CATALOG

Building the First Integrated DNA Storage and Compute Platform

“The biological example of writing information on a small scale has inspired me to think of something that should be possible. Biology is not simply writing information; it is doing something about it”*

-- Richard Feynman, *There’s Plenty of Room at the Bottom*

*”And that something is compute”
CATALOG

Presenter Notes
2022-03-08 18:24:09

Here’s a couple quotes to motivate the work we do at CATALOG. The first, from Bill Gates … it’s a one you have probably heard before. DNA is a digital sequence of symbols that exists in every living cell and stores genetic information. Not only does it store the information but encodes the machinery that process that information and creates function. The second, from Richard Feynman, is more of a call to action.
it is from his seminal lecture
There’s Plenty of Room at the Bottom
In the lecture he discussed the possibility of manipulating individual atoms to store information
he urges people to consider biological systems as a source of inspiration for what should be possible to build, not just for storing data, but for executing programs. This sort of thinking is what guides our work at CATALOG

History of data storage in DNA

(Mikhail Neiman, 1964)

"The biophysical information systems and processes open favorable prospects in the direction on microminiaturisation of information storage and processing devices. These processes are, in particular, in the recording of the hereditary information in single-chain polymer molecules of DNA."

First artificial data stored: **(Joe Davis, 1988) Designed and synthesized an 18-bp message and transformed into *E. coli***

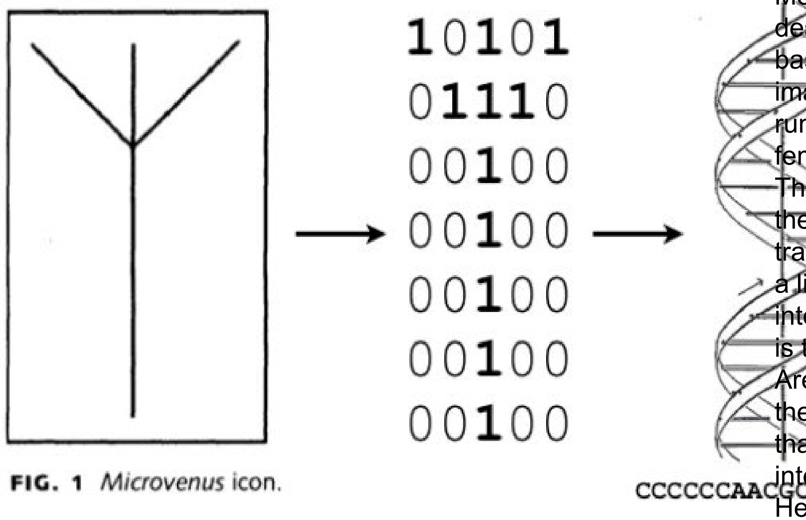


FIG. 1 *Microvenus* icon.

Agapakis, *Scientific American*, 2012

Presenter Notes
2022-03-08 18:24:09

The first record of using DNA for information technology dates back to a paper from a Russian physicist, Michael Neiman, in 1964. This is just about 8 years after the structure of DNA was discovered. But ofcourse, the molecular tools didn't exist to put this theory into practice.

The first one to do it, as far as I can find, is Joe Davis, an artist collaborating with molecular biologist Dana Boyd in Jon Beckwith's lab at Harvard Medical School. In 1988 he designed and synthesized an 18 base-pair message encoding the image of the ancient Germanic rune representing life and the female earth.
The Microvenus message was then pasted into a vector and transformed into E. coli, creating a living work of art. A little interesting tidbit about this work is that it was inspired by the Arecibo telescope message, and there is this interesting article that talks about the artist's intentions if you care to visit it.
Here the image is encoded like asci text, where each pixel is a bit. there are 35 bits which you factor into primes to create the two dimensional image. Then those 35 bits are encoded into DNA using run length encoding. CTAG = run lengths of 1, 2, 3, and 4 respectively.

