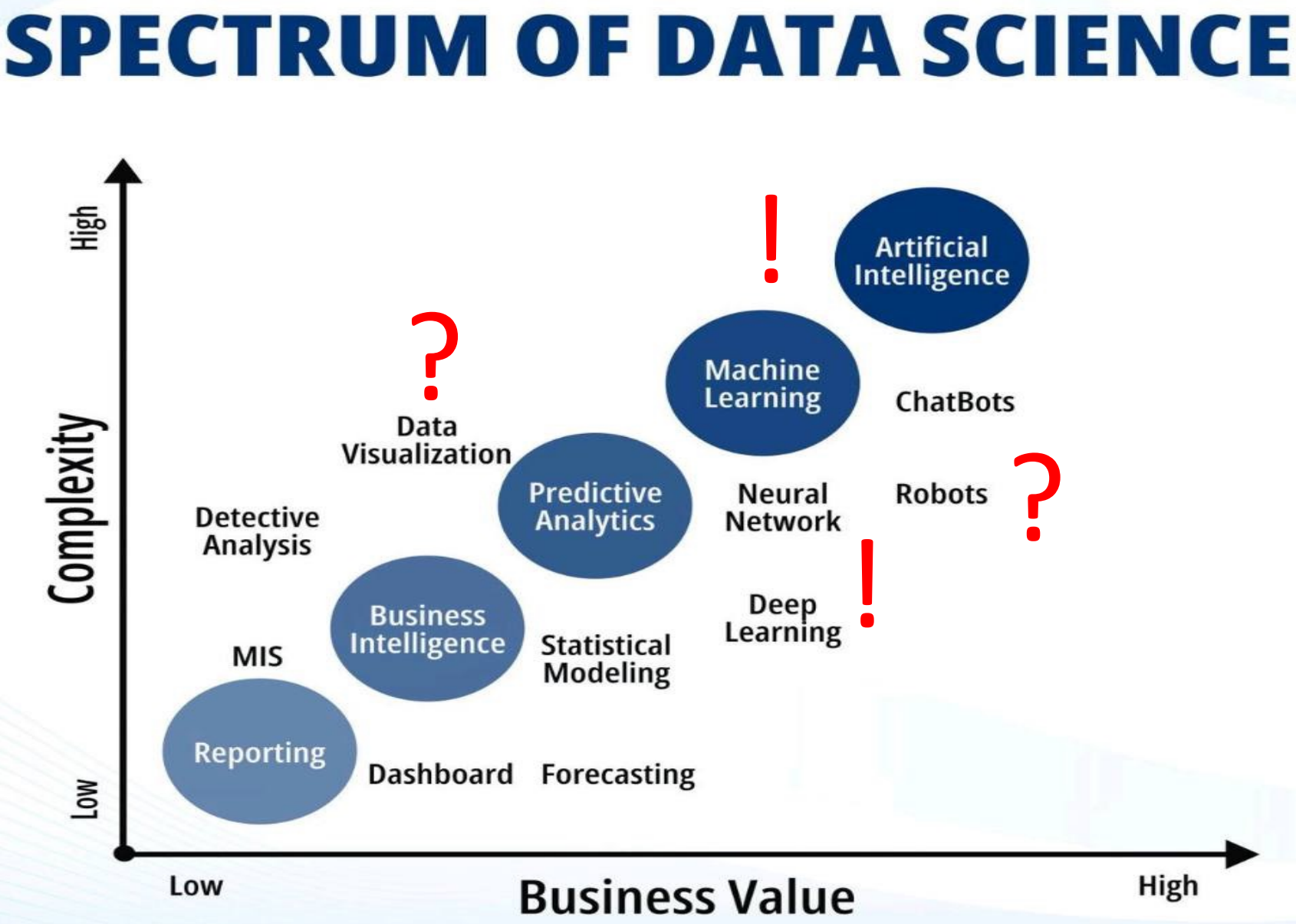
Smallest font

Please turn off and put away your cell phone

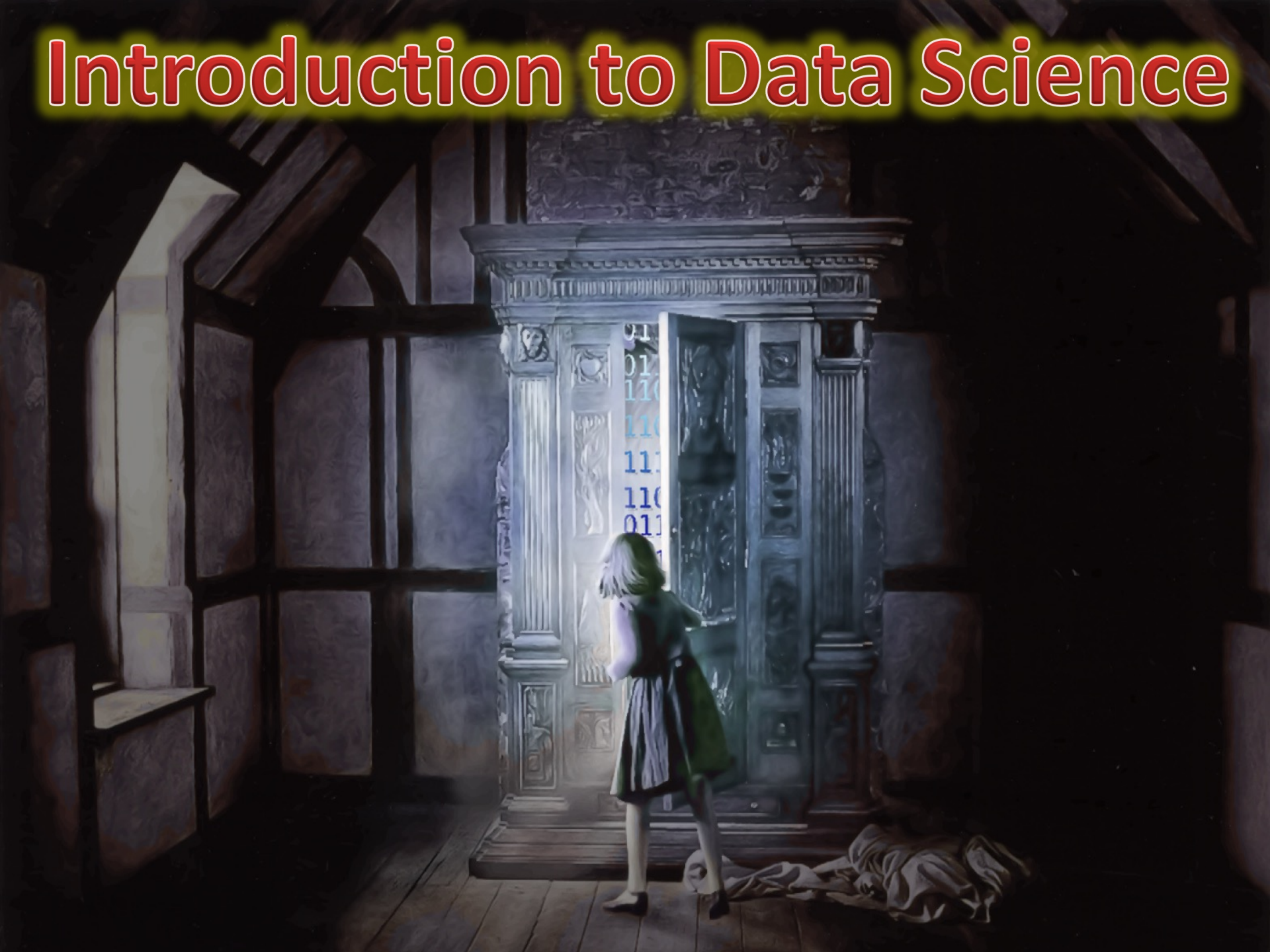
Calibration slide

Smallest font

It’s a daily thing

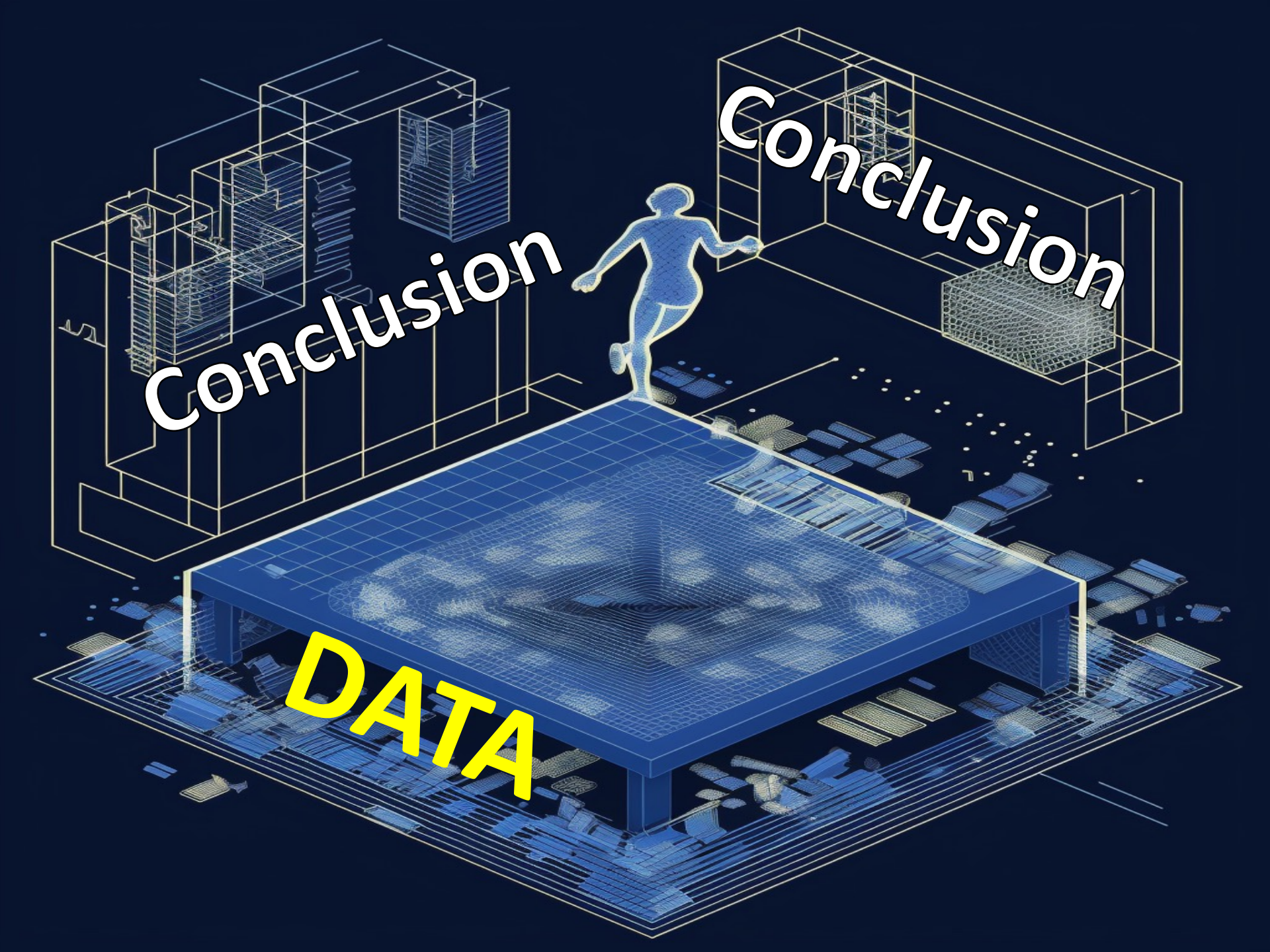


Introduction to Data Science

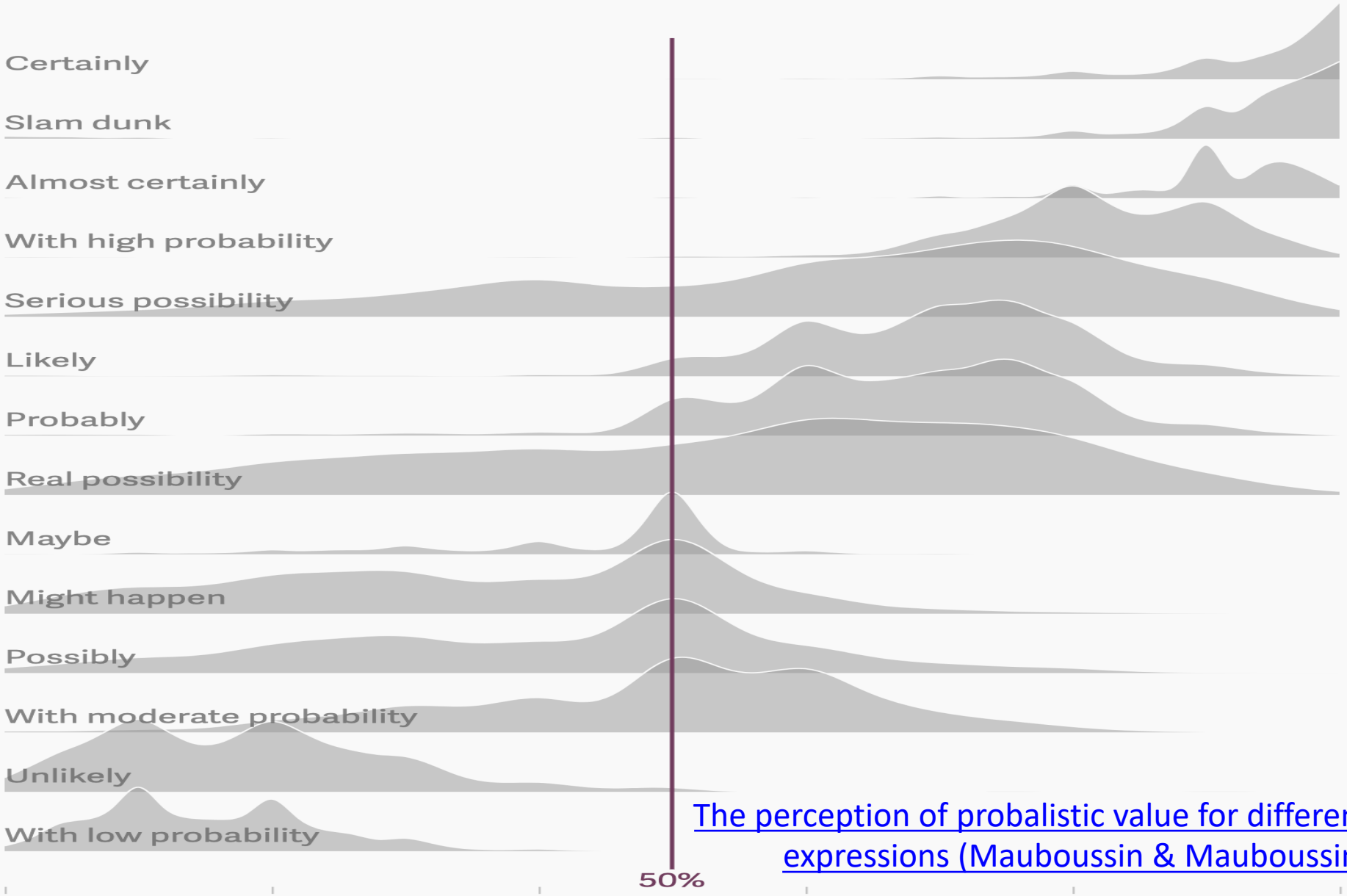


I

Inference


Conclusion
Conclusion
DATA

Foundations : Probability



The perception of probalistic value for different expressions (Mauboussin & Mauboussin)

The need to quantify uncertainty arises from the limitations of formal logic

- Formal logic works well if something is always or never the case, or for all (instances of categories with certain properties) or for none (of them):

