STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

Approximate DNA Storage with High Robustness and Density for Images

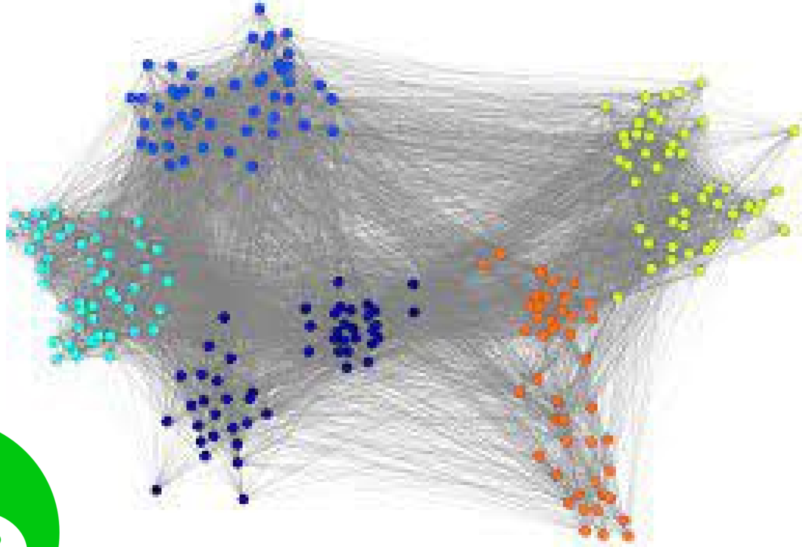
Presented by

Bingzhe Li

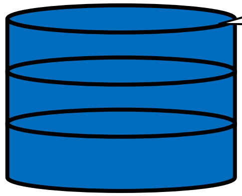
Assistant Professor

University of Texas at Dallas

Big Data



Big Data Era



Data is **doubled** almost every **2 years**
44 Zettabytes in 2020
175 Zettabytes in 2025

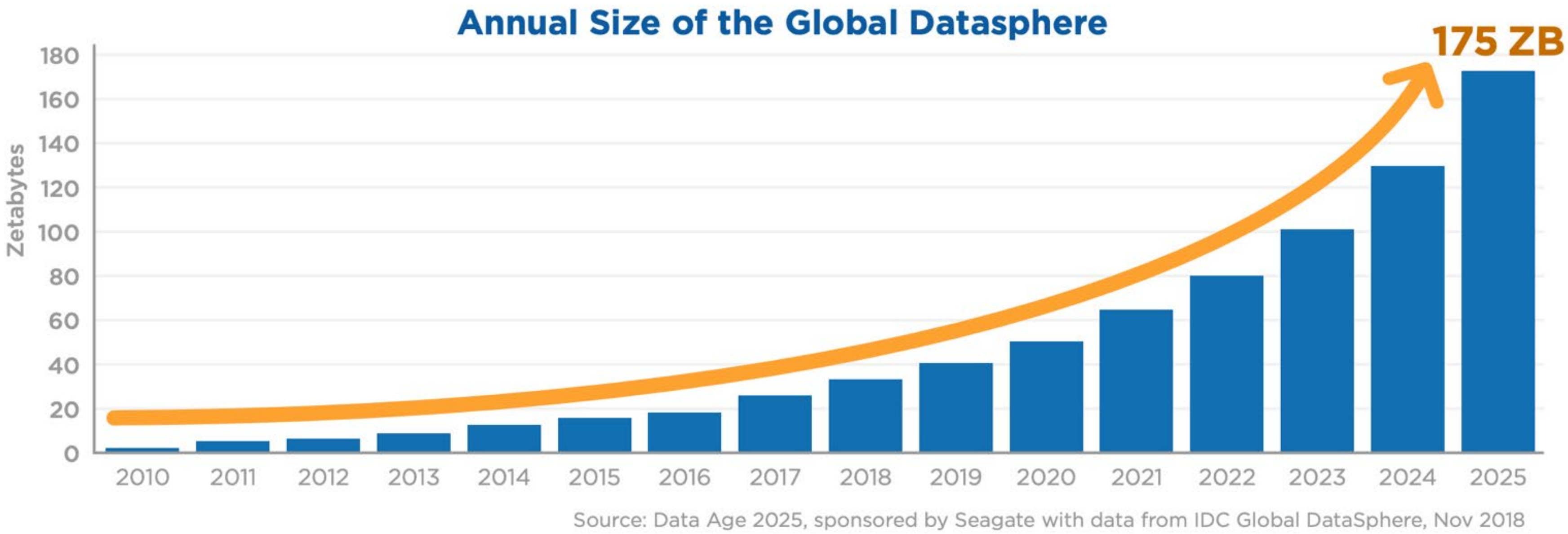
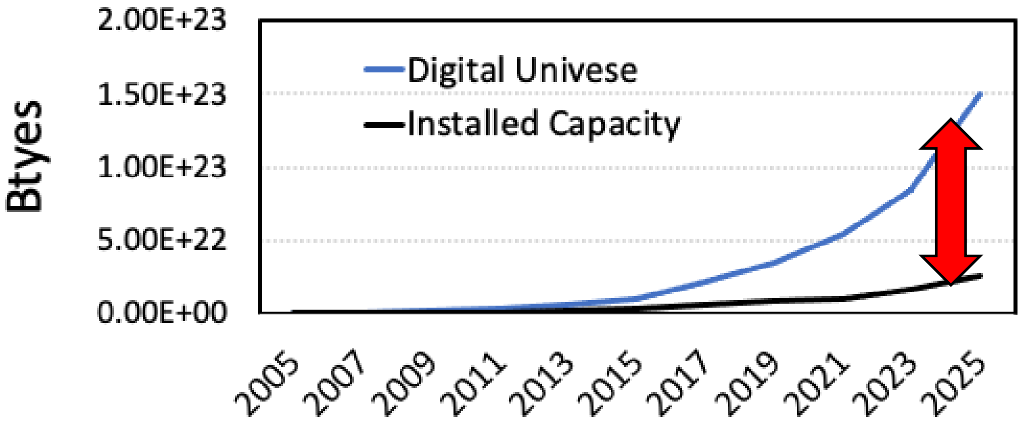


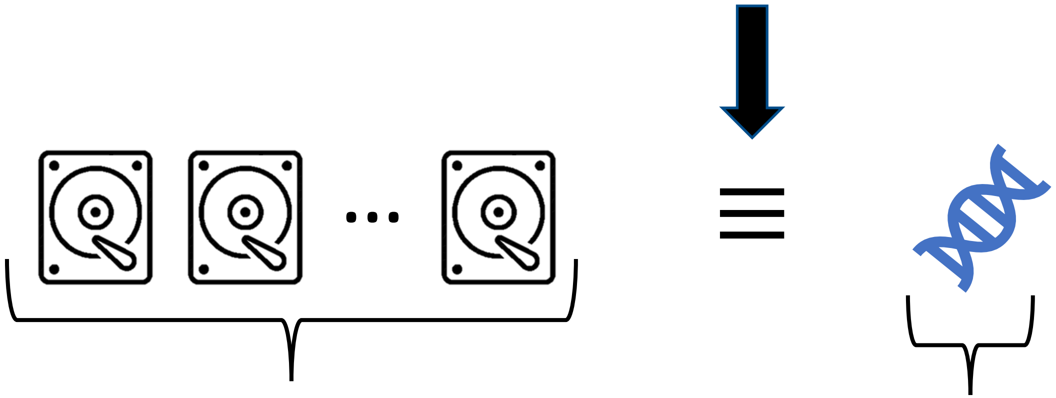
Image from: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Why DNA Storage?



Large gap between generated data and installed storage capacity.

200 PB data



- 25,000 x 8TB HDDs
- 5 – 10 years of warranty

- 1 gram DNA [1]
- Several centuries [2]

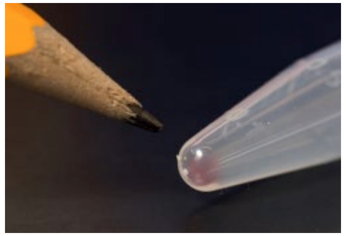
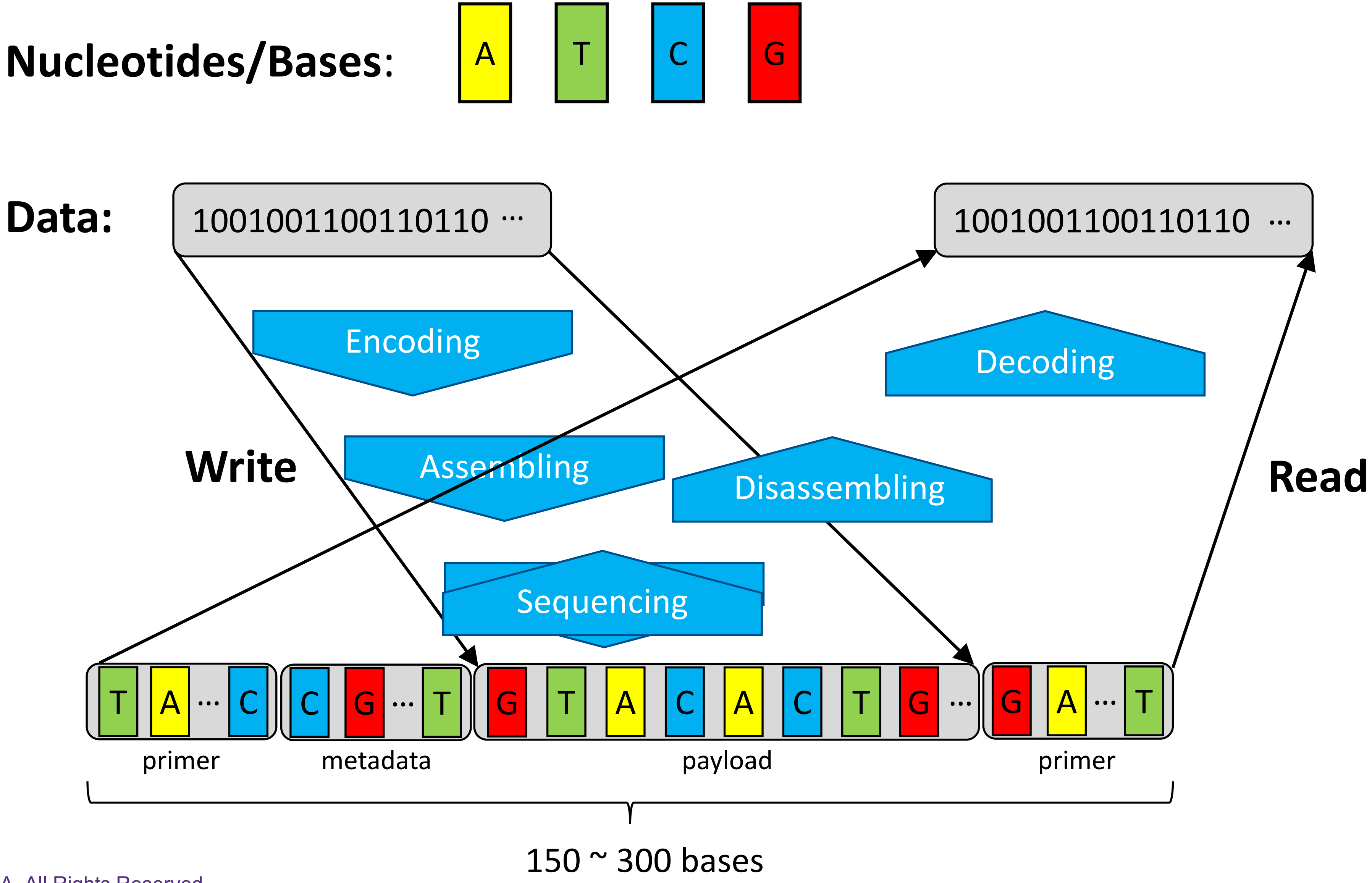


Photo: Tara Brown / UW

[1] *Allentoft et al. The half-life of dna in bone: measuring decay kinetics in 158 dated fossils. Proceedings of the Royal Society B: Biological Sciences, 279(1748):4724–4733, 2012.*
[2] *Grass et al. Robust chemical preservation of digital information on dna in silica with error-correcting codes. Angewandte Chemie International Edition, 54(8):2552–2555, 2015.*
[3] Figure source: IDC

SDC 23

What is DNA Storage?