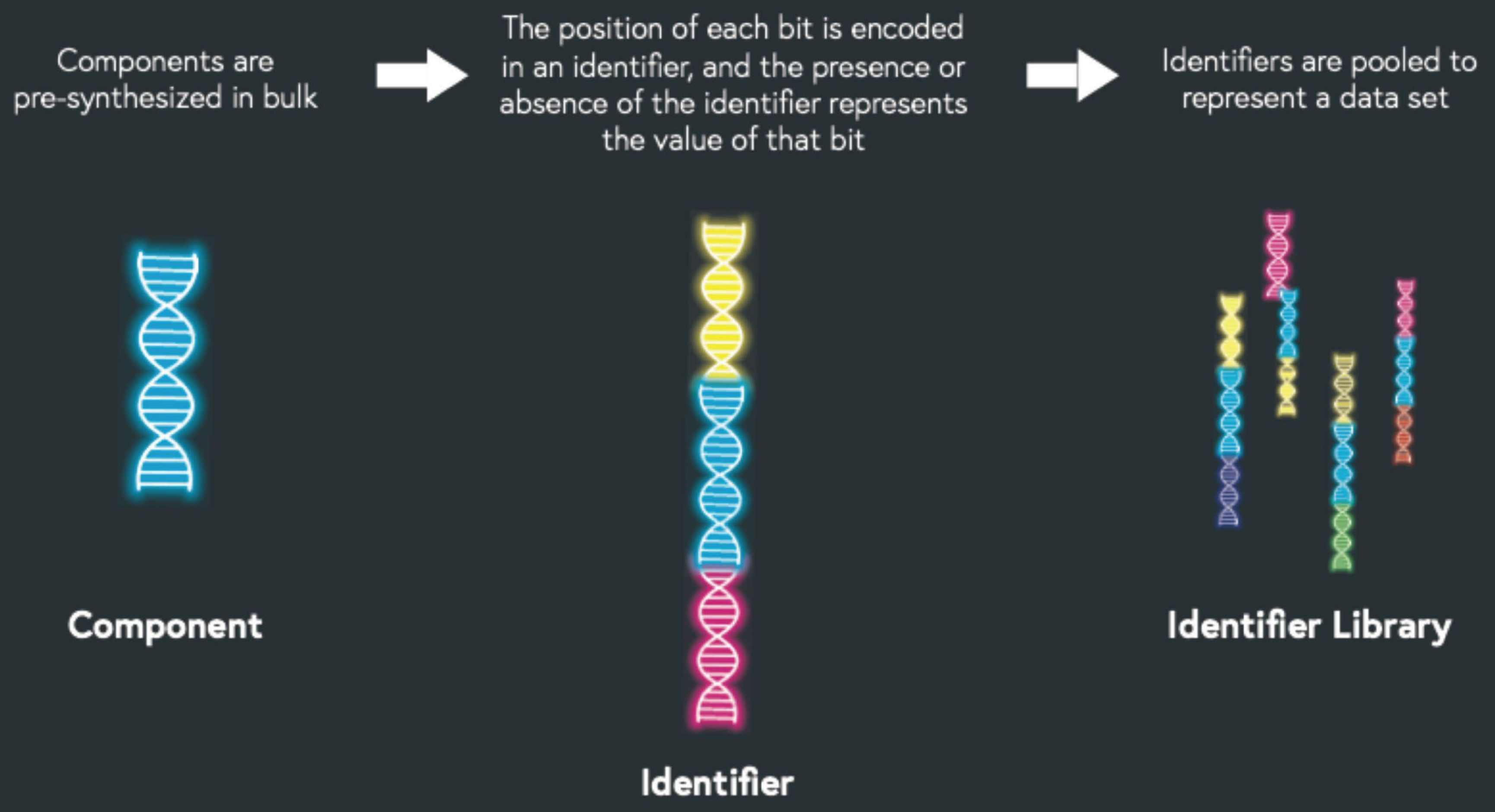
Our writing technology is similar movable type, we are re-using DNA oligos in a combinatorial assembly process to synthesize a large number of unique DNA molecules.

This approach dramatically reduces cost and increases speed of our DNA synthesis.

