Probability and Statistics

Introduction to Probability

谢润烁 Nanjing University, 2023 Fall

The Seminar

- Time: 7:00 p.m. Thu.
- Page: https://leonicatot.github.io/seminars/2023Fall-Probability/
- Textbook: *Probability and Computing*

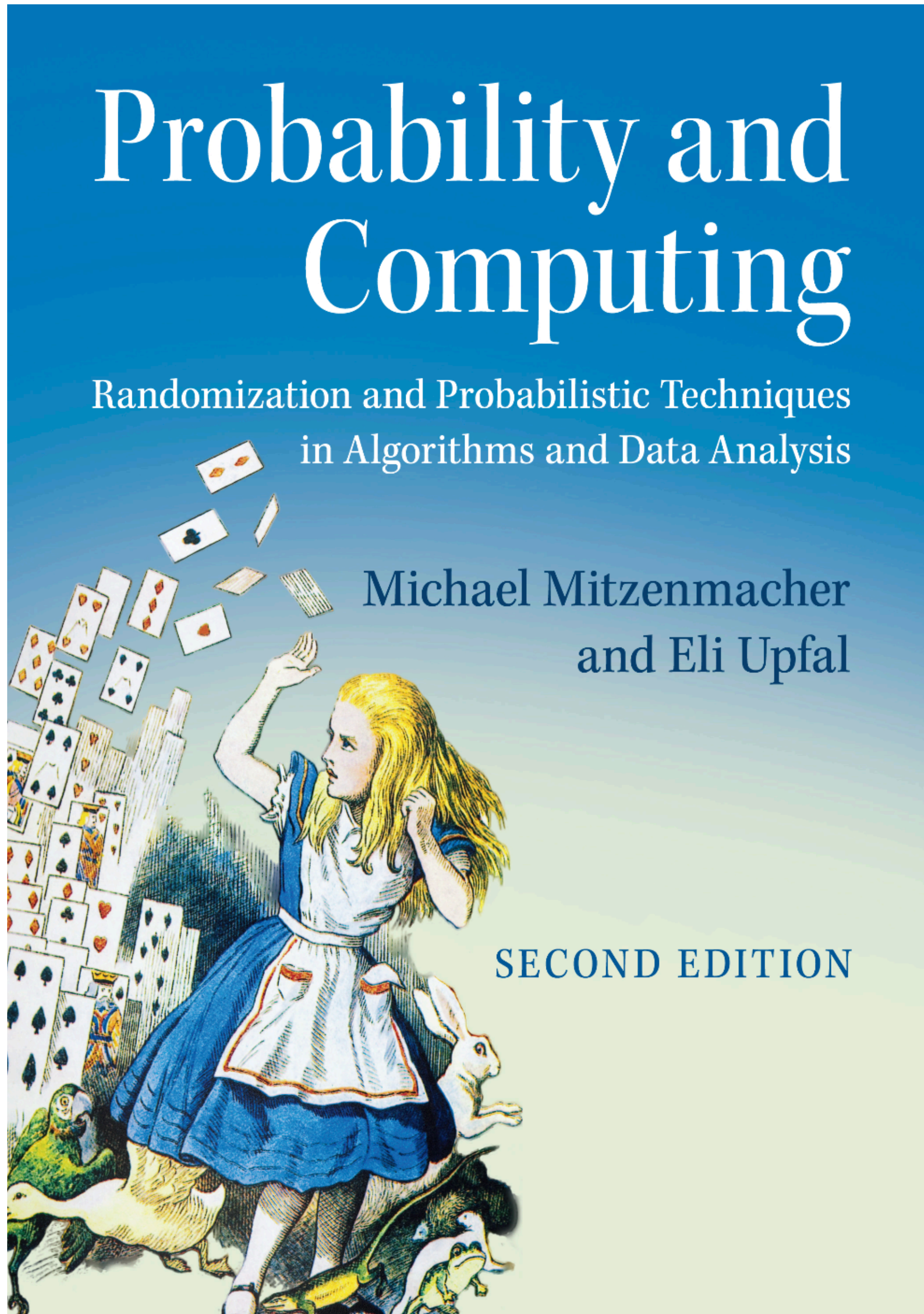
Seminar Information

Content

In this seminar, we'll be using *Probability and Computing* by *Michael Mitzenmacher* and *Eli Upfal* as our main material. Hopefully, we will cover at least **Chapter 1 to Chapter 7** of this book in this seminar, which are:

- Events and Probability
- Discrete Random Variables and Expectation
- Moments and Deviations
- Chernoff and Hoeffding Bounds
- Balls, Bins, and Random Graphs
- The Probabilistic Methods
- Markov Chains and Random Walks

Besides, we might also cover some topics beyond these 7 chapters:



Content

- A quick review
- Our intuition
- History of probability theory
- Probability and algorithms
- Probability and measure theory
- Probability and statistics
- ~~Probability: the logic of science~~

A Quick Review

Randomness Generator

- Dice
- Coin
- Roulette
- ...

Sample Space

- **Sample space Ω**: the set of all outcomes in an experiment
- Flip a coin: $\Omega =$
  - $\{H, T\}$
- Throw a dice: $\Omega =$
  - $\{1,2,3,4,5,6\}$

Events

- **Events Σ**: subset of 2^{Ω} *
  - Event A: the outcome of throwing a dice is even
    - $A =$
      - $\{2,4,6\}$
    - $\Pr(A) =$
      - $\frac{1}{2}$

*: Note that not all subset of 2^{Ω} can be an event

Probability Space

- Sample space Ω
- Event Σ
- **Probability Measure** $\Pr$
  - $\Pr(\emptyset) = 0$ and $\Pr(\Omega) = 1$
  - $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$ for disjoint A_i*
- **Probability Space**: $(\Omega, \Sigma, \Pr)$

*: We will discuss about the definition rigorously in later seminars

More Dices …

- Throw two dice: $\Omega =$
    - $\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$
- $\Pr[(1,1)]$
    - $= \dfrac{1}{36}$
- We don't care about the exact outcome — We only care about the sum of points

Random Variable

- Random Variable
  - $X : \Omega \to \mathbb{R}$
- E.g. X: total points of throwing two dice
  - $X((1,3)) = X((2,2)) = 4$
  - $\Pr(X = 4) = \dfrac{3}{36}$

Problems from Early Times

French Society in the 1650's

- Gambling was popular and fashionable
- Not restricted by law
- As the games became more complicated and the stakes became larger, there was a need for mathematical methods for computing chances.

Adapted from *A Short History of Probability by Dr. Alan M. Polansky*

Division of the Stakes

- We consider a simplified version:
  - Two players *Alice* and *Bob* flip a coin
    - Head: Alice += 1
    - Tail: Bob += 1
    - The first to reach 100 points will win
  - The match is interrupted before finished
  - How to divide the stake?

Reference: https://en.wikipedia.org/wiki/Problem_of_points

Early Solutions

- 1494, Luca Pacioli
  - divide the stakes in proportion to the number of rounds won by each player
  - Consider 1–0
- mid-16th century, Niccolò Tartaglia
  - Base the division on the ratio between the size of the lead and the length of the game
  - Consider 99—89

Pascal and Fermat

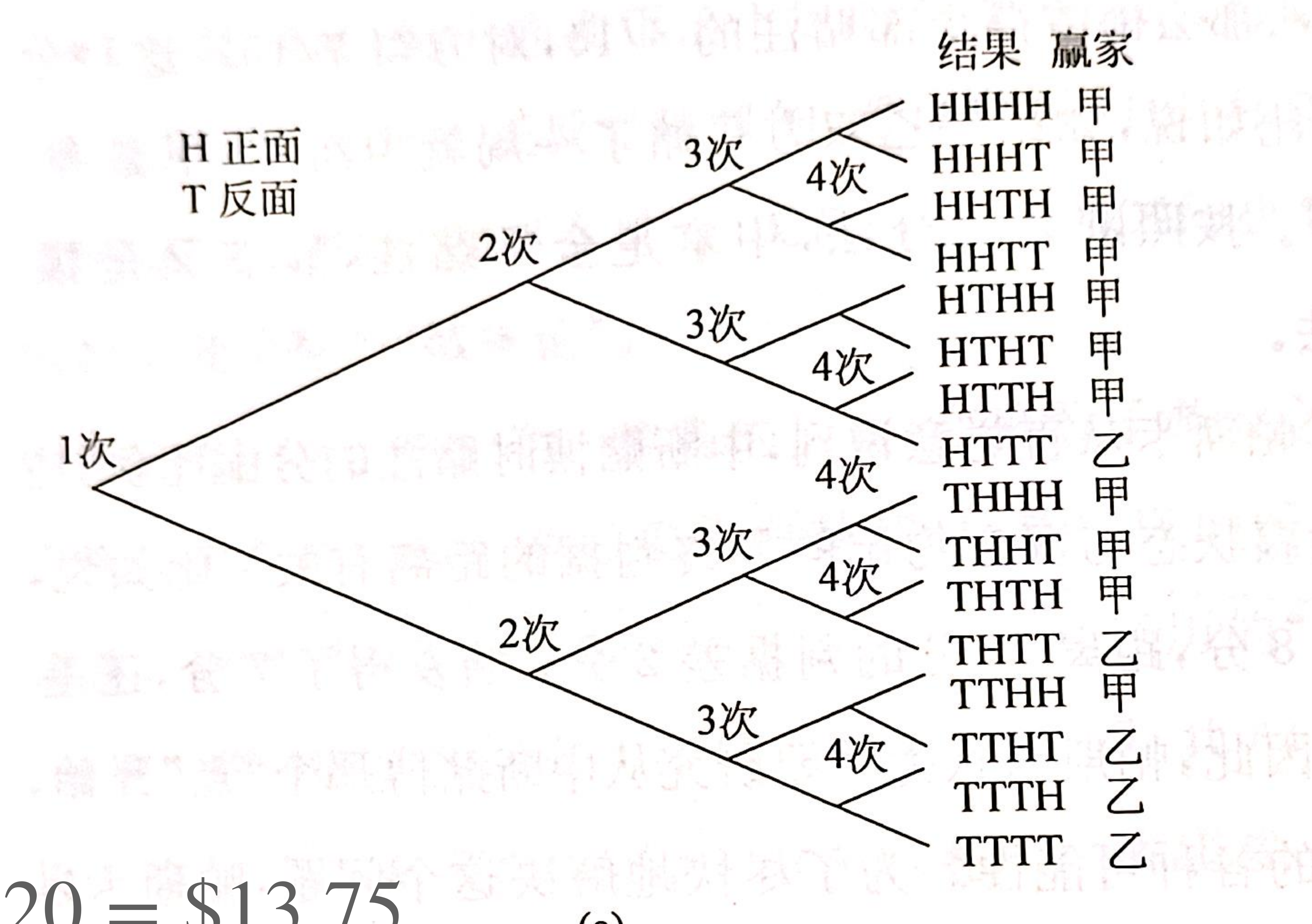
- 1654, Chevalier de Méré posed it to Blaise Pascal
- We consider a simple scene
  - Alice and Bob both place a stake of $10
  - The first to reach 10 points will win
  - When the game is interrupted, Alice : Bob = 8 : 7
  - How to divide the $20?

