

# CMPT 733

# ~~Big Data Programming II~~

---

## Data Science

SLIDES BY:

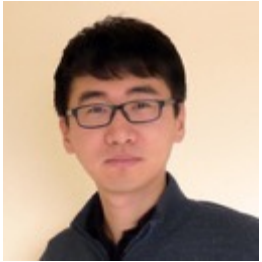
JIANNAN WANG

<https://sfu-db.github.io/bigdata-cmpt733/>



# Who Are We?

---



Jiannan Wang

Associate Professor from SFU  
Postdoc from UC Berkeley AMPLab  
Ph.D. from Tsinghua University

10+ years of research  
experience in the  
**database** field



Steven Bergner

University Research Associate from SFU  
Quantitative Analyst at FINCAD  
Ph.D. and Postdoc from SFU


10+ years of research  
and working experience  
in the **visualization** field

# MPCS Remote Teaching Survey

---

55 Responses

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

- 
1. Slack 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**

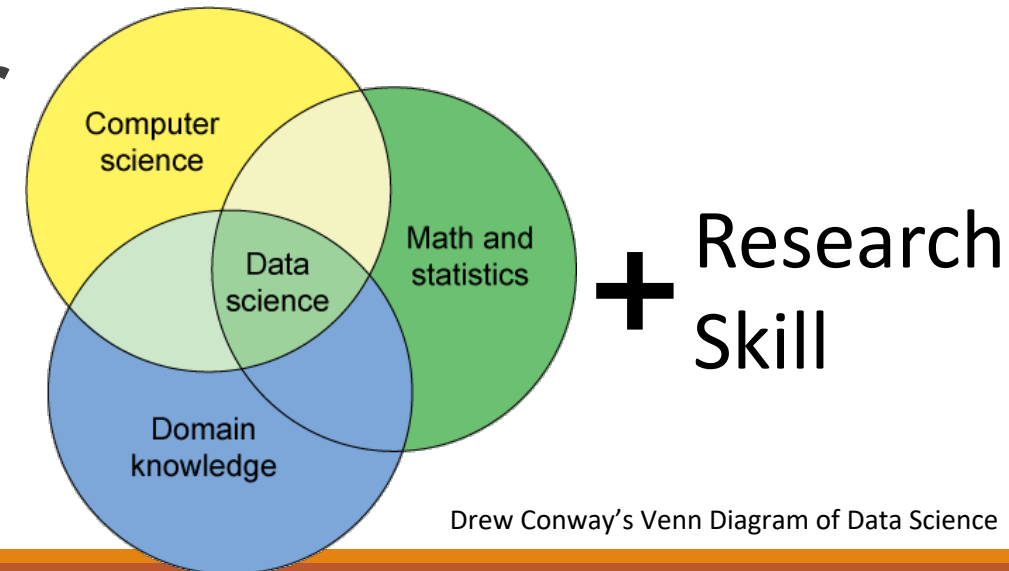
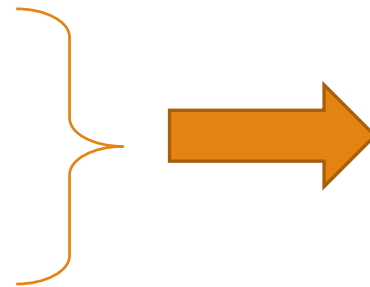
# New Skillset

## 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 PMP Big Data Curriculum

---

Data Scientist, Data Engineer, Machine Learning Engineer, etc.

**Term 4**  
(9 credits)

Three Elective Courses

Go Deeper!!!

