Statistics and Data Science 265

Introductory Machine Learning

Thursday, September 2

Yale

Outline

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

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 wider range of students
- S&DS 365/565 will become “Intermediate Machine Learning,” can be taken as a follow up course.
- Talk to me if unclear or undecided

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

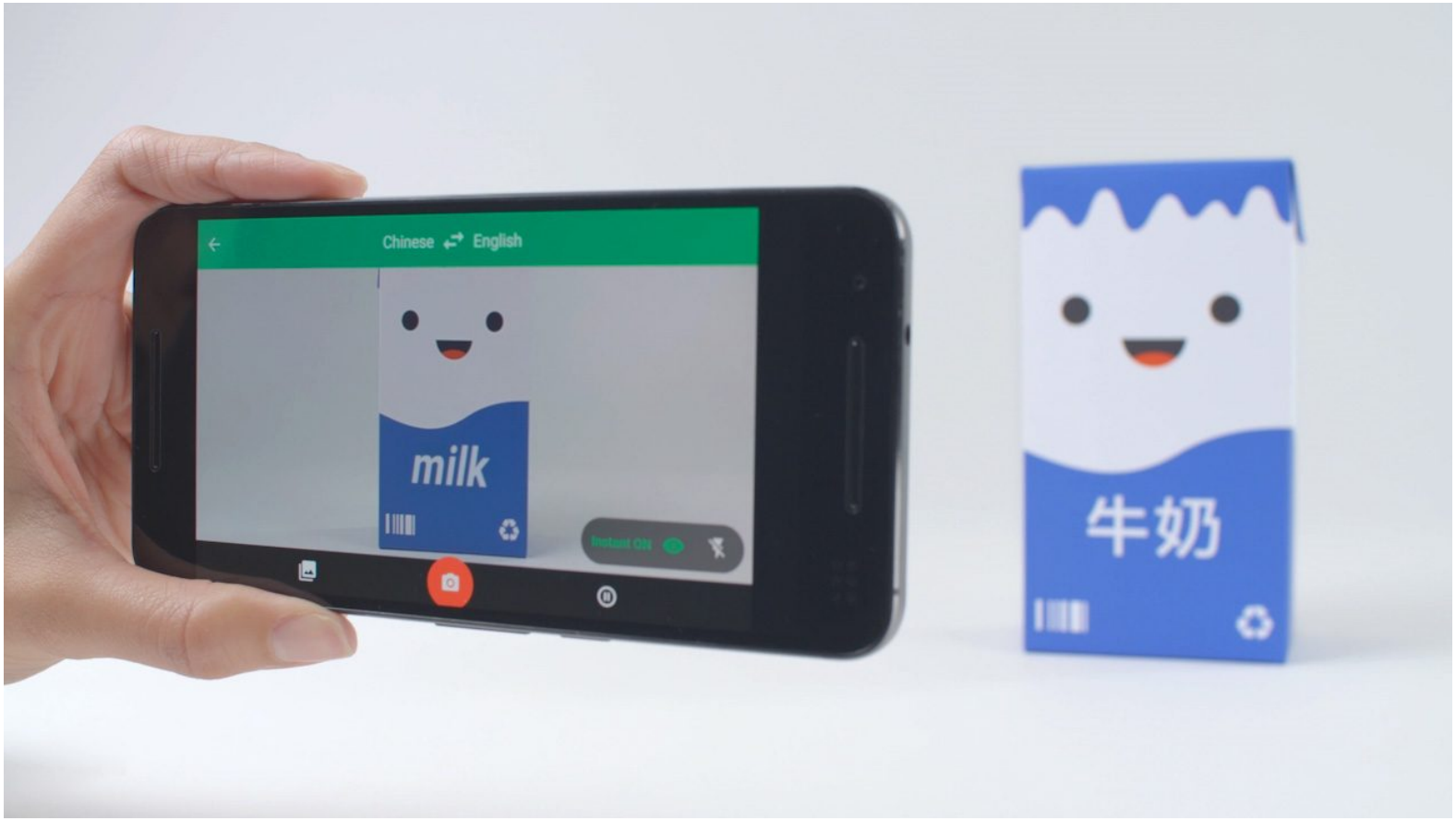
DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

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.

