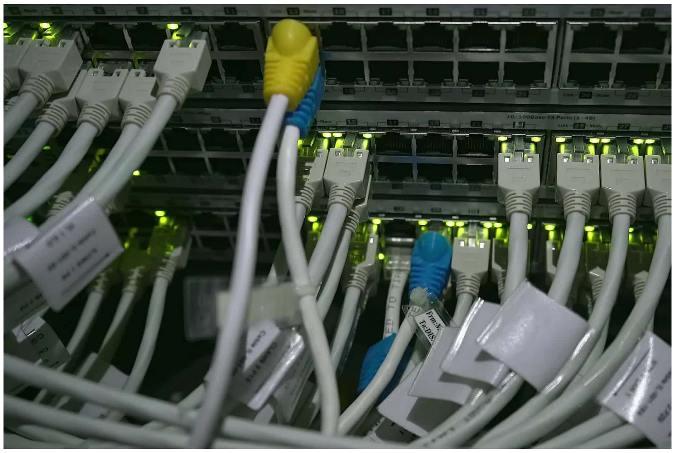
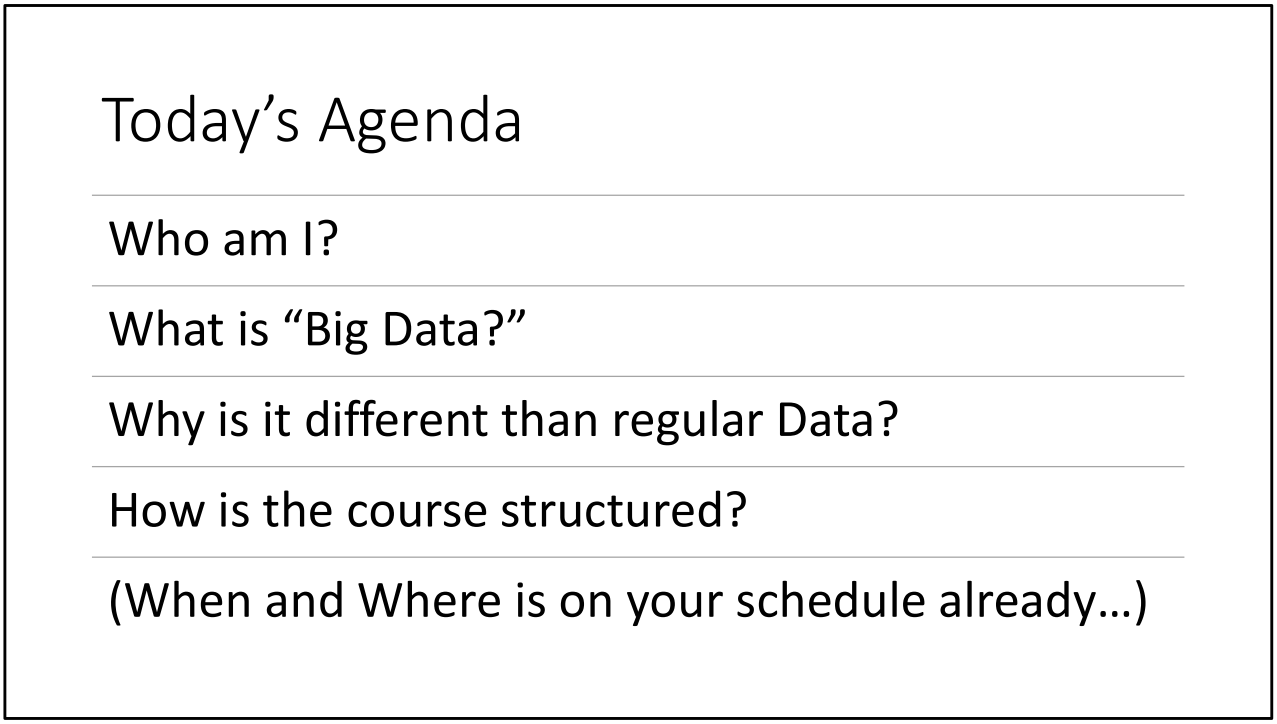
Data-Intensive Distributed Computing CS431/451/631/651

Fall 2024 – Dan Holtby





Today’s Agenda

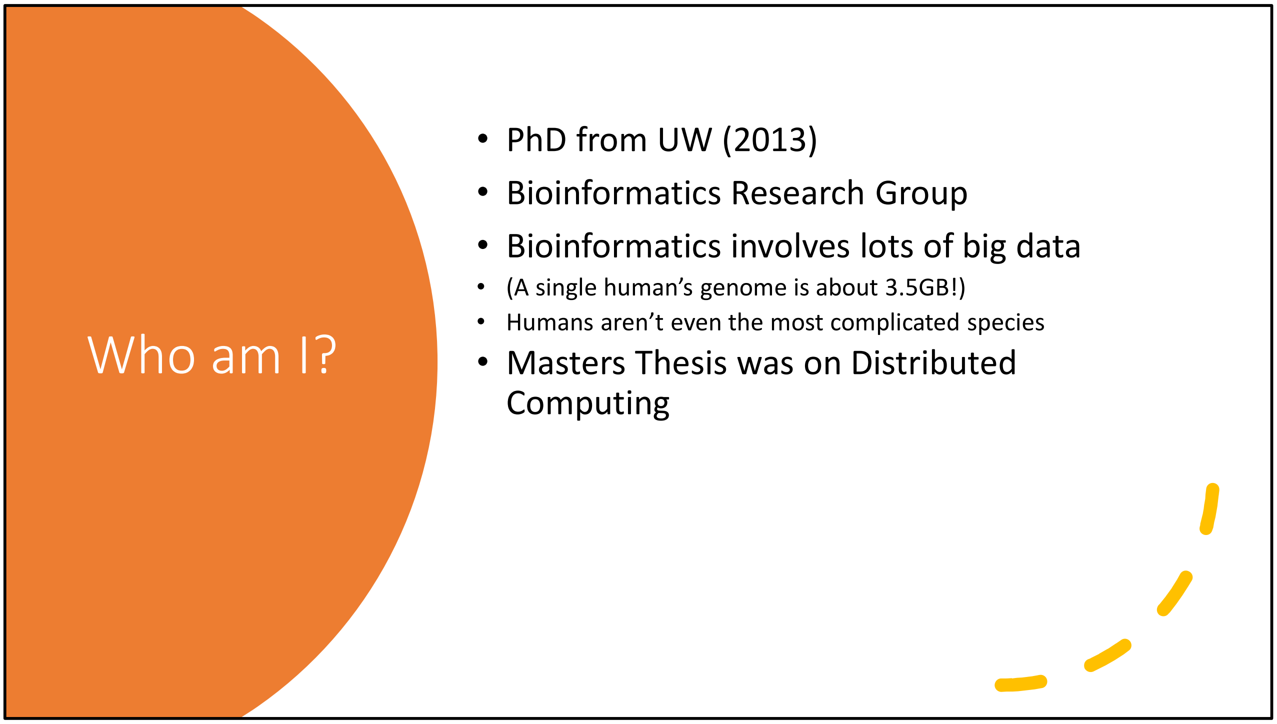
Who am I?

What is “Big Data?”

Why is it different than regular Data?

How is the course structured?

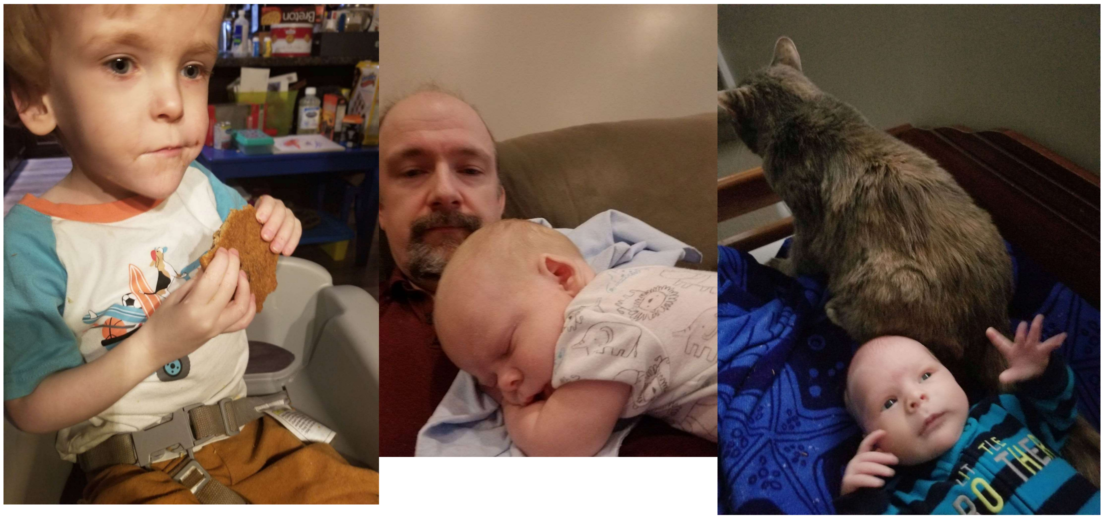
(When and Where is on your schedule already…)



Who am I?

- PhD from UW (2013)
- Bioinformatics Research Group
- Bioinformatics involves lots of big data
- (A single human's genome is about 3.5GB!)
- Humans aren't even the most complicated species
- Masters Thesis was on Distributed Computing

My Family

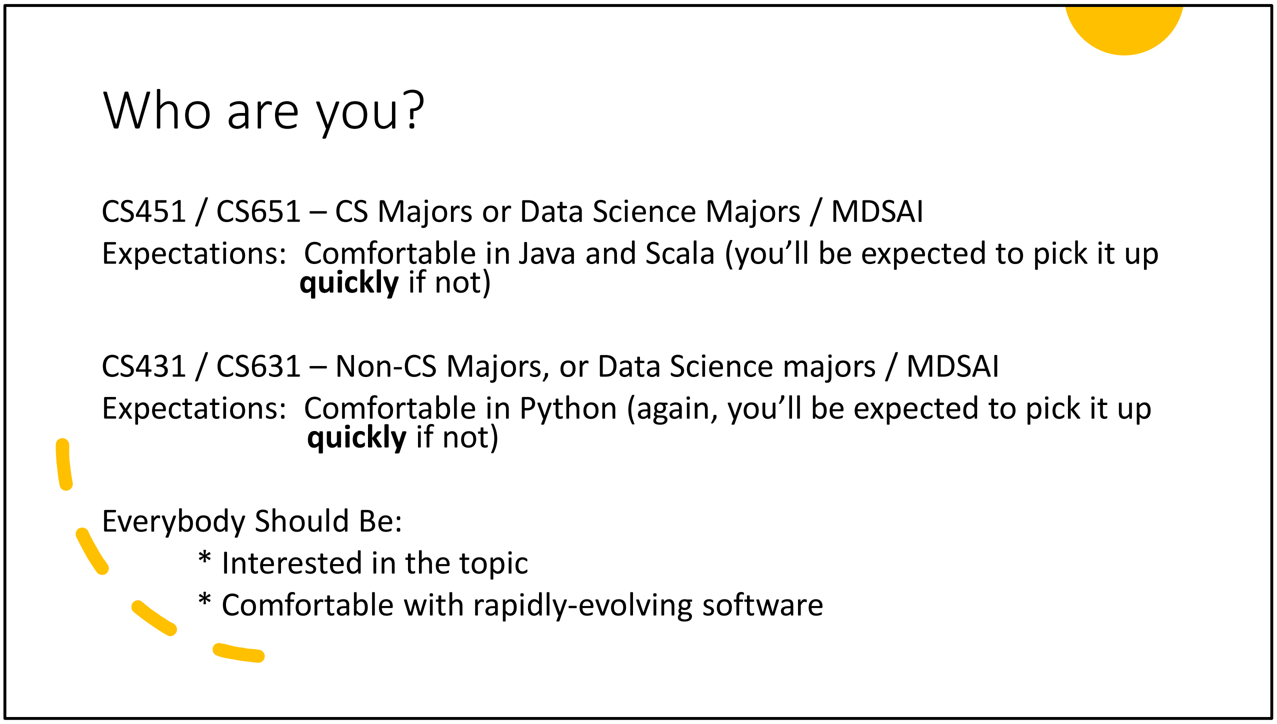


Left to right: Charlie, Me and Ollie, Pixie and Ollie.

