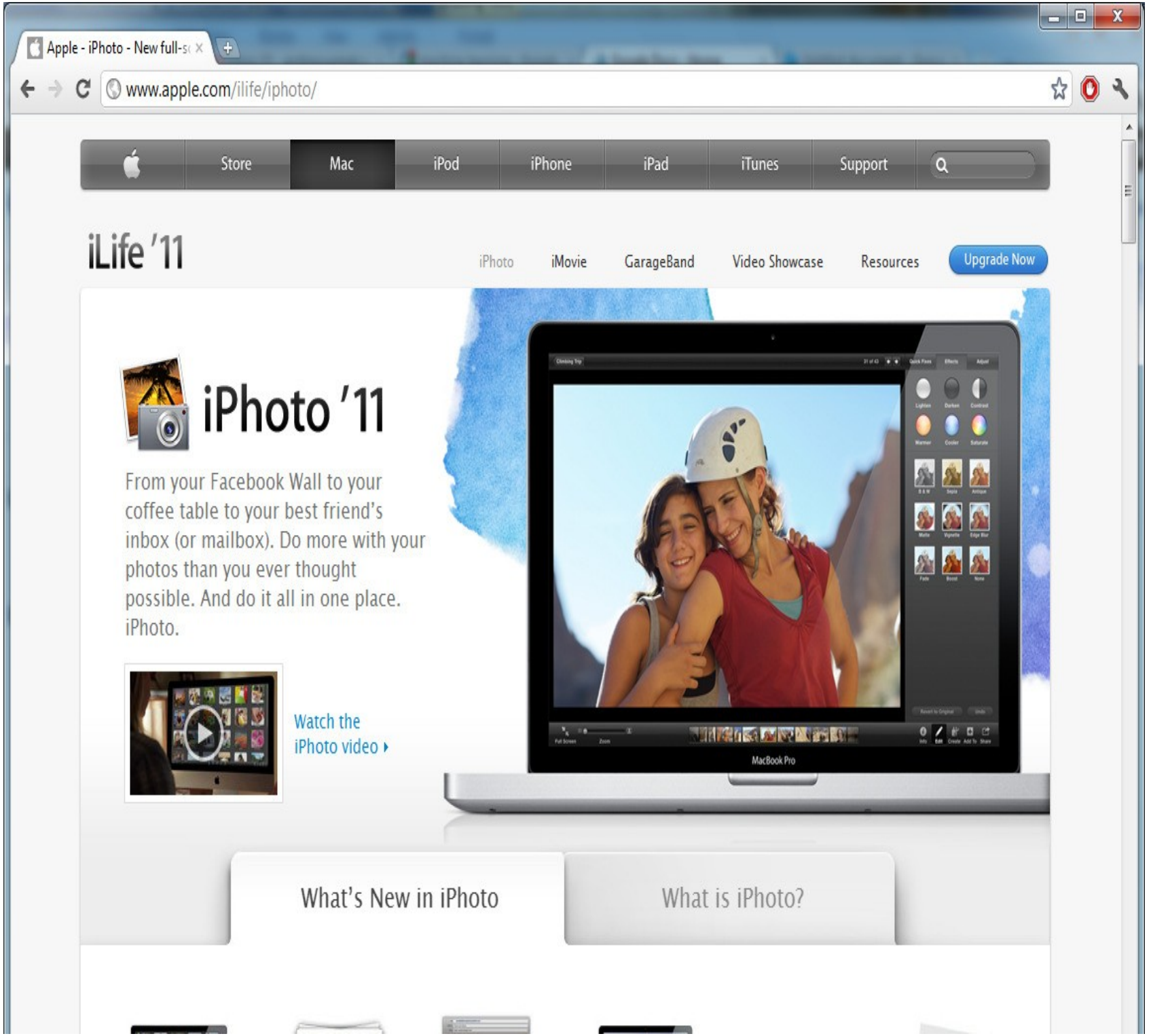
What is AI/MachineLearning?

Introduction


Apple - iPhoto - New full-s
www.apple.com/ilife/iphoto/
Store
Mac
iPod
iPhone
iPad
iTunes
Support
iLife '11
iPhoto
iMovie
GarageBand
Video Showcase
Resources
Upgrade Now
iPhoto '11
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.
Watch the iPhoto video
MacBook Pro
What's New in iPhoto
What is iPhoto?


SPAM

Machine Learning

- Grew out of work in AI
- New capability for computers

Examples:

- Database mining

  Large datasets from growth of automation/web.

  E.g., Web click data, medical records, biology, engineering
- Applications can’t program by hand.

  E.g., Autonomous helicopter, handwriting recognition, mos

  Natural Language Processing (NLP), Computer Vision.

Machine Learning

- Grew out of work in AI

- New capability for computers

Examples:

- Database mining

  Large datasets from growth of automation/web.

  E.g., Web click data, medical records, biology, engineering

- Applications can’t program by hand.
  E.g., Autonomous helicopter, handwriting recognition, most

  Natural Language Processing (NLP), Computer Vision.

- Self-customizing programs

  E.g., Amazon, Netflix product recommendations

Machine Learning

- Grew out of work in AI
- New capability for computers

Examples:

- Database mining

  Large datasets from growth of automation/web.

  E.g., Web click data, medical records, biology, engineering

- Applications can’t program by hand.

  E.g., Autonomous helicopter, handwriting recognition, most c

  Natural Language Processing (NLP), Computer Vision.

- Self-customizing programs

  E.g., Amazon, Netflix product recommendations

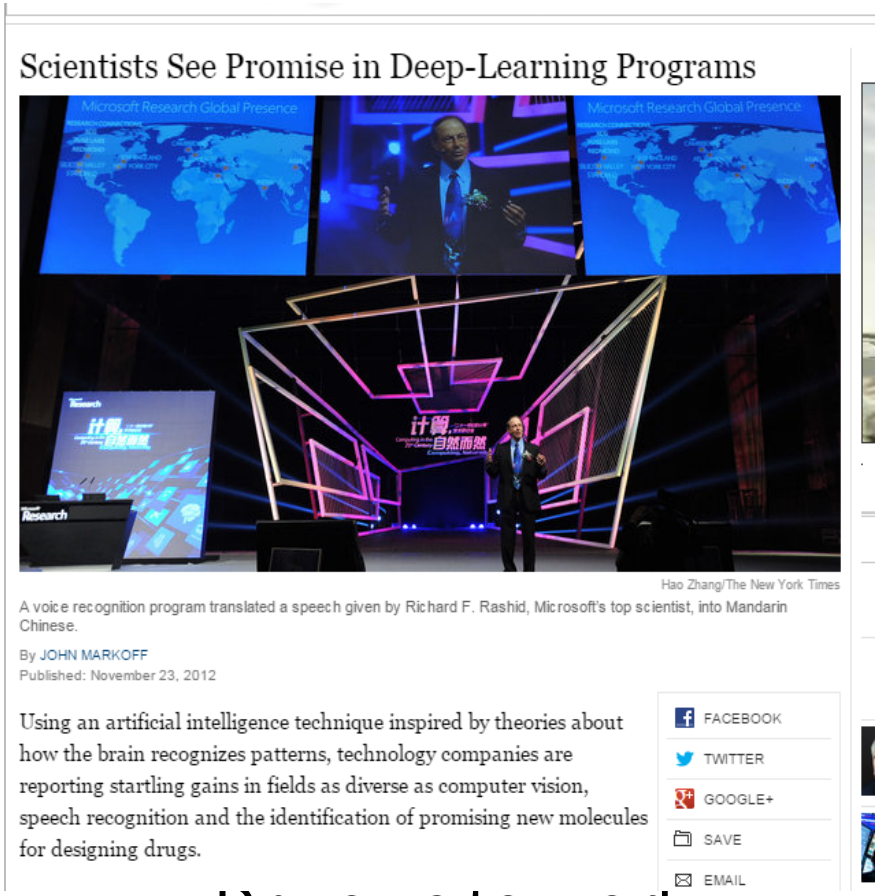
- Understanding human learning (brain, real AI).

What do we want AI to do?

Keep us organized

Guide us to content

Help us find things



Scientists See Promise in Deep-Learning Programs

Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

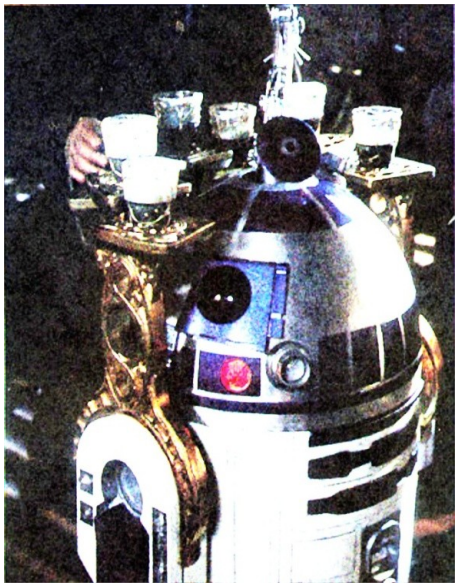
By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

FACEBOOK
TWITTER
GOOGLE+
SAVE
EMAIL

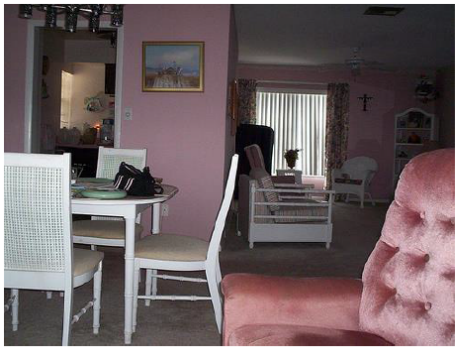
Help us communicate
帮助我们沟通

Drive us to work



Serve drinks?

Two dogs and one person.
A person is playing with a dog.



Six chairs and one table and one person.
The image is taken in a room.

A living room with a carpeting, a cream sofa and chair, and a large door.

The room in the apartment gets some afternoon sun.

Living room with white couch and blue carpeting.

Five bottles and one desk and one person.
A person is sitting on a chair

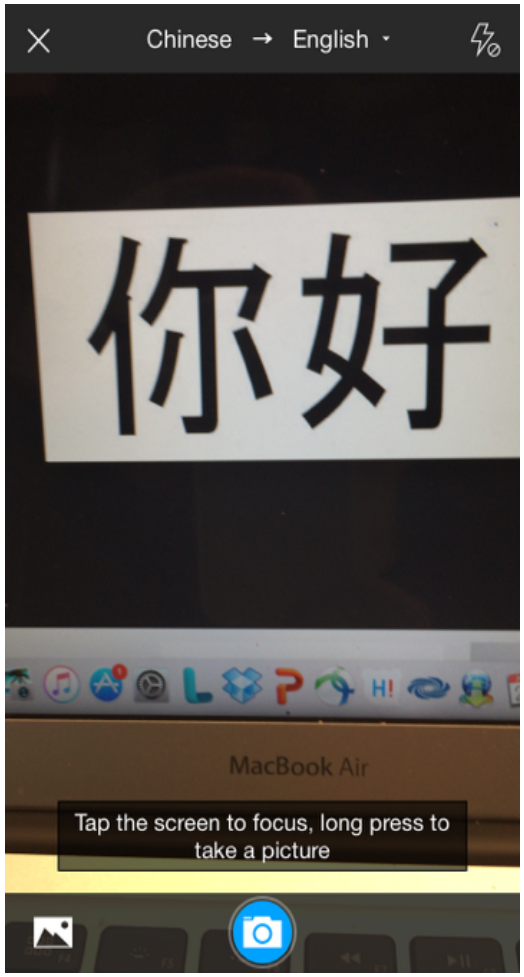
One motorbike and one person.
A person is riding a motorbike.

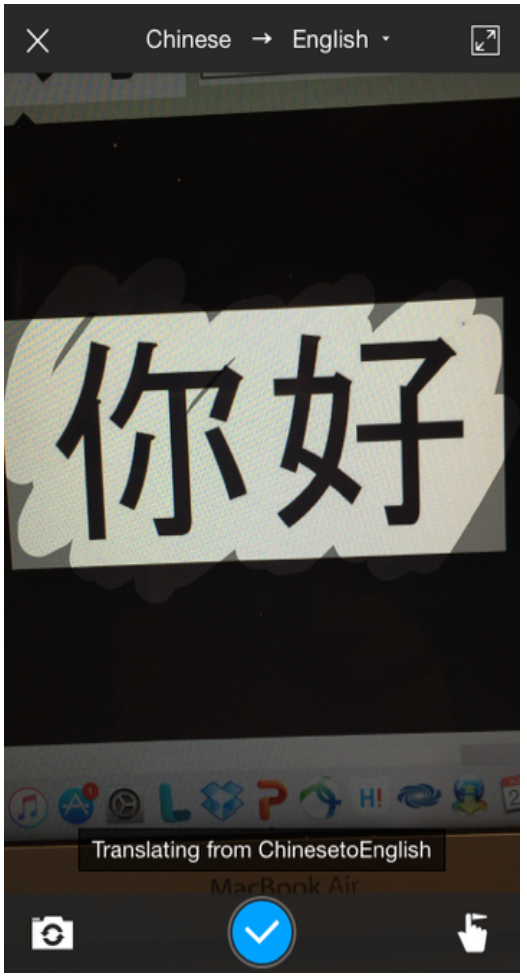
A yellow bus driving down a road with green trees and grass in the background.

School bus on a street.

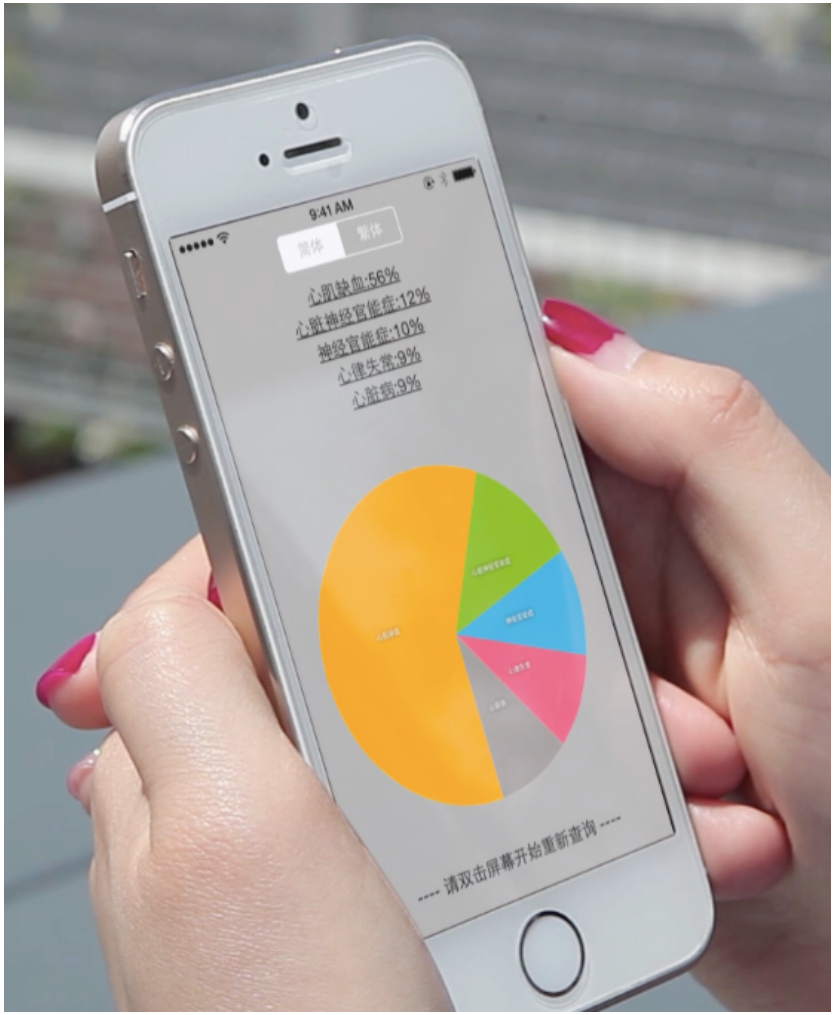
Yellow bus with trees in the background.

OCR-based Translation App



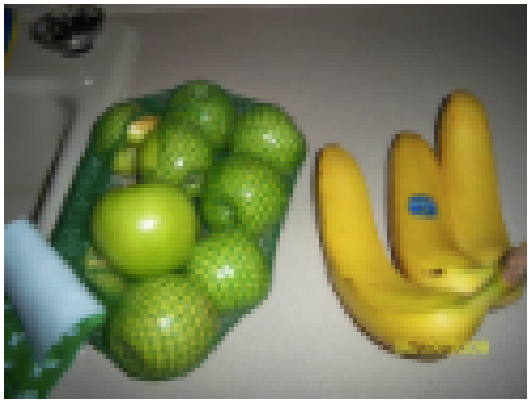
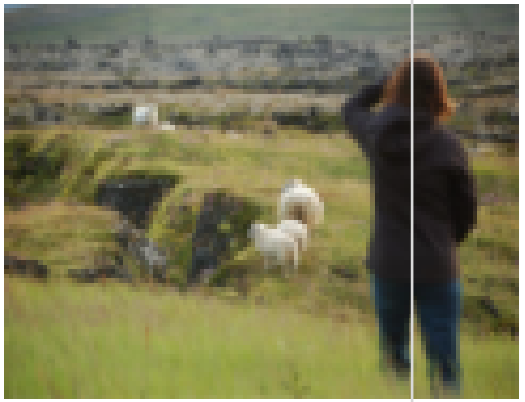


Medical Diagnostics App

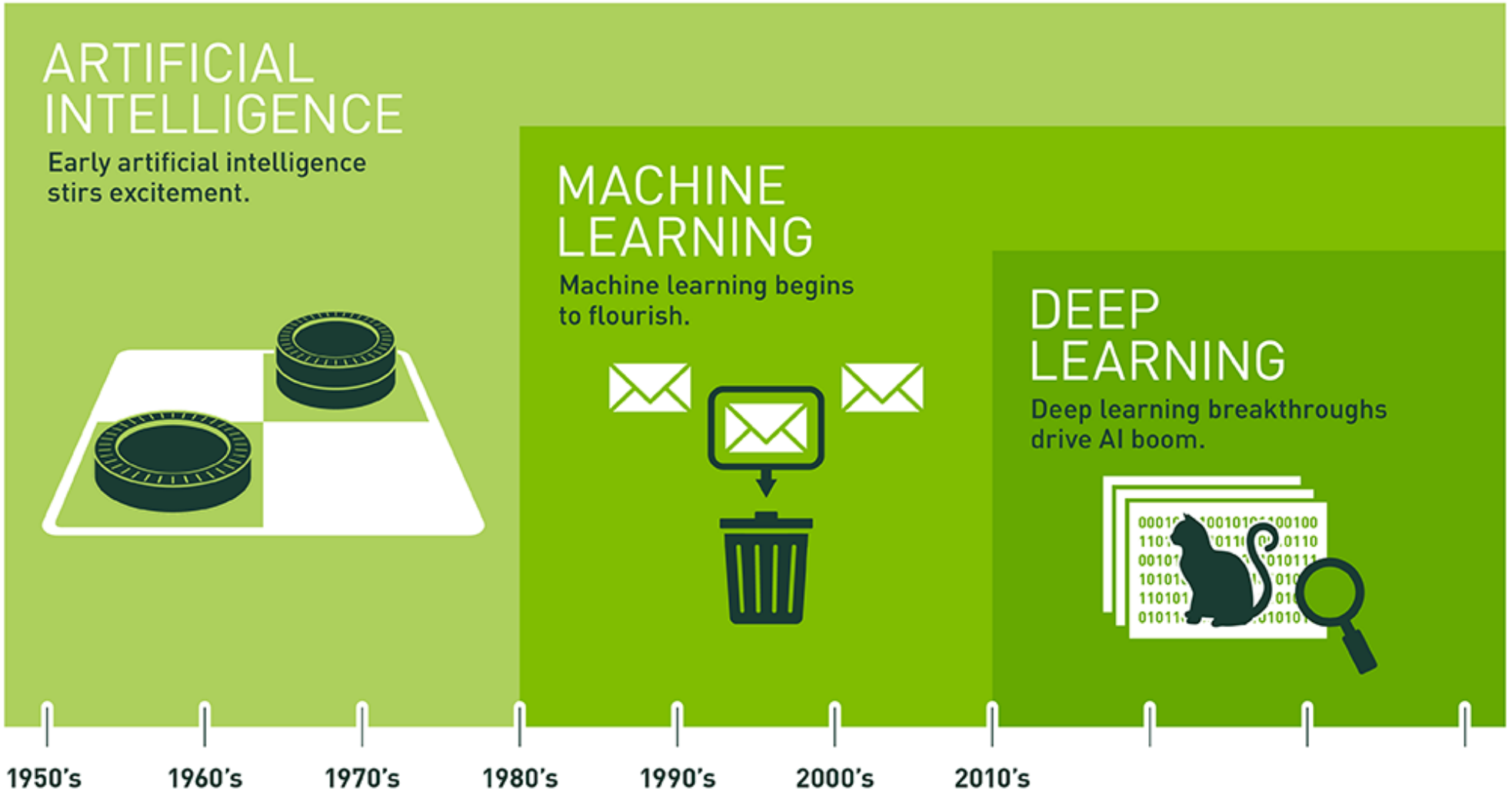


AskADoctor can assess 520 different diseases, representing ~90 percent of the most common medical problems.

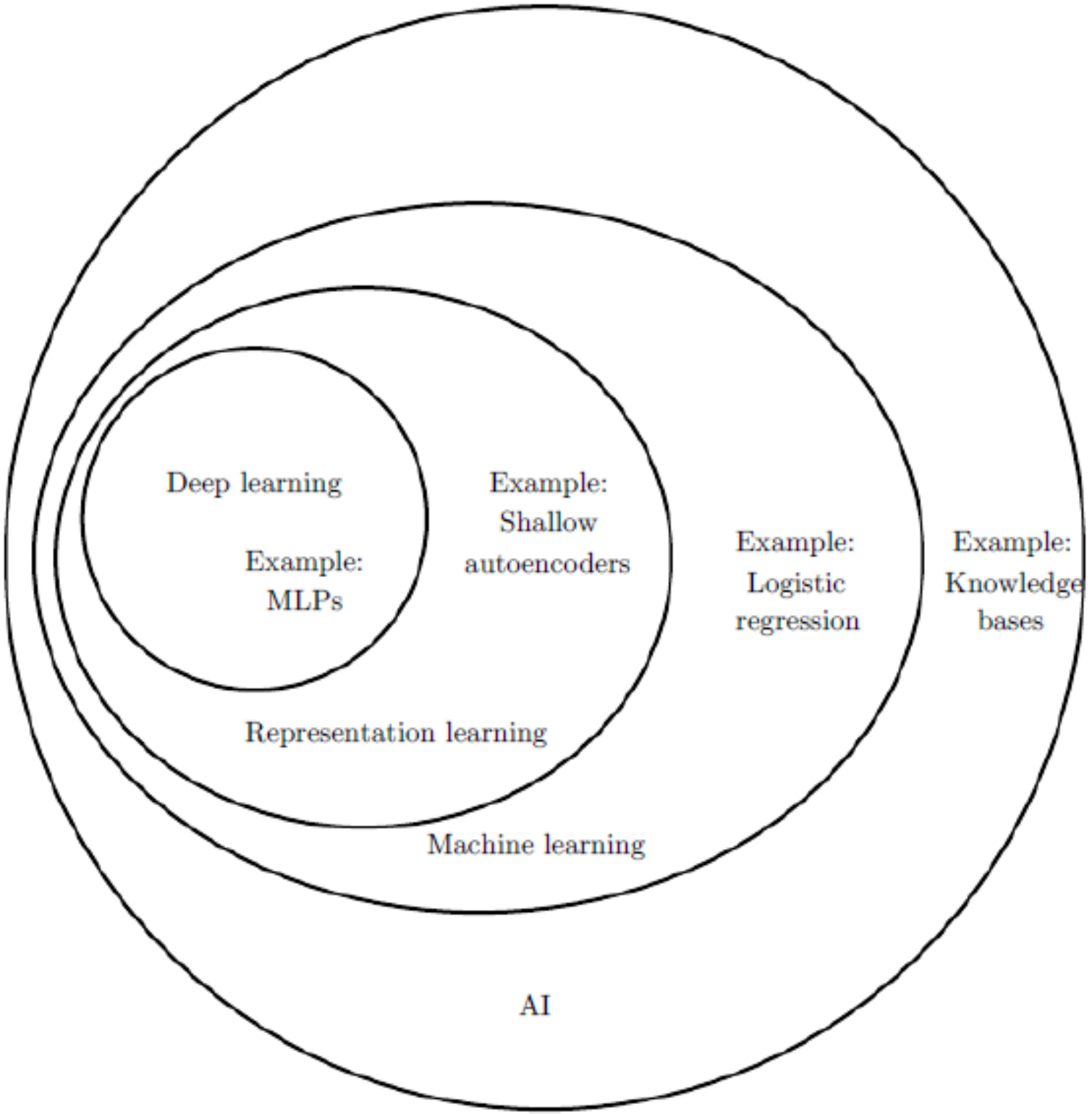
Image Q&A

Image			
Question	公共汽车是什么颜色的？ What is the color of the bus?	黄色的是什么？ What is there in yellow?	草地上除了人以外还有什么动 What is there on the grass, except the person?
Answer	公共汽车是红色的。 The bus is red.	香蕉。 Bananas.	羊。 Sheep.

Sample questions and answers



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.


Deep learning
Example:
MLPs
Example:
Shallow
autoencoders
Representation learning
Example:
Logistic
regression
Machine learning
Example:
Knowledge
bases
AI

Introduction

What is machine learning

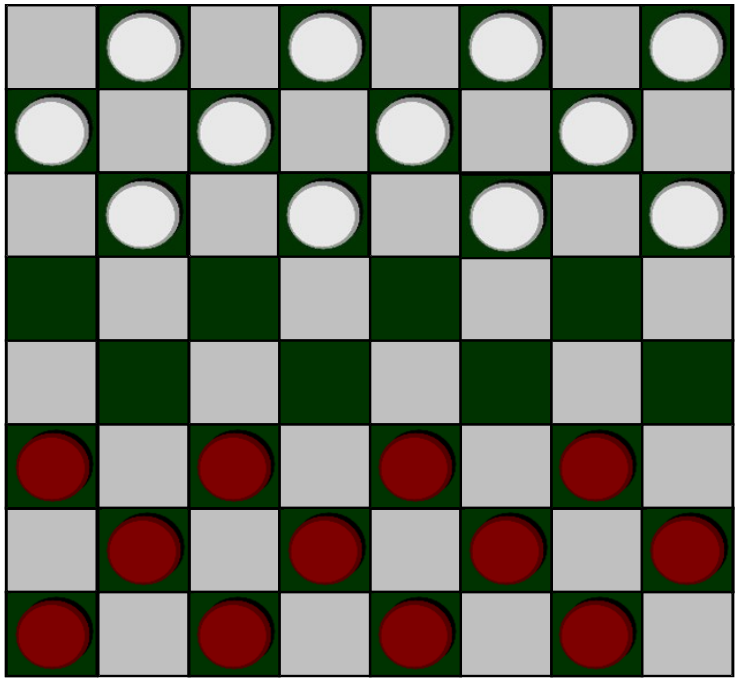
Machine Learning definition

Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.



Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Machine learning algorithms:

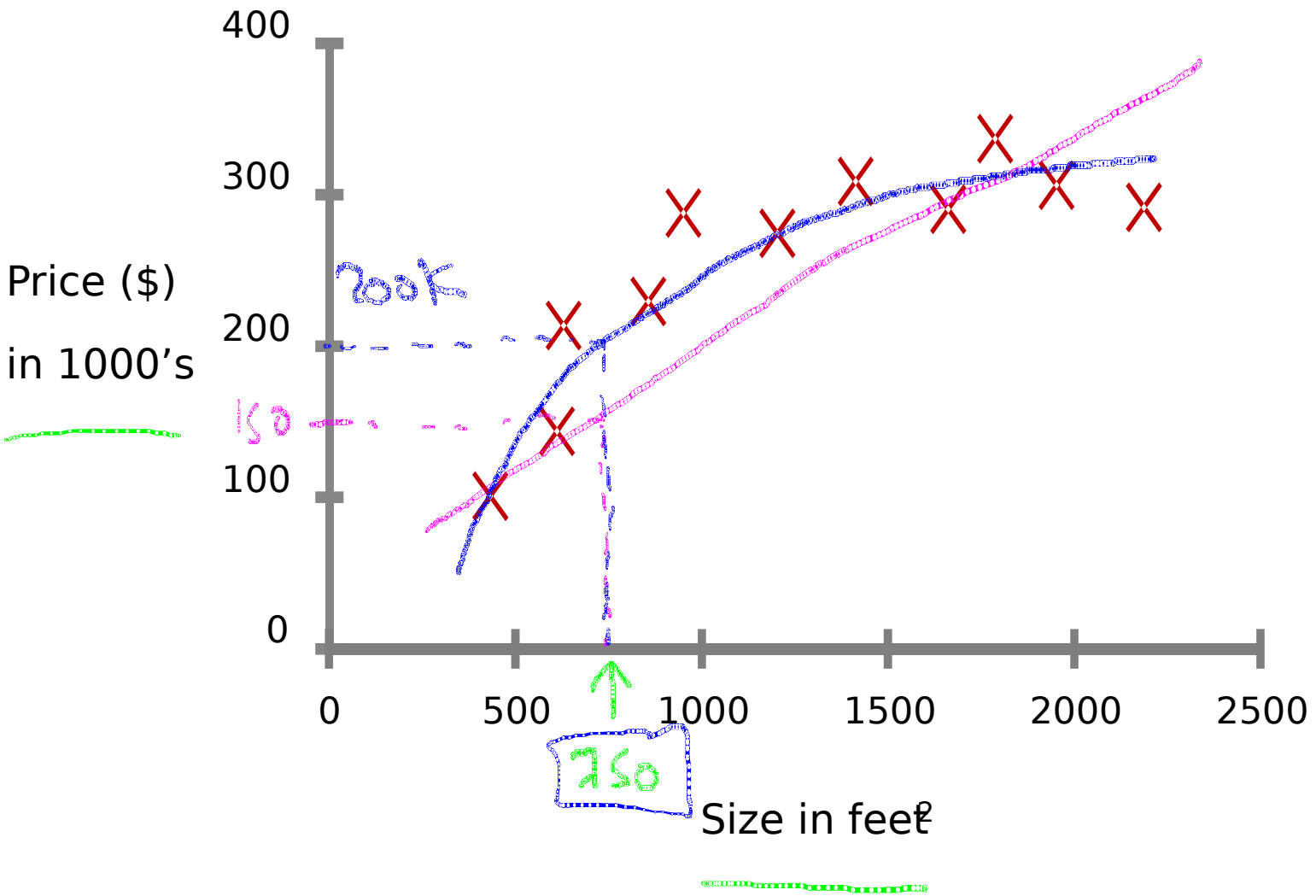
- Supervised learning
- Unsupervised learning

Others: Reinforcement learning, recommender systems.

Also talk about: Practical advice for applying learning algorithms.

Supervised Learning

Housing price prediction.

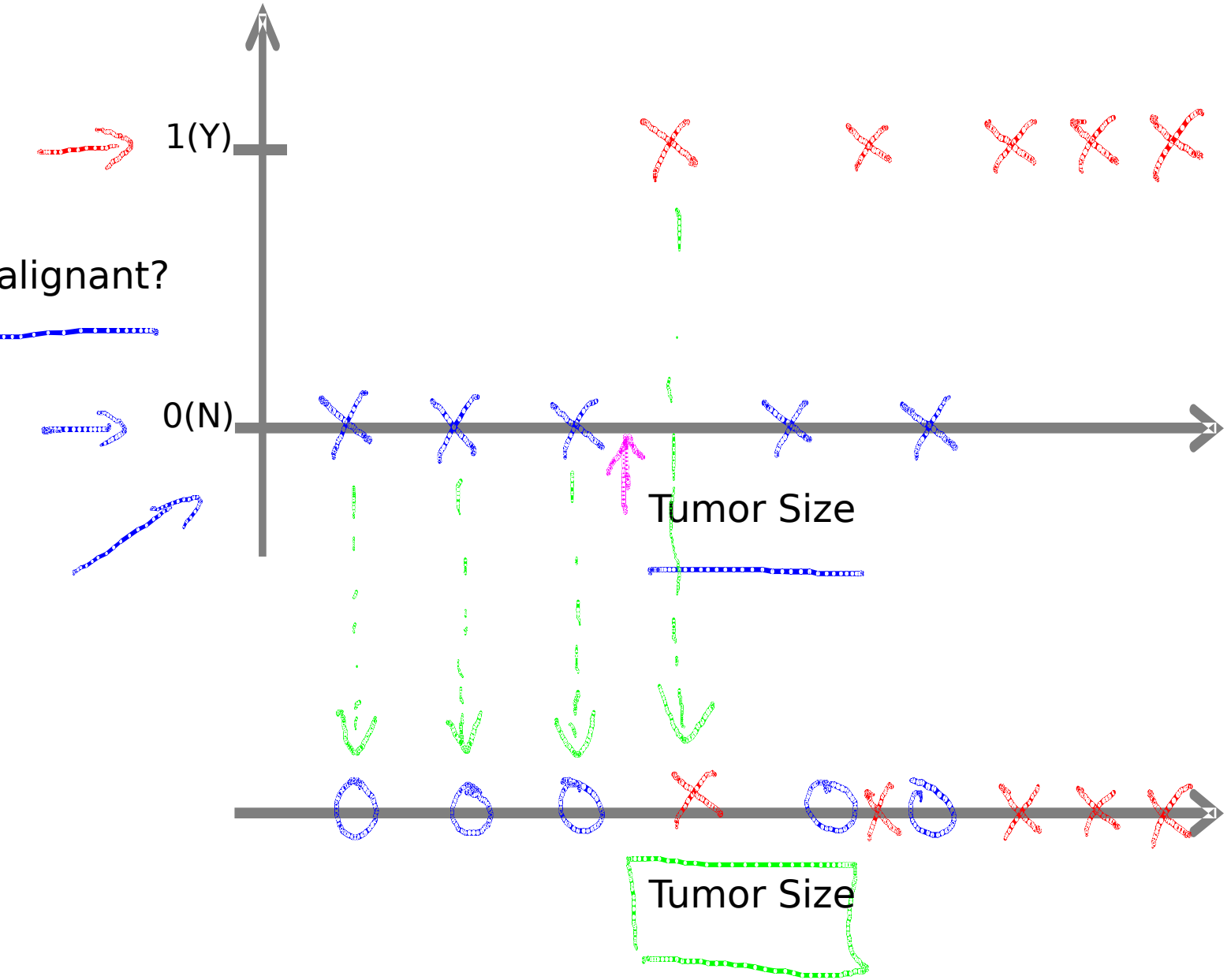


Supervised Learning

"right answers" given

Regression: Predict continuous valued output (price)

Breast cancer (malignant, benign)



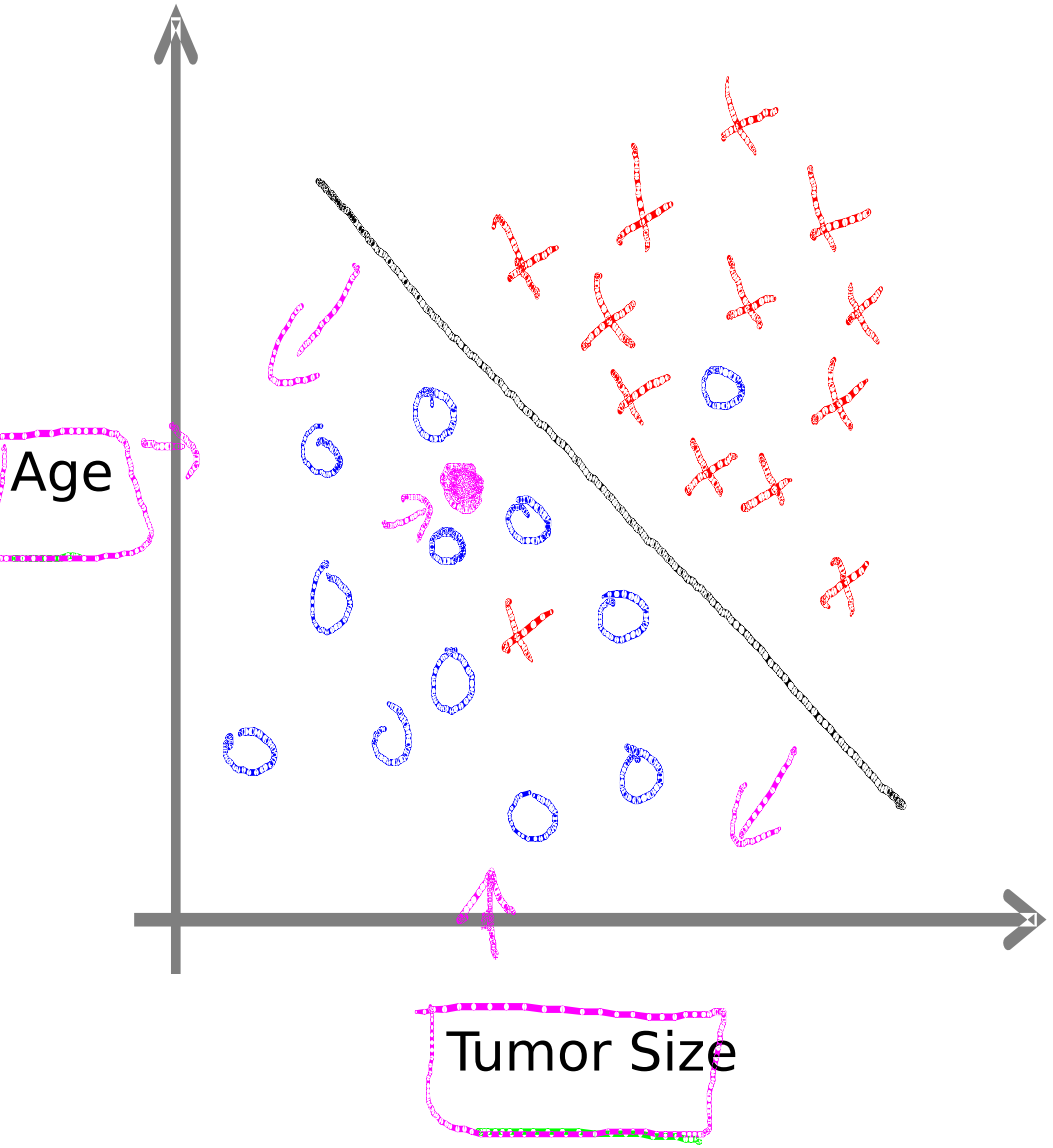
Classification

Discrete value

output (0 or 1)

0, 1, 2, 3, ...

benign, type 1 cancer

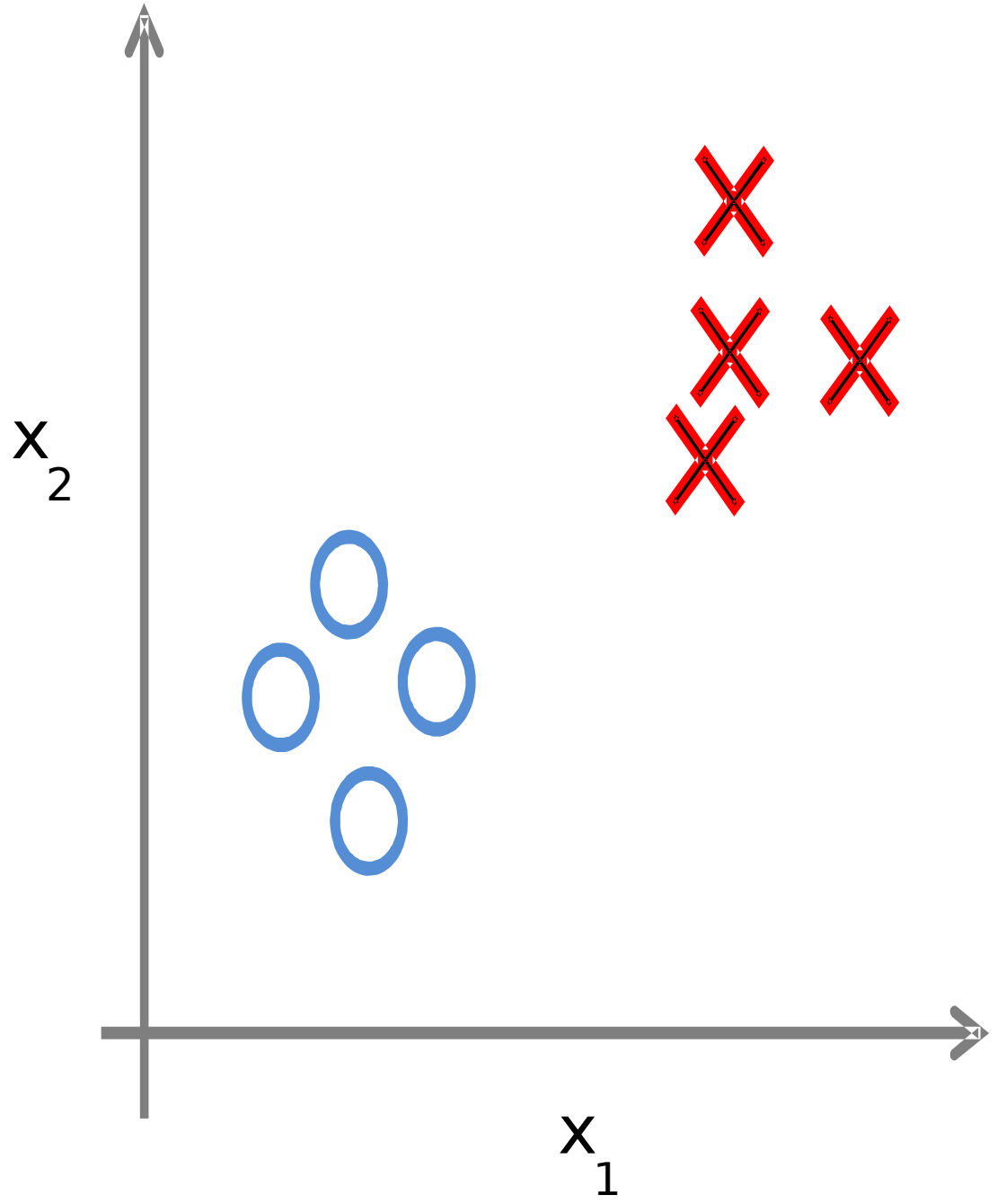


- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

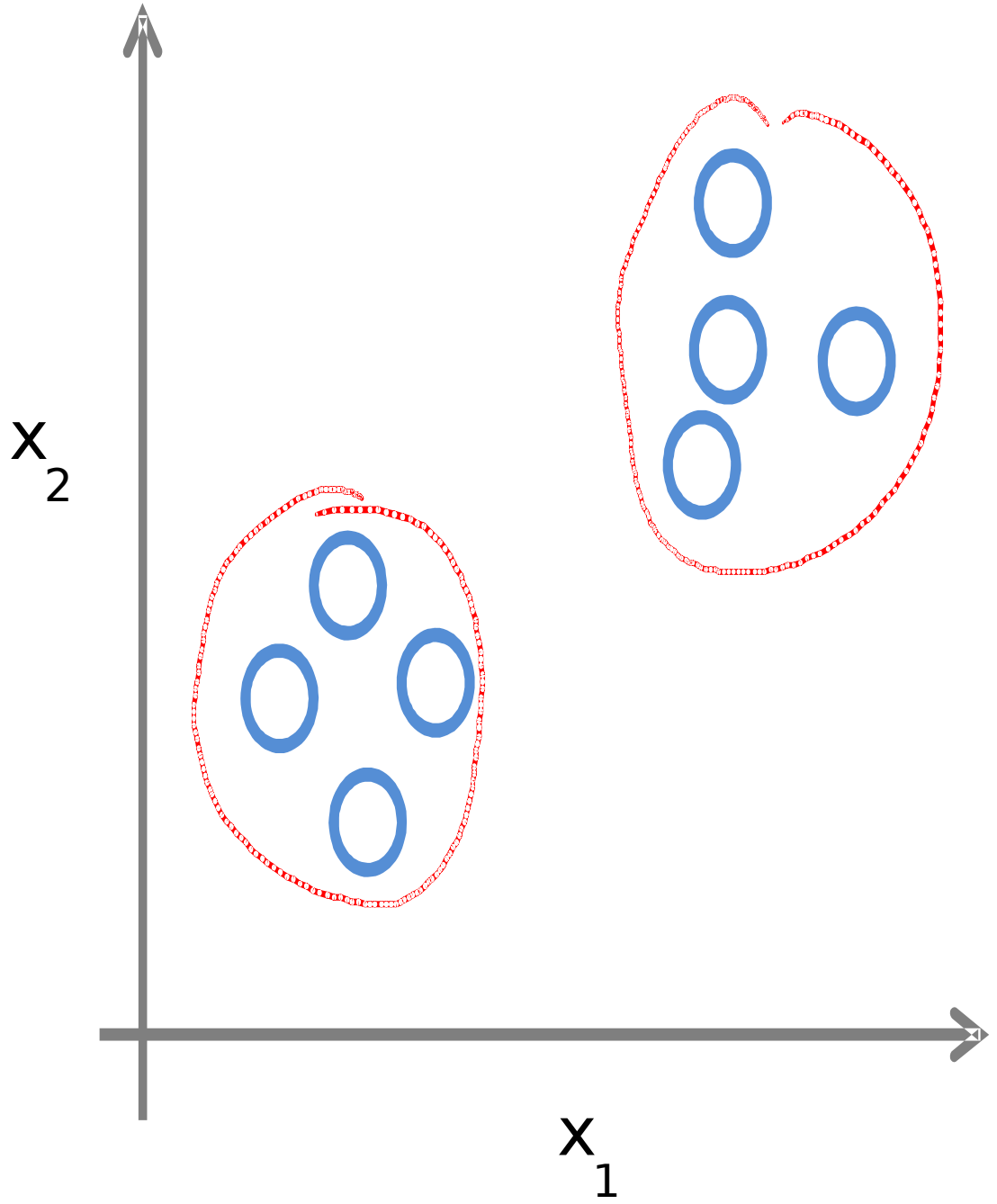
...

