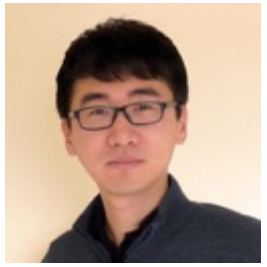
CMPT 733
Big Data Programming II

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

Outline

What is Data Science?

Data Science Lifecycle

4 Questions Data Scientists Can Answer

Is Data Science Over-Hyped?

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

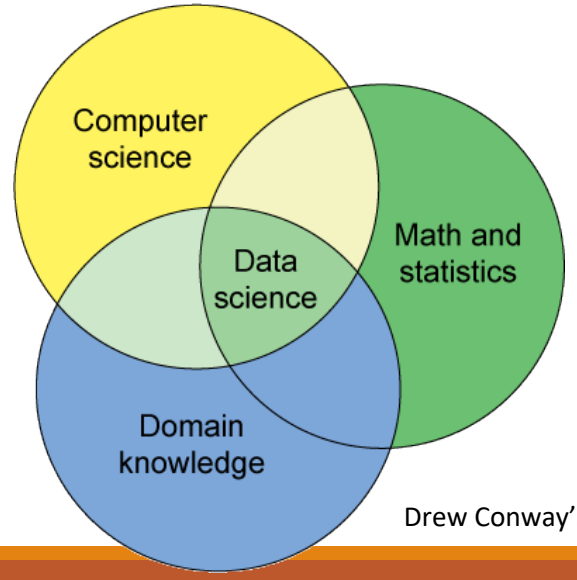
New Skillset

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

Data Science Lifecycle

Data Science Lifecycle (High-Level)

The entire workflow is iterative

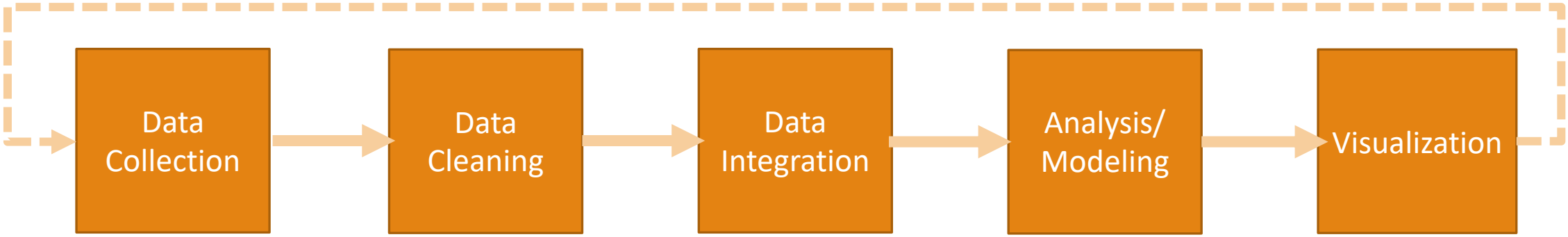
Two ways to produce questions

- 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?

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?

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 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?

Is Data Science Over-Hyped?

