Modern Methods of Statistical Data Analysis

From parameter estimation to deep learning — A guided tour of probability

Lecture 1
-
Organization & Fundamental Concepts I

P.-D. Dr. Roger Wolf
roger.wolf@kit.edu

Dr. Pablo Goldenzweig
pablo.goldenzweig@kit.edu

Dr. Jan Kieseler
jan.kieseler@kit.edu

Dr. Slavomira Stefkova
slavomira.stefkova@kit.edu

Karlsruher Institut für Technologie

Organization

Lectures: Fridays 11^{30} - 13^{00}

- **Lecturers:**
  - Dr. Pablo Goldenzweig (CS 30.23 / R. 9-9) [Lectures 1-8]
  - P.-D. Dr. Roger Wolf (CS 30.23 / R. 9-20) [Lectures 9-12]
  - Dr. Jan Kieseler (CS 30.23 / R. 9-20) [Lectures 13-14]
- **Office hours**: On request

Computerpraktikum: Thursdays 15^{30} - 18^{00}, Fridays 14^{00} - 15^{30}

- **Head tutor:**
  - Dr. Slavomira (Sally) Stefkova (CS 30.23 / R. 9-22)
- **Tutor team of post-docs and PhD students:**
  - Dr. Giacomo de Pietro, Dr. Soureek Mitra, Dr. Raquel Quishpe, Alessandro Brusamolino, Greta Heine, Tim Voigtländer
- **Tutorials and exercise sheets:**
  - First tutorial: **today 21.04 (only organizational material)**
  - Exercise sheets uploaded on Friday, 1 week prior to tutorial
  - First exercise sheet **already uploaded** and will be **discussed on 04.05 & 05.05**

Everything in this color is hyperlinked (in all lectures), *so click for more in your free time*.

- **Sign-up** for the Lecture & Computerpraktikum in **ILIAS** today

Exercises / LP rules

- **10 exercise sheets** with a total of **16 obligatory questions**
  - 6 LP: 11 / 16 Testate
  - 8 LP: 13 / 16 Testate + extra exercise: **“Higgs Challenge”**

- **Details to be discussed this afternoon in the tutorial introduction:**
  - To work on the exercises, a Jupyter Hub server is provided which can be accessed from any device via a browser under the link **https://jupytermachine.etp.kit.edu**. Use your KIT account credentials to log in. Your KIT account must be registered for access of the Physik-Pool. If this is not the case, please register your account under **https://comp.physik.kit.edu/Account/**. On this page you can also find a link for the prolongation of an existing account.

  - Once logged on to the Jupyter Hub, you can spawn a server instance to work on. Choose the datenauswertung image and spawn your server. From this server, you will be able to access your home directory of the Physik-Pool in the File Browser on the left side of your Jupyter Lab window.

  - The exercises will be provided as Python 3 Jupyter Notebooks (e.g. Exercise1.ipynb). These contain the instructions, additional hints and templates you can use to solve the exercises. They will be provided to you by means of the git repository **https://gitlab.etp.kit.edu/Lehre/dataanalysisexercises_forstudents.git**. which will be updated each week to contain the current exercise.
  - It is assumed that you followed the course Rechnernutzung in der Physik and are familiar with the basics of Linux, Python and ROOT. Some links are provided on the web page to refresh this knowledge. Additional hints will be provided in the exercises notebooks.

= yes

- **Exam:**
  - If you plan have this lecture examined before August together with another course, please contact Roger and Pablo via E-Mail ASAP.

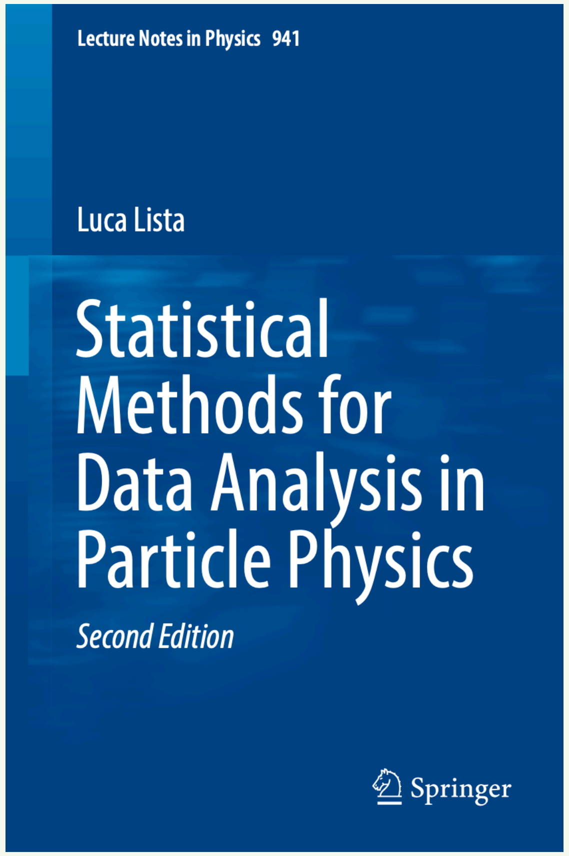
Course material

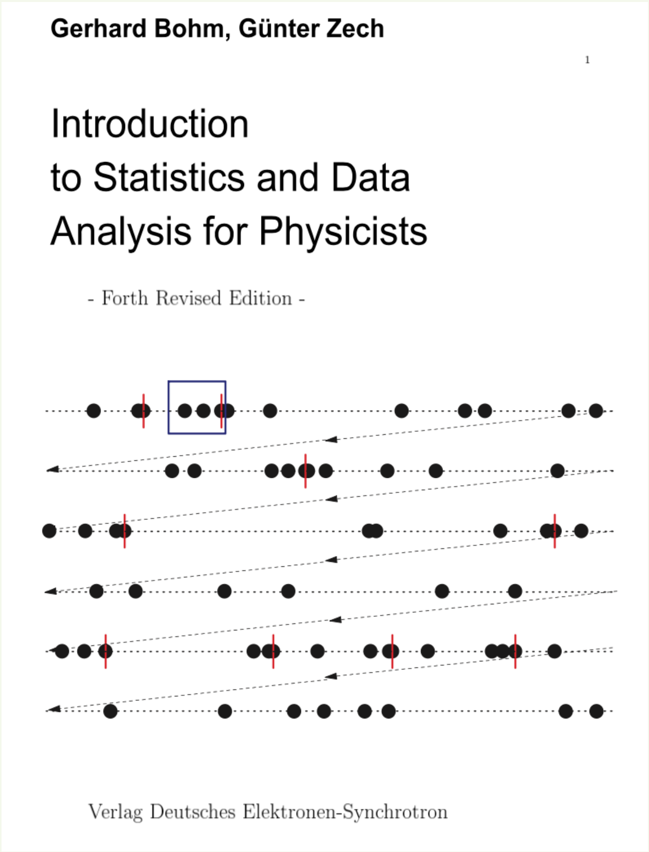
- Annotated lecture slides will be uploaded to ILIAS after each lecture
  - /Lecture material
- Reading material will be uploaded to ILIAS & assigned weekly
  - /Reading material
    - /Textbooks
    - Folders for each lecture with relevant papers to review (some elective)
      - /L01
      - /L02
      - …

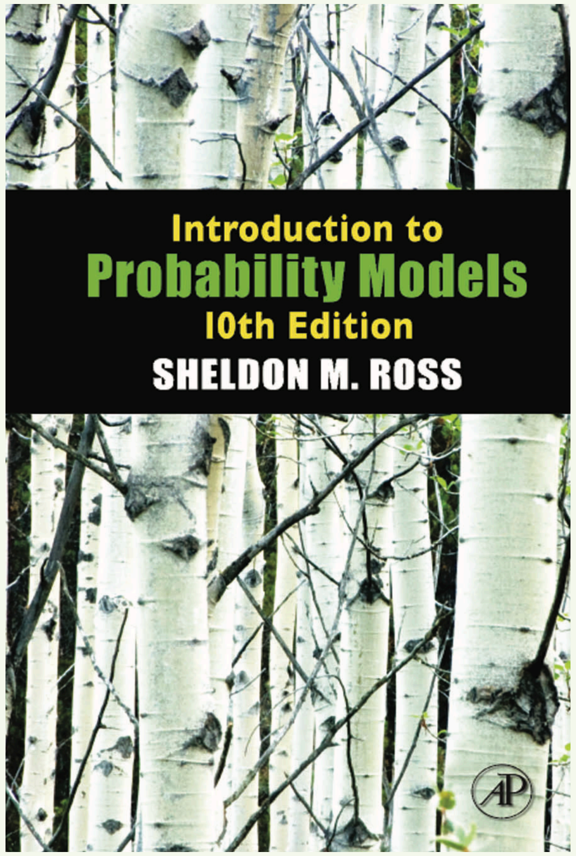
Recommended books

Digital copies uploaded to ILIAS: /Reading material/Textbooks

Heavily sourced







Slides, videos, and exercises on Cowan's website

Typical lecture: 11^{30} - 13^{00}

- First half (11^{30} - 12^{10}): **40’**
  - **5’** recap of last lecture
  - **15’** slides
  - **5’** discuss last lectures Quiz
  - **15’** slides
- Break (12^{10} - 12^{15}): **5’**
- Second half (12^{15} - 13^{00}): **45’**
  - slides and quiz

- **Important:**
  - Please tell me to slow down if I am too fast.
  - Feel free to interrupt me if you have a question during lecture.
  - You will ask questions I do not know the answer to. Whenever this happens, I will try to answer your question in the next lecture.

Quizzes during lecture

- Helpful for us to gauge how you’re doing and adjust speed/content accordingly.
- I will rely on **YOU** to monitor your progress. This is important because if you cannot answer the quiz questions by the end of the semester, you will likely not do well if you decide to include this lecture in your oral exam.
- So, we’ll take some time during lecture to take the quizzes and then we will review the answers together the following week.
- Make sure to have a pen and paper ready to write down your answers.
- **You won’t be graded on them.**

Curriculum

#	Lecture date	Lecture Topic
1	21.4	**Fundamental concepts I**
2	28.4	**Fundamental concepts II**
3	5.5	**Monte Carlo method & production of random distributions**
4	12.5	**Parameter estimation & maximum likelihood**
5	19.5	**Chi-square method**
6	26.5	**Hypothesis tests & Neyman Pearson**
		Semester Break
7	9.6	**Confidence intervals**
8	16.6	**Limit setting & unfolding**
9	23.6	**Event classification - Introduction and perceptron**
10	30.6	**Classification with the multilayer perceptron**
11	7.7	**Neural network training**
12	14.7	**Training algorithms & regularization methods**
13	21.7	**Training validation**
14	28.7	**Advanced neural networks**

#	Hand out exercise	Exercise dates (due Fri.)	Computerpraktikum Topic
Ex.1	✓	4.5, 5.5	**Programming and standard error propagation**
Ex.2	5.5	11.5, 12.5	**Priors and Monte Carlo**
Ex.3	12.5	25.5, 26.5	**ML & Chi-square methods**
		Holiday on 18.05	
Ex.4	26.5	15.6, 16.6	**Combination of correlated measurements**
		Semester Break	
		Holiday on 08.06	
Ex.5	16.6	22.6, 23.6	**Data parameterization & minimization**
Ex.6	23.6	29.6, 30.6	**Hypothesis testing & parameter estimation**
Ex.7	30.6	6.7, 7.7	**Confidence intervals**
Ex.8	7.7	13.7, 14.7	**Unfolding**
Ex.9	14.7	20.7, 21.7	**Multivariate classification**
Ex.10	21.7	27.7, 28.7	**Deep learning**

Introduction to tutorials Today @14:00 online: Zoom link (passcode: 347693)

Curriculum in a box

Monte Carlo Method

Resolution: Folding and Unfolding

Fundamentals of statistics (Probability, PDFs, Bayes, Moments)

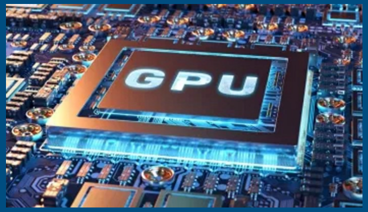
Hypothesis Testing and Neyman-Pearson

Parameter estimation: Likelihood and χ^2

Multivariate Methods: Decision Trees and Neural Networks

Deep Learning

Machine Learning

= Machine (computing power) 

+ Probability **This course**

+ Data ***Everywhere***

Including ourselves

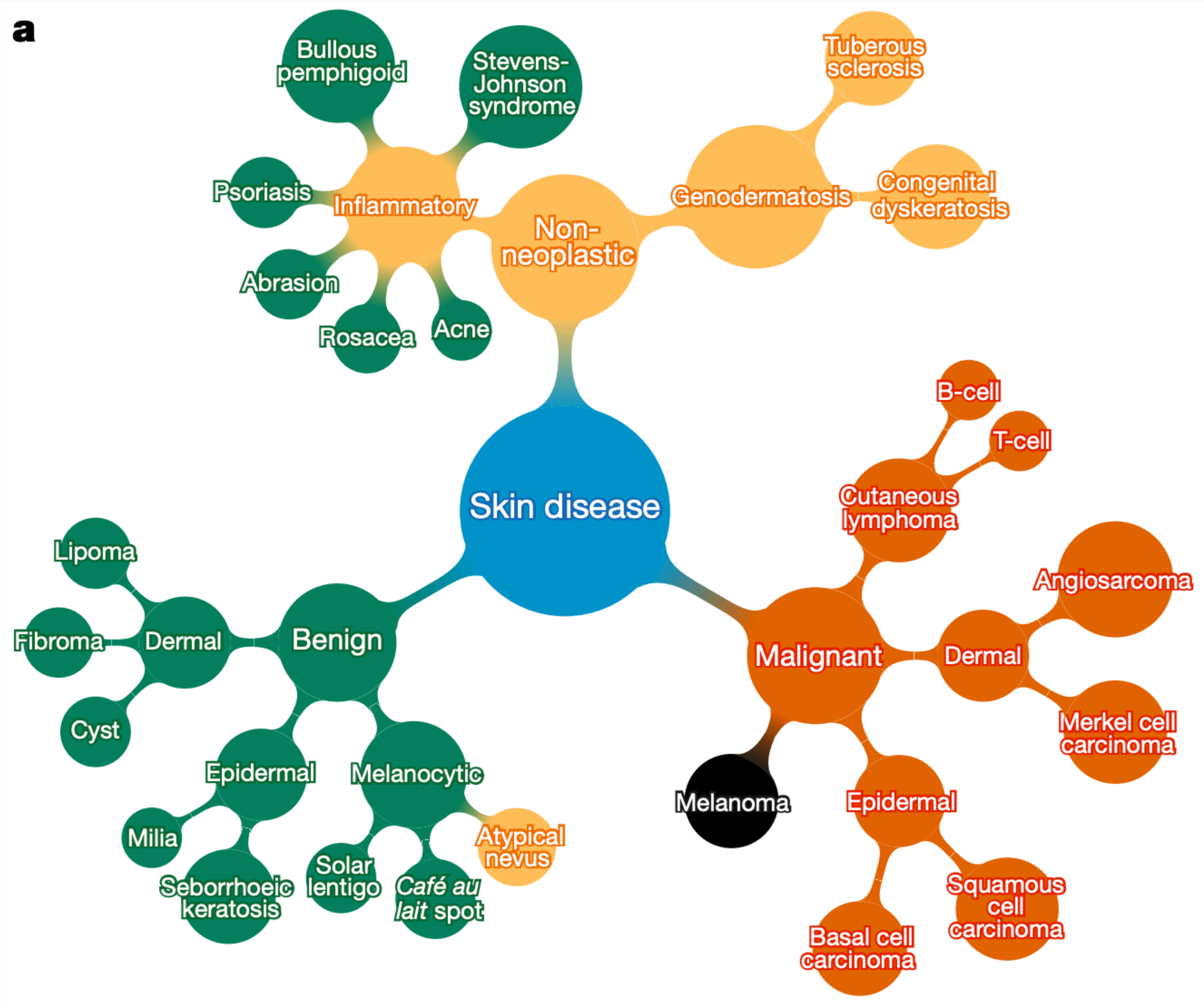
Machine learning algorithm

Data → Build a *probabilistic model* → Do one thing

Classification

Chihuahua or muffin?

Seriously though… Very useful!



A ML algorithm performs better than the best dermatologists

E. Andre et al., "Dermatologist-level classification of skin cancer with deep neural networks" Nature 542.7639 (2017): 115-118

PDF on ILIAS: /Reading material /L01

Decision-making

Over the course of millions of AlphaGo vs AlphaGo games, the system progressively learned the game of Go from scratch, accumulating thousands of years of human knowledge during a period of just a few days. AlphaGo Zero also discovered new knowledge, developing unconventional strategies and creative new moves that echoed and surpassed the novel techniques it played in the games against Lee Sedol and Ke Jie.

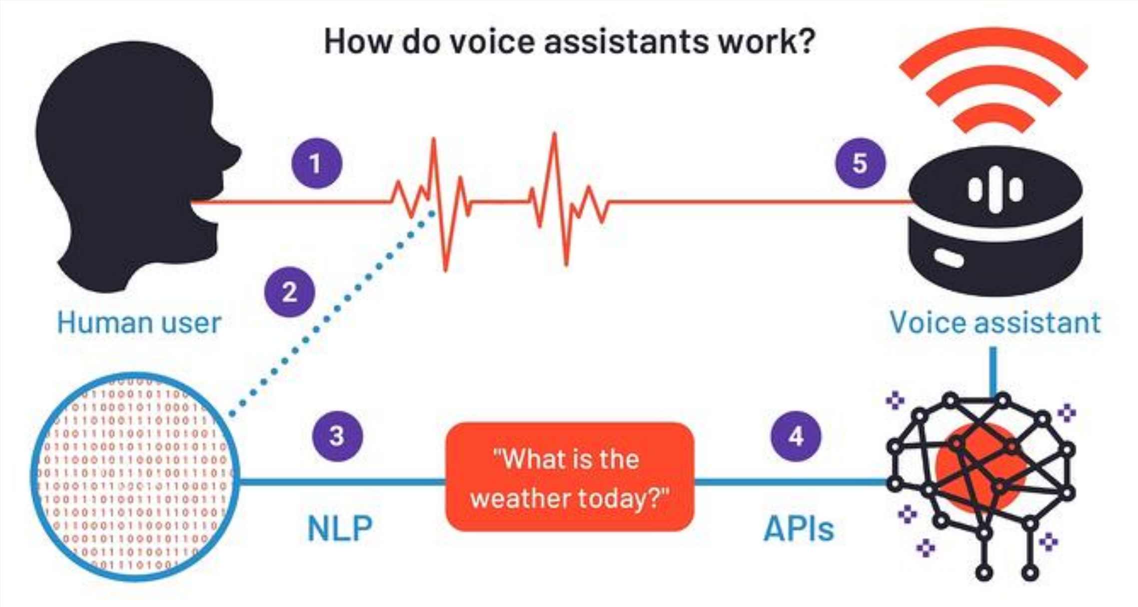
https://deepmind.com/blog/article/alphago-zero-starting-scratch

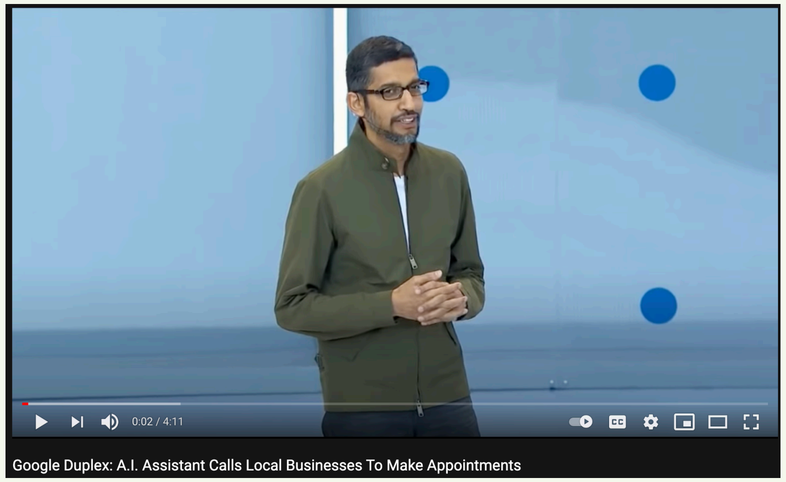
Natural language processing

Augmented reality machine translation

https://www.researchgate.net/publication/224218354_TranslatAR_A_mobile_augmented_reality_translator

Voice assistants:
Voice to text to answer





(Creepy) Google assistant demo
https://www.youtube.com/watch?v=D5VN56jQMWM.

