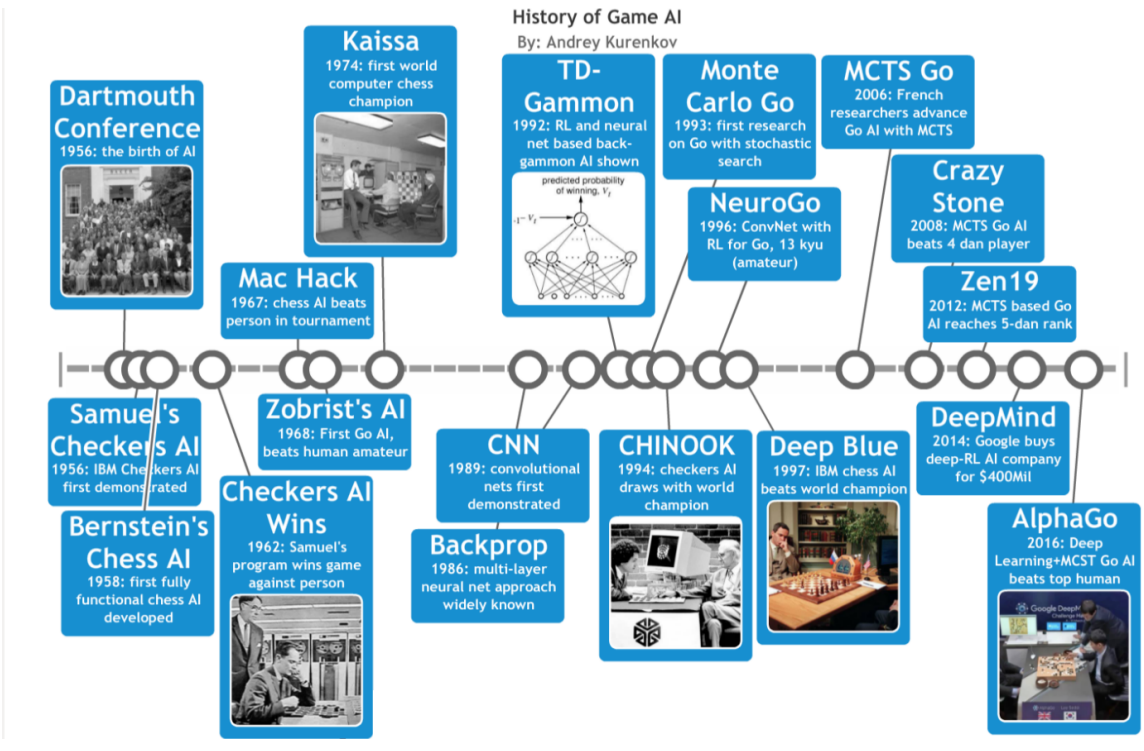
MACHINE LEARNING

THE SCIENCE OF GETTING COMPUTERS TO LEARN FROM DATA WITHOUT HAVING TO BE EXPLICITLY PROGRAMMED BY HUMANS.



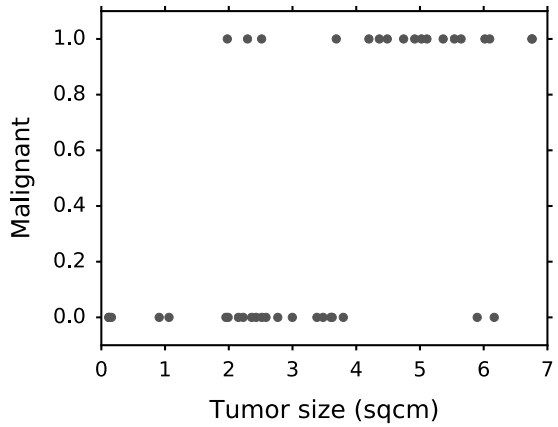
MACHINE LEARNING IS SURROUNDING YOU

- Google search
- Email spam classification
- Spell check

THE MOST BASIC UNDERSTANDING

- It's all about letting computer learns what 'input' is associated to what 'output'.
  - Example: given behavioral profiles of a customer A, algorithms predict the chance this customer is going to leave.
  - Example: given inputs from sensors and cameras, the robotic algorithm pushes out the appropriate movement.

A SIMPLE EXAMPLE



- A doctor has seen hundreds of patients with tumors. They measured the size of tumors and test whether they are malignant.
- The historical data let the doctor make prediction that bigger tumor is more likely to be malignant.

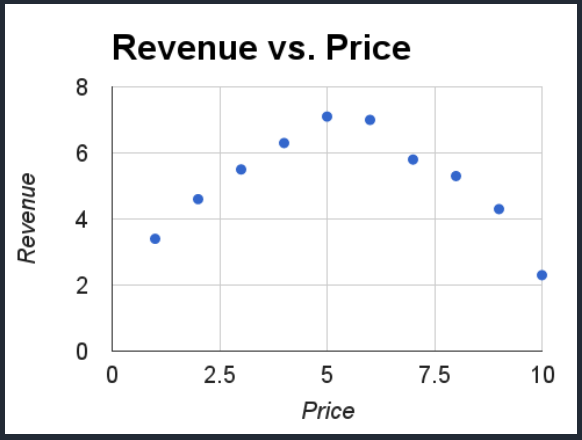
INTELLIGENT SYSTEM WITH MACHINE LEARNING

- People provide Facebook the images and tags of names in the photos.
- Over time Facebook learned to associate names with faces and can automatically recognize these people.

ANOTHER VIEW OF MACHINE LEARNING

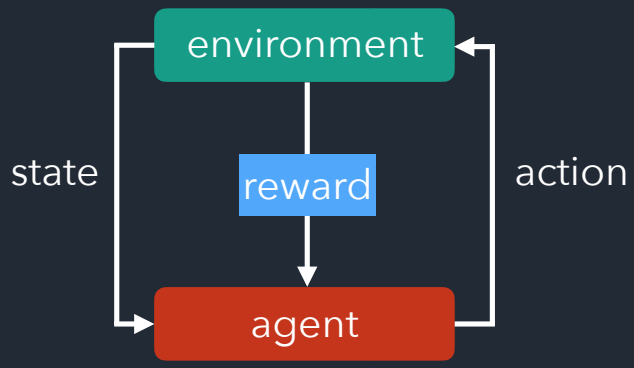
TEACHING THE COMPUTER TO LEARN FROM EXPERIENCES AND OPTIMIZE A GIVEN PERFORMANCE INDEX AS THEY PRACTICE.

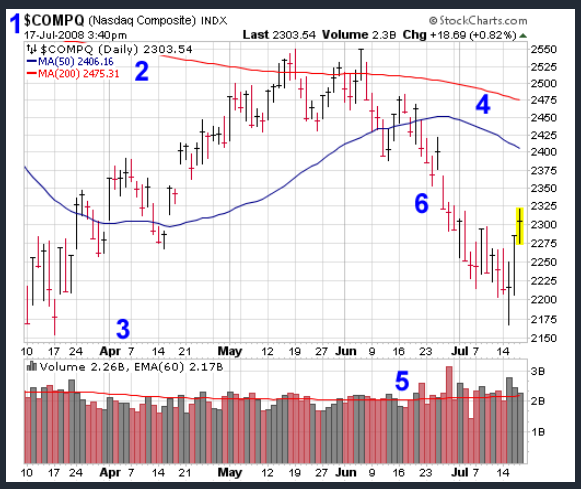
A SIMPLE EXAMPLE



- What is the price we should set for product A in order to maximize revenue?
- What production plan should we use to optimize profit?
- How much should we bid for an advertising spot?

