An Information Theory for Out-of-Order Information:
DNA Data Storage and Beyond

Ilan Shomorony
University of Illinois at Urbana-Champaign

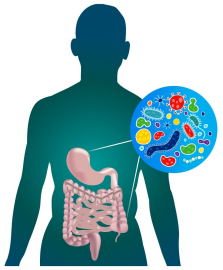
Stanford ISL Colloquium – 03/16/23

The wonders of high-throughput sequencing

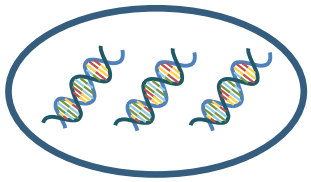
Genome sequencing

RNA-seq

Metagenomics



DNA storage



high-throughput sequencer

TCGCGATT TTCATTCG
CGCATCGA GCGCGATT
TTCTTCGC GATCATTC
ACGCATTC

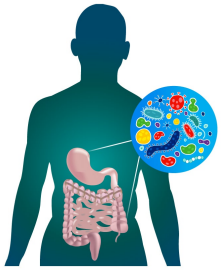
- No location/ordering information!
- Reordering problem based on:
  - Reference (side information)
  - Overlaps
  - Coding

The wonders of high-throughput sequencing

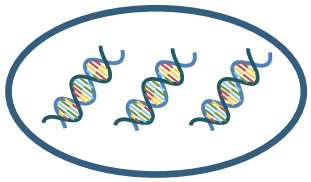
Genome sequencing

RNA-seq

Metagenomics



DNA storage



high-throughput sequencer

TCGCGATT TTCATTCG
CGCATCGA GCGCGATT
TTCTTCGC GATCATTC
ACGCATTC

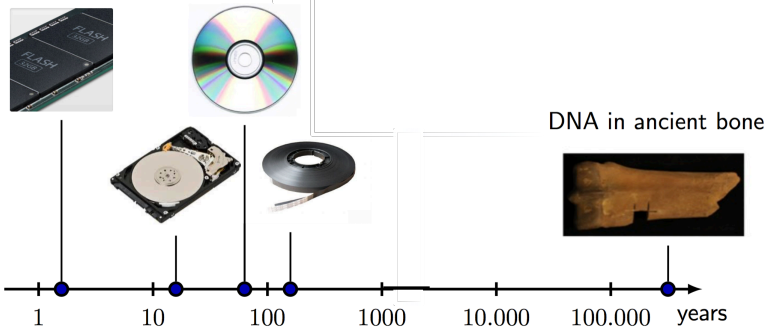
- No location/ordering information!

Outline:

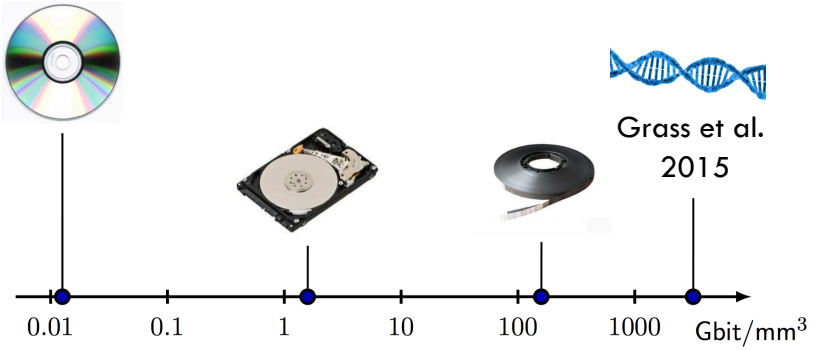
1. Information-theoretic limits of out-of-order channels
2. Efficiently finding overlaps

Why store data in DNA?

Longevity

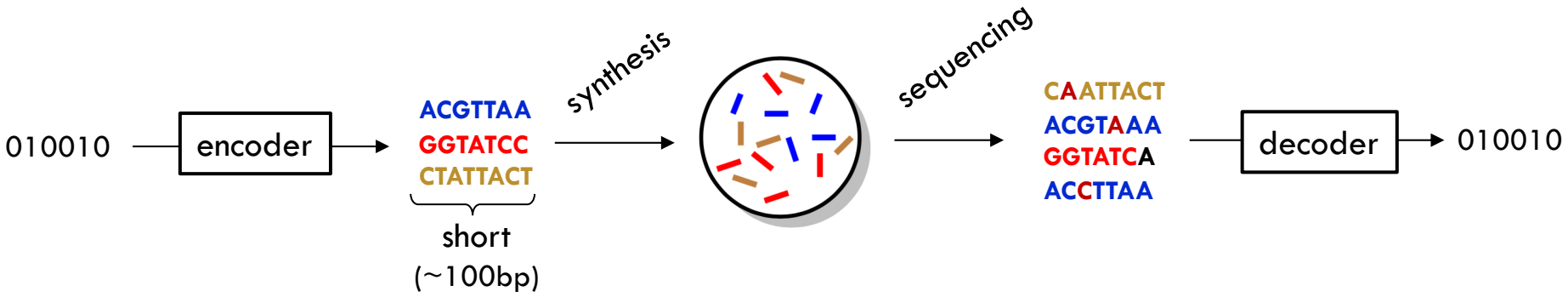


High information density



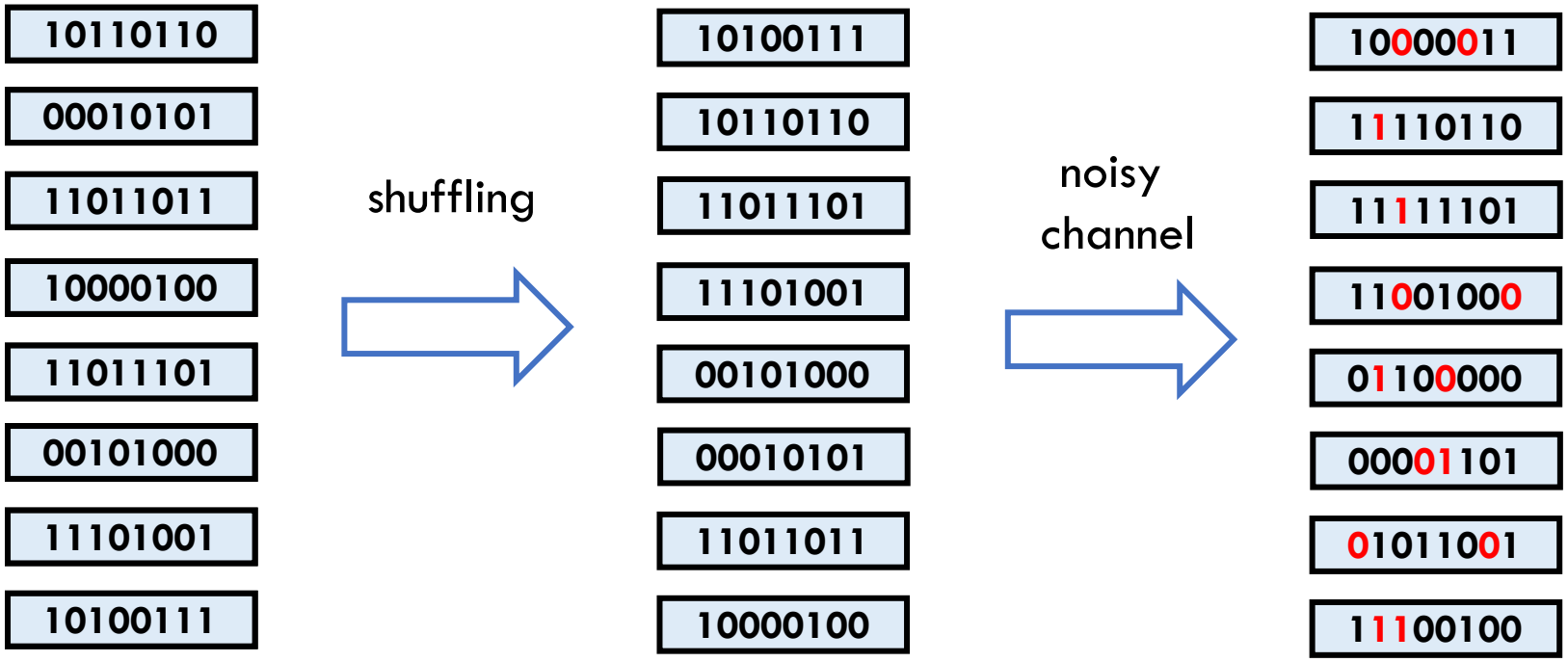
- Good candidate for archival storage

Storing data on DNA

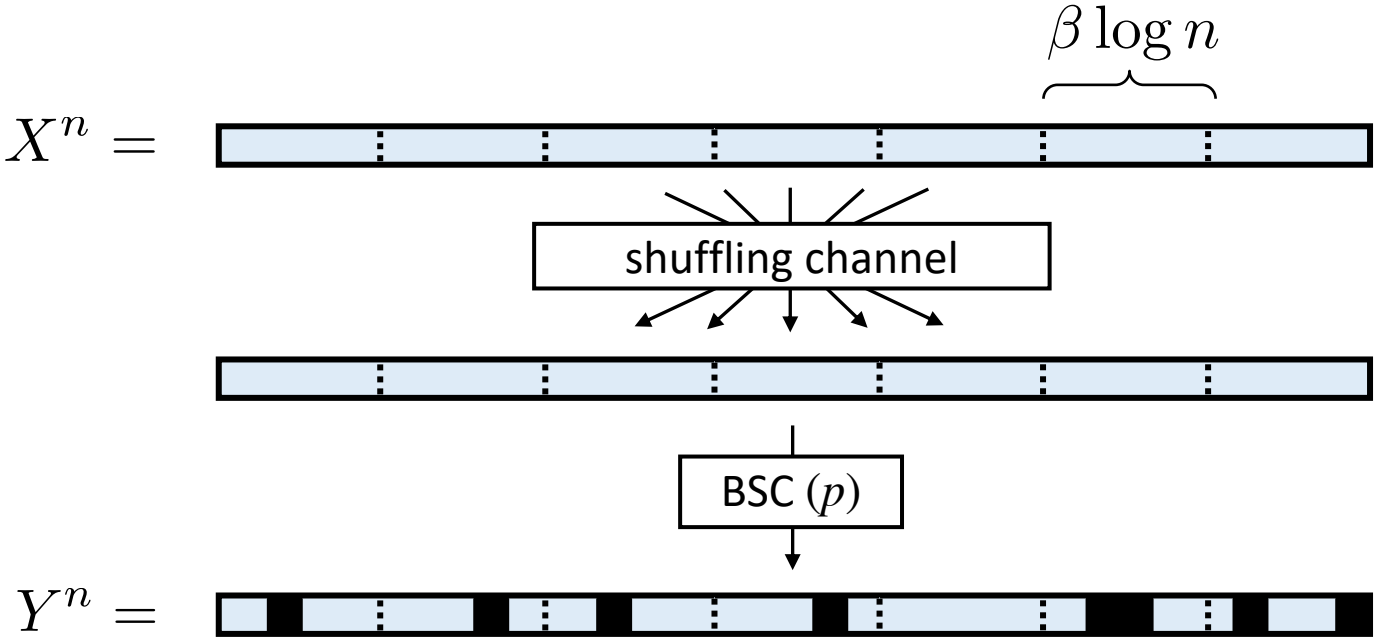


- Can we study the capacity of this channel?
- Distinctive property: out-of-order (shuffled) observations