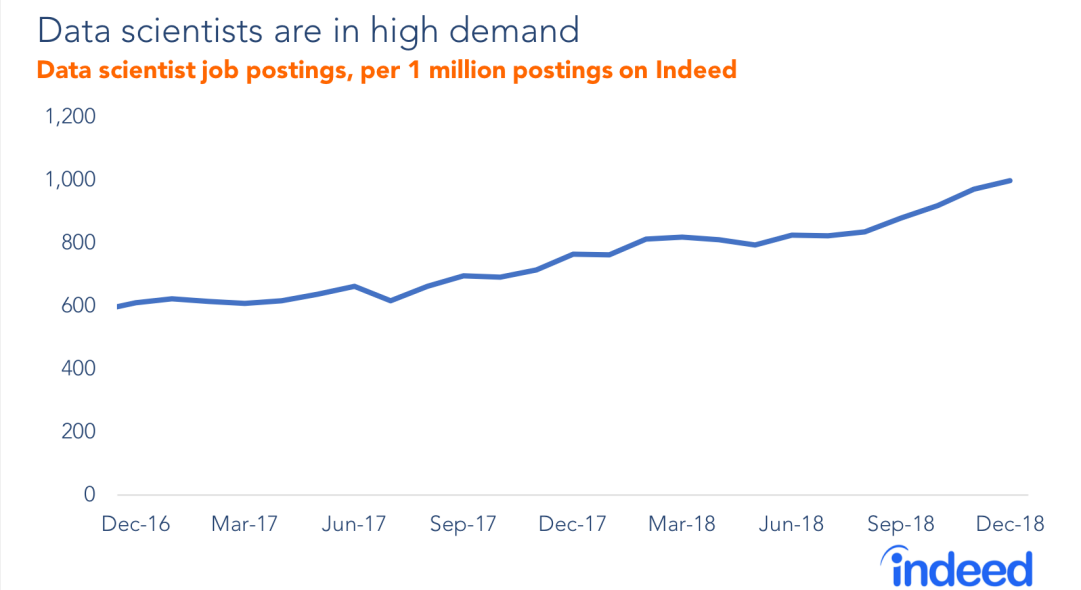
Is Data Science a Buzzword? YES

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

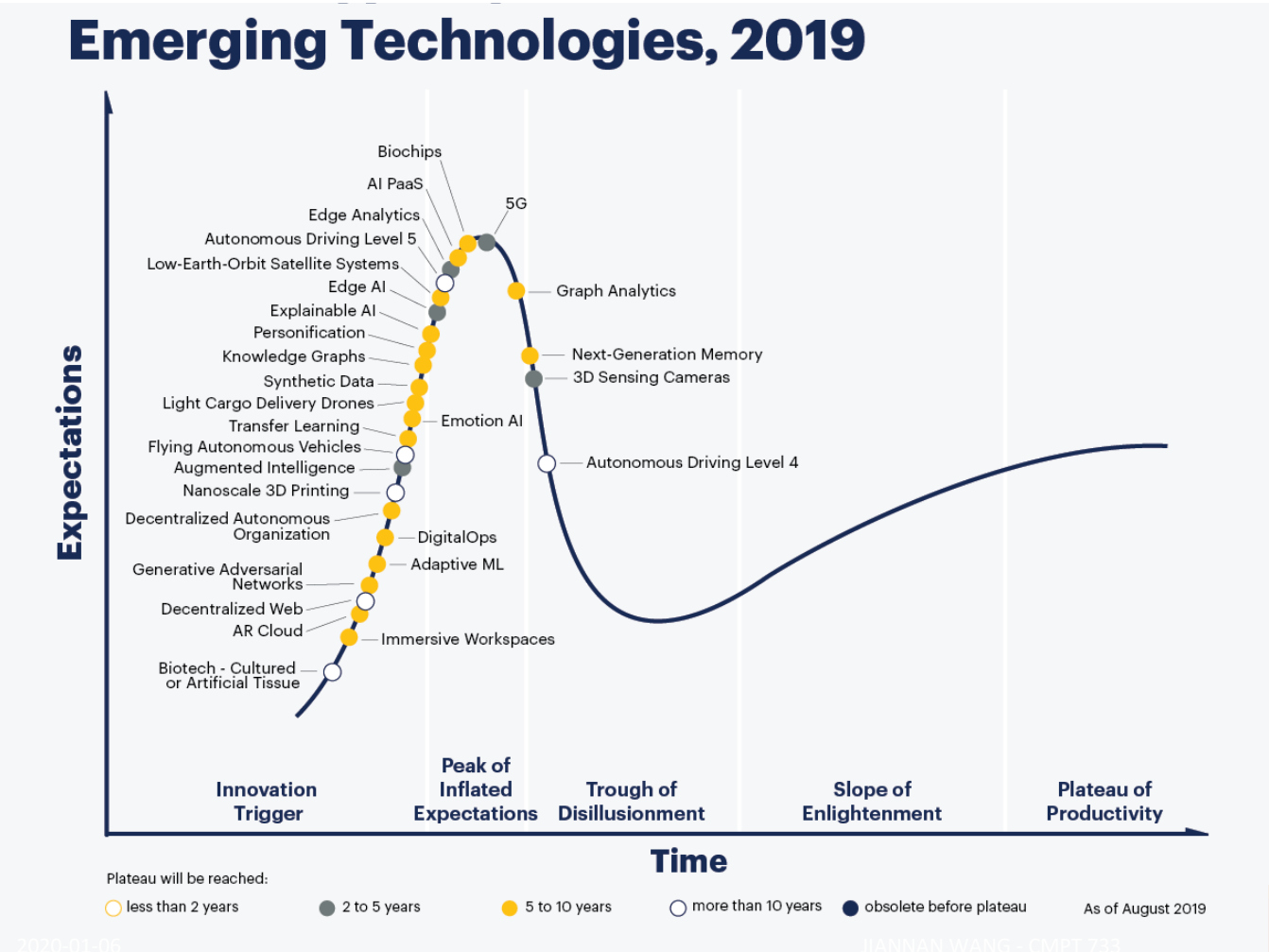
Is Data Science Only a Buzzword? NO

