

INFO 1998: Introduction to Machine Learning



CDS Education

We explore, learn, and educate big minds.

Lecture 10: Real-World Applications of Data Science

INFO 1998: Introduction to Machine Learning

*“B****es be yearning my earnings concerning **machine learning**,
Your girl started flirting when she saw my code churning”*

Young’s Modulus



CDS Education

We explore, learn, and educate big minds.

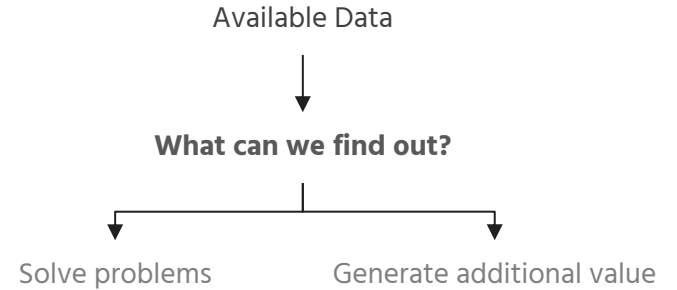
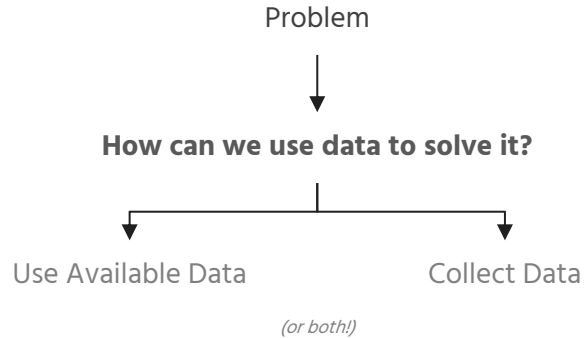
Agenda

- Data-Driven Thinking
- Data Science in the Real World
- An Important Note on Ethics
- Ideating Side Projects
- Next Steps
- Courses at Cornell
- Careers in Data Science



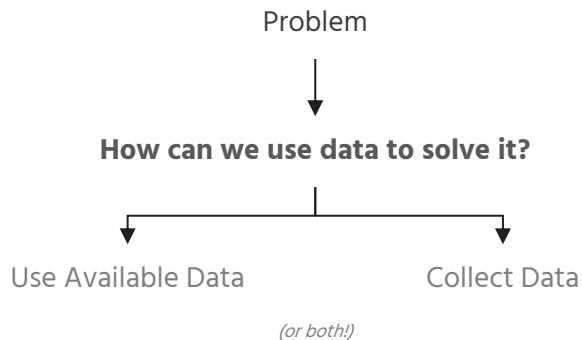
Data-Driven Thinking

Going beyond traditional problem-solving



Data-Driven Thinking

Traditional Approach



Sample Problems

1. Who will win the 2020 Elections?

[FiveThirtyEight](#)

2. Does a patient have lung cancer?

[Data Science Bowl '17](#)

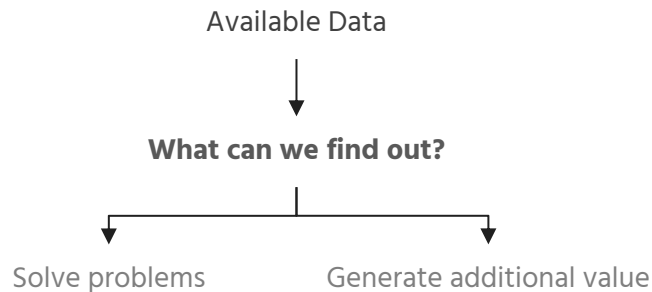
3. Roads are unsafe with increasing traffic.

[DataKind & Vision Zero](#)



Data-Driven Thinking

The New Approach



Sample Data

1. What are the interests of internet user X?

[Advertising](#)

2. All Traffic Data in a city

Optimizing signals, opening up a new business, traffic sign placement

3. All hip-hop music lyrics ever

[RapStats](#), [Rap Analysis Project](#)



Let's think data!