INTELLIGENT SYSTEM WITH MACHINE LEARNING





BIG DATA LEADS TO BETTER MODELS

BIG DATA

↓

BETTER MODEL

↓

HIGHER PRECISION

THIS MEANS THAT TODAY WE CAN CREATE A SYSTEM TO AUTOMATE BUSINESS IN THE WAY WE NEVER THOUGHT POSSIBLE!

DATA SCIENCE PROCESS

from data to values

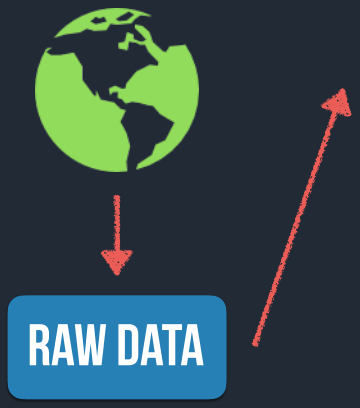
FROM DATA TO VALUES

RAW DATA

- Machine-generated data e.g. server logs, clickstream, POS, kiosk log
- Images & sounds e.g. photographs, videos, handwriting images, voice recordings
- Human-generated (languages) e.g. text messages, tweets, web content
- Records e.g. large automated survey, tax, maintenance log
- Sensors e.g. temperature, accelerometer, geolocation

DATA COLLECTION

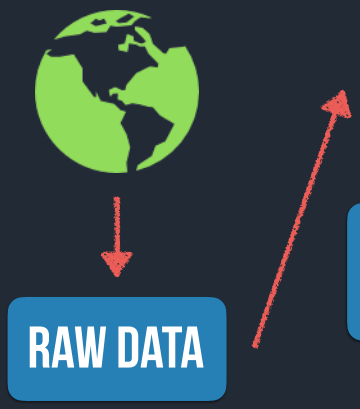
FROM DATA TO VALUES



- linking data from different silos
- selecting which columns we need
- filtering out data that is useless

PRE-PROCESSING, DATA LINKAGE, CLEANING

FROM DATA TO VALUES



- plot data in graphs
- identify predictors
- what models to use?

DESCRIPTIVE STATISTICS, VISUALIZATION, FEASIBILITY CHECK

FROM DATA TO VALUES

RAW DATA

DATA SETS

EXPLORATORY ANALYSIS

FEATURE EXTRACTION

- transform raw data into usable 'features'
- for example transaction -> preference
- this process can be complex

FROM DATA TO VALUES

RAW DATA

DATA SETS

EXPLORATORY ANALYSIS

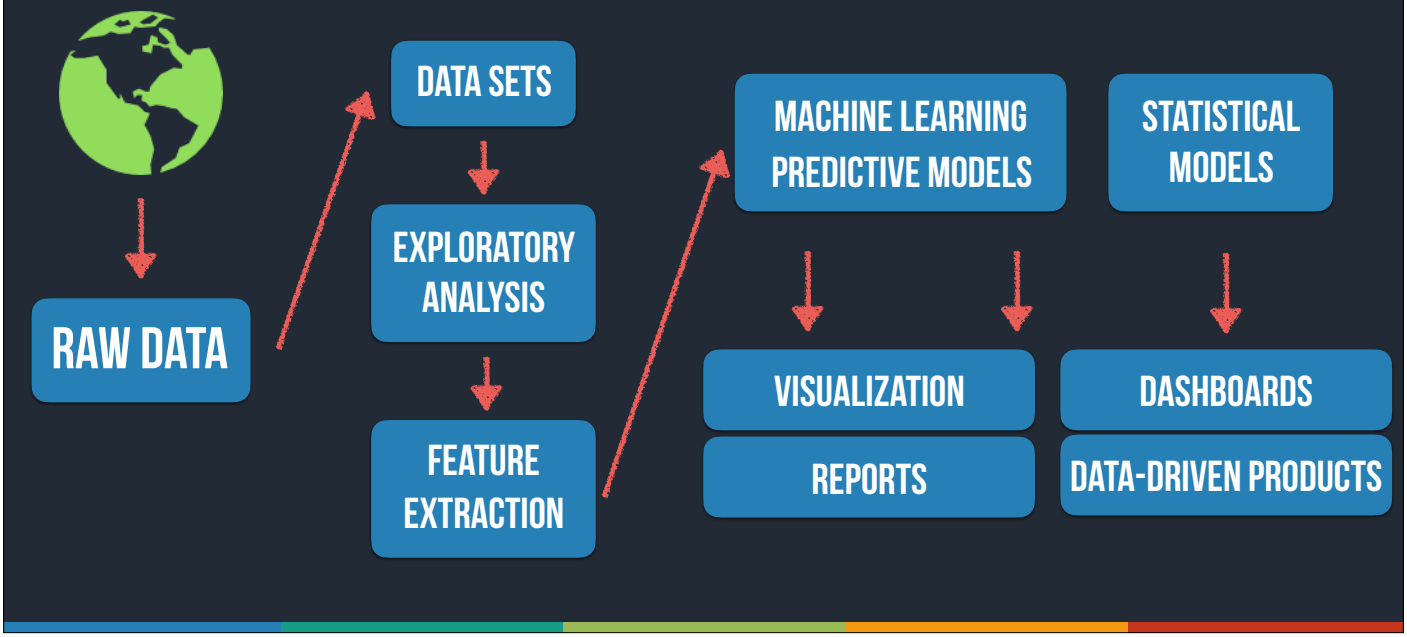
FEATURE EXTRACTION

MACHINE LEARNING PREDICTIVE MODELS

STATISTICAL MODELS

- make predictions
- validate hypotheses you had about the data
- discovering patterns in data

FROM DATA TO VALUES



UNDERSTAND MACHINE LEARNING MODELS

how do they become smart?

TOWARDS MACHINE LEARNING EXPERTISE

- **Machine learning is an applied mathematics field.**
  Understanding of mathematics is required to be able to make proper models. This skill can be acquired through experiences. Though it takes a lot of patience.

- **Coding is somewhat required.**
  As you will see that the model can be arbitrarily complex, therefore it's hard to use GUI tools (Scripting can be done in Matlab, R which are high-level scripts like SPSS or SAS.

- **Easy to do, hard to understand.**

THE MOST BASIC EXAMPLE

An agent has been selling 300 houses in the last years and want to be able to predict the price of a house by just knowing the size of the house.

i	Size (m^2)	Price (Mbaht)
1	50	1.4
2	128	2.6
3	24	0.8
4	78	1.2
i	…	…

THE MOST BASIC EXAMPLE

In general,

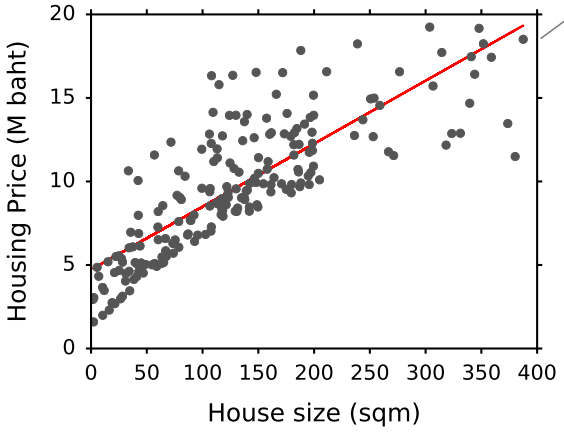
x: feature (input)

y: target (output)

i: sample index

	x	y
i	Size (m^2)	Price (Mbaht)
1	$50 = x_1$	$1.4 = y_1$
2	$128 = x_2$	$2.6 = y_2$
3	$24 = x_3$	$0.8 = y_3$
4	$78 = x_4$	$1.2 = y_4$
i	$\dots = x_i$	$\dots = y_i$

WHAT IS A MODEL



$$h(x) = \theta_0 + \theta_1 x$$

parameters of the model

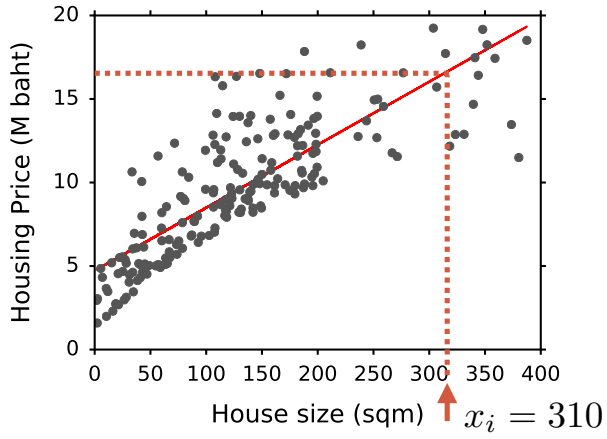
$$\theta_0 = 4.717$$
$$\theta_1 = 0.038$$

A model is a function that takes an input and yields the values we want to predict.

