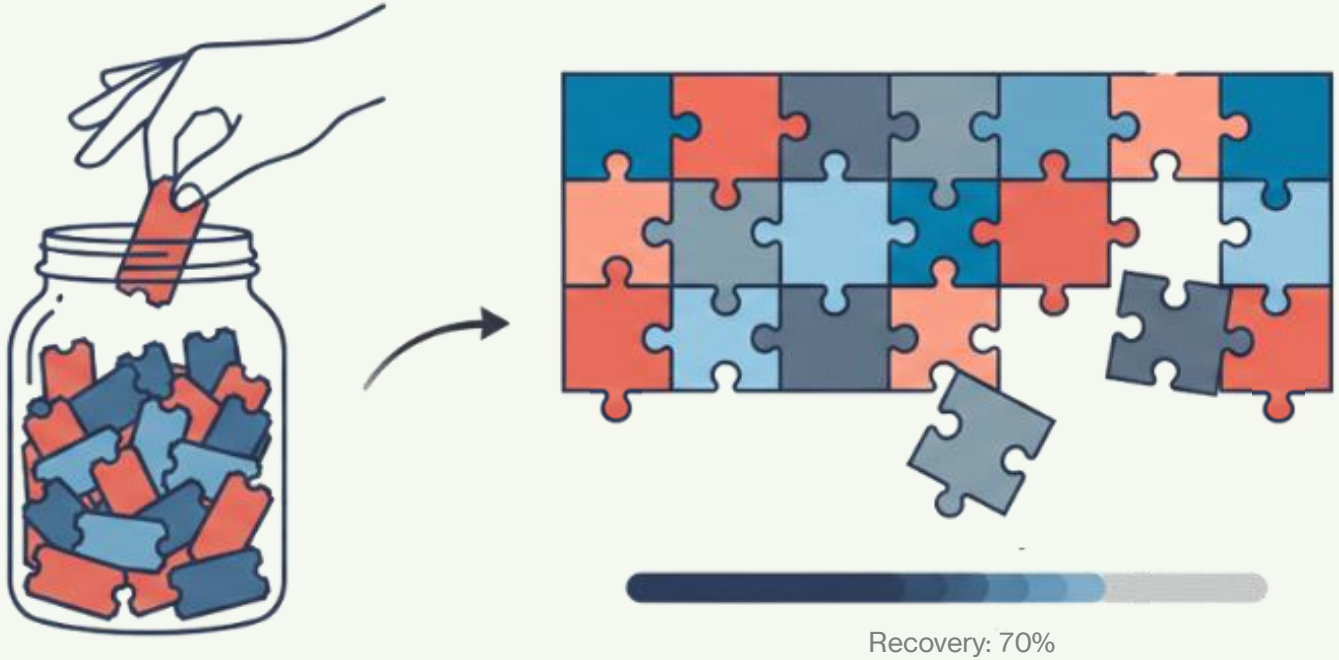
Universität Zürich[UZH]

The Coded Coupons Collector's Problem

Minimizing Sampling Overheads in Probabilistic Information Retrieval



Daniella Bar-Lev (Universität Zürich)

Joint work with: Omer Sabary, Anina Gruica, Ryan Gabrys, Alberto Ravagnani, Ohad Elishco, Eitan Yaakobi

IZS 2026 | February 25, 2026

180 Zettabytes

Global datasphere projection by 2025. Magnetic storage cannot keep pace.

Motivation

DNA-Based Data Storage

The Problem

180 Zettabytes

Global datasphere projection by 2025. Magnetic storage cannot keep pace.

The Solution

Density: Millions of times denser than SSDs

Durability: Lasts thousands of years

Maintenance: Zero power required at rest

The Challenge

Noisy: High error rates
From $< 1\%$ to 25% indels

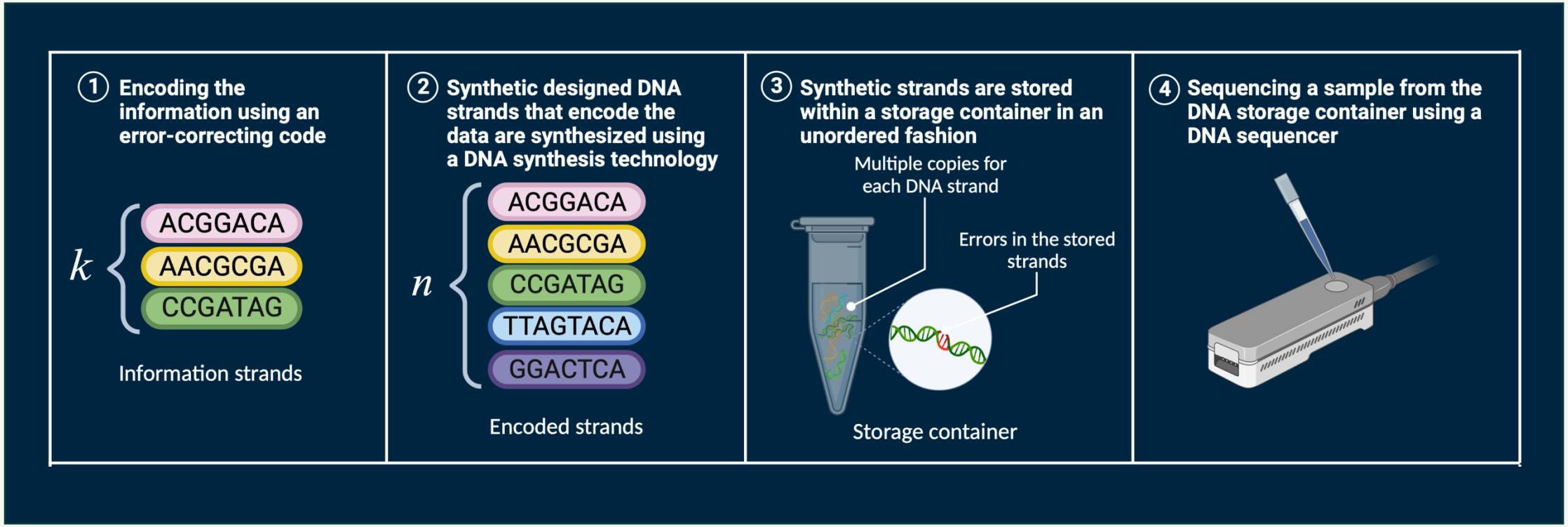
Slow: High r/w latency.
Hours to several days
Unordered data

Expensive: High r/w costs.
Writing: $1\text{MB} = \$4.2\text{K}$
Reading: Approx. $1GB = \$120$

- *MinION:* 50M strands $= \$1\text{K}$
- *Hiseq:* 200M strands $= \$2.5\text{K}$

