Borrowing from Nature to Build Better Computers

Luis Ceze
Molecular Information Systems Laboratory
Paul G. Allen School of Computer Science and Engineering
University of Washington

joint work with Karin Strauss, Georg Seelig, Doug Carmean, Sergey Yejhanin, Lee Organick, Yuan-Juye Chen, Bichlien Nguyen, Chris Takahashi, Sharon Newmann, Cyrus Rashtchian, Siena Ang, David Ward, Randolph Lopez, Max Willsey, Kendall Stewart, James Bornholt.

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

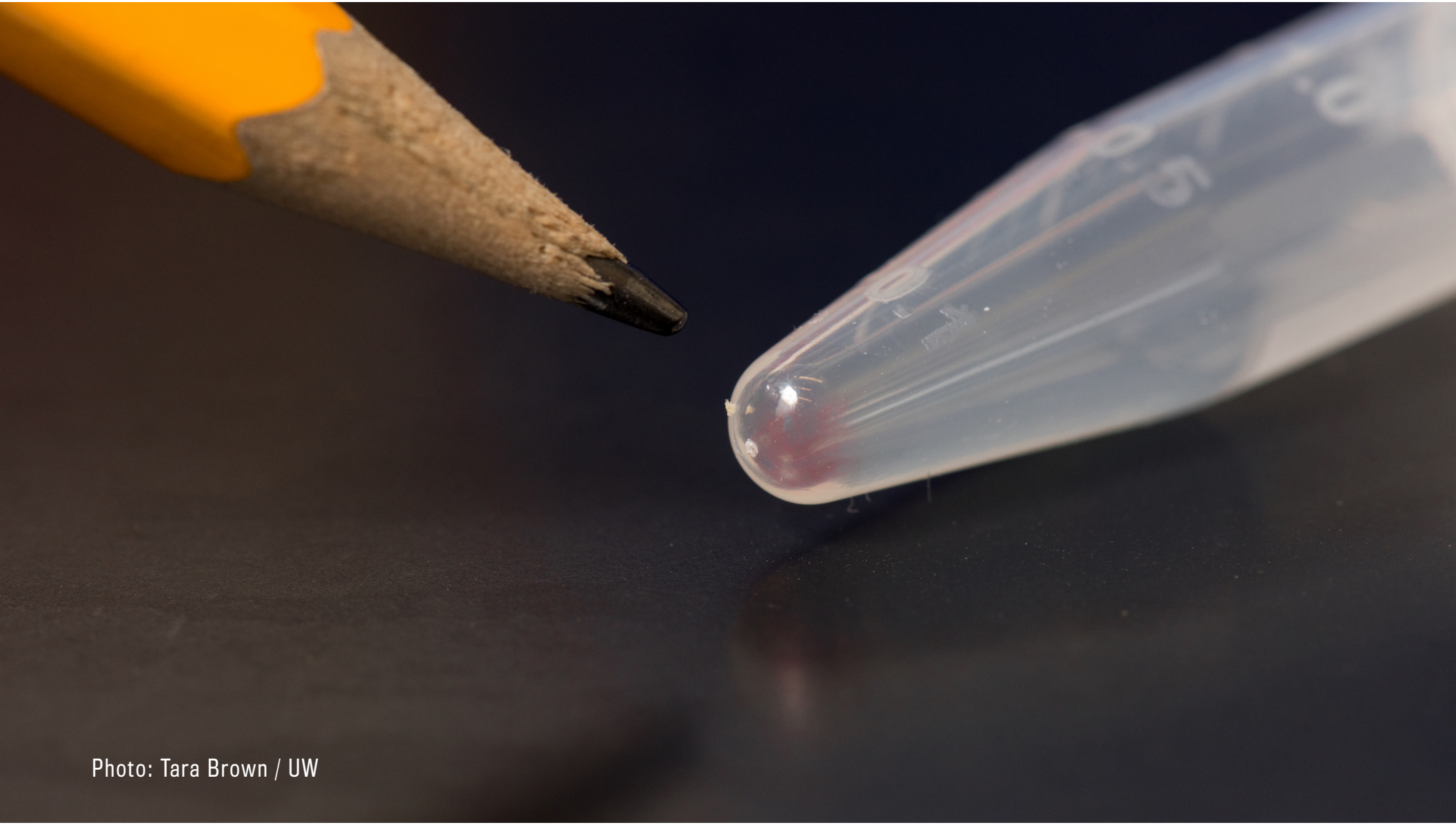
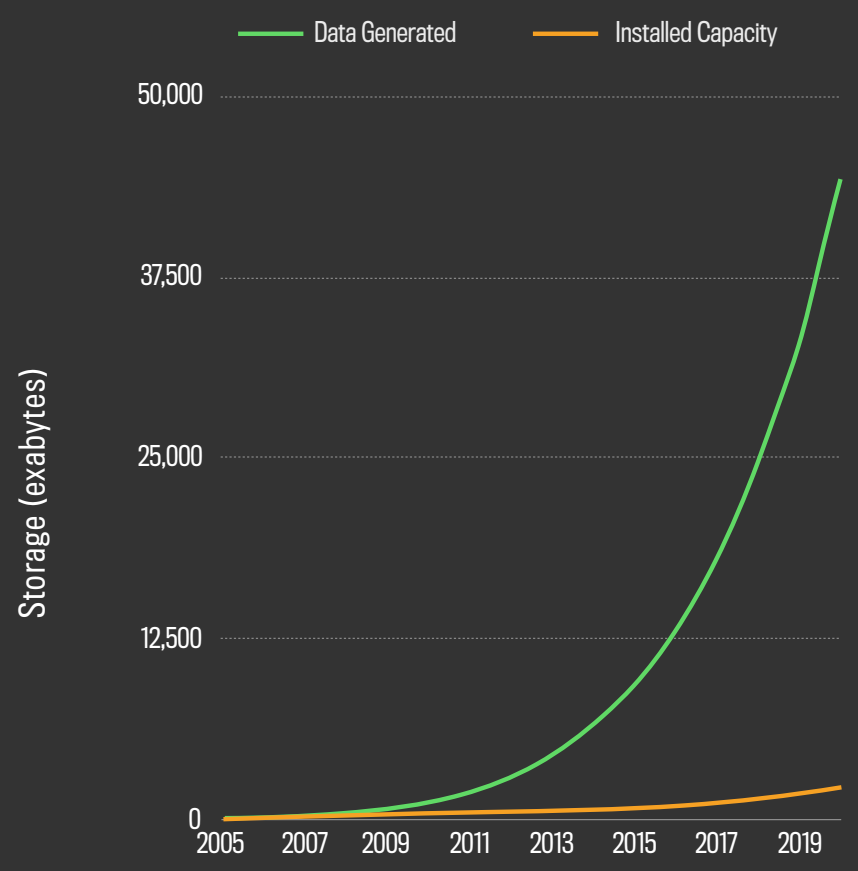
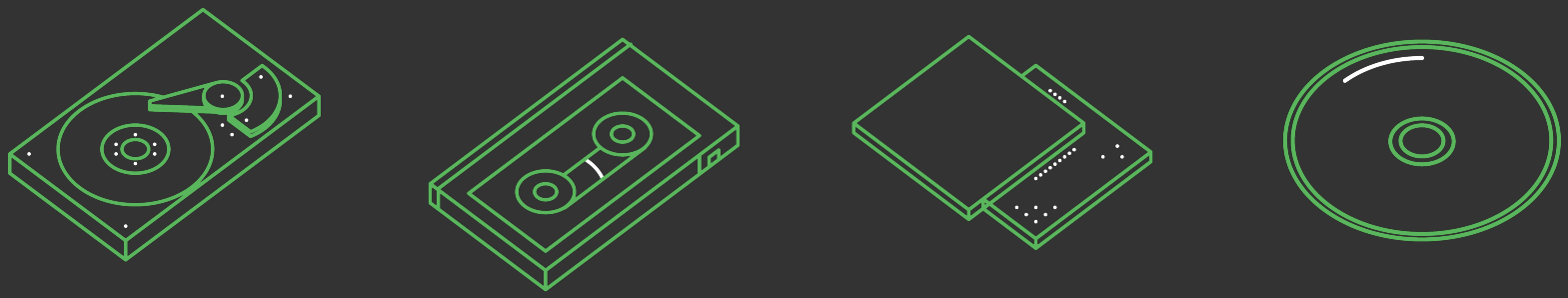
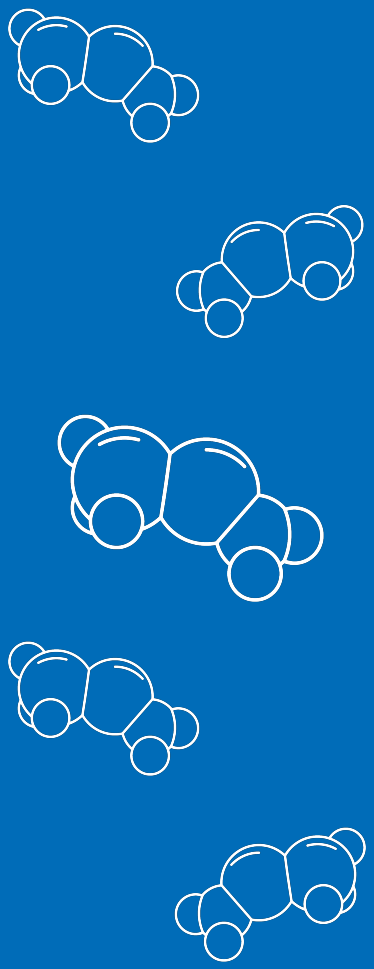


