Fundamentals of Information Theory

Basic Concepts

Yayu Gao

School of Electronic Information and Communications
Huazhong University of Science and Technology
Email: yayugao@hust.edu.cn

Outline

- Model of communication systems
- How to characterize the information source?
- How much information a message contains?
- What is entropy?
- Joint and conditional entropy
- Relative entropy and mutual information
- Entropies in communications
- Chain Rules
- Jensen’s Inequality and Log Sum Inequality
- Entropy rate: from single-outcome to sequence-outcome
- What is a Markov source?
- Differential Entropy: from discrete to continuous

本节学习目标

1. 画出香农提出的通信系统模型
2. 概述≥3种信源的分类方法
3. 说出离散单符号信源的数学模型
4. 概述信息量的建模过程
5. 写出自信息的定义与表达式
6. 说出≥2条自信息的性质
7. 辨别信息量与不确定度的关联与差异
8. 写出信息熵的定义与表达式
9. 说出≥3条信息熵的性质
10. 计算自信息和信息熵

重难点：

- 自信息与信息熵的定义
- 信息量与不确定度的关系
- 自信息与信息熵的性质
- 自信息与信息熵的计算

01

Model of Communication Systems

Model of Communication Systems

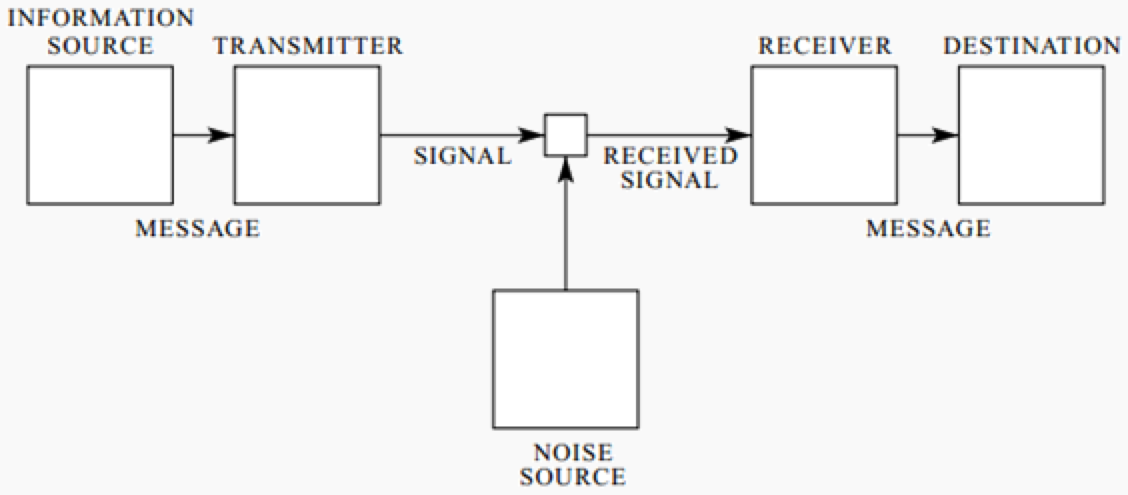
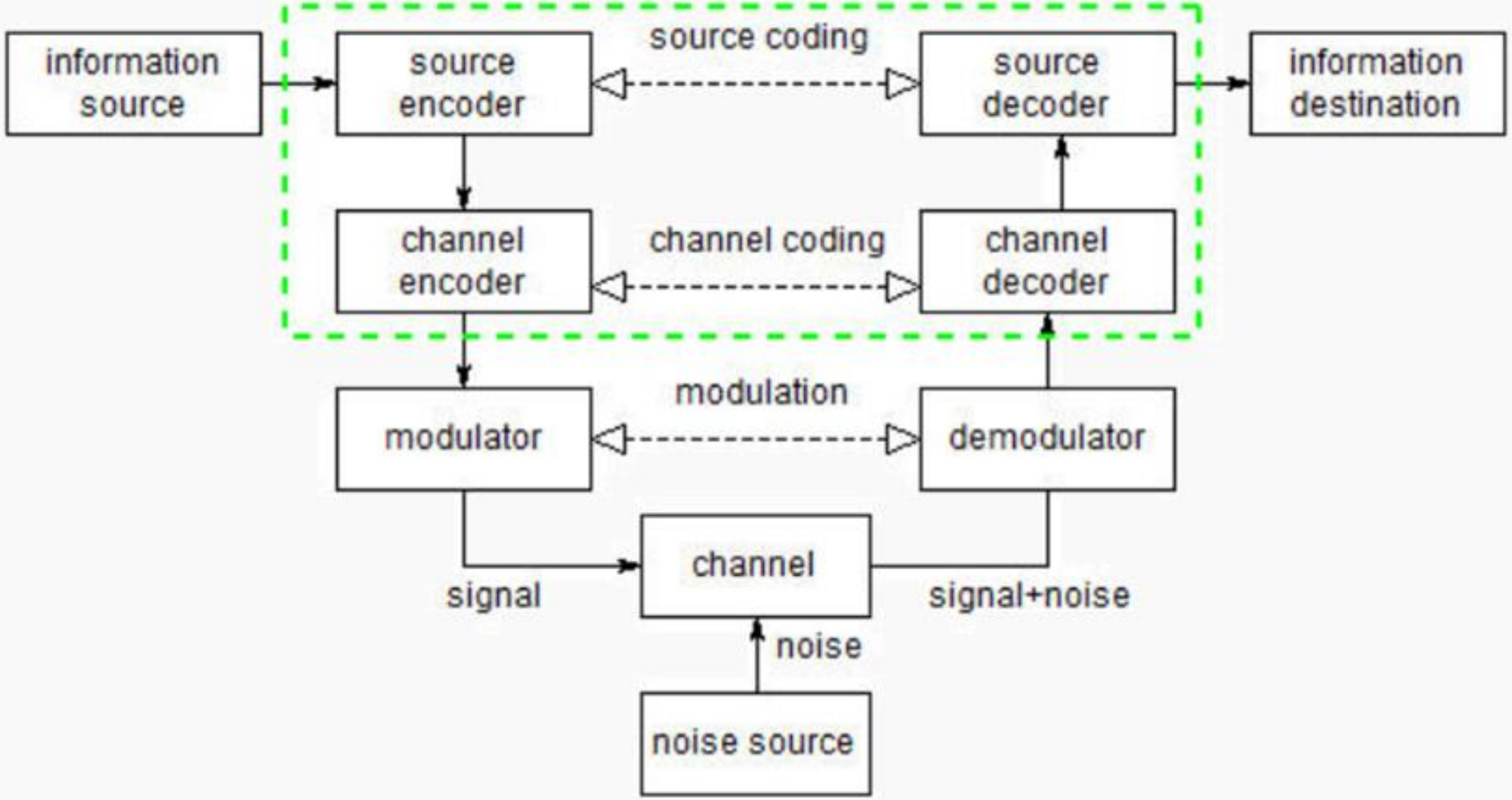


