Introduction to Machine Learning

Augustin Cosse.

Summer 2021

May 24, 2021

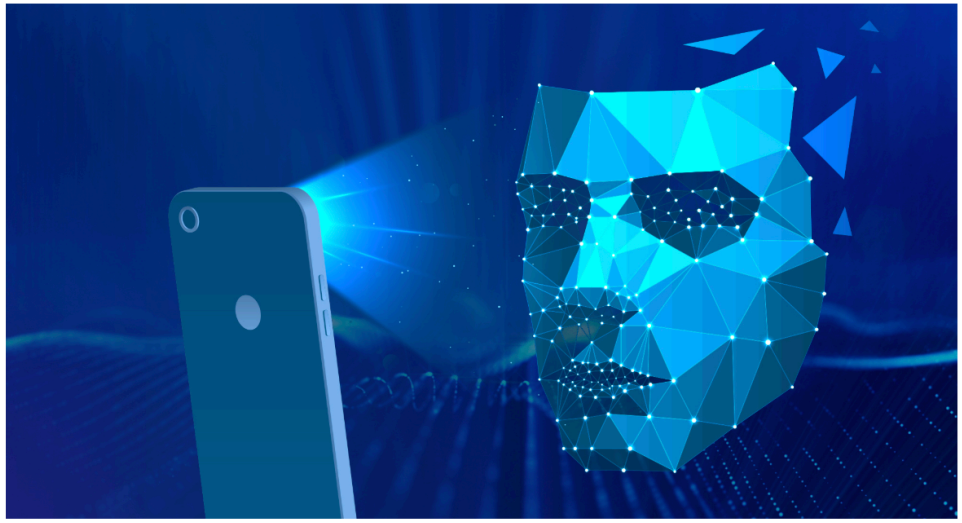
Schedule

- Class and labs: Monday/Tuesday, 3pm-6.30pm + Thursday 3-5pm (CEST).
- Recitation (Mandatory): Probably Tuesday 4pm - 6.30pm.
- Office hour : Thursday 5 - 6pm (CEST), to be confirmed.
- Location: Zoom !
- Combination between programming sessions (python) and lectures
- Final Exam: Midterm: 30%, Final : 30%
- Assignements throughout the semester: 30%
- Independent project: 10%

Course organization

- Notes + Sample exam questions can be found on the course webpage
- Sample exam questions = help you with the study but not comprehensive
- If a section of the notes is not covered in class, you don't have to study it for the exam
- See http://www.augustincosse.com/teaching for details (Scroll down and select "NYU Paris ML Summer School")

Introduction to Machine Learning (NYU Paris, Spring 2020)



NYU

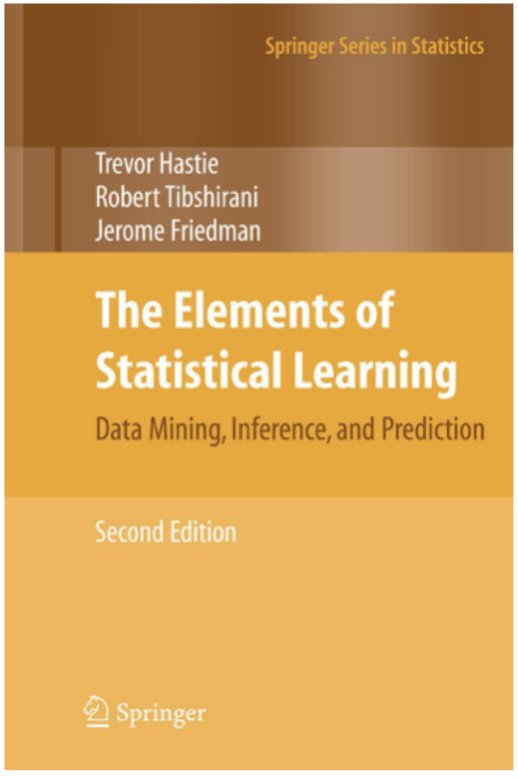
Tentative schedule:

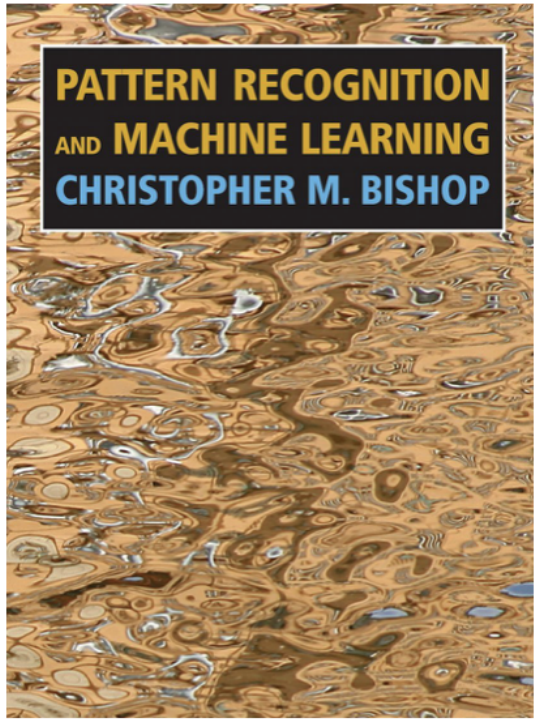
Legend: Lab sessions are in green, Homeworks are in red (right side of the table),

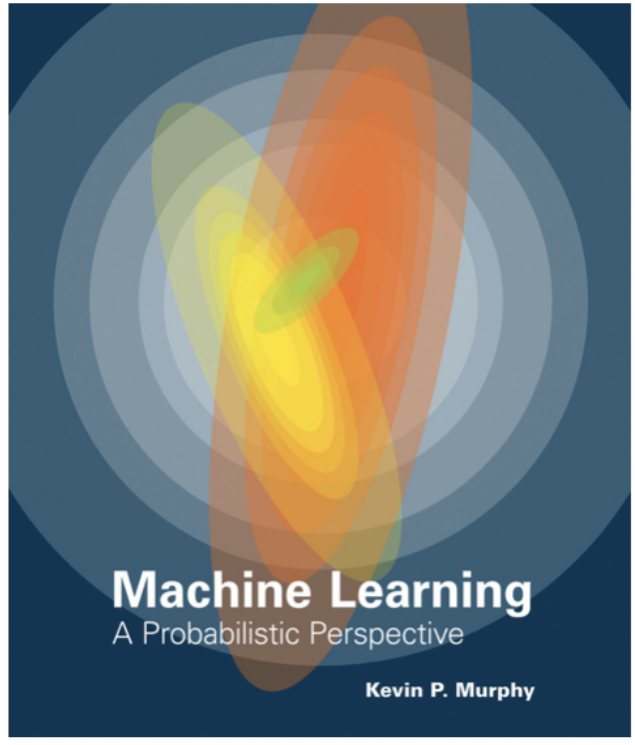
dates related to the project are in orange.

Week #	date	Topic
Week 1	02/03, 02/07	General Intro + reminders on proba and inference. Part I, Part II Lab 1
		Part I : supervised Learning

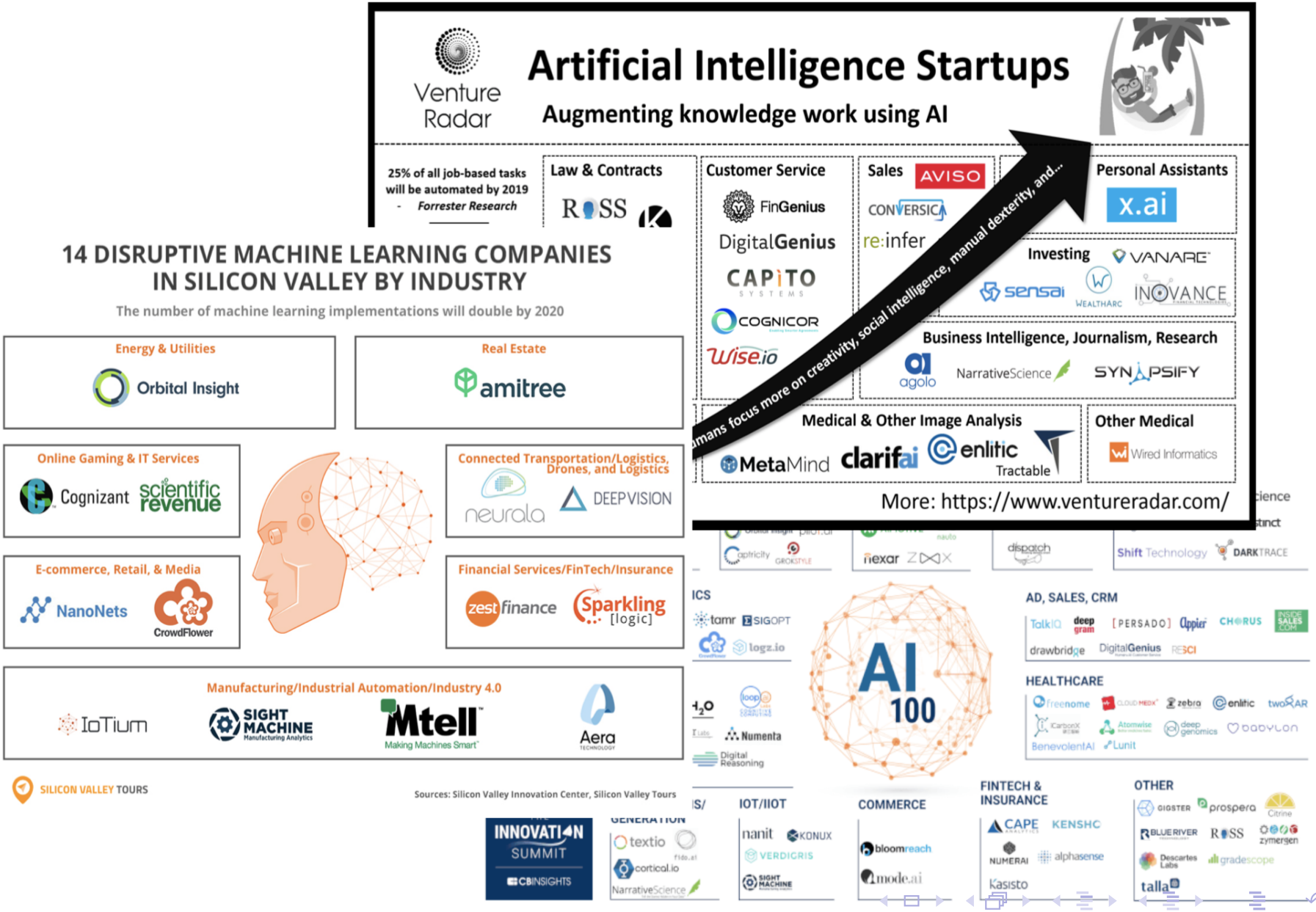
Reference Books







My objective: give you the tools to start working on your own ideas later



And connect you to the ML sphere in Paris

Class + programming sessions

- The objective is to provide you with all the theory/material needed to tackle some practical problems
- Then you will be given the opportunity to deal with those problems through Jupyter notebooks
- Many exercises rely on Scikit learn (free software machine learning library for Python, see https://scikit-learn.org/stable/) which features multiple learning algorithms.

General material (theory), see webpage

- Reminders in Stats/Probability, Inference.
- Supervised learning
  - Classification (Logistic regression, LDA)
  - Regression (Linear regression, regularization)
  - Neural Networks (+ implementation/Keras/TensorFlow)
  - Support Vector Machines
  - ...
- Unsupervised learning
  - Clustering (K-means, K-medoid, EM)
  - PCA, ICA
  - Manifold Learning
- Advanced topics (Adversarial and Reinforcement Learning)

What type of data are we going to use

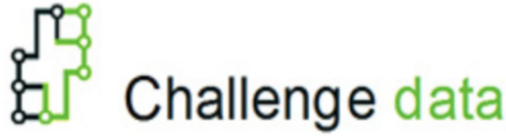
- Today data can be found everywhere
- Some major sources of data available online include
  - Kaggle (https://www.kaggle.com/)
  - UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html)
  - ENS Challenge Data (https://challengedata.ens.fr)
  - Many possibilities to get data directly from python (e.g. pandas, yahoo finance, scikit-learn...)


kaggle


NASA
Microsoft
MasterCard
kaggle


UCI
Machine Learning Repository


YAHOO! FINANCE
STOCK SCREENER


Challenge data


scikit
learn


Pandas

How do the programming sessions work?

- Two key components : Github and Jupyter notebooks
- Most efficient way to teach ML today = notebooks
  - Part of the code is given, remaining part has to be provided by the students
  - Many online examples that can be used directly or modified to generate new programming exercises
- When dealing with Neural nets (including GANs) so far I use TensorFlow/Keras (high-level API to build and train deep learning models)
- Exercises will be given through Github, see https://github.com/acosse/Intro2ML_Summer2021 (will be updated soon, Please fork the repo)

2.3 Convolutional Neural Networks for image classification

(adapted from R.Hon)

In this example we will learn how to train and use convolutional neural network for image recognition. Start by logging on Kaggle and download the MNIST Data : https://www.kaggle.com/c/digit-recognizer/data

Then store the content of the csv files using pandas in a train and a test variables

