

Statistics and Data Science 265 / 565

# **Introductory Machine Learning**

Thursday, August 29

**Yale**

# Outline

- Overview of course
- Perspectives on ML (and AI)
- Syllabus and logistics
- Questions

# **Course objectives**

Gain understanding of and experience with basic machine learning methodology

# Course objectives

- Gain some new perspective
- Appreciate some of the power and limitations of ML
- Reflect on societal implications of AI/ML
- Have fun
- Want to learn more

## Related course

- This course introduced for Certificate in Data Science
- Intended to be accessible intro to ML for wide range of students
- S&DS 365/665 “Intermediate Machine Learning” can be taken as a follow up course; more technical and in-depth
- Happy to chat if unclear this course is right for you

# Common questions

“What’s the difference between AI and Machine Learning?”

“Is Deep Learning the same as Machine Learning?”

“What’s the difference between Statistics and Machine Learning?”

August 31, 1955

John McCarthy, Marvin L. Minsky, Nathaniel Rochester,  
and Claude E. Shannon

A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

*June 17 - Aug. 16*

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

SUNDAY, OCTOBER 19, 1958

## MACHINE TO COPY BRAIN'S METHODS

Huge Computer in London  
to 'Think' Like a Person  
for Study of Learning

Special to The New York Times.

LONDON, Oct. 15—Investigators in neurology at University College here are building a massive automatic computer for the principal purpose of testing theories about the learning capacity of the brain.

The machine will "think"; that is, it will scan shapes such as the letters of the alphabet and simple words and after analyzing and absorbing this visual information it will "say" (through a loudspeaker) what it has seen at precisely the same rate as that of a fairly intelligent human subject.

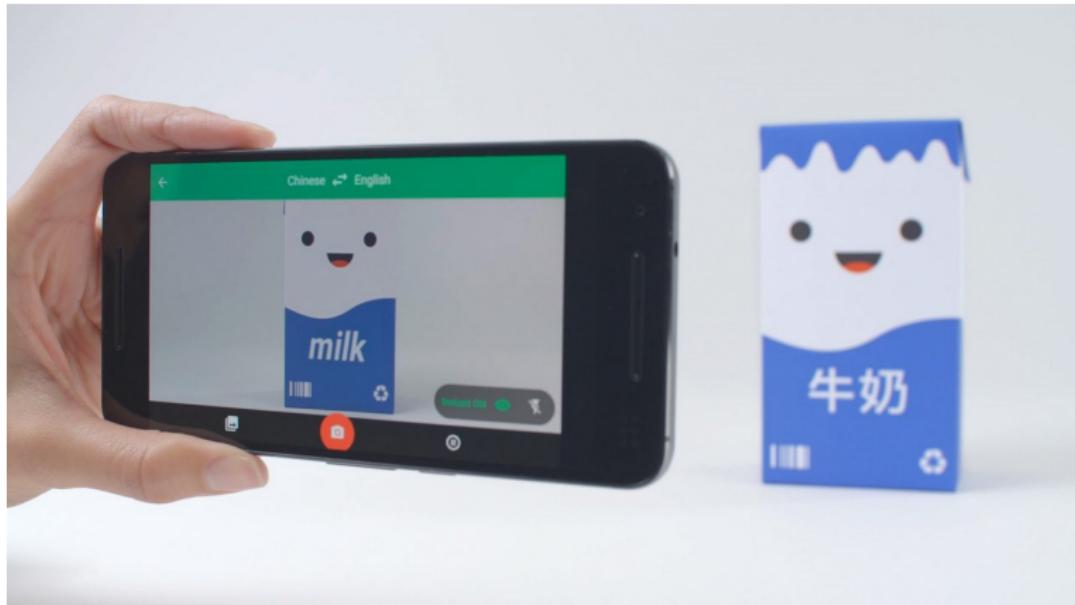
This is being achieved by



# Today: Home assistants



# Translation



<http://www.sciencemag.org/>

# Pricing and recommending homes

## THE WALL STREET JOURNAL.

Subscribe Now | Sign In

\$1 for 2 months

Home World U.S. Politics Economy

**Business**

Tech Markets

Opinion Arts Life

Real Estate



BlackRock,  
Vanguard Mull  
Pressuring Exxon to  
Disclose ...



Ford's New Chief  
Shakes Up  
Management Team



Each Cigna  
Employee to Get Five  
Shares



CIO JOURNAL



## Zillow Develops Neural Network to ‘See’ Like a House Hunter

Granite or stainless steel countertops? Zillow’s visual recognition effort can recognize the difference

By **SARA CASTELLANOS**

Nov 11, 2016 3:29 pm ET

Data scientists at Zillow Group are developing complex computer programs that detect specific attributes in photographs of homes, which could aid in estimating their value. Advances in deep learning, big data and cloud computing have converged to allow the online real estate database firm and others to develop technology that mimics how the human brain [ ]

---

### Recommended Videos

1. Film Clip: Pirates of the Caribbean: Dead Men Tell No Tales'



2. What to do in your 40s to retire a millionaire



<https://blogs.wsj.com/cio/2016/11/11/zillow-develops-neural-network-to-see-like-a-home-buyer/>

# YouTube



---

Amazing ways YouTube uses ML and AI: <https://www.forbes.com/sites/bernardmarr/2019/08/23/the-amazing-ways-youtube-uses-artificial-intelligence-and-machine-learning>

# YouTube

- Each month: 1.9 billion users
- Each day: 1 billion hours of video watched
- Each minute: 300 hours of video uploaded
- ML: Automatically remove objectionable content
- ML: “Up Next” feature

# Translation

HOME > SPORTS

## A Belarusian Olympian who complained about her coaches used Google Translate to relay her plea for help to Japanese police

Lauren Fries Aug 5, 2021, 6:46 PM



Belarusian Olympic sprinter Krystsina Tsimanouskaya said she was taken to the airport against her wishes and would not return home. Reuters

With a zero trust strategy,  you're in the driver's seat

[See how →](#)



IBM

# More translation



## *The Animal Translators*

Scientists are using machine learning to eavesdrop on naked mole rats, fruit bats, crows and whales — and to communicate back.

