



Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems

Joint works with:

Daniella Bar-Lev¹, Omer Sabary¹,
Ryan Gabrys², Anina Gruica³, Alberto Ravnani³

(1) Technion - Israel Inst. of
Technology



(2) University of California,
San Diego



(3) Eindhoven University of
Technology



Limitations of Existing Technologies

Most of the world's data is stored on **magnetic** and **optical** media

Disks are rated for **3-5** years and tapes **10-30**

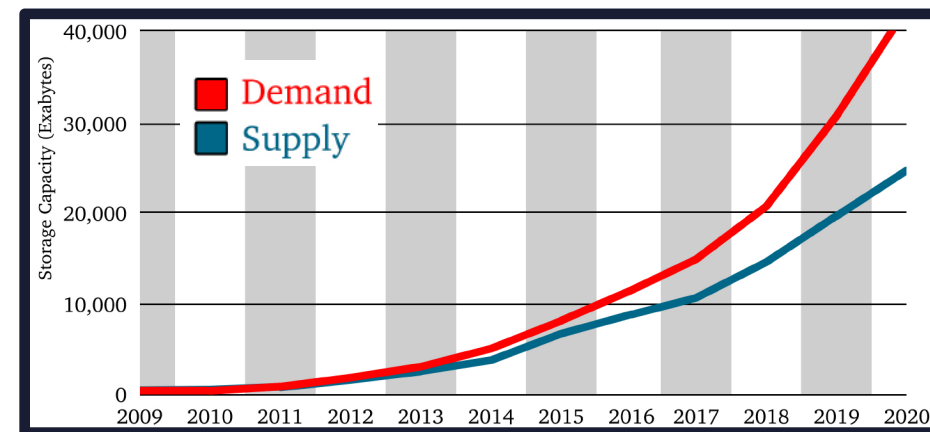
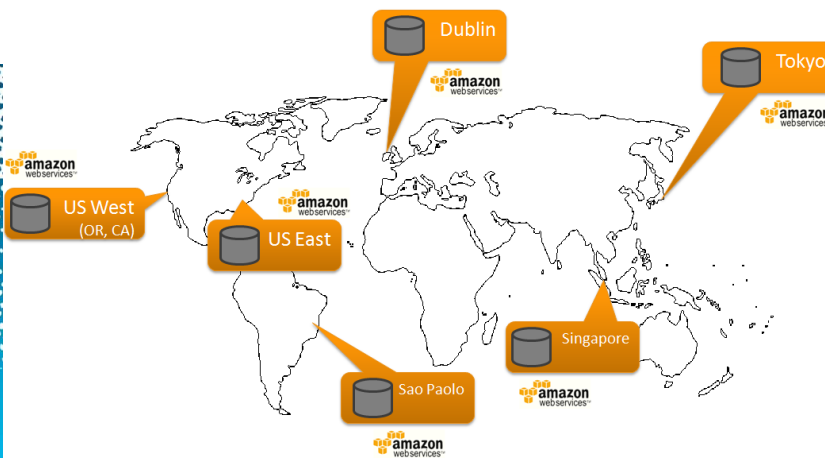
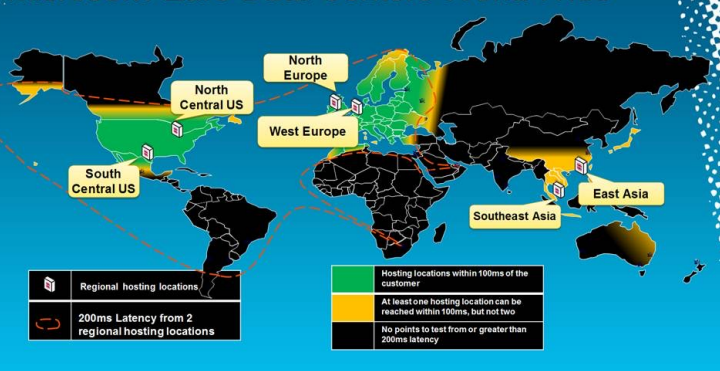


90% OF DATA IN THE WORLD TODAY DIDN'T EXIST TWO YEARS AGO

26X INCREASE IN MOBILE DATA FROM 2010 – 2015 AVERAGING 100% ANNUAL GROWTH

96% OF ORGANIZATIONS ADMITTED THEY COULD DO MORE WITH BIG DATA AND MAKE BETTER USE OF ANALYTICS

Microsoft Azure Data Centers World Wide



DNA as Storage Medium

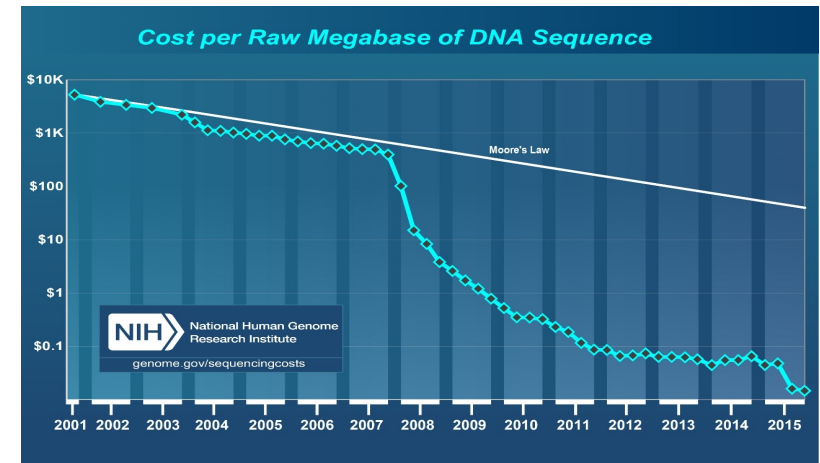
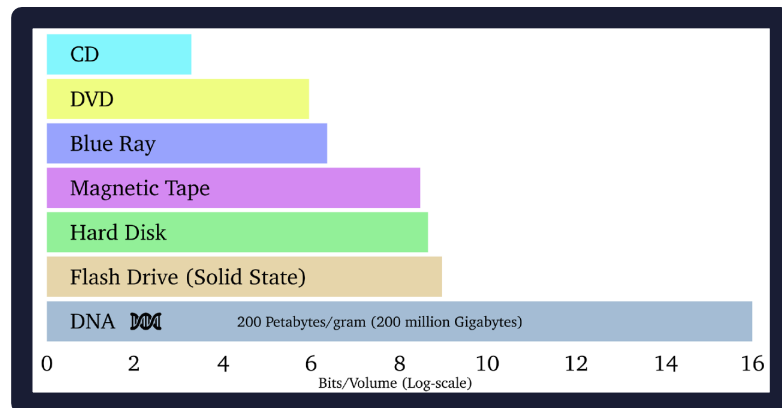
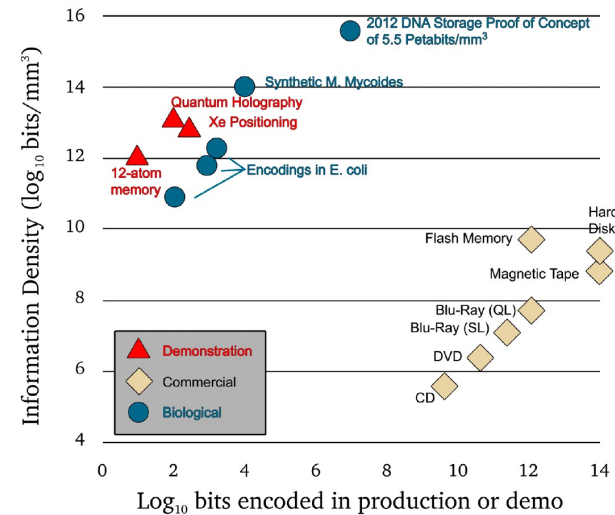
DNA is extremely **durable** - can still recover DNA from mammoths, Neanderthals, and 700,000 old horses!

DNA is **dense**

- Tape: 10-100 GB/mm³
- DNA: **10⁹ GB** /mm³

DNA write (synthesis) and read (sequencing) **costs are decreasing daily**

Can one store user information in DNA?



DNA as Storage Medium

Richard Feynman first proposed the use of macromolecules for storage “*There is plenty of room at the bottom*”

Church et al. (Science, 2012) and **Goldman et al.** (Nature, 2013) stored 643, 739 KB of data in synthetic DNA, resp.

The screenshot shows the Science journal website interface. At the top, the word "Science" is displayed in a large white font on a black background, with "AAAS" in smaller text to its right. Below this is a red navigation bar with links for "Home", "News", "Journals", "Topics", and "Careers". Underneath, a secondary navigation bar lists various scientific fields: "Science", "Science Advances", "Science Immunology", "Science Robotics", "Science Signaling", and "Science Translational Medicine". The main content area features a vertical red bar on the left with the text "Stand up for science" and "SHARE" above social media icons for Facebook, Twitter, and Google+. The article title "Next-Generation Digital Information Storage in DNA" is prominently displayed, followed by the authors "George M. Church^{1,2}, Yuan Gao³, Sriram Kosuri^{1,2,*}". Below the authors, there is a link for "Author Affiliations" and a note: "*To whom correspondence should be addressed. E-mail: sri.kosuri@wyss.harvard.edu". At the bottom, the publication details are listed: "Science 28 Sep 2012; Vol. 337, Issue 6102, pp. 1628; DOI: 10.1126/science.1226355".

The screenshot shows the Nature journal website interface. At the top, the word "nature" is displayed in a white serif font on a dark red background, with "International journal of science" in a smaller white font below it. A "MENU" button with a downward arrow is located to the left of the journal name. Below the journal name, there is a horizontal bar with a color-coded icon (pink, purple, blue, green, yellow, red) and the text "Altmetric: 1190 Citations: 126". To the right of this bar is a link "More detail >>". The main content area features the word "Letter" in a small font, followed by the article title "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA" in a large, bold, black font. Below the title, the authors are listed: "Nick Goldman ✉, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M. LeProust, Botond Sipos & Ewan Birney".

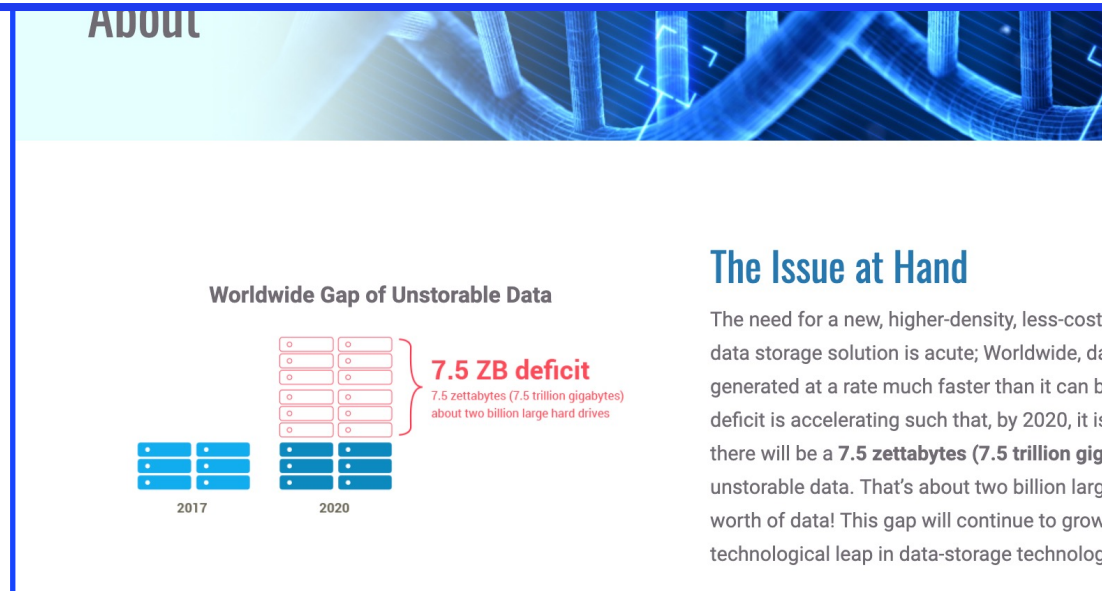
DNA as Storage Medium

- **Richard Feynman** first proposed the use of macromolecules for storage "*There is plenty of room at the bottom*"
- **Church et al.** (Science, 2012) and **Goldman et al.** (Nature, 2013) stored 643, 739 KB in synthetic DNA, resp.
- **Grass et al.:** 2015, 81KB
- **Yazdi et al.:** 2015, random access, rewritable DNA storage system
- **Bornholt et al.:** 2016, 42KB
- **Blawat et al.:** 2016, 22MB
- **Helixworks:** 2016, first commercially available DNA storage medium
- **Erlich & Zielinski:** 2017, 2.11 MB
- **Organick et al.:** 2017, 200MB
- **Yazdi et al.:** 2017, portable and error-free DNA data storage
- **Takahashi et al.:** 2019, end-to-end automation of DNA data storage
- **Tabatabaei et al.:** 2019, DNA punch card
- **Anavy et al.:** 2019, DNA using composite letters
- **DNA Catalog:** 2019, the first to store 16GB of data
- **Iridia:** 2019, complete DNA storage system on a chip
- **Chandak et al.:** 2019, codes for DNA storage using LDPC codes
- **Lee et al.:** 2019, DNA storage using enzymatic synthesis
- **Antkowiak et al.:** 2020, DNA storage using photolithographic synthesis
- **Roquet et al.:** 2021, DNA storage via combinatorial assembly
- **Preuss et al.:** 2021, combinatorial synthesis of DNA shortmers
- **Maes et al.:** 2022, DNA Drive using long double stranded replicative DNA molecules
- **Yan et al.:** 2023, combinatorial synthesis with enzymatically-ligated composite motifs

Twist Bioscience, Illumina and Western Digital Digital Form Alliance with Microsoft to Advance Data Storage in DNA

Press Release November 13, 2020

— Ten Additional Technology Leaders Join Founding Members to Together Advance Industry Roadmap, Set Stage for Widespread Adoption of New Long-term Storage Option —



The Issue at Hand

The need for a new, higher-density, less-cost data storage solution is acute; Worldwide, data generated at a rate much faster than it can be stored. This deficit is accelerating such that, by 2020, it is estimated that there will be a 7.5 zettabytes (7.5 trillion gigabytes) of unstoreable data. That's about two billion large hard drives worth of data! This gap will continue to grow as a result of a technological leap in data-storage technology.

results

Startup packs all 16GB of...

Twist Bioscience, Illumina, Western Digital and Microsoft are leading the DNA Data Storage Alliance as founding members. In addition to the DNA Data Storage Alliance plans to develop and test DNA storage solutions for various industries and industries as well as promote and educate the public about the benefits of DNA storage to promote adoption of this future solution. The following companies have joined the alliance as members:

- [Ansa Biotechnologies](#)
- [CATALOG](#)
- [The Claude Nobs Foundation \(Montreux\)](#)
- [DNA Script](#)
- [EPFL \(École Polytechnique Fédérale de Lausanne\)](#)
- [ETH Zurich Innovation Center \(Montreux Jazz Digital\)](#)
- [ETH Zurich – The Swiss Federal Institute of Technology](#)
- [imec](#)
- [Iridia](#)
- [Molecular Assemblies](#)
- [Molecular Information Systems Lab at the University of Washington](#)

FOUNDERS

- Illumina**: Illumina is improving human health by unlocking the power of the genome. Our focus on innovation has...
- Microsoft**: Microsoft's "Project Scorpion" is a mission to create a new form of digital storage by encoding data in DNA.
- Twist Bioscience**: At Twist Bioscience, we work to create a new form of digital storage by encoding data in DNA.
- Western Digital**: About the company, Western Digital creates innovation for your data. As a leader...