Syllogistic reasoning:
- Major premise
- Minor premise
- Conclusion

- All candy is sweet
- All chocolate is candy
- All chocolate is sweet

Modus ponens:
- If P then Q
- P
- Therefore Q

- If it rains, the street is wet
- It rains
- The street will be wet

Modus tollens:
- If P then Q
- ~Q
- Therefore ~P

- If it rains, the street is wet
- The street is not wet
- It could not have rained

- In other words, it works well for “simple” domains (e.g. numbers) or domains we created (e.g. by coding).
- But reality is complicated:

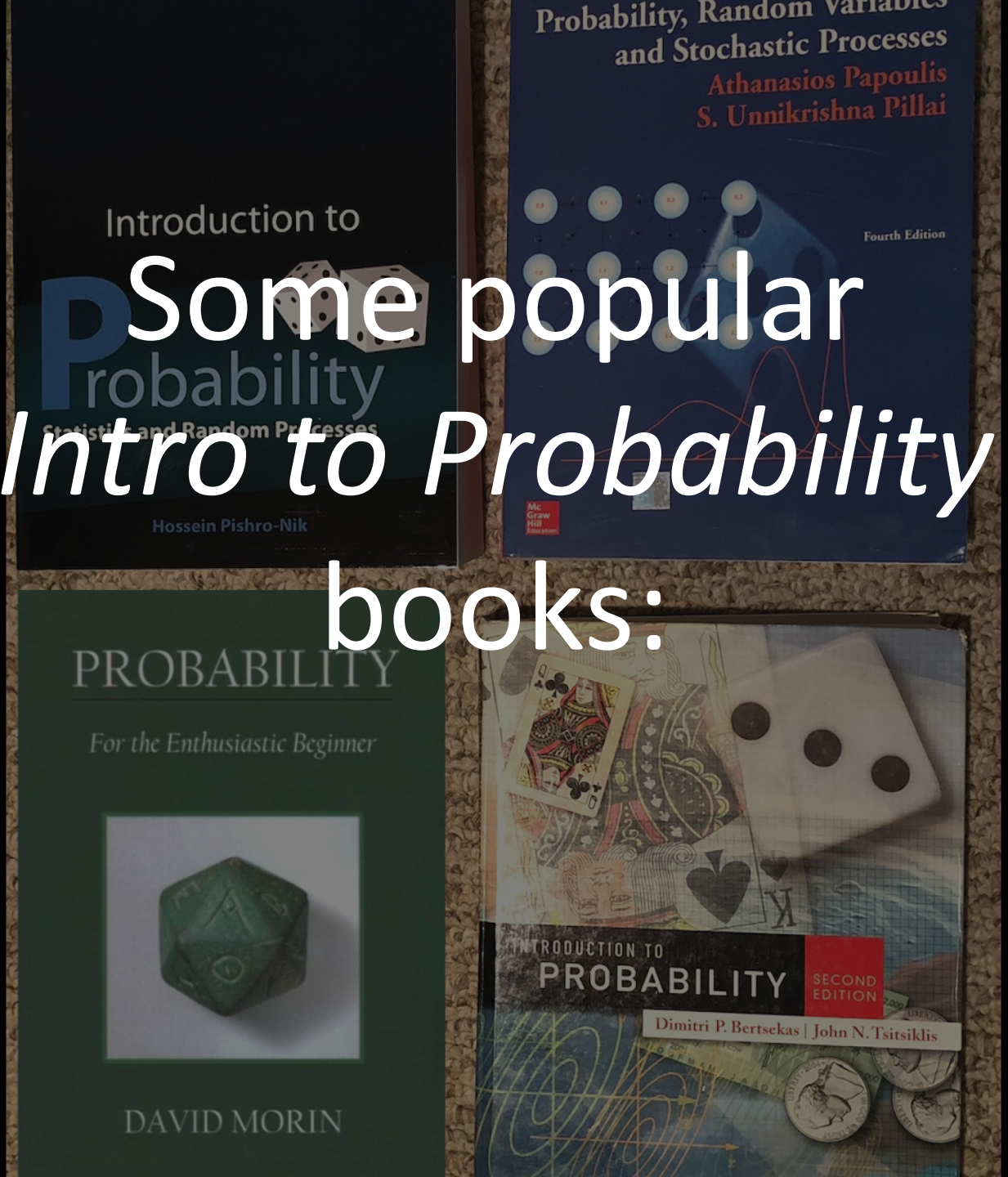
What if any of the the ”all” in the premises above had been “some” instead?

What if it didn’t rain (~P) or what if the street was wet (Q)?

What if it did rain, but the water evaporated, so the street is not wet?

Probability theory as a 3rd way

- When dealing with uncertainty conceptually, we are looking for something that quantifies the likelihood of something happening
- Everyday language seems too fuzzy
- Formal logic is too restrictive
- We need a new approach:
- Probability theory (just right)



Some popular *Intro to Probability* books:

There are historical reasons for this

- Whereas the concept of probability is ancient, it was put on a solid mathematical foundation relatively recently: In the 1660s.
- By analyzing games of chance (mostly card & dice games).
- Kicked off by letters from Pascal to Fermat.
- Kolmogorov defined it axiomatically in the 1930s (based on measure theory)


The emergence of probability


The taming of chance

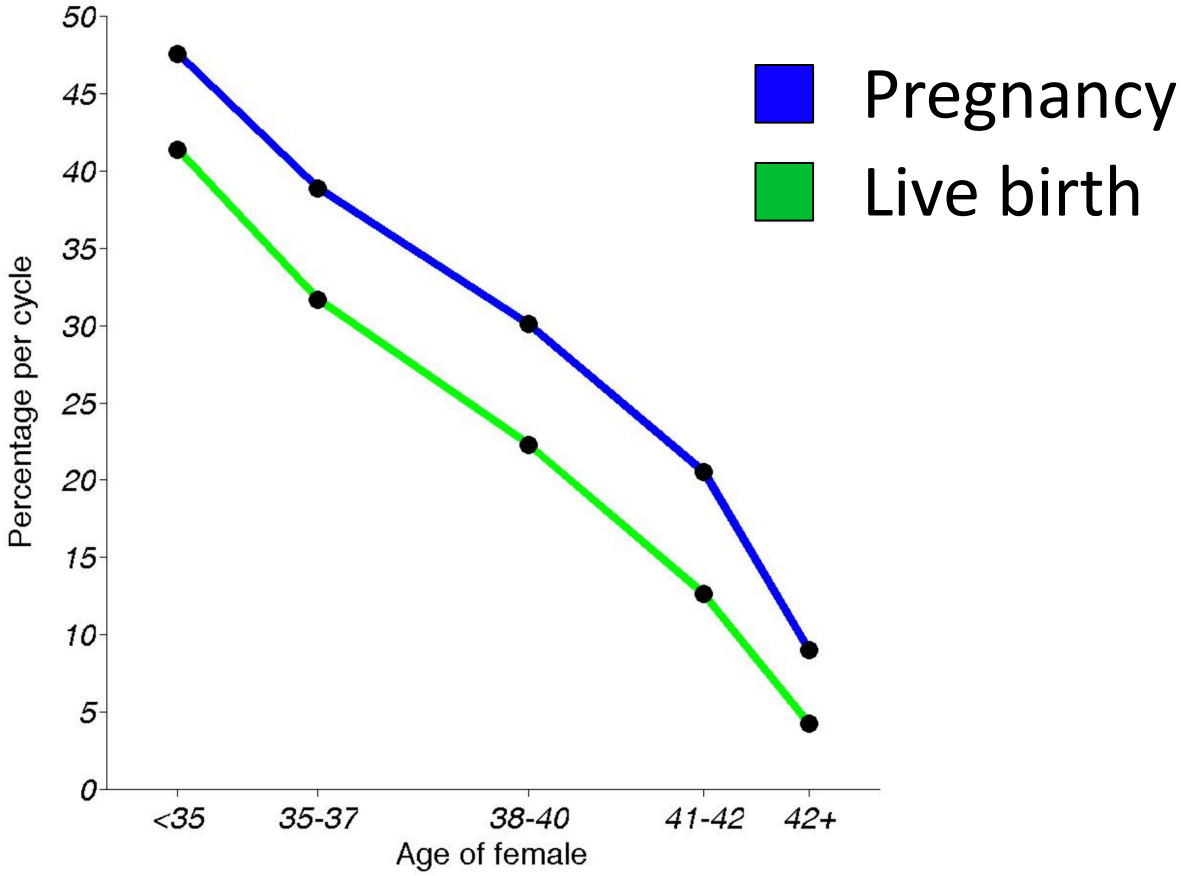
The primary purpose of probability

- Rational **decision making** under uncertainty.
- Uncertainty is very common (in real life).
- The need for rational decisions is as well.
- Probability will be our primary tool underlying data-based **inference**.
- But the scope of probability is much broader:

"Probability is the very guide of life"
-Cicero, De Natura, 5, 12

The example we will use for much of this lecture is based on real life data – and a growing field that relies much on probability: Assisted reproductive technology

IVF success rates as a function of age



A typical case study

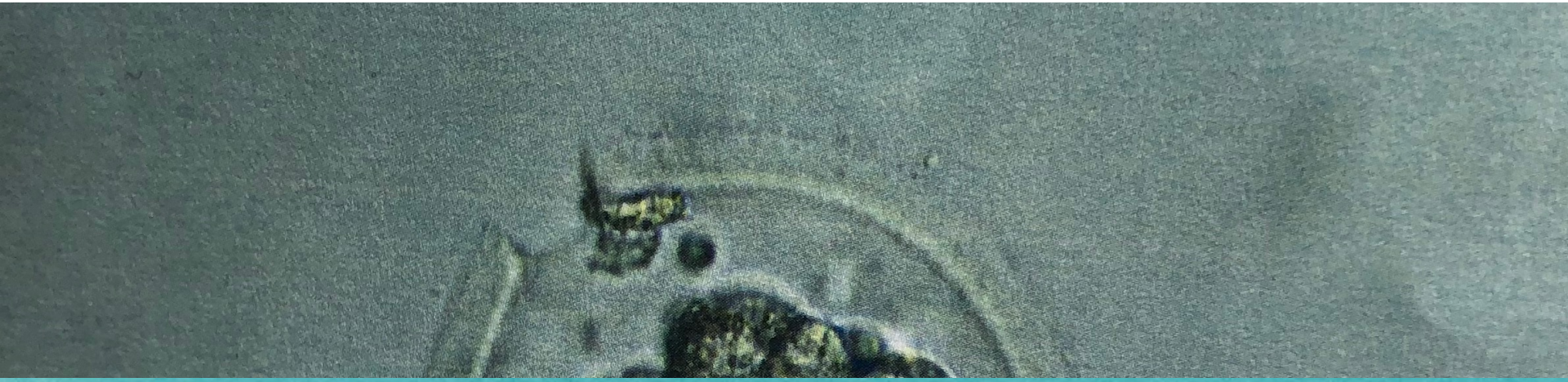
- A 29 year old patient wants to have children with the aid of IVF.
- She undergoes a medicated cycle where she produces – with the aid of hormones - **19** eggs. This is considered a good yield.
- All **19** of eggs (*100%*) are successfully retrieved in a minimally invasive procedure (“egg retrieval”).
- **13** of these **19** eggs (*68%*) are mature enough to attempt fertilization.
- **11** of the **13** remaining eggs (*85%*) actually fertilize.
- **8** of the remaining **11** eggs (*73%*) grow at a fast enough rate to be considered viable embryos.
- At day 5, two decisions have to be made - how many of these embryos to biopsy, and which embryo(s) to transfer, if any.
- Biopsy: Taking a handful of the ~100 embryo cells out for genetic testing.

The results of the genetic testing of the biopsy

ID	Sex	Morphology	Analysis	Interpretation	Result
2	XY	B b	-5[m]	Mosaic	Mosaic
3	XY	B b	-21	Abnormal	Aneuploid
6	XY	B a	All 46 intact	Normal	Euploid
8	XX	B c	+4[m], +9[m], +12[m], +17[m]	Abnormal	Aneuploid
9	XX	B b	All 46 intact	Normal	Euploid
10	?	C b	untested	untested	-
11	XY	C b	-15[m]	Mosaic	Mosaic
12	XY	B c	-12[m]	Mosaic	Mosaic

Now what?

Preview: What the right / rational thing to do is depends on the nature of the sampling (How the biopsy - was done)



ABNORMAL CELLS IN EMBRYOS MIGHT NOT PREVENT IVF SUCCESS

Study shows that chromosomal abnormalities in embryonic cells may be more common than previously thought and these conditions may lead to development of healthy babies during IVF

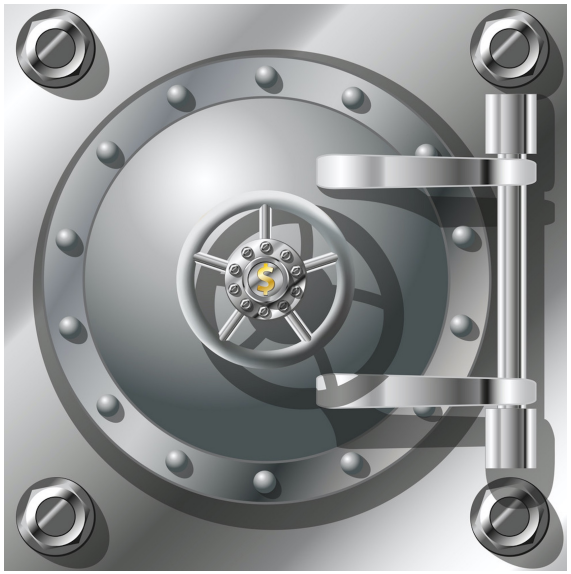
Doug Donovan / Jul 10

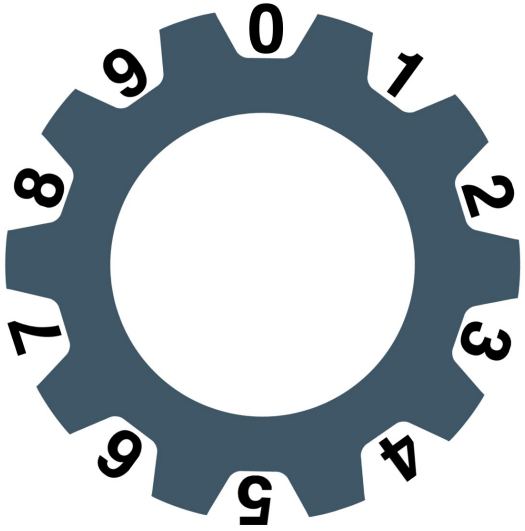
In order to model this situation properly, we first need to introduce terminology and concepts from probability theory

- Preview: The computations are straightforward (just addition, subtraction and multiplication).
- *Given* that the situation is represented properly (distinguishing mutually exclusive from independent from conditional situations)
- *And* thinking clearly about **random variables** and **probability distributions**.

Terminology of probability: The combination lock

- Ω: **Sample space** – the set of all possible outcomes of a random experiment
- Sample space here: {0,1,2,3,4,5,6,7,8,9}
- ω ∈ Ω: One specific **outcome** (lower-case omega) is an element of the sample space
- $F = P$ (Ω): The **event space**, a power set (set of all subsets) of the sample space Ω
- **P** : $F \rightarrow [0, 1]$ is a **probability measure** (or probability), given Kolmogorov's axioms of probability.
- A triple (Ω, F, **P**) forms a **probability space**.