# More translation

## *The Animal Translators*

Scientists are using machine learning to eavesdrop on naked mole rats, fruit bats, crows and whales — and to communicate back.

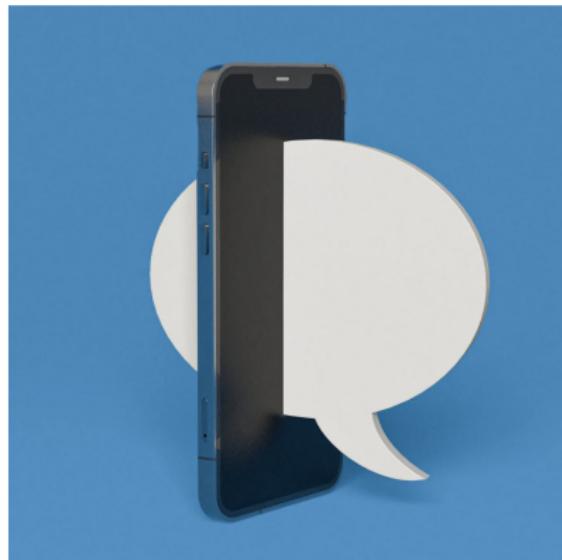
# More translation

The New York Times

OPINION

## Will Translation Apps Make Learning Foreign Languages Obsolete?

July 25, 2023





Prompt: "An AI taking over the world. Polygonal style. Retro Futurist."

# What is Machine Learning?

The study of algorithms and statistical models to develop computer programs that improve with experience.

# What is Machine Learning?

Machine Learning is closely aligned with Statistics, but with a focus on computation, scalability, prediction, representation, and complex problems

- Speech recognition
- Machine translation
- Object recognition and scene classification
- Autonomous driving...

Subproblems of these and other complex problems are concrete, statistical estimation and inference problems that can be studied in isolation.

# AI vs. ML

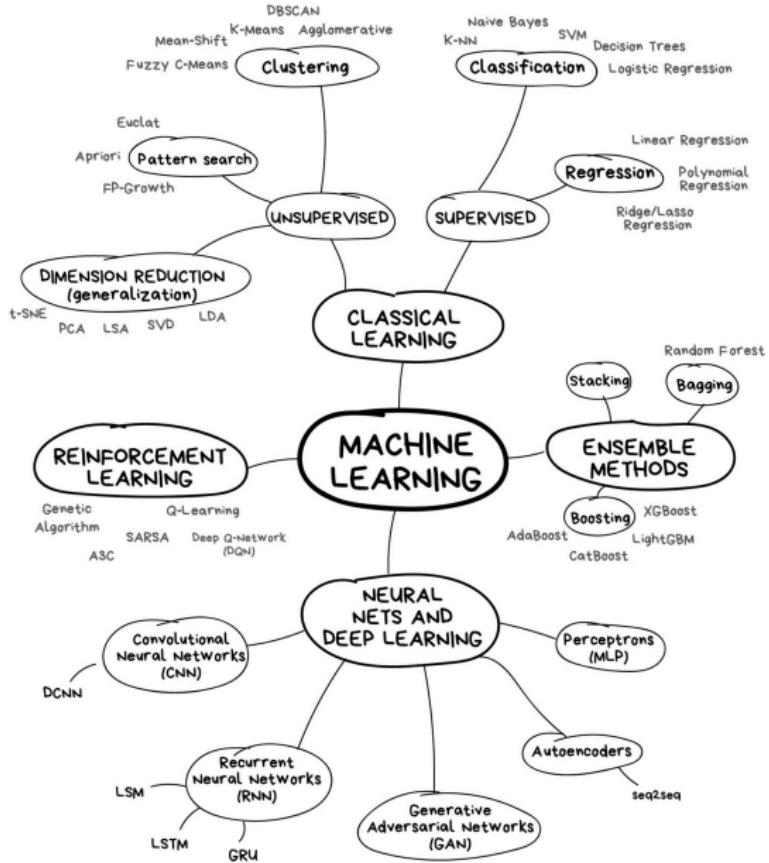
Machine learning focuses on making predictions and inferences from data.