Simple encoding:

Bit	Base
00	A
01	T
10	G
11	C

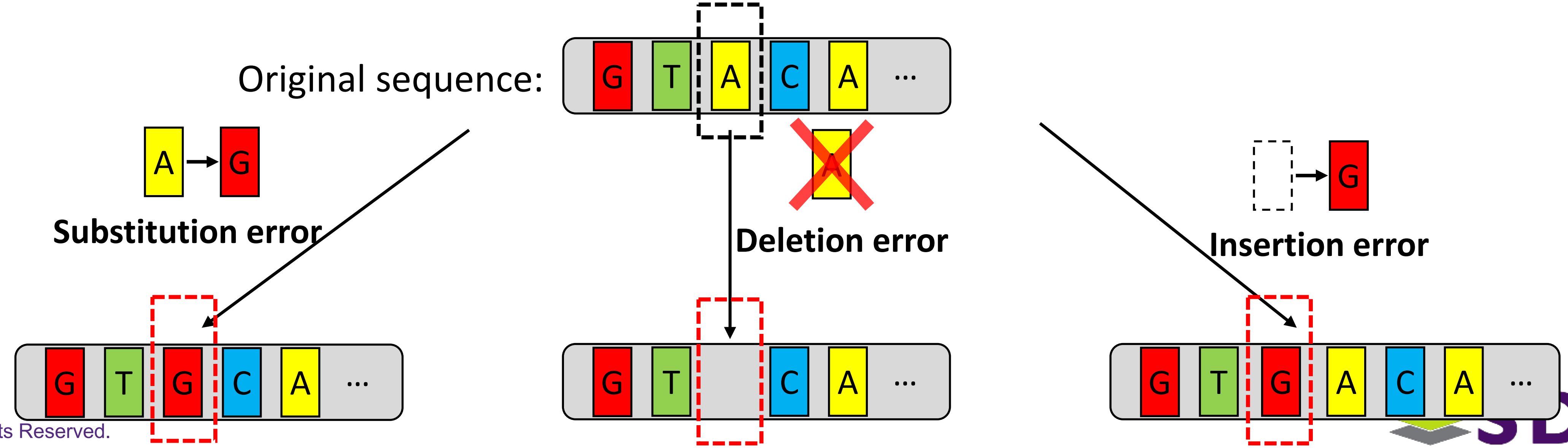
Issues of DNA Storage

DNA storage is

- Error-prone
- Expensive (e.g., $1million/GB)
- Slow (e.g., hours/GB)
- Special preservation
- Low encoding density (ideal one is 2bits/nt)
    - 00->A, 01->T, 10->C, 11->G
- … …

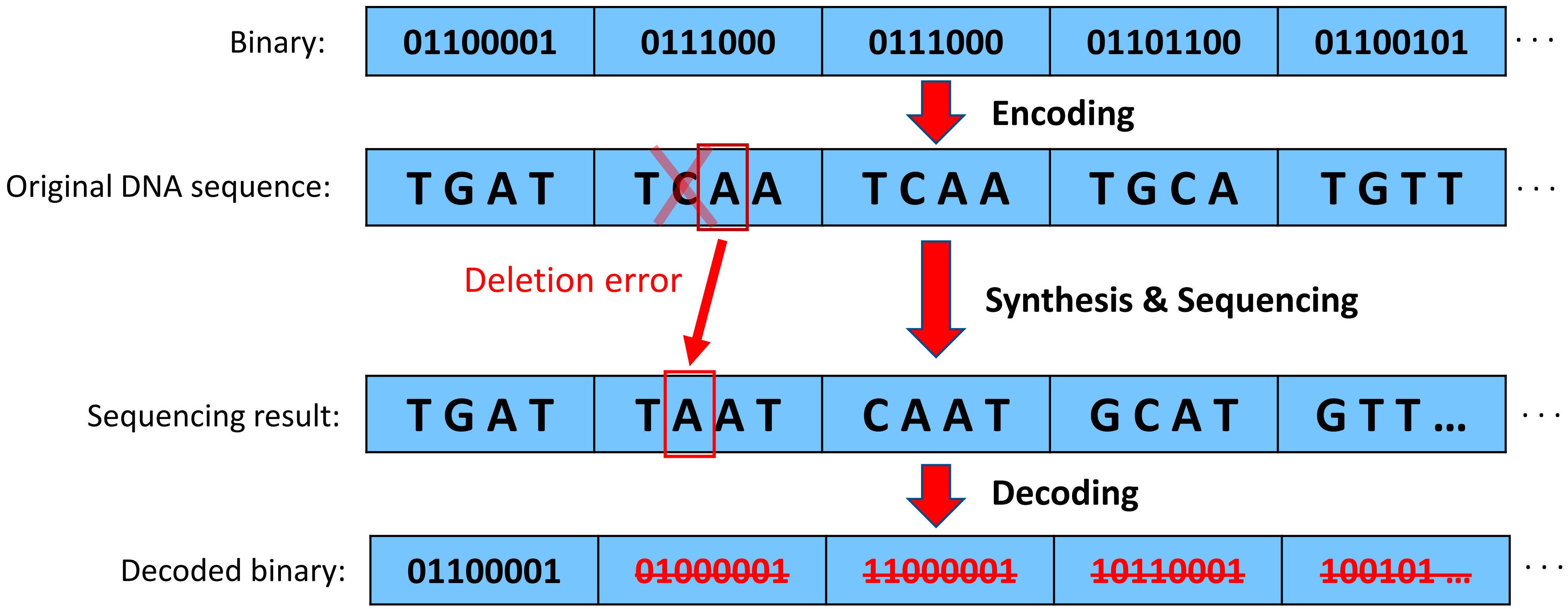
Errors of DNA storage:

- **Some patterns may increase error rates :**
    - Consecutive identical nucleotides (e.g., “AAAA”)
    - Hairpin structure/secondary structure
    - etc.



Error Propagation in DNA Storage

Error propagation:

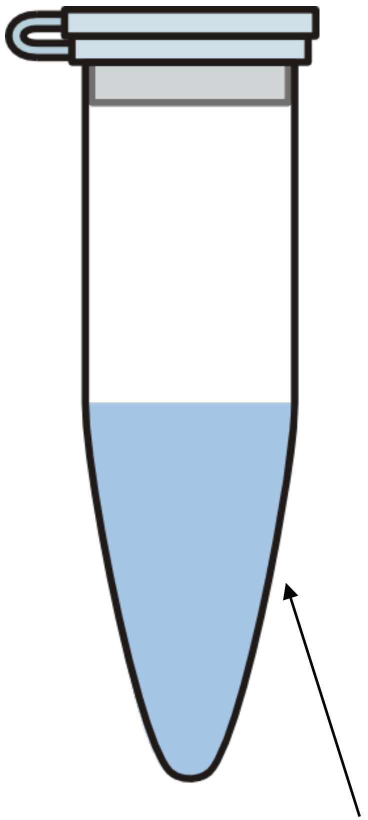


Bit	Base
00	A
01	T
10	G
11	C

Conclusion:

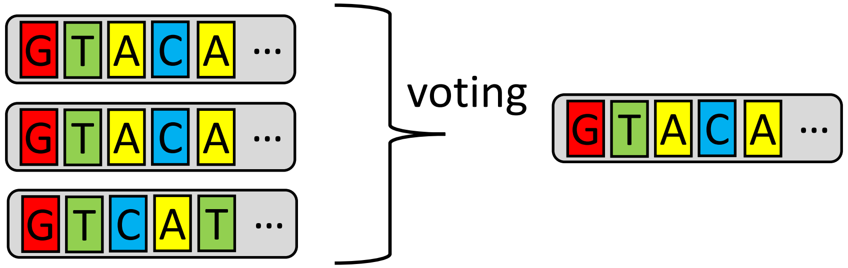
- One nucleotide error causes a series of errors in its subsequence

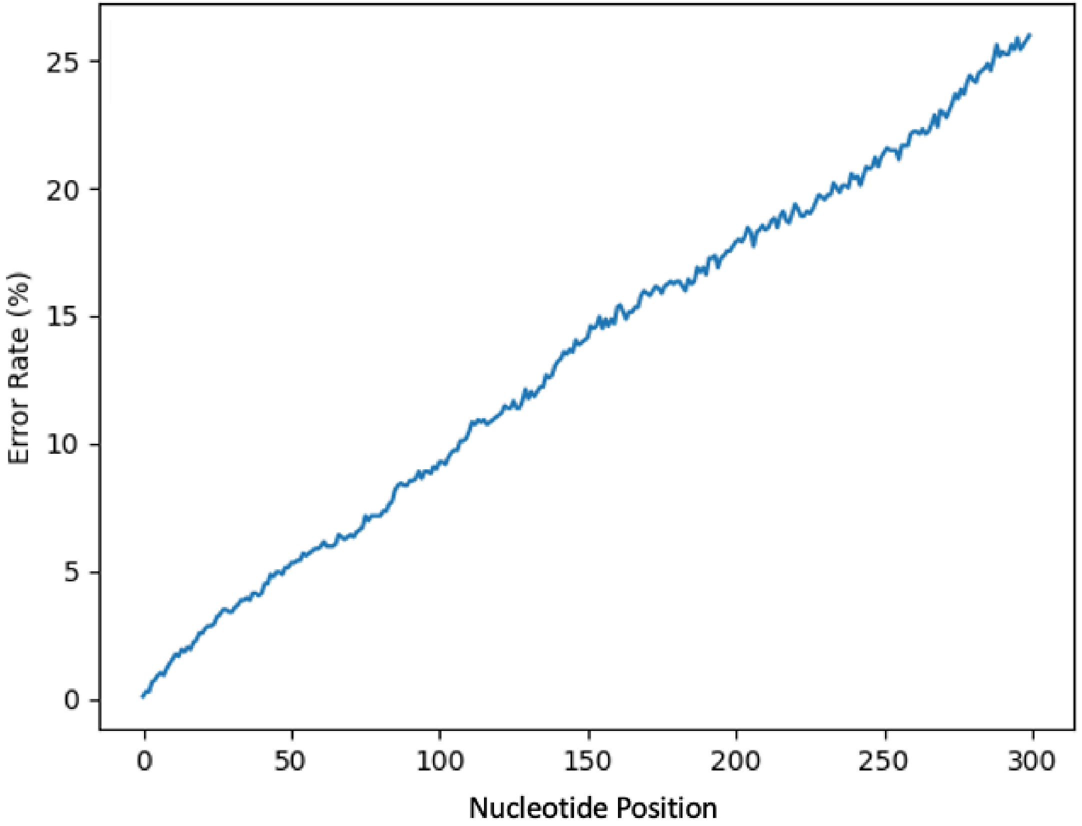
Error Propagation (EP) in DNA Storage cont.



Millions of DNA strands

EP in sequencing:





Conclusion: error propagation in DNA sequence

- One nucleotide error causes a series of errors in its subsequence

Lin, Dehui, Yasamin Tabatabaee, Yash Pote, and Djordje Jevdjic. "Managing reliability skew in DNA storage." In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pp. 482-494. 2022.

Issues of DNA Storage

DNA storage is

- Error-prone
- Expensive (e.g., $1million/GB)
- Slow (e.g., hours/GB)
- Special preservation
- Low encoding density (ideal one is 2bits/nt)
  - 00->A, 01->T, 10->C, 11->G
- … …

IMG-DNA [Systor'21]

DP-DNA [MASCOTS'23]

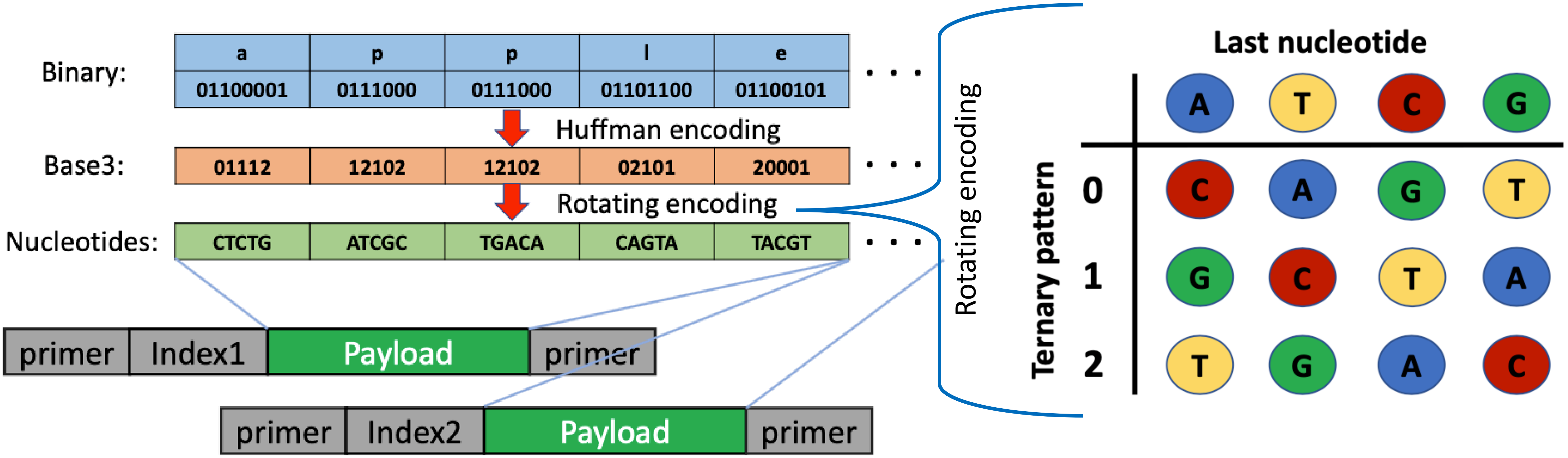
HL-DNA [ICCD'22]

Increase Density of DNA Storage

DP-DNA: A Digital Pattern-Aware DNA Encoding Scheme to Improve Encoding Density of DNA Storage [1]

[1] Bingzhe Li, Li Ou, Bo Yuan, and David Du, "DP-DNA: A Digital Pattern-Aware DNA Encoding Scheme to Improve Encoding Density of DNA Storage", The 31st International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (2023).