MEMBERS

- ANSA**: Ansa Biotechnologies is developing a new way to make DNA that will have faster, cheaper, and more accurate...
- Battelle**: Every day, the people of Battelle apply science and technology to solving what matters most. For more...
- CATALOG**: Founded by MIT scientists, CATALOG is the world's first company to develop a commercial DNA storage...
- The Claude Nobs Foundation**: The Claude Nobs Foundation promotes the creation and commercialization of Swiss high-tech...
- DNA Script**: Founded in 2014 in Paris, DNA Script is a commercial DNA synthesis company engineering...
- EPFL**: Located in Switzerland, EPFL is one of Europe's most vibrant and open universities in science...
- ETH Zurich**: Freedom and individual responsibility, entrepreneurship and open-mindedness: ETH...
- Université Côte d'Azur CNRS I3S Lab**: The research conducted at I3S is a unique blend of scientific disciplines, addressing major challenges in science and society...
- Imagene**: Imagene offers a disruptive technology for non-temperature sensitive preservation of DNA stored...
- imec**: imec is a world-leading research and innovation hub in semiconductor and digital technologies...
- Iridia**: Iridia is a private, venture-backed US company developing an ultra-high density data storage solution for...
- KIOXIA**: Kioxia is a world leader in memory storage solutions, including the development, production and sale of...
- Molecular Assemblies**: Molecular Assemblies Inc. is a private biotech company developing an enterprise DNA storage system through...
- PFU**: PFU provides leading research systems that hold the top number of shares in the world. We will...
- Quantitative Scientific Solutions**: QSS is a private, venture-backed US company developing an ultra-high density data storage solution for...
- Quantum**: Quantum technology will advance human capabilities and create a new digital era...
- SEAGATE**: Seagate is creating the data future with pioneering technology insights and innovation.
- Spectra Logic**: Spectra Logic develops data storage and data management solutions that solve the most complex storage...
- Spectra**: Spectra Logic develops data storage and data management solutions that solve the most complex storage...
- University of Arizona**: At the University of Arizona Center for Applied Research in Cancer and Medicine, we study, we learn, we grow.
- MISL W**: MISL W is a private, venture-backed US company developing an ultra-high density data storage solution for...
- Digital Preservation Coalition**: The Digital Preservation Coalition works to ensure the long-term digital legacy. We make our members' digital...
- Los Alamos National Laboratory**: With 75,000 square meters of research space for long-term digital legacy, we make our members' digital...
- Los Alamos National Laboratory**: Los Alamos National Laboratory is a leader in developing and applying science and technology to meet...
- Cinémathèque Suisse**: The Cinémathèque Suisse is the national film archive of Switzerland. Its collections include over 40,000 titles, acquired over the years...
- 21e8**: 21e8 is a private, venture-backed US company developing an ultra-high density data storage solution for...
- DNALI**: DNALI is a private, venture-backed US company developing an ultra-high density data storage solution for...
- University of Marburg**: Forward-looking topics with high scientific quality are the focus of research at the University of Marburg. The focus is on...
- CENTRILLION**: Centrillion Technologies is a private, venture-backed US company developing an ultra-high density data storage solution for...
- Stanford Compression Forum**: The Stanford Compression Forum (SCF) is a joint effort between academic and industrial leaders in the field of Data Compression...
- Bole State University, NAM Institute**: Bole State is a non-profit public university that provides academic and professional programs for students in a changing world...
- Hyperion Research**: Hyperion Research helps organizations make efficient decisions and foster growth opportunities by providing research...
- Duke**: Duke University is a private, non-profit research university in Durham, North Carolina. It is one of the leading universities in the United States...
- BioMemory**: BioMemory is working with its customers to meet the challenge of global data growth by the coming decade. BioMemory is designed...
- Gupta Lab @ DA-ICT**: Research in our lab currently focuses on the design and development of information processing systems...
- ICMS @ Eindhoven University**: The Institute for Complex Molecular Systems (ICMS) at Eindhoven University of Technology focuses on interdisciplinary research...
- TUM**: The key expertise of the COD group at TUM covers coding theory and security, including fast storage and error correction of parity data, coded...
- Information Storage & Memories @ TU/e**: The Information Storage and Memories (ISM) group at TU/e is a leading center for research in understanding the storage of and applica...
- euroKARE**: euroKARE is a company dedicated to the development of a new form of digital storage by encoding data in DNA. It is a joint effort between academic and industrial leaders in the field of Data Compression...
- CacheDNA**: CacheDNA provides a unique solution for enterprise data storage and access to reduce costs and improve compliance with regulatory requirements for data and...

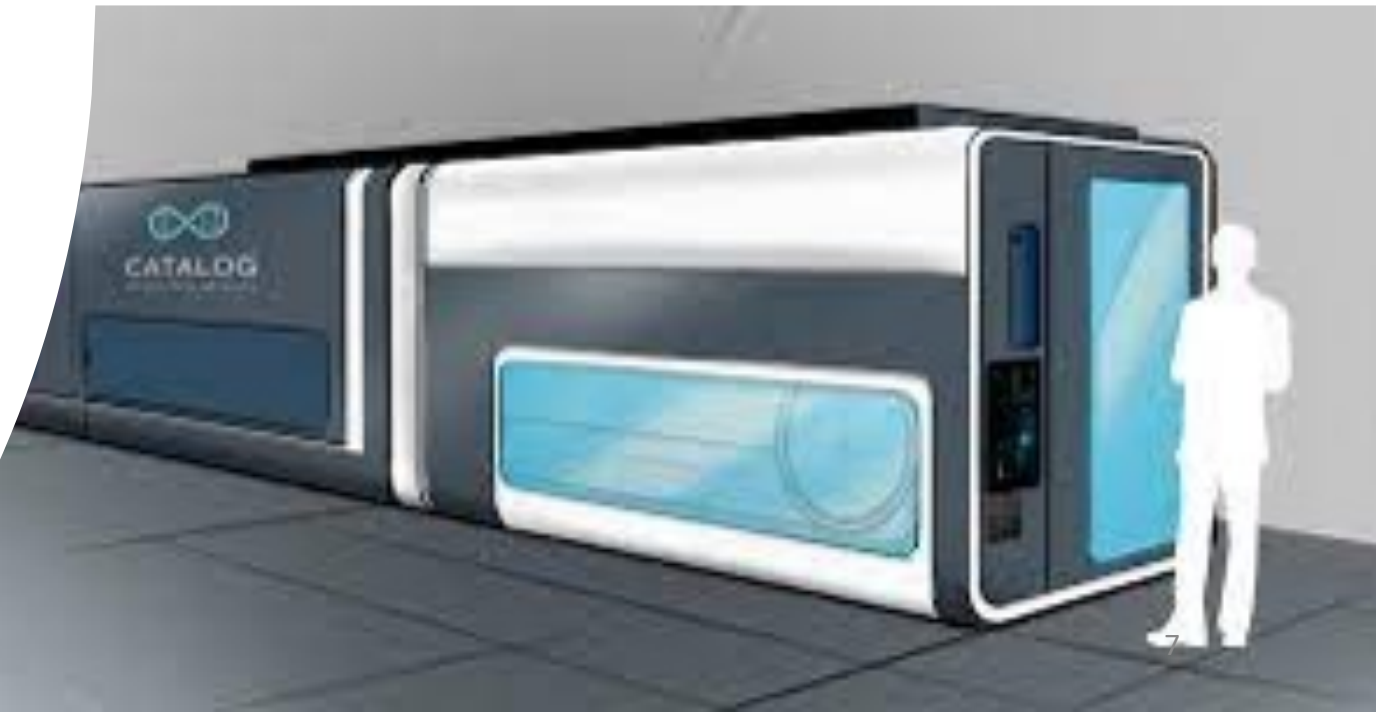
Startup Catalog

CATALOG- Enterprise storage \$35M raised- DNA Archiving

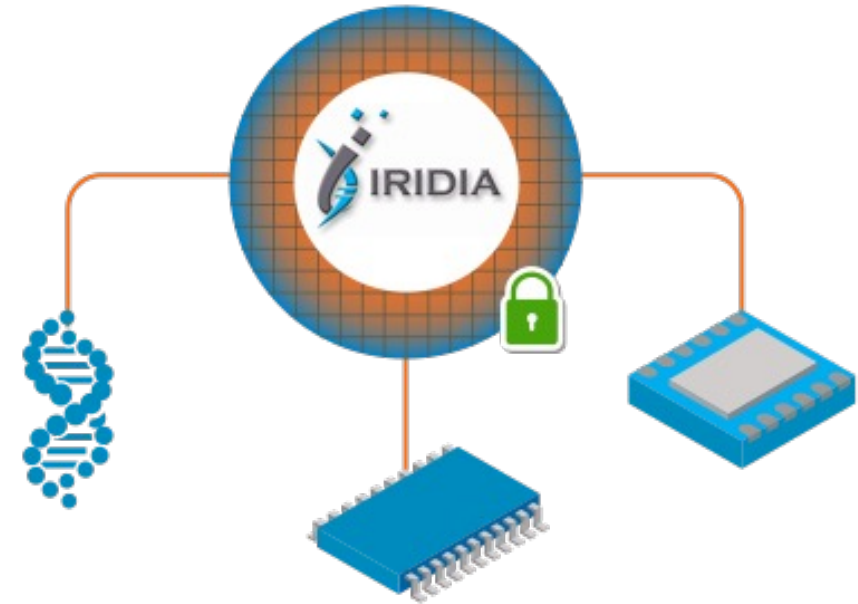
Catalog's technology relies on a device that feeds blank webbing at 16 meters per minute into a modified inkjet printer that deposits drops of synthetic DNA on the web.

That webbing is then moved to an incubation chamber to represent the data, which is then written to a flask of DNA.

Reading the data can be done with a DNA sequencer.



Iridia- Chip scale storage- \$24M Raised

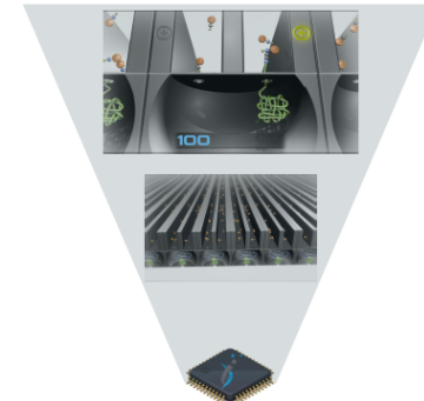


Integrated, High Precision & Distributed Writing, Reading and Storage of DNA on Chip



On Chip "Writing" of DNA

- No moving parts
- No microfluidics
- Longer DNA strands
- No toxic waste (enzyme catalyzed)
- Higher quality DNA (in process QC)
- Less expensive (single molecule)
- Leverages existing semiconductor manufacturing infrastructure
- Massively parallel processing



On Chip "Reading" of DNA

- No PCR required
- No sample prep required
- Don't need to 'assemble' sequences
- Faster
- "Zero" read costs
- Long reads
- Leverages existing semiconductor manufacturing infrastructure
- Massively parallel processing

On Chip Storage of DNA

- Unparalleled data density
- Unparalleled data durability
- Ultra-low power consumption

DNA Storage Companies/Groups



DNA Data Storage: Global Markets and Technologies

- **BBC Research Report**

- The global market for DNA data storage should grow from \$36.4 million in 2020 to \$525.3 million by 2025 with a compound annual growth rate (CAGR) of 70.6% for the period of 2020-2025.
- North American DNA data storage market should grow from \$29.1 million in 2020 to \$340.1 million by 2025 with a compound annual growth rate (CAGR) of 63.5% for the period of 2020-2025.
- European DNA data storage market should grow from \$4.4 million in 2020 to \$95.7 million by 2025 with a compound annual growth rate (CAGR) of 85.1% for the period of 2020-2025.

- **Brandessence Market Research Report**

- At 65.8% CAGR, DNA Data Storage Market Size is Expected to Reach USD 1926.7 Million by 2028

Synthesis and Sequencing Costs

- **Synthesis**

- **Twist/Agilent**

- 100,000 200-base strands cost \approx \$20K (1MB = \$4.2K)

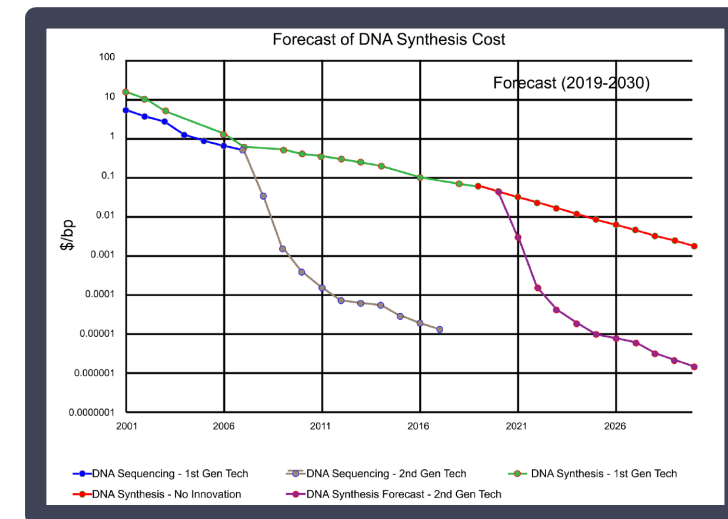
- **Sequencing**

- **Technion Genome Center: Illumina HiSeq**

- \$2500 for 200M strands

- **Oxford Nanopore Technologies MinION sequencer**

- \$1000 for a single run (flow cell) to read 10^{10} bases = 50M strands



DNA as Storage Medium

Goal: Build a fully operational, cost-efficient, real-time, DNA-based storage system

Important challenges:

Cost of synthesis and sequencing

Lack of appropriate coding solutions

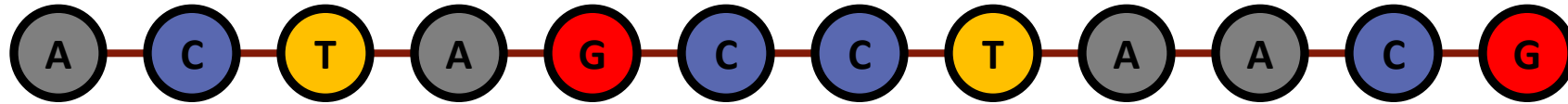


DNA Intro

- DNA consists of 4 bases, aka nucleotides:

Adenine (A) Cytosine (C) Guanine (G) Thymine (T)

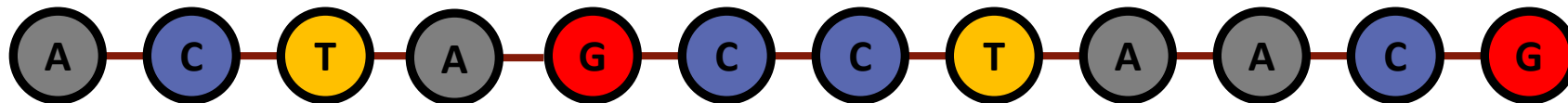
- DNA strand, aka oligonucleotide, is a string of the nucleotides



- Convert a binary sequence into a quaternary sequence

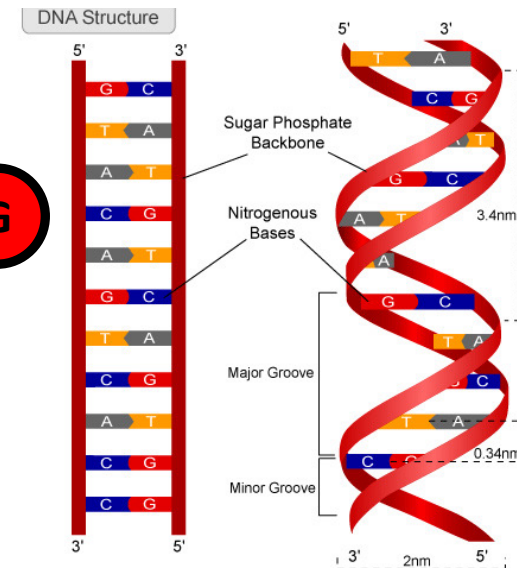
• A = 00 C = 01 G = 10 T = 11

• 00.01.11.00.10.01.01.11.00.00.01.10



- However...

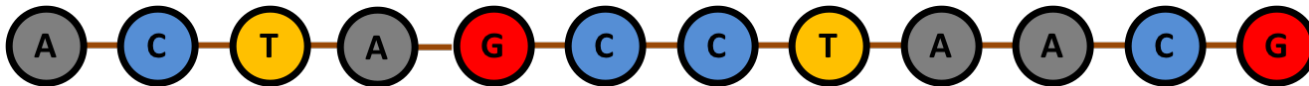
- Strands are limited in their size (~200 bases)
- Strands are not ordered (a soup with many strands)



How to Write Data into DNA?