Photo: Tara Brown / UW

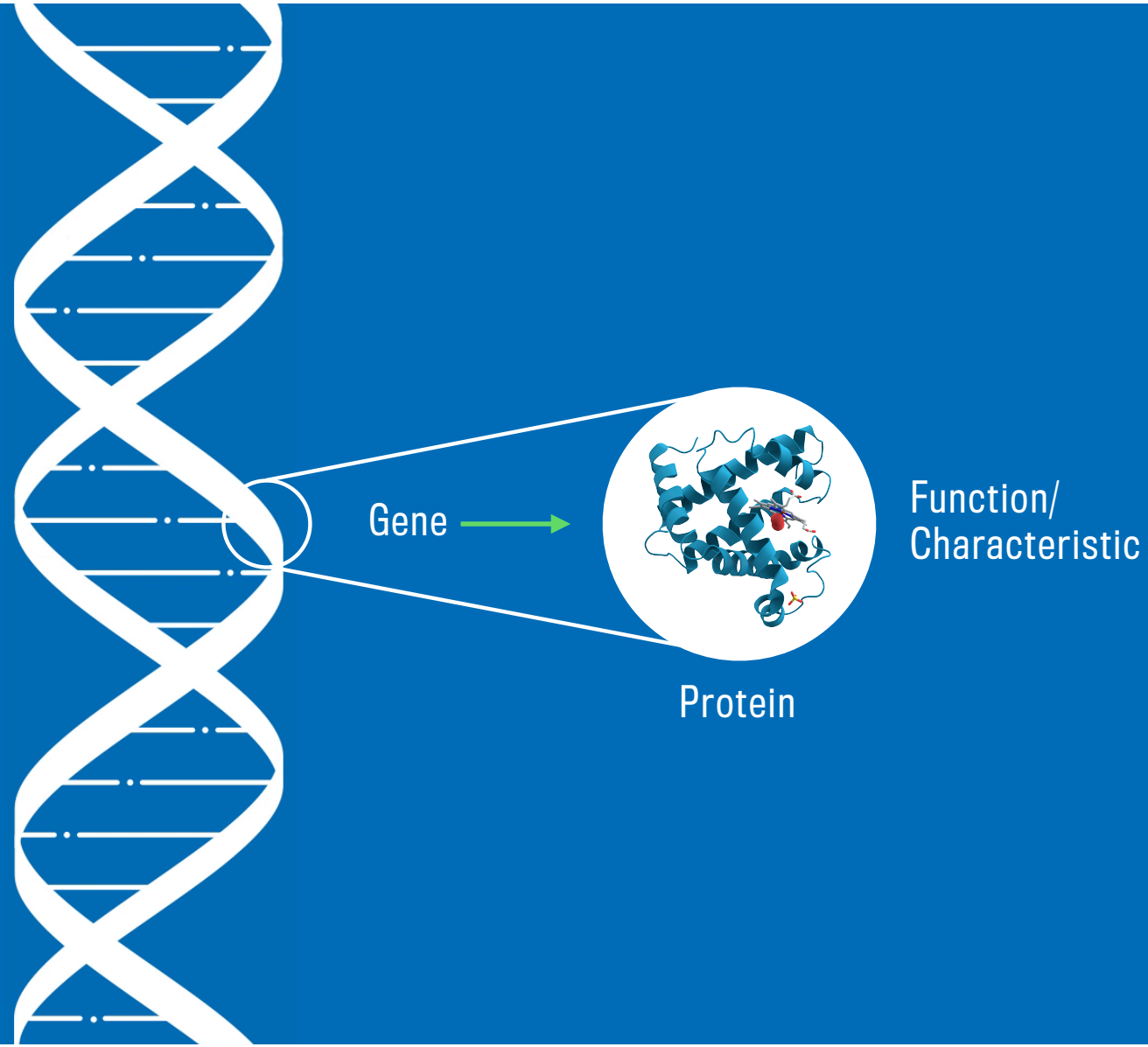

Data Generated
Installed Capacity
Storage (exabytes)
50,000
37,500
25,000
12,500
0
2005
2007
2009
2011
2013
2015
2017
2019

