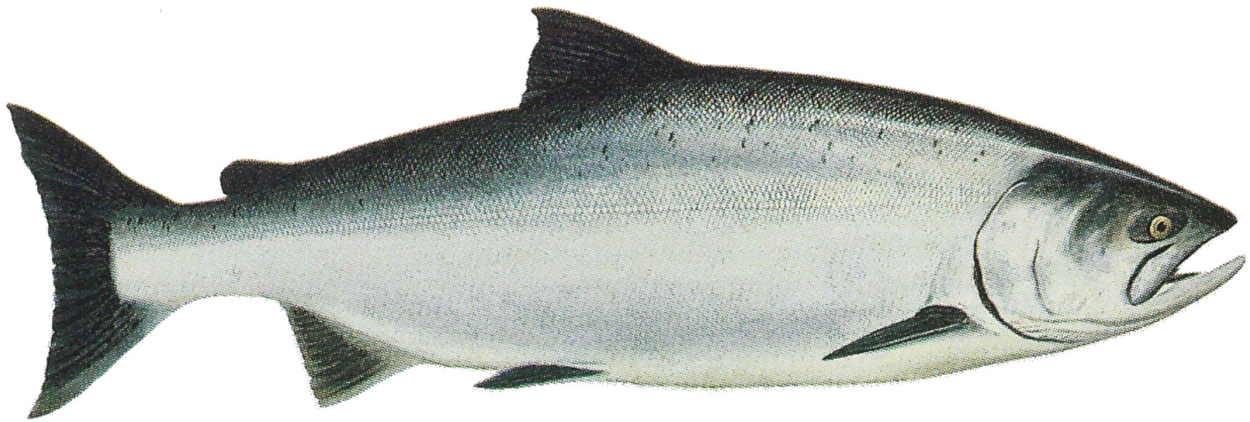
Information Theory
FRIDAY

Mathias Winther Madsen
mathias@gmail.com
github.com/mathias-madsen/nasslli2025/



NASSLLI, June 2025

Marginal, Conditional, and Joint Entropy

Definition

$$H(X) = E\left(\log_2 \frac{1}{p(X)}\right)$$

$$H(X \mid Y) = E\left(\log_2 \frac{1}{p(X \mid Y)}\right)$$

$$H(X, Y) = E\left(\log_2 \frac{1}{p(X, Y)}\right)$$

Theorem

$$H(X, Y) \;=\; H(X \mid Y) + H(Y)$$

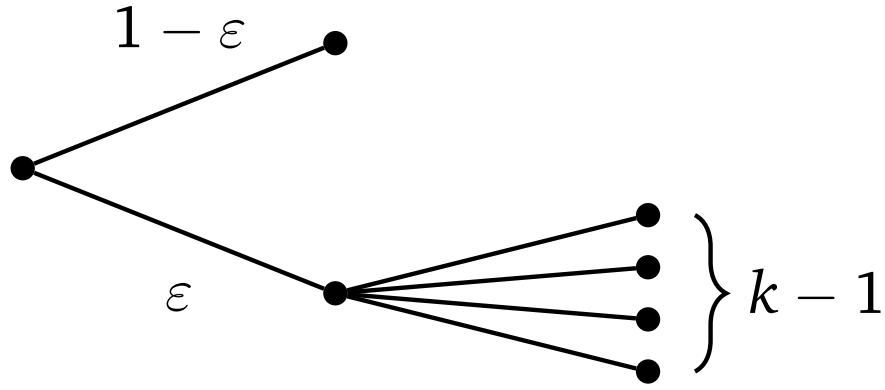
Theorem

$$H(X \mid Y) \;\leq\; H(X)$$

Fano's Inequality

Theorem: Fano's Inequality

Let X be a random variable that can take k different values, one of which has probability $1-\varepsilon$. Then $H(X) \leq 1 + \varepsilon \log_2 k$.



In fact

$$\varepsilon H_2(1/k) \;\leq\; H(X) \;\leq\; 1 + \varepsilon \log_2 k,$$

or equivalently,

$$\frac{H(X)-1}{\log k} \;\leq\; \varepsilon \;\leq\; \frac{H(X)}{H_2(1/k)}.$$

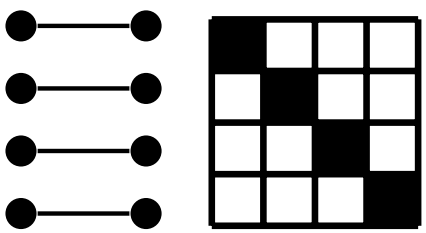
Mutual Information

Definition: Mutual Information

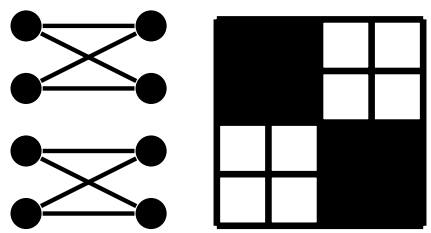
$$I(X;Y) \;=\; H(X) + H(Y) - H(X,Y)$$

Theorem: Decompositions

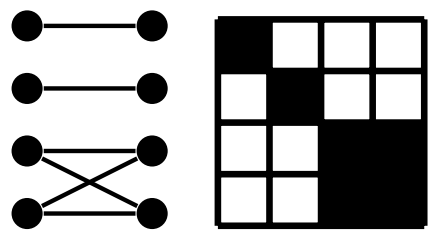
$$\begin{aligned} H(X,Y) &= H(X) + H(Y\,|\,X) &&= H(Y) + H(X\,|\,Y); \\ I(X;Y) &= H(X) - H(X\,|\,Y) &&= H(Y) - H(Y\,|\,X). \end{aligned}$$



