Shannon-inspired research tales on Duality, Encryption, Sampling & Storage

Kannan Ramchandran

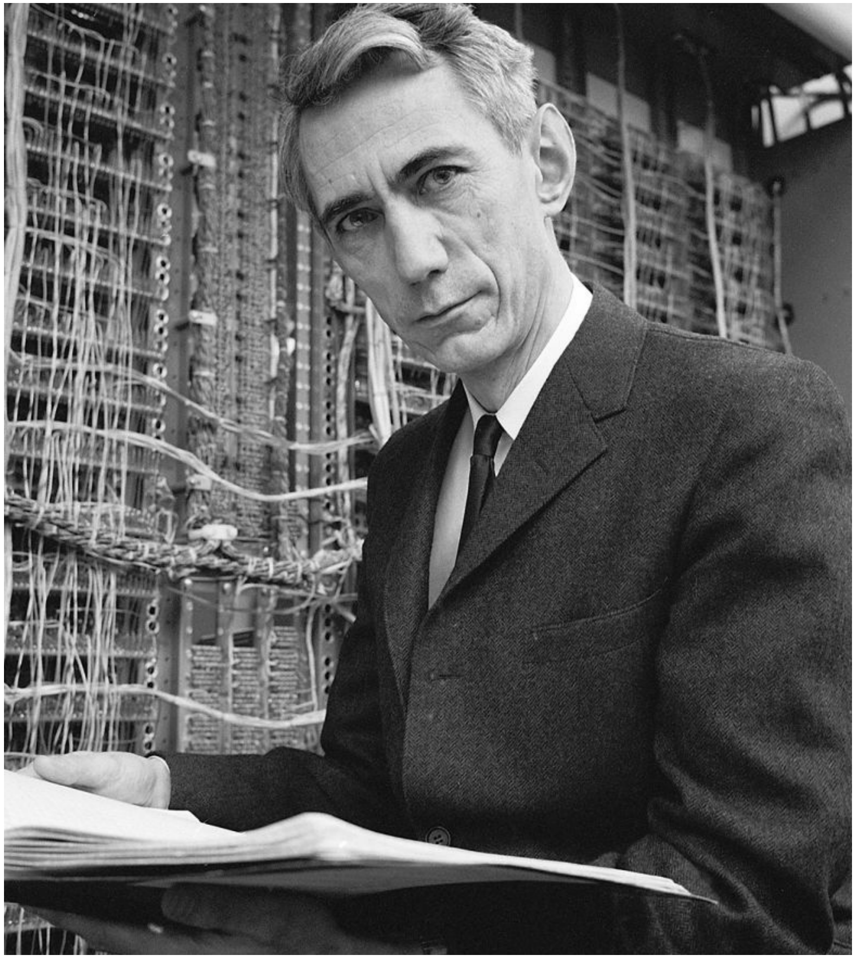
EECS Dept., University of California, Berkeley

Shannon Centenary Day, IIT-Kanpur

October 19, 2016

Shannon’s incredible legacy

- A mathematical theory of communication
- Channel capacity
- Source coding
- Channel coding
- Cryptography
- Sampling theory
- …



(1916-2001)

And many more…

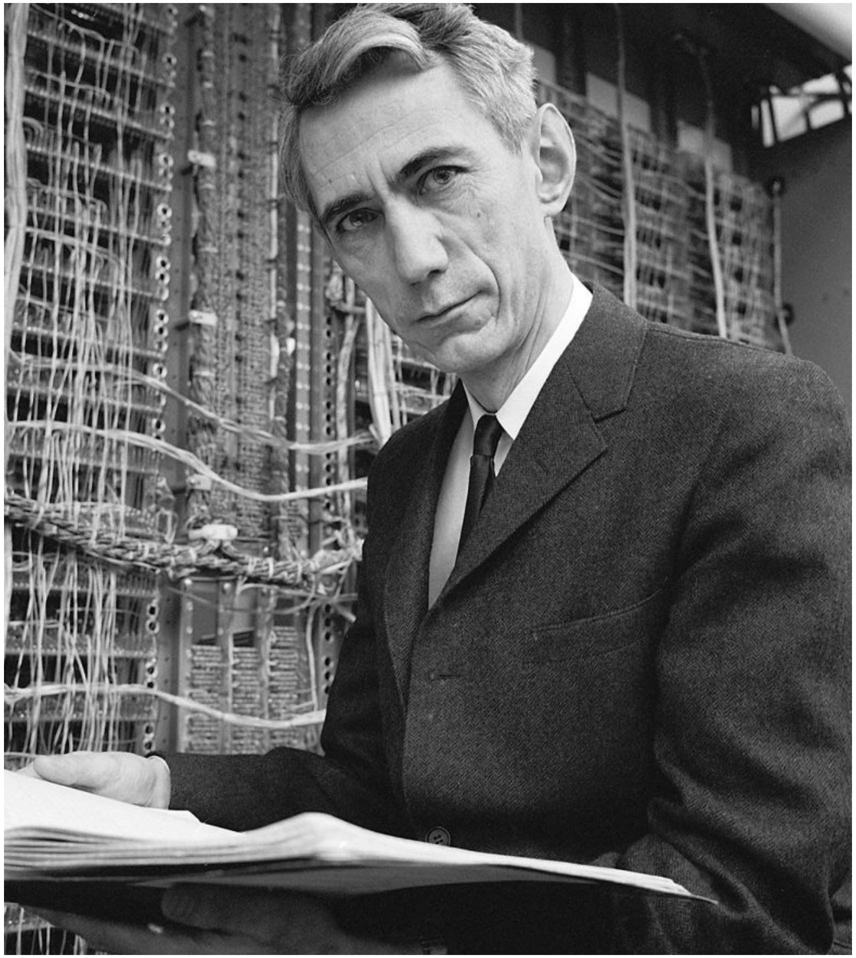
- Boolean logic for switching circuits
  (MS thesis 1937)

- Juggling theorem:
  (F+D)H=(V+D)N

F: the time a ball spends in the air
D: the time a ball spends in a hand, V: the time a hand is vacant
N: the number of balls juggled
H: the number of hands.

- The Ultimate Machine:
  https://www.youtube.com/watch?v=cZ34RDn34Ws

- …

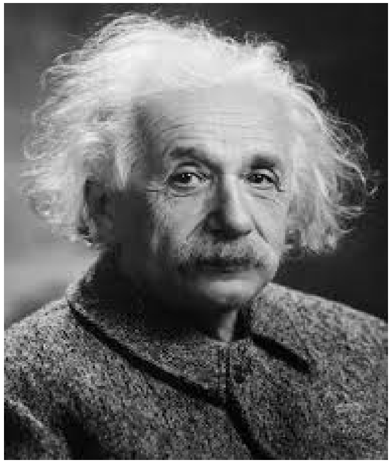


(1916-2001)

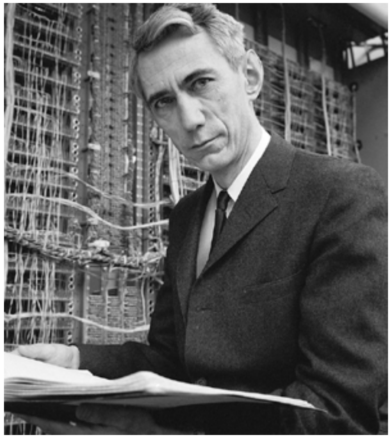
Story: Shannon meets Einstein

As narrated by Arthur Lewbell (2001):

“The story is that Claude was in the middle of giving a lecture to mathematicians in Princeton, when the door in the back of the room opens, and in walks **Albert Einstein.**



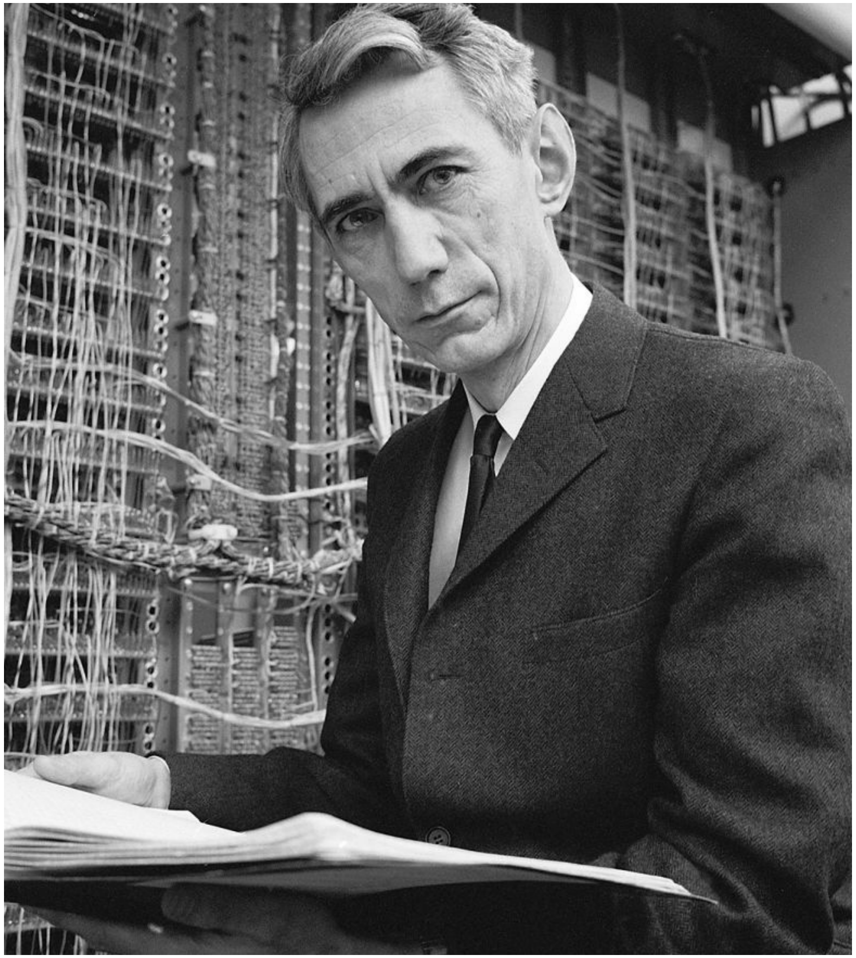
Einstein stands listening for a few minutes, whispers something in the ear of someone in the back of the room, and leaves. At the end of the lecture, Claude hurries to the back of the room to find the person that Einstein had whispered too, to find out what the great man had to say about his work.



The answer: Einstein had asked directions to the men’s room.”

Shannon's incredible legacy

- A mathematical theory of communication
- Channel capacity
- Source coding
- Channel coding
- Cryptography
- Sampling theory
- …



(1916-2001)

Outline: Four "personal" research stories

- *Chapter 1*: **Duality** between source coding and channel coding – with side-information (2002)

- *Chapter 2*: **Encryption** and **Compression** – swapping the order (2003)

- *Chapter 3*: **Sampling** theory meets **Coding** theory – spectrum-blind sampling (2015)

- *Chapter 4*: **Distributed Storage** for massive data centers: network coding (2010)

Sandeep Pradhan

Jim Chou

Chapter 1

Duality

- source coding & channel coding

with side-information

Shannon’s celebrated 1948 paper

The Bell System Technical Journal

Vol. XXVII *July, 1948* *No. 3*

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance

[1] Nyquist, H., “Certain Factors Affecting Telegraph Speed,” *Bell System Technical Journal*, April 1924, p. 324; “Certain Topics in Telegraph Transmission Theory,” *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.

[2] Hartley, R. V. L., “Transmission of Information,” *Bell System Technical Journal*, July 1928, p. 535.

379

general theory of communication

communication system as source/channel/destination

abstraction of the concept of message

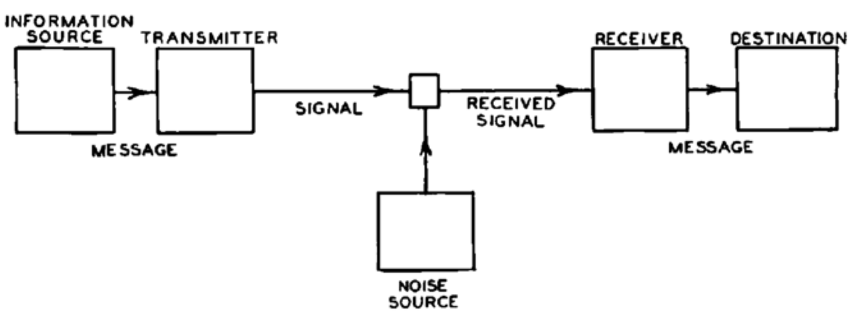
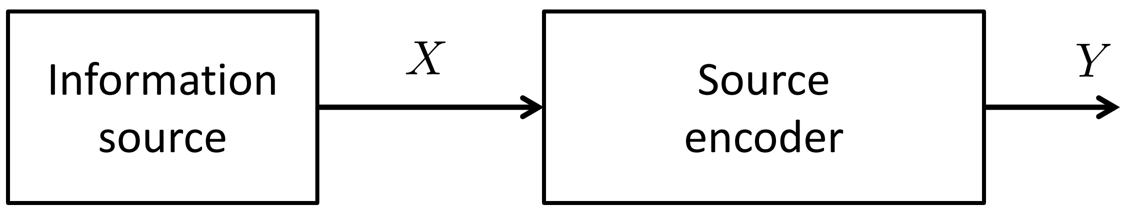


Fig. 1—Schematic diagram of a general communication system.

Source coding



$$H(X) = \mathbb{E}_X\left[\log\left(\frac{1}{p(X)}\right)\right]$$

Entropy of a random variable
= minimum number of bits required to represent the source

Rate-distortion theory - 1948

- Trade-off between lossy compression rate and the distortion

PART V: THE RATE FOR A CONTINUOUS SOURCE

27. Fidelity Evaluation Functions

In the case of a discrete source of information we were able to determine a definite rate of generating information, namely the entropy of the underlying stochastic process. With a continuous source the situation is considerably more involved. In the first place a continuously variable quantity can assume an infinite number of values and requires, therefore, an infinite number of binary digits for exact specification. This means that to transmit the output of a continuous source with *exact recovery* at the receiving point requires, in general, a channel of infinite capacity (in bits per second). Since, ordinarily, channels have a certain amount of noise, and therefore a finite capacity, exact transmission is impossible.

This, however, evades the real issue. Practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a certain tolerance. The question is, can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way. Of course, as the fidelity require-

$$H(X) - H(X|Y)$$

$$\min_{P_{Y|X}(y|x)} \quad I(X;Y)$$

$$\text{subject to} \quad D_P(Y,X) \le D^*$$

distortion measure

Channel coding

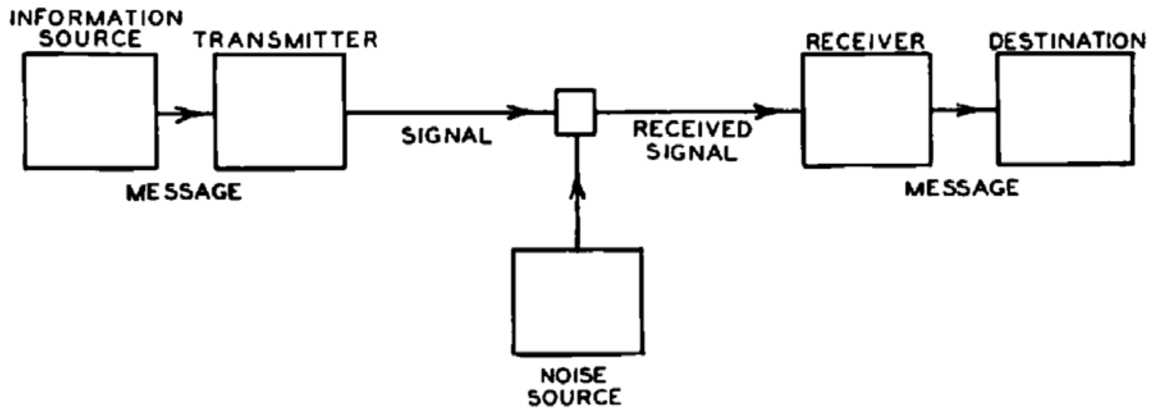


Fig. 1—Schematic diagram of a general communication system.

capacity

$$C = \max_{P_X(x)} I(X;Y)$$

- For rates R < C, can achieve arbitrary small error probabilities
- Used to be thought one needs R → 0

Shannon (1959)