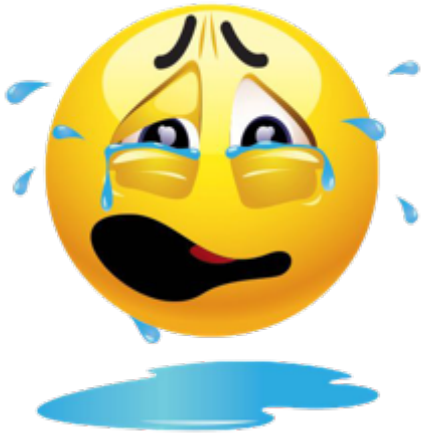


What's New?

- 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





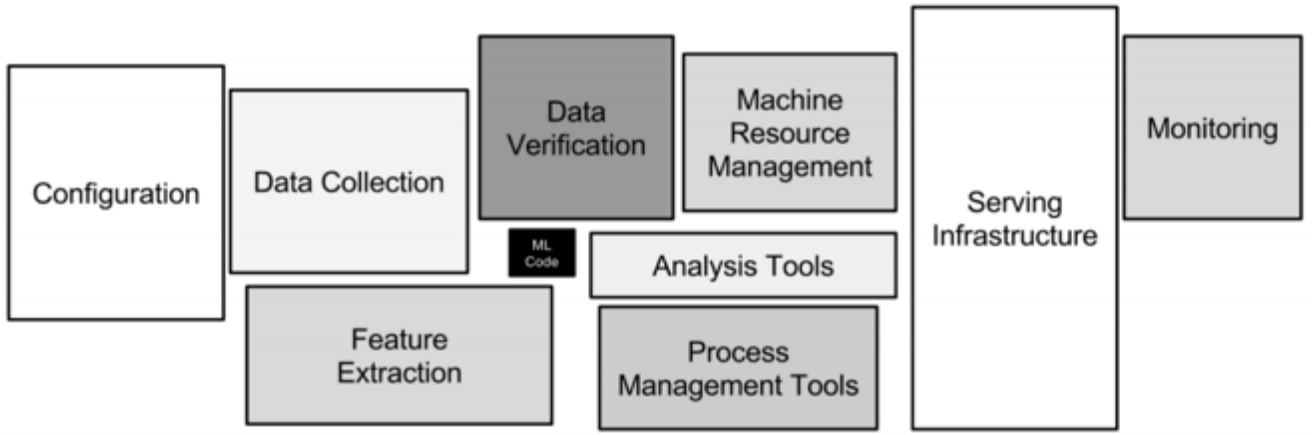
Where is "Data Science"?!
Where is "Big Data"?