WHAT IS A MODEL

After we have a model, we can predict y value from any x value

Model: $h(x_i) = 4.717 + 0.038 * (x_i)$



Notice that:

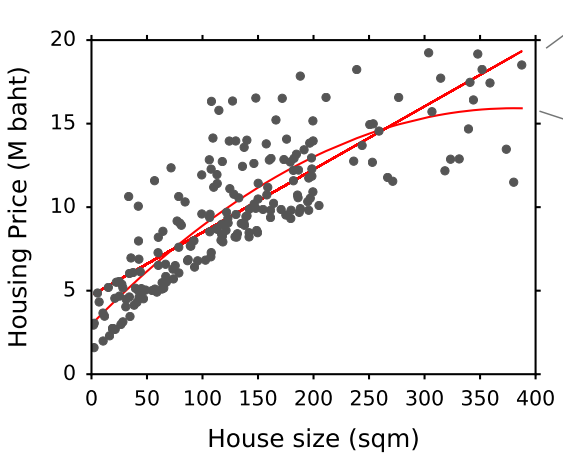
h(x) and y are not the same

h(x): value we predict

y: the actual value

WHAT IS A MODEL

For different models you have different functions, with perhaps different number of parameters. Different models can be fitted to the same data set.



$h(x) = \theta_0 + \theta_1 x$

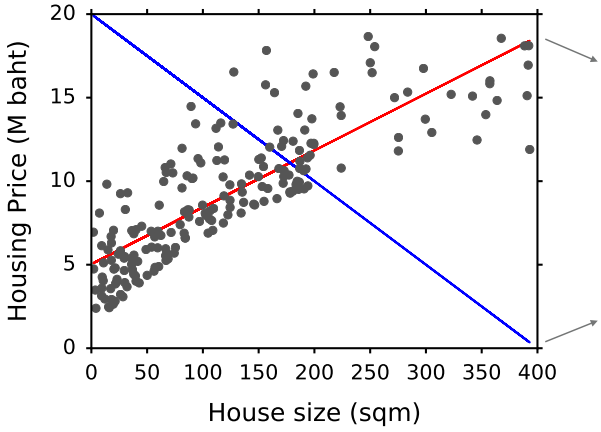
$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

Model is a function of both features (x) and parameter (theta).

$h(x; \theta)$

GOOD AND BAD MODELS

Model: $h(x) = \theta_0 + \theta_1 x$



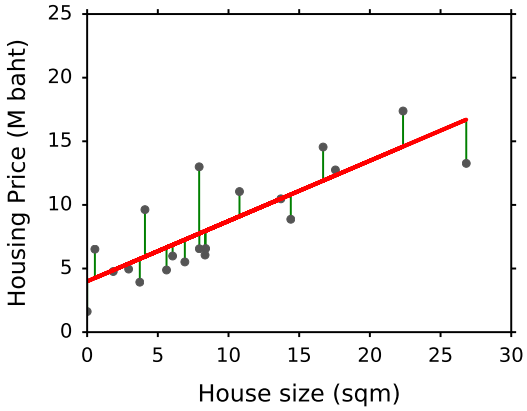
$\theta_0 = 4.717$
$\theta_1 = 0.038$

$\theta_0 = 20$
$\theta_1 = -0.05$

COST FUNCTION

Cost function (or loss function) is a measure of whether a model is a good fit to the data.

For example, a famous cost function is called 'squared error' function that takes the difference between what you predict and the actual data value and square it.