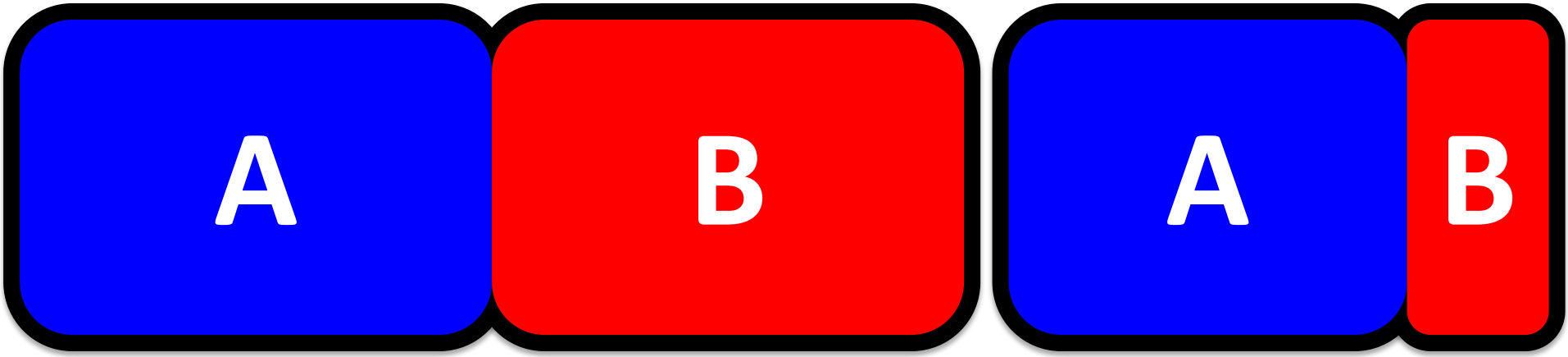
The three axioms of probability (Kolmogorov, 1933)

- Axiom 1: $P(E) \in \mathbb{R}$, ≥ 0 (Probabilities of events are positive real numbers)
- Axiom 2: $P(\Omega) = 1$ (The probability that the outcome of the random process is in the sample space is 1)
- Axiom 3: If there are disjoint subsets (mutually exclusive events), the probability of their union is the sum of the probabilities of the individual events.

These axioms yield the rules for computing with probabilities

1) The addition rule

- If events are mutually exclusive
- p (A or B) = p (A ∪ B) = p(A) + p(B)
- ∪ = union



Examples: Pregnancy, Graduation, ….

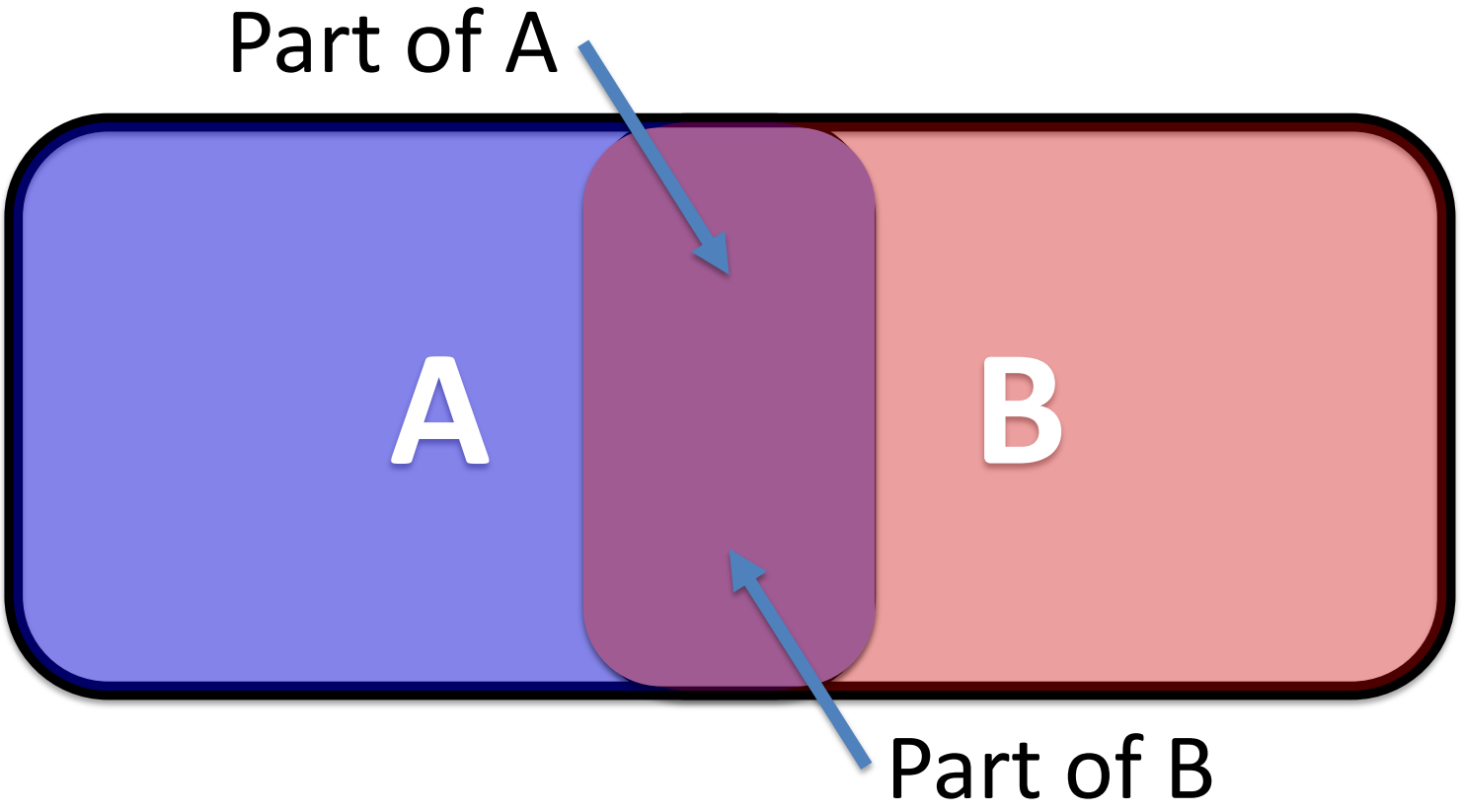
(Only one (mutually exclusive) state at a time)

Not necessarily equal in area

1) The addition rule

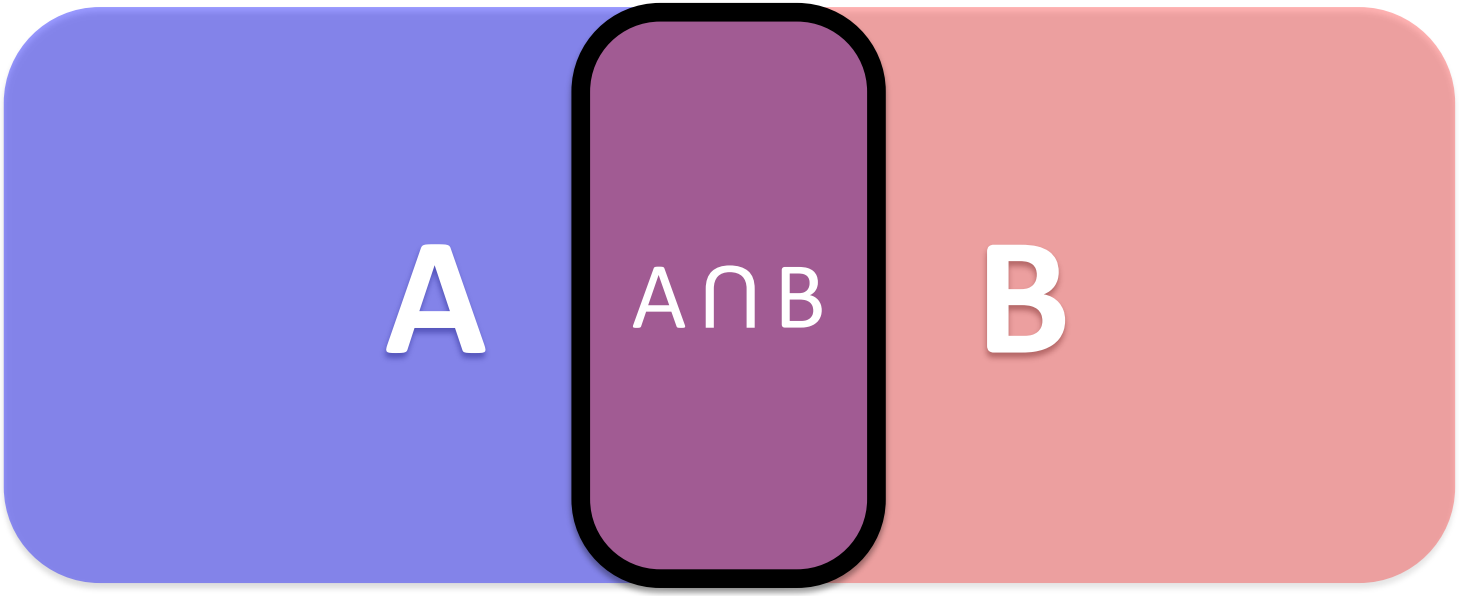
- If events are mutually exclusive
- p (A or B) = p (A $\cup$ B) = p(A) + p(B)
- $\cup$ = union
- If events are not mutually exclusive
- p(A or B) = p (A) + p (B) – p (A and B)

Subtracting avoids double-counting if events are not mutually exclusive



Example: CS major/math major

The intersection



So the mutually exclusive situation is a special case of the more general formulation (no intersection)

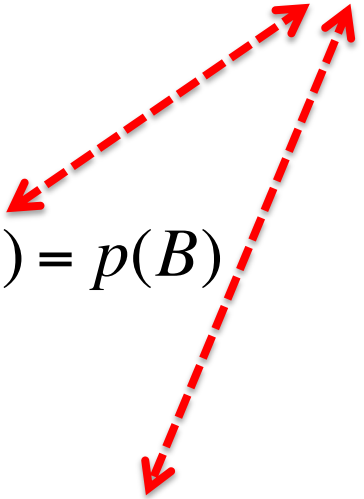
2) The multiplication rule

- If events are independent of each other:
- p (A and B) = p (A ⋂ B) = p (A) * p (B)
- ⋂ = intersection or joint

Statistical independence

Independent $p(A \cap B) = p(A) * p(B)$

Not independent $p(A \cap B) = p(A) * p(B \mid A)$



Independent $p(B \mid A) = p(B)$

Not independent $p(B) \neq p(B \mid A)$

Conditional probability

- How to calculate p (B | A)?
- Probability of B, given A: Solve non-independent case for p (B | A)

Not independent:

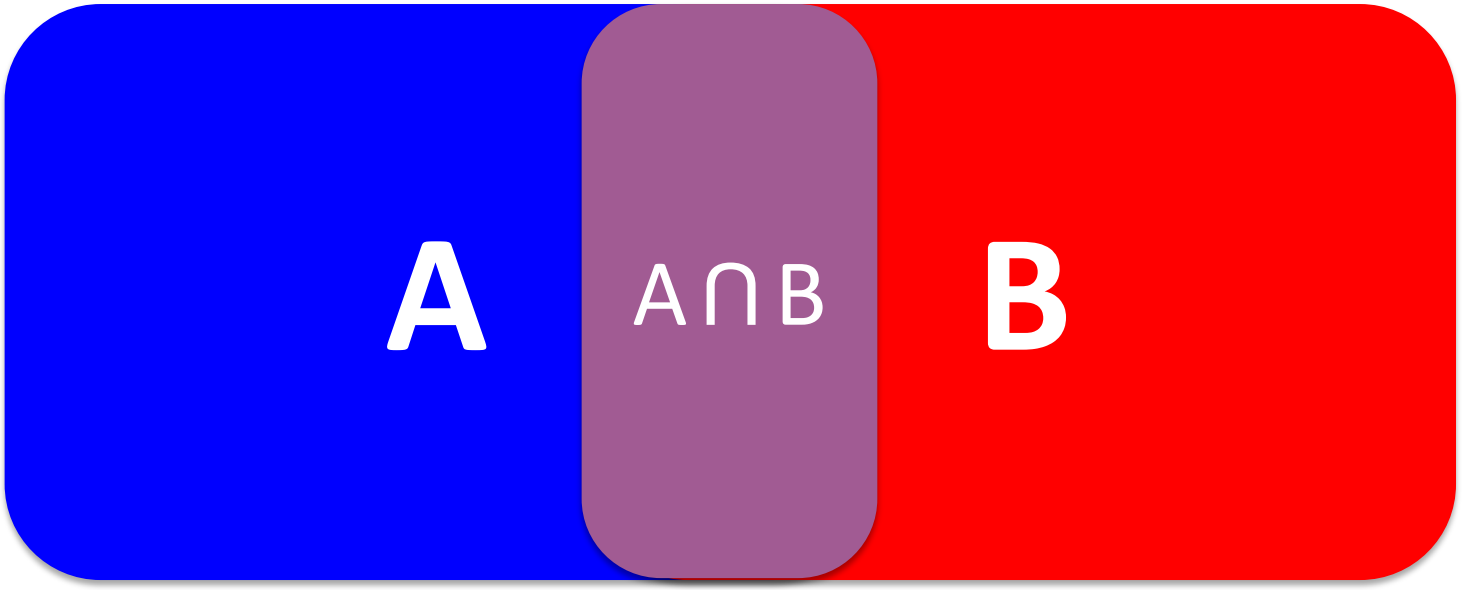
$$p(A \cap B) = p(A) * p(B \mid A)$$

$$\frac{p(A \cap B)}{p(A)} = \frac{p(A) * p(B \mid A)}{p(A)}$$

$$\frac{p(A \cap B)}{p(A)} = p(B \mid A)$$

$$p(B \mid A) = \frac{p(A \cap B)}{p(A)}$$

In conditional probability, the sample space has changed – it shrunk around what happened



We know A happened. So the only B that can still happen is in the intersection with A

$$p(B \mid A) = \frac{p(A \cap B)}{p(A)} = \frac{A \cap B}{A}$$

As a fraction of all the A that can happen

A classic example (Gardner, 1959)

- In situations where events are not independent, conditional probability considerations apply.
- New information about events that happened often changes the sample space.
- Someone has two children, and you know the older is a girl
- What is the probability that all of their children are girls (assuming B/G equally likely, events independent, information is truthful, family was randomly picked, etc.)?

