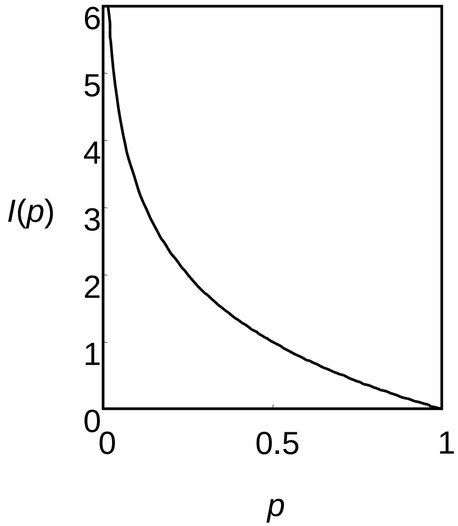
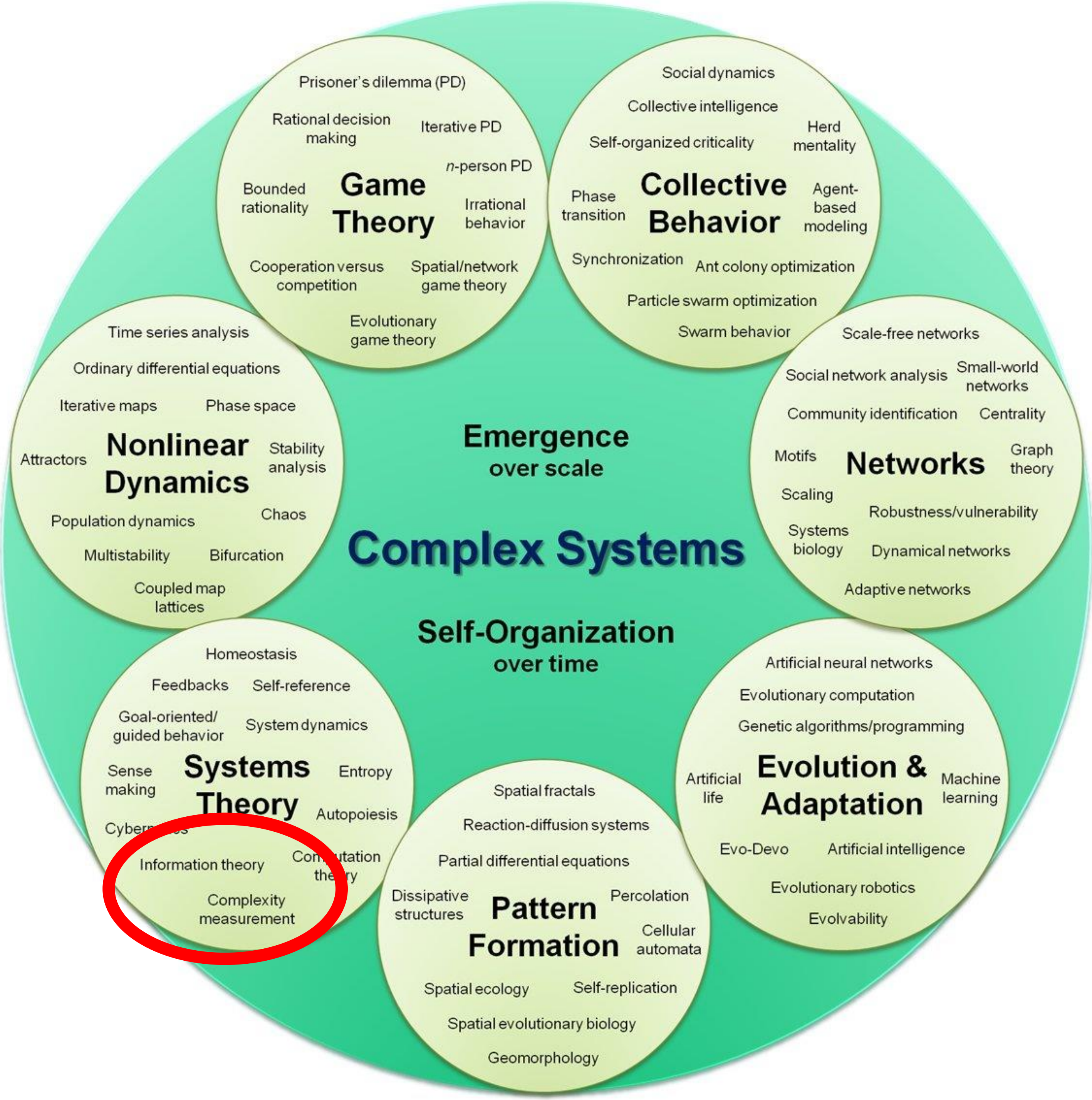
# Introduction to Information Theory





