# 信息论导论

## 第1讲 绪论，熵、联合熵、条件熵

[信息论教材中页码范围] 熵，联合熵，条件熵: p13~18,
熵的链式法则：p22~23

信息学部-信息科学与技术学院 吴绍华
hitwush@hit.edu.cn

# 为什么要开这么课?



- 信息时代
- 信息学部
- 信息学科
- 信息技术
- 信息……

# 从几个最直接的问题开始



1. 什么是“信息”？

2. 什么是“信息论”？

3. 我们可以从这门课中学到什么？

4. 为什么叫“导论”课

# 1. 什么是“信息”？



- **信息的定义**
  - 【英】“Information” —— Something informed
  - 【中】“信息” —— 音讯、**消息（message)**
  - 在中文的日常用语习惯中，通常将信息等同于消息
  - 在信息论中，我们要严格区分：**消息是信息的载体**
  - **并不是所有的消息均包含信息（举例)**
  - 信息的定义（一个简短的定性定义）：
    - 信息就是**不确定性**
  - **严格定义？定量描述，即度量**

# 1. 什么是“信息”？



- **信息的度量**
  - 不确定性 → 随机性； 具有不确定性的事件 → 随机事件。随机事件的研究工具?
  - 所以，应该用“概率”作为基本工具去度量信息。

**定义**

消息的香农信息量定义为消息（随机事件）$x_i$ 的不确定度，可由下式计算：

$$I(x_i) = -\log_2 p(x_i)$$

其中 $p(x_i)$ 为 $x_i$ 的概率，信源的信息量被定义为该信源所有可能消息的平均香农信息量，其在信息论里通常被称作信息熵。具体表达式如下：

$$H(X) = -\sum_{i=1}^{q} p(x_i)\log_2 p(x_i)$$

# 2. 什么是“信息论”？



- **信息论的提出、发展背景**
  - **人类社会运用信息的历史：5个里程碑**

速率 ↑

距离 ↑

- 1st: 语言
- 2nd: 文字
- 3rd: 印刷术
- 4th: 电报，电话 …
- 5th: 数字通信

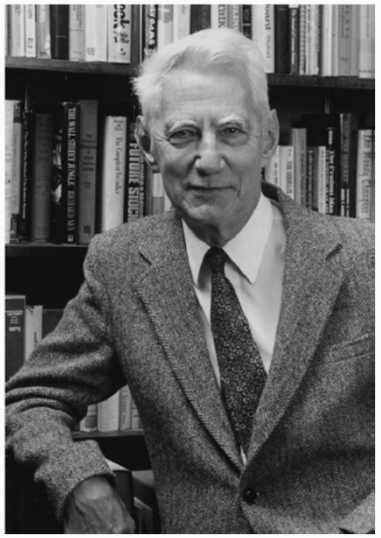
需要**定量**评估系统性能、代价/成本

需要建立起相应**数学理论**作为指导依据

# 2. 什么是“信息论”？



- **信息论的提出、发展背景**
  - **人类社会运用信息的历史：5个里程碑**
  - **信息论的创立**



**20世纪40年代**

- N. Wiener（维纳） —— 控制
- C. Shannon（香农） —— 通信
- R. A. Fisher（费舍尔）—— 统计学

**公认的信息论之父： C. Shannon**

长时间（早期~黄金发展阶段）的贡献与引领（1948~1973）

奠基性著作

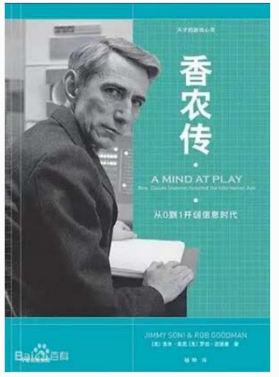
- A mathematical theory of communication, Bell Systems Tech. J, 27: 623–656,1948. （We are sorry to inform you ……）
- Communication in the presence of noise. Proceedings of the IRE, 37(1):10–21, 1949.

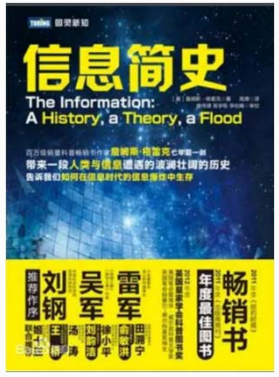
# 2. 什么是“信息论”？



- **信息论的提出、发展背景**
  - **人类社会运用信息的历史：5个里程碑**
  - **信息论的创立**
  - **Claude. Elwood. Shannon其人其事**
    - 1916.4.30－2001.2.26
    - 2016.4.30 —— 香农诞辰100周年纪念研讨会、纪念性论文……
    - “遇见Shannon”系列论文
    - 晚年香农，香农与股票投资……
    - **天赋 + 兴趣**