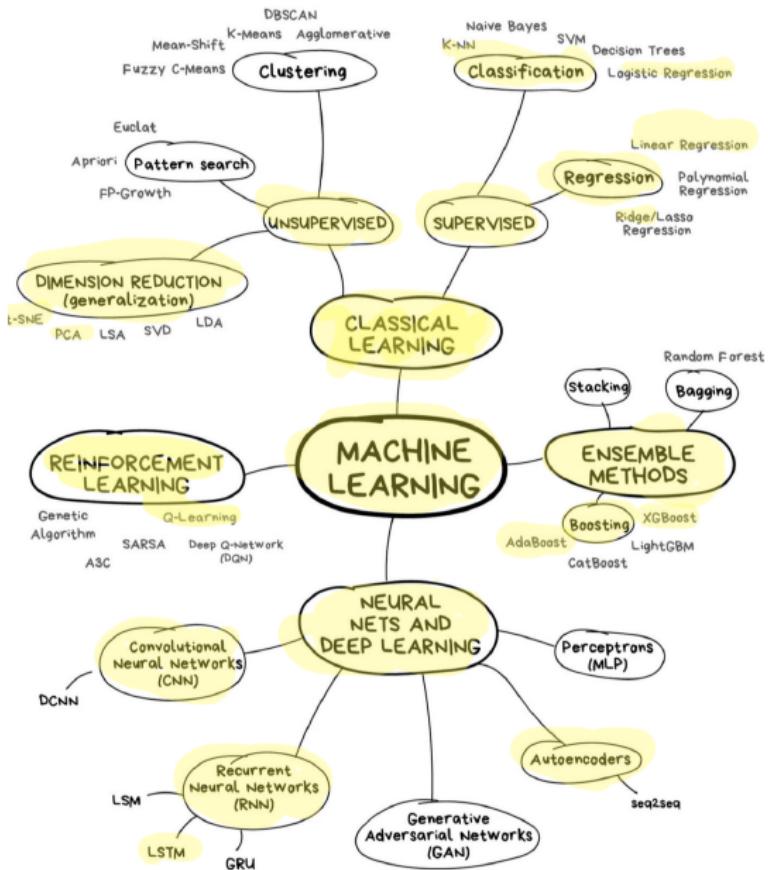
AI combines machine learning components into a larger system that includes a decision making component.

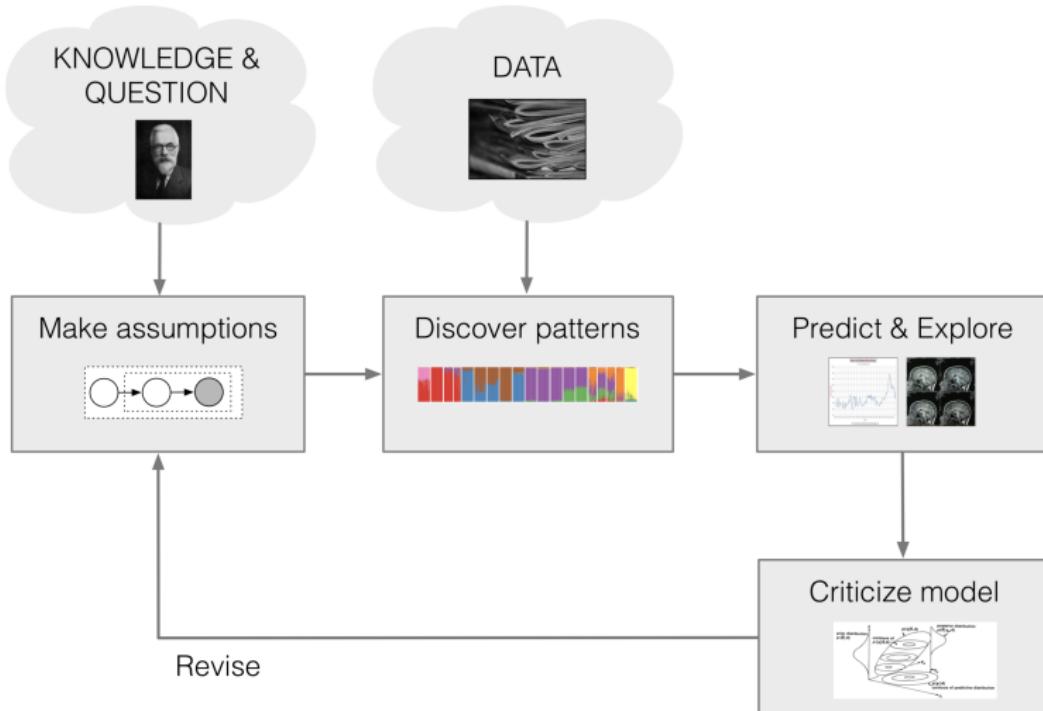
*An AI system exhibits a behavior, resulting from the collective decisions that are made.*