A typical encoding scheme – rotation code



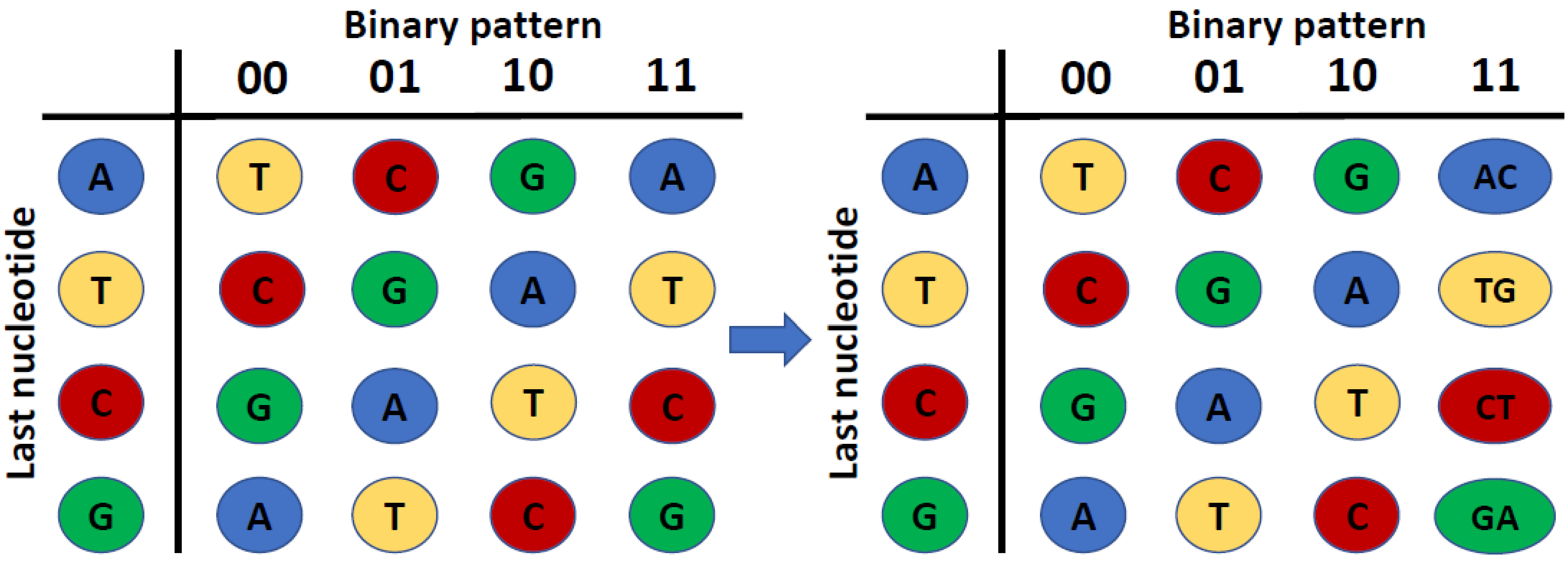
- Avoid long homopolymer
- GC content is roughly maintained

[1] JamesBornholt,RandolphLopez,DouglasMCarmean,LuisCeze,GeorgSeelig, and Karin Strauss. A dna-based archival storage system. In Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, pages 637–649, 2016.

Issues of previous work

- Low encoding density
  - Mapping 8 bits to 5 or 6 trits (base3) ~ 1.57bits/nt
  - Theoretically, encoding density is 2bits/nt, or 1.98bits/nt

Encoding scheme – 2bit-code and unbalance code



(a) 2bit-code

(b) 11-code (unbalanced code)

Issue: how about '111111' for 2bit-code?

- **Long homopolymers**

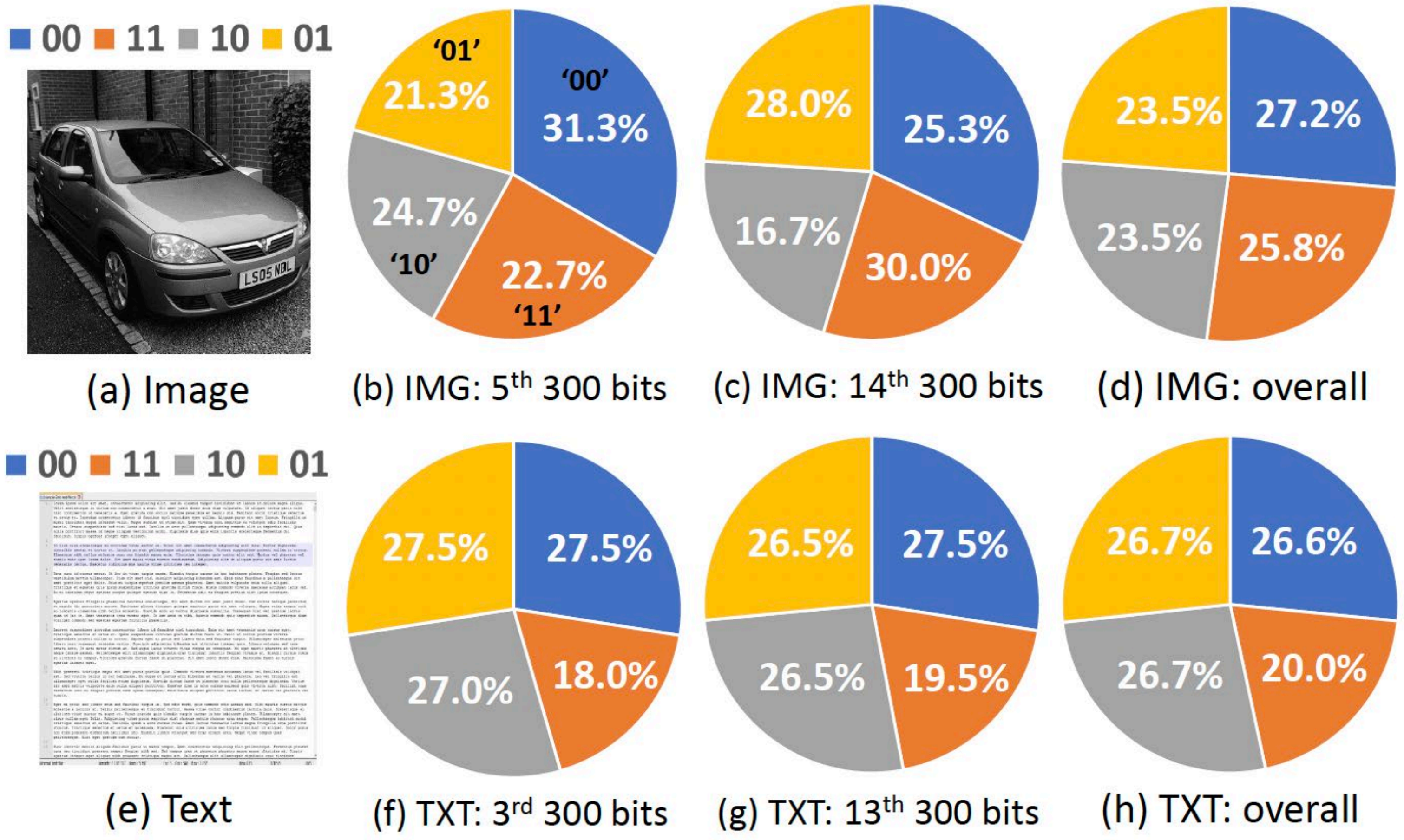
Issue of 11-code

- On average, encoding density is 1.6 bits/nt
- But, an extreme case
  - A sequence of 1111,1111 with an ‘A’ at the beginning
  - Then, DNA sequence will be:
  - **A** – ACAC,ACAC
  - Encoding density is 1bits/nt

Last nucleotide \ Binary pattern	00	01	10	11
A	T	C	G	AC
T	C	G	A	TG
C	G	A	T	CT
G	A	T	C	GA

(b) 11-code (unbalanced code)

Observation: to solve the issue

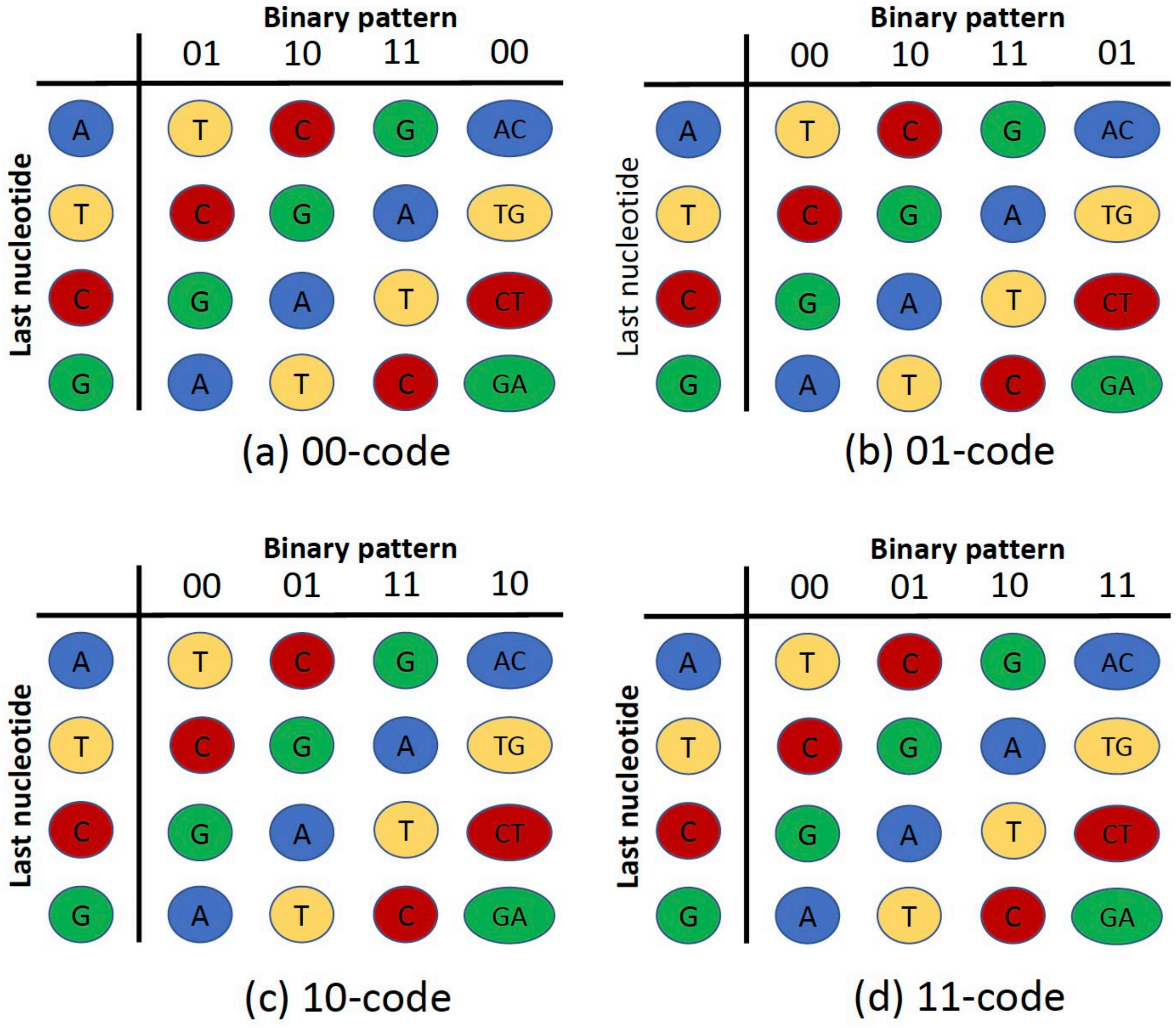


- Four patterns (i.e., 00, 01, 10, and 11) have different distributions among sequences
- 1nt/bit is used for the pattern with the lowest percentage.
- Lower bound case will be 25% for all patterns