Fermat’s Solution

- If one player needs r more rounds to win and the other needs s, the game will surely have been won by someone after $r + s - 1$additional rounds
- In total these rounds have 2^{r+s-1} different possible outcomes
- In some of these possible futures the game will actually have been decided in fewer than $r + s - 1$ rounds
  - but it does no harm to imagine the players continuing to play with no purpose.
- Write down a table of all 2^{r+s-1}possible continuations and counting how many of them would lead to each player winning

1607 ~ 1665
Pierre de Fermat
[pjɛʁ də fɛʁma]

Fermat’s Solution



(a)

$$\frac{11}{16} \times 20 = \$13.75$$

1607 ~ 1665
Pierre de Fermat
[pjɛʁ də fɛʁma]

Pascal's Solution: Expected Value

结果	概率	所得(甲)	概率加权所得
H H	1/4	20	5
H T H	1/8	20	5/2
H T T H	1/16	20	5/4
H T T T	1/16	0	0
T H H	1/8	20	5/2
T H T H	1/16	20	5/4
T H T T	1/16	0	0
T T H H	1/16	20	5/4
T T H T	1/16	0	0
T T T	1/8	0	0

期望值(甲)55/4=$13.75

(b)

1623 ~ 1662
Blaise Pascal
[blɛz paskal]

Pascal's Solution: Expected Value

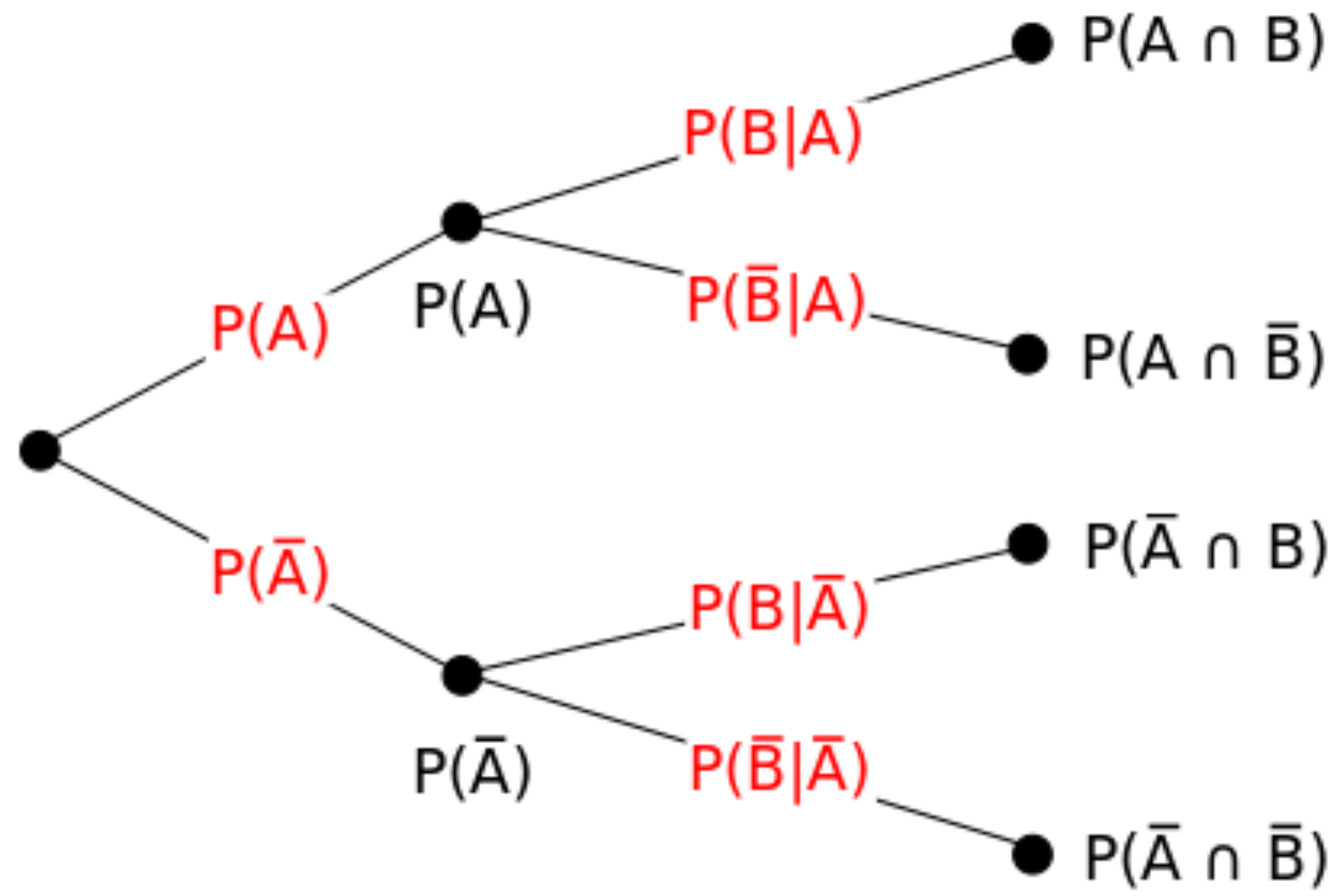
- Through clever manipulation of identities involving what is today known as Pascal's triangle,
- Pascal finally showed that in a game where player a needs r points to win and player b needs s points to win
- the correct division of the stakes between player a and b is:

$$\sum_{k=0}^{s-1} \binom{r+s-1}{k} : \sum_{k=s}^{r+s-1} \binom{r+s-1}{k}$$

1623 ~ 1662
Blaise Pascal
[blɛz paskal]

Visualize using Chance Tree

- A tree diagram may represent a series of independent events or conditional probabilities
- Each node on the diagram represents an event and is associated with the probability of that event.
- The root node represents the certain event and therefore has probability 1.
- Each set of sibling nodes represents an exclusive and exhaustive partition of the parent event.



Reference: https://en.wikipedia.org/wiki/Tree_diagram_(probability_theory)

Expectation/Mean

- For discrete random variable
- $\mathbb{E}(X) = \sum_{x} x \cdot \Pr(X = x)$
- X: The money you can win at the game
- $X < 0$ in casino's scene

St. Petersburg paradox

- If $\mathbb{E}(X) > 0$, is it profitable to play the game?
- Consider the following game:
  - You spend m dollars to play it
  - You can flip a coin until its result becomes tail
  - You get 2^n dollars if you get head n times
- $$\mathbb{E}(X) = \sum_{i=1}^{\infty} \frac{2^n}{2^n} - m = \infty$$

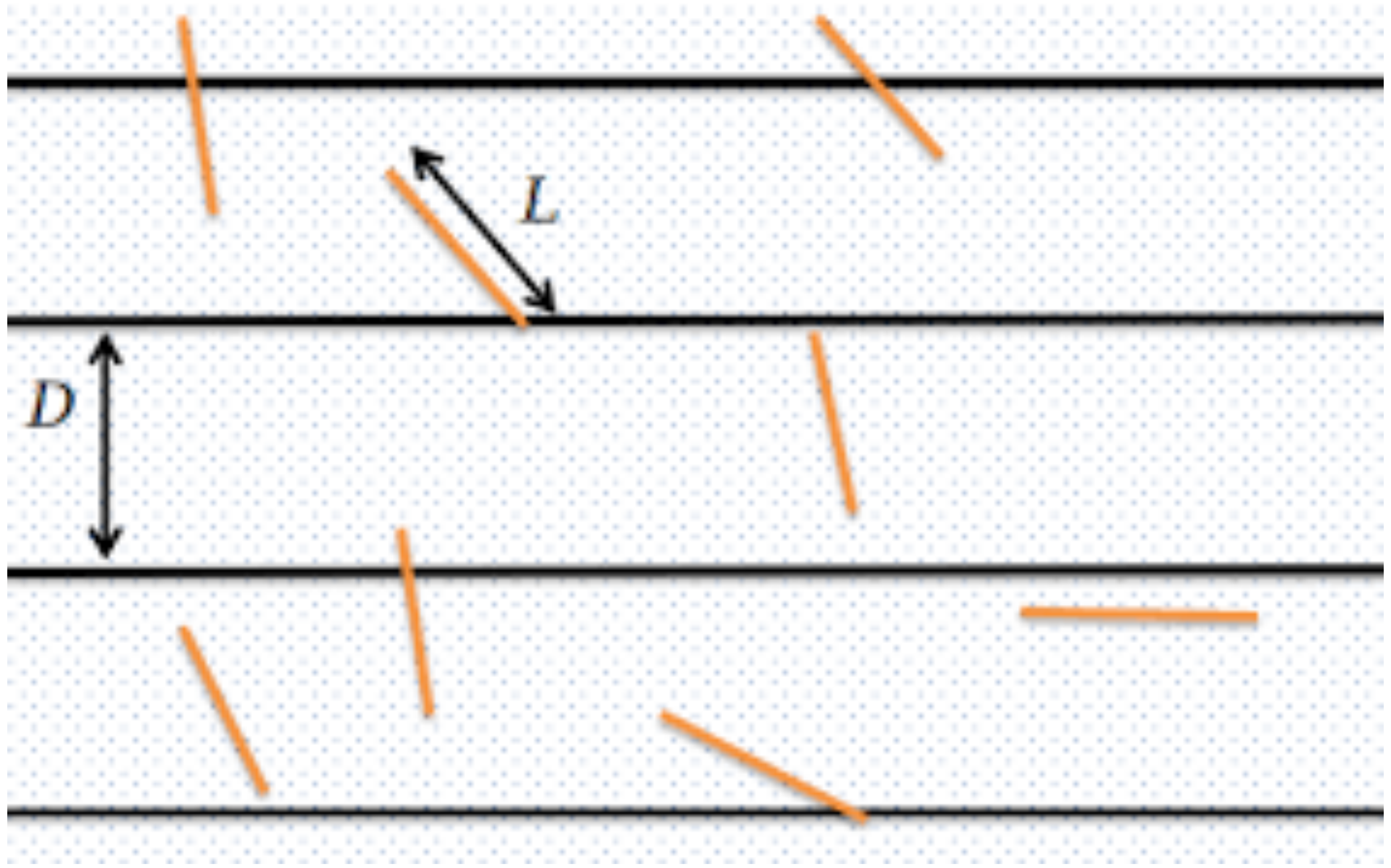
Reference: https://en.wikipedia.org/wiki/St._Petersburg_paradox

Classic and Geometric Probability

- Classical: discrete and uniform
- $\sum_{x \in \Omega} \Pr(x) = 1$
- Geometric: continuous and uniform
- $\int_{\Omega} dF(x) = 1$

Buffon's Needle Problem

- Suppose that you drop a short needle of length L on ruled paper, with distance between parallel lines D ($L < D$).
- What is the probability that the needle comes to lie in a position where it crosses one of the lines?



Reference: https://en.wikipedia.org/wiki/Buffon's_needle_problem

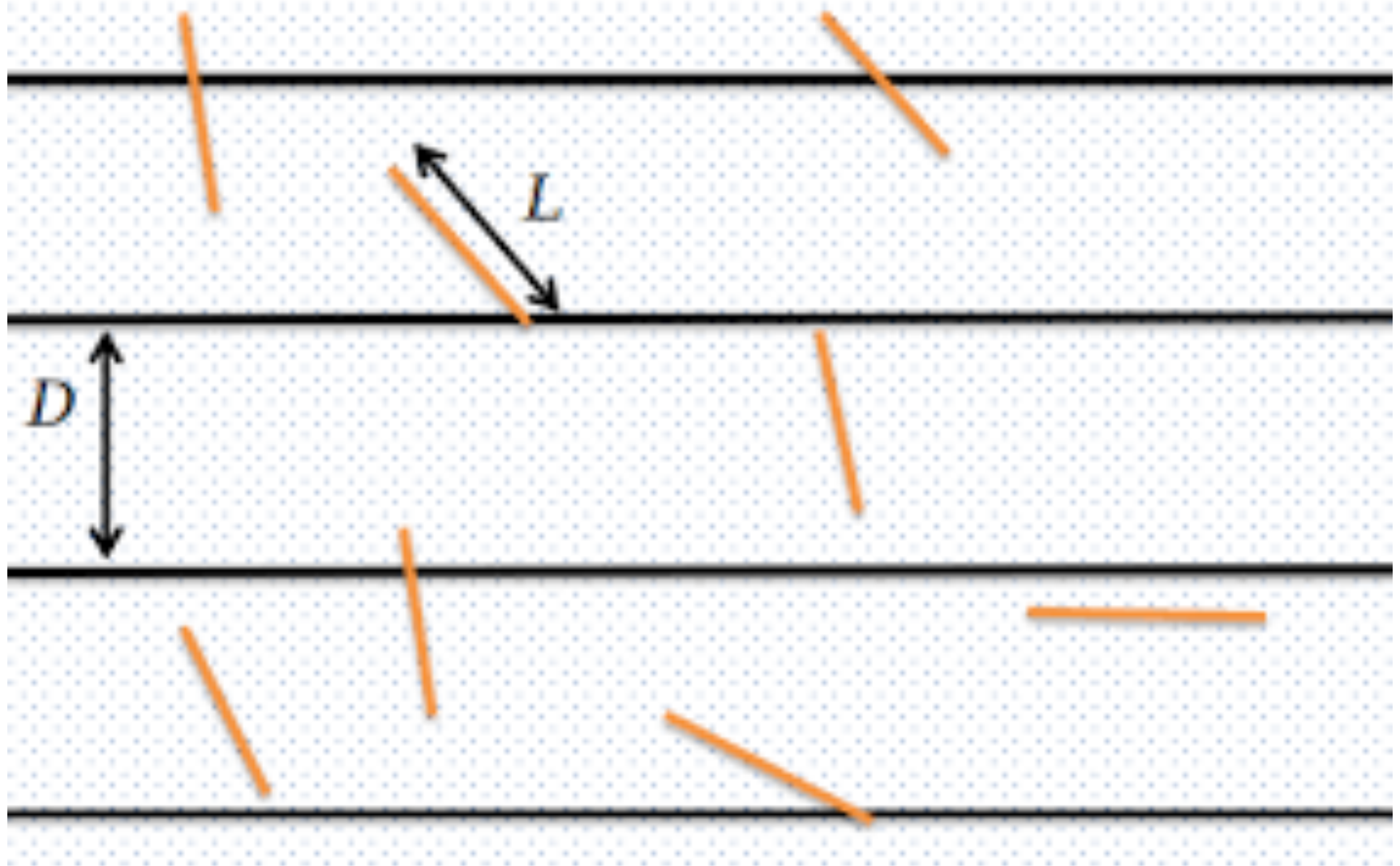
Adapted from Prof. Yitong Yin's slides

Buffon's Needle Problem

- This probability is calculated as:

$$\Pr(A) = \frac{2}{D\pi}\int_0^{\pi}\int_0^{\frac{L}{2}\sin\theta} dx d\theta = \frac{2L}{D\pi}$$

- A *Monte Carlo method** for computing π



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

$\theta \in [0, \pi]$: angle between the needle and the parallel line below it

Event $A = \left\{ (x, \theta) \in [0, \frac{D}{2}] \times [0, \pi] \mid x \leq \frac{L}{2}\sin\theta \right\}$

Reference: https://en.wikipedia.org/wiki/Buffon's_needle_problem

Adapted from Prof. Yitong Yin's slides

Our Intuition

Base Rate Fallacy / Test Paradox

- A rare disease occurs with probability 0.001.
- 5% testing error:
  - A person with the disease tested $\begin{cases} + & 95\,\% \\ - & 5\,\% \end{cases}$
  - A person without the disease tested $\begin{cases} + & 5\,\% \\ - & 95\,\% \end{cases}$
- If a person is tested “+”, what is the probability that he/she is ill?

