Computational Physics Class, November 3, 2021

Big Data and Computing at UConn

Cara Battersby, Assistant Professor of Physics

Slides adapted from Prof. Richard Jones, activity from Ravi Wickramathilake and Prof. Vernon Cormier

What is Big Data?

- **Large** (> terabytes - 10^{12} aka 2^{40} bytes), **complex** (can be structured or unstructured), **diverse** (different types) data.
- Data that has high enough volume (total amount), velocity (speed of reception and/or processing), and/or variety (different types and unstructured) that **traditional data storage and processing techniques are insufficient.**

What is Big Data?

- **Large** (> terabytes - 10^{12} aka 2^{40} bytes), **complex** (can be structured or unstructured), **diverse** (different types) data.
- Data that has high enough volume (total amount), velocity (speed of reception and/or processing), and/or variety (different types and unstructured) that **traditional data storage and processing techniques are insufficient.**
- A widely-used **umbrella term**!

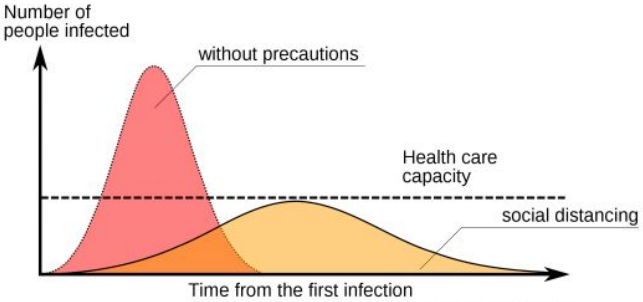
Wikipedia contents under big data

Case Study: Modelling Hospital Demand

Flattening the Curve

Protect the Resources to Protect Us All

- China and Italy were the two hardest hit countries at the beginning of COVID-19
  - Both implemented early lockdowns
- Italy was running out of hospital beds and ventilators
  - Hospitals reported new arrivals every 5 minutes



By Johannes Kalliauer via Wikimedia Commons

Flattening the curve protects resources, but the allocation still has to work

*Slides adapted from Richard Jones, **this slide borrowed from presentation by James P. Howard, II,** 03/01/21, Johns Hopkins Applied Physics Laboratory*

Case Study: Modelling Hospital Demand

Patient Allocation Model

- Goals
  - Maximize utilization of hospital beds
  - Minimize patient travel
  - Minimize displacement of future patients
- Decisions
  - What hospital do we send a patient to?
  - What penalty do we incur for potentially displacing patients?
- Implementation
  - Python and Pyomo
  - Solvable with CBC, CPLEX, GLPK

$$\min \sum_j d_j x_j + \sum_{j,t} p_{jt}$$

$$\text{s.t.} \sum_j x_j = 1 \; \forall j$$

$$x_j \leq b_j x_j - o_j x_j - \sum_{t' \leq t} a_{jt'} x_j + \sum_{t' \leq t} d_{jt'} x_j + p_{jt} \; \forall j, t$$

$$x_j \in \{0,1\}$$

$$p_{jt} \geq 0$$

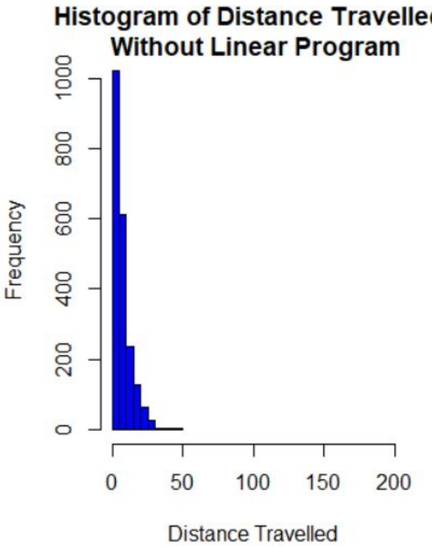
Case Study: Modelling Hospital Demand

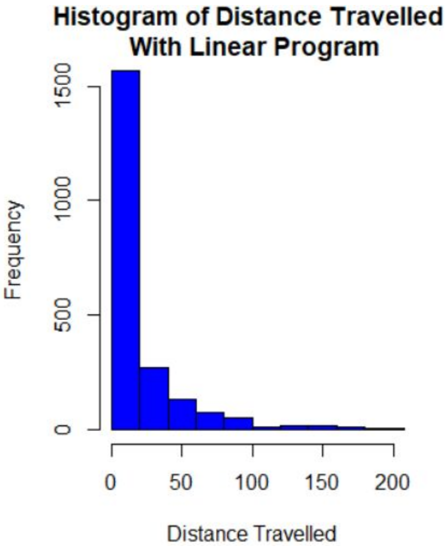
How OSG Helped

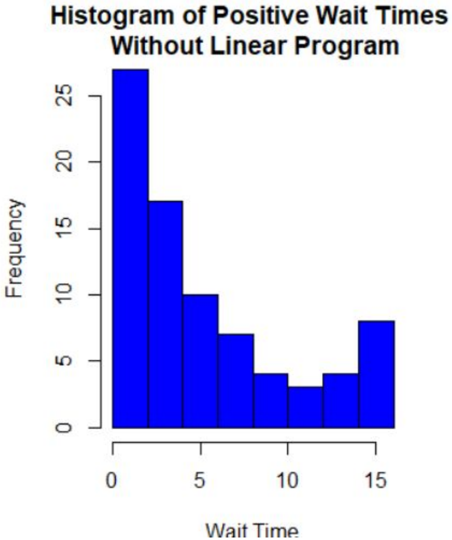
- Monte Carlo simulation
  - Requires many runs
  - All runs are independent of each other
  - Perfect HTC problem
  - OSG made for HTC
- 1600 runs for each policy made
  - Run may take multiple hours
  - Just need to capture metrics

