Statistics and Data Science 265

Introductory Machine Learning

Thursday, September 1

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
- 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 (grad number 665) will become a new course "Intermediate Machine Learning" last spring. 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.

The New York Times

SUNDAY, OCTOBER 19, 1958

MACHINE TO COPY BRAIN'S METHODS

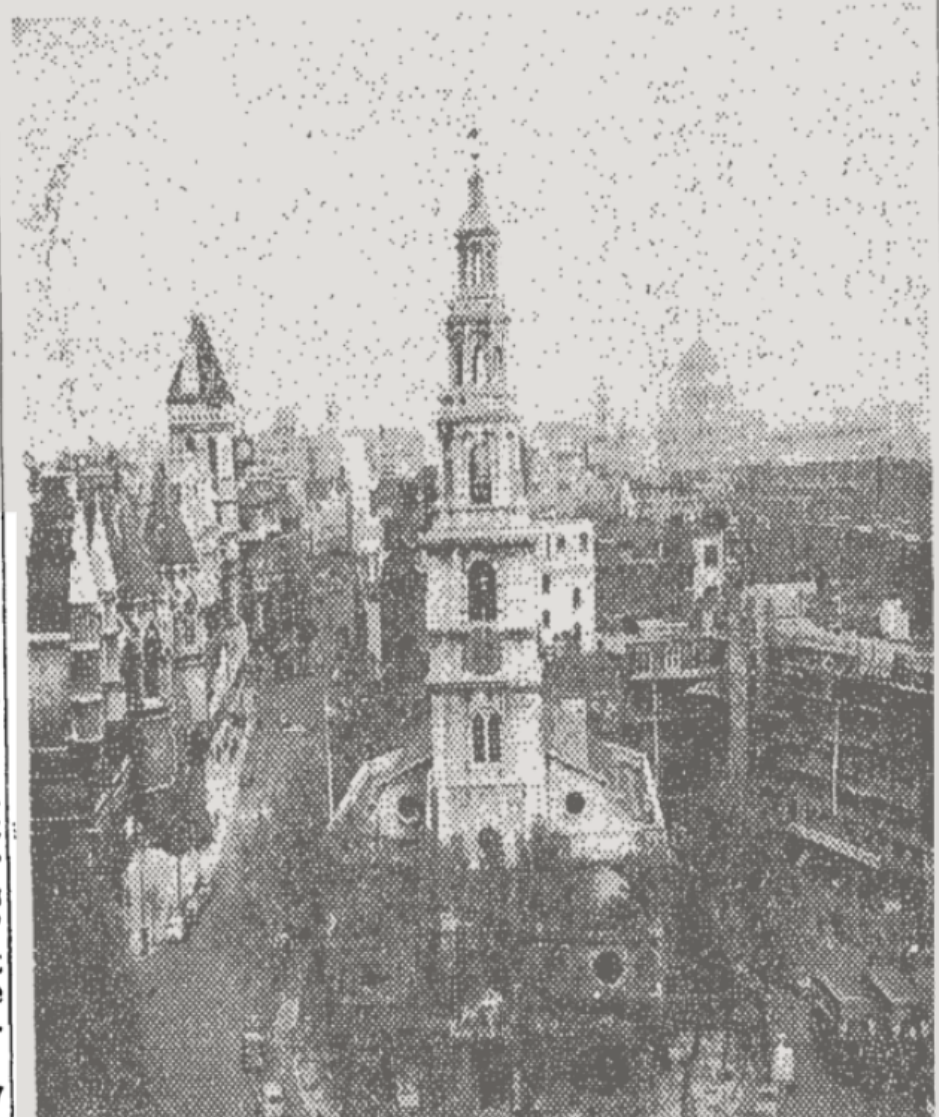
Huge Computer in London to 'Think' Like a Person for Study of Learning

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.

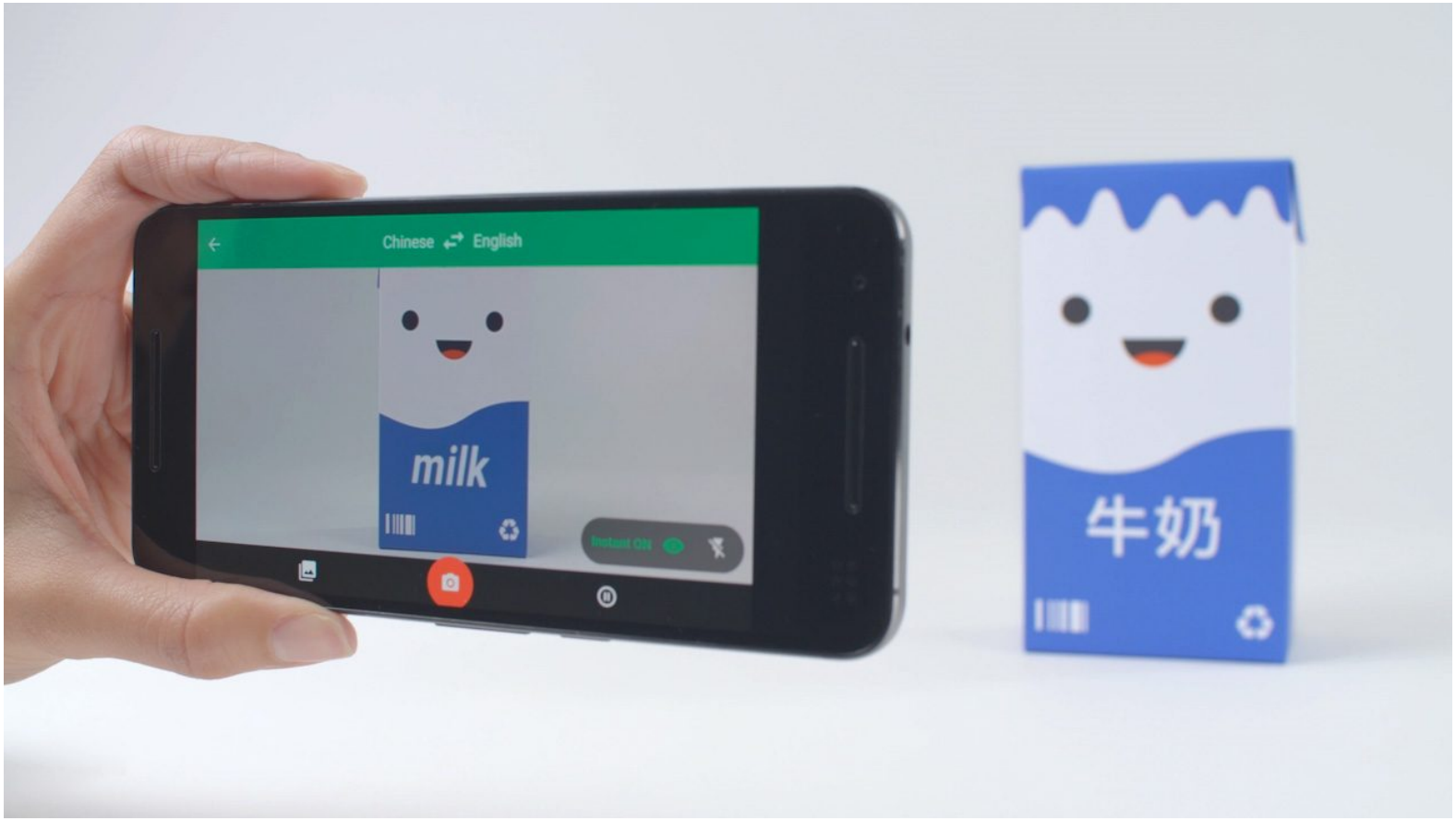
The machine will "think"; that is, it will scan shapes such as the letters of the alphabet and simple words and after analyzing and absorbing this visual information it will "say" (through a loudspeaker) what it has seen at precisely the same rate as that of a fairly intelligent human subject.