My Hobbies

Currently I’m addicted to using Stable Diffusion img2img to Anime-ify family photos.

I have a problem. Please help. There’s so much work to do.



Who are you?

CS451 / CS651 – CS Majors or Data Science Majors / MDSAI

Expectations: Comfortable in Java and Scala (you’ll be expected to pick it up **quickly** if not)

CS431 / CS631 – Non-CS Majors, or Data Science majors / MDSAI

Expectations: Comfortable in Python (again, you’ll be expected to pick it up **quickly** if not)

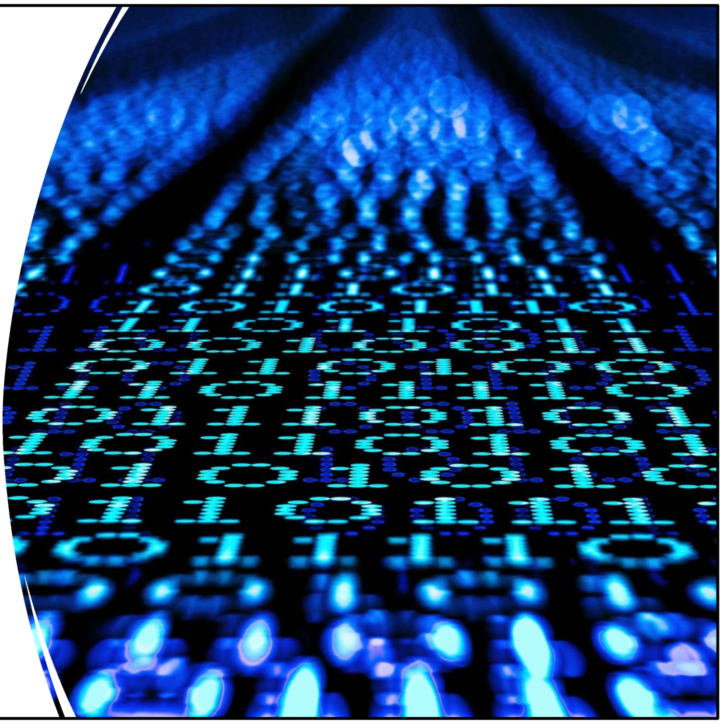
Everybody Should Be:

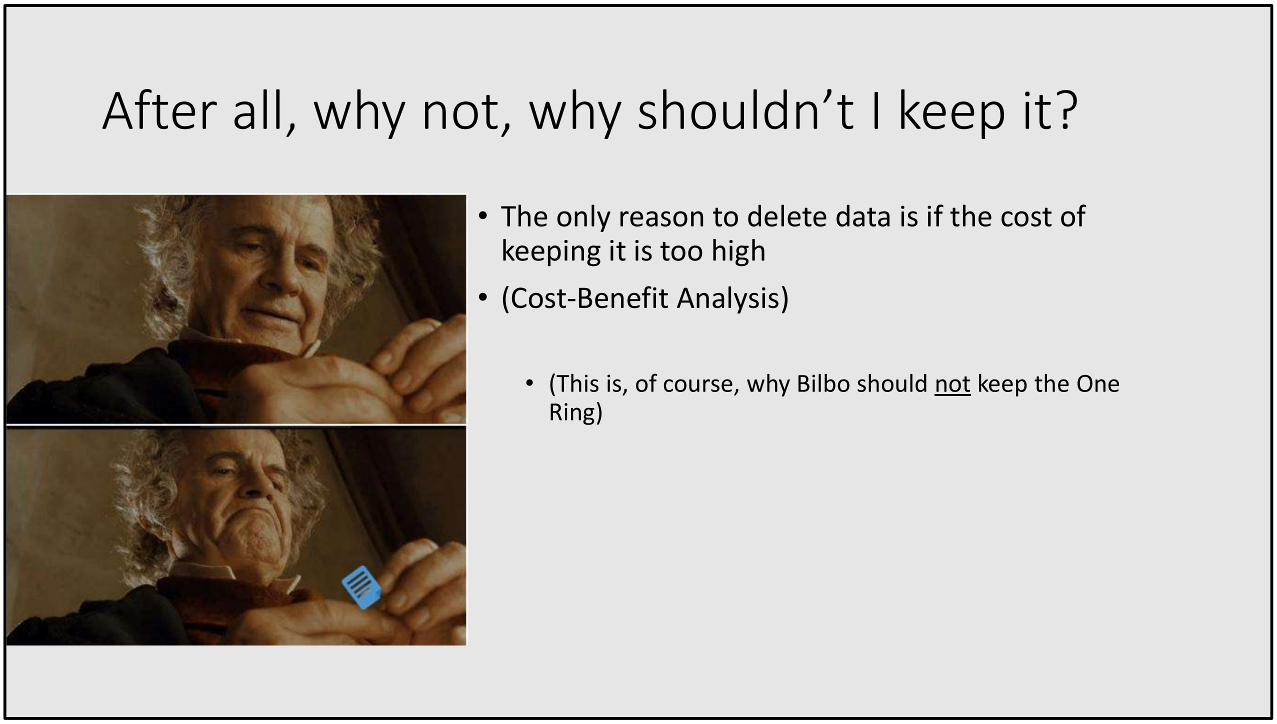
* Interested in the topic
* Comfortable with rapidly-evolving software

“Rapidly-Evolving” means “there might be version update issues and things might break – especially 431 where Python updates are my nemesis)

Big Data

- Question: Why are data so big these days?
- Answer: It's complicated

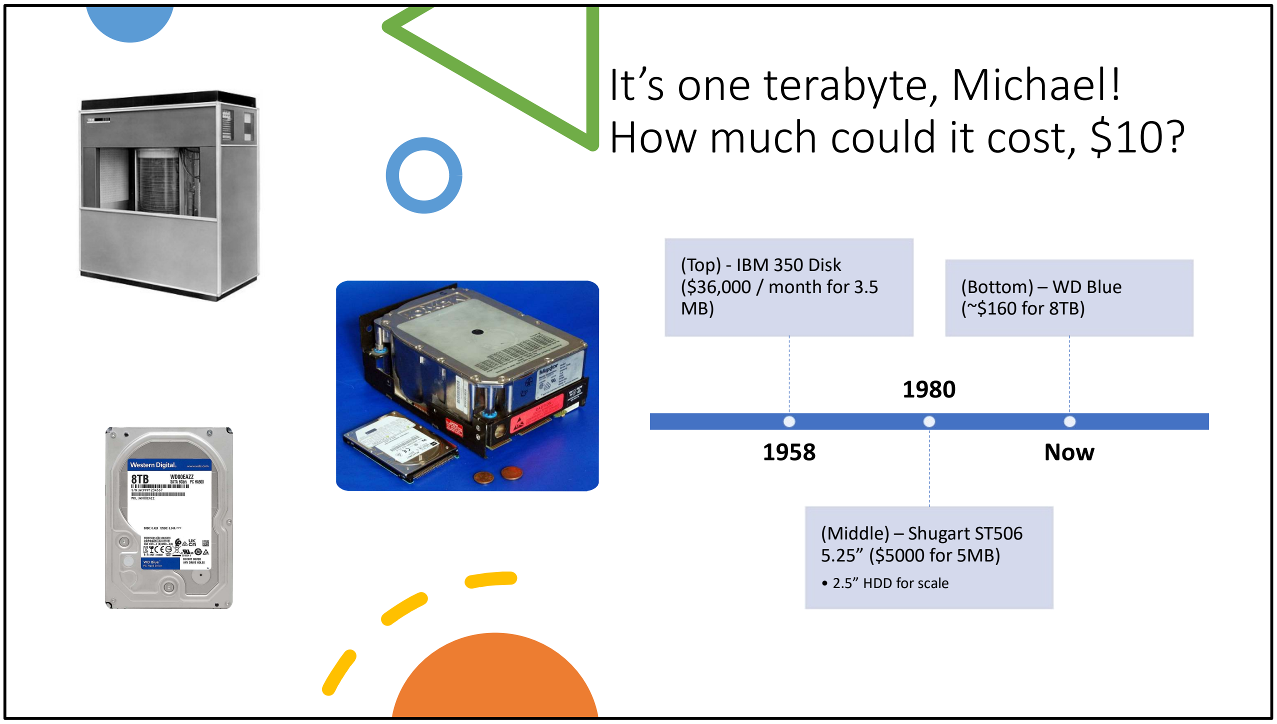




After all, why not, why shouldn’t I keep it?

- The only reason to delete data is if the cost of keeping it is too high
- (Cost-Benefit Analysis)
  - (This is, of course, why Bilbo should not keep the One Ring)

What’s going on here? The cheaper it is to store a GB, the most GB of data you’ll find that are “worth it” to retain.
(In other words, data expands to fill its tank)



It’s one terabyte, Michael! How much could it cost, $10?

(Top) - IBM 350 Disk ($36,000 / month for 3.5 MB)

(Bottom) – WD Blue (~$160 for 8TB)

1980

1958

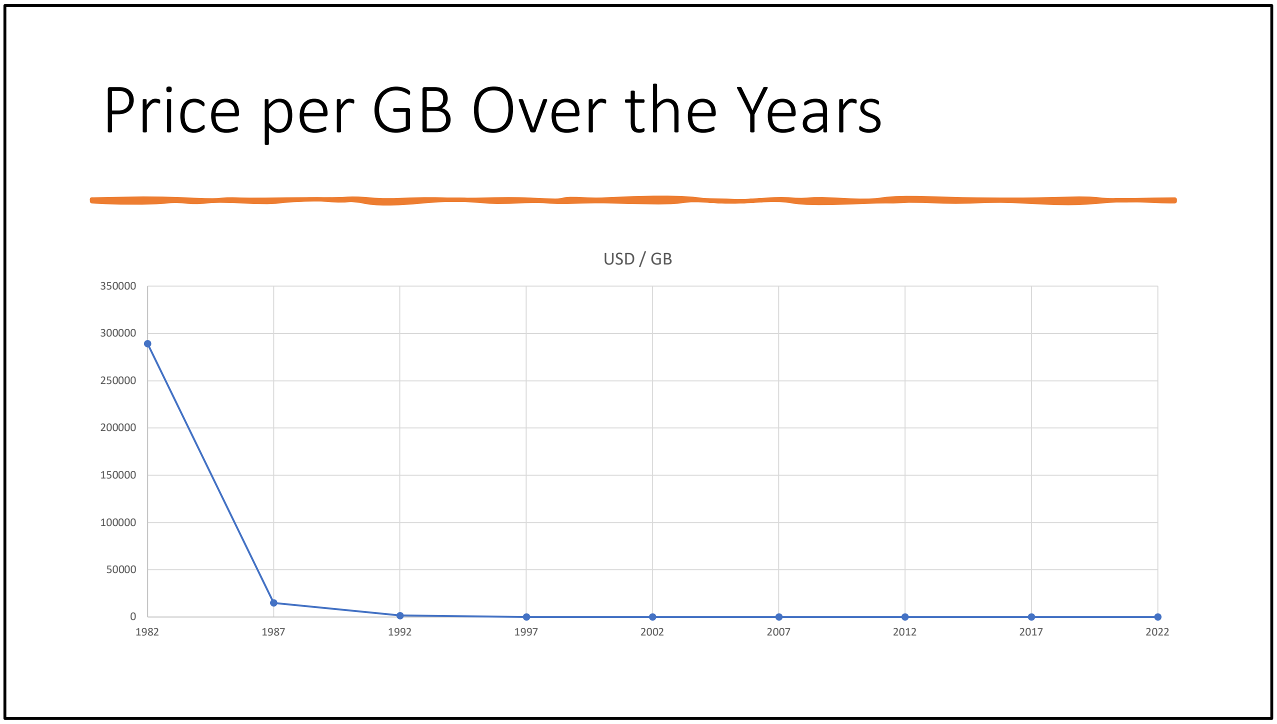
Now

(Middle) – Shugart ST506 5.25” ($5000 for 5MB)