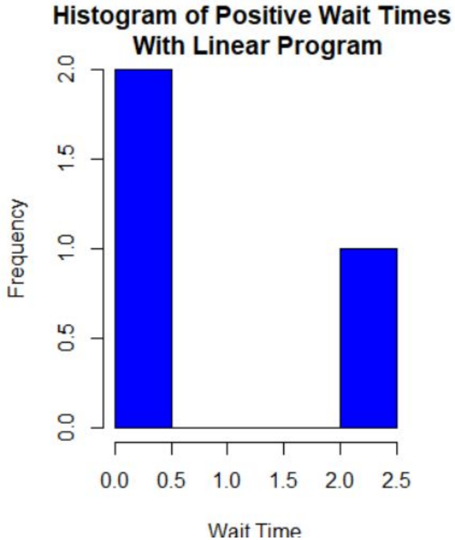
- Python worked well
  - Used virtualenv per OSG documentation
  - All required packages added to the venv
  - Added compiled copy of glpk to venv
    - Felt shady but necessary for Pyomo
  - CSV files with patient stats written
  - Logfiles via spdlog generated
  - Results turned to OSG submit node
    - Processed stats in R
    - Logs provided info to answer questions

Case Study: Modelling Hospital Demand









Capacity Computing for Data-Driven Science

- Can you think of any examples in your field?

Capacity Computing for Data-Driven Science

- Can you think of any examples in your field?

- Why might I need it?
  - It takes too long on my laptop!
  - The data won't fit " " "
  - I have other stuff to do " " "
  - Or, ***what more might I be able to do?*** Check more alternative ideas / models? Do more simulations to understand my data? Try out speculative ideas, just because I can?
- Where might I turn?
  - my research group or Department's dedicated resources
  - the Cloud
  - Central IT (HPC at UConn!) research computing resources
  - National HTC (Open Science Grid) or HPC resources for science

High Performance vs. Throughput Computing (HPC / HTC)

High Performance Computing (HPC):

- Problems that **can't be decomposed into small parts** (e.g. inverting a large matrix, lattice QCD)
- Problems that have **huge memory** and/or **cooperative scheduling** requirements
- Highly tuned complex problems. ***Like a highly tuned Formula 1 car*** -- built for a specific high performance task, but requires special skills to use.

High Throughput Computing (HTC):

- **e.g. UConn's HPC cluster**** https://hpc.uconn.edu/ **or the Open Science Grid,** https://opensciencegrid.org
- HTC tasks **can be decomposed into small parts**, but need to be done many many times (e.g. analyzing millions of individual images).
- ***Like 100,000 compact cars*** -- don't need any special skills to use it, but has a large throughput.
- Most modern-day big data science problems require HTC.

Open Science Grid

*** UConn's cluster is named an **HPC,** because it was when it was first developed 20 years ago, but now would be considered **HTC.***

High Performance vs. Throughput Computing (HPC / HTC)

HPC Cluster - for smaller, flexible jobs

OSG - for bigger jobs

High Performance Computing (HPC):

- Problems that **can't be decomposed into small parts** (e.g. inverting a large matrix, lattice QCD)
- Problems that have **huge memory** and/or **cooperative scheduling** requirements
- Highly tuned complex problems. ***Like a highly tuned Formula 1 car*** -- built for a specific high performance task, but requires special skills to use.

High Throughput Computing (HTC):

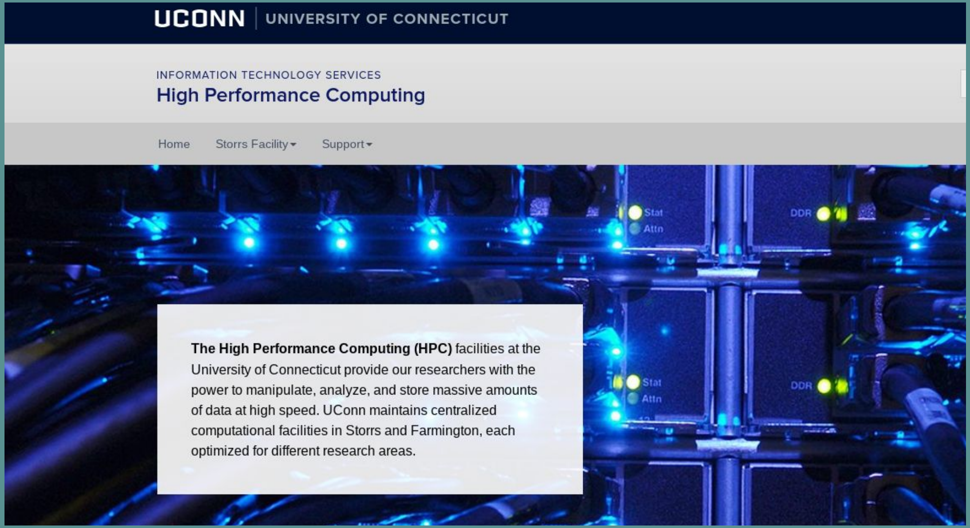
- **e.g. UConn's HPC cluster** https://hpc.uconn.edu/ or the Open Science Grid, https://opensciencegrid.org**
- HTC tasks **can be decomposed into small parts**, but need to be done many many times (e.g. analyzing millions of individual images).
- ***Like 100,000 compact cars*** -- don't need any special skills to use it, but has a large throughput.
- Most modern-day big data science problems require HTC.