This is being achieved by



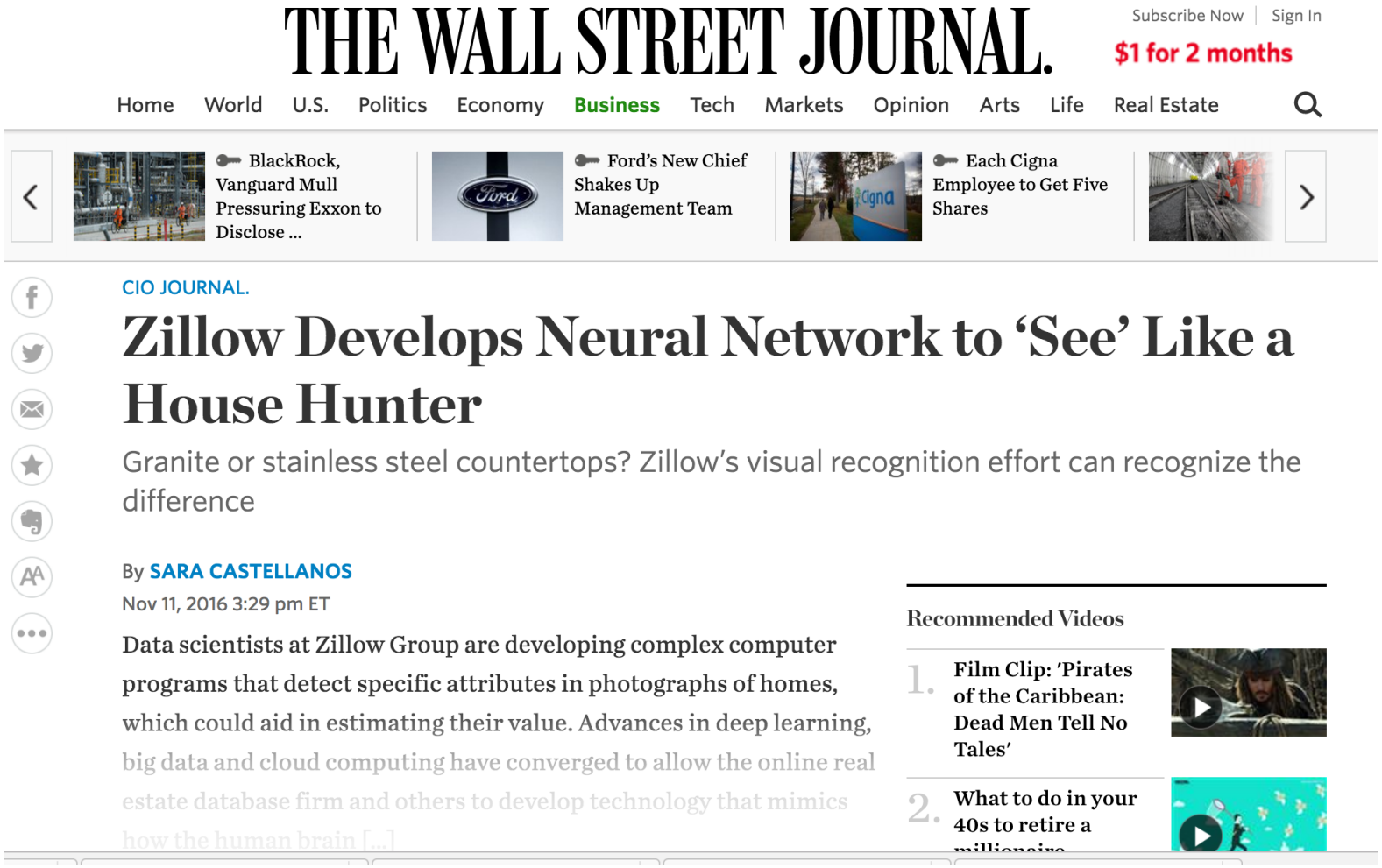
Today: Home assistants

Translation



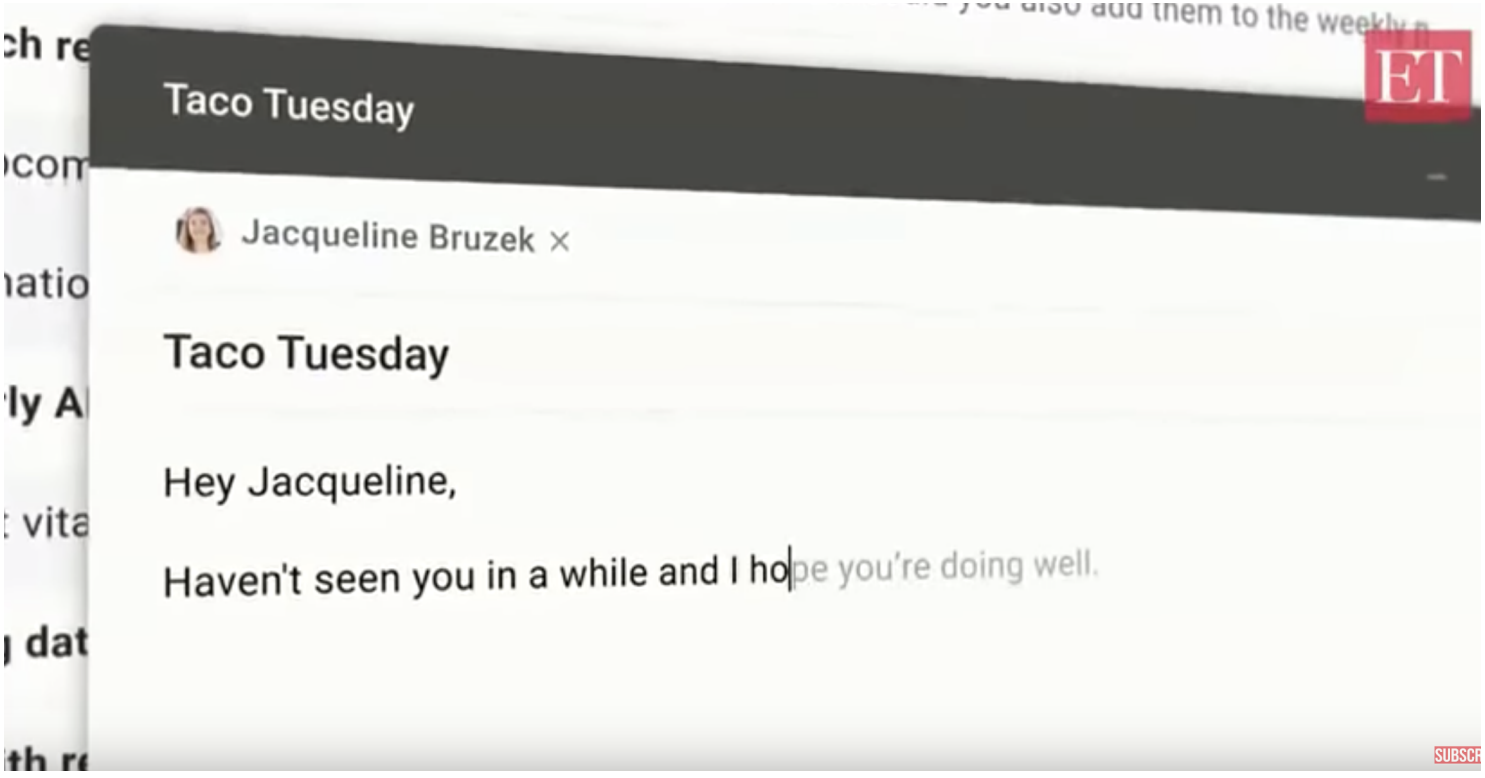
http://www.sciencemag.org/

Pricing and recommending homes



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

Email suggestions



https://www.youtube.com/watch?v=nZ-C8I-8BZw&t=0m16s

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

