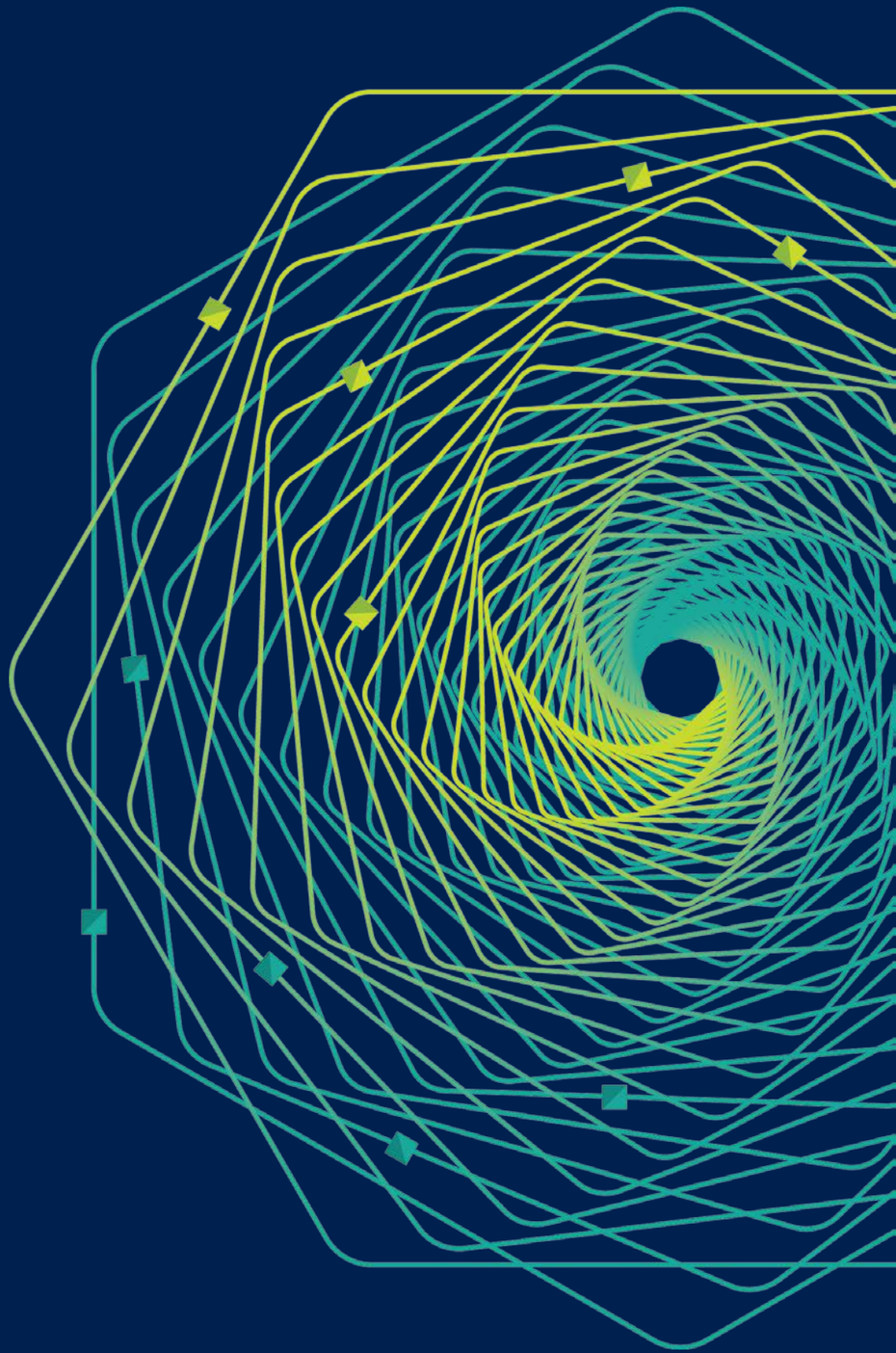
Microsoft

Research Faculty Summit 2018

Systems | Fueling future disruptions



Getting Polymers to Tell a Story: DNA Data Storage and Processing-in-Molecules

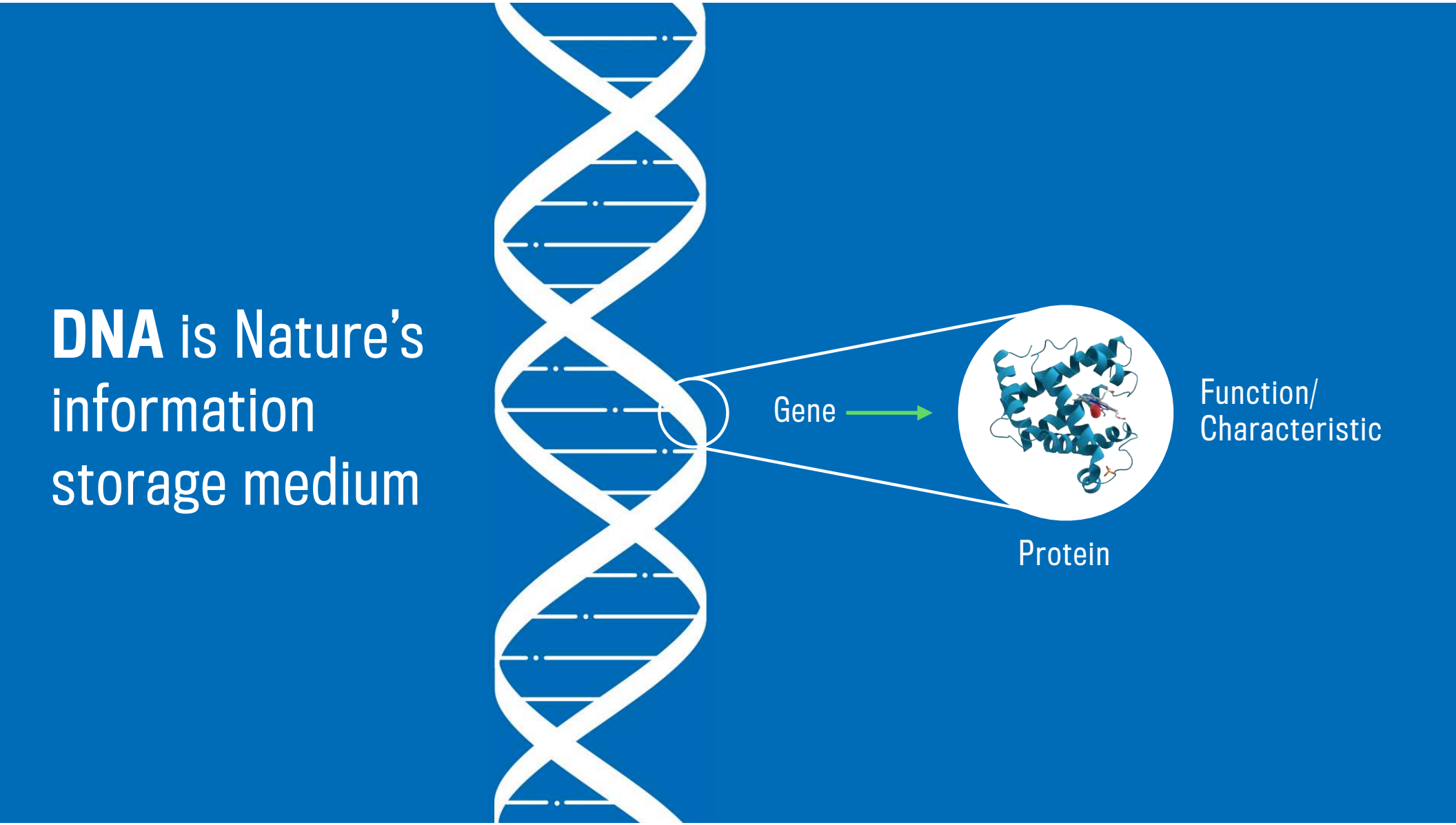
Luis Ceze
Molecular Information Systems Lab
Sampa Lab for Hardware/Software Co-Design
Paul G. Allen School of Computer Science & Engineering
University of Washington

joint work with **Karin Strauss**, 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.

Microsoft Faculty Summit 2018

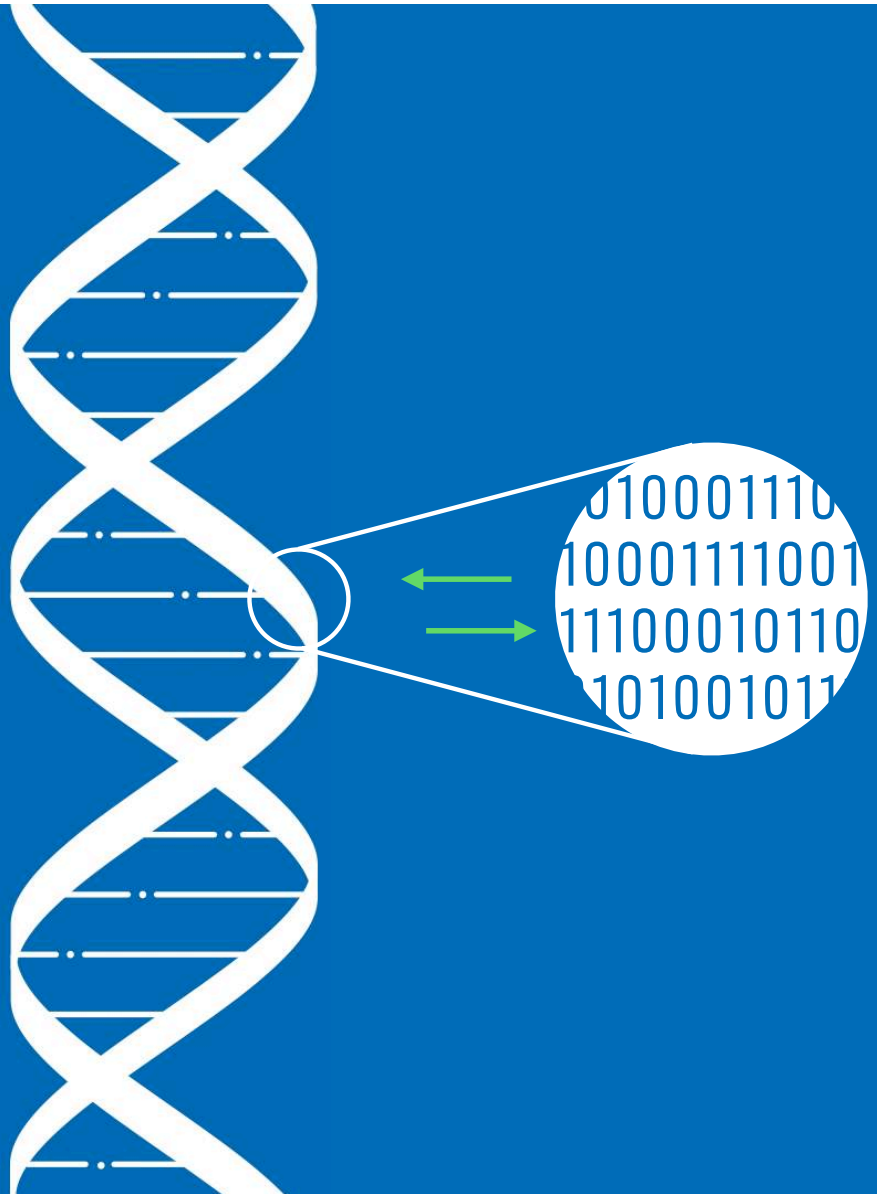
MISL PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING Microsoft