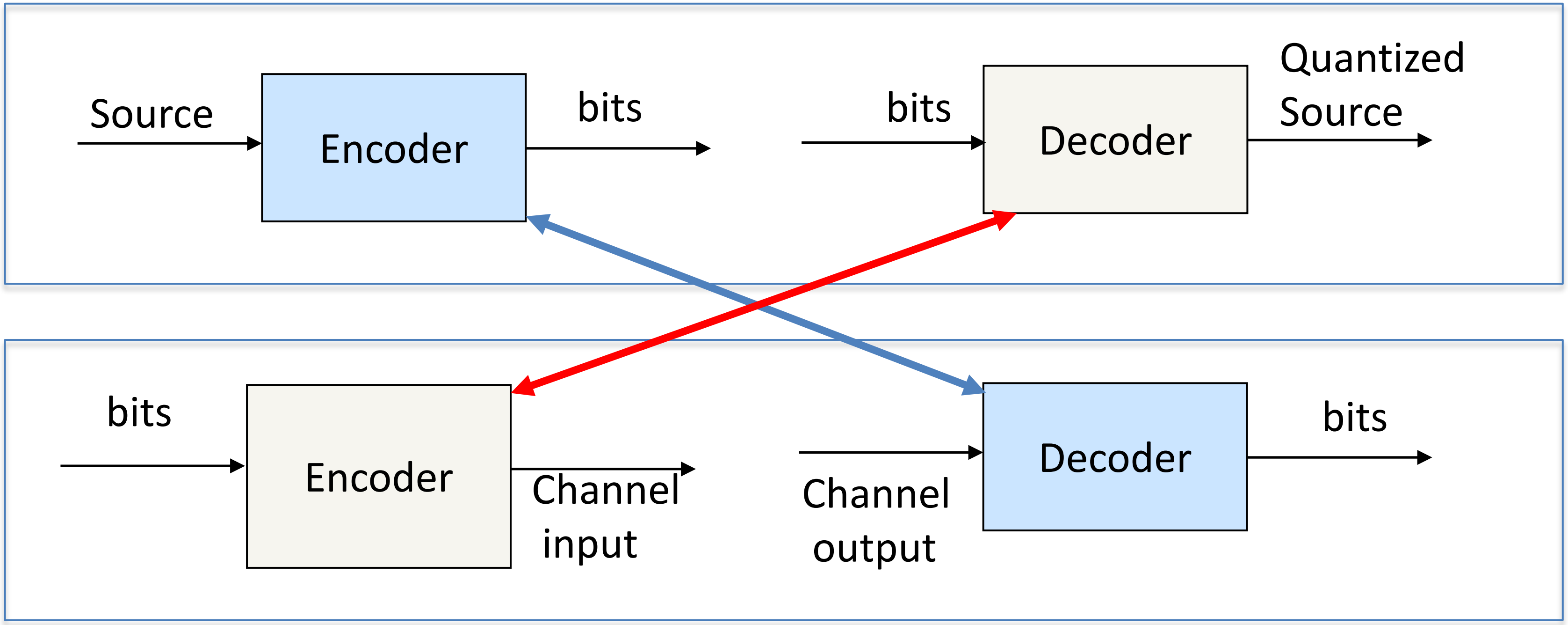
“***There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel.*** This duality is enhanced if we consider channels in which there is a cost associated with the different input letters, and it is desired to find the capacity subject to the constraint that the expected cost not exceed a certain quantity…..

Shannon (1959)

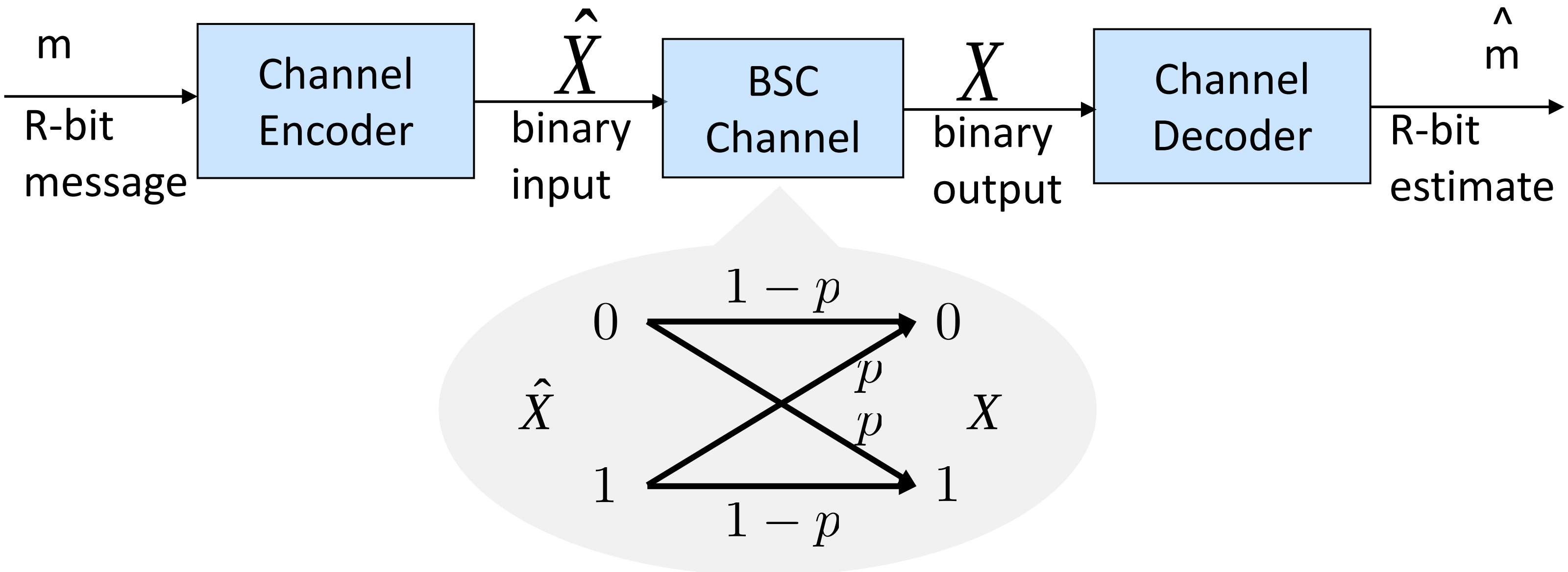
…This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. ***Thus, we may have knowledge of the past but cannot control it; we may control the future but not have knowledge of it.***”

Functional duality

- When is the ***optimal encoder*** for one problem functionally identical to the ***optimal decoder*** for the dual problem?



Duality example: Channel coding



$$p = 0.15$$

$$\text{cost}(0) = ₹0$$

You want to send message m

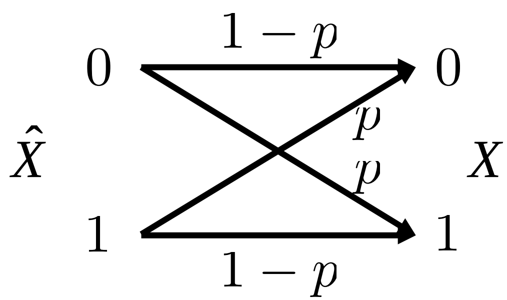
$$\text{cost}(1) = ₹1$$

$$\text{total budget} \leq ₹5,000$$

$$\#\text{channel use} = 10,000$$

How many bits R can you send?

What is the Shannon capacity?



$$\text{capacity}(\text{BSC}_p) = 1 - h(p)$$

$$h(p) = -p\log(p) - (1-p)\log(1-p)$$

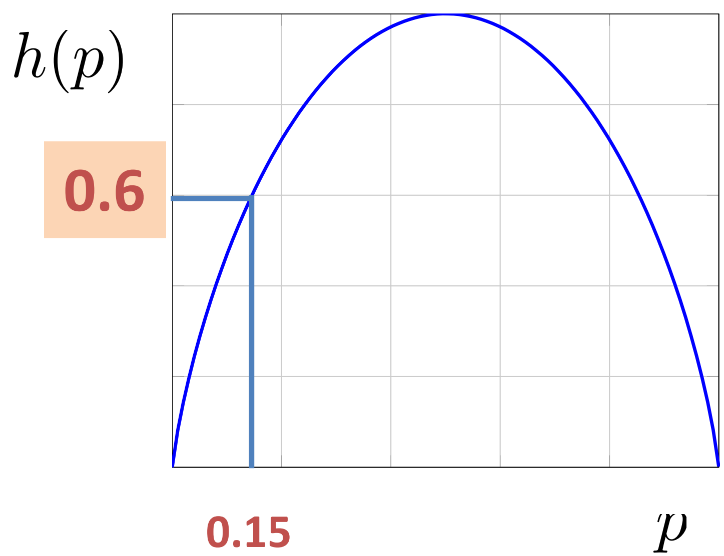
$$p = 0.15$$

$$\text{cost}(0) = ₹0$$

$$\text{cost}(1) = ₹1$$

$$\text{total budget} \leq ₹5,000$$

$$\#\text{channel use} = 10,000$$

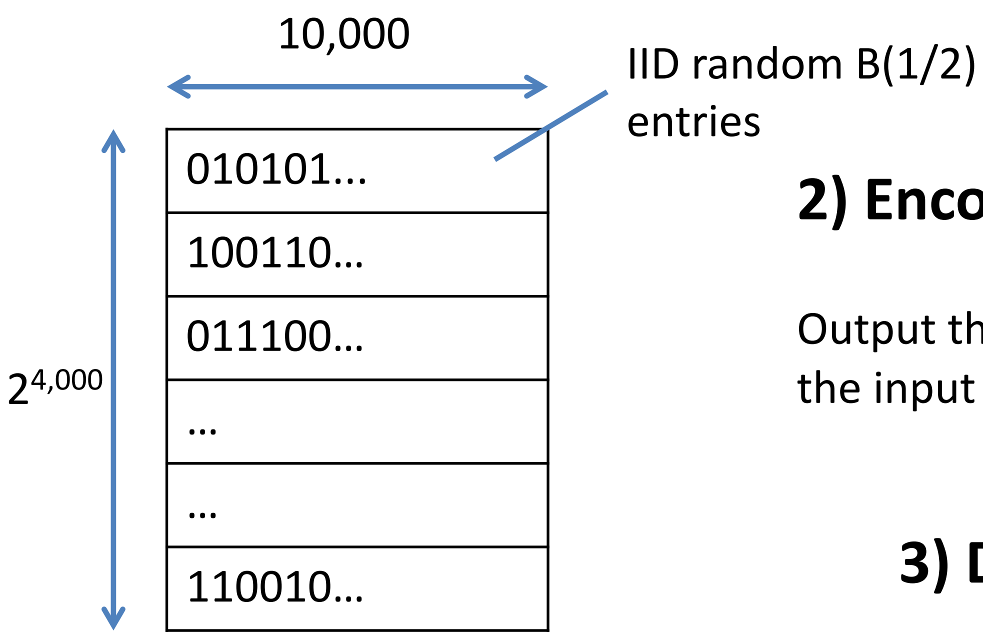


You can send 4,000 bits by using the channel 10,000 times!

How would you do it?

1) Create a codebook

Shannon's random coding argument



Codebook **C** for **channel encoding**

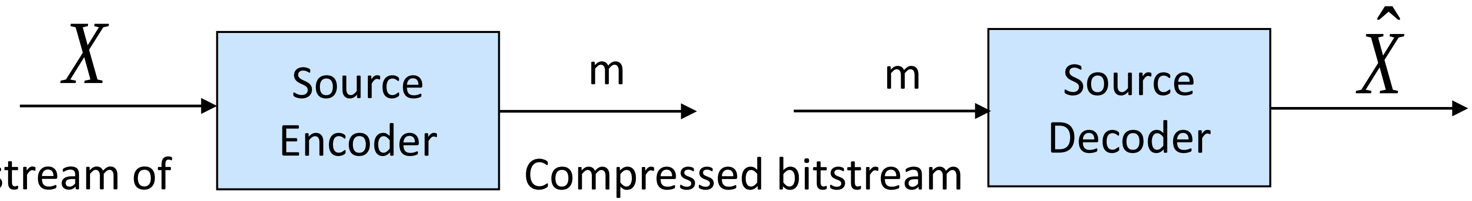
2) Encode your message

Output the codeword in **C** corresponding to the input index

3) Decode your message

Output the index corresponding to the codeword in **C** that is "closest" to the input word

Dual source coding problem



Bitstream of
length 10,000 bits
Distributed B(1/2)

Compressed bitstream
4,000 bits

Want the Hamming distortion to be less than 0.15

$$\frac{1}{10,000} d_H(X, \hat{X}) \leq 0.15$$

How would you do it?

Use **channel decoder** as **source encoder**

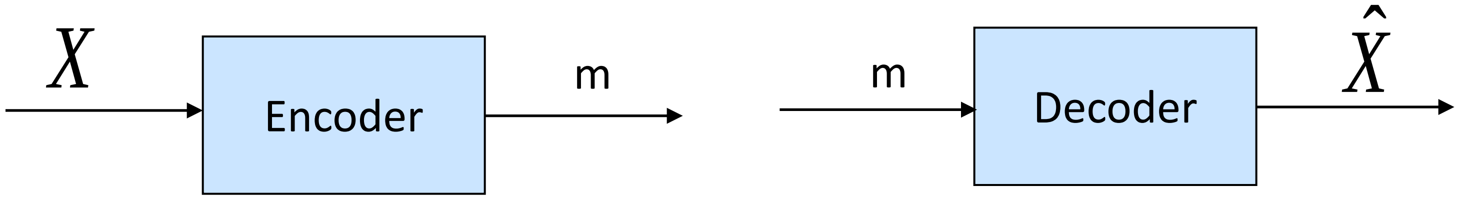
Output the index corresponding to the codeword in **C** "closest" to the input word

010101…
100110…
011100…
…
…
110010…

Use **channel encoder** as **source decoder**

Output the codeword in **C** corresponding to the input index

Source coding



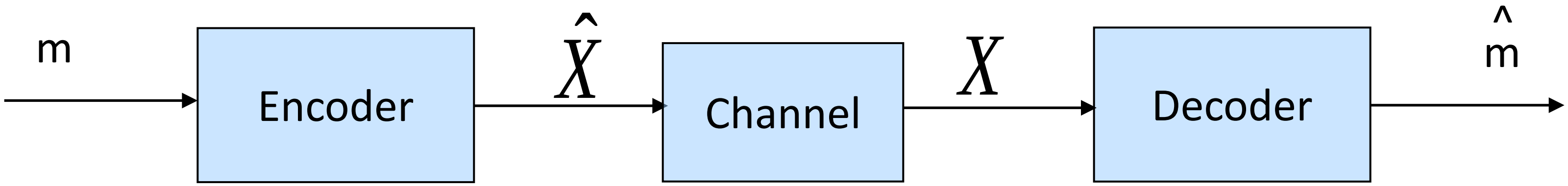
Source distribution: $\overline{p}(x)$

Distortion measure $d(x,\hat{x}): X \times \hat{X} \rightarrow \Re^{+}$

Distortion constraint D: $Ed(x,\hat{x}) \leq D$

Rate-distortion function R(D)= $\min\limits_{p(\hat{x}\mid x)} I(X;\hat{X})$

Channel coding



Channel description: $\bar{p}(x \mid \hat{x})$

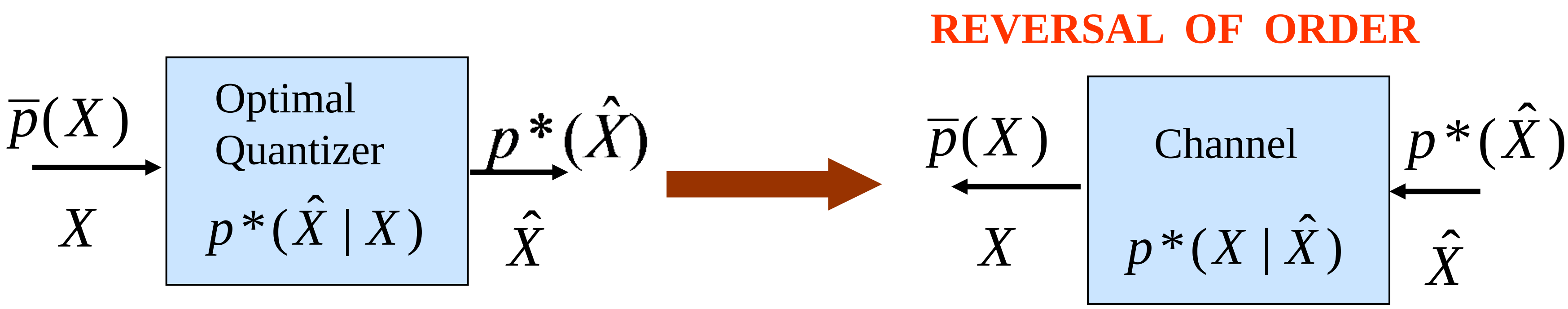
Cost measure $w(\hat{x}) : \hat{X} \rightarrow \Re^{+}$

Cost constraint: $Ew(\hat{x}) \leq W$

Capacity-cost function C(W)= $\max\limits_{p(\hat{x})} I(X;\hat{X})$

- **Source Encoder** & **Channel Decoder** have the same domain and range.
- **Channel Encoder** & **Source Decoder** have the same domain and range.