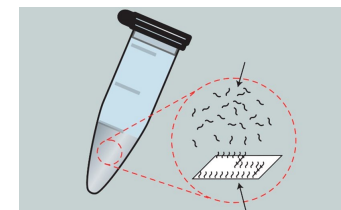
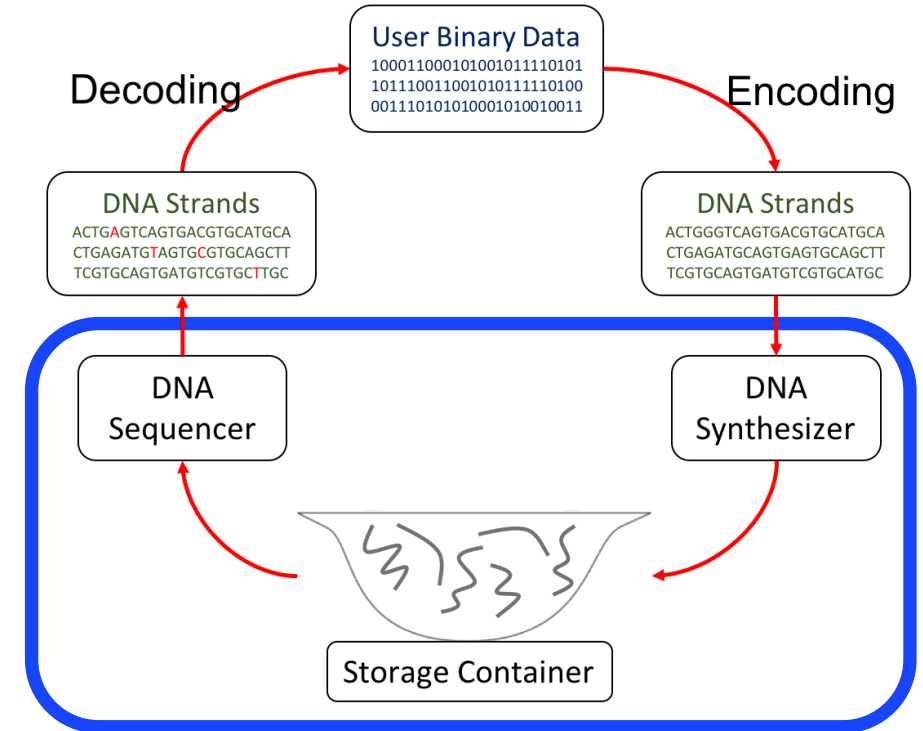
- **DNA Synthesis:** artificially generating DNA strands

- Strands are generated by appending one base at a time
- Typical lengths are **~200** bases (due to technology limitations)
- Each strand has thousands copies

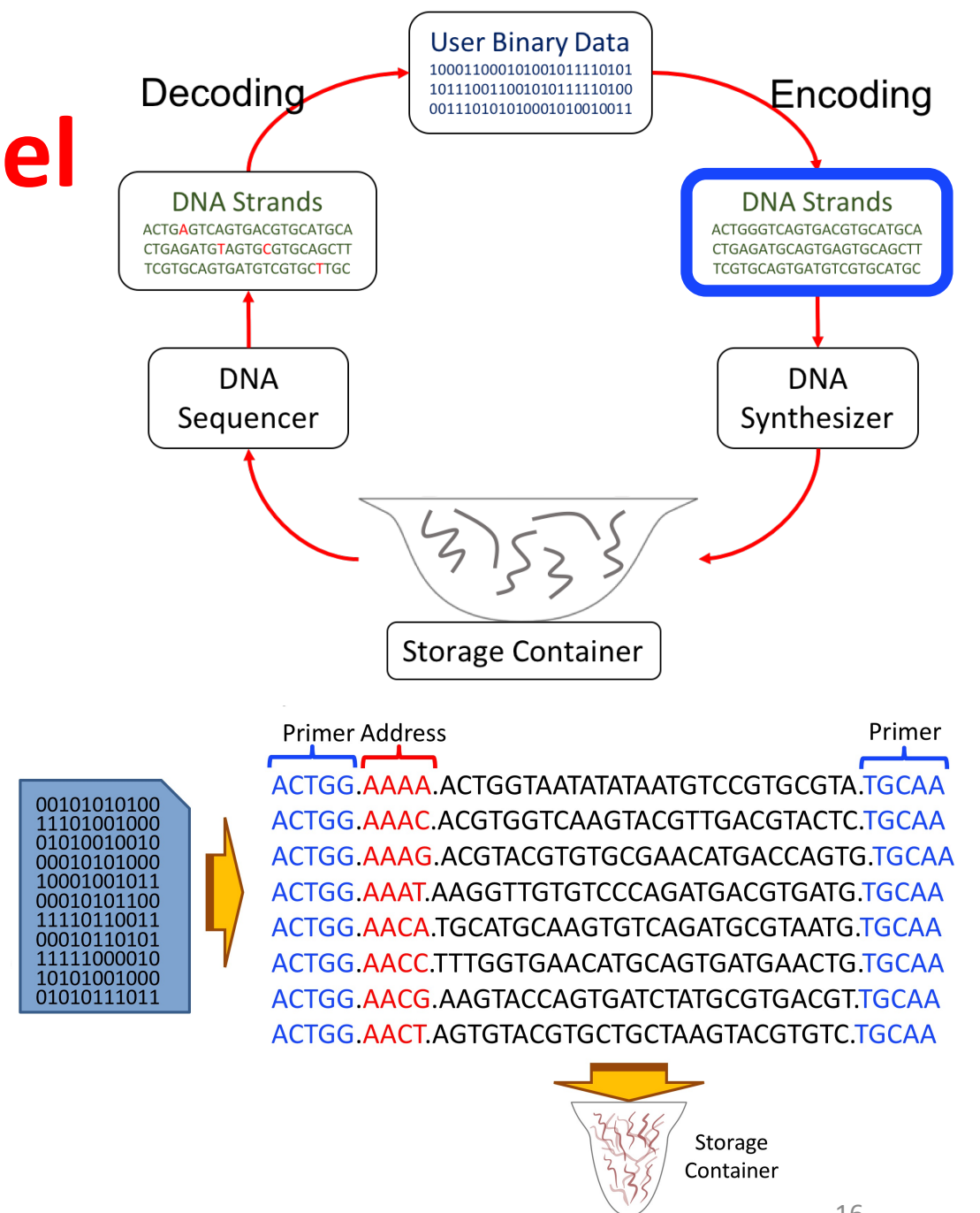
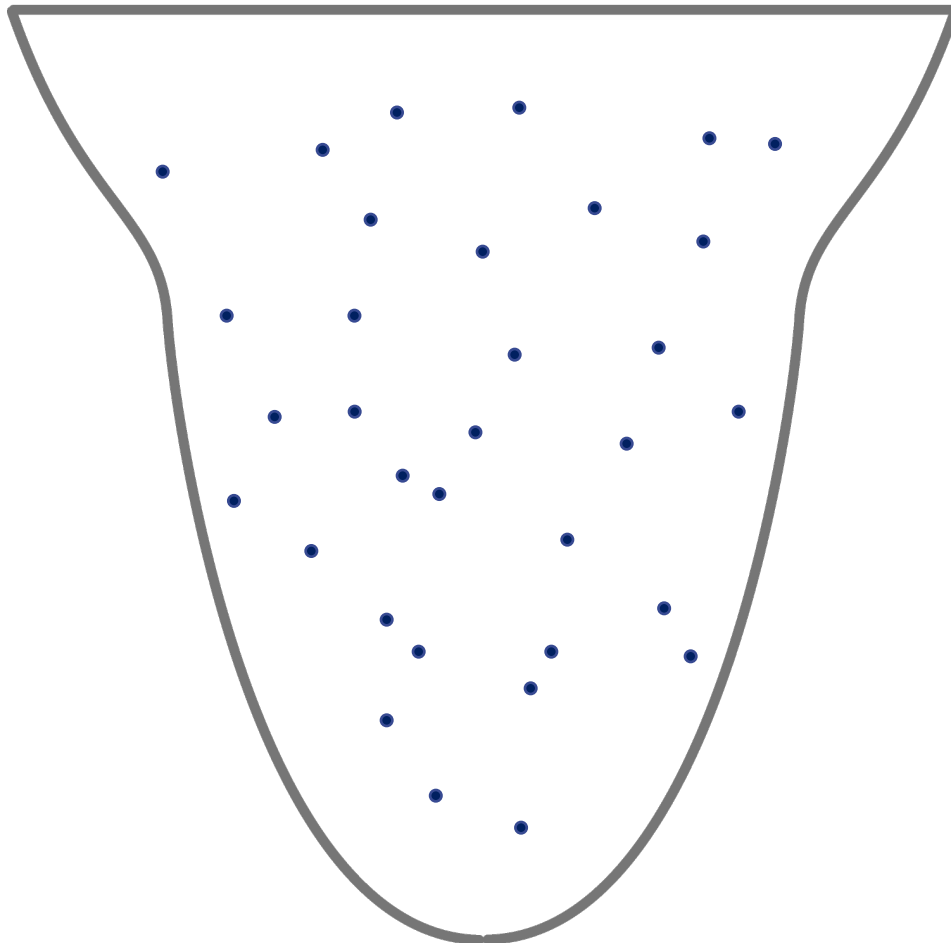


- **DNA Sequencing:** reading DNA strands

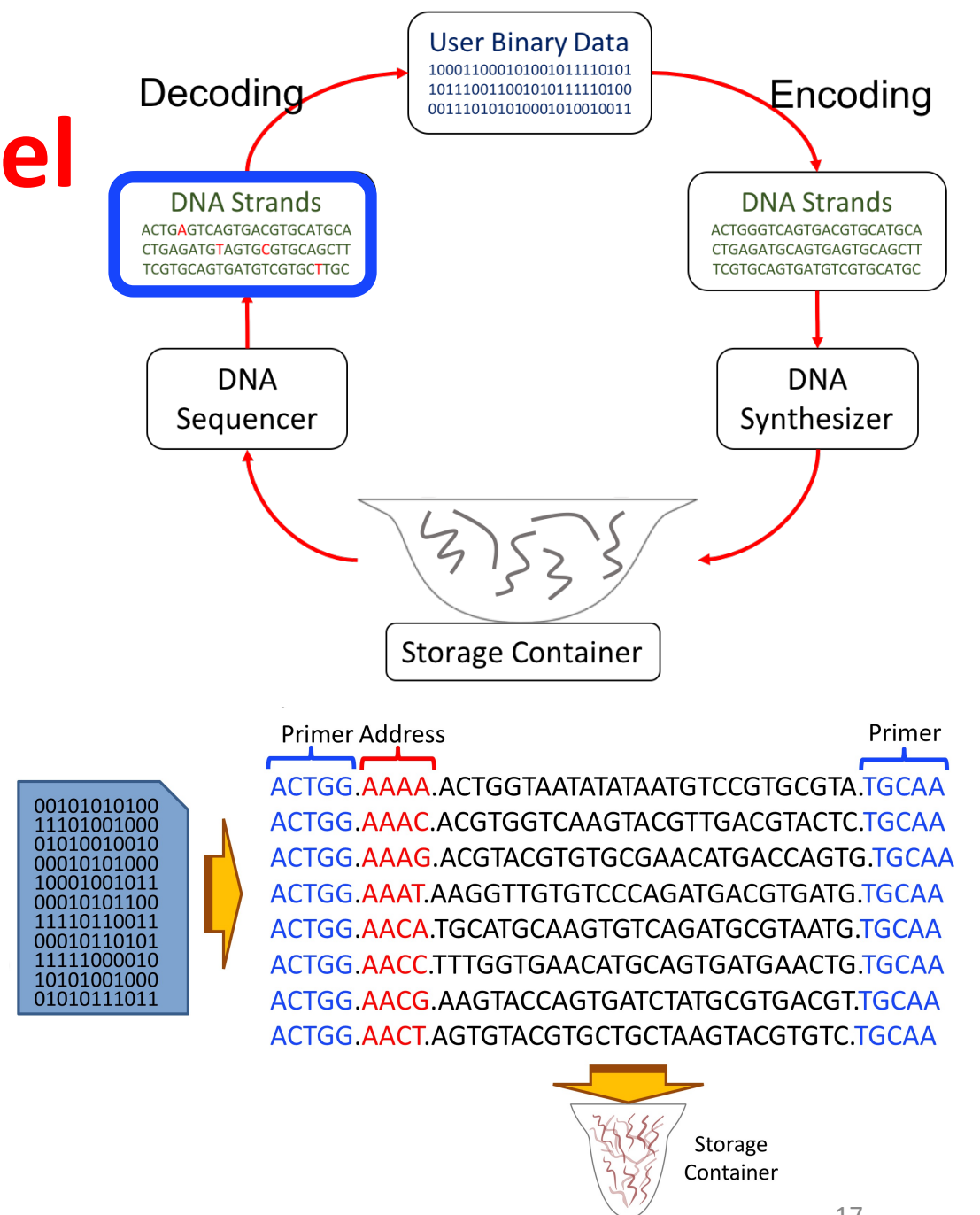
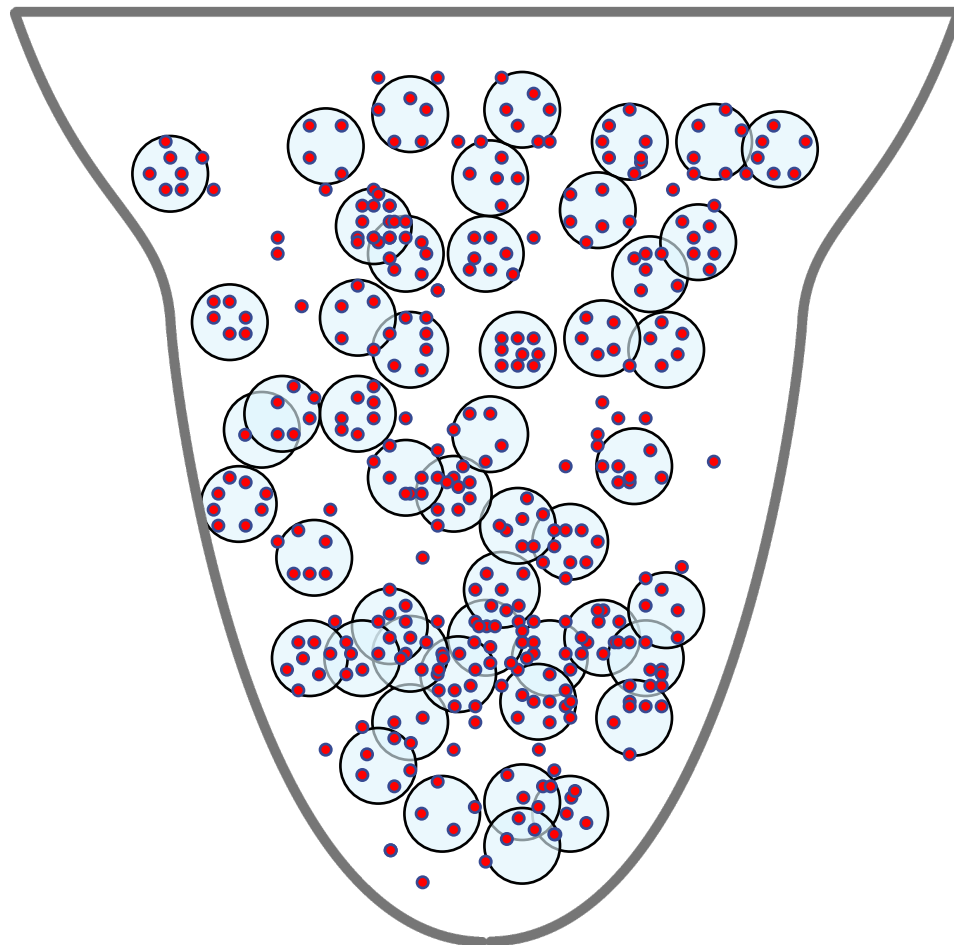
- Generating many reads of each strand
- Less expensive and faster than synthesis (per base)