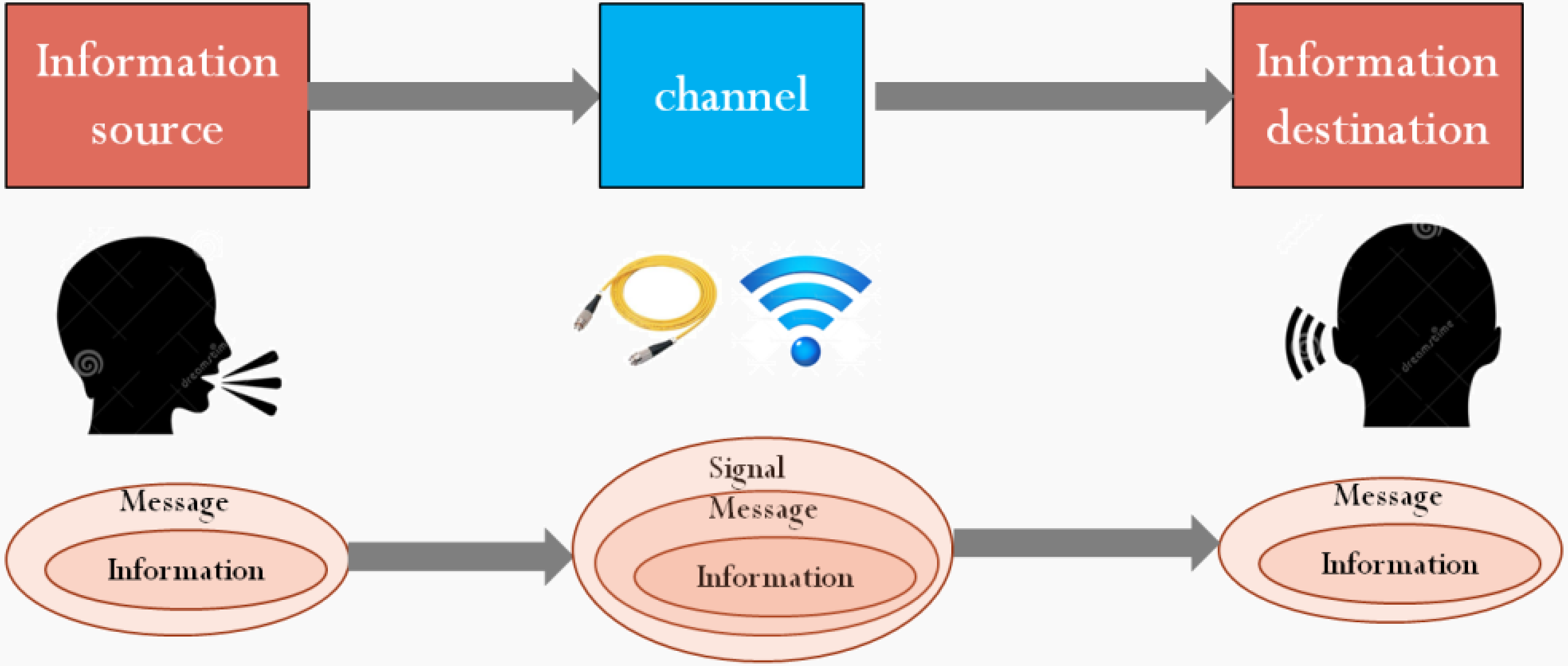
Fig. 1—Schematic diagram of a general communication system.

- Schematic diagram of a **general** communication system given in Shannon's landmark paper.

Block diagram of communication systems



What is the goal of communication?



- The essence of communication is to transmit information.

Major challenges for establishing a communication theory

- How to evaluate the performance of communication systems?
  - **Efficiency** in communication systems
    - How much information can be transmitted through one communication?
  - **Accuracy** for information transmission
    - When transmitting through a noisy channel, how much information is lost?

Fundamental Question:
How to quantify information?

Is information measurable?

身体的营养摄入

到底我们吃了多少?

重量：二两饭，半斤肉...
热量：300卡路里....

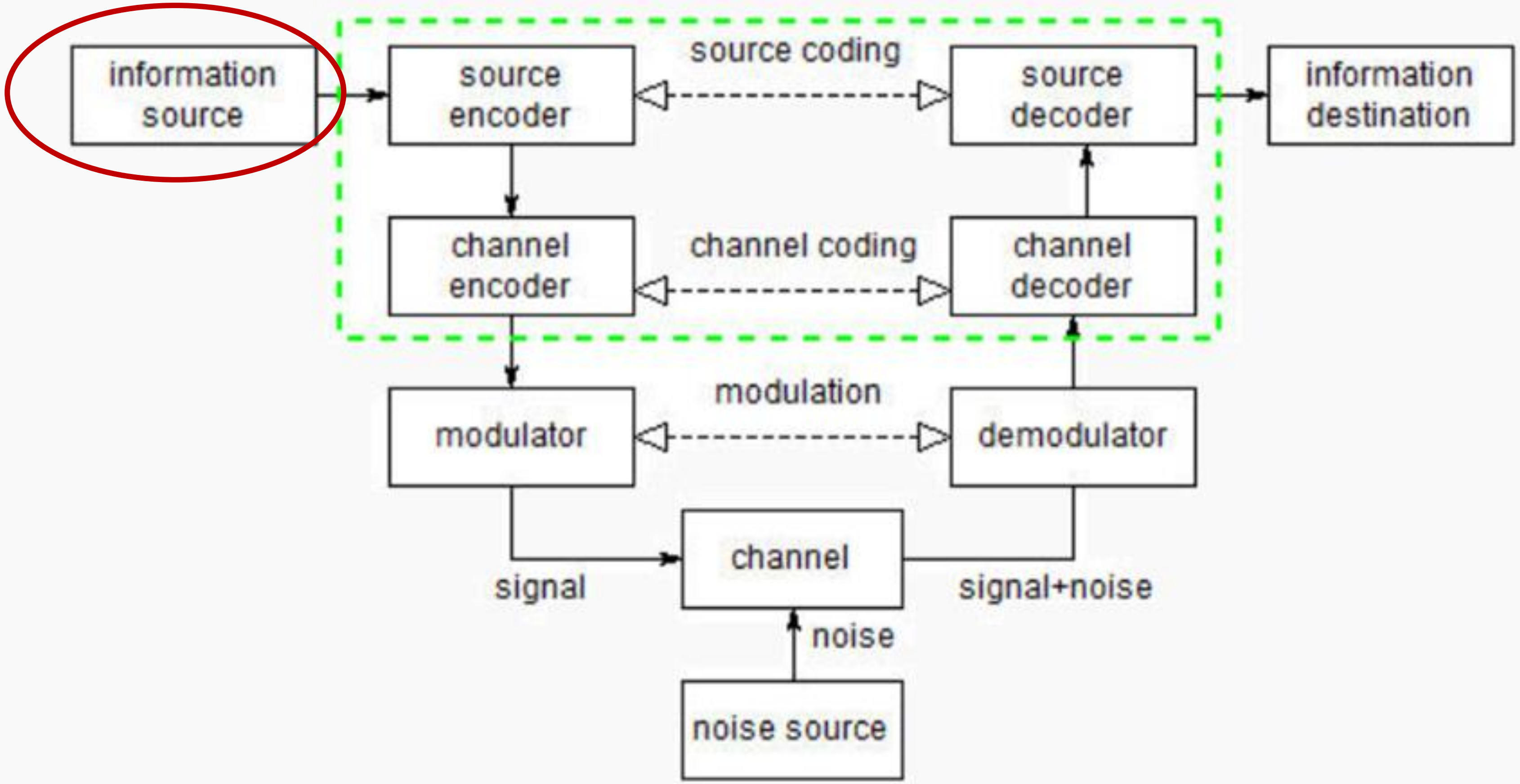
日常生活中的度量经验无法得出信息的度量方法、度量单位。

必须人为引入信息度量的定义。

信息?

灵魂的营养摄入

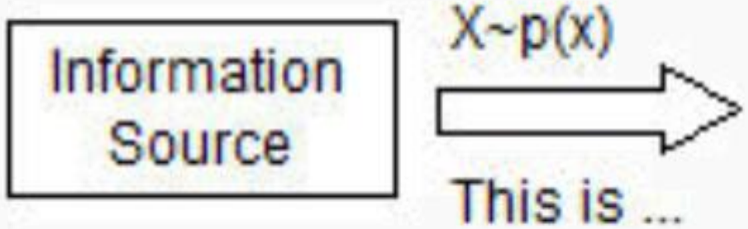
Quantify information– Where to begin?



02

How to characterize the information source?

How to characterize the information source?

