Introduction to Data Mining

Instructor: Alex Thomo
http://webhome.cs.uvic.ca/~thomo

Data Deluge

- 2012
  - every day 2.5 quintillion bytes of data (1 followed by 18 zeros) were created,

- 90% of the world's data created in the last two years alone.
  - More data produced each day than was seen by everyone since the beginning of the earth.

Largest databases

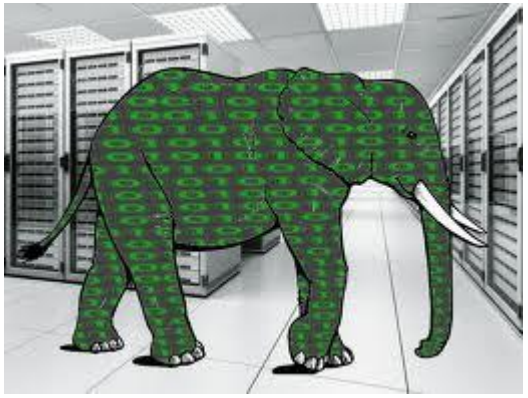
- Largest database: World Data Centre for Climate

  Max Planck Institute and *German Climate Computing Centre*

  - *220 terabytes of data on climate research and climatic trends*
  - *110 terabytes worth of climate simulation data*
  - *6 petabytes worth of additional information stored on tapes*

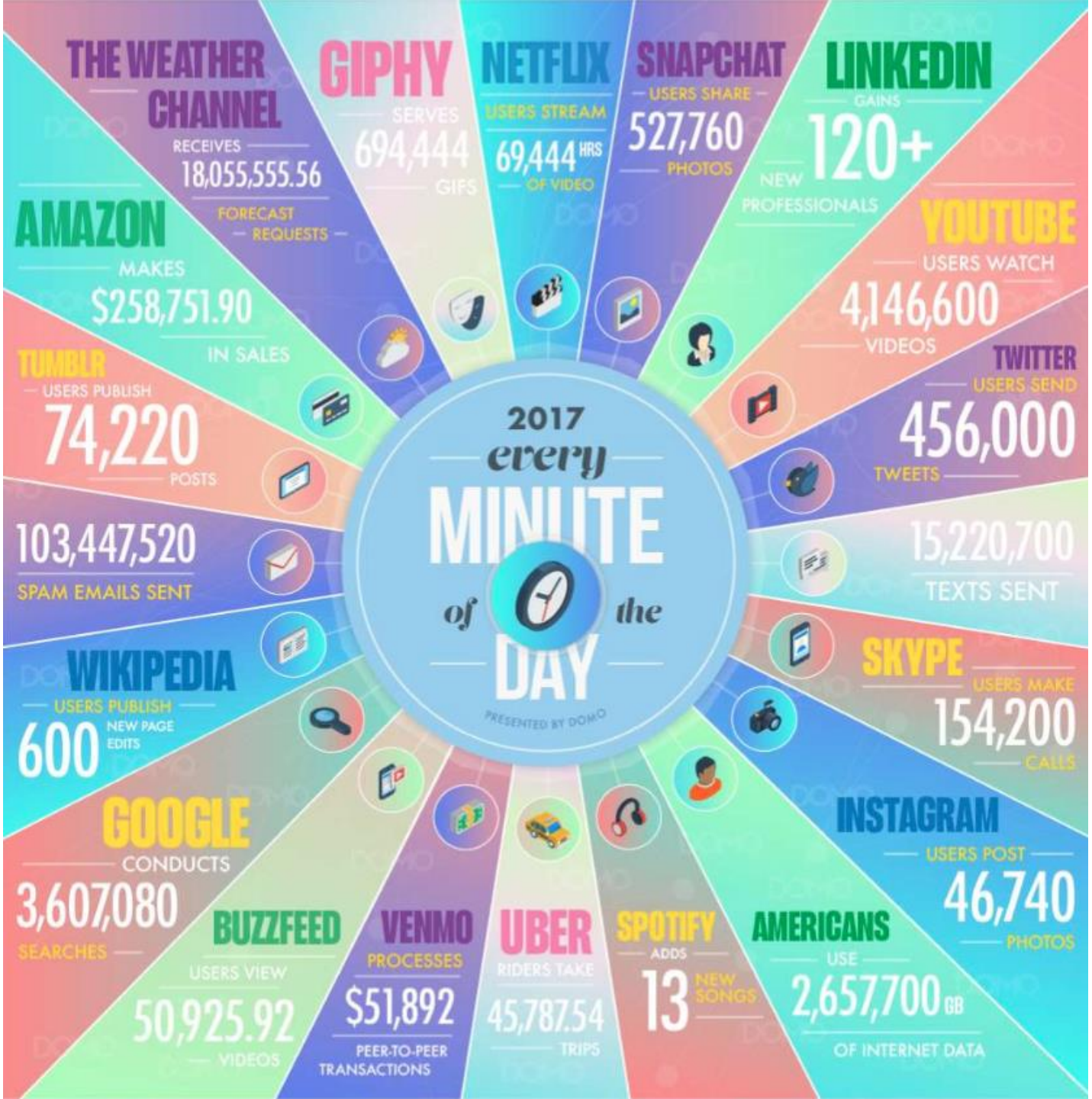
- AT&T
  - *323 terabytes of information*
  - *1.9 trillion phone call records*

- Google
  - *91 million searches per day*
  - *33 trillion database entries a year*



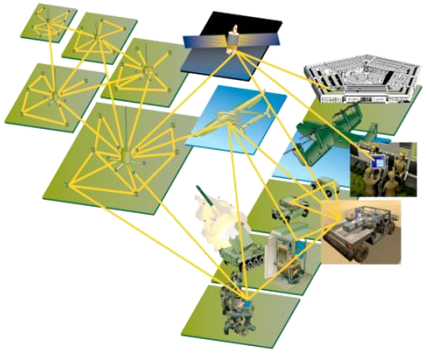
Social Media Data

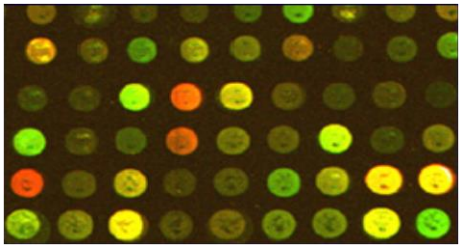
- 144.8 billion emails a day.
- 340 million tweets on Twitter a day.
- 2.5 billion content items shared per day on Facebook.
- 72 hours of new video to YouTube a minute.
- 3125 new photos uploads to Flickr a minute.
- 350 new blog posts on WordPress a minute.

