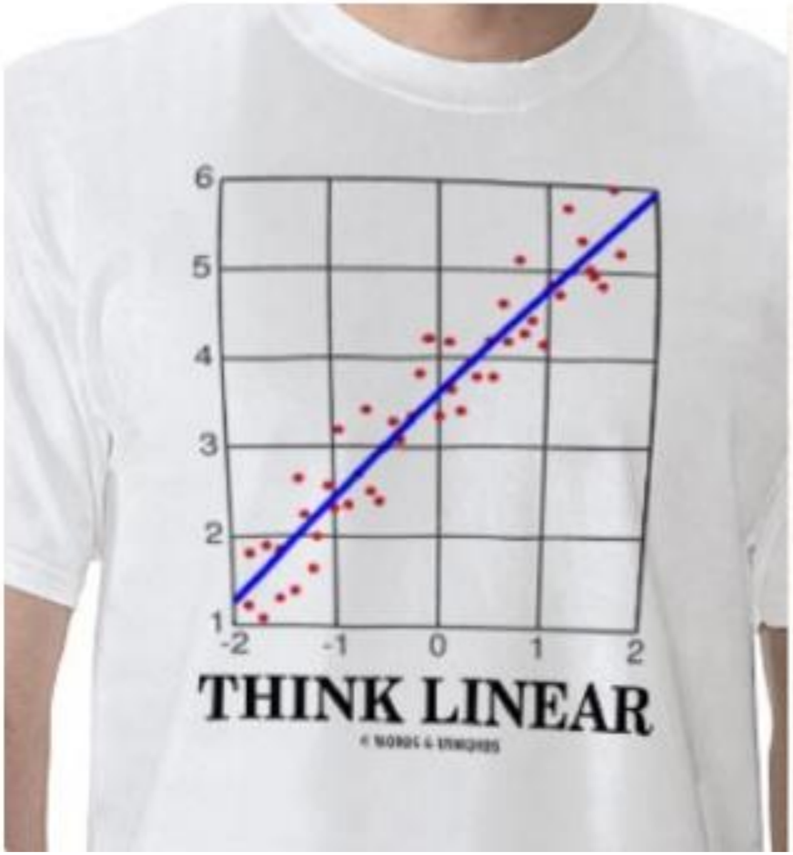
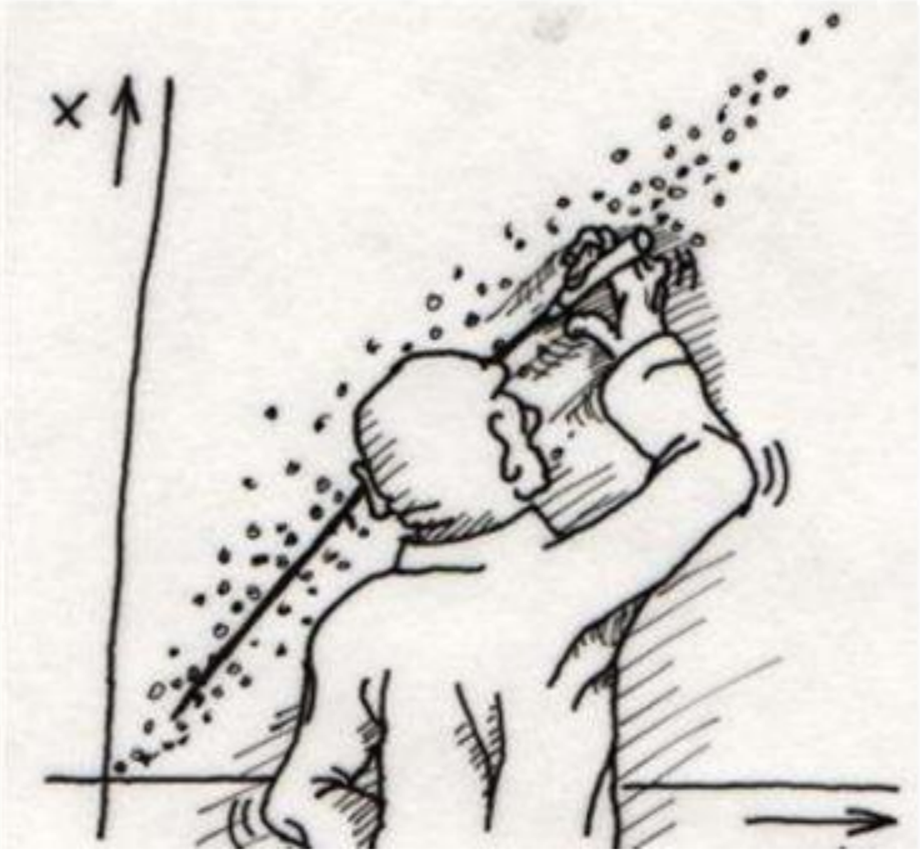
REGRESSION

Presented By:

Rao Muhammad Umer

REGRESSION





REGRESSION

- Section-I:
  - Motivation
- Section-II:
  - Theoretical Bases
- Section-III:
  - Implementation

REGRESSION

PART-I:

Motivation

Outline

- What?
- Why?
- Who?
- How?

Outline

- **What?**
- **Why?**
- **Who?**
- **How?**

Regression

- What about **Machine Learning**?
- What about **Deep Learning**?
- What about **Data Science**?
- What about **Big Data Analysis**?
- What about **Predictive Analytics** ?

Regression

- Predicting a real numeric value for an entity with a given set of features

Deterministic Vs Probabilistic Modeling

- $\mathbf{Y_i = a + b\ X_i}$
  - Substituting a value of X in the equation, we can completely determine a unique value of Y.(Exact relationship)
  - Examples: $\mathbf{F = 32 + (9/5)\ C}$, $\boldsymbol{A = \pi r^2}$

- $\mathbf{Y_i = a + b\ X_i + e_i}$
  - Inexact relationship b/w variables
  - $\mathbf{e_i}$ is known as random error

Types of Regression

- Simple Linear Regression
- Multiple Regression
- Polynomial Regression
- Logistic Regression

Outline

- What?
- **Why?**
- Who?
- How?

Predicting House Price

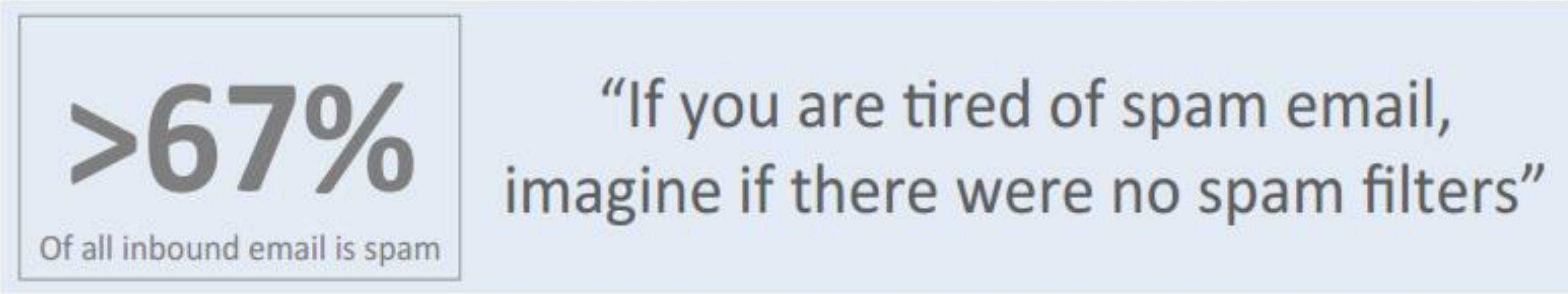
Predicting House Price

Predicting which Television Show will have more viewers for next week

What Causes Retweets?

- How can we help marketers use Twitter data to influence the number of times a tweet is retweeted?
- With over 555 MM registered users who tweet 58 MM times per day, Tweeting is potentially big business.
- We selected the top 30 entertainment industry related accounts to find out, ranging from rock stars, actors, to famous sports figures.
- In particular, we are attempting to understand the factors that contribute to a “retweet”, i.e. an instance where a Twitter account holder’s tweet is shared by another Twitter user.
- This has important implications for marketers.
- By understanding the factors that contribute to a retweet a marketer can maximize the potential audience and can better engage with target audiences.

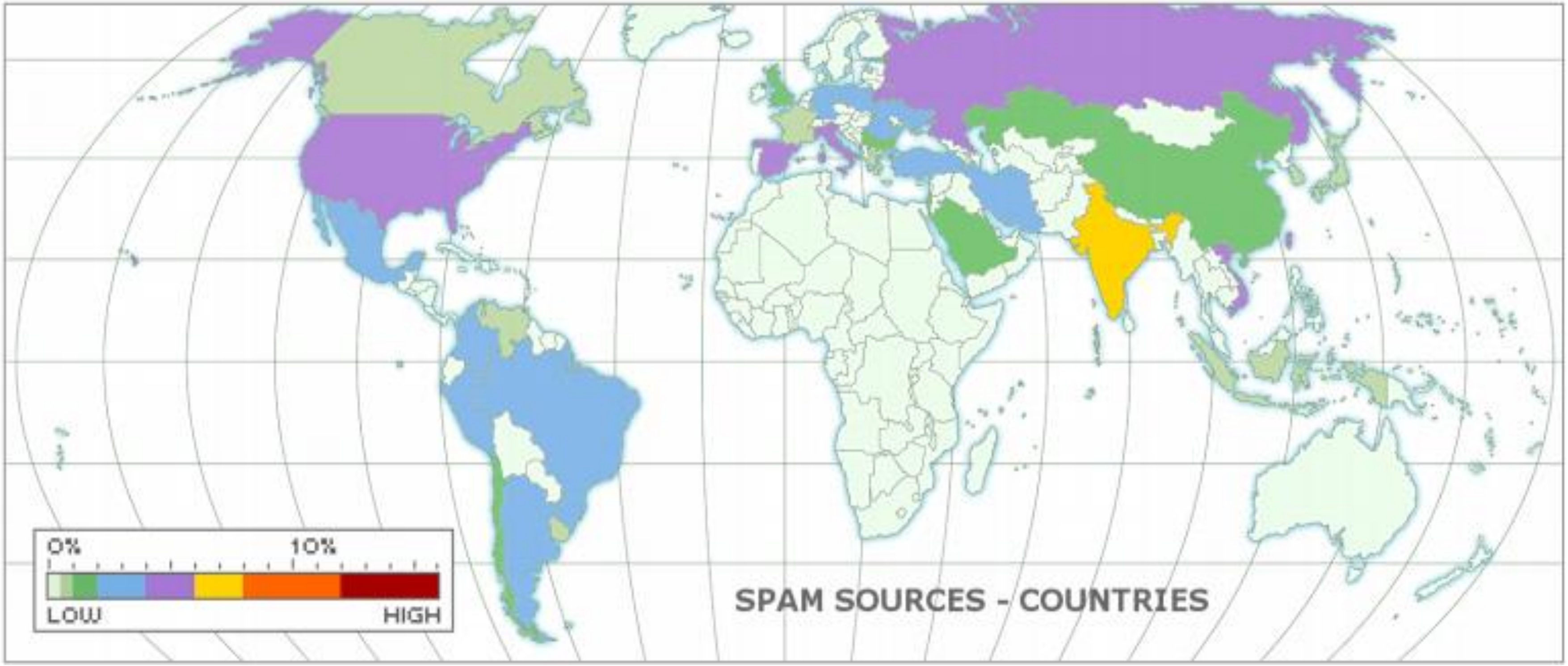
SPAM CLASSIFICATION



>67%

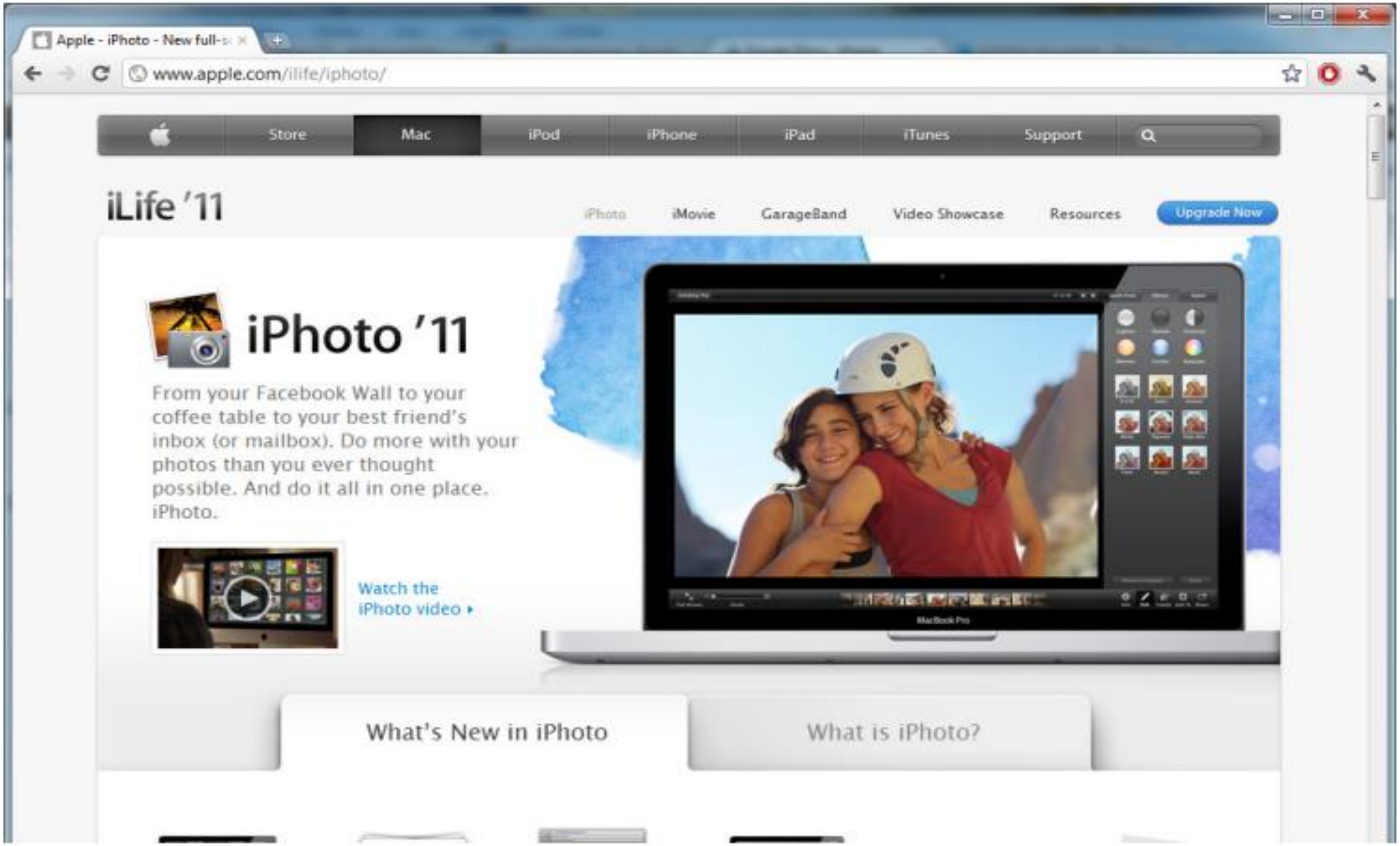
Of all inbound email is spam

"If you are tired of spam email, imagine if there were no spam filters"



SPAM CLASSIFICATION

Tagging a Photo!!!



From your Facebook Wall to your coffee table to your best friend's inbox (or mailbox). Do more with your photos than you ever thought possible. And do it all in one place. iPhoto.