Noisy Shuffling Channel



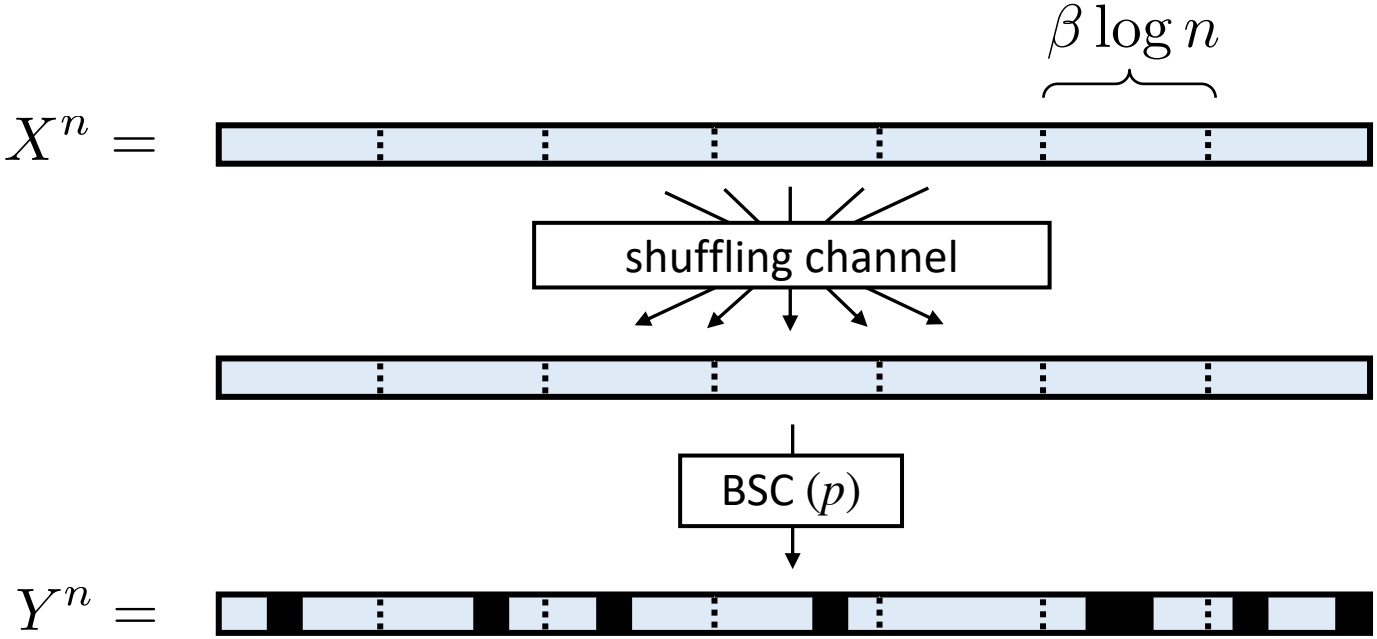
Example: BSC Shuffling Channel



- What is the capacity of this channel?

$$\left(\text{Maximum rate } R = \frac{\log |\mathcal{C}|}{n} \text{ with } P_e \to 0 \text{ as } n \to \infty \right)$$

Example: BSC Shuffling Channel

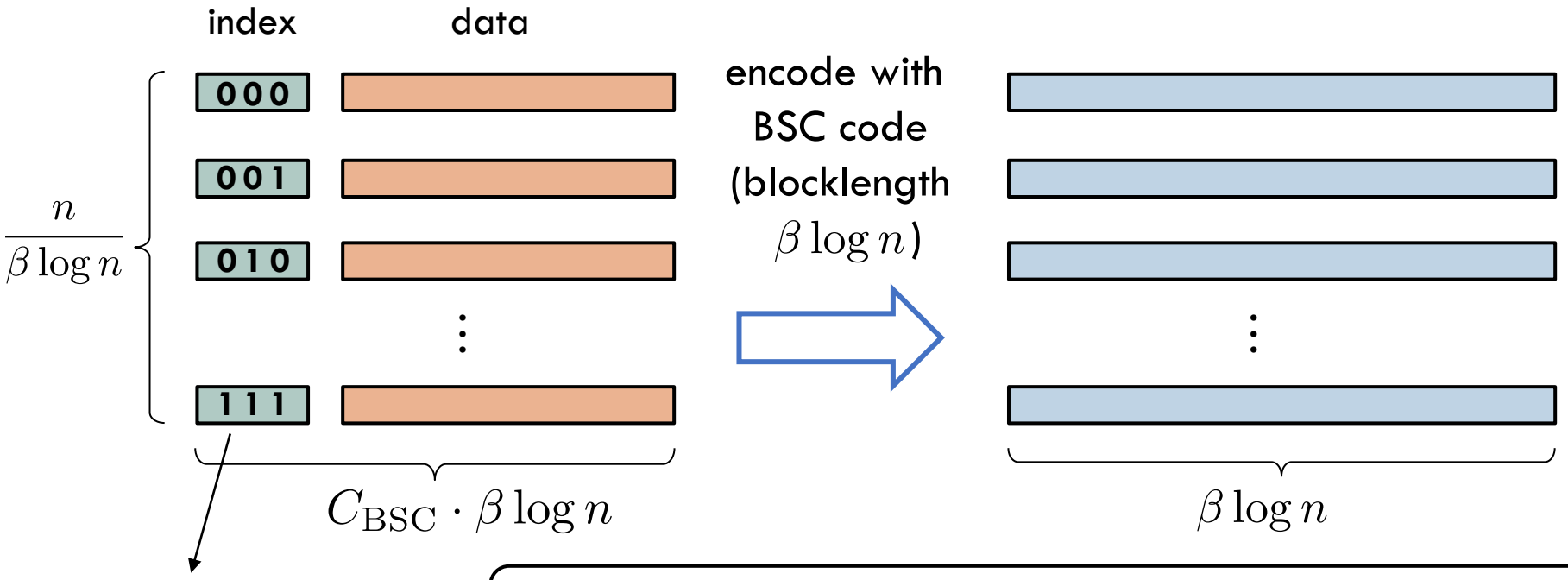


$$C = 1 - H(p) - \frac{1}{\beta}$$

cost of lack of ordering

Simple scheme: index + individual encoding

- Encode **unique index** into each block
- Individually encode block with an optimal BSC code



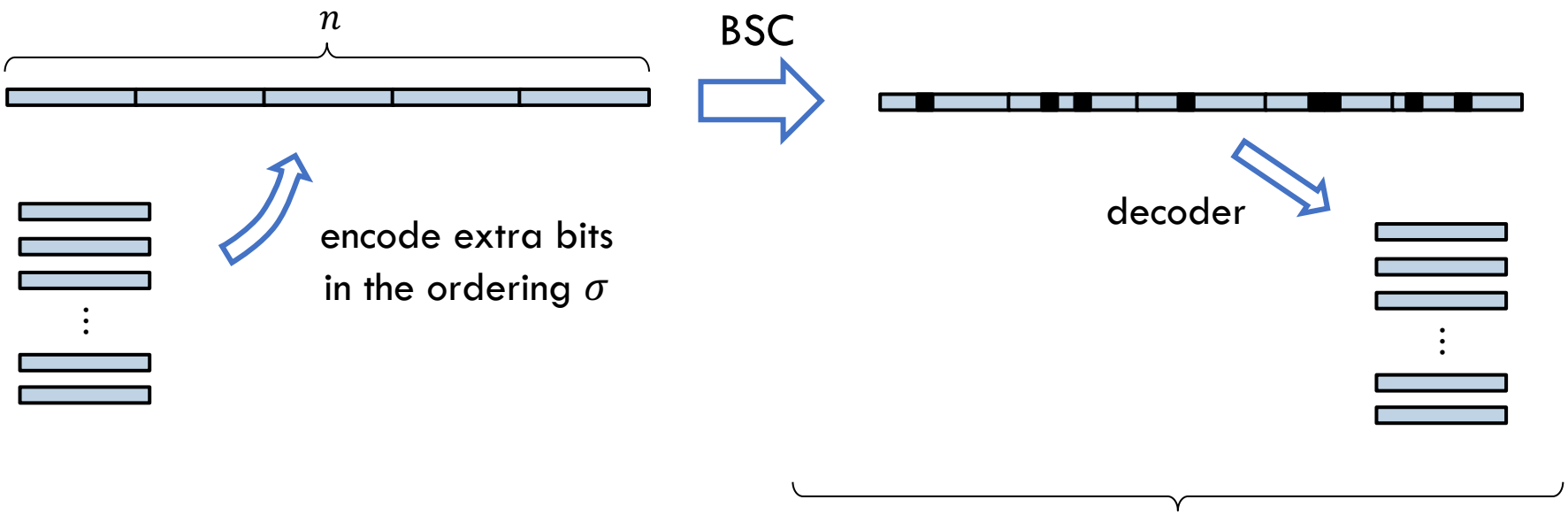
index length:

$$\log\left(\frac{n}{\beta \log n}\right) \approx \log n$$

Rate: $\dfrac{\frac{n}{\beta \log n}(C_{\mathrm{BSC}} \cdot \beta \log n - \log n)}{n}$

Converse idea

- Consider using code for BSC shuffling channel on a regular BSC

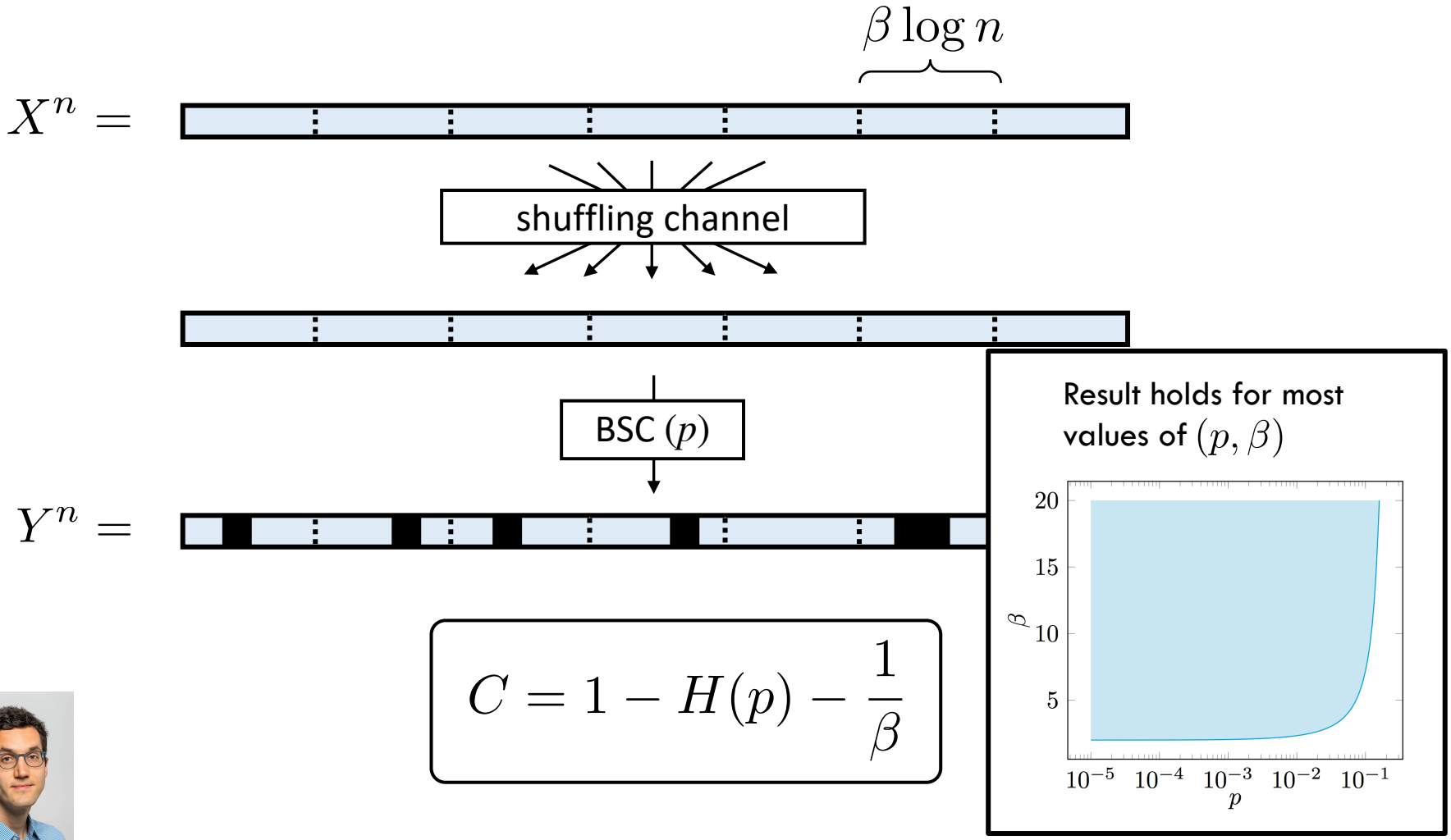


How much of σ can we decode?