Motivation

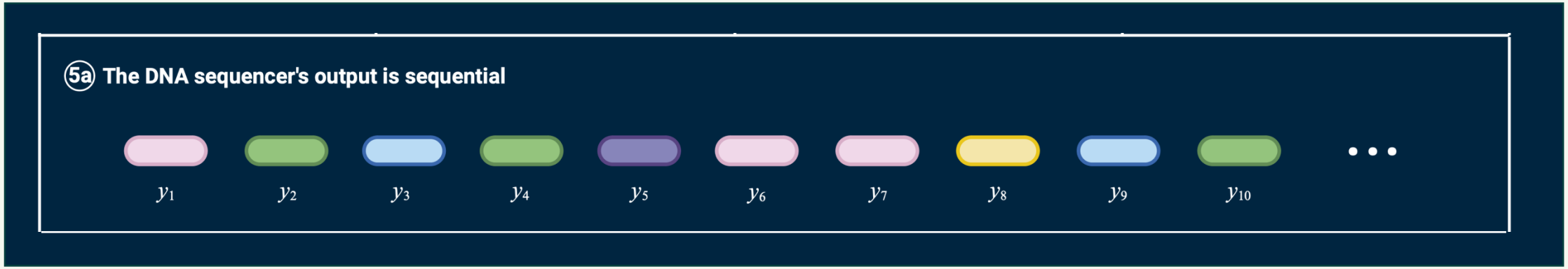
DNA-Based Data Storage: Channel Model



D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

Motivation

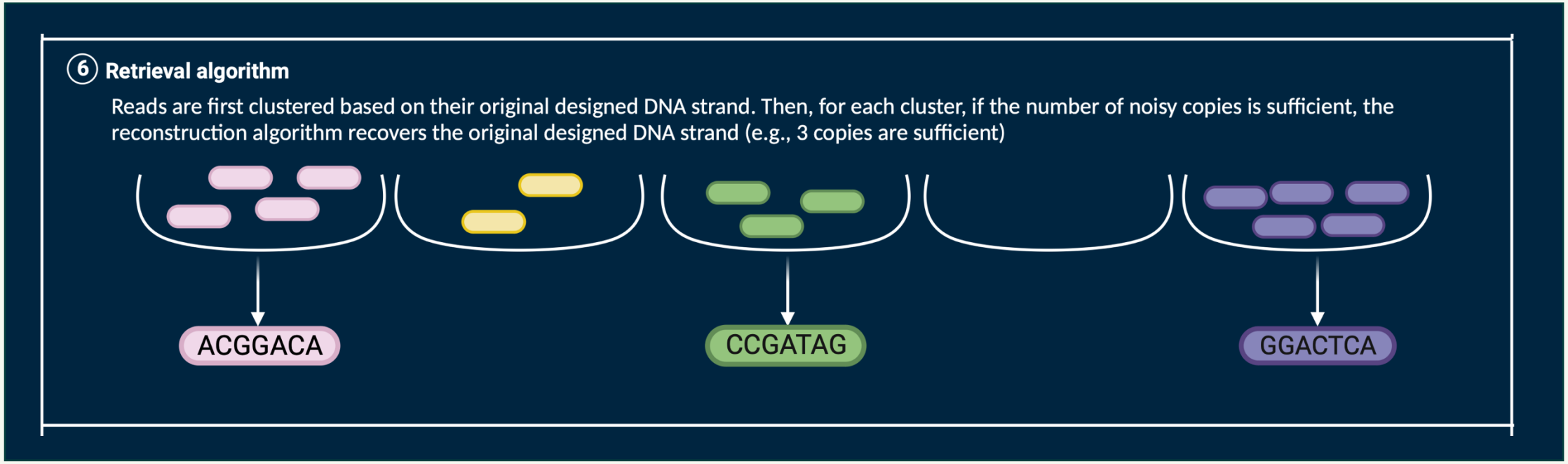
DNA-Based Data Storage: Channel Model



D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, “Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems” *IEEE TIT*, 2025.

Motivation

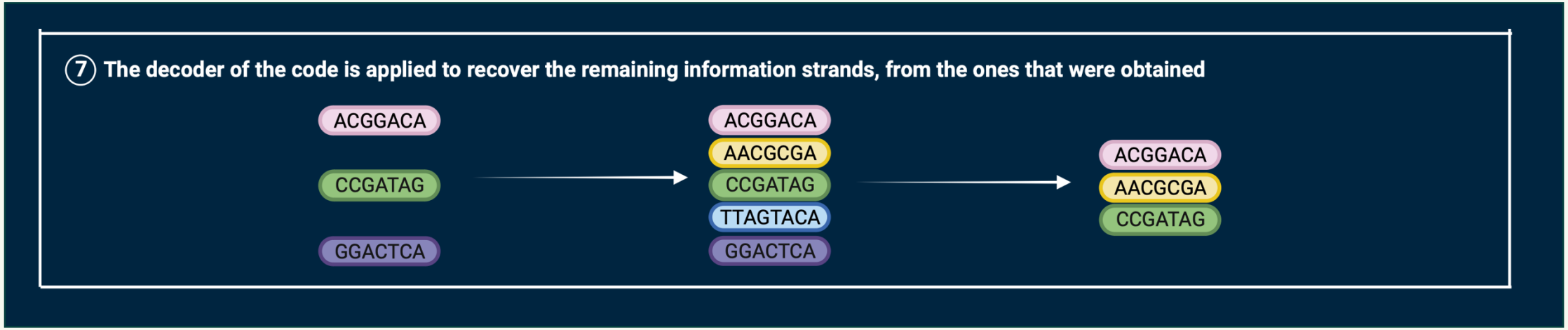
DNA-Based Data Storage: Channel Model



D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

Motivation

DNA-Based Data Storage: Channel Model



Goal: encode the information in a way that minimizes the expected number of reads required for successful retrieval to **reduce latency and cost**

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

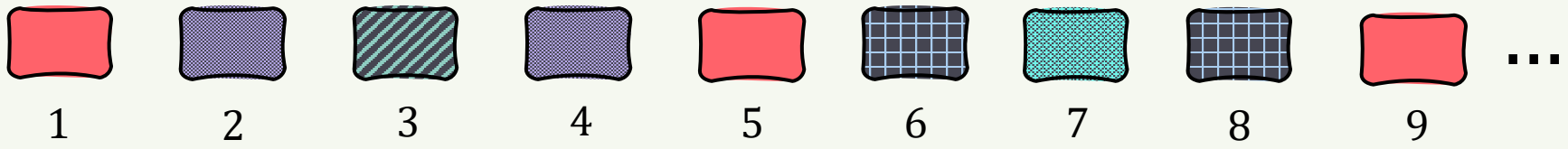
Background

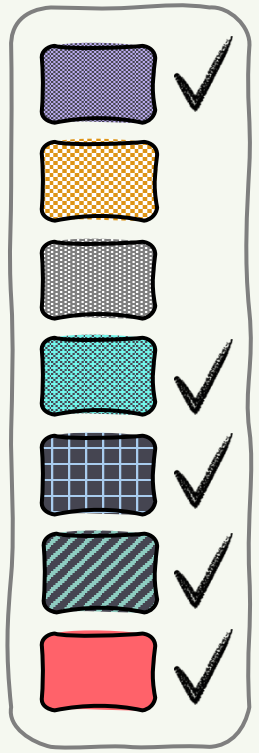
The Coupon's Collector Problem

The Coupon's Collector Problem (First studied by **Feller** in 1967)

Each box of cereal contains one out of n coupons.