In [ ]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# put your code here
```

In this exercice, we will use Keras, which is a high level library which runs on top of Tensor Flow and enable fast experimentation with deep neural networks

In [ ]:
```
from keras.models import Sequential
from keras.layers import Dense, Dropout, Flatten
from keras.layers.convolutional import Conv2D,
from keras.utils import np_utils
from keras.optimizers import RMSprop
from keras.callbacks import ReduceLROnPlateau
from keras.preprocessing.image import ImageDat
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test
```

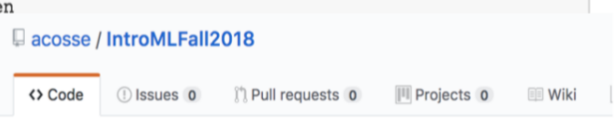
Using the tools form the pandas library, look at your data. W
the first column of each pandas data array contain ?

2.3.1 Displaying the images

Display the first few images from the MNIST dataset using ims

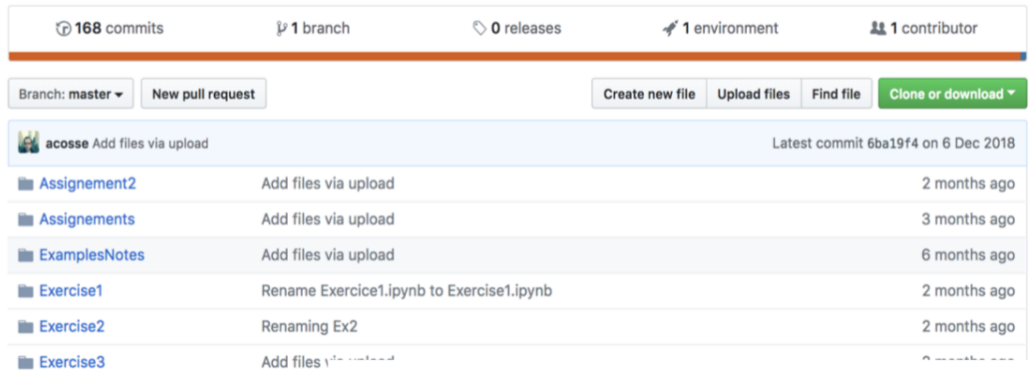
In [ ]:
```
# put your code here
```

Notebook



website for the Fall 2018 NYU Class Introduction to Machine Learning

Manage topics



GitHub

A group project

- A key aspect of the course is the project
- The idea of the project is to give you the opportunity to develop your own ideas (insofar as possible)
- A suggestion of possible topics can be found online. However you are free to come up with new ideas
- View this as a first step towards a startup. If the project works well, you can perhaps extend it later !

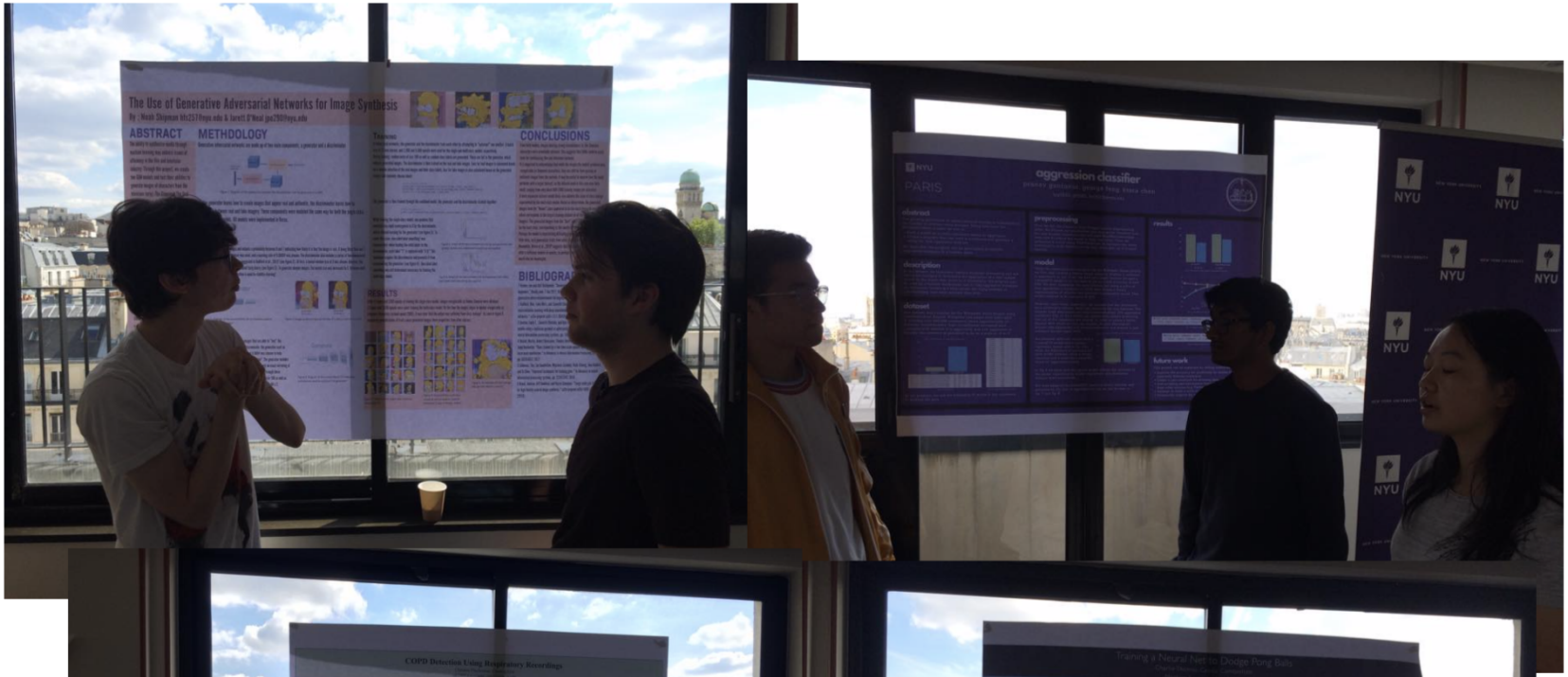
A group project

- You choose one subject in group and start working on it throughout the semester
- The project has to (1) be related to at least one algorithm studied in class and (2) has to exhibit some programming aspects.
- You are free to come during OH with questions on the project
- You then get to present your work at the end of the semester

A group project

Some examples from last year:

- Training Tic Tac Toe
- Training a neural network to win at a pong game
- Using Natural Language processing to detect and remove aggressive comments from Twitter
- Detecting health conditions from medical data
- Use generative adversarial networks to generate cartoons
- ...


The Use of Generative Adversarial Networks for Image Synthesis
ABSTRACT
METHODOLOGY
CONCLUSIONS
BIBLIOGRA
RESULTS

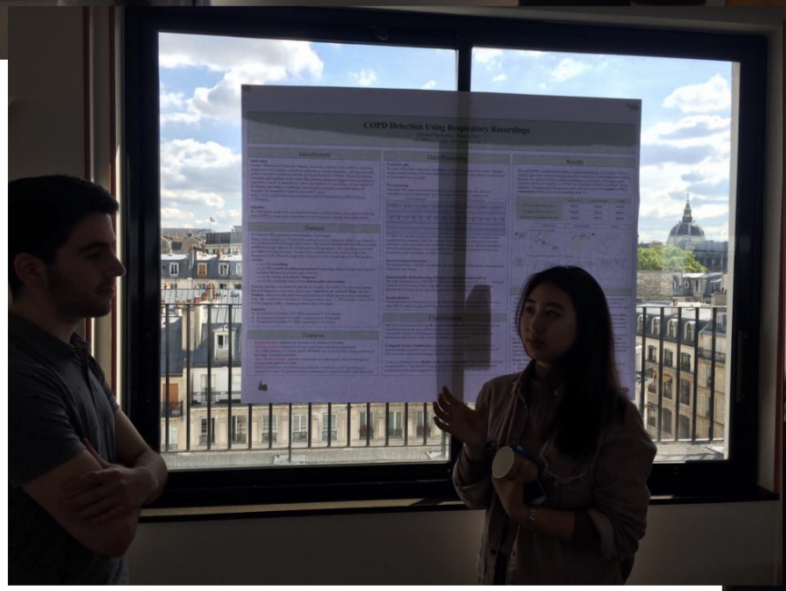
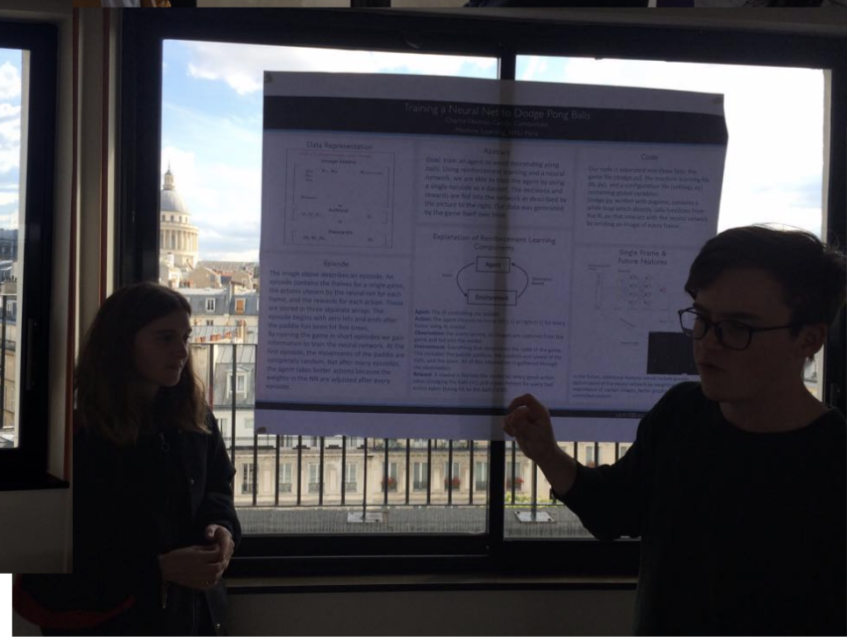
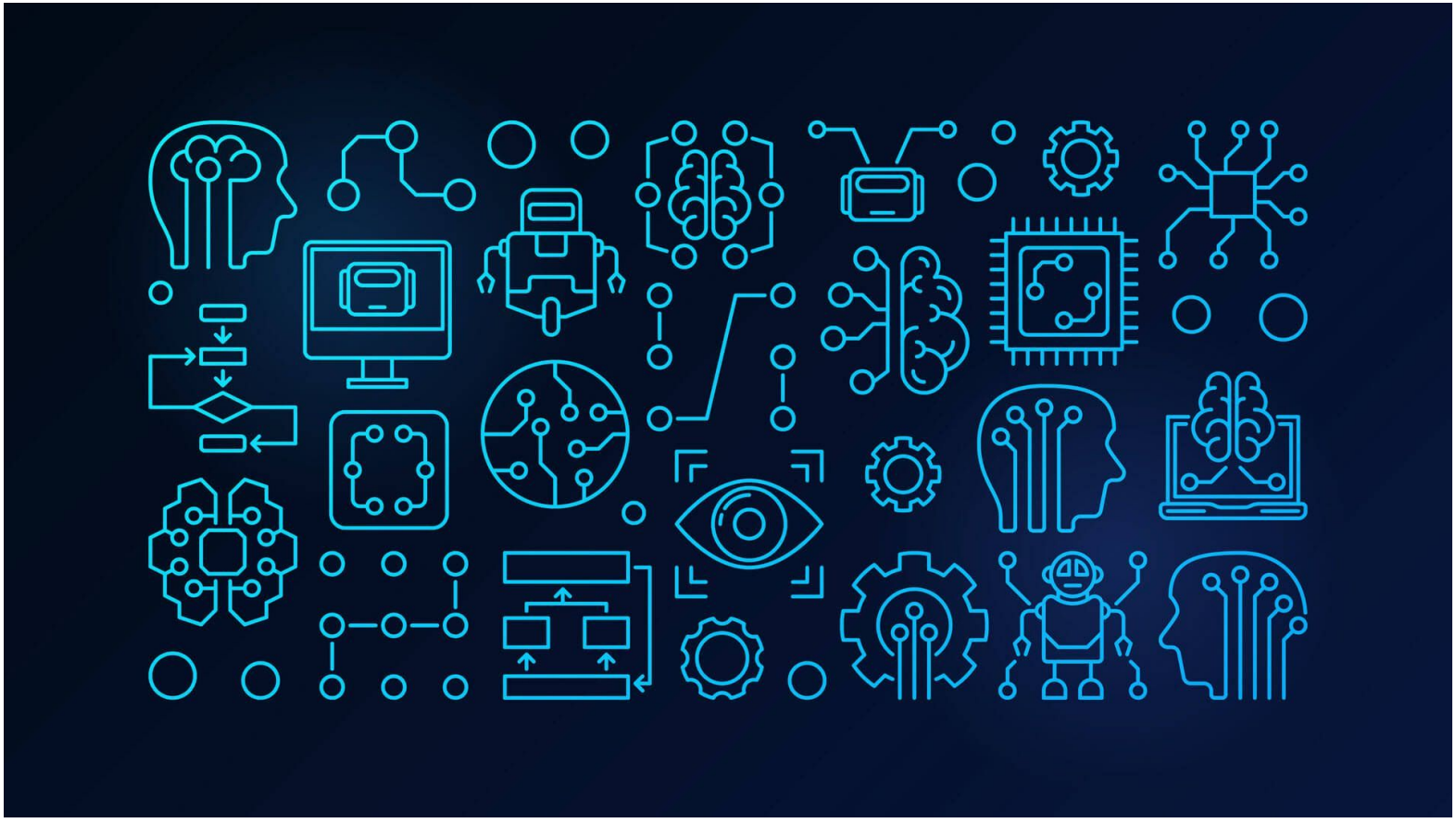
aggression classifier
PARIS
NYU

COPD Detection Using Respiratory Recordings

Training a Neural Net to Dodge Pong Balls
Data Representation
Explanation of Reinforcement Learning Concepts
Episode
Code
Single Frame & Future Features

Table of contents

Machine learning today



Machine learning today

BBC NEWS

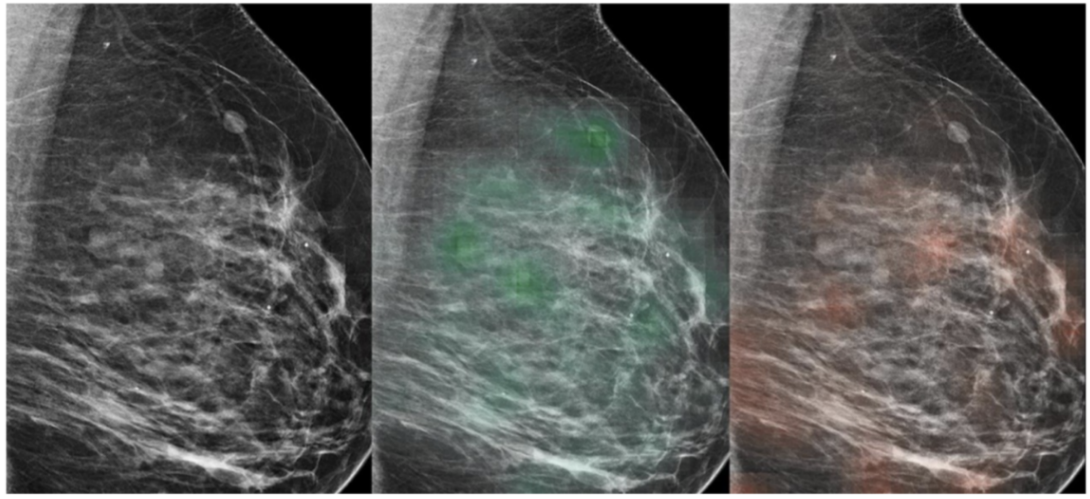
Artificial intelligence-created medicine to be used on humans for first time

By Jane Wakefield
Technology reporter

The drug was much quicker to market than ones developed in more traditi

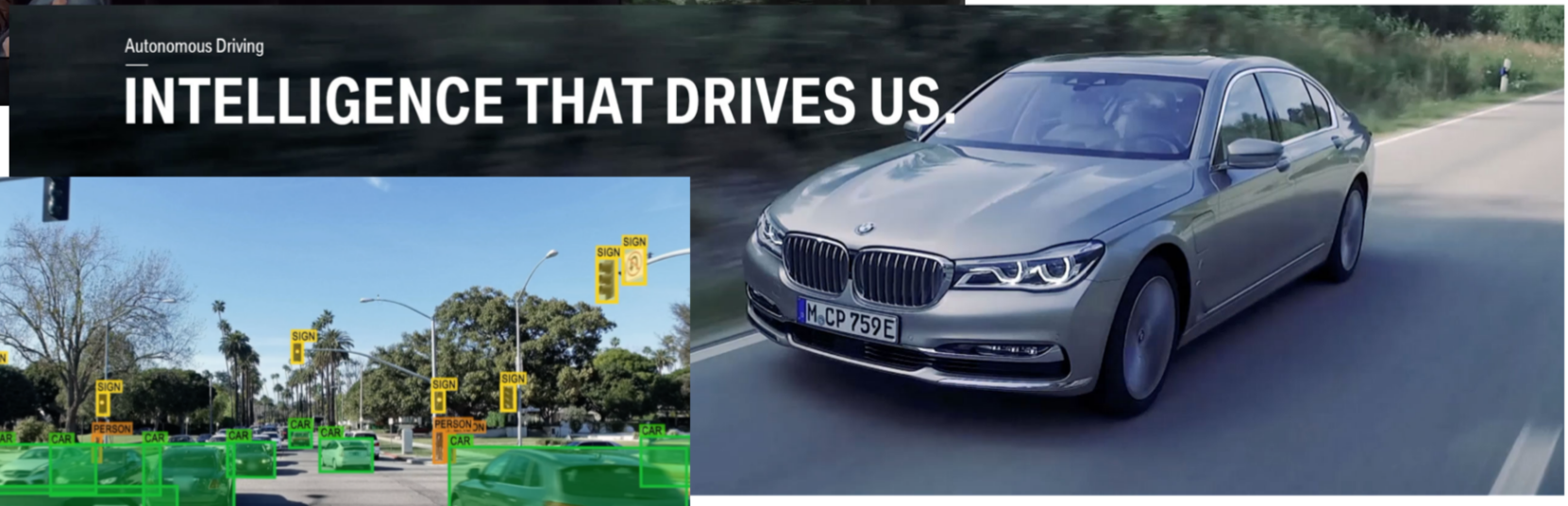
A drug molecule "invented" by artificial intellig human trials in a world first for machine learni

This AI tool helps identifies breast cancer with 90% accuracy rate



The AI tool identifies malignant lesions (red) and benign lesions (green).

Image: NYU School of Medicine

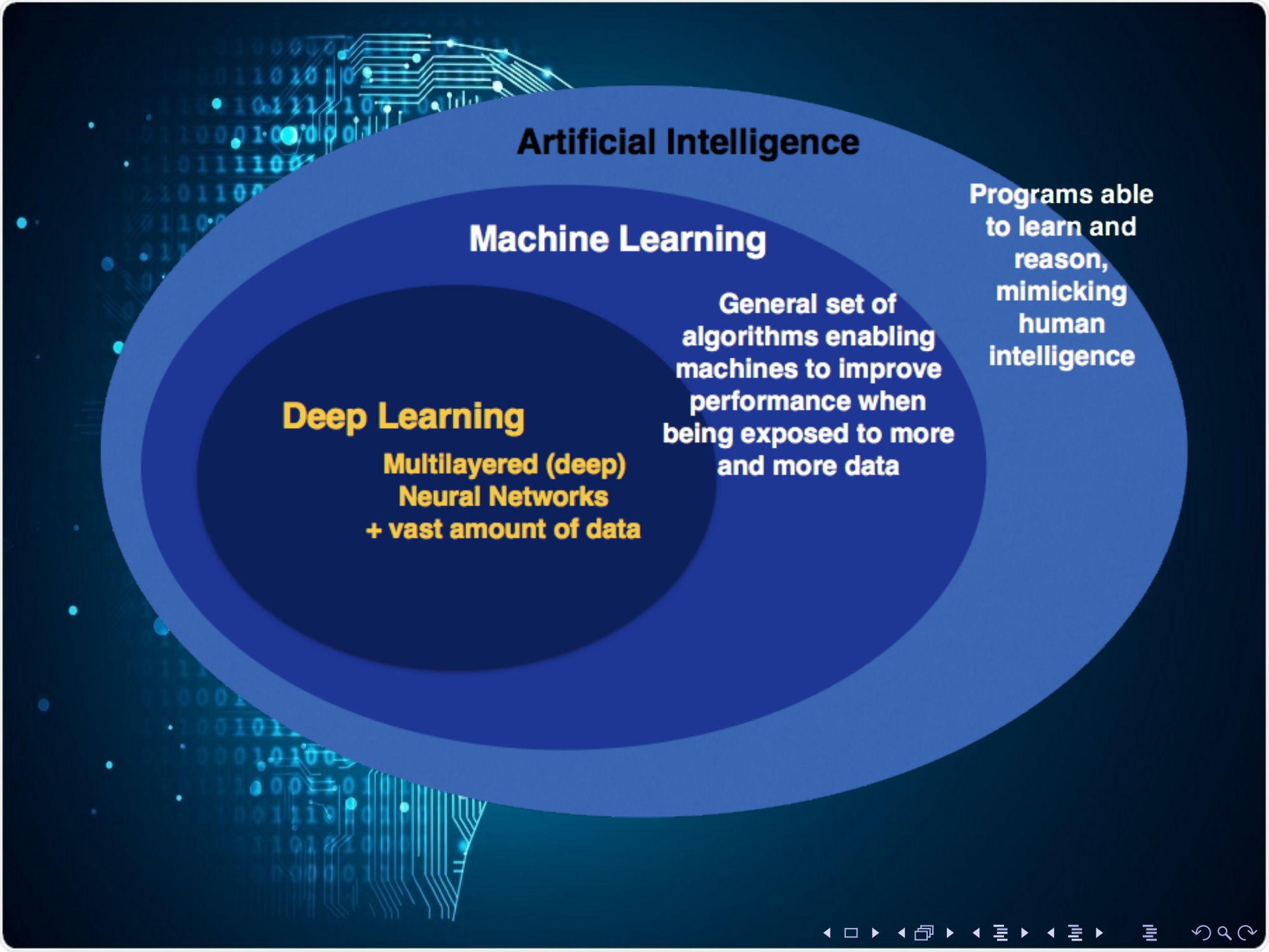


Autonomous truck carts 40,000 lb of butter coast to coast in the US

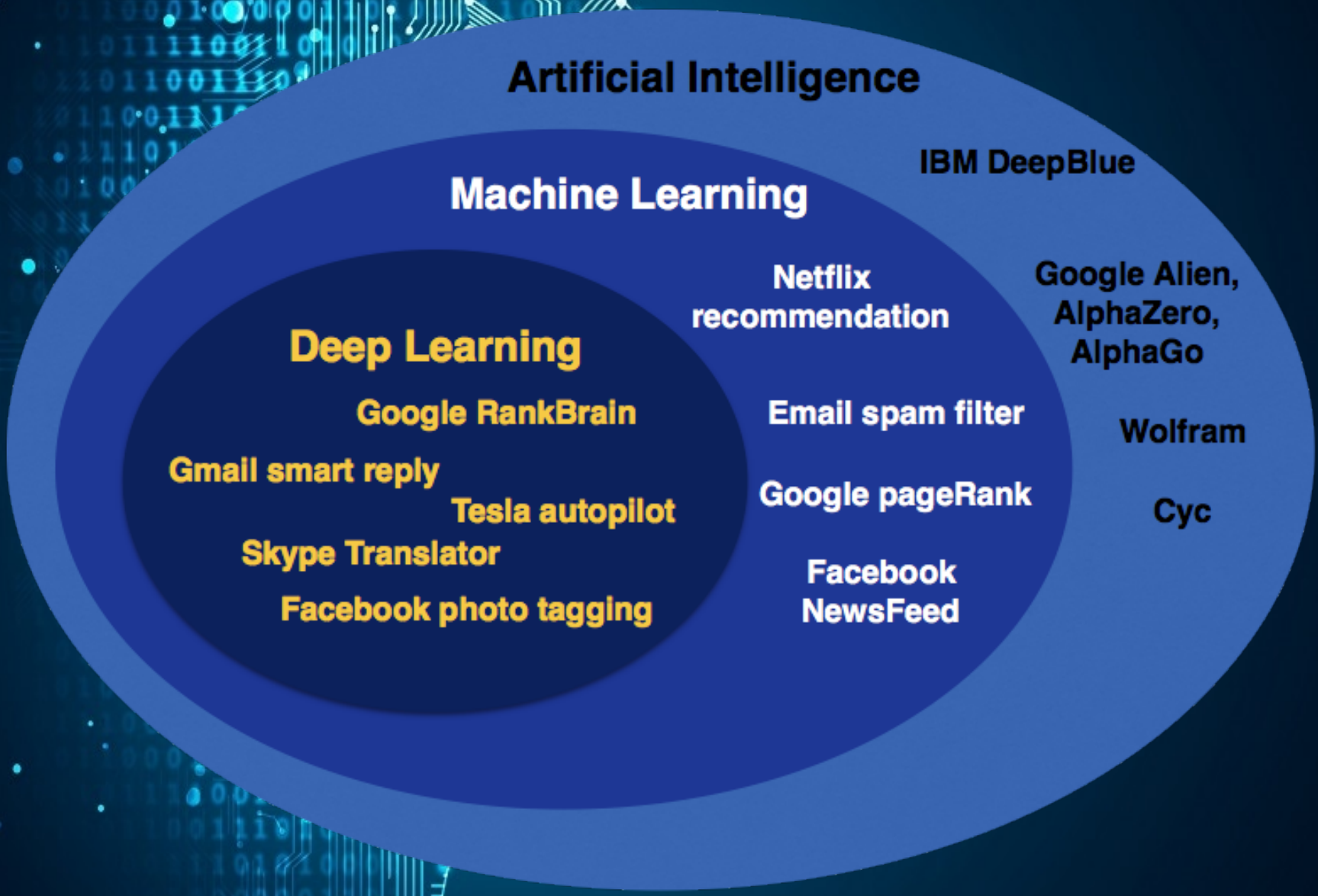
Bloomberg Businessweek

UDACITY

Tesla's Autopilot Could Save the Lives of Millions, But It Will Kill Some People First


Artificial Intelligence
Programs able to learn and reason, mimicking human intelligence
Machine Learning
General set of algorithms enabling machines to improve performance when being exposed to more and more data
Deep Learning
Multilayered (deep) Neural Networks
+ vast amount of data

Some achievements



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW OF PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

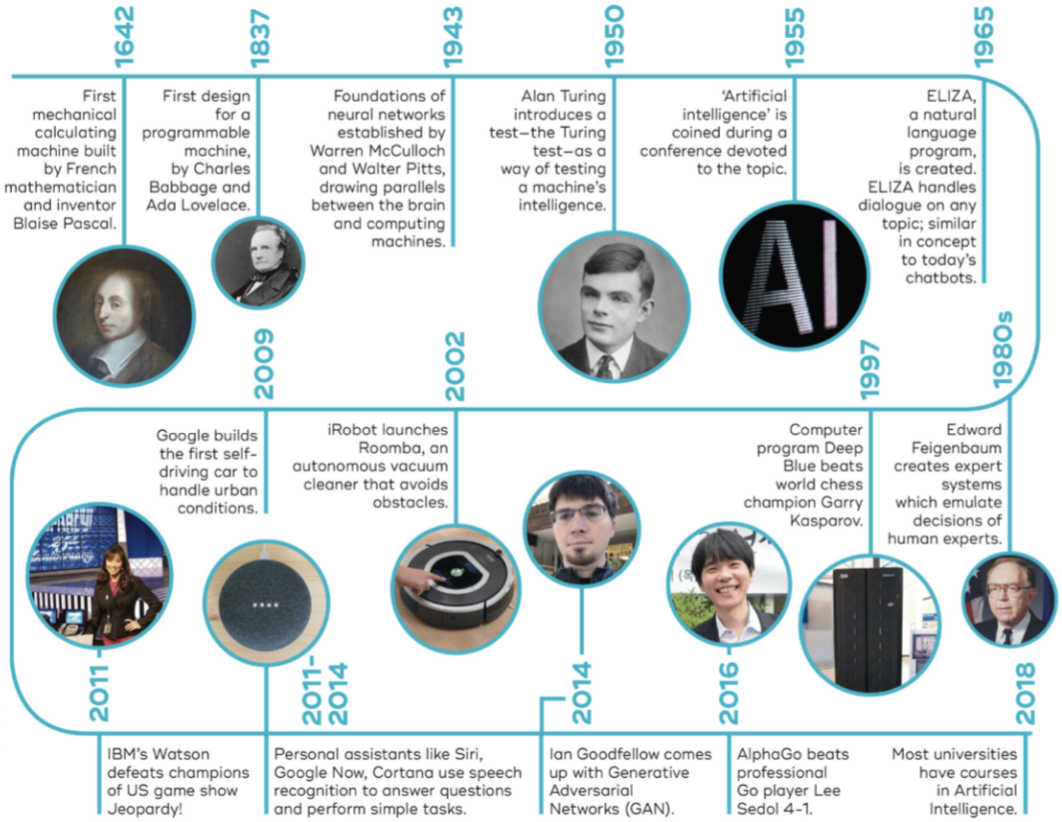
1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's



Lovelace essentially envisioned what computers were capable of before computers really existed.

McKinsey Global Institute

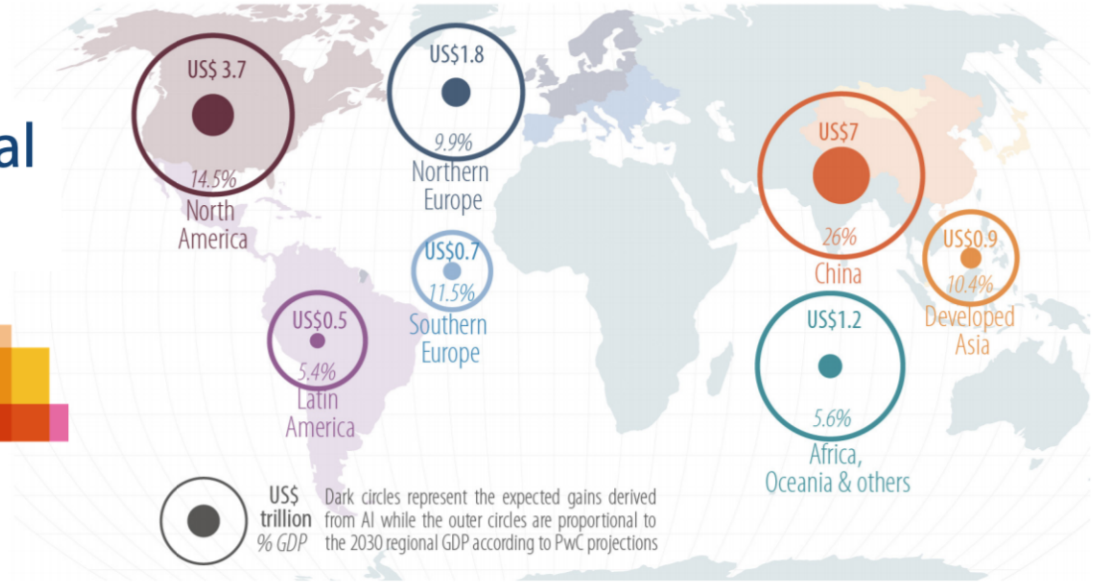
Notes from the AI frontier: Modeling the impact of AI on the world economy

September 2018 | Discussion Paper

Economic impacts of artificial intelligence (AI)

pwc

Figure 2 – Expected gains from AI in the different regions of the world by 2030



Source: The macroeconomic impact of artificial intelligence, PwC, 2018.

PwC's Global Artificial Intelligence Study: Sizing the prize

Healthcare

Three areas with the biggest AI potential

- Supporting diagnosis in areas such as detecting small variations from the baseline in patients' health data or comparison with similar patients.
- Early identification of potential pandemics and tracking incidence of the disease to help prevent and contain its spread.
- Imaging diagnostics (radiology, pathology).

Technology, communications and entertainment

Three areas with the biggest AI potential

- Media archiving and search – bringing together diffuse content for recommendation.
- Customised content creation (marketing, film, music, etc.).
- Personalised marketing and advertising.

Automotive

Three areas with the biggest AI potential

- Autonomous fleets for ride sharing.
- Semi-autonomous features such as driver assist.
- Engine monitoring and predictive, autonomous maintenance.

Intelligent Machines

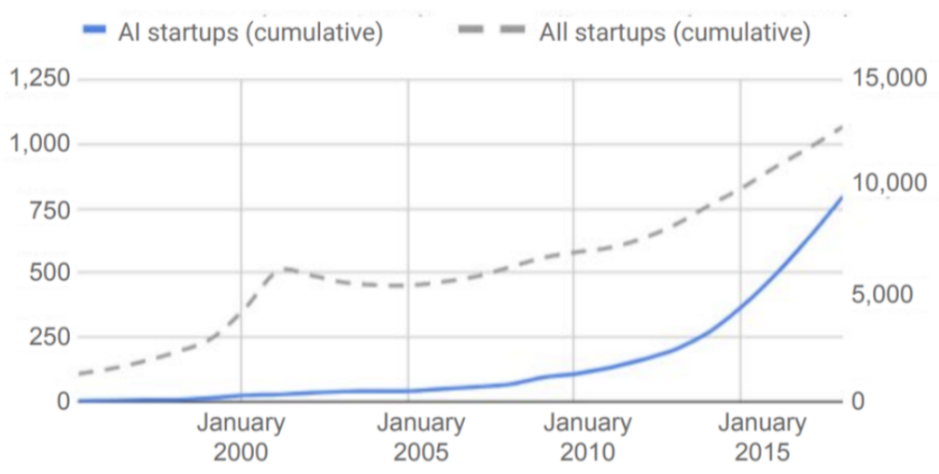
MIT Technology Review

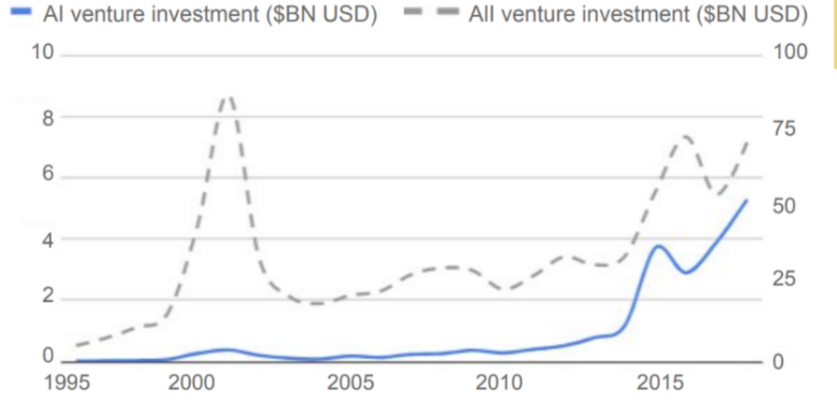
Nine charts that really bring home just how fast AI is growing

Artificial intelligence is booming in Europe, Chi but it's still a very male industry.

by Will Knight December 12, 2018

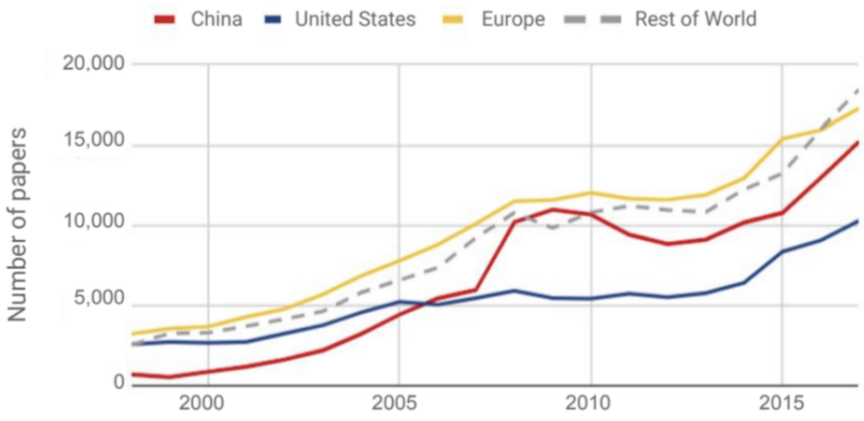
1. AI is being commercialized at a dizzying pace.



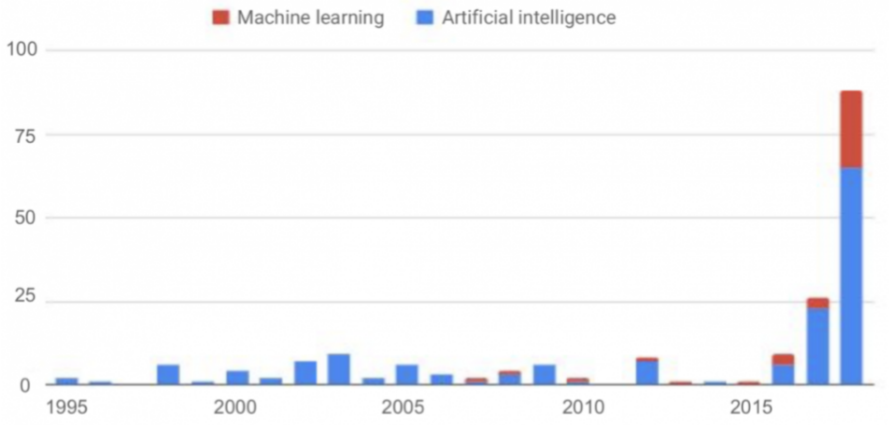


2. The focal points are China and the US, but also Europe.

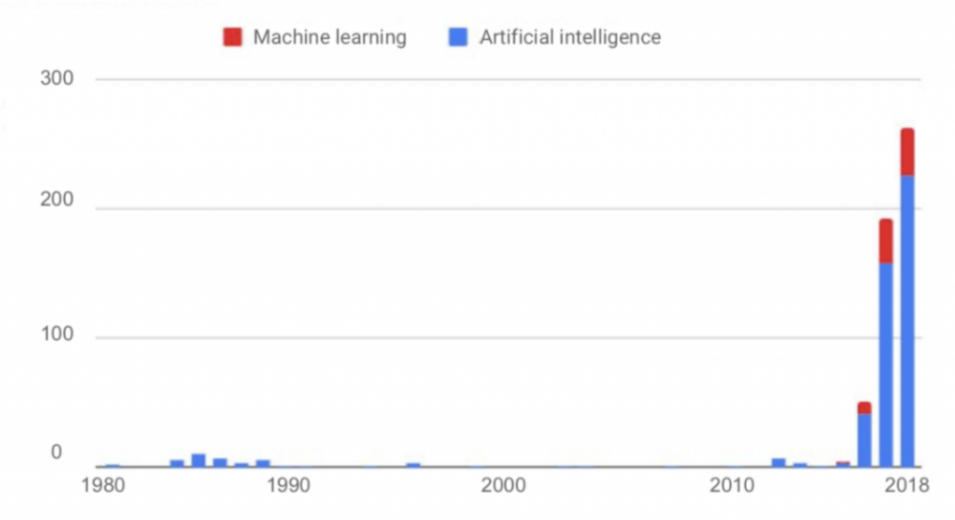
Much has been made of China's rising AI prowess (see "China's AI awakening") and its growing rivalry with the US. As the data shows, Europe is also a huge hub of AI activity. But it seems that three main centers of power are emerging.



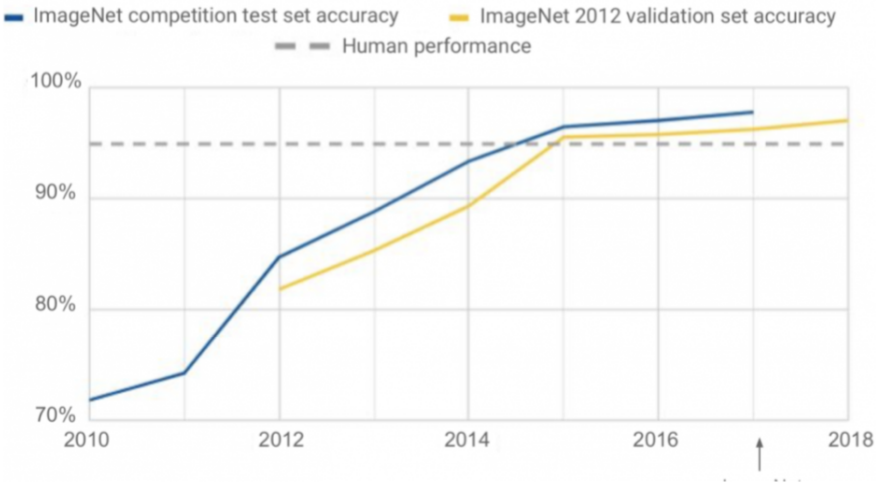
5. Artificial intelligence is a political issue.



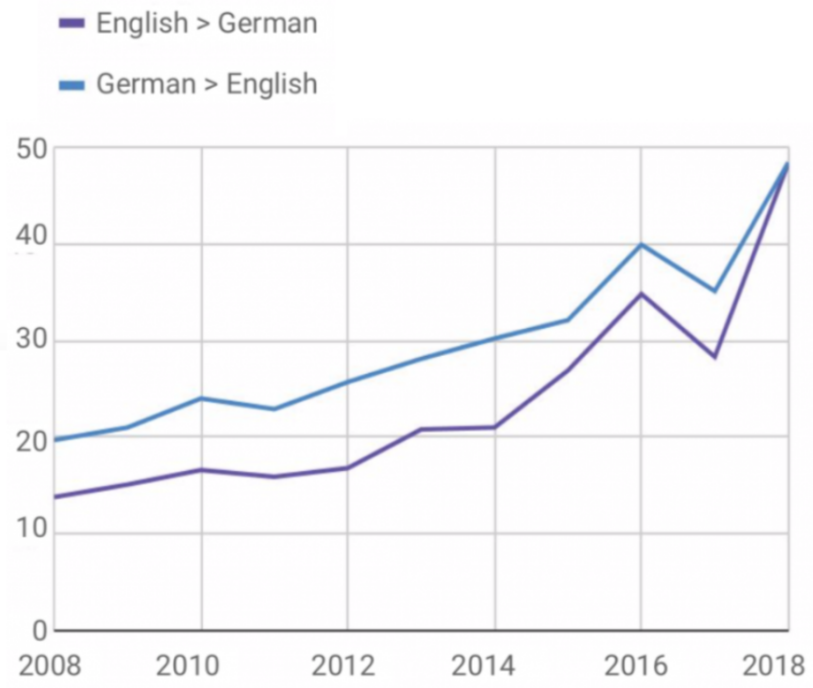
Mentions of artificial intelligence and machine learning in the US Congress (above) and the UK Parliament (below) have exploded in the past few years. This reflects a growing awareness of the technology's economic and strategic importance (see "Canada and France propose an international panel on AI").



4. The state of the art is improving fast.



The report includes several measures of technical progress, including the accuracy of object recognition in images, measured against average human performance (top), and the accuracy of machine translations of news articles, measured using a score assigned by human judges (bottom). These don't mean that the field is getting closer to developing a human-level AI, but they show how key techniques have been honed in recent years.



MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2018

10 Breakthrough Technologies 2017

These technologies all have staying power. They will affect the economy and our politics, improve medicine, or influence our culture. Some are unfolding now; others will take a decade or more to develop. But you should know about all of them right now.

Dueling Neural Networks

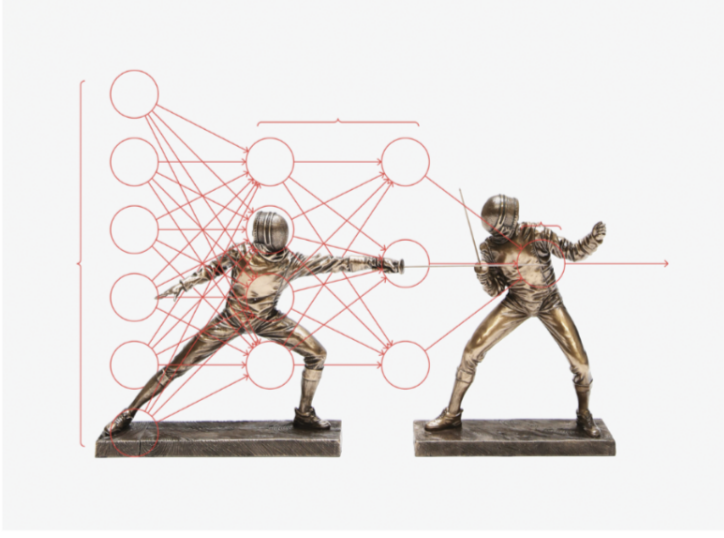


ILLUSTRATION BY DEREK BRAHNEY | DIAGRAM COURTESY OF MICHAEL NIELSEN, "NEURAL NETWORKS AND DEEP LEARNING", DETERMINATION PRESS, 2015

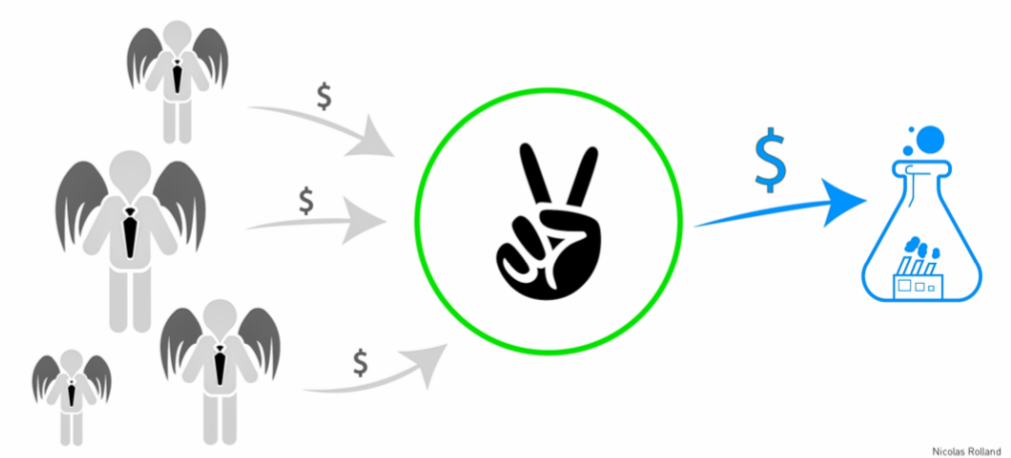
Reinforcement Learning

By experimenting, computers are figuring out how to do things that no programmer could teach them.

Availability: 1 to 2 years

by Will Knight

The Rise of the ‘Unicorns’



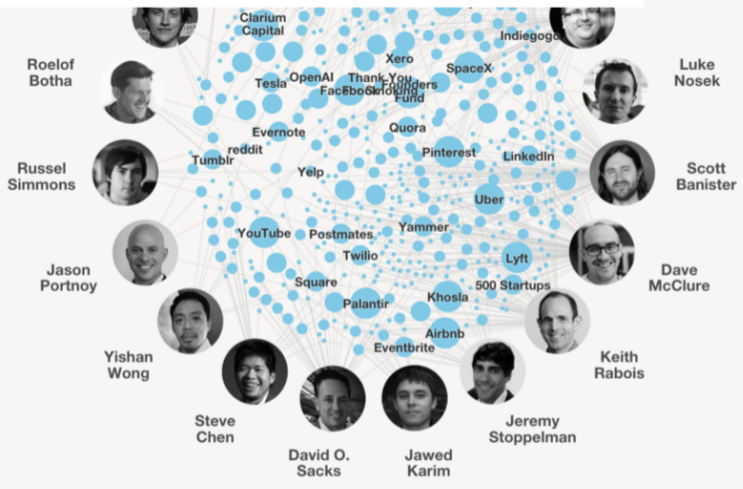
Nicolas Rolland

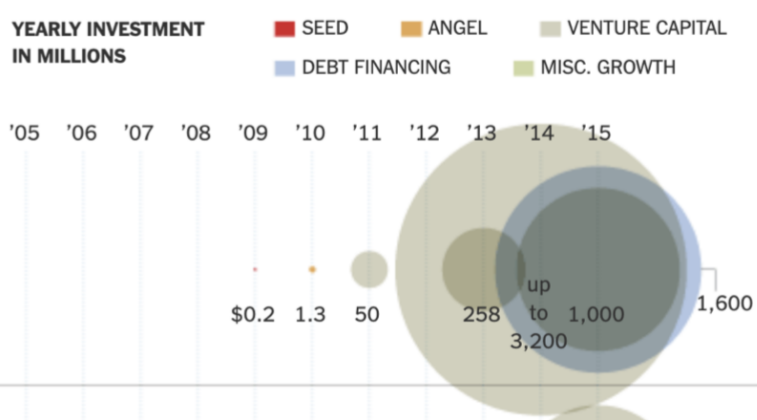
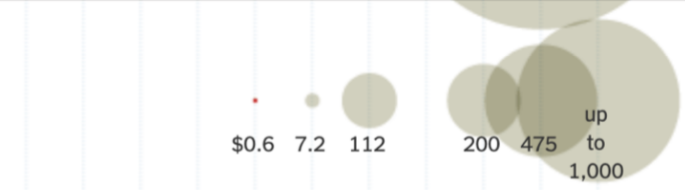
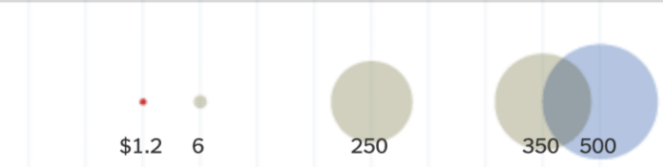
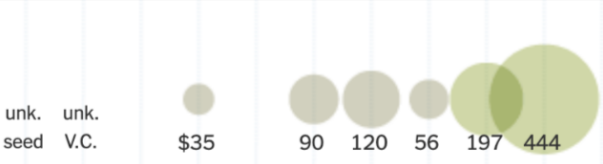
The PayPal Mafia’s Golden Touch

Meet the Dons who run Silicon Valley – the PayPal Mafia.

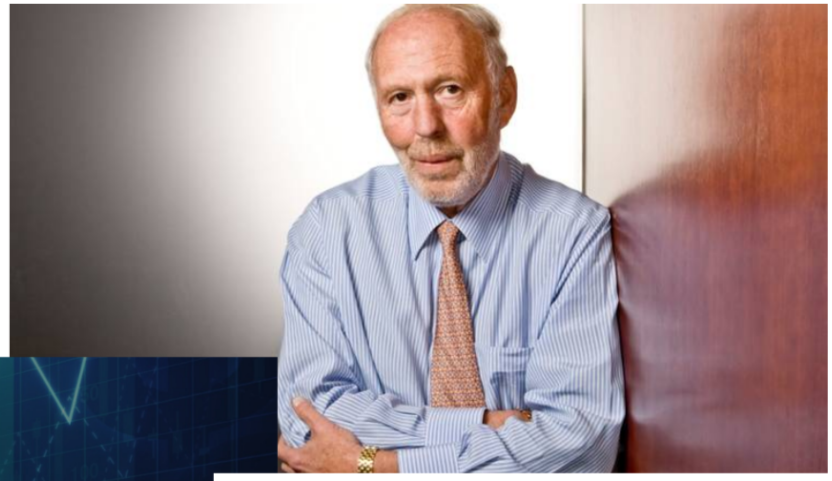
Despite the name, this ‘outfit’ is an entirely legit group of tech entrepreneurs - the founders and former employees of PayPal. The ‘Mafia’ have continued to work and invest together since the company’s 2002 sale to eBay. Between them, they’ve founded, funded or led some of the world’s biggest tech firms.

Household names



COMPANY	MOST RECENT VALUATION	YEARLY INVESTMENT IN MILLIONS	TOTAL AMOUNT RAISED
Uber Uber Technologies' app connects drivers to riders in 55 countries, with pricing that varies with demand.	**$40 billion**		**$6.3 billion**
Airbnb A website that allows users to book apartments, homes and other temporary accommodations.	**$20 billion**		**$1.8 billion**
Snapchat Its messaging app lets users take and share images and videos that disappear once they are viewed.	**$15 billion**		**$809 million**
Dropbox A service that lets users upload and share photos, documents and videos with specific people.	**$10 billion**		**$1.1 billion**
Palantir Palantir Technologies specializes in software used in counterterrorism and the financial industry.	**$15 billion**		**$942 million**

Renaissance



2σ

AI in Finance

D E Shaw & Co

Rich Formula: Math And Computer Wizards Now Billionaires Thanks To Quant Trading Secrets

Nathan Vardi Forbes Staff
Hedge Funds & Private Equity
Following the money trail

By then, quants and others in various industries had embraced Simons' machine learning techniques. Indeed, he and his colleagues at Renaissance had anticipated the transformation in decision-making that's sweeping almost every business and walk of life. More companies and individuals are accepting and embracing models that continuously learn from their successes and failures. As investor Matthew Granade has noted, Amazon, Tencent, Netflix, and others that rely on dynamic, ever-changing models are emerging dominant. The more data that's fed into machines, the smarter they're supposed to become. Just as Jim Simons predicted.

Markets

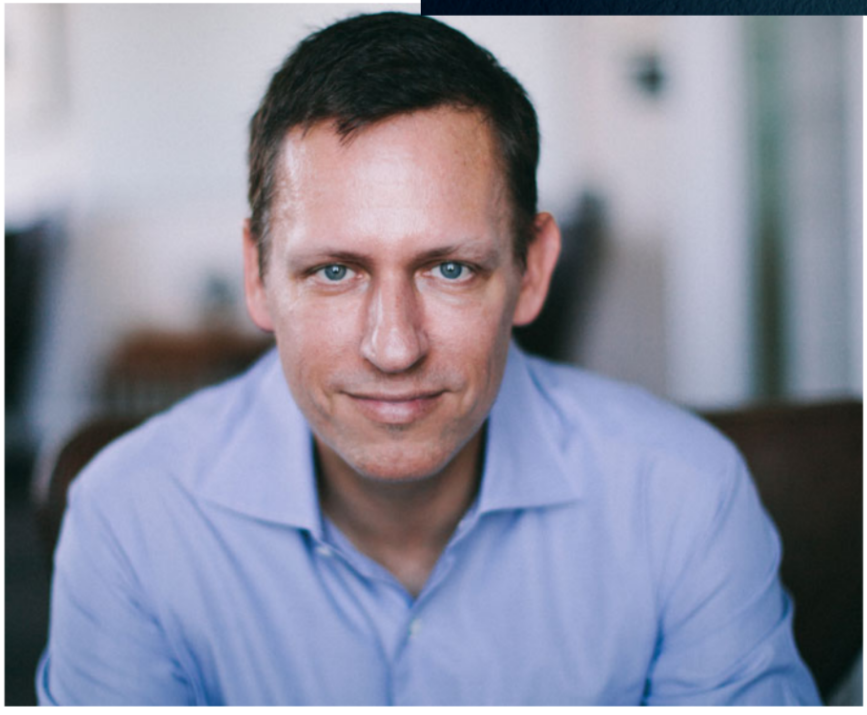
Two Sigma Hires Google Scientist Mike Schuster for AI Expansion

March of the machines

The stockmarket is now run by computers, algorithms and passive managers

Such a development raises questions about the function of markets, how companies are governed and financial stability


Palantir



THE TIMES

Palantir, the tech spooks who found bin Laden

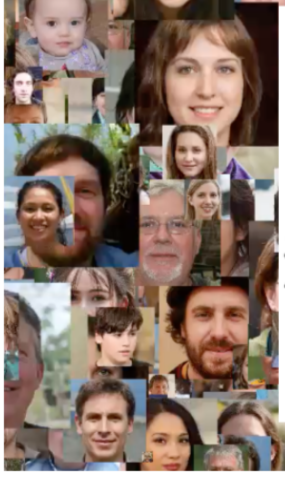
Tech Policy Jan 27

40 groups have called for a US moratorium on facial recognition technology

It's too late to ban face recognition – here's what we need instead

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.



Clearview

In The News

NEW YORK POST

Facial-recognition tech is a win for the little guy

"For all the hysteria over facial-recognition technology, the breathtakingly fast arrest of Larry Griffin II for prompting a major subway panic shows that the tech is a boon to law enforcement."

The New York Times

Video Games and Online Chats Are 'Hunting Grounds' for Sexual Predators

"And so law enforcement officials from across the state took over a building near the Jersey Shore last year, and started chatting under assumed identities as children. In less than a week, they arrested 24 people."

THE WALL STREET JOURNAL.

Have No Fear of Facial Recognition

"In another disturbing case, with explicit video, someone sexually harassed and extorted young girls. Police used facial recognition to identify 14 of 22 victims, who were carefully interviewed. They eventually identified the offender."

Available now for Law Enforcement

Request Access

Stop Searching. Start Solving.

Clearview provides clients with its proprietary technology, database and investigative tools on a subscription basis. A Licensed User's subscription includes:

- ✓ Unlimited Use of CV's Proprietary Research System for its Licensed Users.
- ✓ Unlimited Access to CV's Proprietary Image Database for its Licensed Users.
- ✓ Each Licensed User Account Includes iPhone/Android CV Application
- ✓ Each Licensed User Account Includes Lap/Desktop Versions of CV Program
- ✓ Help-Desk Support

Annual 12-month Subscription Rates 5 Seats: $10,000
10 Seats: $15,000
20 Seats: $25,000
50 Seats: $50,000
125 Seats: $100,000
500 Seats: $250,000
Unlimited License (Unlimited Users): Negotiated Flat Fee

For More Information: Jessica Medeiros Garrison
(e) Jessica@clearview.ai (c) 205.568.4371

Time is law enforcement's most valuable resource. Clearview puts the world's most advanced facial-recognition technology and largest image database into their hands, allowing them to turn a photograph into a solid lead in an instant.

Clearview Ai

Tax ID: 82-2397610 Clearview Ai, Inc.
15 West 72nd St. Suite 23-S New York, NY 10023

Our office spent 12 man hours over a month's time trying to identify a theft suspect. We ran the picture through Clearview and identified the suspect in seconds.

POLICE

If we had Clearview at the time when the report came in, we would not only have identified the suspect sooner, but also would have prevented other thefts that the suspect committed before we arrested him.

John Hodgens
Police Officer

World's best facial-recognition technology combined with the world's largest database of headshots.

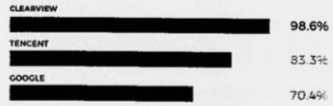
Real-time Results.

On September 24, 2018, The Gothamist published a photo of a man who assaulted two individuals outside a bar in Brooklyn, NY.

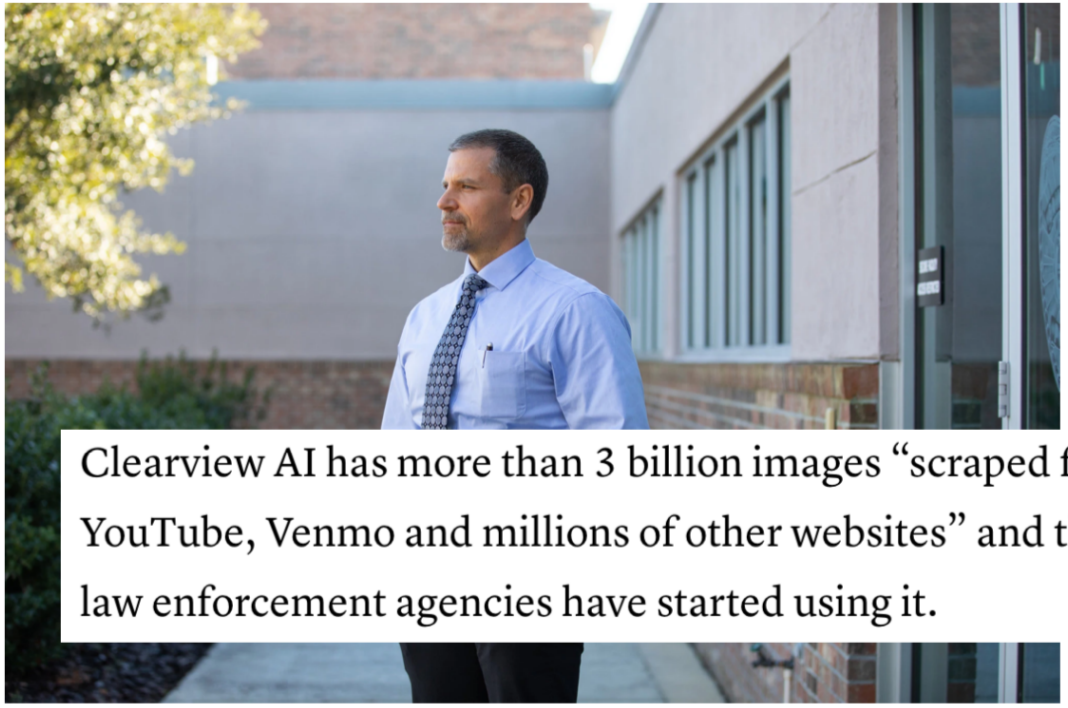
World-Class Accuracy.

Clearview consistently ranks ahead of the world's leading providers in accuracy (Source: NegaFace)

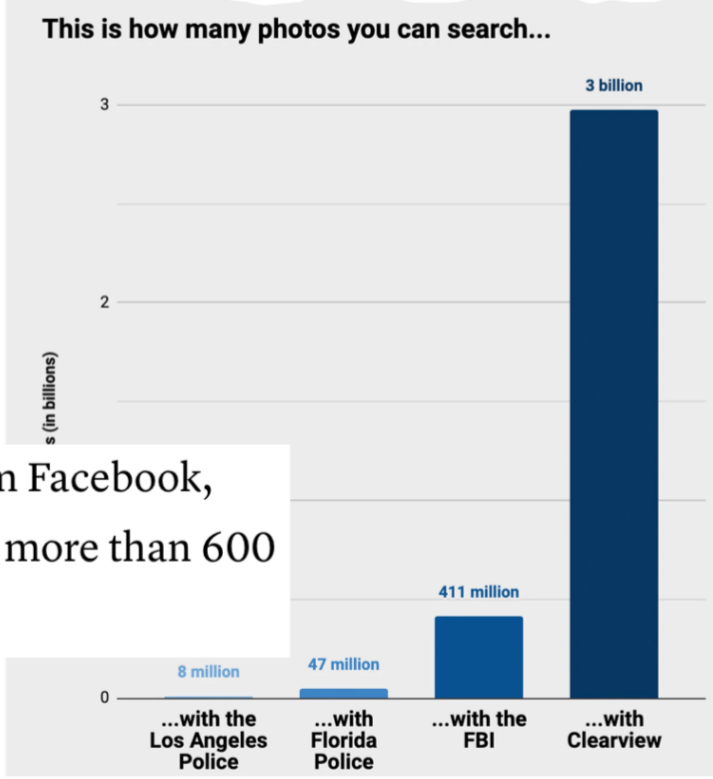
Accuracy finding a match out of 1 Million faces:



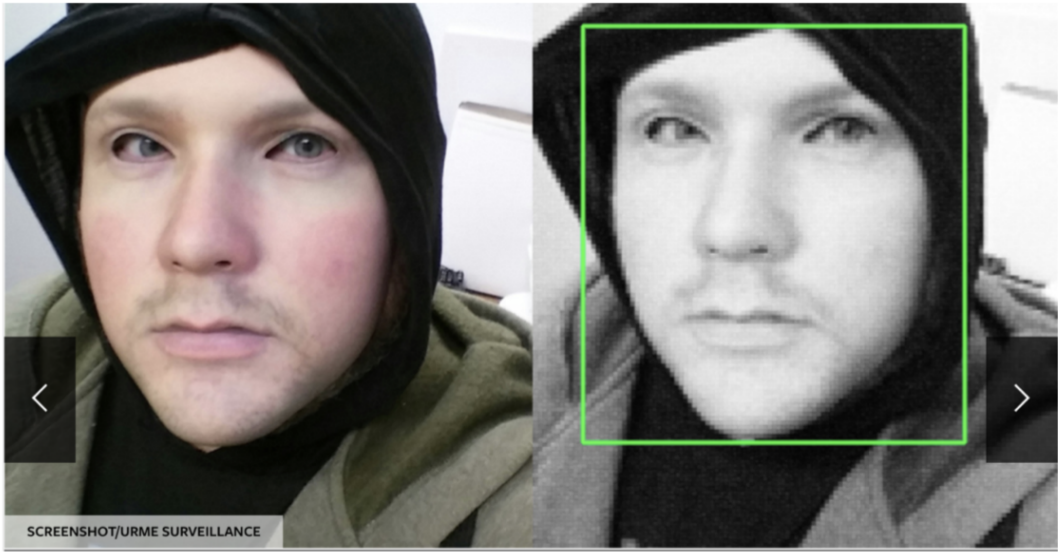
Clearview Ai



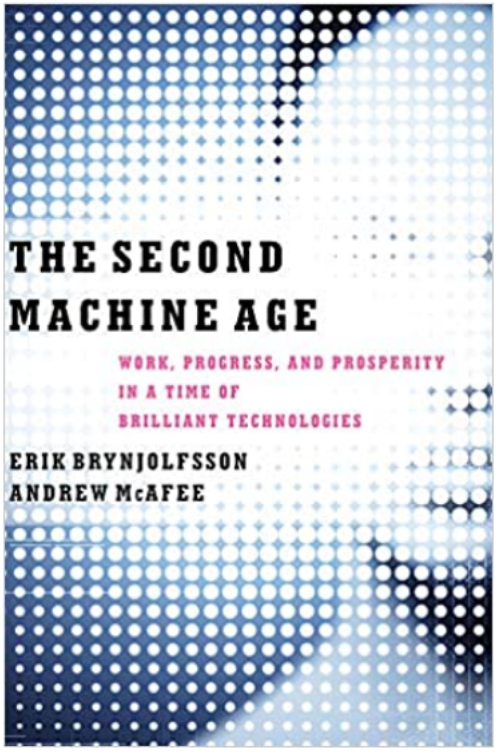