$$R_{extra} = \frac{I(\sigma; Y^n)}{n}$$

$$\Rightarrow \quad R + \frac{1}{\beta} \leq 1 - H(p)$$

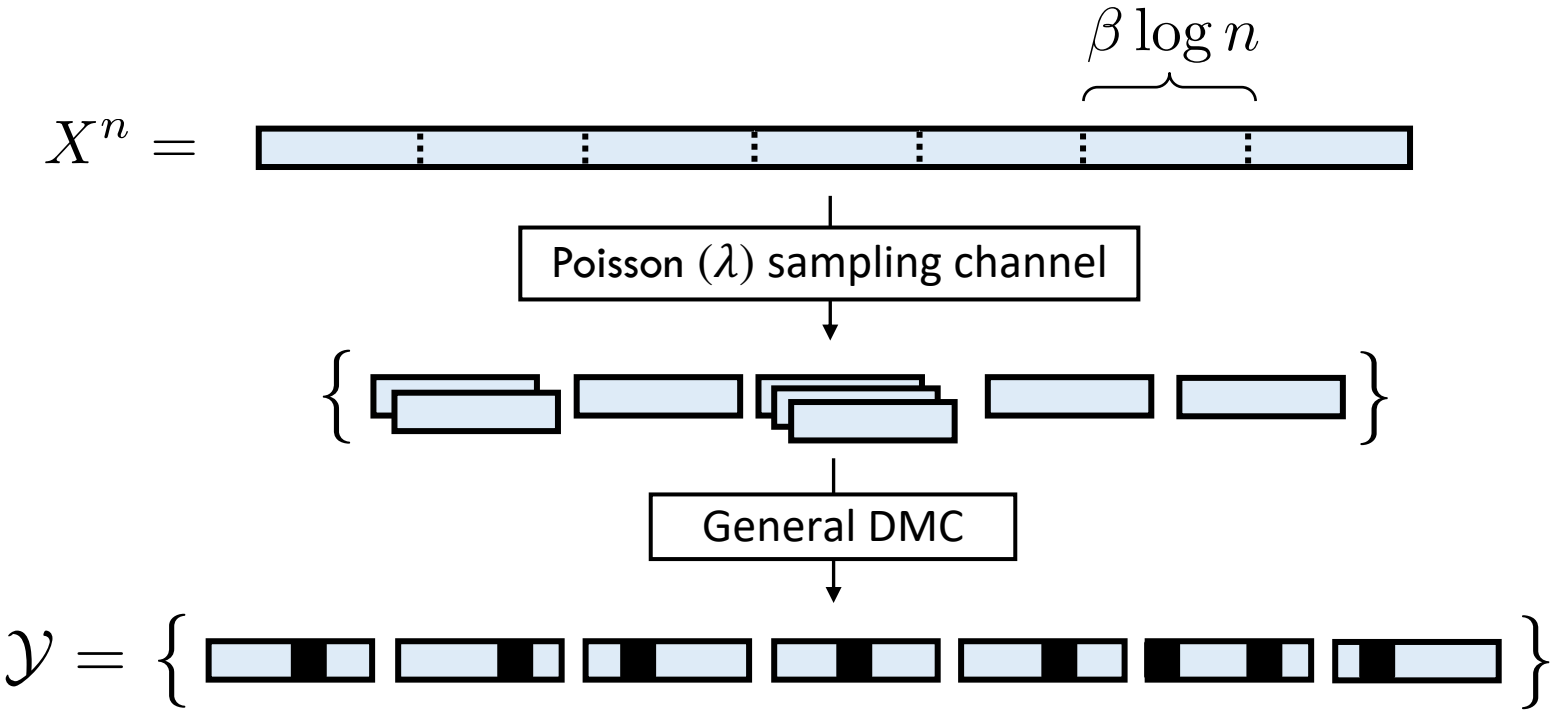
Example: BSC Shuffling Channel



$$C = 1 - H(p) - \frac{1}{\beta}$$

I. Shomorony and R. Heckel, DNA-Based Storage: Models and Fundamental Limits, IT Transactions, 2021

General DNA Storage Channel



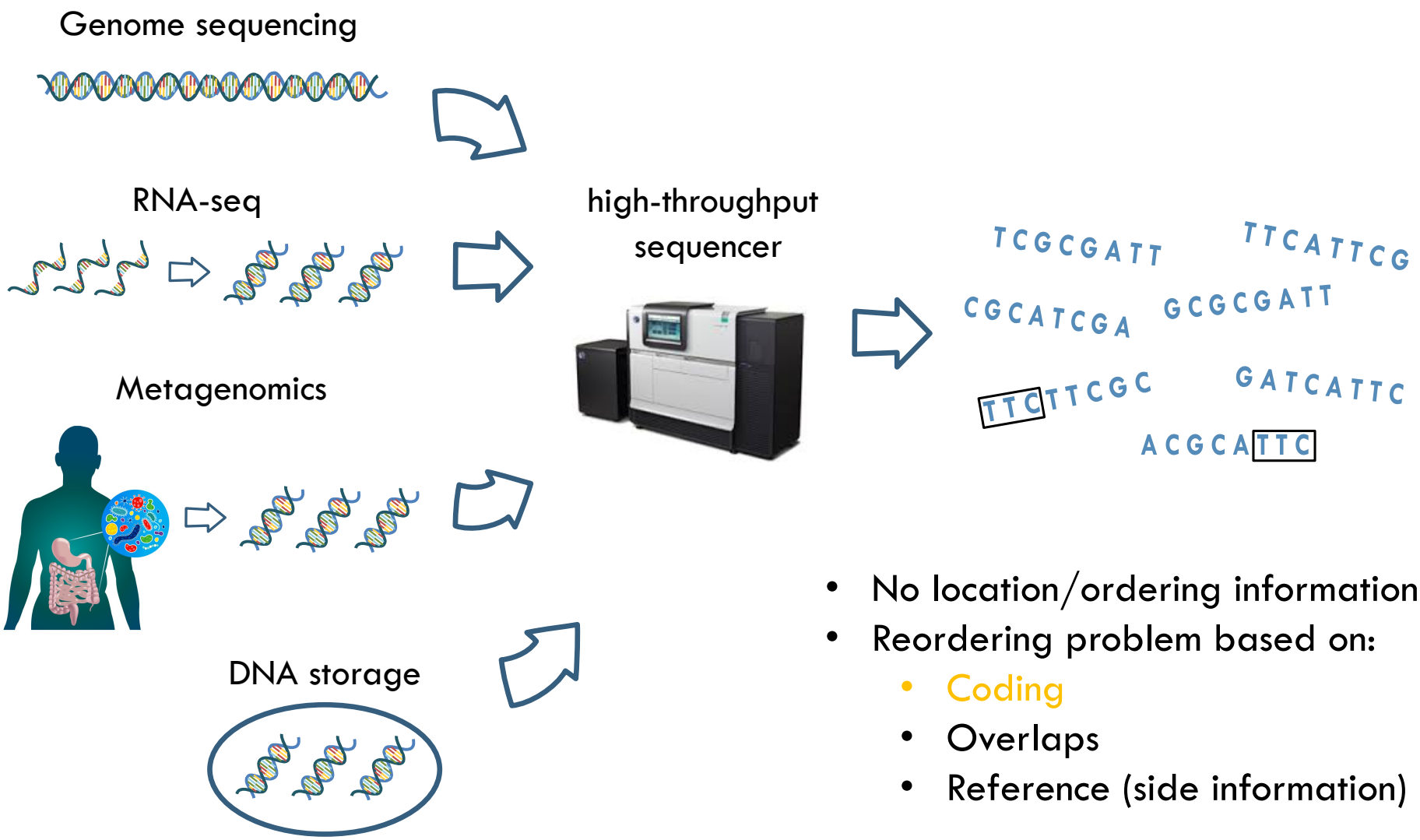
$$C = (1 - e^{-\lambda})(E[C_{\mathrm{DMC}}] - 1/\beta)$$

cost of missing pieces

Lenz, Siegel, Wachter-Zeh, Yaakobi, The Noisy Drawing Channel: Reliable Data Storage in DNA Sequences

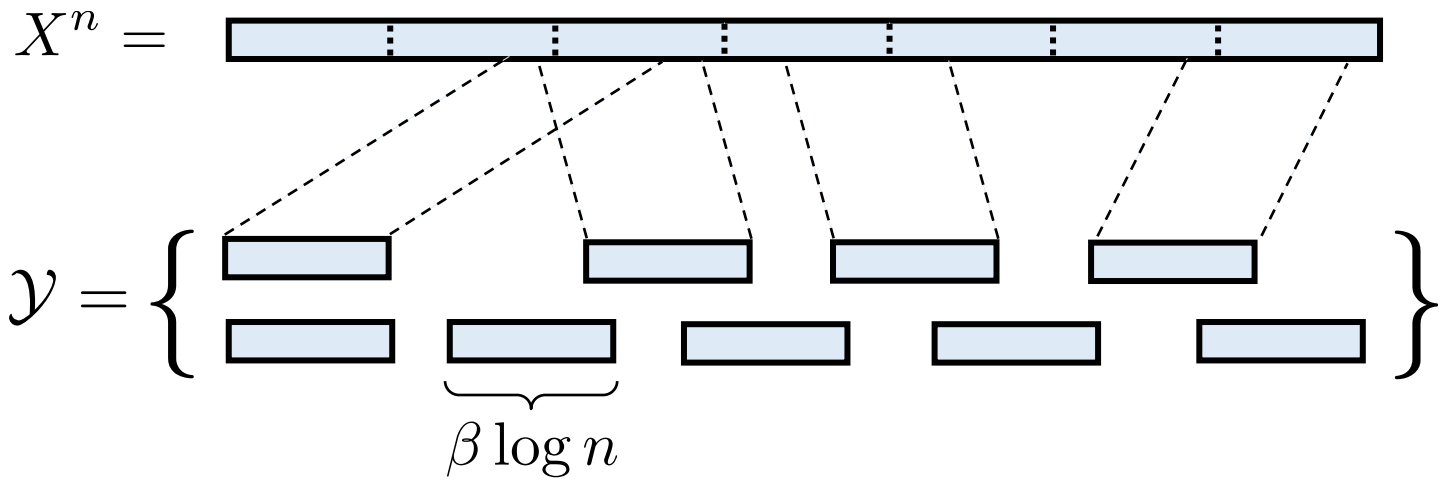
Weinberger and Merhav, The DNA Storage Channel: Capacity and Error Probability Bounds

The wonders of high-throughput sequencing



- No location/ordering information!
- Reordering problem based on:
  - Coding
  - Overlaps
  - Reference (side information)

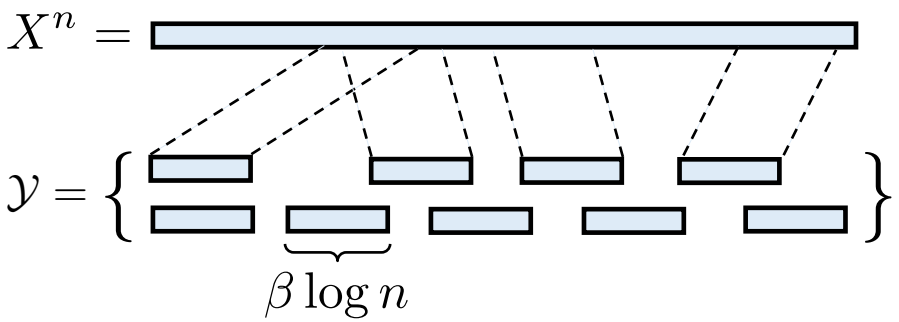
Shotgun Sequencing Channel



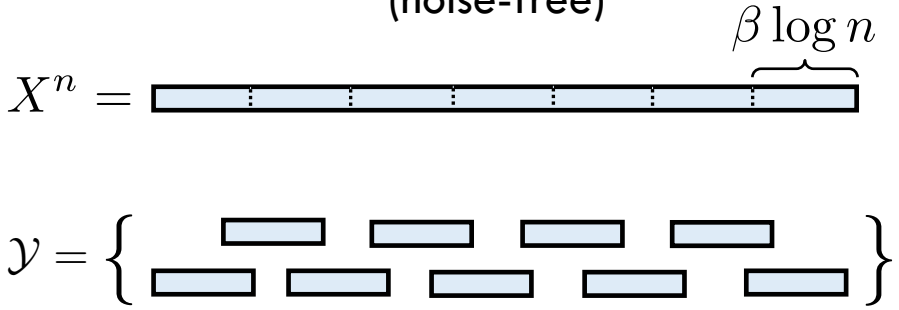
- What is the capacity of this channel?

