Computational Methods for Linguists

Ling 471

Olga Zamaraeva (Instructor)
Yuanhe Tian (TA)
04/29/21

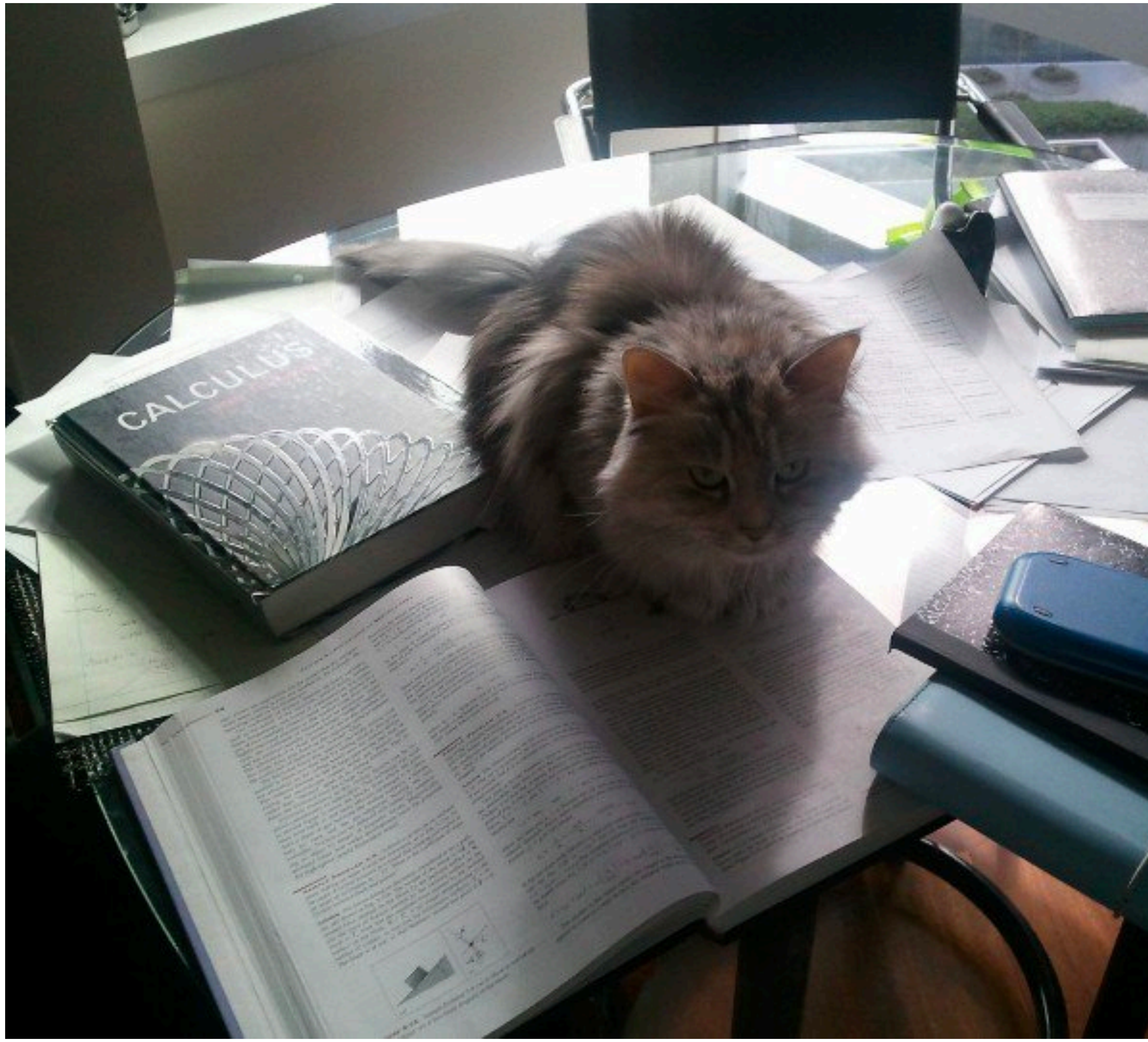
Reminders

- Make sure to create **private** repository copies for the assignments
  - I was **mistaken** to think forks were private
  - You can **delete** your forked copies if you like
- Can clone and copy manually, or use **import**
- Whichever way you choose, **please do not publish any solutions to the HW anywhere**
- Assignment 3:
  - a “short” description **and** a detailed **walkthrough** available
  - ...is **harder** than Assignment 2
- Please fill out Midterm Course Evaluations!

Questions?

Plan for today

- Data science and probability:
  - what's the **connection**?
- Probability theory **basics**
- Statistics: **distributions** and **estimation**
  - time-permitting
- Some of today's and next week's material may be **dense**
  - Goal: Learn **something** about those things
  - Remember, no exams :)
  - Unlikely to ask you to compute something terrible in HW
  - If this is the **first** time you hear about these things:
    - You will understand them better **next** time you hear about them



Probability and Statistics

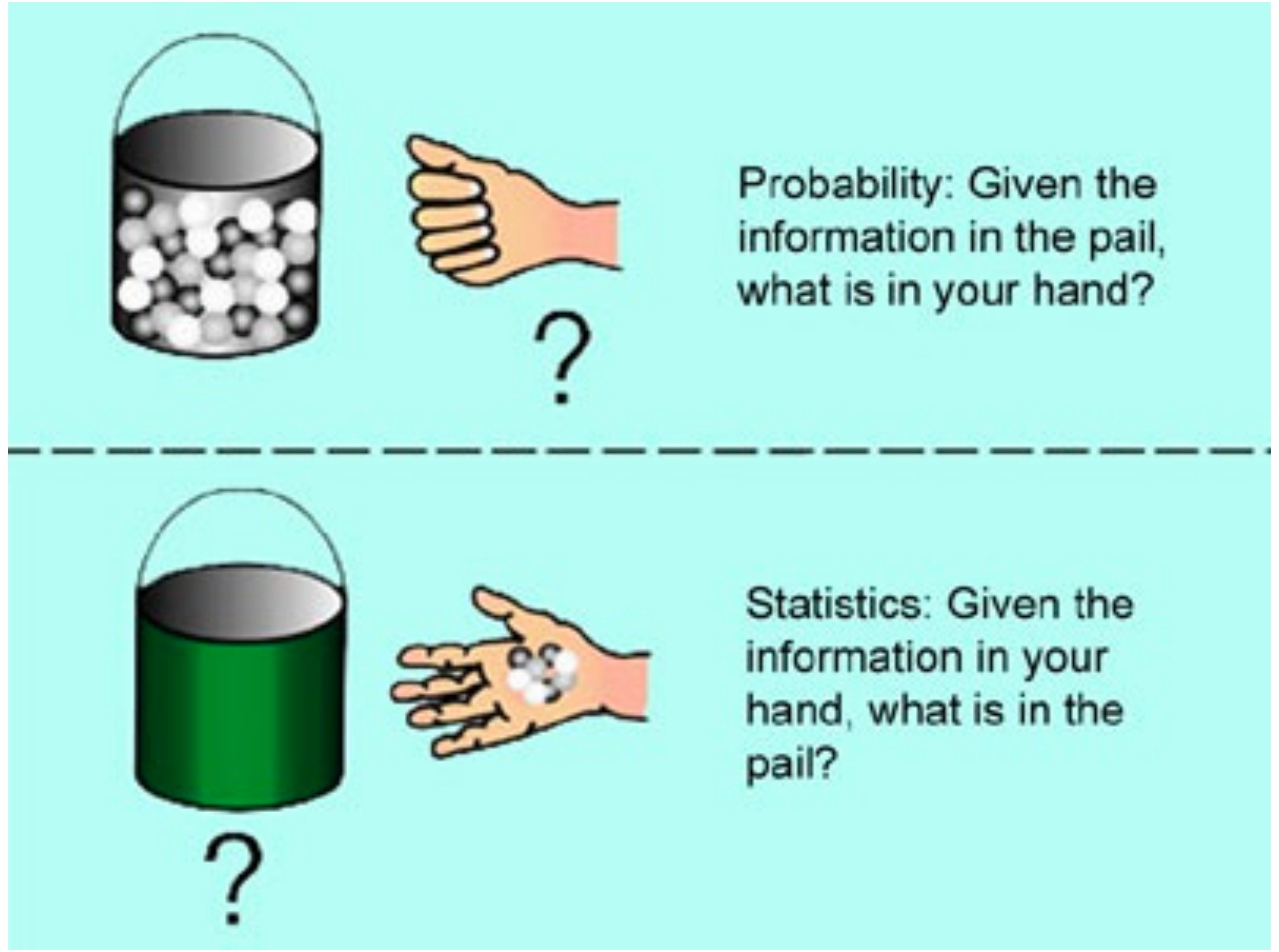
Data science and statistics

- There is a lot of **randomness** and **uncertainty** in the world
- Many processes in our lives are **data-generating**
  - how many times we click on what
  - how many messages we send/receive, of what kind
  - what places we visit and how often
  - etc., etc., etc.
- Statistics:
  - A **science** of making sense of the world by **sampling** data
    - What is true for the sample, is also true for the population
      - ...if the sample is **random** and sufficiently **large**

https://www.scribbr.com/methodology/population-vs-sample/

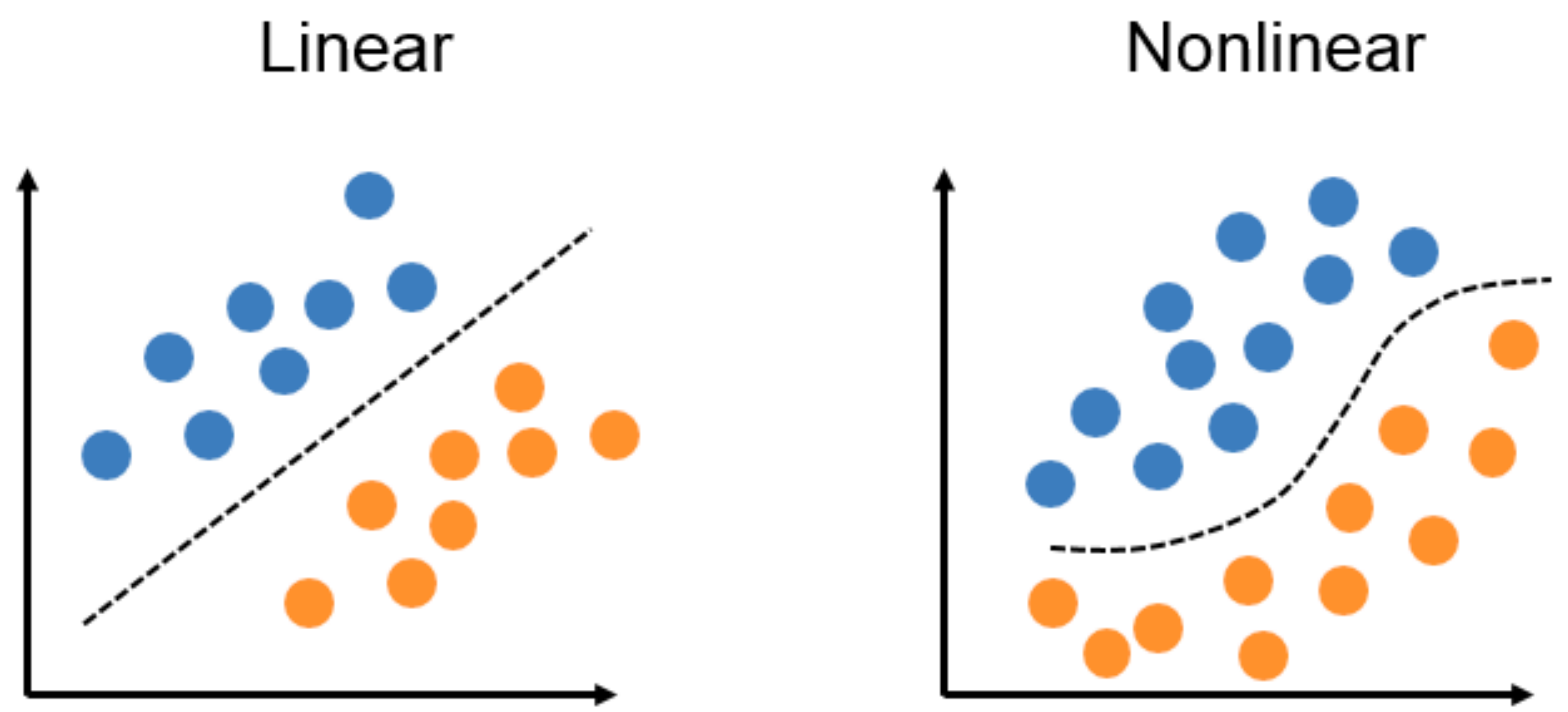
Statistics and Probability Theory
(and Data Science)