Reading: https://en.wikipedia.org/wiki/Base_rate_fallacy

Adapted from Prof. Yitong Yin’s slides

Related Formula

- Conditional Probability

$$\Pr(A \mid B) = \frac{\Pr(AB)}{\Pr(B)}, \Pr(B) \neq 0$$

- Bayes' Theorem

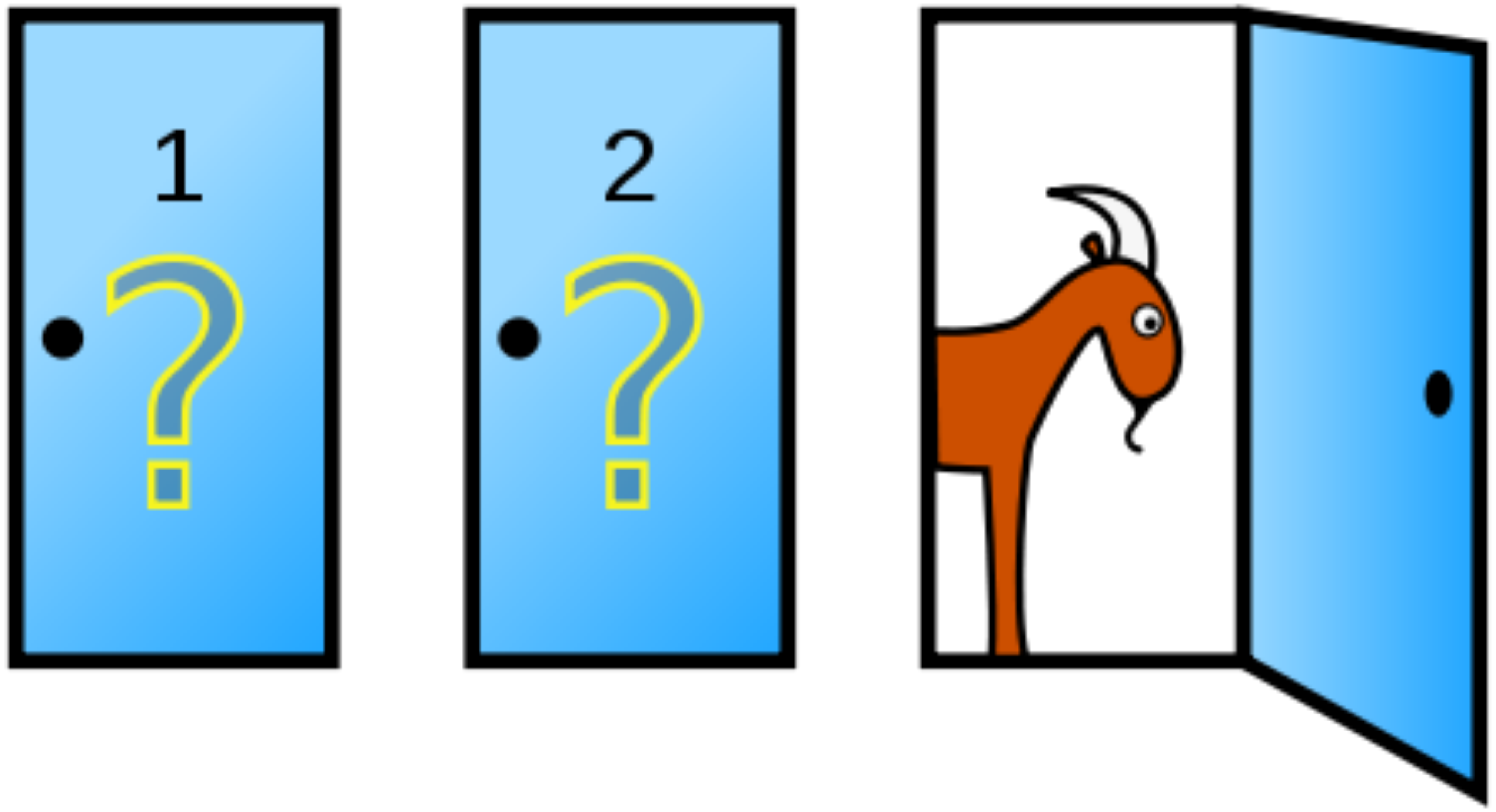
$$\Pr(A \mid B) = \frac{\Pr(AB)}{\Pr(B)} = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

- Law of Total Probability

$$\Pr(A) = \sum_i \Pr(A \mid B_i)\Pr(B_i)$$, where B_i are a partition of Ω

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?



Reference: https://en.wikipedia.org/wiki/Monty_Hall_problem

Adapted from Prof. Yitong Yin's slides

Other Examples

- Birthday Paradox
- Gambler's Fallacy
- Simpson's Paradox
- Random Walk in higher dimensions
  - Shizuo Kakutani: "*A drunk man will find his way home, but a drunk bird may get lost forever.*"
- Benford's Law
- ……

The History of Probability Theory

The History of Probability Theory

- Classical: 1654~1811
- Analysis: 1812~1932
- Modern: 1933~

Reference: 《从博弈问题到方法论学科——概率论发展史研究》 徐传胜著

Classical (1654~1811)

- More on finite and discrete random variable
- Tools
  - Combinatorics
  - Algebra

French Society in the 1650's

- Gambling was popular and fashionable
- Not restricted by law
- As the games became more complicated and the stakes became larger, there was a need for mathematical methods for computing chances.

Adapted from *A Short History of Probability by Dr. Alan M. Polansky*

Correspondence between Pascal and Fermat

- Origin of the mathematical study of probability
- Developed classical approach
- Verified by frequency method

Early Generalizations

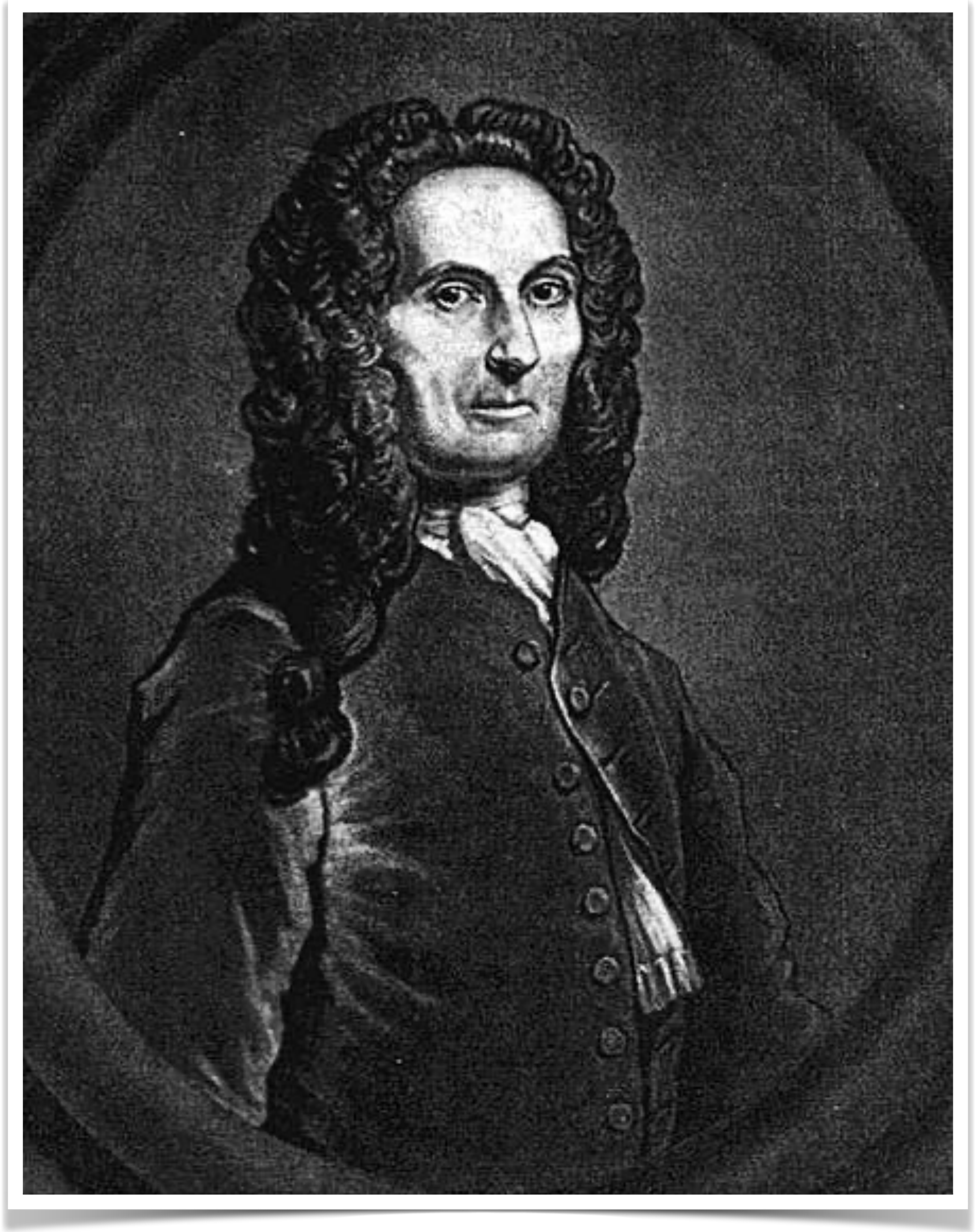
- *Ars Conjectandi*
  - by Jacob Bernoulli in 1713
- Proved that the frequency method and the classical method are consistent
  - Bernoulli's law of large number

1655 ~ 1705
Jacques/Jacob/Jakob/James Bernoulli

*: IPA transcription of people's names will be provided if it is available in Wikipedia.

Early Generalizations

- *The Doctrine of Chances*
  - by Abraham De Moivre in 1718
- Provided many tools to make the classical method more useful
  - Multiplication rule
  - Central limit theorem



1667 ~ 1754
Abraham De Moivre
[abʁaam də mwavʁ] *

From Games to Science

- The 18th century
- The application of probability moved from games of chance to scientific problems
  - Mathematical theory of life insurance - life tables.
  - Biological problems - what is the probability of being born female or male?

Analysis (1812~1932)

- More on continuous random variable
- Tools
  - Characteristic Function
  - Differential Equations
  - Recurrence Relation

Applied Probability

- *Théorie analytique des probabilités*
  - by Pierre-Simon Laplace in 1812
- Presented a mathematical theory of probability with an emphasis on scientific applications

1749 ~ 1827
Pierre-Simon Laplace
[pjɛʁ simɔ̃ laplas]

Stagnation the Frustration

- After the publication of Laplace’s book, the mathematical development of probability stagnated for many years.
- By 1850, many mathematicians found the classical method to be unrealistic for general use and were attempting to redefine probability in terms of the frequency method.
- These attempts were never fully accepted and the stagnation continued.