Self Customizing Programs

- Self customizing programs
  - Netflix
  - Amazon

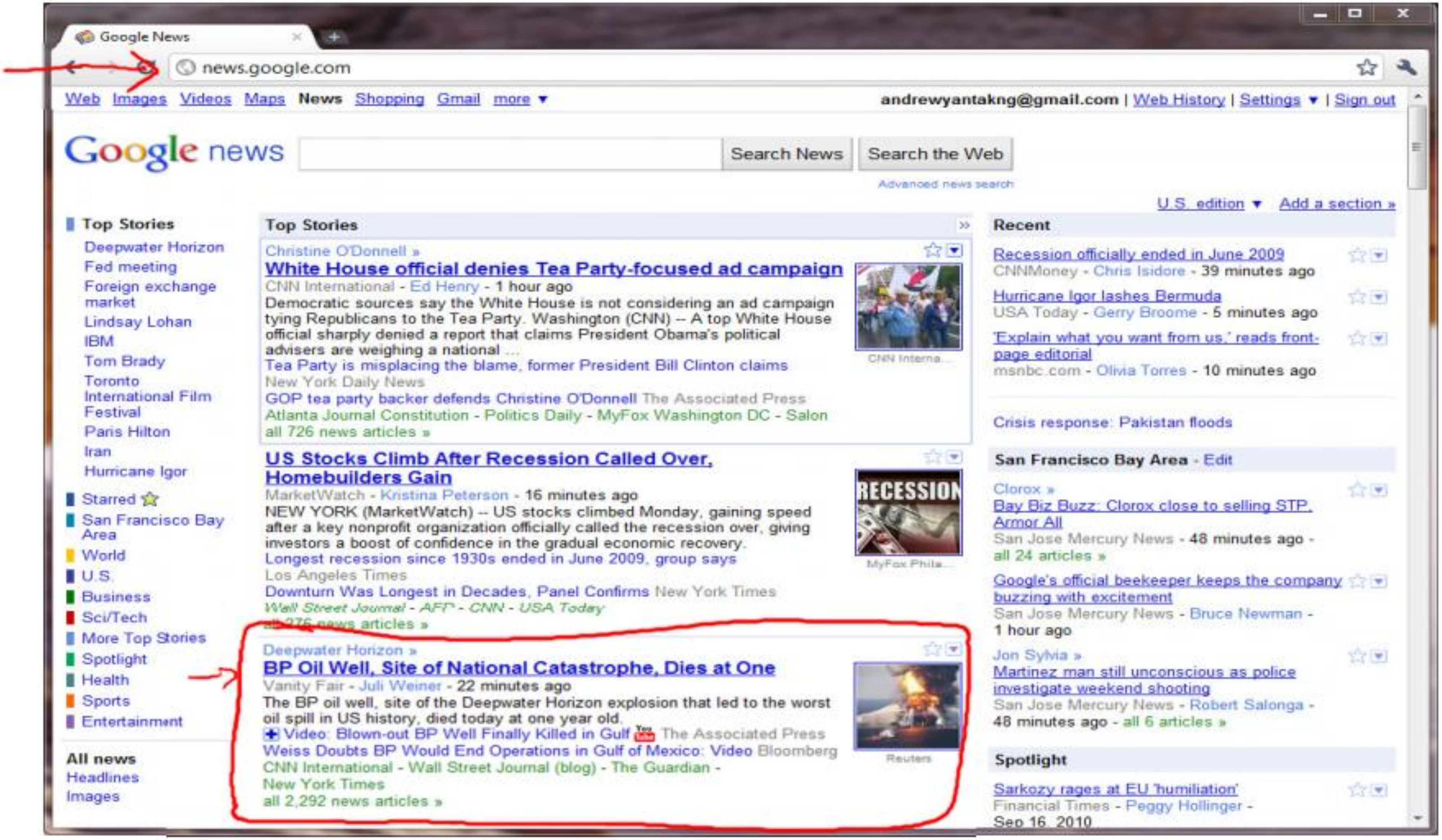
Stock Exchange Price Prediction

- Predicting a continuous response like stock price

Detecting Fraudulent Transaction

- Online Transactions:
  - Fraudulent((Yes(/(No)?
- Credit card fraud
- Online payment fraud
- Spam instant messages etc.

Clustering



Web Images Videos Maps **News** Shopping Gmail more ▾ **andrewyantakng@gmail.com** | Web History | Settings ▾ | Sign out

Top Stories
Deepwater Horizon
Fed meeting
Foreign exchange market
Lindsay Lohan
IBM
Tom Brady
Toronto International Film Festival
Paris Hilton
Iran
Hurricane Igor

Starred ☆
San Francisco Bay Area
World
U.S.
Business
Sci/Tech
More Top Stories
Spotlight
Health
Sports
Entertainment

All news
Headlines
Images

Top Stories

Christine O'Donnell »
White House official denies Tea Party-focused ad campaign
CNN International - Ed Henry - 1 hour ago
Democratic sources say the White House is not considering an ad campaign tying Republicans to the Tea Party. Washington (CNN) -- A top White House official sharply denied a report that claims President Obama's political advisers are weighing a national ...
Tea Party is misplacing the blame, former President Bill Clinton claims New York Daily News
GOP tea party backer defends Christine O'Donnell The Associated Press
Atlanta Journal Constitution - Politics Daily - MyFox Washington DC - Salon
all 726 news articles »
CNN Interna...

US Stocks Climb After Recession Called Over, Homebuilders Gain
MarketWatch - Kristina Peterson - 16 minutes ago
NEW YORK (MarketWatch) -- US stocks climbed Monday, gaining speed after a key nonprofit organization officially called the recession over, giving investors a boost of confidence in the gradual economic recovery.
Longest recession since 1930s ended in June 2009, group says Los Angeles Times
Downturn Was Longest in Decades, Panel Confirms New York Times
Wall Street Journal - AFP - CNN - USA Today
all 276 news articles »
RECESSION
MyFox Phila...

Deepwater Horizon »
BP Oil Well, Site of National Catastrophe, Dies at One
Vanity Fair - Juli Weiner - 22 minutes ago
The BP oil well, site of the Deepwater Horizon explosion that led to the worst oil spill in US history, died today at one year old.
+ Video: Blown-out BP Well Finally Killed in Gulf The Associated Press
Weiss Doubts BP Would End Operations in Gulf of Mexico: Video Bloomberg
CNN International - Wall Street Journal (blog) - The Guardian - New York Times
all 2,292 news articles »
Reuters

Recent

Recession officially ended in June 2009
CNNMoney - Chris Isidore - 39 minutes ago
Hurricane Igor lashes Bermuda
USA Today - Gerry Broome - 5 minutes ago
'Explain what you want from us,' reads front-page editorial
msnbc.com - Olivia Torres - 10 minutes ago

Crisis response: Pakistan floods

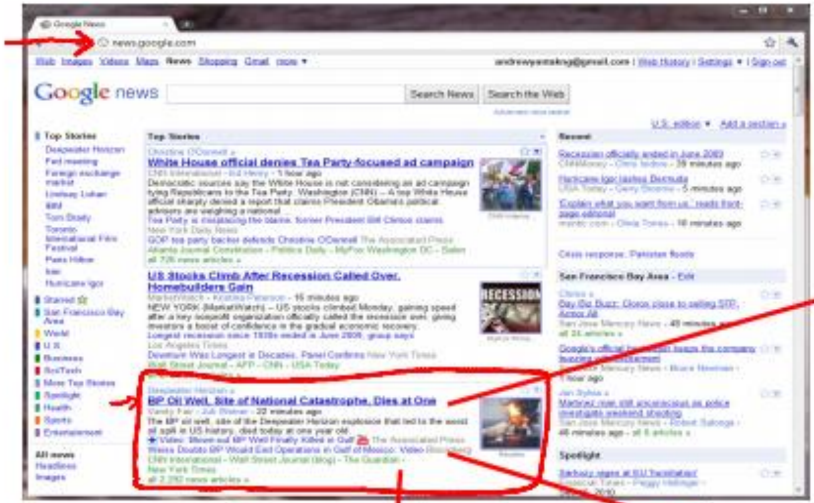
San Francisco Bay Area - Edit

Clorox »
Bay Biz Buzz: Clorox close to selling STP, Armor All
San Jose Mercury News - 48 minutes ago - all 24 articles »
Google's official beekeeper keeps the company buzzing with excitement
San Jose Mercury News - Bruce Newman - 1 hour ago
Jon Sylvia »
Martinez man still unconscious as police investigate weekend shooting
San Jose Mercury News - Robert Salonga - 48 minutes ago - all 6 articles »

Spotlight

Sarkozy rages at EU 'humiliation'
Financial Times - Peggy Hollinger - Sep 16, 2010

Clustering



Google news

Top Stories

White House official denies Tea Party-focused ad campaign

CNN International - 1 hour ago

Democratic sources say the White House is not considering an ad campaign tying Republicans to the Tea Party. Washington (CNN) -- A top White House official sharply denied a report that claims President Obama's political advisers are weighing a national

Tea Party is misplacing the blame, former President Bill Clinton claims

New York Daily News

GOP tea party backer defends Christine O'Donnell The Associated Press

all 726 news articles »

US Stocks Climb After Recession Called Over, Homebuilders Gain

MarketWatch - 15 minutes ago

NEW YORK (MarketWatch) -- US stocks climbed Monday, gaining speed after a key nonprofit organization officially called the recession over, giving investors a boost of confidence in the gradual economic recovery.

Longest recession since 1930s ended in June 2009, group says

Los Angeles Times

Downturn Was Longest in Decades, Panel Confirms New York Times

Wall Street Journal - AFP - CNN - USA Today

BP Oil Well, Site of National Catastrophe, Dies at One

Vanity Fair - 22 minutes ago

The BP oil well, site of the Deepwater Horizon explosion that led to the worst oil spill in US history, died today at one year old.

Video: Blown-out BP Well Finally Killed in Gulf The Associated Press

CNN International - Wall Street Journal (blog) - The Guardian - New York Times

all 2,292 news articles »

Recession officially ended in June 2009

CNNMoney - 39 minutes ago

Hurricane Igor lashes Bermuda

USA Today - 5 minutes ago

'Explain what you want from us,' reads front-page editorial

Crisis response: Pakistan floods

San Francisco Bay Area - Edit

Day Biz Buzz: Clorox close to selling STP, Armor All

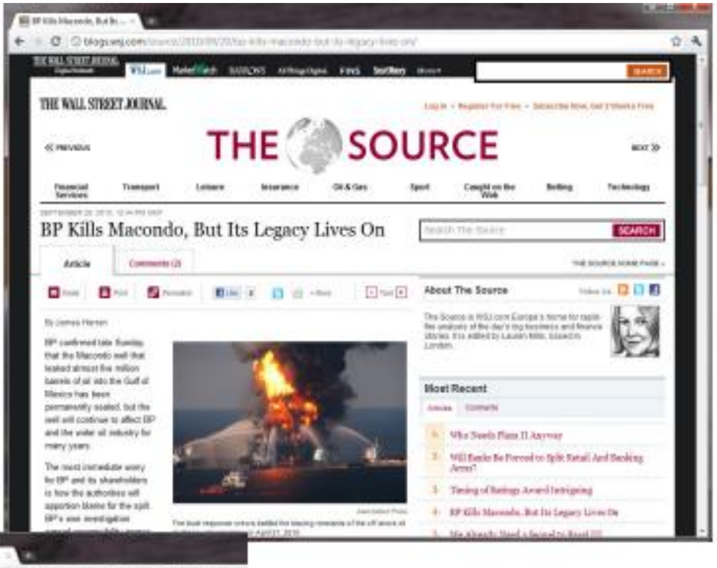
San Jose Mercury News - 49 minutes ago

all 24 articles »

San Jose Mercury News - 1 hour ago

San Jose Mercury News - 46 minutes ago

Sarkozy ...



THE WALL STREET JOURNAL.

THE SOURCE

BP Kills Macondo, But Its Legacy Lives On

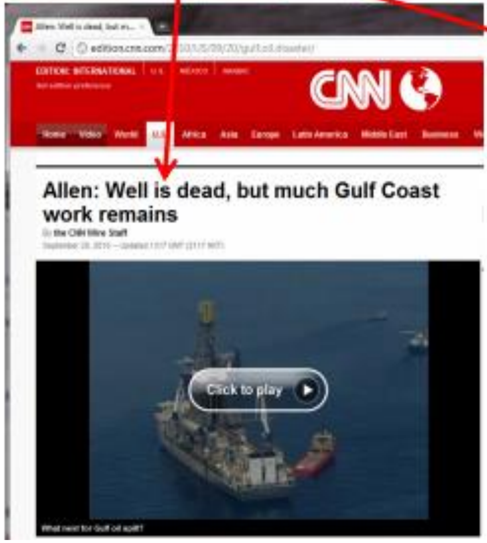
By James Herron

BP confirmed late Sunday that the Macondo well that leaked almost five million barrels of oil into the Gulf of Mexico has been permanently sealed, but the well will continue to affect BP and the wider oil industry for many years.

The most immediate worry for BP and its shareholders is how the authorities will apportion blame for the spill. BP's own investigation

The Source is WSJ.com Europe's home for rapid-fire analysis of the day's big business and finance stories. It is edited by Lauren Mills, based in London.

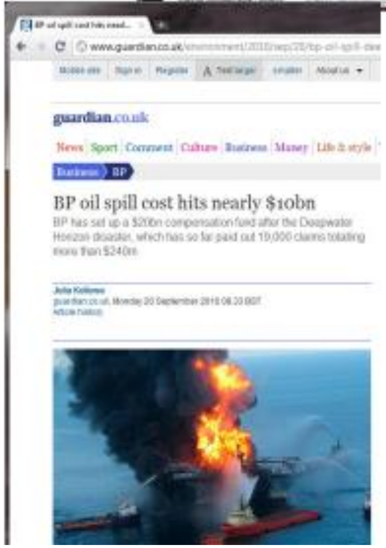
1. Who Needs Plan B Anyway
2. Will Banks Be Forced to Split Retail And Banking Arms?
3. Timing of Ratings Award Intriguing
4. BP Kills Macondo, But Its Legacy Lives On



Allen: Well is dead, but much Gulf Coast work remains

By the CNN Wire Staff

What next for Gulf oil spill?



BP oil spill cost hits nearly $10bn

BP has set up a $20bn compensation fund after the Deepwater Horizon disaster, which has so far paid out 19,000 claims totaling more than $240m

Julia Kollewe

guardian.co.uk, Monday 20 September 2010 08.20 BST

Article history

Outline

- What?
- Why?
- **Who?**
- How?

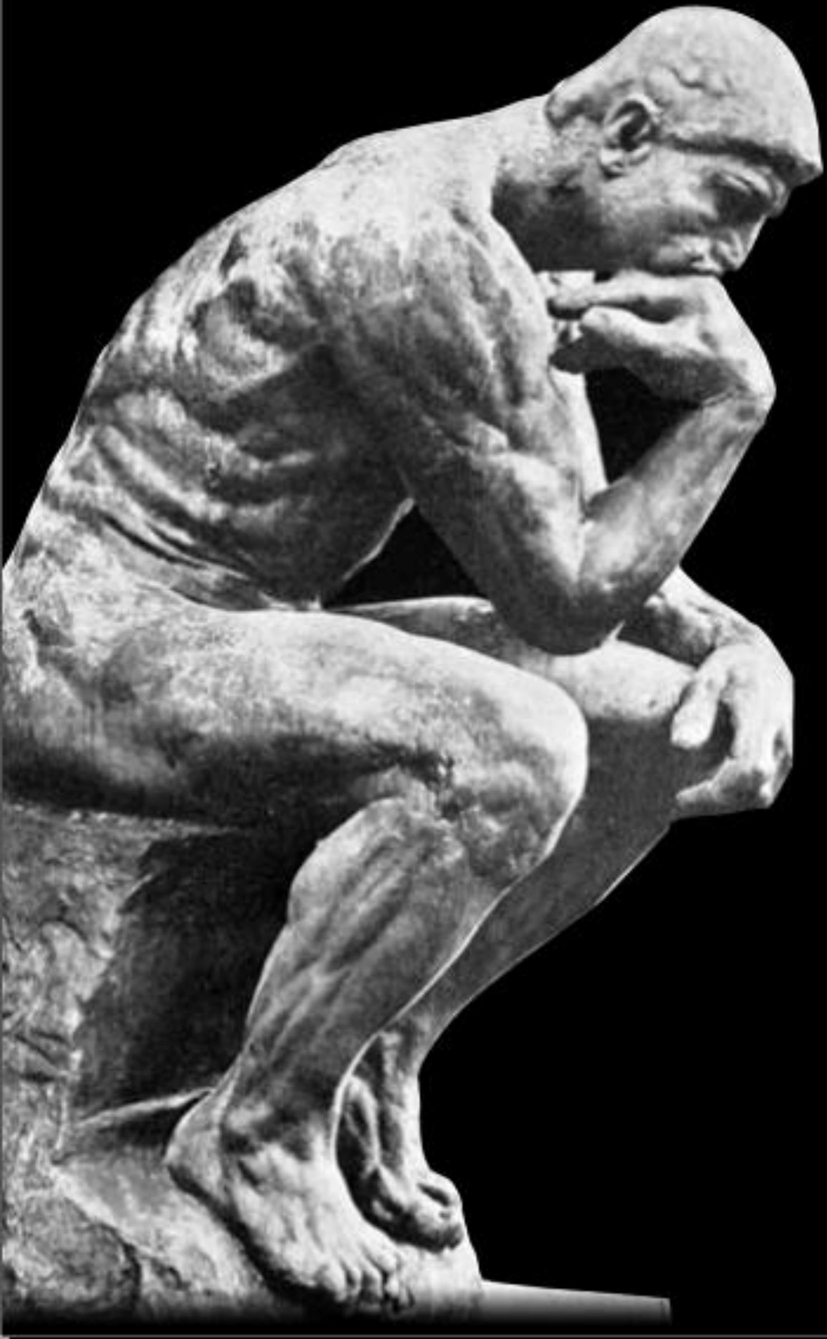
My Background

- **Rao Muhammad Umer**
- MS Computer Science (Continue…)
- Email address: engr.raoumer943@gmail.com
- LinkedIn Profile: https://pk.linkedin.com/in/raomumer
- Github Repository Link: https://github.com/RaoUmer/GPU-Workshop-PIEAS-2016

About You

Outline

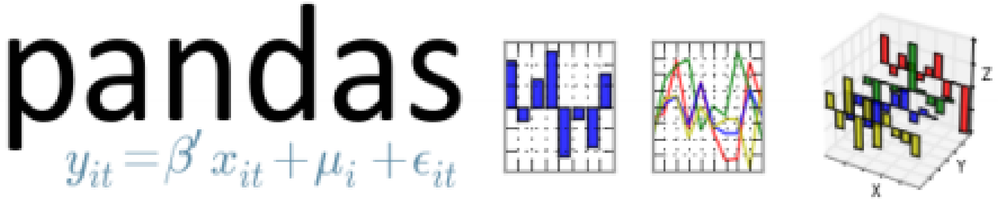
- What?
- Why?
- Who?
- **How?**



Abstractions...

