cloudera

Introduction to Machine Learning

Guru Medasani (@gurumedasani)

Nov 3, 2015

TECHNICAL SUMMIT 2015

About me

- Current - Sr. Solutions Consultant @Cloudera

- Past -BigData Engineer @Monsanto Research and Development

Goals of this talk

- Introduce machine learning

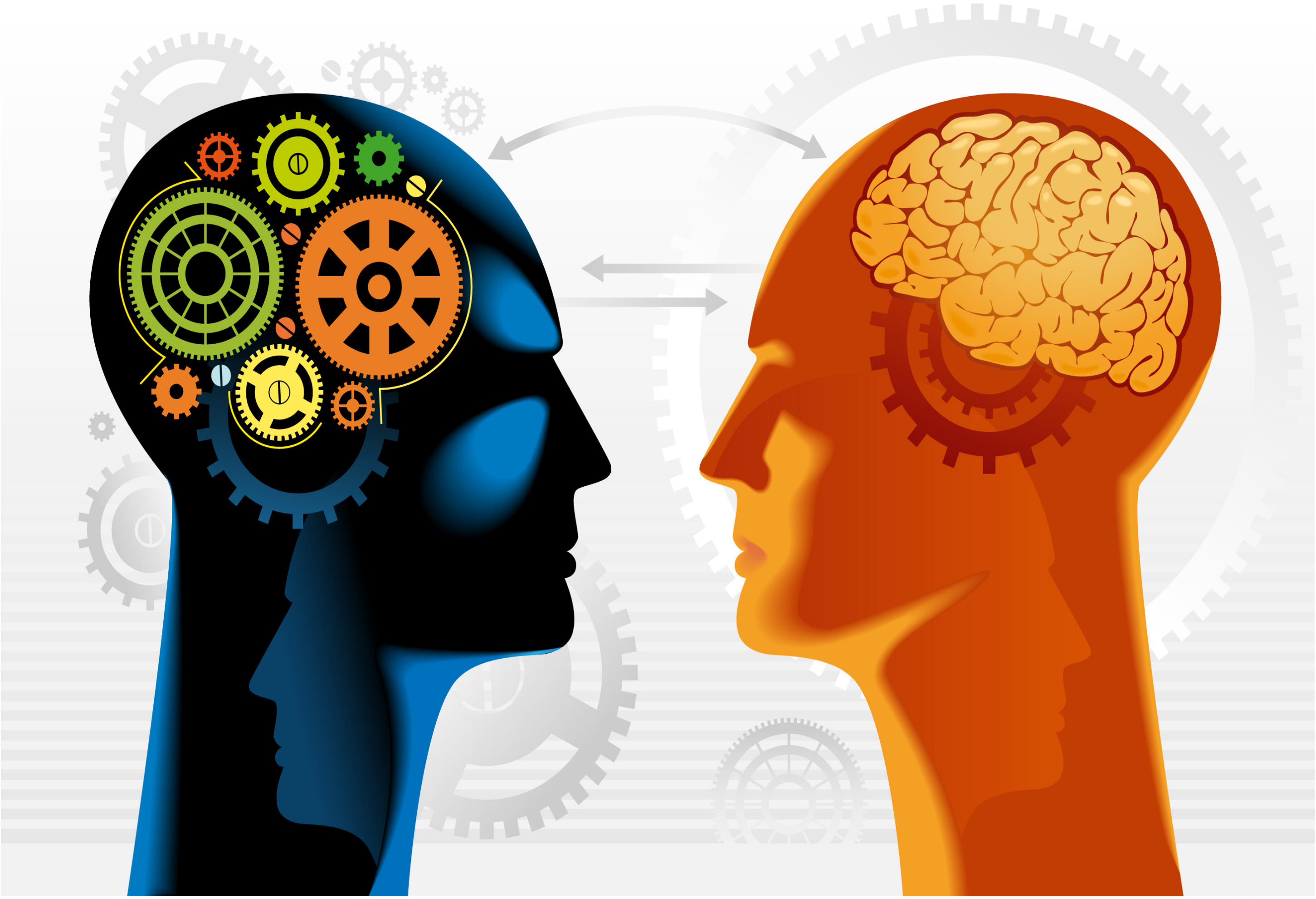
- Get familiar with simple machine learning tools and techniques available.

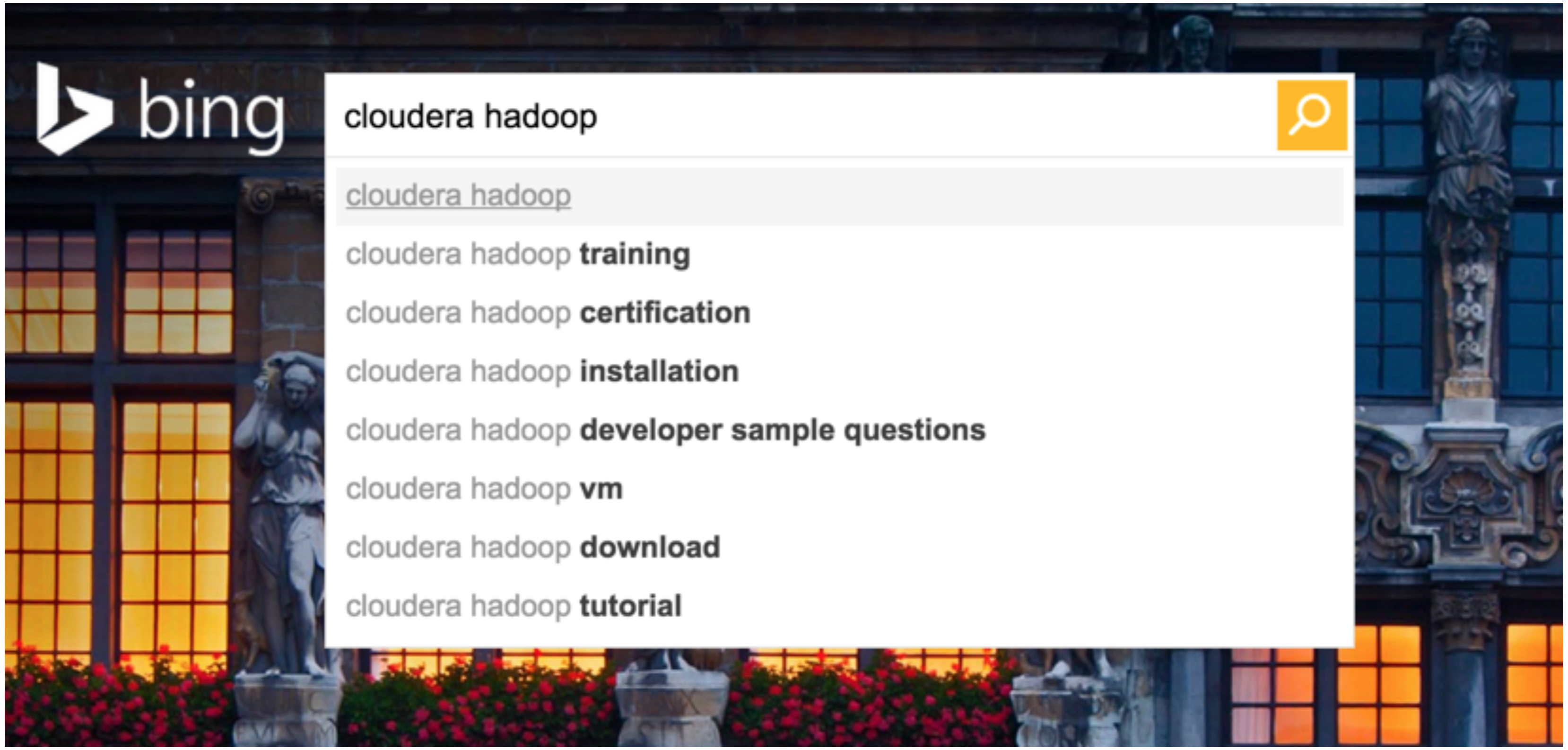
- Enable you to run linear regression algorithms on datasets at any scale.

