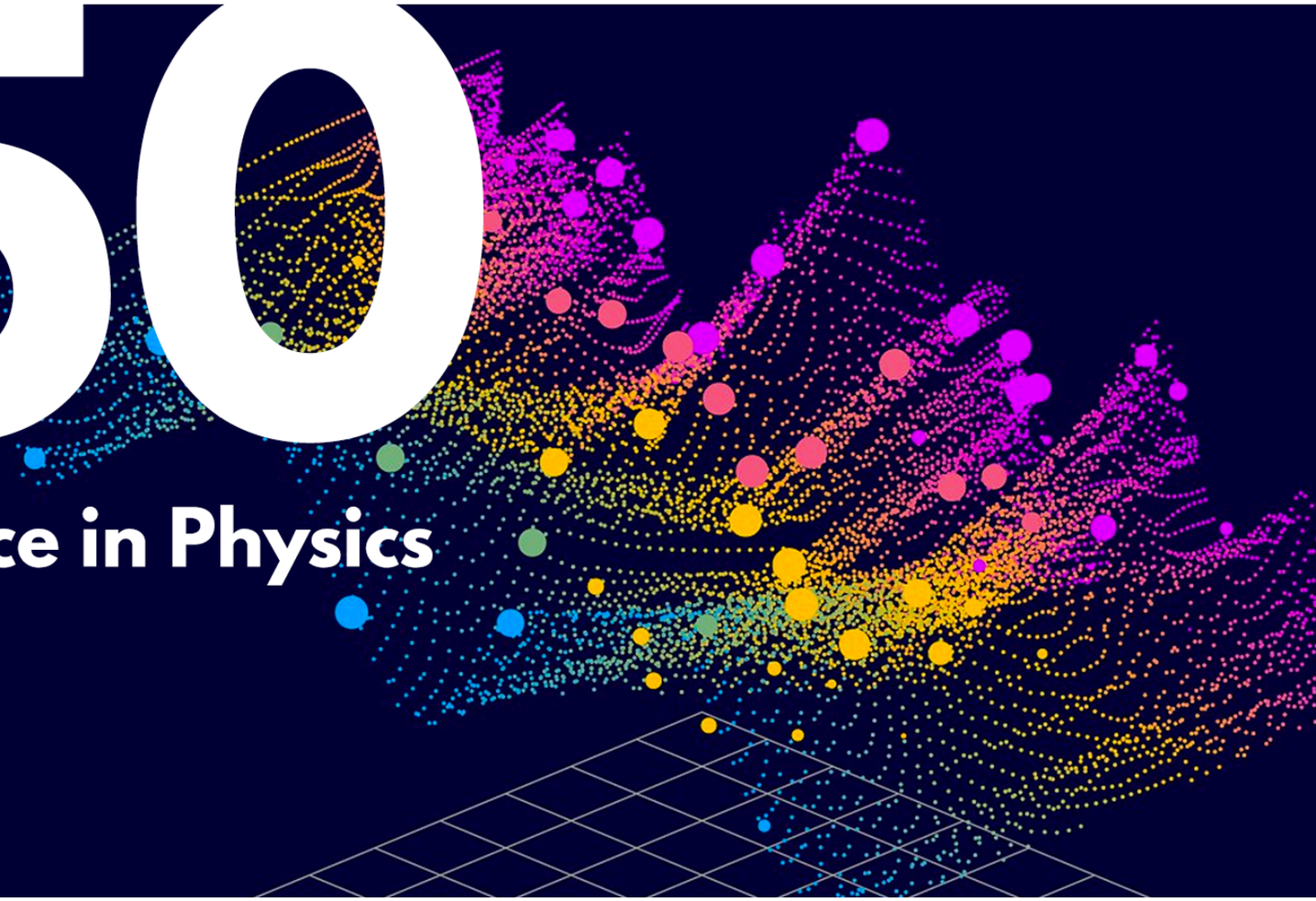
8.S50

Computational Data Science in Physics

IAP 2021
M–F 1-2:30pm



P. Harris, J. Parra, A. Shvonski
J. Dinsmore, A. Chambers, B. Ryan, N. Palladino
O. Kitouni, T. Nguyen,D. Hoang, E. Moreno, S. Mishra-Sharma, A. Nisimara


2022
HAPPY NEW YEAR

Community Values

- We are a strong supporter of the community values
- 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
    - Go over the general statistical tools to analyze data
- We are not going to go into great statistical detail
  - 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 8.03 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 and Expectation, and probability distributions
  - Expect some level of rigor

Cover, but knowing this will help

What you learn

Course description

Computational methods are a critical component to many fields of physics research. With the rise of deep learning and the increased access to large-scale computational facilities, the impact of computation has become increasingly more important. This class aims to present modern computational methods by providing realistic examples of how these computational methods apply to physics research. Topics include: Poisson statistics, error propagation, Fitting, Data analysis statistical measures, Hypothesis testing, Semi-parameter fitting, Deep learning, Monte Carlo Simulation techniques, Markov Chain Monte Carlo, Numerical Differential equations Class is python-based, and will rely heavily on Jupyter notebooks.

Recommended Prerequisite: 8.03/(6.0001+2 or some python) (Can be taken with 8.20)
Class is intended for Sophomore or up

- We are aiming for a one size fits all class
- Difficulty level: Sophomore →Professor

Class Overview

- Class is going to be a project based class
- You will have 3 projects:
    - Projects are on Jupyter notebooks
    - You will have 1 week for each project
- In the last week:
    - You pick a previous project, work on it more
    - You will present your project in the last classes

Real Data
means
Real Problems

Real Data
also means
Real Reseach!

Material

Projects utilize real Data

LIGO

Project 1 :
Gravitational Wave Data
From LIGO

CMS

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

chime

Project 3 :
Cosmic Microwave Background
(simulated) Data

What you learn

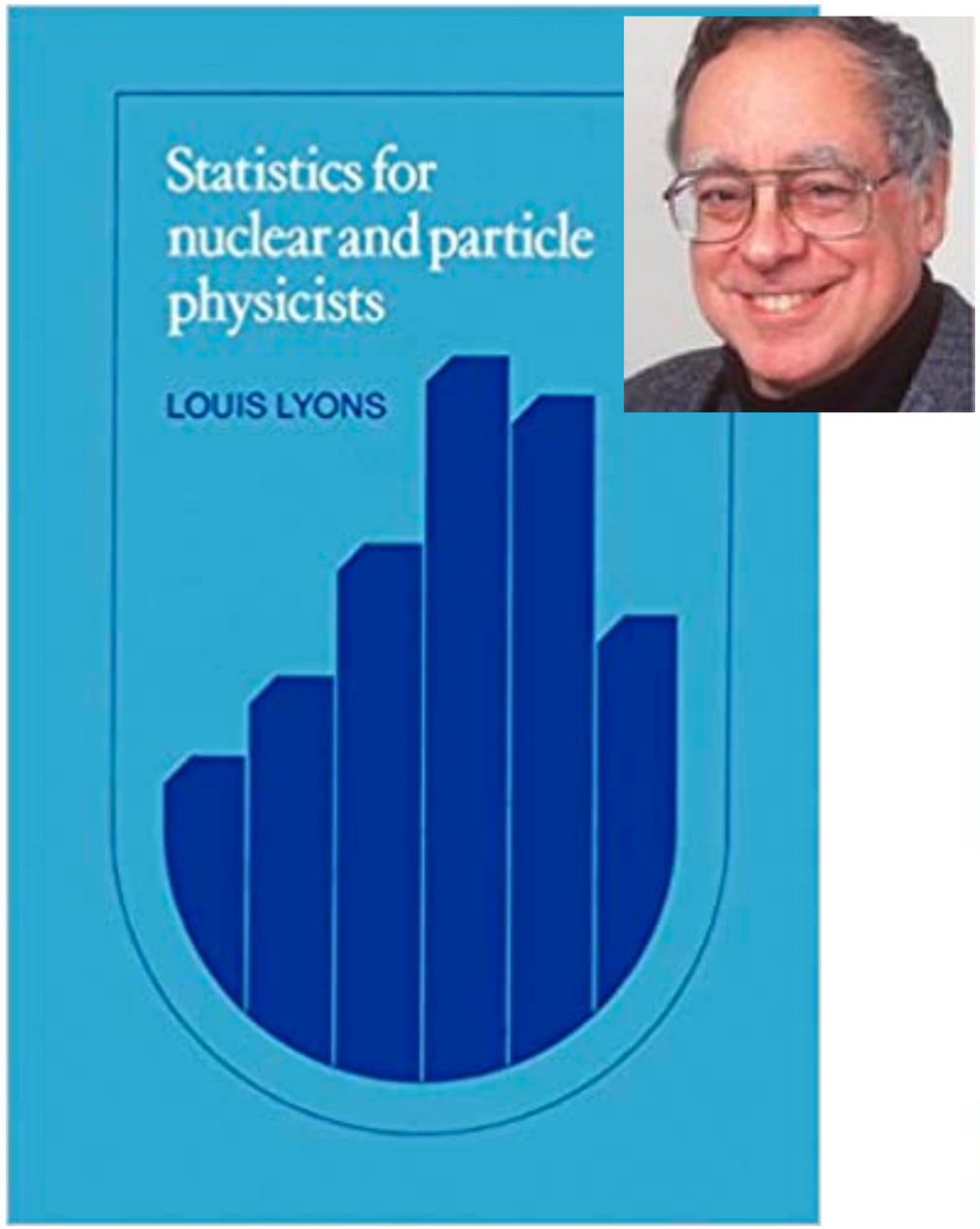
- Aim of this class :
  - Teach you how to use numerical tools to solve physics problems
  - To provide you with numerical tools at Junior Lab level
    - The same level needed for pretty much any Physics UROP
    - Expose you to real data from experiments
      - You can write real, and important papers based on this
    - Expose you to “Data Science” thinking common in physics