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

Translation

HOME > SPORTS

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

Lauren Frias 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

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

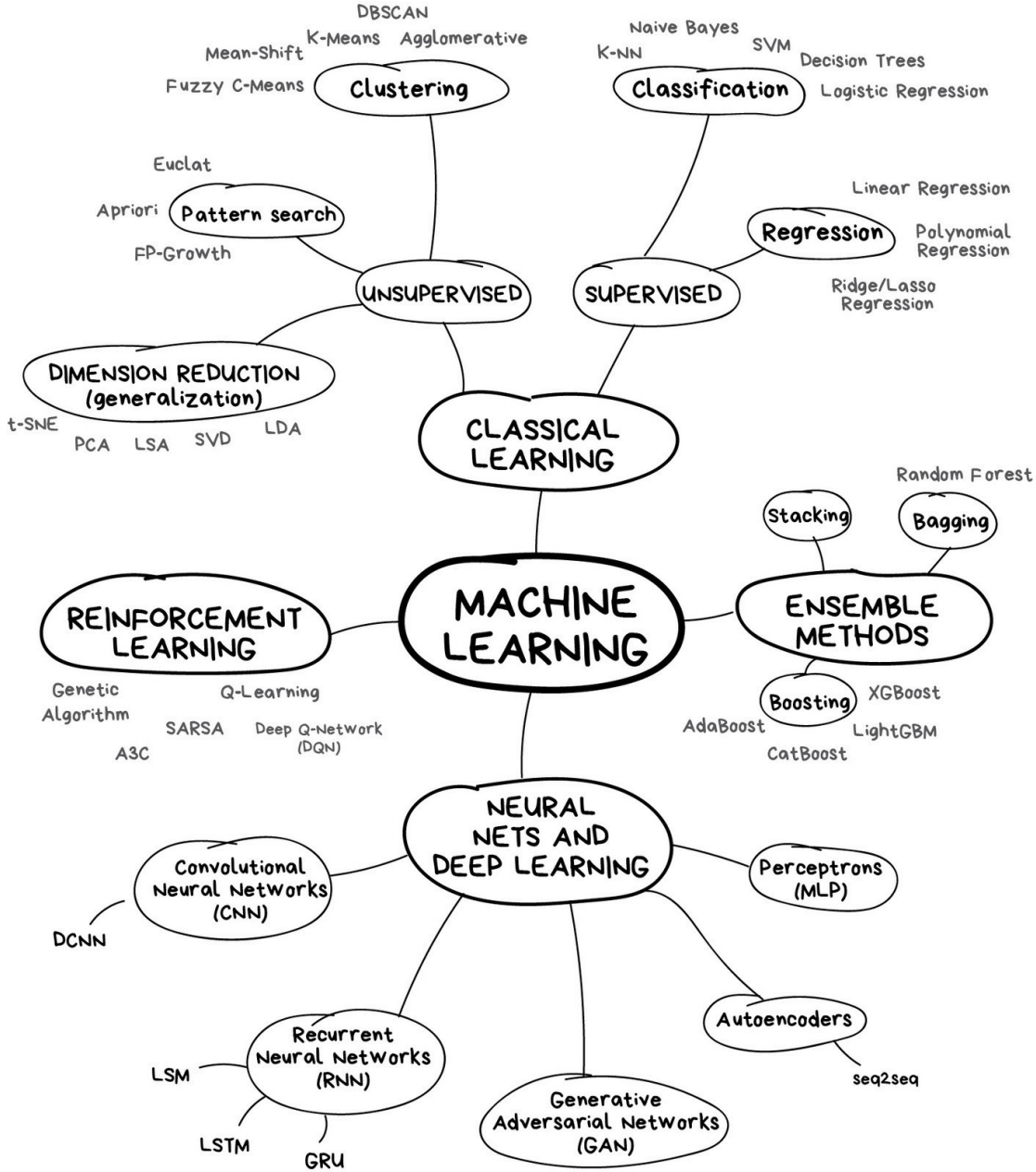
Machine learning focuses on making predictions and inferences from data.