**Hiroki Sayama**

sayama@binghamton.edu


Complex Systems
Emergence over scale
Self-Organization over time
Game Theory
Prisoner's dilemma (PD)
Rational decision making
Iterative PD
n-person PD
Bounded rationality
Irrational behavior
Cooperation versus competition
Spatial/network game theory
Evolutionary game theory
Collective Behavior
Social dynamics
Collective intelligence
Self-organized criticality
Herd mentality
Phase transition
Agent-based modeling
Synchronization
Ant colony optimization
Particle swarm optimization
Swarm behavior
Networks
Scale-free networks
Social network analysis
Small-world networks
Community identification
Centrality
Motifs
Graph theory
Scaling
Robustness/vulnerability
Systems biology
Dynamical networks
Adaptive networks
Nonlinear Dynamics
Time series analysis
Ordinary differential equations
Iterative maps
Phase space
Attractors
Stability analysis
Population dynamics
Chaos
Multistability
Bifurcation
Coupled map lattices
Systems Theory
Homeostasis
Feedbacks
Self-reference
Goal-oriented/guided behavior
System dynamics
Sense making
Entropy
Autopoiesis
Information theory
Complexity measurement
Pattern Formation
Spatial fractals
Reaction-diffusion systems
Partial differential equations
Dissipative structures
Percolation
Cellular automata
Spatial ecology
Self-replication
Spatial evolutionary biology
Geomorphology
Evolution & Adaptation
Artificial neural networks
Evolutionary computation
Genetic algorithms/programming
Artificial life
Machine learning
Evo-Devo
Artificial intelligence
Evolutionary robotics
Evolvability

# Forty Shades of Complexity

- List of "complexities" maintained by MIT professor S. Lloyd
- http://web.mit.edu/esd.83/www/notebook/Complexity.PDF
  - Difficulty of description
  - Difficulty of creation
  - Degree of organization
    - Effective complexity
    - Mutual information

# What is "complexity"?

- # of variables
- Chaos, unpredictability, randomness
- Context/path dependency
- Computational time/space
- Algorithmic "depth"

**Information & Computation**

# Information

# Information?

- **Matter**

  Known since ancient times

- **Energy**

  Knows since 19th century (industrial revolution)

- **Information**

  Known since 20$^{th}$ century (WW's, rise of computers)

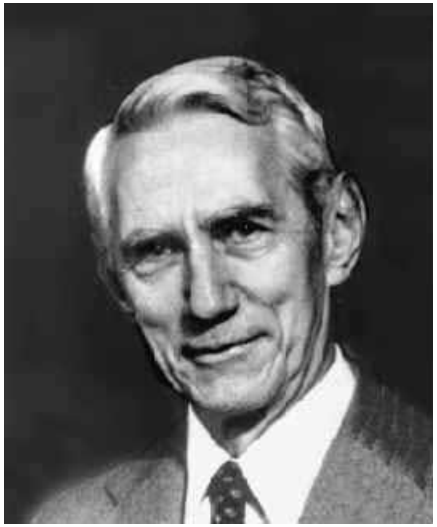
# What is information?