# Machine learning frameworks

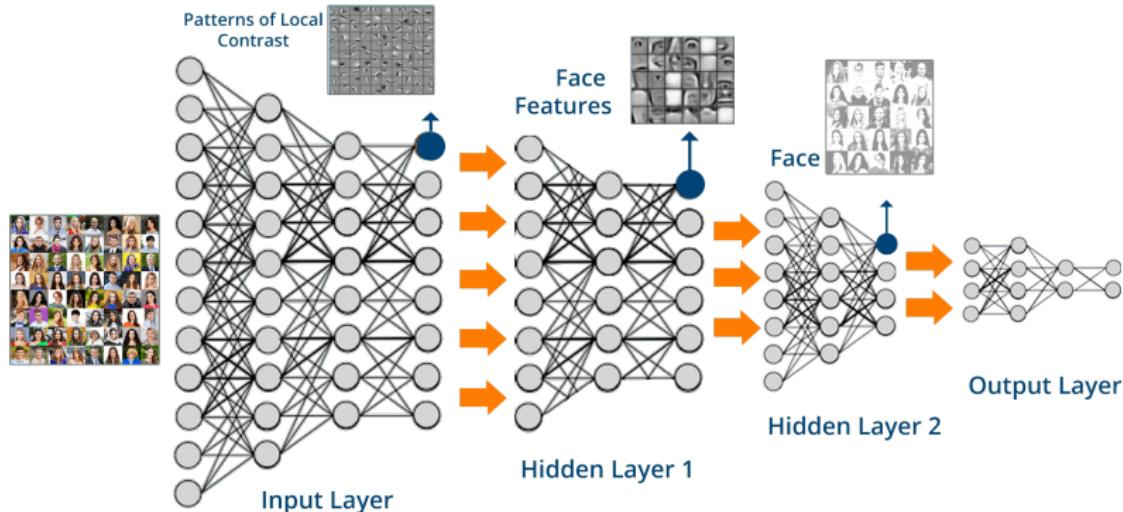
- Supervised, unsupervised, semi-supervised
- Reinforcement learning
- Generative vs. discriminative models
- Representation learning







# Deep learning is a type of machine learning



- Heuristics motivated from simplified view of the brain
- A particular form of nonlinear classification/regression
- Not well-suited to latent variables

# Culture of Code

- Great deal of current AI/ML work is purely engineering based
- Informal input/output reasoning

*“that program gave this output...  
maybe this program will give that output”*

- Deep learning software engineers develop sophisticated intuitions
- The code is the product

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.



[xkcd.com/1838](http://xkcd.com/1838)

# Latent variables: The elephants in the room



# Two types of intelligence

- ① “Neocortical”— acquire semantic and procedural knowledge
  - ▶ Requires extensive data and training
  - ▶ Slow to learn, fast to apply
  - ▶ Well captured by modern deep learning

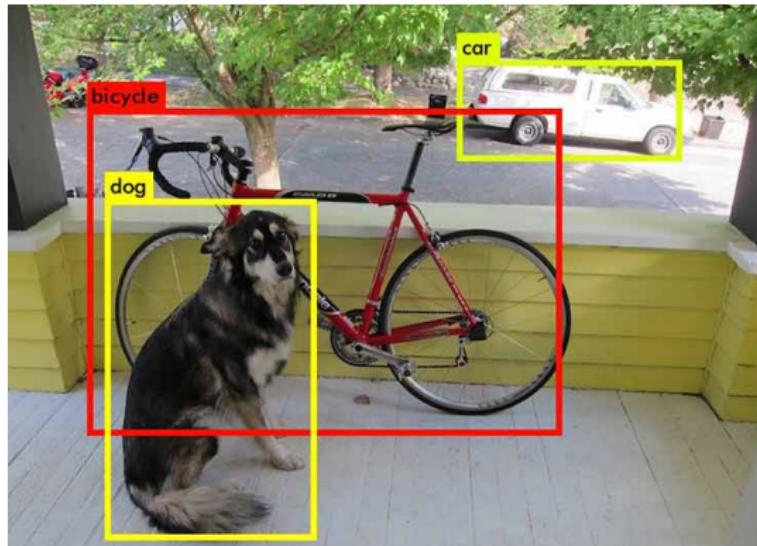
# Two types of intelligence

- ① “Neocortical”— acquire semantic and procedural knowledge



# Two types of intelligence

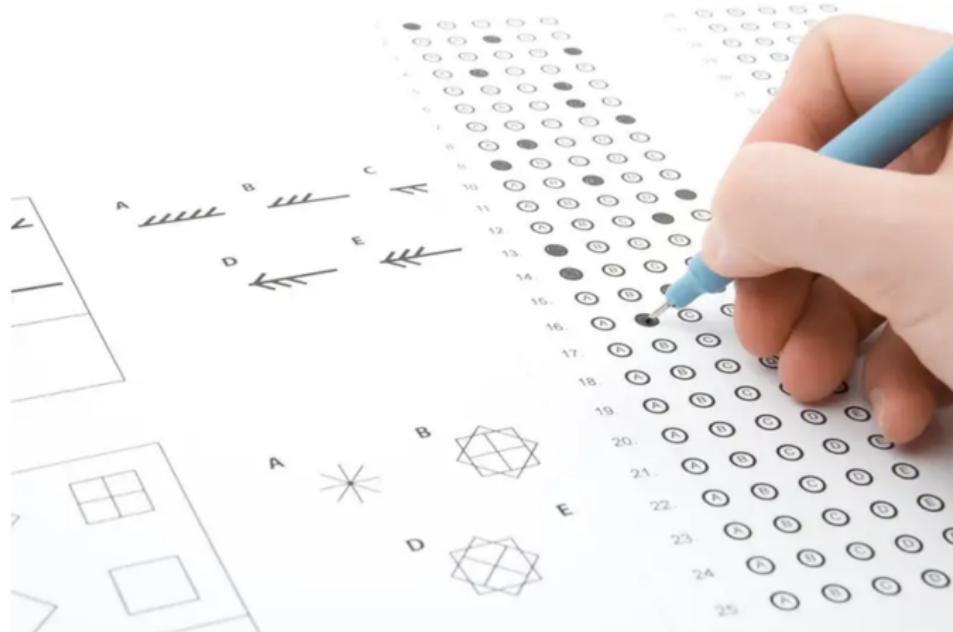
- ① “Neocortical”— acquire semantic and procedural knowledge



