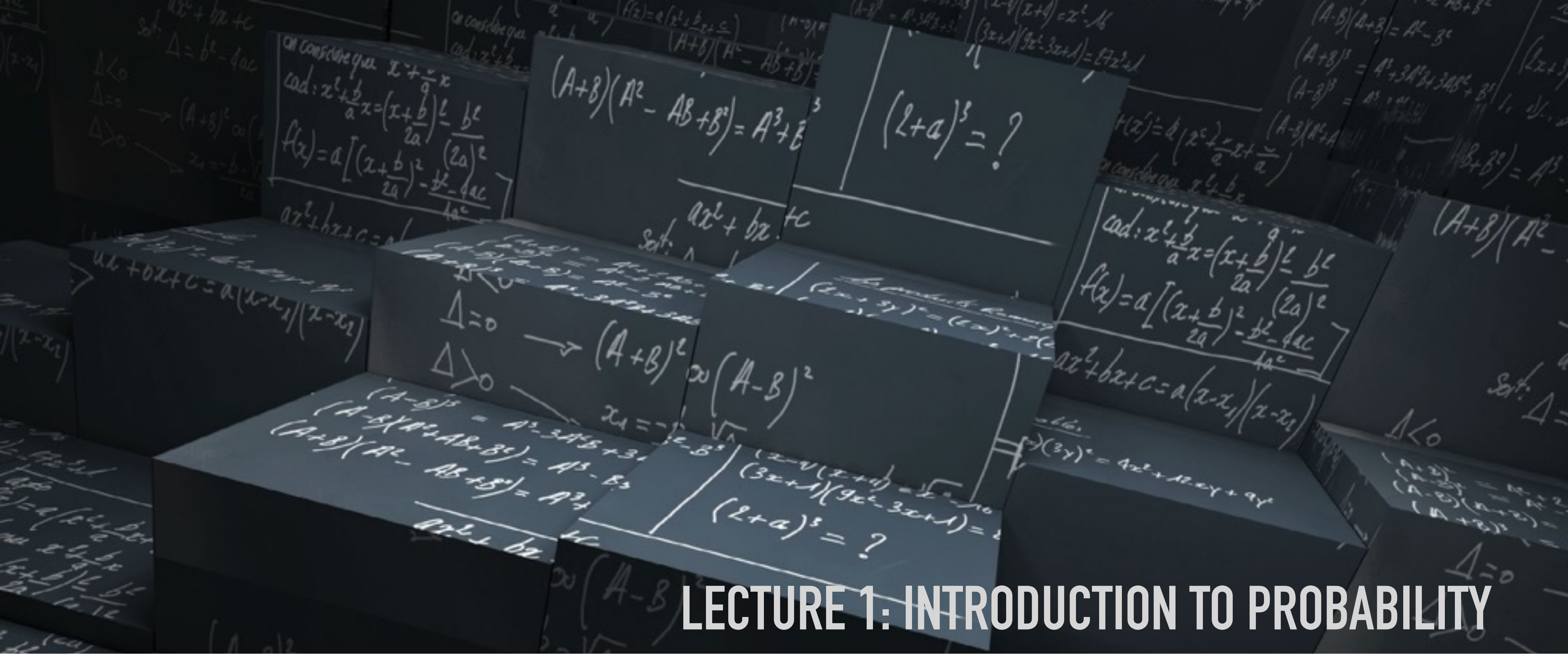


LECTURE 1: INTRODUCTION TO PROBABILITY

STATISTICAL ANALYSIS
IN EXPERIMENTAL PARTICLE PHYSICS

Kai-Feng Chen
National Taiwan University

PROBABILITY?

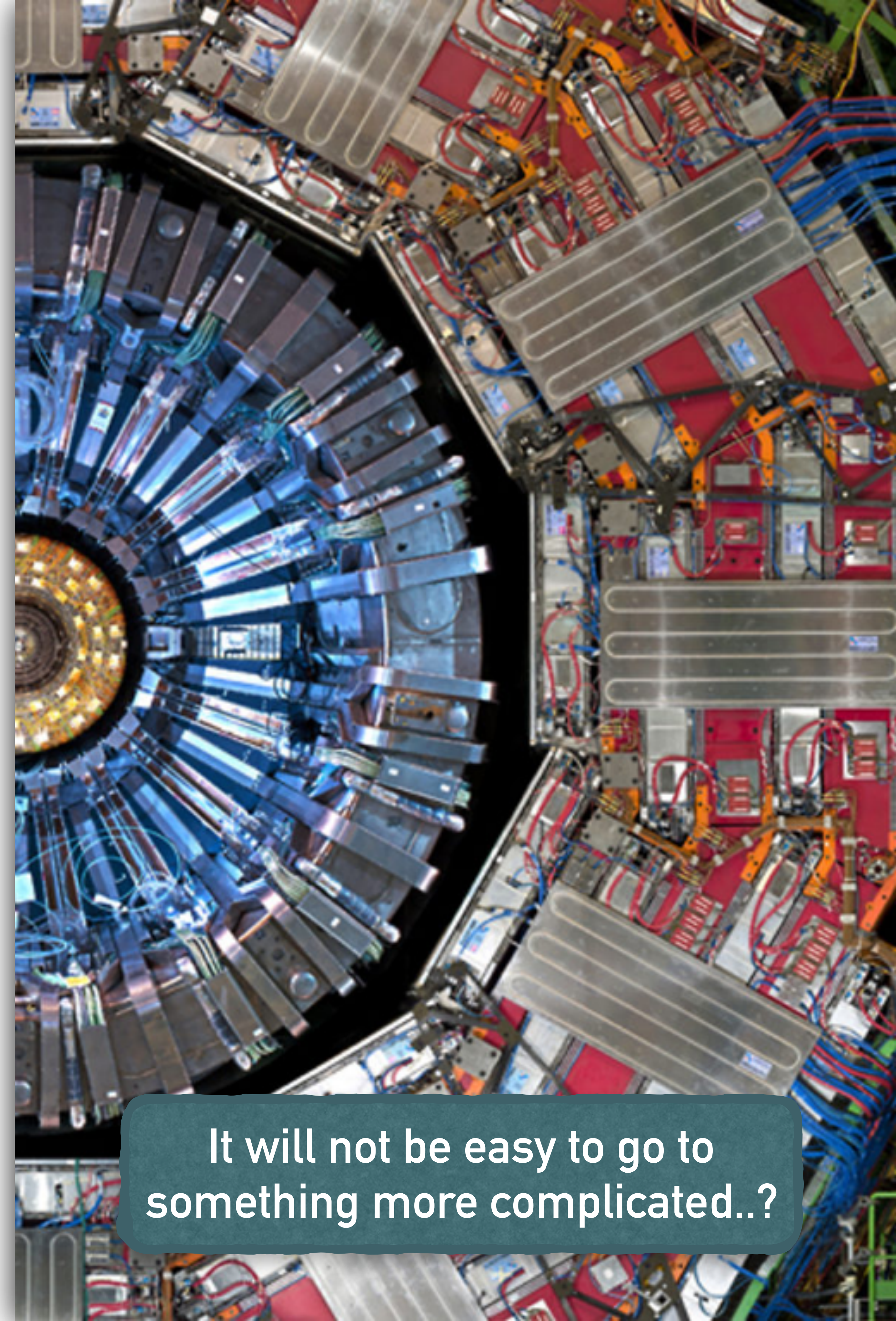
- Probability is the measure of the likelihood that an event will occur, quantified as a number between [0,1]:
  - 0 = impossibility; 1 = certainty.
  - The higher the probability of an event, the more likely it is that the event will occur.
- Classical probability is defined by

$$P = \frac{\text{Number of favorable cases}}{\text{Number of total cases}}$$

  - Assuming all of the cases are *equally possible*.
  - This only works for **discrete cases** rigorously.
  - Problems in continuous cases (*to be discussed*).

P = 1/2 for a fair coin

It it not difficult to discuss a simple system like dice.



It will not be easy to go to something more complicated..?

WARM-UP: PROBABILITY AND COMBINATORIAL

- Complex cases are managed via combinatorial analysis;
- Reduce the event of interest into elementary equiprobable cases.
  For example, sum of two dice below.
- Set algebra may be applied to the sample space.