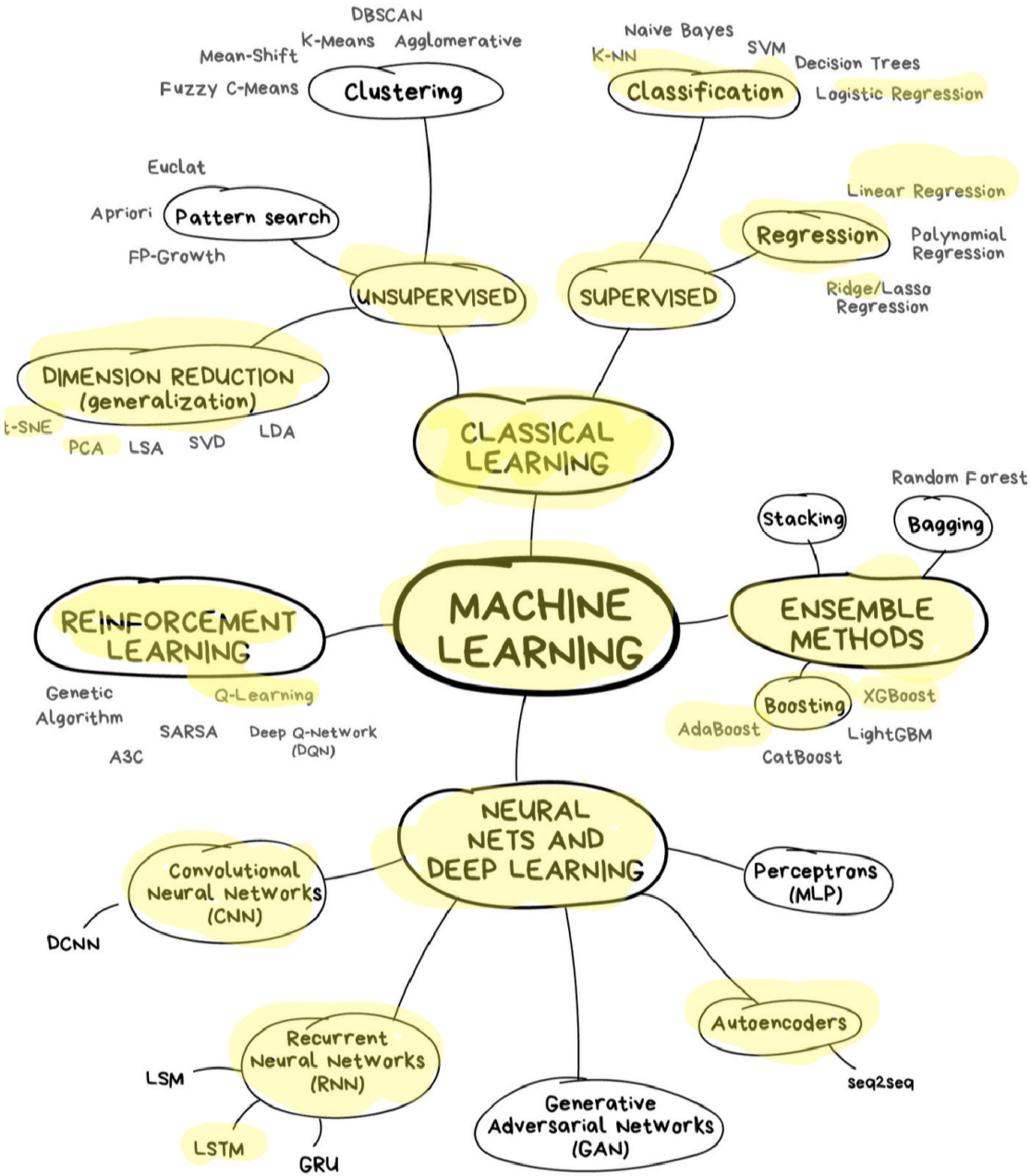
AI combines machine learning components into a larger system that includes a decision making component.

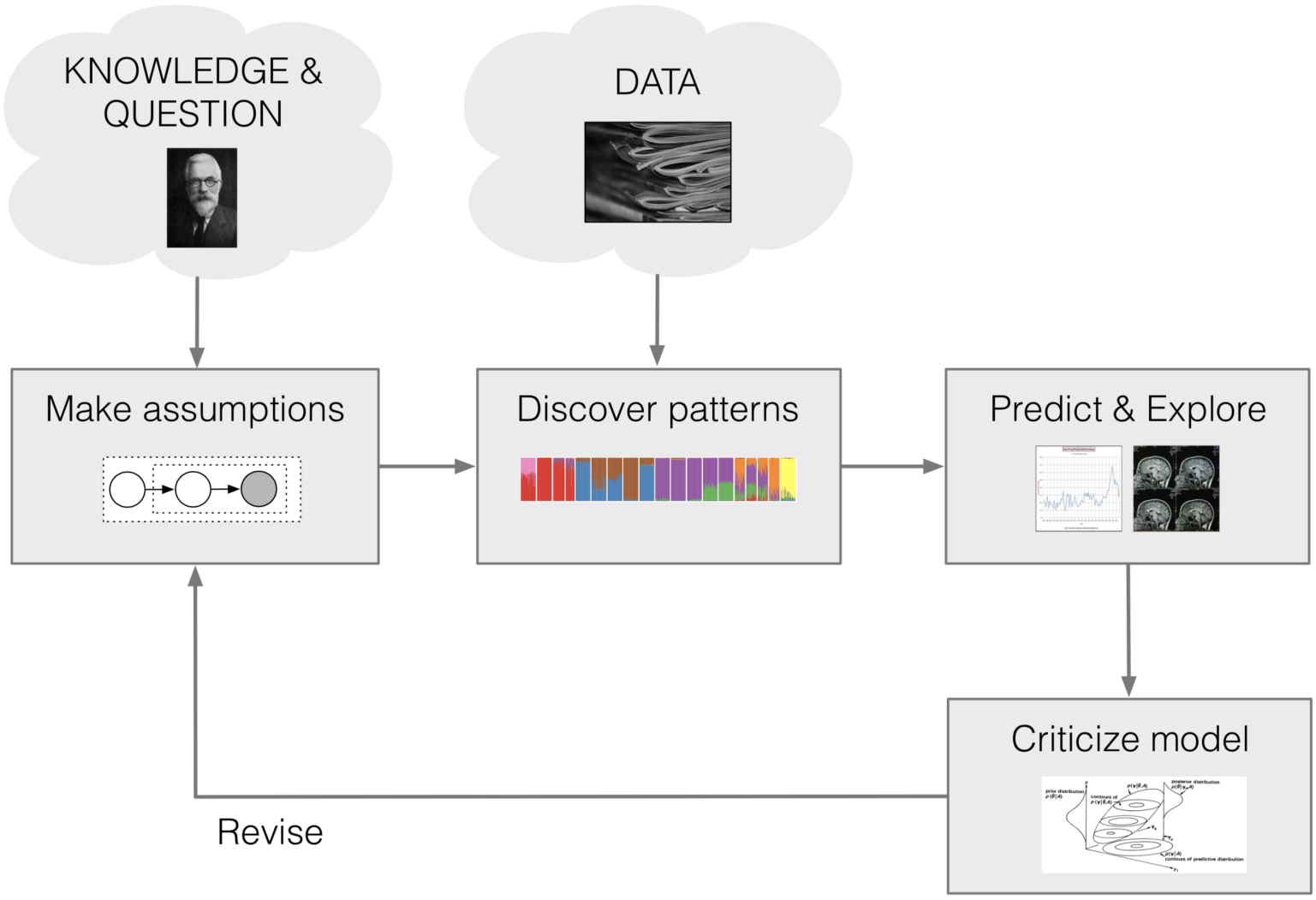
An AI system exhibits a behavior, resulting from the collective decisions that are made.