Exploring Real-World Applications

1. Advertising

- Case Study - Cambridge Analytica: Data Science in Political Campaigning

2. Healthcare

- Case Study – BiliScreen: A Selfie to Diagnose Pancreatic Cancer

3. Media

- Case Study – How Netflix Keeps You Hooked

4. Social Impact

- Case Study – Fighting Human Trafficking with Data

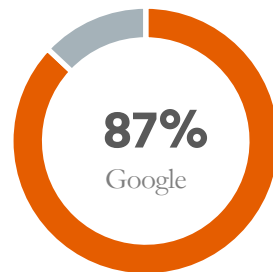


Advertising

Machine Learning: The Modern Mad Men

Context

Some Big Tech giants earn their the bulk of their revenue through ads
One usually earns money when the ad is 'clicked' by the user (this differs!)
Users are most likely to click on ads when the ads are relevant to them
Ads could be tailored to users only when there is data on the users




c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../cutefallskirts	143s	07:56:31
6d1wd34	128.45.313	(62.3, 89.5)	SYR	NY	.../shoestobuy	9s	07:56:35
3d5wf31	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23

Sample Data (Extremely small slice): What can you interpret?



Advertising

c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../cutefallskirts	143s	07:56:31
6d1wd34	128.45.313	(62.3, 89.5)	SYR	NY	.../shoestobuy	9s	07:56:35
3d5wf31	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23



c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../cutefallskirts	143s	07:56:31
	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23
6d1wd34	128.45.313	(62.3, 89.5)	SYR	NY	.../shoestobuy	9s	07:56:35

Objective: Get data on the users



Advertising

c_id	ip	loc	city	state	link	time	timestamp
3d5wf31	128.83.126	(68.3, 98.5)	Hoboken	NJ	../cutefallskirts	143s	07:56:31
	341.34.345	(68.5, 98.6)	NYC	NY	../excelhelp	552s	14:42:23

Hypotheses:

- Lives in NJ and works in NYC
- Lives in area with average rent: \$r
- Lives in area with average income: \$i
- Works in area with average salary: \$s
- Falls in k income bracket (Estimated)
- Takes NJTransit to work
- Takes the 67 Train at 8:05am
- Works at XYZ Company
- Works in Business/Data Analytics
- Is a Female
- Is interested in topics A, B, C

With **enough data** and **testing**, the hypotheses could be affirmed or rejected.



Cambridge Analytica: Data Science in Political Campaigning

Case Study

Overview

Cambridge Analytica combined *data analytics*, *behavioral sciences*, and *innovative ad tech* to influence voters. Widely regarded as instrumental in the result of the 2016 Elections, and many more across the globe.