In this talk

- Machine Learning Motivation
- Machine Learning Definition
- Machine Learning Algorithms
- Linear Regression Motivation
- Linear Regression Deep Dive
- Demo: Linear Regression

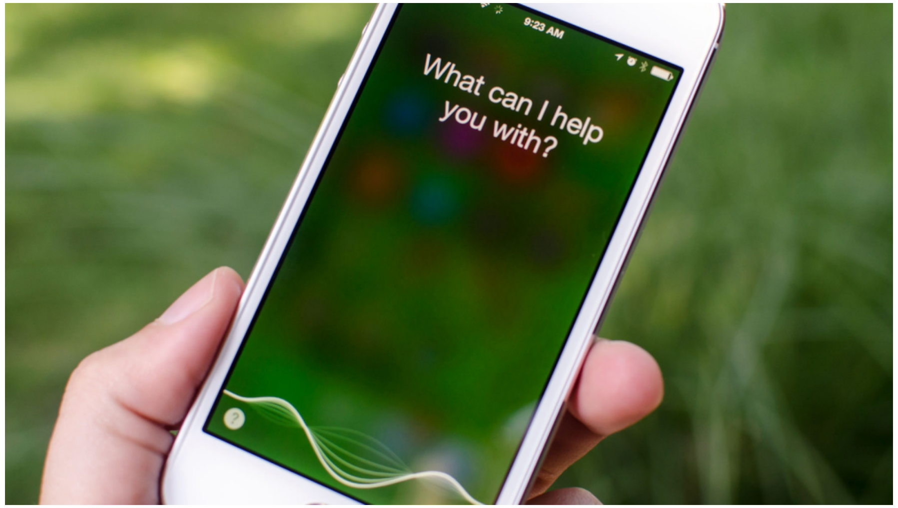
Machine Learning: Motivation




bing
cloudera hadoop
cloudera hadoop
cloudera hadoop training
cloudera hadoop certification
cloudera hadoop installation
cloudera hadoop developer sample questions
cloudera hadoop vm
cloudera hadoop download
cloudera hadoop tutorial


SPAM


Hi. I'm Cortana.
Ask me a question!


9:23 AM
What can I help you with?

ML is Applied Everywhere

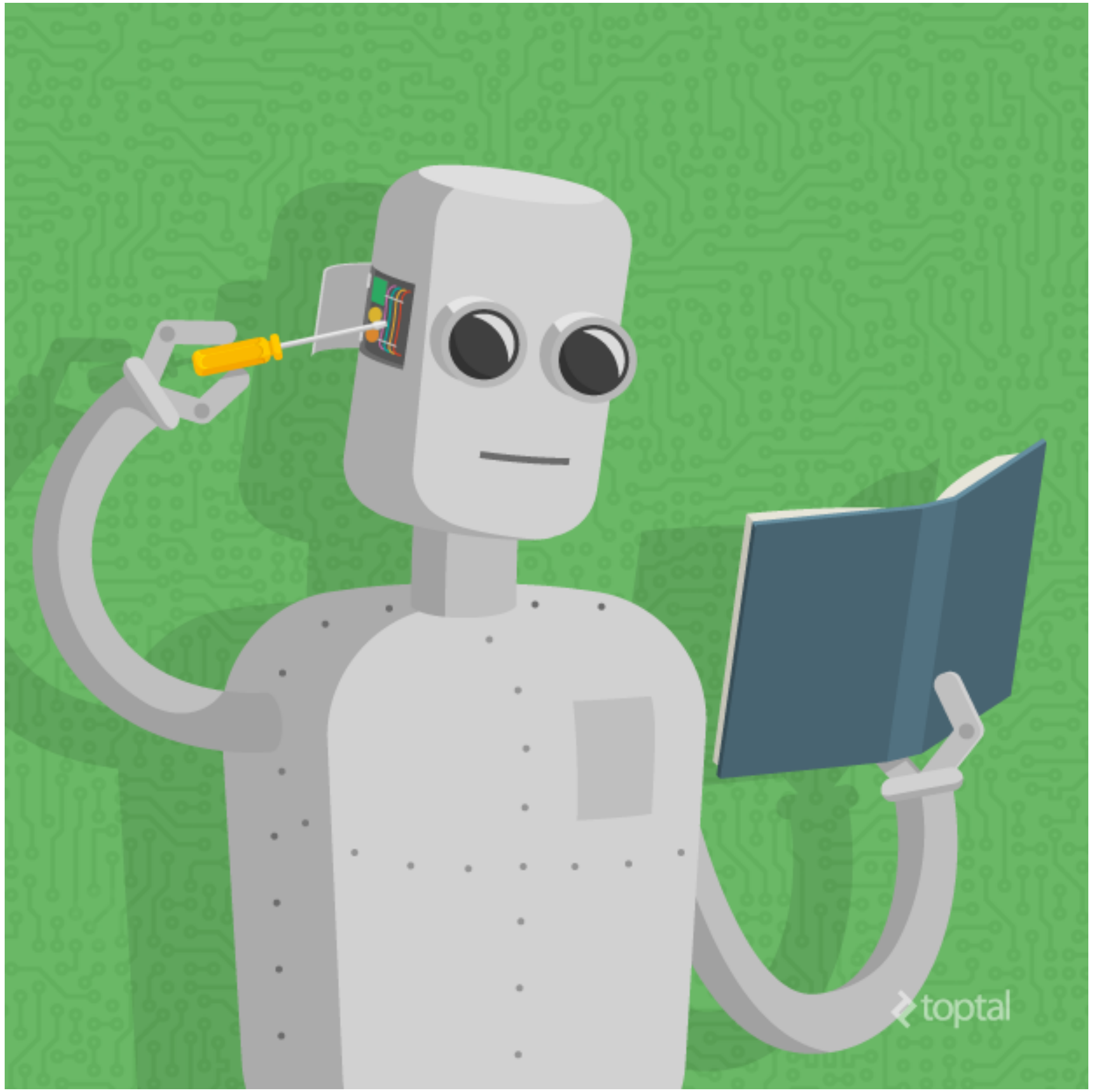
- Data Mining
  - Large datasets from growth of mobile, web, cloud
  - E.g., location data, web click data, genomics
- Applications hard to program by hand
  - E.g., handwriting recognition, computer vision, natural language processing
- Self-customizing programs
  - Netflix, Spotify, Amazon recommendations
- Understanding human learning (brain, real AI)