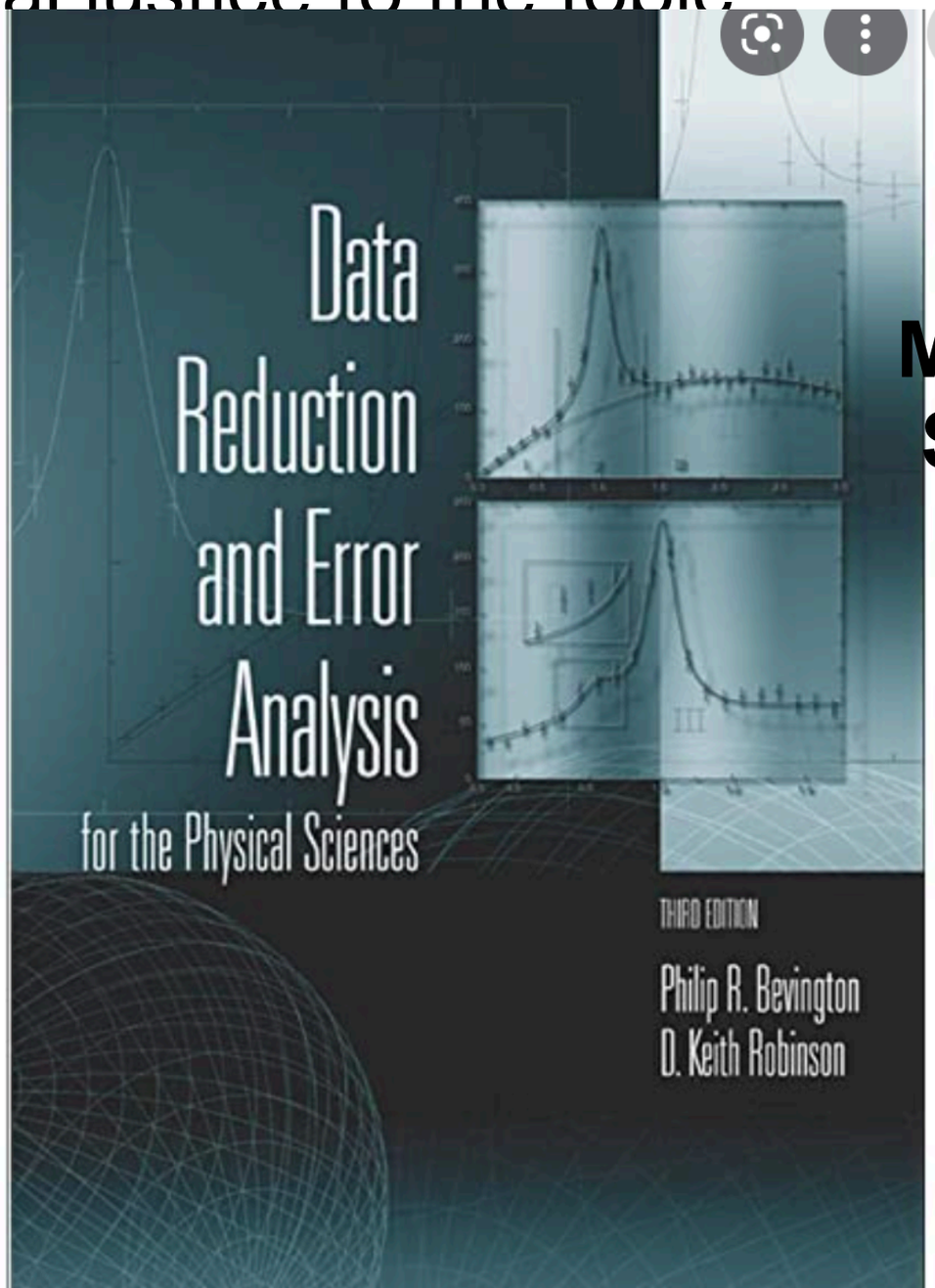
Grading

- Grading for this class will be :
  - A-F grading
- How do you pass the class?
  - Turn in something for each project
    - Give a talk at the end
    - Turn in notes form recitations
    - This class is ***for you to learn***
      - Don’t **stress** about the grades
      - If you have concerns come to my office hours

Textbook

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





MIT Junior Lab Statistics text

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)

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

Workload

Projects use real world data and simulations

You can write scientific papers with this data

- How much effort should you put in?
  - $X = \frac{\text{hours per day}}{24h}, 0 < X < 1$
  - Amount is is up to you
    - You could put 24 hours per day in it if you wanted to
    - Projects are easy to make progress, but nearly impossible to finish
    - For this class, some effort is good 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)

Class Format

- Lectures 1:00-2:30pm on zoom
- Recitations 3:00-4:00pm on zoom
- Class will be full remote
  - You can take this class asynchronously
    - No requirement to attend lectures/recitations in real-time
  - Participation is 15% of grade
    - Participation just means you turn in recitation notes
      - Ideally you go through the recitation exercises

Office Hours

- We will hold office hours everyday from 4-5pm
  - You can ask questions about projects/recitations/lectures
  - Feel free to come by any day
  - Today, we can help ensure you have the software installed

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

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

Data Science Mistakes

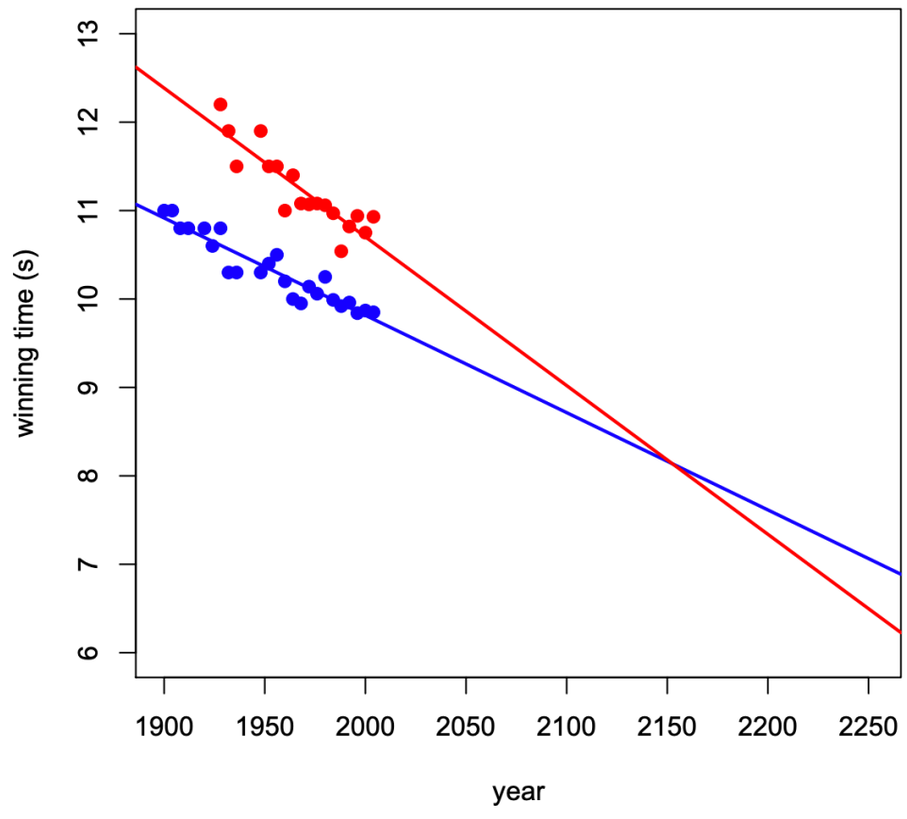
https://hackernoon.com/12-mistakes-that-data-scientists-make-and-how-to-avoid-them-2ddb26665c2d

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

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.

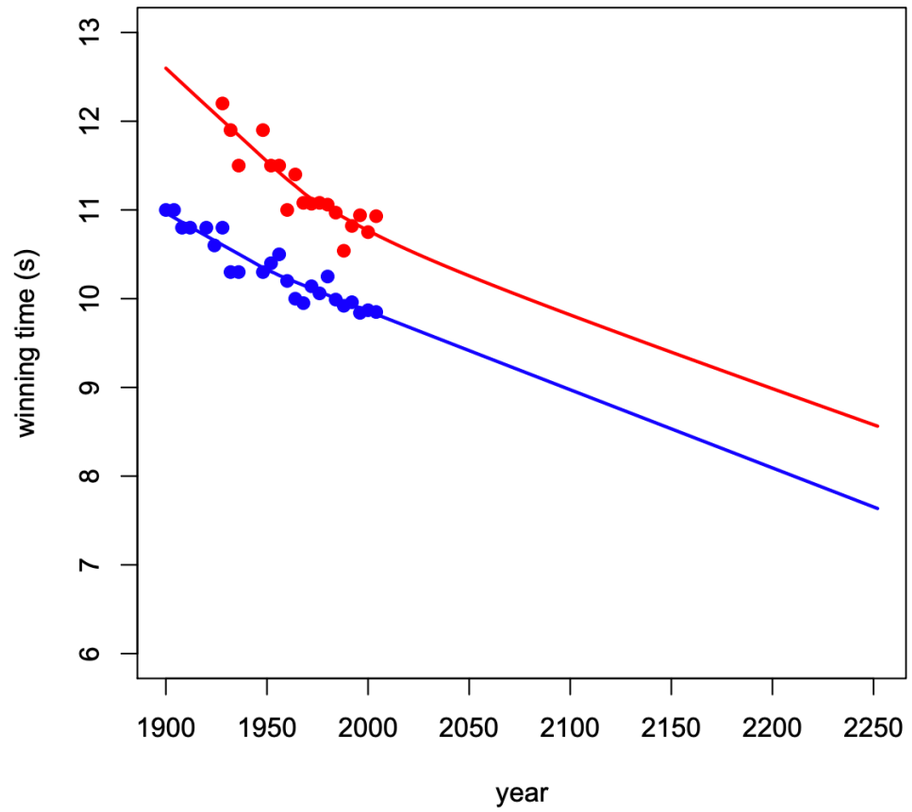


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

- 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.