Duality between source and channel coding:



Given a source coding problem with source distr. $\overline{p}(X)$, optimal quantizer $p^*(\hat{X} \mid X)$ distortion measure $d(x, \hat{x})$ and distortion constraint **D**, (left) ,

$\exists$ a **dual** channel coding problem with channel $p^*(x \mid \hat{x})$, cost measure $w(\hat{x})$, and cost constraint **W** (right) s.t.:

(i) R(**D**)=C(**W**);

(ii) $$p^*(\hat{x}) = \underset{p(\hat{x}):X|\hat{X} \sim p^*(x|\hat{x}), Ew \le W}{\arg\max} I(X; \hat{X}),$$

where $\boxed{w(\hat{x}) \overset{\Delta}{=} c_1 D(p^*(x \mid \hat{x}) \,\|\, \overline{p}(x)) + \theta}$ and $\boxed{W = E_{p^*(\hat{x})} w(\hat{X}).}$

Interpretation of functional duality

For *any* given source coding prob., $\exists$ a **dual** channel coding prob. s.t.

- both problems induce the **same optimal joint distr.** $p^*(x, \hat{x})$

- the **optimal encoder** for one is **functionally identical** to the **optimal decoder** for the other in the limit of large block length

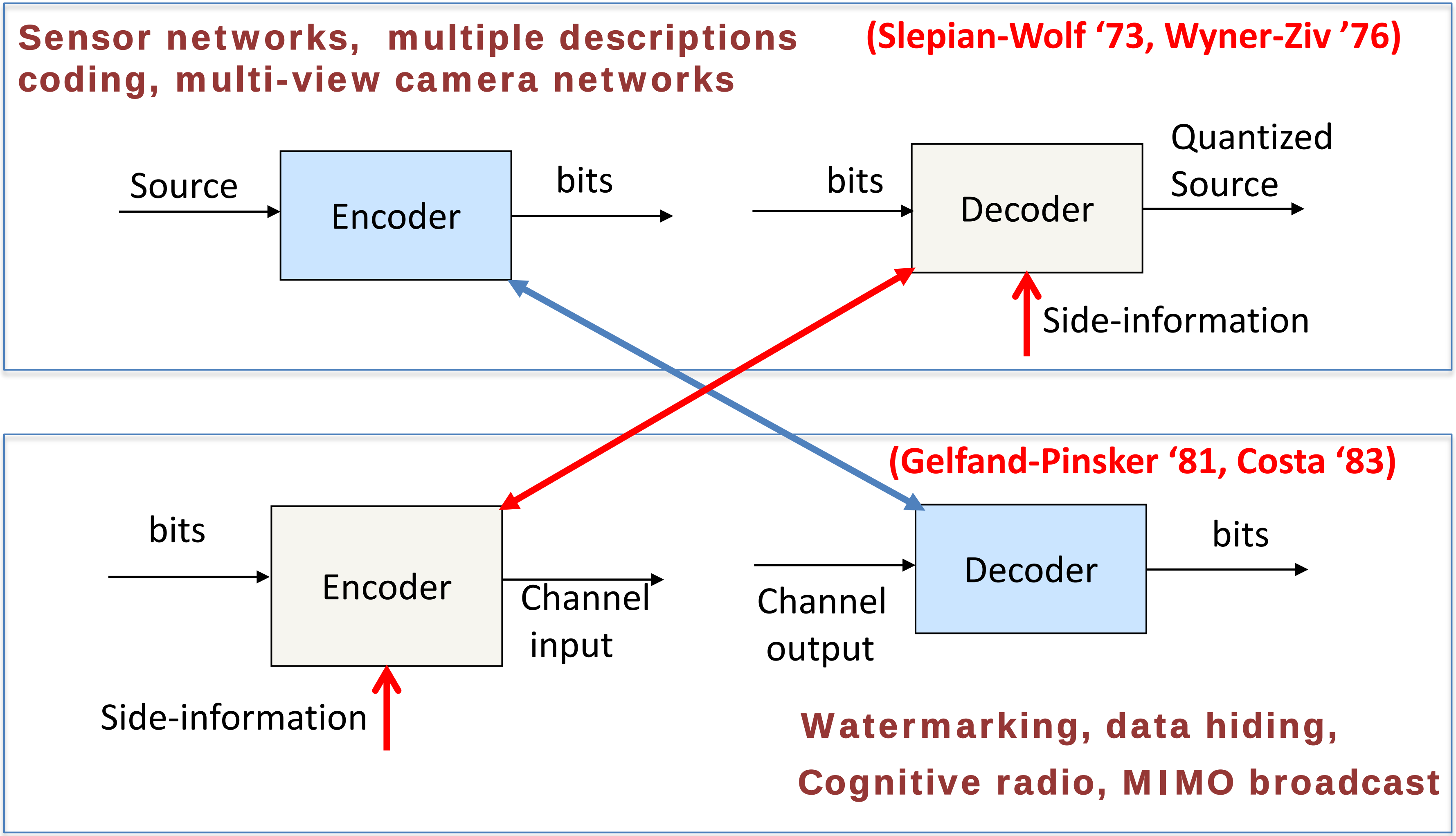
- an appropriate **channel-cost measure** is associated

Key takeaway:

Source coding: **distortion measure** is as important as the **source distribution**

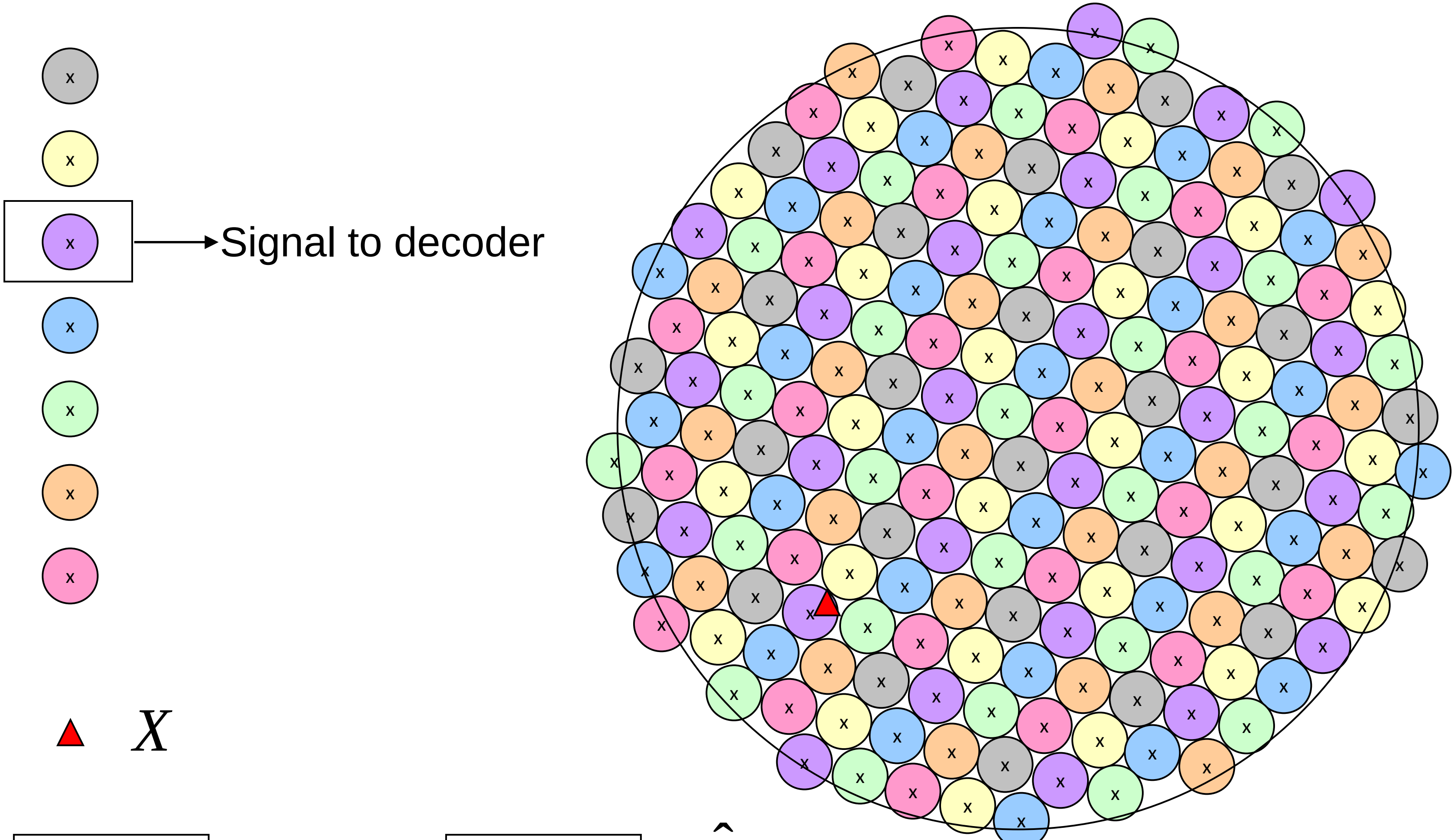
Channel coding: **channel cost measure** is as imp. as the **channel conditional dist.**

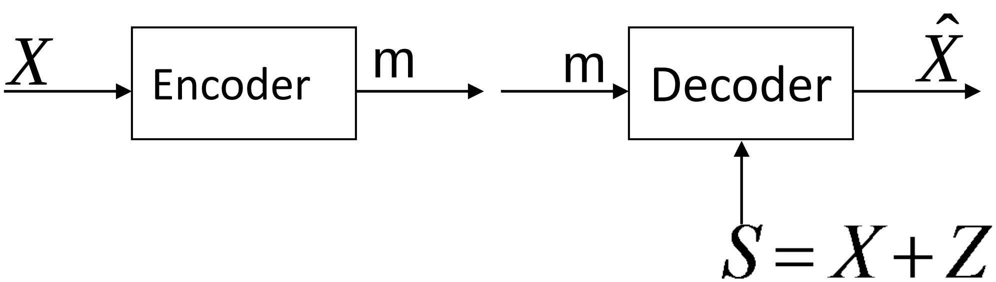
DUALITY BETWEEN SCSI & CCSI



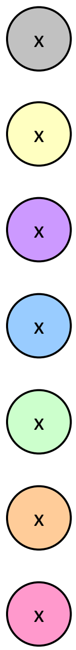
Pradhan, Chou and R, 2003

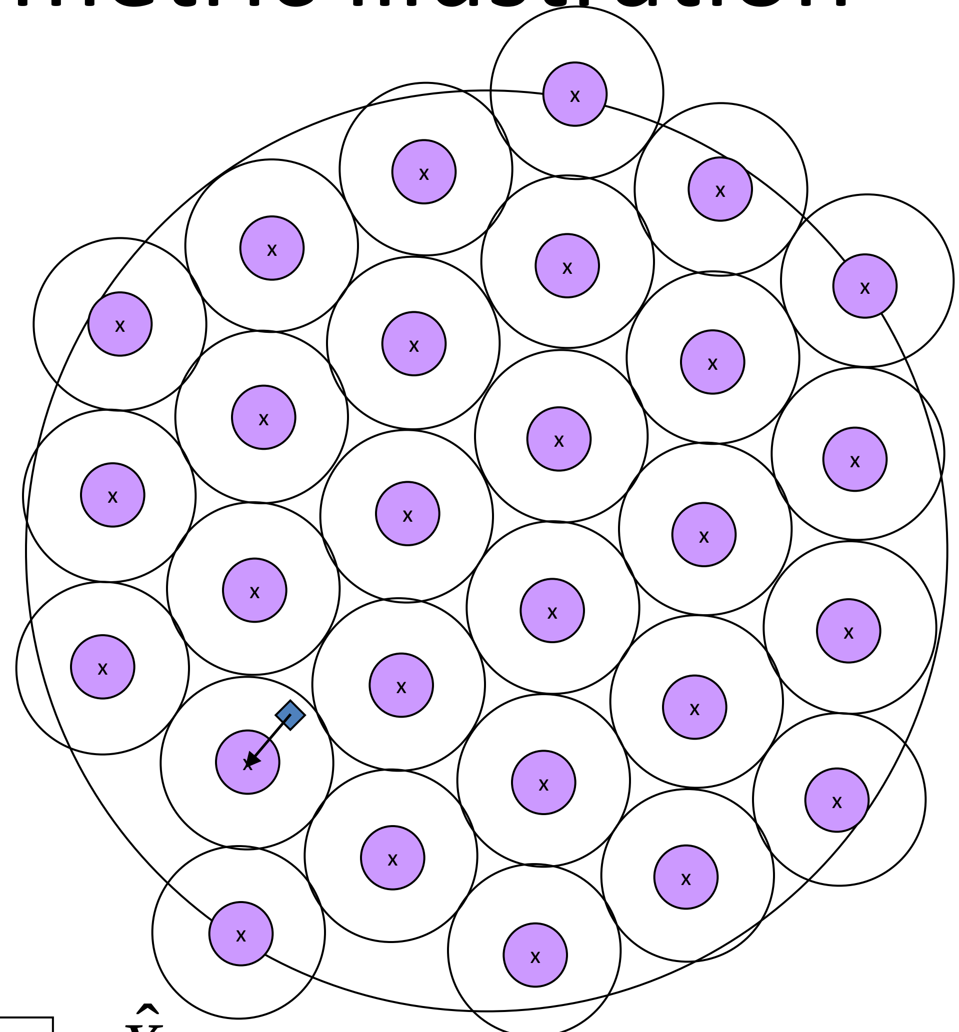
Geometric illustration of SCSI





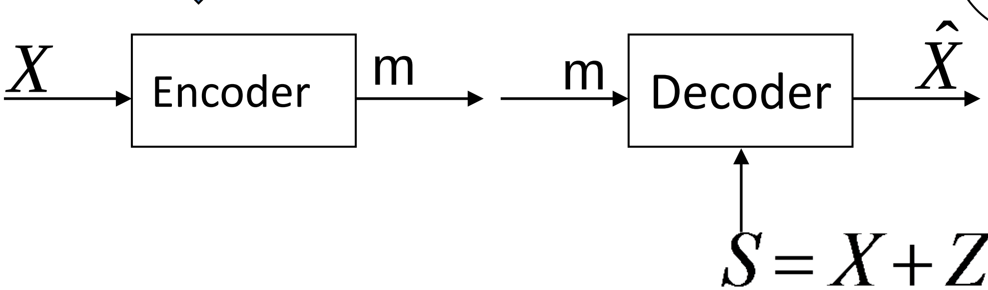
Example: geometric illustration





X

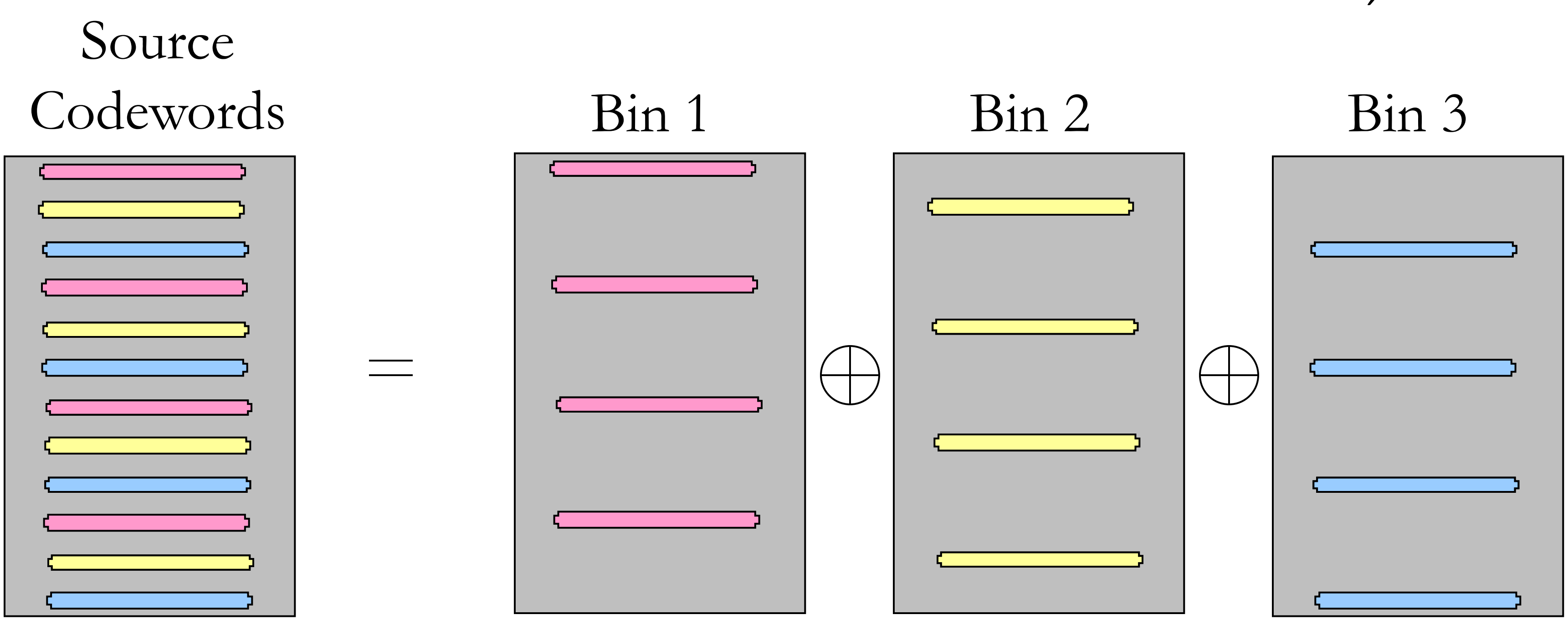
Side information



Practical Code Constructions

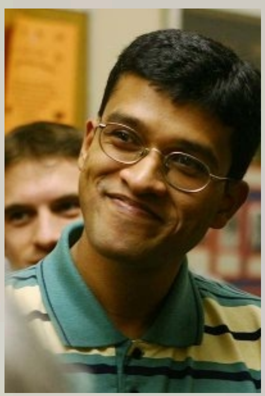
- Use a linear transformation (hash/bin)
- Design cosets to have maximal spacing
  - State of the art linear codes (LDPC codes)
- Distributed Source Coding Using Syndromes (DISCUS)*