(Definition wanted)

# Claude E. Shannon (1916-2001)

**"A mathematical theory of communication"**

**The Bell Sys. Tech. J.**
**27: 379-423, 623-656, 1948**



- Established a formal definition of information and its quantitative measurements
- Proposed mathematical models of information sources and communication channels
- Proved fundamental theorems for both

# An informal definition of information

**Aspects of some physical phenomenon that can be used to select a smaller set of options out of the original set of options**

(Things that reduce the number of possibilities)

- An observer or interpreter involved
- A default set of options needed

# Exercise

- In weather forecast, what are the following?
  - Original set of options for tomorrow's weather
  - Aspects of physical phenomena used for forecasting

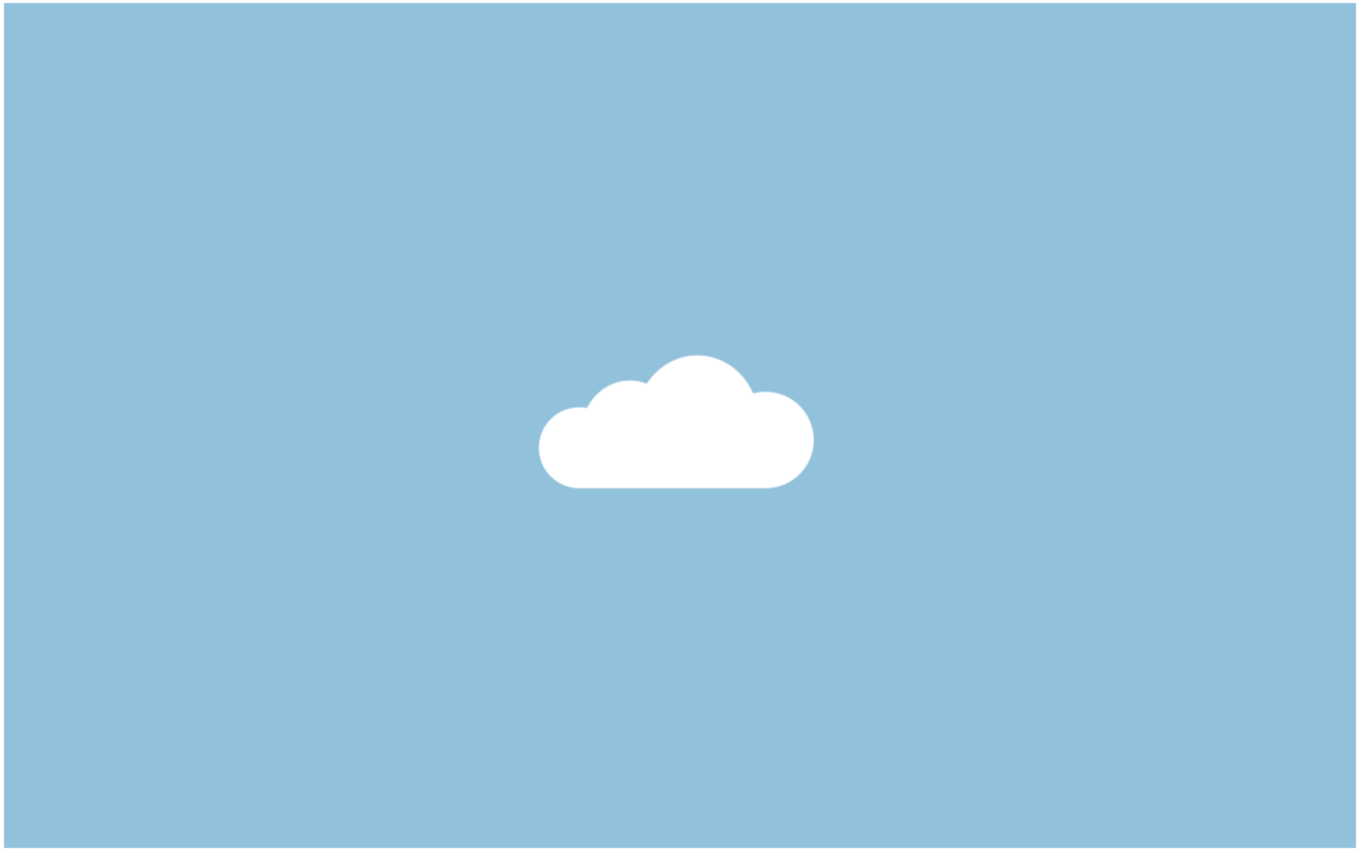
- How do they apply to today's weather forecast in Binghamton?

