# Getting Polymers to Tell a Story: Scaling up DNA Data Storage and Functionality

Luis Ceze and Karin Strauss
Molecular Information Systems Lab

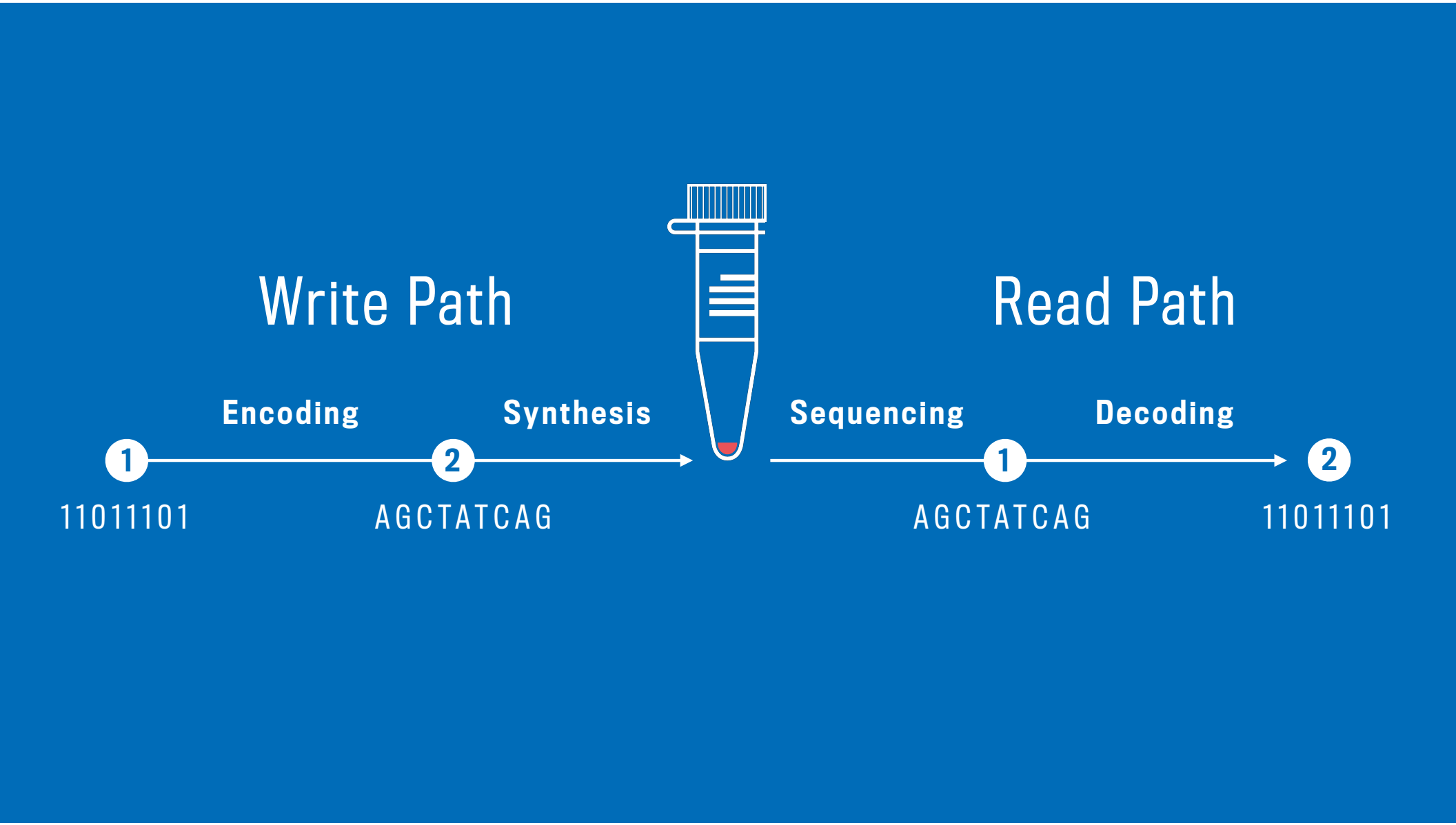
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Microsoft

joint work with Georg Seelig, Doug Carmean, Sergey Yekhanin, Lee Organick, Yuan-Jyue Chen, Bichlien Nguyen, Chris Takahashi, Ashley Stephenson, Pranav Vaid, Sharon Newmann, Cyrus Rashtchian, Miklos Racz, Siena Ang, David Ward, Randolph Lopez, Max Willsey, Kendall Stewart, James Bornholt, Rob Carlson, Hsing-Yeh Parker.