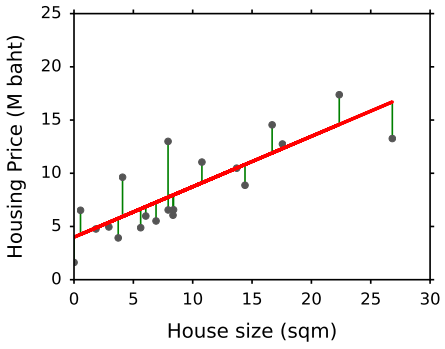


$$\sum_i (h(x_i) - y_i)^2$$

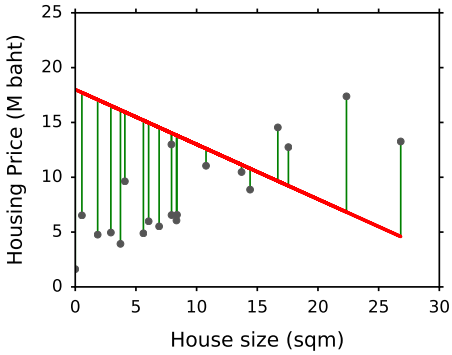
LOWERING COST FUNCTION

Model: $h(x) = \theta_0 + \theta_1 x$

$\theta_0 = 4.717$
$\theta_1 = 0.038$

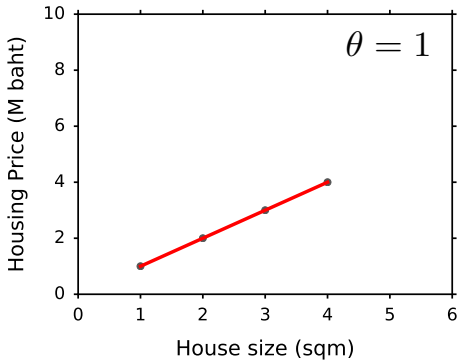


$\theta_0 = 20$
$\theta_1 = -0.05$

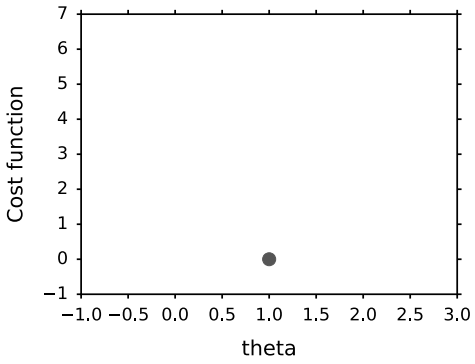


MINIMIZING COST FUNCTION

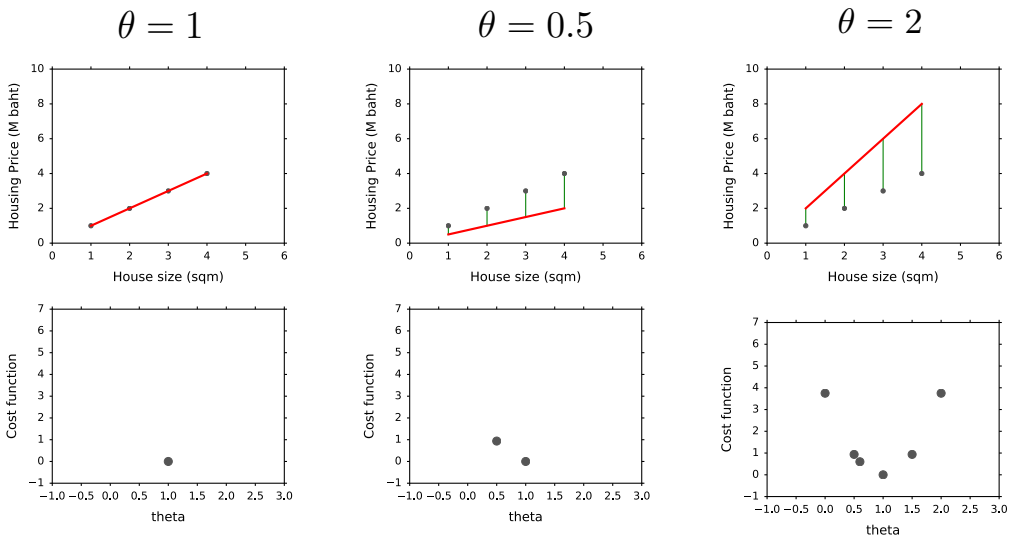
Model: $h(x) = \theta x$



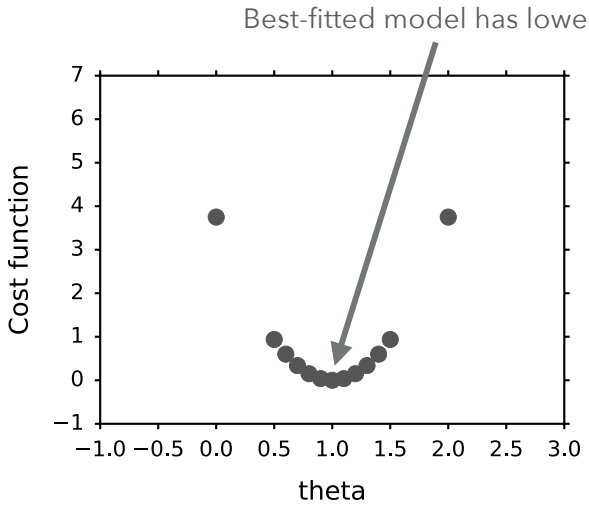
$Cost(\theta)$



MINIMIZING COST FUNCTION



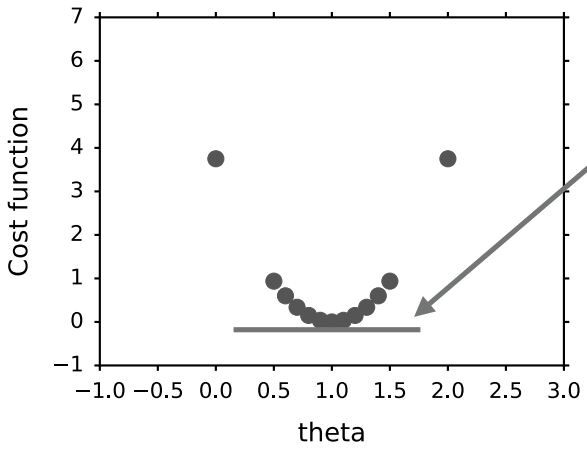
MINIMIZING COST FUNCTION



Training machine learning models = minimizing cost function

Often accomplished by using 'optimization algorithms'

MINIMIZING COST FUNCTION

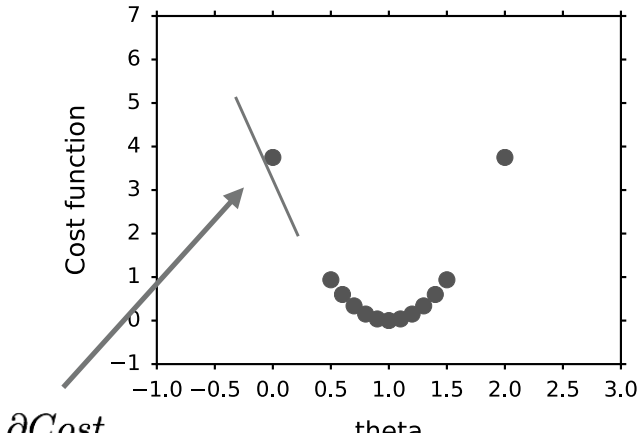


$$\frac{\partial Cost}{\partial \theta} = 0$$

Some optimization algorithm is so simple,

you just solve an equation and done.

MINIMIZING COST FUNCTION



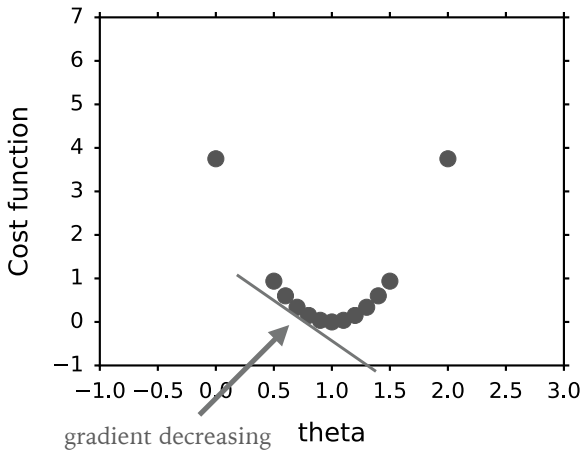
$$\frac{\partial Cost}{\partial \theta} < 0$$

Gradient Descent

- pick a random theta
- find gradient
- update theta following the descending gradient

$$\theta \rightarrow \theta - \alpha * gradient$$

MINIMIZING COST FUNCTION



Gradient Descent

- pick a random theta
- find gradient
- update theta following the descending gradient

$$\theta \rightarrow \theta - \alpha * gradient$$

- gradient will decrease
- repeat until gradient is zero

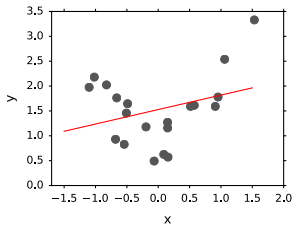
MINIMIZING COST FUNCTION

- Genetic algorithm
- Bootstrapping or alternate optimization
- Second-order methods

OVERFITTING

underfit

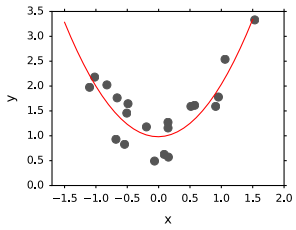
cost = 0.21



$$h(x) = \theta_0 + \theta_1 x$$

just right

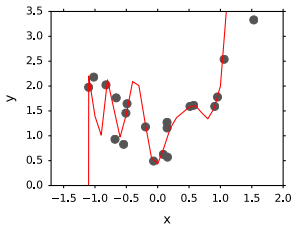
cost = 0.05



$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

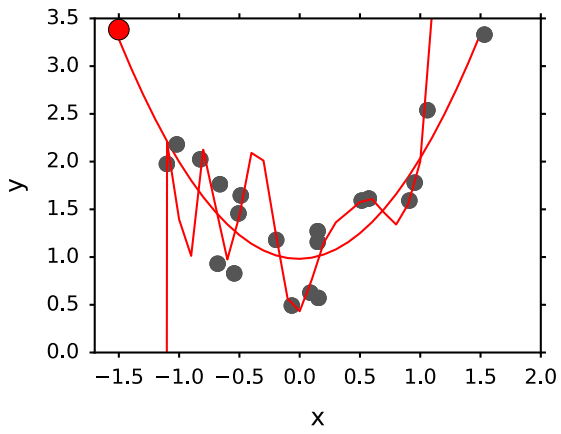
overfit

cost = 0.02



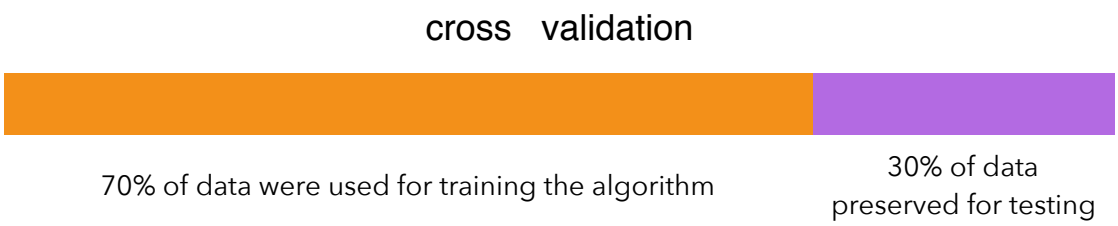
$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_{15} x^{15}$$

