A leading institution at the heart of the digital society

EURECOM
Sophia Antipolis

Microform and Macromolecules: Archiving *digital* data on *analog* and *biological* storage media

Raja Appuswamy

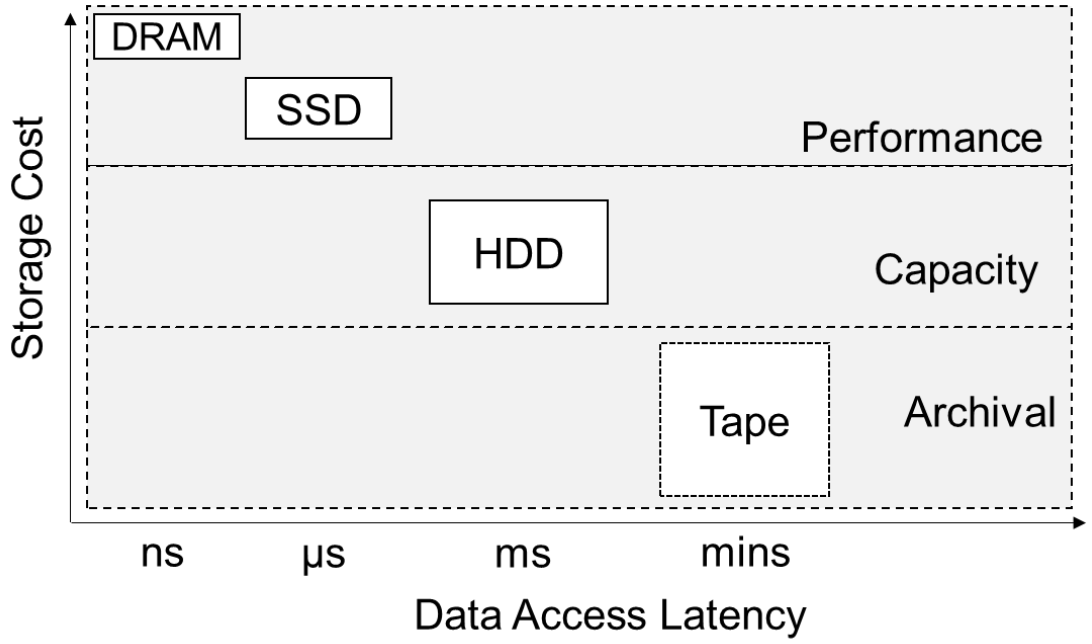
ACAT 2022

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]**



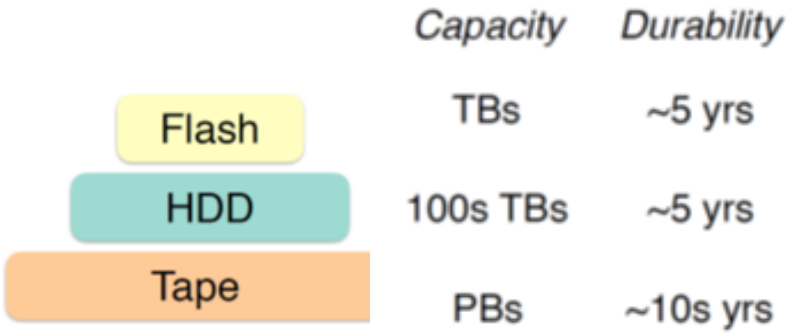
Current tape-based archival suffers from fundamental limitations

Challenges in Long-Term Digital Archival

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

[SNIA 100 Year Archive]