Data: IDC



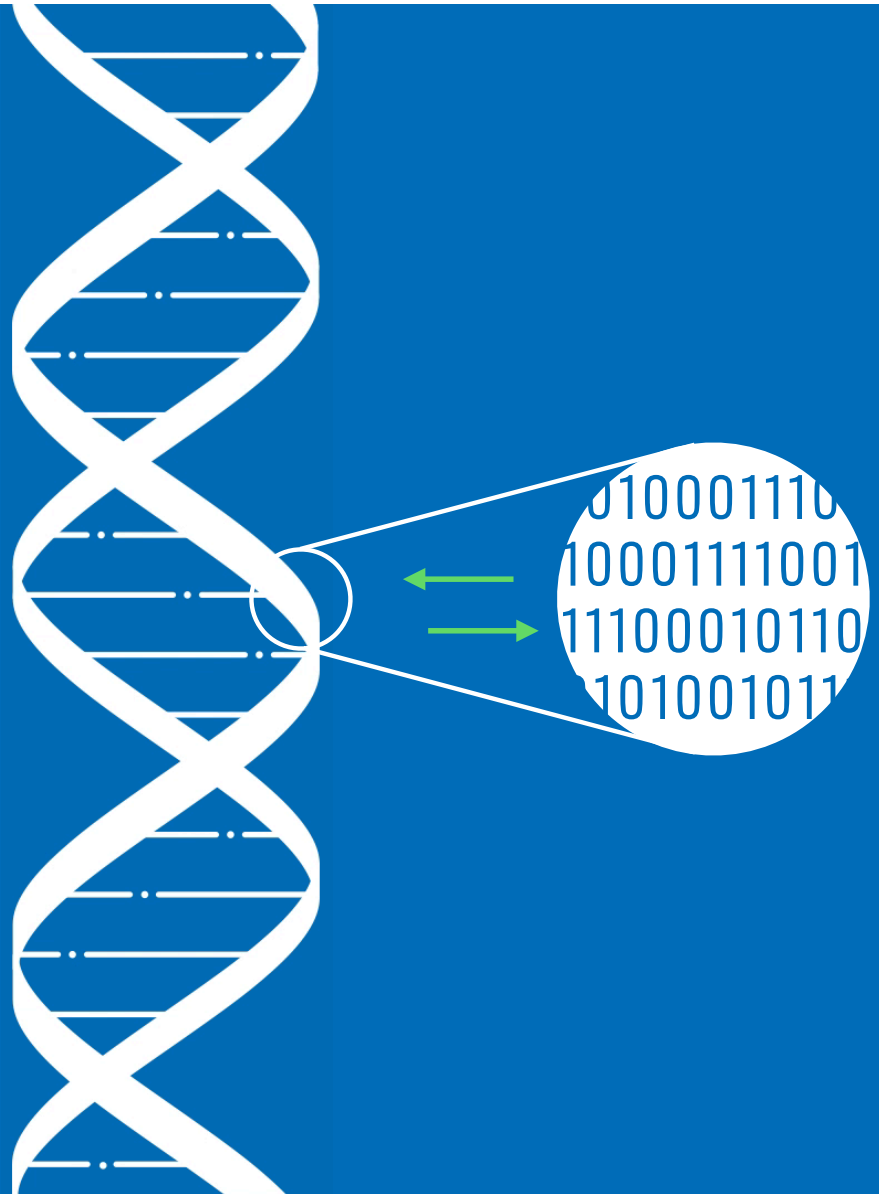


Nature Evolved a Fantastic Storage Medium: **DNA**



Using Synthetic **DNA** for Data Storage

Manufactured DNA



DNA Molecules for Digital Data

Extremely Dense

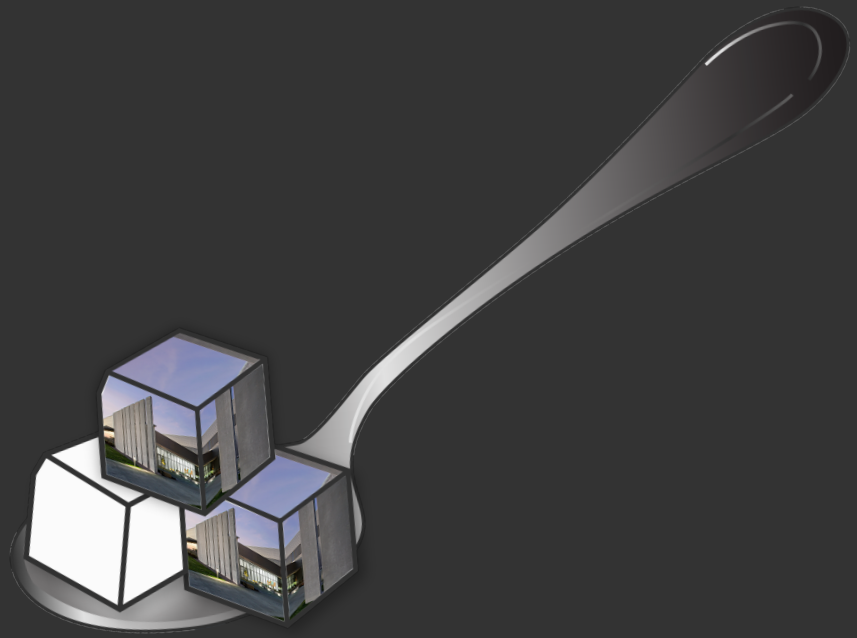
1 Exabyte in 1 in^3

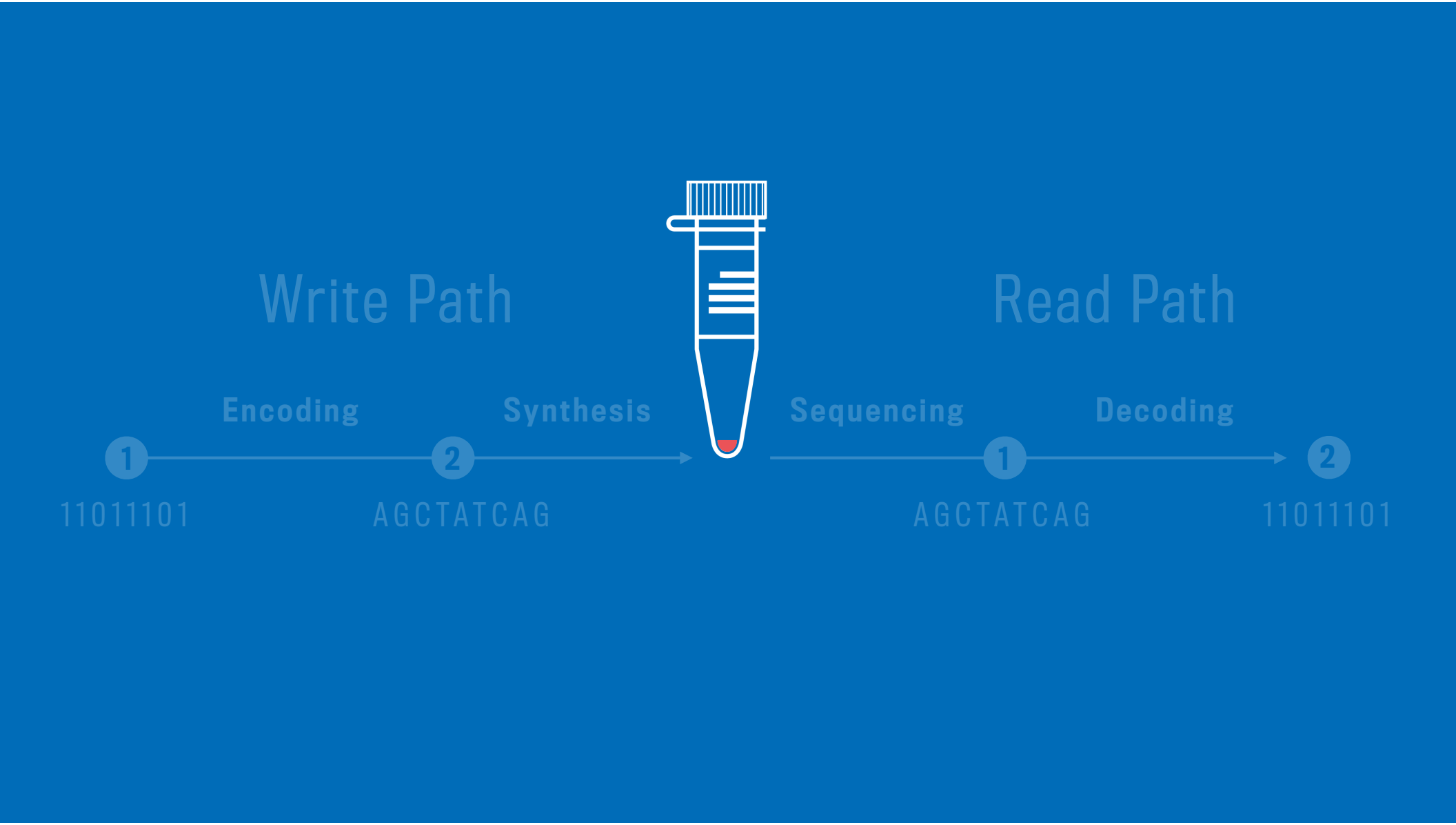
Extremely Durable

100,000s of Years

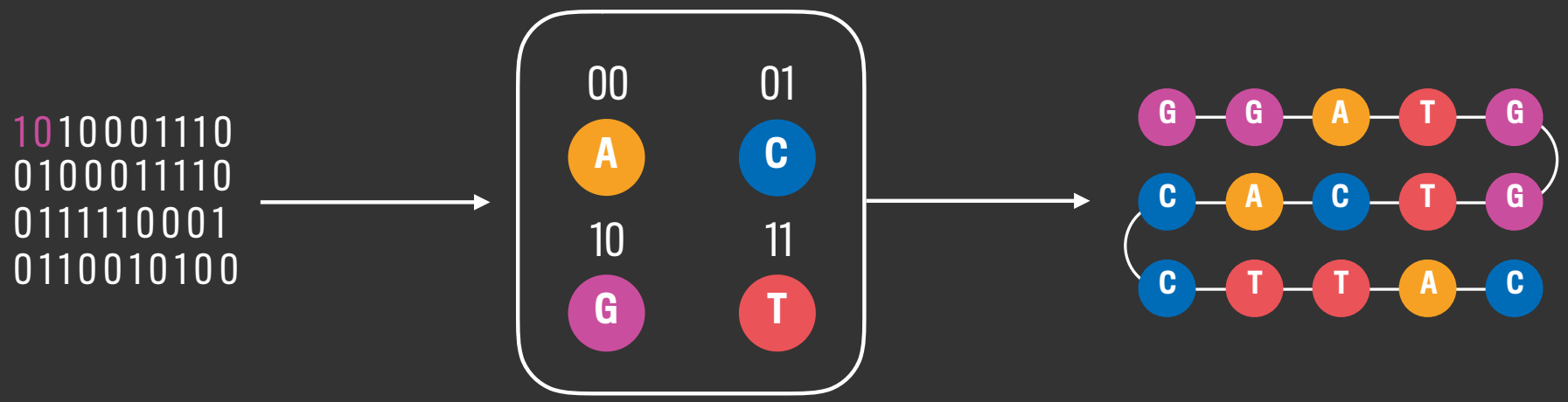
Making Copies Is Nearly Free

Never Gets Obsolete




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

Encoding Digital Data in DNA

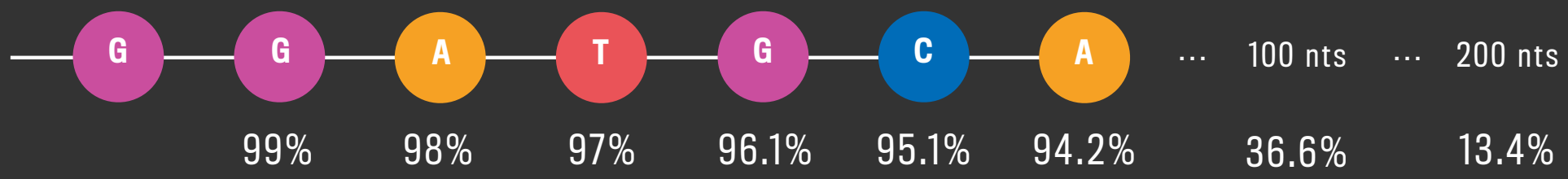


Encoding Digital Data in DNA

Repeated Letters Are Bad: Avoid Them with Clever Coding.