Shotgun Sequencing Channel

Shotgun Sequencing Channel



Shuffling Channel
(noise-free)



- Cannot place unique identifiers
- Overlaps can help reordering

$$C_{\mathrm{SSC}} = 1 - e^{-\lambda(1-1/\beta)}$$

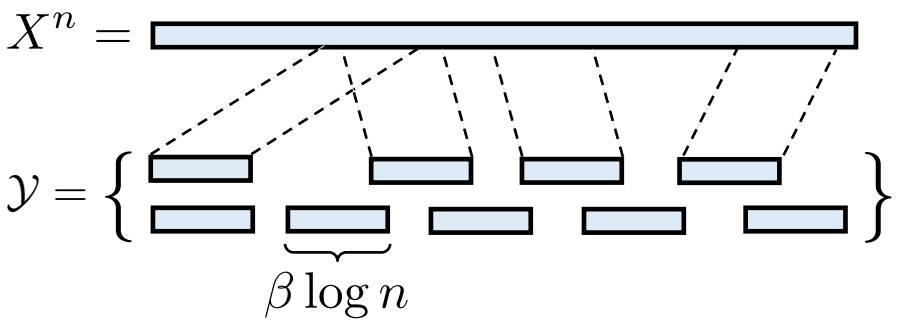
$$C_{\mathrm{shuf}} = (1 - e^{-\lambda})(1 - 1/\beta)$$

sequencing depth: $\lambda = \dfrac{|\mathcal{Y}| \cdot \beta \log n}{n}$

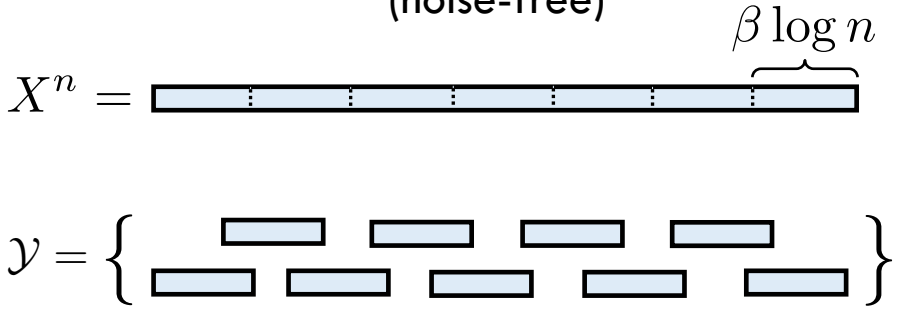
A. Ravi, A. Vahid and I. Shomorony, Coded Shotgun Sequencing, JSAIT, 2022

Shotgun Sequencing Channel

Shotgun Sequencing Channel



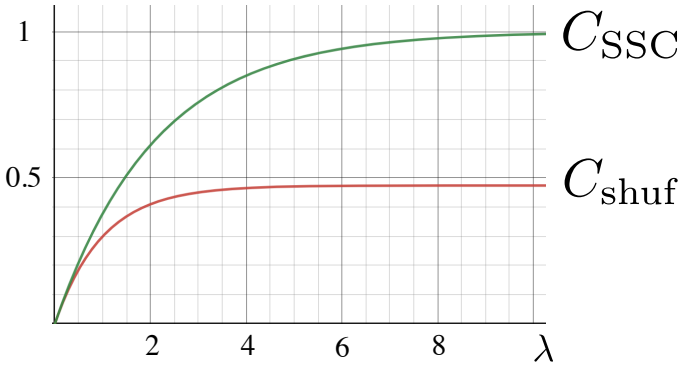
Shuffling Channel
(noise-free)



$$C_{\mathrm{SSC}} = 1 - e^{-\lambda(1-1/\beta)}$$

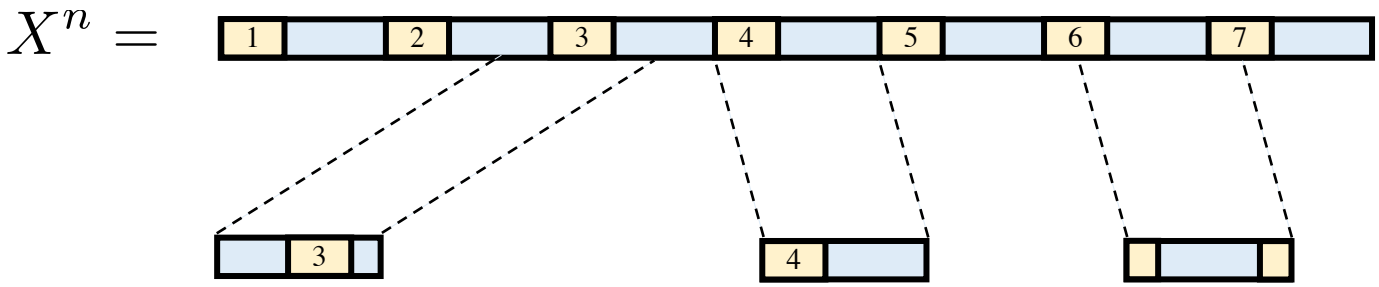
$$C_{\mathrm{shuf}} = (1 - e^{-\lambda})(1 - 1/\beta)$$

overlaps neutralize lack of ordering information

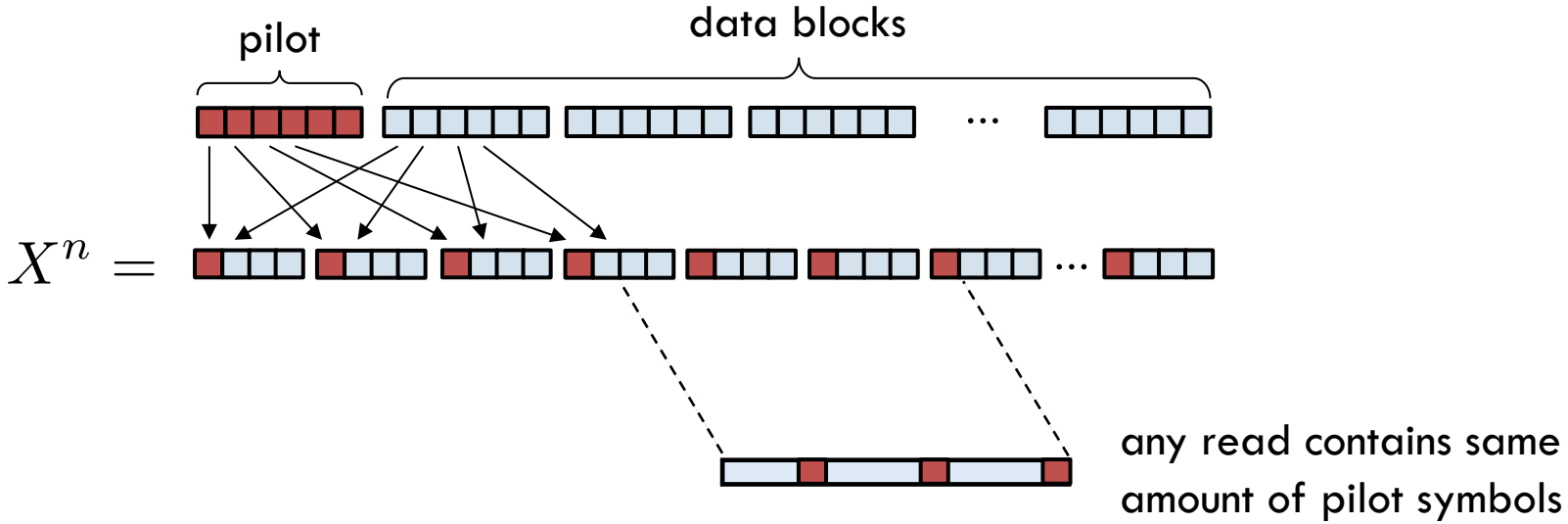


$$\lambda = \frac{|\mathcal{Y}| \cdot \beta \log n}{n}$$

Coding for the Shotgun Sequencing Channel

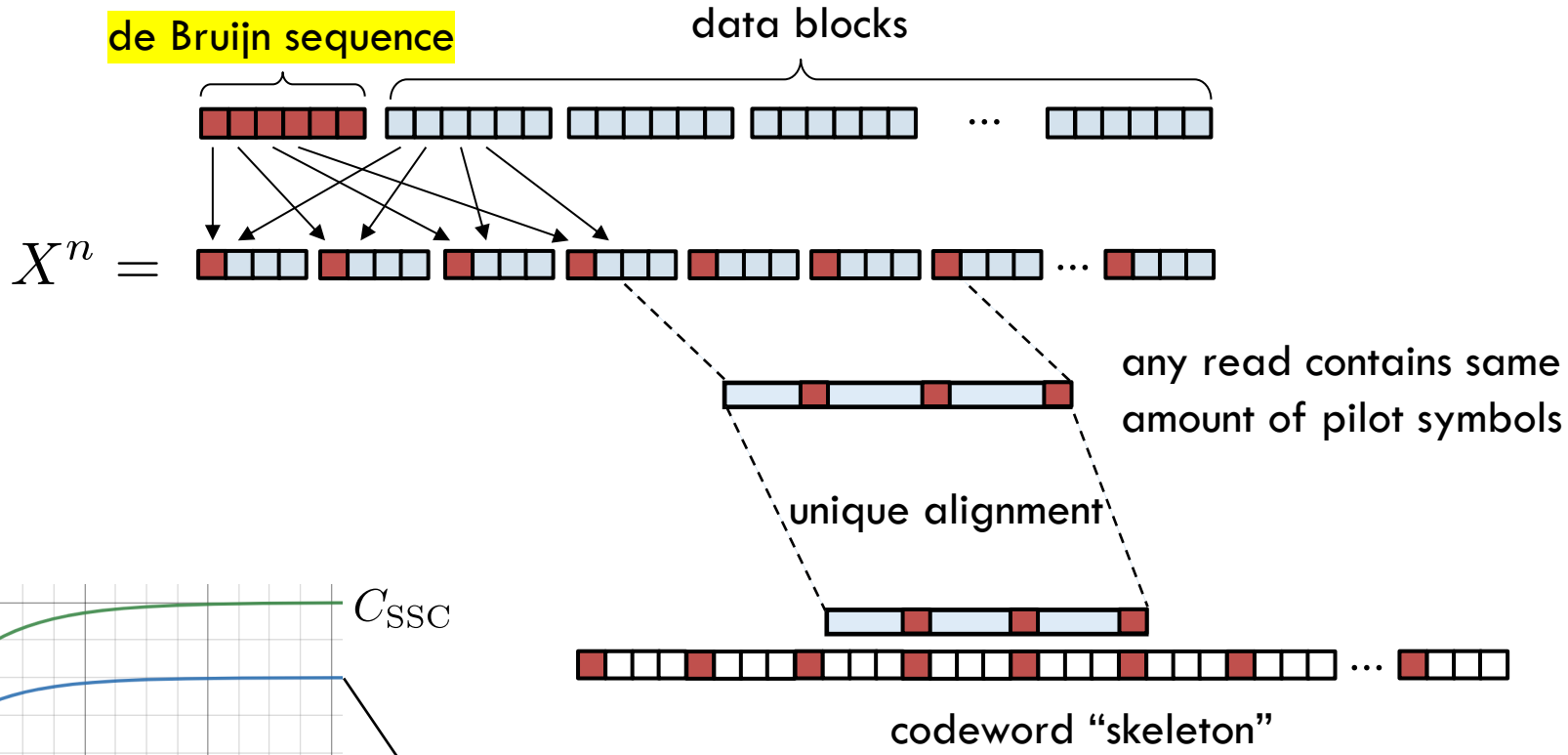


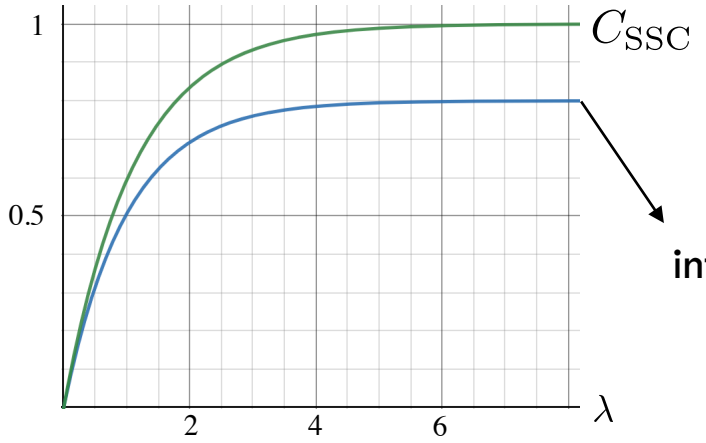
- Better approach: interleave a pilot sequence in X^n



