# Intro to Data Science & Machine Learning

Jamal Madni

CECS 445

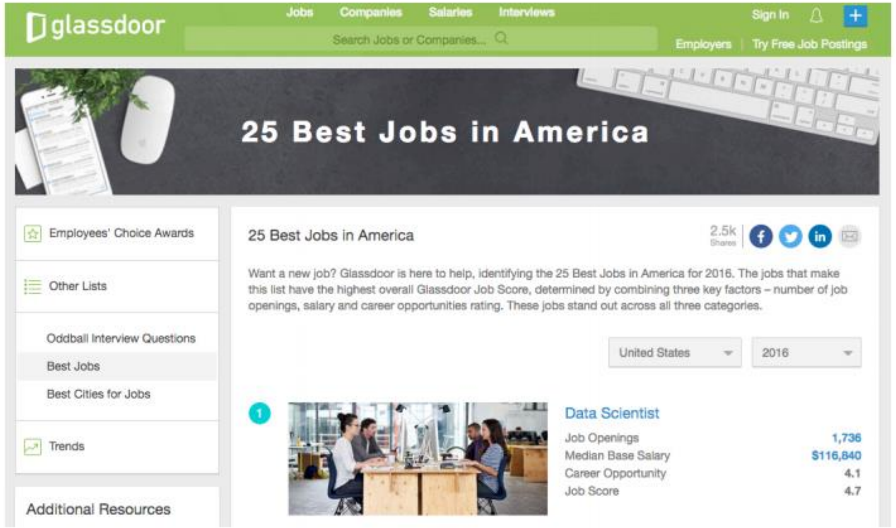
Lecture 16: April $8^{th}$, 2021

# What is Data Science?

Data science is the application of **computational** and **statistical** techniques to address or gain insight into some problem in the **real world**

Data science = statistics +
data processing +
machine learning +
scientific inquiry +
visualization +
business analytics +
big data + …

Data science is the best job in America



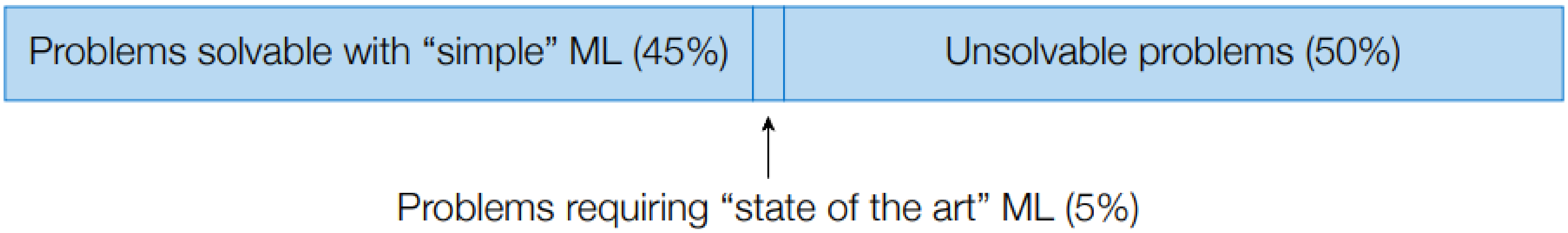
# Data Science is Not Machine learning

Machine learning involves computation and statistics, but has not (traditionally) been very concerned about answering *scientific questions*

Machine learning has a heavy focus on fancy algorithms…

... but sometimes the best way to solve a problem is just by visualizing the data, for instance

**Universe of machine learning problems**



# Data Science is Not Statistics

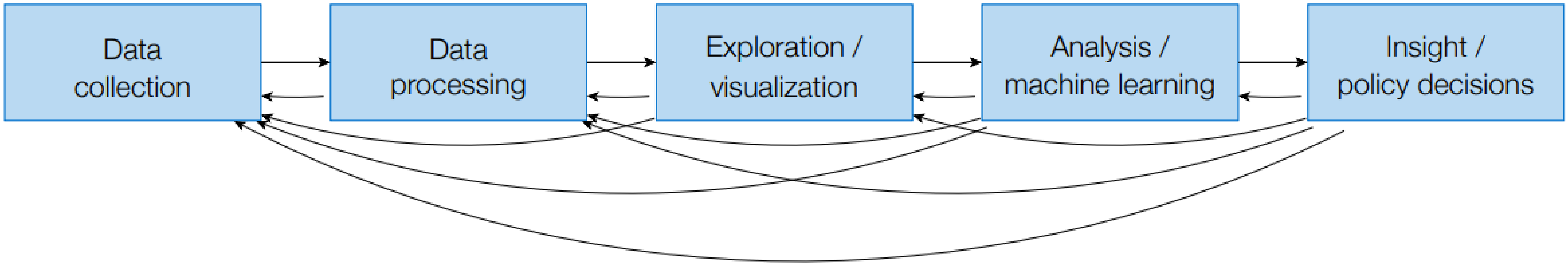
"Analyzing data computationally, to understand some phenomenon in the real world, you say? … that sounds an awful lot like statistics"

Statistics (at least the academic type) has evolved a lot more along the mathematical/theoretical frontier

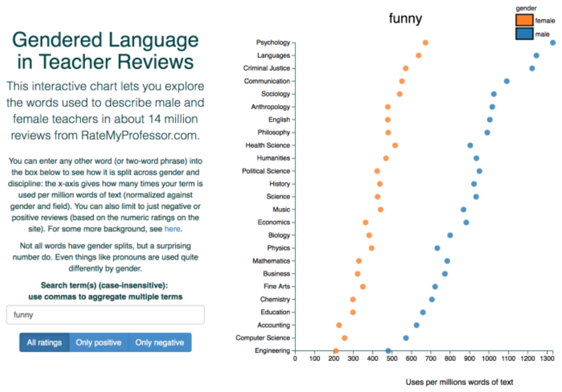
Not many statistics courses have a lecture on e.g. web scraping, or a lot of data processing more generally

Plus, statisticians use R, while data scientists use Python ... clearly these are completely different fields

# What is Data Science?



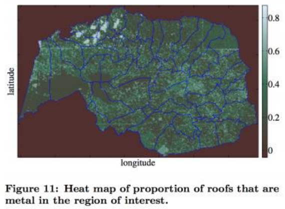

## Gendered language in professor reviews



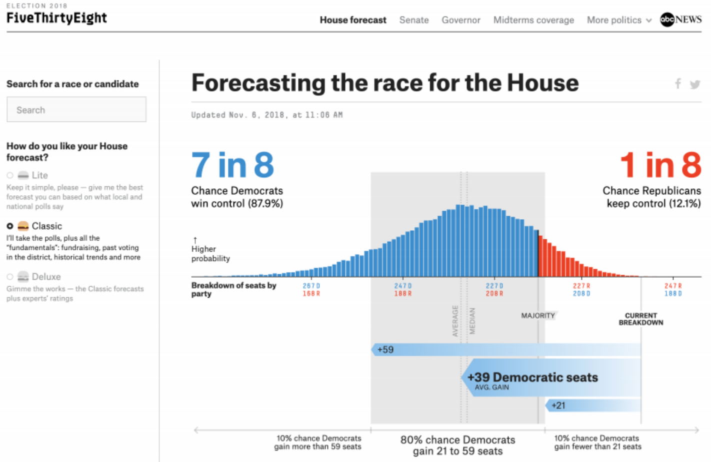

http://benschmidt.org/profGender/

## Poverty Mapping


Abelson, Varshney, and Sun. "Targeting Direct Cash Transfers to the Extremely Poor," 2012




## FiveThirtyEight



# Meddicorp Sales

Meddicorp Company sells medical supplies to hospitals, clinics, and doctor's offices.

Meddicorp's management considers the effectiveness of a new advertising program.

Management wants to know if the **advertisement** in 1999 is related to **sales**.

MediCorp
Health System

*a better way to work.*

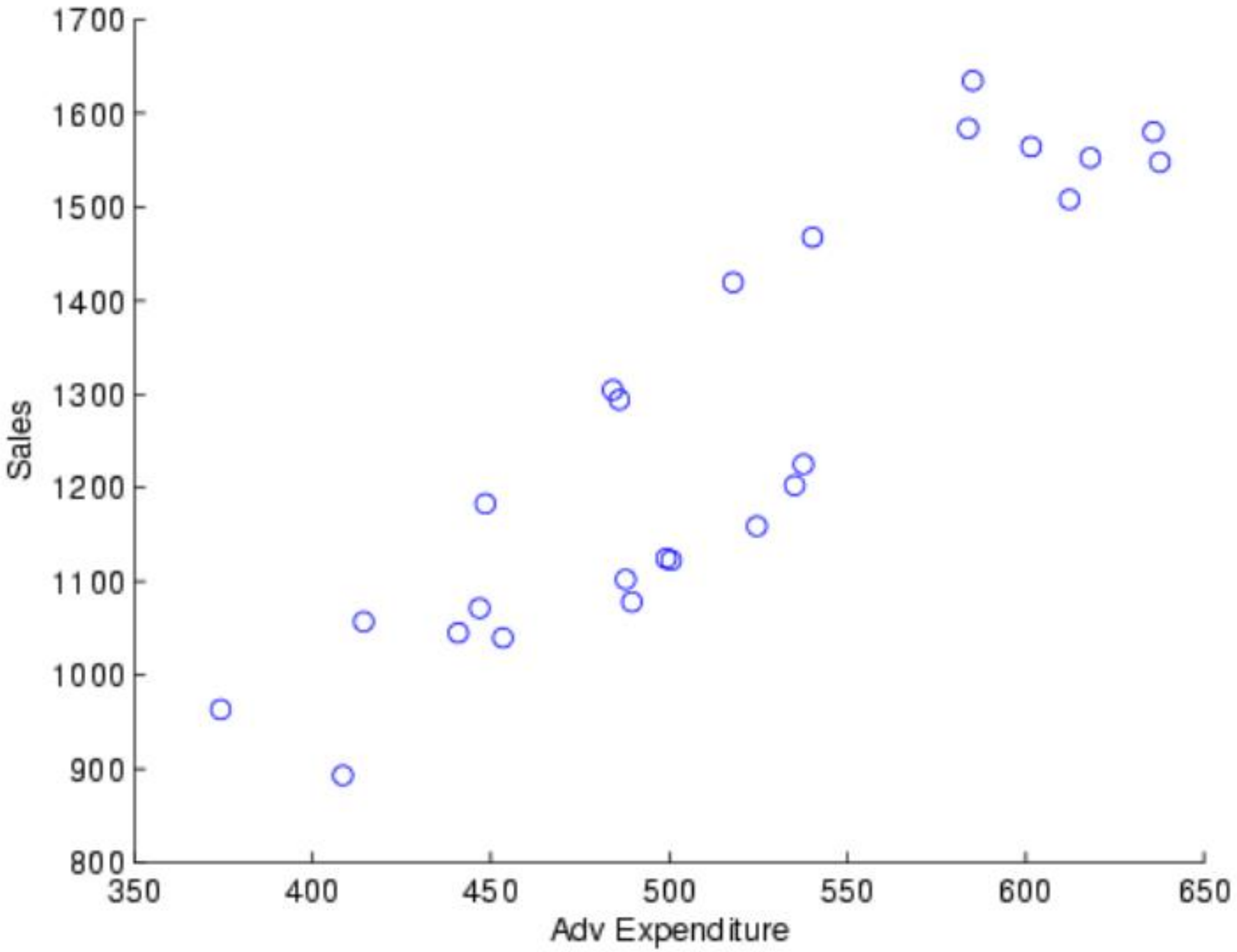
# Data

The company observes for 25 offices the yearly sales (in thousands) and the advertisement expenditure for the new program (in hundreds)

