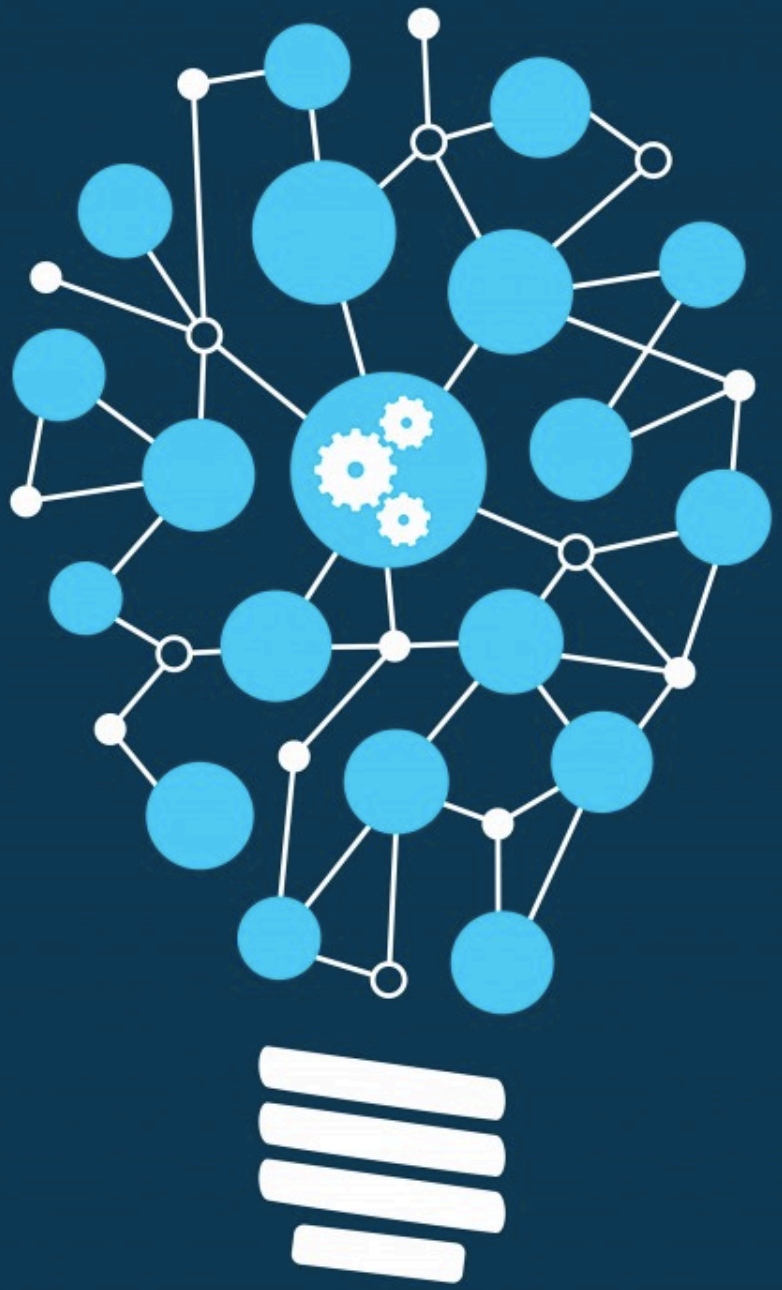
MACHINE LEARNING



Basic Concepts Of Machine Learning

PhD Nussipbekov Abay

Why ML?

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates)

- “Machine learning is the next Internet” (Tony Tether, former director, DARPA)

- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, former CTO, Sun)

- Artificial intelligence is the new electricity (prof. Andrew Ng, former chief scientist at Baidu, Coursera co-founder)

Why ML?

- The time when you can just develop an application which will perform some automation is gone. Now you have to add some “intelligence”.
- ML hype at Silicon Valley stratups
- Jeremy Howard TED talk on how machine learning is changing this world.
  - Read, write, speak, hear better than human
  - The majority of workplaces will be taken by AI machines.

Why ML?

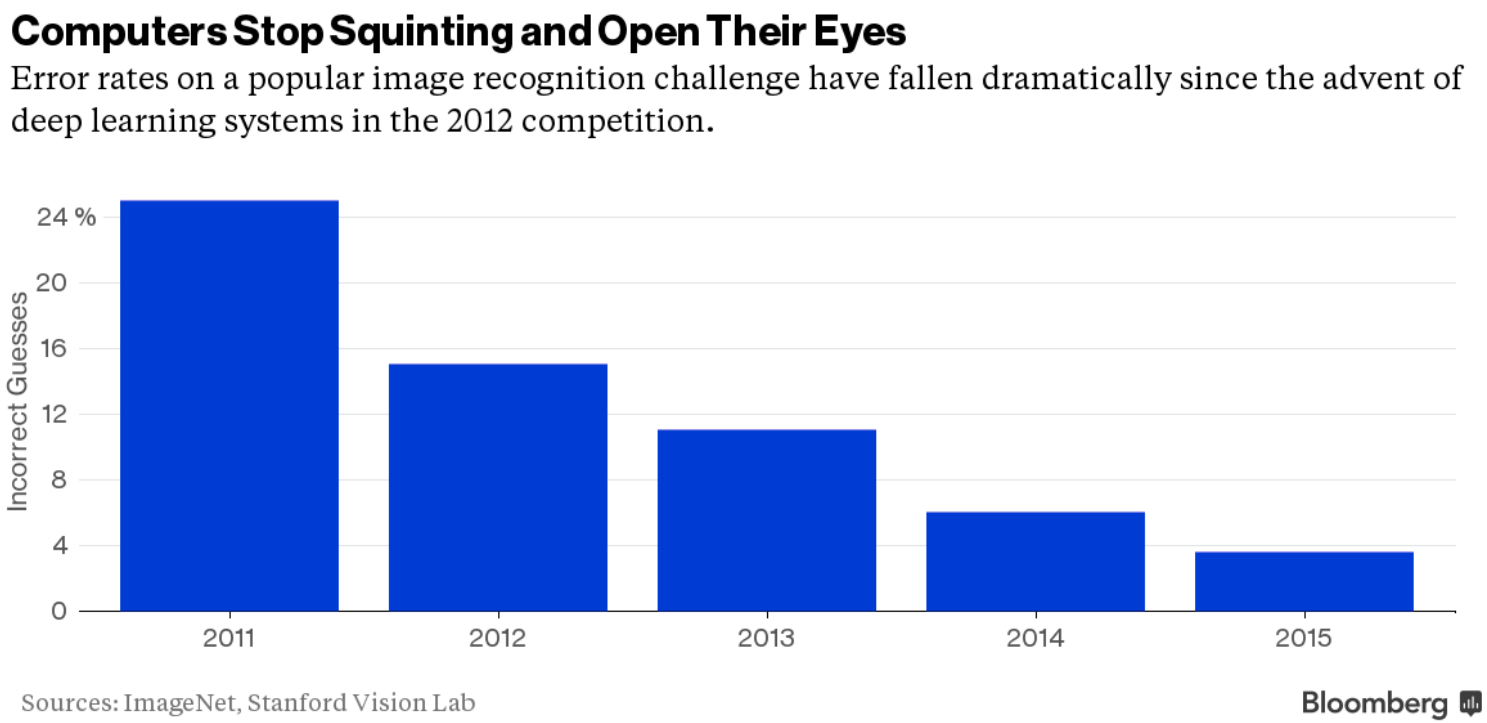
- Now is the era of *knowledge work:*
  - Not just "move this from here to there" but "maximize profits", "minimize risks", "find the best strategy" and etc.
- Hal Varian, chief economist at Google, said:
- *"I keep saying the hottest job in the next ten years will be data scientists. The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."*
- With so much of the economic activity dependent on information, you can't afford to be lost in the data. Machine learning will help you get through all the data and extract some information (insight).

Why ML?