- 2.5” HDD for scale

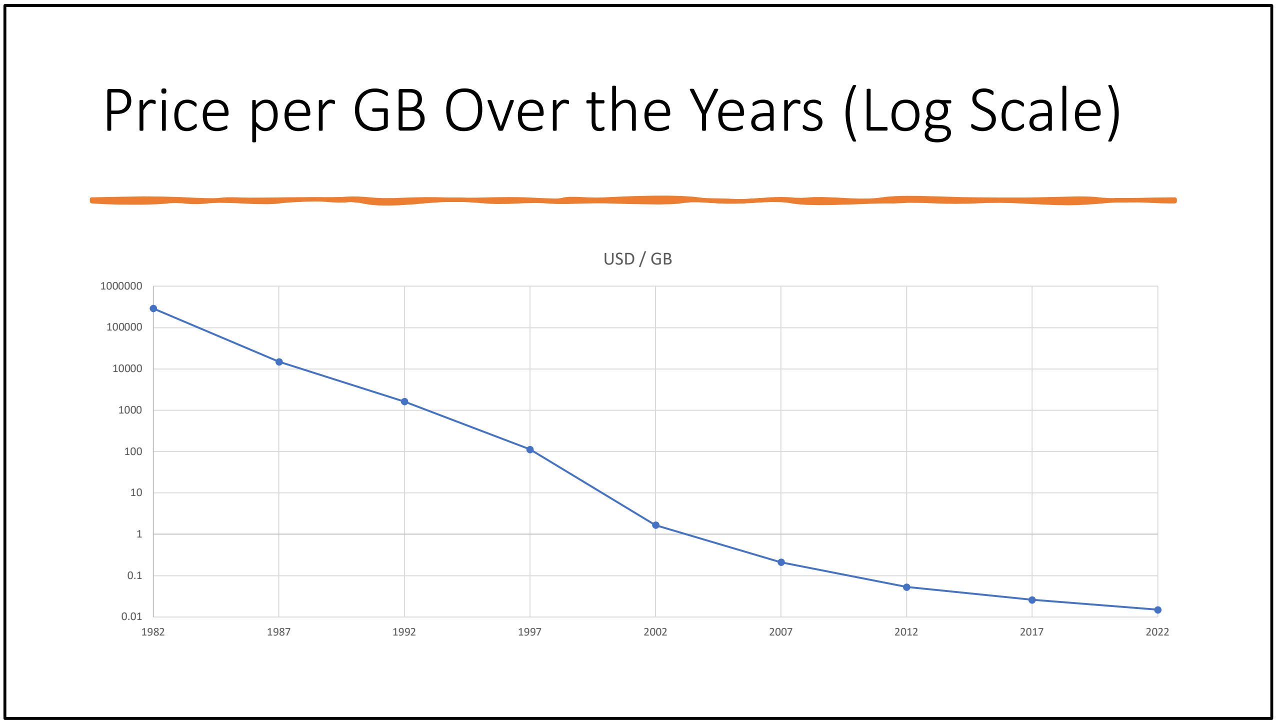
($10CDN / TB is actually pretty close to true now, at least for used enterprise grade gear)

(Aug 14 2023) the WD Blue 8TB is the best GB/$ on amazon.ca if you exclude used. If you don’t, a 4TB WD enterprise grade Refurb is only $60!
(Sep 3 2024) Now the best GB/$ is the Seagate Barracuda 8TB which is $145 on Amazon, or a 12TB HGST Ultrastar $131 refurbished.



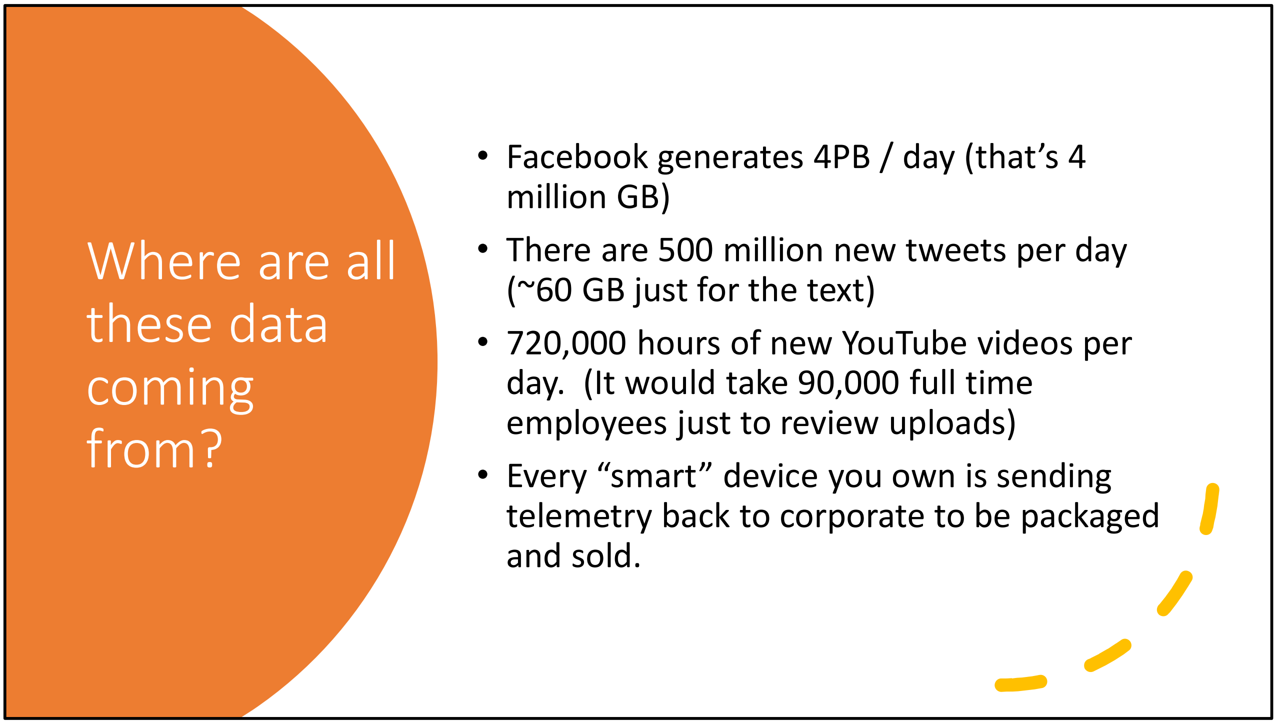
Price per GB Over the Years

Never have a graph that looks like this. If you find that you do, use a log scale! Please!



Price per GB Over the Years (Log Scale)

That’s more like it! The change in slope in the 1997-2002 is the “dot com boom” : economies of scale

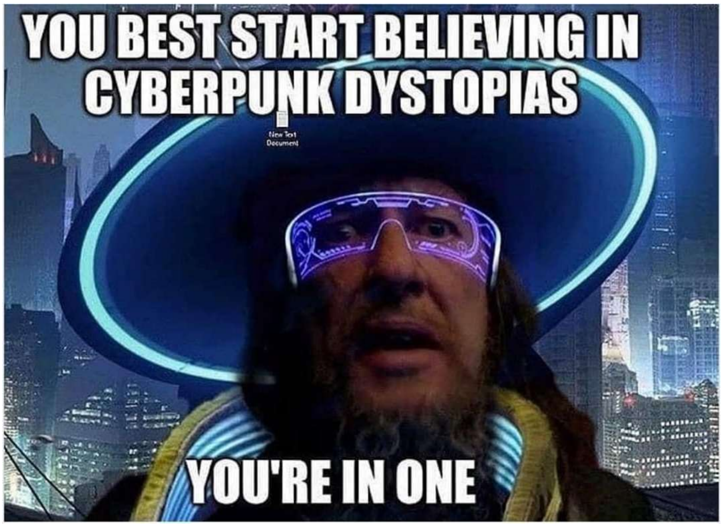


Where are all these data coming from?

- Facebook generates 4PB / day (that’s 4 million GB)
- There are 500 million new tweets per day (~60 GB just for the text)
- 720,000 hours of new YouTube videos per day. (It would take 90,000 full time employees just to review uploads)
- Every “smart” device you own is sending telemetry back to corporate to be packaged and sold.

How much????

- Right now* we generate 400 exobytes (400 million TB) per day
- That’s 48GB / person every day
  - That’s about 500KB / second / person



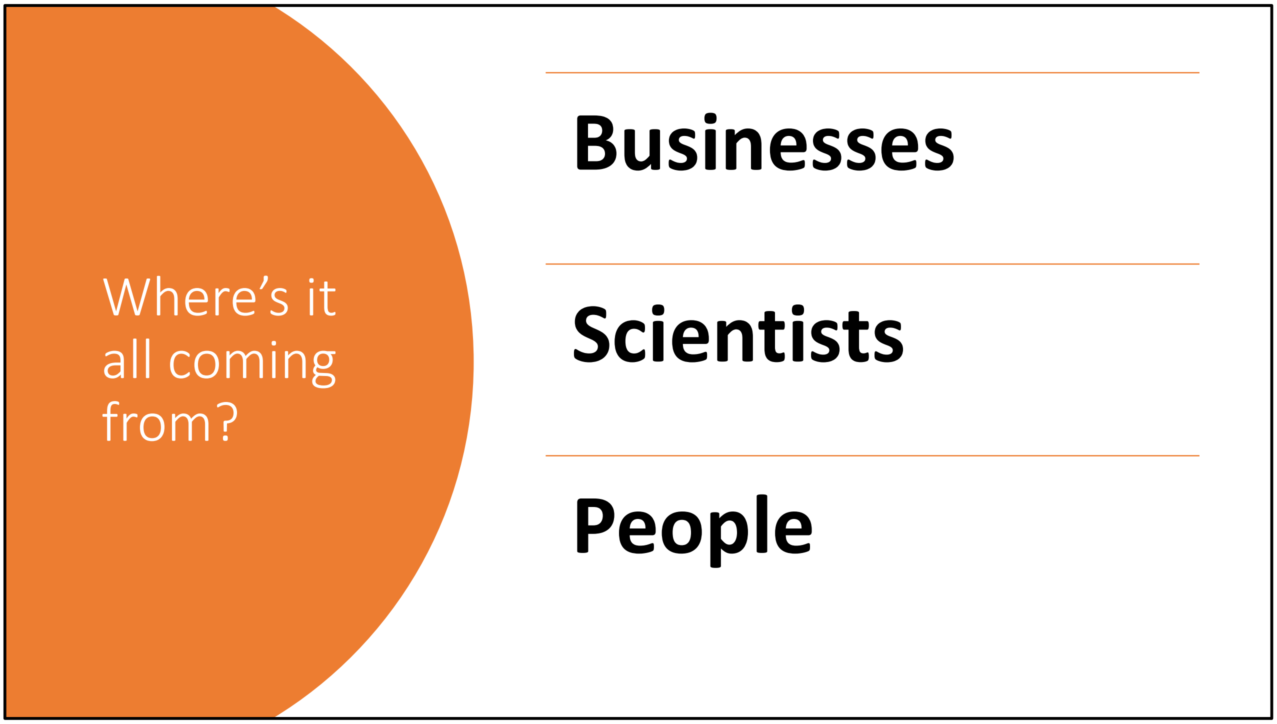
* As of sometime in 2024
https://explodingtopics.com/blog/data-generated-per-day

A lot of that is video so it’s all about averages

400 EXAbytes??? A DAY???

- That might seem like a lot, but it’s nothing compared to what it’s going to be
- Will be up to 500 exabytes / day in 2025 (125 million 4TB HDDs filled per day)

In 2020 the number wasn’t 400, it was 8. The predicted number for 2025 was 500 exabyte number was predicted in 2020 too, and we seem right on track.



Where's it all coming from?

