CPSC 340 and 532M:
Machine Learning and Data Mining

Andreas Lehrmann and Mark Schmidt
University of British Columbia, Fall 2022
https://www.students.cs.ubc.ca/~cs-340

Held on the traditional, ancestral, and
unceded territory of the Musqueam people

Some images from this lecture are taken from Google Image Search, contact me if you want the reference

Big Data Phenomenon

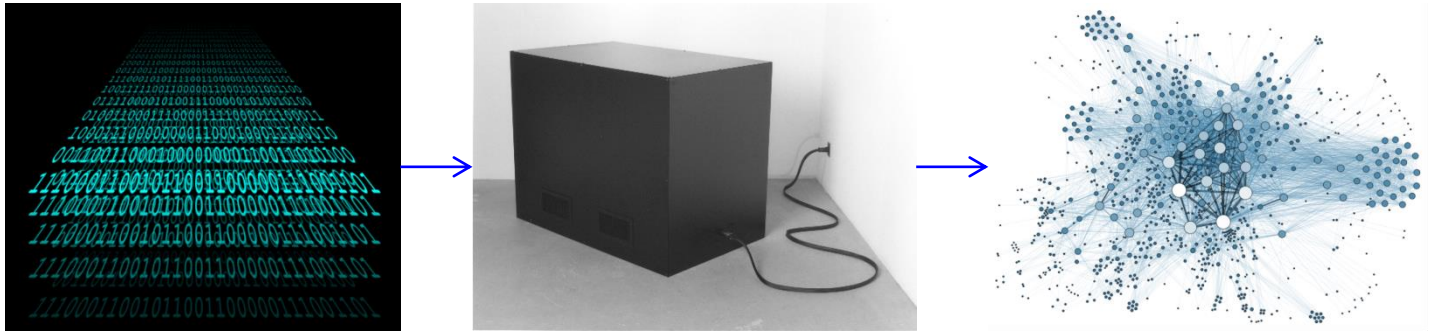
- We are collecting and storing data at an unprecedented rate.
- Examples:
  - YouTube, Facebook, MOOCs, news sites.
  - Credit cards transactions and Amazon purchases.
  - Transportation data (Google Maps, Waze, Uber)
  - Gene/protein sequencing/expression/structures.
  - Maps and satellite data.
  - Camera traps and conservation efforts.
  - Phone call records and speech recognition results.
  - Video game worlds and user actions.

Big Data Phenomenon

- What do you do with all this data?
  - Too much data to search through it manually.
- But there is valuable information in the data.
  - How can we use it for fun, profit, and/or the greater good?
- Data mining and machine learning are key tools we use to make sense of large datasets.

Data Mining

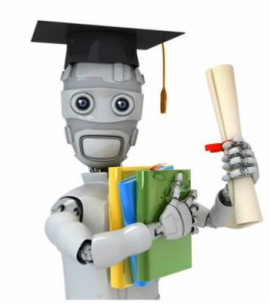
- Automatically extract useful knowledge from large datasets.

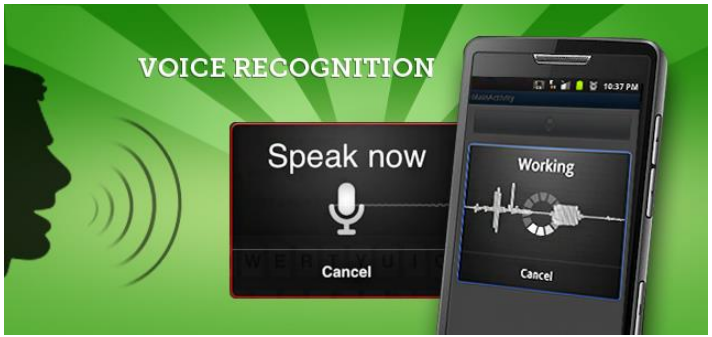


- Usually, to help with human decision making.

Machine Learning

- Using computer to automatically detect patterns in data and use these to make predictions or decisions.