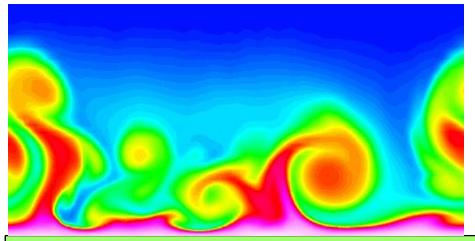

2017
every
MINUTE
of the
DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL
RECEIVES
18,055,555.56
FORECAST REQUESTS
GIPHY
SERVES
694,444
GIFS
NETFLIX
USERS STREAM
69,444 HRS
OF VIDEO
SNAPCHAT
USERS SHARE
527,760
PHOTOS
LINKEDIN
GAINS
120+
NEW PROFESSIONALS
AMAZON
MAKES
$258,751.90
IN SALES
YOUTUBE
USERS WATCH
4,146,600
VIDEOS
TUMBLR
USERS PUBLISH
74,220
POSTS
TWITTER
USERS SEND
456,000
TWEETS
103,447,520
SPAM EMAILS SENT
15,220,700
TEXTS SENT
WIKIPEDIA
USERS PUBLISH
600
NEW PAGE EDITS
SKYPE
USERS MAKE
154,200
CALLS
GOOGLE
CONDUCTS
3,607,080
SEARCHES
INSTAGRAM
USERS POST
46,740
PHOTOS
BUZZFEED
USERS VIEW
50,925.92
VIDEOS
VENMO
PROCESSES
$51,892
PEER-TO-PEER TRANSACTIONS
UBER
RIDERS TAKE
45,787.54
TRIPS
SPOTIFY
ADDS
13
NEW SONGS
AMERICANS
USE
2,657,700 GB
OF INTERNET DATA

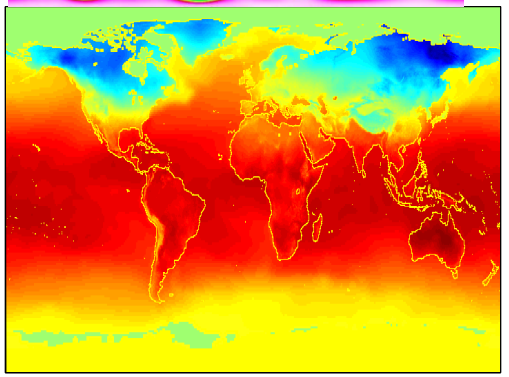
Scientific Data

- Data collected and stored at enormous speeds (GB/hour). E.g.
  - remote sensors on a satellite
  - telescopes scanning the skies
  - scientific simulations generating terabytes of data



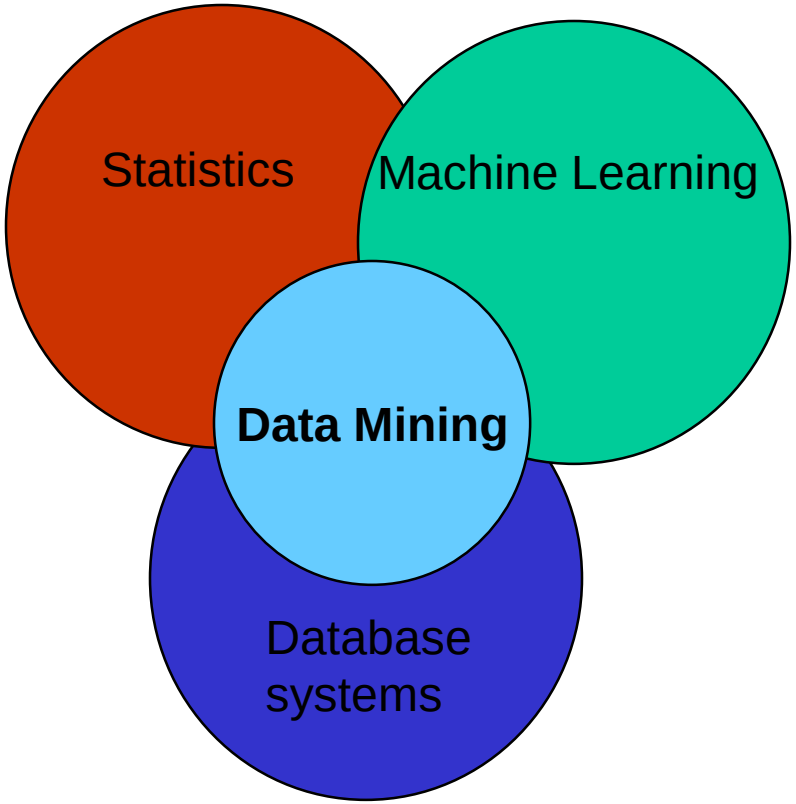






Origins of Data Mining

- Draws ideas from: machine learning/AI, statistics, and database systems



Data Mining Tasks

Data mining tasks are generally divided into two major categories:

- Predictive tasks [Use some attributes to predict unknown or future values of other attributes.]
  - Classification
  - Regression
- Descriptive tasks [Find human-interpretable patterns that describe the data.]
  - Association Discovery
  - Clustering

Predictive Data Mining or Supervised *machine learning*

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.

- Find ("learn") a *model* for the class attribute as a *function* of the values of the other attributes.

- Goal: Assign a class to **previously unseen** records as accurately as possible.

Classification: Fraud Detection

Goal: Predict fraudulent cases in credit card transactions.

Approach:

- Collect data about past transactions
  - **when** does a customer buy,
  - **what** does he buy,
  - **where** does he buy, etc.
- Label some past transactions as **fraud** or **fair** transactions.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

Classification: Direct Marketing

Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new product.

Approach:

- Use the data for a **similar product** introduced before.
  - We know which customers decided to buy and which decided otherwise.
- Collect various **demographic**, **lifestyle**, and other related information about customers.
- Learn a **classifier model**. Use the model to predict whether a customer is likely to adopt the product.
- Send mail only to those predicted as likely.

Finding Associations

The Market-Basket Model

- A large set of *items*, e.g., things sold in a supermarket.
- A large set of *baskets*, e.g., the things one customer buys on one day.

Fundamental problem

- Learn sets of items that are **often bought together**.

Example of an application…

- If a large number of baskets contain both hot dogs and mustard, we can use this information in several ways. How?

On-Line Purchases

amazon.com

- Amazon.com: several million different items for sale, and several tens of millions of customers.

- Basket = Customer,
- Item = Book, DVD, etc.
  - Motivation: Learn what items are bought together.

- Basket = Book, DVD, etc.
- Item = Customer
  - Motivation: Find out similar customers.

- Result: Use for recommender systems.

Words and Documents

- Baskets = sentences;
- Items = words in those sentences.
  - Words that appear together frequently suggest **linked concepts**.

- Baskets = sentences,
- Items = documents containing those sentences.
  - Items that appear together too often could represent **plagiarism**.