Businesses

Scientists

People

That's not to say suits and scientists aren't people…

Business Data

DATA-DRIVEN DECISION-MAKING

DATA-DRIVEN PRODUCT DESIGN

TARGETED ADVERTISING

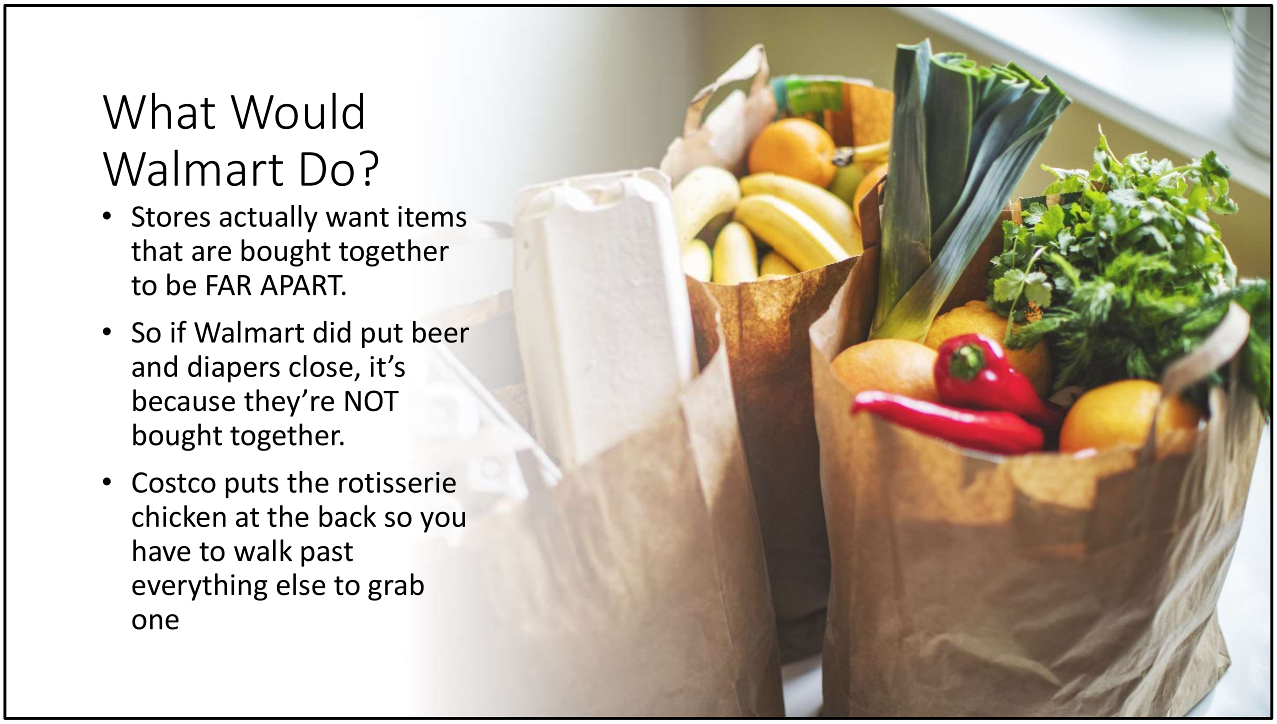
Business Intelligence

- "What worked? What didn't?"
- This isn't a new concept.

Anecdote!

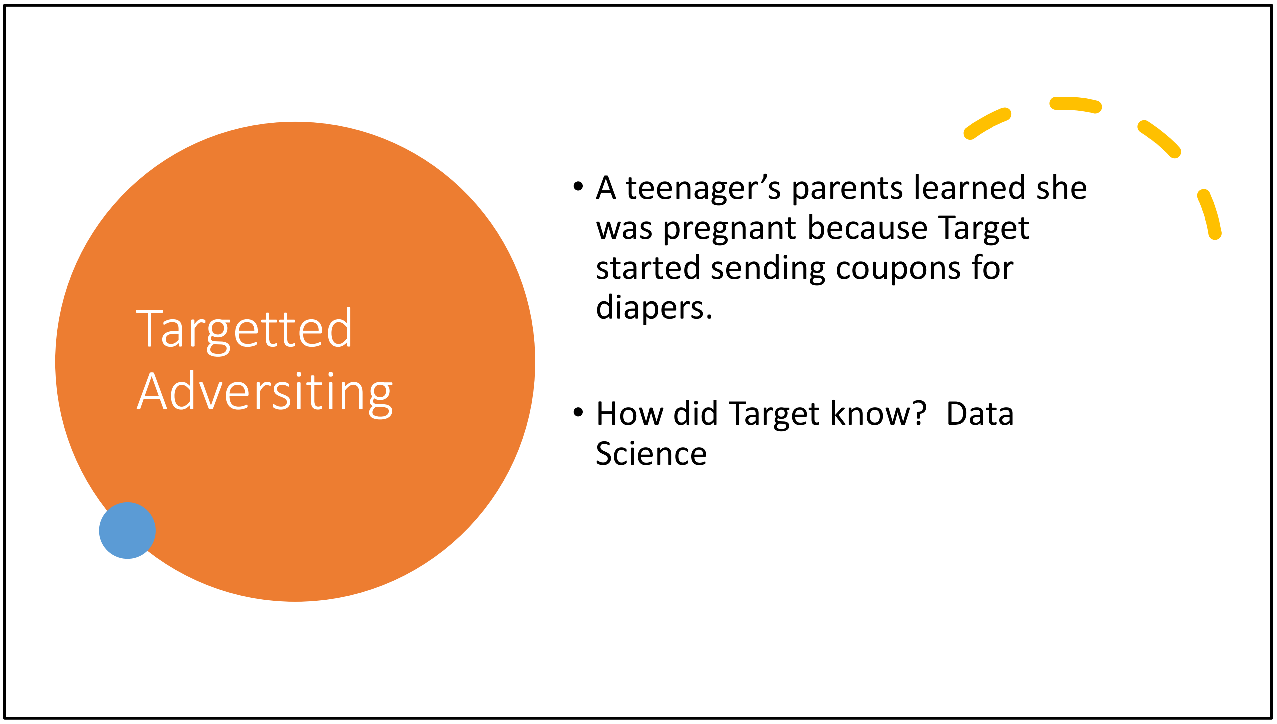
- In the 1990s, Walmart Discovered people tend to buy beer and diapers at the same time, so they put them together.

- PS this isn’t true. Anecdotes rarely are.



What Would Walmart Do?

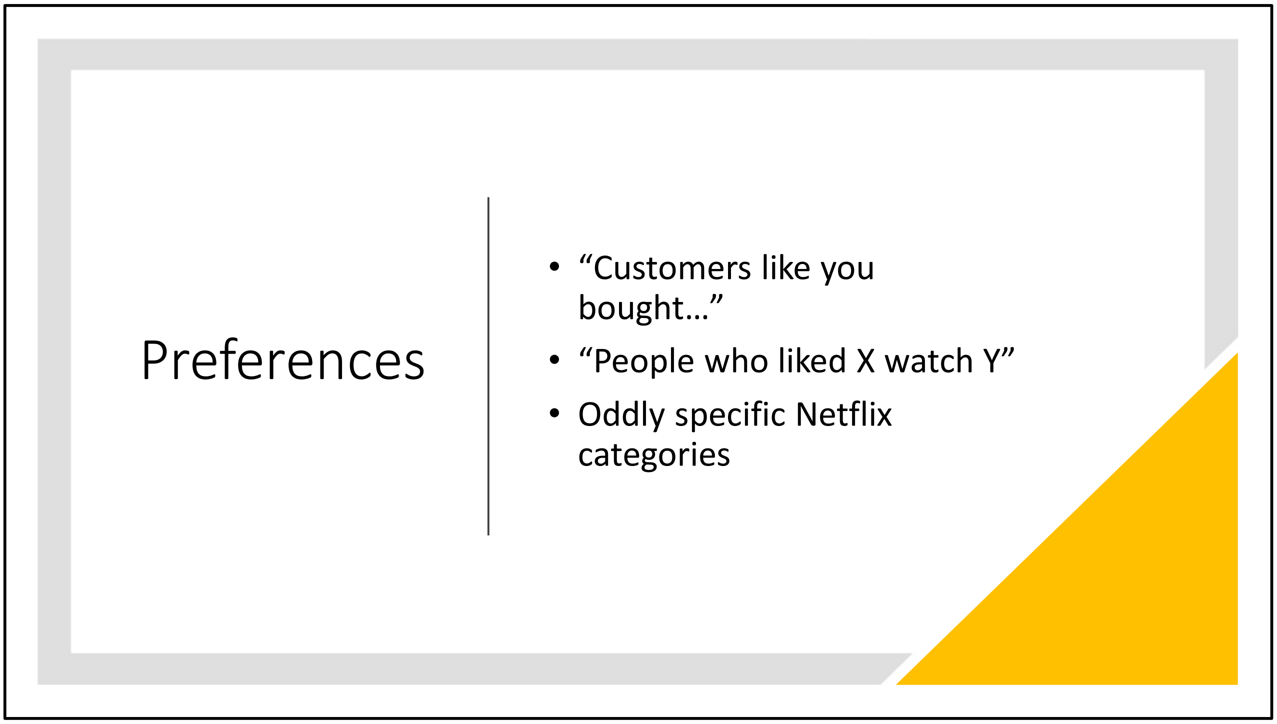
- Stores actually want items that are bought together to be FAR APART.
- So if Walmart did put beer and diapers close, it’s because they’re NOT bought together.
- Costco puts the rotisserie chicken at the back so you have to walk past everything else to grab one



Targetted Adversiting

- A teenager’s parents learned she was pregnant because Target started sending coupons for diapers.
- How did Target know? Data Science

Buying prenatal vitamins is somewhat a strong hint – less strong hint: switching to unscented soaps and lotions. Morning sickness is easily triggered by smells, or so my wife tells me.



Preferences

- "Customers like you bought…"
- "People who liked X watch Y"
- Oddly specific Netflix categories