[5] Marconi: D Tse, Modern wireless communication: When Shannon meets Marconi, 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, May 14-19, 2006, Toulouse, France, 2006.
[6] Nyquist: YX Chen, AJ Goldsmith and YC Eldar, Shannon meets Nyquist: The interplay between capacity and sampling, 49th Annual Allerton Conference on Communication, Control, and Computing, Sept. 28-30, Monticello, IL, USA, 2011.
[7] Nyquist: YX Chen, YC Eldar, and AJ Goldsmith, Shannon meets Nyquist: Capacity of sampled Gaussian channels, IEEE Transactions on Information Theory, 39(8): 4889-4914, 2013.
[8] Wiener: GD Forney, On the role of MMSE estimation in approaching the information-theoretic limits of linear Gaussian channels: Shannon meets Wiener, 41st Annual Allerton Conference on Communication, Control and Computing, Oct. 1-3, 2003, Monticello, IL, USA, 2003; and Shannon meets Wiener II: On MMSE estimation in successive decoding schemes, 2004.
[15] Bellman: S Meyn and G Mathew, Shannon meets Bellman: Feature based Markovian models for detection and optimization, 47th IEEE Conference on Decision and Control, Dec. 9-11, 2008, Cancun, Mexico, 2008.
[16] Nash: RA Berry and DNC Tse, Shannon meets Nash on the interference channel, IEEE Transactions on Information Theory, 57(5): 2821-2836, 2011.
[17] Moore: L Harrison, Moore's law meets Shannon's law: The evolution of the communication's industry, IEEE International Conference on Computer Design: VLSI in Computers and Processors, Sept. 23-26, 2001, Austin, TX, USA, 2001.
[18] Moore: S Scholl, S Weithoffer and N When, Advanced iterative channel coding schemes: When Shannon meets Moore, 9th International Symposium on Turbo Codes and Iterative Information, Sept. 5-9, 2016, Brest, France, 2016.
[19] Kalman: A Gattami, Kalman meets Shannon, arXiv:1404.4350, 2014.



# 2. 什么是“信息论”？



- **信息论的提出、发展背景**
  - **人类社会运用信息的历史：5个里程碑**
  - **信息论的创立**
  - **Claude. Elwood. Shannon其人其事**
  - **信息（论）领域的最高学术奖项**

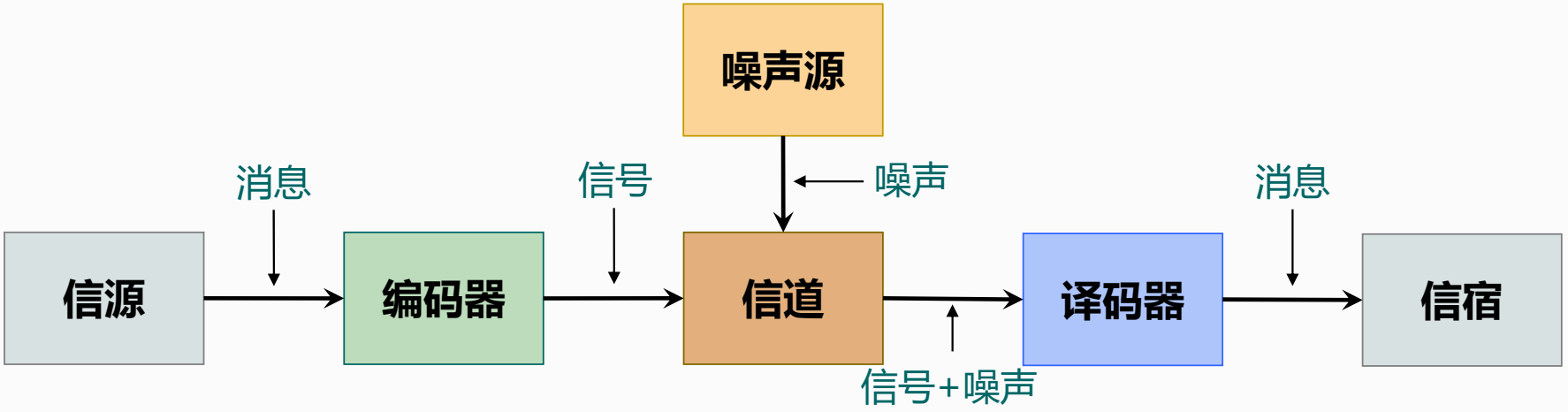
The Claude E. Shannon Award of the IT Society has been instituted to honor consistent and profound contributions to the field of information theory. Each Shannon Award winner is expected to present a Shannon Lecture at the following IEEE International Symposium on Information Theory. The first Shannon Lecturer was Claude Shannon himself.