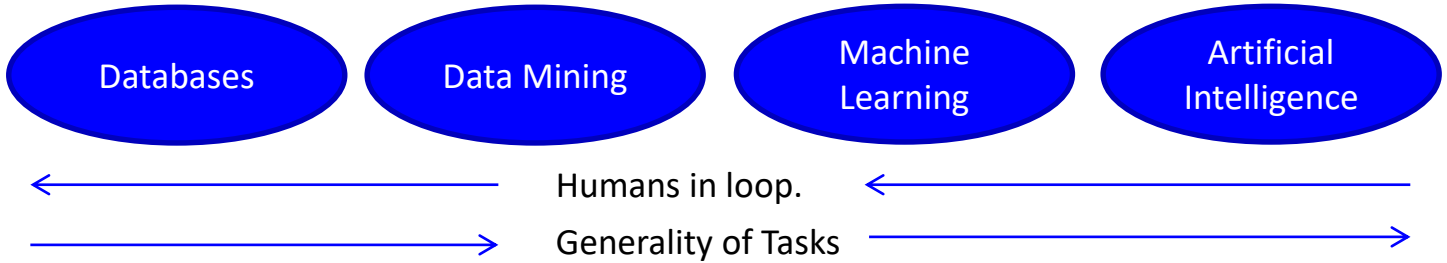




- Most useful when:
  - We want to automate something a human can do.
  - We want to do things a human can't do (look at 1 TB of data).

Data Mining vs. Machine Learning

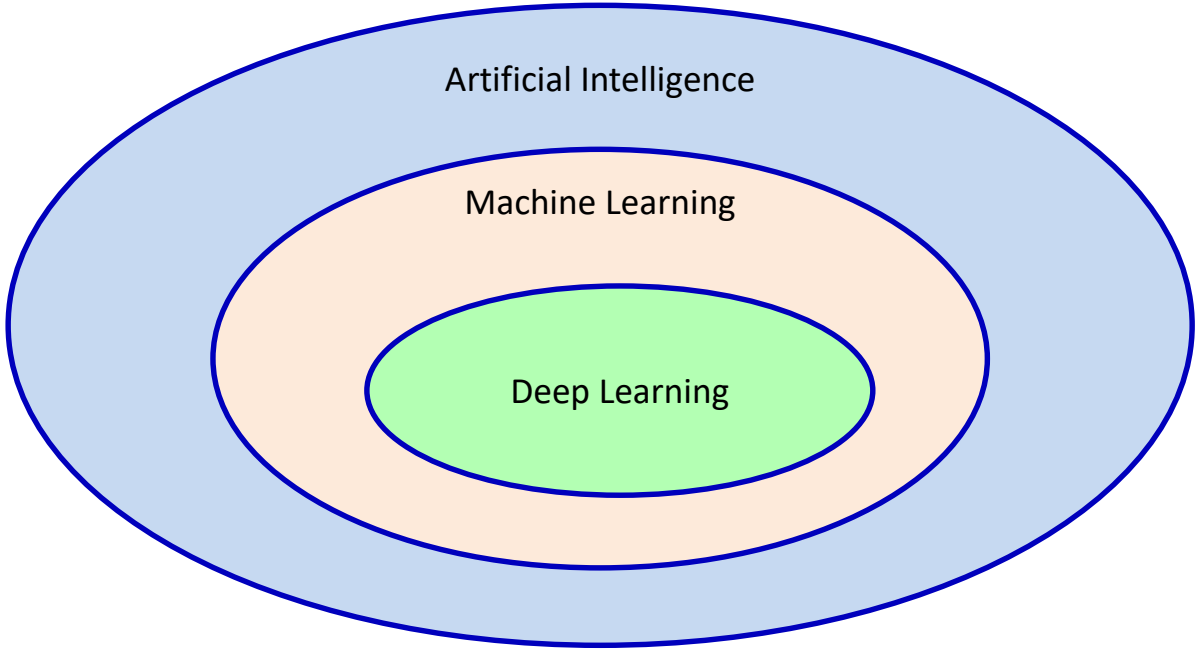
- Data mining and machine learning are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.



- Both are similar to statistics, but more emphasis on:
  - Large datasets and computation.
  - Predictions (instead of descriptions).
  - Flexible models (that work on many problems).