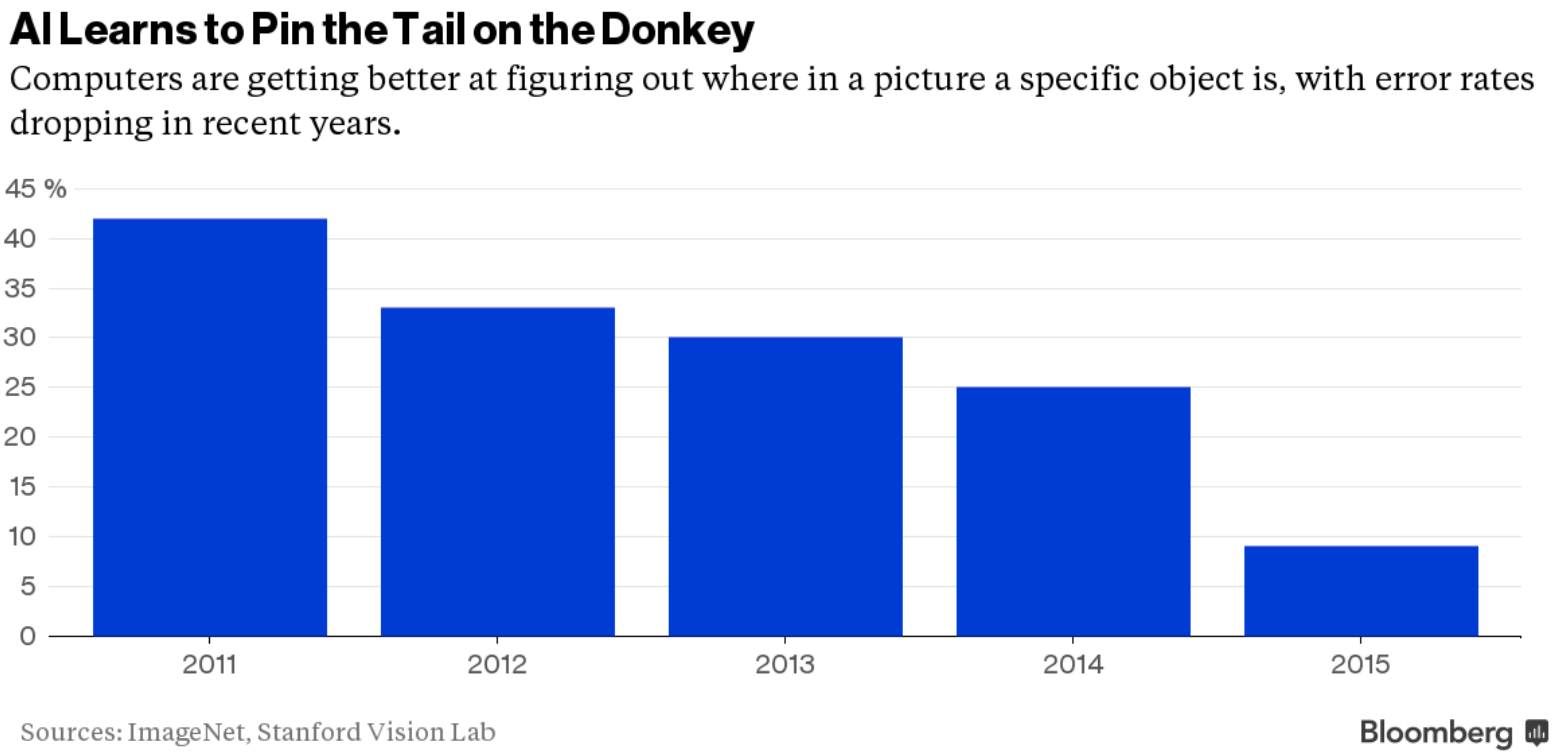
- According to the McKinsey Global Institute (In a May 2011 report): “By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.”

data
science
research
scientists
faculty
environment
fellows
activities
institute
program
support
space
education
UW
bridge
project
section
director
evaluation
graduate
sciences
activity
incubator
escience
students
tools
reproducibility
projects
studio
engineering
career
university
one
budget
metrics
methodology
committee
social
scientific
researchers
working
also
effort
establish
established
campus
participants
impact
year
time
training
participate
paths
software
courses
number
development
appendix
theme
computer
new
work
funded
professor
areas
across
collaborations
focused
area
additional
office
best
domain
among
three
techniques
current
assessed
well
range
many
executive
goal
team
leaders
methodologies
anticipate
annually
rules
individuals
described
breakthrough
see
alternative
open
use
per
community
include
staff
common
people
collaboration
visiting
aligned
administrative
seed
practices
computing
online
success
fields
libraries
events
funds
fellow
permanent
years
long-term
environments
successful
postdocs
flagship
ethnography
washington
steering
positions
universities
initiative
programs
methods
proposal
culture
leaders
center
educational
physical
statistics
groups
made
process
external
rules
annual
funding
group
section
project
bridge
research
graduate
fellows
activities
program
institute
software
paths
appendix
new
work
funded
theme
computer
professor
development
number
incubator
activity
sciences
science
scientists
support
faculty
projects
reproducibility
students
tools
engineering
studio
career
university
one
budget
metrics
environment
methodology
committee
social
scientific
across
areas
researchers
working
also
effort
space
education
UW
director
evaluation
establish
established
campus
participants
impact
year
time
including
training
participate
paths
software
courses
number
development
appendix
theme
computer
new
work
funded
professor
activities
program
institute
fellows
sciences
activity
incubator
escience
bridge
graduate
research
scientists
support
faculty
projects
reproducibility
students
tools
engineering
studio
career
university
one
budget
metrics
environment
methodology
committee
social
scientific
across
areas
collaborations
section
project
director
evaluation
establish
established
campus
participants
impact
year
time
including
training
participate
paths
software
courses
number
development
appendix
theme
computer
new
work
funded
professor
activities
program
institute
fellows
sciences
activity
incubator
escience
bridge
graduate
research

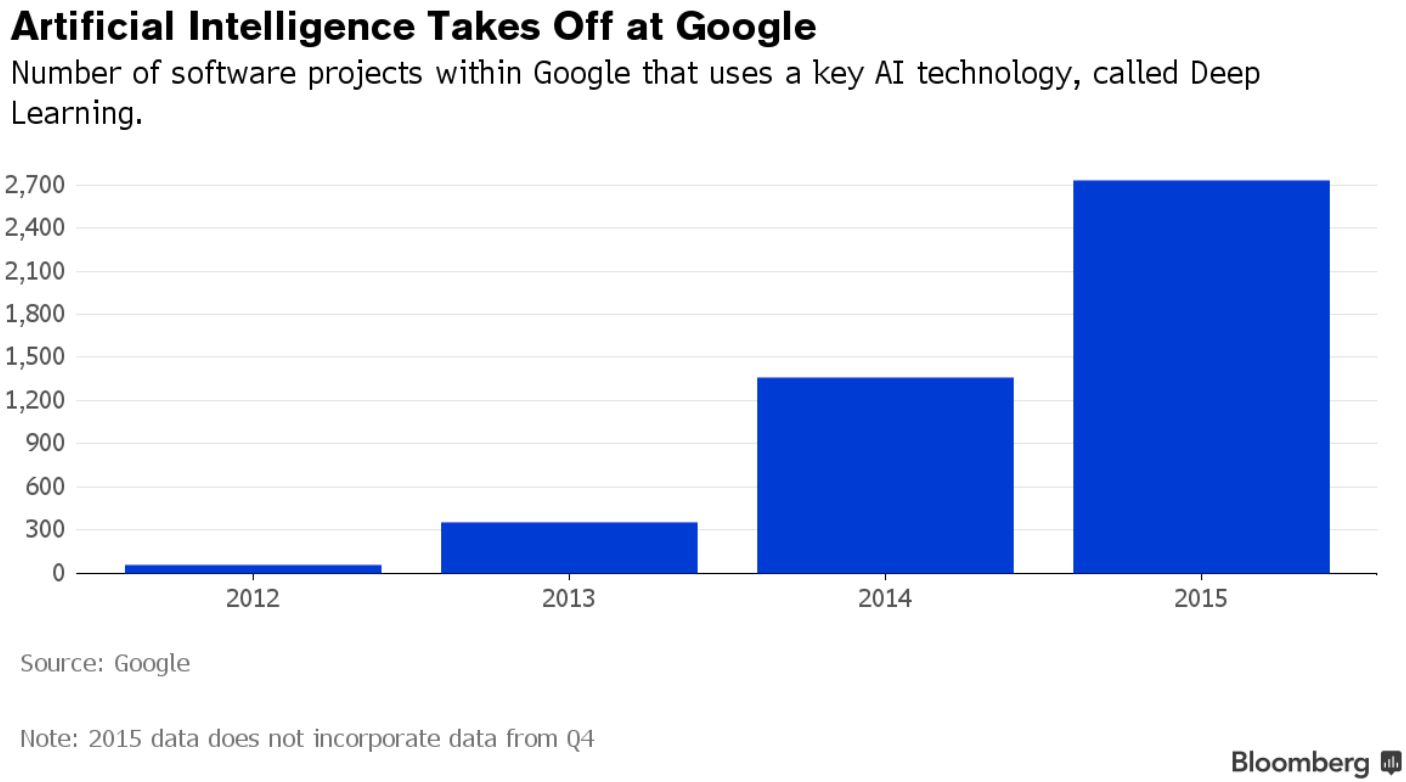
Why ML?


Computers Stop Squinting and Open Their Eyes
Error rates on a popular image recognition challenge have fallen dramatically since the advent of deep learning systems in the 2012 competition.
Incorrect Guesses
24 %
20
16
12
8
4
0
2011
2012
2013
2014
2015
Sources: ImageNet, Stanford Vision Lab
Bloomberg