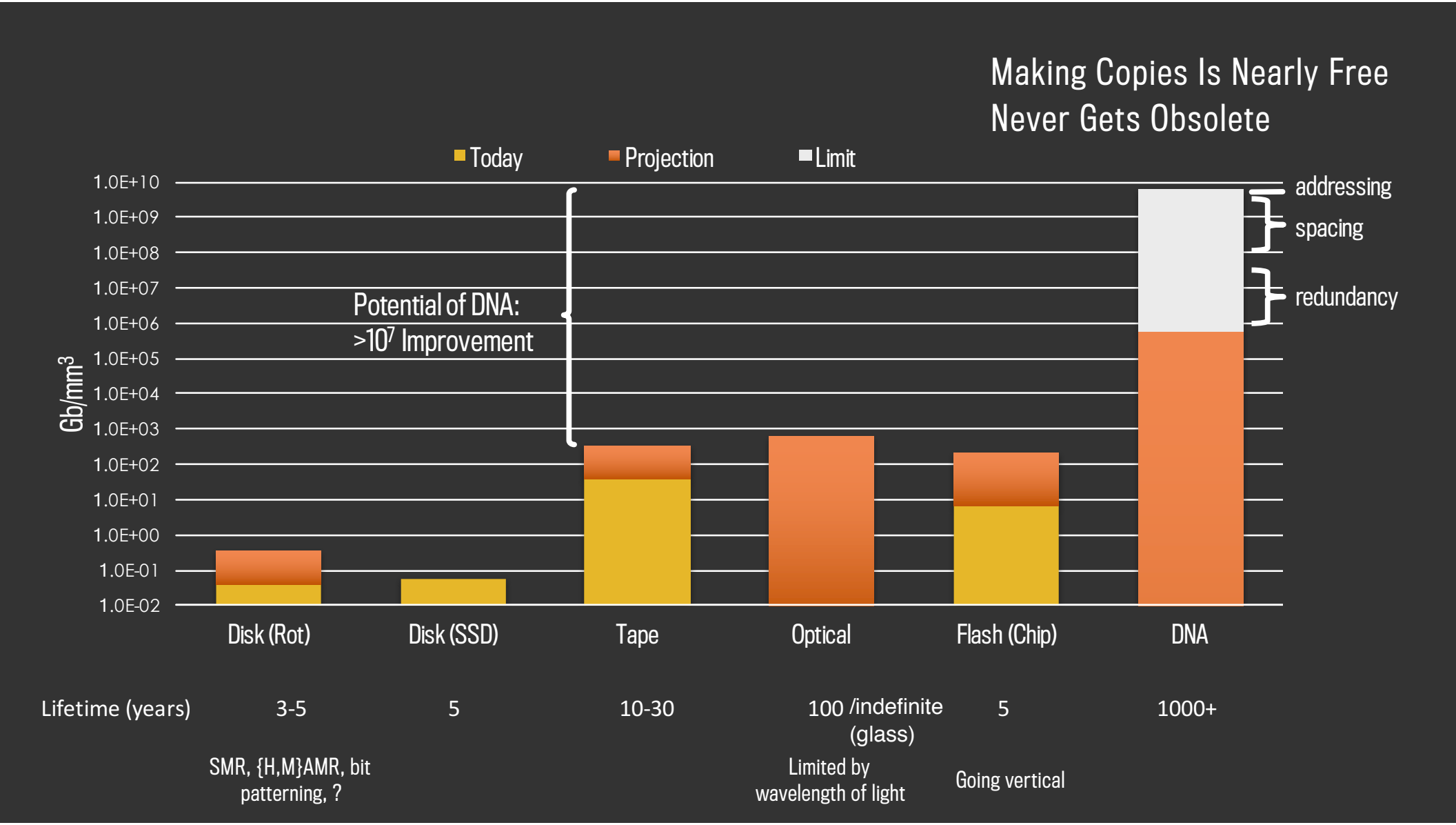
**LoC DSA, September 2018**

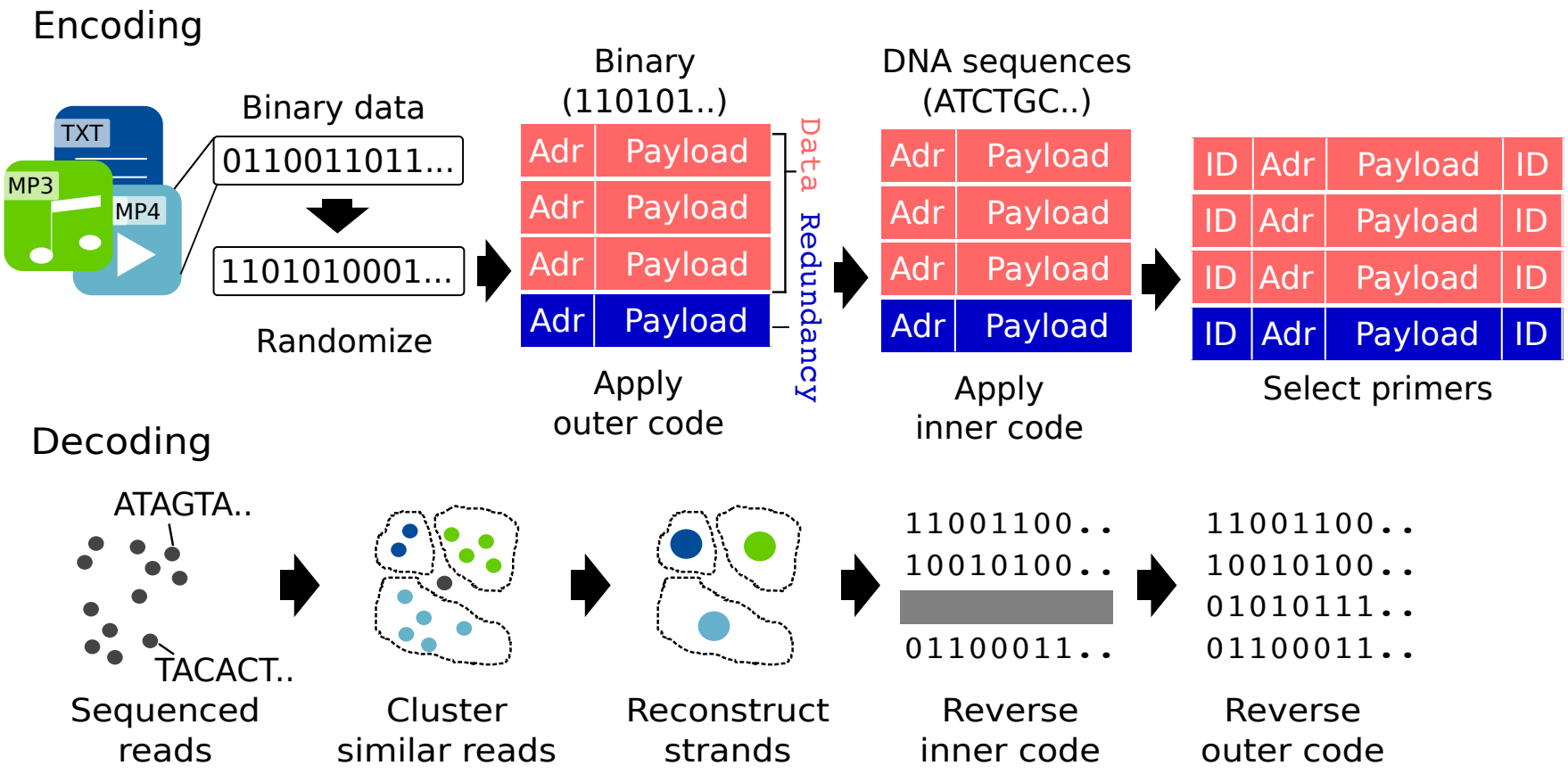
MISL


Write Path
Read Path
Encoding
Synthesis
Sequencing
Decoding
1
2
1
2
11011101
AGCTATCAG
AGCTATCAG
11011101