DNA is Nature's information storage medium



Using Synthetic **DNA** for Data Storage

Manufactured DNA



DNA Molecules
for Digital Data

Extremely Dense

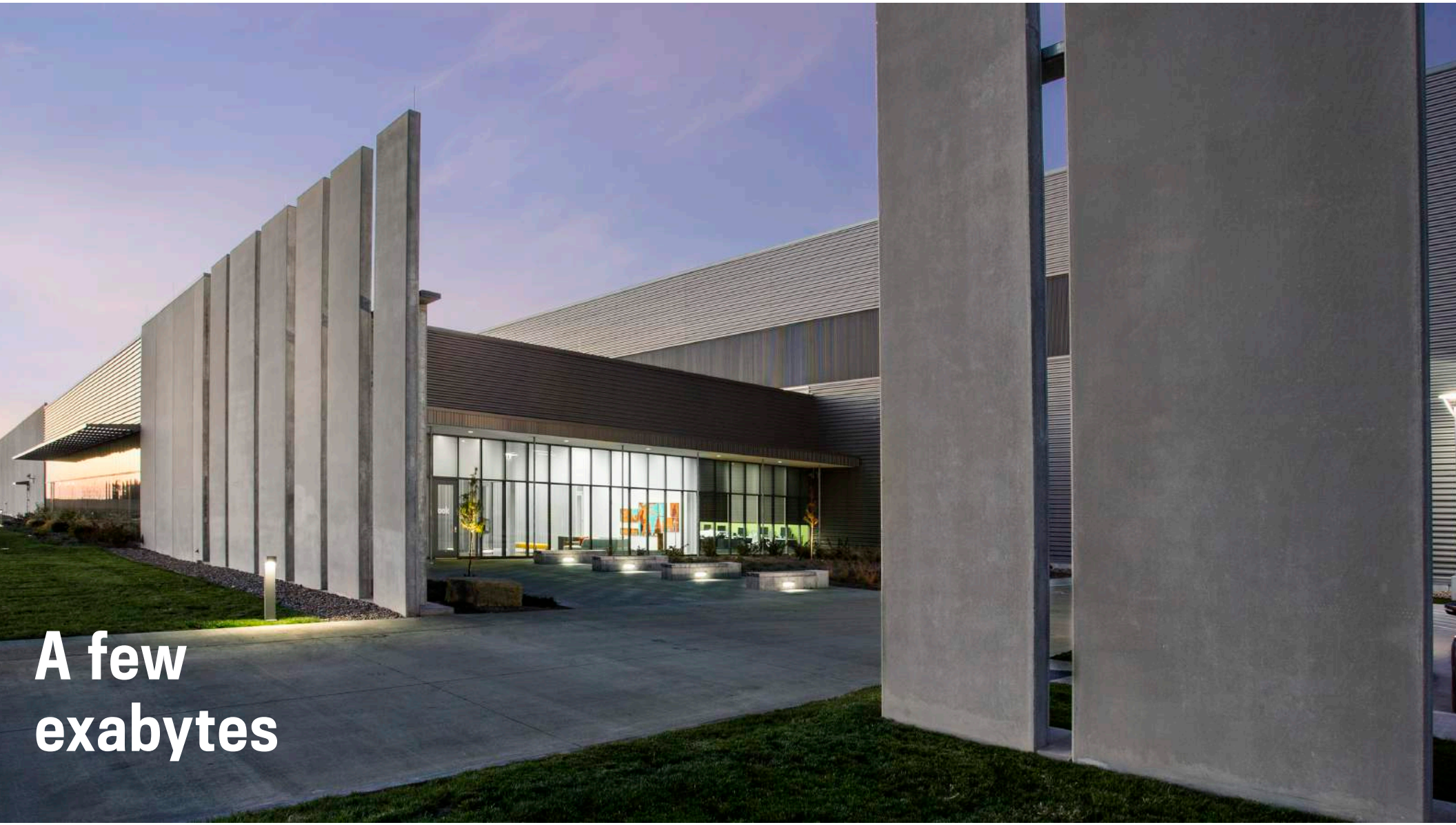
1 Exabyte in 1 in^3

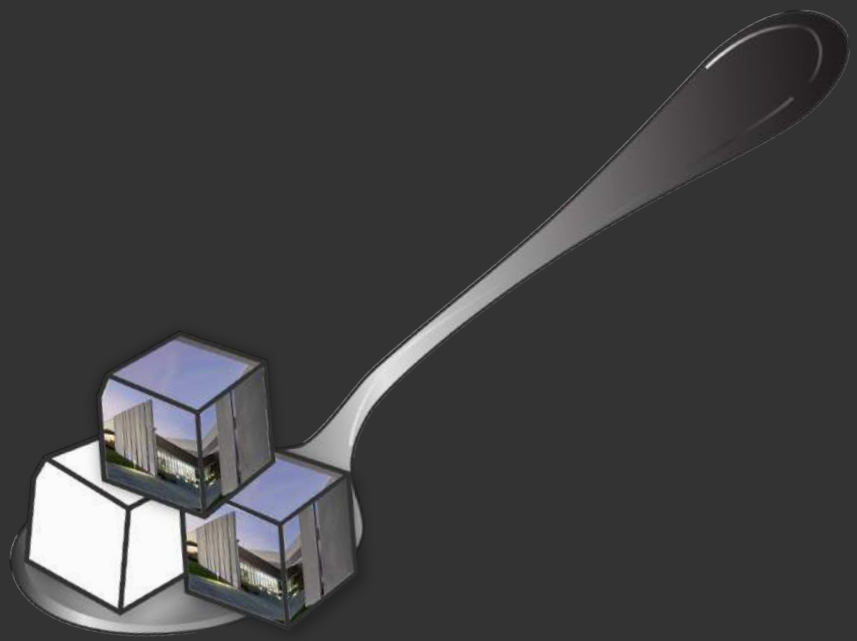
Extremely Durable

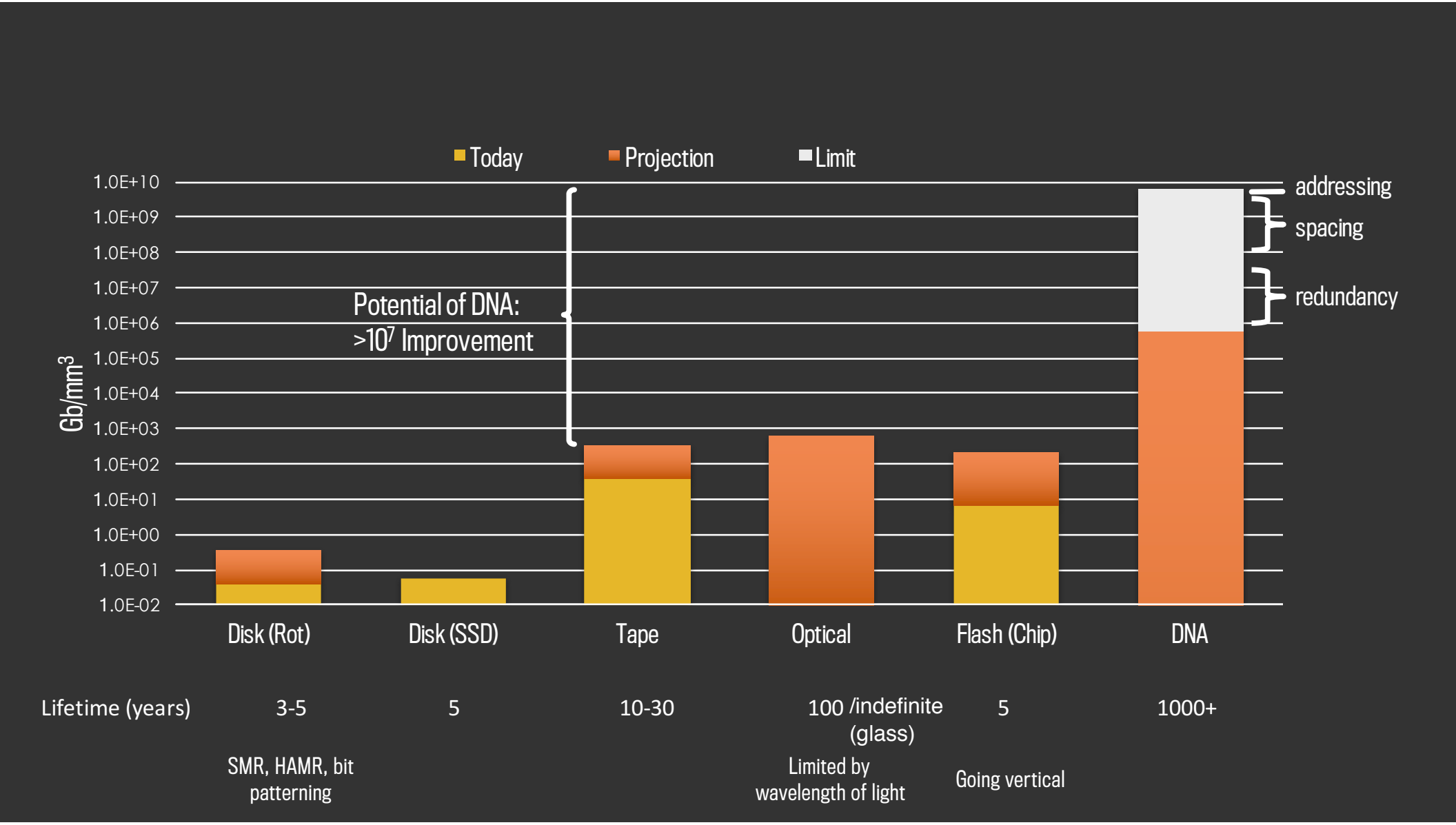
1,000s of Years

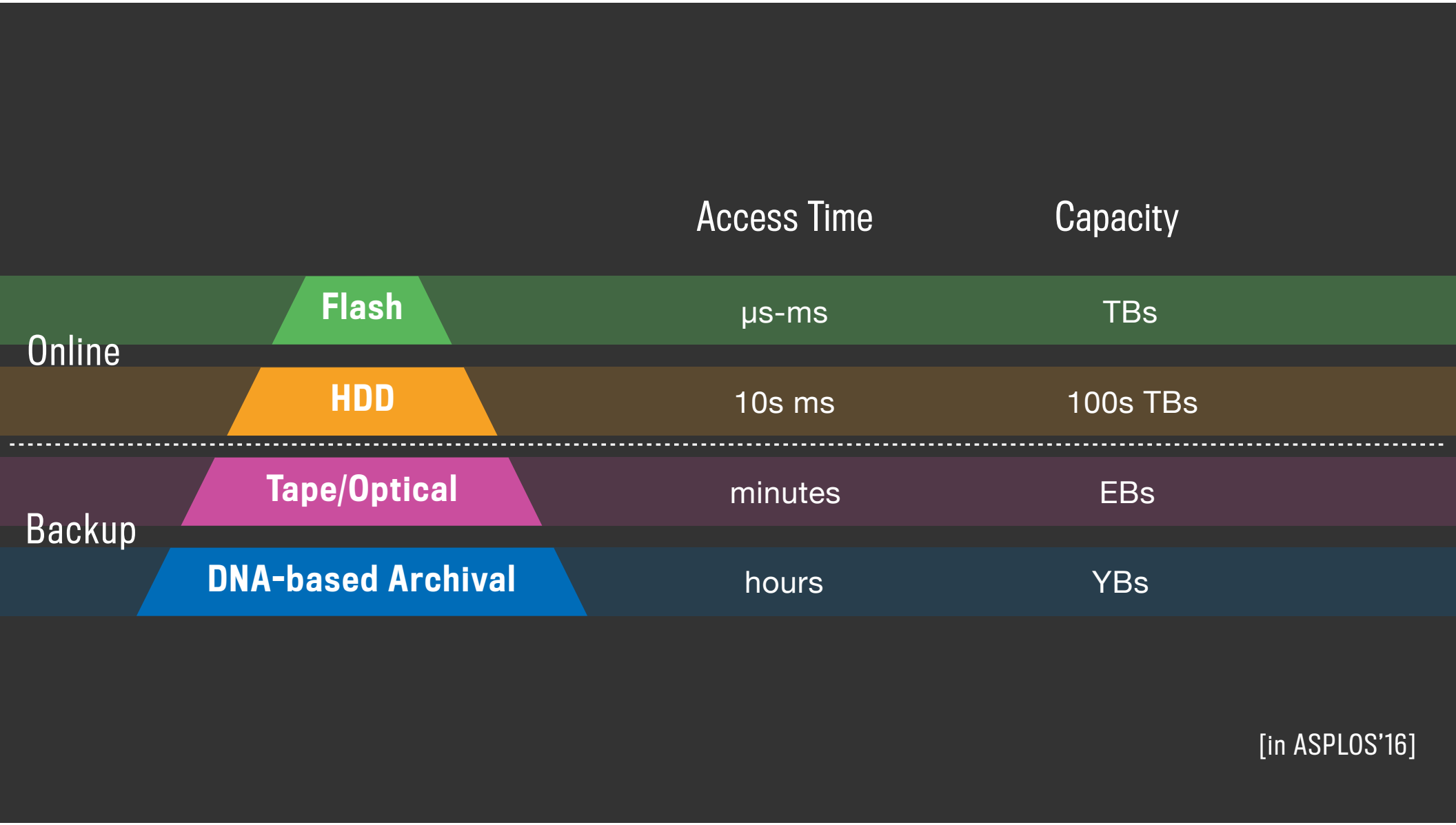
Making Copies Is Nearly Free

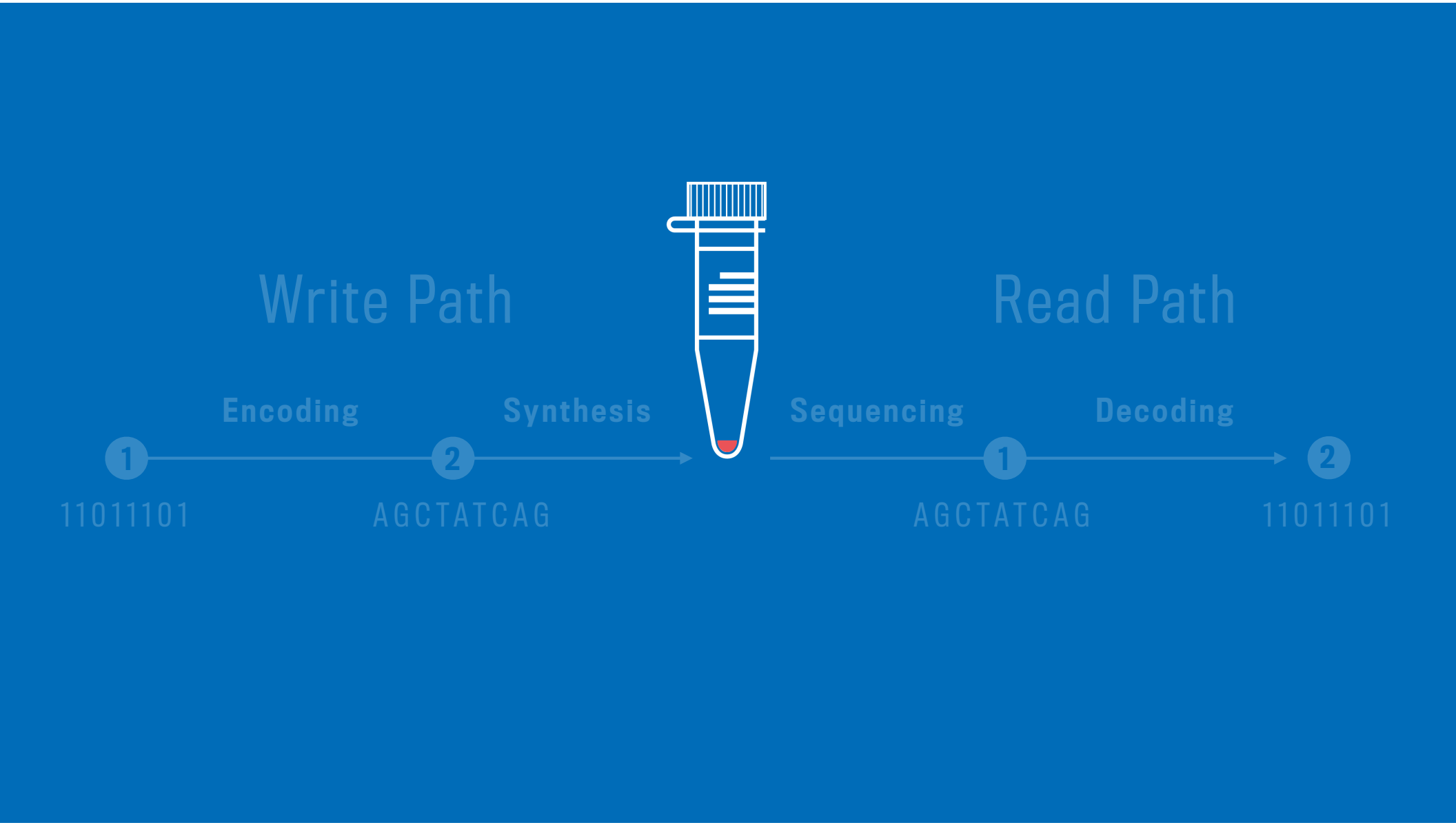
Never Gets Obsolete


A few
exabytes

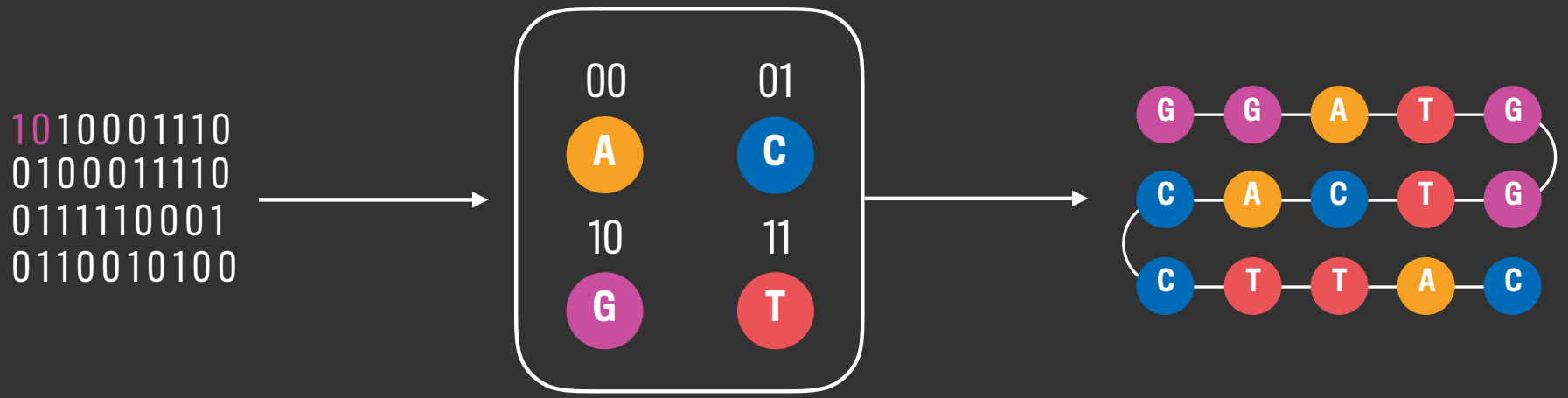



Today
Projection
Limit
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
1.0E+01
1.0E+00
1.0E-01
1.0E-02
Gb/mm3
Potential of DNA:
>10^7 Improvement
addressing
spacing
redundancy
Disk (Rot)
Disk (SSD)
Tape
Optical
Flash (Chip)
DNA
Lifetime (years)
3-5
5
10-30
100 /indefinite (glass)
5
1000+
SMR, HAMR, bit patterning
Limited by wavelength of light
Going vertical


Access Time
Capacity
Online
Flash
µs-ms
TBs
HDD
10s ms
100s TBs
Backup
Tape/Optical
minutes
EBs
DNA-based Archival
hours
YBs
[in ASPLOS'16]

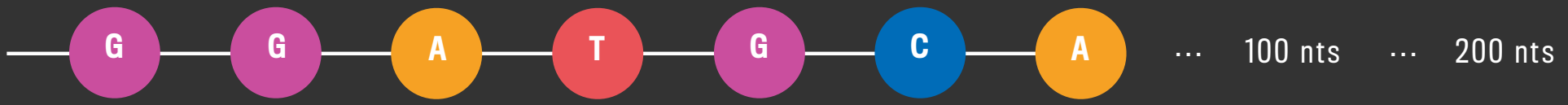

Write Path
Read Path
Encoding
Synthesis
Sequencing
Decoding
1
2
1
2
11011101
AGCTATCAG
AGCTATCAG
11011101

Encoding Digital Data in DNA



Repeated Letters Are Bad: Avoid them with Randomization & Rotation.

P[Attach] = 99%



99% 98% 97% 96.1% 95.1% 94.2% 36.6% 13.4%

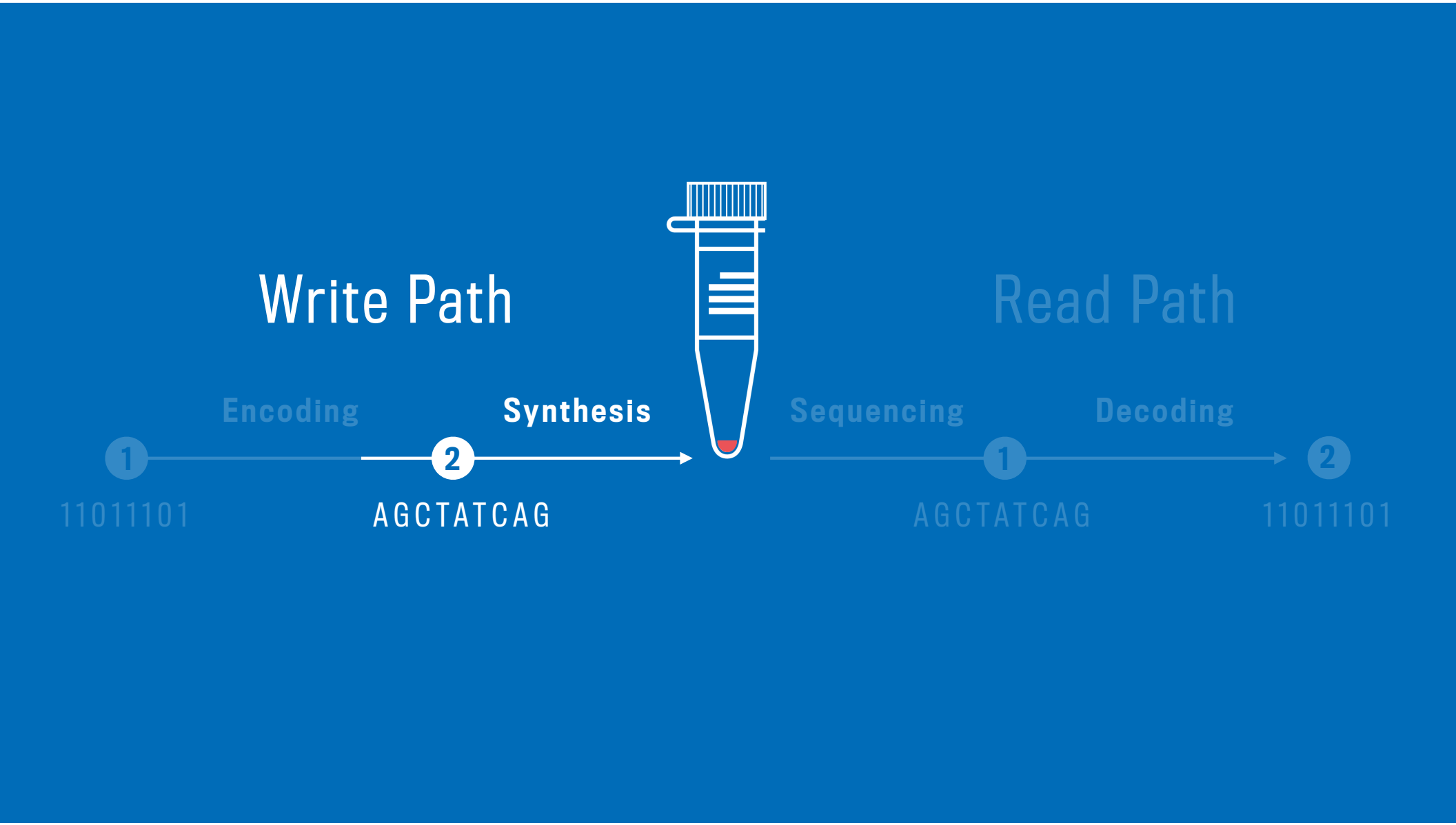
Synthetic DNA has limited length: Break it into chunks.

