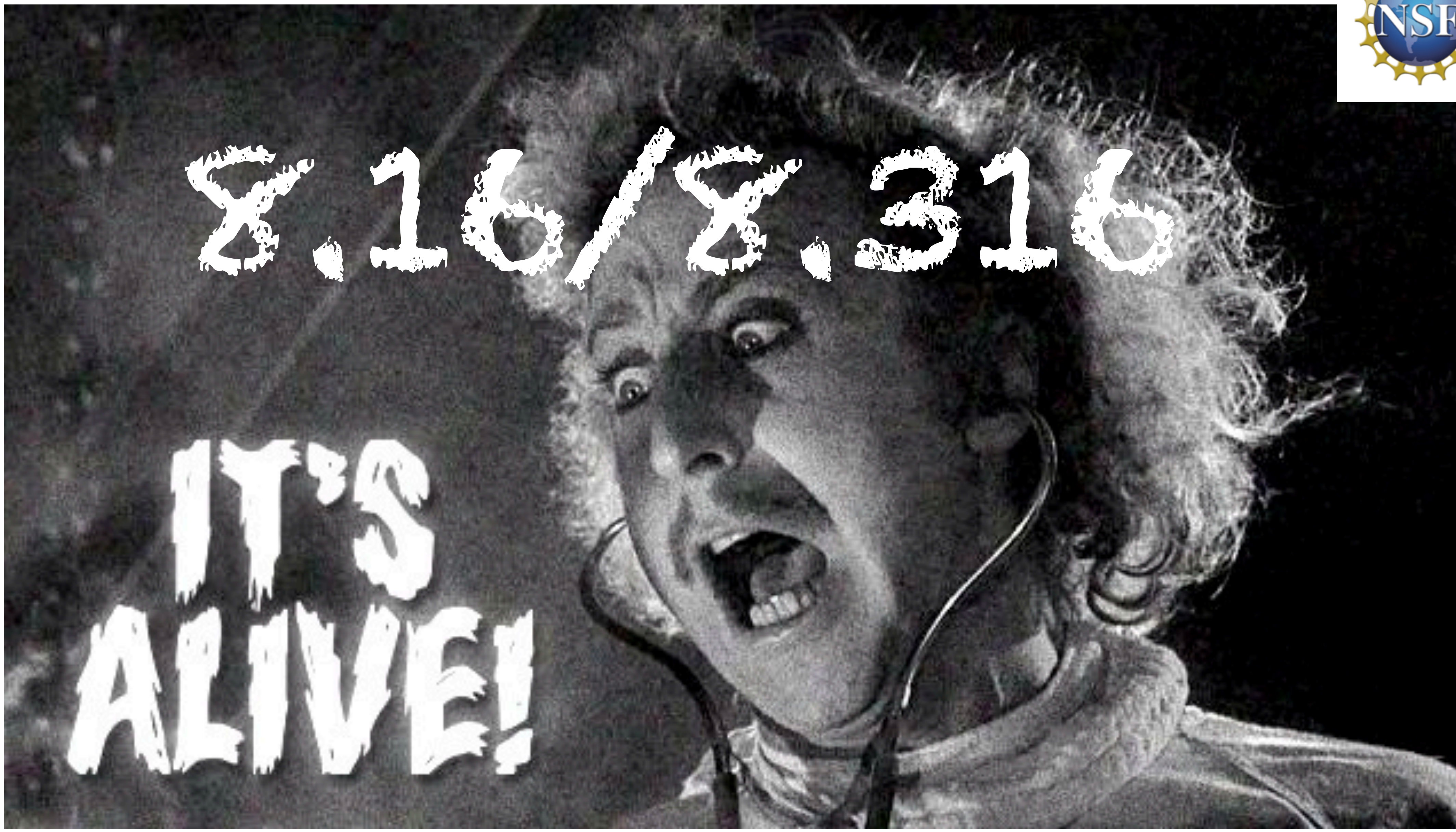


Computational Data Science in Physics

Philip Harris,
Alex Shvonski, Divij Sharma
Matthew Heine

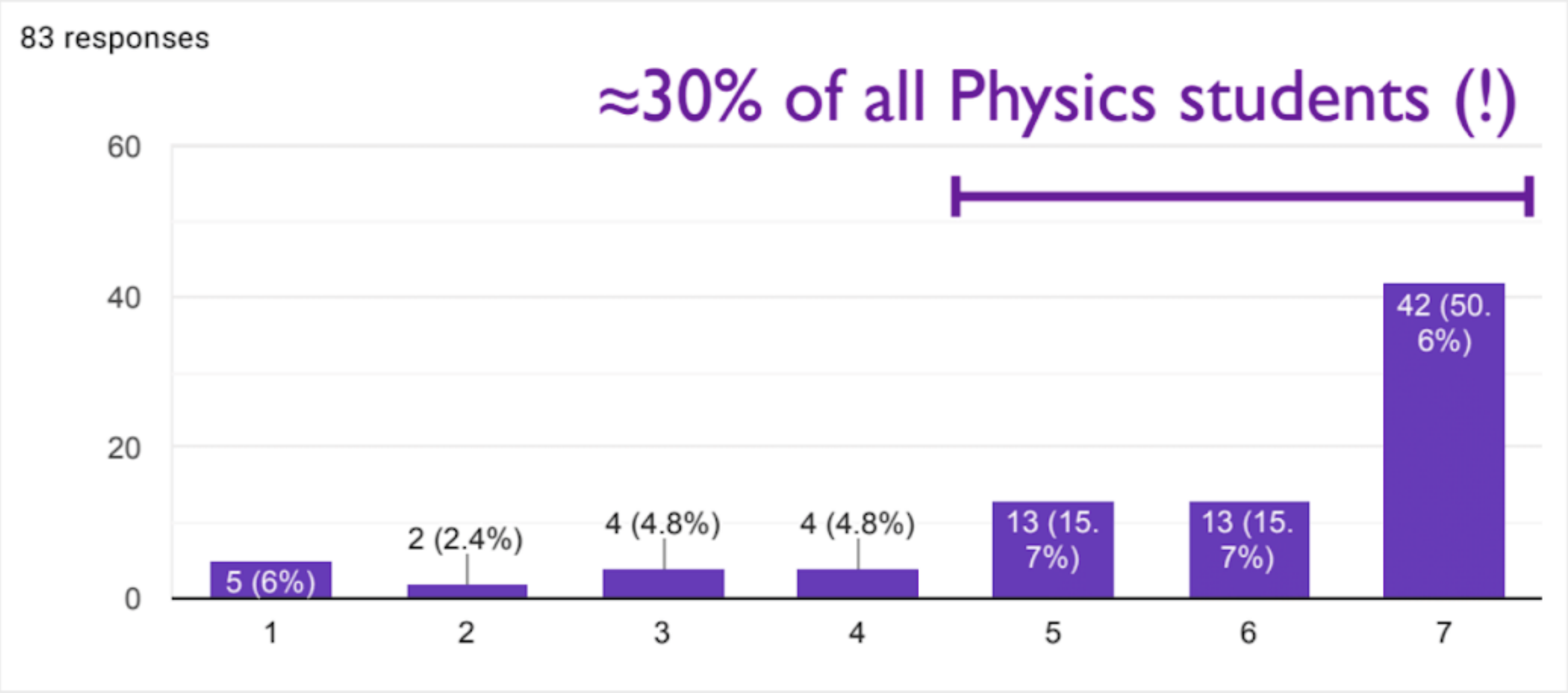
Accelerated AI Algorithms for Data-Driven Discovery

Welcome to this class

- Welcome to 8.16/8.316
    - This is our 4th of class
    - You being here is the result of a 6 year journey
    - The material is rapidly changing, we will stay up to date
- Please bear with me & you about this class
    - This class is run a little differently that others
    - People have very much enjoyed the last few years of it!

After a Long Journey

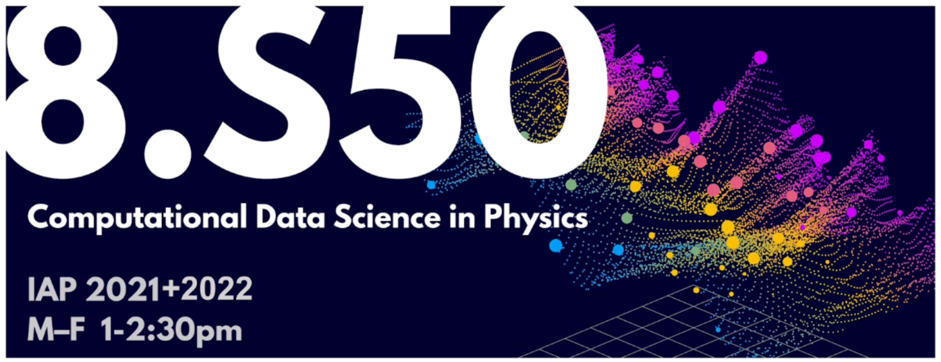
How interested would you be in submitting and defending a PhD thesis that uses statistical methods in a substantial way?



Respondent #11: "I think ML is the most important thing happening in the world right now and should be incorporated into any STEM degree."

- This class is finally coming to fruition
    - Its up to all of us to make sure this is a success

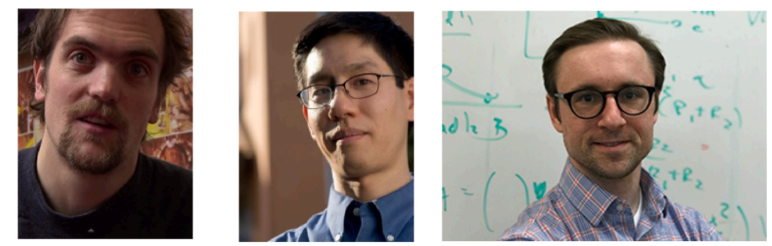
Story of this Class



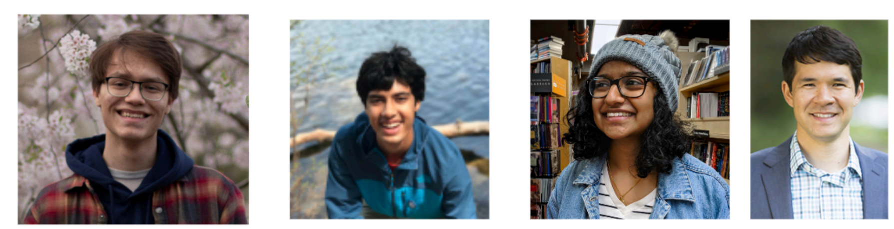
- When COVID hit in 2020, Junior Lab curriculum had to be redesigned to be remote
  - Out of this came several **labs analyzing open data**
- MIT Junior lab has now embraced open data analysis
  - Class has emphasized **python/jupyter ecosystem**
- The reception from the remote projects was very good
  - Tools are very **relevant to current research**
  - A few students managed to do some deep learning
- As a result, Phil Harris put together a class (**8.S50**) that ran during IAP in January 2021 and January 2022
  - Goal to cover **statistical tools up to deep learning**
  - Class targeted fitting/data analysis/deep learning of current data results

MITx MITx 8.S50x
Computational Data Science in Physics

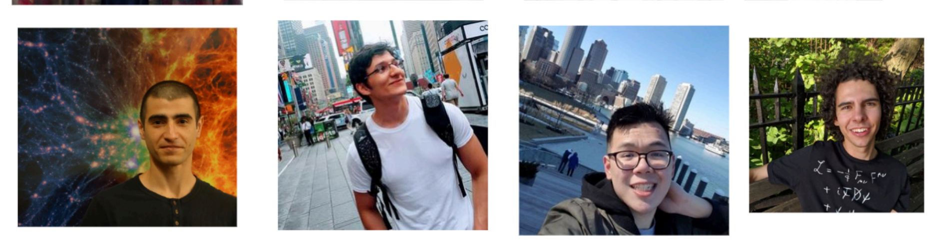
In 2021, **MITx** granted $72k for development of 8.S50 into an online course; will launch in September 2022



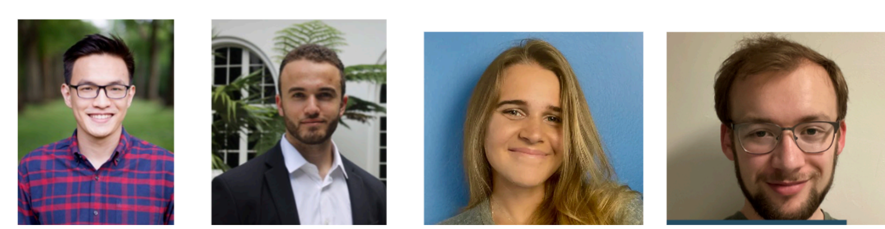
Prof. Phil Harris
Prof. Ike Chuang
Dr. Alex Shvonski



Michael Szurek
Karna Morey
Anjali Nambrath
Prof. Kiyoshi Masui



Juan Parra
Ouail Kitouni
Tri Nguyen
Eric Moreno



Duc Hoang
Aiden Chambers
Brianna Ryan
Jack Dinsmore

Rationale

- Physics Department → Statistics and data science topics minimally covered in other physics classes
- Undergrad physics flex majors → Want to do a focus group in data science in physics
- Undergraduates in physics → Want to learn about statistical methods beyond JLab
- Graduate Students in PhysSDS → Looking for course contributing to interdisciplinary PhD
- MITx → Will need long-term support for running MITx course
- Institute for Artificial Intelligence and Fundamental Interactions → Developing online Micromasters course with SDSC
- Institute for A3D3 → Connect Data Science programs across Universities

To get this knowledge outside of Physics, students would need probability, statistics, ML, and data analysis/algorithms courses (48 units before the science)

Community Values

PHYSICS COMMUNITY VALUES

- We are a strong supporter of the community values
- I want you to succeed just as much as I hope you want me to succeed
- https://physvals.mit.edu/

What is this class?

- This class we are going to introduce you to
  - Data Science and Physics in a practical way
  - Cover the skills you need to analyze data and simulate physics
    - General statistical tools and deep learning for phsyics
- We are not going to go into ultra-statistical detail
  - We will give you the tools and concepts to go deeper
  - But we are going to go more than in classes like J-lab
  - We will provide code based solutions for **everything**

Pre-requisite

- Main Prerequisite : some knowledge of Python
  - 6.0001/6.0002 satisfy this requirement
- Others: We have put 8.04 as well
  - Mostly to say that a good physics foundation is needed
  - Some derivations in this class will require:
    - Special Relativity, Newtonian mechanics
    - Understanding of Fourier transforms
    - Variance Expectation and probability distributions
    - Some quantum mechanics, special relativity, cosmology needed

Cover, but knowing this will help

What you learn

Course Description
Aims to present modern computational methods by providing realistic, contemporary examples of how these computational methods apply to physics research. Designed around research modules in which each module provides experience with a specific scientific challenge. Modules include: analyzing LIGO open data; measuring electroweak boson to quark decays; understanding the cosmic microwave background; and lattice QCD/Ising model. Experience in Python helpful but not required. Lectures are viewed outside of class; *in-class time is dedicated to problem-solving* and discussion. Students taking graduate version complete additional assignments.

- We are going to do lectures in class
  - They are going to be very much active lectures with problems
- We will record them
  - & We will make external recordings available throughout the class

What you learn

Course Description
Aims to present modern computational methods by providing realistic, contemporary examples of how these computational methods apply to physics research. Designed around research modules in which each module provides experience with a specific scientific challenge. Modules include: analyzing LIGO open data; measuring electroweak boson to quark decays; understanding the cosmic microwave background; and lattice QCD/Ising model. Experience in Python helpful but not required. Lectures are viewed outside of class; *in-class time is dedicated to problem-solving* and discussion. Students taking graduate version complete additional assignments.

- Graduate student “Additional” assignments:
  - Really, we just expect you to **go further** in projects
    - Its not extra, but more detail (see later)

What you learn

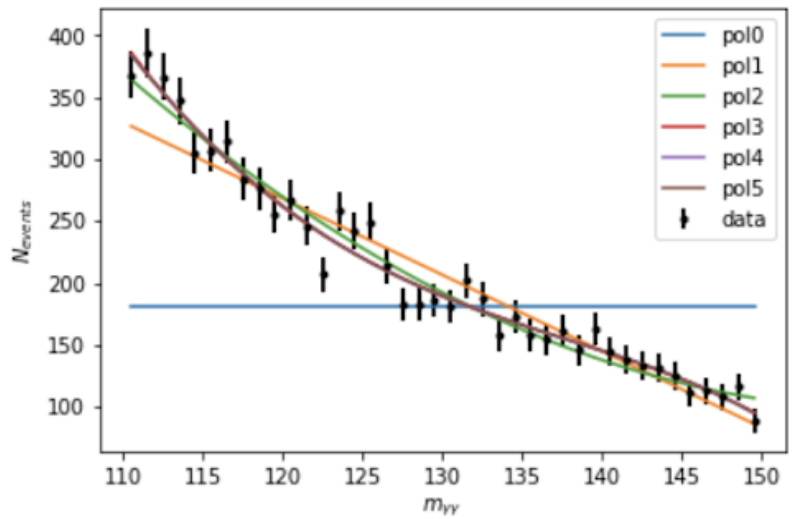
- Aim of this class :
  - Teach you how to use numerical tools to solve physics problems
  - To provide you with numerical tools at a research level
  - You will probably need this in your research
    - Theory & experiment and all domains in physics in this era
  - Expose you to real data from experiments
    - You can write real and **important** papers based on this
  - Expose you to “Data Science” thinking common in physics

Class Overview

- Class is going to be a project based class
- You will have 4 psets : 8% each for a project
  - They are associated to projects
  - They are available on canvas as automated psets
- You will have 4 projects: 15% of grade oneach
  - Projects are on Jupyter notebooks
  - You will have ~3 weeks for each project
- For the 4th project
  - You pick a previous project or do an astro project
    - <u>Goal is to dig deeper</u> => Creativity is key here
  - You will present your project in the last classes (an additional 8%)

Lectures

out2 Test




```
f 1 to 0: 1.1102230246251565e-16
f 2 to 1: 4.937240960511957e-08
f 3 to 2: 0.0035138407452814935
f 4 to 3: 0.8546368491590365
f 5 to 4: 0.9746177621262444
```

So from this looks like a 4th order polynomialgives an f-test above roughly 5% for both the category with the largest yield and the second largest yield. This seems reaonsable for us to use as our background function. Let's proceed with a signal function.

9.4 Fitting a Higgs Signal

Now, to fit a Higgs signal, what we want to do is a hypothesis test like we did above. Except now, we will cast our hypothesis, slightly differently to before.

Null Hypothesis The Higgs signal has a mass of $m_{\gamma\gamma}$ at a specific m_0, and a fixed width 1.2 GeV.

Alternative Hypothesis The Higgs signal is not there.

In [69]:

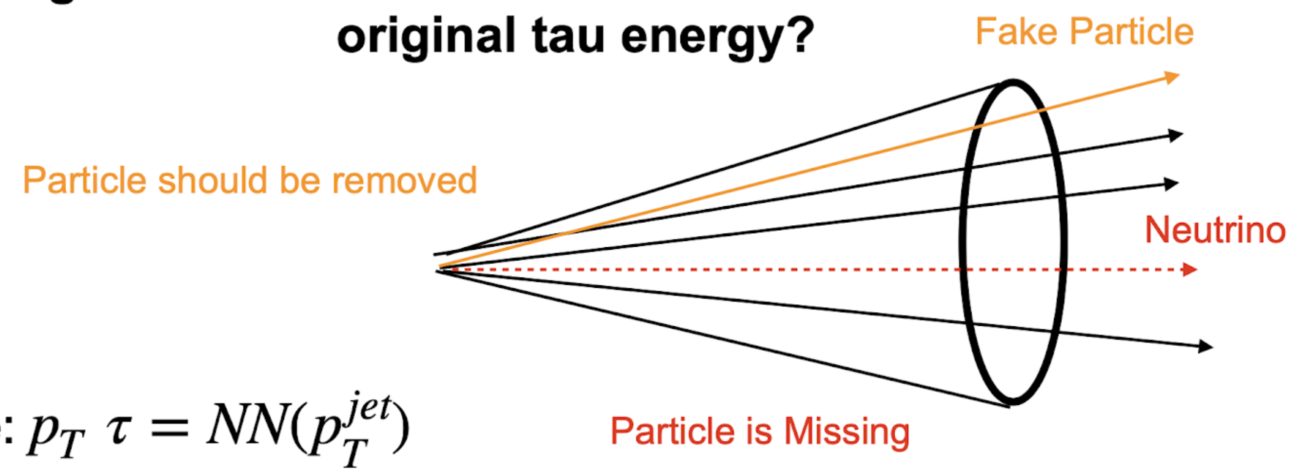
```
def sigpol4(x,p0,p1,p2,p3,p4,amp,mass,sigma):
    bkg=pol4(x,p0,p1,p2,p3,p4)
    sig=amp*stats.norm.pdf(x,mass,sigma)
    return sig+bkg

def fitModel(iX,iY,iWeights,iM,iFunc):
    model  = lmfit.Model(iFunc)
    p = model.make_params(p0=0,p1=0,p2=0,p3=0,p4=0,p5=0,amp=0,mass=iM,sigma=1.2)
    try:
        p["mass"].vary=False
```

We will go into the depths
Of fitting and
Hypothesis Testing

Repeat Nobel Prize discoveries

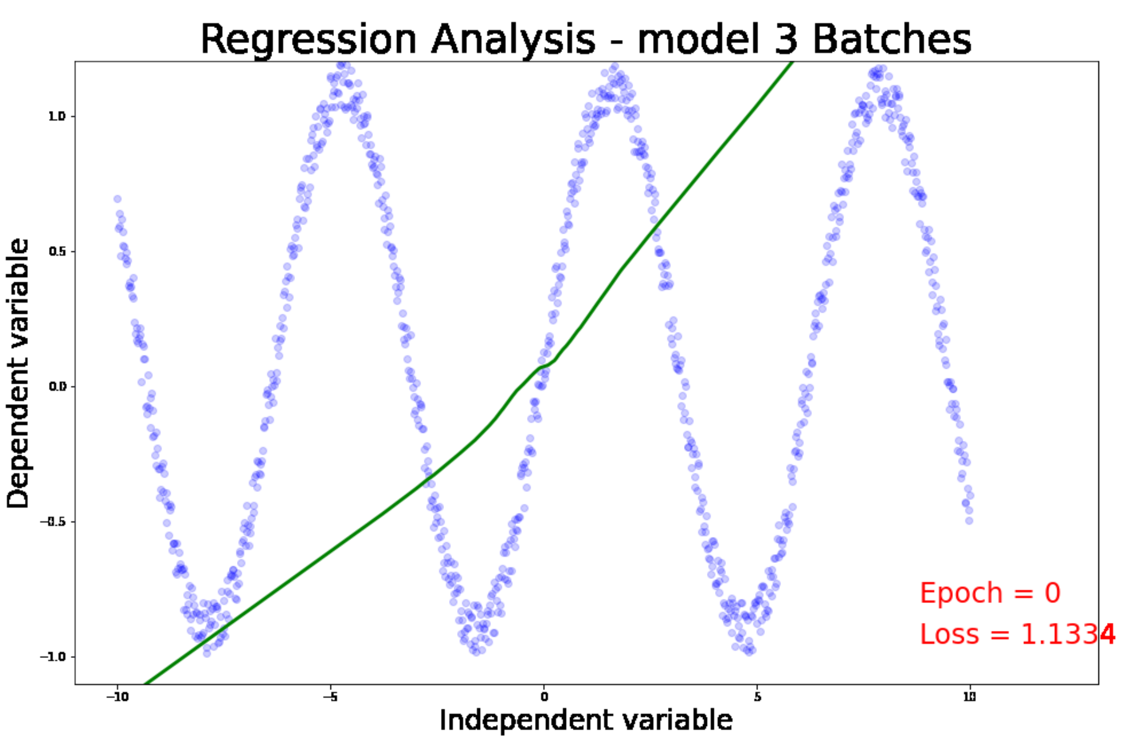
Lectures

Can we guess direction of the neutrinos and reconstruct the original tau energy?