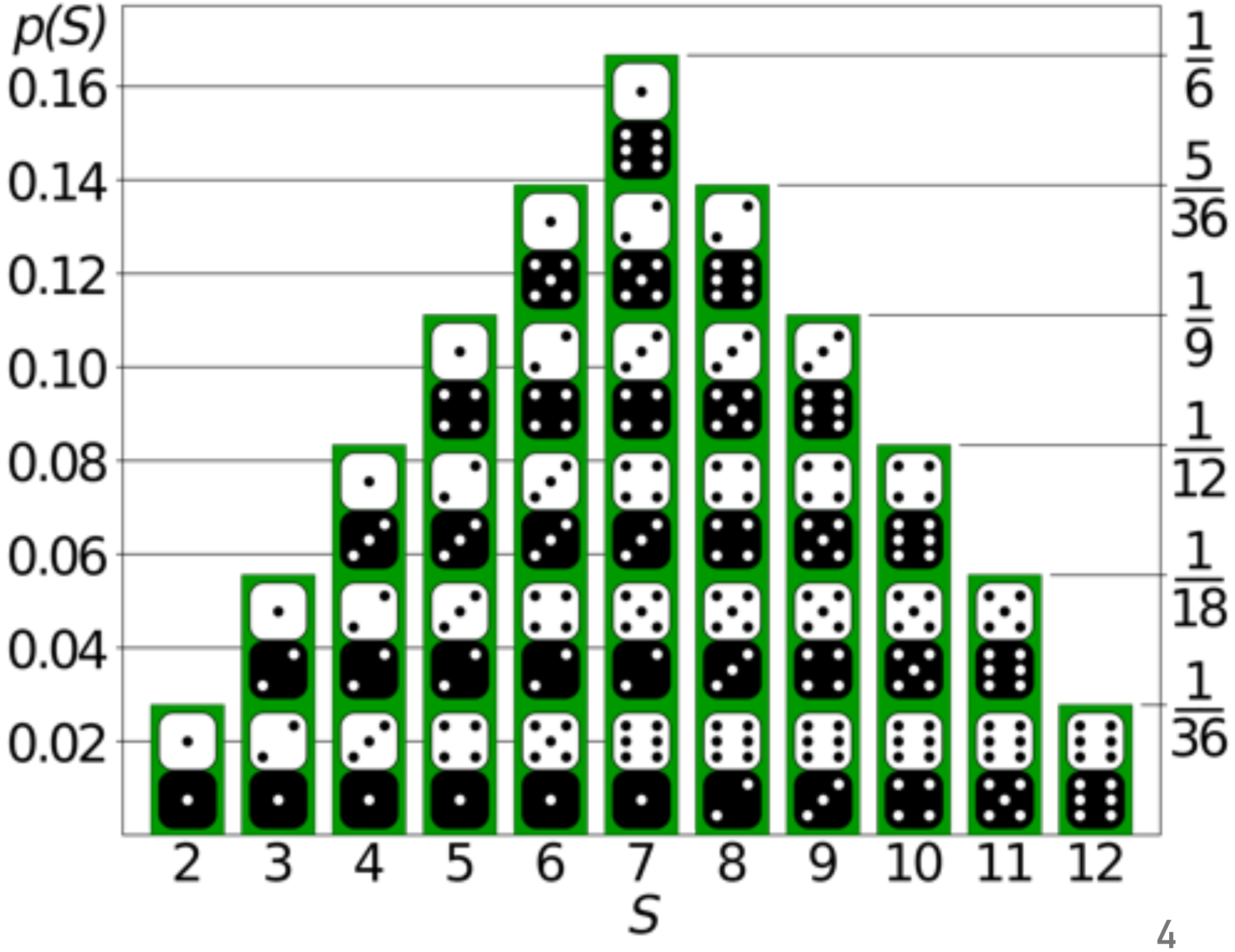
2 = (1,1)

3 = (1,2) or (2,1)

4 = (1,3) or (2,2) or (3,1)

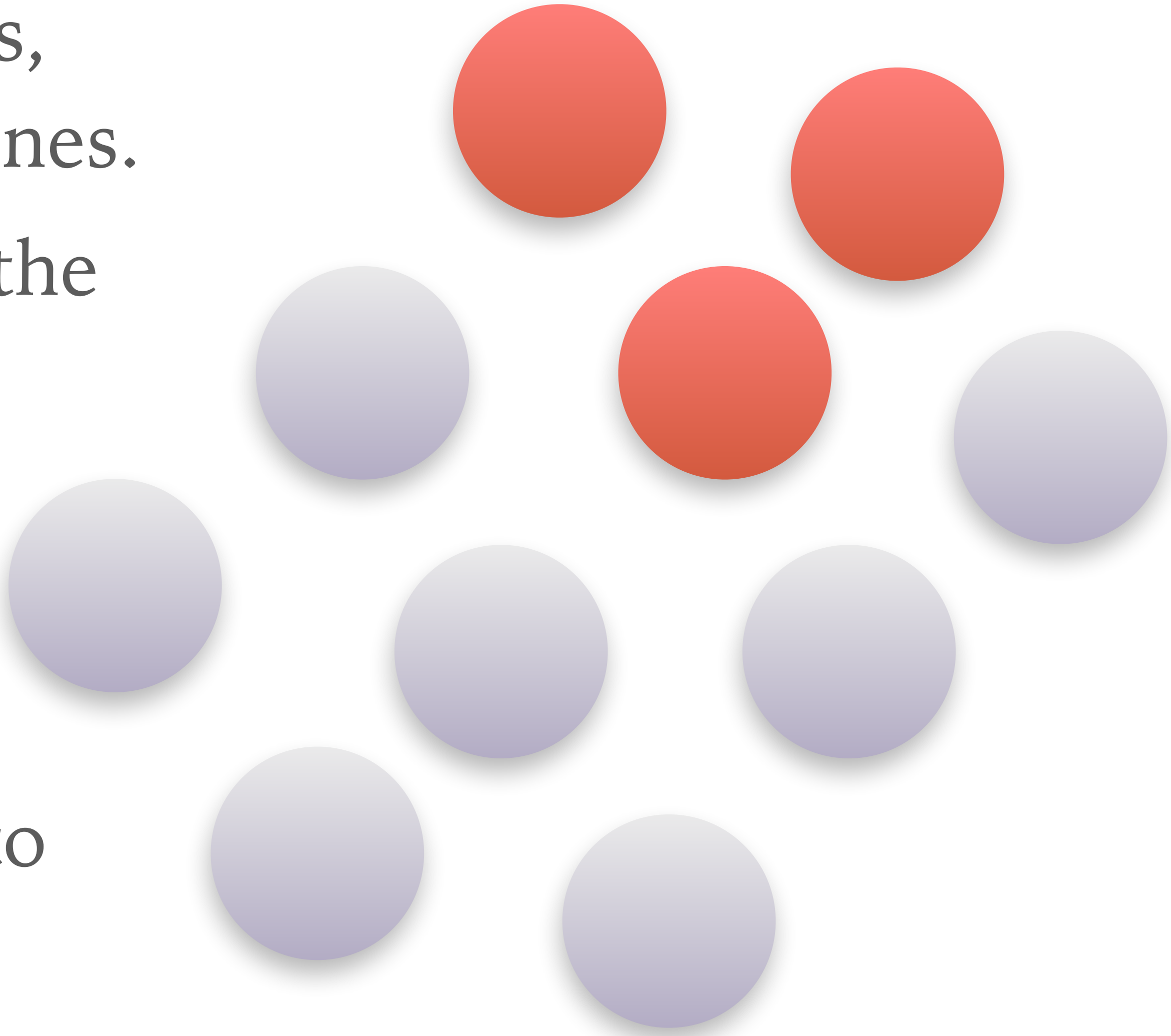
5 = (1,4) or (2,3) or (3,2) or (4,1)

… …



WARM-UP: RANDOM EXTRACTIONS

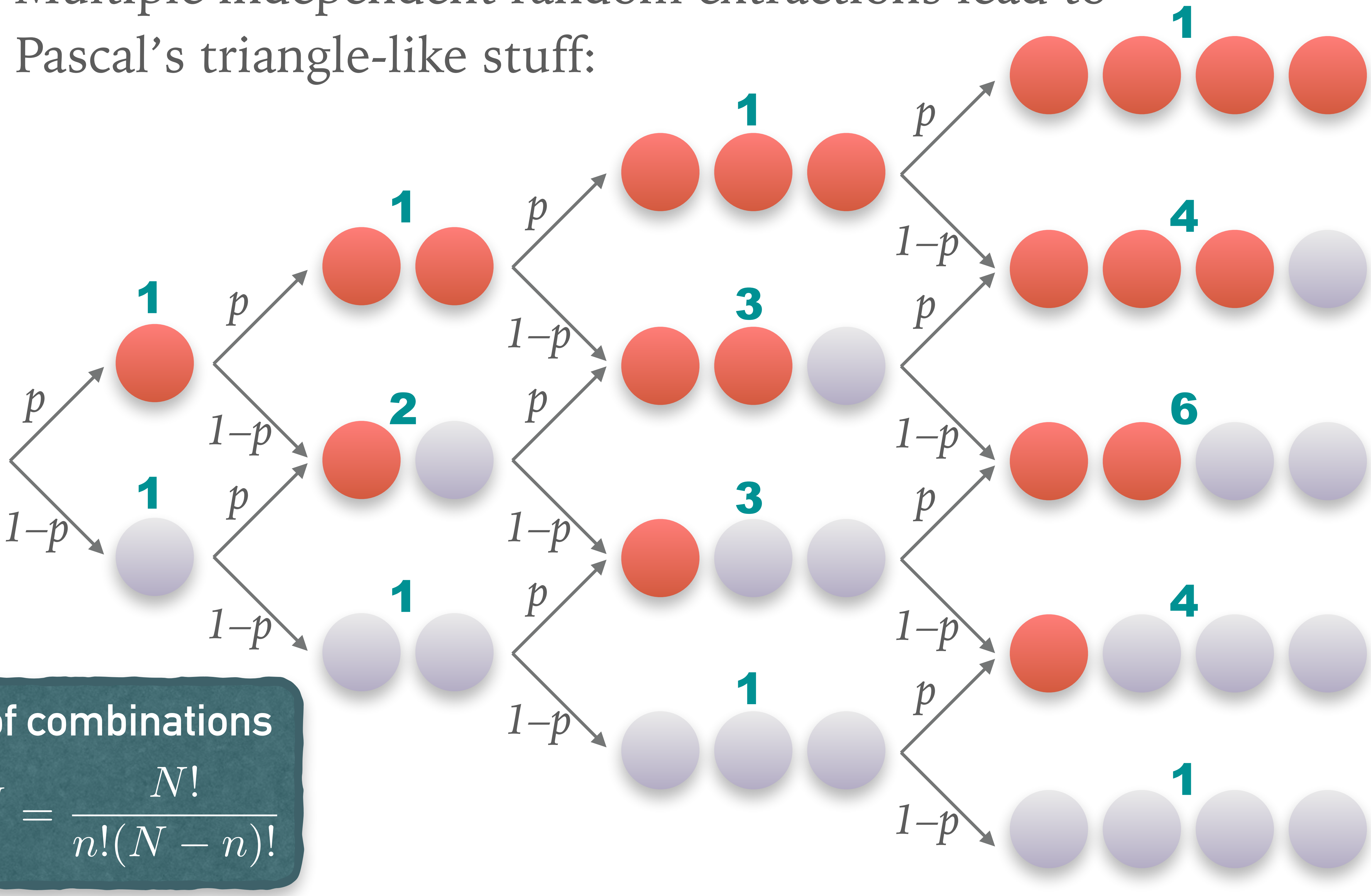
- Suppose you have a bag of marbles, there are 3 red ones and 7 white ones.
- Let's define a "success", which is the extraction of a red marble out of this bag:
  - **Red:** $p = 3/10$
  - **White:** $1 - p = 7/10$
- Classical probability applies only to integer cases,so strictly speaking, p should be a rational number.