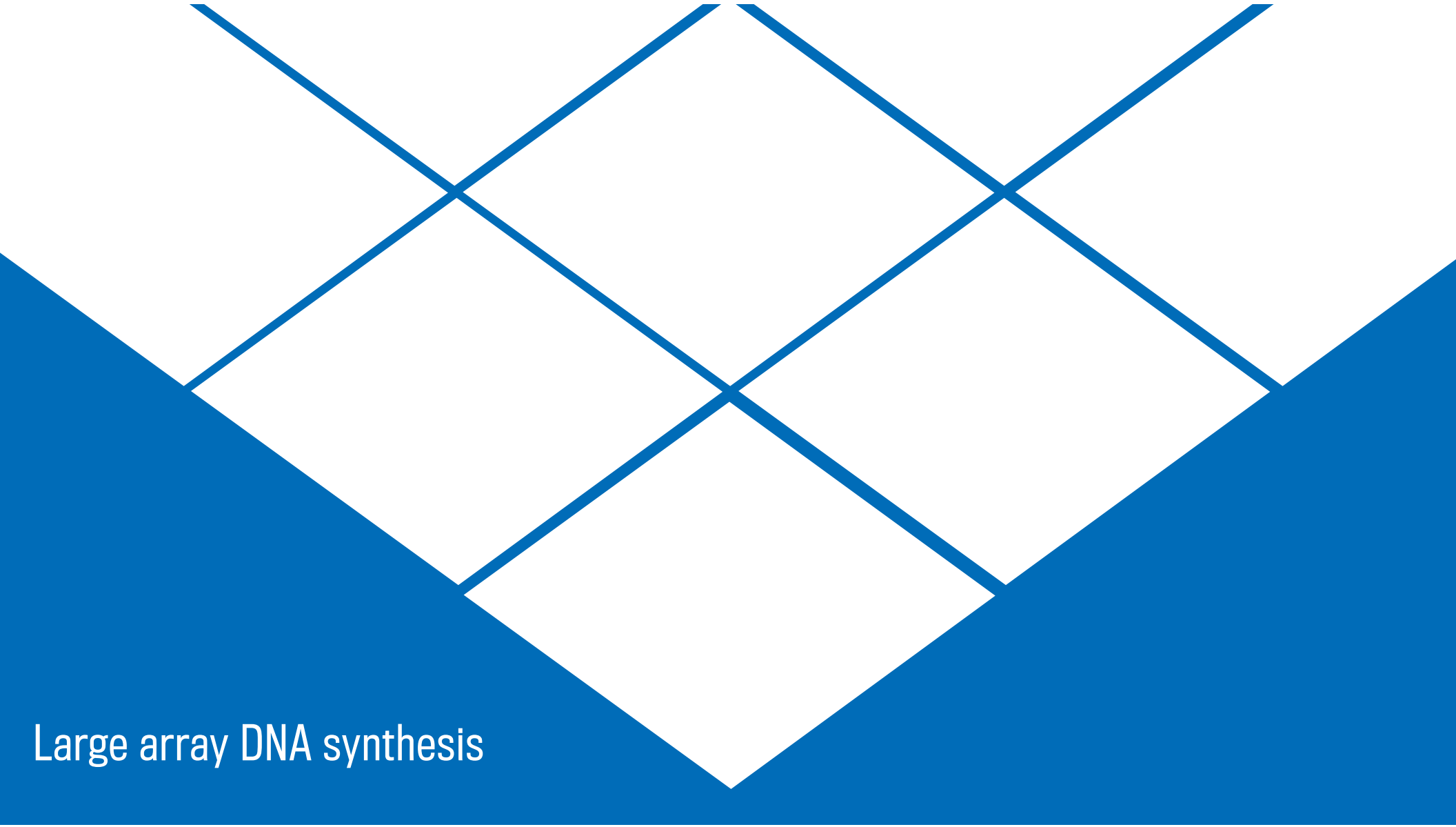
Break into chunks and add redundancy

Key Identifiers ("primers")	Payload Chunks	Addresses Within the File	Key Identifiers ("primers")
ATGTT	CGAT GCAC	ACTA	CATCC
ATGTT	TGCT GACG	ACTC	CATCC
ATGTT	GCTA GCTC	ACTG	CATCC
ATGTT	TCGC TACG	CAAC	CATCC
ATGTT	TGCA ATTC	CAAG	CATCC

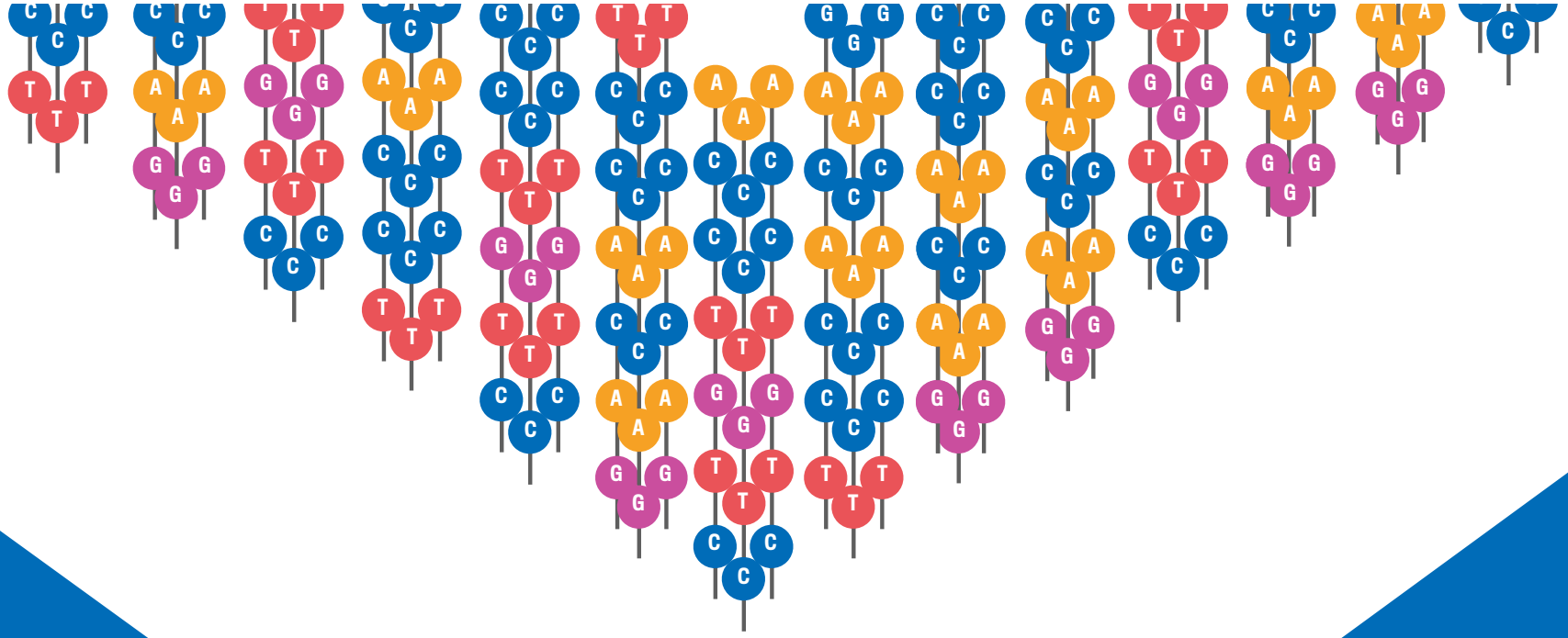
Redundant Data in Additional DNA Strands. Many Possibilities: Parity, Reed Solomon, LDPC, ...

~20 Bytes per 150nt DNA sequence. Many sequences per file. ~1% (in, del, sub) error per base.


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

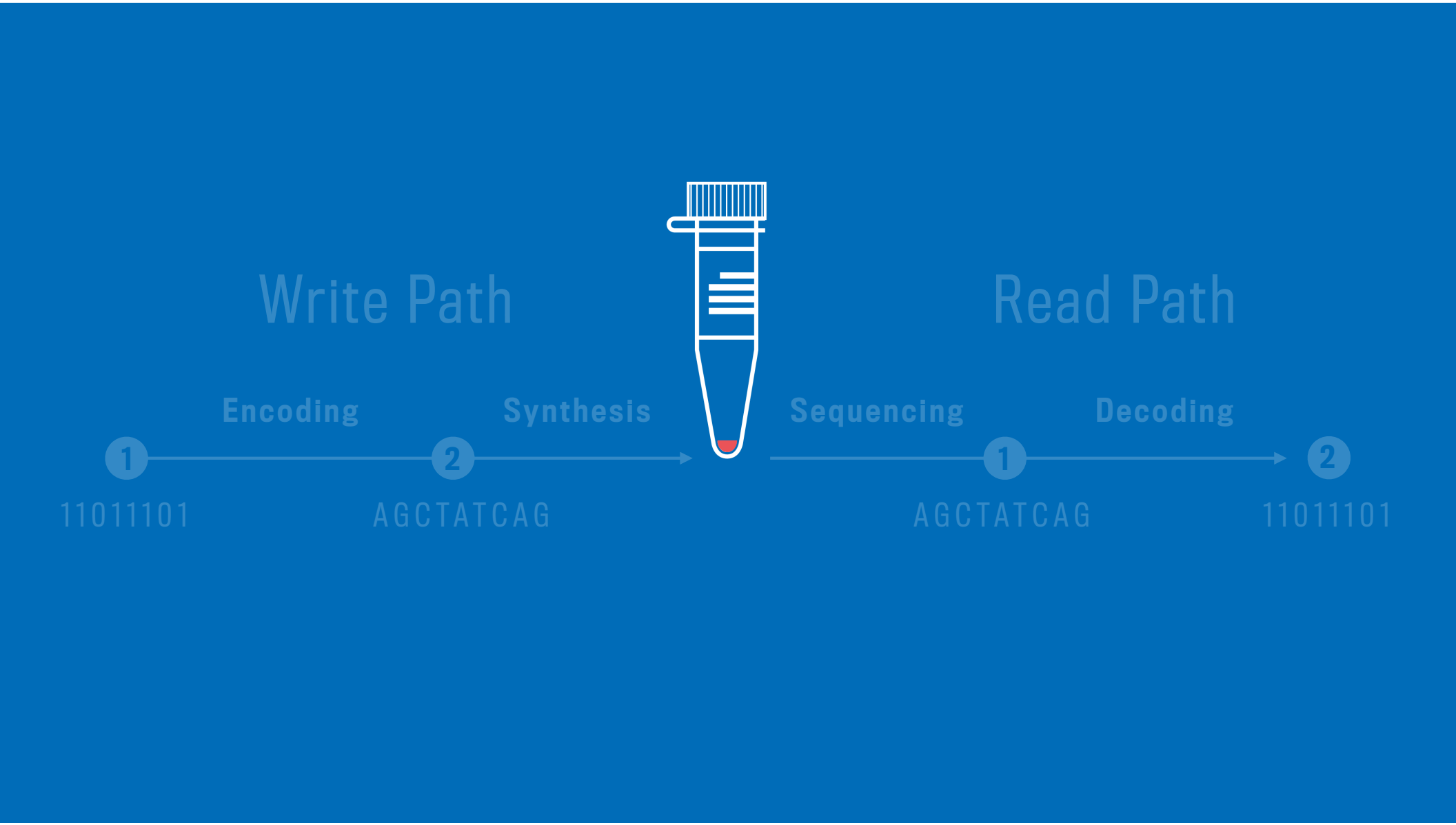


Large array DNA synthesis



Large array DNA synthesis

Each spot grows many copies of a given sequence


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

DNA Sequencing

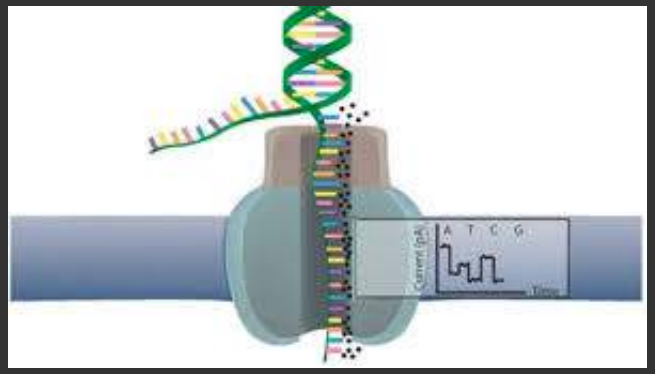
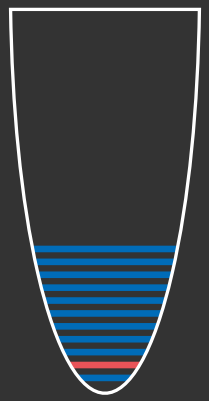
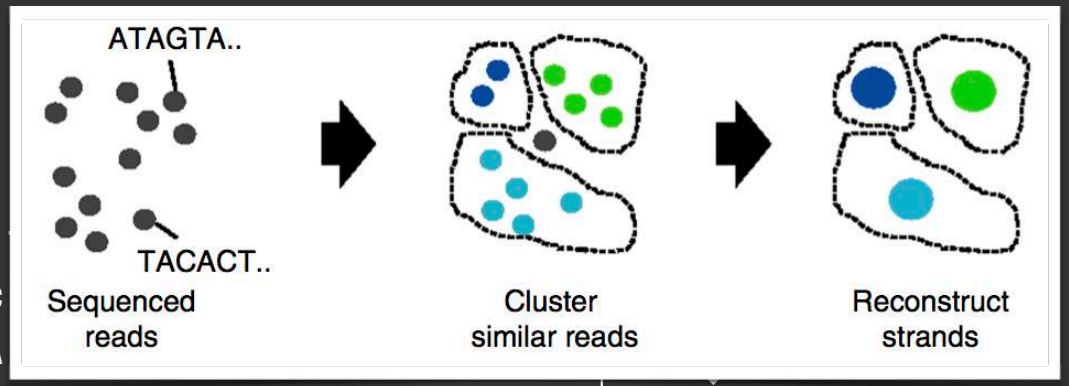
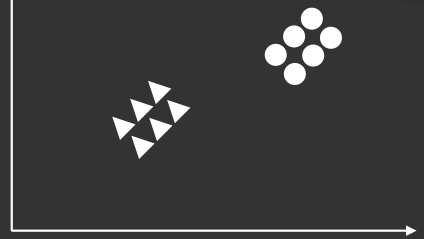


Image credit: Oxford Nanopore



ATGTT TGC
ATGTT GCC
ATGTT GGA
ATGTT TGCT TACC CAAC CATCC
ATGTT GCCA GTTC AAAG CATGC
ATGTT GGAT GCAC AAGA CATCC
ATGTT TGCT TACC CAAC CATCC
ATGTT GCCA GTTC AAAG CATCC



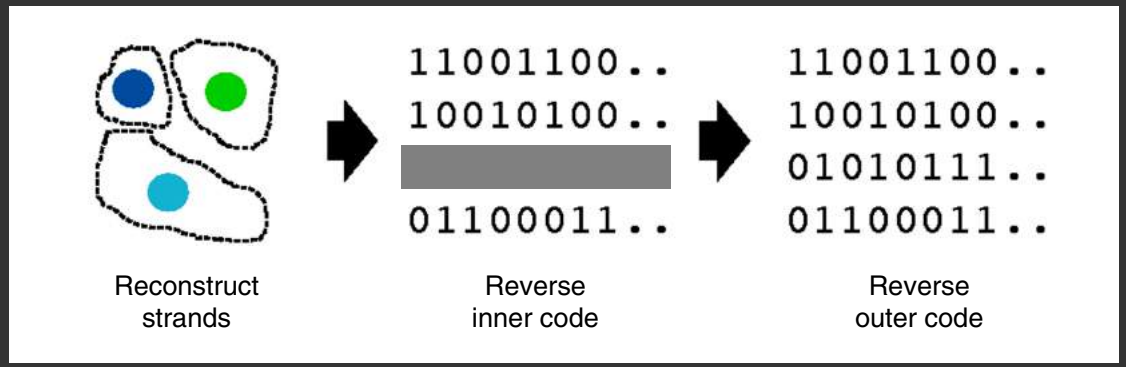


Selected DNA strands

Sequencing (Noisy reads)

