UNIVERSITY OF LATVIA

DATA ANALYSIS

Toni Šćulac

Faculty of Science, University of Split, Croatia

visiting professor at University of Latvia, Latvia

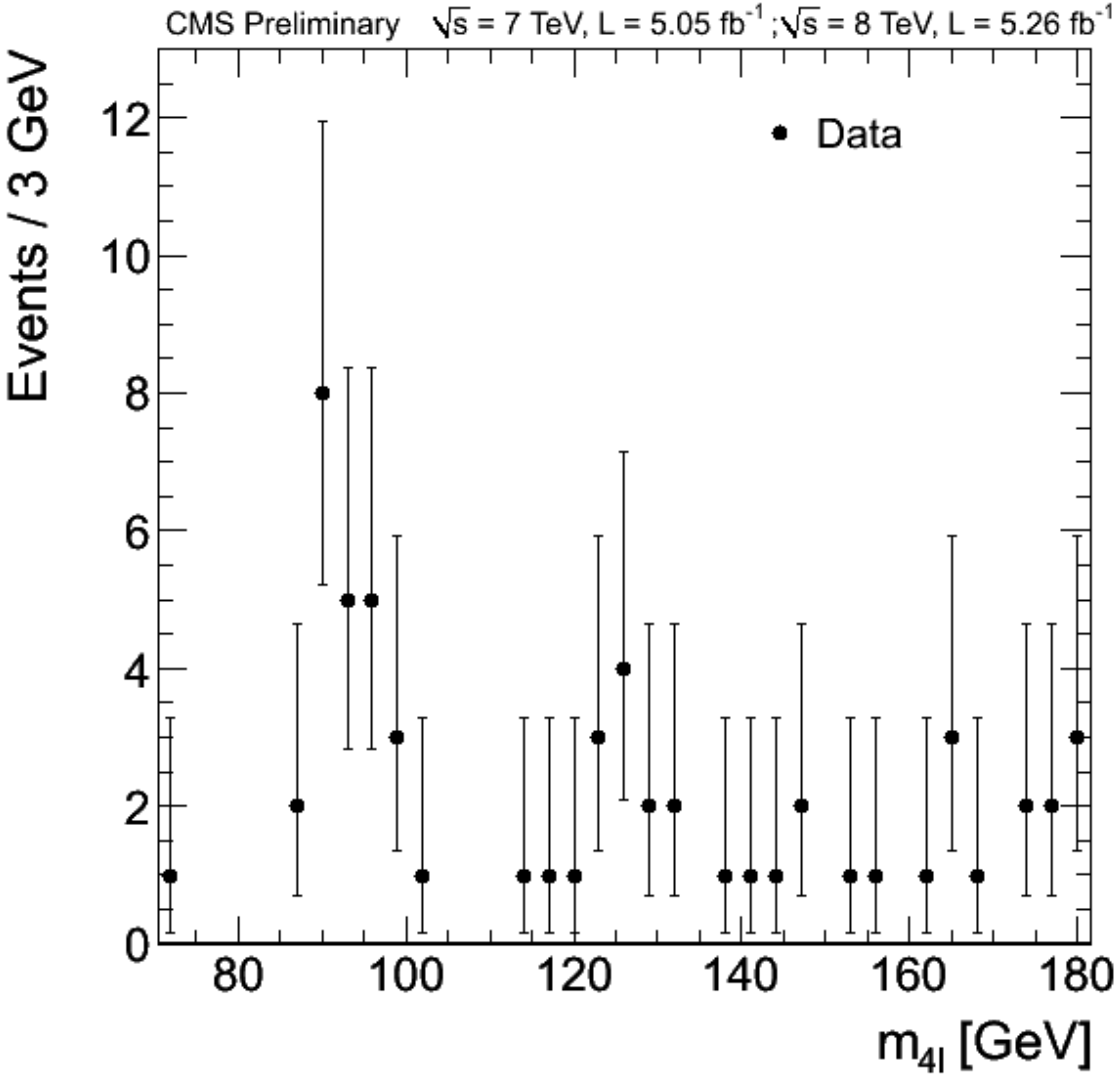
CERN School of Computing 2023, Tartu, Estonia

LECTURES OUTLINE

1) Introduction to Data Analysis
2) Probability density functions and Monte Carlo methods
3) Parameter estimation and Confidence intervals
4) Hypothesis testing and p-value