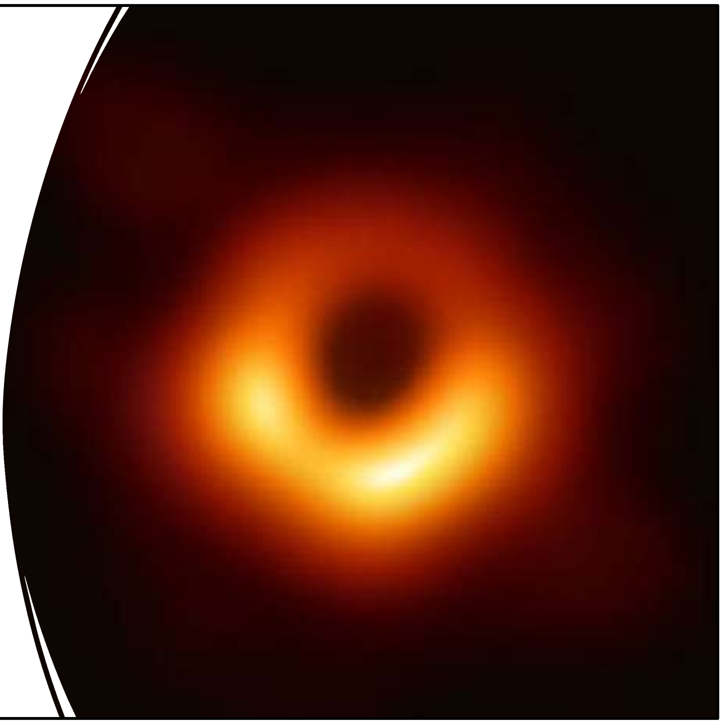
Dark Suspenseful Viral Plague Movies,

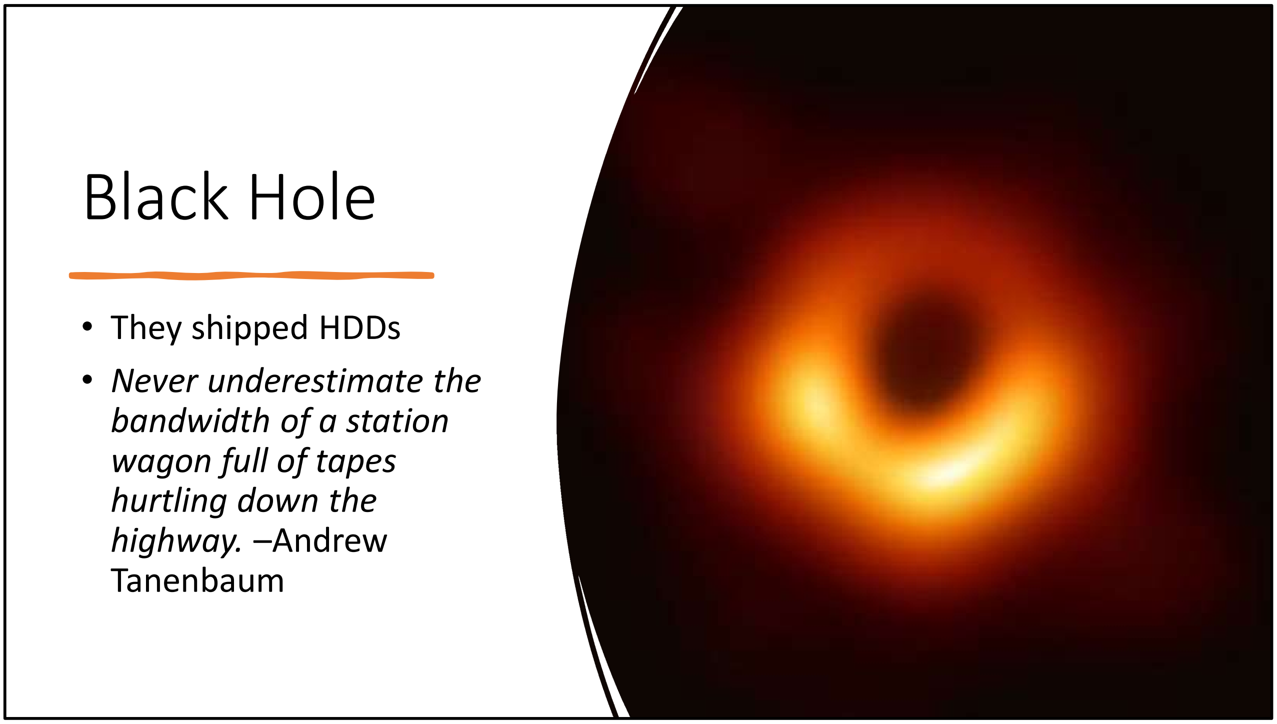
Science!

- Data-Intensive eScience
- Modern Experiments generate BIG DATA

Black Hole

- First Image of a Black Hole (2019)
- 4.5PB of data from 8 telescopes





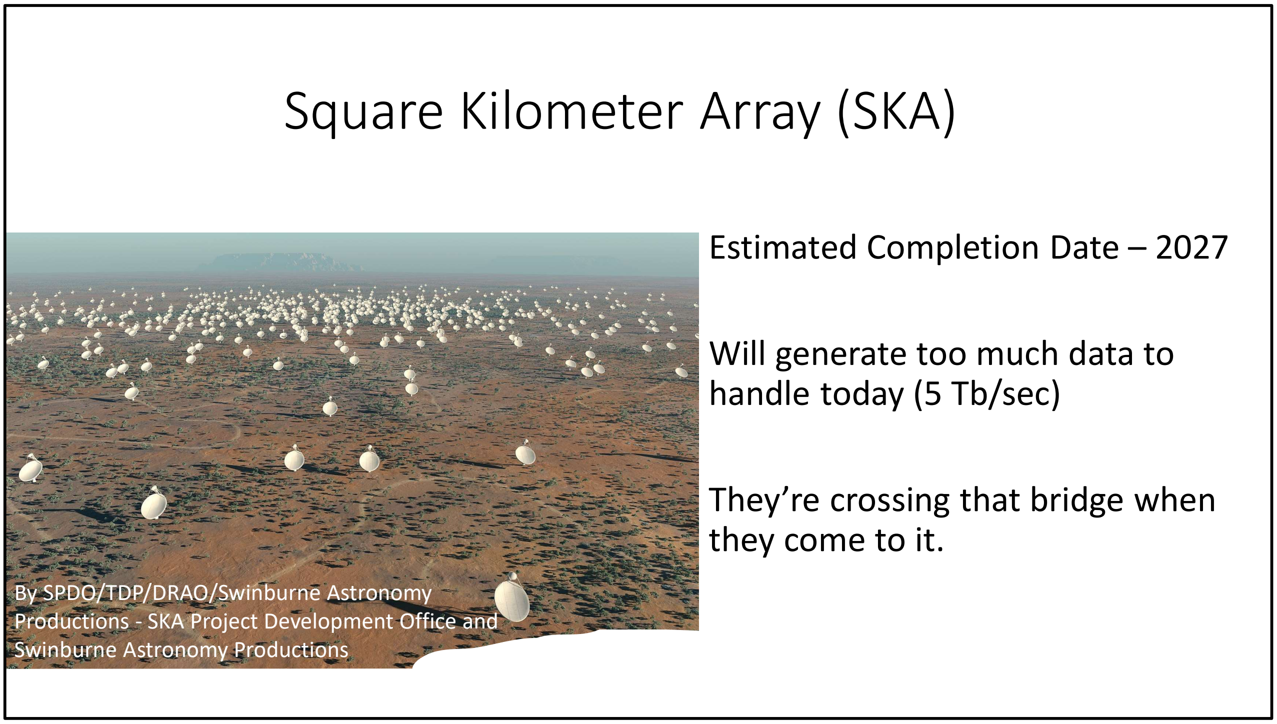
Black Hole

- They shipped HDDs
- *Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway.* –Andrew Tanenbaum

They shipped HDDs by plane, train, and automobile…err..truck. Would have taken years to send over the internet.

JAMES WEBB TELESCOPE

- JWST sends 57GB / day back to Earth
- One pretty picture requires MANY images stitched together

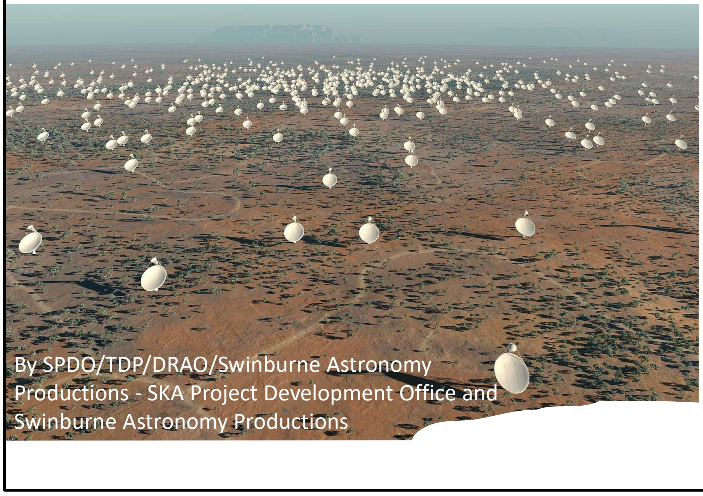


Square Kilometer Array (SKA)

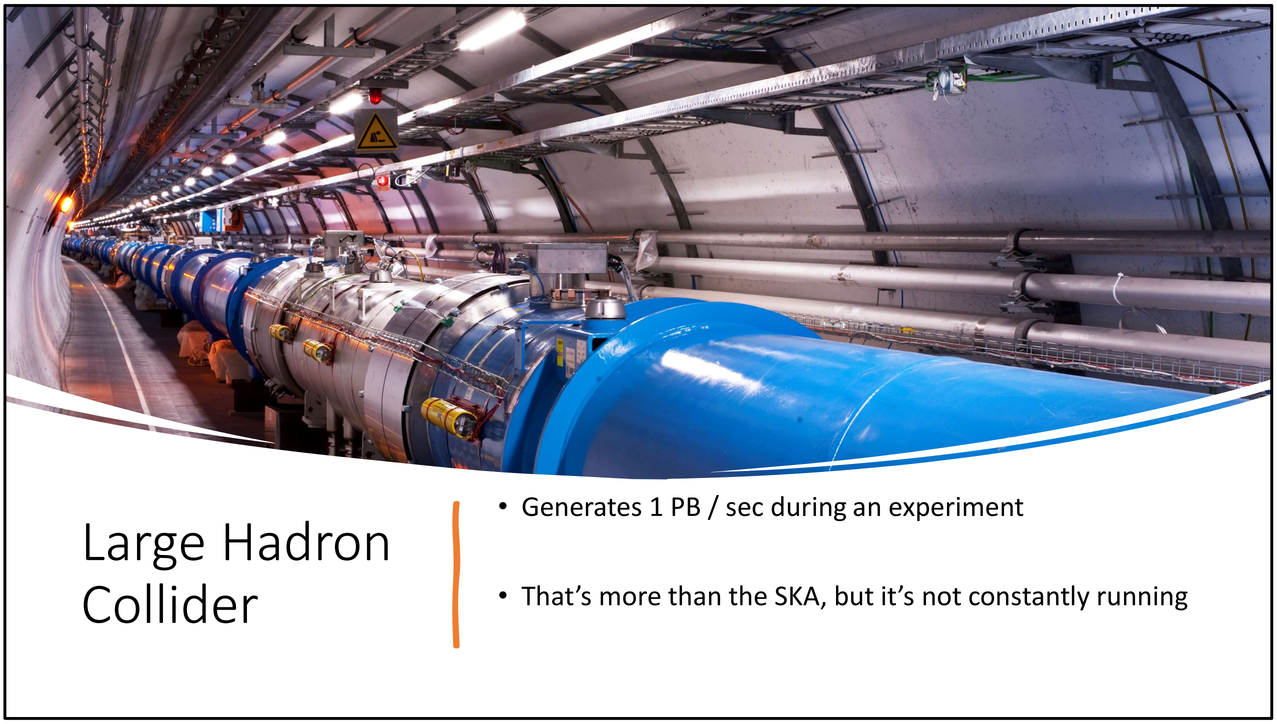
Estimated Completion Date – 2027

Will generate too much data to handle today (5 Tb/sec)

They're crossing that bridge when they come to it.

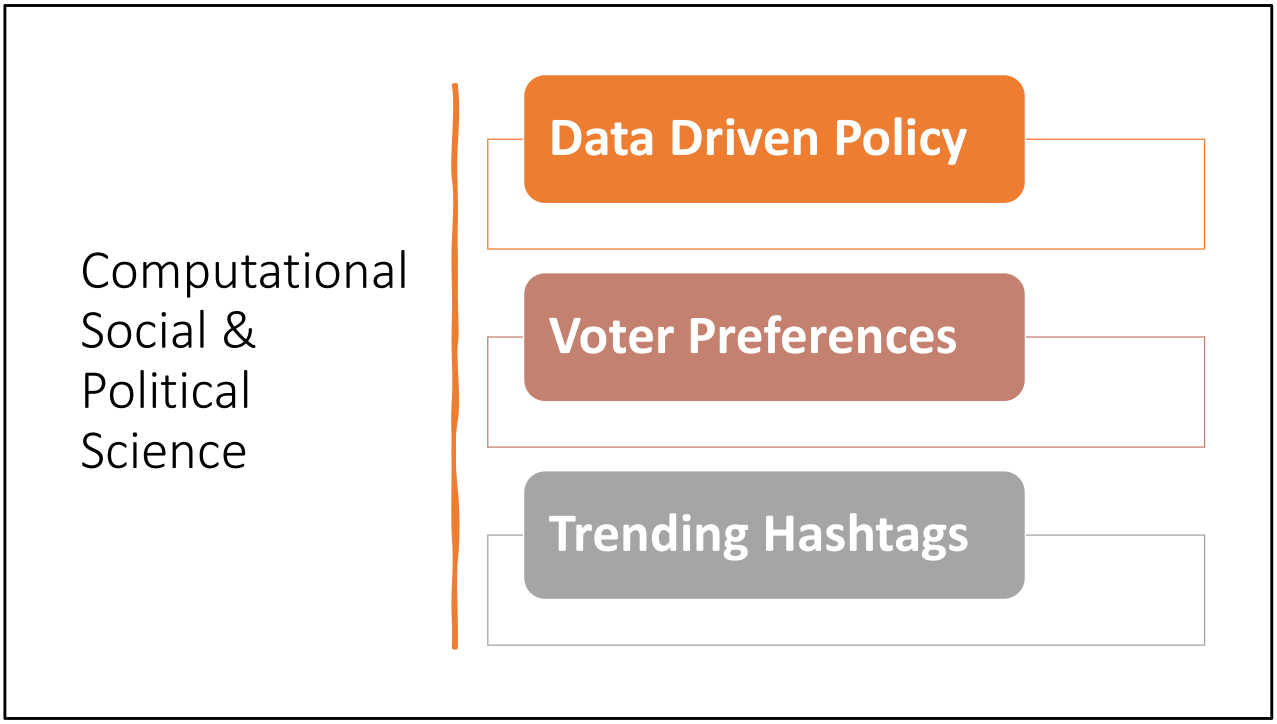


By SPDO/TDP/DRAO/Swinburne Astronomy Productions - SKA Project Development Office and Swinburne Astronomy Productions



