Foundations of Data Science

Course Introduction

刘明谋 Nanjing University, Suzhou, 2025 Fall

Course Info

- Instructors: 刘明谋
  - Emails: lmm@nju.edu.cn
- Office hour: Thu 3-5pm
  - 南雍-西229
  - QQ group：1019436733
- Course homepage:
  - https://tcs.nju.edu.cn/wiki

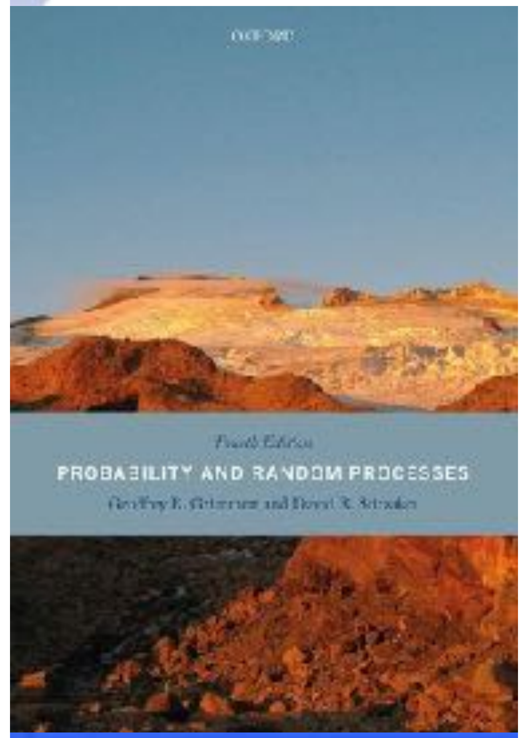
数据科学基础（南京...

群号：1019436733

扫一扫二维码，加入群聊

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）

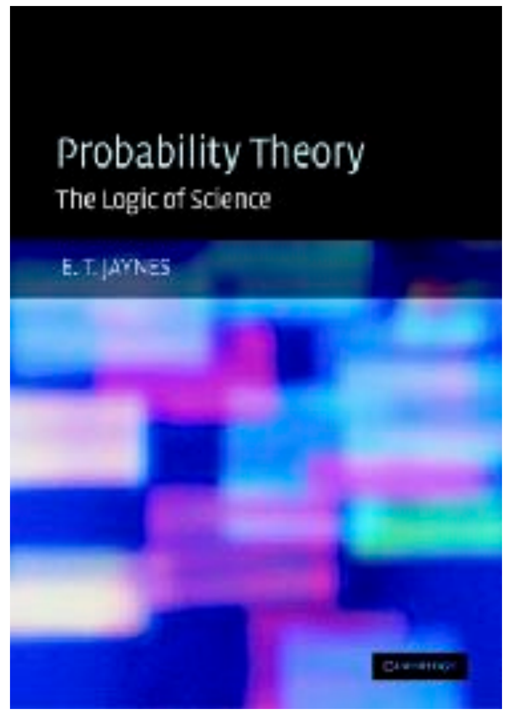


- ***Probability and Random Processes***, *4th edition*.
  Geoffrey Grimmett, David Stirzaker.
  Oxford University Press (2020).

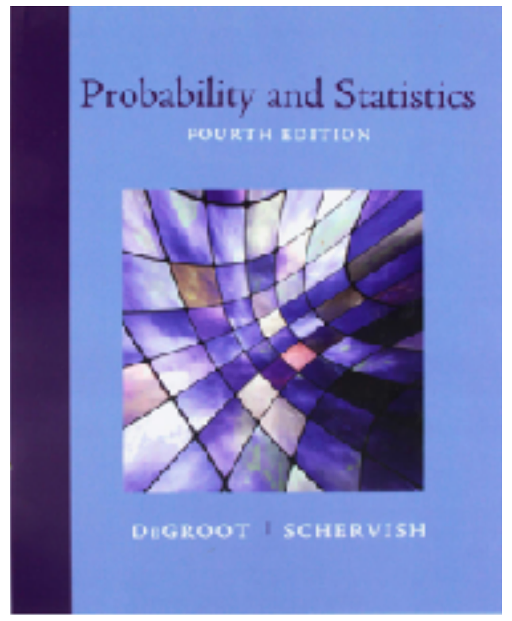
- ***Probability and Computing:*** *Randomization and Probabilistic Techniques in Algorithms and Data Analysis,* *2nd edition*.
  Michael Mitzenmacher, Eli Upfal.
  Cambridge University Press (2017).

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》



- ***Probability and Statistics*** (Classic Version), *4th Edition*. Morris DeGroot, Mark Schervish. Pearson (2018). 《概率统计》

Syllabus

~~Foundations of data science~~

Probability and statistics for computer science

- **经典概率论**： 理解原理
  - 概率空间、随机变量及其数字特征、多维及连续随机变量、极限定理
- **概率与计算**： 求解问题
  - 概率法 (the probabilistic method)、测度集中 (concentration of measure)、离散随机过程
- **数理统计**： 掌握语言
  - 经典与贝叶斯统计的概念和语言（参数估计、假设检验、回归分析……）
- 最终成绩=70%平时作业 + 30%期末考试
  - 作业量+难度较去年低（教室大了）

◉ 光上课：脑子偷懒，“啊对对对”

◉ 在练习中熟悉和掌握

Syllabus

~~Foundations of data science~~

Probability and statistics for computer science

- **费曼学习法**（Feynman Technique）
  - Choose a concept
  - Teach it in simple terms to yourself or someone else
  - Identify gaps by noting where you struggled, and
  - Simplify your explanation, using analogies, until the concept is clear and concise
- 通过输出来“不偷懒”

Richard Feynman (1918-1988)

◉ 光上课：脑子偷懒，“啊对对对”

◉ 在练习中熟悉和掌握

- **What**：什么是随机？什么是概率？
    - (informally) 对不确定性的建模；对可能性的估计
- **Why**：为什么研究随机？为什么研究概率？
- **Belief**：你相信世界是随机的吗？

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

Baruch Spinoza (1632-1677)

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*.)