OVERFITTING MODELS CANNOT GENERALIZE



AVOID OVERFITTING

- Cross-validate your models: train models with one set of data (training set) and test them with another set of data (test set)



AVOID OVERFITTING

- Cross-validate your models: train models with one set of data (training set) and test them with another set of data (test set)

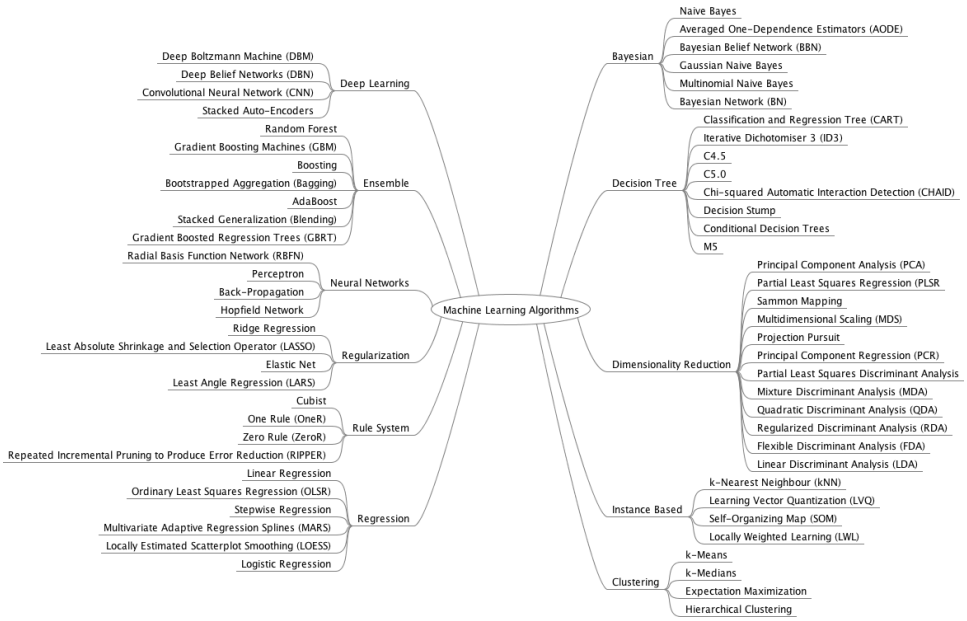
- Reduce the number of features and parameters
  - Manual selection
  - Algorithmic selection (information gain, pruning, stepwise algorithm)
  - Regularization: cost function is split to two parts

$$\sum_i (h(x_i) - y_i)^2 + \lambda \sum_j \theta_j^2$$

DATA SCIENCE PROBLEM DEFINITION

- **Understand the problem:** what do you want to accomplish?
  - A client wants to be able to know how he should price each house for sale.
- **Know the input, the output, and know what data is available.**
  - The client has houses' properties (size, size of the lawn, #bedrooms, #bathrooms) and prices at which the houses were sold.
- **Given the input and output, realize what model to use.**
  - The input and output are continuous variables, we can use regression.

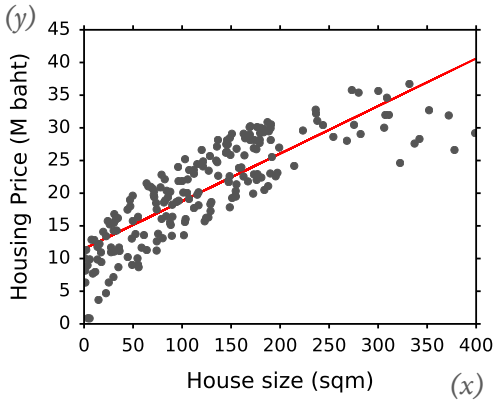
SO MANY ALGORITHMS



SUPERVISED V.S. UNSUPERVISED LEARNING

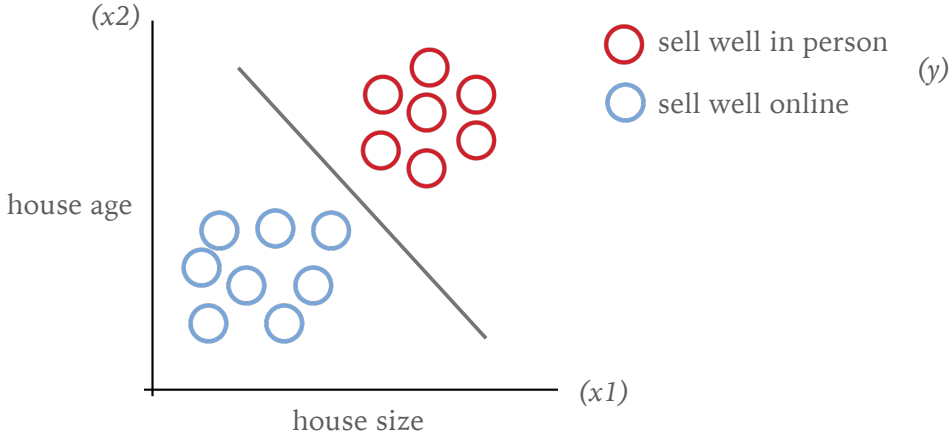
- **Regression Problems:** Predicting **continuous** output from input.
  - Example: predicting prices of houses

Supervised learning means you already have answers in your database.



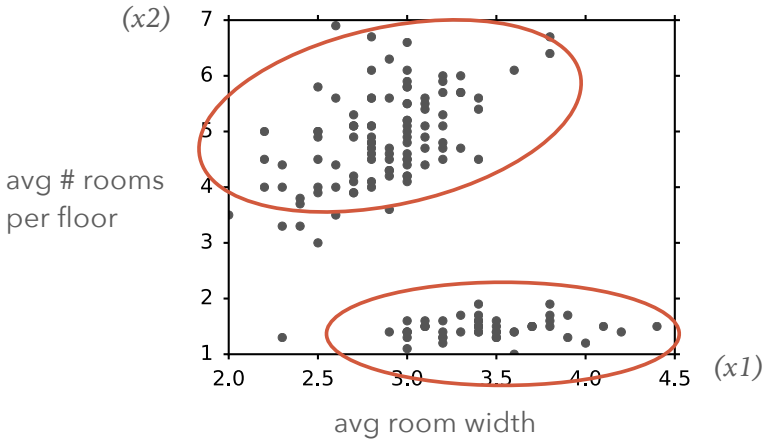
SUPERVISED LEARNING EXAMPLE

- **Classification:** predicting **discrete** output from input
  - Example: predicting whether a house would sell well in person or online



UNSUPERVISED LEARNING EXAMPLE

- **Clustering:** grouping unlabeled input by similarities
  - Example: determining classes of real estates



SUPERVISED V.S. UNSUPERVISED LEARNING

- **Supervised learning**
  - You know the answer or can obtain such answer, note that answers can be numerical or categorical.
- **Unsupervised learning**
  - You would like to discover the categories, you usually don't even know how many categories there are and what they are.
- Note that you can apply both approaches to the same dataset!