- Probability Theory:
  - Formally estimate how likely an **outcome** is
  - Informally: Oriented at predicting **future** events
    - Given what I know about the population, what sample could I draw?
  - Relies on the notion of probability **distribution**
    - How are probabilities of **all** possible outcomes **distributed**?
- Statistics:
  - Use **probability distributions** to make sense of large data **formally**
  - Informally: Oriented at analyzing **past** events
    - Given the samples which I drew, what can I say about the population?
  - No distribution => no statistics!
- Data Science:
  - Probability + Statistics
  - Analyze past events **and** predict future events, **at scale**, in real world



https://www.quora.com/What-is-the-difference-between-probability-and-statistics

Prediction and Probabilities
classification problem



https://jtsulliv.github.io/perceptron/

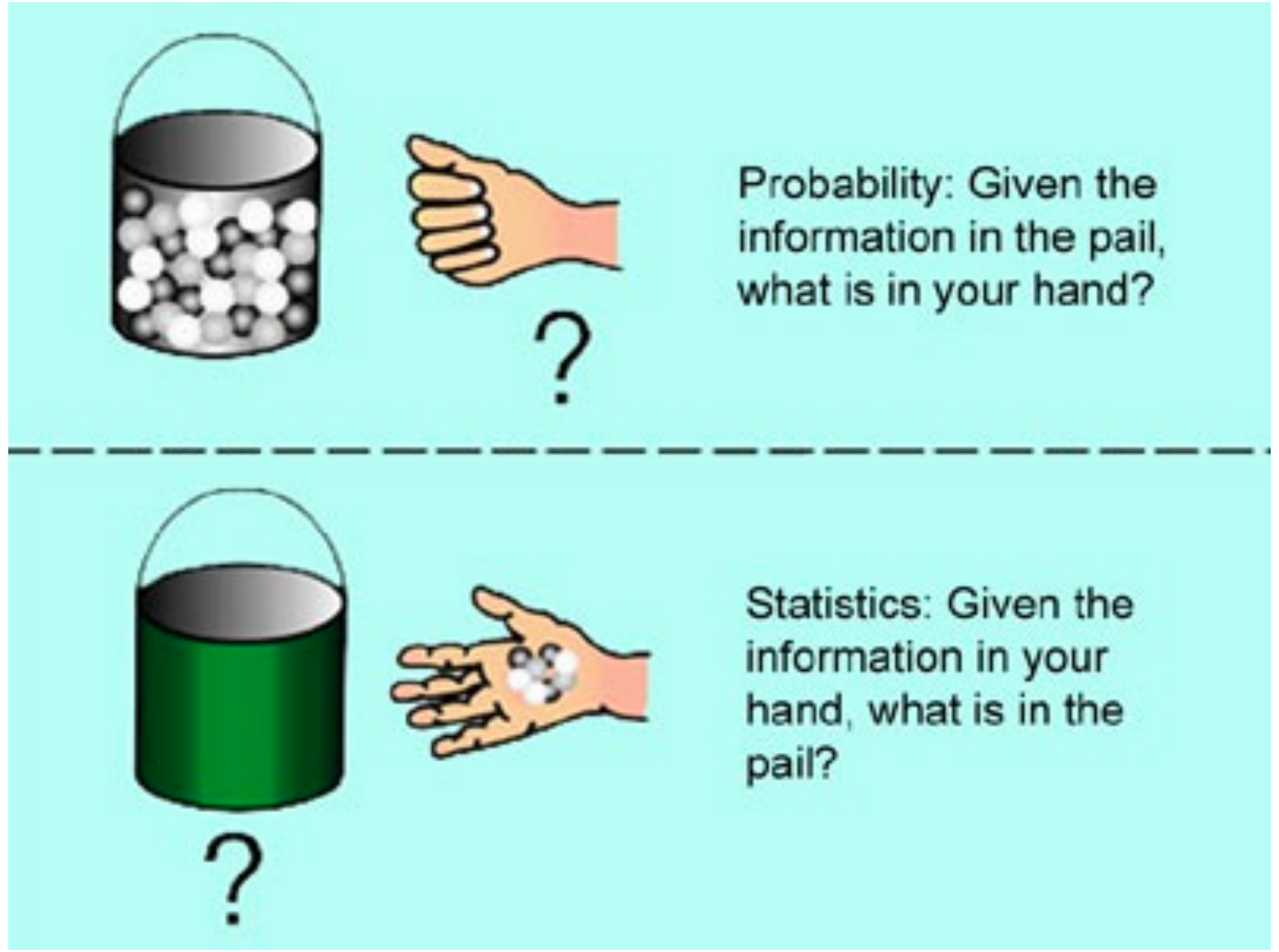
- Predictions in data science and ML need to be quantified
- To predict whether a review is POS or NEG:
  - e.g. compute the **probability** of it being POS
  - predict POS if that probability is **high**
  - predict NEG **otherwise**
- **Conditional** probability: P(Y|X)
  - where Y is the label and X is the observation
    - e.g. Y = POS and X = “this is a good movie!”
  - **How** to learn P(Y|X)?
    - There are mathematical functions which you can use
    - A bit more in our ML-dedicated lectures later…

Probability Theory

- …is notoriously unintuitive and hard
- Our **goal**:
    - get familiar with a **subset** of basic concepts
        - not necessarily in the most formal and exhaustive way
    - …such that we can experiment with some data science models in assignments 4—5

Probability Theory
our goals for this lesson

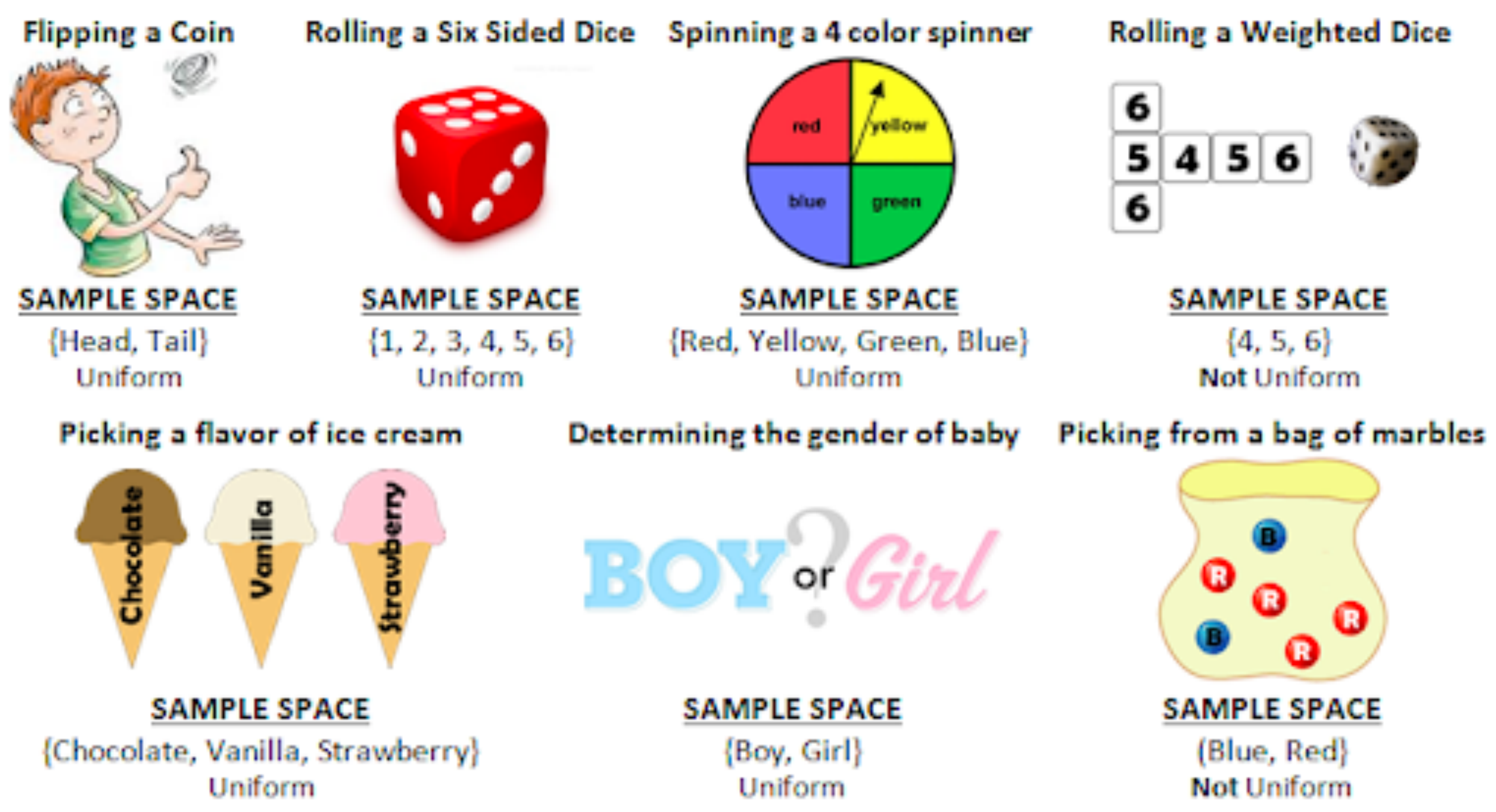
- Definitions:
  - events, outcomes, sample space, random variable
- Mutually exclusive events
- Sequences and independent events
- Joint probabilities
- Conditional probability
- Marginalizing joint probabilities
- Bonus: Maximum Likelihood Estimation