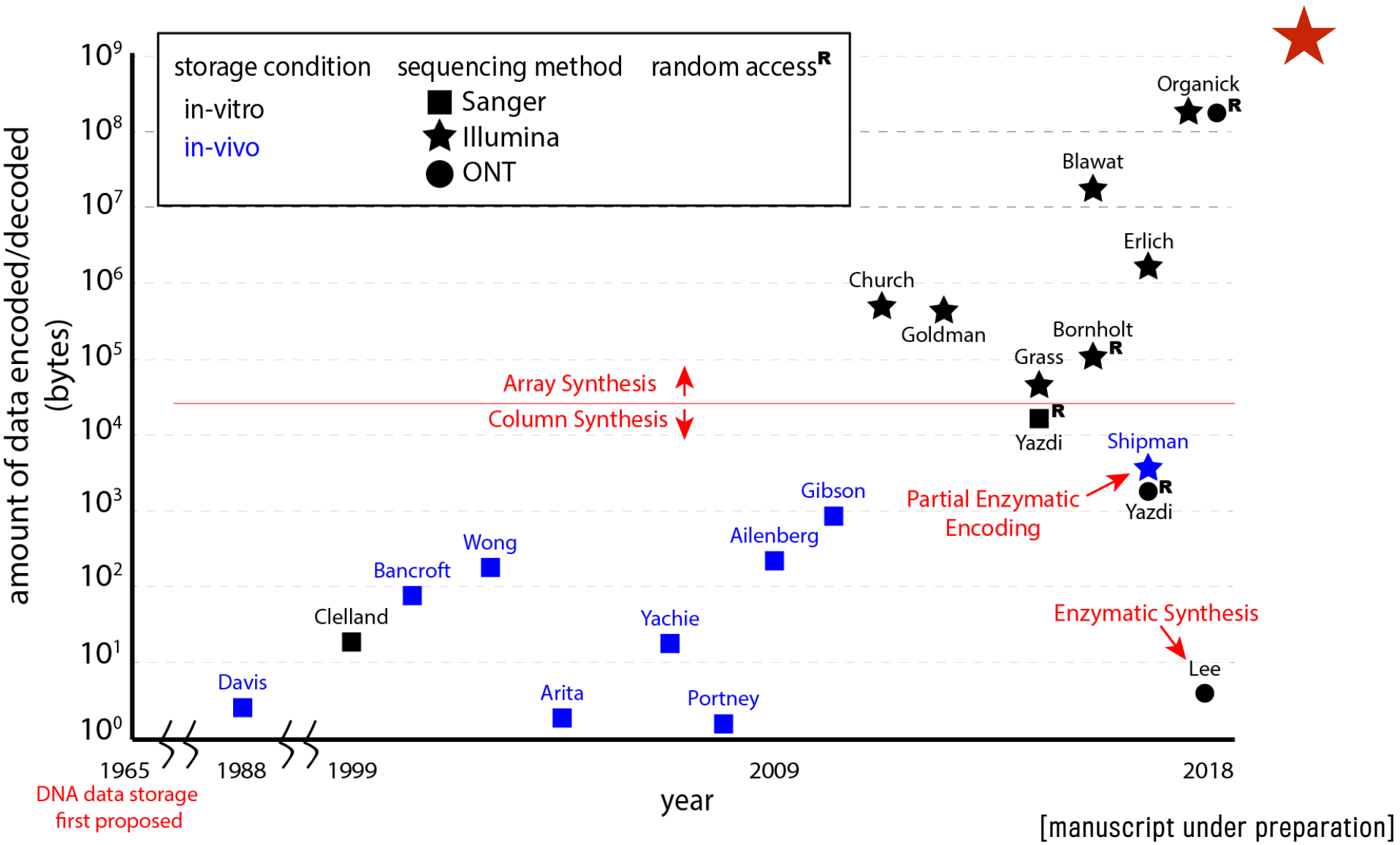
HD Video of OK Go’s This Too Shall Pass — watched 100M+ times


~10 million copies of the HD movie
Photo: Tara Brown / UW


Making Copies Is Nearly Free
Never Gets Obsolete
Today
Projection
Limit
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
1.0E+01
1.0E+00
1.0E-01
1.0E-02
Gb/mm3
Potential of DNA:
>10^7 Improvement
addressing
spacing
redundancy
Disk (Rot)
Disk (SSD)
Tape
Optical
Flash (Chip)
DNA
Lifetime (years)
3-5
5
10-30
100 /indefinite (glass)
5
1000+
SMR, {H,M}AMR, bit patterning, ?
Limited by wavelength of light
Going vertical




[Nature Biotechnology'18, NIPS'17]


storage condition
in-vitro
in-vivo
sequencing method
Sanger
Illumina
ONT
random access R
amount of data encoded/decoded (bytes)
$10^9$
$10^8$
$10^7$
$10^6$
$10^5$
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
Organick
Blawat
Erlich
Church
Goldman
Bornholt
Grass
Array Synthesis
Column Synthesis
Yazdi
Shipman
Gibson
Partial Enzymatic Encoding
Yazdi
Wong
Ailenberg
Bancroft
Clelland
Yachie
Enzymatic Synthesis
Lee
Davis
Arita
Portney
1965
1988
1999
2009
2018
DNA data storage first proposed
year
[manuscript under preparation]