F	S
~~B~~	~~B~~
~~B~~	~~G~~
G	B
G	G

p (AG) = 1 / 4

p (AG | FG)? = 1 / 2

What if you – instead – know that at least one of the children is a girl?

p (AG | OG)?

p (AG | OG) = 1 / 3

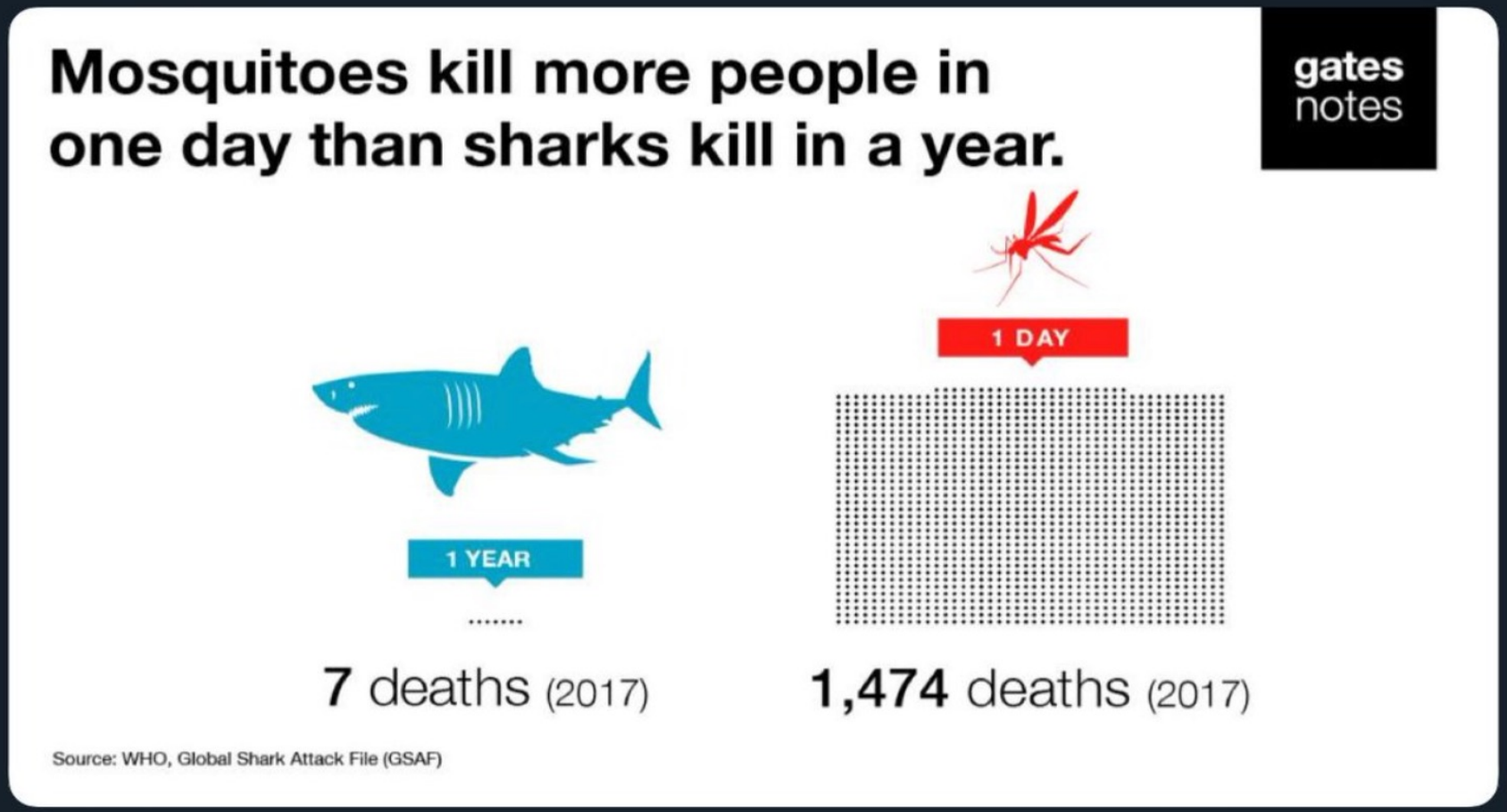
A word of caution:
Probabilistic thinking is not intuitive

Particularly when it comes to conditional probability:

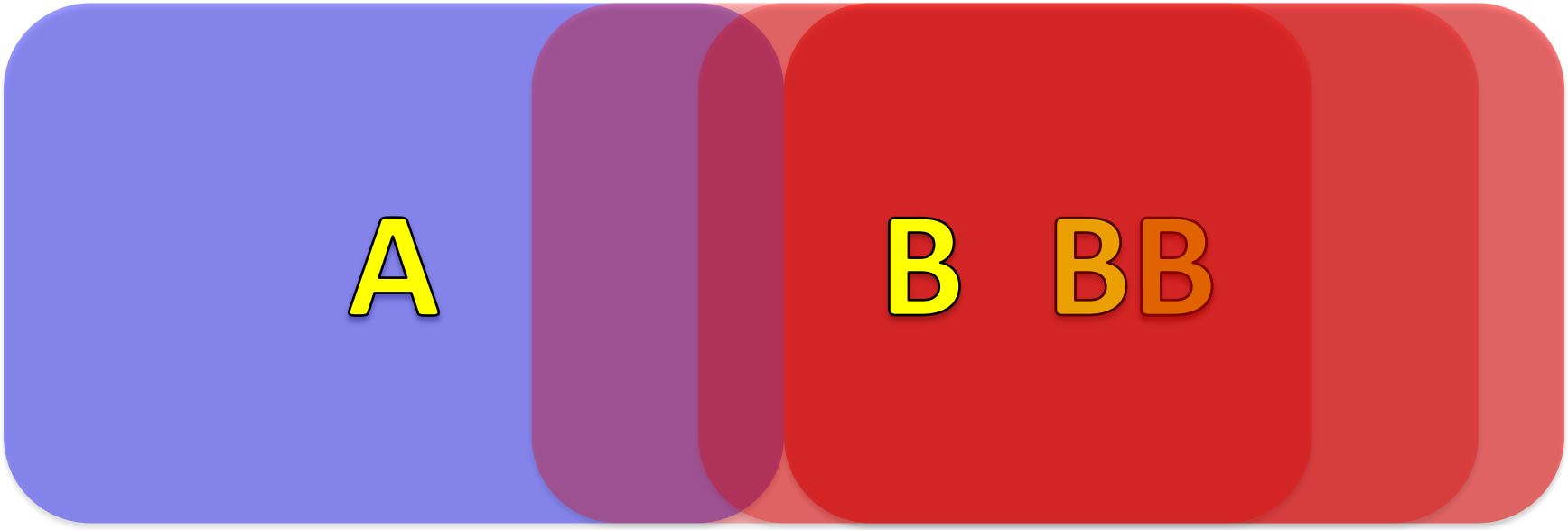
Bill Gates ✓
@BillGates

Why I would rather encounter a shark in the wild than a mosquito: b-gat.es/2XelyuL #MosquitoWeek

Tweet übersetzen



To recap:
The different probability cases



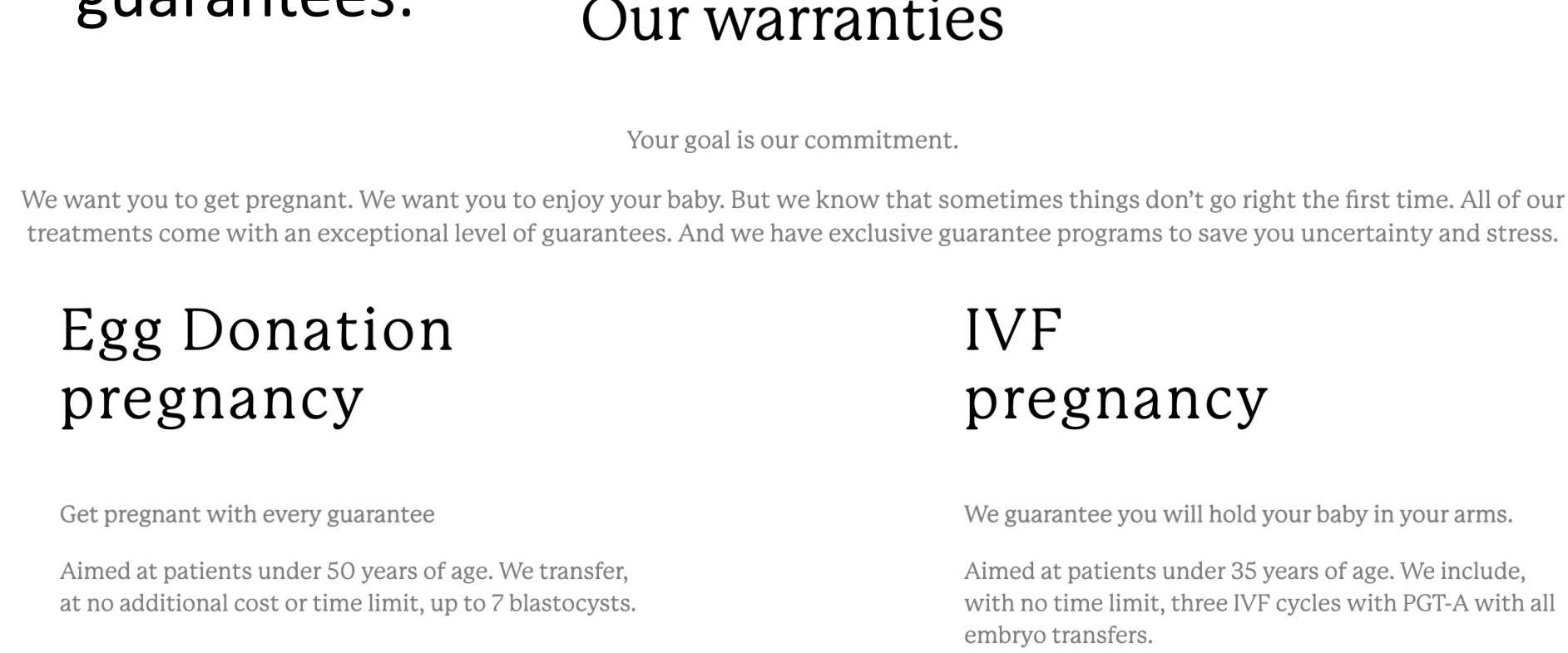
P(A) = 0.5

P(B) = 0.5

P(A
P(B
Co

Moving on to probability distributions

- What we will discuss now is the basis of much of our economy, including all warranties and guarantees:



Our warranties

Your goal is our commitment.

We want you to get pregnant. We want you to enjoy your baby. But we know that sometimes things don't go right the first time. All of our treatments come with an exceptional level of guarantees. And we have exclusive guarantee programs to save you uncertainty and stress.

Egg Donation pregnancy

Get pregnant with every guarantee

Aimed at patients under 50 years of age. We transfer, at no additional cost or time limit, up to 7 blastocysts.

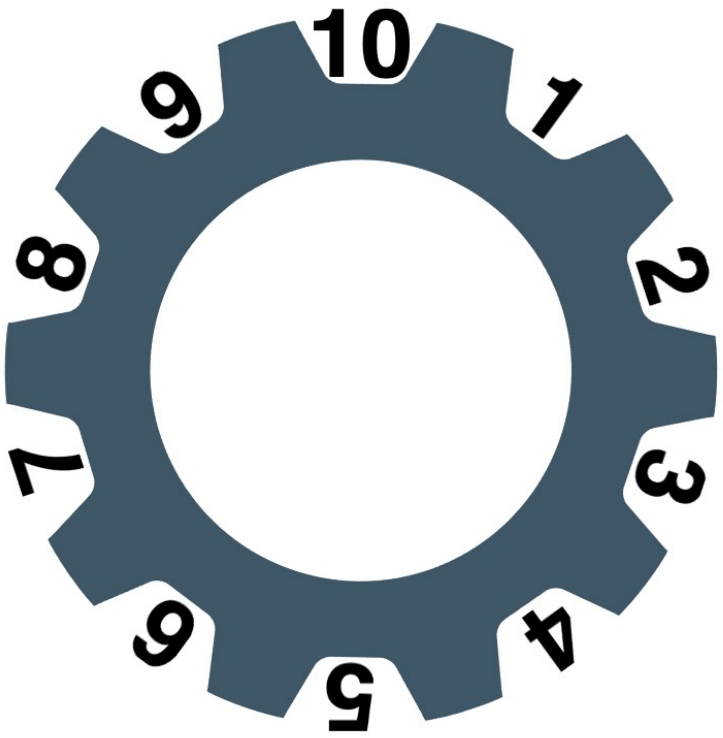
IVF pregnancy

We guarantee you will hold your baby in your arms.

Aimed at patients under 35 years of age. We include, with no time limit, three IVF cycles with PGT-A with all embryo transfers.

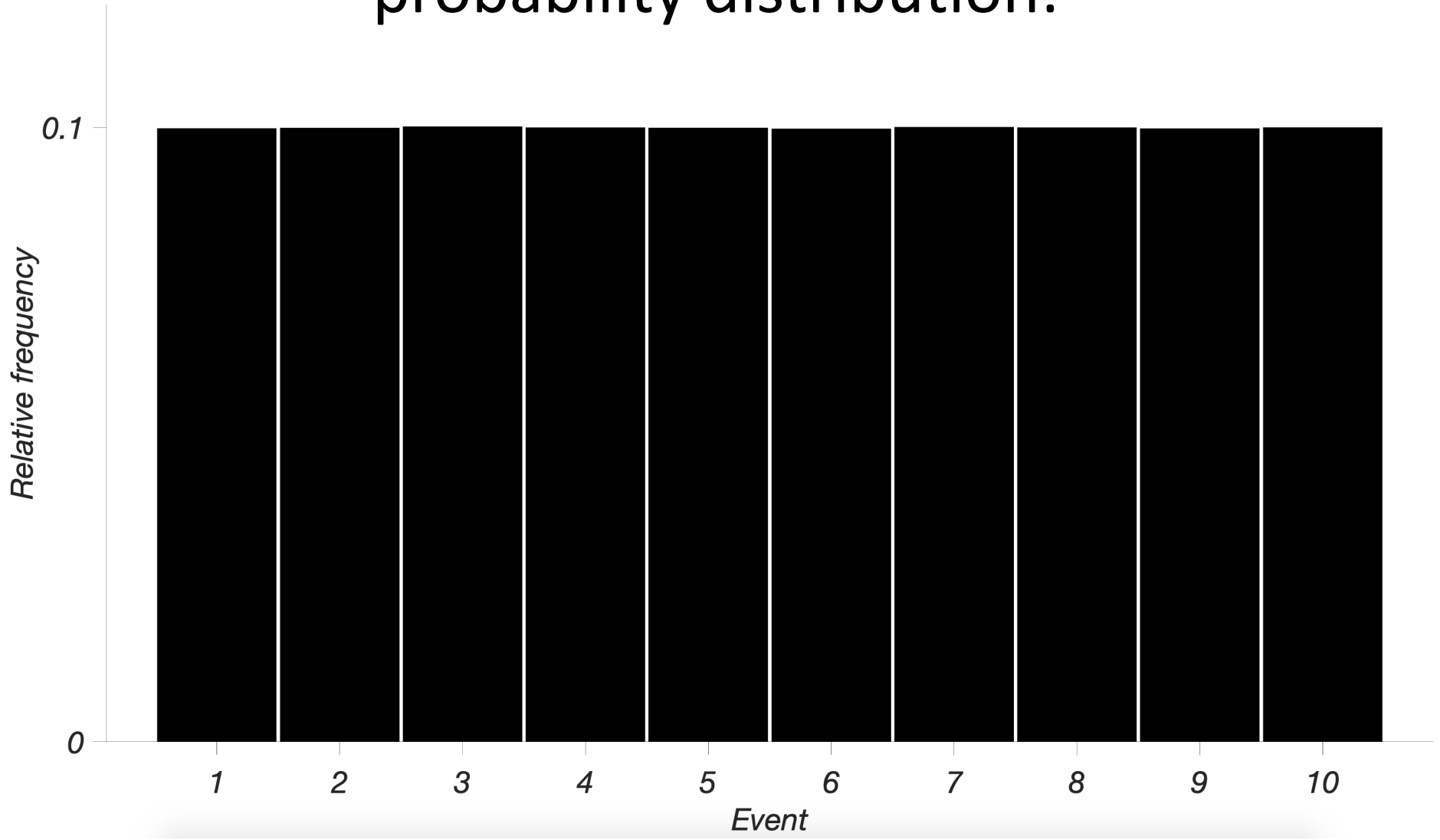
- How can they give such a guarantee?
- We will develop this from simpler principles/examples

The probability space of a random process with mutually exclusive events: The gear of chance