$I = 2$



$I = 1$



$I = 1\frac{1}{2}$

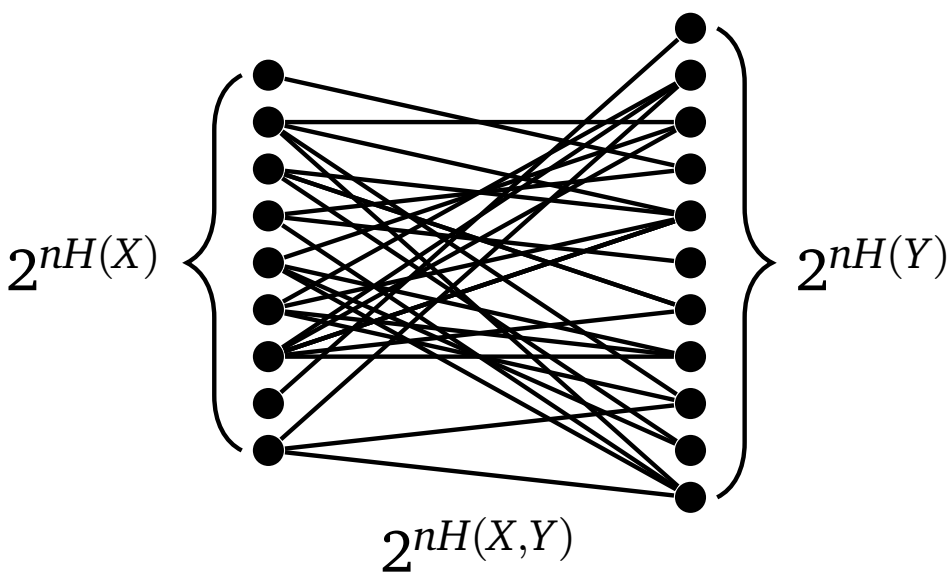
Joint Equipartition

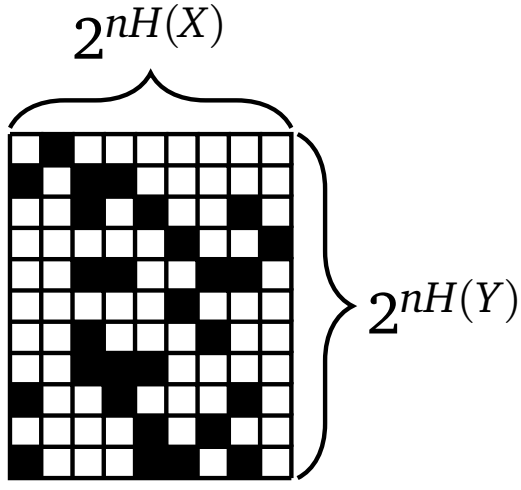
The Asymptotic Equipartition Property

When the input letters $X_1, X_2, \ldots, X_n$ are independent, there are

- $2^{nH(X)}$ typical input strings X^n;
- $2^{nH(Y)}$ typical output strings Y^n;
- $2^{nH(X,Y)}$ typical input-output pairs (X^n, Y^n);

and $P(\text{typical}) \to 0$ as $n \to \infty$.



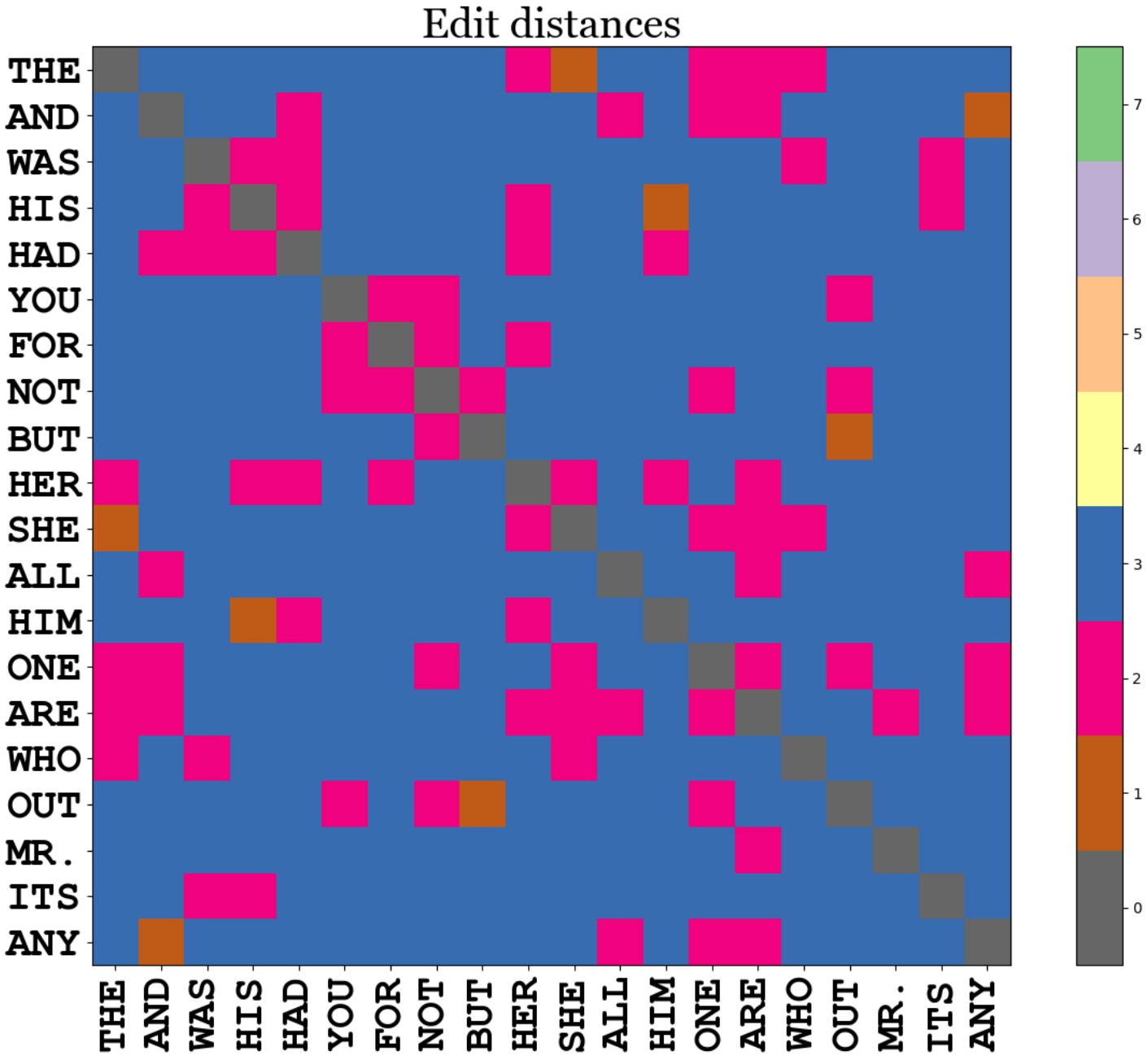


The Redundancy of English

AT SEVENTEEO HE HAD TRIED FPR A GOVERNMENT APPOINTMENT, BUT HE HAD FAILED TO GET IT, BEING POOR AND FRIENRLESS, ANQ FOR THREE YEARS HE HAD WORKED IN THE STINKING LABYRINTH OF TOE MKNDALAY BAZAARS, CEERKING FOR THE RICE MERCHANTS APD SOMETIMES CTEALING. THEN WHEN HE WAS TWENTY A LUCKY STROKE OF BLACKMAIL PUT HIM FN POSSESSION OF FOUR HUNDOED RUPEES, AND HE WENT AT ONCE PO RANGOON ANI BOUGHT HIS WAY INTO A GOVERNMENT CLERKSHIP. THE JGB WAS A LUCRATIVE ONE THOUGH SHE SALARY WAS SMALL.

