Foundations of Data Science

Course Introduction
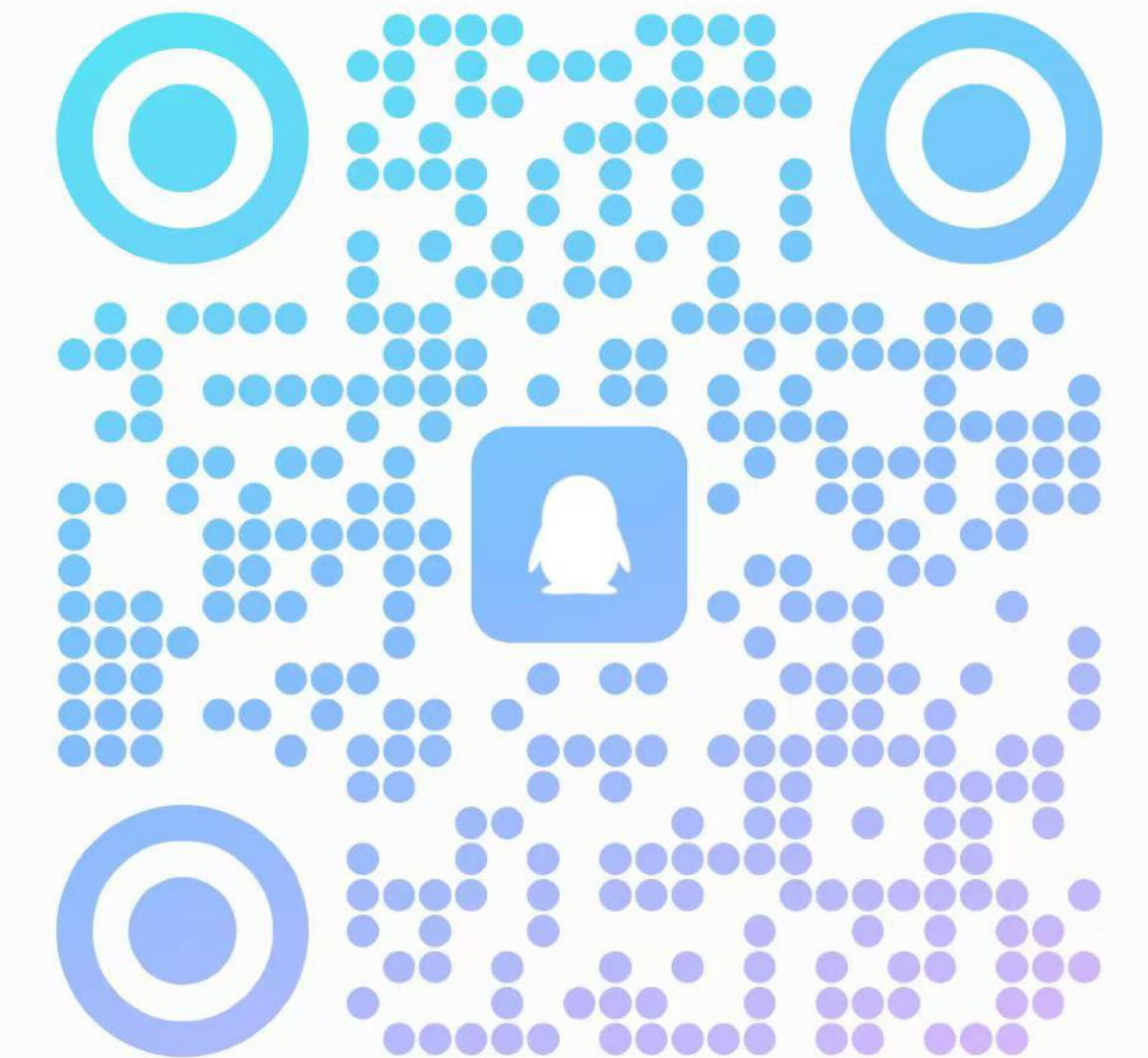
尹一通，刘明谋 **Nanjing University, 2024 Fall**

Course Info
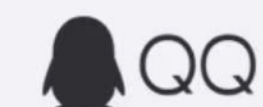
- Instructors: 尹一通，刘明谋

  - Emails: yinyt@nju.edu.cn

- Office hour:

  - Thursday 6-8pm，南雍东207

  - QQ group：302709730

- Course homepage:

  - https://tcs.nju.edu.cn/wiki
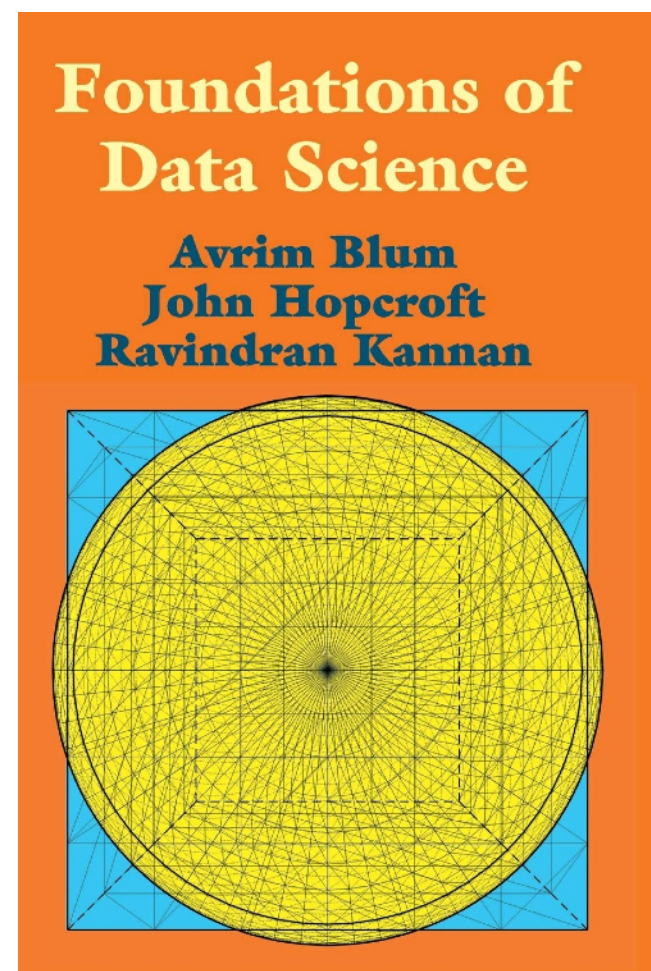
数据科学基础（南京...
群号: 302709730

扫一扫二维码，加入群聊

QQ

Foundations of Data Science

- **Data science**: an interdisciplinary field that focuses on extracting knowledge and insights from data.