Modern (1933~)

- Tool
  - Modern Analysis
  - Set Theory
  - Measure Theory

Axiomatic Development

- *Grundbegriffe der Wahrscheinlichkeitsrechnung*
  - by Andrey Kolmogorov in 1933
- Developed the first rigorous approach to probability

1903 ~ 1987
Andrey Nikolaevich Kolmogorov
Андре́й Никола́евич Колмого́ров
[ɐn'drʲej nʲɪkɐ'lajɪvʲɪtɕ kəlmɐ'ɡorəf]

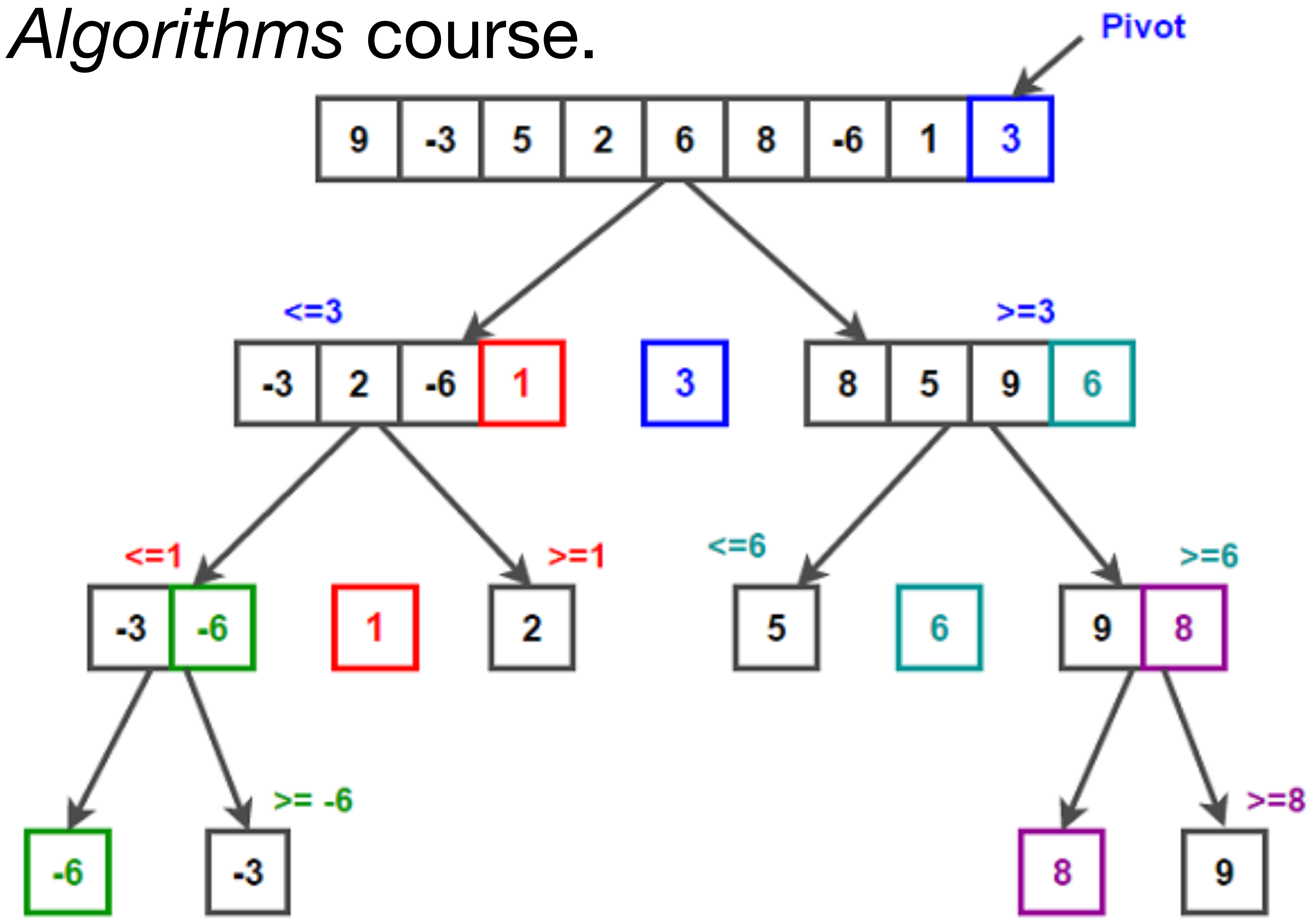
Probability & Algorithms

Analysis of Algorithm

Average Case

- E.g. The average run time of the quick sort

  You'll learn it in your *Data Structure and Algorithms* course.



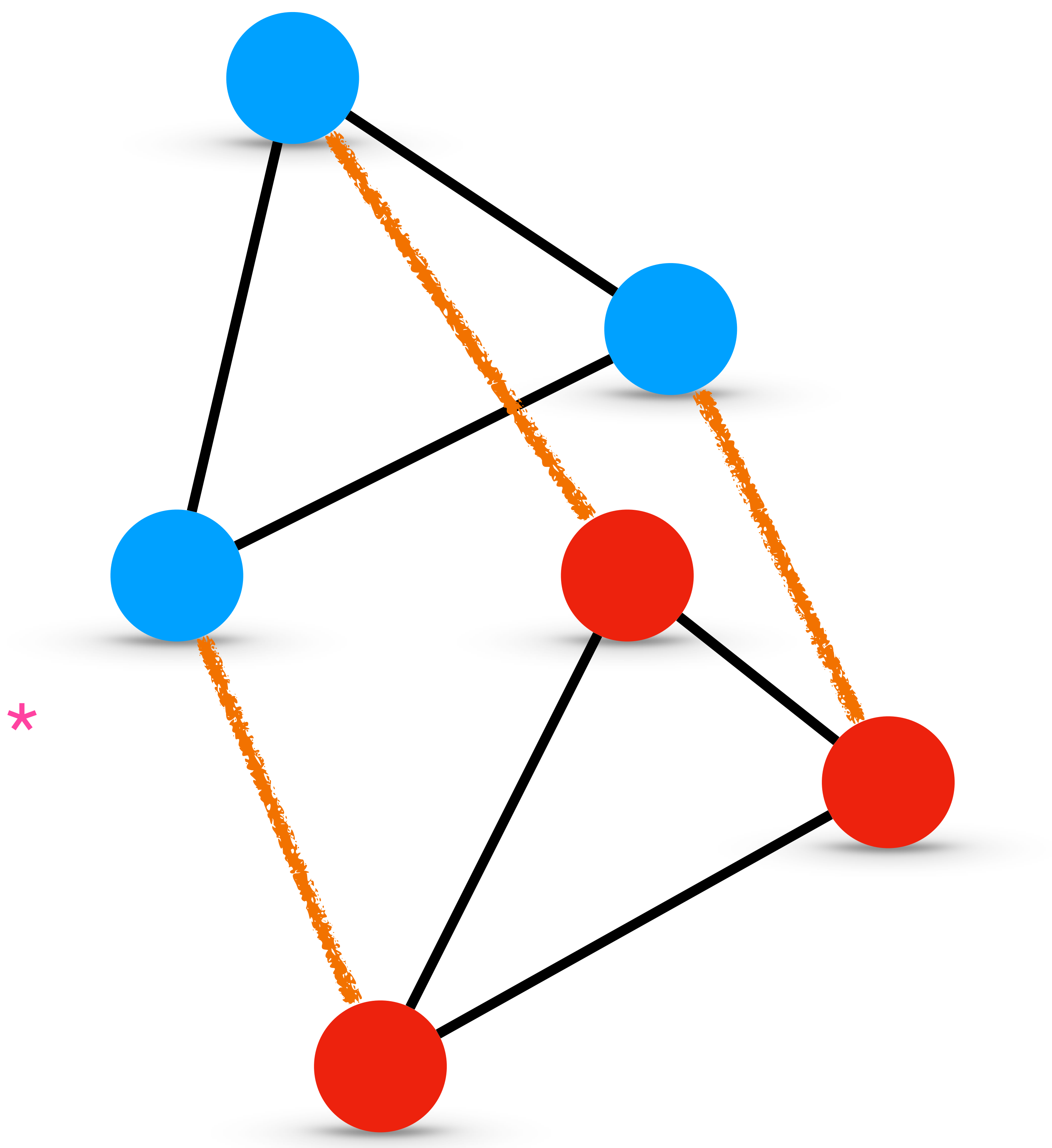
Randomized Algorithms

- Monte Carlo Algorithm
  - Randomized algorithms that may fail or return an incorrect answer
- Las Vegas Algorithm
  - Randomized algorithm that always returns the right answer

* Besides global cuts, there are s-t cuts

Min-Cut

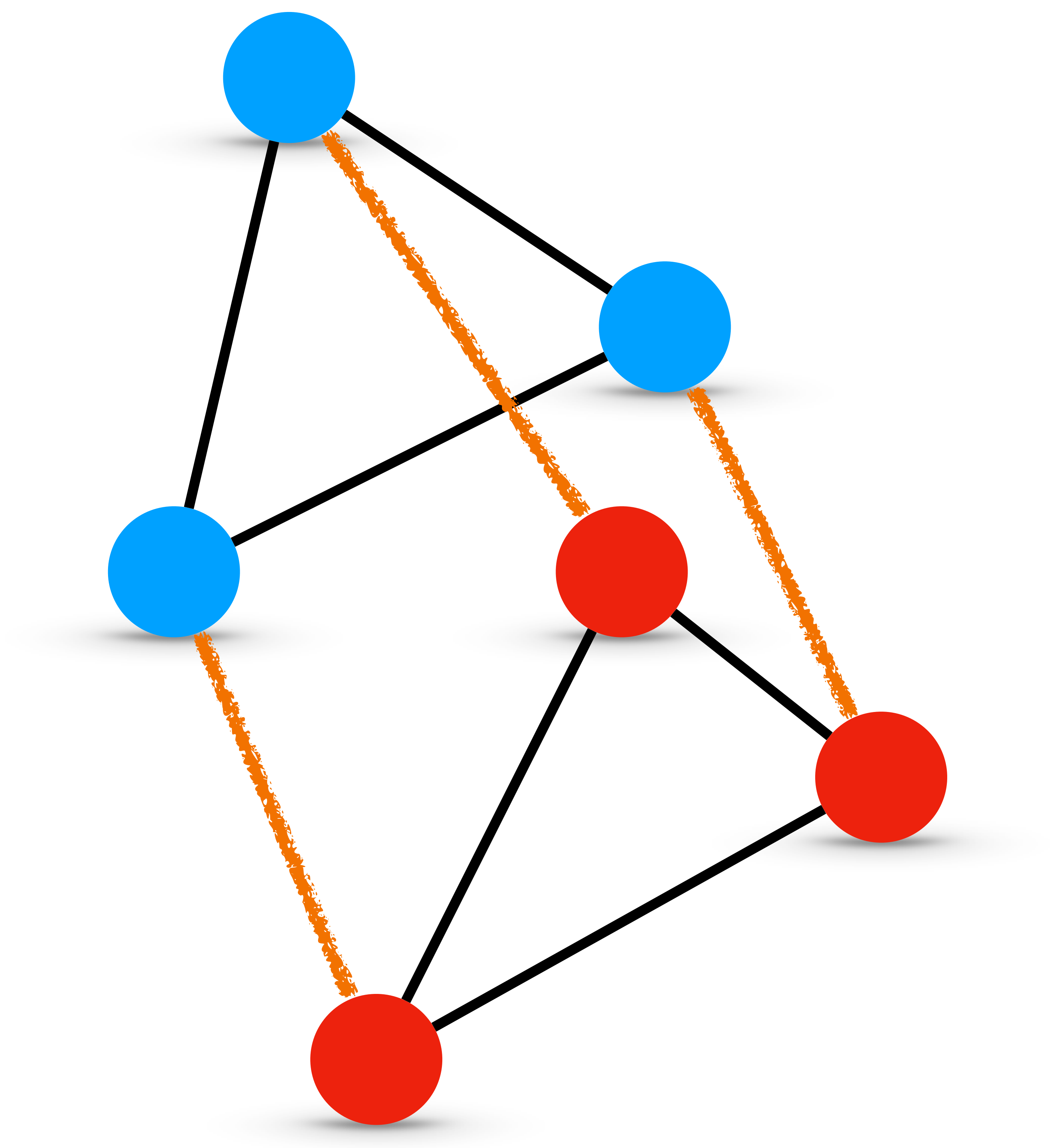
- Undirected graph $G(V, E)$
- Cut: A bi-partition of V into nonempty S and T
  - $C = (S, T)$
- Find a cut set $E(S, T)$ of smallest size (**global** min-cut) *
  - $E(S, T) := \{uv \in E \mid u \in S, v \in T\}$