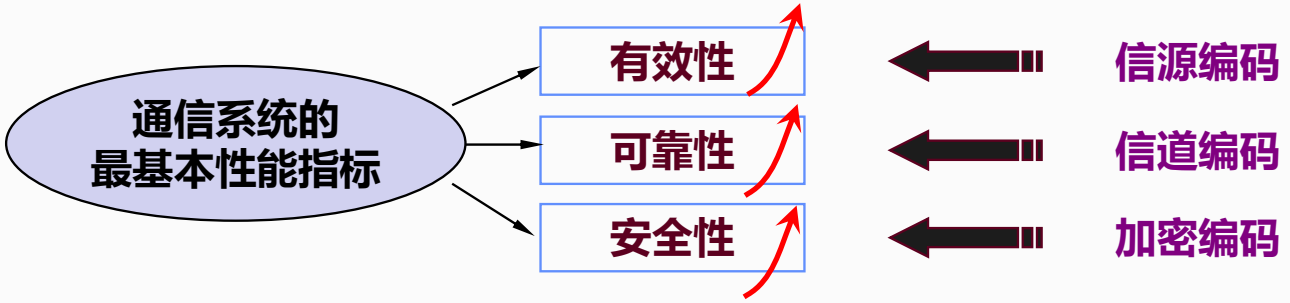
- David S. Slepian, (1923-2007) (1974)
- Robert M. Fano, (1917-2016) (1976)
- Peter Elias, (1923-2001) (1977)
- Mark S. Pinsker (1978)
- Jacob Wolfowitz, (1910-1981) (1979)
- William Wesley Peterson, (1924-2009) (1981)
- Irving S. Reed, (1923-2012) (1982)
- Robert G. Gallager (1983)
- Solomon Golomb, (1932-2016) (1985)
- William L. Root, (1919-2007) (1986)
- James Massey, (1934-2013) (1988)
- Thomas M. Cover, (1938-2012) (1990)
- Andrew J. Viterbi (1991)
- Elwyn Berlekamp (1993)
- Aaron D. Wyner, (1939-1997) (1994)
- G. David Forney, Jr. (1995)
- Imre Csiszár (1996)
- Jacob Ziv (1997)
- Neil J.A. Sloane (1998)
- Tadao Kasami (1999)
- Thomas Kailath (2000)
- Jack Keil Wolf, (1935-2011) (2001)
- Toby Berger (2002)
- Lloyd R. Welch (2003)
- Robert J. McEliece (2004)
- Richard Blahut (2005)
- Rudolf Ahlswede, (1938-2010) (2006)
- Sergio Verdú (2007)
- Robert M. Gray (2008)
- Jorma Rissanen (2009)
- Te Sun Han (2010)
- Shlomo Shamai, (Shitz) (2011)
- Abbas El Gamal (2012)
- Katalin Marton (2013)
- János Körner (2014)
- Robert Calderbank (2015)
- Alexander S. Holevo (2016)
- David Tse (2017)
- Gottfried Ungerboeck (2018)
- Erdal Arıkan (2019)
- Charles Bennett (2020)
- **Alon Orlitsky (2021)**
- **Raymond Wai-Ho Yeung (2022)**
- **Rüdiger Urbanke (2023)**
- **Andrew R Barron (2024)**
- **Peter Shor (2025)**

# 2. 什么是“信息论”？



- **信息论的研究对象**
  - **信息流通系统：信息的感知、获取、存储、传输、处理、应用系统**
  - **最典型的例子：信息传输系统（通信系统），“将信息从一点准确或近似传递到另外一点”**