- ***Foundations of Data Science***:

 (Book by Avrim Blum, John Hopcroft, and Ravi Kannan)

While traditional areas of computer science remain highly important, increasingly researchers of the future will be involved with using computers to understand and extract usable information from massive data arising in applications, not just how to make computers useful on specific well-defined problems. With this in mind we have written this book to cover the theory we expect to be useful in the next 40 years, just as an understanding of automata theory, algorithms, and related topics gave students an advantage in the last 40 years. One of the major changes is an increase in emphasis on probability, statistics, and numerical methods.
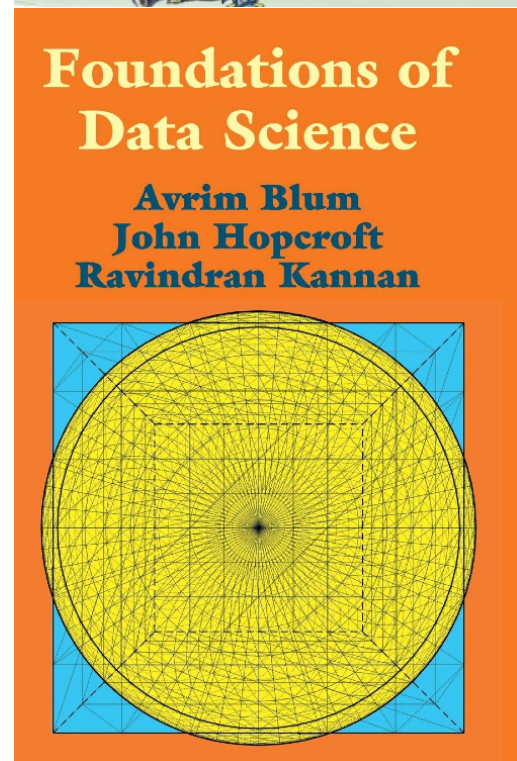
Textbooks
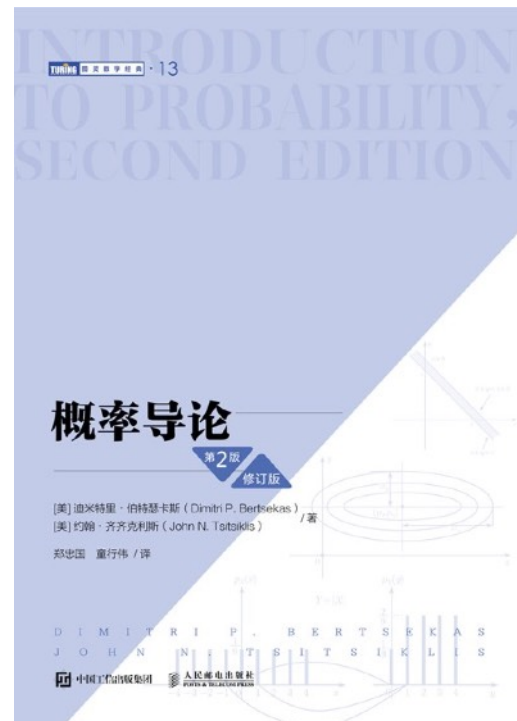
- **概率导论（第2版·修订版）(*Introduction to Probability*)**，
 伯特瑟卡斯 (Dimitri Bertsekas) 齐齐克利斯 (John Tsitsiklis) 著，
  郑忠国，童行伟译，人民邮电出版社（2022）
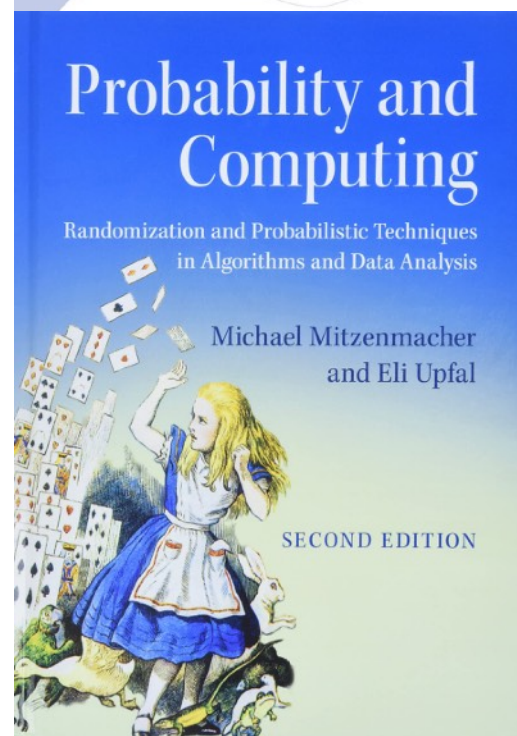
- ***Probability and Computing:* Randomization and Probabilistic Techniques in Algorithms and Data Analysis,** 2nd edition.
  （概率与计算：算法与数据分析中的随机化和概率技术）
  Michael Mitzenmacher, Eli Upfal.
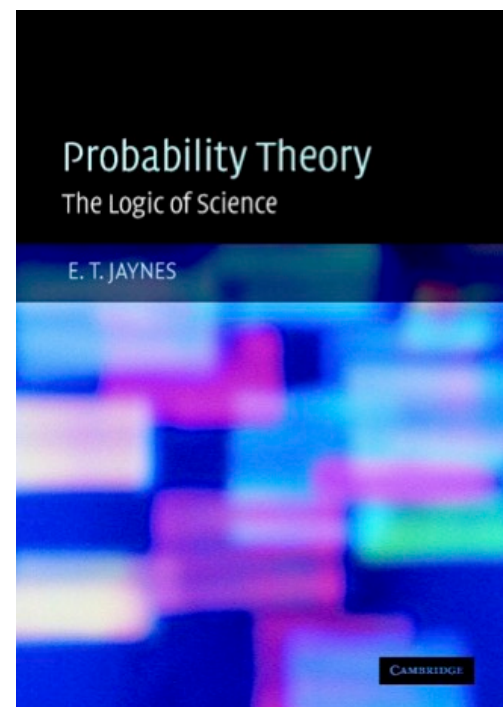 Cambridge University Press (2017).

- ***Foundations of Data Science* (数据科学基础).**
  Avrim Blum, John Hopcroft, and Ravindran Kannan.
  Cambridge University Press (2020).