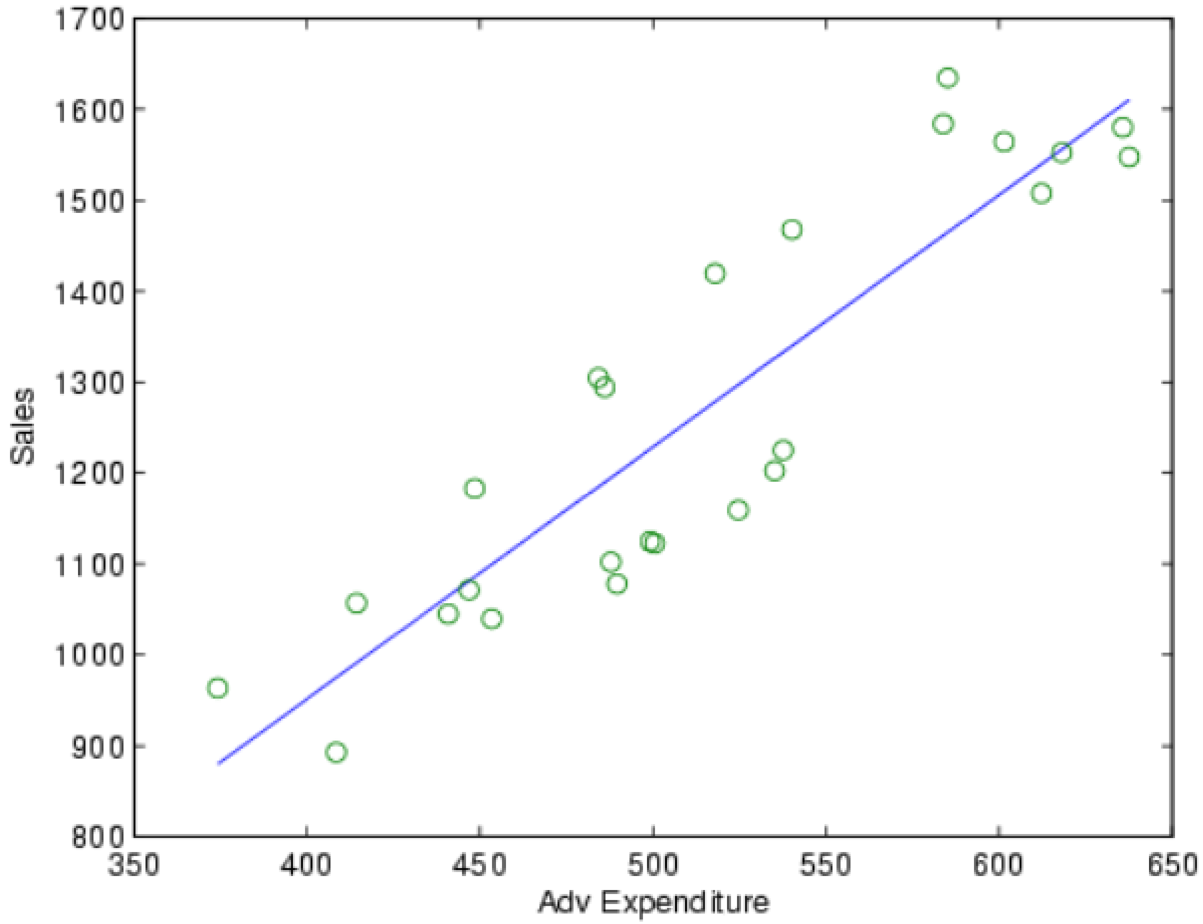
| | SALES | ADV |
|---|---|---|
| 1 | 963.50 | 374.27 |
| 2 | 893.00 | 408.50 |
| 3 | 1057.25 | 414.31 |
| 4 | 1183.25 | 448.42 |
| 5 | 1419.50 | 517.88 |

..........

# Step 1: graphical display of data — scatter plot: sales vs. advertisement cost



# Step 2: find the relationship or association between Sales and Advertisement Cost — Regression

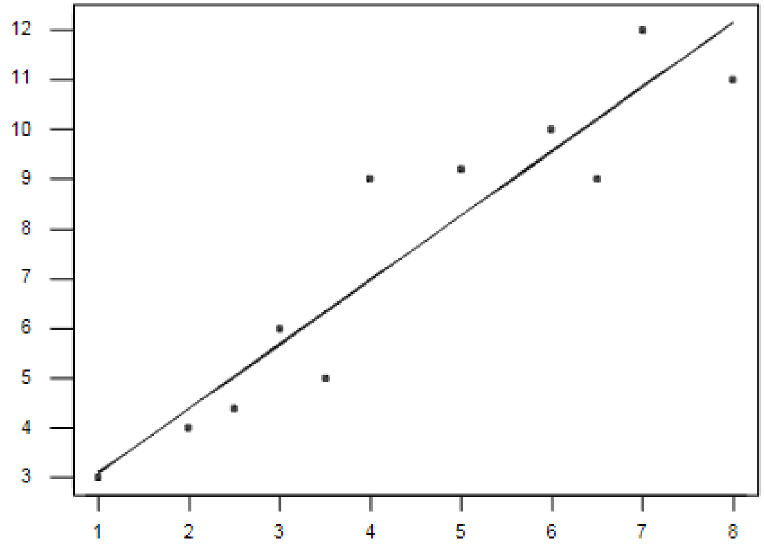


# Simple linear regression

Our goal is to find the best line that describes a linear relationship:

Find $(\beta_0, \beta_1)$ where

$$Y = \beta_0 + \beta_1 X + \varepsilon$$




Unknown parameters:

1. $\beta_0$ *Intercept* (where the line crosses y-axis)
2. $\beta_1$ *Slope* of the line

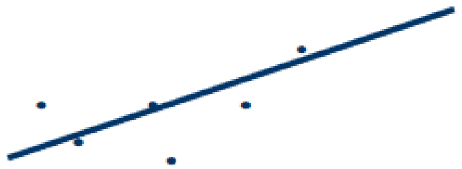
Basic idea

a. Plot observations (X,Y)
b. Find best line that follows plotted points

# Different forms of regression

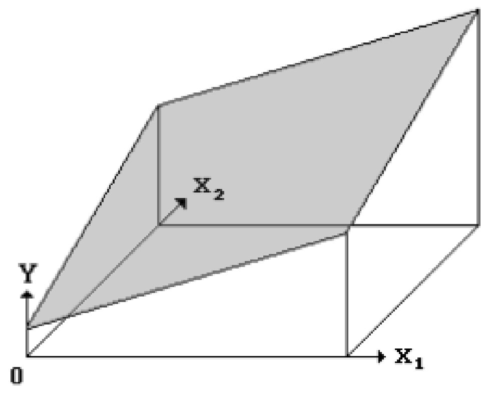
- Simple linear regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



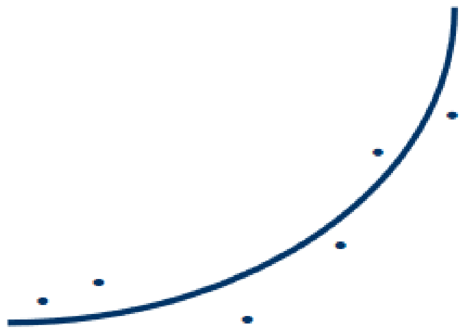
- Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$



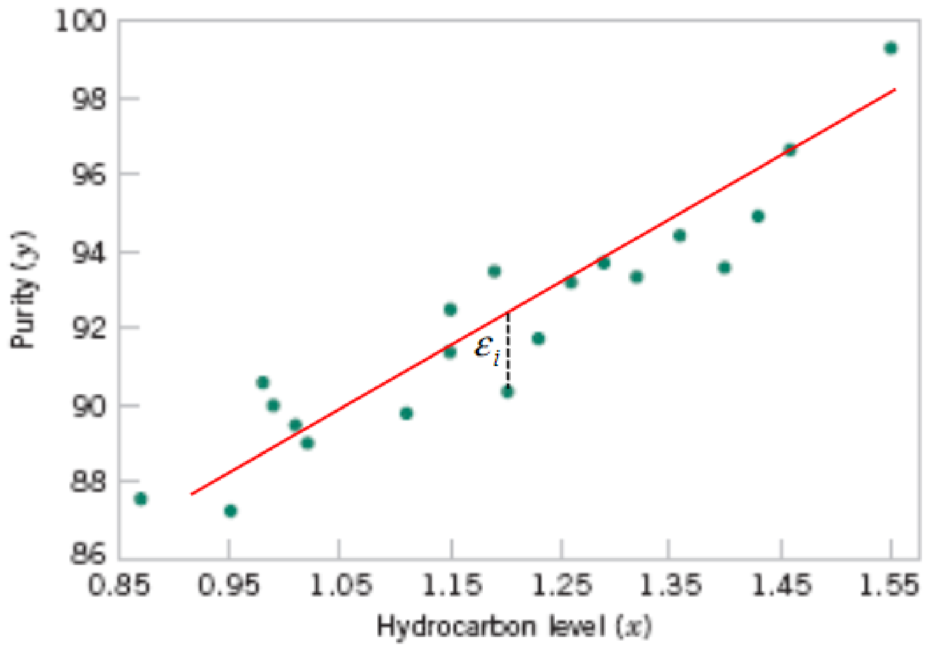

- Polynomial regression

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$



# Summary: simple linear regression

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to X by the following simple linear regression model:




Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

$\varepsilon_i \sim \mathrm{N}(0, \ \sigma^2)$

Intercept

Slope

Random error

where the slope and intercept of the line are called regression coefficients.

