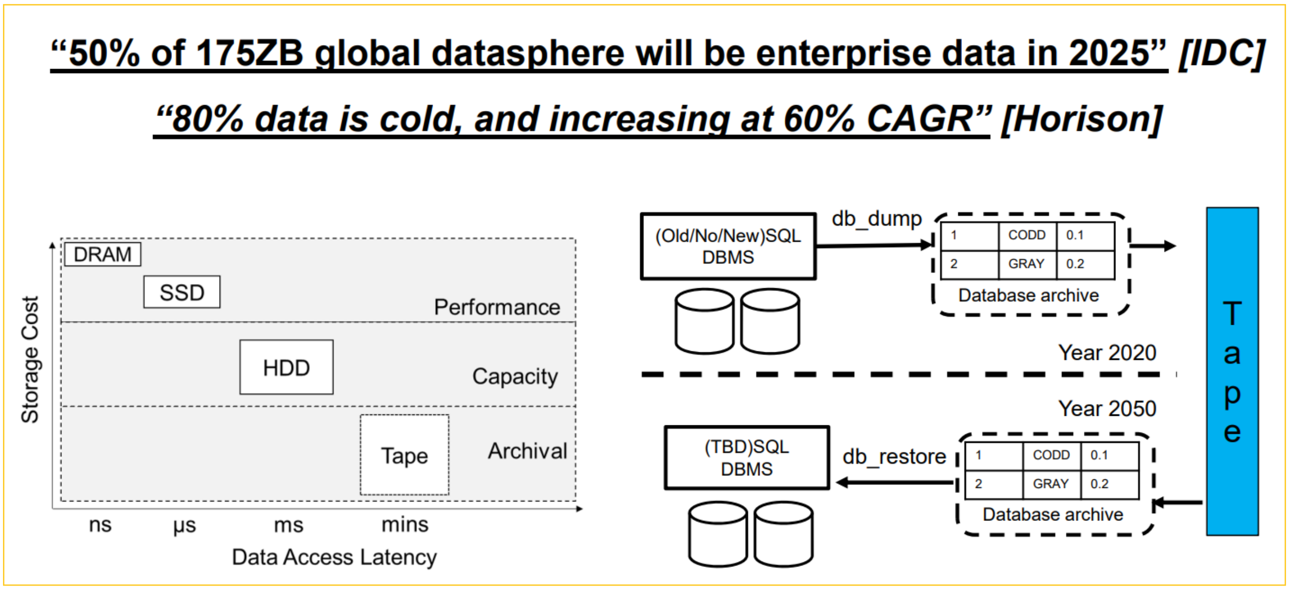
Storing and reviving databases on DNA

Raja Appuswamy (EURECOM research institute, Sophia Antipolis, France)

Raja Appuswamy is currently working as an Assistant Professor at EURECOM. Previously, he worked as a Visiting Professor at EPFL, as a Visiting Researcher in the Systems and Networking group at Microsoft Research, Cambridge, and as a Software Development Engineer in the Windows 7 kernel team at Microsoft, Redmond. He received his PhD in Computer Science from the Vrije Universiteit, Amsterdam, where he worked under the guidance of Prof. Andrew S. Tanenbaum on designing and implementing a new storage stack for the MINIX 3 microkernel operating system. He also holds dual Masters degrees in Computer Science and Agricultural Engineering from the University of Florida.

Appuswamy cited recent studies showing that nearly 80 percent of all stored data are 'cold' (infrequently accessed) and this number is increasing over time (Memishi et al. 2019). Data often needs to be stored and kept for legal compliance reasons, typically on magnetic tape. There are problems with these tapes, because tape vendors typically only support backwards compatibility for two generations. This poses problems if large collections need to be moved to newer tape formats, which hold especially true for audiovisual data (Perlmutter 2021).