# Two types of intelligence

- ② “Prefrontal”— identify novel associations and relations
  - ▶ Fast to learn, slow to apply
  - ▶ Symbolic processing and abstraction
  - ▶ Little explicit training data

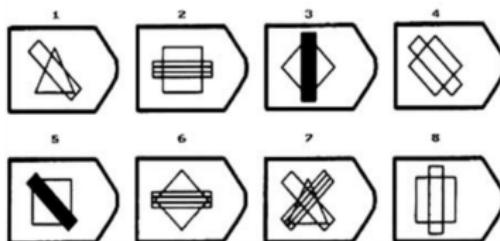
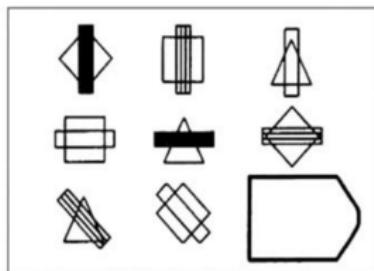
# Two types of intelligence



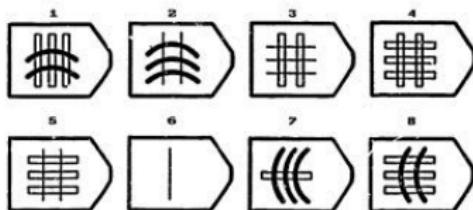
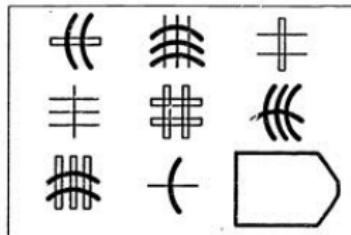
AI struggles with IQ tests

Panther Media GmbH / Alamy Stock Photo

# Two types of intelligence



# Two types of intelligence



# Two types of intelligence

- ① “Neocortical”— efficient function approximation
- ② “Prefrontal”— rapid abstraction, association and generalization

*Can both types be supported in a single architecture?*

# Relations are essential to reasoning

*“Jack and Jill are sitting side by side. The person next to Jack is angry. The person next to Jill is happy. Who is happy, Jack or Jill?”*



# Relations are essential to reasoning

ChatGPT response:

- Based on the information provided, Jill is happy. The statement specifies that the person next to Jill is happy, indicating that Jill is the one who is happy. Therefore, Jill is happy while Jack is angry.

# Relations are essential to reasoning

ChatGPT response:

- Based on the information provided, Jill is happy. The statement specifies that the person next to Jill is happy, indicating that Jill is the one who is happy. Therefore, Jill is happy while Jack is angry.

# Relations are essential to reasoning

ChatGPT response:

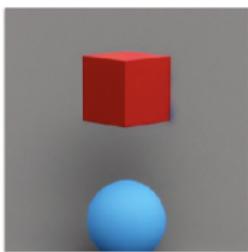
- Based on the information provided, Jill is happy. The statement specifies that the person next to Jill is happy, indicating that Jill is the one who is happy. Therefore, Jill is happy while Jack is angry.

# Relations are essential to reasoning

*“Two shapes are in a room. The shape next to the cube is blue. The shape next to the sphere is red. Which is red, the cube or the sphere?”*

# Relations are essential to reasoning

*“Two shapes are in a room. The shape next to the cube is blue. The shape next to the sphere is red. Which is red, the cube or the sphere?”*



# Relations are essential to reasoning

- Based on the given information, it can be inferred that the sphere is red. The statement specifies that the shape next to the sphere is red, implying that the sphere itself is red. The color of the cube is not mentioned, so we cannot determine its color from the given information.

# Relations are essential to reasoning

- Based on the given information, it can be inferred that the sphere is red. The statement specifies that the shape next to the sphere is red, implying that the sphere itself is red. The color of the cube is not mentioned, so we cannot determine its color from the given information.

# Relations are essential to reasoning

- Based on the given information, it can be inferred that the sphere is red. The statement specifies that the shape next to the sphere is red, implying that the sphere itself is red. The color of the cube is not mentioned, so we cannot determine its color from the given information.

# ARC-AGI Prize

**ARC PRIZE**

AGI PROGRESS HAS STALLED.  
NEW IDEAS ARE NEEDED.

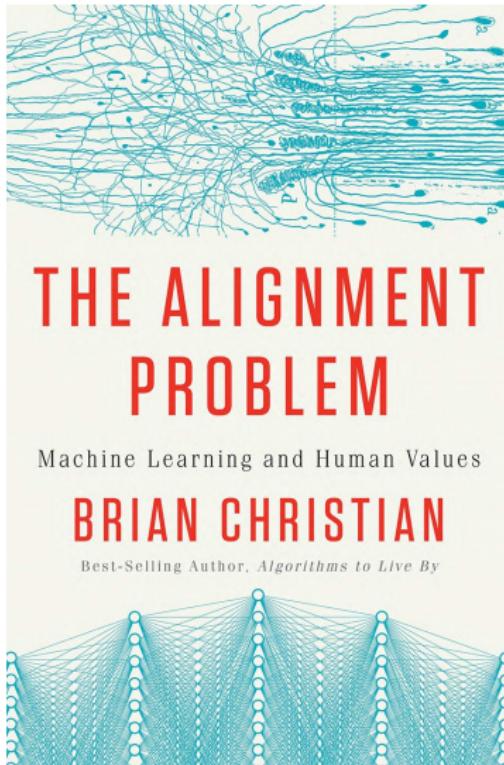
Presented by  WU LAB

- > Home
- > ARC-AGI
- > Leaderboard
- > ARC Prize 2024
- > Technical Guide
- > ARC-AGI-Pub
- > Play
- > Blog