•The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

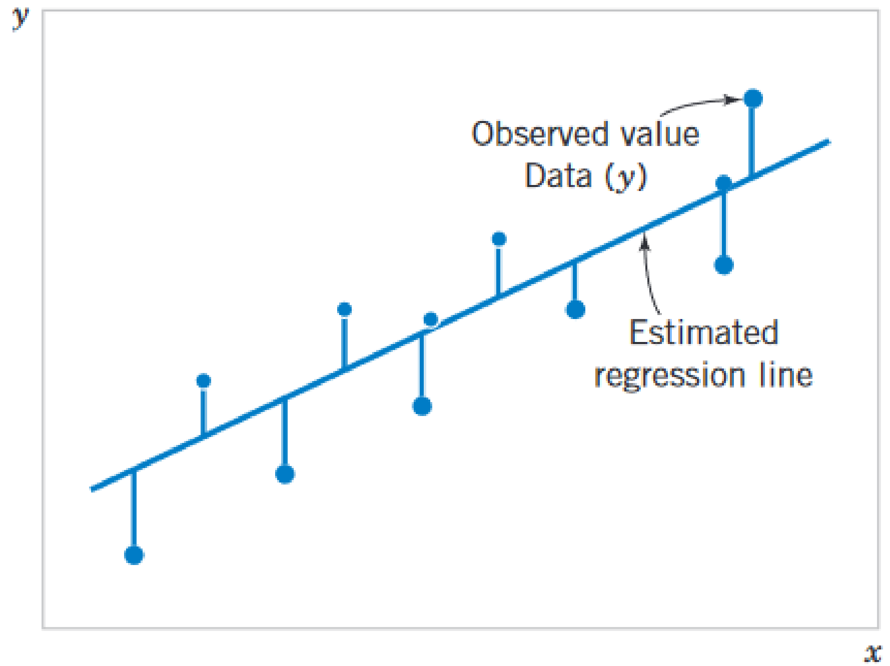
# Estimate regression parameters

To estimate $(\beta_0, \beta_1)$ , we find values that minimize squared error:

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

- derivation: method of least squares

# Method of least squares




$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

To estimate $(\beta_0, \beta_1)$ , we find values that minimize squared error:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators of $\beta_0$ and $\beta_1$, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

# Least square estimates

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} y_i x_i - \frac{\left(\displaystyle\sum_{i=1}^{n} y_i\right)\left(\displaystyle\sum_{i=1}^{n} x_i\right)}{n}}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

# Example: oxygen and hydrocarcon level

**Table 11-1** Oxygen and Hydrocarbon Levels

| Observation Number | Hydrocarbon Level $x$ (%) | Purity $y$ (%) |
|---|---|---|
| 1 | 0.99 | 90.01 |
| 2 | 1.02 | 89.05 |
| 3 | 1.15 | 91.43 |
| 4 | 1.29 | 93.74 |
| 5 | 1.46 | 96.73 |
| 6 | 1.36 | 94.45 |
| 7 | 0.87 | 87.59 |
| 8 | 1.23 | 91.77 |
| 9 | 1.55 | 99.42 |
| 10 | 1.40 | 93.65 |
| 11 | 1.19 | 93.54 |
| 12 | 1.15 | 92.52 |
| 13 | 0.98 | 90.56 |
| 14 | 1.01 | 89.54 |
| 15 | 1.11 | 89.85 |
| 16 | 1.20 | 90.39 |
| 17 | 1.26 | 93.25 |
| 18 | 1.32 | 93.41 |
| 19 | 1.43 | 94.98 |
| 20 | 0.95 | 87.33 |

Question: fit a simple regression model to related purity (y) to hydrocarbon level (x)

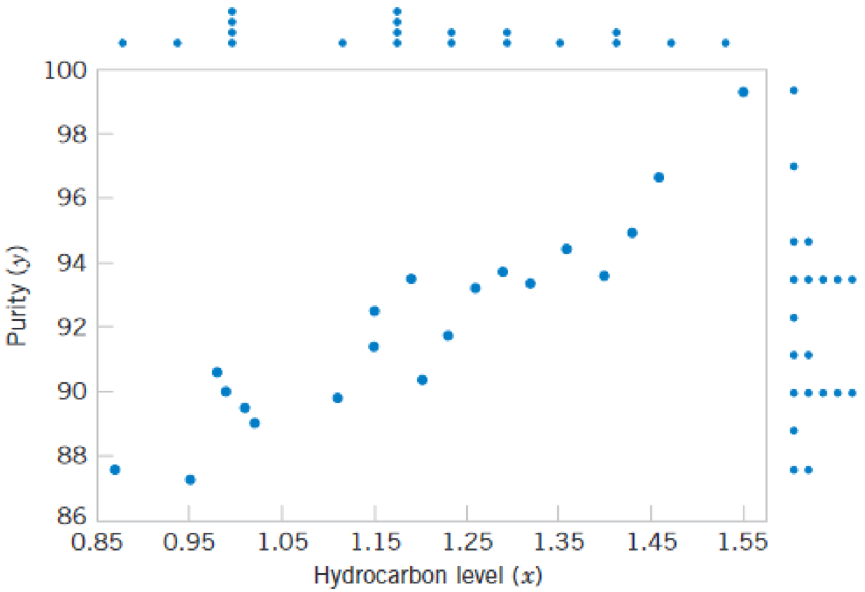



**Figure 11-1** Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

# Interpretation of regression model

- Regression model

$$\hat{y} = 74.283 + 14.947x$$

$$\hat{y} = 89.23\% \text{ when the hydrocarbon level is } x = 1.00\%.$$

- This may be interpreted as an estimate of the true population **mean** purity when $x = 1.00\%$.
- The estimates are subject to error

# Estimation of variance

- Using the fitted model, we can estimate value of the response variable for given predictor

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Residuals: $r_i = y_i - \hat{y}_i$
- Our model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,\ldots,n$, $Var(\varepsilon_i) = \sigma^2$
- Unbiased estimator (MSE: Mean Square Error)

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

- oxygen and hydrocarcon level example $\hat{\sigma}^2 = 1.18$

# Example: Oil Well Drilling Costs

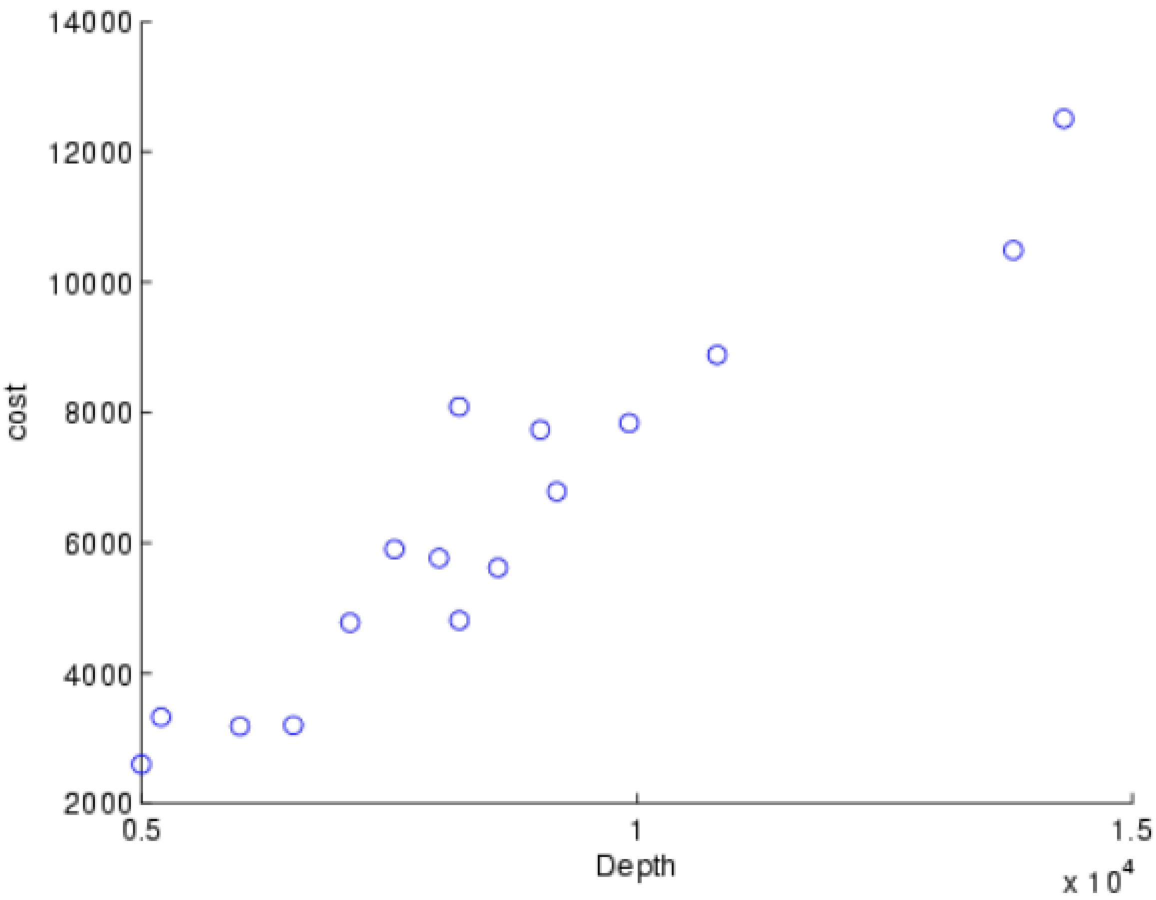
Estimating the costs of drilling oil wells is an important consideration for the oil industry.