Clearview AI has more than 3 billion images “scraped from Facebook, YouTube, Venmo and millions of other websites” and that more than 600 law enforcement agencies have started using it.



Detective Sgt. Nick Ferrara in Gainesville, Fla., said he had used Clearview’s app to identify dozens of suspects. Charlotte Kesl for The New York Times



The patterns on the goods in this shop are designed to trigger Automated License Plate Readers, injecting junk data in to the systems used by the State and its contractors to monitor and track civilians and their locations.

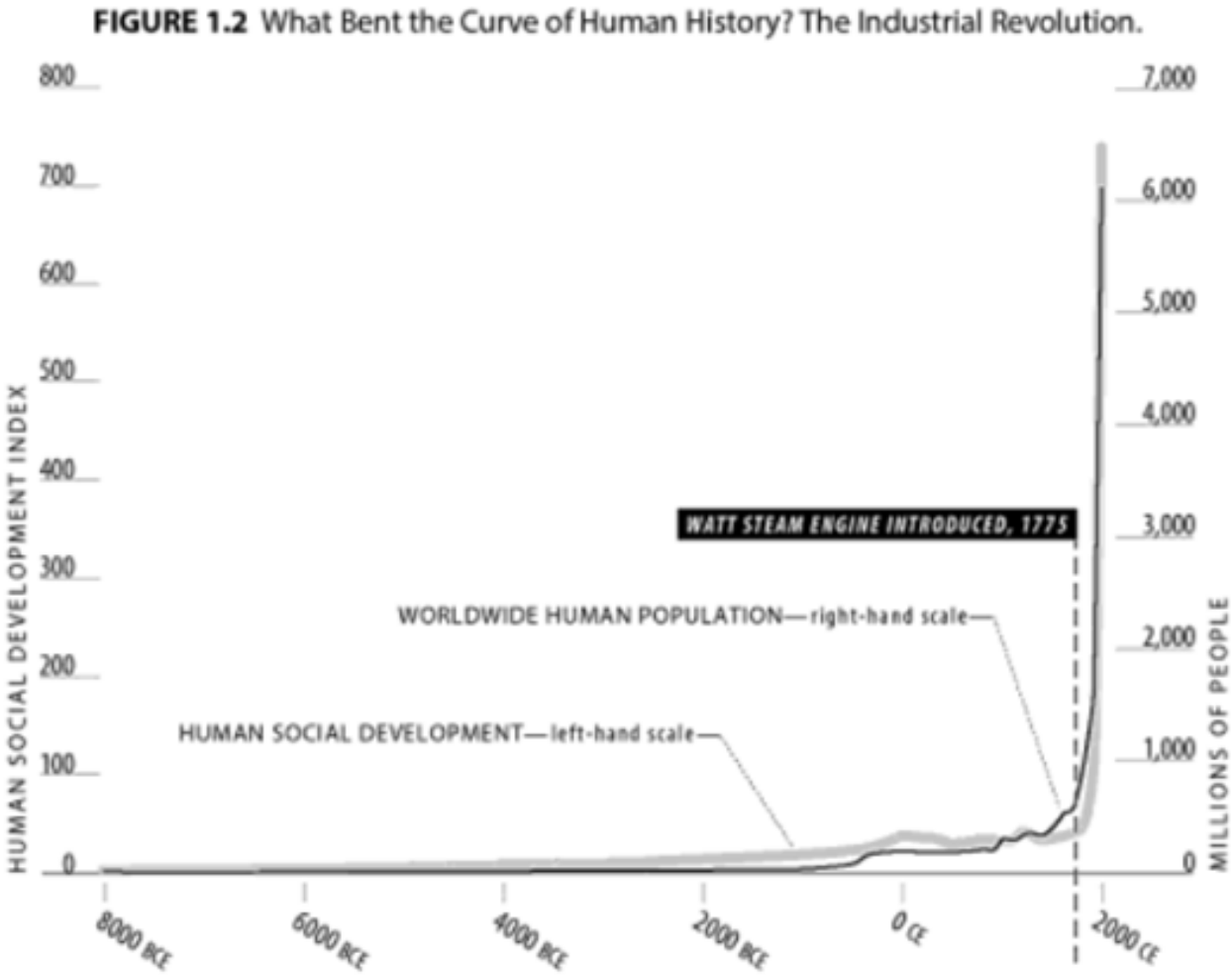


"Artificial intelligence and the core technology of machine learning are likely to be a general purpose technology of the scale and scope similar to electricity or even the steam engine before it - fundamentally revolutionizing many, many sectors of the economy."

ERIK BRYNJOLFSSON
DIRECTOR OF THE MIT INITIATIVE ON THE DIGITAL ECONOMY

Perhaps as important as steam engines back in the days..

Erik Brynjolfsson, ICLR 2018



Back to the industrial revolution

from Erik Brynjolfsson, ICLR 2018

- What can history tell us?
- Steam engine was classified by Bresnahan et Trajtenberg (1996) as belonging to the so-called General Purpose Technologies
- Those technologies are characterized by 3 features:
  - Pervasive
  - Able to be improved over time
  - Able to spawn complemetary innovations
- Does that remind you of something ?

Back to the industrial revolution

from Erik Brynjolfsson, ICLR 2018

- Technology is not neutral
- You can do a small pox vaccine...

Back to the industrial revolution

from Erik Brynjolfsson, ICLR 2018

- Technology is not neutral
- You can do a small pox vaccine...But you can also create a nuclear weapon
- In fact let's compare..

BUSINESS

Is 'Progress' Good for Humanity?

Rethinking the narrative of economic development, with sustainability in mind

JEREMY CARADONNA SEP 9, 2014

Rage against the machine: Luddites smashing a loom. (CHRIS SUNDE / WI

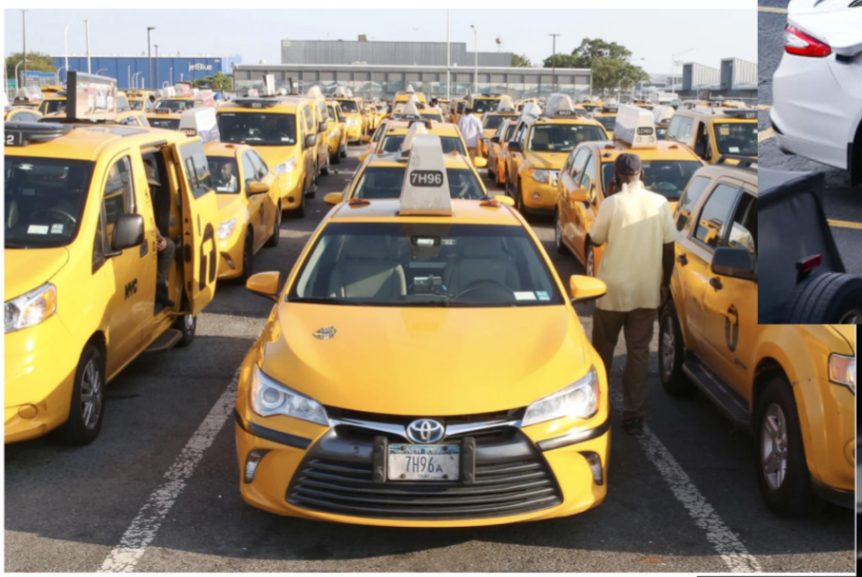
Colin Stretch, general counsel for Facebook, Sean Edgett, acting general counsel for Twitter, Richard Salgado, director of law enforcement and information security at Google, testify before the Senate Judiciary Committee's hearing on 'Extremist Content and Russian Disinformation Online: Working with Tech to Find Solutions' on Capitol Hill in Washington DC on Oct. 31, 2017. Shawn Thew / EPA

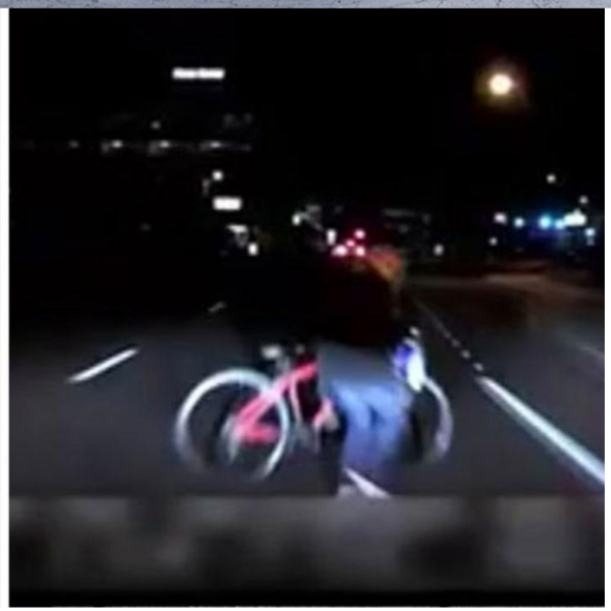
NEW YORK POST

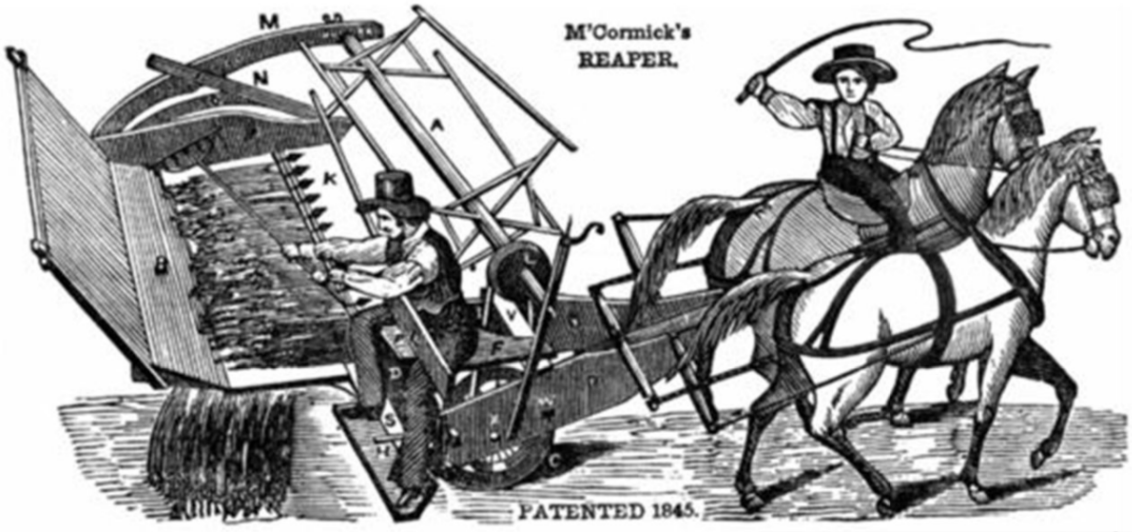
The challenges of driving a yellow cab in the age of Uber

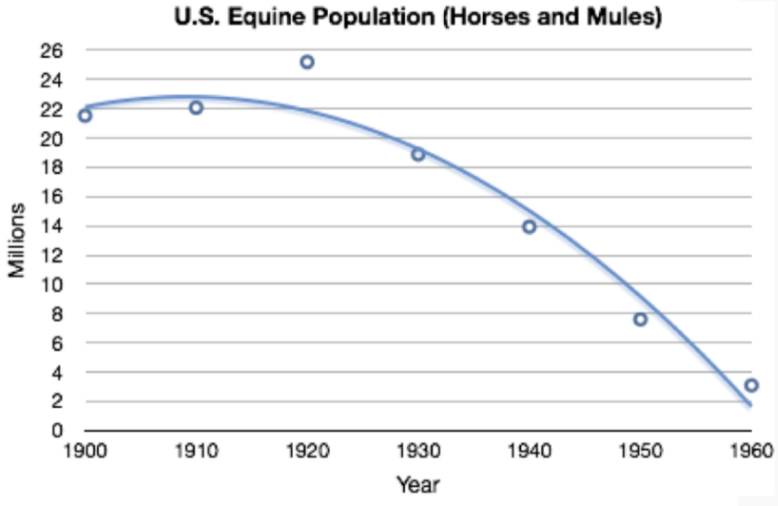
By John Crudele

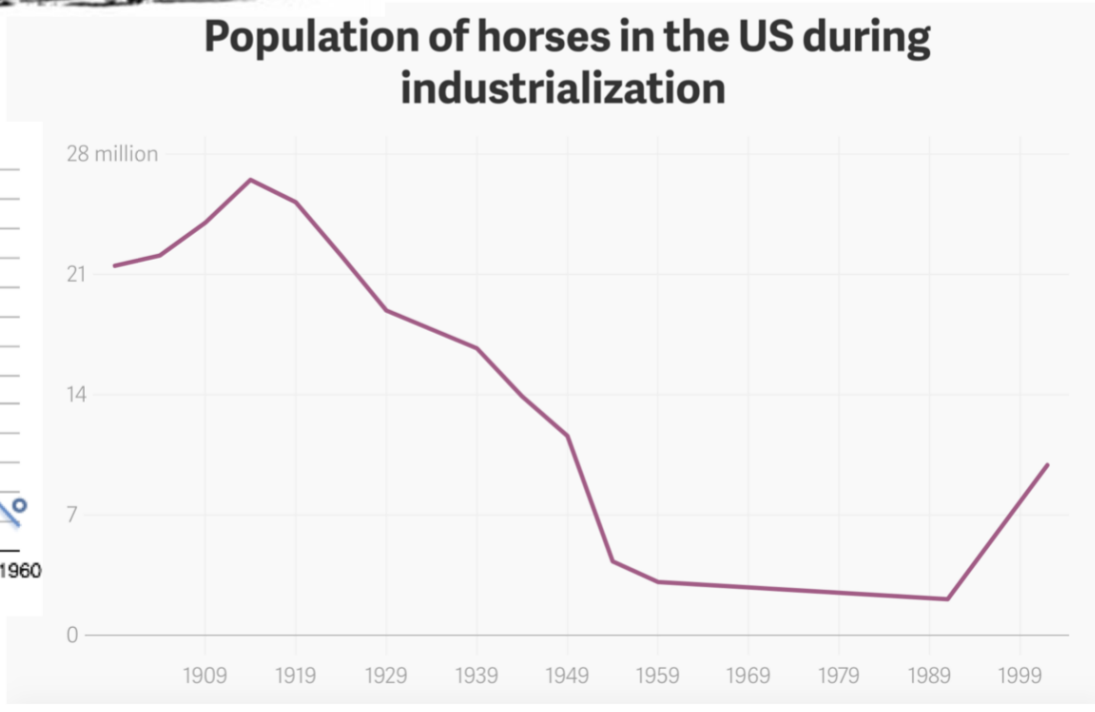
September 18, 2017 | 10:35pm

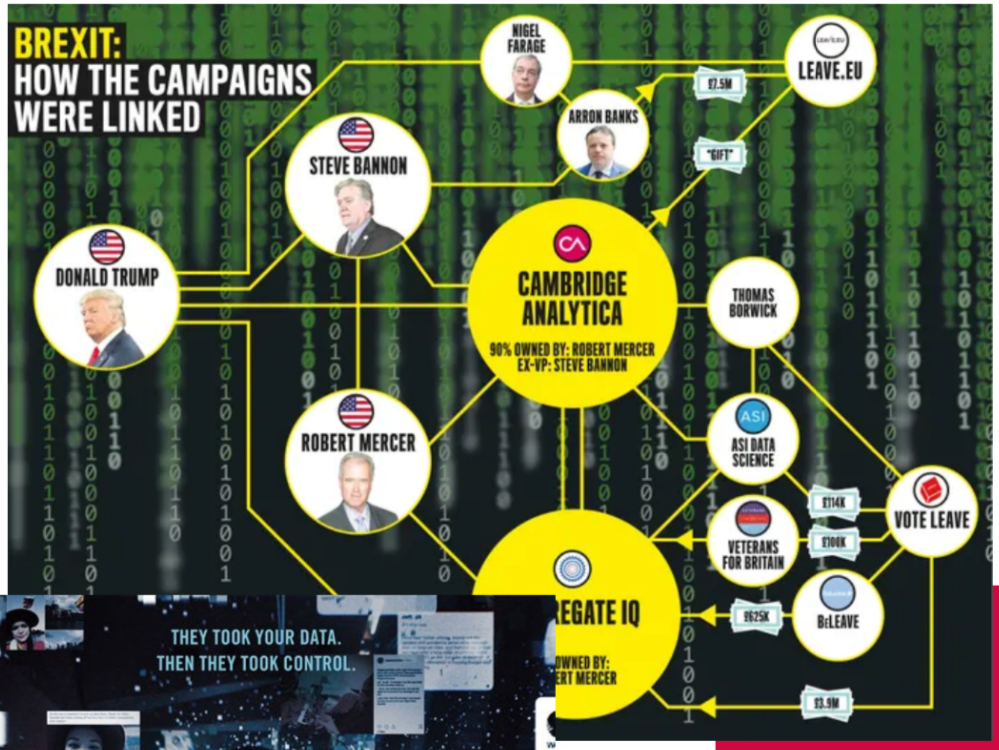





M'Cormick's
REAPER.
PATENTED 1845.


U.S. Equine Population (Horses and Mules)
Millions
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1900
1910
1920
1930
1940
1950
1960
Year


Population of horses in the US during industrialization
28 million
21
14
7
0
1909
1919
1929
1939
1949
1959
1969
1979
1989
1999



Cambridge Analytica

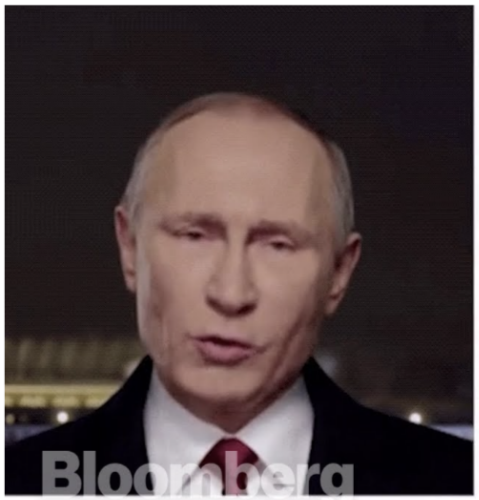
Facebook Bans Deepfake Videos That Could Sway Voters, But Is It Enough?

Forbes

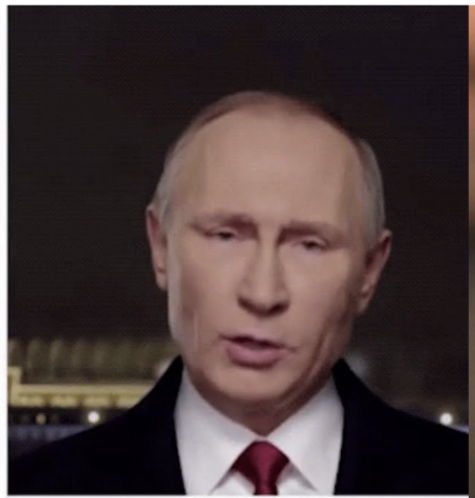
John Brandon Contributor ⓘ
Social Media
John Brandon covers social media trends. @jmbrandonbb

The Telegraph

'Deep fake' imagery manipulation poses threat to society not just military, US General warns



Reference

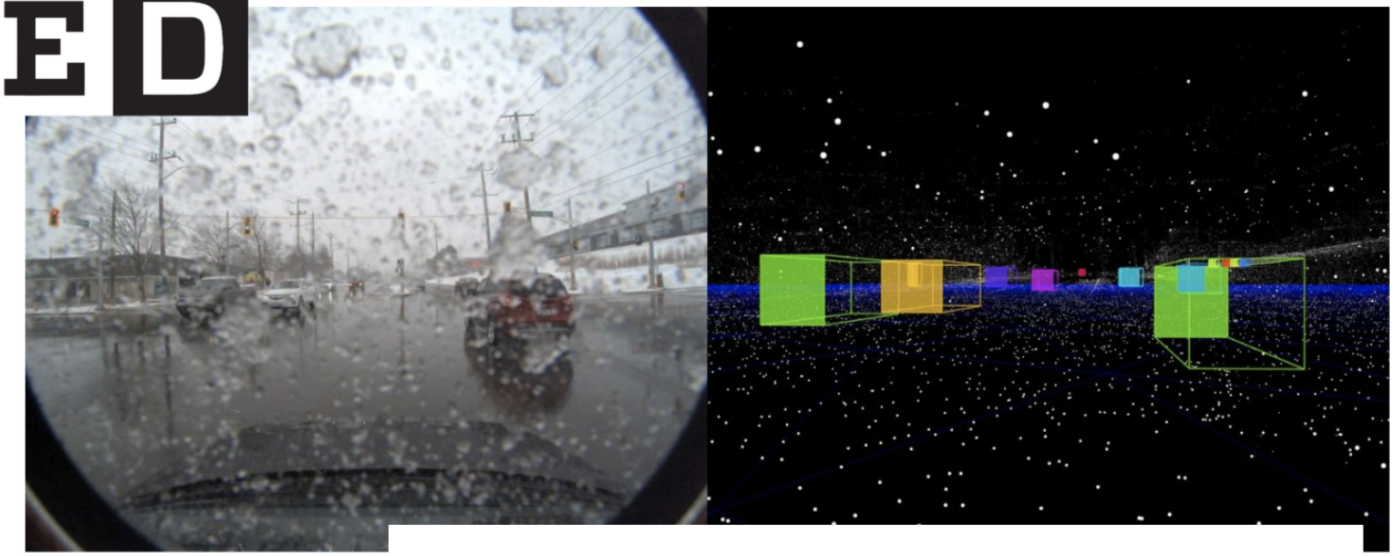


Our Result

Snow and Ice Pose a Vexing Obstacle for Self-Driving Cars

Most testing of autonomous vehicles until now has been in sunny, dry climates. That will have to change before the technology will be useful everywhere.

WIRED

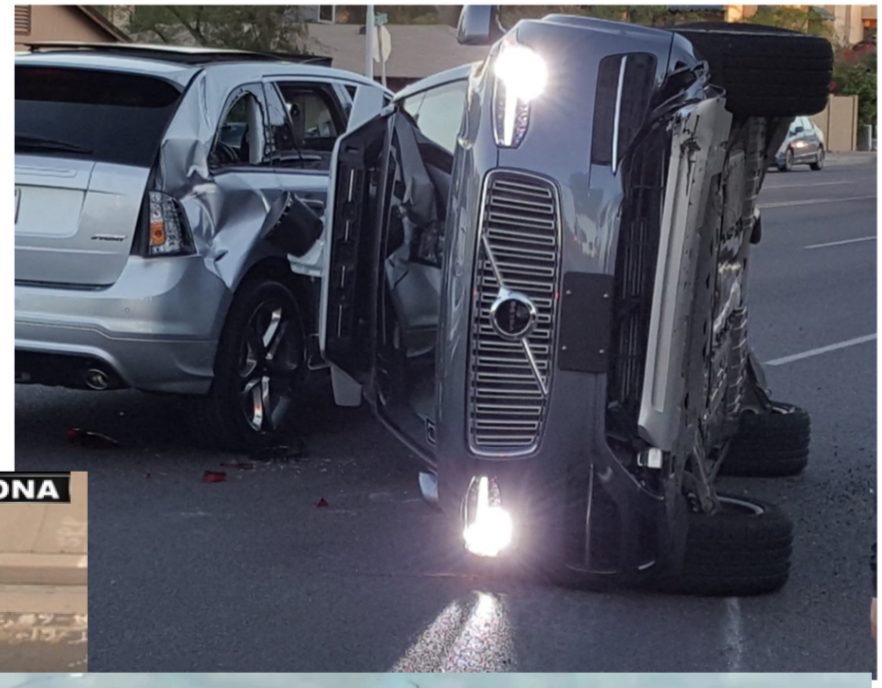


Self-driving cars will need to identify

The Canadian driving data set includes lane markings and vehicles that are covered with snow. COURTESY OF JEFF HILNBRAND

Are All Levels Of Autonomous Vehicles Equally Safe?

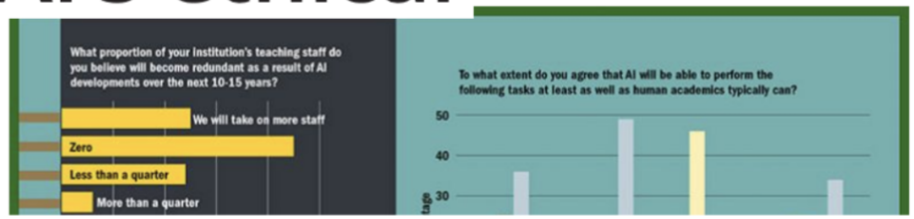
Kathleen Walch Contributor
COGNITIVE WORLD Contributor Group
AI



Horses, Equine Law And The Future Of The Autonomous Vehicle Legal Framework

Rahul Razdan Contributor ⓘ
Transportation
I focus on the impact of autonomous systems on society.

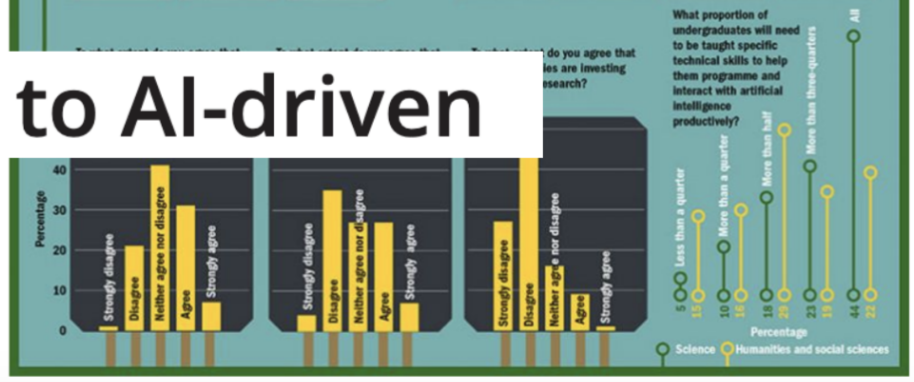
Universities told to be AI's ethical watchdogs



Industry and government alone cannot oversee new technology, experts say

TIMES HIGHER EDUCATION WORLD UNIVERSITY RANKINGS

Engineers will be crucial to AI-driven economies



Just as the AI revolution calls for more computer scientists, engineers will be needed to develop next-generation AI hardware, says Bashir M. Al-Hashimi

CIO JOURNAL.

What Machine Learning Can and Cannot Do

Jul 27, 2018 1:56 pm ET

A doctor examines a magnetic resonance image of a human brain during a Beijing neuroimaging competition between human doctors and AI, June 30, 2018. **PHOTO:** MARK SCHIEFELBEIN / ASSOCIATED PRESS

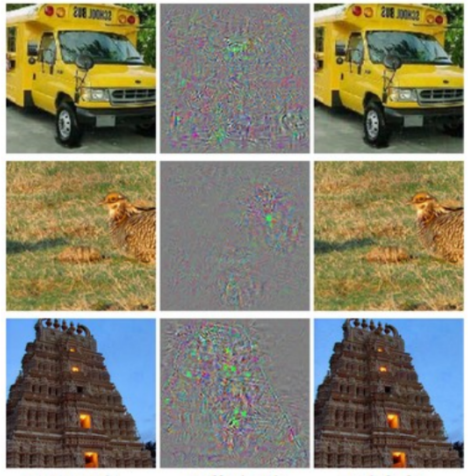
What Machine Learning can and cannot do

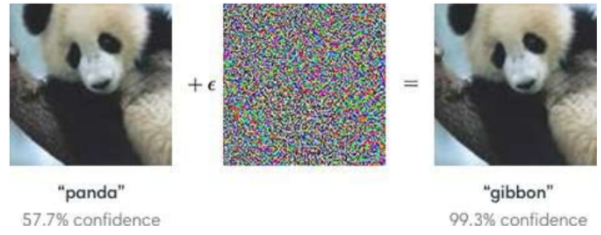
- We have seen many achievements (essentially in vision, language)..
- In particular Supervised learning has been quite successful
- But there are still plenty of tasks that computers still cannot handle (see Lex Friedman, MIT Sloan lecture)
  - Awareness of self
  - Emotion
  - Imagination
  - Morality
  - Consciousness
  - high level reasoning

Still many tasks that machines cannot do

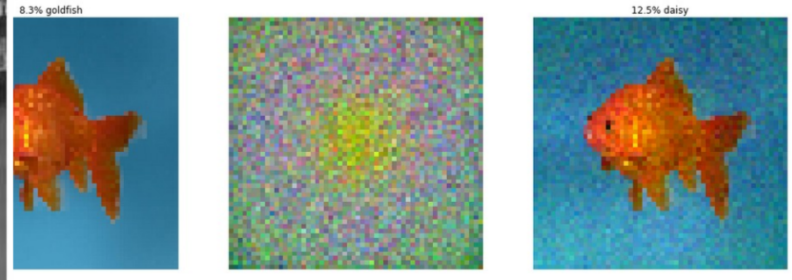
from Erik Brynjolfsson, ICLR 2018 keynote.

Immediate Challenges





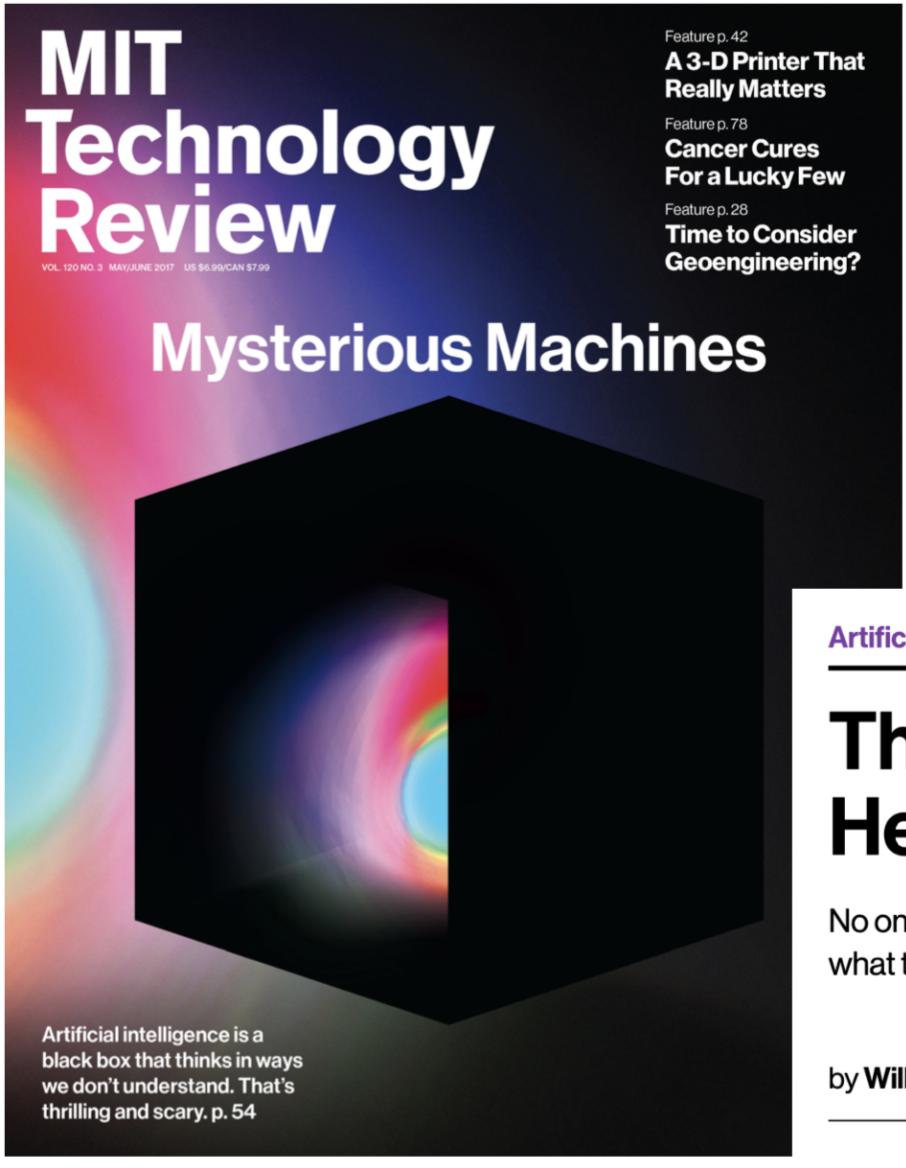
Neural Nets and machine vision can be fooled by very small distortions

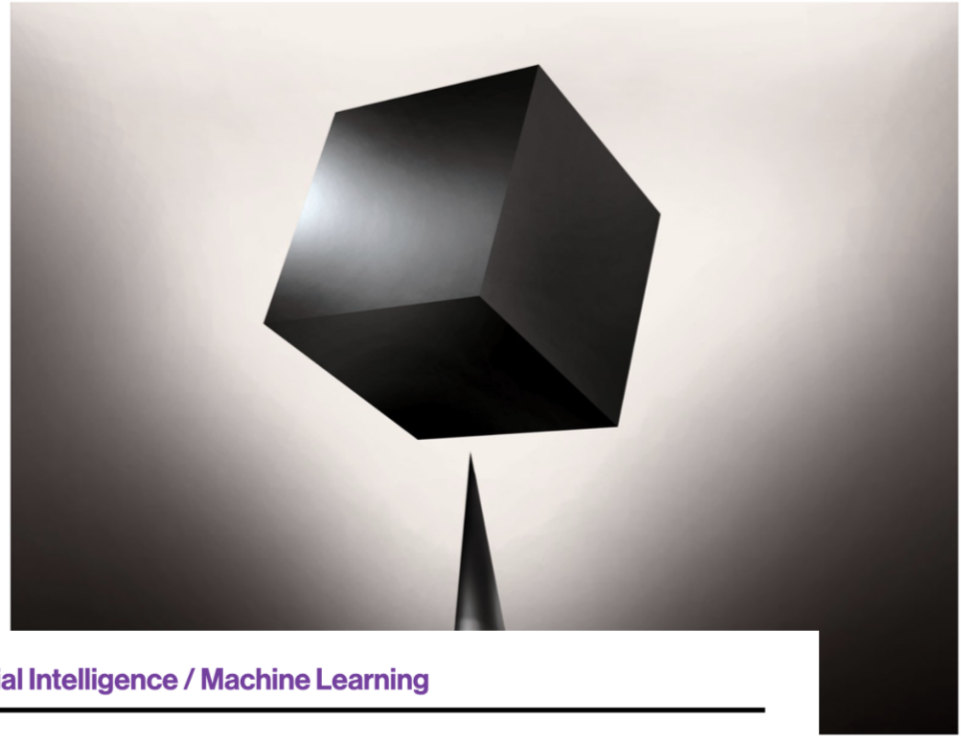


Immediate Challenges

(Lex Friedman, MIT Sloan)

- Occlusions
- Sensor spoofing (camera, Lidar)
- adversarial noise
- Risk quantification
- Data is costly $\Rightarrow$ Ideally, we would like to move on to more unsupervised learning.





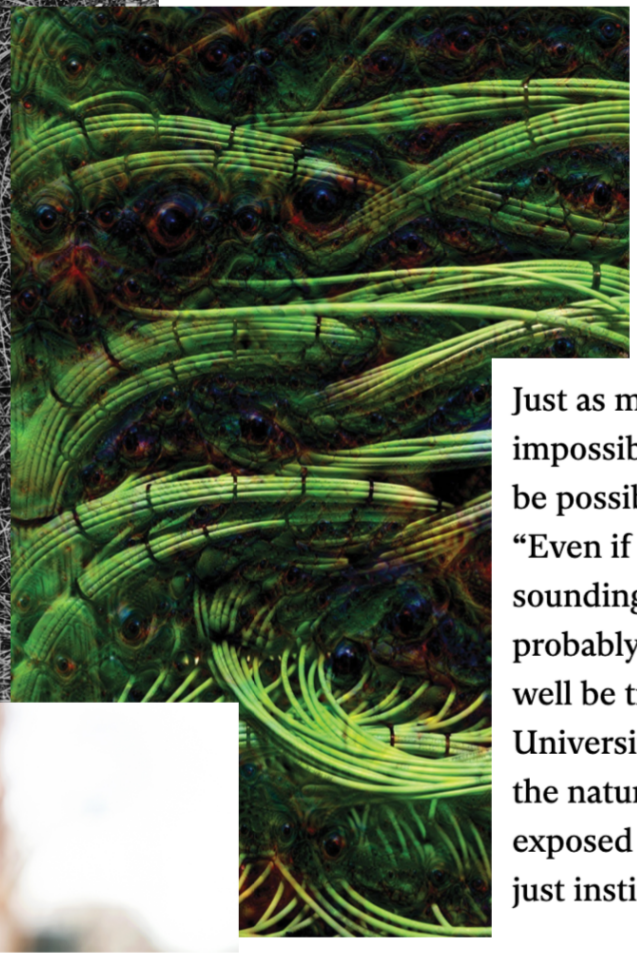
Artificial Intelligence / Machine Learning

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by **Will Knight**

"We can build these models, but we don't know how they work."



How well can we get along with machines that are unpredictable and inscrutable?

Just as many aspects of human behavior are impossible to explain in detail, perhaps it won't be possible for AI to explain everything it does. "Even if somebody can give you a reasonable-sounding explanation [for his or her actions], it probably is incomplete, and the same could very well be true for AI," says Clune, of the University of Wyoming. "It might just be part of the nature of intelligence that only part of it is exposed to rational explanation. Some of it is just instinctual, or subconscious, or inscrutable."

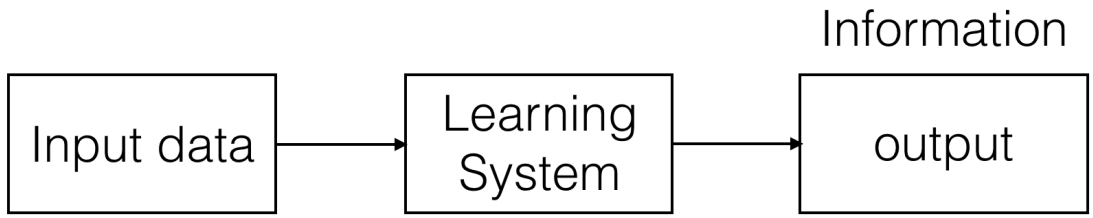
Deep neural networks are coming to your phone