Further References and Readings
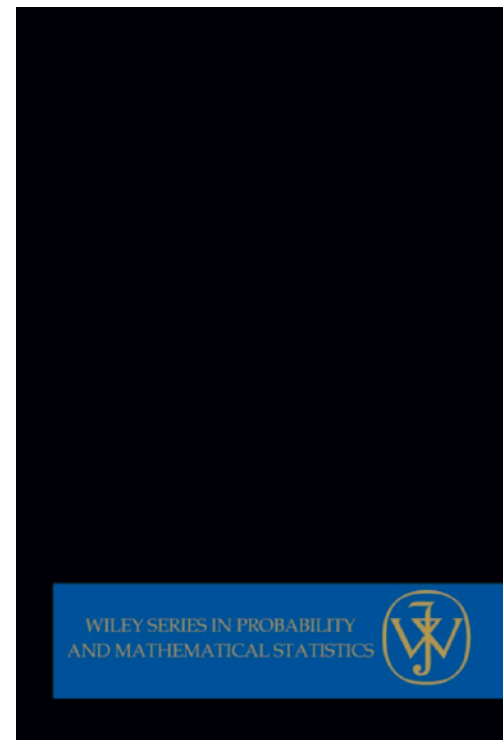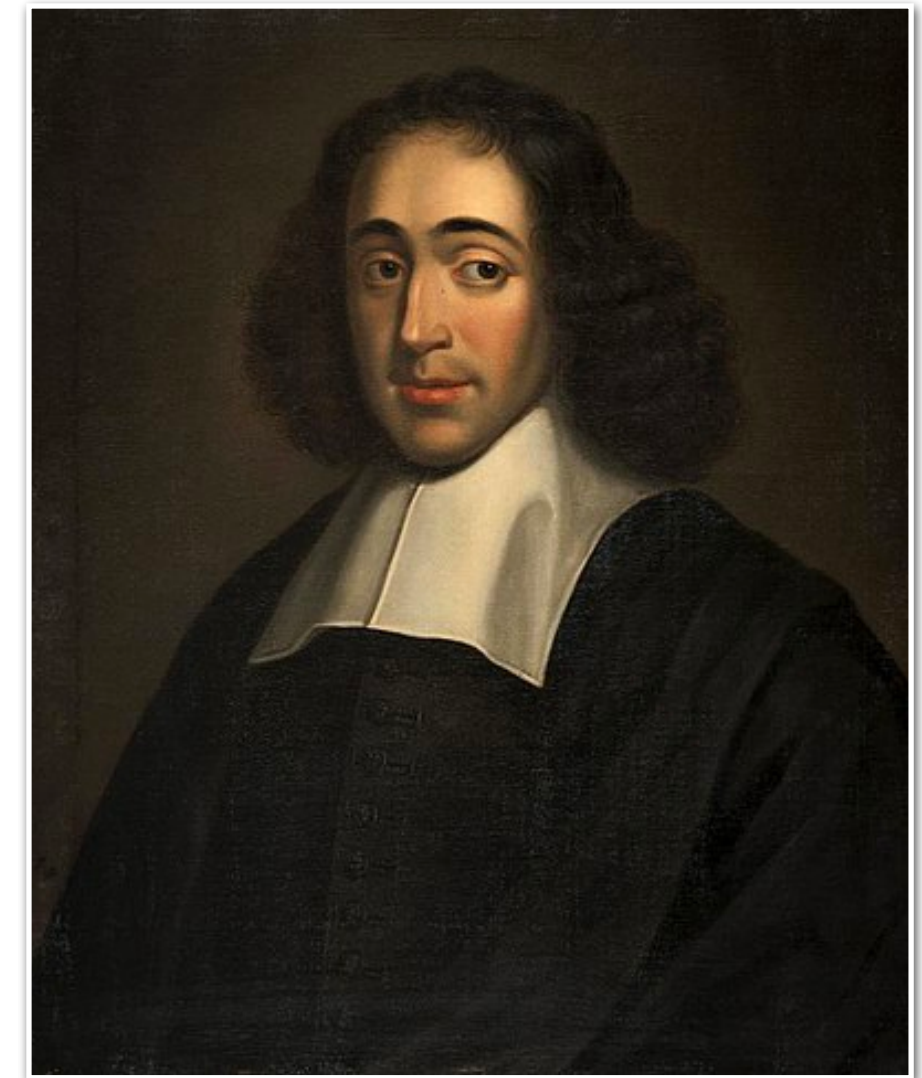
- ***An Introduction to Probability Theory and Its Applications**, Vol. 1, 3rd Ed*. William Feller. Wiley (1968). 《概率论及其应用》

- ***Probability Theory: The Logic of Science***. E. T. Jaynes. Cambridge University Press (2003). 《概率论沉思录》

"*In practical life we are compelled to follow what is most* *probable*; *in speculative thought we are compelled to follow* *truth*."

Baruch de Spinoza（巴鲁赫·斯宾诺莎　十七世纪哲学家）

Probability : *truth about probables*

- *"La théorie des probabilités n'est, au fond, que le bon sens réduit au calcul (Probability theory is nothing but common sense reduced to calculation.)"*

　　　　　　　　　　　　　　　　　　　——皮埃尔·西蒙·拉普拉斯（1819）

- *"The true logic of this world is the calculus of probabilities."*

　　　　　　　　　　　　　　　　　　——詹姆斯·克拉克·麦克斯韦（1850）

- "概率论是对计算机专业最'有用'的一门数学课"—— 许多计算机专业的师生

Interpretation of Probability
(This course is not about this.)

$$\text{“某件事 } A \text{ 发生的概率是 } p\text{”}$$

- **Frequentism**: the probability of an event is its *relative frequency* over time (by repeating an experiment a large number of times under the same condition), e.g. *"硬币正面朝上的概率是1/2"*

- **Subjectivism** (**Bayesian probability**): the probability is a measure of the *"degree of belief"* of the individual assessing the uncertainty of a particular situation, e.g. *"'黑猴'年度游戏稳了"、"《红楼梦》的作者八成是曹雪芹"*

- 以及若干其它的诠释。概率的诠释是一个**哲学问题**:
  - https://plato.stanford.edu/entries/probability-interpret/

History of Probability
(*This course is not about this either.*)

- **16世纪**：文艺复兴时期的博学家 (polymath) Gerolamo Cardano (卡尔达诺) 在一本小册子中研究了骰子等概率问题

- **17世纪**：Blaise Pascal (帕斯卡) 与 Pierre de Fermat (费马) 之间的通信 the Pascal-Fermat correspondence 为概率论的许多概念奠定了基础；Christiaan Huygens (惠更斯) 发表了最早的概率论教科书 *De Ratiociniis in Ludo Aleae (On reasoning in games of chance)*

- **18世纪**：Jacob Bernoulli (伯努利) 发表了 *Ars Conjectandi (The Art of Conjecturing)*，给出了最早的大数定律；Abraham de Moivre (棣莫弗) 发表了 *The Doctrine of Chances*，引入了正态分布并证明了首个中心极限定理

- **19世纪**：高斯将概率应用于天文预测；Pierre-Simon de Laplace (拉普拉斯) 发表了 *Théorie analytique des probabilités (Analytic Theory of Probability)* 巩固并进一步奠定概率论与数理统计的基础；19世纪末 Ludwig Boltzmann (玻尔兹曼) 和 Josiah Willard Gibbs (吉布斯) 利用概率工具发展了统计力学