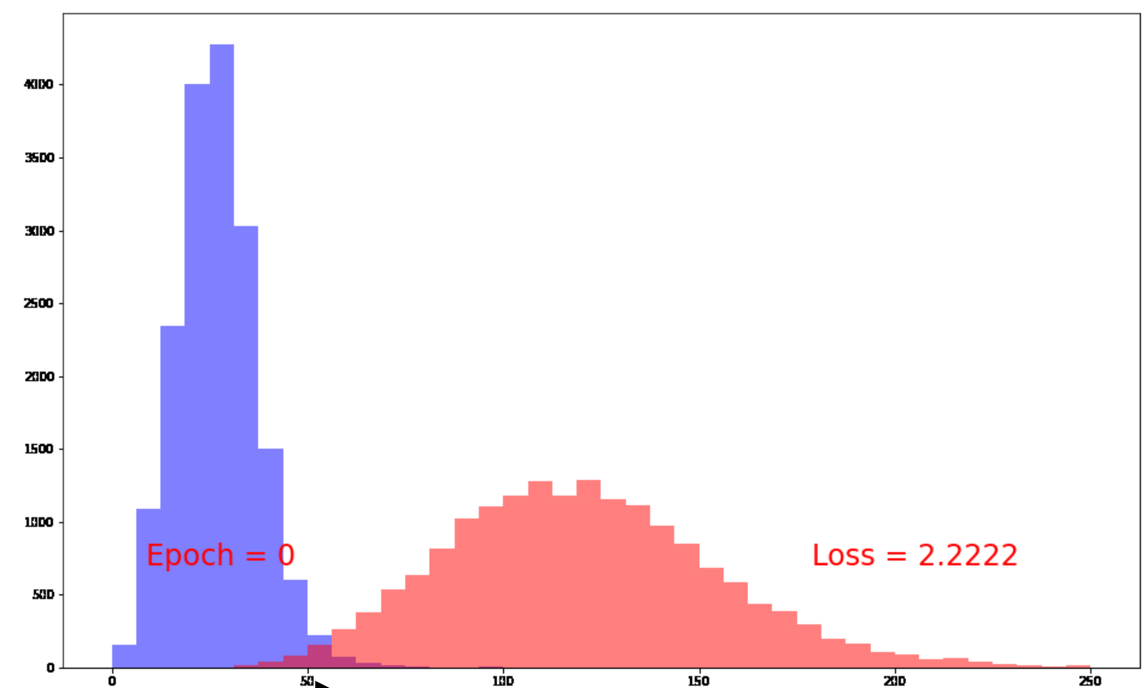


- simple: $p_T\ \tau = NN(p_T^{jet})$

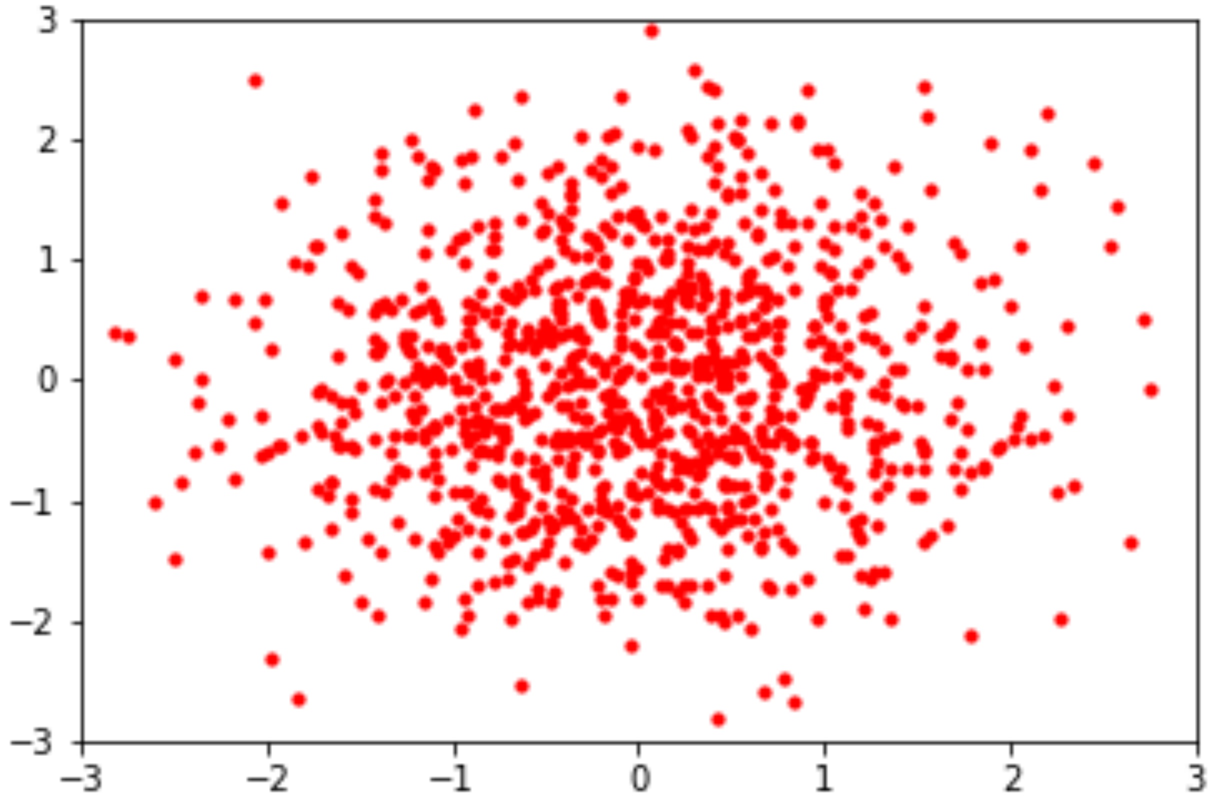
Concept (NN Regressions)

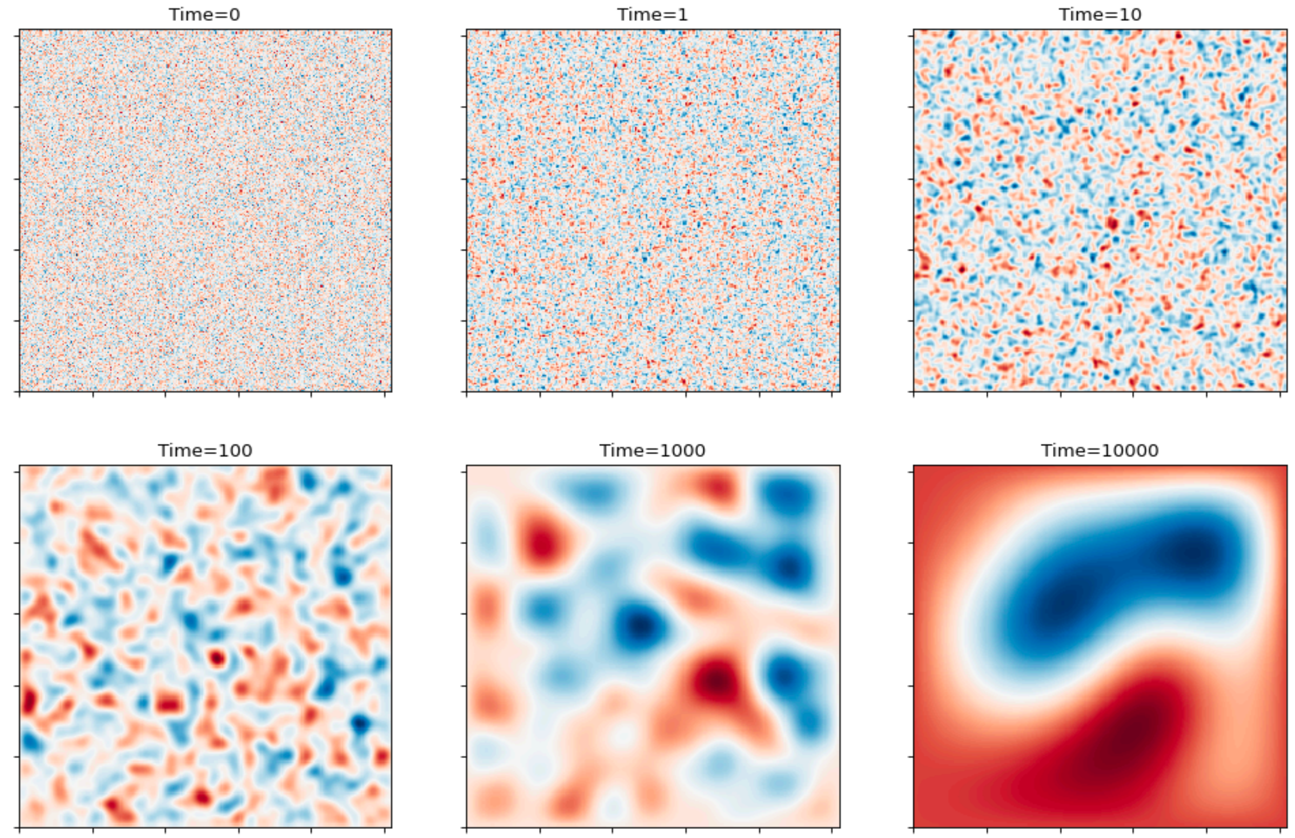


Physics
NN Regressions for Physics



Later Lectures





- Later on we will work on simulation
  - How do we model probability distributions
  - How do we use NNs to make this better!
- These are not as polished, but I think they are really cool!

Material

Projects utilize real Data

Real Data
means
Real Problems

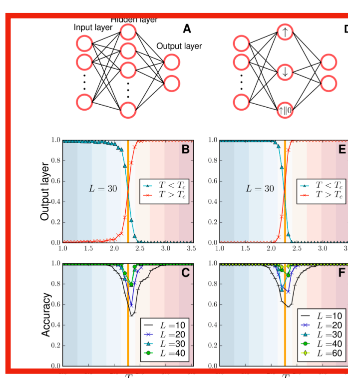
Real Data
also means
Real Reseach!

LIGO

Project 1 :
Gravitational Wave Data
From LIGO

CMS

Project 2 :
Collider Physics Data
from the Compact Muon Solenoid
on the Large Hadron Collider



Project 3 :
Ising Model/Lattice QCD
With embedded ML inside

chime

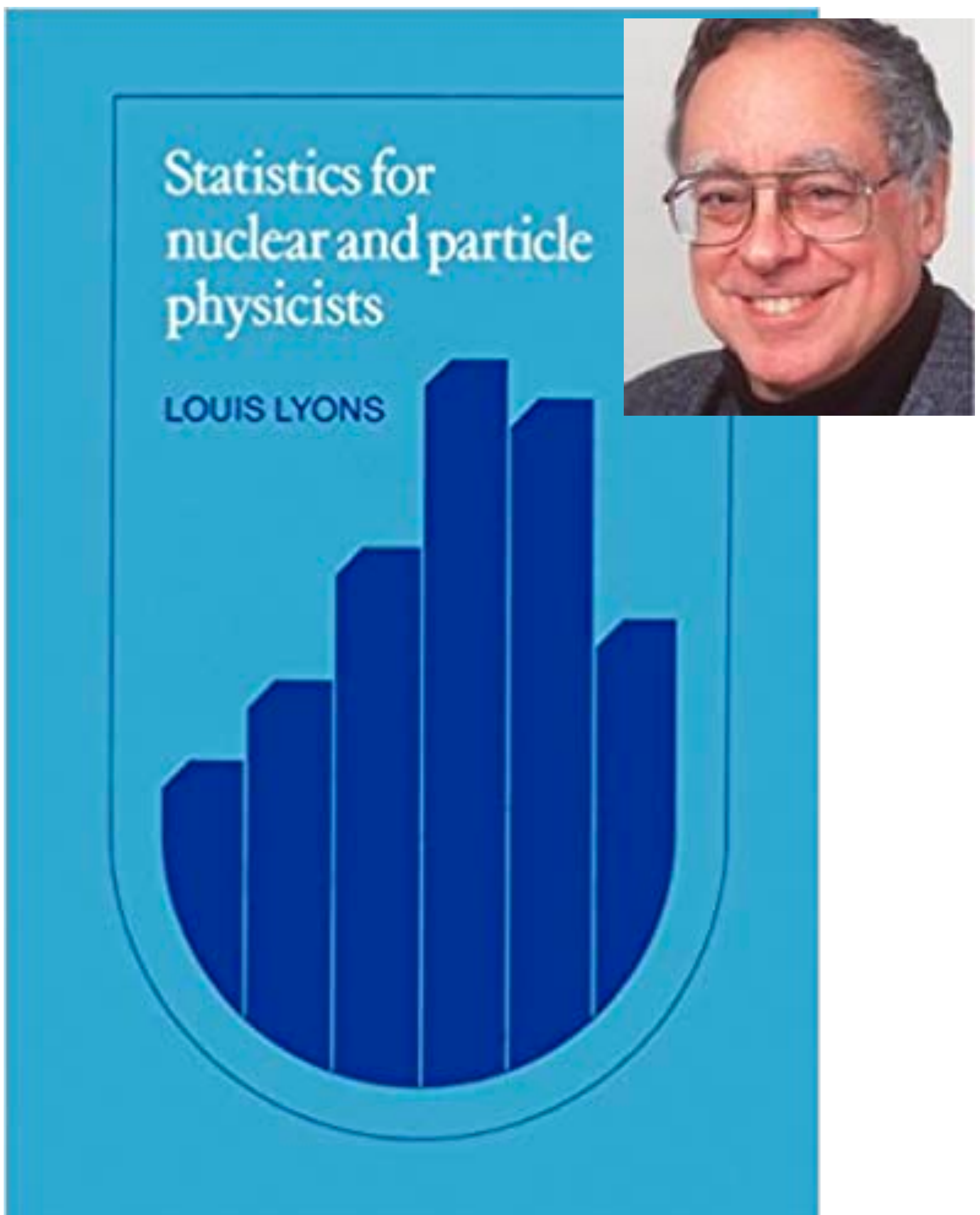
Project 4 : Your own or
Cosmic Microwave
Background
(simulated) Data

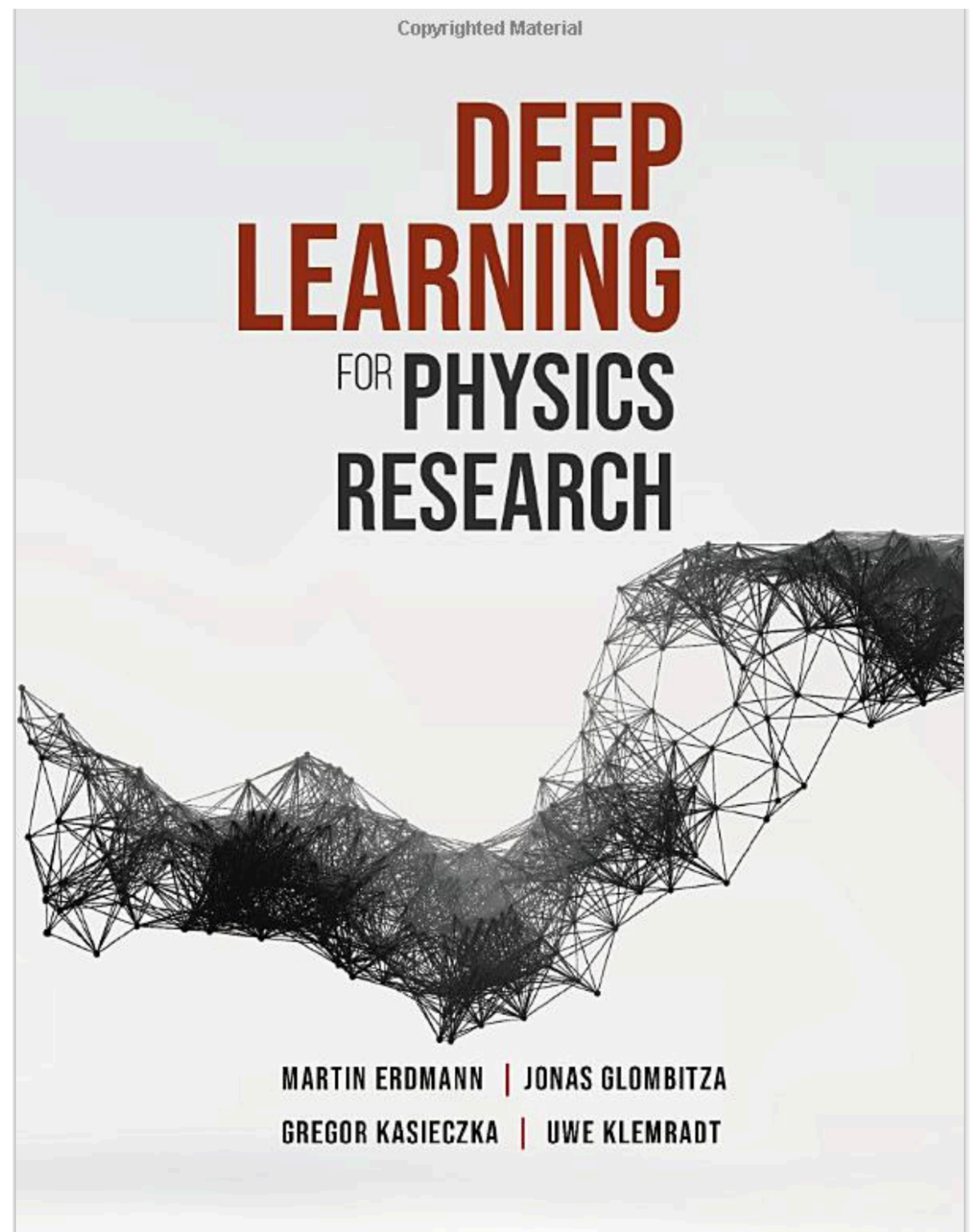
Grading

- Grading for this class will be :
  - A-F grading
- How do you do well in the class?
  - Turn in **something** for each project
  - Give a talk at the end
  - Do the problem sets
  - This class is ***for you to learn***
    - Don’t **stress** about the grades (really don’t)
    - If you have concerns come to my office hours

Textbook

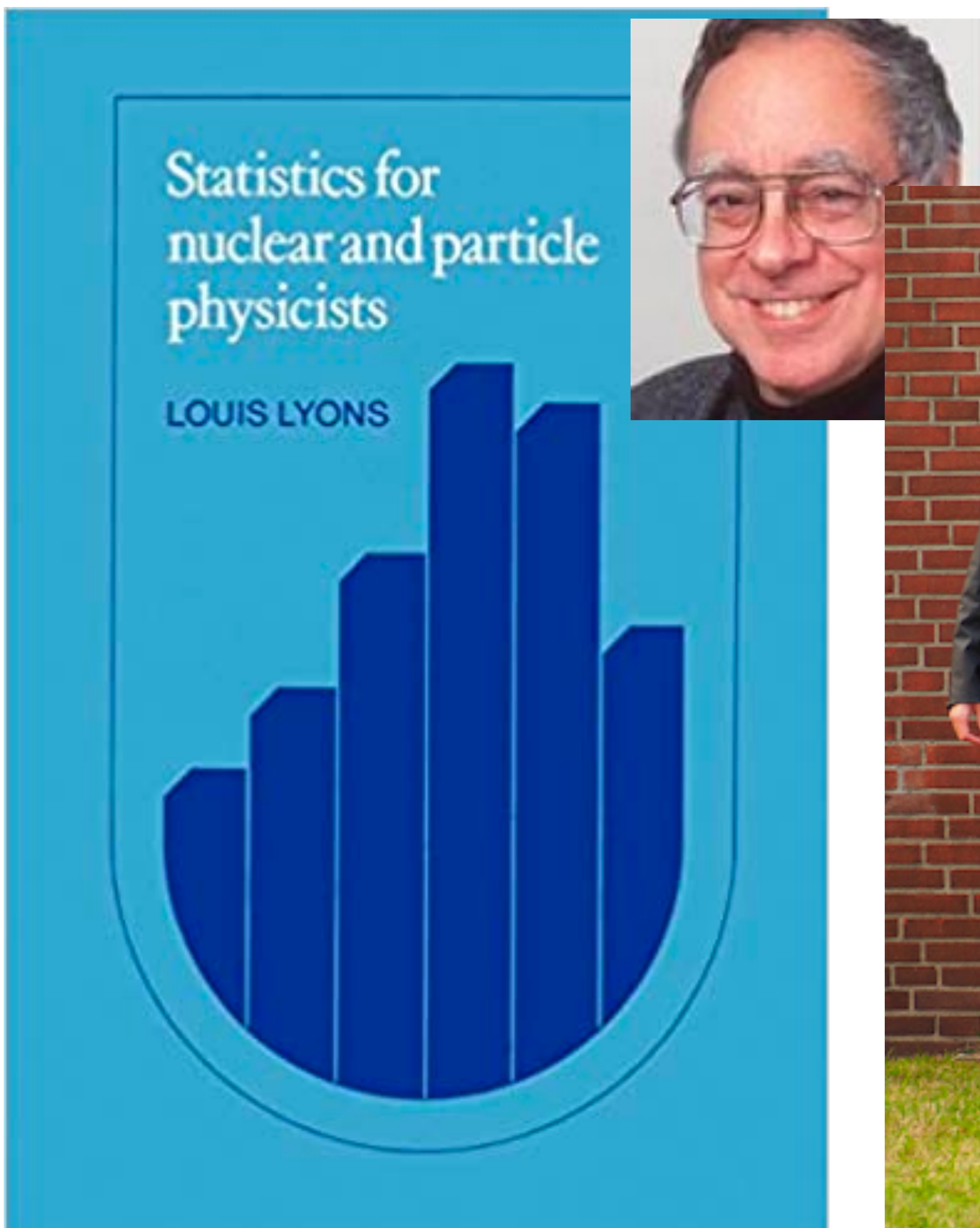
- Have a few suggested texts:
  - None of them do real justice to the topic





Textbook

- Have a few suggested texts:
  - None of them do real justice to the topic



Textbook

- Have a few suggested texts:
    - None of them do real justice to the topic

Other Resources I have used

Introduction to statistics and measurement analysis for physicists

https://inspirehep.net/literature/704473

MIT 18.05 Lecture notes:

http://www-math.mit.edu/~dav/05.dir/05.html

Advanced Methods in Applied Statistics:

Class Notes(Niels Bohr Insittute)

Kyle Cranmer's book

http://theoryandpractice.org/stats-ds-book/intro.html

More!

MITx course: https://github.com/mitx-8s50/nb_LEARNER

UIUC Data Analyis and machine learning : https://illinois-mla.github.io/syllabus/

UCSD Data Science Capstone: https://dsc-capstone.github.io

CMS Collaboration, “2020 CMS Data Analysis School": https://lpc.fnal.gov/programs/schools-workshops/cmsdas.shtml

2020 Hands-on Advanced Tutorial Sessions at the LPC: https://lpc.fnal.gov/programs/schools-workshops/hats.shtml

Computational and data science training for high energy physics.: https://codas-hep.org

2021 Machine Learning and the Physical Sciences Workshop.: https://ml4physicalsciences.github.io/2021 P. Calafiura, D. Rousseau and K. Terao, Artificial Intelligence for High Energy Physics, World Scientific (2022), 10.1142/12200
UCSD “Particle Physics and Machine Learning.” https://jduarte.physics.ucsd.edu/capstone-particle-physics-domain 10.5281/zenodo.4768815

G. Cowan, “Statistics for Particle Physicists.” https://cds.cern.ch/record/2773595

The 2020 US-ATLAS Computing Bootcamp website : https://indico.cern.ch/event/933434

BU “Machine Learning for Physicists.” : http://physics.bu.edu/~pankajm/PY895-ML.html

UMN “Big Data in Astrophysics.” : https://github.com/mcoughlin/ast8581_2022_Spring

UIUC Fundamentals of Data science: https://github.com/gnarayan/ast596_2020_Spring

Vanderbilt Astrostatistics: https://github.com/VanderbiltAstronomy/astr_8070_s21

Drexel Big Data Physics: Methods of Machine Learning: https://github.com/gtrichards/PHYS_440_540

Caltech Astroinformatics: https://www.astro.caltech.edu/ay119/

GROWTH summer school: http://growth.caltech.edu/growth-school-2019.html

AURA winter school: http://www.aura-o.aura-astronomy.org/winter_school/ - go to Past Years.