The Redundancy of English

HHH	HHD	HHO	HHI	HHM	HHN	HDH	HDD	HDO	HDI	HDM	HDN
HOH	HOD	HOO	HOI	HOM	HON	HIH	HID	HIO	HII	HIM	HIN
HMH	HMD	HMO	HMI	HMM	HMN	HNH	HND	HNO	HNI	HNM	HNN
DHH	DHD	DHO	DHI	DHM	DHN	DDH	DDD	DDO	DDI	DDM	DDN
DOH	DOD	DOO	DOI	DOM	DON	DIH	DID	DIO	DII	DIM	DIN
DMH	DMD	DMO	DMI	DMM	DMN	DNH	DND	DNO	DNI	DNM	DNN
OHH	OHD	OHO	OHI	OHM	OHN	ODH	ODD	ODO	ODI	ODM	ODN
OOH	OOD	OOO	OOI	OOM	OON	OIH	OID	OIO	OII	OIM	OIN
OMH	OMD	OMO	OMI	OMM	OMN	ONH	OND	ONO	ONI	ONM	ONN
IHH	IHD	IHO	IHI	IHM	IHN	IDH	IDD	IDO	IDI	IDM	IDN
IOH	IOD	IOO	IOI	IOM	ION	IIH	IID	IIO	III	IIM	IIN
IMH	IMD	IMO	IMI	IMM	IMN	INH	IND	INO	INI	INM	INN
MHH	MHD	MHO	MHI	MHM	MHN	MDH	MDD	MDO	MDI	MDM	MDN
MOH	MOD	MOO	MOI	MOM	MON	MIH	MID	MIO	MII	MIM	MIN
MMH	MMD	MMO	MMI	MMM	MMN	MNH	MND	MNO	MNI	MNM	MNN