Adapted from Prof. Yitong Yin's slides

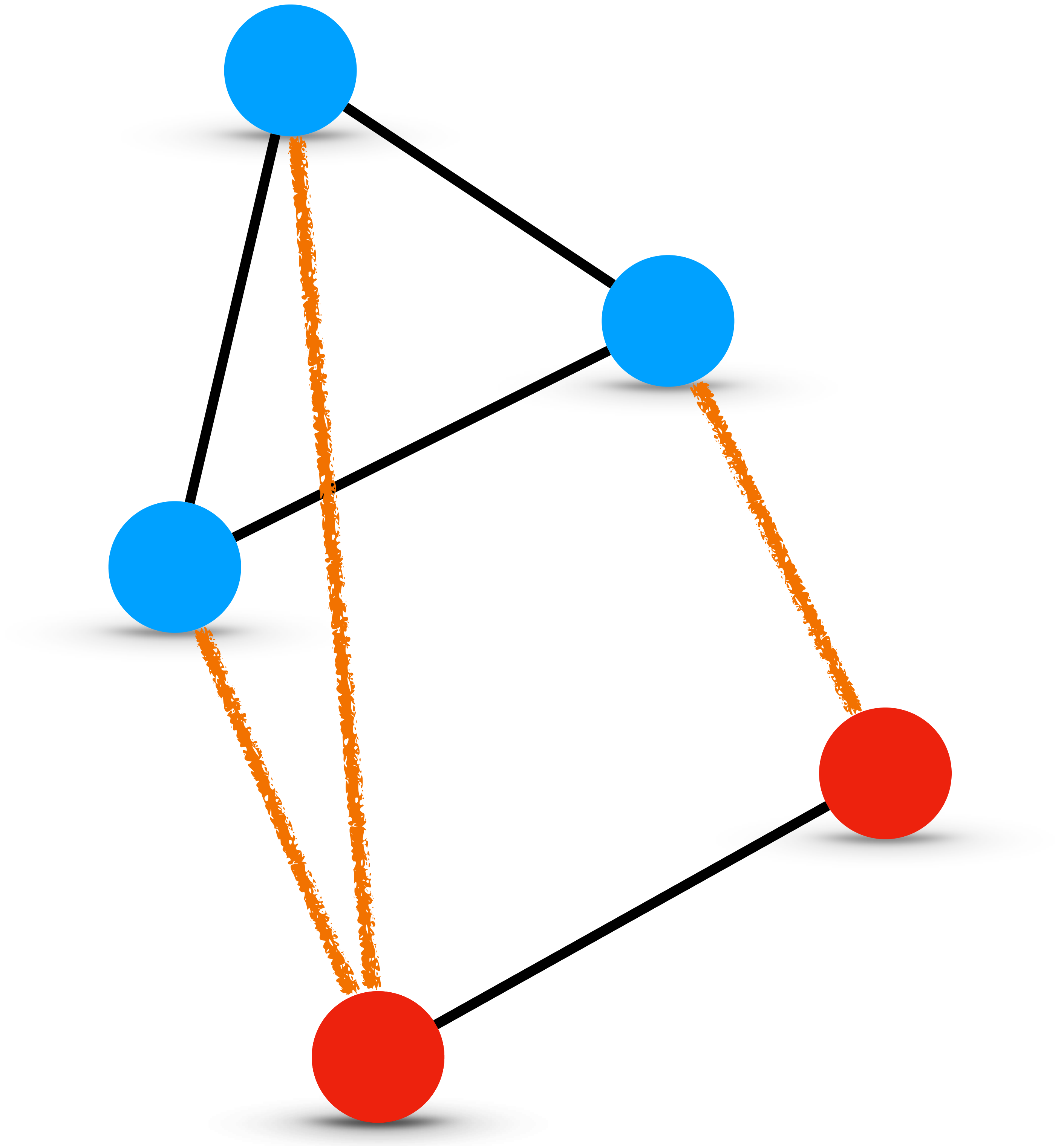
Karger's Algorithm

- contract(e)



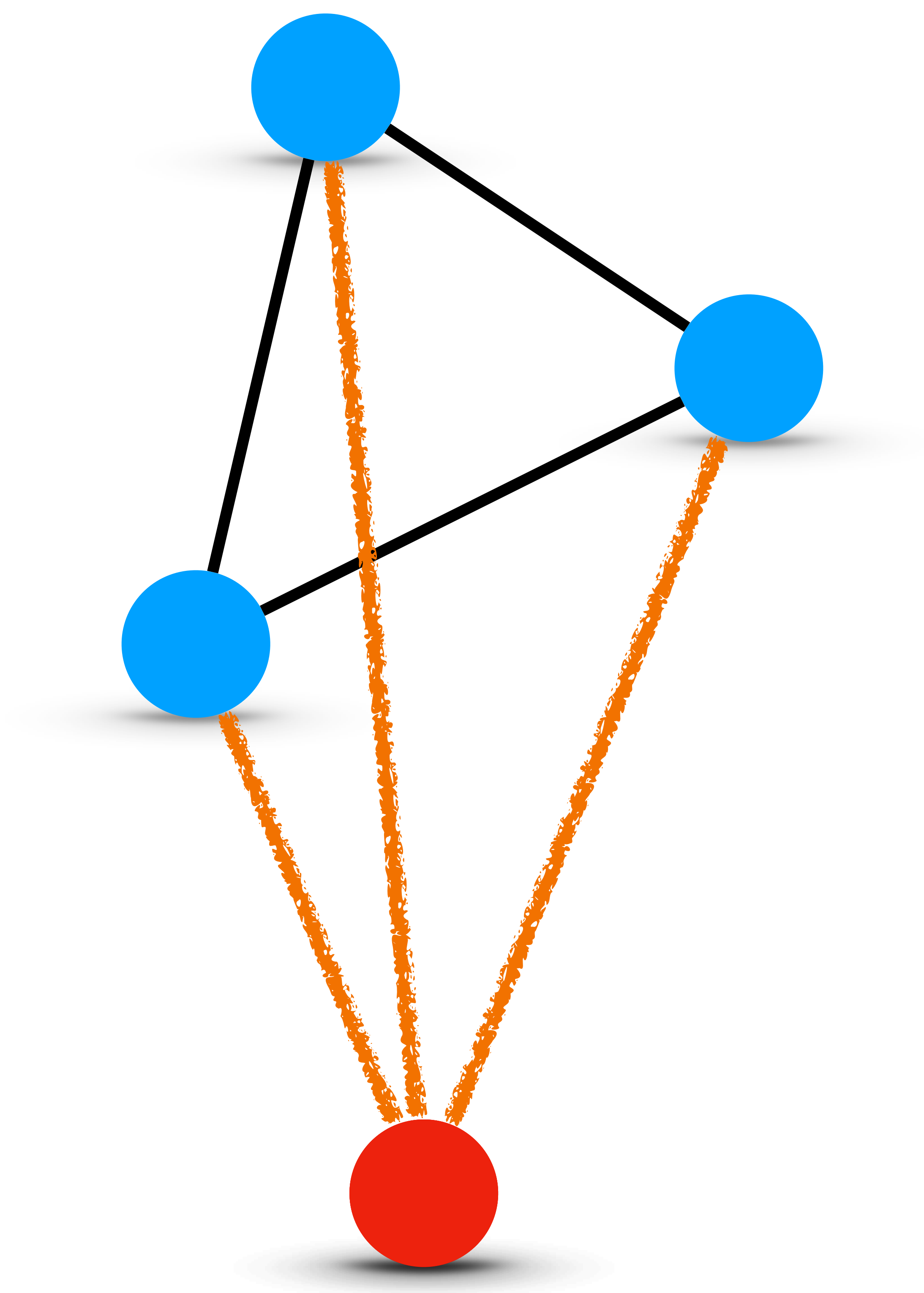
Karger's Algorithm

- contract(e)



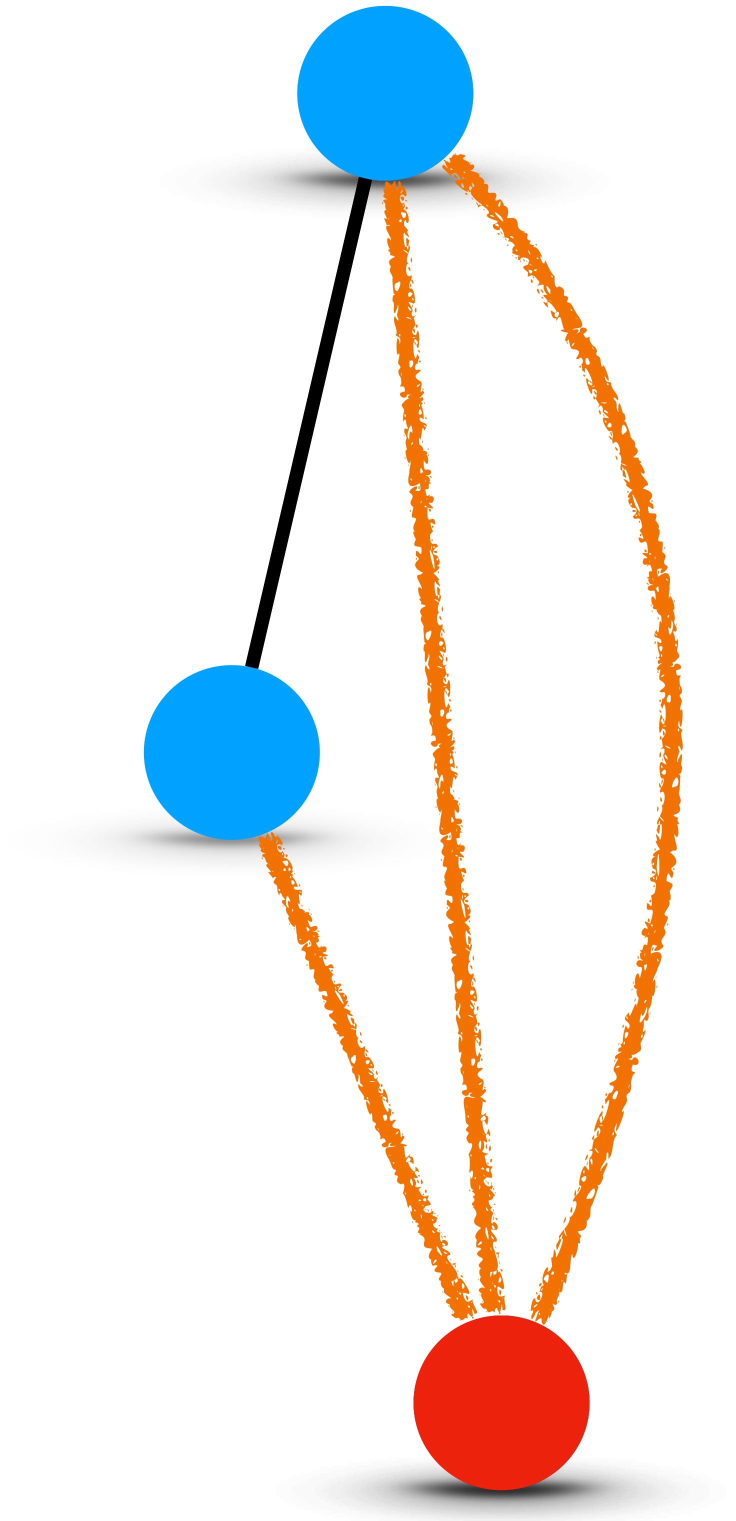
Karger's Algorithm

- contract(e)



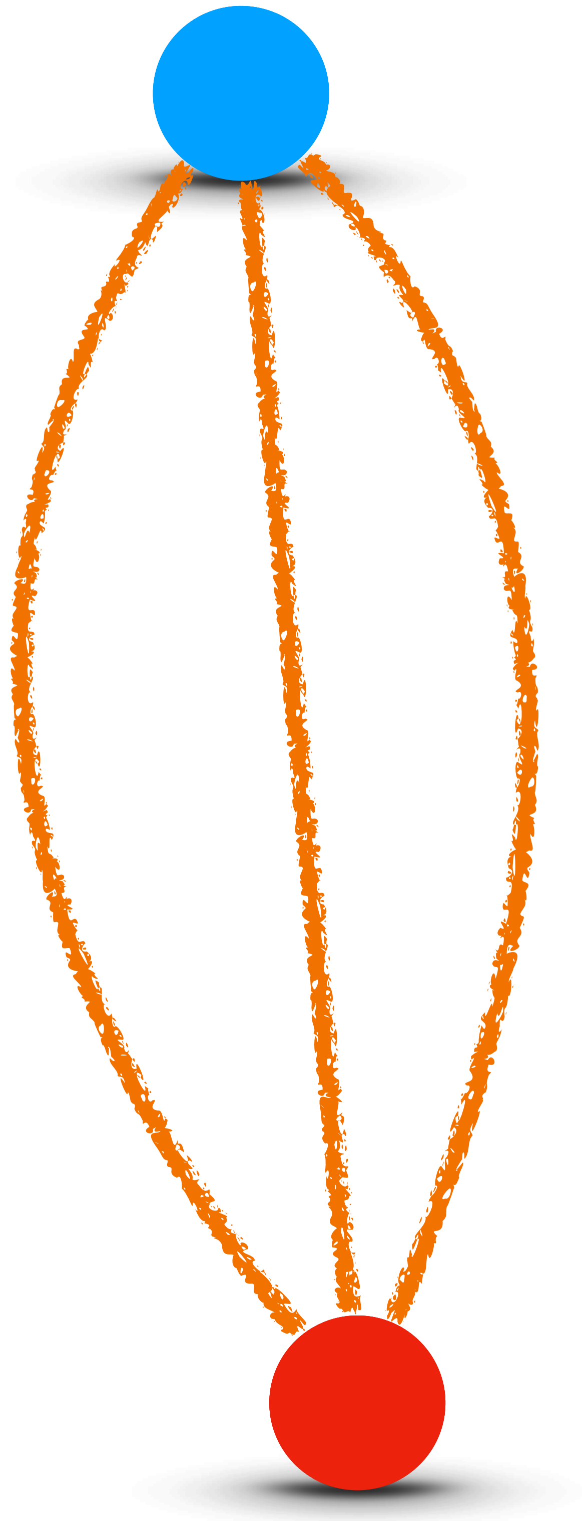
Karger’s Algorithm

- contract(e)