YouTube Neural Networks: https://www.youtube.com/watch?v=aircAruvnKk

Workload

Projects use real world data and simulations

You can write scientific papers with this data

- How much effort should you put in?
- $X = \dfrac{\text{hours per day}}{24h}, \; 0 < X < 1$
- Amount is is up to you
  - You could put 24 hours per day in it if you wanted to, so be careful
  - Projects are easy to make progress, but **nearly impossible to finish**
  - For this class, *some* effort is enough for success

Software Requirements

- This class will rely on Jupyter notebooks to run
  - https://jupyter.org/install
  - Be sure to get it installed as soon as possible
  - Lectures, projects, recitaations are all in jupyter
- Additionally, you need some standard python packages
  - scipy, numpy, matplotlib,gwpy (project1), uproot (project2)
- If you don’t know you can use Google Collab
  - Let me show you! https://colab.research.google.com/notebook

Class Format

- Lectures 2:30-4:00pm
- Class will recorded
  - We will record everything in class
  - They will be available on the canvas site
  - Lecture pdfs will be available on GitHub
    - You can follow much of this class remotely
    - Class attendance is **not mandatory**, lose ability to ask questions
- Please ask questions in class
  - This is your time to get the feedback

8.16/8.316 Divide

- This is an interesting divide
    - Its hard to know where to draw the line
- At this stage our/my view:
    - Preserve the assignments
        - Assignments have additional tasks
    - Change the level of grading for 8.16 vs 8.316
        - We will be **clear** about where the stopping point is for each level
- Please note that every project has an advanced project on top
    - Advanced project is very much graduate level
    - We will not make this a requirement for the class, its **good start for project 4!**

Lets get lectures

- We are going to put everything on GitHub!
- If you have GitHub already installed:
    - https://github.com/orgs/mit-physics-data/repositories
    - git clone git@github.com:mit-physics-data/lectures.git
    - git clone git@github.com:mit-physics-data/psets.git
    - git clone git@github.com:mit-physics-data/projects.git
- We will release everything on GitHub first
    - Psets and projects will be available on canvas too

Why this class?

I once had dinner with
Prof. Jerry Friedman
Nobel Prize 1990

He told me for his Ph.D
Thesis he did a fit to data

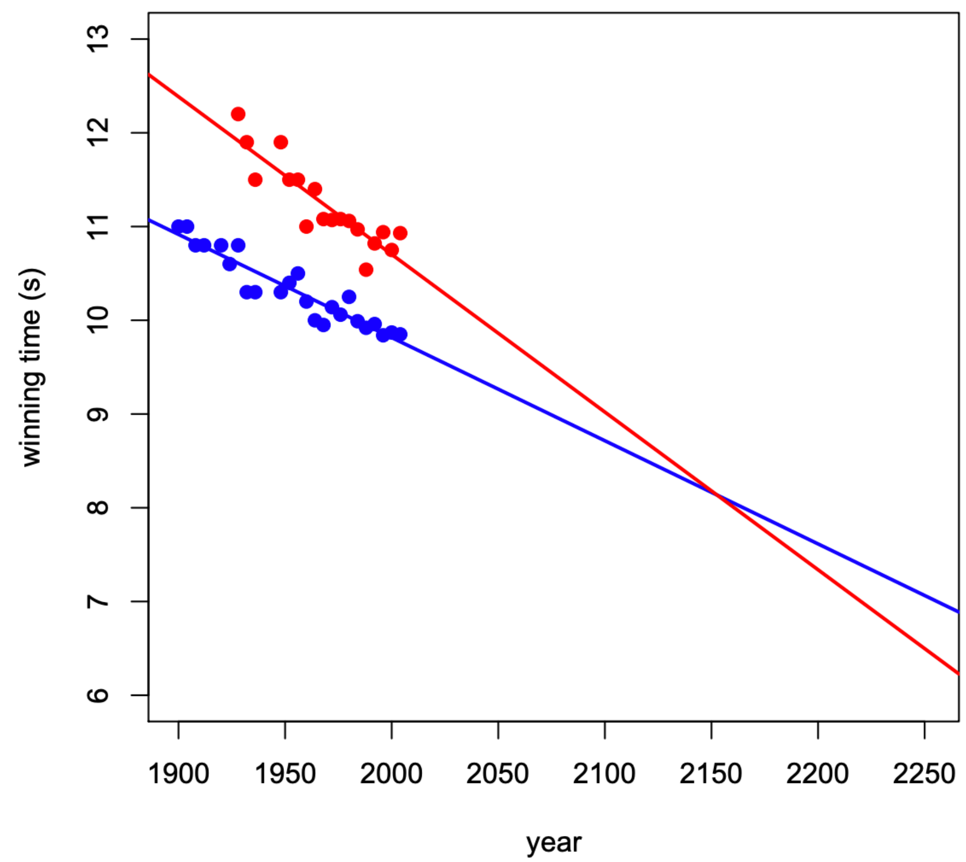
It took him a **whole summer**
Enrico Fermi was his advisor

https://www.nobelprize.org/prizes/physics/1990/friedman/biographical/
https://www.youtube.com/watch?v=iLupedvSsFA

- The same thing Prof. Friedman did now takes 5 min
- There is a data science revolution underway

Whats wrong?

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.

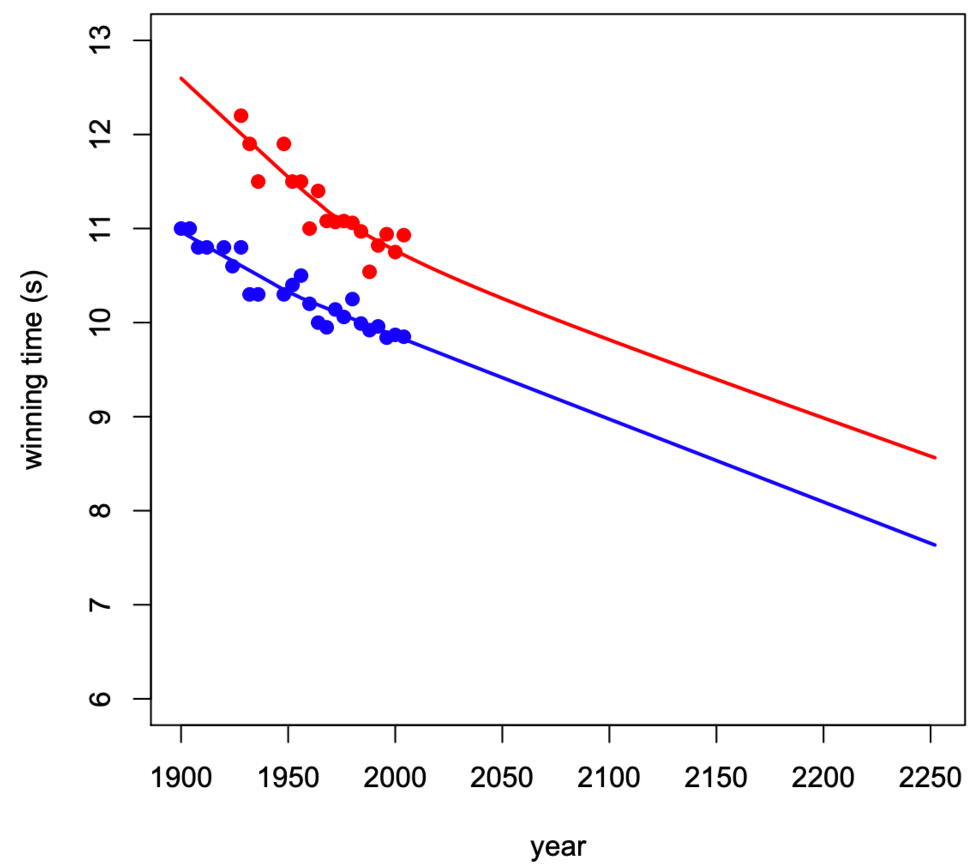


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

- Does this make sense?

Whats wrong?

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.

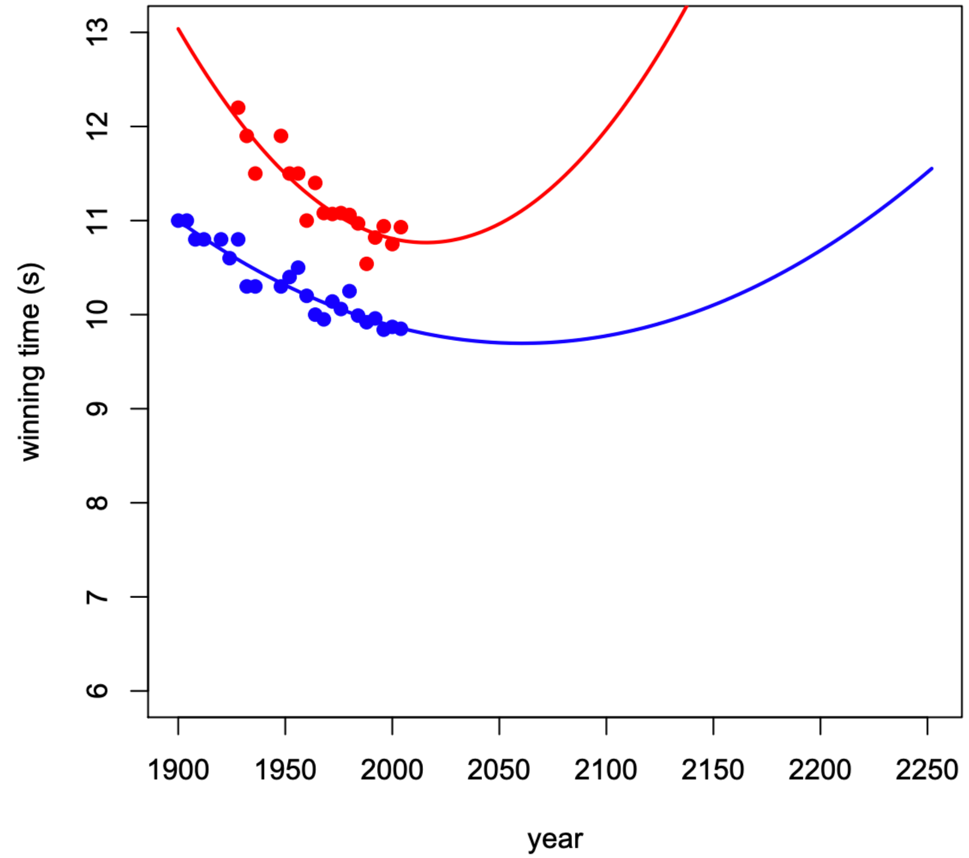


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

- Does this make sense?

Whats wrong?

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.



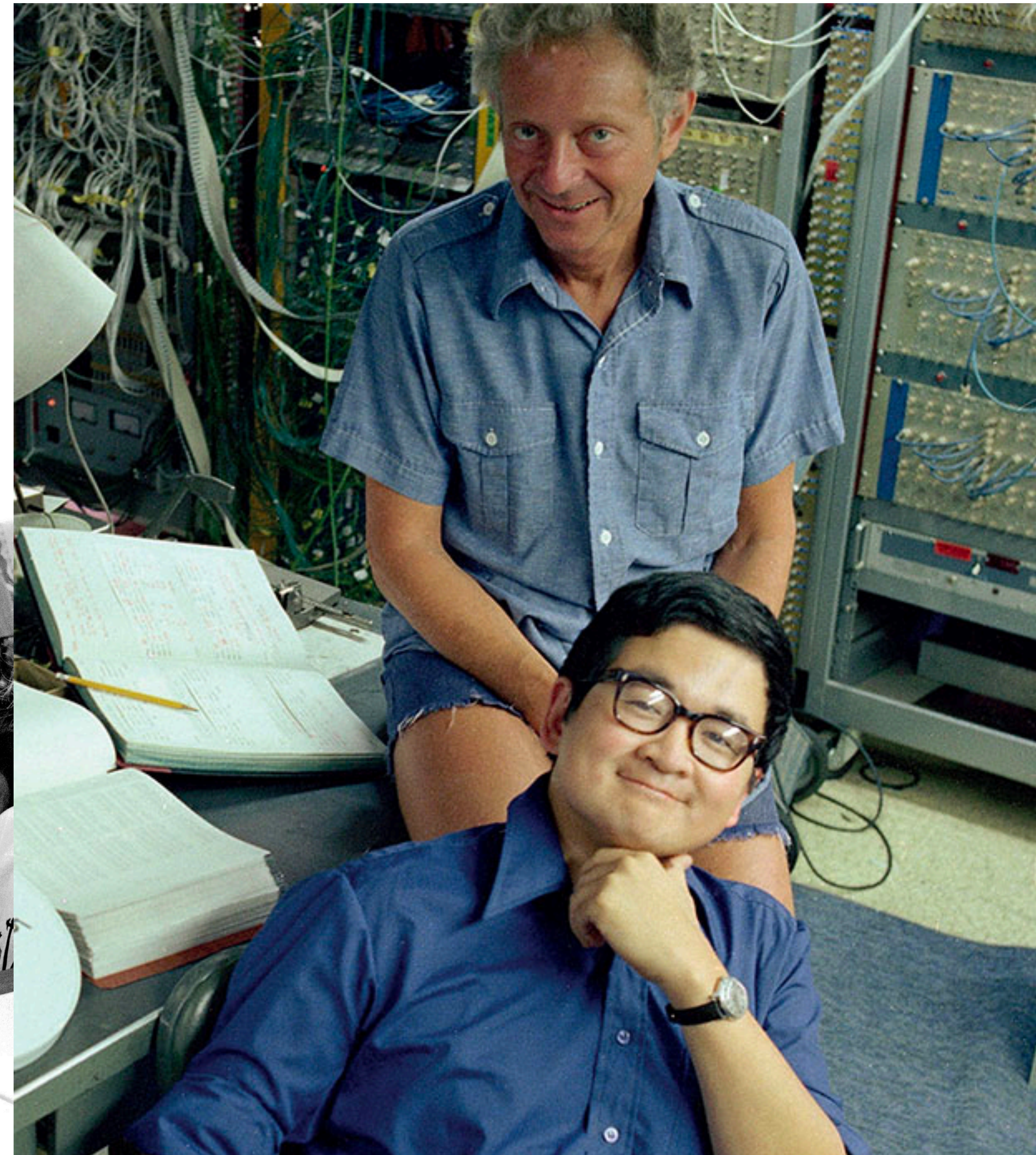
Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