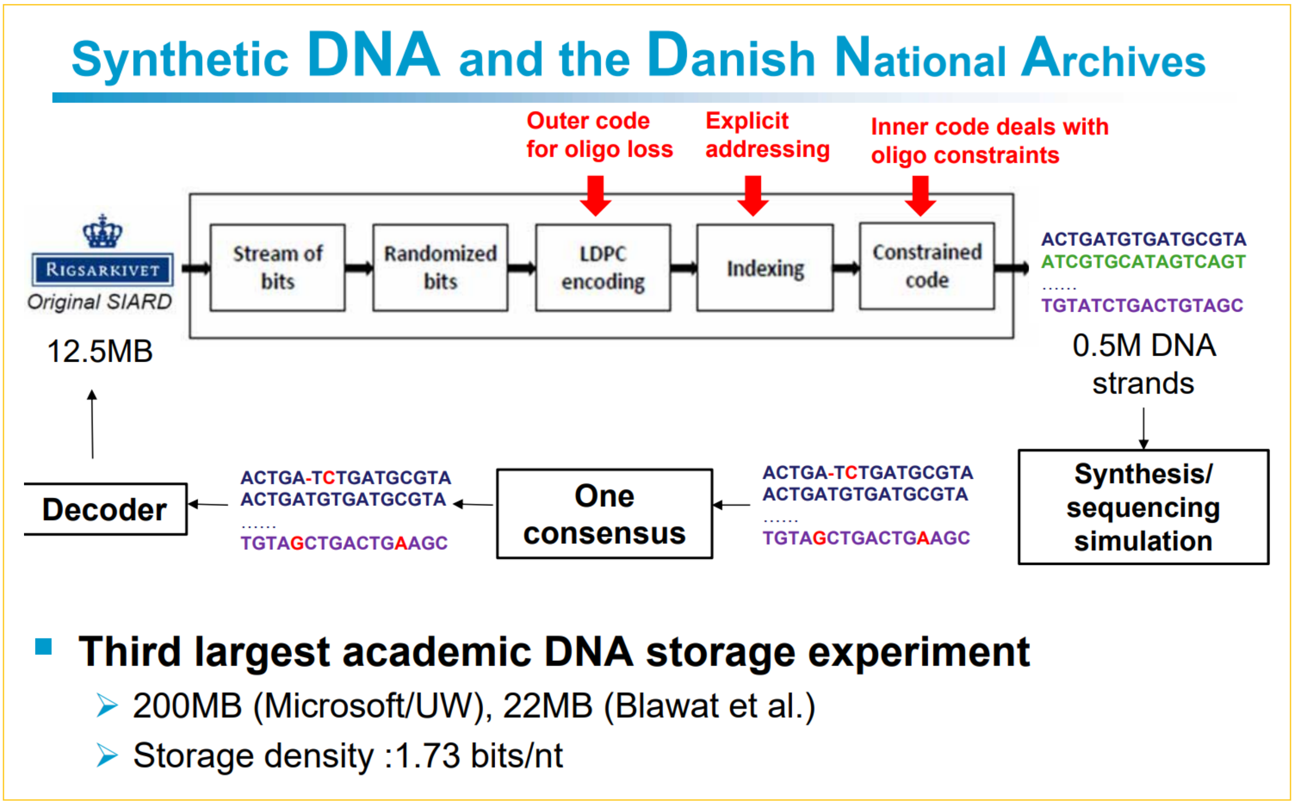
"50% of 175ZB global datasphere will be enterprise data in 2025" *[IDC]*

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

1	CODD	0.1
2	GRAY	0.2

1	CODD	0.1
2	GRAY	0.2

Figure 6: Scenarios on growth of data and proportion of "cold" data.



Synthetic DNA and the Danish National Archives

Outer code for oligo loss

Explicit addressing

Inner code deals with oligo constraints

RIGSARKIVET

Original SIARD

12.5MB

Stream of bits

Randomized bits

LDPC encoding

Indexing

Constrained code

ACTGATGTGATGCGTA
ATCGTGCATAGTCAGT
......
TGTATCTGACTGTAGC

0.5M DNA strands

Synthesis/ sequencing simulation

ACTGA-TCTGATGCGTA
ACTGATGTGATGCGTA
......
TGTAGCTGACTGAAGC

One consensus

ACTGA-TCTGATGCGTA
ACTGATGTGATGCGTA
......
TGTAGCTGACTGAAGC

Decoder

- Third largest academic DNA storage experiment
  - 200MB (Microsoft/UW), 22MB (Blawat et al.)
  - Storage density :1.73 bits/nt

Figure 7: The DNA4DNA collaboration

Magnetic media is only one possibility for storing data. DNA (desoxyribonucleic acid) is another option. It is a molecule made from four different nucleotides which can be sorted into one big string. DNA storage refers to single strand DNA manufactured and stored outside the body – no living beings are involved.

Sequencing technology to read the DNA is available. Storage density on DNA is much higher than on tape. DNA is also very durable, can survive for a long time (thousands of years) and in harsh conditions. The project Oligoarchive focuses on using DNA as an intelligent storage medium. The project is looking at how DNA can be used to store files.

There are economic limits today, as writing DNA is very costly, while reading devices are already becoming affordable. In spite of this, the application of the concept to databases has been tested. The Danish National Archives have encoded a 12.5 megabyte SIARD database file to DNA and reread it successfully (Marinelli et al. 2021).

The OligoArchives project cooperates with the company EUPALIA that has long experience with emulation and traditional persistent data carriers like microfilm.

There are some complications.

- The need to encode the data. The data in the DNA needs to store metadata too. The metadata and the data can be decoded into the data for the database. DNA synthesis is very expensive (10^7 times more expensive than tape) currently. DNA does not solve media or format obsolescence.
- The need to preserve the decoder (and its mechanisms) by self-explaining instructions.
- For decoding a DNA data carrier, you have to inject a liquid and by that destroy it. DNA today is a write-once, read-once medium. This shortcoming can be overcome by keeping several samples of a data object (tiny physical particles) in one data container and by re-writing new copies at certain intervals.

Questions and discussion

- The strong need for a comprehensive way of declaring the content and meaning of an arbitrary byte sequence extracted from DNA might induce the Big Data industry to work on improvements in the database archiving sector.
- Having “bootstrap routines” that help reviving data out of its own metadata might become commercially important.
- DNA storage researchers thus welcome academic and industry people who want to get involved.