Large Hadron Collider

- Generates 1 PB / sec during an experiment
- That’s more than the SKA, but it’s not constantly running



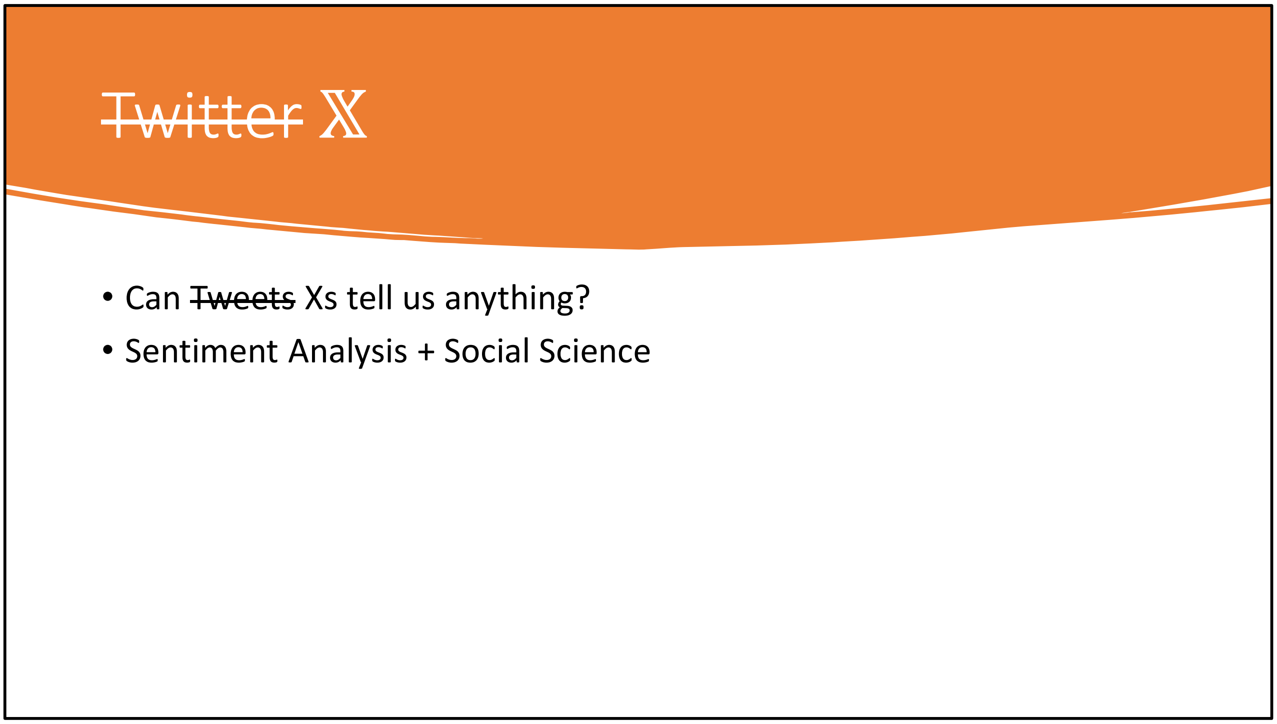
Computational Social & Political Science

- Data Driven Policy
- Voter Preferences
- Trending Hashtags

Humans as Sensors

Humans record their thoughts on social media.

What can we do with all those data?



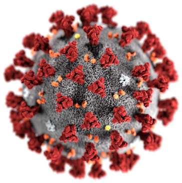
~~Twitter~~ 𝕏

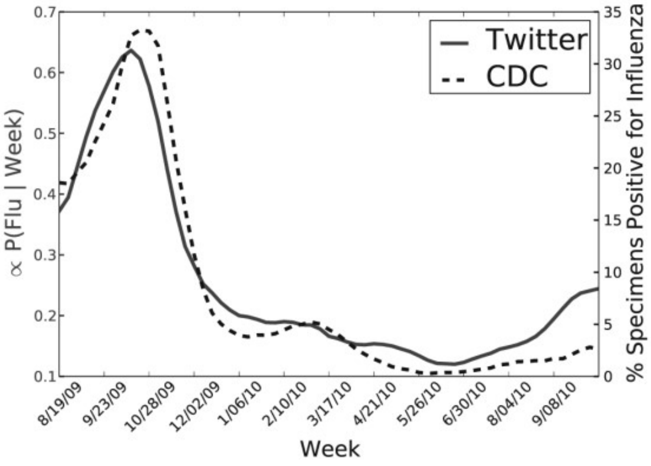
- Can ~~Tweets~~ Xs tell us anything?
- Sentiment Analysis + Social Science

I will never stop calling it Twitter. It’s Twitter.

Predicting *X* with ~~Twitter~~ 𝕏

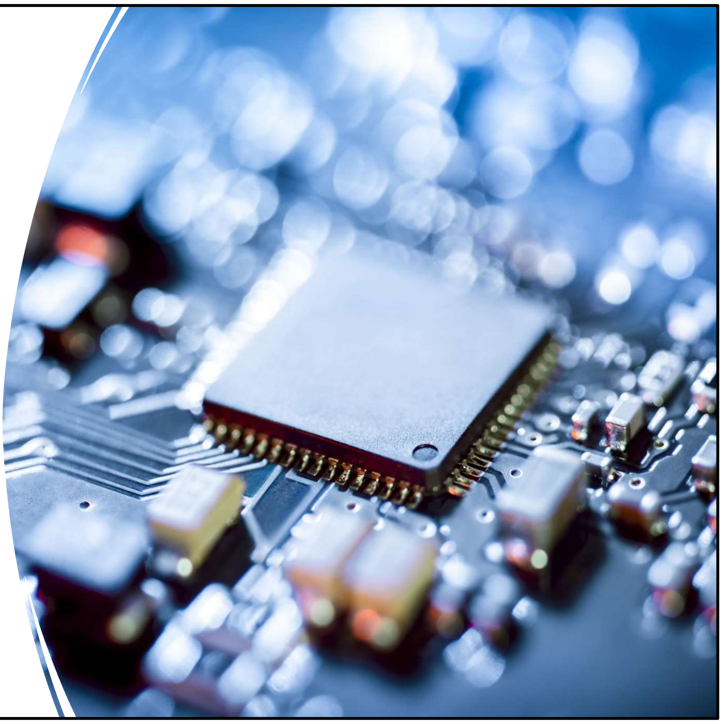
Fall 2020 Project : Predicting COVID with Twitter





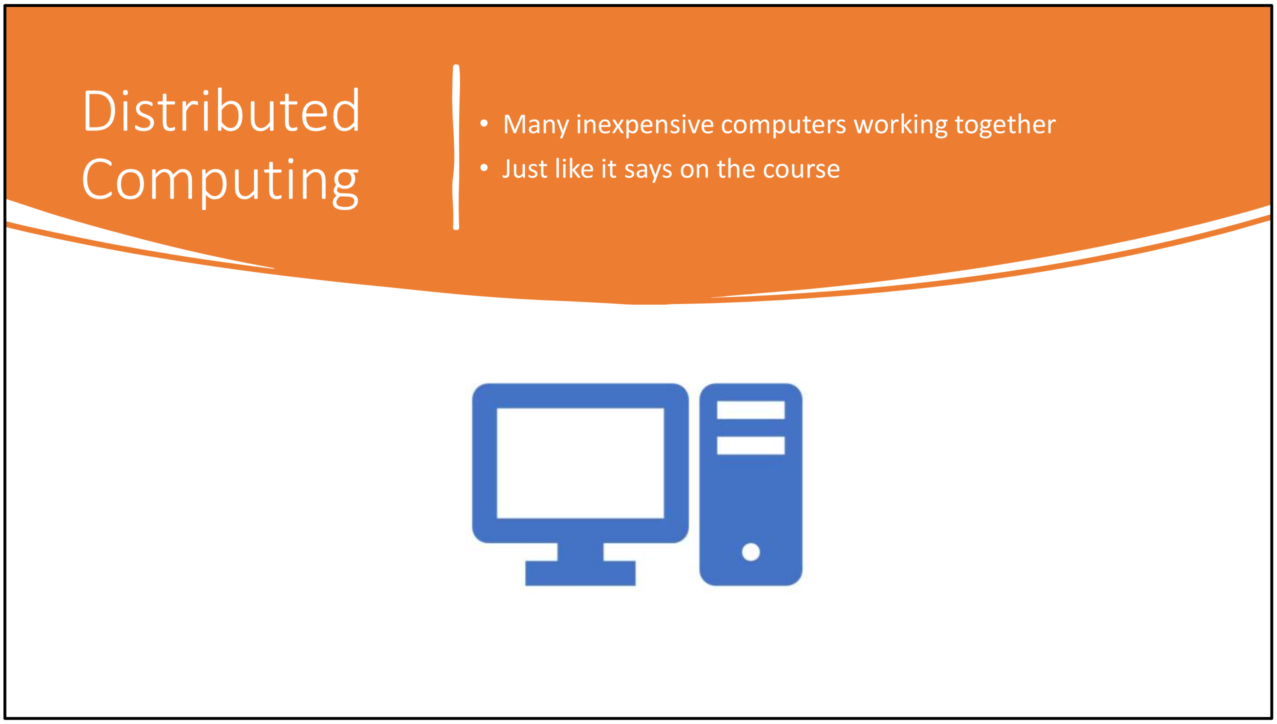
Big Data, Big Computer?

- Vertical Scaling – More RAM, Disk, CPU
- Return of the Mainframe?
- Expensive!
- Limited!



Big Data, Big Network!

- Horizontal Scaling
- Cheap computers, just more of them

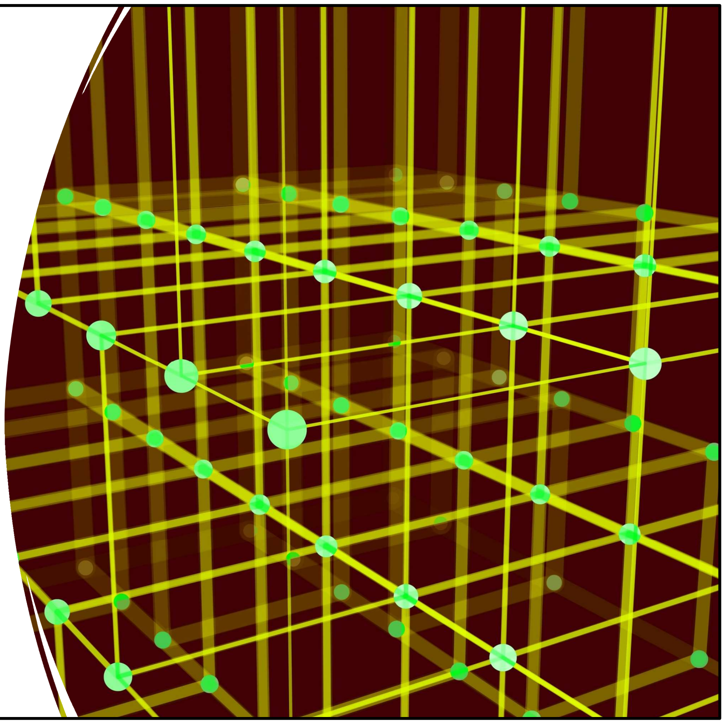


Distributed Computing

- Many inexpensive computers working together
- Just like it says on the course

Parallelization is hard

- Deadlocks, Livelocks, Race Conditions, oh my!
- That’s just on one computer. What if they’re remote?