Open Science Grid

*** UConn's cluster is named an **HPC,** because it was when it was first developed 20 years ago, but now would be considered **HTC.***

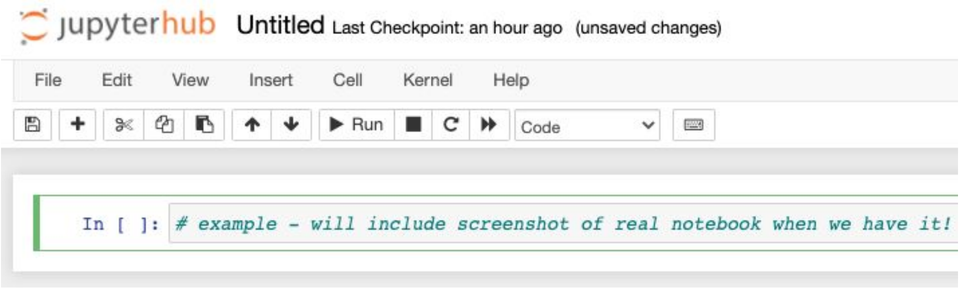
UConn’s HPC Cluster

- https://hpc.uconn.edu/
- Anyone can make an account!
- No proposal required for jobs!
- Run locally!
- We will run a jupyter notebook on the cluster today



Instructions to access the HPC cluster

Jupyter Hub
https://cn410.storrs.hpc.uconn.edu:48000/



Terminal

```
$ ssh
<your_netid>@login.storrs.hpc.uconn.edu
password: <your_netid_password>

... <write your code>

... <test your code>

... <write yourscript.sh to automate running
of your code>

[netid@cn01 ~]$ sbatch -p generalepyc
<yourscript.sh>   #submits your job

[netid@cn01 ~]$ squeue
#watches your job run

... <look at your results, fix the problems,
try it again> ...
```

High Performance vs. Throughput Computing (HPC / HTC)

HPC Cluster - for smaller, flexible jobs

OSG - for bigger jobs

High Performance Computing (HPC):

- Problems that **can't be decomposed into small parts** (e.g. inverting a large matrix, lattice QCD)
- Problems that have **huge memory** and/or **cooperative scheduling** requirements
- Highly tuned complex problems. ***Like a highly tuned Formula 1 car*** -- built for a specific high performance task, but requires special skills to use.

High Throughput Computing (HTC):

- **e.g. UConn's HPC cluster**** https://hpc.uconn.edu/ or the **Open Science Grid,** https://opensciencegrid.org
- HTC tasks **can be decomposed into small parts**, but need to be done many many times (e.g. analyzing millions of individual images).
- ***Like 100,000 compact cars*** -- don't need any special skills to use it, but has a large throughput.
- Most modern-day big data science problems require HTC.

Open Science Grid

*** UConn's cluster is named an **HPC,** because it was when it was first developed 20 years ago, but now would be considered **HTC.***

The Open Science Grid

Open Science Grid

- https://opensciencegrid.org
- **Anyone (**researchers at any US academic, government, or non-profit organization) can make an account! https://www.osgconnect.net/ → **sign up!**
- **Free access to the Open Science Pool** via an OSG-supported access point, no proposal / allocation necessary
- **Includes:**
  - Initial consultation with an OSG research computing facilitator
  - Online documentation and examples
  - Access to OSG's central software modules
  - (roughly) unlimited scratch, space for staging large datasets and software, with cache across OSG

Open Science Grid

Submit locally, run globally.

US Researchers Can Use Open Science Pool, NOW!