# Challenges

## to mainstream availability

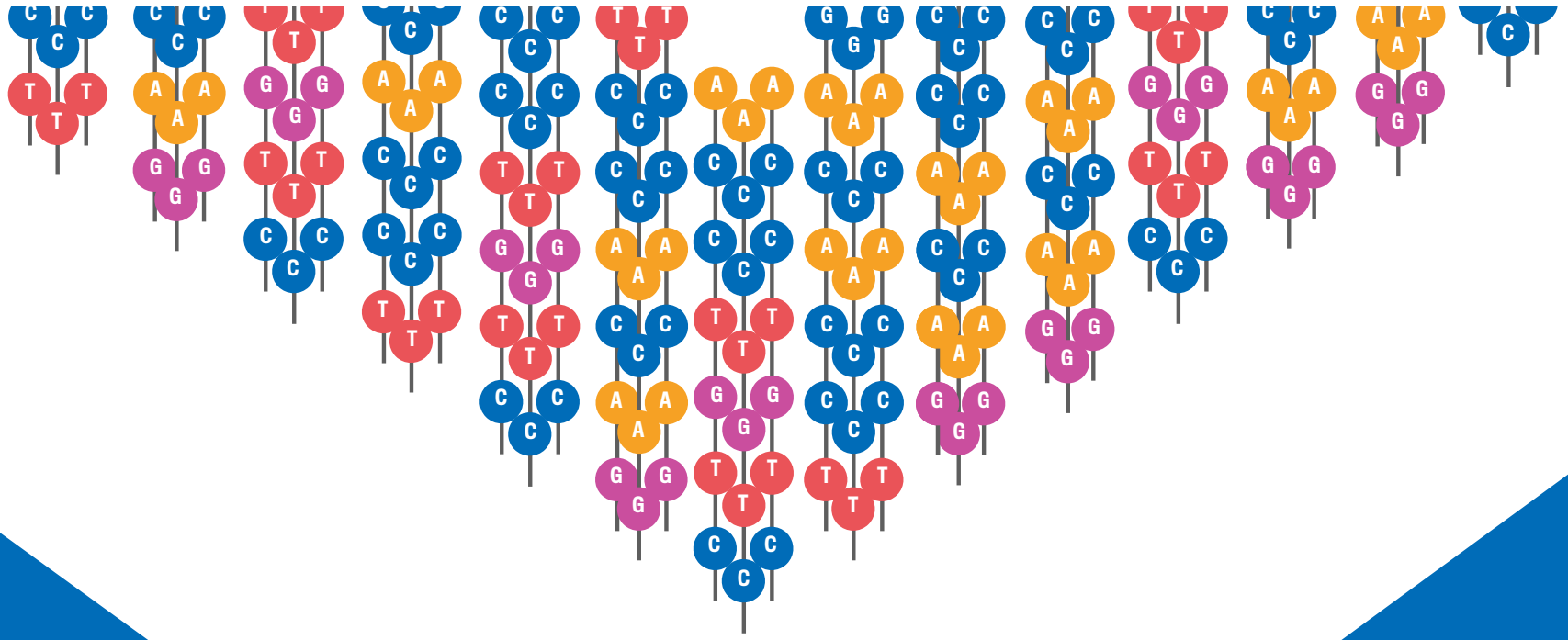
Bandwidth goal is O(TB/s), today writes are at O(KB/s)

Very high throughput and low cost writing and reading

Large-scale fluidics automation

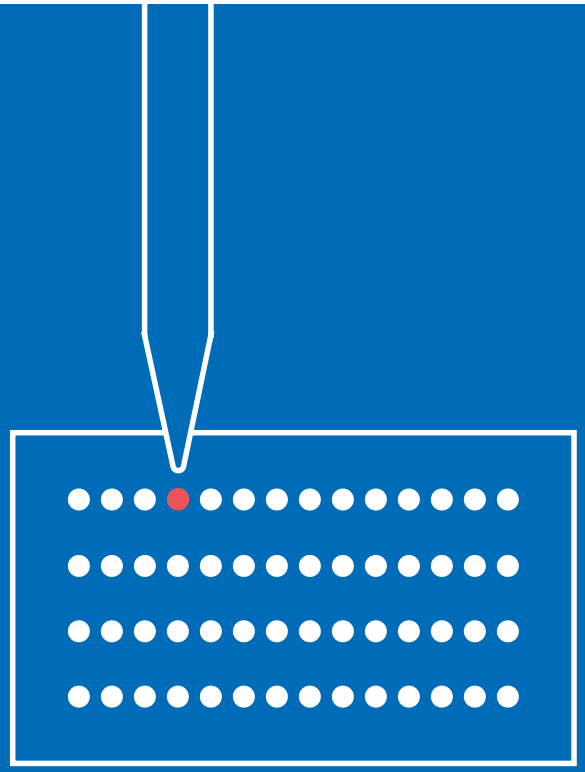
Scalable DNA physical organization and retrieval

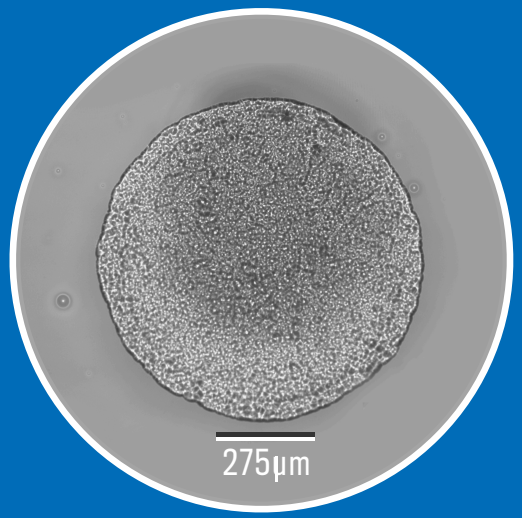
Computational costs



Large array DNA synthesis

Each spot grows many copies of a given sequence. Many spots.