Next level AI

Probability and music

Decoding Beethoven’s music style using data science

https://actu.epfl.ch/news/decoding-beethoven-s-music-style-using-data-scienc/

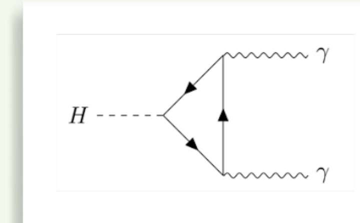
“The study finds that very few chords govern most of the music, a phenomenon that is also known in linguistics, where very few words dominate language corpora.... It characterizes Beethoven's specific composition style for the String Quartets, through a distribution of all the chords he used, how often they occur, and how they commonly transition from one to the other.”

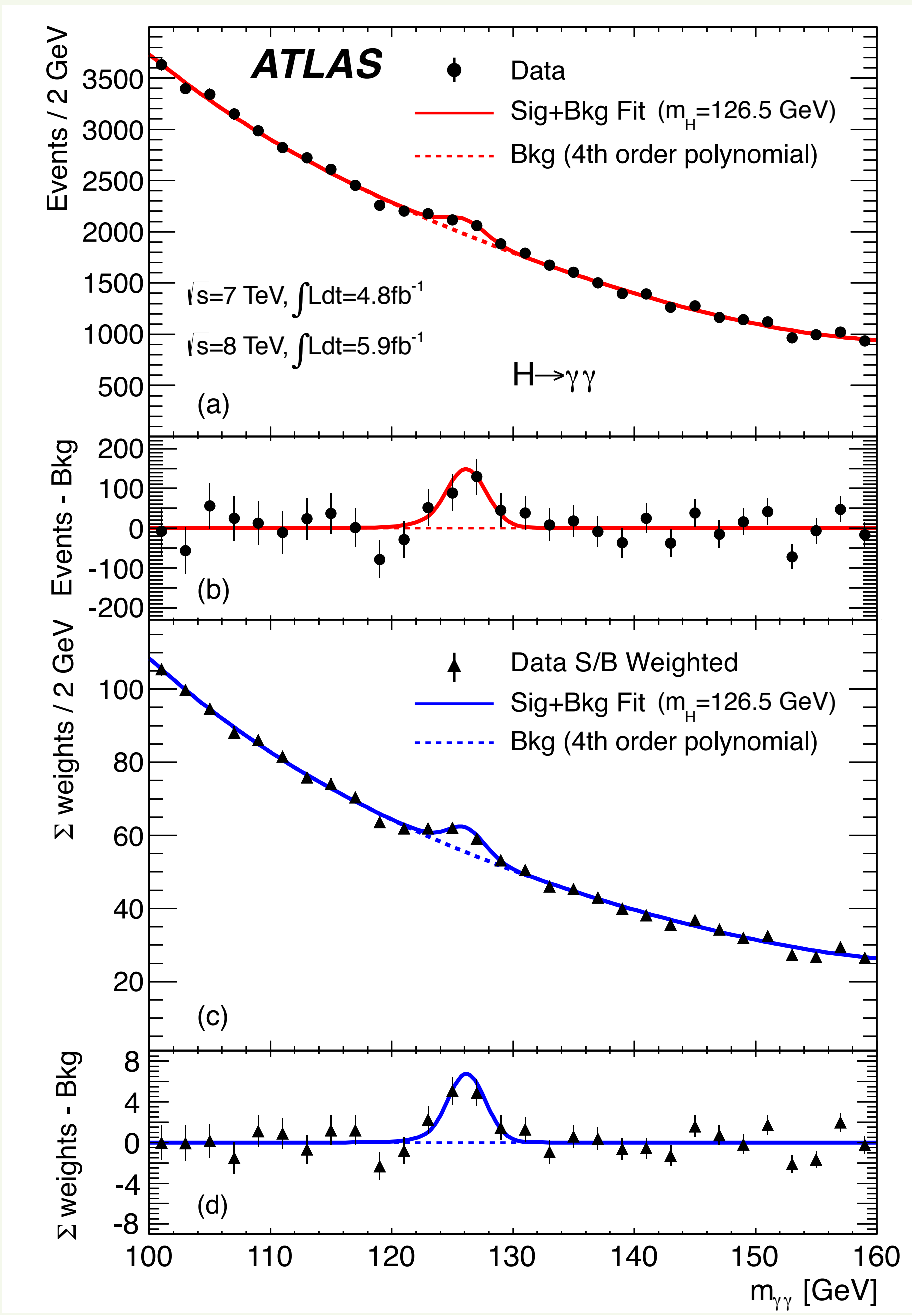
Manuscript of Große Fuge in b flat major

Probability is *more* than just machine learning

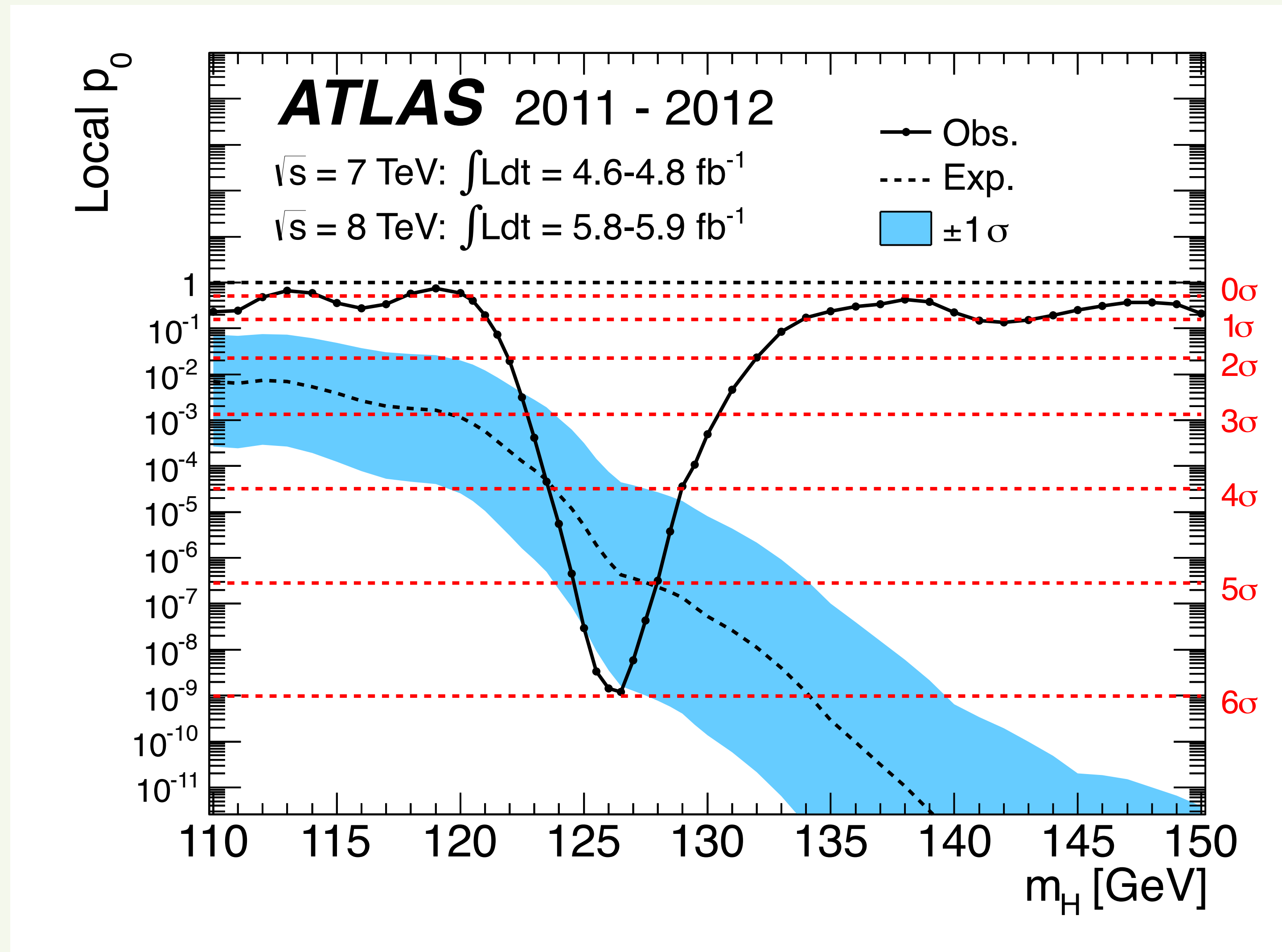
Why is statistical data analysis important?

- Fundamental for many things:
  - Discovery of the **Higgs boson** and gravitational waves





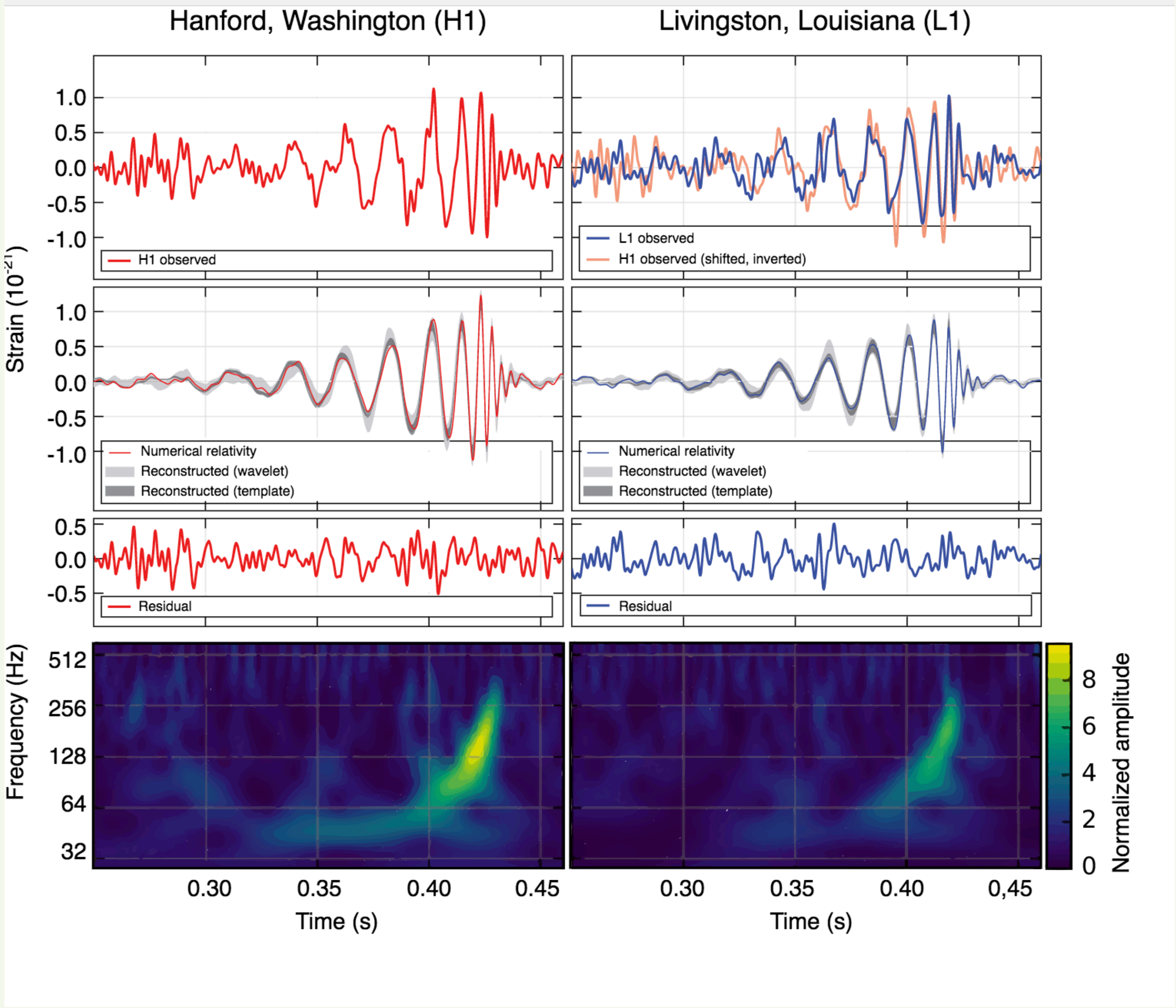
https://doi.org/10.1016/j.physletb.2012.08.020

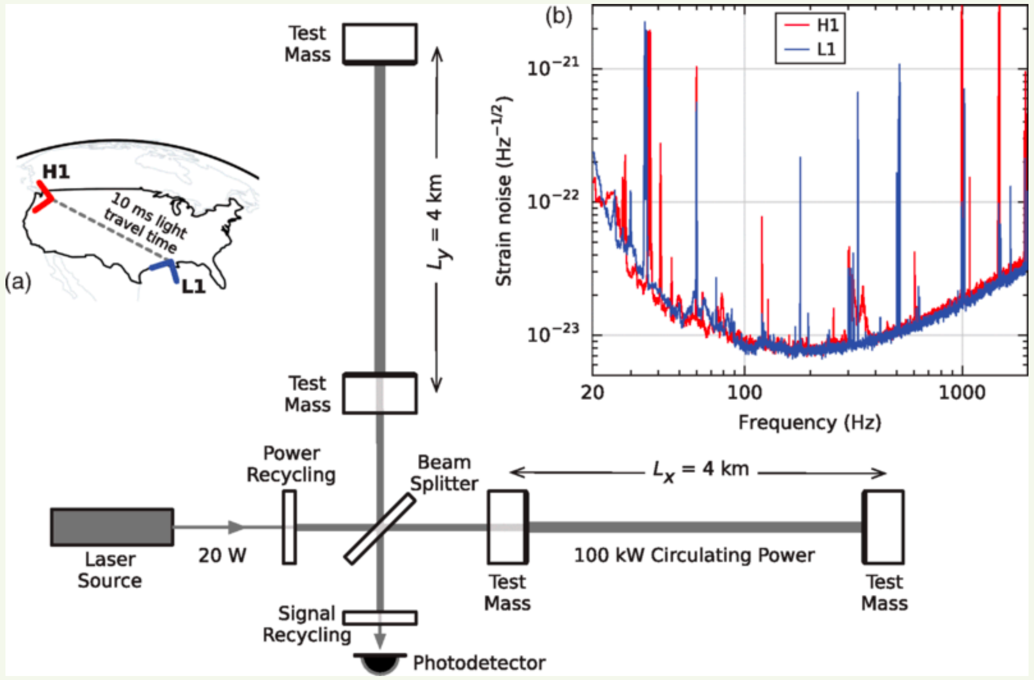


A nice article on why 5σ for discovery

Why is statistical data analysis important?

- Fundamental for many things:
  - Discovery of the Higgs boson and **gravitational waves**