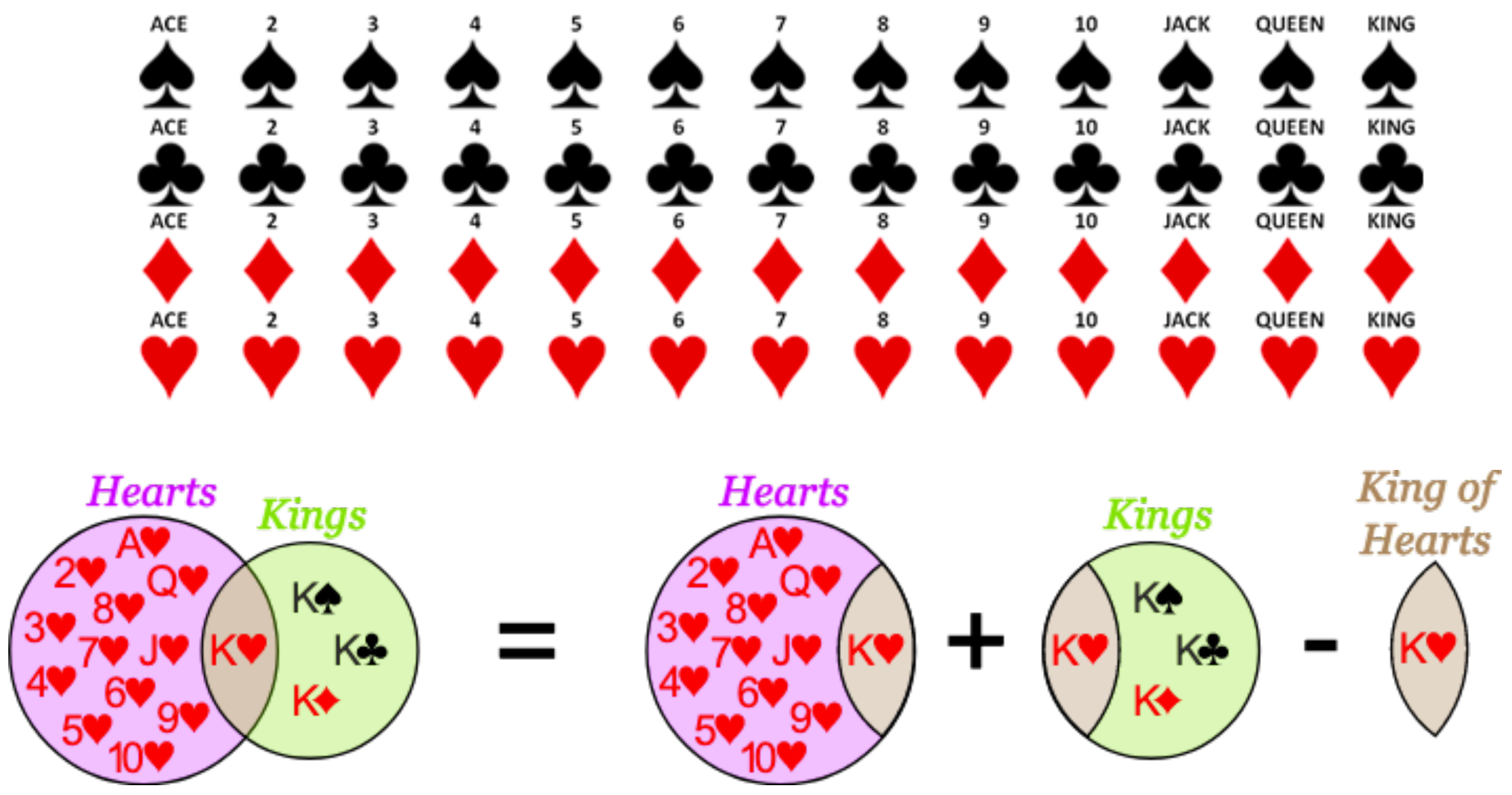
https://www.quora.com/What-is-the-difference-between-probability-and-statistics

Probability
basic intuitions

- How likely is something to happen?
  - Well, we don't know!
  - But, we can **estimate**
    - based on **prior observations** or base on what we **assume** about the situation
- **Out of n** experiments (the "sample space"), **how many** resulted in a **specific** outcome?
  - this ratio is the **probability** of that specific outcome
    - turns out, you can show formally that it **is** the ratio (MLE)
  - Understanding what the "sample space" is **exactly** is **crucial**
  - The probability will be different based on what the sample space actually is
  - Often times, need to subtract things from what intuitively seems like it's the sample space
    - particularly **conditional** probabilities
  - That's the main reason why probability is often unintuitive



http://www.geometrycommoncore.com/content/unit6/gcp1/studentsnotes1.html

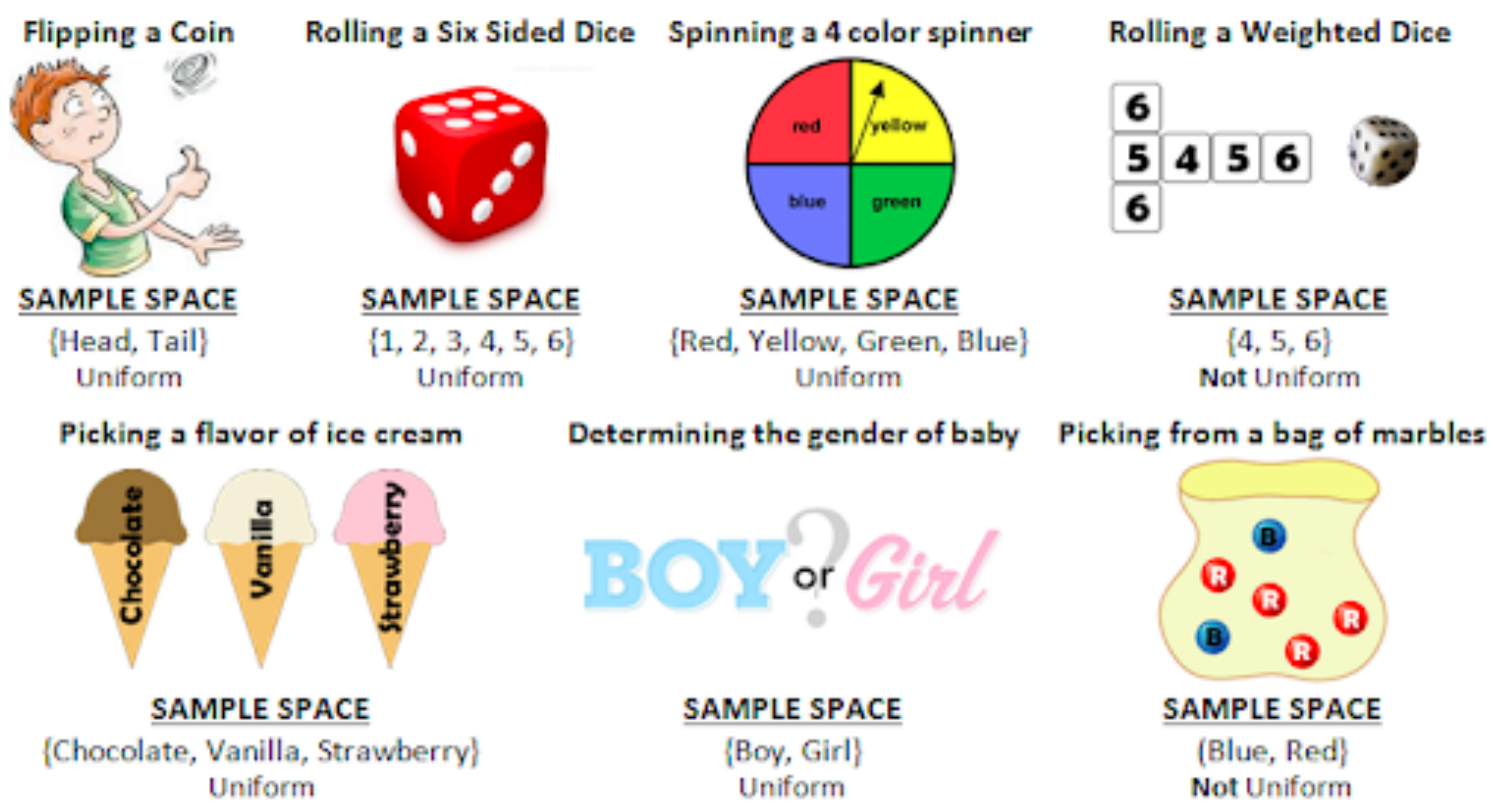


https://www.mathsisfun.com/data/probability-events-mutually-exclusive.html

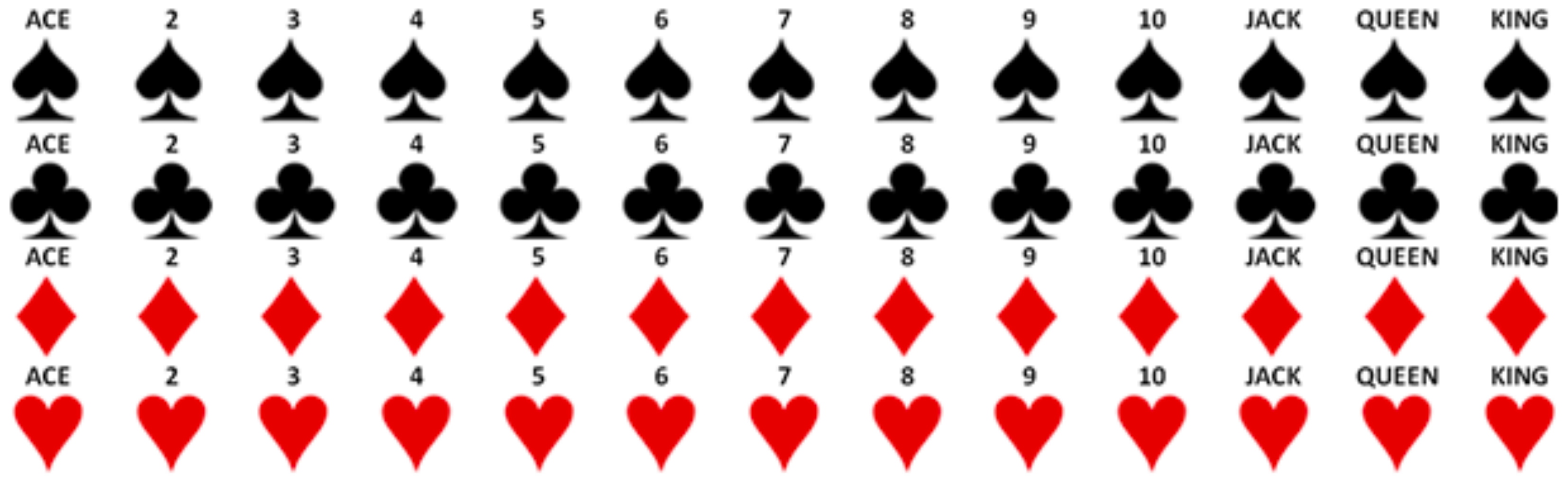
Coin toss

the classic probability example

- Sample space
- Experiment
- Outcome



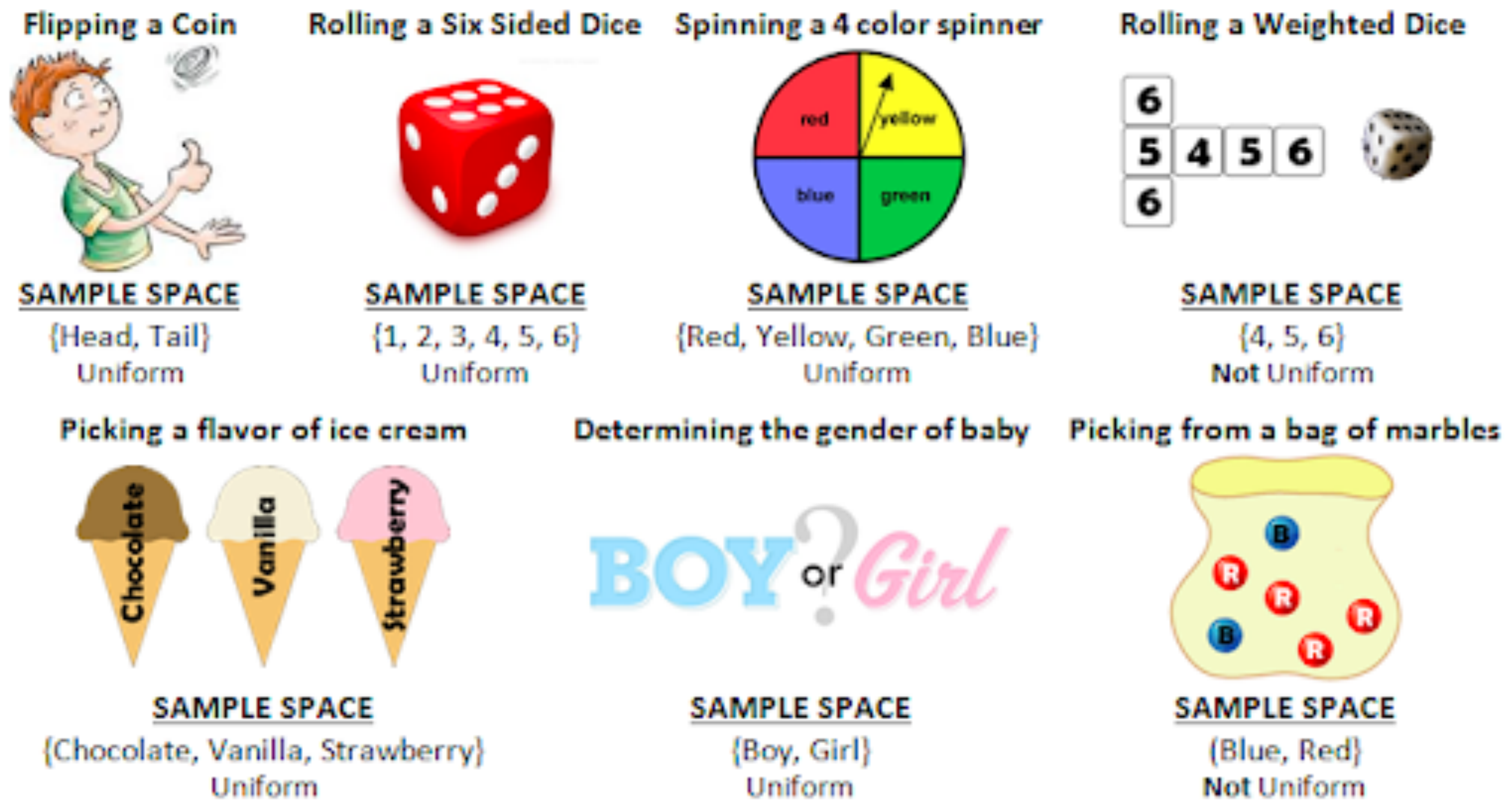
http://www.geometrycommoncore.com/content/unit6/gcp1/studentsnotes1.html