Digital Pattern aware code (DP-DNA)

- Find the lowest-frequency pattern
- Use the corresponding code

- For example, '11' has the lowest frequency in a binary sequence
- Then, use 11-code



(a) 00-code (b) 01-code (c) 10-code (d) 11-code

- Worst case:
  - All patterns evenly show in a sequence
  - Encoding density is 1.60 bits/nt > 1.57bit/nt

Adding 2bit-code and Using Variable Length

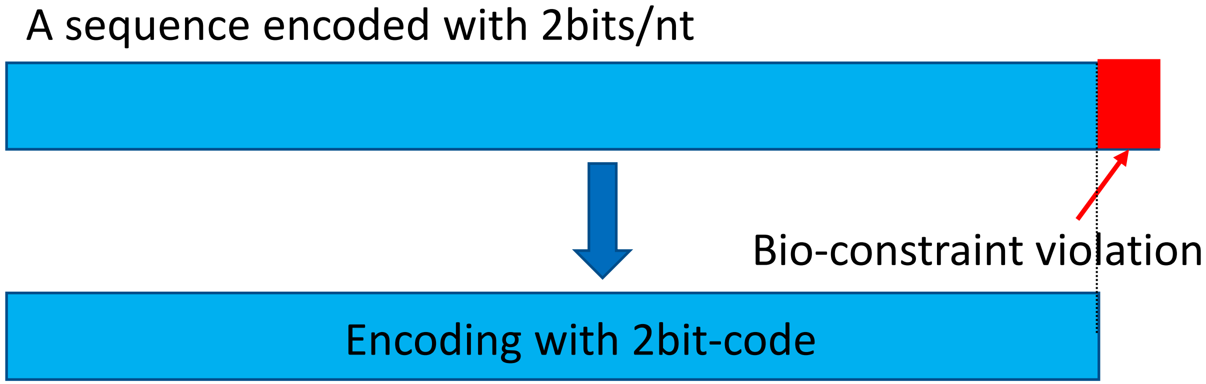
Adding 2bit-code:

- Ideal encoding density (2bits/nt)
- If some sequences encoded with 2bits-code have **no bio-constraint violations**, we can encode those sequences with 2bit-code

Encoding density ↑

Variable Length

- Ideal encoding density (2bits/nt)

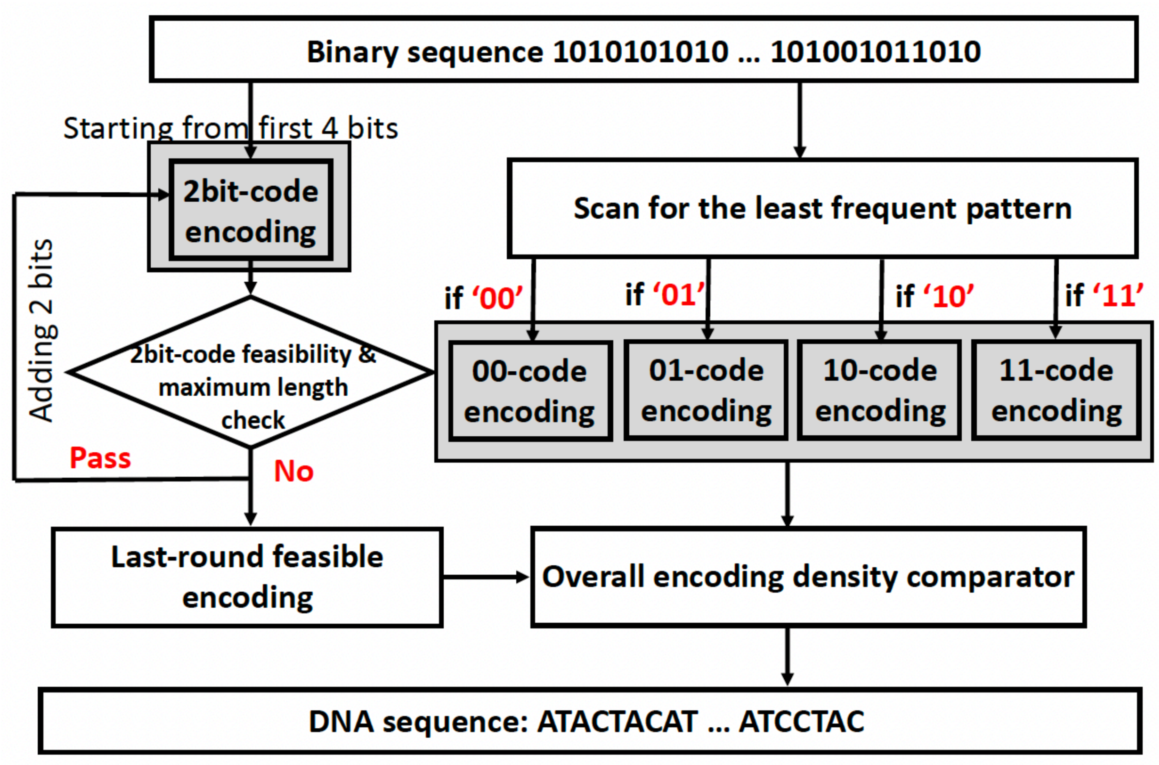


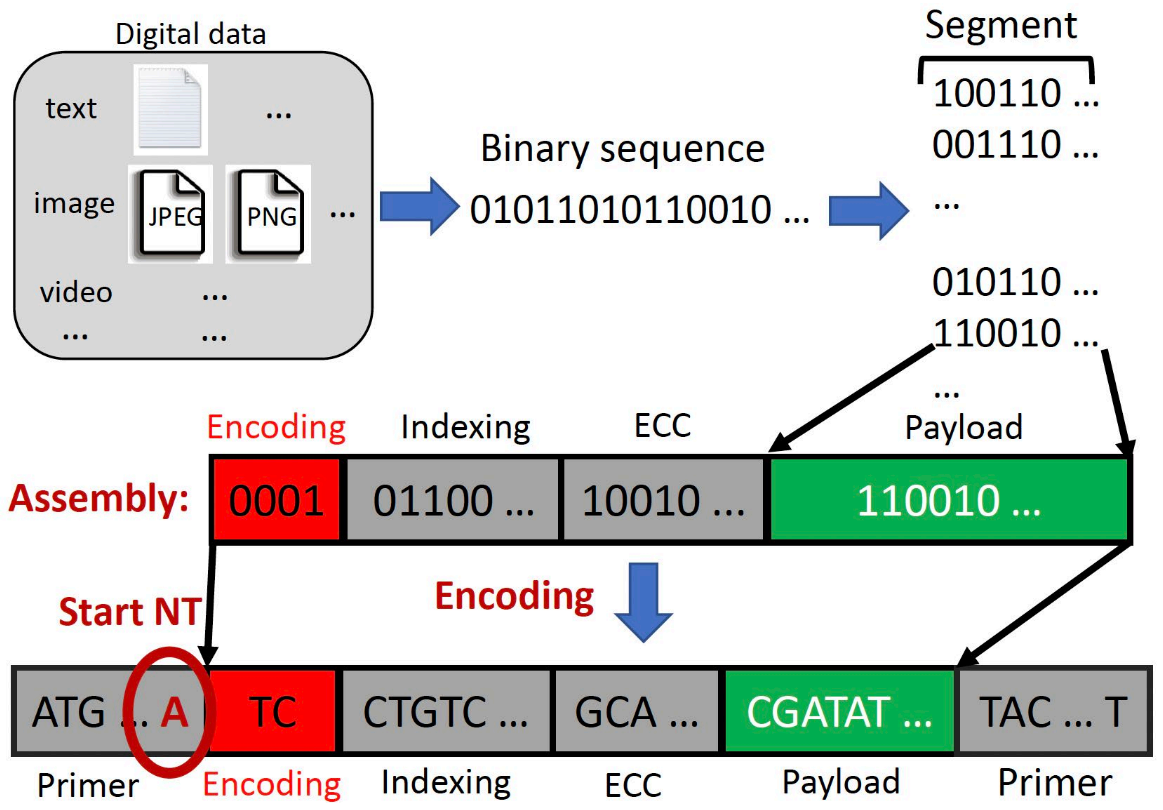
$$\frac{L}{L/\varepsilon_1 + L_{meta}} < \frac{L - M}{(L - M)/\varepsilon_2 + L_{meta}}$$

where ε_1 and ε_2 indicate the code densities of the low-density and high-density codes, respectively. L is the default length of the binary sequence to be encoded. M indicates how many bits are excluded for the high-density code. L_{meta} refers to the number of nucleotides used for metadata such as primer pairs and internal index in DNA strands.

SDC 23

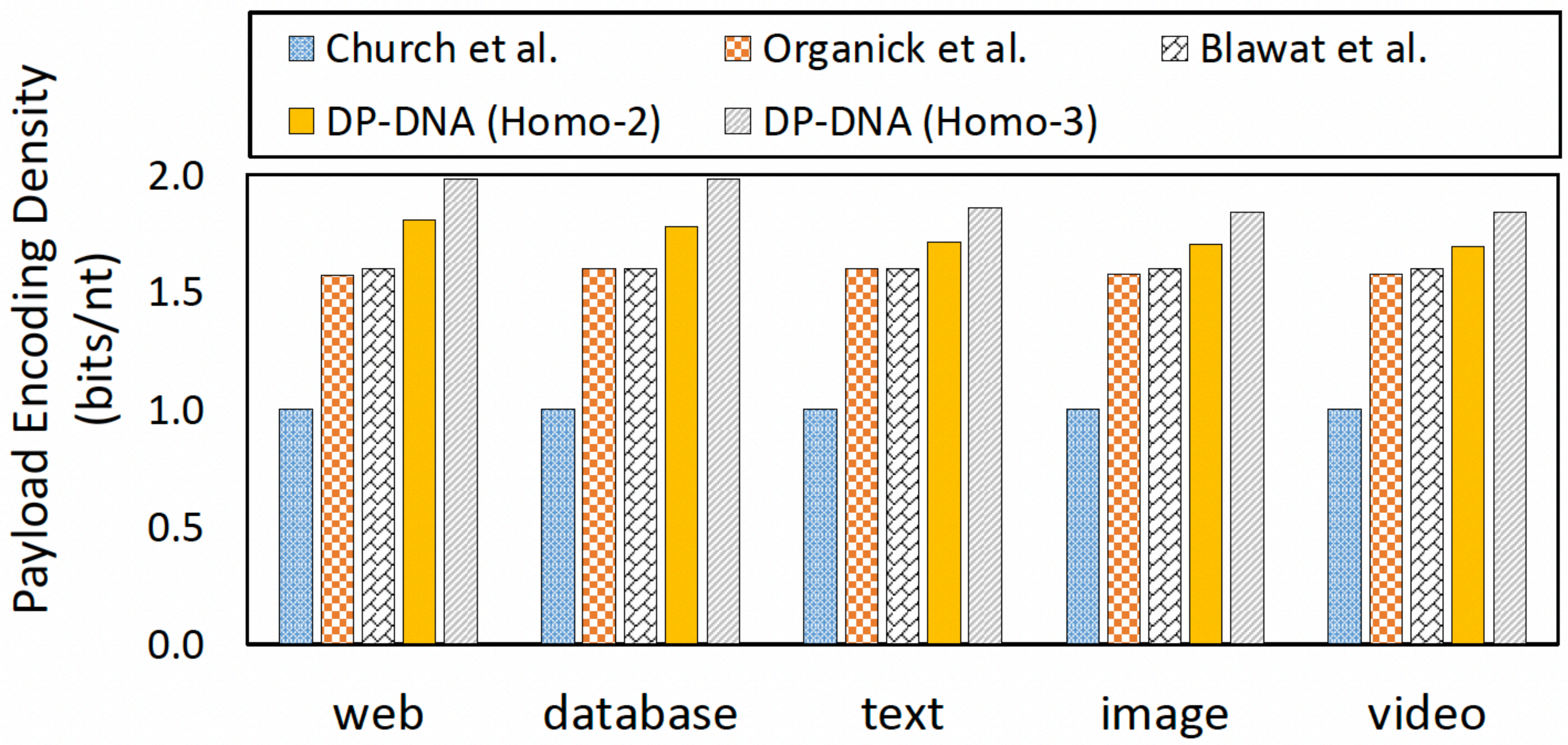
DP-DNA overall design





Experimental results

- Dataset
  - Web
  - Database
  - Text
  - Image
  - Video



Increase Robustness of DNA Storage for Images

IMG-DNA: approximate dna storage for images[1]

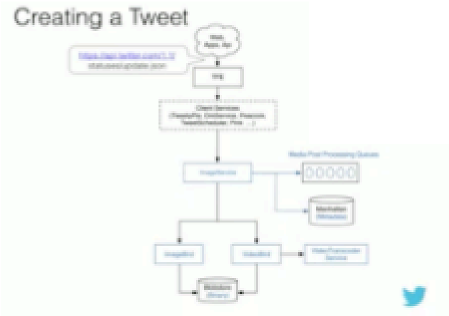
[1] Bingzhe Li, Li Ou, and David Du. "IMG-DNA: approximate dna storage for images." Proceedings of the 14th ACM International Conference on Systems and Storage. 2021.