“某件事 A 发生的概率是 p”

- **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.“《黑神话》年度最佳游戏稳了”、“《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**

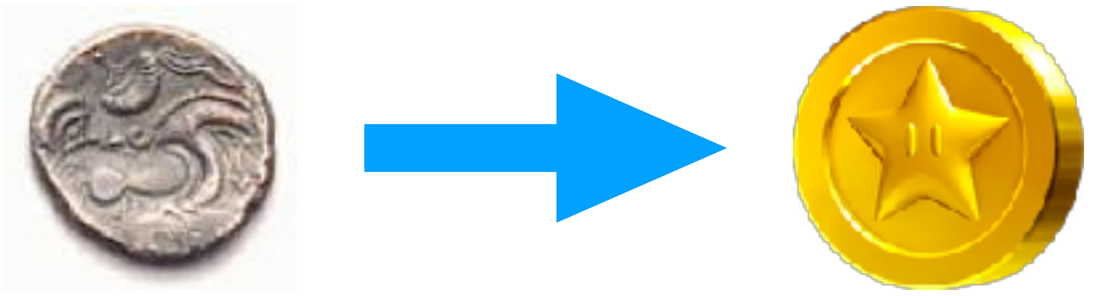
  Gian-Carlo Rota (1998) in MIT's course 18.313 on *Probability*: "The idea of probability is one of the great ideas developed in this (20th) century. You can find it in all the sciences."

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 $\mathbf{HEADS}$, say p, is **unknown**. How can you use this coin to generate unbiased (fair) coin-flips."



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

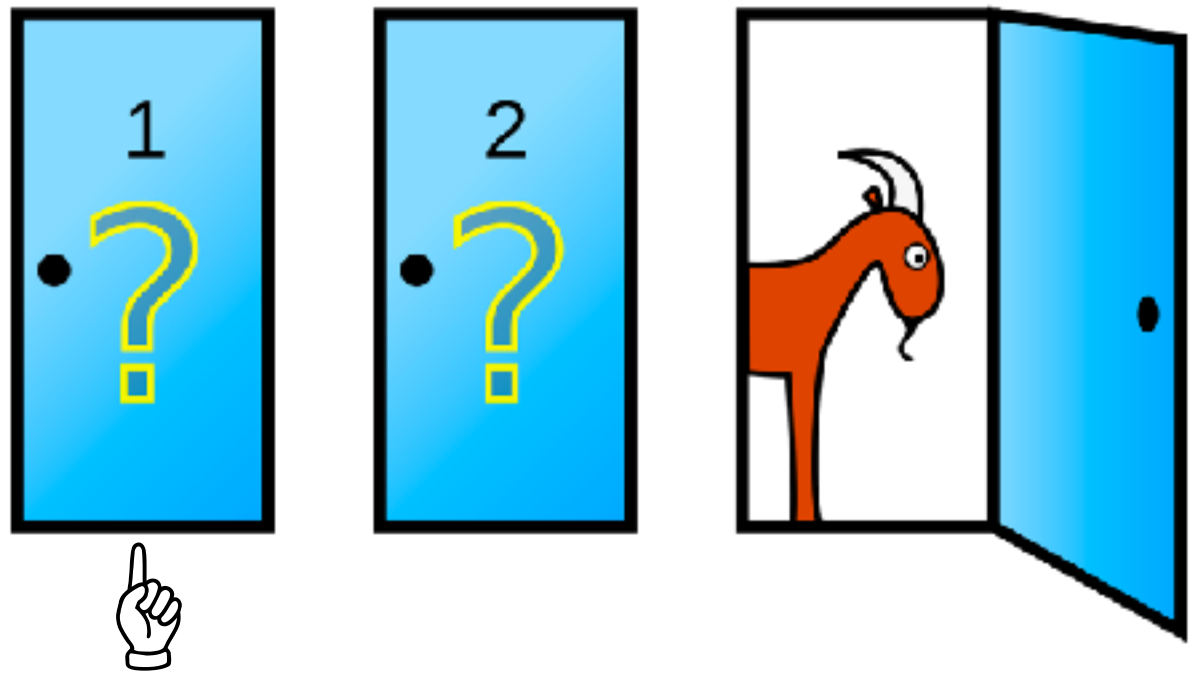
The Two Child Problem
(boy or girl paradox)

- 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.

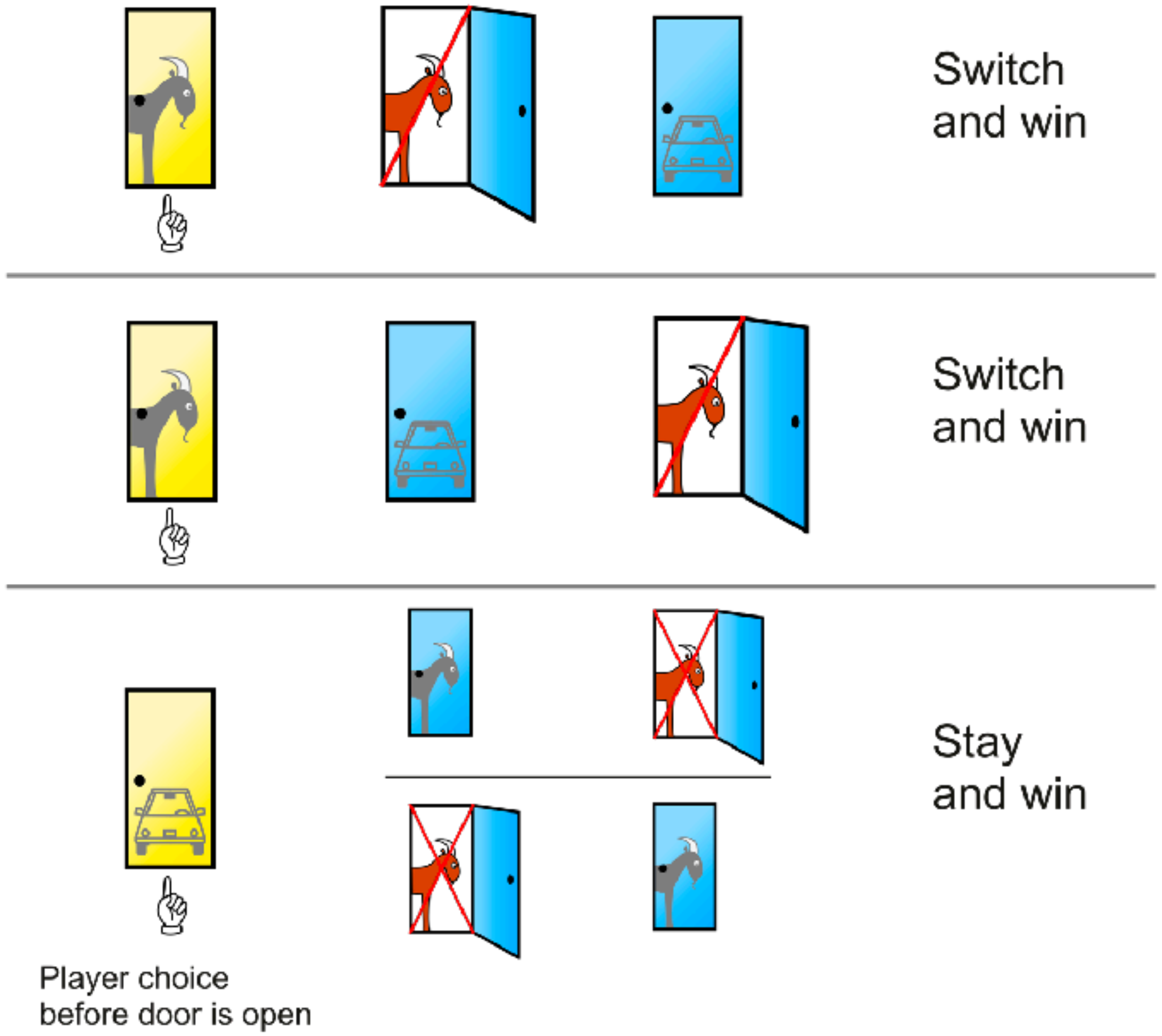
- 💡: Understand the probability space.

