# CS258: Information Theory

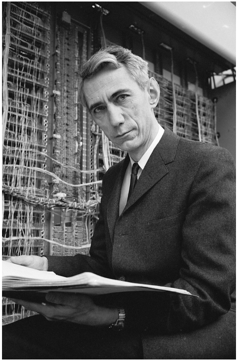
Fan Cheng

Spring, 2018. chengfan@sjtu.edu.cn

- Instructor: Fan Cheng, Rm 3-513, SEIEE
  (`http://www.cs.sjtu.edu.cn/~chengfan/`)
  Office hour: By appointment
  TA: TBA
- Textbook: David J.C. MacKay, "Information Theory, Inference, and Learning Algorithms," Cambridge Press, 2005 (http://www.inference.org.uk/itprnn/book.html)
- 16 Weeks := 14 lectures + 1 in-class midterm + 1 Q&A
- Grade policy := 50% final + 30% midterm + 10% attendance + 10% homework

# Birth of Information Theory

"A Mathematical Theory of Communication," Bell System Technical Journal, 27 (3): 379-423, July, 1948.



Claude. E. Shannon
(1916-2001)
https://en.wikipedia.org/wiki/Claude_Shannon

*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.*

C. E. Shannon, 1948

**IEEE Information Theory Society**

http://www.itsoc.org

**IEEE Transactions on Information Theory**

http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=18

# Shannon: father of information theory

## Mathematician

Ph.D. in Mathematics from MIT. Worked at AT&T Bell Labs and RLE in MIT

## Electrical engineer

Mater's Thesis: electrical applications of Boolean algebra could construct any logical, numerical relationship

## Cryptographer

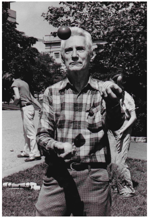
"A Mathematical Theory of Cryptography," 1949.

## Friend of Turing

For two months early in 1943, Shannon came into contact with the leading British mathematician Alan Turing. Shannon and Turing met at teatime in the cafeteria. Turing showed Shannon his 1936 paper that defined what is now known as the "Universal Turing machine"

# Shannon: life

Magnetic mouse



Juggling

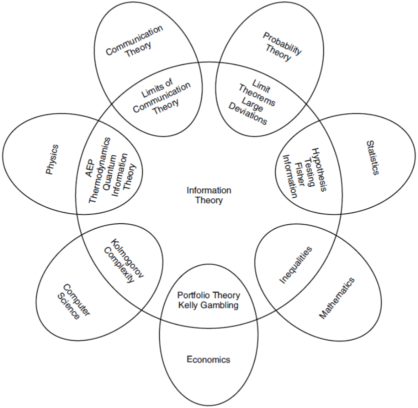
Unicycling

# Topics in IT

Big Data Analytics
Coding for Communication and Storage
Coding Theory
Combinatorics and Information Theory
Communication Theory
Complexity and Computation Theory
Compressed Sensing and Sparsity
Cryptography and Security
Detection and Estimation
Distributed Storage
Emerging Applications of Information Theory
Information Theory and Statistics
Information Theory in Biology
Information Theory in Computer Science
Statistical/Machine Learning
Network Coding and Applications
Network Data Analysis
Network Information Theory
Optical Communication
Quantum Information and Coding Theory
Shannon Theory
Signal Processing
Source Coding and Data Compression
Wireless Communication and Networks

https://www.isit2018.org/authors/call-for-papers/

# Information theory to other fields




[1]

