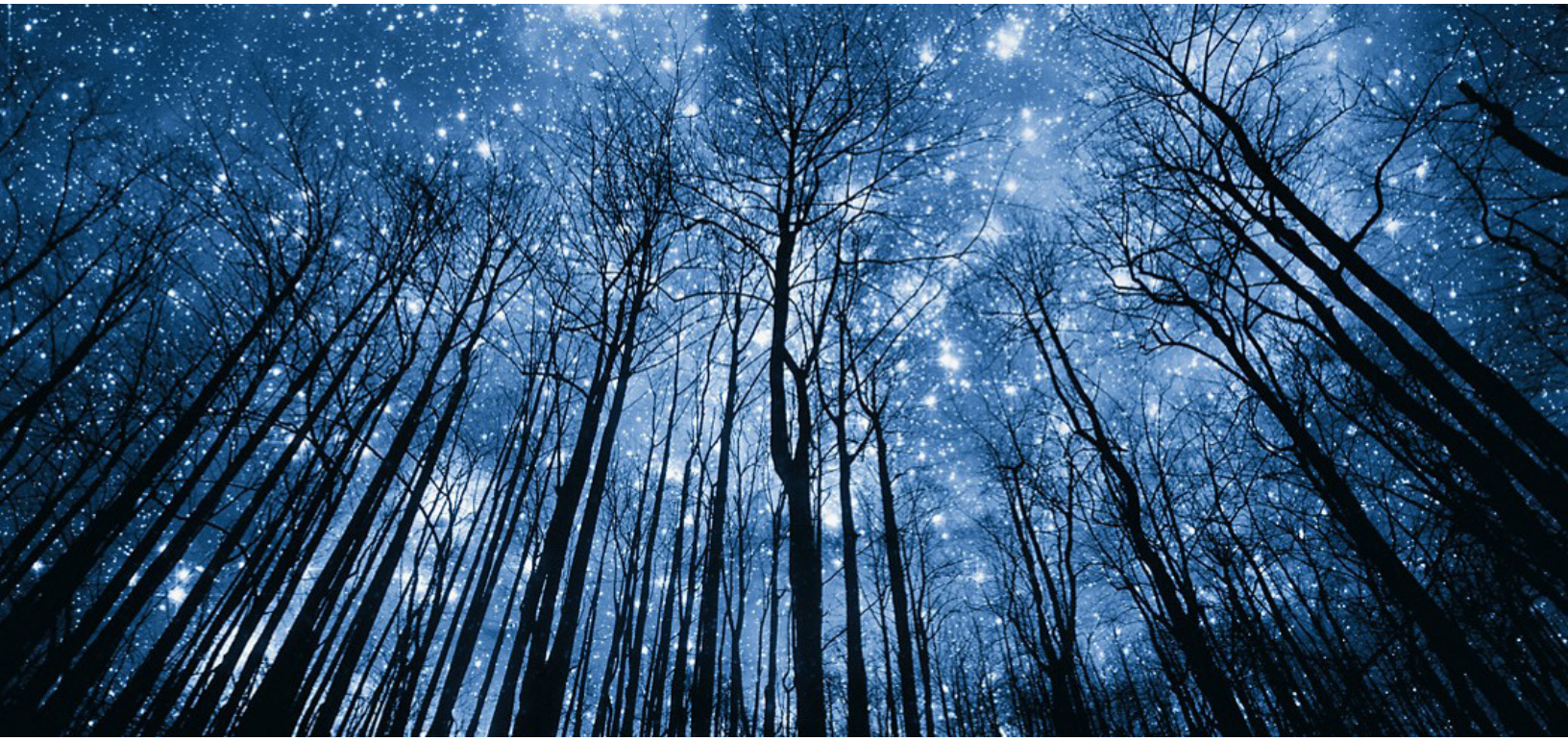
DELL Technologies

DNA: FUTURE OF DATA STORAGE



Rekha Sivakumar
Specialist 2, Inside Product
Dell Technologies

Sreejith S. Pai
Specialist 2, Inside Product
Dell Technologies

Adithyan Rajeev
Specialist 2, Inside Product
Dell Technologies

DELL Technologies
Proven Professional

Knowledge Sharing Article

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged and Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers. Learn more at www.dell.com/certification

Table of Contents

1. Introduction

The population of the world is at 8 billion and is constantly growing. With more people beginning to work from home and remote locations, technologies like Internet of Things (IoT), Artificial Intelligence (AI), edge computing have become a part of our day-to-day lives. Usage of these technologies has resulted in data explosion. With each person generating about 1.7MB of data per second on an average [1], it is estimated by IDC that there would be approximately 175 Zetta Bytes (ZB) of data as of 2025 [2]. This presents a new problem – how do we store all this data?