Methodology



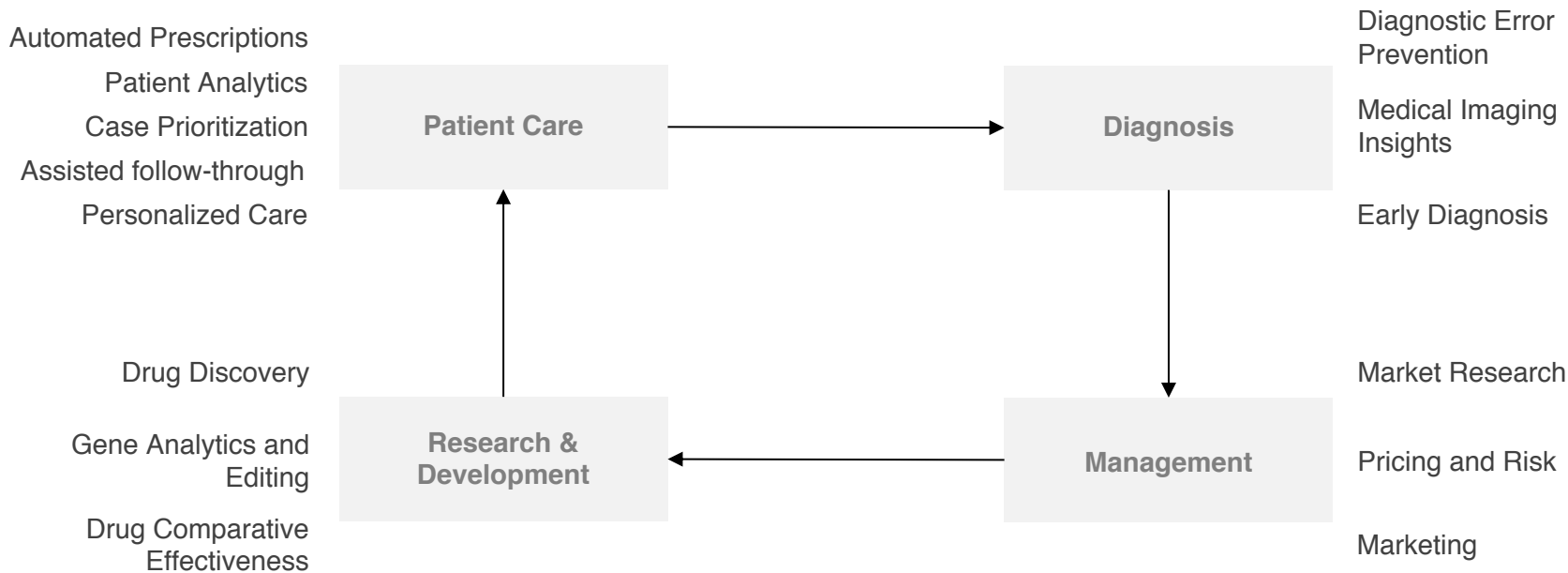
Example



Source: Cambridge Analytica

Healthcare

All-round betterment in the healthcare industry



Source: <https://blog.appliedai.com/healthcare-ai/>

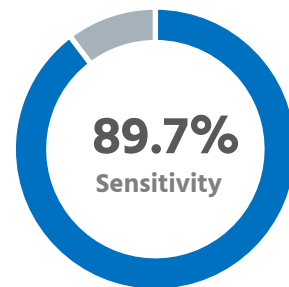
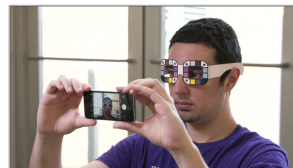
BiliScreen: A Selfie to Diagnose Pancreatic Cancer

Case Study

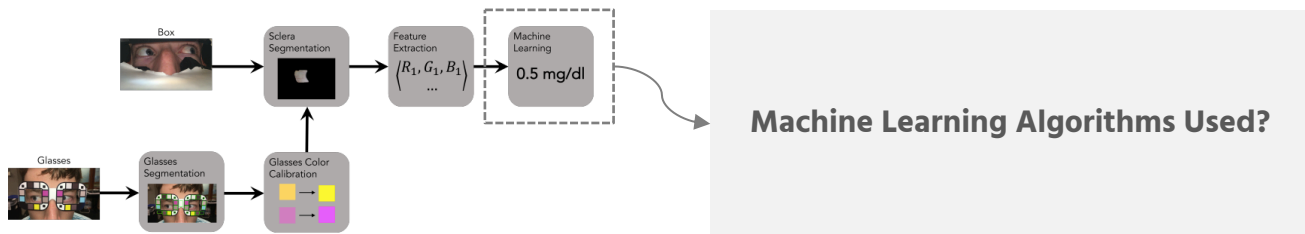
Overview

A smartphone app that captures pictures of the eye and produces an estimate of a person's bilirubin level

Uses: (1) A 3D-printed box that controls the eyes' exposure to light
(2) Paper glasses with colored squares for calibration



Methodology



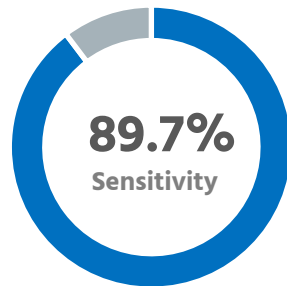
BiliScreen: A Selfie to Diagnose Pancreatic Cancer