# Another informal statement about information in a system

- The amount of information contained in a system is the length of description needed to specify the system's state

# Exercise

- Describe the following picture in words



# Exercise

- Describe the following picture in words

# Note

- Shannon's information theory is purely based on probability theory

- Semantics (meaning) of information is left out of consideration
  - To consider semantics, one would need to take into account the mappings between symbols and other external things

# Quantitative Definition of Information

# Amount of information


- There is an apparent difference in the amount of information
- How can we quantify it?

# Quantitative definition of information: Basic idea

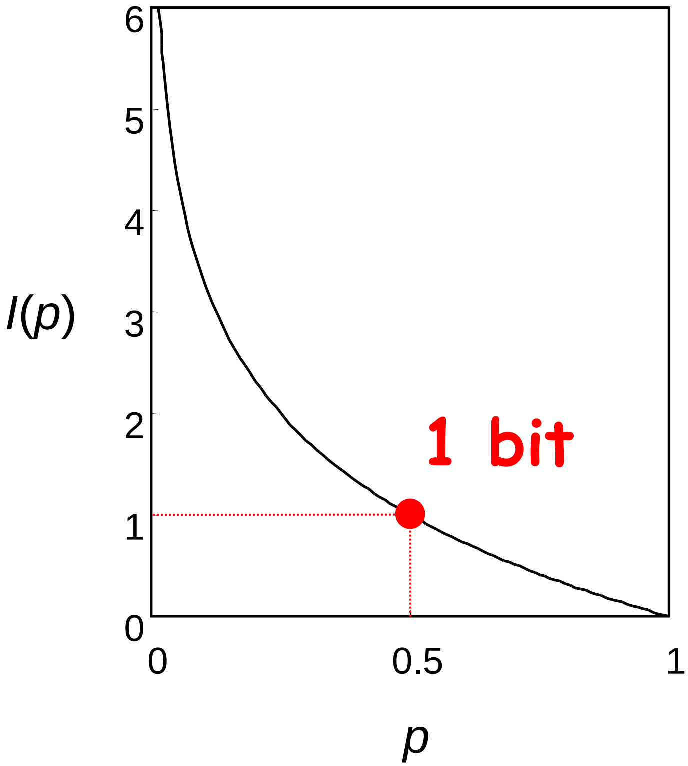
- If something is expected to occur almost certainly, its occurrence should have nearly zero information
- If something is expected to occur very rarely, its occurrence should have very large information
- If an event is expected to occur with probability $p$, the information produced by its occurrence (self-information) is given by

$$I(p) = -\log p$$

# Information measured in bits

$$I(p) = -\log p$$

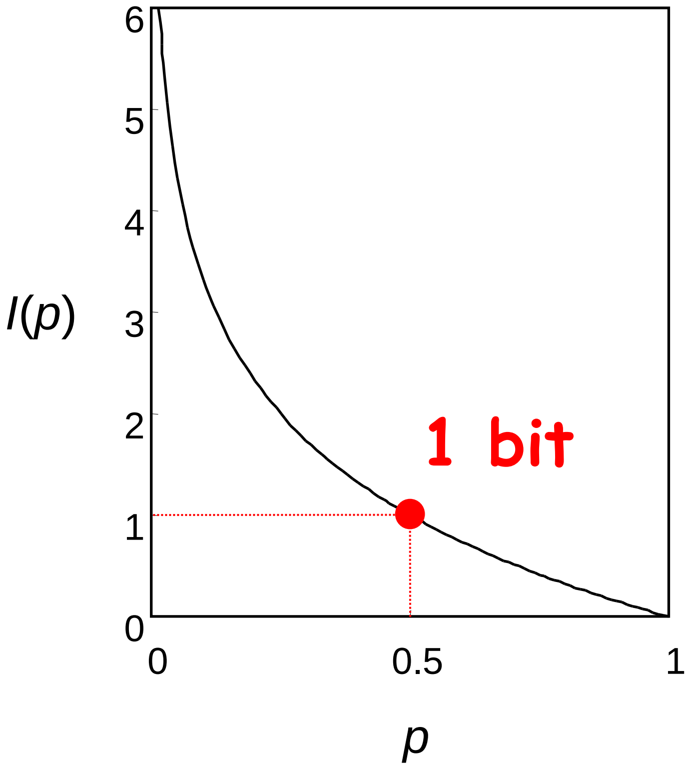
- 2 is often used as the base of log
  - Unit of information is bit (binary digit)