We have seen an evolution in storage technologies right from punch cards to powerful high density NVMe drives but the constant growth of data creates a requirement for more optimized modes of storage. Over the past decades, researchers and scientists have worked to create a method of storing data on a medium that is more dense, robust, and durable. However, our modern storage techniques have a few flaws: they are not robust, have a low information density and each media requires a special device to read and write data. This article aims at exploring the potential of an unconventional method of storing data, using the capabilities of Deoxyribose Nucleic Acid (DNA).

2. History of Data storage

When we look back to about 64,000 years ago, the only modes of data storage humans had, were paintings and carvings to store and convey information to others. A writing script was not developed at that time. Gradually over time, scriptures were developed, and data began to be stored in the written format. Over time, about 1000 years ago, printing technology was invented and soon it was possible for data to be printed onto paper. But over time the storage requirement became very high, and the technology of that time was limited and inconvenient.

Gradually, a few hundred years ago, scientists and inventors started finding ways to store data digitally which were far more convenient and much more reliable than the conventional methods. Various developments and technology have risen in the past few decades and some of the advancements have been listed below:

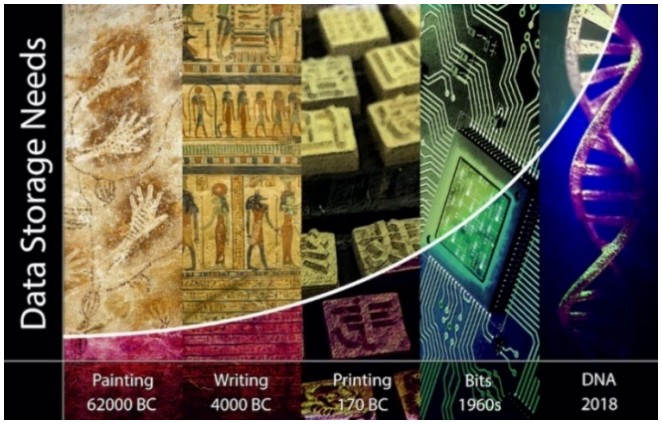


Fig 1: Growth of Data Storage in history
Image source: https://wyss.harvard.edu/technology/dna-data-storage/

1890: Punched Cards & Punched Paper Tapes
Punched cards and punched paper tapes are the original storage form of mechanized information. In 1890, Herman Hollerith, an American statistician invented the punch-card tabulating machine that can record up to 960 bits for collecting and counting census data. The beginning of the era of semi-automatic data processing systems was marked there. Punched cards and punched paper tapes are used in punch card tabulation machines as input and output devices of early computers. Programs and data were converted into binary codes. Holes were considered as 1 and non-holes were considered as 0, and then input into the computer through photoelectric scanning [3].

1951: Magnetic Tape-The Beginning of The Era of Magnetic Storage
Fritz Pfleumer, a German Engineer invented the audio tape which could store analog signals. Magnetic tapes were prepared by gluing crushed magnetic particles to a paper strip. 1951 was the year in which magnetic tapes were used for computer storage. 10,000 punched paper cards were equivalent to 1 magnetic tape [3].

1956: The Hard Disk Drive
The first hard disk ever designed was the IBM Model 350 Disk File in the year 1956. The Disk drive was designed for the IBM 305 RAMAC computers. It had a total storage capacity of 5 million characters of data which were stored in 50 quantities of 24-inch magnetically coated metal discs [4].

1978: LaserDisc
The LaserDisc (LD) is an optical disc storage technology that was introduced as "Discovision" and co-owned by MCA and Philips. Though it was better in terms of audio and video quality when compared to its competitors, the major drawback with LD was its high price. The LaserDisc remained significant for a good part of the 1980s and set the stage for the CD and DVD which became highly popular storage format a few years later [4].