- **20世纪**: **Andrey Kolmogorov (柯尔莫哥洛夫) 引入概率论公理系统**

Gian-Carlo Rota (1998) in MIT's course 18.313 on *Probability*: "The idea of probability is one of the great ideas developed in this (20th) century. You can find it in all the sciences."
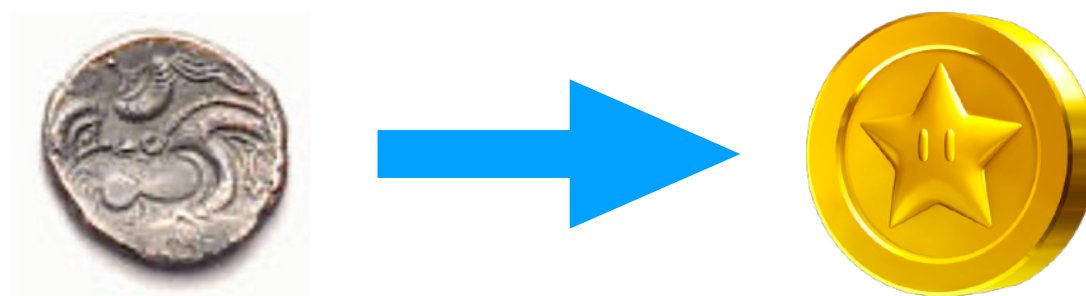
Sample Problems

Fair Coins from Biased
(von Neumann's Bernoulli factory)

- John von Neumann (1951): "Suppose you are given a coin for which the probability of HEADS, say p, is **unknown**. How can you use this coin to generate unbiased (fair) coin-flips."

- 💡: Construct two equally probable events (without knowing p).

The Two Child Problem
(boy or girl paradox)

- Martin Gardner (1959): "Knowing that I have two children and at least one of them is girl, what is the probability that both children are girls?"

- **Assumption:**

  - Each child is either girl or boy;

  - each child has the same chance of being female as of being male;

  - the gender of each child is independent of the gender of the other.

- 💡 : Understand the probability space.

Monty Hall Problem
(three doors problem)

- Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

- 💡 : Consider a 1-car-100-door variant.
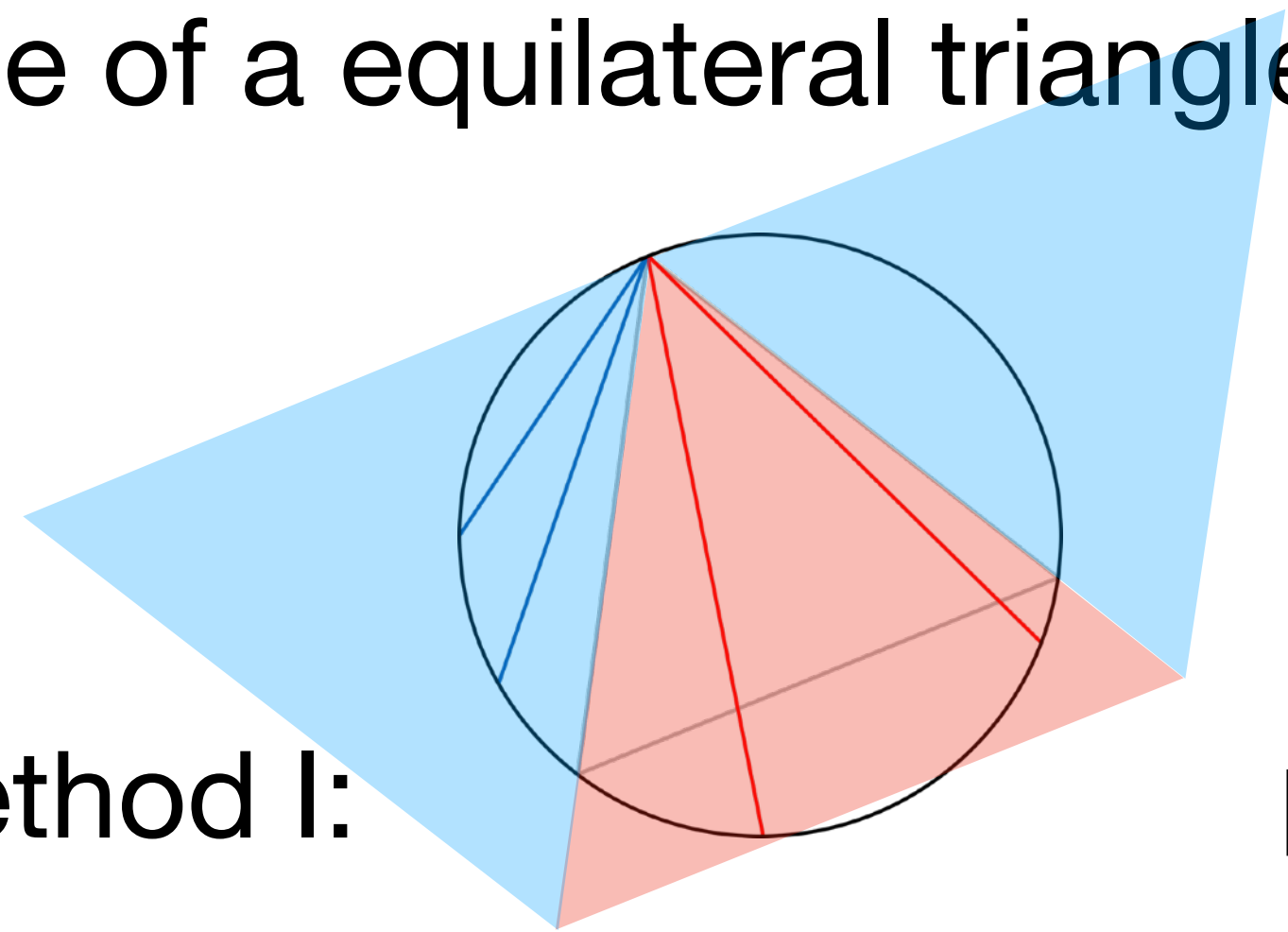