FREQUENTLY USED MODEL FAMILY

more advanced machine learning techniques

POPULAR MODEL FAMILY: GLM

Generalized Linear Model:
predicting y from a linear function of x variables.

$$h(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n)$$

Most famous members

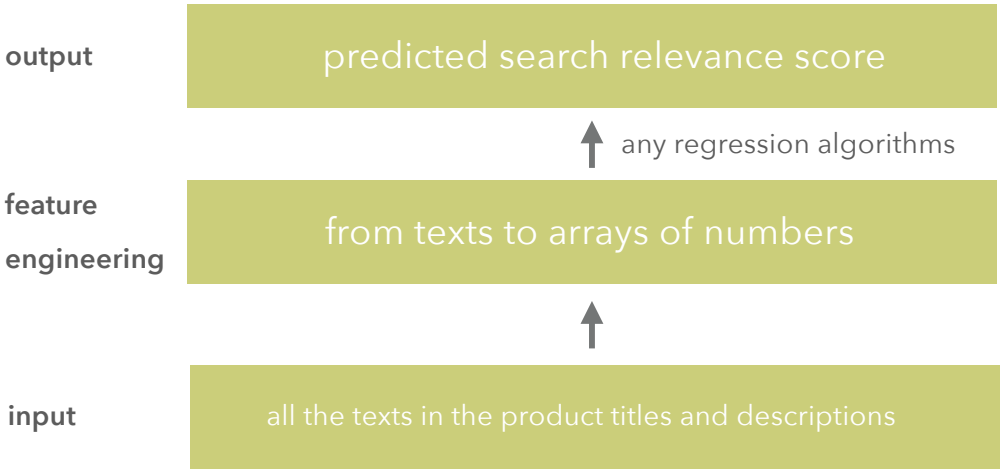
Linear Regression: x and y are continuous variables

Logistic Regression: x is continuous, y is categorical

Stepwise Regression, Ridge Regression

GLM EXAMPLE

HomeDepot challenge: predict search relevance score from titles of product page and product description.



CLASSIFICATION

Classification: y is a categorical variable. Predict the probability that data point x_i belongs to each category.

$$y \in \{1, 2, 3, ...\} \qquad h(x) = P(y = c|x; \theta)$$

h(x) approximates the probability that y belongs to category c.

Most famous members

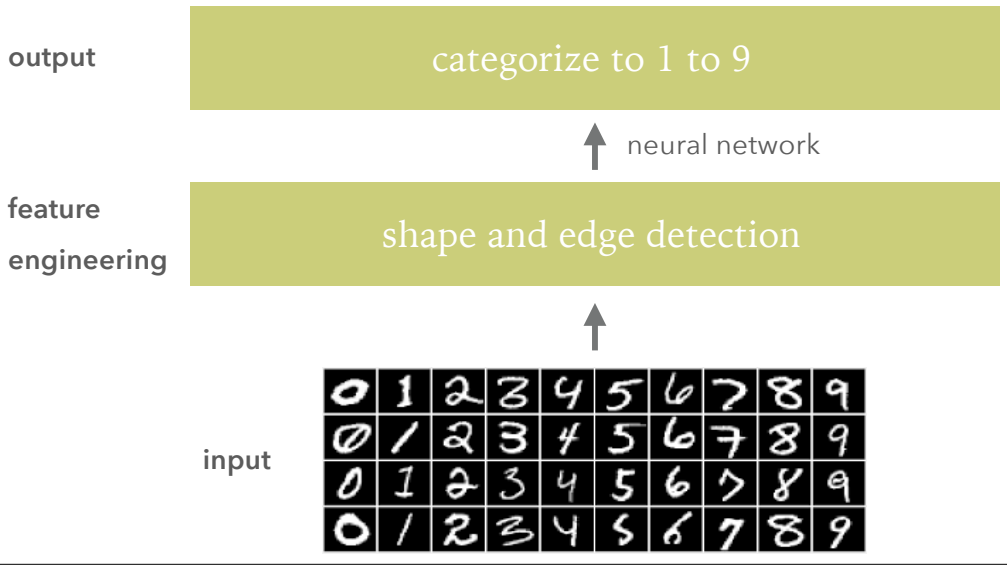
Instance-based learning, Linear discriminant analysis, Support vector machine, Naive Bayes classifier, Neural networks

COMMON APPROACHES

- Instance-based (nearest neighbor) methods
- Kernel method (such as SVM)
- Decision tree and random forest methods
- Bayesian methods
- Neural network methods

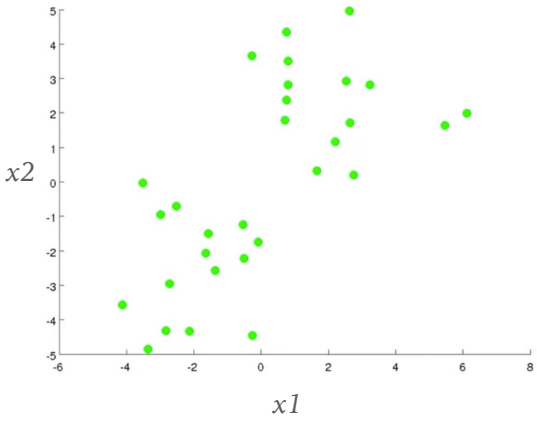
COMMON APPROACHES

Reading numbers out of an image of a check.



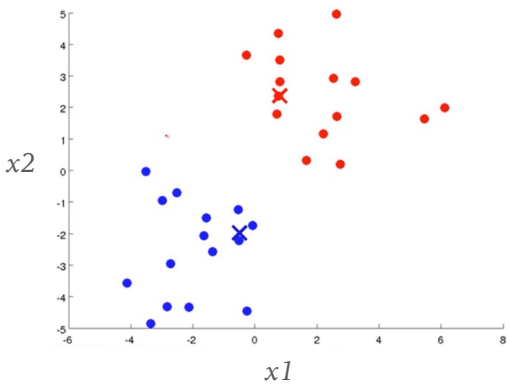
POPULAR MODEL FAMILY: CLUSTERING

Clustering: given a set of data points, find the optimal way to group these data points together.



POPULAR MODEL FAMILY: CLUSTERING

Clustering: given a set of data points, find the optimal way to group these data points together.

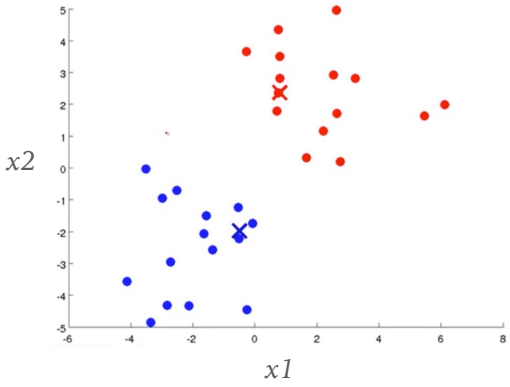


The main idea is to find **centers of categories or centroids...**

Dots that are closer to centroid red belong to category red.

POPULAR MODEL FAMILY: CLUSTERING

Clustering: given a set of data points, find the optimal way to group these data points together.

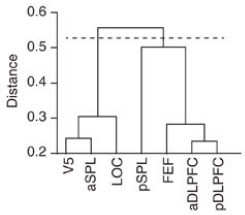


- First you might pick random dots to be centroids of red and blue.
- Keep updating the locations intelligently until the cost function is lowest.
- Cost function here is the total distance from all data points to the centers of categories they belong to.