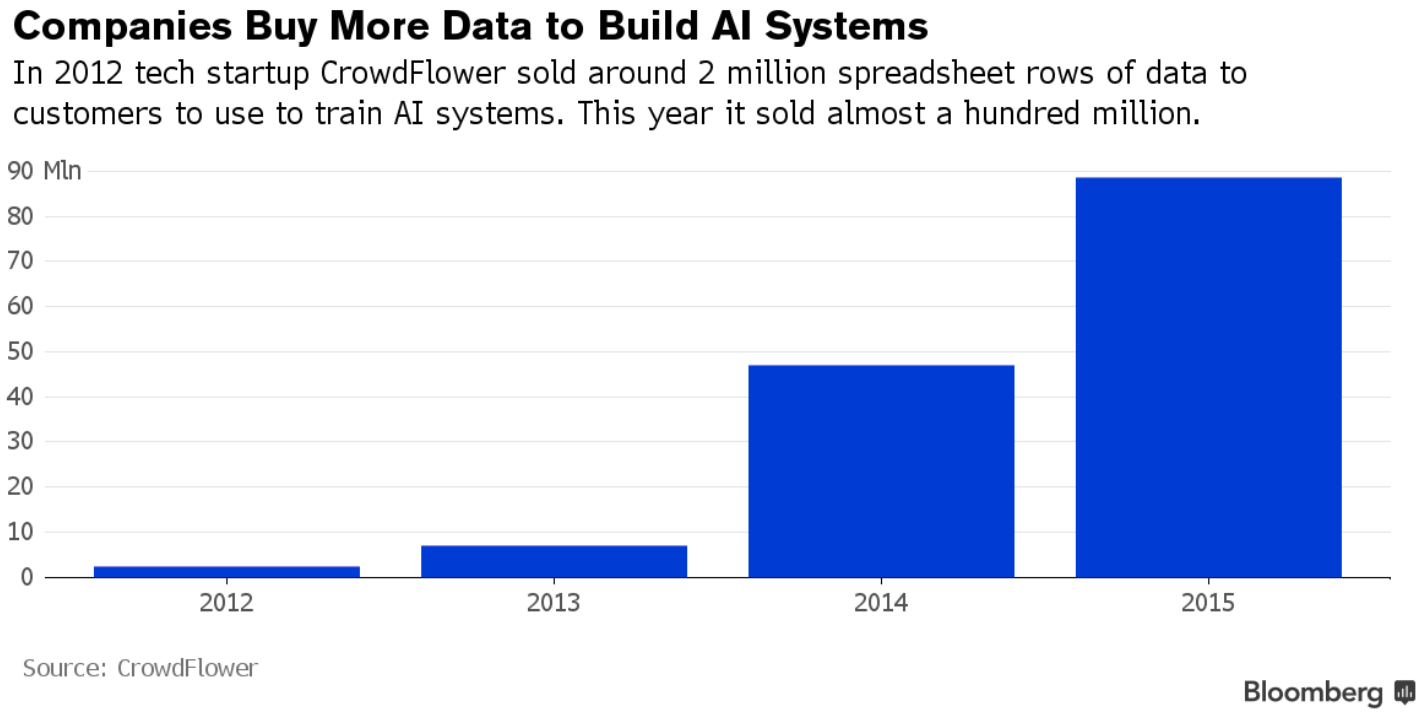
Why ML?


AI Learns to Pin the Tail on the Donkey
Computers are getting better at figuring out where in a picture a specific object is, with error rates dropping in recent years.
45 %
40
35
30
25
20
15
10
5
0
2011
2012
2013
2014
2015
Sources: ImageNet, Stanford Vision Lab
Bloomberg

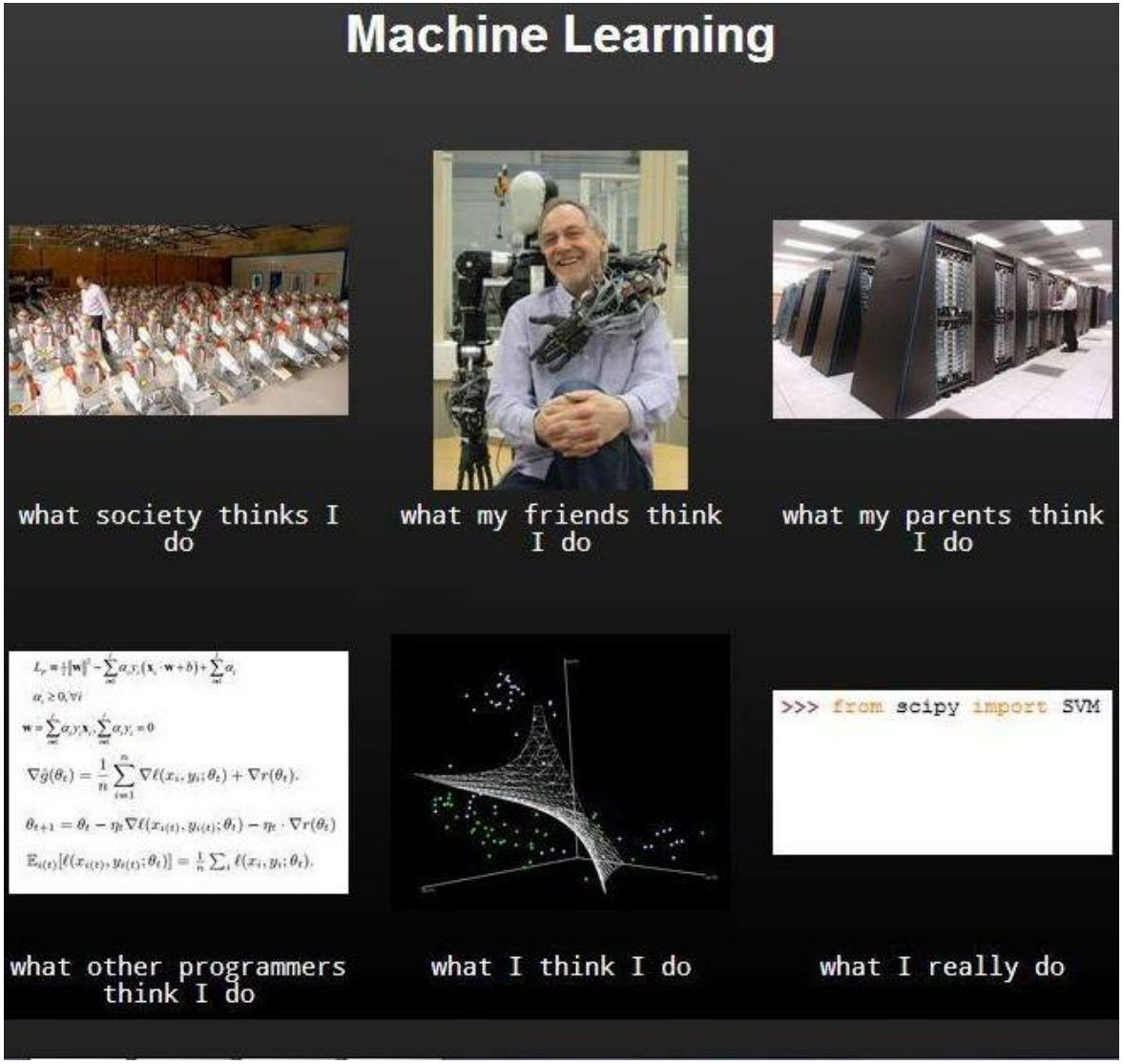
Why ML?


Artificial Intelligence Takes Off at Google
Number of software projects within Google that uses a key AI technology, called Deep Learning.
2,700
2,400
2,100
1,800
1,500
1,200
900
600
300
0
2012
2013
2014
2015
Source: Google
Note: 2015 data does not incorporate data from Q4
Bloomberg

Why ML?


Companies Buy More Data to Build AI Systems
In 2012 tech startup CrowdFlower sold around 2 million spreadsheet rows of data to customers to use to train AI systems. This year it sold almost a hundred million.
90 Mln
80
70
60
50
40
30
20
10
0
2012
2013
2014
2015
Source: CrowdFlower
Bloomberg

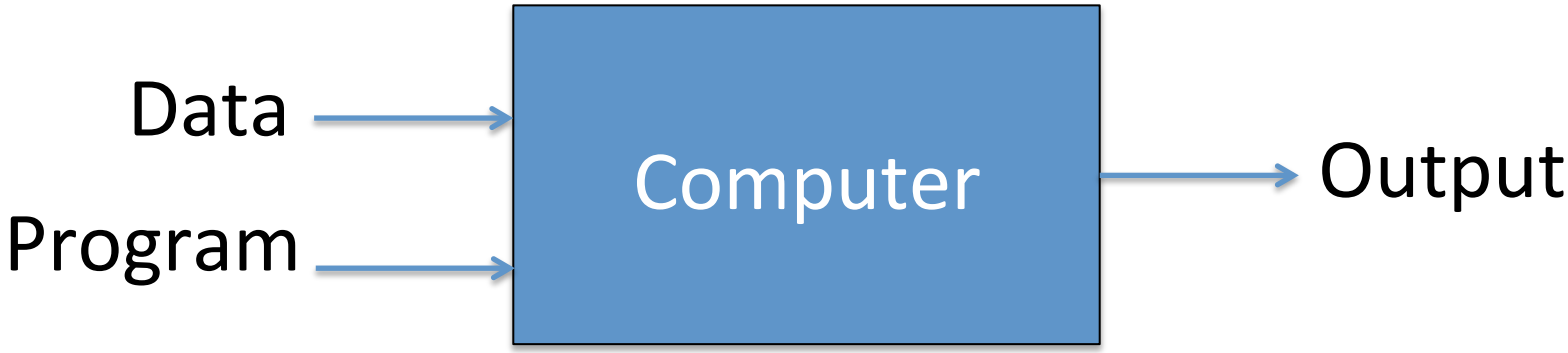
What is ML?