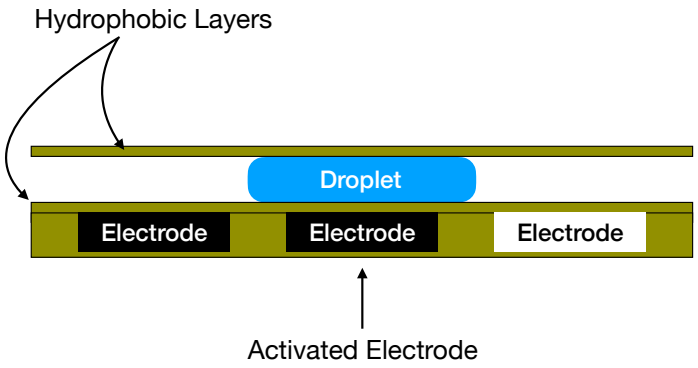


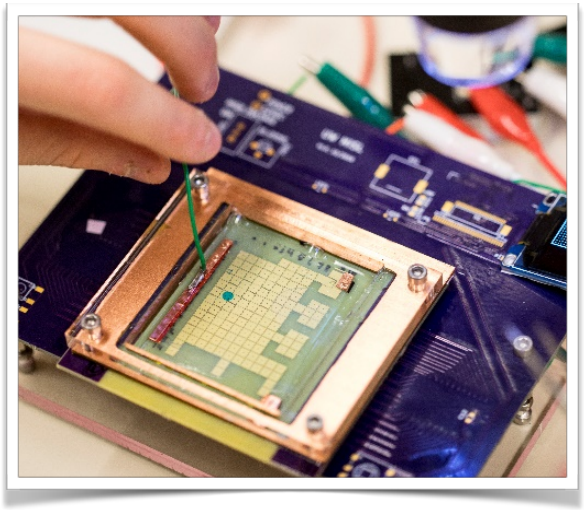



~100TB per spot

[manuscript under submission]

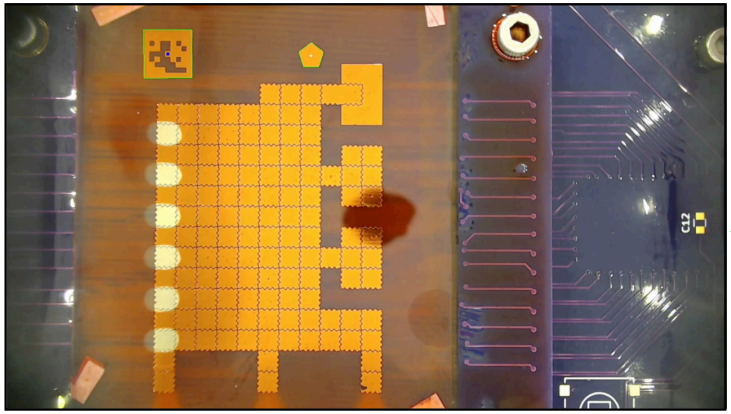
# Digital microfluidics



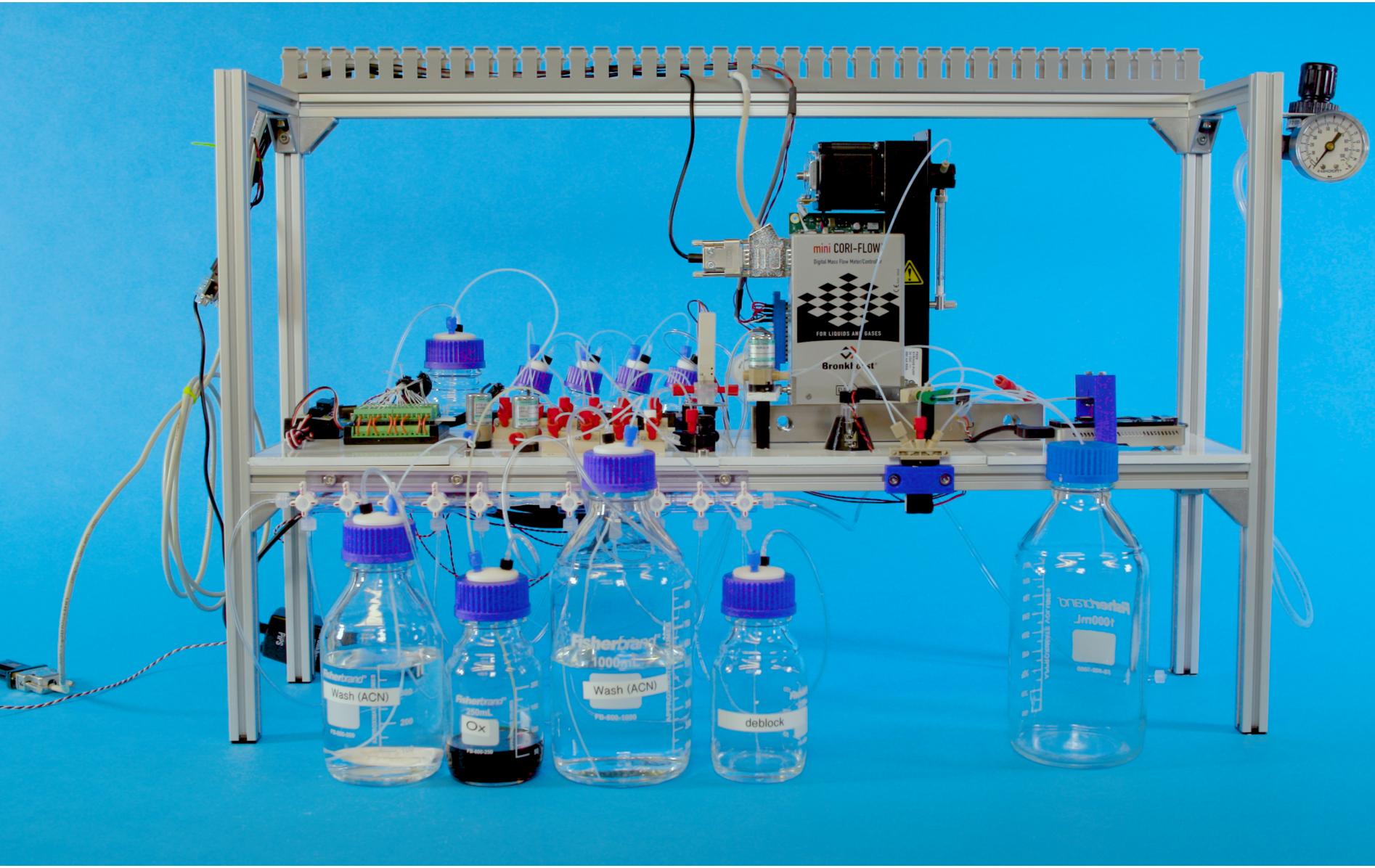
Hardware, software, wetware :)