High Demand for Storing Images

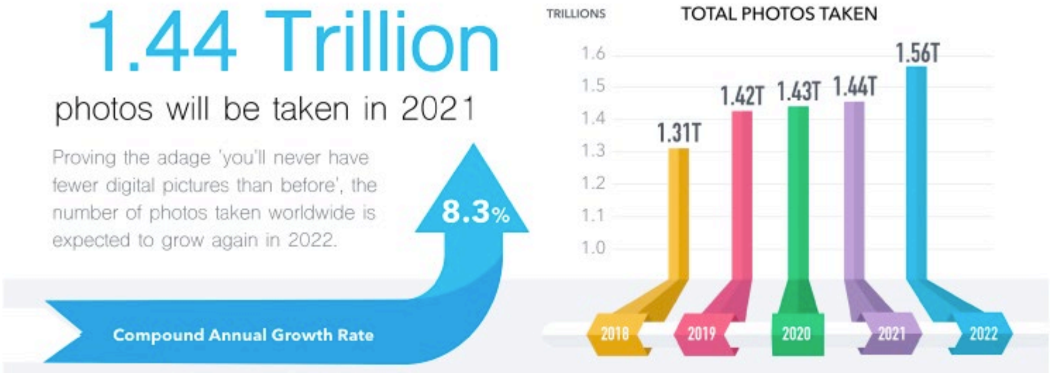
How Twitter Handles 3,000 Images Per Second

WEDNESDAY, APRIL 20, 2016 AT 8:56AM

Today Twitter is creating and persisting 3,000 (200 GB) images per second. Even better, in 2015 Twitter was able to save $6 million due to improved media storage policies.



It was not always so. Twitter in 2012 was primarily text based. A Hogwarts without all the cool moving pictures hanging on the wall. It's now 2016 and Twitter has moved into to a media rich future. Twitter has made the transition through the development of a new *Media Platform* capable of supporting photos with previews, multi-photos, gifs, vines, and inline video.

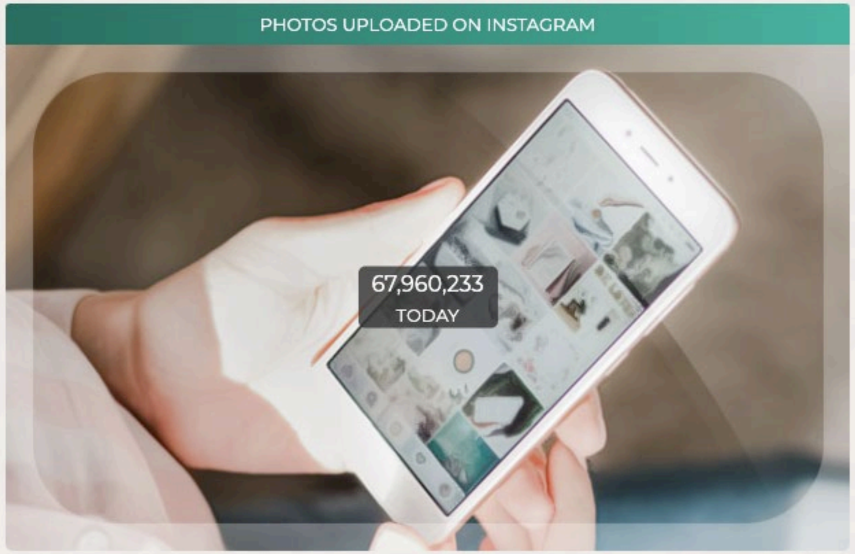


HOME > TECH

Facebook Users Are Uploading 350 Million New Photos Each Day

Cooper Smith Sep 18, 2013, 7:00 AM

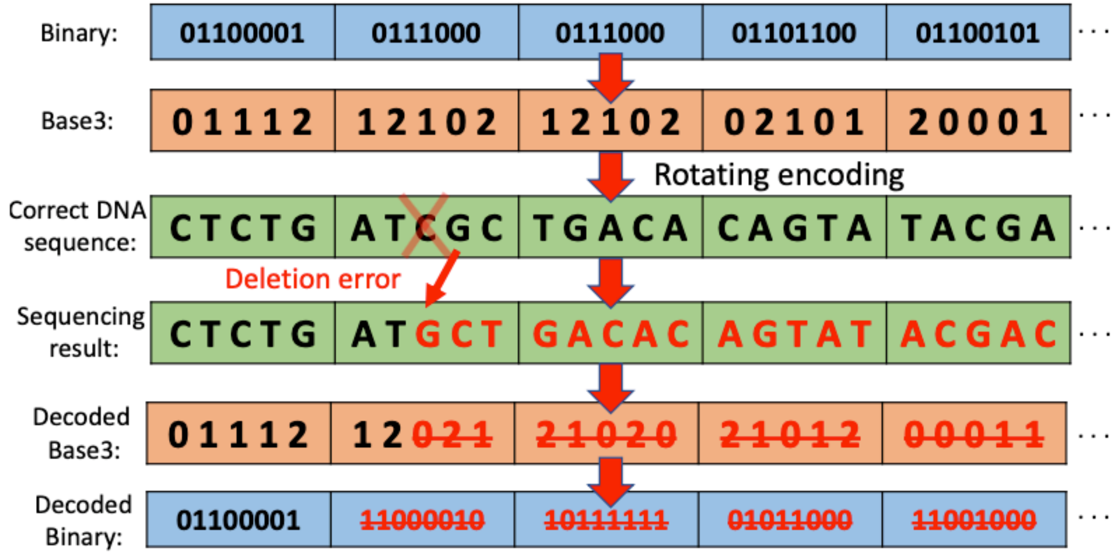
***Social Media Insights** is a daily newsletter from Business Insider that collects and delivers the top social media news first thing every morning. You can sign up to receive Social Media Insights here or at the bottom of this post.*



Instagram has captured a large piece from the social media users and as today, there are 500 million active daily users. There are 995 photos uploaded every second and since the beginning of Instagram and by today, there are more than 50 billion uploaded images that is keep increasing. Instagram was originally created by a group of young people in 2010 but not too long after Instagram has become very popular, Facebook has purchased it for $1 billion and owns it since. The most followed Instagram user is Cristiano Ronaldo with over 203 million followers.

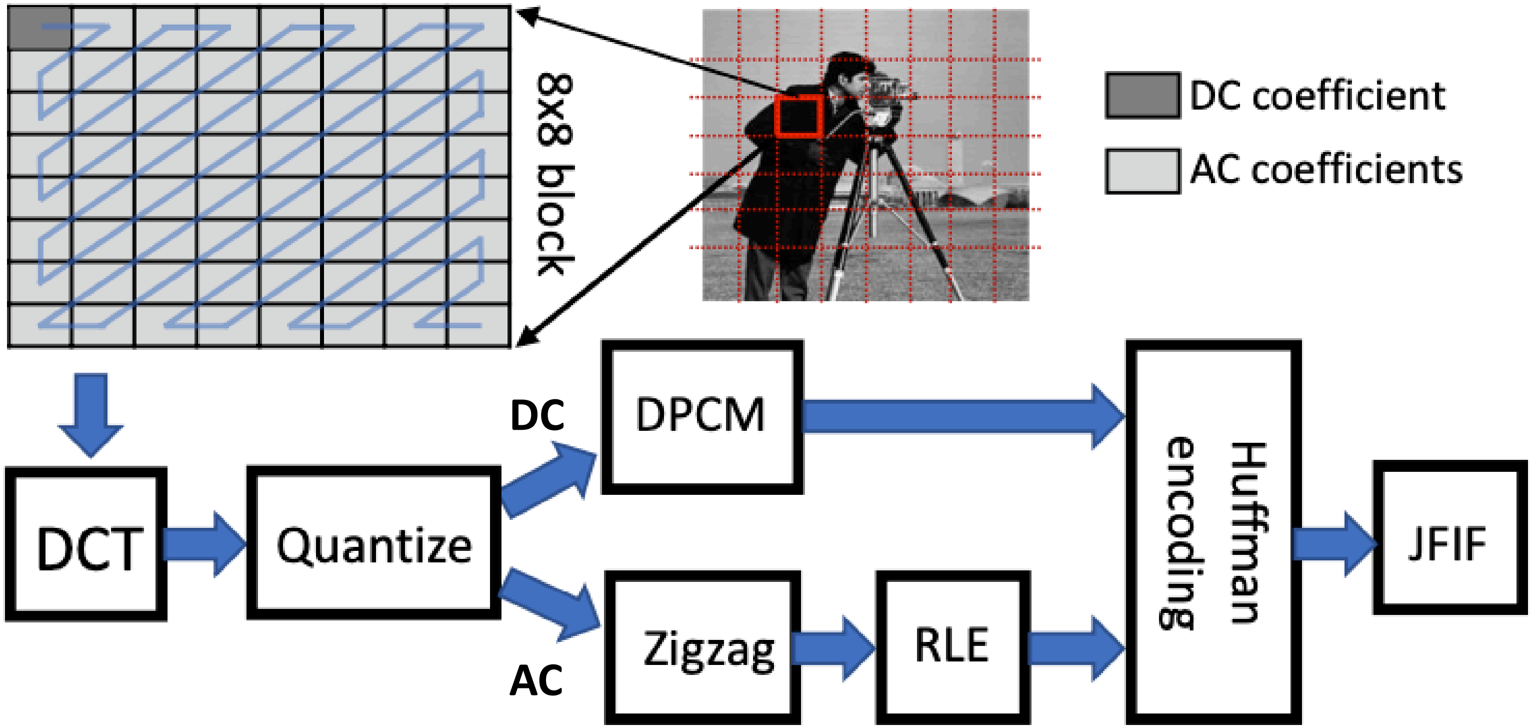
Observations of DNA Storage Encoding

- Small practical tube capacity
  - About 230GB per tube for random-access based DNA storage [1]
- Error prone:
  - Propagation errors [2]: One nucleotide error causes a series of errors in its subsequence



[1] Y. Wei, B. Li, and D. H. Du, "Dna storage: A promising large scale archival storage?" arXiv preprint arXiv:2204.01870, 2022.
[2] B. Li, L. Ou, and D. Du, "Img-dna: approximate dna storage foXr images," in Proceedings of the 14th ACM International Conference on Systems and Storage, 2021, pp. 1–9.

