CMPT 733
Big Data Programming II

Instructors Steven Bergner and Zhengjie Miao

Course website https://coursys.sfu.ca/2025sp-cmpt-733-g1/pages/

CMPT 733
~~Big Data Programming II~~
Data Science

Instructors Steven Bergner and Zhengjie Miao

Course website https://coursys.sfu.ca/2025sp-cmpt-733-g1/pages/

Instructor

Steven Bergner [he/him]

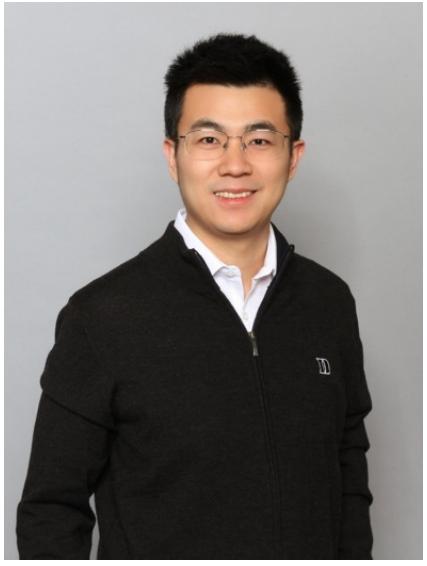
PhD, Computing Science | Simon Fraser University

Term Lecturer | Computing Science

Director of Data and Analytics | SFU's Big Data Hub

10+ years of research and working experience in scientific visualization, machine learning, and data science

Instructor



Zhengjie Miao [he/him]

PhD, Computing Science | Duke University

Assistant Professor | Computing Science | Simon Fraser University

Assistant Professor at SFU specializing in data management, AI, and human-data interaction. Innovator of algorithms and tools that enhance data science workflows.

MPCS Remote Teaching Survey

55 Responses

Questions	Yes
Are you satisfied with **lab courses**?	91%
Are you satisfied with **co-op office support?**	87%
Are you satisfied with **academic team support**?	91%
Do you **feel a sense of community** in your cohort?	72%
Are you happy with your decision to **pursue the program**?	93%

MPCS Remote Teaching Survey

55 Responses

Questions	Yes
Are you satisfied with **lab courses**?	91%
Are you satisfied with **co-op office support?**	87%
Are you satisfied with **academic team support**?	91%
Do you **feel a sense of community** in your cohort?	72%
Are you happy with your decision to **pursue the program**?	93%

1. Discord workspace
2. In-Class Discussion
3. Group Assignments

Outline

What is Data Science?

Data Science Lifecycle

4 Questions Data Scientists Can Answer

The “Data Science” term: buzzword?

Course Structure

What is Data Science?

Computer Science vs. Data Science

What	When	Who	Goal
Computer Science	1950-	Software Engineer	Write software to make computers work

Plan ➔ Design ➔ Develop ➔ Test ➔ Deploy ➔ Maintain

What	When	Who	Goal
Data Science	2010-	Data Scientist	Extract insights from data to answer questions

Collect➔ Clean ➔ Integrate ➔ Analyze ➔ Visualize ➔ Communicate

New possible trend: AI-enhanced Data Science ?

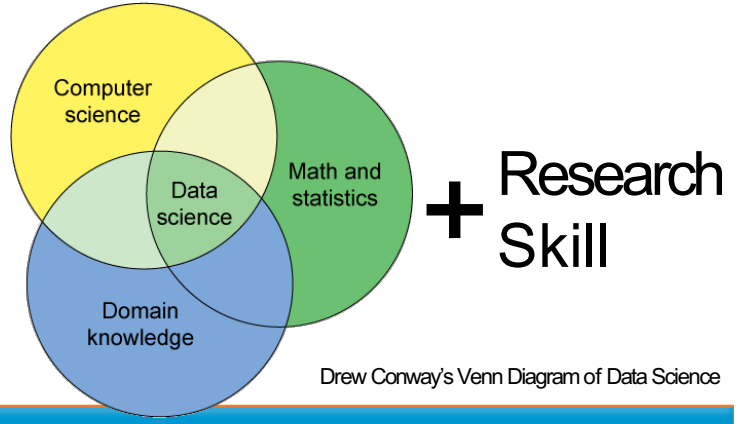
New Skillset

Example Questions

- How popular will this new product be? (Predictive Model)
- Which features should be added? (A/B Testing)
- Who are the potential customers? (Recommendation System)
- ...