**Pradhan & R, '03*



Mark Johnson

Prakash Ishwar



Vinod Prabhakaran

Chapter 2

Cryptography

- Compressing encrypted data

Cryptography – 1949

- Foundations of ***modern cryptography***
- All theoretically unbreakable ciphers must have the properties of one-time pad

Communication Theory of Secrecy Systems*

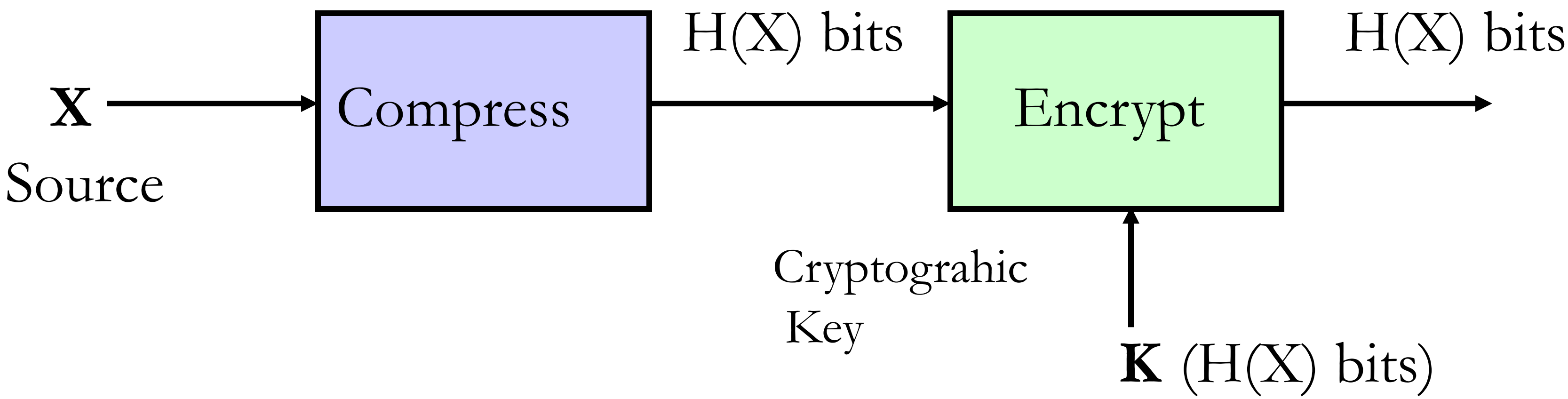
By C. E. SHANNON

1. INTRODUCTION AND SUMMARY

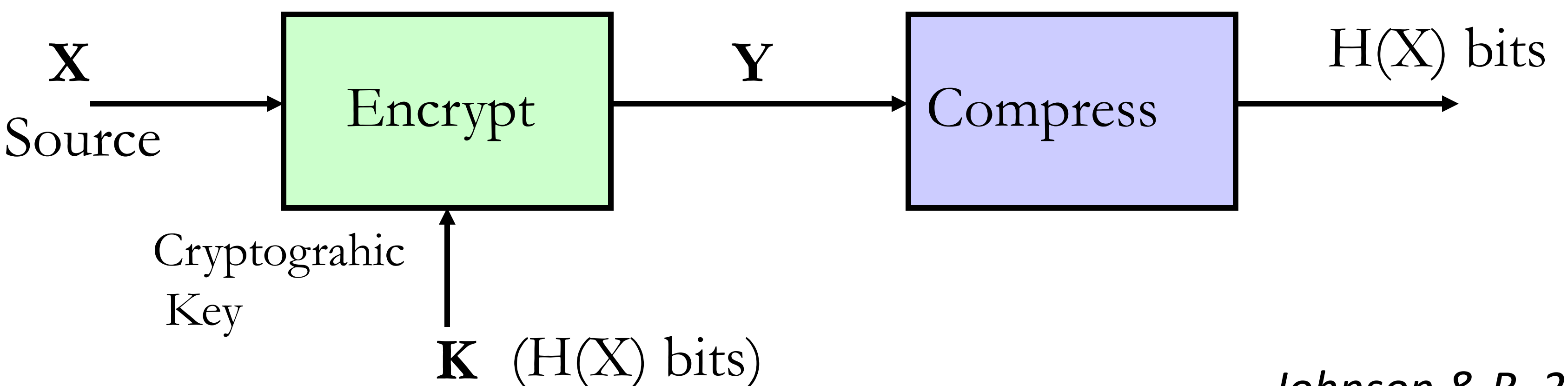
THE problems of cryptography and secrecy systems furnish an interesting application of communication theory.[1] In this paper a theory of secrecy systems is developed. The approach is on a theoretical level and is intended to complement the treatment found in standard works on cryptography.[2] There, a detailed study is made of the many standard types of codes and ciphers, and of the ways of breaking them. We will be more concerned with the general mathematical structure and properties of secrecy systems.

Compressing Encrypted Data

“Correct” order



Wrong order?



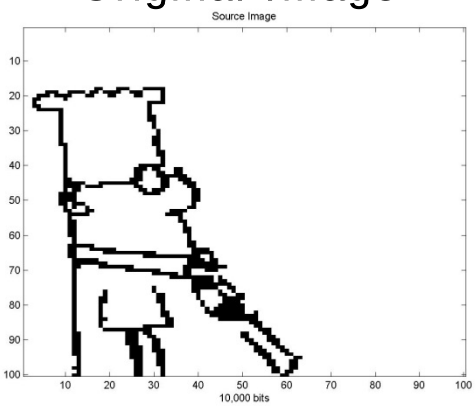
Johnson & R, 2003

Example

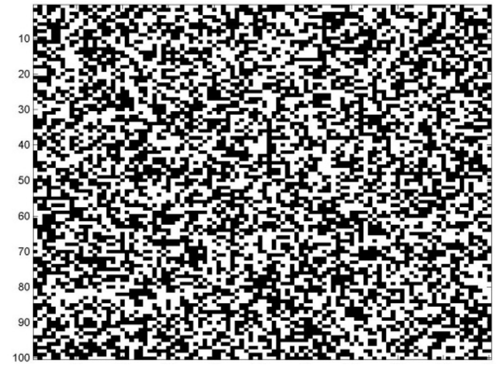
10,000 bits

5,000 bits

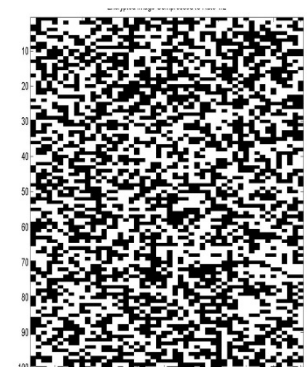
Original Image



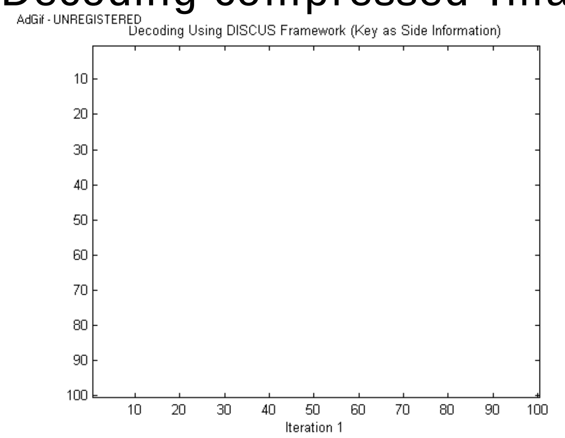
Encrypted Image



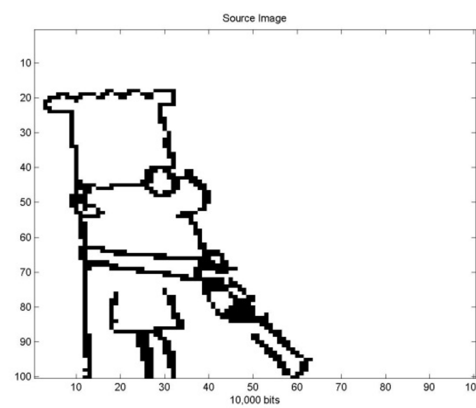
Compressed Encrypted Image

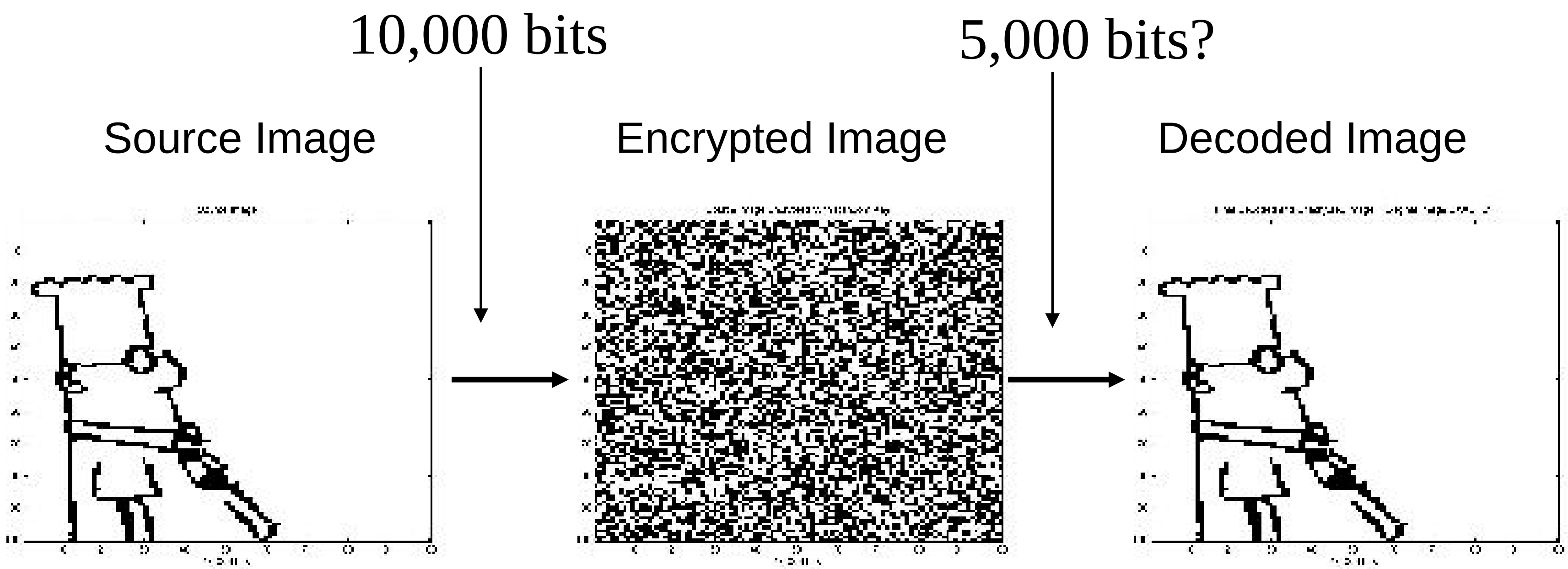


Decoding compressed Image

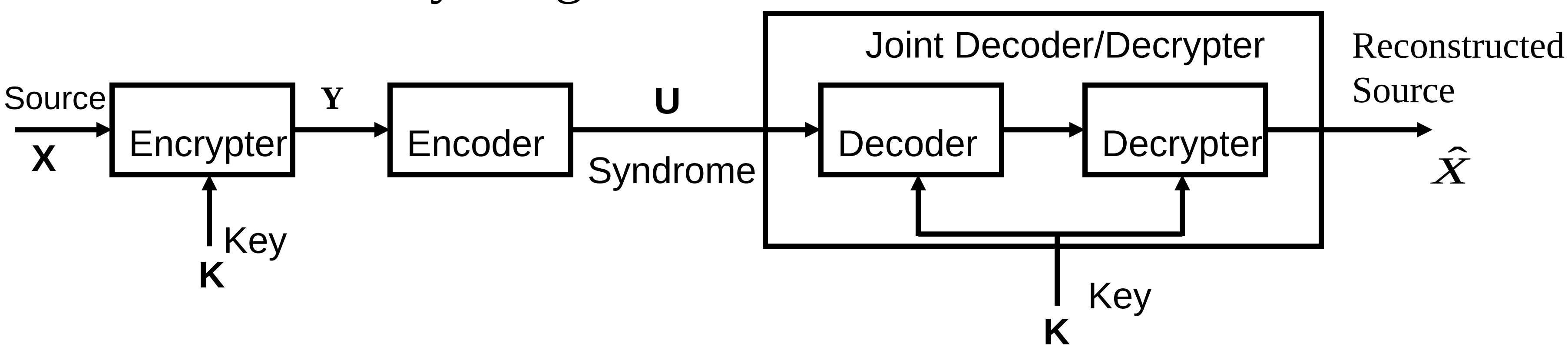


Final Reconstructed Image




10,000 bits
5,000 bits?
Source Image
Encrypted Image
Decoded Image

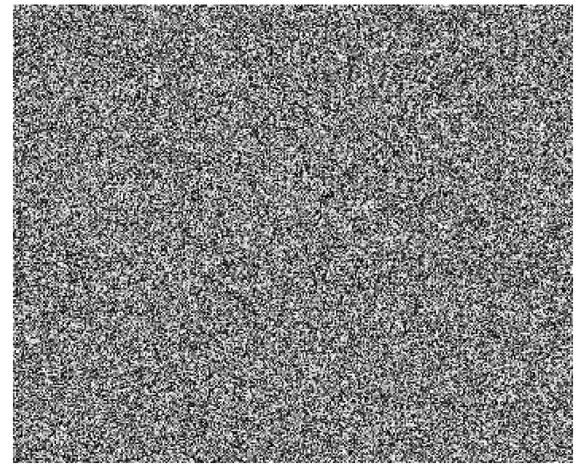
Key Insight!


Source
X
Encrypter
Key
K
Y
Encoder
U
Syndrome
Joint Decoder/Decrypter
Decoder
Decrypter
Key
K
Reconstructed
Source
$\hat{X}$

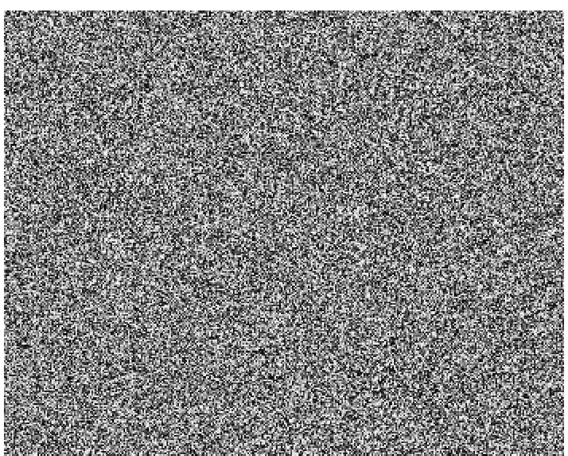
Compression of encrypted video

- Video offers both temporal and spatial prediction
  - Decoder has access to unencrypted prior frames

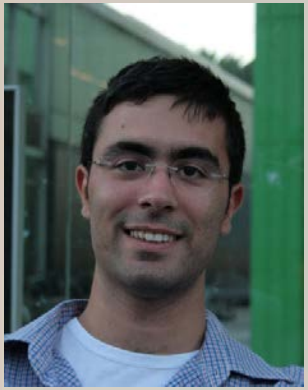
$\oplus$



=



Saves 33.00%



Orhan Ocal

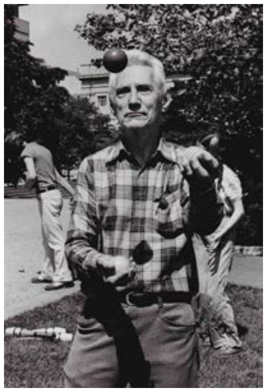
Xiao Li

Chapter 3

Sampling

- Sampling theory & coding theory: an unexplored union

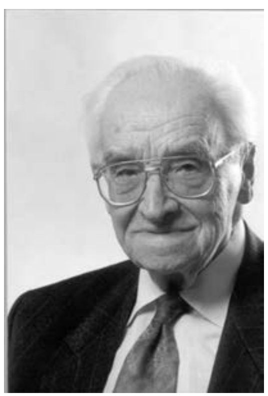
Sampling theorem



Shannon
1949

Nyquist
1928

Whittaker
1915



Kotelnikov
1933

Communication in the Presence of Noise

CLAUDE E. SHANNON, MEMBER, IRE

Theorem 1: If a function $f(t)$ contains no frequencies higher than W cps, it is completely determined by giving its ordinates at a series of points spaced $1/2\ W$ seconds apart.

pointwise sampling!