You can't have data science without science

- Choice of model requires some intuition within the field

Outside Chicago 1976

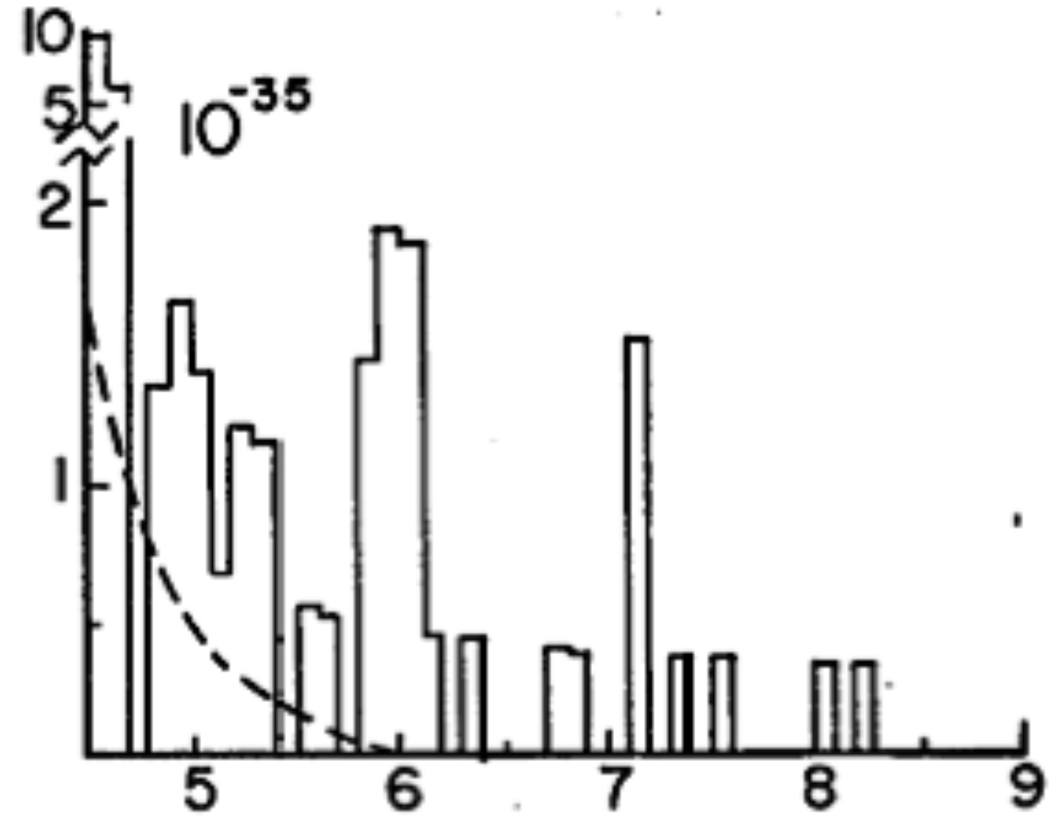
Outside Chicago 1976



Physics Blunders

https://en.wikipedia.org/wiki/Oops-Leon

- Cover the statistical tools
- For you to understand integrity



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

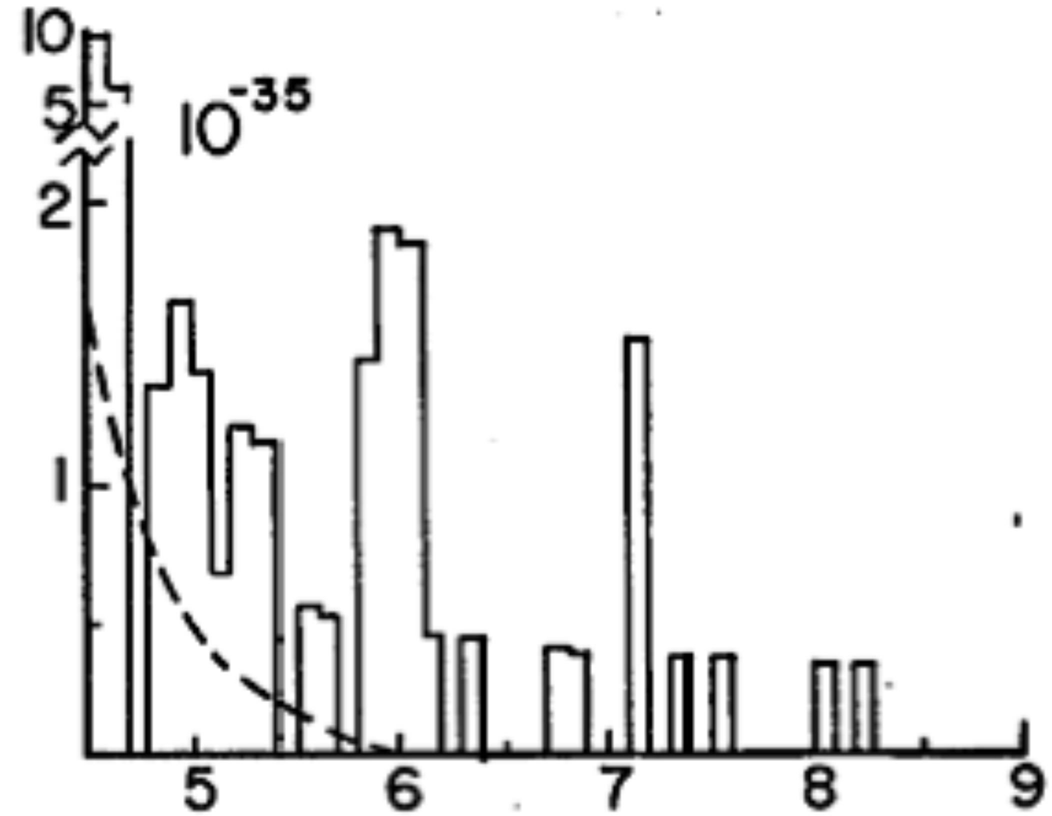
Hope is to build intuition

Physics Blunders

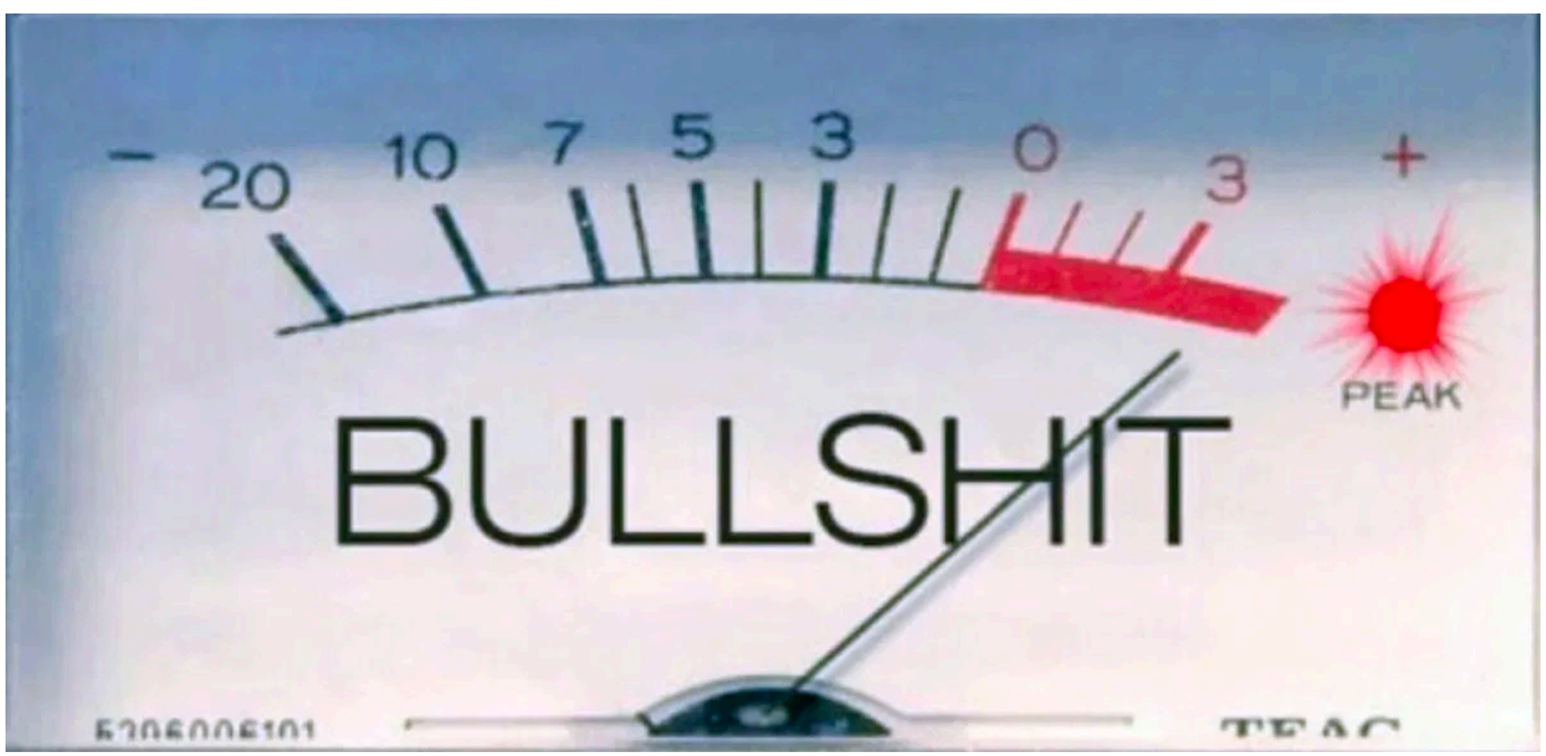
https://en.wikipedia.org/wiki/Oops-Leon

- Cover the statistical tools

- For you to understand integrity



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



Hope is to build intuition

Geneva 2022

Geneva December,2022

Some Excitement

OUT THERE

As the Large Hadron Collider Revs Up, Physicists' Hopes Soar

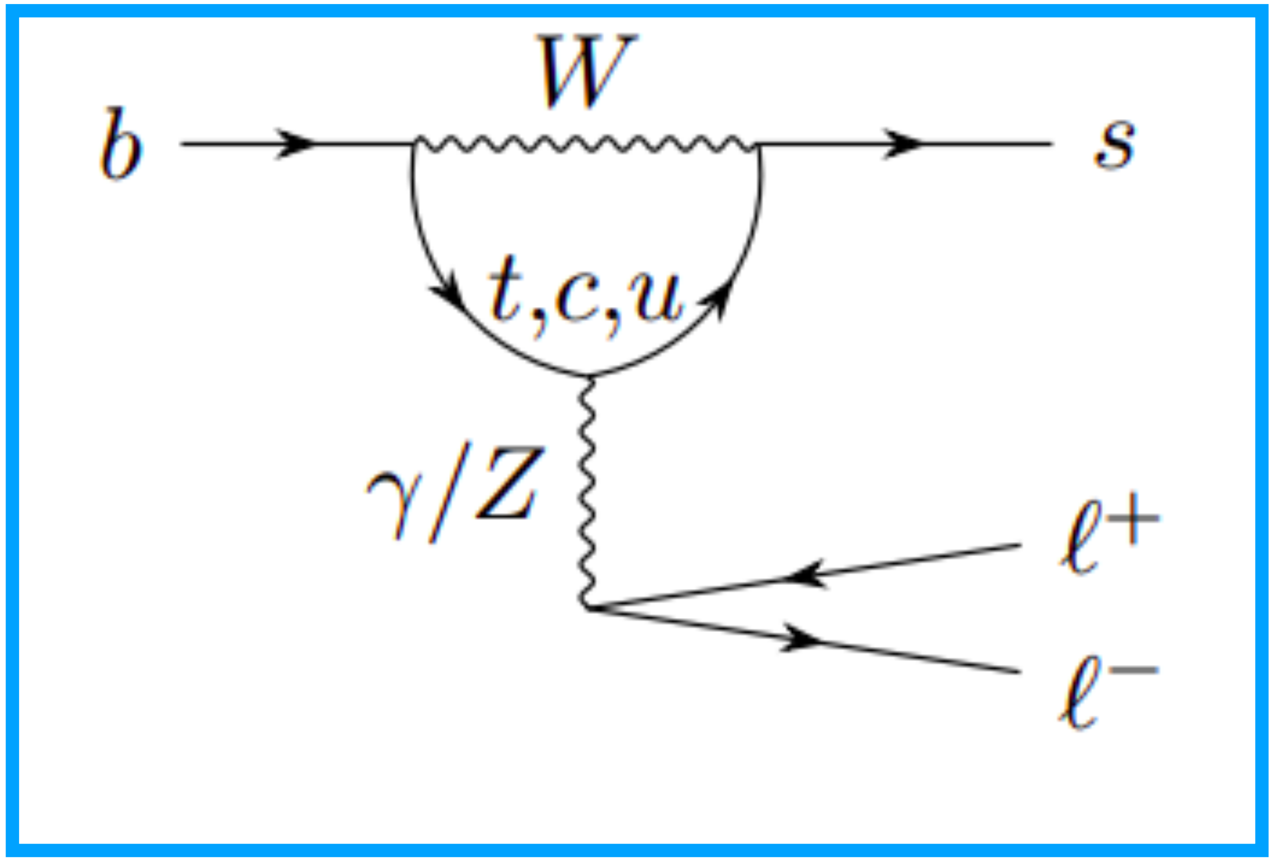
The particle collider at CERN will soon restart. "There could be a revolution coming," scientists say.

Give this article

Inside the Large Hadron Collider near Geneva, a worker uses a bicycle to navigate its 17 miles of tunnels during maintenance in 2020. Valentin Flauraud/Agence France-Presse

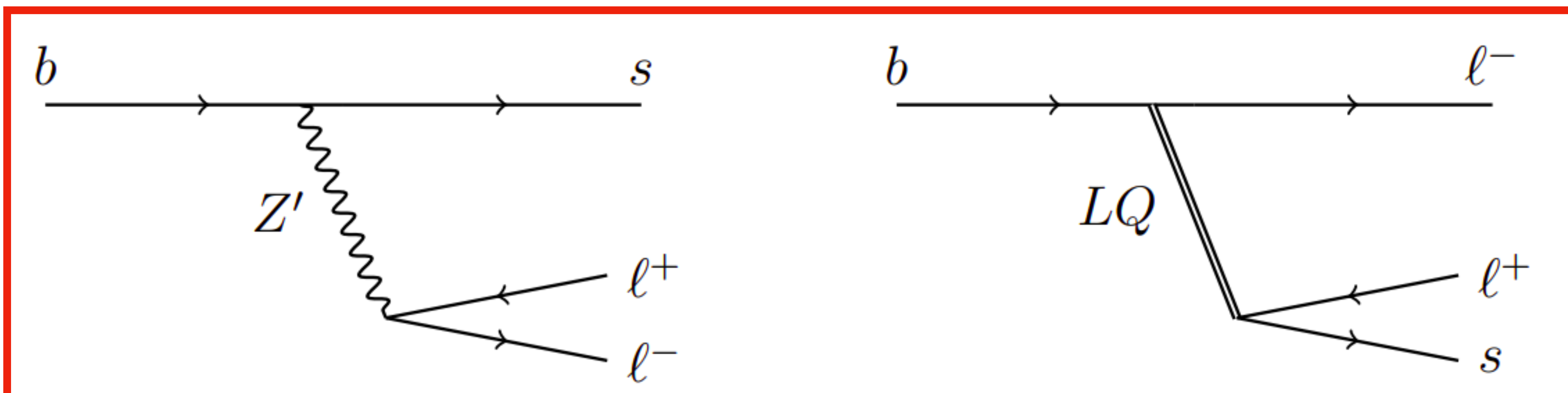
https://www.nytimes.com/2022/06/13/science/cern-hadron-collider-muon-leptoquark.html

Over Past few years excitement

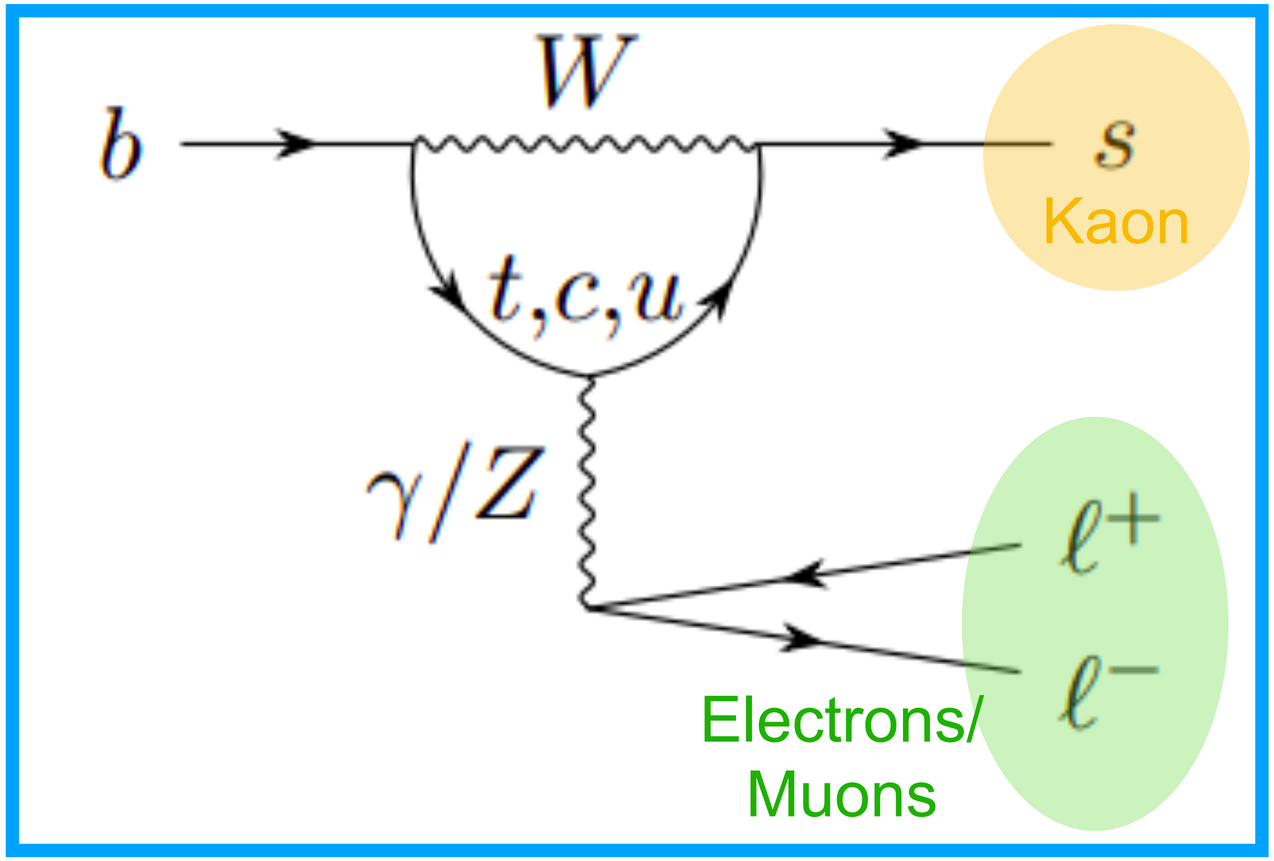


Expected

New Physics



Over Past few years excitement

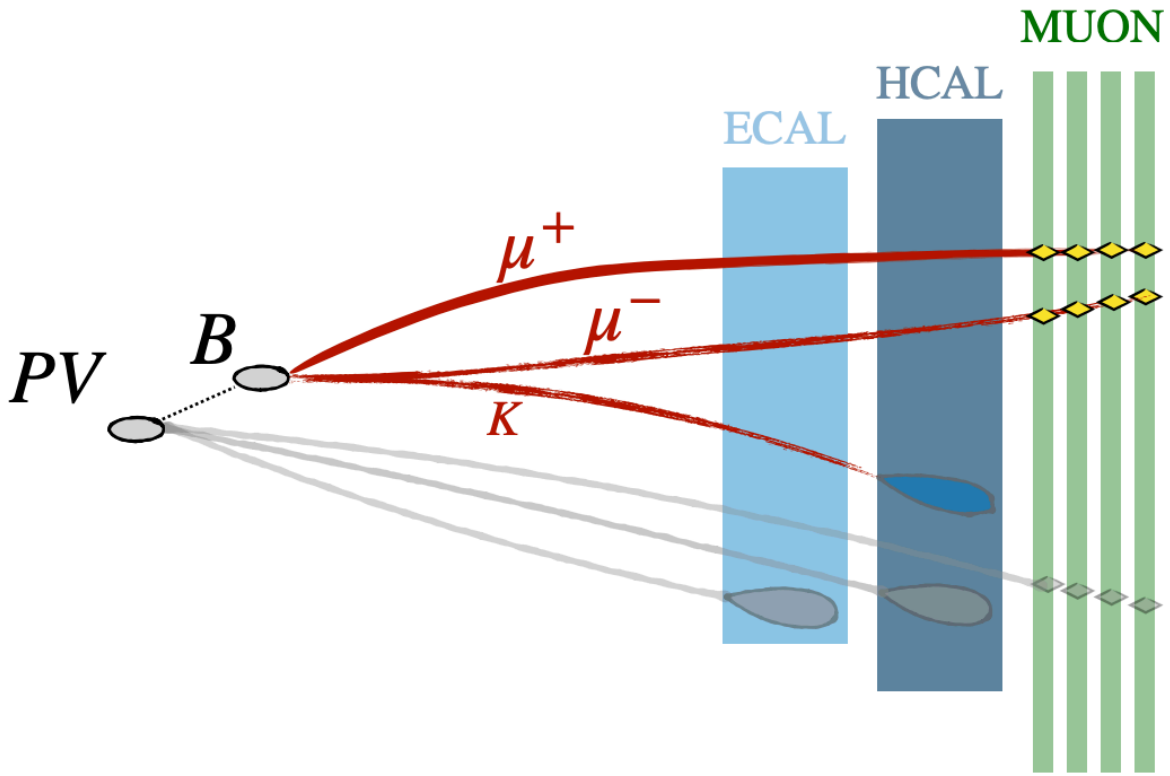


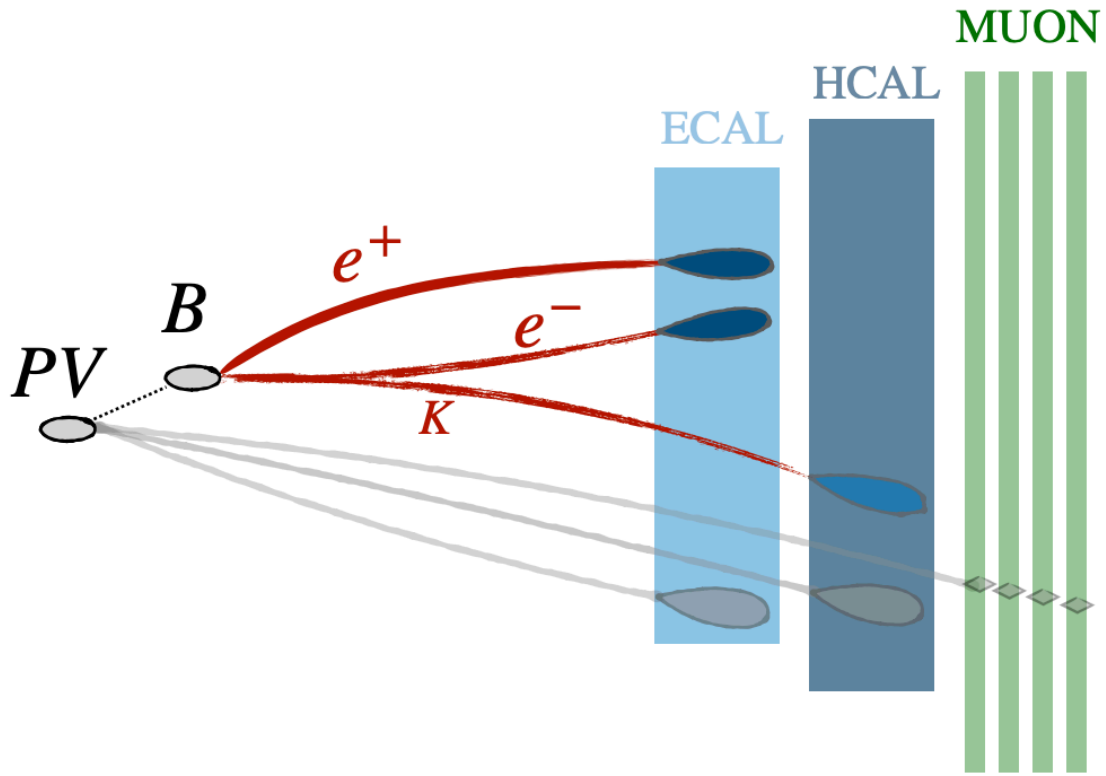
Expected

$$R_K = \frac{b \to K e^+ e^-}{b \to K \mu_+ \mu_-} = 1$$

Ratio of Electrons to Muons

Over Past few years excitement

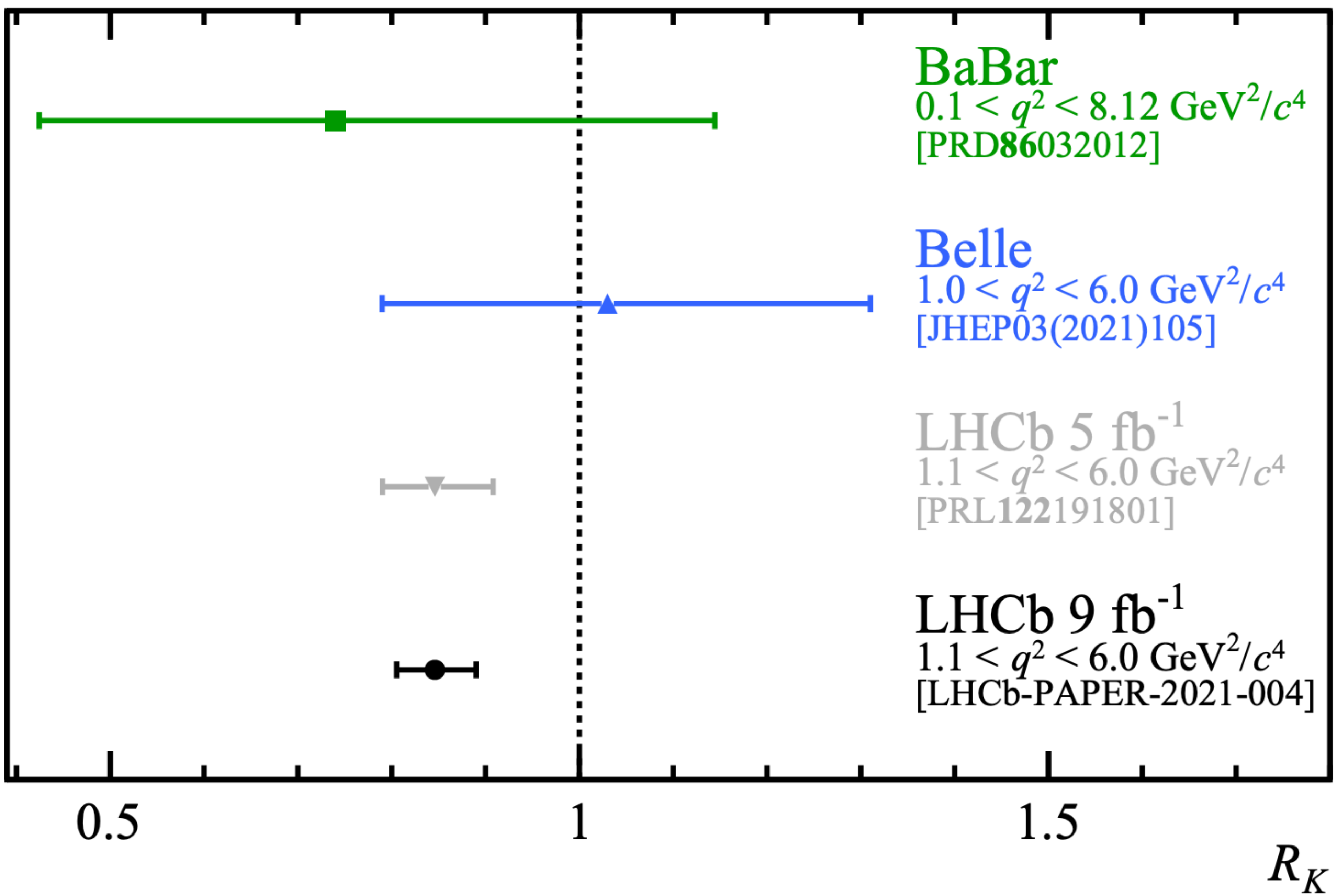




$$R_K = \frac{b \to K e^+ e^-}{b \to K \mu_+ \mu_-} = 1$$

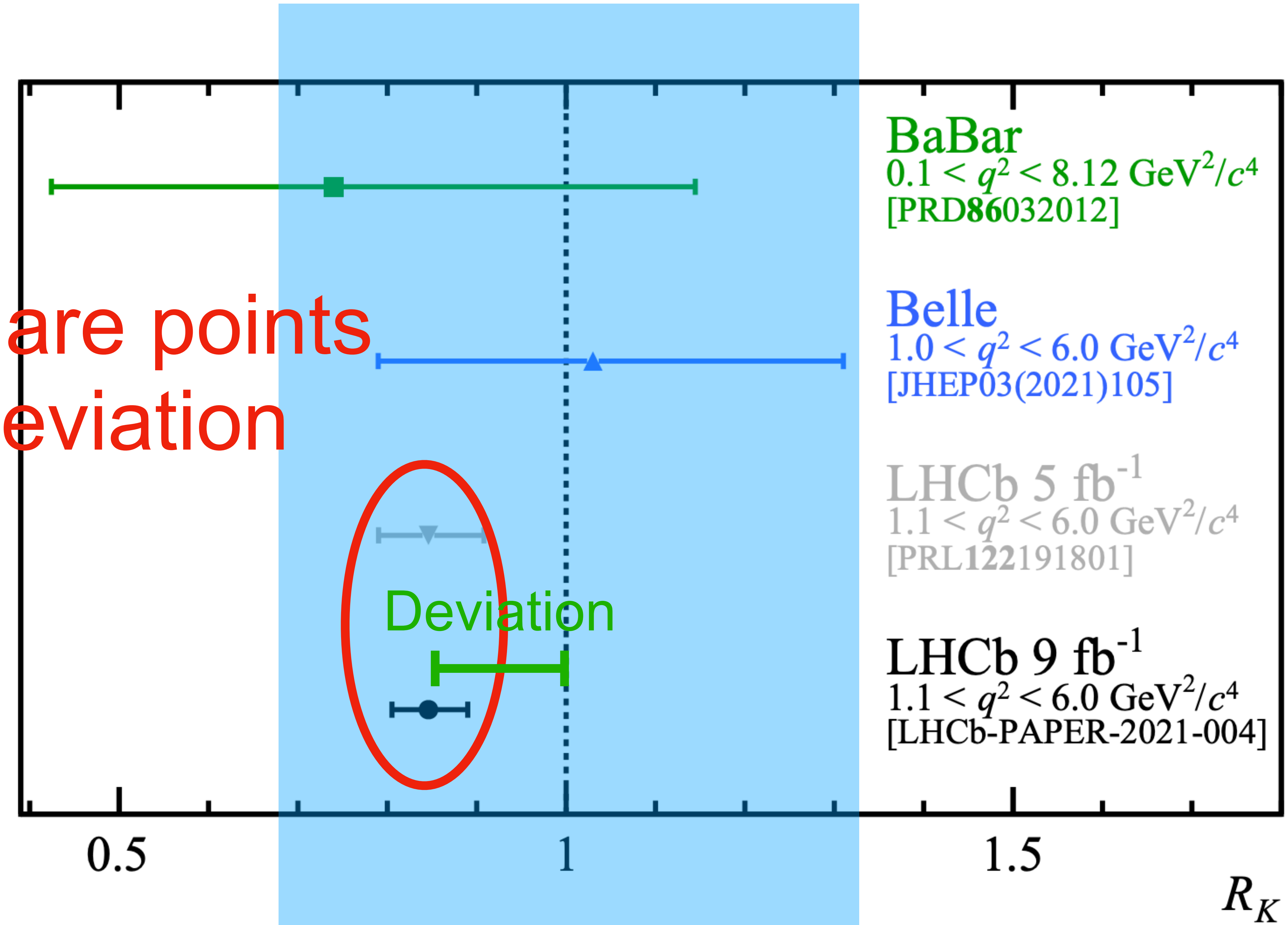
Ratio of
Electrons to
Muons

Over Past few years excitement



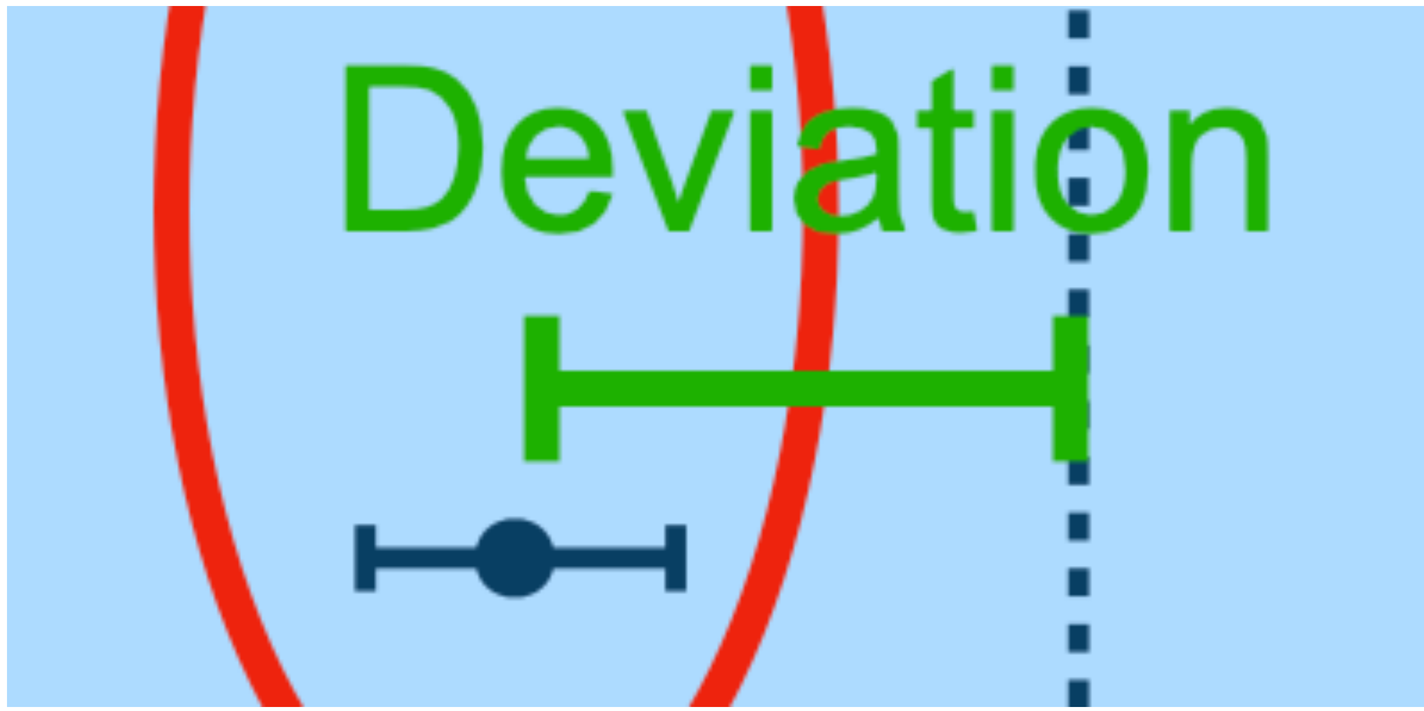
Anatomy of a Plot

There are points
That deviation
from 1



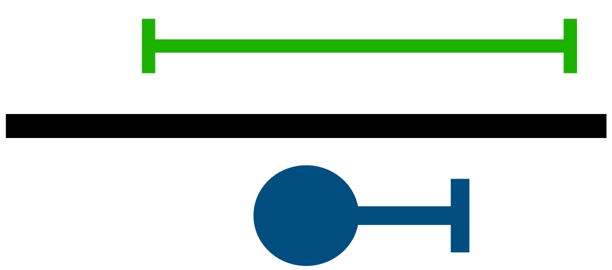
This is the ratio it should be 1

Anatomy of a Plot



Magnitude of Deviation is units of unc.