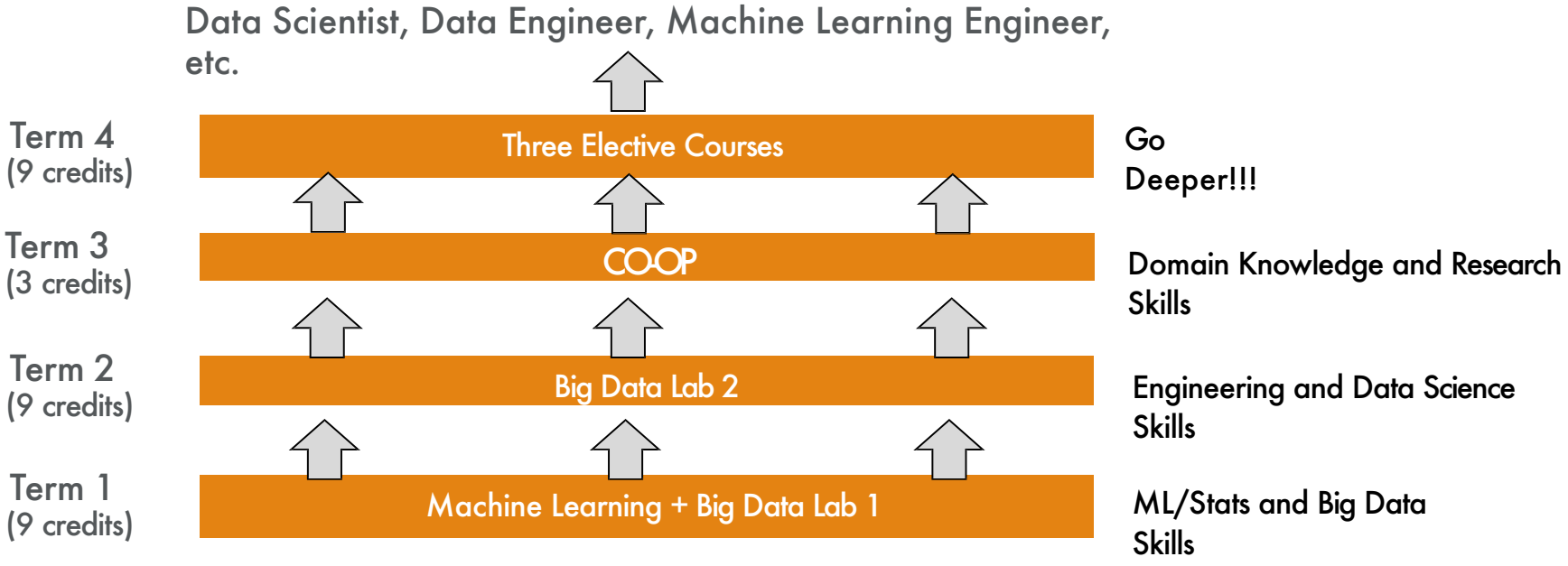
What skills are needed to answer these questions?

- Programming Skills
- Machine Learning/Statistics
- Domain Knowledge



Drew Conway's Venn Diagram of Data Science

SFU MPCS Big Data Curriculum



Data Science Lifecycle

Data Science Lifecycle (High-Level)

The entire workflow is **iterative**

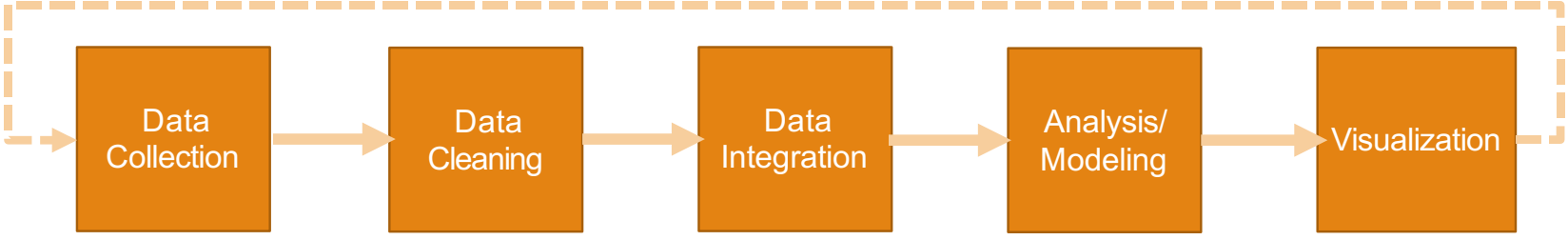
Two ways to produce questions

- Start with questions and then collect the related data
- Start with data and then think about the questions that can be answered

Data Processing Pipeline

What you think you do?