**SIGN UP**

# Shortcomings are masked

- Recent innovation with ChatGPT hides these deficiencies
- System is trained to convince us
- Over-confidence and bogus deductive reasoning
- After this course you'll have an understanding of main components of ChatGPT



# **Example of representation learning: Word embeddings**

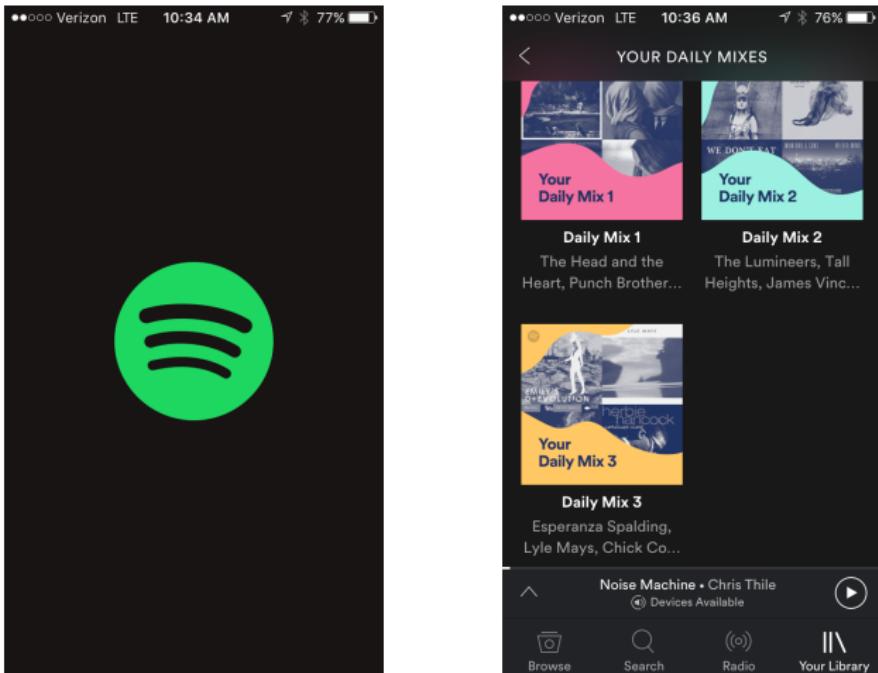
- Each word in vocab is mapped to 100 or 500 dimensional vector
- Based solely on co-occurrence statistics in corpus of text

# Example of representation learning: Word embeddings

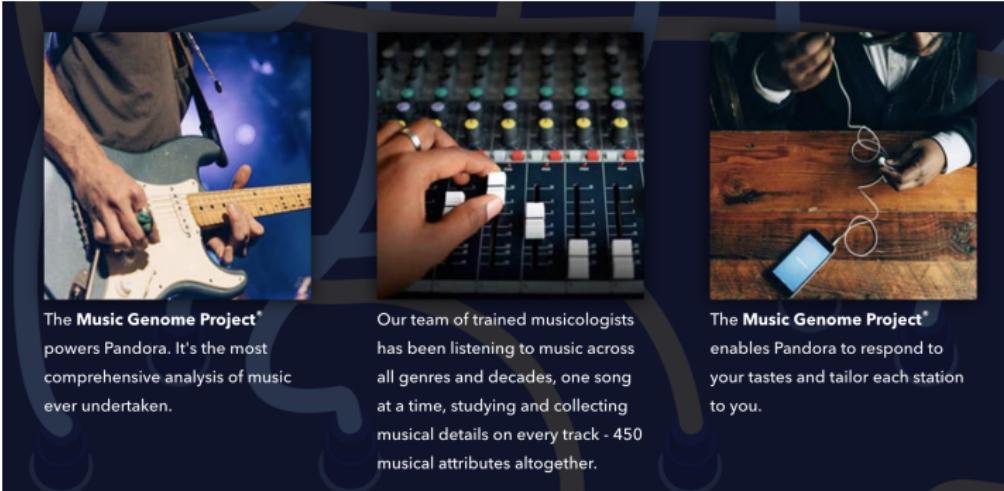
Yale:

```
[ 0.78310001, 0.51717001, -0.38207 , -0.23722 , -0.31615999, 0.30805001, 0.76389998, 0.064106 , -0.74913001,  
 0.60585999, -0.23871 , -0.16876 , -0.25634 , 1.07270002, -0.29967999, 0.020095 , 0.54500997, -0.17847 , -0.26675999,  
 -0.11798 , -0.48692 , 0.22712 , 0.017473 , -0.4747 , 0.44861001, -0.084281 , -0.30412999, -1.13510001, -0.14869 , -0.11182 ,  
 -0.32530001, 1.0029 , -0.35742 , 0.35148999, -1.10679996, -0.064142 , -0.72284001, 0.14114 , -0.41247001, -0.16184001,  
 -0.54576999, -0.12958001, -0.88356 , -0.089722 , 0.10555 , -0.12288 , 0.92851001, 0.50032002, 0.1349 , 0.21457 ,  
 0.35073999, -0.73132998, 0.39633 , -0.43239999, -0.38815999, -1.34669995, 0.37463999, -0.79386002, 0.11185 , 0.18007 ,  
 -0.75142998, 0.24975 , -0.094948 , -0.36341 , 0.24869999, -0.22667 , 0.32289001, 1.29489994, 0.42658001, 1.29120004,  
 -0.13954 , 0.68976003, 0.21586999, 0.13715 , -1.00919998, 0.028827 , 0.11011 , -0.1912 , -0.073198 , -0.52449 , 0.49199 ,  
 0.14463 , -0.18844 , -0.75536001, -0.28704 , 0.019113 , 0.30349001, -0.74425 , -0.072221 , -0.40647 , 0.26899001, -0.28318  
, 0.72409999, 0.50796002, -0.37845999, -0.13008 , -0.13808 , 0.098928 , 0.16215999, 0.16293 ]
```