Deviation

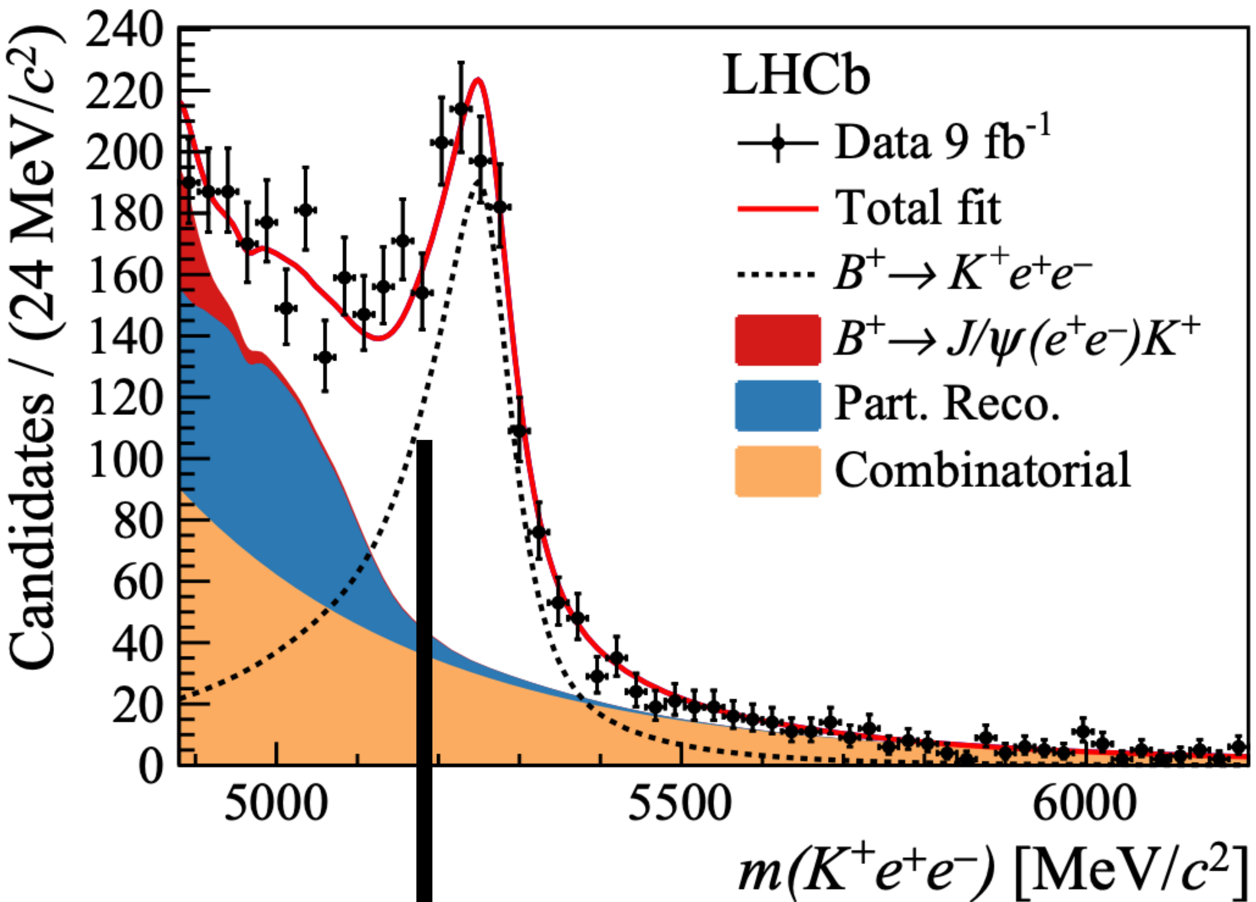


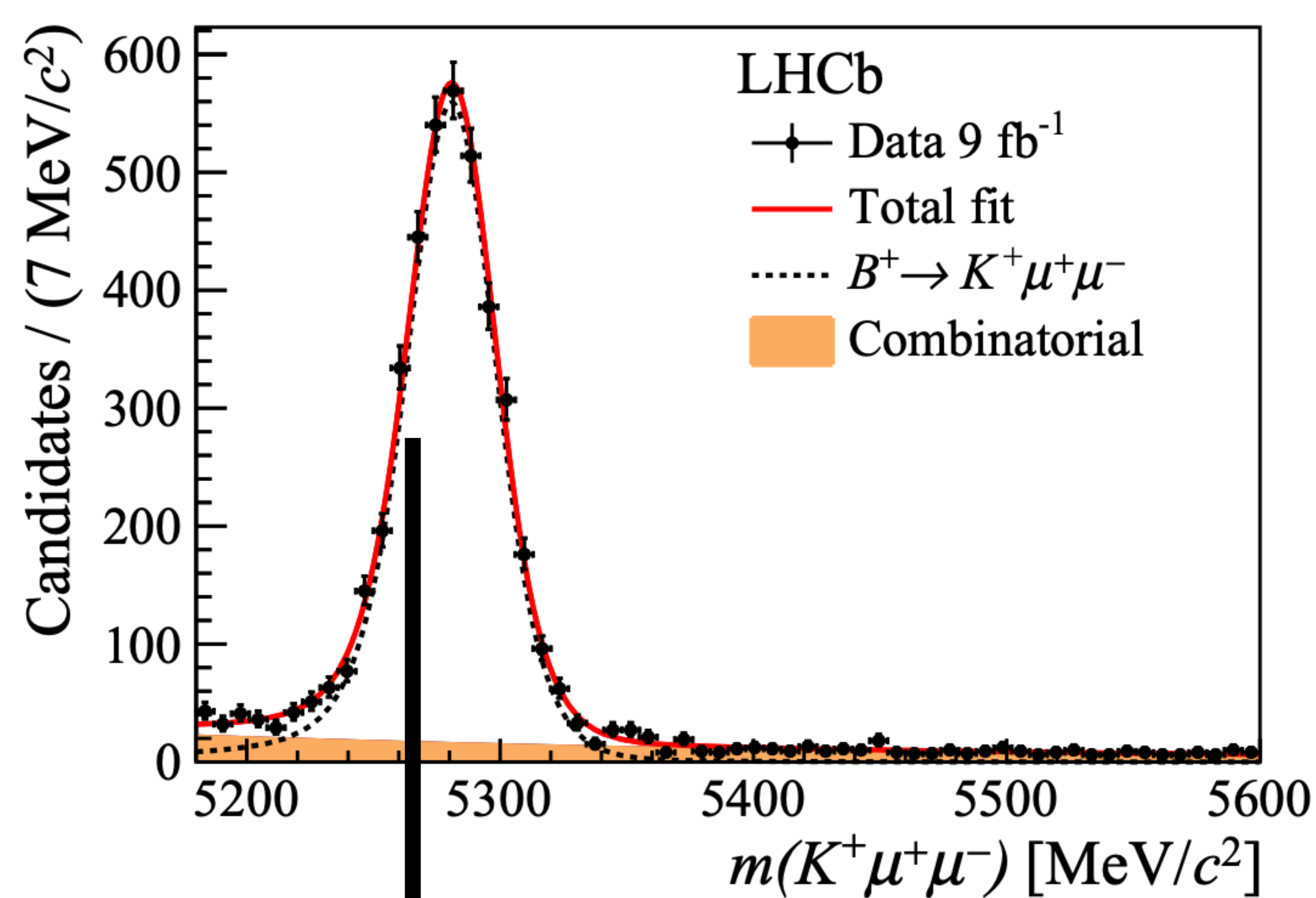
Measurement Uncertainty

$\approx 4\sigma$

4 Uncertainties
Is a lot of deviation!

The Actual Measurement



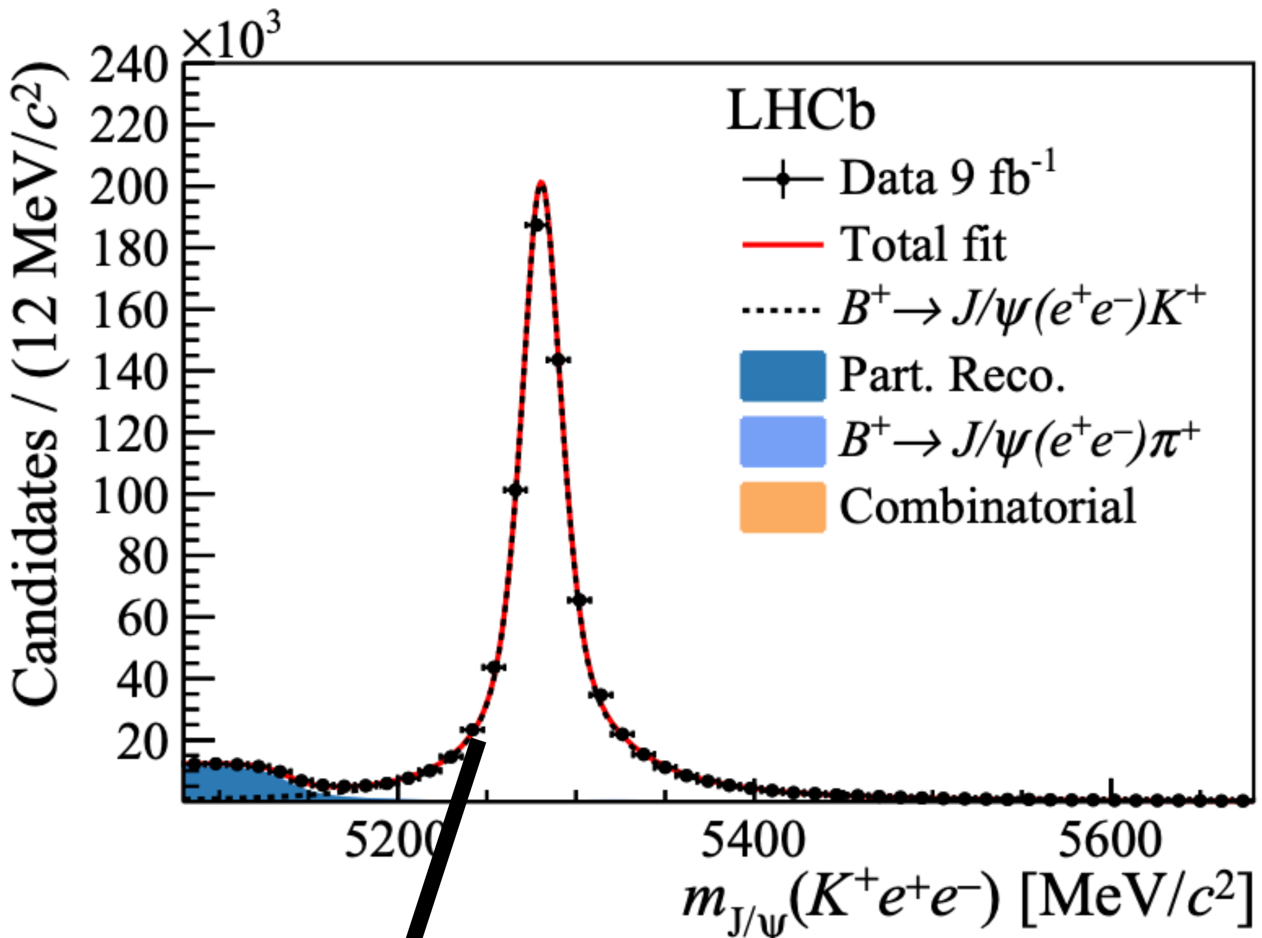


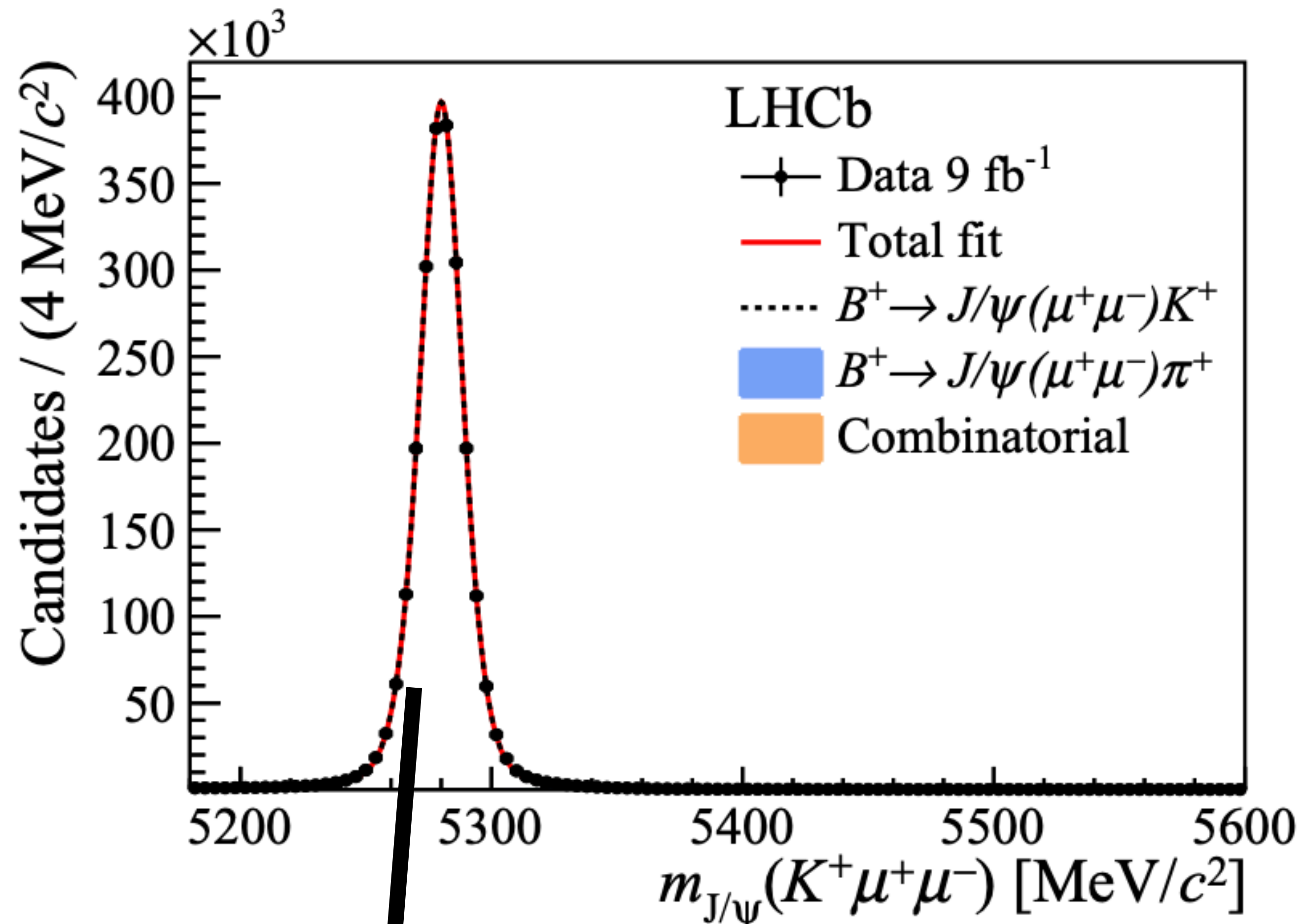
Integrate this dashed line
And correct for selection Biases

Integrate this dashed line
And correct for selection Biases

Compute the rates of these guys

The Check it works



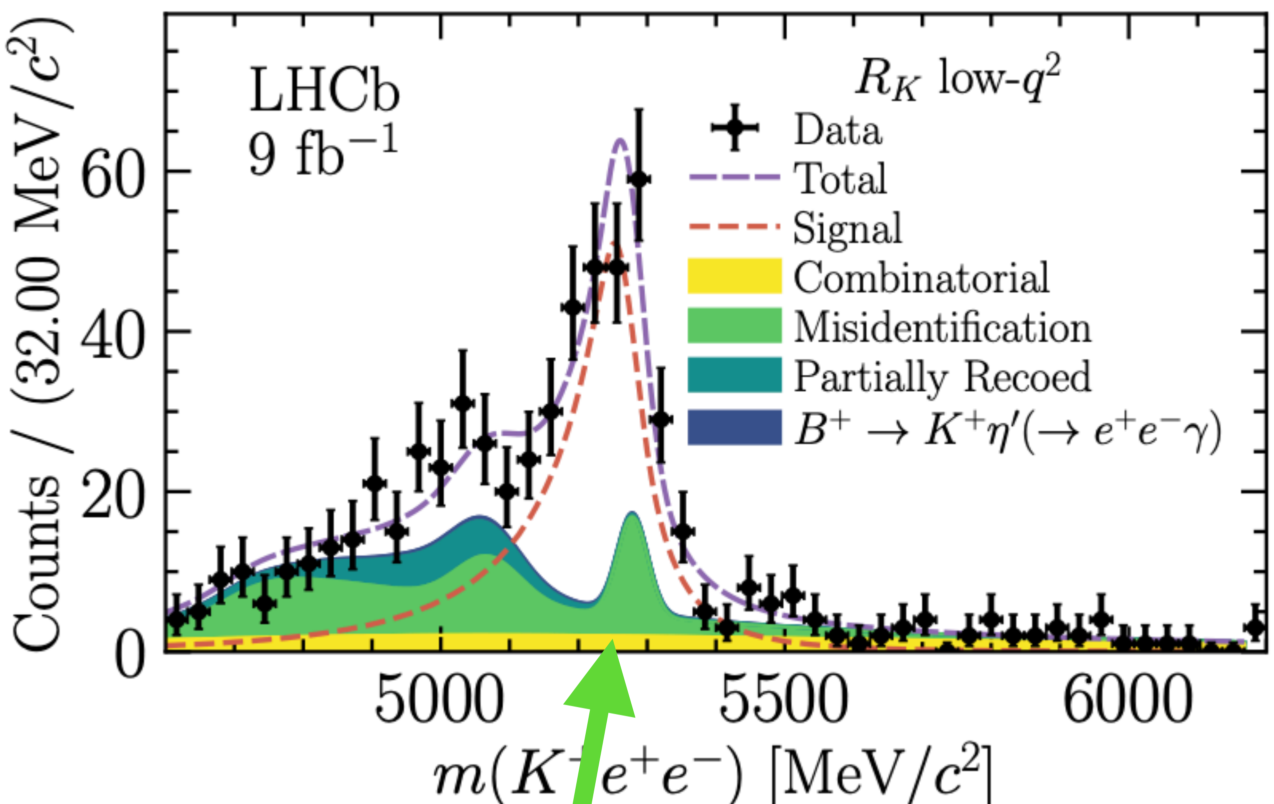


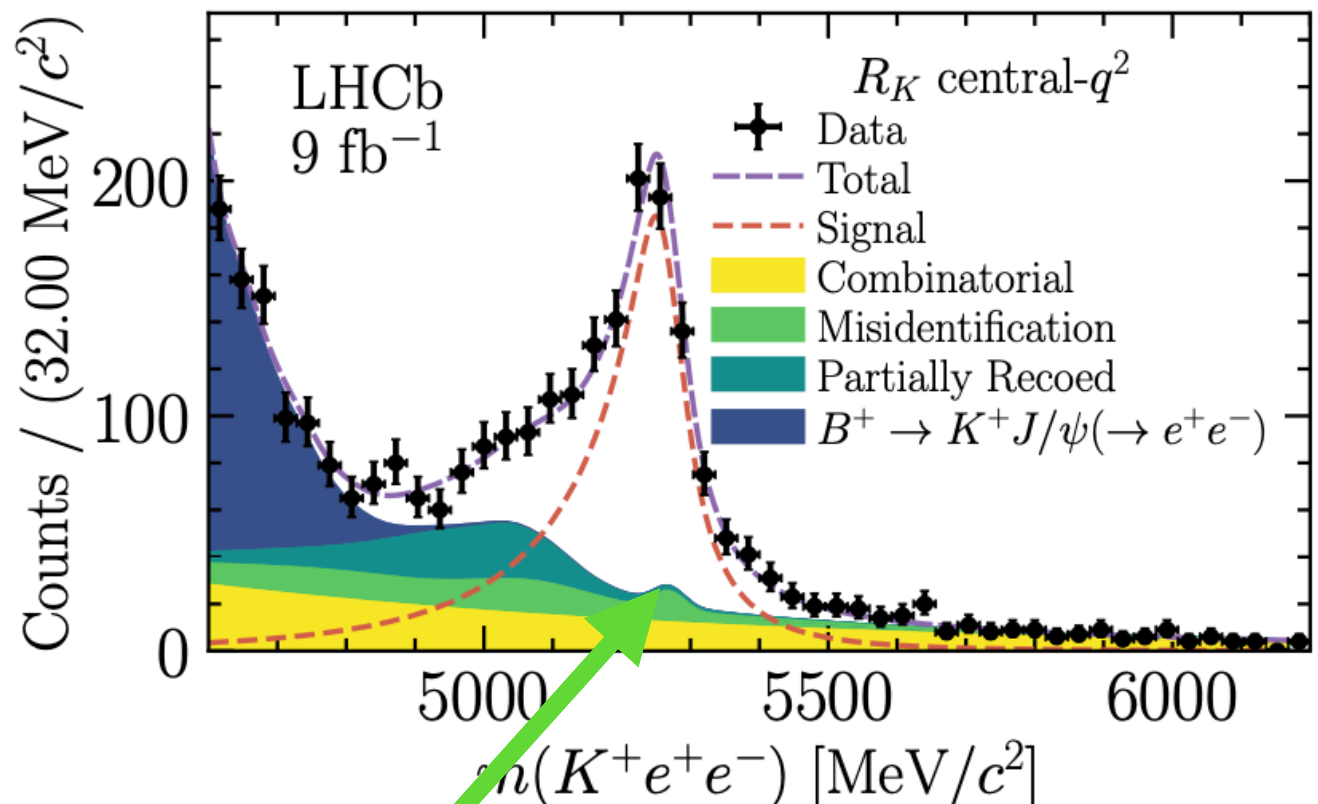
Integrate this dashed line
And correct for selection Biases

