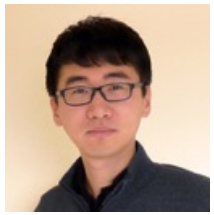
CMPT 733
Big Data Programming II

SLIDES BY:

JIANNAN WANG

https://www.cs.sfu.ca/~jnwang/

Who Are We?



Jiannan Wang

Assistant 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

Cohort 1 (2014)

Cohort 2 (2015)

Cohort 3 (2016)

Cohort 4 (2017)

Cohort 5 (2018)

Outline

What is Data Science?

Data Science Lifecycle

4 Questions Data Scientists Can Answer

Is Data Science Over-Hyped?

Course logistics

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

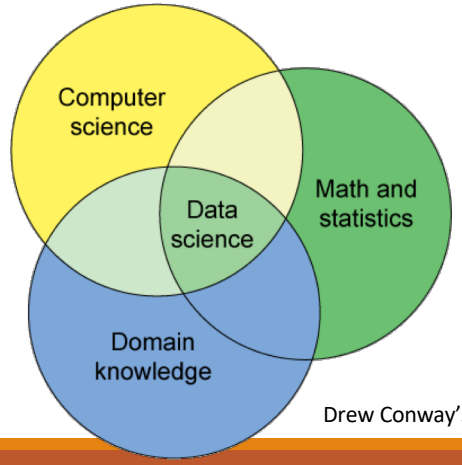
New Skillset

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

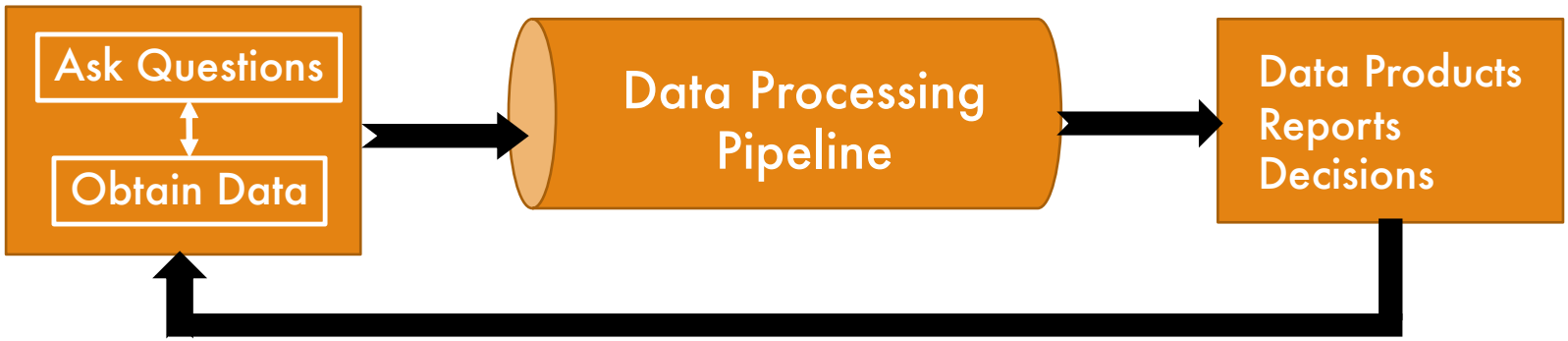


Drew Conway's Venn Diagram of Data Science

Data Science Lifecycle

Data Science Lifecycle (High-Level)

The entire workflow is iterative



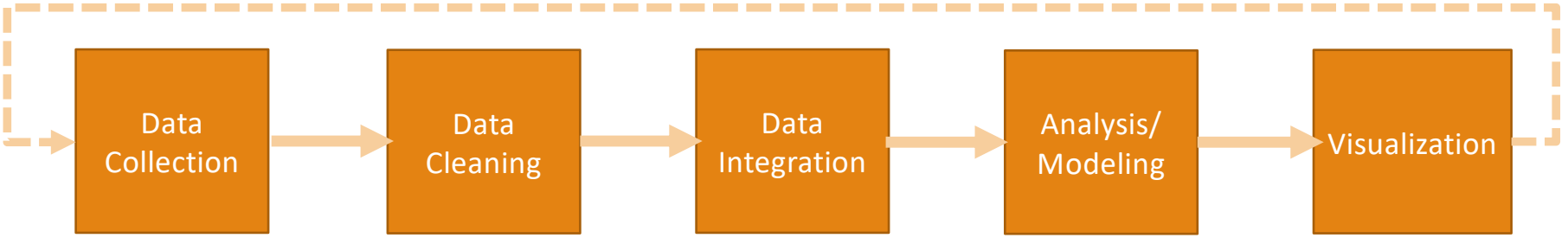
Two ways to come up with questions

- 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://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?

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?

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?

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?

Is Data Science Over-Hyped?