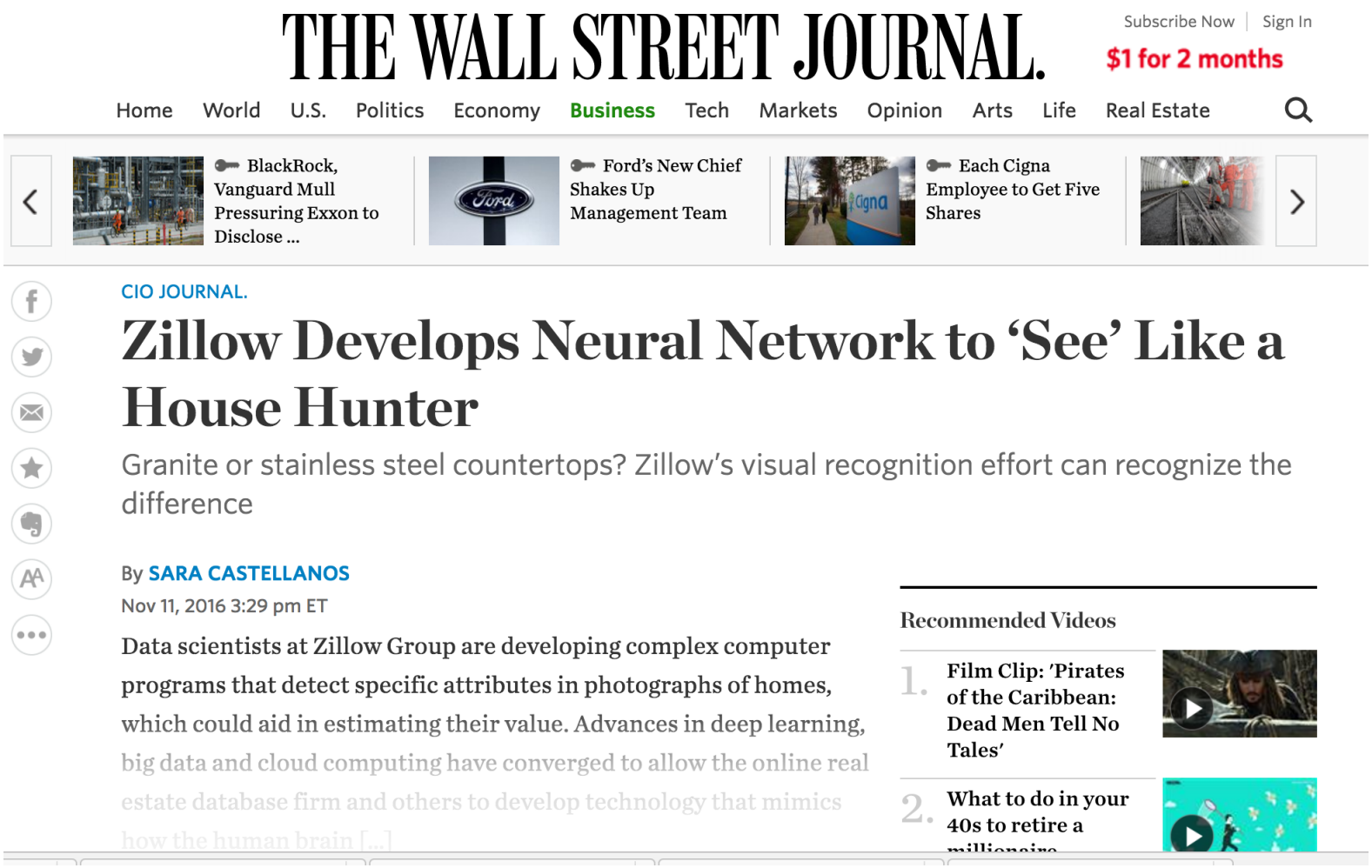
Today: Home assistants

Translation



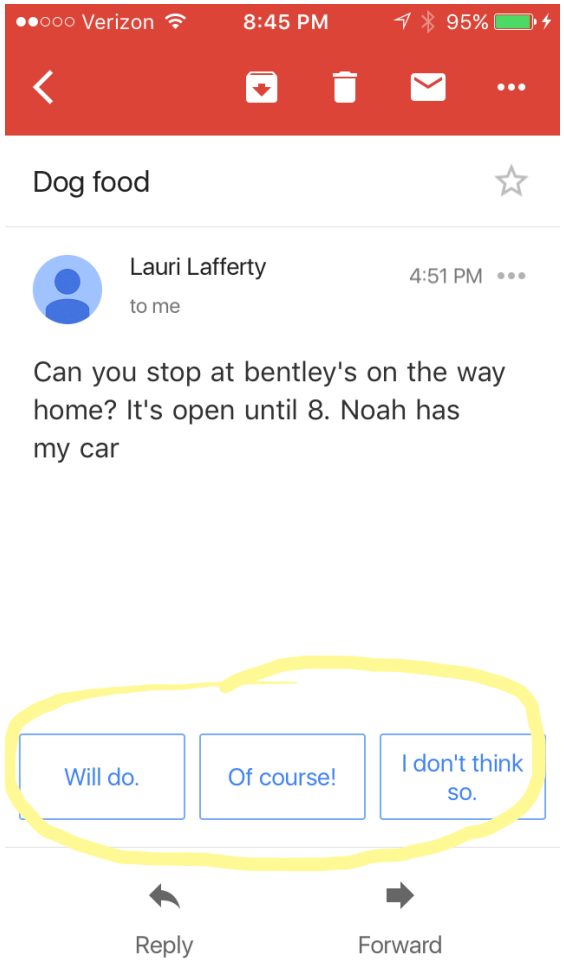
http://www.sciencemag.org/

Pricing and recommending homes

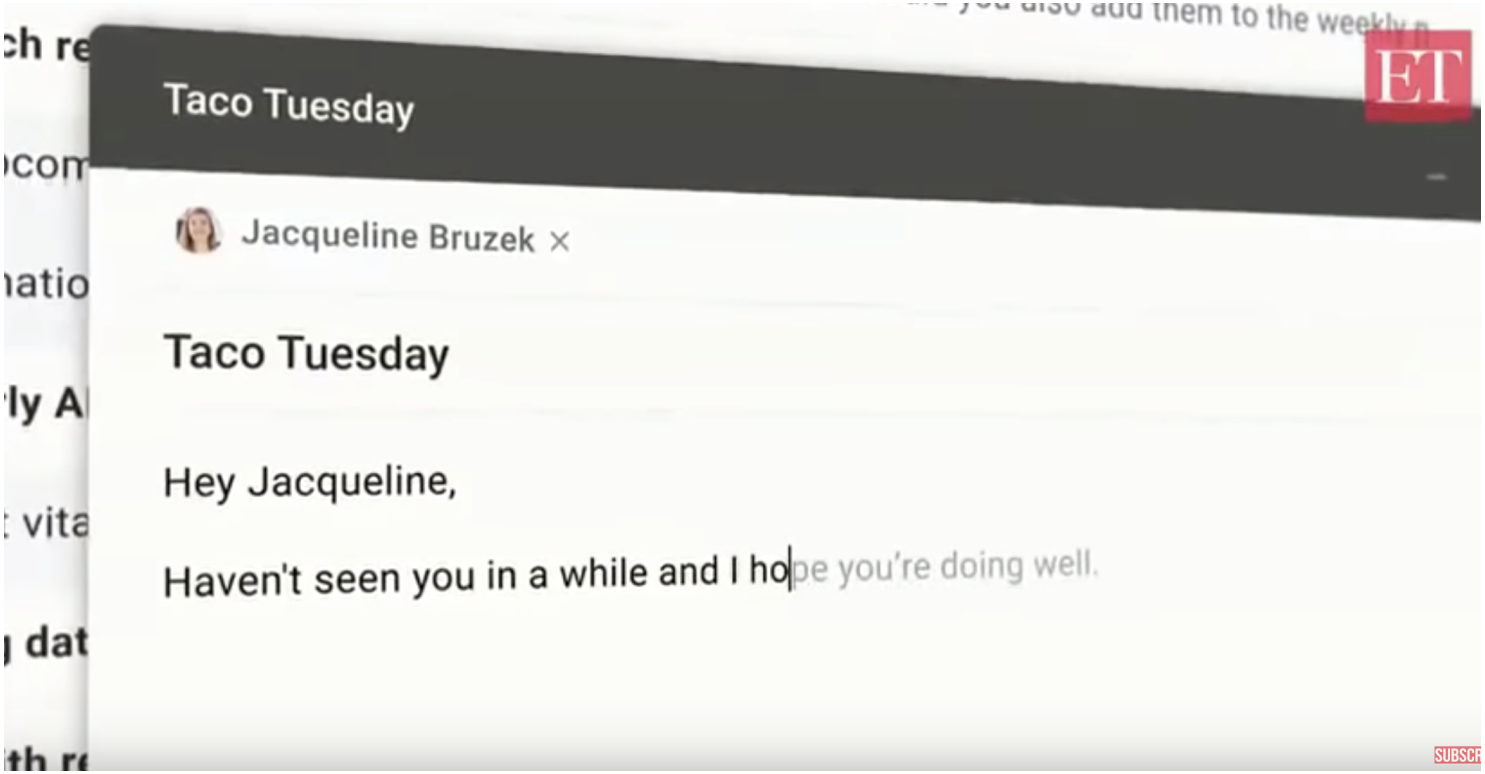


https://blogs.wsj.com/cio/2016/11/11/zillow-develops-neural-network-to-see-like-a-home-buyer/