**Term 3**  
(3 credits)

CO-OP

Domain Knowledge and Research Skills

**Term 2**  
(9 credits)

Distributed & Cloud Systems + Big Data Lab 2

Engineering and Data Science Skills

**Term 1**  
(9 credits)

Machine Learning + Big Data Lab 1

ML/Stats and Big Data Skills

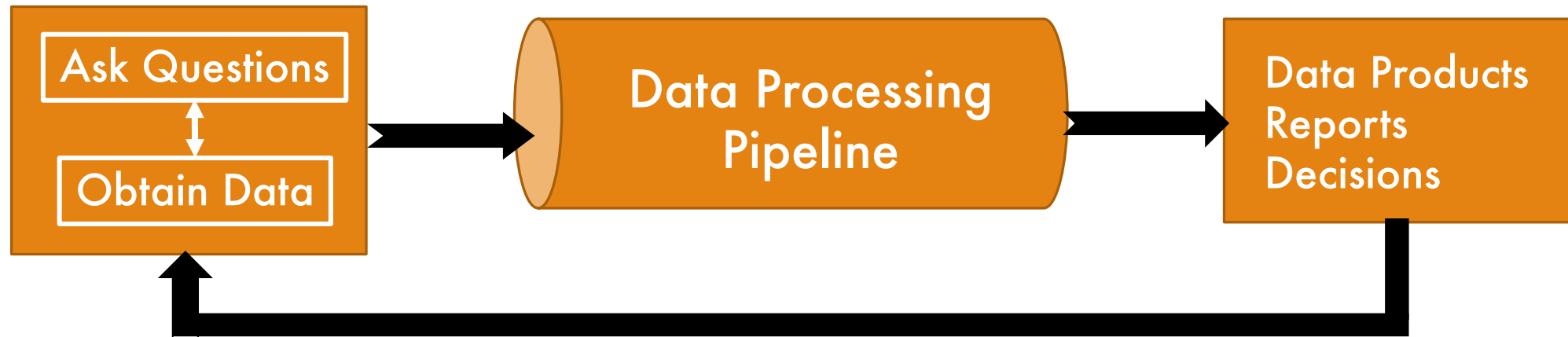


---

# **Data Science Lifecycle**

# Data Science Lifecycle (High-Level)

The entire workflow is **iterative**



## Two ways to produce questions

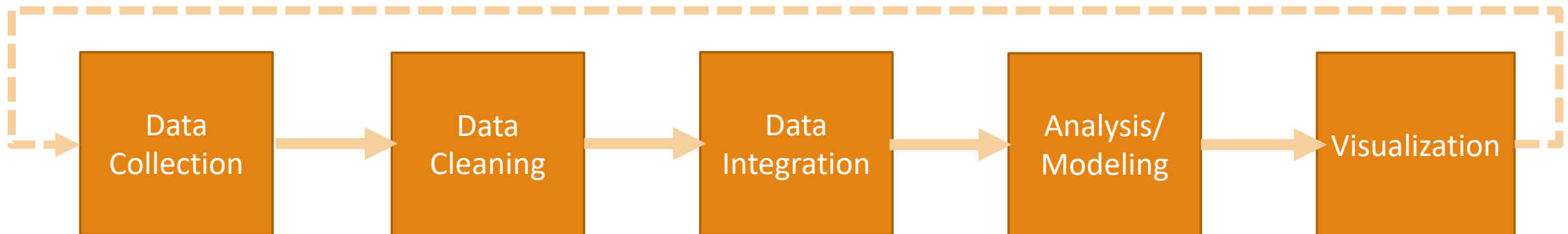
- Start with questions and then collect the related data
- Start with data and then think about the questions that can be answered

# Data Processing Pipeline

What you think you do?



What you really do?