...and Tools

IP[y]:
IPython
Interactive Computing


pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

scikits
learn
machine learning in Python

NumPy

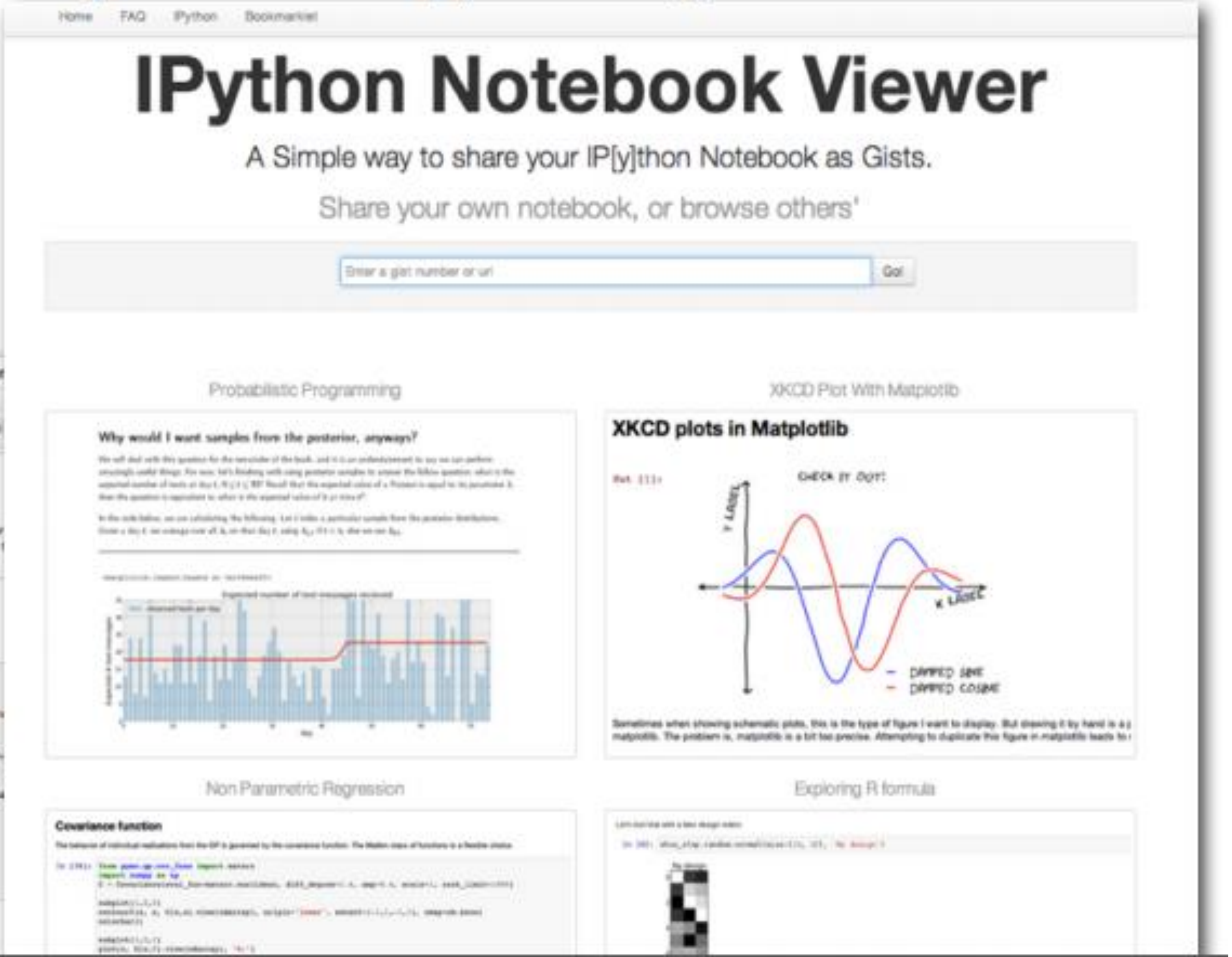
SciPy.org
Sponsored By
ENTHOUGHT

matplotlib

mrjob

IPython Notebooks

http://nbviewer.ipython.org/



IPython Notebook Viewer

A Simple way to share your IP[y]thon Notebook as Gists.

Share your own notebook, or browse others'

Probabilistic Programming

Why would I want samples from the posterior, anyways?

XKCD Plot With Matplotlib

XKCD plots in Matplotlib

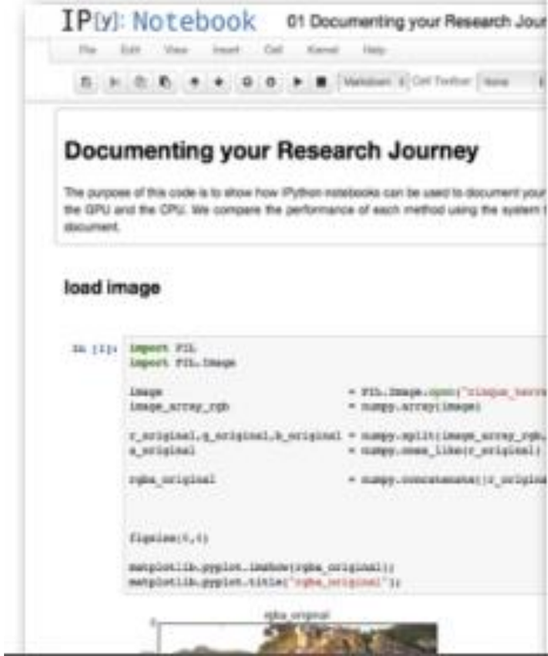
Out [11]:

Sometimes when showing schematic plots, this is the type of figure I want to display. But drawing it by hand is a
matplotlib. The problem is, matplotlib is a bit too precise. Attempting to duplicate this figure in matplotlib leads to

Non Parametric Regression

Covariance function

Exploring R formula



Documenting your Research Journey

The purpose of this code is to show how IPython notebooks can be used to document your
the GPU and the CPU. We compare the performance of each method using the system
document.

load image

```
In [1]: import PIL
        import PIL.Image

        image                            = PIL.Image.open("
        image_array_rgb                  = numpy.array(image)

        r_original,g_original,b_original = numpy.split(image_array_rgb,
        a_original                       = numpy.ones_like(r_original)

        rgba_original                    = numpy.concatenate((r_original



        figsize(4,4)

        matplotlib.pyplot.imshow(rgba_original);
        matplotlib.pyplot.title("rgba_original");
```

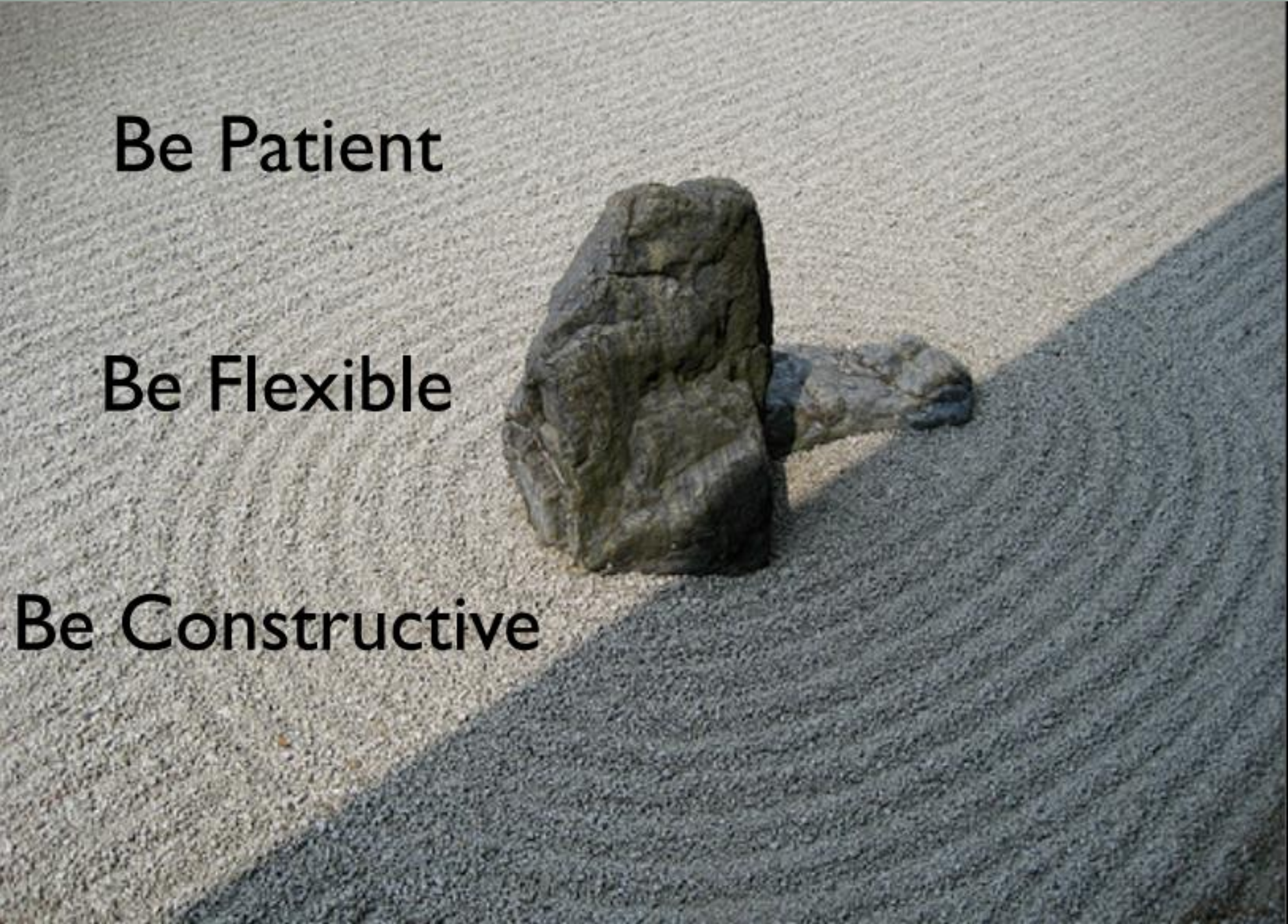
Is this possible for me ???

Prerequisites

- Programming experience

  C, C++, Java, Python, **Cuda C** etc.
- Basic statistical knowledge

  Mean, Standard Deviation, Probability etc.
- Willingness to learn new software & tools
  - This can be time consuming
  - You will need to read online documentation