Machine Learning
what society thinks I do
what my friends think I do
what my parents think I do
$L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$
$\alpha_i \geq 0, \forall i$
$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{l} \alpha_i y_i = 0$
$\nabla \hat{g}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$
$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$
$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$
>>> from scipy import SVM
what other programmers think I do
what I think I do
what I really do

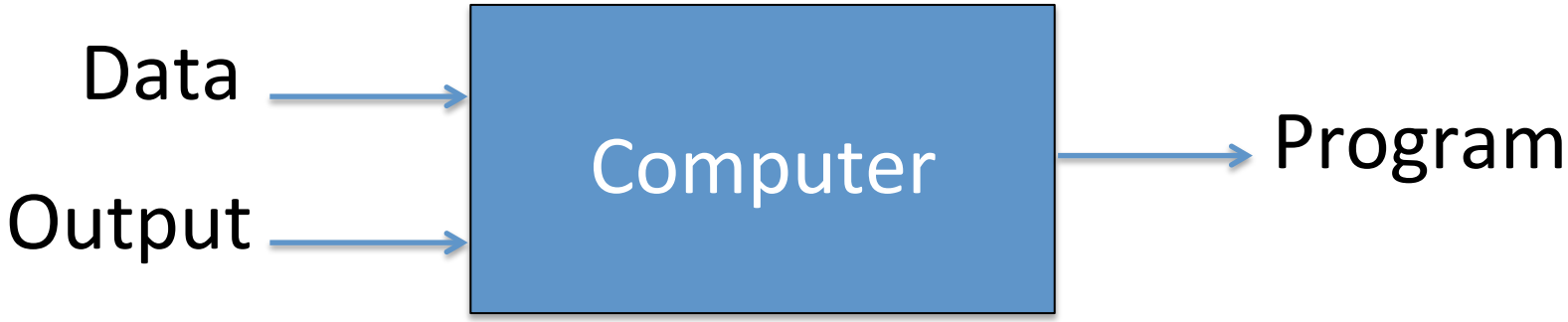
What is ML

- Machine learning is the science of getting computers to act without being explicitly programmed. (Asoc. Prof. Andrew Ng, Stanford University)
- Machine learning is about turning data into information. (Peter Harrington, chief scientists for Zillabyte Inc.)
- Machine learning is automating automation (Prof. Pedro Domingo, University of Washington)

Traditional Programming



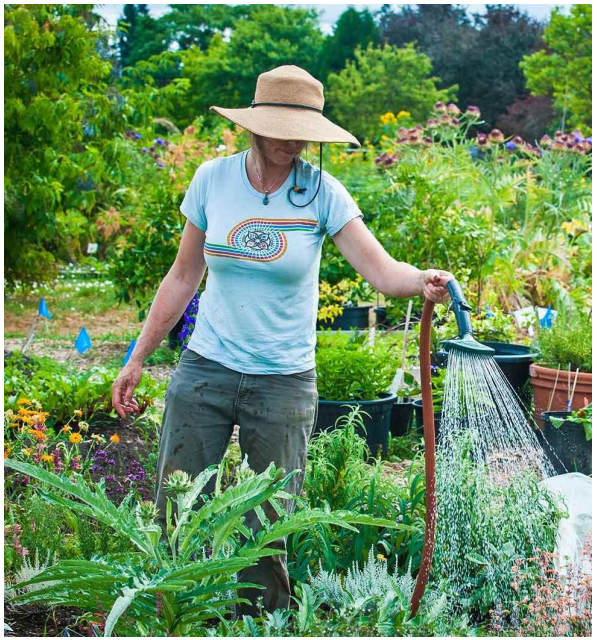
Machine Learning



✓ Getting computers program themselves

Analogy

- Seeds = algorithms
- Nutrients = data
- Gardener = you
- Plants = programs

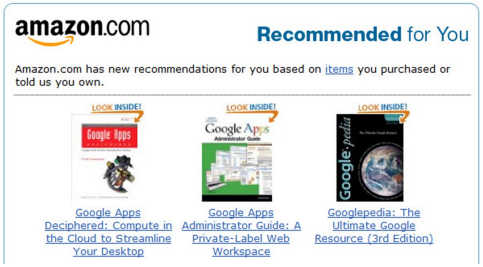


All the work is done by “nature”

Applications of ML

- Web search

- Recommendation systems

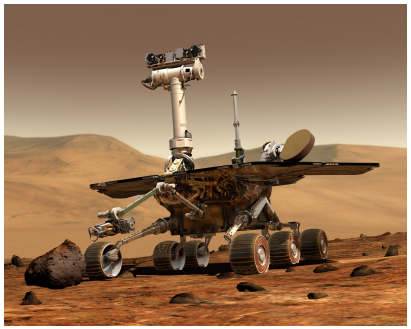


- Computational biology

- Finance

Applications of ML

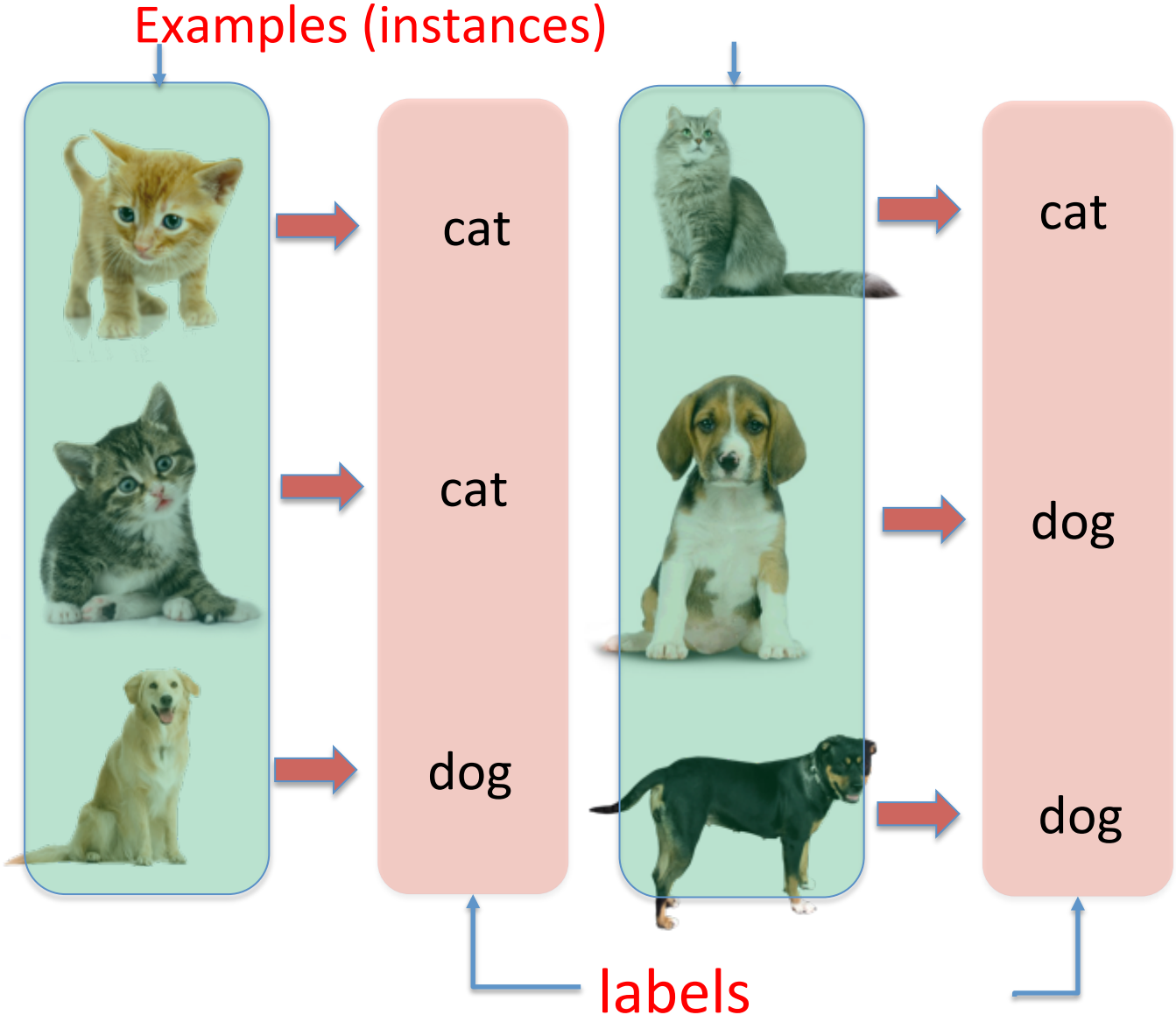
- Space exploration
- Robotics
- Social networks
- Many other…



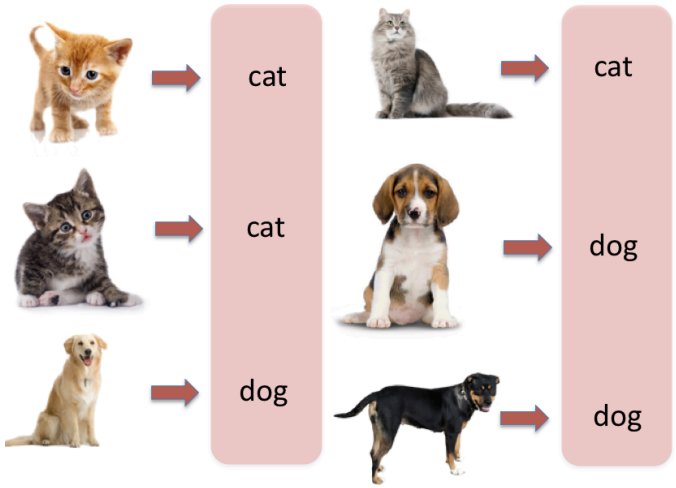
Types of learning

- Supervised learning
  - Training data includes desired outputs
  - Classification, regression
- Unsupervised learning
  - Training data does not include desired outputs
  - Find hidden structure in data
  - Clustering, dimensionality reduction
- Semi-supervised learning
  - Training data includes a few desired outputs
- Reinforcement learning
  - The learner interacts with the world via "actions" and tries to find an optimal policy of behavior with respect to "rewards"(feedback) it receives from the environment
  - Learning with a critic who merely states that something is right or wrong, but does not say specifically how it is wrong.

Supervised learning



Supervised learning



Generalization
via
induction

Supervised learning

Well learned model

cat

dog

Features

- Color (categorical features – list)
- Weight (real number)
- Number of legs (numerical number)
- Tail (binary value – boolean)

- Informative features
  - Muzzle lenght
  - Nose color
- Feature extraction cost
  - Distance between eyes

Dataset

		Weight (kg)	Muzzle length (cm)	Tail	Color	Label
Training set	1	10.1	3.0	Yes	black	cat
	2	15.7	7.0	Yes	gray	dog
	3	3.0	6.0	No	brown	dog
	4	3.5	3.5	Yes	white	cat
Testing set	5	3.0	4.5	Yes	yellow	dog
	6	4.0	2.5	Yes	gray	cat
		5.0	3.0	Yes	black	?