Coding for the Shotgun Sequencing Channel

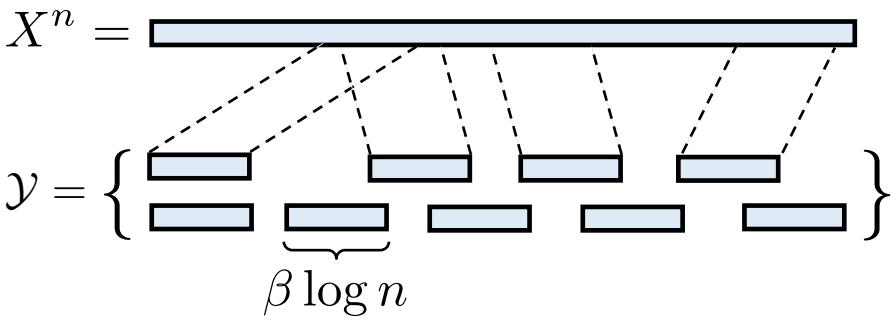
- Better approach: interleave a pilot sequence in X^n





interleaved pilot: doesn't take advantage of overlaps

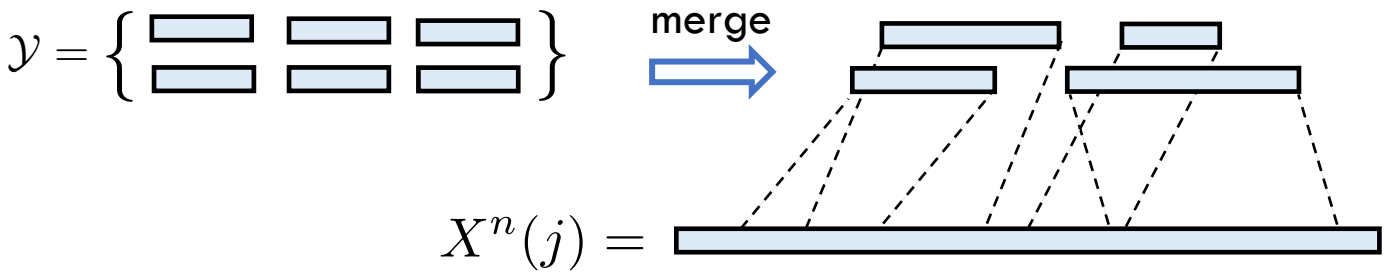
Coding for the Shotgun Sequencing Channel



$$C_{\mathrm{SSC}} = 1 - e^{-\lambda(1-1/\beta)}$$

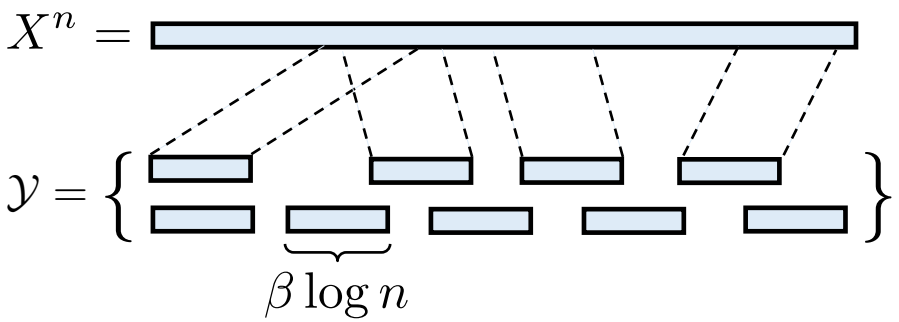
$$\lambda = \frac{|\mathcal{Y}| \cdot \beta \log n}{n}$$

- Achievability
  - Random code: i.i.d. $\mathrm{Ber}(1/2)$ codewords
  - Decoding: merge reads into "islands" + align to codewords

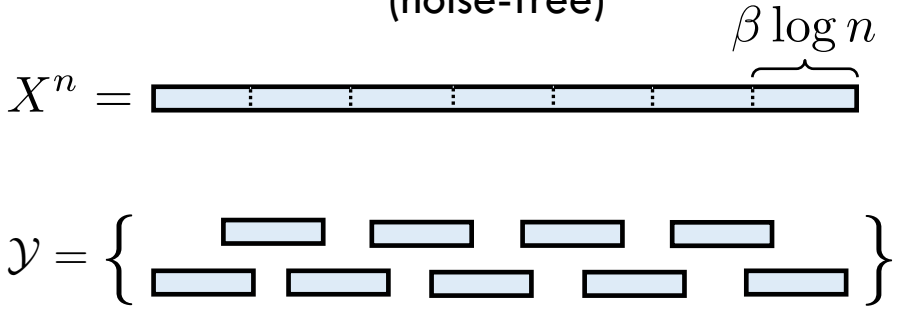


Shotgun Sequencing Channel

Shotgun Sequencing Channel

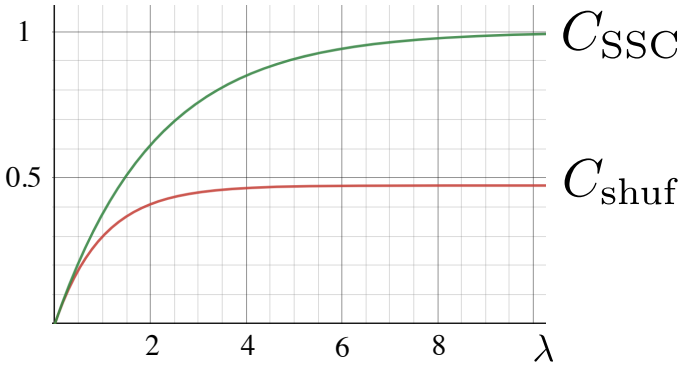


Shuffling Channel
(noise-free)



$$C_{\mathrm{SSC}} = 1 - e^{-\lambda(1-1/\beta)}$$

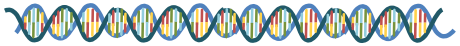
$$C_{\mathrm{shuf}} = (1 - e^{-\lambda})(1 - 1/\beta)$$



$$\lambda = \frac{|\mathcal{Y}| \cdot \beta \log n}{n}$$

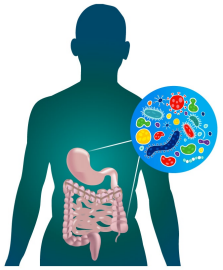
The wonders of high-throughput sequencing

Genome sequencing

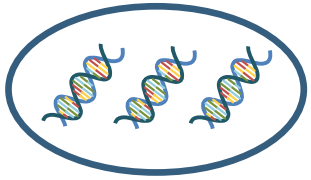


RNA-seq

Metagenomics



DNA storage



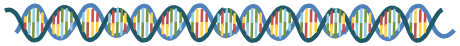
high-throughput sequencer

TCGCGATT TTCATTCG
CGCATCGA GCGCGATT
TTCTTCGC GATCATTC
ACGCATTC

- No location/ordering information!
- Reordering problem based on:
  - Coding
  - Overlaps
  - Reference (side information)

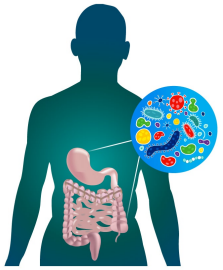
The wonders of high-throughput sequencing

Genome sequencing

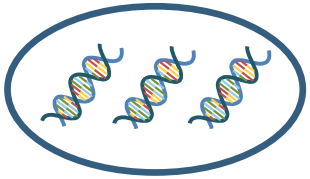


RNA-seq

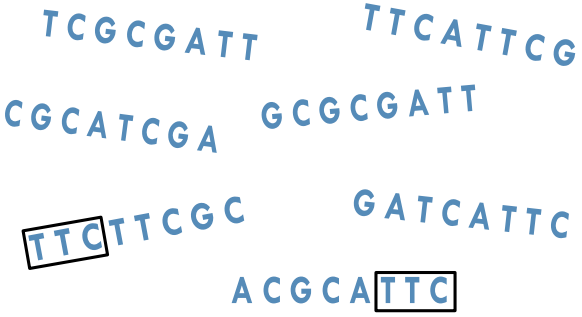
Metagenomics



DNA storage



high-throughput sequencer

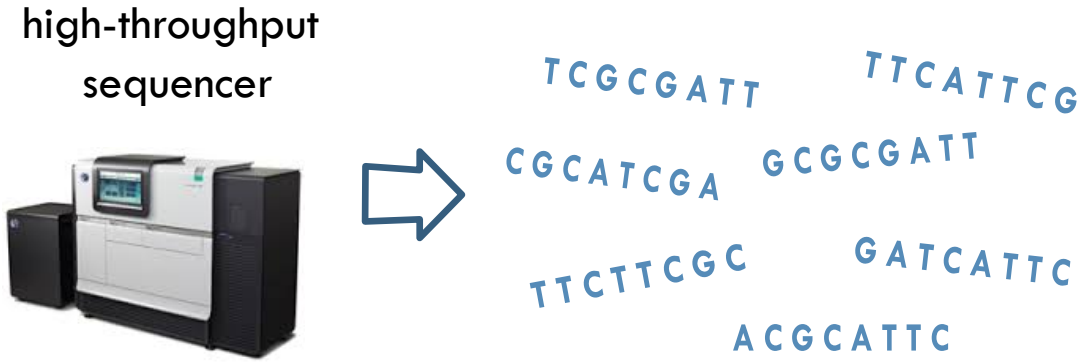


- No location/ordering information!

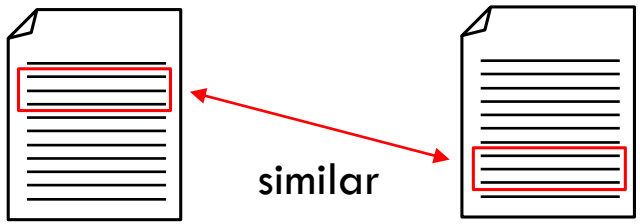
Outline:

1. Information-theoretic limits of out-of-order channels
2. Efficiently finding overlaps

Finding pairwise overlaps



- Computational bottleneck in bioinformatics
- Other applications: document comparison (plagiarism detection)



Finding pairwise overlaps

high-throughput sequencer

TCGCGATT TTCATTCG
CGCATCGA GCGCGATT
TTCTTCGC GATCATTC
ACGCATTC

m

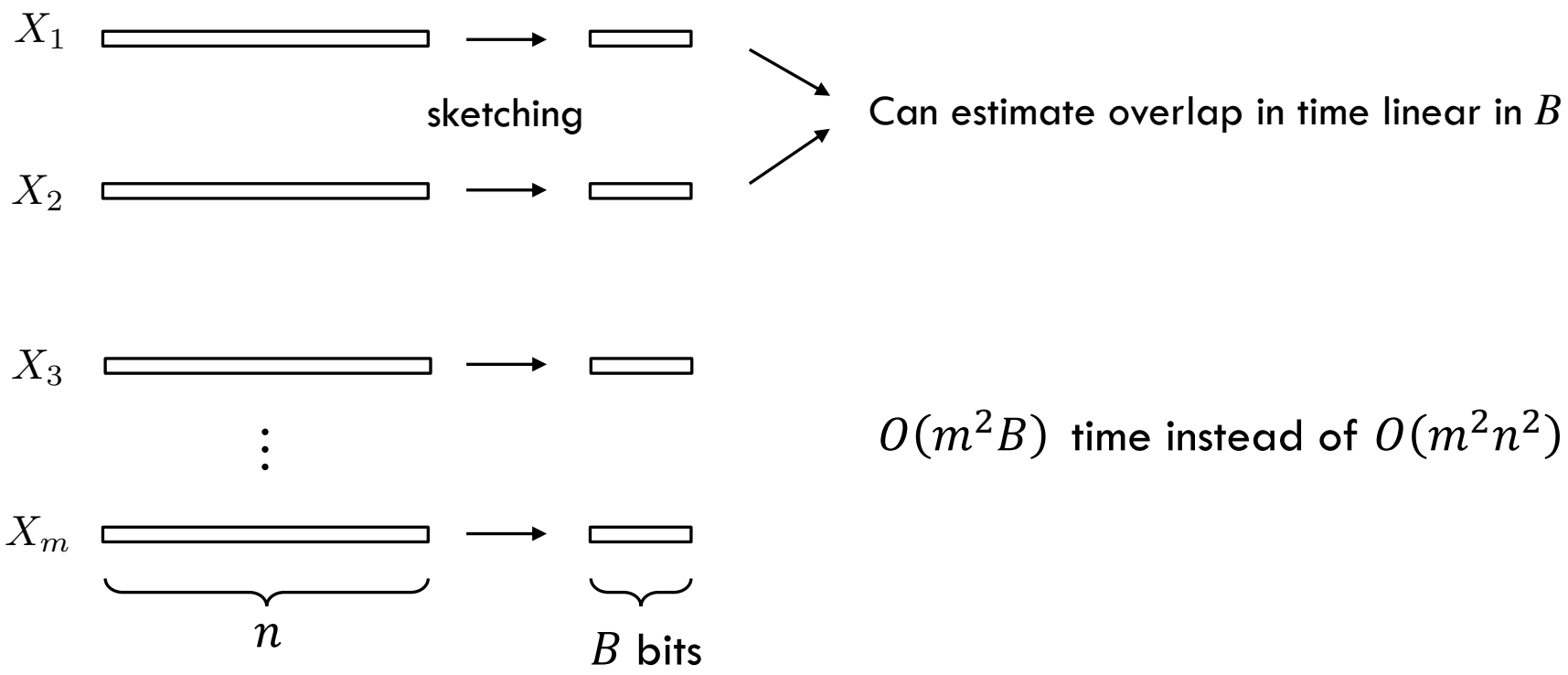
- Given two noisy sequences, find best overlap between them