Data: the **total costs** and the **depths** of 16 off-shore oil wells located in Philippines.

| Depth | Cost |
|---|---|
| 5000 | 2596.8 |
| 5200 | 3328.0 |
| 6000 | 3181.1 |
| 6538 | 3198.4 |
| 7109 | 4779.9 |
| 7556 | 5905.6 |
| 8005 | 5769.2 |
| 8207 | 8089.5 |

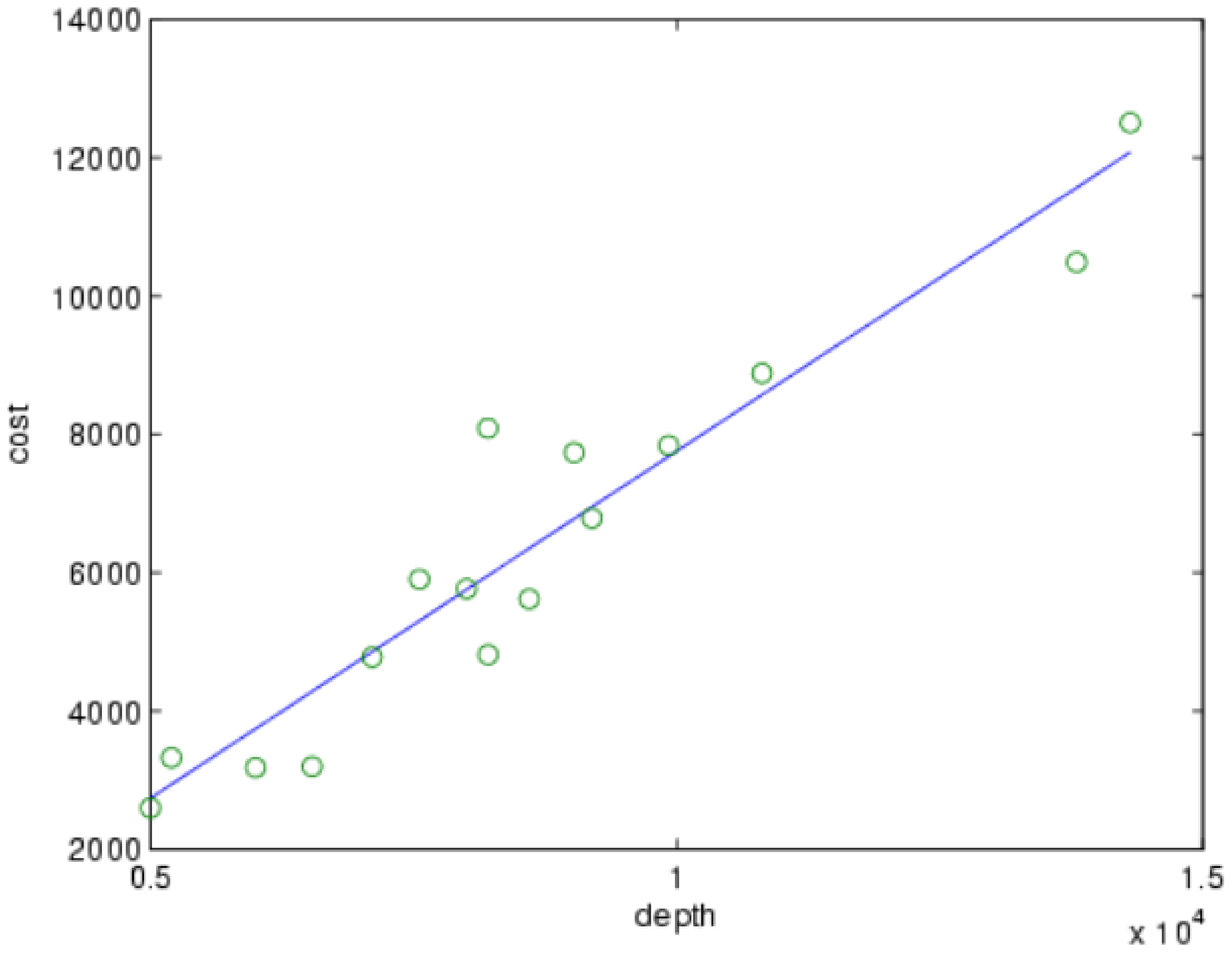
| Depth | Cost |
|---|---|
| 8210 | 4813.1 |
| 8600 | 5618.7 |
| 9026 | 7736.0 |
| 9197 | 6788.3 |
| 9926 | 7840.8 |
| 10813 | 8882.5 |
| 13800 | 10489.5 |
| 14311 | 12506.6 |

# Step 1: graphical display of the data




R code: plot(Depth, Cost, xlab= “Depth”, ylab = “Cost”)

# Step 2: find the relationship between Depth and Cost



# Results and use of regression model

1. Fit a linear regression model:

Estimates $(\beta_0,\beta_1)$ are (-2277.1, 1.0033)

2. What does the model predict as the cost increase for an additional depth of 1000 ft?

If we increase X by 1000, we increase Y by $1000\beta_1$ = \$1003

3. What cost would you predict for an oil well of 10,000 ft depth?

X = 10,000 ft is in the range of the data, and estimate of the line at x=10,000 is $\hat{\beta}_0 + (10{,}000)\hat{\beta}_1$ = -2277.1 + 10,033 = \$7753

4. What is the estimate of the error variance? Estimate $\sigma^2 \approx 774{,}211$

# Summary

- Simple linear regression

$$Y = \beta_0 + \beta_1 X$$

- Estimate coefficients from data: method of least squares

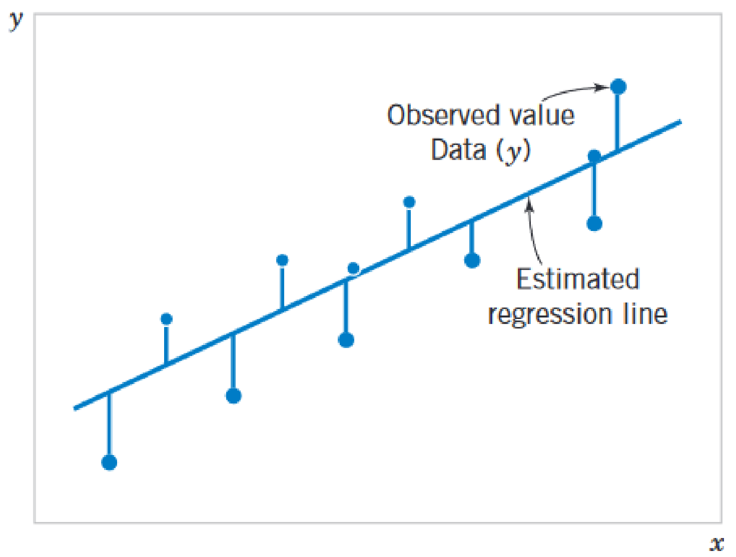
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Fitted (estimated) regression model

- Estimate of variance



# Manual implementation of linear regression

Create data matrices:

```
# initialize X matrix and y vector
X = np.array([df["Temp"], df["IsWeekday"], np.ones(len(df))]).T
y = df_summer["Load"].values
```

Compute solution:

```
# solve least squares
theta = np.linalg.solve(X.T @ X, X.T @ y)
print(theta)
# [ 0.04747948  0.22462824 -1.80260016]
```

Make predictions:

```
# predict on new data
Xnew = np.array([[77, 1, 1], [80, 0, 1]])
ypred = Xnew @ theta
print(ypred)
# [ 2.07794778  1.99575797]
```

# Scikit-learn

By far the most popular machine learning library in Python is the scikit-learn library (http://scikit-learn.org/)

Reasonable (usually) implementation of many different learning algorithms, usually fast enough for small/medium problems

**Important:** you *need* to understand the very basics of how these algorithms work in order to use them effectively

Sadly, a lot of data science in practice seems to be driven by the default parameters for scikit-learn classifiers…

# Linear regression in scikit-learn

## Fit a model and predict on new data

```
from sklearn.linear_model import LinearRegression

# don't include constant term in X
X = np.array([df_summer["Temp"], df_summer["IsWeekday"]]).T
model = LinearRegression(fit_intercept=True, normalize=False)
model.fit(X, y)

# predict on new data
Xnew = np.array([[77, 1], [80, 0]])
model.predict(Xnew)
# [ 2.07794778  1.99575797]
```

## Inspect internal model coefficients

```
print(model.coef_, model.intercept_)
# [ 0.04747948  0.22462824]  -1.80260016
```

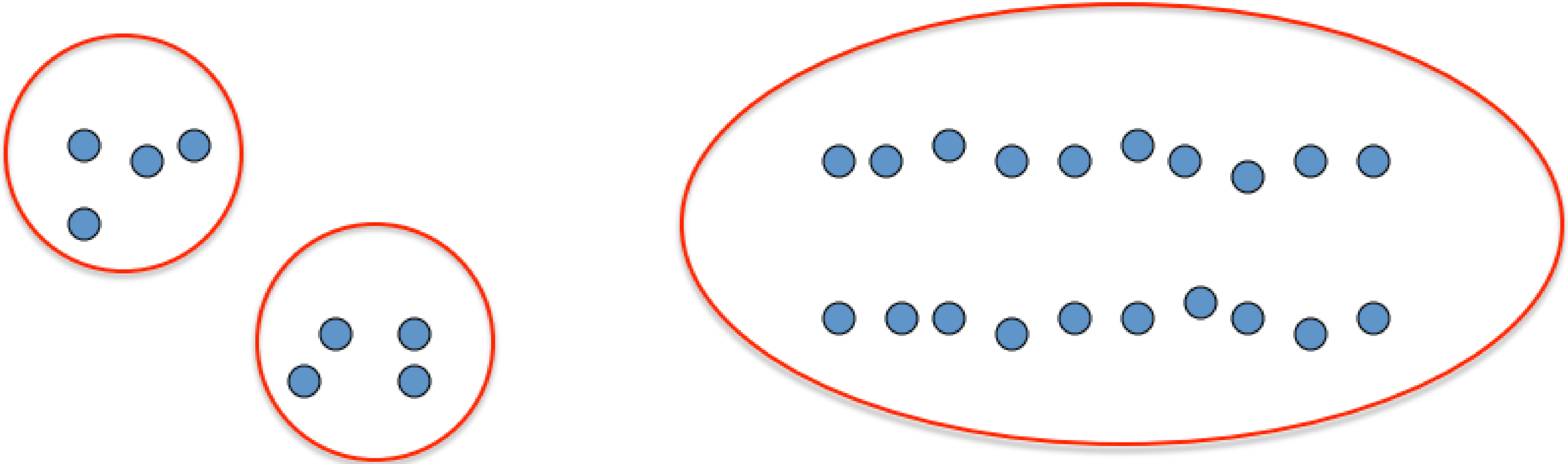
# Clustering

Clustering:

- Unsupervised learning
- Requires data, but no labels
- Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Useful when don't know what you're looking for
- But: can get gibberish