Karger's Algorithm

- contract(e)



Karger's Algorithm

- contract(e)

Karger's Algorithm

while $|V| > 2$ do:

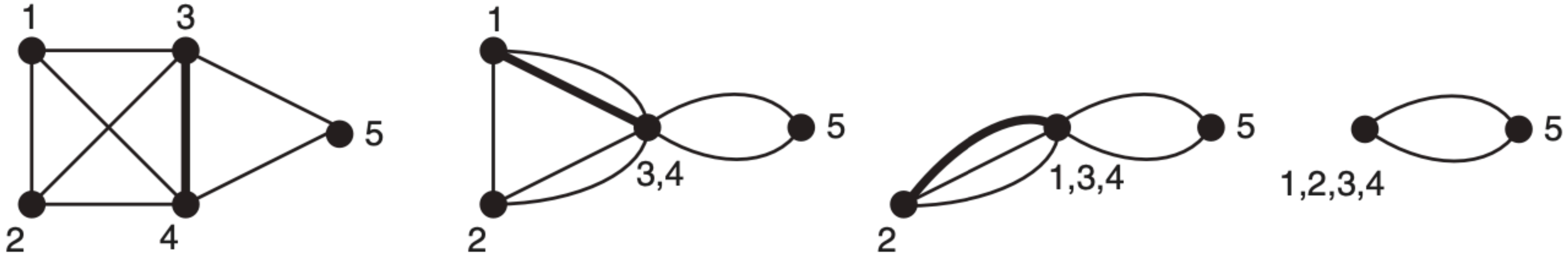
  pick random $e \in E$;

  contract(e);

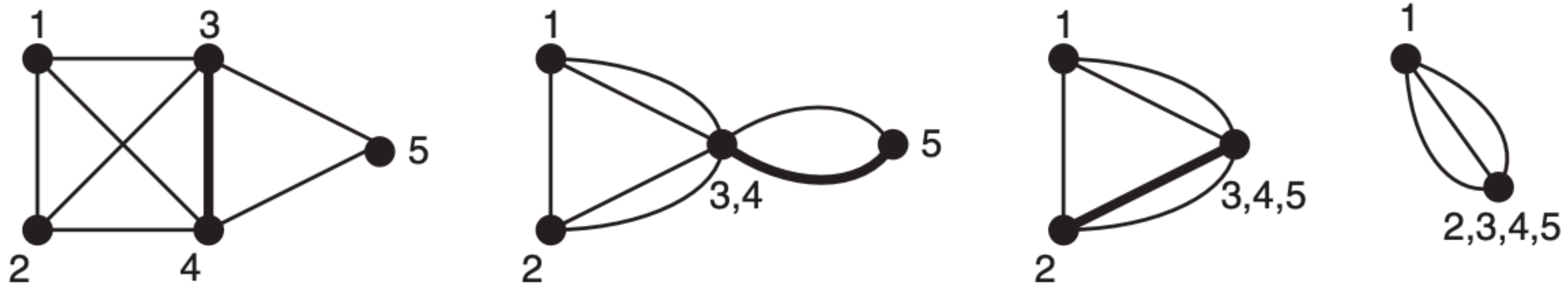
return remaining edges;

Adapted from Prof. Yitong Yin's slides

Karger's Algorithm



(a) A successful run of min-cut.



(b) An unsuccessful run of min-cut.

Karger's Algorithm

Theorem (Karger 1993).

$$\Pr[\text{a min-cut is returned}] \geq \frac{2}{n(n-1)}$$

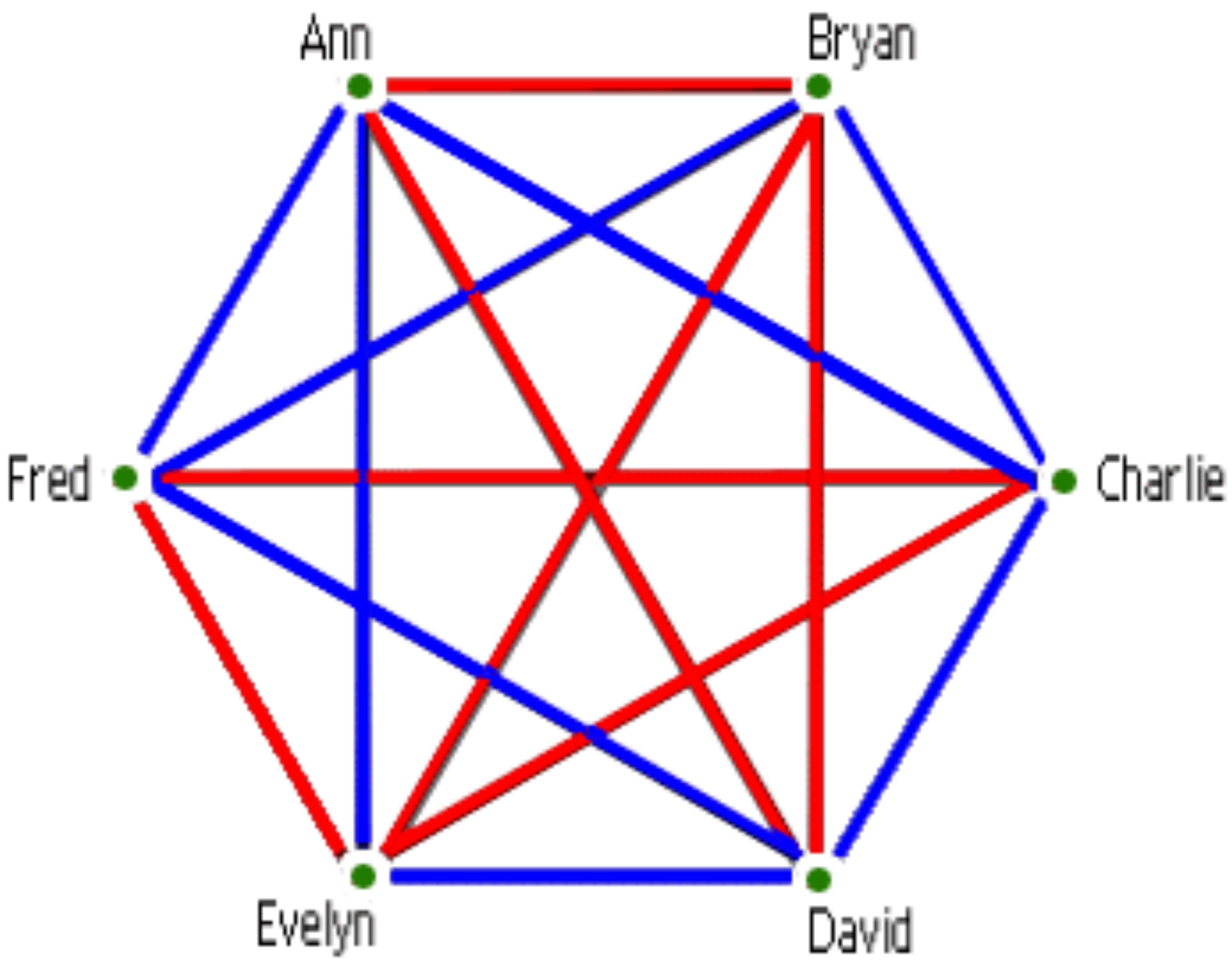
Observation:

- Any cut-set of a graph in an intermediate iteration of the algorithm is also a cut-set of the original graph.
- The output of the algorithm is always a cut-set of the original graph but not necessarily the minimum cardinality cut-set.

The Probabilistic Method

Ramsey Number $R(k, k)$

In any party of six people, either at least three of them are mutual strangers or at least three of them are mutual acquaintances



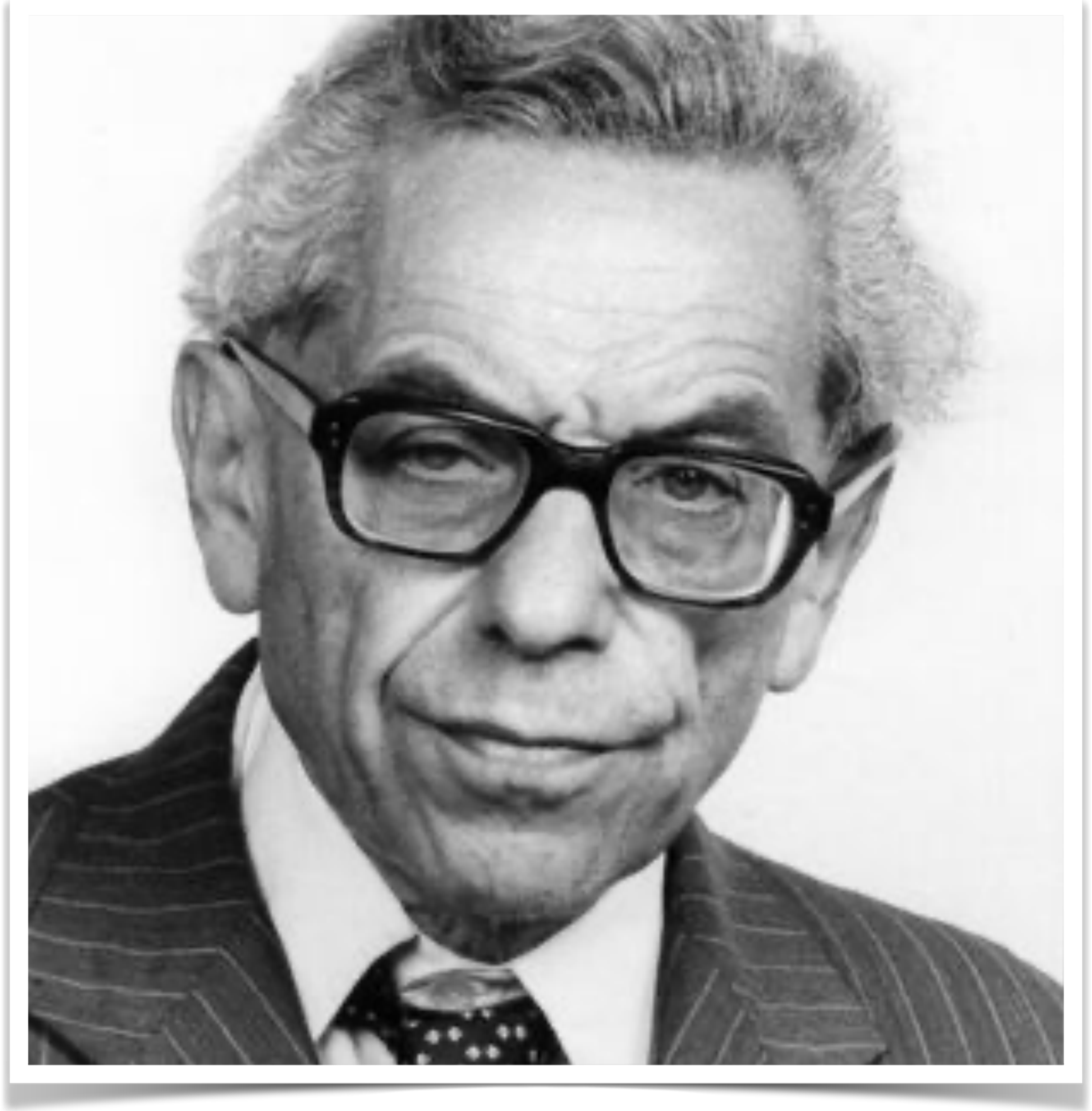
Ramsey Theorem

If $n \geq R(k, k)$, for any edge-2-coloring of K_n, there is a monochromatic K_k

Adapted from Prof. Yitong Yin's slides

Theorem (Erdős 1947)

If $\binom{n}{k} \cdot 2^{1-\binom{k}{2}} < 1$ then it is possible to color the edges of K_n with 2 colors so that there is no monochromatic K_k subgraph.



