CPSC 340:
Machine Learning and Data Mining

Danica Sutherland (101) and Mike Gelbart (103)

University of British Columbia, Fall 2021

https://github.com/ubc-cs/cpsc340-2021w1

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

- Welcome to the course!

Lectures

- All slides will be posted online (before lecture, and final version after).
  - There are also Jupyter notebook companions to some lectures, will also post.
- Please ask questions: you probably have similar questions to others.
  - I may deflect to the next lecture or Piazza for certain questions.
- Be warned that the course we will move fast and cover a lot of topics:
  - Big ideas will be covered slowly and carefully.
  - But a bunch of other topics won't be covered in a lot of detail.
- Isn't it wrong to have only have shallow knowledge?
  - In this field, it's better to know many methods than to know 5 in detail.
    - This is called the "no free lunch" theorem: different problems need different solutions.

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’re useful if you want to take 440 or work in this area.

- I’ll use this colour of background on bonus slides.

Videos from Previous Offering

- Videos of Mike’s January 2018 offering of the course:
  - https://www.youtube.com/playlist?list=PLWmXHcz_53Q02ZLeAxigki1JZFfCO6M-b

- You may find these useful:
  - Material is very similar
  - Though notation has changed in a few cases

Essential Links

- Please bookmark the course webpage:
  - https://github.com/ubc-cs/cpsc340-2021w1
  - Contains lecture slides, assignments, optional readings, additional notes.

- You should sign up for Piazza:
  - http://piazza.com/ubc.ca/winterterm12021/cpsc340
  - Can be used to ask questions about lectures/assignments/exams.
  - Also used for course announcements.
  - Most questions should be "public" and not "private";
    I will switch viewability of generally-relevant questions to "public".
  - Use Piazza instead of e-mail for questions.

- Canvas:
  - https://canvas.ubc.ca/courses/78047/assignments/syllabus
  - A list of all the places you might need to look for things

Different Sections of 340

- Section 101, 4-5pm, taught by Danica Sutherland
- Section 103, 2-3pm, taught by Mike Gelbart
  - Both sections have the same webpage, assignments, and exams.

- Lectures will cover almost exactly the same set of topics.
  - You will only be tested on material that appears in both sections.

Textbooks

- No required textbook.

- I’ll post relevant sections out of these books as optional readings:
  - Artificial Intelligence: A Modern Approach (Rusell & Norvig)
  - Introduction to Data Mining (Tan et al.)
  - The Elements of Statistical Learning (Hastie et al.)
  - Mining Massive Datasets (Leskovec et al.)
  - Machine Learning: A Probabilistic Perspective (Murphy)
- Most of these are on reserve in the ICICS reading room.
- List of related courses on the webpage, or you can use Google.

- Good online textbook covering mathematical background:
  - Mathematics for Machine Learning (Deisenroth et al.), Chapters 1-3 and 5-6.

Assignments

- There will be 6 Assignments worth 30% of final grade:
  - Usually a combination of math, programming, and very-short answer.

- Assignment 1 is posted (or about to be), and is due next Friday.
  - Submission instructions posted on webpage.
  - The assignment should give you an idea of expected background.
  - Make sure to submit before the deadline and check your submission.

- Start early, there is a lot there.
  - Don’t wait to see you if get off the waiting list to start.
  - You should be able to do the first few questions already.

Working in Teams for Assignments

- Assignment 1 must be done individually.

- Assignments 2-6 can optionally be done in pairs.
  - See submission instructions for how to specify partnership.
  - You don’t need to have the same partner for all assignments.
  - Generally “working together” is more effective than “divide and conquer”.

Programming Language: Python

- 3 most-used languages in these areas: Python, Matlab, and R.

- We will be using Python which is a free high-level language.
  - See some Python resources course webpage.
  - Expected to be able to learn a programming language on your own.

Late Day Policy for Assignments

- Assignments will be due at midnight (11:59:59pm + 1 second) on the due date.

- If you can’t make it, you can use “late days”:
  - For example, if assignment is due on a Friday:
    - Handing it in Saturday is 1 late day.
    - Handing it in Sunday is 2 late days.

  - There is no penalty for using “late days”,
    but you will get a mark of 0 on an assignment if you:
    - Use more than 2 late days on an assignment.
    - Use more than 4 late days across all assignments.

- We’ll release solutions to assignments after 2 “late days”.
  - We’ll try to put grades up within 10 days of this.

Midterm and Final

- Midterm worth 19% and a (cumulative) final worth 50%
  - TBD about open vs. closed-book.
  - No need to pass the final to pass the course (but recommended).

- Midterm is scheduled for 6:00-7:30pm October 21.
  - Let us know if you have a conflict that cannot be resolved.
  - Midterm will be computer-based, TBD about where.
- I don't control when the final is; don't make travel plans before December 22.

- There will be two types of questions:
  - 'Technical' questions requiring things like pseudo-code or derivations.
    - Similar to assignment questions, and will only be related topics covered in assignments.
  - 'Conceptual' questions testing understanding of key concepts.
    - All lecture slide material except "bonus slides" is fair game here.

Syllabus Quiz

- In addition to the assignments (30%), midterm (19%) and final exam (50%), the last 1% of you grade is the Syllabus Quiz.
- This is available on Canvas. You have multiple attempts.
- Due September 24th.

Reasons NOT to take this class

- Compared to typical CS classes, there is a lot more math:
  - Requires linear algebra, probability, and multivariate calculus (at once).
  - “I think the prerequisites for this course should require that students have obtained at least 75% (or around there) in the required math courses. As someone who who did not excel at math, I felt severely under prepared and struggled immensely in this course, especially seeing that I have taken CPSC courses in the past with similar math requirements, but were not nearly as math heavy as CPSC340.”
- If you’ve only taken a few math courses (or have low math grades), this course will ruin your life for the next 4 months.
- It’s better to improve your math, then take this course later.
  - A good reference covering the relevant math is here (Chapters 1-3 and 5-6).

Reasons NOT to take this class

- This is not a class on “how to use scikit-learn or TensorFlow or PyTorch”.
  - You will need to implement things from scratch, and modify existing code.

- Instead, this is a 300-level computer science course:
  - You are expected to be able to quickly understand and write code.
  - You are expected to be able to analyze algorithms in big-O notation.

- If you only have limited programming experience,
  this course will ruin your life for the next 4 months.
- It’s better to get programming experience, then take this course later.
  - Take CPSC 310 and/or 320 instead, then take this course later.

Reasons NOT to take this class

- Do NOT take this course expecting a high grade with low effort.

- Many people find the assignments very long and very difficult.
  - You will need to put time and effort into learning new/difficult skills.
  - If you aren’t strong at math and CS, they may take all of your time.

- Class averages have only been high because of graduate students.
  - NOT because this is an “easy” course; for most people it’s not.

CPSC 330 vs. CPSC 340

- There is a less-advanced ML course, CPSC 330: Applied ML
  - 330 emphasizes “how to use” tools, 340 emphasizes “how they work”.