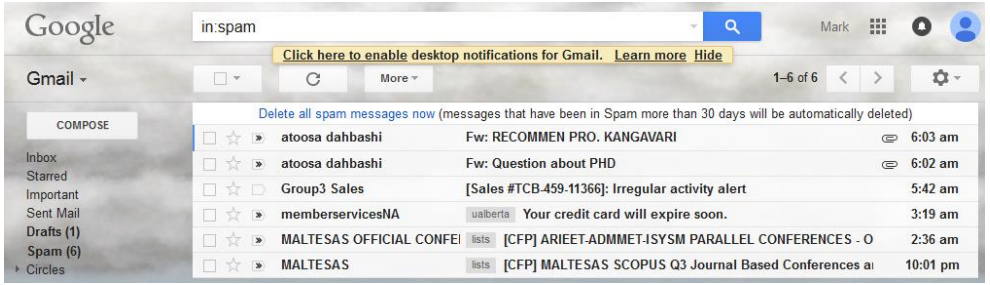
Deep Learning vs. Machine Learning vs. AI

- Traditional we've viewed ML as a subset of AI.
  - And "deep learning" as a subset of ML.

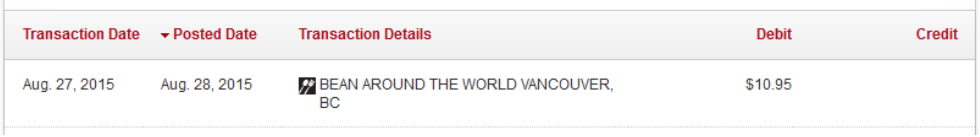


Applications

- Spam filtering:



- Credit card fraud detection:

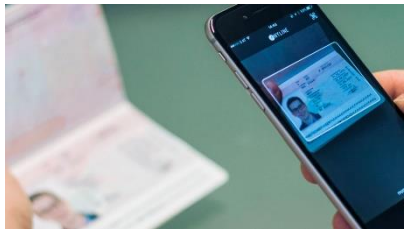


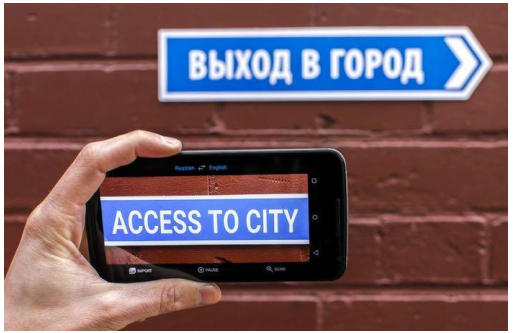
Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

- Product recommendation:

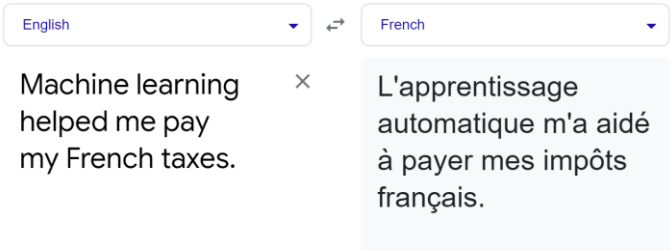
Applications

- Optical character recognition:





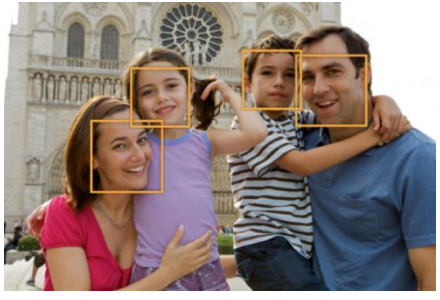
- Machine translation:



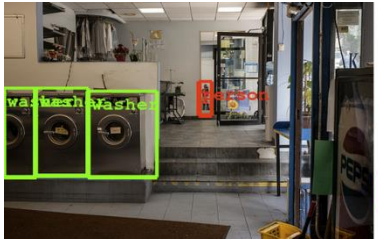
- Speech recognition:

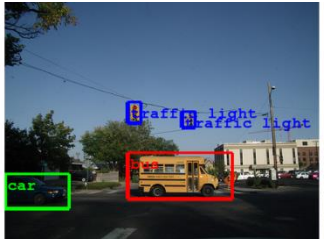
Applications

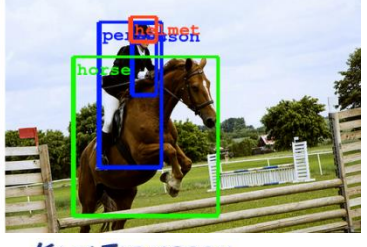
- Face detection/recognition:



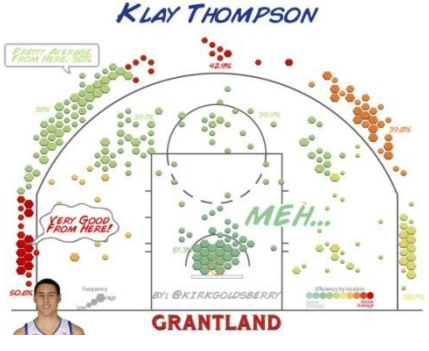
- Object detection:





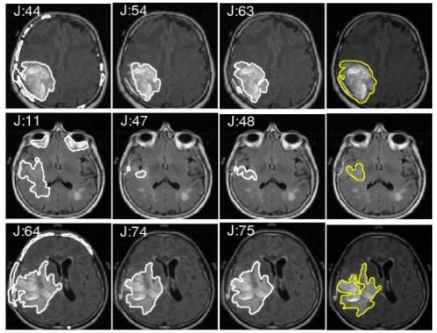


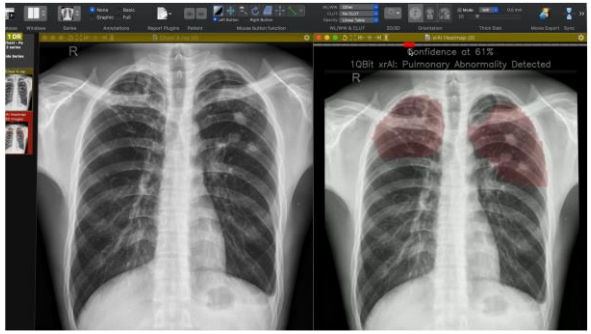
- Sports analytics:



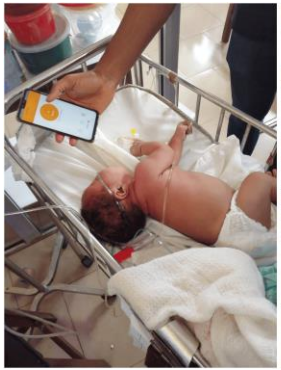
Applications

- Medical imaging:





- Medical diagnostics:



- Self-driving cars:

Applications

- Image completion:

- Image annotation:

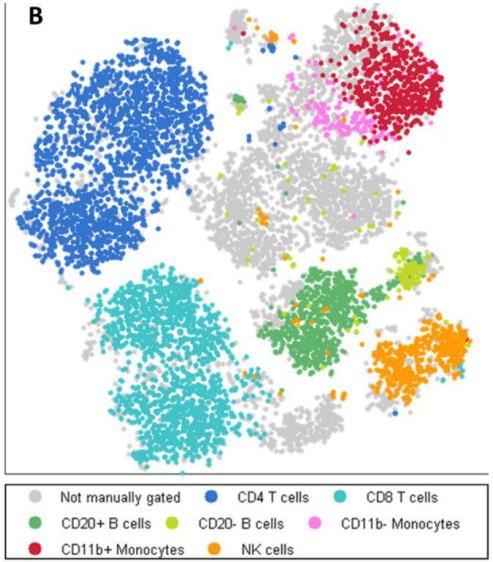
a cat is sitting on a toilet seat
logprob: -7.79

a display case filled with lots of different types of donuts
logprob: -7.78

a group of people sitting at a table with wine glasses
logprob: -6.71

Applications

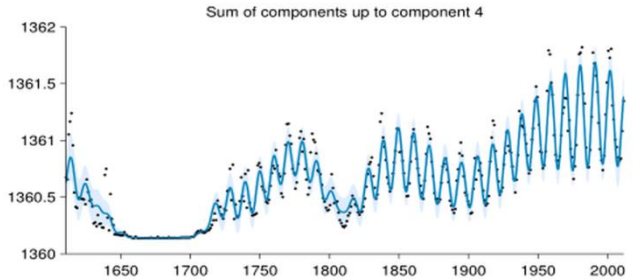
- Discovering new cancer subtypes:



- Automated Statistician:

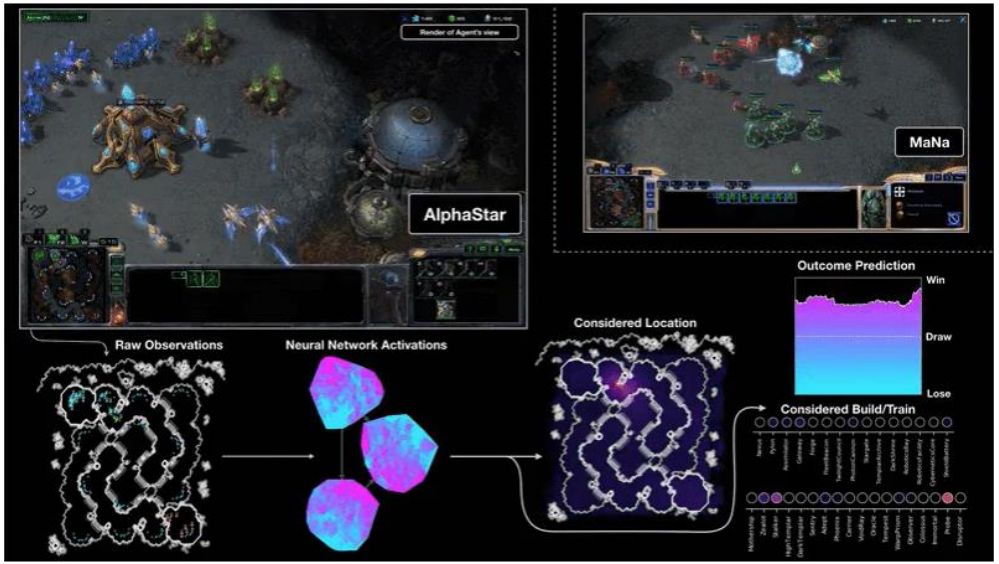
2.4 Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.



Applications

- Beating humans in Go and Starcraft:

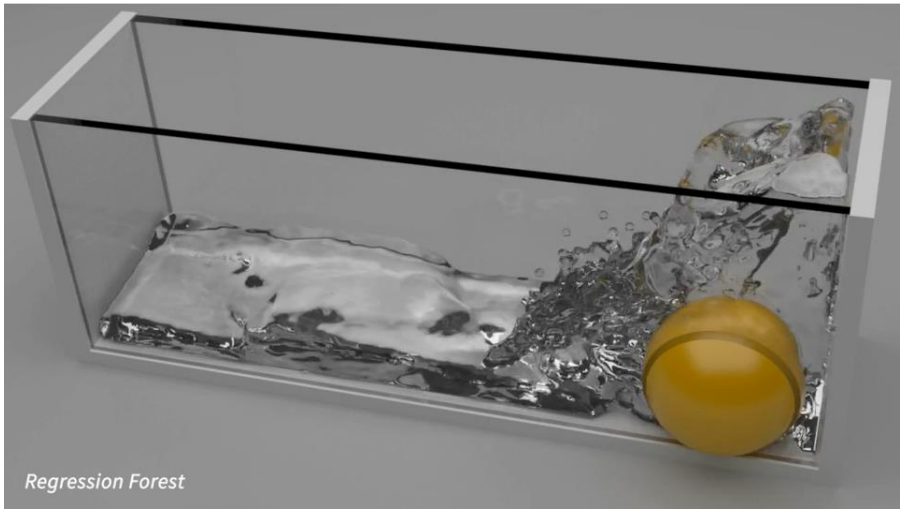


Applications

- Mimicking artistic styles (video).