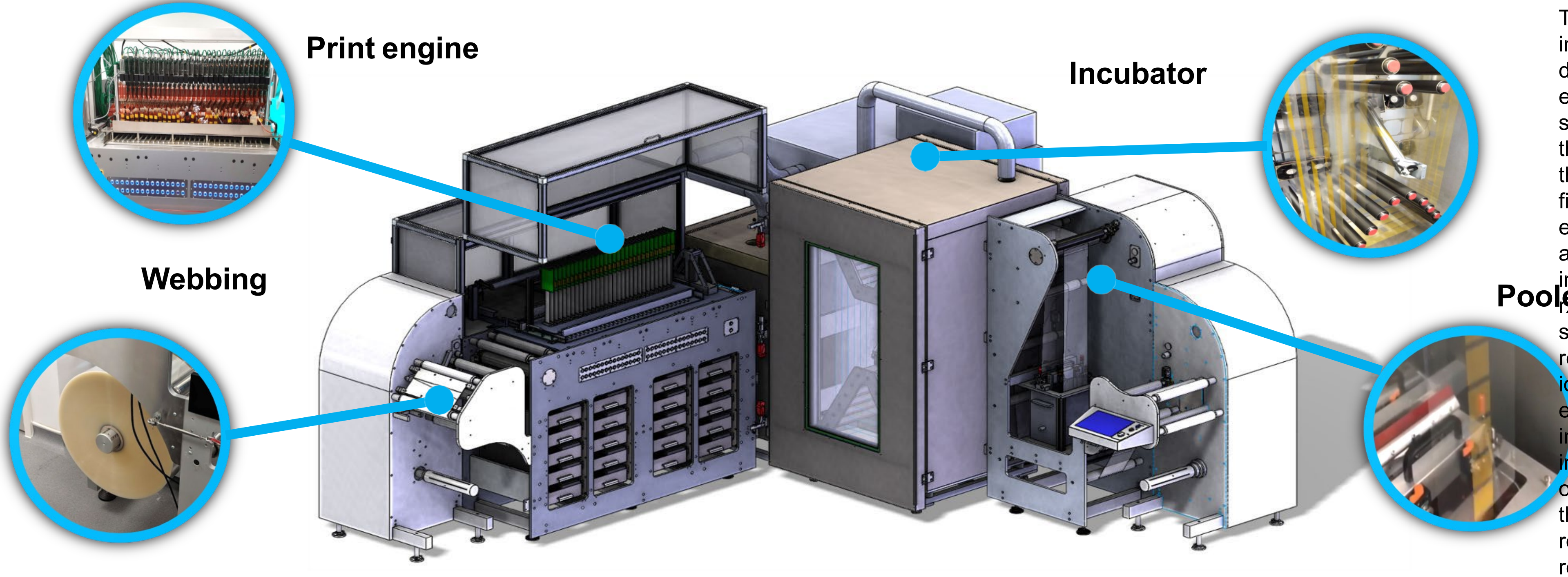
Presenter Notes
_2022-03-08 18:24:11_

The predominant method for synthesizing DNA uses phorphoramidate chemistry. In phorphoramidate chemistry, single nucleotides are added sequentially through a multi-step organic synthesis process. Although this process was developed 50 years ago, the time required for a single cycle remains around 2 minutes. This is ~5 orders of magnitude away from the speed needed to be competitive in the storage market. Now, 8 years later, this approach is still considered state of the art but continues to have limitations. CATALOG's method sidesteps the primary bottlenecks by using prefabricated DNA molecules, it is faster because DNA is available immediately and it is cost effective because DNA is ordered in large quantities.
In our encoding scheme, each bit in a bit string is mapped to a unique DNA molecule and the bit value "1" or "0" is determined by the molecule's presence or absence. We overcome the synthesis limitations (speed and cost) that other DNA-based storage groups have by using pre-fabricated DNA parts. This is similar to the concept of movable type, where movable components are re-used to make words. Our approach allows us to synthesize a large number of unique DNA molecules using a combinatorial assembly process.

Moveable type encoding unlocks write speed and compute



Shannon: Mbps write speed instrument



Collocate components → Add assembly mix → incubate → pool

Presenter Notes
2022-03-08 18:24:12

We've developed a custom instrument that allows us to do this. There are 4 modules to the system. Chassis provides the substrate, a slightly hydrophobic polypropylene webbing, that traverses the entire instrument. This substrate is where DNA is dispensed and reaction drop/spots are formed. The Print Engine is an array of industrial inkjet printheads that dispense pL size volumes of each DNA components in specific locations that result in the creation of ligation reactions that assemble identifiers. The final printhead is reserved for an enzyme that catalyzes the assembly reaction. The instrument creates about 500K rxns/second, and we've done studies showing that each reaction can assemble up to 32 identifiers. For appropriate environmental conditions, an incubator is used. In the incubator, the reactions located on the webbing or substrate, are threaded through a series of rollers that extend the time for reactions to remain in this environment and eventually empty into a basin that pools all Identifiers. This is the instrument that we've previously mentioned being shipped from the UK and is currently being re-built. It brought us from off-the-shelf instrumentation which was much less than Kb/s write speeds to Mb/s write speeds which is where we are currently at, and

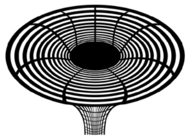
Computing on DNA Encoded Data

- **Two Critical Aspects of DNA for Computing**
  - **Random Access**
  - **Massive Parallelism (courtesy of easy replication)**

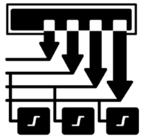
- **Value Based Search**
  - **Lower cost of retrieval (time invariant with data volume)**
- **Other Types of Chemical Instructions**
  - **New algorithms**
- **Storage Becomes “Active”**
  - **Compute and storage merged into single platform**

CATALOG is building the world's first DNA-based Data Storage and Computing Platform

Limitless Storage



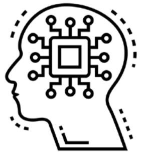
Hyper Dense: 1,000,000x denser than SSD (solid state drive)



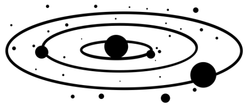
Massive redundancy: DNA is easily replicable into multitudes of copies for simultaneous computing/query

Ultra-persistent: Stable for 1000s of years – once archived will last forever

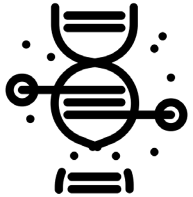
Limitless Compute



In-Storage: Compute directly on stored data without costly movement between memory tiers



Scale-Free: The time and cost required to process a GB of data in DNA is the same as that required to process a PB of data



DNA-Native: Rely on structure / physical properties inherent in DNA to perform unique computing operations