AI is the new hype, but…

Google
NIPS 2015

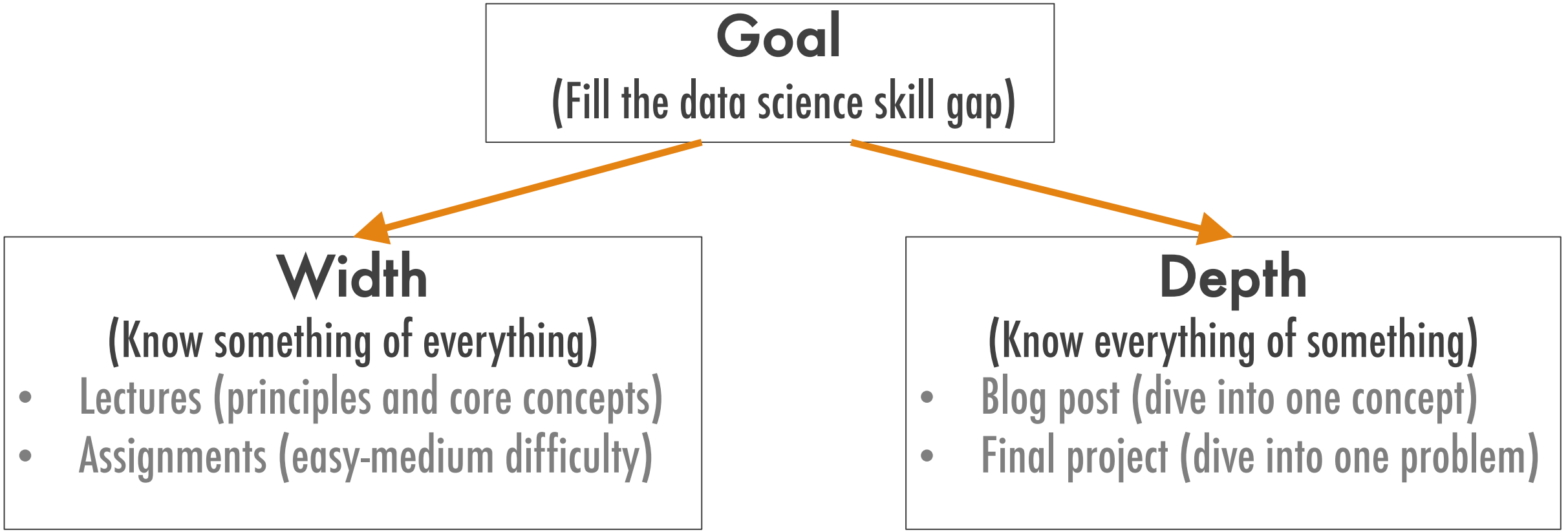
Hidden Technical Debt in Machine Learning Systems

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)

800+ 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 :).

Final Project Showcase

Presented in Vancouver Downtown

Open to all students and industrial people

Best Project Awards (10,000 CAD)

Get feedback from PMP Big Data Advisors

Course Topics

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

Data Preparation

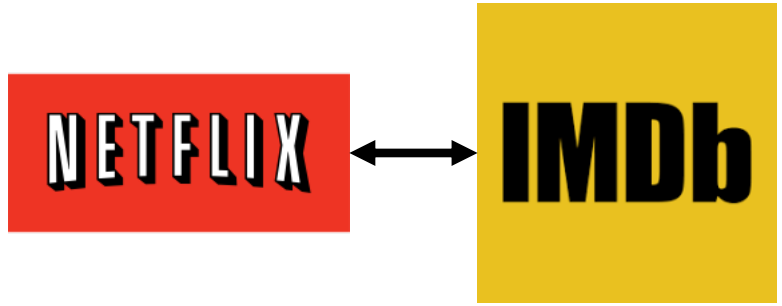
Do you know data integration?

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.