How many cereal boxes one should expect to buy to collect all n coupons?





Solution: T: #draws, t_i: time to collect the i-th new coupon

$$T = t_1 + t_2 + t_3 + \cdots + t_n$$

Each t_i is a geometric random variable

$$E[T] = E[t_1] + E[t_2] + \cdots + E[t_n] = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1}$$

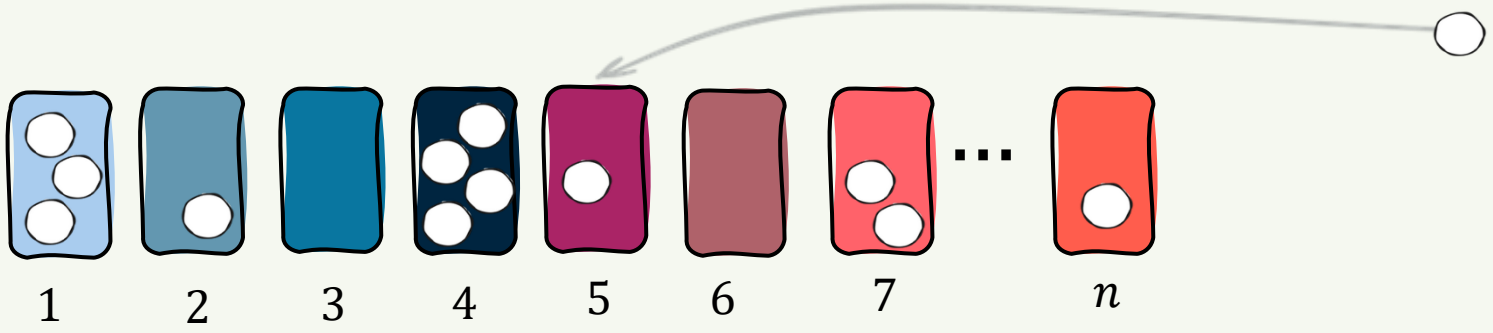
$$= n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + \frac{1}{1}\right) = nH_n = n\log(n) + \gamma n + 0.5 + O\left(\frac{1}{n}\right),$$

$\gamma \approx 0.57$ (Euler-Mascheroni const.)

Background

The Coupon's Collector Problem

The Dixie Cup Problem/The Urn Problem (First studied by **Newman** in 1960 and later by **Flajolet et al.** in 1992.)
Identical balls are thrown into n urns randomly.
What is the expected number of thrown balls that is needed to have at least t balls in each urn?

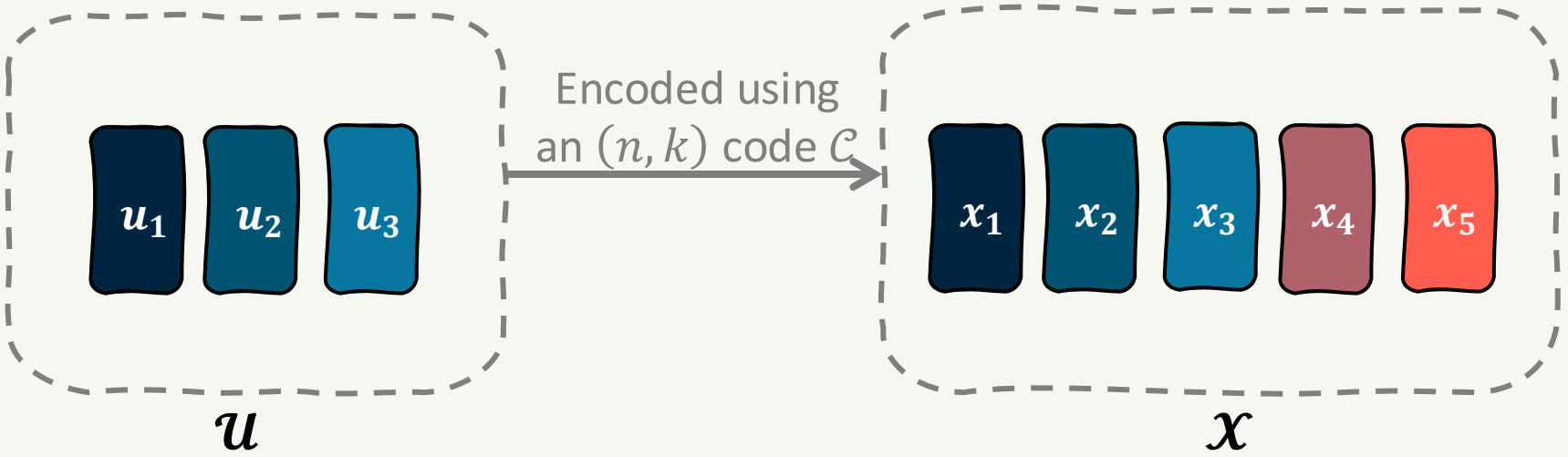


Other extensions

- 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 Coded Coupon's Collector Problem

Problem Setup



Main goal: Study the expected number of samples M to guarantee successful decoding of $\boldsymbol{u}$ and find codes that minimize this number.

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Full Recovery: Problem Definition & Main Results

Problem 1 [The MDS coverage depth problem]: For any k, n, and $\boldsymbol{p}$ find:

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

Uniform is optimal + upper and lower bounds on $\mathbb{E}[v_t(n,k)/n]$ + more…

Problem 2 [The coding coverage depth problem]: For any k find:

1. Given n and $\boldsymbol{p}$ find an (n,k) code $\mathcal{C}$ that minimizes value $\mathbb{E}[\nu_t^{\boldsymbol{p}}(\mathcal{C})]$
2. The minimum value of $\mathbb{E}[\nu_t^{\boldsymbol{p}}(\mathcal{C})]$ over all possible codes $\mathcal{C}$, $\boldsymbol{p}$. That is, find