POPULAR MODEL FAMILY: CLUSTERING

Most famous members

K-means clustering, hierarchical clustering



Example

News categories, gene profiles, market segmentation, social network analysis

POPULAR MODEL FAMILY: DIMENSIONALITY REDUCTION

Dimensionality Reduction: take very high-dimensional features and transform them into lower-dimension features without losing the quality of the data.

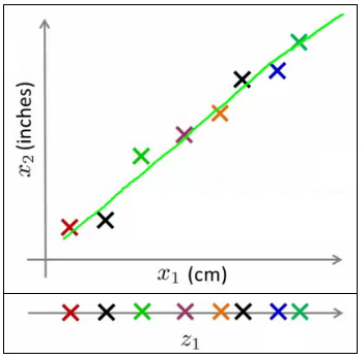
This is usually done by optimizing a cost function that quantifies goodness of components.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_{100} \end{bmatrix} \rightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Most famous members

Principle component analysis, singular value decomposition, independent component analysis

POPULAR MODEL FAMILY: DIMENSIONALITY REDUCTION



PCA

- Starting with high-dimensional data
- Find new axes where data points can be projected on
- Get data on a new lower dimension form

More Realistic Example

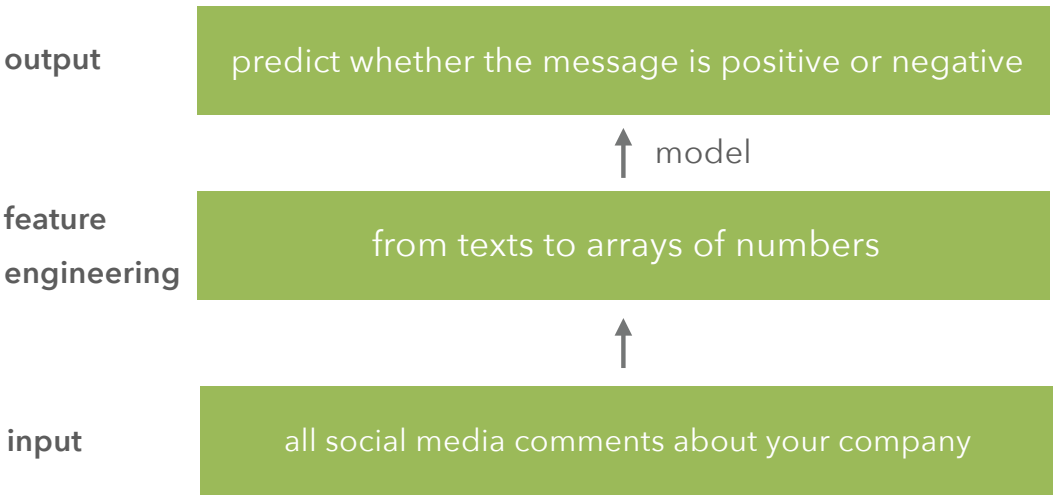
We collected restaurant ratings from 1M people rating 1000 restaurants (this dataset has 1000x1000 dimensions). We don't want to build model in 1000 dimensions (too computationally expensive).

FEATURE ENGINEERING

Data rarely comes in the appropriate form so one of the most important tasks in machine learning is to transform those data.

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

FEATURE ENGINEERING



FEATURE ENGINEERING

- Preprocessing
- Feature selection
- Linear transformation (dimensionality reduction)
- Feature extraction

MACHINE LEARNING BASIC SUMMARY

- Model is a function that takes a a given sample feature (x) and returns the prediction (h(x)) of value we are interested (y).
- Model is defined with a set of parameters which can be adjusted to fit the data.
- Given a model, you want to define cost function and find a set of parameters that minimize cost function.
- Models with really, really low cost function is not necessarily good. A good model must be able to generalize to data it has not seen.
- A lot of machine learning tasks focus not on the models, but on engineering the input.

MACHINE LEARNING TOOLS

- Language and platforms
  - **Python**, Java, C++
  - Matlab, **R**
- Libraries (https://goo.gl/Sf9hYK)
  - **Python datascience**: scipy, numbpy, scikit-learn, matplotlib, pandas
  - **Python neural net and deep learning**: Theano, Tensorflow, Keras
  - Java: Weka, Caffe
  - C++: MultiBoost, Shogun
- Large-scale
  - Spark (scala, java, **python**), Mahout (java)

SOME TIPS

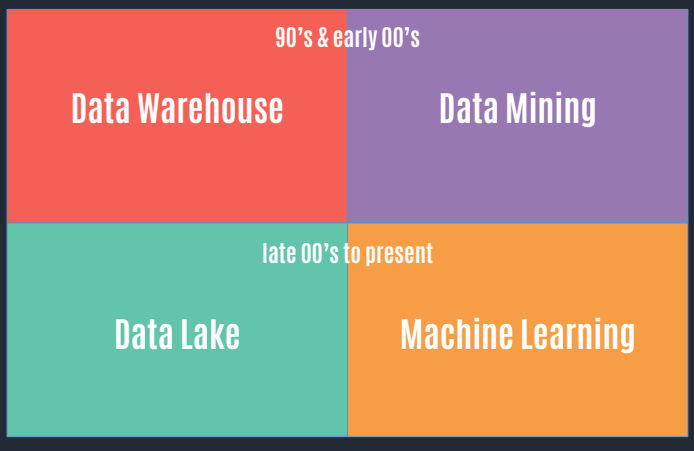
- Get to know more models.
- Don't just use libraries blindly. Try to understand what your black-box ML algorithms are doing, so you know how to improve them or fix them when things go wrong.
- Experimentations are crucial.

MACHINE LEARNING AT SCALE

more advanced machine learning techniques

ANALYTICS IS NOT NEW IT JUST GOT MUCH BETTER

WHAT IS THE DIFFERENCE BETWEEN DATA MINING AND MACHINE LEARNING?

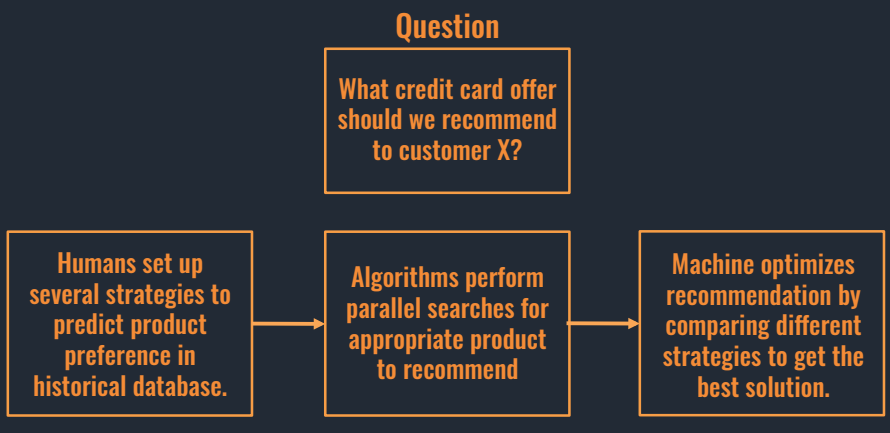


OLD ANALYTICS - HYPOTHESIS DRIVEN

Question	Hypothesis	Data mining
What product should we recommend to customer X?	X is 50 years old and has kids, he should prefer product Y.	Segment customers based on age and kids. Prove the hypothesis.

Hypothesis-driven analytics relies on humans to make hypotheses. This often means few variables are added to a rigid math model.

NEW ANALYTICS - DATA DRIVEN



Machine learning algorithm is designed with less constraints, exploring larger space of solutions to solve problem.