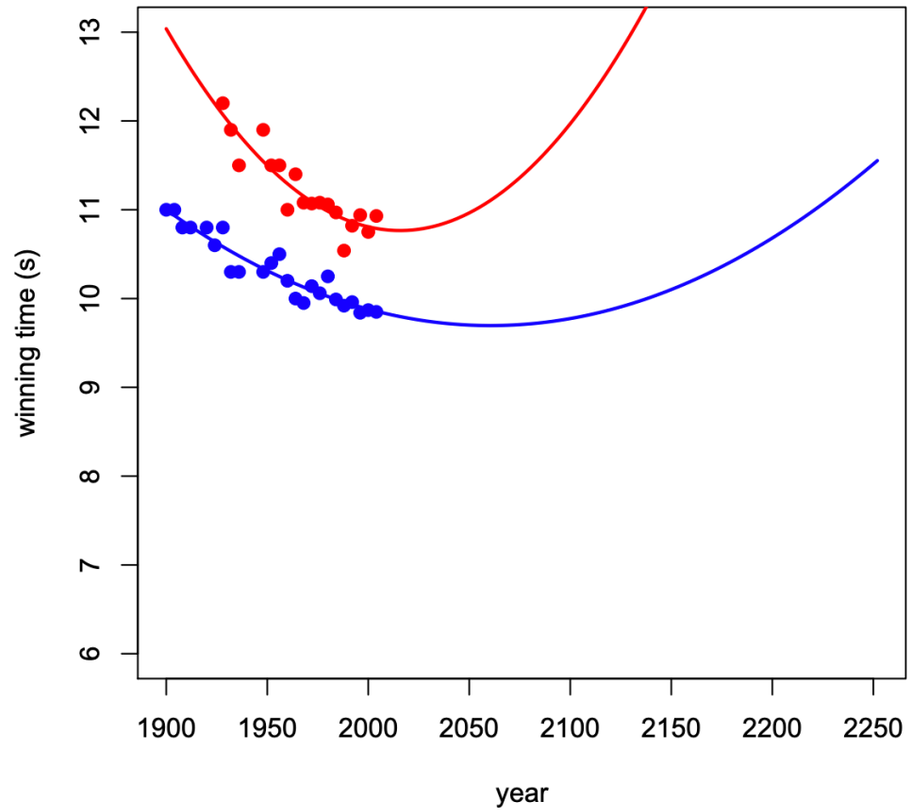


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

- 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.

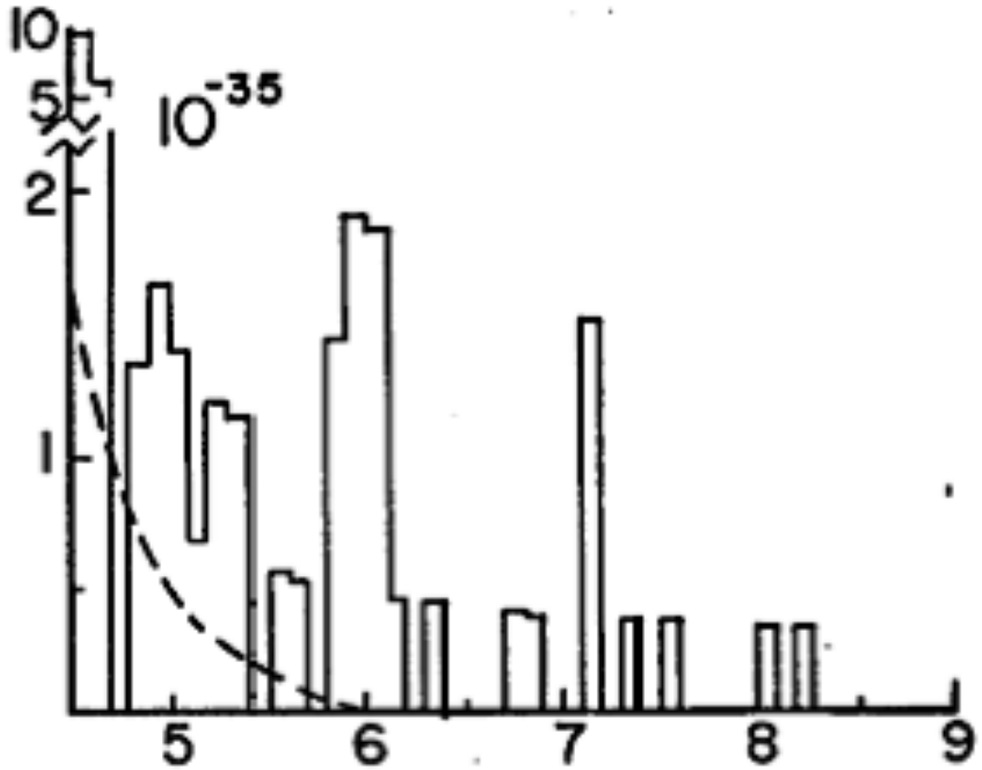
You can't have data science without science

- Choice of model requires some intuition within the field

Physics Blunders

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

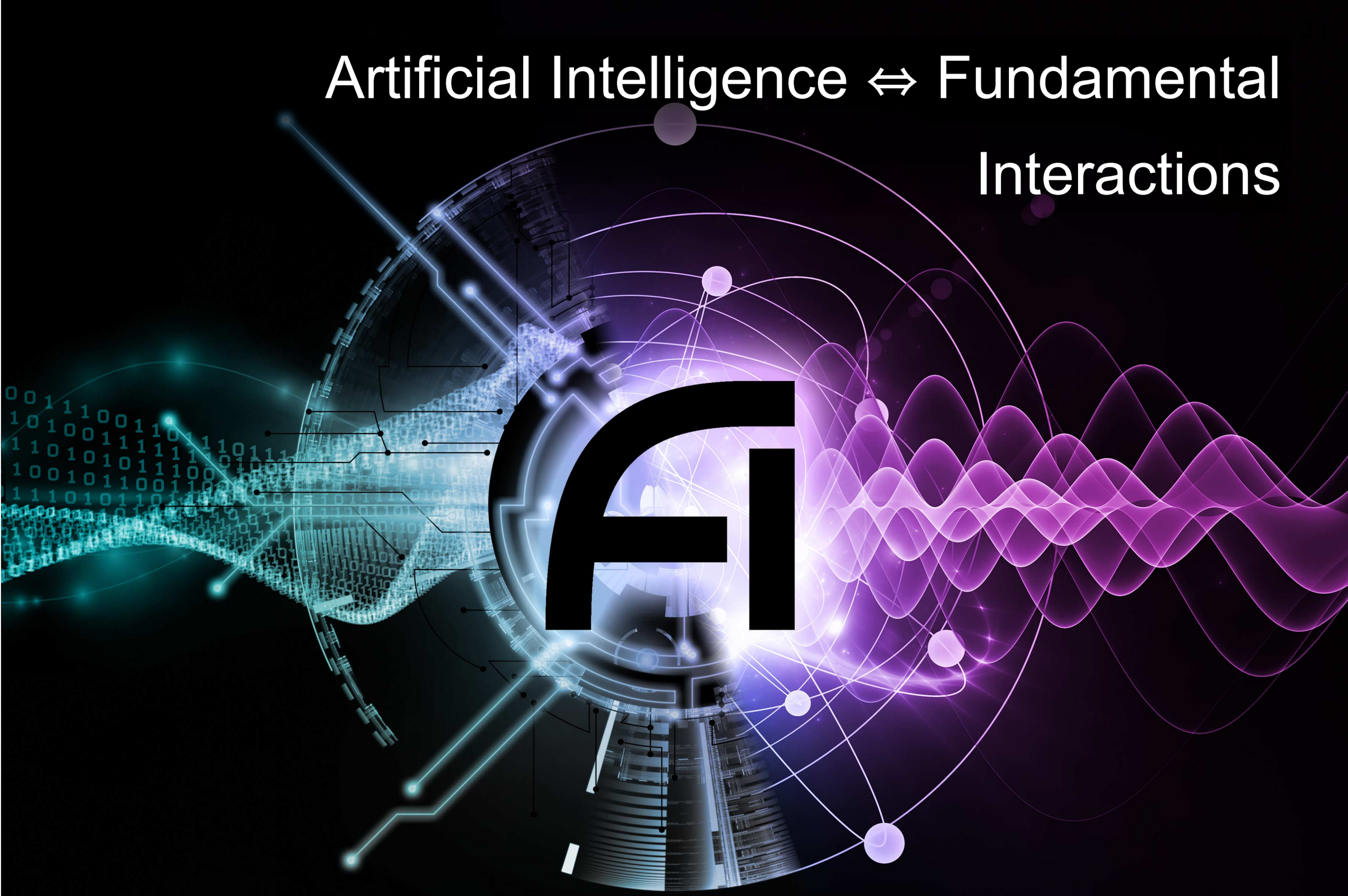
- Cover the statistical tools
- For you to understand integrity



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

Hope is to build intuition

Artificial Intelligence ⇔ Fundamental Interactions



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

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

"I-φ"