Clustering (No reference)

[in NIPS'17]



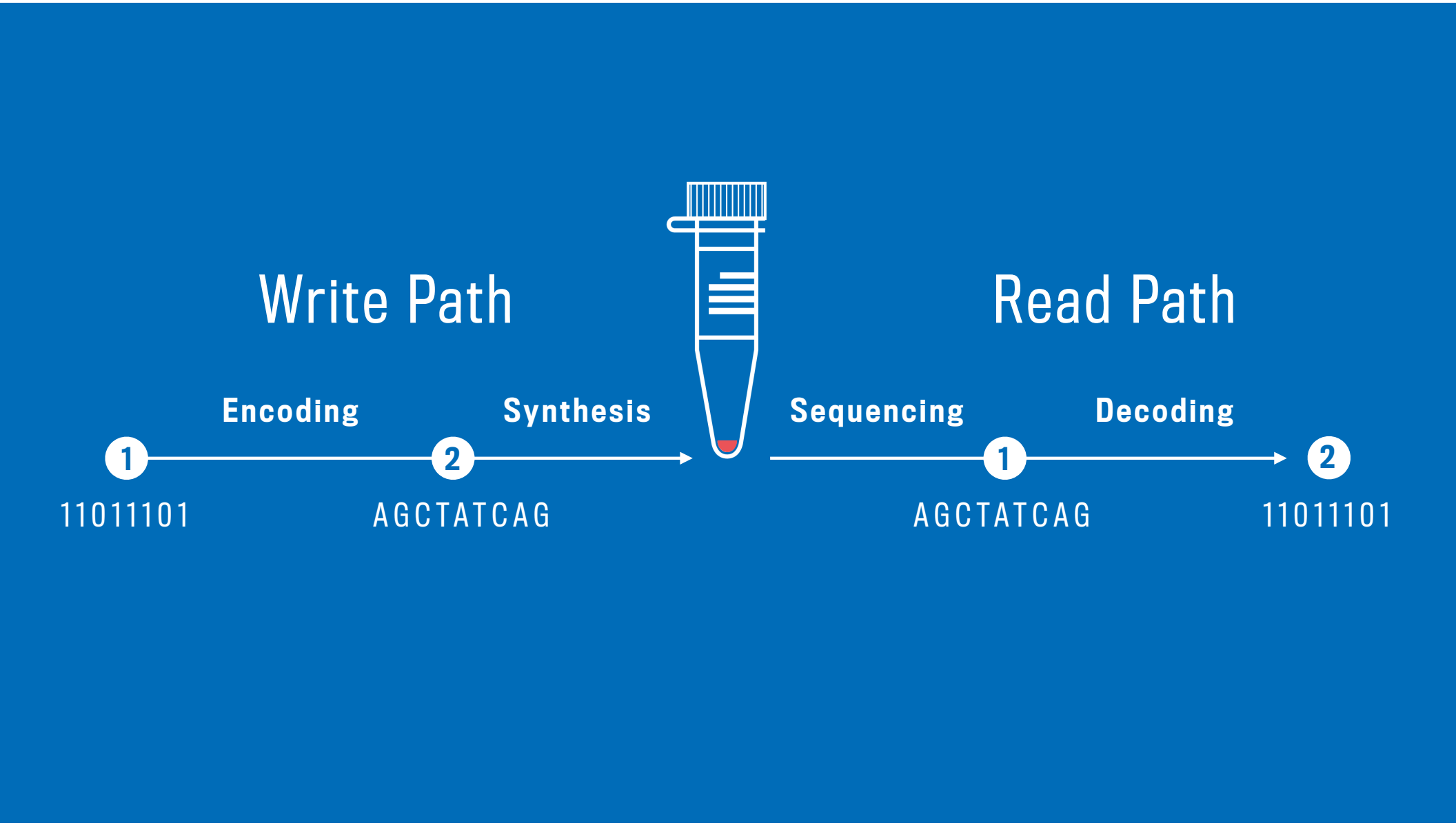
...
11011
01011
01...

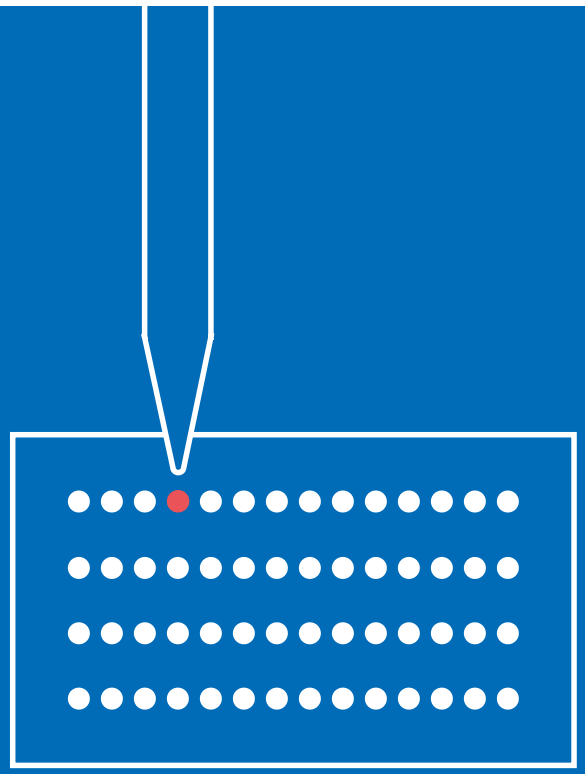
■■■▲▲▲●●●

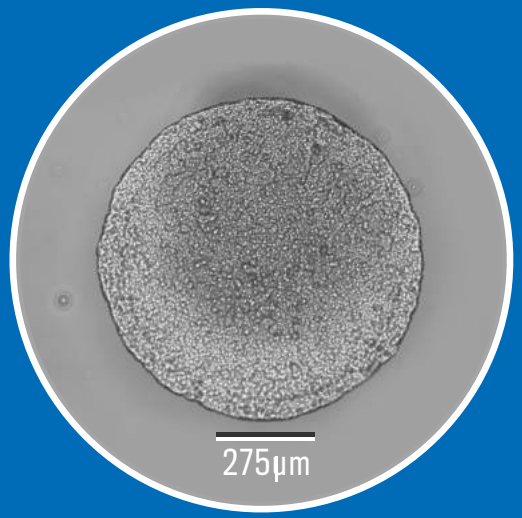
Reassemble Data

Error Correction

Error-Free Data

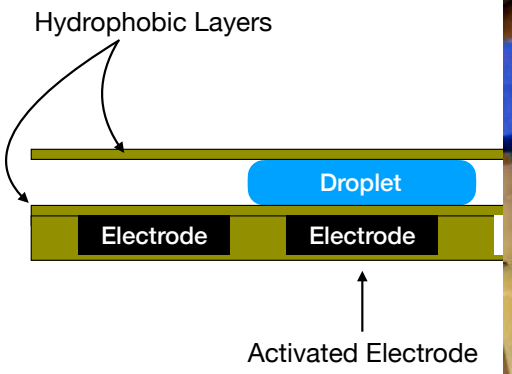

Write Path
Read Path
Encoding
Synthesis
Sequencing
Decoding
1
2
1
2
11011101
AGCTATCAG
AGCTATCAG
11011101

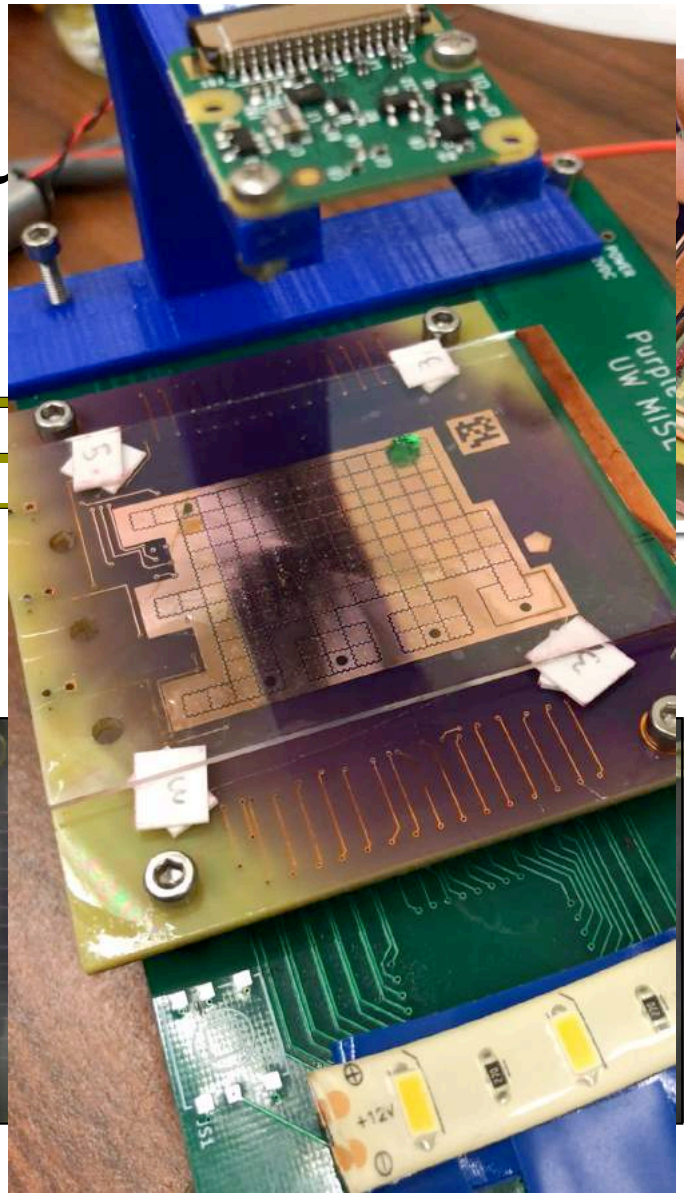


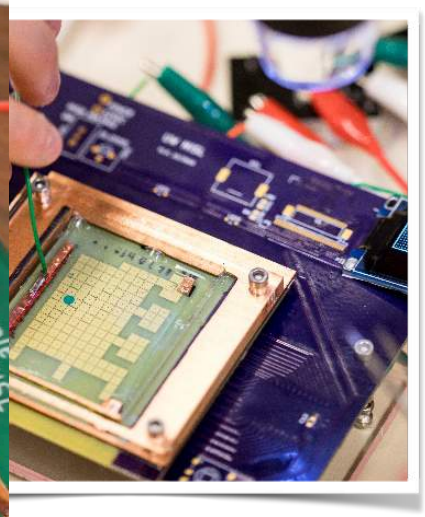


~100TB per spot

Digital microflu





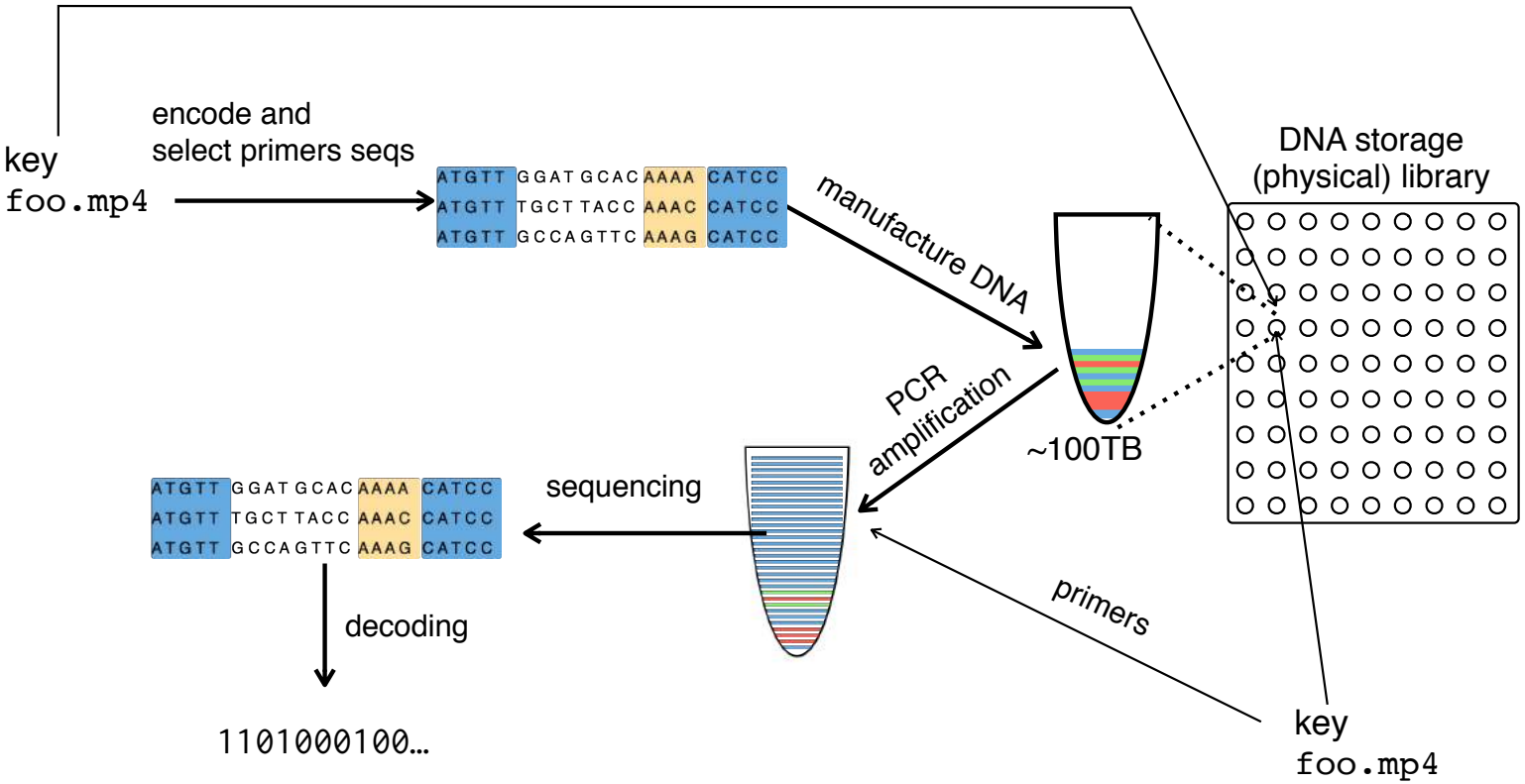


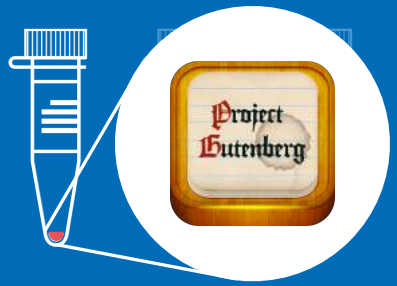
Molecular domain

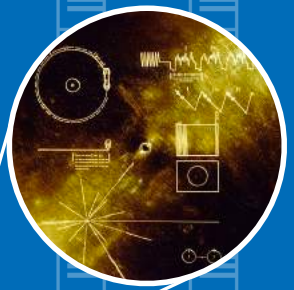
Electronic domain

Putting it all together as (key, value) store

Data address specifies physical location and primer for random access.