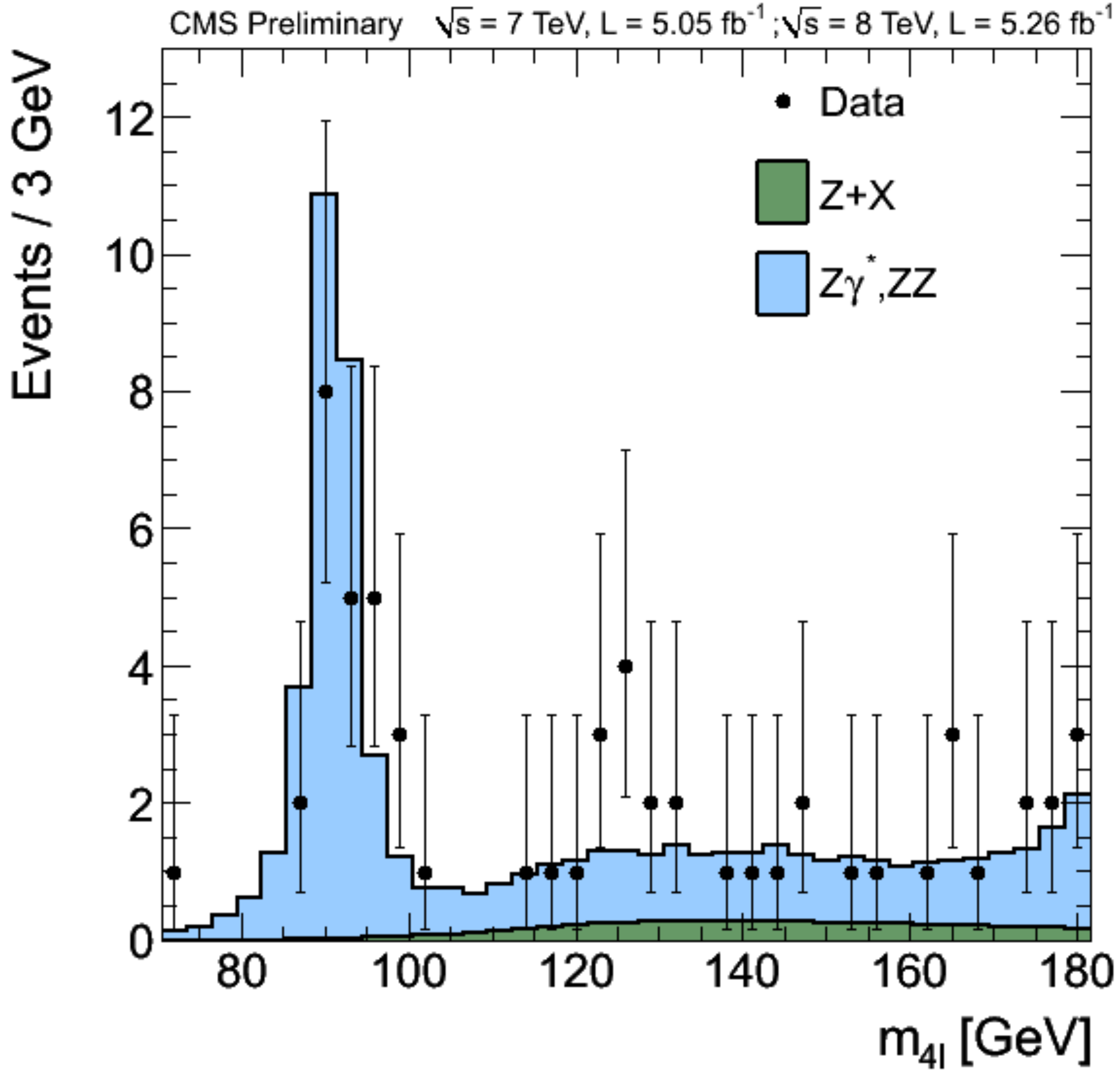
INTRODUCTION TO DATA ANALYSIS

EVENT DISTRIBUTION



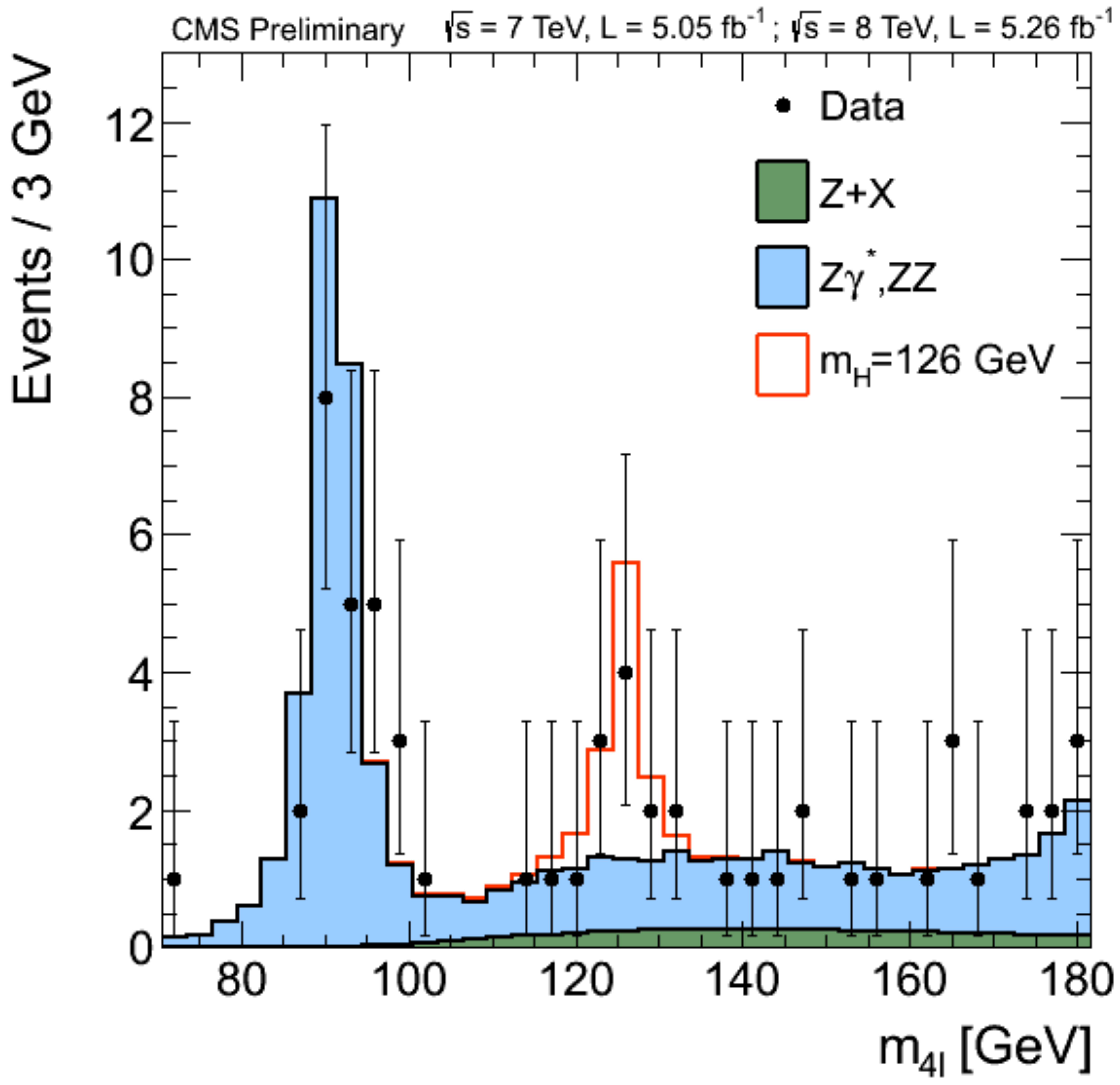
- Does the observed data agree with our expectations from the Standard Model?

EVENT DISTRIBUTION



- We can not tell until we can compare to the expected distribution
- Is there any place where data does not agree with the expectation? Where? How significant?

EVENT DISTRIBUTION



- When can we tell that we have discovered something new?
- Can we ever be 100% sure?
- What is the mass of a newly discovered particle?

WHAT IS DATA ANALYSIS?

*“Data analysis is a process for obtaining **raw data** and converting it into information useful for decision-making by users. Data are collected and analyzed to answer questions, test hypotheses or disprove theories.”*

- Data analysis uses statistics for presentation and interpretation (explanation) of data
- A mathematical foundation for statistics is the probability theory

DATA ANALYSIS IN THE INDUSTRY

RAW DATA → **DATA ANALYSIS** → **USABLE INFORMATION**

$(\text{search string}_1, \text{location}_1)^{\text{user 1}}$
$(\text{search string}_2, \text{location}_2)^{\text{user 1}}$
...
$(\text{search string}_n, \text{location}_n)^{\text{user 1}}$
$(\text{search string}_1, \text{location}_1)^{\text{user 2}}$
...
$(\text{search string}_m, \text{location}_m)^{\text{user 2}}$
$(\text{search string}_1, \text{location}_1)^{\text{user 3}}$
...
$(\text{search string}_1, \text{location}_1)^{\text{user k}}$
...