Synthetic DNA Sequences Have Limited Length: Break It Up

P[Attach] = 99%

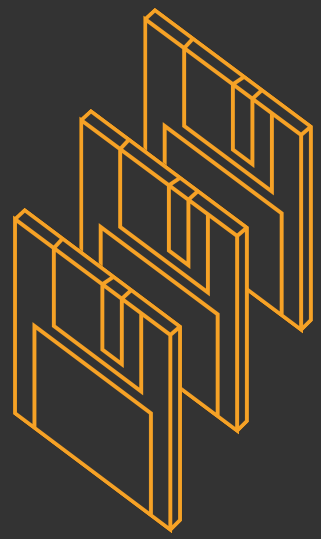


Breaking up Data into Chunks (~150nts)

CGAT GCAC TGCT GACG GCTA GCTC

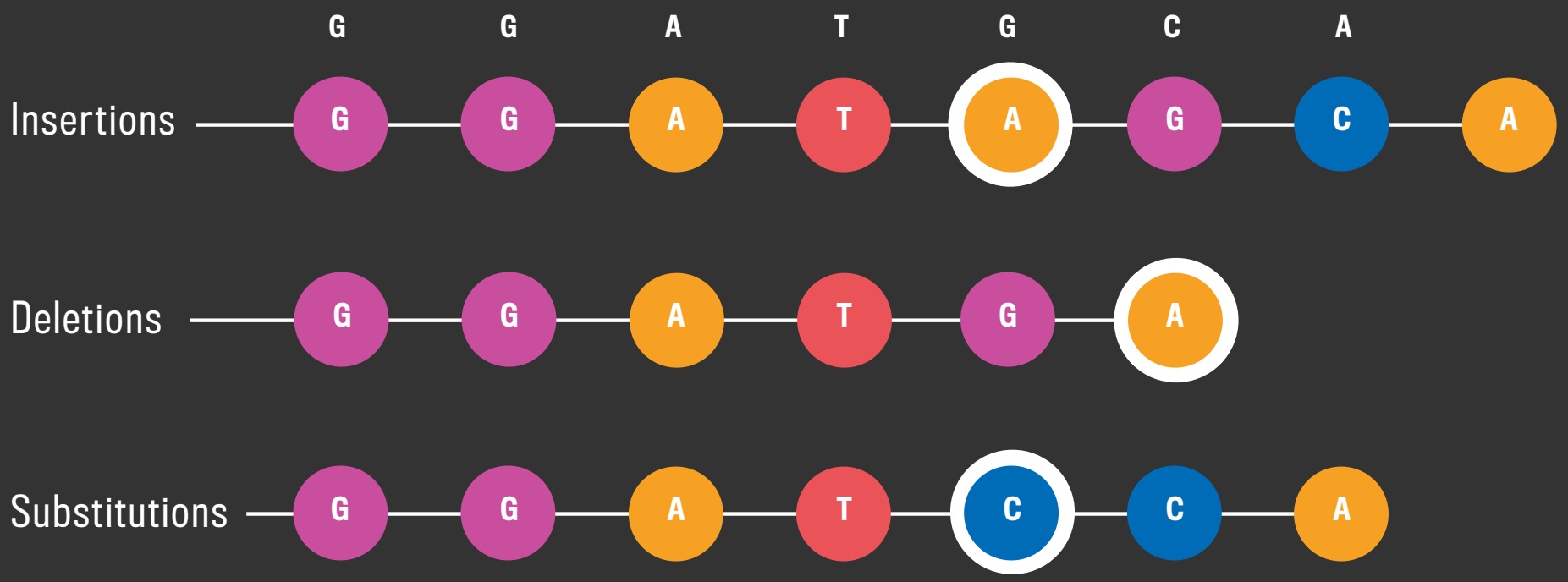
Breaking up Data into Chunks (~150nts)

Key Identifiers ("Primers")			Addresses Within the File	Key Identifiers ("Primers")
ATGTT	CGAT	GCAC	AAAA	CATCC
ATGTT	TGCT	GACG	AAAC	CATCC
ATGTT	GCTA	GCTC	AAAG	CATCC



~ 20 Bytes per DNA Strand. Many Strands per File.