Be Patient

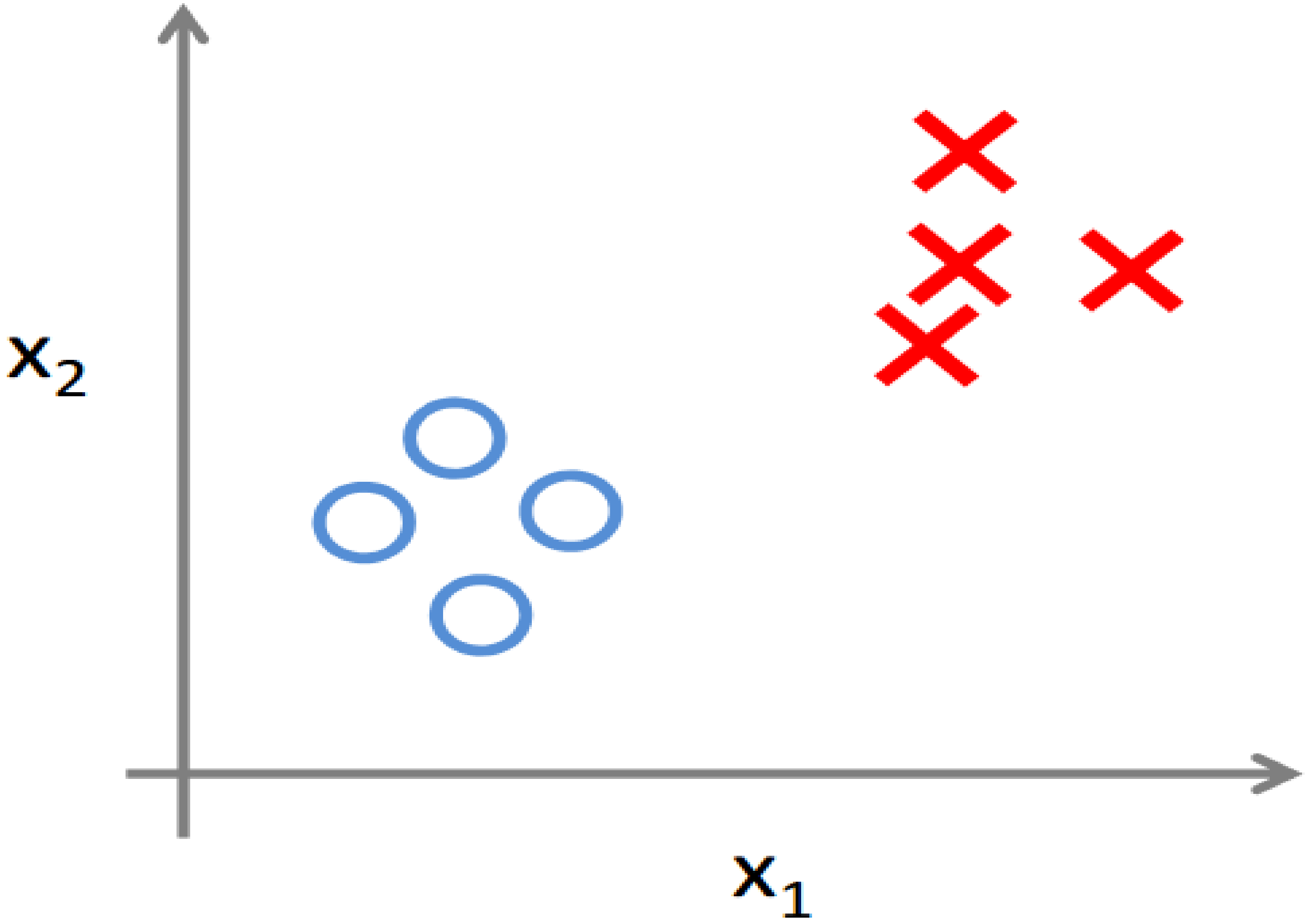
Be Flexible

Be Constructive

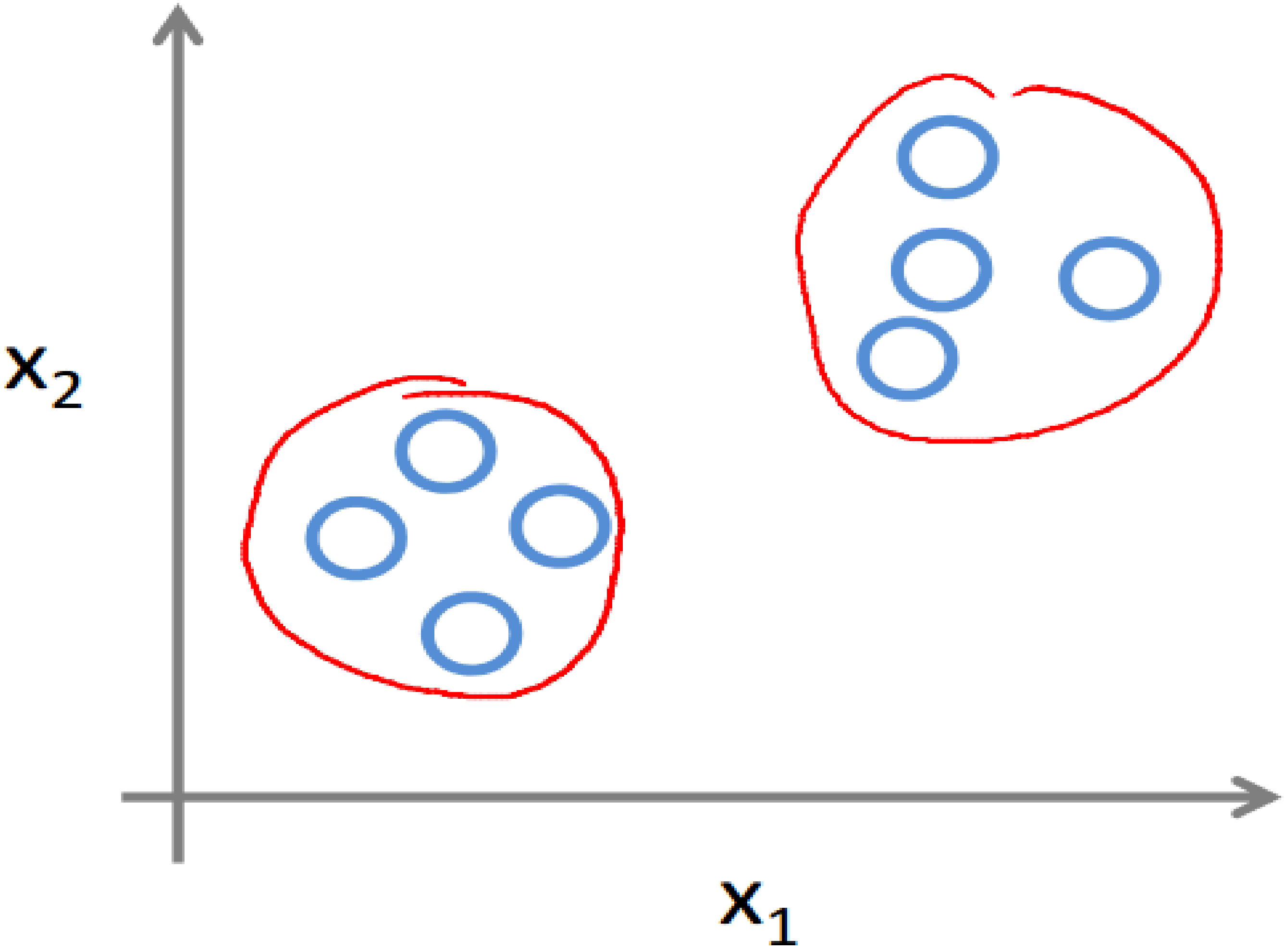
REGRESSION

PART-II:
Theoretical Basics

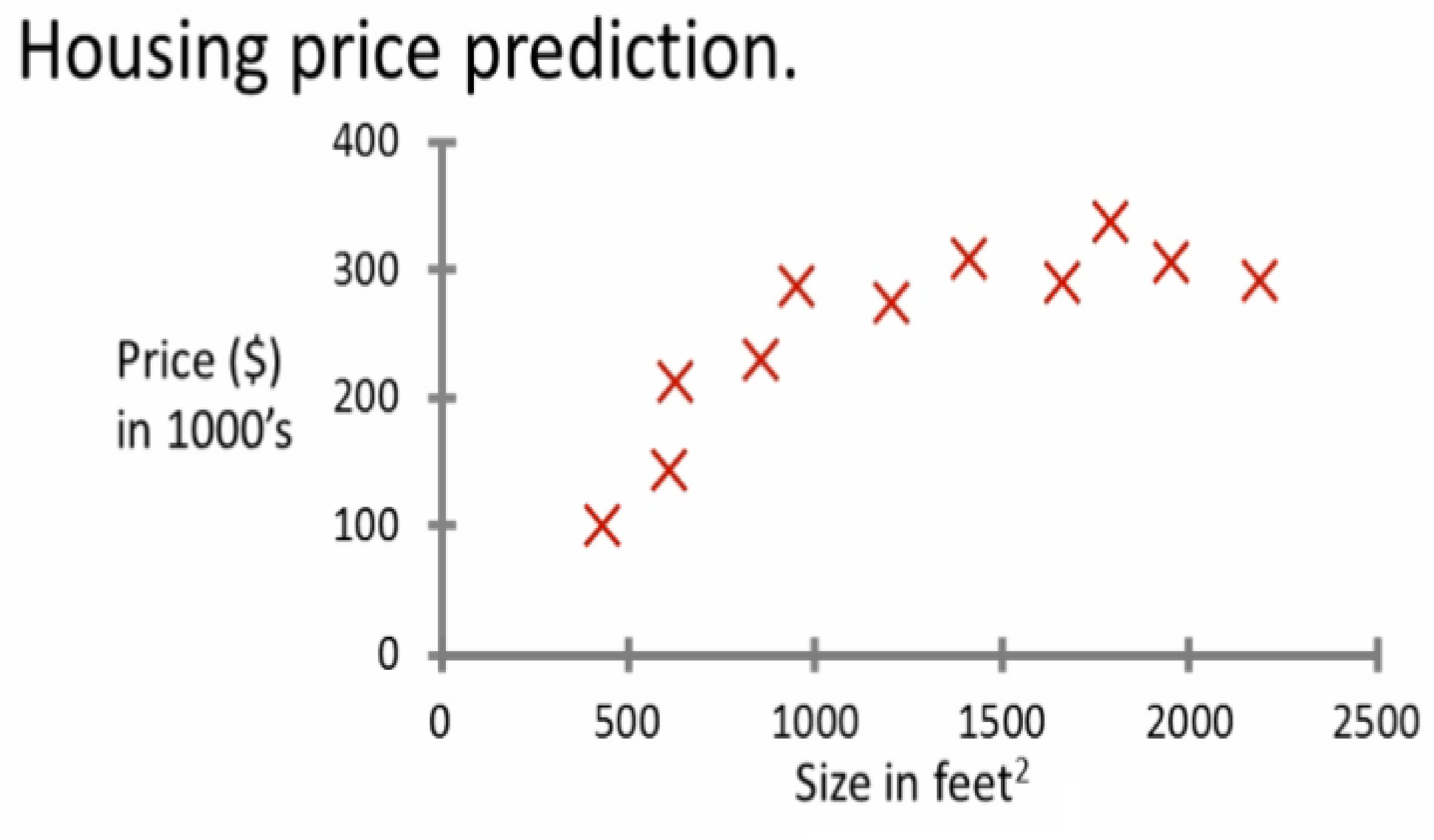
Supervised Learning



Unsupervised Learning



Linear Regression

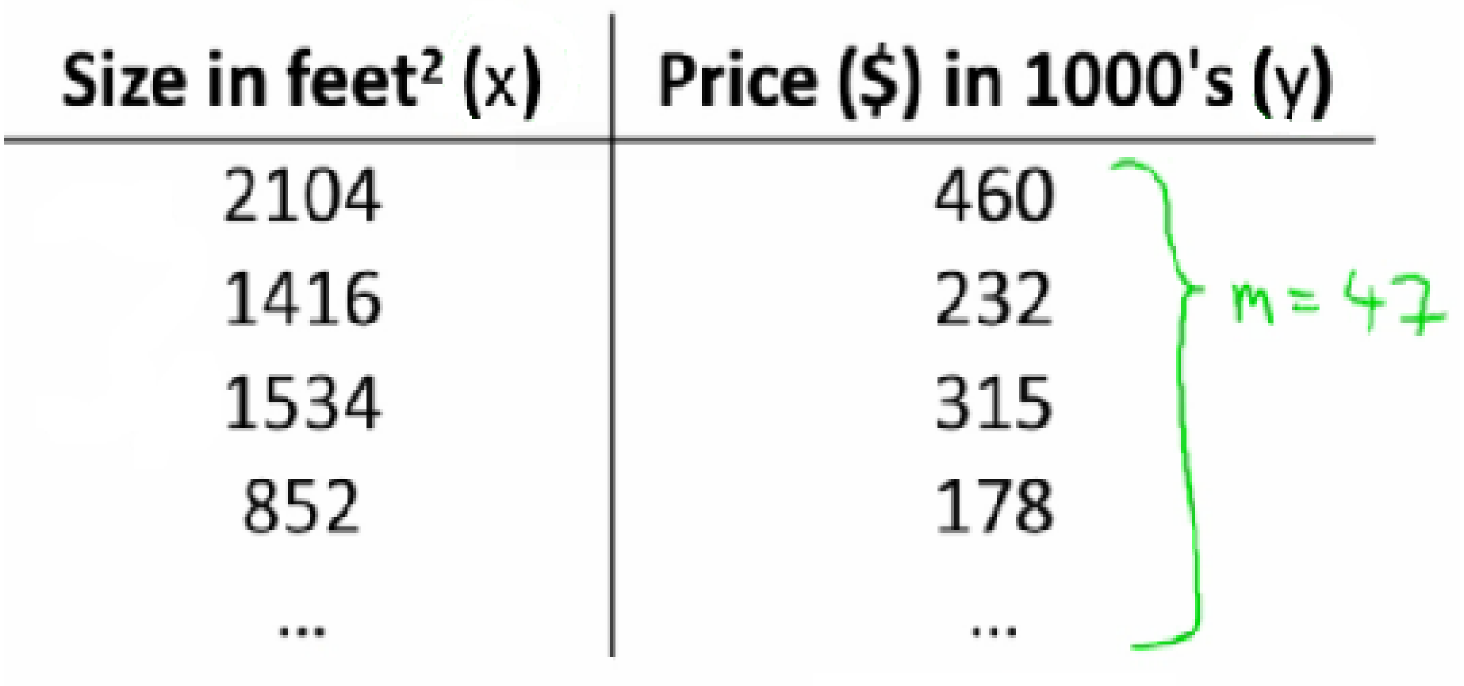


Housing price prediction.

Housing Price Prediction

- Supervised Learning
  - “right answers” given
- Regression: Predict continuous valued output (price)

Training set of Housing Prices

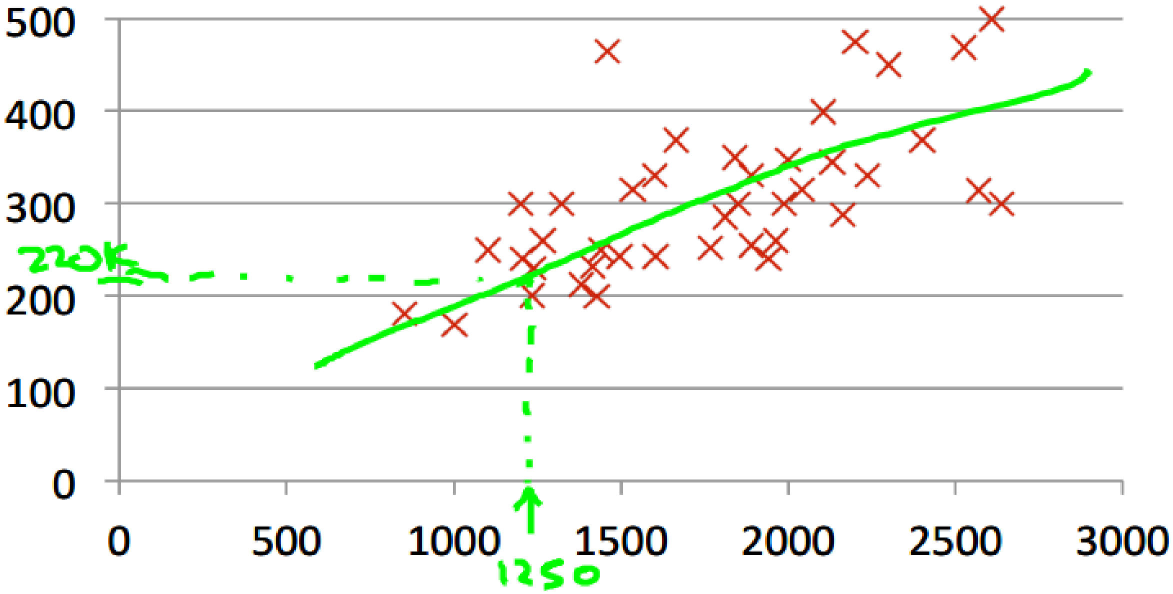


Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

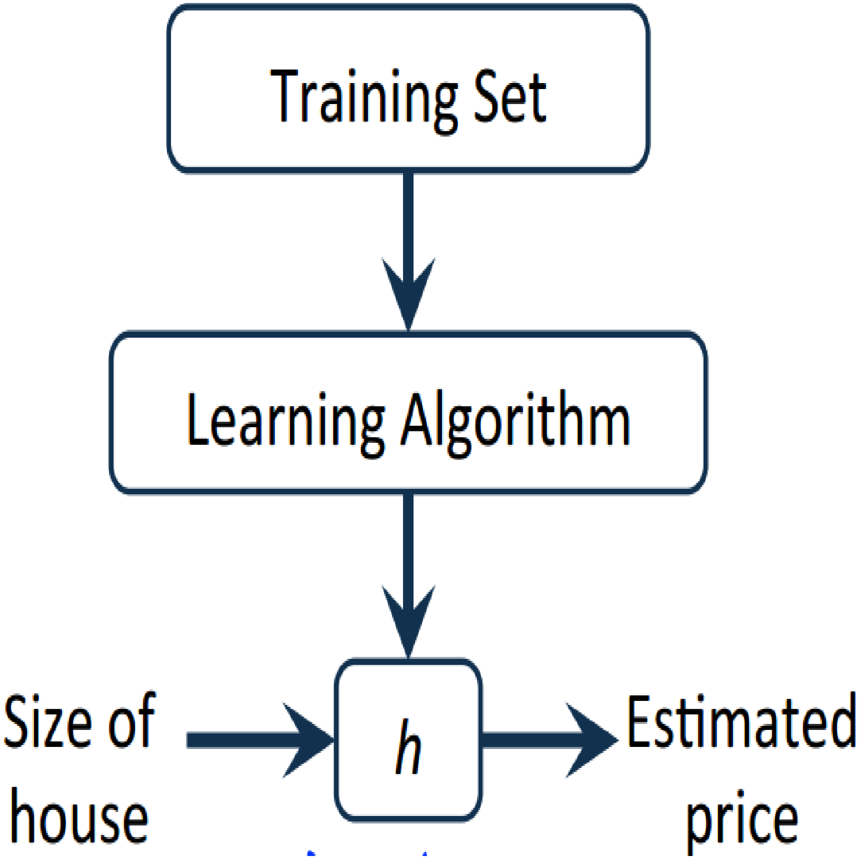
$m = 47$

Simple Linear Regression

- Housing Prices



- Supervised Learning
  - Given the “right answer” for each example in the data
- Regression Problem
  - Predict real-valued output