Already, mathematical models are being used to help determine who makes parole, who's approved for a loan, and who gets hired for a job. If you could get access to these mathematical models, it would be possible to understand their reasoning. But banks, the military, employers, and others are now turning their attention to more complex machine-learning approaches that could make automated decision-making altogether inscrutable.

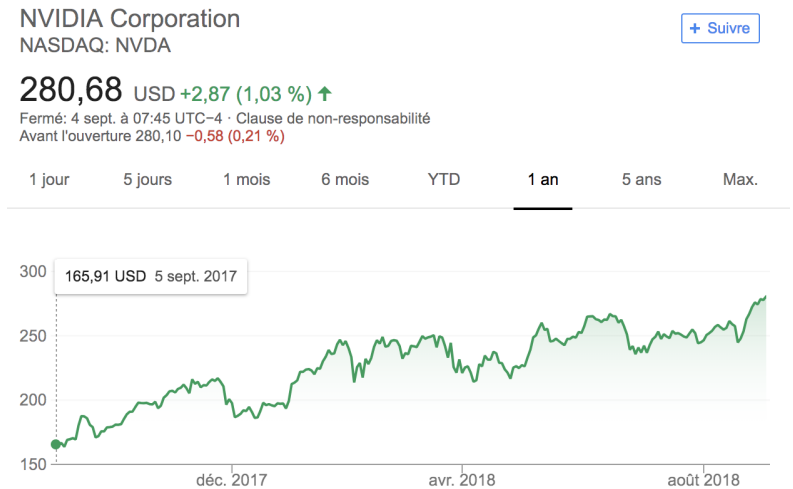
"Whether it's an investment decision, a medical decision, or maybe a military decision, you don't want to just rely on a 'black box' method."

Machine Learning is not new..

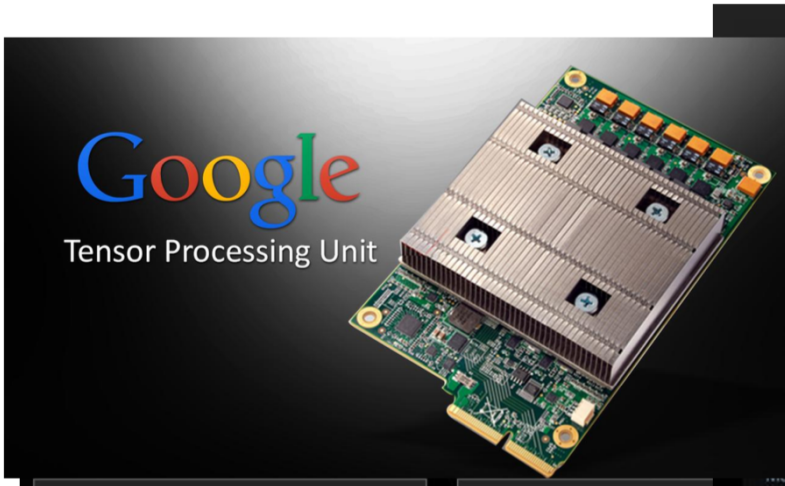
- General principle of machine learning is very simple

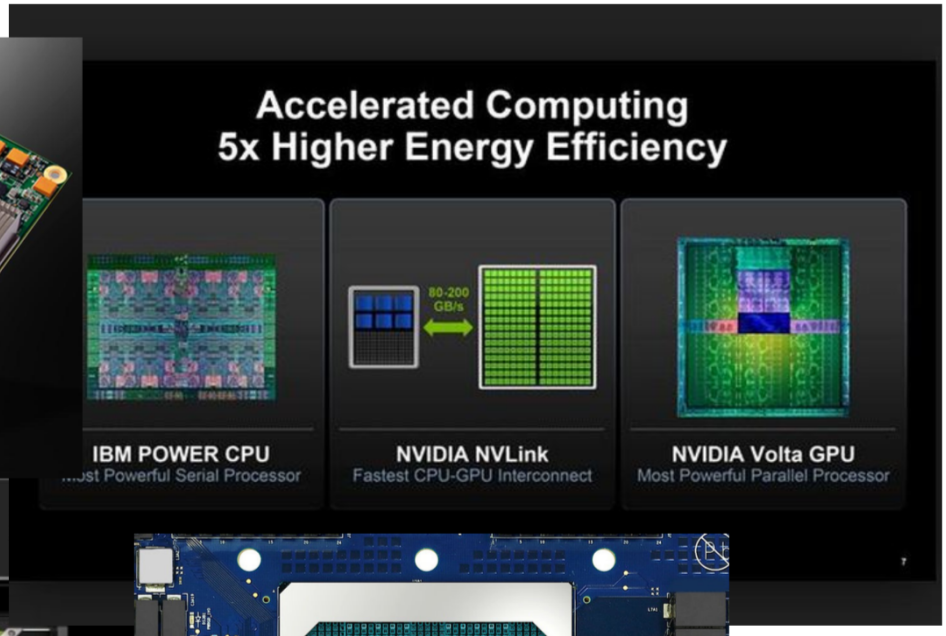


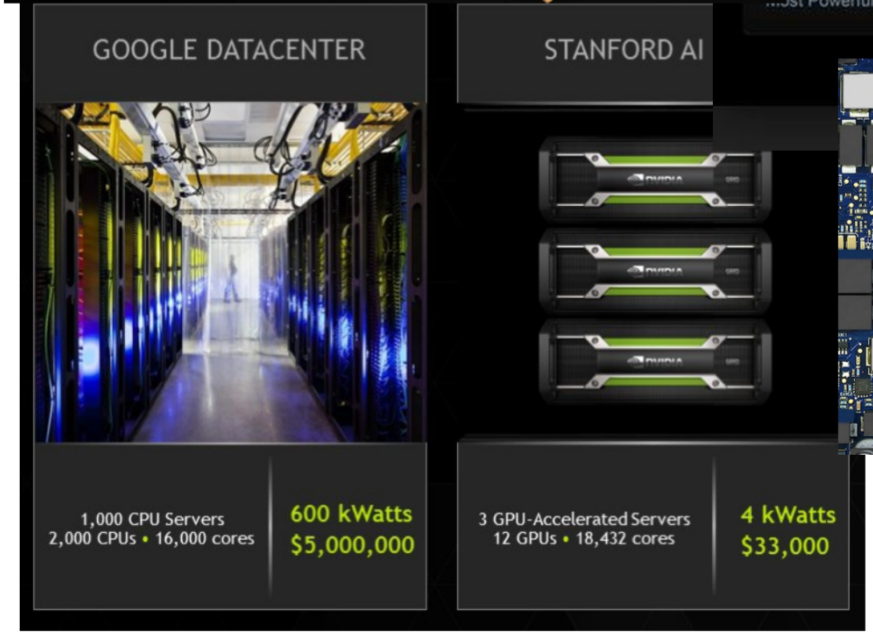
- One of the reason for the renewed excitement is Massive parallelism through Graphical Processing Units $\Rightarrow$ Essential for neural network training on massive databases (think of imageNet, GoogleNet,..)

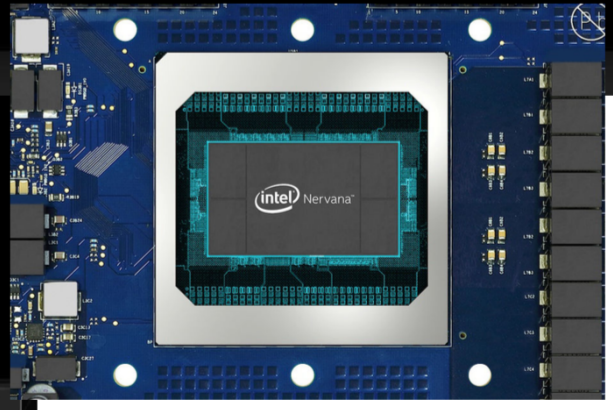


Some new architectures are coming









Issue: costly
still Energy greedy

The general picture: Supervised vs Unsupervised

- Supervised learning tries to understand the relation between data $\mathcal{D} = \{x_i\}_{i=1}^{N}$ and the associated labels (targets) $\{y_i\}$ based on a subset of samples for which the labels are known. The goal is thus to find a mapping between the points $x_j \in \mathcal{D}$ and their targets y_j. During training all labels are known.

  Ex: Handwriting recognition. Data = images from MNIST, labels,knowledge = actual numbers displayed

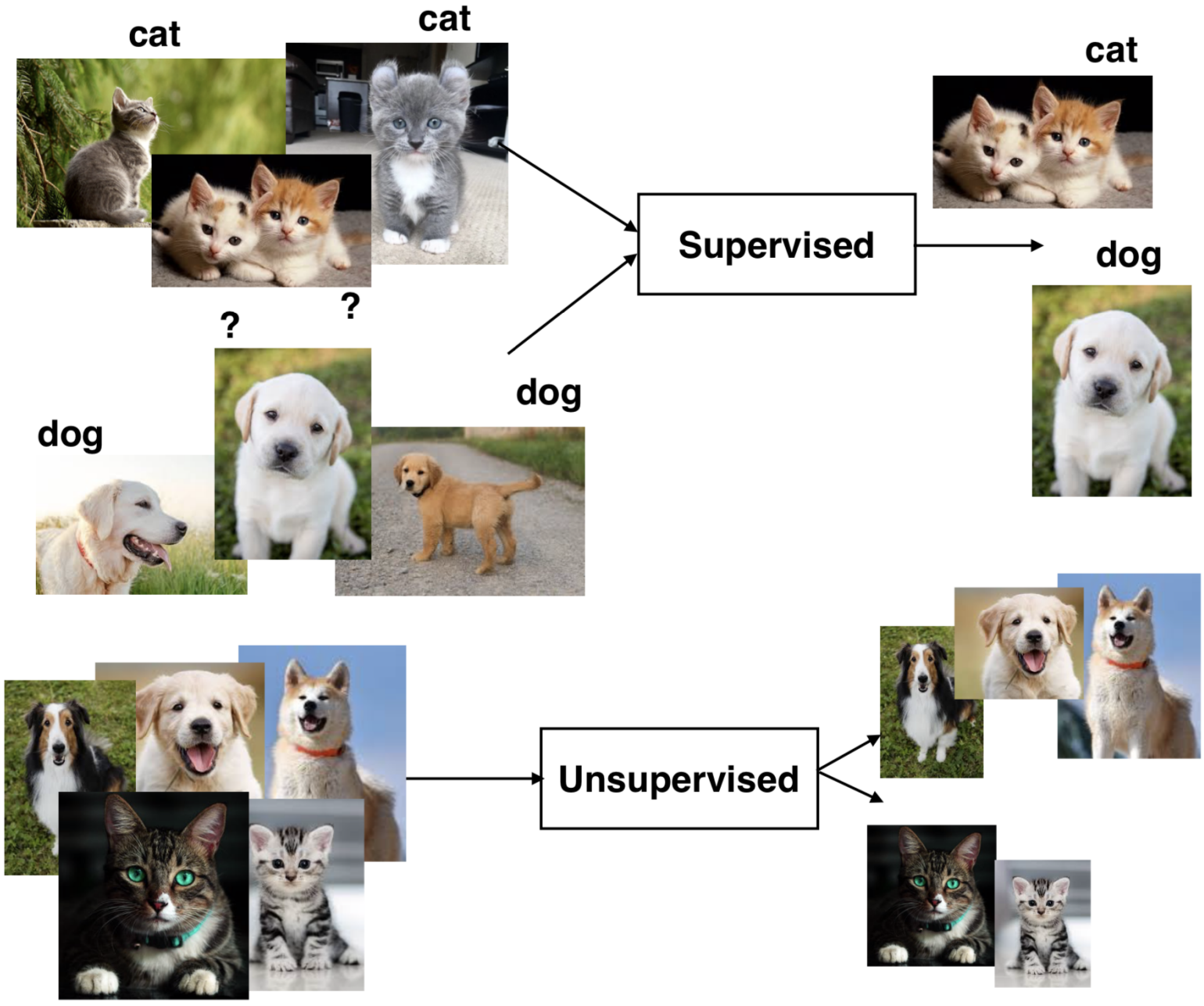
- Unsupervised learning tries to find structure within a given (unlabeled) dataset

  Ex: customized advertising (cluster users in groups in order to send specific advertising to each group)

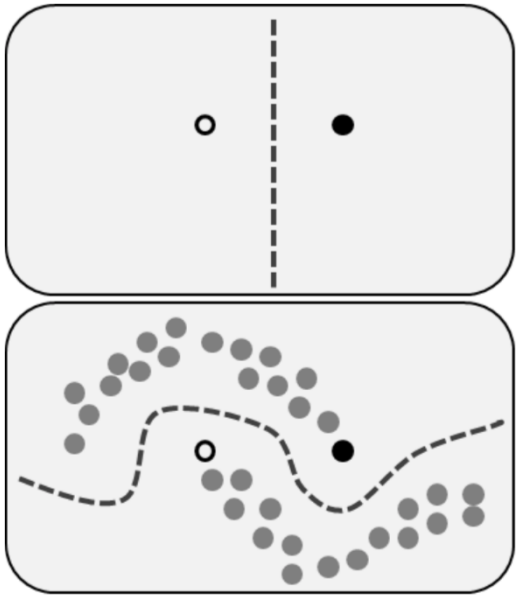
The big picture: Supervised vs Unsupervised

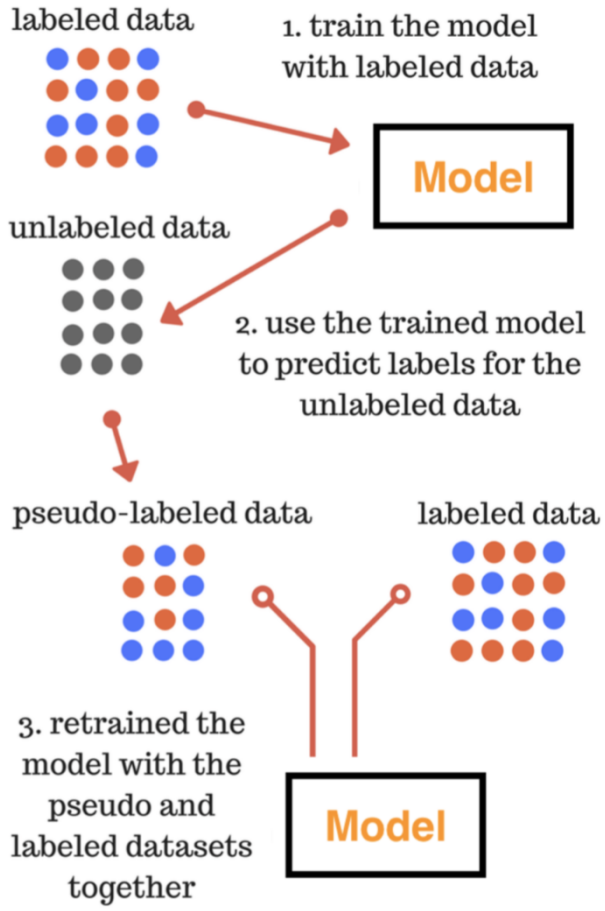
- Semi-supervised Learning (SSL): small amount of labeled data with large amount of unlabeled data. Other examples of partial supervision can involve constraints on the prototypes x_i (such as requiring subgroups of prototypes to have the same target)

- Reinforcement: The machine looks for suitable actions, in a given situation, in order to maximize a "reward" (e.g. Neural Network backgammon playing : board position + dice value $\Rightarrow$ move + reward)


cat
cat
cat
Supervised
dog
?
?
dog
dog
Unsupervised

Semi-supervised: How unlabelled data can help?





Supervised learning

Supervised learning can be subdivided into two main classes of problems:

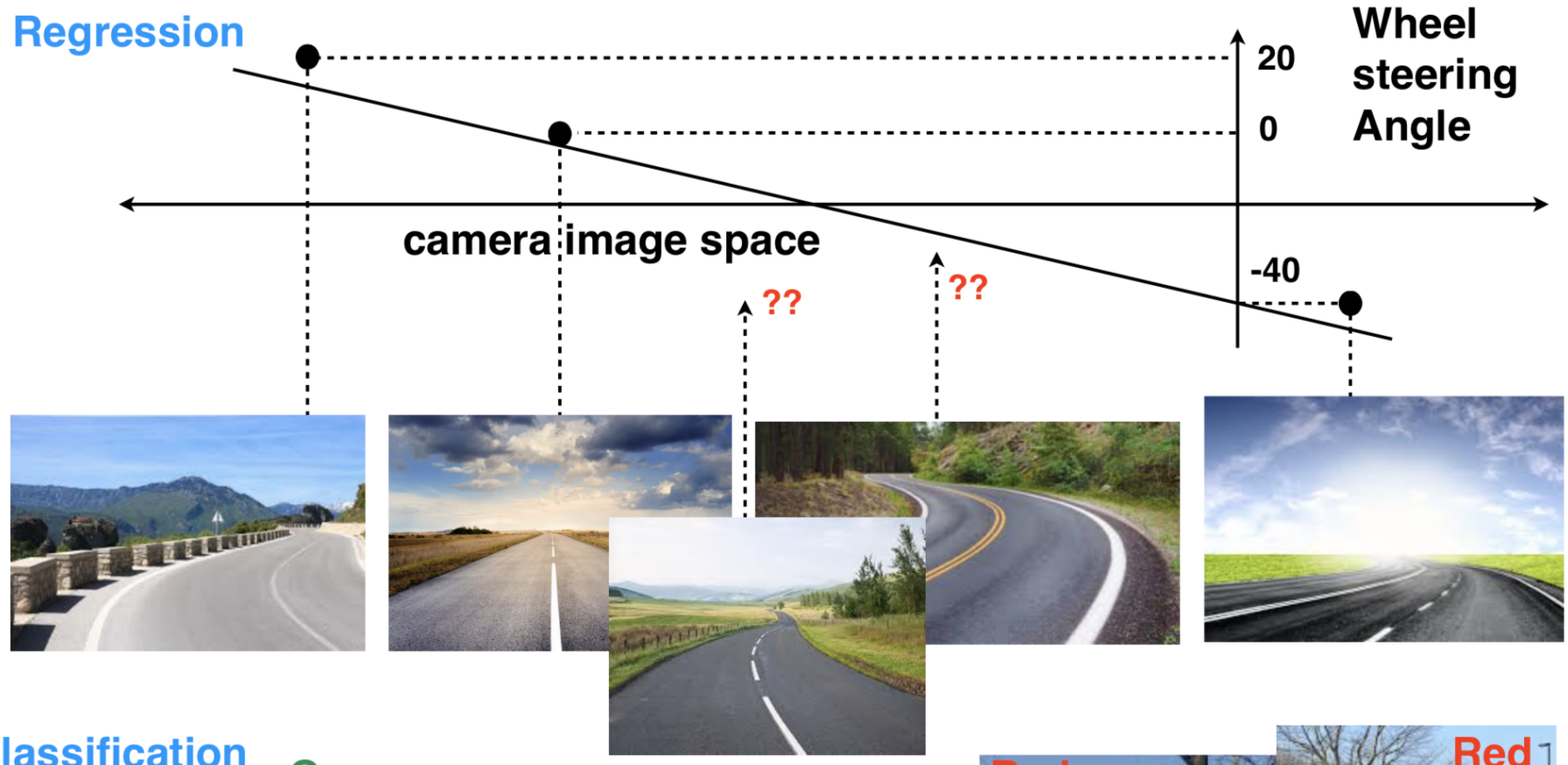
- Regression problems (approx. a mapping $f(x_\mu)$ from inputs x_μ to continuous output variables y_μ)

$$Ex. \quad \hat{\beta} = \min_{\beta} \sum_{i=1}^{N} (y_i - \langle x_i, \beta \rangle)^2 \quad \text{(residual SS)}$$

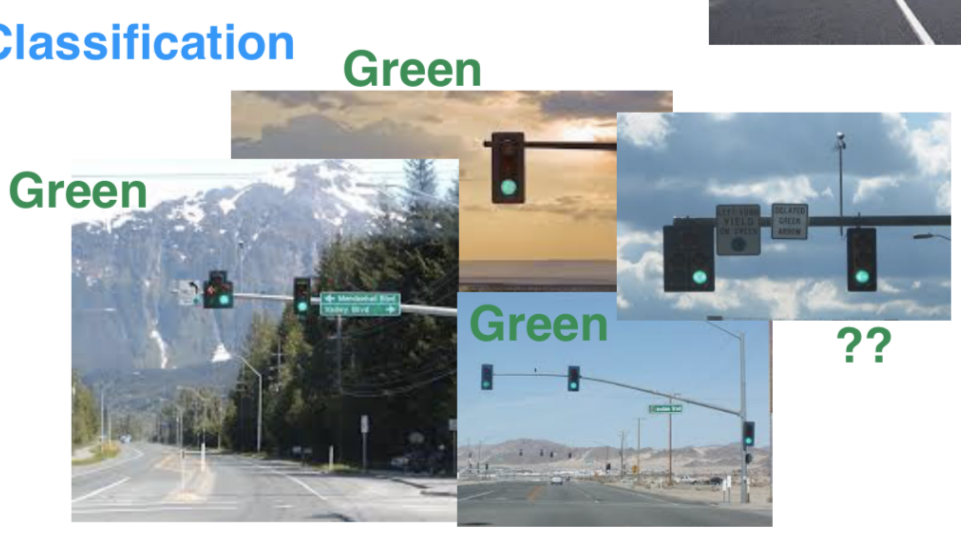
- Classification problems (approx. a mapping $f(x_\mu)$ from inputs x_μ to discrete output variables y_μ)

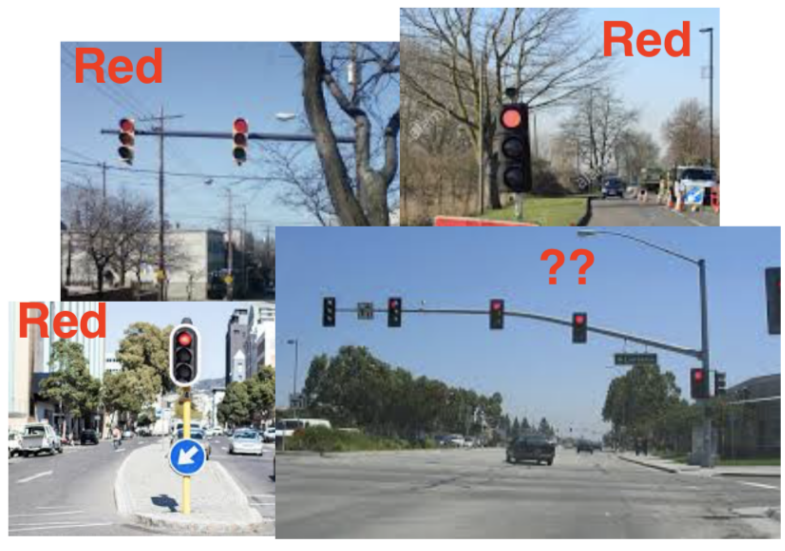
$$Ex. \quad p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\boldsymbol{x})} = \frac{N_{\delta,K}}{N_\delta}, \quad \text{(Nearest Neighbors)}$$

Regression



Classification





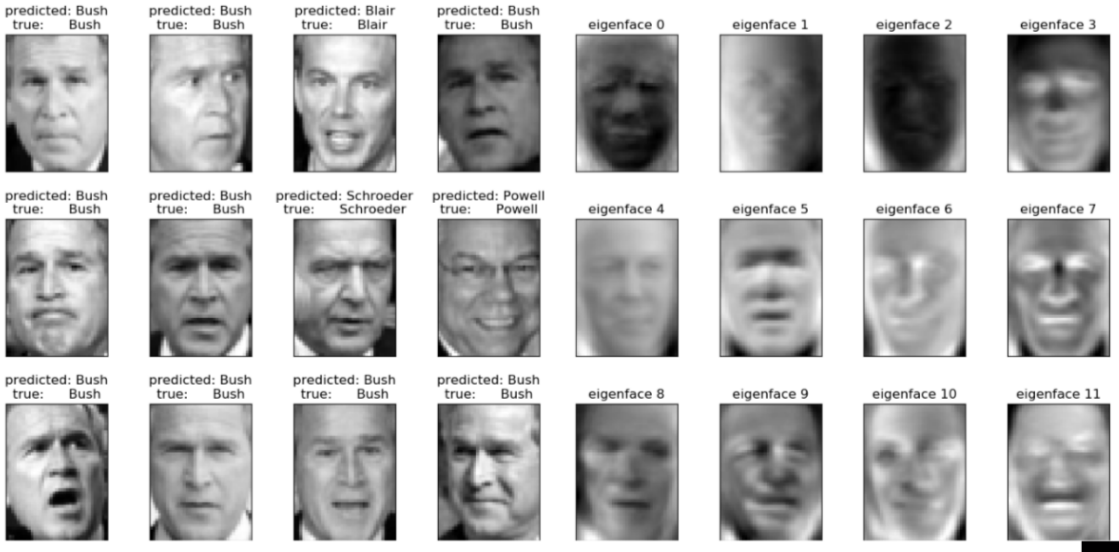
Examples of supervised learning algorithms

Many possible regression/classification models (some models can be used in both frameworks)

- Linear Regression
- Neural networks
- Support vector machines (SVMs)
- K-nearest neighbors
- Tree based models
- ...

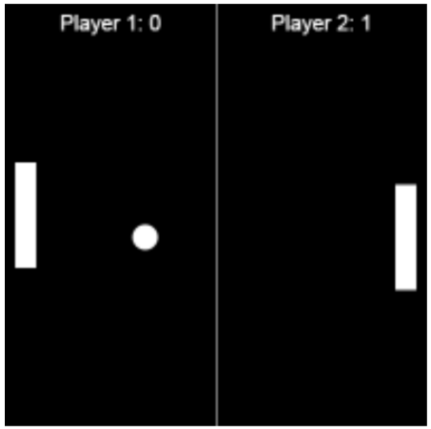
Supervised Learning: What are we going to learn?

SVMs for face recognition



Logistic regression, neural nets for handwritten digits recognition

Neural nets pong training



Parametric vs non-parametric

- Fixed number of parameters = parametric
  $+$: faster to use
  $-$: stronger assumptions regarding data distribution.

  Ex. linear regression

- Complexity of the model grows with size of training dataset = non-parametric
  $+$: more flexible regarding data.
  $-$: often computationally intractable for large datasets

  Ex. KNN

Parametric vs non-parametric

- From Hastie, Tibshirani, Friedman

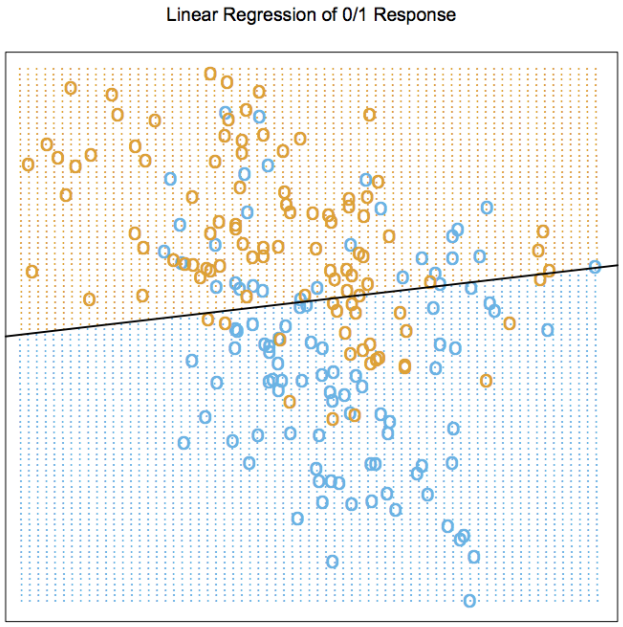


FIGURE 2.1. *A classification example in two dimensions. The classes are as a binary variable (BLUE = 0, ORANGE = 1), and then fit by linear regre The line is the decision boundary defined by $x^T\hat{\beta} = 0.5$. The orange shaded r denotes that part of input space classified as ORANGE, while the blue regi classified as BLUE.*

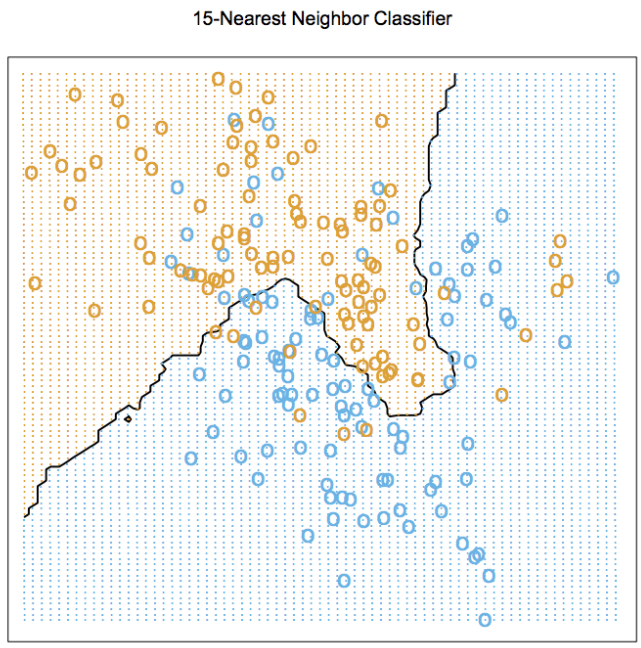


FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.*

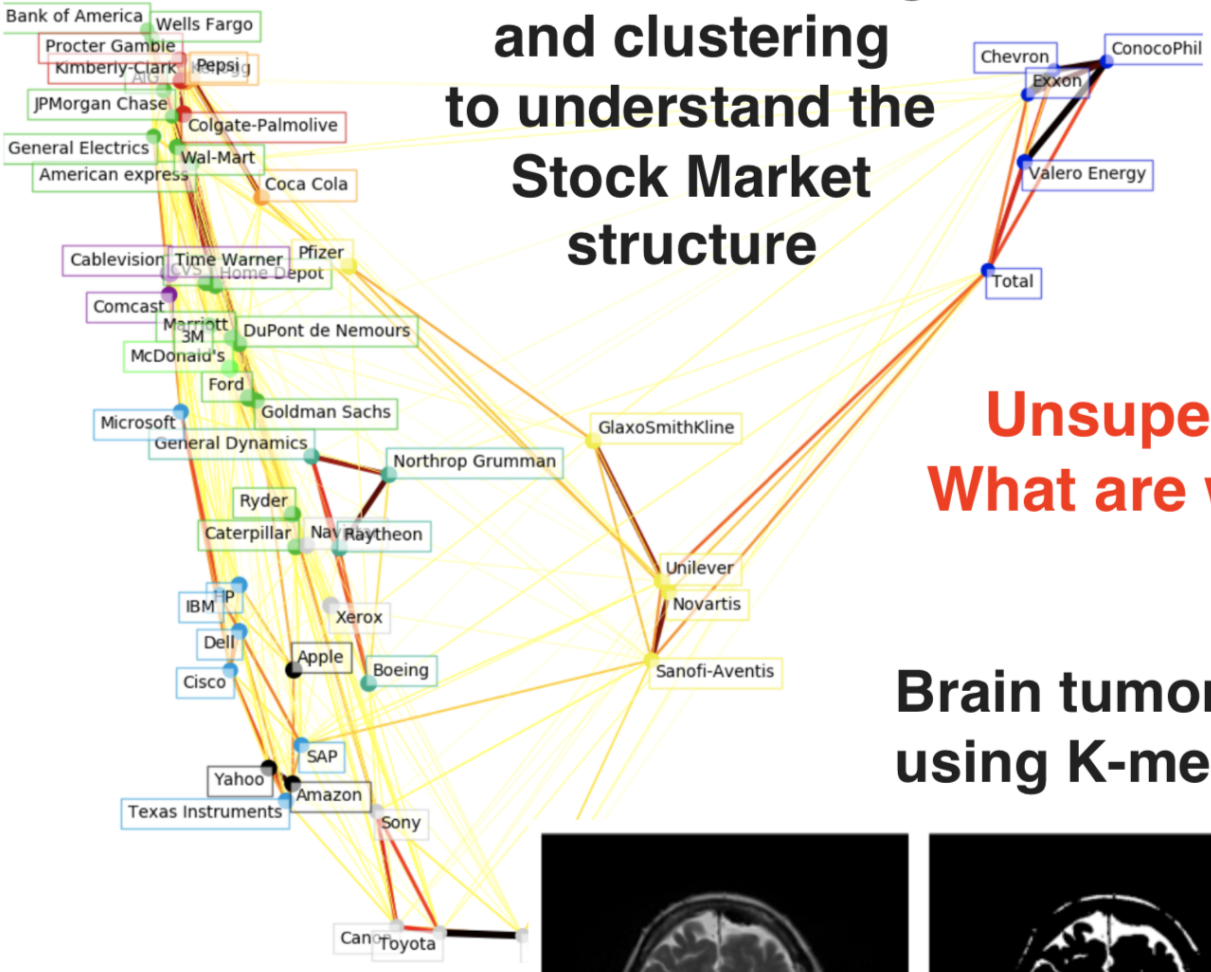
Unsupervised learning

In unsupervised learning, we are only given prototypes x_μ and we want to extract pattern from the data.

Examples of unsupervised learning approaches include

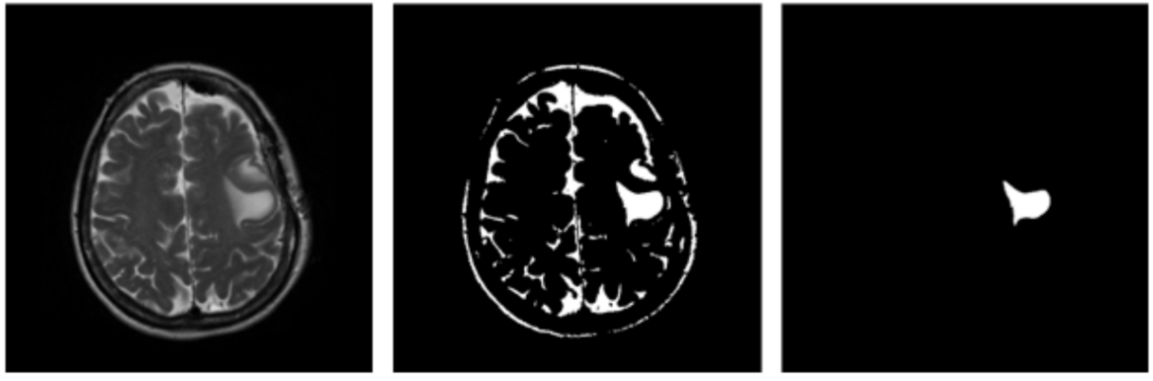
- Clustering (K-means, K-medoid)
- Manifold learning/Latent variable models
- Factor Analysis, Principal and Independent Component Analysis (PCA/ICA)
- Gaussian mixture models (GMM)

Manifold learning and clustering to understand the Stock Market structure



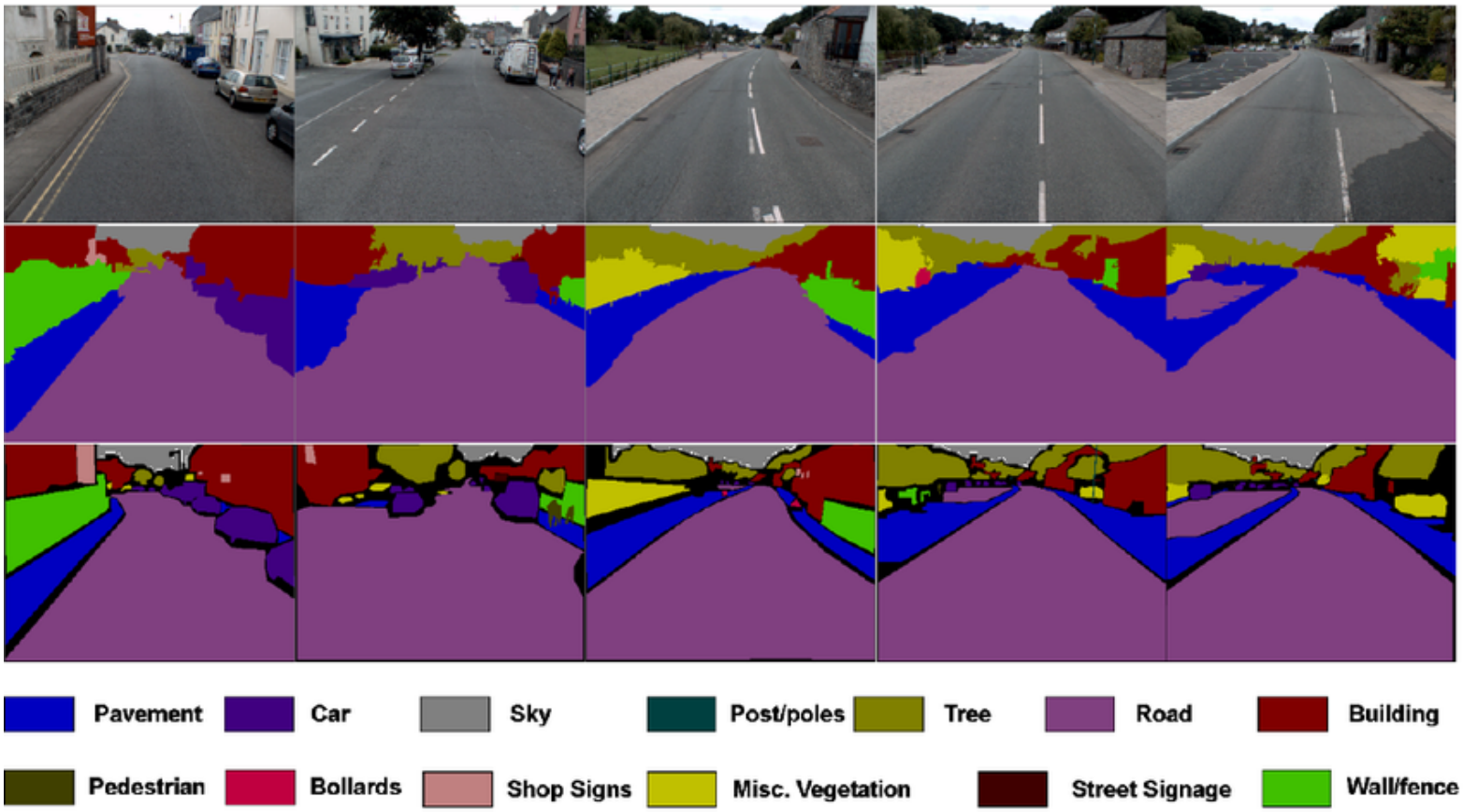
Unsupervised Learning: What are we going to learn?

Brain tumor segmentation using K-means



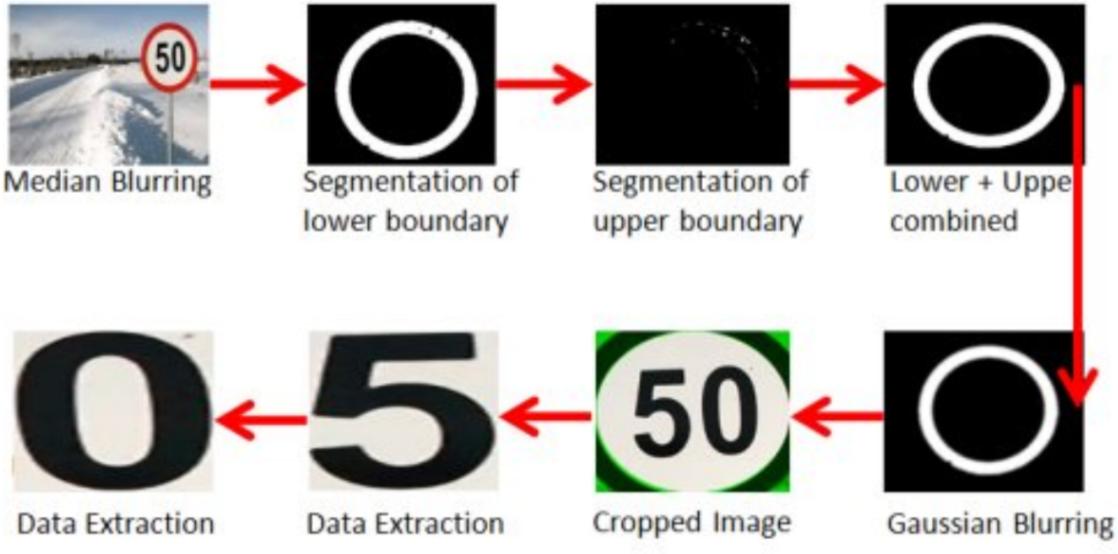
(a) (b) (c)

Image segmentation through clustering



You Can Trick Self-Driving Cars by Defacing Street Signs

By **Catalin Cimpanu**



What are we going to learn?
Combined supervised and unsupervised

Samir Bhatt, Bhaskar Trivedi, Ankur Devani, Hemang Bhimani (eInfochips)

Short reminder on Linear Algebra: motivation

Why are Linear algebra and Differential Calculus useful in Machine Learning?

1. Linear algebra is useful in both supervised and supervised learning to derive the expression of the model's parameters given the training data.
   - Several models in unsupervised learning rely on matrix factorization. Singular value decomposition, for example can be used for dimensionality reduction (PCA, ICA)
   - When looking for an optimal model, given a particular dataset $\mathcal{D}$, we will often have to find the solution to multivariate systems of linear equations. Linear algebra will be useful to understand how we can compute the solution efficiently when those systems are overdetermined or even ill conditioned.

2. The tools from linear algebra will also be useful when will need concepts from multivariate statistics

Short reminder on Linear Algebra

(also see the notes of Z. Kolter)

- Vector products. Given two vectors $\boldsymbol{x}$ and y one can define the quantity $\boldsymbol{x}^T\boldsymbol{y}$ which is called the inner product or scalar product of $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T\boldsymbol{y} \in \mathbb{R} = [x_1, x_2, \ldots, x_n] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

- Norms. A norm is any function $N : \mathbb{R}^n \mapsto \mathbb{R}$ that satisfies the following properties
  1. For all $x \in \mathbb{R}^n$, $N(x) \geq 0$
  2. $N(x) = 0$ if and only if $x = 0$