Each of the rows in table is an instance made up of features.

It is represented as a feature vector $x = \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix}$

Supervised learning

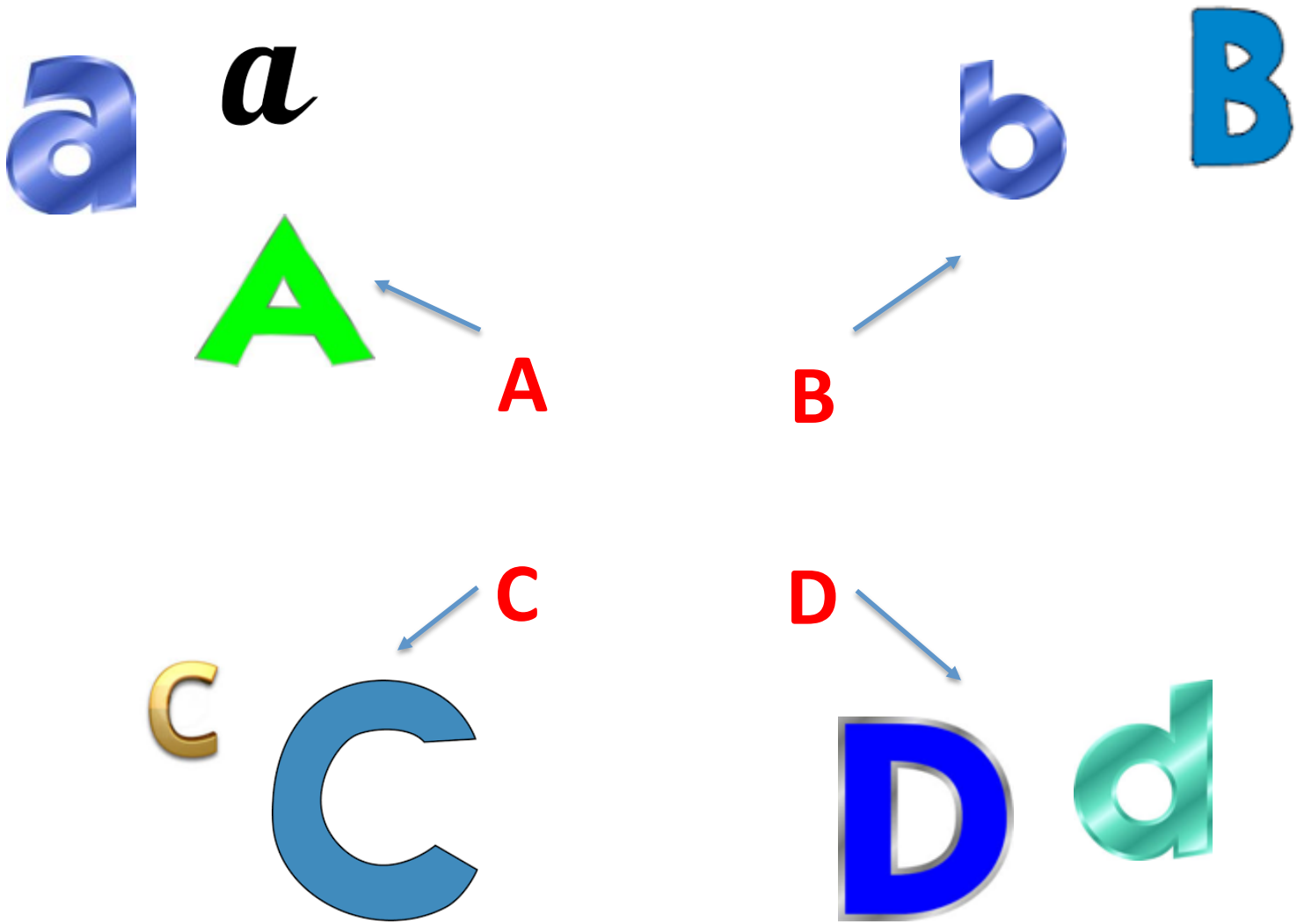
	Weight (kg)	Muzzle length (cm)	Tail	Color	Label
	5.0	3.0	Yes	black	?

- The target variable (label) is what we'll be trying to predict with our machine learning algorithms.
- In classification the target variable takes on a nominal value, and in the task of regression its value could be continuous.
- The machine learns by finding some relationship between the features and the target variable.

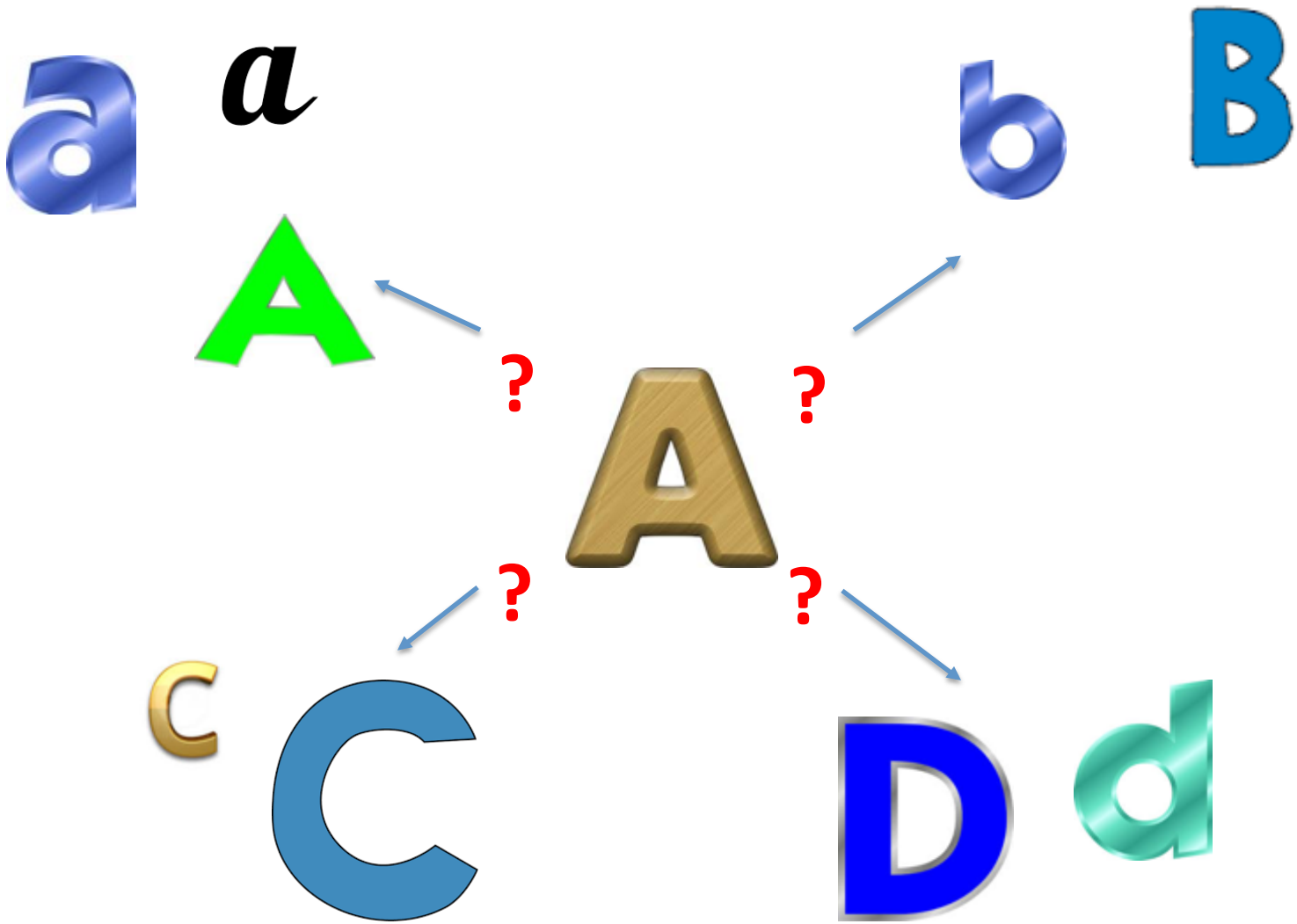
Supervised learning tasks

- Classification - predict what class an instance of data should fall into
- Regression prediction of numeric value.

Classification



Classification



Classification: detection

Regression

Input	1	2	3	4	5	6
Output	1	4	9	16	25	?