# Note on self-information

$$I(p) = -\log p$$

- This is no longer the length of bit strings!
- It can take non-integer values as well



# Exercise

- Calculate the amount of self-information of the following events:
  - You throw a die and the face "6" appears
  - You throw two dice and the sum of their faces is 6
  - You keep throwing a die and the face "6" appears for the first time in the tenth throw

# Why log?

- To fulfill the additivity of information
  - For independent events A and B:

Self-information of "A happened": $I(p_A)$

Self-information of "B happened": $I(p_B)$

$$\Downarrow$$

Self-information of "A and B happened":

$$I(p_A p_B) = I(p_A) + I(p_B)$$

"$I(p) = -\log p$" satisfies this additivity

# Exercise

- You pick up a card from a well-shuffled deck of cards (w/o jokers):
  - How much self-information does the event "the card is of spade" have?
  - How much self-information does the event "the card is a king" have?
  - How much self-information does the event "the card is a king of spades" have?

# Information Entropy

# Some terminologies

- **Event:** An individual outcome (or a set of outcomes) to which a probability of its occurrence can be assigned
- **Sample space:** A set of all possible individual events
- **Probability space:** A combination of sample space and probability distribution (i.e., probabilities assigned to individual events)

# Defining quantitative information on a stochastically behaving system

- Self-information $I(p) = -\log p$ is defined for each individual event observed

- Is it possible to measure the amount of information for a stochastically behaving system (probability space) even before making observations?

# Expected self-information

- Probability distribution in probability space X: $p_i$ ($i = 1 \ldots n$, $\Sigma_i\, p_i = 1$)
- Expected self-information H(X) when one of the individual events happened:

$$H(X) = \Sigma_i\, p_i\, I(p_i)$$

$$= -\,\Sigma_i\, p_i \log p_i$$

# Exercise

- Calculate H(X) for the following probability distribution:

$$\{ p_i \} = \{1/3,\ 1/3,\ 1/3\}$$

$$\{ p_i \} = \{1/2,\ 1/4,\ 1/4\}$$

$$\{ p_i \} = \{1/4,\ 1/4,\ 1/4,\ 1/4\}$$

# What does H(X) mean?

- Average amount of self-information the observer could obtain by one observation
- Average "newsworthiness" the observer should expect for one event
- Ambiguity of knowledge the observer had about the system before observation
- Amount of "ignorance" the observer had about the system before observation

# What does H(X) mean?

- It quantitatively shows the lack of information (*not* the presence of information) before observation

## Information Entropy

# Information entropy

- Similar to thermodynamic entropy both conceptually and mathematically
  - Entropy is minimal if the system state is uniquely determined with no fluctuation
  - Entropy increases as the randomness increases within the system
  - Entropy is maximal if the system is completely random (i.e., if every event is equally likely to occur)

# Relationship with Hartley's I(A)

$$I(A) = K \log_b |A|$$

- Hartley's information measure is a special case of information entropy with:
  - The assumption that each element in $A$ occurs with equal probability ($p_i = 1/|A|$)
  - $K = 1$, $b = 2$