1992: SSD module
SanDisk manufactured the commercial flash-based Solid-State Drive module for laptops and computers. They used non-volatile memory chips to replace the existing spinning disks. Later in 1998, they introduced the 2.5-inch, and the 3.5-inch form factor drives for PATA interface devices [4].

2000: USB Flash drives
USB Flash drives consist of flash memory with an USB interface. They can be used to store, back up or transfer of files between devices. USB drives came with many advantages – they had a small factor, were faster and had higher storage capacity. Owing to their affordable prices and higher resiliency, PC and laptop manufactures started to include USB ports instead of floppy and CD slots in their devices [4].

2006: Storage in The Cloud
'As-a-service' offering for data storage picked up pace in the late 2000s with Amazon Web Services launching AWS S3. It became possible to store data digitally in the form of logical pools which can be present off-site with AWS (i.e., the cloud service provider) being responsible for accessibility and availability of the data. Main advantages of the cloud-based model include lesser time spent on maintenance and lower capital expenditure [4].

2011: Non-Volatile Media express (NVMe)
NVMe is a new storage access and transport protocol for systems that utilize SSDs. It has a faster response time and delivers a higher throughput. NVMe addresses the needs of Enterprise and Client systems that utilize PCI Express based solid-state storage. NVM Express (NVMe) is a host controller interface which is optimized and provides high-performance and scalability. NVMe provides efficient access to storage devices which have non-volatile memory including current technologies like NAND flash to memory technologies of the future [5].

After many such developments in technology, we can store up to 4 Terabytes of data easily on a single SD card which is so compact that it can be held on our fingertips. However, to keep up with the rapid data growth, we need better and more robust storage methods. The latest advancements in the field of data storage have led us to one of the oldest forms of storage - the DNA.

3. What is DNA?

Deoxyribonucleic acid is basically a hereditary material consisting of a group of molecules carrying genetic information about a living organism. It is composed of two polynucleotide chains that coil around each other to form a double helix and has genetic instructions for the biological processes of all known organisms and many viruses.

These two DNA strands are called polynucleotides. They are made up of monomeric units known as nucleotides. Each nucleotide is composed of one of four nucleobases (adenine [A] thymine [T], cytosine [C] or guanine [G]), and a sugar phosphate backbone [6].

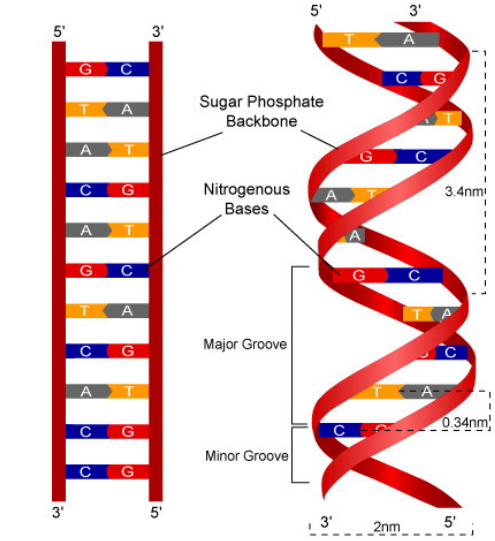


Fig 2: DNA Structure and Bases
Image source: https://www2.nau.edu/lrm22/lessons/dna_notes/dna_notes.html

DNA provides the genetic code for all living things on the planet and hence is considered the basic element of life. The genetic code in DNA has the capability to store base information that has been passed down through many generations. DNA is also known to be at least 1000-fold denser than the most compact solid-state hard drives [7]. As a result, DNA is a very efficient storage medium that has been developed over the years to store significant amounts of data. The small size and high-density features of DNA makes it an effective storage solution. Approximately, 1 gram of dry DNA can easily store about 455 exabytes of raw data [8].

4. Advantages of DNA for Data Storage

As the DNA has half-life of more than 500 years, the data stored on DNA strand won't degrade over time or become obsolete like conventional cassette tapes or hard drives. The information stored on a DNA molecule can also be replicated, compressed, and secured similar to that of any other existing traditional storage device. Various research has shown that if maintained under favorable conditions, it is possible to retain information stored on DNA molecules for many centuries [8].