Training Set

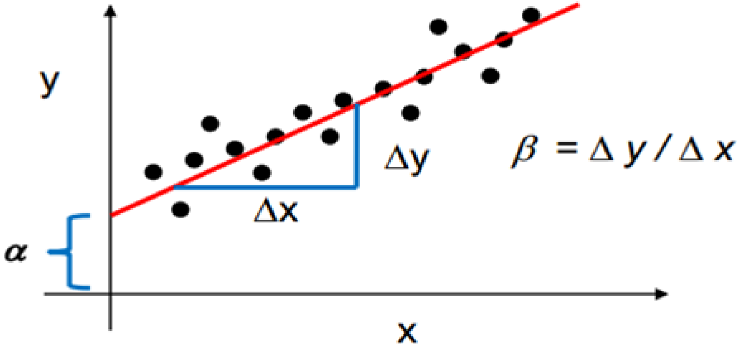
Learning Algorithm

Size of house

h

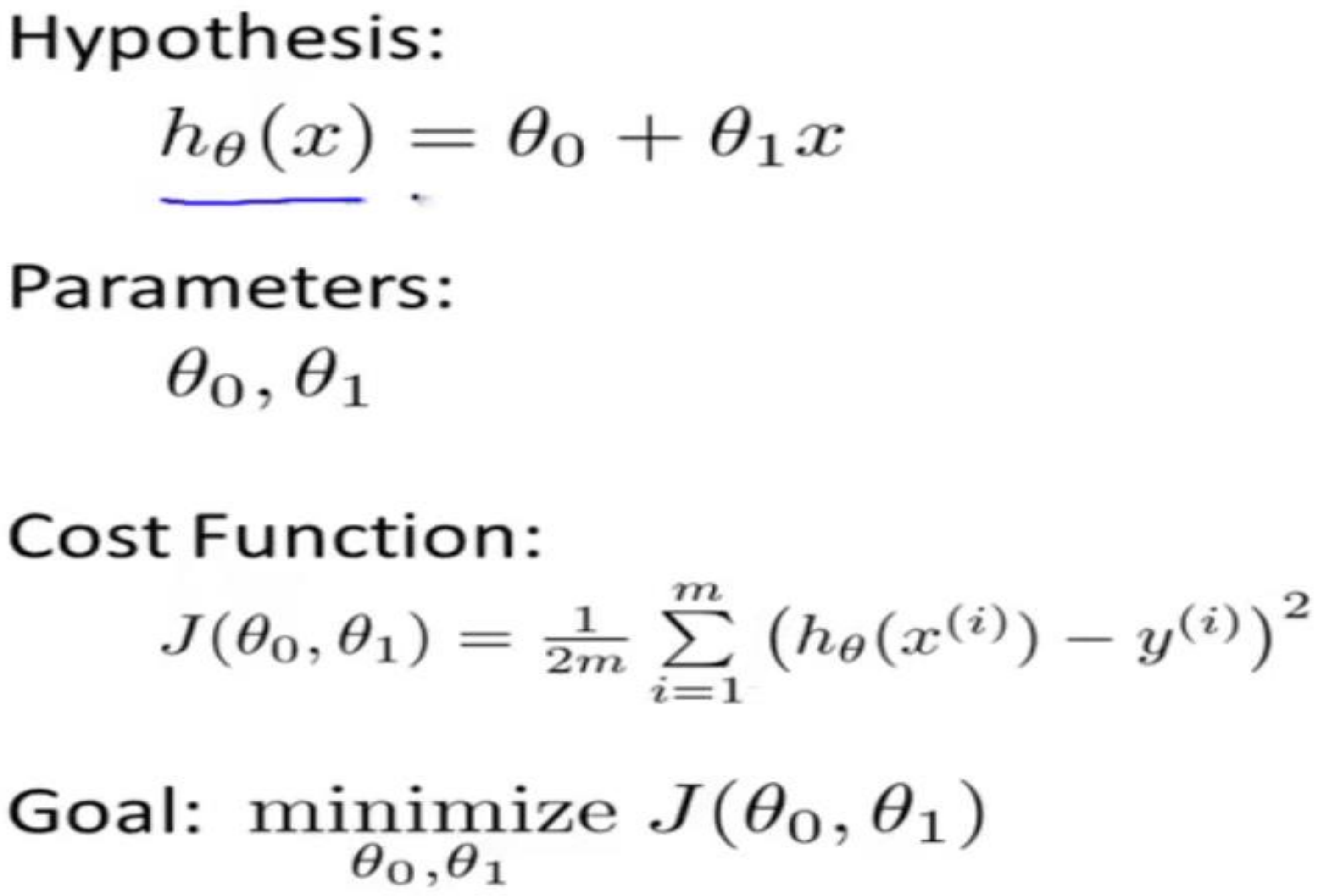
Estimated price

- **How do we represent h?**



- Linear regression with one variable
- Univariate linear regression

Cost function



Hypothesis:

$$\underline{h_\theta(x)} = \theta_0 + \theta_1 x$$

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

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

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

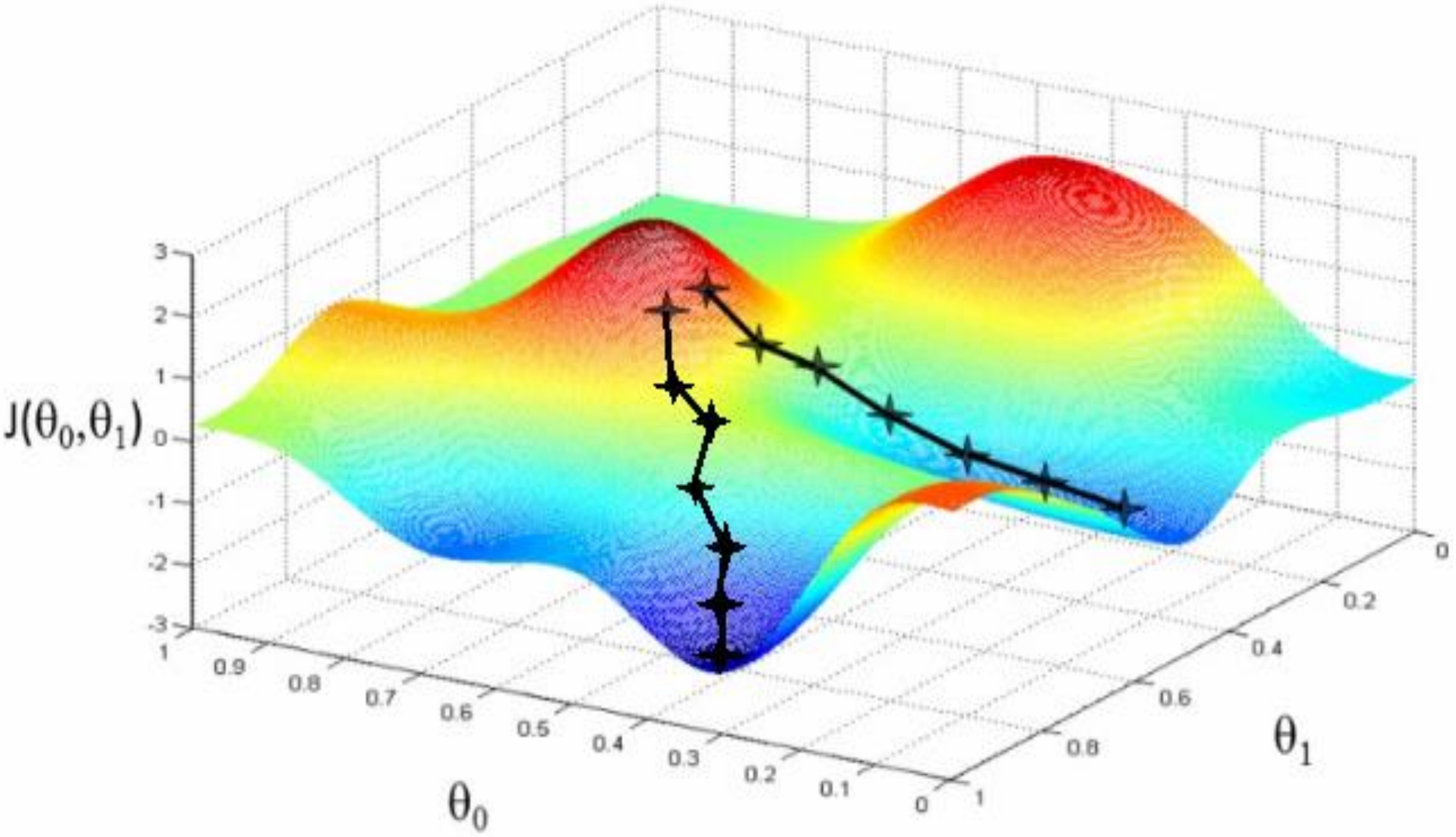
Gradient Descent

- Repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$

}

Gradient Descent



Gradient Descent for Linear Regression

Gradient descent algorithm

$$\text{repeat until convergence } \{$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

$$(\text{for } j = 1 \text{ and } j = 0)$$

$$\}$$

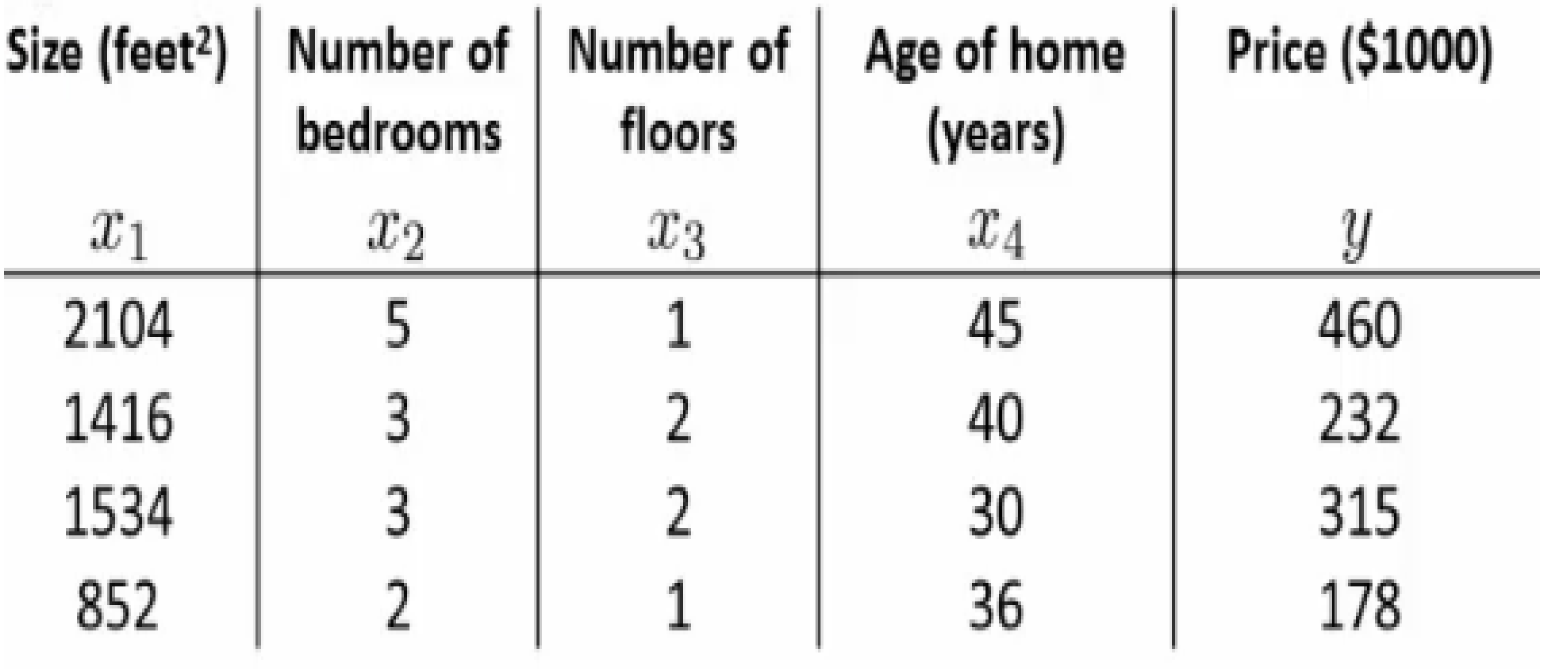
Linear Regression Model

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

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

Multiple Linear Regression

- Linear Regression with Multiple variables



Size (feet^2)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
x_1	x_2	x_3	x_4	y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178

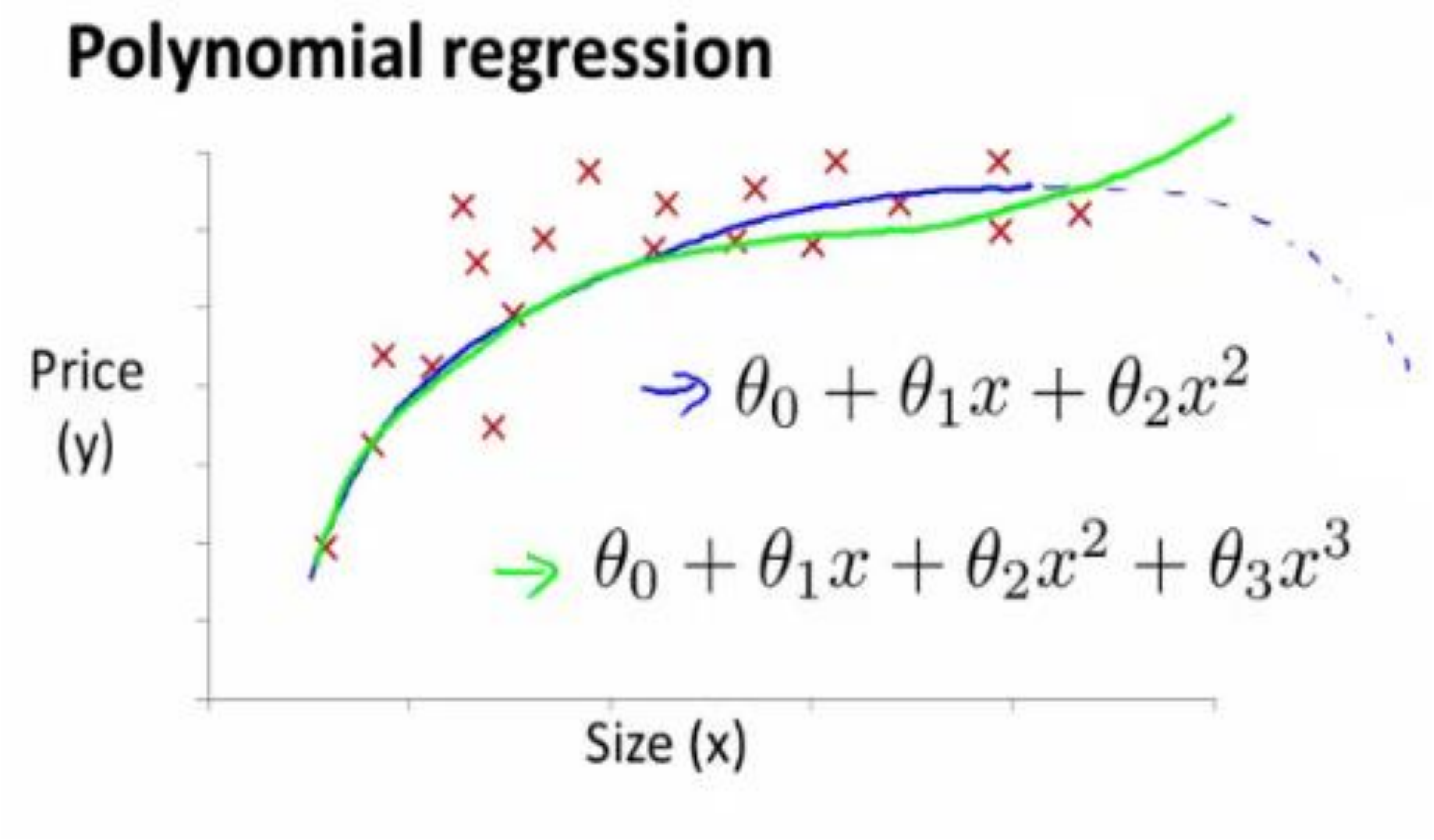
$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Polynomial Regression

- Housing Price Prediction

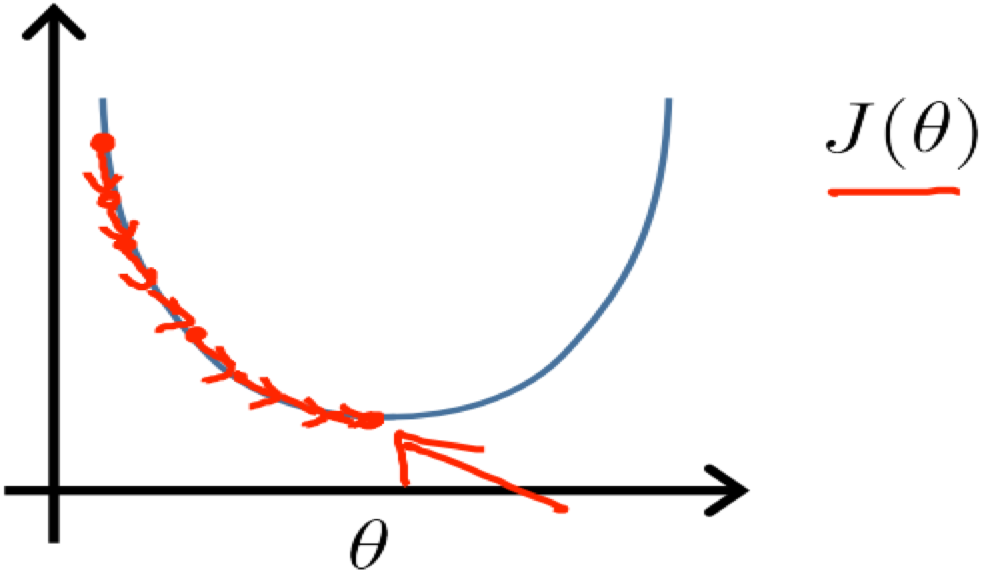
$$h_\theta(x) = \theta_0 + \theta_1 \times \underbrace{\boxed{frontage}}_{x_1} + \theta_2 \times \underbrace{\boxed{depth}}_{x_2}$$