Coin toss

the classic probability example

- Sample space
  - {T,H}
- Experiment
  - **one** toss
- Outcome
  - **either** H **or** T



http://www.geometrycommoncore.com/content/unit6/gcp1/studentsnotes1.html

Coin toss series
the classic probability example

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

- Sample space
  - **depends** on the number of tosses
  - for **2**: {HH, HT, TH, TT}
- Experiment
  - A number of tosses
- Outcome
  - A **sequence** of Hs and Ts
- Statistically, the P(H) is estimated by a large number of experiments
  - toss the coin a billion times
  - compute **how many H** you got (**N**)
  - **N/billion** is the statistical/empirical estimate of **P(H)**
    - and you can actually **prove it formally**
      - Maximum Likelihood Estimation (MLE)

A Fair Coin

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

- A **fair** coin is a coin such that **P(H) = 1/2**
- In other words, you can toss it a billion times and expect H to come up ~500 mln times
    - what if I actually did it and got 500,000,001 Heads?
    - 500,000,001/1,000,000,000 = 0.500000001
        - for all practical purposes, that's **still 1/2** :)

Probability and Frequency

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

- How probable is some outcome?
  - e.g. H or T
- How frequent is some outcome?
  - e.g. H or T
- What's the difference?
  - Frequency is observed
  - Probability is estimated

Probability of sequence
in NLP

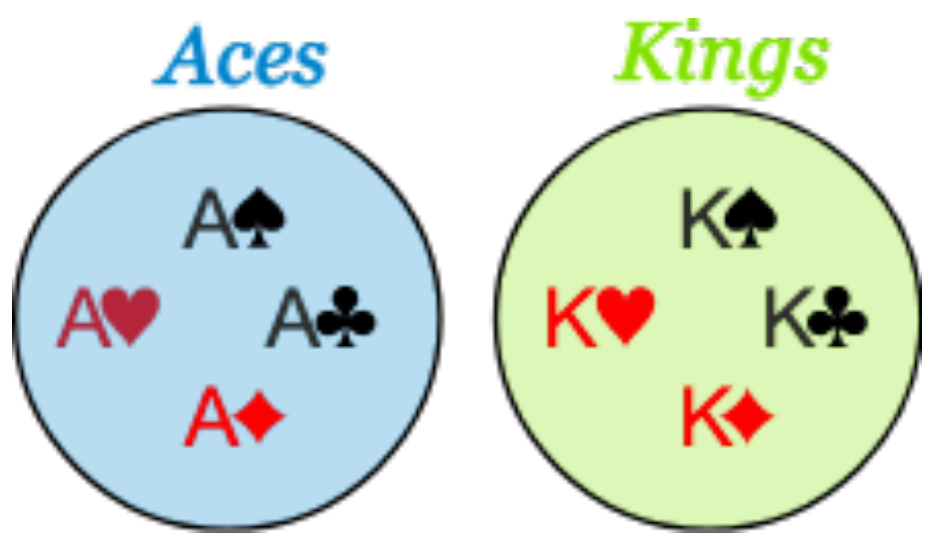
https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

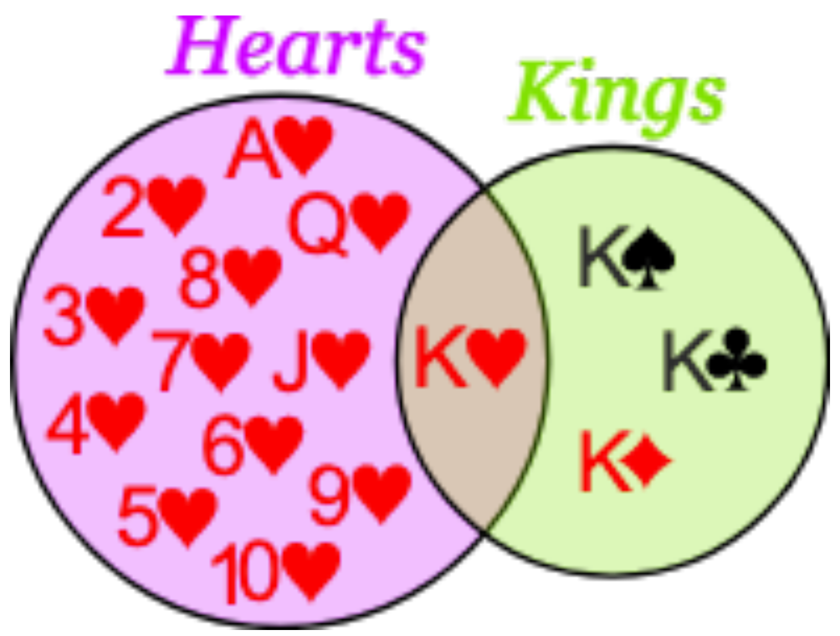
- Very important in data science and NLP!
  - ...because, we usually deal with **many** events
  - ...because, **texts** are **sequences** :)
    - ...of words, characters, syllables, sentences, paragraphs...
    - **language modeling**:
      - estimating probabilities of textual sequences
      - given what we've seen before, what is the **most likely** continuation?

“Probabilities sum to 1”

...for mutually exclusive events

- Why? What does that mean?
  - This refers not to any set of probabilities but only to those which account for **all possible outcomes** in a specific setting
  - Just a convention/definition
  - 1 = **100%**
  - Consider all possible outcomes in the **coin toss** setting
    - e.g. {H,T}
    - when you toss a coin, it **must** result in H or T
    - ...There is a 100% probability that ONE of the possible outcomes will be observed
  - Notation: P(H) + P(T) = 1

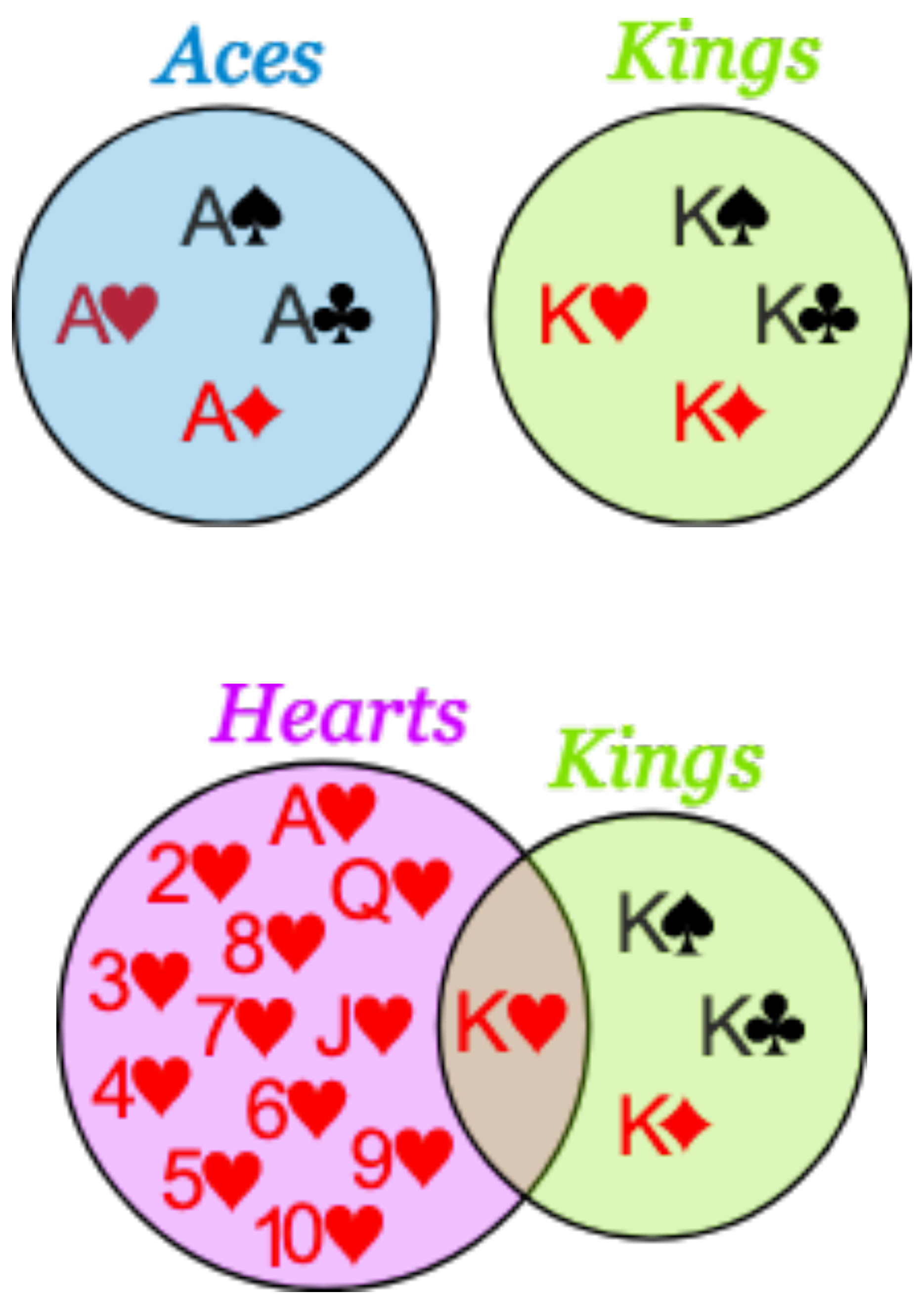




https://www.mathsisfun.com/data/probability-events-mutually-exclusive.html

Mutually exclusive events

- e.g. H and T in a coin toss
  - P (H and T) = 0
    - for one coin toss
- e.g. P(King and Ace) = 0
  - if drawing **one** card
- but: P(King and Hearts) > 0