Errors in Writing/Reading DNA

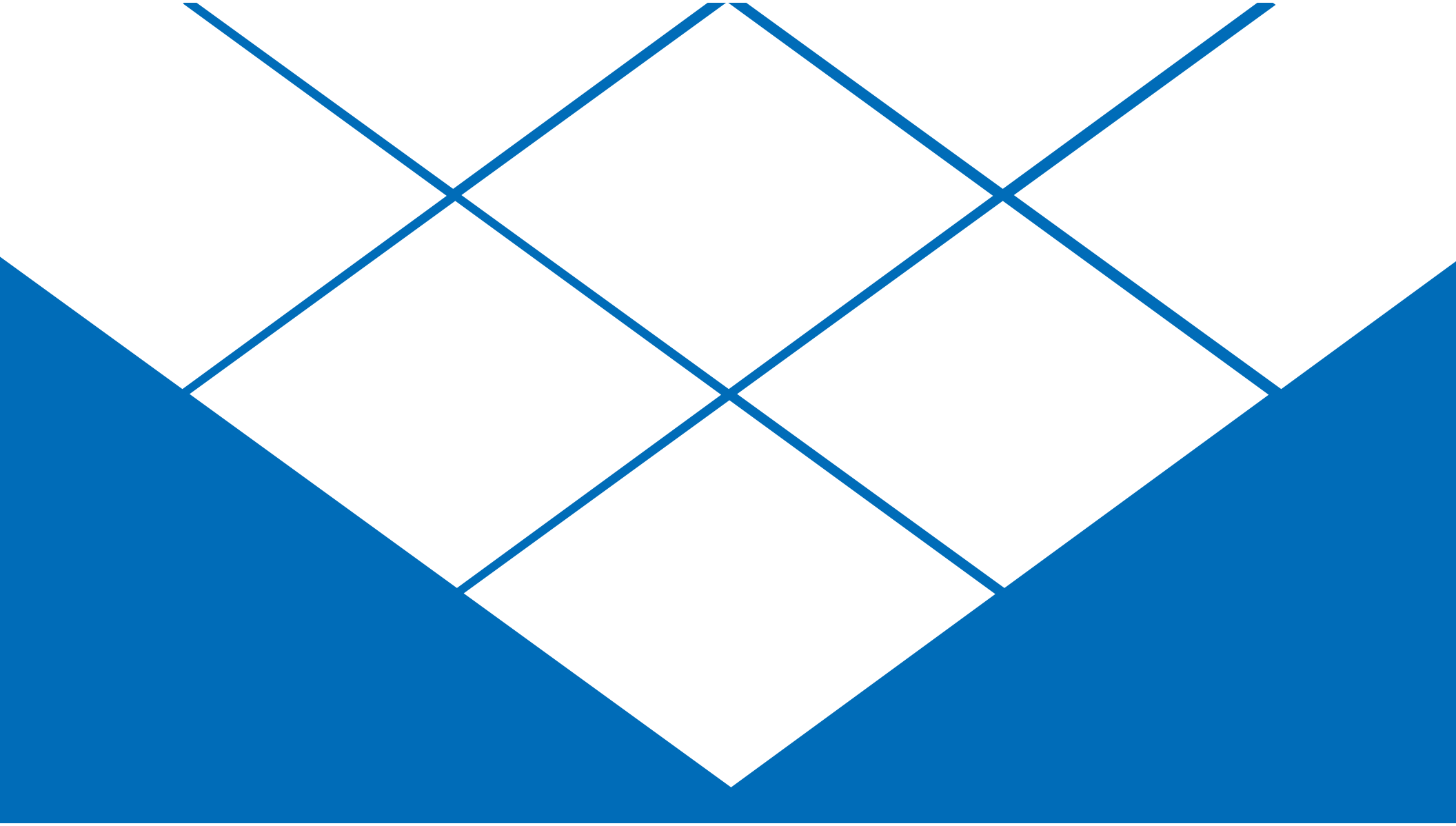


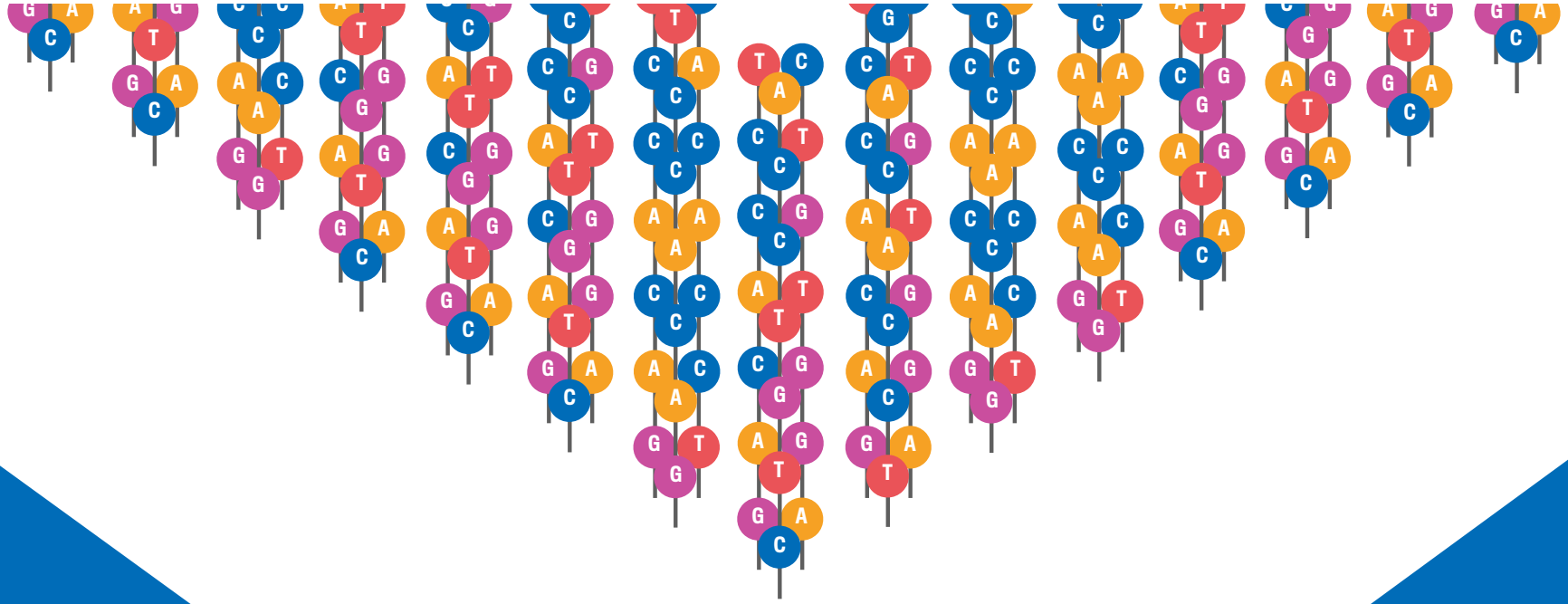
Aggregate Error Rates ~1%

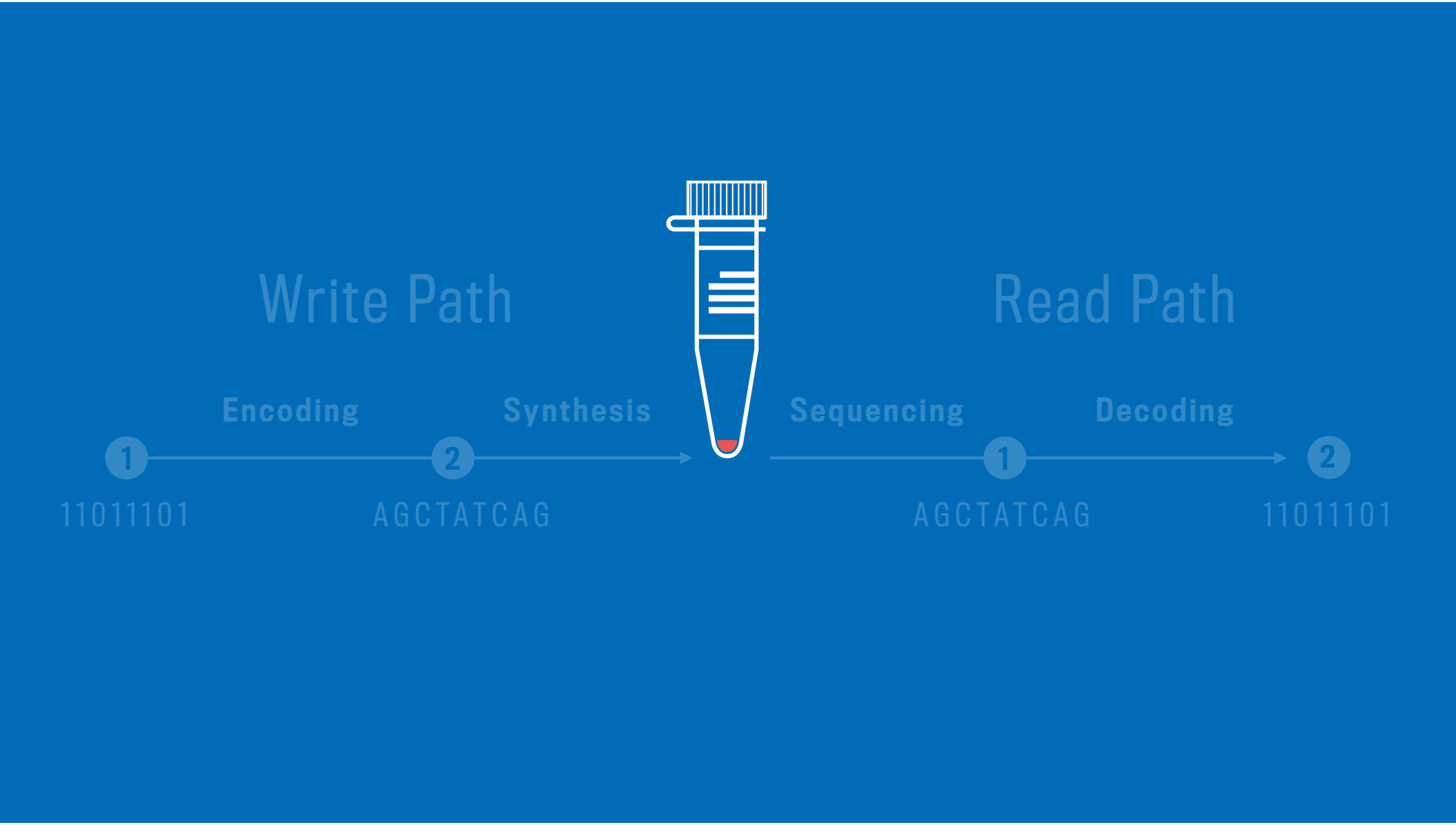
Adding Redundant Information

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

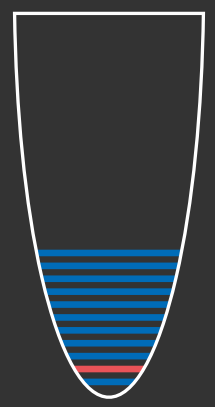
Redundant Data in Additional DNA Strands. Many Possibilities: Parity, Reed Solomon, LDPC, ...