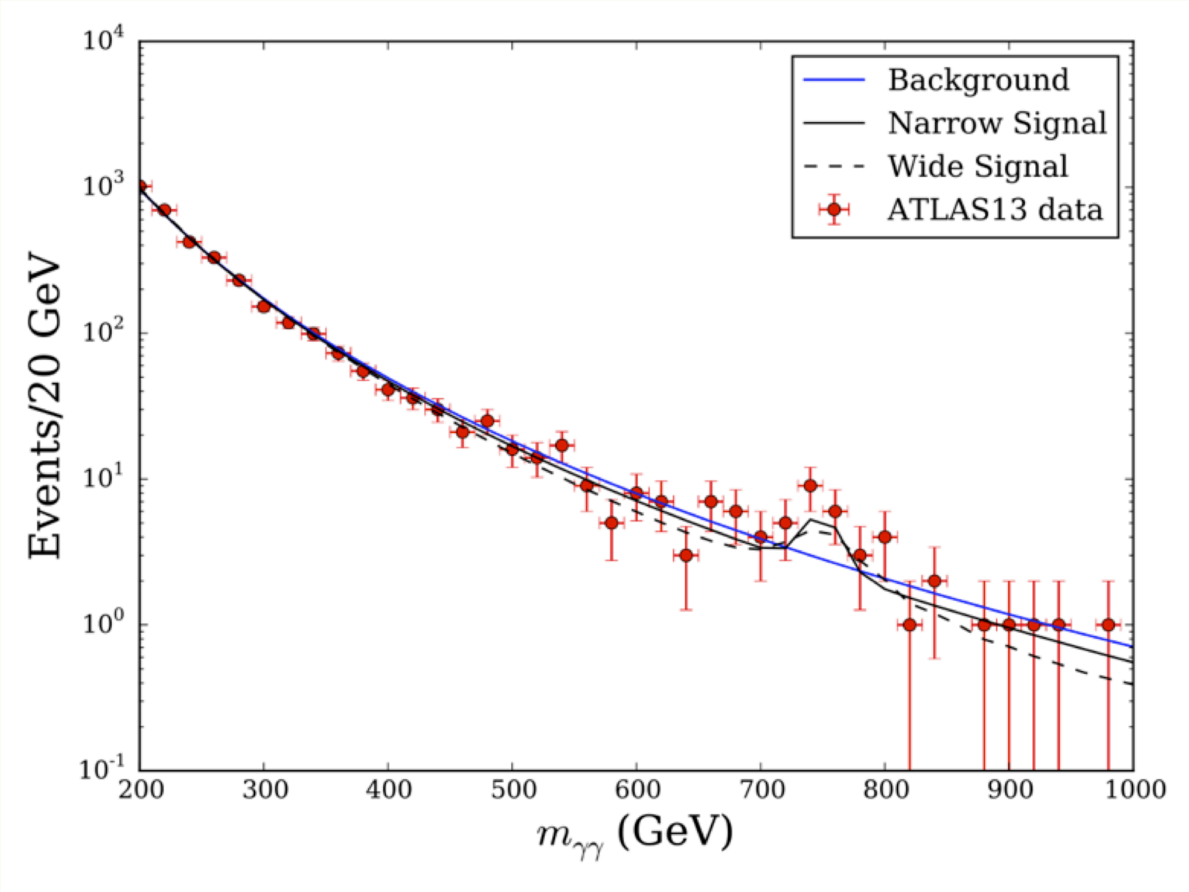


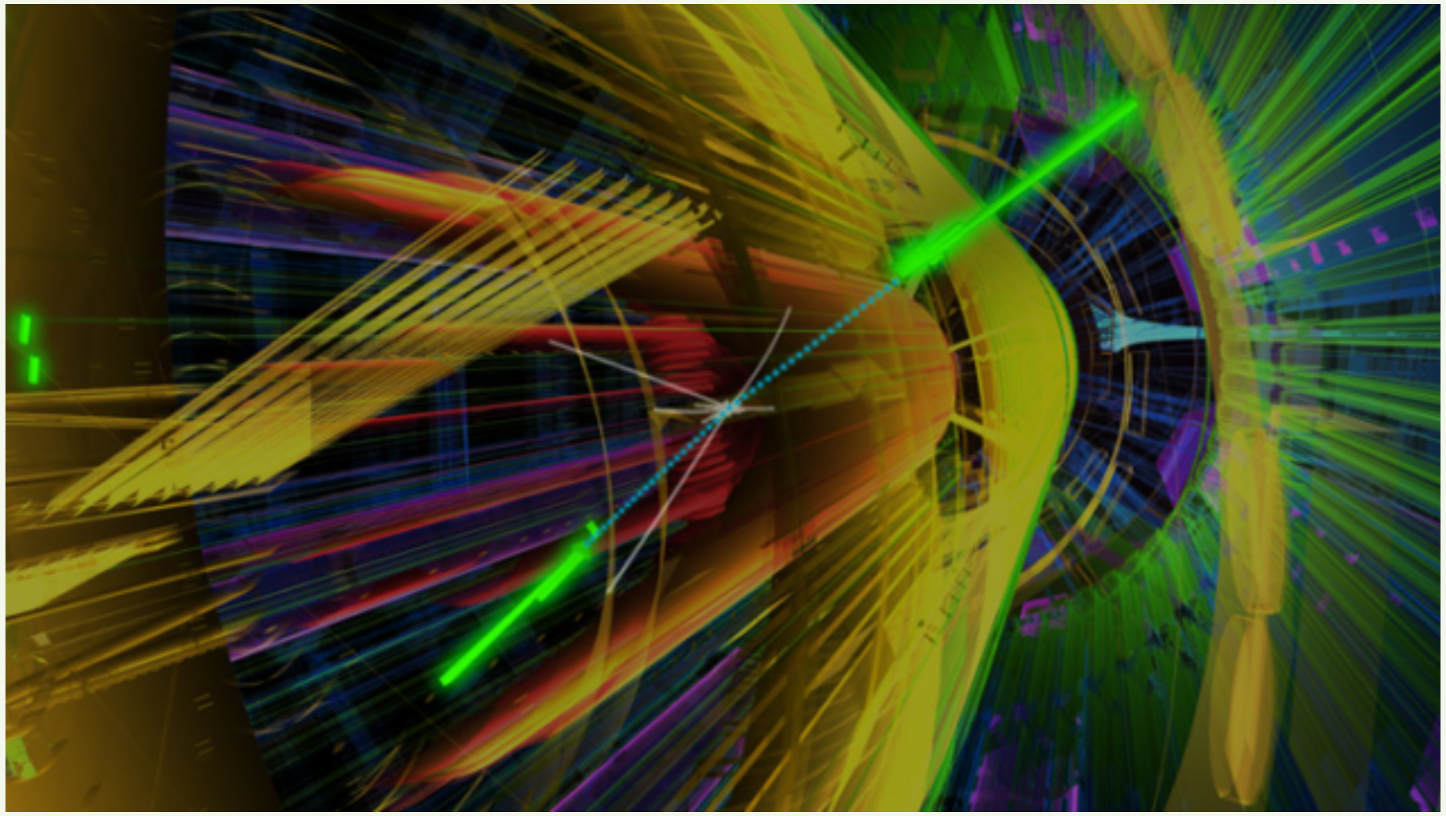


Why is statistical data analysis important?

- Fundamental for many things:
  - Stringent test for claiming new phenomena.

LHC's 750 GeV 'bump'





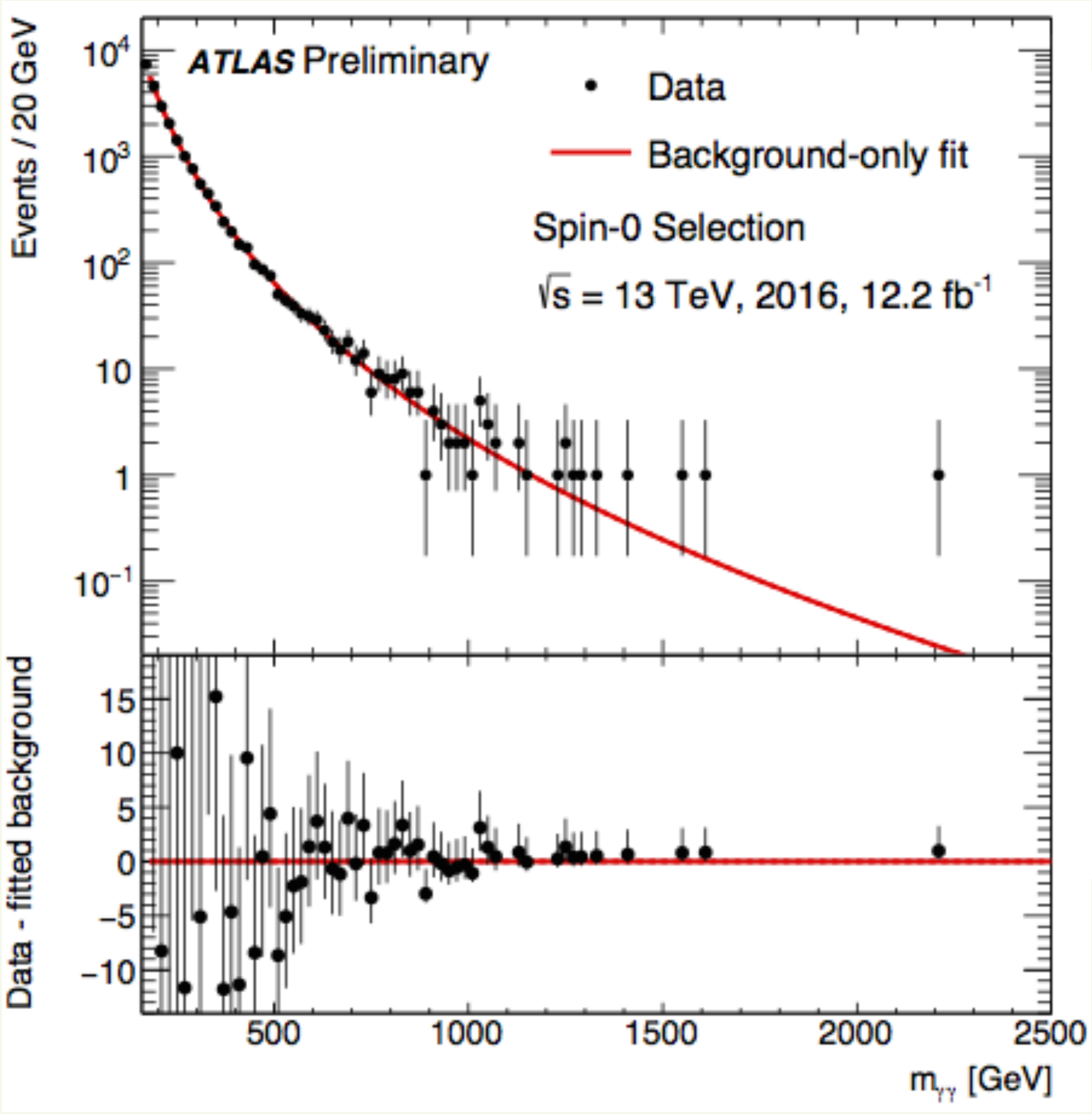
>250 papers submitted to arXiv in first 6 months!

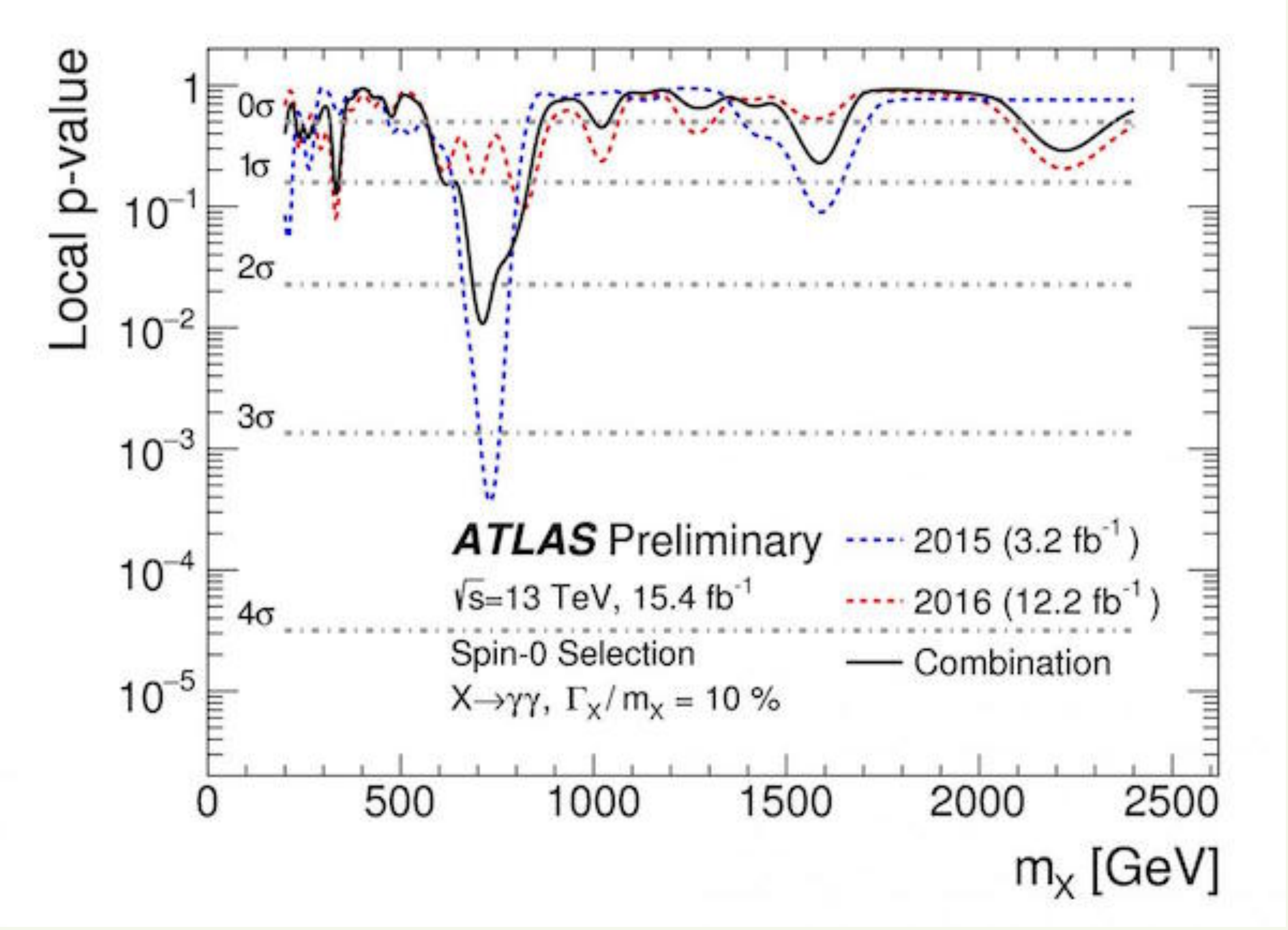
https://physicsworld.com/a/theorizing-about-the-lhcs-750-gev-bump/

Why is statistical data analysis important?

- Fundamental for many things:
  - Stringent test for claiming new phenomena.

LHC's 750 GeV 'bump'





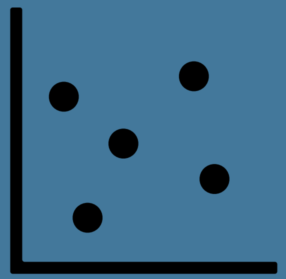
http://bostonreview.net/books-ideas/matthew-buckley-search-new-physics-cern-part-8

And much more!

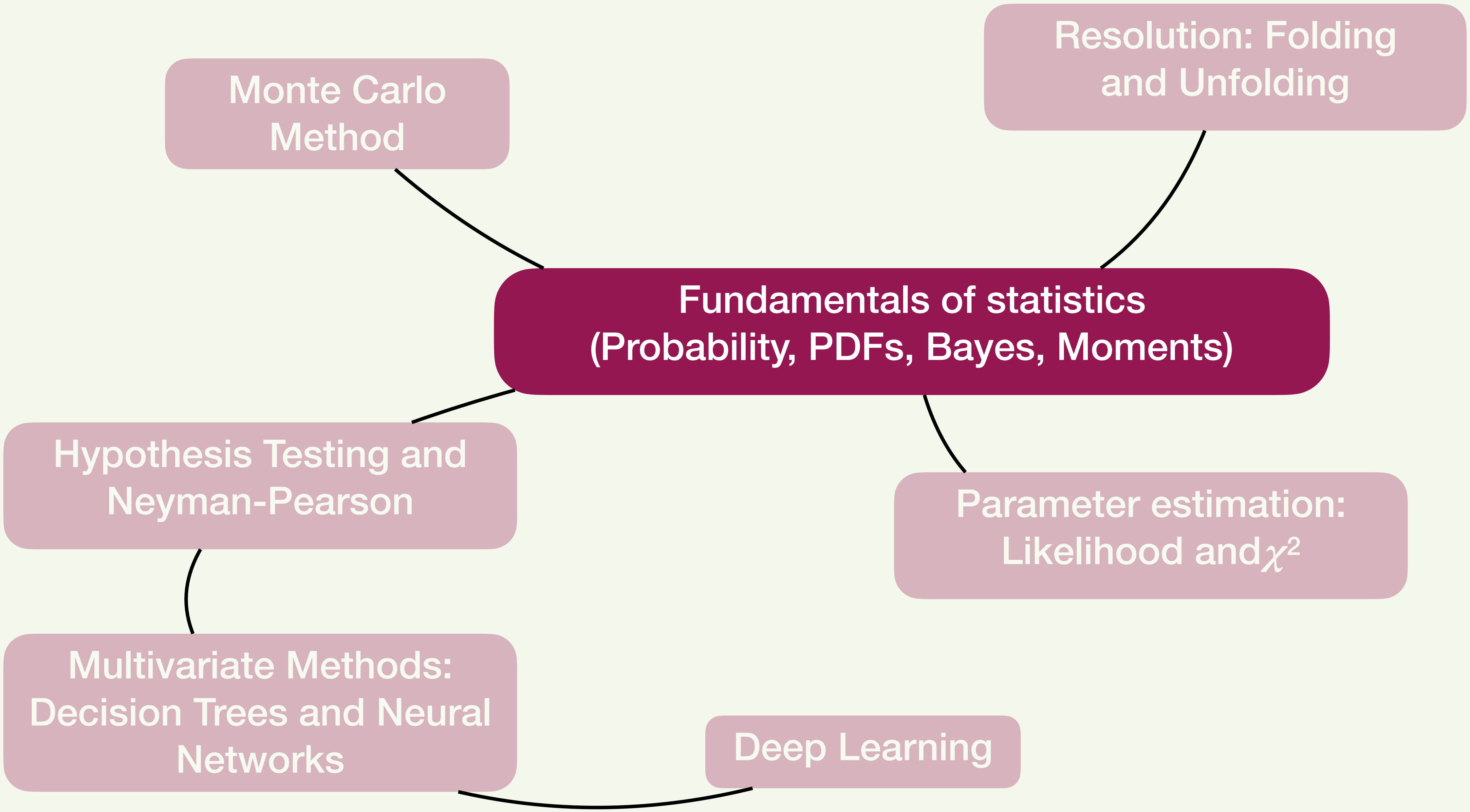
We'll get there,

but fundamentals first…

Lecture 1: Fundamental Concepts

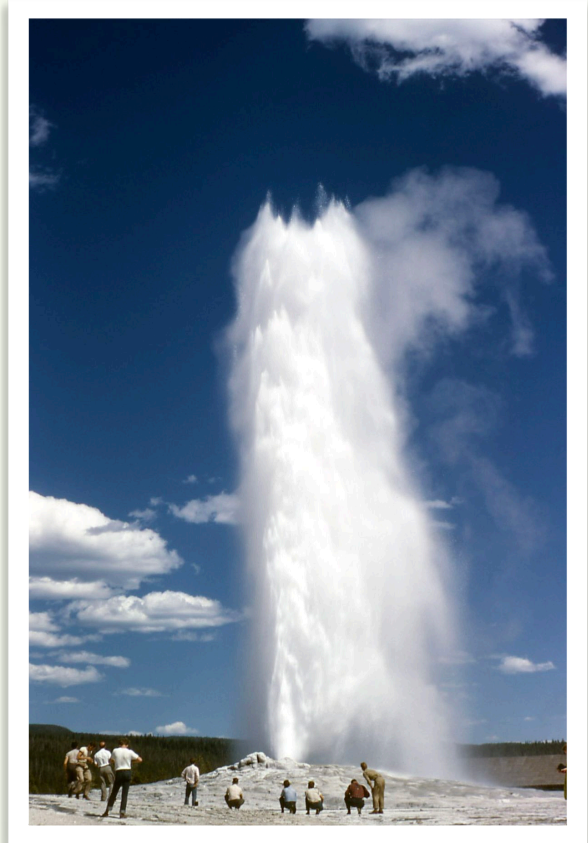


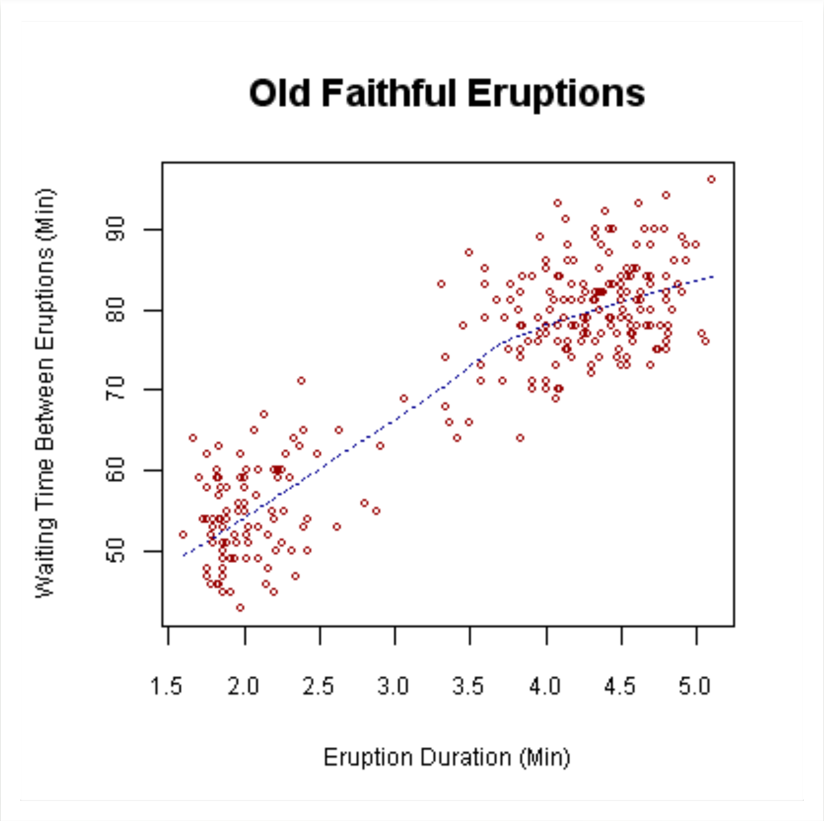
A brief recap of what you should already know