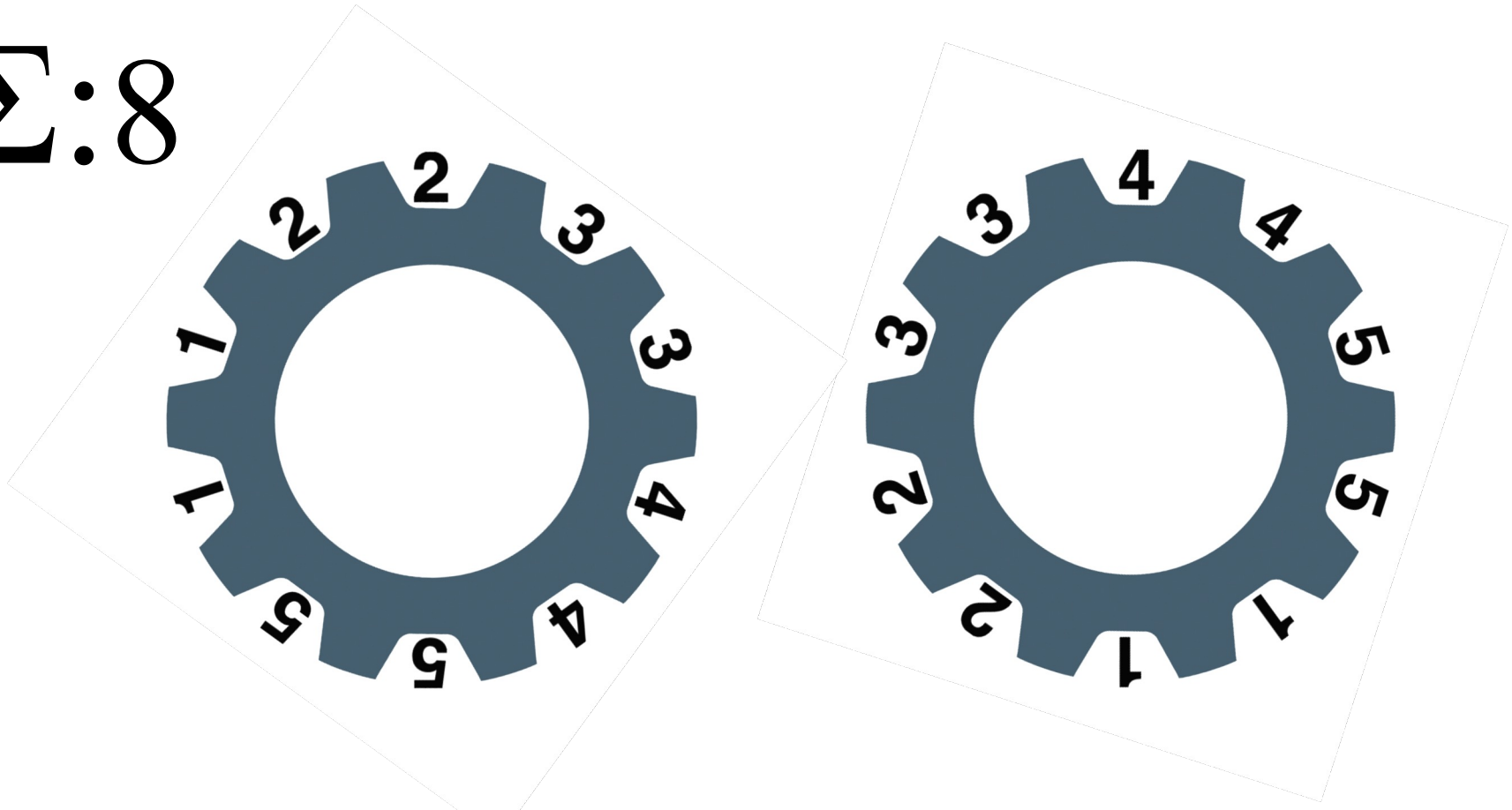
Event	1	2	3	4	5	6	7	8	9	10
Probability	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10

If probabilities are stationary, and the process is repeated often, this leads to a characteristic probability distribution:



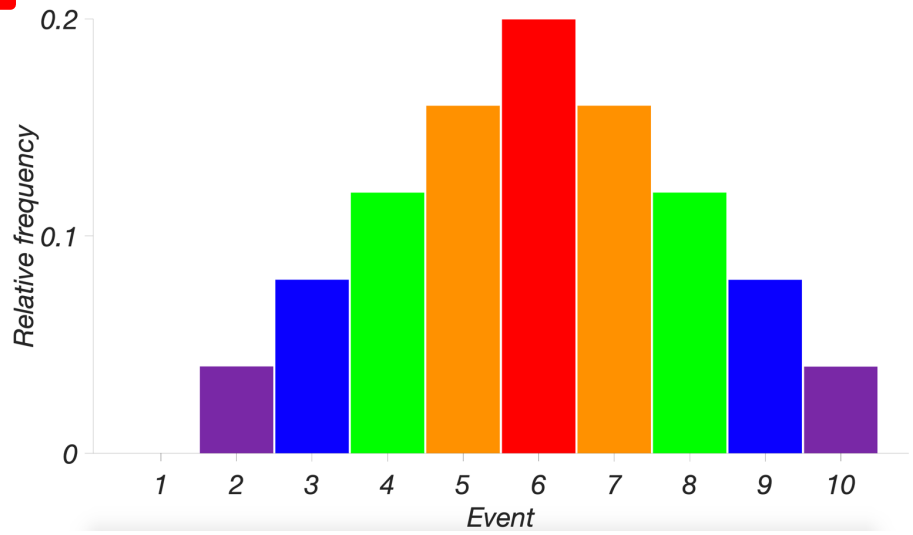
Does the fine structure of the random process matter, i.e. whether the outcomes are produced by the same single wheel as before or by two similar wheels with an individual sample space from 1 to 5, where the overall outcome is the **sum** from the individual wheels?

$\Sigma:8$

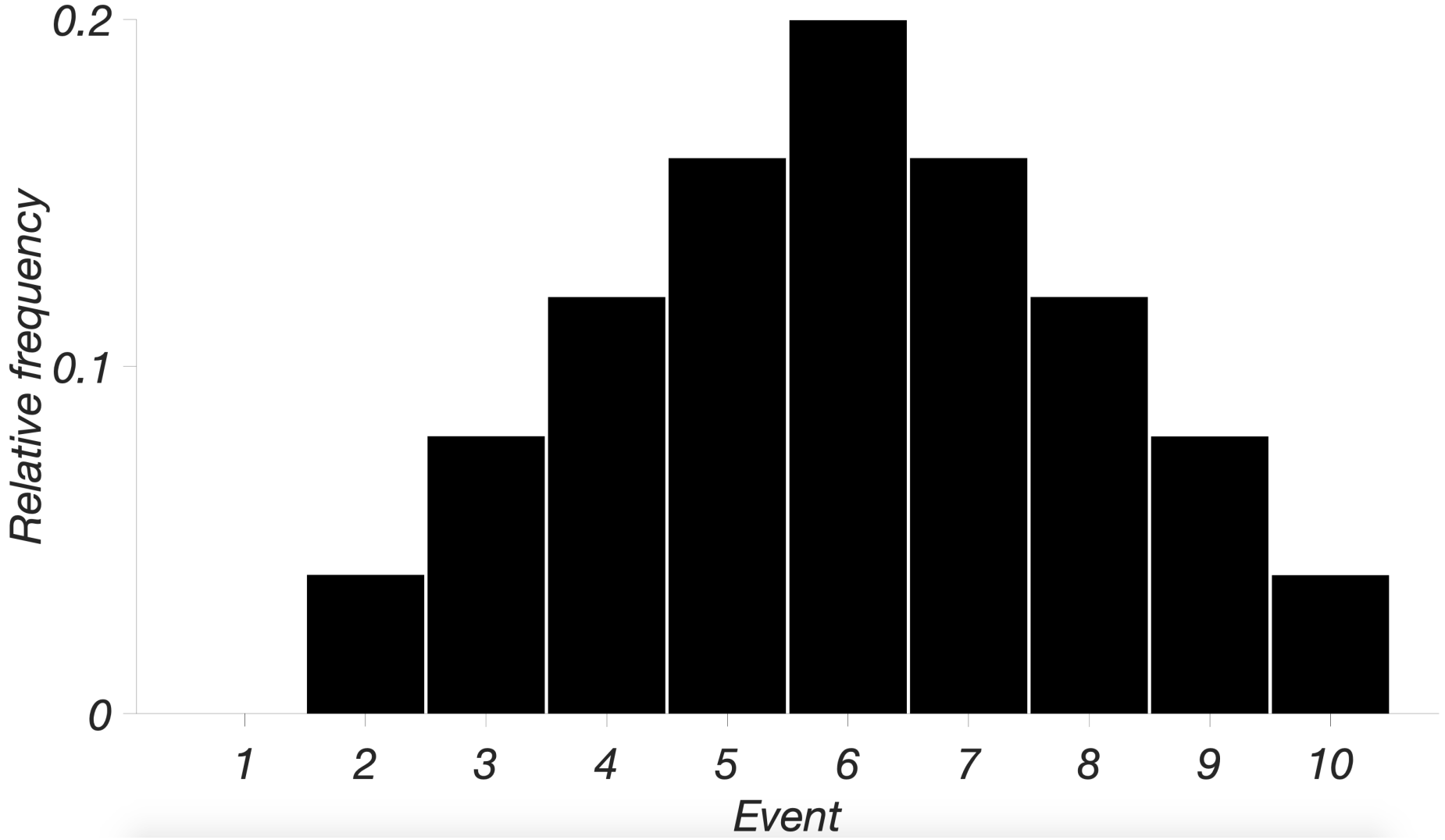


The new sample space:

1,1	2,1	3,1	4,1	5,1
1,2	2,2	3,2	4,2	5,2
1,3	2,3	3,3	4,3	5,3
1,4	2,4	3,4	4,4	5,4
1,5	2,5	3,5	4,5	5,5

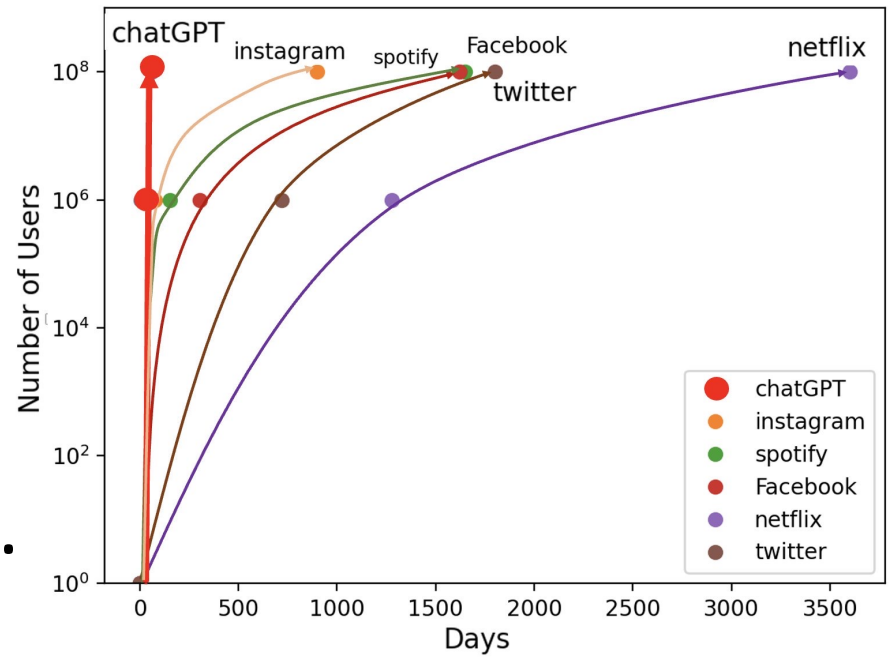


The new – warped - probability distribution



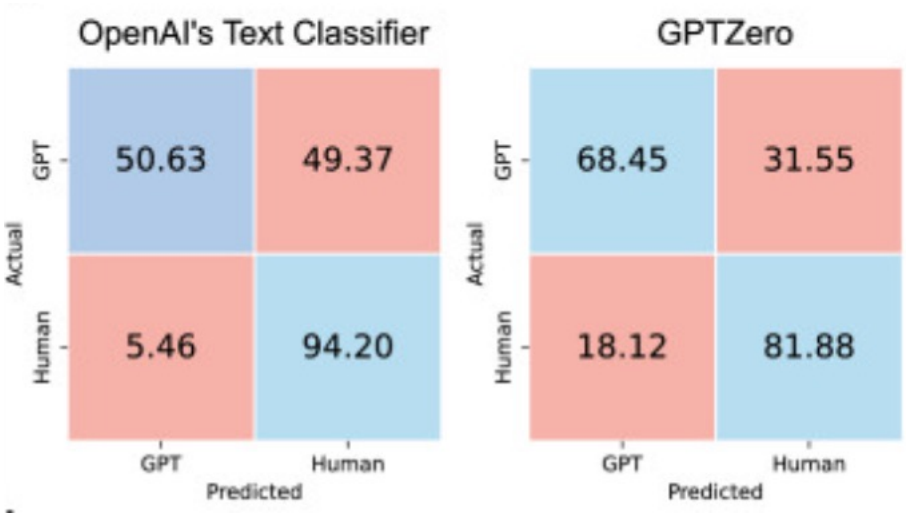
Why this matters: Watermarking in generative AI

- Viable generative AI via large language models exists now (since Nov 2022).
- The widespread adoption of such tools has been extremely rapid.
- This raises concerns about the use of generative AI in educational contexts.
- In particular in classes where AI is demonstrably better than beginners, e.g. creative writing.



A student has been accused of using chatGPT in a writing class

- The student was supposed to write an essay to reflect on “War and Peace”, but their essay was flagged by TurnItIn as likely generated by chatGPT.
- However, the student denies all allegations, and points out that TurnItIn is based on GPTZero (which analyzes writing style), and their false positive rate is demonstrably way too high to be useful:
- The professor sees their point, but is unwilling to let this go, as the student is unable to answer even the most basic questions about the book.



The student faces the tribunal on academic integrity

The tribunal is advised by experts

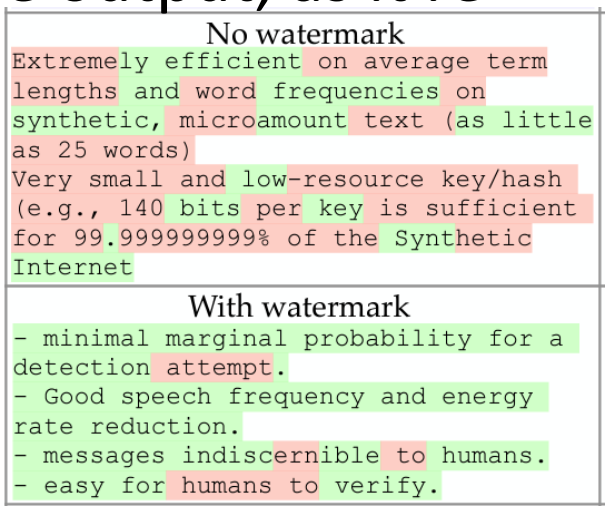
- These experts point out that there are characteristic watermarks embedded in large language models that is designed to detect AI generated text.
- These statistical fingerprints are based on how LLMs work – they are essentially predicting the most likely next word (token) based on the text it has just seen, so essentially conditional probability:

He had a wonderful…

p (the |He had a wonderful) = 0
p (life |He had a wonderful) = 0.5
p (theory |He had a wonderful) = 0.05
p (insight |He had a wonderful) = 0.1
p (nightmare | He had a wonderful) = 0.001
p (time | He had a wonderful) = 0.2

How the watermarking works

- The LLM has a corpus of n words.
- Based on the last word of the text (e.g. “wonderful”) as a seed, this corpus is randomly divided into a “whitelist” and a “blacklist” of size eg n/2 (no quality decrement bc synonyms)
- Which words are on the whitelist and which on the blacklist changes randomly with each word of the output, as it re-seeds the random number generator.
- The LLM will not pick blacklisted words.
- However, sometimes, the next word is so overwhelmingly likely, e.g.:
  p (concern |To whom it may) = 0.99
  that the blacklist is ignored.