- Note the "success" can be also finding an event passing your selection criteria, or a successfully reconstructed physics object, for example, a track, a EM cluster, etc.

MULTIPLE RANDOM EXTRACTIONS

- ➤ Multiple independent random extractions lead to Pascal's triangle-like stuff:



# of combinations

$$C_n^N = \frac{N!}{n!(N-n)!}$$

BINOMIAL DISTRIBUTION

- Consider a distribution of "# of successes" with ***N*** trials, while each trial has a probability of success ***p***:

$$P(n;N,p)=\frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}$$

  - Average: $\boldsymbol{E = \langle n\rangle = Np}$
  - Variance: $\boldsymbol{V = \sigma^2 = \langle n^2\rangle - \langle n\rangle^2 = Np(1-p)}$

- Frequently used for efficiency estimation with a limited size of sample, in this case the efficiency $\boldsymbol{\varepsilon = \langle n\rangle/N = p}$, the uncertainty is given by

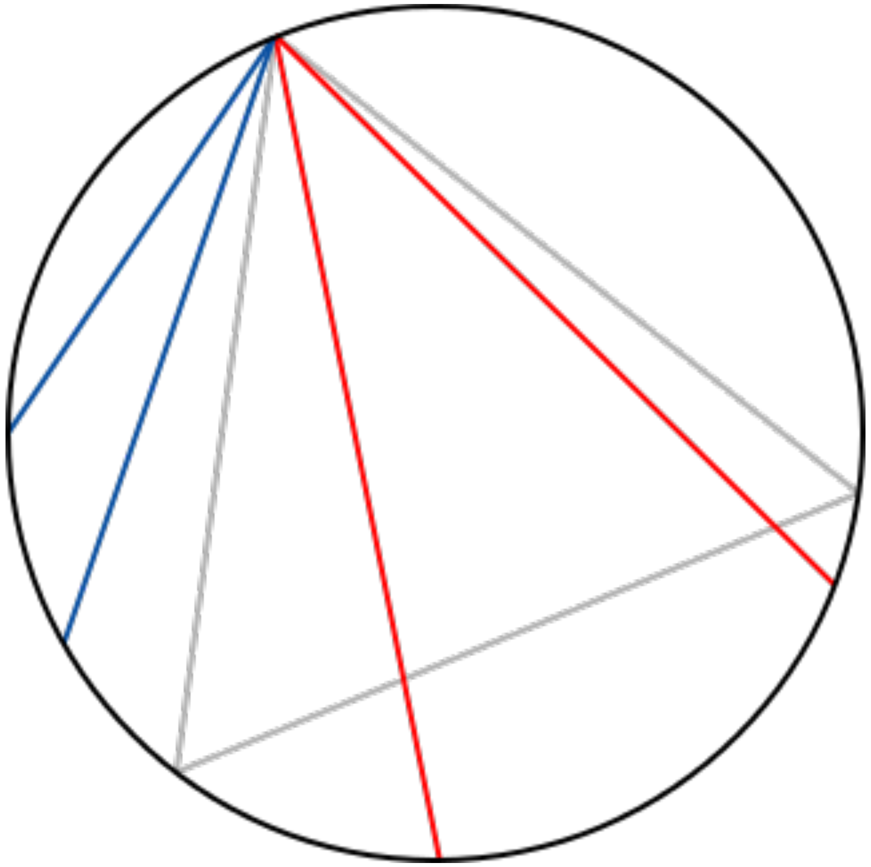
$$\sigma_\epsilon = \sqrt{\frac{\epsilon(1-\epsilon)}{N}}$$

Remark:

σ_ε → 0 when ε → 0 or 1

IN CONTINUOUS CASES?

- A typical example of problematic probability definition in non-discrete cases – Bertrand’s paradox:
  - Given a randomly chosen chord on a circle, what is the probability that the chord’s length is larger than the side of the inscribed triangle?



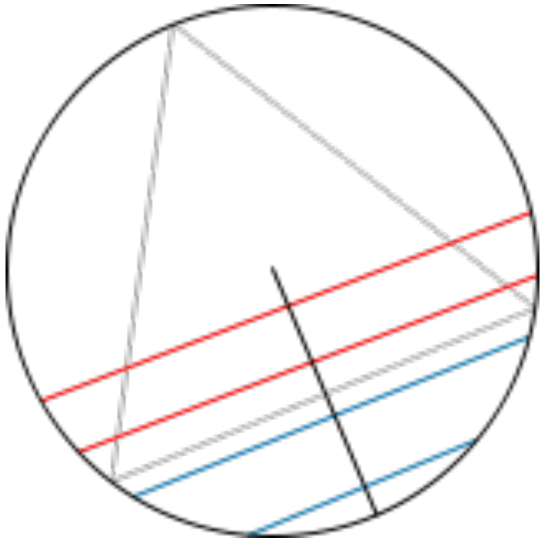
*The "**random endpoints**" method*

Choose two random points on the circumference of the circle and draw the chord joining them. The probability that a random chord is longer than a side of the inscribed triangle is 1/3.

…is this always true?

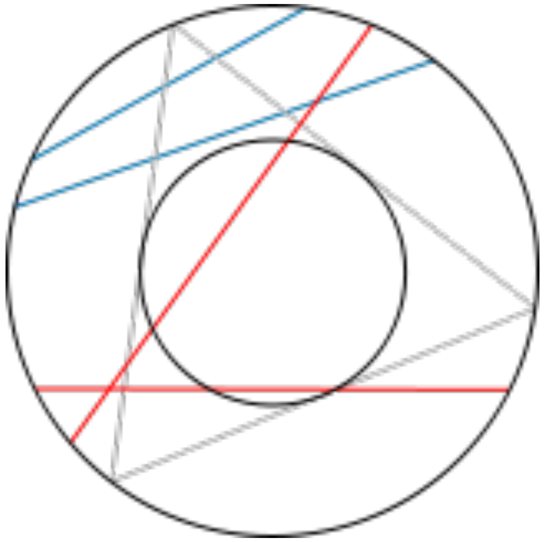
IN CONTINUOUS CASES? (CONT.)

- If one considers slightly different methods, for example:



*The "**random radius**" method*

Choose a radius of the circle, choose a point on the radius and construct the chord through this point and perpendicular to the radius. The side of the triangle bisects the radius, therefore the probability is 1/2.



*The "**random midpoint**" method*

Choose a point anywhere within the circle and construct a chord with the chosen point as its midpoint. The chord is longer if the chosen point falls within a circle of radius 1/2. Thus the probability is 1/4.

*“**Random choice**” is not a well defined concept in this case; some classical probability concepts become arbitrary until we move to discuss the **probability density functions**.*

FORMAL DEFINITION OF PROBABILITY

- **Mathematical probability** – define **Ω** to be the set of all possible *elementary events* $\boldsymbol{X_i}$, which are exclusive (ie. occurrence of one of them implies none of others occurs). The probability of the occurrence of $\boldsymbol{X_i}$, $\boldsymbol{P(X_i)}$, to obey the **Kolmogorov axioms**:

$$(a) P(X_i) \geq 0 \text{ for all } i$$

$$(b) P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$$

$$(c) \sum_{\Omega} P(X_i) = 1$$

more complex probability expressions can be deduced for non-elementary events.

We require operational definitions which allows us to measure probabilities: Frequentist probability and Bayesian probability. Both of them satisfy the Kolmogorov axioms.