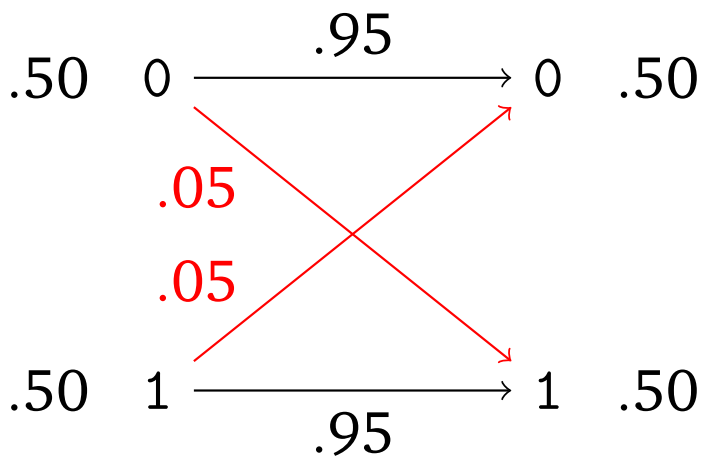

Edit distances
THE
AND
WAS
HIS
HAD
YOU
FOR
NOT
BUT
HER
SHE
ALL
HIM
ONE
ARE
WHO
OUT
MR.
ITS
ANY
7
6
5
4
3
2
1
0

Noisy Channels

$$X^{20} = 00010010111110110111$$

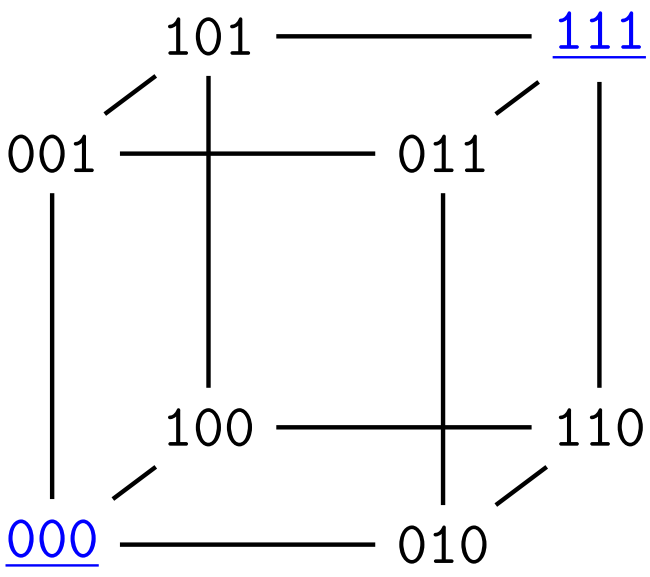
$$\downarrow$$

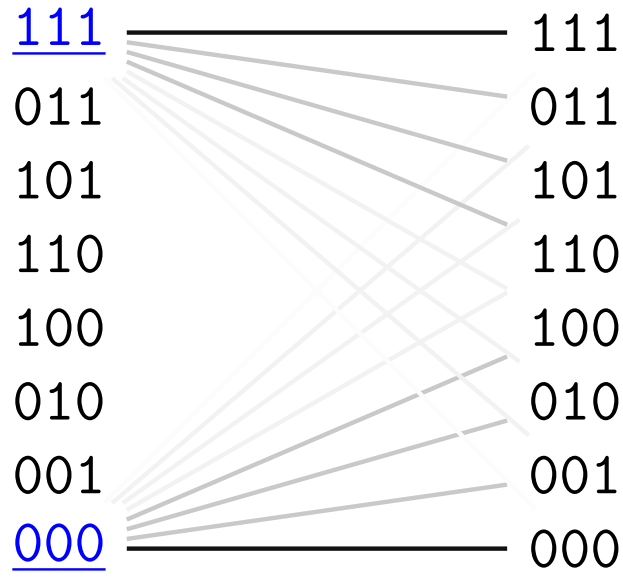
$$Y^{20} = 00010010110110111011$$



$P(X,Y)$	$X=0$	$X=1$	$P(Y)$
$Y=0$	.45	.05	.50
$Y=1$	.05	.45	.50
$P(X)$	.50	.50	1

Repetition Codes





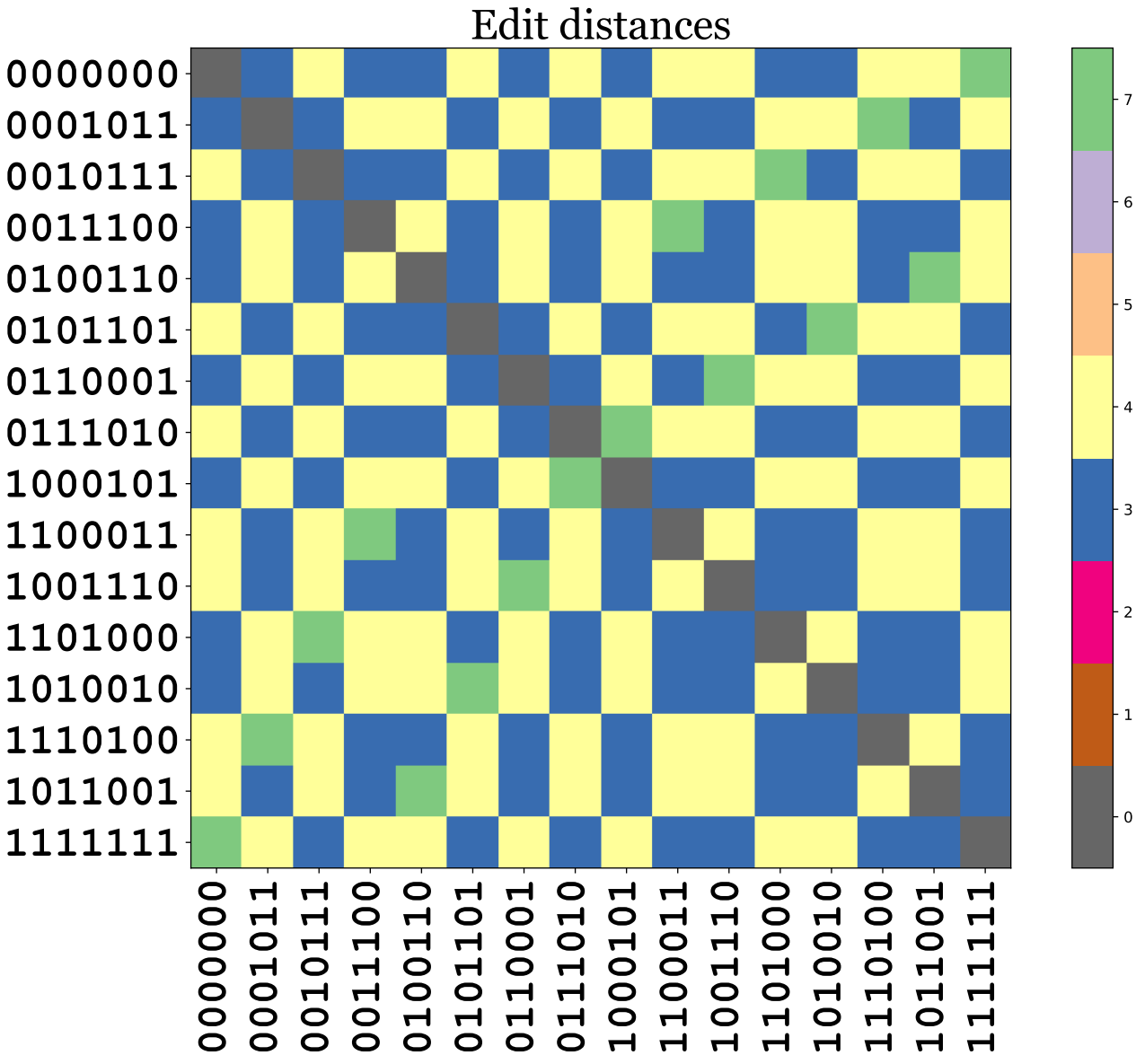
	000	001	010	100	011	101	110	111
000	.485	.005	.005	.005	.000	.000	.000	.000
111	.000	.000	.000	.000	.005	.005	.005	.485

Hamming Codes

0000000	0000001	0000010	0000011	0000100	0000101	0000110	0000111
0001000	0001001	0001010	0001011	0001100	0001101	0001110	0001111
0010000	0010001	0010010	0010011	0010100	0010101	0010110	0010111
0011000	0011001	0011010	0011011	0011100	0011101	0011110	0011111
0100000	0100001	0100010	0100011	0100100	0100101	0100110	0100111
0101000	0101001	0101010	0101011	0101100	0101101	0101110	0101111
0110000	0110001	0110010	0110011	0110100	0110101	0110110	0110111
0111000	0111001	0111010	0111011	0111100	0111101	0111110	0111111
1000000	1000001	1000010	1000011	1000100	1000101	1000110	1000111
1001000	1001001	1001010	1001011	1001100	1001101	1001110	1001111
1010000	1010001	1010010	1010011	1010100	1010101	1010110	1010111
1011000	1011001	1011010	1011011	1011100	1011101	1011110	1011111
1100000	1100001	1100010	1100011	1100100	1100101	1100110	1100111
1101000	1101001	1101010	1101011	1101100	1101101	1101110	1101111
1110000	1110001	1110010	1110011	1110100	1110101	1110110	1110111
1111000	1111001	1111010	1111011	1111100	1111101	1111110	1111111

Richard W. Hamming: “Error Detecting and Error Correcting Codes,” *Bell System Technical Journal*, 1950.