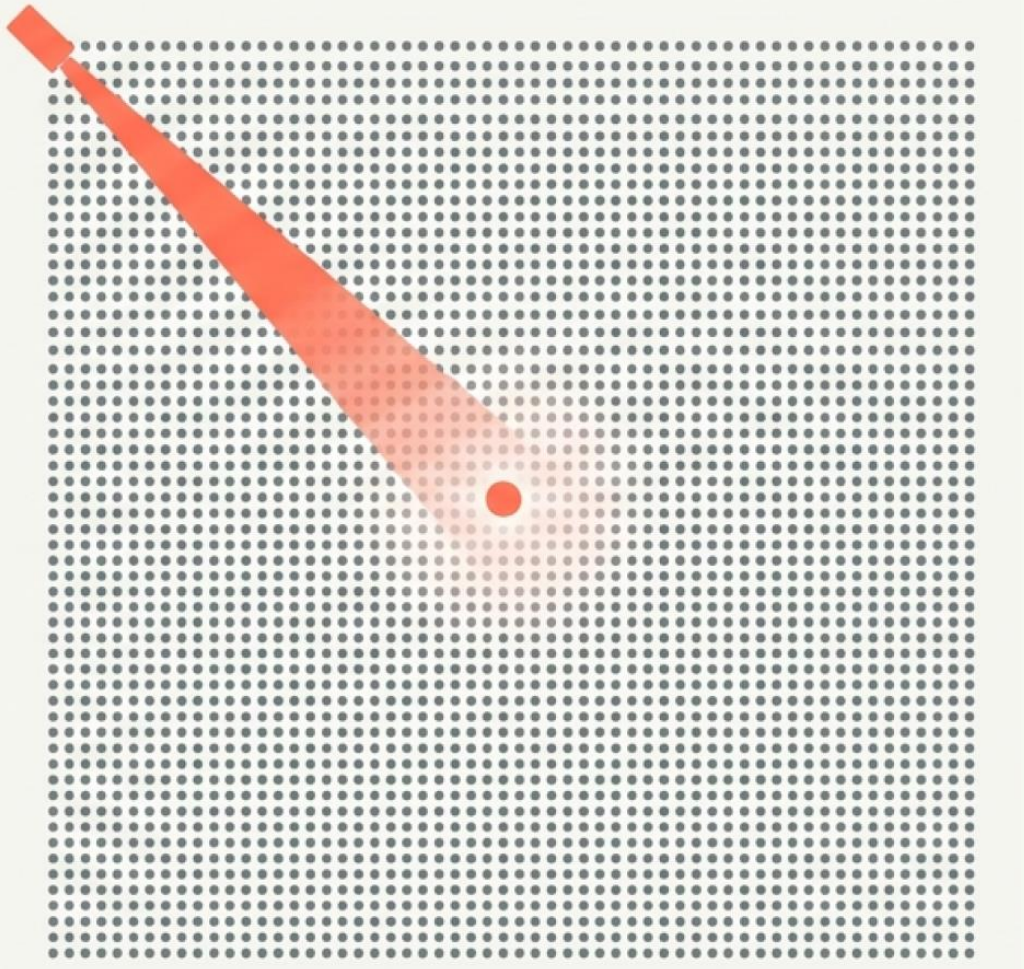
$$M^{\mathrm{opt}}(k) \triangleq \liminf_{\mathcal{C},\boldsymbol{p}} \left\{ \mathbb{E}\left[\nu_t^{\boldsymbol{p}}(\mathcal{C})\right] \right\}$$

We fully solves Problem 2 for the noiseless channel with uniform distribution.

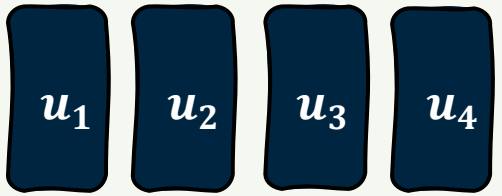
D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

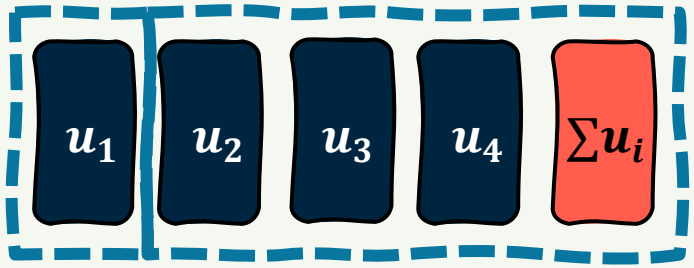
Random Access



Users rarely retrieve an entire exabyte archive. They usually need **specific files**



$\mathbb{E}[\tau_1] = n = k = 4$

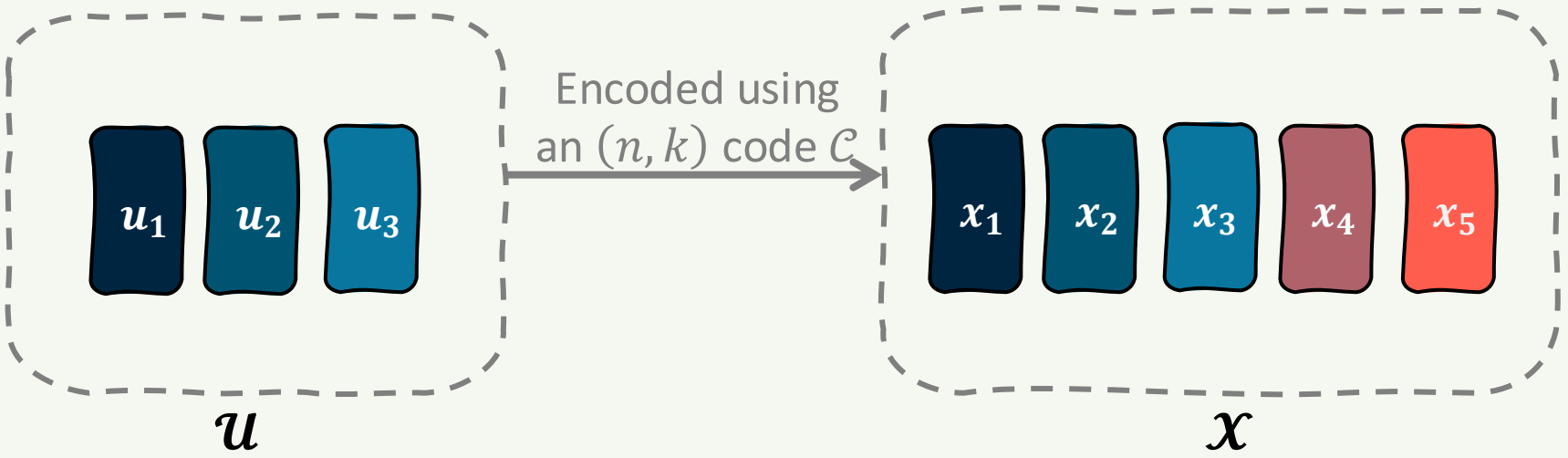


$\mathbb{E}[\tau_1] = ?$

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.
A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Problem Definition



The user **wishes to retrieve a single information symbol** out of the k information symbols