Polynomial Regression



Normal Equation

- Gradient Descent



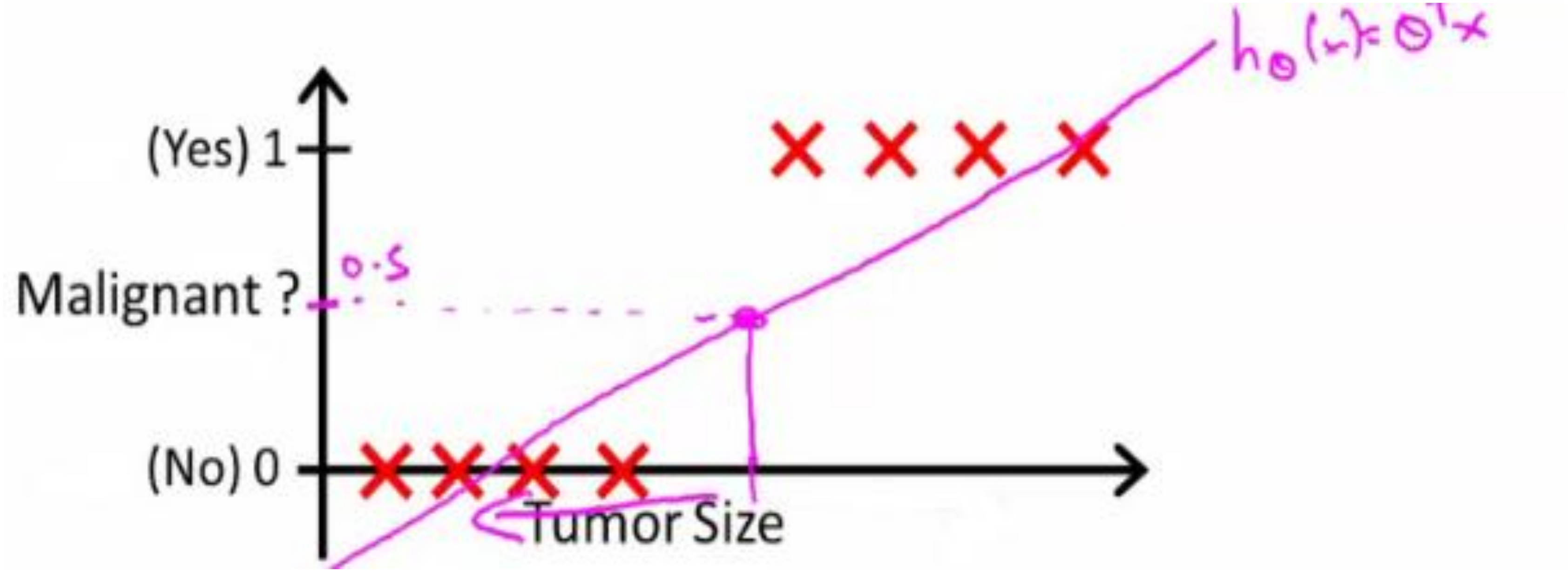
- Normal equation
  - Method to solve for θ analytically

Normal Equation

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

- pinv(X’ * X) *X’ * y

Logistic Regression



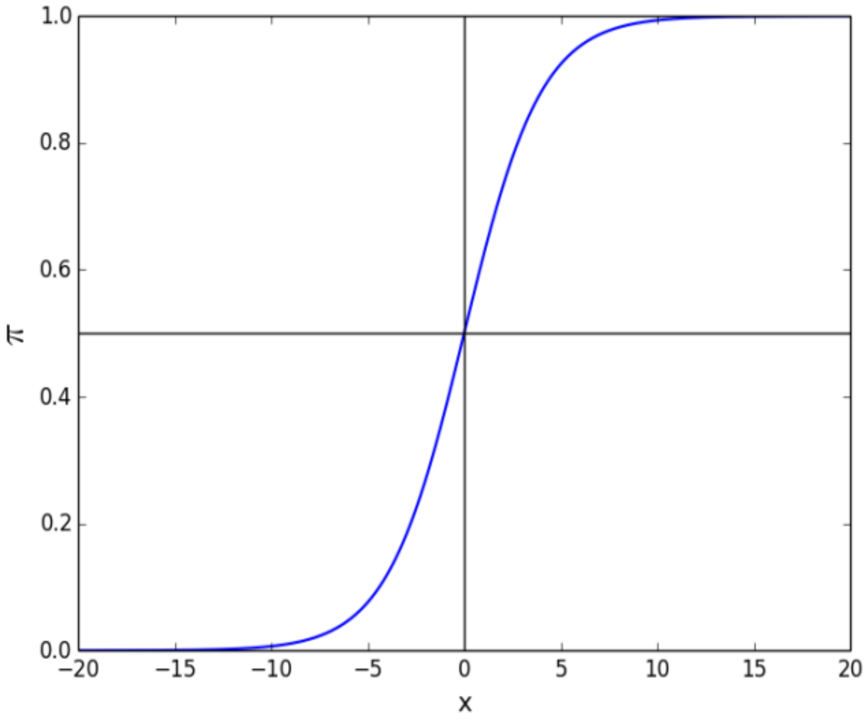
Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict “y = 1”

If $h_\theta(x) < 0.5$, predict “y = 0”

Logistic Regression Model

- Sigmoid Function
- Logistic Function



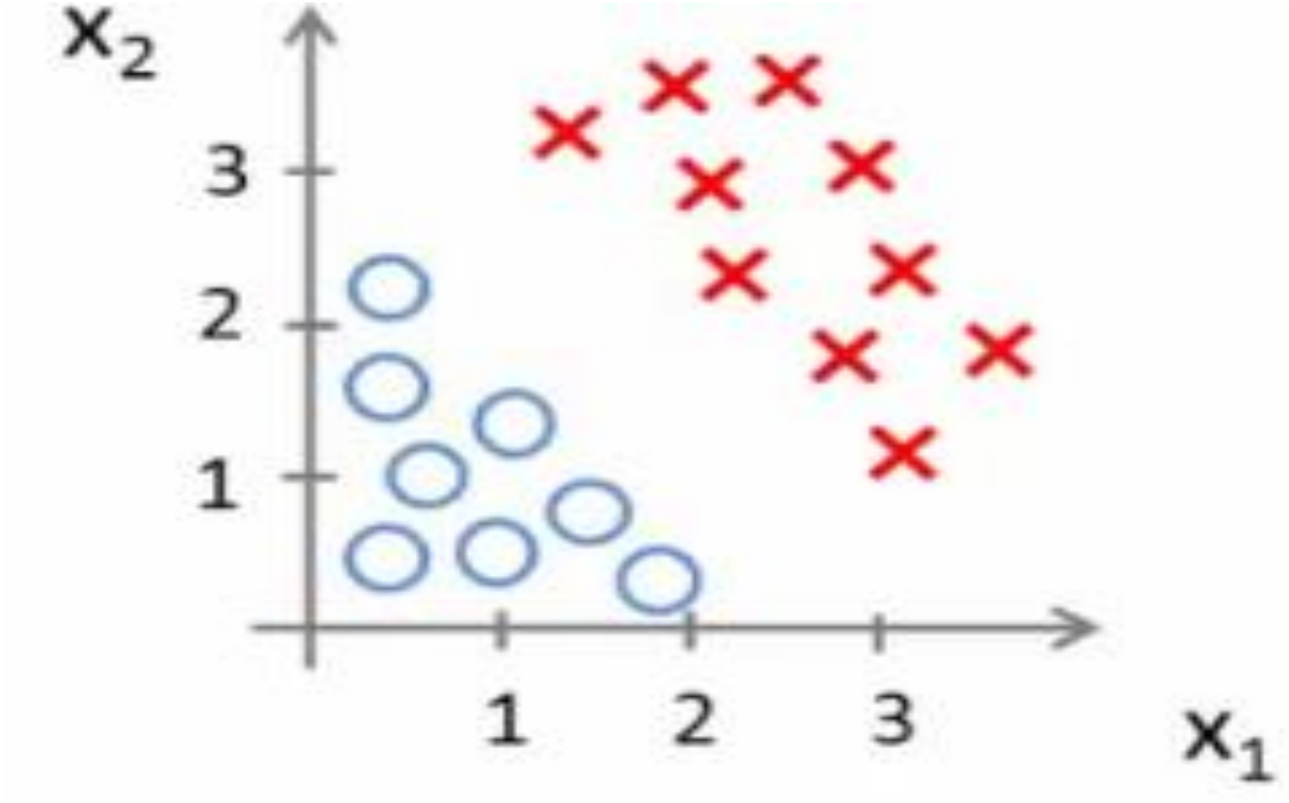
$$h_\theta(x) = g(\theta^T x)$$

$$g(z) = \frac{1}{1+e^{-z}}$$

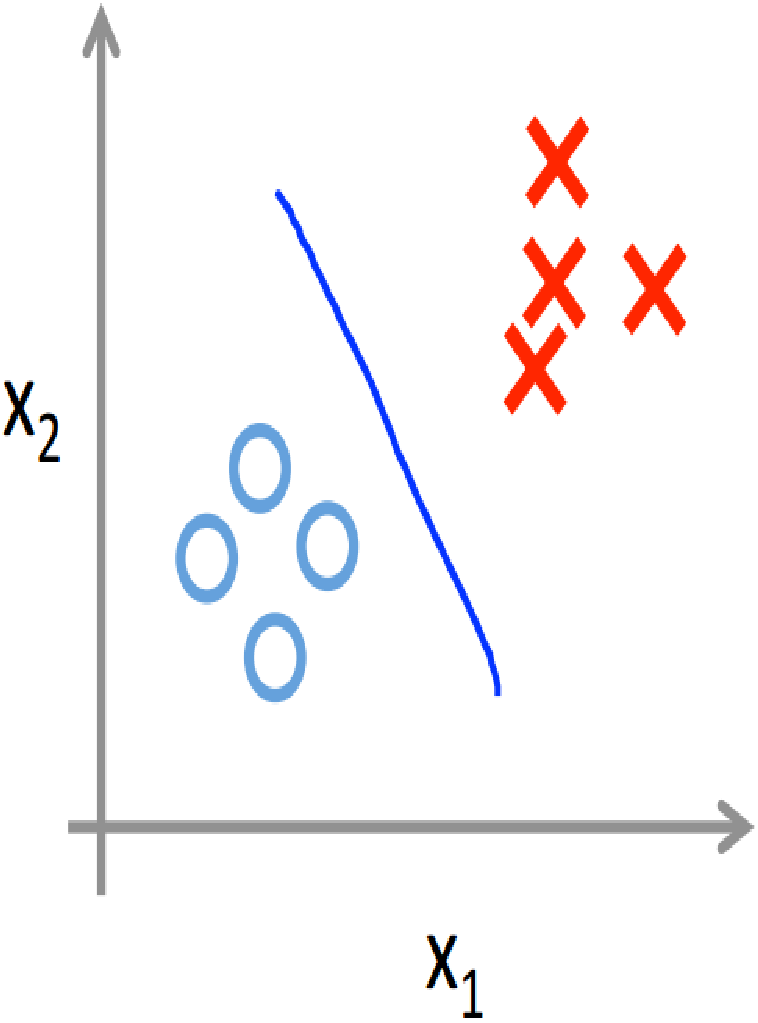
Suppose predict "$y = 1$" if $h_\theta(x) \geq 0.5$

predict "$y = 0$" if $h_\theta(x) < 0.5$

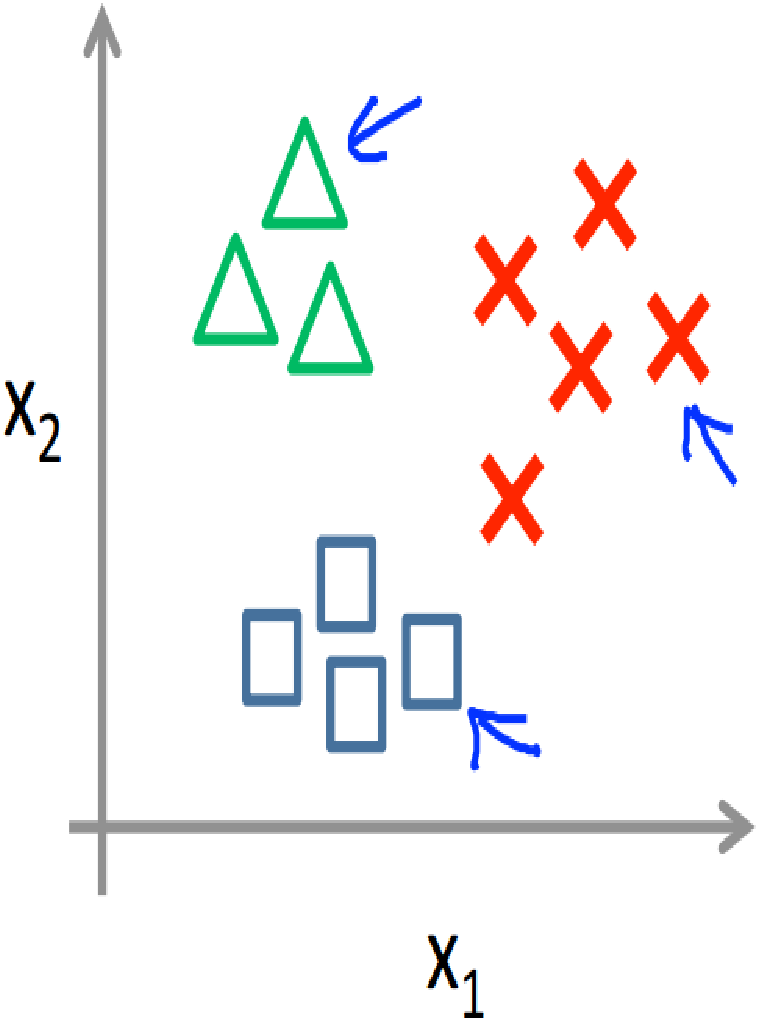
Classification



Binary classification:



Multi-class classification:

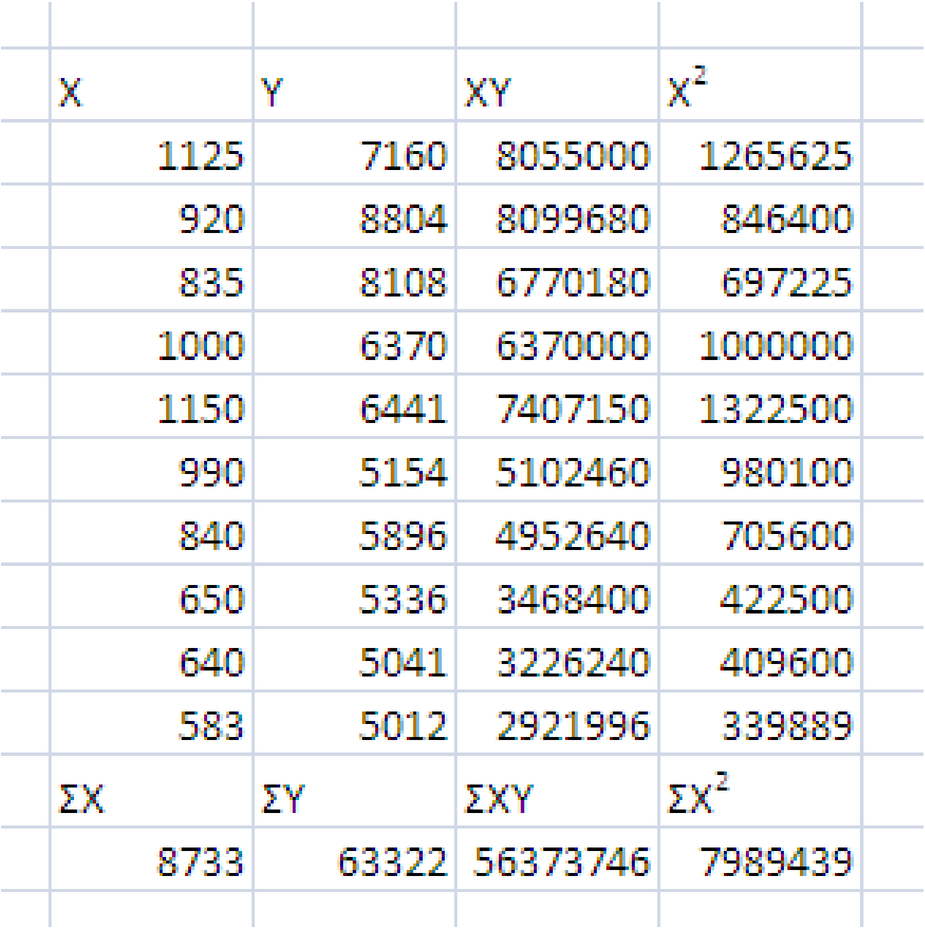


REGRESSION

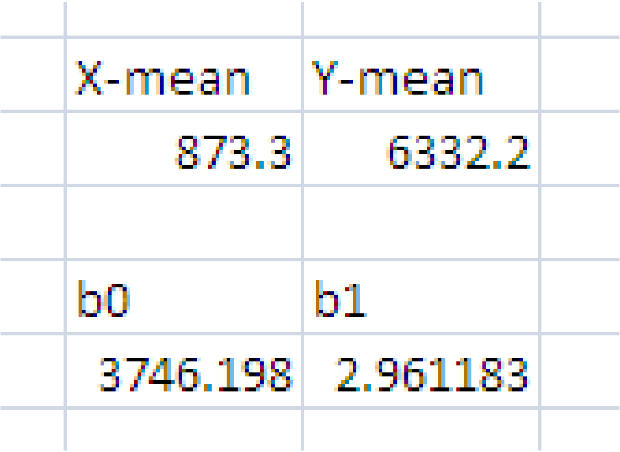
PART-III: Implementation

Predicting House Price

LINEAR REGRESSION EXAMPLE



X	Y	XY	X^2
1125	7160	8055000	1265625
920	8804	8099680	846400
835	8108	6770180	697225
1000	6370	6370000	1000000
1150	6441	7407150	1322500
990	5154	5102460	980100
840	5896	4952640	705600
650	5336	3468400	422500
640	5041	3226240	409600
583	5012	2921996	339889
ΣX	ΣY	ΣXY	$ΣX^2$
8733	63322	56373746	7989439