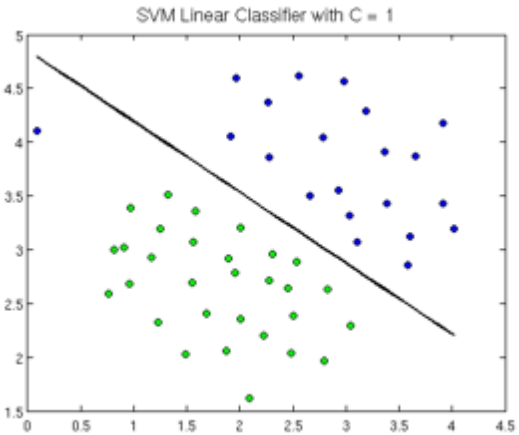
$$\hat{y} = f(x_1, \ldots, x_n)$$

Stock market

Temperature prediction

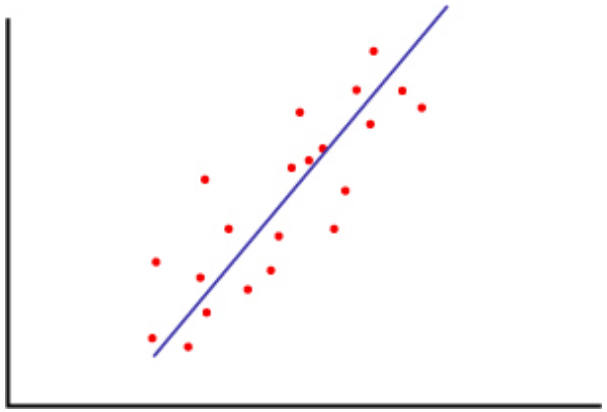
Classification vs. regression

classification



- find a decision boundary which separates groups
- output - class

regression

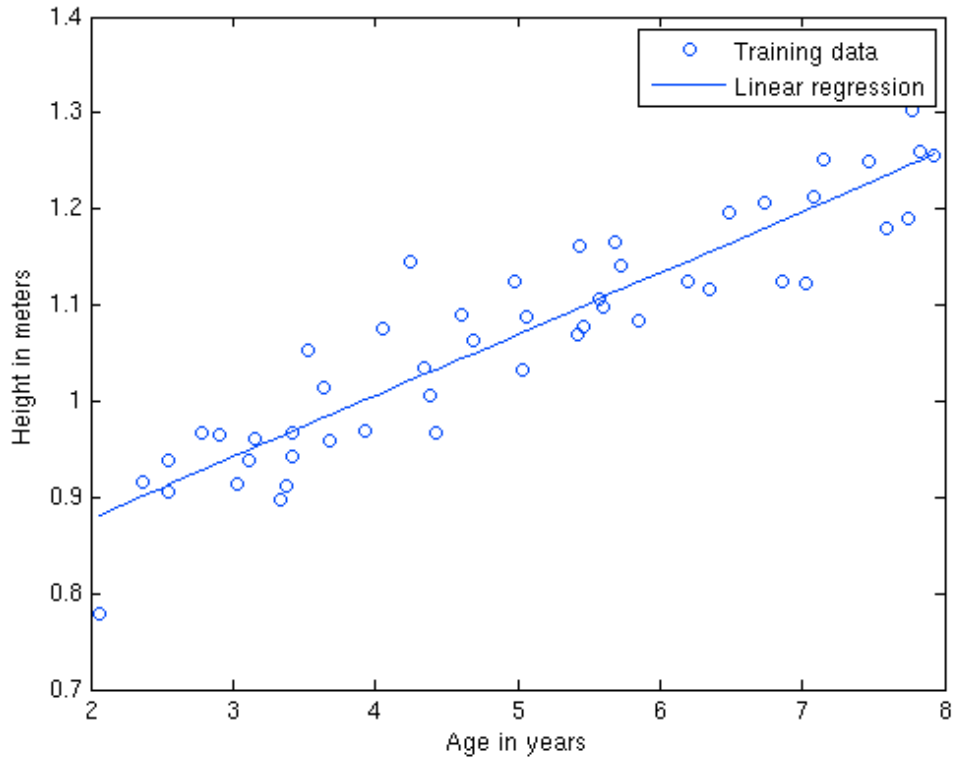


- find a line (function) which describes trend
- output – real number

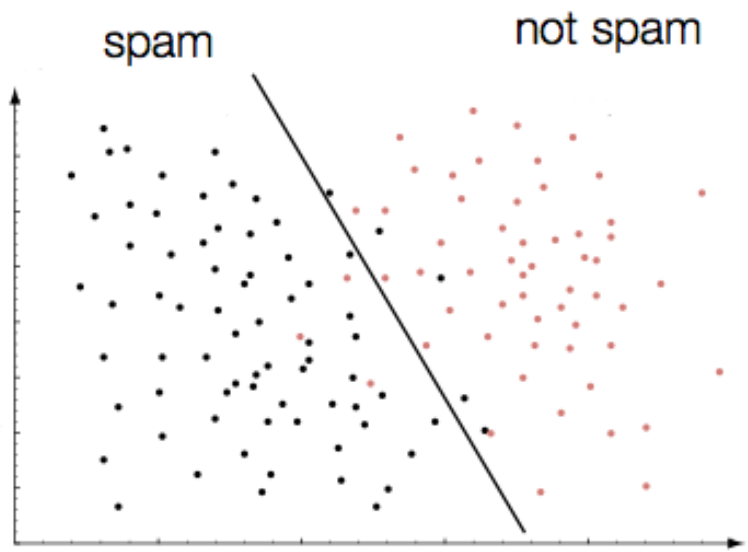
What is learning actually?

- Mostly learning means calculating model parameters.
- Model – function (f) that maps input variables (X) to output variables (Y).

$$Y = f(x)$$

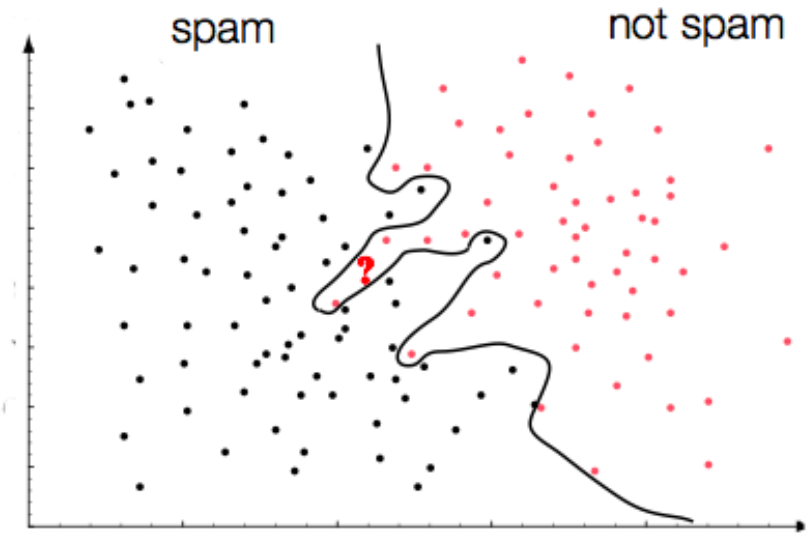


Generalization and overfitting



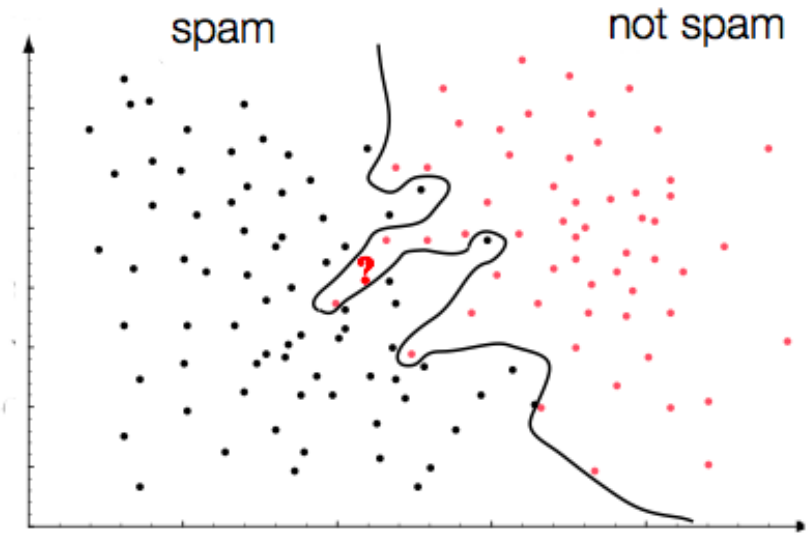
E-mails are divided by decision boundary into two classes (spam, not spam)

Generalization and overfitting



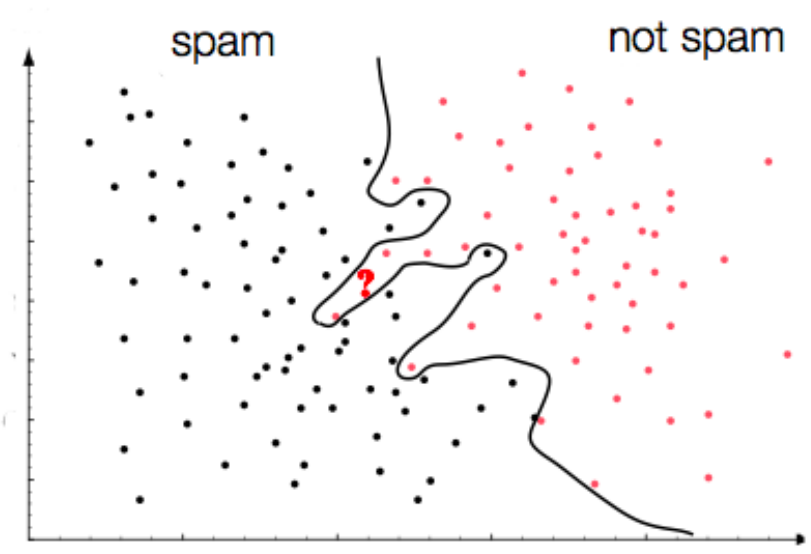
If our models is extremely complicated, our classifier would have a decision boundary more complex than the simple straight line. In that case all the training patterns would be separated perfectly.