Machine Learning

- In all these applications, it is very hard to write a computer program to perform these tasks.

- We would like the computer to learn by itself how to perform these various tasks

Machine Learning: Definition



Machine Learning: Definition -1

- Arthur Samuel in 1959 described it as:

  " The field of study that gives computers the ability to learn without being explicitly programmed."

- This is an older, informal definition

https://en.wikipedia.org/wiki/Machine_learning

Machine Learning: Definition -2

- Tom Mitchell in 1998 described a well-posed learning problem as:

  "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with Experience E."

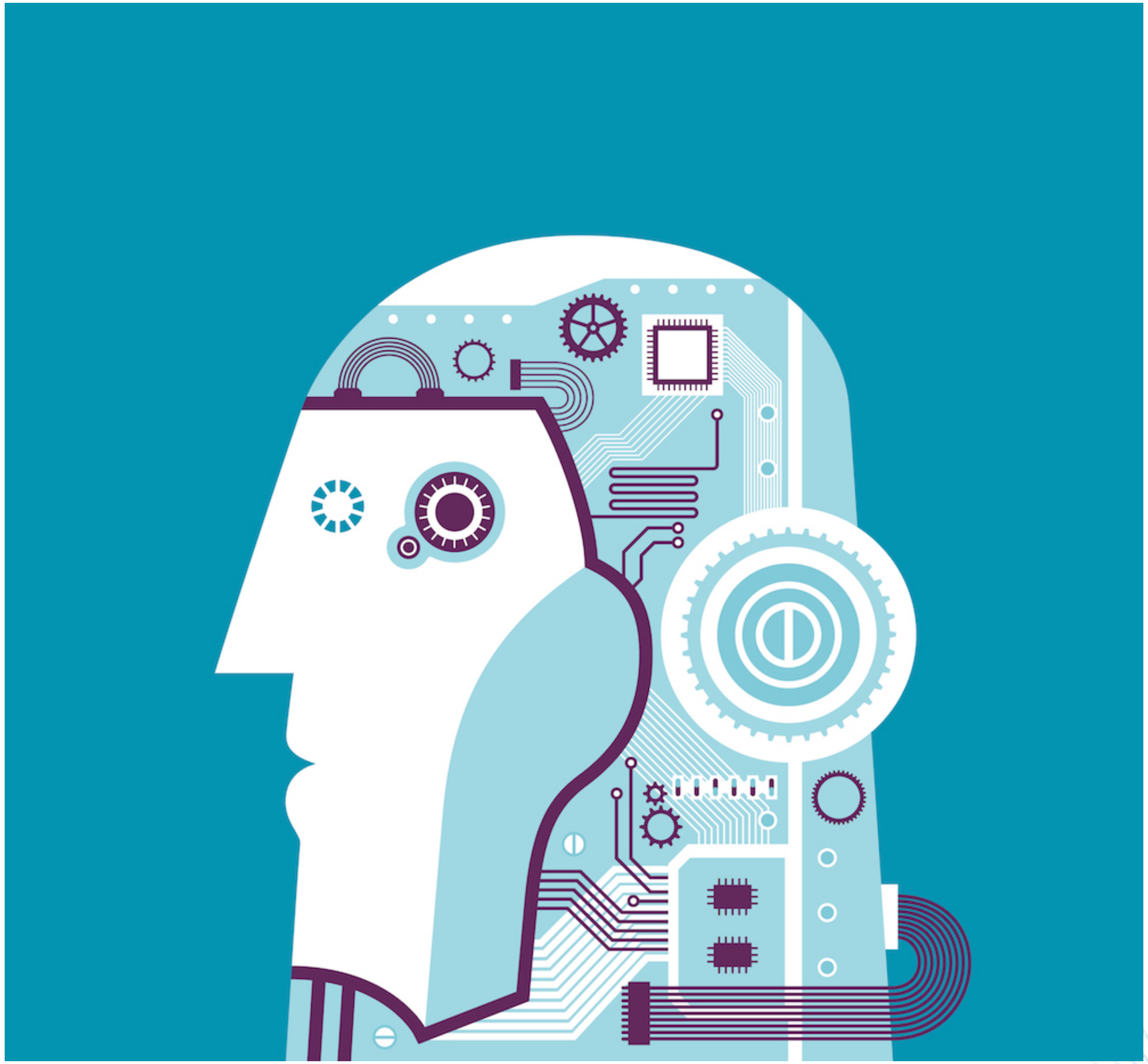
- This is a more modern definition

https://en.wikipedia.org/wiki/Machine_learning

Example 1: Spam Detection

- E = Watching you label emails as spam or not spam
- T = Task of classifying emails as spam or not spam
- P = The number (or fraction) of emails correctly classifies as spam/not spam