Machine learning frameworks

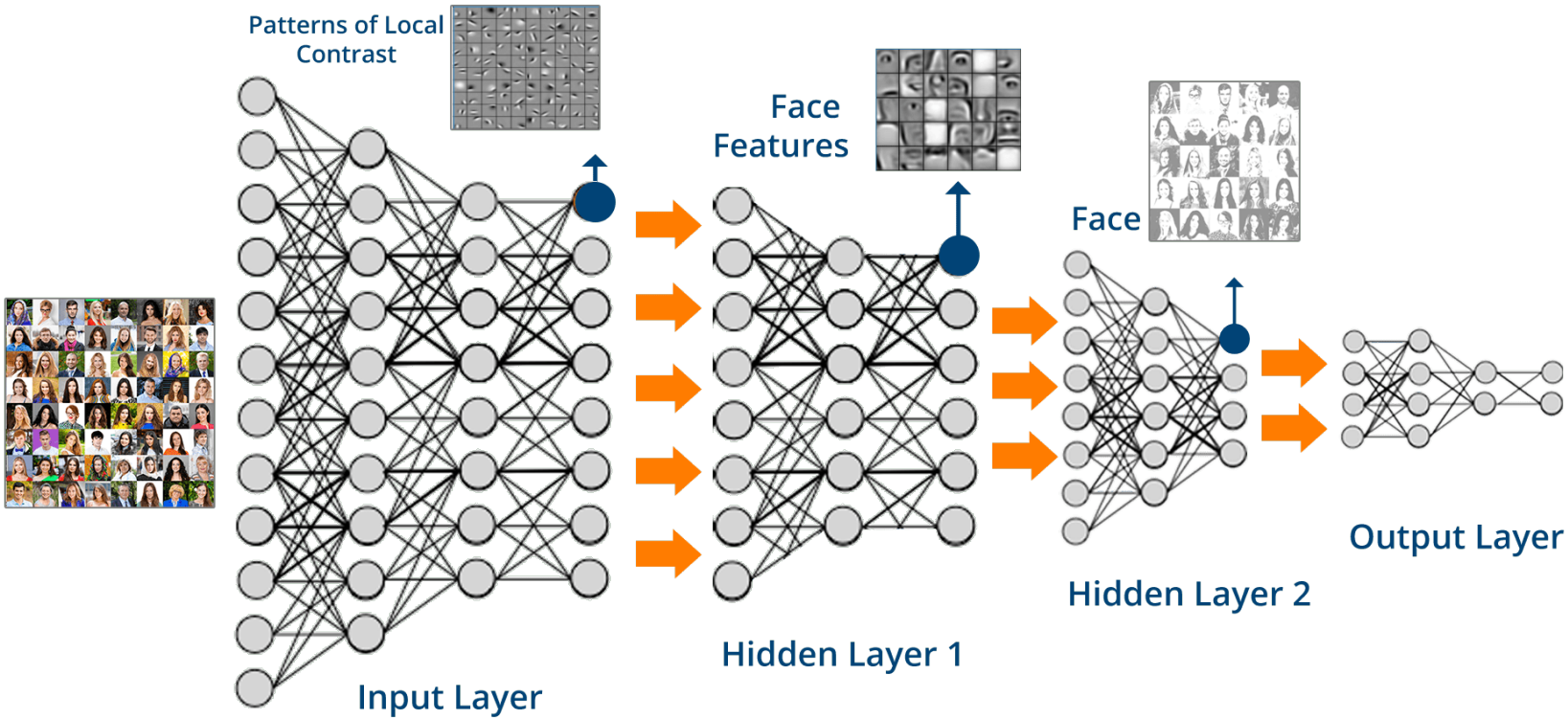
- Supervised, unsupervised, semi-supervised
- Reinforcement learning
- Generative vs. discriminative models
- Representation learning


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Pattern search
Euclat
Apriori
FP-Growth
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
SUPERVISED
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
REINFORCEMENT LEARNING
Genetic Algorithm
A3C
SARSA
Q-Learning
Deep Q-Network (DQN)
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq


KNOWLEDGE &
QUESTION
DATA
Make assumptions
Discover patterns
Predict & Explore
Criticize model
Revise

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

xkcd.com/1838

Latent variables: The elephants in the room

DALL·E 2

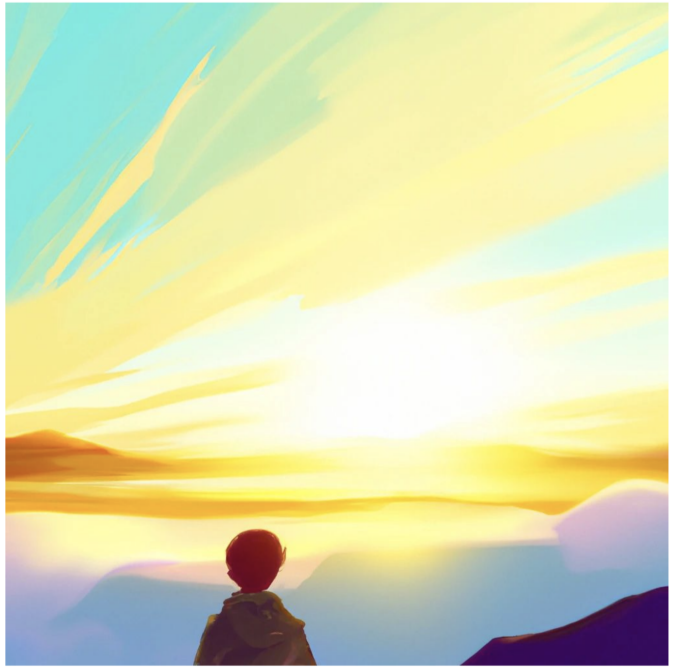
DALL·E 2 is a new AI system that can create realistic images and art from a description in natural language.