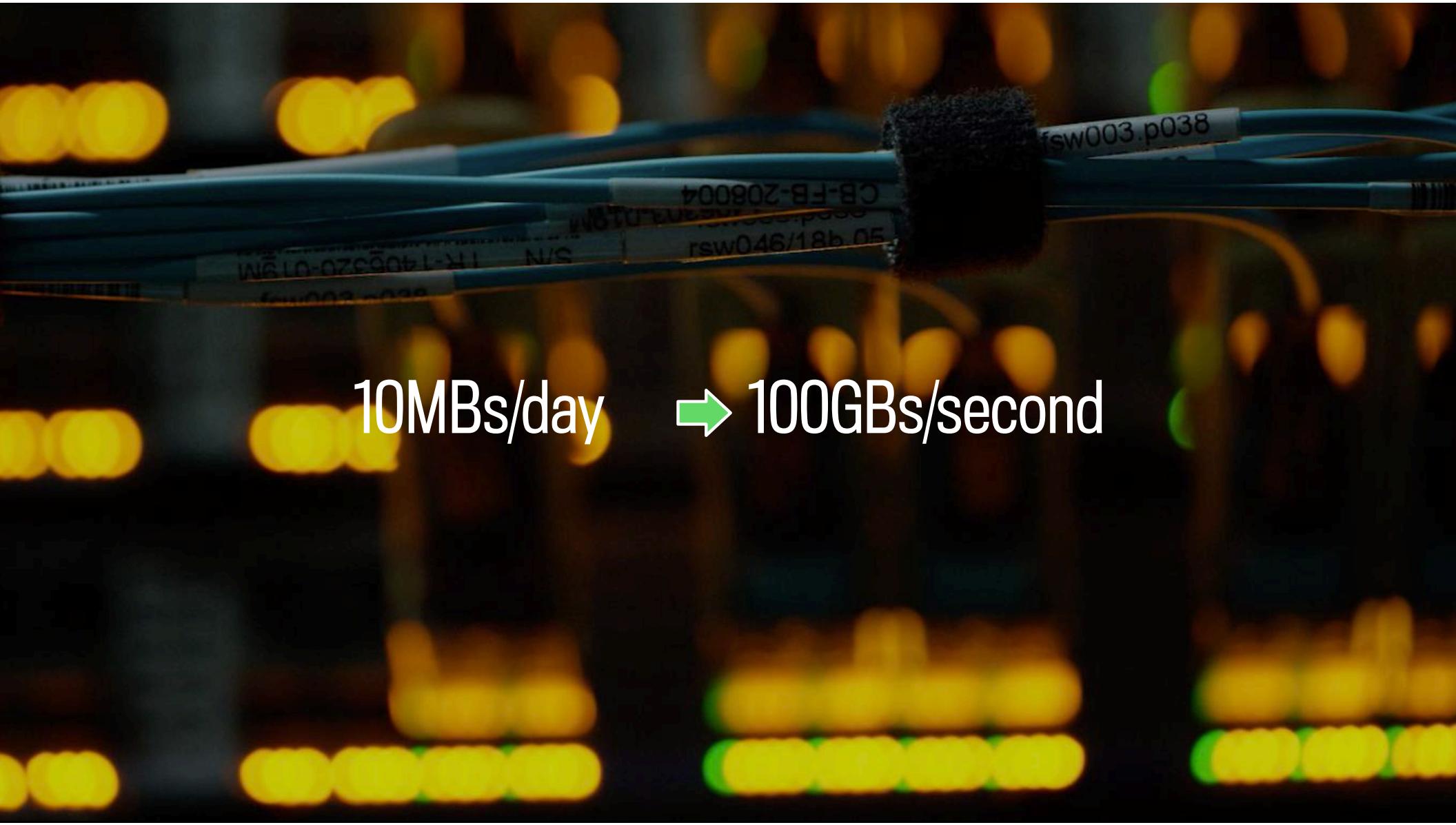


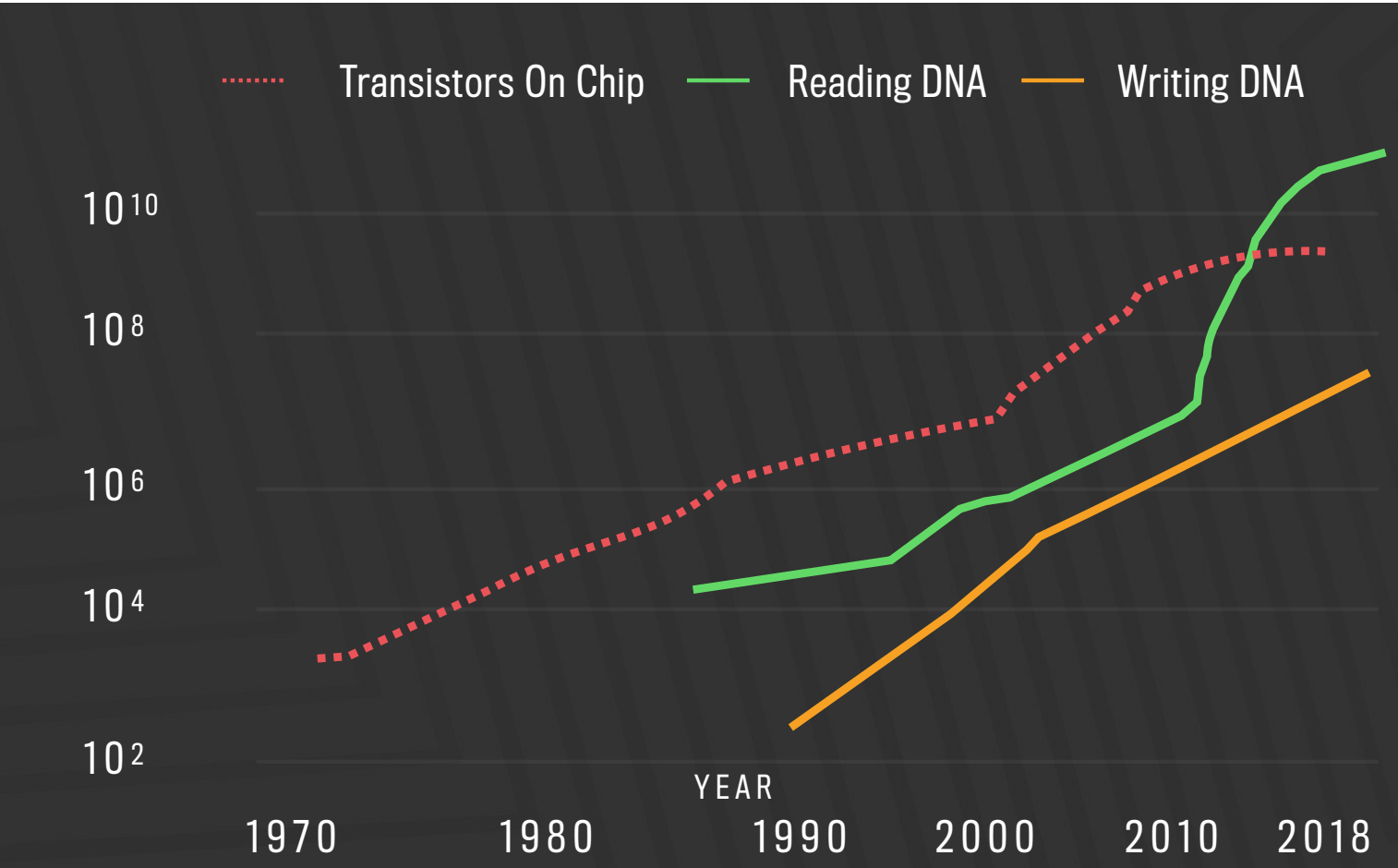




Over 700MB. 50M+ sequences. 9B+ Nucleotides,
4B+ reads. Demonstrated random access w/ 40+ objects.
Illumina and Nanopore sequencing readout.

[Nature Biotechnology'18]


fsw003.p038
10MBs/day
100GBs/second



	Quality
Life sciences	perfect
Data storage	arbitrarily tolerant

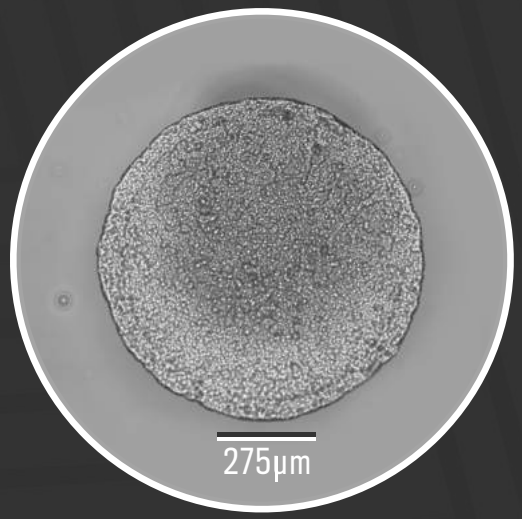
Source: Robert Carlson

Scalability

- Data centers offer perfect abstraction for “exotic technologies” (Carmean)
- Large-scale fluidics for synthesis, manipulation and sequencing
- Throughput of ~1TB/s at the data-center level
- Computational cost significant
  - Today:~2.8KB/s encode, ~1KB/s decode on 16 core Xeon.

Beyond DNA Data Storage

DNA “computing" in the age of big data

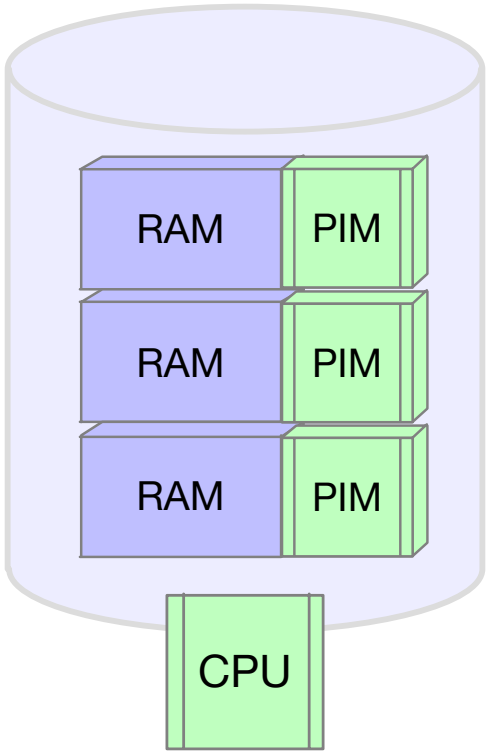


~100TB per spot

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

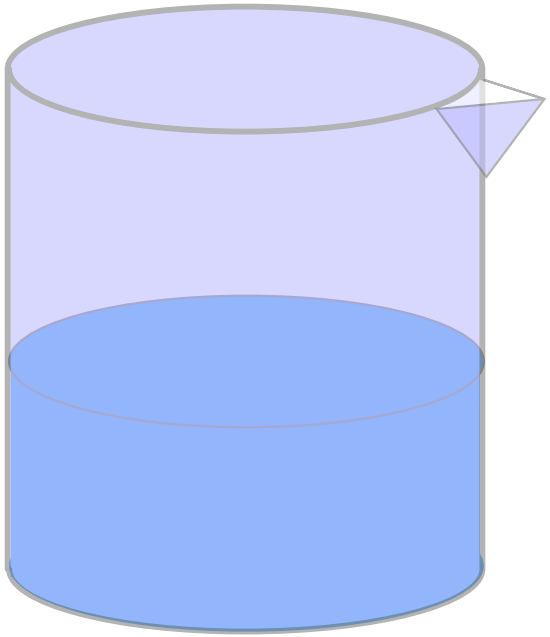
Extremely parallel and energy efficient

Processing-in-Memory



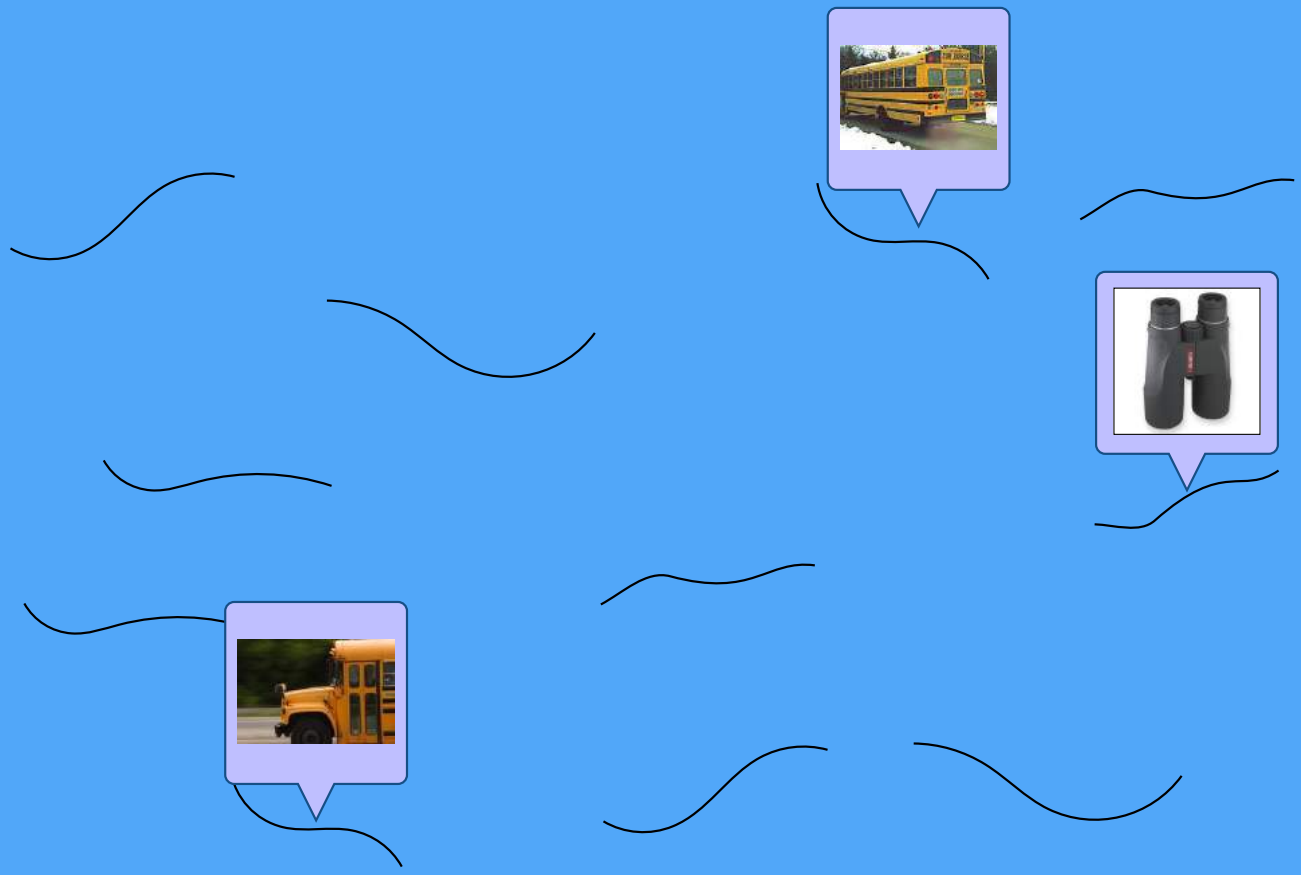
Slide credit: Kendall Stewart

Processing-in-Molecules

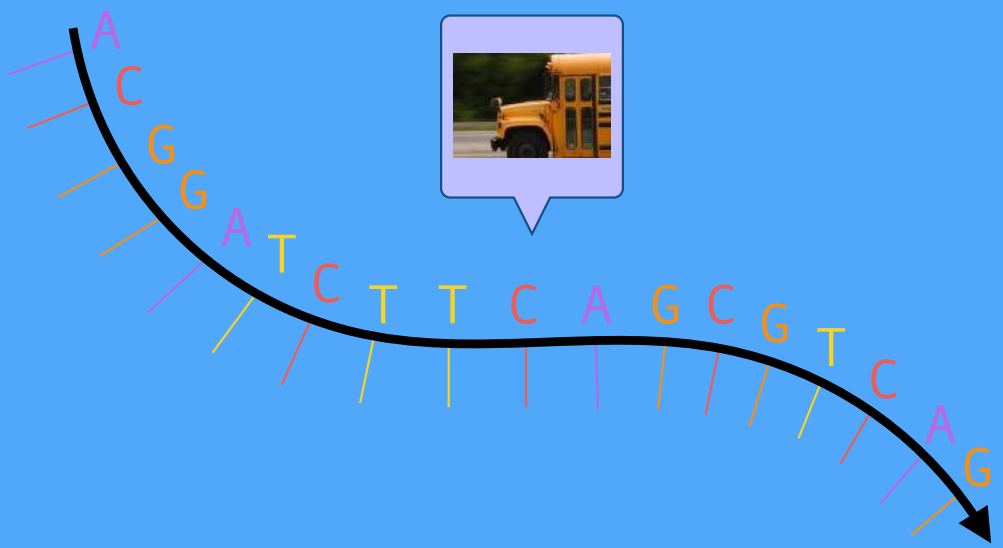


Slide credit: Kendall Stewart

Storage *and* Processing-in-Molecules (DNA)