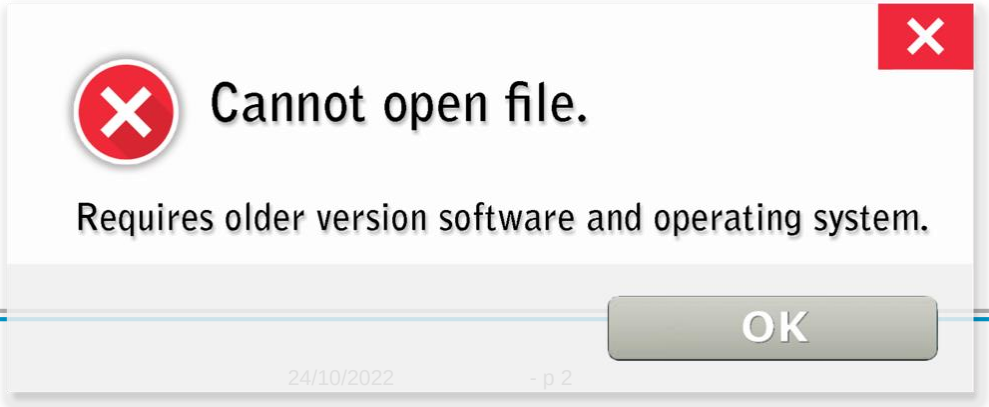
Media decay



Media obsolescence

	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

Format obsolescence



Net Effect: Migration-based Active Preservation

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.***"

Enterprise DBMS archives will soon face obsolescence issues

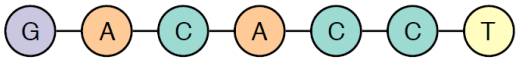
DNA as a digital storage media

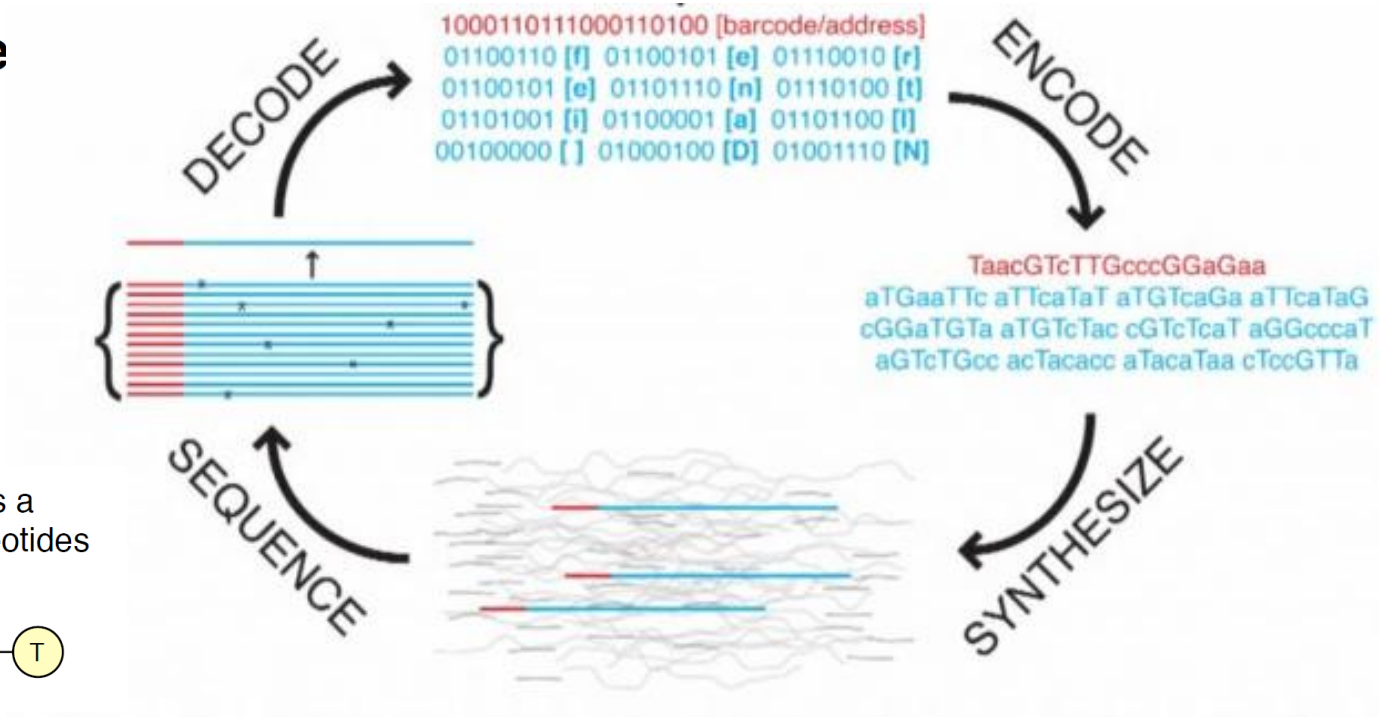
DNA molecule

Four nucleotides:

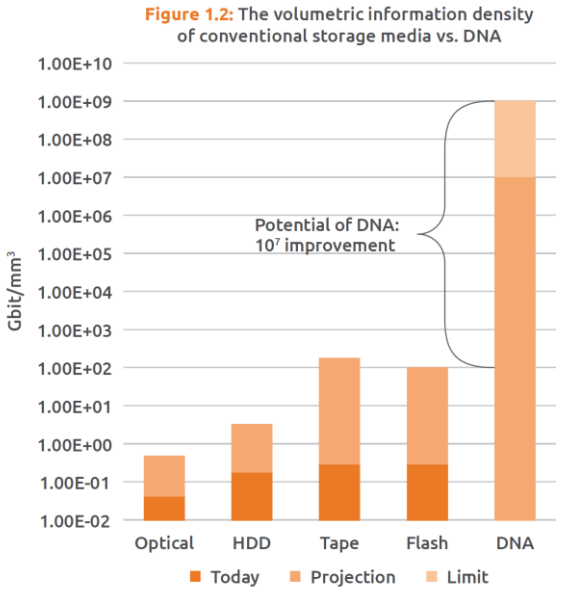
- A Adenine
- C Cytosine
- G Guanine
- T Thymine

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





Why DNA



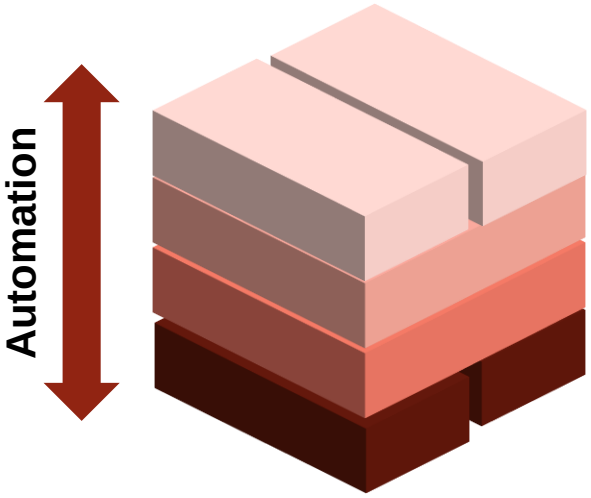
10^7 higher density

EU FET project Oligoarchive focuses on using DNA as an intelligent storage medium

OLIGO ARCHIVE


Woolly mammoth on verge of resurrection, scientists reveal

Scientist leading 'de-extinction' effort says Harvard team could create hybrid mammoth-elephant embryo in two years



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

OS Layer
Advanced access paths (block, fs, …)