Maximum Likelihood fit
Significance
Hypothesis testing
P-value
Neural Networks

DATA ANALYSIS IN HEP

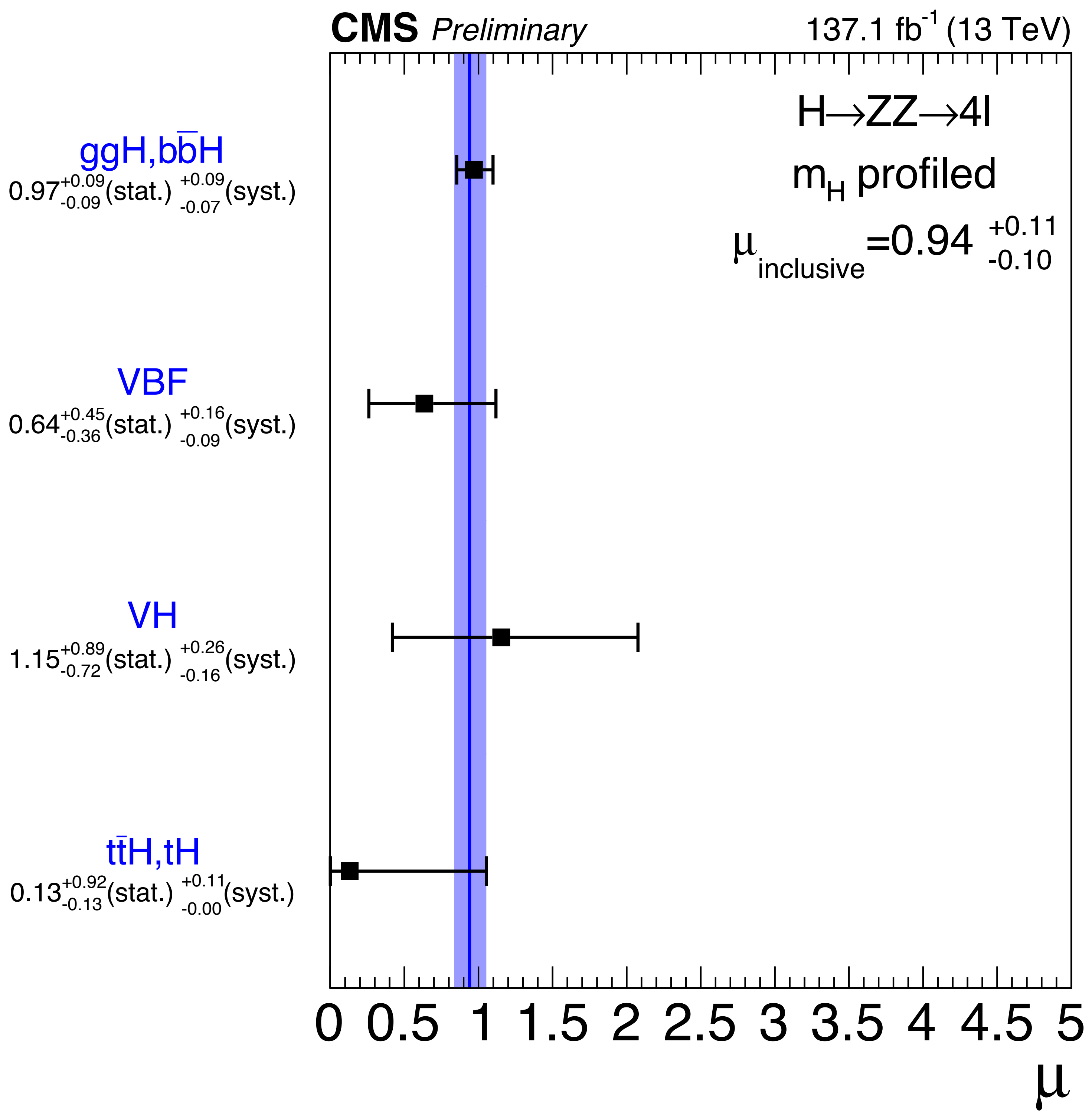
RAW DATA

DATA ANALYSIS →

USABLE INFORMATION

$(p_{x1}, p_{y1}, p_{z1}, E_1)^{\text{event 1}}$
$(p_{x2}, p_{y2}, p_{z2}, E_2)^{\text{event 1}}$
...
$(p_{xn}, p_{yn}, p_{zn}, E_n)^{\text{event 1}}$
$(p_{x1}, p_{y1}, p_{z1}, E_1)^{\text{event 2}}$
...
$(p_{xm}, p_{ym}, p_{zm}, E_m)^{\text{event 2}}$
$(p_{x1}, p_{y1}, p_{z1}, E_1)^{\text{event 3}}$
...
$(p_{x1}, p_{y1}, p_{z1}, E_1)^{\text{event k}}$
...

Maximum Likelihood fit
Significance
Hypothesis testing
P-value
Neural Networks



DATA ANALYSIS IN HEP

- **Main goals** are:
  - estimate (measure) the parameters
  - quantify the uncertainty of the parameter estimates
  - test the extent to which the predictions of a theory are in agreement with the data
- Use of **statistics** for presentation and interpretation (explanation) of data
- A mathematical foundation for statistics is **the probability theory**
- Why is statistics even needed?
  - theory predictions in quantum mechanics are not deterministic
  - finite size of data sample
  - imperfection of the measurement

DATA ANALYSIS GENERAL PICTURE

1 Physical phenomena

Described by a theory

Described by PDFs, depending on unknown parameters with true values

$\theta^{true}=(m_H^{true},\Gamma_H^{true},\ldots,\sigma^{true})$

2 Sampling reality

EXPERIMENT

3 Data sample

$x = (x_1,x_2,\ldots,x_N)$

x is a multivariate random variable

4 DATA ANALYSIS

5 Results

- **parameter estimates**
- **confidence limits**
- **hypothesis tests**

PROBABILITY DEFINITION

What is probability anyway?

> *"Unfortunately, statisticians do not agree on basic principles."*
> *- Fred James*