# Exercise

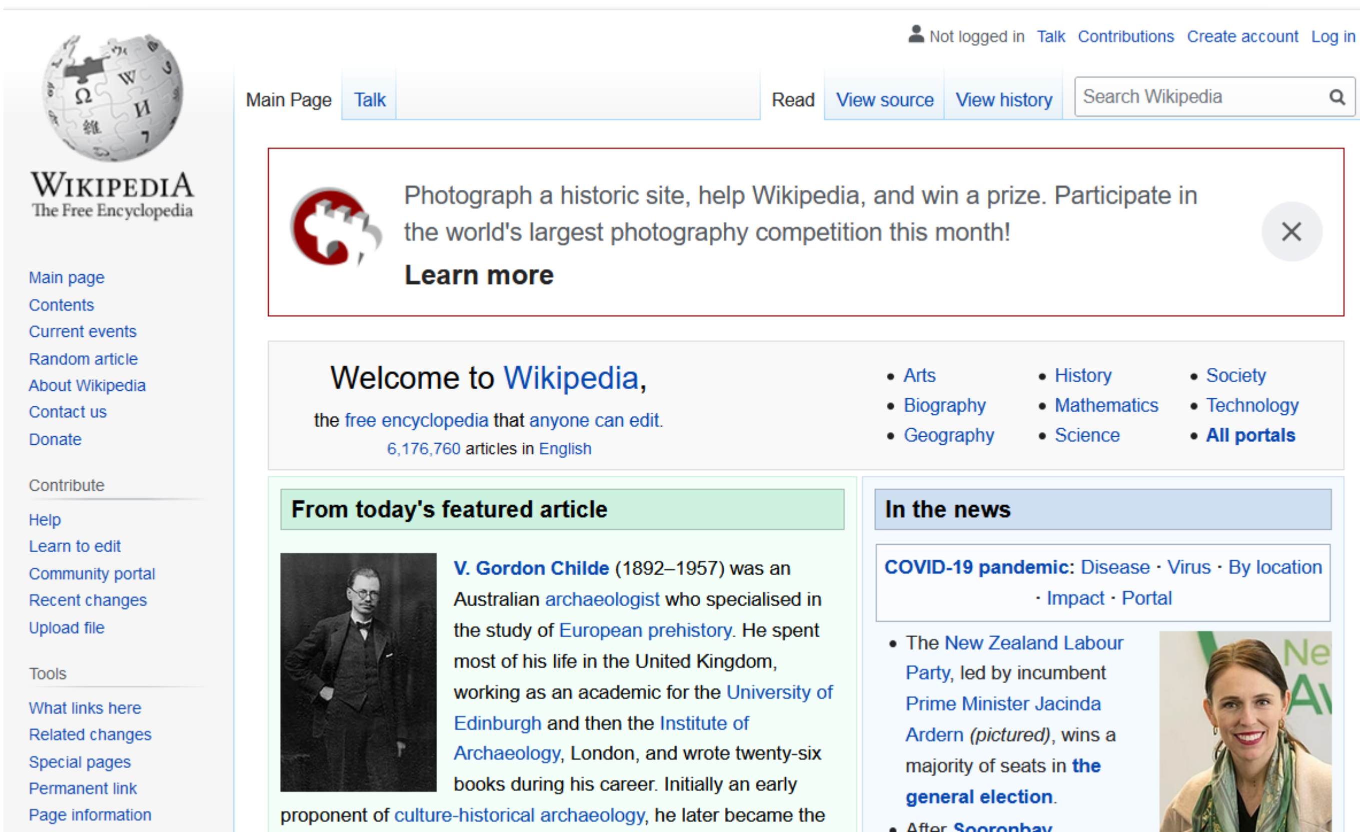
- Calculate the information entropy of a random variable that showed the following behavior:

**A, B, B, A, C, B, A, C, A, B, C, A, A, A, A, A, A, A, C, A**

(Assuming that the number of occurrences of each event accurately represents its probability)

# Exercise

- **Calculate the information entropy in the distribution of frequencies of words that appear on the top page of English Wikipedia**



# Exercise

- What is the information entropy with the following probability distribution?

$$\{ p_i \} = \{1/3,\ 1/3,\ 0,\ 1/3,\ 0\}$$

# Exercise

- Prove the following:

  Entropy is maximal if the system is completely random (i.e., if every event is equally likely to occur)