Applications

- Fast physics-based animation:

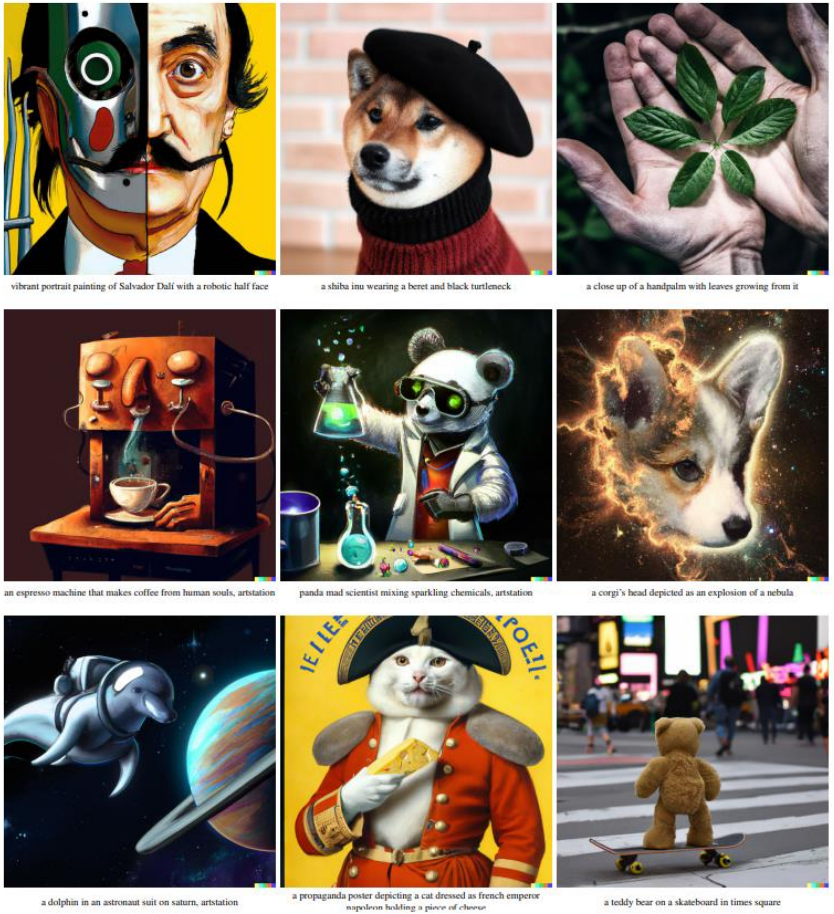


- Character animation (video):

- Recent work on generating text/music/voice/poetry/dance.

