CMPT 733
Big Data Programming II

Instructor  Steven Bergner
Course website  https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages/
Instructor

Steven Bergner [he/him]
PhD, Computing Science
Term Lecturer | Computing Science | SFU
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
# MPCS Remote Teaching Survey

## 55 Responses in 2021

<table>
<thead>
<tr>
<th>Questions</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are you satisfied with lab courses?</td>
<td>91%</td>
</tr>
<tr>
<td>Are you satisfied with co-op office support?</td>
<td>87%</td>
</tr>
<tr>
<td>Are you satisfied with academic team support?</td>
<td>91%</td>
</tr>
<tr>
<td>Do you feel a sense of community in your cohort?</td>
<td>72%</td>
</tr>
<tr>
<td>Are you happy with your decision to pursue the program?</td>
<td>93%</td>
</tr>
</tbody>
</table>
## MPCS Remote Teaching Survey

### 55 Responses

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

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

<table>
<thead>
<tr>
<th>Computer Science</th>
<th>1950-</th>
<th>Software Engineer</th>
<th>Write software to make computers work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science</td>
<td>2010-</td>
<td>Data Scientist</td>
<td>Extract insights from data to answer questions</td>
</tr>
</tbody>
</table>

## When

- Plan ➔ Design ➔ Develop ➔ Test ➔ Deploy ➔ Maintain
- Collect ➔ Clean ➔ Integrate ➔ Analyze ➔ Visualize ➔ Communicate

## Who

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
SFU MPCS Big Data Curriculum

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

Term 4
(9 credits)
Three Elective Courses

Term 3
(3 credits)
CO-OP

Term 2
(9 credits)
Big Data Lab 2

Term 1
(9 credits)
Machine Learning + Big Data Lab 1

Go Deeper!!!
Domain Knowledge and Research Skills
Engineering and Data Science Skills
ML/Stats and Big Data Skills
Data Science Lifecycle
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?

Data Collection → Data Cleaning → Data Integration → Analysis/Modeling → Visualization
4 Questions Data Scientists Can 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**

Hype Cycle for Emerging Tech, 2022

Where is “Data Science”?! Where is “Big Data”?!
AI is the new hype, but…

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?

Goal
(Fill the data science skill gap)

Breadth
(Know something of everything)

Depth
(Know everything of something)
Demystifying Random Forest
A deep dive into Random Forest

Tushar Chand Kapoor
Mar 2, 2019 · 7 min read

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

Object Detector Android App Using PyTorch Mobile Neural Network
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 (1.5 weeks)
7. Cloud Computing (0.5 week)
8. Responsible Data Science (1 week)
Data Preparation

Do you know data integration?

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.

Disclaimer: 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**?

Which model are you going to choose?

A

Because it has wings and a beak

Bird: 99.0%

B

Because it is white and the background is blue

Bird: 99.9%
Marking Scheme

Assignments: 11 x 4% = 44%
In-lab exercise: 9%
Final Project: 47%
  ◦ Proposal (2%), Milestone (15%), Poster (15%), Report (15%)
## Major Deadlines

<table>
<thead>
<tr>
<th>When</th>
<th>What</th>
</tr>
</thead>
<tbody>
<tr>
<td>Every Friday</td>
<td>Assignment Due</td>
</tr>
<tr>
<td>Tuesday Jan 16</td>
<td>Form a team (3-5 members)</td>
</tr>
<tr>
<td>Friday Feb 16</td>
<td>Final Project Proposal</td>
</tr>
<tr>
<td>Thursday Mar 9</td>
<td>Final Project Milestone</td>
</tr>
<tr>
<td>April 9</td>
<td>Final Project Presentation Session</td>
</tr>
<tr>
<td>April 11</td>
<td>Final Project Video/Code/Report Submission</td>
</tr>
</tbody>
</table>
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 1:30 PM to 3:20 PM (TAs)
- Lab G102: Wed 1:30 PM to 3:20 PM (Instructor + TA)
  Fri 11:30 AM to 1:20 PM (TA)

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/2024sp-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 2024)

CourSys Discussion Forum
- https://coursys.sfu.ca/2024sp-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

Steven Bergner - CMPT 733
Assignment 1: Web Scraping

2 Part Assignment, obtain info from 2 sources:

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