...

Mathematically, this process can be described as follows. Let x_n be the nth sample. Then the function $f(t)$ is represented by

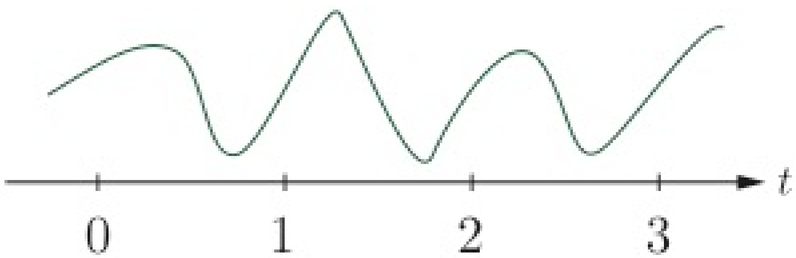
$$\boxed{f(t) = \sum_{n=-\infty}^{\infty} x_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}.} \qquad (7)$$

linear interpolation!

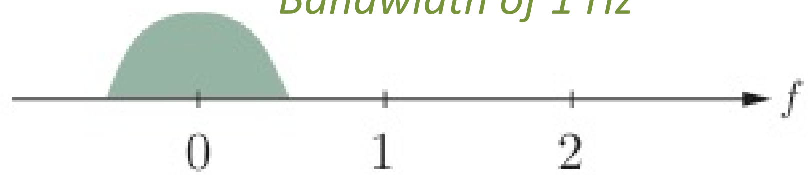
Sampling theorem illustration

Time domain | **Frequency domain**

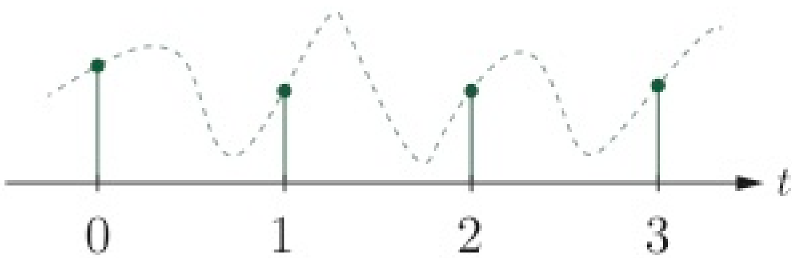
Input signal



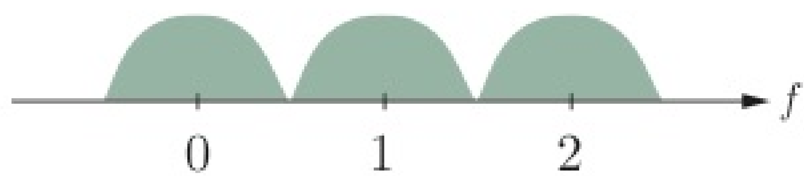
Bandwidth of 1 Hz



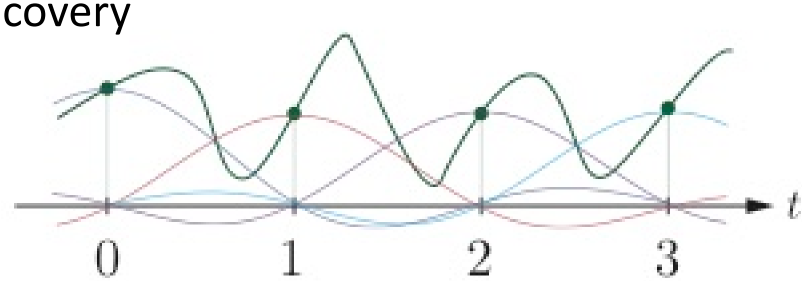
Sampling at rate 1



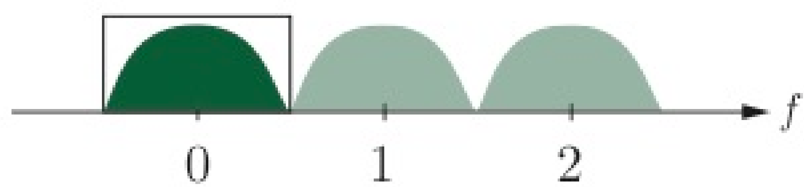
No aliasing



Recovery



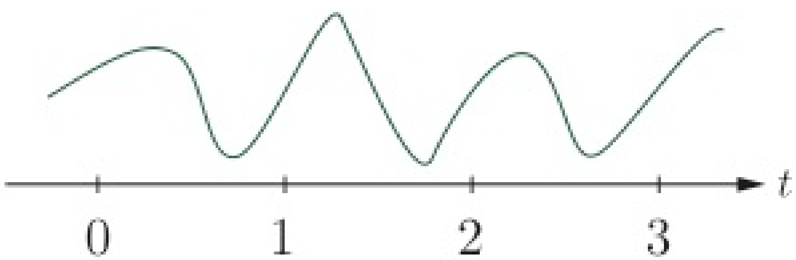
*Recovery by **linear** filtering*

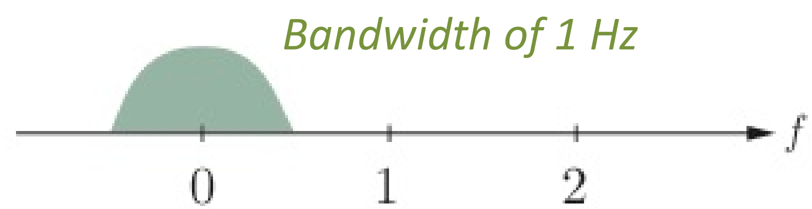


Aliasing phenomenon

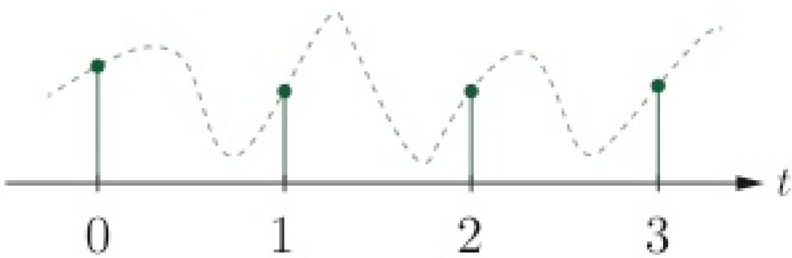
Time domain | **Frequency domain**

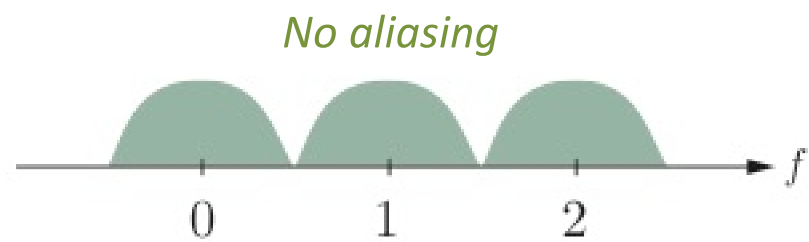
Input signal



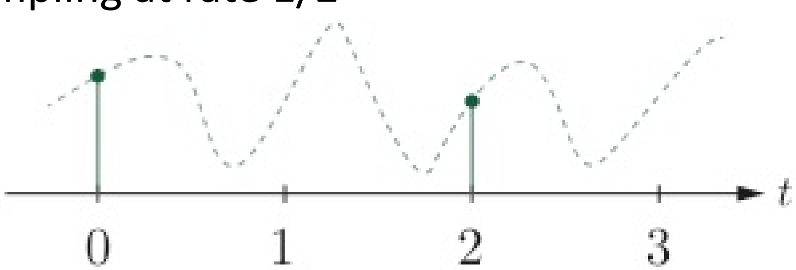


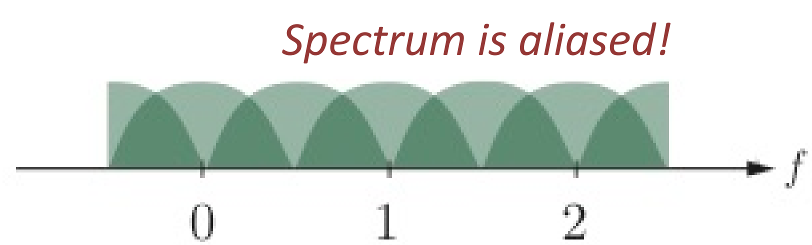
Sampling at rate 1





Sampling at rate 1/2





But what if the spectrum is sparsely occupied?

Frequency domain

f

W_1 W_2 W_3 W_4 W_5

$$f_{occ} = \sum_{i=1}^{5} W_i = 100\text{MHz}$$

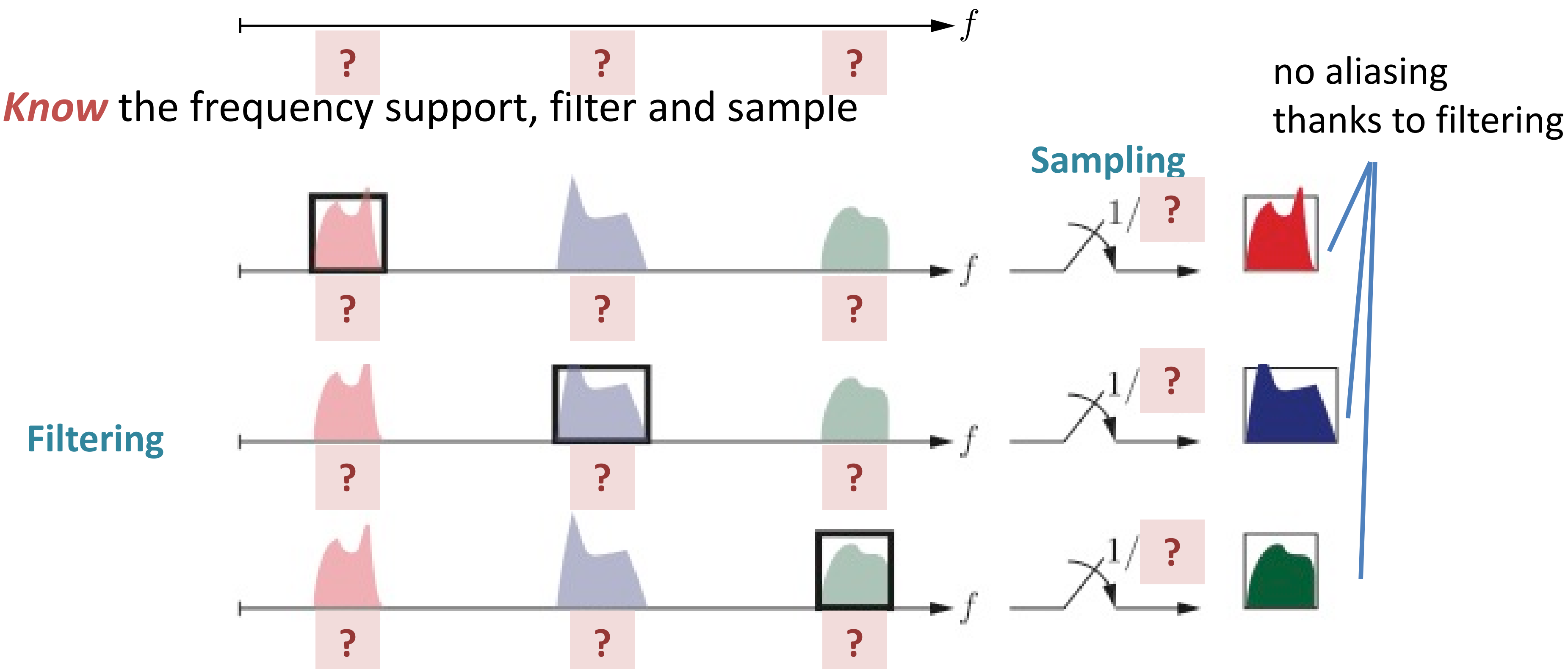
Henry Landau [1967]

- Know the frequency support
- Sample at rate *"occupied bandwidth"* f_{occ} ***(Landau rate)***

Filter bank approach

Input in frequency domain

Know the frequency support, filter and sample



Sampling ***spectrum-blind?*** Requires $2f_{OCC}$ samples (Lu &Do, '08)

Q) Can we design a constructive scheme? (Ocal, Li & R, '15)

Key insights for spectrum-blind sampling

subsampling → aliasing

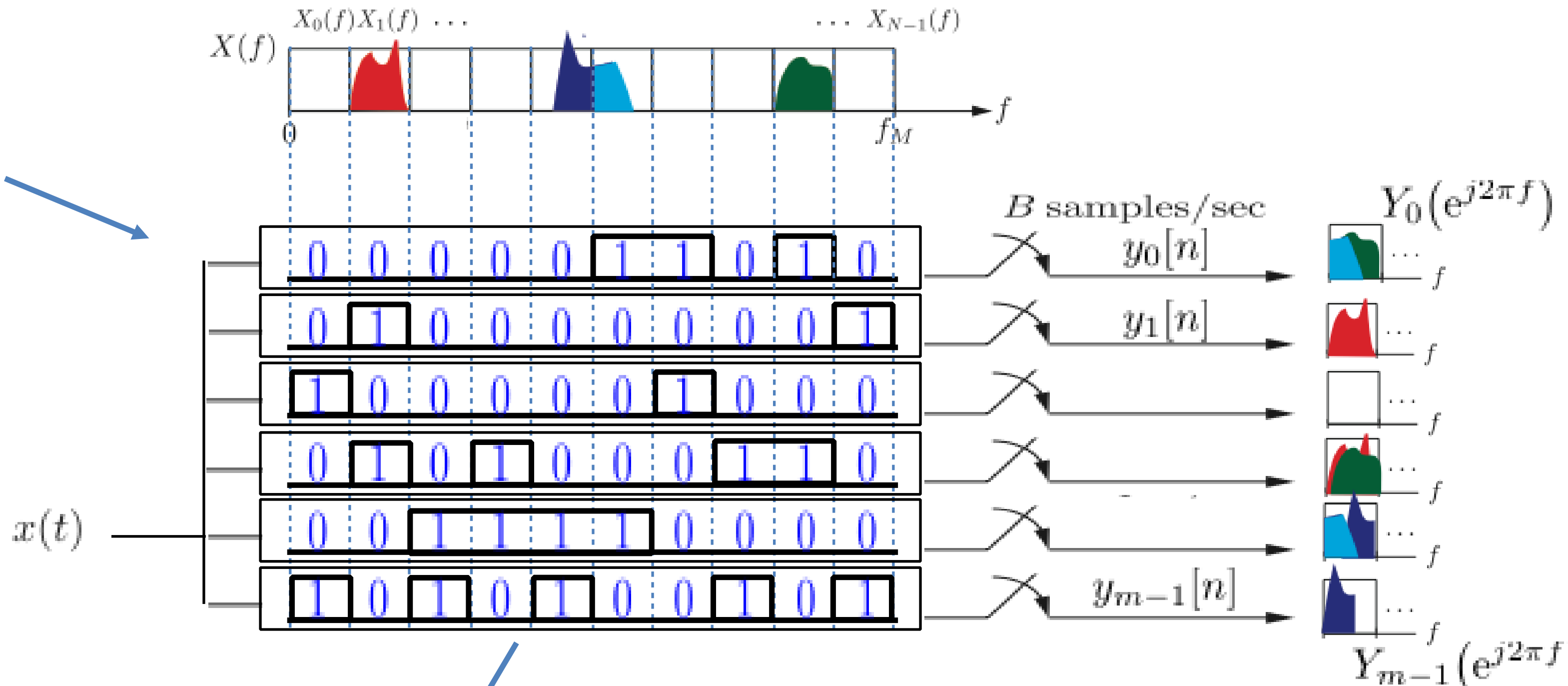
``smart" filtering/subsampling → "removable" aliasing

- No need to avoid aliasing: linear interpolation
- Just to remove it: *nonlinear channel decoding*

- ***Filter bank*** design ←→ ***capacity-achieving LDPC codes***
- Aliasing removed by non-linear ***fast peeling-decoding***

'Sparse-graph-coded' filter bank

m filters
N bands



$$\vec{Y}(e^{j2\pi f}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \vec{X}(Bf) \quad \text{where} \quad \vec{X}(f) = \begin{pmatrix} X_0(f) \\ \vdots \\ X_{n-1}(f) \end{pmatrix}$$

$m \times N$ matrix

Main result

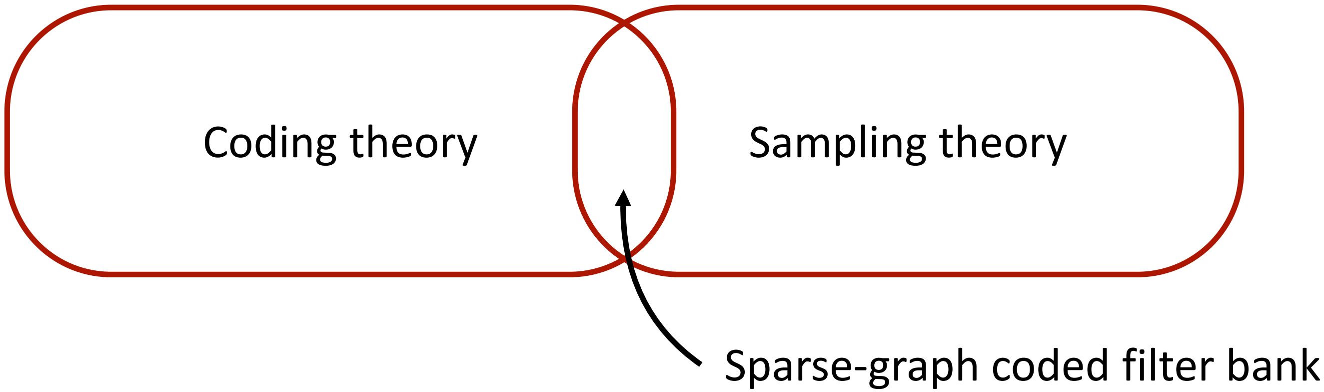
Any bandlimited signal $x(t) \in \mathbb{C}$ whose spectrum has occupancy f_{occ} can be sampled asymptotically at rate $f_s = 2f_{occ}$ by a randomized "*sparse-graph-coded filter bank*" with probability 1 using $O(f_{occ})$ operations per unit time.