Background of JPEG-based Image



DCT: Discrete Cosine Transform
DPCM: Differential Pulse Code Modulation
JFIF: JPEG File Interchange Format

Two observation [1, 2]:

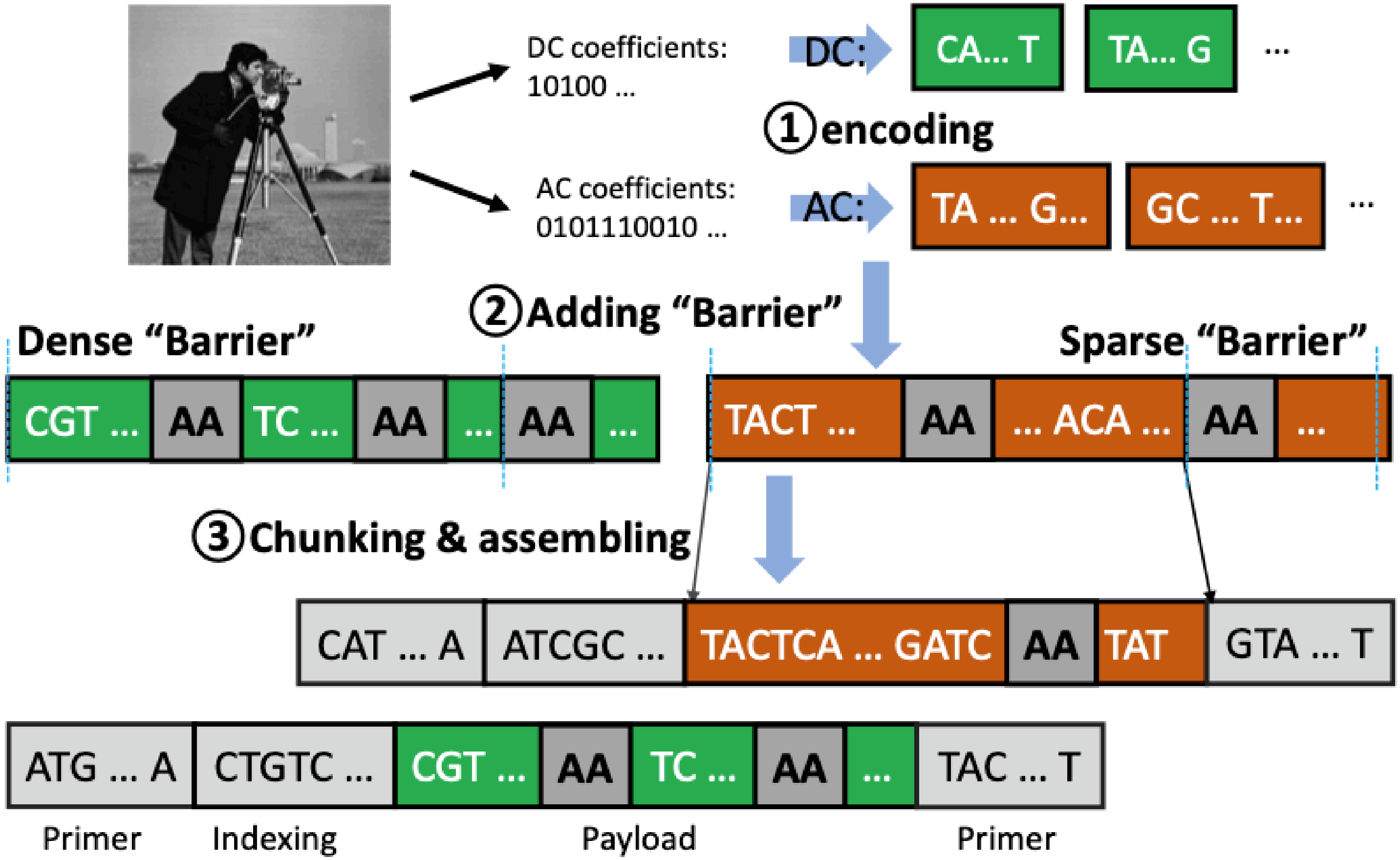
- Fault tolerance
- DC and AC coefficients have different influence on the quality of images

[1] Yu-ChunKuo,Ruei-FongChiu,andRen-ShuoLiu.Long-termjpegdataprotection and recovery for nand flash-based solid-state storage. In 2019 35th Symposium on Mass Storage Systems and Technologies (MSST), pages 141–147. IEEE, 2019.
[2] Qianqian Fan, David J Lilja, and Sachin S Sapatnekar. Adaptive-length coding of image data for low-cost approximate storage. IEEE Transactions on Computers, 69(2):239–252, 2019.

Our Contributions

- Image-based DNA Storage Architecture
- AC/DC Coefficient Separation at DNA Level
- Adding ‘Barriers’
- Asymmetric Barriers for AC/DC Coefficients

Image-based DNA Storage Architecture



1. AC/DC separation
2. Encoding
3. Adding barrier
4. Chunking & assembling

The 14th ACM International Systems and Storage Conference (Systor'21)

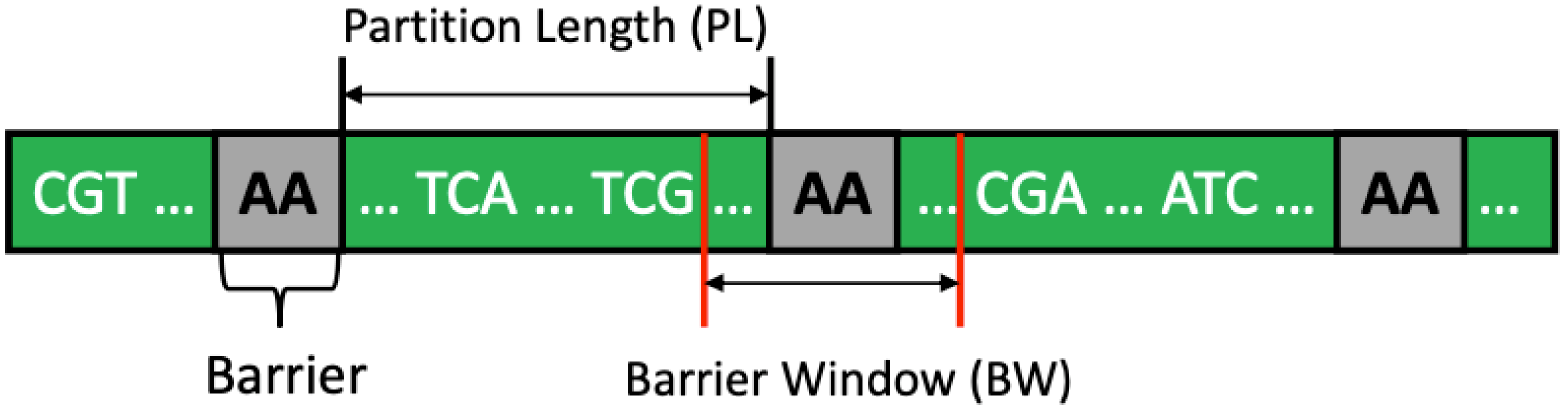
Adding ‘Barriers’ and Asymmetric ‘Barriers’

”AA” as a barrier keeps the error propagation within a partition

- No two consecutive identical “A” in the rotation encoding scheme
- The probability of generating ”AA” caused by errors is low
- Barrier window is used for preventing the errors of insertion and deletion

Asymmetric ‘Barriers’ for AC/DC coefficients

- **Quality:** AC/DC have different influence on the quality of images
- **Overhead:** The number of ACs is much more than that of DC



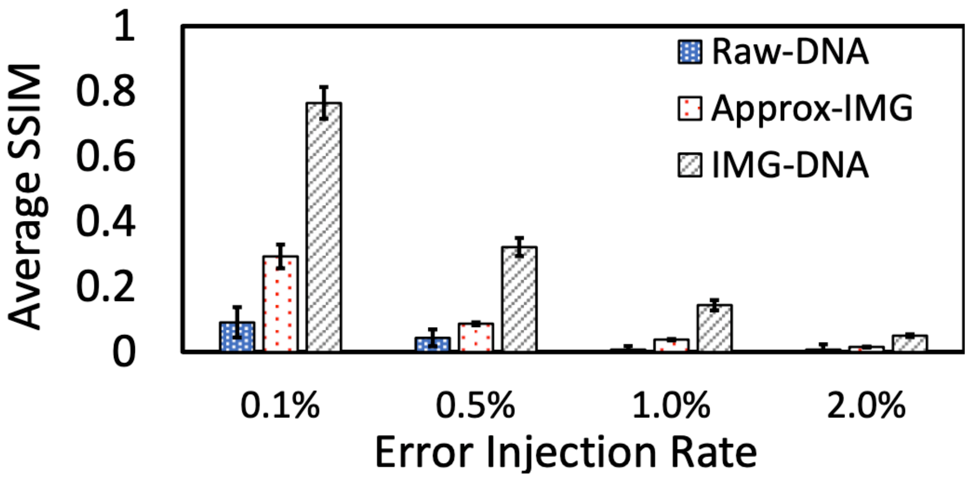
Experimental Results

- Dataset: ImageNet
- Baselines: 1) Raw-DNA; 2) Approx-IMG; 3) IMG-DNA
- Metric: SSIM (structural similarity index metric)
- DNA strand length 250bp

- Environment:
  - A system with Intel i-7-47900 CPU@3.6GHz and 8GB memory
  - MATLAB2020a

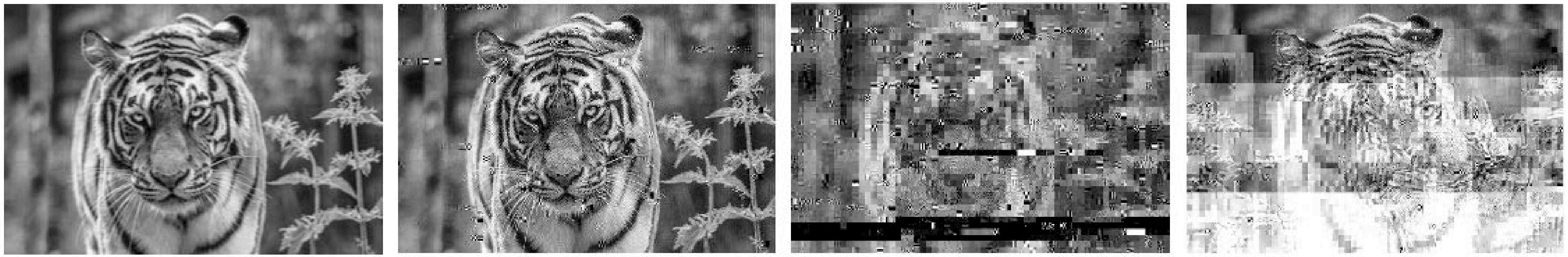
Robustness of Image-based DNA System

Overall comparison:



The SSIM is higher, the quality of images is better

A graphic view of an image with different encoding schemes (0.1% error rate):



(a) Original (b) IMG-DNA (SSIM=0.9078) (c) Approx-DNA (SSIM=0.1604) (d) Raw-DNA (SSIM=0.0561)

More results are shown in the paper

The 14th ACM International Systems and Storage Conference (Systor'21)

Increase Robustness and Density of DNA Storage for Images

HL-DNA: A Hybrid Lossy/Lossless Encoding Scheme to Enhance DNA Storage Density and Robustness for Images[1]

[1] Yi Li, David HC Du, Li Ou, and Bingzhe Li. "HL-DNA: A Hybrid Lossy/Lossless Encoding Scheme to Enhance DNA Storage Density and Robustness for Images." In *2022 IEEE 40th International Conference on Computer Design (ICCD)*, pp. 434-442. IEEE, 2022.