DNA Storage Channel Model

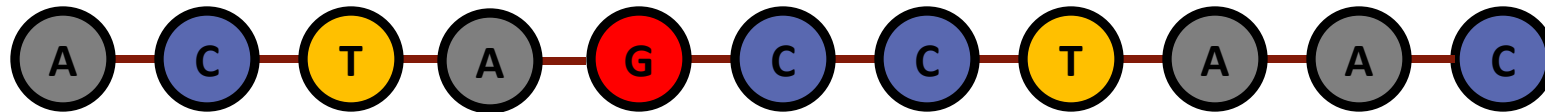


DNA Storage Channel Model

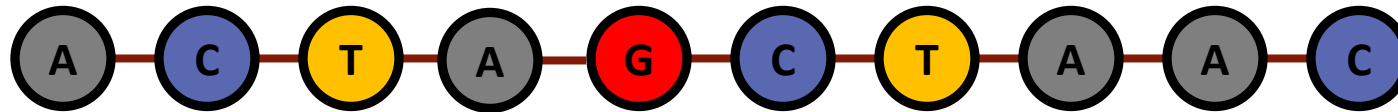


Errors in DNA

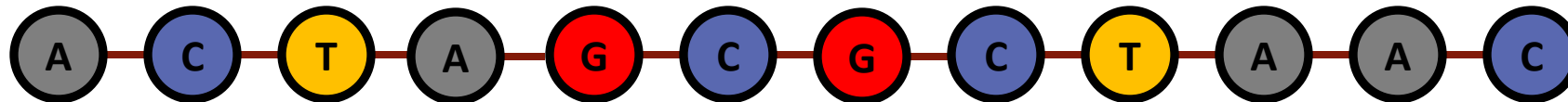
- Both synthesis and sequencing can cause errors



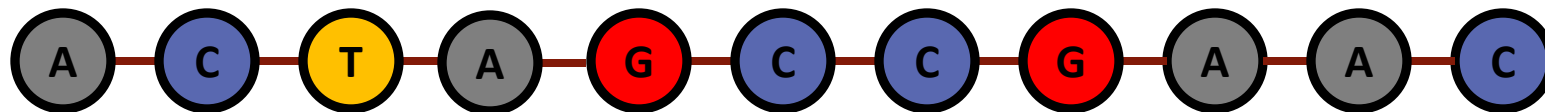
Deletions



Insertions

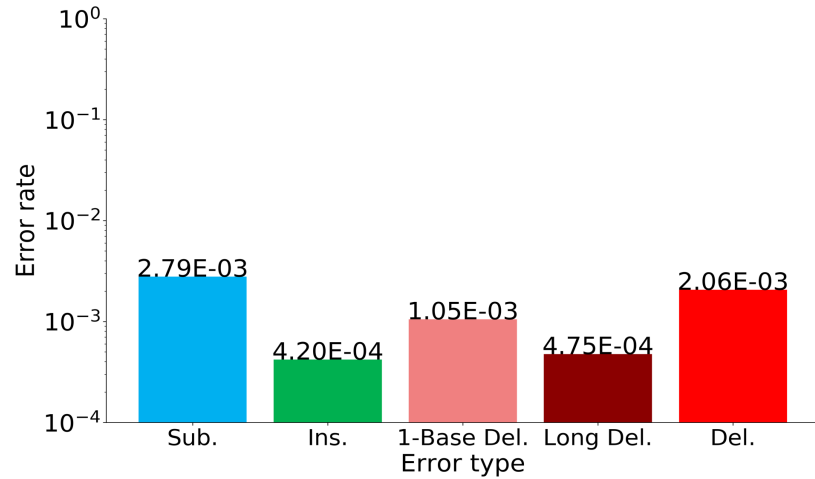


Substitutions

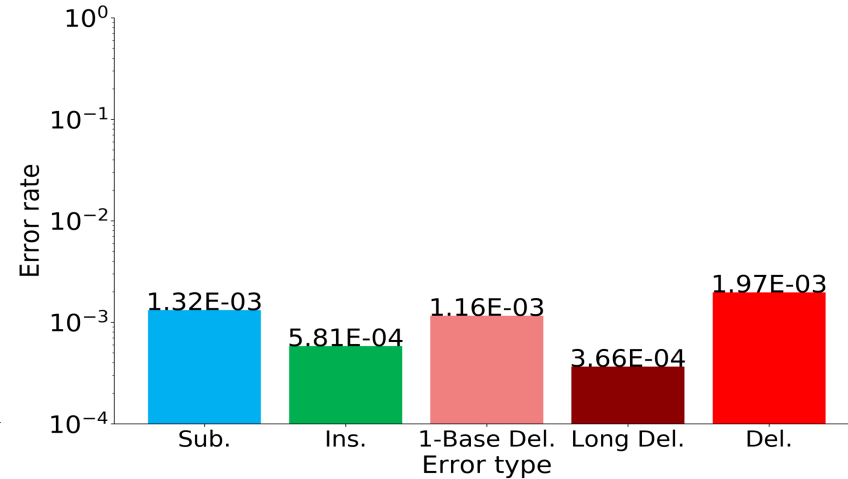


Error Characterization

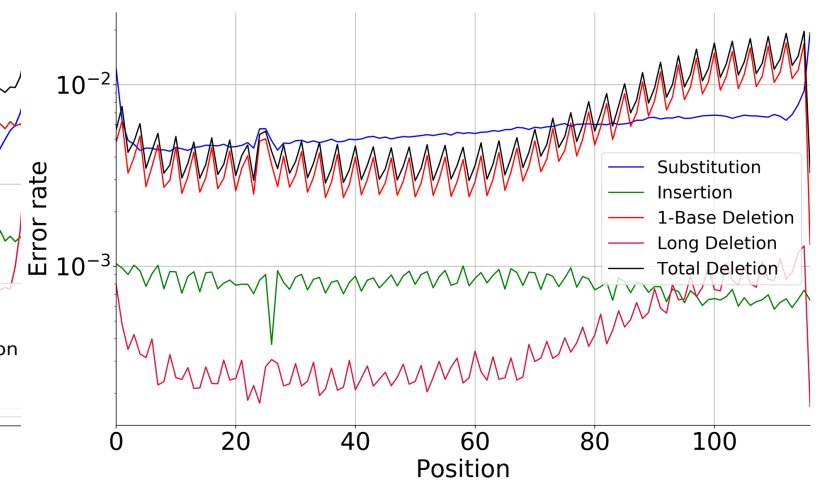
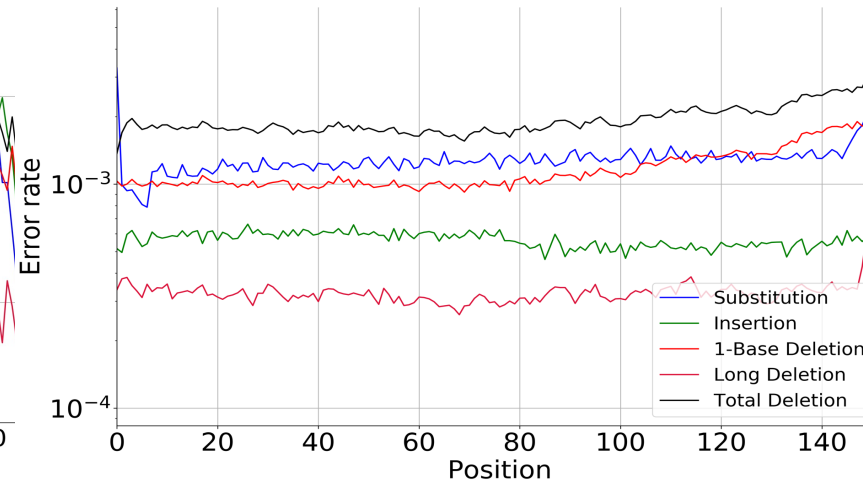
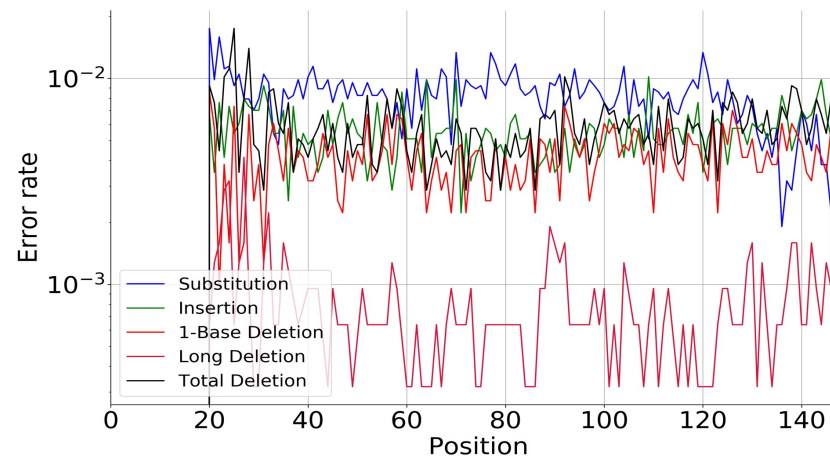
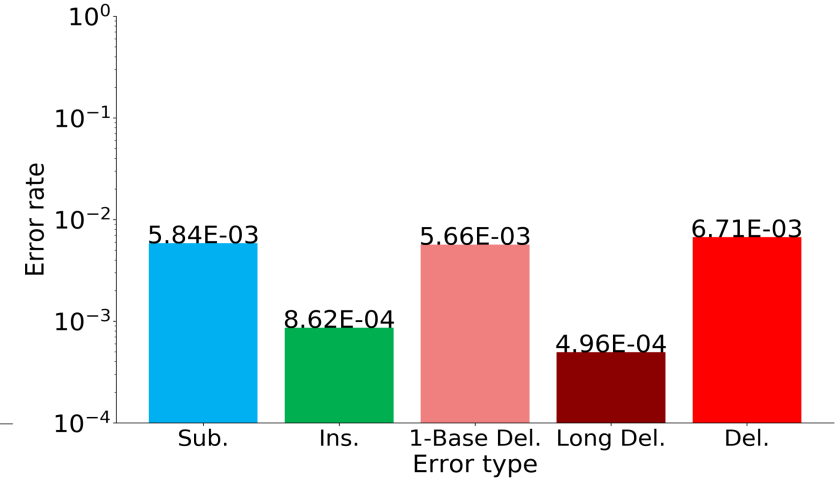
Organick et al., 22MB



Erlich & Zielinski, 2.11 MB



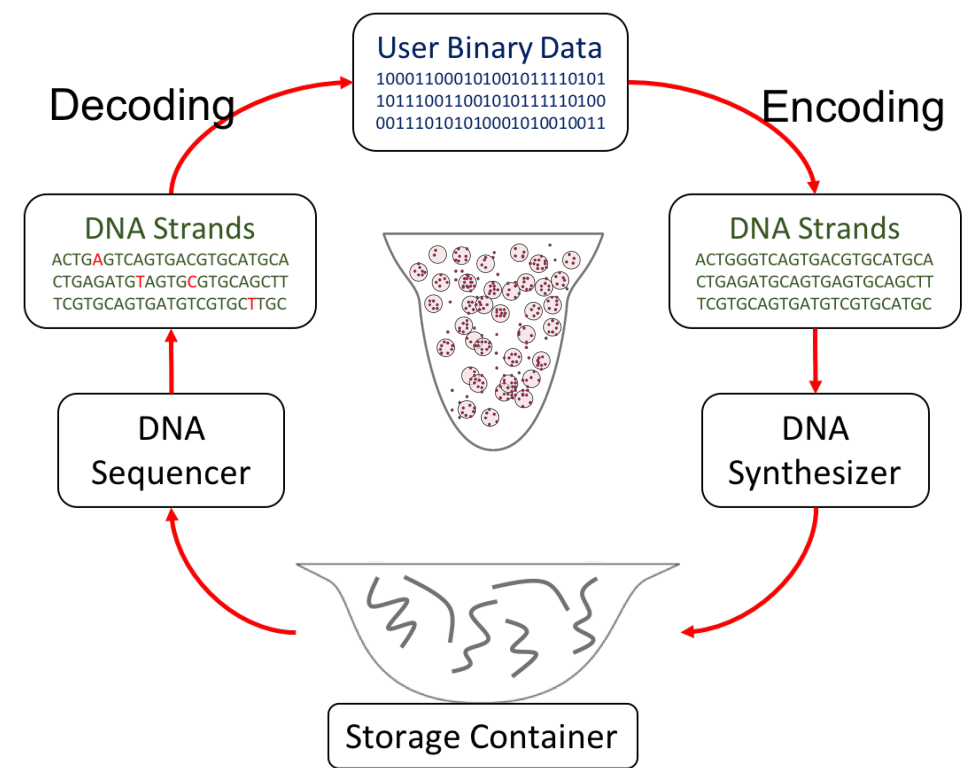
Grass et al., 81KB



Coding Problems

- **Main goals of coding for DNA-storage**

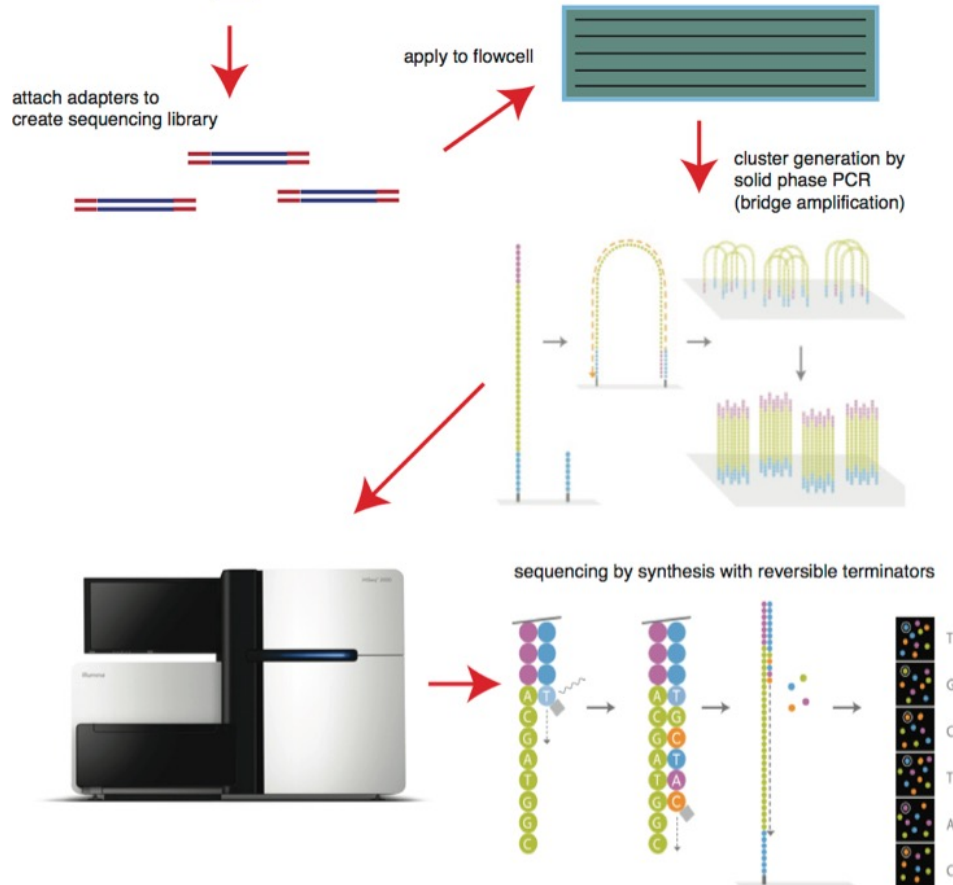
- Clustering algorithms
Clustering specifically for the errors in DNA
- Reconstruction of sequences
Reconstruction of different sequences together
- Constrained codes
Avoiding the specific bad patterns in DNA such as long homopolymers and GC content
- Codes correcting insertions/deletions
Codes correcting combinations of deletions, insertions, and substitutions



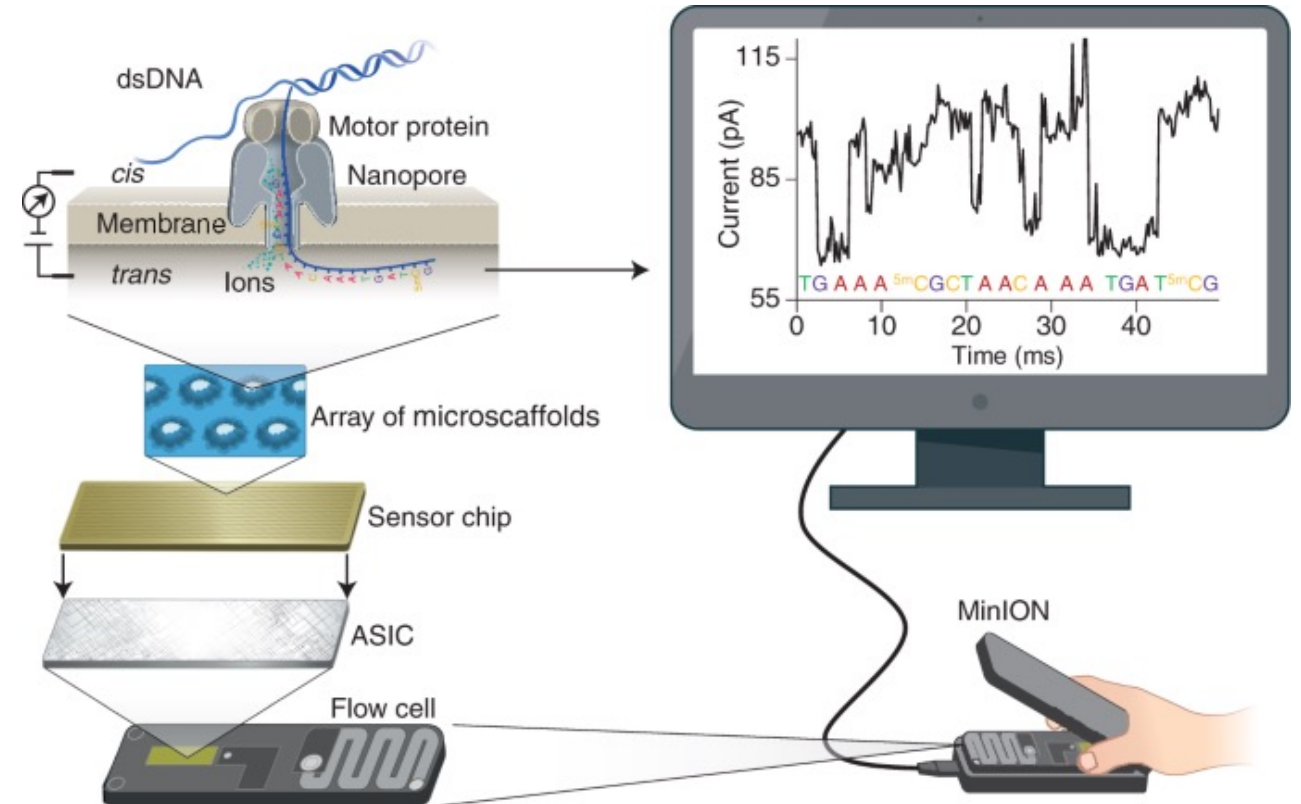
How to Sequence DNA Strands?