Remarks

- Computational cost $O(f_{occ})$ ***independent of bandwidth***
- Requires mild assumptions (genericity)
- Can be made robust to sampling noise

Ocal, Li & R, 2016

Beautiful connection



- *Minimum-rate spectrum-blind* sampling
- *Coding theory* and *sampling theory*
  - Capacity-approaching codes for erasure channels
  - Minimum-rate blind sampling of multiband signals

Shannon's inspiration

- Pre-Shannon Communication:
  - *Linear filtering* (Wiener) at receiver to remove noise
- Post-Shannon Communication:
  - Capacity-approaching codes
  - *Non-linear estimation* (MLE) at receiver

Reliable transmission at rates approaching channel capacity

Alex Dimakis

Rashmi Vinayak

Nihar Shah

Chapter 4

Distributed Storage

- Network coding for distributed storage

The Big Data Age

Data Centers

Dropbox

Cloud Storage

Google+

facebook

twitter

Social Networks

YouTube™

hulu

NETFLIX

Video on Demand

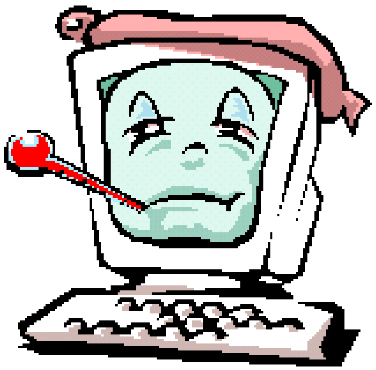
Cloud Computing

- Web search
- Recommendation sys.
- Healthcare
- Finance
- Genomics
- Particle physics,...

Distributed storage systems form the backbone of most big data applications

Fault Tolerance is Essential

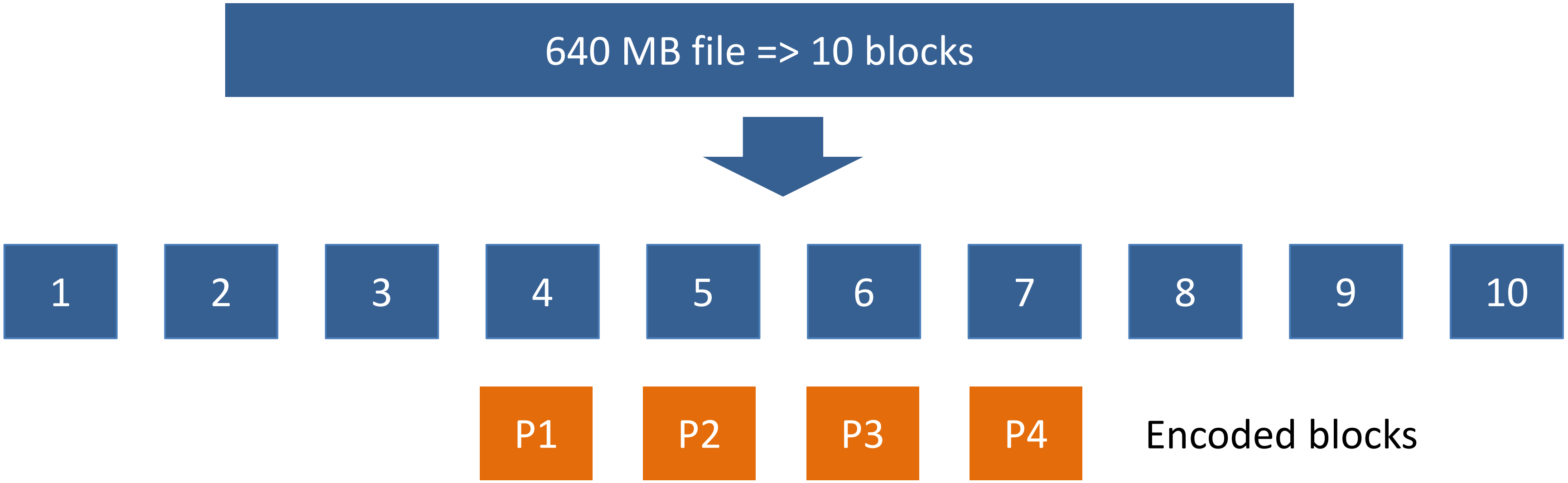
- Machines become unavailable for various reasons
  - unreliable components
  - software issues
  - power glitches
  - maintenance operations



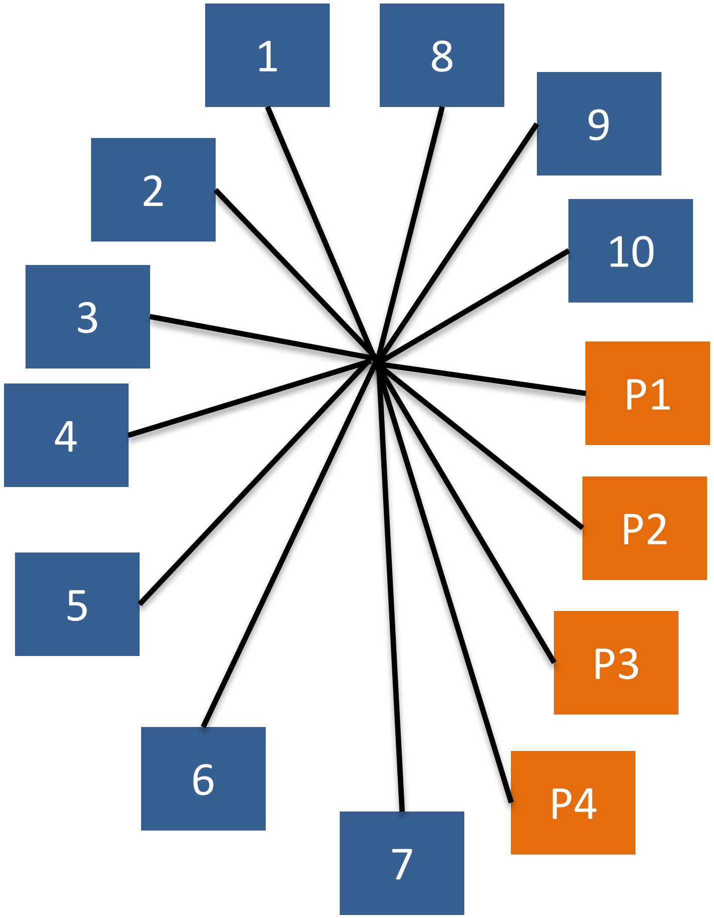
- ***Redundant storage*** needed for *data reliability* and *availability*
  - Current default solution (3x replication)
  - Storage efficiency becomes critical

MDS codes

- The most popular, and also most efficient storage codes
  - E.g. Reed-Solomon codes
- (n, k) MDS code:
  - A file is encoded into n blocks
  - From any k blocks, one can recover the file
- E.g. (14, 10) MDS code:



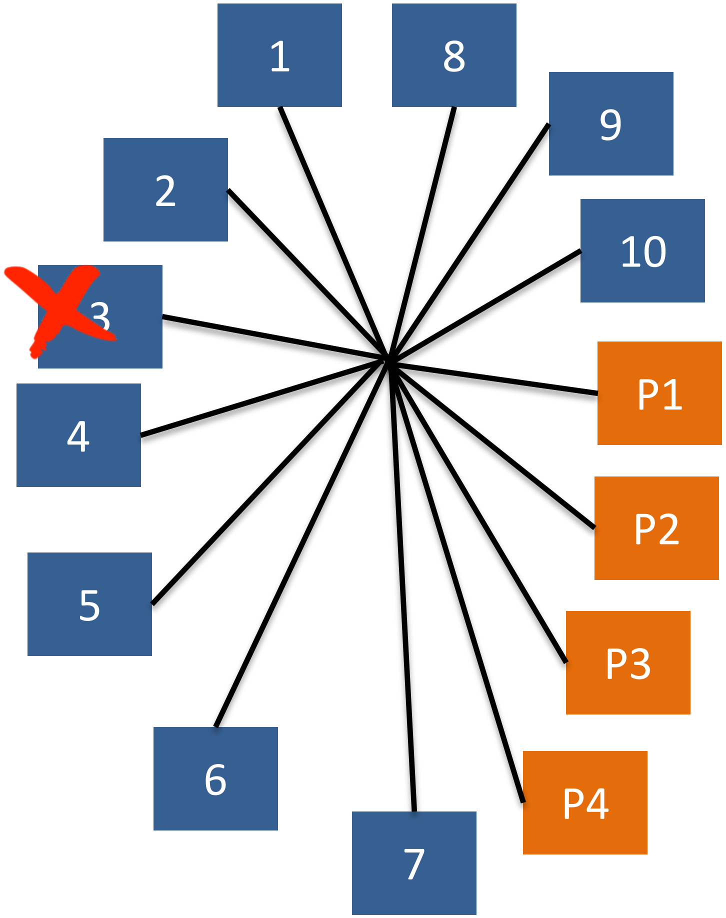
MDS codes



Good news:

We can now tolerate 4 node failures.

MDS codes

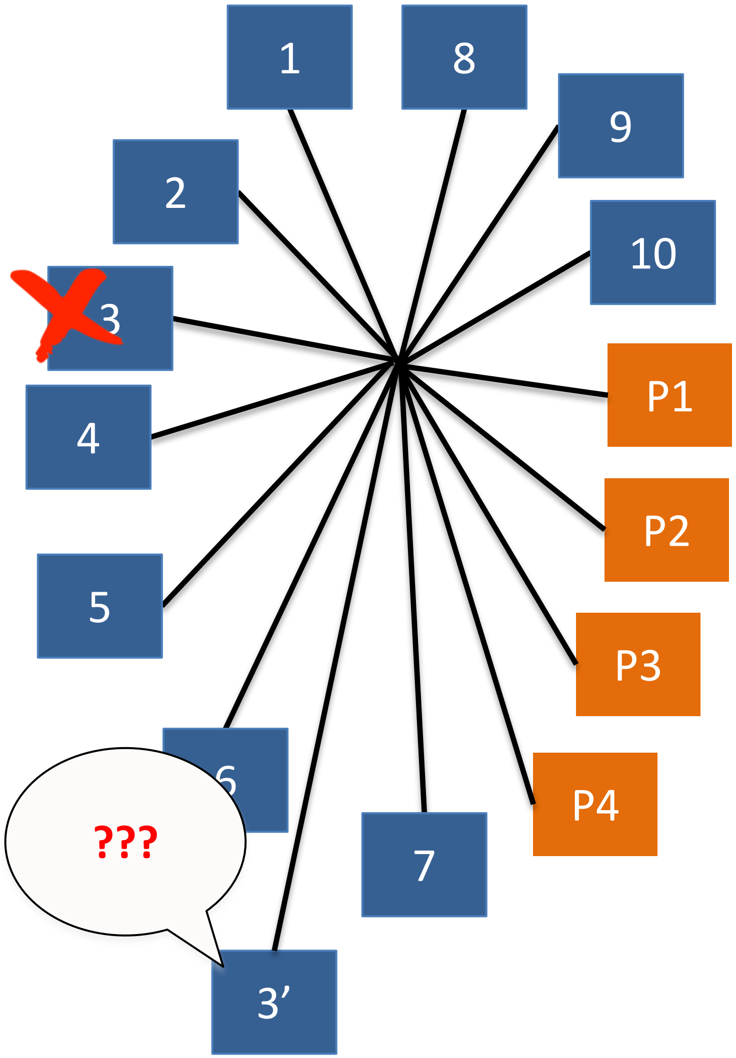


Good news:

We can now tolerate 4 node failures.

Most of the time we start with a single failure.

MDS codes

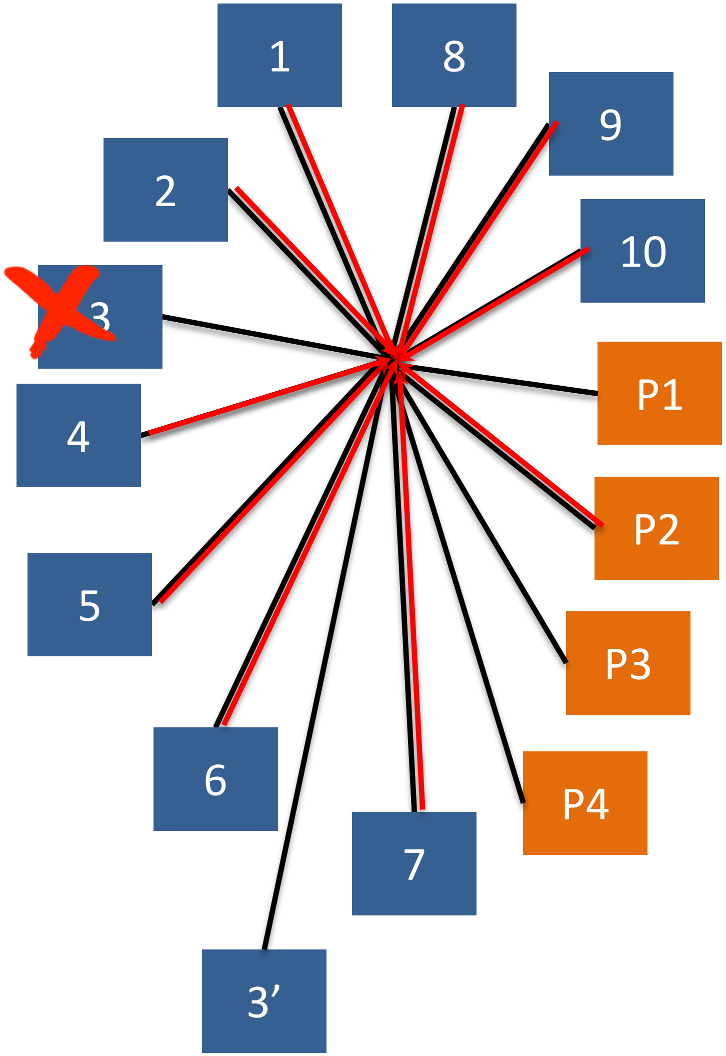


Good news:

We can now tolerate 4 node failures.

Most of the time we start with a single failure.