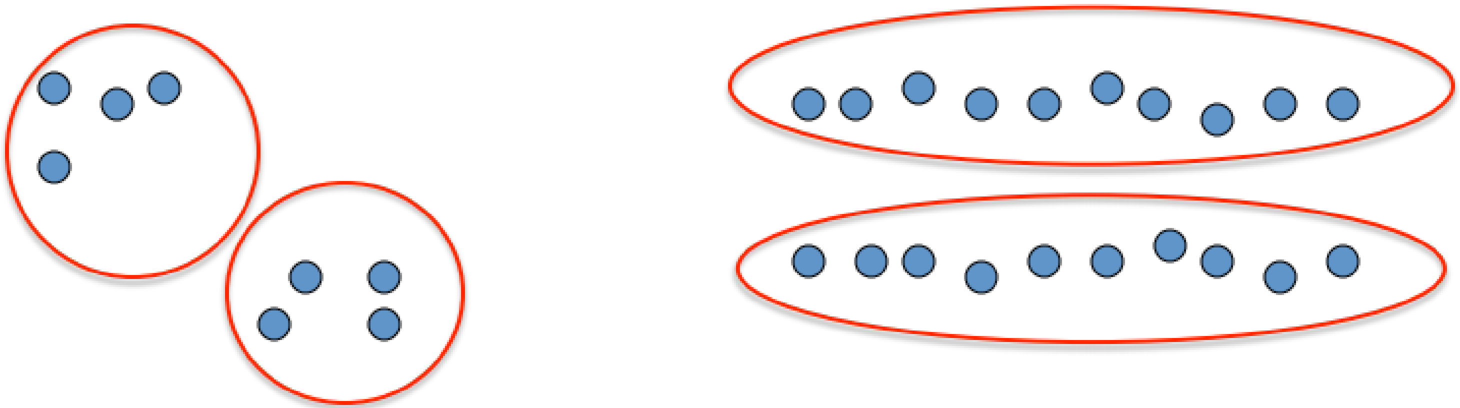
# Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



# Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



- What could "similar" mean?
  - One option: small Euclidean distance (squared)

$$\mathrm{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$

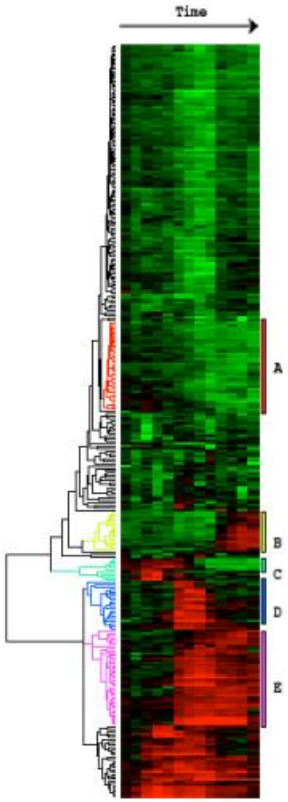
  - Clustering results are crucially dependent on the measure of similarity (or distance) between "points" to be clustered

**Image segmentation**

Goal: Break up the image into meaningful or perceptually similar regions

[Slide from James Hayes]