GCCGTAGGCGATTTGCCCG-CG
CGGTTTGTCCGACGGCGGCTAG

n

- Dynamic programming solution: $O(m^2n^2)$ time

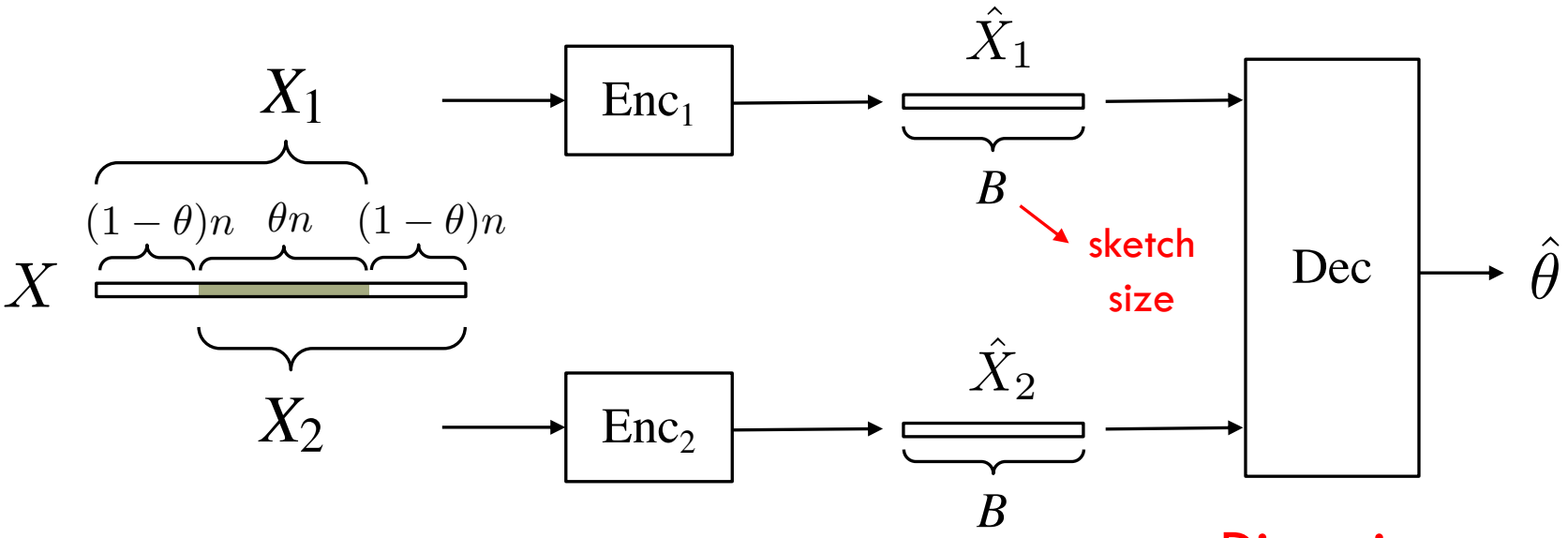
Finding overlaps in practice



- **Fundamental tradeoff:** Sketch size and overlapping accuracy

Distributed source coding formulation

- We focus on estimating the **overlap** size $\theta \in (0,1)$



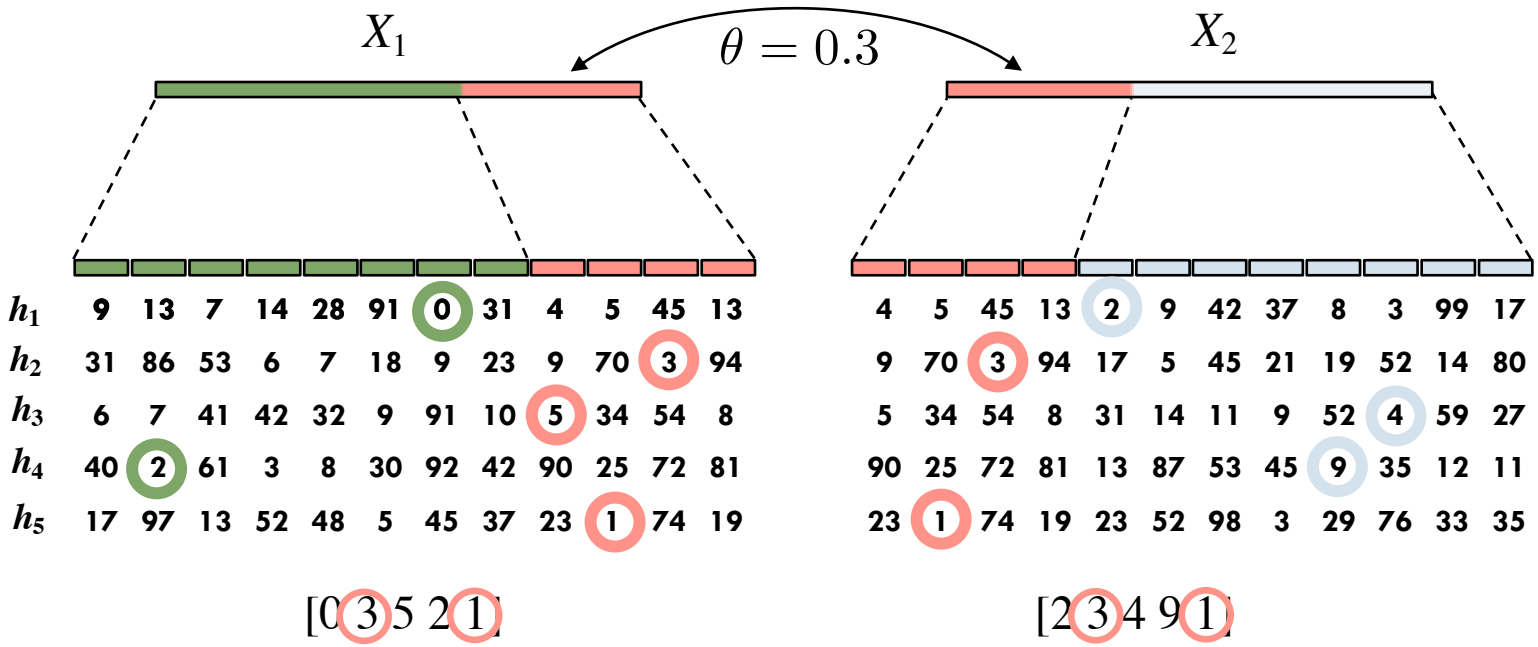
Distortion:

$$E[(\theta - \hat{\theta})^2] \leq D$$

- $B(D)$: Minimum B required to achieve distortion D (as $n \to \infty$)
- Related to literature on distributed parameter estimation
  (Amari et al., "Statistical inference under multiterminal data compression," 1998)

Standard approach to sketching

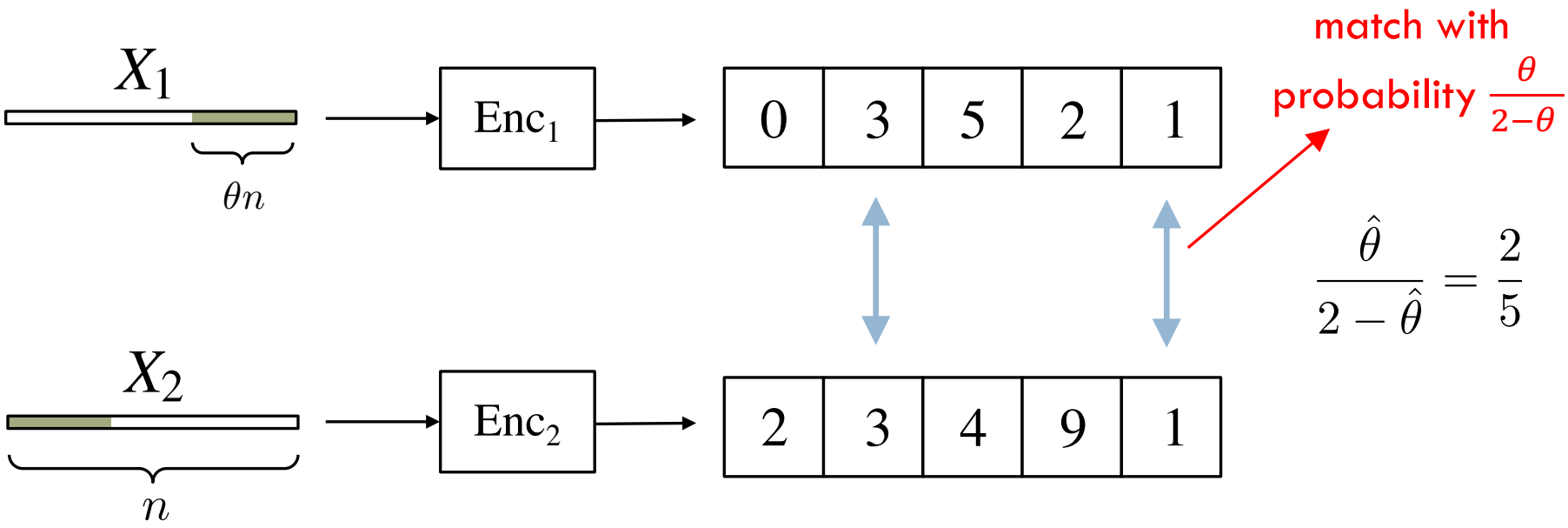
- Sketches based on locality-sensitive hashing (min-hash)



overlap estimate = 2/5

Standard approach to sketching

- Sketches based on locality-sensitive hashing (min-hash)



- Achievable sketch size: $B \leq \frac{3 \log n}{D}$

Fundamental Limits of Sketching

- Previous approach based on min-hash achieves

$$B(D) \leq \frac{3 \log n}{D}$$

$B(D)$: minimum sketch size

D: distortion

Theorem: The function $B(D)$ satisfies

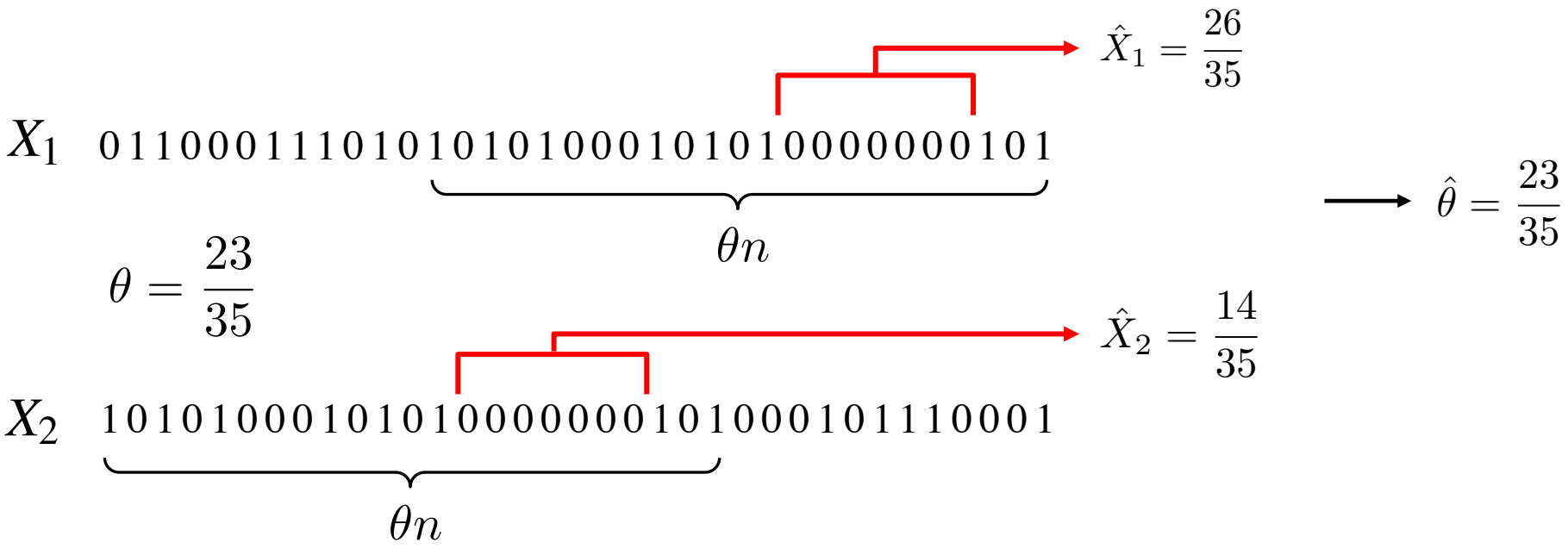
$$K_1 \cdot \log\left(\frac{1}{D}\right) \leq B(D) \leq K_2 \cdot \log^2\left(\frac{1}{D}\right)$$