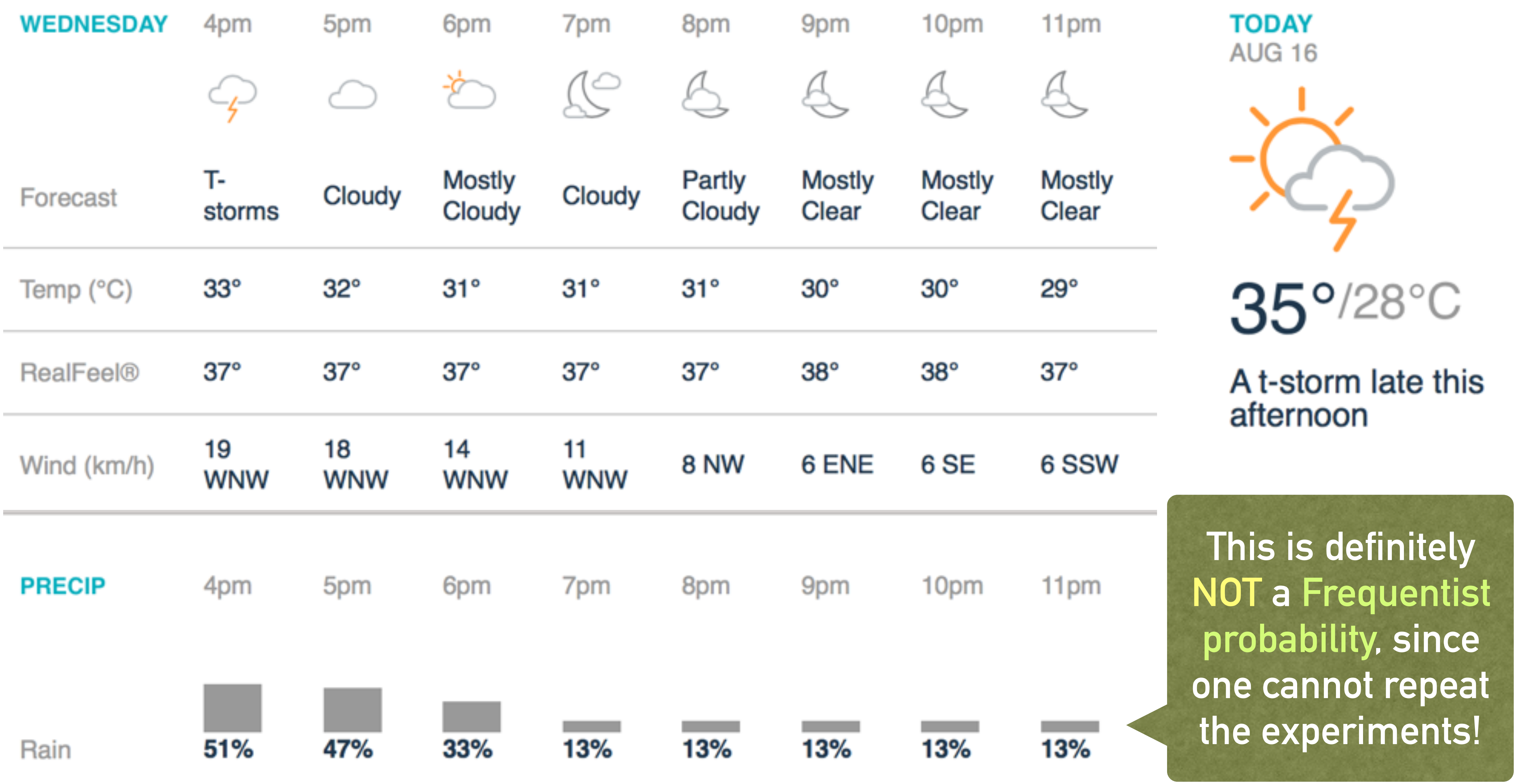
FREQUENTIST PROBABILITY

- **Frequentist probability** is in fact, defined along ***experiments***. Consider # of events of type X is ***n***, and total # of events is ***N*** obtained from a series of experiments, then the frequentist probability that any single event will be of type X can be defined as

$$P(X) = \lim_{N\to\infty} \frac{n}{N}$$

- Obviously this definition requires an infinite number of experiments, and it cannot be the real case! But as long as it is in principle possible always to perform one more experiments, a targeting accuracy can be obtained.
- However, this definition implies an important restriction: it can be only applied to **repeatable experiments**!

A FAMILIAR "BROKEN" CASE?



WEDNESDAY	4pm	5pm	6pm	7pm	8pm	9pm	10pm	11pm
Forecast	T-storms	Cloudy	Mostly Cloudy	Cloudy	Partly Cloudy	Mostly Clear	Mostly Clear	Mostly Clear
Temp (°C)	33°	32°	31°	31°	31°	30°	30°	29°
RealFeel®	37°	37°	37°	37°	37°	38°	38°	37°
Wind (km/h)	19 WNW	18 WNW	14 WNW	11 WNW	8 NW	6 ENE	6 SE	6 SSW

PRECIP	4pm	5pm	6pm	7pm	8pm	9pm	10pm	11pm
Rain	51%	47%	33%	13%	13%	13%	13%	13%

TODAY
AUG 16

35°/28°C

A t-storm late this afternoon

This is definitely NOT a Frequentist probability, since one cannot repeat the experiments!

Unless you want to talk about multiverse…

BAYESIAN PROBABILITY

- In order to define a probability that can be applied to non-repeatable experiments, we have to replace it by something else: the ***degree of belief***, which is the basis of **Bayesian probability**.
- The idea is to determine how strongly a person believes that X will occur by determining how much he would be willing to bet on it, assuming that he wins of fixed amount of X does later occur and nothing if it fails to occur.
- $P(X)$ is defined as the largest amount he would willing to bet, divided by the amounts he stands to win.
- Although all these statement may sound strange, this definition does obey the Kolmogorov axioms.

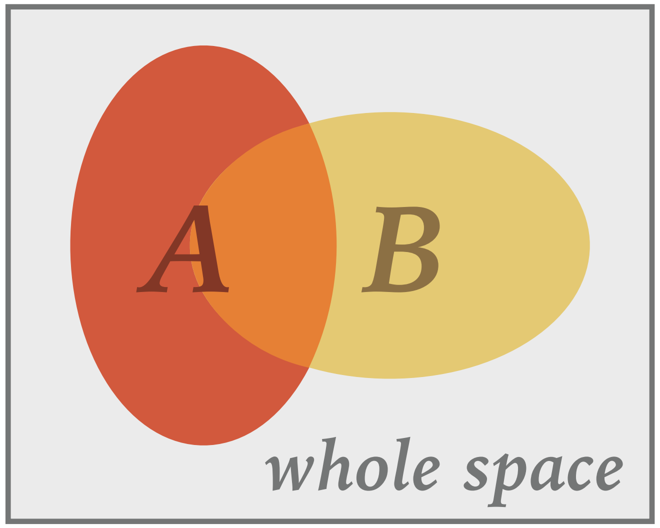
BAYESIAN PROBABILITY (CONT.)

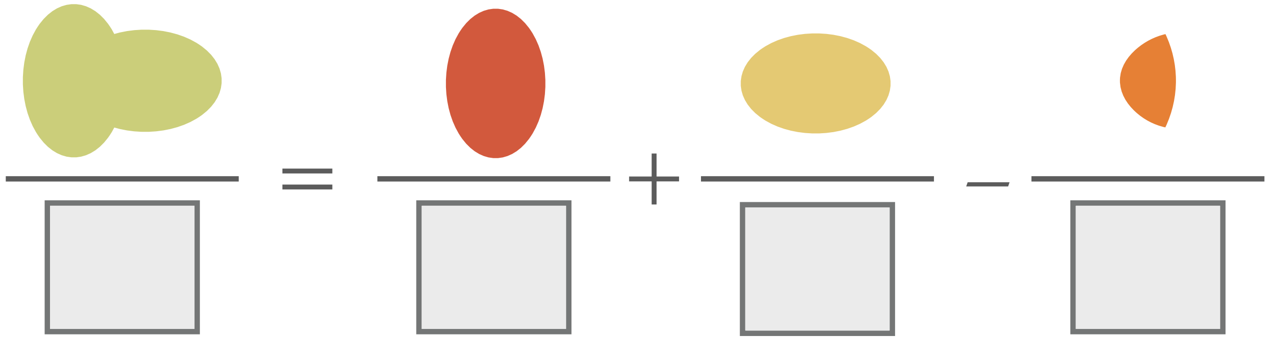
- Bayesian probability is an interpretation of the concept of probability, which is interpreted as reasonable expectation representing **a state of knowledge** or as **quantification of a personal belief**.
- Properties of (*subjective*) Bayesian probability:
  - It is as much a property of observer as it is of the system being observed.
  - It depends on the state of the observer's knowledge, and will in general change as the observer obtains more knowledge.
- For example, **P(tomorrow is a raining day)** and **P(SUSY is true)** do exist, which cannot be defined in frequentist way.

PROPERTIES OF PROBABILITY

- For any probability satisfies Kolmogorov axioms, the following discussions do apply.
- Consider a set ***A*** of elementary event $\boldsymbol{X_i}$, we denote ***P(A)*** as the probability that an $\boldsymbol{X_i}$ in set ***A*** occurs.
- For two non-exclusive sets A and B, the probability of an event occurring in A or in B, or in both can be obtained by the addition law:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