- ↗ JOIN WAITLIST
- ↓ EXPLORE
- ▶ WATCH VIDEO
- VIEW RESEARCH
- FOLLOW ON INSTAGRAM

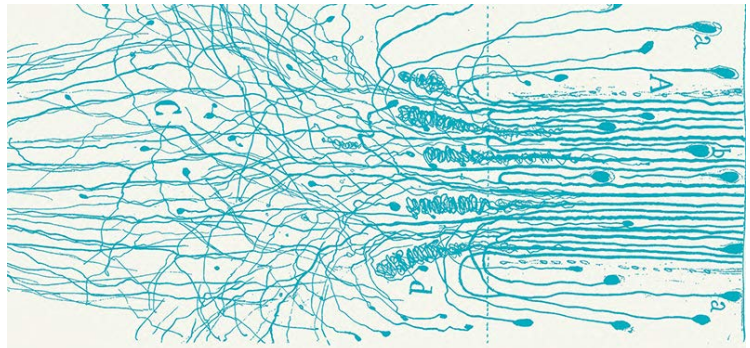
DALL·E Now Available in Beta

We'll invite 1 million people from our waitlist over the coming weeks. Users can create with DALL·E using free credits that refill every month, and buy additional credits in 115-generation increments for $15.

JOIN DALL·E 2 WAITLIST ↗



July 20, 2022
3 minute read

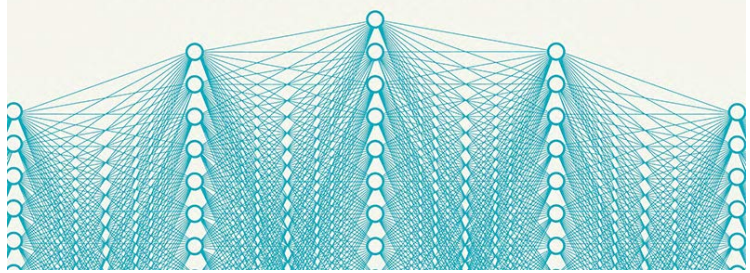


THE ALIGNMENT PROBLEM

Machine Learning and Human Values

BRIAN CHRISTIAN

Best-Selling Author, *Algorithms to Live By*



Example of representation learning: Word embeddings

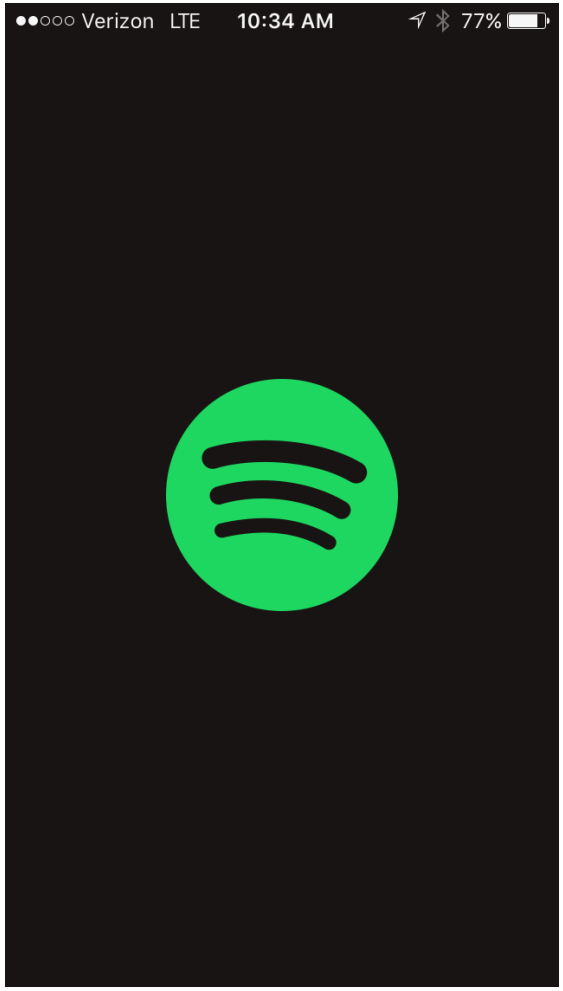
- Each word in vocab is mapped to 100 or 500 dimensional vector
- Based solely on co-occurrence statistics in corpus of text

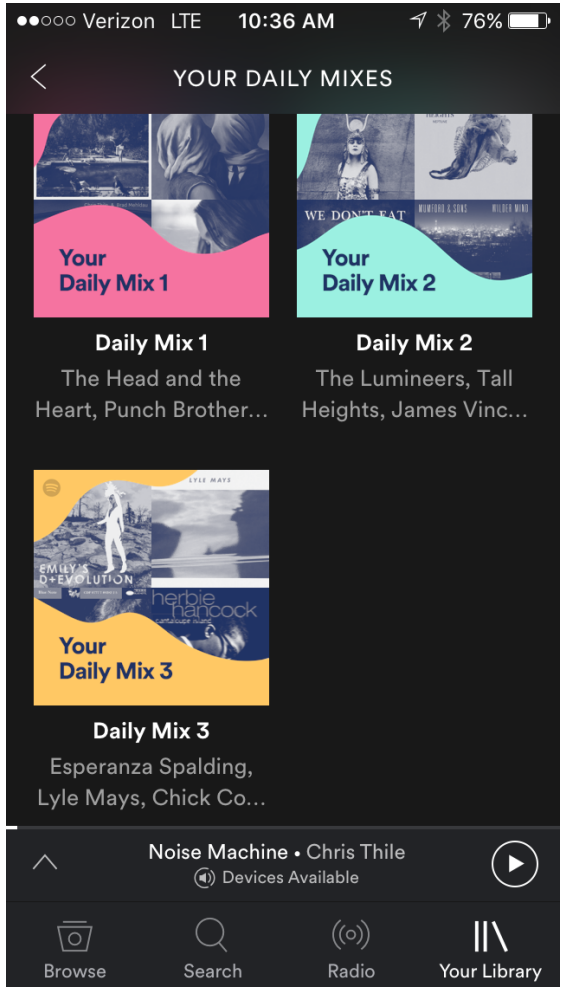
Example of representation learning: Word embeddings

Yale:

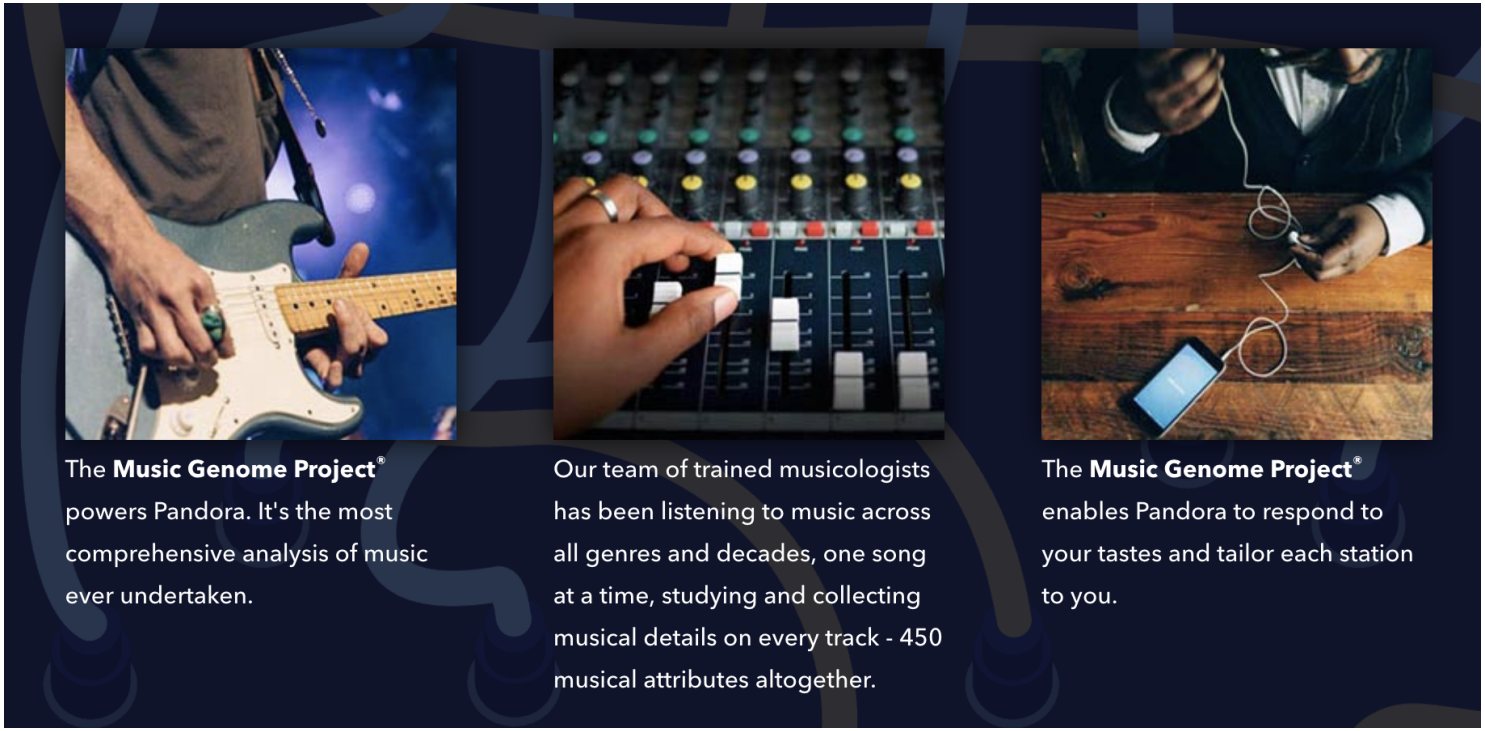
[0.78310001, 0.51717001, -0.38207 , -0.23722 , -0.31615999, 0.30805001, 0.76389998, 0.064106 , -0.74913001, 0.60585999, -0.23871 , -0.16876 , -0.25634 , 1.07270002, -0.29967999, 0.020095 , 0.54500997, -0.17847 , -0.26675999, -0.11798 , -0.48692 , 0.22712 , 0.017473 , -0.4747 , 0.44861001, -0.084281 , -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



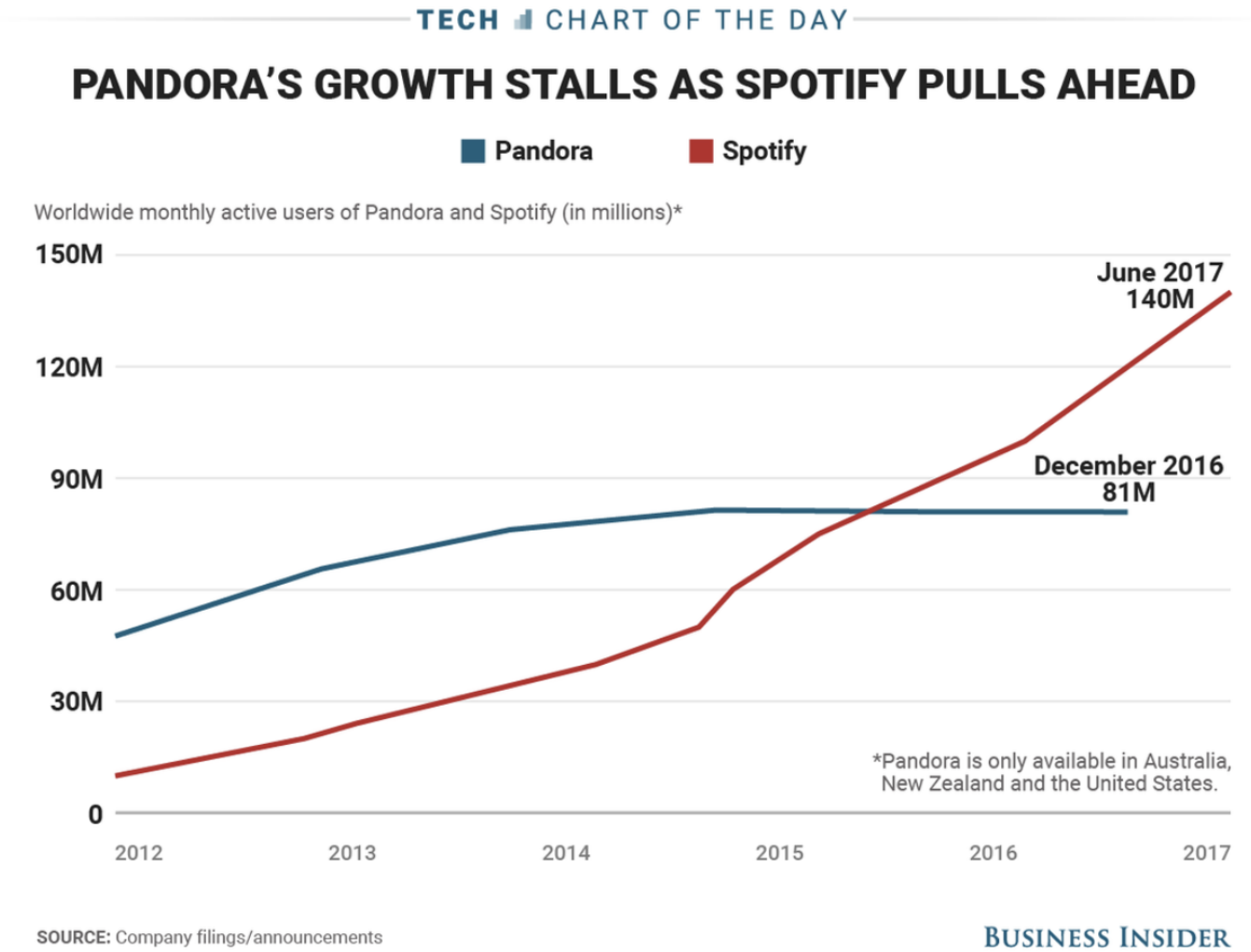
- Pandora's "Music genome": Over 450 musical attributes
- Melody, harmony, rhythm, form, composition, lyrics...