MDS codes



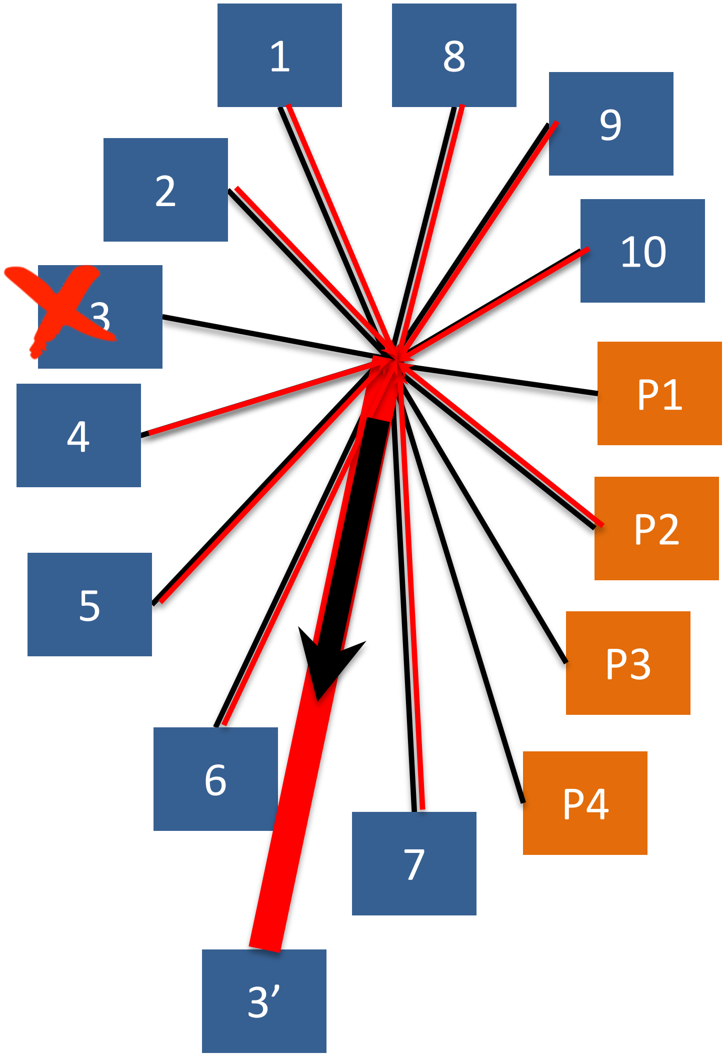
Good news:

We can now tolerate 4 node failures.

Most of the time we start with a single failure.

Read from any 10 nodes, send all data to 3’ who can repair the lost block.

MDS codes



Good news:

We can now tolerate 4 node failures.

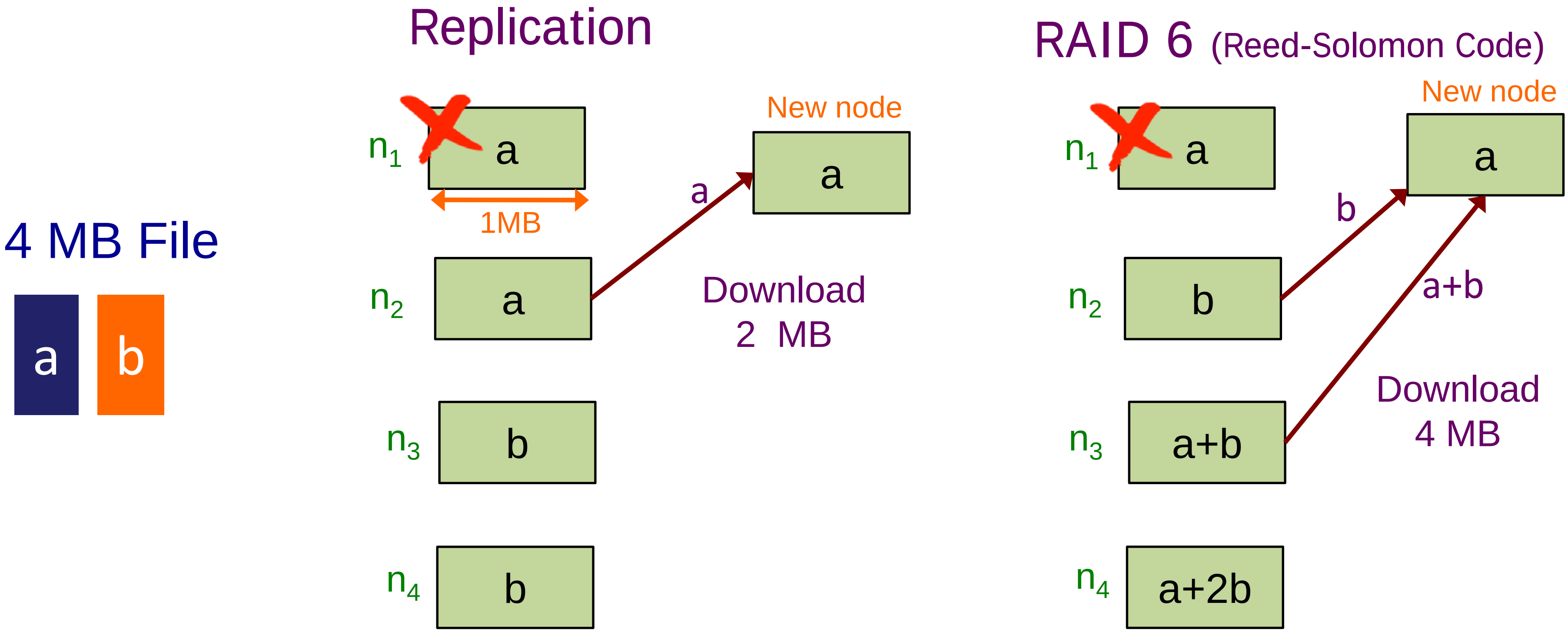
Most of the time we start with a single failure.

Read from any 10 nodes, send all data to 3’ who can repair the lost block.

Bad news:

- High network traffic
- High disk read (10x more than the lost information)

Replication vs. Erasure Codes



tolerates only 1 failure

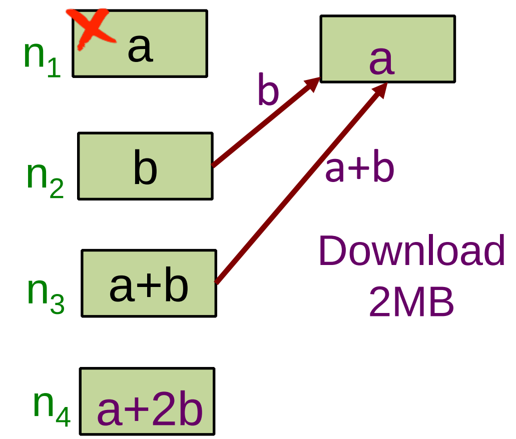
tolerates 2 failures

Reliability		
Bandwidth		

Best of both worlds possible ?

Can we have

- **Storage eff./Reliability** of codes
- **Bandwidth eff.** of replication



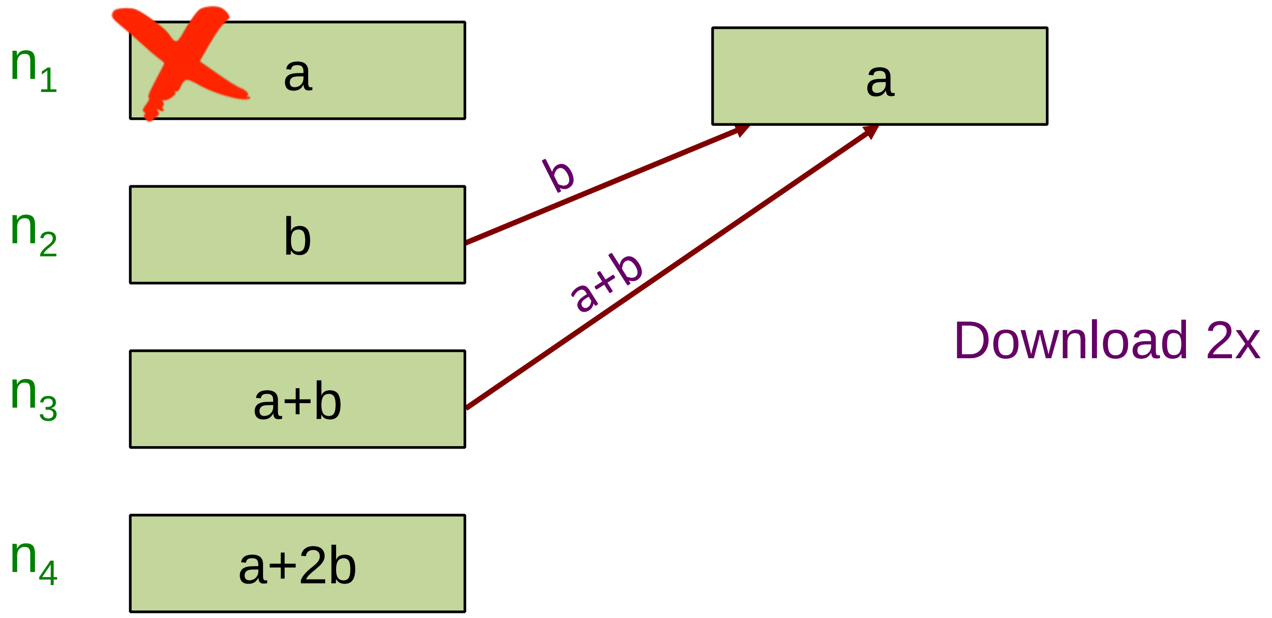
Almost…

… there exists (an optimal) tradeoff

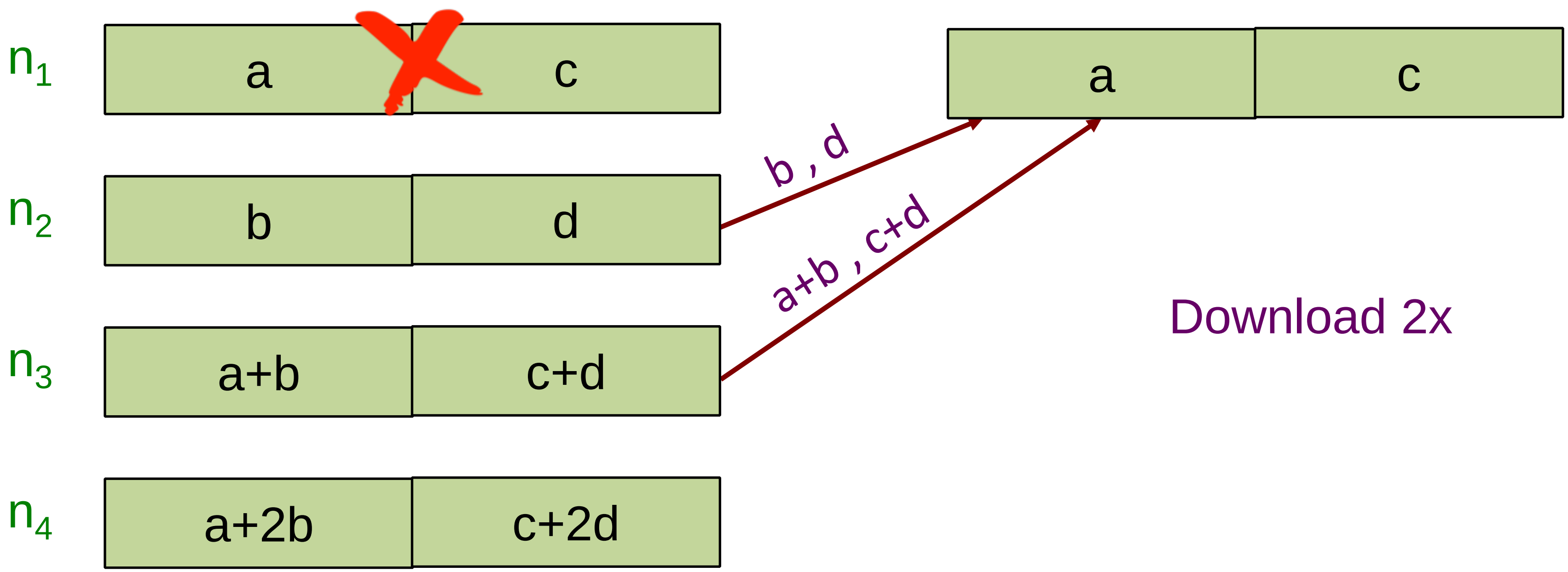
Regenerating Codes

Dimakis, Godfrey, Wainwright & R, 2010

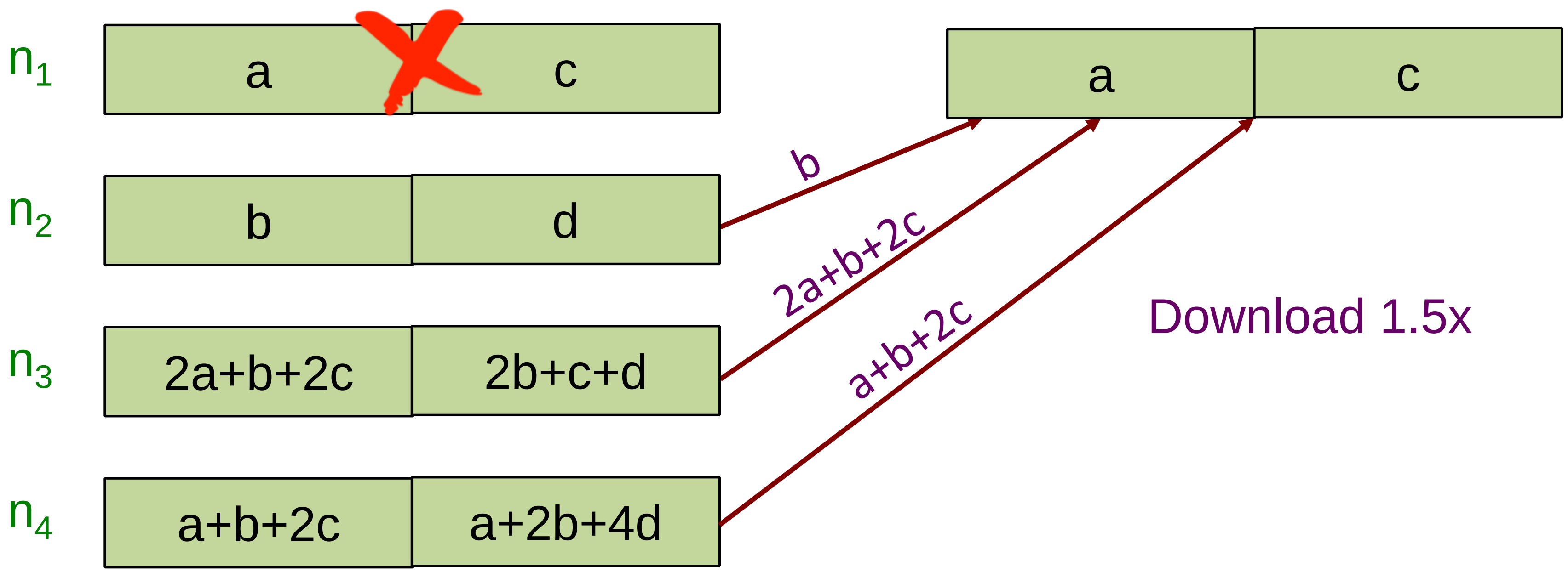
RAID-6 (Reed-Solomon)



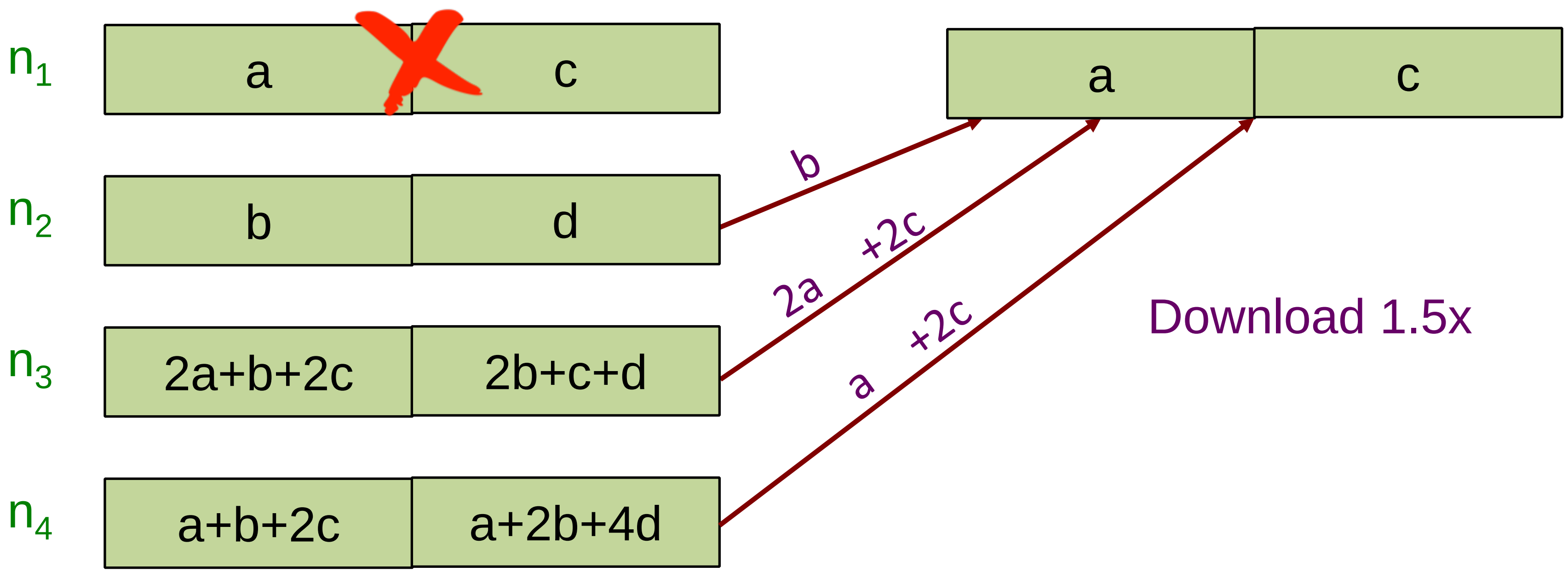
RAID-6 (Reed-Solomon)