Unsupervised Learning

Supervised Learning



Unsupervised Learning



news

Search News | Search the Web

Advanced news search

Web Images Vi

U.S. edition ▼ Add a section »

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 Interna...

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 ...

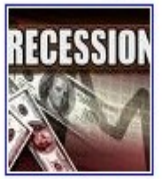
Tea Party is misplacing the blame, former President Bill Clinton claims New York Daily News

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

Atlanta Journal Constitution - Politics Daily - MyFox Washington DC - Salon

all 726 news articles »

US Stocks Climb After Recession Called Over, Homebuilders Gain



MyFox Phila...

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 »

Deepwater Horizon »

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

Vanity Fair - Juli Weiner - 22 minutes ago

all 276 news articles »

Deepwater Horizon »

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

Reuters

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 »

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

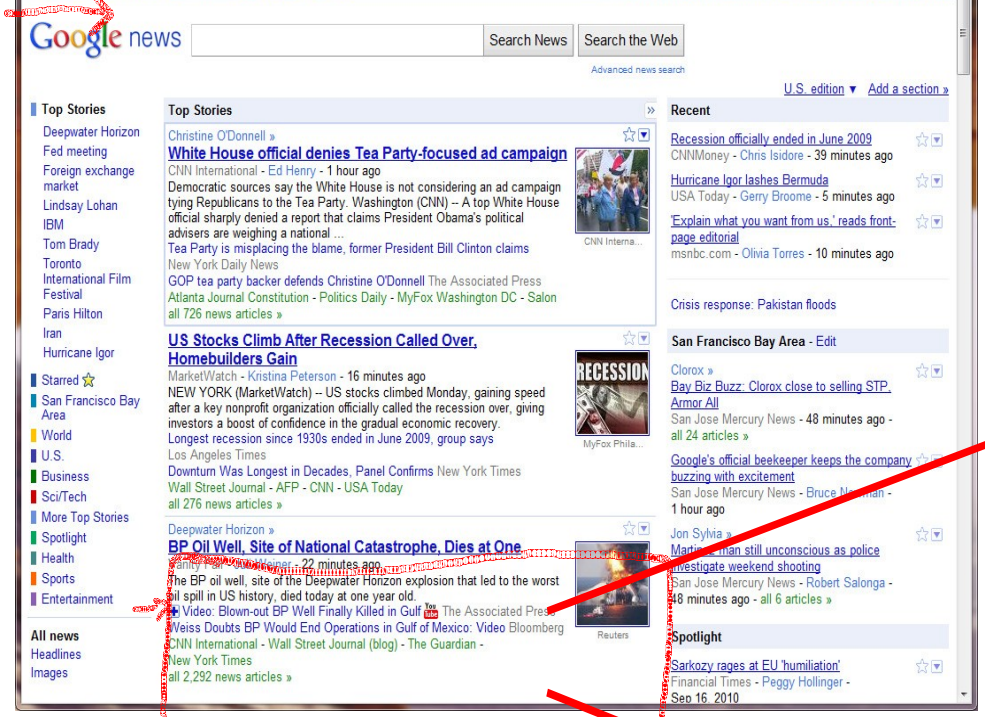
Clorox »
Bay Biz Buzz: Clorox close to selling STP, Armor All
San Jose Mercury News - 48 minutes ago - all 24 articles »

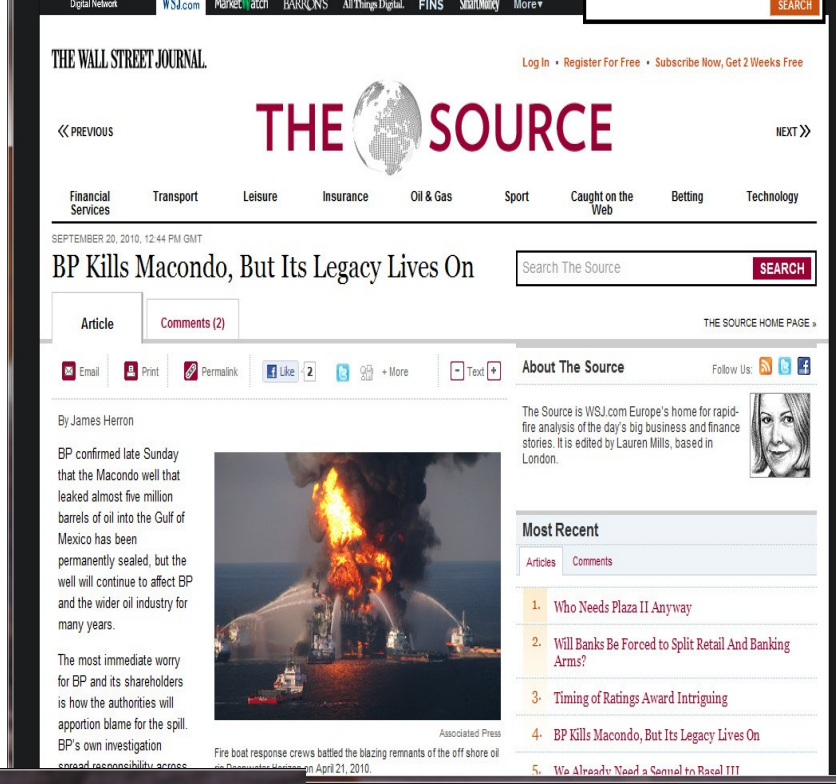
Google's official beekeeper keeps the company buzzing with excitement
San Jose Mercury News - Bruce Newman - 1 hour ago

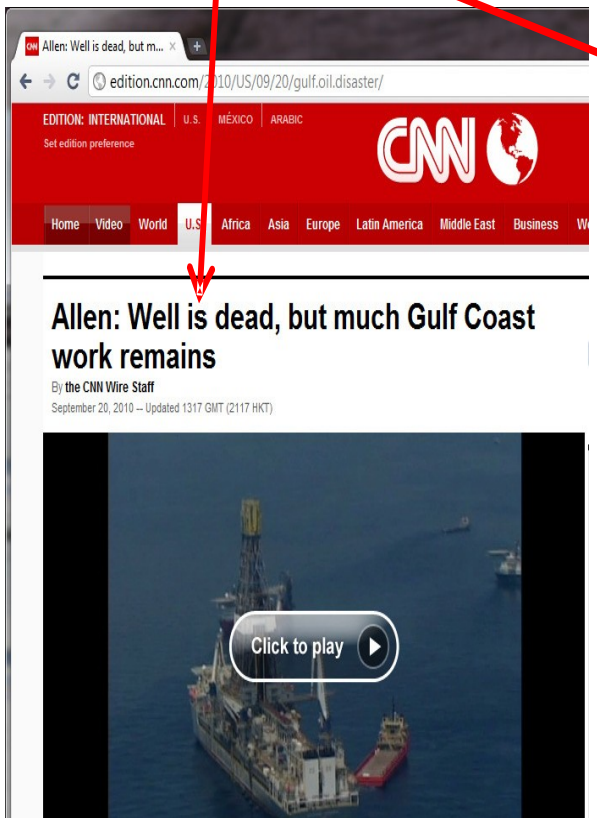
Jon Sylvia »
Martinez man still unconscious as police investigate weekend shooting
San Jose Mercury News - Robert Salonga - 48 minutes ago - all 6 articles »

Spotlight

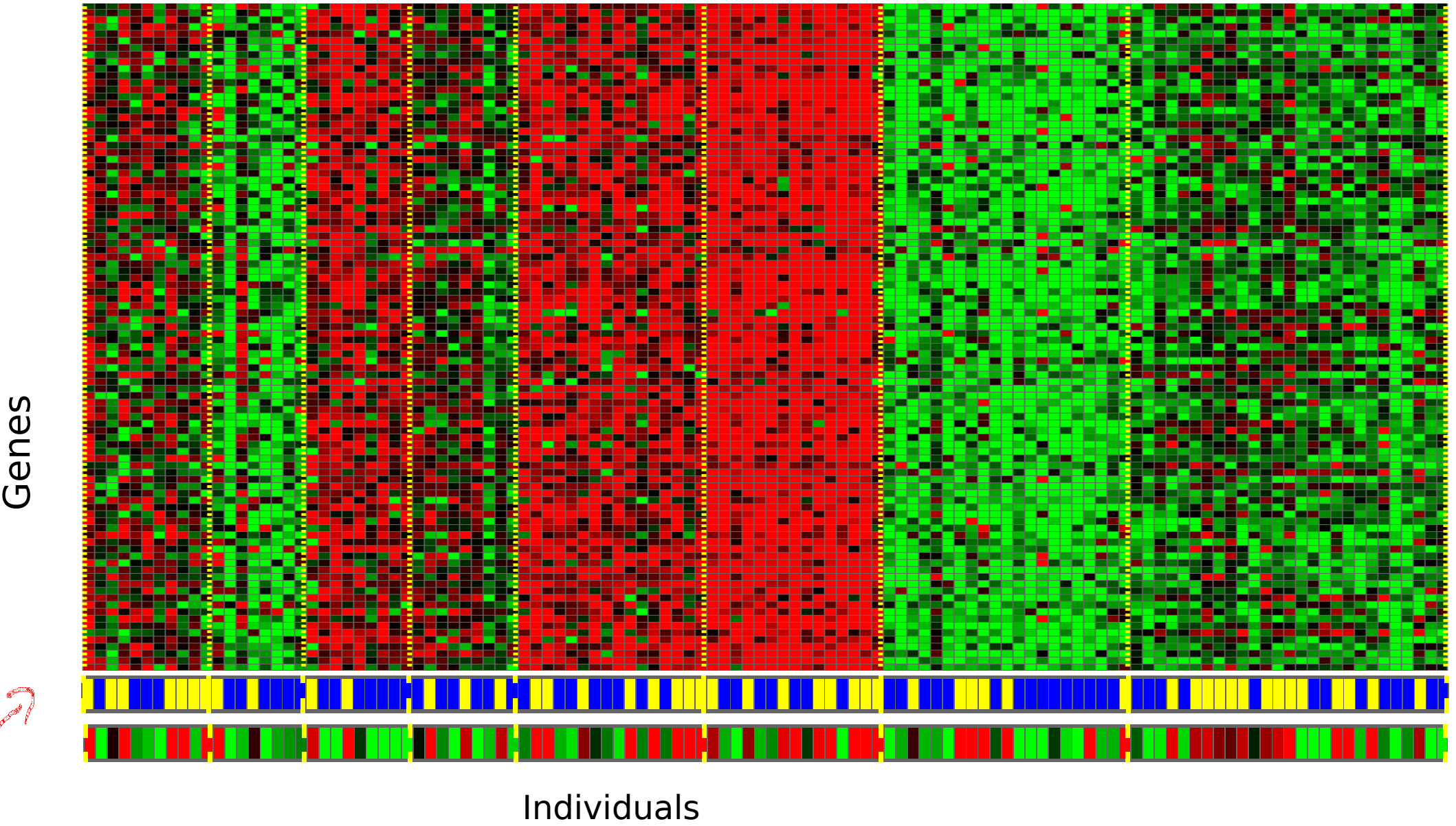
Sarkozy rages at EU 'humiliation'
Financial Times - Peggy Hollinger - Sep 16, 2010

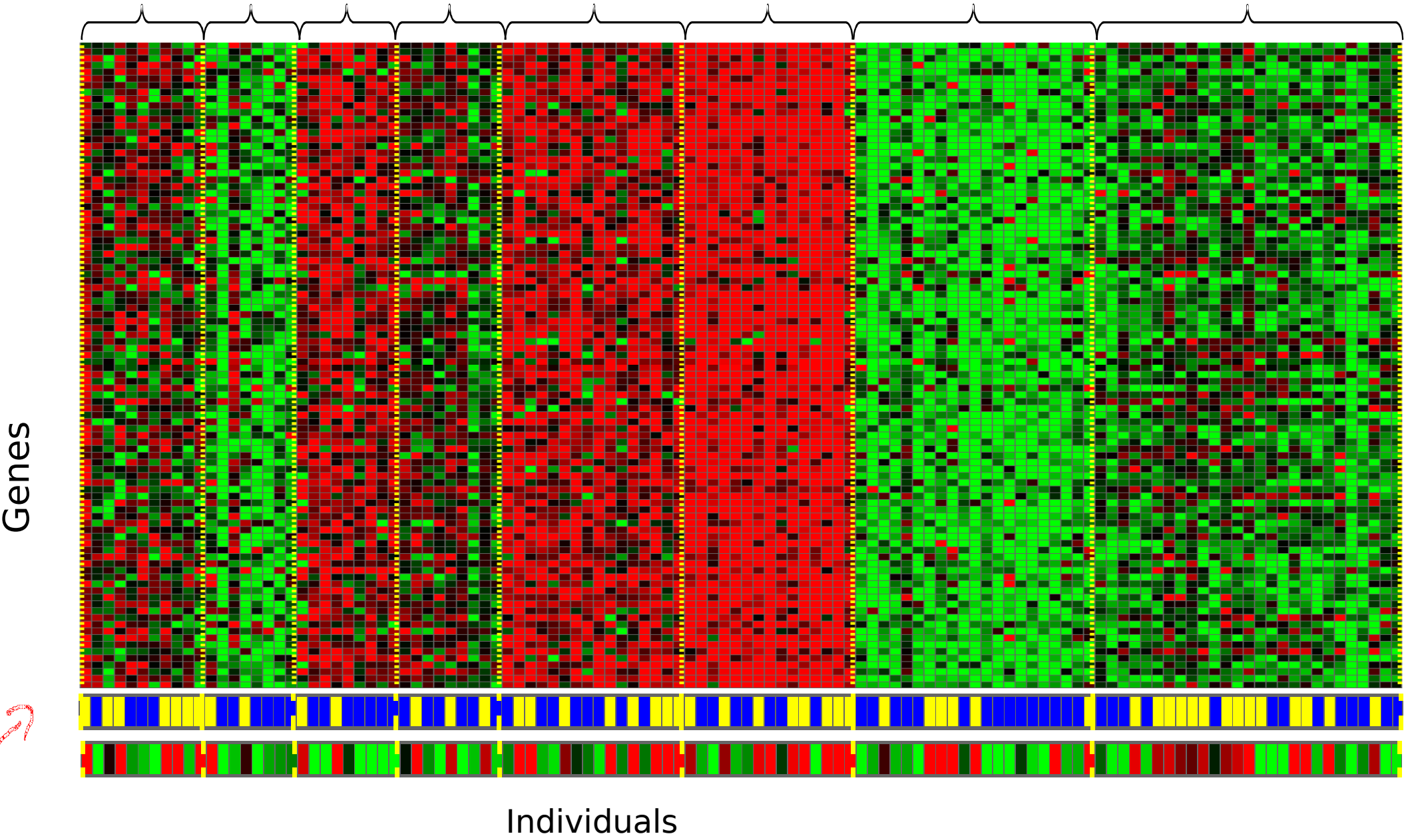

Google news
Search News
Search the Web
Advanced news search
U.S. edition ▼ Add a section »
Top Stories
Deepwater Horizon
Fed meeting
Foreign exchange market
Lindsay Lohan
IBM
Tom Brady
Toronto International Film Festival
Paris Hilton
Iran
Hurricane Igor
Starred
San Francisco Bay Area
World
U.S.
Business
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 »
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 »
Deepwater Horizon »
BP Oil Well, Site of National Catastrophe, Dies at One
Reuters - 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 »
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


THE WALL STREET JOURNAL.
Log In • Register For Free • Subscribe Now, Get 2 Weeks Free
« PREVIOUS
THE SOURCE
NEXT »
Financial Services
Transport
Leisure
Insurance
Oil & Gas
Sport
Caught on the Web
Betting
Technology
SEPTEMBER 20, 2010, 12:44 PM GMT
BP Kills Macondo, But Its Legacy Lives On
Search The Source
SEARCH
Article
Comments (2)
THE SOURCE HOME PAGE »
Email
Print
Permalink
Like 2
+ More
Text
About The Source
Follow Us:
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.
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
Associated Press
Fire boat response crews battled the blazing remnants of the off shore oil
Most Recent
Articles
Comments
1. Who Needs Plaza II 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
5. We Already Need a Sequel to Basel III

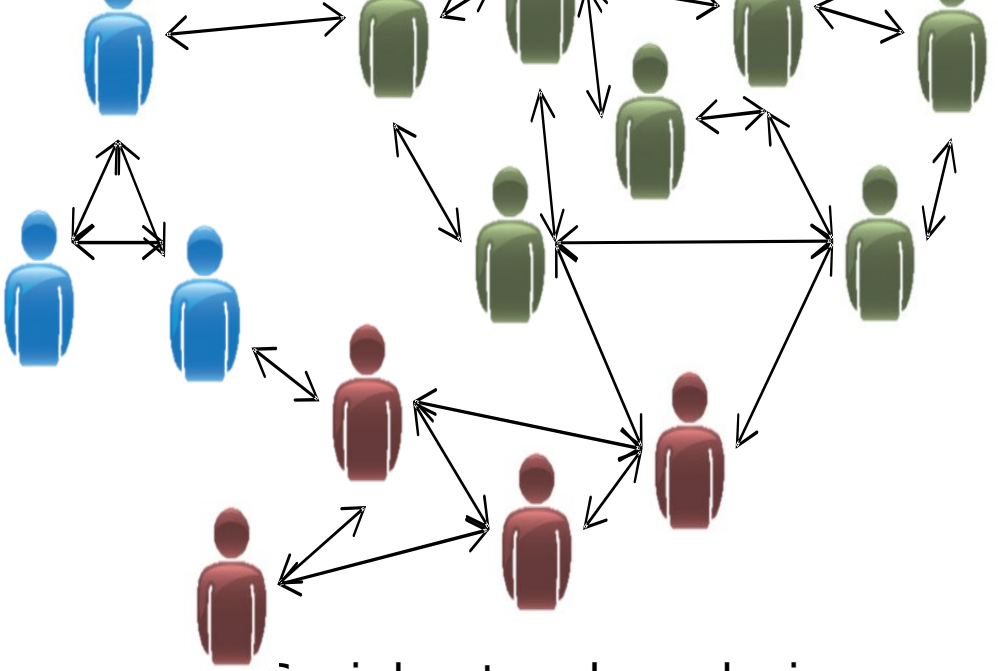

Allen: Well is dead, but m...
edition.cnn.com/2010/US/09/20/gulf.oil.disaster/
EDITION: INTERNATIONAL
U.S.
MÉXICO
ARABIC
Set edition preference
CNN
Home
Video
World
U.S.
Africa
Asia
Europe
Latin America
Middle East
Business
Allen: Well is dead, but much Gulf Coast work remains
By the CNN Wire Staff
September 20, 2010 -- Updated 1317 GMT (2117 HKT)
Click to play


BP oil spill cost hits nearl...
www.guardian.co.uk/environment/2010/sep/20/bp-oil-spill-dee
Mobile site
Sign in
Register
Text larger
smaller
About us
guardian.co.uk
News
Sport
Comment
Culture
Business
Money
Life & style
Business
BP
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 totalling more than $240m
Julia Kollewe
guardian.co.uk, Monday 20 September 2010 08.33 BST
Article history

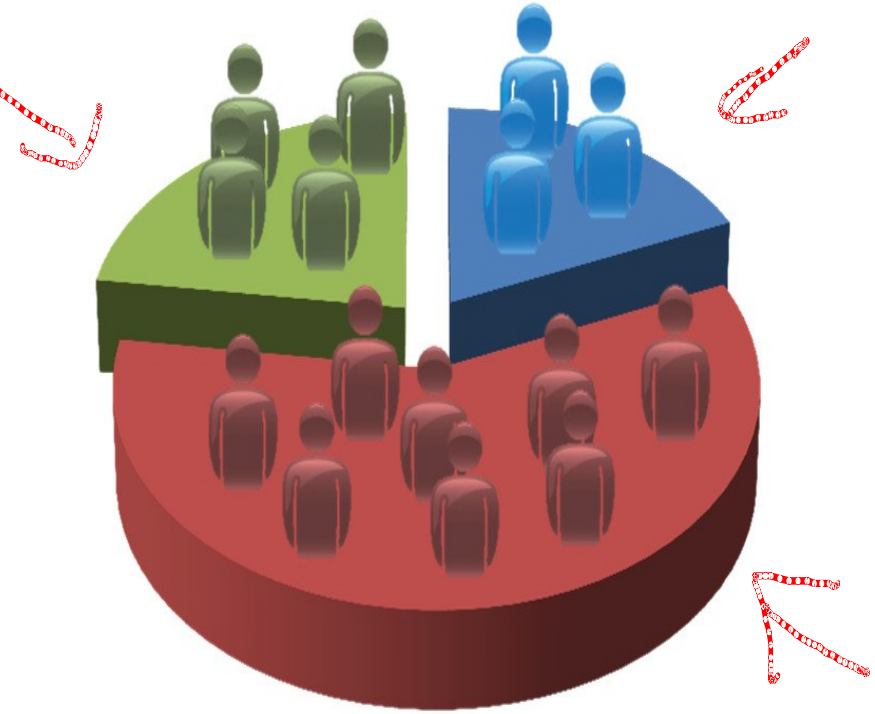

Genes
Individuals


Genes
Individuals

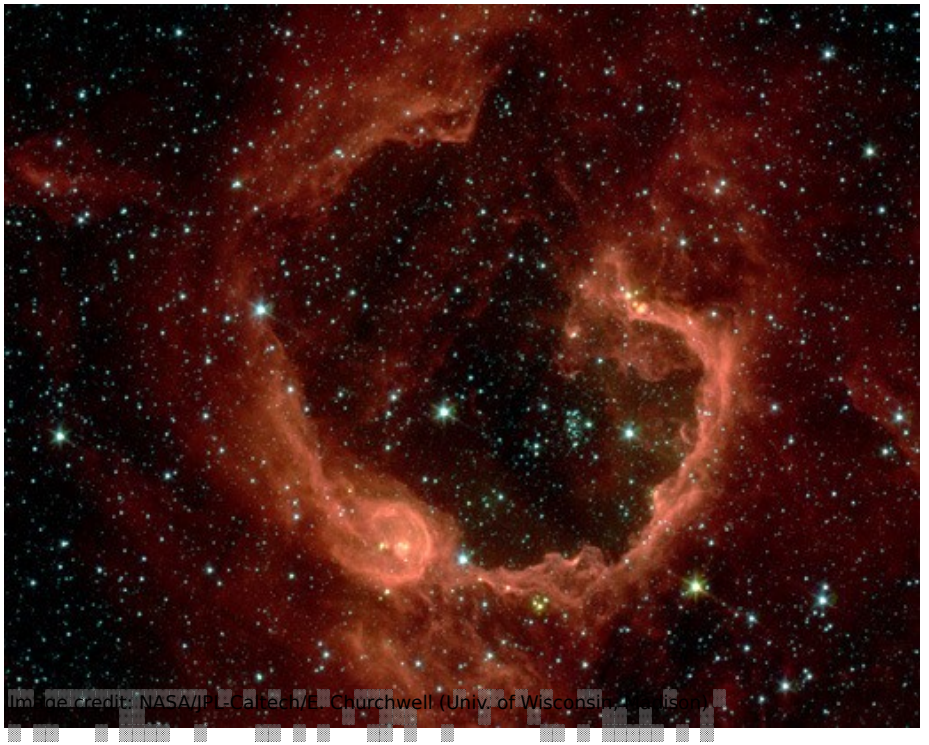
Organize computing clusters



Social network analysis

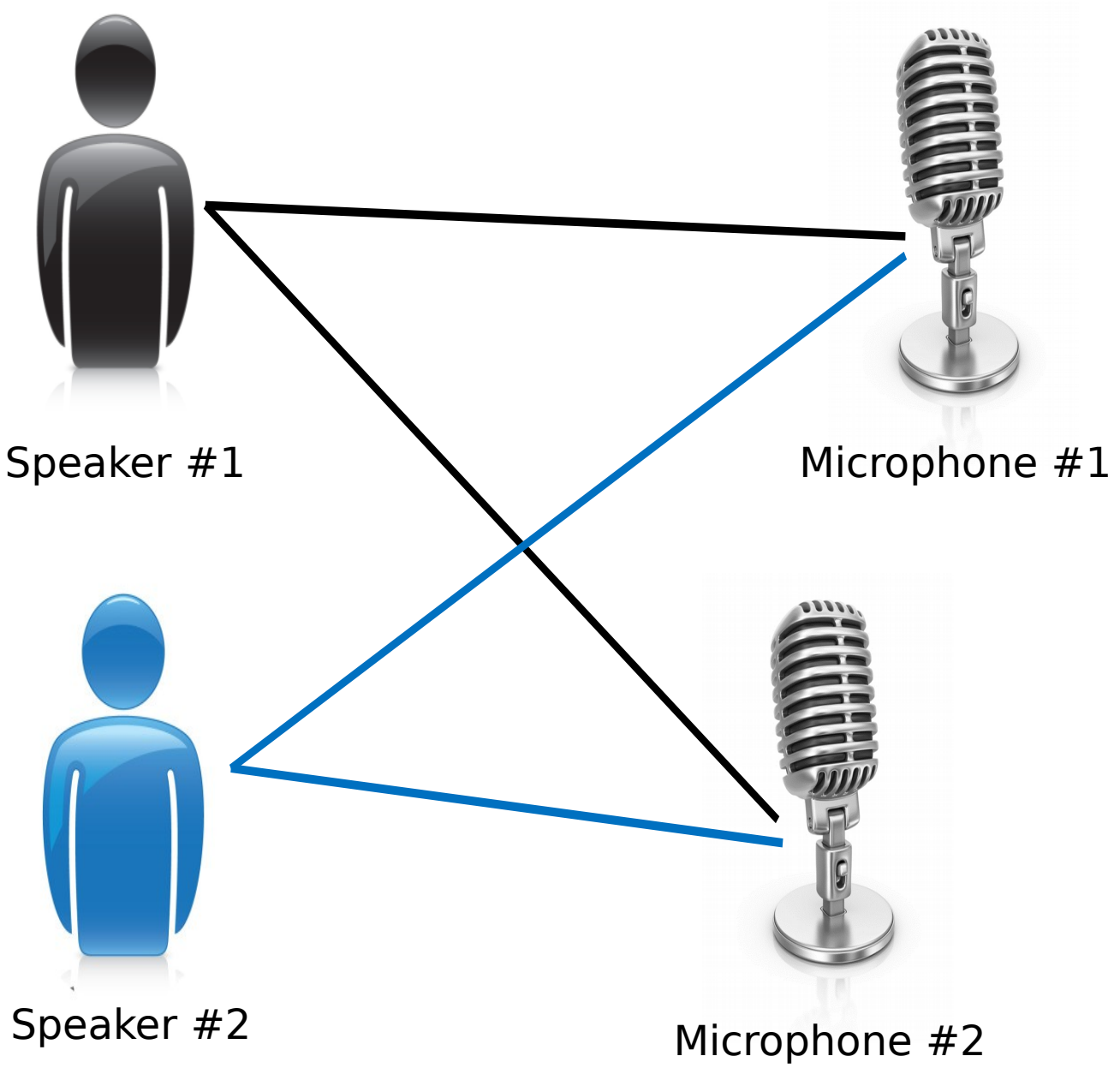


Market segmentation



Astronomical data analysis

Cocktail party problem

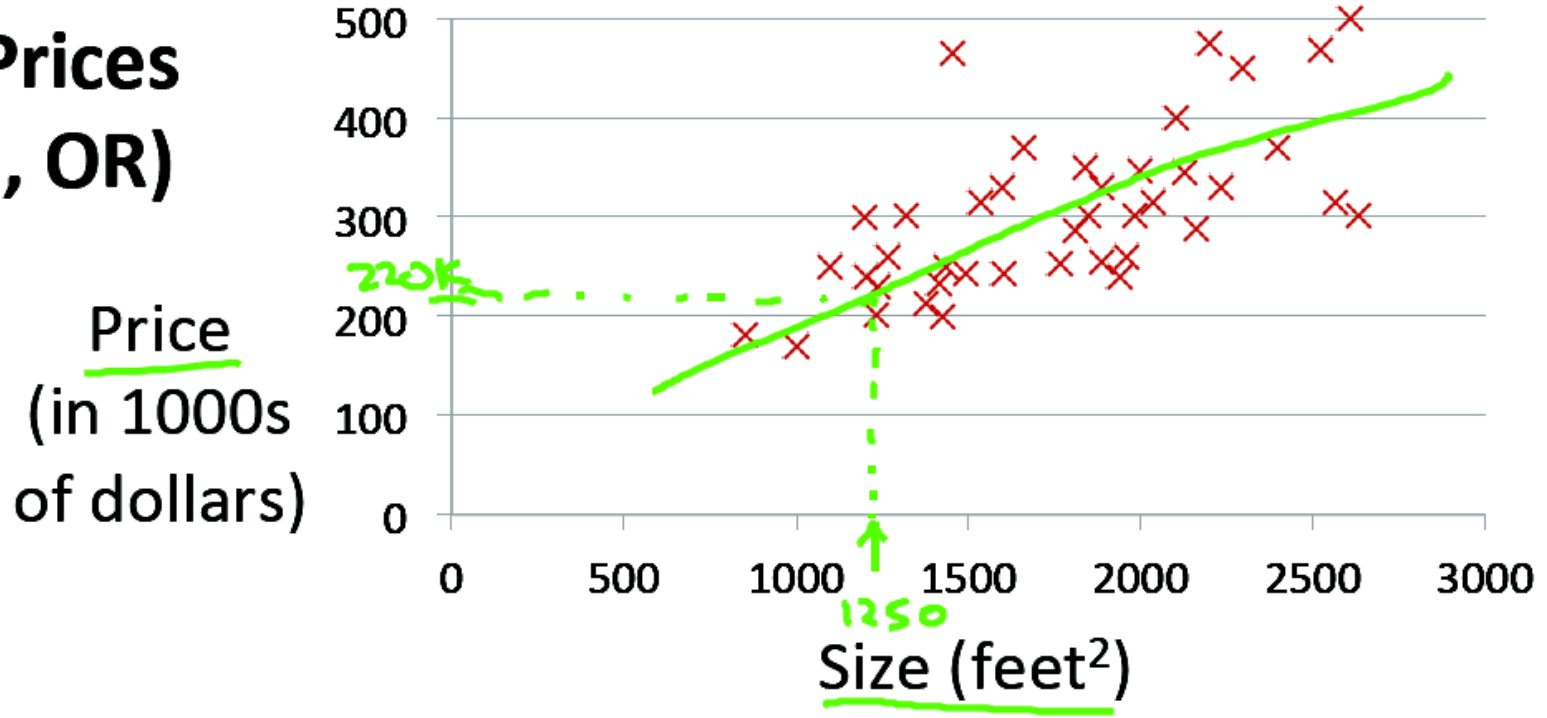


Cocktail party problem algorithm

```
[W,s,v] = svd((repmat(sum(x.*x,1),size(x,1),1).*x)*x');
```

Linear Regression

Housing Prices (Portland, OR)



Supervised Learning

Given the "right answer" for each example in the data.

Regression Problem

Predict real-valued output

Classification: Discrete-valued output

Linear Regression

Training set of housing prices (Portland, OR)

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

m = 47

Notation:

- **m** = Number of training examples
- **x**'s = "input" variable / features
- **y**'s = "output" variable / "target" variable

$x^{(1)} = 2104$

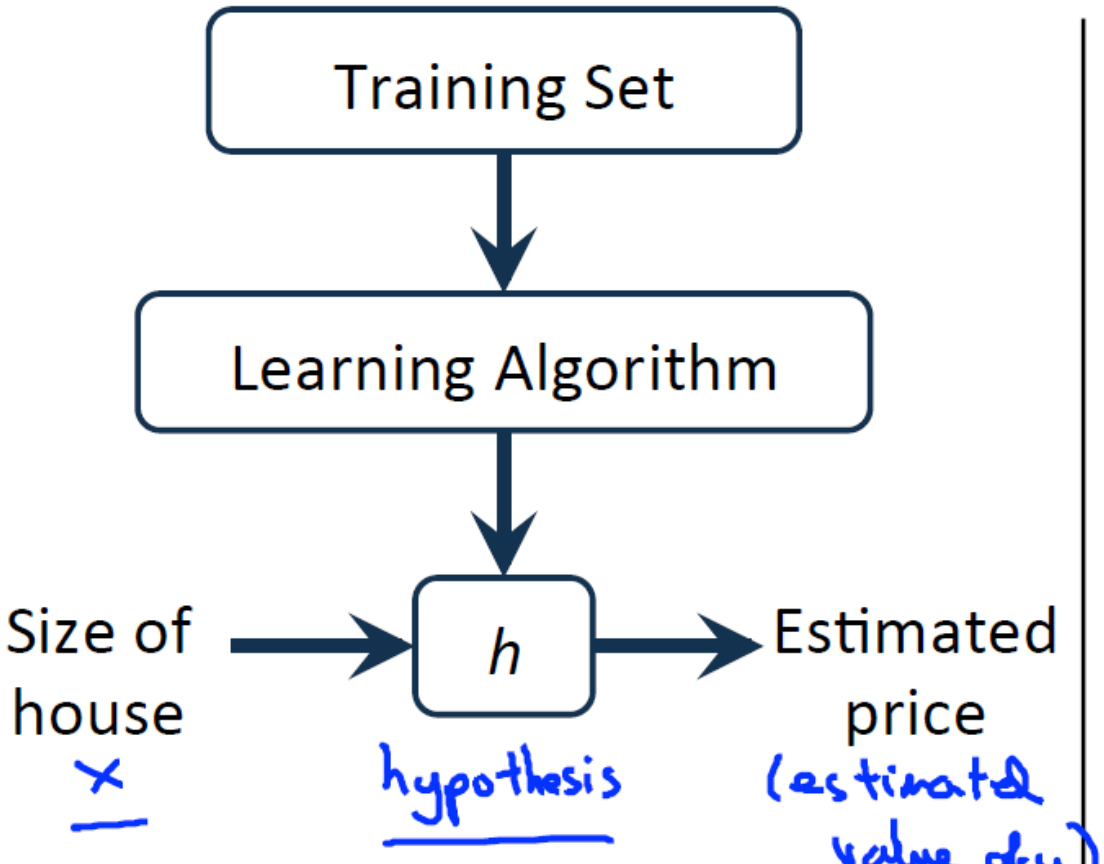
$x^{(2)} = 1416$

$y^{(1)} = 460$

(x, y) - one training example

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

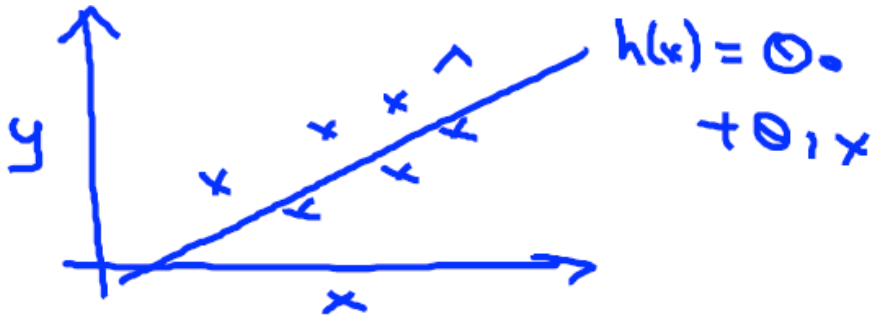
Linear Regression



How do we represent *h* ?

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

Shorthand: $h(x)$



Linear regression with one variable. (x)

Univariate linear regression.

└ one variable

Linear Regression:Cost Function

Training Set

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

m = 47

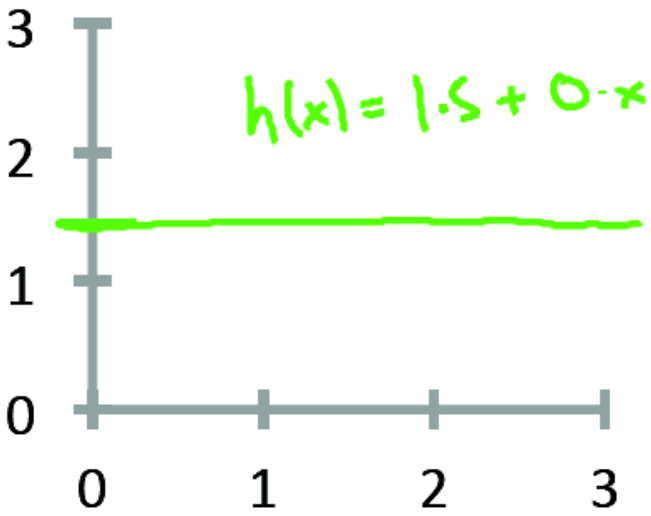
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

θ_i's: Parameters

How to choose θ_i's ?