Machine Learning: Algorithms



Types of algorithms

- Supervised learning
- Unsupervised learning
- Others: reinforcement learning, recommender systems

Supervised Learning

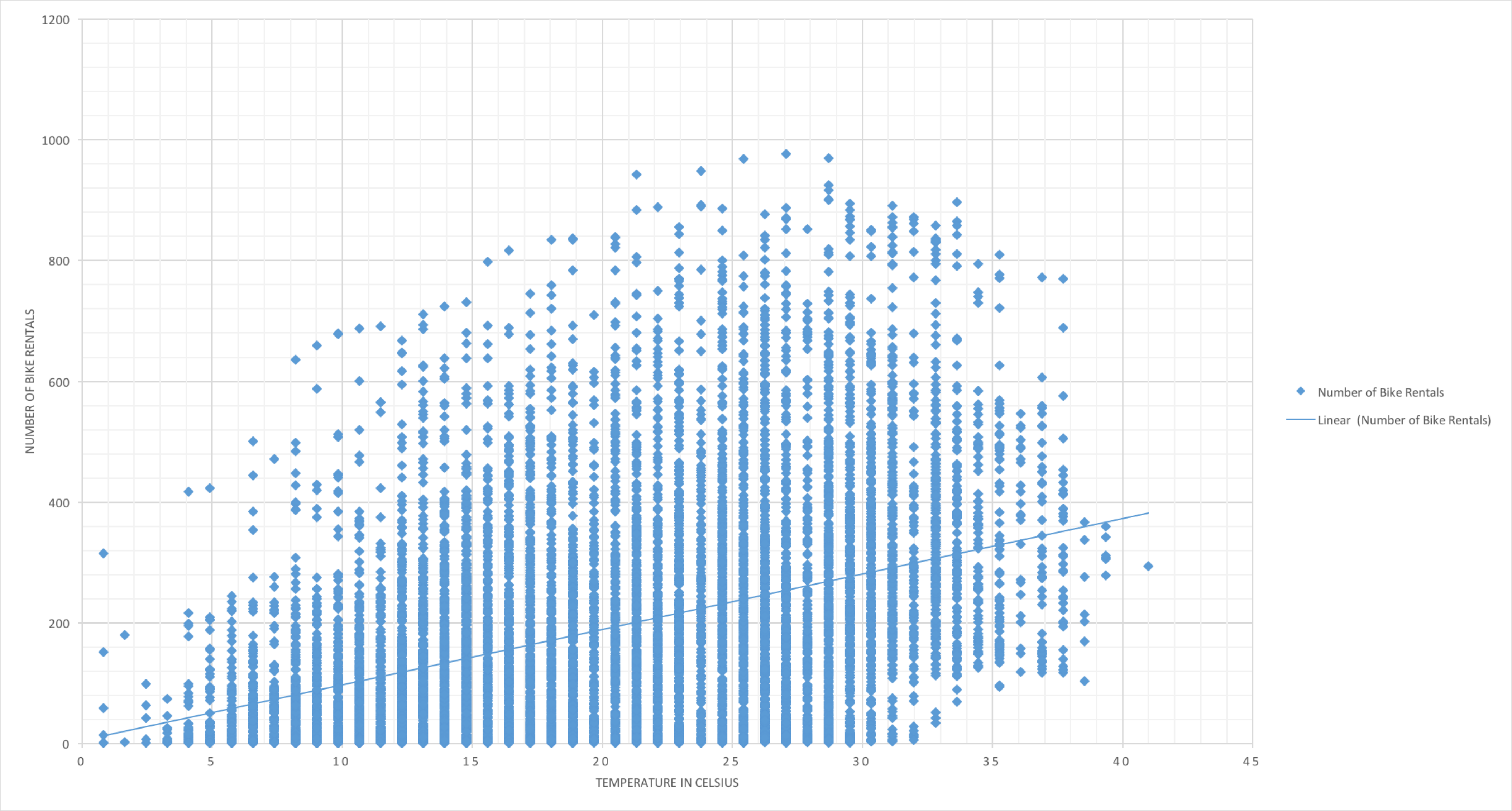
- Definition (Wikipedia): “Supervised learning is the machine learning task of inferring a function from labeled training data.”

- Two common examples:
  1. Regression: map input variables to some continuous function
  2. Classification: map input variables into discrete categories

Regression Example:

Given the data about the temperature and number of bike rentals on various days, try to predict the number of bike rentals in the future based on temperature

Temperature in Celsius	Number of Bike Rentals
9.84	16
9.02	40
9.84	13
9.02	17
15.58	36
20.10	34
42.10	37
8.2	3
13.12	??
20.33	??