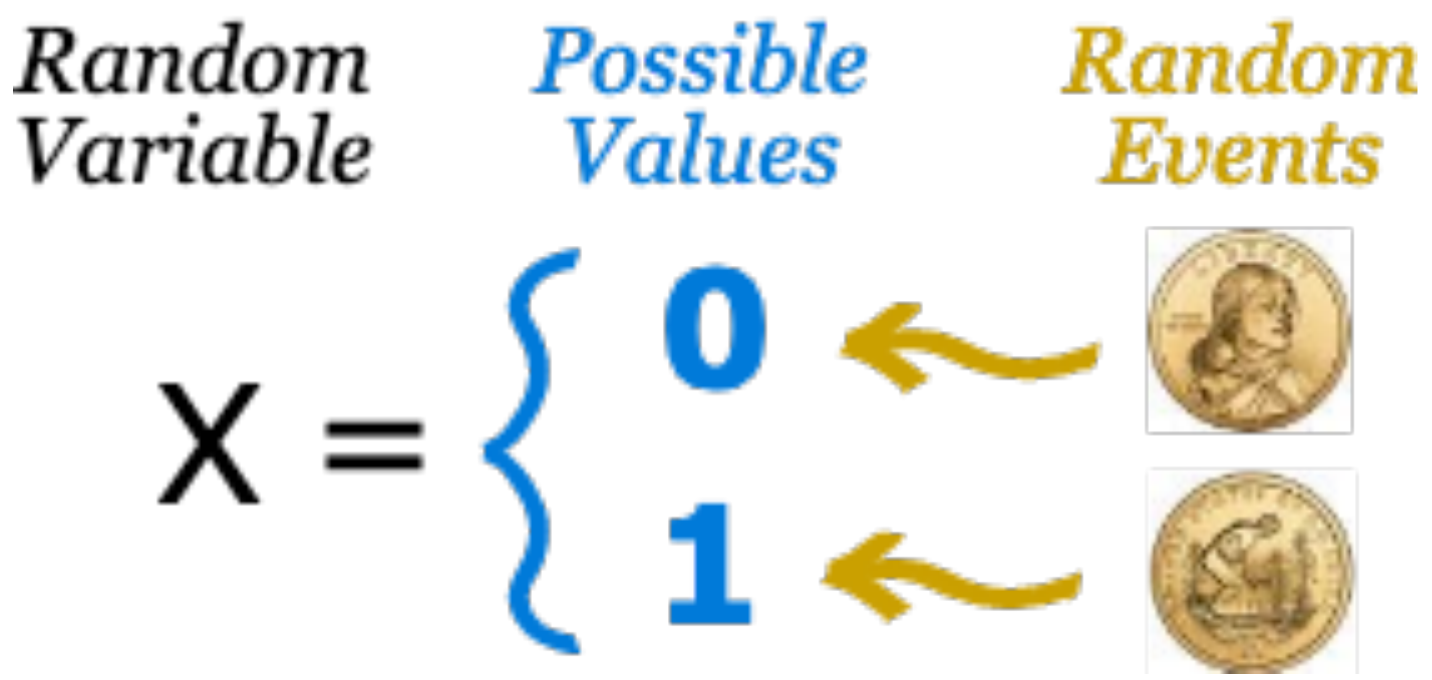
https://www.mathsisfun.com/data/probability-events-mutually-exclusive.html

Probability of sequence
of independent events

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

- Suppose you toss a fair coin twice
- What's the sample space?
  - {HH, HT, TT, TH}
- What's P(HH)?
  - 1/4
  - observe: this is P(H) * P(H)
  - Probability of a sequence is a **product**
- What's P(HT, in this order)?
  - 1/4
- What's P of getting one H and one T, any order?
  - 1/2
  - observe: this is P(HT) + P(TH)!
  - you want to estimate the P of getting one OR the other!
  - Probability of a disjunction is a **sum**

Random variables



https://www.mathsisfun.com/data/random-variables.html

- Set of possible values from a probabilistic experiment
  - e.g. {H, T}
  - we can call H=1 and T=0, or any other arbitrary value!
  - the point is, there is two of them and they are mutually exclusive
- Potentially confusing:
  - What do people mean when saying P(X) or P(A)?
    - it depends, but most often they mean:
      - if A is a random variable and the values are e.g. {1,2,3,4,5,6}
      - then P(A) may refer specifically to P(A=1) or P(A=5)

Independent events

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

- One event does not affect the other
  - e.g. coin toss/die roll etc.
- P(A and B) = P(A)* P(B) **only** if A and B are independent

- P(1)?
- P(2)?

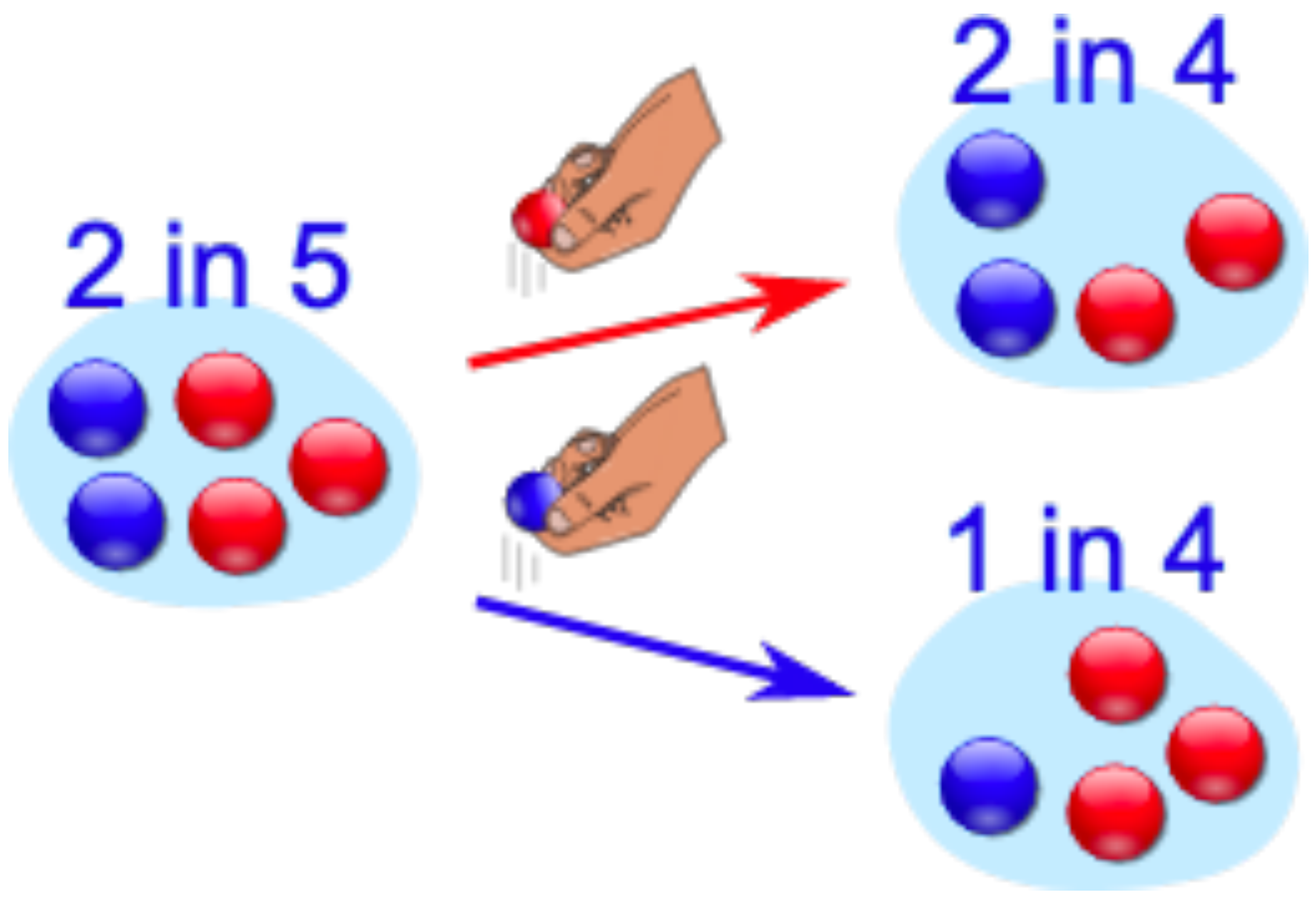
Independent events

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

- One event does not affect the other
  - e.g. coin toss/die roll etc.
- P(A and B) = P(A)* P(B) **only** if A and B are independent

- P(1) = 1/1024
- P(2) = 1/1024
  - whaaaat?!
- This is unintuitive, because we were not comparing P(1) to P(2)
  - we were comparing P(1) with something more like 1 - P(1)