Integrate this dashed line
And correct for selection Biases

This does get us to 1 as it should

Over Past few years excitement





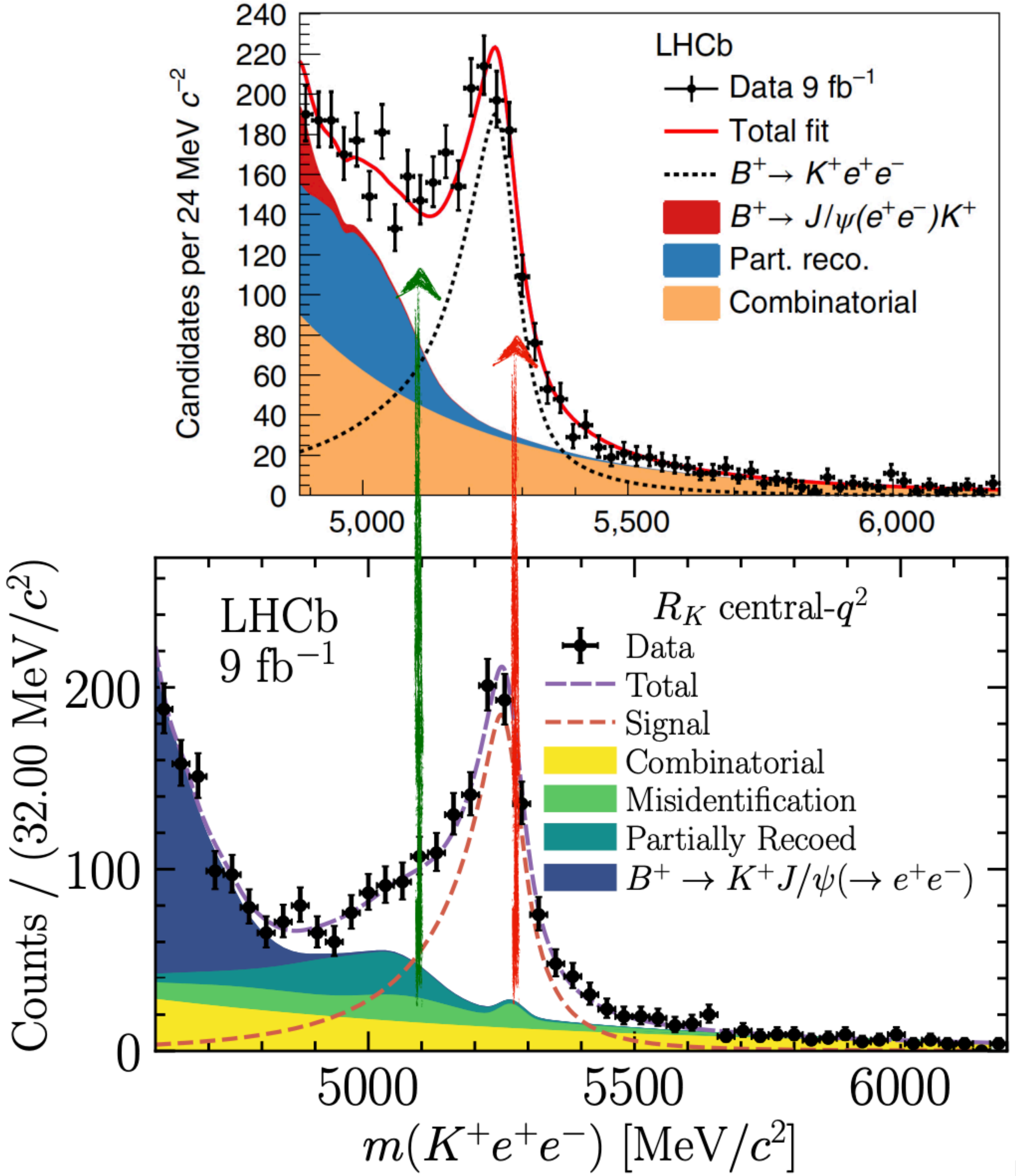
These lime green processes were forgotten

This was a very unwelcome problem

The High Energy Physics community is aghast at what’s going on

What we learn in this class?

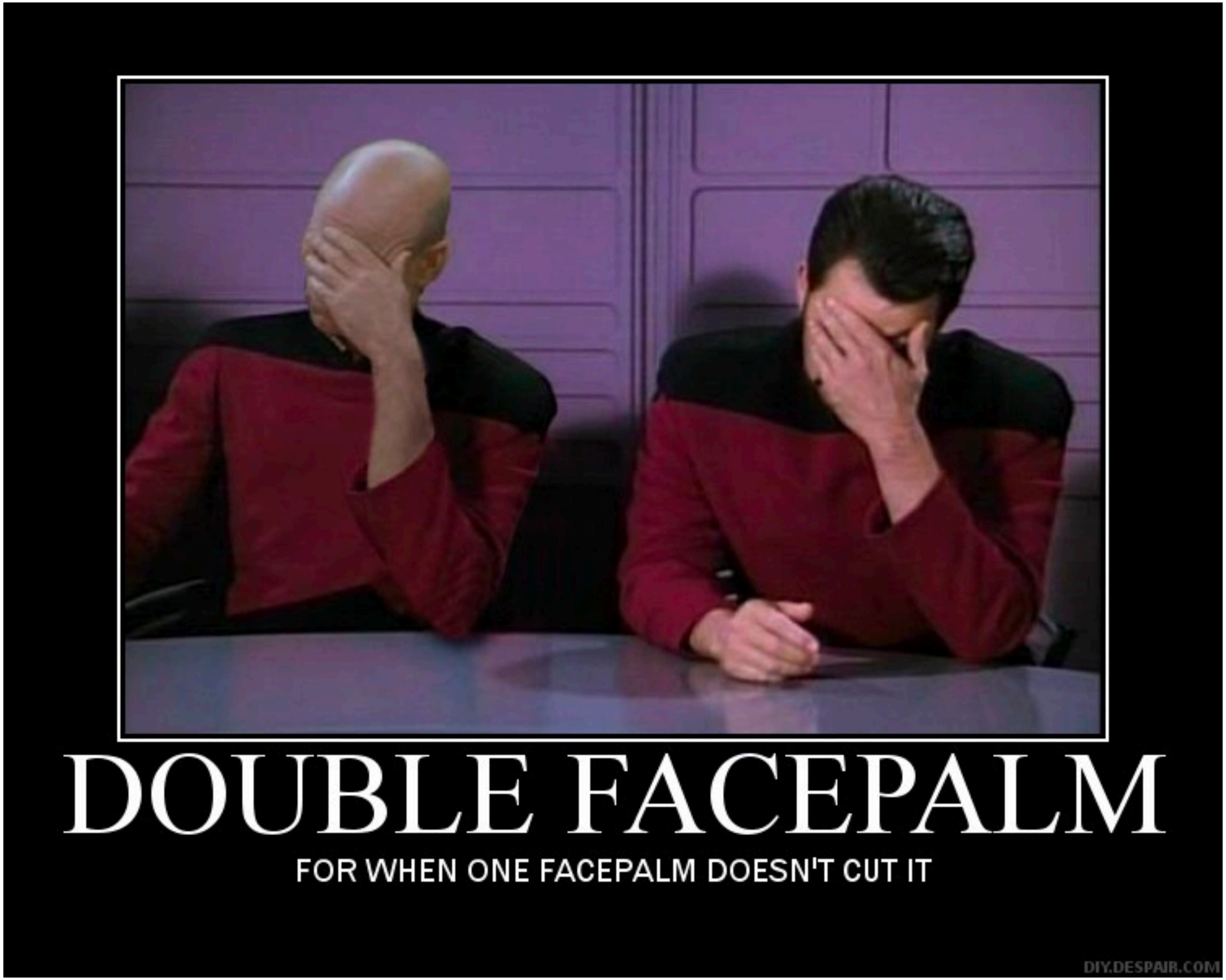
[Nat. Phys. 18, 277-282 (2022)]



I don't think I could teach you to immediately find a mistake

However, this class will give you the experience to have a healthy amount of scepticism for what is going on

Another one bites the dust



Data Science Mistakes

https://hackernoon.com/12-mistakes-that-data-scientists-make-and-how-to-avoid-them-2ddb26665c2d

1. Spending huge time on theory without practical application
2. Coding too many algorithms without learning the prerequisites
3. Jumping into Deep End
4. Focusing on Accuracy over Understanding how model works
5. Giving Preference to Tools over Problem
6. Overestimating Value of Academic Degrees
7. Thinking that if You don't code well, You can't be a Data Scientist
8. Using too many Data Science Terms in your Resume
9. Learning Multiple Tools at Once
10. Not Having a Structured Approach to Problem Solving
11. Not Working Consistently
12. Not working on Communication Skills

In this class we focus on how we apply data science to physics

Artificial Intelligence ⇔ Fundamental Interactions

[http://iaifi.org/, MIT News Announcement]

The NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)

"ι-φ"

Senior Investigators: 20 Physicists + 7 AI Experts
Junior Investigators: ≈20 PhD Students, ≈7 IAIFI Fellows in steady state

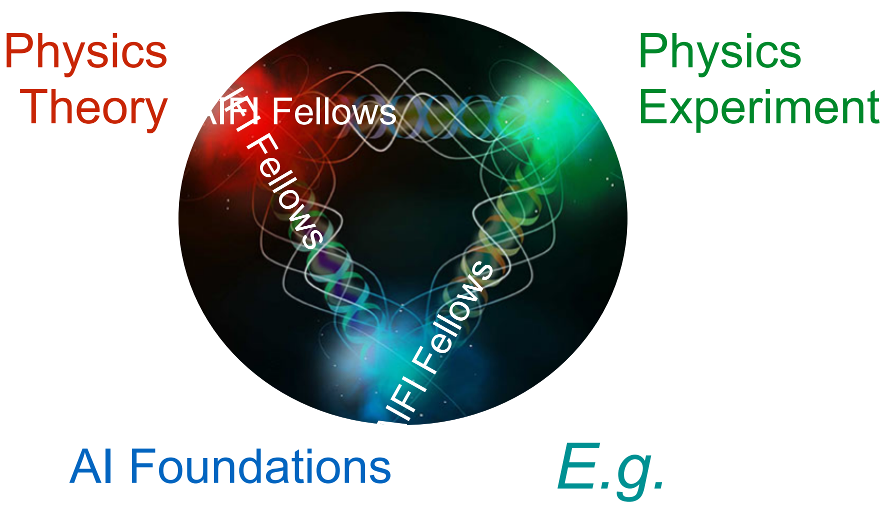
MIT		Harvard	Northeastern / Tufts
Pulkit Agrawal	Phiala Shanahan	Demba Ba	James Halverson
Lisa Barsotti	Tracy Slatyer	Edo Berger	Brent Nelson
Isaac Chuang	Marin Soljacic	Cora Dvorkin	
William Detmold	Justin Solomon	Daniel Eisenstein	
Bill Freeman	Washington Taylor	Doug Finkbeiner	
Philip Harris	Max Tegmark	Matthew Schwartz	
Kerstin Perez	Jesse Thaler	Yaron Singer	
Alexander Rakhlin	Mike Williams	Todd Zickler	Taritree Wongjirad

Boston Area: Critical Mass for Transformative Ab Initio AI Research

The NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)

"I-φ"

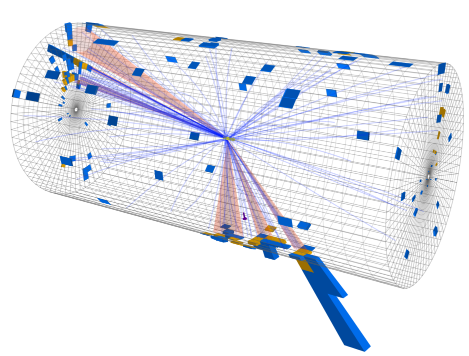
Advance physics knowledge — from the smallest building blocks of nature to the largest structures in the universe — and galvanize AI research innovation



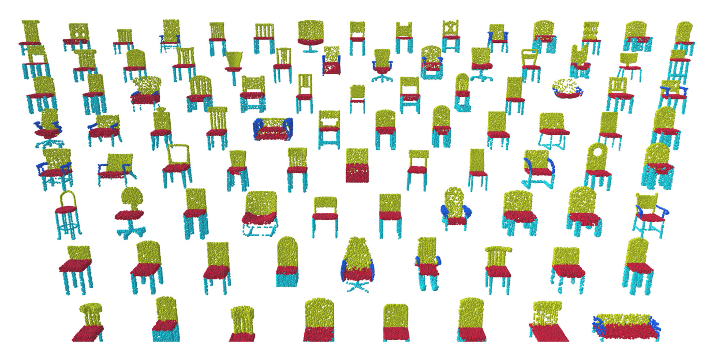
Training, education & outreach at Physics/AI intersection
Cultivate early-career talent (e.g. IAIFI Fellows)
Foster connections to physics facilities and industry
Build strong multidisciplinary collaborations
Advocacy for shared solutions across subfields

E.g.

Analyzing Collisions ⇔ Geometric Data Processing

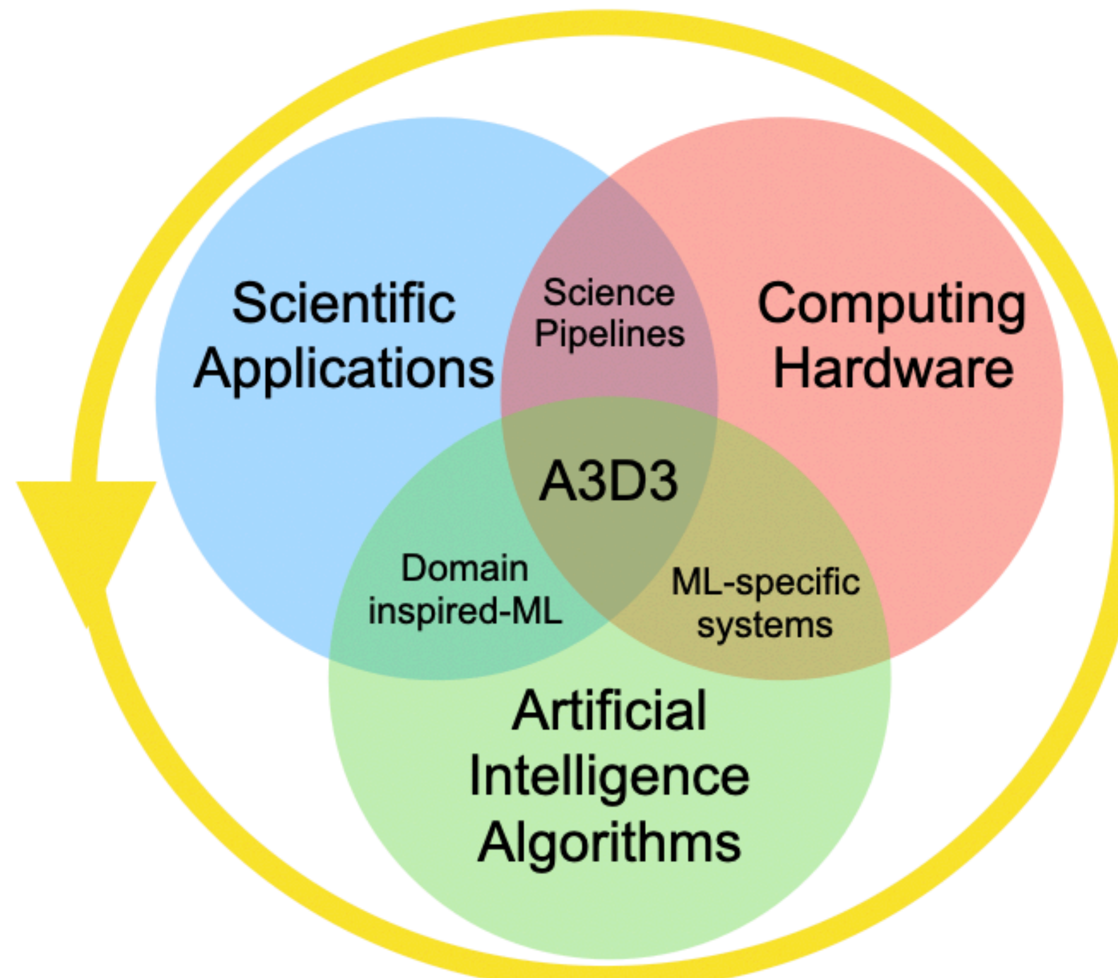


[**Harris**, **Schwartz**, **JDT**, **Williams**]



[Wang, Sun, Liu, Sarma, Bronstein, **Solomon**, TOG 2019]

Accelerated AI Algorithms for Data-Driven Discovery



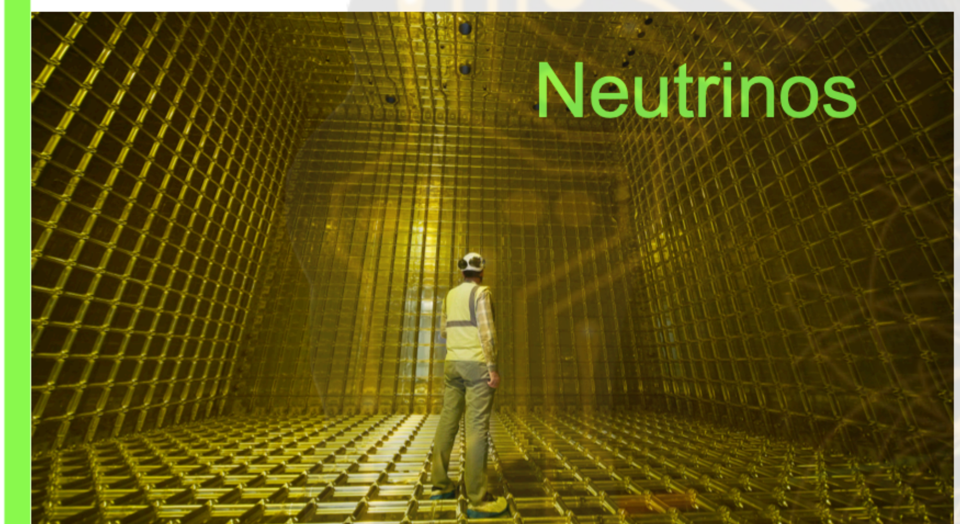
https://a3d3.ai/

https://news.mit.edu/2021/taming-data-deluge-1029

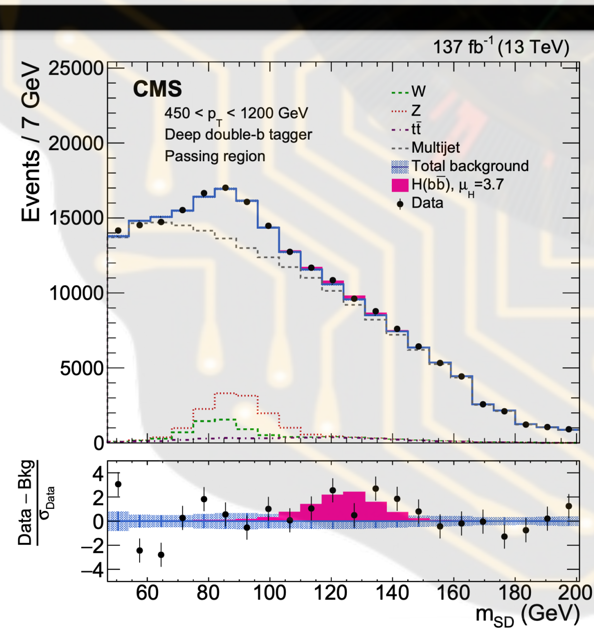
Real-time AI Institute: **A3D3**

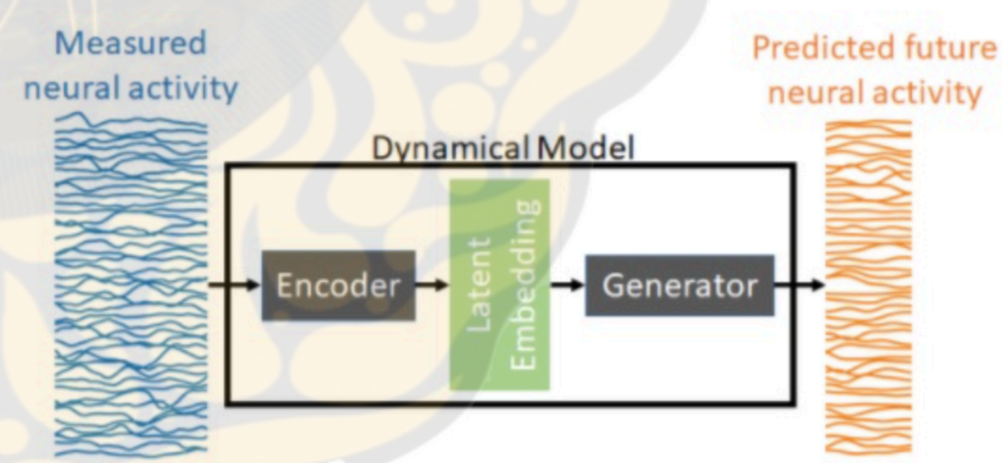
- We have been awarded a new institute to explore real-time AI
  - Accelerated AI Algorithms for Data Driven Discovery (A3D3)

New Types of Computing



LHC Physics





NeuroScience

Whats going on at MIT?

- As part of IAIFI and now A3D3:
  - Are exploring new ways to teach:
    - the overlap of physics and artificial intelligence
    - Broadly extends to all statistical analysis and physics
- **This is the test of how to do this**

Accelerated AI
Algorithms for
Data-Driven
Discovery

Getting Credit?

- This counts for a second breadth requirement for Grad students
  - All domains! Hooray!
- Also :
  - Statistics/Data Science Phd
    - Comp & Stat or Data science

Let me know if you have concerns

Accelerated AI Algorithms for Data-Driven Discovery

Online Component

- You will notice that the notes are done in great detail
  - This is because we are working to put this class online
  - Eventually, we want to host a version of this on edX
  - This class is fully availble in 3 parts on MITx
- Also means that you can follow this class asynchronously
  - Notes have problems that will guide you through lectures

Accelerated AI
Algorithms for
Data-Driven
Discovery

Who Am I?

Philip Harris

https://www.youtube.com/watch?v=UTXc-2agiUo

https://www.symmetrymagazine.org/article/october-2014/cern-people-series-tells-it-like-it-is

https://news.mit.edu/2021/taming-data-deluge-1029

Additional Support

Divij Sharma
TA

Alex Shvonski
Online Course Expert

Gaia Grosso
IAIFI Fellow
Guest Lecturer

We will have a few guest lectures later on in the class!