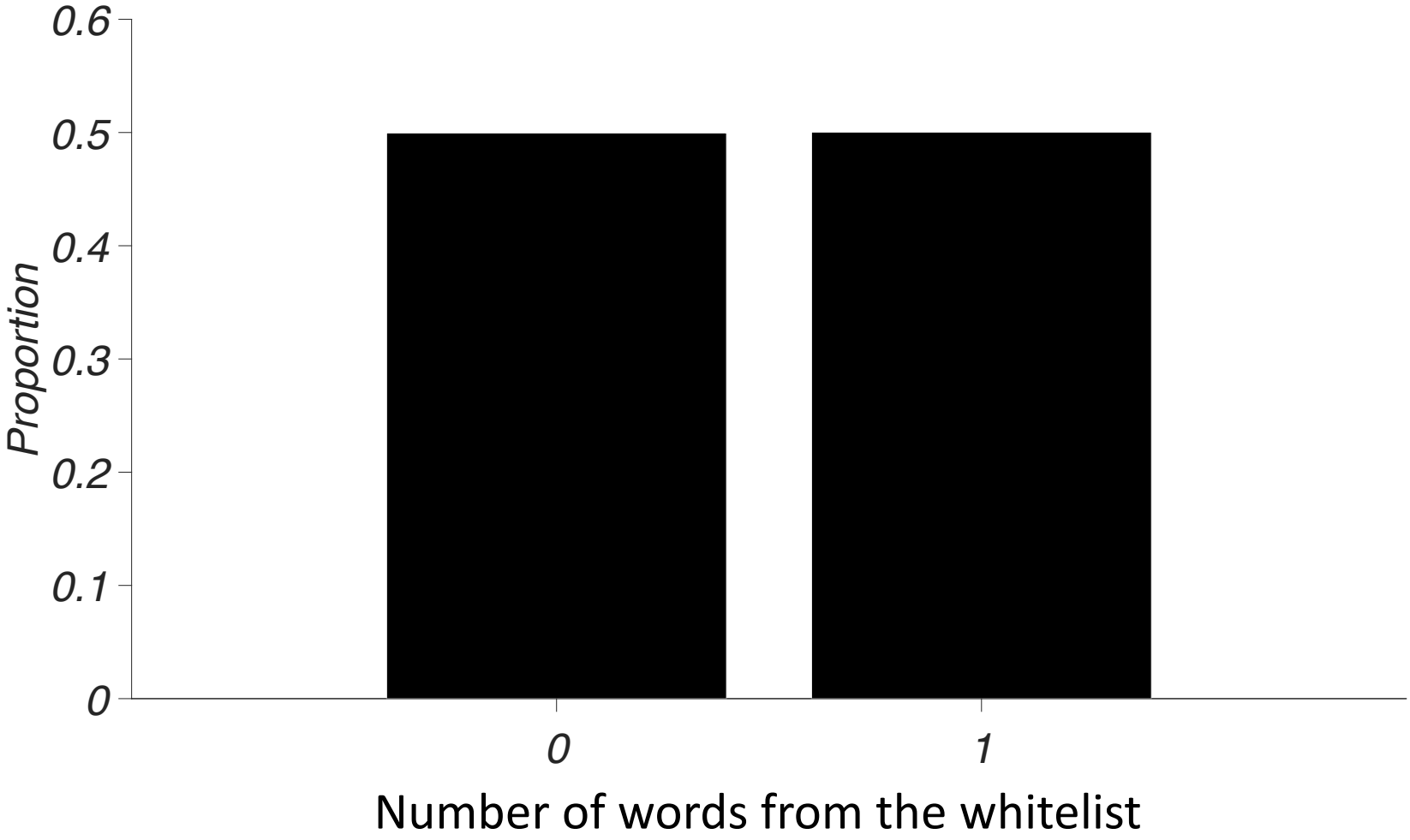


- So an AI generated text will have more words from the whitelist than one would expect from chance (this can be calibrated such that the chance performance – a given word coming from the whitelist if not using an LLM – is 0.5)

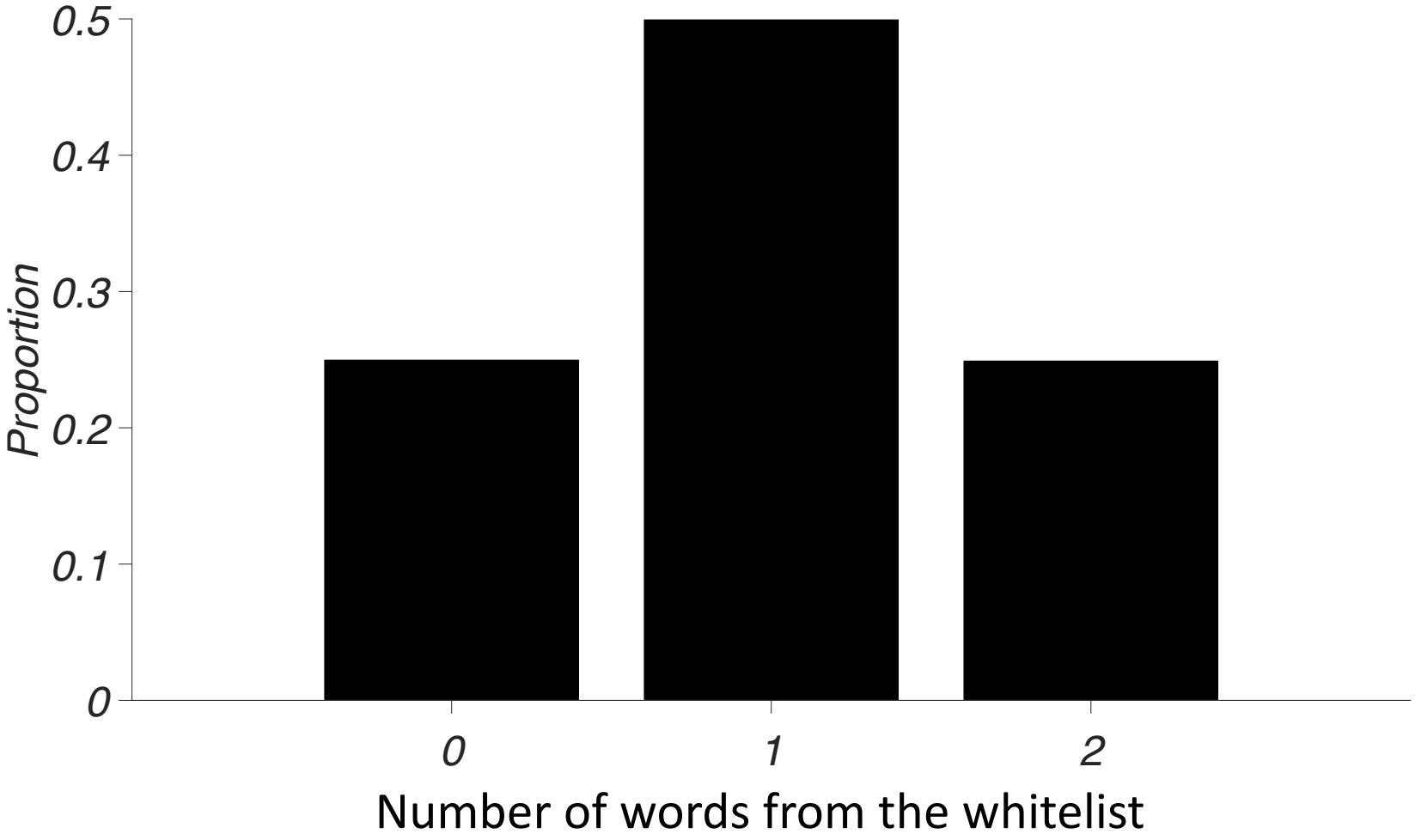
The tribunal announces the findings of their watermarking analysis

- The magistrate points out that the student essay contained 305 words, 211 of which were from the whitelist, more than the 152.5 one would have expected by chance, given the calibration of watermarking.
- The student is stunned by these revelations, but quickly regains their composure and points out that this translates to about 70% of words from the whitelist. In other words, one would expect 5 out of 10 words from the whitelist if chatGPT was not used. By chance, they ended up with 7 out of 10 words from the whitelist.
- That's the defense. They emphasize that given that this is the sum total of the evidence, the tribunal has no hard evidence and given that there are many doubts, has to release them.
- You are a data scientist advising the tribunal. What's your take?

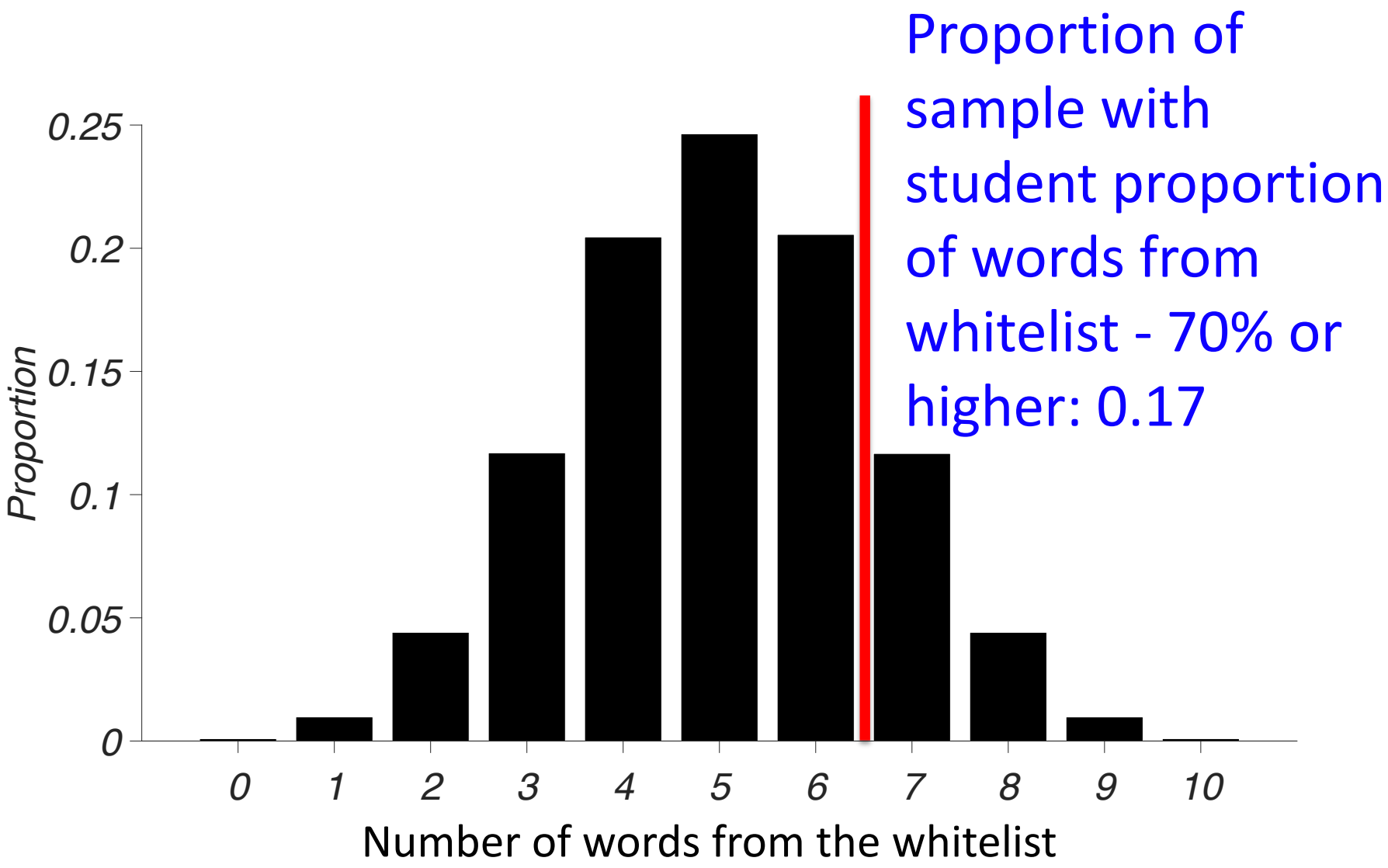
What can one expect if writing one word / token (if one is not using chatGPT)



What can one expect if writing two words / tokens (if one is not using chatGPT)

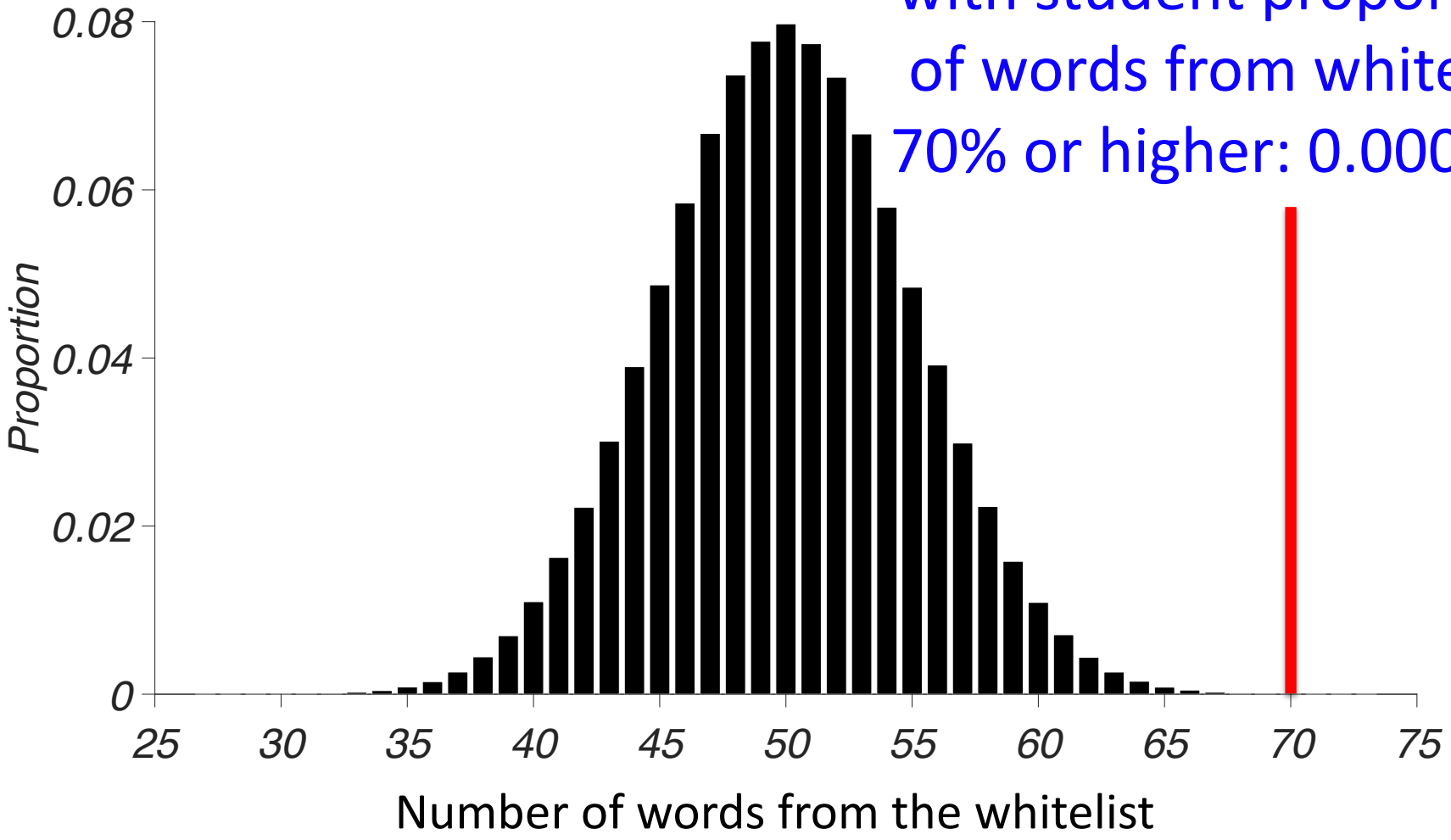


What can one expect if writing 10 words / tokens (if one is not using chatGPT)