Conditional probability

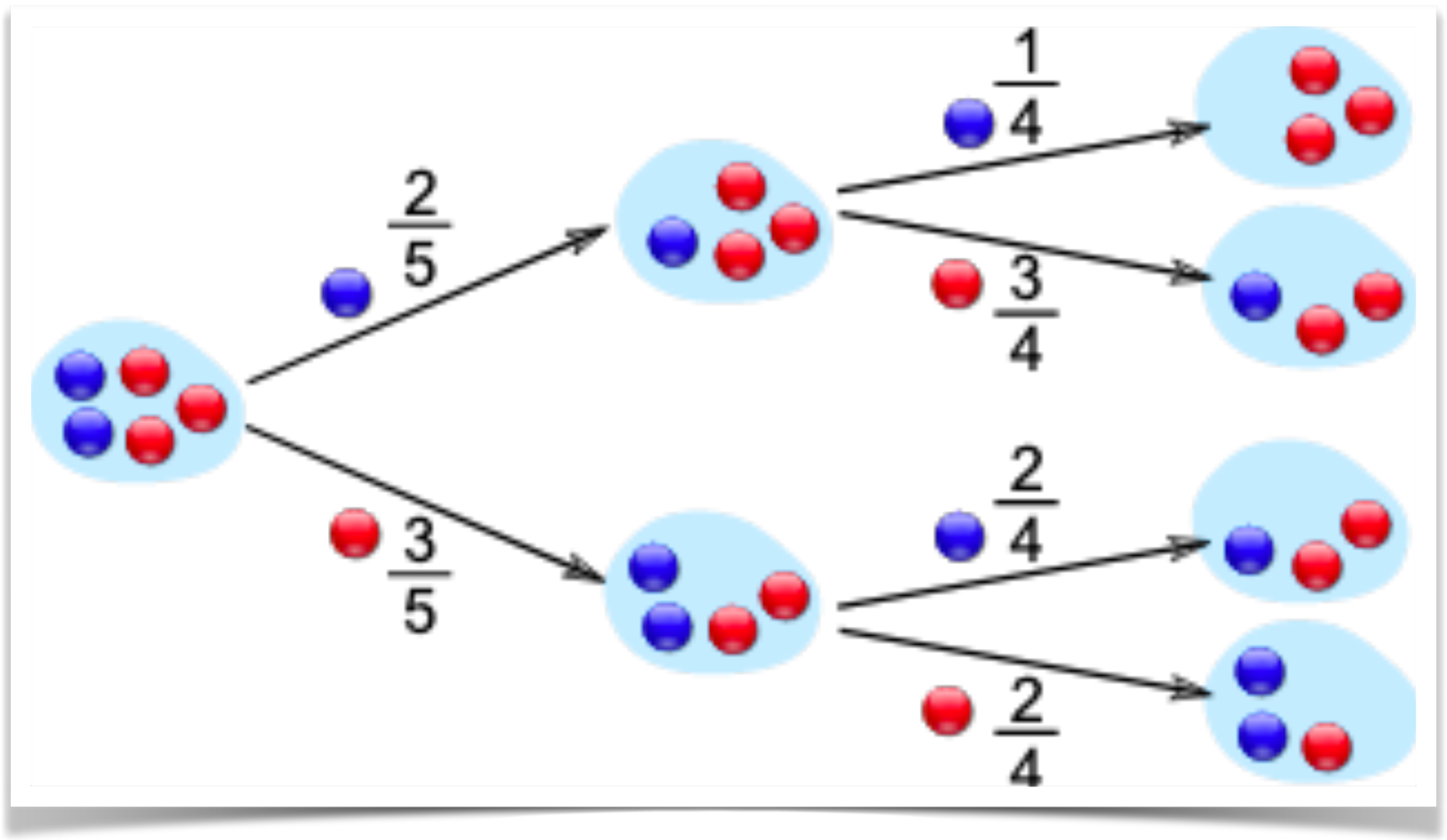


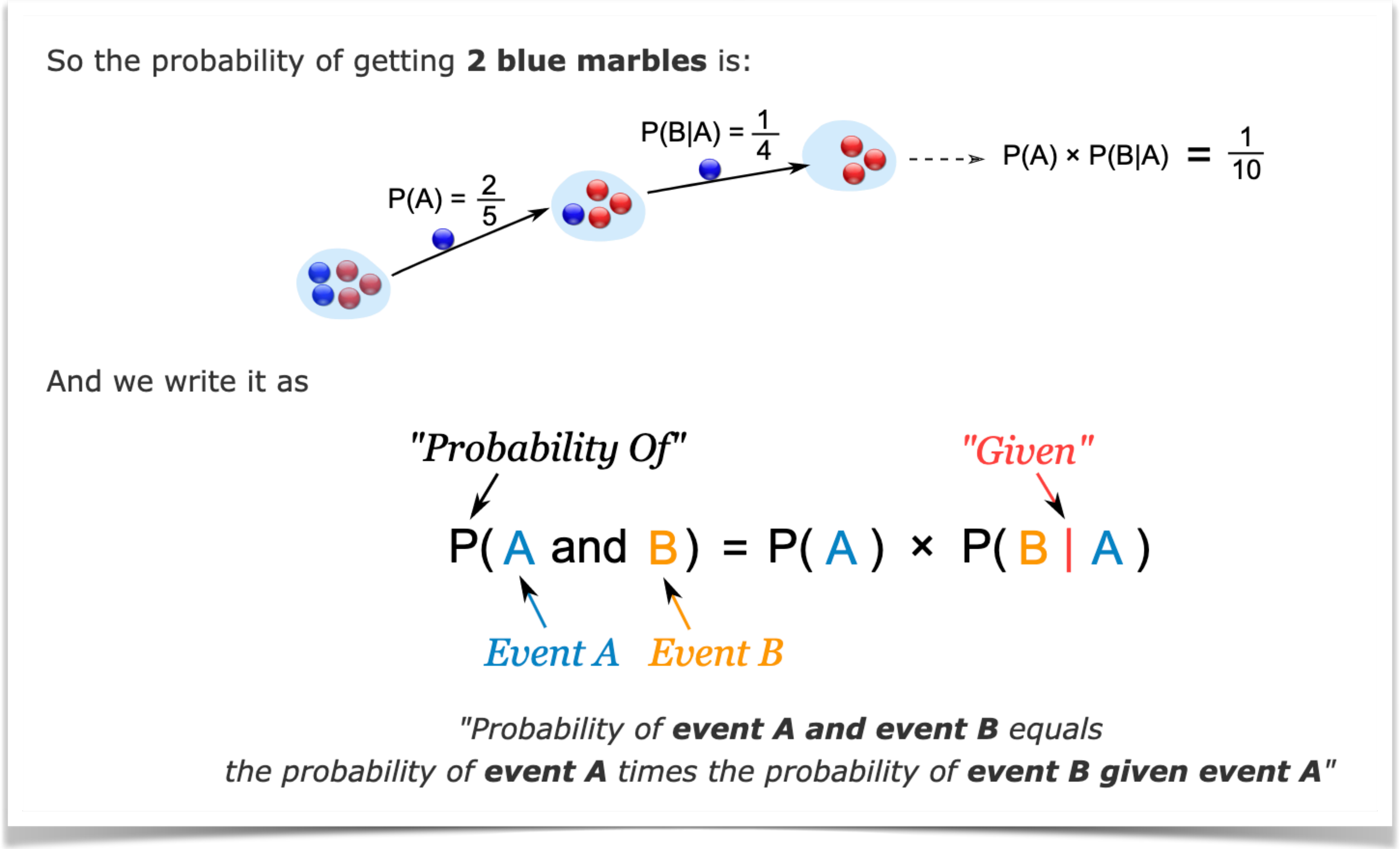
https://www.mathsisfun.com/data/probability-events-conditional.html

- What's the probability of A **given** B?
  - e.g., if it is very sunny, is it more or less likely that it will rain in 30 minutes?
    - (compared to when it is **not** sunny)
  - e.g. if you see lightning, is it more or less likely that you hear thunder in a few seconds?
    - (compared to when you **don't** see a lightning)
  - Formal example: removing marbles from a bag
    - consider the **sample space**

Conditional probability
definition

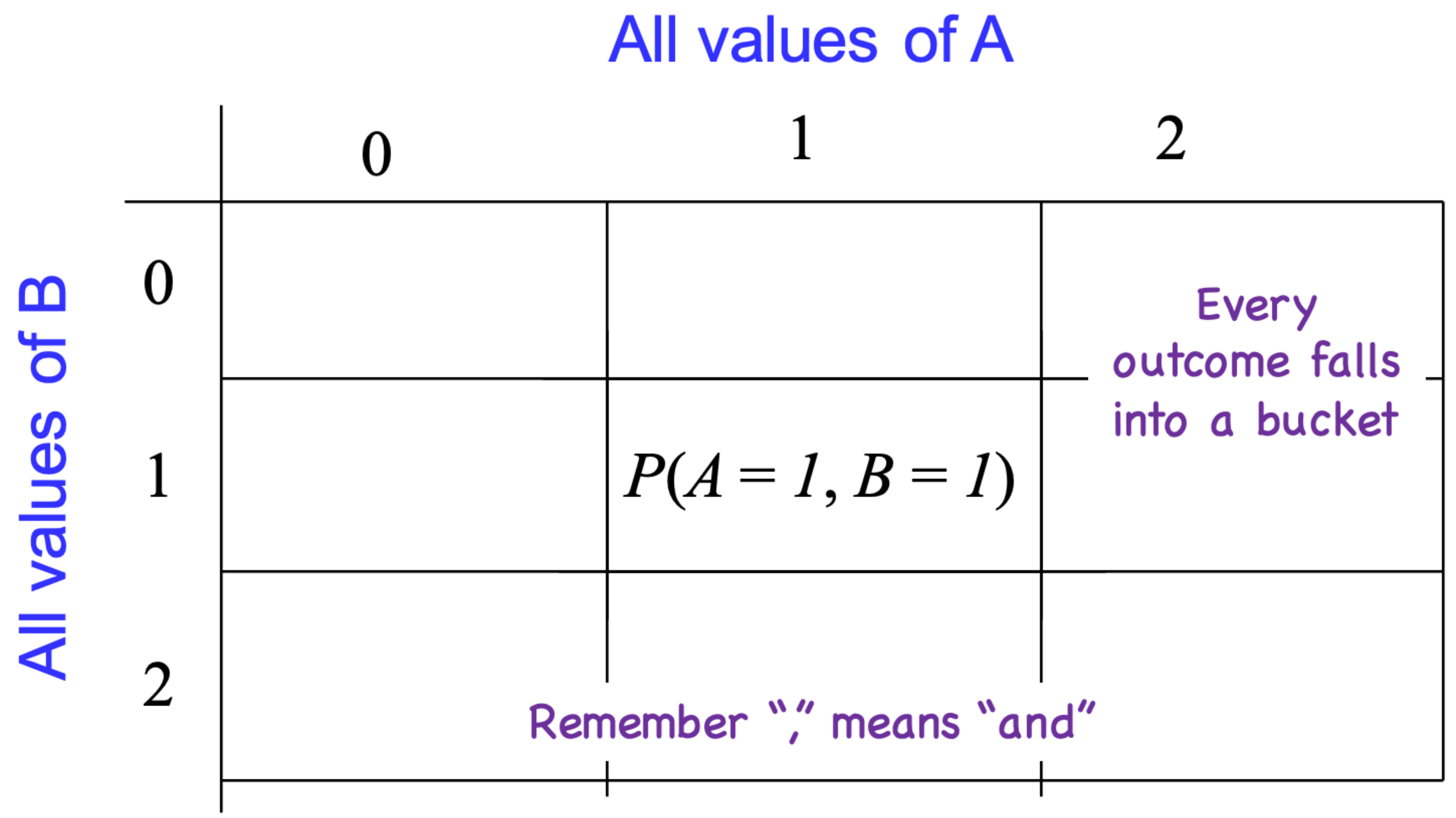
- P(thunder | ligntning) = P(L and T)/P(L)
  - P(L and T):
    - estimated by counting all occurrences when **both** things occurred
  - P(L):
    - estimated by counting all occurrences when **L** occurred
- Conditional prob. is crucial in the **Bayes** Theorem
  - and the Naive Bayes classification algorithm
  - the bread and butter of many data science techniques
  - Assignment 4





https://www.mathsisfun.com/data/probability-events-conditional.html

Marginal probabilities



- Prepresent conditional probabilities in tables
  - the table has joint probabilities in it, of two events
  - to marginalize a probability of A is to compute P(A) by removing any dependencies on other events
    - by summing along row or column
      - e.g. 0.24 is the P of being a Freshman
      - e.g. 0.45 is the P of being Single
    - the marginals should sum up to 1
      - across row and separately along column
      - why?

	Joint Probability Table			
	Single	In a relationship	It's complicated	**Marginal Year**
Freshman	0.13	0.09	0.02	0.24
Sophomore	0.16	0.10	0.02	0.28
Junior	0.12	0.10	0.02	0.23
Senior	0.01	0.09	0.00	0.10
5+	0.03	0.12	0.01	0.15
Marginal Status	0.45	0.48	0.07	

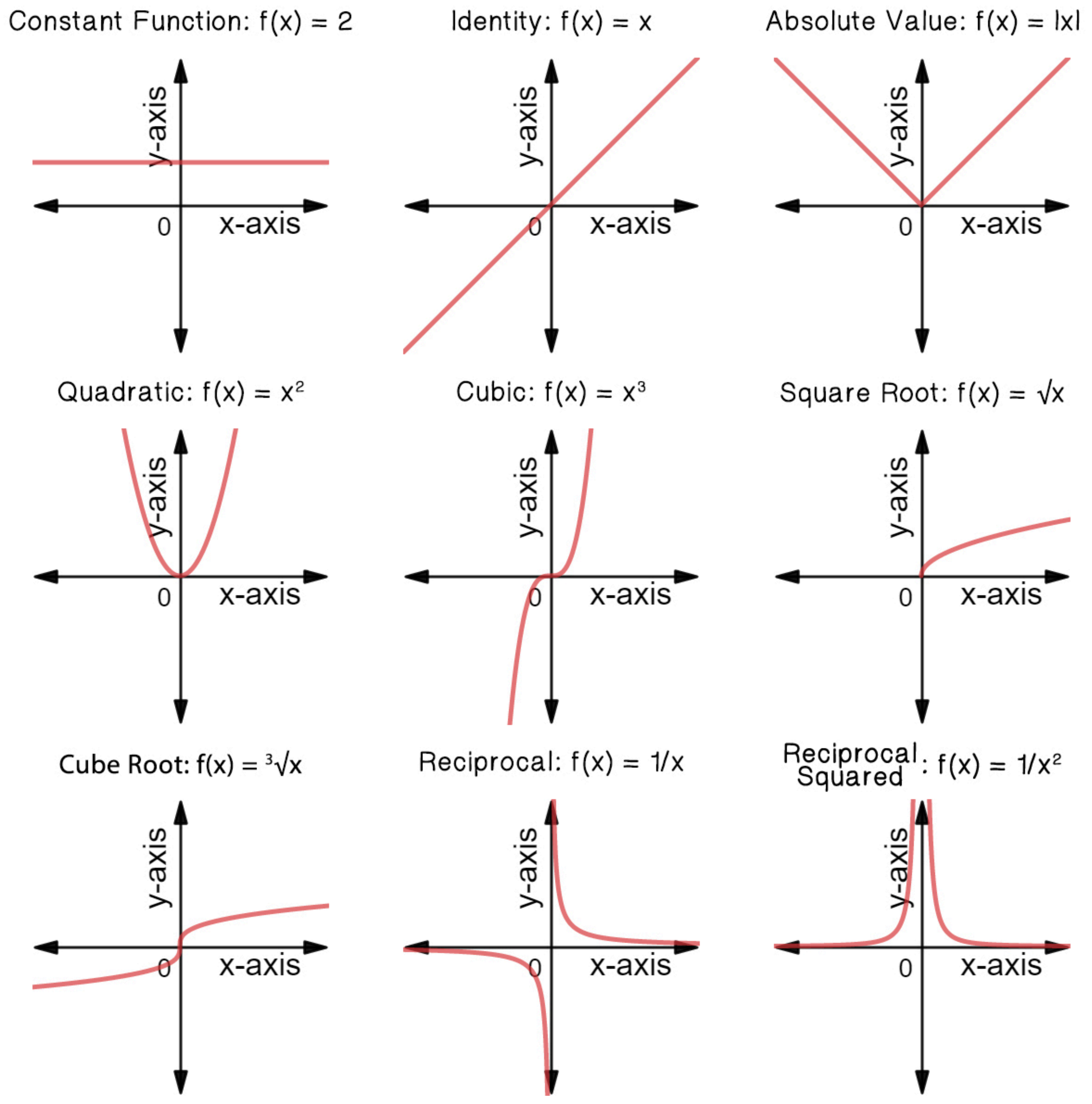
https://web.stanford.edu/class/archive/cs/cs109/cs109.1176/lectures/12-ContinuousJoint.pdf