Illumina

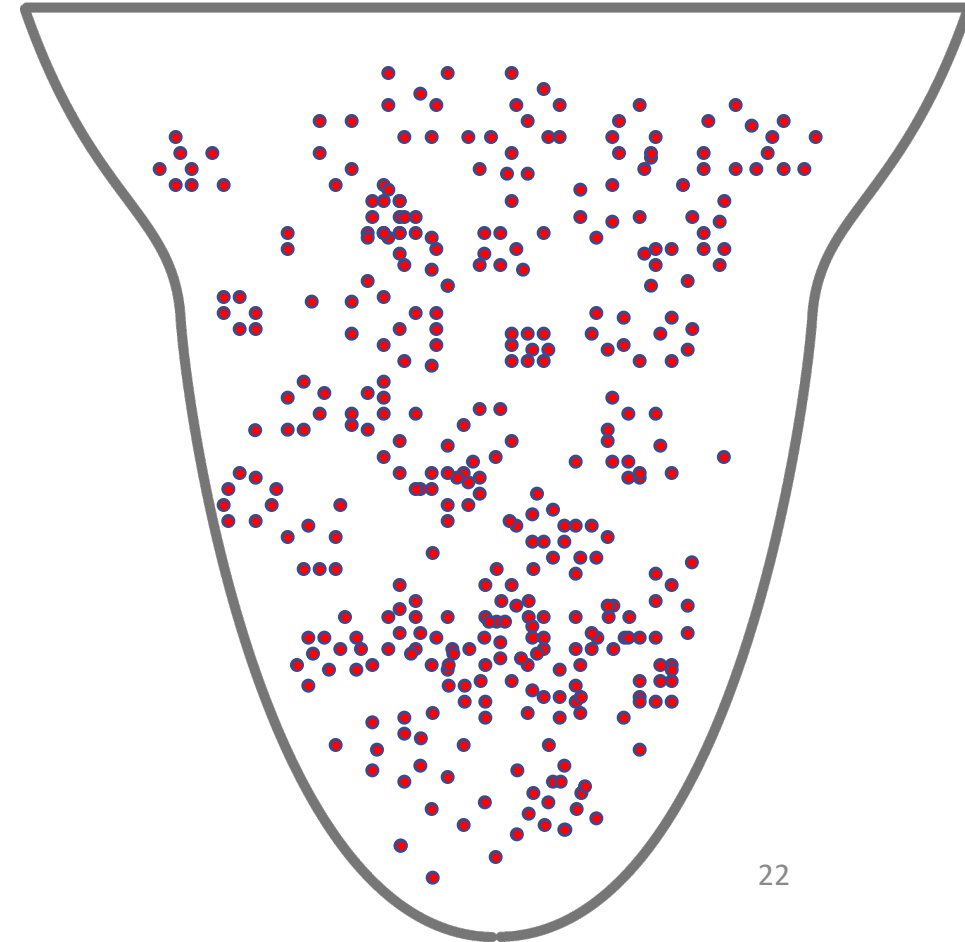
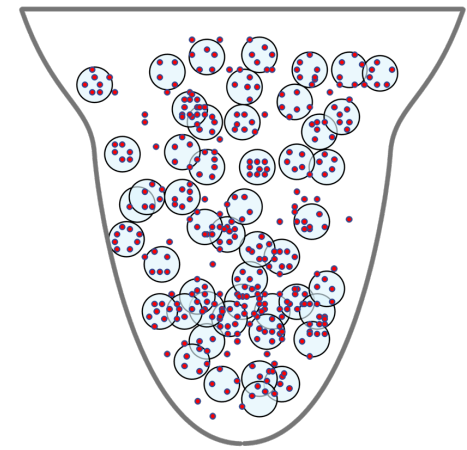


Nanopore



The Coverage Depth Problem

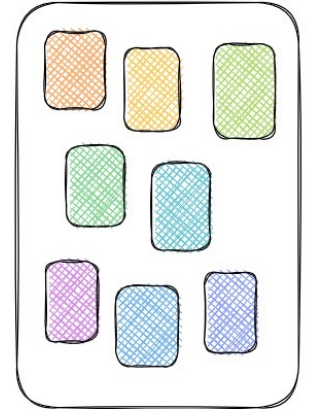
- **Assumptions:**
 - The file is encoded into n strands, each has millions of copies
 - During sequencing, the strands are randomly read until the file is decoded
- **The problem:** Find the expected number of reads and the probability to decode the file
- The answer depends upon:
 - The code
 - The noise model
 - The reading distribution of the strands



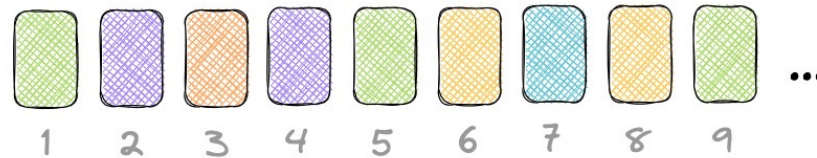
The Coupon's Collector Problem

- First studied by **Feller** in 1967
- **The problem:** If each box of cereal contains one out of n coupons, how many cereal boxes one should expect to buy to collect all n coupons?

How many coupons do you expect you need to draw with replacement before having drawn each coupon at least once?



There are n different coupons



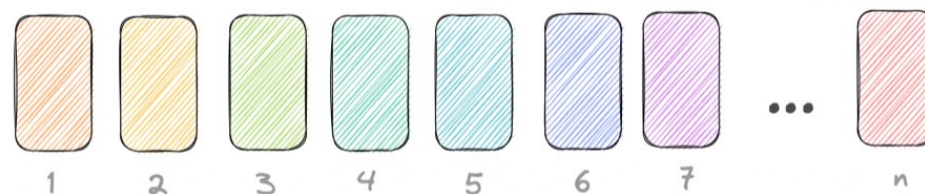
- **Solution:**

- T : #draws, t_i : time to collect the i -th new coupon
- $T = t_1 + t_2 + t_3 + \dots + t_n$
- Each t_i has geometric dist. w/ succ. prob. $p_i = \frac{n-i+1}{n}$ and expectation $\frac{1}{p_i} = \frac{n}{n-i+1}$
- $E[T] = E[t_1] + E[t_2] + \dots + E[t_n] = \frac{n}{n} + \frac{n}{n-1} + \dots + \frac{n}{1} = n\left(\frac{1}{n} + \frac{1}{n-1} + \dots + \frac{1}{2} + \frac{1}{1}\right)$
 $= nH_n = n\log(n) + \gamma n + 0.5 + O\left(\frac{1}{n}\right)$, $\gamma \approx 0.57$ the Euler-Mascheroni const.

The Dixie Cup Problem/The Urn Problem

- First studied by **Newman** in 1960
- **The problem:** Given n urns, what is the expectation of the number of thrown balls in order to have at least t balls in each urn?

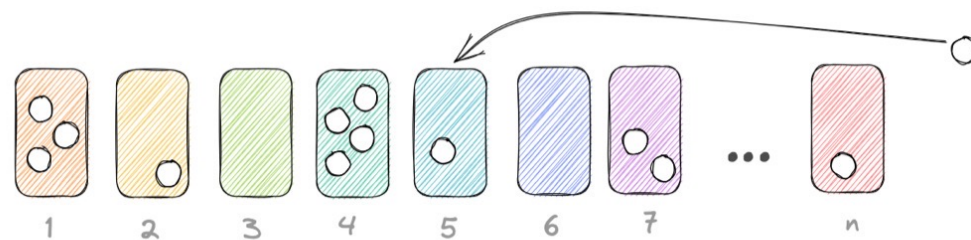
There are n different urns



Identical balls are thrown into the urns and in each round one ball is thrown into one of the urns randomly.
How many balls do you expect you need to throw into the urns, with replacement, before having all the urns not empty?

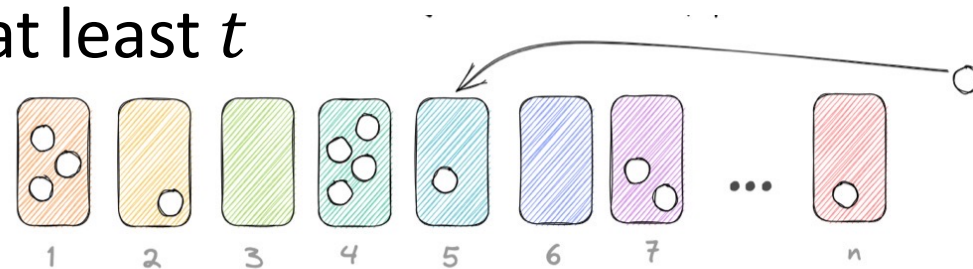
- **Other extensions:**

- It is sufficient to have only k out of the n urns, each with at least t balls
- Different distributions to throw balls to the urns



The Dixie Cup Problem/The Urn Problem

- First studied by **Newman** in 1960
- **The problem:** Given n urns, what is the expectation of the number of thrown balls in order to have at least t balls in each urn?



- **Known results:**

- $k = n, t = 1: nH_n = n \log(n) + \gamma n + 0.5 + O\left(\frac{1}{n}\right)$

- $k < n, t = 1: n(H_n - H_{n-k}) \approx n \log\left(\frac{n}{n-k}\right)$

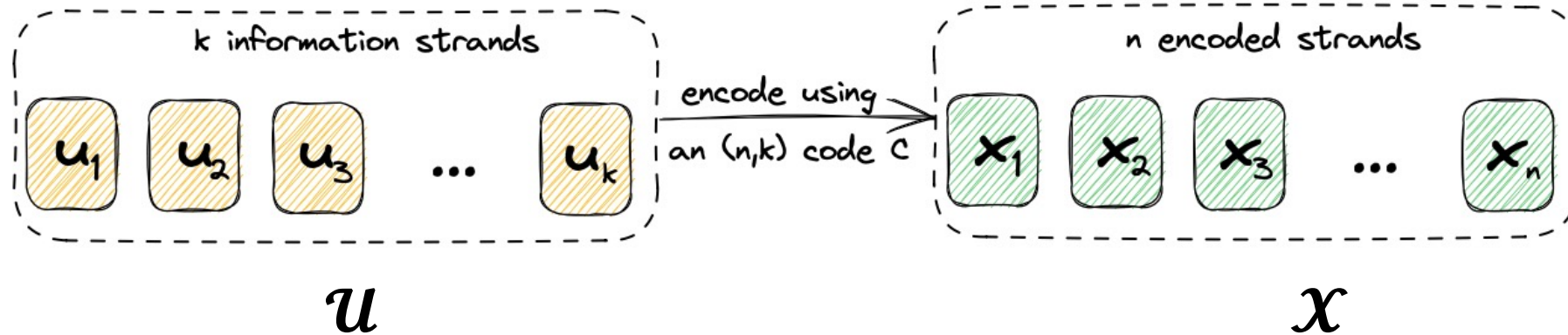
- $k = n, t > 1: n \log n + n(t-1) \log \log n + nC_t + o(n)$

- $k < n, t > 1: \sum_{q=0}^{k-1} \int_0^{\infty} [u^q] \prod_{i=1}^n (e_{t-1}(p_i v) + u(e^{p_i v} - e_{t-1}(p_i v))) e^{-v} dv$

$$e_t(x) = \sum_{i=0}^t \frac{x^i}{i!}$$

The Coverage Depth Problem

k information strands are encoded into n strands using an (n, k) code \mathcal{C}



Main goal: Study the required **sample size** M to guarantee successful decoding of \mathcal{U}

$v_t^{\mathbf{p}}(\mathcal{C})$ - r.v. of the number of samples for successful decoding of \mathcal{U}

$v_t^{\mathbf{p}}(n, k)$ - when \mathcal{C} is an MDS code

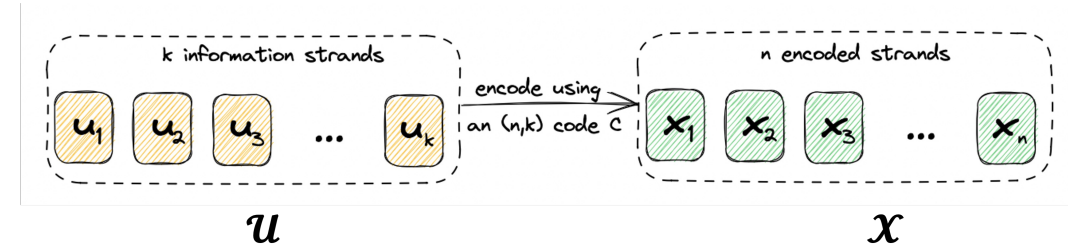
If \mathbf{p} is the uniform distribution, it is removed from the notation

The Coverage Depth Problem

Problem 1 - The MDS coverage depth problem

For any k, n , find:

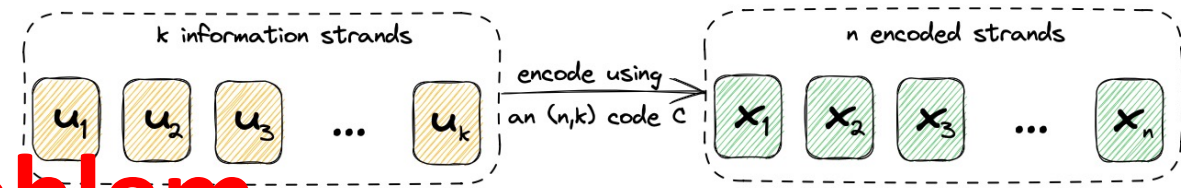
- 1 The expectation value $\mathbb{E}[v_t(n, k)]$
- 2 The probability distribution of $v_t(n, k)$, i.e., for any $m \in \mathbb{N}$ find the value $P[v_t(n, k) > m]$



Problem 2 - The coding coverage depth problem

For any k, n , find:

- 1 Given n, p , find an (n, k) code C that minimizes $\mathbb{E}[v_t^p(C)]$
- 2 The minimum value of $\mathbb{E}[v_t^p(C)]$ over all possible C, p .
That is, the value $M^{\text{opt}}(k) \triangleq \liminf_{C, p} \{\mathbb{E}[v_t^p(C)]\}$