$\forall i \in [k]$: $\tau_i(\mathcal{C})$ - r.v. of the #samples for successful decoding of the i-th information symbol (assuming that $\boldsymbol{p}$ is the uniform and the channel is noiseless, i.e., $t = 1$)

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Problem Definition

Problem 3 [The random access coverage depth problem]: Given an (n, k) code $\mathcal{C}$ find:

1. The expectation value $\mathbb{E}[\tau_i(\mathcal{C})]$ and the prob. distribution $\mathrm{P}[\tau_i(\mathcal{C}) > r]$, for any $r \in \mathbb{N}$.
2. The maximal expected number of samples to retrieve an information symbols, i.e.,

$$T_{\max}(\mathcal{C}) \triangleq \max_{i \in [k]} \mathbb{E}[\tau_i(\mathcal{C})] \,.$$

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.
A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

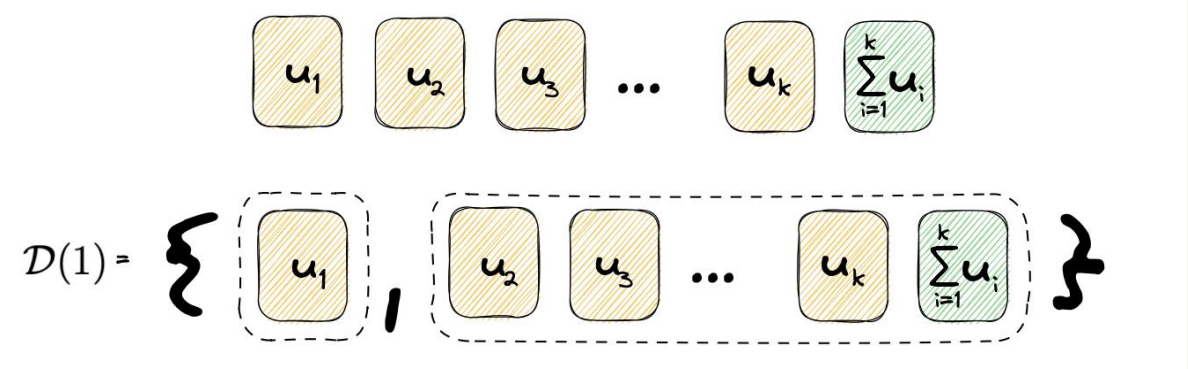
The Coded Coupon's Collector Problem

Random Access: Retrieval Sets

Definition: A set $J \subseteq [n]$ is a **retrieval set** of the i-th information symbols, $\boldsymbol{u}_i$, if it is possible to decode $\boldsymbol{u}_i$ from the encoded strands whose indices belong to J.

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

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



D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Retrieval Sets

Theorem: For any (n,k) code $\mathcal{C}$, if $\mathcal{D}(i)=\{A,B\}$ for two disjoint retrieval sets $A\cap B=\emptyset$, then

$$\mathbb{E}[\tau_i(\mathcal{C})]=n\cdot\left(H_{|A|}+H_{|B|}-H_{|A|+|B|}\right)$$

Proof: Denote

$\lambda_J(r-1)$- # ways to draw $r-1$ symbols s.t. $\exists j\in J$ for which the j-th strand wasn't drawn

$\lambda(r-1)$- # ways to draw $r-1$ symbols s.t. $\boldsymbol{u}_i$ cannot be retrieved from the drawn symbols

We have that $\lambda_{A\cup B}(r-1)=\lambda_A(r-1)+\lambda_B(r-1)-\lambda(r-1)$.

$$\lambda_A(r-1)=\sum_{j=1}^{|A|}\binom{|A|}{j}(-1)^{j+1}(n-j)^{r-1}$$

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Retrieval Sets

Theorem: For any (n,k) code $\mathcal{C}$, if $\mathcal{D}(i)=\{A,B\}$ for two disjoint retrieval sets $A\cap B=\emptyset$, then

$$\mathbb{E}[\tau_i(\mathcal{C})]=n\cdot\left(H_{|A|}+H_{|B|}-H_{|A|+|B|}\right)$$

Proof: Denote

$\lambda_J(r-1)$- # ways to draw $r-1$ symbols s.t. $\exists j\in J$ for which the j-th strand wasn't drawn

$\lambda(r-1)$- # ways to draw $r-1$ symbols s.t. $\boldsymbol{u}_i$ cannot be retrieved from the drawn symbols

We have that $\lambda_{A\cup B}(r-1)=\lambda_A(r-1)+\lambda_B(r-1)-\lambda(r-1)$.

$$\mathbb{E}[\tau_i(\mathcal{C})]=\sum_{r=1}^{\infty}\frac{\lambda(r-1)}{n^{r-1}}=\sum_{r=1}^{\infty}\sum_{j=1}^{|A|}\frac{\binom{|A|}{j}(-1)^{j+1}(n-j)^{r-1}}{n^{r-1}}+\sum_{r=1}^{\infty}\sum_{j=1}^{|B|}\frac{\binom{|B|}{j}(-1)^{j+1}(n-j)^{r-1}}{n^{r-1}}-\sum_{r=1}^{\infty}\sum_{j=1}^{|A|+|B|}\frac{\binom{|A|+|B|}{j}(-1)^{j+1}(n-j)^{r-1}}{n^{r-1}}$$

$$=n\cdot\left(H_{|A|}+H_{|B|}-H_{|A|+|B|}\right)$$

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

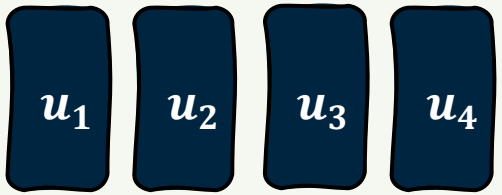
The Coded Coupon's Collector Problem

Random Access: Retrieval Sets

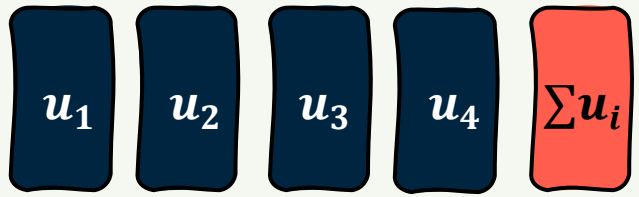
Theorem: For any (n, k) code $\mathcal{C}$, if $\mathcal{D}(i) = \{A, B\}$ for two disjoint retrieval sets $A \cap B = \emptyset$, then

$$\mathbb{E}[\tau_i(\mathcal{C})] = n \cdot \left(H_{|A|} + H_{|B|} - H_{|A|+|B|}\right)$$

Corollary: Assume $\mathcal{C}$ is the $[n = k + 1, k]$ simple parity code.
Then, for any i, $T^{\mathcal{C}}_{\max} = \mathbb{E}[\tau_i(\mathcal{C})] = n(H_1 + H_k - H_{k+1}) = n\left(1 - \frac{1}{k+1}\right) = k + 1 - \frac{k+1}{k+1} = k$.