Mathematical (axiomatic) definition

Classical definition

Frequentist definition

Bayesian (subjective) definition

MATHEMATICAL DEFINITION

- Developed in 1933 by Kolmogorov in his “Foundations of the Theory of Probability”

- Define an exclusive set of all possible elementary events x_i
  - Exclusive means the occurrence of one of them implies that none of the others occurs

- For every event x_i, there is a probability $P(x_i)$ which is a real number satisfying the Kolmogorov Axioms of Probability:
  I) $P(x_i) \geq 0$
  II) $P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$
  III) $\sum P(x_i) = 1$

- From these properties more complex probability expressions can be deduced
  - For non-elementary events, i.e. set of elementary events
  - For non-exclusive events, i.e. overlapping sets of elementary events

- Entirely free of meaning, does not tell what probability is about

CLASSICAL DEFINITION

"Probability = N(favourable) / N"

- My free translation *of the original definition of Pierre-Simon Laplace, A Philosophical Essay on Probabilities*

FREQUENTIST DEFINITION

- Experiment performed N times, outcome x occurs N(x) times
- Define probability: $P(x) = \lim_{N\to\infty} \frac{N(x)}{N}$
- Such a probability has big restrictions:
  - depends on the sample, not just a property of the event
  - experiment must be repeatable under identical conditions
  - For example one can't define a probability that it'll snow tomorrow
- Probably the one you're implicitly using in everyday life
- Frequentist statistics is often associated with the names of *Jerzy Neyman* and *Egon Pearson*

BAYESIAN DEFINITION

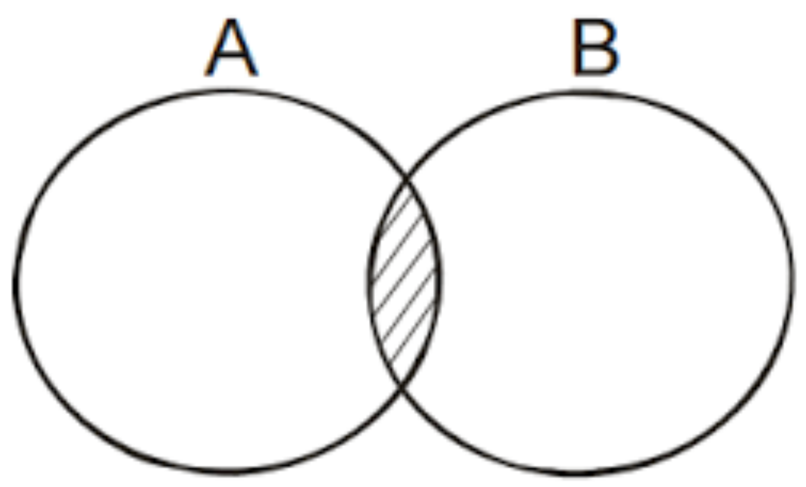
- Define probability: P(x) = **degree of belief** that x is true