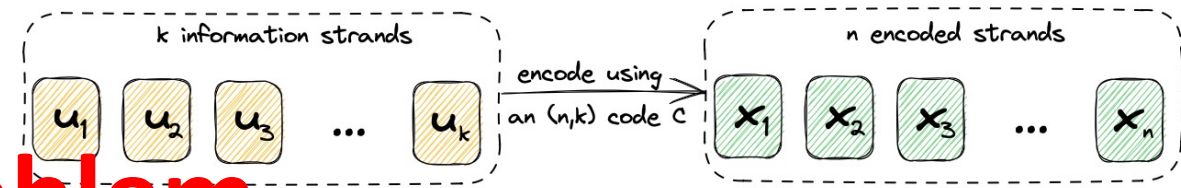
The Coverage Depth Problem

The Noiseless Channel ($t = 1$)

- **The uncoded case:** There are n strands, and all of them should be sampled
Solution: Coupon collector's problem: $\mathbb{E}[v_1(n, n)] = n \log(n) + \gamma n + O(1)$

- **The coded case:** k of the n strands should be sampled

$$\mathbb{E}[v_1(n, k)] = n \log \left(\frac{n}{n-k} \right) = \frac{k}{R} \log \left(\frac{1}{1-R} \right), \quad \mathbb{E}[v_1^p(n, k)] = \int_0^\infty e^{-nv} \cdot \sum_{q=0}^{k-1} \left(\sum_{\substack{I \subseteq [n] \\ |I|=q}} \prod_{i \in I} (e^{p_i v} - 1) \right) dv$$



The Coverage Depth Problem

The Noiseless Channel ($t = 1$)

- **The uncoded case:** $\mathbb{E}[\nu_1(n, n)] = n \log(n) + \gamma n + O(1)$
- **The coded case:** $\mathbb{E}[\nu_1(n, k)] \approx n \log\left(\frac{n}{n-k}\right)$, $\mathbb{E}[\nu_1^p(n, k)] = \int_0^\infty e^{-nv} \cdot \sum_{q=0}^{k-1} \left(\sum_{\substack{I \subseteq [n] \\ |I|=q}} \prod_{i \in I} (e^{p_i v} - 1) \right) dv$
- **Claim:** For all $n \geq k$, $\mathbb{E}[\nu_1(n, k)] \geq \mathbb{E}[\nu_1(n+1, k)]$
- **Claim:** If \mathcal{C} is not an MDS code, then $\mathbb{E}[\nu_1^p(n, k)] \leq \mathbb{E}[\nu_1^p(\mathcal{C})]$
- **Theorem:** For any p , $\mathbb{E}[\nu_1^p(n, k)] \geq \mathbb{E}[\nu_1(n, k)] \approx n \log\left(\frac{n}{n-k}\right)$
- **Theorem:** $\liminf\{\mathbb{E}[\nu_1(n, k)] : n \in \mathbb{N}\} = \begin{cases} k \log(e) & \text{If } \frac{k}{n} = \Theta(1) \\ k & \text{Otherwise} \end{cases}$

The MDS Coverage Depth Problem

The Noisy Channel ($t > 1$)

Assumptions:

- ✦ \mathcal{C} is an $[n, k]$ MDS code and \mathbf{p} is the uniform distribution
- ✦ Each strand \mathbf{x}_i can be retrieved given $t > 1$ samples
- ✦ $\sum_{q=0}^{k-1} \int_0^{\infty} [u^q] \prod_{i=1}^n (e_{t-1}(p_i v) + u (e^{p_i v} - e_{t-1}(p_i v))) e^{-v} dv$

Lemma: For any ϵ and n s.t. $n > e^{6t2^{t-1}/\epsilon} \geq 16$, it holds

$$P[v_t(n, k) \leq r(n, k, t)] \geq 1 - \epsilon$$

$$r(n, k, t) = n \log \left(\frac{n}{n-k} \right) + nt \log \log n + 2n \log(t+1)$$

Lemma: For any $c > 0$, it holds: $P \left[v_t(n, k) \leq n \log \left(\frac{n}{n-k} \right) - nc \right] \leq e^{-c} \left(\frac{n-k+1}{n-k} \right)$

The MDS Coverage Depth Problem

The Noisy Channel ($t > 1$)

Assumptions:

- ✦ \mathcal{C} is an $[n, k]$ MDS code and \mathbf{p} is the uniform distribution
- ✦ Each strand \mathbf{x}_i can be retrieved given $t > 1$ samples
- ✦ $\sum_{q=0}^{k-1} \int_0^\infty [u^q] \prod_{i=1}^n (e_{t-1}(p_i v) + u (e^{p_i v} - e_{t-1}(p_i v))) e^{-v} dv$

Theorem: For any ϵ and n large enough, it holds

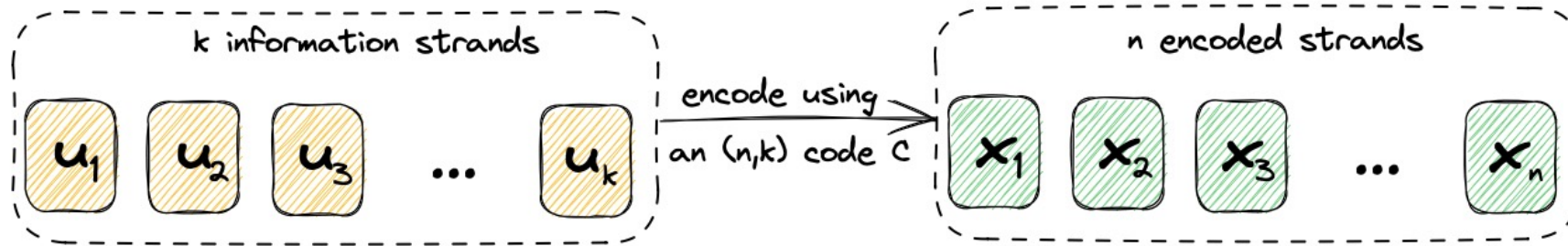
$$\log \left(\frac{1}{1-R} \right) + f_c(n, R) \leq \mathbb{E} \left[\frac{\nu_t(n, k)}{n} \right] \leq \left(\log \left(\frac{1}{1-R} \right) + t \log \log n + 2 \log(t+1) \right) \cdot (1 + 2\epsilon)$$

where $f_c(n, R) = \frac{1}{2n} \left(1 - \frac{1}{1-R} \right) - \sum_{h=1}^{\infty} \frac{B_{2h}}{2hn^{2h}} \left(1 - \frac{1}{(1-R)^{2h}} \right) = \mathcal{O}\left(\frac{1}{n^2}\right)$

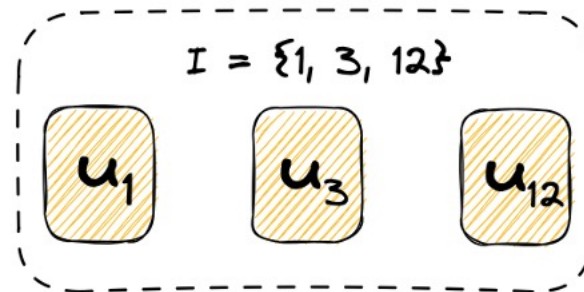
and B_h is the h -th Bernoulli number.

The Random Access Problem

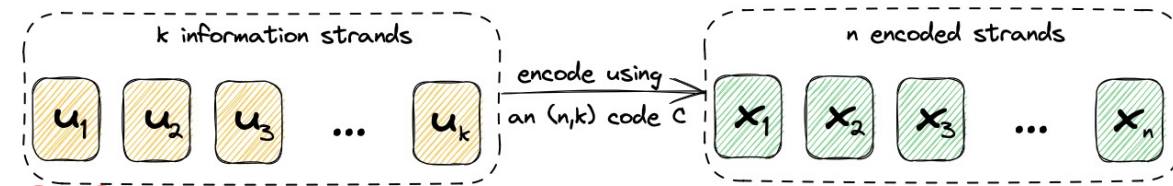
k information strands are encoded into n strands using some (n, k) code \mathcal{C}



The user wishes to retrieve a subset of the k information strands



We consider the singleton case, i.e., $|I| = 1$



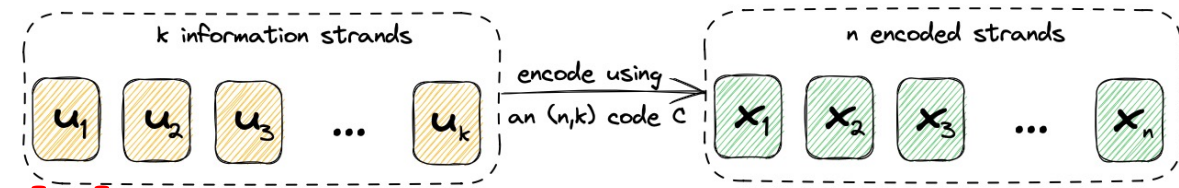
The Random Access Problem

Problem 3 - The singleton coverage depth problem

- \mathcal{C} - an (n, k) code
- $\tau_i(\mathcal{C})$ - r.v. for the number of samples to recover the i -th info. strand

- 1 Find the expectation value $\mathbb{E}[\tau_i(\mathcal{C})]$ and the probability distribution $P[\tau_i(\mathcal{C}) > r]$ for any $r \in \mathbb{N}$
- 2 Find the maximal expected number of samples to retrieve an information strand

$$T_{\max}^{\mathcal{C}} \triangleq \max_{1 \leq i \leq k} \mathbb{E}[\tau_i(\mathcal{C})]$$



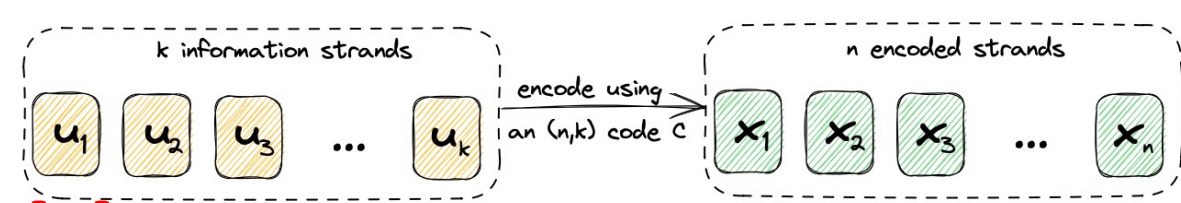
The Random Access Problem

Solve Problem 3 in case $n = k$ and no coding is used

Lemma: For $n \geq 1$ and $1 \leq i \leq n$, the following hold

① $\mathbb{E}[\tau_i] = n$ and $T_{\max} = n$

② For any $r \in \mathbb{N}$ we have that $P[\tau_i > r] = \left(1 - \frac{1}{n}\right)^r$ and $P[\tau_i = r] = \frac{1}{n} \cdot \left(1 - \frac{1}{n}\right)^{r-1}$



The Random Access Problem

Solve Problem 3 in case $n = k$ and no coding is used

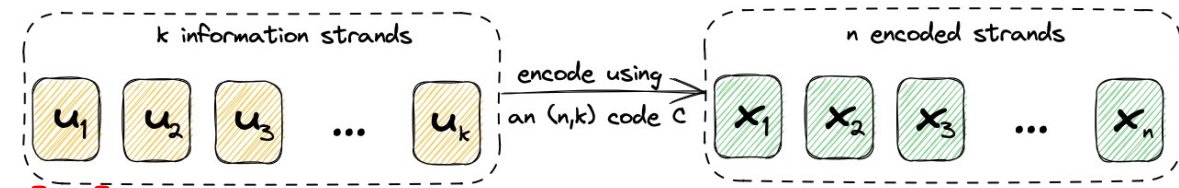
Lemma: For $n \geq 1$ and $1 \leq i \leq n$, the following hold

- ① $\mathbb{E}[\tau_i] = n$ and $T_{\max} = n$
- ② For any $r \in \mathbb{N}$ we have that $P[\tau_i > r] = \left(1 - \frac{1}{n}\right)^r$ and $P[\tau_i = r] = \frac{1}{n} \cdot \left(1 - \frac{1}{n}\right)^{r-1}$

Proof:

- ① τ_i has geometric distribution with success probability $p = \frac{1}{n}$. Hence,

$$T_{\max} = \max_{1 \leq i \leq k} \mathbb{E}[\tau_i] = p^{-1} = n$$



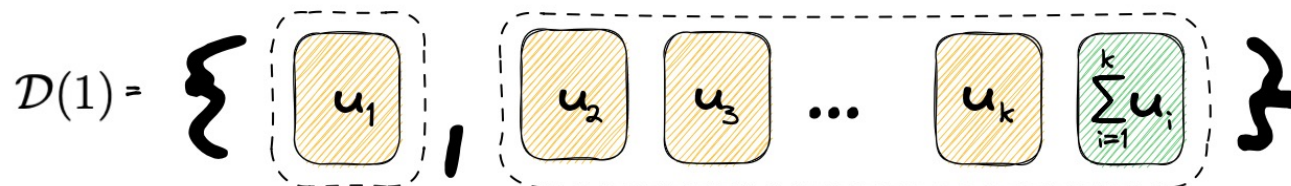
The Random Access Problem

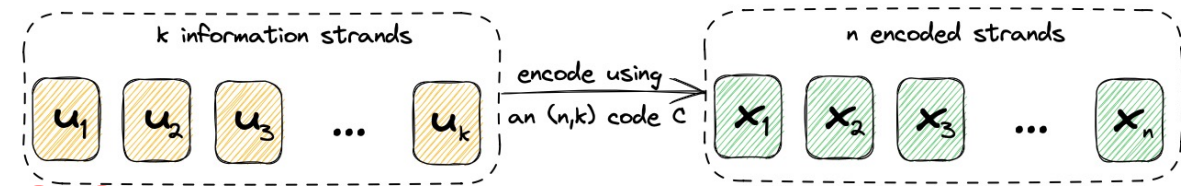
Definition: A set $J \subseteq [n]$ is a **retrieval set** of the i -th information strand, u_i , if it is possible to decode u_i from the encoded strands whose indices belong to J

$\widehat{\mathcal{D}}(i)$ - The set of all retrieval sets of u_i

$\mathcal{D}(i)$ - The set of all **minimal retrieval sets** of u_i (with respect to inclusion)

Example: For the $[k + 1, k]$ simple parity code:





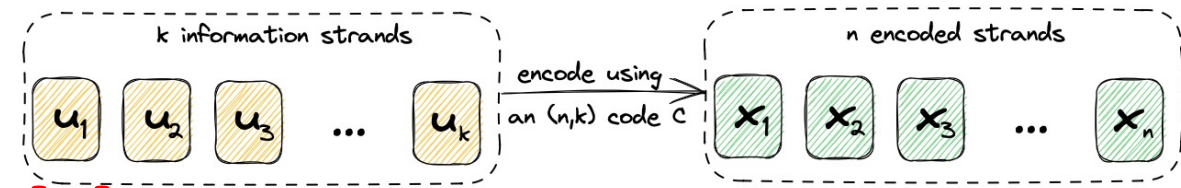
The Random Access Problem

Solve Problem 3 in case $n = k$

Claim: For any $(n = k, k)$ code \mathcal{C} it holds that $T_{\max}^{\mathcal{C}} \geq T_{\max} = n$.
 In particular, if ρ_i is the size of the smallest retrieval set of u_i , then

- 1 $\mathbb{E}[\tau_i(\mathcal{C})] = nH_{\rho_i}$
- 2 $T_{\max}^{\mathcal{C}} = nH_{\rho}$, where $\rho = \max_i \rho_i$

Observation: Since $n = k$, given any set of strands $\{x_i: i \in J\}$ we can recover at most $|J|$ information strands



The Random Access Problem

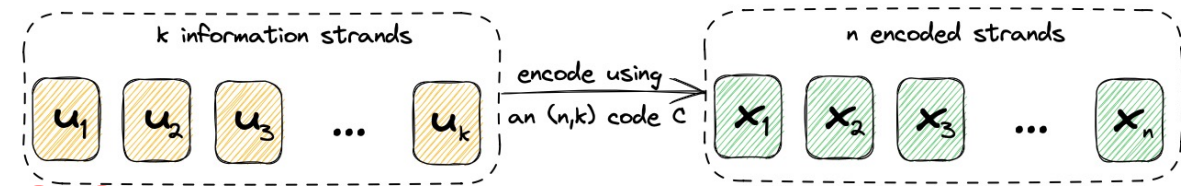
Theorem: For any (n, k) code \mathcal{C} , if $\mathcal{D}(i) = \{A, B\}$ for two disjoint retrieval sets $A \cap B = \emptyset$, then $\mathbb{E}[\tau_i(\mathcal{C})] = n \cdot (H_{|A|} + H_{|B|} - H_{|A|+|B|})$

Corollary 1: For any (n, k) code \mathcal{C} , if $\mathcal{D}(i) = \{A_1, \dots, A_v\}$ for mutually disjoint retrieval sets, then,

$$\mathbb{E}[\tau_i(\mathcal{C})] = n \cdot \sum_{s=1}^v (-1)^{s+1} \sum_{1 \leq j_1 < \dots < j_s \leq v} H_{|A_{j_1}| + \dots + |A_{j_s}|}$$

Corollary 2: For the $[n = k + 1, k]$ simple parity code:

$$\text{For any } i, T_{\max}^{\mathcal{C}} = \mathbb{E}[\tau_i(\mathcal{C})] = (k + 1) \cdot (H_1 + H_k - H_{k+1}) = k$$



The Random Access Problem

Question: Is it possible to have $T_{\max}^{\mathcal{C}} < k$?