Motivation

- Images are error tolerant
- DNA storage is error-prone

Consider them together

Lossless code design

- DNA strands need to follow some bio-constraints to avoid high errors
- Rotation code helps avoid homopolymers (e.g., ‘AAAA’)

- Lossless code design
  - High density area: 2bits/nt
  - Low density area: 1bits/nt

Ternary pattern	Last nucleotide: A	T	C	G
0	C	A	G	T
1	G	C	T	A
2	T	G	A	C

	Binary pattern (Coding manner) C0	C1	Last nucleotide: A	T	G	C
High Density	00	10	C	A	T	G
High Density	01	11	G	C	A	T
Low Density	10	00	TC	GA	CT	AG
Low Density	11	01	TG	GC	CA	AT

Lossless code design

- DNA strands need to follow some bio-constraints to avoid high errors
- Rotation code helps avoid homopolymers (e.g., ‘AAAA’)

- Lossless code design
  - High density area: 2bits/nt
  - Low density area: 1bits/nt

Last nucleotide

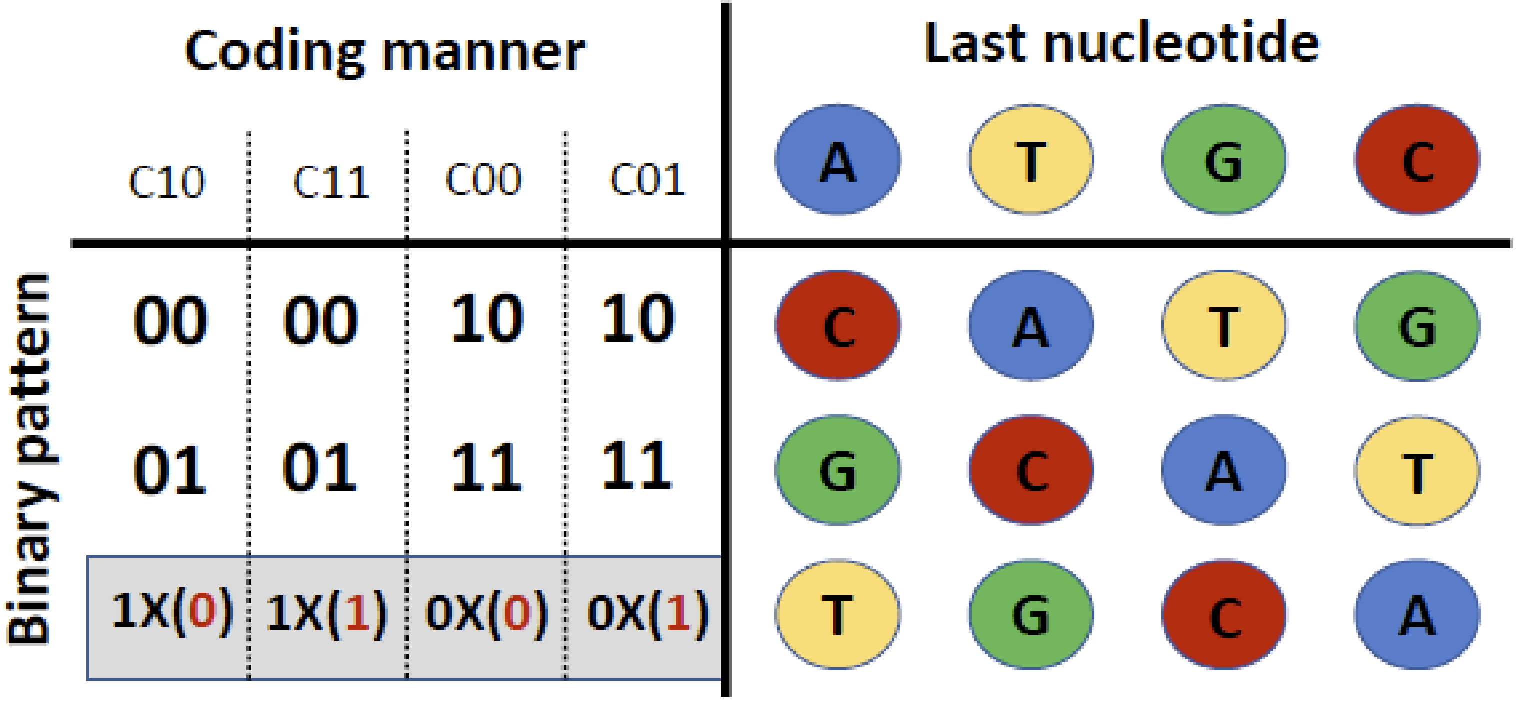
Ternary pattern	A	T	C	G
0	C	A	G	T
1	G	C	T	A
2	T	G	A	C

Binary pattern: Coding manner C0	Binary pattern: Coding manner C1	Last nucleotide A	Last nucleotide T	Last nucleotide G	Last nucleotide C	
00	10	C	A	T	G	High Density
01	11	G	C	A	T	High Density
10	00	TC	GA	CT	AG	Low Density
11	01	TG	GC	CA	AT	Low Density

Common first nucleotide

Lossy code design

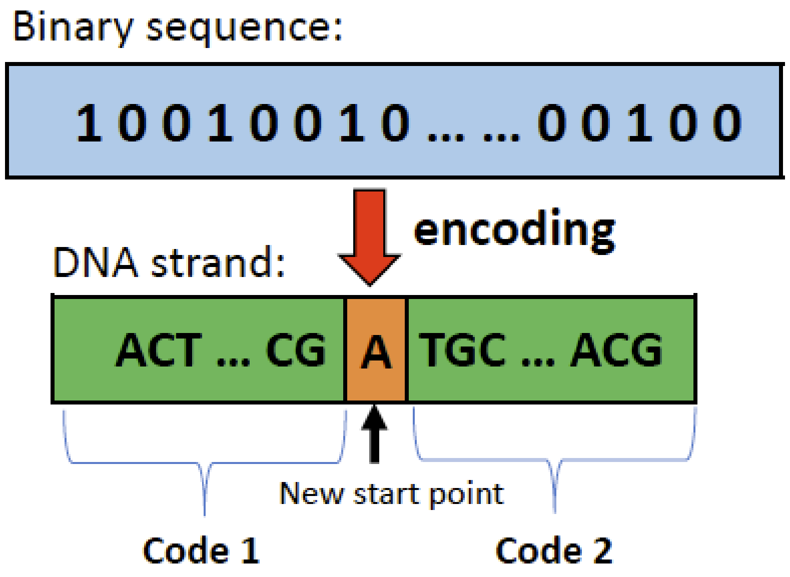
- Combine two low density rows together
- Using four different codes (C10, C11, C00, and C01)
  - Four codes have different error preferences
- 1X(0) indicates 11 and 10 are both encoded into the same nucleotides but will be decoded back to 10



Partition Scheme: Adding ‘Barrier’

”A” as a barrier indicator

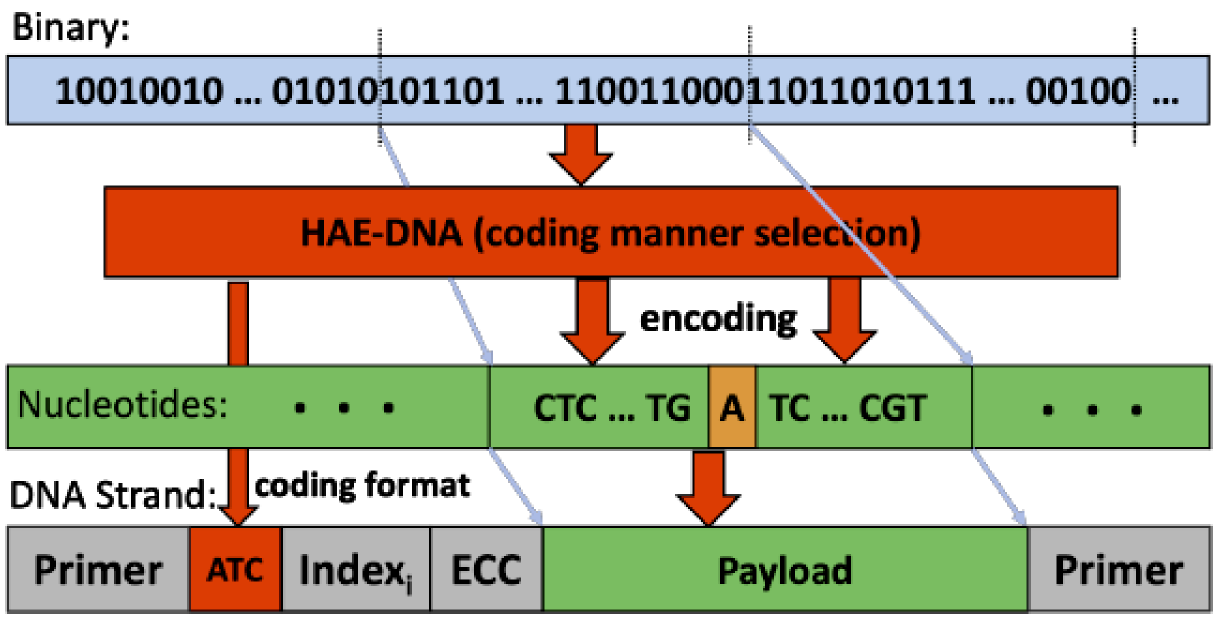
- Improve the robustness of DNA storage like [1]
  - Restricts the error propagation in a partition
- Enable multiple encodings in the same DNA strand to improve the encoding density/reduce error rates induced by the lossy encoding



[1] *B. Li, L. Ou, and D. Du, “Img-dna: approximate dna storage foXr images,” in Proceedings of the 14th ACM International Conference on Systems and Storage, 2021, pp. 1–9.*

Overall Design of HL-DNA

1. Encode binary to nucleotides based on encoding scheme
    - Based on density lossy to select which encoding is used
2. Insert ”barrier” to the DNA sequence
3. Adding the corresponding metadata such as primers, index, ECC, etc.
4. Coding format to indicate multiple encodings in the DNA strand

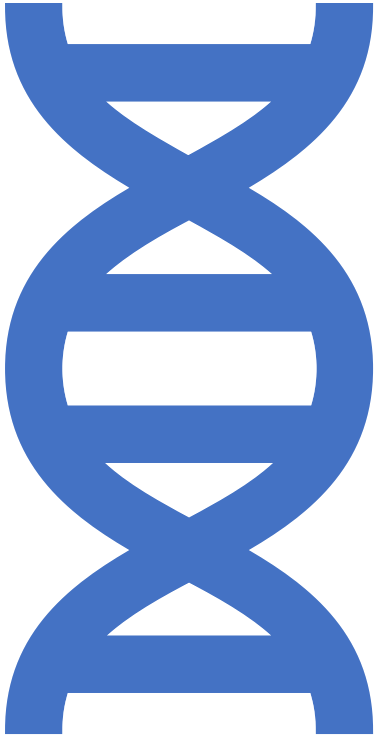


Algorithm 1 HL-DNA Algorithm

Inputs: BinarySeqs //**Binary sequences**//
Outputs: DNASeqs //**DNA sequences**//
procedure HL-DNA ENCODING ALGORITHM(binary sequence BinarySeqs)
  binary_len = length(BinarySeqs)
  for i in binary_len **do**
    Compute frequencies f_{xx} of four binary patterns '11', '10', '01', '00'
    if $f_{11} + f_{10} \geq f_{00} + f_{01}$ **then**
      density_lossless = $\frac{binary\_len}{length(C1(i))}$
      DNA_lossless = C1(i)
    else
      density_lossy = $\frac{binary\_len}{length(C0(i))}$
      DNA_lossless = C0(i)
    if $f_{00} == min(f_{00}, f_{01}, f_{10}, f_{11})$ **then**
      density_lossy = $\frac{binary\_len}{length(C01(i))}$
      DNA_lossy = C01(i)
    else if $f_{01} == min(f_{00}, f_{01}, f_{10}, f_{11})$ **then**
      density_lossy = $\frac{binary\_len}{length(C00(i))}$
      DNA_lossy = C00(i)
    else if $f_{10} == min(f_{00}, f_{01}, f_{10}, f_{11})$ **then**
      density_lossy = $\frac{binary\_len}{length(C11(i))}$
      DNA_lossy = C11(i)
    else
      density_lossy = $\frac{binary\_len}{length(C10(i))}$
      DNA_lossy = C10(i)
    $err = \frac{min(f_{00}, f_{01}, f_{10}, f_{11})}{f_{00}+f_{01}+f_{10}+f_{11}}$
    if density_lossless $\leq$ 1.65bits/nt or $err < Threshold$ **then**
      DNASeqs[i] = DNA_lossy
    else
      DNASeqs[i] = DNA_lossless
Note: C0(), C1(), C00(), C01(), C10(), and C11() are the functions of encoding manners in Fig. 3 and Fig. 4.