Generalization and overfitting



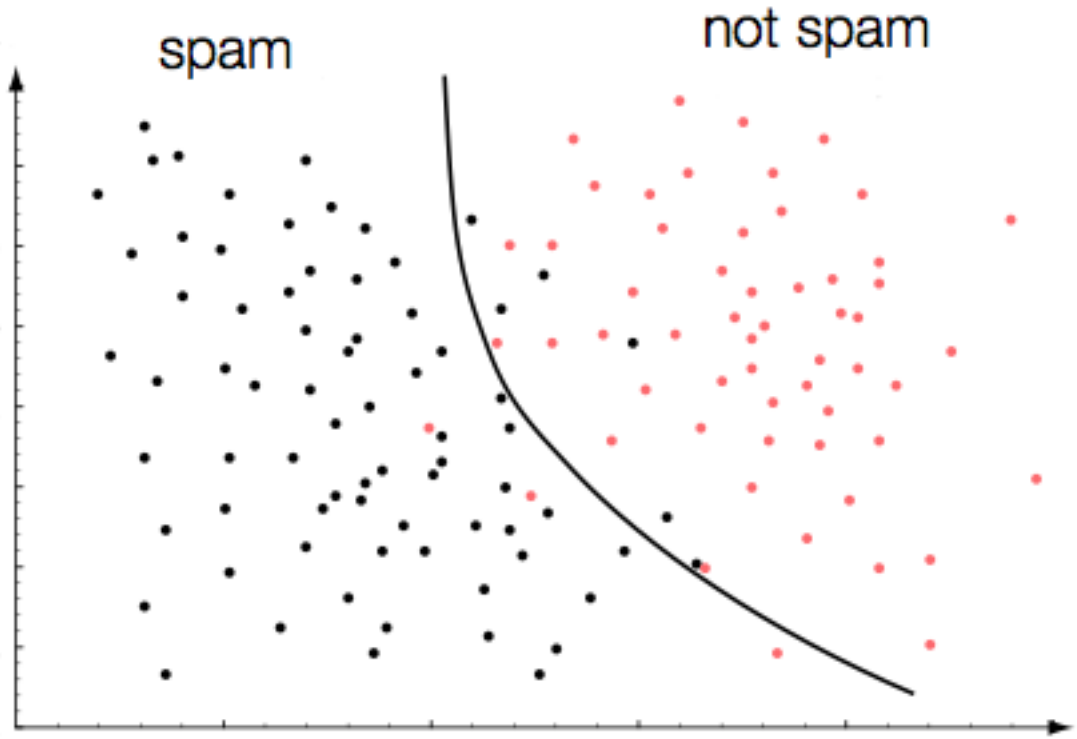
The central aim of designing a classifier is to suggest actions when presented with novel patterns, i.e., e-mail not yet seen. This is the issue of generalization. It is unlikely that the complex decision boundary in would provide good generalization, since it seems to be "tuned" to the particular training samples, rather than some underlying characteristics or true model of all the e-mails that will have to be separated.

Generalization and overfitting



This complicated model is called overfitted model.

Generalization and overfitting



May be “simplification” will be the solution and we might be satisfied with the slightly poorer performance on the training samples if it means that our classifier will have better performance on novel patterns.

Unsupervised learning

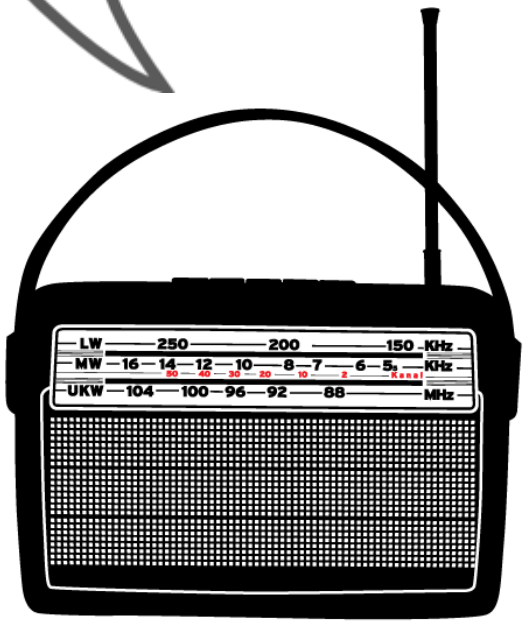
- Training set does not contain desired output
- Find hidden structure
- Clustering, dimensionality reduction

Unsupervised learning

Bizin online mağazasında iPhone 6 telefonun fiyatlar düştü

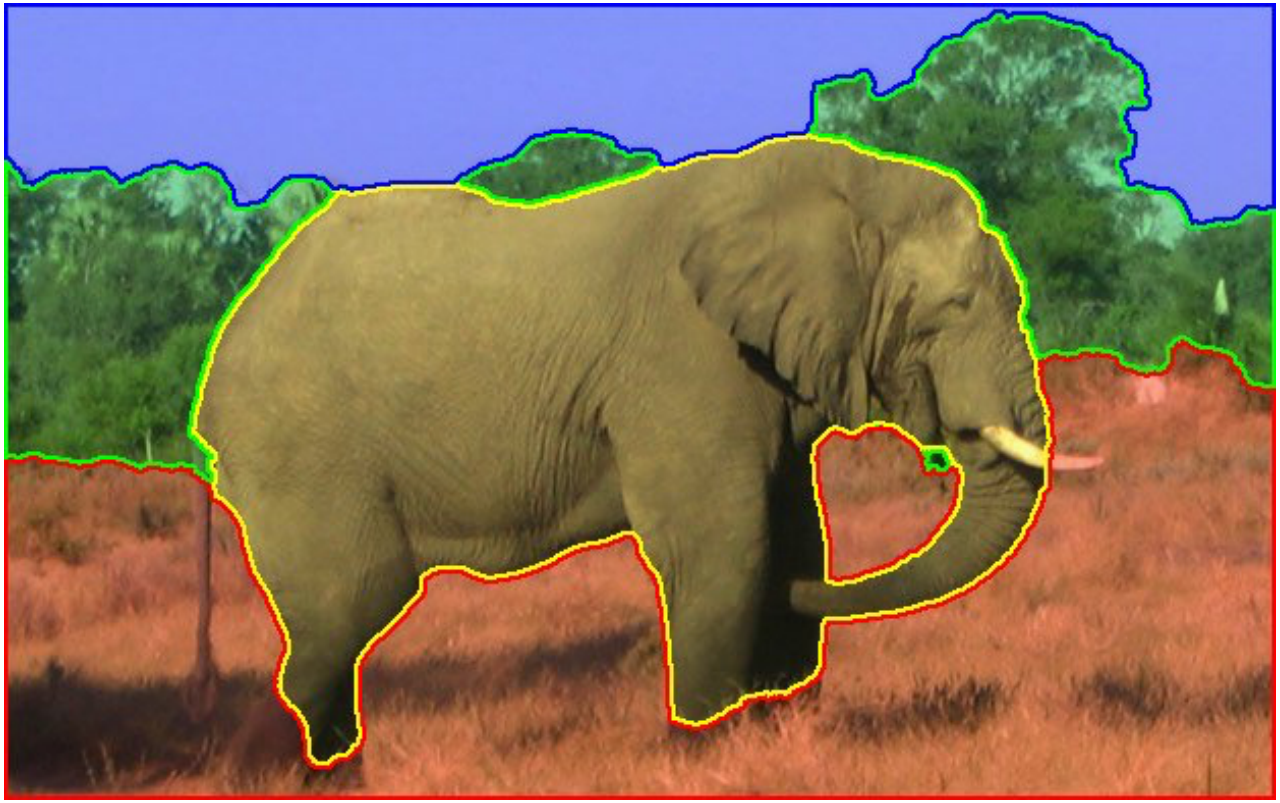
Merhaba değerli seyiriciler bügün biz makine öğrenme hakkında konuşacağız!

After listening thousands of podcasts you start recognizing patterns, expects particular sounds, understand from the context and etc.



Listening to podcasts on foreign language, which you do not understand –
you don't have dictionary, nobody will tell

Clustering



Clustering

← 2012 November 8, 2016 2020 →

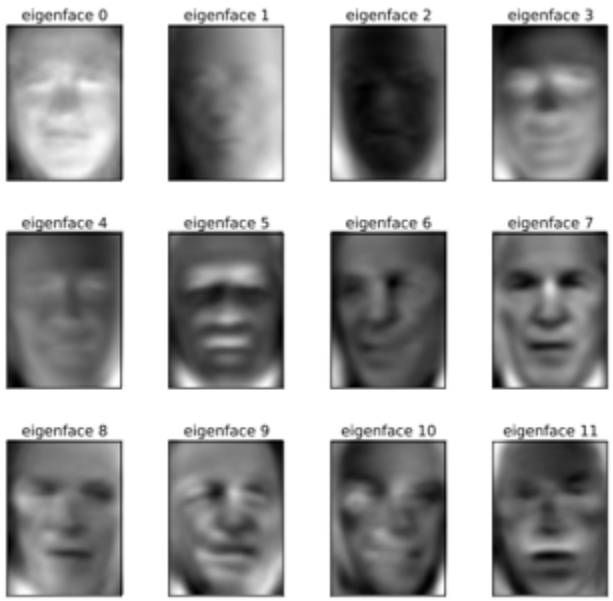
538 members of the Electoral College
270 electoral votes needed to win

Turnout	54.7% (estimated)[1] ▼ 0.2 pp	
Nominee	Donald Trump	Hillary Clinton
Party	Republican	Democratic
Home state	New York	New York
Running mate	Mike Pence	Tim Kaine
Electoral vote	304[a][2]	227[a][2]
States carried	30 + ME-02	20 + DC
Popular vote	62,984,825[3]	65,853,516[3]
Percentage	46.1%	48.2%