---

[1] Elements of Information Theory, Thomas M. Cover and Joy A. Thomas

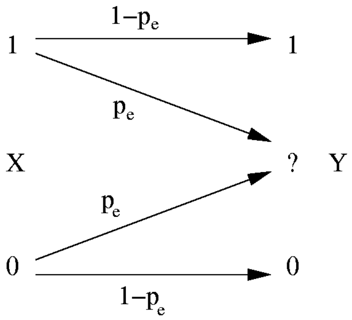
# Classical textbooks

- Information Theory and Reliable Communication, 1st Edition, Robert G. Gallager
- Elements of Information Theory, 2nd Edition (Wiley Series in Telecommunications and Signal Processing), Thomas M. Cover, Joy A. Thomas
- Information Theory: Coding Theorems for Discrete Memoryless Systems, 2nd Edition, Imre Csiszar, Janos Korner
- Information Theory, Inference and Learning Algorithms, David J. C. MacKay
- A First Course in Information Theory (Information Technology: Transmission, Processing and Storage), 1st Edition, Raymond W. Yeung

# Course plan

- Elements: entropy, mutual information, information divergence, etc.
- Data compression
- Noisy-channel coding
- Probability and inference
- Neural networks
- Low-density parity-check codes

Prerequisite: Probability theory, mathematical analysis, matrix theory
In class: pen and paper




Binary Erasure Channel

# Probability

For random variable $X$ define on alphabet $\mathscr{X}$, its mean and variance is defined as

$$\mathscr{E}(X) := \sum_{x \in \mathscr{X}} x p(x)$$

$$Var(x) := \mathscr{E}(X^2) - (\mathscr{E}(X))^2$$

- $\mathscr{E}(X_1 + X_2) = \mathscr{E}(X_1) + \mathscr{E}(X_2)$
- If $X_1$ and $X_2$ are independent, then $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$

Some probability distributions: Bernoulli, Binomial, Poisson, Gauss, etc.

A function $f$ on $(a, b)$ is called convex iff for any $x_1$, $x_2$ in $(a, b)$

$$f(\frac{x_1 + x_2}{2}) \leq \frac{f(x_1) + f(x_2)}{2}$$

- If $f(x)$ is twice differentialable, then $f(x)$ is convex iff $f''(x) \geq 0$
- If $f(x)$ is twice differentialable, then $f(x)$ is minimized iff $f'(x) = 0$
- (Jesen's inequality) $f(x)$ is convex in $(a, b)$,

$$f(\mathscr{E}(X)) \leq \mathscr{E} f(X)$$

Take $f := e^x, sin(x), cos(x), x^2, x^3$ for example

# Binomial distribution

A bent coin has probability $f$ of coming up heads. The coin is tossed $N$ times. What is the probability distribution of the number of heads, $r$? What are the mean and variance of $r$?

$$
\begin{aligned}
p(r|f,N) &= \binom{N}{r} f^r (1-f)^{N-r} \\
\mathscr{E}(r) &= Nf \\
\mathrm{Var}(r) &= Nf(1-f)
\end{aligned}
$$

# Approximating $x!$ and $\binom{N}{r}$

## Stirling's approximation

$$x! \simeq x^x e^{-x}\sqrt{2\pi x} \Longleftrightarrow \ln x! = x\ln x - x + \frac{1}{2}\ln 2\pi x$$

- Poisson distribution: $P(r|\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$
- When $\lambda$ is large and $r \to \lambda$, $P(r|\lambda) \to \frac{1}{\sqrt{2\pi\lambda}}e^{-\frac{(r-\lambda)^2}{2\lambda}}$
- Plug $r = \lambda$

$$\ln\binom{N}{r} \simeq (N-r)\ln\frac{N}{N-r} + r\ln\frac{N}{r}$$

## Binary Entropy Function

$$H_2(x) = -x\log x - (1-x)\log(1-x)$$

$$\binom{N}{r} \simeq 2^{NH_2(r/N)}$$

# Exercise

- Plot $H_2(x)$ in python
- $H_2(x)$ is symmetric at $x = \frac{1}{2}$
- $H_2(x)$ is maximized at $x = \frac{1}{2}$
- Let $H_2^{-1}(x)$ be the inverse of $H_2(x)$ and $H_2^{-1}(x) \in [0, 1/2]$. For $p \in [0, 1]$, define $p * x := (1-p)x + p(1-x)$. Prove that $H_2(p * H_2^{-1}(x))$ is convex in $x$