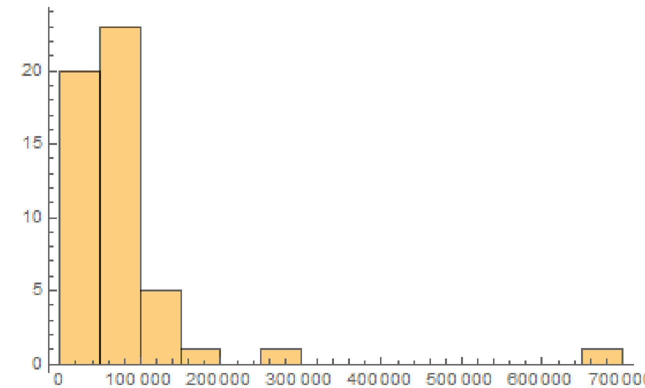
- Show that $f(p_1, p_2, \ldots, p_n) = -\sum_{i=1 \sim n} p_i \log p_i$ (with $\sum_{i=1 \sim n} p_i = 1$) takes its maximum with $p_i = 1/n$
  - Remove one variable using the constraint
  - Or use the method of Lagrange multiplier

# Information Entropy of Continuous Variables

# How to calculate information entropy for continuous variables?

- Simple idea: Binning
  - Example: Areas of 50 US states

{52 420.1 $mi^2$, 665 384. $mi^2$, 113 990. $mi^2$, 53 178.5 $mi^2$, 163 695. $mi^2$,
104 094. $mi^2$, 5543.42 $mi^2$, 2488.73 $mi^2$, 68.34 $mi^2$, 65 757.7 $mi^2$, 59 425.2 $mi^2$,
10 931.7 $mi^2$, 83 568.9 $mi^2$, 57 913.5 $mi^2$, 36 419.5 $mi^2$, 56 272.8 $mi^2$,
82 278.4 $mi^2$, 40 407.8 $mi^2$, 52 378.1 $mi^2$, 35 379.7 $mi^2$, 12 405.9 $mi^2$,
10 554.4 $mi^2$, 96 713.5 $mi^2$, 86 935.8 $mi^2$, 48 431.8 $mi^2$, 69 707. $mi^2$, 147 040. $mi^2$,
77 347.8 $mi^2$, 110 572. $mi^2$, 9349.16 $mi^2$, 8722.58 $mi^2$, 121 590. $mi^2$, 54 555. $mi^2$,
53 819.2 $mi^2$, 70 698.3 $mi^2$, 44 825.6 $mi^2$, 69 898.9 $mi^2$, 98 378.5 $mi^2$, 46 054.3 $mi^2$,
1212. $mi^2$, 32 020.5 $mi^2$, 77 115.7 $mi^2$, 42 144.2 $mi^2$, 268 596. $mi^2$, 84 896.9 $mi^2$,
9616.36 $mi^2$, 42 774.9 $mi^2$, 71 298. $mi^2$, 24 230. $mi^2$, 65 496.4 $mi^2$, 97 813. $mi^2$}