Email replies



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

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

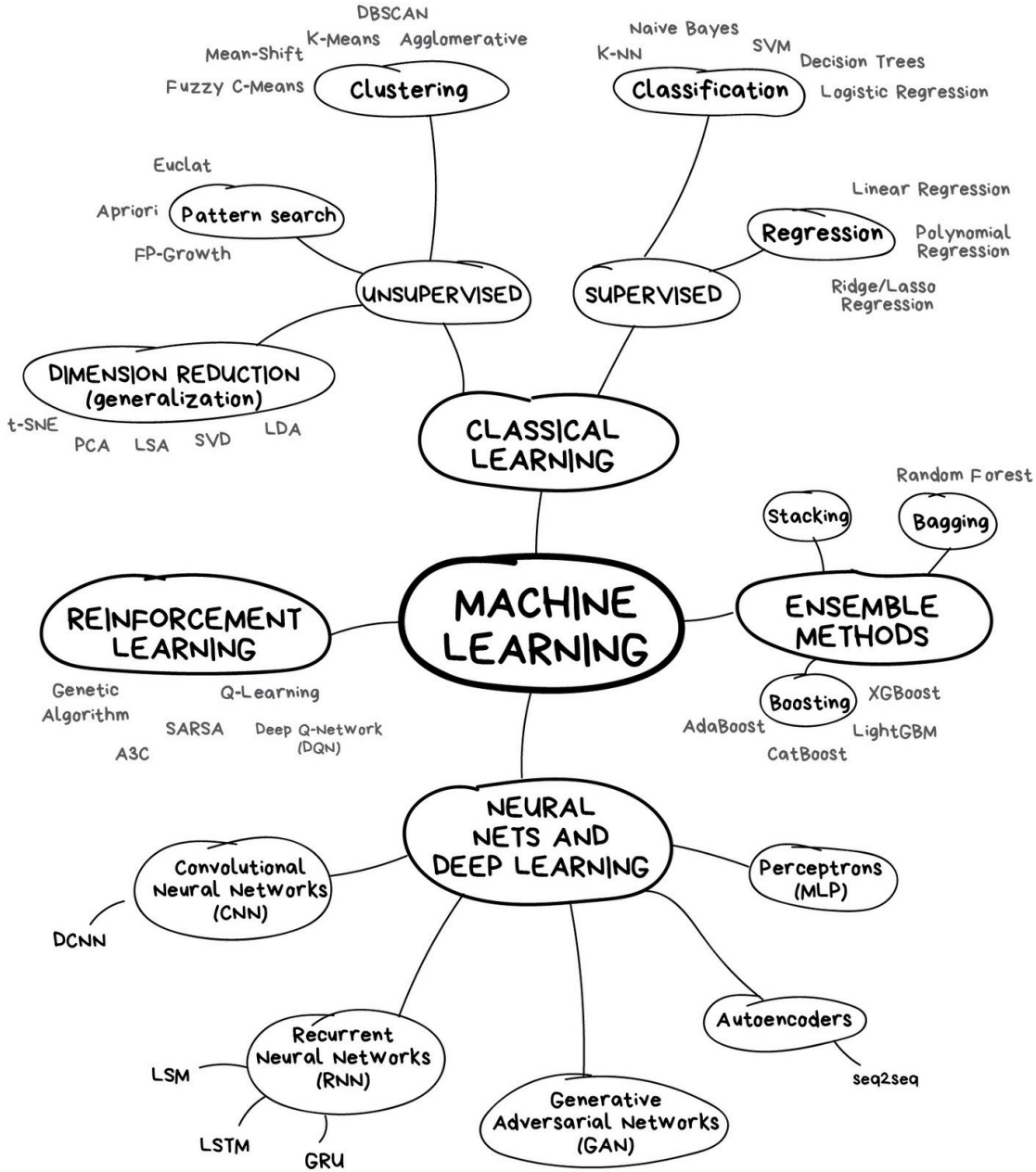
Machine learning focuses on making predictions and inferences from data.

