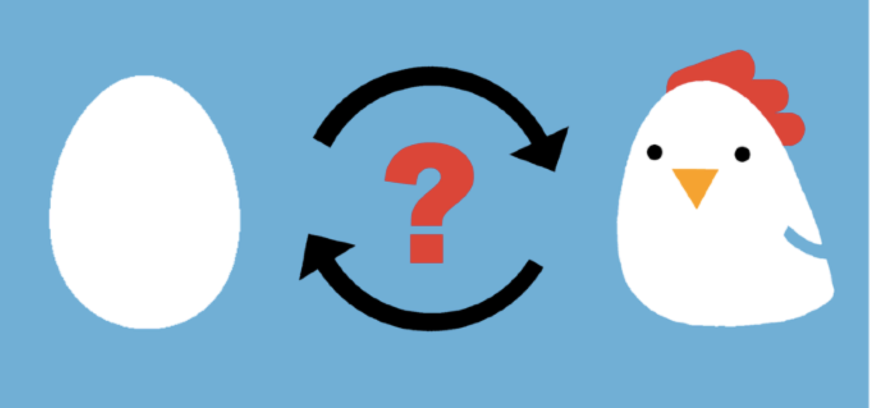


Causality and Machine Learning *(80-816/516)*

Classes 4 (Jan 23, 2025)

From Statistics to ML: Settings, Assumptions, Basic Methods, and Model Selection

Instructor:

Kun Zhang (kunz1@cmu.edu)

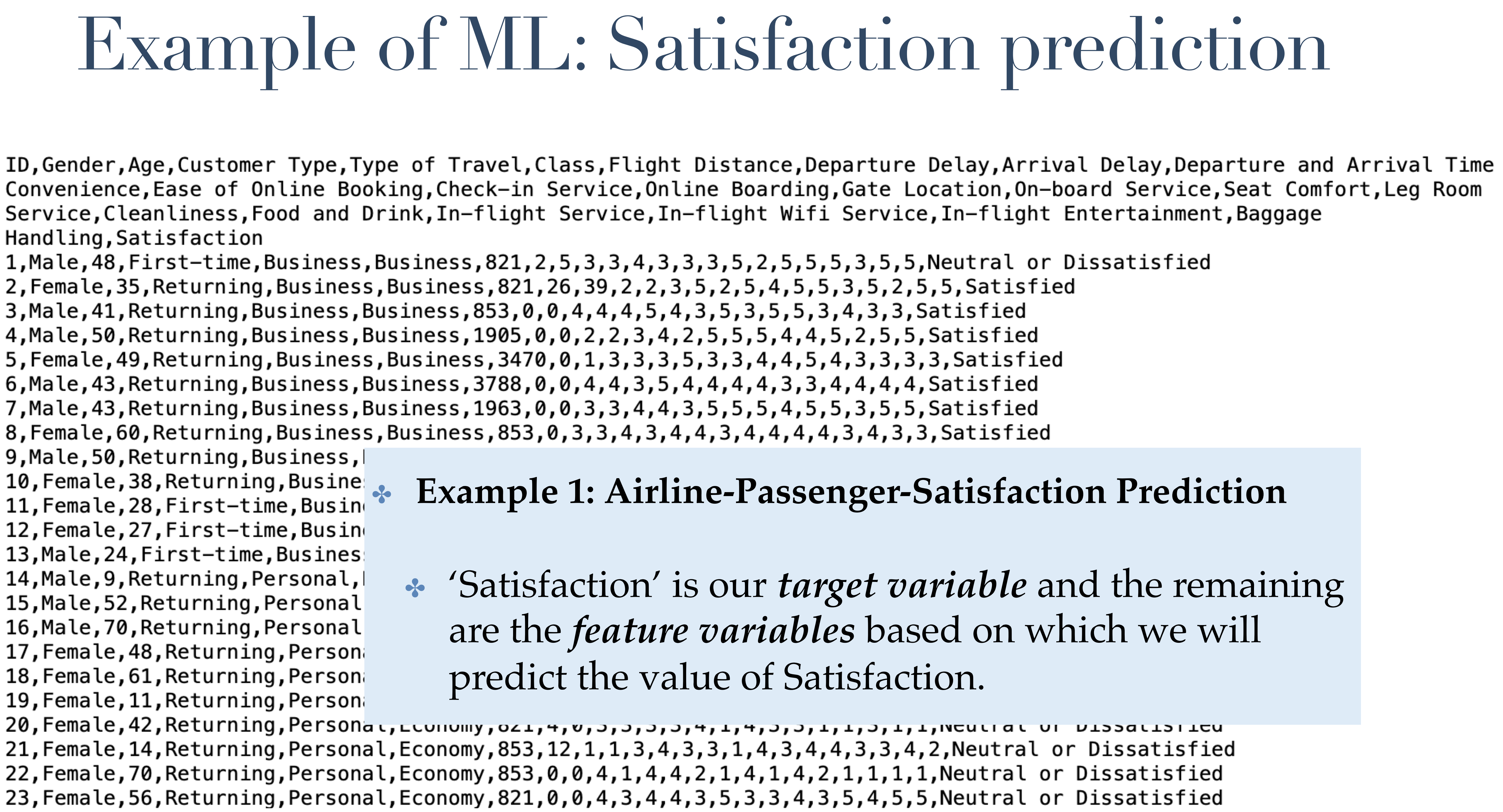
Zoom link: https://cmu.zoom.us/j/8214572323)

Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

Outline

- Supervised learning
    - From linear regression to nonlinear methods
        - Properties of regression
    - From parametric models to nonparametric models
    - Model selection: Why? What? How?
    - Classification
- Unsupervised learning
    - Clustering ↓
    - Dimensionality reduction… →

Remember This Example?