$\mathbb{E}[\tau_1] = n = k = 4$



$\mathbb{E}[\tau_1] = \mathbf{4}$

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Alternative Definition

Problem 3 [The random access coverage depth problem]: Given an (n, k) code $\mathcal{C}$ find:

1. The expectation value $\mathbb{E}[\tau_i(\mathcal{C})]$ and the prob. distribution $\mathrm{P}[\tau_i(\mathcal{C}) > r]$, for any $r \in \mathbb{N}$.
2. The maximal expected number of samples to retrieve an information symbols, i.e.,

$$T_{\max}(\mathcal{C}) \triangleq \max_{i \in [k]} \mathbb{E}[\tau_i(\mathcal{C})] .$$

Problem 3 [The random access coverage depth problem]: Given an (n, k) code with a generator matrix $G \in \mathbb{F}_q^{k \times n}$, let $\tau_i(G)$ denote the random variable that counts the minimum number of columns of G that are drawn until the standard basis vector e_i is in their $\mathbb{F}_q$-span. Find:

1. The expectation value $\mathbb{E}[\tau_i(G)]$ and the prob. distribution $\mathrm{P}[\tau_i(G) > r]$, for any $r \in \mathbb{N}$.
2. The maximal expected number of samples to retrieve an information symbols, i.e.,

$$T_{\max}(G) \triangleq \max_{i \in [k]} \mathbb{E}[\tau_i(G)] .$$

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.
A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Alternative Definition

Problem 3 [The random access coverage depth problem]: Given an (n, k) code with a generator matrix $G \in \mathbb{F}_q^{k\times n}$, let $\tau_i(G)$ denote the random variable that counts the minimum number of columns of G that are drawn until the standard basis vector e_i is in their $\mathbb{F}_q$-span. Find:

1. The expectation value $\mathbb{E}[\tau_i(G)]$ and the prob. distribution $\mathrm{P}[\tau_i(G) > r]$, for any $r \in \mathbb{N}$.
2. The maximal expected number of samples to retrieve an information symbols, i.e.,

$$T_{\max}(G) \triangleq \max_{i\in[k]} \mathbb{E}[\tau_i(G)] \,.$$

Definition: Denote the j-th column of G by g_j. For $i \in [k]$ and $1 \le s \le n$, let

$$\alpha_i(s) \coloneqq \left|\left\{S \subseteq [n]\colon\ |S| = s,\ \ e_i \in < g_j\colon\ j \in S >\right\}\right|$$

Lemma: For $G \in \mathbb{F}_q^{k\times n}$ and for all $i \in [k]$ we have

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

A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Alternative Definition

Definition: Denote the j-th column of G by g_j. For $i \in [k]$ and $1 \leq s \leq n$, let

$$\alpha_i(s) := \left|\left\{S \subseteq [n]:\ |S| = s,\ \ e_i \in < g_j:\ j \in S >\right\}\right|$$

Lemma: For $G \in \mathbb{F}_q^{k\times n}$ and for all $i \in [k]$ we have

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

Example: Let $G = \begin{pmatrix} 10101 \\ 01011 \end{pmatrix} \in \mathbb{F}_2^{2\times 5}$ we have

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

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

A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Alternative Definition

Definition: Denote the j-th column of G by g_j. For $i \in [k]$ and $1 \leq s \leq n$, let

$$\alpha_i(s) \coloneqq \left|\left\{S \subseteq [n]:\ |S| = s,\ \ e_i \in < g_j:\ j \in S >\right\}\right|$$

Lemma: For $G \in \mathbb{F}_q^{k\times n}$ and for all $i \in [k]$ we have

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

Corollary: Let $G \in \mathbb{F}_q^{k\times n}$ be a systematic generator matrix of an MDS code. For all $i \in [k]$ we have

$$T_{\max}(G) = \ \mathbb{E}[\tau_i(G)] = k$$

A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Alternative Definition

Definition: For $i \in [n]$, we let $\widetilde{\tau}_i$ denote the random variable that counts the minimum number of columns of G that are drawn until the i-th column of G belongs to their $\mathbb{F}_q$-span.

Observation: If $G \in \mathbb{F}_q^{k\times n}$ is systematic, then for all $i \in [k]$ we have $\tau_i(s) = \widetilde{\tau}_i(s)$.

Lemma: For a code $\mathcal{C}$ with generator matrices $G, G' \in \mathbb{F}_q^{k\times n}$, for all $i \in [n]$ we have

$$\mathbb{E}[\widetilde{\tau}_i(G)] = \mathbb{E}[\widetilde{\tau}_i(G')].$$

Theorem : Let $\mathcal{C} \subseteq \mathbb{F}_q^n$ be a code of dimension k with generator matrix G. We have that

$$\sum_{i=1}^{n} \mathbb{E}[\widetilde{\tau}_i(G)] = nk.$$

A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Alternative Definition

Theorem : Let $\mathcal{C} \subseteq \mathbb{F}_q^n$ be a code of dimension k with generator matrix G. We have that

$$\sum_{i=1}^{n} \mathbb{E}[\widetilde{\tau_i}(G)] = nk.$$

Definition: Let $\mathcal{C}$ be a code with generator matrix G. We call $\mathcal{C}$ a recovery balanced code if

$$\mathbb{E}[\widetilde{\tau_1}(G)] = \mathbb{E}[\widetilde{\tau_2}(G)] = \cdots = \mathbb{E}[\widetilde{\tau_n}(G)]$$

Corollary : Let $\mathcal{C} \subseteq \mathbb{F}_q^n$ be a recovery balanced code with a systematic generator matrix G. For all $i \in [k]$ we have