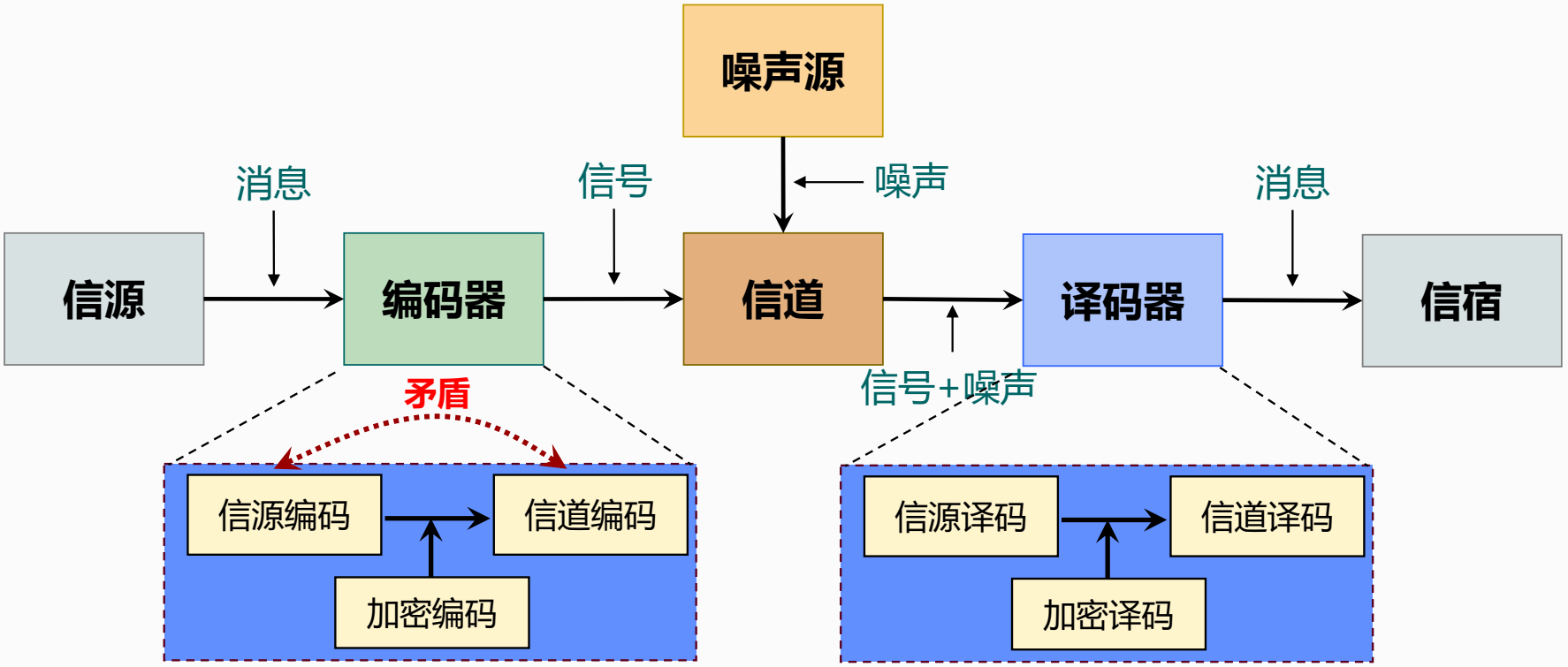


# 2. 什么是“信息论”？



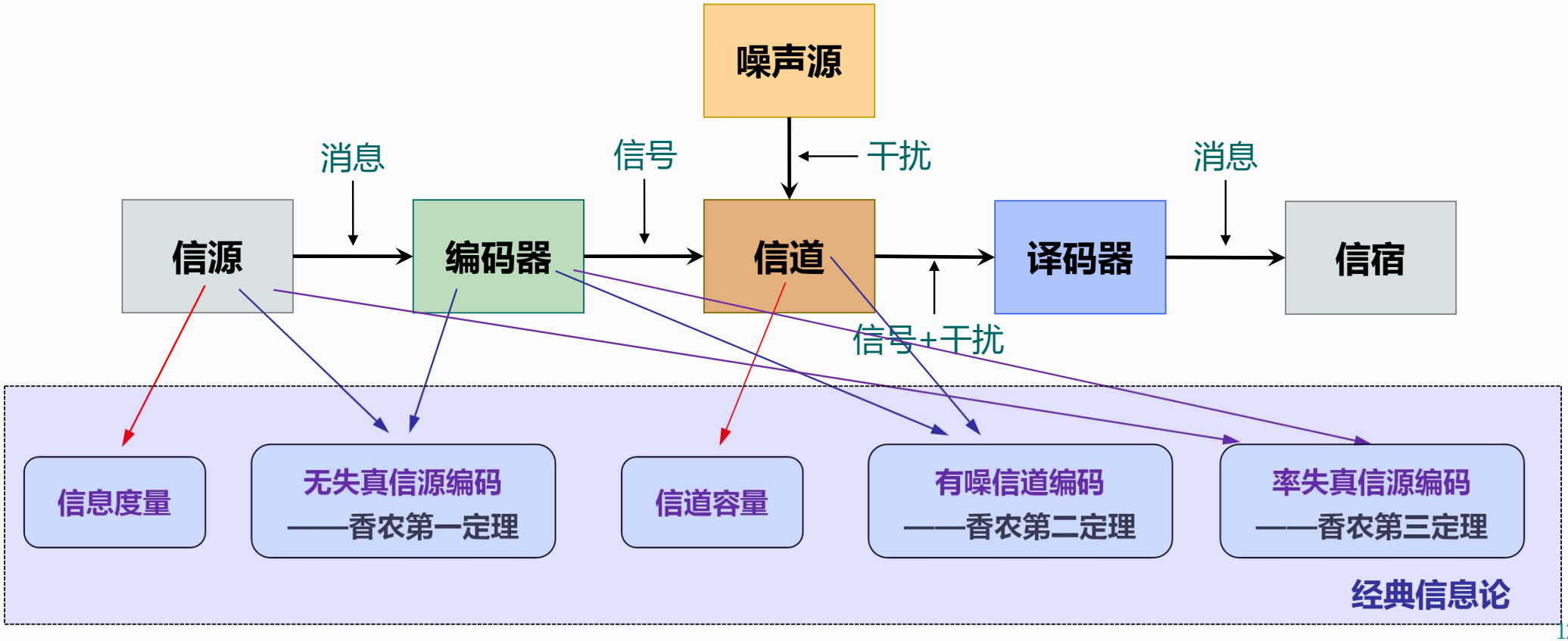
- **信息论的研究对象**
  - **信息流通系统：信息的感知、获取、存储、传输、处理、应用系统**
  - **最典型的例子：信息传输系统（通信系统），“将信息从一点准确或近似传递到另外一点”**