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

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

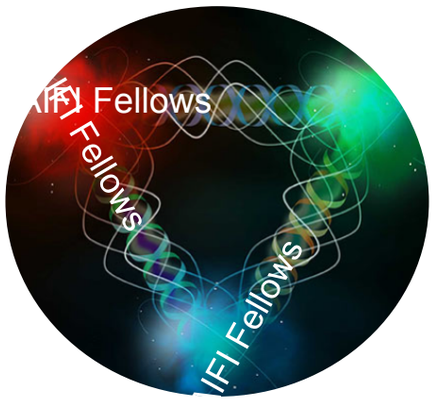
Boston Area: Critical Mass for Transformative Ab Initio AI Research

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

Physics Theory



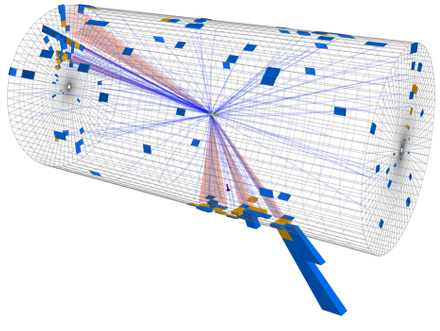
Physics Experiment

AI Foundations

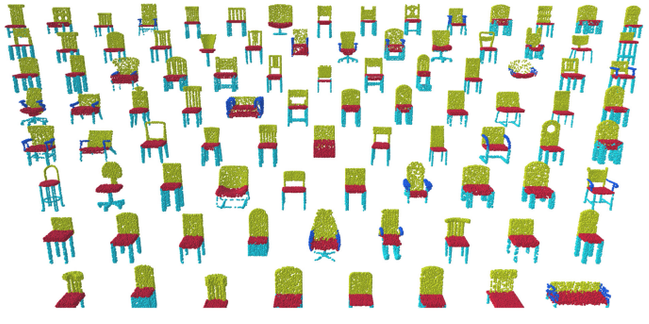
E.g.

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

Analyzing Collisions ⇔ Geometric Data Processing

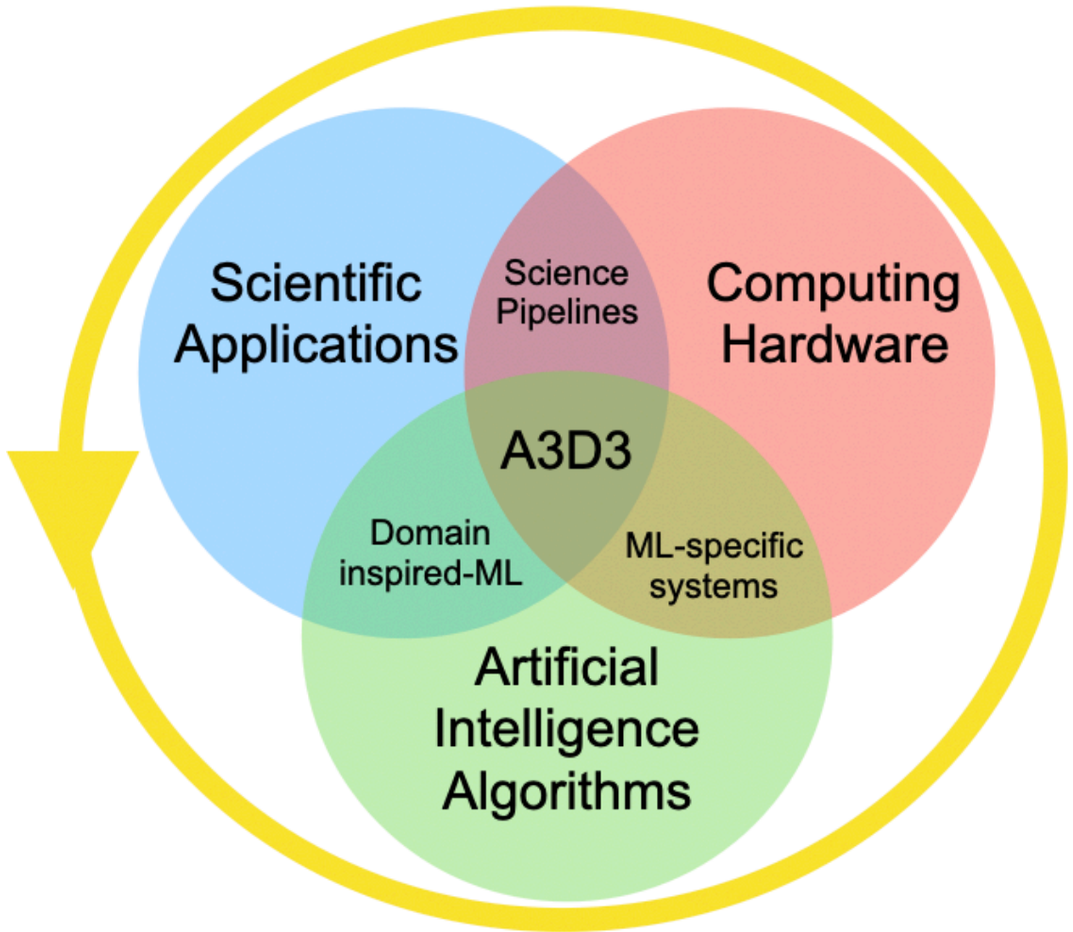


[Harris, Schwartz, JDT, Williams]



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

Accelerated AI Algorithms for Data-Driven Discovery



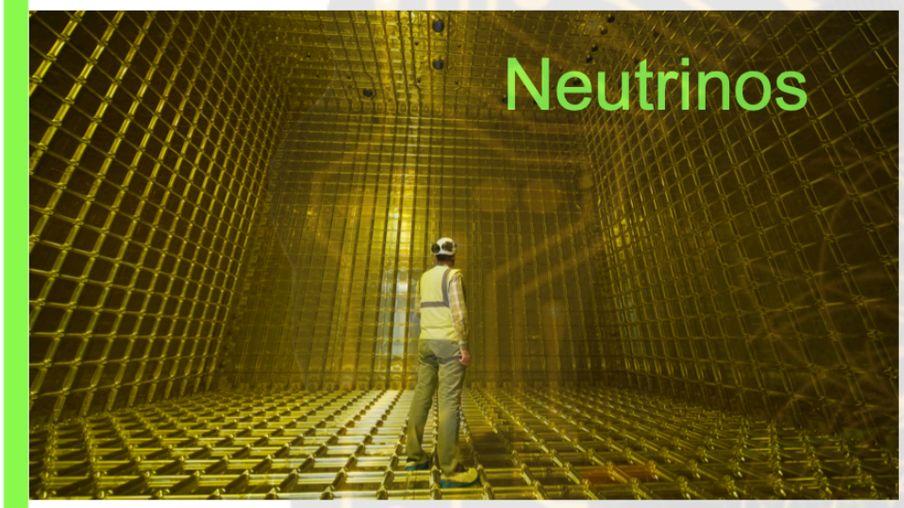
https://a3d3.ai/

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

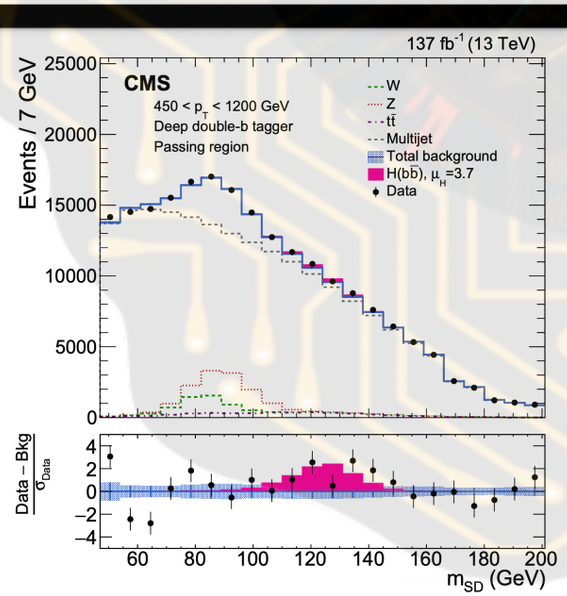
A New Institute: **A3D3**

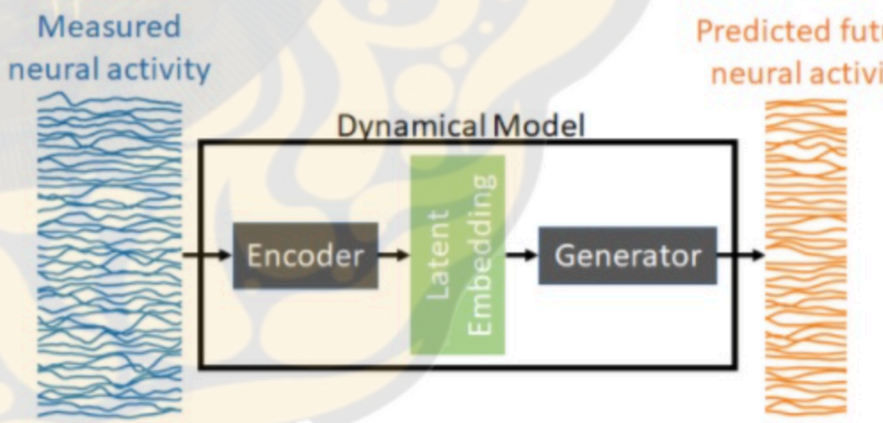
- We have been awarded a new institute to explore real-time AI
  - Accelerated AI Algorithms for Data Driven Discovery (A3D3)

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 a first tester class of how to do this**
  - Hoping to build this class into a full semester class

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
- We will record the lectures
- Also means that you can follow this class asynchronously
  - Everything can be followed on your own time

Who Are We?