  - Fewer prerequisites:
    - 330 spends more time on how-to and has basically no equations.
      - More “learning by doing” and less discussion of fundamental principles.
    - 330 spends more time on data cleaning, communicating results, and so on.
      - More emphasis on the entire “pipeline” of data of analysis.
    - 330 cannot be used as a prereq for the more-advanced CPSC 440.

  - You can take both for credit (better to take 330 first or at same time).

CPSC 340 vs. CPSC 540

- There is also a more-advanced ML course, CPSC 440:
  - Starts where this course ends.
  - More focus on theory/implementation, less focus on applications.
  - More prerequisites and higher workload.

- For almost all students, CPSC 340 is the better class to take:
  - CPSC 330/340 focus on the most widely-used methods in practice.
    - It covers much more material than standard ML classes like Coursera.
  - CPSC 440 focuses on less widely-used methods and research topics.
    - It is intended as a continuation of CPSC 340.
    - You'll miss important topics if you skip CPSC 340.

Waiting List and Auditing

- Right now only CS students can register directly.
  - All other students need to sign up for the waiting list to enroll.

- We're going to start registering people from the waiting list.
  - Being on the waiting list is the only way to get registered:
    - https://www.cs.ubc.ca/students/undergrad/courses/waitlists
  - You might be registered without being notified, be sure to check!
    - They might also ask to submit a prereq form; let me know if you have issues.

- Because the room is full, we may not have seats for auditors.
  - If there is space, I'll describe (light) auditing requirements then.

Getting Help

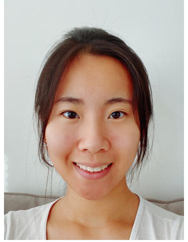
- Many students find the assignments long and difficult.
- But there are many sources of help:
  - TA office hours and instructor office hours.
    - Starting in the second week of class.
    - Times will be posted on the course webpage.
  - Piazza (for general questions).
  - Weekly tutorials (optional).
    - Starting in second week of class.
    - Will go through provided code, review background material, review big concepts, and/or do exercises.
  - Other students (ask your neighbor for their e-mail).
  - The web (almost all topics are covered in many places).

TA Cheat Sheet

- Daniel Ajisafe

- Dylan Green

- Helen Zhang



- Michael Liu

- Lironne Kurzman

- Anubhav Garg
- Ruiyu Gou

- Niloofar Khoshsiyar

- Paul Lin

Cheating and Plagiarism

- Read about UBC's policy on "academic misconduct" (cheating):
  - http://www.calendar.ubc.ca/Vancouver/index.cfm?tree=3,54,111,959
- When submitting assignments, acknowledge all sources:
  - Put "I had help from Sally on this question" on your submission.
  - Put "I got this from another course's answer key" on your submission.
  - Put "I copied this from the Coursera website" on your submission.
  - Otherwise, this is plagiarism (course material/textbooks are ok with me).
- At Canadian schools, this is taken very seriously.
  - Automatic grade of zero on the assignment.
  - Could receive 0 in course, be expelled from UBC, or have degree revoked.
  - We have actually given 0 to people before.

Code of Conduct

- Do not post offensive or disrespectful content on Piazza.
- If you have a problem or complaint, let me know (maybe we can fix it).
- Do not distribute any course materials without permission.
- Do not record lectures without permission.

- Think about how/when to ask for help:
  - Don't ask for help after being stuck for 10 seconds. Make a reasonable effort to solve your problem (check instructions, Piazza, and Google).
  - But don't wait until the 10th hour of debugging before asking for help.
    - If you do, the assignments could take all of your time.
  - Recommended length of time to struggle before asking: 10-30 min

- There will be no post-course grade changes based on grade thresholds:
  - 48% will not be rounded to 50%, and 70% will not be rounded to 72%, and so on.

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 Slide: “Machine Learning” vs. “Data Mining”

- 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

Next Topic: Covid Protocols

Covid Protocols

- Masks in indoor spaces
  - If you are not wearing a mask (properly), I will ask you to put one on.
  - If you do not, I will ask you to leave (unless you have an exemption).
  - Please don’t eat in class; if you bring a drink, lift your mask for each sip.
- Vaccination
  - We cannot require vaccination, but PLEASE get vaccinated.
  - Please.
- Stay home if you are sick

  … no matter what you think you have

  … even if you’re sure it’s not Covid

  PLEASE DO NOT COME TO CAMPUS

Covid Protocols

- **Attendance is NOT mandatory!**
- If you come, try to sit in the same area every time.
- Lectures will be recorded (and hopefully live-streamed)
  - Section 103 (2pm): https://ubc.ca.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=9ac1674d-46f5-4903-b28e-ad970125e84a
  - Section 101 (4pm): https://ubc.ca.panopto.com/Panopto/Pages/Sessions/List.aspx#folderID=%2224f3b12e-8d95-41b4-a024-ad97012a2e1d%22
- Tutorials are optional (and may also be recorded)

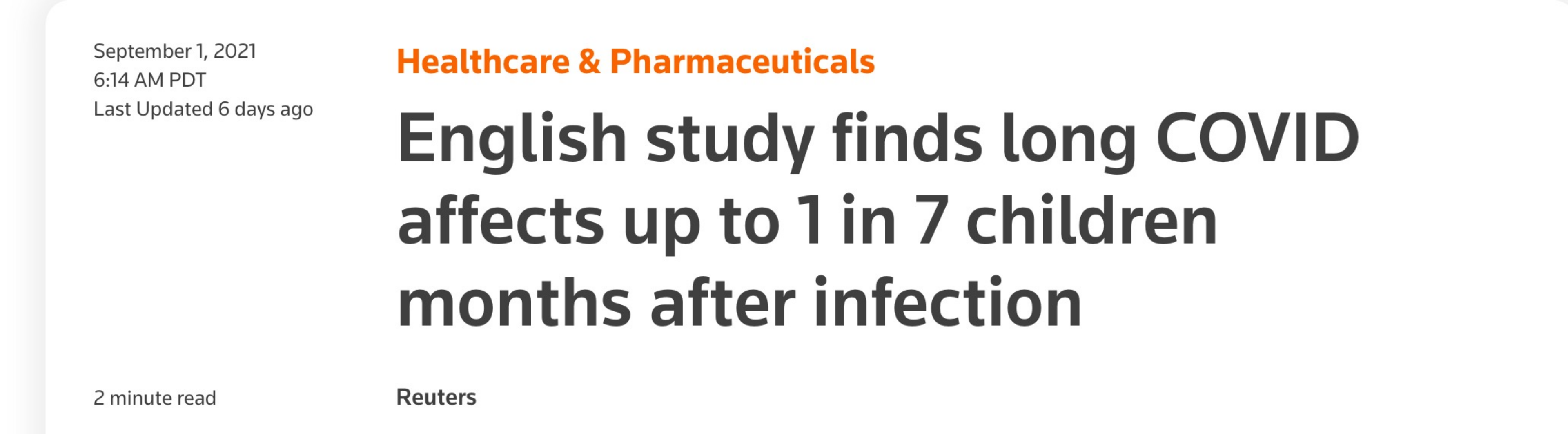
A personal plea…

Sophie Gelbart:

NOT VACCINATED

Alexander Gelbart:

NOT VACCINATED



September 1, 2021
6:14 AM PDT
Last Updated 6 days ago

Healthcare & Pharmaceuticals

English study finds long COVID affects up to 1 in 7 children months after infection

2 minute read

Reuters

- They each have an expected 80+ years ahead of them
- Our decisions next week might affect others next month/year.
- Although you’ve never met these people, your lives are entangled and they are relying on you.