# 2. 什么是“信息论”？

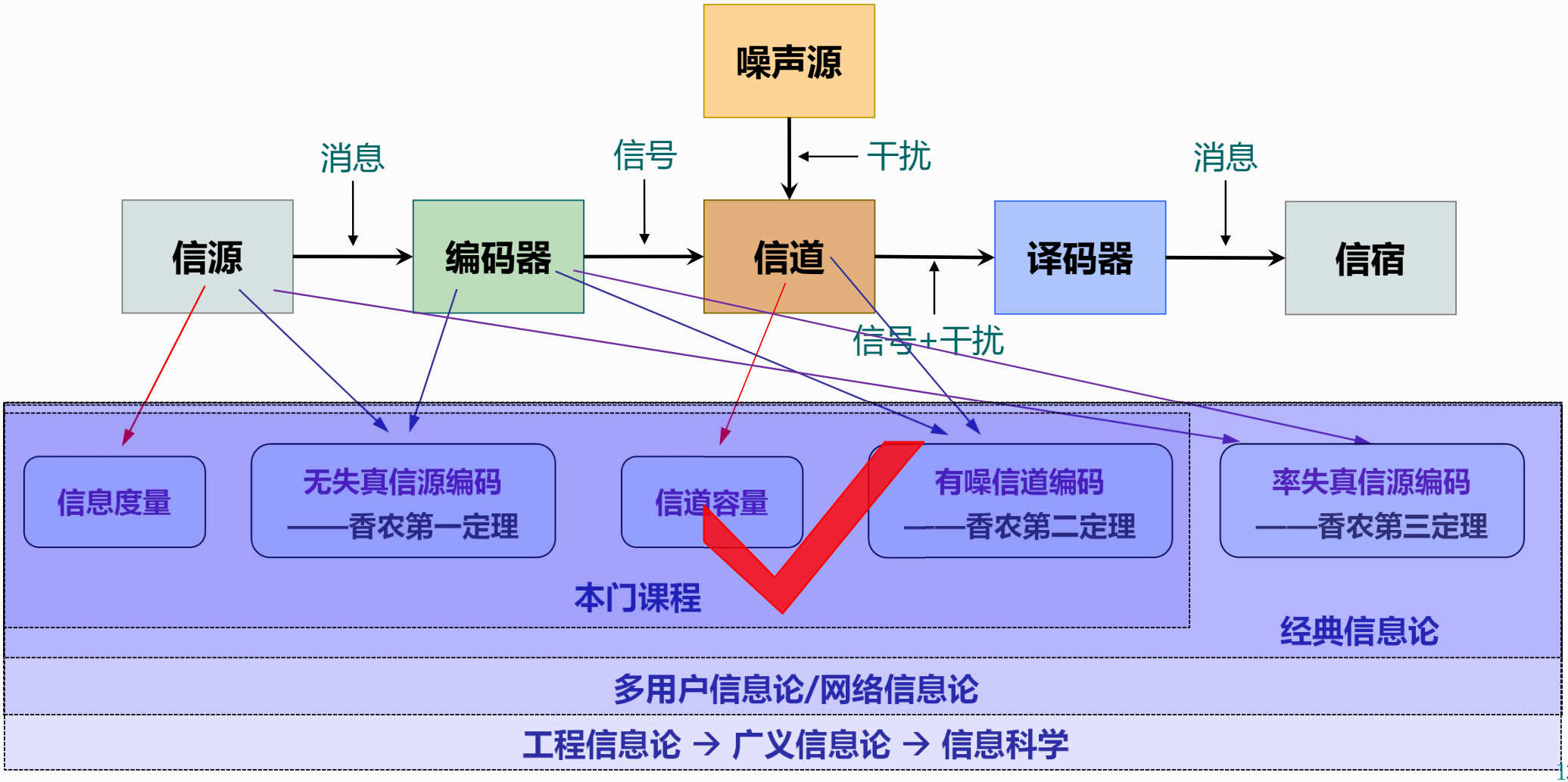


- 信息论的研究内容
  - 信息流通系统（e.g. 通信系统）中的一些最根本性问题的分析与解答



# 3. 我们可以从这门课中学到什么?





# 4. 为什么叫“导论”课?



- **大二下学期（承上启下时间点），需要有这么一门课**
- **通过此门课：**为后续信息类学科（通信、计算机……）知识的学习奠定基础。
- **16学时课程、不以“全面、系统、深入”为目标，期望结课时：**
  - 能有如下认识：1）信息论这门课就是概率（随机变量/过程、期望……）、统计（弱大数定律……）的延伸；2）信息论的典型研究范式是“极限 + 逼近极限的方法”，如“H(X), L→H(X)”、“C，R→C”……
  - 能产生对信息基础理论的兴趣，理解一些基本概念、基本结论，为今后科研中可能碰到的一些基础问题留下线索 ……
  - 能感受到经典信息理论的简洁、优美，并以此为目标设立自己今后的学术、科研“质量标准”

# 一些课程信息

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

**参考教材：**

**Thomas Cover，Elements of Information Theory（2nd edition），Wiley，2006**

**参考书：**

① Stefan M Moser，Information Theory Lecture Notes，ETH Zurich，Switzerland，2018

② Yury Polyanskiy and Yihong Wu，Information Theory From Coding to Leaning，Cambridge University Press，2022

③ 姜丹等，信息论与编码基础，电子工业出版社，2013年

④ 顾学迈等，信息与编码理论，哈尔滨工业大学出版社，2014年

**考核：**

A — 平时作业（课后 + 随堂）及表现30分

B — 期末考查70分（**期末课程报告**）

总成绩：A + B

**助教、课程QQ群：**

金鑫，2024级硕士生；QQ群（10~13班）---------->

# 关于期末课程报告

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

## 可在如下类型课程报告中任选一类：

- **关于经典信息论的局限性的思考 —— 副标题**
- **信息论在信息学科前沿领域的研究与应用 —— 副标题**
- **信息论在我####竞赛/课题中的角色 —— 副标题**
- **关于信息论课程中“####”问题的思考** （注：可能会在课堂上留下一些开放性思考题）