---

At Least

# 4 Questions Data Scientists Can Answer

<https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-the-5-questions-data-science-answers>

# 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**

## Hype Cycle for Emerging Technologies, 2020



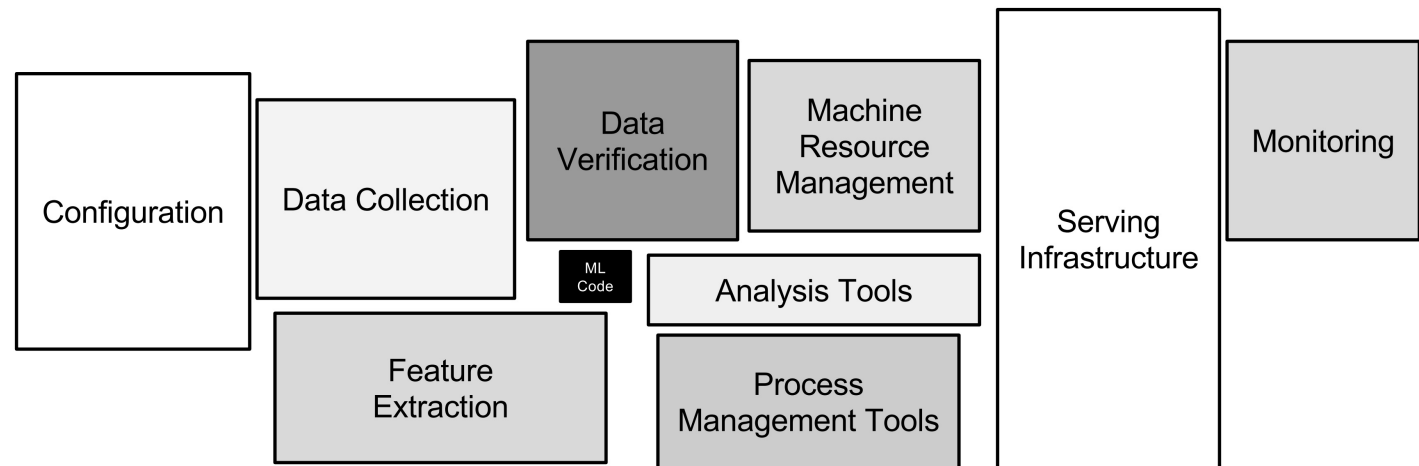
Where is "Data Science"?!  
Where is "Big Data"?