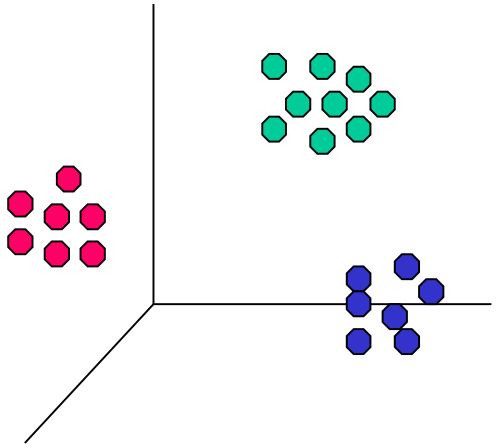
Clustering

- Given a set of data points, find **clusters** such that
  - Data points in **one** cluster are **more similar** to one another.
  - Data points in **separate** clusters are **less similar** to one another.

☒*E.g. Euclidean Distance Based Clustering in 3-D space.*

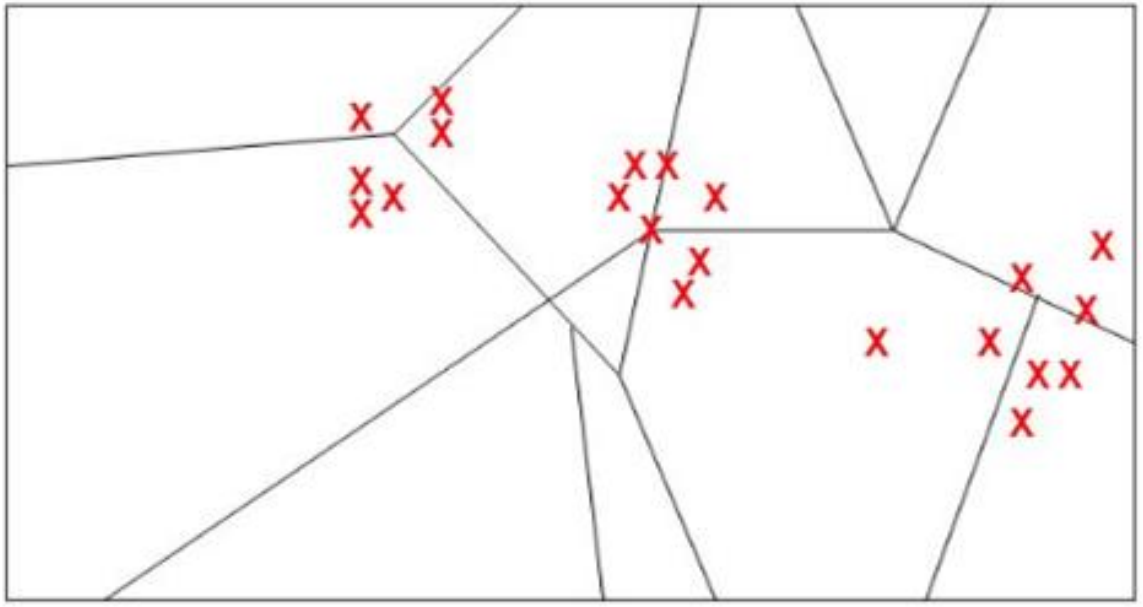
Intracluster distances are minimized

Intercluster distances are maximized



Example: a cholera outbreak in London

Many years ago, during a cholera outbreak in London, a physician plotted the location of cases on a map. Properly visualized, the data indicated that cases clustered around certain intersections, where there were polluted wells, not only exposing the cause of cholera, but indicating what to do about the problem.



Clustering: Application 1

- Market Segmentation:
  - **Goal**: divide market into **distinct subsets** of customers. **Target** each subset with a **distinct marketing campaign**.
  - **Approach**:
    - Collect different attributes of customers based on their **geographical** and **lifestyle** related information.
    - Find clusters of similar customers.

Clustering: Application 2

- Document Clustering:
  - **Goal**: Find **groups** of documents that are similar to each other based on important words appearing in them.
  - **Approach**:
    - Identify frequently occurring **words** in each document.
    - Form a **similarity measure** based on the **word frequencies**. Use it to cluster.

Outline

- Topics:
  - Predictive data mining
  - Data Analytics
  - Visualization
  - Association Analysis
  - Clustering
  - Web mining
  - Recommender Systems

Tools

- Weka

- Python ecosystem
  - Numpy
  - Matplotlib
  - Seaborn
  - Pandas
  - Scikit-learn

- Languages
  - Java
  - Python

CARTOON INTRODUCTION TO SOME ALGORITHMS OF PREDICTIVE DATA MINING (SUPERVISED MACHINE LEARNING)