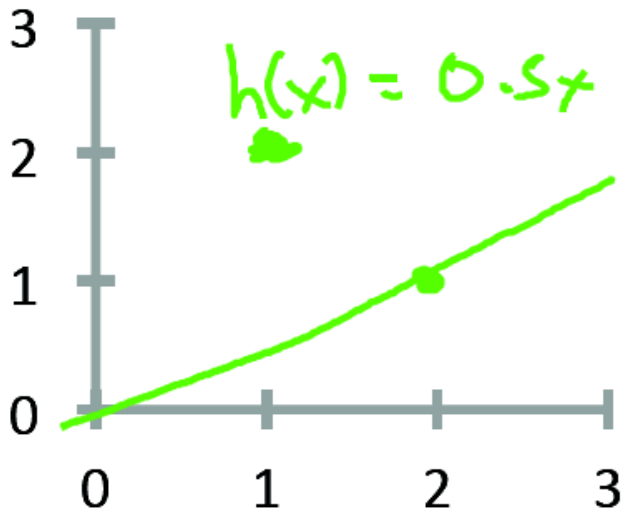
Linear Regression:Cost Function

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



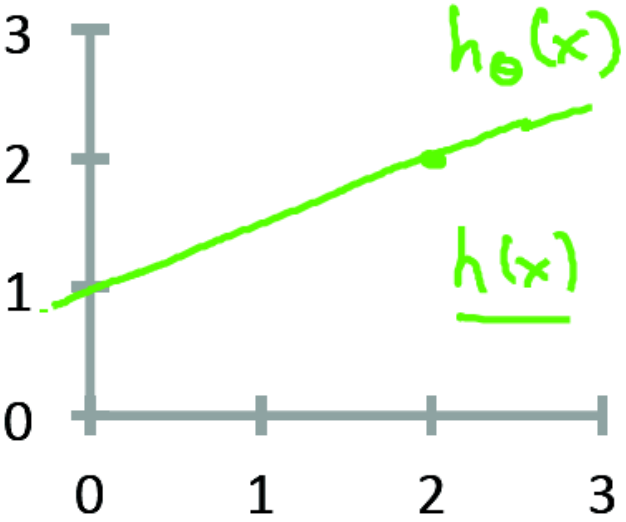
→ $\theta_0 = 1.5$

→ $\theta_1 = 0$



→ $\theta_0 = 0$

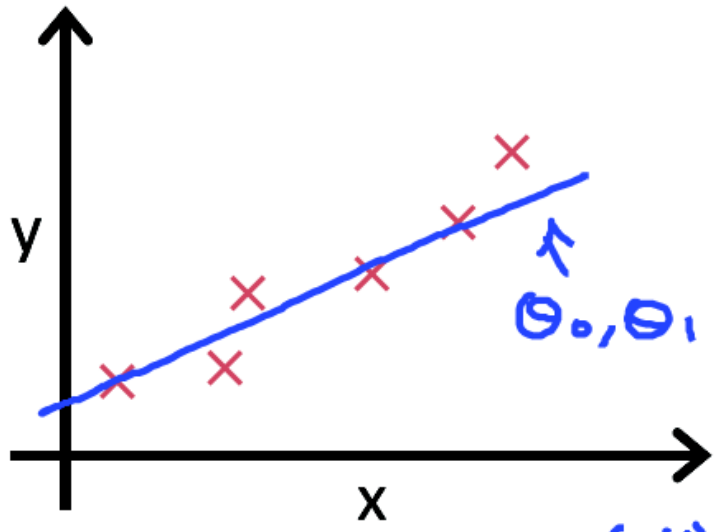
→ $\theta_1 = 0.5$



→ $\theta_0 = 1$

→ $\theta_1 = 0.5$

Linear Regression:Cost Function



#training examples

minimize θ_0, θ_1 $\frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

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

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

Idea: Choose θ_0, θ_1 so that $h_\theta(x)$ is close to y for our training examples (x, y)

x, y

Minimize θ_0, θ_1 $J(\theta_0, \theta_1)$

Cost function

Squared error function

Linear Regression:Cost Function:intuitively

Hypothesis:

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

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

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

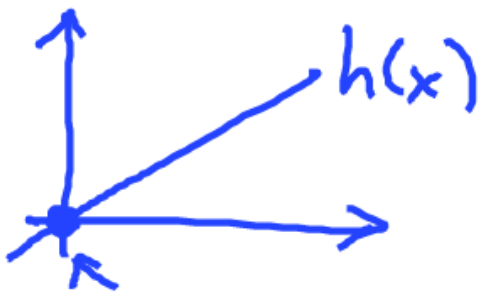
Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Simplified

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

$\theta_0 = 0$

θ_1



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

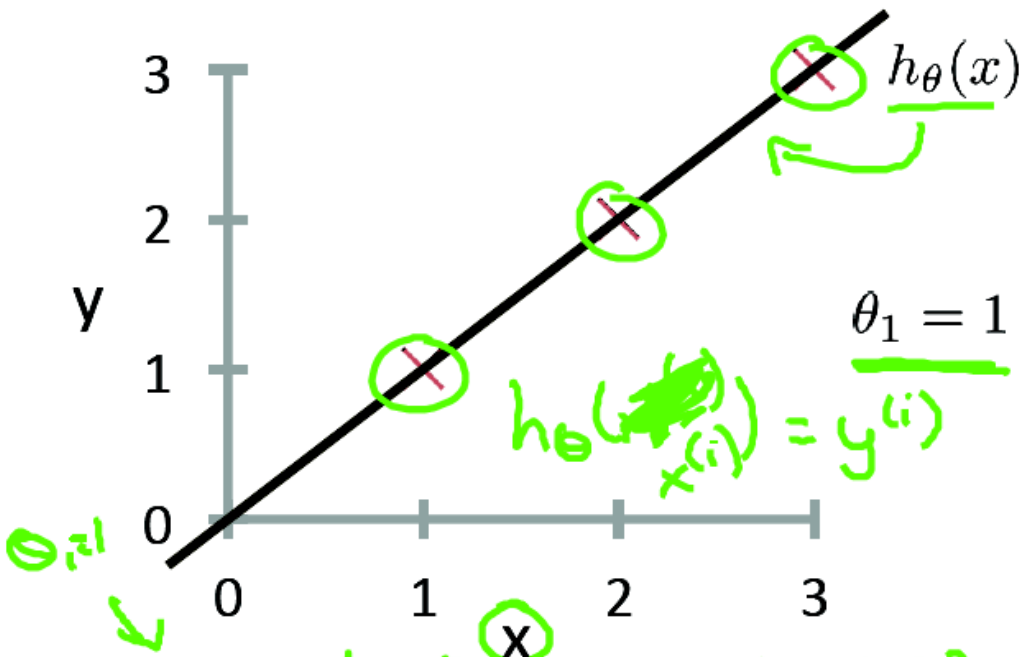
$\theta_1, x^{(i)}$

$$\underset{\theta_1}{\text{minimize}}\ J(\theta_1)$$

Linear Regression:Cost Function:intuitively

→ $h_\theta(x)$

(for fixed θ_1, this is a function of x)



$h_\theta(x^{(i)}) = y^{(i)}$

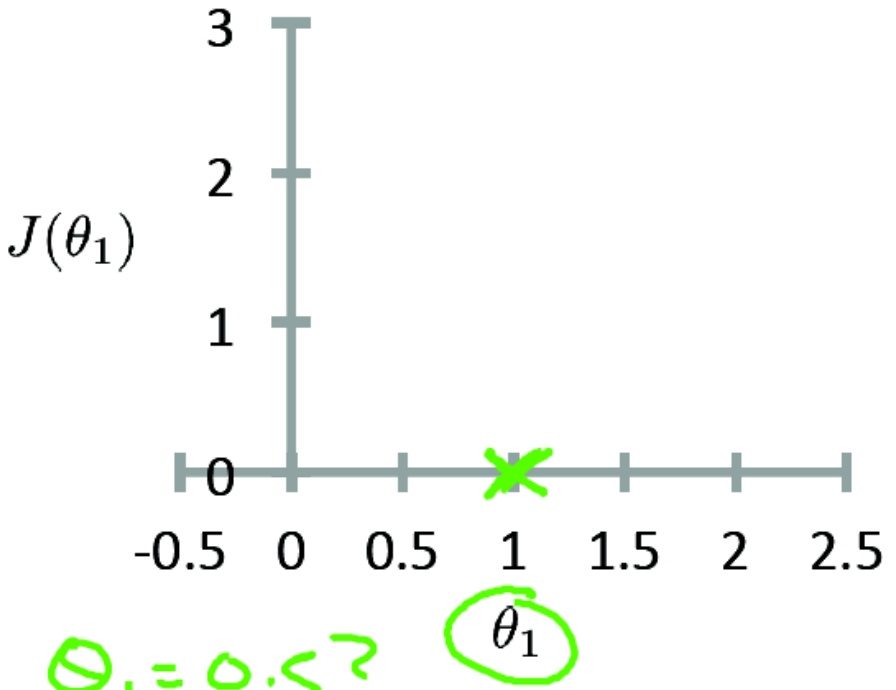
$\theta_1 = 1$

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

$= \frac{1}{2m} \sum_{i=1}^{m} (\theta_1 x^{(i)} - y^{(i)})^2 = \frac{1}{2m}(0^2 + 0^2 + 0^2) = 0$

→ $J(\theta_1)$

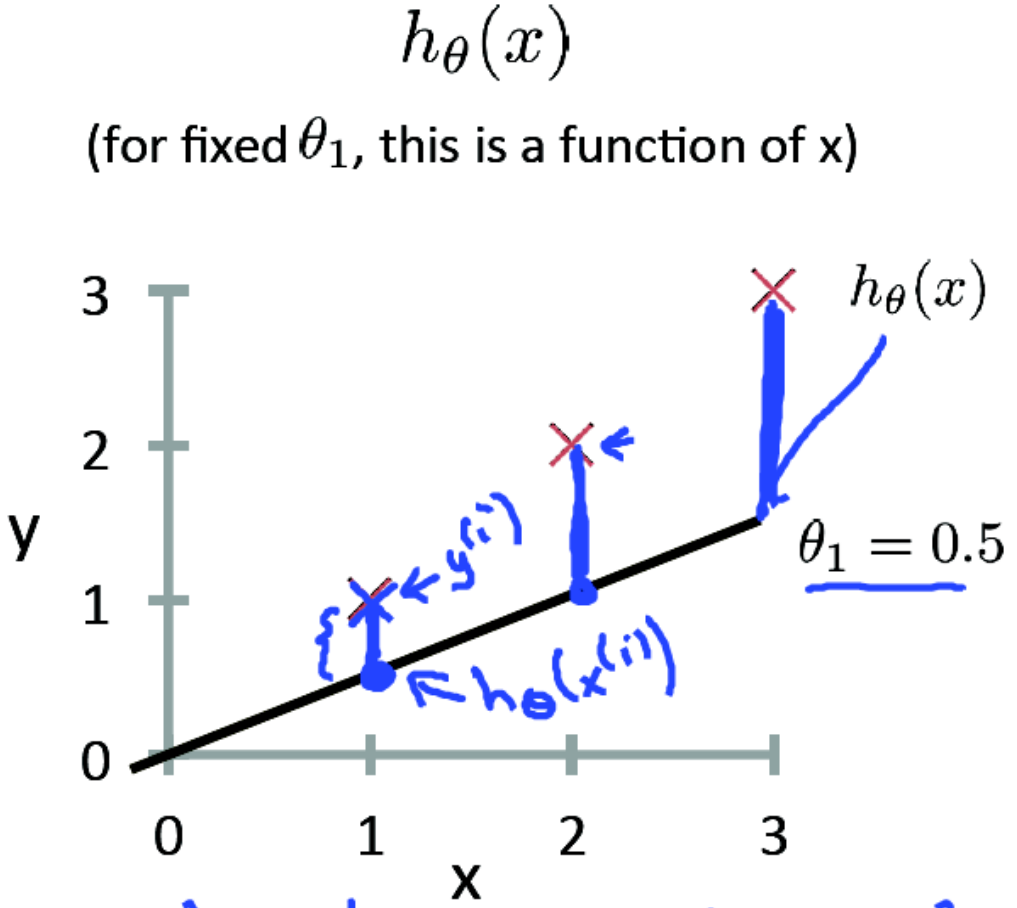
(function of the parameter θ_1)

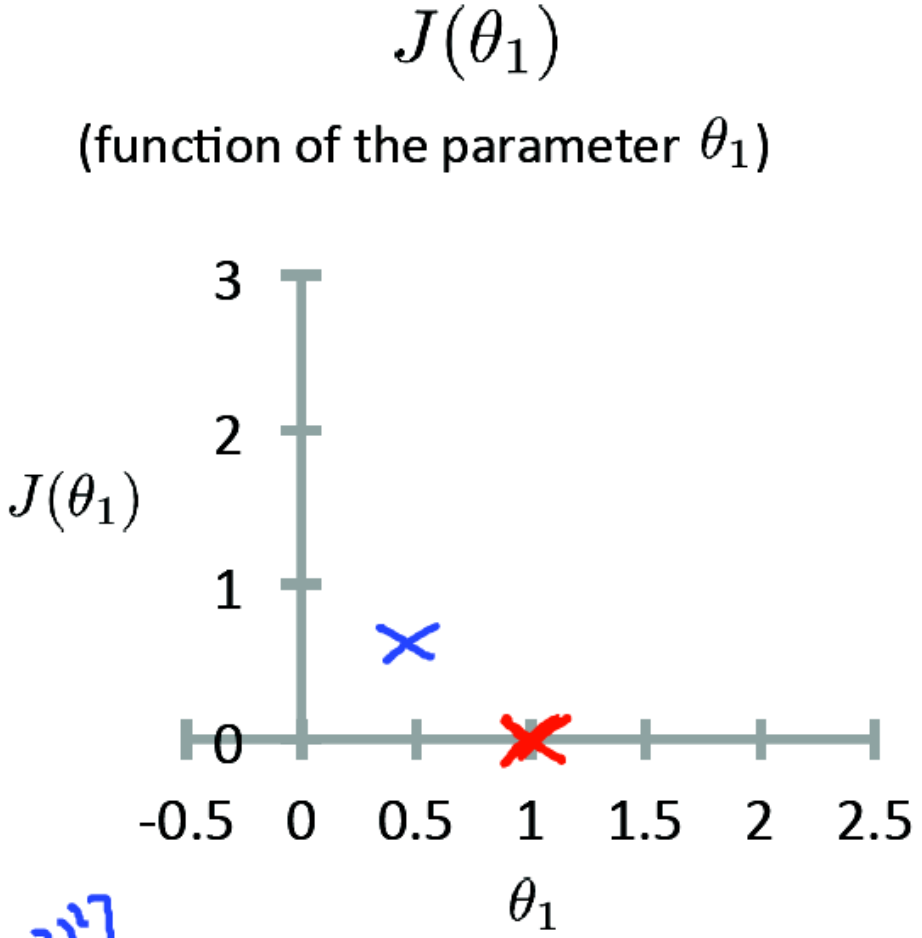


$\theta_1 = 0.5$?

$J(1) = 0$

Linear Regression:Cost Function:intuitively





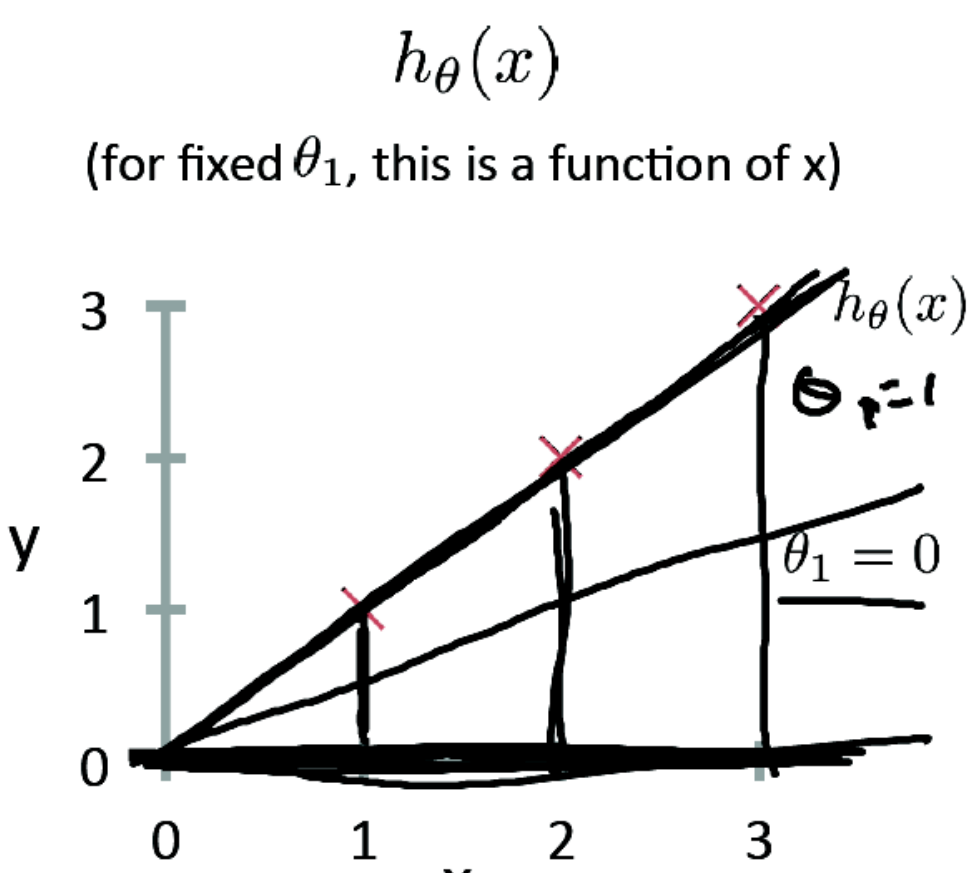
$J(0.5) = \frac{1}{2m}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right]$

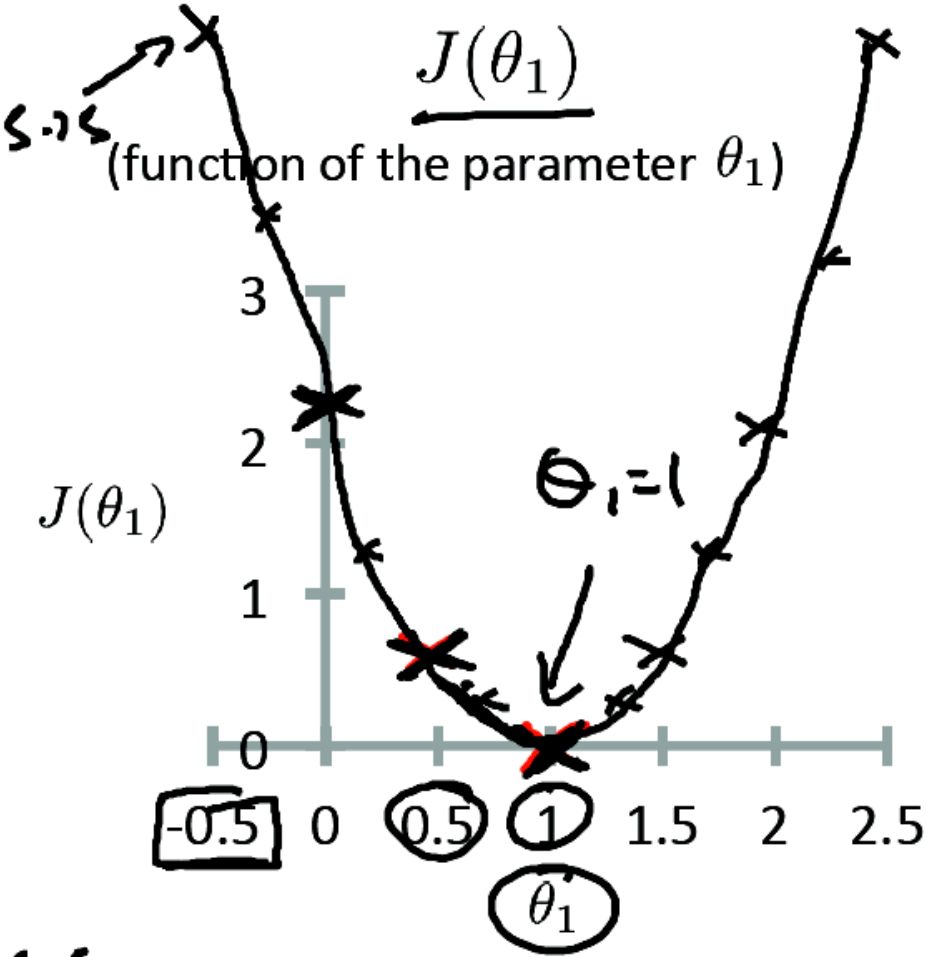
$= \frac{1}{2\times 3}(3.5) = \frac{3.5}{6} \approx 0.58$

$\theta_1 = 0?$

$J(0) = ?$

Linear Regression:Cost Function:intuitively





$$J(0) = \frac{1}{2m}(1^2 + 2^2 + 3^2)$$
$$= \frac{1}{6} \cdot 14 \approx 2.3$$

$h(x) = -0.5x$

Minimize $J(\theta_1)$
θ_1

Linear Regression:Gradient Descent

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

Parameters: θ_0, θ_1

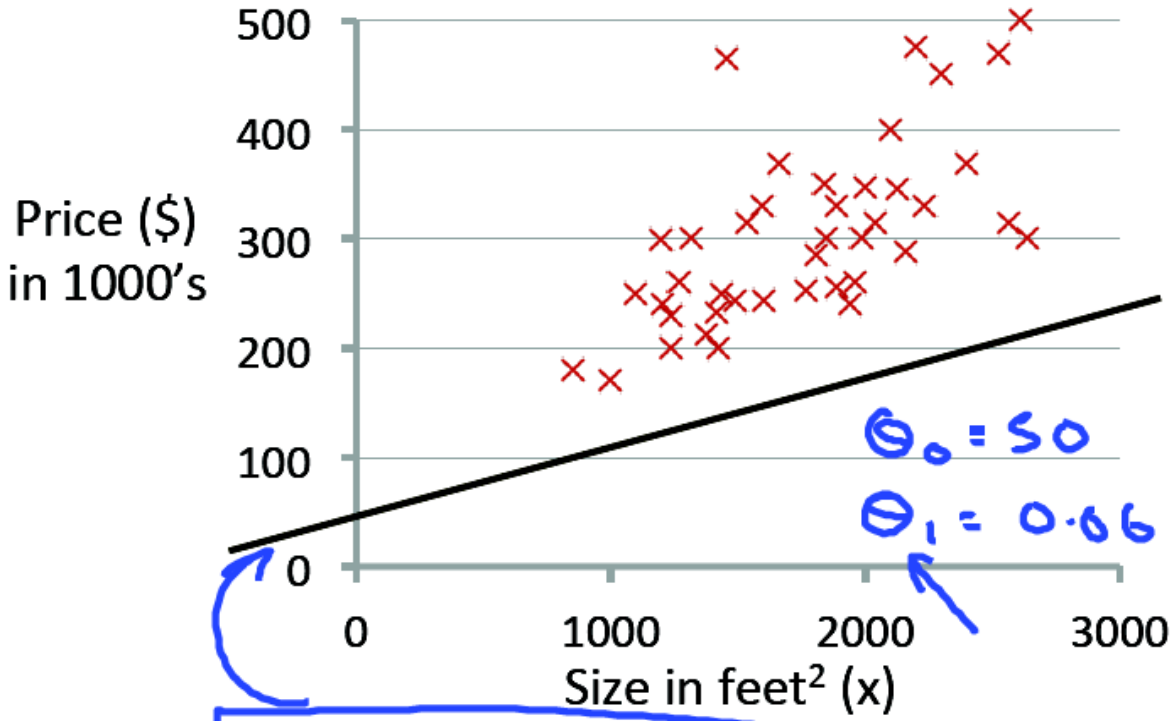
Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

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

Linear Regression:Gradient Descent

$h_\theta(x)$

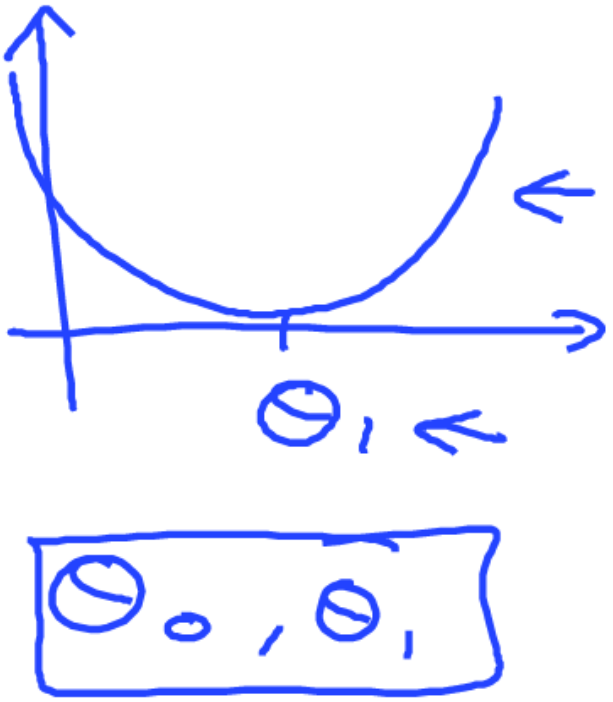
(for fixed θ_0, θ_1, this is a function of x)



$h_\theta(x) = 50 + 0.06x$

$J(\theta_0, \theta_1)$

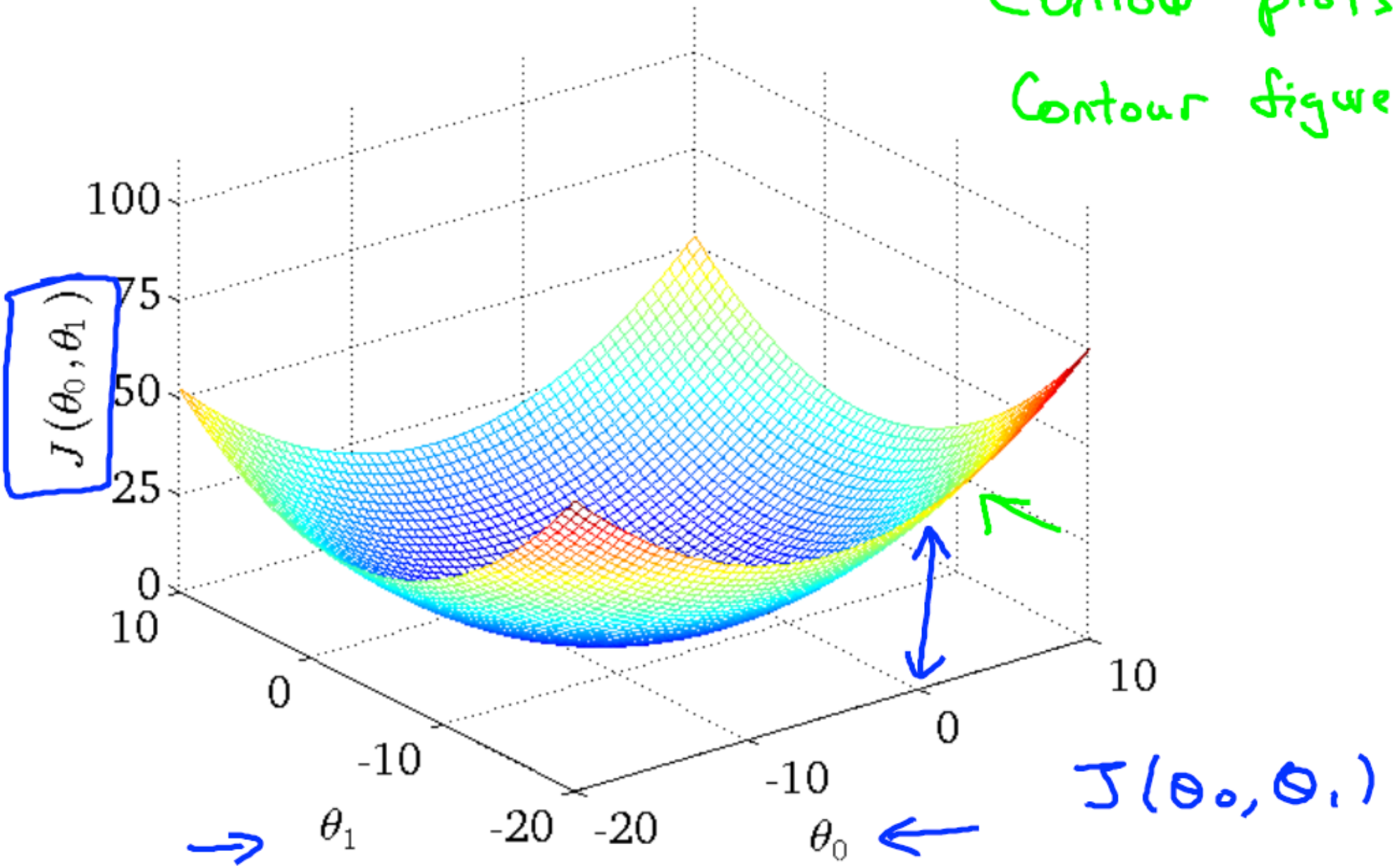
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent

Contour plots

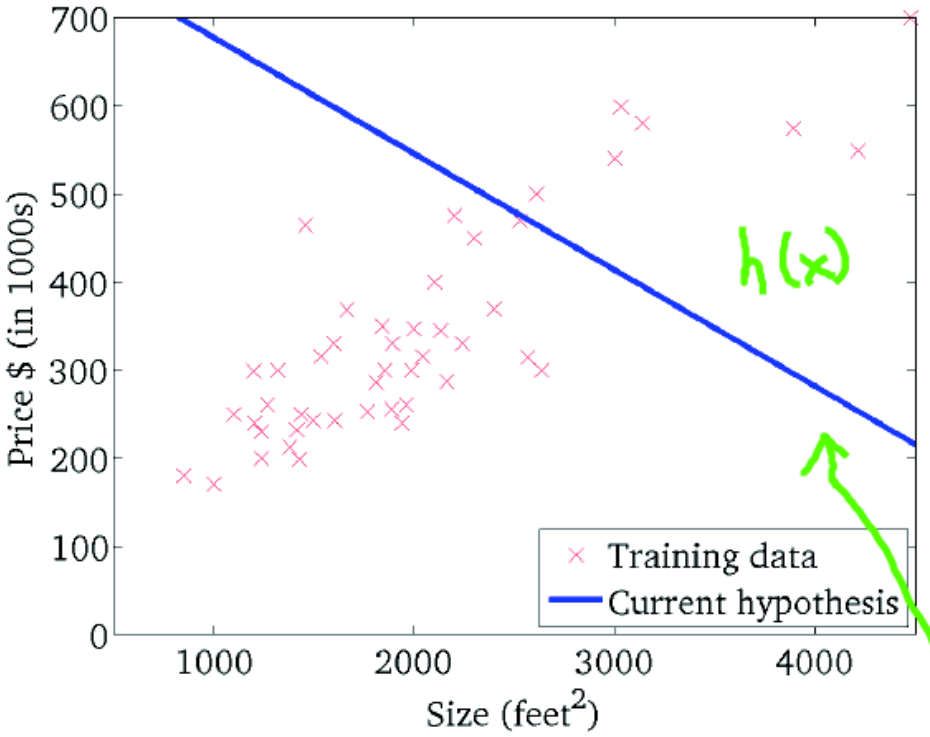
Contour figures -



Linear Regression:Gradient Descent

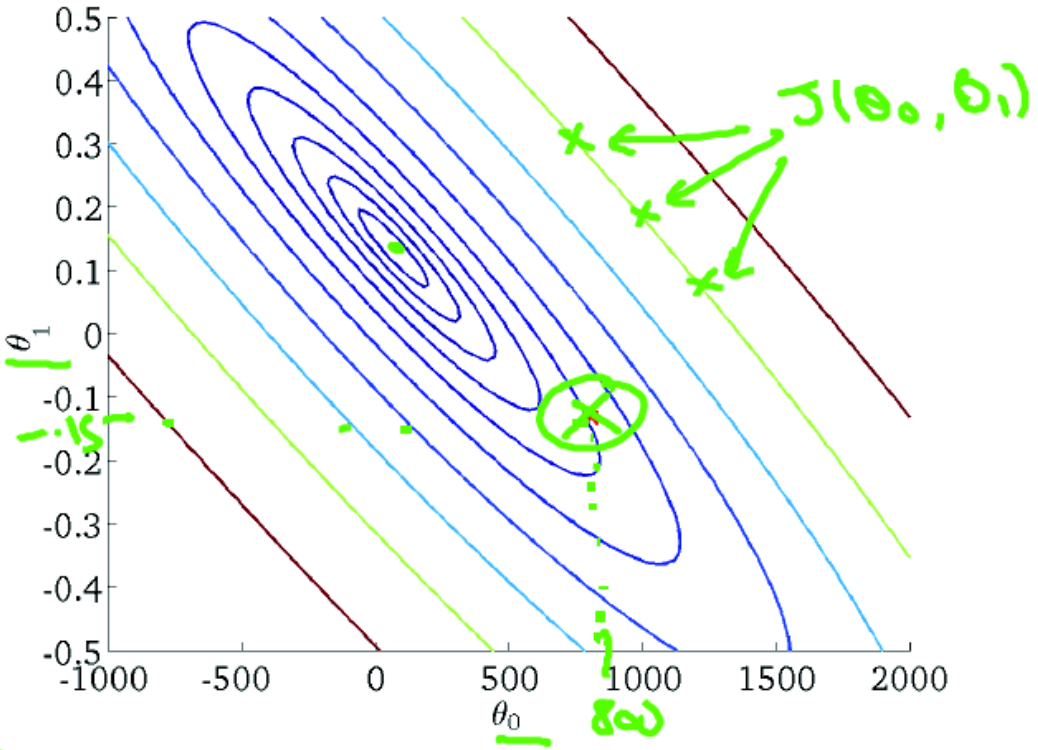
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



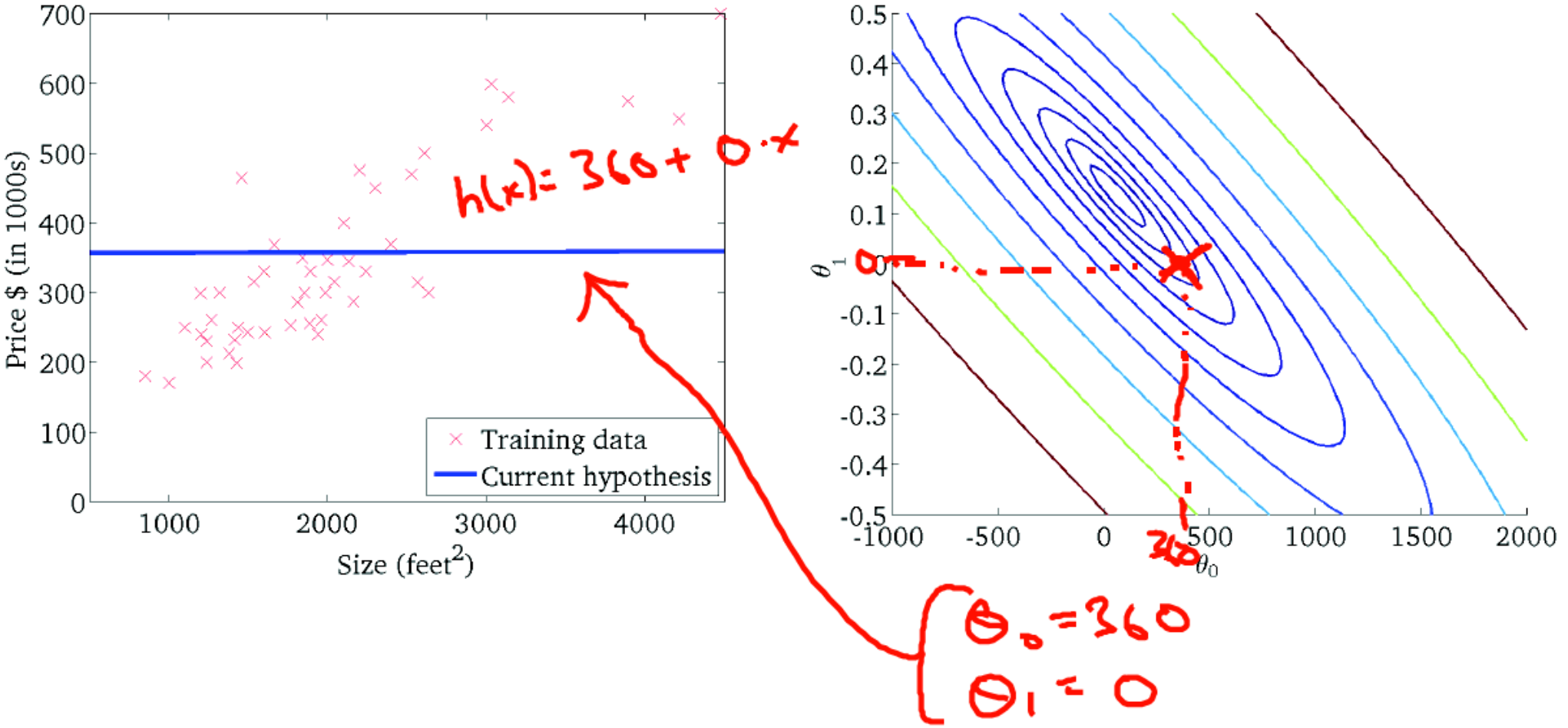
Linear Regression:Gradient Descent

$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)

$J(\theta_0, \theta_1)$

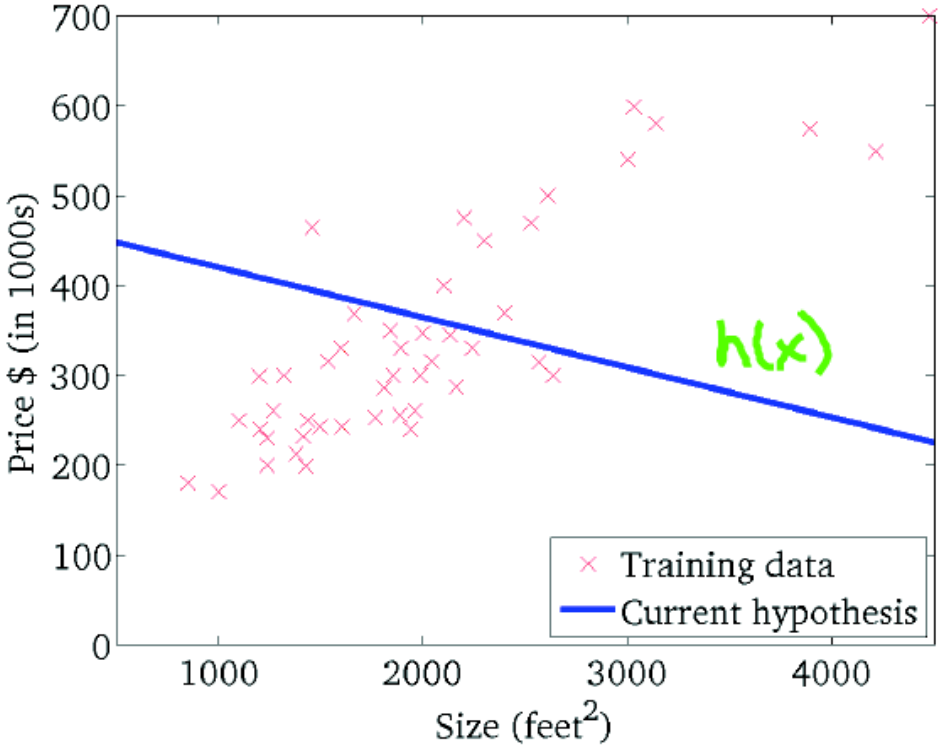
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent

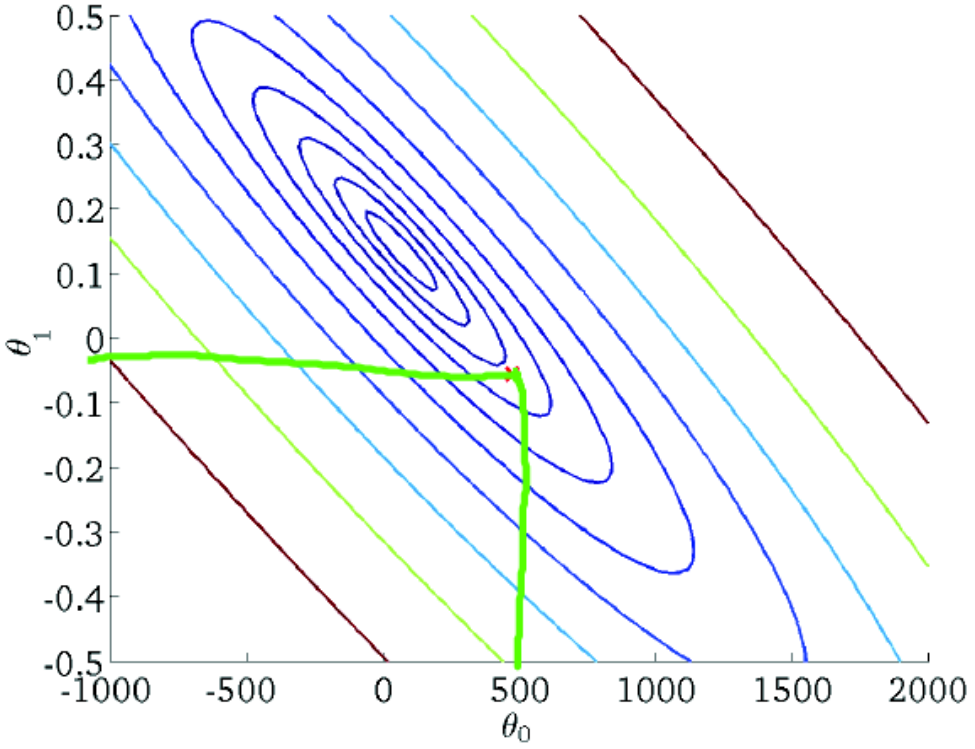
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

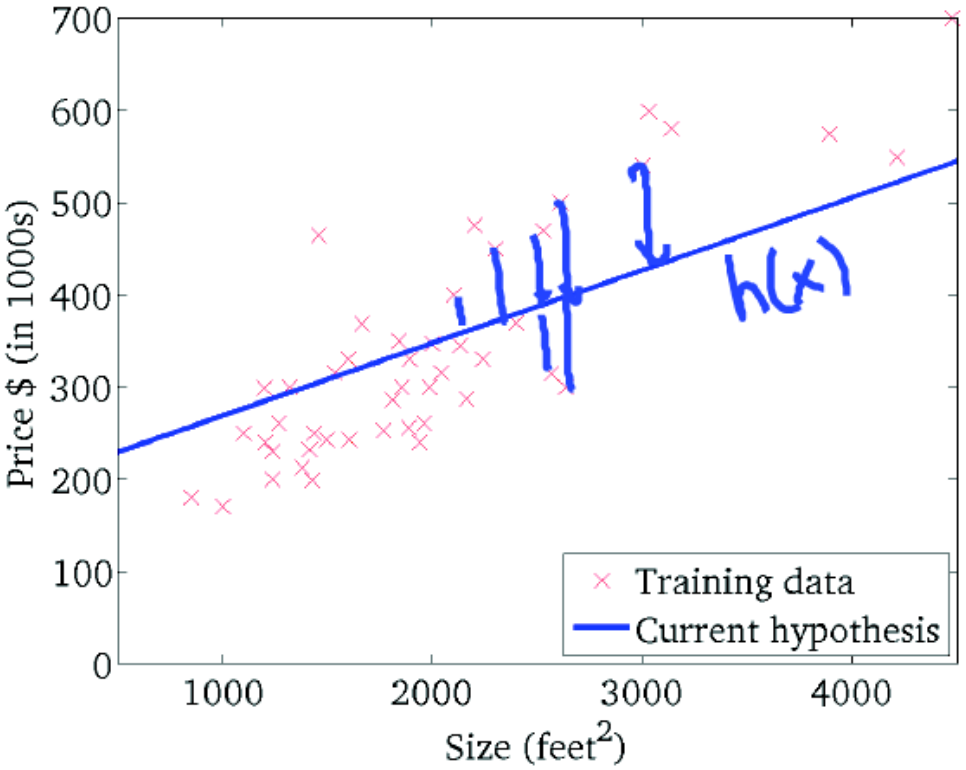
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent

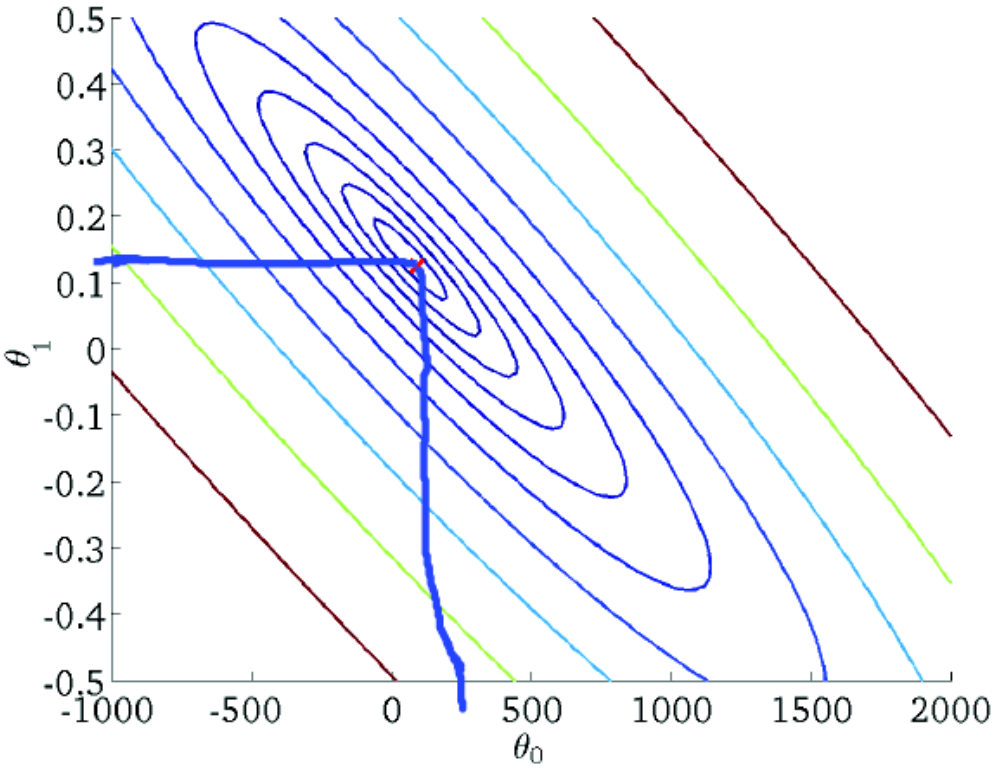
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent

Have some function $J(\theta_0, \theta_1)$

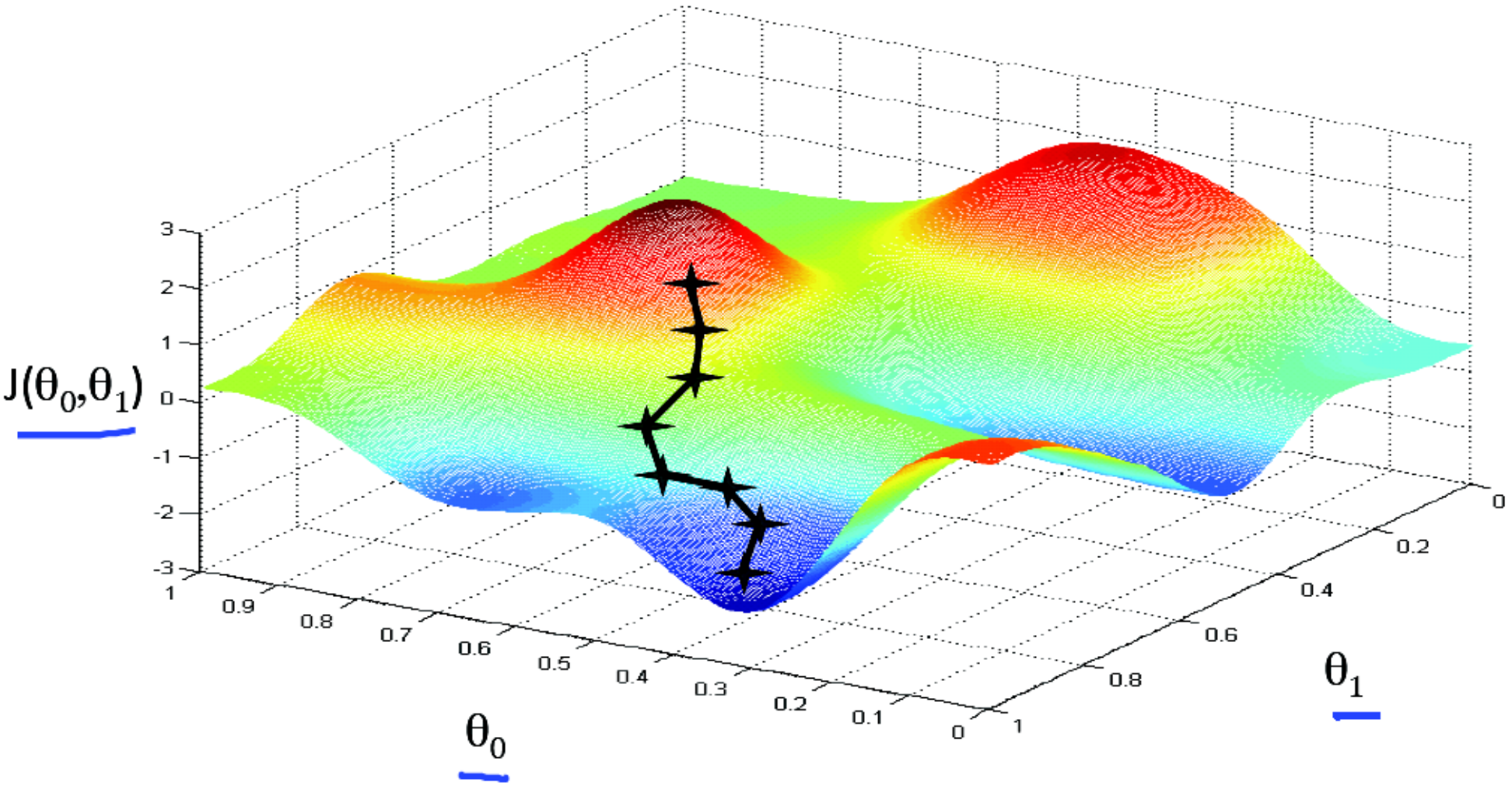
Want $\min_{\theta_0,\theta_1} J(\theta_0, \theta_1)$

Outline:

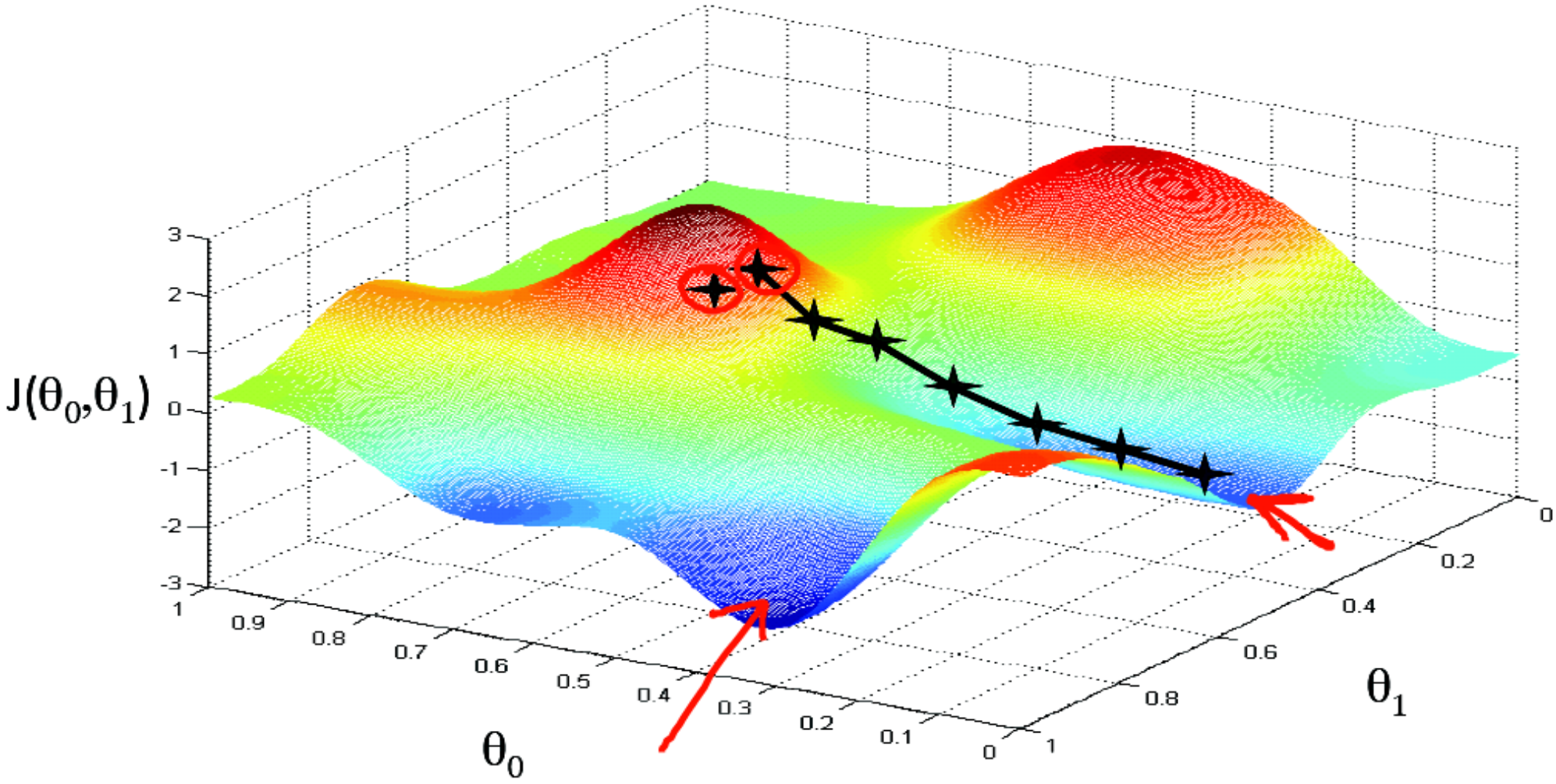
- Start with some θ_0, θ_1
- Keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$

  until we hopefully end up at a minimum

Linear Regression:Gradient Descent



Linear Regression:Gradient Descent



Linear Regression:Gradient Descent Algorithm

Gradient descent algorithm

repeat until convergence {

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

}

learning rate

Simultaneously update θ_0 and θ_1

Assignment

$a := b$

$a := a+1$

Truth assertion

$a = b$

$a = a+1$ ✗

Correct: Simultaneous update

$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$

$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$

$\theta_0 := \text{temp0}$

$\theta_1 := \text{temp1}$

Incorrect:

$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$

$\theta_0 := \text{temp0}$

$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$

$\theta_1 := \text{temp1}$

Linear Regression:Gradient Descent Algorithm

Gradient descent algorithm

repeat until convergence {

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

(simultaneously update $j = 0$ and $j = 1$)

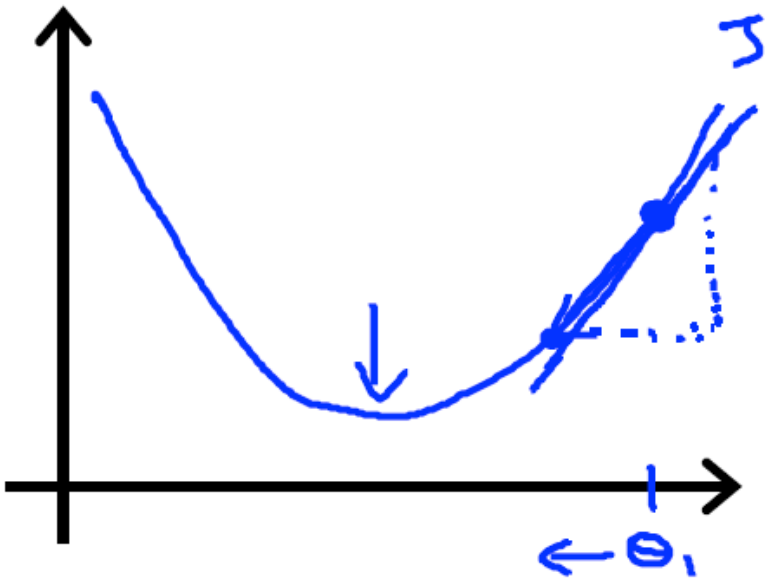
}

learning rate

derivative

$\min_{\theta_1} J(\theta_1)$ $\quad \theta_1 \in \mathbb{R}$.