5. How DNA can be used to store data

We know the DNA spiral constitutes of alternating pairs of four bases - adenine, guanine, cytosine, and thymine. The data can be encoded into these nucleobases (A, C, G, T). The four-lettered alphabet (nucleobases) of DNA can thus be mapped to a code. As data is artificially infused into DNA, DNA data storage medium does not create nor require any living entities.

As an alternative to storing bits of data on silicon or magnetic devices using optical or electromagnetic mechanisms; the DNA data storage uses simple chemical processes to infuse digital data into the DNA molecules. Since the DNA molecules are artificially created from scratch—it has been discovered that it is possible to write or specify long sequences of information on DNA strands using the encoded letters A, C, G and T (or their combinations) and then read those sequences back from the nucleobases. This process is equivalent to that of a computer storing its data into binary format and from there, it is a short conceptual step of encoding the data into a molecule.
DNA storage consists of the following processes: coding the data, DNA synthesis, DNA sequencing, and decoding it.

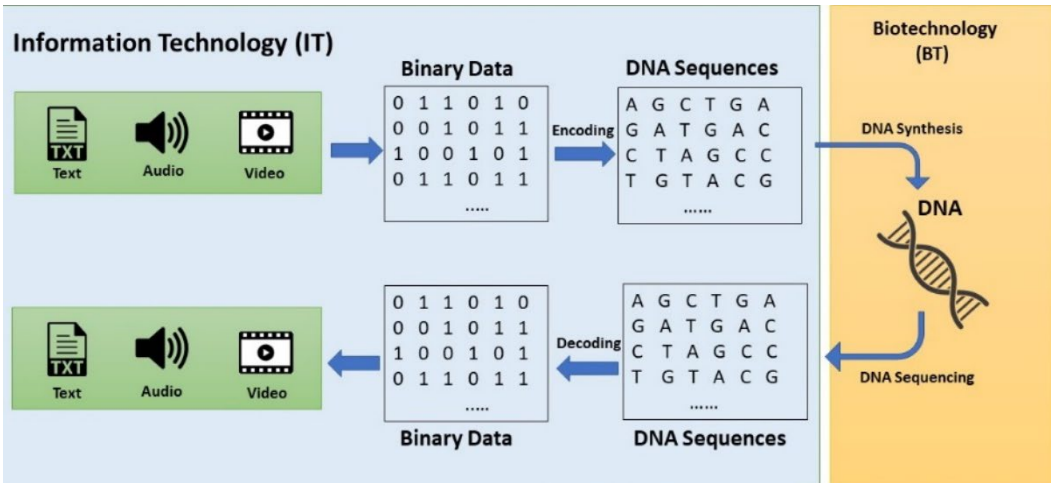


Fig 3: DNA Data storage process
Image source: https://link.springer.com/article/10.1007/s42514-022-00094-z/figures/1

5.1. Encoding and decoding

There are many methods for encoding text on to a DNA Strand. Most of these involve mapping each letter of a sequence into a corresponding sequence of nucleotides in a lookup table. This unique sequence is called a *codon*.

5.1.1. Encoding arbitrary data

Firstly, the data is usually converted into binary (base 2), or ternary (base 3) data as done by computers. These converted digits are called bits or trits respectively.

a) Bits to nucleotides

The binary encoding is based on simple mappings. Since DNA is composed of 4 nucleotides (Adenine, Cytosine, Guanine, Thymine), each nucleotide is assigned to a two-bit. A sample lookup table would be:

Table 1.1: DNA Bases to Binary lookup table

Nucleotide	Binary
A	00
T	01
G	10
C	11

For example, the number 3 would first be converted to binary resulting in the bits 11. This would directly map to the nucleotide “C”. Similarly, the alphabets would be mapped to the respective ASCII values. For example, the letter “R” with ASCII value as 82, would be represented as 01010010 in the binary form. This would be mapped as combination of nucleotides - “TTAG”.

b) Trits to nucleotides

Like binary methods, trits are converted to a nucleotide using a lookup table. However, repeating nucleotides also called homopolymers, can result in errors when determining the accurate sequences. To prevent this, the lookup table is formatted such that the result also depends on the preceding nucleotide.

Table 1.2: Trits to DNA Bases lookup table

Previous	0	1	2
A	C	G	T
T	A	C	G
G	T	A	C
C	G	T	A

For example, considering Table 1.2, if the previous nucleotide in the sequence is A (adenine), and the trit is 0, the next nucleotide will be C (Cytosine).