Monty Hall Problem

(three doors problem)



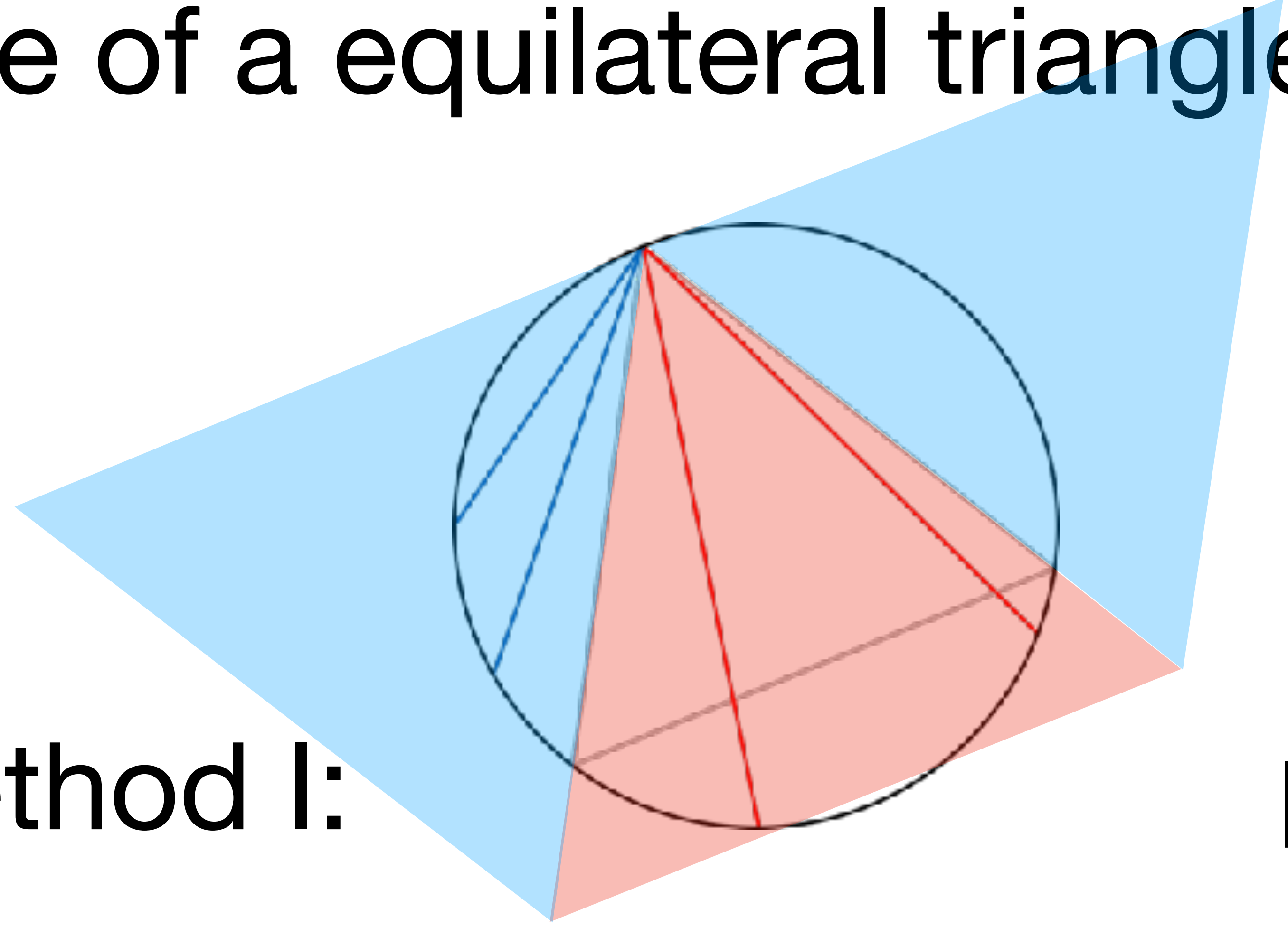
- Suppose you’re on a game show, and you’re given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what’s behind the doors, opens another door, say No. 3, which has a goat. He then says to you, “Do you want to pick door No. 2?” Is it to your advantage to switch your choice?
- 💡: Consider a 1-car-100-door variant.