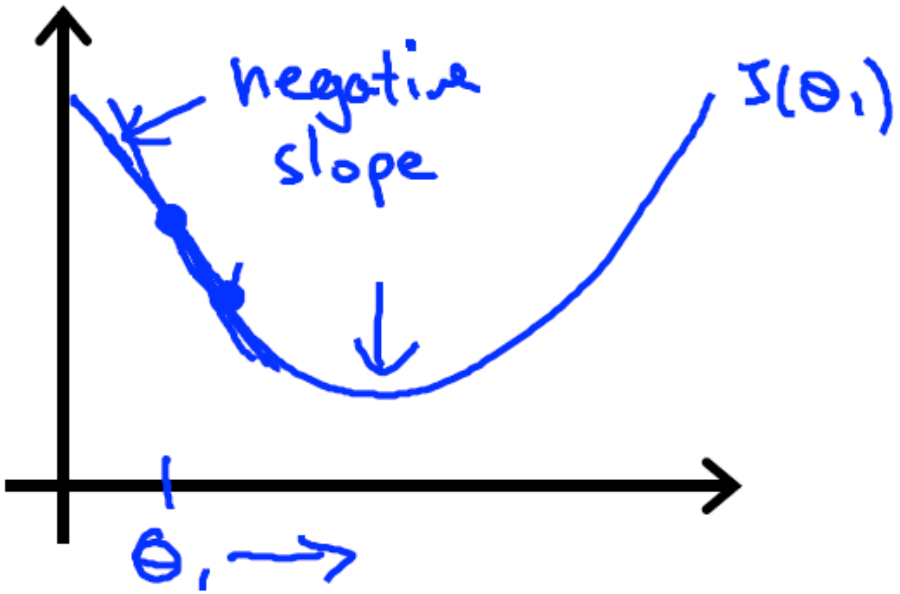
Linear Regression:Gradient Descent Algorithm



$(\theta_1 \in \mathbb{R})$

$$\theta_1 := \theta_1 - \alpha \underbrace{\frac{d}{d\theta_1} J(\theta_1)}_{\geq 0}$$

$$\theta_1 := \theta_1 - \alpha \cdot (\text{positive number})$$



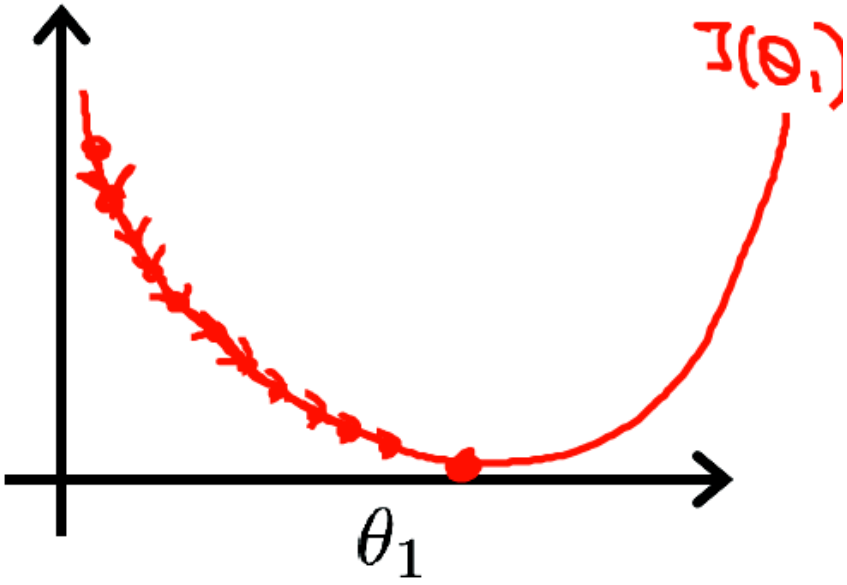
$$\underbrace{\frac{d}{d\theta_1} J(\theta_1)}_{\leq 0}$$

$$\theta_1 := \theta_1 - \alpha \, (\text{negative number})$$

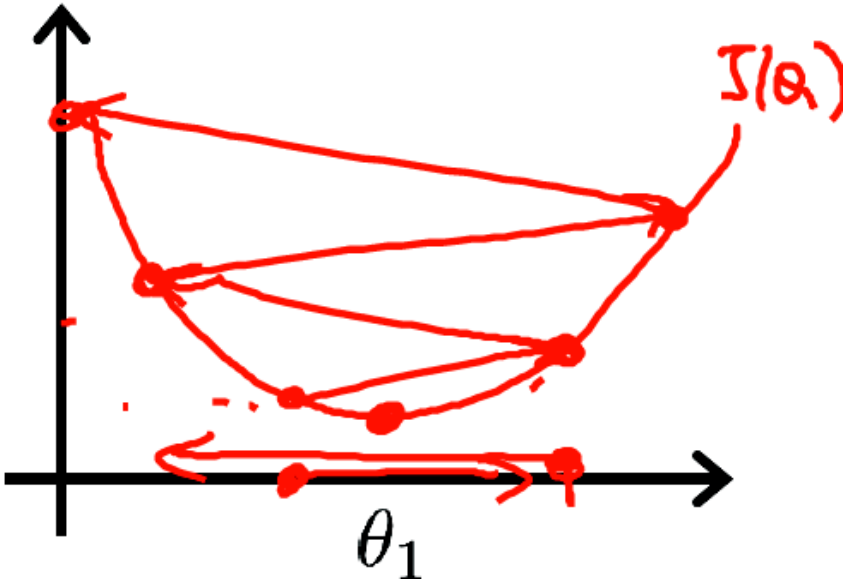
Linear Regression:Gradient Descent Algorithm

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

If α is too small, gradient descent can be slow.



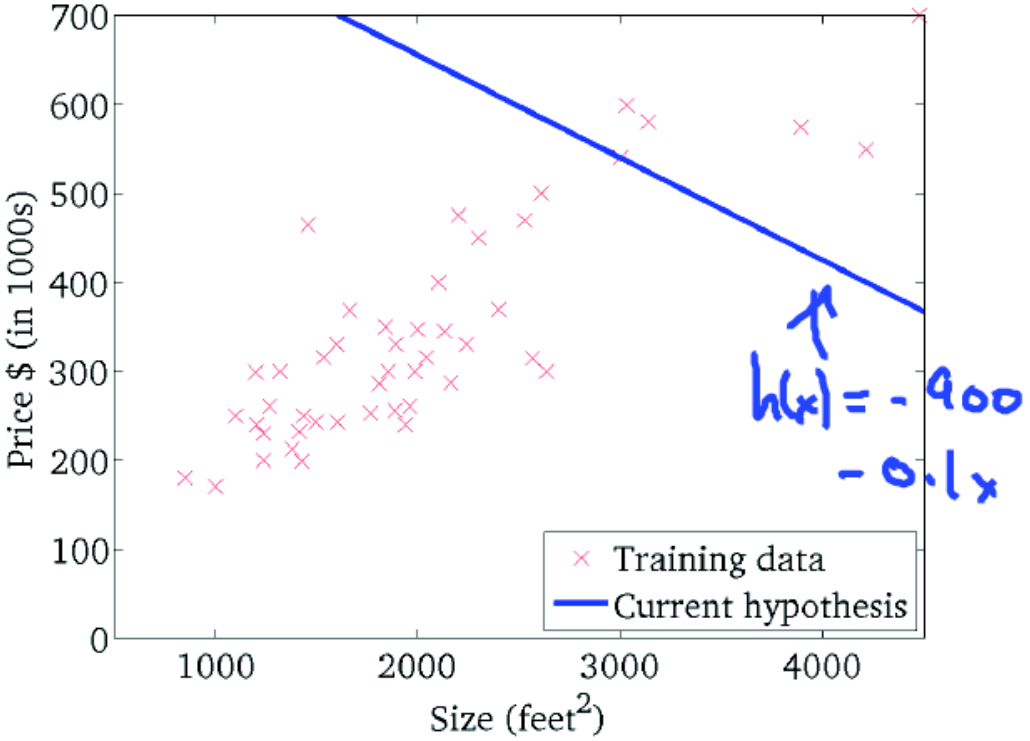
If α is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.



Linear Regression:Gradient Descent Algorithm:intuitively

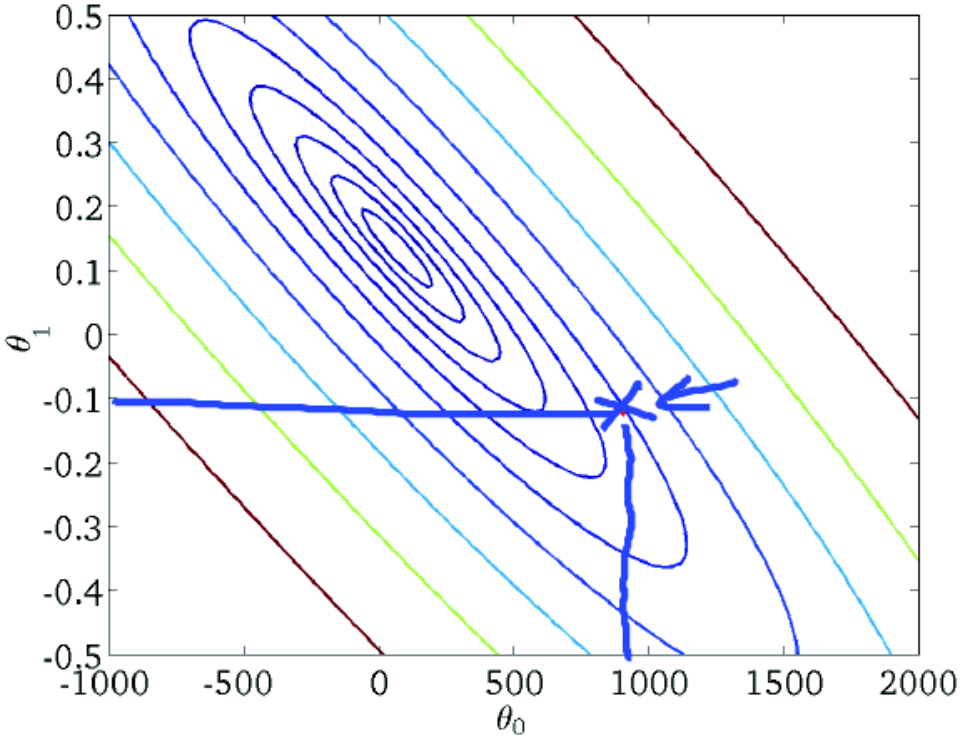
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

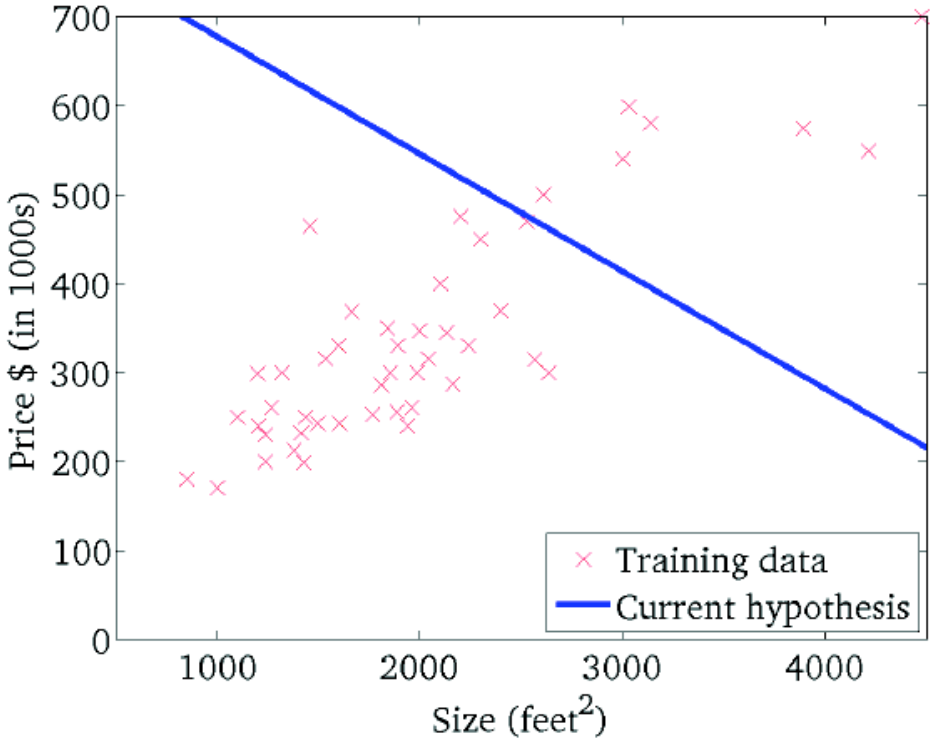
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

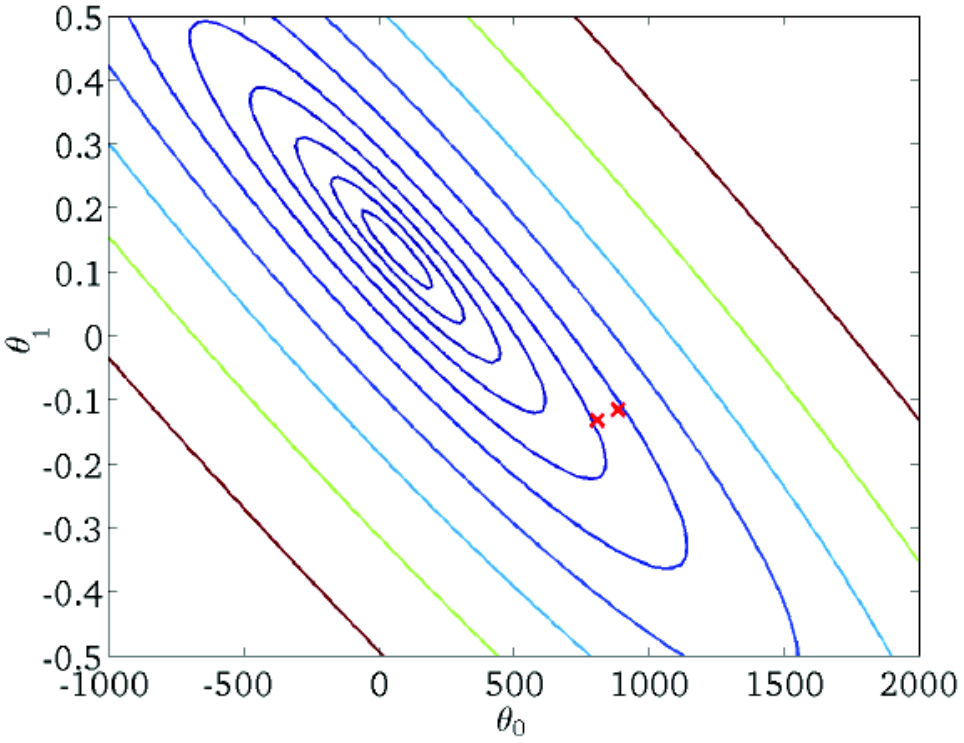
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

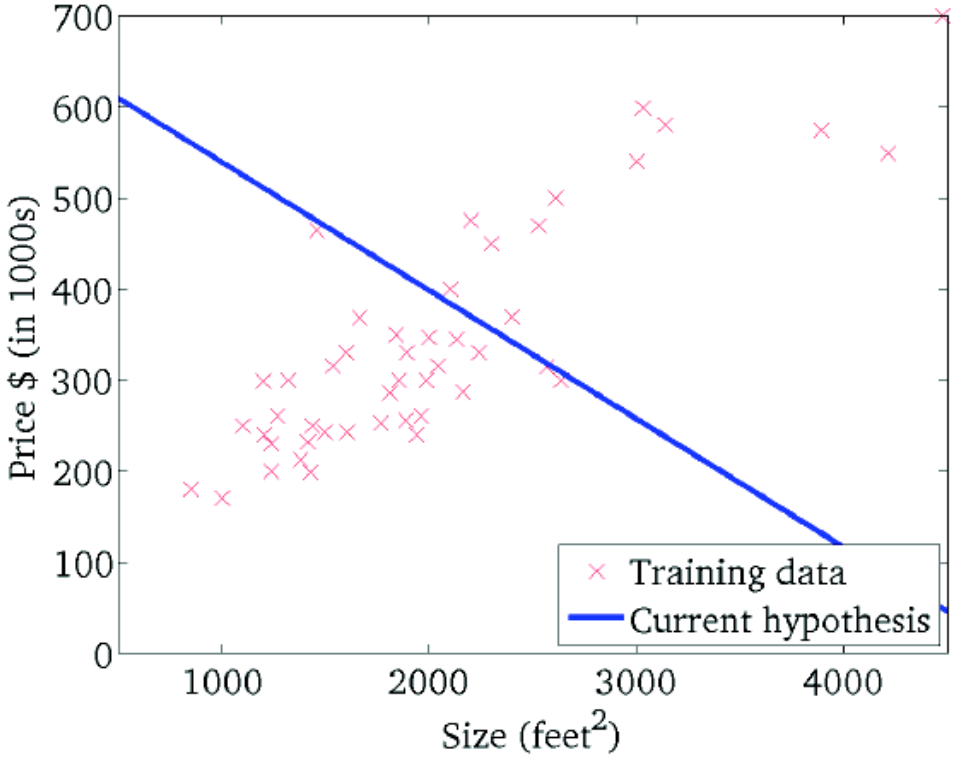
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

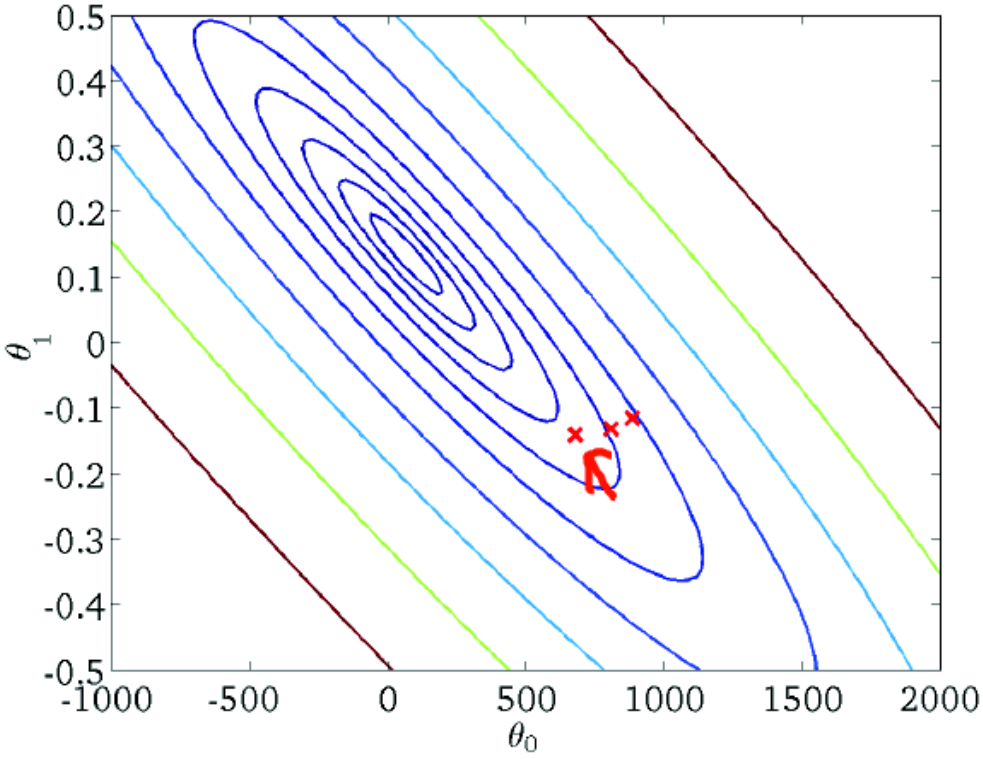
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

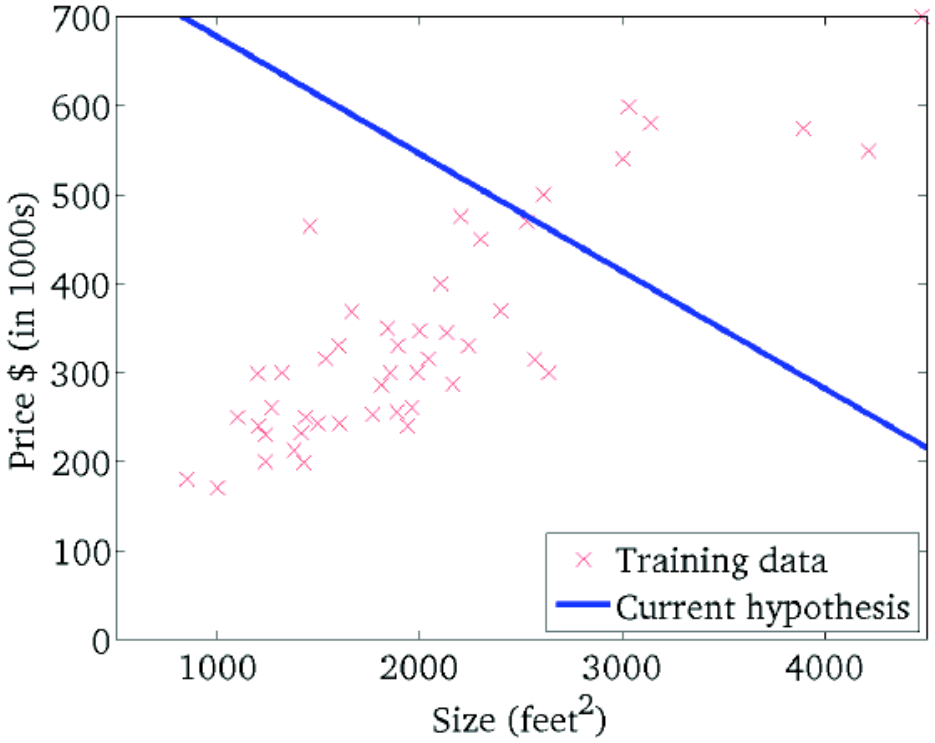
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

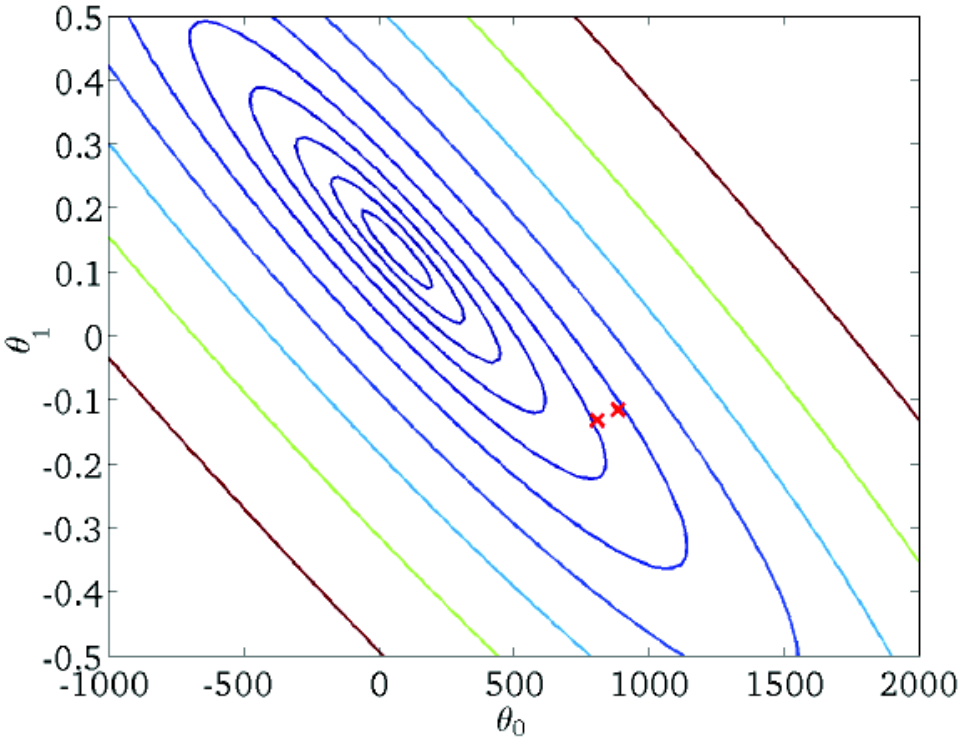
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

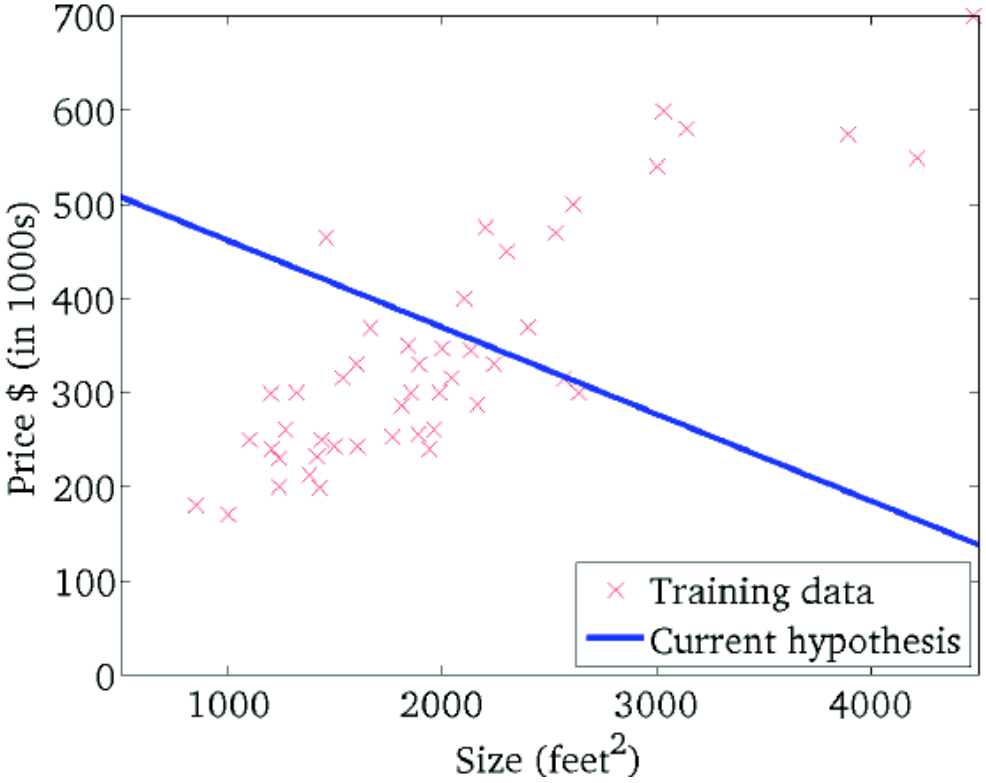
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

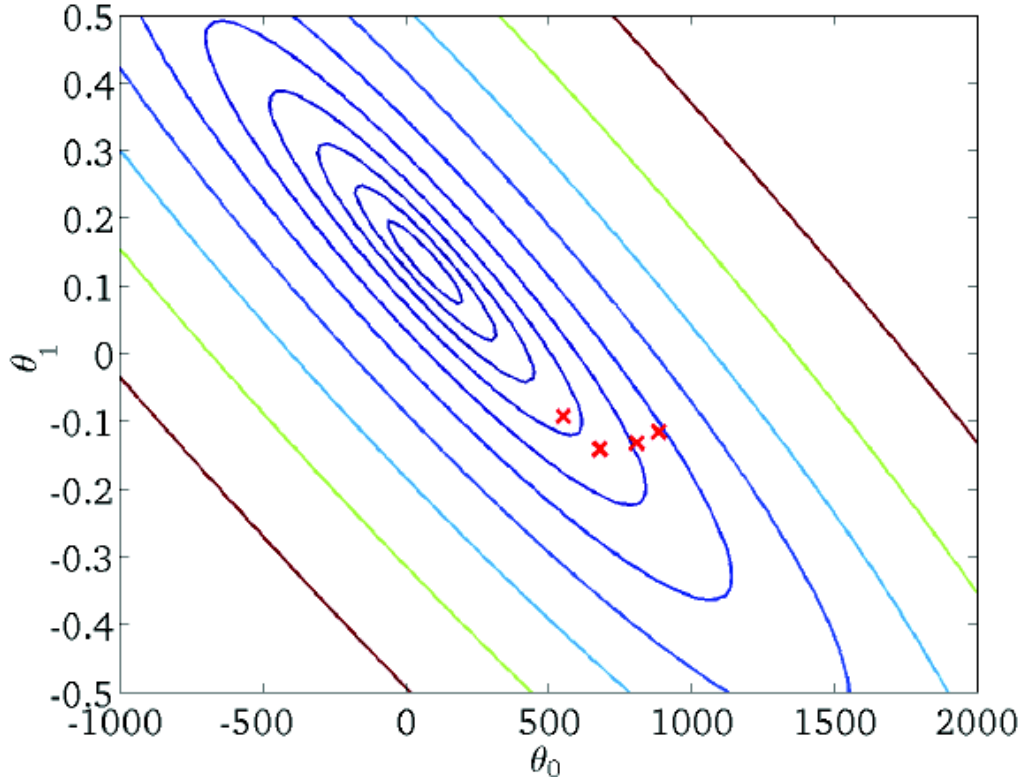
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

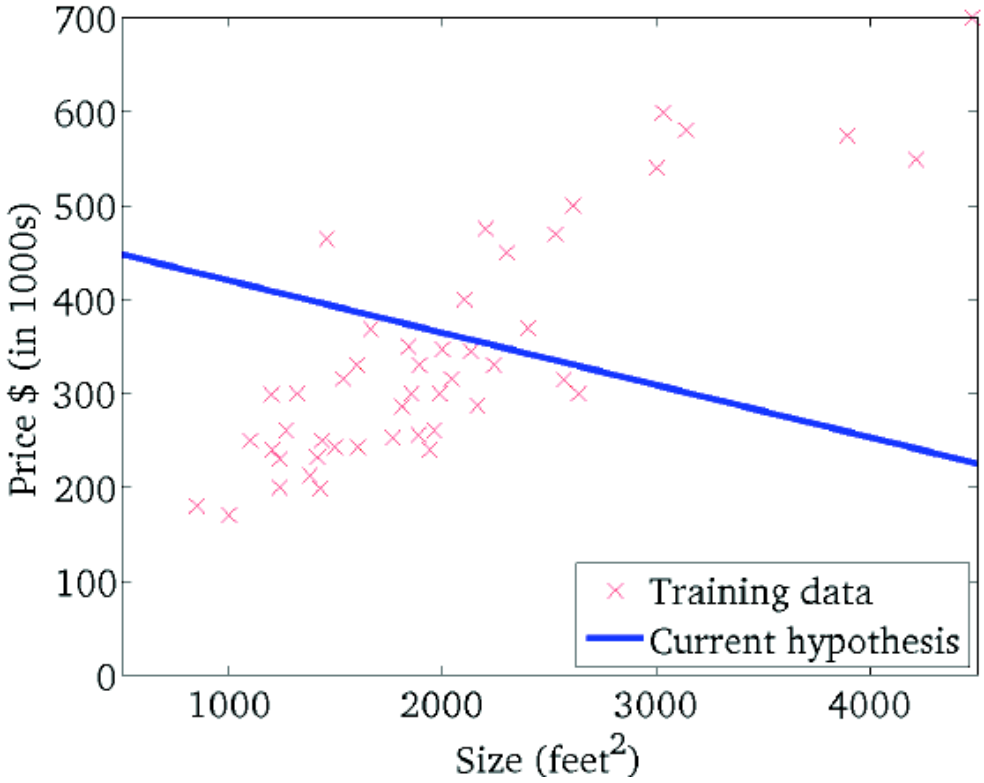
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

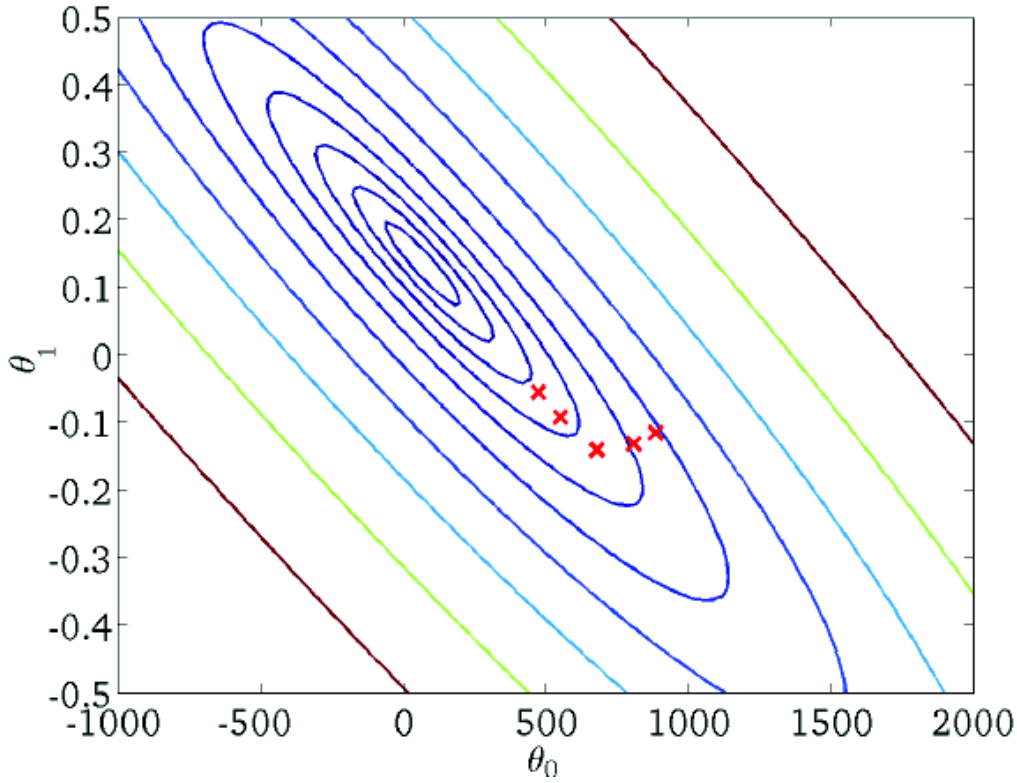
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

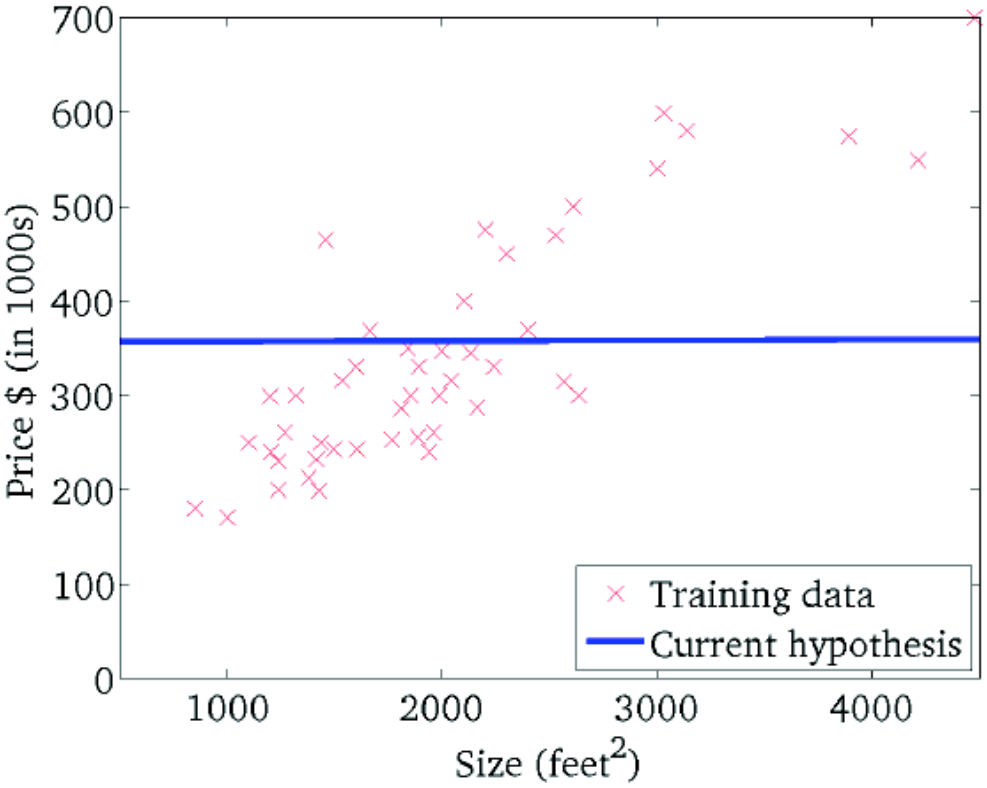
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

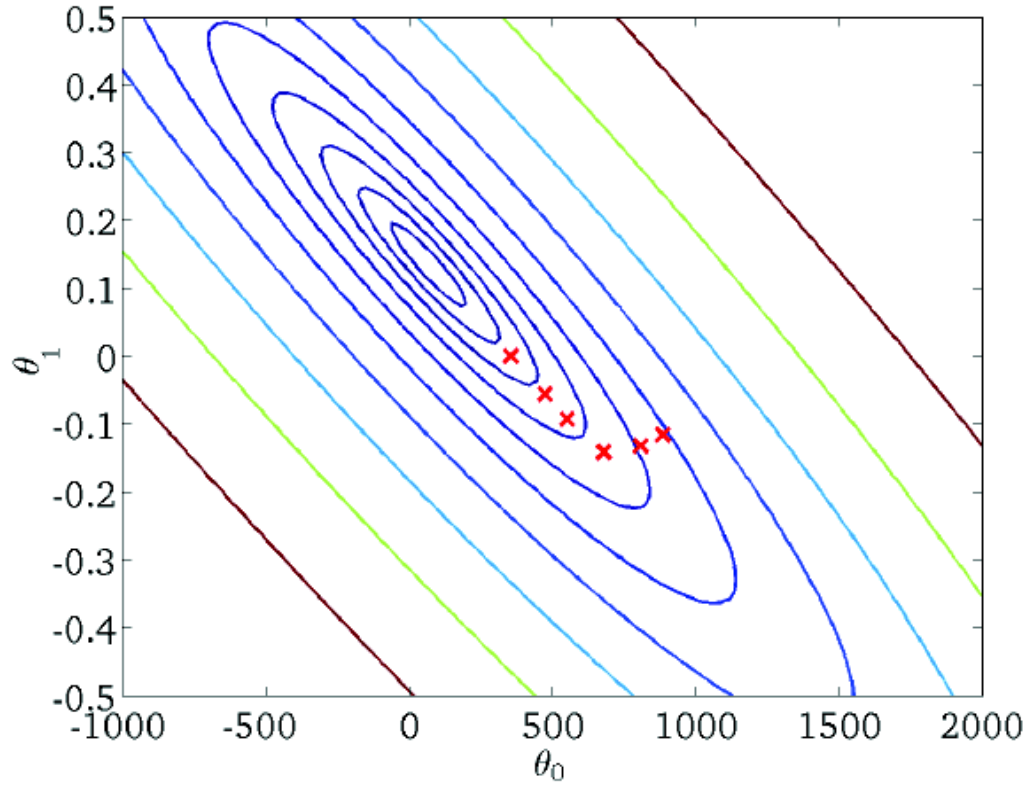
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