What you really do?



4 Questions Data Scientists Can Answer

https://thelead.io/data-science/what-types-of-questions-can-data-science-answer/

Is This A or B?

Classification Algorithms

Examples

- Is this an image of a cat or a dog?
- Will this customer renew their subscription?
- Will this tire fail in the next thousand miles?

1. Which company do you work at?
2. Why does your company care about this question?
3. What data do you need to answer this question?
4. How do you evaluate how good your solution is?
5. What data product do you plan to build?

Is This Weird?

Anomaly Detection Algorithms

Examples

- Is this transaction a fraud?
- Is this combination of purchases very different from what this customer has made in the past?
- Are these voltages normal for this season and time of day?

1. Which company do you work at?
2. Why does your company care about this question?
3. What data do you need to answer this question?
4. How do you evaluate how good your solution is?
5. What data product do you plan to build?

How much or How Many?

Regression Algorithms

Examples

- How many new followers will I get next week?
- What will the temperature be next Tuesday?
- What will my fourth quarter sales in Canada be?

1. Which company do you work at?
2. Why does your company care about this question?
3. What data do you need to answer this question?
4. How do you evaluate how good your solution is?
5. What data product do you plan to build?

How Is This Organized?

Clustering Algorithms

Examples

- Which shoppers have similar tastes in products?
- Which viewers like the same kind of movies?
- Which printer models fail the same way?

1. Which company do you work at?
2. Why does your company care about this question?
3. What data do you need to answer this question?
4. How do you evaluate how good your solution is?
5. What data product do you plan to build?

The “Data Science” term: buzzword?

What is a Buzzword?

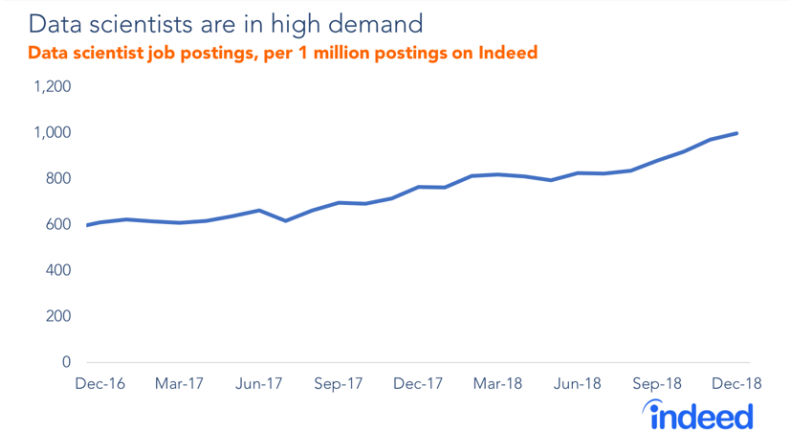
No clear definition

No big breakthrough on the technical side

No respect for the people who have been working on this kind of stuff for years

Data Science was a Buzzword (before 2018)