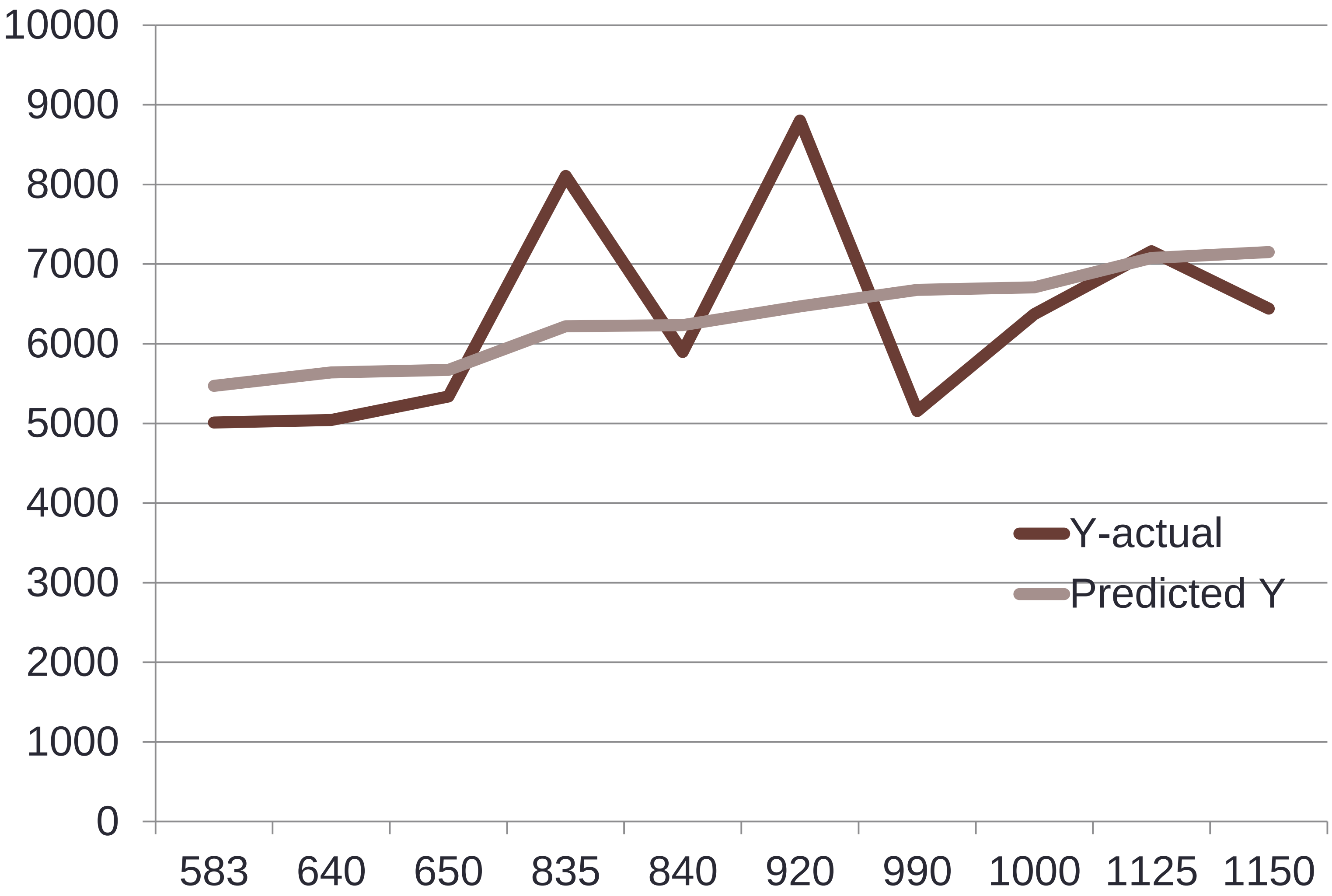


X-mean	Y-mean
873.3	6332.2
b0	b1
3746.198	2.961183

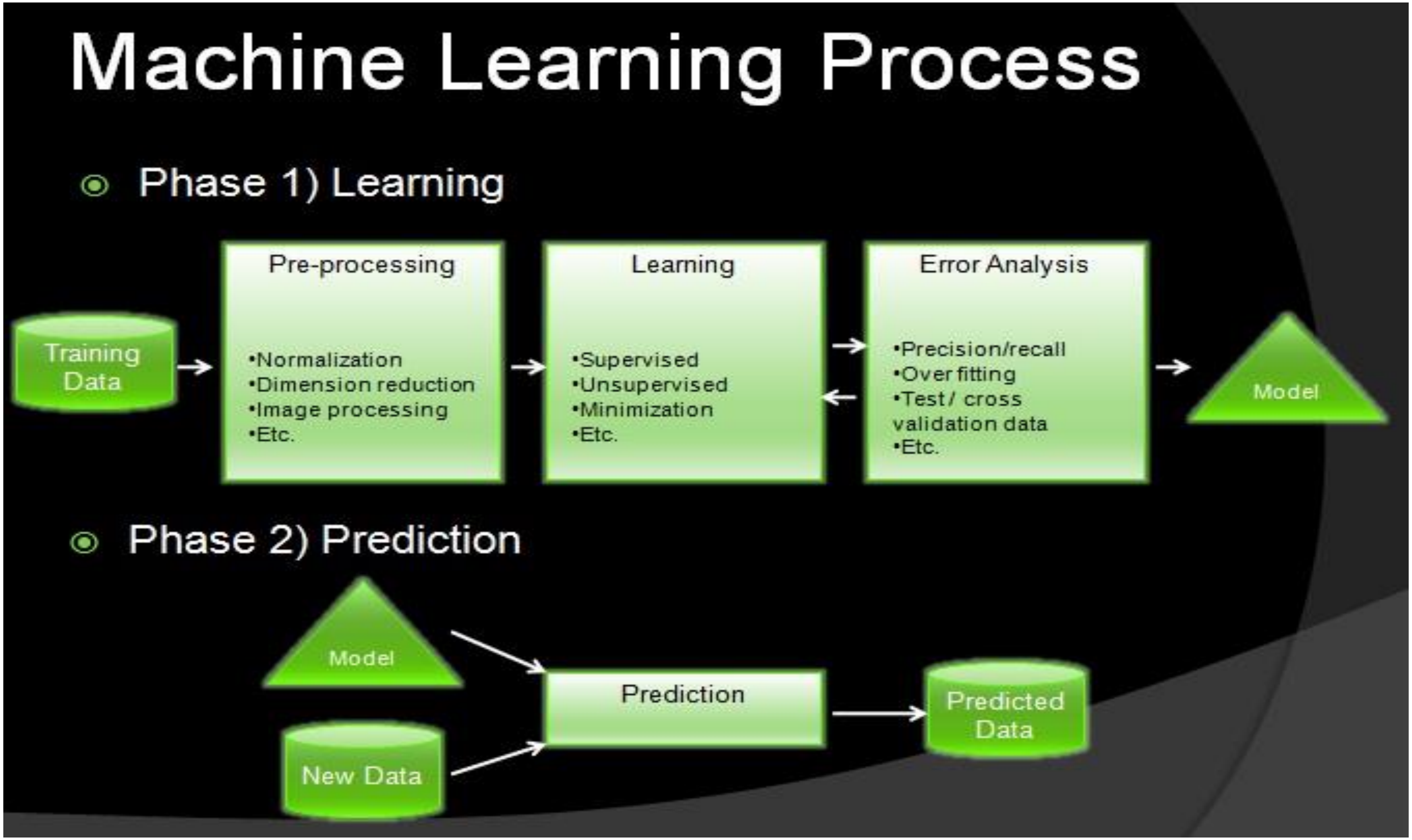
$$\beta_0 = \frac{(\sum X^2)(\sum Y) - (\sum X)(\sum XY)}{n\sum X^2 - (\sum X)^2}$$

$$\beta_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

LINEAR REGRESSION EXAMPLE



Machine Learning



Machine Learning Process

- Phase 1) Learning

Training Data → Pre-processing → Learning ⇄ Error Analysis → Model

Pre-processing
- Normalization
- Dimension reduction
- Image processing
- Etc.

Learning
- Supervised
- Unsupervised
- Minimization
- Etc.

Error Analysis
- Precision/recall
- Over fitting
- Test / cross validation data
- Etc.

- Phase 2) Prediction

Model → Prediction

New Data → Prediction

Prediction → Predicted Data

Machine Learning

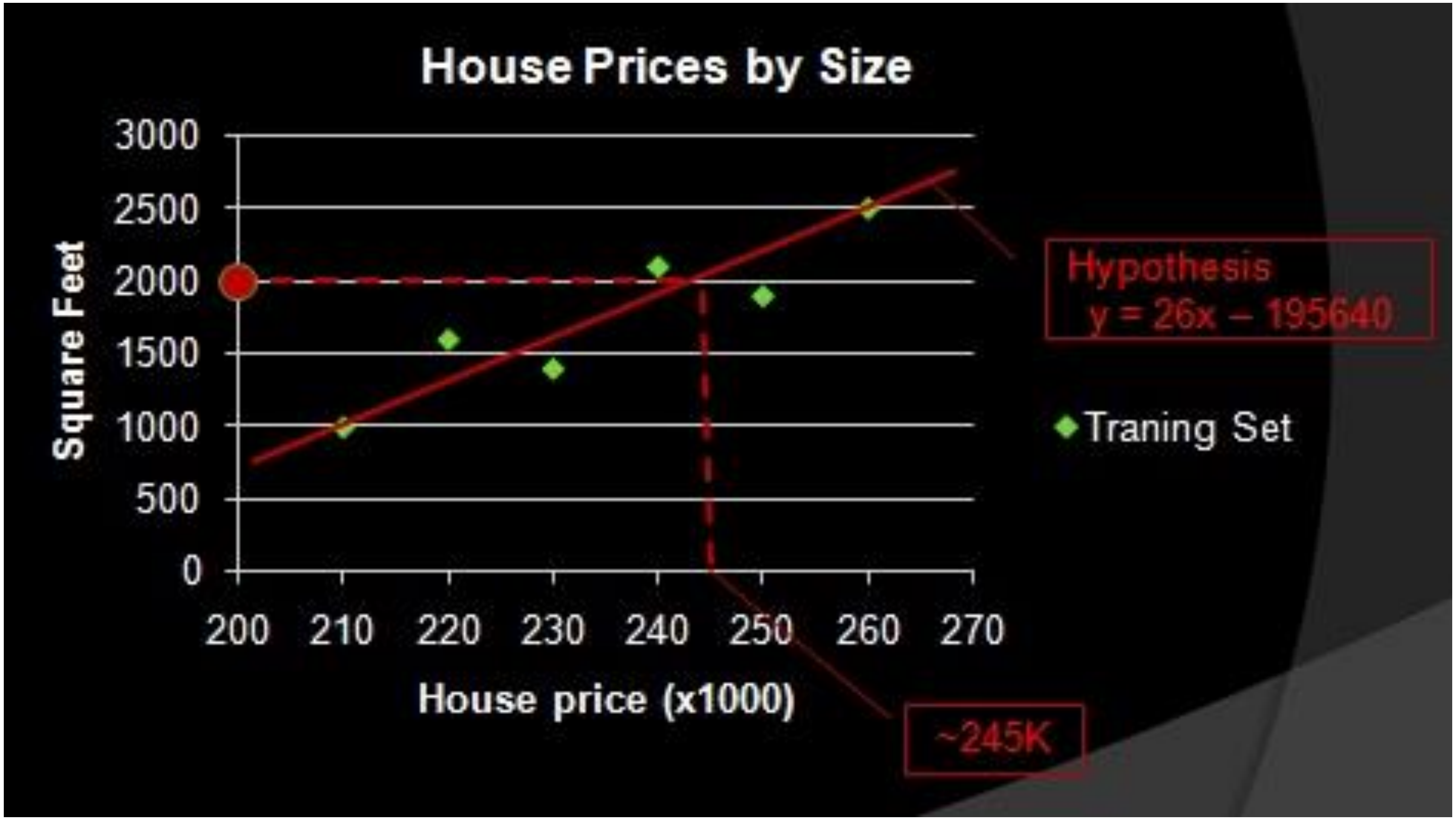
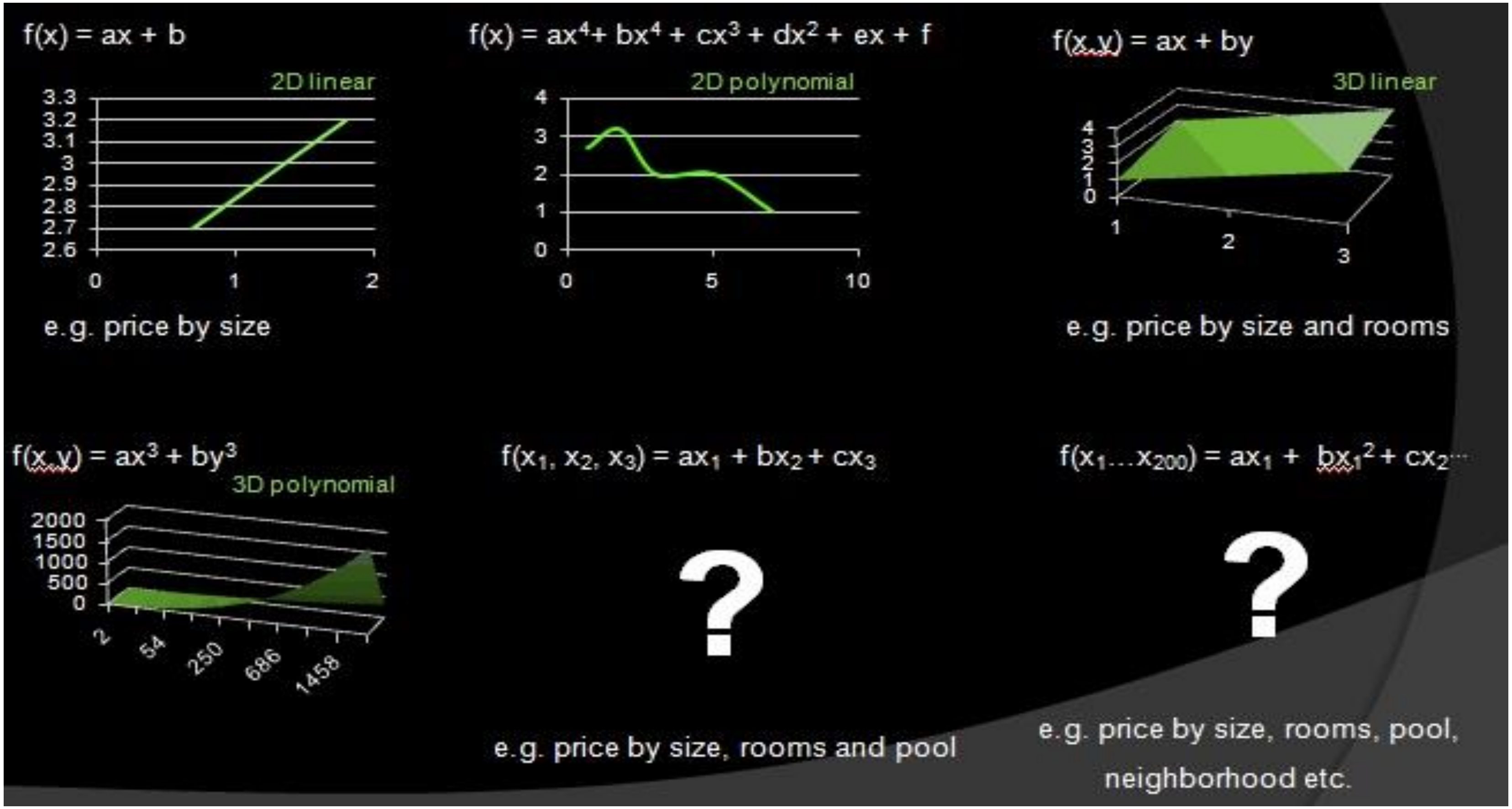