$$T_{\max}(G) = \mathbb{E}[\tau_i(G)] = k$$

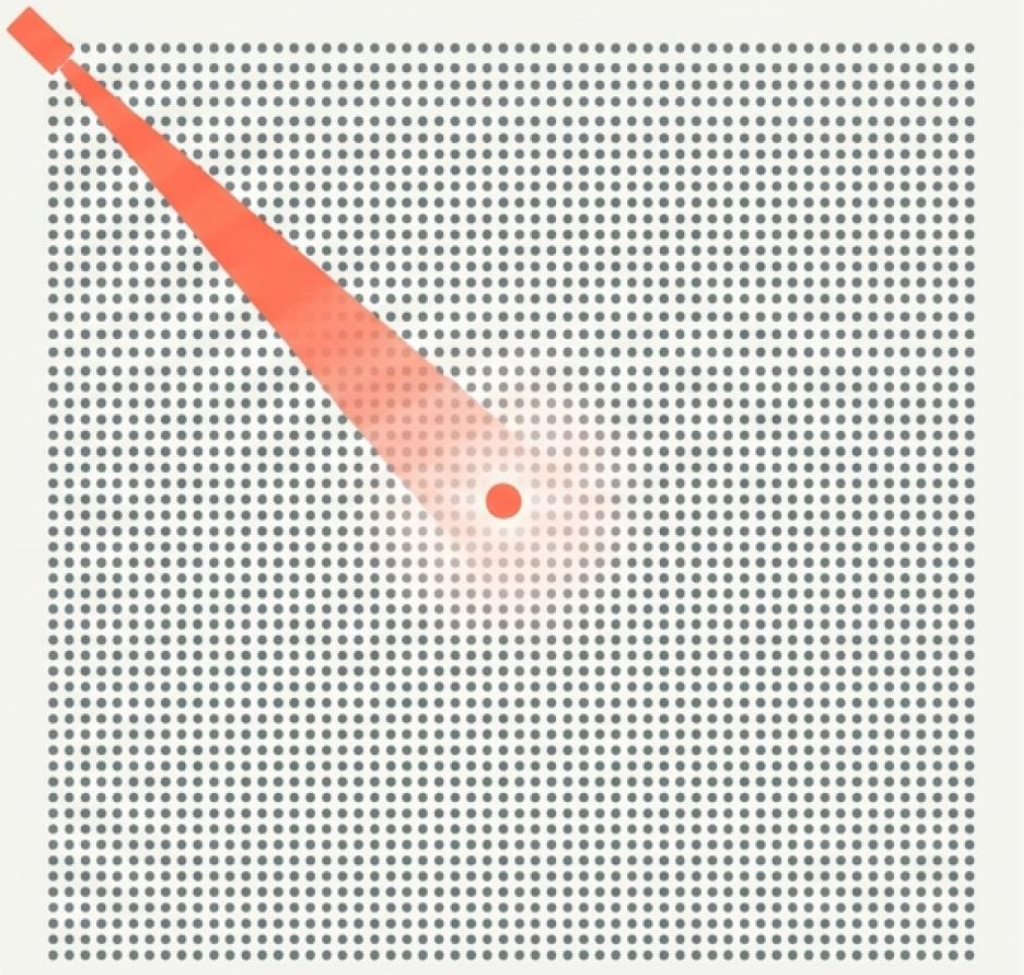
In the paper we give three different sufficient conditions for a code to be recovery balanced

Conjecture: A code $\mathcal{C}$ is recovery balanced if and only if its dual code $\mathcal{C}^{\perp}$ is recovery balanced.

A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Retrieval Sets



Users rarely retrieve an entire exabyte archive. They usually need **specific files**

Selected results:

- Identity code: $\mathbb{E}[\tau_i(\mathcal{C})] = k$
- Simple parity code $\mathbb{E}[\tau_i(\mathcal{C})] = k$
- Non-systematic $[n, k]$ MDS code achieves $\mathbb{E}[\tau_i(\mathcal{C})] > k$
- Systematic $[n, k]$ MDS code $\mathbb{E}[\tau_i(\mathcal{C})] = k$
- Simplex code: $\mathbb{E}[\tau_i(\mathcal{C})] = k.$
- Hamming code: $\mathbb{E}[\tau_i(\mathcal{C})] = k.$

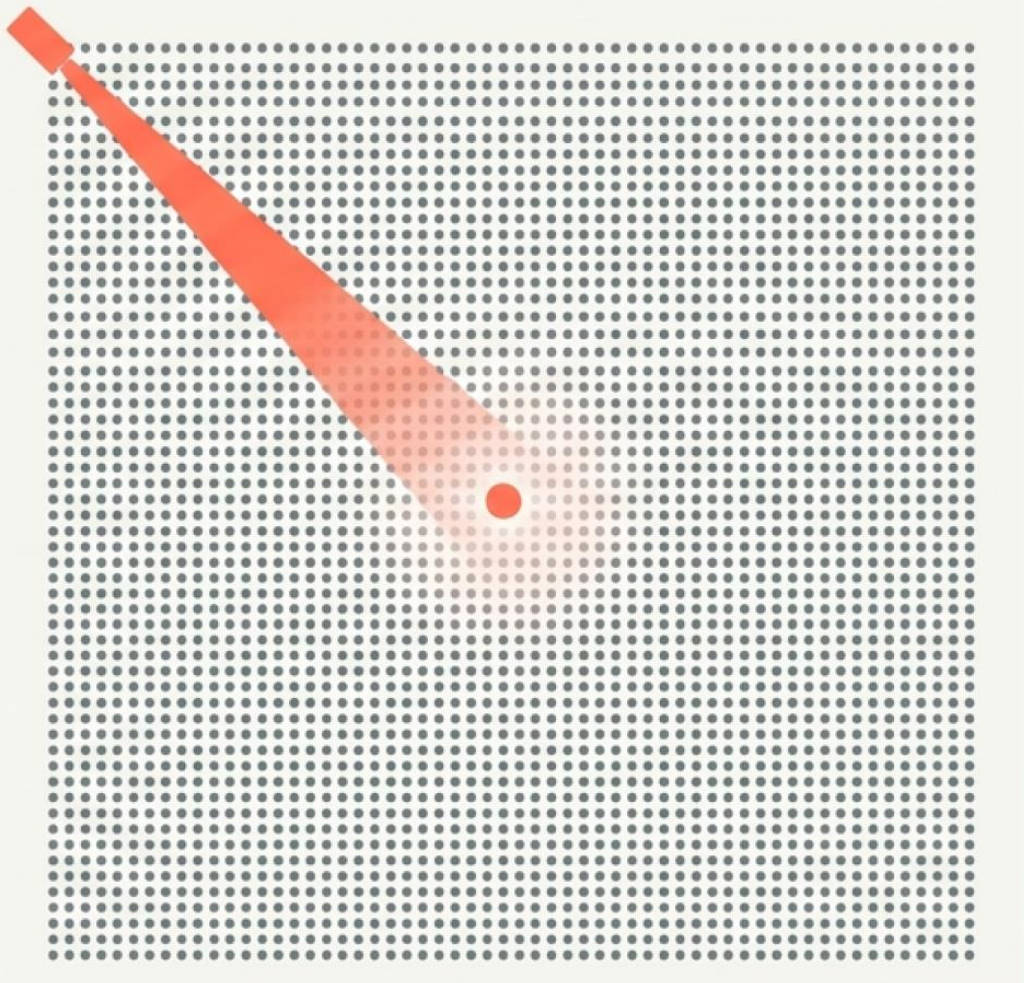
} Gruica et al.