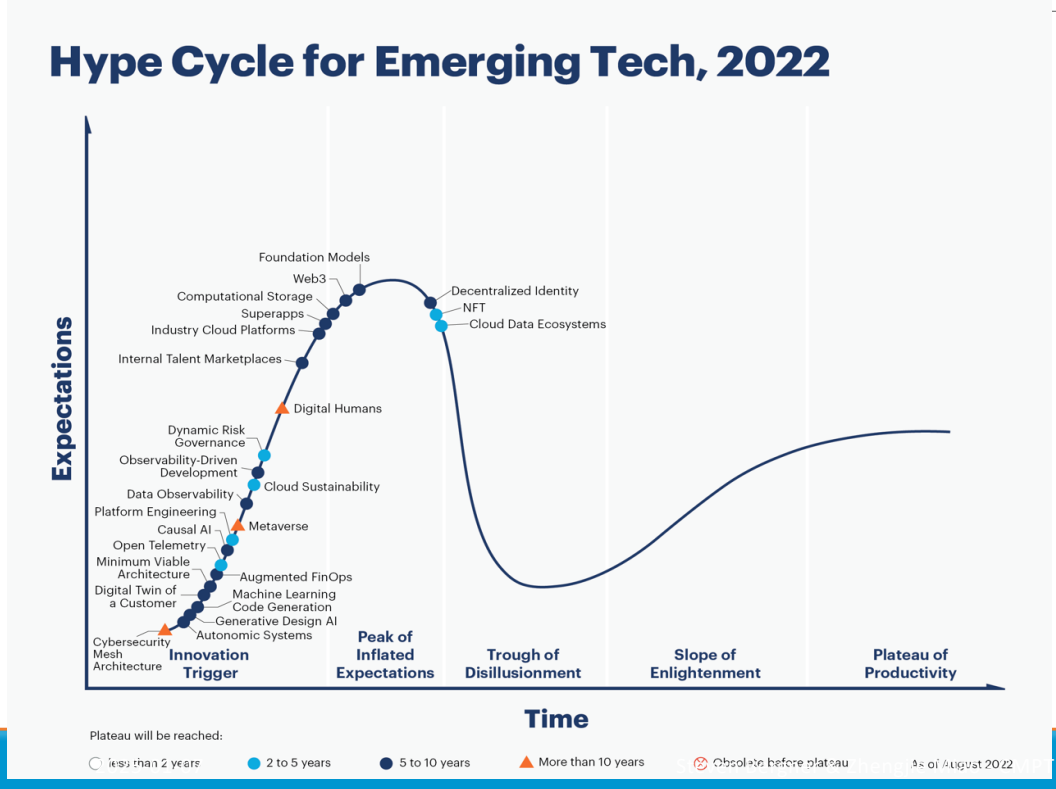
Is Data Science Only a Buzzword?



What's New?

- The combination of the three skills
- Lots of data about many aspects of our lives
- Infinite computing power (due to cloud computing)
- The need for data science is not only in the tech giant, but everywhere

Is Data Science Over-Hyped? Not Any More



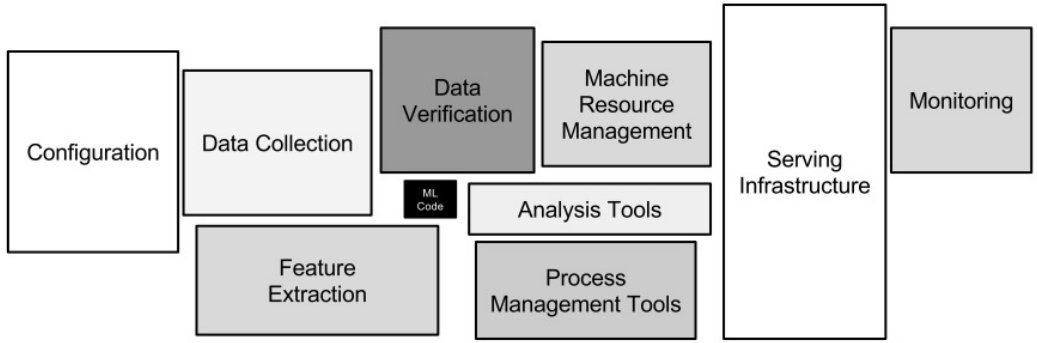
Where is “Data Science”?!
Where is “Big Data”?