- I’m not the only person here with vulnerable people at home.
- Please keep us safe.

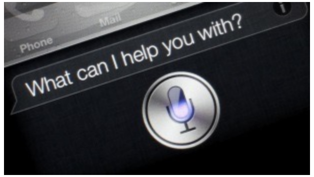
Course entry survey

- This helps us get to know you.
- Please take the next 5 min to fill it out: https://bit.ly/340-surv

Up next: Motivation

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 expression data and protein interaction assays.
  - Maps and satellite data.
  - Large hadron collider and surveying the sky.
  - Phone call records and speech recognition results.
  - Video game worlds and user actions.


What can I help you with?

Big Data Phenomenon

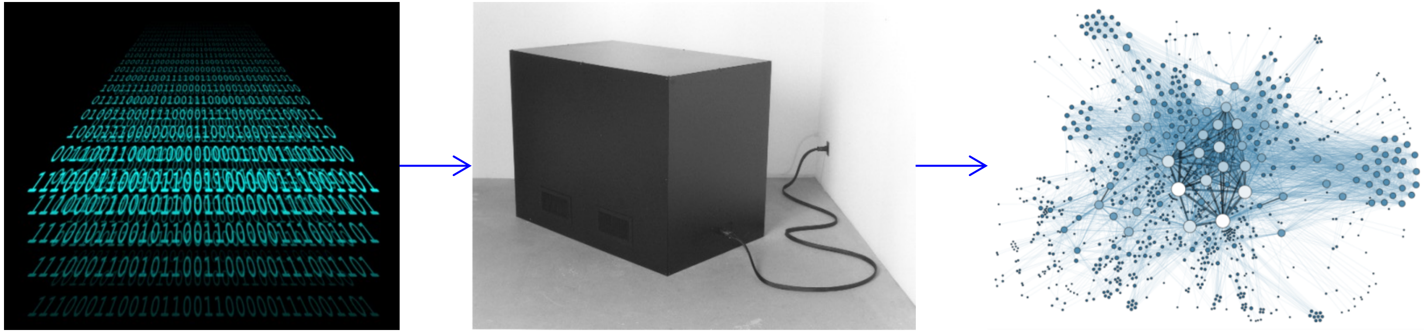
- What do you do with all this data?
  - Too much data to search through it manually.

- 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

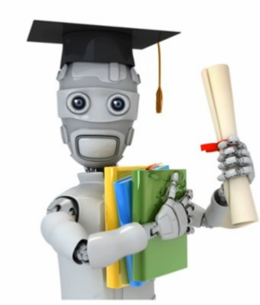
- Automatically extract useful knowledge from large datasets.

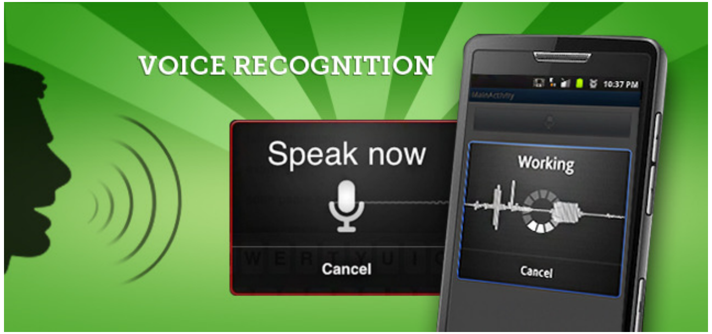


- Usually, to help with human decision making.

Machine Learning

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

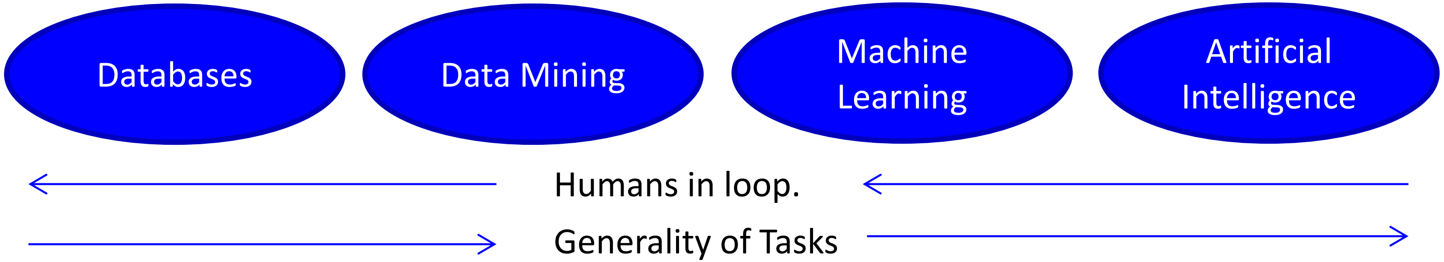




- Most useful when:
  - We want to automate something a human can do.
  - We want to do things a human can’t do (look at 1 TB of data).