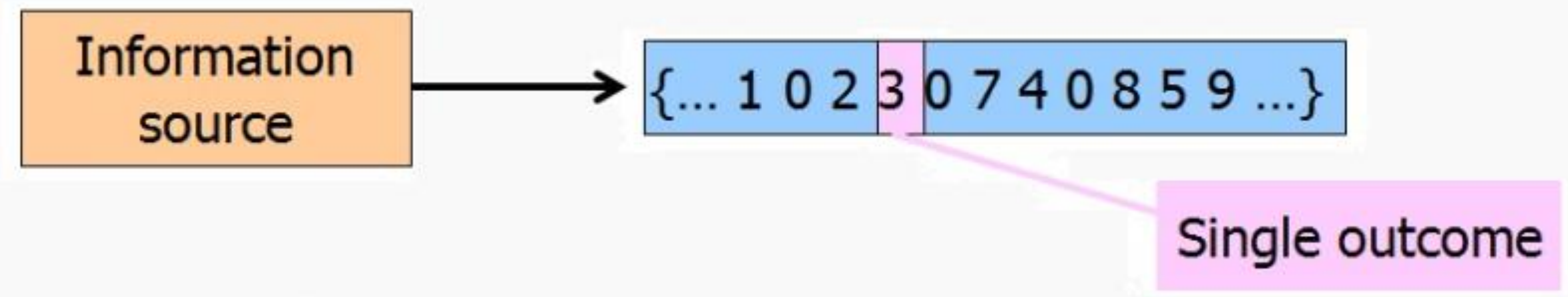


- The output sequence of information source is **stochastic**:
  - “We can think of a discrete source as generating the message, symbol by symbol.
  - It will choose successive symbols according to certain probabilities...
  - A mathematical model of a system, which produces such a sequence of symbols governed by a set of probabilities, is known as a stochastic process.
  - Any stochastic process which produces a discrete sequence of symbols chosen from a finite set may be considered as a discrete source."

--C.E. Shannon, 1948

Classification of sources: outcome of the source

- **Single** outcome or outcome **sequence**

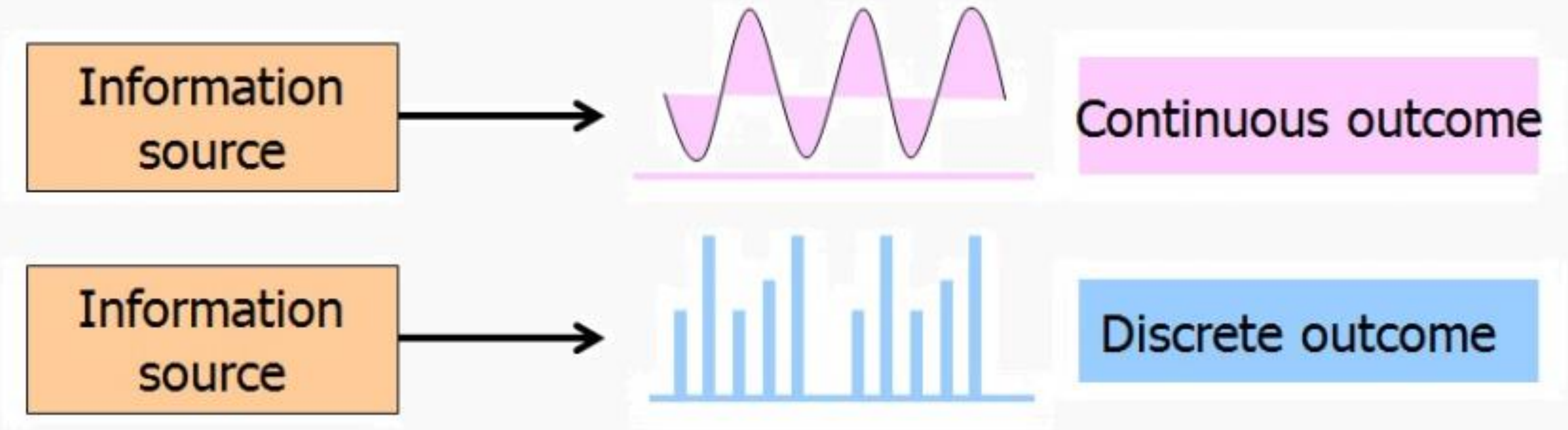


Information source

{... 1 0 2 3 0 7 4 0 8 5 9 ...}

Single outcome

- **Continuous** or **Discrete**



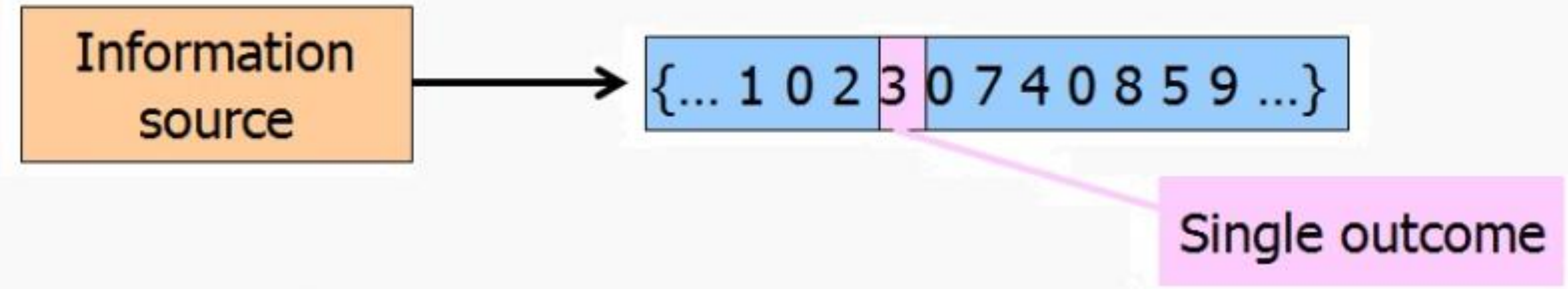
Information source

Continuous outcome

Information source

Discrete outcome

Modeling single outcome



Information source

{... 1 0 2 3 0 7 4 0 8 5 9 ...}

Single outcome

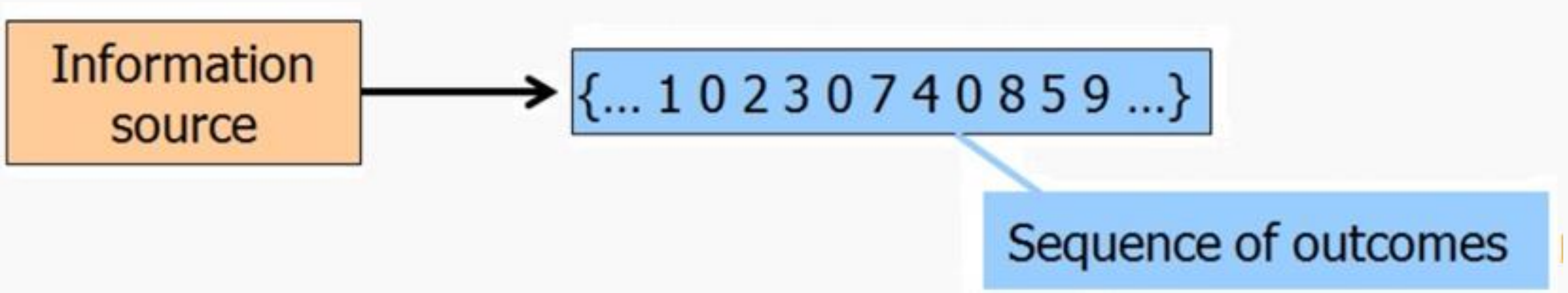
- Continuous Source

$$\left[\begin{array}{c} R \\ p(x) \end{array}\right], \int_R p(x)dx = 1$$

- Discrete Source

$$\left[\begin{array}{c} X \\ P(X) \end{array}\right] = \left[\begin{array}{cccc} a_1, & a_2, & \cdots & a_q \\ P(a_1), & P(a_2), & \cdots & P(a_q) \end{array}\right], \sum_{i=1}^{q} P(a_i) = 1$$

Modeling outcome sequence



Information source

{... 1 0 2 3 0 7 4 0 8 5 9 ...}

Sequence of outcomes

- Waveform Source
  - Continuous in both time and amplitude
  - Modeled as a continuous stochastic process $\{x(t)\}$

- Sequence Source
  - Sampled from waveform source
  - Discrete in time or space
  - Modeled as a stochastic sequence $\{X_i(t_i)\}$

Classification of sources

- **Stationary**: whether the distribution changes with time?
  - Stationary Source: **good** source (easy to analyze)
  - Un-stationary source: sometimes can be simplified as Markov source
- **Memory**: whether the variables in sequence have relationship?
  - Source without memory: **good** source (easy to analyze)
  - Source with memory: can be modeled as Markov source
- . . .