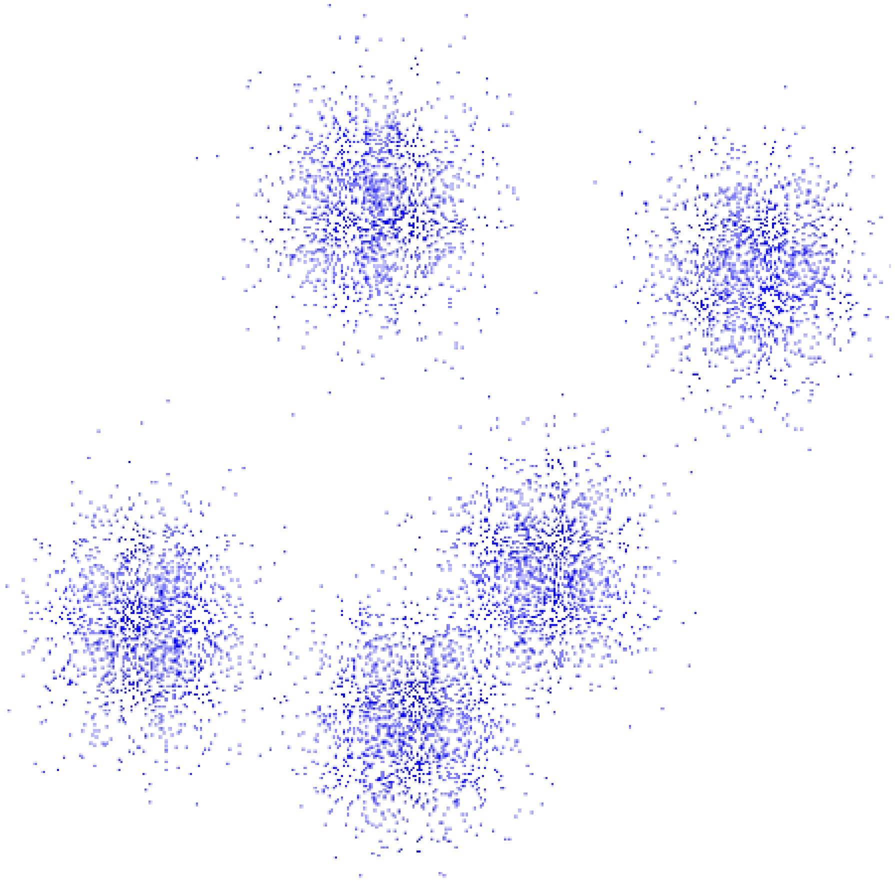
## Clustering gene expression data



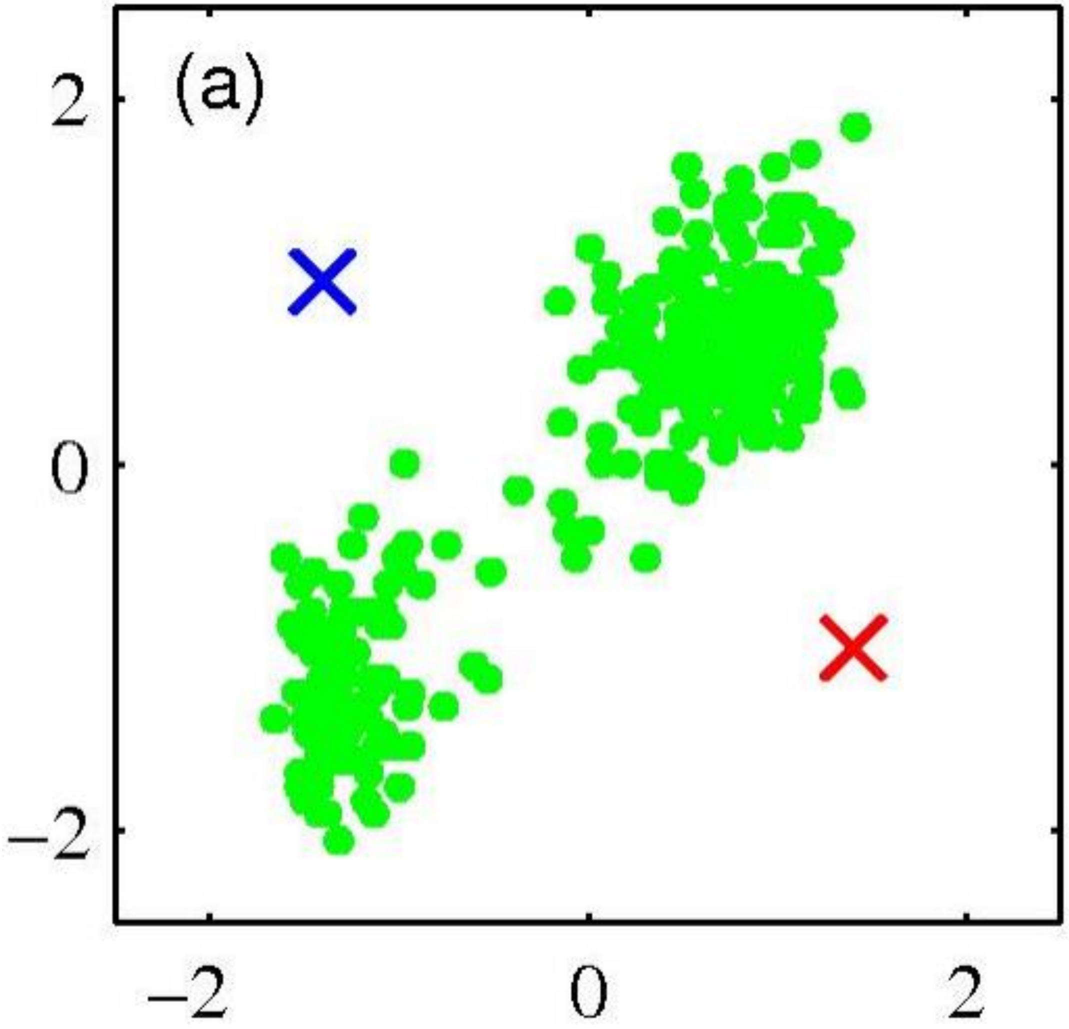

Eisen et al, PNAS 1998

# K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



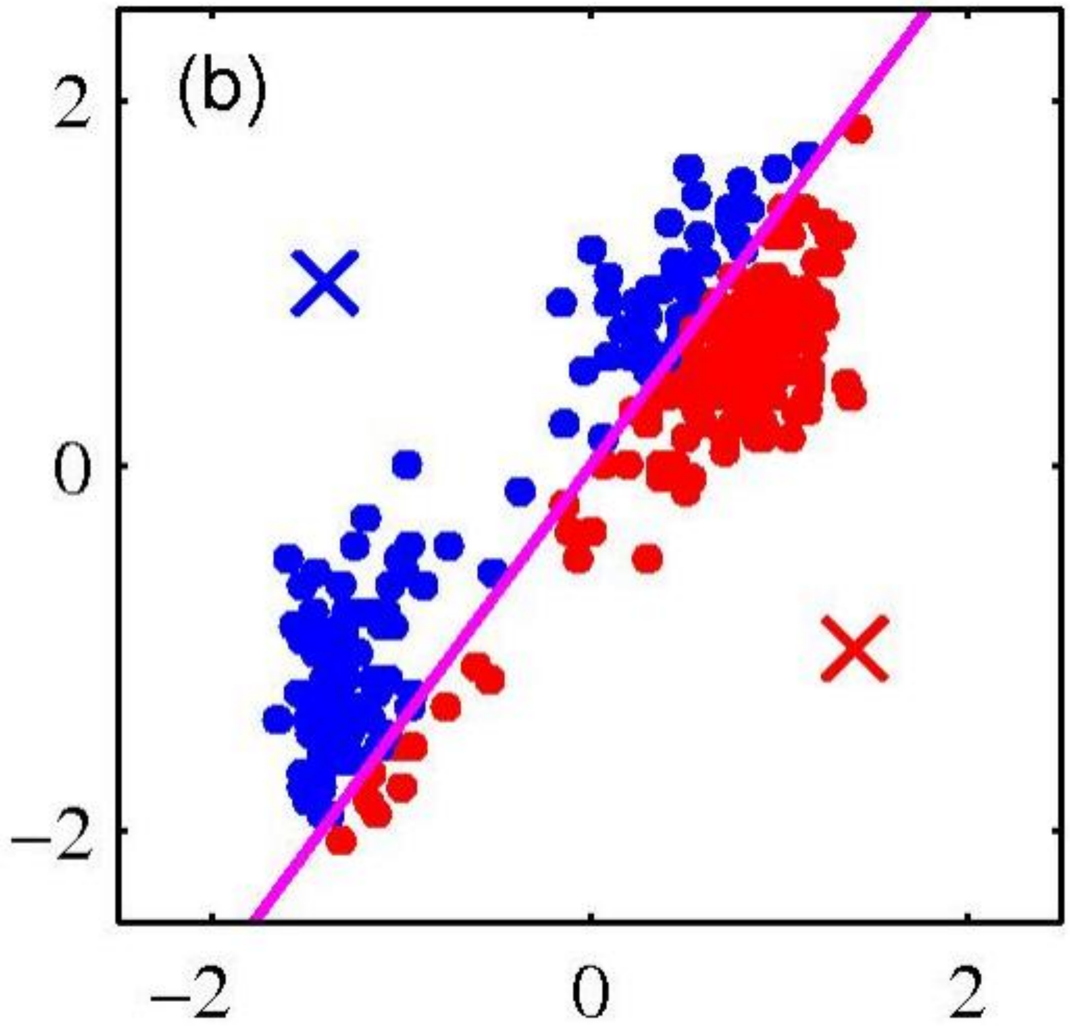
# K-means clustering: Example




- Pick $K$ random points as cluster centers (means)

Shown here for $K$=2

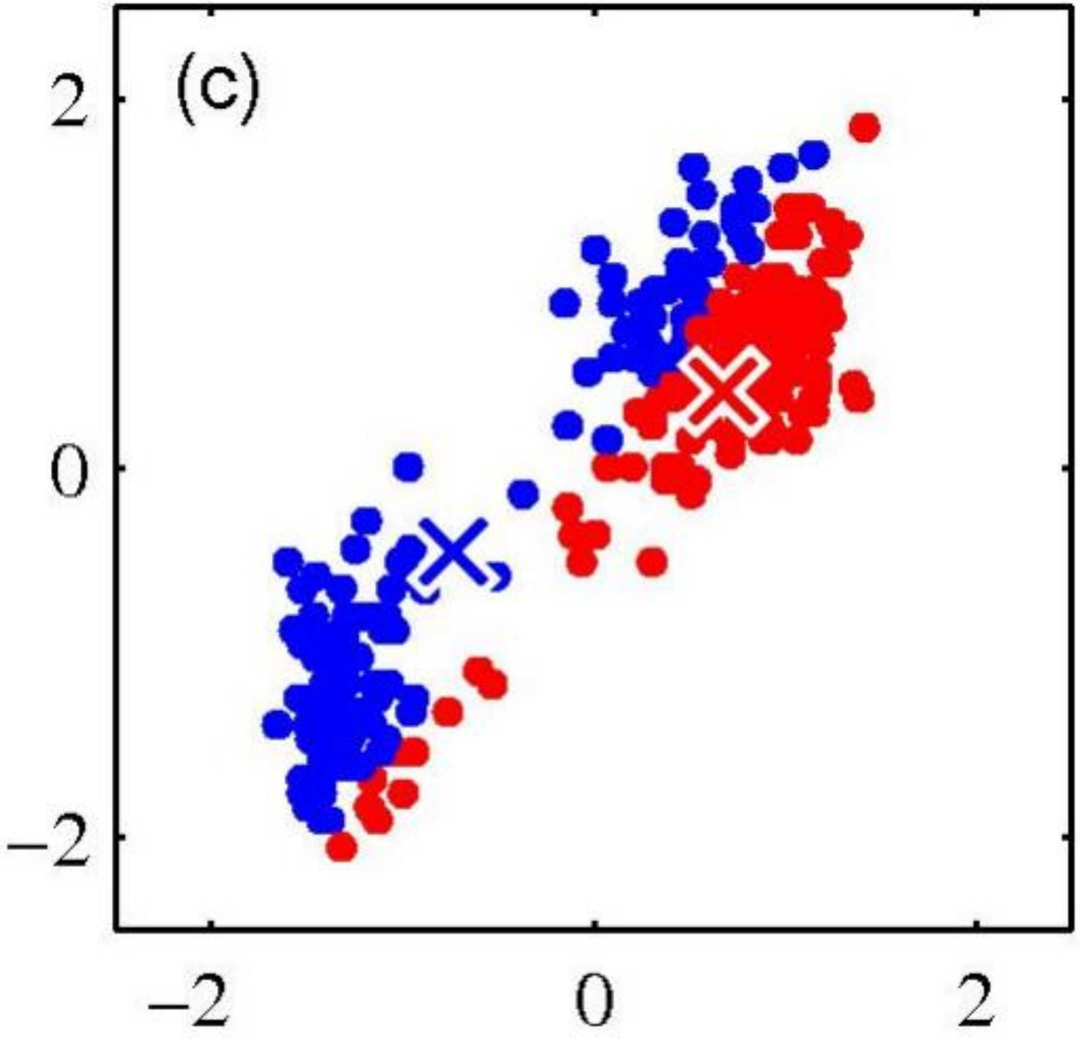
# K-means clustering: Example




Iterative Step 1

- Assign data points to closest cluster center

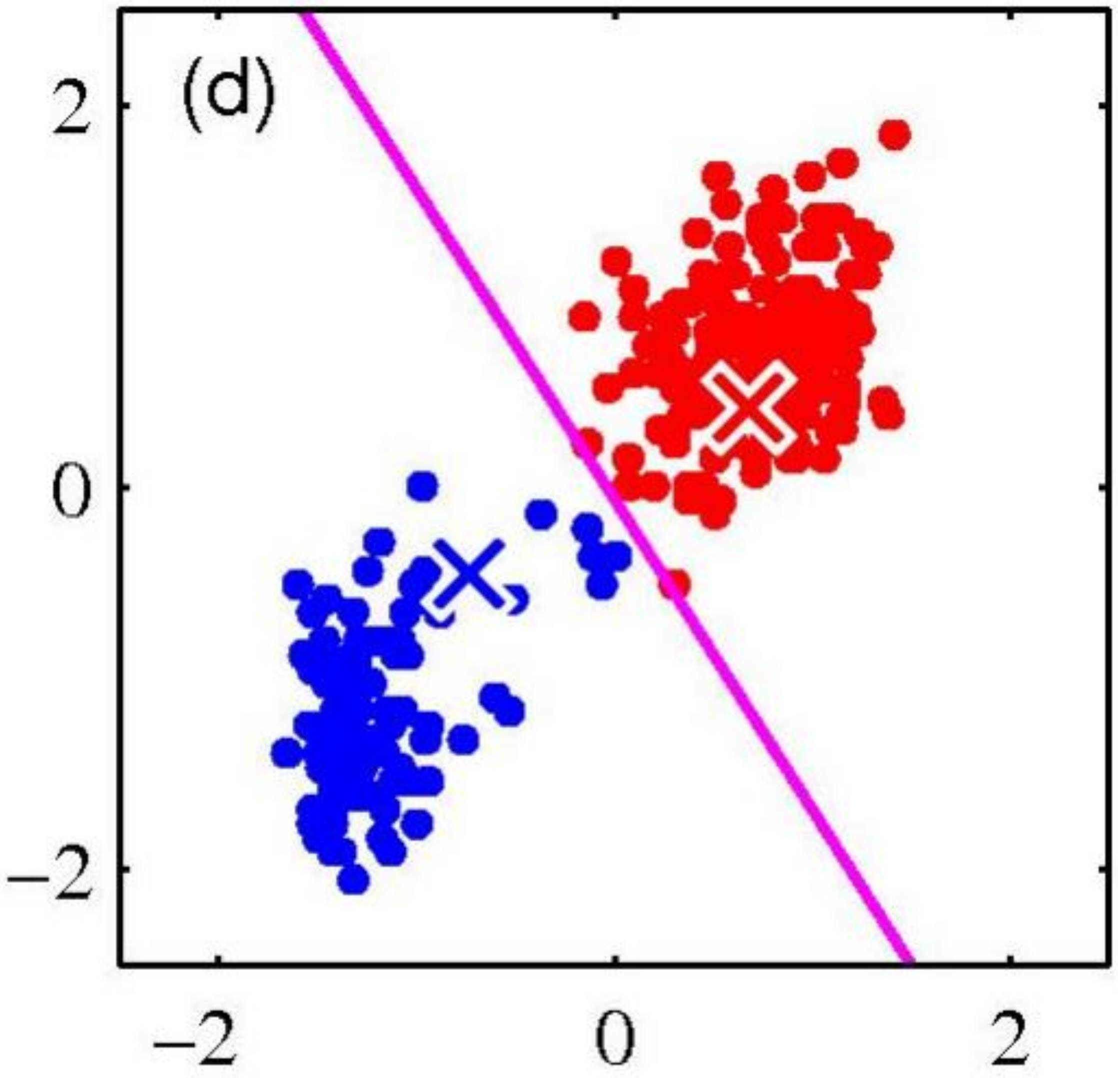
# K-means clustering: Example




Iterative Step 2

- Change the cluster center to the average of the assigned points

# K-means clustering: Example




- Repeat until convergence

# Properties of K-means **algorithm**

- Guaranteed to converge in a finite number of iterations

- Running time per iteration:
  1. Assign data points to closest cluster center

     O(KN) time

  2. Change the cluster center to the average of its assigned points

     O(N)