1913 ~ 1996
Paul Erdős
Erdős Pál
[ˈɛrdøːʃ ˈpaːl]

Probability & Measure Theory

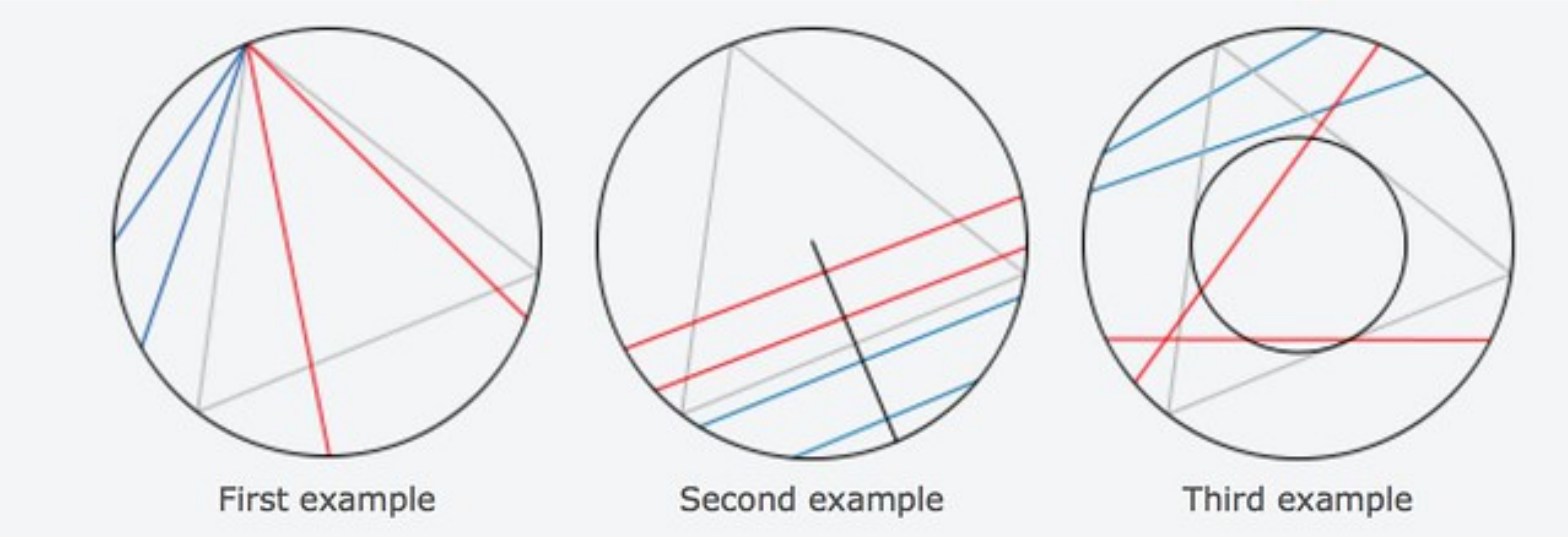
Two Questions

- If I randomly select a number from $\mathbb{N}$, what is the probability that this number is odd? Assume each number has the same chance of being selected.
- If I randomly select a number from $[0,1]$, what is the probability that this number is a rational number? Assume each number has the same chance of being selected.
  - consider Dirichlet function

Bertrand Paradox

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?



Adapted from Prof. Yitong Yin's slides

Measure Theory for Probability Theory

- Axiomatic Foundation of Probability Theory
- σ-algebra / σ-field
- Borel set
- Lebesgue Integral

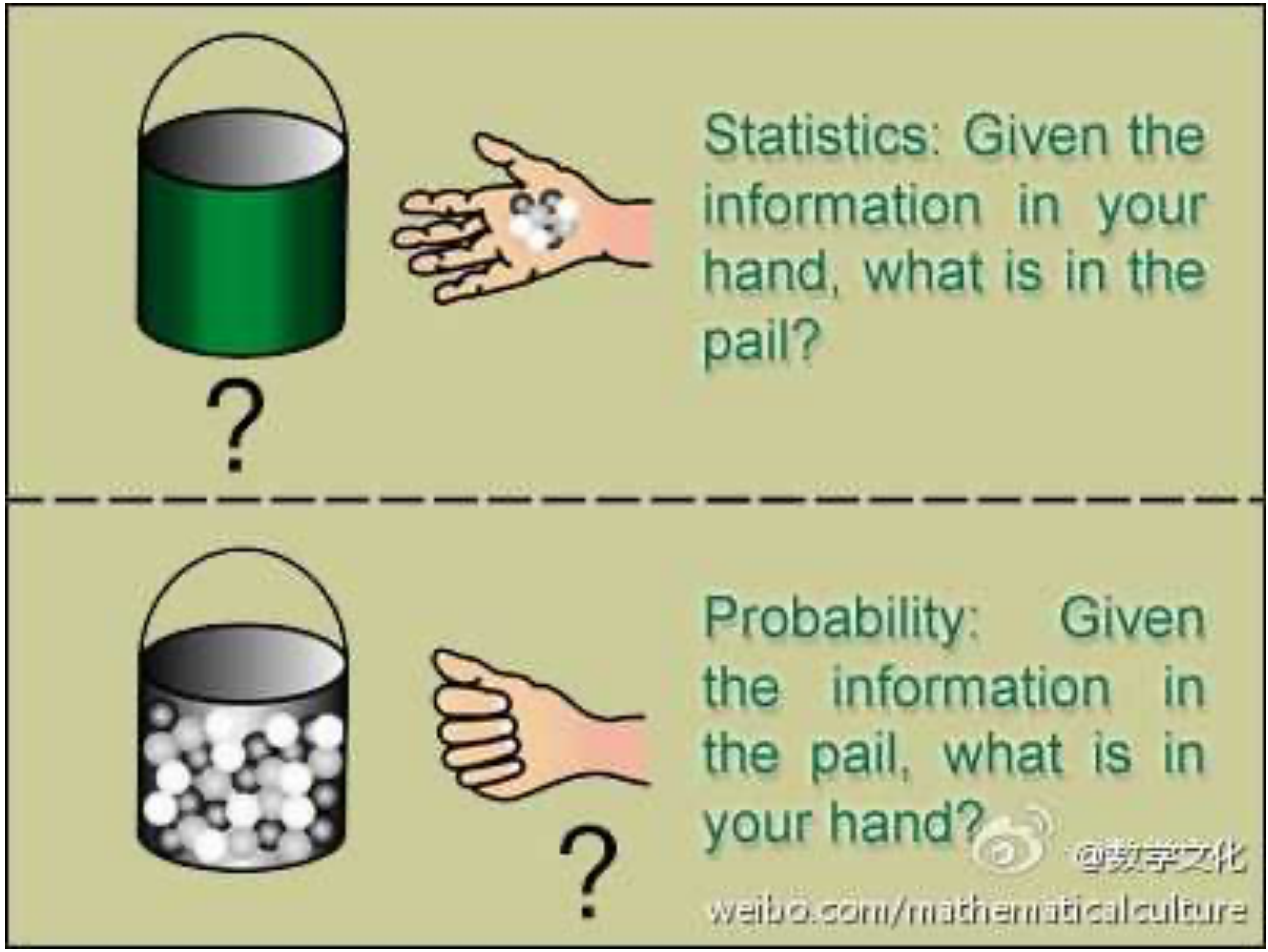
Reading: 以测度为基础的高等概率论和普通的概率论到底有什么区别？ - 知乎 https://www.zhihu.com/question/40145175

Probability & Statistics

Probability & Statistics

The basic problem that we study in **probability** is:
Given a data generating process, what are the properties of the outcomes?

The basic problem of **statistical inference** is the **inverse** of probability:
Given the outcomes, what can we say about the process that generated the data?



Reference: 概率论与统计学的关系是什么？ - bigcloud的回答 - 知乎 https://www.zhihu.com/question/20269390/answer/19994114

Statistics

- Two main statistical methods
  - Descriptive Statistics
    - Mean
    - Deviation (e.g. Variance)
  - Inferential Statistics
    - Parameter Estimation
    - Hypothesis Testing

Benford’s Law

Aka Newcomb–Benford law / first-digit law

- An observation that in many real-life sets of numerical data, the leading digit is likely to be small.
  - 1 appears as the leading significant digit about 30% of the time
  - 9 appears as the leading significant digit less than 5% of the time

Reference: https://en.wikipedia.org/wiki/Benford's_law

Benford's Law

- An observation that in many real-life sets of numerical data, the leading digit is likely to be small.
  - 1 appears as the leading significant digit about 30% of the time
  - 9 appears as the leading significant digit less than 5% of the time

TABLE I

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

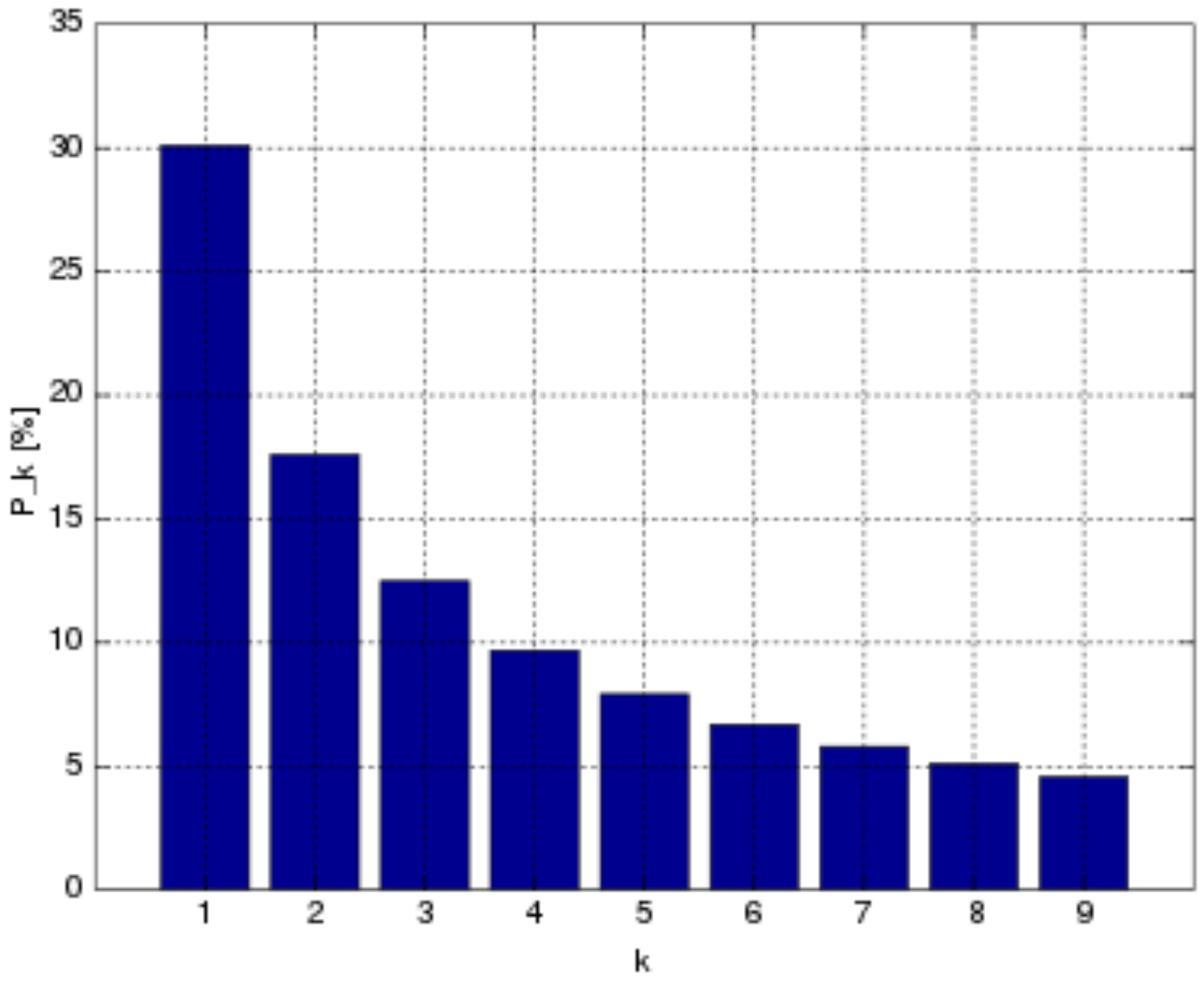
Group	Title	First Digit 1	2	3	4	5	6	7	8	9	Count
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
	Average........	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
	Probable Error	±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

* Hill, Theodore P. (1995). "Base-invariance implies Benford's law". *Proceedings of the American Mathematical Society*. **123** (3): 887–895

Benford’s Law

Aka Newcomb–Benford law / first-digit law

- $\Pr(d) = \log_b(d+1) - \log_b(d)$
- Proved by Ted Hill in 1995 *



d	$P(d)$	Relative size of $P(d)$
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

Reference: https://en.wikipedia.org/wiki/Benford's_law

Detecting Fabricated Data

Financial Fraud of Kevin Lawrence

Possibly the **biggest financial fraud** in Washington State's history

- Kevin Lawrence claimed that his startup would be an industry innovator that integrated fitness and health care into one business model.
- Flush with investor money, Lawrence floated two companies – Znetix Inc and Health Maintenance Centers Inc.
- In reality, there was no evidence that Znetix/HMC could make the business operation pay for itself.