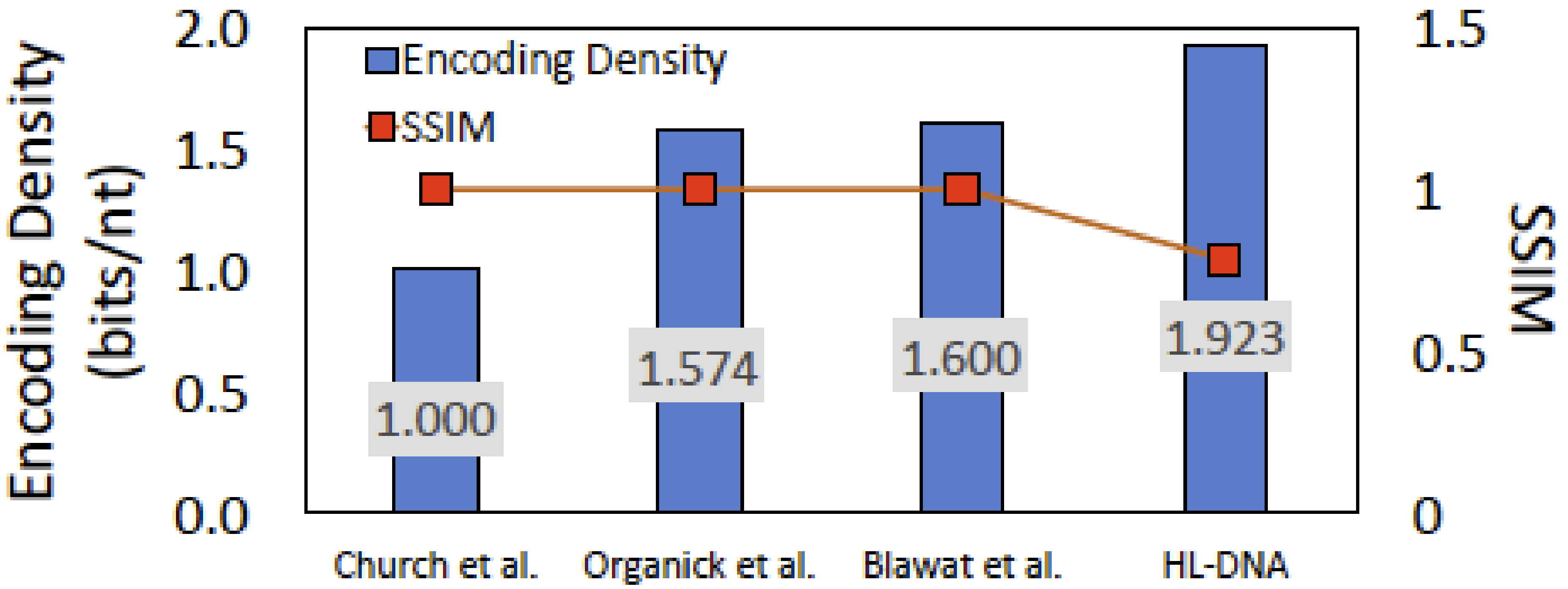
Experimental Setup

- Dataset: ImageNet
- Four schemes:
  - Church et al. [1], Organick et al. [2], Blawat et al. [3], and HL-DNA
- Metric:
  - Encoding density (bits/nt)
  - SSIM (structural similarity index metric)
- DNA strand length 300bp



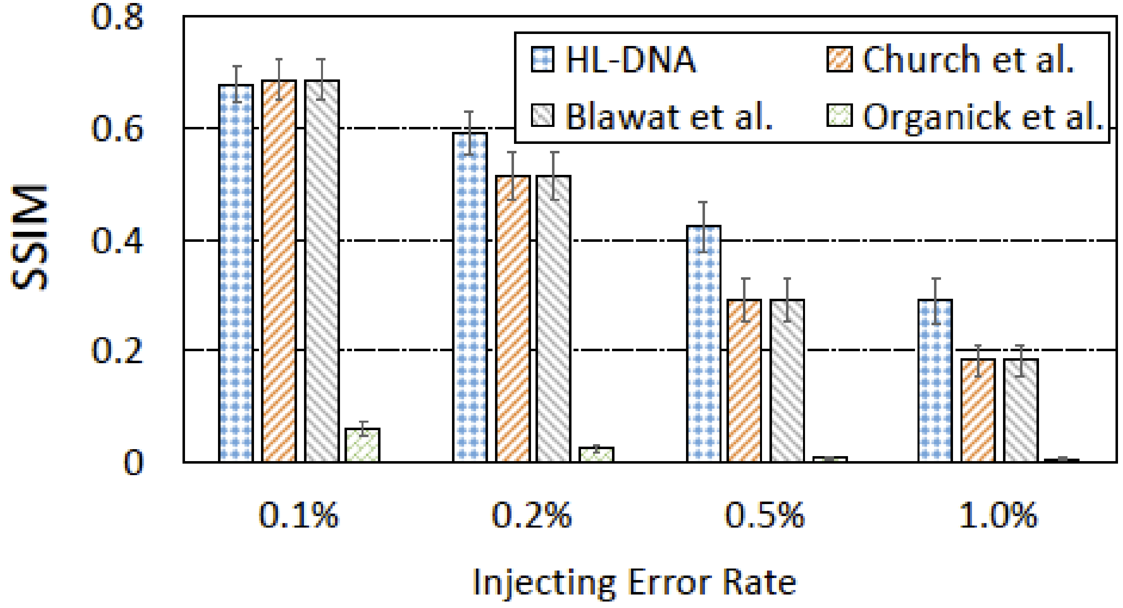
[1] G. M. Church, Y. Gao, and S. Kosuri, “Next-generation digital information storage in dna,” Science, vol. 337, no. 6102, pp. 1628–1628, 2012.
L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin,
[2] K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen et al., “Random access in large-scale dna data storage,” Nature biotechnology, vol. 36, no. 3, p. 242, 2018.
[3] M. Blawat, K. Gaedke, I. Huetter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, and G. M. Church, “Forward error correction for dna data storage,” Procedia Computer Science, vol. 80, pp. 1011–1022, 2016.

Overall encoding density comparison



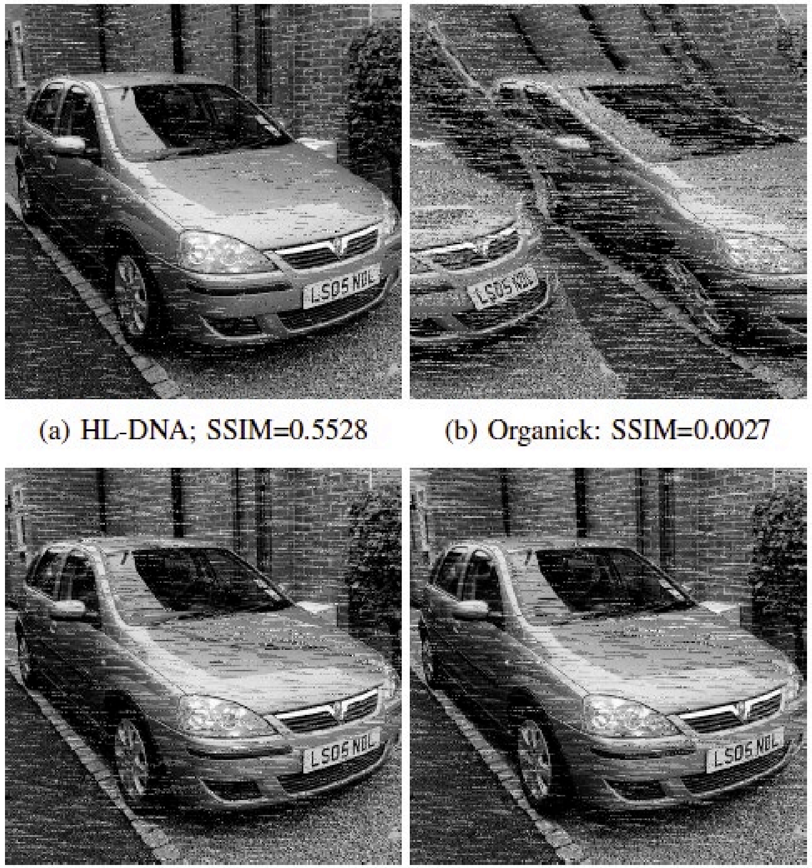
- HL-DNA increases the average encoding density of the previous studies by about 20.2% - 89.4%.
- HL-DNA achieves the highest SSIM, which indicates the best robustness among different schemes.

Robustness of Image-based DNA System



The higher the SSIM is, the better the quality of images is.

A graphic view of an image with different encoding schemes (0.5% error rate):

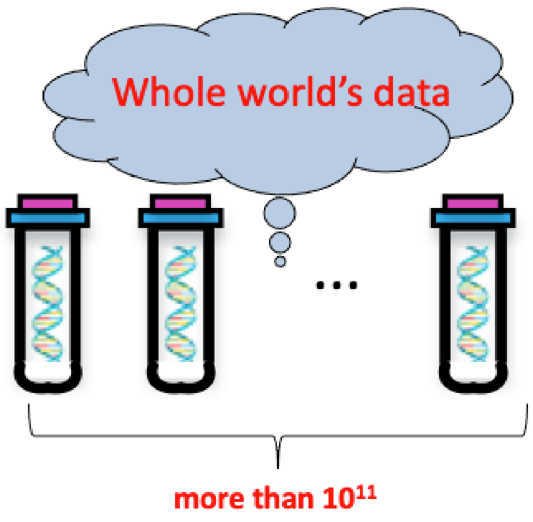


(a) HL-DNA; SSIM=0.5528 (b) Organick: SSIM=0.0027

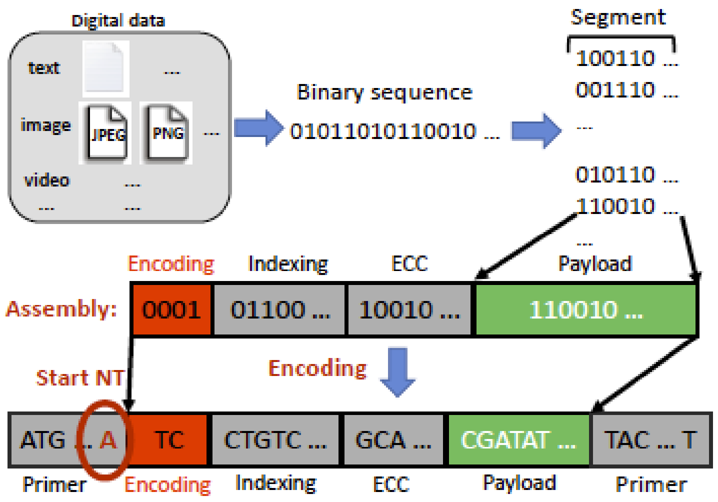
(c) Church: SSIM=0.4482 (d) Blawat: SSIM=0.4567

Potential DNA storage research

Scalability



Capability



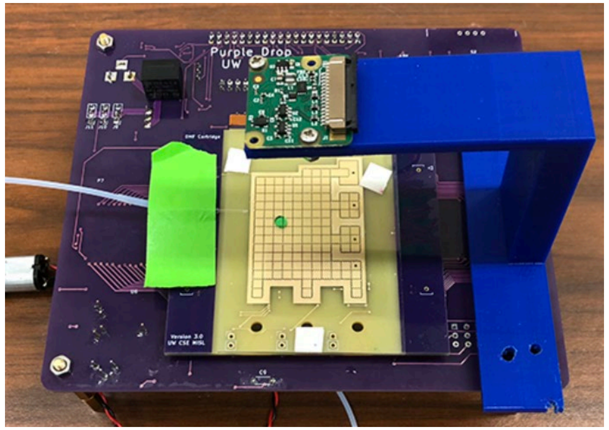
More issues:

- DNA storage preservation
- Issue of limited read number
- Performance of sequencing/synthesis
- API to users

Encoding/ECC

Binary pattern: Coding manner C0	C1	Last nucleotide: A	T	G	C	
00	10	C	A	T	G	High Density
01	11	G	C	A	T	
10	00	TC	GA	CT	AG	Low Density
11	01	TG	GC	CA	AT	

Microfluidic system



Conclusions

- DP-DNA for increase areal density
- IMG-DNA is a robust architecture of DNA storage for images
- A hybrid lossy/lossless encoding based DNA storage architecture called HL-DNA
- Potential DNA storage research directions

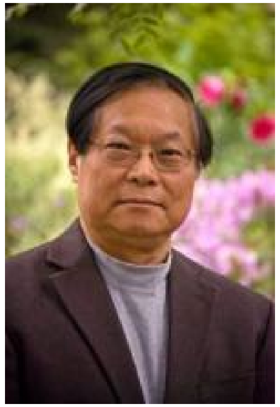
Thanks!
Q&A

UTD

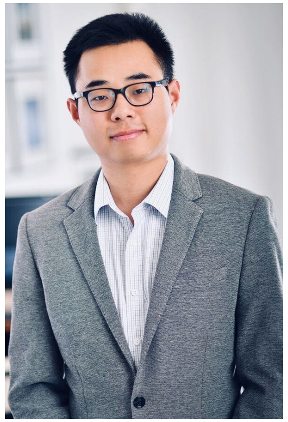
Lab4ISC
Lab for Intelligent Storage and Computing

SDC 23

Acknowledgement



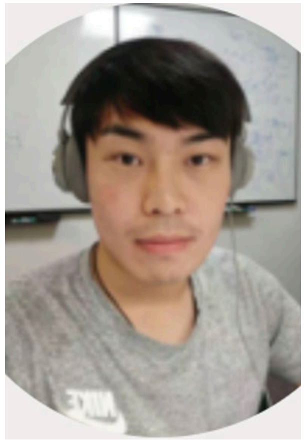
Prof. David Du



Dr. Li Ou

Yi Li

Alex Sensintaiffar



Yixun Wei

NSF

DNA DATA STORAGE ALLIANCE

SNIA®

Lab4ISC

Lab for Intelligent Storage and Computing

Please take a moment to rate this session.

Your feedback is important to us.