Philip Harris

Juan Parra

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/2019/revolutionary-radio-telescope-detects-bevy-fast-radio-bursts-0109

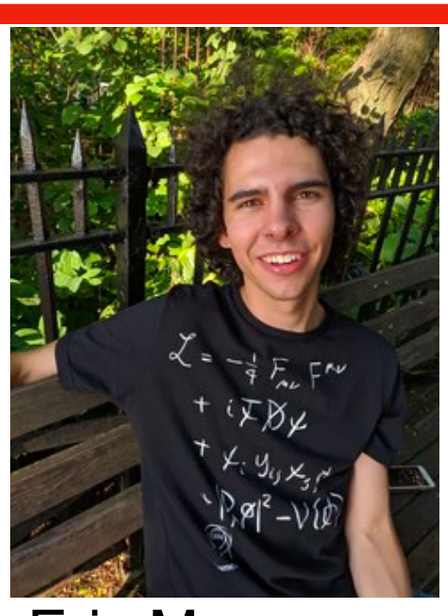
Additional Lectures

Tri Nguyen
LIGO/Gaia

Ouail Kitouni (TA)
LHCb

Noah Paladino
CMS

Siddharth Mithra-Sharma
MIT/Harvard IAIFI Fellow



Eric Moreno
CMS/LIGO

Duc Hoang
CMS

- Project leaders + Juan

Alex Shvonski
Online Course
Expert

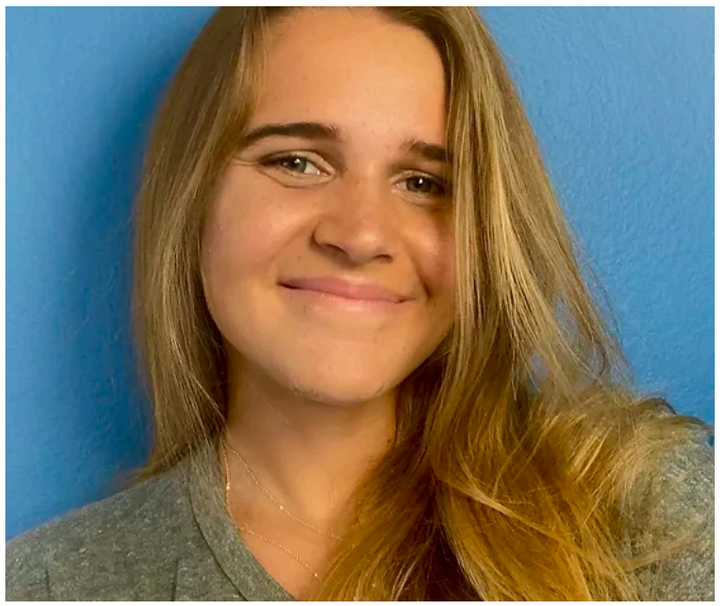
Jesse Thaler
IAIFI director

Dylan Rankin
CMS (MIT alum)

Undergrad TAs

Jack Dinsmore

Aidan Chambers



Brianna Ryan

- Recitations will be run by the undergrad/grad TAs
- Recitations will be hands on examples for projects
  - **Recitations will focus on getting the projects to work**

Communication

- We will communicate on Slack:
- mit-8S50-iap.slack.com
- Also we have canvas
  - Big announcements will be made on canvas
  - Day to day work, we intend to be on slack

How I got into physics

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

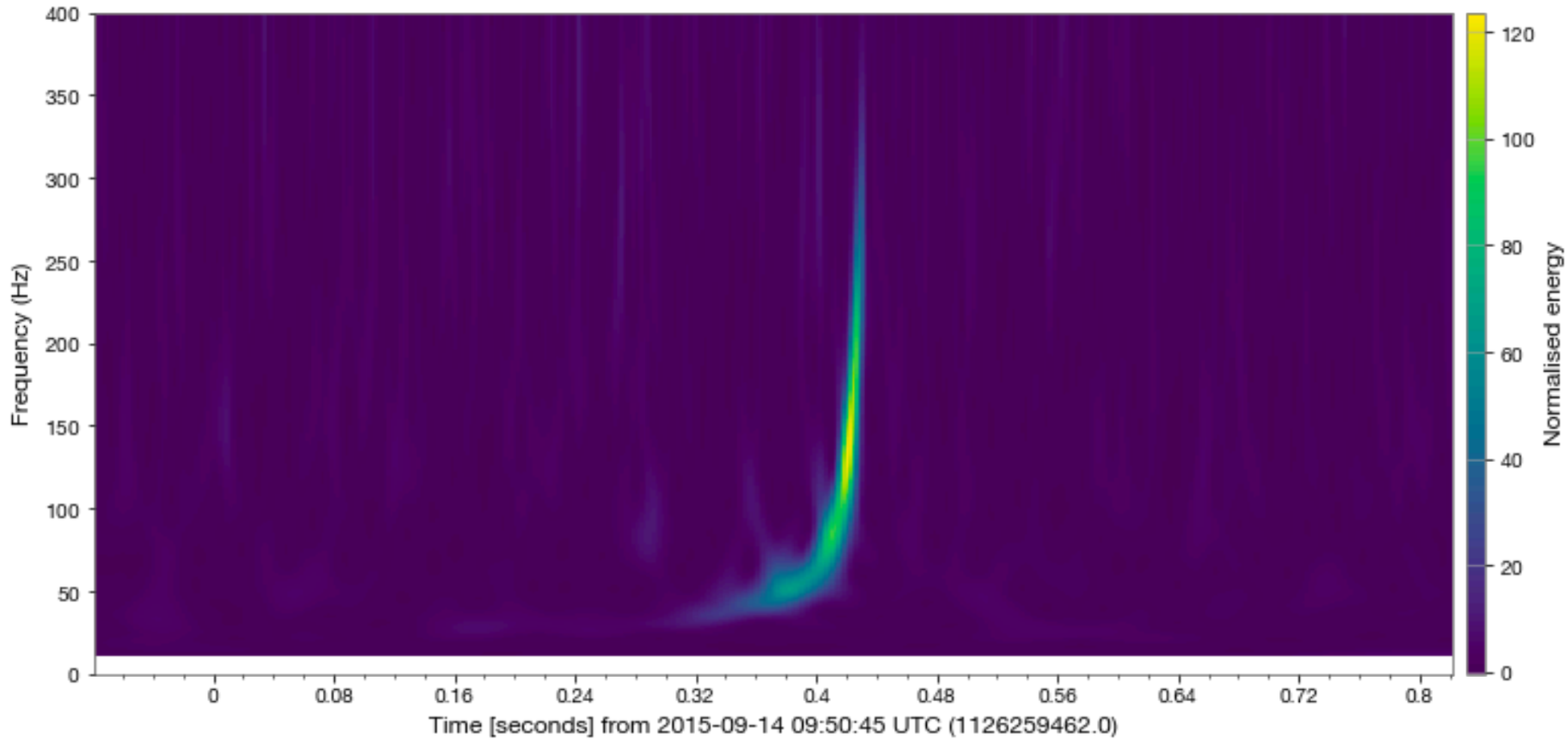
- I have a serious problem
  - I am addicted to data analysis
  - As a student I became obsessed with analyzing data
  - The most fun was building complex analyses
    - This class is a modern take on analyzing data
    - Now with all of the modern tools at hand

Projects

- All of these projects are going to use public 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