- It can be quantified with betting odds:
  - What's amount of money one's willing to bet based on their belief on the future occurrence of the event

- In particle physics frequency interpretation often most useful, but Bayesian probability can provide more natural treatment of non-repeatable phenomena

BAYES' THEOREM

- Define conditional probability: $P(A|B) = P(A \cap B)/P(B)$
  - probability of A happening given B happened
  - for independent events $P(A|B) = P(A)$, hence $P(A \cap B)=P(A)P(B)$



- From the definition of conditional probability Bayes' theorem states:

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)}$$

- T is a **theory** and D is the **data**
- P(T) is the **prior probability** of T: the probability that T is correct before the data D was seen
- P(D|T) is the **conditional probability** of seeing the data D given that the theory T is true.
  - P(D|T) is called the likelihood.
- P(D) is the **marginal probability** of D.
  - P(D) is the prior probability of witnessing the data D under all possible theories
- P(T|D) is the **posterior probability**: the probability that the theory is true, given the data and the previous state of belief about the theory

BONUS PROBLEM - 1

Some rules to follow:

1. In every lecture there will be one bonus problem presented
2. If you have good knowledge in stats and everything I am presenting is known to you feel free to start working on the problem now!
3. Otherwise, work on the problem after the lectures.
4. Solutions won't be provided, you have to come and talk to me to check if your answer is correct or if you need hints!
5. Google/AI assistance is not allowed. These are problems that I want you to think about on your own

Some disease is affecting 0.1% of the total population. You have developed a test to check for the presence of this disease with the following performance:

- For people affected by the disease, the test will be positive 98% of the times
- For people unaffected, the test will still be positive 3% of the times

A patient tests positive, what is the probability that he or she is affected by the disease?

EXAMPLE

- You meet an old friend in a pub. He proposes that the next round should be payed by whoever of the two extracts the card of lower value from a pack of cards
- This situation happens many times in the following days. What is the probability that your friend cheats if you end up paying *N* consecutive times?*
- You assume:
  - P(*cheat*) = 5% and P(*honest*) = 95% (surely an old friend is an unlikely cheater…)
  - P(*N* | *cheat*) = 1 and P(*N* | *honest*) = 2^{-N}
- Bayesian solution:

$$P(cheat\,|\,N) = \frac{P(N\,|\,cheat)P(cheat)}{P(N\,|\,cheat)P(cheat) + P(N\,|\,honest)P(honest)}$$

$$P(cheat\,|\,0) = \frac{1 \cdot P(cheat)}{1 \cdot P(cheat) + 2^{-0}P(honest)} = \frac{0.05}{0.05 + 0.95} = 5\,\%$$

$$P(cheat\,|\,5) = \frac{1 \cdot P(cheat)}{1 \cdot P(cheat) + 2^{-5}P(honest)} = \frac{0.05}{0.05 + 0.03} = 63\,\%$$

*taken from G.D'Agostini, Bayesian Reasoning in HEP, Principles and Applications, CERN-99-03, 1999

LEARNING BY EXPERIENCE

- If you started with P(*cheat*) = 5% and you end up paying for 5 drinks in a row, what should you do when you meet your old "friend" again after 2 years?
- You should learn from your experience and take your prior to be P(*cheat*)=63%!
- If you now end up paying 5 more consecutive drinks:

$$P(cheat\,|\,5) = \frac{1 \cdot P(cheat)}{1 \cdot P(cheat) + 2^{-5} P(honest)} = \frac{0.63}{0.63 + 0.012} = 98\,\%$$

P(cheat)	**P(cheat I N)**		
%	N=5	N=10	N=15
1	24%	91%	99.7%
5	63%	98%	99.94%
50	97%	99.9%	99.99%

When you learn from the experience, your conclusion does not longer depend on the initial assumptions!