Source: https://www.youtube.com/watch?v=IpGxLWOIZy4

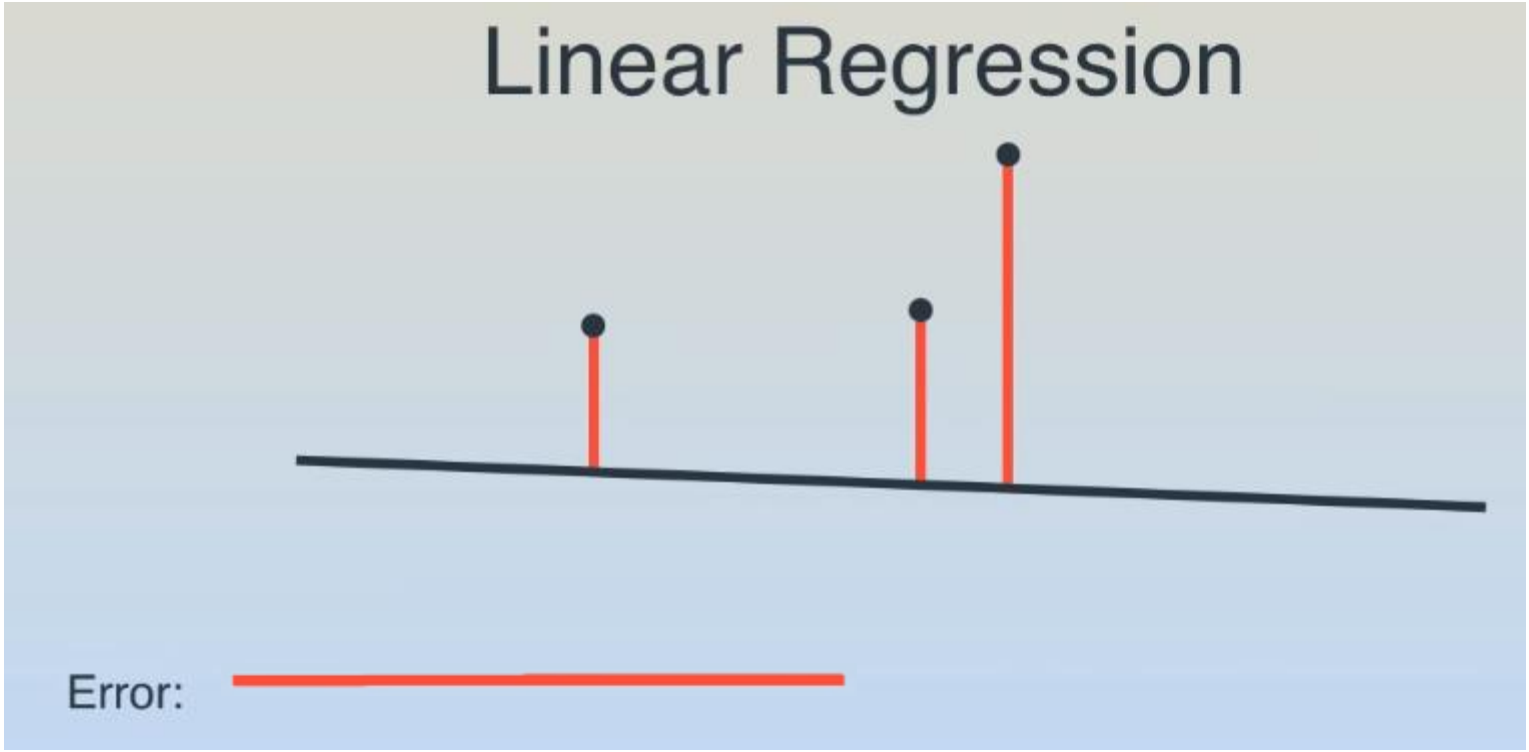
What's Machine Learning

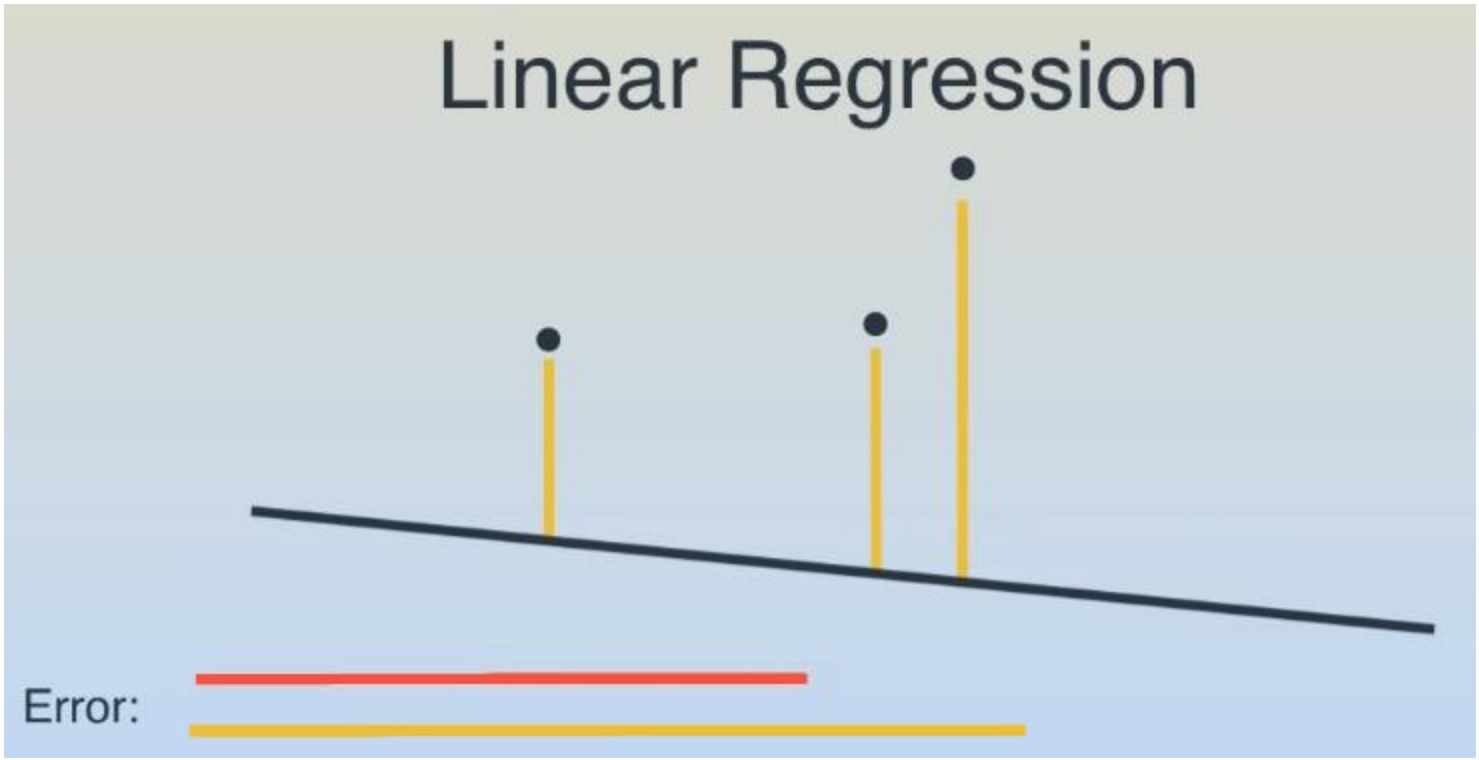
Learn from experience

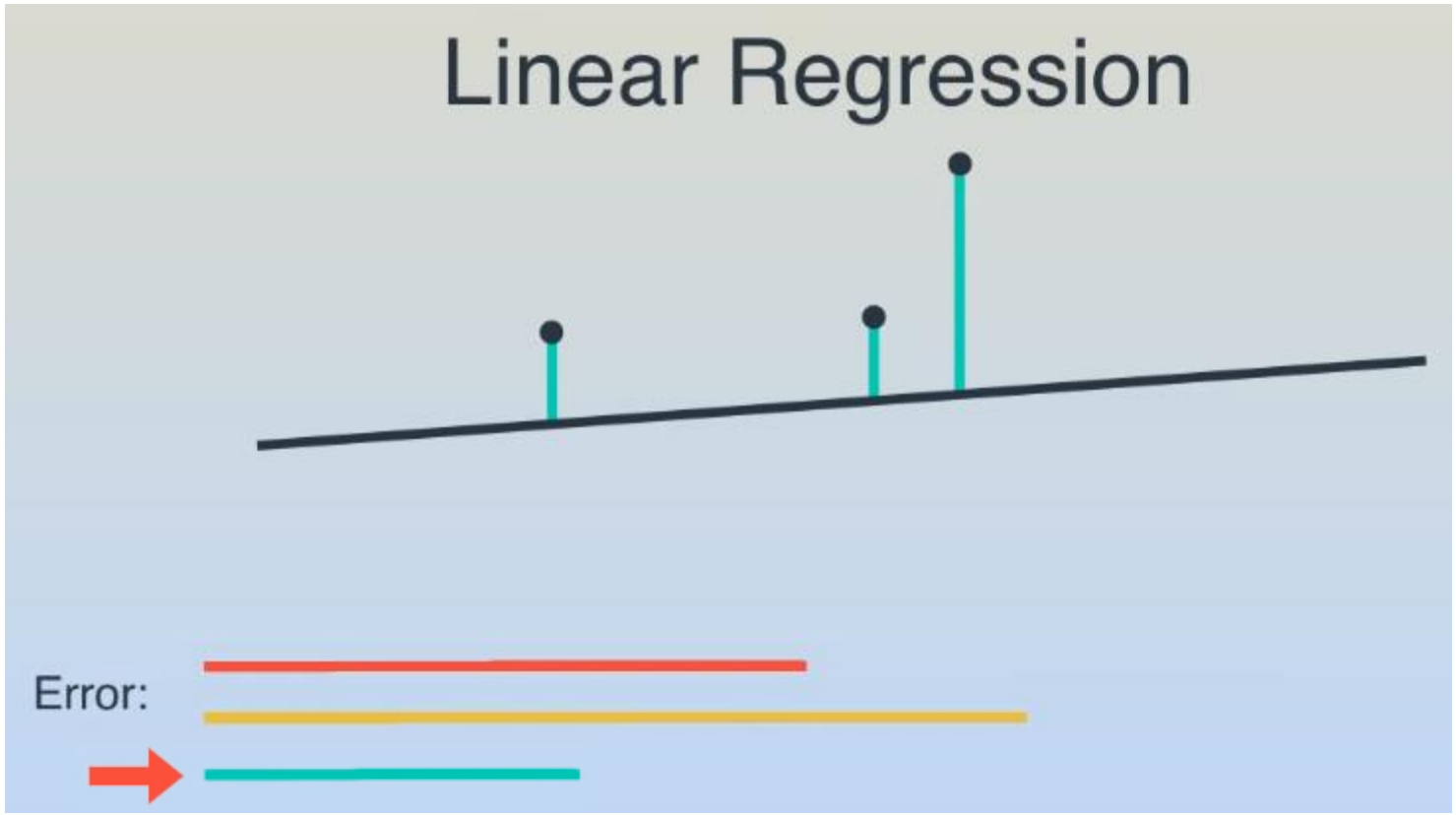
Learn from ~~experience~~ data

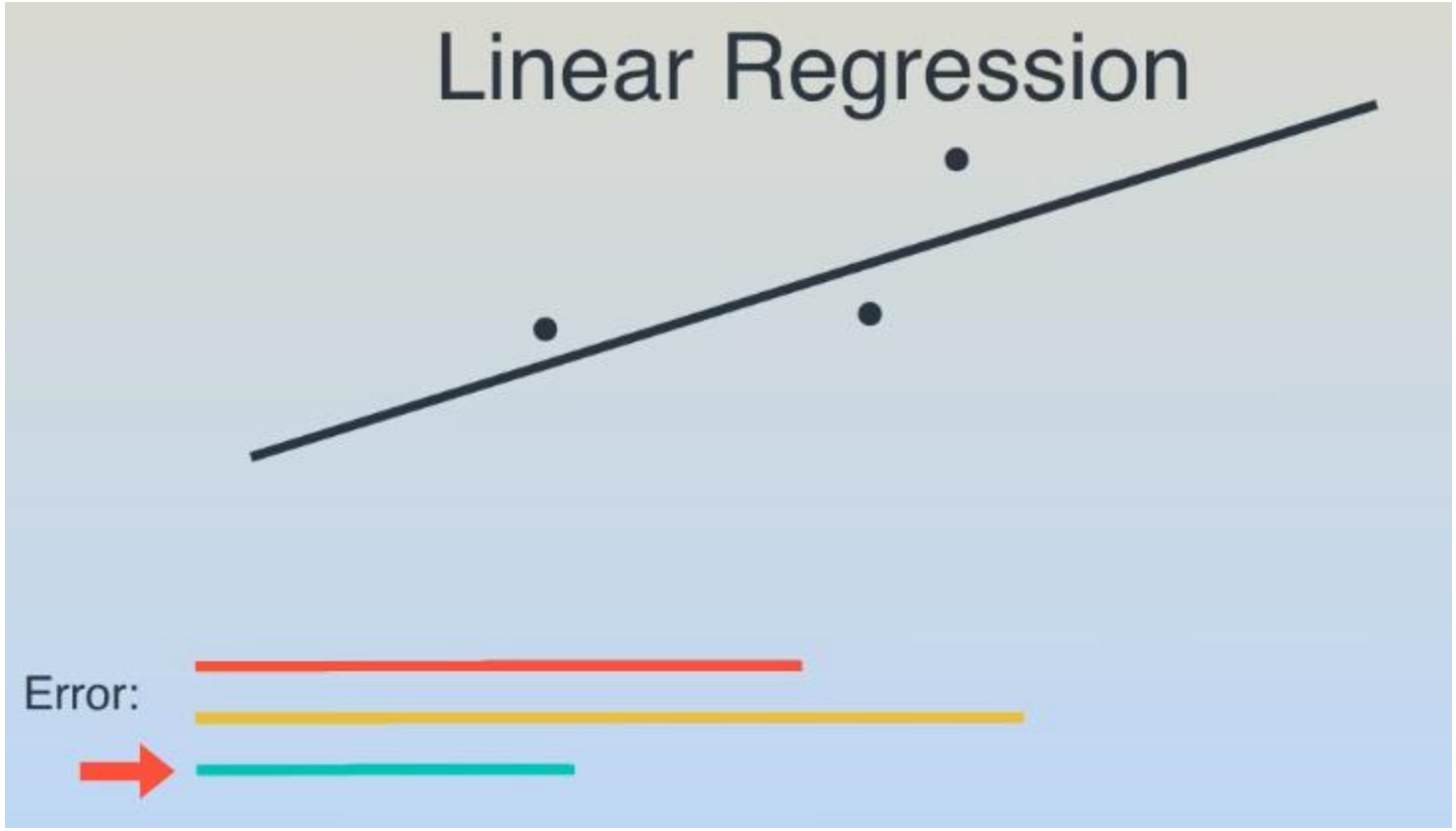
Follow instructions


Price of a house
Price ($10,000)
20
15
10
5
?
12

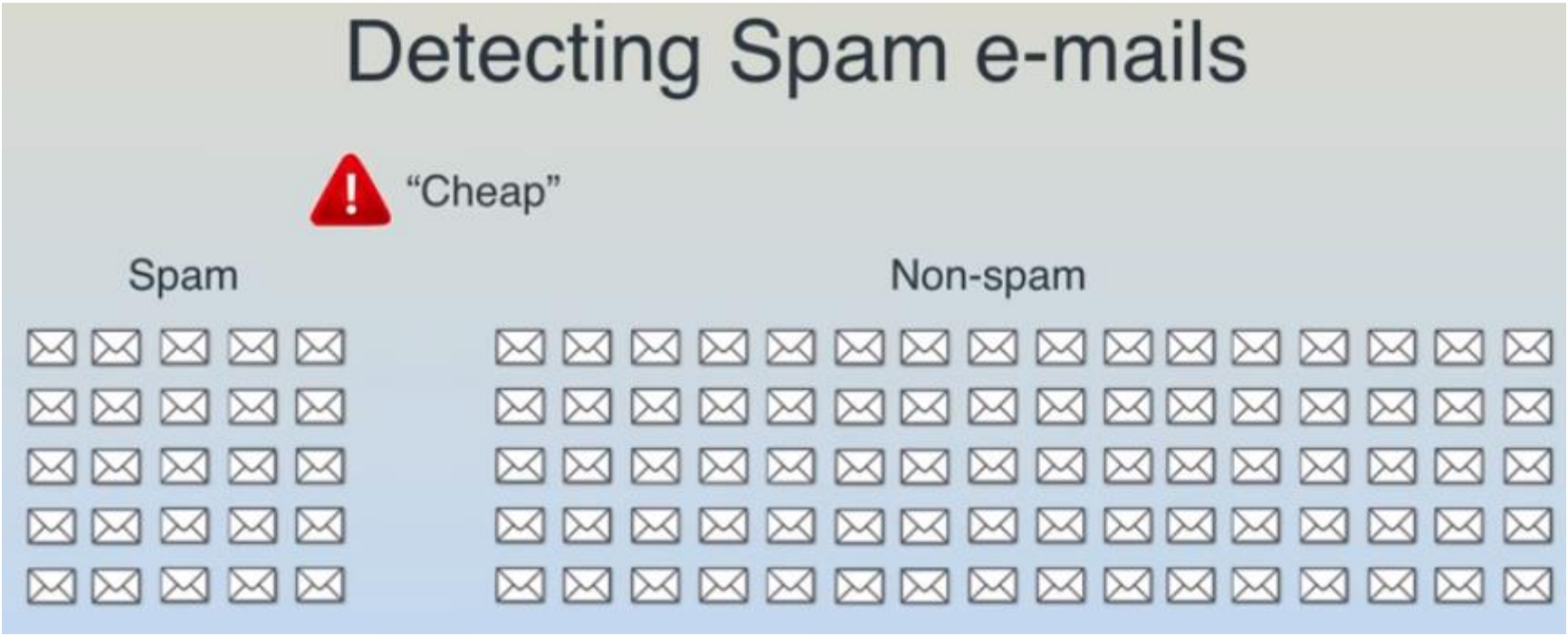

Linear Regression
Error:


Linear Regression
Error:

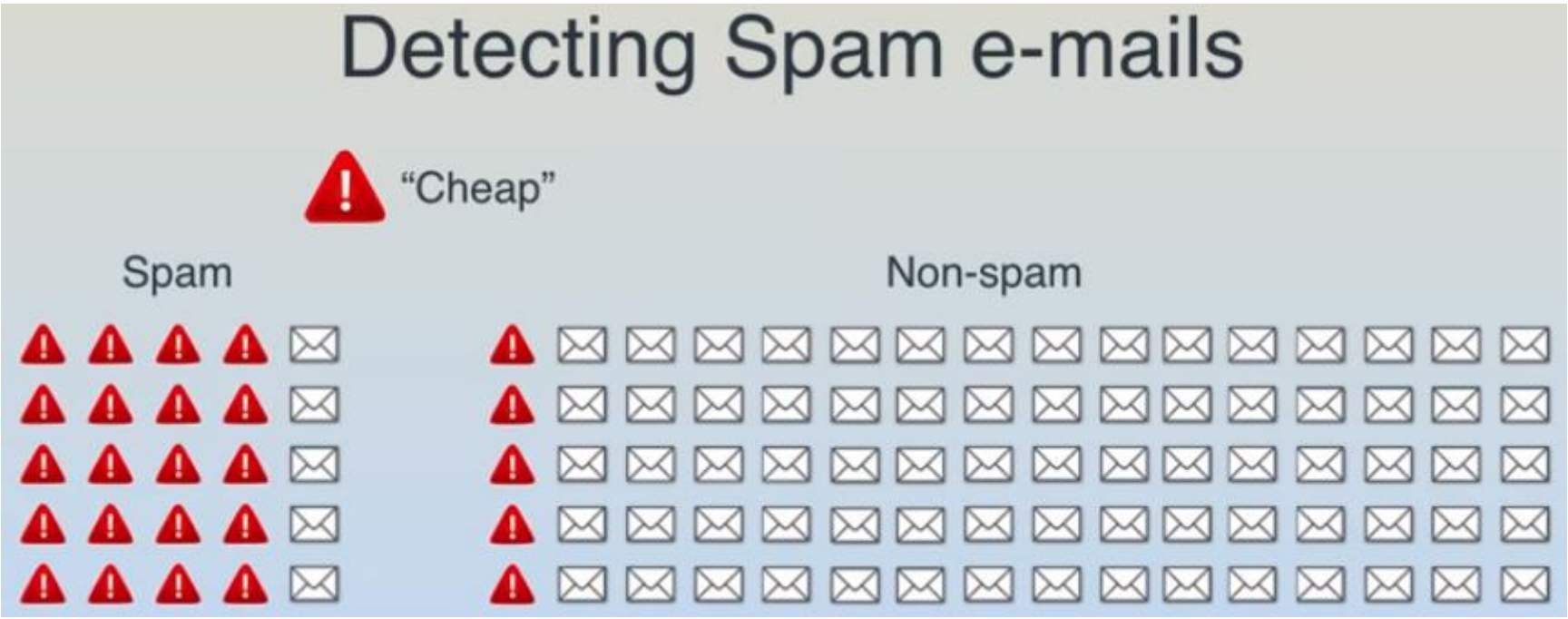

Linear Regression
Error:


Linear Regression
Error:

Detecting Spam e-mails



Detecting Spam e-mails



Detecting Spam e-mails

"Cheap"

Spam

Non-spam

Quiz: If an e-mail contains the word "cheap", what is the probability of it being spam?

- ○ 40%
- ○ 60%
- ○ 80%

Naive Bayes Algorithm


"Cheap" ⟶ 80%


Spelling mistake → 70%


Missing title ⟶ 95%


etc…

Quiz: If an e-mail contains the word "cheap", what is the probability of it being spam?

- ○ 40%
- ○ 60%
- ● 80%

Conclusion:
If the e-mail contains the word "cheap", The probability of it being spam is 80%

Recommending Apps

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Recommending Apps

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Quiz: Between Gender and Age, which one seems more decisive for predicting what app will the users download?

- ○ Gender
- ○ Age

Recommending Apps

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Quiz: Between Gender and Age, which one seems more decisive for predicting what app will the users download?