Case Study

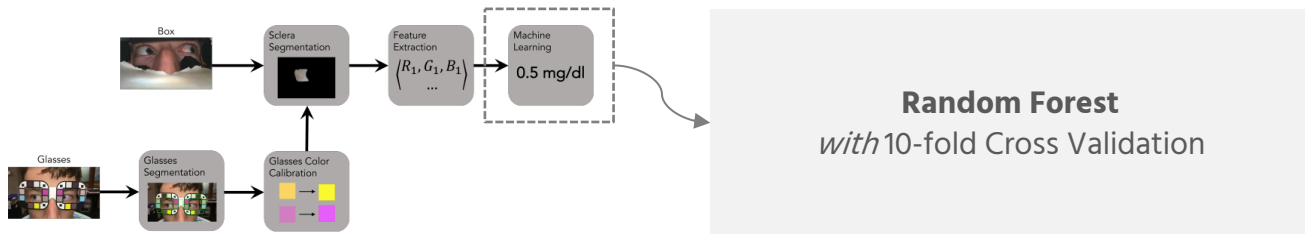
Overview

A smartphone app that captures pictures of the eye and produces an estimate of a person's bilirubin level

Uses: (1) A 3D-printed box that controls the eyes' exposure to light
(2) Paper glasses with colored squares for calibration



Methodology



Media: Recommender Systems

How Netflix keeps you hooked

Overview

Most of Netflix's views (~80%) come through recommendations

The famous Netflix Challenge offered \$1m to the participant that could do better than Netflix's recommender system

These algorithms are relatively simple and intuitive, but extremely effective

c_id	movie	tags	time	duration	rating
A	Avengers	Action, Superhero	07:56:31	112m	5/5
	Mr. Bean	Comedy	07:36:35	3m	2/5
B	Batman	Superhero	14:42:23	59m	4/5
	Black Mirror	Sci-Fi	07:56:34	142m	5/5

Sample: What would you recommend A next?

Usually, many other features and tags for the movies/shows would exist in the database as well

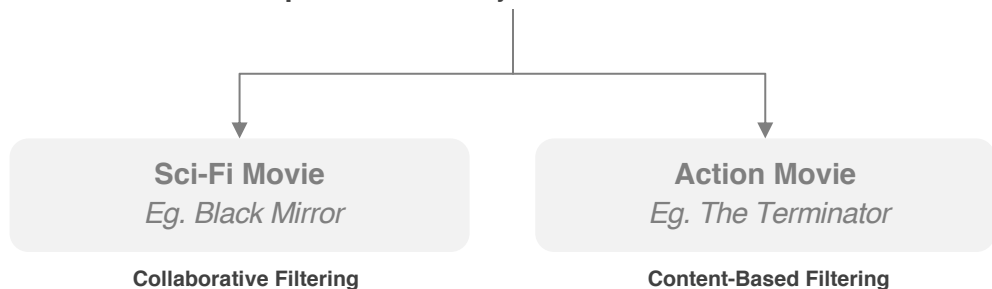


Media: Recommender Systems

How Netflix keeps you hooked

c_id	movie	tags	time	duration	rating
A	Avengers	Action, Superhero	07:56:31	112m	5/5
	Mr. Bean	Comedy	07:36:35	3m	2/5
B	Batman	Superhero	14:42:23	59m	4/5
	Black Mirror	Sci-Fi	07:56:34	142m	5/5

Sample: What would you recommend A next?



Where else are recommender systems applicable?



Social Impact

Data Science for Social Good

Overview

Advanced analytics for social impact is becoming increasingly popular due to innumerable low-cost and high-impact applications

- [Marine Data Science](#)
- [Data Science in Agriculture](#)
- [Big Data for Refugee Resettlement](#)
- [Saving Water in Drought-Stricken California](#)
- [Expanding Economic Opportunity for low-income people](#)
- [Data Science to Combat Trafficking](#)



Predicting End Location: Tackling Human Trafficking

Case Study

Overview

Human trafficking is a great cause of concern, especially in developing countries
ML could be leveraged to aid ground rescue operations for trafficking victims



Predicting End Location: Tackling Human Trafficking

Case Study

Overview

Human trafficking is a great cause of concern, especially in developing countries
ML could be leveraged to aid ground rescue operations for trafficking victims



Other Applications

Education

Adaptive-learning technology that could **recommend** material based on student's success and engagement

Public Sector

Identifying tax-fraud using alternate data such as browsing history, retail data, or payments history.