RANDOM VARIABLES

- **Random event** is an event having more than one possible outcome
  - Each outcome may have associated probability
  - Outcome not predictable, only the probabilities known

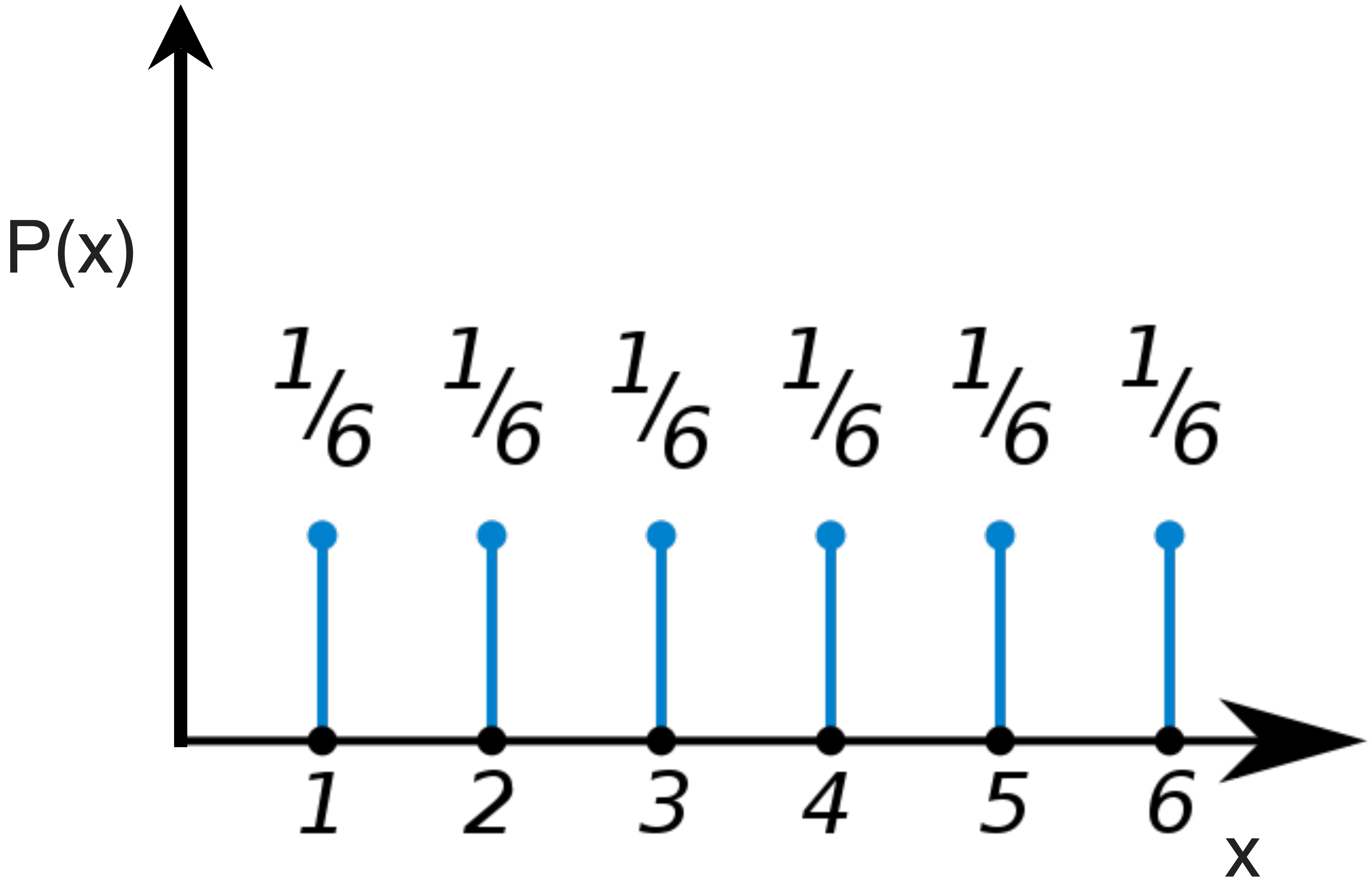
- Different possible outcomes may take different possible numerical values x_1, x_2, ...
- The corresponding probabilities $P(x_1)$, $P(x_2)$, ... form a **probability distribution**