New sketching idea:

Lexicographic Hashing

I. Shomorony and G. Kamath, Sketching and Sequence Alignment: A Rate-Distortion Perspective, ISIT, 2022

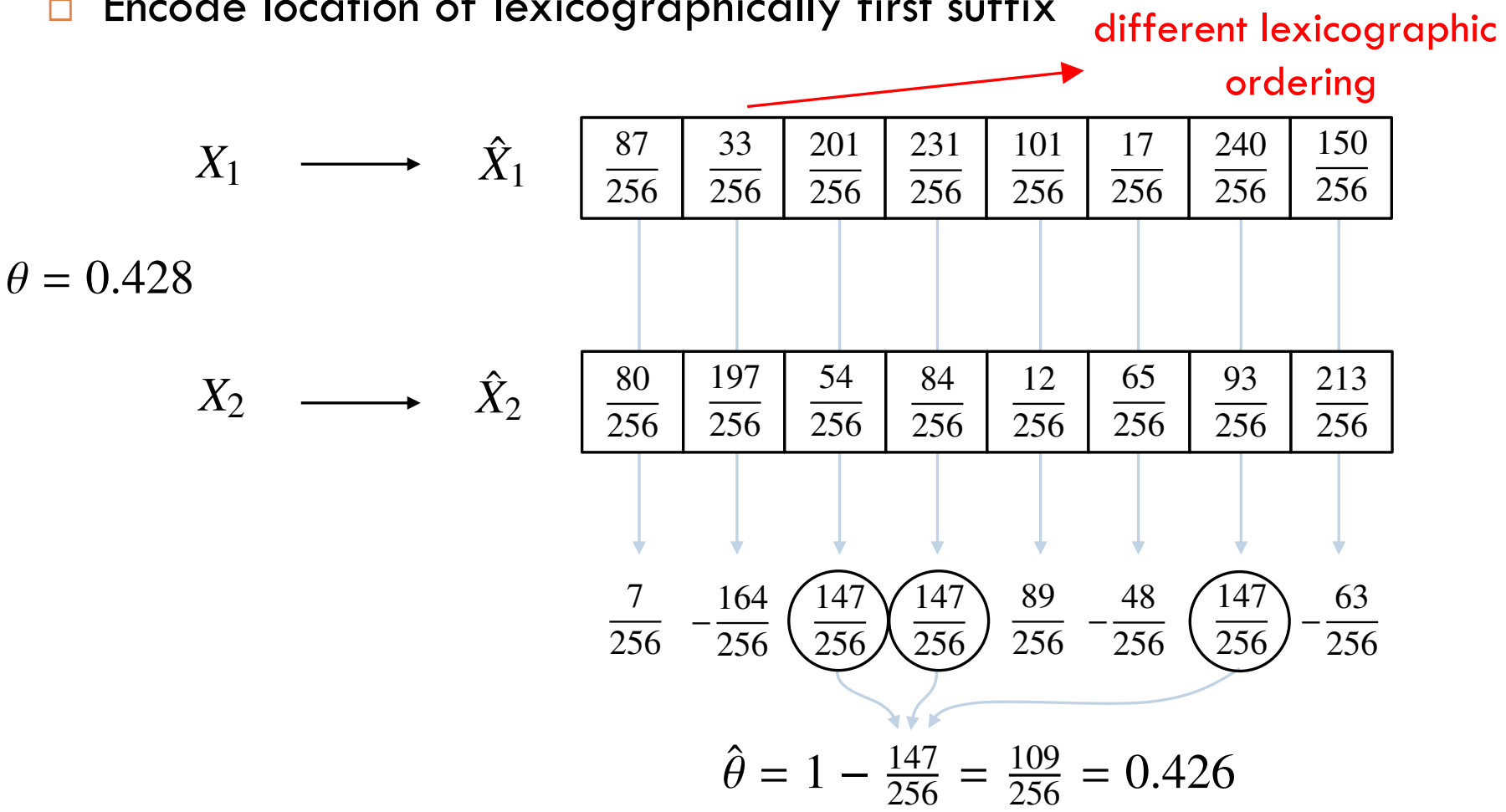
New idea: Lexicographic Hashing



- Encode location of longest run of zeros
- Overlap estimate: $\hat{\theta} = 1 - (\hat{X}_1 - \hat{X}_2)$

New idea 2: Multiple lexicographic orderings

- Encode location of lexicographically first suffix



Fundamental Limits of Sketching

- Previous approach based on min-hash achieves

$$B(D) \leq \frac{3 \log n}{D}$$

$B(D)$: minimum sketch size

D: distortion

Theorem: The function $B(D)$ satisfies

$$K_1 \cdot \log\left(\frac{1}{D}\right) \leq B(D) \leq K_2 \cdot \log^2\left(\frac{1}{D}\right)$$

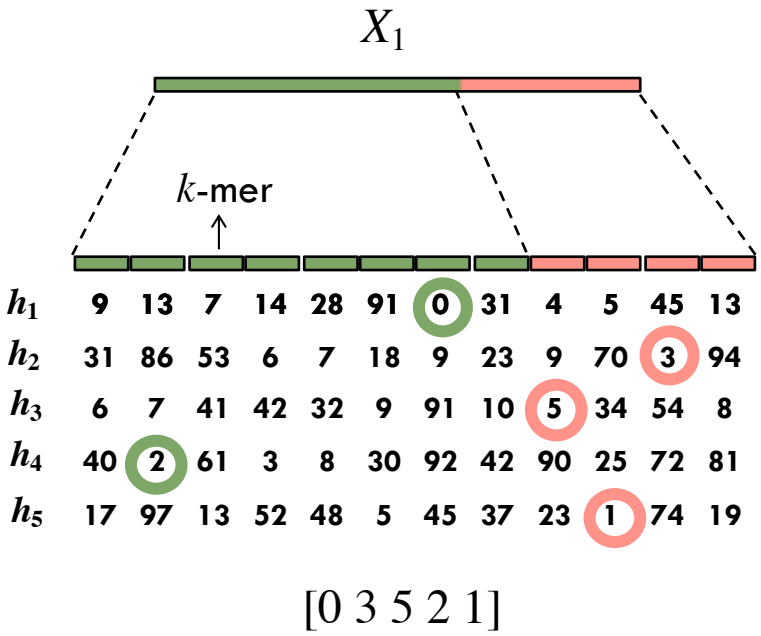
New sketching idea:

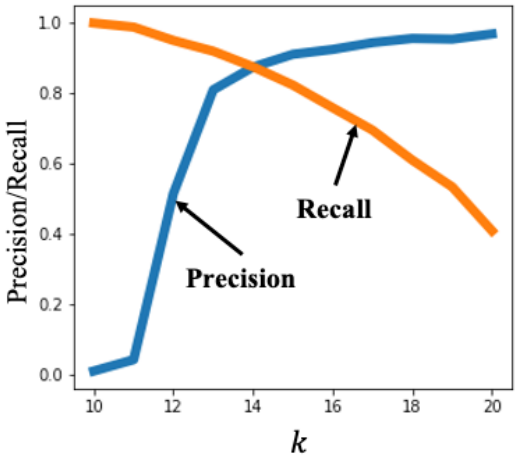
Lexicographic Hashing

I. Shomorony and G. Kamath, Sketching and Sequence Alignment: A Rate-Distortion Perspective, ISIT, 2022

Practical overlap finding: LexicHash

- Sequence alignment tool based on lexicographic hashing
- More accurate overlap estimates
- Avoids need to select parameter k from min-hash

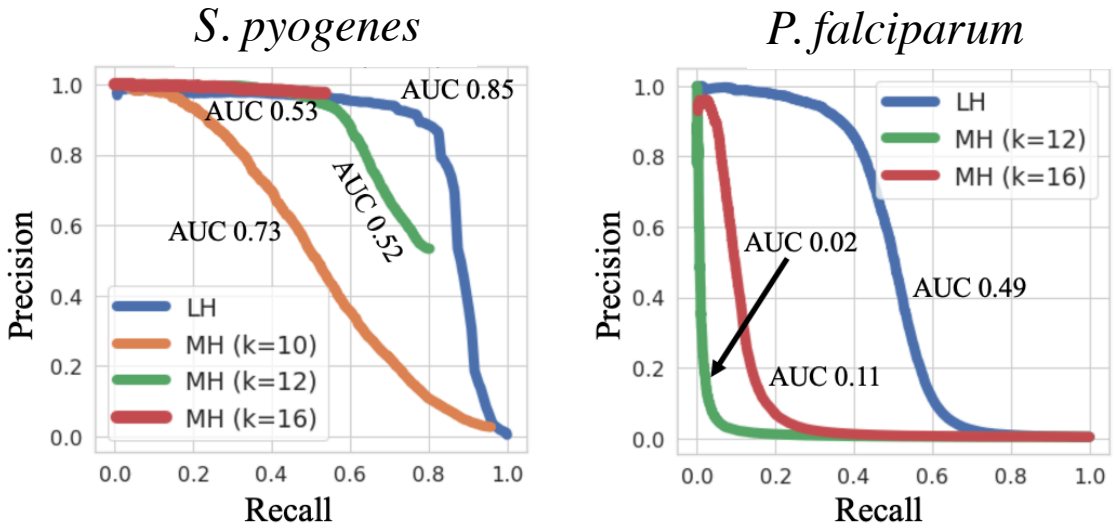




G. Greenberg, A. Ravi, I. Shomorony, LexicHash: Sequence Similarity Estimation via Lexicographic Comparison of Hashes (submitted)

Practical overlap finding: LexicHash

- Sequence alignment tool based on lexicographic hashing
- More accurate overlap estimates
- Avoids need to select parameter k from min-hash
- Significant improvements in overlap classification:



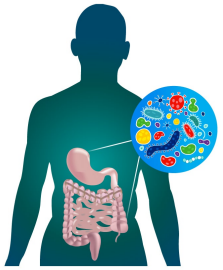
G. Greenberg, A. Ravi, I. Shomorony, LexicHash: Sequence Similarity Estimation via Lexicographic Comparison of Hashes (submitted)