Hamming Codes



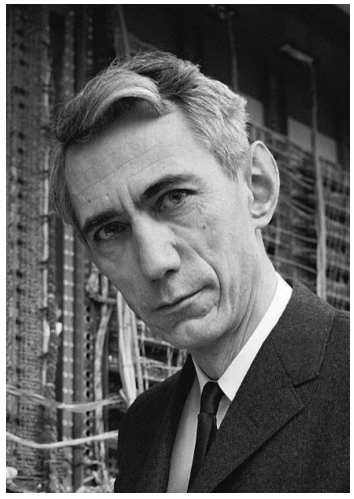
The Channel Coding Theorem

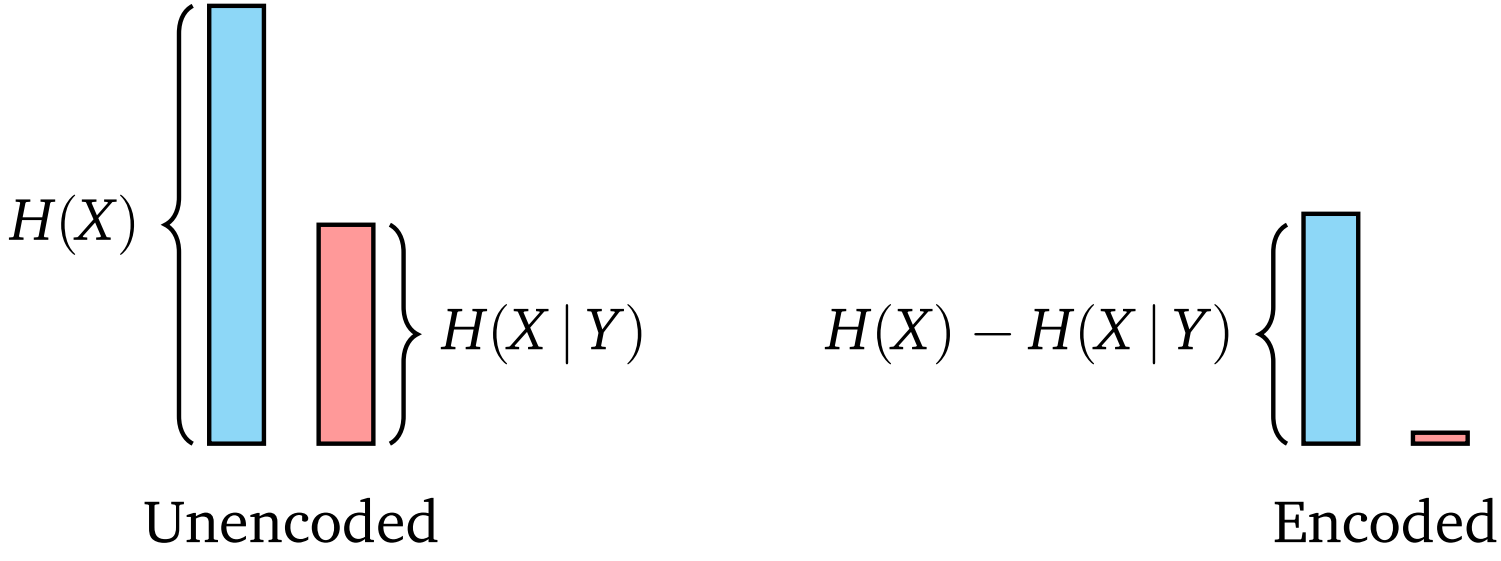
The Channel Coding Theorem

We can communicate at any rate

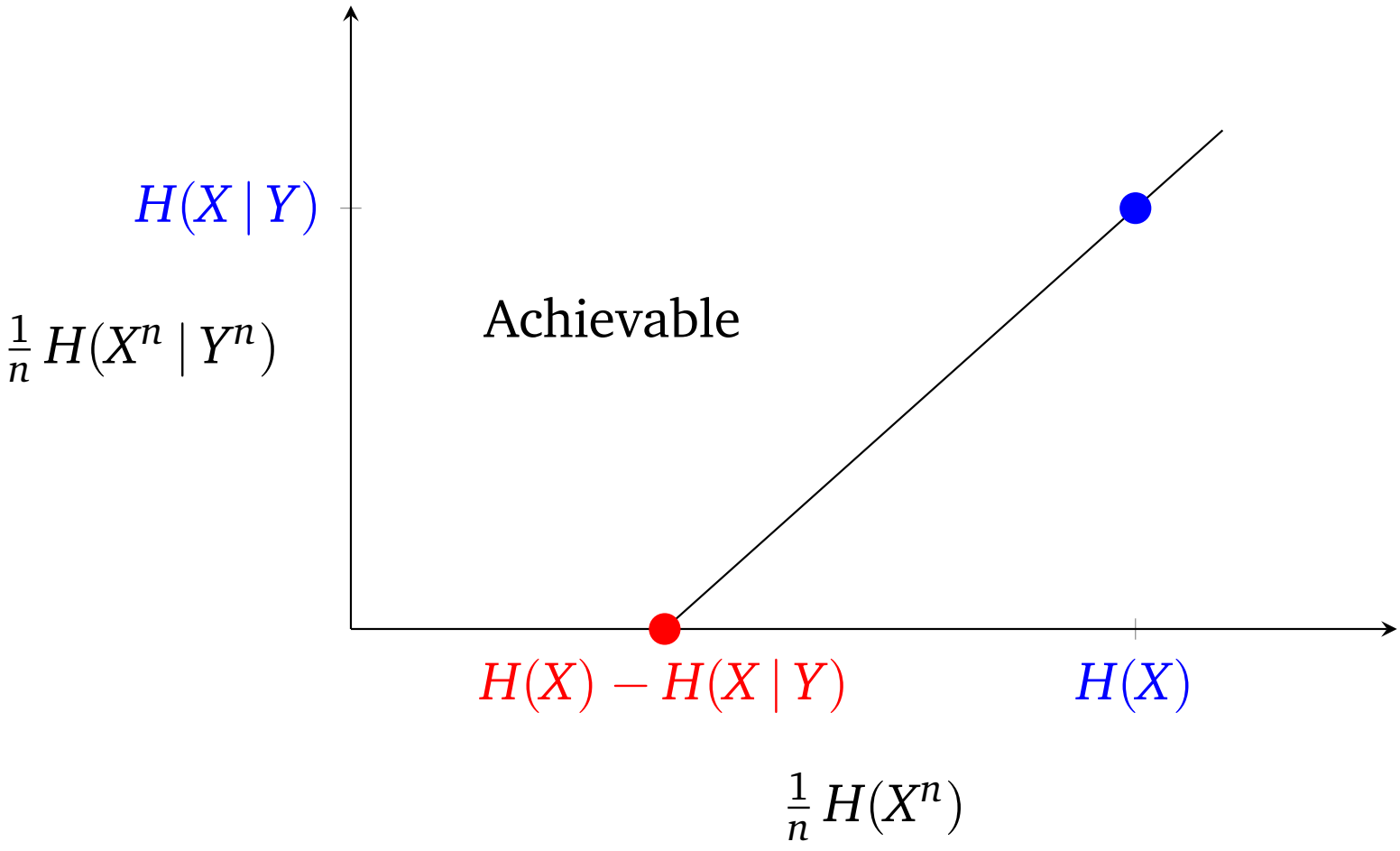
$$R < H(X) - H(X \mid Y)$$

with neglible probability of error.