osgconnect.net > Sign Up

Questions

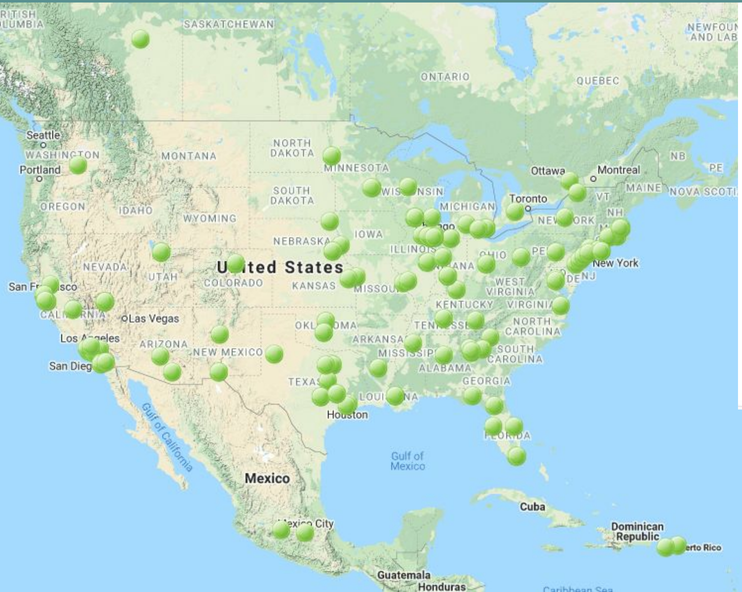
Lauren Michael, lmichael@wisc.edu

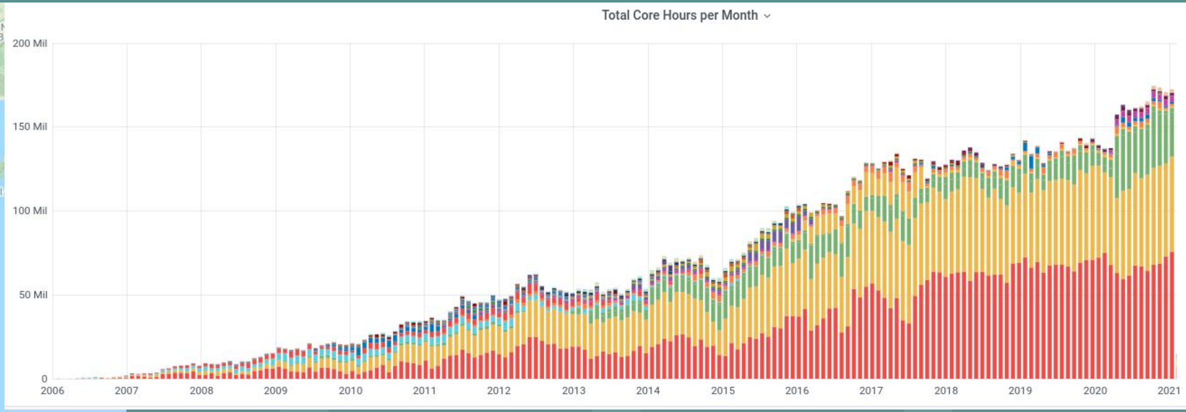
Research Facilitation Lead (campuses and researchers), OSG

This work was partially supported by the NSF grants OAC-2030508, OAC-1841530, OAC-1836650, and MPS-1148698

The Open Science Grid

- *NOT* a replacement for traditional *HPC*
- most science problems do not require *HPC*, just a lot of time to run ordinary process over a larger volume of data or simulations -- *high-throughput computing*

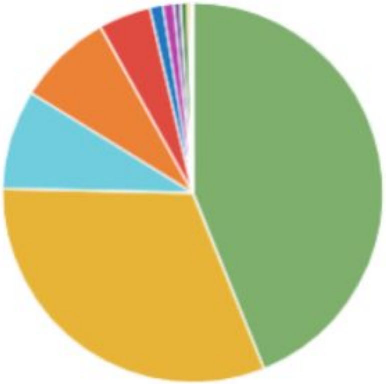




Open Science Grid

Open Science Pool in 2020

Core Hours by Field of Science



	total
Biological Sciences	133.4 Mil
Physics	95.7 Mil
Astronomy	25.7 Mil
Chemistry	24.7 Mil
Engineering	13.47 Mil
Integrative Activities	3.34 Mil
Mathematics	2.834 Mil
Agricultural Sciences	1.712 Mil
Health	1.679 Mil
Education	746 K
Computer Sciences	546 K
Other	90.4 K
Economics	87.4 K
Earth and Ocean Sciences	13.70 K
Other Social Sciences	0.0047

https://gracc.opensciencegrid.org/d/000000077/open-science-pool-all-usage?orgId=1

Covid19 research

Dark matter simulations

Core Hours by Institution

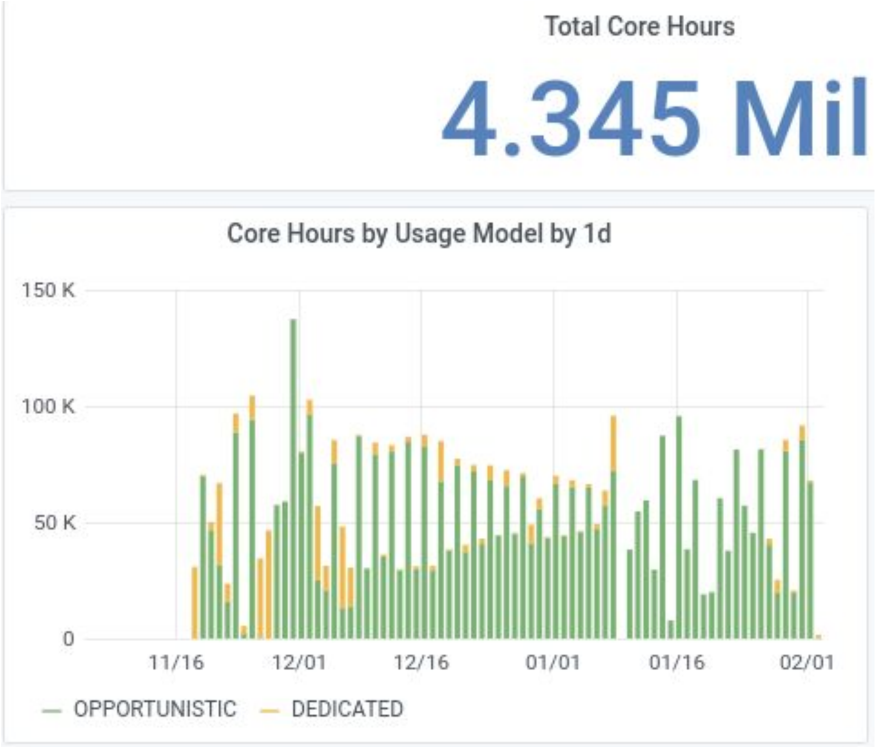
	total
Folding@Home Consortium (FAHC)	86.7 Mil
Massachusetts Institute of Technology	35.4 Mil
Stanford University	26.4 Mil
University of Pittsburgh	22.1 Mil
University of Hawaii at Manoa	15.0 Mil
Fermilab	12.39 Mil
Rochester Institute of Technology	12.31 Mil
New Mexico State University	12.14 Mil
Wayne State University	11.62 Mil
University of Chicago	8.40 Mil
University of Pennsylvania	8.35 Mil
University of North Carolina at Chapel Hill	7.09 Mil
Lancaster University	6.01 Mil
University of Arizona	5.64 Mil
Arizona State University	5.60 Mil
University of Wisconsin-Madison	5.04 Mil
LSU School of Public Health	3.69 Mil
Georgia Institute of Technology	3.49 Mil
Rutgers, The State University of New Jersey	2.150 Mil
Brookhaven National Laboratory	2.079 Mil