Controller Layer
Near-molecule query processing

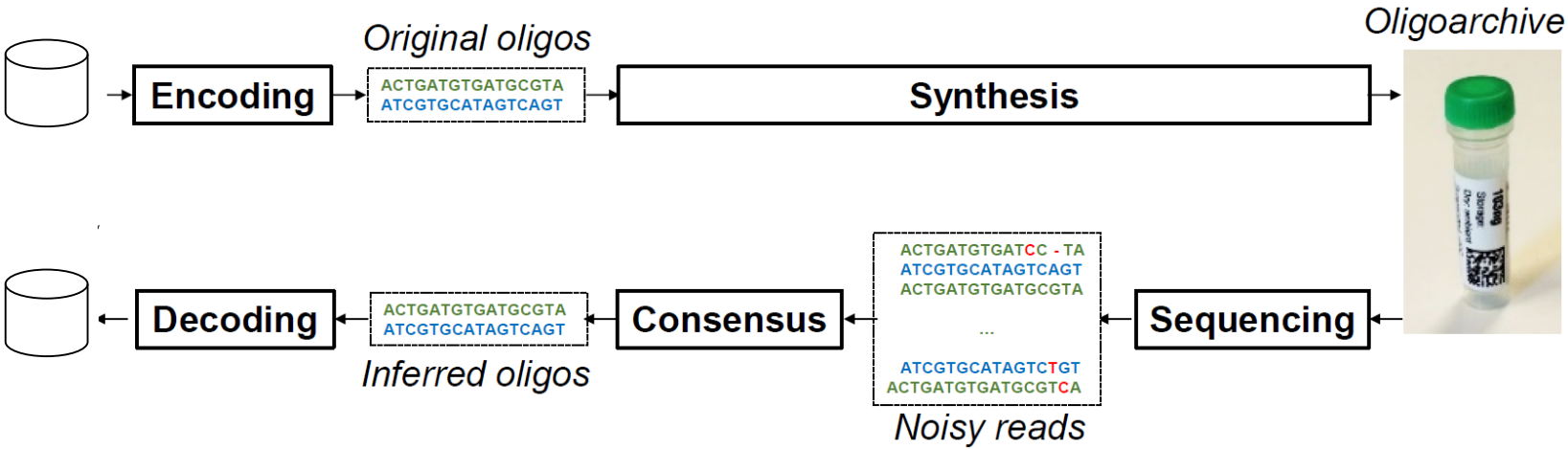
Media Layer
Synthesis and Sequencing

Durable, eternally relevant

EURECOM
Sophia Antipolis

DNA Archival & Restoration: Challenges

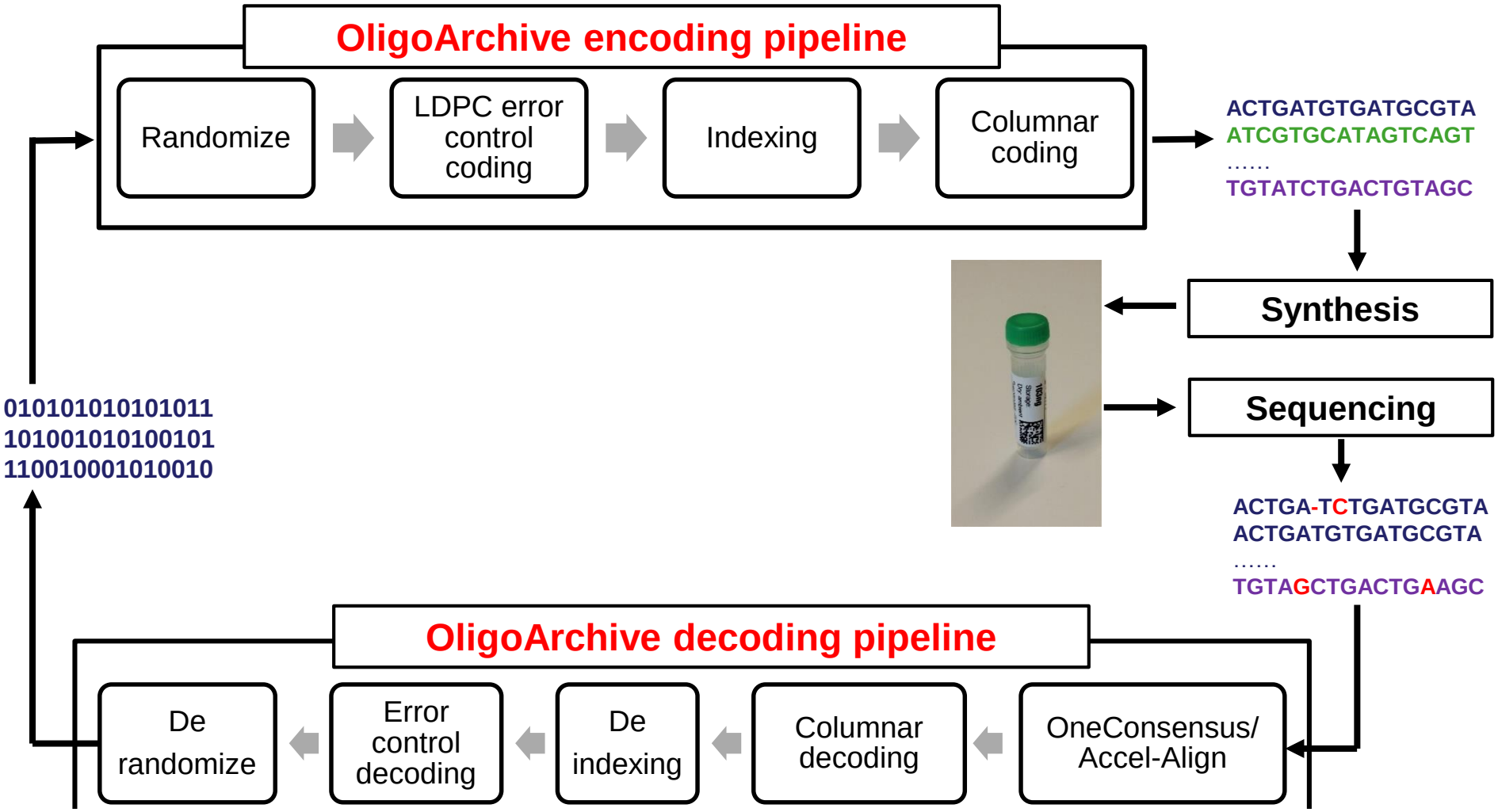
- **Each DNA is limited to a few hundred nucleotides**
  - Data spread out across millions of DNA
- **Not all DNA are created equal**
  - G-C content limitations, homopolymers
- **DNA has no addressing**
  - Need to add ordering information in DNA