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

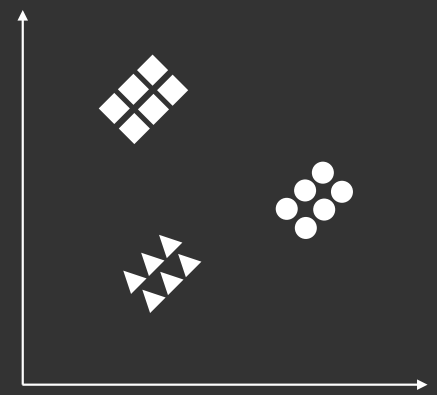
DNA Sequencing



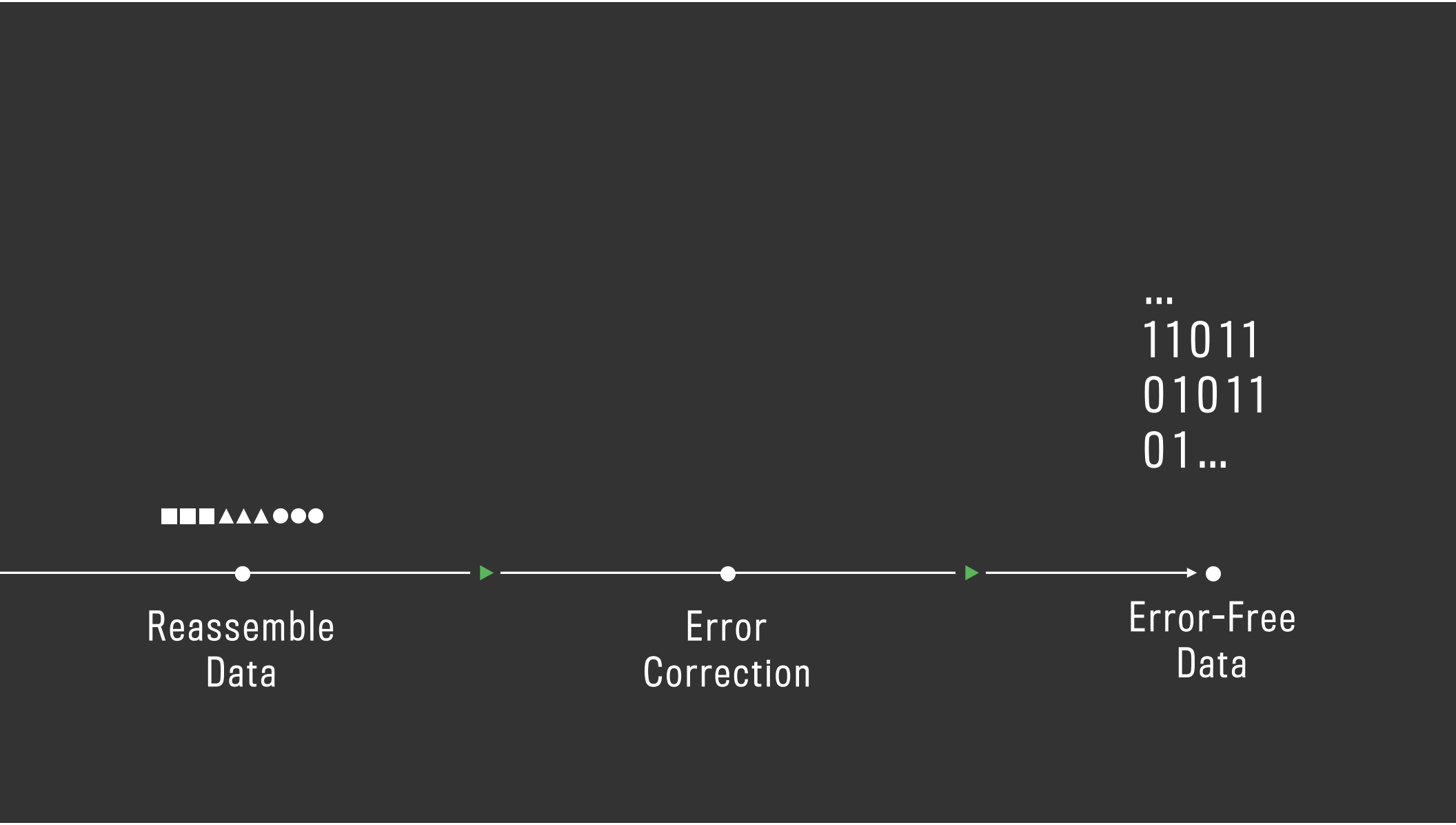
Strands with Desired Primer

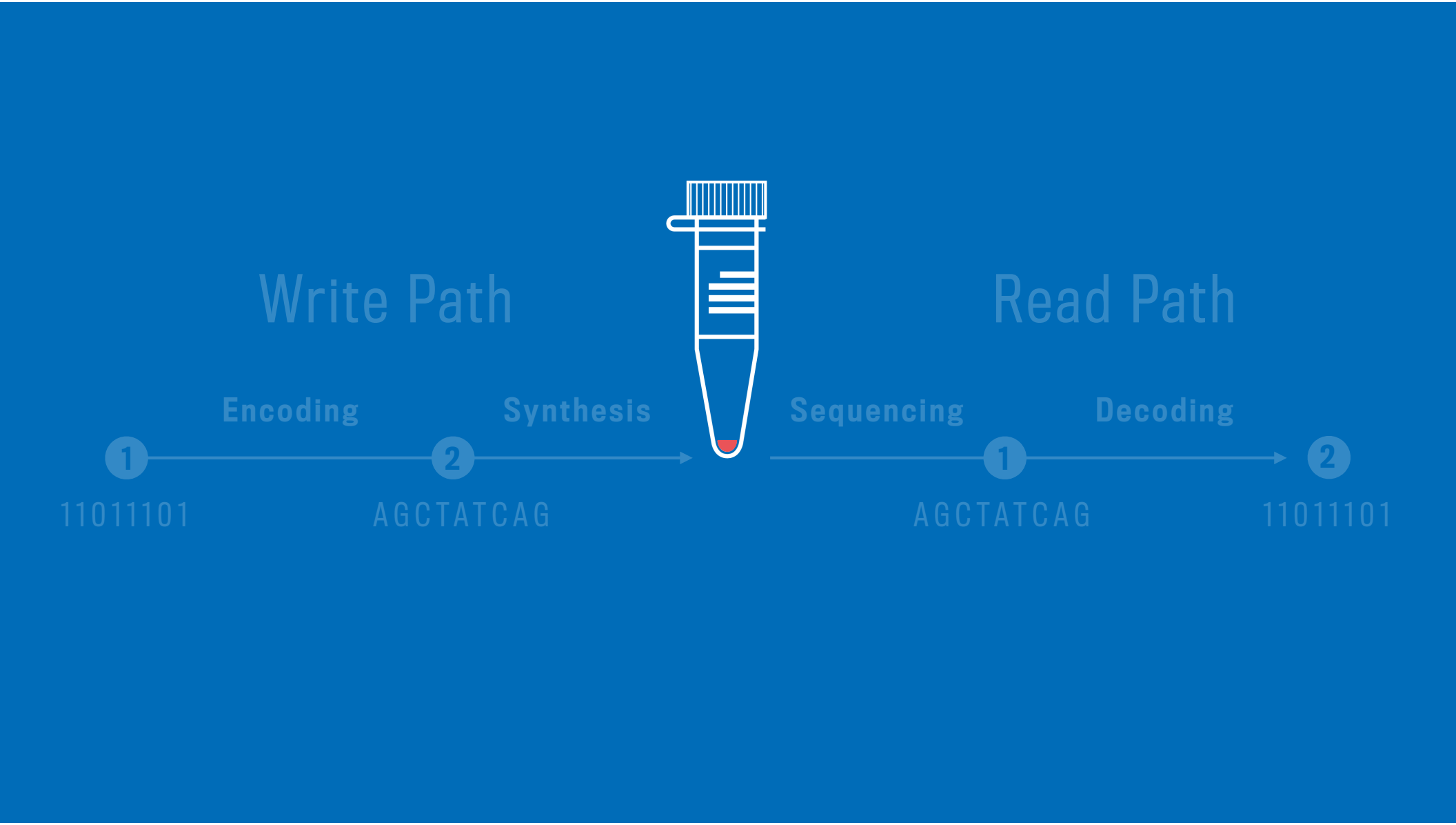
ATGTT TGCT TACC AAAC CATGC
ATGTT GCCA GTTC ACAG CATCC
ATGTT GGAT GCAC AAGA CATCC
ATGTT TGCT TACC CAAC CATCC
ATGTT GCCA GTTC AAAG CATGC
ATGTT GGAT GCAC AAGA CATCC
ATGTT TGCT TACC CAAC CATCC
ATGTT GCCA GTTC AAAG CATCC