## 报告形式与要求：

① **模板**统一（由助教提供）

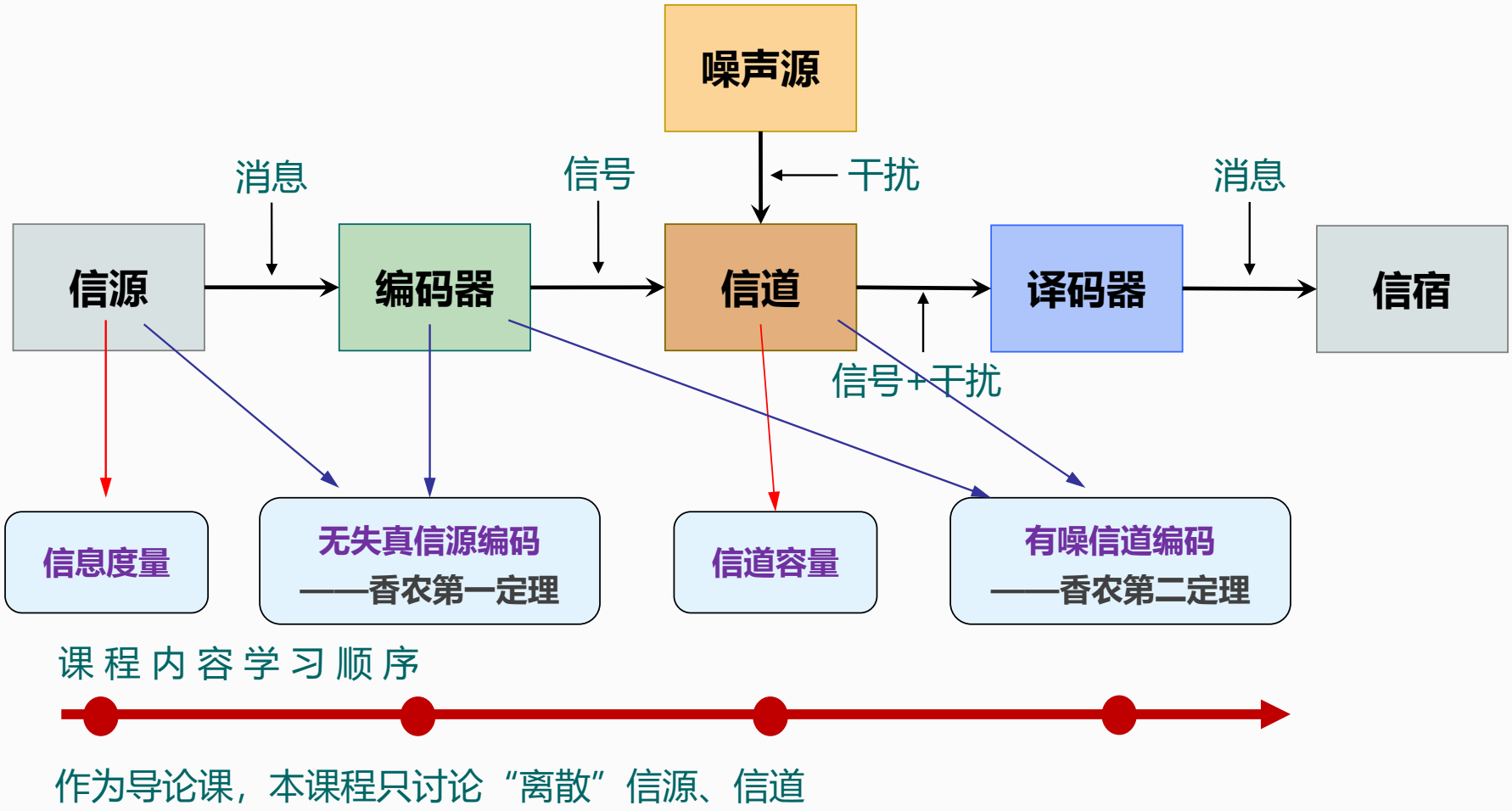
② **篇幅**不限，要有**摘要、结论、参考文献**

## 报告提交时间、提交方式：

- **提交时间**： **2025.6.19，24:00之前**，逾期不收（以助教邮箱接收时间戳为准）
- **提交方式**： 电子版报告附上个人**电子签名**，邮箱发送至助教邮箱，以**收到助教邮件回复**为“报告成功提交”确认依据