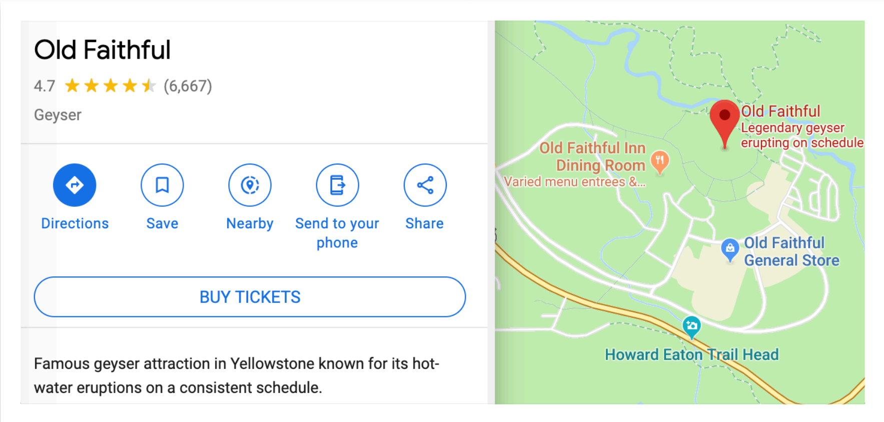

Fundamentals of statistics
(Probability, PDFs, Bayes, Moments)
Monte Carlo Method
Resolution: Folding and Unfolding
Hypothesis Testing and Neyman-Pearson
Parameter estimation: Likelihood and χ^2
Multivariate Methods: Decision Trees and Neural Networks
Deep Learning

Descriptive statistics

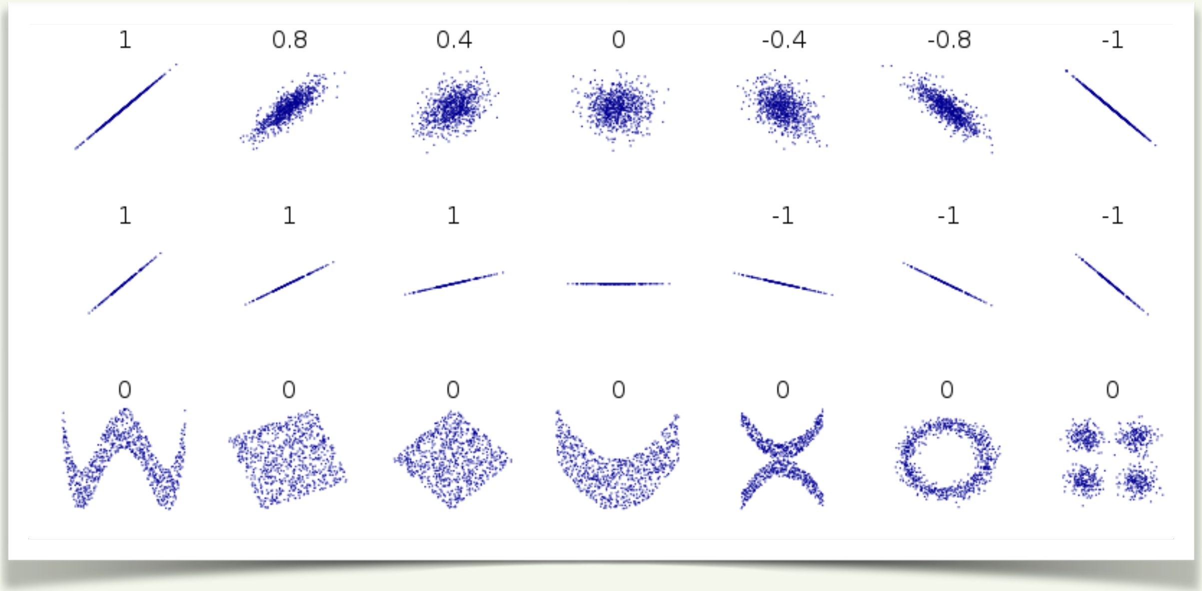
- Statistical methods used to describe and analyze measured data from experiments
  - **Description of data (descriptive statistics)**
    - Describe data w/o knowing fundamental principles of what is measured
    - Characterize data with few but descriptive parameters, but no aim to use data to learn about the population that the data is thought to represent.

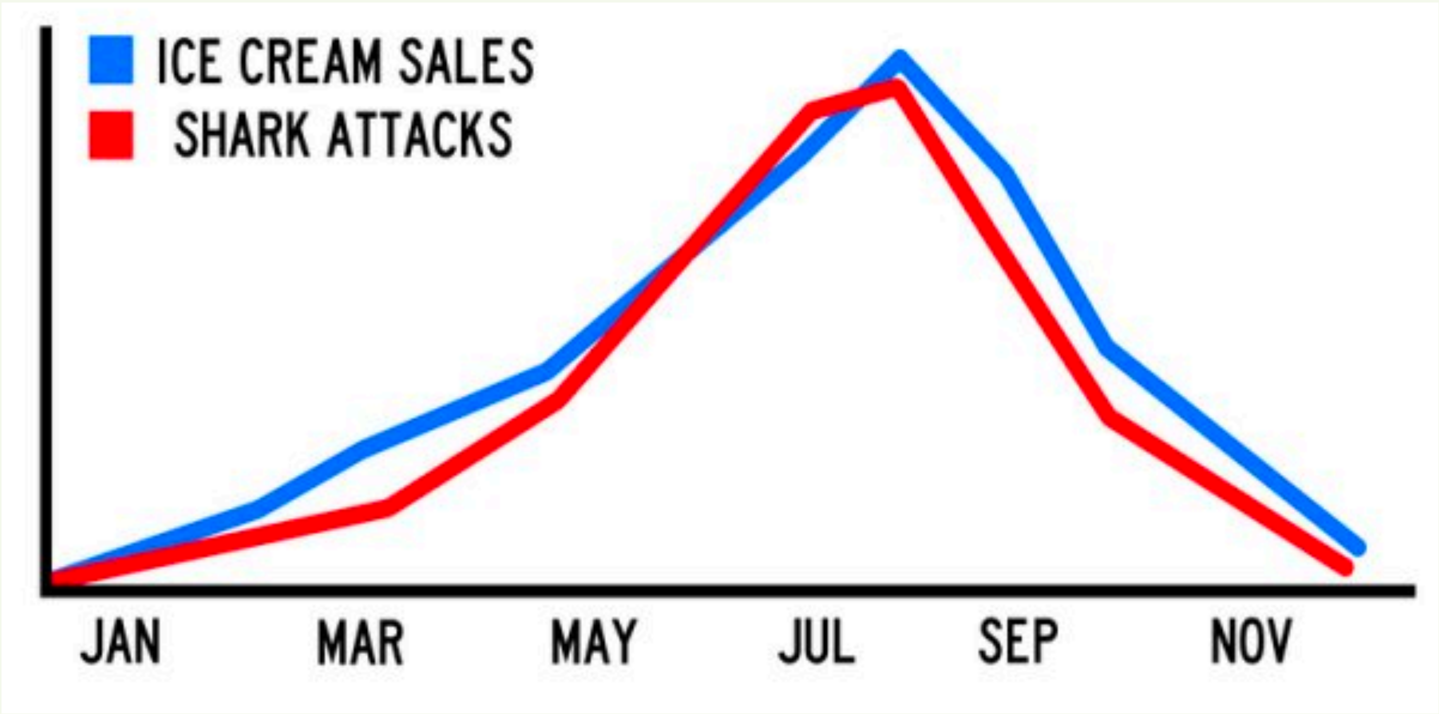


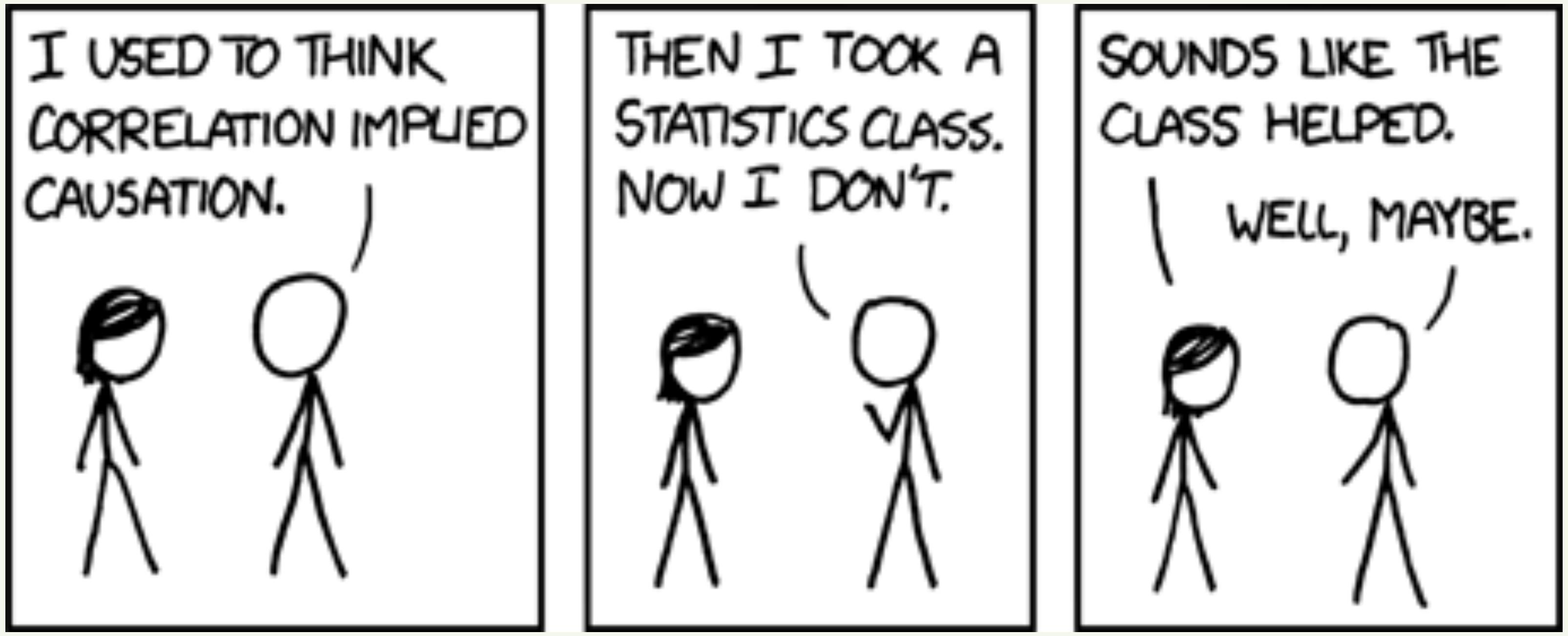




Descriptive statistics: *Careful!*





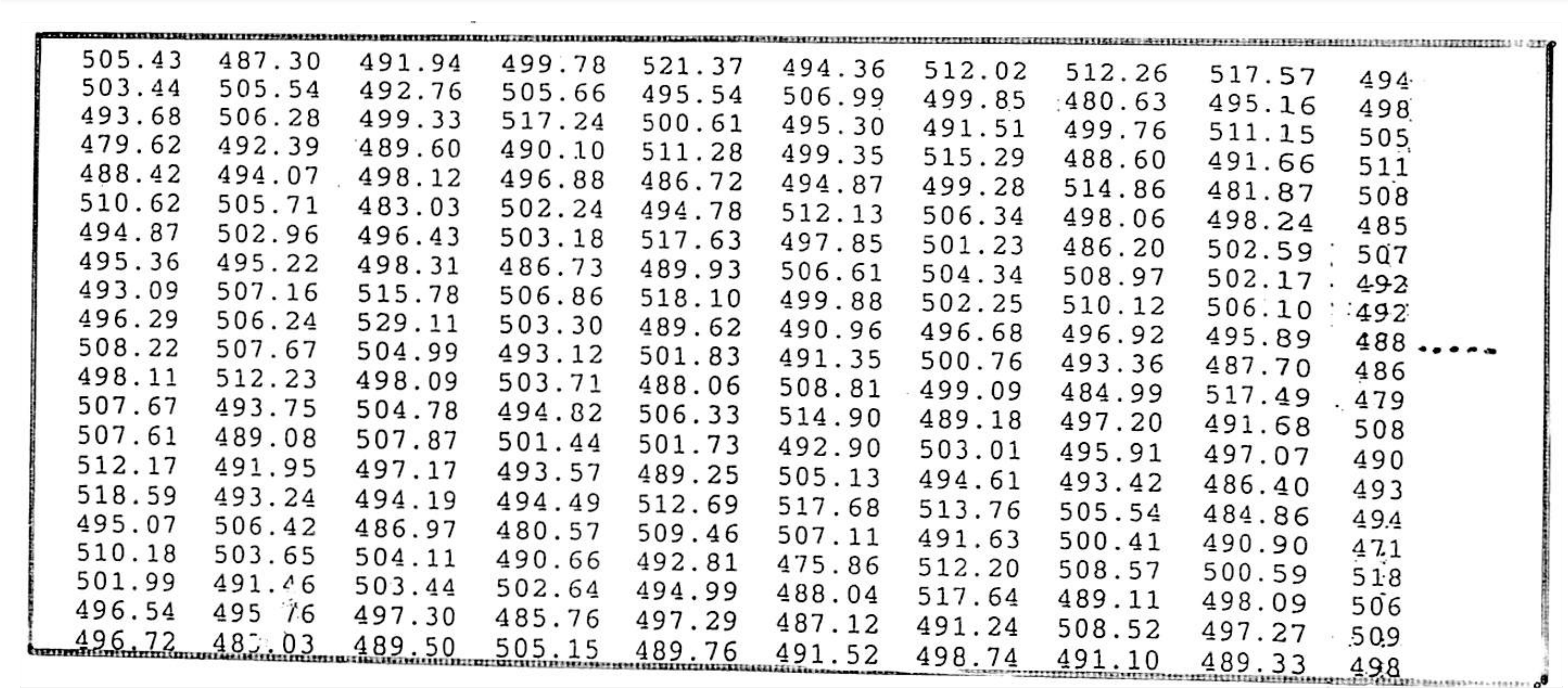


Physical measurement

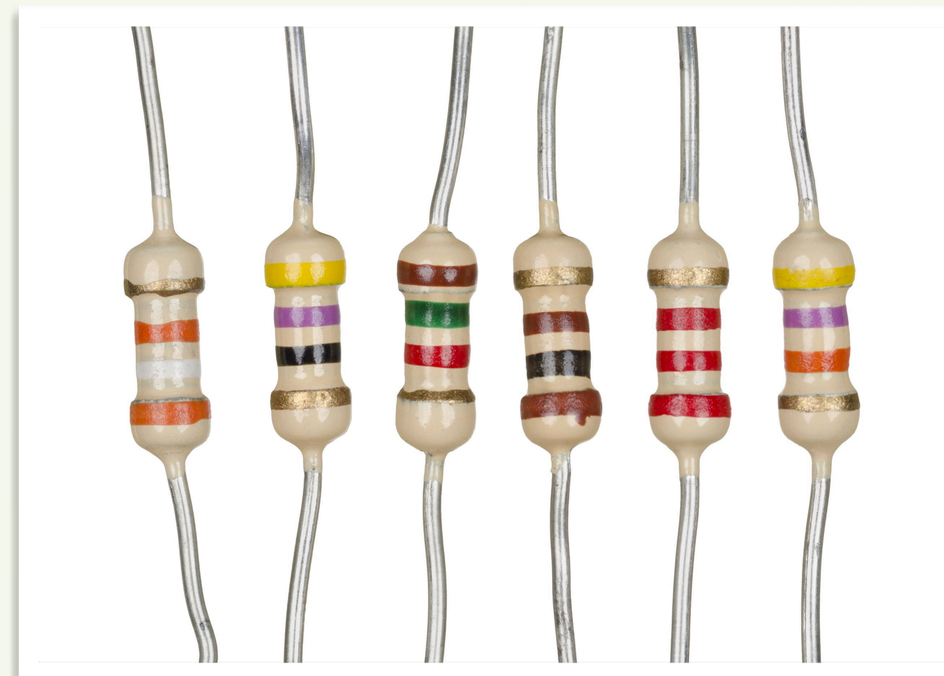
- Statistical methods used to describe and analyze measured data from experiments
  - **Description of data (descriptive statistics)**
    - Describe data w/o knowing fundamental principles of what is measured
    - Characterize data with few but descriptive parameters
  - **Physical Measurements (inductive statistics)**
    - Every physical measurement has an associated uncertainty
      - Typically: experimental resolution

Physical measurement: *Example*

- Block of data: Measured resistance in Ohm:

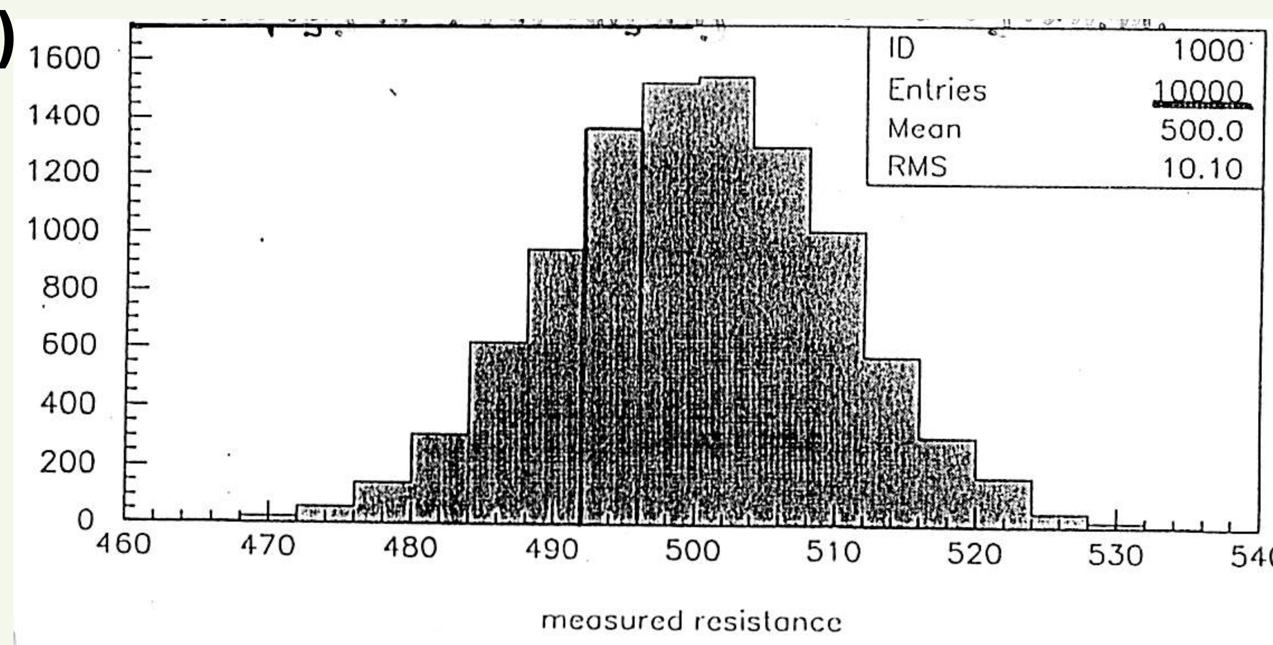


505.43	487.30	491.94	499.78	521.37	494.36	512.02	512.26	517.57	494
503.44	505.54	492.76	505.66	495.54	506.99	499.85	480.63	495.16	498
493.68	506.28	499.33	517.24	500.61	495.30	491.51	499.76	511.15	505
479.62	492.39	489.60	490.10	511.28	499.35	515.29	488.60	491.66	511
488.42	494.07	498.12	496.88	486.72	494.87	499.28	514.86	481.87	508
510.62	505.71	483.03	502.24	494.78	512.13	506.34	498.06	498.24	485
494.87	502.96	496.43	503.18	517.63	497.85	501.23	486.20	502.59	507
495.36	495.22	498.31	486.73	489.93	506.61	504.34	508.97	502.17	492
493.09	507.16	515.78	506.86	518.10	499.88	502.25	510.12	506.10	492
496.29	506.24	529.11	503.30	489.62	490.96	496.68	496.92	495.89	488
508.22	507.67	504.99	493.12	501.83	491.35	500.76	493.36	487.70	486
498.11	512.23	498.09	503.71	488.06	508.81	499.09	484.99	517.49	479
507.67	493.75	504.78	494.82	506.33	514.90	489.18	497.20	491.68	508
507.61	489.08	507.87	501.44	501.73	492.90	503.01	495.91	497.07	490
512.17	491.95	497.17	493.57	489.25	505.13	494.61	493.42	486.40	493
518.59	493.24	494.19	494.49	512.69	517.68	513.76	505.54	484.86	494
495.07	506.42	486.97	480.57	509.46	507.11	491.63	500.41	490.90	471
510.18	503.65	504.11	490.66	492.81	475.86	512.20	508.57	500.59	518
501.99	491.46	503.44	502.64	494.99	488.04	517.64	489.11	498.09	506
496.54	495.76	497.30	485.76	497.29	487.12	491.24	508.52	497.27	509
496.72	48[illegible].03	489.50	505.15	489.76	491.52	498.74	491.10	489.33	498



- Probability distribution (**Histogram**):

N(x)

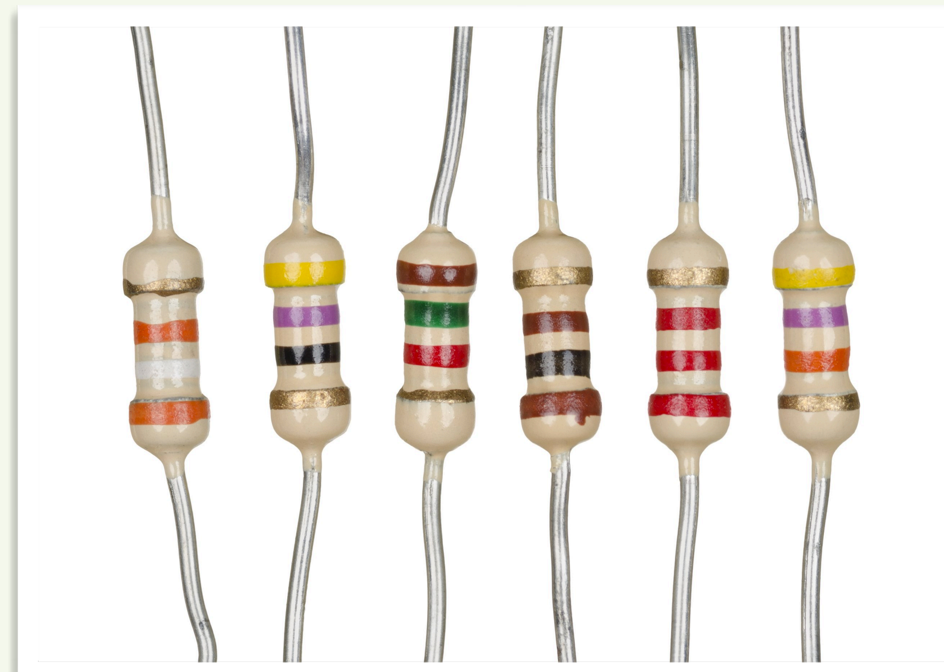


[Ohm]

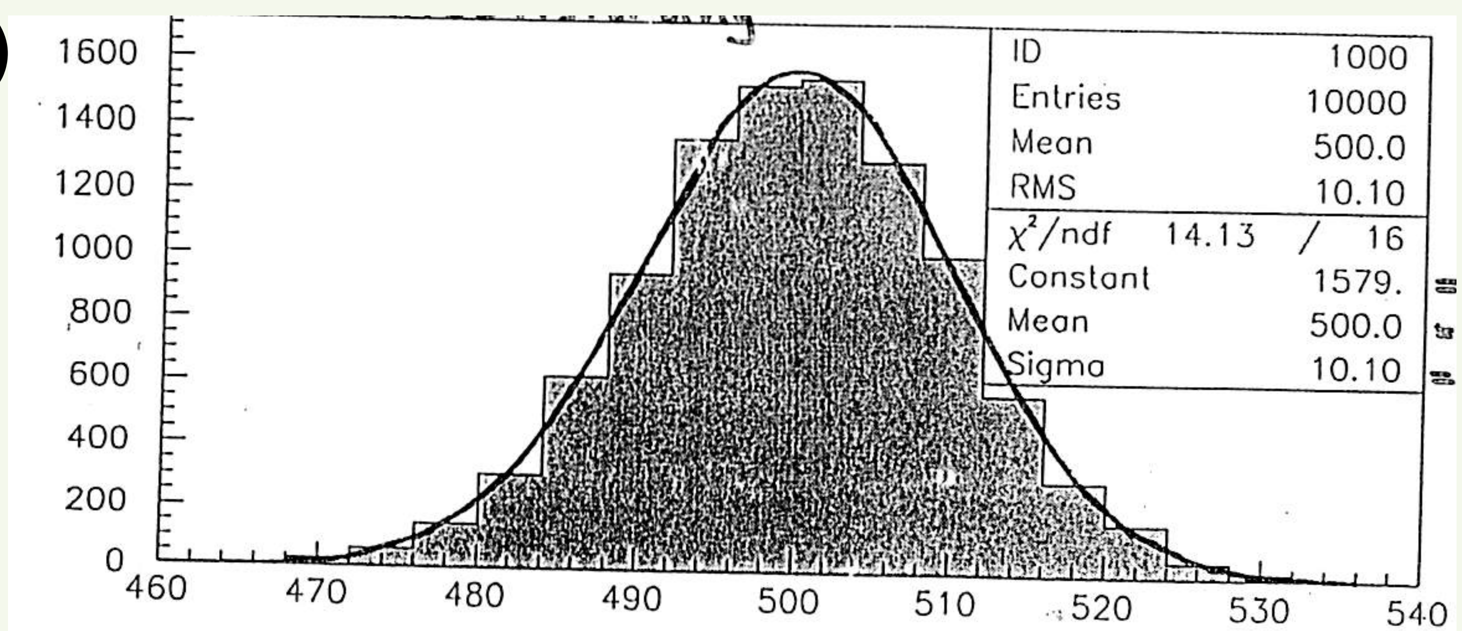
Physical measurement: *Example*

- Block of data: Measured resistance in Ohm:

505.43	487.30	491.94	499.78	521.37	494.36	512.02	512.26	517.57	494
503.44	505.54	492.76	505.66	495.54	506.99	499.85	480.63	495.16	498
493.68	506.28	499.33	517.24	500.61	495.30	491.51	499.76	511.15	505
479.62	492.39	489.60	490.10	511.28	499.35	515.29	488.60	491.66	511
488.42	494.07	498.12	496.88	486.72	494.87	499.28	514.86	481.87	508
510.62	505.71	483.03	502.24	494.78	512.13	506.34	498.06	498.24	485
494.87	502.96	496.43	503.18	517.63	497.85	501.23	486.20	502.59	507
495.36	495.22	498.31	486.73	489.93	506.61	504.34	508.97	502.17	492
493.09	507.16	515.78	506.86	518.10	499.88	502.25	510.12	506.10	492
496.29	506.24	529.11	503.30	489.62	490.96	496.68	496.92	495.89	488
508.22	507.67	504.99	493.12	501.83	491.35	500.76	493.36	487.70	486
498.11	512.23	498.09	503.71	488.06	508.81	499.09	484.99	517.49	479
507.67	493.75	504.78	494.82	506.33	514.90	489.18	497.20	491.68	508
507.61	489.08	507.87	501.44	501.73	492.90	503.01	495.91	497.07	490
512.17	491.95	497.17	493.57	489.25	505.13	494.61	493.42	486.40	493
518.59	493.24	494.19	494.49	512.69	517.68	513.76	505.54	484.86	494
495.07	506.42	486.97	480.57	509.46	507.11	491.63	500.41	490.90	471
510.18	503.65	504.11	490.66	492.81	475.86	512.20	508.57	500.59	518
501.99	491.46	503.44	502.64	494.99	488.04	517.64	489.11	498.09	506
496.54	495.76	497.30	485.76	497.29	487.12	491.24	508.52	497.27	509
496.72	48[illegible].03	489.50	505.15	489.76	491.52	498.74	491.10	489.33	498



- Parametrize the distribution



$$N(x) = A \cdot e^{-\frac{1}{2}\left(\frac{x-B}{C}\right)^2}$$

$= A$ normalisation

$= B$ estimator for true resistance

$= C$ dispersion parameter / estimator for true resolution

Fundamentals

- A couple of terms we should introduce formally:
  - **Uncertainty**: *Outcome of measurement varies unpredictably upon repetition due to*
    - **Resolution**
    - **Errors** in **measuring device** (Systematic Uncertainty)
    - **Fundamental property** of the system (e.g. QM)
  - **Random**: *Characterization of a system is random if*
    - it is **not known** or
    - **cannot** be **predicted with absolute certainty**

▶ Degree of randomness quantifiable with concept of **probability**

Quiz 1

Don't worry if you can't answer everything

- Five warm up questions in 5'
  - What's the probability to not toss five heads in a row in coin toss?
  - Name the following distributions:

$$Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad f(k;\lambda) = \Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

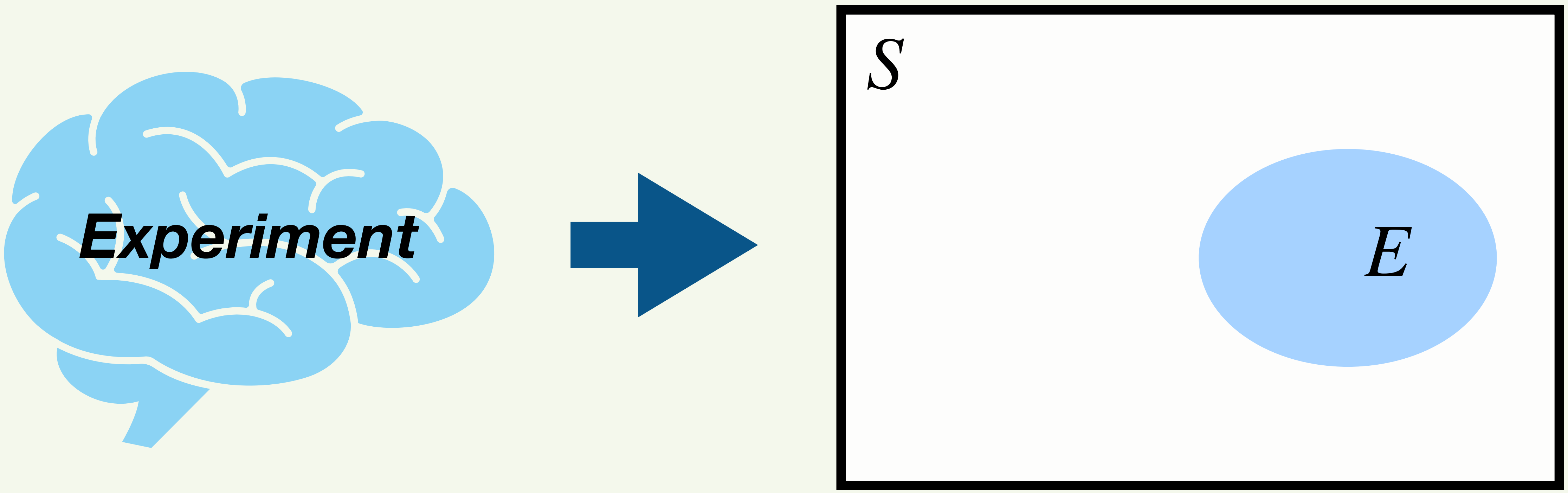
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]} \qquad f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$$

  - Mean and variance of a *n* independent random variables x_i
  - I know what Bayes' theorem is about (Yes/No).
  - Write down the definition of a χ^2 function.

Defining Probability

Key definitions

An experiment in probability:



Sample space, S: The set of all possible **outcomes** of an **experiment**

Event, E: Some subset of S $(E \subseteq S)$

Key definitions

Sample space, S

- Coin flip
  S = {Heads,Tails}

- Flipping 2 coins
  S = {(H,H), (H,T), (T,H), (T,T)}

- Roll of 6-sided die
  S = {1, 2, 3, 4, 5, 6}

- # of emails in a day
  $S = \{x \mid x \in Z, x \geq 0\}$

- FB hours in a day
  $S = \{x \mid x \in R, 0 \leq x \leq 24\}$

Event, E

- Flip lands heads
  E = {Heads}

- $\geq$ 1 head on 2 coin flips
  E = {(H,H), (H,T), (T,H)}

- Roll is 3 or less
  E = {1, 2, 3}

- Low email day ($\leq$ 20 mails)
  $E = \{x \mid x \in Z, 0 \leq x \leq 20\}$

- Wasted day ($\geq$ 5 FB hours)
  $E = \{x \mid x \in R, 5 \leq x \leq 24\}$

What is a probability?

A number between 0 and 1
to which we ascribe meaning.*

**our belief that an event* E *occurs.*

What is a probability?

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

n = # of total trials

$n(E)$ = # of trials where E occurs

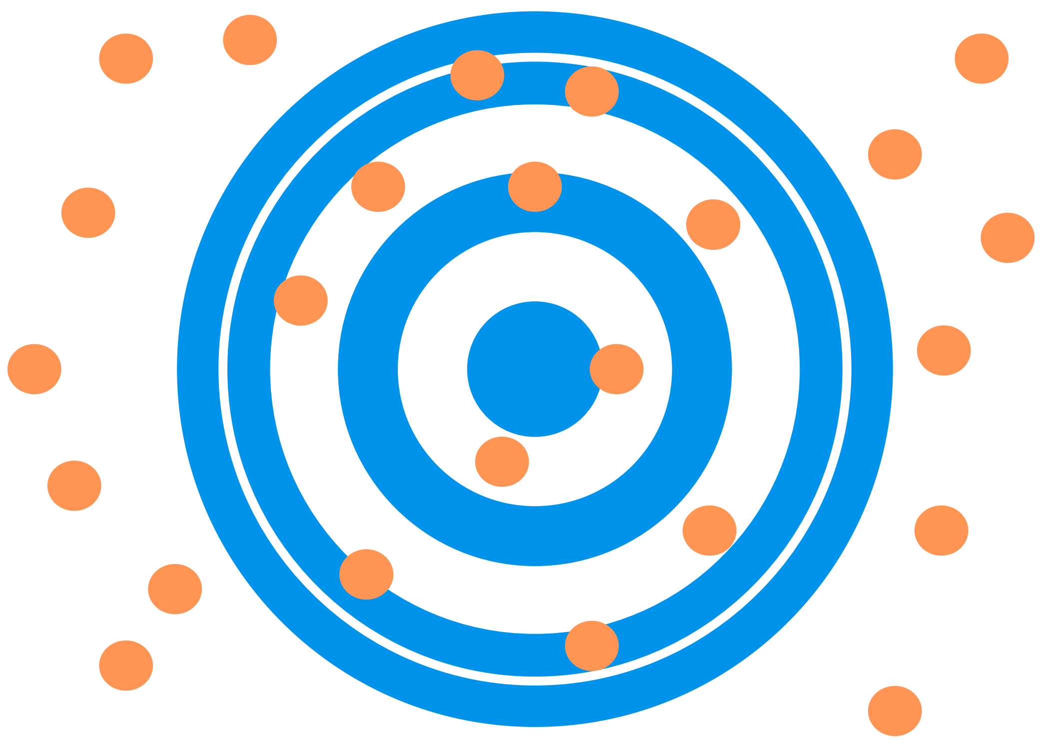
Frequentist

Relative frequency interpretation of probability

(We'll discuss the subjective probability interpretation later)

Let E = the set of outcomes when you hit the target

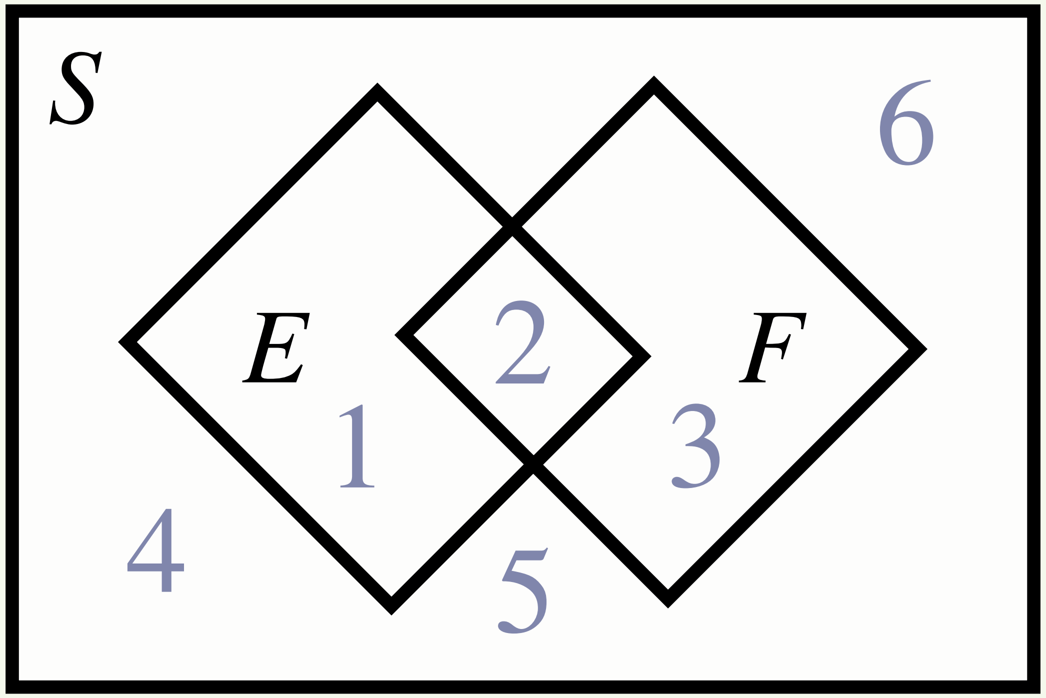
Hit: 11
Thrown: 24



$$P(E) \approx 0.458$$

Axioms of Probability

Quick review of sets



E and F are events in S.