Switch and win

Switch and win

Stay and win
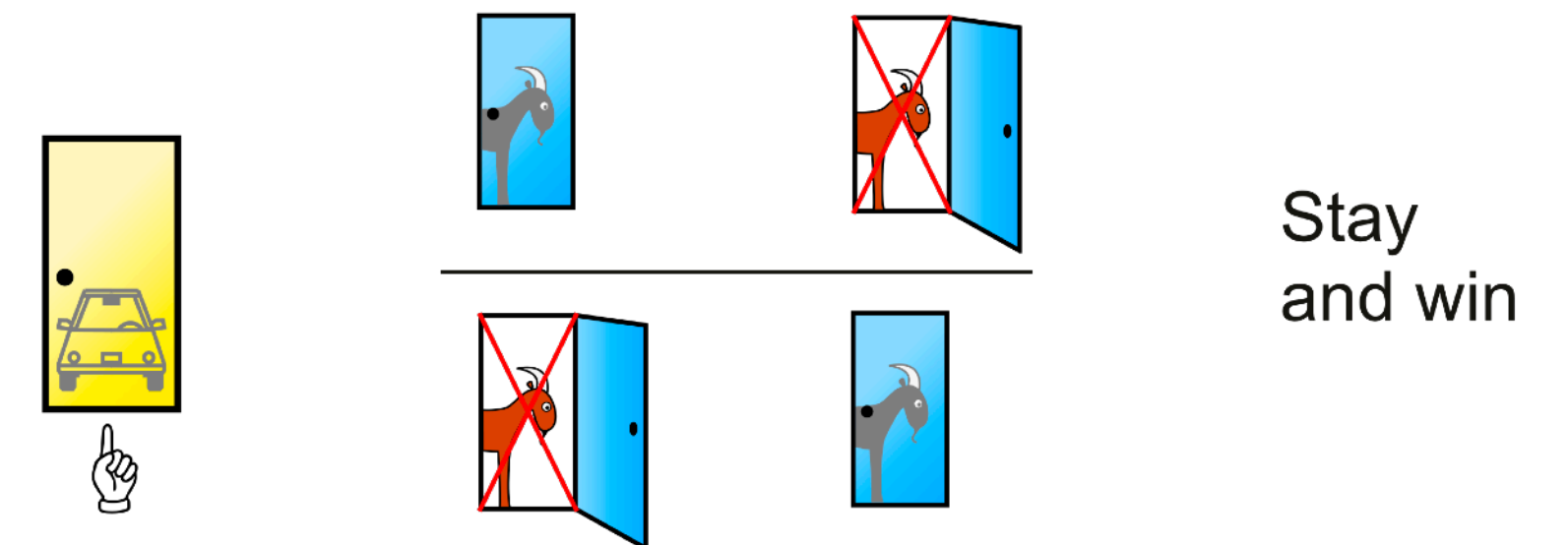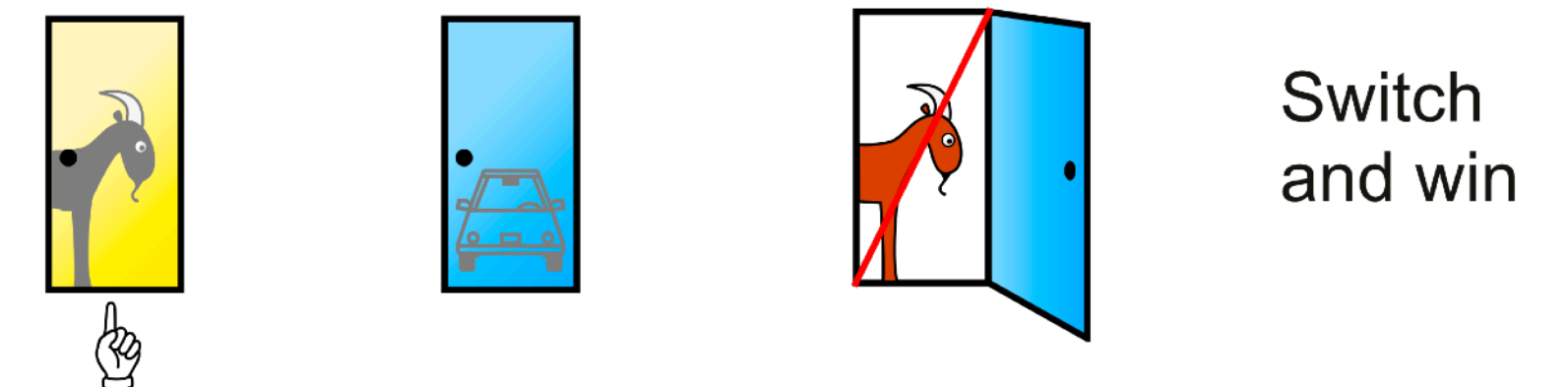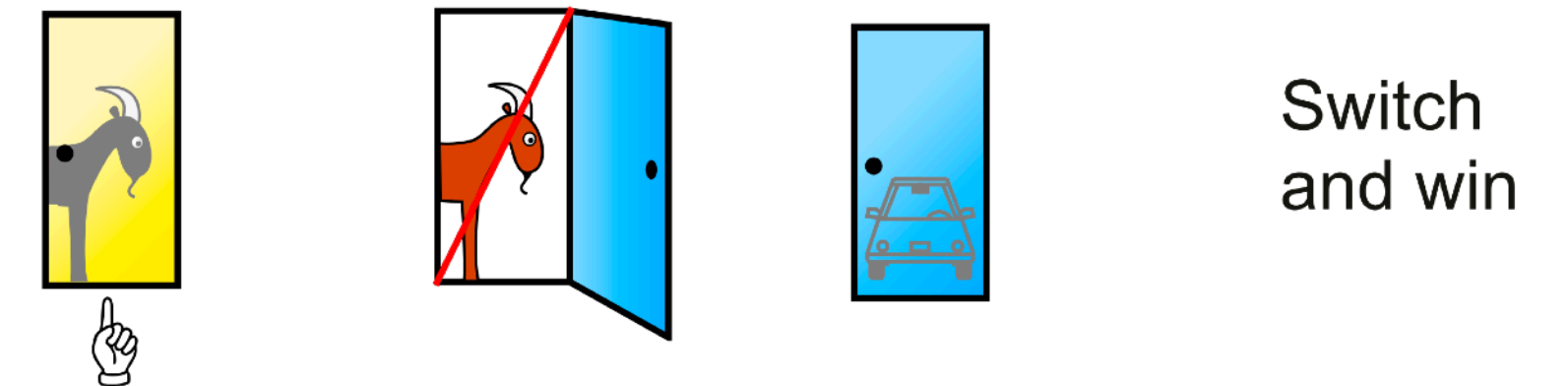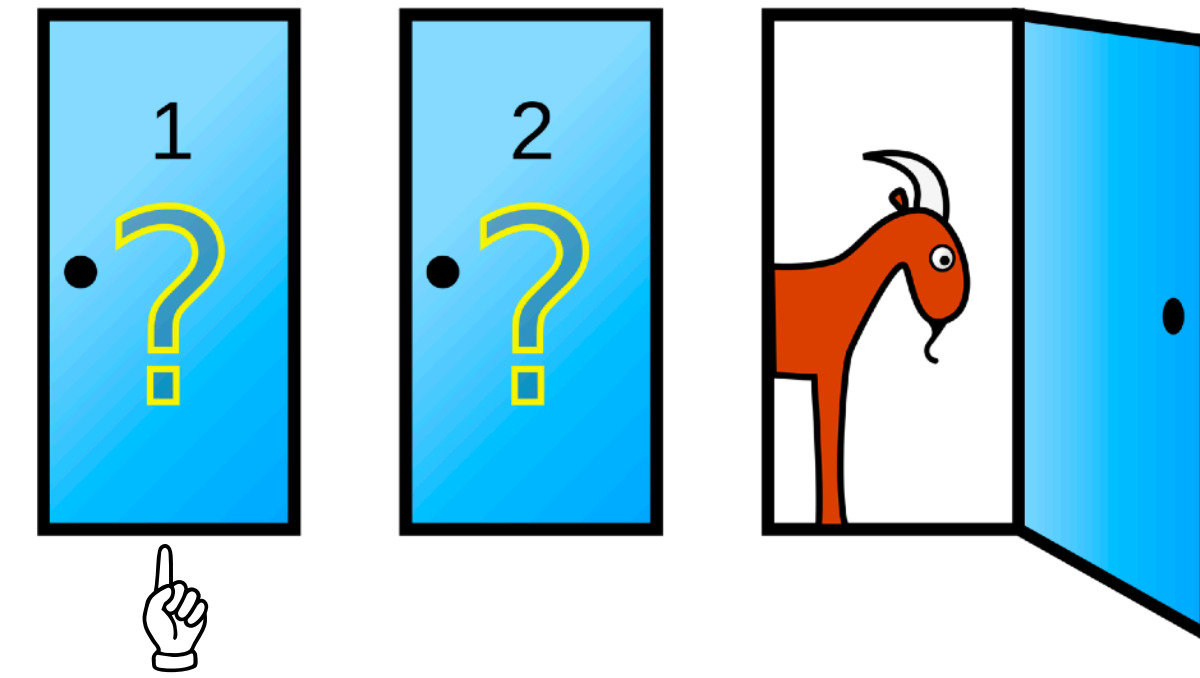
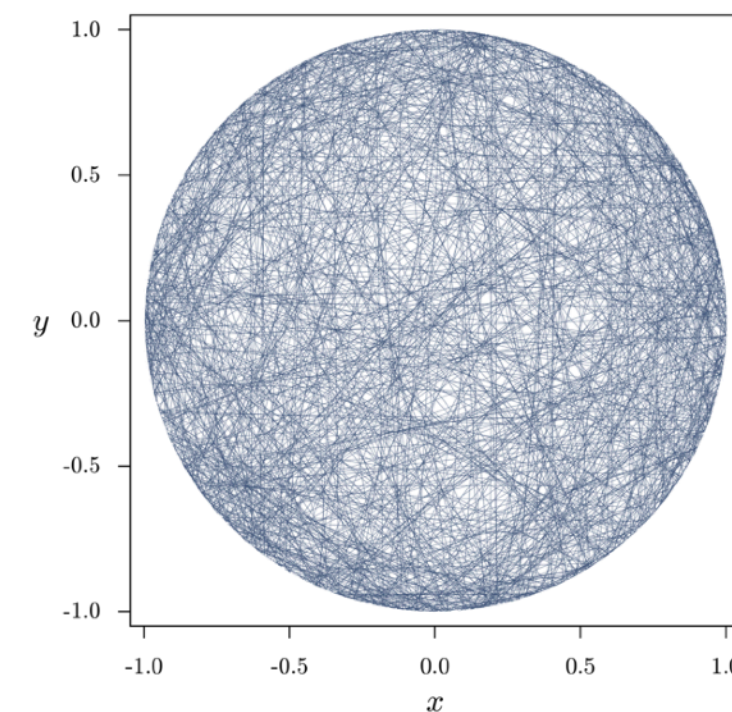Player choice before door is open