UConn and The Open Science Grid

- began at UConn Health
- at Storrs: 2009 US DOE grant (R. Jones, PI) to explore possibilities
  - formed "virtual organization" of users (Gluex)
  - set up UConn Storrs as "grid site" on OSG
  - outward-facing services (VOMS, CondorCE, StachCache, gFTP, xrootd)
  - accounting and reporting (Gratia, GRACC)
  - long-lived outgoing connections (hours)
- 2019 - NSF equipment grant ($400k) for new OSG resource UConn-HTC
  - shared use model -- letter of support from UConn CIO
  - 38 new nodes housed in the data center (HPC racks)
  - different network requirements from existing HPC-Storrs cluster

https://gracc.opensciencegrid.org

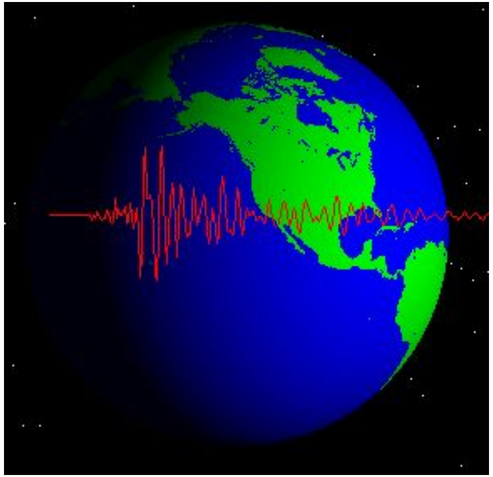
UConn-HTC operations



Nuclear Physics	1.16 M
Engineering	1.15 M
Biological and Biomedical Sci.	911.44 K
Physics	835.11 K
Astronomy	589.36 K
Chemistry	513.12 K
Astrophysics	507.64 K
Astronomy and Astrophysics	338.65 K
High Energy Physics	323.10 K
Bioinformatics	141.57 K
Statistics	121.36 K
Comp. Architecture/Comp. Eng.	101.91 K
Materials Science	73.26 K
Evolutionary Biology	55.14 K
Biological Sciences	54.73 K
Computer Sciences	36.74 K
Education	22.94 K
Biochemistry	18.16 K
Mathematical Sciences	8.25 K
Physical Therapy	4.96 K
Elementary Particle Physics	3.29 K
Computer Science	2.70 K
Biophysics	2.37 K
Information Science and Eng.	1.78 K
Agricultural Sciences	917.36
Geographic Information Sci.	56.91
Computer and Information Sci.	27.92
Biomedical research	3.18
Evolutionary Sciences	1.24
Computer and Info. Services	0.73
Health	0.49
Multi-Science Community	0.07

Real World Example - Using the global seismic wavefield to understand Earth's interior

- Use 100,000 individual seismograms to **estimate the radius of Earth's core!**

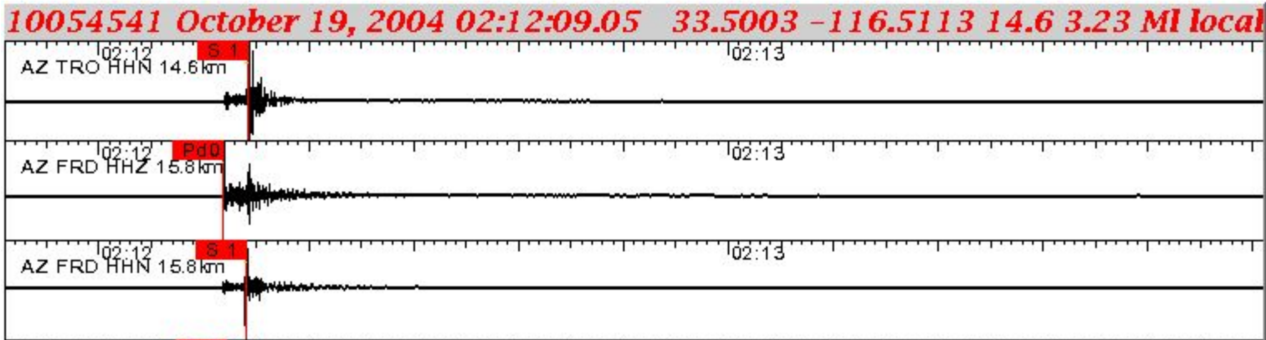


From Harvey Mudd College Seismology:
http://www.physics.hmc.edu/research/geo/seismo.html

Real World Example - Using the global seismic wavefield to understand Earth's interior

Global Seismic Wavefield

- An **individual seismogram** records ground motion at a single location as a function of time. They are essential to detect the location and magnitude of earthquakes
- A global seismic wavefield incorporates these from across the globe and stacks them over time.