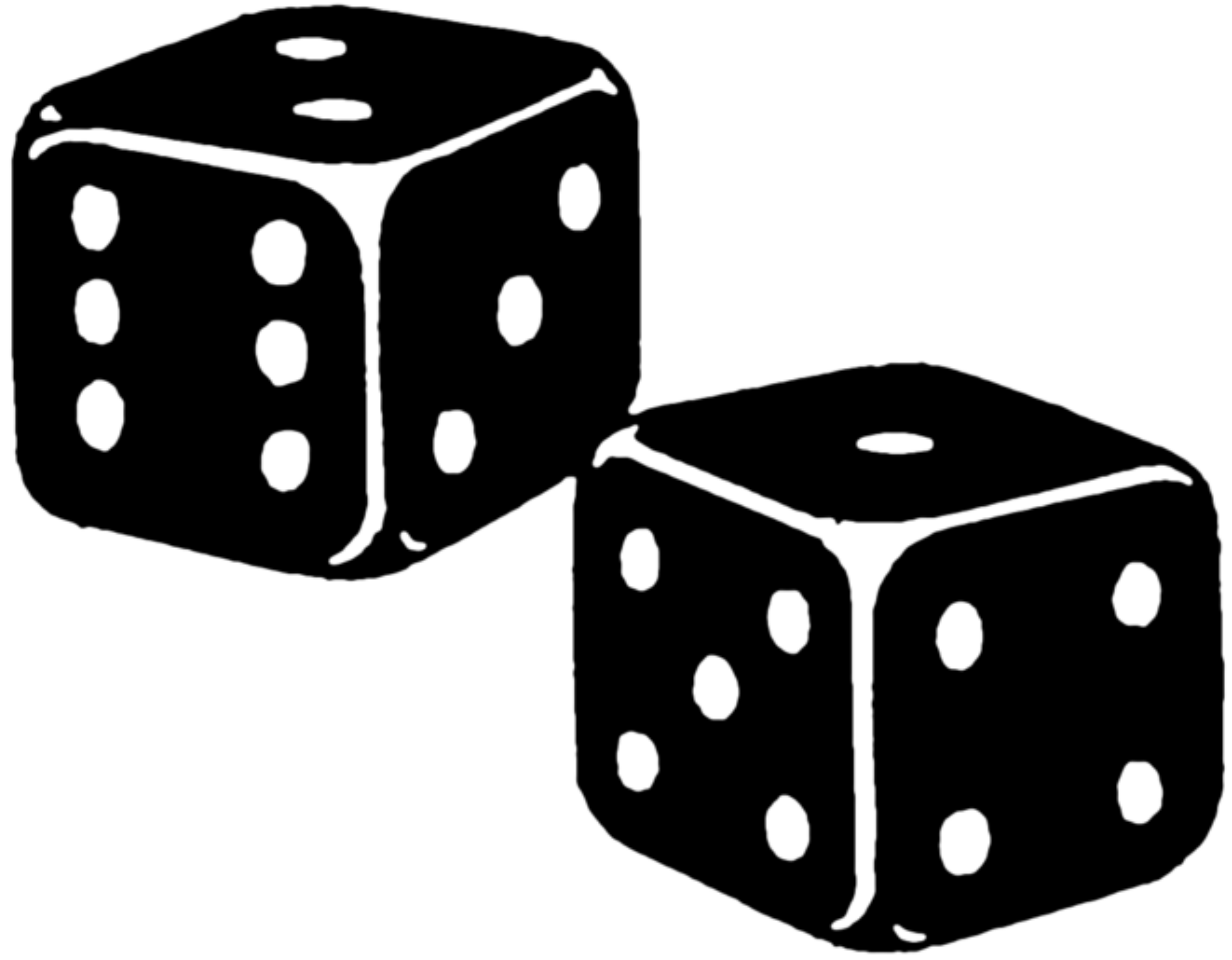
- If observations are independent the distribution of each random variable is unaffected by knowledge of any other observation
- When an experiment consists of N repeated observations of the same random variable x, this can be considered as the single observation of a random vector **x**, with components x_1, x_2, …, x_N

DISCRETE RANDOM VARIABLES

- Rolling a die:
  - Sample space = {1,2,3,4,5,6}
  - Random variable x is the number rolled

- Discrete probability distribution:





CONTINUOUS RANDOM VARIABLES

- A spinner:
  - Can choose a real number from [0,2n]
  - All values equally likely
  - x = the number spun
  - Probability to select any real number = 0
  - Probability to select any **range** of values > 0
    - Probability to choose a number in [0,n] = 1/2
  - Probability to select a number from any range Δx is Δx/2n
  - Now we say that **probability density** p(x) of x is 1/2n

- More general: $$P(A < x < B) = \int_A^B p(x)dx$$