5.2. Synthesis

Once the data is encoded, the next step is to synthesize DNA strands. DNA synthesis is the process of creating DNA molecules. While DNA synthesis is a naturally occurring process in all eukaryotes and prokaryotes, as well as some viruses, it is possible to create DNA artificially too. However, in order to prevent DNA mutations, precise DNA synthesis is crucial. DNA synthesis and the *in vitro* (occurring outside a living organism, typically in an artificial environment) options involved have been the subject of numerous research over the years. The most commonly used synthesis process are Artificial Gene Synthesis and Oligonucleotide synthesis.

5.3. Sequencing

DNA sequencing is a technique that is used in labs and clinics to determine the accurate sequence of the DNA molecule (nucleotides). The sequence of the four bases (A, C, T, G) encodes the information that is stored in the DNA. DNA data post synthesis can be selectively retrieved from the source DNA fragment in a simple process called random access. The process of random access utilizes PCR-based enhancement of the DNA source using the primer pairs that link to the data generated during the encoding process. The automated sequencing instruments are then used to read the respective DNA molecules and are then decoded back to the format used in the original data.

There has been a lot of research and development in the field of genome sequencing and various techniques have been introduced in the industry, each of which has its own process and characteristics. The common sequencing techniques that are presently used in the industry are Sanger Sequencing and Massively parallel DNA sequencing. Recent developments have also introduced a sequencing method called Nanopore DNA sequencing which reads individual DNA strands that are passed through a tiny pore within the protein membrane [9].

5.4. DNA Data Protection

For example, if it is required to store 3 numbers: 7, 19 and 17. It should be possible to recover any of the three of them, even if one of them is corrupted. To do that, we store the sum of all three as well (7 + 19 + 17 = 43). This is also called as the parity number. Say the number 17 gets corrupted, we can recalculate it by subtracting the sum of the other numbers from the parity (43-(7+19)). This is an example of an error correcting code.
Similarly, by adding fountain codes (i.e., parity) to a DNA encoded data strand, we can ensure that it can be read back even if some A, T, C or Gs can't be read correctly.

5.4.1. High Availability

Whenever a data is stored, it should be available in case of a disaster or loss of data. The DNA having a double-stranded molecule helps to protect the genetic code from damage. One strand can serve as a template for repair in case the other strand gets damaged or broken. For additional protection against enzymatic attack, there are proteins present around the DNA strands.

5.4.2. DNA Replication

DNA replication is the process of producing two identical replicas of DNA from an existing DNA molecule. DNA replication is a naturally occurring process in all living organisms and plays an important role in genetic inheritance. It is also possible to perform DNA replication artificially (in vitro, outside a living organism). DNA contains DNA polymerases whose main function is to replicate DNA content during cell division. Such isolated enzymes and artificial DNA primers can be used to initiate DNA synthesis in a sample DNA molecule. Ligase Chain Reaction (LCR), Transcription-Mediated Amplification (TMA) and Polymerase chain reaction (PCR) are a few examples of DNA replication methodologies.

Traditionally with hard drives, every drive has to be read and every single bit has to be copied to another one. DNA on the other hand, can easily be replicated millions of times using any of the above-mentioned reaction methods which uses synthetic DNA as data storage. The only downside to this is that, while copying DNA there is usually addition of some noise and the quality is reduced. But this can be overcome by utilizing the error-correcting codes. Ultimately, even if one copy is corrupted or lost, there would still be plenty of copies left.

5.5. DNA of Things (DoT)

In DNA of Things, we take the DNA and fuse it to a functional material, and thereby create an object that will have a property with immutable memory. The concept of DoT can thus help to store crucial and robust information into day-to-day materials that we use. In the experiment mentioned in article [10], the researchers have taken a video sample and tried to encode the data onto a DNA molecule in a manner that it is robust to errors. This was achieved by encapsulating the DNA molecule in silica nanoparticles, thereby forming silica particle-encapsulated DNA (SPED) [11]. This SPED particle helps to maintain the quality of DNA and therefore not degrade it when we mix with any functional material. The encapsulated DNA was then mixed with polymethyl methacrylate (PMMA), a widely used transparent plastic compound. It was then shaped into a lens and mounted on a glass frame. This hid the DNA information on the lens but ensured that it could be retrieved from just a fragment of the lens. This experiment highlights the capabilities of DNA as a data storage medium, which enables a broad range of objects, from buttons to keys, to help securely store and carry data.