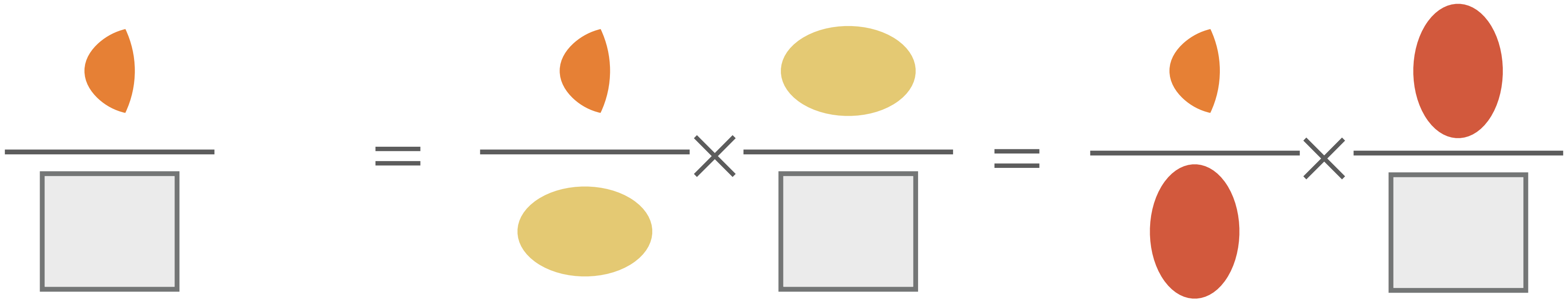


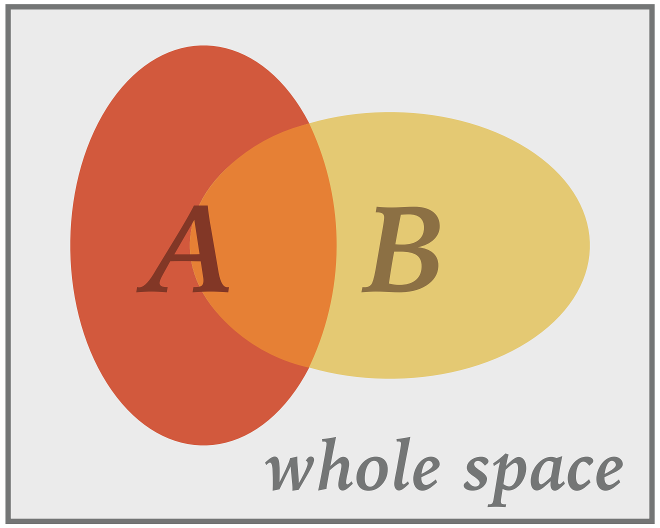


CONDITIONAL PROBABILITY

- Then the conditional probability, ***P(A|B)***, the probability that an elementary event, known to belong to the set ***B***, and is also a member of set ***A***:

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$





Sets A and B are said to be independent (occurrence of B is irrelevant to the occurrence of A) if

$$P(A|B) = P(A)$$

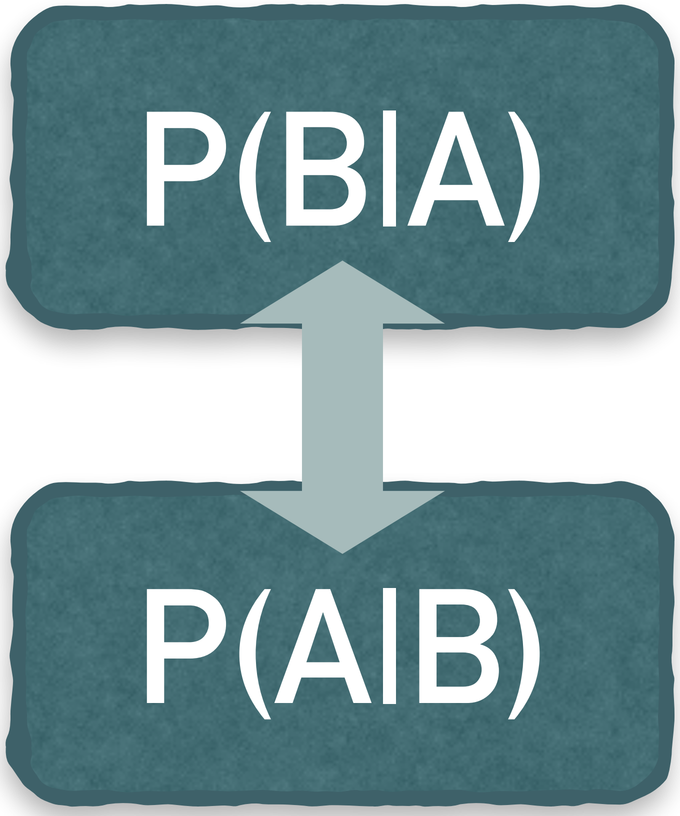
or

$$P(A \text{ and } B) = P(A)P(B)$$

BAYES THEOREM

- **Bayes theorem** describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
- A common usage is to invert conditional probabilities.

$$P(A|B) = P(B|A) \cdot P(A)/P(B)$$


P(B|A)

P(A|B)

the likelihood of observing event B given that A is true.

e.g. A: typhoon is landing
B: it is raining

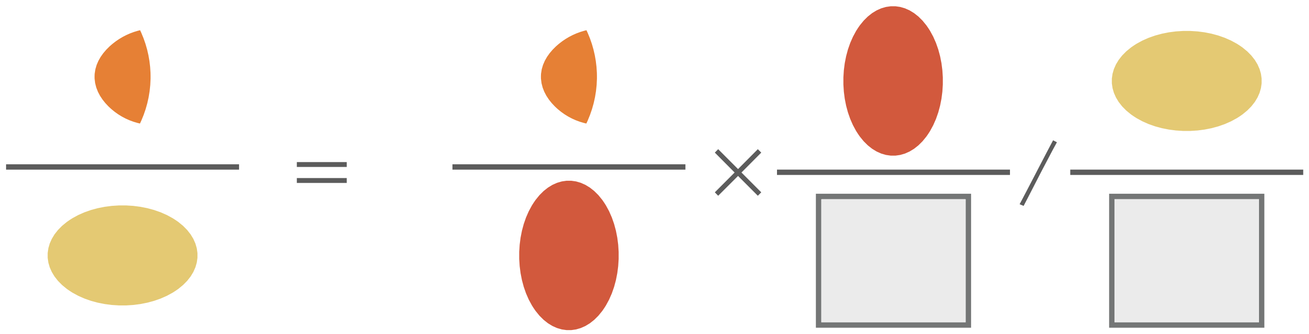
the posterior probability of A is true given observing B.

e.g. B: it is raining
A: typhoon is landing

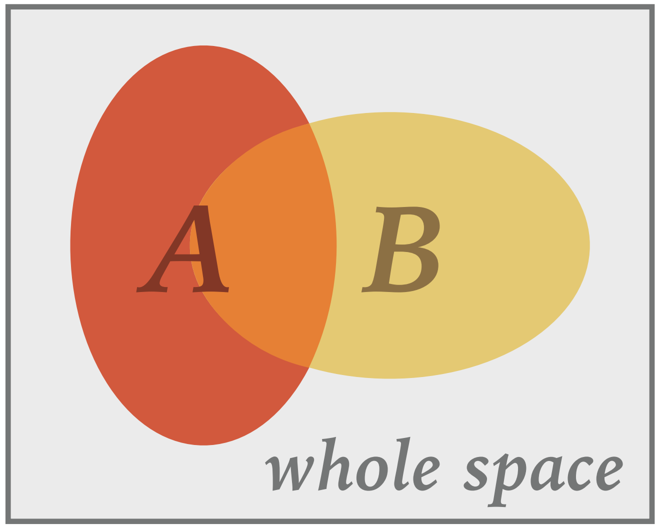
BAYES THEOREM FOR DISCRETE EVENTS

- The theorem which links ***P(A|B)*** to ***P(B|A)*** is the **Bayes theorem**, which follows the definition of conditional probability:

$$P(A|B) = P(B|A) \cdot P(A)/P(B)$$



This is kind of obvious from the formulation, but it may be totally straightforward if one considers a real case.



Remark: using the above Bayes theorem does not imply you are using a Bayesian probability. The Bayes theorem applies to different probabilities as well.

DRUG TEST RESULTS

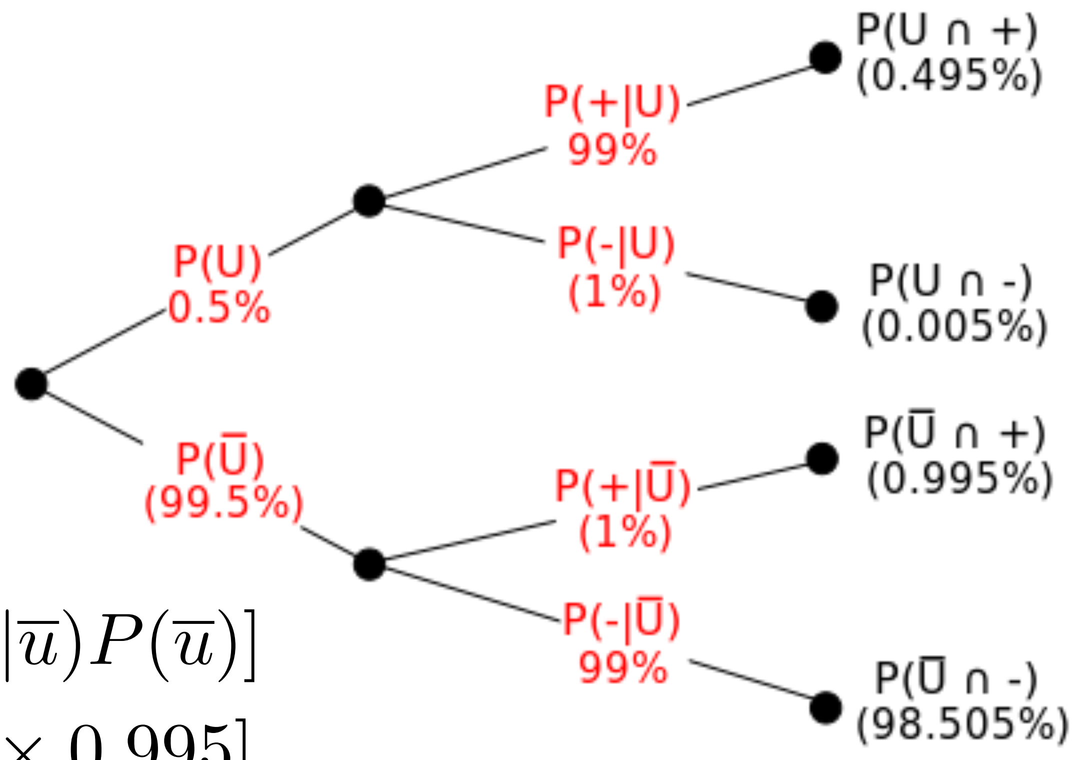
PASSED

FOR INTERNAL USE ONLY!