Crisis

Predicting the progression of wildfires to optimize the response of firefighters.



An Important Note on Ethics

The [ACM Code of Ethics](#) and [the Ethical Guidelines for Statistical Practice](#) (American Statistical Association) are good places to start. It's easy to get caught up in the technical challenge, but it is important to know that your work may affect other people directly or indirectly, now or in the future. Ask yourself the following questions often:

- Does your data or analysis impede on anyone's privacy?
- Did the people give consent for their data to be used?
- Were the people given the option to opt out?
- Who has the right of access to your data?
- Who owns the data?
- Was the data anonymized sufficiently?
- Was there any bias in your dataset against certain sections of the society?
- Are you introducing any bias?
- Should you include any features that may be discriminatory?
- Is your analysis transparent?
- Are the end users aware of shortcomings?

'Anonymous' Data? [Think again.](#)



Looking Forward



Ideating Side Projects

[Towards Data Science](#) is a good place to start for quick reads. You could also follow pages and personalities on your preferred social media.

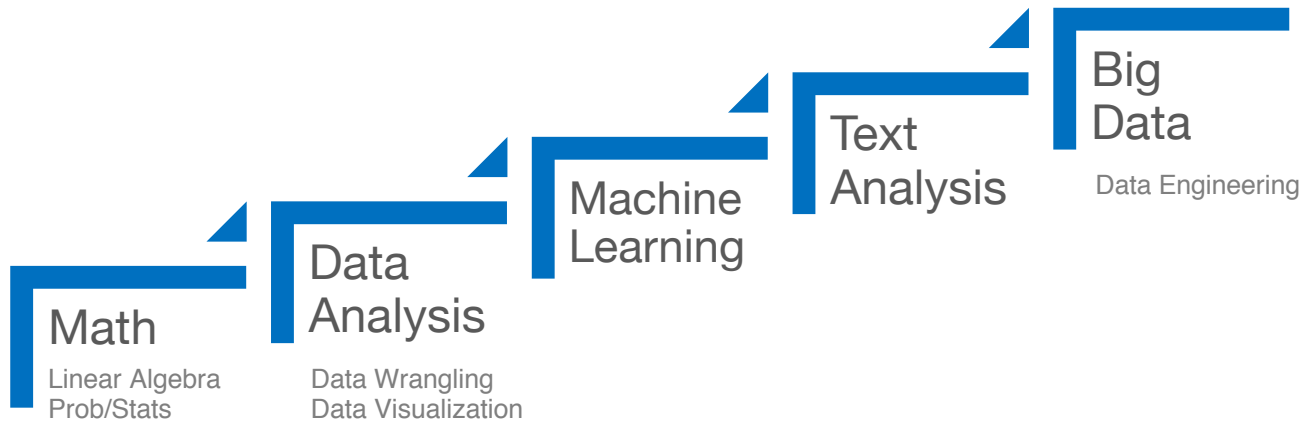
I recommend [Cassie Kozyrov](#)'s articles!

1. Dig into **your own data** – Health, Messages, Spotify, etc.
2. Make **something you'd use**.
3. Look at issues from a **social/economic/political** lens.
4. ...There's always Kaggle and data.gov