Bertrand Paradox (贝特朗悖论)

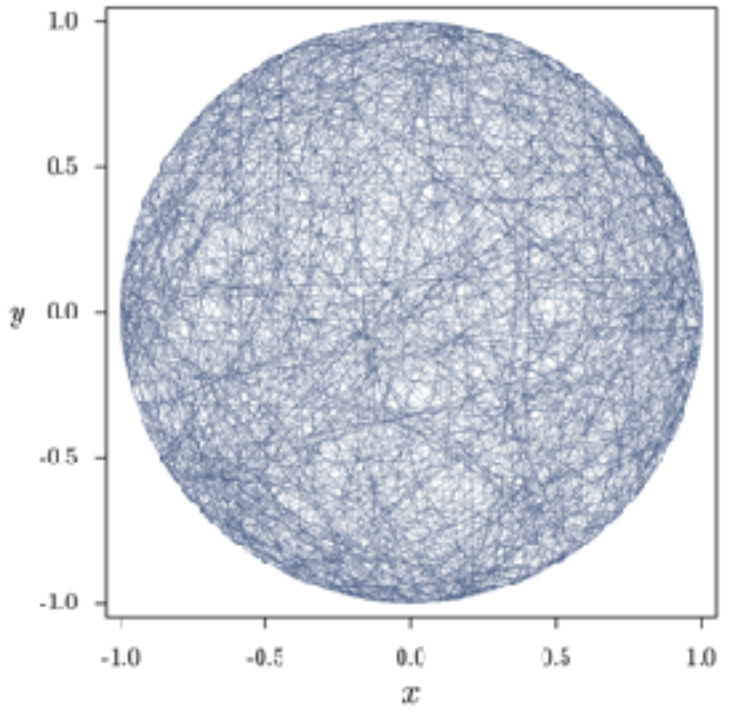
introduced in *Calcul des probabilités* (1889) by Joseph Bertrand

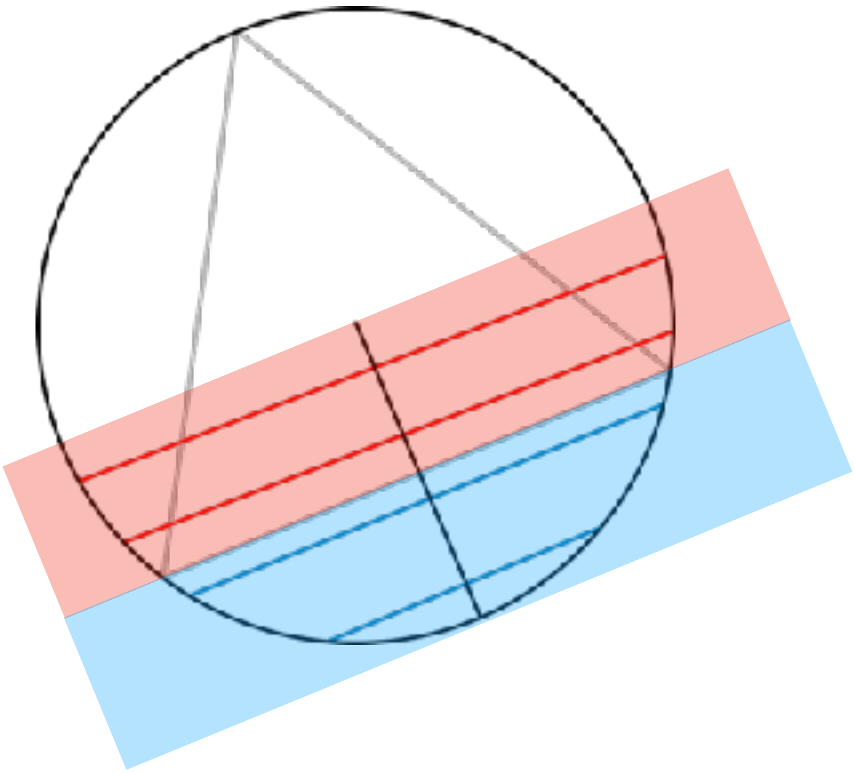
- What is the probability that a ***random chord of a circle***, is longer than the side of a equilateral triangle inscribed in a circle?



- method I:

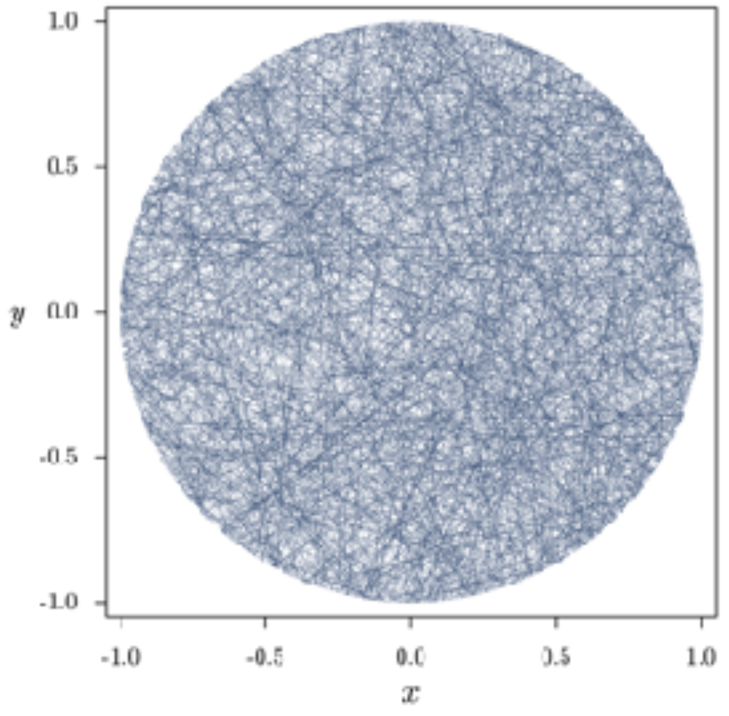
Prob=$1/3$





- method II:

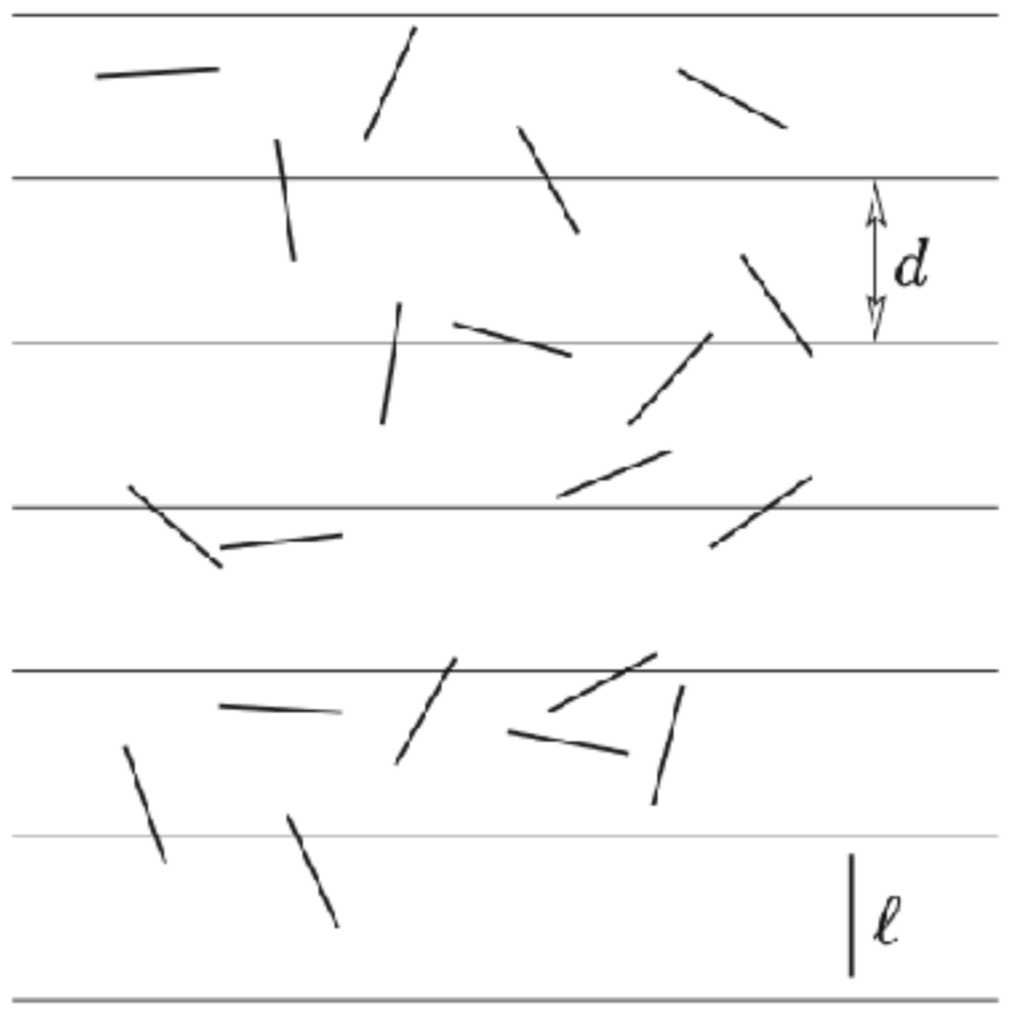
Prob=$1/2$



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

Buffon's Needle Problem (蒲丰投针问题)

- Suppose that you drop a short needle of ℓ on ruled paper, with distance d between parallel lines.
- What is the probability that the needle comes to lie in a position where it crosses one of the lines?
- For $\ell < d$, this probability is calculated as:



$$\frac{2}{d}\frac{2}{\pi}\int_0^{\pi/2}\frac{\ell}{2}\sin\theta\,\mathrm{d}\theta = \frac{2\ell}{d\pi}$$

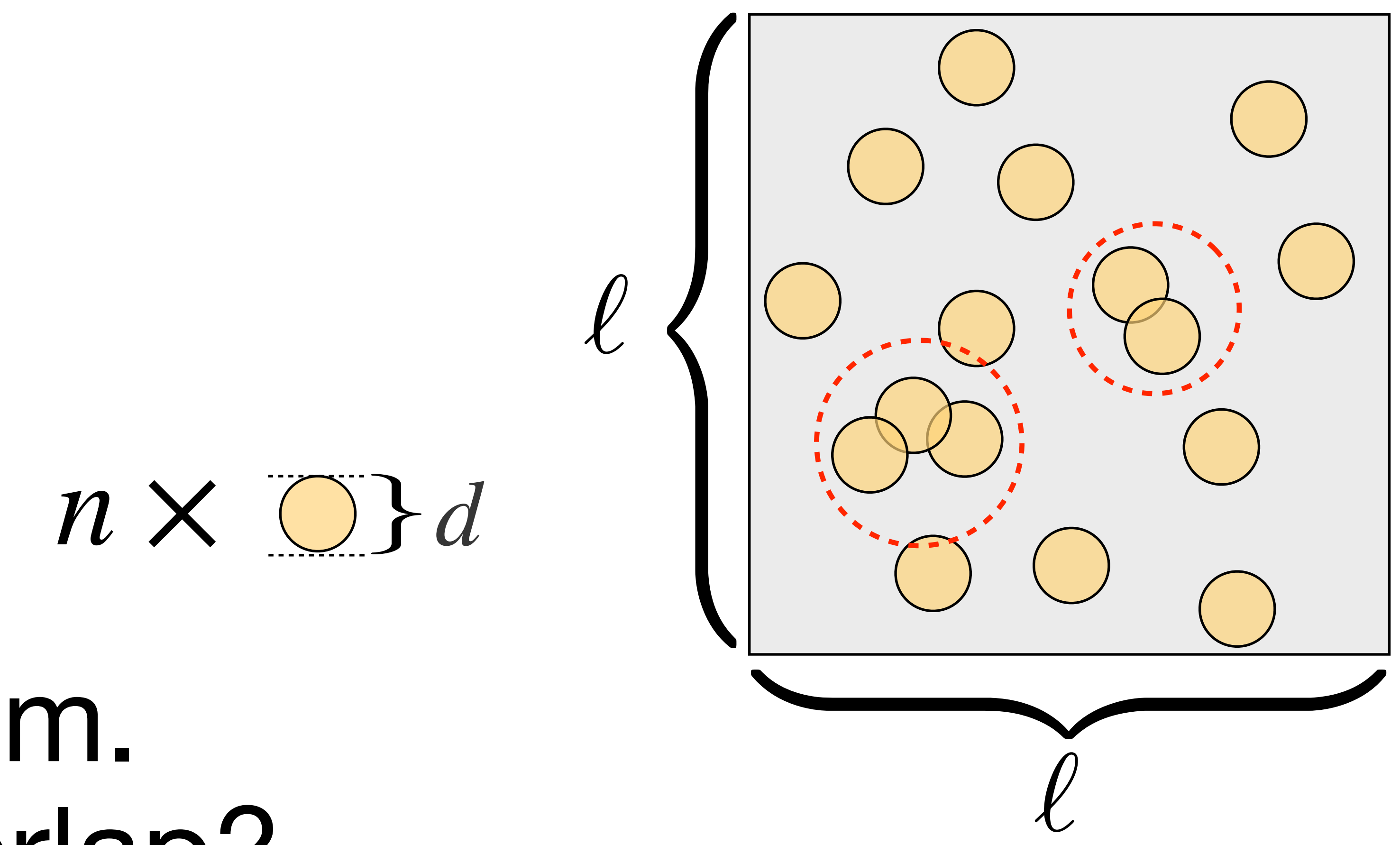
$x \in [0, d/2]$: distance from the center of the needle to the closest parallel line