cloudera

TECHNICAL SUMMIT 2015

Classification: Example

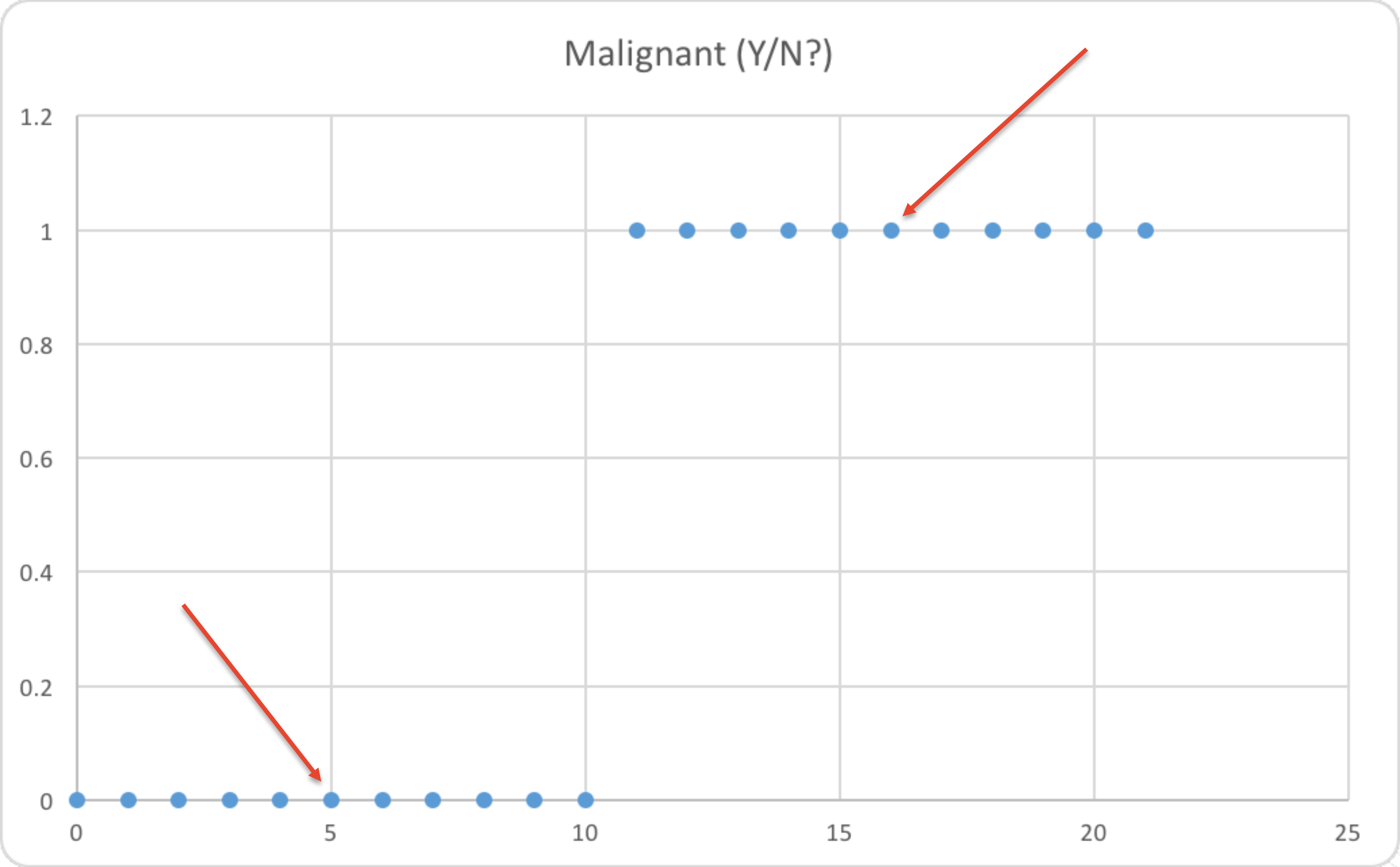
Given the data about the various cancer tumor sizes and whether they are malignant or not, try to predict if a tumor is benign or malignant based on tumor size

Tumor Size	Malignant (Y/N)?
0	N
1	N
3	N
4	N
5	N
11	Y
13	Y
14	Y
16	??
17	??

1 (Y)

Malignant (Y?)

0 (N)



Tumor Size

Unsupervised Learning

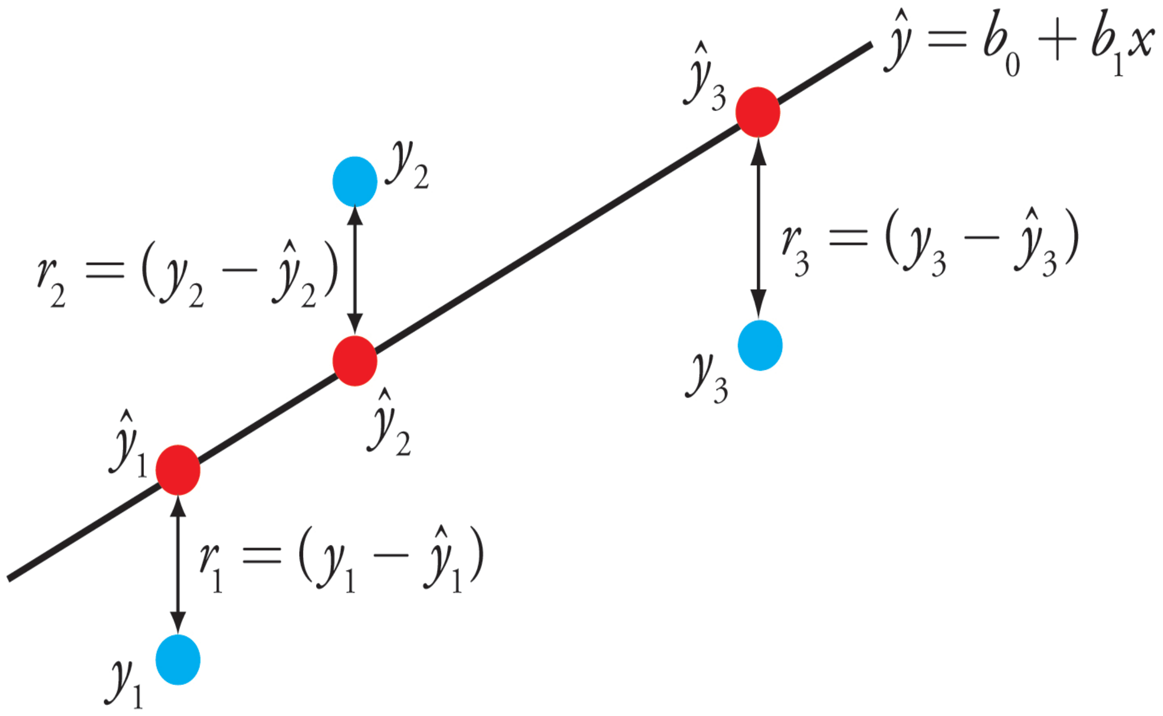
- Definition (Wikipedia): Find hidden structure in unlabeled data.

- Used in exploratory data analysis or as a preprocessing step for supervised task

- Two common examples:
  1. Clustering: group set of similar objects in the same group (called a cluster)
  2. Dimensionality Reduction: reduce the number of variables or features under consideration

Ok. Lets get to the real stuff

Linear Regression: Motivation



Why are we learning linear regression?