Communication

- We will communicate on Canvas Discussions:
  - https://canvas.mit.edu/courses/31223/discussion_topics
- Canvas is where everything is
  - Big announcements will be made on canvas
  - Day to day work will be on discusison forum
- If you need a fast answer ask me on slack (Philip Harris MIT)
  - If I don’t reply its because I am sleeping or eating
- For slow answers send me an email

Office Hours

- Phil will hold office hours **Thursdays** **from 3-4 pm**
  - **My office 24-502 (can do zoom too)**
    - https://mit.zoom.us/j/92904381940
  - You can ask questions about projects/recitations/lectures
  - Feel free to come by just to chat
  - Importantly you should get all set with everything
- Divij will hold office hours from **????**
  - It will be at one of the physics derpartment rooms in building 8

How I got into physics

https://www.urbandictionary.com/define.php?term=Data%20Junkie

- I have a serious problem
    - I am addicted to data analysis
- As a student I became obsessed with analyzing data
    - It would keep me up all night
- The most fun was building complex analyses
    - This class is a modern take on analyzing data
    - Now with all of the modern tools at hand

Problem Sets

- These are here to make sure you follow along
  - We would like your feedback about these
  - https://canvas.mit.edu/courses/31223/assignments/383726

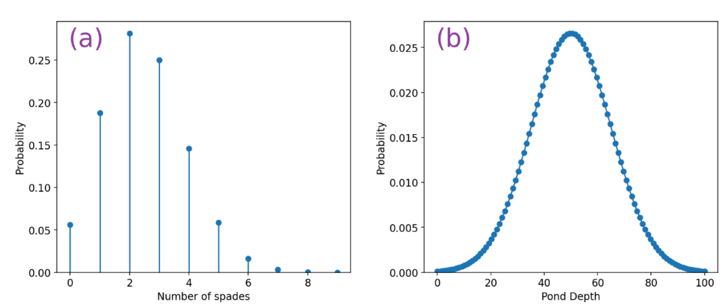
Pset 1

Problem 1.1.1

Which of the following distributions represents a pdf? Choose from the two plot options.

- The left plot (a) shows the probability of finding a certain number of spades when drawing 10 cards from a deck.
- The right plot (b) shows the probability of finding a pond with a particular depth among a selection of ponds.

After submitting your answer, you can look at the solution to see the code that generated these plots.



--

Save Submit View Answer

As staff, you are always allowed to submit. If you were a student, you would see the following:

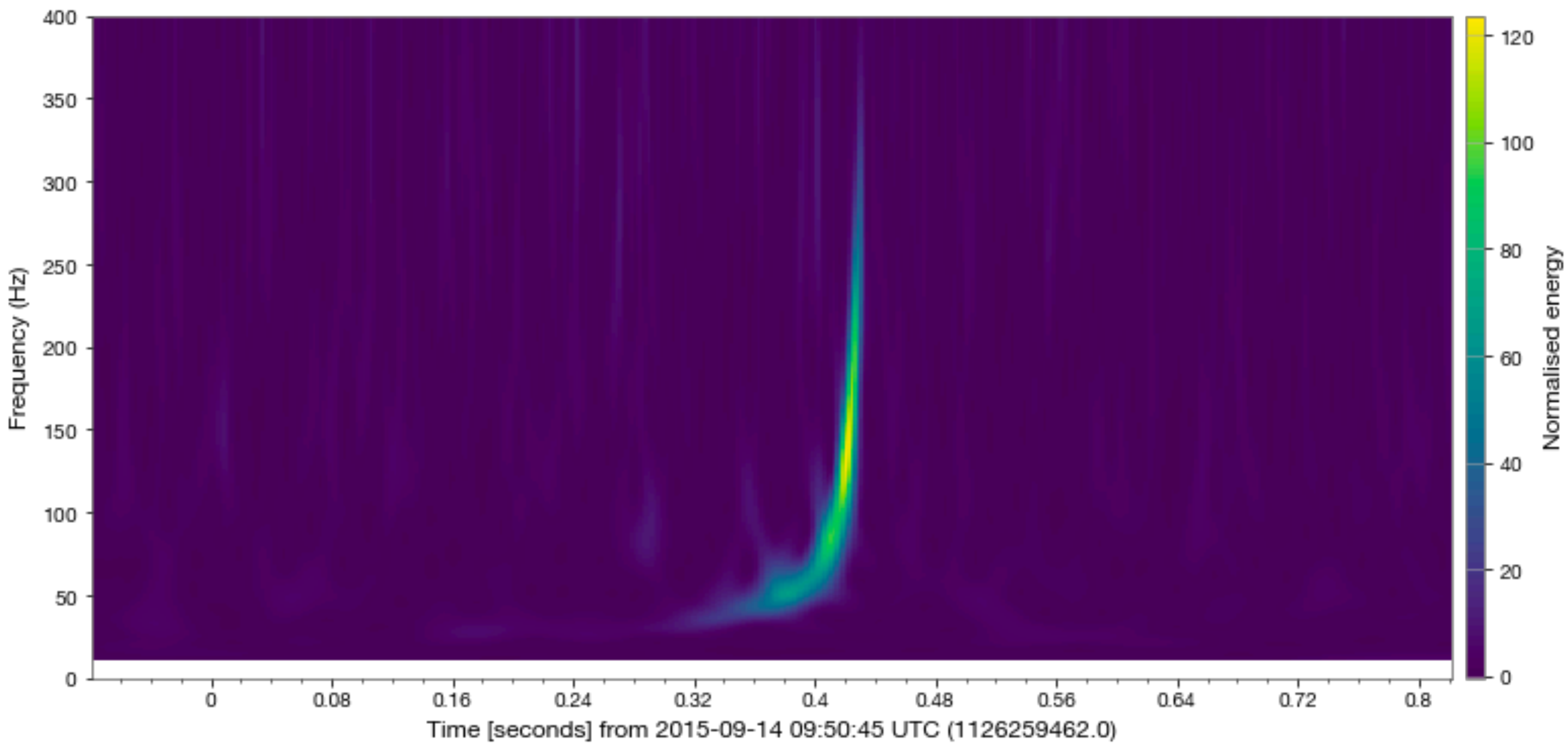
You have infinitely many submissions remaining.

Projects

- All of these projects are going to use public real data
    - This is **real** data from real experiments
- Some of the projects have led to important papers
    - All of them have led to papers published in the last 5 years
    - Some of these projects are still open to interpretation
- While the focus of this class is the data science behind it
    - We will talk about the physics implications of all of this

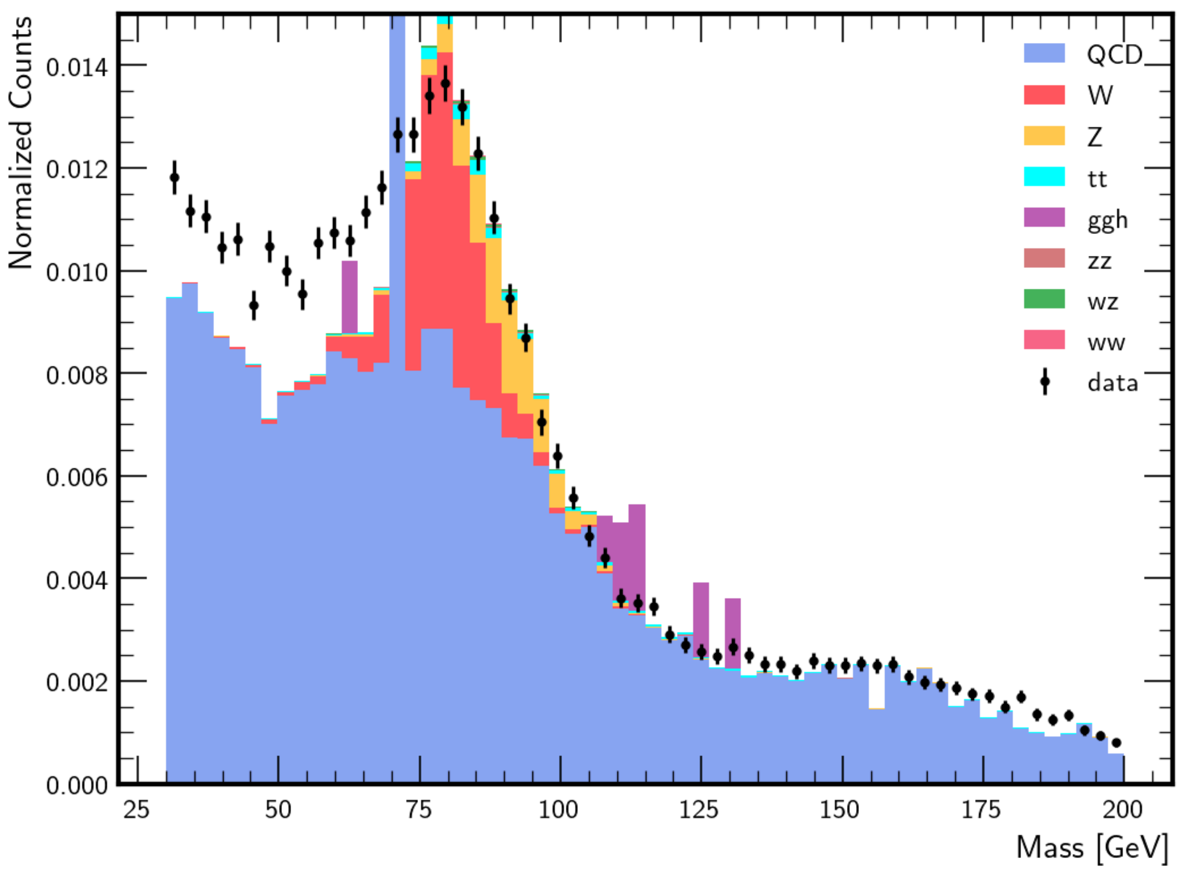
Project #1

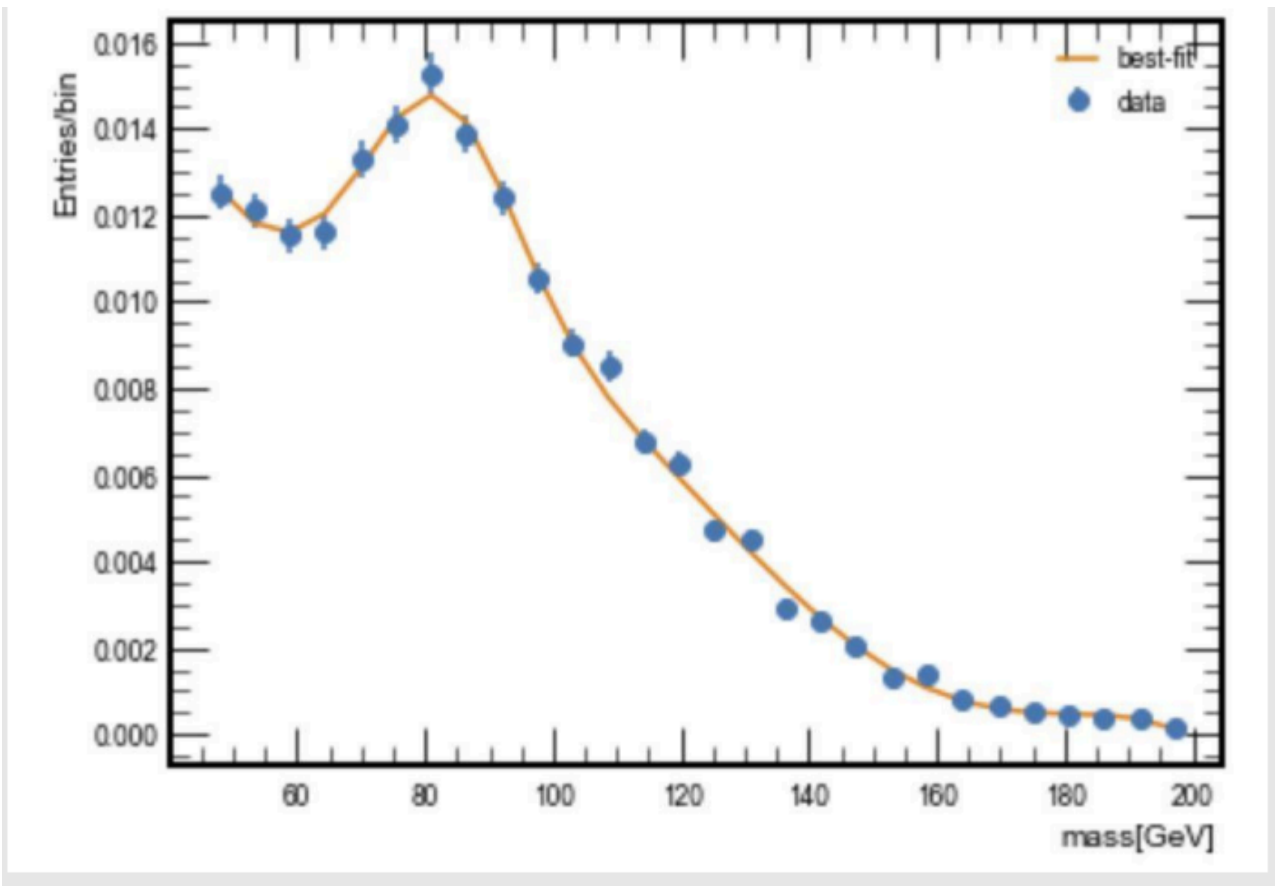
- Project 1 : Discovering Gravitational Waves



How do you discover a gravitational wave?
What are the parameters of the wave?

Project #2



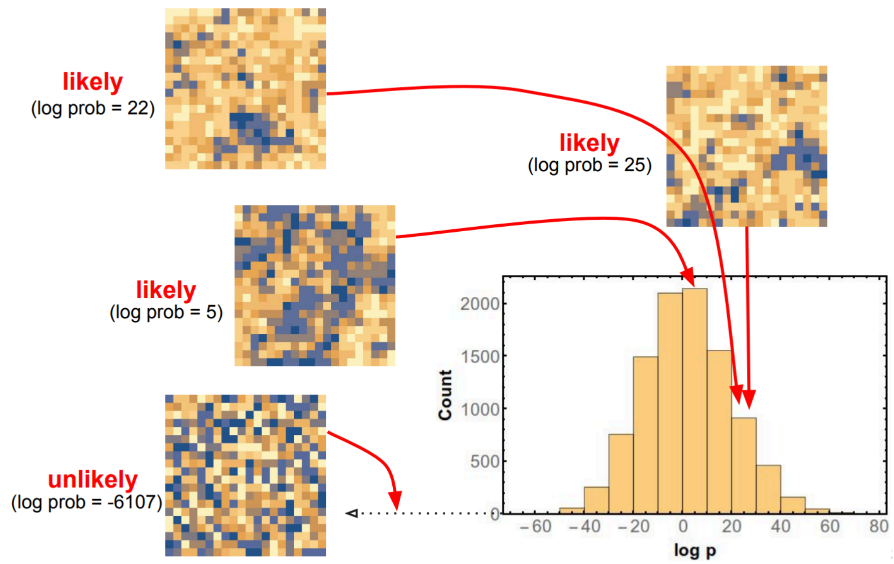


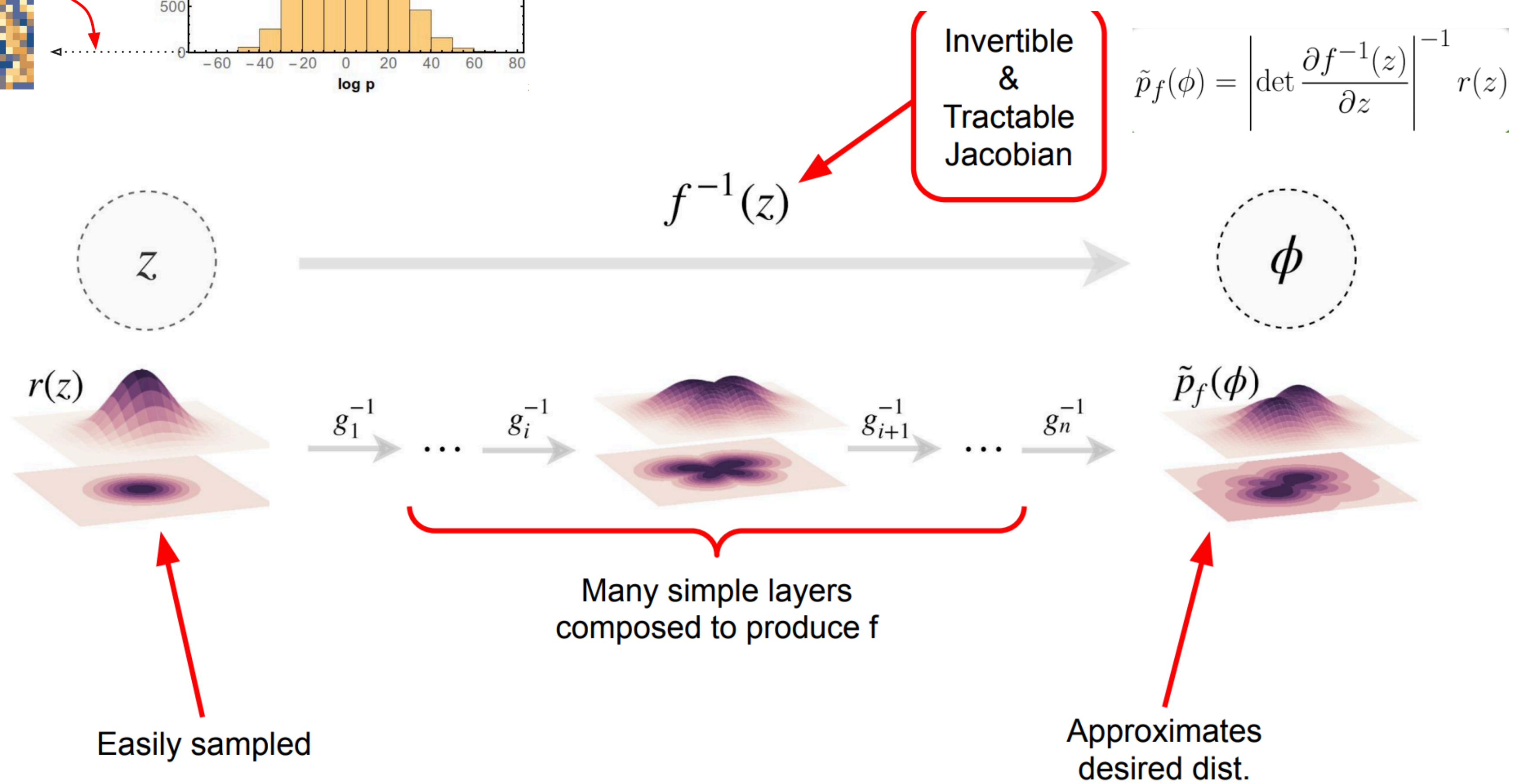
- Discover the W and Z boson decaying to quarks
- Try to enhance this measurement with deep learning

Project #3

Generate field configurations $\phi(x)$ with probability

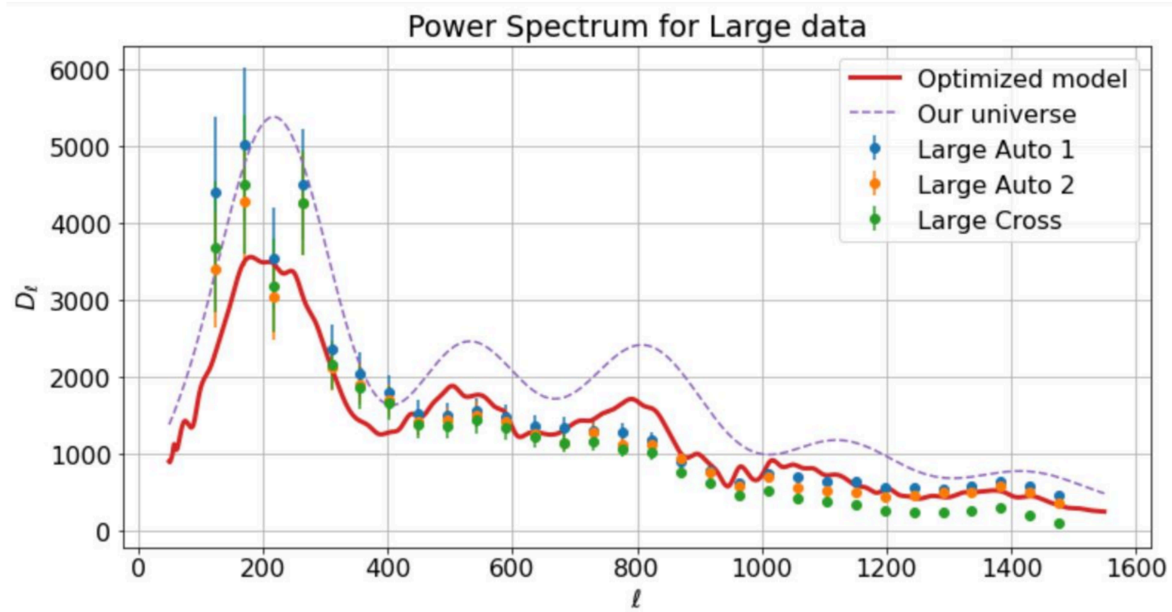
$$P[\phi(x)] \sim e^{-S[\phi(x)]}$$

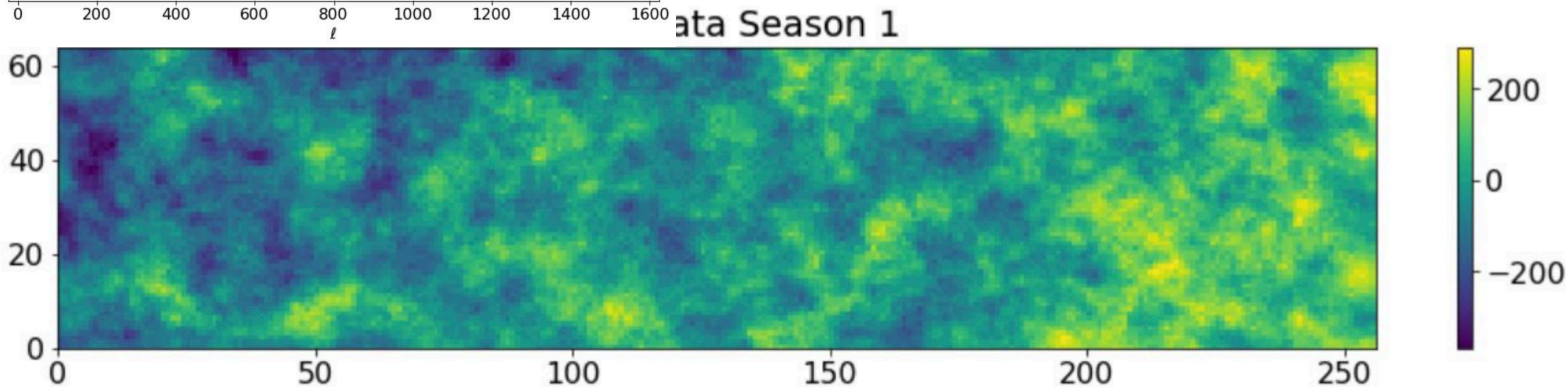


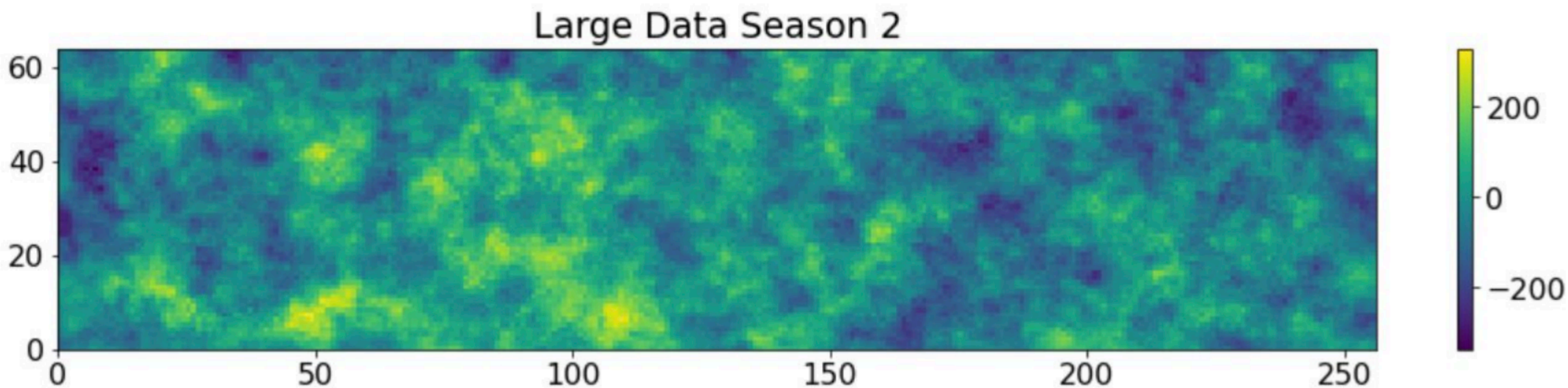


- Newer Project: Students like this project
  - Teaches the core of modern simulation