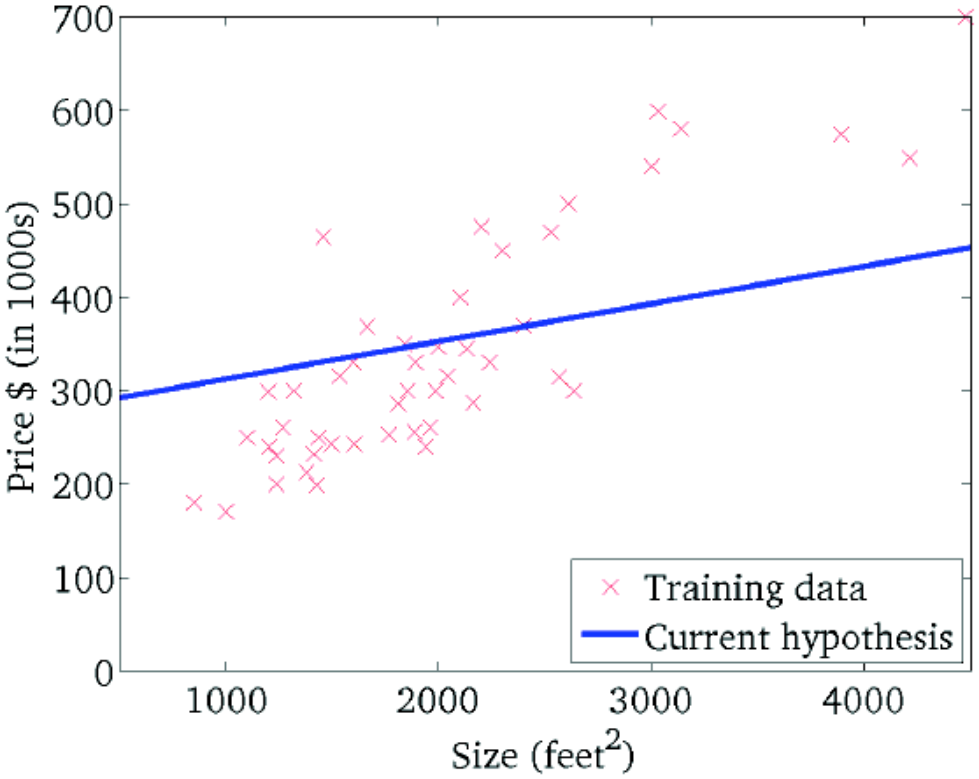
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

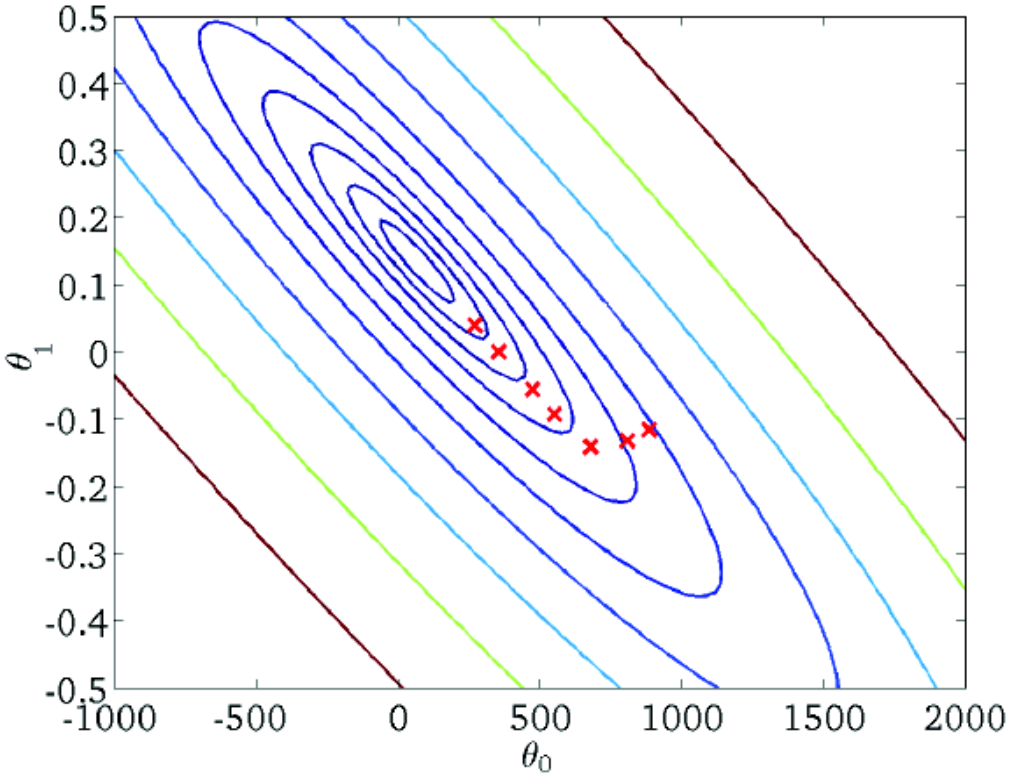
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

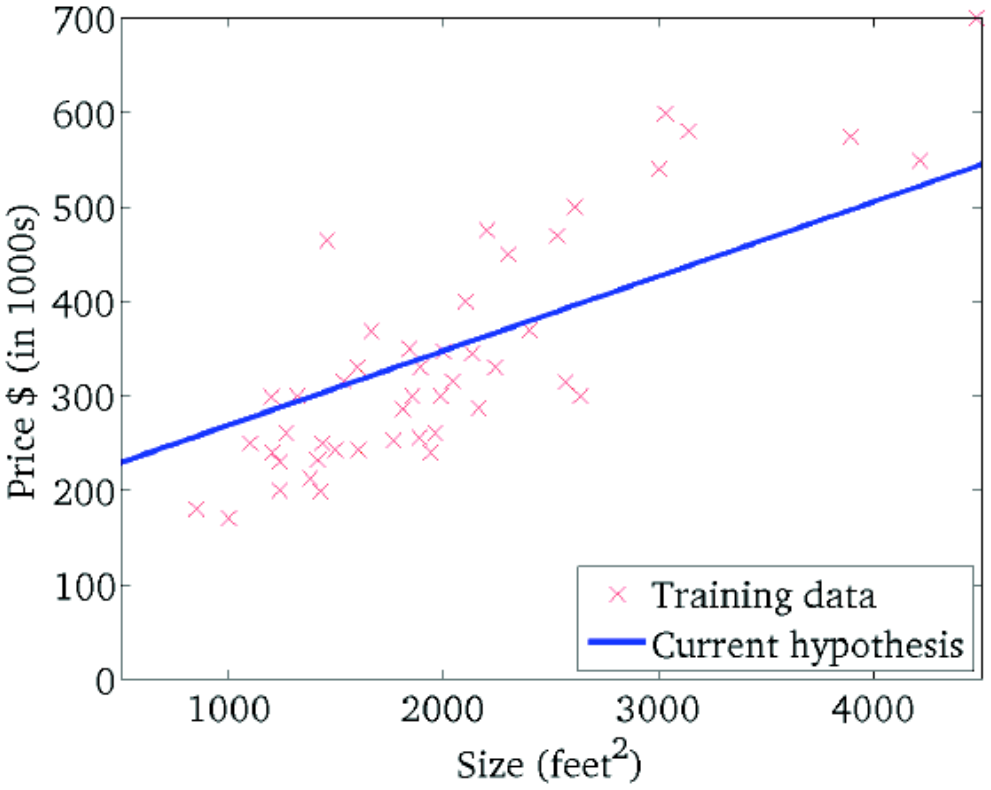
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

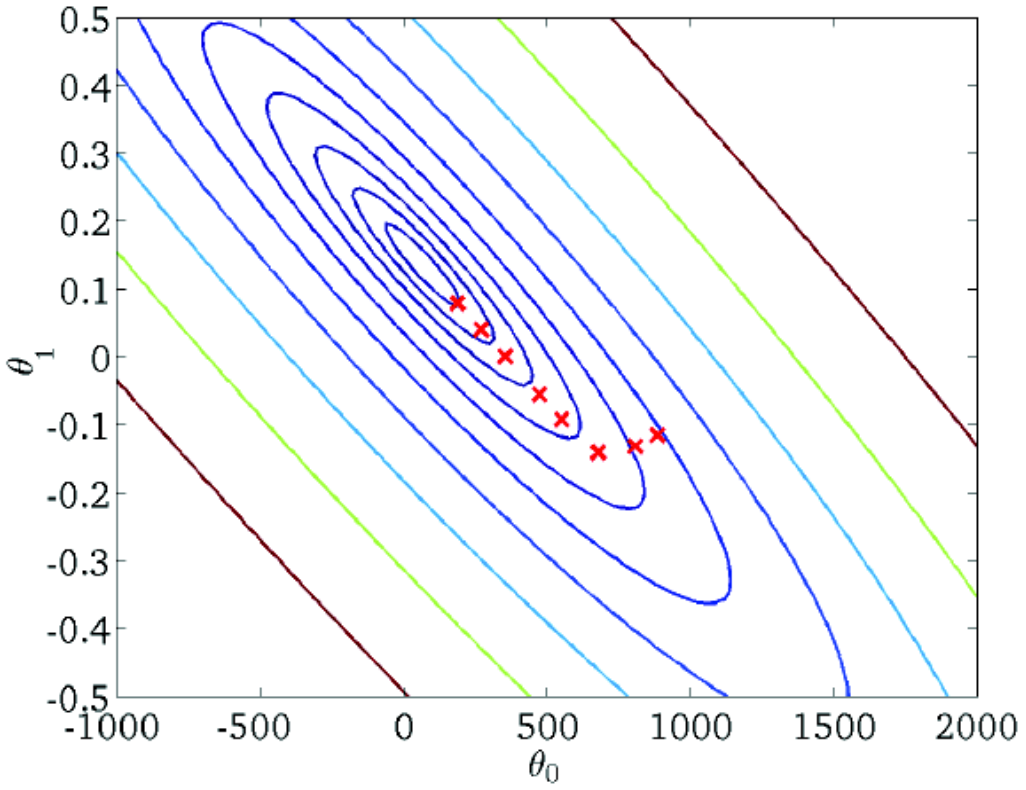
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

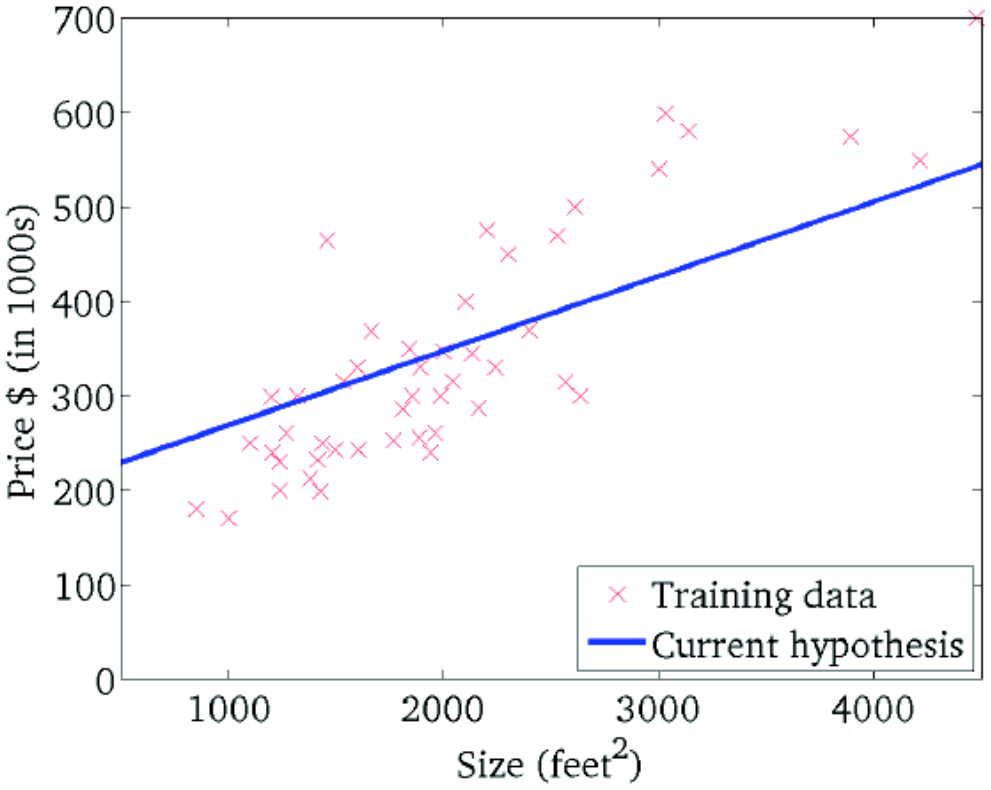
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

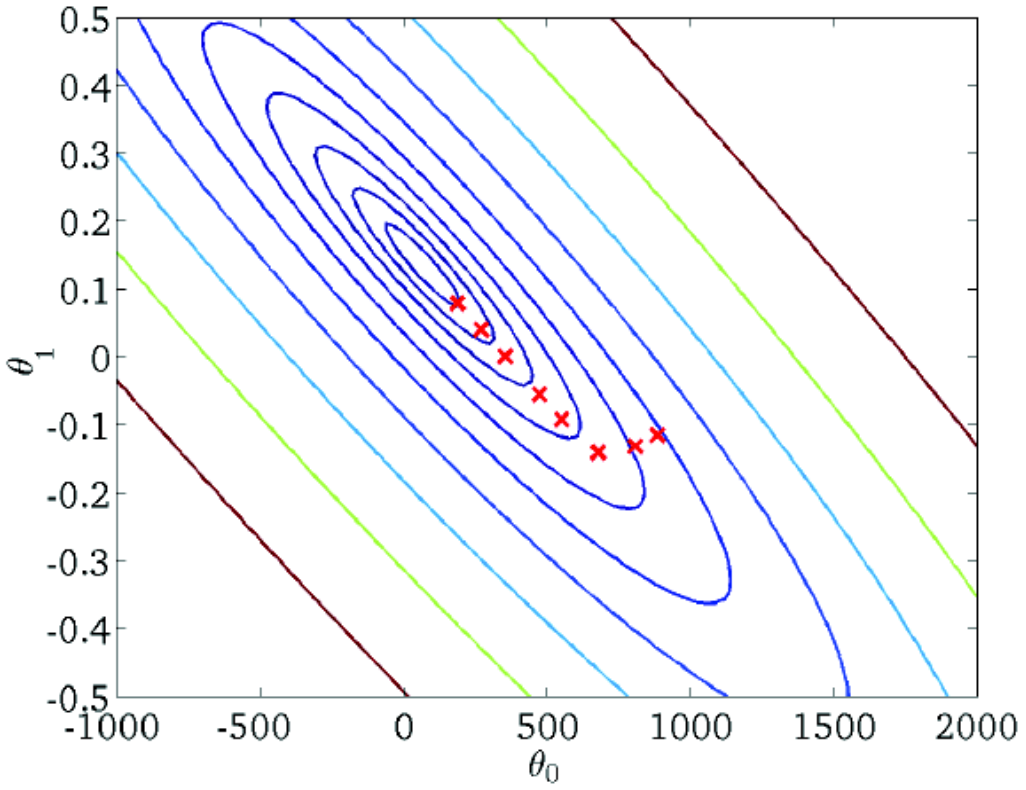
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

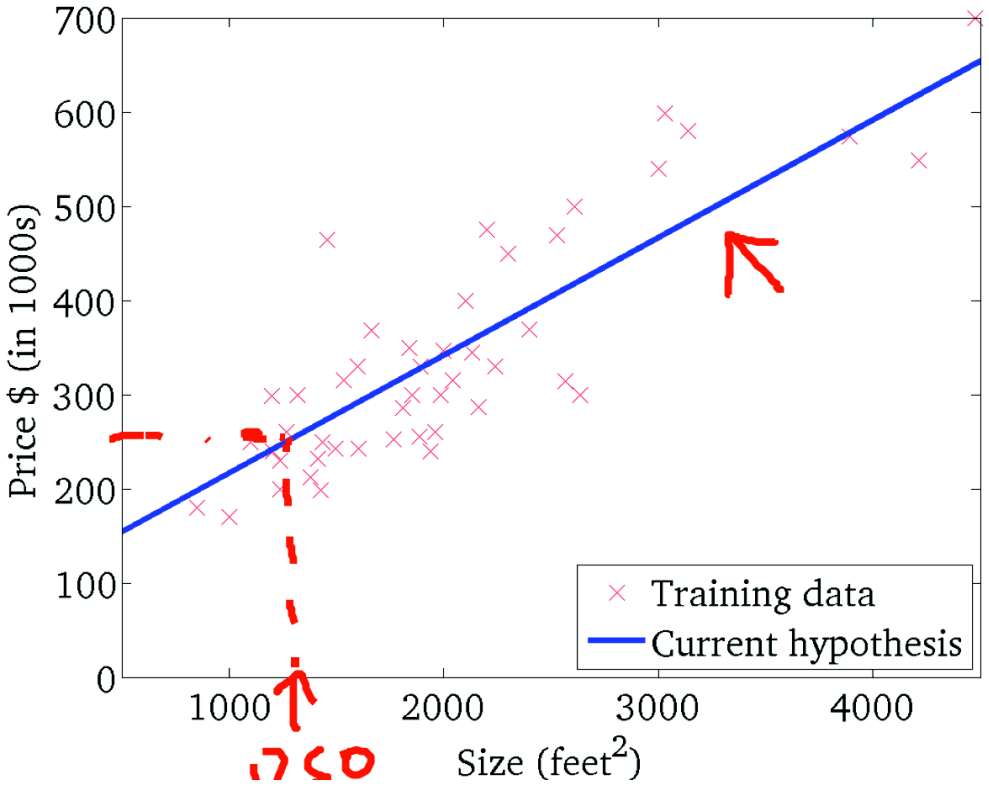
(function of the parameters θ_0, θ_1)



Linear Regression:Gradient Descent Algorithm:intuitively

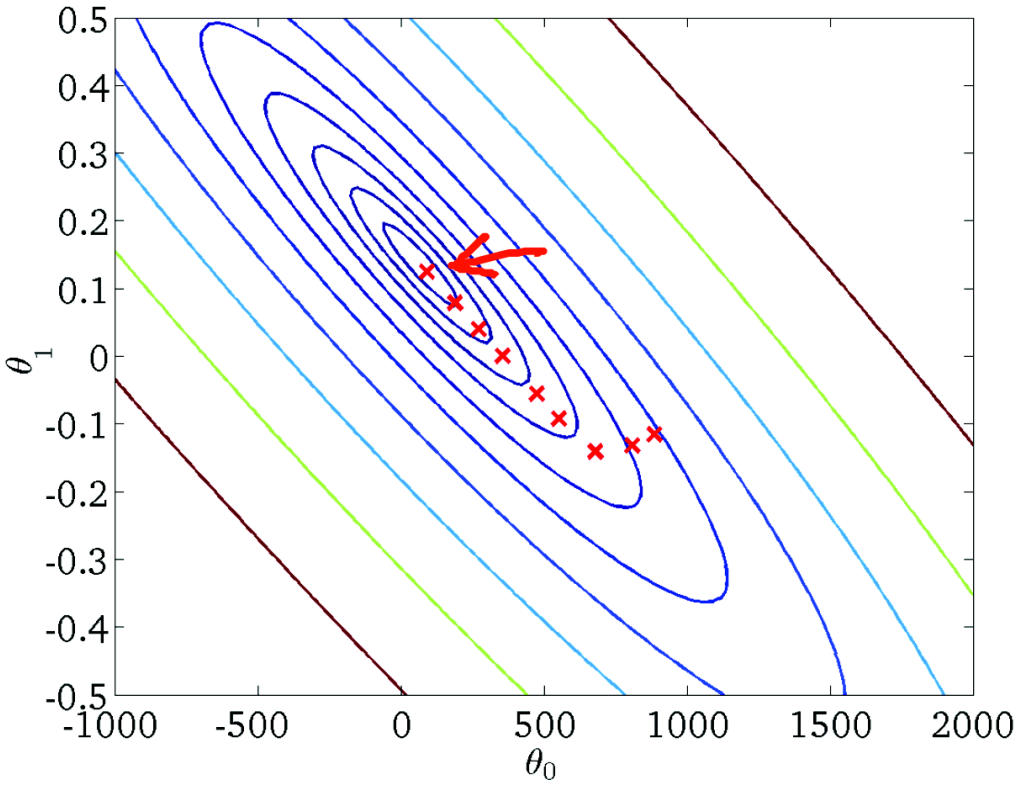
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



Logistic Regression

- Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression.

Logistic Regression

Classification

- Email: Spam / Not Spam?
- Online Transactions: Fraudulent (Yes / No)?
- Tumor: Malignant / Benign ?

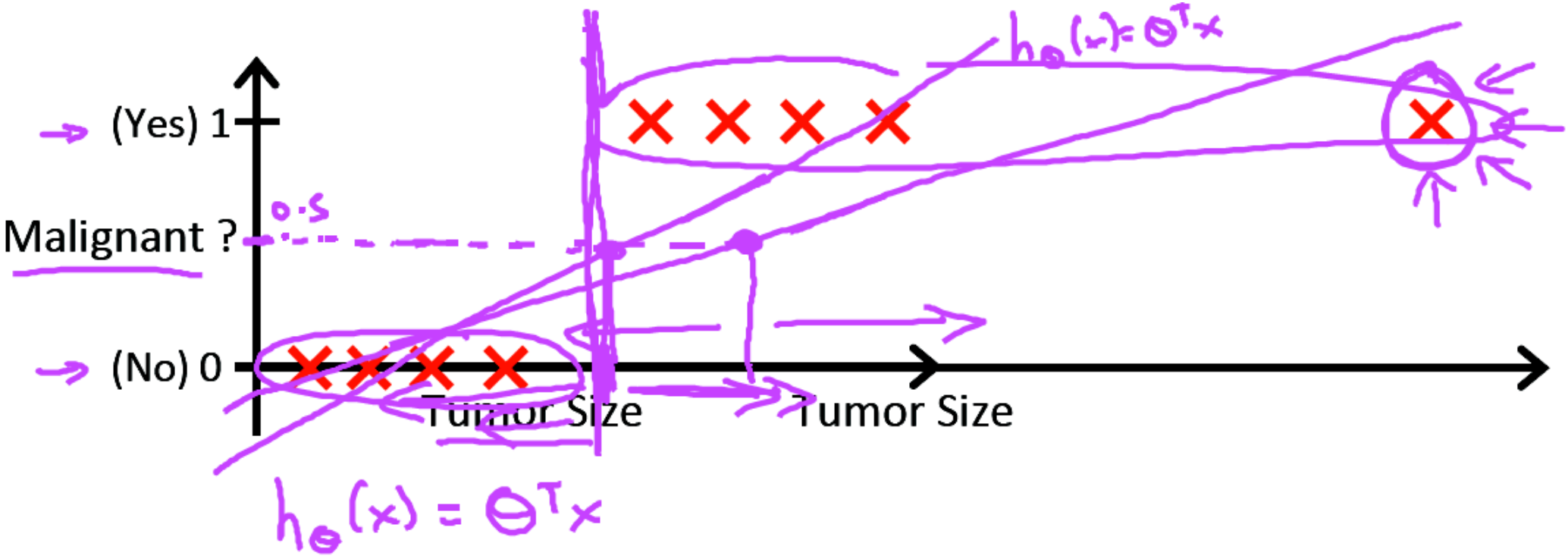
$y \in \{0, 1\}$

0: "Negative Class" (e.g., benign tumor)

1: "Positive Class" (e.g., malignant tumor)

$y \in \{0, 1, 2, 3\}$

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

Classification: $y = 0$ or 1

$h_\theta(x)$ can be > 1 or < 0

Logistic Regression: $0 \leq h_\theta(x) \leq 1$

Classification

Logistic Regression

Logistic Regression Model

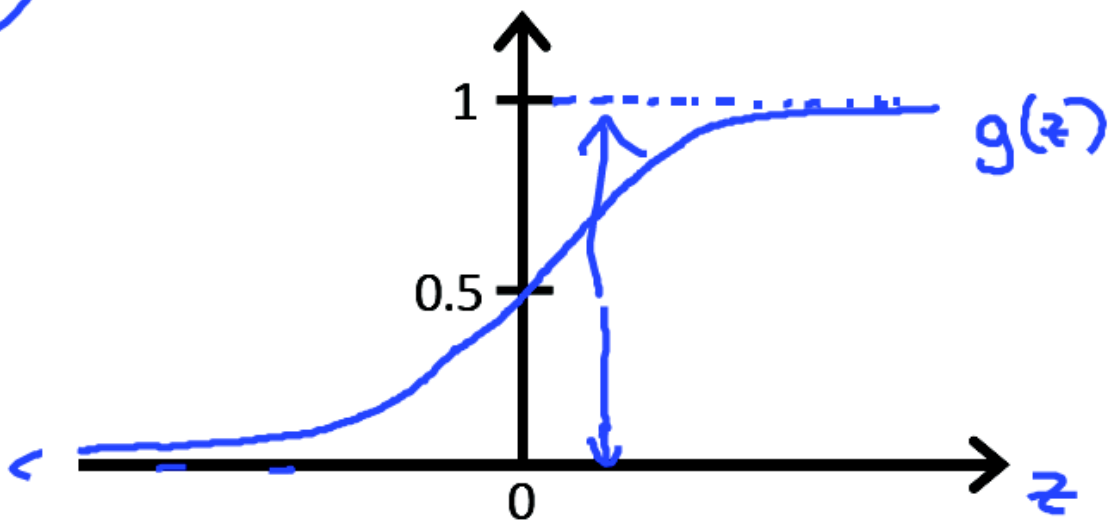
Want $0 \leq h_\theta(x) \leq 1$

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

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

$z = \theta^T x$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$



Parameters θ.

- Sigmoid function
- Logistic function

Logistic Regression

Logistic Regression Model

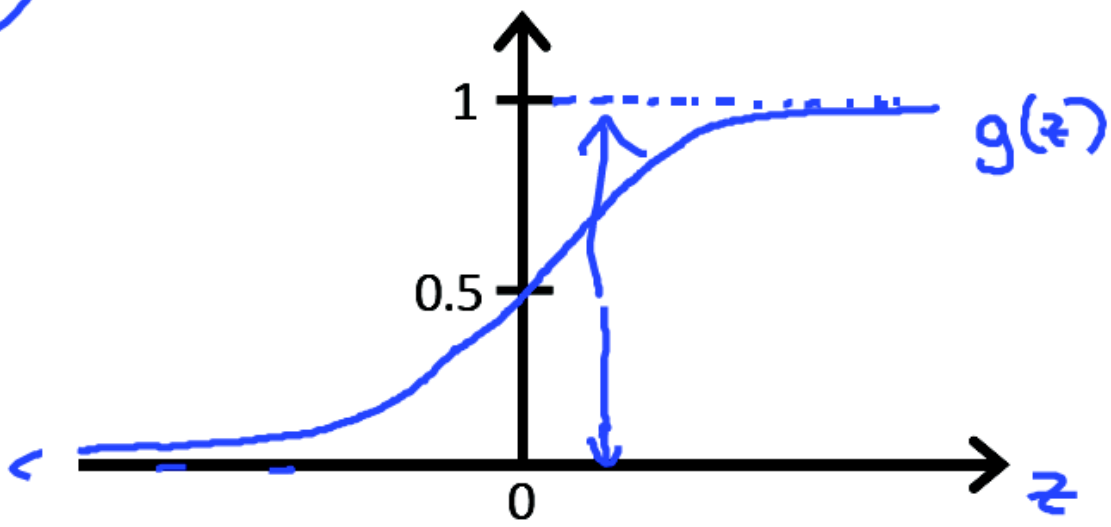
Want $0 \leq h_\theta(x) \leq 1$

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

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

$z = \theta^T x$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

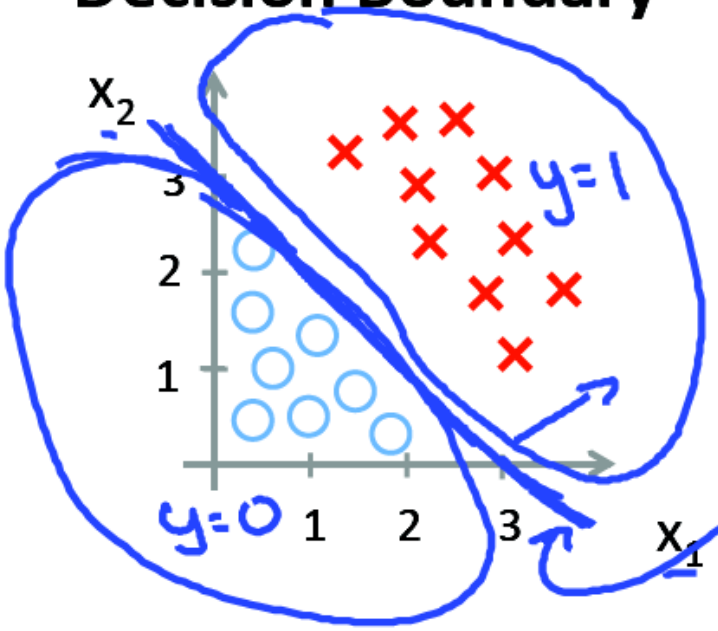


Parameters θ.

- Sigmoid function
- Logistic function

Logistic Regression

Decision Boundary



$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$\theta_0 = -3$, $\theta_1 = 1$, $\theta_2 = 1$

Decision boundary

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$\theta^T x$

$x_1 + x_2 \geq 3$

x_1, x_2

$h_\theta(x) = 0.5$

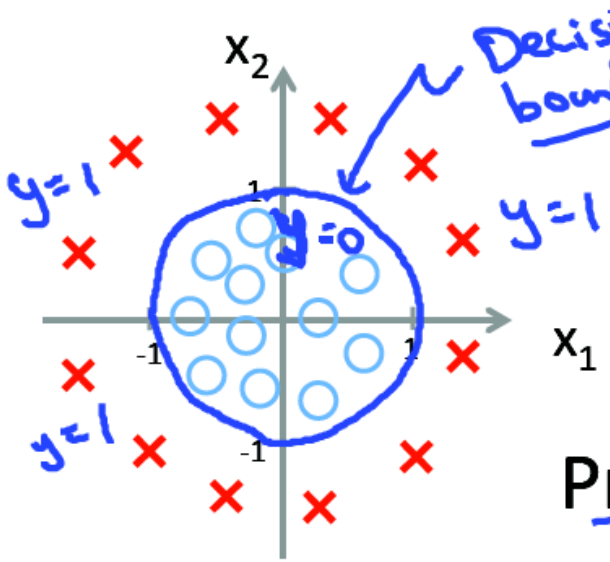
$x_1 + x_2 = 3$

$x_1 + x_2 < 3$

$y = 0$

Logistic Regression

Non-linear decision boundaries



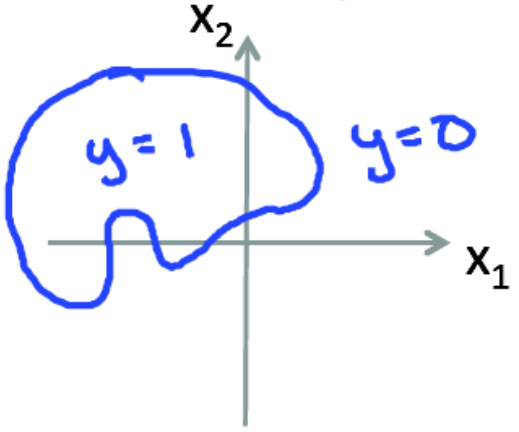
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$x_1^2 + x_2^2 = 1$

$x_1^2 + x_2^2 \geq 1$



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \ldots)$$

Logistic Regression

Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

m examples $x \in \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix}$ $\mathbb{R}^{n+1}$ $x_0 = 1, y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

How to choose parameters θ ?

Logistic Regression

Cost function

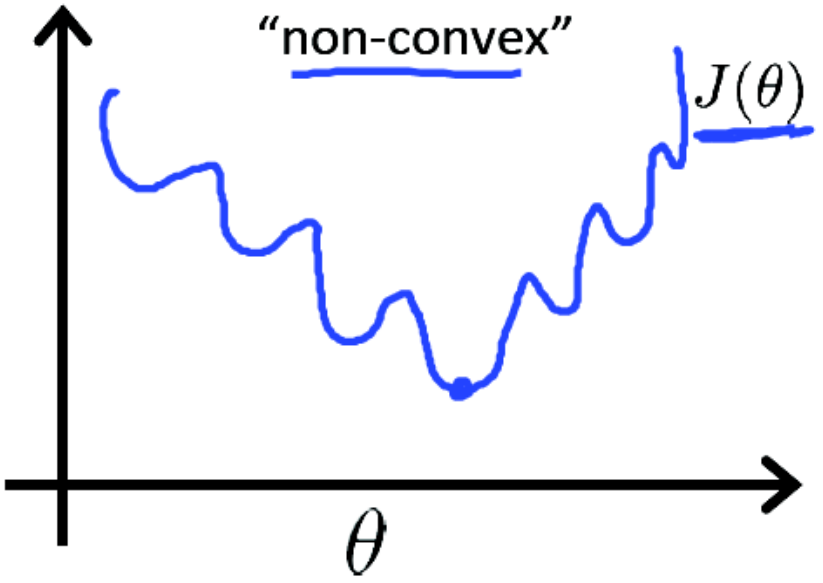
→ ~~Linear~~ regression: $J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

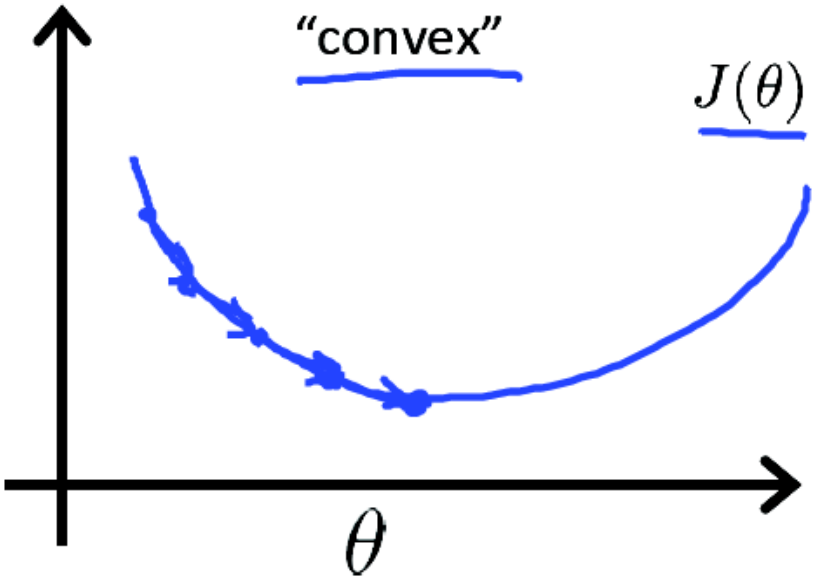
Logistic

$\rightarrow \text{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

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

$\frac{1}{1 + e^{-\theta^T x}}$

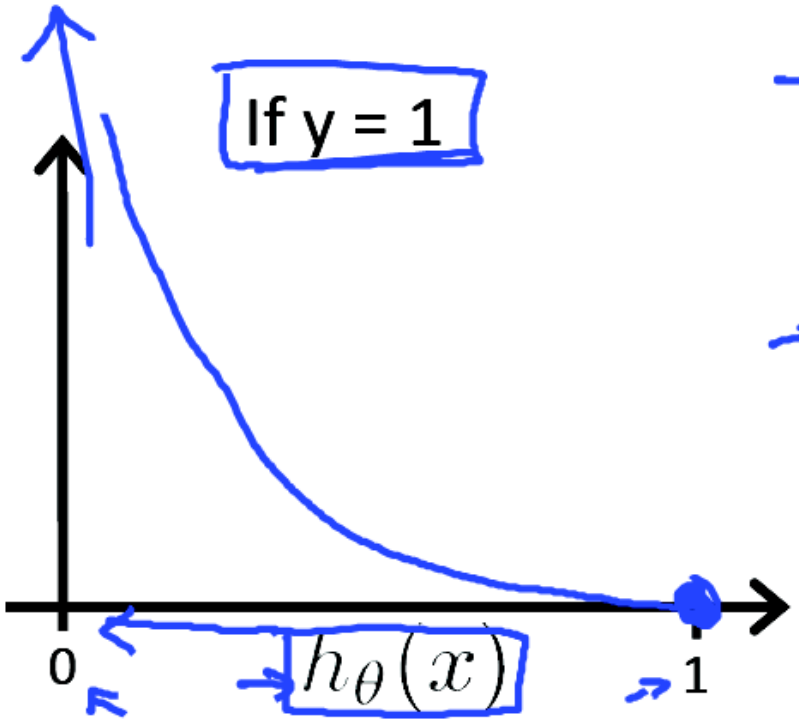




Logistic Regression

Logistic regression cost function

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



→ $\text{Cost} = 0$ if $y = 1, h_\theta(x) = 1$

But as $h_\theta(x) \to 0$

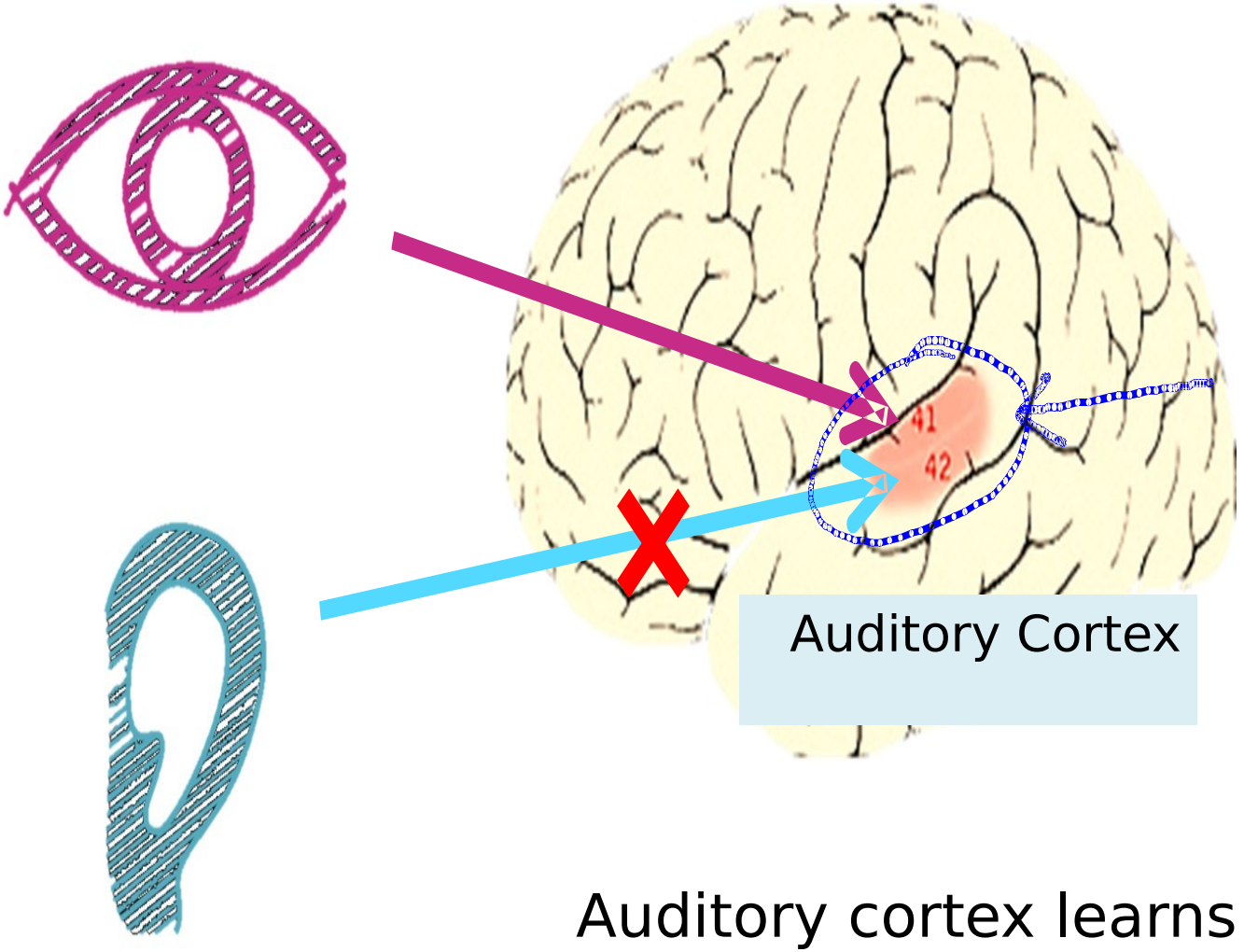
$Cost \to \infty$

→ Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x; \theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

Neural Networks

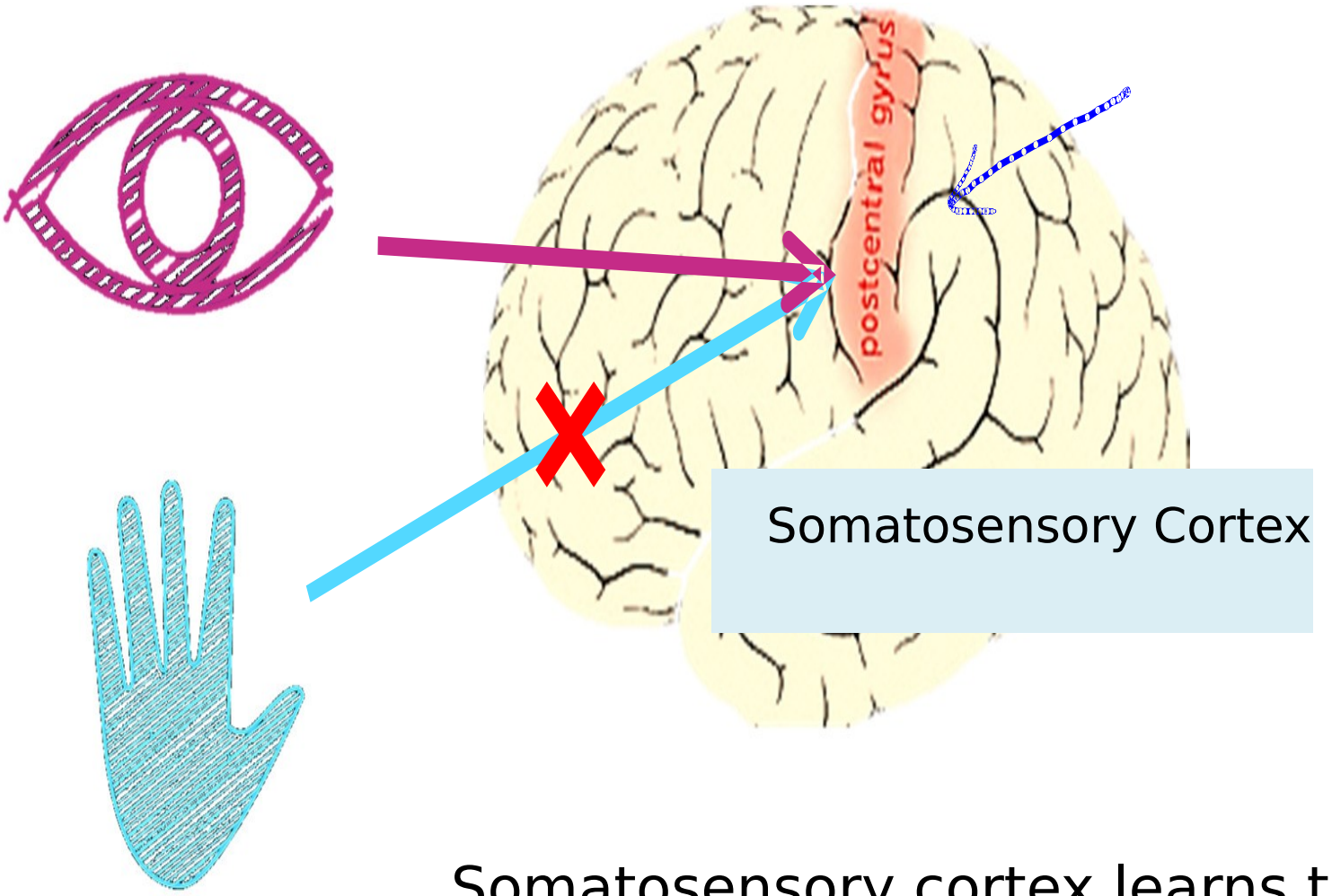
- Origins: Algorithms that try to mimic the brain.
- Was very widely used in 80s and early 90s; popularity diminished in late 90s.
- Recent resurgence: State-of-the-art technique for many applications