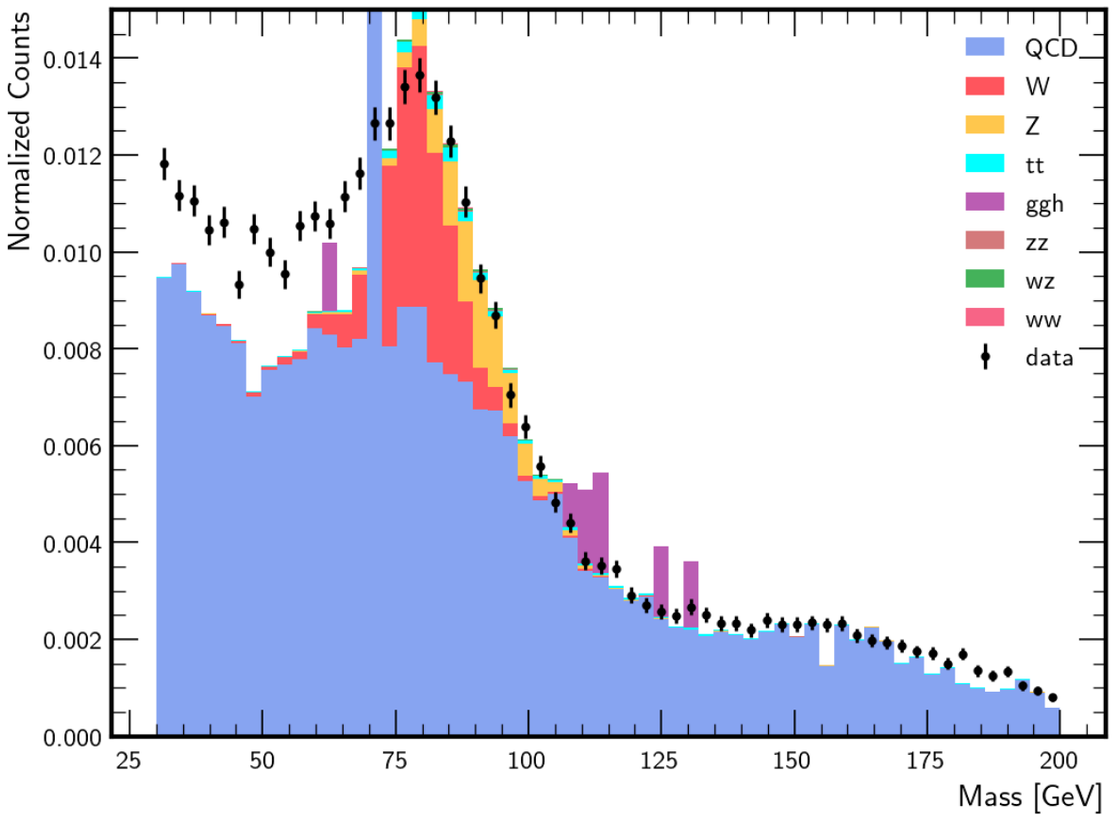
Project #1

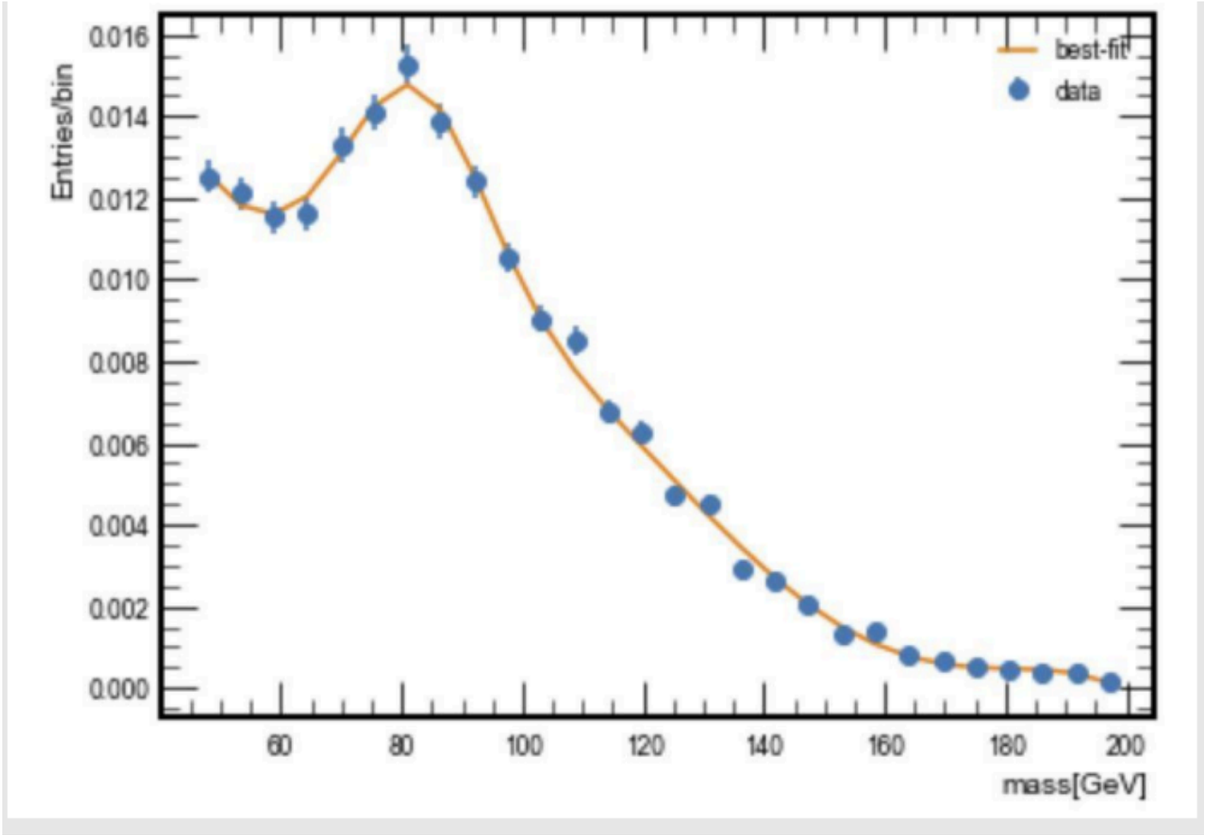
- Project 1 : Discovering Gravitational Waves



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

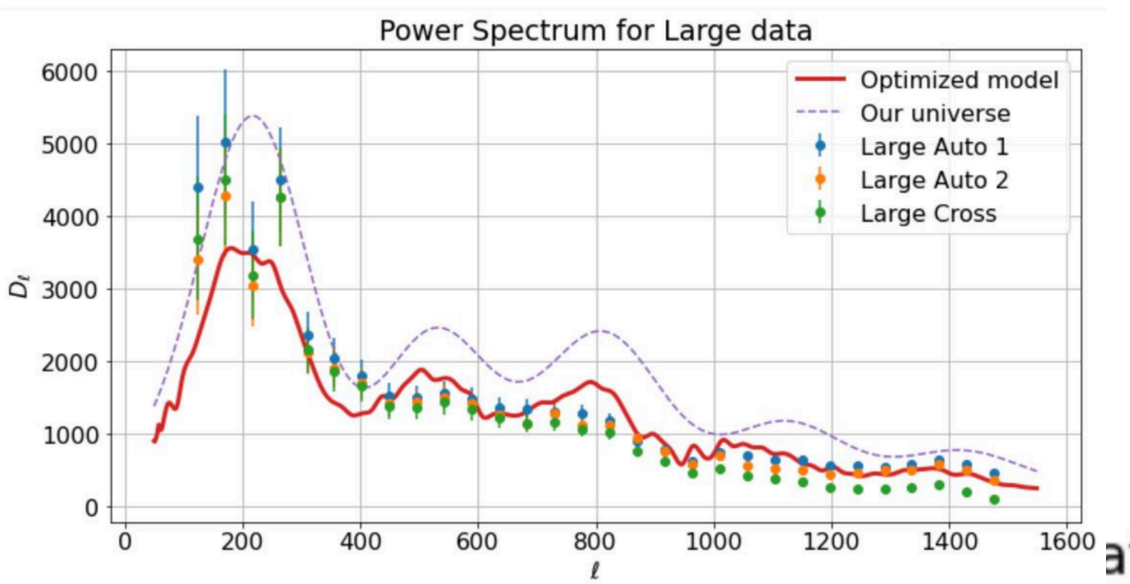
Project #2

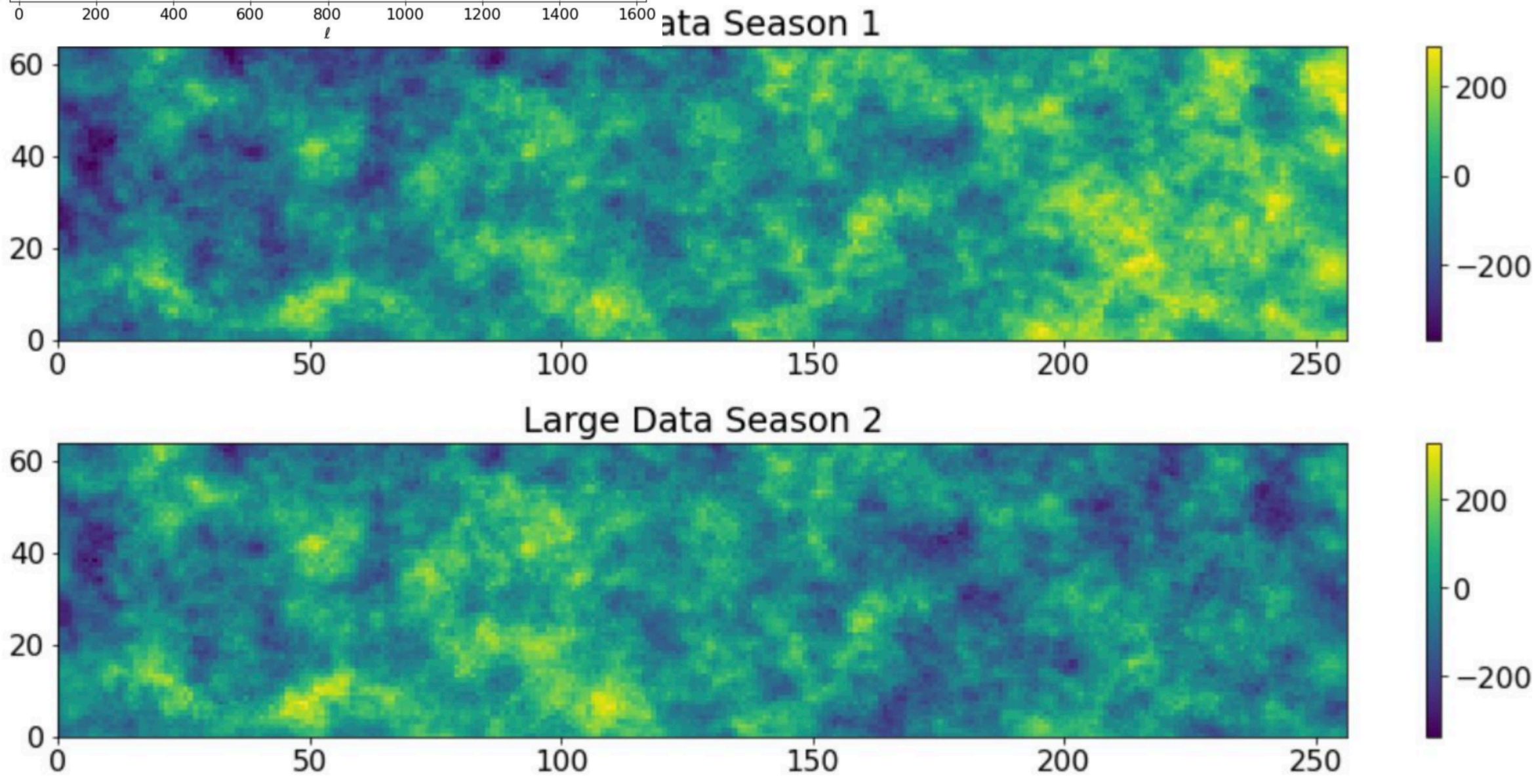




- Discover the W and Z boson decaying to quarks
- Try to enhance this measurement with deep learning