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



Hacking ML Systems

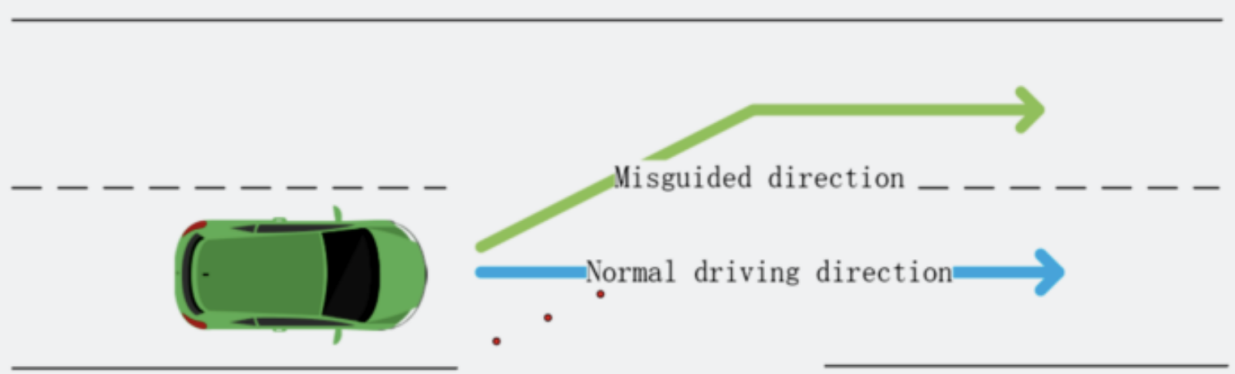
ars TECHNICA SUBSCRIBE SIGN IN

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
- Users are hashed; employees don’t have access to friends’ data

Reasons for optimism

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

Questions or discussion?

Team

- Instructor
  John Lafferty
- Teaching Fellows
  Zhehao Xu (PhD student in S&DS)
  Hannah Wang (Graduate student, School of the Environment)
  Wendy Luo (Graduate student, Biomedical Engineering)
- ULAs
  Yuxuan Geng, Lucas Zheng, Jacob Alvarado, Edward Hu, Kaitlin Flores, Mofeed Nagib, Teckhua Chiang
- Course manager
  Hanwen Gu

Course materials

Materials posted to `https://iml.ydata123.org`; sometimes to Canvas

Please use Ed Discussion for any questions about lectures, homework, etc. first, before email!

For email

- For logistical issues (e.g. adding to Canvas): E-mail Yiyang, copy me, Chris, Wendy
- All other: E-mail me, copy TAs

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
- Have fun
- Want to learn more

Evaluation

- Six assignments (50%)
- Mid-semester exam (20%)
- Six quizzes (10%)
- Final exam: 20%

Lowest assignment 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.

Week	Dates	Topics	Demos & Tutorials	Lecture Slides	Readings and Notes	Assignments & Exams
1	Sept 1	Course overview		Sept 1: Course overview		
2	Sept 6, 8	Python and background concepts	Python elements Covid trends	Sept 6: Python elements Sept 8: Pandas and linear regression	Data8 Chapters 3, 4, 5	Thu: Quiz 1
3	Sept 13, 15	Linear regression and classification	Covid trends (revisited) Classification examples	Sept 13: Regression concepts Notes on regression Sept 15: Classification	Notes on classification	Thu: Assn1 out
4	Sept 20, 22	Stochastic gradient descent	SGD examples	Sept 20: Classification (continued) Sept 22: Stochastic gradient descent		Thu: Quiz 2
5	Sept 27, 29	Bias and variance, cross-validation	Bias-variance tradeoff Covid trends (revisited) California housing	Sept 27: Bias and variance Sept 29: Cross-validation		Thu: Assn 1 in Assn2 out
6	Oct 4, 6	Tree-based methods	Trees and forests Visualizing trees	Oct 4: Trees Oct 6: Forests		Thu: Quiz 3
7	Oct 11, 13	PCA and dimension reduction	PCA examples PCA revisited Used for regression	Oct 11: PCA Oct 13: PCA and review		Thu: Assn 2 in Assn3 out

9	Oct 25, 27	Language models, word embeddings	Word embeddings	Oct 25: Language models Oct 27: Word embeddings		Thu: Assn 3 in Assn4 out
10	Nov 1, 3	Bayesian inference, topic models	Mixtures Bayesian inference Topic models	Nov 1: Bayesian inference Nov 3: Bayes and topic models	Notes on Bayesian inference Notes on simulation	Thu: Quiz 4
11	Nov 8, 10	Introduction to neural networks	Minimal neural network Regression examples	Nov 8: Topic models Nov 10: Neural networks		Thu: Assn 4 in Assn5 out
12	Nov 15, 17	Deep neural networks	Tensorflow playground Autoencoder examples	Nov 15: Neural networks (continued) Nov 17: Autoencoders	Notes on backpropagation	Thu: Quiz 5
13	Nov 22, 24	No class, Thanksgiving break				
14	Nov 29, Dec 1	Reinforcement learning	Q-learning	Nov 29: Reinforcement learning Dec 1: Deep reinforcement learning		Thu: Assn 5 in Assn 6 out
15	Dec 6, 8	Societal issues for machine learning		Dec 6: Societal issues Dec 8: Course wrap up		Thu: Quiz 6
16	Dec 15					Thu: Assn 6 in
		Mon, Dec 19, 7pm	Final exam		Registrar: Final exam schedule Practice final, sample solution	

Auditing

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

Questions?