- The identity code achieves $T_{\max}^{\mathcal{C}} = \mathbb{E}[\tau_i(\mathcal{C})] = k$
- The simple parity code achieves $T_{\max}^{\mathcal{C}} = \mathbb{E}[\tau_i(\mathcal{C})] = k$
- A non-systematic $[n, k]$ MDS code achieves $T_{\max}^{\mathcal{C}} \approx n \log \left(\frac{n}{n-k} \right) > k$
- What about systematic MDS codes...?

• **Theorem:** For any (n, k) MDS code \mathcal{C} , $k > n$, it holds $T_{\max}^{\mathcal{C}} = \mathbb{E}[\tau_i(\mathcal{C})] = k$

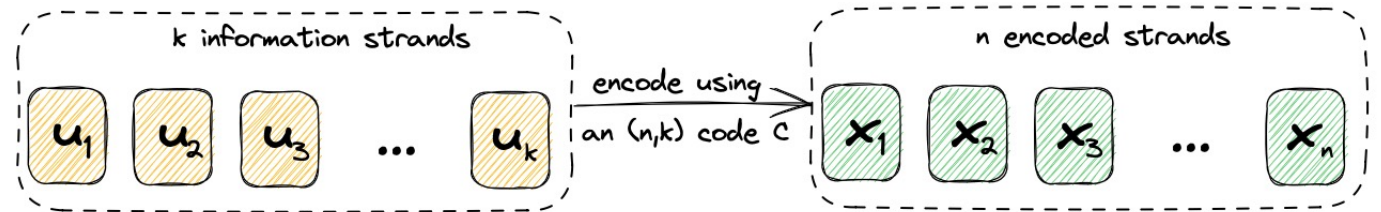
• **Lemma:** For the **Hamming** code \mathcal{C} , it holds $T_{\max}^{\mathcal{C}} = \mathbb{E}[\tau_i(\mathcal{C})] = k$

• **Lemma:** For the **Simplex** code \mathcal{C} , it holds $T_{\max}^{\mathcal{C}} = \mathbb{E}[\tau_i(\mathcal{C})] = k$

• **Lemma:** For the **Product** code \mathcal{C} , it holds $T_{\max}^{\mathcal{C}} = \mathbb{E}[\tau_i(\mathcal{C})] = k$

The Random Access Problem

k information strands are encoded into n strands using some (n, k) code \mathcal{C}
with a parity check matrix G

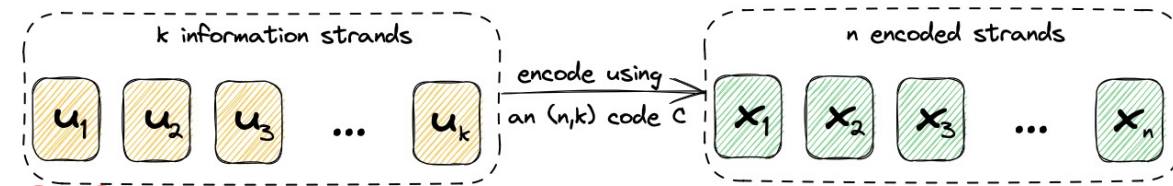


The user wishes to retrieve **one** of the k information strands

Problem 3' - The singleton coverage depth problem

- \mathcal{C} - an (n, k) code **with a parity check matrix G**
- $\tau_i(G)$ - r.v. for the number of **column samples from G** to decode the **i -th unit vector e_i**
- Find the maximal expected number of samples to retrieve any unit vectors

$$T_{\max}^G \triangleq \max_{1 \leq i \leq k} \mathbb{E}[\tau_i(G)]$$



The Random Access Problem

Problem 3' - The singleton coverage depth problem

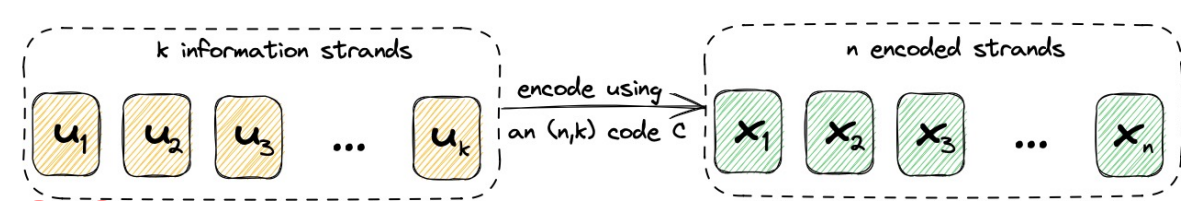
- \mathcal{C} - an (n, k) code **with a parity check matrix G**
- $\tau_i(G)$ - r.v. for the number of **column samples from G** to decode the **i -th unit vector e_i**
- Find the maximal expected number of samples to retrieve any unit vector

$$T_{\max}^G \triangleq \max_{1 \leq i \leq k} \mathbb{E}[\tau_i(G)]$$

Example:

- $\mathcal{C}: (x_1, x_2) \rightarrow (x_1, x_2, x_1, x_2, x_1 + x_2)$
- $\mathbb{E}[\tau_1(G)] = \mathbb{E}[\tau_2(G)] = 1.917 < 2$

$$G = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$



The Random Access Problem

- **Theorem:** Given a parity check matrix G of a code \mathcal{C} ,

$$\text{let } \alpha_i(s) = |\{S \subseteq [n] : |S| = s, e_i \in \langle g_j : j \in S \rangle\}|.$$

$$\text{Then, } E[\tau_i(G)] = nH_n - \sum_{s=1}^{n-1} \frac{\alpha_i(s)}{\binom{n-1}{s}}.$$

- **Example:**

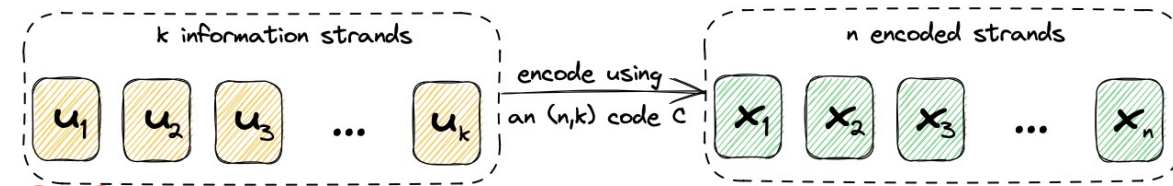
- $\mathcal{C}: (x_1, x_2) \rightarrow (x_1, x_2, x_1, x_2, x_1 + x_2)$

- $\mathbb{E}[\tau_1(G)] = \mathbb{E}[\tau_2(G)] = 1.917 < 2$

$$G = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

$$\alpha_1(1) = 2, \alpha_1(2) = 9, \alpha_1(3) = \binom{5}{3}, \alpha_1(4) = \binom{5}{4}$$

$$\mathbb{E}[\tau_1(G)] = 5H_5 - \sum_{s=1}^4 \frac{\alpha_1(s)}{\binom{4}{s}} = \frac{23}{12} \approx 1.917$$



The Random Access Problem

- **Theorem:** Given a parity check matrix G of a code \mathcal{C} ,

$$\text{let } \alpha_i(s) = |\{S \subseteq [n] : |S| = s, e_i \in \langle g_j : j \in S \rangle\}|.$$

$$\text{Then, } E[\tau_i(G)] = nH_n - \sum_{s=1}^{n-1} \frac{\alpha_i(s)}{\binom{n-1}{s}}.$$

- **Example:** Assume \mathcal{C} is an MDS code with a systematic generator matrix G .

$$\alpha_i(s) = \begin{cases} \binom{n-1}{s-1} & \text{if } s \in [k-1] \\ \binom{n}{s} & \text{if } s \geq k. \end{cases}$$

$$\mathbb{E}[\tau_i(G)] = nH_n - \sum_{s=1}^{k-1} \frac{\binom{n-1}{s-1}}{\binom{n-1}{s}} - \sum_{s=k}^{n-1} \frac{\binom{n}{s}}{\binom{n-1}{s}} = nH_n - \sum_{s=1}^{k-1} \frac{s}{n-s} - \sum_{s=k}^{n-1} \frac{n}{n-s} = k$$

The Average Expectation

- $\tilde{\tau}_i(G)$ - r.v. counting the number of drawn columns of G until the i th column of G is recovered.

$$G = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

- **Theorem:** $\sum_{i=1}^n E[\tilde{\tau}_i(G)] = kn.$

- A code \mathcal{C} is called recovery balanced if $E[\tilde{\tau}_1(G)] = \dots = E[\tilde{\tau}_n(G)].$

- **Corollary:** If G is a systematic generator matrix of a recovery balanced code \mathcal{C} , then $E[\tilde{\tau}_i(G)] = k$ for $i \in [n]$ and $T_{\max}^{\mathcal{C}} = k.$

- For a systematic MDS code \mathcal{C} with systematic generator matrix G , it holds $E[\tilde{\tau}_i(G)] = k$ for $i \in [n]$ and $T_{\max}^{\mathcal{C}} = k.$

Breaking the Balance of MDS Codes

- **Theorem:** Let $G = (I_k | R)$ be a systematic generator matrix of an MDS code. For $x \geq 1$, let $G^x = (I_k | \cdots | I_k | R)$ (x copies of the identity matrix). Then,

$$T_{\max}(G^x) = 1 + \sum_{s=1}^{N-1} \frac{\binom{N-x}{s}}{\binom{N-1}{s}} - \sum_{s=k}^{N-1} \sum_{a=0}^{k-1} \frac{\binom{k-1}{a}}{\binom{N-1}{s}}$$

$$\sum_{m=0}^{s-k} \binom{n-k}{s-a-m} \sum_{t=0}^a (-1)^t \binom{a}{t} \binom{(a-t)x}{m+a}.$$

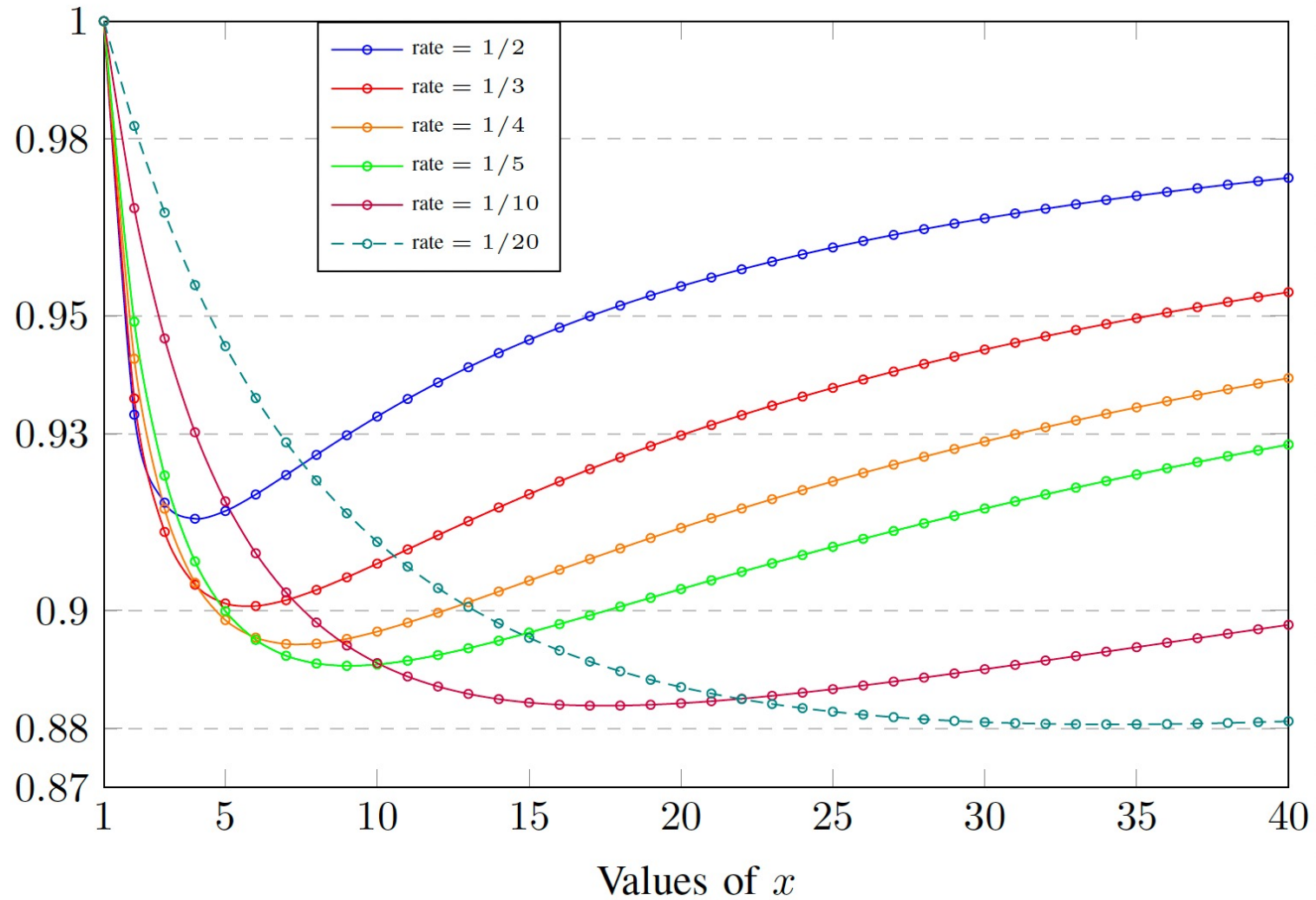
- **Example:**

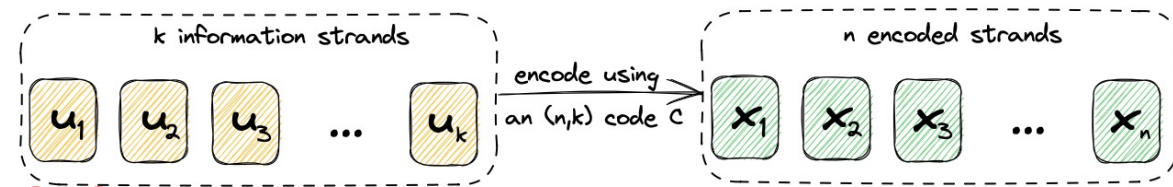
- $\mathcal{C}: (x_1, x_2) \rightarrow (x_1, x_2, x_1, x_2, x_1 + x_2)$

- $\mathbb{E}[\tau_1(G)] = \mathbb{E}[\tau_2(G)] = 1.917 < 2$

$$G = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Breaking the Balance of MDS Codes





The Random Access Problem

Question: Is it possible to have $T_{\max}^{\mathcal{C}} < k$?