Slide credit: Kendall Stewart


A
C
G
G
A
T
C
T
T
C
A
G
C
G
T
C
A
G

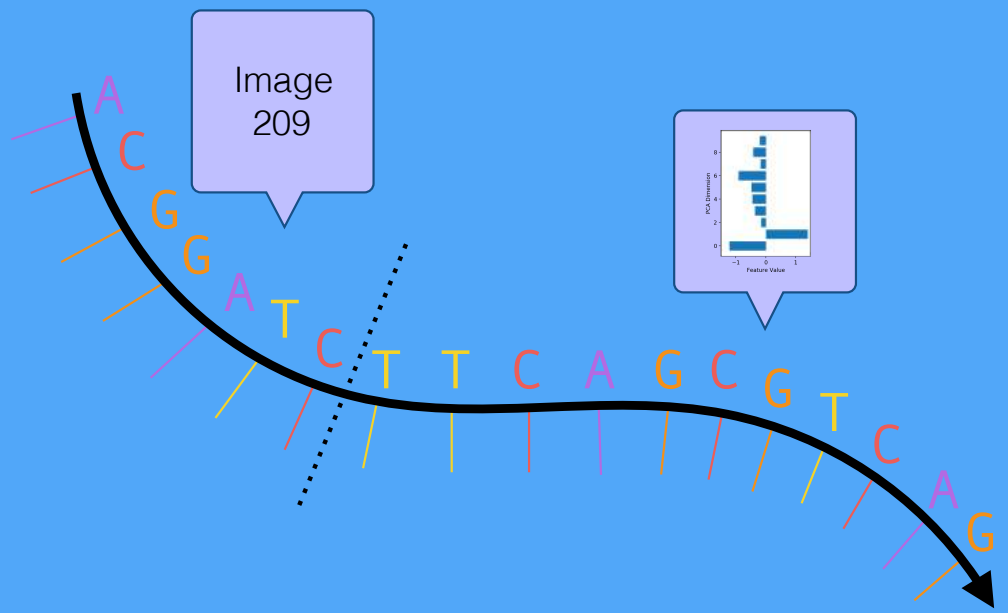


Image
209
A
C
G
G
A
T
C
T
T
C
A
G
C
G
T
C
A
G

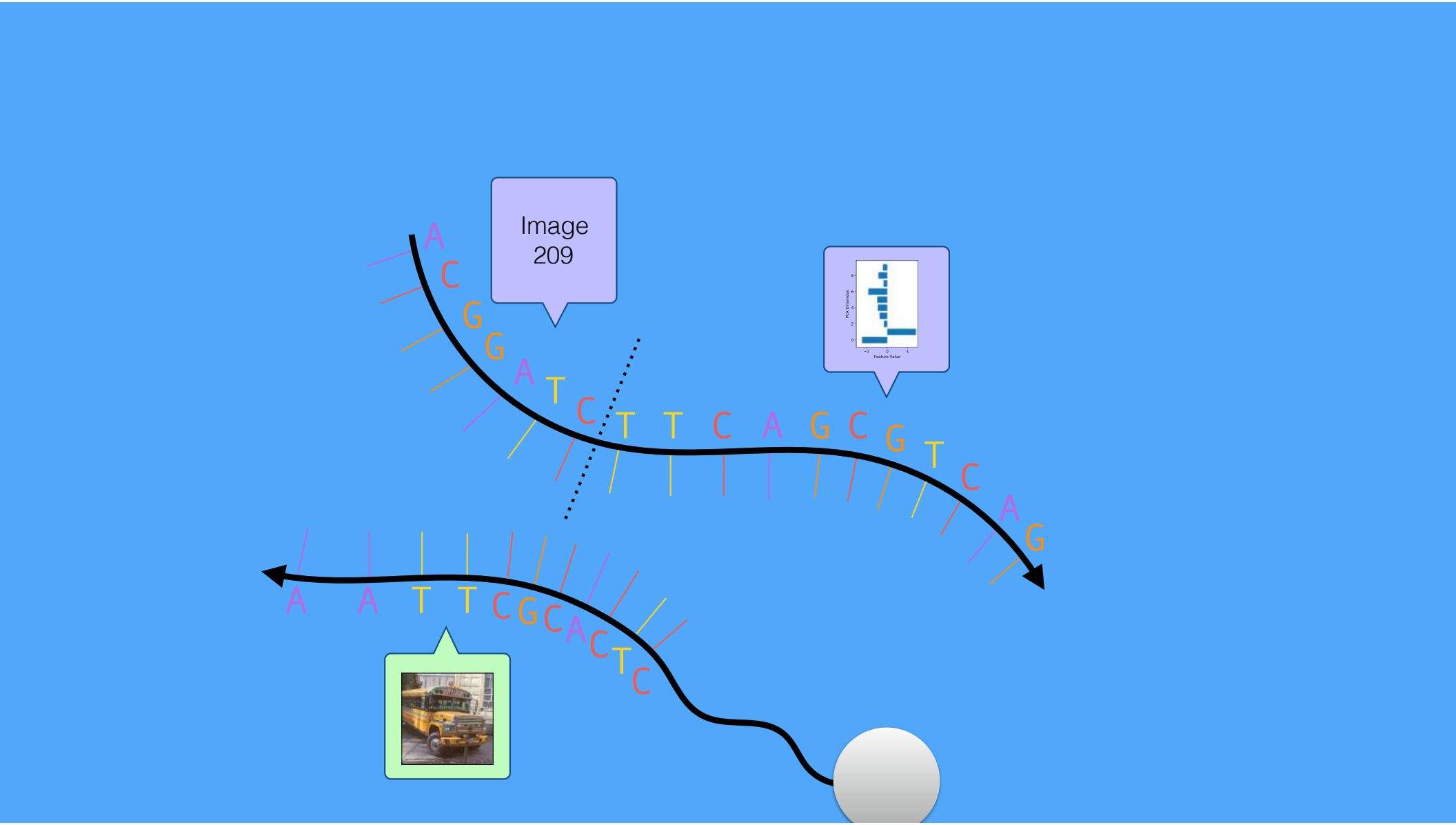


Image
209
A C G G A T C T T C A G C G T C A G
A A T T C G C A C T C

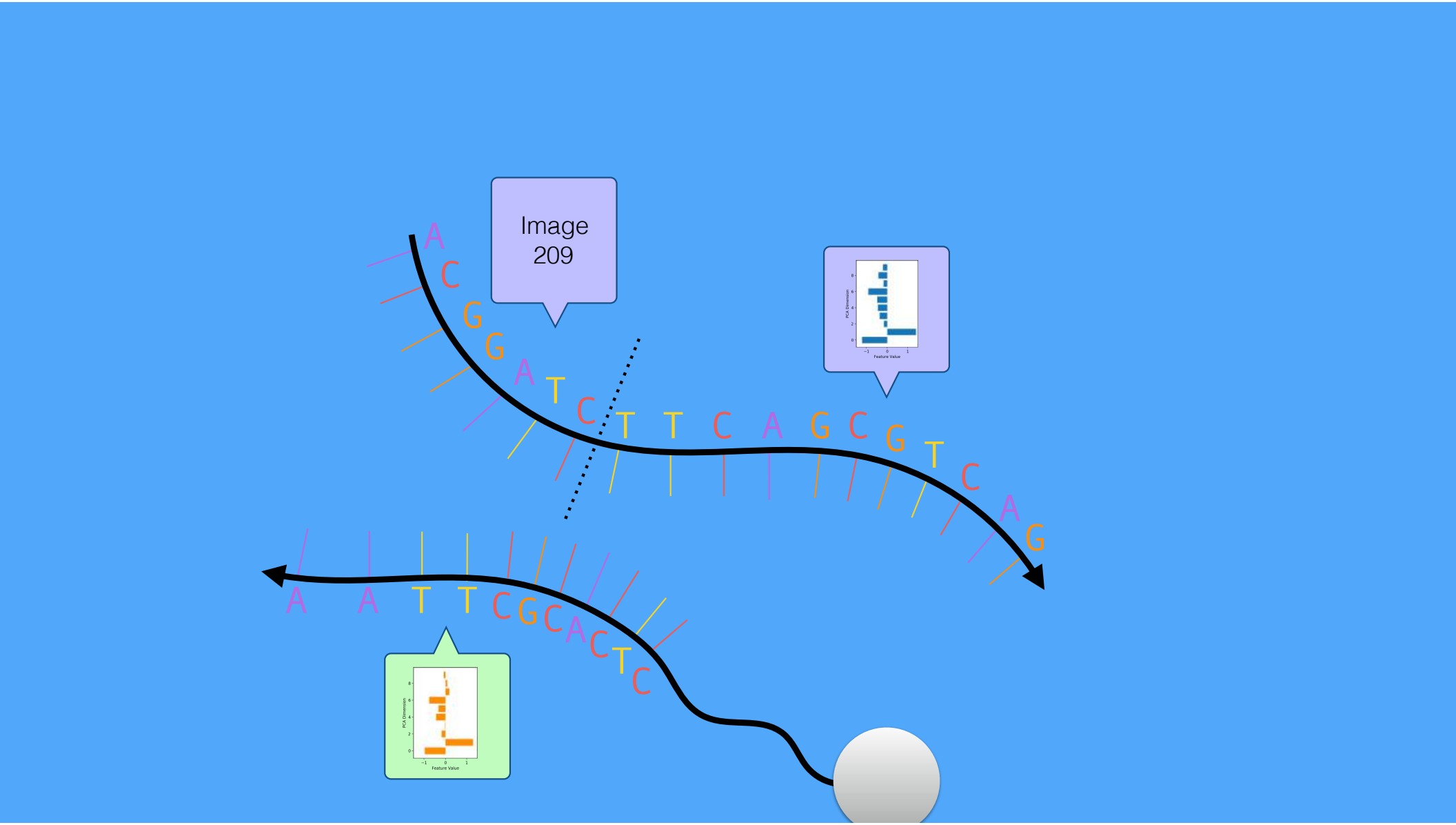
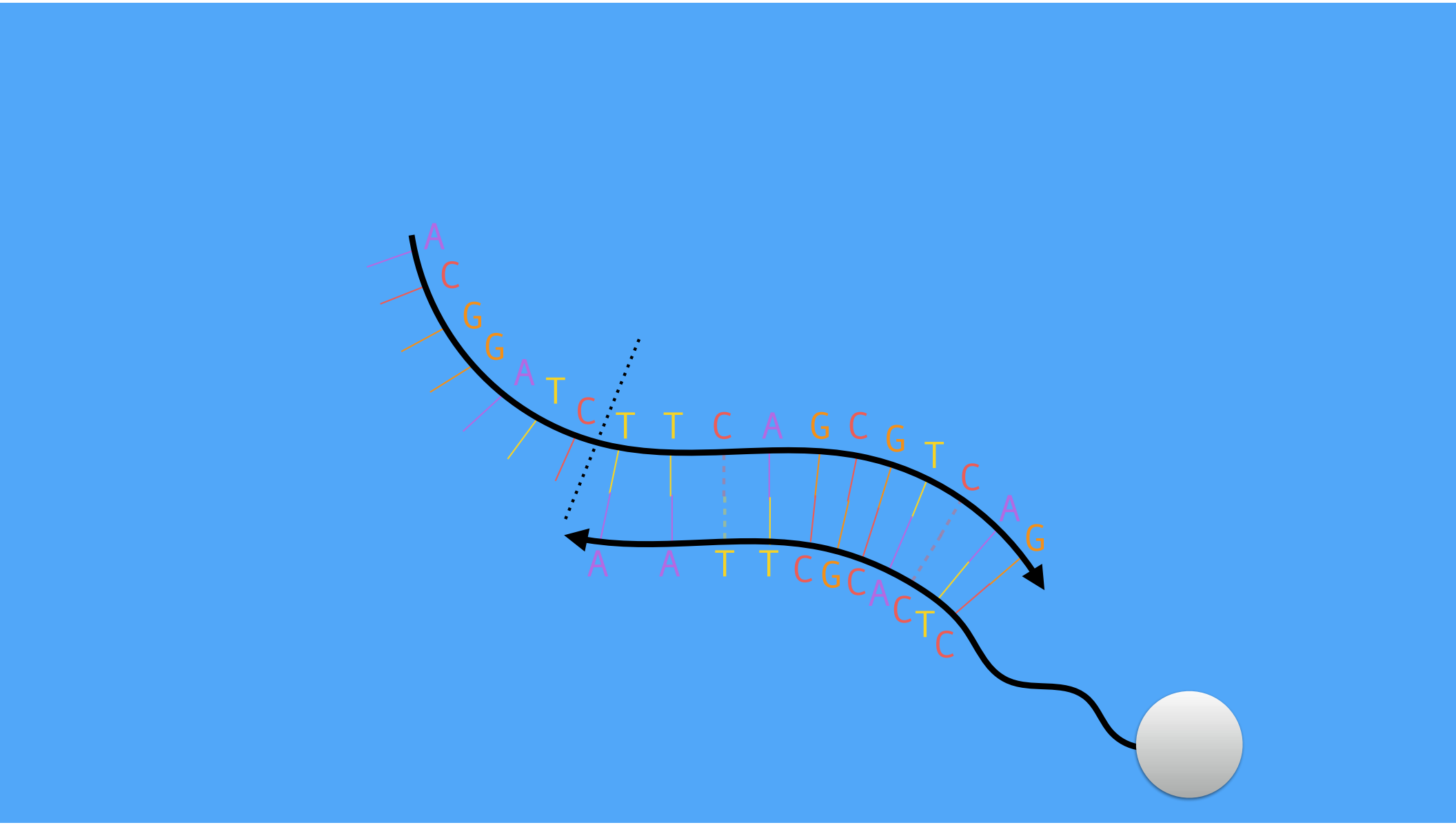
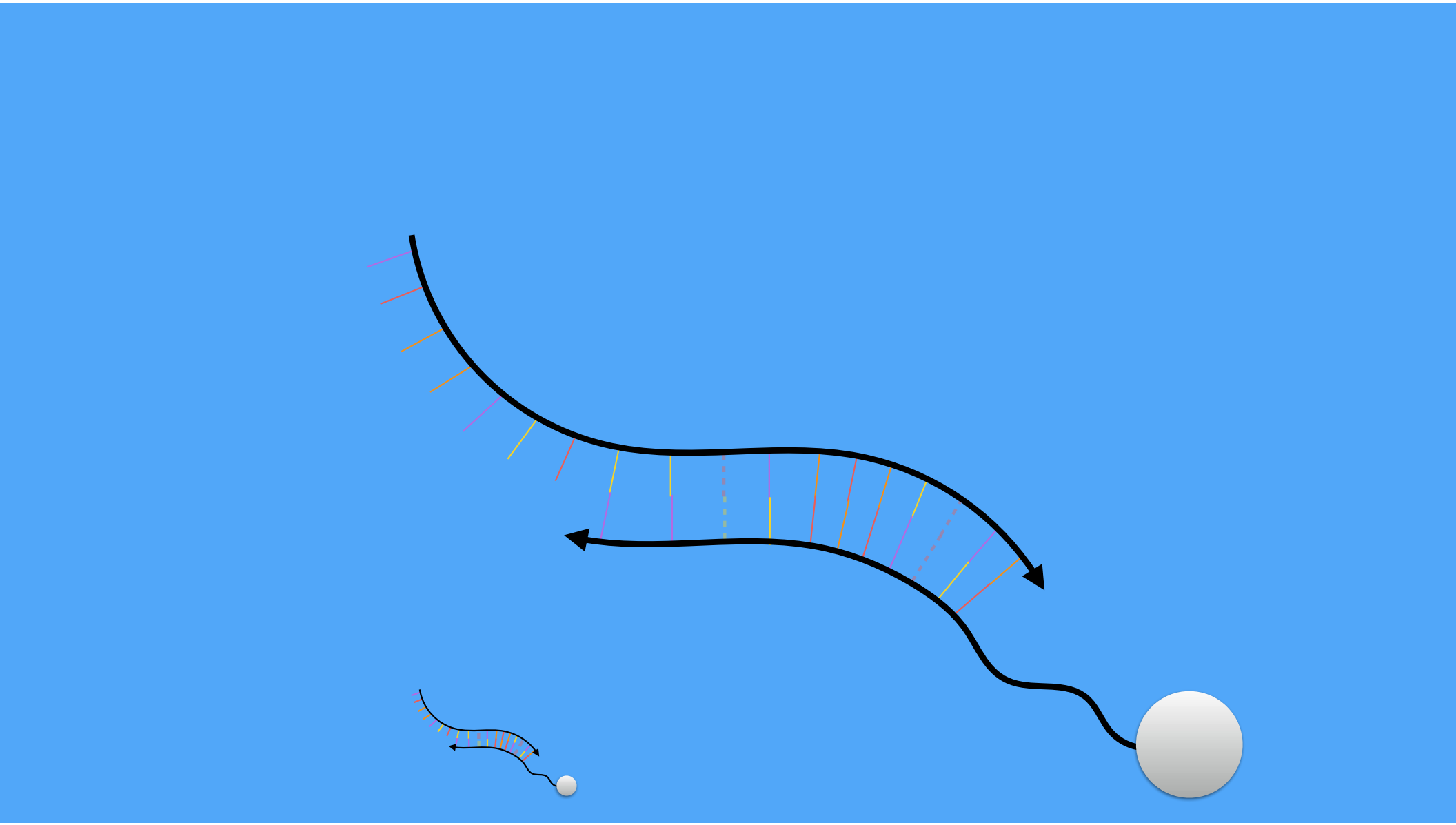
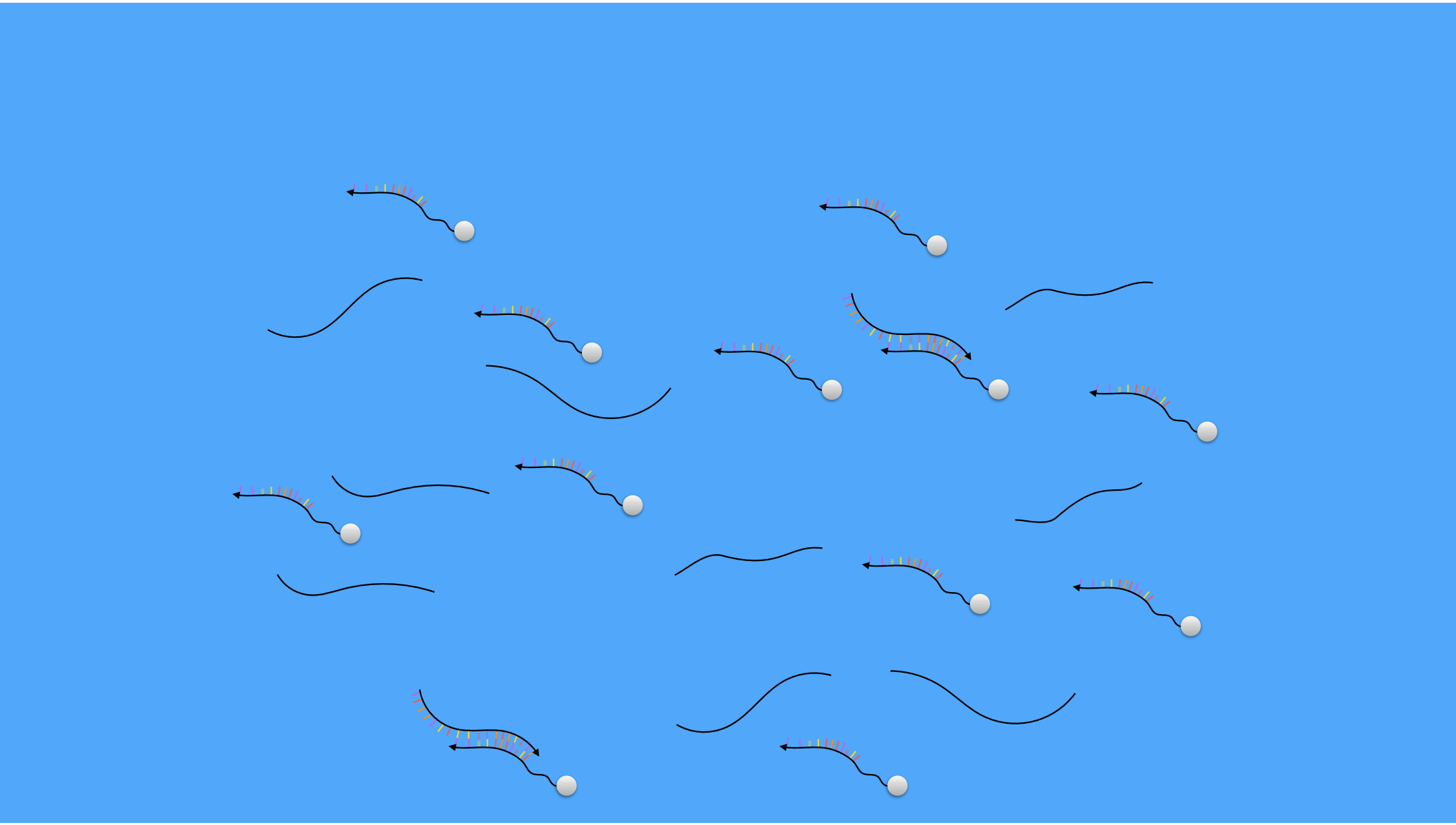
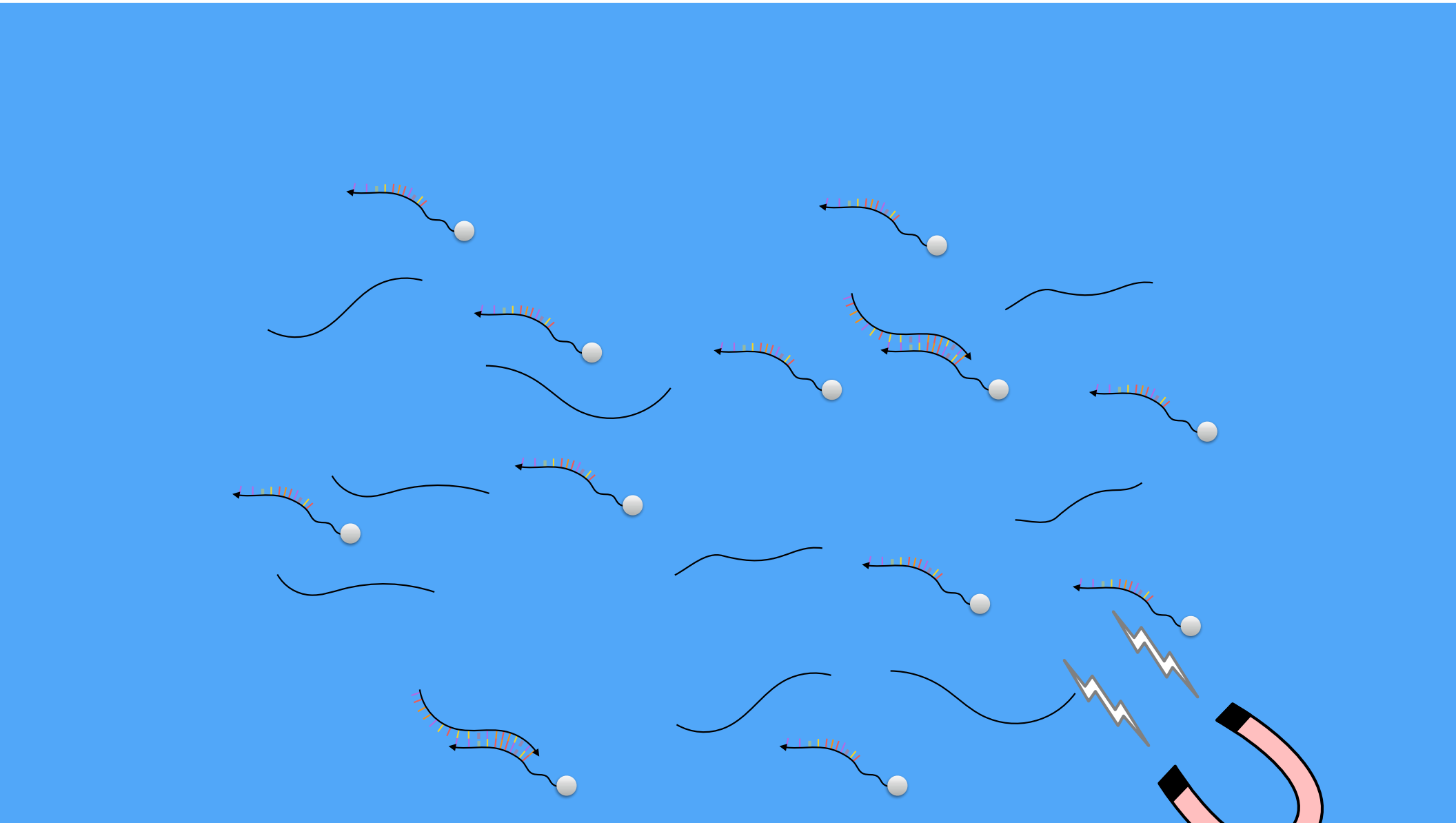