- ○ Gender
- ● Age

Recommending Apps

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Recommending Apps

Gender	Age	App
F	25	
M	32	
F	40	

Acceptance at a University

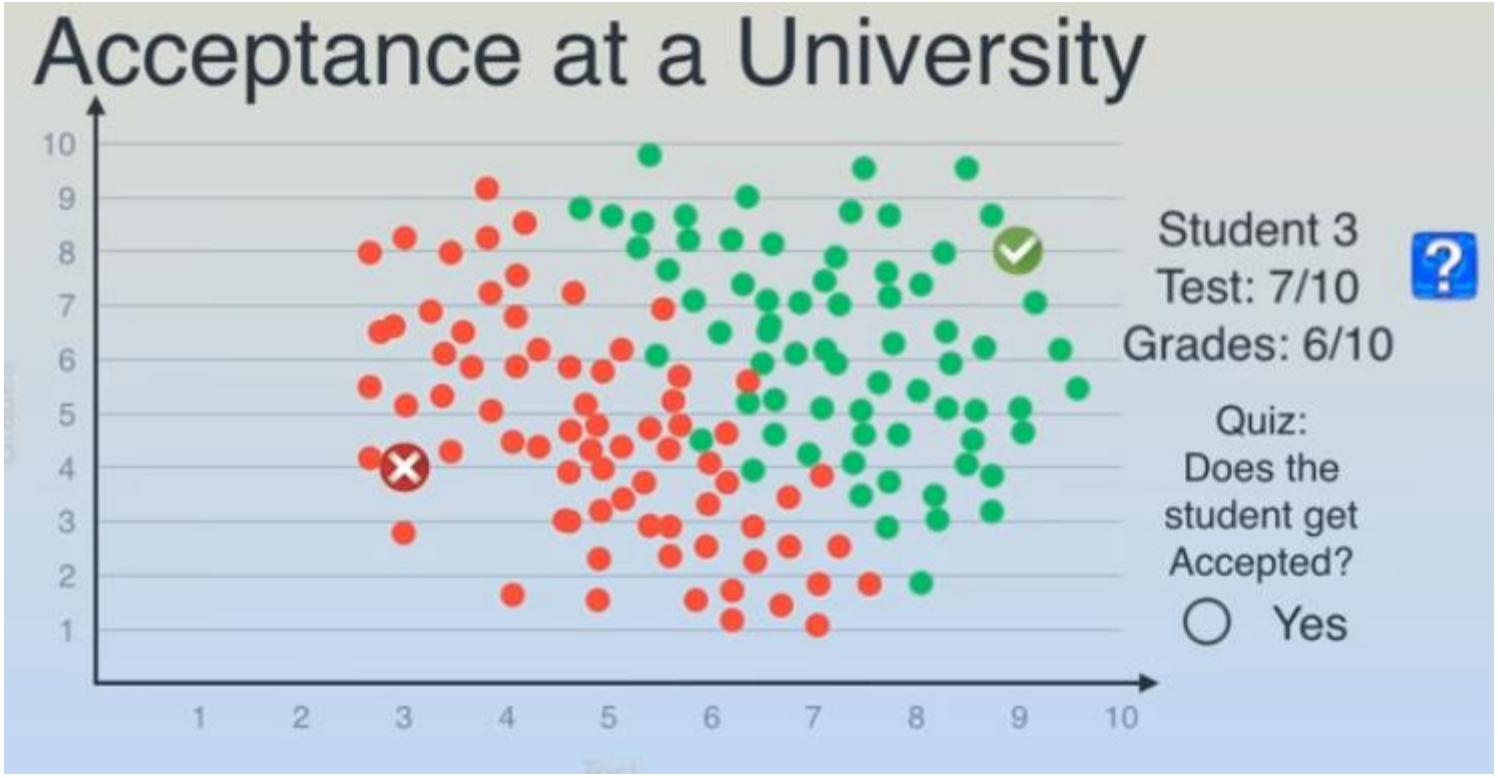
Test

Grades

Student 1
Test: 9/10
Grades: 8/10

Student 2
Test: 3/10
Grades: 4/10

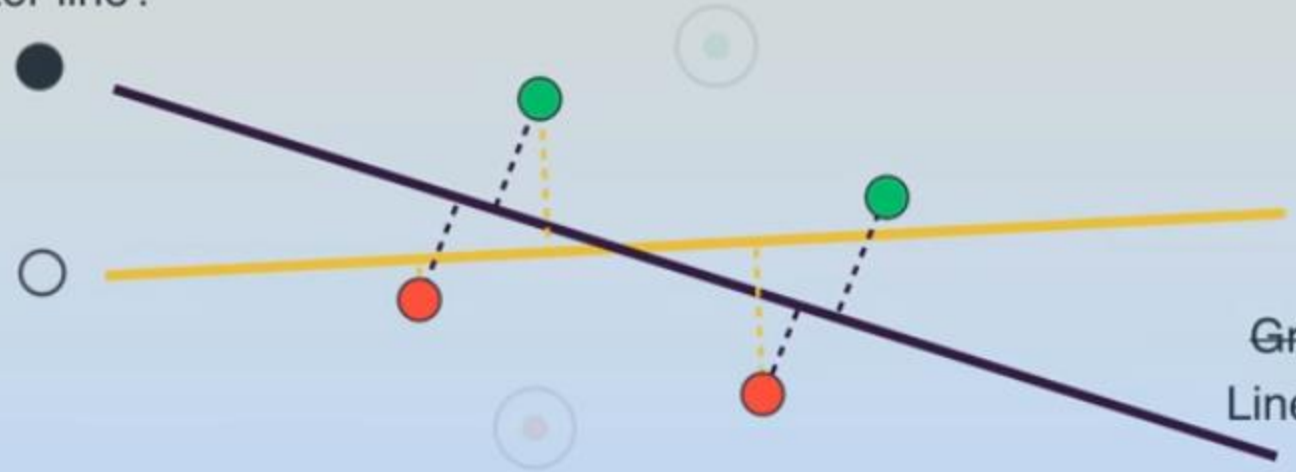
Student 3
Test: 7/10
Grades: 6/10


Acceptance at a University
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
Student 3
Test: 7/10
Grades: 6/10
Quiz:
Does the
student get
Accepted?
Yes


Logistic Regression
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
?
Student 3
Test: 7/10
Grades: 6/10
Quiz:
Does the
student get
Accepted?
Yes

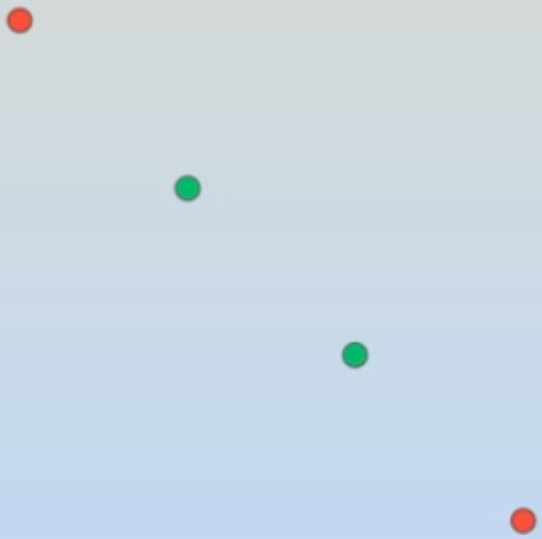
Support Vector Machine

Quiz:
Which one is a better line?