Data Mining vs. Machine Learning

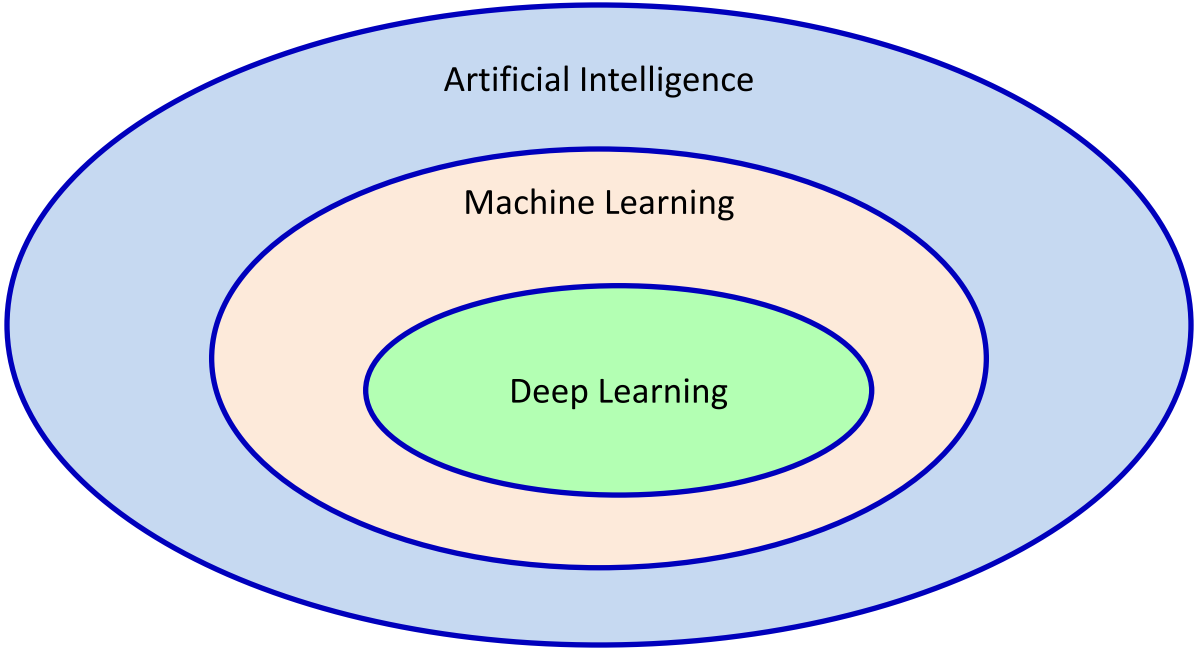
- Data mining and machine learning are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.



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

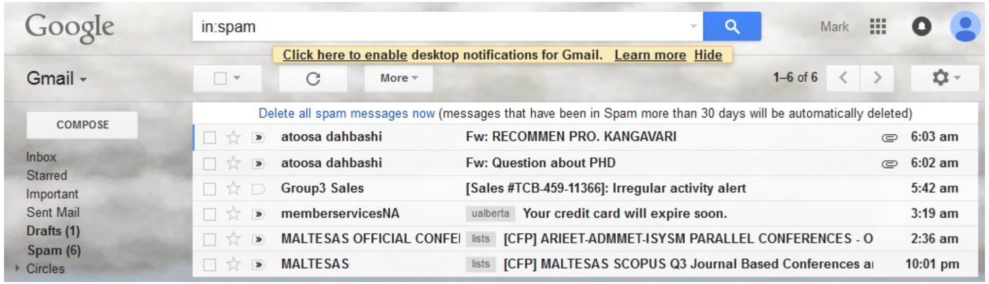
Deep Learning vs. Machine Learning vs. AI

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



Applications

- Spam filtering:

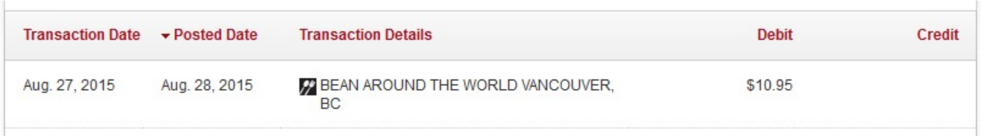


Click here to enable desktop notifications for Gmail. Learn more Hide

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

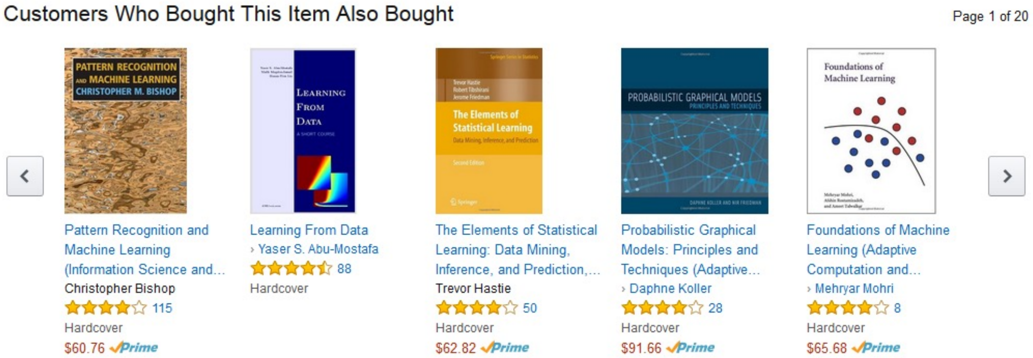
atoosa dahbashi		Fw: RECOMMEN PRO. KANGAVARI	6:03 am
atoosa dahbashi		Fw: Question about PHD	6:02 am
Group3 Sales		[Sales #TCB-459-11366]: Irregular activity alert	5:42 am
memberservicesNA	ualberta	Your credit card will expire soon.	3:19 am
MALTESAS OFFICIAL CONFEI	lists	[CFP] ARIEET-ADMMET-ISYSM PARALLEL CONFERENCES - O	2:36 am
MALTESAS	lists	[CFP] MALTESAS SCOPUS Q3 Journal Based Conferences ai	10:01 pm

- 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:



Customers Who Bought This Item Also Bought

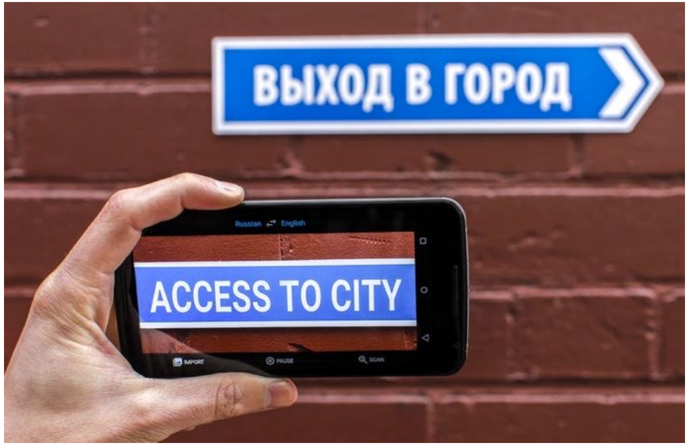
Page 1 of 20

Pattern Recognition and Machine Learning (Information Science and...	Learning From Data	The Elements of Statistical Learning: Data Mining, Inference, and Prediction,...	Probabilistic Graphical Models: Principles and Techniques (Adaptive...	Foundations of Machine Learning (Adaptive Computation and...
Christopher Bishop	› Yaser S. Abu-Mostafa	Trevor Hastie	› Daphne Koller	› Mehryar Mohri
115	88	50	28	8
Hardcover	Hardcover	Hardcover	Hardcover	Hardcover
$60.76 Prime		$62.82 Prime	$91.66 Prime	$65.68 Prime

Applications

- Motion capture:

- Optical character recognition and machine translation:

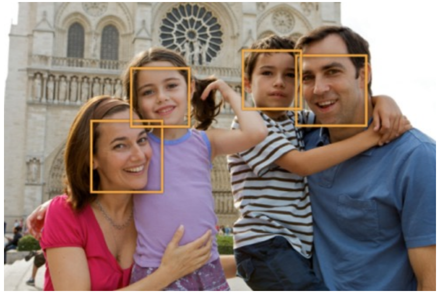


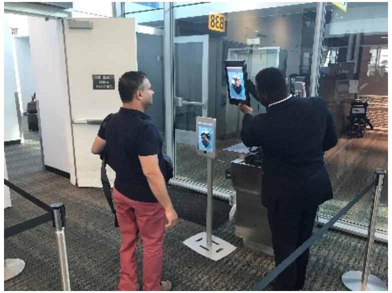
- Speech recognition:

What can I help you with?