Sources studied in our course

- We study the ideal sources with **good properties**, then use them to approximate real sources.
  - Discrete Source
    - Single Outcome Discrete Source
    - Outcome sequence Discrete Source
      - Discrete stationary memoryless source
      - Discrete stationary source with memory
  - Continuous source
    - Waveform source

Always start from some **simple (naive) examples**

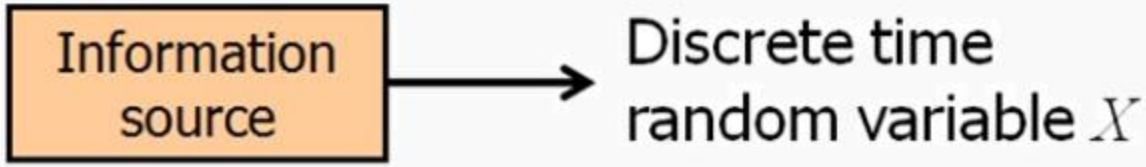
- Why study ideal sources? Why not real sources?

Information source model

Any examples?

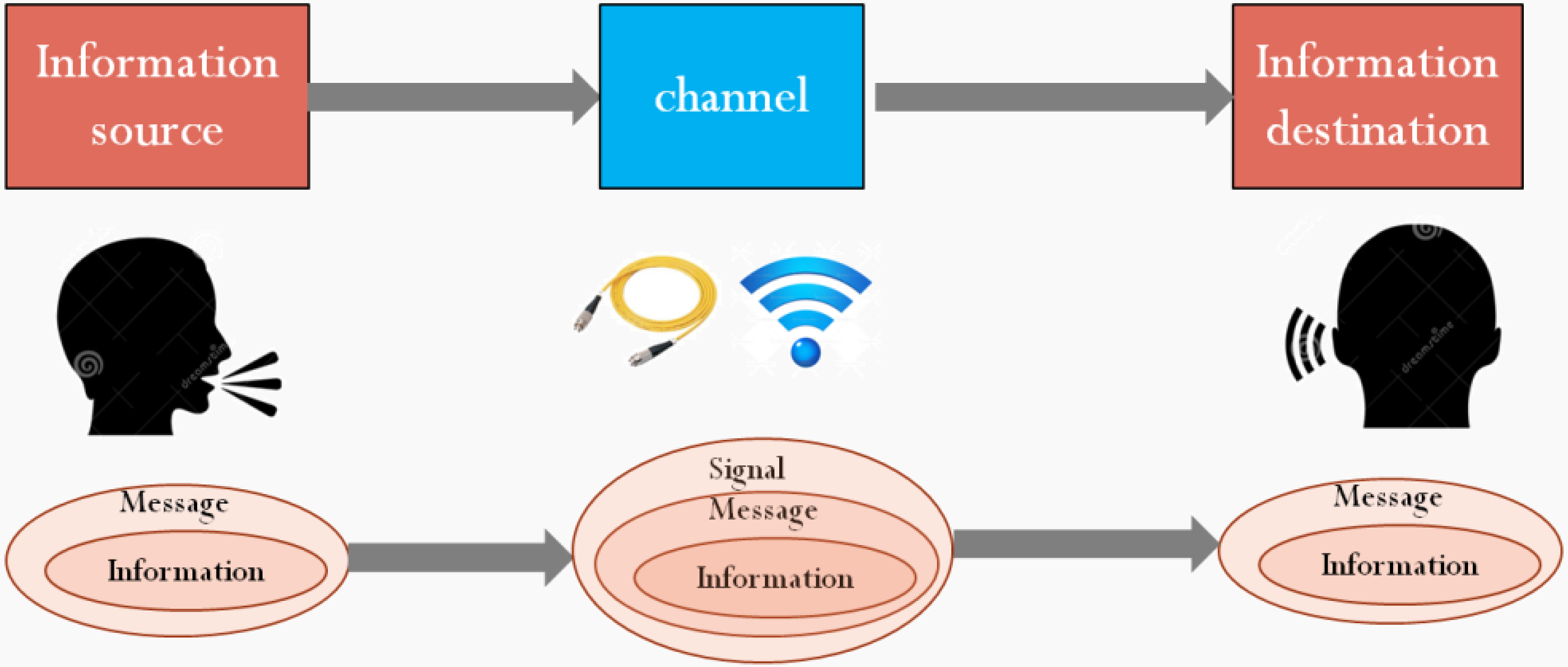
微助教讨论

- Consider a discrete single outcome source.



- Notations
  - Sample space: $\mathcal{X}$
  - Random variable ($r.v.$): X
  - Outcome of $\mathcal{X}$ or realization of X: x
  - Cardinality of set **X** (the number of elements): $|\mathcal{X}|$
- Probability mass function (p.m.f.)
  - $P(x) = Pr[X = x]$, $x \in \mathcal{X}$
  - $P(x, y) = Pr[X = x, Y = y]$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$

What is the goal of communication?

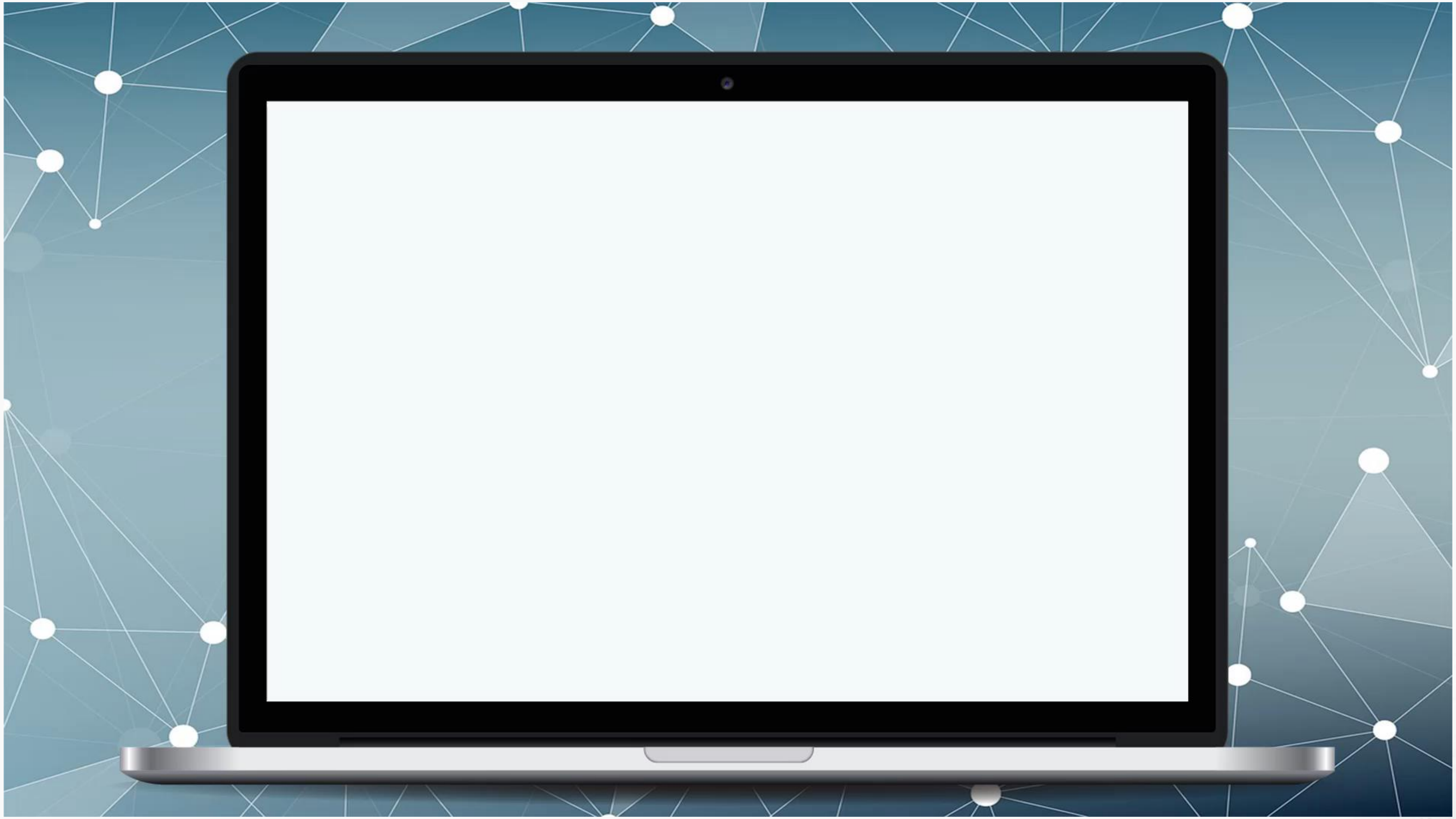


- How much information contained in a message generated by a source?