Bertrand Paradox (贝特朗悖论)
introduced in *Calcul des probabilités* (1889) by Joseph Bertrand

- What is the probability that a *random chord of a circle*, is longer than the side of a equilateral triangle inscribed in a circle?
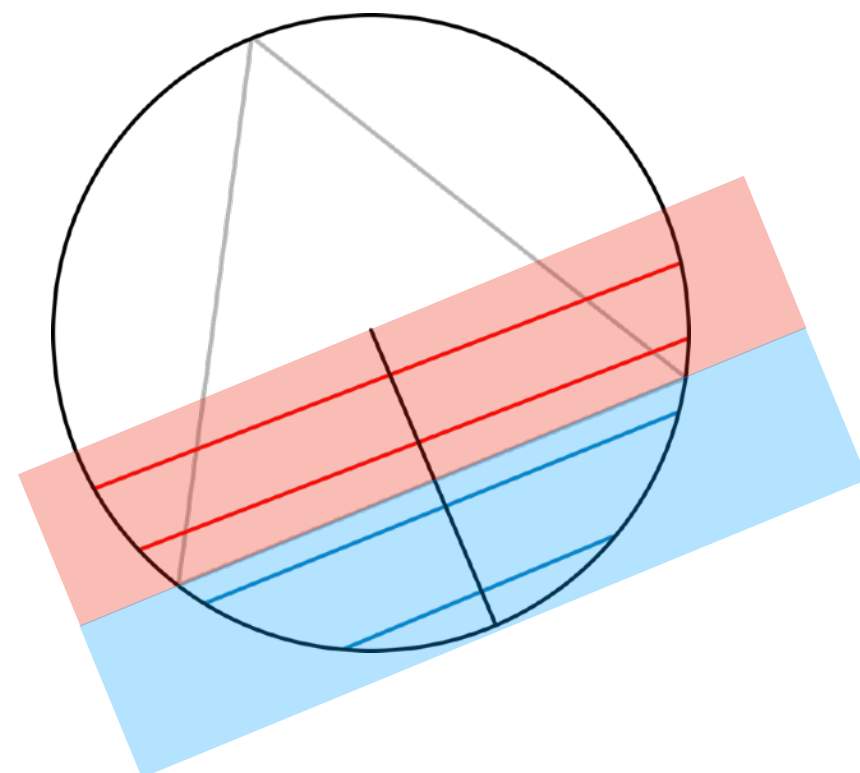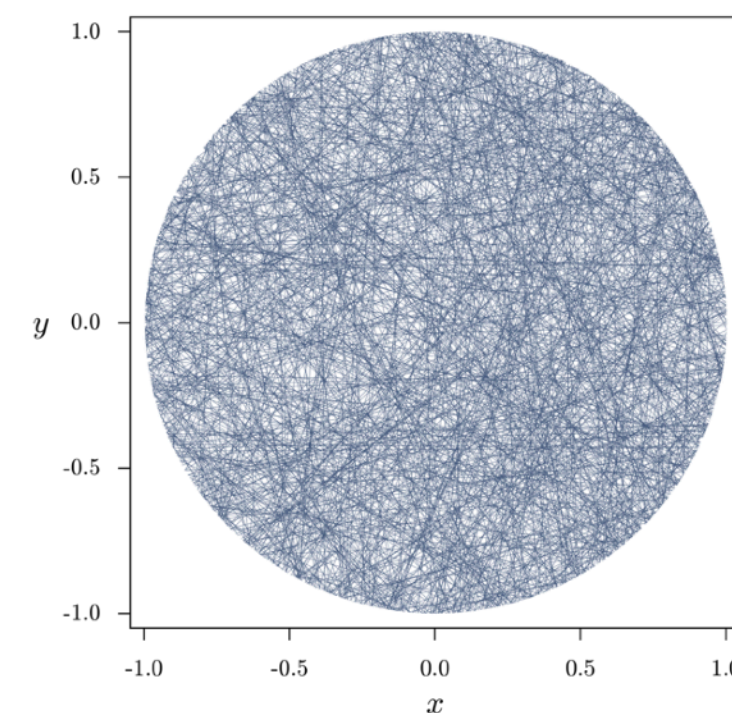