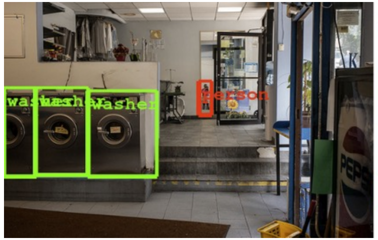
Applications

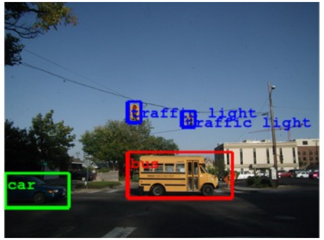
- Face detection/recognition:

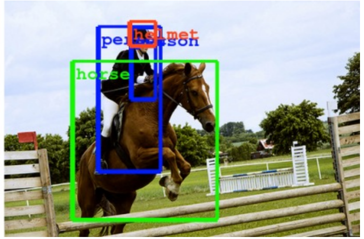




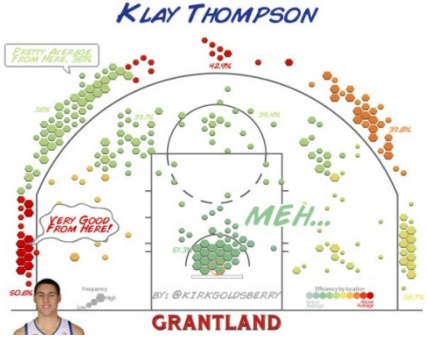
- Object detection:





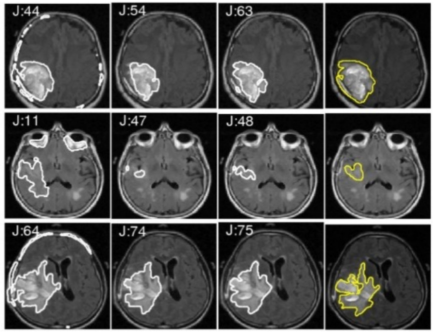


- Sports analytics:



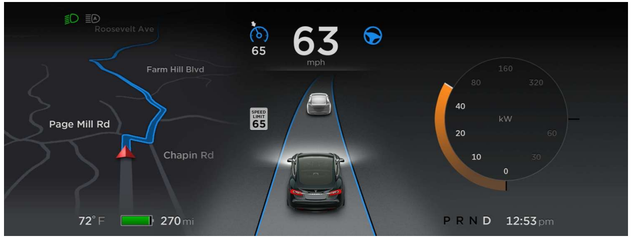
Applications

- Medical imaging:



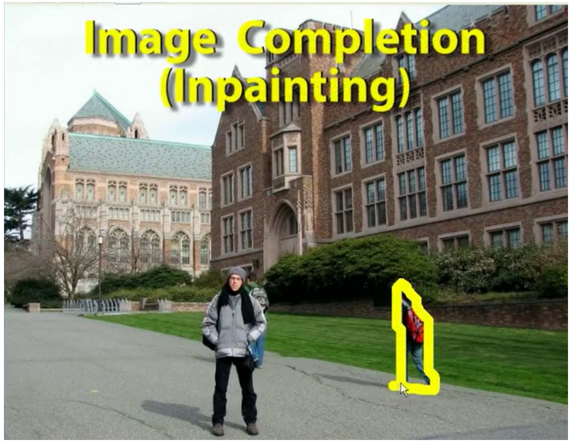
- Medical diagnostics:

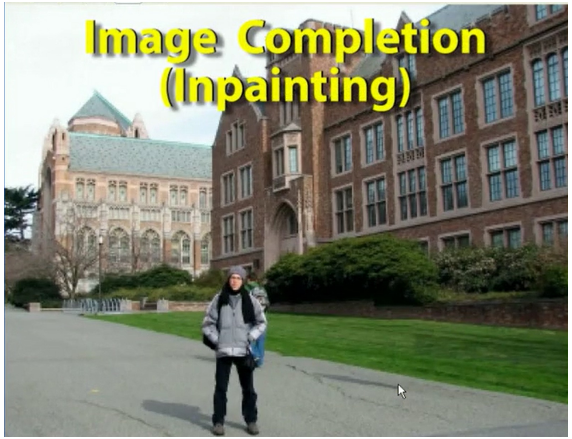
- Self-driving cars:



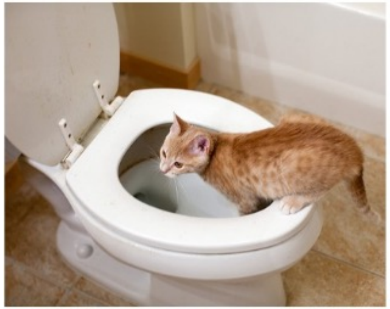
Applications

- Image completion:

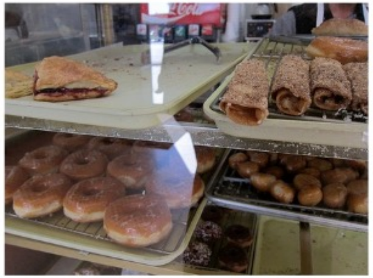




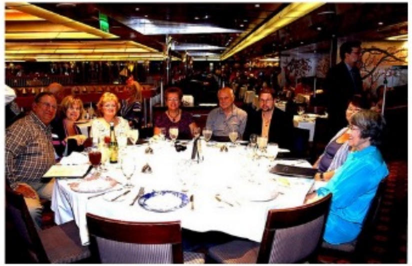
- Image annotation:



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



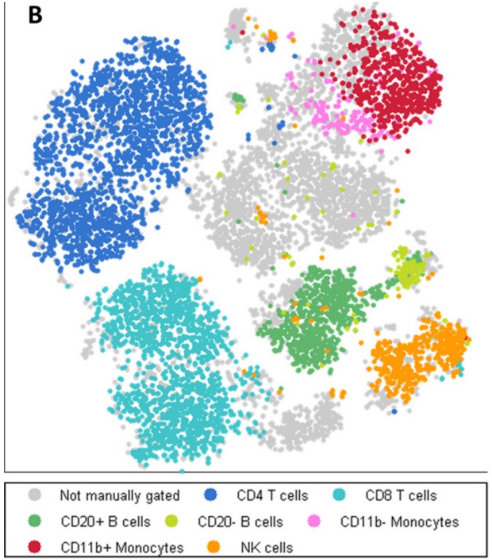
a display case filled with lots of different types of donuts
logprob: -7.78