AI is the new hype, but…

Google
NeurIPS 2015

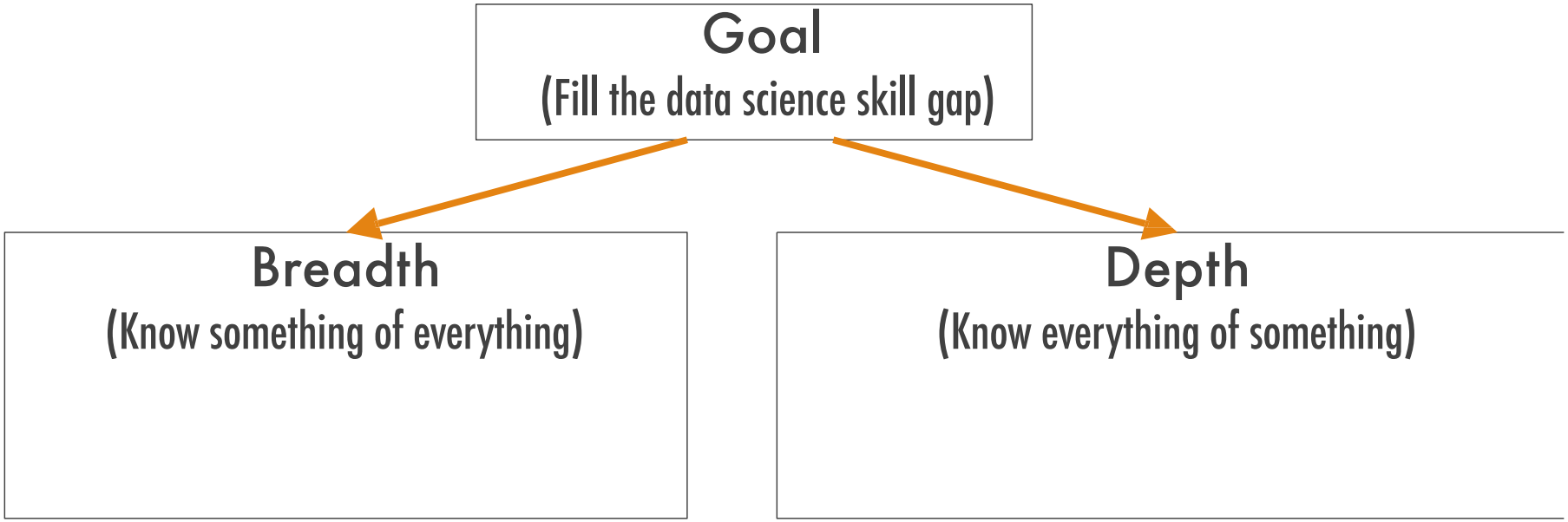
Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.



Course Structure

What's This Course About?



SFU Big Data Science Publication

(https://medium.com/sfu-big-data) **1300+** Followers; **100,000** visits in 3 months;

Demystifying Random Forest

A deep dive into Random Forest

Tushar Chand Kapoor
Mar 2, 2019 · 7 min read ★

Demystifying Random Forest

Distributed by curators in MACHINE LEARNING

Lifetime summary

Published on March 2, 2019 in SFU Big Data Science

VIEWS	EARNINGS	AVERAGE READING TIME
14.6K	$14.83	1 min 1 sec

Tushar Chand Kapoor
Data Engineer at Best Buy CHQ | Machine Learning | Big Data | Azure | tusharck.com

NOV 12, 2019

Tushar Chand Kapoor · 10:21 pm
Hi Professor,

Thanks for introducing us to the world of writing articles on medium. This has really helped me along the way.

Regards

Jiannan Wang · 10:50 pm
I am so glad to hear this. You have the special talent of writing articles on medium. :)

Tushar Chand Kapoor · 11:06 pm
Thanks you very much :).

towards
data science

Glimpse into PyTorch3D: An open-source 3D deep learning library

Tushar Chand Kapoor Feb 9, 2020 · 2 min read ★

Object Detector Android App Using PyTorch Mobile Neural Network

Tushar Chand Kapoor Nov 18, 2019 · 4 min read ★

Final Project

Proposal Phase (1wk)

Milestone (3wks)

- Student presentation

Final Project Presentation (4wks)

- Best Project Awards
- Get feedback from MPCS Big Data Advisors

Course Topics

1. Introduction to Data Science (1 week)
2. Data Preparation (1 week)
3. Visualization (2 weeks)
4. Statistics (2 weeks)
5. Practical Machine Learning (2 weeks)
6. Deep Learning / Gen AI (2 weeks)
7. Responsible Data Science (1 week)

Data Preparation

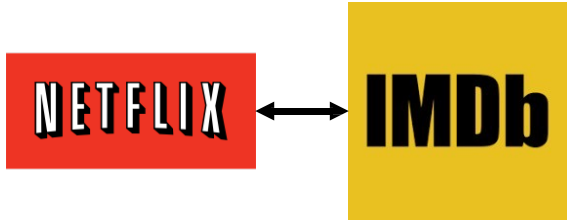
Do you know data integration?