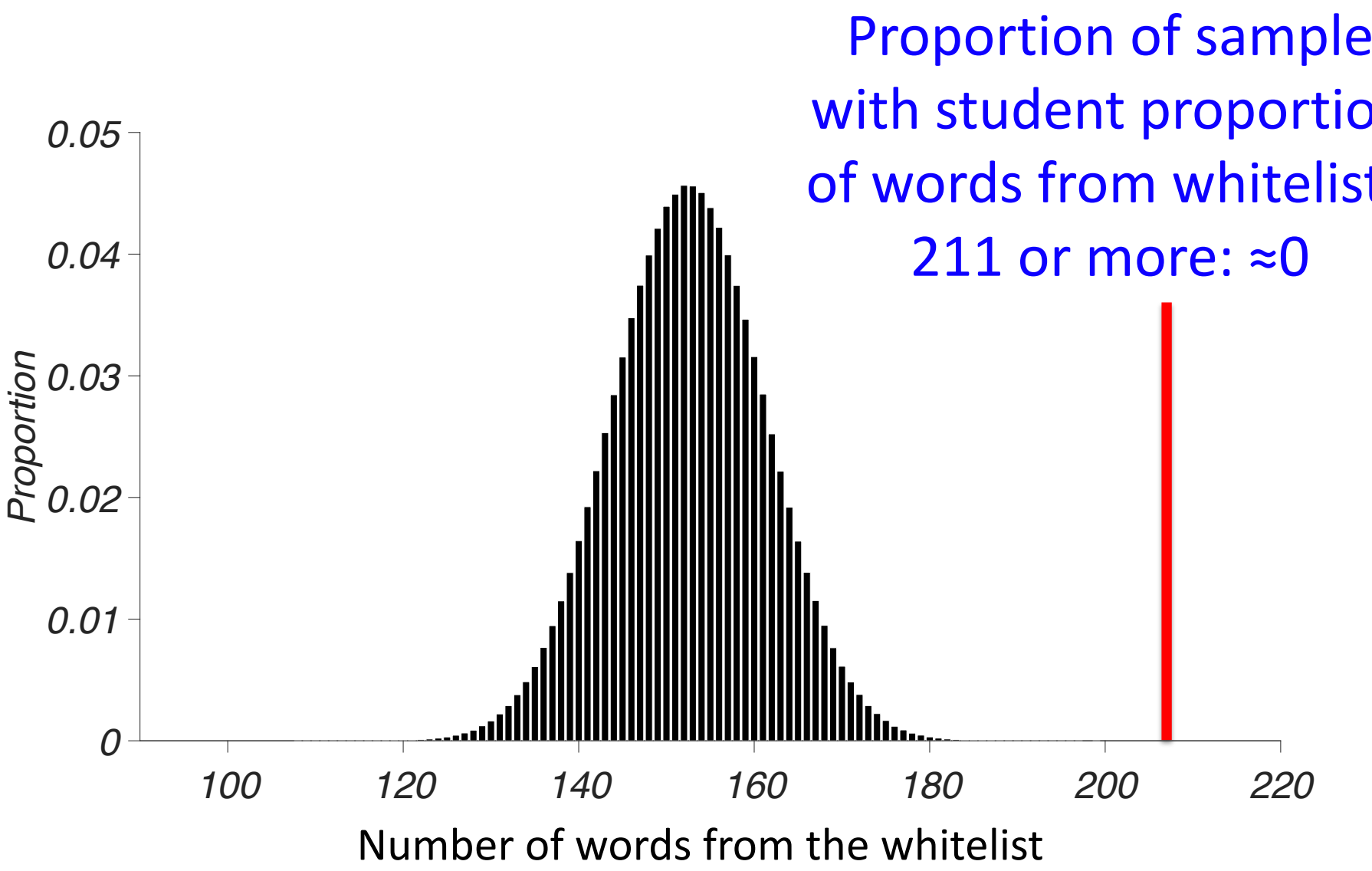


What can one expect if writing 100 words / tokens (if one is not using chatGPT)

Proportion of sample with student proportion of words from whitelist 70% or higher: 0.000034



What can one expect if writing 305 words / tokens (the student number, if one is not using chatGPT)

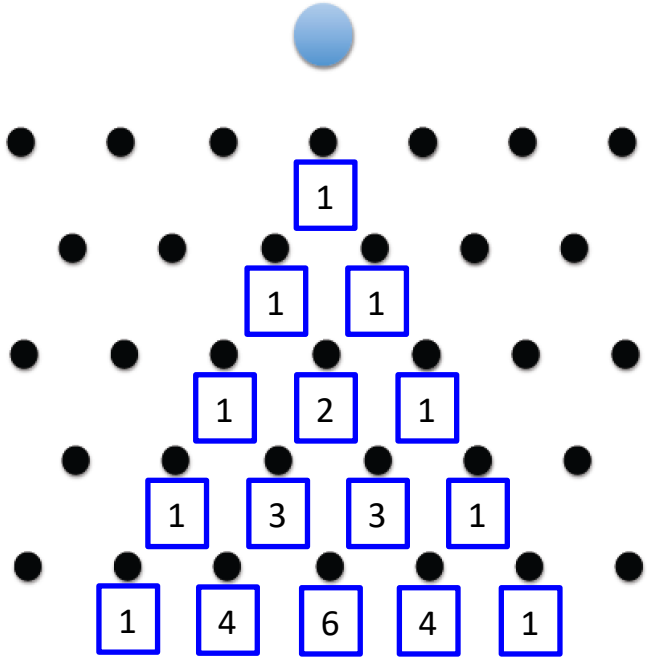


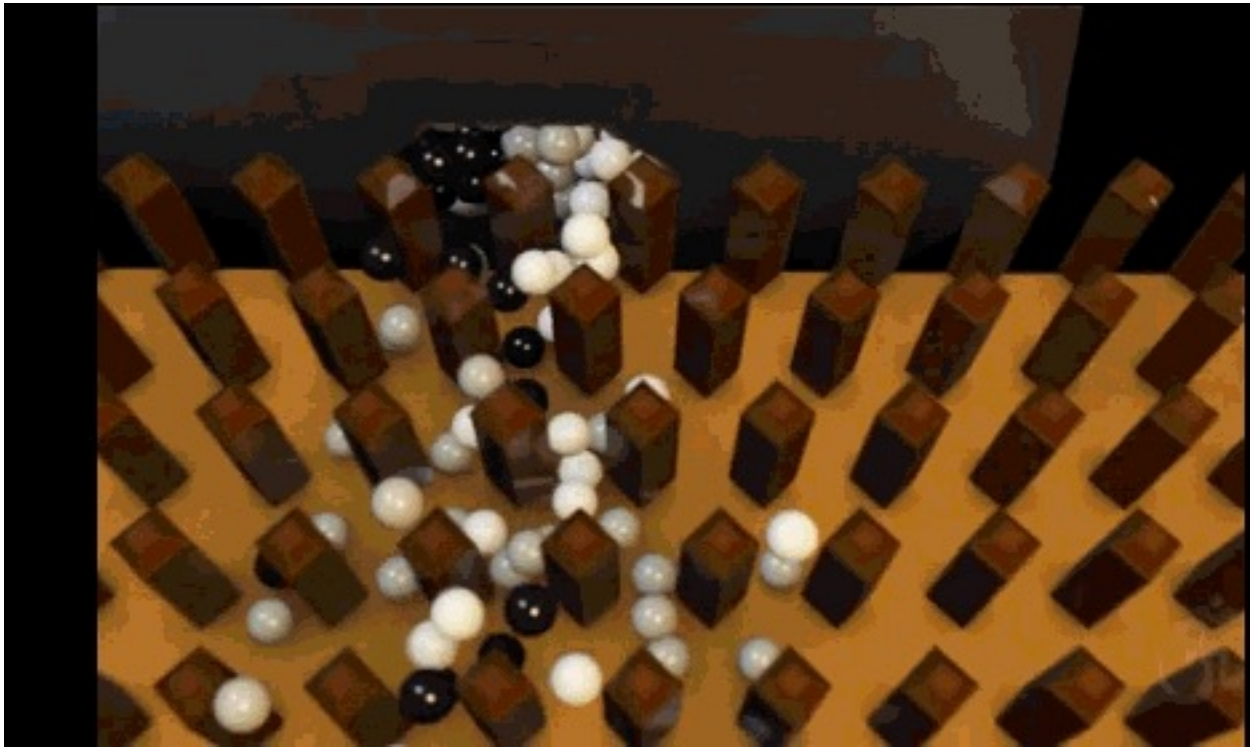
Resolution and Relevance

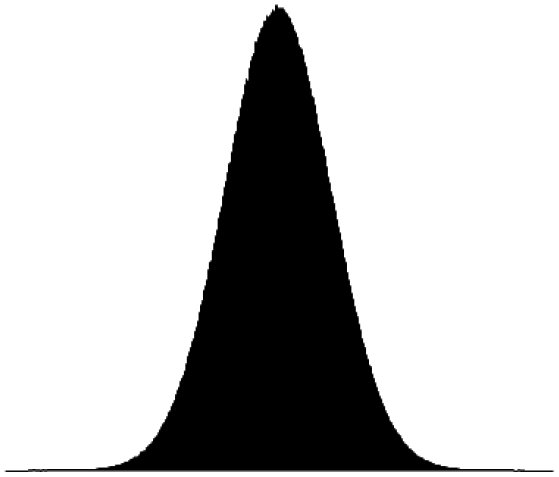
- After vehemently denying any wrongdoing, confronted with this overwhelming evidence, the student finally confesses that they used chatGPT to write the essay.
- The student is forced to accept the judgment of the tribunal and face the consequences of their actions.
- This basic principle – that clear expectations emerge from complete uncertainty – goes well beyond cheating detection in student essays.
- When combining many independent events, is the basis of many industries, including gambling, insurance, actuaries, etc. – and even the fundamental principle underlying science itself (more on this next time).
- It is not the case that the house always wins. It does not win every game. But the game is designed so the house wins in the long run when many independent games are played.
- This also explains why insurance usually doesn't cover for correlated events (e.g. war, natural disasters, etc.) – this math no longer works.

How normal distributions come about in general

The distribution is a physical manifestation of the relative number of possible paths to reach a given location, as described by Pascal's triangle







Necessary:

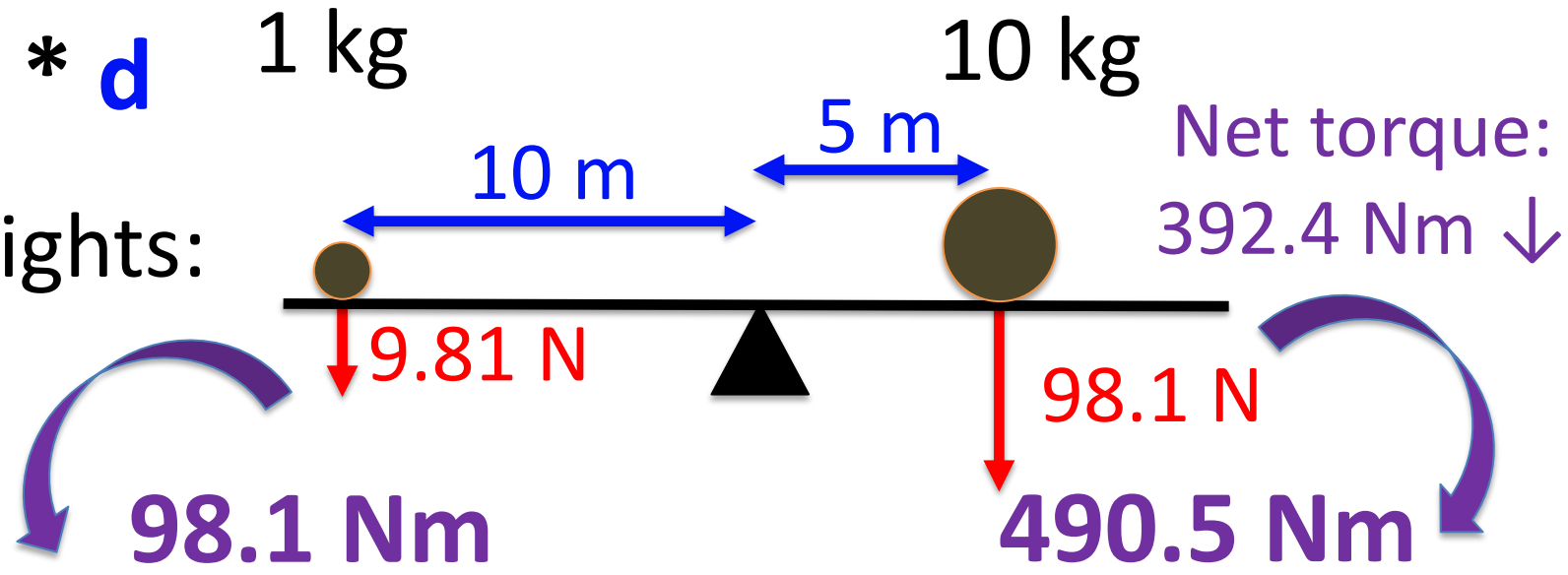
1) Factors combining independently
2) Lots of them

Probability distributions are characterized by “moments”

- The concept of a “moment” comes to us from physics – and latin: “momentum”
- Specifically movement about some axis of rotation
- In other words, the turning force around a pivot:

For example: Weights on a seesaw

M = F * d

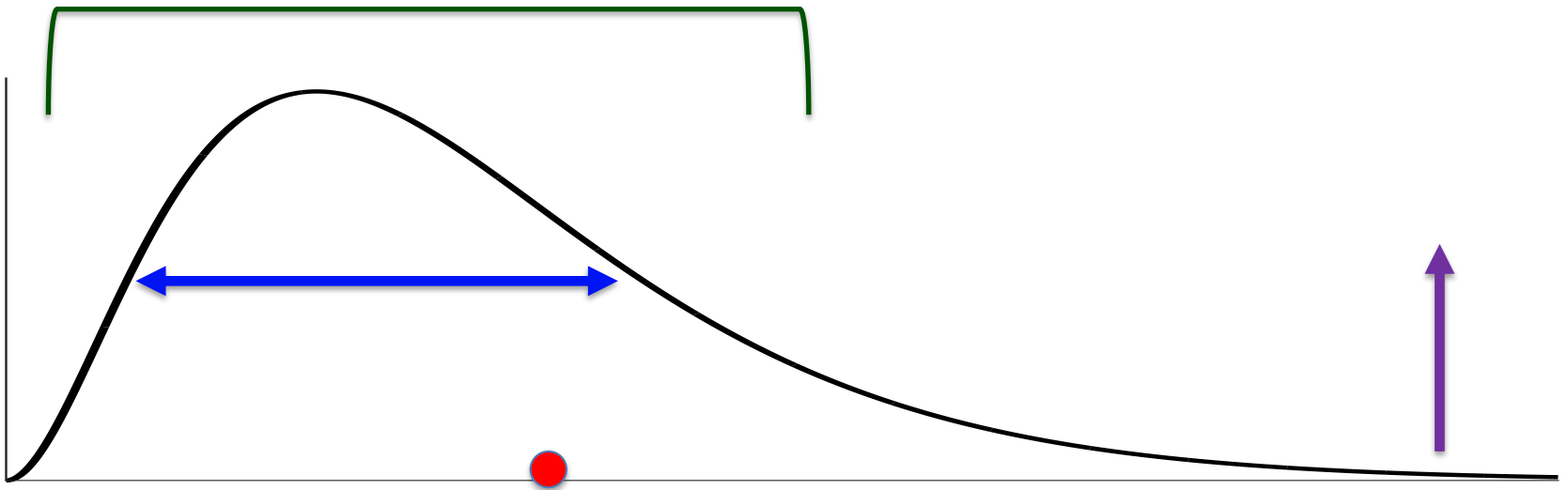


To balance (=equal and opposite moments), the fulcrum has to move to the mean of the weight distribution