03

How much information a message contains?

How much information a message contains?



How to measure information?

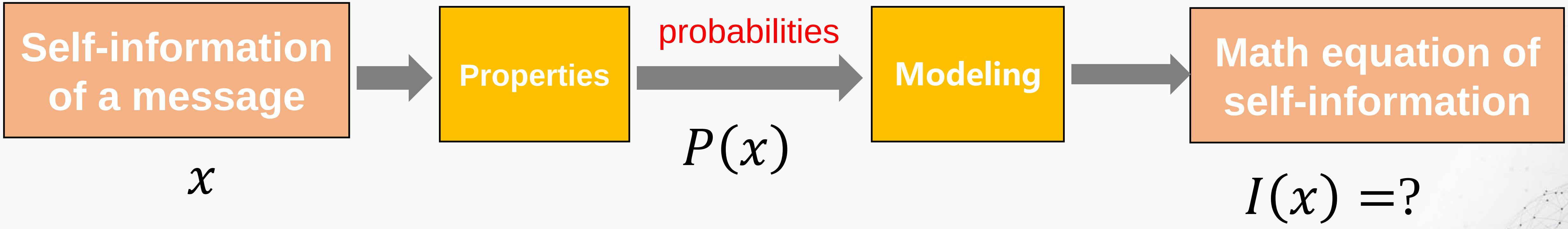
- The essence of information is to **eliminate uncertainty**.
  - The more uncertainty is eliminated after receiving the message, the more information is transmitted.

- How to use mathematical tools to model information?
  - Consider statistical information.
  - Uncertainty can be described by probability theory, stochastic process, and so on.

How to model the information contained in a message that is generated by a source?

Mathematical modeling of self-information

- Definition: **Self-information *I*(*x*)**
  - the amount of information contained in the message *x*
- Modeling steps:
  - investigate the properties of information
  - model them in probabilities



Self-information of a message

x

Properties

probabilities

$P(x)$

Modeling

Math equation of self-information

$I(x) = ?$

Mathematical modeling of self-information

- Property 1
  - Information contained in events ought to be defined in terms of some measure of the **uncertainty** of the events.

- Modeling
  - A nature measure of uncertainty of event is the **probability** of *x*, *P*(*x*).
  - Define the information in terms of *P*(*x*).

$$\boxed{I(x) = f(P(x))}$$

Mathematical modeling of self-information

- Property 2
  - Less certain events ought to contain more information than more certain events.

- Modeling
  - **Inversely proportional** to the probability

$$\boxed{P(a) > P(b) \to I(a) < I(b)}$$

Mathematical modeling of self-information

- Property 3
  - For a message with probability 1, what is its information?
  - For a message with probability 0, what is its information?

- Modeling
  - Non-linear mapping from probability to information

$$P(x) = 1 \rightarrow I(x) = 0$$
$$P(x) = 0 \rightarrow I(x) = \infty$$

Mathematical modeling of self-information

- Property 3
  - The total information of independent events should be the sum of the information of each event

- Modeling
  - Suppose a and b are two independent events. $P(a) = p_1 , P(b) = p_2$
  - (a, b) is considered together as a single event: $P(a, b) = p_1 p_2$

$$\boxed{\begin{gathered} I(a,b) = I(a) + I(b) \\ f(p_1 \cdot p_2) = f(p_1) + f(p_2) \end{gathered}}$$

What is the answer?

$$I(x) = f(P(x))$$

$$P(a) > P(b) \rightarrow I(a) < I(b)$$

$$P(x) = 1 \rightarrow I(x) = 0$$
$$P(x) = 0 \rightarrow I(x) = \infty$$

$$f(p_1 \cdot p_2) = f(p_1) + f(p_2)$$

Exponential,
Logarithm
....

$$\boldsymbol{I(x)} = ?$$

Self-information: definition

- Suppose the event x with probability p(x), its **self-information** is defined by

$$I(x) = -\log\left[p(x)\right] = log\left[\frac{1}{p(x)}\right].$$

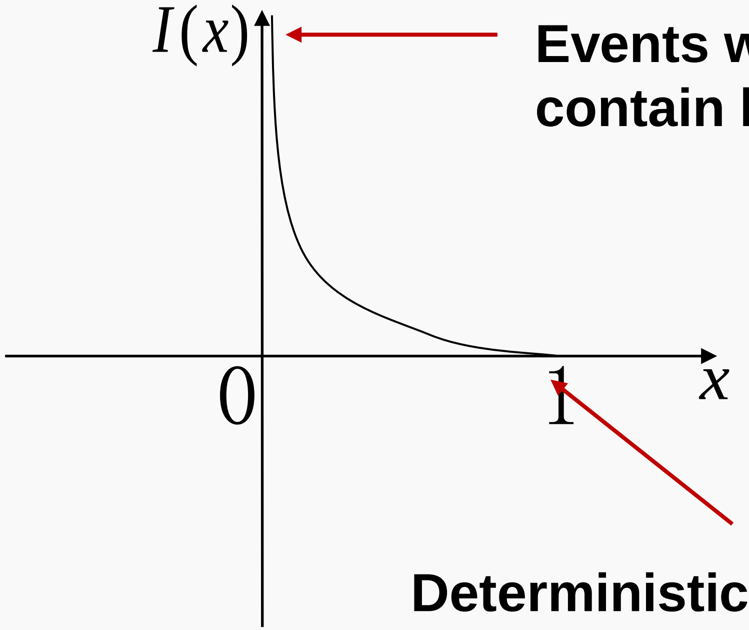
- The **only form** that satisfies all the properties of information.

- The base of logarithm can be any
  - 2: information measured in binary units (**bits**)
  - e: information measured in natural units (nats)
  - By default, base is 2.

Self-information: Summary

- Self-information

$$I(x) = -\log\left[p(x)\right] = log[\frac{1}{p(x)}].$$



Events with low prob. contain huge information.

Deterministic events contain no information.

非负性：

- 随机事件的发生**总能提供一些信息，最差是0**.
- 不会因为事件发生而使不确定性增大

递减性：

- 概率越大的事件，不确定性越小，发生提供的信息量越小

Information vs. Uncertainty

- For an event, its self-information and uncertainty has the same mathematical definition:

$$I(x) = -\log[p(x)] = log[\frac{1}{p(x)}].$$

- Question: Any **difference** between them?

- **例：袋子里一共有100个手感相同的球。已知其中99个是红球，只有1个是白球。现随机抽取1个球**
  - **问：抽球之前：抽出来的球会是红球吗?**
  - **回答：不能确定，不过多半是。**

Information vs. Uncertainty

- **Uncertainty 不确定度：**

在随机实验进行前，关于某随机事件在这次实验中是否会发生的不确定程度。

- **“抽中的是红球 ” 不确定度** $=\log\left(1/\,p(\text{红球})\right)=\log(1/0.99)\approx 0.0145$ **比特**
- **“抽中的是白球 ” 不确定度** $=\log\left(1/\,p(\text{白球})\right)=\log(1/0.01)\approx 6.644$ **比特**

- **Information 信息量：**

抽球之后，某次抽出的是红球，并明确地告诉你答案。计算你所获得的信息量。

- **获得的信息量 =（抽球前，对于抽中红球的不确定度）—（抽球后，对于抽中红球仍存在的不确定度）=0.0145-0=0.0145比特**
- **类似地，当抽中的是白球，获得的信息量为6.644比特。**

Summary: Information vs. Uncertainty

- **Uncertainty 不确定度：**

  在随机实验进行前，关于某随机事件在这次实验中是否会发生的不确定程度。

- **Information 信息量：**

  某次随机实验完成后，出现某个随机事件时所获得的信息量。

根据自信息的定义，重新回答之前的问题

新闻里说中国男乒乓球队战胜了巴西男乒乓球队，和中国男足战胜了巴西男足，分别获得多少信息?