Statistics

Let’s work with probabilities to estimate what the world looks like!

Functions
review

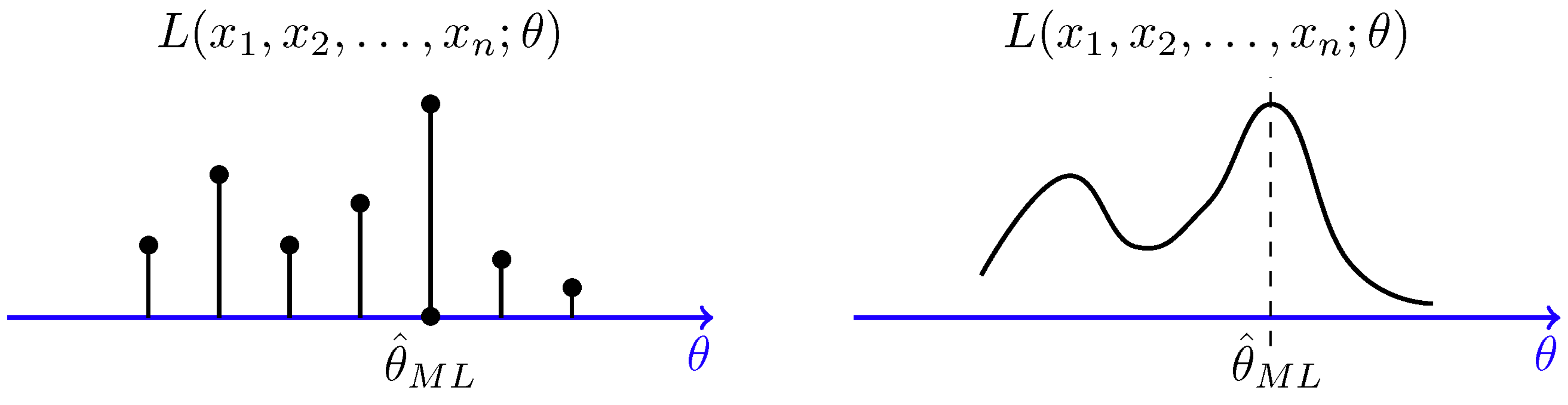
- Functions are bread and butter of statistics
- Function:
  - input—output
  - given **x**, what is the value of **y**?
  - f(x)
  - e.g f (x): y = 2x
- Function equations can be visualized as lines and curves (in 2D)
- **Probabilities** can be seen as functions
  - what is the **probability** of observing **datapoint** x?
  - ...need to know how datapoints are **distributed**
  - **probability functions** describe such **distributions**



https://www.expii.com/t/classifying-common-functions-4320

Maximum (log) Likelihood

Maximum likelihood
estimation



https://www.probabilitycourse.com/chapter8/8_2_3_max_likelihood_estimation.php

- Goal:
  - Represent probabilities **abstractly**, as formulae
  - Prob. of each outcome is a **parameter**
  - Parameters can be **unknown**; we want to **estimate** their values
    - e.g. (weighted, non-fair) coin toss
    - What's the P(H)?
      - we don't know, so we will use an abstract parameter
      - θ
      - then P(T) = 1 - θ
      - then P(HT) = θ*(1-θ)
      - then P(HHHTT) = $\theta^3 * (1 - \theta)^2$
      - What is θ ?

Maximum likelihood

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

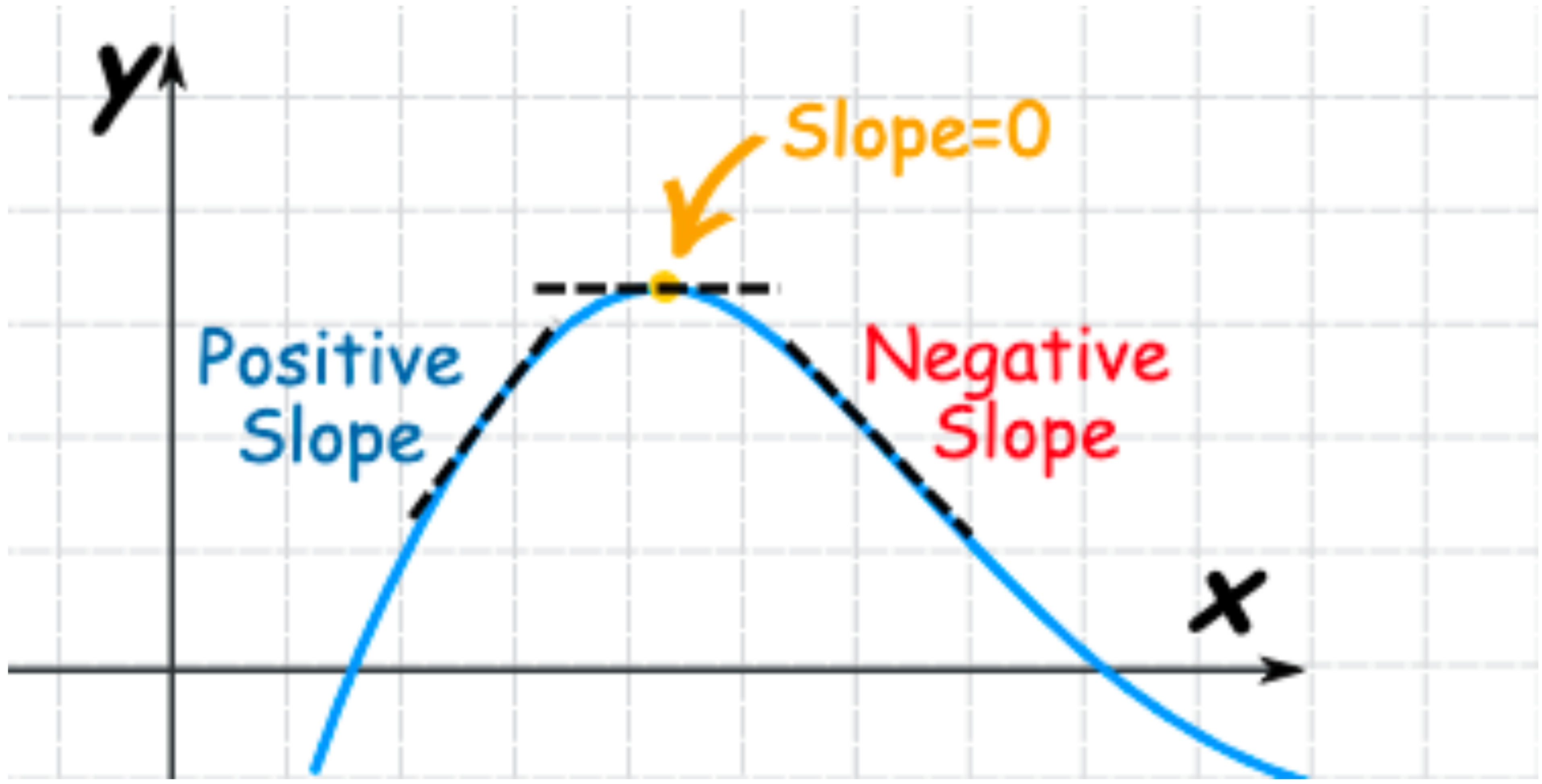
- Suppose we tossed a non-fair coing 5 (billion) times:
  - result{H,H,H,T,T}
  - what's the P(H)?
    - 3/5
    - This is by definition, which is theoretical
    - Can we get some practical evidence for this?

Maximum likelihood estimation

https://www.air-worldwide.com/blog/posts/2019/3/understanding-probability-are-you-asking-the-right-questions/

- Yes!
- We know there is some P of getting H:
  - call it θ
- What do we know about P(T)?
  - it has to be 1-θ
- D = {HHHTT}
  - What's P(D)?
  - P(D) is the **product** of the probabilities