Biochemical errors

- substitution, insertions, deletions,
- Bias & duplication

OligoArchive DNA Storage Pipeline



OligoArchive enables high-density digital archival on DNA

DNA does not solve format obsolescence issues

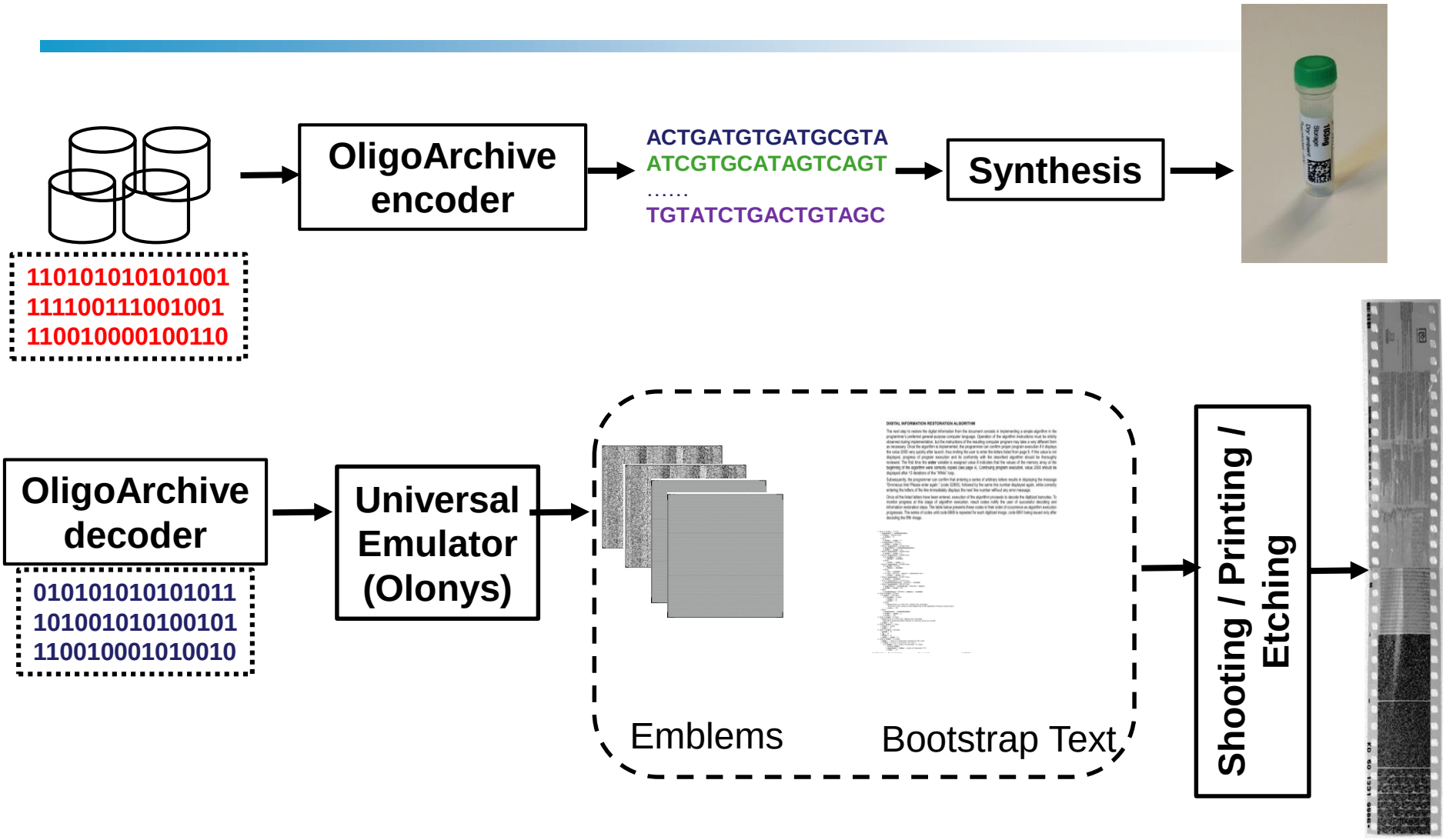
DNA: New Format Obsolescence Issues

- **New media imposes a new format**
  - Storing data on DNA requires encoding data into oligos
  - Retrieving data 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 decoders**
  - Otherwise, can sequence oligos, but not decode
- **Analog media + emulation to bootstrap DNA storage**
  - Ongoing collaboration with Vincent Joguin (EUPALIA), Martin Kunze (MoM/CERAMICRO)