A. 前者多　　✓ 后者多　　C. 一样多

在男子乒乓球赛事中，中国与巴西共交战16次，巴西队仅获胜1次。
在男子足球赛事中，中国队从未战胜过巴西队。

$$P(\mathrm{A}) = \frac{15}{16}$$

$$I(\mathrm{A}) = -\log\left[\frac{15}{16}\right] = 0.093\ \mathrm{bits}$$

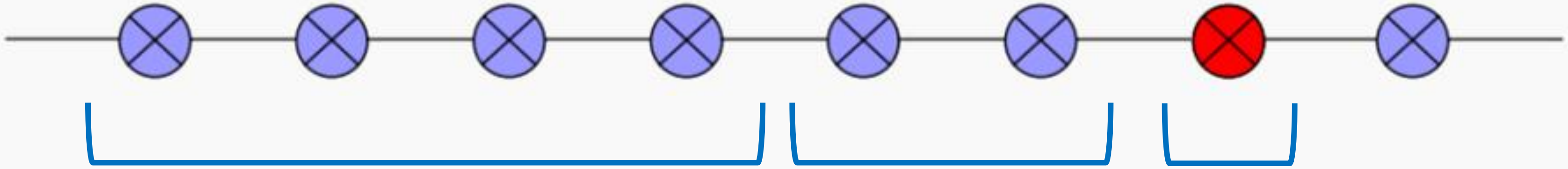
Example: meaning of 1 bit

- In information theory, one bit is typically defined as the uncertainty of a binary random variable that is 0 or 1 with equal probability.
- Or the information that is gained when the value of such a variable is known.

Always start from some **simple (naive) examples**

Example: meaning of bits

- A series circuit with 8 bulbs, one of which burns out.



- How many times at least do you need to identify the burnt-out one?
- Now see it from the **perspective of information theory**

$$I(x) = -\log\left(\frac{1}{8}\right) = 3\ bits$$

Perspective of Information Theory

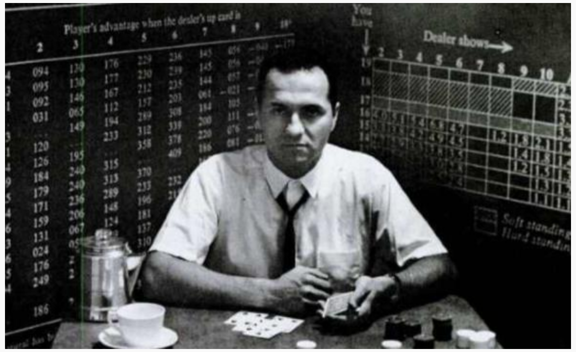
- One is given 24 coins. It is known that precisely one coin is fake, which weights differently compared with genuine coins. It is not clear whether this fake coin is heavier or lighter, though. Now we use a balance to identify the fake coin, but we do not have the weights for this balance.
- **What is the minimum number of the weighting operations in order to identify this fake coin?**

How much uncertainty?

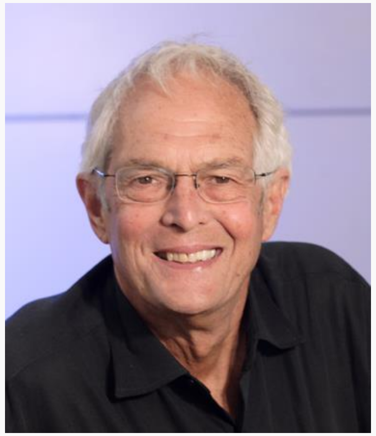
How much information can be obtained from each experiment?

Perspective of Information Theory

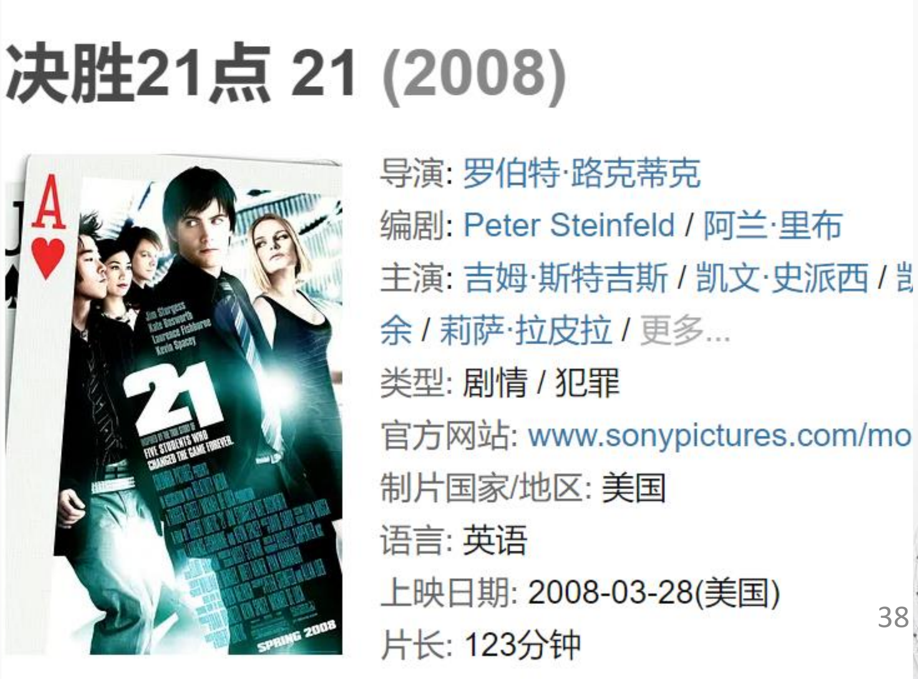
Claude Shannon



Ed Thorp



Thomas M. Cover



决胜21点 21 (2008)

导演: 罗伯特·路克蒂克

编剧: Peter Steinfeld / 阿兰·里布

主演: 吉姆·斯特吉斯 / 凯文·史派西 / 凯

余 / 莉萨·拉皮拉 / 更多...

类型: 剧情 / 犯罪

官方网站: www.sonypictures.com/mo

制片国家/地区: 美国

语言: 英语

上映日期: 2008-03-28(美国)

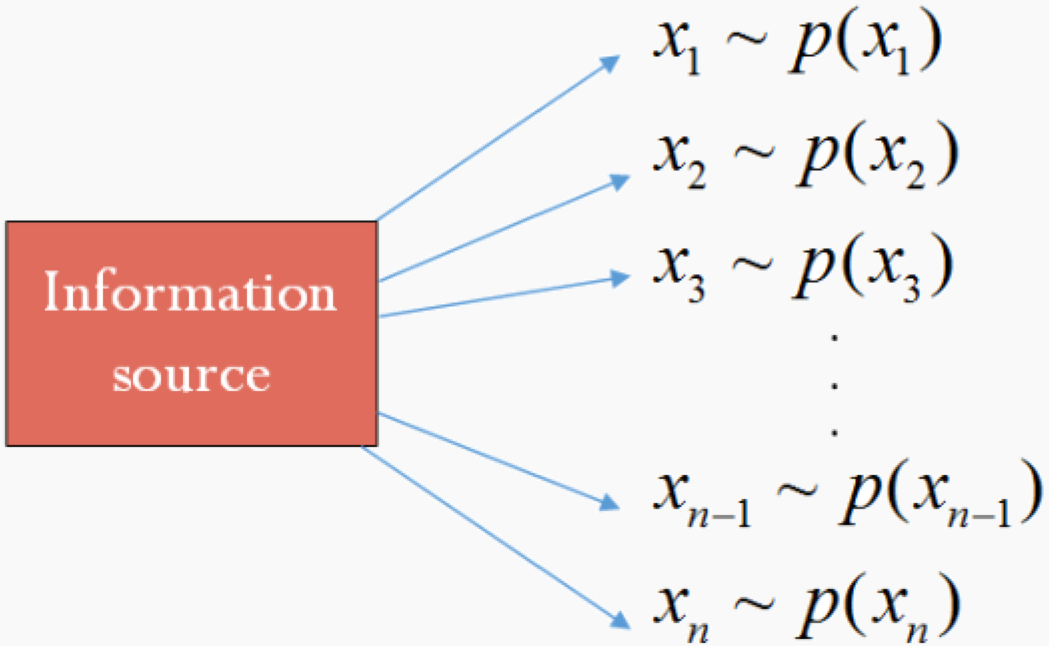
片长: 123分钟

04

What is entropy?