Molecular domain



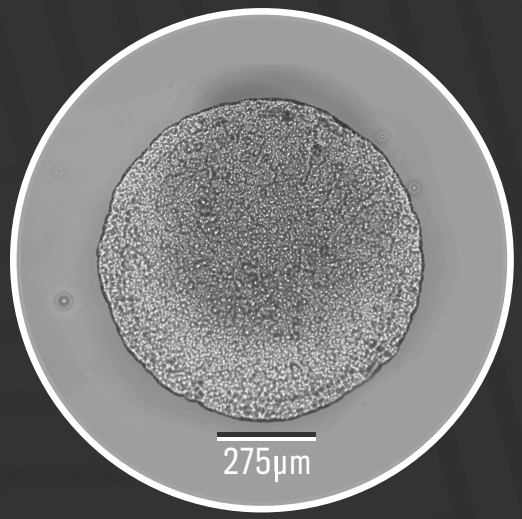
Electronic domain

# And now....


mini CORI-FLOW
Wash (ACN)
Ox
Wash (ACN)
deblock

# Beyond just data storage with DNA...

# DNA “computing" in the age of big data



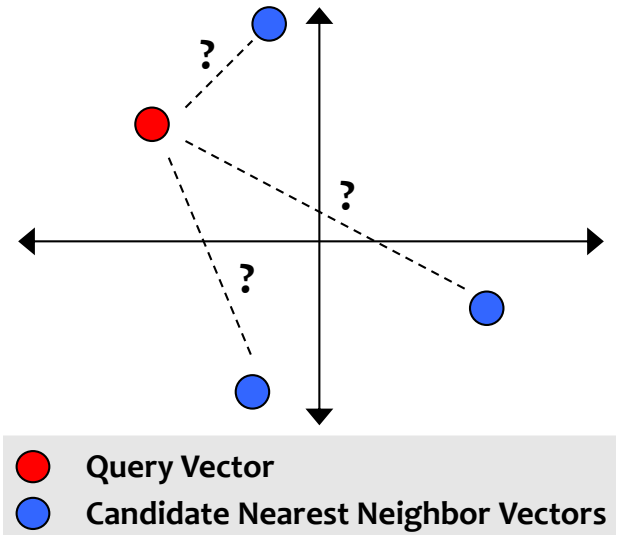

~100TB per spot

If DNA data storage succeeds, what if we could process data directly in DNA?

Extremely parallel and energy efficient

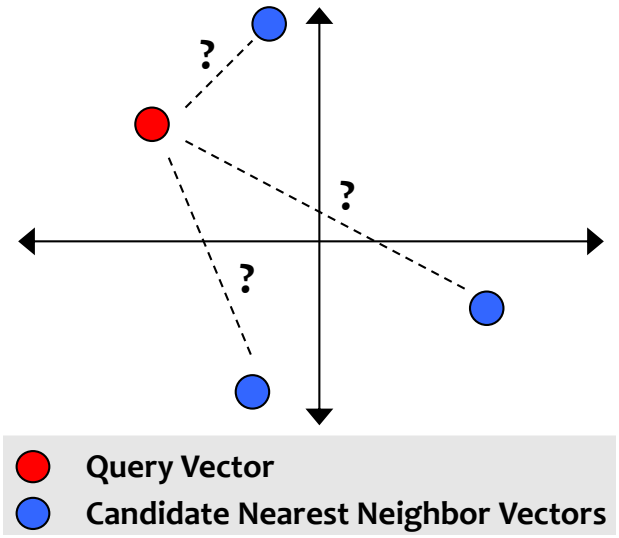
# Content-based media search

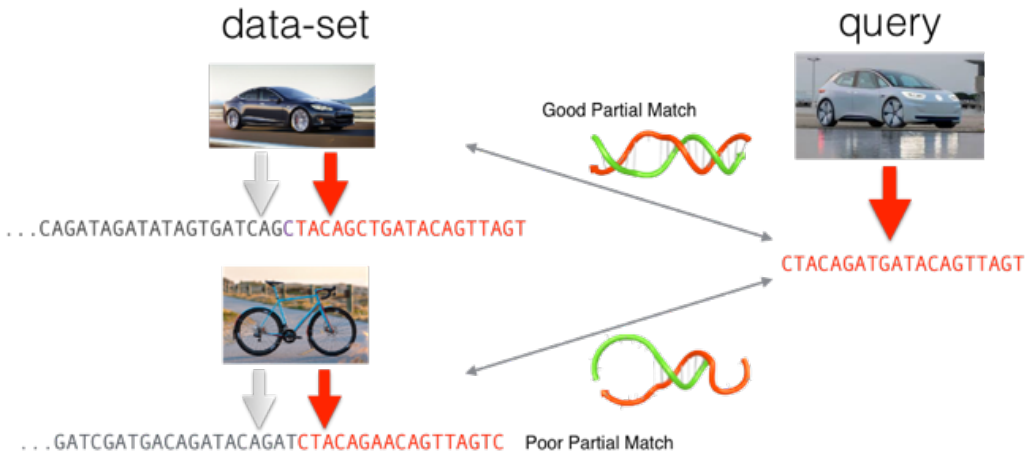
Extract feature vector, search in a high-dimensional space