The “one learning algorithm” hypothesis



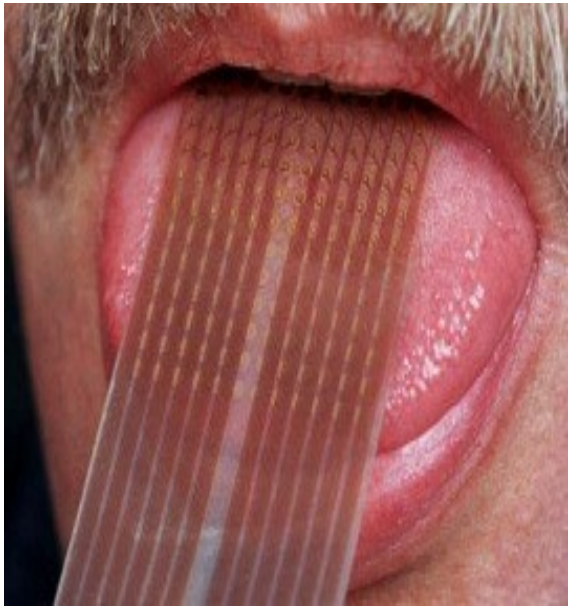
Auditory cortex learns to see

The “one learning algorithm” hypothesis



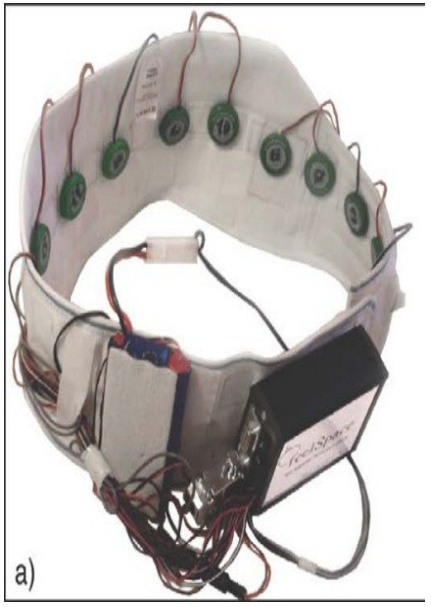
Somatosensory cortex learns to see

Sensor representations in the brain

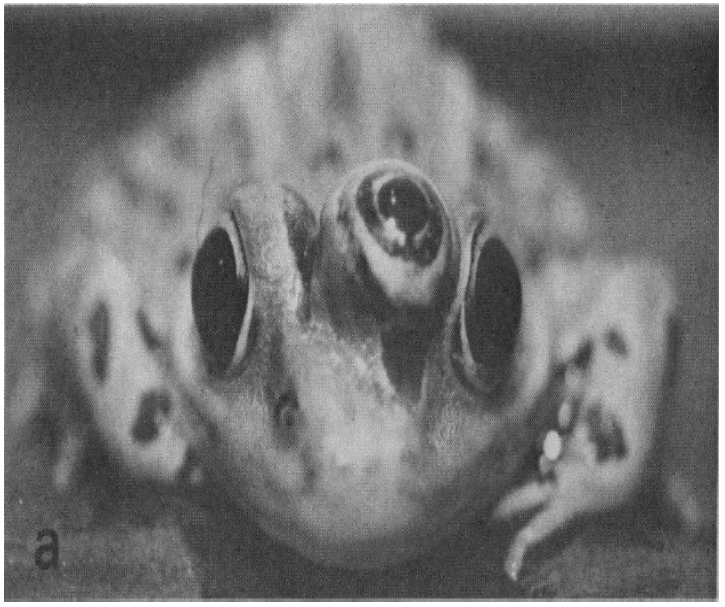


Seeing with your tongue

Human echolocation (sonar)

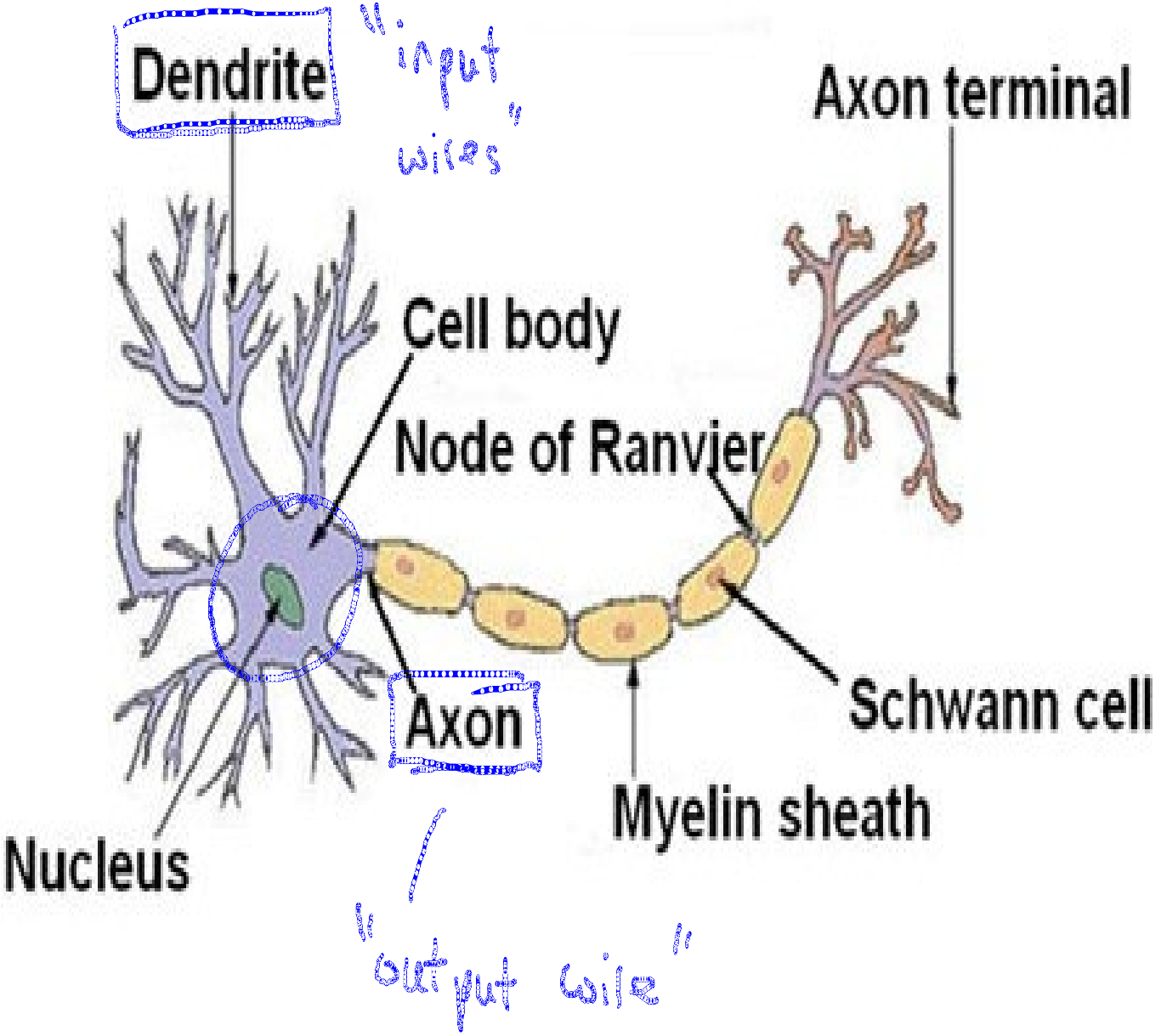


Haptic belt: Direction sense

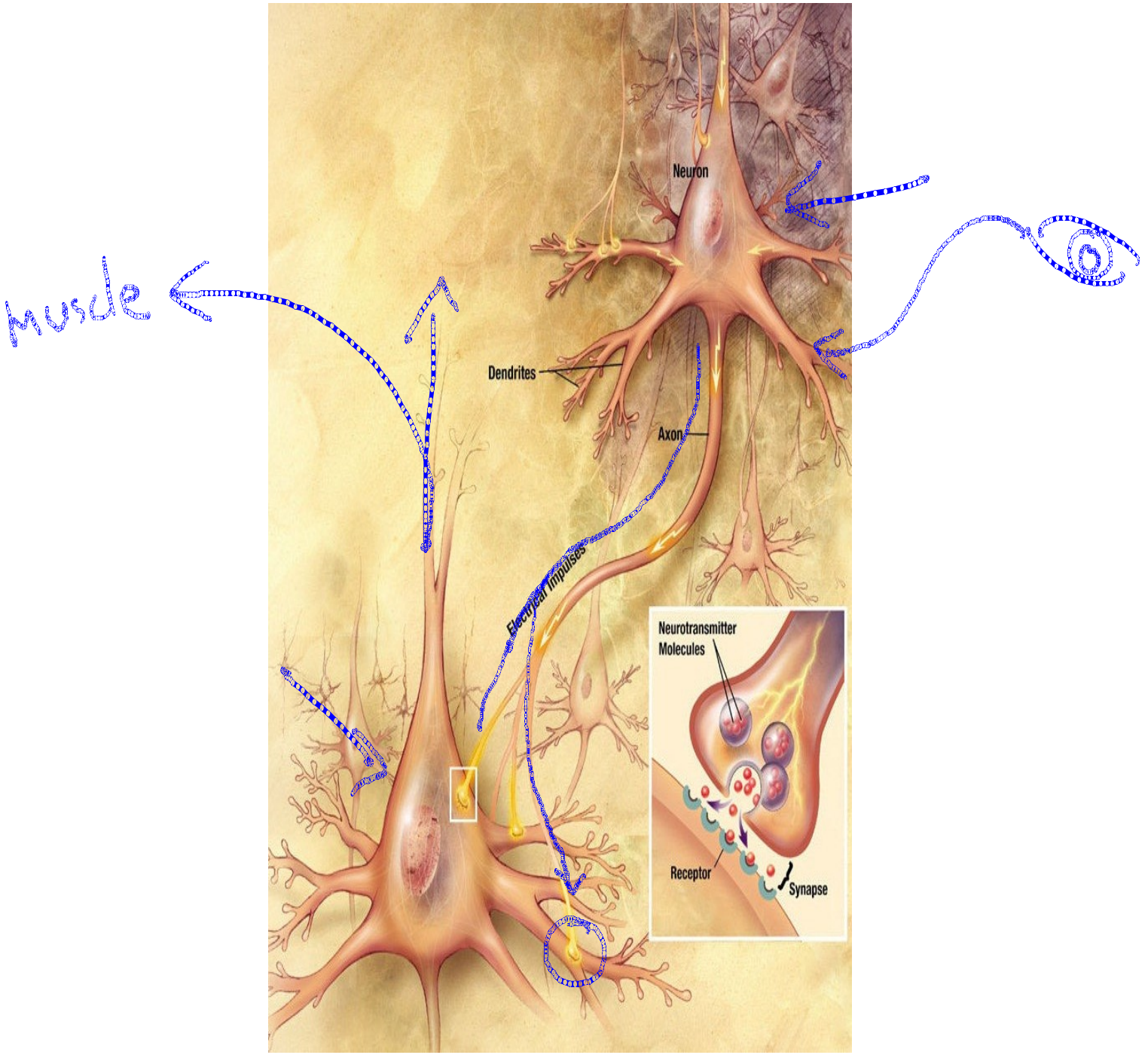


Implanting a 3rd eye

Neuron in the brain



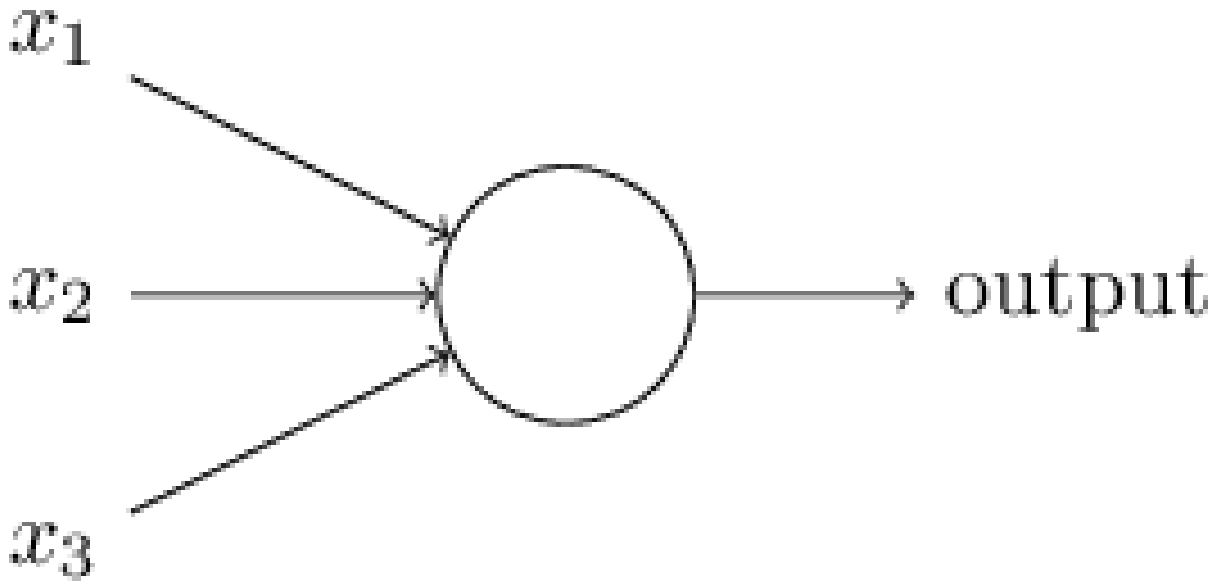
Neurons in the brain



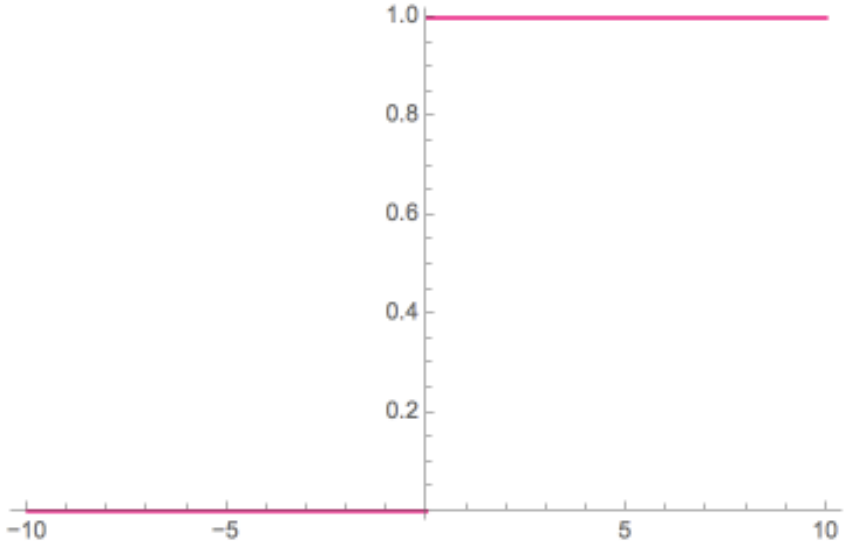
Artificial Neurons:Perceptrons

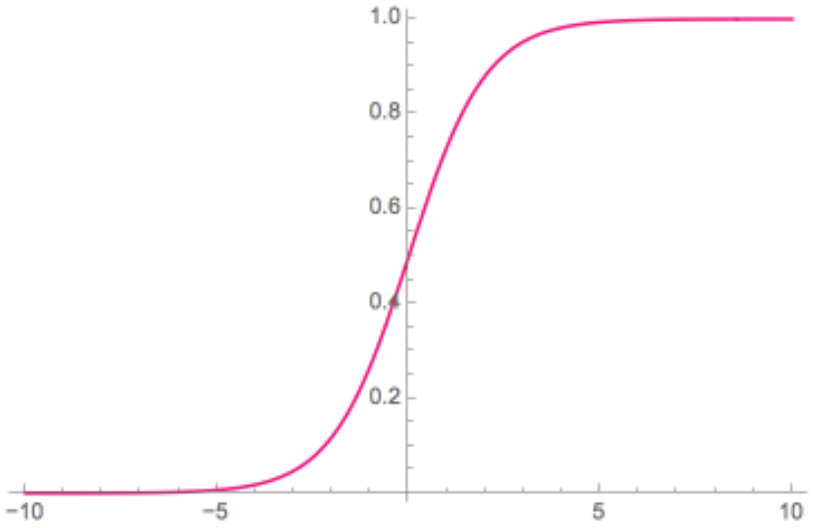
- Suppose the weekend is coming up, and you've heard that there's going to be a food festival in your city.
- You might make your decision by weighing up three factors:
  - Is the weather good?
  - Does your boyfriend or girlfriend want to accompany you?
  - Is the festival near public transit? (You don't own a car).

Artificial Neurons:Perceptrons

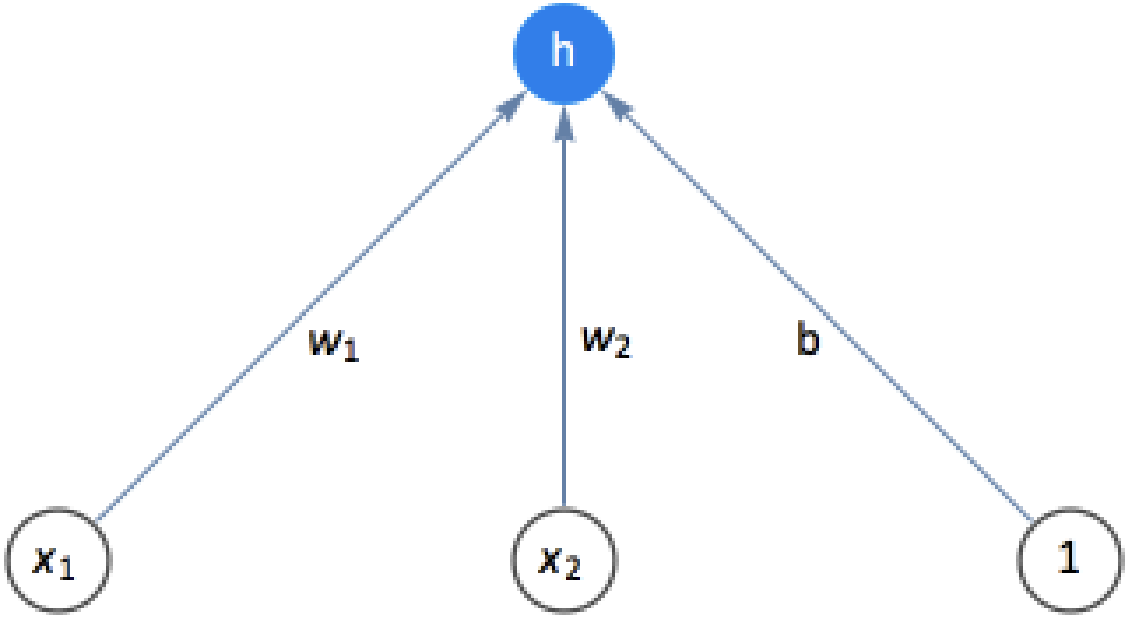


Artificial Neurons:Sigmoid Neuron



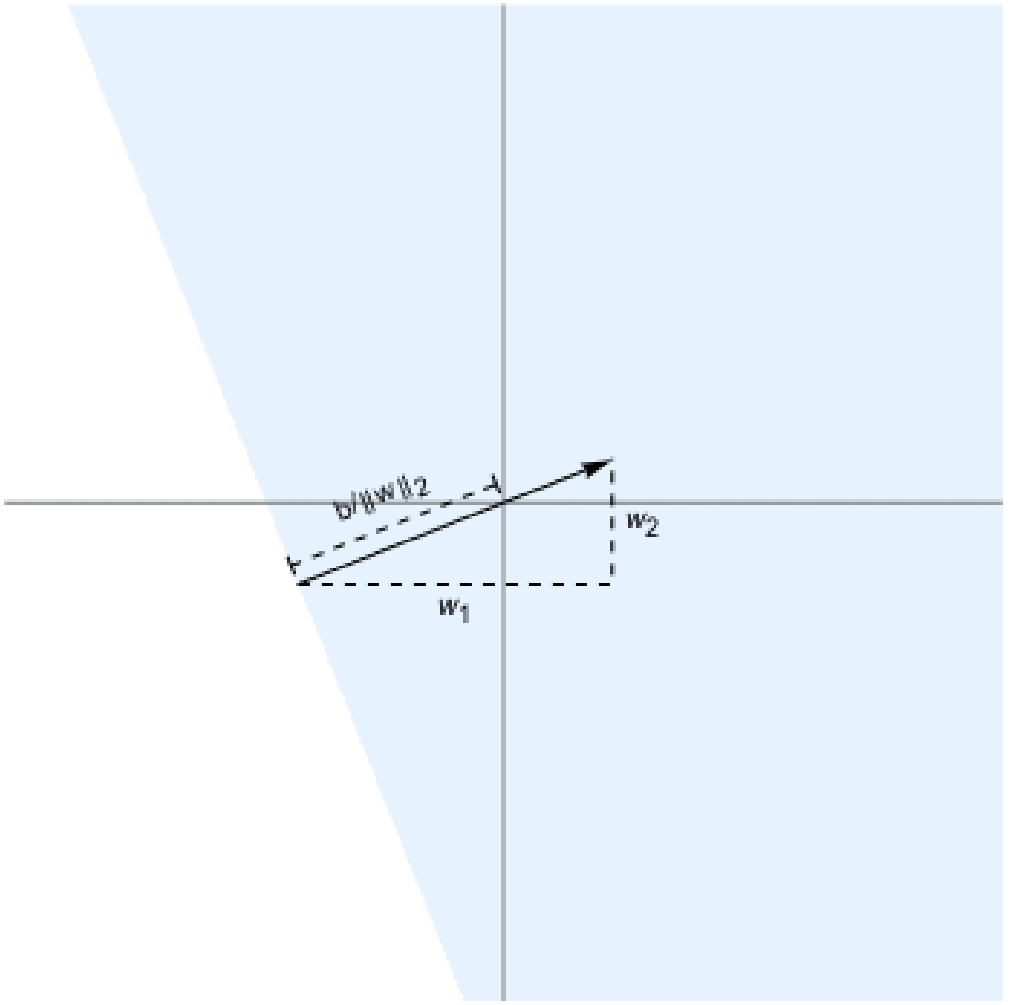


Artificial Neurons:Sigmoid Neuron

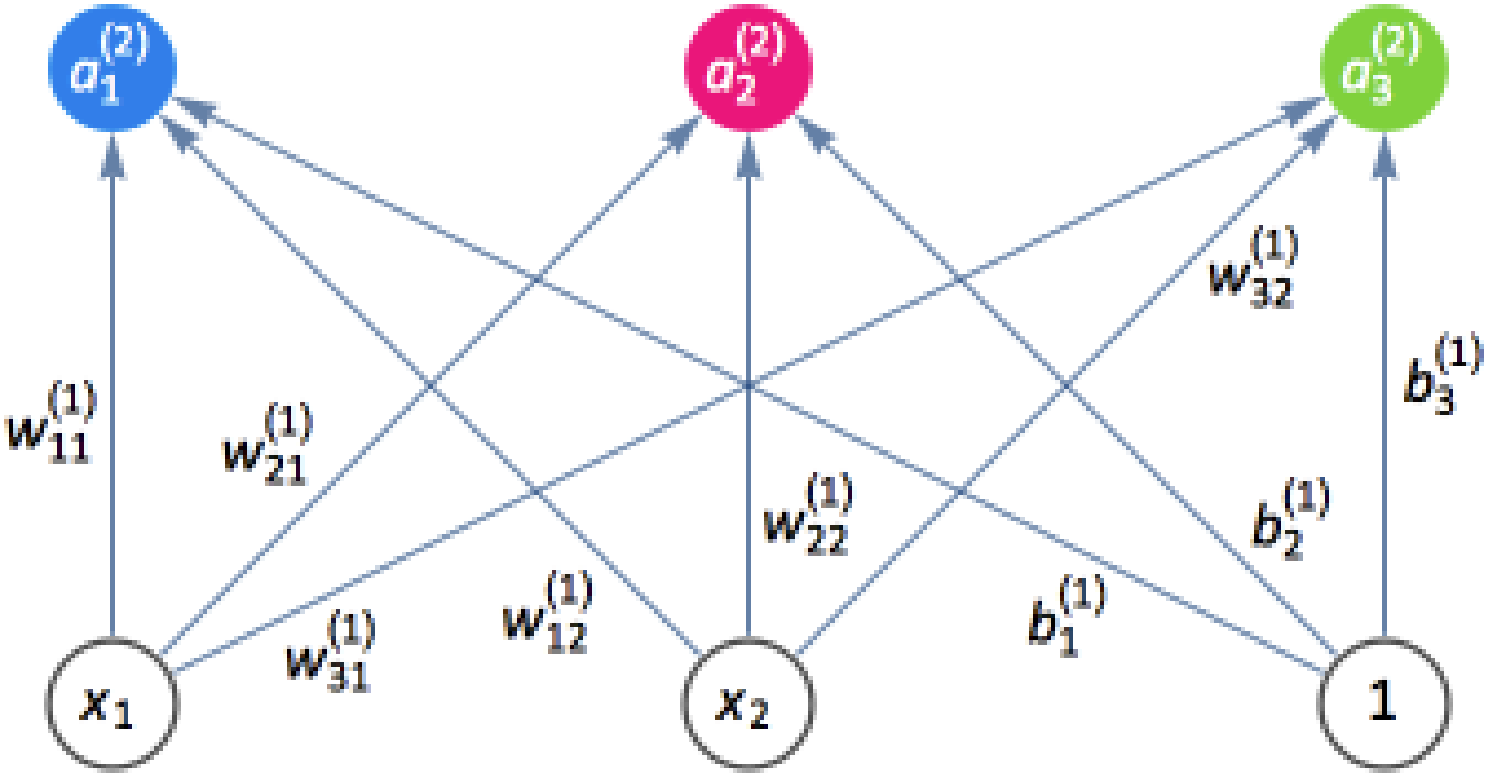


$$h = f\left(w_1 x_1 + w_2 x_2 + b\right)$$

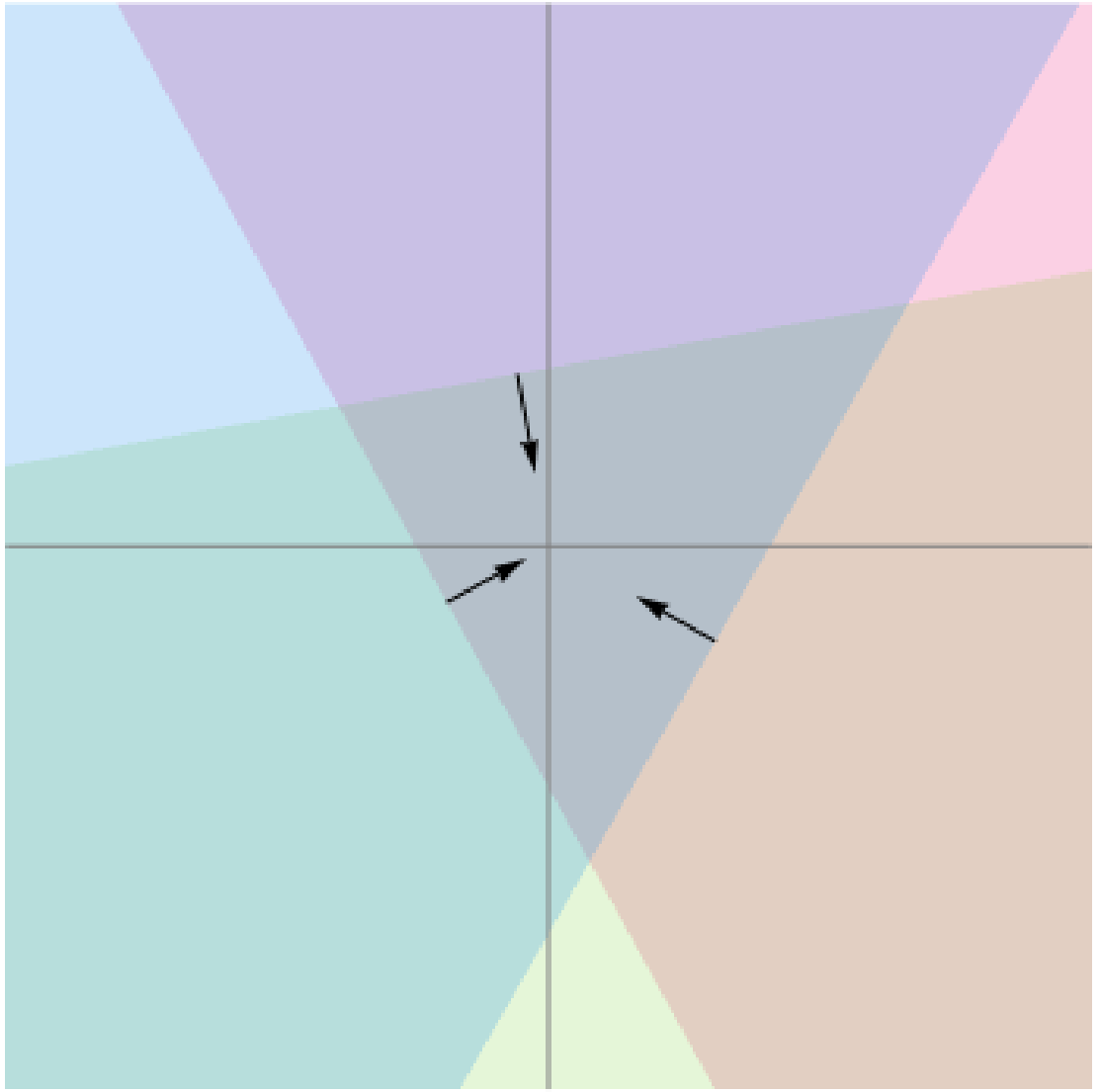
Artificial Neurons:Sigmoid Neuron



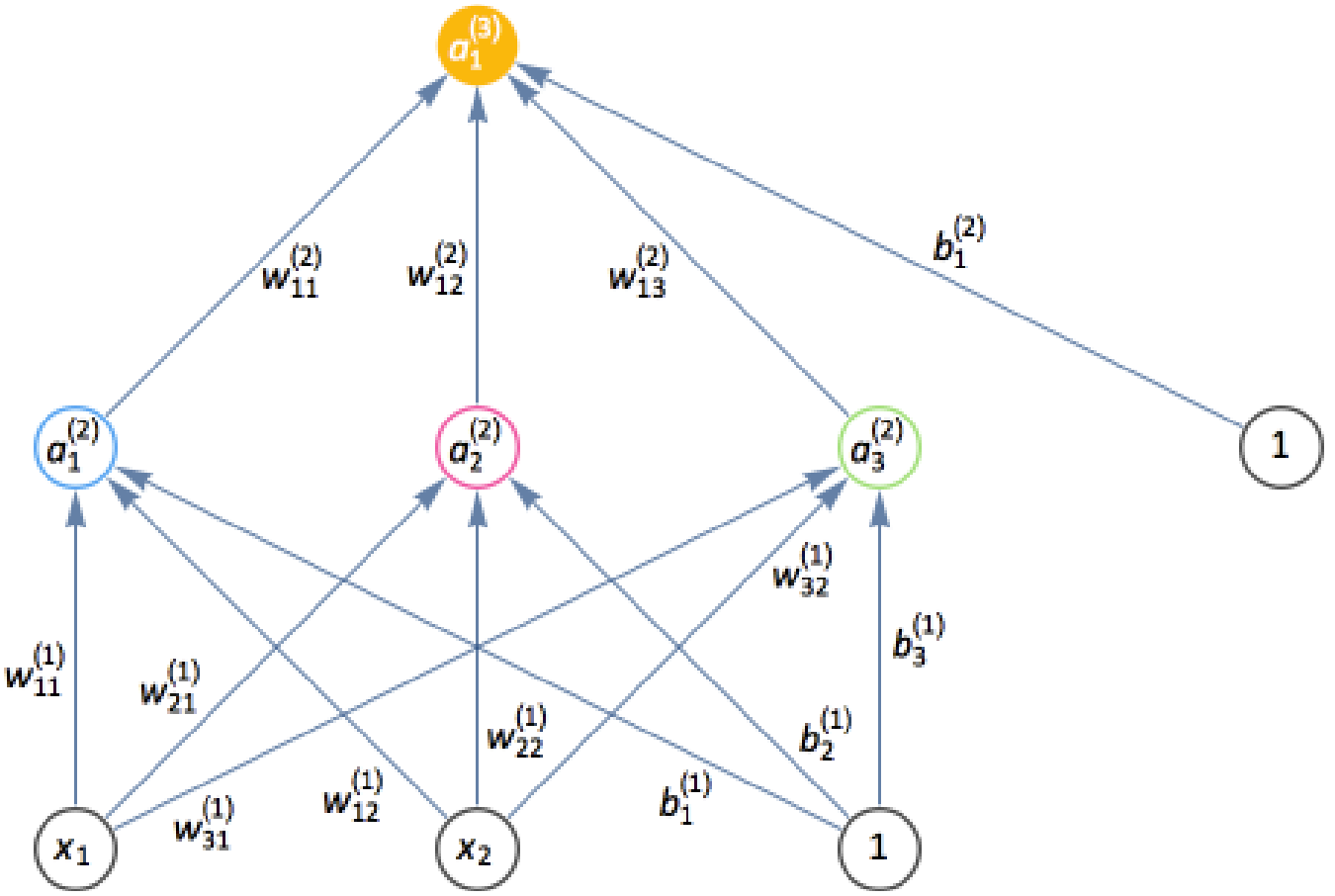
Deep Neural Network



Deep Neural Network

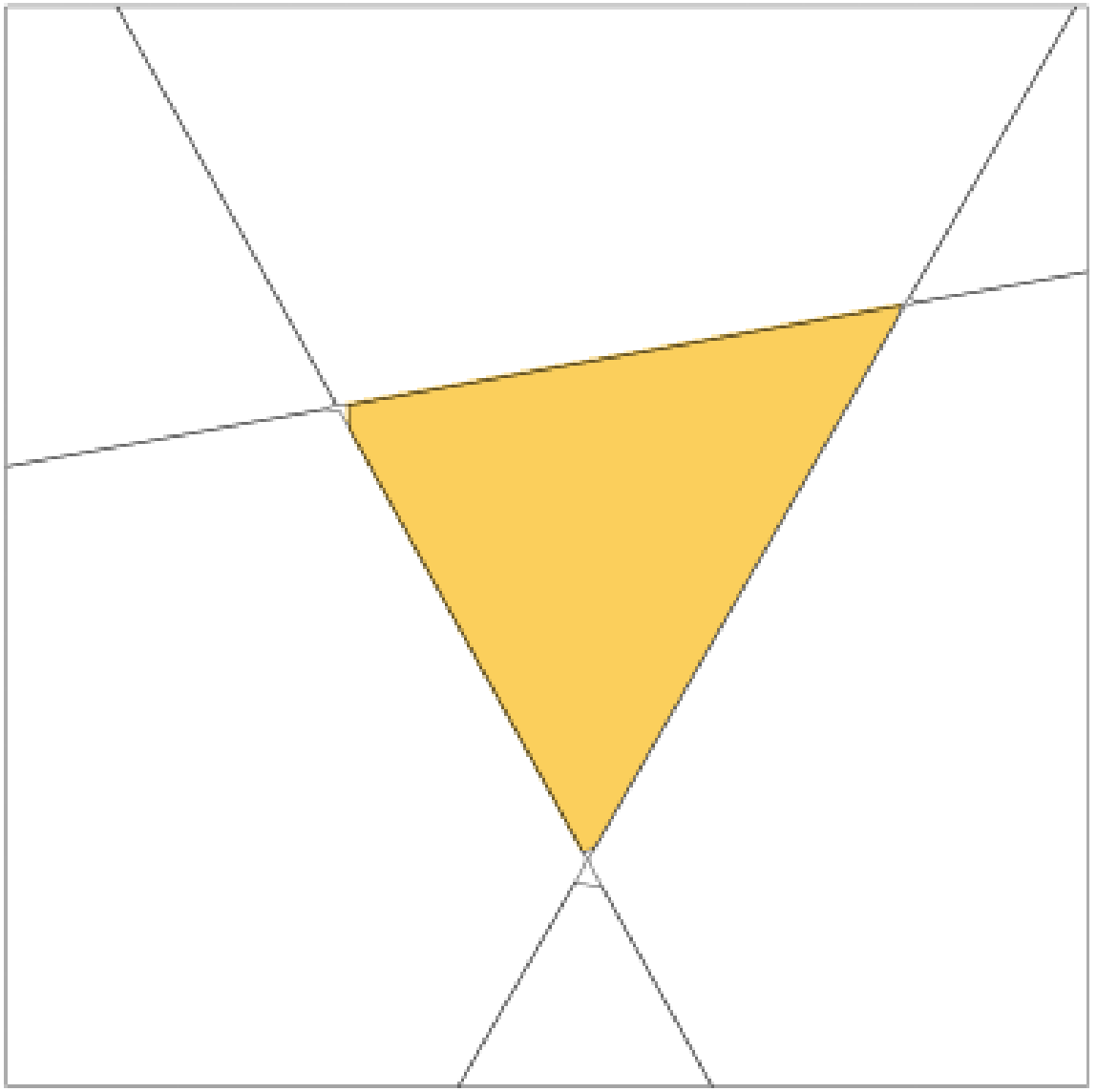


Deep Neural Network



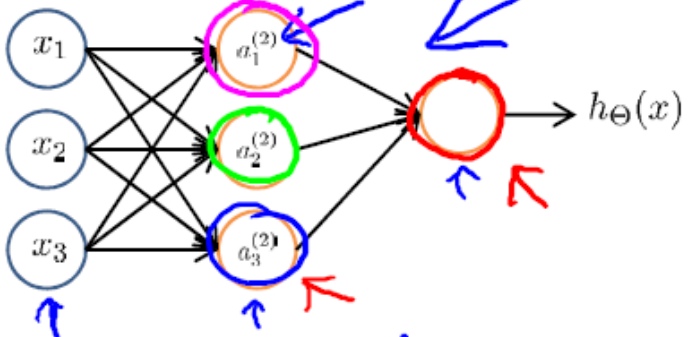
$$a_1^{(3)} = f\left(w_{11}^{(2)} a_1^{(2)} + w_{12}^{(2)} a_2^{(2)} + w_{13}^{(2)} a_3^{(2)} + b_1^{(2)}\right)$$

Deep Neural Network



Deep Neural Network

Neural Network



→ $a_i^{(j)}$ = "activation" of unit i in layer j

→ $\Theta^{(j)}$ = matrix of weights controlling function mapping from layer j to layer $j+1$

$\Theta^{(1)} \in \mathbb{R}^{3\times 4}$

$h_\Theta(x)$

→ $a_1^{(2)} = g(\Theta_{10}^{(1)} x_0 + \Theta_{11}^{(1)} x_1 + \Theta_{12}^{(1)} x_2 + \Theta_{13}^{(1)} x_3)$

→ $a_2^{(2)} = g(\Theta_{20}^{(1)} x_0 + \Theta_{21}^{(1)} x_1 + \Theta_{22}^{(1)} x_2 + \Theta_{23}^{(1)} x_3)$

→ $a_3^{(2)} = g(\Theta_{30}^{(1)} x_0 + \Theta_{31}^{(1)} x_1 + \Theta_{32}^{(1)} x_2 + \Theta_{33}^{(1)} x_3)$

$\Theta^{(2)}$

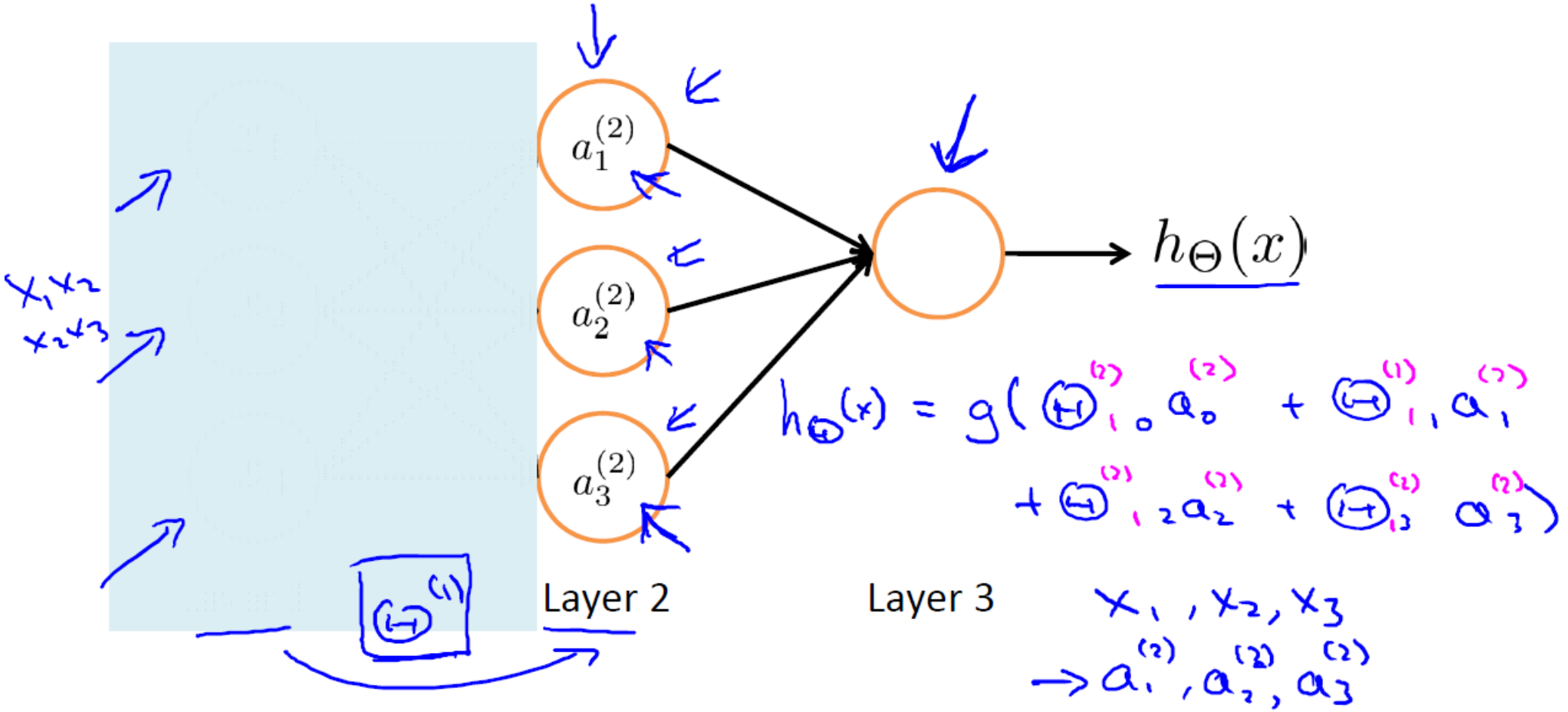
→ $h_\Theta(x) = a_1^{(3)} = g(\Theta_{10}^{(2)} a_0^{(2)} + \Theta_{11}^{(2)} a_1^{(2)} + \Theta_{12}^{(2)} a_2^{(2)} + \Theta_{13}^{(2)} a_3^{(2)})$

→ If network has s_j units in layer j, s_{j+1} units in layer $j+1$, then $\Theta^{(j)}$ will be of dimension $s_{j+1} \times (s_j + 1)$.