HOW BIG DATA GENERATE VALUES

BIG DATA

↓

BETTER MODEL

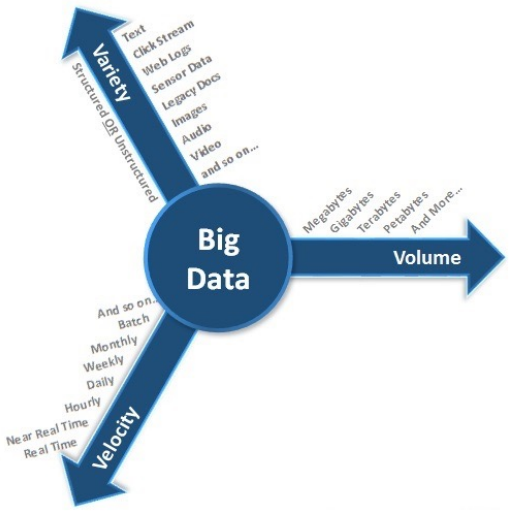
↓

HIGHER PRECISION

THIS MEANS THAT TODAY WE CAN CREATE A SYSTEM TO AUTOMATE BUSINESS IN THE WAY WE NEVER THOUGHT POSSIBLE!

SCALING TO BIG DATA

The true power of machine learning comes when we scale it to the full potentials.

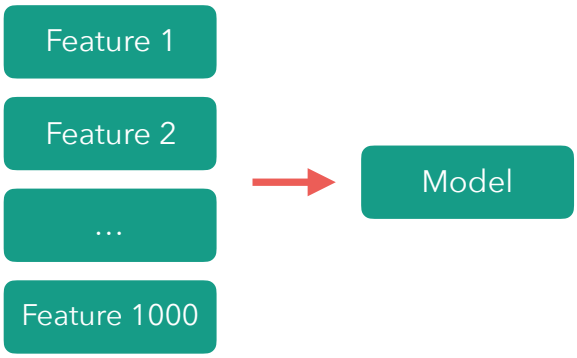


CHALLENGES OF BIG DATA

- Volume: large amount of data / limited computing resources, the algorithm must scale to deal with large volume of data
- Velocity: algorithm must be portable to make predictions on the fly
- Variety: algorithms must be able to take in a large variety of inputs

HOW MACHINE LEARNING DEALS WITH BIG DATA

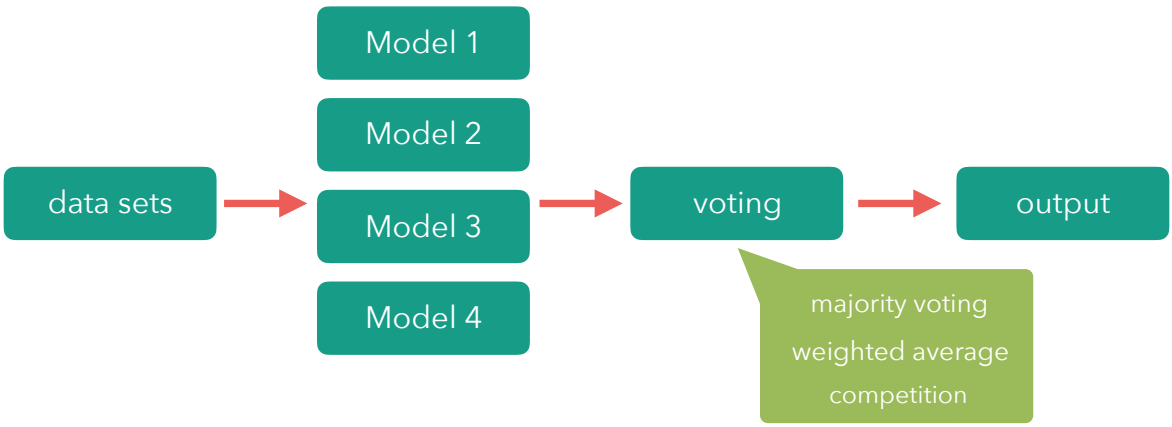
Machine learning can tackle large feature spaces



Example: some customer analytics model is reported to have over 70,000 features.

HOW MACHINE LEARNING DEALS WITH BIG DATA

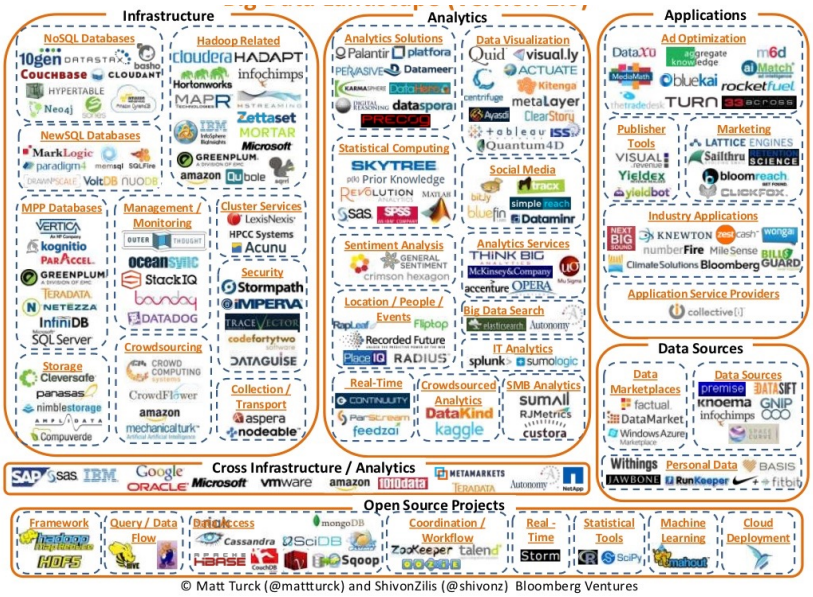
Multiple machine learning models attack one problem



Example: some customer analytics model is reported to use over 130 models on one problem.

HOW MACHINE LEARNING DEALS WITH BIG DATA

By exploiting big data architecture



© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

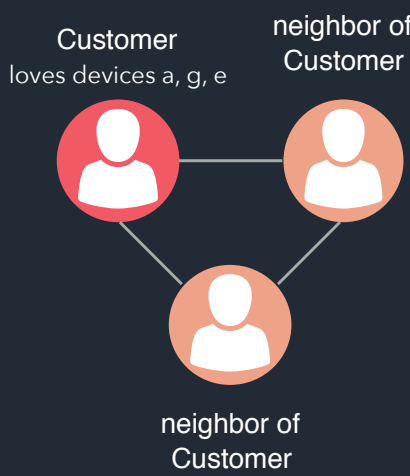
POWER MACHINE LEARNING WITH PARALLEL COMPUTING

If you properly implement machine learning on big data, there will be a few instances where computations are just really massive it will take a few months to run all the computation you'd like.

RECOMMENDATIONS

Will customer A like shop B and C?

RECOMMENDATIONS



Find customers with similar behavior as customer A.

If neighbors love shop B&C, customer A should love shop B&C too.

POWER MACHINE LEARNING WITH PARALLEL COMPUTING

Total Computation	Assuming 1 process takes 1 ms
$10^6 \times 10^6$	1 core	3.2 months
	48 cores	2 days

Parallel computing makes all the difference.

MODERN MACHINE LEARNING

The development of modern machine learning algorithms were facilitated by many factors such as cheaper computing cost, yielding new methods with many advantages.

1. **LARGE FEATURE SPACE**
   Models are designed to scale. Machine learning model can handle large numbers of variables compared to traditional methods.
2. **MORE COMPLEX ALGORITHMS**
   Machine learning models are often more complex, think nonlinear neural network in comparison to regression. Larger parameter spaces.
3. **LARGE NUMBER OF ALGORITHMS**
   Support multiple models to compete or collaborate to find the optimal results. This often leads to models with higher accuracy.
4. **FLEXIBILITY AND SCALABILITY**
   Models can be engineered to fit any problems and to be run in distributed environment.