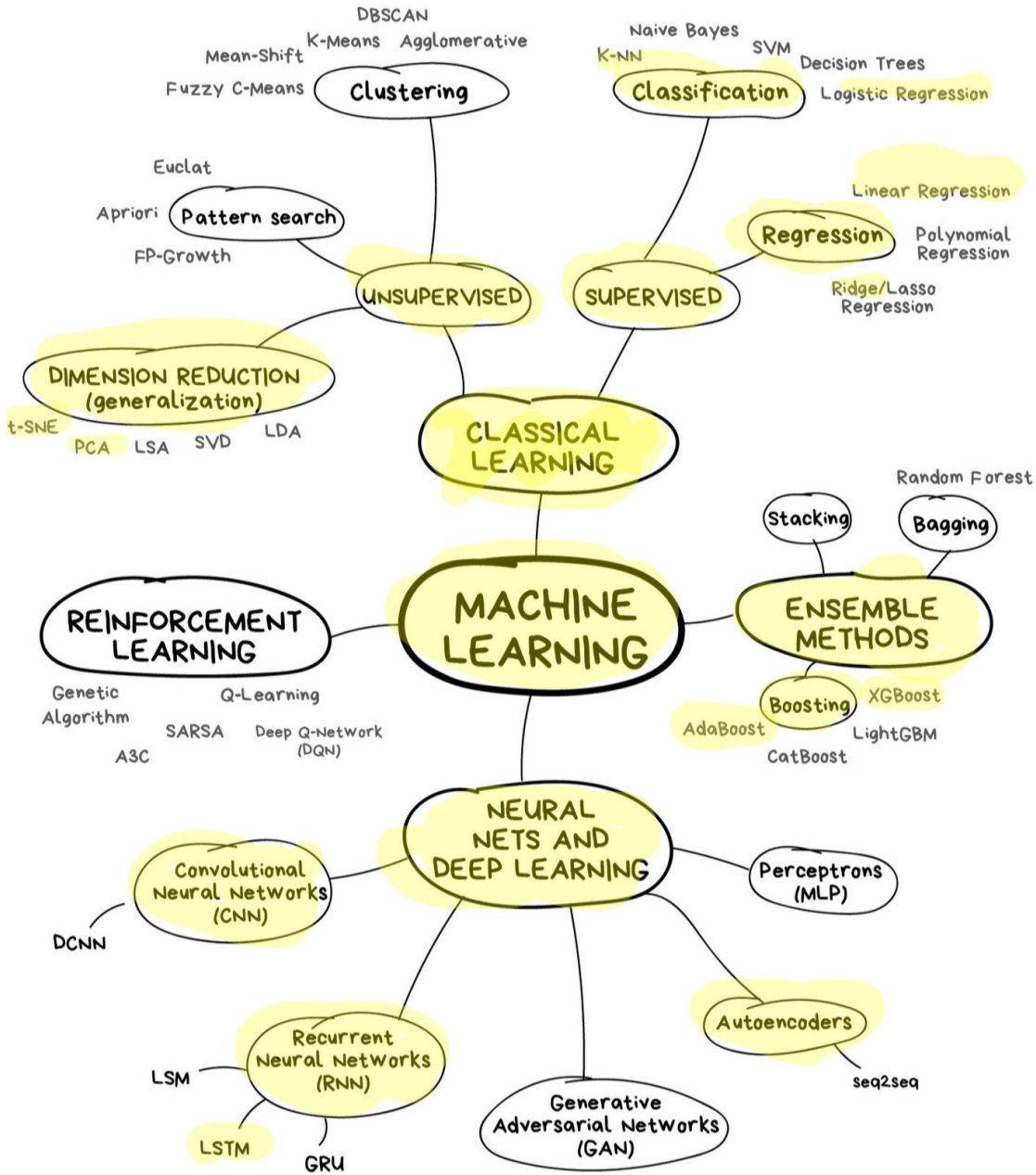
AI combines machine learning components into a larger system that includes a decision making component.

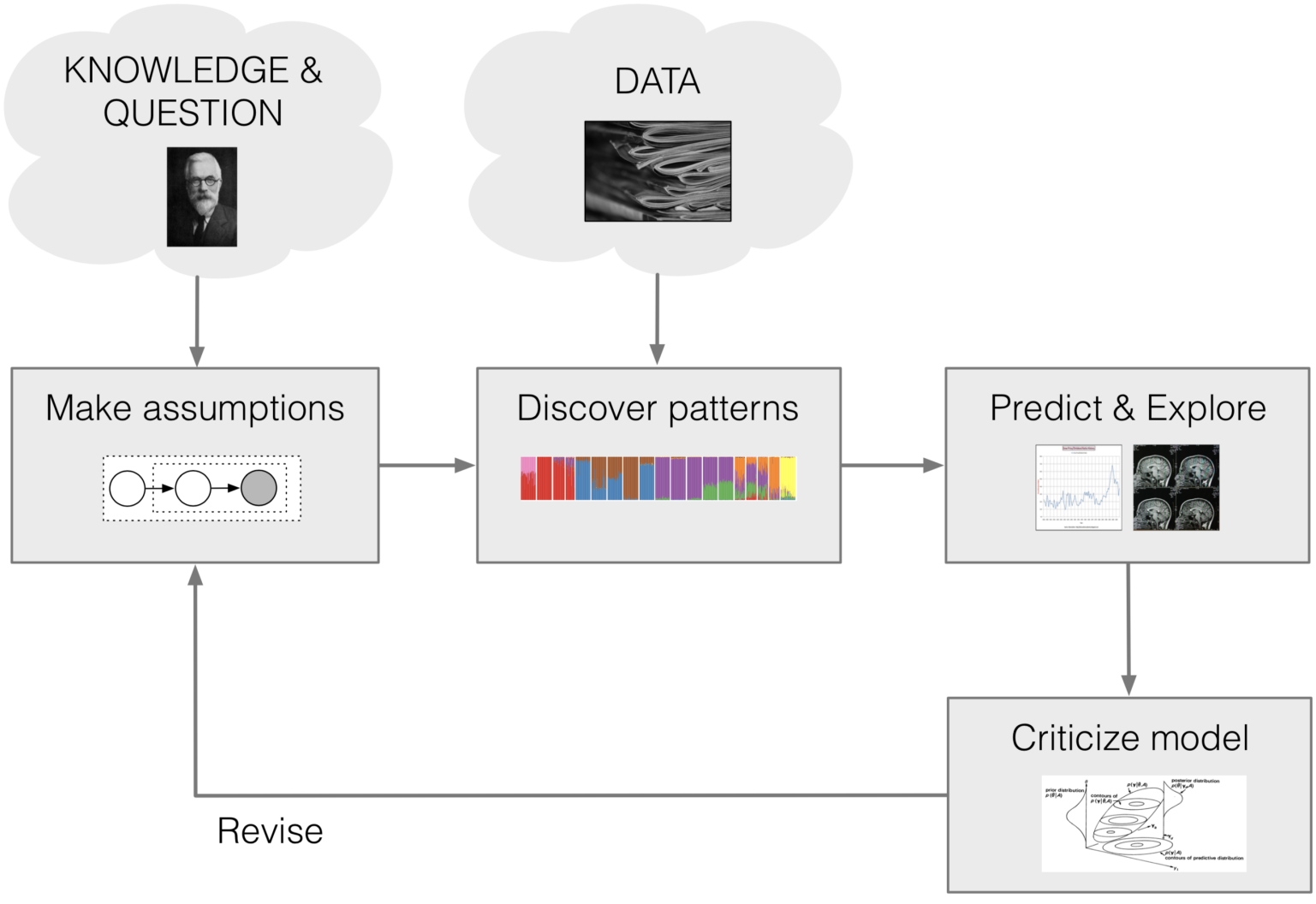
An AI system exhibits a behavior, resulting from the collective decisions that are made.