Next Steps

Path to becoming a data scientist



Gathering, EDA, Deployment

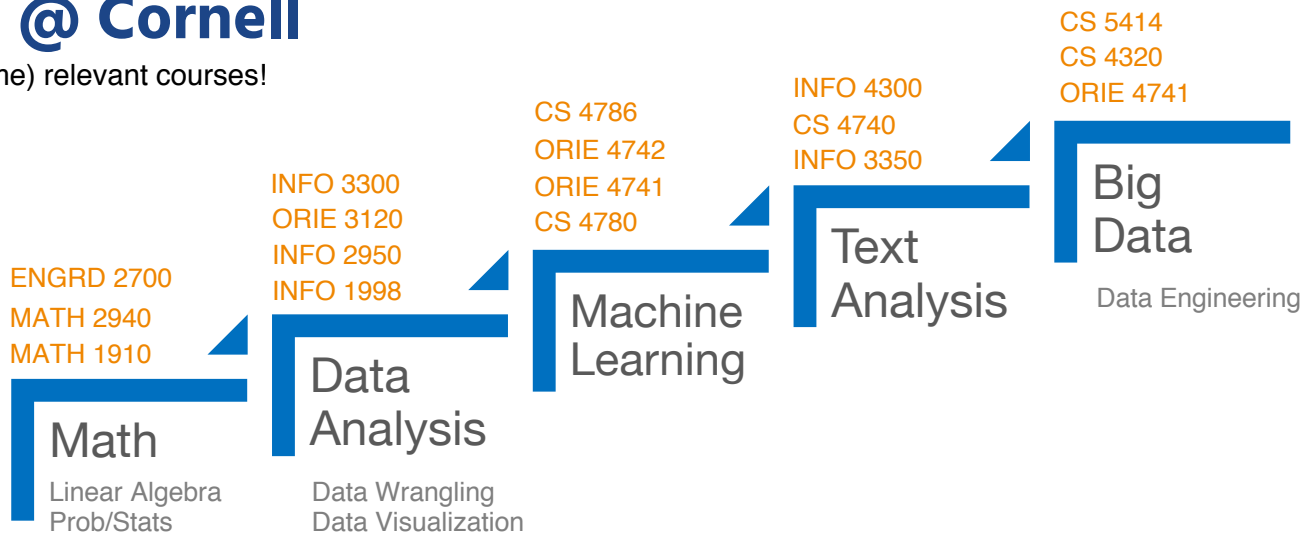
Software Engineering Skills

Business Acumen



Courses @ Cornell

Examples of (some) relevant courses!



Other: CS 4700, CS 4670, CS 4787, etc.

Gathering, EDA, Deployment **INFO 2950**

Software Engineering Skills **CS 1110 CS 2110 CS 5150**

Business Acumen & Domain Knowledge **Read!**



Note: This is not an official list, and does not represent the views of Cornell Data Science.

Careers

Common roles and their meanings

Data Analyst

These are typically the roles right out of undergrad. You'll likely be working with SQL/Excel (and maybe a little bit of Python/R).

Data Scientist

This role typically covers responsibilities additional to those that data analysts have. You'll be expected to have a strong understanding of math fundamentals, and machine learning models. It's also a good idea to be well-versed in programming.

Data Engineer

As a data engineer, you'll be managing the data infrastructure – building data pipelines, pushing code into production, etc. You would ideally like to be well-versed in software development and have exposure to other software and tools your target companies use.

Machine Learning Engineer

This is similar to the data scientist role, but is more specific to building machine learning models. You would like be required to have a robust knowledge of applied math and software development.



Careers

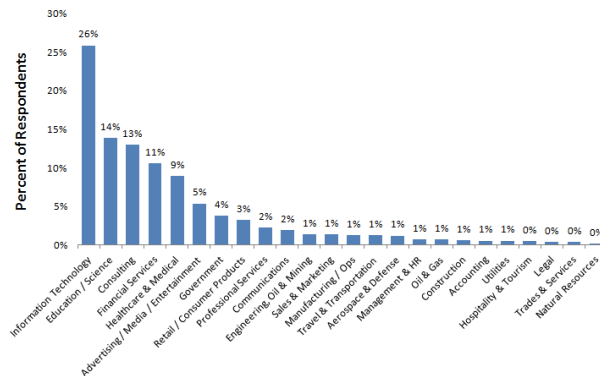
Product Analytics vs Business Intelligence

Product Analytics

Focused on a certain product and the behaviors of the user's product. For example, you may be working on boosting customer engagement using clickstream data.

Business Intelligence

Focused on creating business insights from your products/services and informing internal decisions. For example, you may be generating reports of number of users on your platform.



Source: Business Broadway

That's all folks!

Just Kidding

- **Final Project Due:** May 13, 2020
- **Course Feedback Form** out soon!
- **Course Staff Invitations** out in summer
- **Office Hours go on** until May 13, 2020
- Stay tuned for **CDS Recruitment** next semester!
- **Get in touch:** tb444@cornell.edu

Thank you all for taking this class, and for an incredible semester.

Good luck on finals, and stay safe!



CDS Education

We explore, learn, and educate big minds.