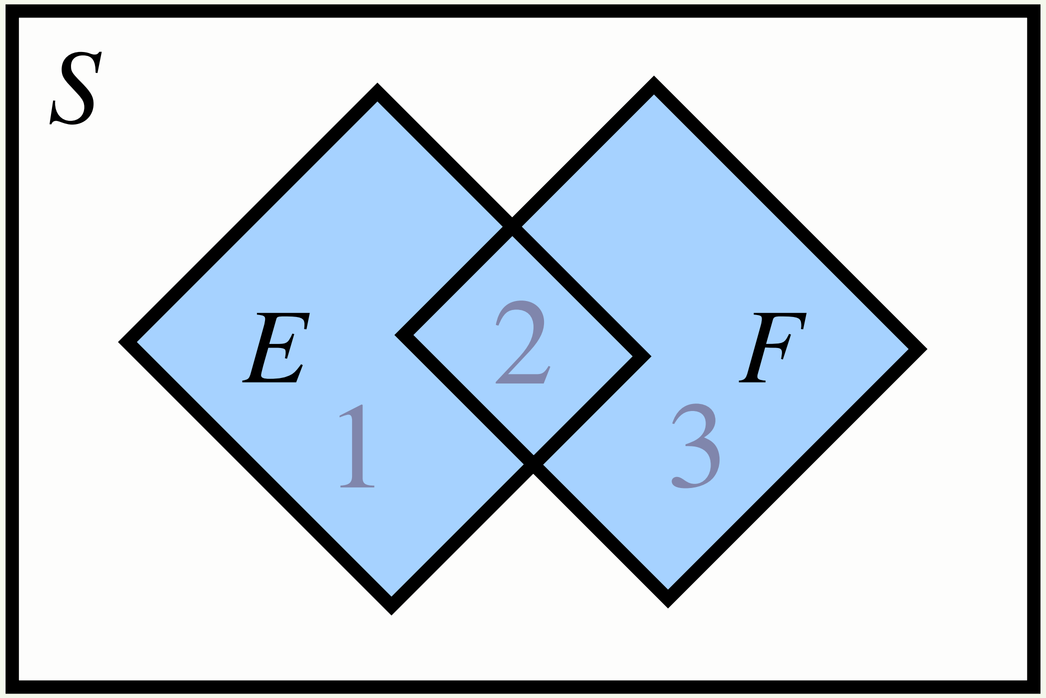
Experiment:

Die roll

S= {1,2,3,4,5,6}

Let E= {1,2}, and F= {2,3}

Quick review of sets



E and F are events in S.

Experiment:

Die roll

$S = \{1,2,3,4,5,6\}$

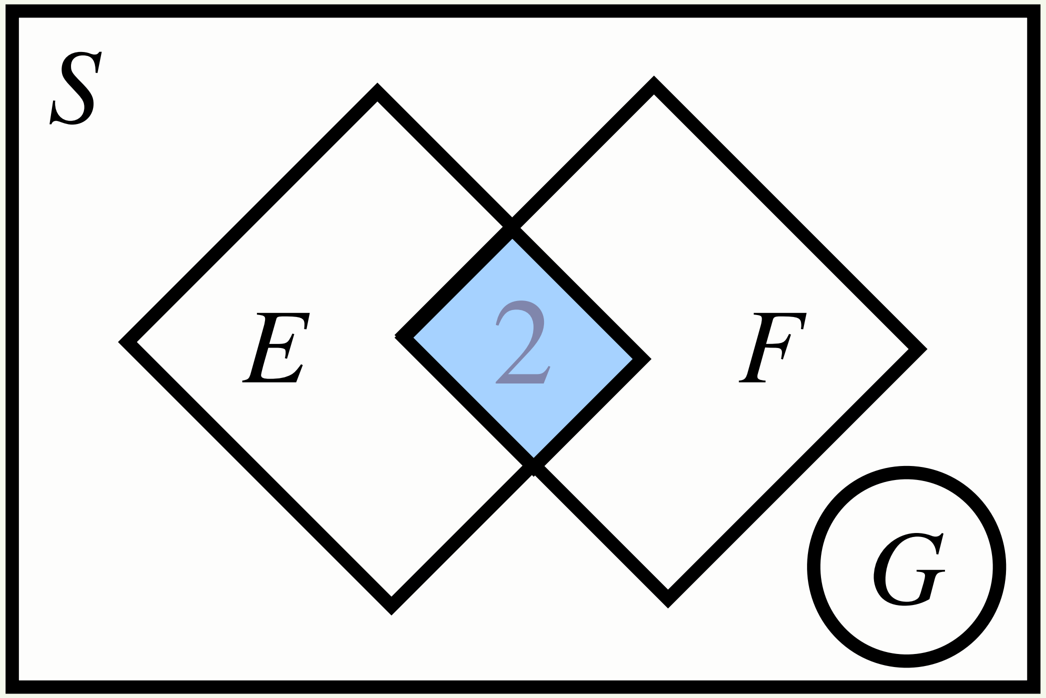
Let $E = \{1,2\}$, and $F = \{2,3\}$

def **Union** of events, $E \cup F$

The event containing all outcomes in E ***or*** F.

$E \cup F = \{1,2,3\}$

Quick review of sets



E and F are events in S.

Experiment:

Die roll

S = {1,2,3,4,5,6}

Let E = {1,2}, and F = {2,3}

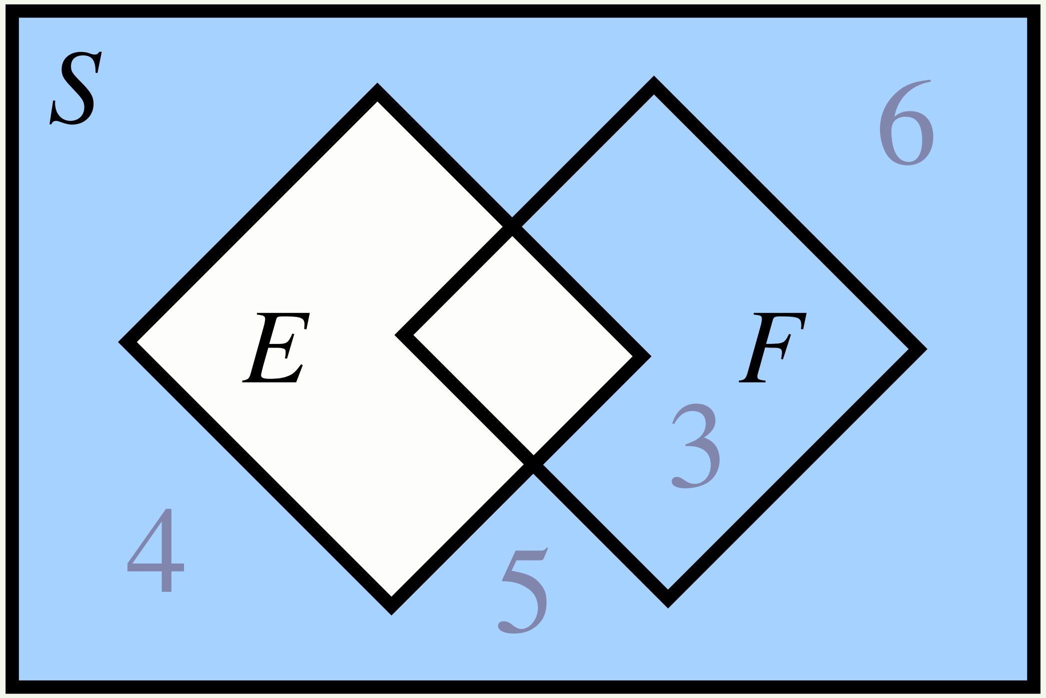
<u>def</u> **Intersection** of events, $E \cap F$

The event containing all outcomes in E ***<u>and</u>*** F.

$E \cap F = EF = \{2\}$

<u>def</u> **Mutually exclusive** events, F and G means that $F \cap G = \emptyset$

Quick review of sets



E and F are events in S.

Experiment:

Die roll

S= {1,2,3,4,5,6}

Let E= {1,2}, and F= {2,3}

<u>def</u> **Complement** of event E, E^C

The event containing all outcomes that are ***not*** in E.

E^C = {3,4,5,6}

3 axioms of probability

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

Axiom 1: $0 \le P(E) \le 1$

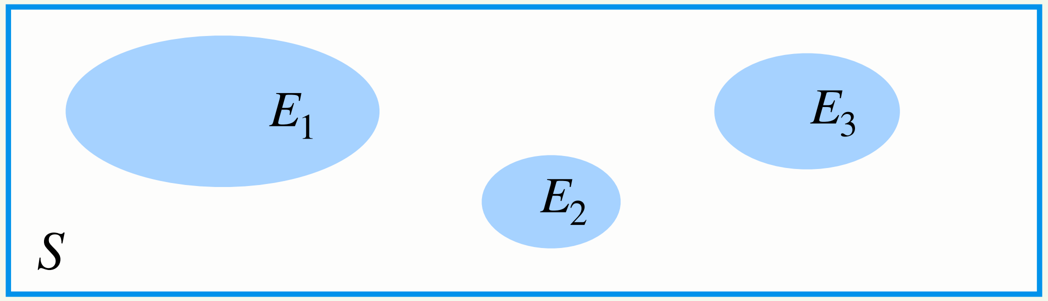
Axiom 2: $P(S) = 1$

Axiom 3: If E and F are mutually exclusive $E \cap F = \varnothing$, then $P(E \cup F) = P(E) + P(F)$

The analytically useful Axiom

More generally, for any sequence of mutually exclusive events $E_1, E_2, \ldots$:

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

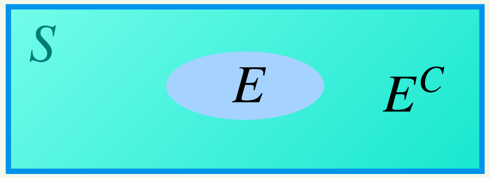


(like the Sum Rule of counting, but for probabilities)

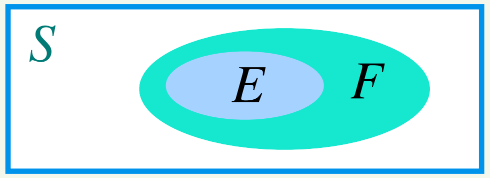
Corollaries of Probability

3 corollaries of axioms of probability

Corollary 1: $P(E^C) = 1 - P(E)$



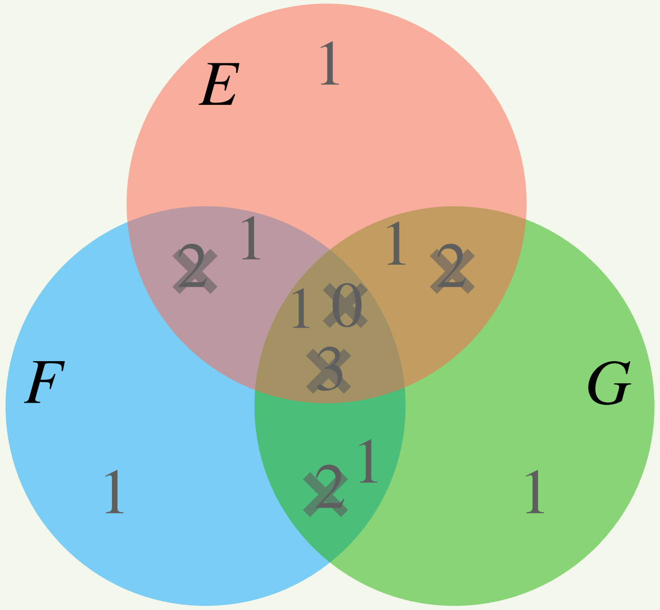
Corollary 2: If $E \subseteq F$, then $P(E) \leq P(F)$



Corollary 3: $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

(Inclusion-Exclusion Principle for Probability)

General form: $$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \ldots < i_r} P\left(\bigcap_{j=1}^{r} E_{i_j}\right)$$



$$P(E \cup F \cup G) =$$
$$P(E) + P(F) + P(G) \checkmark$$
$$-P(E \cap F) - P(E \cap G) - P(F \cap G) \checkmark$$
$$+P(E \cap F \cap G) \checkmark$$

Take an example: Selecting programmers

- P(student programs in Python) = 0.28
- P(student programs in C++) = 0.07
- P(student programs in Python and C++) = 0.05

What is P(student does not program in C++ or Python)?

1. **Define events & state goals**
2. **Identify known probabilities**
3. **Solve**

Given: E: python
F: C++

Find: $P((E \cup F)^C) = ?$

$$P(E) = 0.28$$
$$P(F) = 0.07$$
$$P(E \cap F) = 0.05$$

Corollary 1

$$P((E \cup F)^C) = 1 - P(E \cup F)$$

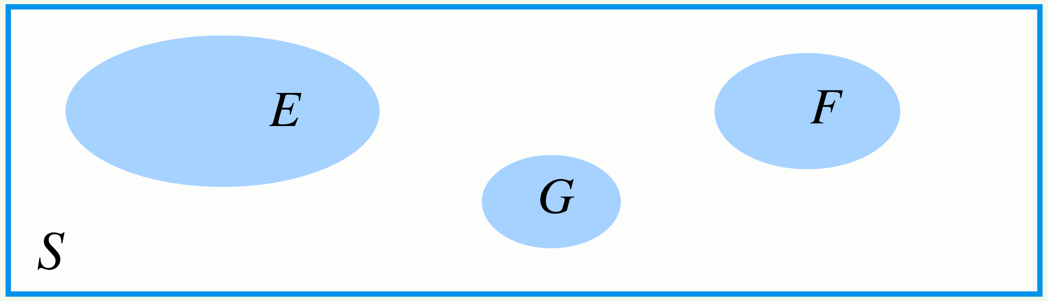
Corollary 3

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$
$$= 0.28 + 0.07 - 0.05$$
$$= 0.3$$

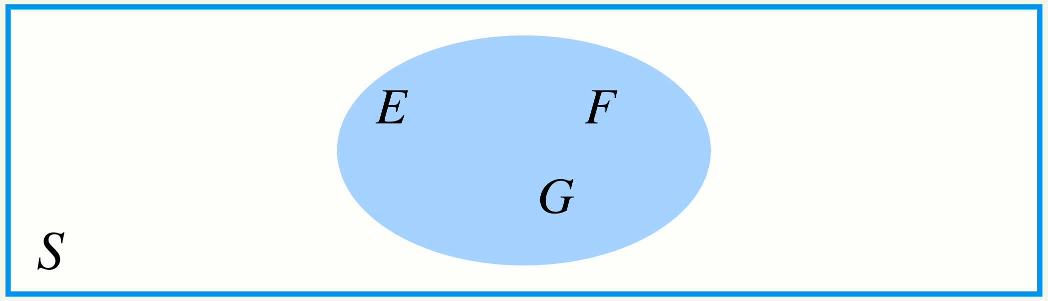
$$\Rightarrow 1 - 0.3 = 0.7$$

Takeaway: Union of events

Axiom 3,
Mutually exclusive events



Corollary 3,
Inclusion-exclusion principle



The challenge of probability is in defining events.

Some event probabilities are easier to compute than others.

Conditional Probability

Conditional probability

The conditional probability of E given F is the probability that E occurs given that F has already occurred. This is known as ***conditioning*** on F.

Written as:	$P(E \mid F)$
Means:	"$P(E$, given F already observed)"
Sample space →	all possible outcomes consistent with F (i.e. $S \cap F$)
Event →	all outcomes in E consistent with F (i.e. $E \cap F$)

General def of conditional probability:

$$P(E \mid F) = \frac{P(EF)}{P(F)}$$

The Chain Rule (aka Product Rule):

$$P(EF) = P(F)\ P(E \mid F)$$

Generalized: $P(E_1 E_2 \ldots E_n) = P(E_1)P(E_2 \mid E_1) \ldots P(E_n \mid E_1 E_2 \ldots E_{n-1})$

Law of Total Probability

$$P(EF)$$

Chain rule
(Product rule)

Definition of
conditional probability

$$P(E \mid F)$$

Law of Total Probability

$$P(E)$$

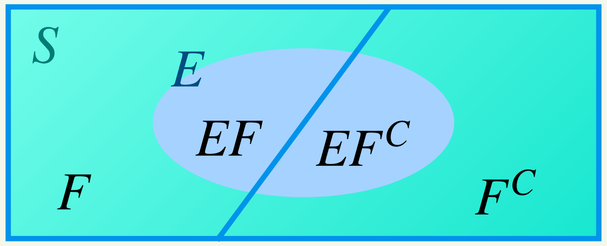
Law of total probability

Thm Let F be an event where $P(F) > 0$. For any event E,

$$P(E) = P(E \mid F)P(F) + P(E \mid F^C)P(F^C)$$

Proof

1. F and F^C are disjoint s.t. $F \cup F^C = S$ — *Def. of complement*



2. $E = (EF) \cup (EF^C)$

3. $P(E) = P(EF) + P(EF^C)$ — *Axiom 3*

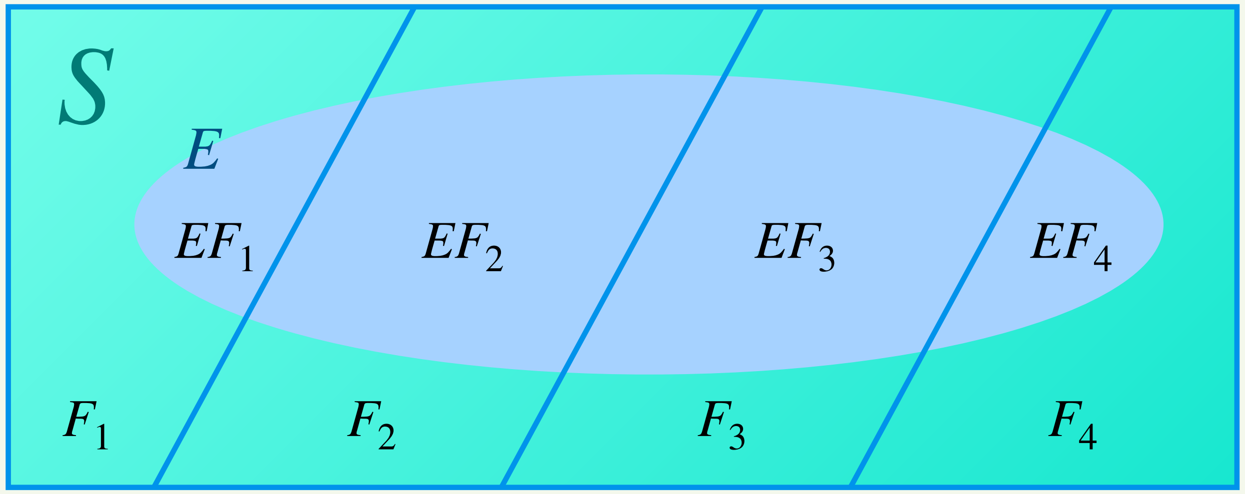
4. $P(E) = P(E \mid F)P(F) + P(E \mid F^C)P(F^C)$ — *Chain rule*

General law of total probability

Thm For mutually exclusive events $F_1, F_2, \ldots, F_n$ s.t. $F_1 \cup F_2 \cup \ldots \cup F_n = S$,

$$P(E) = \sum_{i=1}^{n} P(E \mid F_i) P(F_i)$$

Proof



Finding $P(E)$ from $P(E \mid F)$

Law of total probability

$$P(E) = P(E \mid F)P(F) + P(E \mid F^C)P(F^C)$$

- Flip a fair coin.
- If heads: roll a fair 6-sided die.
- Else: roll a fair 3-sided die.