# 开始正式进入对课程知识的学习

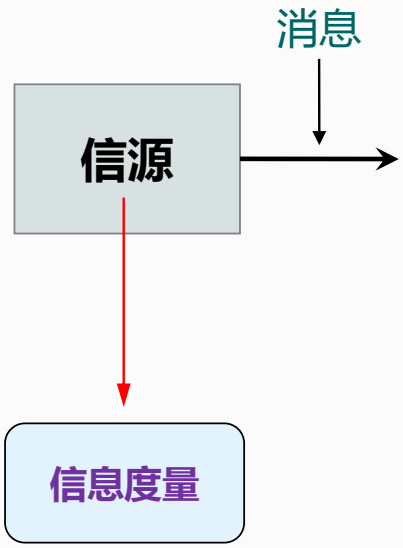





作为导论课，本课程只讨论“离散”信源、信道

# 离散信源：离散随机变量






- 设 $X$ 是一个离散随机变量，其字母表（即取值集合）用 $\mathscr{X}$ 表示，$X$ 的概率质量函数记为 $p(x) = \Pr\{X = x\}, x \in \mathscr{X}$。
- 期望：设 $g(x)$ 为关于 $X$ 的函数，其期望值可以用如下数学表达式计算：

$$E_X g(X) = \sum_{x \in \mathscr{X}} p(x)g(x)$$

概率质量向量

**例**

$\mathscr{X} = [1; 2; 3; 4; 5; 6], \mathbf{p} = [\frac{1}{6}; \frac{1}{6}; \frac{1}{6}; \frac{1}{6}; \frac{1}{6}; \frac{1}{6}]$

$EX = 3.5 = \mu$

$EX^2 = 15.17 = \mu^2 + \sigma^2$

$E\sin(0.1X) = 0.338$

$E - \log_2 p(x) = 2.58$

*X* 的“熵”

# 香农信息量



- **定义** 若某一个取值 $x$（对应信源 $X$ 产生的某一个消息）出现的概率为 $p(x)$，那么这个取值的香农信息量计算为 $-\log_2(p(x))$。

思考：我们为什么使用 $\log_2(\cdot)$ 作为信息的度量函数?

**例**

例 1: 投掷硬币
随机变量$\mathscr{X}$ = [Heads; Tails]，对应的概率分布 $\mathbf{p} = [\frac{1}{2}; \frac{1}{2}]$，对应的香农信息量为[1; 1]bits
例 2: 今天是我的生日吗?
随机变量$\mathscr{X}$ = [No; Yes]，对应的概率分布$\mathbf{p} = [\frac{364}{365}; \frac{1}{365}]$，对应的香农信息量为[0.004; 8.512]bits

**越不可能出现的结果，信息量越大**

# 熵（Entropy）



**定义** 随机变量 $X$ 的熵用 $H(X)$ 表示，定义为：

$$H(X) = E(-\log_2 p(x)) = -\sum_{x\in\mathscr{X}} p(x)\log_2 p(x)$$

对 $H(X)$ 的几点基本理解、拓展：

- 它表示信源 $X$ 的平均香农信息量
- 对于随机变量 $X$，从未知其值到获知其值，这个过程所获得的平均信息量
- 对于随机变量 $X$，若允许我们用一系列“二元问题”去确定它的值，则所需要的平均问题数量在区间 $[H(X),\ H(X)+1)$ 内

**例**

随机变量 $\mathscr{X} = [\text{circle}; \text{square}; \text{rhombus}; \text{triangle}]$，对应的概率分布 $\mathbf{p} = [\frac{1}{2}; \frac{1}{4}; \frac{1}{8}; \frac{1}{8}]$

$$H(X) = -\sum_{x\in\mathscr{X}} p(x)\log_2 p(x)$$

$$= -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8}) = 1.75$$

**思考：平均需要多少个 “Yes-No” 问题，可以猜对我的生日是几月几日?**

# 熵（Entropy）



- 熵 $H(X)$ 的大小仅依赖于概率分布 $\mathbf{p}$，而和随机变量的取值集合 $\mathscr{X}$ 无关，因此，也常用 $H(\mathbf{p})$ 来表示 $H(X)$
- 通常用 $\log(x) \equiv \log_2(x)$，即默认信息的度量单位是“比特”
  - 如果使用以 $e$ 为底的对数来描述信息熵，则信息的度量单位为奈特 (nat)
  - $1 \text{ nat} = \log_2(e) \text{ bits} = 1.44 \text{ bits}$