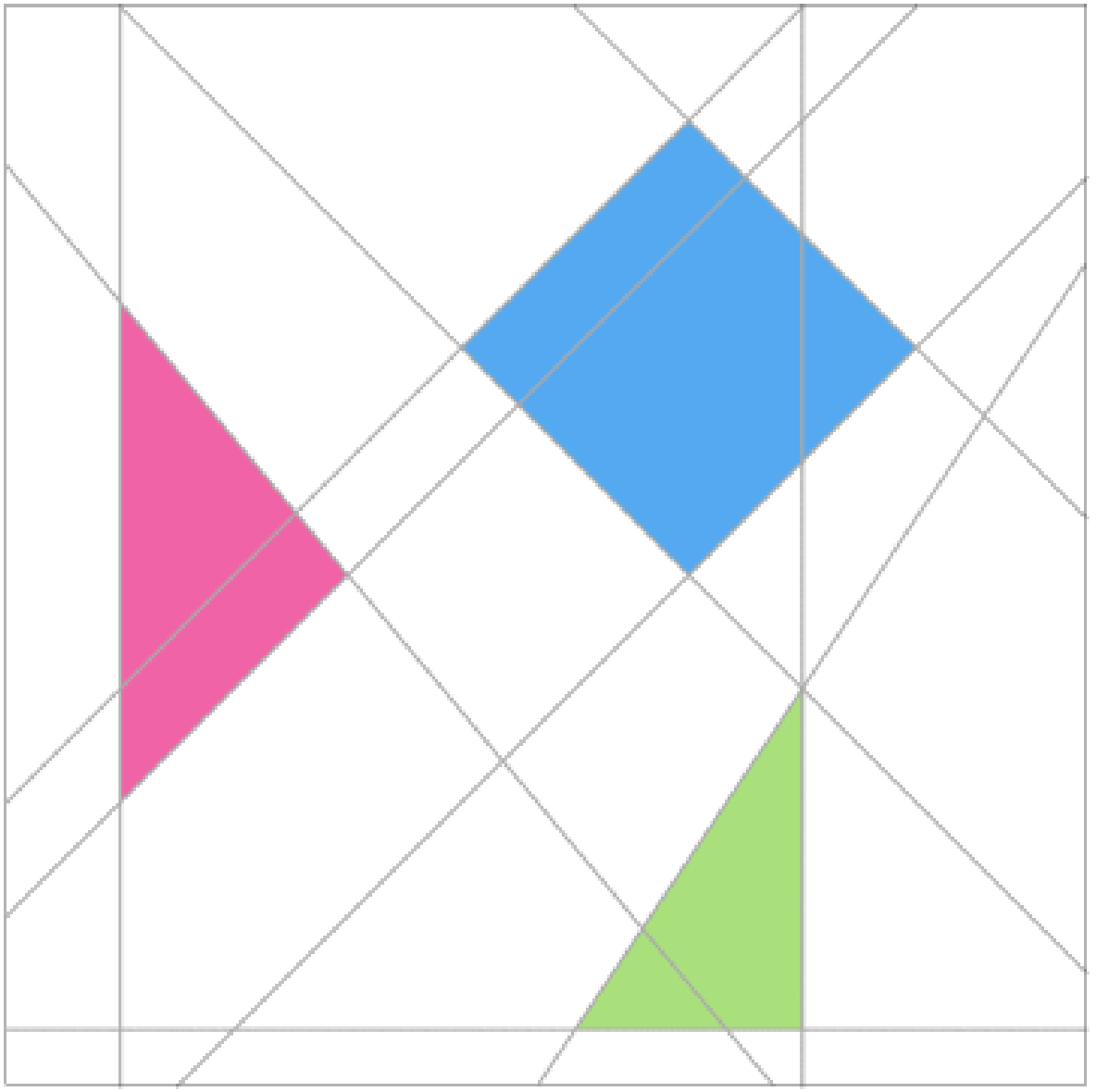
$s_{j+1} \times (s_j + 1)$

Deep Neural Network

Neural Network learning its own features



Deep Neural Network

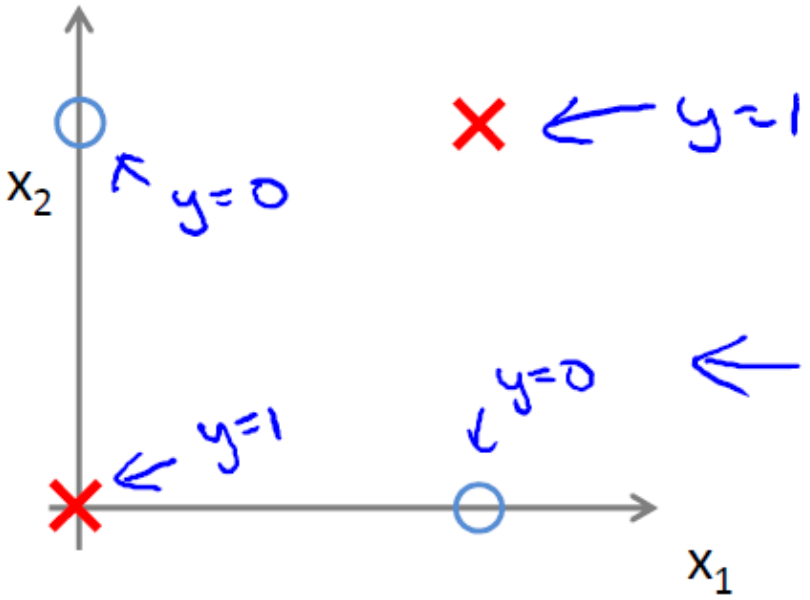


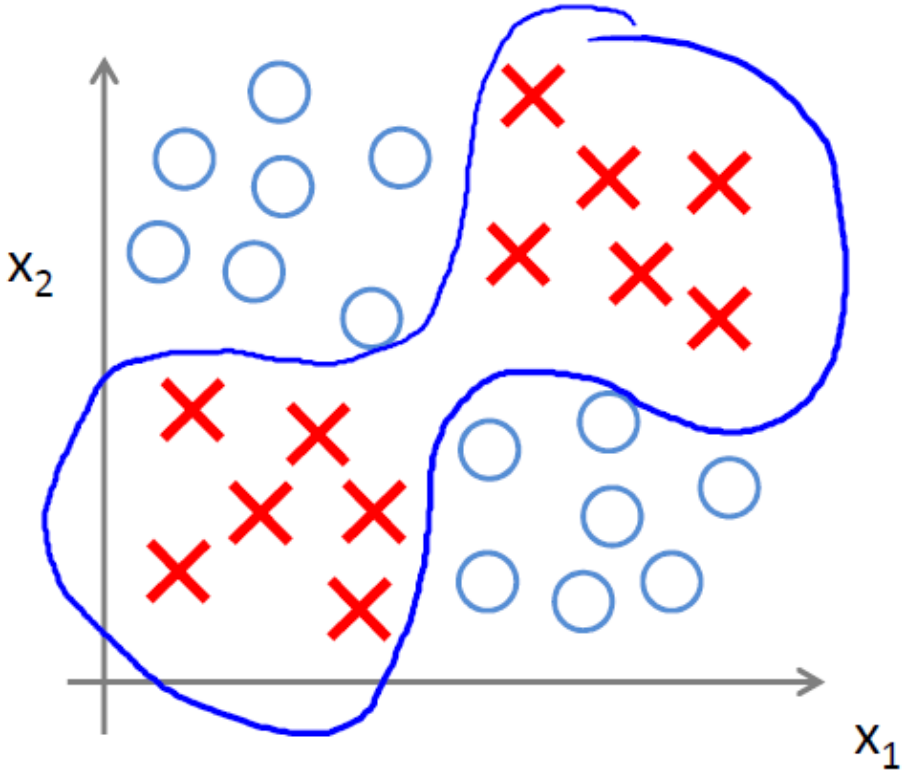
Deep Neural Network

Deep Neural Network

Non-linear classification example: XOR/XNOR

x_1, x_2 are binary (0 or 1).





$y = x_1 \text{ XOR } x_2$

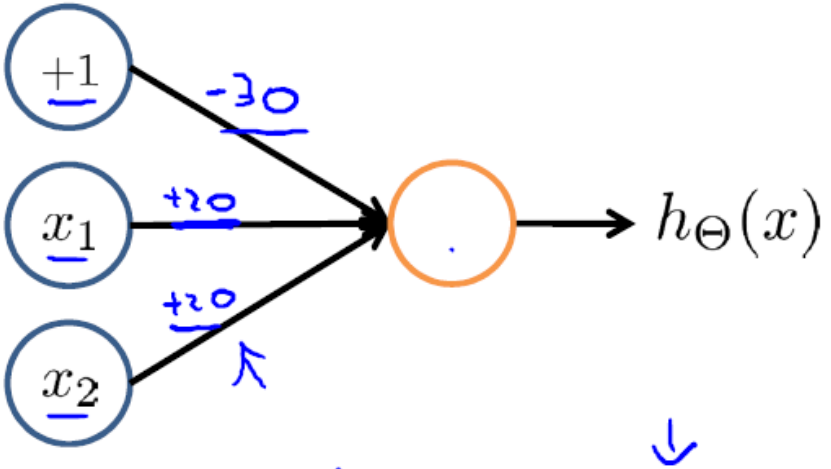
$x_1 \text{ XNOR } x_2$

$\text{NOT } (x_1 \text{ XOR } x_2)$

Deep Neural Network

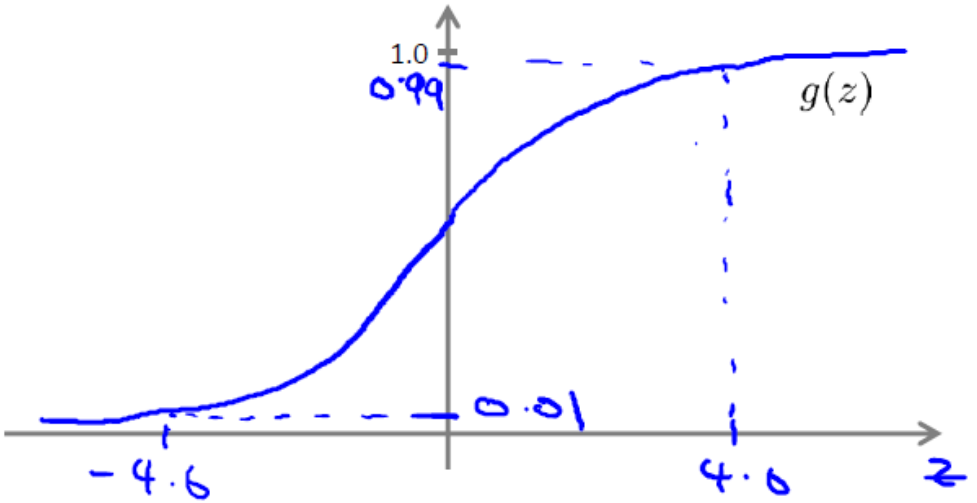
Simple example: AND

- $x_1, x_2 \in \{0, 1\}$
- $y = x_1$ AND x_2



$h_\Theta(x) = g(-30 + 20x_1 + 20x_2)$

$\Theta^{(1)}_{10}$ $\Theta^{(1)}_{11}$ $\Theta^{(1)}_{12}$

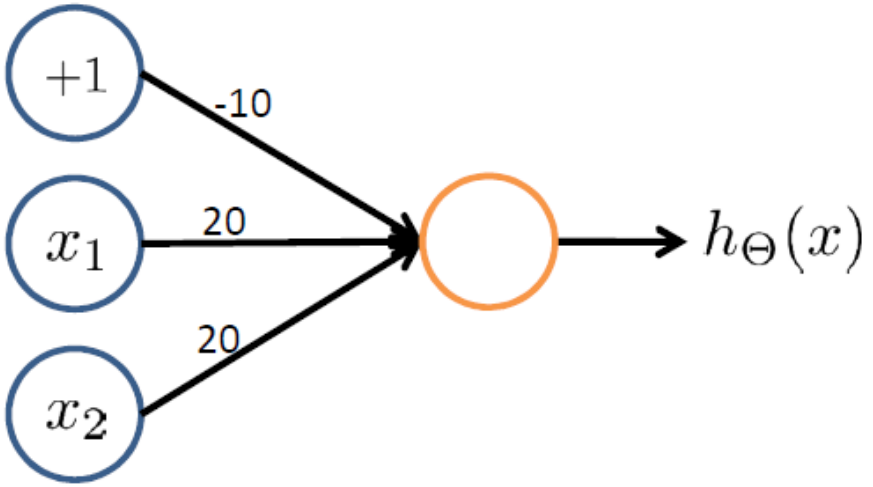


x_1	x_2	$h_\Theta(x)$
0	0	$g(-30) \approx 0$
0	1	$g(-10) \approx 0$
1	0	$g(-10) \approx 0$
1	1	$g(10) \approx 1$

$h_\Theta(x) \approx x_1$ AND x_2

Deep Neural Network

Example: OR function



$g(-10 + 20x_1 + 20x_2)$

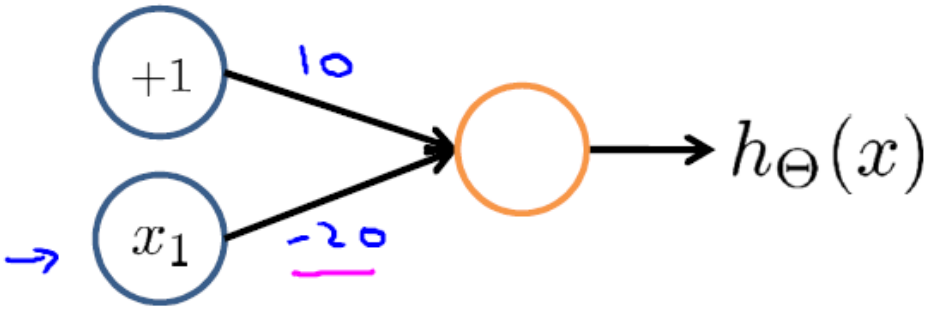
x_1	x_2	$h_\Theta(x)$
0	0	$g(-10) \approx 0$
0	1	$g(10) \approx 1$
1	0	≈ 1
1	1	≈ 1

Deep Neural Network

→ x_1 AND x_2 → x_1 OR x_2 {0,1}.

Negation:

NOT x_1



x_1	$h_\Theta(x)$
0	$g(10) \approx 1$
1	$g(-10) \approx 0$

$h_\Theta(x) = g(10 - 20x_1)$

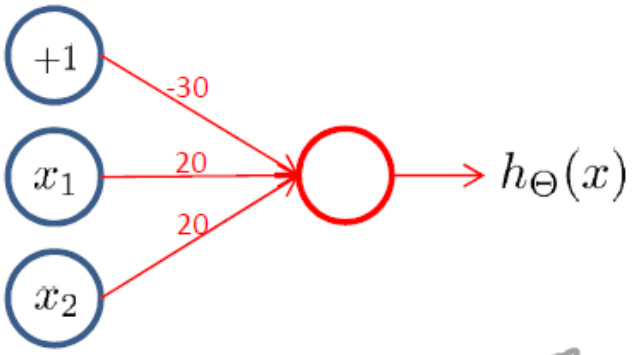
→ (NOT x_1) AND (NOT x_2)

(=1 if and only if

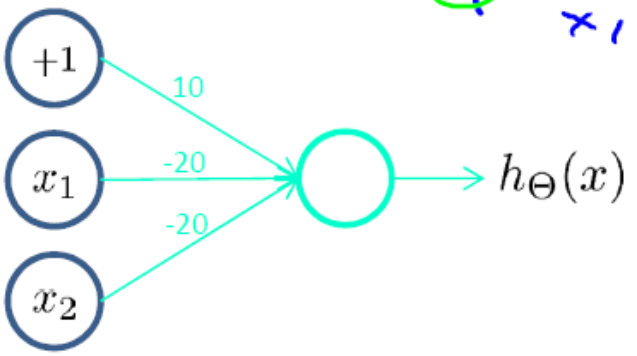
→ $x_1 = x_2 = 0$

Deep Neural Network

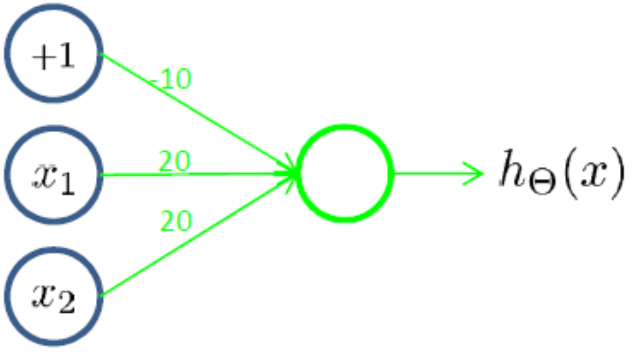
Putting it together: x_1 XNOR x_2



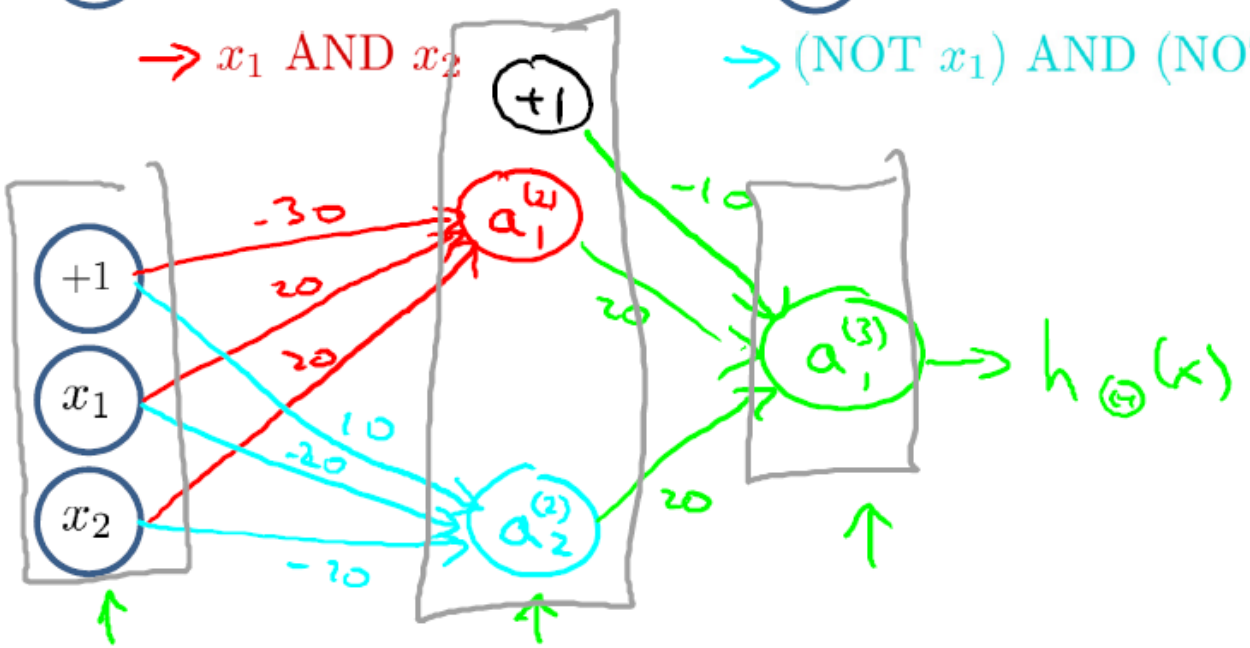
→ x_1 AND x_2



→ (NOT x_1) AND (NOT x_2)

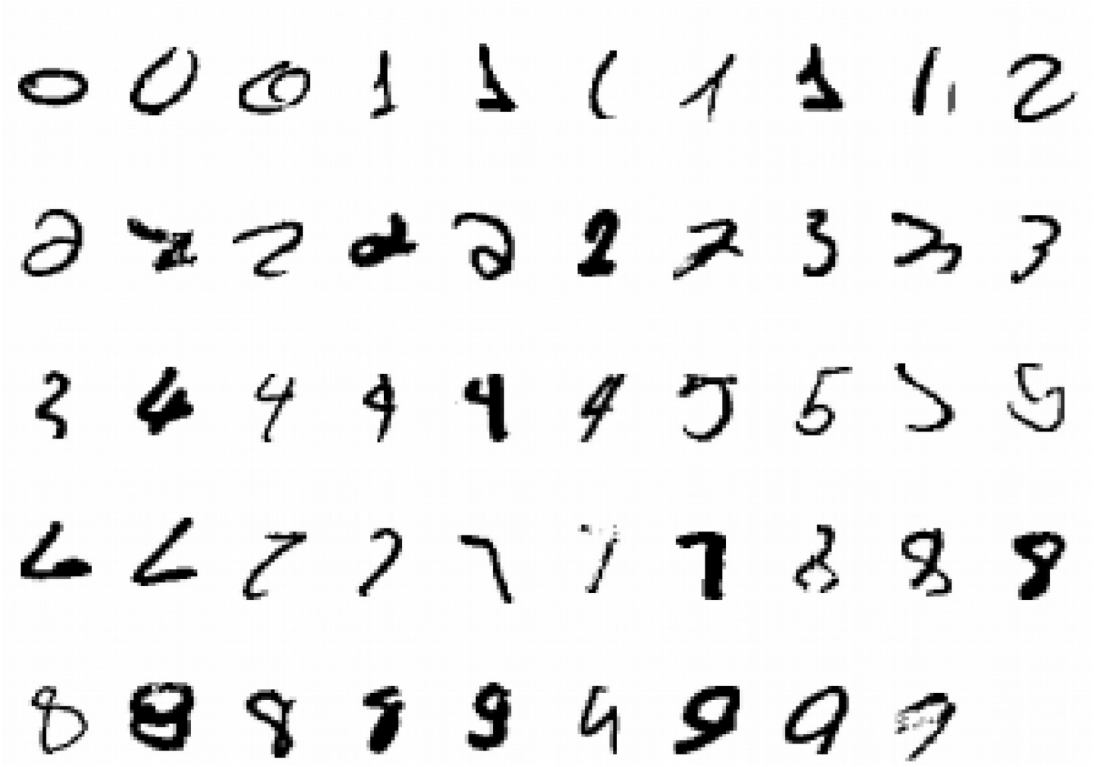


→ x_1 OR x_2



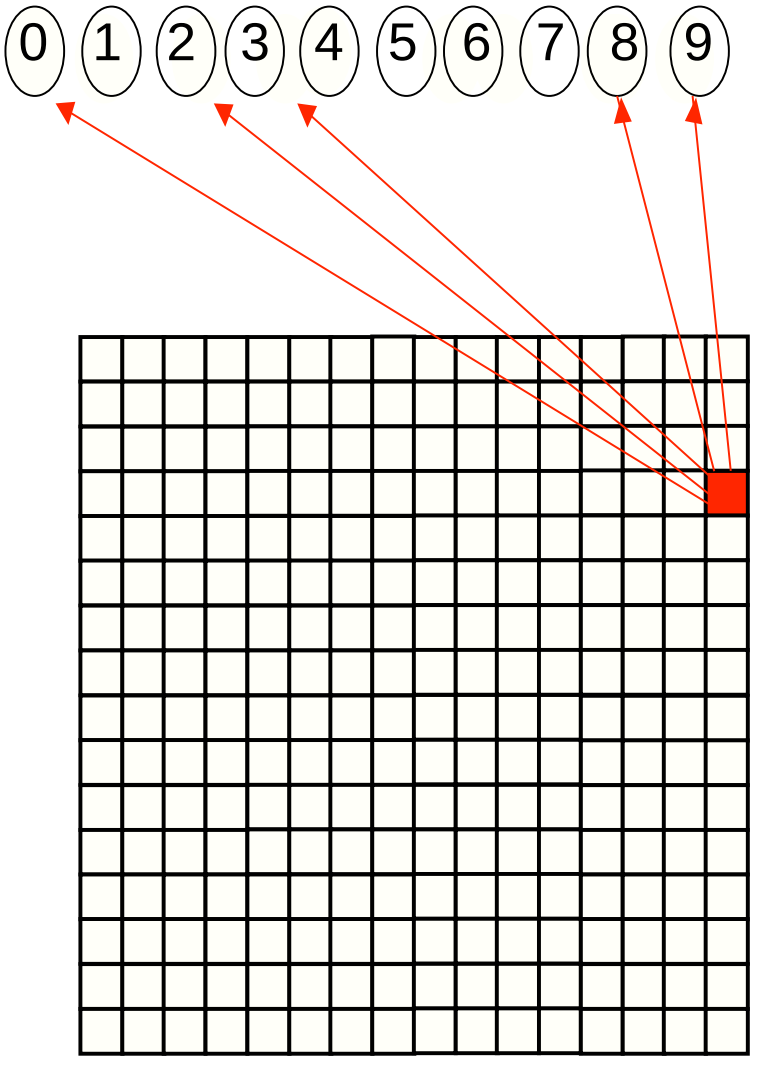
x_1	x_2	$a_1^{(2)}$	$a_2^{(2)}$	$h_\Theta(x)$
0	0	0	1	1
0	1	0	0	0
1	0	0	0	0
1	1	1	0	1

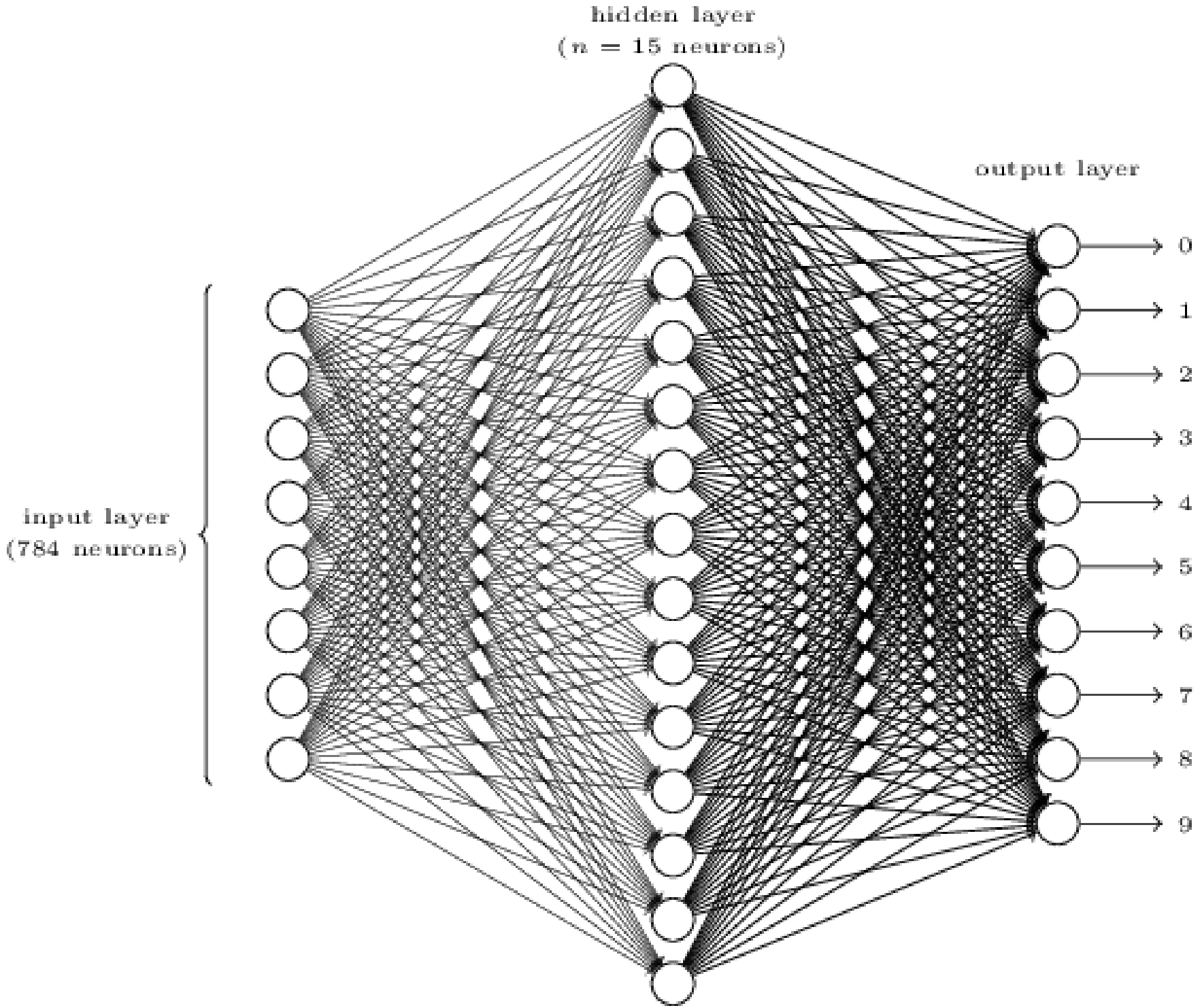
It is very hard to say what makes a 2



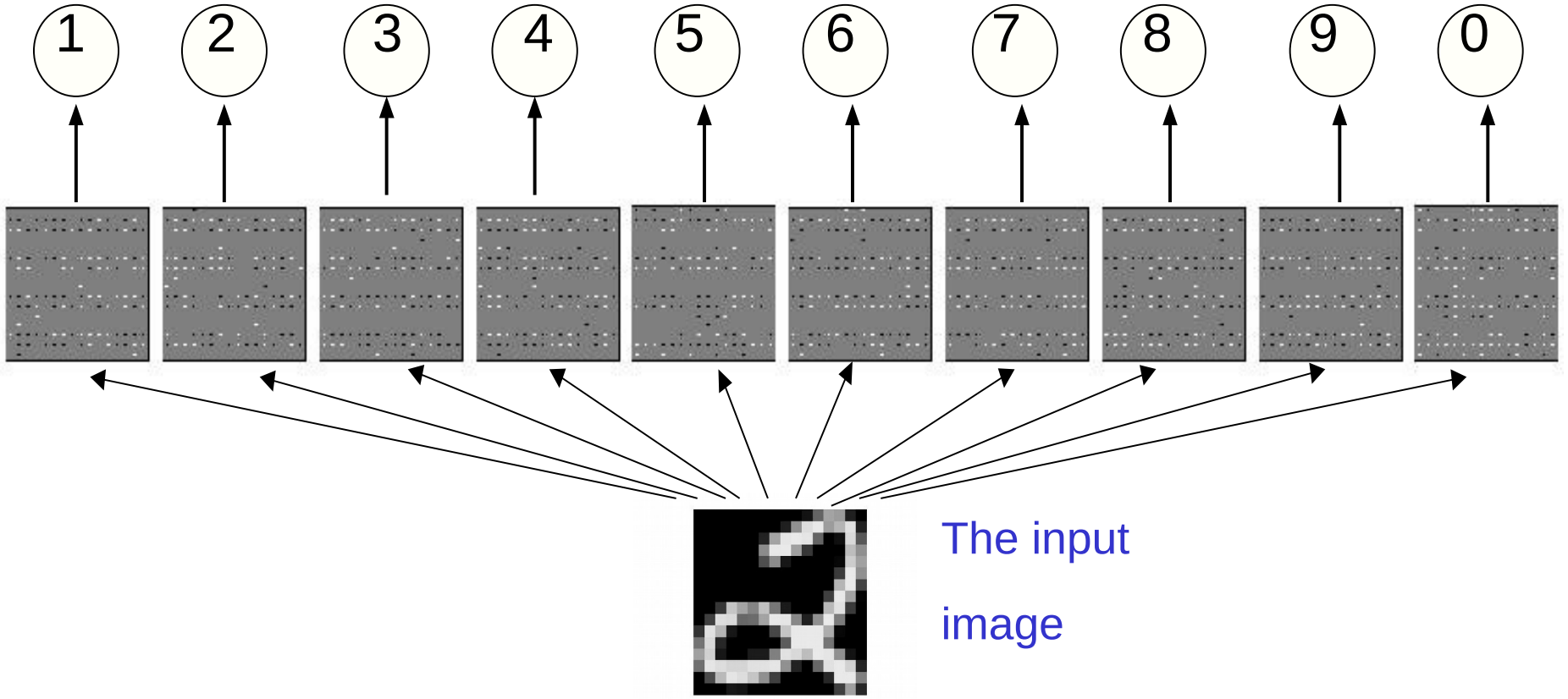
A very simple way to recognize handwritten shapes

- ? Consider a neural network with two layers of neurons.
  - ? neurons in the top layer represent known shapes.
  - ? neurons in the bottom layer represent pixel intensities.
- ? A pixel gets to vote if it has ink on it.
  - ? Each inked pixel can vote for several different shapes.
- ? The shape that gets the most votes wins.




hidden layer
(n = 15 neurons)
output layer
input layer
(784 neurons)
0
1
2
3
4
5
6
7
8
9

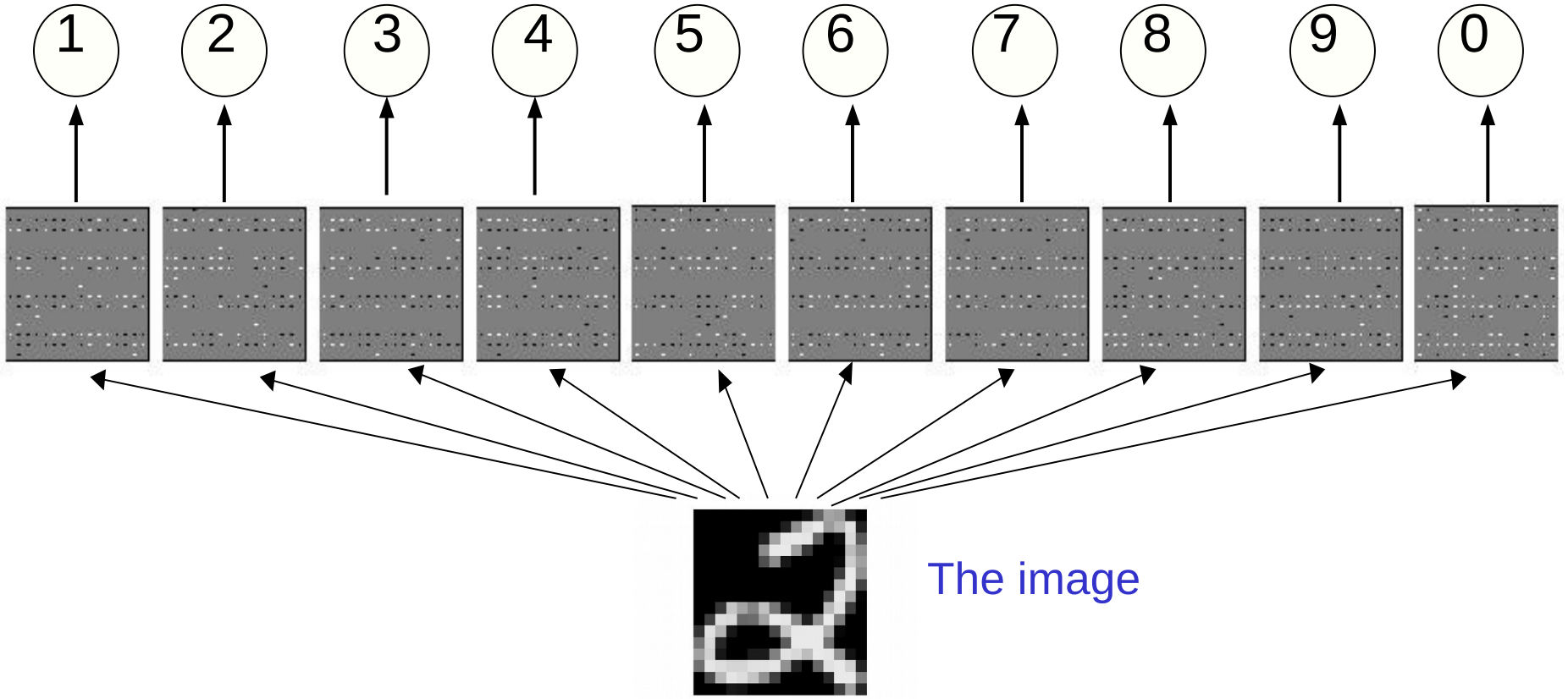
How to display the weights



Give each output unit its own “map” of the input image and display the weight coming from each pixel in the location of that pixel in the map.

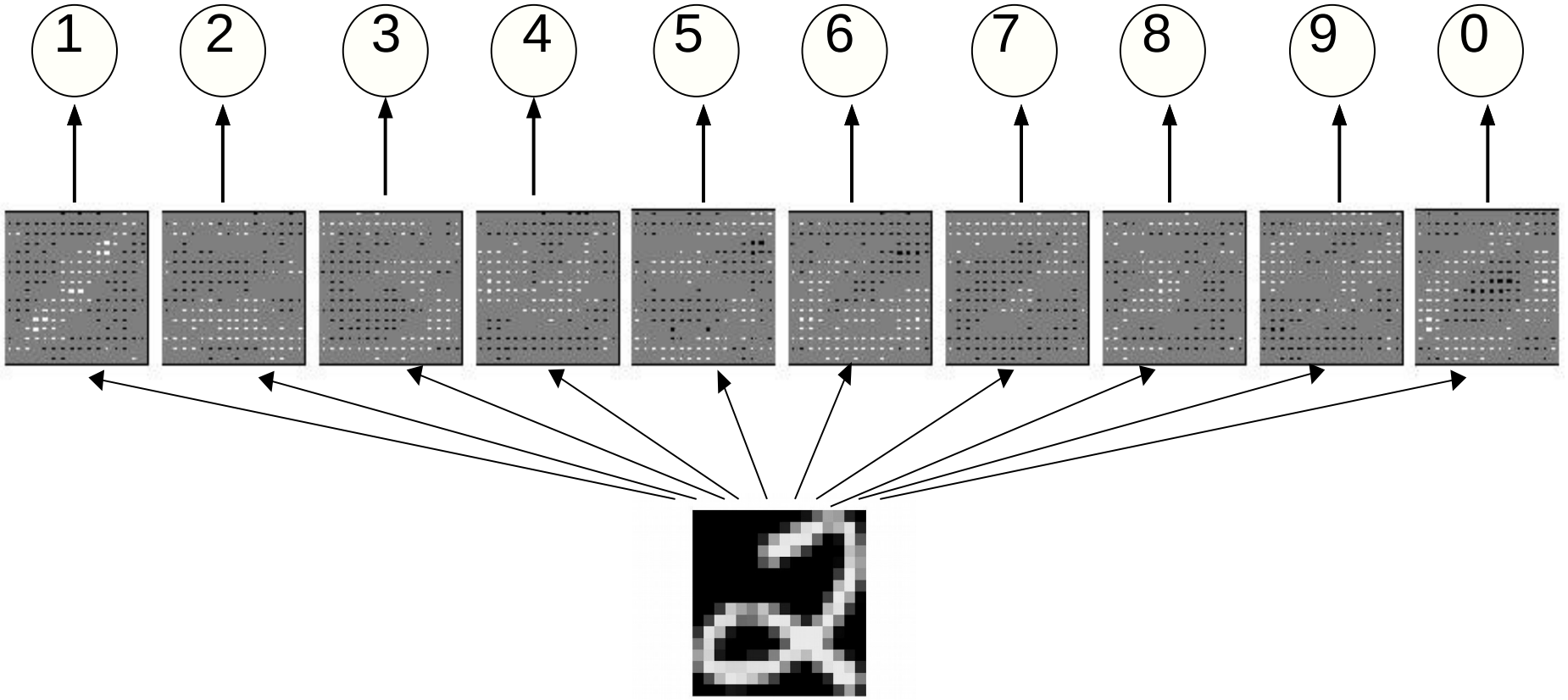
Use a black or white blob with the area representing the magnitude of the weight and the color representing the sign.

How to learn the weights

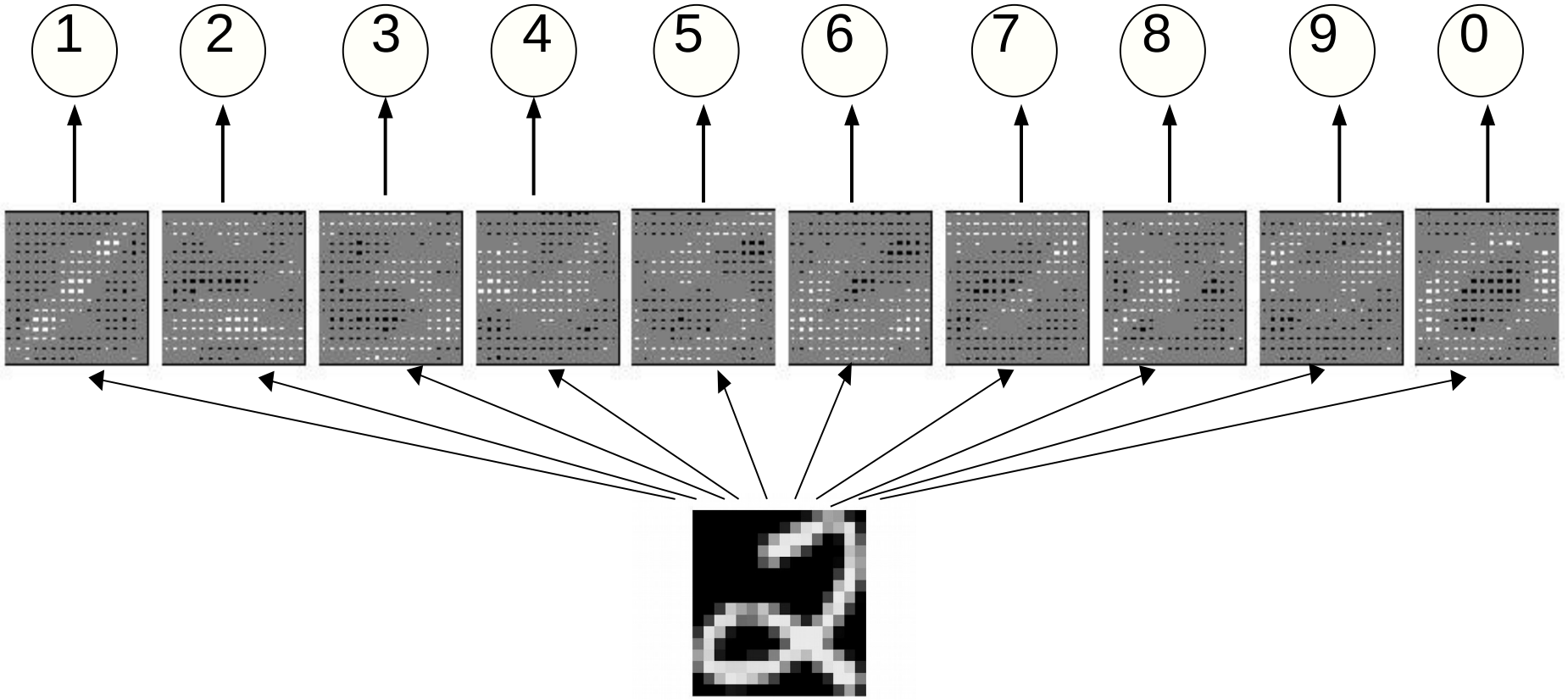


Show the network an image and increment the weights from active pixels to the correct class.

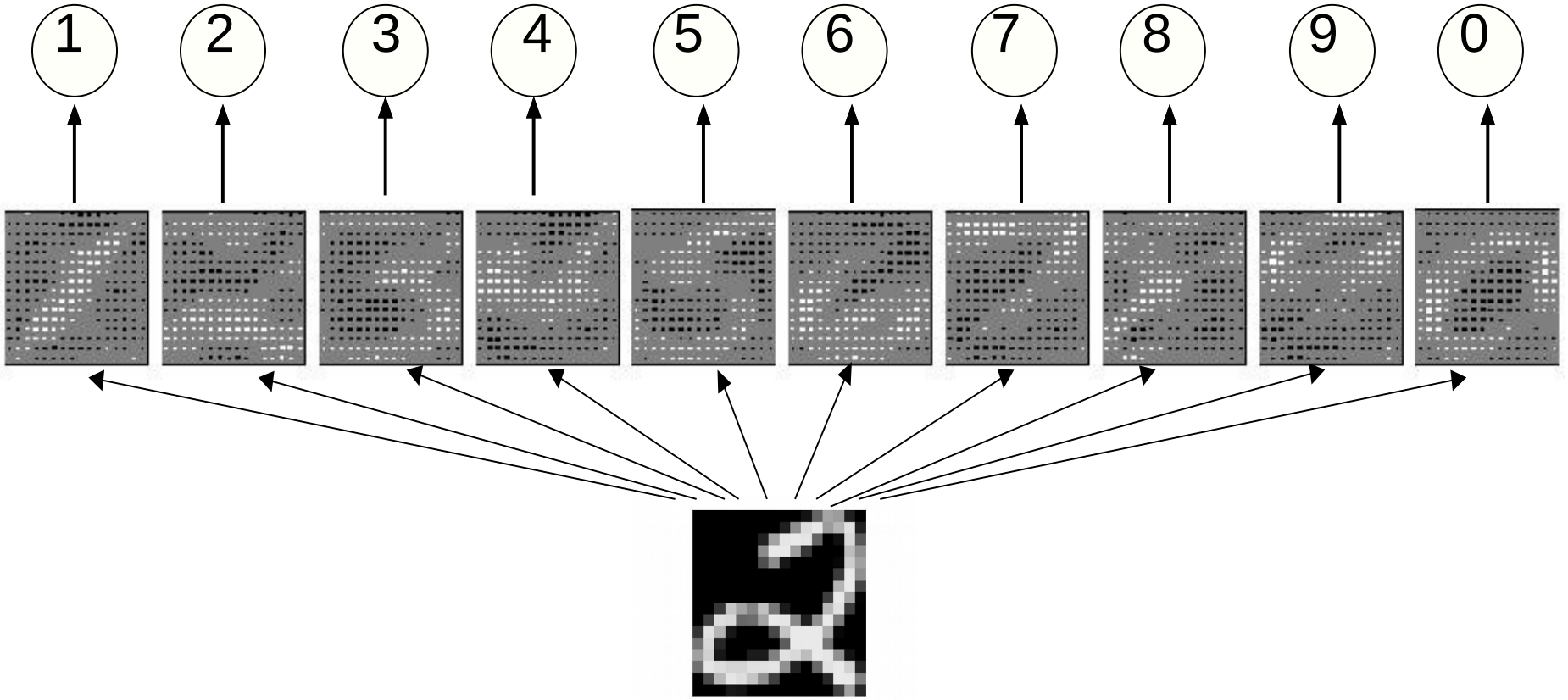
Then decrement the weights from active pixels to whatever class the network guesses.