Moments of probability distributions

- Moments characterize the shape of probability distributions
- Physics analogy: The distribution of the probability mass
- The kth moment of RV *X* with density function *f(x)* is given by
- 0th moment: 1 ("Total mass") = 1
- 1st moment: Expected Value ("Location")
- 2nd moment: Variance ("Scale")
- 3rd moment: Shape ("Skewness")
- 4th moment: Tailedness ("Kurtosis")

$$m_k = E\left(X^k\right) = \int_{-\infty}^{\infty} x^k f(x)dx$$



A closer look at different kinds of moments

$$m_k = E\left(X^k\right) = \int_{-\infty}^{\infty} x^k f(x) dx$$

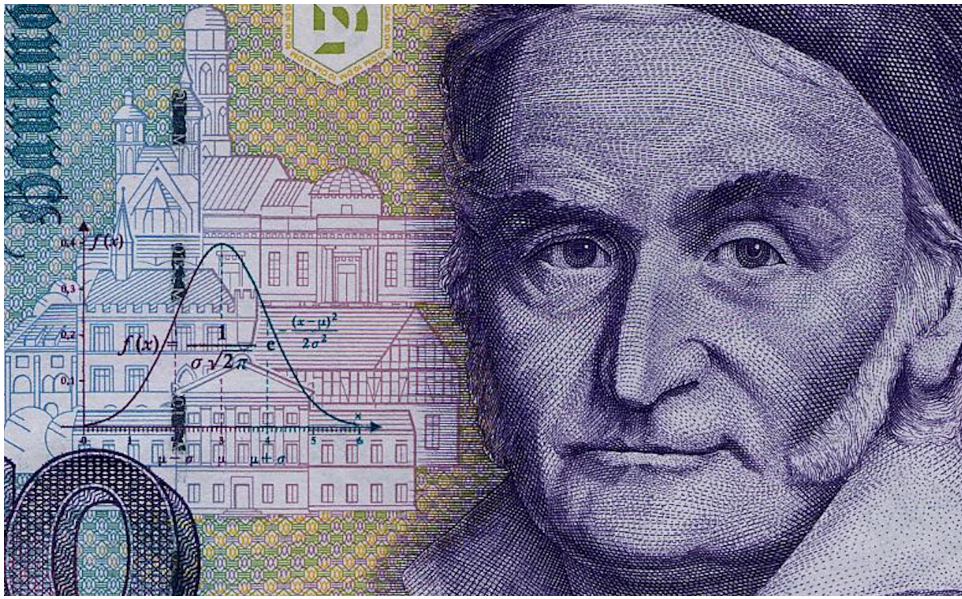
k[th] raw moment: $m_k = E\left(X^k\right)$

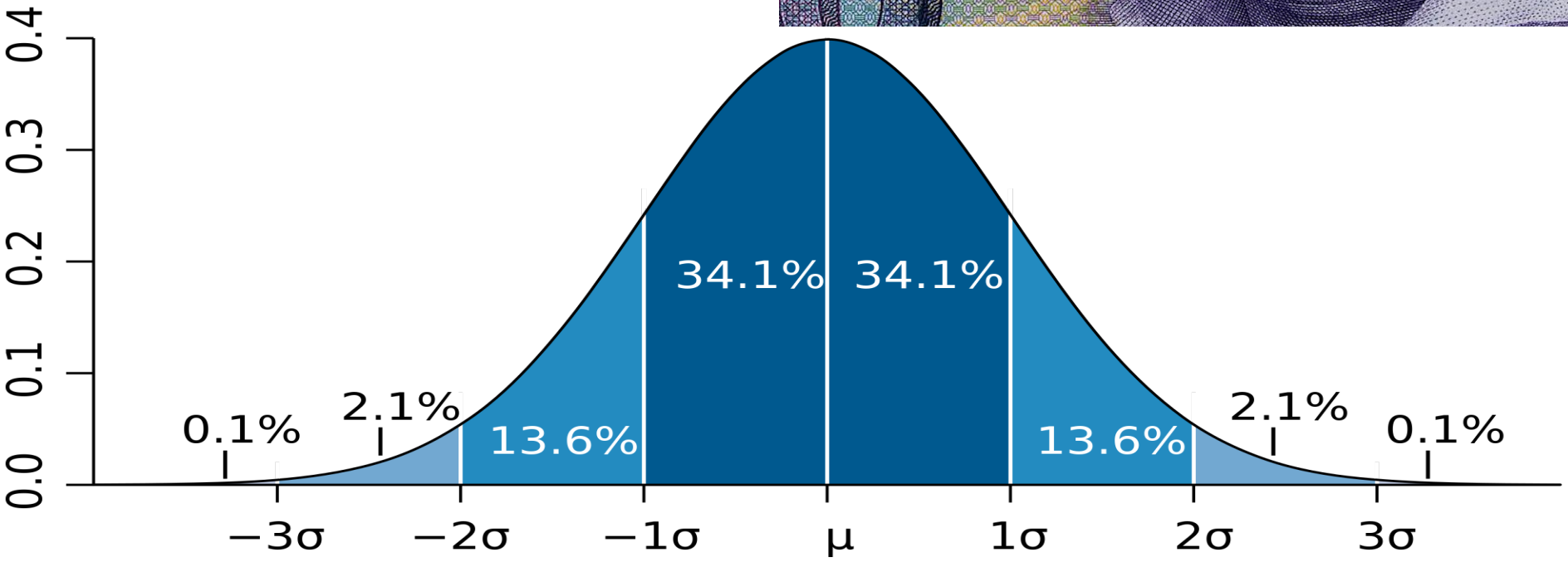
k[th] centralized moment: $m_k = E(X - \mu)^k$

k[th] standardized moment: $m_k = E\left(\frac{(X-\mu)}{\sigma}\right)^k$

The normal distribution is fully determined by the first two moments and is well behaved: Extreme events are very unlikely

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$





Finally: How is the sample space mapped to real numbers (probabilities)?

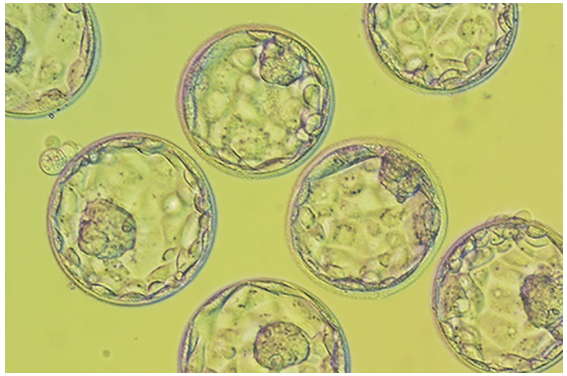
- We need to distinguish the concept of a **constant**, a **variable** and a **random variable**.
- $x + 3 = 10$
- Is x a variable?
- No, x is a constant in disguise, the only value that x can take for this equation to be true is 7.
- $y = x + 5$
- Now, x is a genuine variable. It can take many values, and y is a function of this variable x: y(x)
- If this is a variable, what is a random variable?
- A variable that takes random values?

Random variables

- A **random variable (RV)** is actually a *function* that maps all elements of the sample space **Ω** to real numbers.
- They are often written in upper case letters, e.g. X
- However, this is very confusing, as matrices are also conventionally written in capital letters.
- To avoid this confusion, we will use upper case fraktur to denote random variables: $\mathfrak{X}$
- But functions are deterministic. Where does the randomness come from?
- From the sample space, which comes from the random experiment.
- Thus, random variables are the standard way to model random processes in Data Science

A practical example to illustrate the versatility of random variables to model uncertainty

- You run an IVF clinic, as many medical schools do.
- You run a study to determine how many viable embryos are yielded by a medicated cycle.
- You run your clinic as normal, but record the outcome of every cycle.
- The outcome of any given cycle is uncertain.



Possible outcomes: {0,1,2,3,4,5,6,7,8,9}

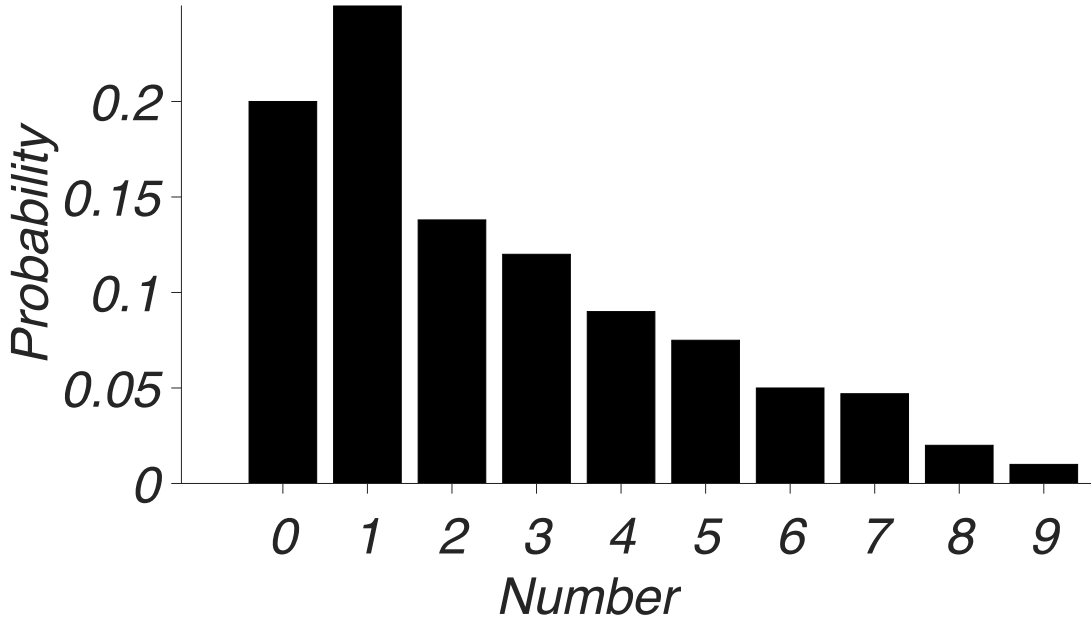
Sample space Ω: {0,1,2,3,4,5,6,7,8,9}

As this is a random process, we characterize it with a random variable that maps this sample space to probabilities

We don't know its properties a priori, so we do this empirically, from 1000 cycles during this time

Number of viable embryos per cycle as a **discrete** random variable $\mathfrak{X}$

Number of viable embryos (x)	Cycles
0	200
1	250
2	138
3	120
4	90
5	75
6	50
7	47
8	20
9	10



Random variables have **moments** that characterize them.

The first moment is the mean, or expected value:

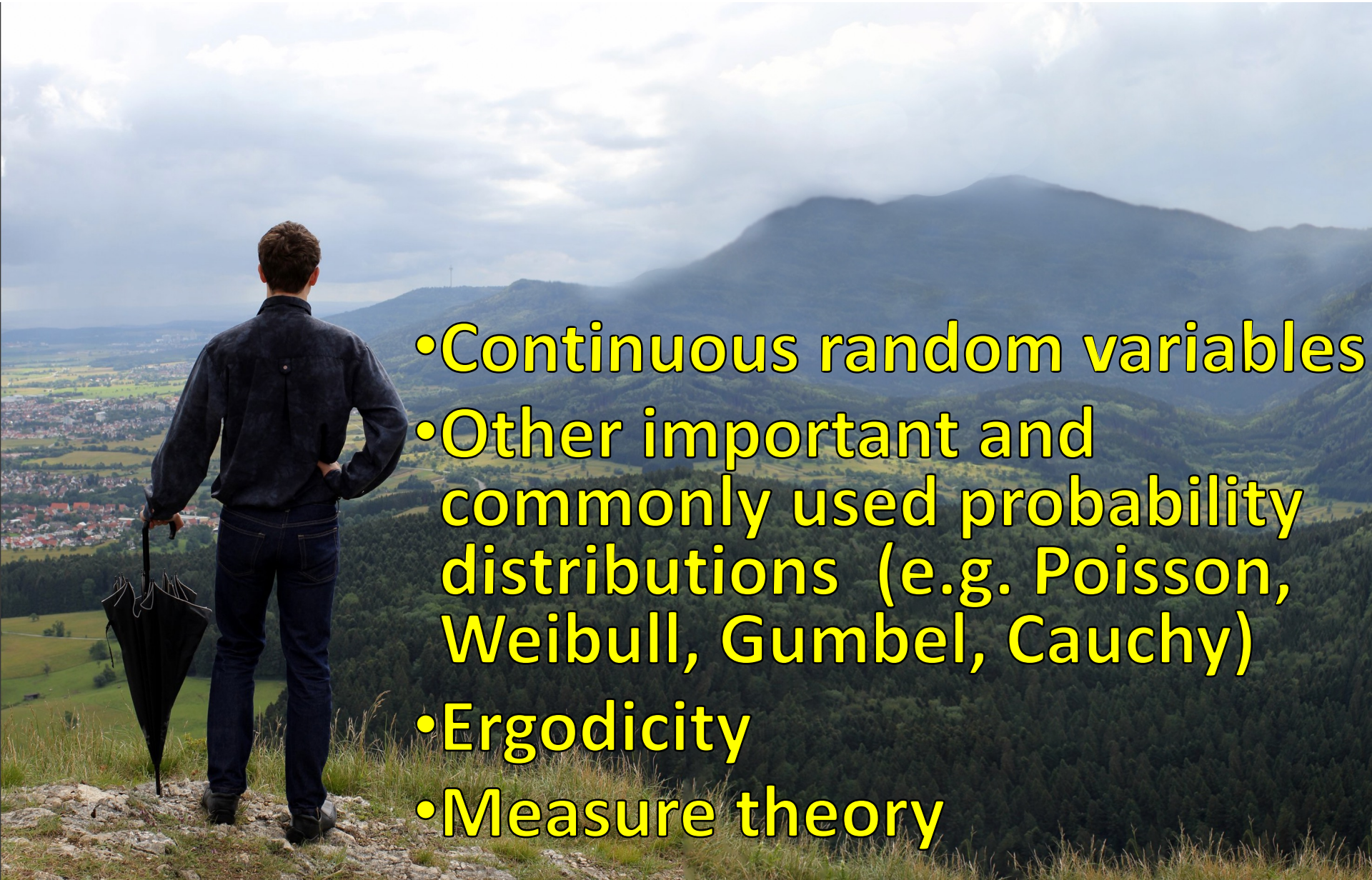
$E(\mathfrak{X}) = \text{sum}(x*p(x)) = 2.5$

We can use the RV to calculate:

$p(\mathfrak{X}>5) = 0.05 + 0.047 + 0.02 + 0.01$

$p(\mathfrak{X}>5) = 0.127$

Kn-ow(e)-l-edge



- Continuous random variables
- Other important and commonly used probability distributions (e.g. Poisson, Weibull, Gumbel, Cauchy)
- Ergodicity
- Measure theory

Thanks for the clarification questions and the engagement

Now:

Are there any more general questions?