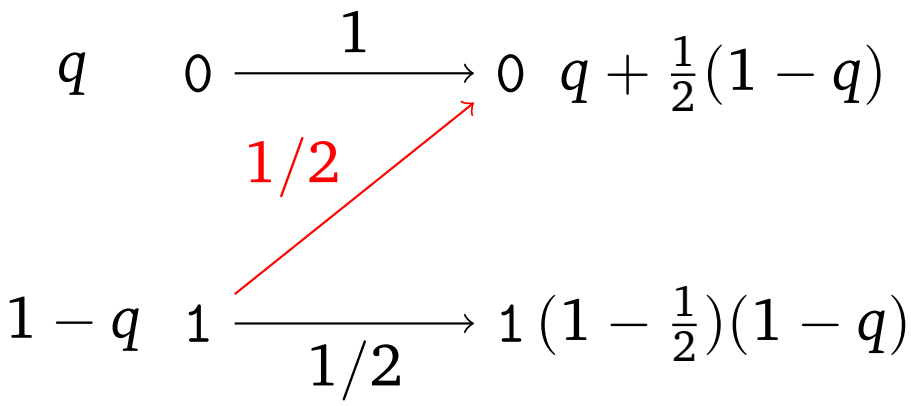


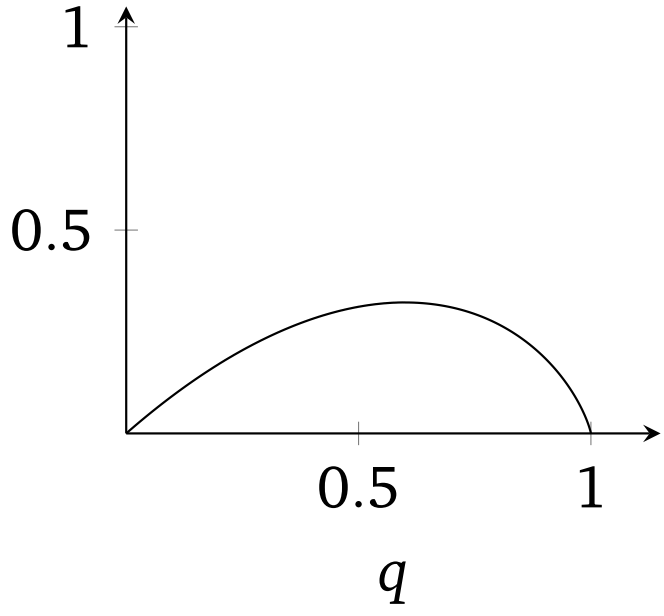


The Channel Coding Theorem

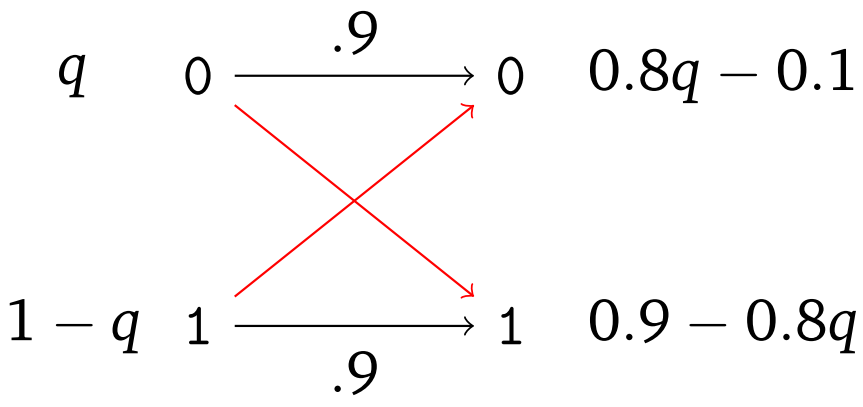


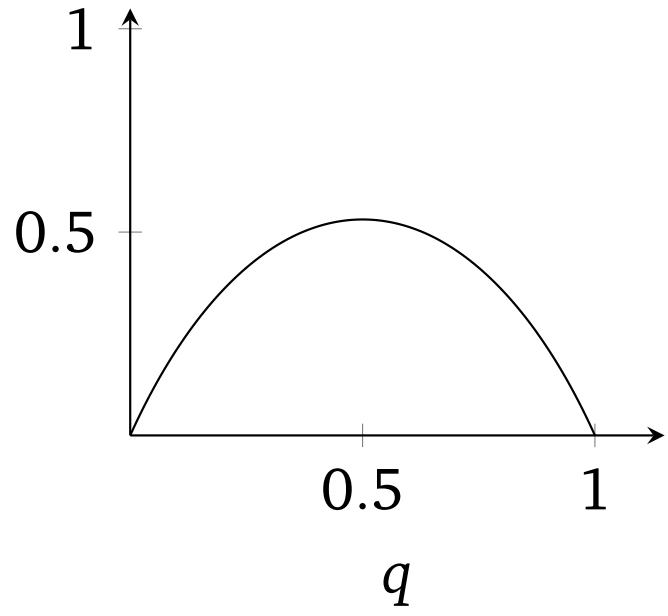
The Channel Coding Theorem





The Channel Coding Theorem



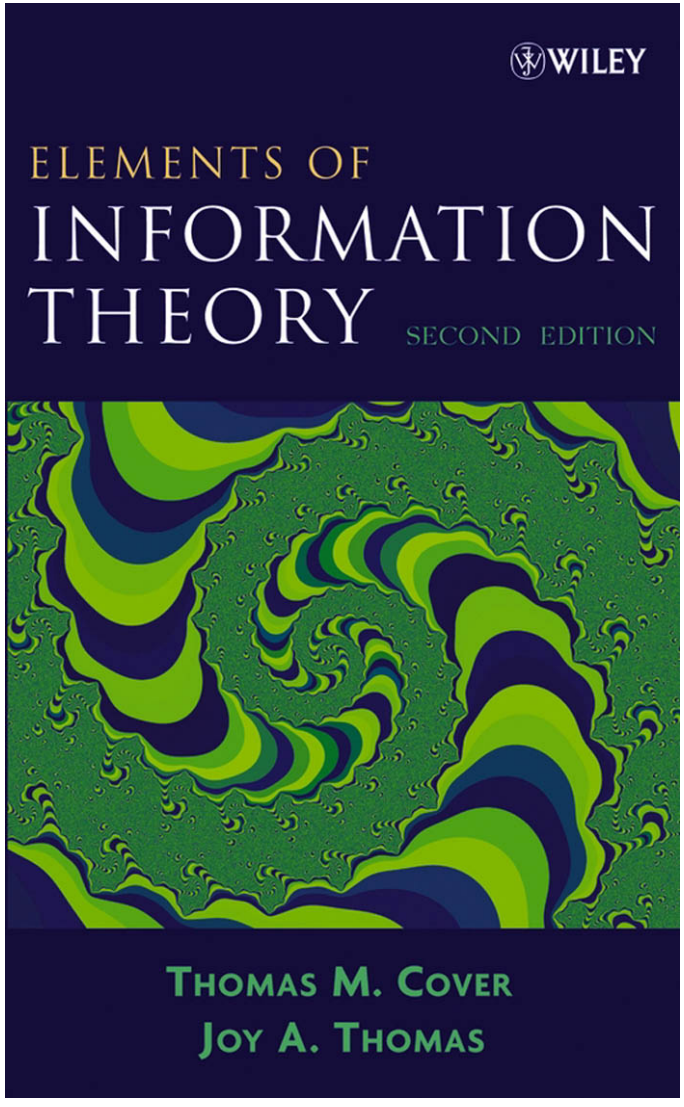


Random Codes

000000	000001	000010	000011	000100	000101	000110	000111
001000	001001	001010	001011	001100	001101	001110	001111
010000	010001	010010	010011	010100	010101	010110	010111