# Example: K-Means for Segmentation

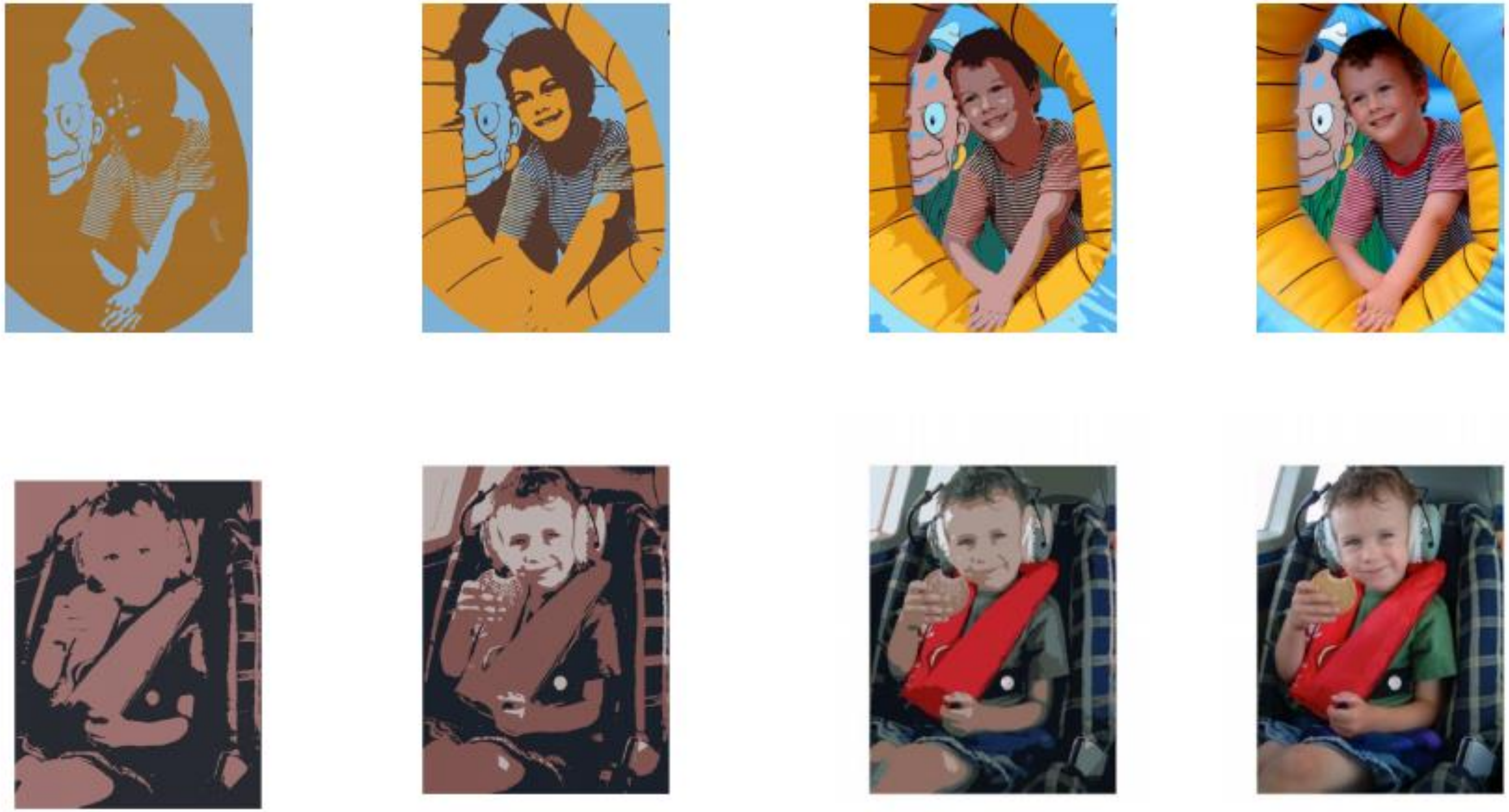
K=2

**Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.**

Original

# Example: K-Means for Segmentation

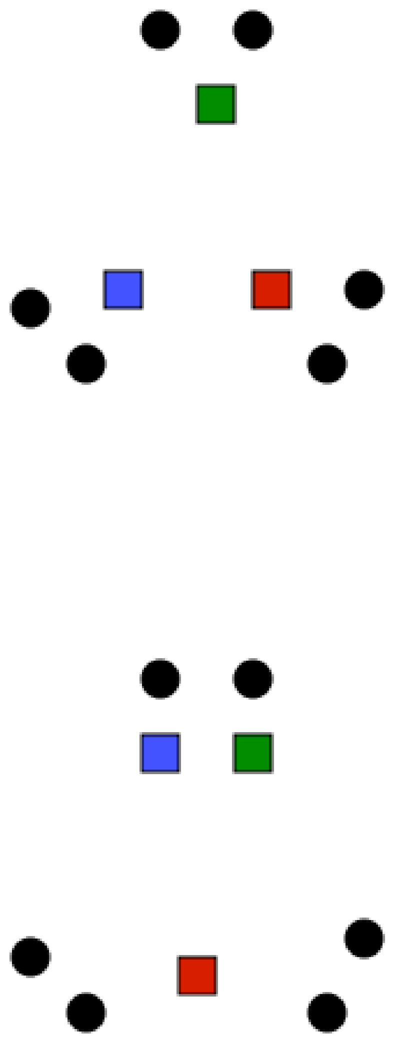
K=2 K=3 K=10 Original



4% 8% 17%

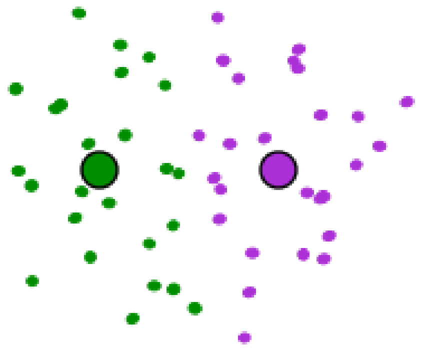
# Initialization

- K-means **algorithm** is a heuristic
  - Requires initial means
  - It does matter what you pick!
  - What can go wrong?
  - Various schemes for preventing this kind of thing: variance-based split / merge, initialization heuristics

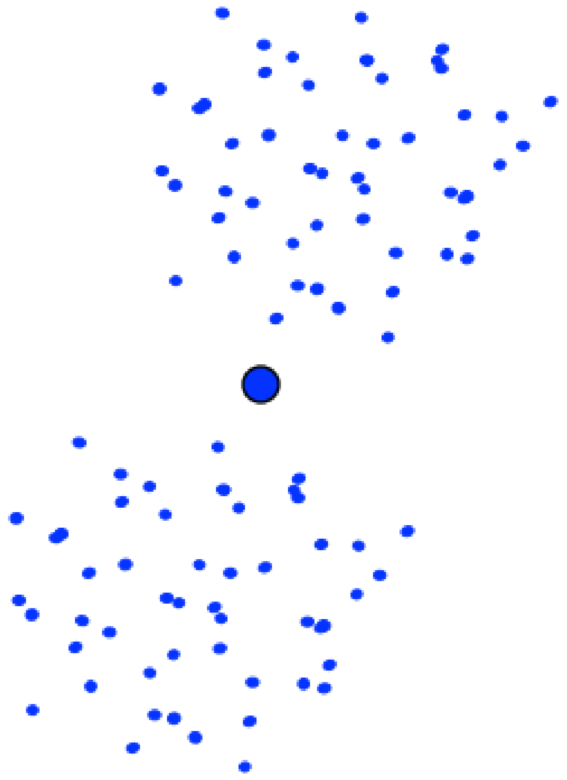


# K-Means Getting Stuck

A local optimum:

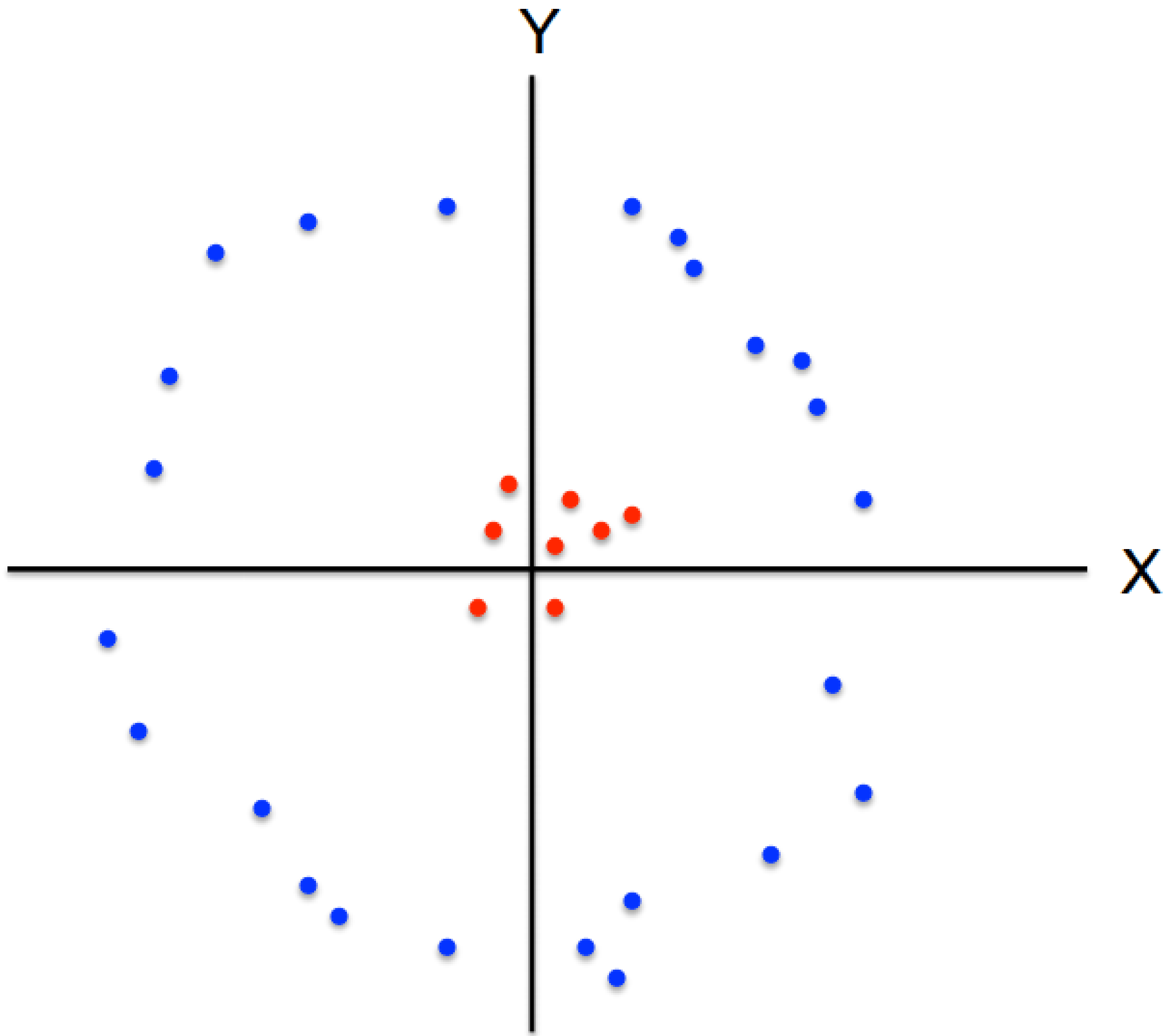


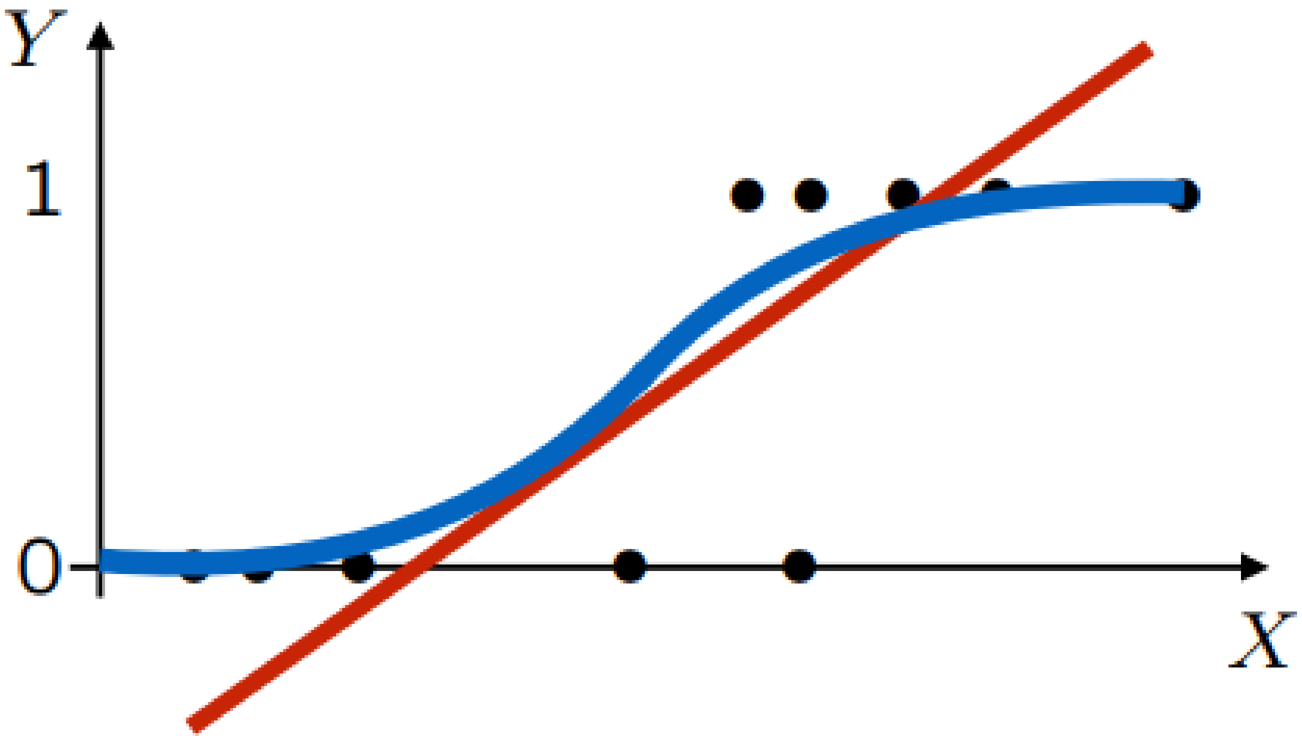
Would be better to have
one cluster here

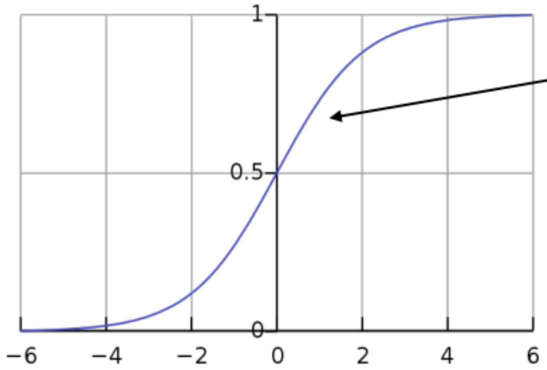


… and two clusters here

# K-means not able to properly cluster






$$\frac{e^x}{e^x + e^{-x}}$$

shift and scale

$$\frac{e^{\theta_1 x}}{e^{\theta_0 x} + e^{\theta_1 x}}$$

$$\mathbb{P}(Y = k \mid \mathbf{X}) = \frac{\exp\{\boldsymbol{\theta}_k^T \mathbf{X}\}}{\sum_s \exp\{\boldsymbol{\theta}_s^T \mathbf{X}\}}$$

- **A better idea:** Fit a curve that ranges between 0 and 1 (must be nonlinear)
  - interpret as (estimated) probability $\mathbb{P}(Y = 1 \mid \mathbf{X})$

# Logistic Regression - Simple Example

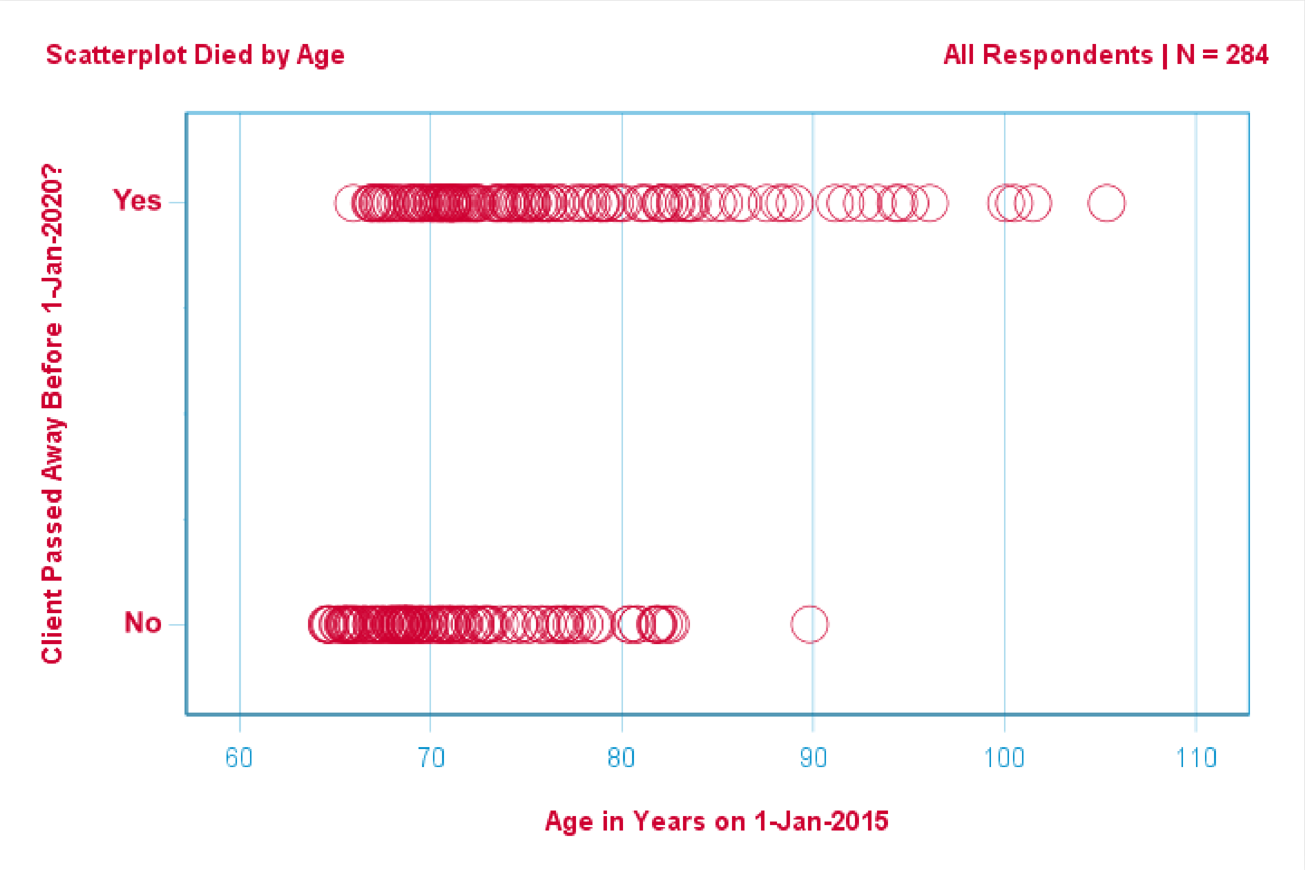
A nursing home has data on N = 284 clients' sex, age on 1 January 2015 and whether the client passed away before 1 January 2020. The raw data are in this Googlesheet, partly shown below.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Raw data | | | | |
| 2 | id | male | age | died | |
| 3 | 17734 | 1 | 86 | 1 | |
| 4 | 17742 | 0 | 83 | 0 | |
| 5 | 17748 | 0 | 66 | 0 | |
| 6 | 17753 | 1 | 72 | 1 | |

Let's first just focus on age:

**can we predict death before 2020 from age in 2015?**

And -if so- precisely *how*? And to what extent? A good first step is inspecting a scatterplot like the one shown below.



## Logistic Regression Example Curves

$$P(Y_i) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i})}}$$

where

- $P(Y_i)$ is the predicted probability that $Y$ is true for case $i$
- $e$ is a mathematical constant of roughly 2.72;
- $b_0$ is a constant estimated from the data;
- $b_1$ is a b-coefficient estimated from the data;
- $X_i$ is the observed score on variable $X$ for case $i$.