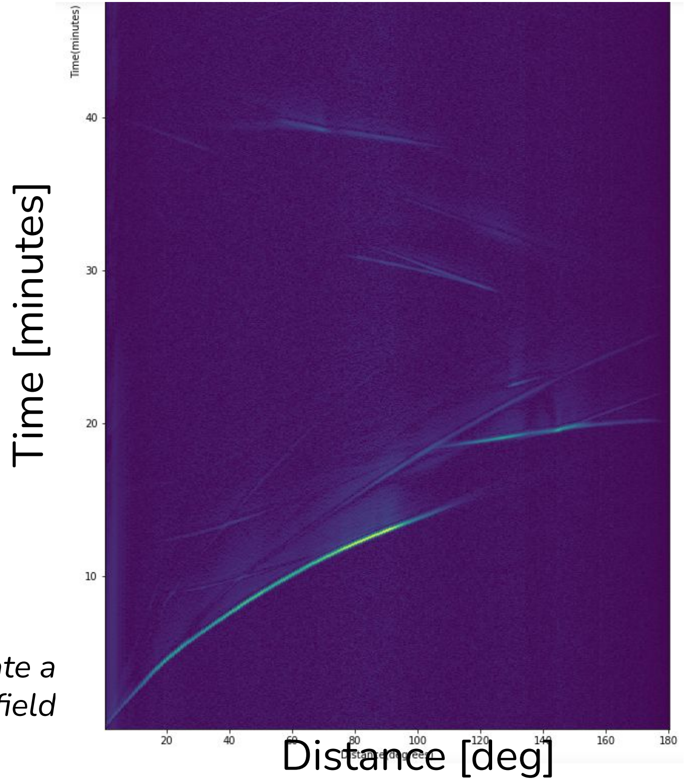


Example seismogram of an earthquake from the USGS (wikipedia) at 1 location with 3 components of motion (vertical (Z), north-south, and east-west.

Real World Example - Using the global seismic wavefield to understand Earth's interior

Global Seismic Wavefield

- An individual seismogram records ground motion at a single location as a function of time. They are essential to detect the location and magnitude of earthquakes
- A **global seismic wavefield** incorporates these from across the globe and stacks them over time.

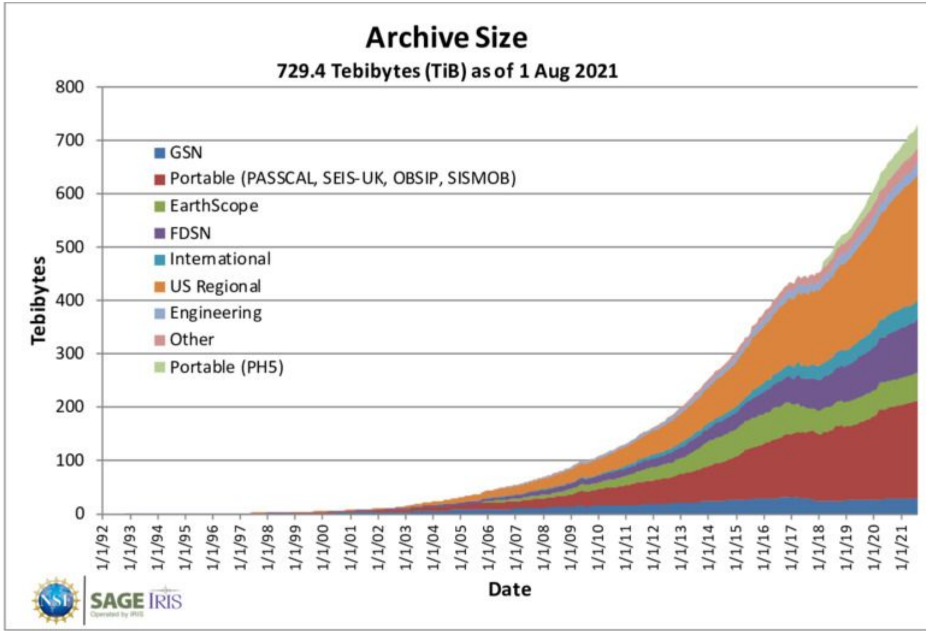


100,000 Stacked seismograms to create a global seismic wavefield

Real World Example - Using the global seismic wavefield to understand Earth's interior