It is a **crucial technique to learn** for many reasons:

- widely used and well-understood
- easy to use because minimal “tuning” is required.
- highly “interpretable”, meaning that it’s easy to explain to others
- basis for many other machine learning techniques

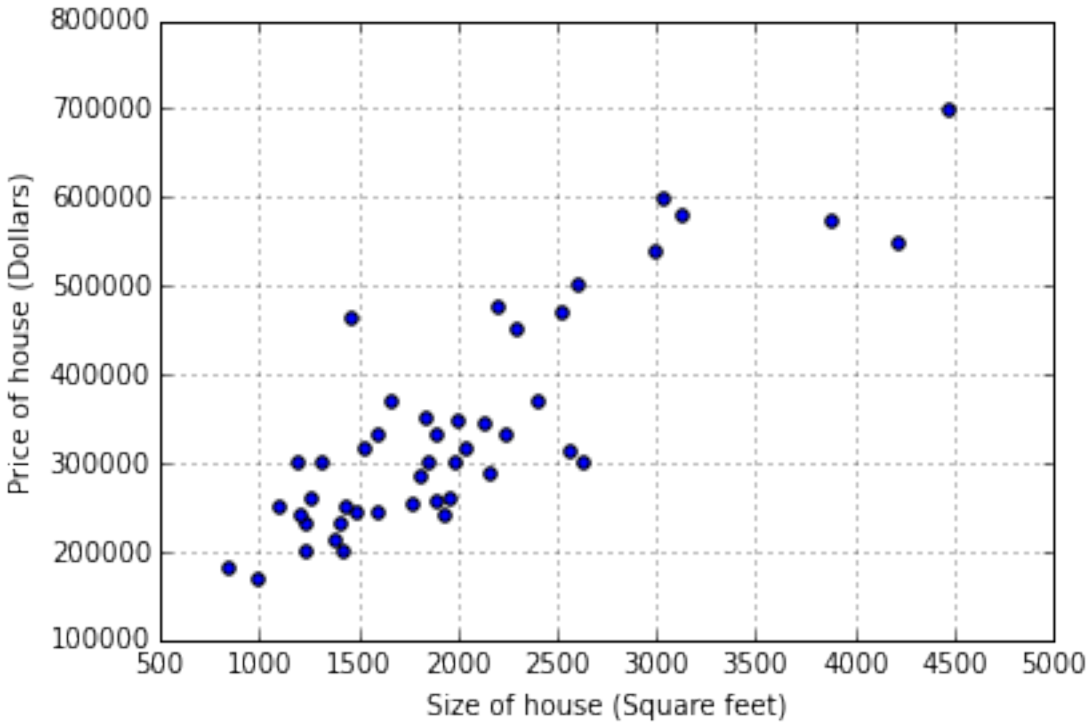
“Importance of having a good understanding of linear regression before studying more complex methods cannot be overstated”

[An Introduction to Statistical Learning – Trevor Hastie, Robert Tibshirani, Gather James and Daniela Witten]

Linear Regression: Deep Dive

Linear Regression: Model Representation - 1

Data: Housing Prices (Portland, Oregon)



Motivation: Predict the house price given the size of a house

Supervised Learning – **"right answers"** are given for each example in the data set

Regression Problem – predict **real-valued** output

Linear Regression: Model Representation - 2

Housing price data

Size of house in square feet (x)	Price of house in \$ (y)
2104	399900
852	179900
2040	314900
4215	549000

x's – "input" variable/features

y's – "output" variable/ "target" variable

m – number of observations (number of training examples). E.g., m = 4

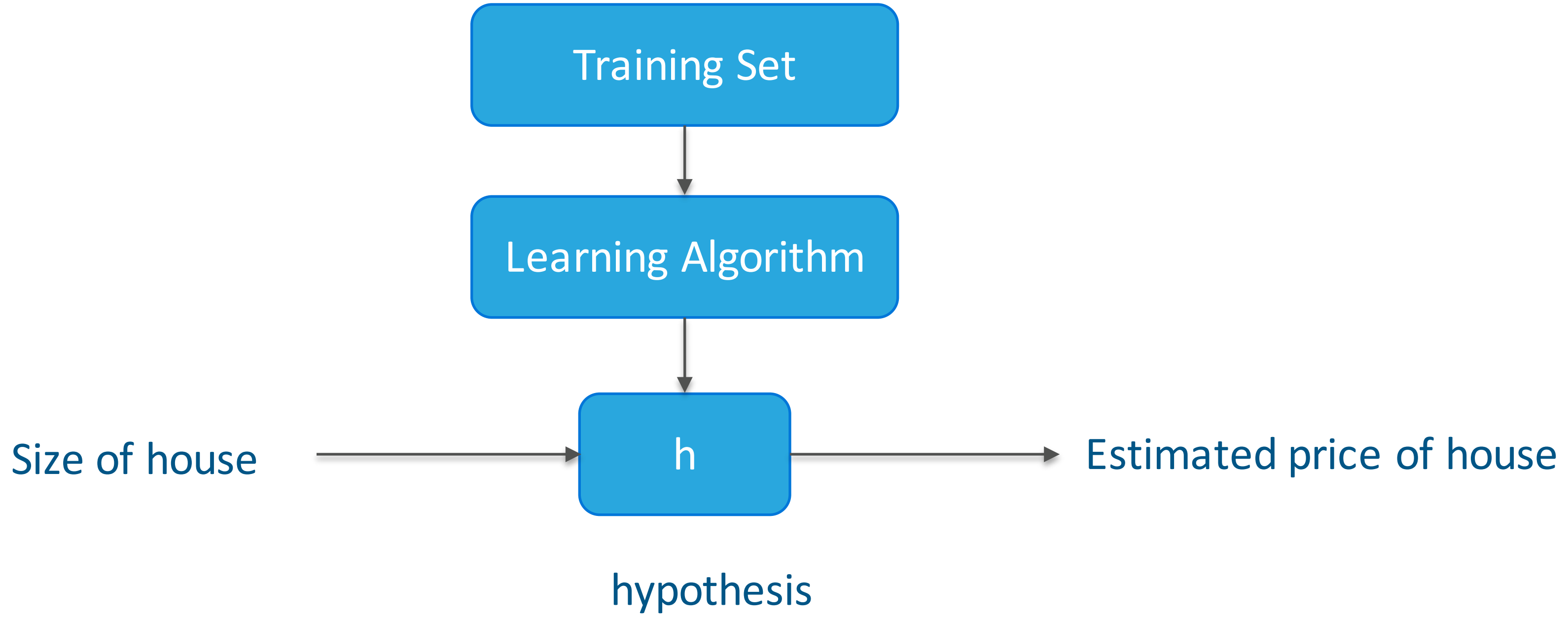
n – number of input variables/features. E.g., n = 1