Machine learning frameworks

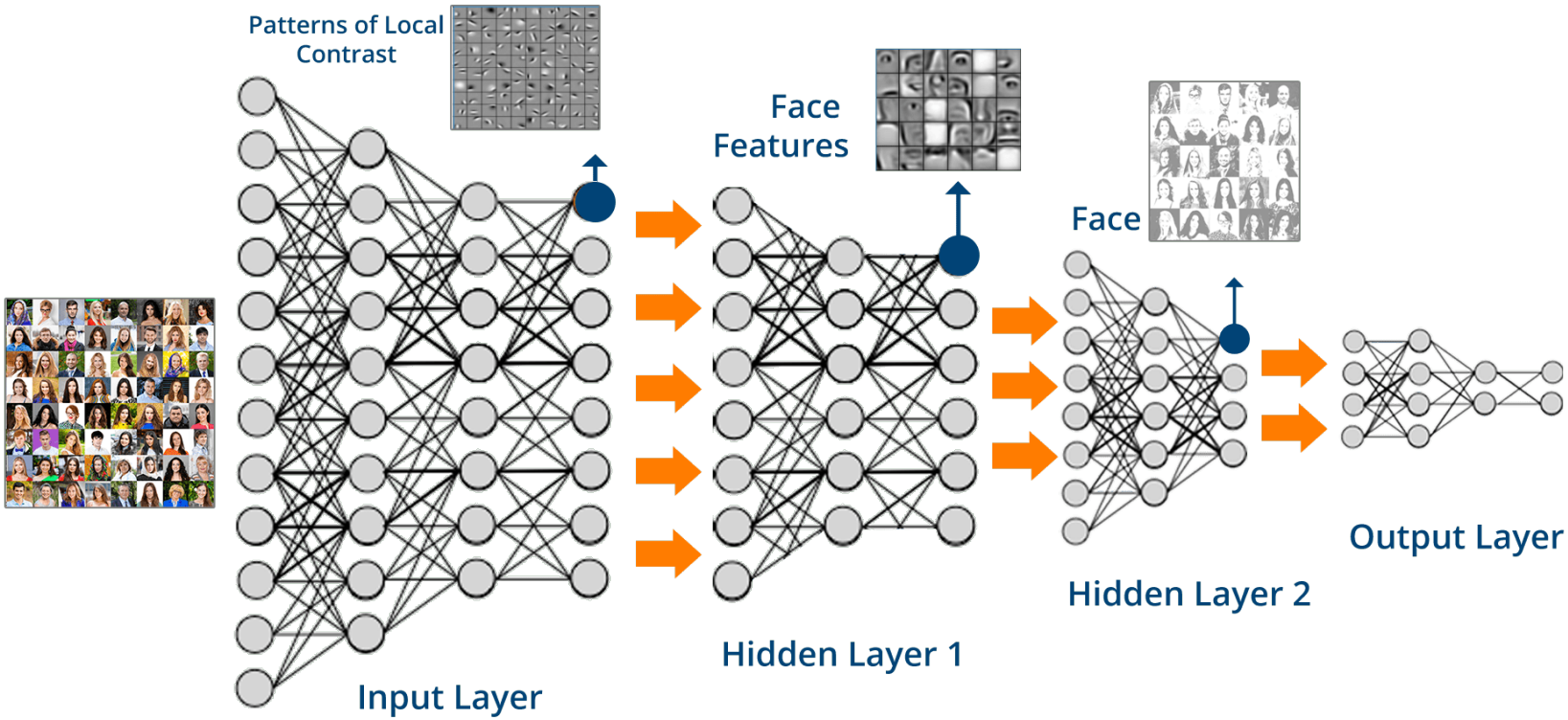
- Supervised, unsupervised, semi-supervised
- Reinforcement learning
- Generative vs. discriminative models
- Representation learning


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
SVM
Decision Trees
K-NN
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
Q-Learning
SARSA
Deep Q-Network (DQN)
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
AdaBoost
XGBoost
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

xkcd.com/1838

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
- Have patterns replaced principles?