EMPLOYEE INFORMATION
EMPLOYEE NAME: John Doe
STREET ADDRESS: 12345 Main Street
here STATE: Anywhere ZIP CODE: 123456
0) 456-7890

BAYES THEOREM EXAMPLE

- Let’s consider a classical problem – suppose a drug test is 99% sensitive and 99% specific:
  - 99% positive results for drug users (=1% failing detection)
  - 99% negative results for non-drug users (=1% false alarm)
- Suppose that 0.5% of people are users of the drug. What is the probability that a randomly selected individual with a positive test is a user?



$$
\begin{aligned}
P(u|+) &= P(+|u)P(u)/P(+) \\
&= P(+|u)P(u)/[P(+|u)P(u) + P(+|\overline{u})P(\overline{u})] \\
&= 0.99 \times 0.005/[0.99 \times 0.005 + 0.01 \times 0.995]
\end{aligned}
$$

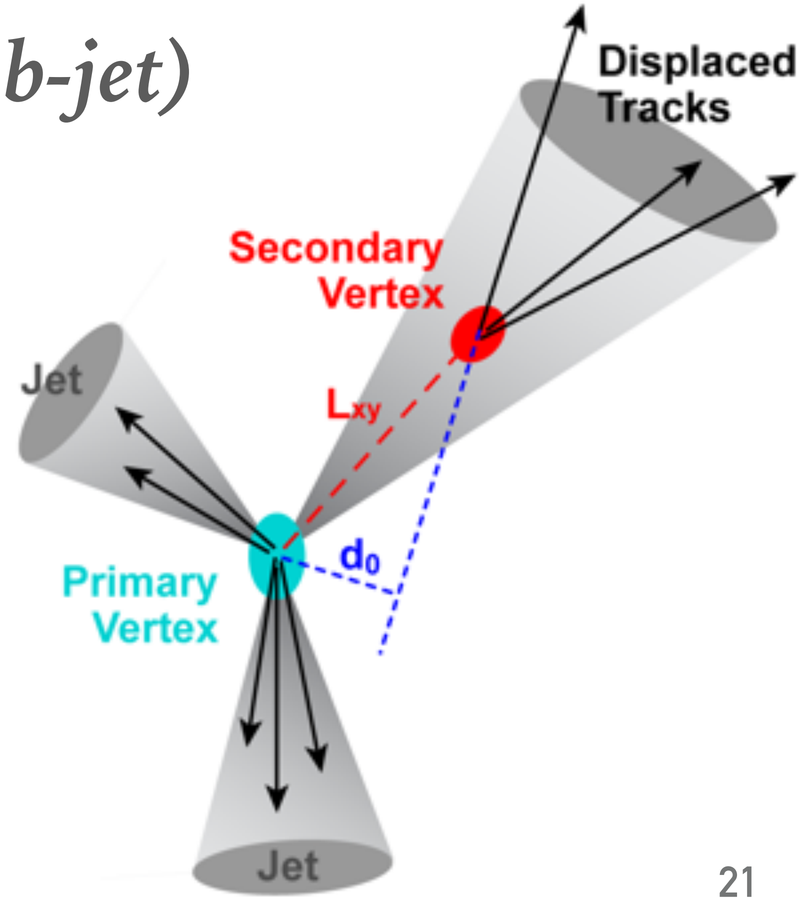
$$\approx 33\%$$

It's as low as 33% in fact!

USING THE BAYERS THEOREM

- Considering a b-tag algorithm that has been developed, one measures the following probabilities:
  - ***P(b-tag|b-jet)***: efficiency for b-tagging (*probability of true b-jet passing the b-tag criteria*)
  - ***P(b-tag|not b-jet)***: efficiency for background
  - ***P(not b-tag|b-jet) = 1 – P(b-tag|b-jet)***
  - ***P(not b-tag|not b-jet) = 1 – P(b-tag|not b-jet)***

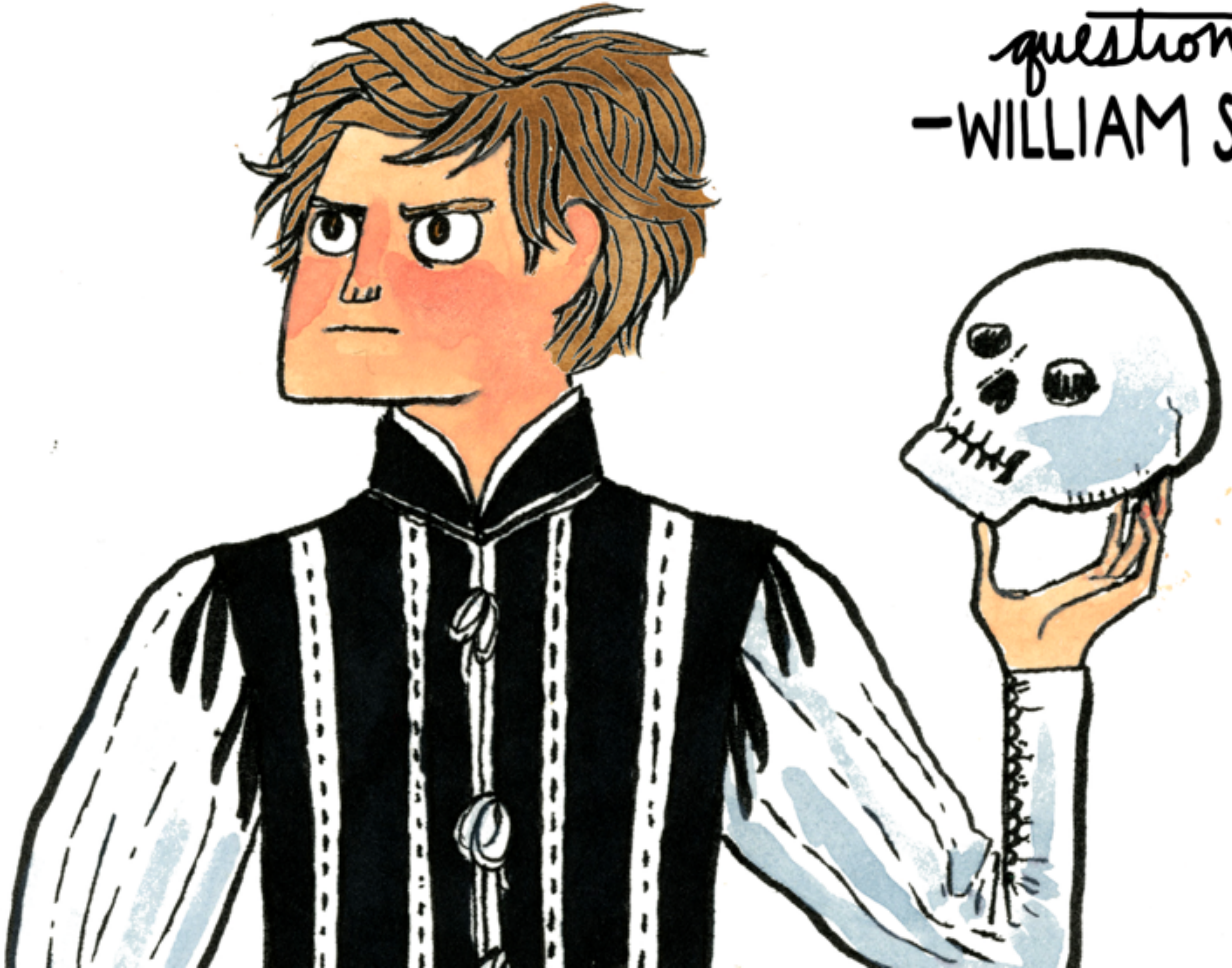
- Question: **given a selection of jets tagged as b-jets, what fraction of them is b-jets?** ie. what is *P(b-jet|b-tag)*, which is usually called as <u>the purity of b-tagged jets</u>?



TO BE OR NOT TO BE?

That is the question."

-WILLIAM SHAKESPEARE



USING THE BAYERS THEOREM (II)

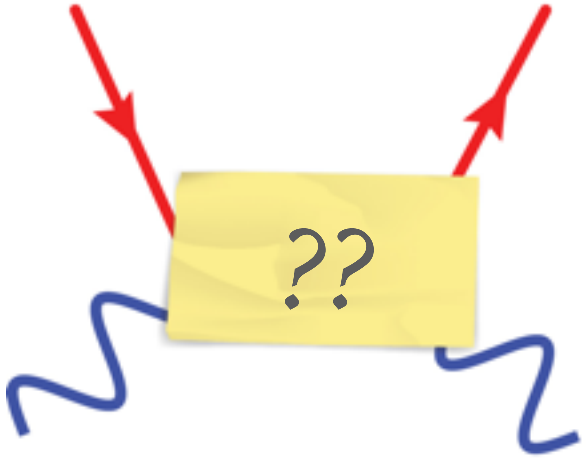
- Answer: **nope, we cannot.** Missing information of ***P(b-jet)*** as:

$$P(b\text{-}jet\,|\,b\text{-}tag) = P(b\text{-}tag\,|\,b\text{-}jet) \times P(b\text{-}jet)\ /\ P(b\text{-}tag)$$

- ***P(b-tag)*** is known to some extent since you know how many candidates passing the b-tag requirement.
- ***P(b-jet)*** is the true fraction of all jets that are b-jets, which is unknown in the current scope.
- It is not straightforward to invert ***P(b-tag|b-jet)***, the efficiency of b-tagging, to ***P(b-jet|b-tag)***, the purity of b-tagged jets.
- And you may noticed some of the "***P***" we are discussing in this example is in fact follows frequentist definition.