You win if you roll a 6. What is P(winning)?

1. Define events & state goals

2. Identify <u>known</u> probabilities

3. Solve

Let: E: win
F: flip heads

Find: $P(\text{win}) = P(E)$

$$P(\text{win} \mid H) = P(E \mid F) = \frac{1}{6}$$

$$P(H) = P(F) = \frac{1}{2}$$

$$P(\text{win} \mid T) = P(E \mid F^C) = 0$$

$$P(T) = P(F^C) = 1 - \frac{1}{2} = \frac{1}{2}$$

$$P(E) = \frac{1}{6} \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{12}$$

$$P(EF)$$

Chain rule
(Product rule)

Definition of
conditional probability

$$P(E \mid F)$$

Law of Total
Probability

**Bayes'
Theorem**

$$P(E)$$

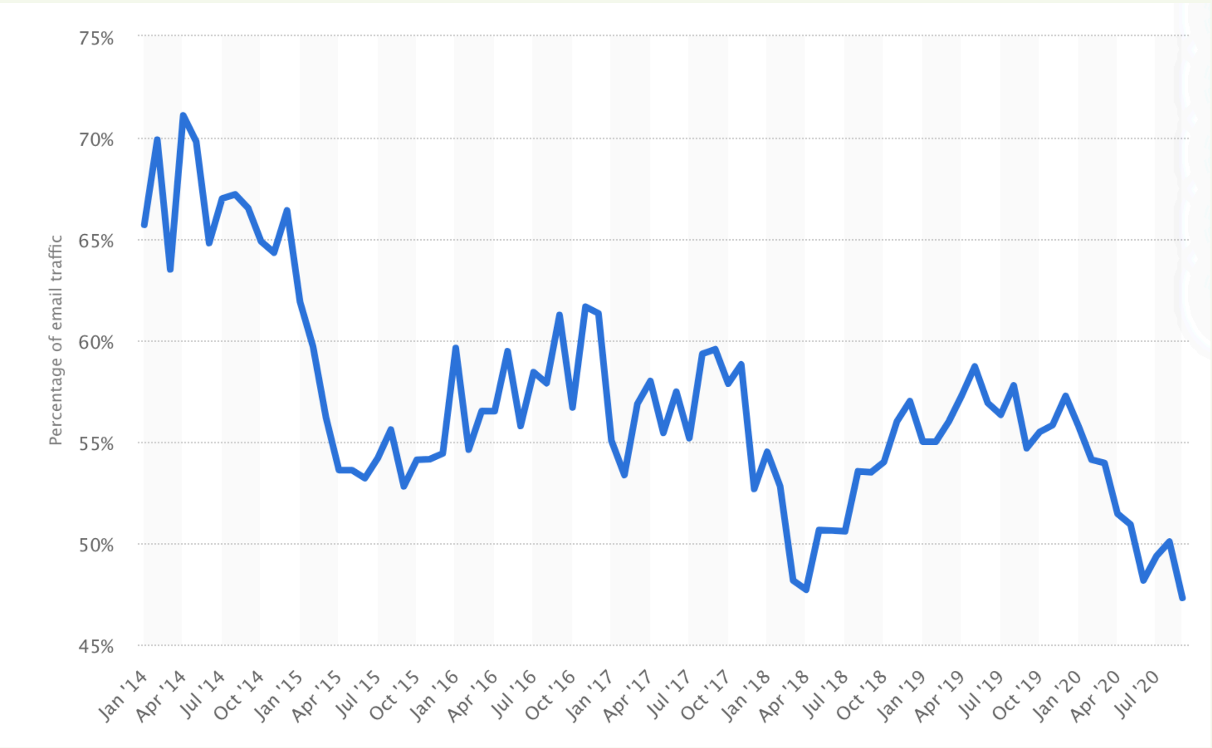
$$P(F \mid E)$$

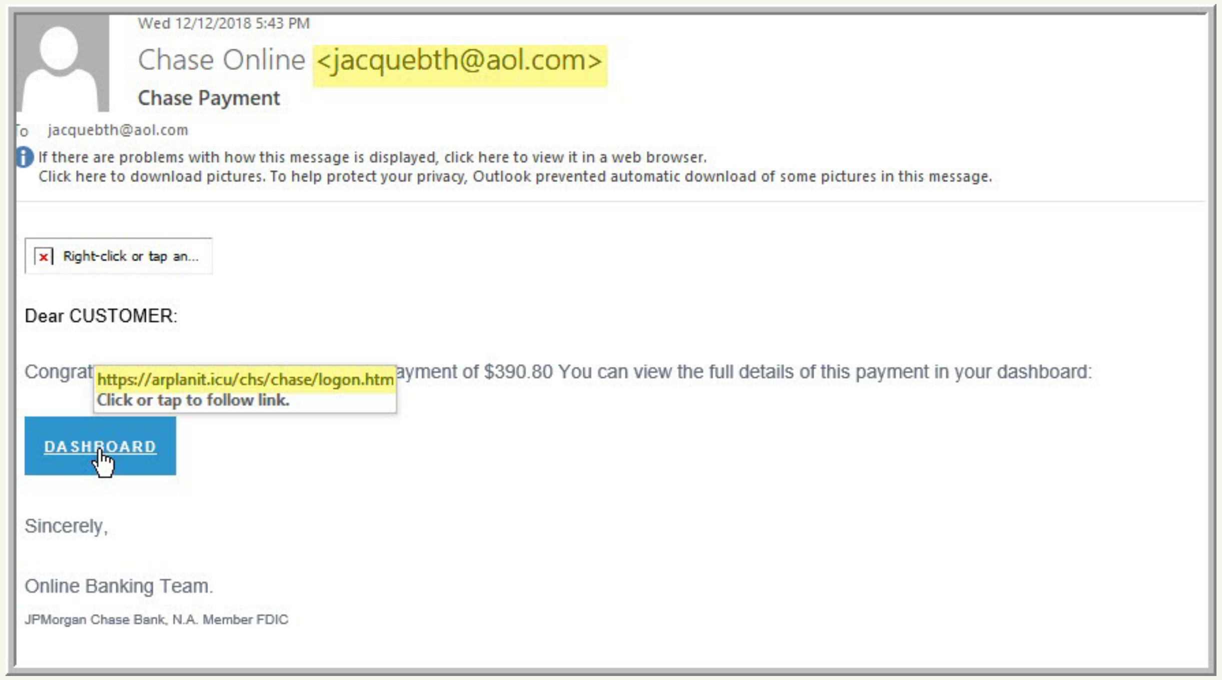
***Subjective probability
interpretation***

Take 5

Detecting spam email

Global spam volume as % of total email traffic





One can easily calculate how many spam emails contain "Dear":

$$P(E\,|\,F) = P(\text{“Dear”}\,|\,\text{Spam email})$$

But what is the probability that an email containing "Dear" is spam?

$$P(F\,|\,E) = P(\text{Spam email}\,|\,\text{“Dear”})$$

Bayes' Theorem

$P(E\,|\,F) \;\Rightarrow\; P(F\,|\,E)$

Thm For any events E and F where $P(E) > 0$ and $P(F) > 0$,

$$P(F\,|\,E) = \frac{P(E\,|\,F)\;P(F)}{P(E)}$$

Proof

1) $P(F\,|\,E) = \dfrac{P(EF)}{P(E)}$ [def. of conditional probability]

2) $= \dfrac{P(E\,|\,F)P(F)}{P(E)}$ [chain rule]

Expanded form

$$P(F\,|\,E) = \frac{P(E\,|\,F)\;P(F)}{P(E\,|\,F)\;P(F)\;+\;P(E\,|\,F^C)\;P(F^C)}$$

Proof $P(E)$ [law of total probability]

Detecting spam email

$$P(F \mid E) = \frac{P(E \mid F)\ P(F)}{P(E \mid F)\ P(F)\ +\ P(E \mid F^C)\ P(F^C)}$$

- 60% of all email in 2016 is spam. $P(F) = 0.6$
- 20% of spam has the word "Dear." $P(E \mid F) = 0.2$
- 1% of non-spam has the word "Dear." $P(E \mid F^C) = 0.01$

You receive an email with the word "Dear" in it.

What is the probability that the email is spam?

1. **Define events & state goals**
2. **Identify known probabilities**
3. **Solve**

Let: E: "Dear"

F: spam

Find: $P(\text{spam} \mid \text{Dear}) = P(F \mid E)$

$$P(F \mid E) = \frac{0.2 \cdot 0.6}{0.2 \cdot 0.6 + 0.01 \cdot 0.4}$$

$$\approx 0.97$$

Bayes' Theorem terminology

- 60% of all email in 2016 is spam. $P(F)$ *prior*
- 20% of spam has the word "Dear." $P(E \mid F)$ *likelihood*
- 1% of non-spam has the word "Dear." $P(E \mid F^C)$

You receive an email with the word "Dear" in it.

What is the probability that the email is spam? $P(F \mid E)$ *posterior*

$$\underbrace{P(F \mid E)}_{\text{posterior}} = \frac{\overbrace{P(E \mid F)}^{\text{likelihood}} \; \overbrace{P(F)}^{\text{prior}}}{\underbrace{P(E)}_{\text{normalization constant}}}$$

Why is Bayes’ so important?

It links belief to evidence in probability

$$P(E = \text{Evidence} \,|\, F = \text{Fact})$$

(collected from data)

Bayes’

$$P(F = \text{Fact} \,|\, E = \text{Evidence})$$

(categorize a new data point)

Given new evidence E, update belief of fact F

Prior belief $\rightarrow$ Posterior belief

$P(F) \rightarrow P(F \,|\, E)$

A topical example

You carry the disease (yes/no)

You tested positive

+

You tested negative

−

If a test returns positive, what is the likelihood that you have the disease?

Interpreting test results can be confusing

Fact, F	Has disease	*take test* →	Evidence, E	Test positive
or F^C	No disease		or E^C	Test negative

Take a moment to fill in the confusion matrix

		Fact	
		F, disease +	F^C, disease -
Evidence	E, test +	True positive $P(E\,\vert\,F)$	False positive $P(E\,\vert\,F^C)$
	E^C, test -	False negative $P(E^C\,\vert\,F)$	True negative $P(E^C\,\vert\,F^C)$

If a test returns positive, what is the likelihood that you have the disease?

Covid testing

$$P(F\,|\,E) = \frac{P(E\,|\,F)\;P(F)}{P(E\,|\,F)\;P(F)\;+\;P(E\,|\,F^C)\;P(F^C)}$$

- A test is 98% effective at detecting Covid ("true positive"). $P(E\,|\,F)$
- The test has a "false positive" rate of 1%. $P(E\,|\,F^C)$
- 0.5% of the population has Covid. $P(F)$

What is the likelihood you have Covid if you test positive?

1. **Define events & state goals**
2. **Identify known probabilities**
3. **Solve**

Let: E: you test positive
F: you actually have covid

Find: $P(\text{covid}\,|\,\text{test }+) = P(F\,|\,E)$

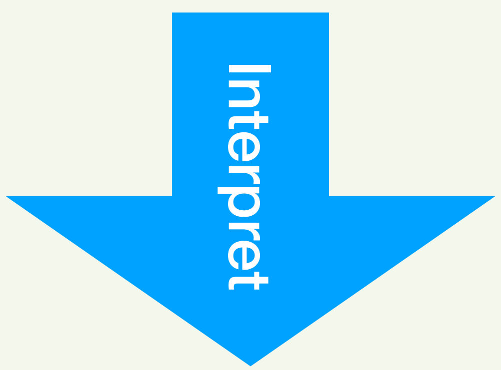
$$P(F\,|\,E) = \frac{0.98 \cdot 0.005}{0.98 \cdot 0.005 + 0.01 \cdot 0.995}$$

$$\approx 0.33$$

Bayes' Theorem intuition

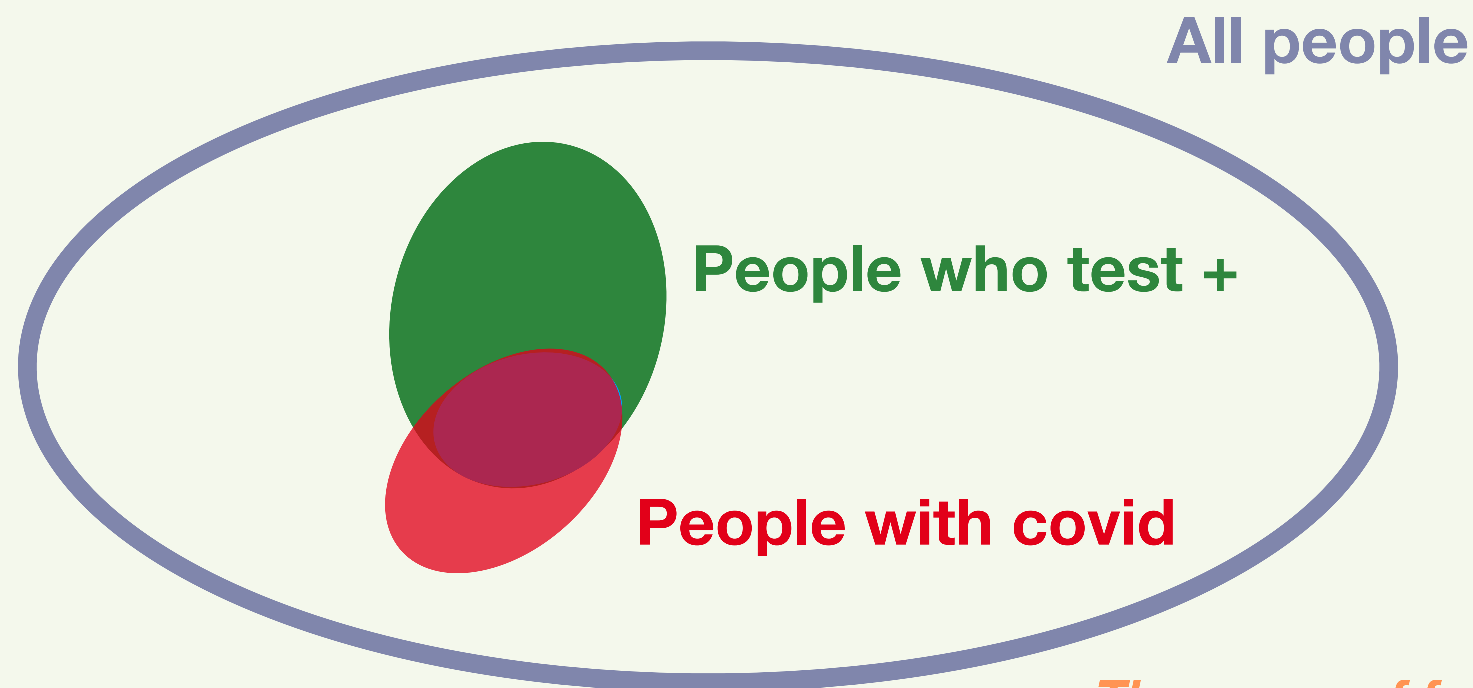
<u>Original question</u>:

What is the likelihood that you have Covid if you test positive?



<u>Interpretation</u>:

Of the people who test positive, how many actually have Covid?



The space of facts

The space of facts conditioned on a positive test result

Why you should still get tested

- A test is 98% effective at detecting Covid (“true positive”).
- The test has a “false positive” rate of 1%.
- 0.5% of the population has Covid.

E = you test positive.
F = you actually have the disease.

E^C = you test negative.

What is $P(F \mid E^C)$?

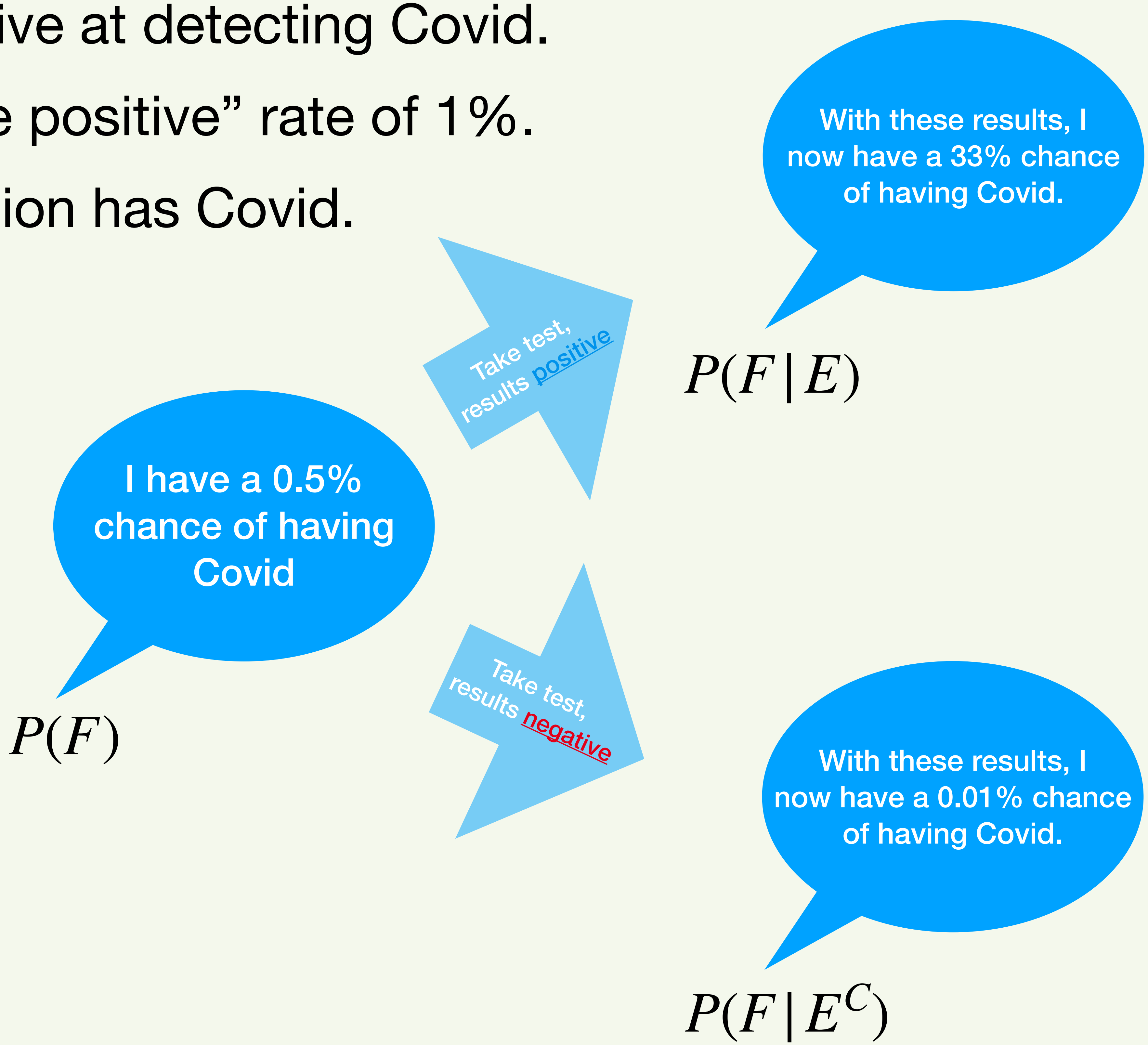
		Fact	
		F, disease +	F^C, disease -
Evidence	E, test +	True positive $P(E \mid F)$	False positive $P(E \mid F^C)$
	E^C, test -	False negative $P(E^C \mid F)$	True negative $P(E^C \mid F^C)$

$$P(F \mid E^C) = \frac{P(E^C \mid F)\ P(F)}{P(E^C \mid F)\ P(F)\ +\ P(E^C \mid F^C)\ P(F^C)} = 0.0001$$

Putting it all together

- A test is 98% effective at detecting Covid.
- The test has a "false positive" rate of 1%.
- 0.5% of the population has Covid.