# AI is the new hype, but...

## Hidden Technical Debt in Machine Learning Systems

Google  
NeurIPS 2015

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

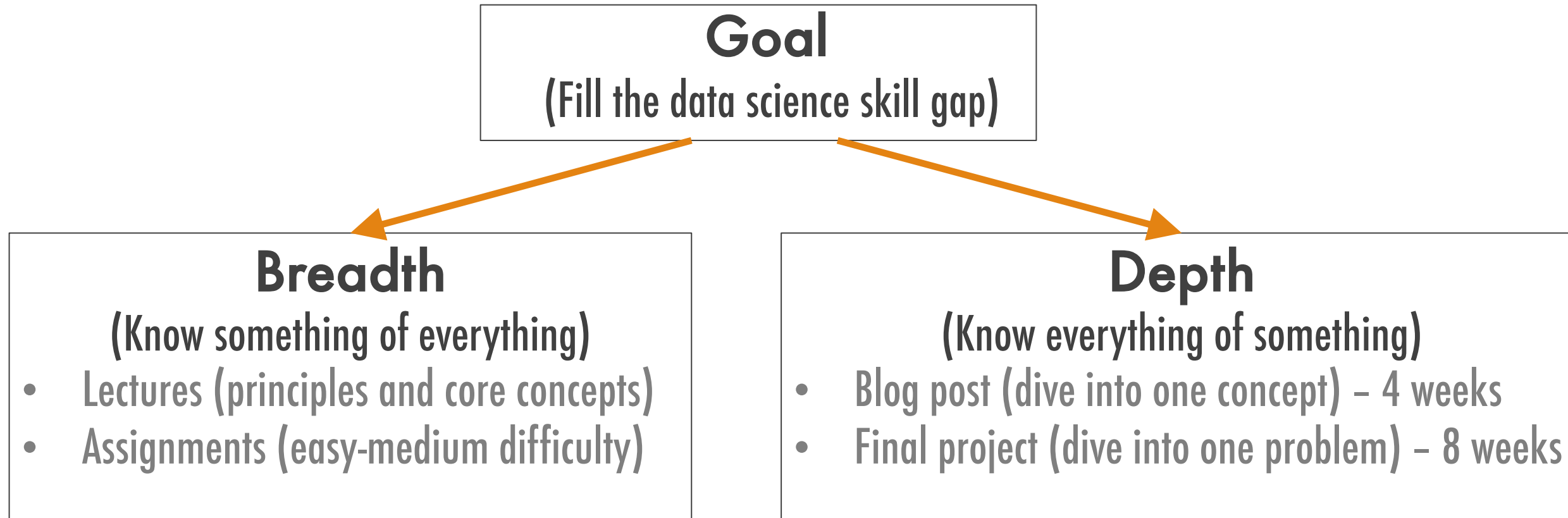


---

# 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

**14.6K**

EARNINGS ⓘ

**\$14.83**

AVERAGE READING TIME ⓘ

**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 (10,000 CAD)
- Get feedback from **MPCS Big Data Advisors**



**Beier Cai**

Co-Founder  
*Commit*



**Jesse Calderon**

SVP, Product Development  
*Tableau*



**George Chow**

Data Engineer  
*Self-employed*



**Dennis Loktionov**

Director of Digital  
*Tech Resources*



**Aly Sidi**

Vice President, Software  
*Neurio Technologies*



**Ken Wong**

Senior Director of Product  
Management  
*Tableau*



**Kate Wright**

VP, Product & Development  
*SAP*



**Matthew St. John**

Director, Analytics  
Centre of Excellence  
*Best Buy Canada*

# 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 (1.5 weeks)
7. Cloud Computing (0.5 week)
8. Responsible Data Science (1 week)

# Data Preparation

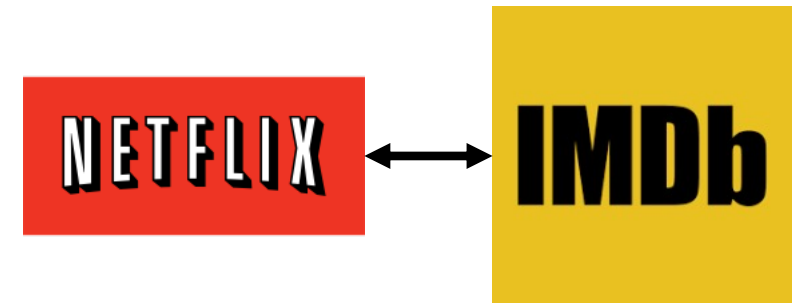
## Do you know data integration?

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 of the Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

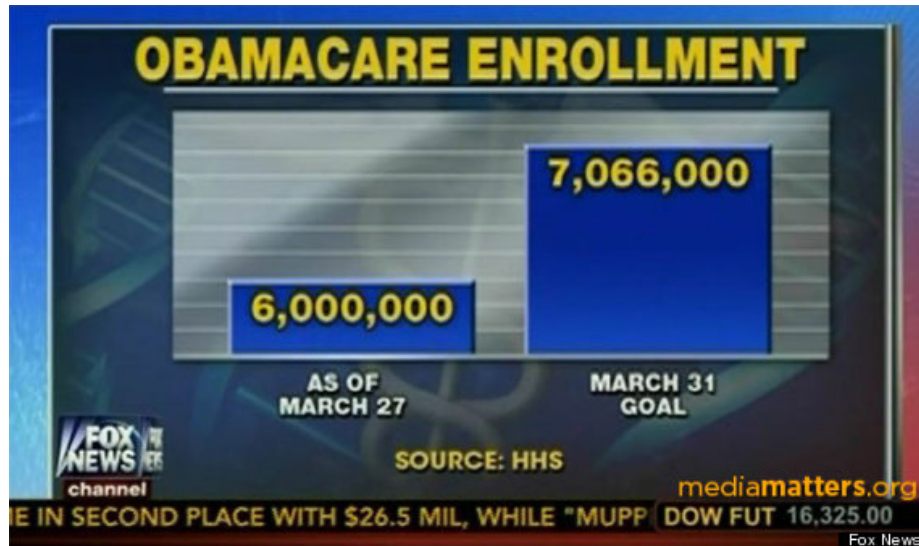


**Disclaim:** The point is to show the power of data integration rather than encourage you to work on De-Anonymization.

# Visualization

Do you know visualization principles?