The wonders of high-throughput sequencing

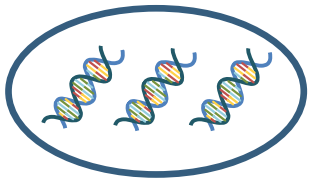
Genome sequencing

RNA-seq

Metagenomics



DNA storage



high-throughput sequencer

TCGCGATT TTCATTCG
CGCATCGA GCGCGATT
TTCTTCGC GATCATTC
ACGCATTC

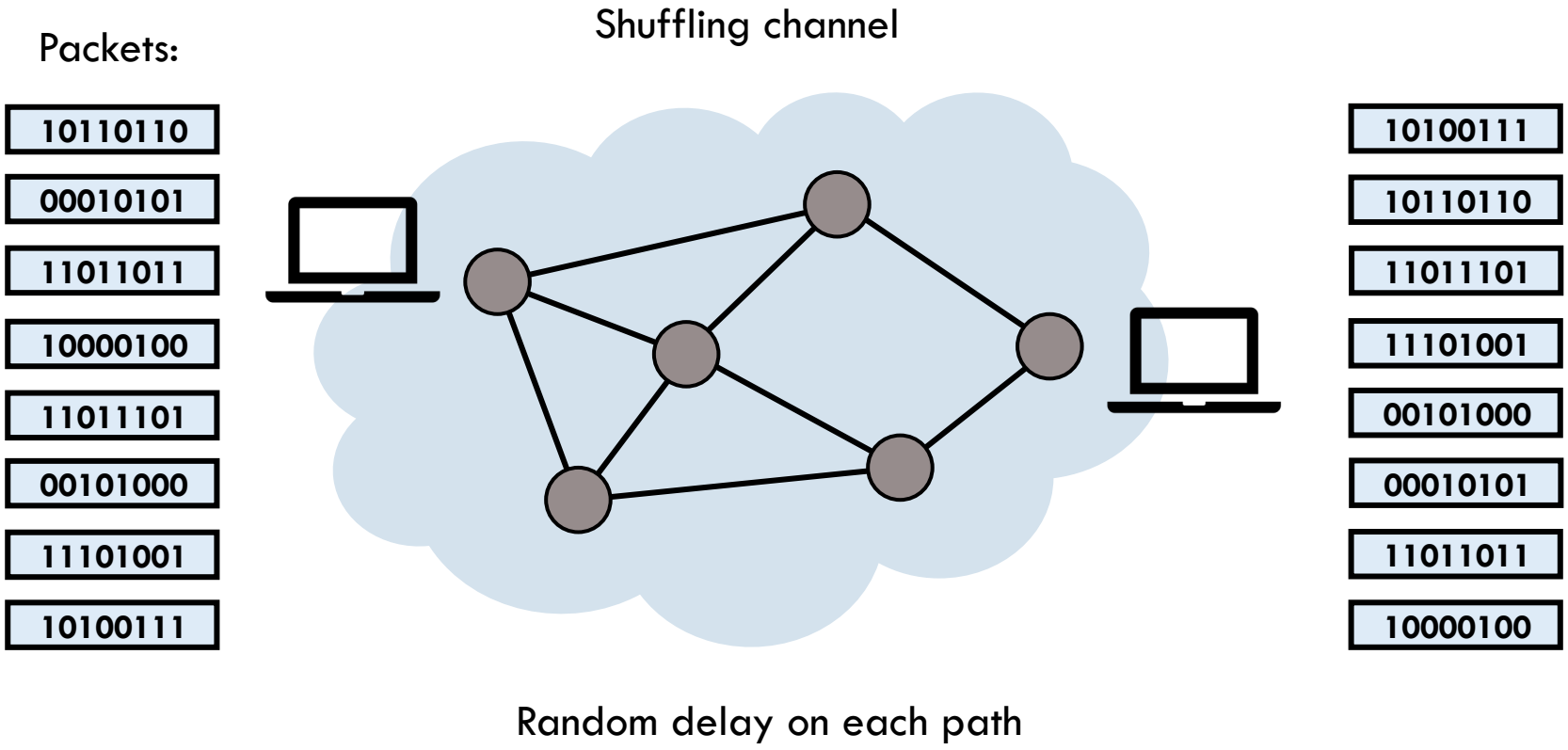
- No location/ordering information!

Outline:

1. Information-theoretic limits of out-of-order channels
2. Efficiently finding overlaps

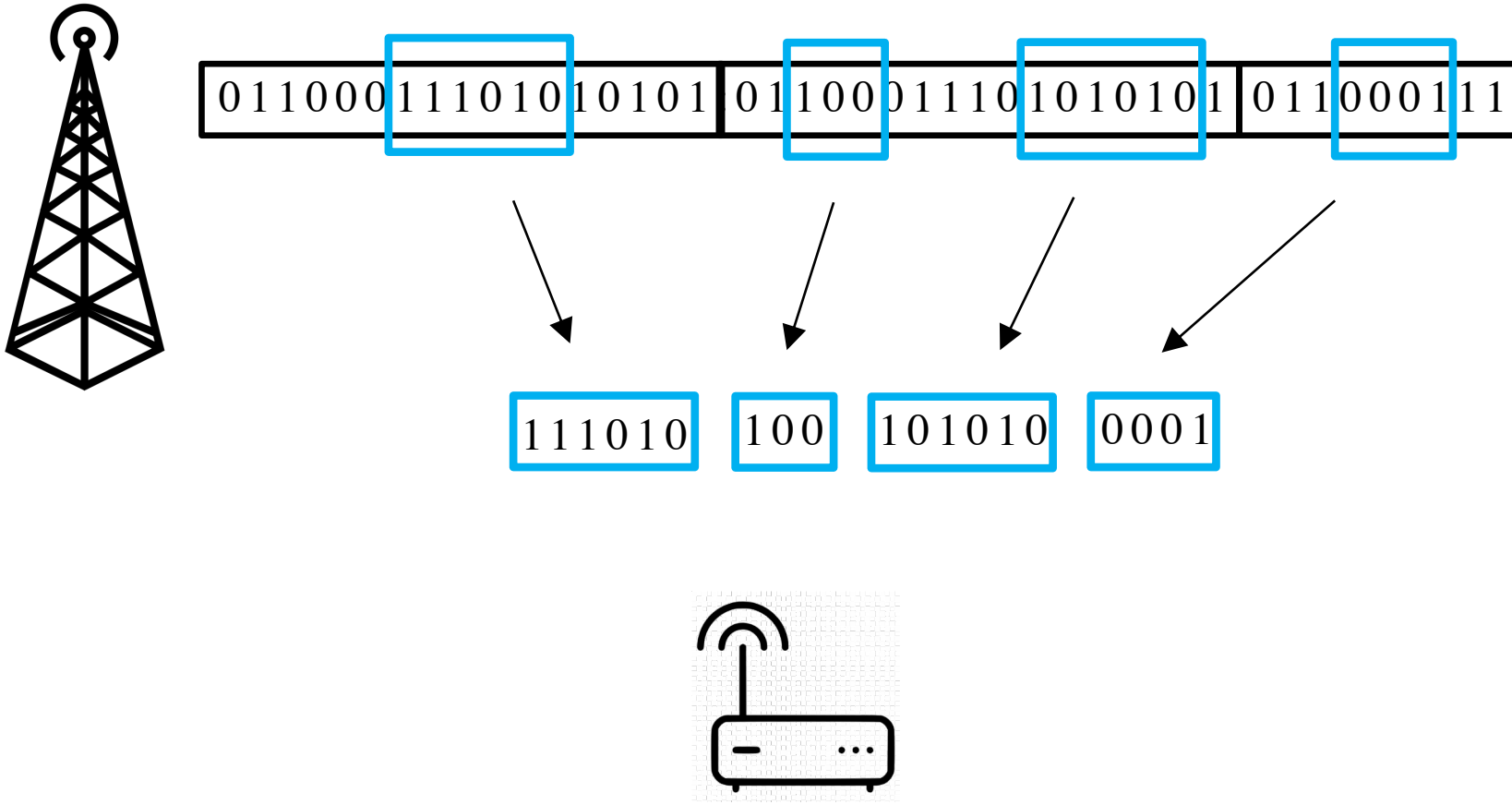
Out-of-order information?

- Packet networks



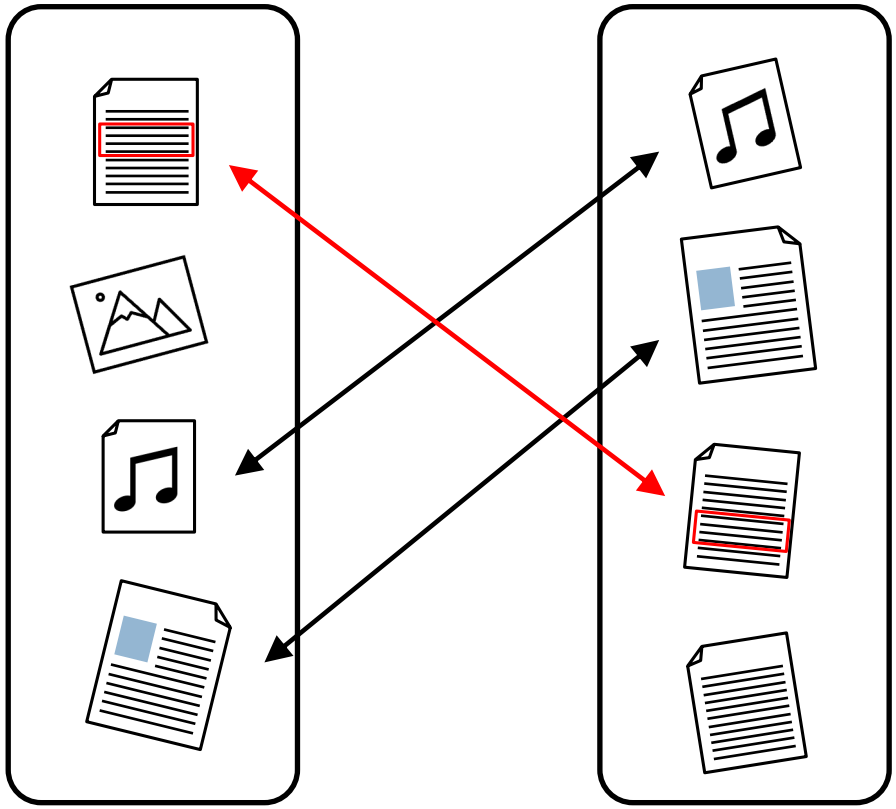
Out-of-order information?

- Intermittent broadcast communication?

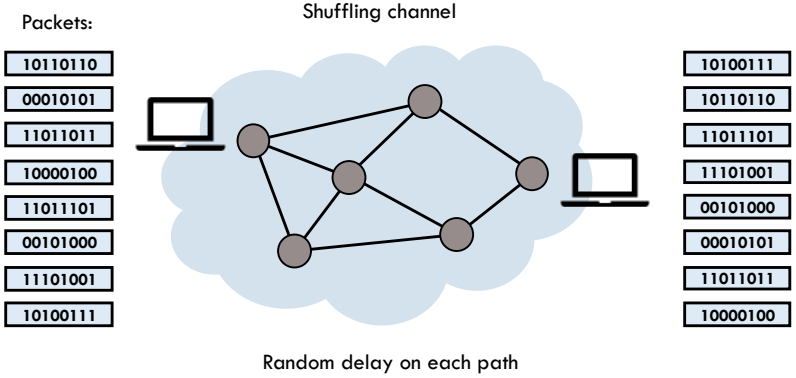


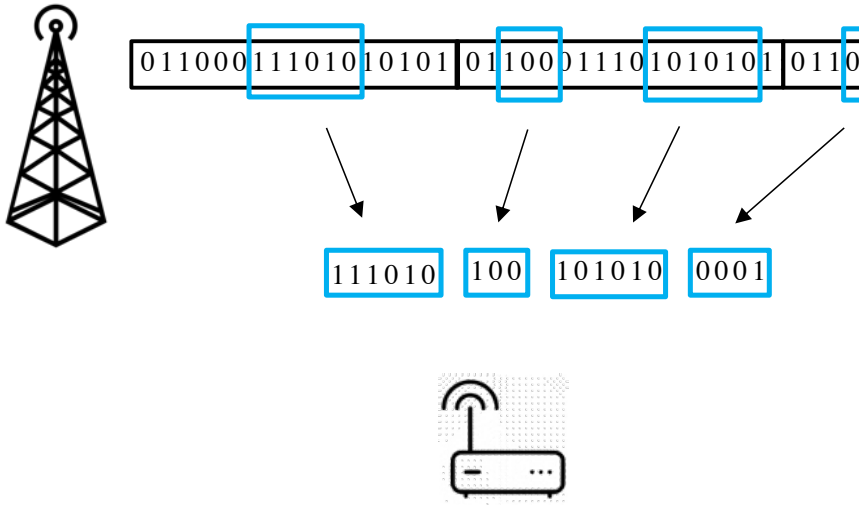
Out-of-order information?

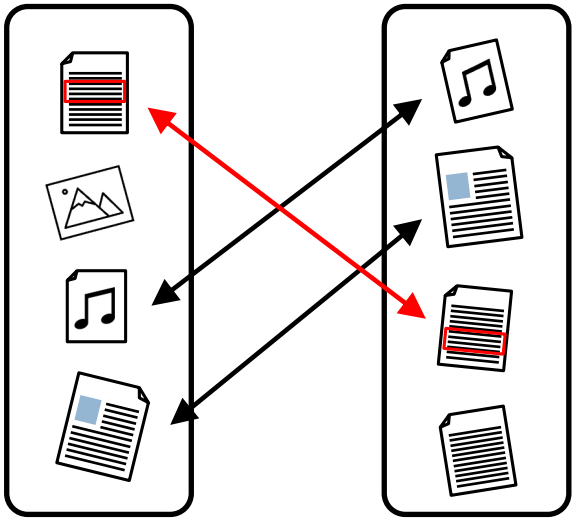
- Dataset alignment?



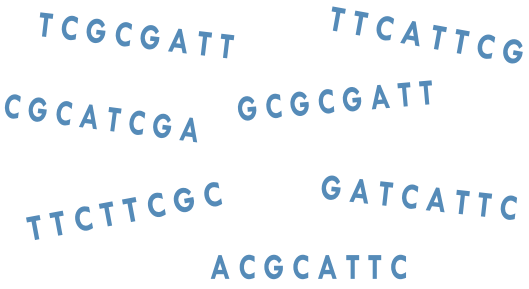
Out-of-Order Information Theory?







high-throughput sequencer



Thank you!