  3. For all $x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $N(\alpha x) = |\alpha| N(x)$
  4. For all $x, y \in \mathbb{R}^n$, $N(x + y) \leq N(x) + N(y)$

Short reminder on Linear Algebra

(also see the notes of Z. Kolter)

- Common examples of norms include the ℓ_2 norm $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$, the ℓ_1 norm $\|x\|_1 = \sum_{i=1}^n |x_i|$ and the ℓ_∞ norm $\|x\|_\infty = \max_i |x_i|$.
- More generally, the familly of ℓ_p norms is defined as $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$

Short reminder on Linear Algebra

(also see the notes of Z. Kolter)

- Similar ideas hold for matrices. For two matrices A and B, one can define an inner product $\langle A, B \rangle$ as $\langle A, B \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} B_{ij}$.

- Just as we define the ℓ_p norms for vectors, we can define spectral p-norms on matrices (which are jsut ℓ_p on the vector of singular values)

$$\|A\|_1 = \sum_{i=1}^{n} \sigma_i(A)$$

$$\|A\|_2 = \sqrt{\sum_{i=1}^{n} \sigma_i^2(A)}$$

$$\|A\|_{\text{inf}} = \sup_{1 \le i \le n} \sigma_i(A)$$

Short reminder on Linear Algebra

(also see the notes of Z. Kolter)

- Linear independence. A set of vectors is said to be linearly independent if no vector can be represented as a linear combination of the remaining set of vectors.
- The rank of a matrix A is the largest subset of linearly independent columns of A.
- The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ is the unique matrix A^{-1} such that $A^{-1}A = I = AA^{-1}$
- Two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ are orthogonal if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$. A vector $\boldsymbol{x}$ is normalized if $\|\boldsymbol{x}\| = 1$
- A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if $U^T U = UU^T = I$.

Eigenvalues and eigenvectors

(also see the notes of Z. Kolter)

- Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of A and $\boldsymbol{x} \in \mathbb{C}^n$ is its corresponding eigenvector if

$$A\boldsymbol{x} = \lambda \boldsymbol{x}, \quad \boldsymbol{x} \neq 0$$

- Given a symmetric matrix $A \in \mathbb{S}^{n \times n}$, all the eigenvalues of A are real. Moreover, the eigenvectors of A are orthonormal. If we store the eigenvectors in U and the eigenvalues in the diagonal matrix Λ, A can read as

$$A = U \Lambda U^*$$

- From this, for any vector $x \in \mathbb{R}^n$, we thus have $x = \sum_i t_i u_i + \overline{x}$ where $\langle \overline{x}, u_i \rangle = 0$ for all u_i which gives

$$x^T A x = \sum_i \lambda_i \left(\sum_j t_j u_j^T \right) u_i u_i^T \left(\sum_j t_j u_j \right) = \sum_i |t_i|^2 \lambda_i$$

Eigenvalues and eigenvectors

- In particular, when $\|x\|^2 = 1 = \sum_i t_i^2 \|u_i\|^2 = \sum_i t_i$you see that the quantity

$$x^T A x = \sum_i \lambda_i \left(\sum_j t_j u_j^T \right) u_i u_i^T \left(\sum_j t_j u_j \right) = \sum_i |t_i|^2 \lambda_i$$

  is maximized when $x = t_j u_j$ where u_j is the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$.

- The solution to the maximization problem

$$\max_{x \in \mathbb{R}^n} x^T A x \quad \text{subject to} \quad \|x\|_2^2 = 1$$

  is thus the eigenvector corresponding to the largest eigenvalue

- We will use this later when discussing latent variable models in unsupervised learning.

(Matrix) differential calculus

- Given a function $f : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$, that takes as input the matrix of unknowns $\boldsymbol{X}$ the gradient of $f(\boldsymbol{X})$ with respect to $\boldsymbol{X}$ is defined as fhte matrix

$$\nabla_{\boldsymbol{X}} f(\boldsymbol{X}) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(\boldsymbol{X})}{\partial X_{11}} & \frac{\partial f(\boldsymbol{X})}{\partial X_{12}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial X_{1n}} \\ \frac{\partial f(\boldsymbol{X})}{\partial X_{21}} & \frac{\partial f(\boldsymbol{X})}{\partial X_{22}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial X_{2n}} \\ \vdots & \cdots & \ddots & \vdots \\ \frac{\partial f(\boldsymbol{X})}{\partial X_{m1}} & \frac{\partial f(\boldsymbol{X})}{\partial X_{m2}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial X_{mn}} \end{bmatrix}$$

- When $\boldsymbol{X}$ is a vector, we recover

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

(Matrix) differential calculus

- When looking for the optimal model in terms of empirical risk function, we will often have to find the value x that minimizes the sum of squred residuals

$$\min_x \|Ax - b\|_2^2 = x^T A^T Ax - 2b^T Ax + b^T b$$

- The solution in this case can be computed in closed form by first taking the derivative with respect to x,

$$\nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) = 2A^T Ax - 2A^T b$$

- Setting the derivative to 0, we get

$$x = (A^T A)^{-1} A^T b$$

provided that $A^T A$ is invertible.

(Matrix) differential calculus

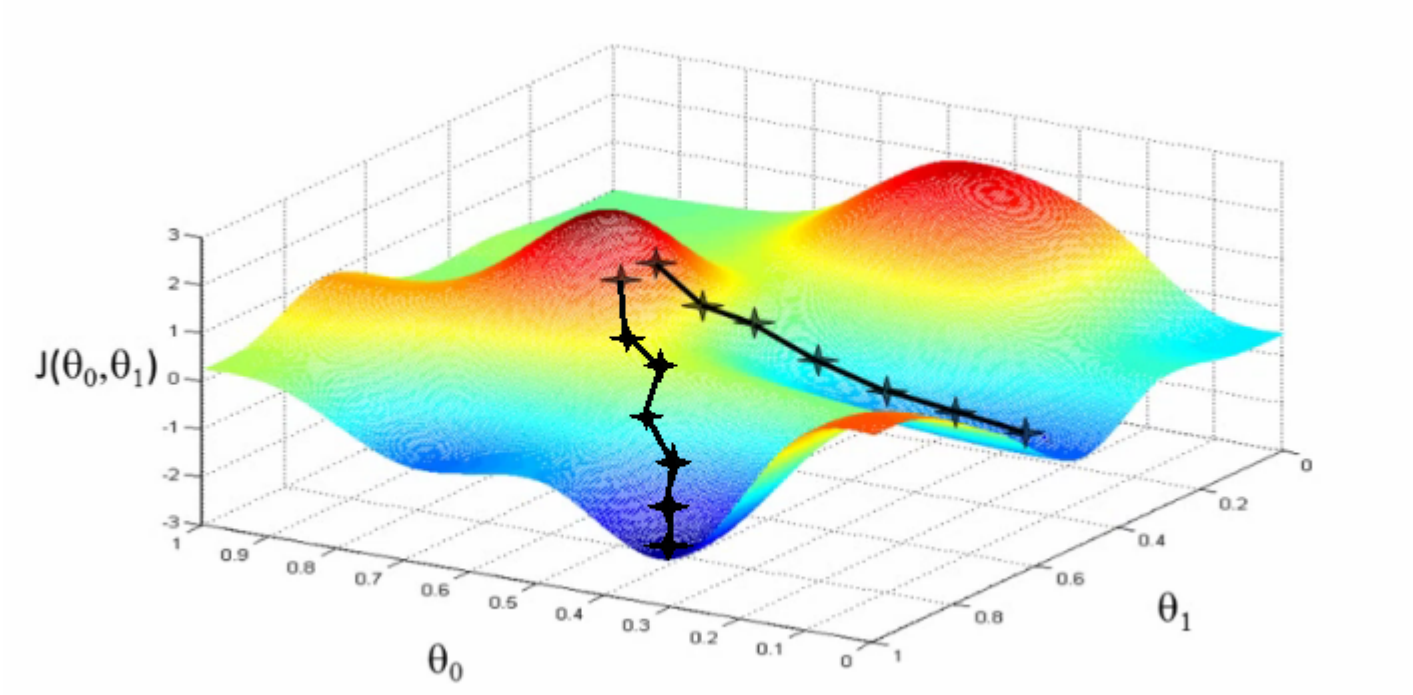
- When deriving a closed form solution is not possible, the model (A, b) can be learned through gradient descent
- For any given function $f(\boldsymbol{x})$, the gradient gives a vector pointing in the direction of greatest increase of the function. Hence $-\nabla f(\boldsymbol{x})$ is a vector pointing in the direction of greatest decrease of $f(\boldsymbol{x})$
- Starting from an initial guess $\boldsymbol{x}^{(0)}$ for the solution of $\min_{\boldsymbol{x}} f(\boldsymbol{x})$, one can thus define an iterative procedure (known as gradient descent) as

$$\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} - \alpha \nabla f\left(\boldsymbol{x}^{(t)}\right)$$

α, which represents the step size, is also known as the learning rate. When $\boldsymbol{x} \in \mathbb{R}^n$ is a vector, recall that we have

$$\nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right]$$

Gradient descent



Reminders on Statistics and probability

- Why using stats/proba?
- Machine Learning relies on complex distributions (cancerous cells, possible moves in Go, Existing sign roads, possible evolutions of stocks,...)
- Only a few samples are usually available
- $\Rightarrow$ We need a way to measure how well those samples are representing the underlying (unknown) distribution

Why is that important?

Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam

A woman crossing Mill Avenue at its intersection with Curry Road in Tempe, Ariz. on Monday. A pedestrian was struck and killed by a self-driving Uber vehicle at the intersection a night earlier.
Caitlin O'Hara for The New York Times

Reminders (I)

(Discrete sets of events)

- Sum rule $p(X) = \sum_Y p(X|Y)$
- Product rule $p(X, Y) = p(X|Y)p(Y)$
- Bayes theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

(continuous sets of events)

- density $p(x)$,

$$p(x \in [a, b]) = \int_a^b p(x)dx, \quad p(x) = \int p(x, y)dy$$

Reminders (II)

- Cumulative distribution Function (CDF) $F(z) = \int_{-\infty}^{z} p(x)\, dx$
- Expectation $\mathbb{E}[x] = \int x p(x) dx$, $\mathbb{E}[x] = \sum_i x_i p(x_i)$
- Conditional expectation $\mathbb{E}_x f(x|y) = \sum_x f(x) p(x|y)$