Lawrence and his pals tried to cover their tracks by moving investors' money through a complex web of bank accounts and shell companies to give the appearance of a bustling and growing business.

Source: https://www.safalniveshak.com/latticework-of-mental-models-benfords-law/
https://dfi.wa.gov/news/press/judge-sentences-kevin-l-lawrence-20-years-prison-znetixhmc-stock-scam

Detecting Fabricated Data

Financial Fraud of Kevin Lawrence

Lawrence bought several properties including

- a home in Hawaii.
- twenty personal watercrafts (including a 22-foot Bombardier speedboat),
- forty-seven luxury cars (five Hummers, four Ferraris, two DeThomaso Panteras,
- three Dodge Vipers, two Cadillac Escalades, a Lamborghini Diablo),
- Rolex watches, expensive diamond jewelry for his girlfriend(s) and a $200,000 Samurai sword.

Source: https://www.safalniveshak.com/latticework-of-mental-models-benfords-law/
https://dfi.wa.gov/news/press/judge-sentences-kevin-l-lawrence-20-years-prison-znetixhmc-stock-scam

Detecting Fabricated Data

Financial Fraud of Kevin Lawrence

- Darrell Dorrell, a suspicious forensic accountant
  - compiled a list of over 70,000 numbers representing their various checks and wire transfers
  - compared the distribution of digits with **Benford's law**.
- On 25 November 2003, Kevin Lawrence was sentenced to 20 years in prison

Source: https://www.safalniveshak.com/latticework-of-mental-models-benfords-law/
https://dfi.wa.gov/news/press/judge-sentences-kevin-l-lawrence-20-years-prison-znetixhmc-stock-scam

Detecting Fabricated Data

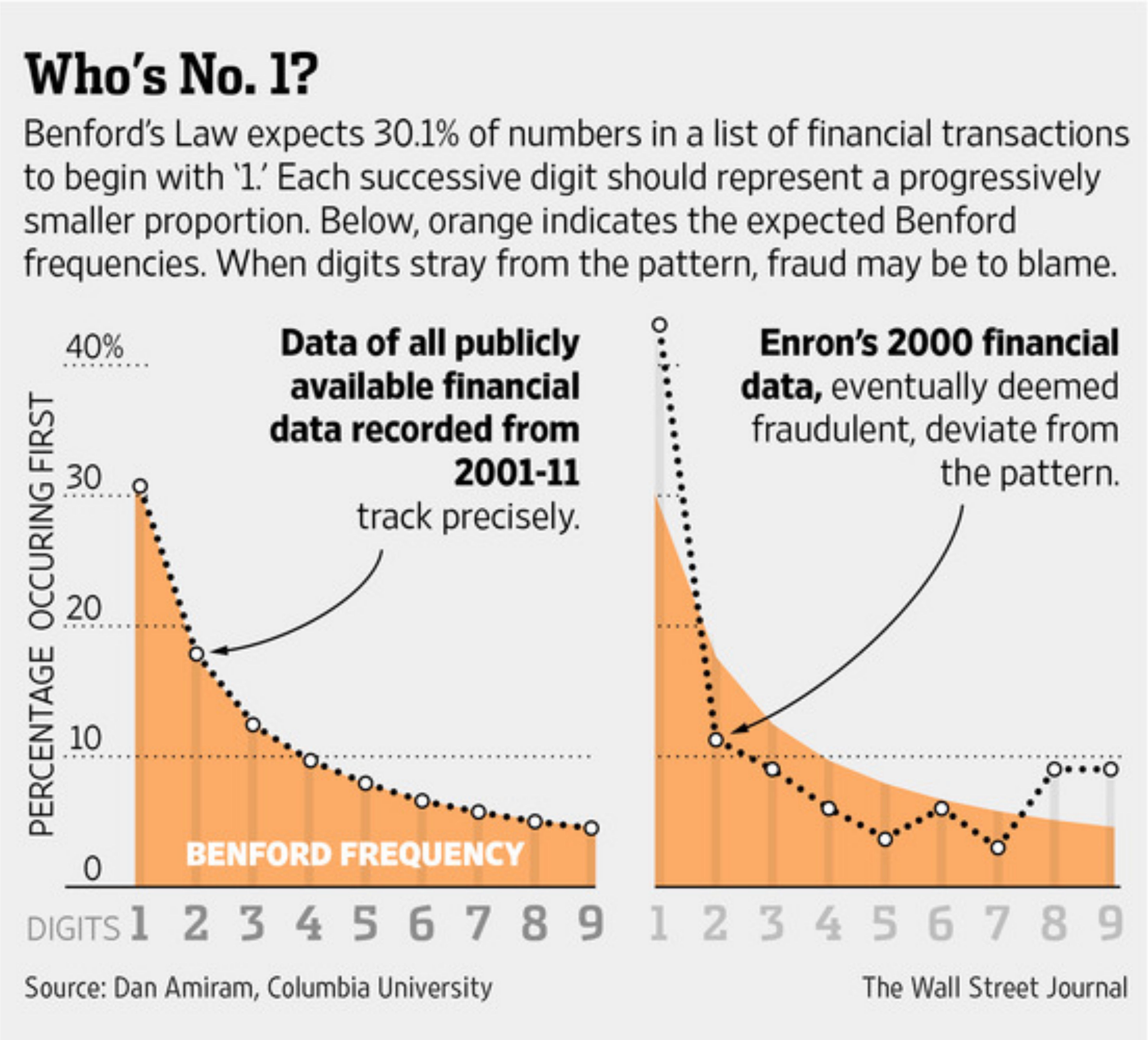
Enron scandal

- Enron Corporation, an American energy company based in Houston, Texas
- In October 2001, the company declared bankruptcy.
- In addition to being the largest bankruptcy reorganization in U.S. history at that time, Enron was cited as the biggest audit failure.

Reference: https://en.wikipedia.org/wiki/Enron_scandal 【突破审计陷阱】安然公司财务造假案例之介绍

Detecting Fabricated Data

Enron scandal



From The Wall Street Journal

Reference: https://en.wikipedia.org/wiki/Enron_scandal 【突破审计陷阱】安然公司财务造假案例之介绍

The 1970 draft lottery

- In 1970, during the Vietnam War
- The American army used a lottery system based on birth dates to determine who would be called up for service in the military forces.
- Draftees were called up for service based on the draft number assigned to their dates of birth.
- Those receiving low draft numbers were called up first.

Is it Fair?

Table 3.3. *Draft numbers assigned by lottery.*

day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1	305	086	108	032	330	249	093	111	225	359	019	129
2	159	144	029	271	298	228	350	045	161	125	034	328
3	251	297	267	083	040	301	115	261	049	244	348	157
4	215	210	275	081	276	020	279	145	232	202	266	165
5	101	214	293	269	364	028	188	054	082	024	310	056
6	224	347	139	253	155	110	327	114	006	087	076	010
7	306	091	122	147	035	085	050	168	008	234	051	012
8	199	181	213	312	321	366	013	048	184	283	097	105
9	194	338	317	219	197	335	277	106	263	342	080	043
10	325	216	323	218	065	206	284	021	071	220	282	041
11	329	150	136	014	037	134	248	324	158	237	046	039
12	221	068	300	346	133	272	015	142	242	072	066	314
13	318	152	259	124	295	069	042	307	175	138	126	163
14	238	004	354	231	178	356	331	198	001	294	127	026
15	017	089	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	055	274	120	044	207	254	107	096
17	235	189	033	260	112	073	098	154	255	288	143	304
18	140	292	332	090	278	341	190	141	246	005	146	128
19	058	025	200	336	075	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	063	192	185	135
21	186	363	334	062	250	060	027	291	204	243	156	070
22	337	290	265	316	326	247	153	339	160	117	009	053
23	118	057	256	252	319	109	172	116	119	201	182	162
24	059	236	258	002	031	358	023	036	195	196	230	095
25	052	179	343	351	361	137	067	286	149	176	132	084
26	092	365	170	340	357	022	303	245	018	007	309	173
27	355	205	268	074	296	064	289	352	233	264	047	078
28	077	299	223	262	308	222	088	167	257	094	281	123
29	349	285	362	191	226	353	270	061	151	229	099	016
30	164		217	208	103	209	287	333	315	038	174	003
31	211		030		313		193	011		079		100

The 1970 draft lottery

Table 3.3. *Draft numbers assigned by lottery.*

day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1	305	086	108	032	330	249	093	111	225	359	019	129
2	159	144	029	271	298	228	350	045	161	125	034	328
3	251	297	267	083	040	301	115	261	049	244	348	157
4	215	210	275	081	276	020	279	145	232	202	266	165
5	101	214	293	269	364	028	188	054	082	024	310	056
6	224	347	139	253	155	110	327	114	006	087	076	010
7	306	091	122	147	035	085	050	168	008	234	051	012
8	199	181	213	312	321	366	013	048	184	283	097	105
9	194	338	317	219	197	335	277	106	263	342	080	043
10	325	216	323	218	065	206	284	021	071	220	282	041
11	329	150	136	014	037	134	248	324	158	237	046	039
12	221	068	300	346	133	272	015	142	242	072	066	314
13	318	152	259	124	295	069	042	307	175	138	126	163
14	238	004	354	231	178	356	331	198	001	294	127	026
15	017	089	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	055	274	120	044	207	254	107	096
17	235	189	033	260	112	073	098	154	255	288	143	304
18	140	292	332	090	278	341	190	141	246	005	146	128
19	058	025	200	336	075	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	063	192	185	135
21	186	363	334	062	250	060	027	291	204	243	156	070
22	337	290	265	316	326	247	153	339	160	117	009	053
23	118	057	256	252	319	109	172	116	119	201	182	162
24	059	236	258	002	031	358	023	036	195	196	230	095
25	052	179	343	351	361	137	067	286	149	176	132	084
26	092	365	170	340	357	022	303	245	018	007	309	173
27	355	205	268	074	296	064	289	352	233	264	047	078
28	077	299	223	262	308	222	088	167	257	094	281	123
29	349	285	362	191	226	353	270	061	151	229	099	016
30	164		217	208	103	209	287	333	315	038	174	003
31	211		030		313		193	011		079		100

Table 3.4. *Average draft number per month.*

January	201.2	July	181.5
February	203.0	August	173.5
March	225.8	September	157.3
April	203.7	October	182.5
May	208.0	November	148.7
June	195.7	December	121.5

Is it a Coincidence?

Hypothesis Testing

- Let's start out with the hypothesis that the lottery was fair.
- If we can show that the outcomes are extremely improbable under the hypothesis
- We can reject our hypothesis and conclude that the lottery was most probably unfair.

Building the Model

- The expected value of the average draft number for a given month is 183.5 for each month.
- G_i: the average draft number for month i
- We want to know

$$\Pr\left(\sum_{i=1}^{12} |G_i - 183.5| \geq 272.4\right)$$

Monte Carlo Method

$$\Pr\left(\sum_{i=1}^{12} |G_i - 183.5| \geq 272.4\right)$$

- Deriving a versatile mathematical formula for this probability seems like an endless task.
- In a **Monte Carlo** study with 100,000 simulation runs, we came out with a simulated value of 0.012 for the probability in question.

Another Way to Test

Table 3.5. *Index numbers for the 1970 draft lottery.*

month	1	2	3	4	5	6	7	8	9	10	11	12
index	5	4	1	3	2	6	8	9	10	7	11	12

- Under the hypothesis that the lottery is fair, this permutation would have to be a "random" permutation.
- For a random permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_{12})$ of the numbers $1,\ldots,12$,
- We define the distance measure $d(\sigma)$ by $d(\sigma) = \sum_{i=1}^{12} |\sigma_i - i|$
- It holds that $0 \leq d(\sigma) \leq 72$
- $d(\sigma^*) = 18$, for σ^* from Table 3.5

Monte Carlo Again

- Now we want to know

$$\Pr(d(\sigma) \leq 18)$$

- A Monte Carlo study with 100,000 generated random permutations led us to an estimate of 0.0009 for our sought-after probability.

- This is strong evidence that the 1970 draft lottery did not proceed fairly.