Regenerating Codes

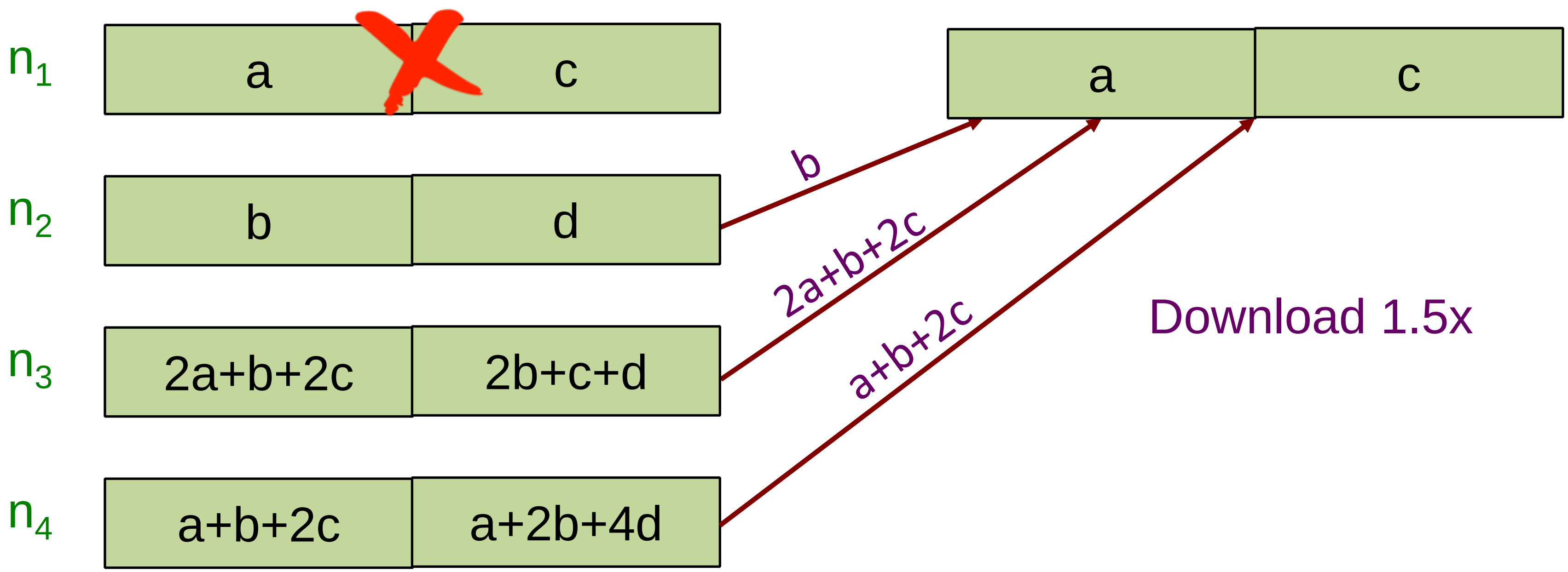


Regenerating Codes



- 25% savings in network bandwidth; much higher in general
- Same reliability as RAID/RS

Key Idea

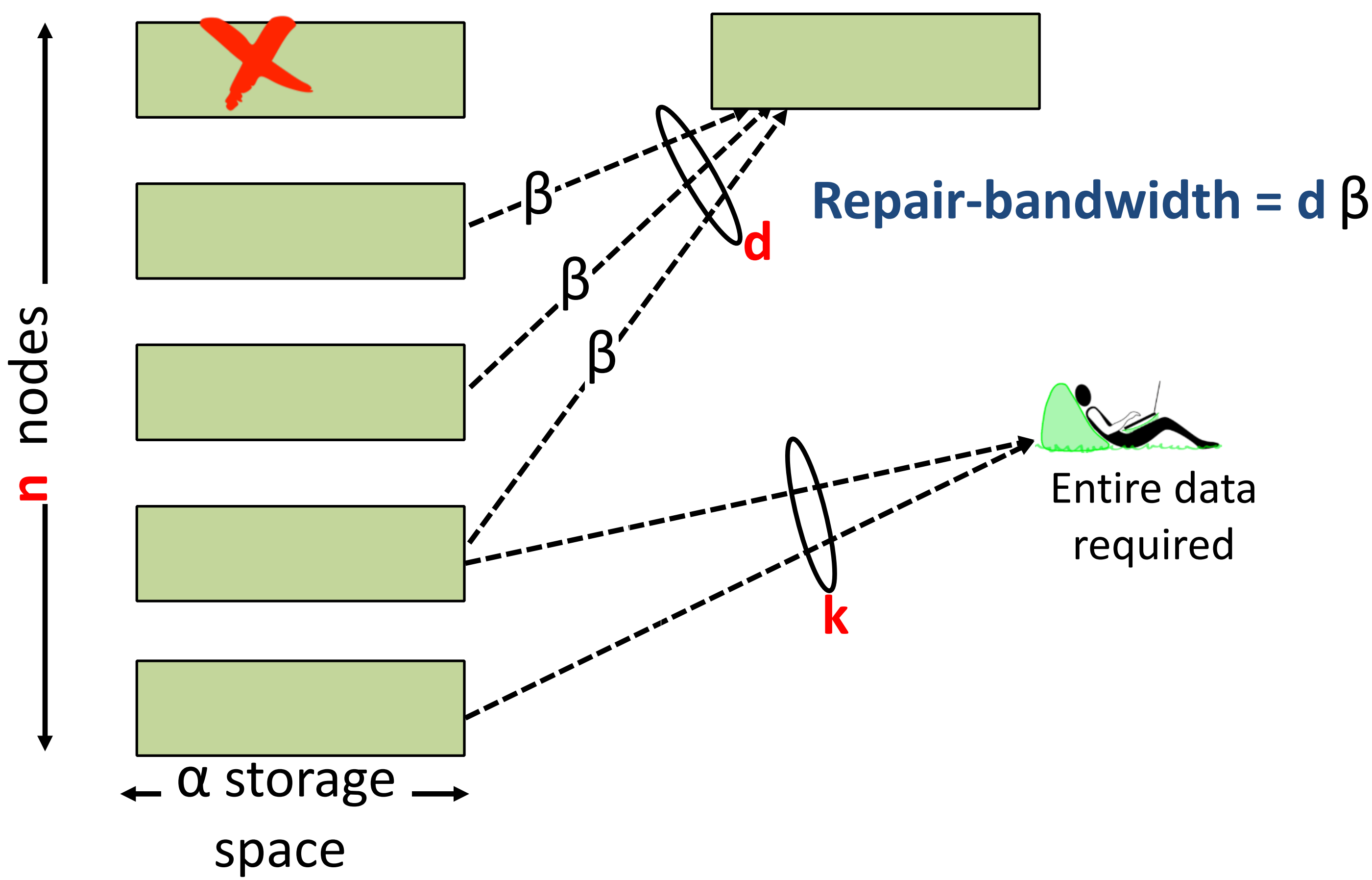


- Code in blocks, and code cleverly *across* multiple blocks
- Connect to more nodes & download less from each

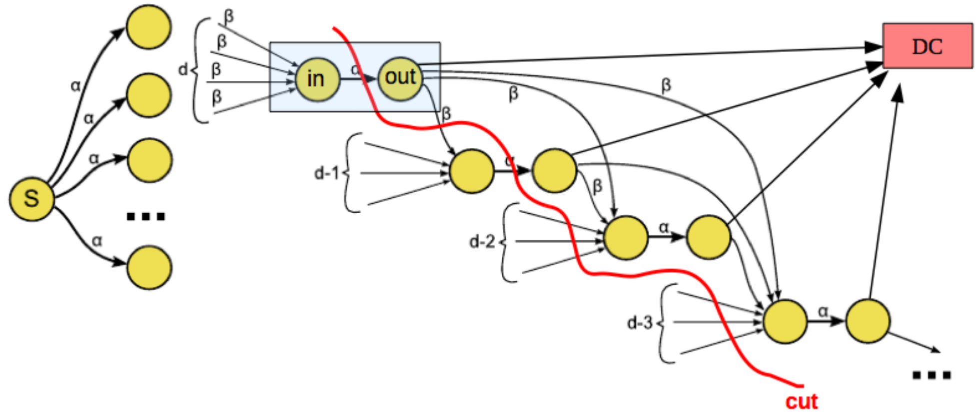
Regenerating Codes

[n, k, d] , { B, α, β}

Source data size is **B**



Storage-Bandwidth tradeoff

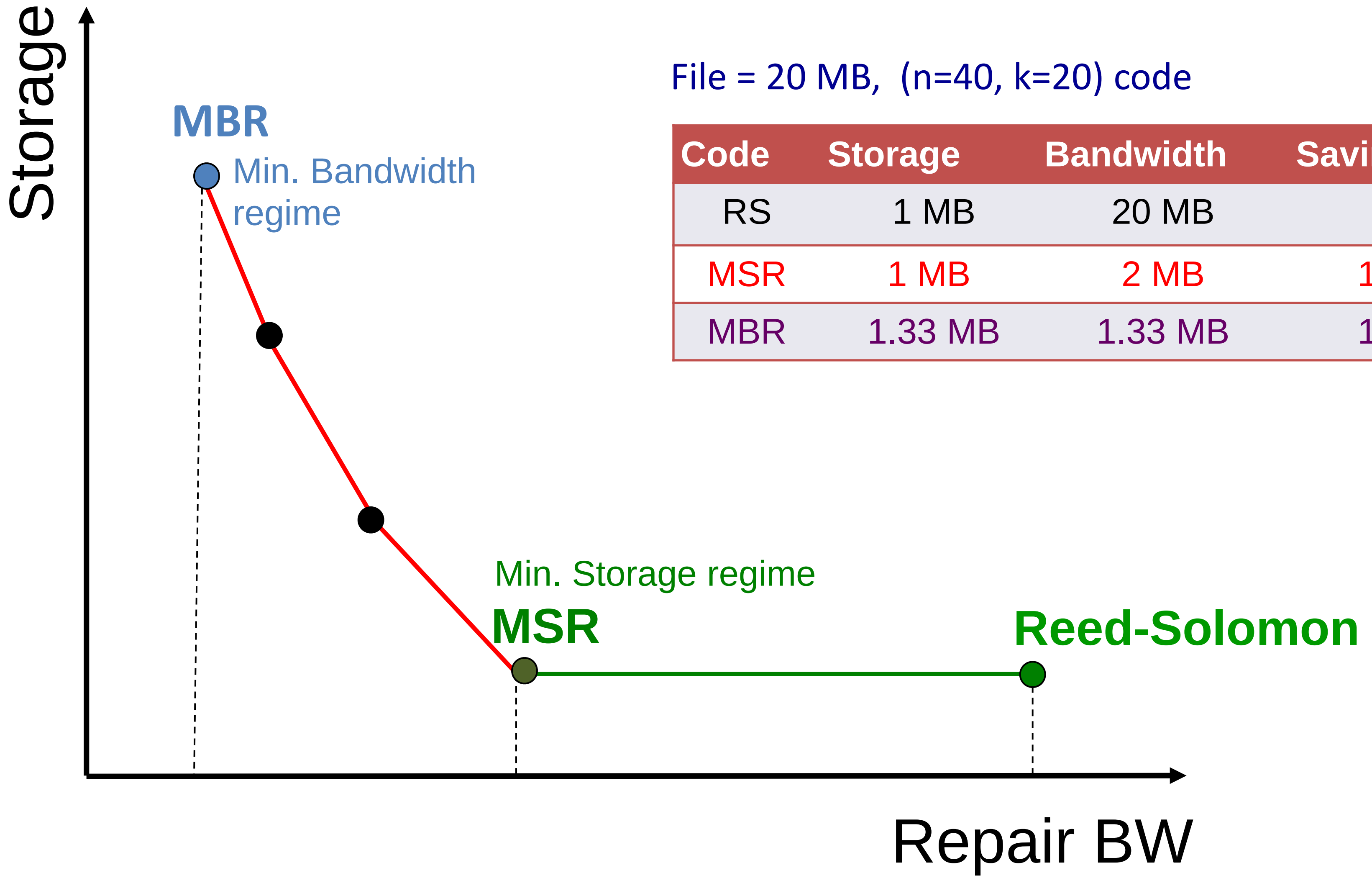


Cut-set bound of network coding:

$$B \leq \sum_{i=0}^{k-1} \min\{\alpha,\ (d-i)\beta\}$$

Tradeoff between storage α and bandwidth β

Storage-Bandwidth tradeoff



File = 20 MB, (n=40, k=20) code

Code	Storage	Bandwidth	Savings
RS	1 MB	20 MB	
MSR	1 MB	2 MB	10x
MBR	1.33 MB	1.33 MB	15x

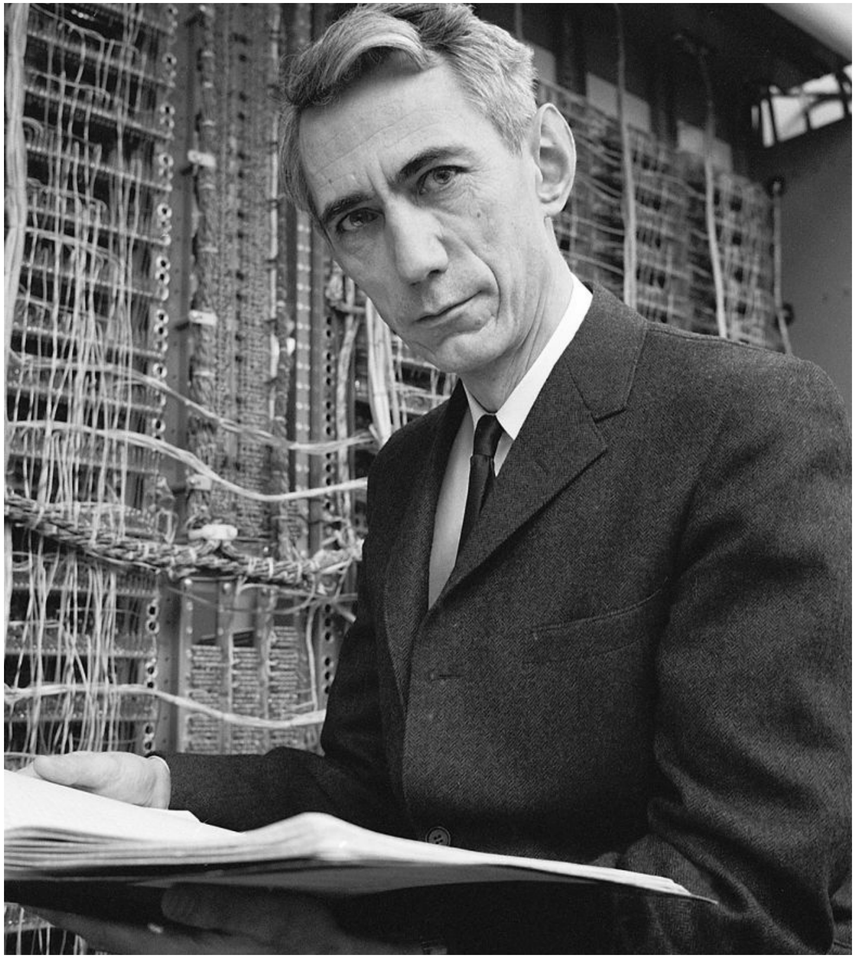
From Shannon to Hadoop: Hitchhiker codes

- Erasure coded storage system built on top of Hadoop Distributed File System (HDFS)
- Rides on top of the RS-based HDFS
  - Reduces network transfer by 25-45% with same storage space and fault tolerance
  - For (14,10) saves *35% disk reads and network transfers*
- Hitchhiker will be a part of future releases of Apache Hadoop 3.0

Conclusion : Shannon’s incredible legacy

- A mathematical theory of communication
- Channel capacity
- Source coding
- Channel coding
- Cryptography
- Sampling theory
- …

His legacy will last many more centuries!



(1916-2001)