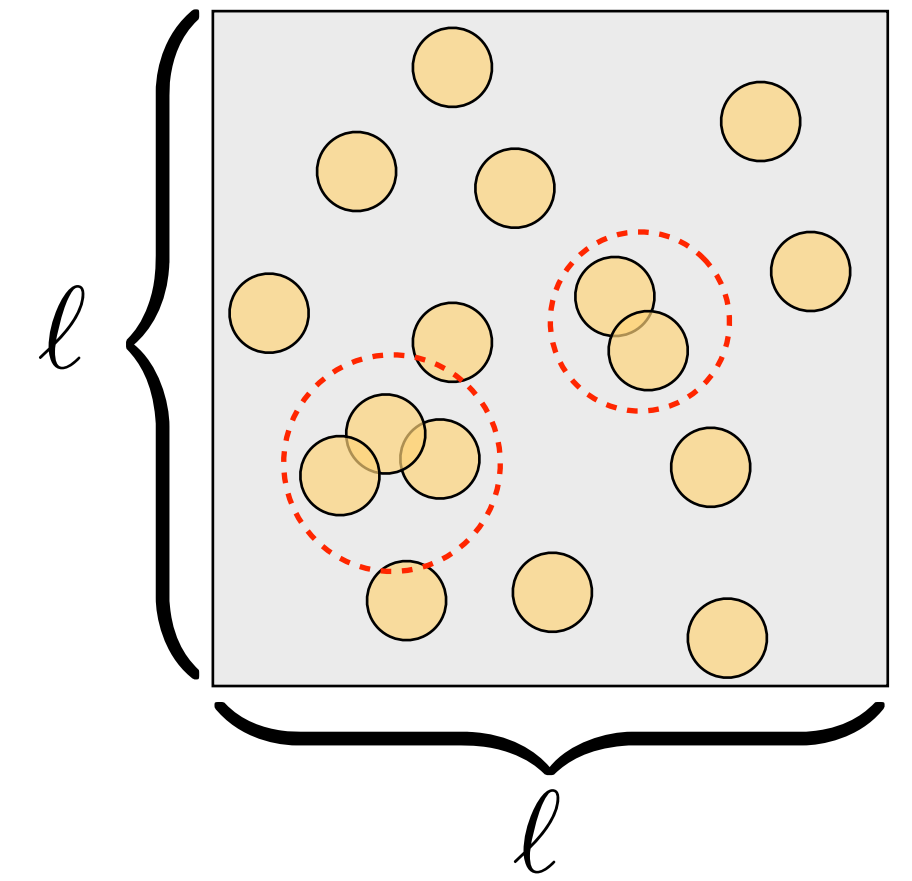- method I: Prob=$1/3$

- method II: Prob=$1/2$

- 💡: The notion of "random chord of a circle" shall be more rigorously defined.

Pennies on a Carpet
(hard spheres in 2D square)

$n \times$ $\}d$

- Drop n pennies on a square-shape carpet at random. What is the probability that no two pennies will overlap?

- In 1-dimension (n needles on a line segment):

$$\begin{cases} \left(\dfrac{\ell - nd}{\ell - d}\right)^n & \text{if } \ell \geq nd \\ \\ 0 & \text{otherwise} \end{cases}$$

- In 2-dimension: Nothing is known about this problem. as of 1979–98.

硬球(hard spheres)模型: This problem is one of the most important problems of statistical mechanics. If we could answer it we would know, for example, why water boils at 100ºC, on the basis of purely atomic computations.
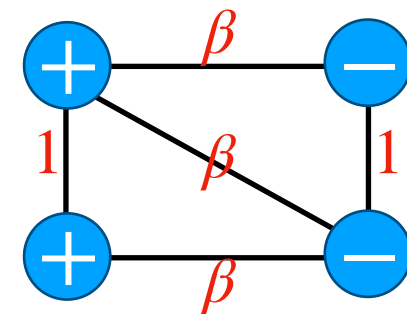

Gian-Carlo Rota teaching 18.313

M. Jerrum, H. Guo (2021): 高效数值求解hard spheres模型的蒙特卡罗算法

Phase Transition of Ferromagnetism

(critical behavior of Ising model)

- Ising model (Lenz 1920):

- Given a graph $G(V, E)$ and $\beta \in (0,1)$,
  each $\sigma \in \{-1, +1\}^V$ is assigned a weight:

$$w(\sigma) = \beta^{\sum_{uv \in E} |\sigma_u - \sigma_v|/2}$$

- Generate a random $\sigma \in \{+1, -1\}^V$
  with probability $\propto w(\sigma)$.

- "Resolve" the model:
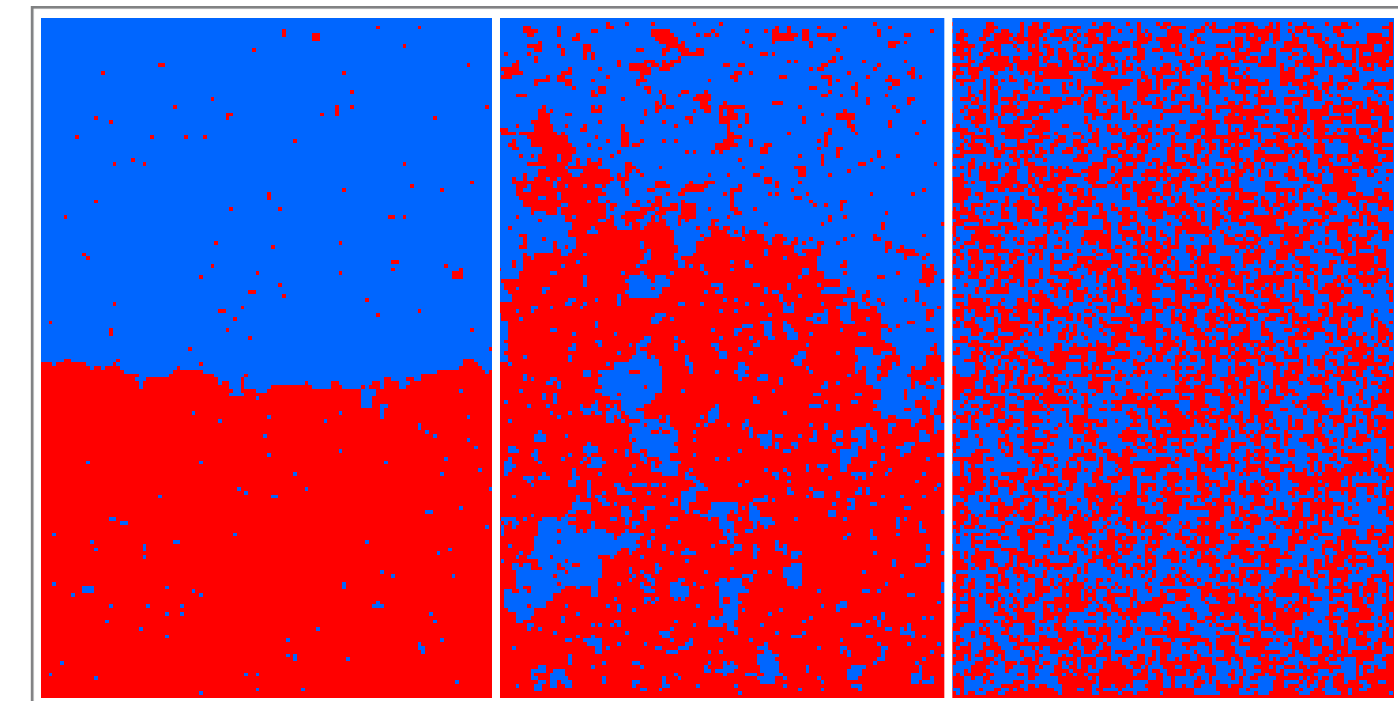
  - Ising (1D), Onsager (2D)
  - Jerrum-Sinclair (1989): Monte Carlo algorithm