Project #3



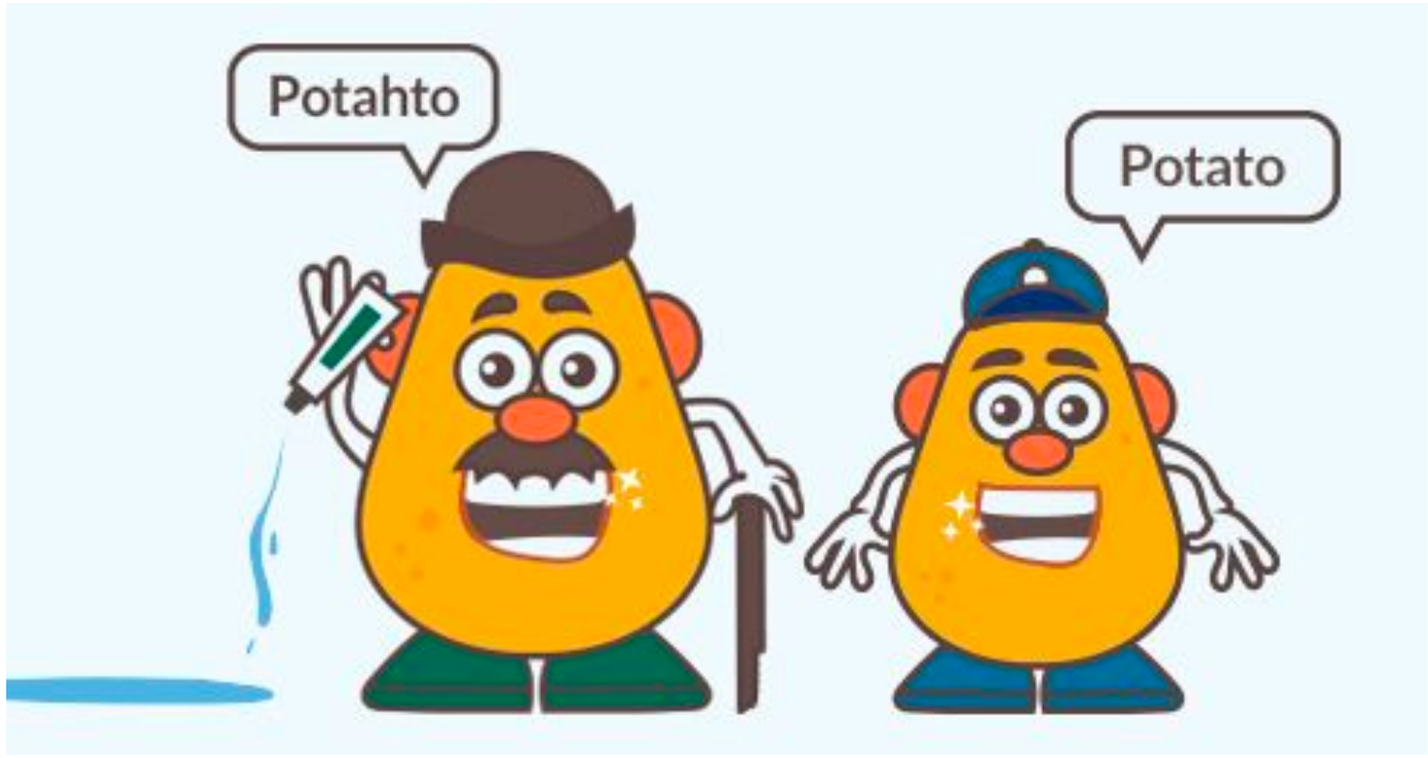


- How do we model the expansion of the universe

Reports

- We would like you to turn in a jupyter notebook
- Notebook should show some work
- We are still trying to figure out how to give feedback
- Turning in something is enough to pass the class

Solutions



- We will release worked out solutions
- Grading will be done by peer grading
- 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 booked the last two classes for you to present
- Thats a total of 180min
  - Depending on the enroll we will divide up the time
  - Talks will be 5-10 minutes depending on enrollment
  - You can post your talk on youtube before
- We will have a competition for the best talk in the class
  - Everybody will vote on the best talk
  - Winner will get $100 and a chance to present at an IAIFI seminar

Grading

- Focus of this class is on having fun
  - Don’t stress about grades!
- 20% Project 1,2,3
- 25% Final Talk
- 15% Participation
  - Participation means just turning in recitation notes

Assignment Format

- We will use github to perform the assignments
  - https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners
  - Assignments will be turned in as pull requests
    - Lectures will all be available on github as well
- Github is the standard toolkit for data science projects
  - You will have to learn it at some point
- Assignments are due on Monday at midnight AOE timezone

Syllabus

- Syllabus is posted on Canvas

Week 1: Project LIGO data analysis.
Goals : Plotting, Fitting, Parameter extraction, Time series analysis, statistical significance

Day 1: Class overview, Jupyter setup, making plots, Expectations, Variance
Day 2: LIGO Project, Error propagation, Poisson Statistics
Day 3: Minimization
Day 4: Uncertainty
Day 5: Normal distributions, confidence intervals, z-scores, non-gaussian distributions

- Week 1: The goal is to really understand fitting
- Why do we fit and what do want to do with it?

Syllabus

Week 2: **Project LHC Jet Physics Open Data analysis**

Goals : Hypothesis testing, interpolation, introduction to deep learning

Day 1: Correlations/Covariance, Convolutions **Project 1 due**

Day 2: Introduction to jets and collider physics

Day 3: Hypothesis testing, Bayesian vs Frequentist vs Hybrid

Day 4: f-tests/gaussian processes + semi-parametric methods

Day 5: Deep Learning, Guest Lecturer : Dylan Rankin

- Week 2: Goal is to understand modern data analysis techniques
- Leading up to why and how deep learning is used in physics

Syllabus

Week 3: **Project Cosmic microwave background analyses Lecturer: Juan Parra**
Goals : Image denoising strategies, template morphing, complex fitting strategies (MCMC)
Monday : MLK Day
Day 1: Overview of CMB measurements **Project 2 due**

- CMB and CMB data analysis pipelines: data -> maps -> power spectra -> parameters
- Bayes' Theorem and maximum likelihood techniques
- Map making and linear estimators
- Project problem assigned.

Day 2: Denoising Radio telescope data to make CMB maps
Day 3: Quadratic estimators and power-spectrum estimation
Day 4: Parameter Estimation with Markov Chain Monte Carlo

- Week 3: Goal is to perform a complicated astrophysical analysis
- Project builds on all the ideas from before but at a larger scale

Syllabus

Week 4: Pick your favorite project and make a presentation (review of last 3 weeks)
Goals : Applications of Deep Learning, review
Day 1 : Deep Learning Regression strategies
Day 2 : Probabilistic Programming from Guest Lecturer from Siddharth Mishra-Sharma
Day 3 : Guest lecture from Jesse Thaler
Day 4 : Presentations on selected projects
Day 5 : Presentations on selected projects

- Week 4: Goal is have fun and be creative
- This week, we focus on fun lectures, and getting projects done