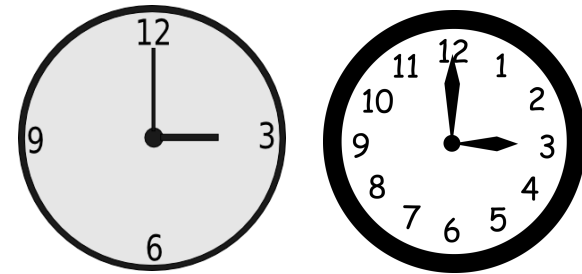
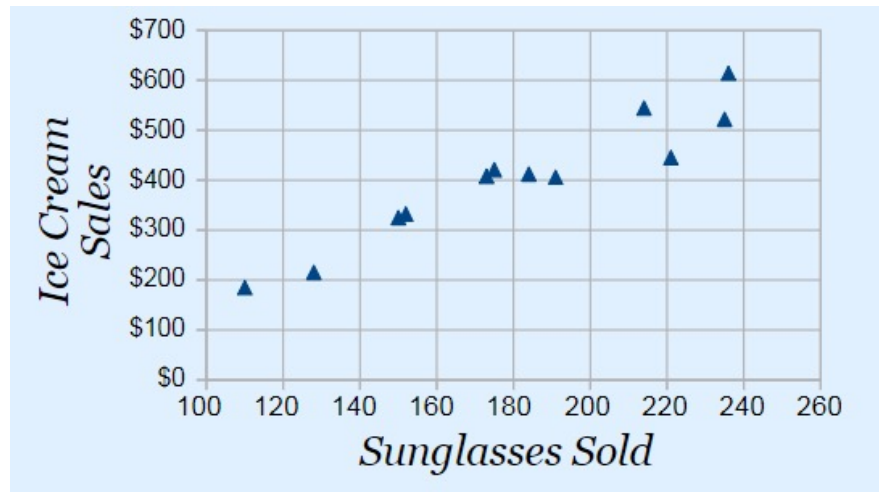
Without knowing the principles,  
you might make a lot of mistakes like this!



# Statistics

---

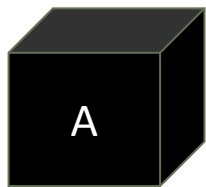
Do you know correlation  $\neq$  causality?



# Practical Machine Learning

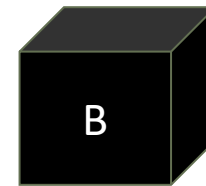
---

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?

# Marking Scheme

---

Assignments:  $11 \times 3\% = 33\%$

Blog Post: 20%

- Depth (10%), Popularity (10%)

Final Project: 47%

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

**Bonus:** Contribute to [dataprep.ai](https://dataprep.ai) (0.5%)

- Create an issue (0.2%)
- Send a pull request (0.3%)

# Major Deadlines

---

When	What
Every Monday	Assignment Due
Monday Jan 17	Form a team (3-5 members)
Monday Feb 11	Blog Post Submission
Tuesday Feb 15	Final Project Proposal
Thursday Mar 10	Final Project Milestone
Monday Mar 28	Blog Post Popularity
April 11	Final Project Presentation Session
April 17	Final Project Video/Code/Report Submission



# Lectures/Labs

---

## Lectures

- Monday 1:30 - 3:20

## “Lab” Hours

- Lab G101+G103: Tue 1:30 PM - 3:20 PM, Thu 3:30 PM - 5:20 PM
- Lab G102+G103: Tue 3:30 PM - 5:20 PM, Fri 3:30 PM - 5:20 PM

You can use your own computer for most of the work in the course.