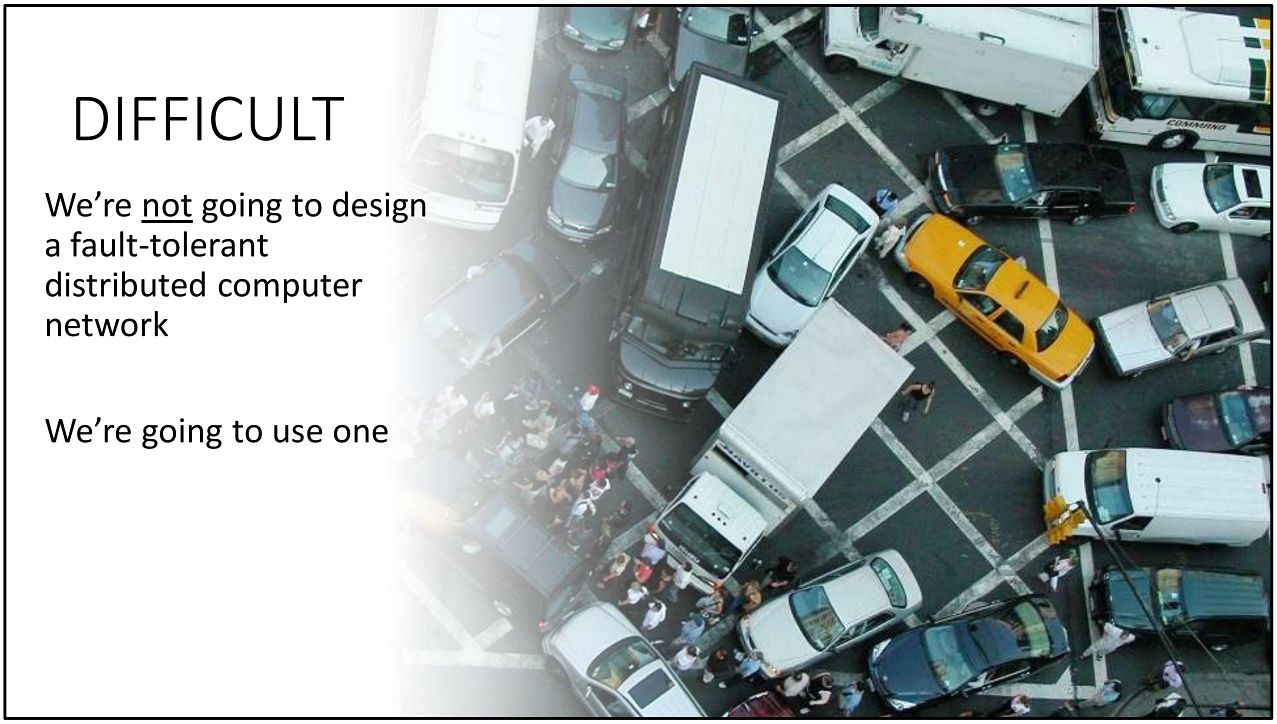


Scaling Out!

- A datacenter of many machines?
- Many datacenters???
- Fault tolerance

ALL HARDWARE FAILS

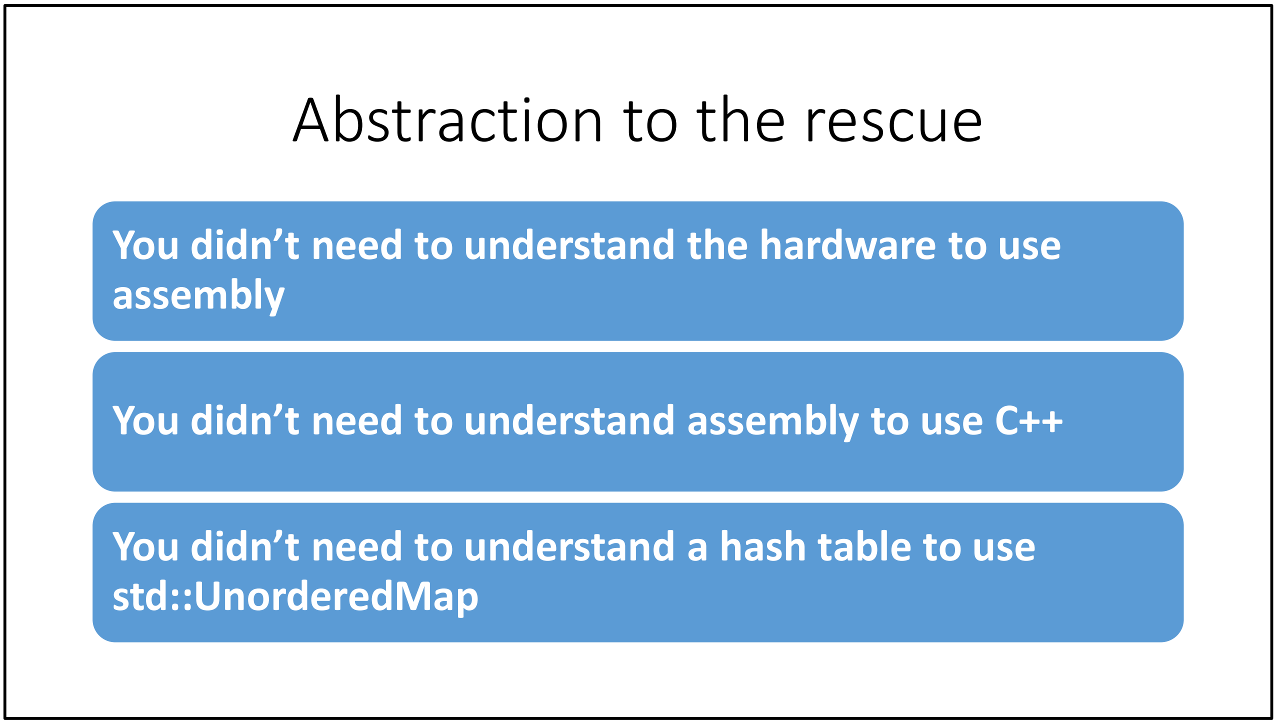
I'VE NEVER HAD A COMPUTER FAIL
YOU'VE NEVER HAD A COMPUTER FAIL SO FAR
imgflip.com



DIFFICULT

We’re not going to design a fault-tolerant distributed computer network

We’re going to use one

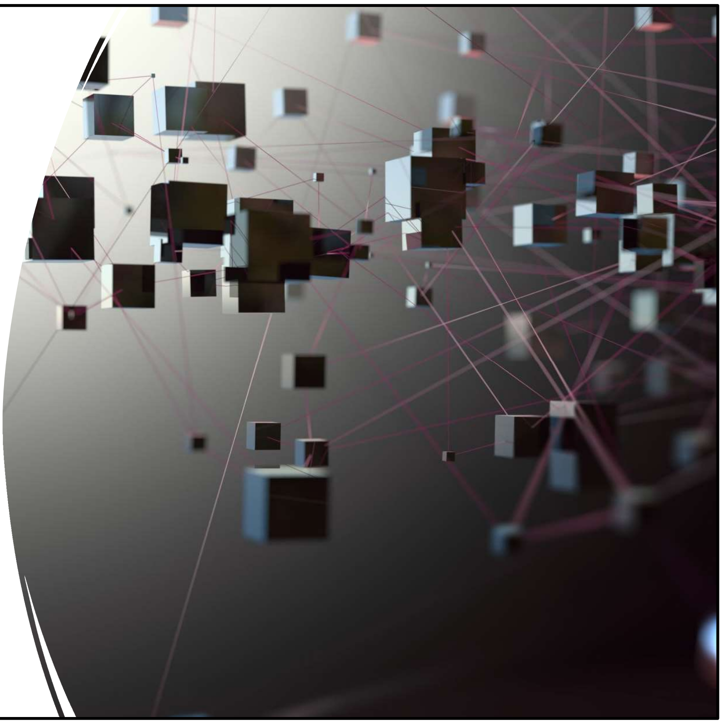


Abstraction to the rescue

- You didn’t need to understand the hardware to use assembly
- You didn’t need to understand assembly to use C++
- You didn’t need to understand a hash table to use std::UnorderedMap

What’s the Next Layer?

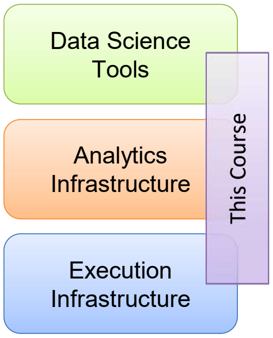
- How can we abstract a distributed network?
- (That’s the topic of the next few lectures)



What’s CS431/CS451?

A little helping of

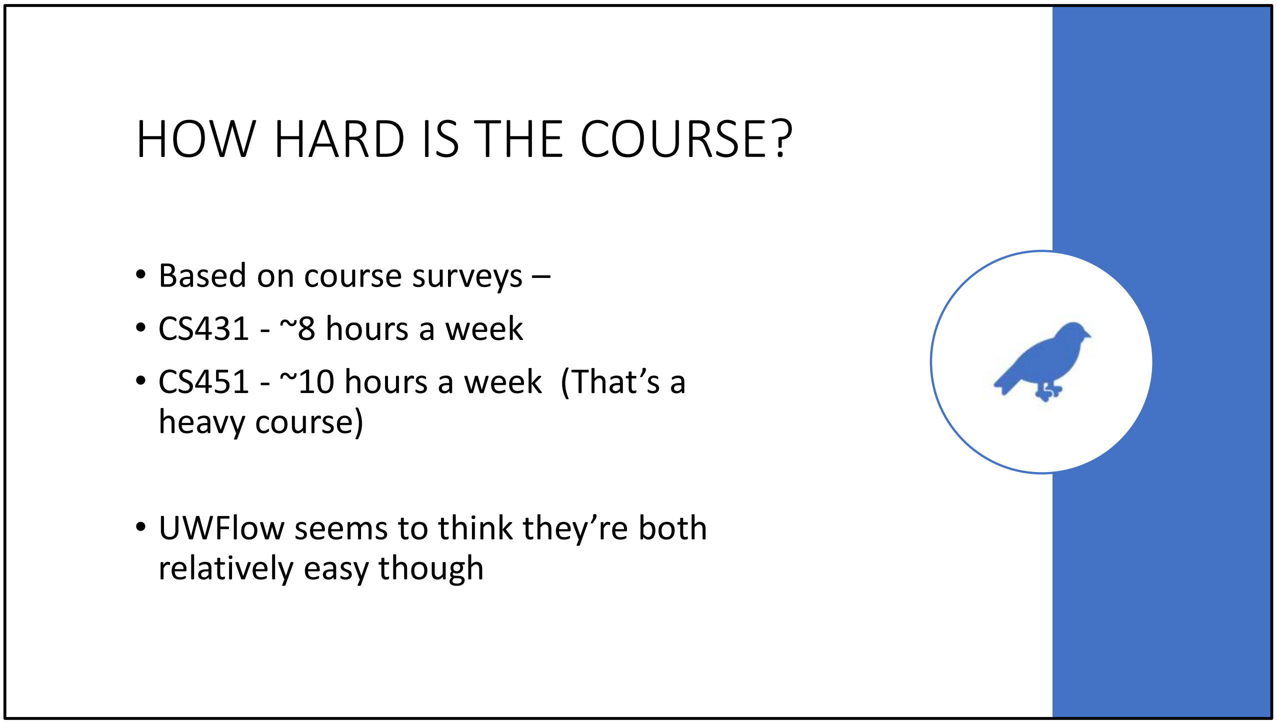
- **Data Science**
- **Distributed Analytics**
- **Distributed Execution**



More Buzzwords Please!

You got it!