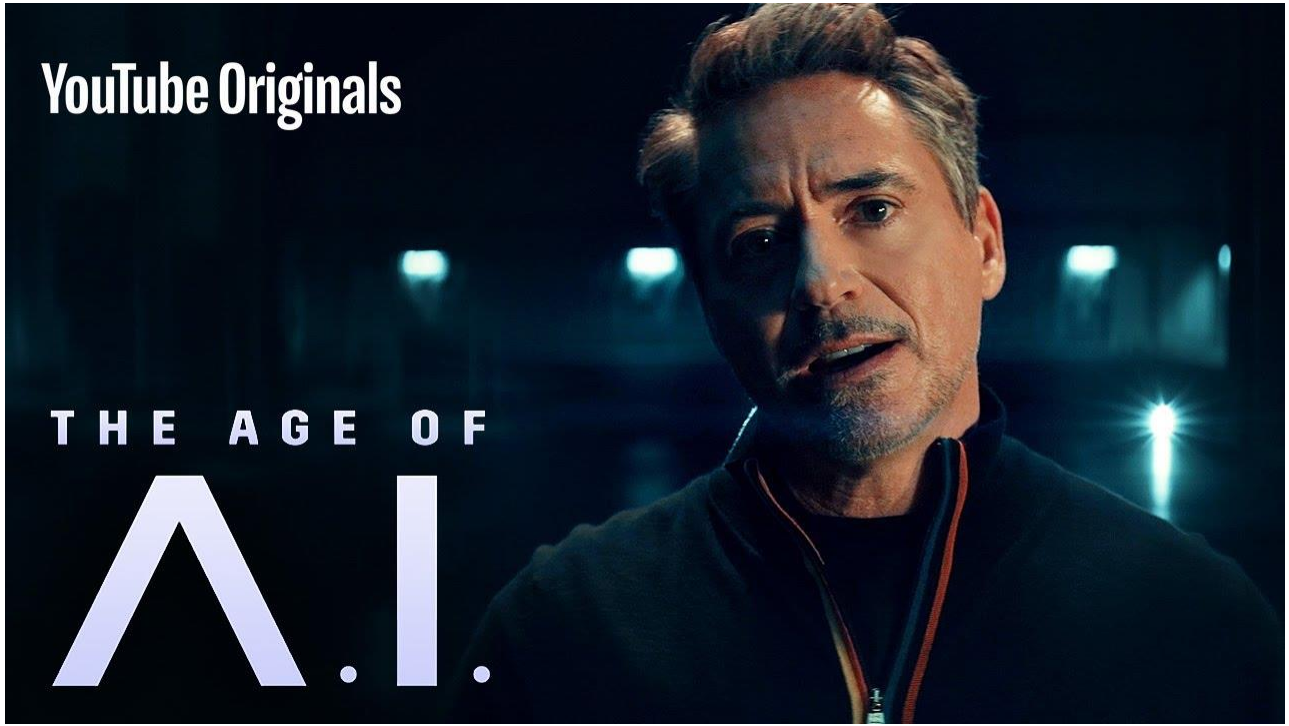
Applications

- Generating images from text:



Applications

- “Age of AI” YouTube series:

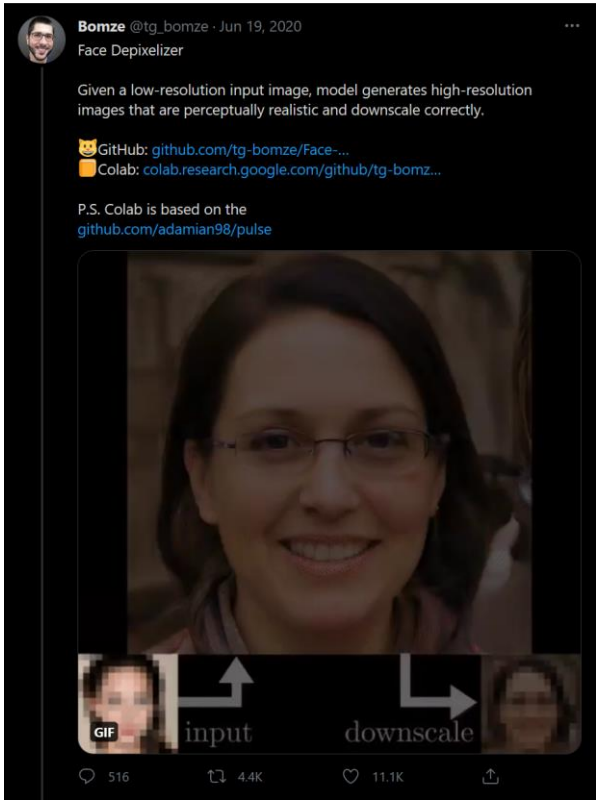


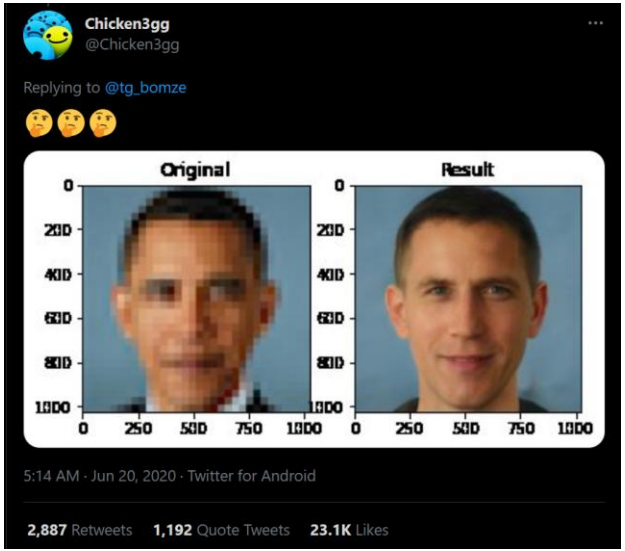
- Summary:
  - There is a lot you can do with a bit of statistics and a lot data/computation.