You can also access the lab cluster (<http://cluster.cs.sfu.ca/>) (Credit: Greg Baker)

# Communications

---

## Web page

- Link: <https://sfu-db.github.io/cmpt733>
- Course information, lecture notes, and assignments

## Google form

- Link: <http://tiny.cc/733-feedback>
- Provide anonymous feedback to improve courses (Available from Jan - Apr 2022)

## Slack Channel

- Link: [https://join.slack.com/t/cmpt-733-spring-2022/shared\\_invite/zt-11hh0b91d-Bp7hdws97vzxHaFZnB~FBw](https://join.slack.com/t/cmpt-733-spring-2022/shared_invite/zt-11hh0b91d-Bp7hdws97vzxHaFZnB~FBw)
- Questions and discussions outside of lab times

# Policy

---

## Don't be Late

- Everyone has a budget of 2 days to be used on assignments
- Once it is used up, 20% 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

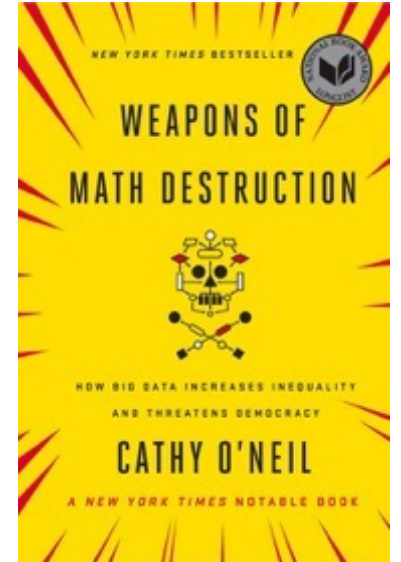
## Data science could be harmful

- Kill jobs, increase inequality, threaten democracy

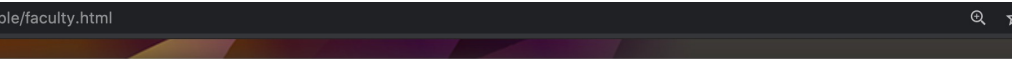
Don't be evil!



or



# Assignment 1-1: Web Scraping



[Home](#) / [People](#) / Faculty

## FACULTY

**Emeriti Faculty Members**

**Adjunct Professors**

**University Research Associates**

**Associate Members**



YAGIZ AKSOY, ASSISTANT PROFESSOR

**Area:** Computational photography, computer graphics, computer vision and deep learning

[Profile & Contact Information](#) | [Home Page](#)



ALAA ALAMELDEEN, ASSOCIATE PROFESSOR

**Area:** Computer architecture, computer systems, memory systems/security

[Profile & Contact Information](#) | [Home Page](#)



SABA ALIMADADI, ASSISTANT PROFESSOR

**Area:** Software engineering

[Profile & Contact Information](#) | [Home Page](#)



OULDOOZ BAGHBAN KARIMI, LECTURER

**Area:** Data & Networks

[Profile & Contact Information](#) | [Home Page](#)

faculty\_table

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

# Assignment 1-2: Web APIs

---

azure

dblp

etsy

finnhub

guardian

mapquest

musixmatch

openweathermap

spoonacular

spotify

times

twitch

twitter

wikia

yelp

youtube

## Yelp -- Collect Local Business Data

- ▶ What's the phone number of Capilano Suspension Bridge Park?
- ▶ Which yoga store has the highest review count in Vancouver?
- ▶ How many Starbucks stores in Seattle and where are they?
- ▶ What are the ratings for a list of restaurants?

Group Assignment: <https://coursys.sfu.ca/2022sp-cmpt-733-g1/pages/Web-API-Assignment/view>