# Embeddings for music recommendations



# Experts vs. Data: The case of Pandora vs. Spotify



The Music Genome Project<sup>\*</sup> powers Pandora. It's the most comprehensive analysis of music ever undertaken.

Our team of trained musicologists has been listening to music across all genres and decades, one song at a time, studying and collecting musical details on every track - 450 musical attributes altogether.

The Music Genome Project<sup>\*</sup> enables Pandora to respond to your tastes and tailor each station to you.

- Pandora's "Music genome": Over 450 musical attributes
- Melody, harmony, rhythm, form, composition, lyrics...

<https://arstechnica.com/tech-policy/2011/01/digging-into-pandoras-music-genome-with-musicologist-nolan-gasser/>

# Experts vs. Data: The case of Pandora vs. Spotify

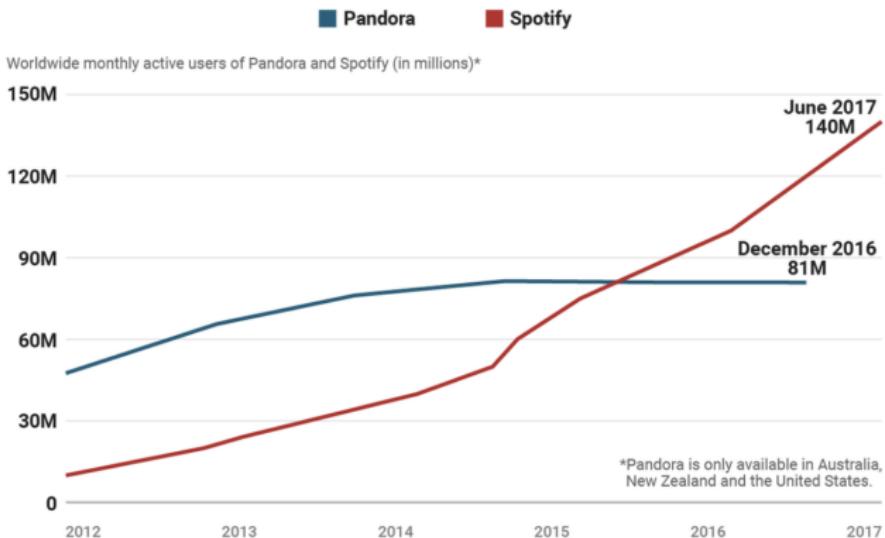


*Spotify: Word embeddings trained from playlists*

# Experts vs. Data: The case of Pandora vs. Spotify

TECH ■ CHART OF THE DAY

## PANDORA'S GROWTH STALLS AS SPOTIFY PULLS AHEAD



SOURCE: Company filings/announcements

BUSINESS INSIDER

# Hacking ML Systems



TECHNICA

SUBSCRIBE



SIGN IN ▾

TESLA AUTOPILOT —

## Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

DAN GOODIN - 4/1/2019, 8:50 PM

Keen Security Lab



# Machine learning at a large Internet company

- Typical project lifetime: 6 months to 1 year
- Ads projects involve thousands of software engineers
- Often adding new “feature” to existing black box model
- No single person understands entire model
- Not interpretable
- Potential for large breaches of personal information

# Reasons for optimism

- Increasingly part of academic research across disciplines
- Engaging a broad community; wide public awareness
- We're still in very early stages

**Questions or discussion?**

# Course materials

Materials posted to <http://introml.ydata123.org>; sometimes to Canvas

Please use Ed Discussion for any content questions about lectures, homework, etc.

**Email all questions about logistics to** sds265@yale.edu

# Syllabus

*Introductory Machine Learning* covers the key ideas and techniques in machine learning without the use of advanced mathematics. Basic methodology and relevant concepts are presented in lectures, including the intuition behind the methods and a more formal understanding of how and why they work. Assignments give students hands-on experience with the methods on different types of data.

# Syllabus

Topics include linear regression and classification, tree-based methods, topic models, word embeddings, recurrent neural networks, deep learning and reinforcement learning. Examples come from a variety of sources including political speeches, archives of scientific articles, real estate listings, natural images, and several others. Programming is central to the course, and is based on the Python programming language.

# Prerequisites

- At least two of the following courses: S&DS 230, 238, 240, 241 and 242
- Previous programming experience (e.g., R, Matlab, Python, C++), Python preferred. The course will make extensive use of Python programming, using Jupyter notebooks.