Example

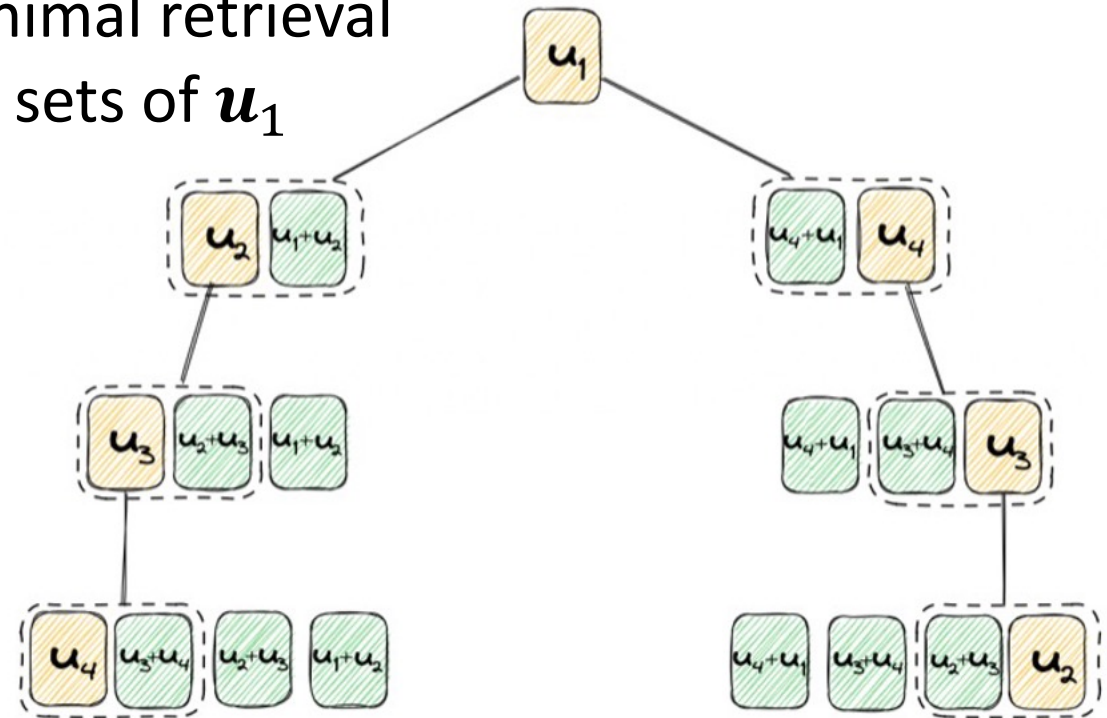
Information word



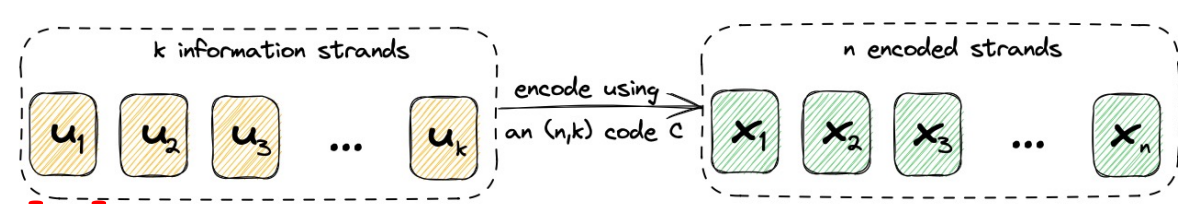
Codeword



Minimal retrieval sets of u_1



$$\mathbb{E}[\tau_1(\mathcal{C})] = \sum_{r=1}^{\infty} P[\mathcal{T}_1^{\mathcal{C}} \geq r] = \frac{403}{105} \approx 3.838.$$



The Random Access Problem

Question: Is it possible to have $T_{\max}^{\mathcal{C}} < k$?

Example

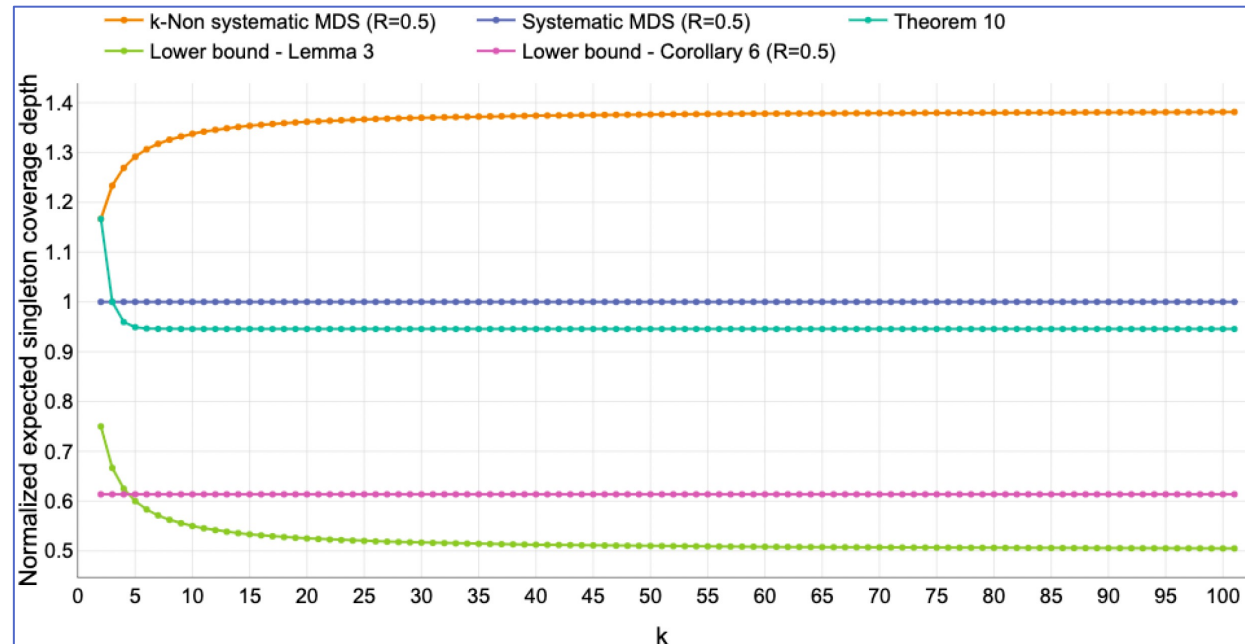
$$\mathbb{E}[\tau_j(\mathcal{C}_{(2k,k)})] = 1 + \sum_{i=1}^{2k-3} B(k, i) \cdot \frac{2k}{(2k-i) \binom{2k}{i}}$$

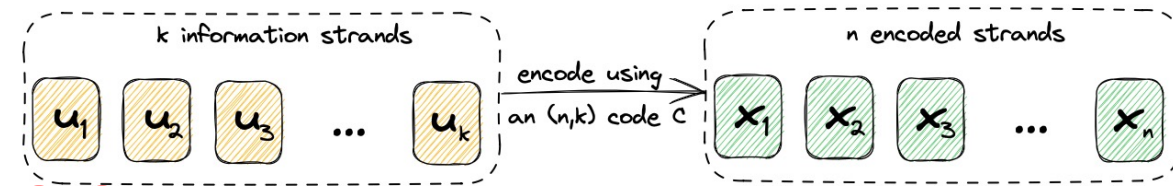
$$B(k, i) = \begin{cases} \binom{2k-1}{i} + 2B(k-1, i-1) - B(k-2, i-2) & k \geq 2, i \geq 2 \\ 1 & k \geq 0, i = 0 \\ 2k + 1 & k \geq 0, i = 1 \\ 1 & k = 1, i = 2 \\ 0 & k = 0, i \geq 2 \\ 0 & k = 1, i \geq 3 \end{cases}$$

Information word



Codeword





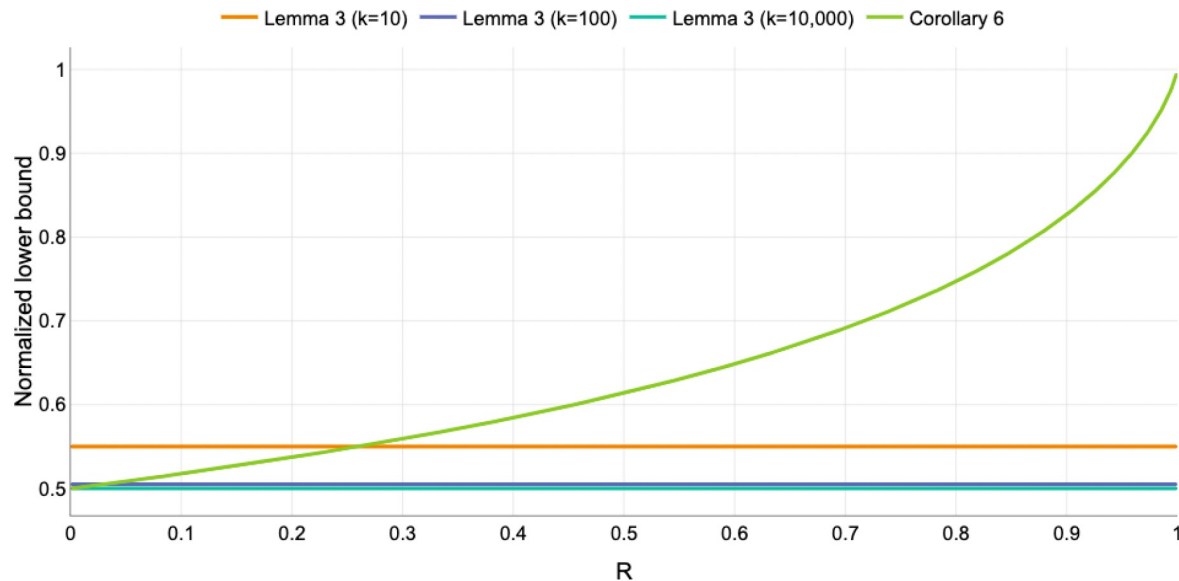
The Random Access Problem

- **Theorem:** There exists an $(n, 2)$ code \mathcal{C} s.t. $T_{\max}^{\mathcal{C}} = 1.83 = 0.914 \cdot 2$.
There exists an $(n, 3)$ code \mathcal{C} s.t. $T_{\max}^{\mathcal{C}} = 2.67 = 0.89 \cdot 3$.
- For an (n, k) code \mathcal{C} , \mathcal{C}^γ is the $(\gamma n, \gamma k)$ code consisting of γ copies of \mathcal{C} .
- **Theorem:** $T_{\max}^{\mathcal{C}^\gamma} = \gamma T_{\max}^{\mathcal{C}}$
- **Corollary:** There exists an $(\gamma n, 2\gamma)$ code \mathcal{C} s.t. $T_{\max}^{\mathcal{C}} = 0.914 \cdot 2\gamma$.
There exists an $(\gamma n, 3\gamma)$ code \mathcal{C} s.t. $T_{\max}^{\mathcal{C}} = 0.89 \cdot 3\gamma$.

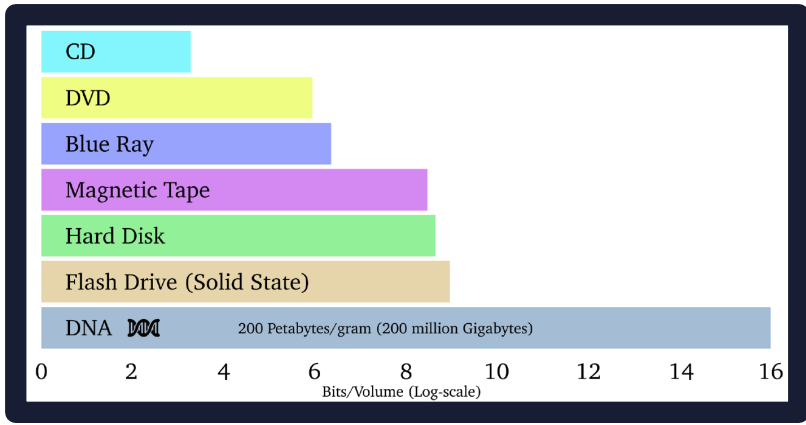
Lower Bounds

Theorem: For any (n, k) code \mathcal{C} , it holds: $T_{\max}^{\mathcal{C}} \geq \frac{k+1}{2}$

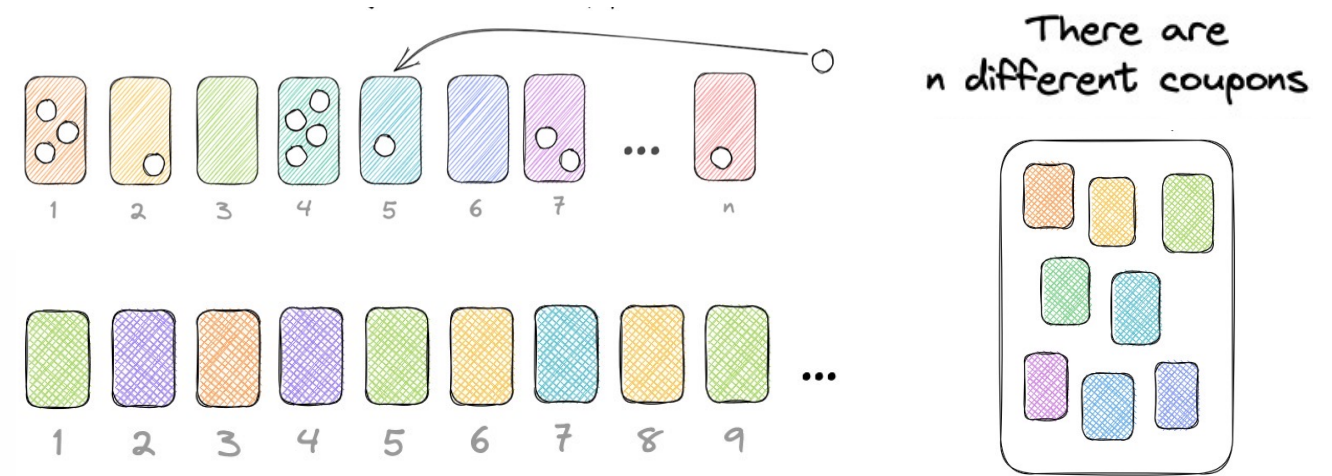
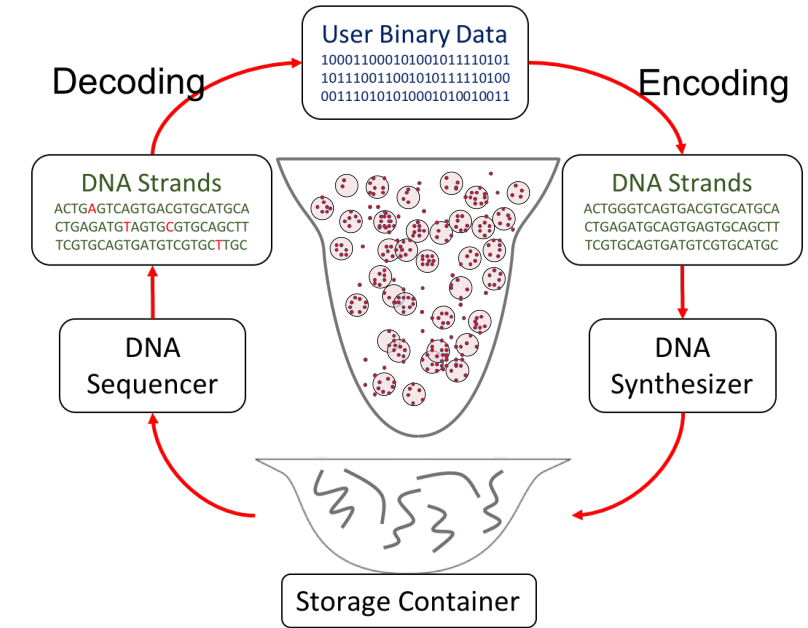
Theorem: For any (n, k) code \mathcal{C} , it holds: $T_{\max}^{\mathcal{C}} \geq \frac{n}{k} \cdot \sum_{i=0}^k \frac{k-i}{n-i} = n - \frac{n(n-k)}{k} \cdot (H_n - H_{n-k})$
 $= k \left(\frac{1}{R} + \frac{1-R}{R^2} \cdot \log(1-R) \right)$



Summary



- The DNA storage channel
- The coverage depth problem
- The random access problem
- Many interesting open problems...



Funded by the European Union

Coding Theory and Algorithms for DNA-based Data Storage

Call for Contributions

SUNDAY, JULY 7, 2024

ATHENS, GREECE

The workshop will focus on coding theory and algorithms for DNA-based data storage. It will consist of invited and contributed talks, as well as poster presentations, from researchers and experts. The workshop is organized as a satellite workshop of the 2024 IEEE International Symposium on Information Theory (ISIT2024).

- Jointly organized with Dave Landsman from the DNA Data Storage Alliance.
- Contribution deadline: April 15, 2024.
- Designed to foster collaboration.