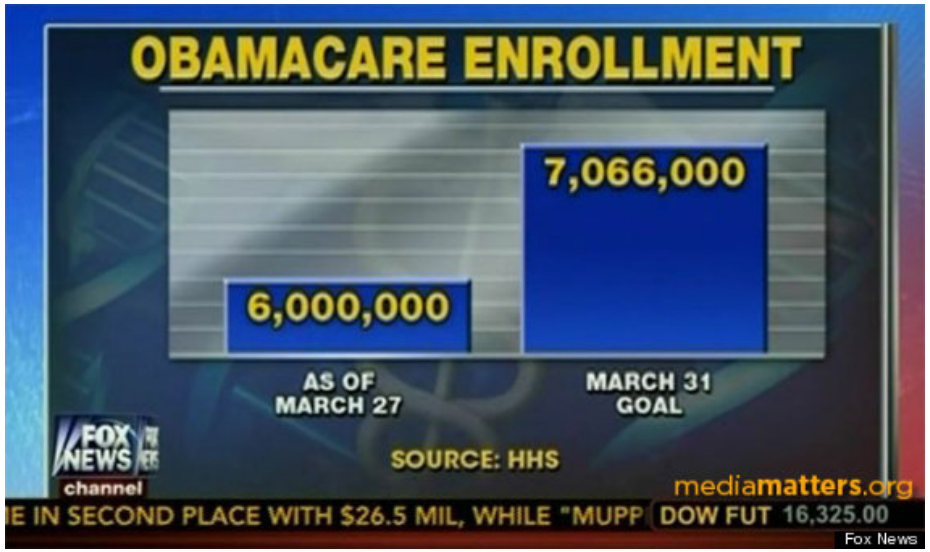


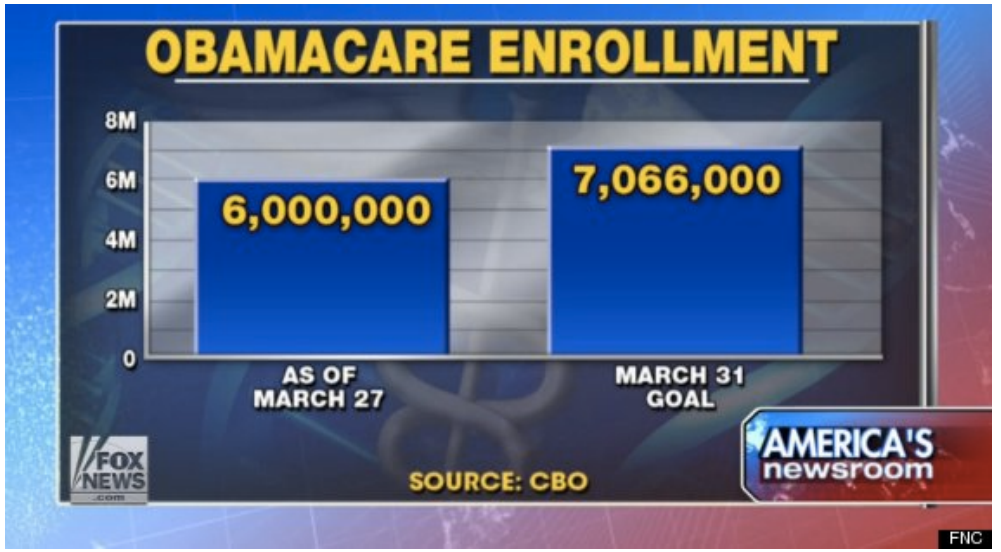
Disclaim: The point is to show the power of data integration rather than encourage you to work on De-Anonymization.

Visualization

Do you know visualization principles?

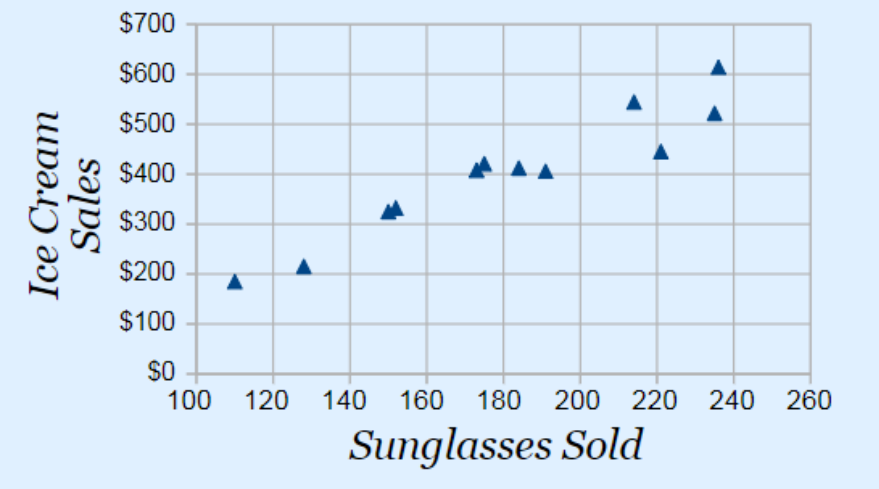
Without knowing the principles,
you might make a lot of mistakes like this!