Is Data Science a Buzzword? YES

No clear definition

No big breakthrough on the technical side

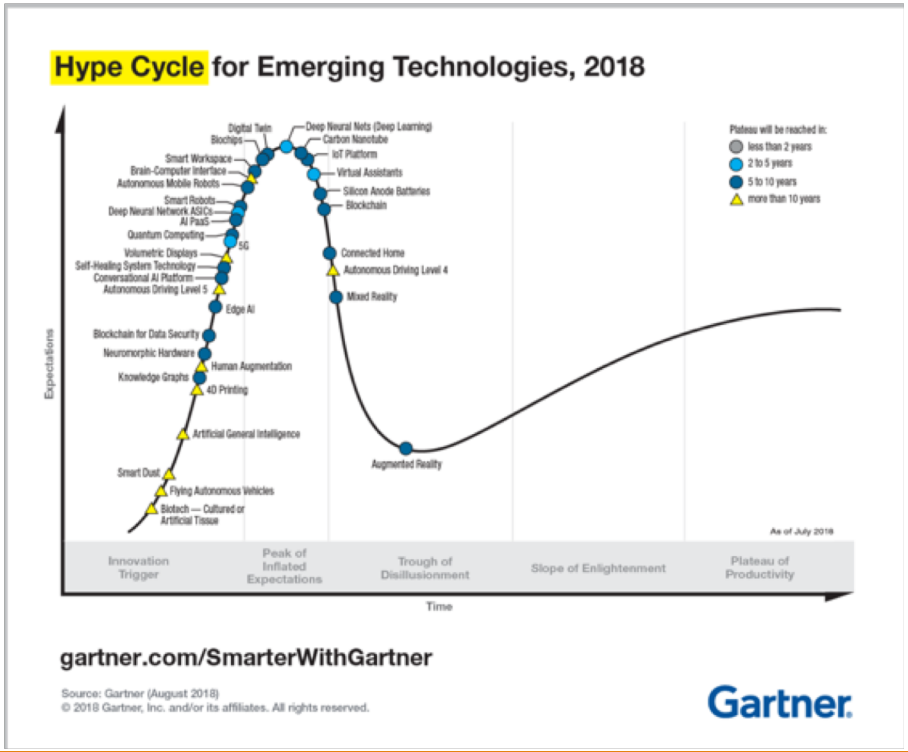
No respect for the people who have been working on this kind of stuff for years

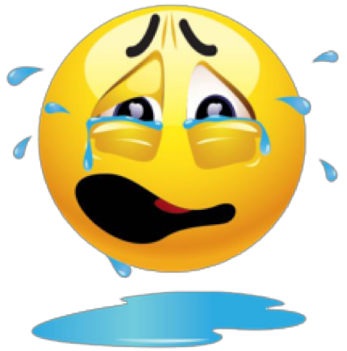
Is Data Science Only a Buzzword? NO

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

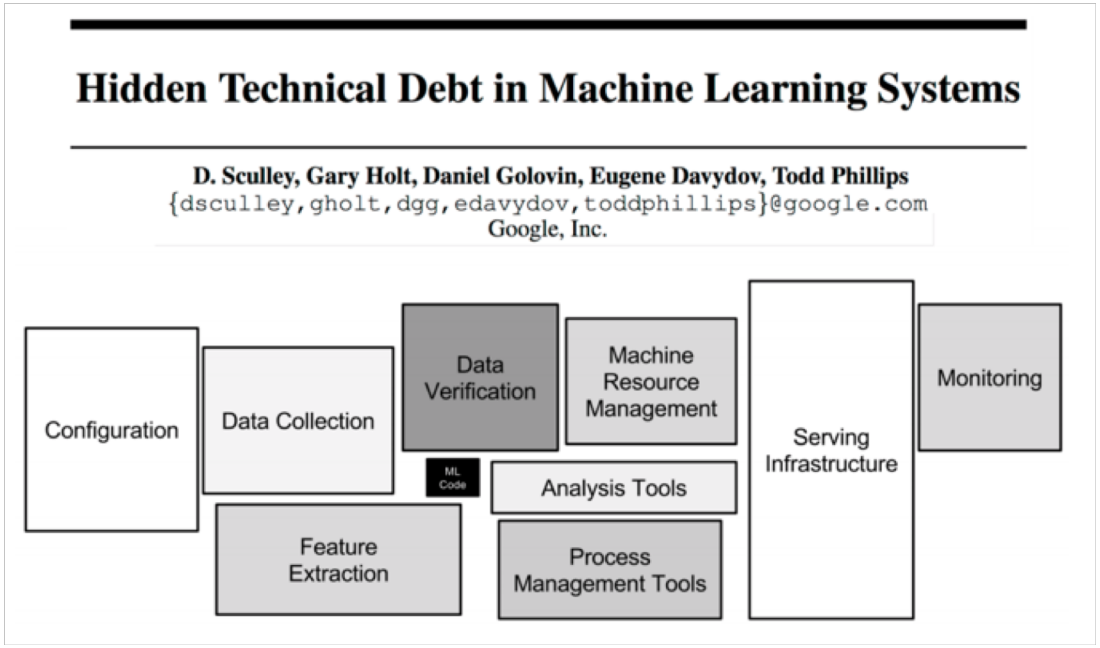




Where is “Data Science”?!
Where is “Big Data”?