$\theta \in [0, \pi/2]$: acute angle between the needle and one of the parallel lines

Event: $\left\{(x,\theta) \in \left[0,\frac{d}{2}\right] \times \left[0,\frac{\pi}{2}\right] : x \leq \frac{\ell}{2}\sin\theta\right\}$

Pennies on a Carpet

(hard spheres in 2D square)



- 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):

$$\begin{cases} \prod_{i=1}^{n-1} \frac{\ell - i \cdot d}{\ell - d} & \text{if } \ell \geq nd \\ 0 & \text{otherwise} \end{cases}$$

- 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 100°C, on the basis of purely atomic computations.

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

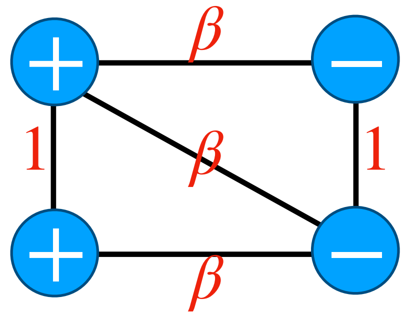
Gian-Carlo Rota teaching 18.313

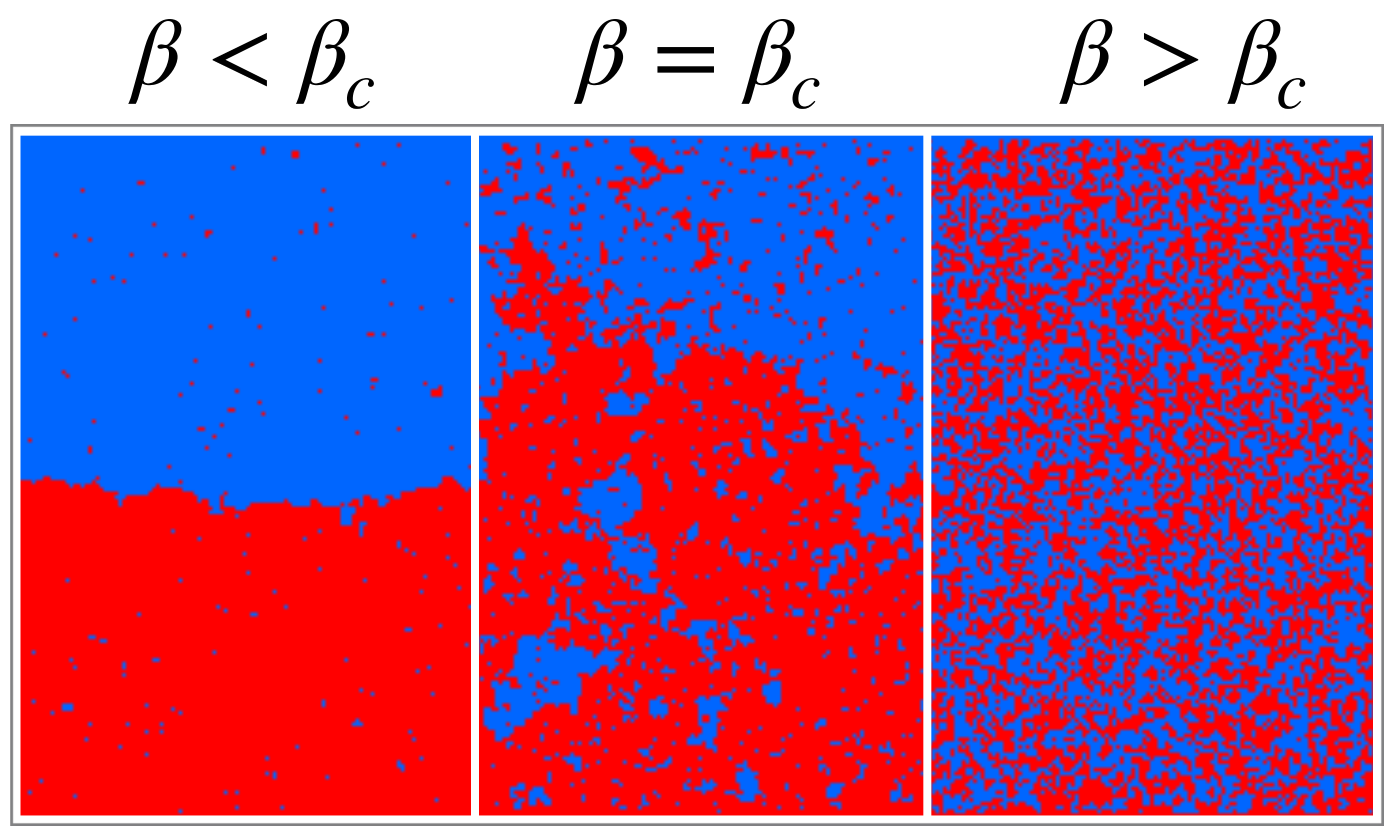
Phase Transition of Ferromagnetism

(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

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

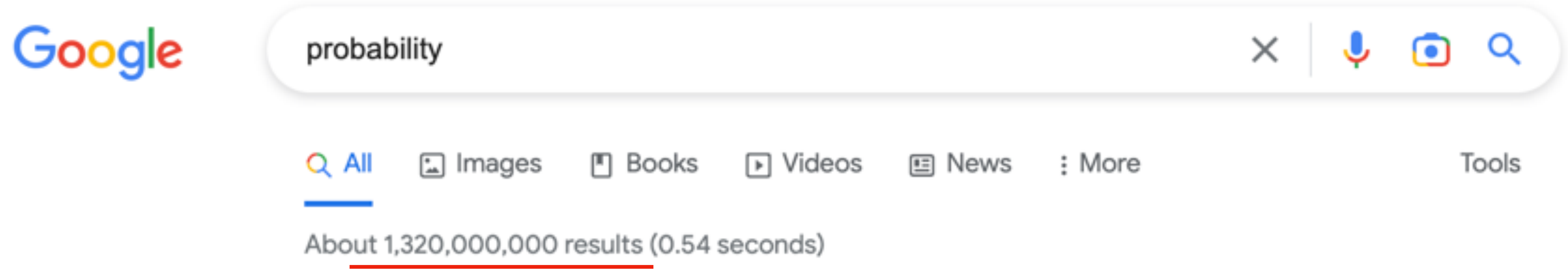
Kaufman, Onsager, Yang, Kac, Ward, Potts, Montroll, Hurst, Green, Kasteleyn, McCoy, Wu, Vdovichenko, Fisher, Baxter, ...

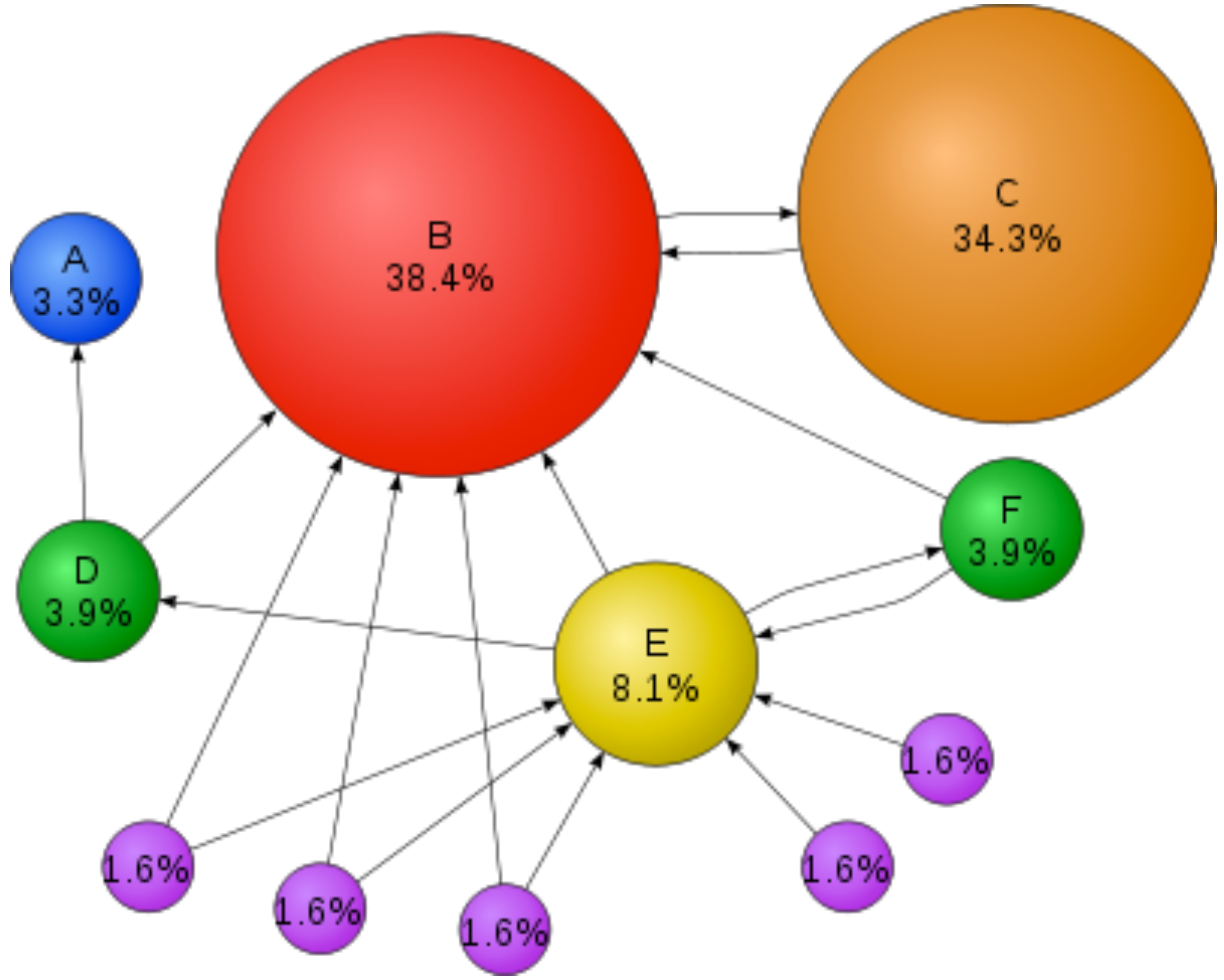
PageRank

- 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!





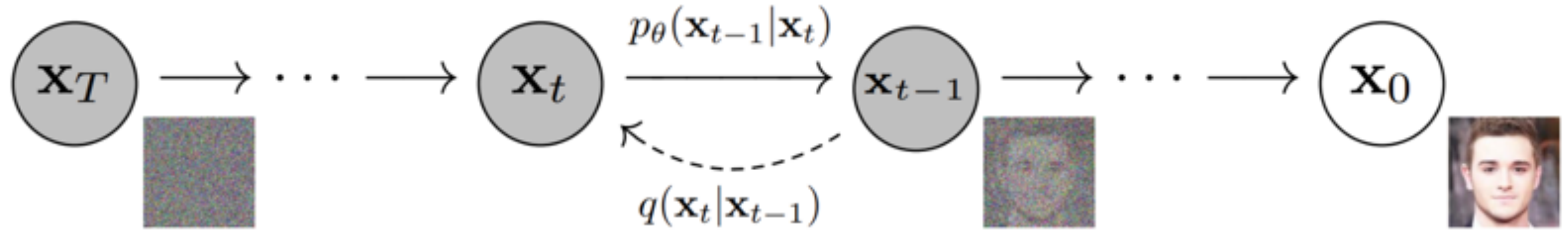
Directed graph $G(V, E)$

OpenAI's *Sora*

(Creating video from text)

- *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: 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.

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.