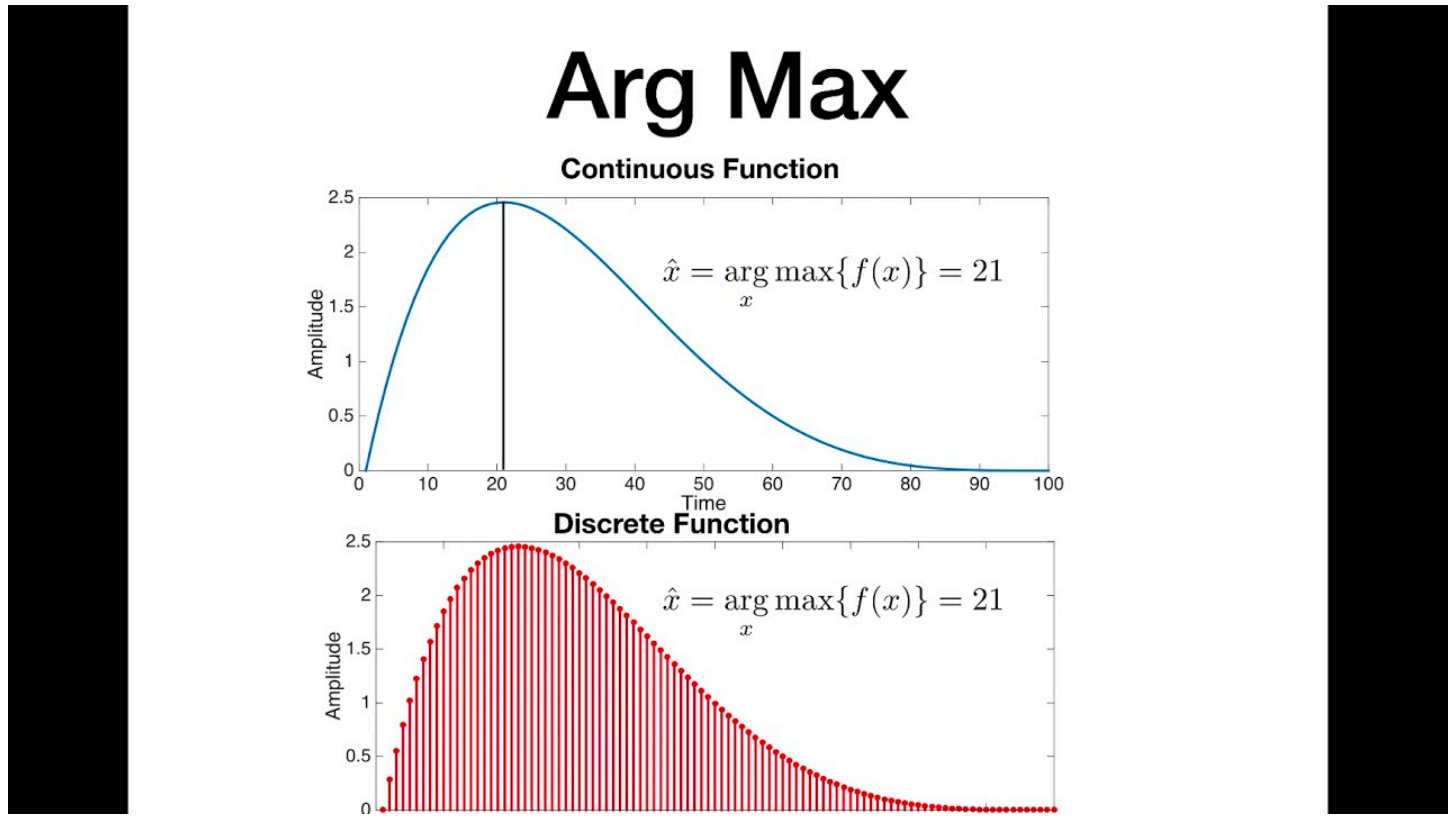
Maximum likelihood
estimation



- D = {HHHTT}
  - P(D) = $\theta^3 * (1 - \theta)^2$
- What are we after here?
  - θ (aka P(H))
- We want a value for θ **such that** P(D) is **max**!
  - how to find the **maximum** point of a function?
  - think of functions as **curves**
  - a curve becomes **flat** at its maximum
  - a curve's **slope** is its **derivative,** and derivative = **0** at the flat point
  - which may be directly **computable** (calculus)
  - we know how to compute derivatives for a range of functions
    - we just **look it up**
  - for functions for which we **can't** compute the derivatives:
    - we estimate by **other means** ("gradient descent")

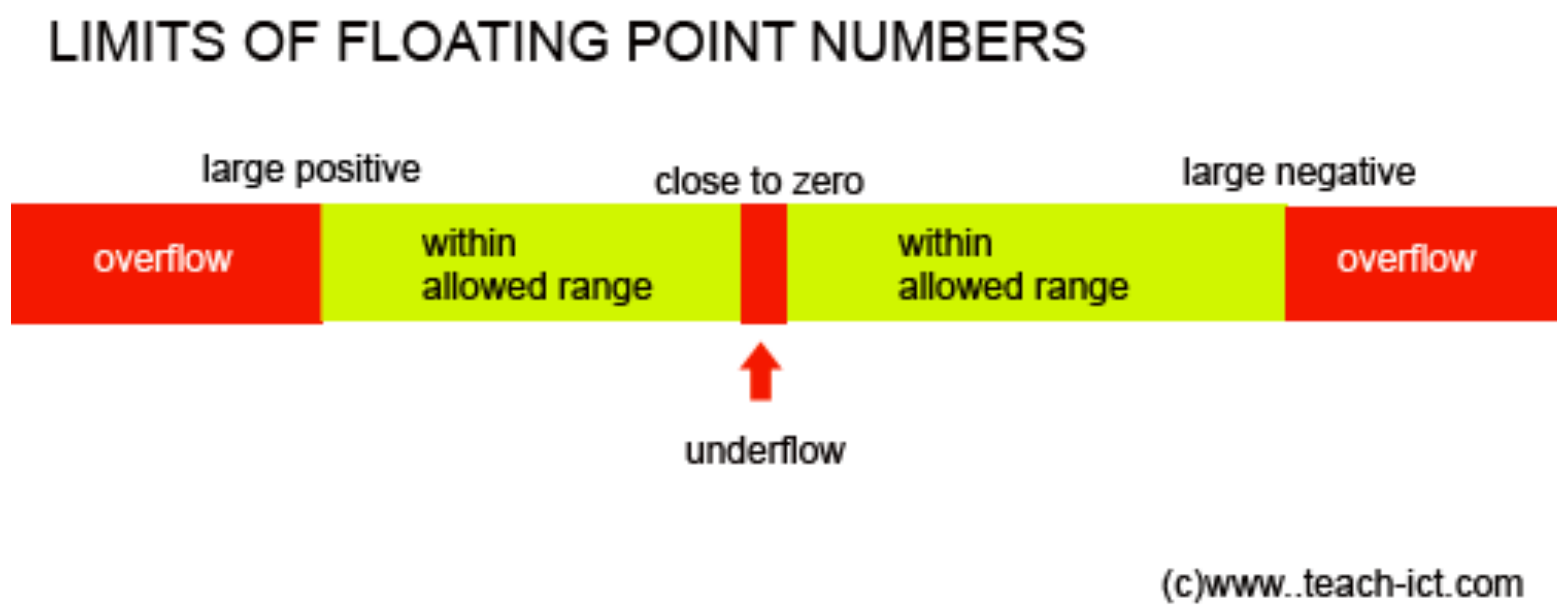
Before we continue: Two additional pieces

arg max



- functions look like **curves** (in 2D)
- Those curves have **maxima** along the **Y**-axis
- The point on the **X-axis** where Y is maximum:
    - is the **arg max**
- Why is this important:
    - We want to find parameters for probability functions given our observations
    - If the function has parameter θ, **which** value for θ results in **maximum** probability for the **observed** sequence/data?

Logarithms and Products

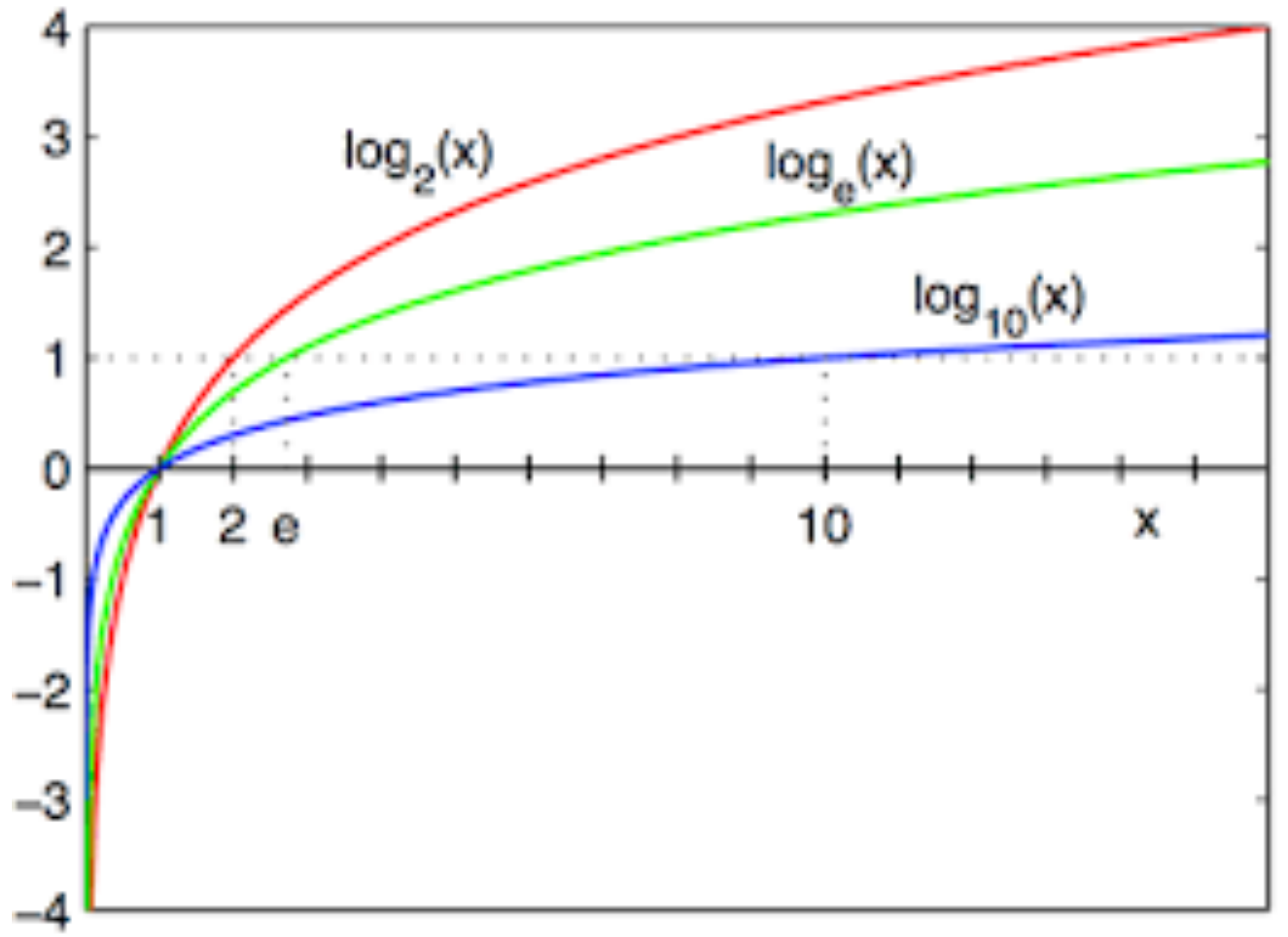


https://www.teach-ict.com/as_as_computing/ocr/H447/F453/3_3_4/floating_point/miniweb/pg9.htm

- Probabilities range from **0 to 1**
- Suppose you have a **looooong** sequence of events
- What happens if you multiply **many-many** numbers **each** ranging between 0 and 1?
  - your number becomes **so small** that the computer **cannot represent** it
  - **logs** to the rescue!

Logarithms and Products

- log(x*y) = log(x) + log(y)
- Due to certain **properties of the log**:
  - Can use log(P(A)) **in place** of P(A)
    - for likelihood estimation
    - arg max of P(D) will be where arg max for *log*(P(D)) is!
      - and *ln*(P(D))
  - => Can use **sum of logs instead** of product



https://en.wikipedia.org/wiki/Logarithm

- Reminder:
  - log is inverse function to exponent
  - e.g. 10^2 = 100
  - => $log_{10}(100) = 2$
  - *ln* is "natural log"; it is "base 2.71828" (*e*)

Maximum likelihood

for calculus fans

- D = {HHHTT}
  - P(D) = $\theta^3 * (1-\theta)^2$
- What are we after here?
  - θ (aka P(H))
- We want a value for θ such that P(D) is max!
  - we know the derivative for natural log of x
    - as well as for ln(1-x)
    - use θ as x

$$P(D) = \theta^3(1-\theta)^2$$

$$\hat{\theta} = \arg\max_\theta P(D;\theta) = \arg\max_\theta \ln(\theta^3(1-\theta)^2) =$$

$$\frac{d}{d\theta}\ln(\theta^3(1-\theta)^2) = \frac{d}{d\theta}\ln(\theta^3) + \ln(1-\theta)^2$$

$$= \frac{d}{d\theta}\ln\theta^3 + \frac{d}{d\theta}\ln(1-\theta)^2 =$$

$$3\frac{d}{d\theta}\ln\theta + 2\frac{d}{d\theta}\ln(1-\theta)$$

$$3\cdot\frac{1}{\theta} + 2\cdot\left(-\frac{1}{1-\theta}\right) = 0$$

$$\frac{3}{\theta} - \frac{2}{1-\theta} = 0$$

$$\frac{3}{\theta} = \frac{2}{1-\theta}$$

$$3(1-\theta) = 2\theta$$

$$3 - 3\theta = 2\theta$$

$$3 = 5\theta$$

$$\theta = 3/5$$

Lecture survey in the chat