AI is the new hype, but…

Google
NIPS 2015



Course Logistics

What's This Course About?

Goals

- Fill the data science skill gap

Lecture style

- More "why" less "how"

Assignment style

- Problem centric instead of tool centric

Final Project

- Start from Week 4 to Week 12

Course Topics

1. Introduction to Data Science (1w)
2. Data Preparation (1w)
3. Visualization (2w)
4. Statistics (2w)
5. Deep Learning (2w)
6. Practical Machine Learning (2w)
7. Communication (1w)

Marking Scheme

Assignments: 9 x 6% = 54%

- Work in individual

Blog Post: 10%

- Work in group
- Topics: alumni interview, tech trend, dive into X, etc.

Project: 36%

- Work in group
- Proposal (2%), Presentation (6%), Poster (14%), Report (14%)

Lectures/Labs

Lectures (2 hour/week)

- Monday 9:30-11:20

Labs (4 hours/week)

- Group A: Tues 9–10:50, Thurs 9–10:50
- Group B: Wed 11:30–13:20, Fri 11:20–13:20
- Group C: Wed 13:30–15:20, Fri 13:30–15:20

1 TA

1 TA + 1 Instructor

Communications

Web page

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

Piazza

- Sign up: https://piazza.com/sfu.ca/spring2019/cmpt733
- THE place to ask course-related questions
- Log in today and enable notifications

Google form

- Link: https://goo.gl/forms/kWgkfta0sRvcPsfC2
- Provide anonymous feedback to improve courses

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

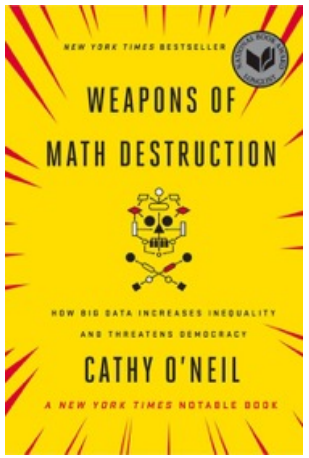
- We will do plagiarism check
- If you got caught, your final mark would be deducted by 30%

If you are struggling, let us know!

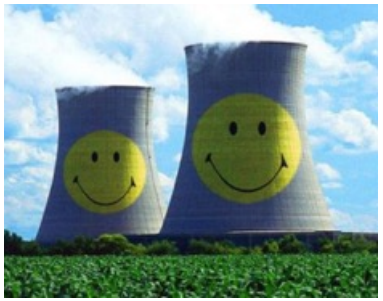
The Last But Not The Least

Data science could be harmful

- Kill jobs, increase inequality, threaten democracy



Don't be evil!

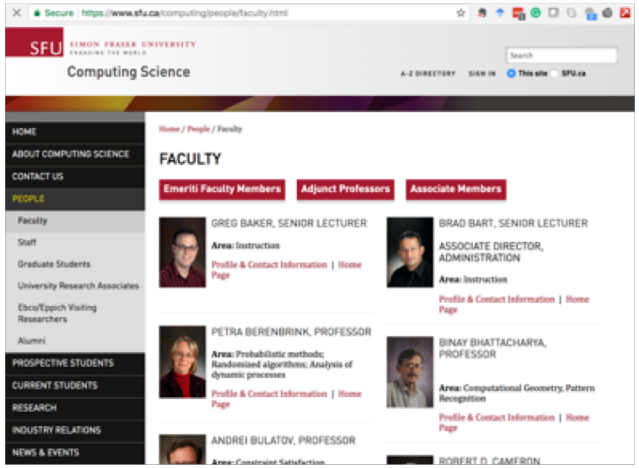


or

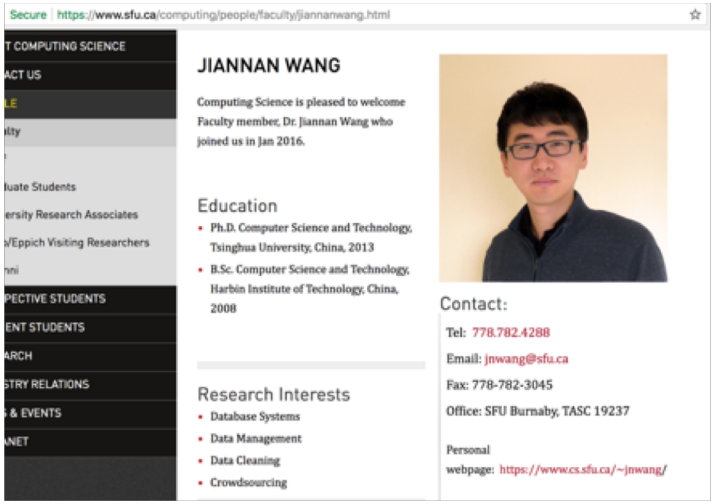
Assignment 1: Web Scraping

(http://tiny.cc/cmpt733-a1)

Task 1. Who are the CS faculty members?



Task 2. What are their research interests?



Deadline: 23:59pm, Jan 14