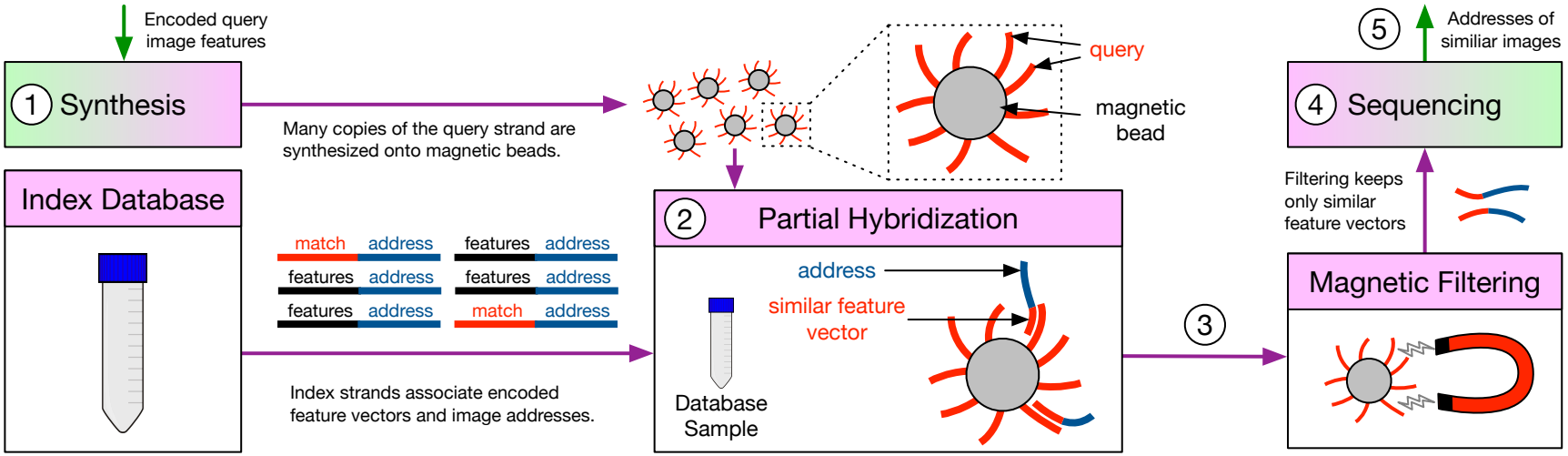
# Content-based media search … in DNA

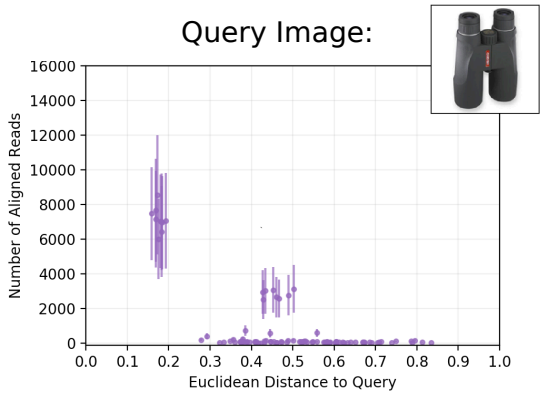
Map features vector do DNA such that molecules mapping to similar vectors “stick”

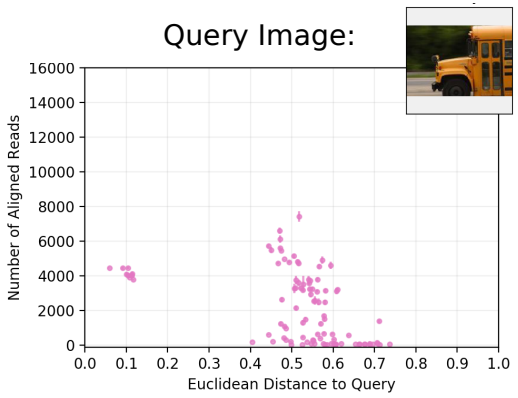


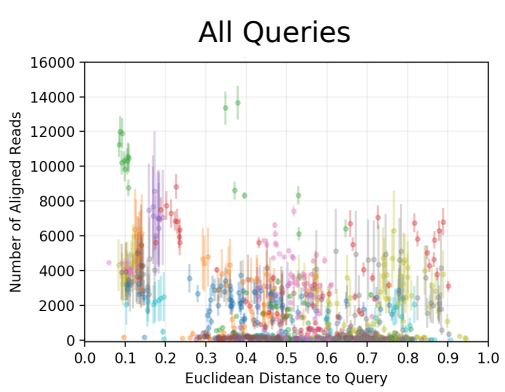




[in DNA’18]