BRUCE SCHNEIER SECURITY 12.12.07 09:00 PM

Why 'Anonymous' Data Sometimes Isn't

LAST YEAR, NETFLIX published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

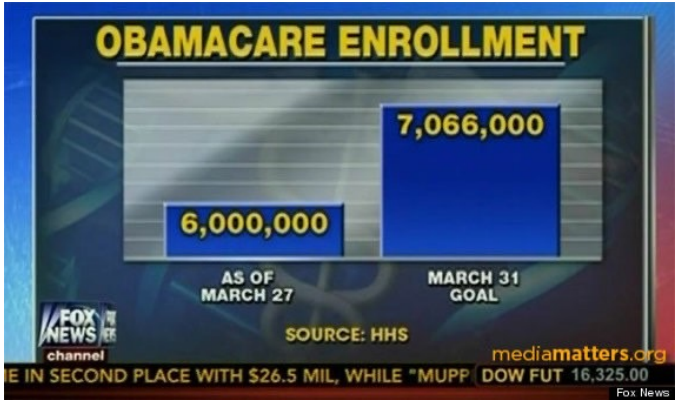
Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, de-anonymized some ofthe Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.



Disclaim: The point is to show the power of data integration rather than encourage you to work on De-Anonymization.

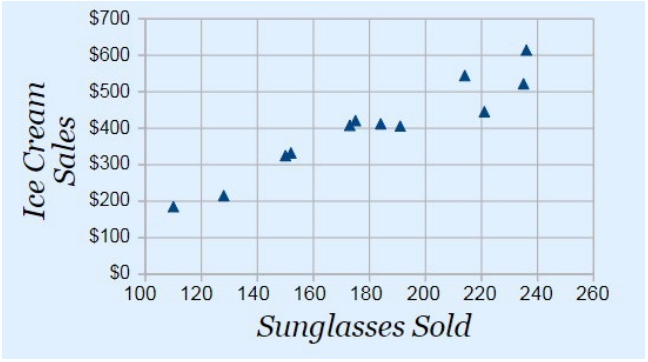
Visualization

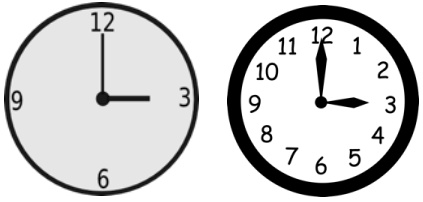
- Do you know visualization principles?
  - Without knowing the principles,
  - you might make mistakes like this!



Statistics

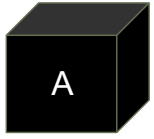
Do you know correlation ≠ causality?





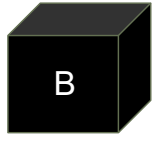
Practical Machine Learning

Do you know ML explanation?



Because it has wings and a beak

Bird: 99.0%



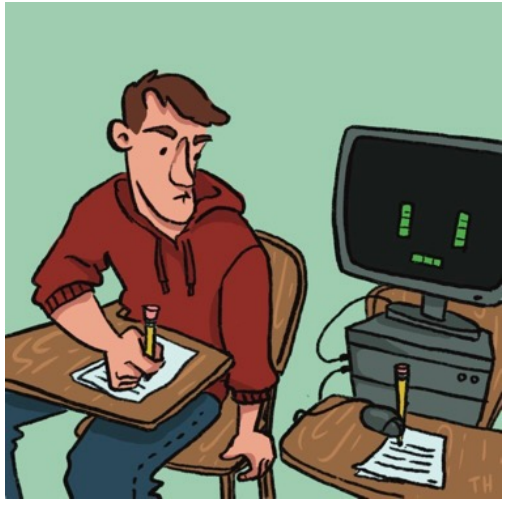
Because it is white and the background is blue

Bird: 99.9%

Which model are you going to choose?

On the use of Generative AI tools

- LLM – Large Language Model (e.g., ChatGPT)
  - 40% of respondents in Anaconda's survey say their companies are working on internal generative AI tools, such as LLMs
  - About half data practitioners reported they have spent more time on GenAI tools compared to last year
- It is still a tool, like any tool it has strengths and weaknesses
  - You may explore its uses
  - However, do **NOT** take its responses blindly
    - For some questions it works well and some **very poorly**
    - You need to know when it is wrong



Tuesday Hadden | Head of Illustration

Marking Scheme

Assignments: 10 x 4% = 40%

Blog post: 8%

In-lab exercise: 5%

Final Project: 47%

- Proposal (2%), Milestone (15%), Poster (15%), Report (15%)

Major Deadlines

When	What
Every Friday	Assignment Due
Tuesday, Jan 14	Form a team (3-5 members)
Friday, Feb 7	Blog Post on Medium
Friday Feb, 14	Final Project Proposal
Wednesday, Mar 5	Final Project Milestone
Tuesday, April 8	Final Project Presentation Session
April 11	Final Project Video/Code/Report Submission

