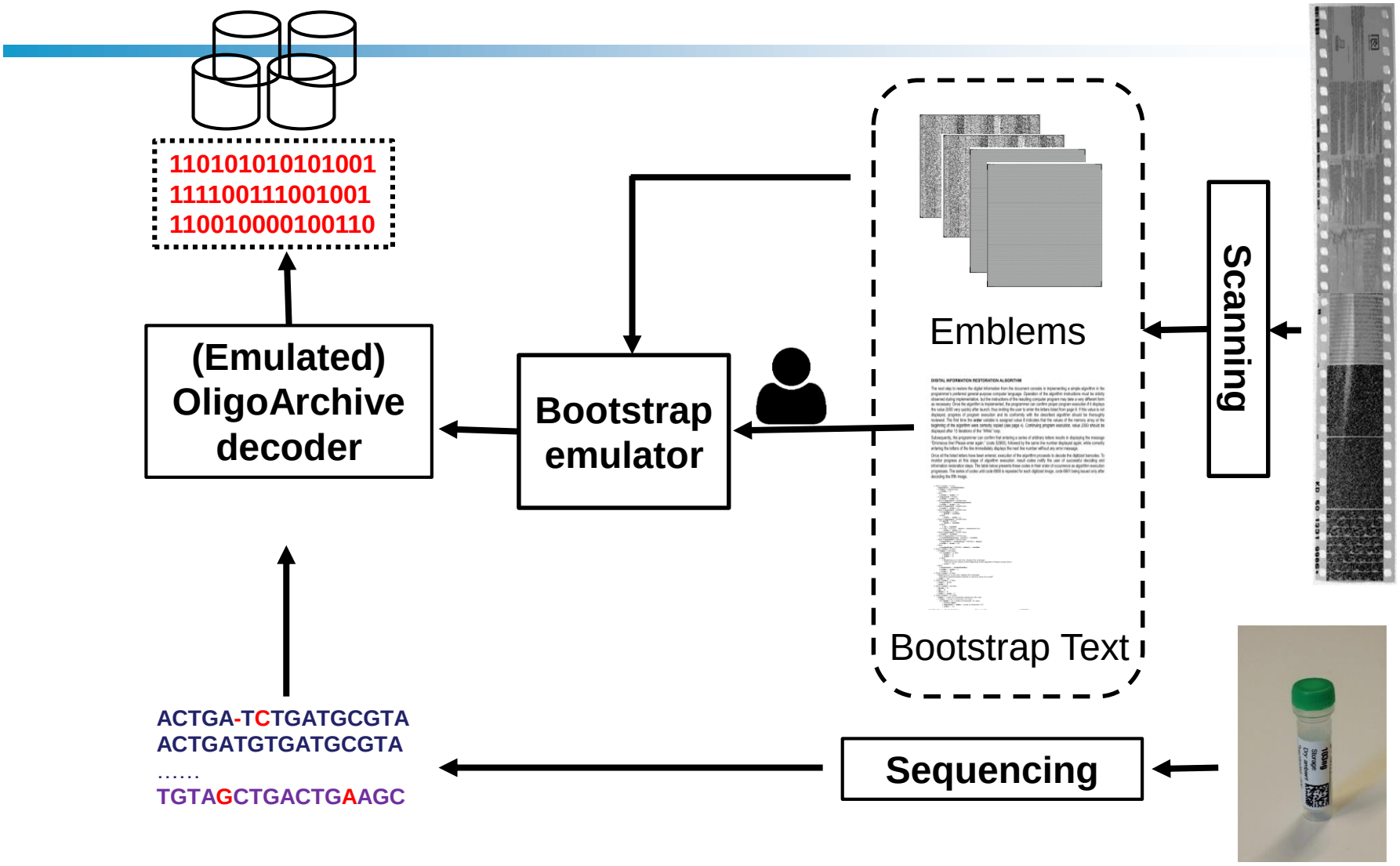
Taking a Page from Digital Preservation

- **Emulation**
  - Technology used to simulate one hardware environment using another
  - Emulation used in software preservation for getting old software to run on modern computing environments
- **Universal Emulation**
  - Observation: Often only need to preserve application logic, not current hardware/software stack
  - Develop a virtual software processor with a very simple ISA that can be easily emulated. Develop software to target this virtual ISA.
- **Central idea: Universal Layout Emulation**
  - Use a universal emulator to archive layout decoders with data
  - Collaboration with Vincent Joguin@EUPALIA

Analog Bootstrap for DNA Archives



Restoration Using Analog Bootstrap



Migration-Free, End-to-end Passive Preservation of Digital Data with Analog + Biological Media

Portability and Programmability of Bootstrap

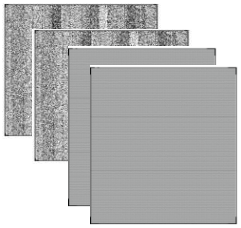
- **How hard is it to bootstrap the Olonys emulator?**
  - We requested first-year UG students (Lycee Bonaparte, Toulon), engineers (CNES), and researchers (EURECOM), to implement the VeRisc emulator
  - The emulator was implemented on Windows and Linux in JavaScript, Python, C++, and C# in less than 1 week

- **How portable is the Olonys emulator?**
  - Olonys also ported to Raspberry Pi and GameBoy Advance (ARM), TI-85 calculator (Z80), Atari Falcon (68030), and Palm PDA (68000).

Putting it all together: Towards Holistic Passive Preservation

Synthetic DNA

Analog bootstrap



Emblems

Bootstrap Text

Solve media decay issues with DNA

Solve media obsolescence with analog bootstrap

Conclusion

- **Contemporary magnetic media suffers from decay and obsolescence**
  - Continuous migration expensive for long-term archival/preservation
- **DNA provides a biological alternative**
  - Dense, durable, eternal relevance (solves media decay)
  - OligoArchive & MOSS enable the use of DNA as a digital media
- **End-to-end passive preservation is feasible**
  - *Synthetic DNA*: High-density, decay-free digital archival media
  - *Analog media* + *emulation*: Bootstrap for archiving DNA decoders

We are recruiting PhD students, research engineers & postdocs to work on DNA storage and beyond!
Please get in touch!

UAG
UGA
UAA