Latent variables: The elephants in the room

Example of representation learning: Word embeddings

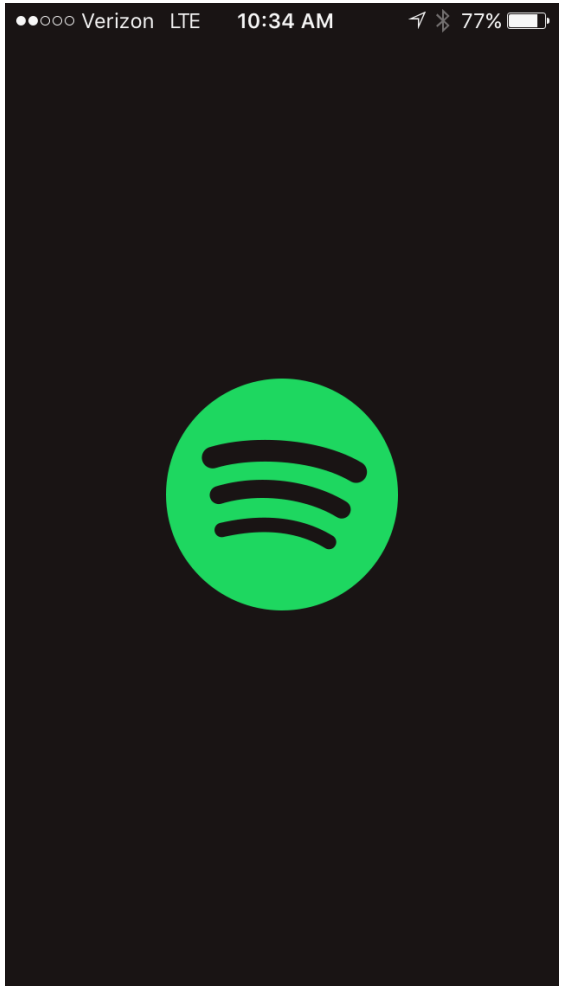
- Each word in vocab is mapped to 100 or 500 dimensional vector
- Based solely on co-occurrence statistics in corpus of text

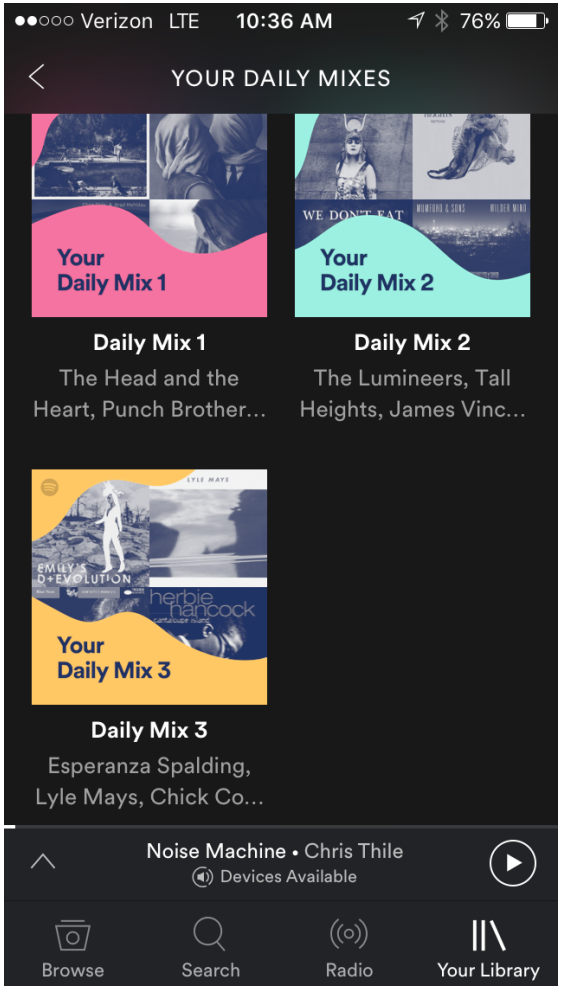
Example of representation learning: Word embeddings

Yale:

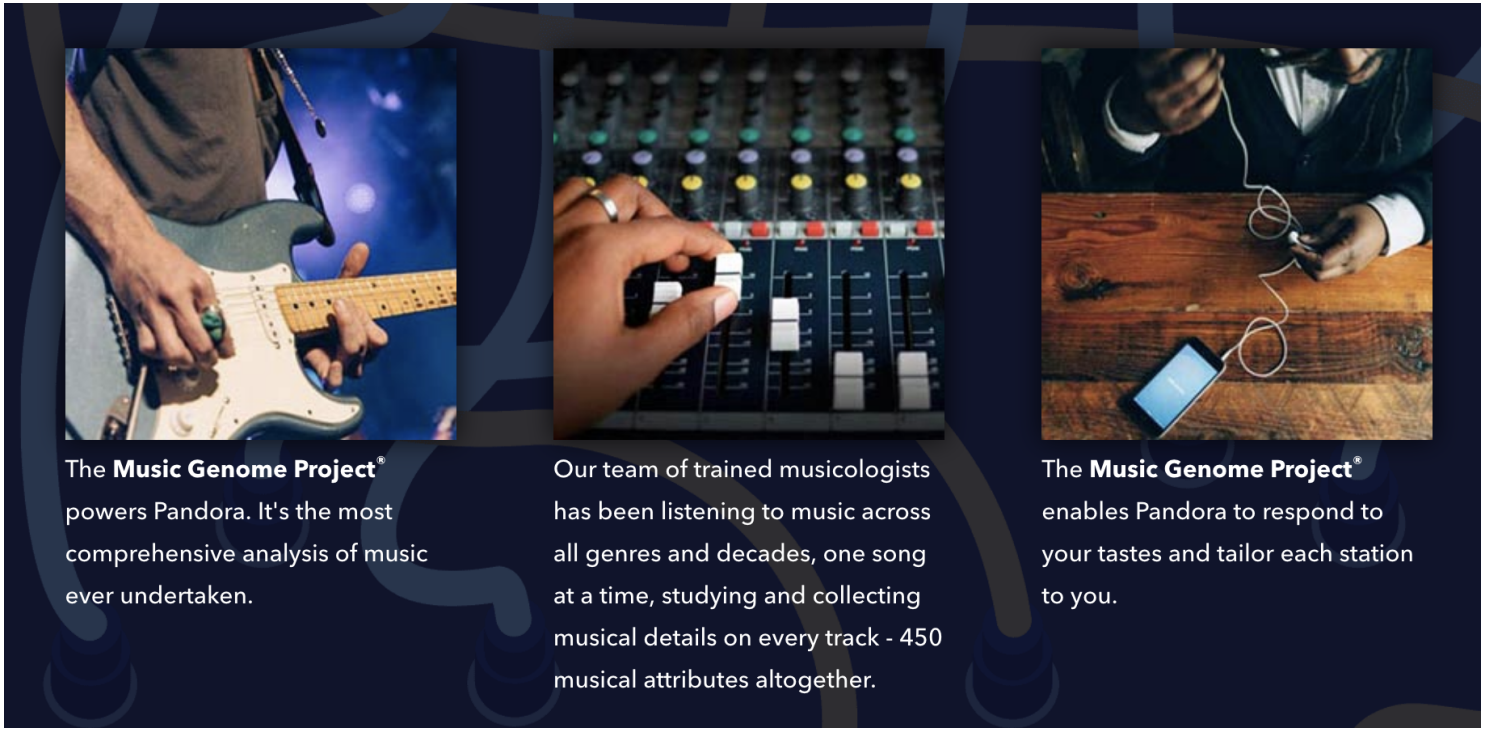
[0.78310001, 0.51717001, -0.38207 , -0.23722 , -0.31615999, 0.30805001, 0.76389998, 0.064106 , -0.74913001, 0.60585999, -0.23871 , -0.16876 , -0.25634 , 1.07270002, -0.29967999, 0.020095 , 0.54500997, -0.17847 , -0.26675999, -0.11798 , -0.48692 , 0.22712 , 0.017473 , -0.4747 , 0.44861001, -0.084281 ,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



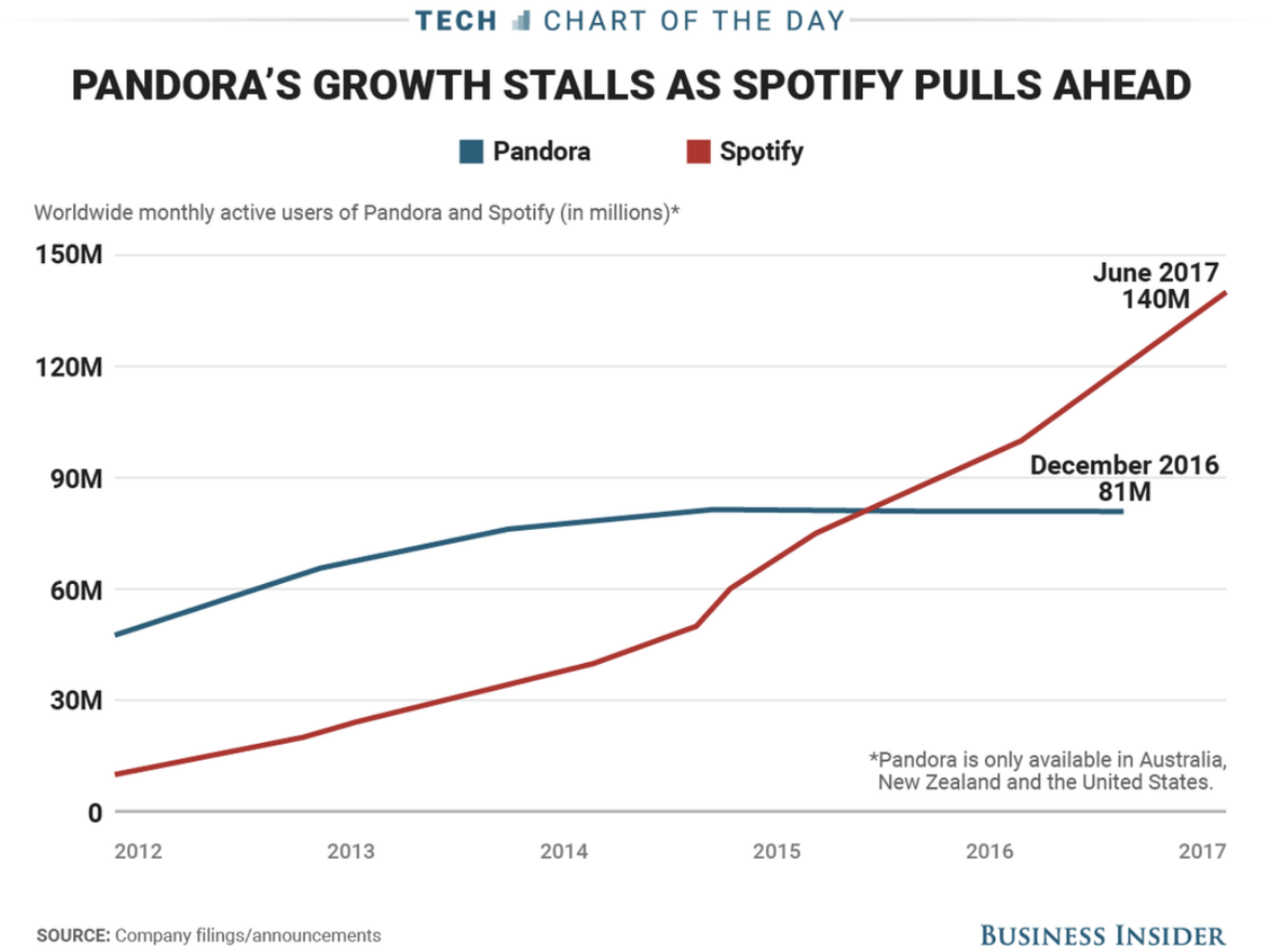
- Pandora's "Music genome": Over 450 musical attributes
- Melody, harmony, rhythm, form, composition, lyrics...