USING THE BAYERS THEOREM (III)

- How about an obviously Bayesian probability case?
- Consider a background free counting experiment, a theorist proposed a model which predicts a signal with Poisson mean of 3 events. The experiment has been performed and zero events are observed. From Poisson distribution we know:
  - *P(0 event|model true)* = $3^0e^{-3}/0!$ = 0.05
  - *P(0 event|model false)* = 1
  - *P(>0 event|model true)* = 0.95
  - *P(>0 event|model false)* = 0



- Question: **Given the result of the experiment, what is the probability that the proposed model is true?** ie. what is *P(model true|0 event)*?

USING THE BAYERS THEOREM (IV)

- Answer: **now you can see that it is not possible to invert the probability** due to the missing ***P(model true),*** the **degree of belief** in the model *prior* to the experiment. Such as

$$P(\text{model true}|0\text{ event}) = P(0\text{ event}|\text{model true}) \times \frac{P(\text{model true})}{P(0\text{ event})}$$

- Let's apply the theorem to the opposite case:

$$P(\text{model false}|0\text{ event}) = P(0\text{ event}|\text{model false}) \times \frac{P(\text{model false})}{P(0\text{ event})}$$

- Remember:

$$\begin{aligned} P(\text{model false}|0\text{ event}) &= 1 - P(\text{model true}|0\text{ event}) \\ P(\text{model false}) &= 1 - P(\text{model true}) \\ P(0\text{ event}|\text{model true}) &= 0.05 \\ P(0\text{ event}|\text{model false}) &= 1 \end{aligned}$$

It is straightforward to obtain this relation:

$$P(\text{model true}|0\text{ event}) = \frac{0.05 \times P(\text{model true})}{1 - 0.95 \times P(\text{model true})}$$

USING THE BAYERS THEOREM (V)

- So the probability that the proposed model is true does depend on the ***result of the experiment*** as well as the ***prior probability P(model true)***!
- Let the "model" to be something nearly possible, for example possibly the SM itself: ***P(model true) = 1 – ε***.
  - This gives ***P(model true| 0 event) = 1 – 20ε***, still very likely to be true even the ***P(0 event|model true)*** is only as low as 5%!
- Let the "model" to be something nearly impossible, for example possibly some crazy new physics: ***P(model true) = ε***.
  - This gives ***P(model true| 0 event) = 0.05ε***, a low prior probability gives a very low posterior probability.
- You may find this interpretation if kind of odd, given the prior is something you cannot avoid when introducing Bayesian probability!

INTERMISSION

- Somebody gave you a “magic coin”, and claimed that it only shows head in coin tosses. But you only allowed to perform the experiment (toss it!) for 3 times and all tosses indeed show the head, e.g.
  - **P(all 3 heads|magic coin) = 1**
  - **P(all 3 heads|normal coin) = $(0.5)^3$ = 0.125**
  - **P(not 3 heads|magic coin) = 0**
  - **P(not 3 heads|normal coin) = 1 – 0.125 = 0.875**

- Given the experimental result, what is the probability of this coin is really a magic coin? ie. what is **P(magic coin|all 3 heads)**, by considering the following two priors:
  - If you really believe in magic, e.g. **P(magic coin)~0.99**
  - If you do not believe in magic, e.g. **P(magic coin)~0.01**

A MORE GENERALIZED VIEW

- When involving hypotheses testing (*e.g. model true*) instead of just sets of events (*e.g. b-tag*), we are entering the Bayesian framework. Therefore the Bayes theorem can be written as

$$P(\theta_i|X^0) = P(X^0|\theta_i) \times P(\theta_i) \,/\, P(X^0)$$

- $P(\theta_i|X^0)$: the posterior probability for hypothesis θ_i, given data X^0 have been observed.
- $P(X^0|\theta_i)$: the probability of obtaining the observed data X^0, given hypothesis θ_i, which must be known.
- $P(\theta_i)$: the prior probability and represents the knowledge or degree of brief before the experiment was performed.
- $P(X^0)$: normalization, since $\Sigma_i\, P(\theta_i|X^0) = P(X^0)$, but this may not be known. If this is the case, a weaker form is usually given by

$$P(\theta_i|X^0) \propto P(X^0|\theta_i) \times P(\theta_i)$$

COMMENT: THE PRIOR

- The degrees of brief in a hypothesis depends on the experimental results and the prior probability before the experiment. Or, one can say that **Bayesian statistics is subjectivity by definition**.
- Surely for the physicists this is not very appealing; people tried very hard to look for a way to avoid introducing prior into the experiments, but without a real success.
  - Bayesian may comment that it is actually intersubjective, i.e. the real nature of learning and knowing physics.
- **Frequentist approach is generally preferred by a large fraction of physicists** (*probably the majority, but Bayesian statistics is getting more and more popular in many application, also thanks to its easier application in many of the cases*).

CONTINUOUS RANDOM VARIABLE

- A random event may be associated a random variable $\boldsymbol{X}$, which takes different possible numerical values $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, …, corresponding to the different possible outcome.
- Those probabilities $\boldsymbol{P(X_1)}$, $\boldsymbol{P(X_2)}$, …, form a **probability distribution**.
- When an experiment consists of $\boldsymbol{N}$ repeated observations of the same random variable $\boldsymbol{X}$, it can be considered as the single observation of a random vector $\boldsymbol{X = \{X_1, X_2, \ldots., X_N\}}$.
- Instead of probability for discrete cases, now we can generalize probabilities of events to probability distributions of random variable, using the tools like **probability density functions**.

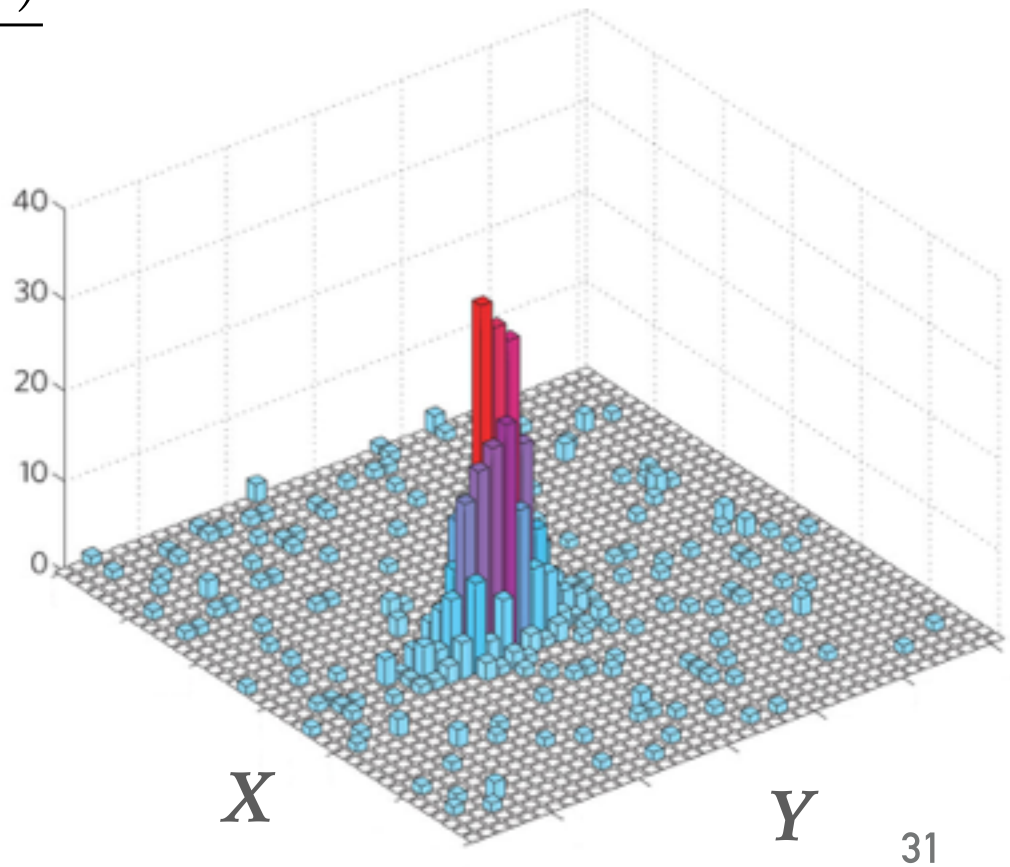
PROBABILITY DENSITY FUNCTION (PDF)

- Consider a random 2D histogram of ***X*** and ***Y***, collecting the data with a particle gun.
- The probability distribution of finding particles at ***X*** and ***Y*** is denoted by ***P(X and Y)***, which is still discrete.
- However it is more convenient to describe this using a continuous function ***f(X, Y)*** by introducing infinite small steps:

$$f(X,Y) = \lim_{\Delta X \to 0, \Delta Y \to 0} \frac{P(X \text{ and } Y)}{\Delta X \Delta Y}$$