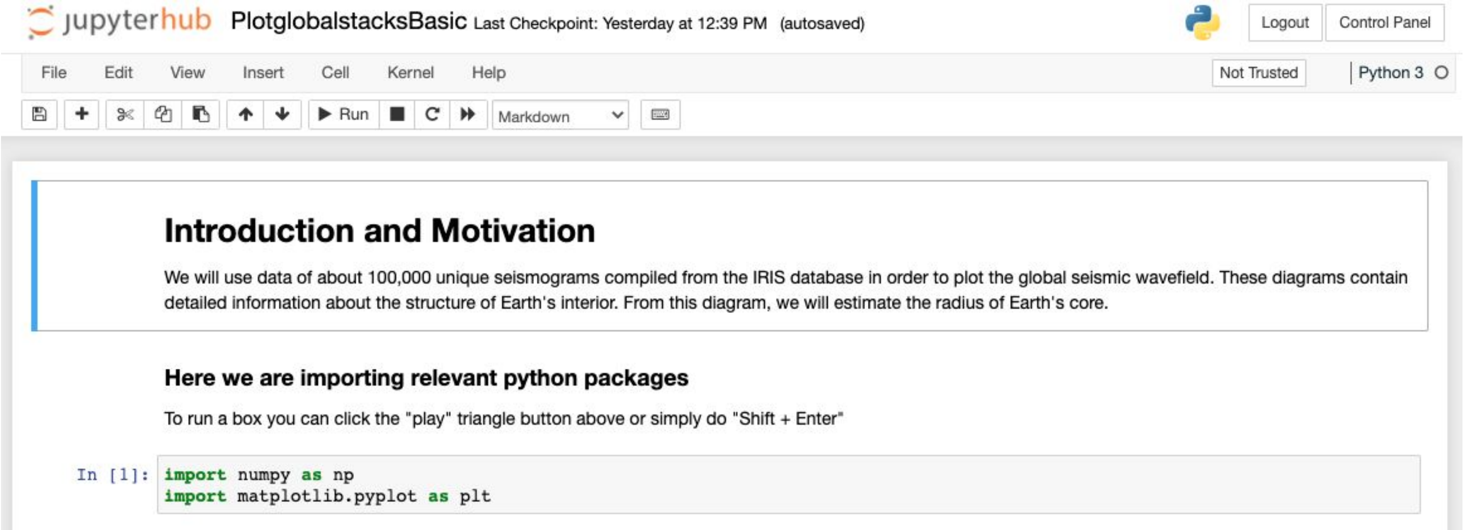
The Data

- These data are from IRIS (Incorporated Research Institutions for Seismology) and comprise data from 1000s facilities over 29 years.
- Archive is now more than 700 Tebibytes (1 TiB = 1.1 TB, so 770 TB), aka HUGE!



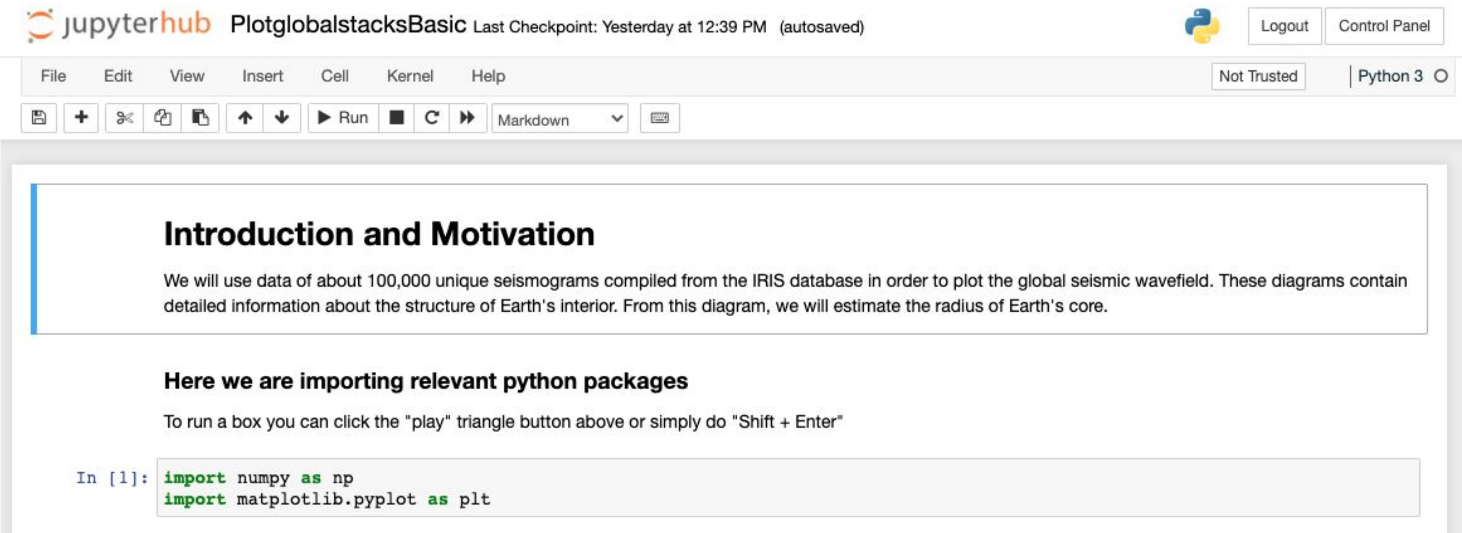
Real World Example - Using the global seismic wavefield to understand Earth's interior

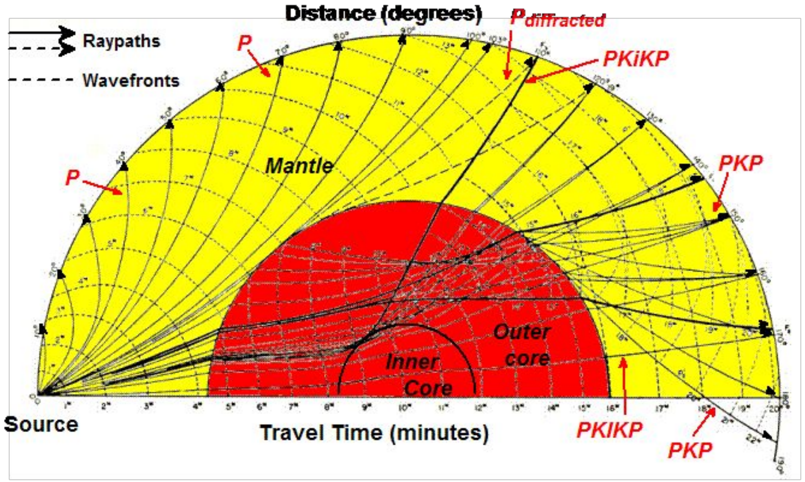
- We will compile the data from 100,000 unique seismograms, retrieved from the IRIS database to plot the **global seismic wavefield**
- This tells us about Earth's interior
- We will **estimate the radius of the Earth's core**