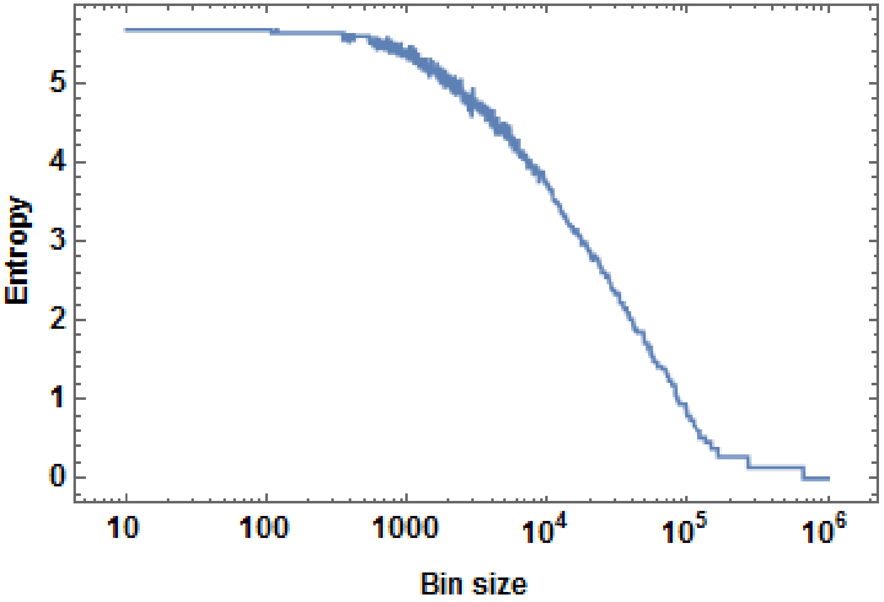

$$H(X) = 1.70987$$

# Problem in simple binning

- The result of entropy calculation will depend on bin size

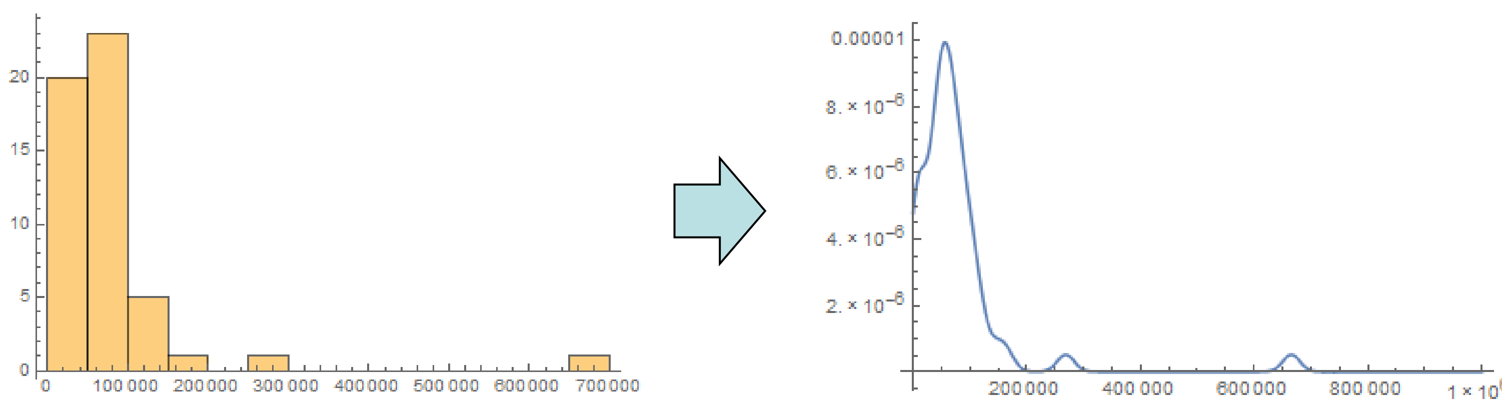
With large $\Delta x$:
  Entropy $\rightarrow 0$

With small $\Delta x$:
  Entropy $\rightarrow \log n$



# Solution: Probability density function

- Representing the sample distribution with a continuous probability density function (PDF) avoids convergence to trivial “log n” as $\Delta x \to 0$
  - E.g. Gaussian kernel method



# Problem with continuous PDF

$$H(X) = -\,\Sigma_i\, p_i \log p_i$$

$$= -\,\Sigma_x\, \mathrm{pdf}(x)\,\Delta x \log(\,\mathrm{pdf}(x)\,\Delta x\,)$$

$$\rightarrow -\int_x \mathrm{pdf}(x) \log \mathrm{pdf}(x)\, dx$$

$$- \log \Delta x$$

- Information entropy diverges to infinity as $\Delta x \rightarrow 0$ !!

# Differential entropy

$$H_{dif}(X) = - \int_x pdf(x) \log pdf(x)\, dx$$

- Just ignore the "$- \log \Delta x$" term
- No longer the same quantity as the original entropy, but still useful for comparing two systems, etc.

# Note on differential entropy

- Its value can be negative!
- Its magnitude does not tell by itself the amount of information (uncertainty) in the variable
  - Though a difference between two differential entropies does

# Exercise

- Calculate the differential entropy of the following PDFs:
  - Uniform PDF in [0,1]
  - Uniform PDF in [0, 0.5]
  - Gaussian PDF with mean 0 and s.d. 1
  - Gaussian PDF with mean 0 and s.d. 0.1