- The **probability density function** ***f(X,Y)*** represents a probability density per unit array length of ***X*** and unit length of ***Y***. The normalization condition must be held:

$$\iint_{\Omega} f(X,Y)dXdY = 1$$



CHANGE OF VARIABLE

- Consider such a transformation of ***X*** ⇒ ***Y***, ***f(X)*** ⇒ ***g(Y)***, and maps the interval [***X***,***X***+***dX***] ⇒ [***Y***,***Y***+***dY***], what would be the expression for ***g(Y)***?
- If the transformation is one-to-one with ***Y* = *h(X)***, one has

$$g(Y)dY = f(X)dX \quad \Rightarrow \quad g(Y) = \frac{f(X)}{|h'(X)|}$$

- The ***h'(X)*** is the derivative of the transformation. If ***X*** and ***Y*** are vectors, it would be just the Jacobian of the transformation, i.e. a matrix of the elements

$$J_{ij} = \frac{\partial h_i}{\partial X_j}$$

- If the transformation is NOT one-to-one, i.e., multiple segment of [***X***,***X***+***dX***] mapping into the same [***Y***,***Y***+***dY***]. Thus one has to sum over all such segments, ie.

$$g(Y) = \sum \frac{f(X)}{|h'(X)|}$$

CUMULATIVE AND CONDITIONAL DISTRIBUTIONS

- **Cumulative distribution** ***F(X)*** is defined by

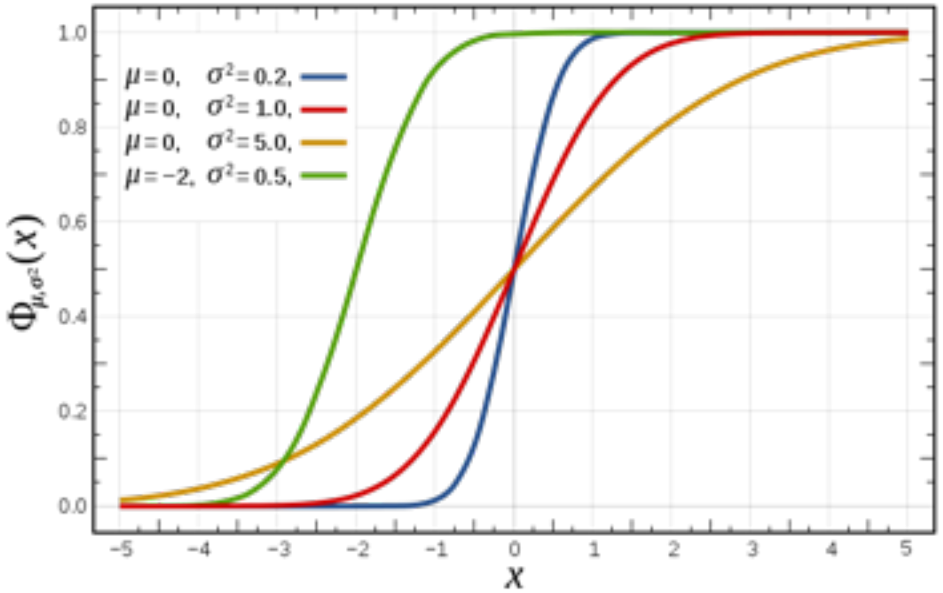
$$F(X) = \int_{X_{\min}}^{X} f(X')dX'$$

by construction:

$$F(X_{\min}) = 0, \quad F(X_{\max}) = 1$$

- **Conditional distribution**: the normalized section through the density function ***f(X,Y)*** at ***X = X₀*** gives the conditional density function of ***Y***:

$$f(Y|X_0) = \frac{f(X_0, Y)}{\int f(X_0, Y)dY}$$



Cumulative distribution function for the normal distribution

COMMENT: BAYES THEOREM FOR CONTINUOUS VARIABLE

- Consider ***N*** independent observation of a continuous variable ***X_i***, and for a continuous hypothesis ***θ*** (for example, a physics parameter like particle mass). The PDF for ith variable is *$f_i(X_i|\theta)$*. The joint density function is

$$p(X|\theta) = \prod_{i=1}^{N} f_i(X_i|\theta)$$

- Question: having made ***N*** observations from the distributions *$f_i(X_i|\theta)$*, what can one say about the value of ***θ***?

- Answer: **classically *θ* has a fixed true value**. So in principle when fits (for example, maximum likelihood fits, will be discussed in the upcoming lecture) applied, the value of ***θ*** can be estimated. But this **cannot be carried out with Bayes theorem**.

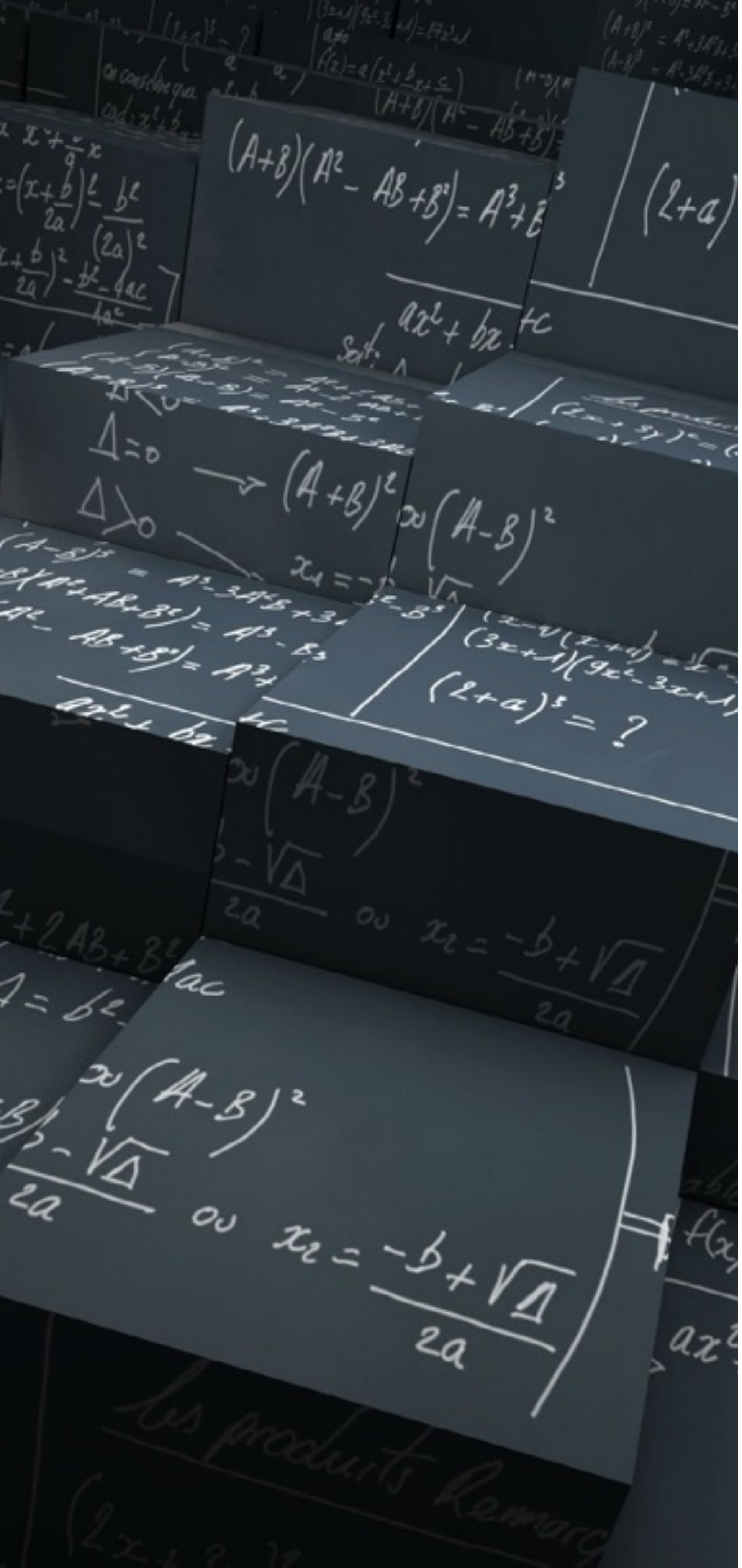
COMMENT: BAYES THEOREM FOR CONTINUOUS VARIABLE (CONT.)

- However with Bayesian methods introduced, the distributions of $\boldsymbol{\theta}$ (*using PDF of* $\boldsymbol{\theta}$) can be taken to represent the degree of belief in different possible value of $\boldsymbol{\theta}$.
- We can obtain the form of Bayes theorem used in Bayesian parameter estimation for a particular set of data, X^0:

$$p(\theta|X^0) = \frac{p(X^0|\theta)p(\theta)}{\int p(X^0|\theta)p(\theta)d\theta}$$

where

- $\boldsymbol{p(\theta|X^0)}$ is posterior probability density for $\boldsymbol{\theta}$.
- $\boldsymbol{p(X^0|\theta)}$ is the likelihood function (*not a PDF!*)
- $\boldsymbol{p(\theta)}$ is the prior probability density for $\boldsymbol{\theta}$. Again this is the major problem in the evaluation. Will be discussed in a later lecture.
- The integration in the denominator is just a normalization factor.



SUMMARY

- The definitions of different probabilities: **Mathematical** / **Frequentist** / **Bayesian** probabilities are introduced.
- You may find it it very interesting one requires different definitions of probabilities to face different problems!
- We have introduced Bayers theorem and how to use it in some special cases, as well as the limitation.
- Next we are going to introduce/discuss several commonly used PDFs.