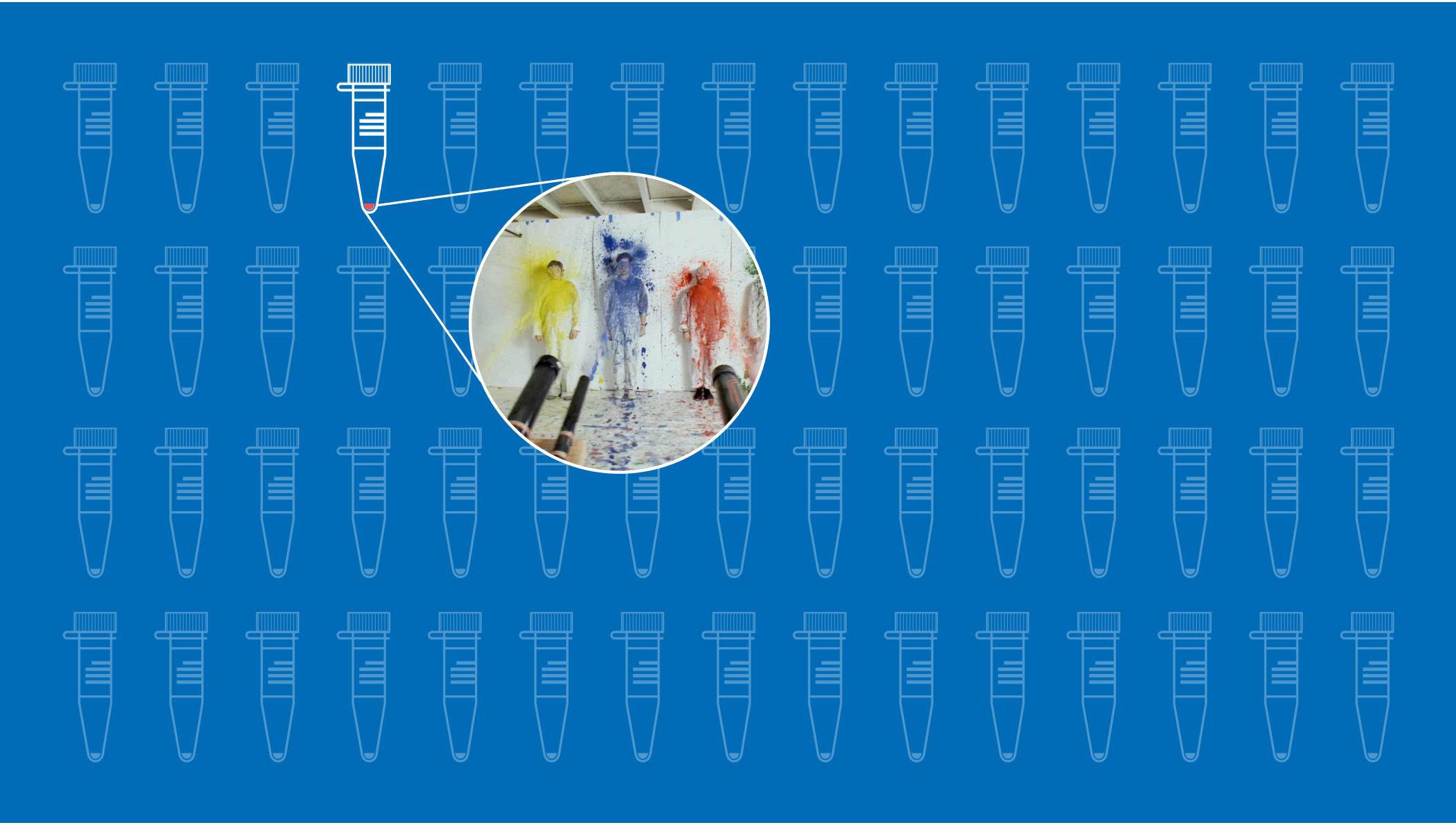
Sequencing

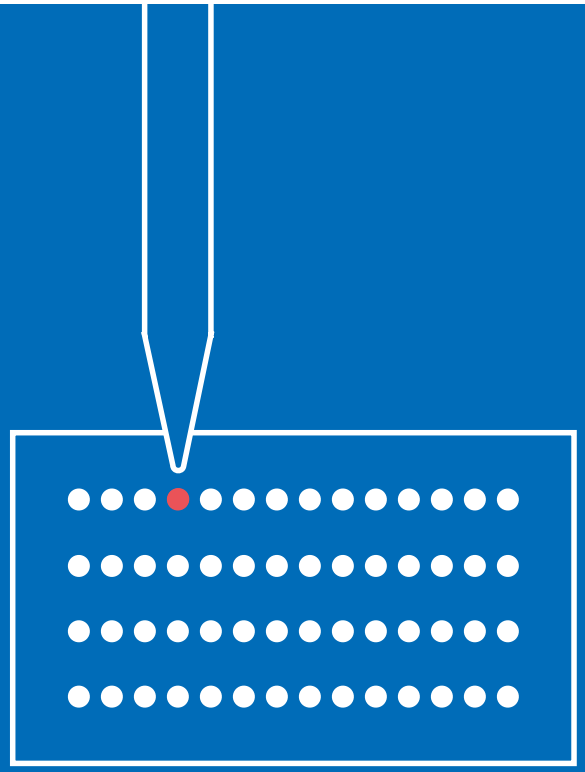


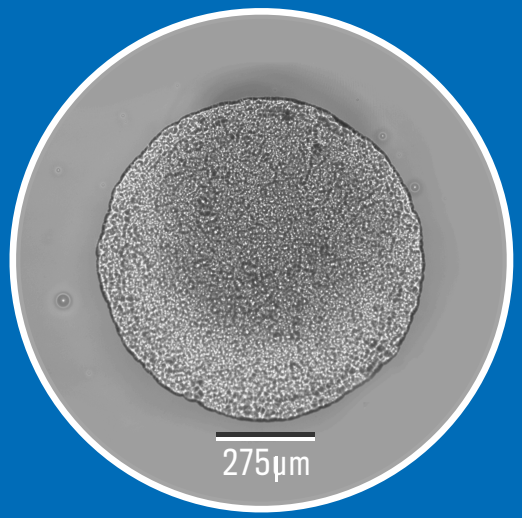
Clustering


...
11011
01011
01...
Reassemble Data
Error Correction
Error-Free Data

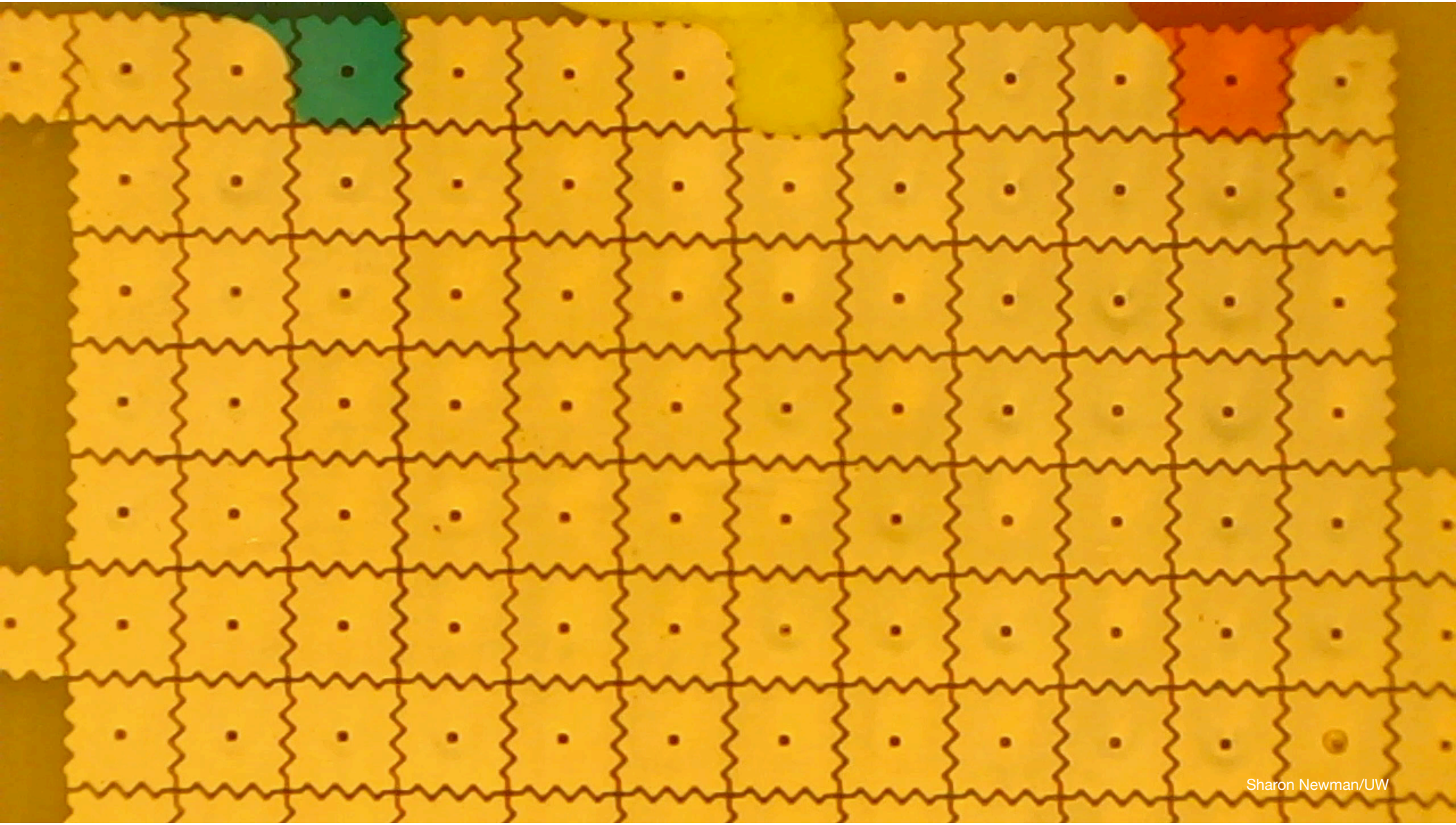

Write Path
Read Path
Encoding
Synthesis
Sequencing
Decoding
1
2
1
2
11011101
AGCTATCAG
AGCTATCAG
11011101