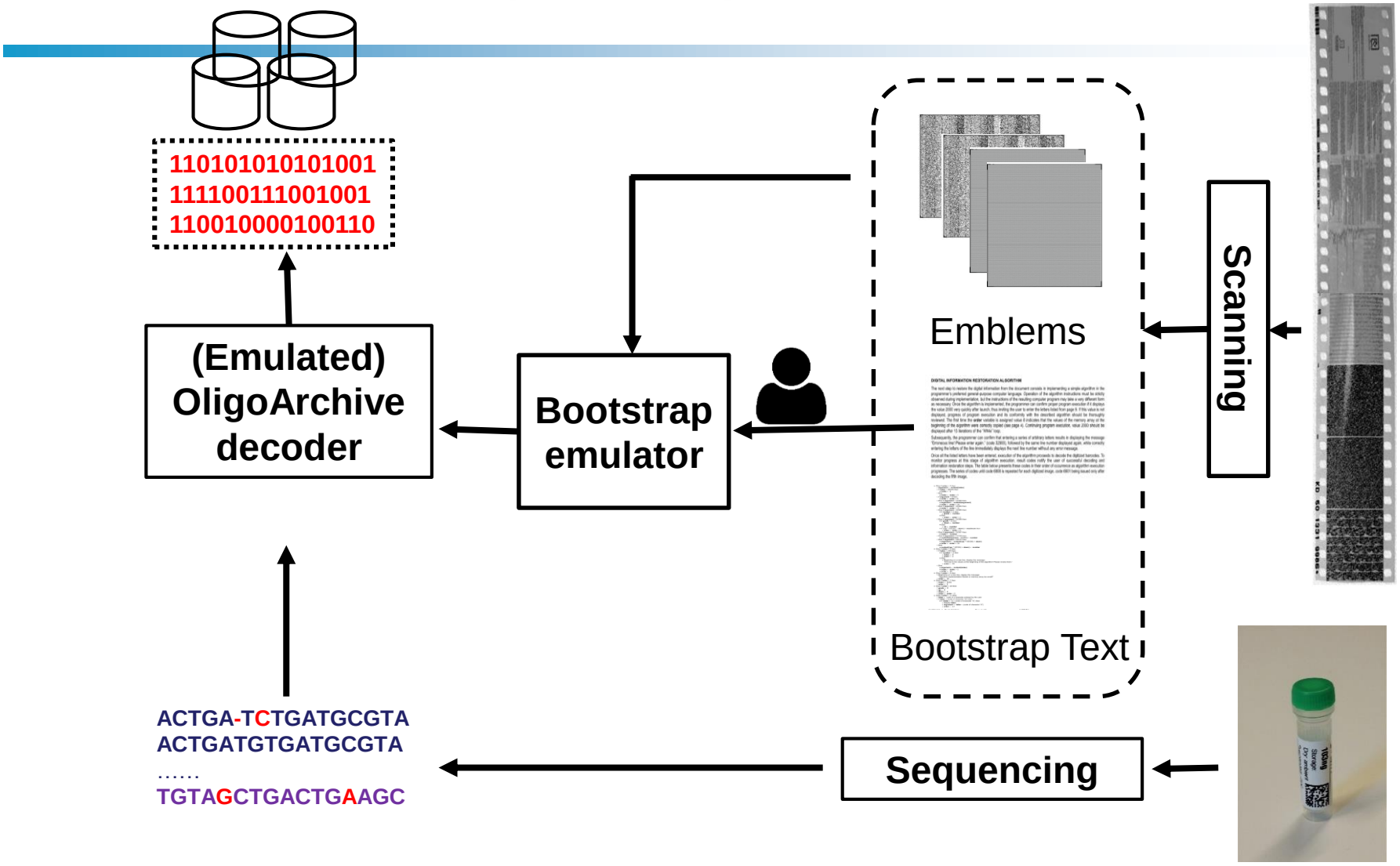
Taking a Page from Digital Preservation

- **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 decoders with data

Analog Bootstrap for DNA Archives



Restoration Using Analog Bootstrap



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

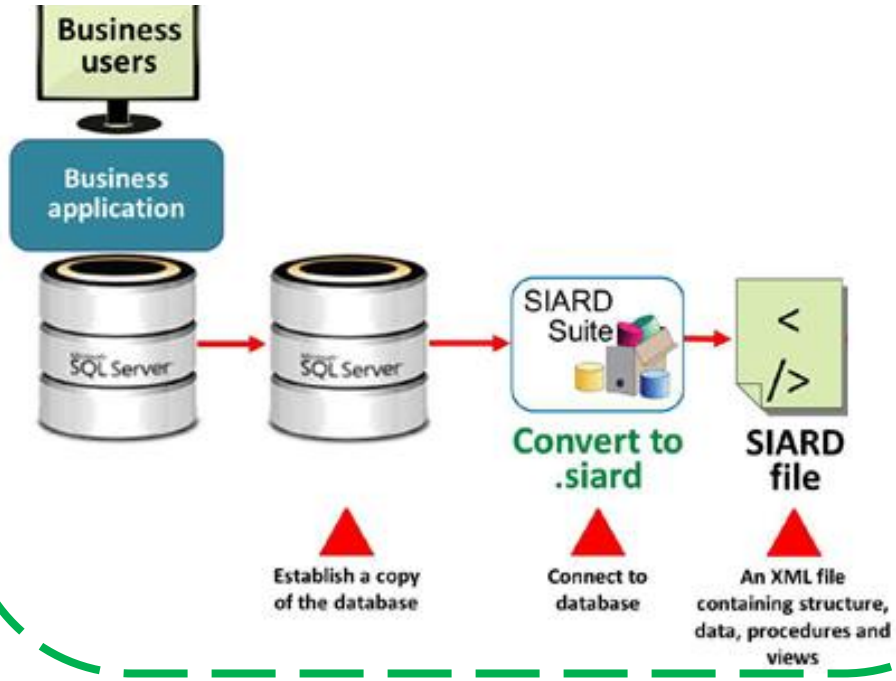
Digital Preservation with Synthetic DNA: Danish National Archive Example

- **Danish National Archive**
  - Preservation of digitally created/retro-digitized data since 1970
- **Digitized hand drawings of King Cristian IV**
  - Actual drawings date back to 1583-1591
  - Material ranked as having unique national significance
  - Part of a larger archival unit (TIFF, metadata)

Towards Holistic Passive Preservation

DILCISBoard/**SIARD**

SIARD (Software Independent Archiving of Relational Databases) - an open file format for the long-term archiving of relational databases



Synthetic DNA

Analog bootstrap

Solve format obsolescence with standards

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 enables the use of DNA as a digital media
- **End-to-end passive preservation is feasible**
  - *Standard file formats (SIARD)* to solve format obsolescence
  - *Synthetic DNA*: High-density, decay-free digital archival media
  - *Analog media + emulation*: Bootstrap for archiving DNA decoders

UAG
UGA
UAA