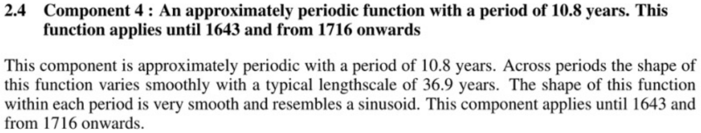
a group of people sitting at a table with wine glasses
logprob: -6.71

Applications

- Discovering new cancer subtypes:

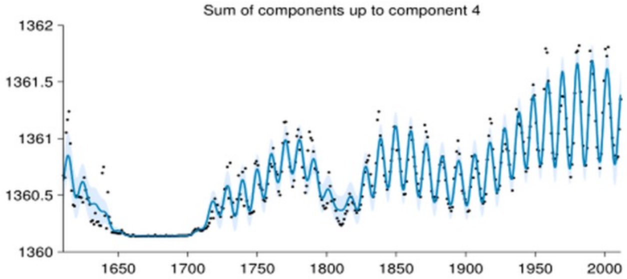


- Automated Statistician:



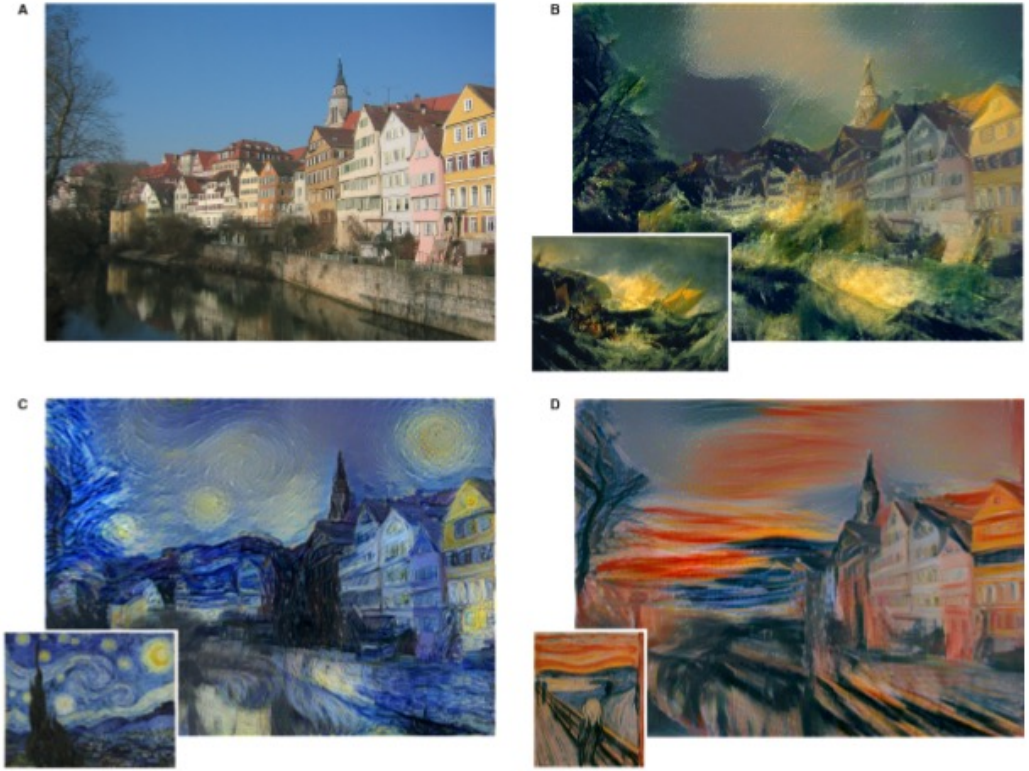
2.4 **Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards**

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.



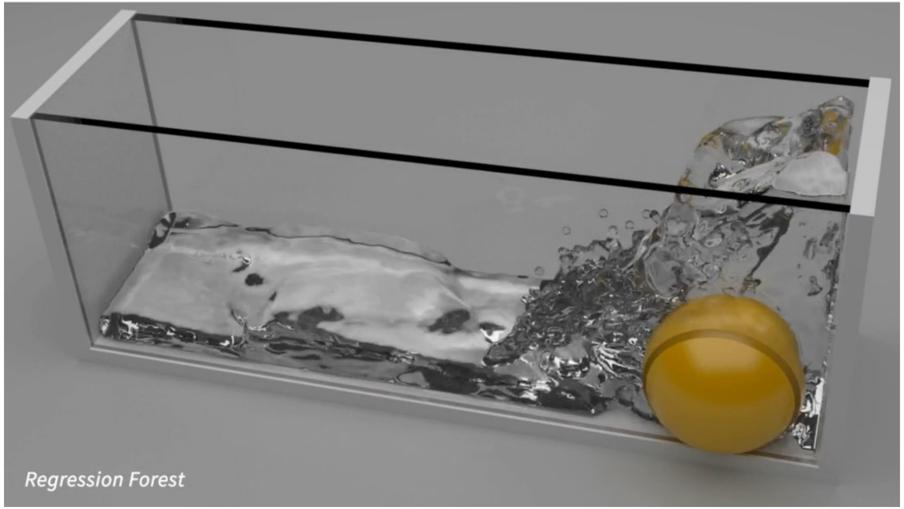
Applications

- Mimicking artistic styles:



Applications

- Fast physics-based animation:

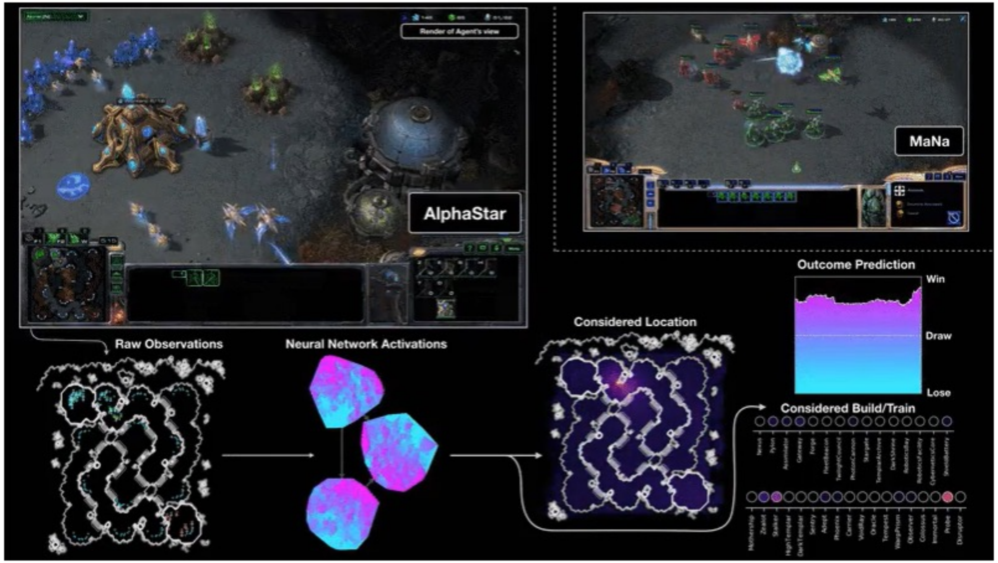


- Character animation:

- Mimicking art style in video.
- Recent work on generating text/music/voice/poetry/dance.