6. DNA Data Storage: use cases and developments

DNA Data Storage has been a topic of interest for many researchers over the years and there are enough success stories to demonstrate its potential. Some of these are listed below:
In 2012, scientists have been able to successfully encode an audio clip of Martin Luther King Jr.'s famous speech "I Have a Dream", a copy of Francis Crick and James Watson's scientific paper "double helix" from 1953 and all of Shakespeare's 154 sonnets on DNA. They later were able to retrieve the information with 99.99% accuracy [12].

Another example of DNA data storage was in 2015, when Nick Goldman from the European Bioinformatics Institute (EBI) announced the Davos Bitcoin Challenge. During the presentation, DNA tubes were handed out to the audience, with a message that each tube contained the passkey for one bitcoin, all coded in DNA. The challenge was to sequence and decode the DNA within 3 years and the one to do so could claim the bitcoin. In 2018, it was announced that a Belgian PhD student, Sander Wuyts, had successfully completed the challenge and retrieved the instructions to claim the bitcoin [13]. Later on, in 2016, researchers from Microsoft and University of Washington were able to use binary coding to encode 200MB of data which even included Universal declaration of human rights into DNA strands [14].

7. Challenges of DNA Data Storage

Despite the advances and success, one of the main disadvantages with DNA data storage is the cost. Apart from the cost, current chemical synthesis methods still have drawbacks like synthetic errors and production toxic by-products. DNA data storage also has environmental impacts - inefficient and harmful production methods are against green and clean data storage. Another issue is that encoding data in DNA is usually a time-consuming process, which is much slower than timescales in a silicon memory chip. But with the evolution of technology, there has been progress and the ability to store data into DNA at megabit per second write speeds would be possible in the future.

8. Conclusion

While it is clear that plenty of challenges remain before DNA storage could become a cheap and reliable commercial option, DNA data storage certainly can be a potential solution to today's storage problems. Some companies who want to preserve extensive archives of information that are not accessed frequently are already using DNA Storage. Unfortunately, it will be quite some time before DNA storage becomes a common and affordable storage option available in the market. And when it does, it will allow us to store incredible amounts of data in a very small space and we'll able to archive data for generations to come. In the meantime, we should carefully pick the best storage solutions for reliable long-term data storage.

9. References

[1] Bernard Marr & Co. - https://bernardmarr.com/how-much-data-is-there-in-the-world/

[2] Gartner - The Digitization of the World From Edge to Core by David Reinsel, John Gantz, John Rydning, An IDC White Paper – #US44413318

[3] Utmel: History of storage-https://www.utmel.com/blog/categories/memory%20chip/the-evolution-history-of-storage-devices

[4] Computer History Museum - https://www.computerhistory.org/timeline/memory-storage/#169ebbe2ad45559efbc6eb357207cbc8

[5] NVMe - https://nvmexpress.org/

[6] DNA - https://medlineplus.gov/genetics/understanding/basics/dna/

[7] DNA Data Storage - DNA Data Storage (harvard.edu)

[8] DNA Storage capacity - DNA Storage Record Broken: 1 Gram Could Hold As Much as 455 Exabytes (scitechdaily.com)

[9] DNA sequencing - https://www.genome.gov/genetics-glossary/DNA-Sequencing

[10] Koch, J., Gantenbein, S., Masania, K., Stark, W.J., Erlich, Y., & Grass, R.N. (2019). A DNA-of-things storage architecture to create materials with embedded memory. *Nature Biotechnology, 38*, 39-43.

[11] Paunescu, D., Puddu, M., Soellner, J. O. B., Stoessel, P. R. & Grass, R. N. Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA 'fossils'. Nat. Protoc. 8, 2440–2448 (2013).

[12] National Geographic DNA data storage - https://www.nationalgeographic.com/science/article/shakespeares-sonnets-and-mlks-speech-stored-in-dna-speck

[13] Microsoft: The AI Blog- https://blogs.microsoft.com/ai/synthetic-dna-storage-milestone/

[14] EMBL-EBI Communications - https://www.ebi.ac.uk/about/news/announcements/belgian-phd-student-decodes-dna-wins-bitcoin/