011000	011001	011010	011011	011100	011101	011110	011111
100000	100001	100010	100011	100100	100101	100110	100111
101000	101001	101010	101011	101100	101101	101110	101111
110000	110001	110010	110011	110100	110101	110110	110111
111000	111001	111010	111011	111100	111101	111110	111111

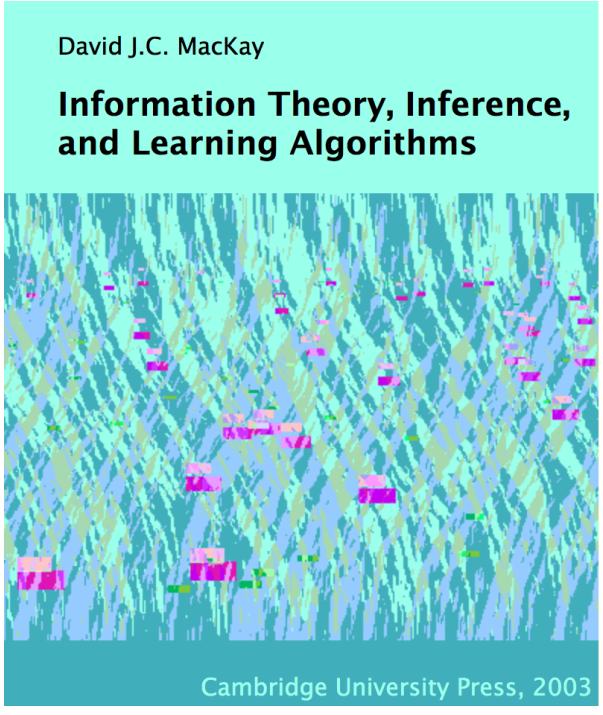
Randomly downsample the input space to 2^{nR} codewords

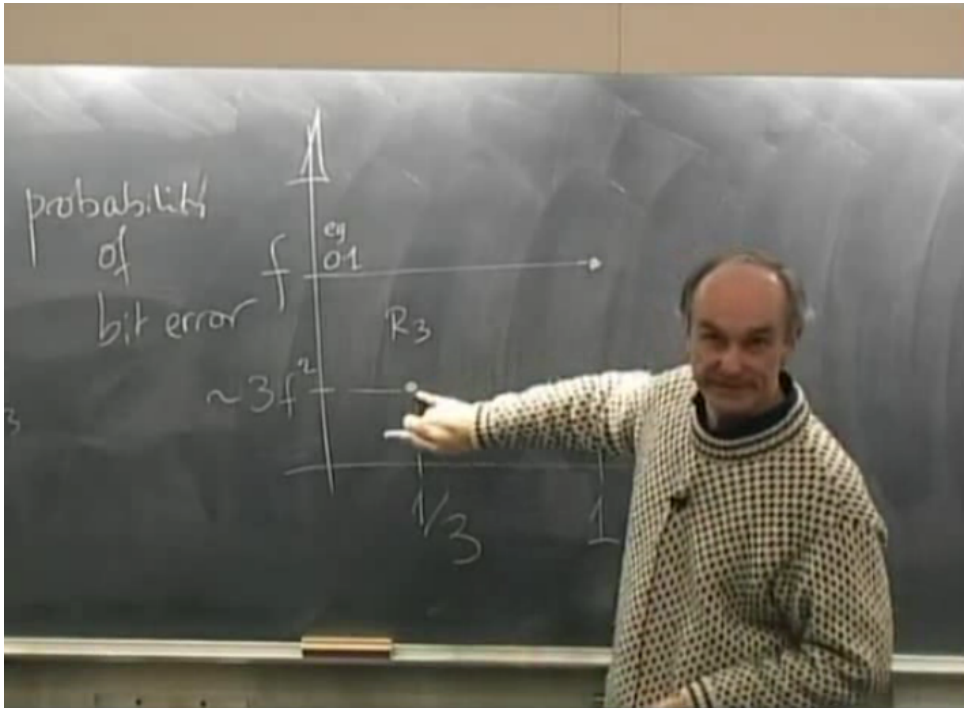
Where To Go Next



Thomas Cover and Joy Thomas:
Elements of Information Theory
(1991, 2006)