- Small group of people may change the result (microtargeting, Fornier, Sosnik and Dowd, *Applebee's America)*
- How to find that small group with limited budget?
- **Clustering**
  - Collect any information, which can influence their opinion
  - Use that information for clustering
  - Come up with a speech for each cluster and start a campaign

Dimensionality reduction

- Eigenfaces:
  - They way people and computers see faces
  - 20.000 features (pixels) -> 100 main components



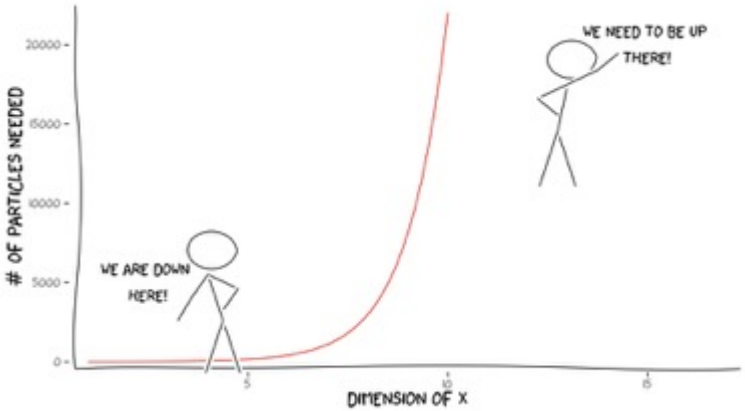
Dimensionality reduction

- Having a lot of information is really cool. Imagine facebook!
- BUT **much harder!**
- Dimensionality!

Dimensionality reduction

- Curse of dimensionality!

- If N objects for 1 feature,
  then N^{10} for 10 features!



If people could «see» in multiple dimensions,
then ML algorithms would not be needed 😀

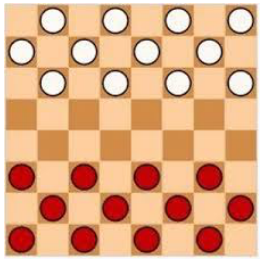
Reinforcement learning

- Learning based on sequence of actions and results, in order to find optimal game rules

backgammon

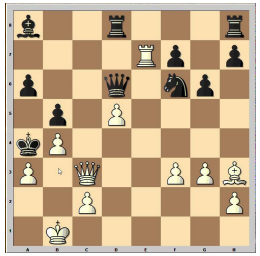
1992

checkers



1994

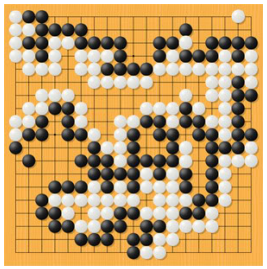
chess



1997

Harry Kasparov vs.
Deep Blue

go



12 March 2016!

- Google's Deep Mind Alpha Go vs. Lee Se-dol
- It was believed that it is almost impossible for computer to win human in Go game
- The number of possible positions is greater than the number of atoms in the universe!

Steps in developing ML application

1. *Collect data*
   - Scraping a website
   - RSS feed
   - API
   - Read from different devices (wind speed, blood glucose levels, etc.)
   - Use publicly available data
2. *Prepare the input data*
   - Make sure it's in a usable format (lists, matrices, dictionaries, etc.)
   - Algorithm-specific formatting (e.g. need features in a special format, target variables and features as strings or integers)
   - The choice of the distinguishing features is critical. Depends on problem. Needs prior knowledge.

Steps in developing ML application

3. *Analyze the input data*
   - Looking at the data simply in a text editor to make sure you don't have empty values or if you can recognize any patterns or if there's anything obvious, such as a few data points that are vastly different from the rest of the set.
   - Plotting data in one, two, or three dimensions can help. But most of the time you'll have more than three features.
   - If it's a production system and you know what the data should look like, or you trust its source, you can skip this step. This step is human involvement which is not for automated system. It makes you understand you don't have garbage coming in.

Steps in developing ML application

4. *Train the algorithm*

- This is where the ML takes place. This step and the next step are where the "core" algorithms lie, depending on the algorithm. You feed the algorithm good clean data from the first two steps and extract knowledge or information.
- Unsupervised algorithms skip this step

Steps in developing ML application

5. Test the algorithm

- Evaluate it to see how well it does. In the case of supervised learning, you have some known values you can use to evaluate the algorithm. In unsupervised learning, you may have to use some other metrics to evaluate the success.
- In either case, if you're not satisfied go to step 4 or 1
- Cost – depends on problem domain.

Steps in developing ML application

6. *Use it*

- Here you make a real program to do some task, and once again you see if all the previous steps worked as you expected. You might encounter some new data and have to revisit steps 1–5.

Main components of ML

- Tens of thousands of machine learning algorithms
- Hundreds new every year
- Every machine learning algorithm has three components:
  - Representation
  - Evaluation
  - Optimization

Representation

- Representation
- Decision trees
- Sets of rules / Logic programs
- Instances
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Etc.

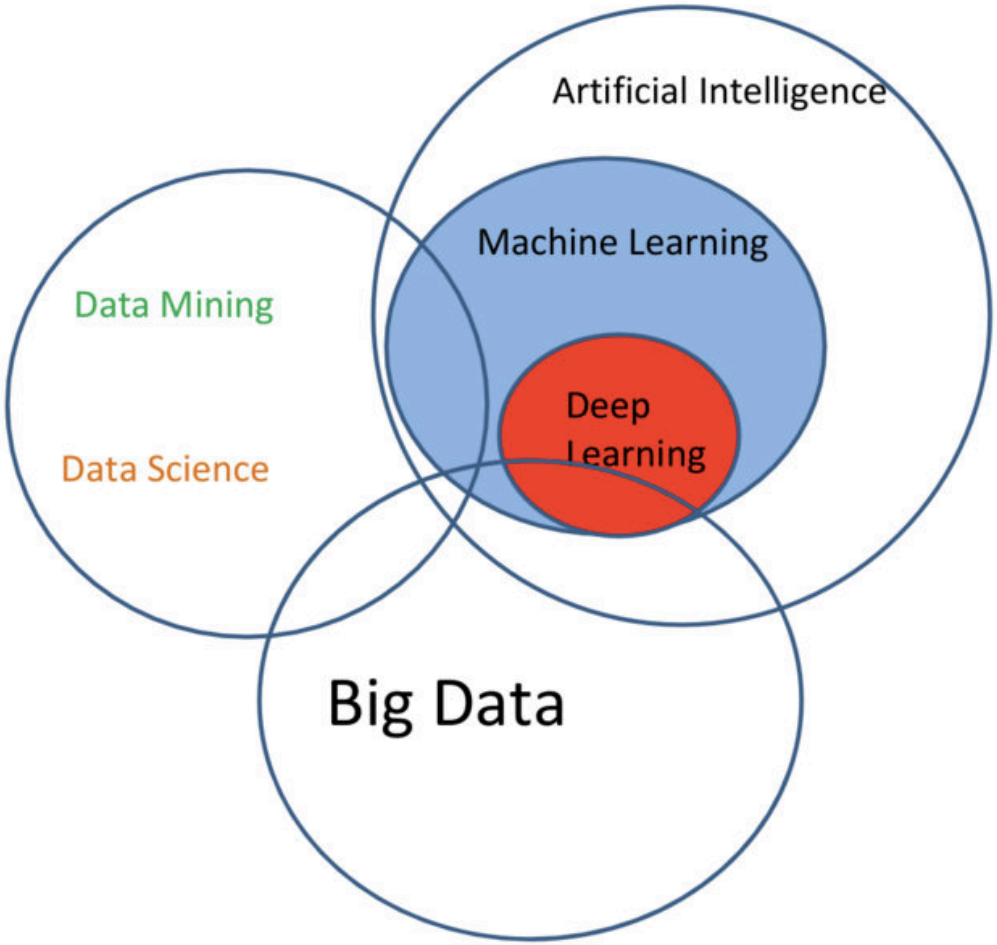
Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost/Utility
- Margin
- Entropy
- Etc.

Optimization

- Combinatorial optimization
  - E.g.: Greedy search
- Convex optimization
  - E.g.: Gradient descent
- Constrained optimization
  - E.g.: Linear programming

Difference between ML, AI, Data Science



Why python?

- Clear and simple syntax. A wide range of predefined data structures (lists, tuples, dictionaries, etc.)
- Any style: OOP, procedural and etc.
- Scripting language that can be used in web application, also has a lot of GUI libraries.
- A wide range of scientific libraries: SciPy and NumPy (compiled at low level using C and Fortran), Matplotlib, Panda, Scikit and etc.
- One of the most popular languages in ML.
- Famous companies like Google use it in their data analysis projects.
- Other strong competitors are R, SAS, Matlab but they have their own disadvantages.

Suggested Resources

- “Programming Collective Intelligence”, Toby Segaran
- “Patter Recognition and Machine Learning”, C. Bishop.
- Machine Learning (coursera), Stanford University, Andrew Ng
- Machine Learning (udacity), Sebastian Thrun

What We'll Cover

- **Supervised learning:** k-Nearest Neighbors, Decision tree, Bayes classifier, Logistic regression, Support vector machines, AdaBoost, Linear Regression, Neural networks, etc.
- **Unsupervised learning:** k-Means clustering, Dimensionality reduction, etc.
- **Additional:** Principal component analysis, singular value decomposition, Big data and MapReduce
- **General machine learning concepts and techniques:** Feature selection, cross-validation, visualization and etc.
- **Your responsibility:**
  – Linear algebra, Statistics and Probability theory, Python

Lecture guests

- Google
- Apple
- LinkedIn
- Other companies & startups
- PhD students