- Several more codes all achieve $\mathbb{E}[\tau_i(\mathcal{C})] \geq k$

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.
A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

The Coded Coupon's Collector Problem

Random Access: Retrieval Sets



Users rarely retrieve an entire exabyte archive. They usually need **specific files**

Question: Is it possible to have $\max_{1 \le i \le k} \mathbb{E}[\tau_i(\mathcal{C})] < k$?

Answer: YES!

New explicit construction allows us to beat this baseline, enabling targeted retrieval with minimal overhead

D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems" *IEEE TIT*, 2025.
A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025.

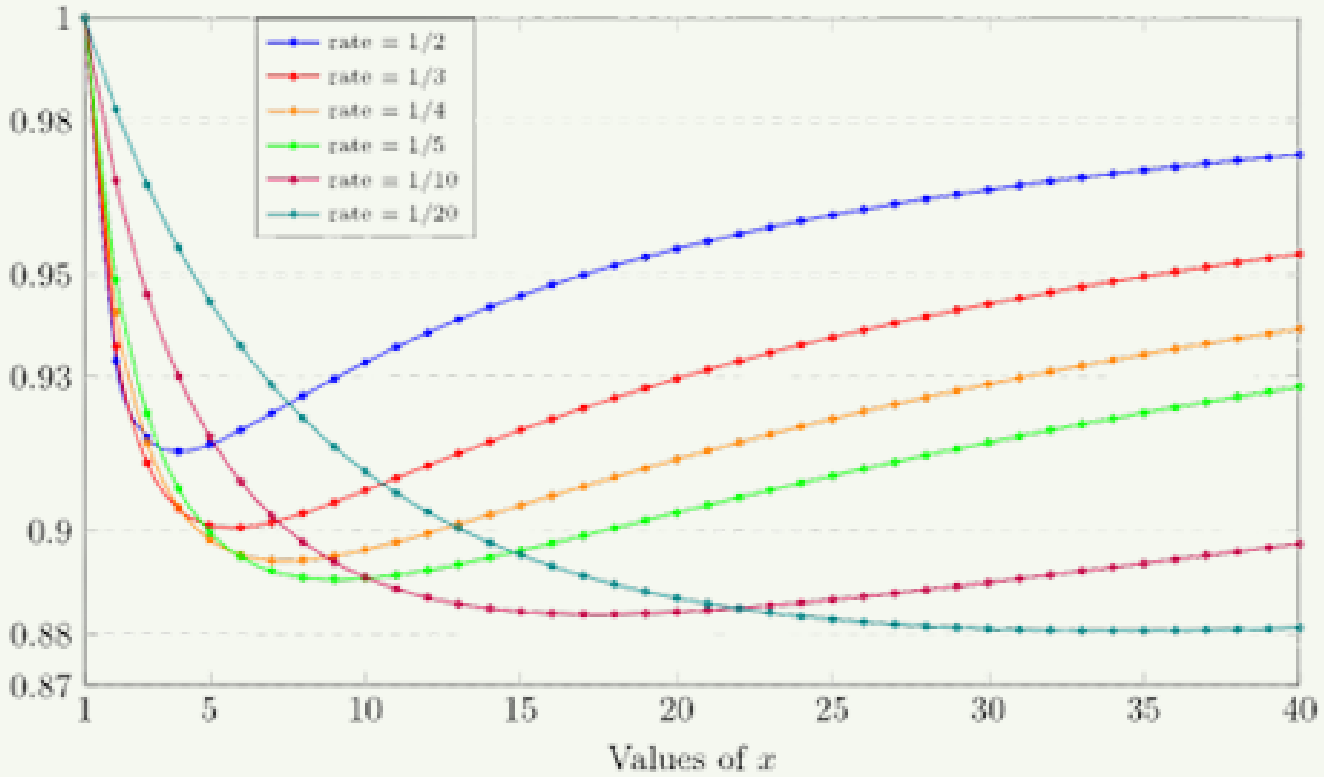
The Coded Coupon's Collector Problem

Random Access: Breaking the Balance

$$G = (I_k|R) \in \mathbb{F}_q^{k\times n}$$
$$G^x = (\underbrace{I_k|I_k|\ldots|I_k}_{x \text{ times}}|R) \in \mathbb{F}_q^{k\times N}$$

$$G = \begin{pmatrix}101\\011\end{pmatrix} \qquad G^2 = \begin{pmatrix}10101\\01011\end{pmatrix}$$

$$\mathbb{E}[\tau_i(G)] = 2 \qquad \mathbb{E}[\tau_i(G^2)] = \frac{23}{12}$$



Normalized random access coverage depth $T_{\max}(G^x)$ from for $k = 5$ and various rates

A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage." *IEEE TIT*, 2025..

The Coded Coupon's Collector Problem

Related Works

Classical CCP

[ER61] Erdős and Rényi, "On a classical problem of probability theory," Publ. Math. Inst. Hung. Acad. Sci., Ser. A, 1961.

[Fel67] Feller, "An introduction to probability theory and its applications," Wiley, 1967.

[FGT67] Flajolet, Gardy, and Thimonier, "Birthday paradox, coupon collectors, caching algorithms and self-organizing search," Discrete Applied Mathematics, 1992.

[New60] Newman, "The Double Dixie Cup Problem," The American Mathematical Monthly, 1960.

Full Recovery

[CTL+19] Chandak et al., "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes," Allerton, 2019.

[PGYA24] Preuss, Galili, Yakhini, and Anavy, "Sequencing coverage analysis for combinatorial DNA-based storage systems," IEEE Transactions on Molecular, Biological, and Multi-Scale Communications, 2024.

[SAC+24] Sokolovskii,. Agarwal, Croquevielle, Zhou, and Heinis, "Coding over coupon collector channels for combinatorial motif-based DNA storage. IEEE Transactions on Communications," 2024.

[Han25] Hanna, "On the Reliability of Information Retrieval From MDS Coded Data in DNA Storage", IEEE ISIT, 2025.

[CC25] Cao and Chen, "Optimizing Sequencing Coverage Depth in DNA Storage: Insights From DNA Storage Data," IEEE ISIT, 2025.

[BRY25] Bertuzzo, Ravagnani, and Yaakobi, "The Coverage Depth Problem in DNA Storage Over Small Alphabets," IEEE ISIT, 2025.

Random Access

[GMZ24] Gruica. Montanucci, and Zullo, " The Geometry of Codes for Random Access in DNA Storage," arXiv preprint, 2024..

[BEG+25] Boruchovsky, Elishco, Gabrys, Gruica, Tamo, and Yaakobi, "Making it to First: The Random Access Problem in DNA Storage," arXiv preprint, 2025.

[BLL+25] Bodur, Lia, López, Ludhani, Ravagnani, and Seccia "The Random Variables of the DNA Coverage Depth Problem", arXiv preprint, 2025.

[WY26] Wang and Yaakobi, "Random Access in DNA Storage: Algorithms, Constructions, and Bounds," arXiv preprint, 2026

And more…

Thank you!