Figure shows how one feature can be used on a linear regression problem to predict new house prices

Machine Learning



On Figure how the complexity can grow easily from 2 dimensions linear to hundreds of dimensions polynomial

Gradient Descent

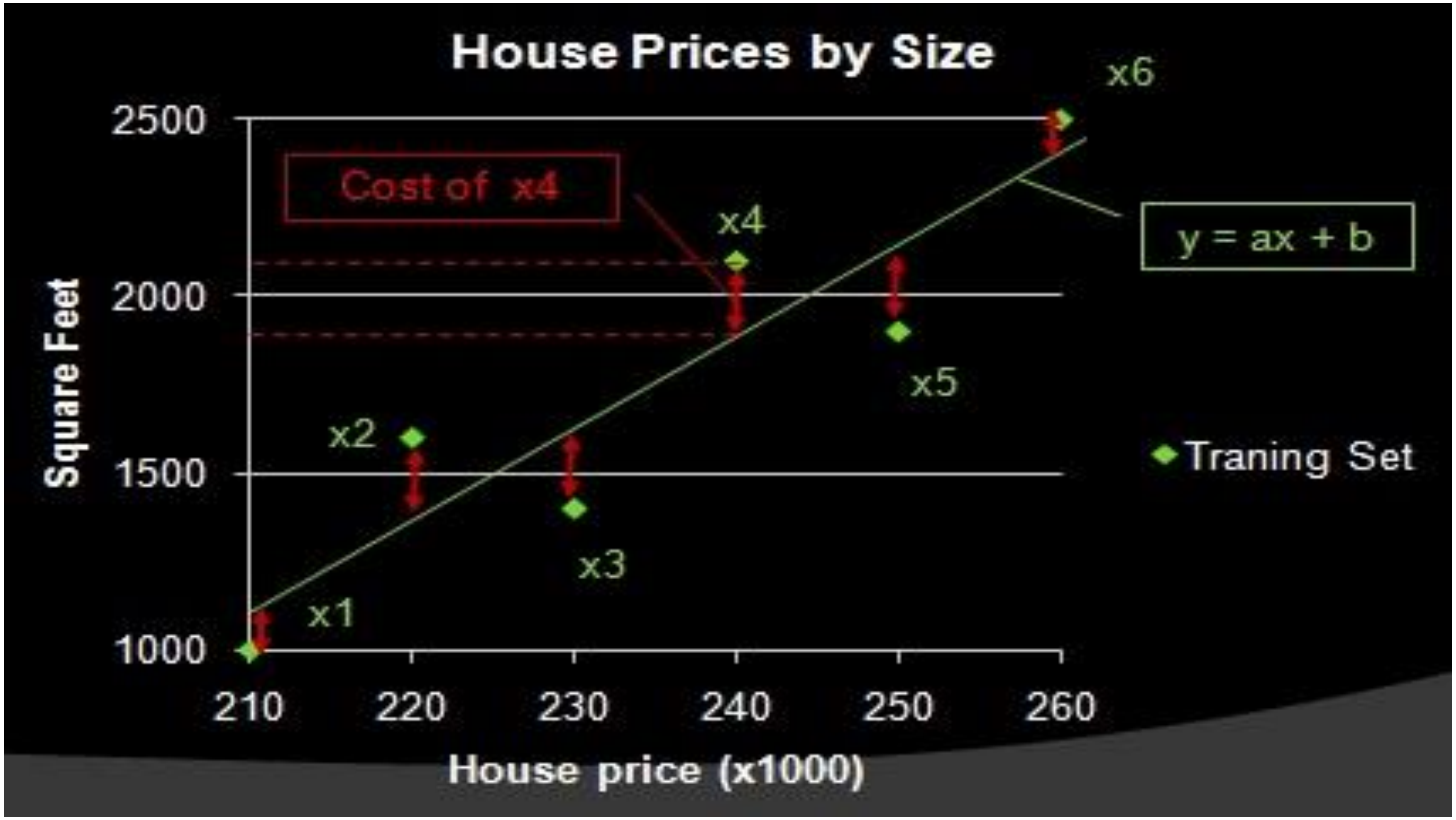
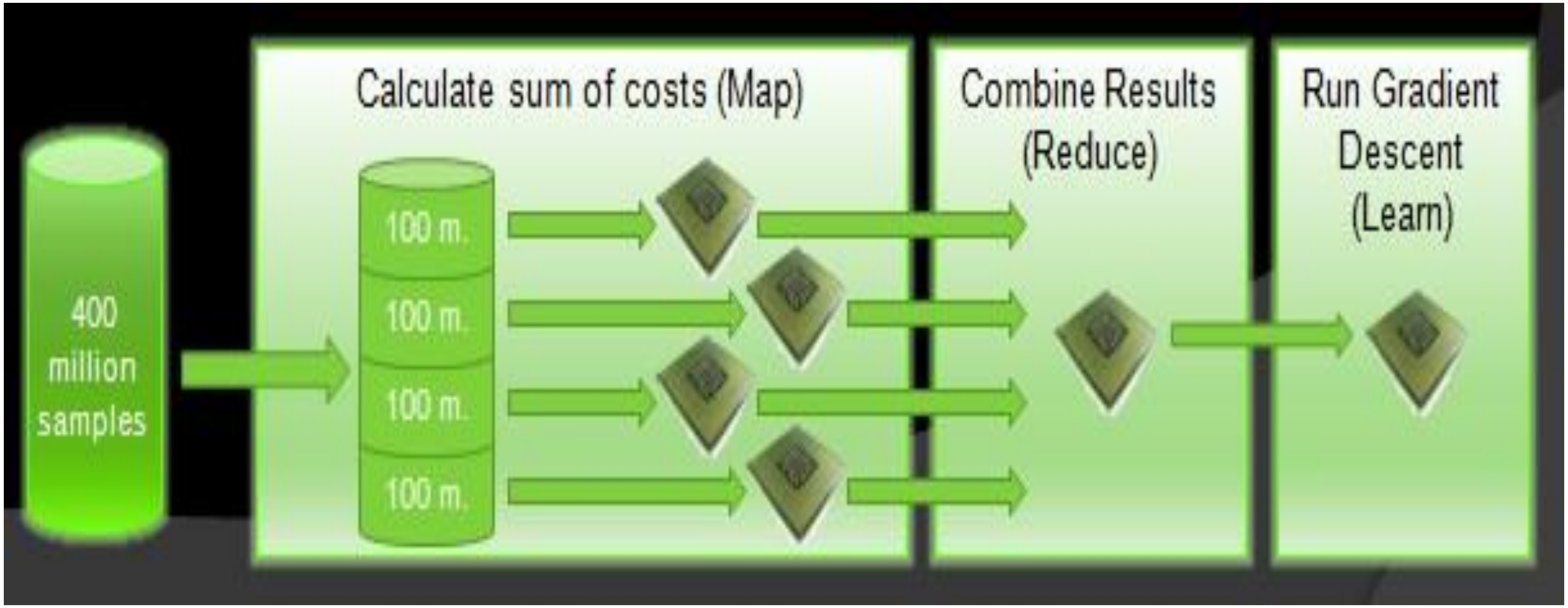


Figure shows how a training data is plot and the error is calculated

Large Scale Machine Learning



In Figure we can see how the data is split into four parts and fed into four different processors

Large Scale Machine Learning

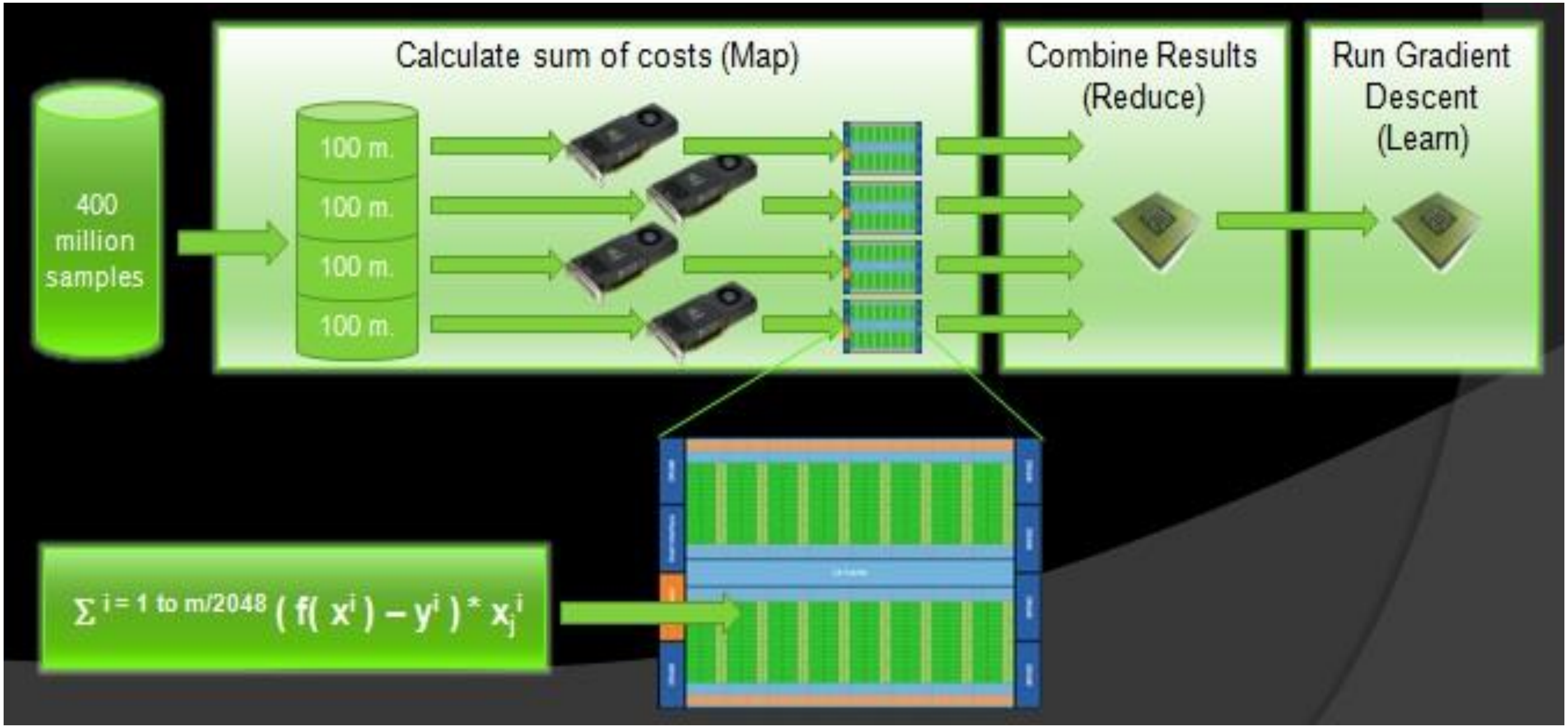
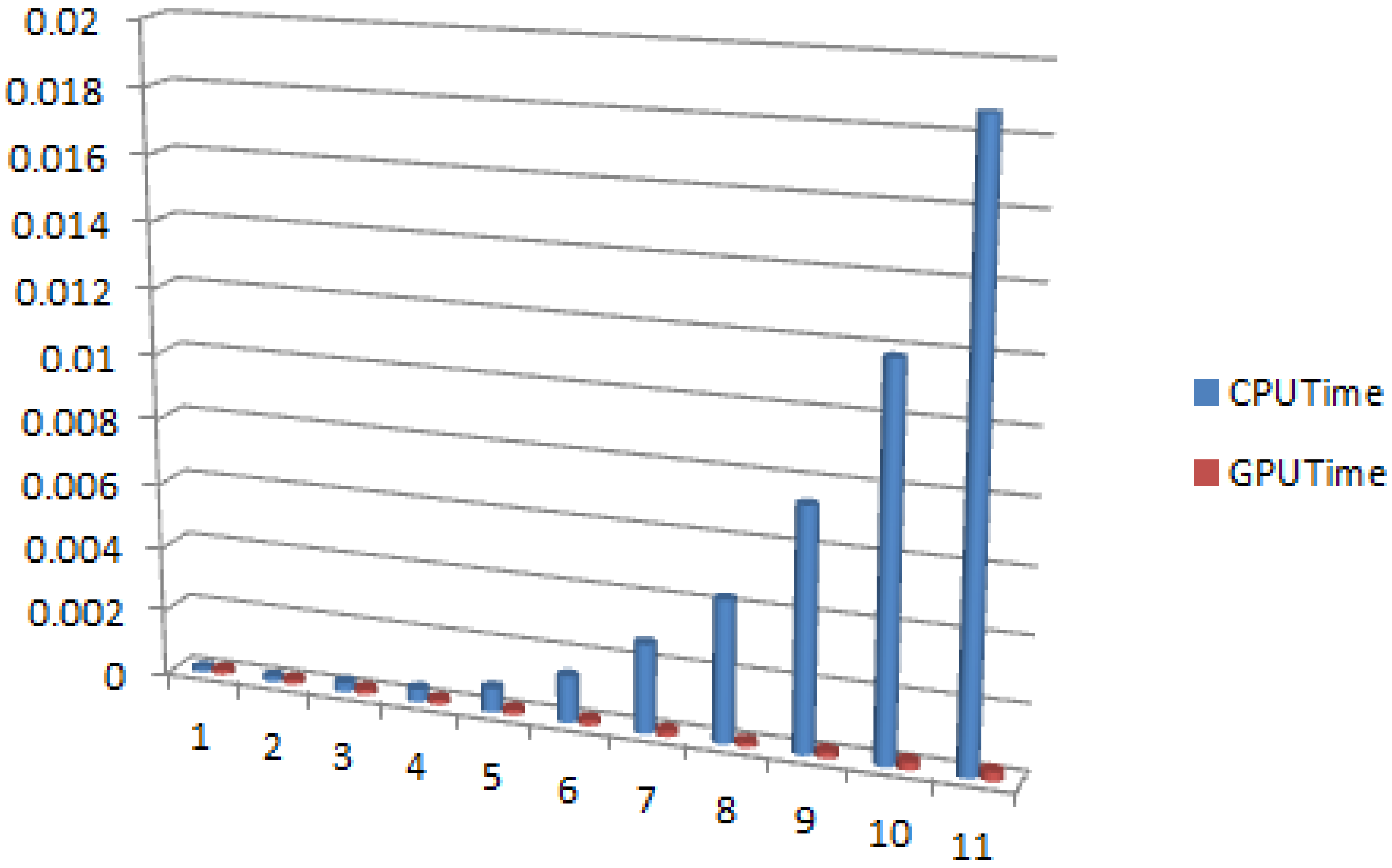

400 million samples
Calculate sum of costs (Map)
100 m.
100 m.
100 m.
100 m.
Combine Results (Reduce)
Run Gradient Descent (Learn)
$\Sigma^{i=1 \text{ to } m/2048} (f(x^i) – y^i) * x_j^i$

Figure shows this configuration along with the parallelized part on the GPU cores

GPGPU

- Utilizing GPUs to enable dramatic increases in computing performance of general purpose scientific and engineering computing is named GPGPU.
- NVIDIA is providing a parallel computing platform and programming model named CUDA to develop GPGPU software on C, C++ or Fortran which can run on any NVIDIA GPU.

CPU vs GPU Time



Conclusion

- **GPGPU**, **Machine Learning** and **Big Data** are three rising fields in the IT industry.
- As much as we get deeper into these fields ,we figure out how well they fit together.
- I hope this sample application gave you some basic idea and maybe just one perspective how you can use **NVIDIA CUDA** easily on machine learning problems.

References

- http://gpgpu.org/
- https://en.wikipedia.org/wiki/Linear_regression
- https://docs.nvidia.com/cuda/
- https://en.wikipedia.org/wiki/Polynomial_regression
- https://en.wikipedia.org/wiki/Gradient_descent
- https://en.wikipedia.org/wiki/Overfitting
- https://en.wikipedia.org/wiki/Regularization_(mathematics)
- http://www.nvidia.com/object/machine-learning.html
- https://developer.nvidia.com/nvidia-nsight-visual-studio-edition
- https://docs.nvidia.com/cuda/thrust/index.html
- https://adnanboz.wordpress.com/2012/02/25/large-scale-machine-learning-using-nvidia-cuda/

References

- http://statsmodels.sourceforge.net/
- http://www.nature.com/ejhg/journal/v17/n8/full/ejhg2009 5a.html
- http://www.alsharif.info/#!iom530/c21o7