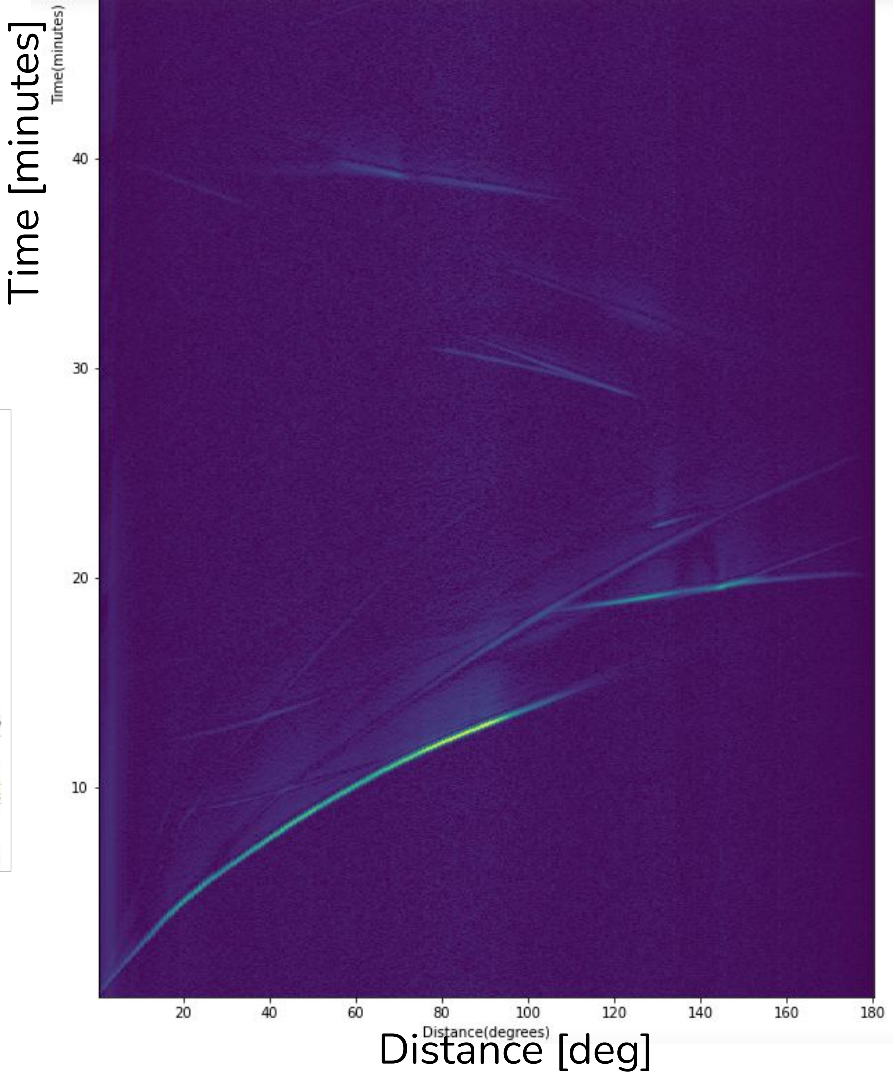


Real World Example - Using the global seismic wavefield to understand Earth's interior

- Log in to jupyter hub with your netID: https://cn410.storrs.hpc.uconn.edu:48000/
- Follow along
- Work in groups of 2-3 (and ask questions!) to estimate the radius of Earth's core



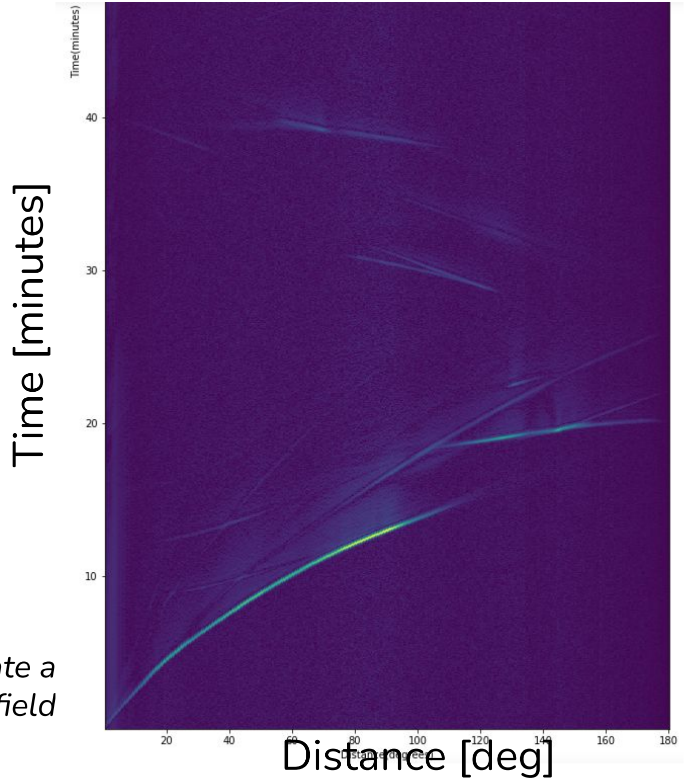

Distance (degrees)
Raypaths
Wavefronts
P
P
$P_{diffracted}$
PKiKP
PKP
Mantle
Outer core
Inner Core
Source
Travel Time (minutes)
PKIKP
PKP


Time [minutes]
Time(minutes)
40
30
20
10
20
40
60
80
100
120
140
160
180
Distance(degrees)
Distance [deg]

Real World Example - Using the global seismic wavefield to understand Earth's interior

Global Seismic Wavefield

- Final result
- Interpretation
- Questions?
- Feedback if you have time: https://docs.google.com/forms/d/e/1FAIpQLSeyQXn32AhHv4HmJhM5q4NyJxSC_q8on3JZ3xEpgjwhhvejCw/viewform?usp=sf_link



100,000 Stacked seismograms to create a global seismic wavefield