~~Gradient descent~~
Linear Optimization

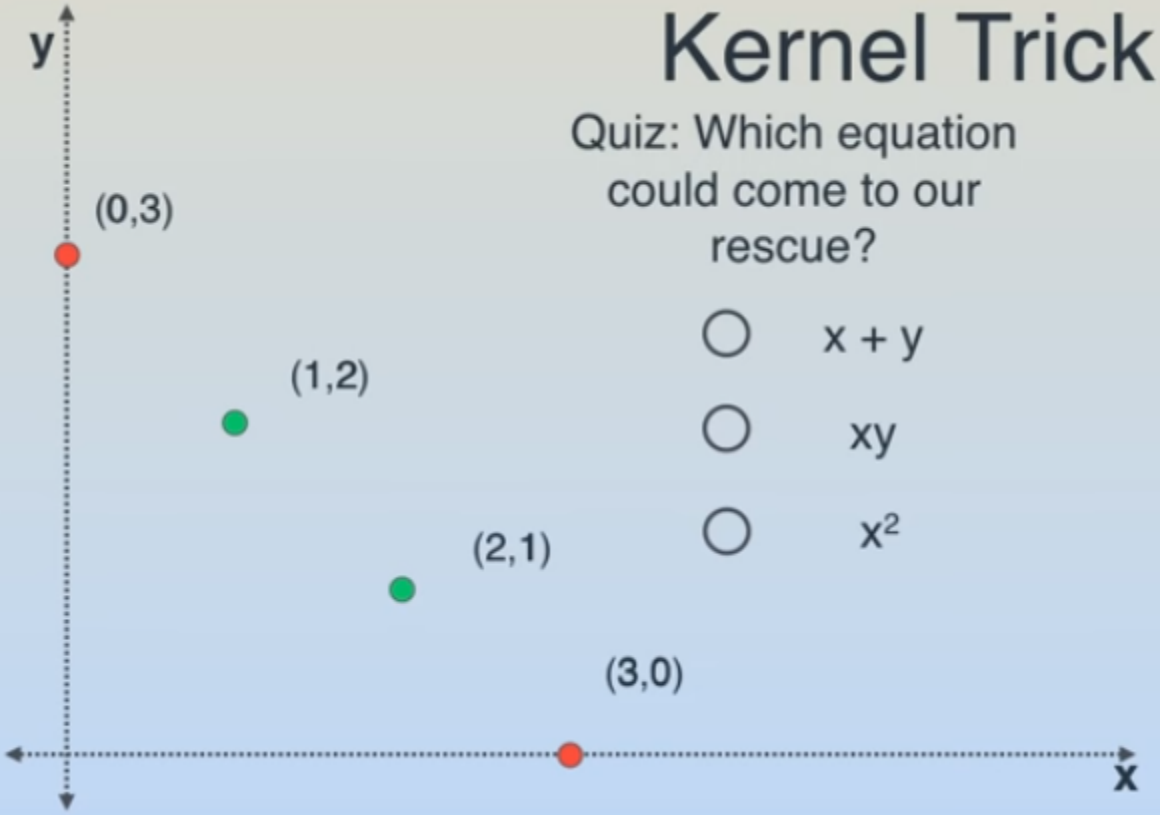
When a line is not enough…



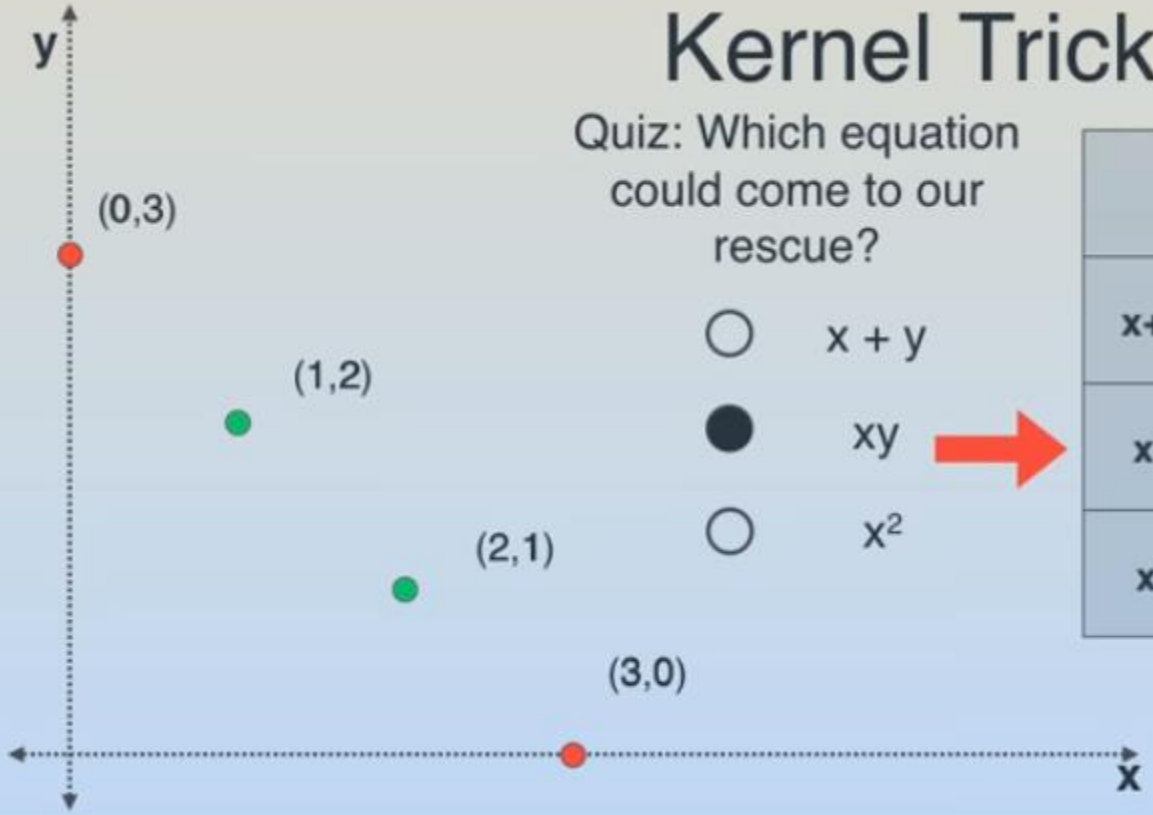
Kernel Trick

Quiz: Which equation could come to our rescue?

- ○ $x + y$
- ○ xy
- ○ x^2



Kernel Trick



Quiz: Which equation could come to our rescue?

- ○ x + y
- ● xy
- ○ x^2

	(0,3)	(1,2)	(2,1)	(3,0)
x+y	3	3	3	3
xy	0	2	2	0
x^2	0	1	4	9

Kernel Trick