# Installing Jupyter

- See installation guide on course Canvas site: Files > Getting started
- Use Python 3.x version

# Course goals

Gain understanding of and experience with basic machine learning methodology

# Course goals

- Gain some new perspective
- Appreciate some of the power and limitations of ML
- Reflect on societal implications of AI/ML
- Have fun
- Want to learn more

# Evaluation

- Five assignments (40%)
- Mid-semester exam (25%)
- Five quizzes (10%)
- Final exam: (25%)

Lowest assignment and quiz score will be dropped. Late assignments not accepted.

# Assignments

- Roughly every two weeks
- Due at midnight (11:59pm), typically Thursdays
- Submitted using Gradescope
- Mix of problem solving and data analysis
- Prepared using Python notebooks

# Collaboration

Collaboration on homework assignments with fellow students is encouraged. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had any discussions concerning the problem. You may *not* share written work or code—after discussing a problem with others, the solution must be written by yourself.

# Using Chatbots

Use ChatGPT or other AI tools such as Codex, CoPilot on assignments if you find them useful. Acknowledge your use of them—just as for collaboration in general.

However, exams and quizzes will have coding questions, and such tools are expressly forbidden for these evaluations.

# Exams

- Midterm exam: October 15 in class
- Final exam: December 16 at 2pm
- No rescheduling

Week	Dates	Topics	Demos & Tutorials	Lecture Slides	Readings and Notes	Assignments & Exams
1	Aug 31	Course overview		Thu: Course overview		
2	Sept 5, 7	Python and background concepts	<a href="#">Python elements</a> <a href="#">Covid trends</a>	Tue: Python elements Thu: Pandas and linear regression	<a href="#">Data8 Chapters 3, 4, 5</a>	Quiz 1 <a href="#">Assn 1 out</a>
3	Sept 12, 14	Linear regression and classification	<a href="#">Covid trends (revisited)</a> <a href="#">Classification examples</a>	Tue: Regression concepts Thu: Classification	ISL Sections 3.1, 3.2, 3.5 Notes on <a href="#">regression</a> ISL Sections 4.3, 4.4 Notes on <a href="#">classification</a>	
4	Sept 19, 21	Stochastic gradient descent	<a href="#">SGD examples</a>	Tue: Classification (continued) Thu: Stochastic gradient descent	ISL Section 6.2.2 ISL Section 10.7.2	Assn 1 in <a href="#">Assn 2 out</a>
5	Sept 26, 28	Bias and variance, cross-validation	<a href="#">Bias-variance tradeoff</a> <a href="#">Covid trends (revisited)</a> <a href="#">California housing</a>	Tue: Bias and variance Thu: Cross-validation	ISL Section 2.2 ISL Section 5.1	Quiz 2

6	Oct 3, 5	Tree-based methods	<ul style="list-style-type: none"> <li>○ Trees and forests</li> <li>○ Visualizing trees</li> <li>○ Bagging operations</li> </ul>	Tue: Trees Thu: Forests	ISL Sections 8.1, 8.2	Assn 2 in ○ Assn 3 out
7	Oct 10, 12	PCA and dimension reduction	<ul style="list-style-type: none"> <li>○ PCA examples</li> <li>○ PCA revisited</li> <li>○ Used for regression</li> </ul>	Tue: PCA Thu: PCA and review	ISL Section 12.2	Quiz 3
8	Oct 17	Midterm exam (in class)			On Canvas: Practice midterms / Sample solns Midterm / Sample soln	
9	Oct 24, 26	Language models, word embeddings	<ul style="list-style-type: none"> <li>○ GPT-3 demo</li> <li>○ Word embeddings</li> </ul>	Tue: Language models Thu: Word embeddings	OpenAI: Better language models (GPT-2)	Assn 3 in ○ Assn 4 out
10	Oct 31, Nov 2	Bayesian inference, topic models	<ul style="list-style-type: none"> <li>○ Mixtures</li> <li>○ Bayesian inference</li> <li>○ Topic models</li> </ul>	Tue: Bayesian inference Thu: Topic models	Notes on Bayesian inference	Quiz 4
						

X

11	Nov 7, 9	Introduction to neural networks	<a href="#">CO Sanity check</a> <a href="#">CO Minimal neural network</a> <a href="#">CO Regression examples</a>	Tue: Neural networks Thu: Neural networks	ISL Sections 10.1, 10.2	Assn 4 in <a href="#">CO Assn 5 out</a>
12	Nov 14, 16	Reinforcement learning	<a href="#">CO Q-learning</a>	Tue: Reinforcement learning Thu: Deep reinforcement learning		Quiz 5
13	Nov 21, 23	No class, Thanksgiving break				
14	Nov 28, 30	Deep neural networks	<a href="#">Tensorflow playground</a> <a href="#">CO Autoencoder examples</a>	Tue: Deep neural networks Thu: Autoencoders	ISL Section 10.7 <a href="#">Notes on backpropagation</a>	Assn 5 in
15	Dec 5, 7	Societal issues for machine learning		Tue: Societal issues Thu: Course wrap up		Quiz 6
16	Fri, Dec 15, 2pm, Room TBA	Final exam			Registrar: Final exam schedule Practice final	

# Auditing

- Auditors are welcome!
- Full access to Canvas
- Just expected to regularly attend class

# **Questions?**