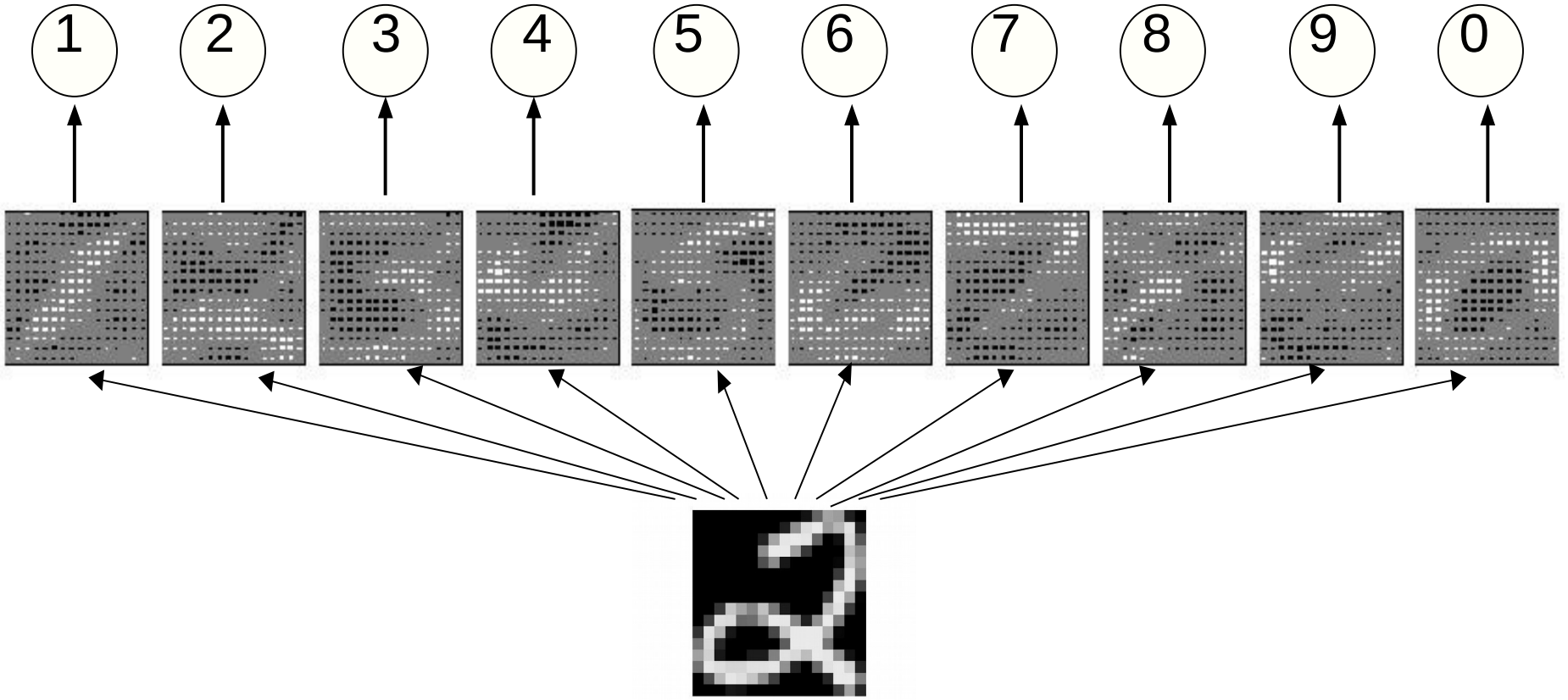
The image



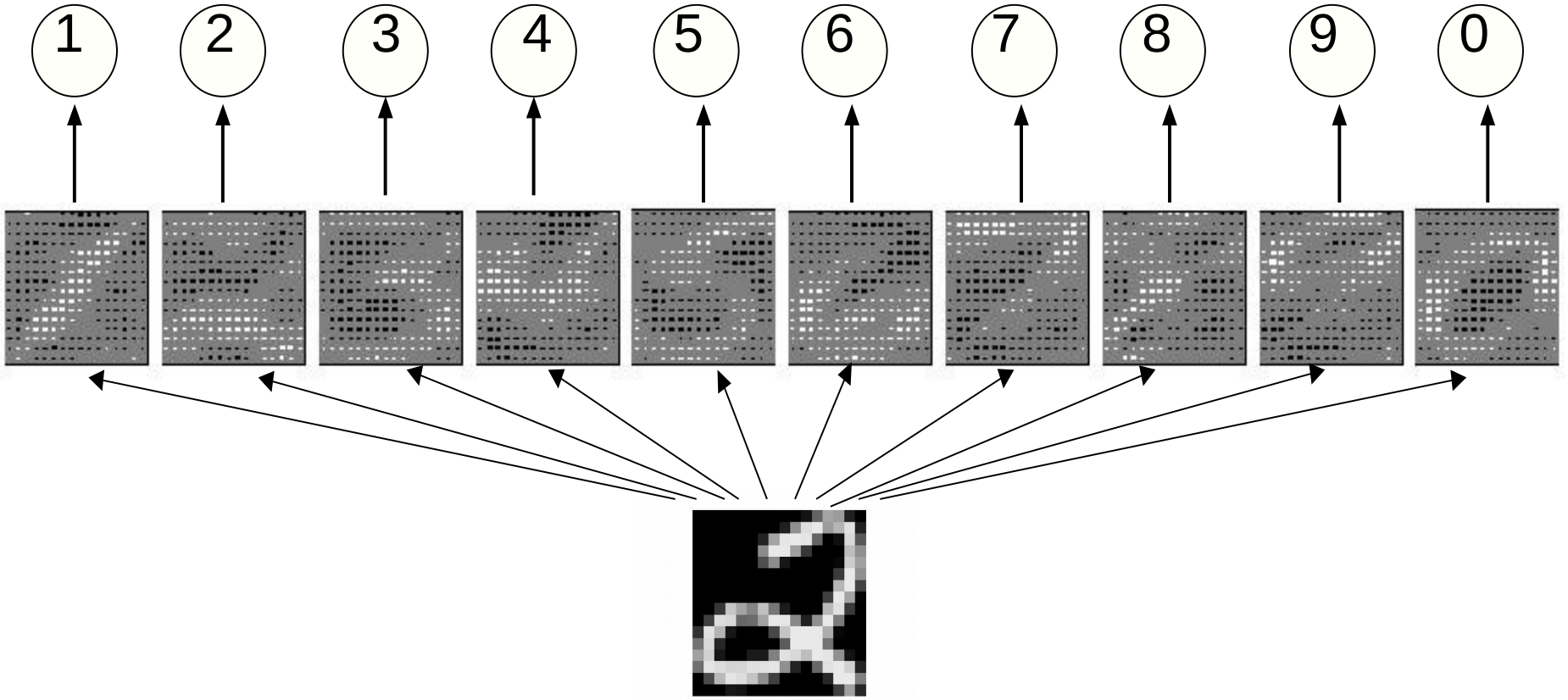
The image



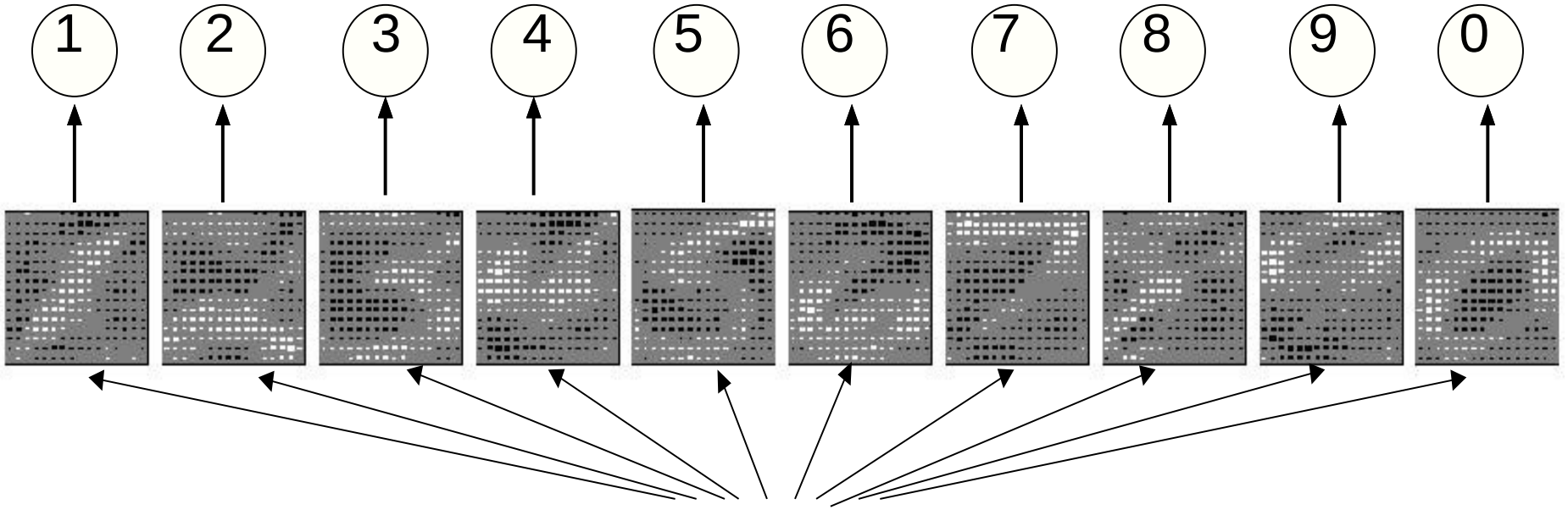
The image



The image

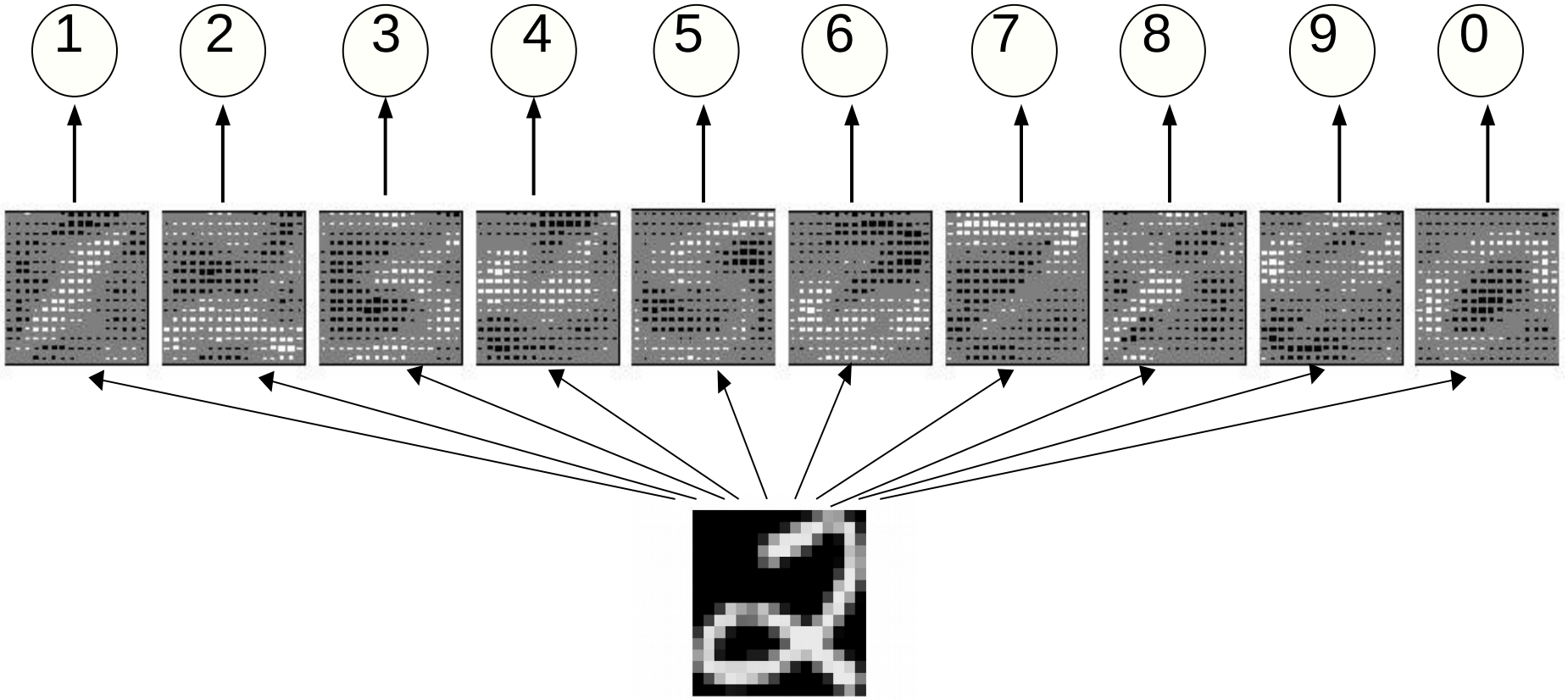


The image

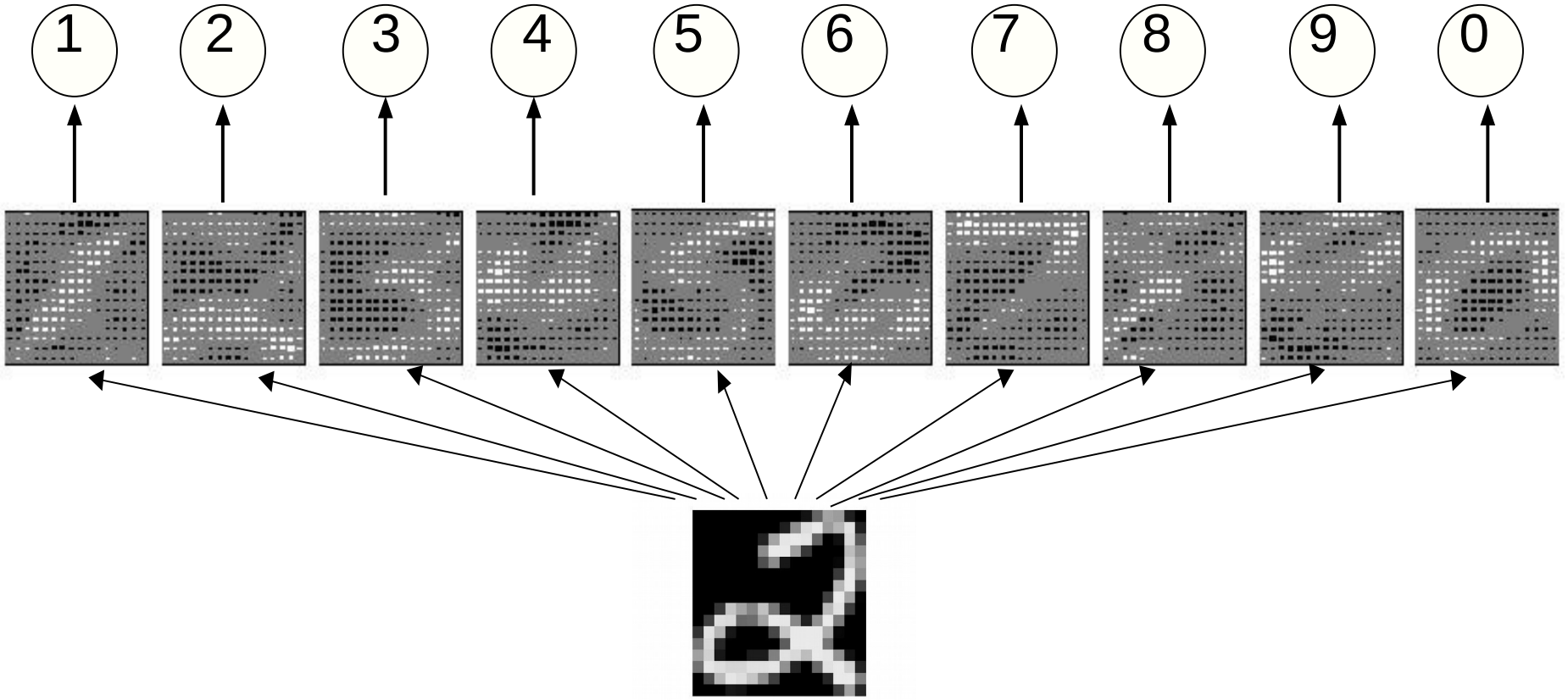


The image

The learned weights



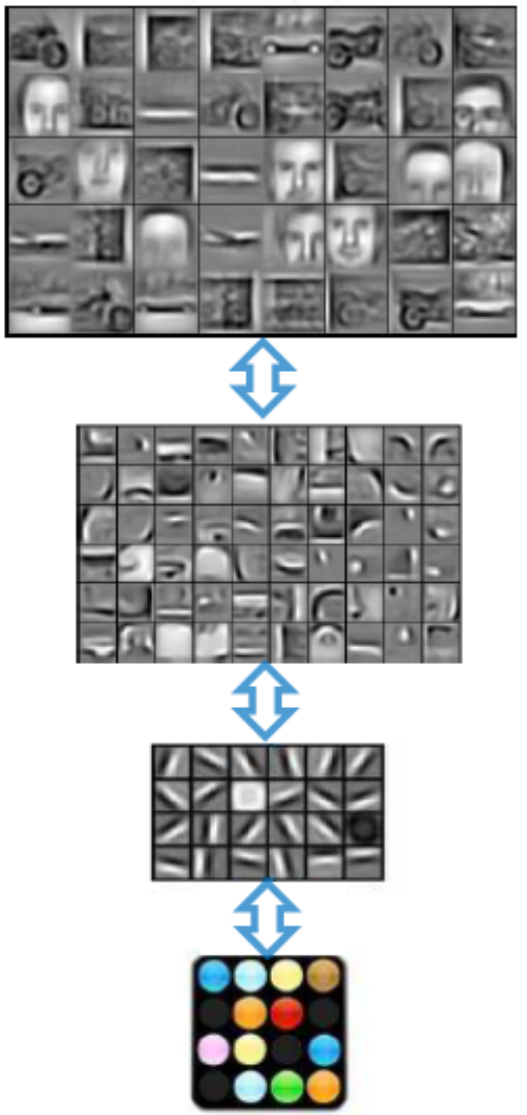
The image

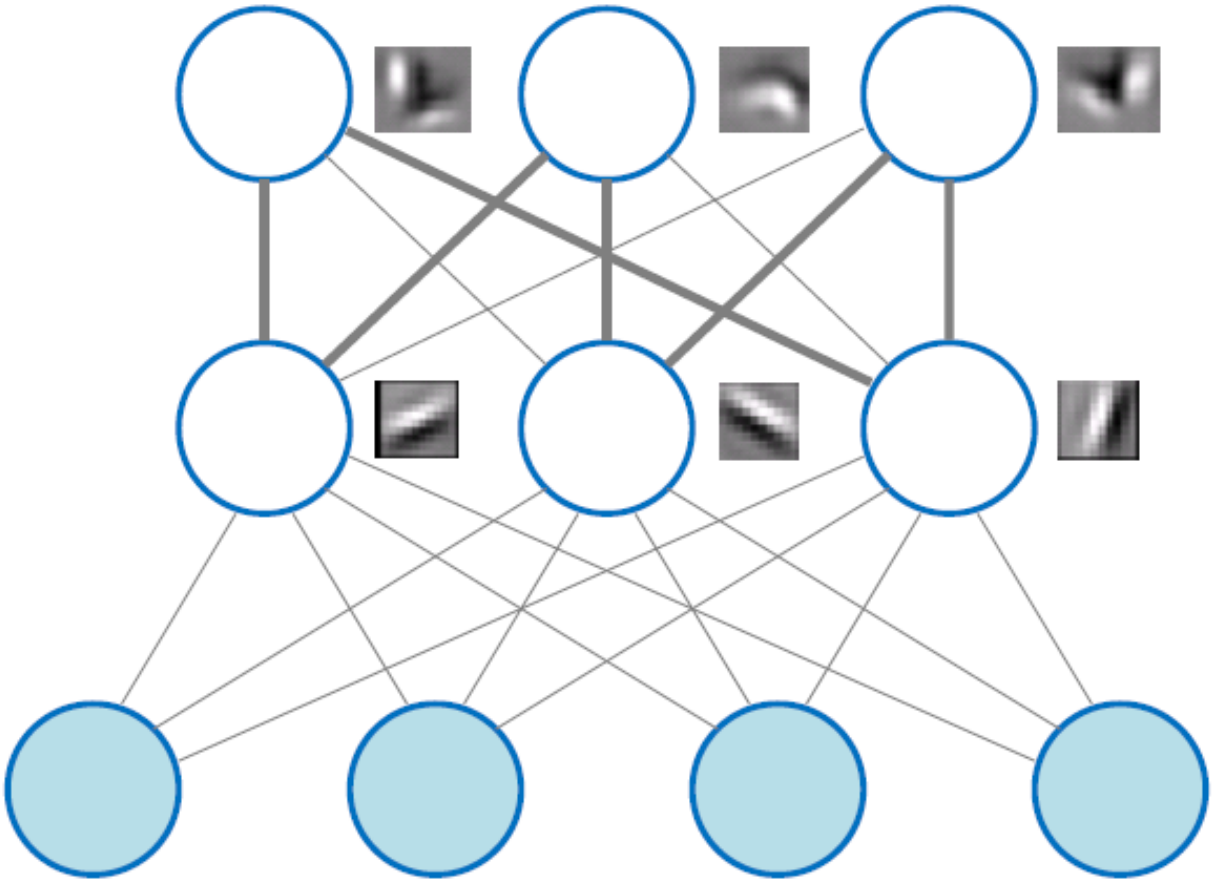


The image

Learning Feature Hierarchy

- Deep Learning
  - Deep architectures can be representationally efficient.
  - Natural progression from low level to high level structures.
  - Can share the lower-level representations for multiple tasks.





Higher layer: DBNs
(Combinations
of edges)

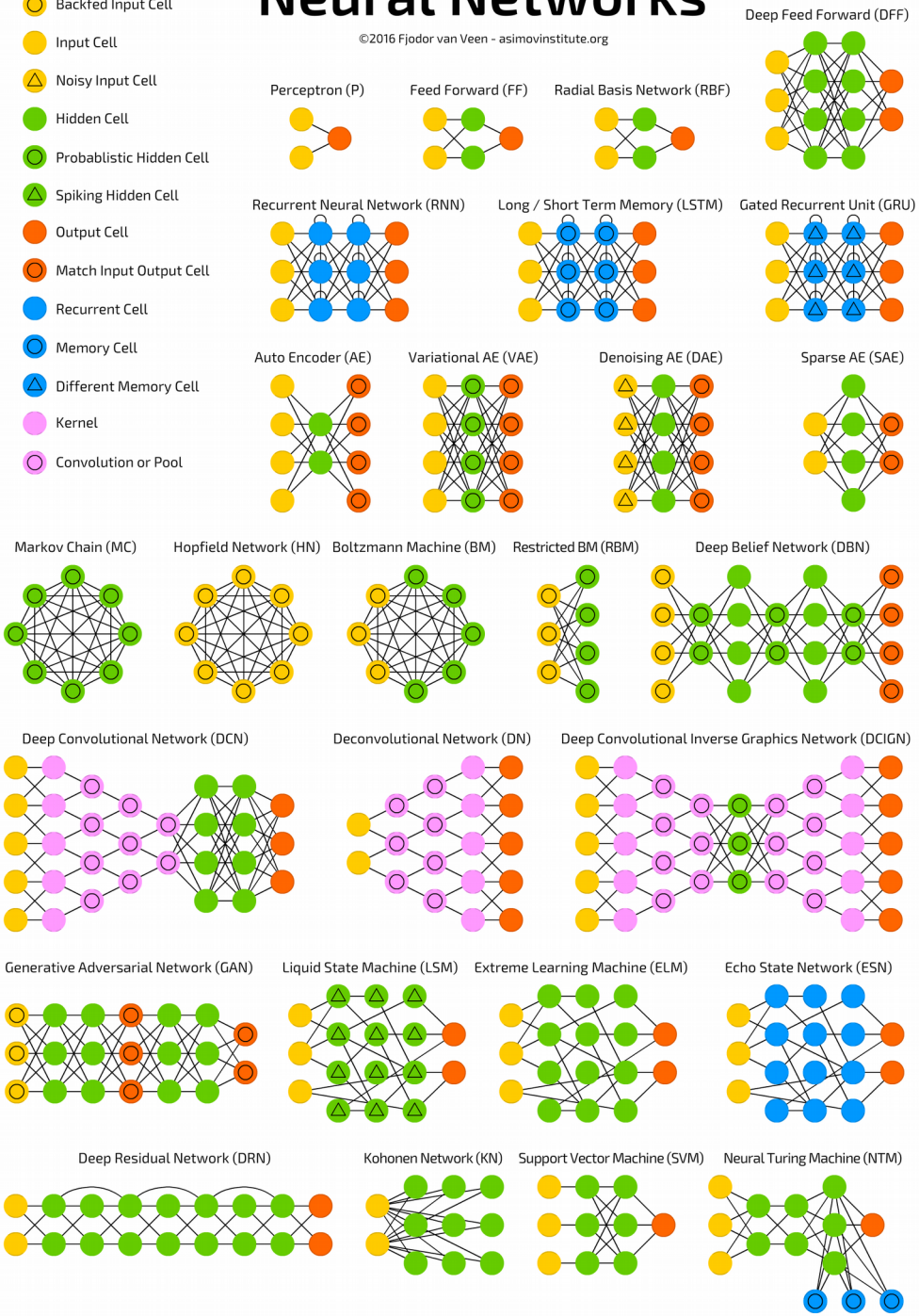
First layer: RBMs
(edges)

Input image patch
(pixels)

A mostly complete chart of

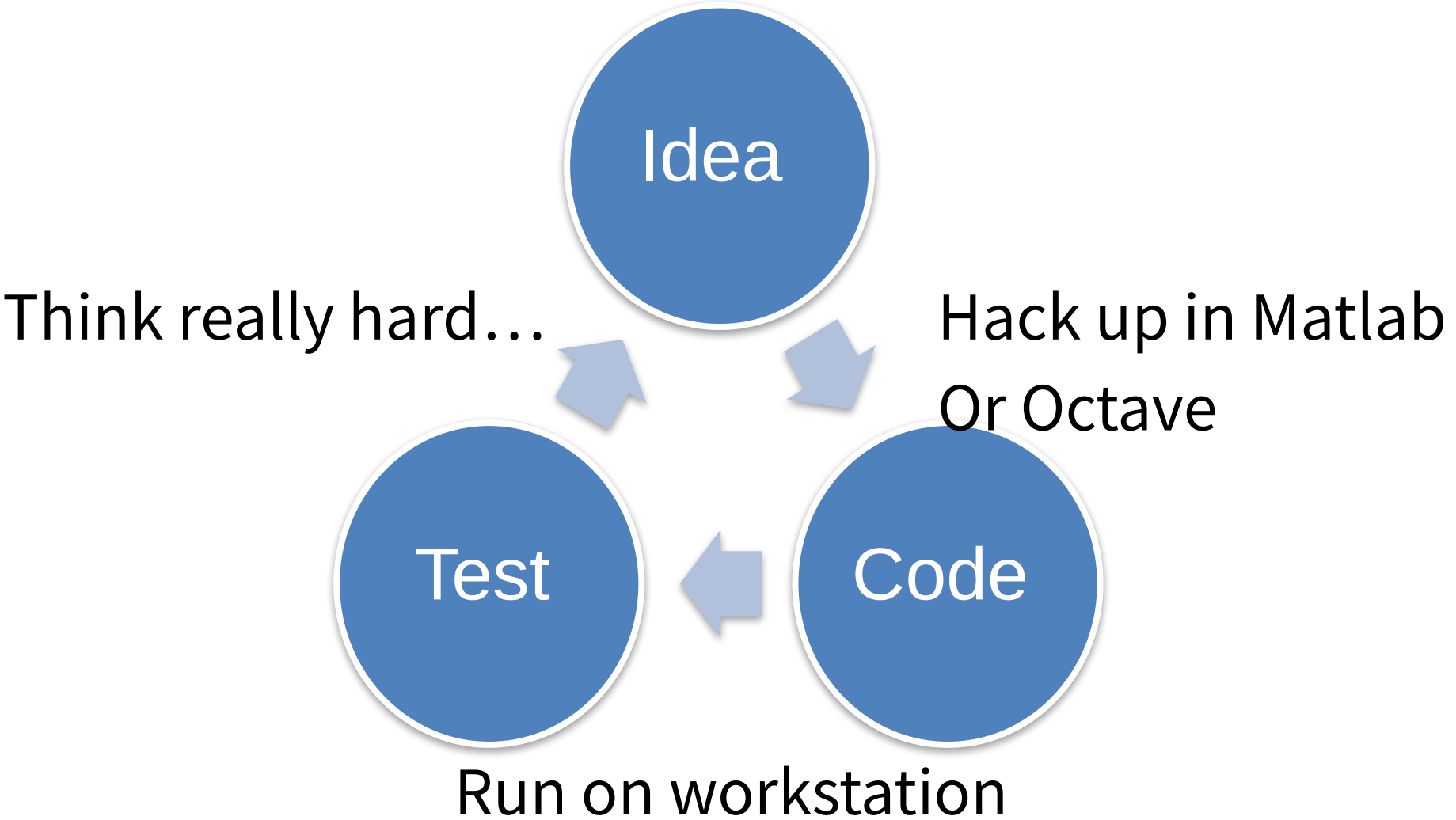
Neural Networks

- Backfed Input Cell
- Input Cell
- Noisy Input Cell
- Hidden Cell
- Probablistic Hidden Cell
- Spiking Hidden Cell
- Output Cell
- Match Input Output Cell
- Recurrent Cell
- Memory Cell
- Different Memory Cell
- Kernel
- Convolution or Pool



AI in practice

- Enormous amounts of research time spent inventing new features.

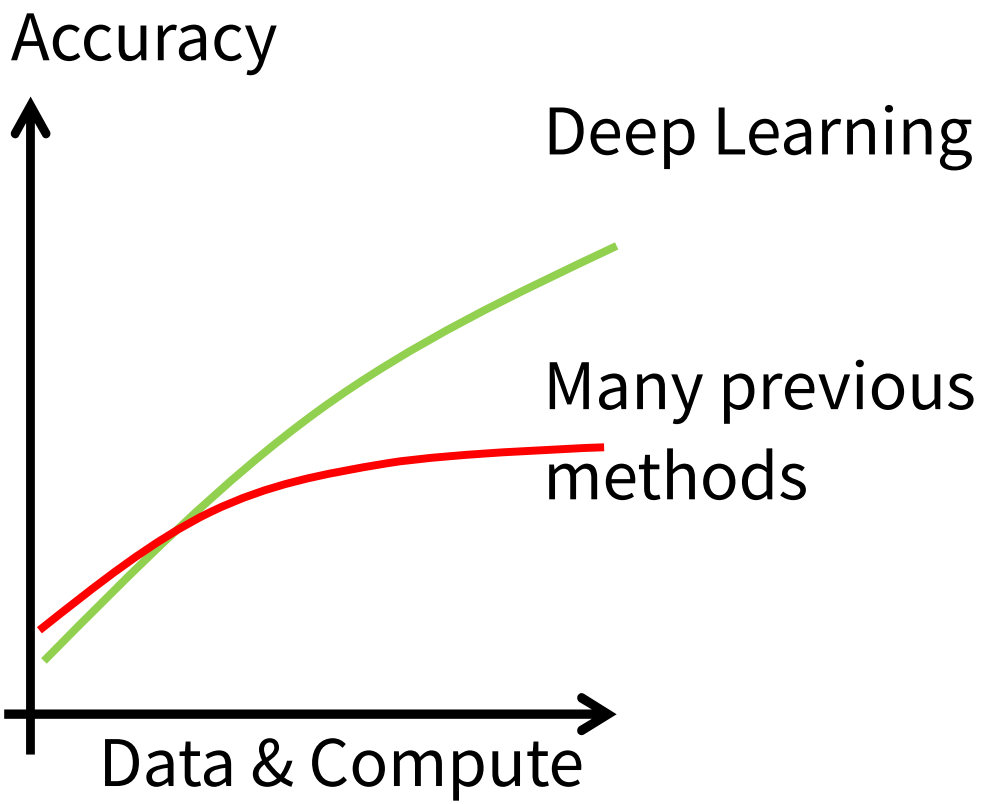


Why Deep Learning?

1. Scale Matters
   - Bigger models usually win

2. Data Matters
   - More data means less cleverness necessary

3. Productivity Matters
   - Teams with better tools can try out more ideas

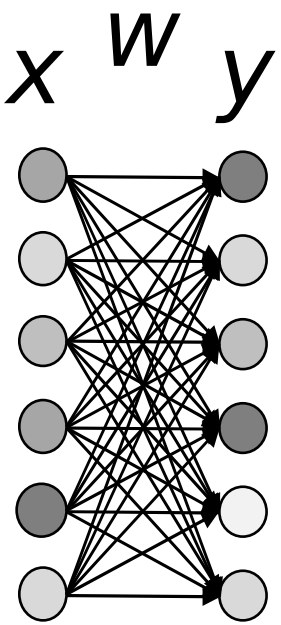


Scaling up

- Make progress on AI by focusing on systems
  - Make models bigger
  - Tackle more data
  - Reduce research cycle time
    - Accelerate large-scale experiments

Training Deep Neural Networks

$$y_j = f\left(\sum_i w_{ij}\, x_i\right)$$



- Computation dominated by dot products
- Multiple inputs, multiple outputs, batch means GEMM
  - Compute bound
- Convolutional layers even more compute bound

Computational Characteristics

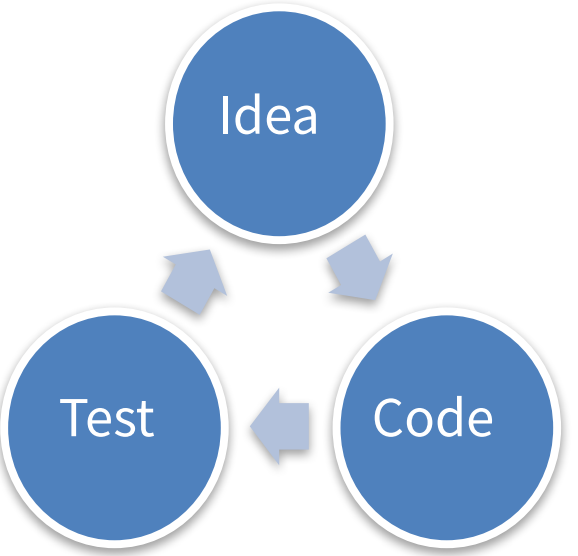
- High arithmetic intensity
  - Arithmetic operations / byte of data
  - O(Exaflops) / O(Terabytes) : 10^6
    - In contrast, some other ML training jobs are O(Petaflops)/O(Petabytes) = 10^0

- Medium size datasets
  - Generally fit on 1 node
  - HDFS, fault tolerance, disk I/O not bottlenecks

Training 1 model: ~20 Exaflops

Deep Neural Network training is HPC

- Turnaround time is key
- Use most efficient hardware
  - Parallel, heterogeneous computing
  - Fast interconnect (PCIe, Infiniband)
- Push strong scalability
  - Models and data have to be of commensurate size



Infrastructure

- Software: CUDA, MPI, Majel (internal library)
- Hardware:

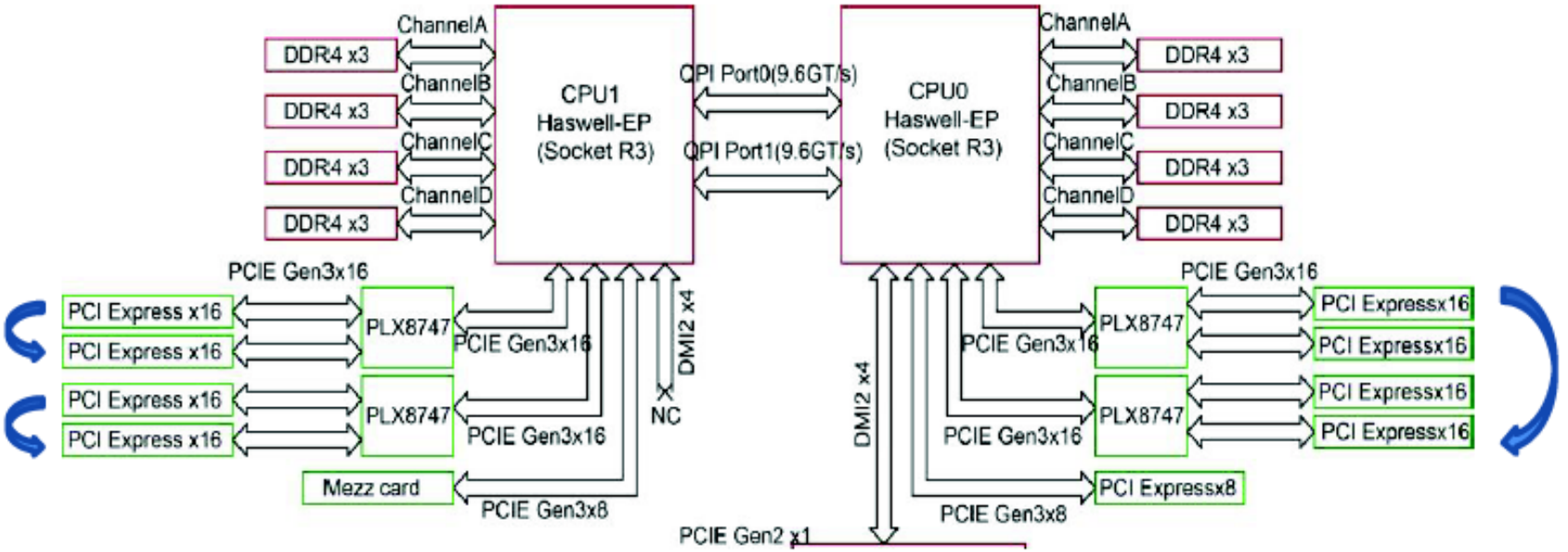
NVIDIA GeForce GTX Titan X

Titan X x8

Mellanox Interconnect

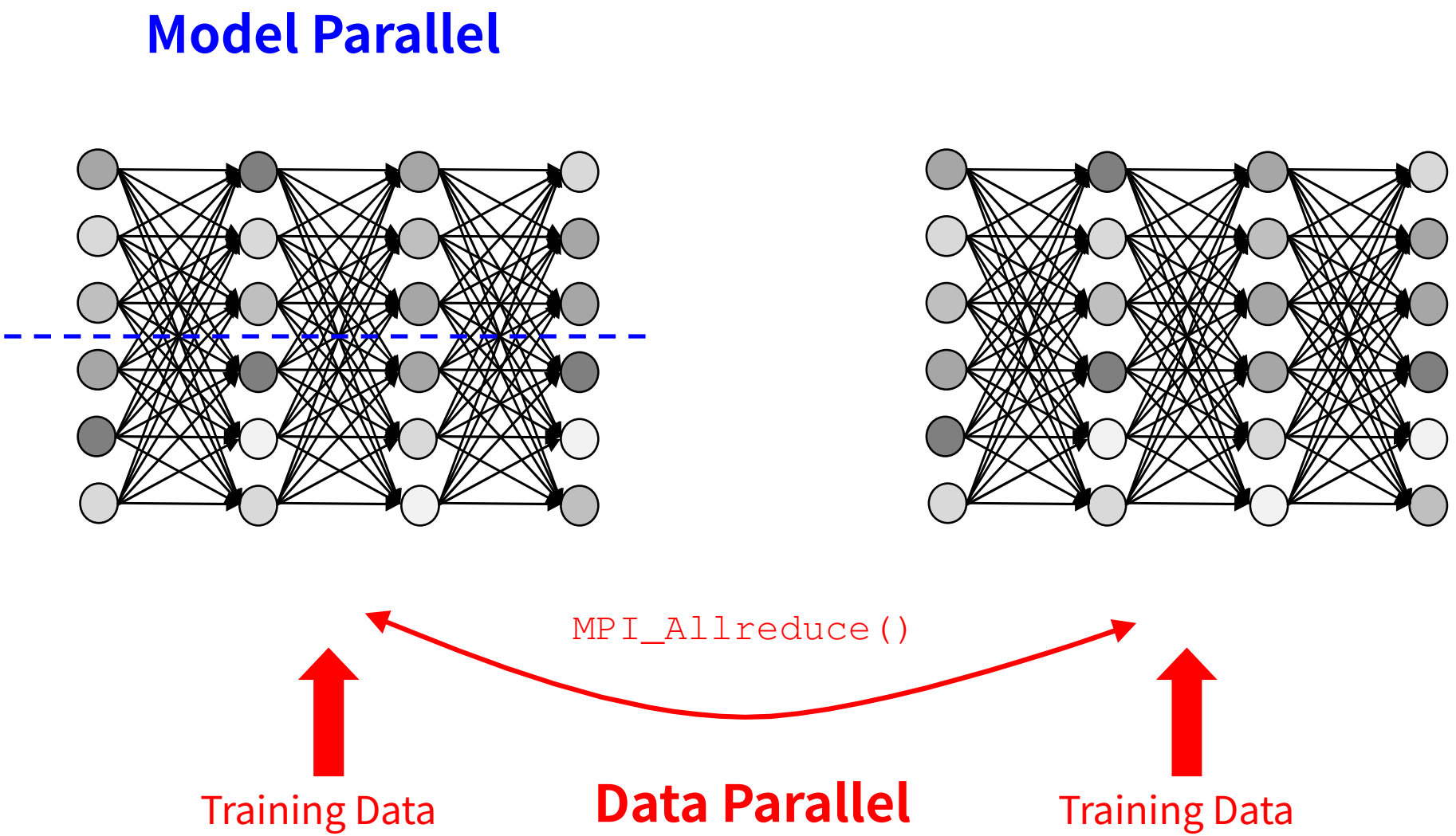
http://www.tyan.com

Node Architecture



- All pairs of GPUs communicate simultaneously over PCIe Gen 3 x16
- Groups of 4 GPUs form Peer to Peer domain
- Avoid moving data to CPUs or across QPI

Parallelism



Determinism

- Determinism very important
- So much randomness,
  hard to tell if you have a bug
- Networks train despite bugs,
  although accuracy impaired
- Reproducibility is important
  - For the usual scientific reasons
  - Progress not possible without reproducibility

Conclusions

- Computationally dense processors (like GPUs) required
- Programmability
  - We don't know the algorithms of the future
- Lower precision
  - But not too low
  - Interesting algorithm/dataset engineering here
- We need better support for multi-GPU
  - E.g. Atomics between GPUs, collectives
  - Looking forward to NVLink

Conclusions

- Deep Learning is solving many hard problems
- Training deep neural networks is an HPC problem
- Scaling brings AI progress!

How can we help fintechs?

- Discover regulatory and compliance issues before violations happen, preventing fines and enforcement investigations.
- Advanced pattern recognition detects activity likely to be high-risk, flagging hidden threats well before disaster strikes.
- Examines profitability, cost drivers, and industry-specific risk factors to incentivize behaviors and improve firm performance.
- Finds outlier events, by understanding historical trends, environmental context and correlation within firm activity.
- We show financial institutions the true power of artificial intelligence by delivering clearly prioritized calls to action. This empowers decision makers to solve challenging business problems.

How can we help fintechs?

- SPOOFING SIMILARITY
- ABUSIVE MESSAGING DETECTION
- MOMENTUM IGNITION DETECTION
- SUSPICIOUS PRICE MOVEMENT DETECTION
- PINGING AND PHISHING DETECTION
- WASH TRADE DETECTION
- CROSS TRADE DETECTION
- CLOSING PERIOD ABUSE DETECTION