$(x^{(i)}, y^{(i)})$ - i^{th} training example/observation

$(x^{(3)}, y^{(3)})$ - 3^{rd} observation- (2040,314900)

Linear Regression: Model Representation - 3

Pipeline for supervised learning



h maps from x's to y's

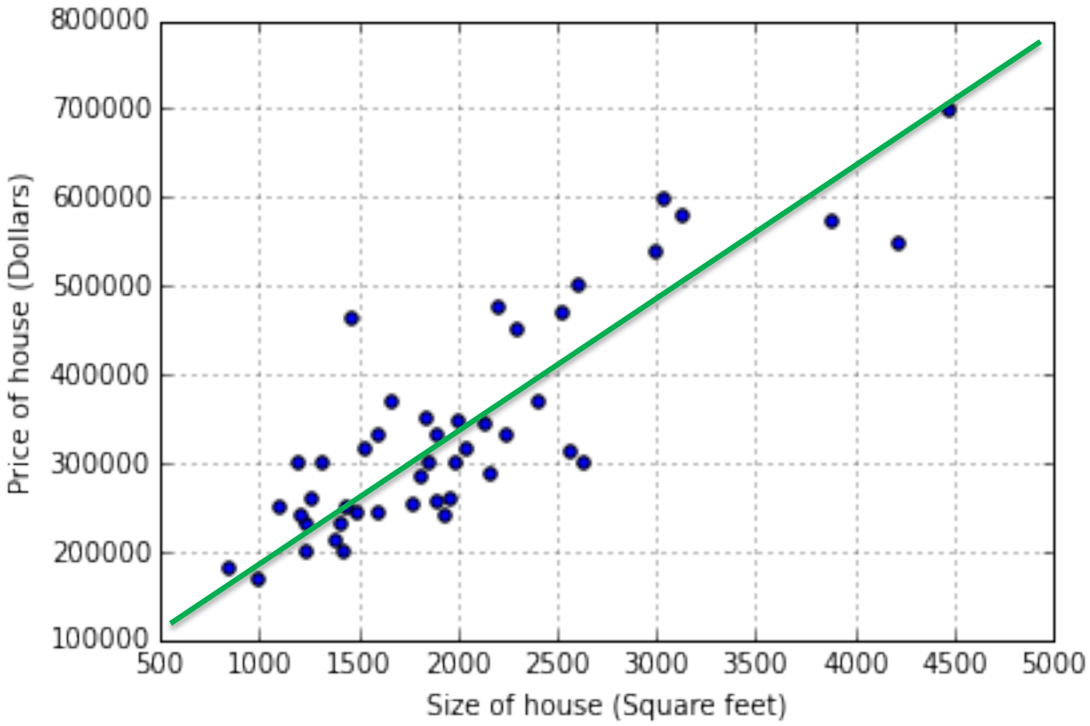
Linear Regression: Model Representation - 4

- Fit a straight line through the data

- Straight line equation

$$h_\theta(x) = \theta_0 + \theta_1 x$$

- Univariate linear regression
  - x – only one input variable



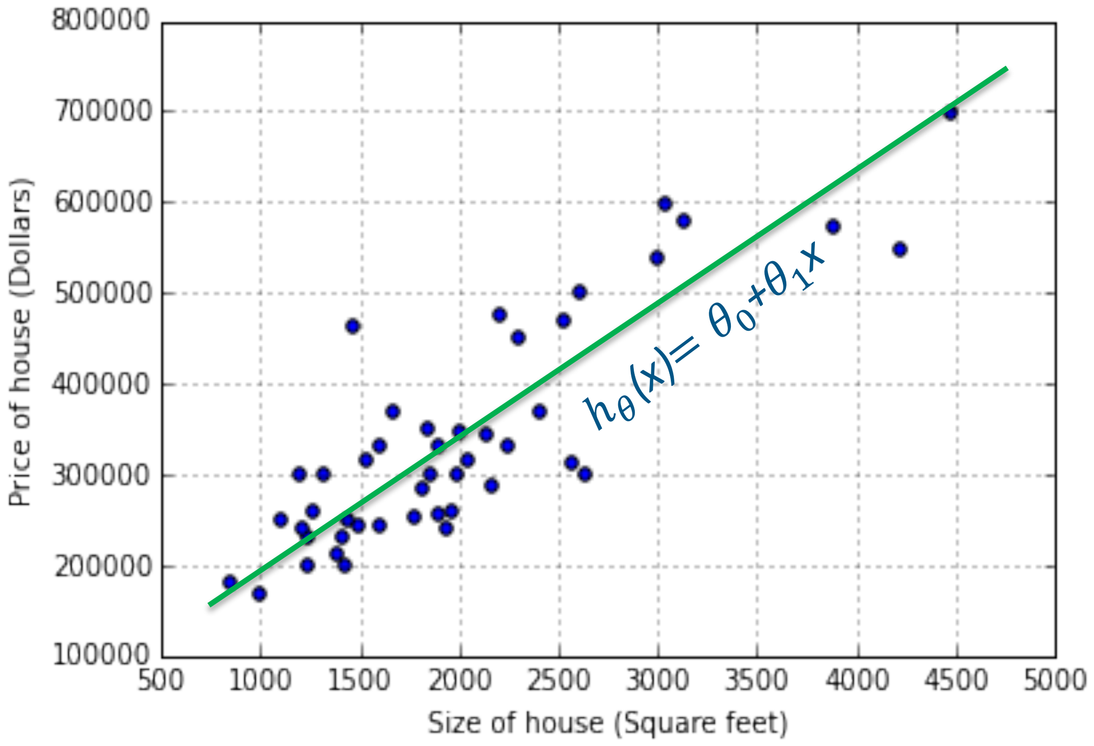
This is very interesting - meow

Linear Regression: Model Representation - 5

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

θ_0, θ_1 - parameters

But how can we find the right parameters?