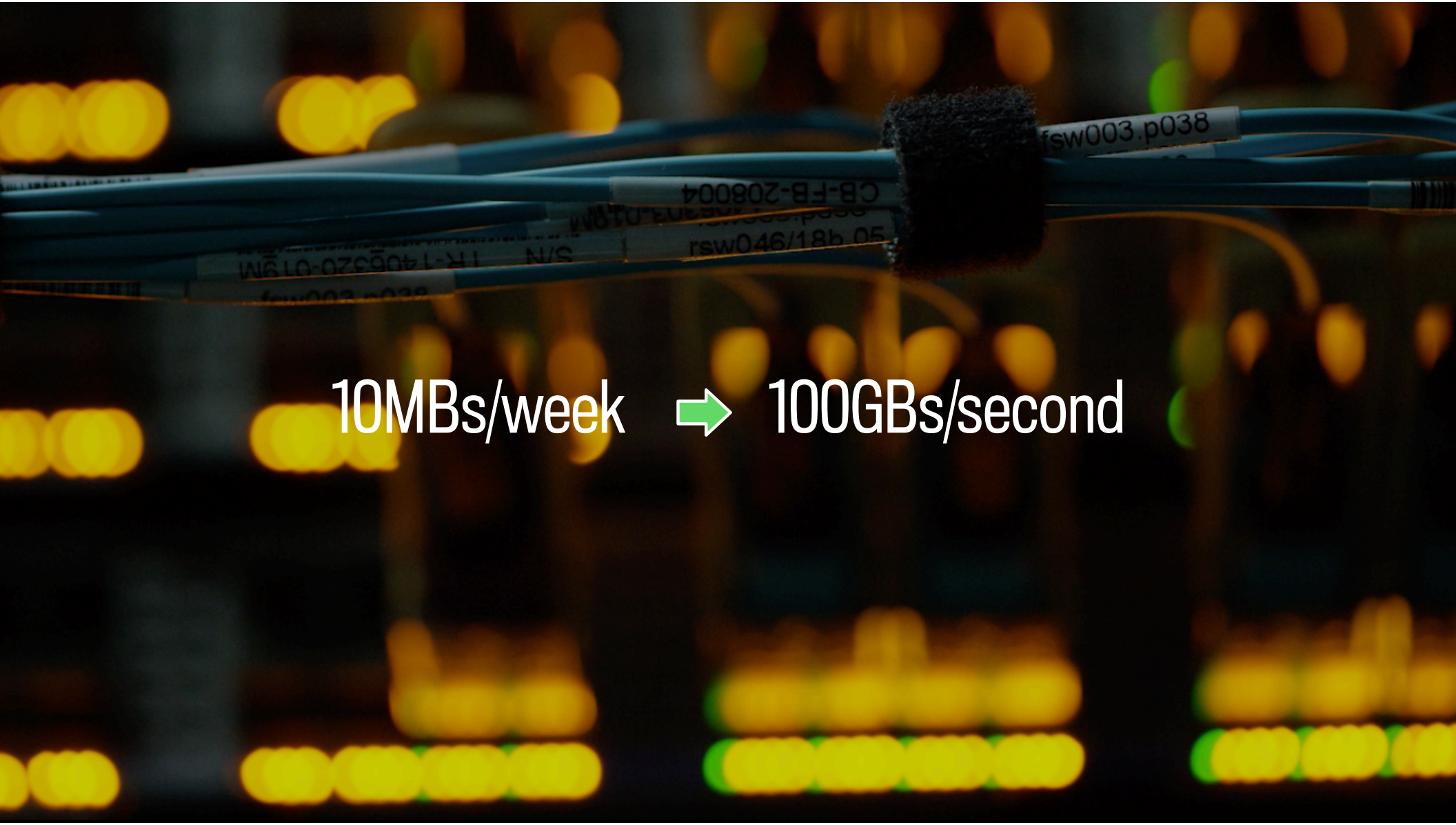


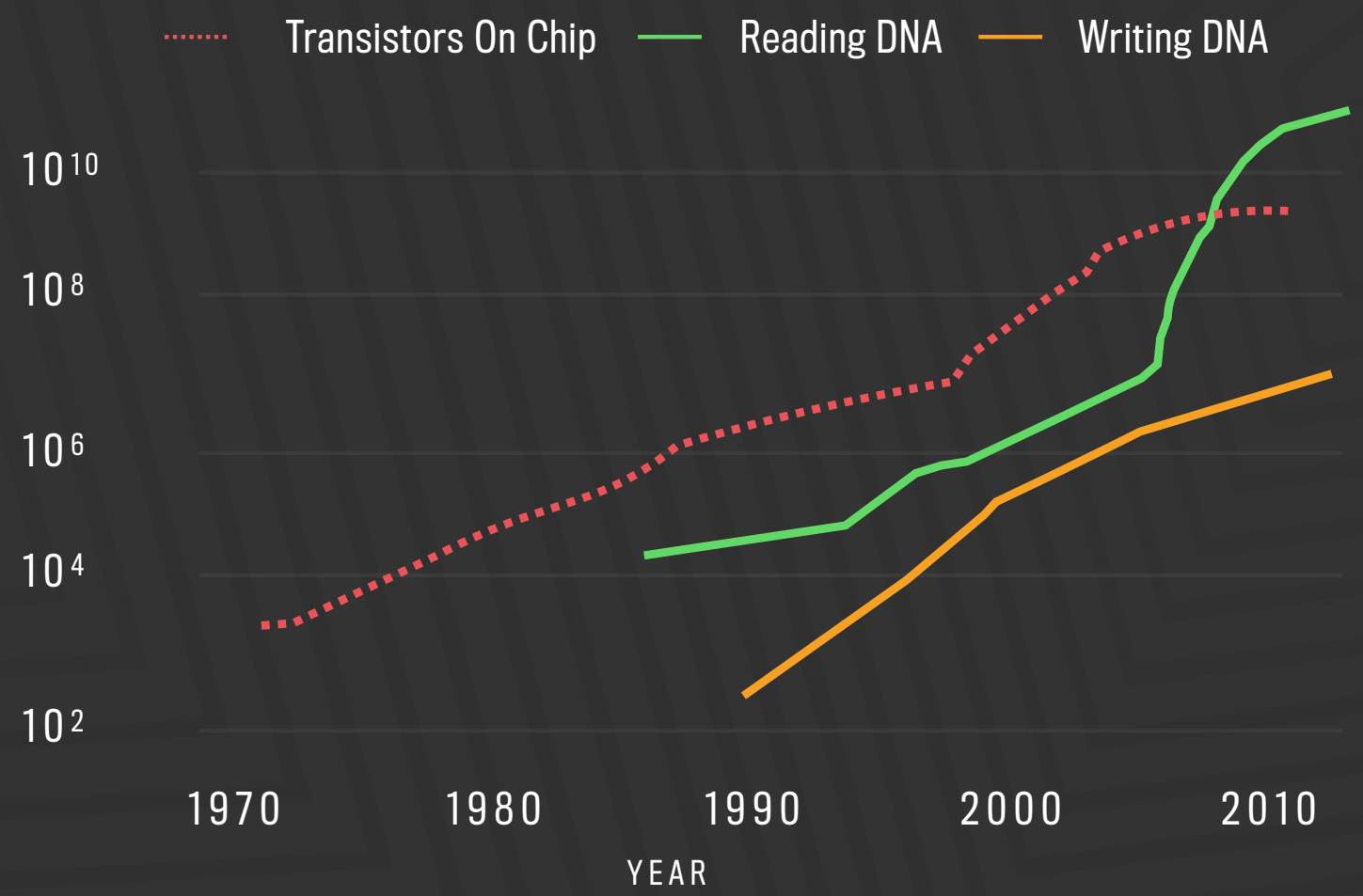


~100TB per spot

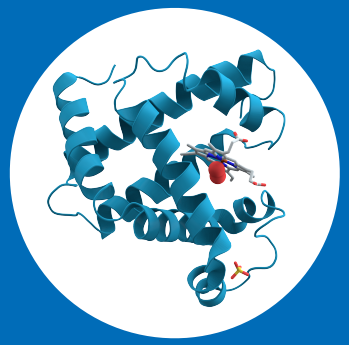

Sharon Newman/UW

CROP TRUST
SECURING OUR FOOD, FOREVER
UNIVERSAL DECLARATION OF HUMAN RIGHTS
Project Gutenberg
Over 400MB, 5B+ Nucleotides,
Demonstrated Random Access.


fsw003.p038
CB-FB-208004
rsw046/18b.05
S/N
10MBs/week
100GBs/second



Source: Robert Carlson



+

MISL

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Microsoft