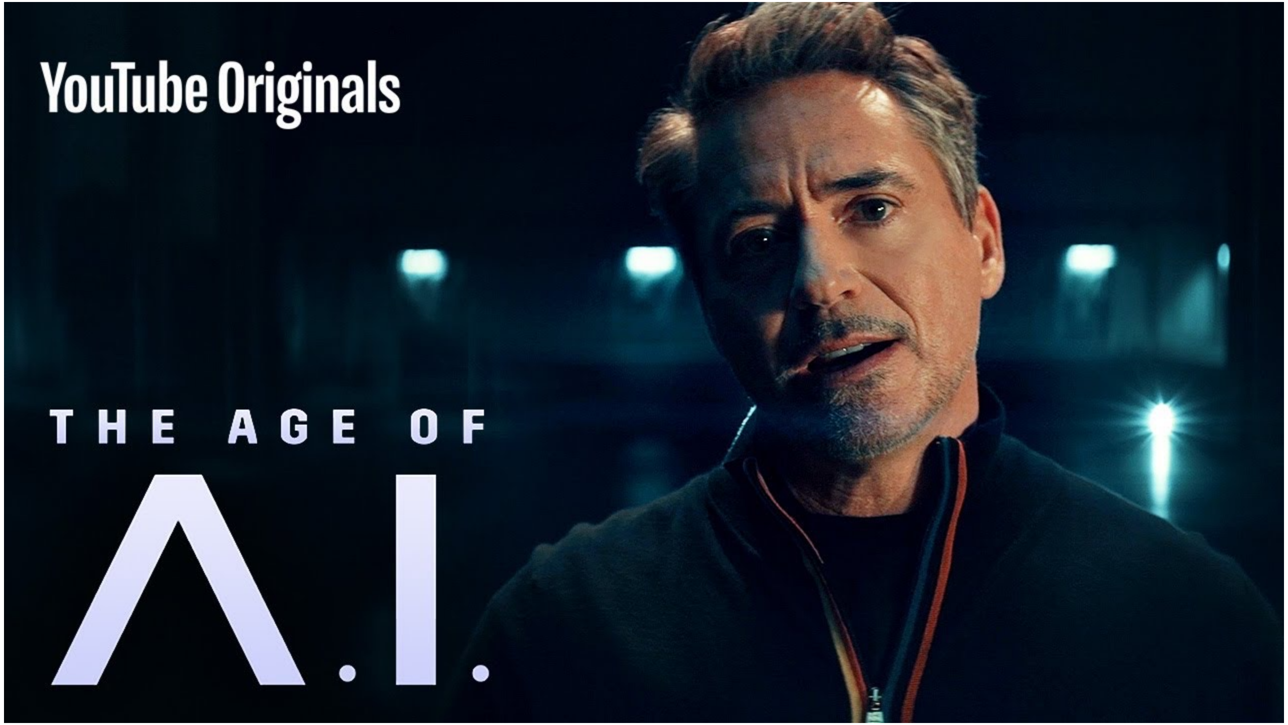
Applications

- Beating humans in Go and Starcraft:



Applications

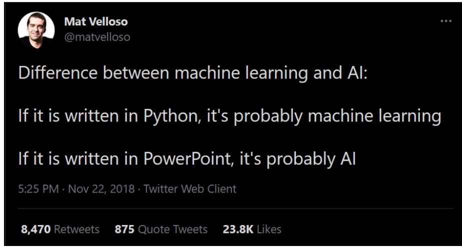
- "Age of AI" YouTube series:



- 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 fields like speech recognition and computer vision.
  - Things are changing a lot on the timescale of 3-5 years.
  - NeurIPS conference sold out in ~11 minutes in 2018.
  - 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.
  - “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.” – John Tukey
  - A huge number of people applying ML are just “overfitting”.
    - Or don’t understand the assumptions needed for them to work.
    - Their methods do not work when they are released “into the wild”.



Mat Velloso
@matvelloso

Difference between machine learning and AI:

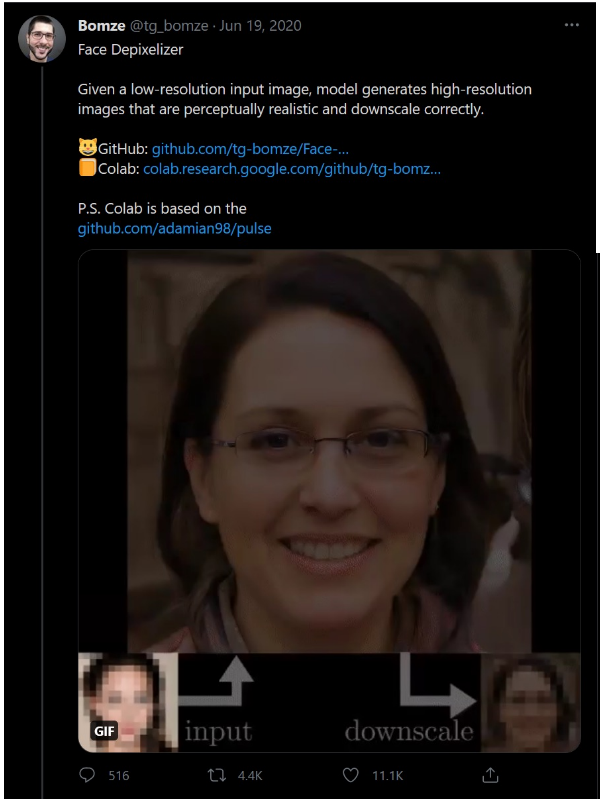
If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

5:25 PM · Nov 22, 2018 · Twitter Web Client

8,470 Retweets 875 Quote Tweets 23.8K Likes

Failures of Machine Learning



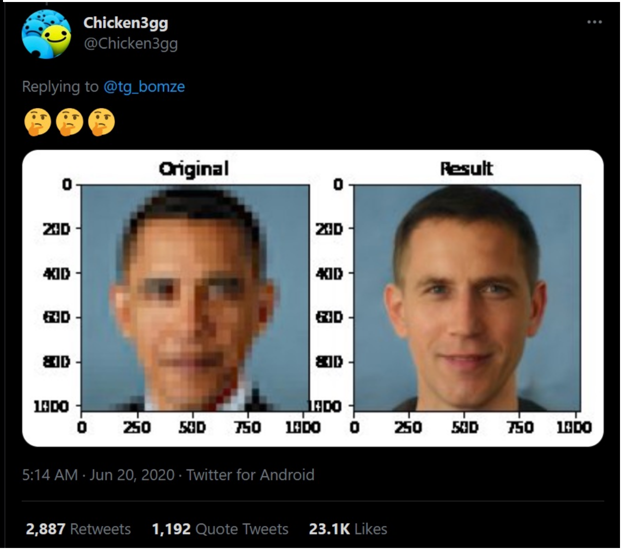
Bomze @tg_bomze · Jun 19, 2020

Face Depixelizer

Given a low-resolution input image, model generates high-resolution images that are perceptually realistic and downscale correctly.