E = you test positive.
F = you actually have the disease.



Generalized version of Bayes' Theorem

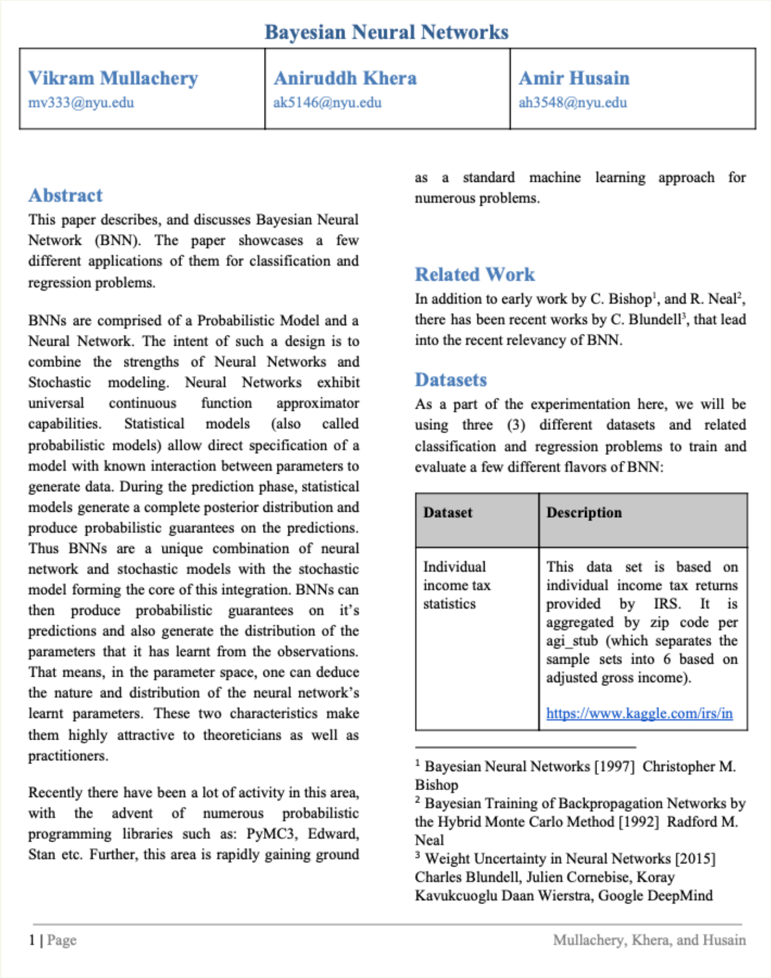
- Today we derived Bayes' Theorem for 2 subsets in a sample space.
- Can be generalized to multiple disjoint subsets.
- For HW reading, go through the **Bayes' handout** and memorize the proof.

You never know when you may be asked…

Why Bayes' *again*?

Preview

- Well, *it's fundamental*, and very important!
- Pops up everywhere
- Here are some nice examples (ILIAS: Reading material /L01):
  - Bayesian Neural Networks (Covered in Lecture 14)
  - NeuroBayes algorithm (developed at ETP)



Bayesian Neural Networks

Vikram Mullachery mv333@nyu.edu	Aniruddh Khera ak5146@nyu.edu	Amir Husain ah3548@nyu.edu

Abstract

This paper describes, and discusses Bayesian Neural Network (BNN). The paper showcases a few different applications of them for classification and regression problems.

BNNs are comprised of a Probabilistic Model and a Neural Network. The intent of such a design is to combine the strengths of Neural Networks and Stochastic modeling. Neural Networks exhibit universal continuous function approximator capabilities. Statistical models (also called probabilistic models) allow direct specification of a model with known interaction between parameters to generate data. During the prediction phase, statistical models generate a complete posterior distribution and produce probabilistic guarantees on the predictions. Thus BNNs are a unique combination of neural network and stochastic models with the stochastic model forming the core of this integration. BNNs can then produce probabilistic guarantees on it's predictions and also generate the distribution of the parameters that it has learnt from the observations. That means, in the parameter space, one can deduce the nature and distribution of the neural network's learnt parameters. These two characteristics make them highly attractive to theoreticians as well as practitioners.

Recently there have been a lot of activity in this area, with the advent of numerous probabilistic programming libraries such as: PyMC3, Edward, Stan etc. Further, this area is rapidly gaining ground as a standard machine learning approach for numerous problems.

Related Work

In addition to early work by C. Bishop[1], and R. Neal[2], there has been recent works by C. Blundell[3], that lead into the recent relevancy of BNN.

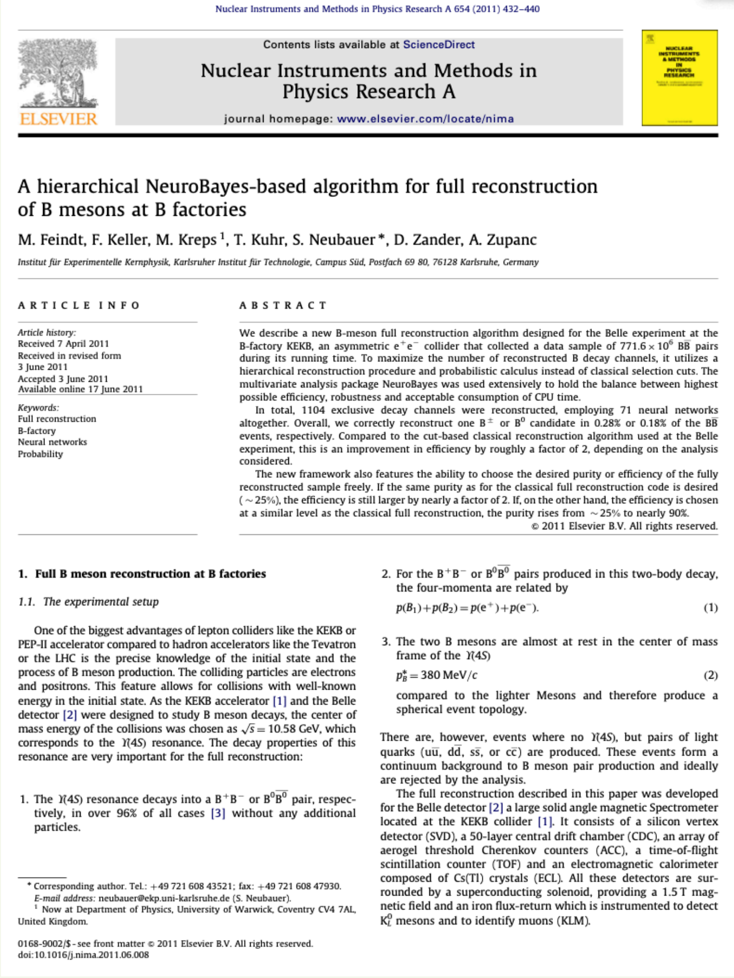
Datasets

As a part of the experimentation here, we will be using three (3) different datasets and related classification and regression problems to train and evaluate a few different flavors of BNN:

Dataset	Description
Individual income tax statistics	This data set is based on individual income tax returns provided by IRS. It is aggregated by zip code per agi_stub (which separates the sample sets into 6 based on adjusted gross income). https://www.kaggle.com/irs/in

[1] Bayesian Neural Networks [1997] Christopher M. Bishop
[2] Bayesian Training of Backpropagation Networks by the Hybrid Monte Carlo Method [1992] Radford M. Neal
[3] Weight Uncertainty in Neural Networks [2015] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu Daan Wierstra, Google DeepMind

1 | Page Mullachery, Khera, and Husain



Nuclear Instruments and Methods in Physics Research A 654 (2011) 432–440

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

journal homepage: www.elsevier.com/locate/nima

ELSEVIER

A hierarchical NeuroBayes-based algorithm for full reconstruction of B mesons at B factories

M. Feindt, F. Keller, M. Kreps [1], T. Kuhr, S. Neubauer *, D. Zander, A. Zupanc

Institut für Experimentelle Kernphysik, Karlsruher Institut für Technologie, Campus Süd, Postfach 69 80, 76128 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 7 April 2011
Received in revised form 3 June 2011
Accepted 3 June 2011
Available online 17 June 2011

Keywords:
Full reconstruction
B-factory
Neural networks
Probability

ABSTRACT

We describe a new B-meson full reconstruction algorithm designed for the Belle experiment at the B-factory KEKB, an asymmetric e^+e^- collider that collected a data sample of 771.6×10^6 $B\overline{B}$ pairs during its running time. To maximize the number of reconstructed B decay channels, it utilizes a hierarchical reconstruction procedure and probabilistic calculus instead of classical selection cuts. The multivariate analysis package NeuroBayes was used extensively to hold the balance between highest possible efficiency, robustness and acceptable consumption of CPU time.

In total, 1104 exclusive decay channels were reconstructed, employing 71 neural networks altogether. Overall, we correctly reconstruct one $B^\pm$ or B^0 candidate in 0.28% or 0.18% of the $B\overline{B}$ events, respectively. Compared to the cut-based classical reconstruction algorithm used at the Belle experiment, this is an improvement in efficiency by roughly a factor of 2, depending on the analysis considered.

The new framework also features the ability to choose the desired purity or efficiency of the fully reconstructed sample freely. If the same purity as for the classical full reconstruction code is desired ($\sim$25%), the efficiency is still larger by nearly a factor of 2. If, on the other hand, the efficiency is chosen at a similar level as the classical full reconstruction, the purity rises from $\sim$25% to nearly 90%.

© 2011 Elsevier B.V. All rights reserved.

1. Full B meson reconstruction at B factories

1.1. The experimental setup

One of the biggest advantages of lepton colliders like the KEKB or PEP-II accelerator compared to hadron accelerators like the Tevatron or the LHC is the precise knowledge of the initial state and the process of B meson production. The colliding particles are electrons and positrons. This feature allows for collisions with well-known energy in the initial state. As the KEKB accelerator [1] and the Belle detector [2] were designed to study B meson decays, the center of mass energy of the collisions was chosen as $\sqrt{s}=10.58$ GeV, which corresponds to the $\Upsilon(4S)$ resonance. The decay properties of this resonance are very important for the full reconstruction:

1. The $\Upsilon(4S)$ resonance decays into a B^+B^- or $B^0\overline{B^0}$ pair, respectively, in over 96% of all cases [3] without any additional particles.
2. For the B^+B^- or $B^0\overline{B^0}$ pairs produced in this two-body decay, the four-momenta are related by
   $$p(B_1)+p(B_2)=p(e^+)+p(e^-). \quad (1)$$
3. The two B mesons are almost at rest in the center of mass frame of the $\Upsilon(4S)$
   $$p_B^* = 380\ \mathrm{MeV}/c \quad (2)$$
   compared to the lighter Mesons and therefore produce a spherical event topology.

There are, however, events where no $\Upsilon(4S)$, but pairs of light quarks ($u\bar{u}$, $d\bar{d}$, $s\bar{s}$, or $c\bar{c}$) are produced. These events form a continuum background to B meson pair production and ideally are rejected by the analysis.

The full reconstruction described in this paper was developed for the Belle detector [2] a large solid angle magnetic Spectrometer located at the KEKB collider [1]. It consists of a silicon vertex detector (SVD), a 50-layer central drift chamber (CDC), an array of aerogel threshold Cherenkov counters (ACC), a time-of-flight scintillation counter (TOF) and an electromagnetic calorimeter composed of Cs(Tl) crystals (ECL). All these detectors are surrounded by a superconducting solenoid, providing a 1.5 T magnetic field and an iron flux-return which is instrumented to detect K_L^0 mesons and to identify muons (KLM).

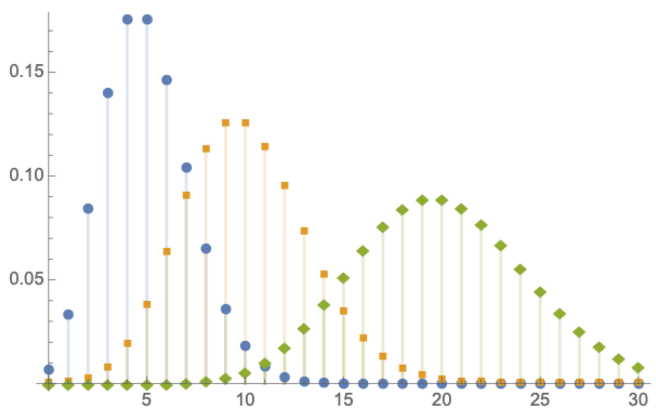
* Corresponding author. Tel.: +49 721 608 43521; fax: +49 721 608 47930.
E-mail address: neubauer@ekp.uni-karlsruhe.de (S. Neubauer).
[1] Now at Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom.

0168-9002/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2011.06.008

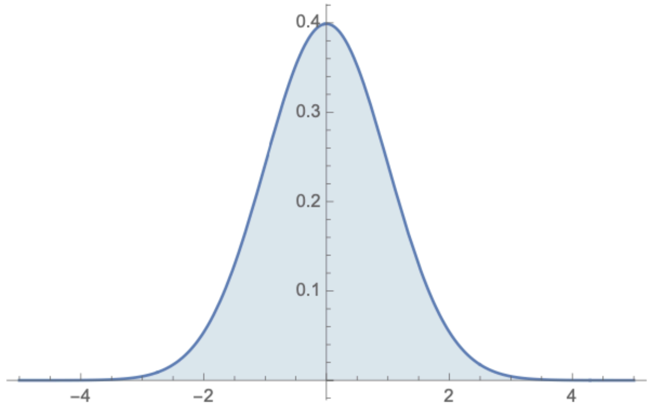
Quiz Time: 2nd Round

We will discuss the solutions next time

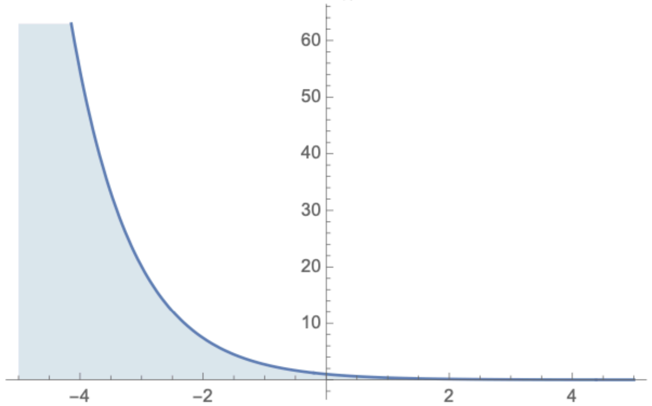
1. Which of these plots show probability density functions (or short PDFs)? The sampling space for (a) is $k \in \mathbb{N}_0$, for (b,c,d) $x \in [-\infty, \infty]$ and for (e) $x \in [0, 10]$.



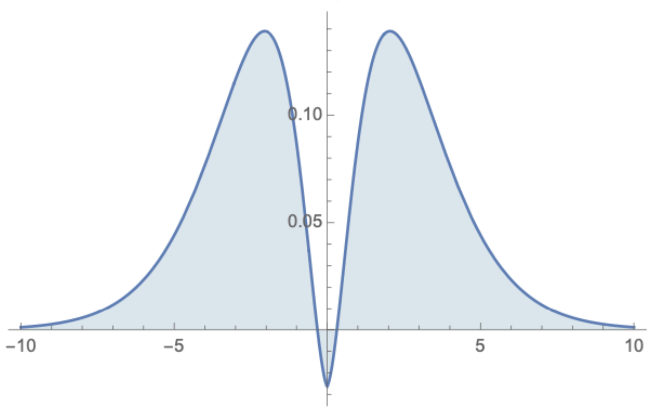
(a) $\frac{e^{-\lambda}\lambda^k}{k!}$



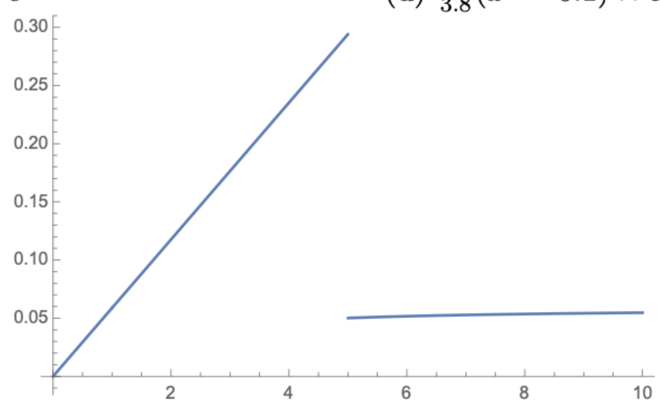
(b) $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$



(c) e^{-x}



(d) $\frac{1}{3.8}(x^2 - 0.1) \times \exp(-|x|)$ with $x \in \mathbb{R}$



(e) $\frac{1}{a}\left(\Theta(5-x)\left(\frac{1}{2} + \frac{\arctan(x-1)}{\pi}\right) + x\,\Theta(x-5)\right)$ with Θ denoting the step function and $a = 17.0033$.

2. Are mammographies useful?

A mammography is a x-ray-scan that tries to diagnose breast cancer by looking for abnormal tissue structures. Using Bayes' theorem assess the probability for a correct positive, $P(C|+)$, for patients at ages 30 and 70 with the following probabilities:

$$P(+|C) = 80\%\,, \qquad P(+|\bar{C}) = 7\%$$

denoting the probability of a positive diagnose in case of cancer (C) or not cancer $(\bar{C})$. The incidence of cancer, $P(C)$, in the general population is about 1.5% at the age of 30 and 4% at the age of 70.

For next time

- Required reading
  - Cowan textbook: chapter 1
  - Bayes’ handout: /Reading material / L01 / BayesThmForDisjointSets

- Extra reading for fun: /Reading material / L01 /
  - SkinCancerClassificationWithDNNs
  - TheRoleOfStatsHiggsDiscovery_DDyk
  - BayesianNeuralNetworks
  - NeuroBayes

Next time:

- Fundamental concepts II:
  - (Conditional) independence
  - Random variables
  - Expectation values, variance
  - Tour of important probability mass (density) functions

Bibliography

- Part of the material presented in this lecture is adapted from the following sources. See the active links (when available) for a complete reference
  - **2019 KIT Data Analysis** (offline) by Prof. Dr. F. Bernlochner (Bonn): slides 9, 24-30
  - **Probability for CS** (Stanford): slides 11-16, 33-66