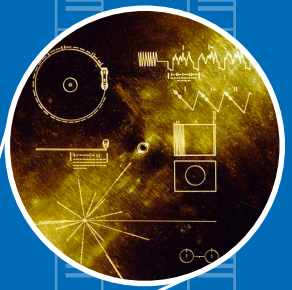


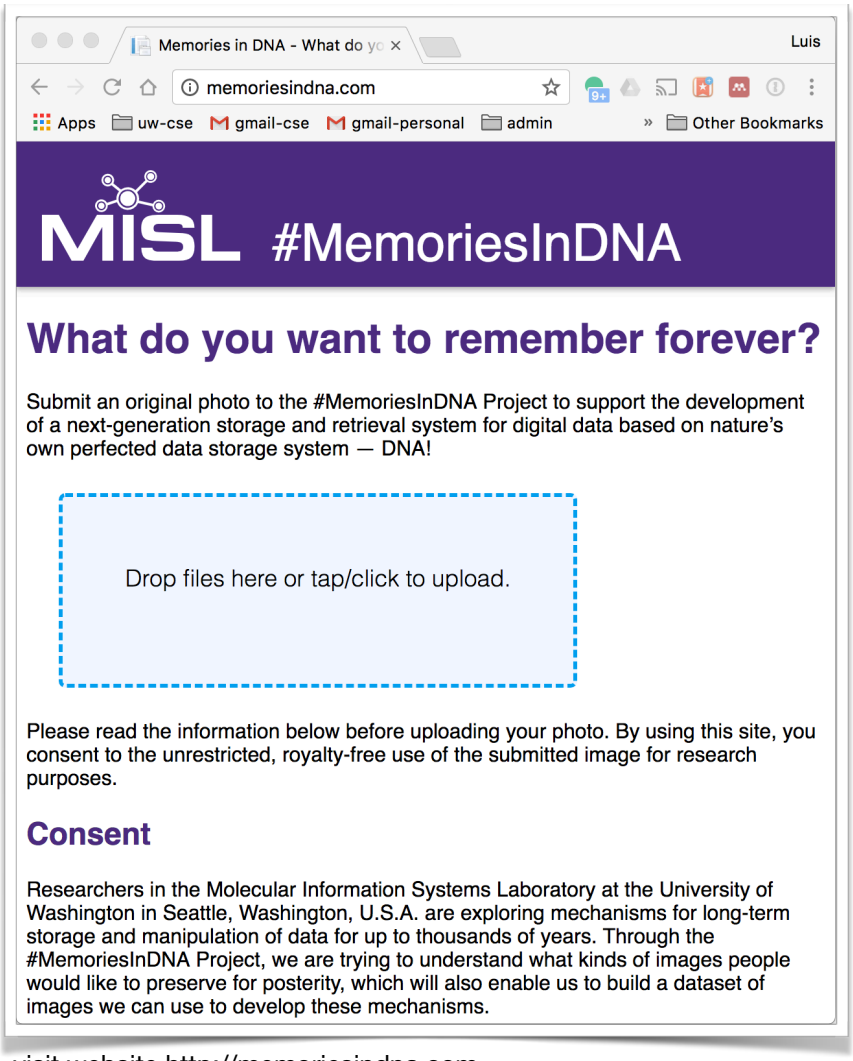



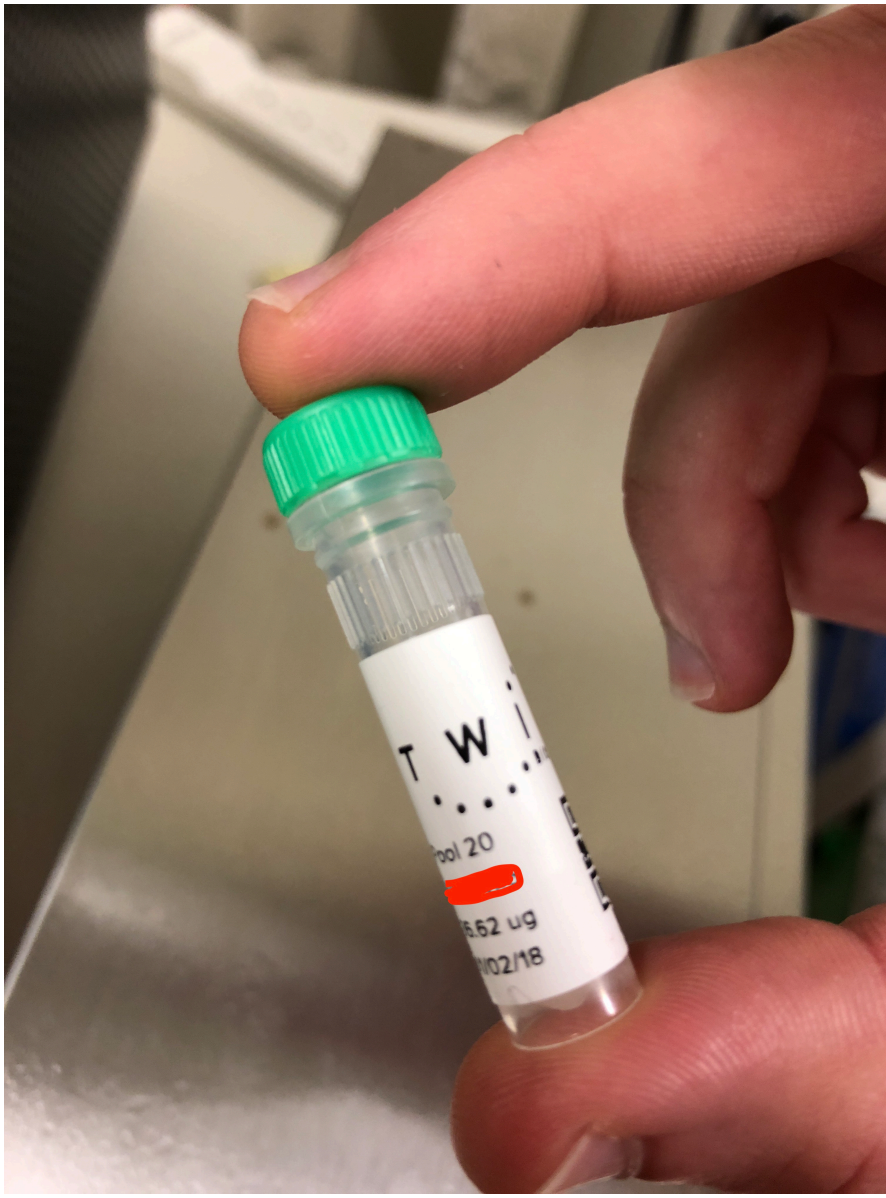

# Wetlab results


Over 700MB. 50M+ sequences. 9B+ Nucleotides,
5B+ reads. Demonstrated random access w/ 40+ objects.
Illumina and Nanopore sequencing readout.




visit website http://memoriesindna.com
or send email to submissions@memoriesindna.com



MISL

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Microsoft