😺GitHub: github.com/tg-bomze/Face-...
📙Colab: colab.research.google.com/github/tg-bomz...

P.S. Colab is based on the
github.com/adamian98/pulse



Chicken3gg
@Chicken3gg

Replying to @tg_bomze

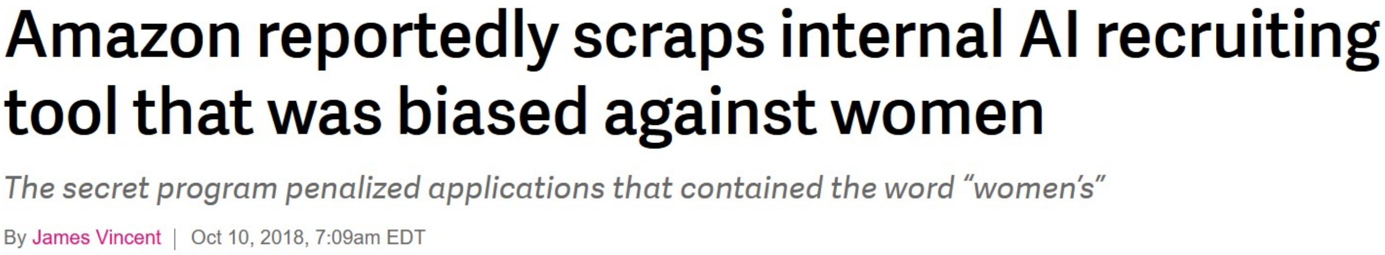
🤔🤔🤔

5:14 AM · Jun 20, 2020 · Twitter for Android

2,887 Retweets **1,192** Quote Tweets **23.1K** Likes

Racial bias

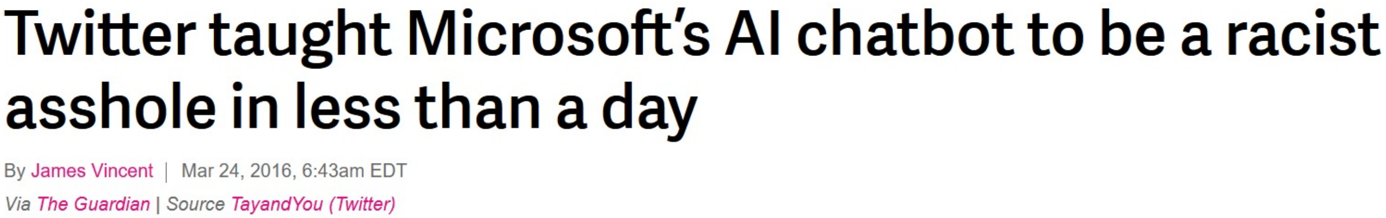
Failures of Machine Learning



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

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

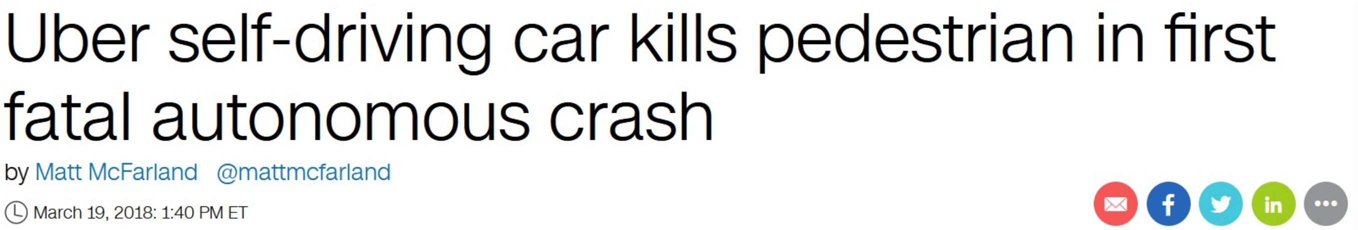
By James Vincent | Oct 10, 2018, 7:09am EDT



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

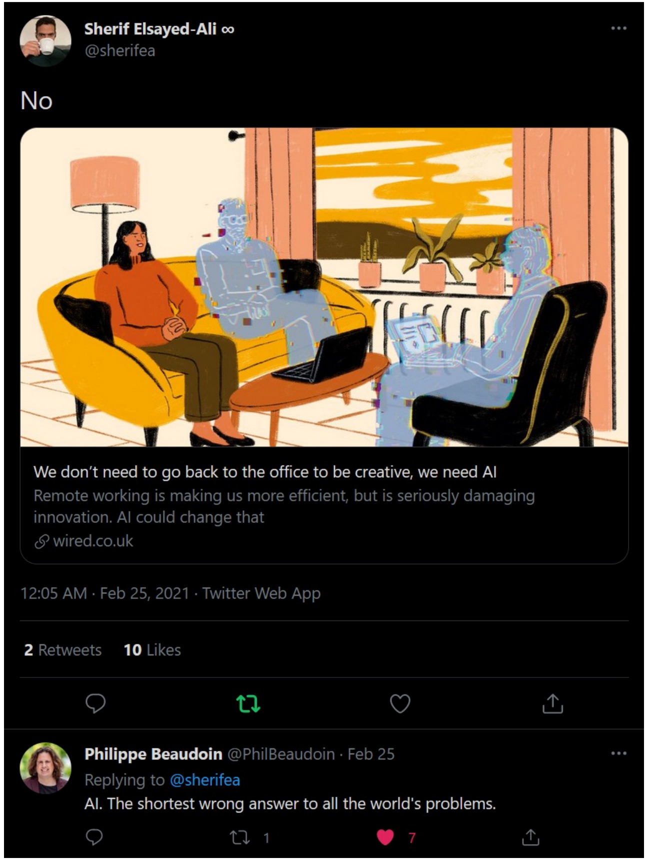
Via The Guardian | Source TayandYou (Twitter)



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

by Matt McFarland @mattmcfarland

March 19, 2018: 1:40 PM ET



Sherif Elsayed-Ali ∞
@sherifea

No

We don't need to go back to the office to be creative, we need AI
Remote working is making us more efficient, but is seriously damaging innovation. AI could change that
wired.co.uk

12:05 AM · Feb 25, 2021 · Twitter Web App

2 Retweets **10** Likes

Philippe Beaudoin @PhilBeaudoin · Feb 25
Replying to @sherifea
AI. The shortest wrong answer to all the world's problems.

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