- Variance $\mathrm{Var}[x] \equiv \mathbb{E}\left\{(x - \mathbb{E}x)^2\right\}$
- Covariance $\mathrm{Cov}[x, y] \equiv \mathbb{E}\left\{(x - \mathbb{E}x)(y - \mathbb{E}y)\right\}$

Reminders (III) A few important distributions

- The gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Uniform distribution: $P(y) = \frac{1}{|b-a|}, \quad y \in [a,b]$

- χ^2 distribution: $\chi^2 \sim \sum_{i=1}^{N} {Z_i}^2$ with Z_i independent standard normal RV.

Reminders (IV) A few important distributions

- Binary variables: Bernoulli and Rademacher,

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}, \quad x = \begin{cases} 1 \\ 0 \end{cases}, 0 \leq \mu \leq 1$$

$$\text{Rademacher:} \quad \varepsilon(x) = \begin{cases} (1/2), & x = +1 \\ (1/2), & x = -1 \\ 0, & \text{otherwise} \end{cases}$$

The exponential family

- Many of the distributions we have discussed are part of a general family called The exponential family
- The exponential family has many interesting properties
  - It is the only family of distribution with finite-sized sufficient statistics (see next slides)
  - It is the only family with known conjugate priors
  - It is at the core of generalized linear models
  - it is at the core of variational inference
- We will come back to these notions later

The exponential family

- A pdf $p(\boldsymbol{x}|\theta)$ is said to be in the exponential family for $\boldsymbol{x} = (x_1, \ldots, x_m)$ and $\theta \subseteq \mathbb{R}^d$ if

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp(\theta^T \phi(\boldsymbol{x}) \\ &= h(\boldsymbol{x}) \exp(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})) \end{aligned}$$

- $Z(\theta)$ and $A(\theta)$ are defined as

$$\begin{aligned} Z(\theta) &= \int_{\mathcal{X}^m} h(\boldsymbol{x}) \exp[\theta^T \phi(x)] \, dx \\ A(\theta) &= \log(Z(\theta)) \end{aligned}$$

- $Z(\theta)$ is called the partition function, θ are the mutual parameters, $\phi(x) \in \mathbb{R}^d$ is the vector of sufficient statistics, $A(\theta)$ is the log partition function or cumulant function.

The exponential family

- Two examples
  - Bernoulli

$$\begin{aligned}\text{Ber}(x|\mu) &= \mu^x(1-\mu)^{1-x} = \exp(x\log(\mu) + (1-x)\log(1-\mu)) \\ &= \exp(\phi(x)^T\theta)\end{aligned}$$

  - Univariate Gaussian

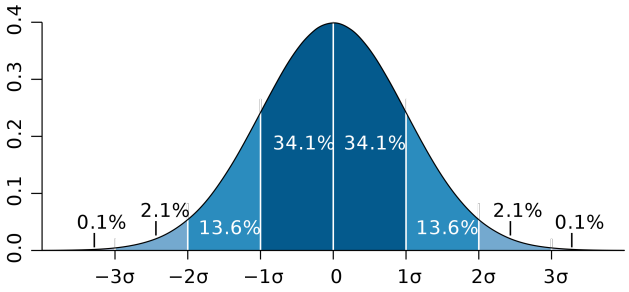
- The Uniform distribution does not belong to the exponential family

Parameter inference: What does it mean?

(Unknown) Population

Known sample

(Approximate) Distribution



Prediction

Background vector created by freepik - www.freepik.com

Parameter/model inference: Bayesian vs frequentist

- Several models in ML are special instance of a more general idea called model inference
- Inference can be used in both supervised (learn new labels from training labels) and unsupervised (learn parameters from distribution) frameworks
- Two main approaches: frequentist and Bayesian.

Parameter/model inference: Bayesian vs frequentist

- Bayesian statistics.
  - Considers the (distribution) parameters as random
  - Relies heavily on the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$
  - dominated statistical practice before 20^{th} century
  - Ex: MAP $\underset{\theta}{\text{argmax}}\, P(\mathcal{D}|\theta)P(\theta)$
- Frequentist statistics (a.k.a classical stat.)
  - Parameters $\boldsymbol{\theta}$ viewed as fixed, sample $\mathcal{D}$ as random (Randomness in the data affects the posterior)
  - Relies on the likelihood or some other function of the data
  - dominated statistical practice during 20^{th} century
  - Ex. MLE : $\underset{\theta}{\text{argmax}}\, P(\mathcal{D}|\theta)$

Bayesian statistics: Some vocabulary

- We saw Bayesian inference relies on the posterior $p(\theta|\mathcal{D})$
- The posterior reads from the Bayes rule as

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

- $p(\theta)$ is called the prior, $p(\mathcal{D}|\theta)$ is called the likelihood function and $Z = p(\mathcal{D})$ is the normalizing constant (independent of θ)
- Given a set of patterns $(\boldsymbol{x}_\mu, \boldsymbol{y}_\mu)$, probabilistic classifiers are usually of two types:
  - Generative (learn model for $p(\boldsymbol{x}, \boldsymbol{y}|\theta)$)
  - Discriminative (learn model for $p(\boldsymbol{y}|\boldsymbol{x}, \theta)$)

Bayesian statistics: Some vocabulary

- An example of discriminative classifier : Logistic regression
  - Here we take $\mu(\boldsymbol{x}) = \text{sigm}(\boldsymbol{w}^T \boldsymbol{x})$ and define the classifier as a Bernoulli distribution

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Ber}(y|\mu(\boldsymbol{x}))$$

  - Good when the output is binary

- An example of generative classifier :
  - relies on the assumption that the features (hidden variables) are independent

$$p(\boldsymbol{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} p(x_j|y = c, \theta_{jc})$$

  - $\theta_{j,c}$ is the parameters of the distribution of class c for j^{th} entry in the D-dimensional pattern vector $\boldsymbol{x} \in \{1, \ldots, K\}^D$.

- We will study those models in further detail when discussing classifiers.

Bayesian statistics

- In Bayesian statistics, randomness is most often used to encode uncertainty
- The posterior $p(\boldsymbol{\theta}|\mathcal{D})$ summarizes all we know on the parameters
- Bayesian inference is not always the right choice because of the following
  - The Mode is not a typical point in the distribution
  - MAP estimator depends on re-parametrization

Bayesian statistics: Drawbacks and solutions

- A solution to the first part is to use a more robust loss function $\ell(\hat{\theta}, \theta)$
- A solution to the second part is to replace the MAP with the following estimator (when available)

$$\hat{\theta} = \underset{\theta}{\text{argmax}} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) |\boldsymbol{I}(\boldsymbol{\theta})|^{-1/2} \tag{1}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fischer information matrix

Fischer information matrix

- For a generative model $p(\boldsymbol{x}|\theta)$, we let $g(\theta, \boldsymbol{x})$ denote the Fischer score

$$g(\theta, \boldsymbol{x}) = \nabla_\theta \log(p(\boldsymbol{x}|\theta))$$

- the Fischer Kernel is the defined as

$$k(\boldsymbol{x}, \boldsymbol{x}') = g(\boldsymbol{\theta}, \boldsymbol{x})^T \boldsymbol{F}^{-1} g(\boldsymbol{\theta}, \boldsymbol{x}')$$

- The matrix $\boldsymbol{F}$ is called the Fischer matrix and defined as

$$\boldsymbol{F} = \mathbb{E}_{\boldsymbol{x}} \left\{ g(\boldsymbol{\theta}, \boldsymbol{x}) g(\boldsymbol{\theta}, \boldsymbol{x})^T \right\}$$

- Note that it is often computed empirically as

$$\boldsymbol{F} \approx \frac{1}{N} \sum_{n=1}^{N} g(\theta, \boldsymbol{x}) g(\boldsymbol{\theta}, \boldsymbol{x})^T$$

Occam's razor and Model selection

- Only looking for the best model often leads to overfitting (we will see that later in more details)
- Bayesian framework offers and alternative called Bayesian model selection
- For a family of models, we can express the posterior