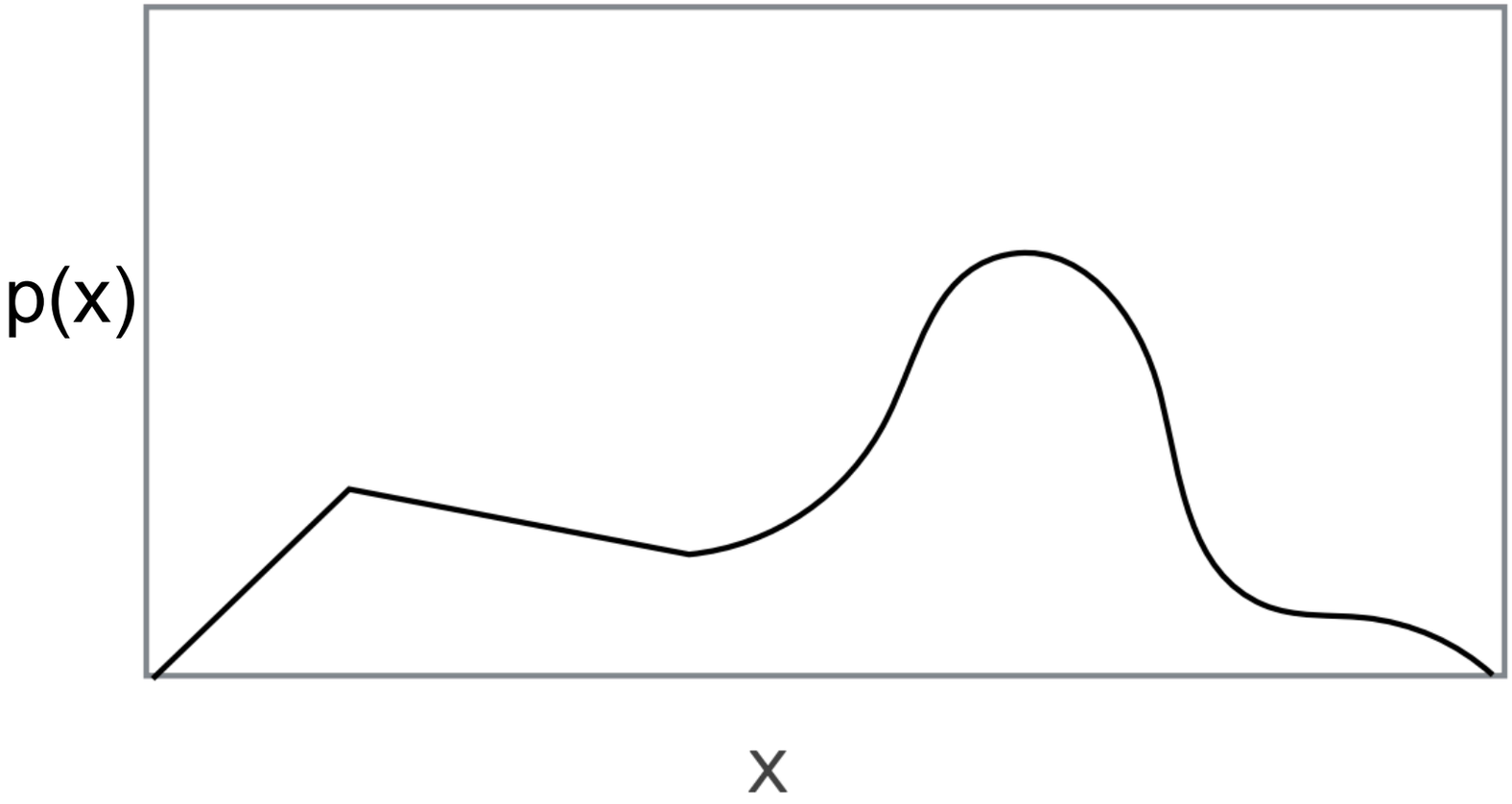
Disclaimer

- This is a fairly new class (2nd time we offered it)
  - Please be mindful of the fact that this is an experiment
  - We would really like to turn this into a full semester class
    - 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
- 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-8s50/course
- We are going to use jupyter and python for this class
  - https://github.com/MIT-8s50/course/blob/main/Recitation0/Recitation_0_setup.ipynb
  - https://www.dropbox.com/s/v6xk9z11vnp49jf/8.012%20Intro%20to%20Coding.ipynb?dl=0

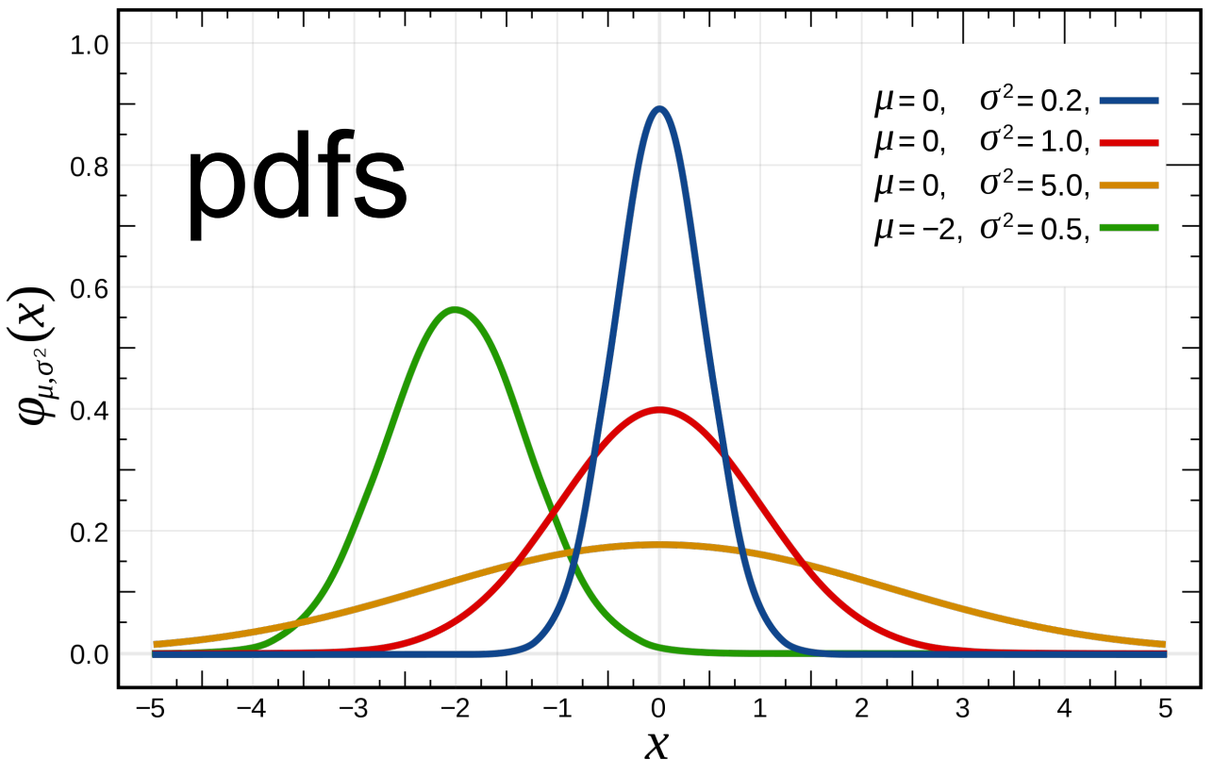
PDFs

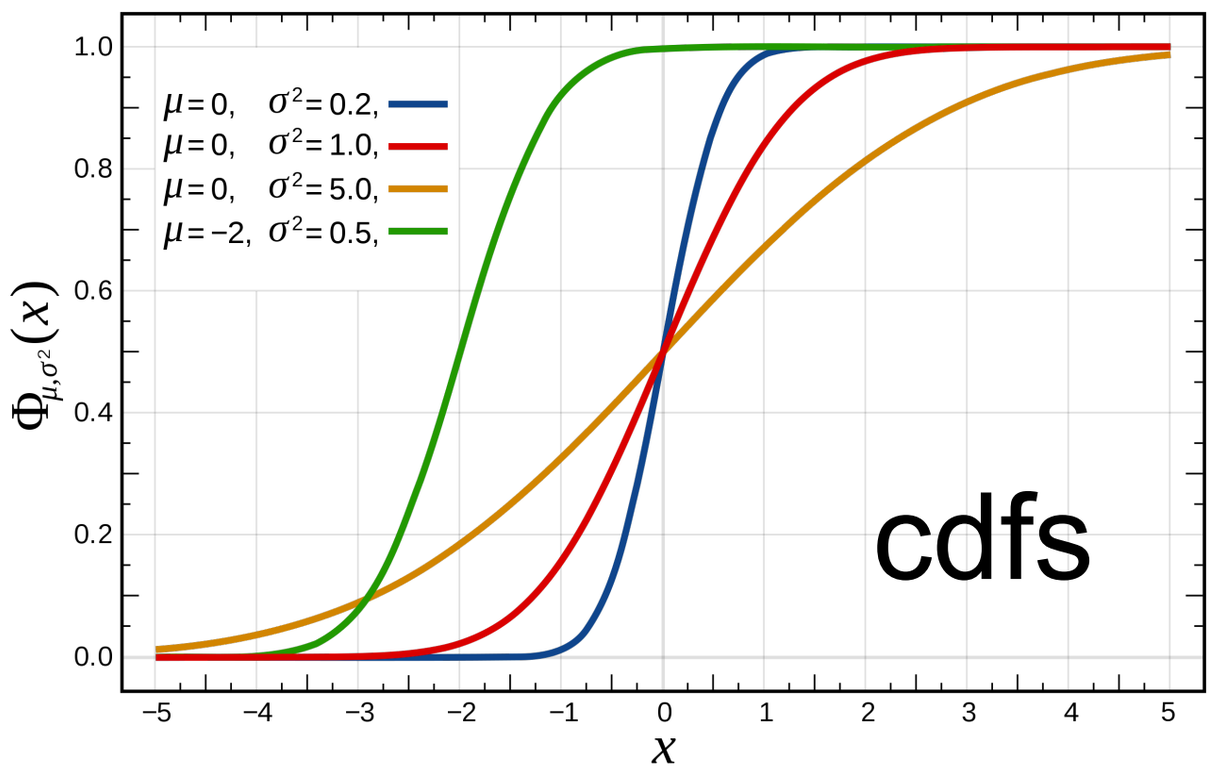


- Probability distribution(density) function *p(x)* sometimes *f(x)*
  - Probability of being between *x* and *x+dx*
  - $P(x \in [x, x + dx]) = p(x)dx$
  - $P(x \in [a, b]) = \int_a^b p(x)dx$

Probability can be disjoint

CDFs





- Cumulative distribution(density) functions or sometime CDFs

- $$\mathrm{cdf}(\mathrm{p}(\mathrm{x}), \mathrm{a}) = \int_{-\infty}^{\mathrm{a}} \mathrm{p}(\mathrm{x})\mathrm{dx}$$

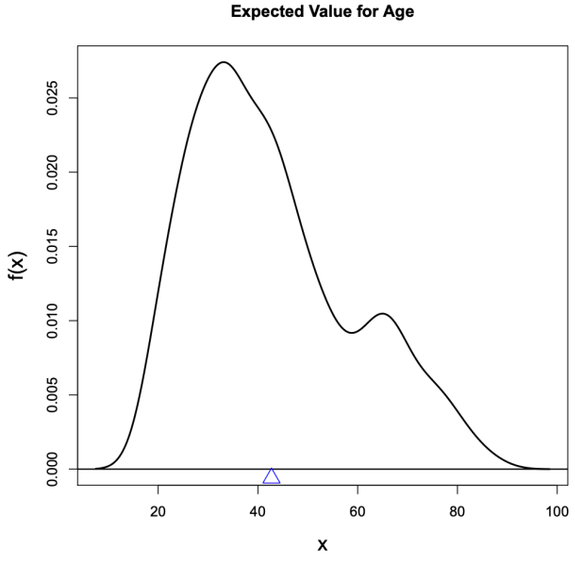
Expectation

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

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

Questions?