- We are in exciting times.
  - Major recent progress in many fields:
    - Speech recognitio, computer vision, natural language processing, iamge generation.
  - Things are changing a lot on the timescale of 3-5 years.
  - NeurIPS conference sold out in ~11 minutes in 2019 (switched to lottery).
  - A bubble in ML investments (most “AI” companies are just doing ML).

- But it is important to know the limitations of what you are doing.
  - A huge number of people applying ML are just “overfitting”.
    - Their methods do not work when they are released “into the wild”.

Failures of Machine Learning





Failures of Machine Learning

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent | Oct 10, 2018, 7:09am EDT

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

Via The Guardian | Source TayandYou (Twitter)

Uber self-driving car kills pedestrian in first fatal autonomous crash

by Matt McFarland @mattmcfarland

March 19, 2018: 1:40 PM ET

- ML/AI worship is not healthy.
- Learn how things work "under the hood", and have a healthy dose of skepticism!

Course Outline

- Next class discusses “exploratory data analysis”.

- After that, the remaining lectures focus on five topics:
  1) Supervised Learning.
  2) Unsupervised learning.
  3) Linear prediction.
  4) Latent-factor models.
  5) Deep learning.

- “What is Machine Learning?” (overview of many class topics)

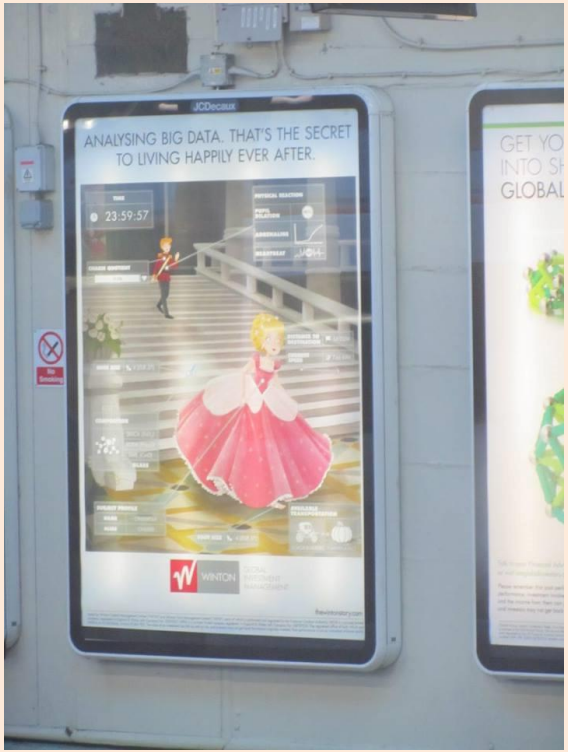
Bonus Slides

- I will include a lot of “bonus slides”.
  - May mention advanced variations of methods from lecture.
  - May overview big topics that we don’t have time for.
  - May go over technical details that would derail class.

- You are not expected to learn the material on these slides.
  - But they are useful if you want to take 440 or work in this area.

- I will use this colour of background on bonus slides.

Photo I took in the UK on the way home from the “Optimization and Big Data” workshop:



Less-inspirational quote: “Without data you're just another person with an opinion.” W.E. Deming

This is the end of the lecture.
(Future lectures will end on a “Summary” slide.)

The slides after the “Summary” slide are typically “bonus” material related to the topics of the lecture.

Bonus Slide: “Machine Learning” vs. “Data Minig”

- Machine learning and data mining have many similarities (as do other fields like statistics and signal processing), and the similarity is increasing due to the 'arXiv' effect (people from both fields can now easily read each other's papers and are using standard notation).
- However, as a subjective answer I would say that the focuses are different. Data mining is broader in scope and includes things like how to organize data, models that simply look up answers or are based on counting (KNN and naive Bayes are also often covered in data mining, and in data mining there is a greater focus on interpretable models), and tasks like information visualization. Machine learning is more narrow, focusing largely on the modeling aspect, generalization error, and using methods that rely on numerical optimization or high-dimensional integration (that may not necessarily be interpretable).
- Another subjective comment would be that data mining often focuses on tools that help professionals analyze their data, while machine learning often focuses on automating data analysis. For example, here is a recent very-interesting project by some machine learning folks from Cambridge and MIT:
  - http://www.automaticstatistician.com