Lectures/Labs

Lectures

- Tuesday 10:30 AM - 12:20 PM

"Lab" Hours

- **Lab G101:** Wed 11:30 AM to 1:20 PM (Instructor + TA)
  Fri 11:30 AM to 1:20 PM (TA)
- **Lab G102:** Wed 1:30 PM to 3:20 PM (Instructor + TA)
  Fri 1:30 PM to 3:20 PM (TAs)

You can use your own computer for most of the work in the course.
You can use the lab machine – lots of space, compute, and availability.
You can also access the lab cluster (http://cluster.cs.sfu.ca/) **(Credit: Greg Baker)**

Communications

Web page

- Link: https://coursys.sfu.ca/2025sp-cmpt-733-g1/pages/
  Course information, lecture notes, and assignments

Google form

- Link: http://tiny.cc/9qw2vz
- Provide anonymous feedback to improve courses (Available from Jan - Apr 2025)

CourSys Discussion Forum

- https://coursys.sfu.ca/2025sp-cmpt-733-g1/forum/
- Questions and discussions outside of lab times

Policy

Don't be Late

- Everyone has a total budget of 2 days to be used on assignments
- Once it is used up, 20% penalty per day for each late day

Don't Cheat

- We will do plagiarism check
- If you got caught, your final mark would be deducted by 30%

If you are struggling, let us know!

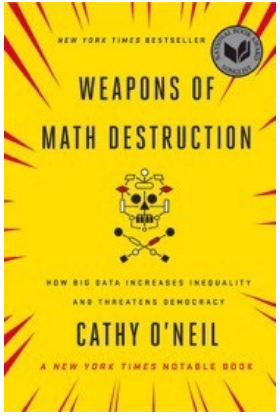
The Last But Not The Least

Data science could be harmful

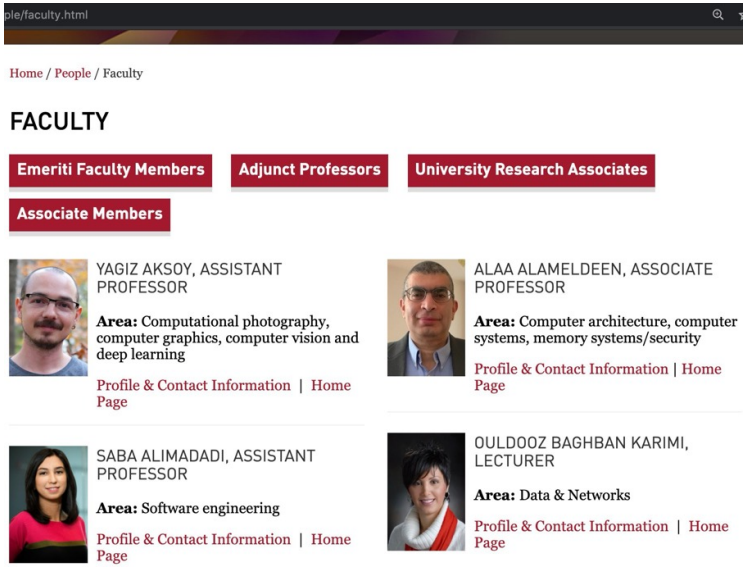
- Kill jobs, increase inequality, threaten democracy

Don't be evil!

or



Assignment 1: Web Scraping



faculty_table

name	rank	area	profile
Yagiz Aksoy	Assistant Professor	Computational Photography, Computer Graphics, Co	http://www.sfu.ca/con
Saba Alimadadi	Assistant Professor	Software Engineering	http://www.sfu.ca/con
Brad Bart	Senior Lecturer	Instruction	http://www.sfu.ca/con
Andrei Bulatov	Professor	Constraint Satisfaction, Complexity Of Computation	http://www.sfu.ca/con
Sheelagh Carpendale	Professor	Information Visualization	http://www.sfu.ca/con
Angel Chang	Assistant Professor	Natural Language Processing, Artificial Intelligence, C	http://www.sfu.ca/con
Victor Cheung	Limited-Term Lecturer	Human-Computer Interaction, Interface And Interactic	http://www.sfu.ca/con

2 Part Assignment, obtain info from 2 sources:

a) From departmental website
b) From course outline DB via web interface