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(m|\mathcal{D})} \propto p(\mathcal{D}|m)p(m)$$

where $p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$ is called the marginal likelihood, integrated likelihood or evidence

Occam's razor

- Integrating the parameters $\boldsymbol{\theta}$ such as in

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

  acts as a natural regularization and prevents overfitting when solving for $\max_m p(m|\mathcal{D})$. This idea is known as Bayesian Occam's razor

- The evidence $p(\mathcal{D}|m)$ can be understood as the probability to generate a particular dataset from a family of model (all values of the parameters included).

- When the family of models is too simple, or too complex, this probability will be low.

Bayesian decision theory

- How do we resolve the lack of robustness of Bayesian estimators vis a vis the distribution (recall the bimodal distribution)?
- Statistical decision theory can be viewed as a game against nature.
- Nature has a parameter value in mind and gives us a sample
- We then have to guess what the value of the parameter is by choosing an action a
- As an additional piece of information, we also get a feedback from a loss function $L(y, a)$ which tells us how compatible our action is w.r.t Nature's hidden state.

Bayesian decision theory

- The goal of the game is to determine the optimal decision procedure,

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}} \, \mathbb{E}\left\{L(y, a)\right\}$$

- In economics $L(y, a) = U(y, a)$ and leads to the Maximum utility principle which is considered as rational behavior

$$\delta(x) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathbb{E}\left\{U(y, a)\right\}$$

- In the Bayesian framework, we want to minimize the loss over the models compatible with the observations $\{\boldsymbol{x}_\mu\}$

$$\delta(x) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \mathbb{E}_{p(\theta|\{\boldsymbol{x}_\mu\})}\left\{L(\theta, a)\right\} = \sum_{\theta \in \Theta} L(\theta, a) p(\theta|\left\{\boldsymbol{x}_\mu\right\}_\mu)$$

Bayesian decision theory (continued)

- The MAP is equivalent to minimizing a 0/1 loss

$$L(\hat{\theta}, \theta) = \mathbb{1}_{\theta \neq \hat{\theta}} = \begin{cases} 0 & \text{if } \hat{\theta} \neq \theta \\ 1 & \text{if } \hat{\theta} = \theta. \end{cases}$$

  we then have

$$\mathbb{E}L(\hat{\theta}, \theta) = p(\theta \neq \hat{\theta} | \{\boldsymbol{x}_\mu\}_\mu) = 1 - p(\hat{\theta} = \theta | \{\boldsymbol{x}_\mu\}_\mu)$$

  which is maximized for $\hat{\theta} = \theta$ with θ taken as

$$\theta^*(\{\boldsymbol{x}_\mu\}_\mu) = \underset{\hat{\theta}}{\text{argmax}}\, p(\theta | \{\boldsymbol{x}_\mu\}_\mu)$$

- The 0/1 loss means that each time you have an outcome different from your estimator, you are maximally penalized

What other losses can we choose?

- Is it possible to take more robust losses?
- ℓ_2 loss, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|^2$ gives posterior mean

$$\mathbb{E}\left\{(\hat{\theta} - \theta)^2 | \boldsymbol{x}_\mu\right\} = \mathbb{E}[\theta^2 | \boldsymbol{x}_\mu] - 2\hat{\theta}\mathbb{E}[\theta | \boldsymbol{x}_\mu] + \hat{\theta}^2$$

- setting derivative to 0, $\partial_{\hat{\theta}}\mathbb{E}\{\hat{\theta} | \boldsymbol{x}_\mu\} = 0$, we get

$$-2\mathbb{E}\left\{\theta | \boldsymbol{x}_\mu\right\} + 2\hat{\theta} = 0$$

$$\hat{\theta} = \int \theta p(\theta | \boldsymbol{x}_\mu)\, d\theta$$

What other losses can we choose? (continued)

- Is it possible to take more robust losses?
- ℓ_1 loss, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ gives posterior median
- The value $\hat{\theta}$ such that

$$p(\theta < \hat{\theta} | \boldsymbol{x}_\mu) = p(\theta \geq \hat{\theta} | \boldsymbol{x}_\mu) = 1/2$$

What other losses can we choose? (continued)

- Now assume $\hat{\theta}$ defines the value of some hidden variable y (e.g. the class of a point $\boldsymbol{x}_\mu$ defined by a gaussian mixture $\hat{\boldsymbol{\theta}}$).
- Finding the optimal parameters (or equivalently estimate the hidden state) can be done by considering the error

$$\begin{aligned} L_g(\theta, \hat{\theta}) &= \mathbb{E}_{(\boldsymbol{x}_\mu, y_\mu) \sim p(\boldsymbol{x}_\mu, y_\mu | \theta)} \left\{ \ell(\theta, f_{\hat{\theta}}) \right\} \\ &= \sum_{\boldsymbol{x}_\mu} \sum_{y_\mu} \ell(y_\mu, f_{\hat{\theta}}(x_\mu)) p(x_\mu, y_\mu | \boldsymbol{\theta}) \end{aligned}$$

- The Bayesian approach then minimizes the posterior expected loss

$$\underset{\hat{\theta}}{\operatorname{argmin}} \int p(\theta | \mathcal{D}) L_g(\theta, \hat{\theta}) \, d\theta$$

- Note that here the model is fixed and we want to learn the parameters ($><$ model selection)

How to pick up the priors?

- The controversial aspect of Bayesian statistics are the priors
- The main argument of Bayesians is that we most often know something about the world
- When it is possible, it makes things easier to pick up a prior from the same family as the likelihood function

Parameter/model inference: Biased vs Unbiased

- Imagine that we have access to a set of obervations $x_i \in \mathbb{R}^n$ and we can reasonably assume those samples are drawn independently from gaussian distributions.
- Because the observations are i.i.d, we can write the expression for the probability of oberving the x_i given the common μ and σ^2,

$$p(x|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \tag{2}$$

- A reasonably good idea to derive estimates for μ and σ is then to maximize this likelihood

Parameter/model inference: Biased vs Unbiased

- Since the log is a monotonically increasing function,

$$\underset{\mu,\sigma^2}{\text{argmax}} \quad p(x|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

is equivalent to maximizing the log likelihood function

$$\underset{\mu,\sigma^2}{\text{argmax}} \quad \log\left(p(x|\mu,\sigma^2)\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log(\sigma^2) - \frac{N}{2}\log(2\pi).$$

Parameter/model inference: Biased vs Unbiased

- Maximizing the log likelihood function with respect to μ first and then σ^2 gives the maximum likelihood estimators

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2$$

- Those two estimates are functions of the data set $x_1, \ldots, x_N$

Parameter/model inference: Biased vs Unbiased

- Remember the ML estimators

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\hat{\sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2$$

- Now take the expectation of those estimators with respect to the known distribution,

$$\mathbb{E}\hat{\mu}_{ML} = \mu$$

$$\mathbb{E}\hat{\sigma}^2_{ML} = \left(\frac{N-1}{N}\right)\sigma^2$$

Parameter/model inference: Biased vs Unbiased

- On average, the MLE will get you the right mean, but will underestimate the variance

$$\mathbb{E}\hat{\mu}_{ML} = \mu$$
$$\mathbb{E}\hat{\sigma}^2_{ML} = \left(\frac{N-1}{N}\right)\sigma^2$$

- This problem is called bias and is related to the problem of overfitting

- In fact it turns out that a better estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2$$