Information of a **source**

- For a discrete random variable, how to define the measure of information?



Information source

$x_1 \sim p(x_1)$

$x_2 \sim p(x_2)$

$x_3 \sim p(x_3)$

$\vdots$

$x_{n-1} \sim p(x_{n-1})$

$x_n \sim p(x_n)$

- Each outcome x has its self-information of I(x)

- The average information of the source
  - the **expectation** of self-information

$$\sum p(x) \cdot I(x)$$

Entropy: definition

- The average information of *r.v.* X is called the **entropy** of X

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log [p(x)]$$

熵

- A (convenient) **measure of uncertainty** of a *r.v.*.
- Entropy is a function of the **probability distribution**
    - Independent of the outcomes of the *r.v.* itself
    - Only the distribution matters
- Logarithm
    - Base can by any, i.e., 2 by default. (bits)
    - Base b is sometimes marked as $H_b(X)$.

Entropy: basic properties

- Entropy is the **expected** value of self-information

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

- Entropy $H(X)$ is **non-negative**.

$$0 \leq p(x) \leq 1 \Rightarrow \log\left(\frac{1}{p(x)}\right) \geq 0$$

$$H(X) = E\left[\log \frac{1}{p(x)}\right] \geq 0$$

- Change of bases, since

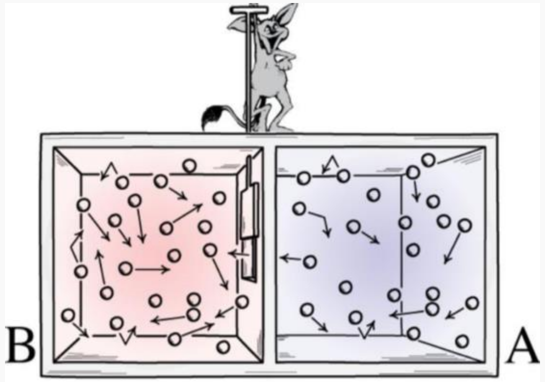
$$H_b(X) = [\log_b a] \cdot H_a(X)$$

$$\log_b p = [\log_b a] \cdot \log_a p$$

Entropy: physical meaning

- Entropy *H*(*X*) has three physical meanings
  - **Before** the outcome of source, the average uncertainty the source has.
  - **After** the outcome of source, the average information each message can provide.
  - Refection of the randomness of the *r.v. X*.

- Question: Why did Shannon call it **entropy**?



Maxwell's demon

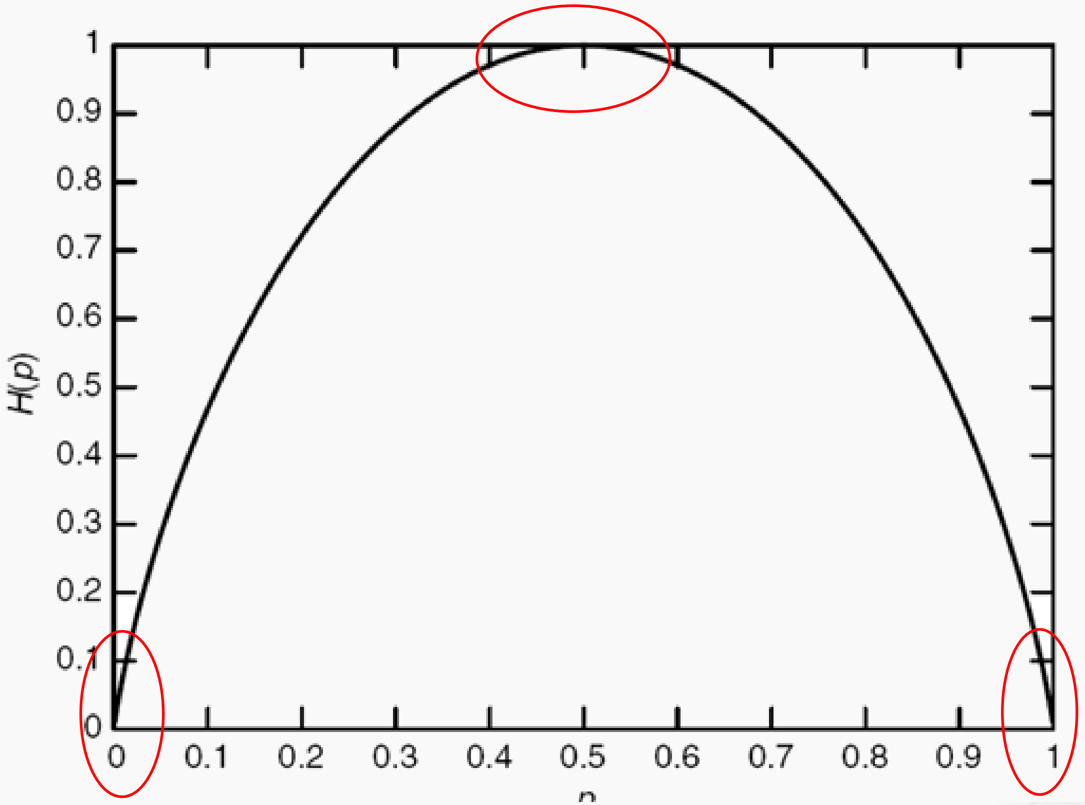
Entropy: examples

- For a binary random variable *X*

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ p & 1-p \end{bmatrix}$$

- Its entropy *H*(*X*) is given by

$$H(X) = -p \log p - (1-p)\log(1-p)$$



$p = 0, H(X) = 0$ → Deterministic system doesn’t contain information.

$p = 0.5, H(X) = 1$ → Uniform distribution maximizes the entropy?

Entropy: Example #1

- **设某信源输出为掷一非均匀骰子的点数，若其任一面出现的概率与该面的点数成正比。试求该信源的信源熵?**

- **解：首先要求解信源模型**
  - **设出现1点的概率为p，因概率分布满足归一性:**

如果是均匀骰子呢?

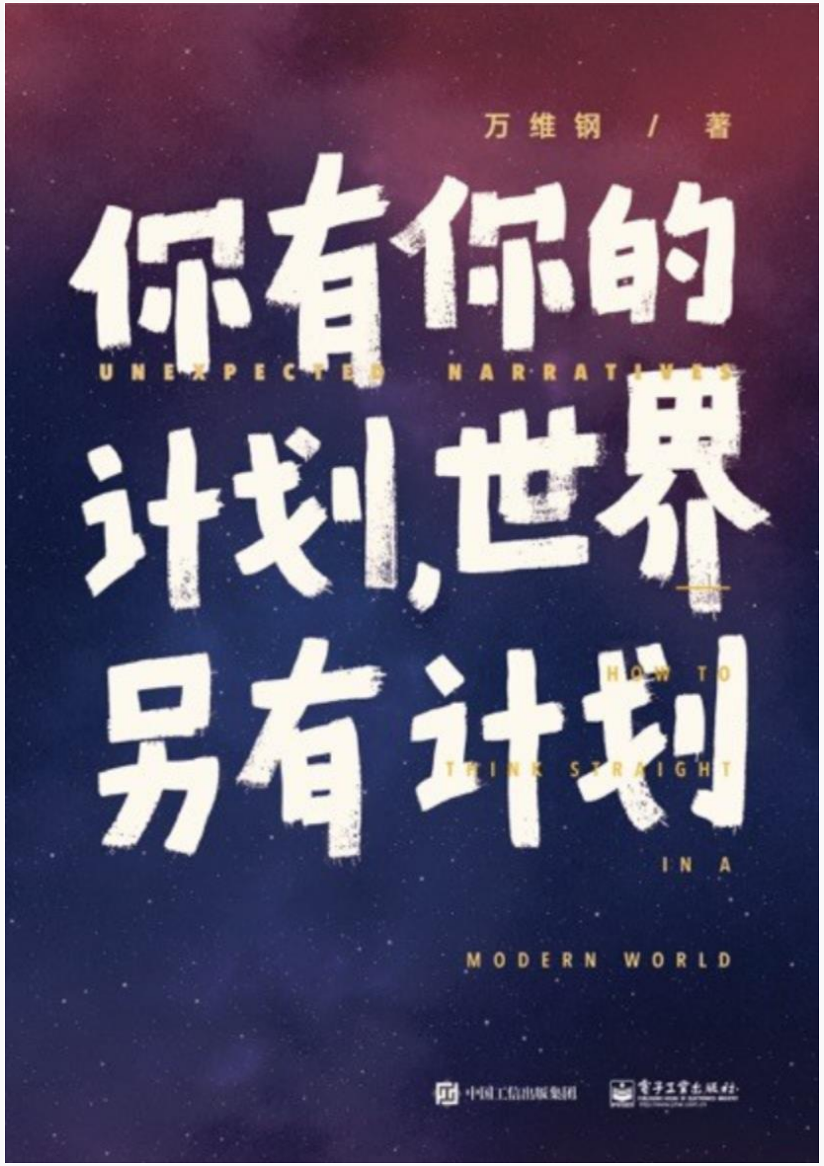
$$p+2p+3p+4p+5p+6p=1 \Rightarrow p=\frac{1}{21}$$