Project #4





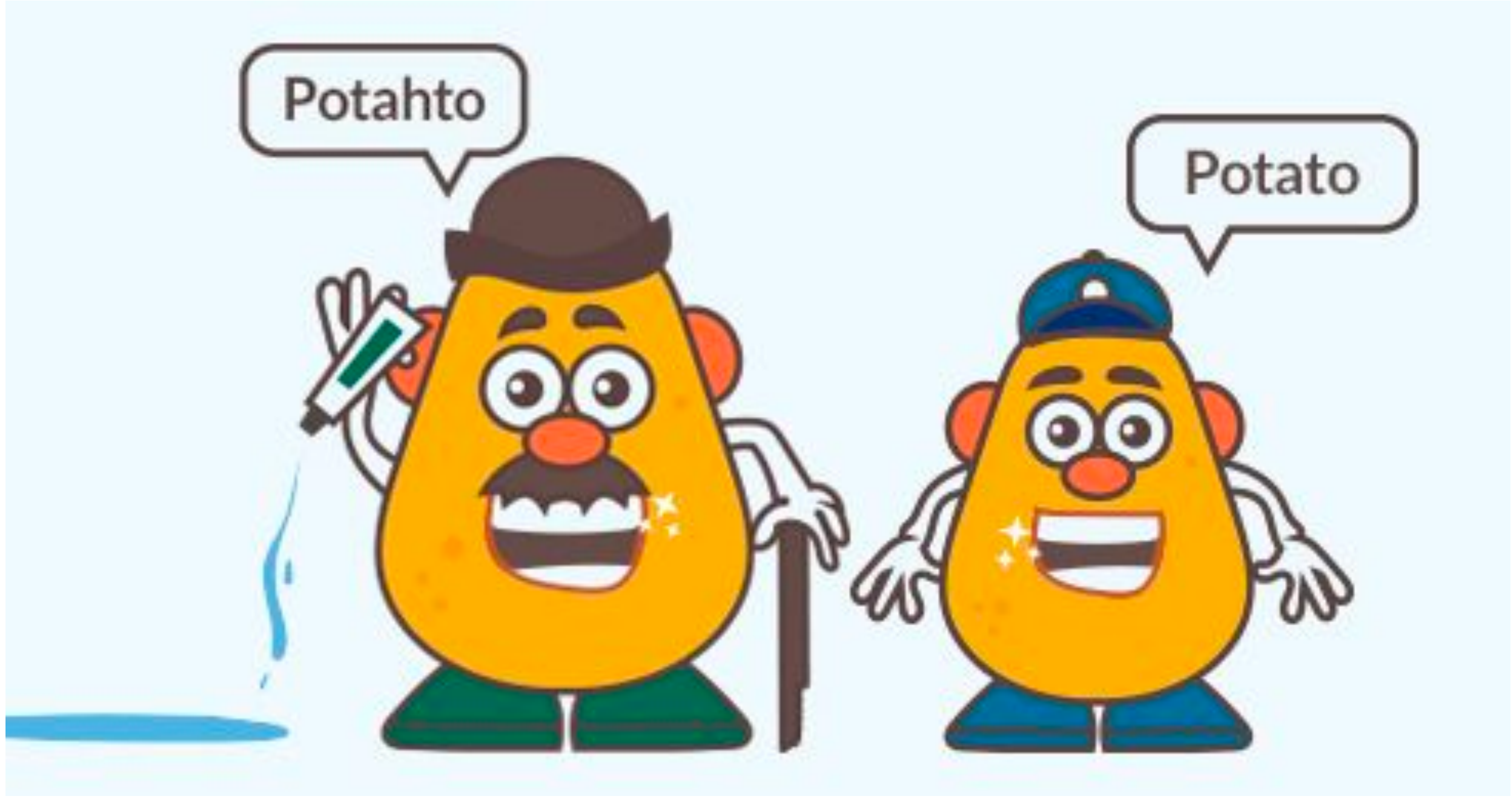


- How do we model the expansion of the universe

Project Reports

- We would like you to turn in a jupyter notebook
- Notebook should show some work
- We will grade the. Notebooks and give you feedback
- Here are some examples of what I should look like:
    - https://github.com/jmduarte/iaifi-summer-school/blob/main/book/2.1_getting_started.ipynb
    - https://github.com/FAIR4HEP/hbb_interaction_network/blob/main/notebooks/1.0-ar-Baseline-Model-Inspection.ipynb
    - https://github.com/FAIR4HEP/hbb_interaction_network/tree/main/notebooks

Project Solutions



- We will release worked out solutions
- Project Grading will be done by Me and the TAs
- But this is real data, there is no truly correct solution
- These projects can go on and on **(We don't expect that)**

Final Talk

- We have reserved the last two classes for you to present
- Thats a total of 180min
    - Depending on the enrollment we will divide up the time
    - Talks will be 10-ish minutes depending on enrollment
    - You can post your talk on youtube before
        - If you can't make it for whatever reason
- We would like to highlight a talk at an IAIFI meeting after class

Grading

- Focus of this class is on having fun
  - Don't stress about grades!
- 8% Pset 1,2,3,4
- 15% Project 1,2,3,4
- 8% Final Talk on project 4
- If you have concerns, please discuss with me
  - This is the first time we teach this
  - I want to make sure you are enjoying this class & Learning

Material Format

- We will use github to post the projects
  - https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners
  - Lectures will all be available on github as well
  - Assignments will be turned in as pdfs on canvas
- Github is the standard toolkit for data science projects
  - You will have to learn it at some point
- Assignments are due on Friday at Midnight!

Syllabus

Week 1: Basic Statistics, Project #1 Intro

Feb 2 Day 1: Class overview, Jupyter setup, making plots, Expectations, Variance
Feb 4 Day 2: Binomial, Poisson, Gaussian Distributions, Error propagation

Week 2: Distribution and Fitting Pset 1 due Friday Feb 13th

Feb 9 Day 1: LIGO Project
Feb 11 Day 2: Gradient Descent, Minimization, Introduction to Fitting
Feb 13 **Pset 1 due : Fitting data**

Week 3: Uncertainty and interpreting uncertainty Pset 2 due Friday Feb 20th

Feb 17 (Tuesday) Day 1: Extracting Uncertainty from a fit and goodness of fit
Feb 18 Day 2: Normal, confidence intervals, z-scores, non-gaussian distributions
Feb 20: **Pset #2 Due: Fourier analyses**

Week 4: **Project LHC Jet Physics Open Data analysis Project 1 due Friday, Feb 27th**

Feb 23 Day 1: Correlations/Covariance, Principle Component Analysis (Zoom)
Feb 25 Day 2: Introduction to jets and collider physics (Zoom)
Feb 27: **Project #1 Due: Fitting LIGO Gravitational Wave Data**

Syllabus

Week 5: Bayesian approaches

March 2 Day 1: Bayesian vs. Frequentist, Convolutions
March 4 Day 2: Hypothesis testing Intro

Week 6: Hypothesis Testing + Deep Learning Pset 3 due Friday March 15th

March 9 Day 1: Hypothesis testing II, f-tests/gaussian processes + semi-parametric metho
March 11 Day 2: Markov Chain Mone Carlo
March 13 **Pset #3 Due: Measuring Higgs properties with full likelihoods**

Week 7: Deep Learning Project 2 due Friday March 22nd

March 16 Day 1: Deep Learning Introduction
March 18 Day 2: Deep Learning Regression (zoom)

March 20 Project 2: **Measuring the Sinθ W with Jets**

Spring Break

Syllabus

Week 8: Intro to Simulation

March 31st Day 1: Finite Numerical Differential Equations and Integration Methods
April 2rd Day 2: Physics Informed NNs, Multi-body simulation (3-body problem)

Week 9: Numerical Simulation Strategies

April 7 Day 1: Introduction to Lattice QCD (tbc)
April 9 Day 2: tree methods, galaxy simulations, MC Methods

Week 10: Towards Lattice/Multi-body techniques Numerical Boltzman Sim Pset #4 Du

April 13 Day 1: Monte Carlo Methods
April 15 Day 2: Deep Learning Monte Carlo Methods
April 17 **Pset #4 Due :** N-body simulations

Week 11: Simulation Based Inference

April 20 Day 1: No Class (Patriots Day Holiday)
April 22 Day 2: Simulation Based Inference/Likelihood Free Inference
April 24 **Project #3 Due : Lattice QCD Project Due**

Syllabus

Week 12: Advanced deep learning techniques

April 27 Day 1: AI based anomaly detection
April 29 Day 2: Autodifferentiable methods (up for debate)

Week 13: Advanced parameter estimation techniques

May 4 Day 1 : Generative Models (Diffusion/…) (up for deba
May 6 Day 2: Final Presentations

Week 14:

May 11 Final Presentations on this last day
May 16 **Project #4 Due: Extracting CMB Parameters from simulated Cosmic Data project of choice on previous topics chosen for the last class**

https://docs.google.com/document/d/1ezQSL1WHEvK7rBZyo6mUnUmjsrfdSEWu4Zp551Q1CPo/edit?usp=sharing

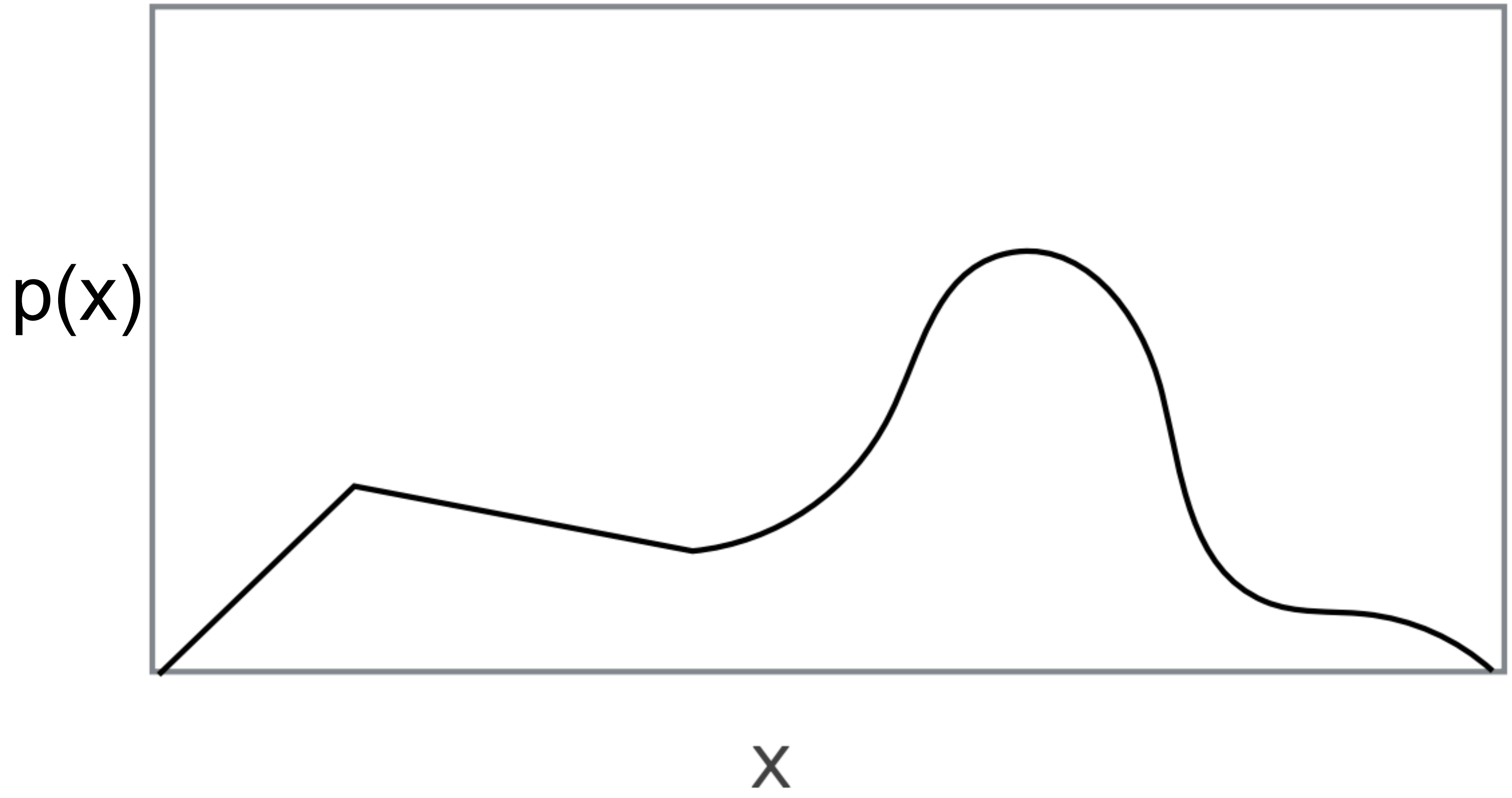
Disclaimer

- Please be mindful of the fact that this relative
- We are going to solicit feedback
- Your input is critical to making this happen!
- There are similar classes at other universities
  - This is a new, creative, take on such a class
  - Thinking about writing a textbook based on this!
- Goal here is to really take advantage of recent developments
- Your feedback is crucial to making this a great class

Lets Make a Plot

- Before we go into details of this lets make a plot
- All of our notes are documented on github
  - https://github.com/mit-physics-data/lectures
  - https://github.com/mit-physics-data/psets
  - https://github.com/mit-physics-data/projects
- We are going to use jupyter and python for this class
  - Warmup:
    - https://github.com/mit-physics-data/psets/tree/main/pset0
  - https://www.dropbox.com/s/v6xk9z11vnp49jf/8.012%20Intro%20to%20Coding.ipynb?dl=0

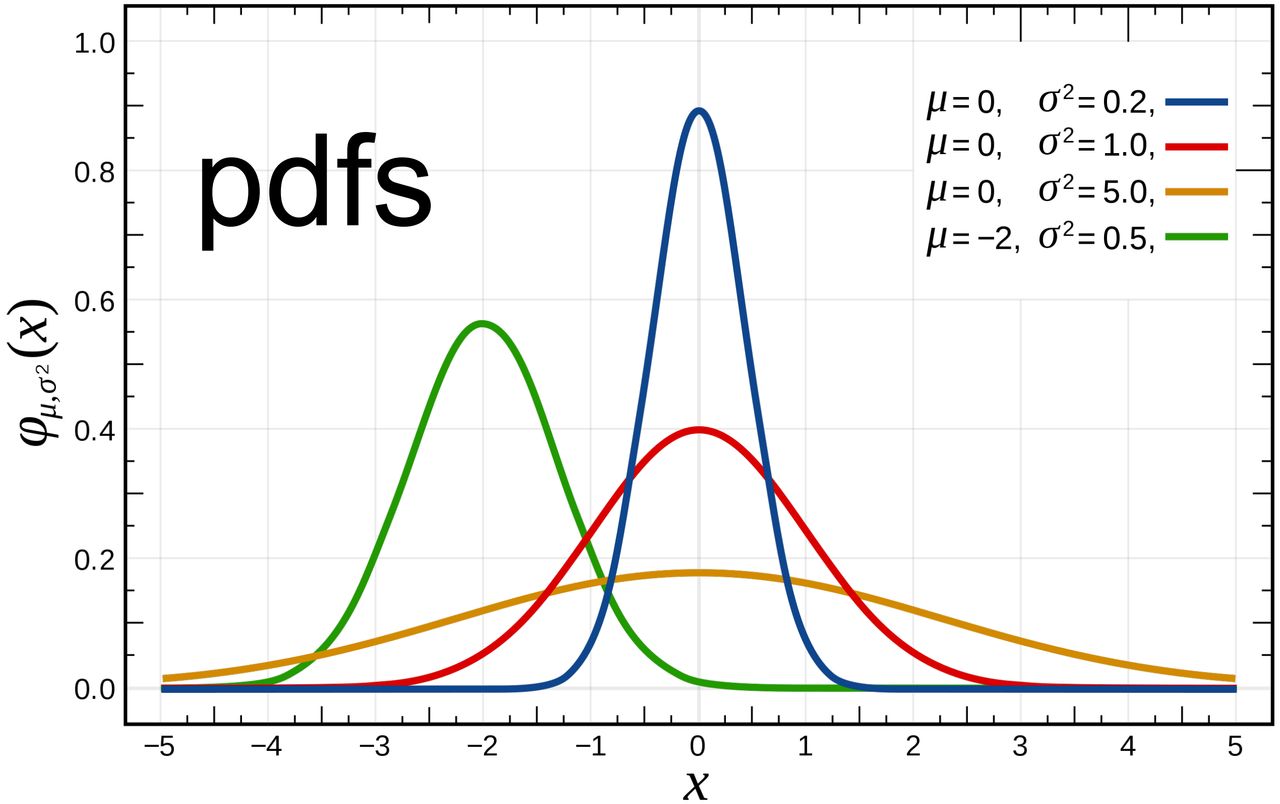
PDFs

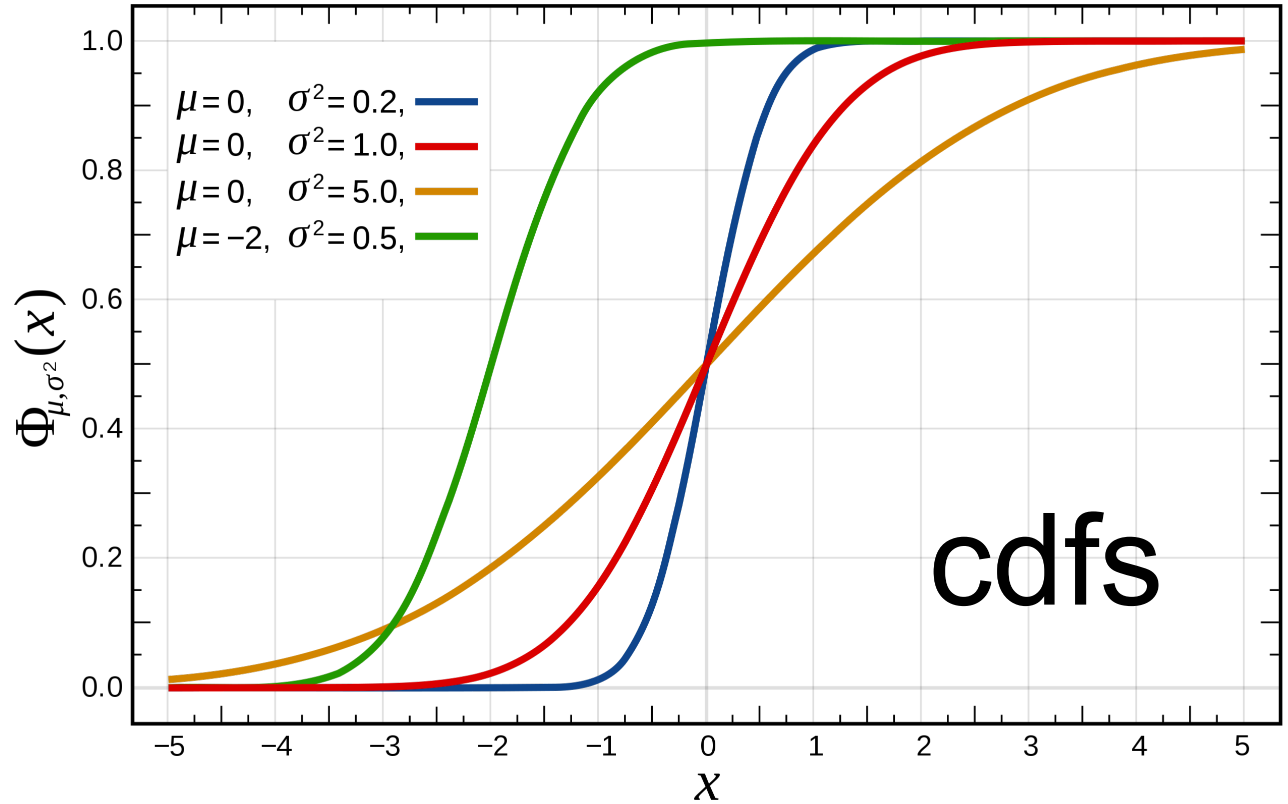


- Probability distribution(density) function *p(x)* sometimes *f(x)*
  - Probability of being between *x* and *x+dx*
  - $P(x \in [x, x + dx]) = p(x)dx$
  - $P(x \in [a, b]) = \int_a^b p(x)dx$

Probability can be disjoint

CDFs





- Cumulative distribution(density) functions or sometime CDFs
- $$\mathrm{cdf}(\mathrm{p}(\mathrm{x}), \mathrm{a}) = \int_{-\infty}^{\mathrm{a}} \mathrm{p}(\mathrm{x})\mathrm{dx}$$

Expectation

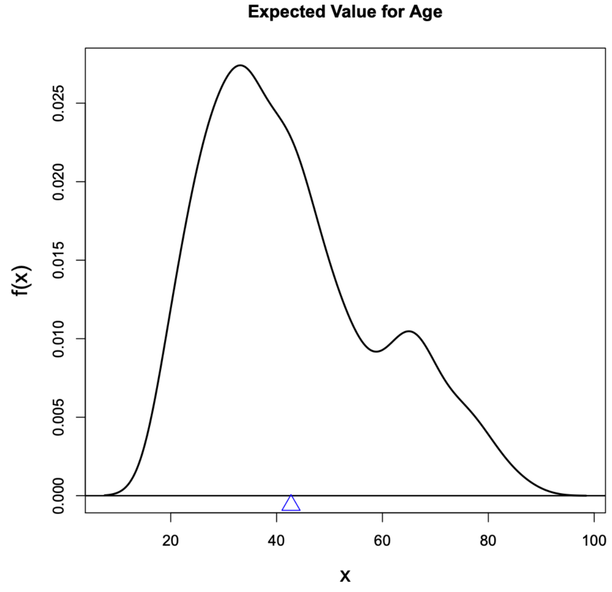
The expected value of a random variable X is denoted by $E[X]$ and is a measure of central tendency of X. Roughly speaking, an expected value is like a weighted average (weighted by probability of occurrence).

The expected value of a discrete random variable X is defined as

$$E[X] = \sum_{\text{all } x} x \cdot f_X(x).$$

The expected value of a continuous random variable X is defined as

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx.$$



Expectation Balance Point of a distribution

Expectation

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{x} = \sum_{\text{all } x_i} x_i \cdot f(x_i), \text{ where } f(x_i) = \frac{1}{N}$$

$$E[b] = b$$

$$E[aX] = aE[X]$$

$$E[aX + b] = aE[X] + b$$

$$E\left[\sum_{i=1}^{k} X_i\right] = E[X_1] + \cdots + E[X_k]$$

Variance

The expected value of a function $g()$ of the random variable X, written $g(X)$, is denoted by $E[g(X)]$ and is a measure of central tendency of $g(X)$.

The variance is a special case of this, and the variance of a random variable X (a measure of its dispersion) is given by

$$V[X] = E[(X - E[X])^2]$$

It is the expectation of the squared distances from the mean.

Variance is a measure of the width of our distribution

Variance

For a discrete random variable X

$$V[X] = \sum_{\text{all } x} (x - E[X])^2 f_X(x)$$

For a continuous random variable X

$$V[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx$$

Suppose a and b are constants and X is a random variable. Then

$$V[b] = 0$$

$$V[aX] = a^2 V[X]$$

$$V[aX + b] = a^2 V[X] + 0$$

Variance

Suppose we have k independent random variables $X_1, \ldots, X_k$. If $V[X_i]$ exists for all $i = 1, \ldots, k$, then

$$V\left[\sum_{i=1}^{k} X_i\right] = V[X_1] + \cdots + V[X_k]$$

Standard Deviation is defined as $\sigma = \sqrt{V[X_i \ldots .]}$

It is a measure of the width of a distribution

Label standard deviation to imply that we have chosen this for our uncertainty

Standard deviation is what we often use for **uncertainty**

Questions?