Where To Go Next

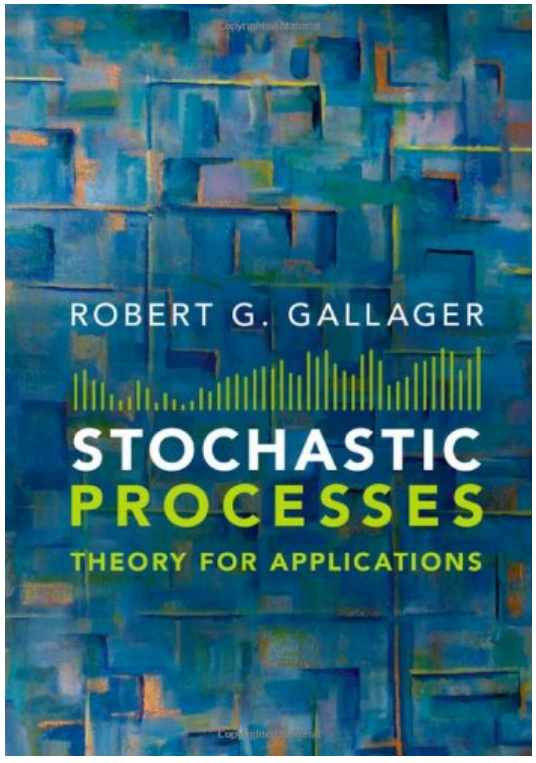


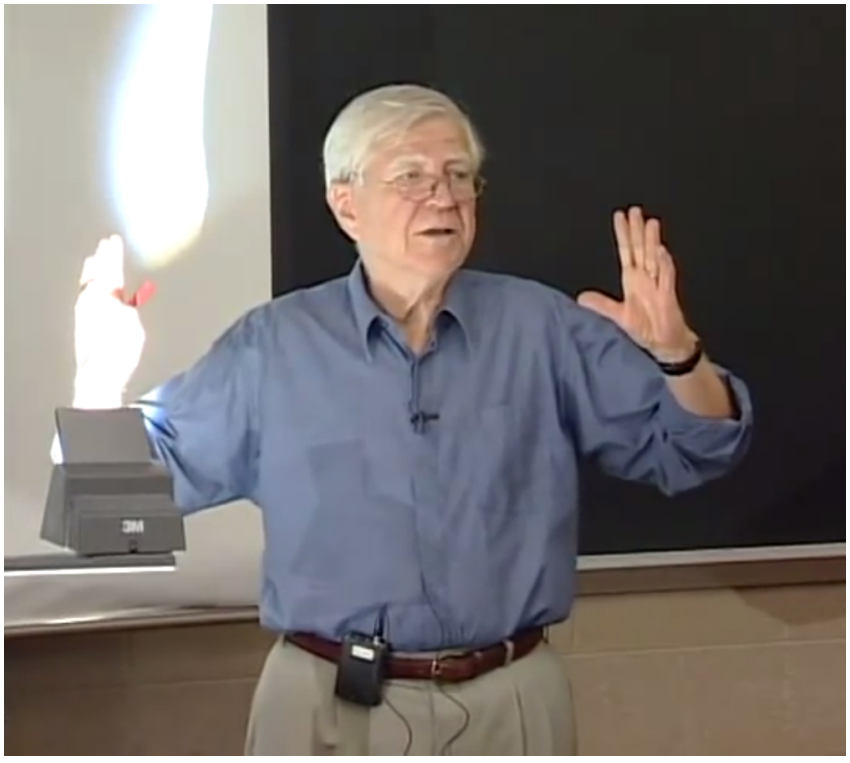


David MacKay: *Information Theory, Inference, and Learning Algorithms* (2003)

`www.inference.org.uk/mackay/itila/`

Where To Go Next





Robert Gallager: *Discrete Stochastic Processes: Theory for Applications* (2013)

`https://ocw.mit.edu/courses/6-262-discrete-stochastic-processes-spring-2011/`

Where To Go Next

These results are the main justification for the definition of C and will now be proved.

Theorem 11. Let a discrete channel have the capacity C and a discrete source the entropy per second H. If $H \leq C$ there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If $H > C$ it is possible to encode the source so that the equivocation is less than $H - C + \epsilon$ where ϵ is arbitrarily small. There is no method of encoding which gives an equivocation less than $H - C$.

The method of proving the first part of this theorem is not by exhibiting a coding method having the desired properties, but by showing that such a code must exist in a certain group of codes. In fact we will average the frequency of errors over this group and show that this average can be made less than ϵ. If the average of a set of numbers is less than ϵ there must exist at least one in the set which is less than ϵ. This will establish the desired result.

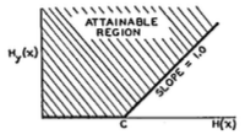


Fig. 9—The equivocation possible for a given input entropy to a channel.

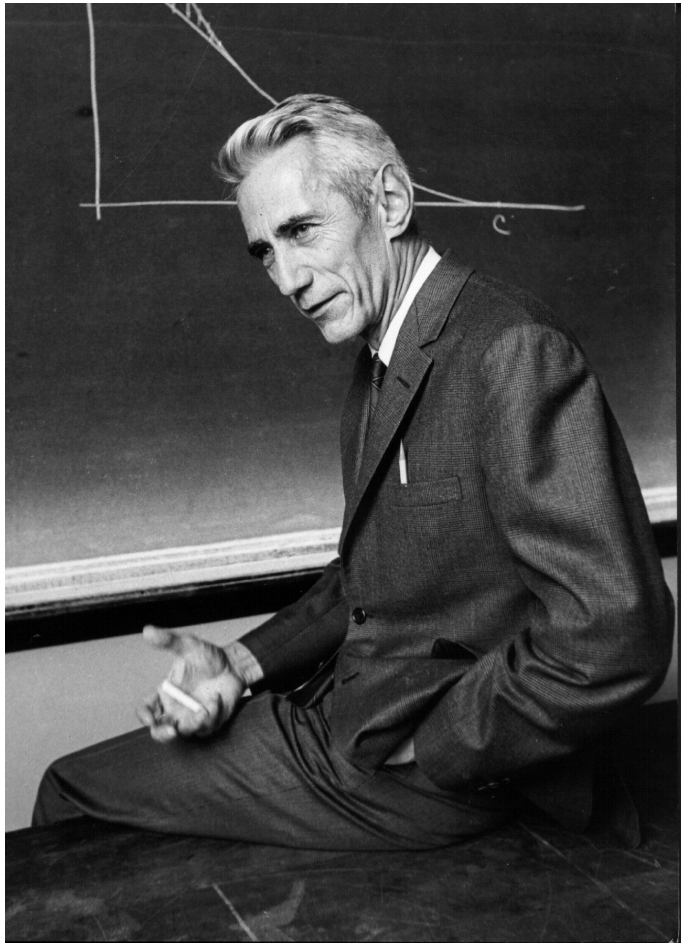
The capacity C of a noisy channel has been defined as

$$C = \text{Max}\,(H(x) - H_y(x))$$

where x is the input and y the output. The maximization is over all sources which might be used as input to the channel.

Let S_0 be a source which achieves the maximum capacity C. If this maximum is not actually achieved by any source let S_0 be a source which approximates to giving the maximum rate. Suppose S_0 is used as input to the channel. We consider the possible transmitted and received sequences of a long duration T. The following will be true:

1. The transmitted sequences fall into two classes, a high probability group with about $2^{TH(x)}$ members and the remaining sequences of small total probability.
2. Similarly the received sequences have a high probability set of about $2^{TH(y)}$ members and a low probability set of remaining sequences.
3. Each high probability output could be produced by about $2^{TH_y(x)}$ inputs. The probability of all other cases has a small total probability.



Shannon: “A Mathematical Theory of Communication” (1948)

Where To Go Next

mathias.winther@gmail.com
github.com/mathias-madsen