**引理 2.1.1**

$$H(X) \geq 0$$

**引理 2.1.2**

$$H_b(X) = (\log_b a) H_a(X)$$

# 伯努利（Bernoulli）熵

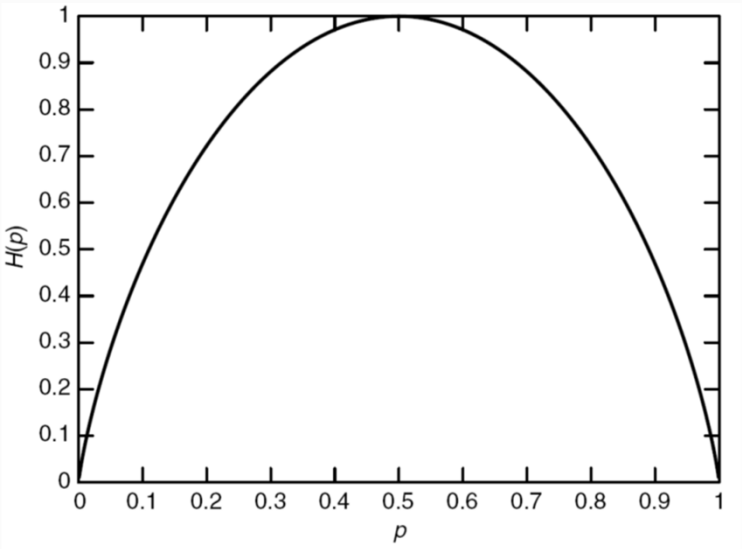


**例**

随机变量$\mathscr{X}=[1;0]$，对应的概率分布 $\mathbf{p}=[p;1-p]$ 则有：

$$H(X)=H([p;1-p])=-p\log p-(1-p)\log(1-p)$$

对于Bernoulli信源，其熵$H([p; 1-p])$通常简记为$H(p)$




- 当 $p=0$ 或 $p=1$ 时，$H(p)$ 的值为 0
  ——确定性事件的信息量为 0
- $H(p)$ 是关于 $p$ 的凹函数
- 当 $p=0.5$，即该伯努利信源服从均匀分布时，$H(p)$ 取最大值
- $H(X)$ 的取值范围为 $0\leq H(X)\leq\log|\mathscr{X}|$

# 联合熵和条件熵



**定义**

对于服从联合分布$p(x,y)$ 的一对离散随机变量$(X,Y)$，其联合熵 (Joint Entropy) $H(X,Y)$ 定义为：

$$H(X,Y) = -E\log p(x,y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

**定义**

对于服从联合分布$p(x,y)$ 的一对离散随机变量$(X,Y)$，其条件熵 (Conditional Entropy)$H(Y|X)$ 定义为:

$$H(Y|X) = -E\log p(y|x) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

不是 ***p(y\|x)***

# H(X, Y) 与 H(Y|X) —— 例题



$p(x, y)$

| Y \ X | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | $\frac{1}{8}$ | $\frac{1}{16}$ | $\frac{1}{32}$ | $\frac{1}{32}$ |
| 2 | $\frac{1}{16}$ | $\frac{1}{8}$ | $\frac{1}{32}$ | $\frac{1}{32}$ |
| 3 | $\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ |
| 4 | $\frac{1}{4}$ | 0 | 0 | 0 |

$$H(X,Y) = \frac{27}{8} \text{ bits}$$

$$H(Y|X) = \frac{13}{8} \text{ bits}$$

# 对条件熵的进一步理解



- 变形结果 1: 平均“列（行）熵”

$$
\begin{aligned}
H(Y|X) &= -E\log p(y|x) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} -p(x)p(y|x)\log p(y|x) \\
&= \sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} -p(y|x)\log p(y|x) \\
&= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x)
\end{aligned}
$$

$H(Y|X=x)$ → 列熵

以 ***p(x)*** 为权重，计算条件概率表中各列分布对应的熵（简称“列熵”）的加权平均值

用本页方法重新计算前面例题中的 ***H(Y|X)***

# 对条件熵的进一步理解

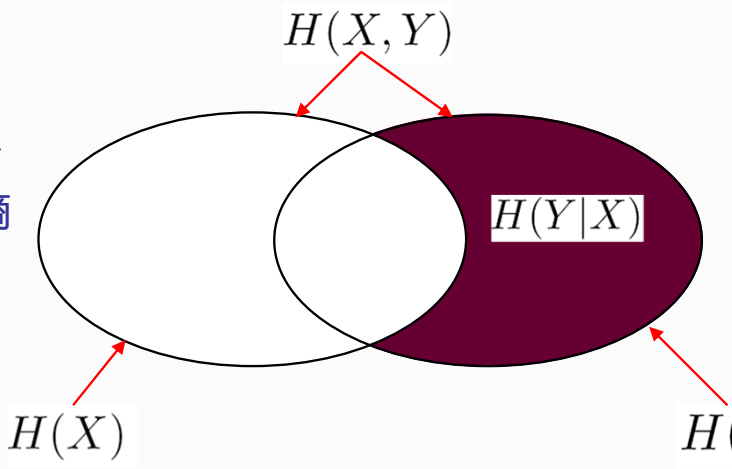


- 变形结果 2: “剩余”的熵

$$
\begin{aligned}
H(Y|X) &= E(-\log p(y|x)) = E(-\log \frac{p(x,y)}{p(x)}) \\
&= E(-\log p(x,y)) - E(-\log p(x)) \\
&= H(X,Y) - H(X)
\end{aligned}
$$

***H(Y|X)*** 是在已知**X**的条件下，**Y** “剩余” 的熵

用本页方法再重新计算前面例题中的 ***H(Y|X)***




$$H(Y|X) \neq H(X|Y)$$

$$
\begin{aligned}
&H(Y) - H(Y|X) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

# 熵的链式法则



**定理 2.2.1**

$$H(X,Y) = H(X) + H(Y|X)$$

**推论**

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

**用韦恩图验证此推论?**

**定理 2.5.1(熵的链式法则)**

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1)$$

# 阶段小结



- **熵**：$H(X) = E(-\log_2 p(x)) = -\sum_{x \in \mathscr{X}} p(x)\log_2 p(x)$
- **联合熵**：$H(X,Y) = E(-\log_2 p(x,y))$
- **条件熵**：

$$\begin{aligned} H(Y|X) &= E(-\log p(y|x)) \\ &= \sum_{x \in \mathscr{X}} p(x) H(Y|X=x) = H(X,Y) - H(X) \end{aligned}$$

- **熵的链式法则**：

$$H(X,Y) = H(X) + H(Y|X)$$

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1)$$

# 结束