Solution:

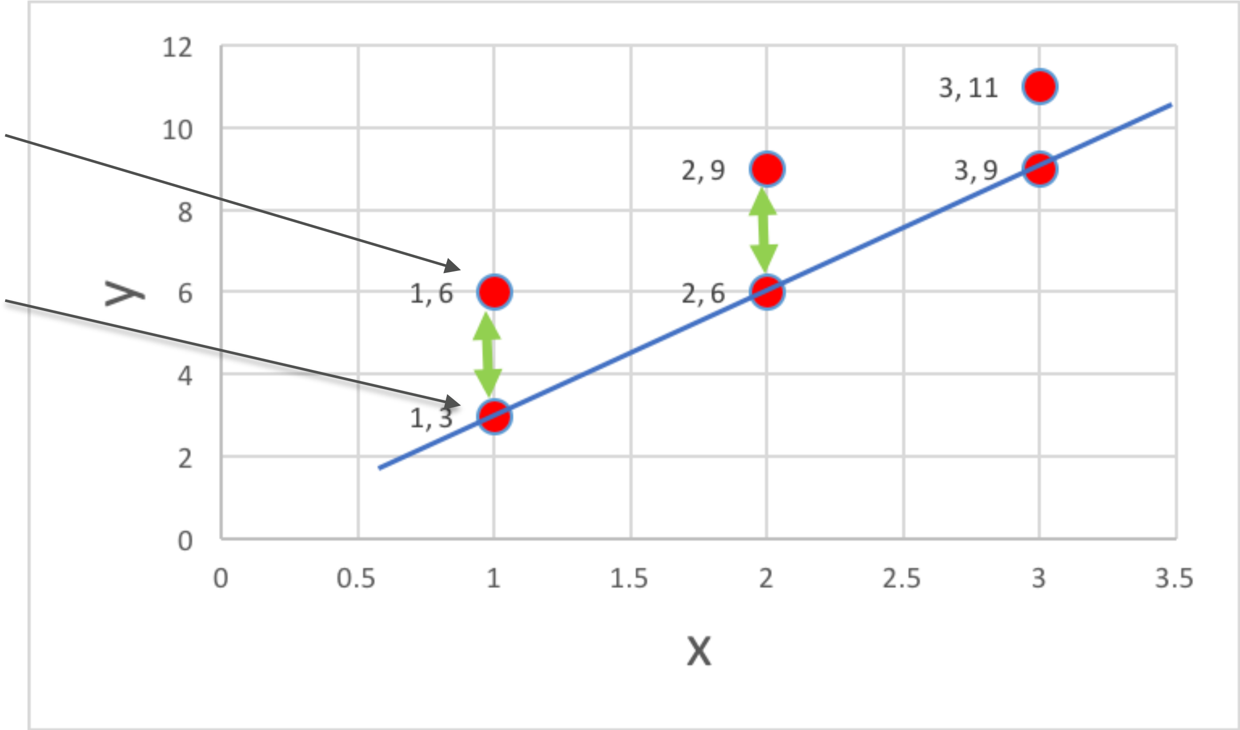
Find θ_0, θ_1 such that the $h_\theta(x)$ is **close** to y for given (x, y) training examples

Linear Regression: Model Representation - 6

What do we mean by <u>**close**</u>?

$(x^{(i)}, h_\theta(x^{(i)}))$

$(x^{(i)}, y^{(i)})$



Euclidean distance between above two points is

$$d = \sqrt{(x^{(i)} - x^{(i)})^2) + (h_\theta(x^{(i)}) - y^{(i)})^2}$$

Squared distance, $d^2 = (h_\theta(x^{(i)}) - y^{(i)})^2$

Linear Regression: Model Representation - 7

- Different θ_0, θ_1 parameters give different hypothesis

- Our goal is to minimize the squared error distance between predicted values (hypothesis) $h_\theta(x^{(i)})$ and the actual values $y^{(i)}$ in our training data set.

- For m training examples, squared error $= \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

- Average/Mean Squared Error = $\frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Linear Regression: Model Representation -8

- Cost Function / Loss Function $J(\theta_0, \theta_1)$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

- Formalize the goal - $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

- This is called Least Squares Regression algorithm

cloudera

TECHNICAL SUMMIT 2015

Please.. can't wait to get my hands on this algorithm

Linear Regression: Model Recap

- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$
- Parameters: θ_0, θ_1
- Cost Function / Loss Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$
- Goal: $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$
- Algorithm Name: Least Squares Regression

Linear Regression: Solution

- Closed-form Solution:
  - Normal Equation Method

- Iterative Solution:
  - Batch Gradient Descent
  - Stochastic Gradient Descent

Linear Regression: Intuition

Training data

Size of house (x)	Price of house(y)
2104	399990
1600	369000
1427	198999
1380	212000

number of training examples m = 4

number of input features/variables n = 1

$$X = \begin{array}{c} \begin{bmatrix} 1 & 2104 \\ 1 & 1600 \\ 1 & 1427 \\ 1 & 1380 \end{bmatrix} \\ \quad x_0 \qquad\quad x_1 \end{array} \qquad y = \begin{bmatrix} 399990 \\ 369000 \\ 198999 \\ 212000 \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

X -> $m \times (n+1)$

y -> m $dimensional\ vector$

θ -> $n+1$ $dimensional\ vector$

x_0 is a dummy/extra feature.

→ $h_\theta(x^{(1)}) = \theta_0\ x_0 + \theta_1 x_1$

→ $h_\theta(x^{(1)}) = \theta_0 \times 1 + \theta_1 \times x_1$

→ $h_\theta(x^{(1)}) = \theta_0 + \theta_1 x_1$

Linear Regression: Normal Equation

- Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$
- Goal: $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$
- Minimization (from calculus)
  - Set partial derivatives of cost function $J(\theta_0, \theta_1)$ with respect to θ_0, θ_1 to zero and solve for θ_0, θ_1

$$\theta = (X^T X)^{-1} X^T y$$

Lets get to the demo

Summary: what we've covered so far

- Machine Learning Motivation
- Machine Learning Definition
- Machine Learning Algorithms
- Linear Regression Motivation
- Linear Regression Deep Dive
- Demo: Linear Regression

Questions?

Thank you

email: gmedasani@cloudera.com

twitter: @gurumedasani