$\beta < \beta_c$ $\beta = \beta_c$ $\beta > \beta_c$

Ising (1924), Onsager (1944), Lee-Yang (1952)

Kaufman, Onsager, Yang, Kac, Ward, Potts,
Montroll, Hurst, Green, Kasteleyn, McCoy, Wu,
Vdovichenko, Fisher, Baxter, ...

PageRank

- Rank pages in the World Wide Web:

  - A page has higher rank if pointed by more high-rank pages.

  - High-rank pages have greater influence.

  - Pages pointing to few others have greater influence.

- Rank of page $v \in V$:

$$r(v) = \sum_{(u,v) \in E} \frac{r(u)}{d_+(u)}$$
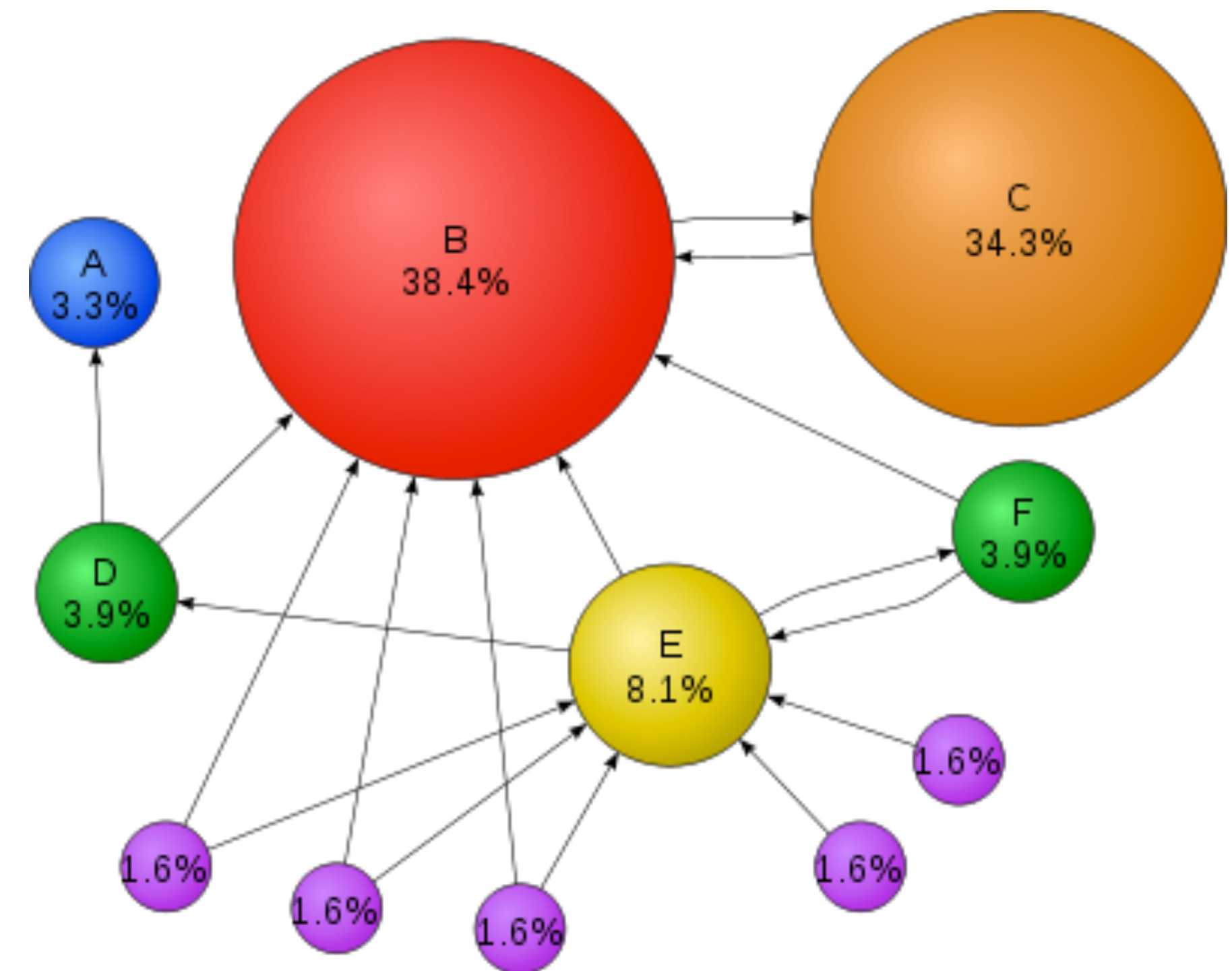
- *Stationary distribution* of random walk!

Directed graph $G(V, E)$

OpenAI's *Sora*
(Creating video from text)

Prompt: Several giant wooly mammoths approach treading through a snowy meadow, their long wooly fur lightly blows in the wind as they walk, snow covered trees and dramatic snow capped mountains in the distance, mid afternoon light with wispy clouds and a sun high in the distance creates a warm glow, the low camera view is stunning capturing the large furry mammal with beautiful photography, depth of field.

- *Sora* is an AI model that can create realistic and imaginative scenes from text instructions.

- It generates videos from text by inferring in *diffusion processes* (a class of random processes).

Prompt: Drone view of waves crashing against the rugged cliffs along Big Sur's garay point beach. The crashing blue waters create white-tipped waves, while the golden light of the setting sun illuminates the rocky shore. A small island with a lighthouse sits in the distance, and green shrubbery covers the cliff's edge. The steep drop from the road down to the beach is a dramatic feat, with the cliff's edges jutting out over the sea. This is a view that captures the raw beauty of the coast and the rugged landscape of the Pacific Coast Highway.

A Shiba Inu dog wearing a beret and black turtleneck.