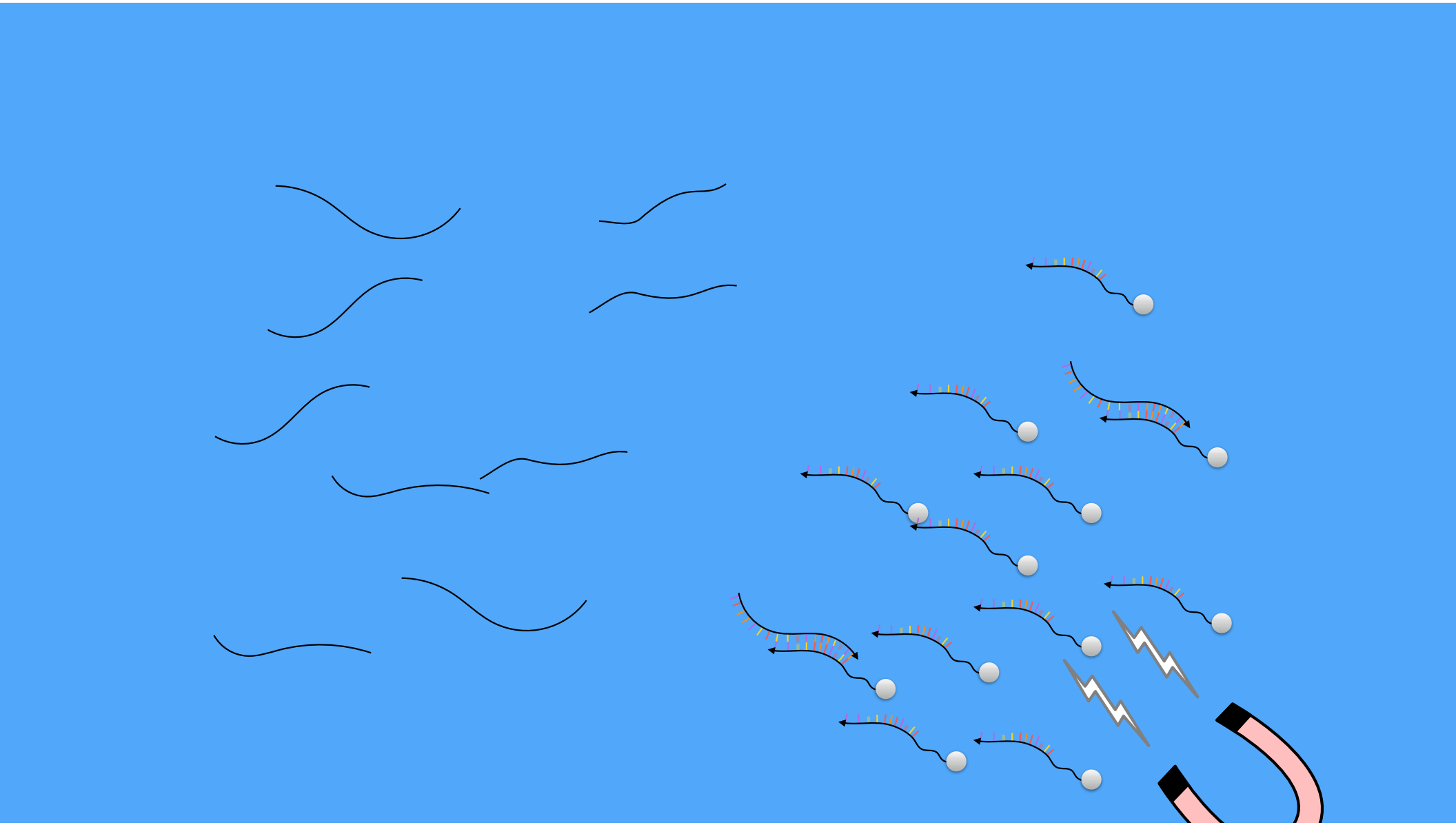
Image
209
A C G G A T C T T C A G C G T C A G
A A T T C G C A C T C
PCA Dimension
Feature Value