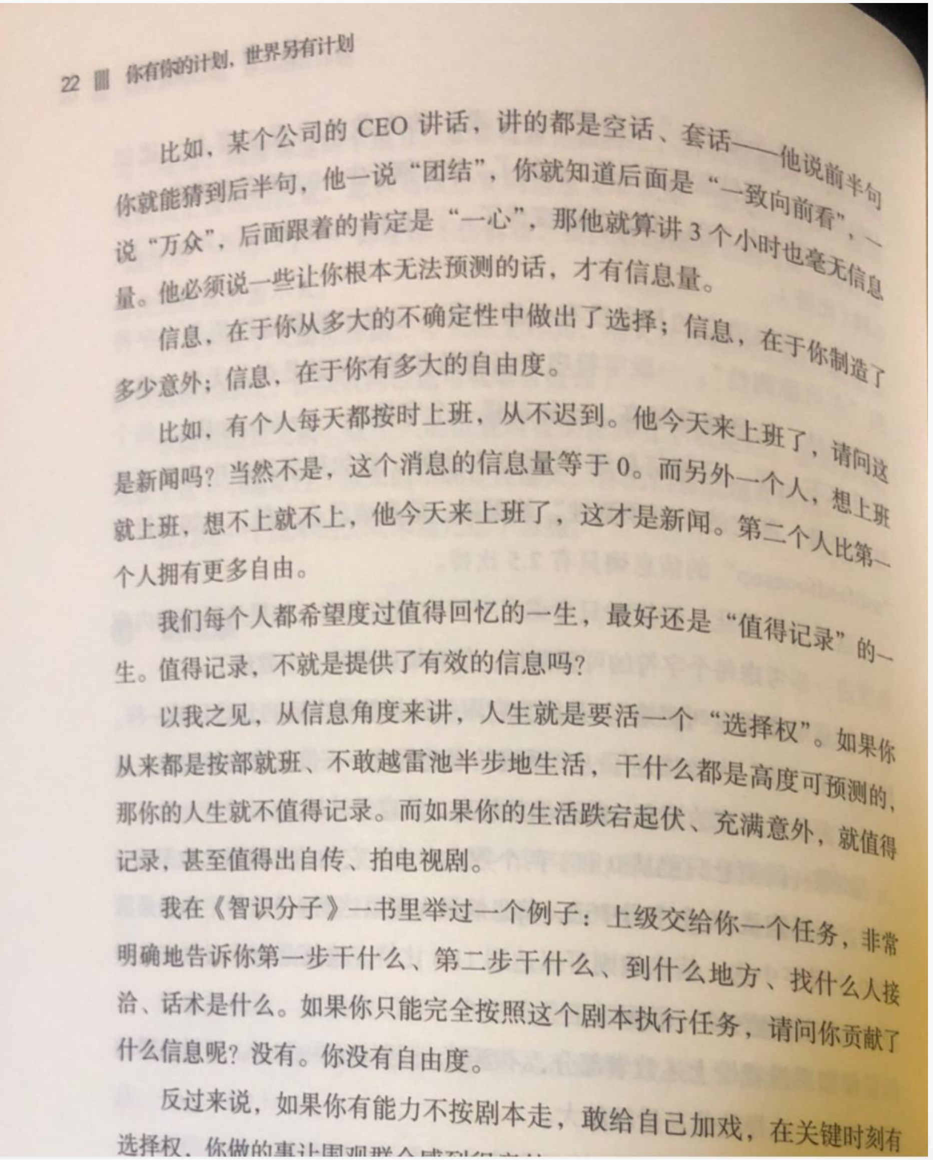
$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{21} & \frac{2}{21} & \frac{3}{21} & \frac{4}{21} & \frac{5}{21} & \frac{6}{21} \end{bmatrix}$$

$$\therefore\ H(X)=-\frac{1}{21}\cdot\log\frac{1}{21}-\frac{2}{21}\cdot\log\frac{2}{21}-\cdots-\frac{6}{21}\cdot\log\frac{6}{21}$$

$$\approx 2.4$$ 比特/符号

Entropy: examples





22 ▥ 你有你的计划，世界另有计划

比如，某个公司的 CEO 讲话，讲的都是空话、套话——他说前半句你就能猜到后半句，他一说“团结”，你就知道后面是“一致向前看”，一说“万众”，后面跟着的肯定是“一心”，那他就算讲 3 个小时也毫无信息量。他必须说一些让你根本无法预测的话，才有信息量。

信息，在于你从多大的不确定性中做出了选择；信息，在于你制造了多少意外；信息，在于你有多大的自由度。

比如，有个人每天都按时上班，从不迟到。他今天来上班了，请问这是新闻吗？当然不是，这个消息的信息量等于 0。而另外一个人，想上班就上班，想不上就不上，他今天来上班了，这才是新闻。第二个人比第一个人拥有更多自由。

我们每个人都希望度过值得回忆的一生，最好还是“值得记录”的一生。值得记录，不就是提供了有效的信息吗？

以我之见，从信息角度来讲，人生就是要活一个“选择权”。如果你从来都是按部就班、不敢越雷池半步地生活，干什么都是高度可预测的，那你的人生就不值得记录。而如果你的生活跌宕起伏、充满意外，就值得记录，甚至值得出自传、拍电视剧。

我在《智识分子》一书里举过一个例子：上级交给你一个任务，非常明确地告诉你第一步干什么、第二步干什么、到什么地方、找什么人接洽、话术是什么。如果你只能完全按照这个剧本执行任务，请问你贡献了什么信息呢？没有。你没有自由度。

反过来说，如果你有能力不按剧本走，敢给自己加戏，在关键时刻有选择权，你做的事让围观群众感到很

Entropy for Life

“人活着就是在**对抗熵增定律**，生命以负熵为食。”

——薛定谔《生命是什么》

“如果地球毁灭了，我们怎么能够在一张名片上写下地球文明的全部精髓，让其他文明知道我们曾有过这个文明。”

$1+1=2$（代表了数学文明）

$E=mc^2$（爱因斯坦的质能方程）

$S=-\sum P\ln P$ (熵的定义)

——吴军《谷歌方法论》

Summary

- Model of communication systems
- How to characterize the information source?
- How much information a message contains?
- What is entropy?
- Joint and conditional entropy
- Relative entropy and mutual information
- Entropies in communications
- Chain Rules
- Jensen’s Inequality and Log Sum Inequality
- Entropy rate: from single-outcome to sequence-outcome
- What is a Markov source?
- Differential Entropy: from discrete to continuous

本节学习目标

1. 画出香农提出的通信系统模型
2. 概述≥3种信源的分类方法
3. 说出离散单符号信源的数学模型
4. 概述信息量的建模过程
5. 写出自信息的定义与表达式
6. 说出≥2条自信息的性质
7. 辨别信息量与不确定度的关联与差异
8. 写出信息熵的定义与表达式
9. 说出≥3条信息熵的性质
10. 计算自信息和信息熵

重难点：

- ➢ 自信息与信息熵的定义
- ➢ 信息量与不确定度的关系
- ➢ 自信息与信息熵的性质
- ➢ 自信息与信息熵的计算

Thank you!

My Homepage

Yayu Gao

School of Electronic Information and Communications
Huazhong University of Science and Technology
Email: yayugao@hust.edu.cn