Probability Space

Sample Space (样本空间)

- <u>Sample space</u> Ω: set of all possible outcomes of an experiment (<u>samples</u>).
  - **Example:** all sides of a dice; all outcomes of a sequence of coin tosses; …
- A family $\Sigma \subseteq 2^{\Omega}$ of subsets of Ω, called <u>events</u> (事件), satisfies:
  - $\varnothing$ and Ω are events (the <u>*impossible event*</u> and <u>*certain event*</u>);
    “不可能事件” “必然事件”
  - if A is an event (i.e. if $A \in \Sigma$), then so is its complement $A^c = \Omega \backslash A$;
  - if $A_1, A_2, \ldots$ are events (i.e. if $A_1, A_2, \ldots \in \Sigma$), then so is $\bigcup_i A_i$ (and $\bigcap_i A_i$)

σ-Algebra (σ-代数)

- A family $\Sigma \subseteq 2^\Omega$ of subsets of Ω is called a **σ-algebra**, if:
  - $\varnothing \in \Sigma$;
  - $A \in \Sigma \implies A^c \in \Sigma$ (where $A^c = \Omega \backslash A$);
  - $A_1, A_2, \ldots \in \Sigma \implies \bigcup_i A_i \in \Sigma$.
- **Examples**: $\Sigma = 2^\Omega$; $\Sigma = \{\varnothing, \Omega\}$; $\Sigma = \{\varnothing, A, A^c, \Omega\}$ for any $A \subseteq \Omega$

Sets as Events

Notation	Set interpretation	Event interpretation
$\omega \in \Omega$	Member of Ω	Sample, elementary event
$A \subseteq \Omega$	Subset of Ω	Event A occurs
A^c	Complement of A	Event A does not occur
$A \cap B$	Intersection	Both A and B
$A \cup B$	Union	Either A or B or both
$A \backslash B$	Difference	A, but not B
$A \oplus B$	Symmetric difference	Either A or B, but not both
$A \subseteq B$	Inclusion	A implies B
$A \cap B = \varnothing$	Set disjointness	A and B cannot both occur
$\varnothing$	Empty set	Impossible event
Ω	Whole space	Certain event

Probability Measure (概率测度)

- Let $\Sigma \subseteq 2^{\Omega}$ be a <u>σ-algebra</u>.
- A <u>**probability measure**</u> (also called <u>**probability law**</u> 概率律) $\Pr$ on sample space Ω (with events Σ) is a function $\Pr : \Sigma \to [0,1]$ satisfying:
  - (*normalized*) $\Pr(\Omega) = 1$;
  - (*additive*) for **disjoint (不相容)** $A_1, A_2, \ldots \in \Sigma$: $\Pr\left(\bigcup_i A_i\right) = \sum_i \Pr(A_i)$.
- The triple $(\Omega, \Sigma, \Pr)$ is called a <u>**probability space**</u>.

Basic Properties of Probability

- $\Pr(A^c) = 1 - \Pr(A)$
- $\Pr(\varnothing) = 0$
- $\Pr(A \backslash B) = \Pr(A) - \Pr(A \cap B)$
- $A \subseteq B \implies \Pr(A) \leq \Pr(B)$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- “随机自然数是偶数的概率为1/2” 这句话没有意义 (但是“[0,1]中随机实数是有理数的概率为0”是对的)
- All can be deduced from the **axioms** of probability space.

Union Bound

- **Union bound** (Boole's inequality): for events $A_1, A_2, \ldots A_n \in \Sigma$

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i)$$

- **Example:** A machine has n parts, each of which breaks down with prob. p

$\Pr[$ all parts are fine $]$

$= 1 - \Pr[\exists$ a part breaks down $] \leq 1 - np$

Holds unconditionally.

(tight if all bad events are disjoint)

Principles of Inclusion-Exclusion

- **Principle of inclusion-exclusion**: for events $A_1, A_2, \ldots A_n \in \Sigma$,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i) - \sum_{i<j} \Pr(A_i \cap A_j) + \sum_{i<j<k} \Pr(A_i \cap A_j \cap A_k) - \cdots$$

$$= \sum_{\substack{S \subseteq \{1,2,\ldots,n\} \\ S \neq \emptyset}} (-1)^{|S|-1} \Pr\left(\bigcap_{i \in S} A_i\right)$$

- **Boole-Bonferroni Ineqaulity**: for events $A_1, A_2, \ldots A_n \in \Sigma$, for any $k \geq 0$

$$\sum_{\substack{S \subseteq \{1,2,\ldots,n\} \\ 1 \leq |S| \leq 2k}} (-1)^{|S|-1} \Pr\left(\bigcap_{i \in S} A_i\right) \leq \Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{\substack{S \subseteq \{1,2,\ldots,n\} \\ 1 \leq |S| \leq 2k+1}} (-1)^{|S|-1} \Pr\left(\bigcap_{i \in S} A_i\right)$$

Derangement (错排)
(le problème des rencontres)

- The probability that a random permutation $\pi : [n] \to [n]$ has no fixed point (i.e. there is no $i \in [n]$ such that $\pi(i) = i$).

- Let A_i be the event that $\pi(i) = i$.

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} \sum_{S \in \binom{\{1,2,\ldots,n\}}{k}} (-1)^{k-1} \Pr\left(\bigcap_{i \in S} A_i\right) = \sum_{k=1}^{n} \binom{n}{k} (-1)^{k-1} \frac{(n-k)!}{n!} = -\sum_{k=1}^{n} \frac{(-1)^k}{k!}$$

$$\Pr\left(\bigcap_{i=1}^{n} A_i^c\right) = 1 - \Pr\left(\bigcup_{i=1}^{n} A_i\right) = 1 + \sum_{k=1}^{n} \frac{(-1)^k}{k!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \to \frac{1}{\mathrm{e}} \text{ as } n \to \infty$$

Classical Examples of Probability Space

- 古典概型 (classic probability): ***discrete uniform probability law***

  Finite sample space Ω, each outcome $\omega \in \Omega$ has equal probability.

  For every event $A \subseteq \Omega$: $\Pr(A) = \dfrac{|A|}{|\Omega|}$.

- 几何概型 (geometric probability): continuous probability space where $\Pr(A)$ is calculated as some geometric measure of A.
  - Bertrand’s paradox
  - Buffon’s needle problem

$\Pr \propto \angle$