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



News from last year

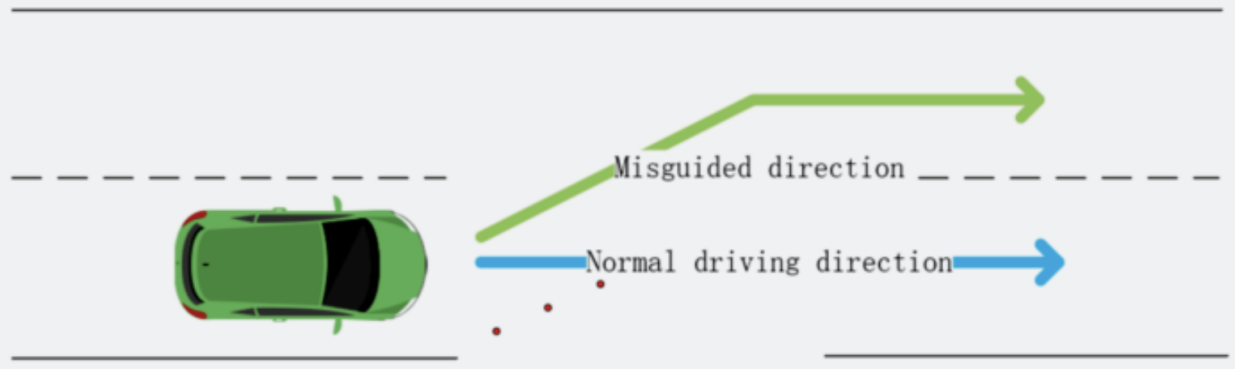
ars TECHNICA SUBSCRIBE SIGN IN

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
- This course and you!

Let’s pause for questions and discussion

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

Team

- Instructor: John Lafferty (Prof, DS2 and CS)
- Teaching Fellows

  Wendy Luo (PhD student in Biomedical Engineering)
  Jerome Yu (PhD student in Comp. Bio. and Bioinformatics)

- ULAs (4-5)

Office hours (tentative)

Johnny Xu (ULA)	Sunday	6:00 pm
Andrew Wei (ULA)	Monday	8:00 pm
John Lafferty	Tuesday	3:00 pm
Daniel Zhao (ULA)	Tuesday	8:00 pm
SK Bong (ULA)	Wednesday	2:00 pm
Wendy Luo (TF)	Thursday	3:00 pm
Jerome Yu (TF)	Thursday	8:00 pm

Zoom links will be posted to Canvas

Piazza

Materials posted to Canvas. Discussion and copies of some materials on Piazza.

`https://piazza.com/yale/fall2020/sds355555`

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

For email

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

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 and deep 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

- A beginner's guide to installing Python and Jupyter on your computer is here: `https://bit.ly/22KVCfsV`
- See installation guide on course Canvas site
- Use Python 3.x version

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

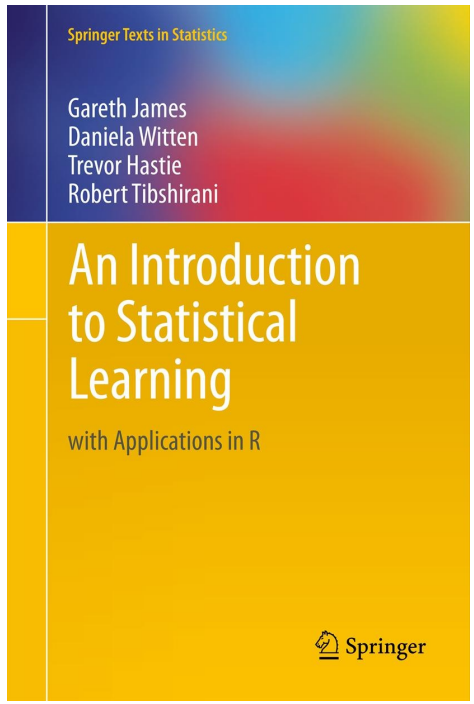
Evaluation

- Seven assignments (50%)
- Two mid-semester exams (25%)
- Several short quizzes (10%)
- Final exam: 15%

Lowest assignment score will be dropped. Late assignments not accepted

Possibly useful reference

- "An Introduction to Statistical Learning," by G. James, D. Witten, T. Hastie, and R. Tibshirani, Springer (2013), `http://www-bcf.usc.edu/~gareth/ISL`



Assignments

- Roughly every 1.5 weeks
- Due at 11:59pm on the day
- Submitted using Canvas
- 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	Date	Topic	Out/Due
1	Sept 1 Sept 3	course introduction background concepts, programming framework	
2	Sept 8 Sept 10	linear regression and classification	assn 1 out
3	Sept 15 Sept 17	stochastic gradient descent	quiz 1 assn 1 in; assn 2 out
4	Sept 22 Sept 24	bias and variance, cross-validation	
5	Sept 29 Oct 1	tree-based methods	assn 2 in; assn 3 out
6	Oct 6 Oct 8	PCA and dimension reduction	quiz 2 assn 3 in; assn 4 out
7	Oct 13 Oct 15	mixtures and Bayes	

Week	Date	Topic	Out/Due
8	Oct 20 Oct 22	topic models	midterm exam 1
9	Oct 27 Oct 29	language models, word embeddings	assn 4 in; assn 5 out
10	Nov 3 Nov 5	introduction to neural networks	quiz 3 assn 5 in; assn 6 out
11	Nov 10 Nov 12	autoencoders and multilayer networks	assn 6 in; assn 7 out
12	Nov 17 Nov 19	sequence models	midterm exam 2 assn 6 in; assn 7 out
13	Nov 21 Nov 23	no class (Thanksgiving break)	
14	Dec 1 Dec 3	further study in machine learning review	quiz 4 assn 7 in

Auditing

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

Questions?