- Analytics
- Business intelligence
- Data warehousing
- MapReduce, Hadoop, Spark, Pig, Hive, NoSQL, Pregel, Giraph, Storm/Heron
- Thinking at scale

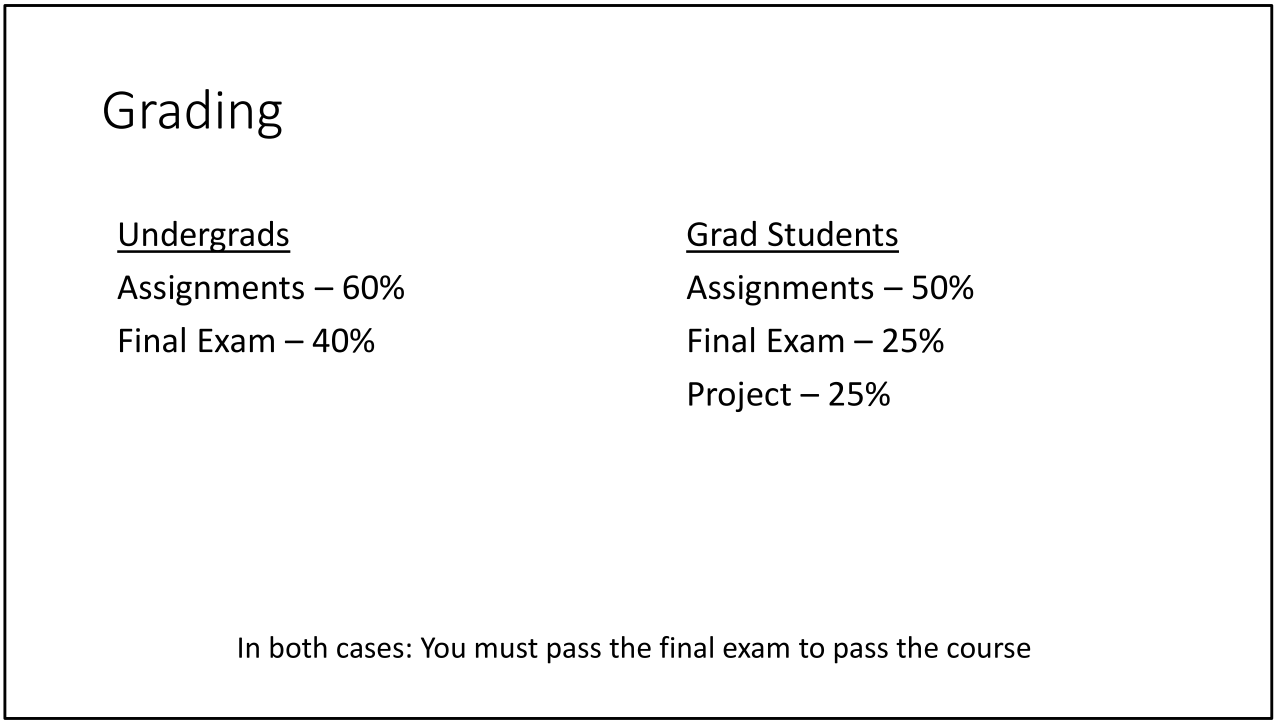


HOW HARD IS THE COURSE?

- Based on course surveys –
- CS431 - ~8 hours a week
- CS451 - ~10 hours a week (That's a heavy course)

- UWFlow seems to think they're both relatively easy though

The assignments are heavy for CS451 / 651. That doesn't per se mean they are challenging, but that some of them are long. Honestly most of the length is due to learning the tools and frameworks involved, which is good, that's the point of having assignments!



Grading

Undergrads	Grad Students
Assignments – 60%	Assignments – 50%
Final Exam – 40%	Final Exam – 25%
	Project – 25%

In both cases: You must pass the final exam to pass the course

Don’t worry, the exam average is usually around 70 and failing it is rare!

Course Info and Help

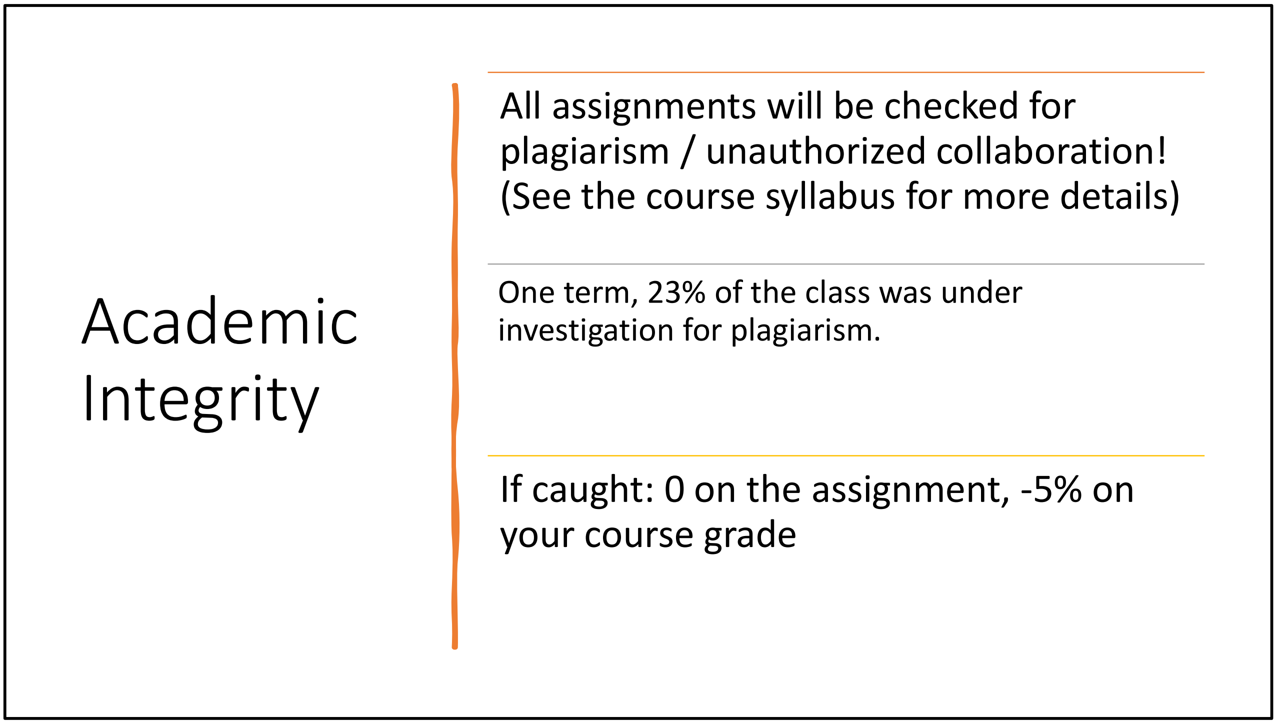
Course Website: https://www.student.cs.uwaterloo.ca/~cs451

(Yes, even if you're in CS431)

Piazza (you should have been emailed an invite)

Online Office Hours: Microsoft Teams

In-Person Office Hours: See website.



Academic Integrity

All assignments will be checked for plagiarism / unauthorized collaboration! (See the course syllabus for more details)

One term, 23% of the class was under investigation for plagiarism.

If caught: 0 on the assignment, -5% on your course grade

Assignment Mechanics (CS451/651)

Java

Scala ⚠

We'll be using <u>private</u> Git repos for assignments

Complete your assignments, push to GitLab
We'll pull your repos at the deadline and grade

Late assignments will get 0

Assignment Mechanics (CS431)

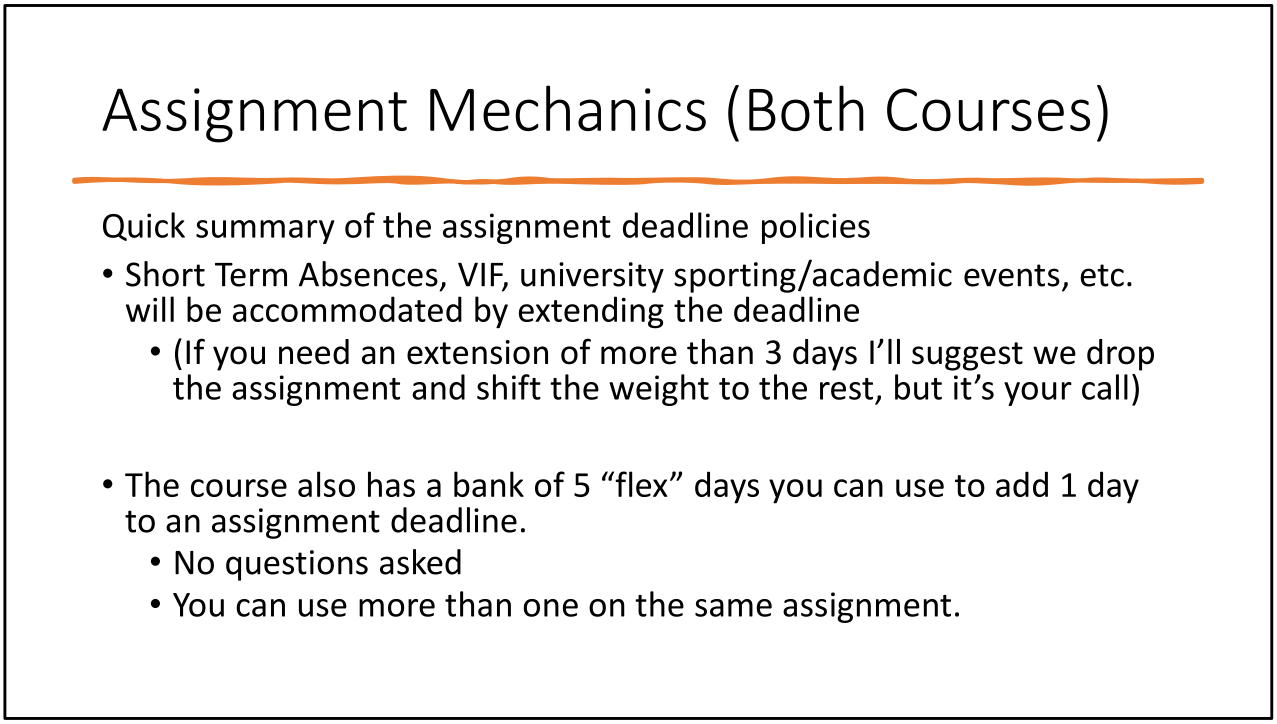
Assignments will use Python and Jupyter (Google Colab)
Everything you need to know is in the assignment itself

Assignments will be submitted using private Git repositories
Details are on the course website for the appropriate assignment

APACHE Spark

Python

Late assignments will get 0



Assignment Mechanics (Both Courses)

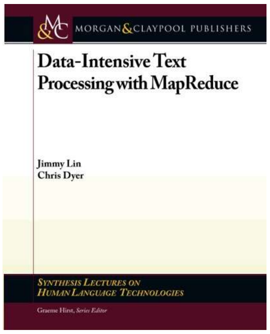
Quick summary of the assignment deadline policies

- Short Term Absences, VIF, university sporting/academic events, etc. will be accommodated by extending the deadline
  - (If you need an extension of more than 3 days I'll suggest we drop the assignment and shift the weight to the rest, but it's your call)

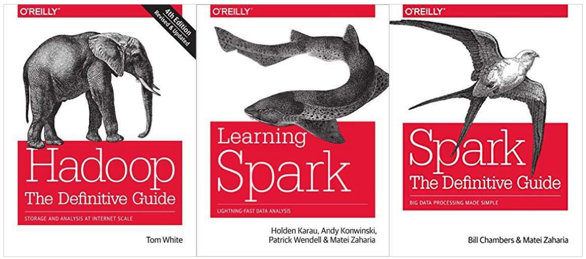
- The course also has a bank of 5 "flex" days you can use to add 1 day to an assignment deadline.
  - No questions asked
  - You can use more than one on the same assignment.

Course Materials

One (required) textbook +
Three (optional but recommended) books +
Additional readings from other sources as appropriate


MORGAN & CLAYPOOL PUBLISHERS
Data-Intensive Text Processing with MapReduce
Jimmy Lin
Chris Dyer
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES
Graeme Hirst, Series Editor

(optional but recommended)


O'REILLY
4th Edition
Revised & Updated
Hadoop
The Definitive Guide
STORAGE AND ANALYSIS AT INTERNET SCALE
Tom White
O'REILLY
Learning
Spark
LIGHTNING-FAST DATA ANALYSIS
Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia
O'REILLY
Spark
The Definitive Guide
BIG DATA PROCESSING MADE SIMPLE
Bill Chambers & Matei Zaharia

Note: 4th Edition