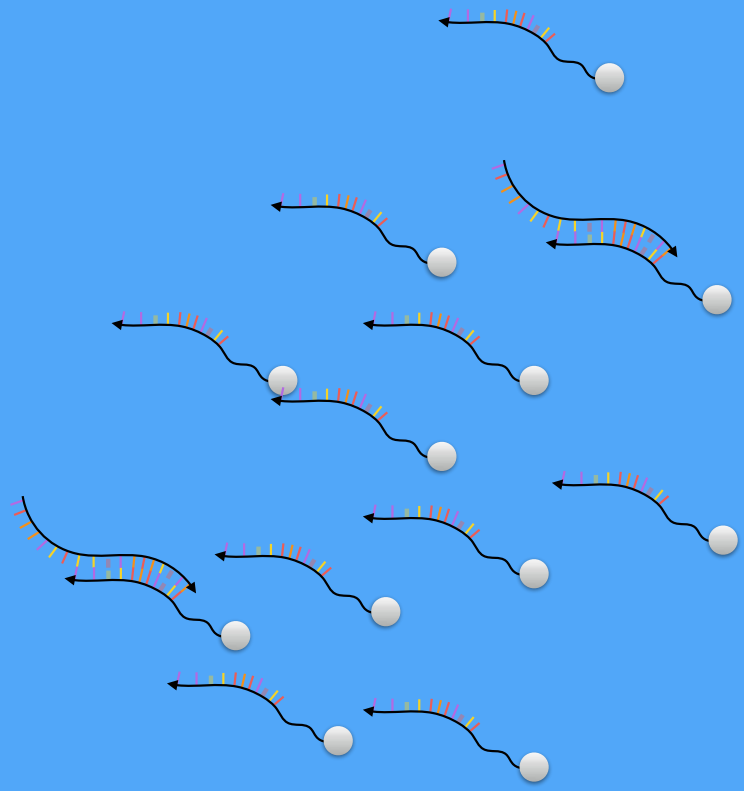

A C G G A T C T T C A G C G T C A G
A A T T C G C A C T C

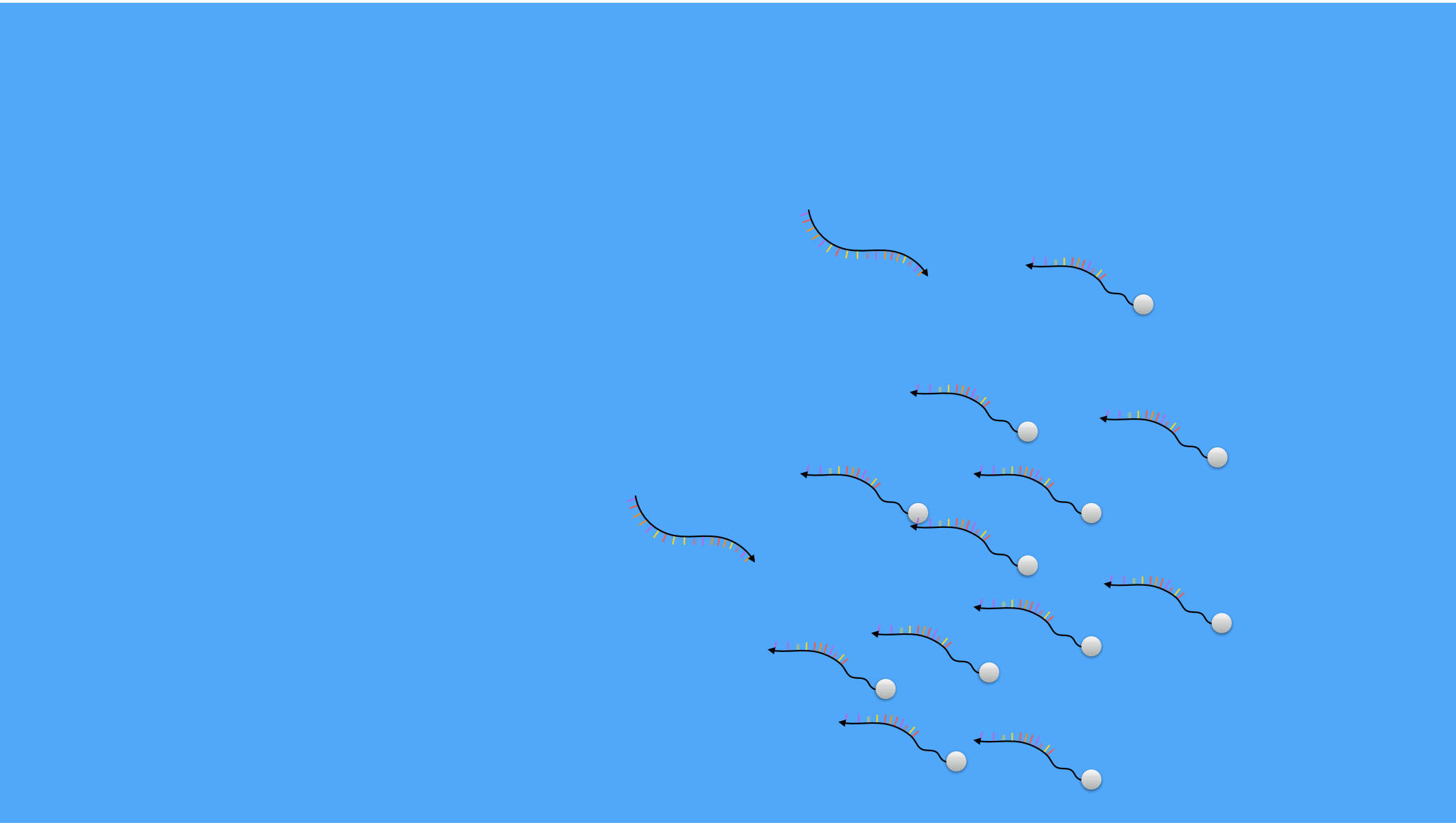


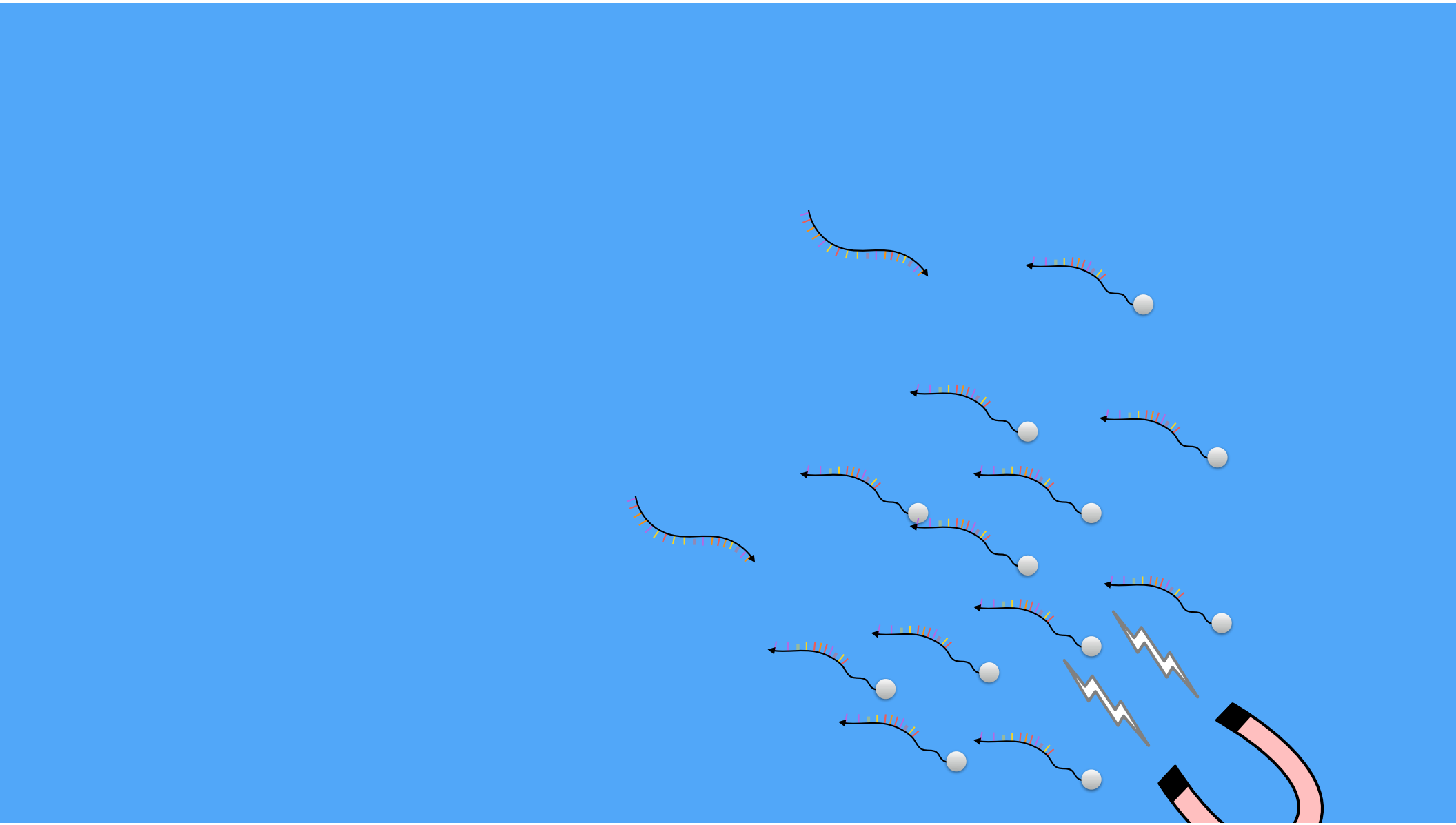




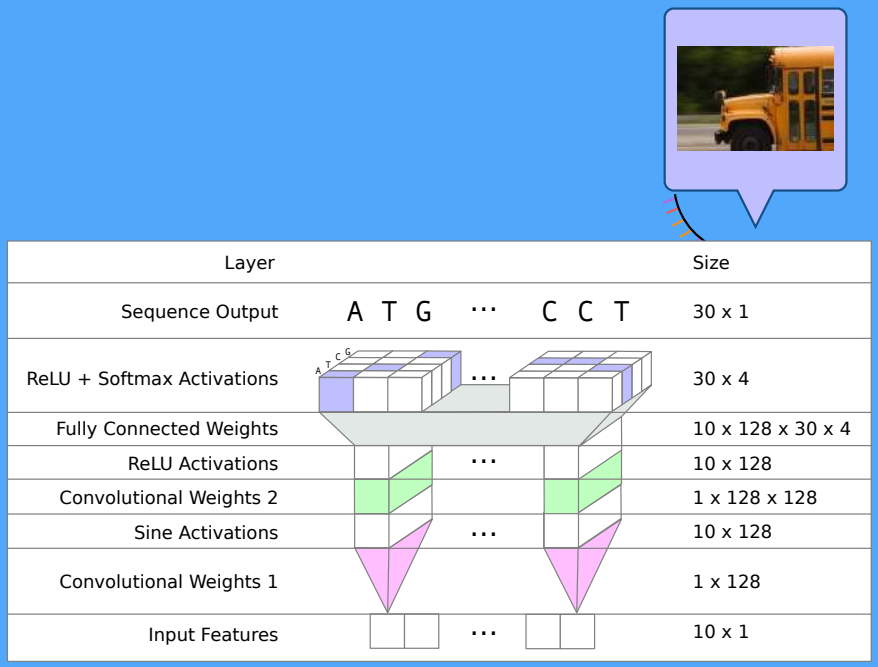


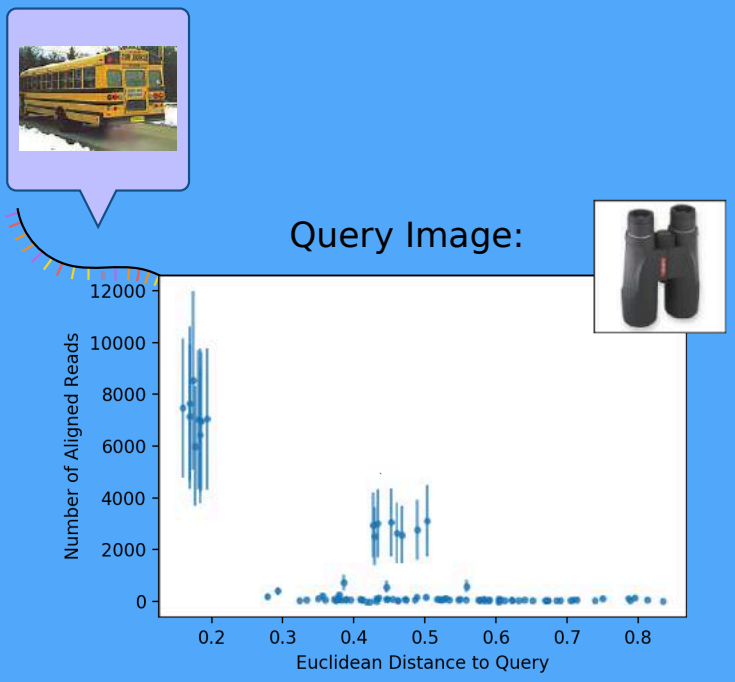


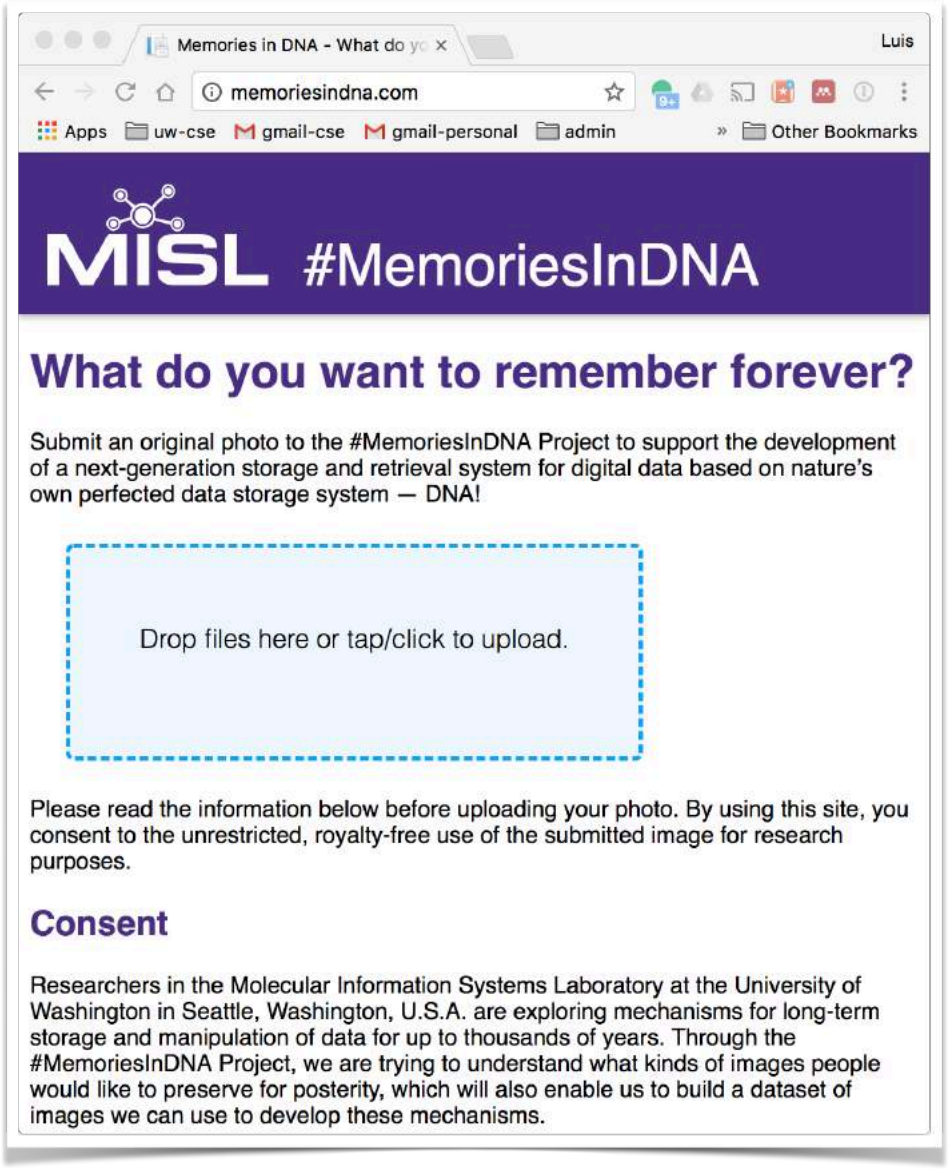


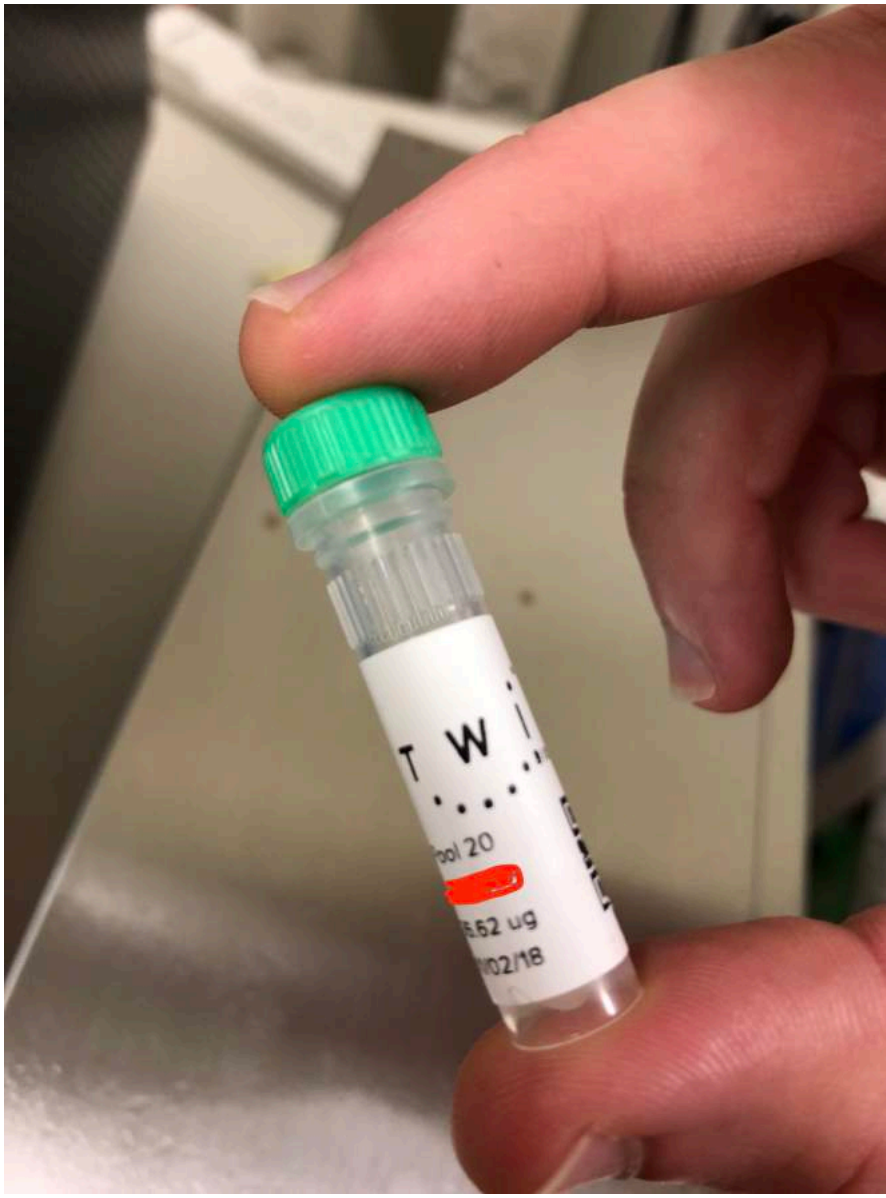


Yottabyte-scale Associate Memories?






Memories in DNA - What do yo
Luis
memoriesindna.com
Apps
uw-cse
gmail-cse
gmail-personal
admin
Other Bookmarks
MISL
#MemoriesInDNA
What do you want to remember forever?
Submit an original photo to the #MemoriesInDNA Project to support the development of a next-generation storage and retrieval system for digital data based on nature's own perfected data storage system — DNA!
Drop files here or tap/click to upload.
Please read the information below before uploading your photo. By using this site, you consent to the unrestricted, royalty-free use of the submitted image for research purposes.
Consent
Researchers in the Molecular Information Systems Laboratory at the University of Washington in Seattle, Washington, U.S.A. are exploring mechanisms for long-term storage and manipulation of data for up to thousands of years. Through the #MemoriesInDNA Project, we are trying to understand what kinds of images people would like to preserve for posterity, which will also enable us to build a dataset of images we can use to develop these mechanisms.



MISL

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Microsoft

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

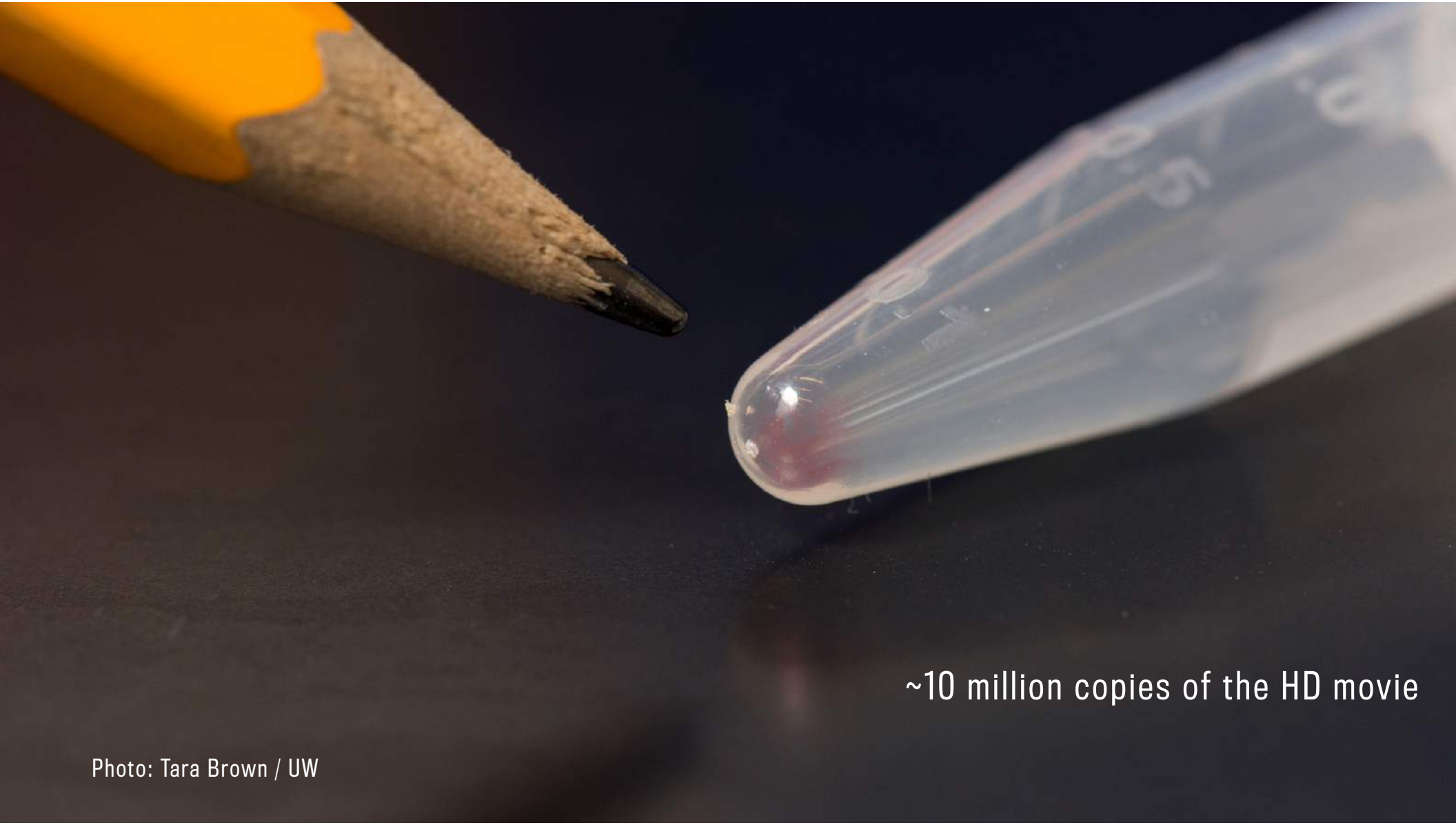


Photo: Tara Brown / UW


0:30

Thank you!

Microsoft