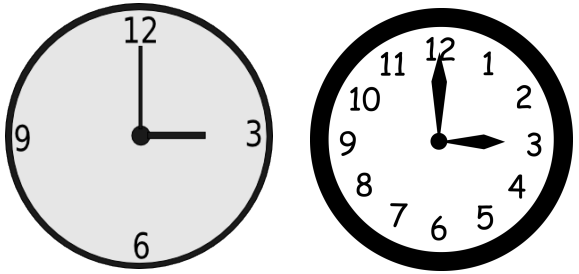




Statistics

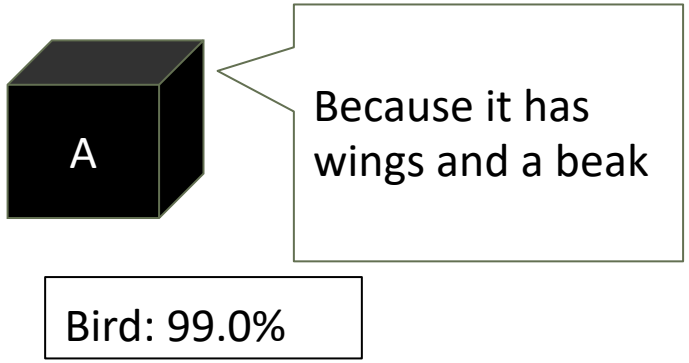
Do you know <u>correlation ≠ causality</u>?

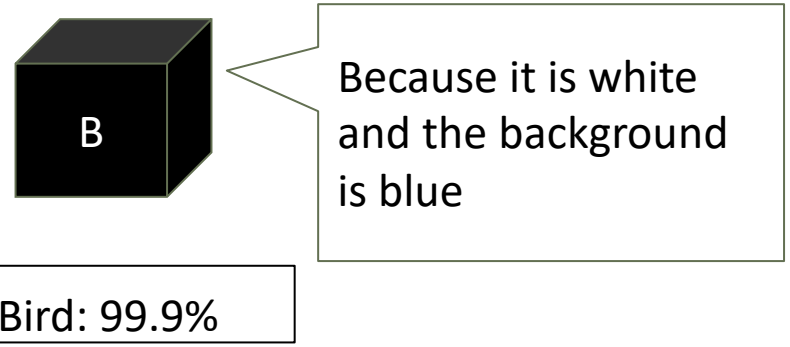




Practical Machine Learning

Do you know ML explanation?





Which model are you going to choose?

Marking Scheme

Assignments: 11 x 4% = 44%

Blog Post:16%

- Depth (8%), Popularity (8%)

Final Project: 40%

- Proposal (2%), Milestone (8%), Poster (15%), Report (15%)

Bonus: Contribute to 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 13	Form a team (3-5 members)
Monday Feb 3	Blog Post Submission
Monday Feb 10	Final Project Proposal
Monday Mar 9	Final Project Milestone
Monday Mar 30	Blog Post Popularity
April 6 - 19	Final Project Poster Session
April 6 - 19	Final Project Video/Code/Report Submission

Lectures/Labs

Lectures (2 hours/week)

- Monday 12:30 - 2:20

Labs (4 hours/week)

- Group A: Tues 9–10:50, Thurs 9–10:50
- Group B: Wed 11:30–13:20, Fri 10:30–12:20
- Group C: Wed 13:30–15:20, Fri 12:30–14:20

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

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

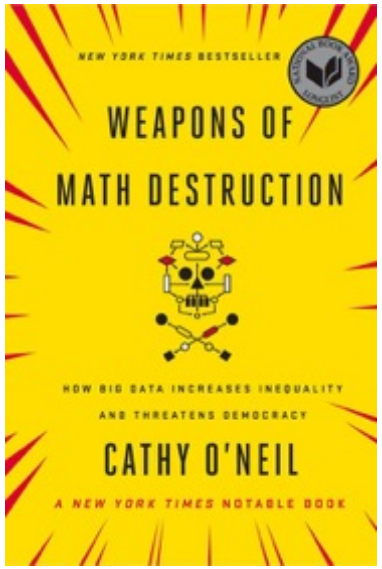
- We will do plagiarism check
- If you got caught, your final mark would be deducted by 30%

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