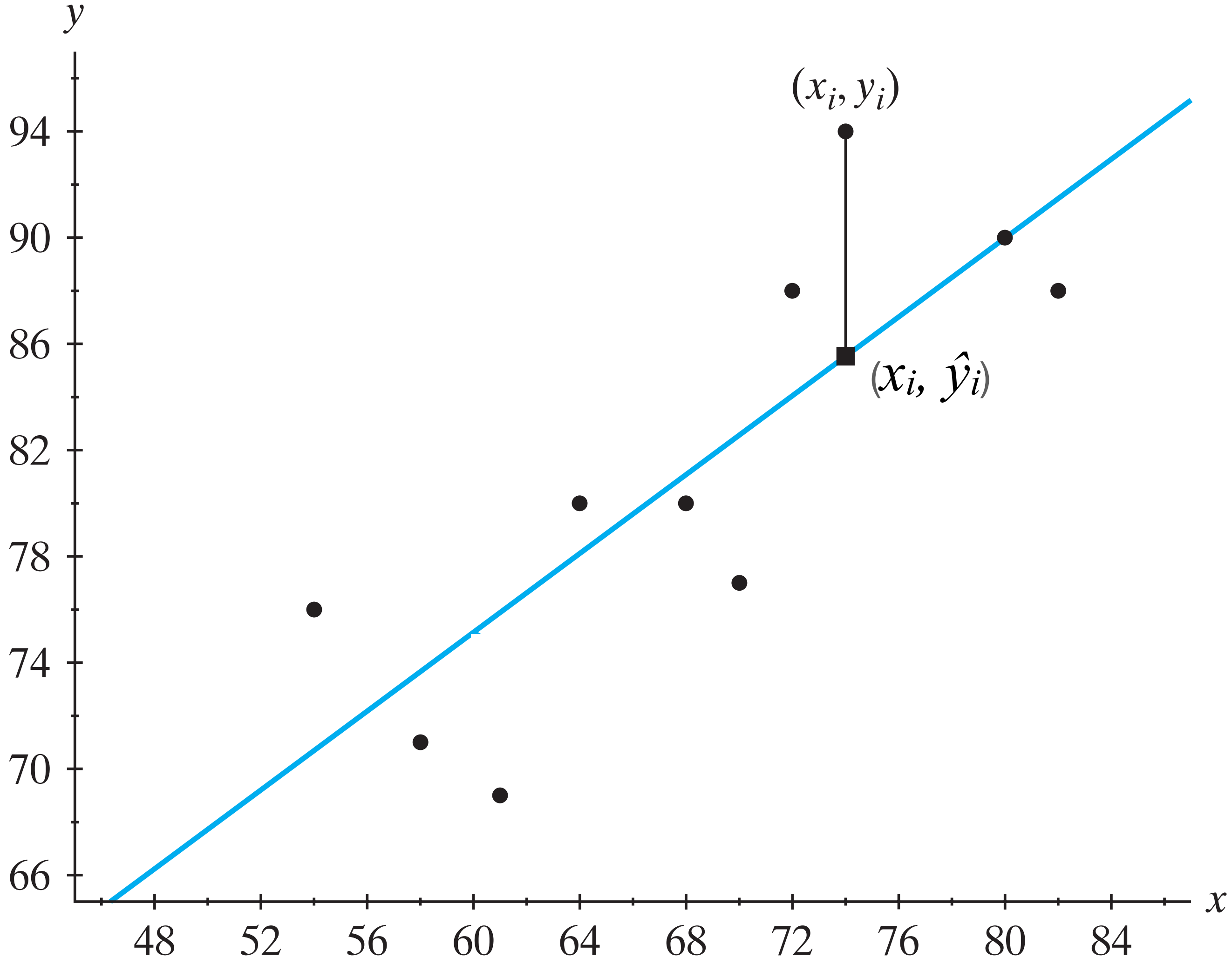
Example of ML: Satisfaction prediction

```
ID,Gender,Age,Customer Type,Type of Travel,Class,Flight Distance,Departure Delay,Arrival Delay,Departure and Arrival Time
Convenience,Ease of Online Booking,Check-in Service,Online Boarding,Gate Location,On-board Service,Seat Comfort,Leg Room
Service,Cleanliness,Food and Drink,In-flight Service,In-flight Wifi Service,In-flight Entertainment,Baggage
Handling,Satisfaction
1,Male,48,First-time,Business,Business,821,2,5,3,3,4,3,3,3,5,2,5,5,5,3,5,5,Neutral or Dissatisfied
2,Female,35,Returning,Business,Business,821,26,39,2,2,3,5,2,5,4,5,5,3,5,2,5,5,Satisfied
3,Male,41,Returning,Business,Business,853,0,0,4,4,4,5,4,3,5,3,5,5,3,4,3,3,Satisfied
4,Male,50,Returning,Business,Business,1905,0,0,2,2,3,4,2,5,5,5,4,4,5,2,5,5,Satisfied
5,Female,49,Returning,Business,Business,3470,0,1,3,3,3,5,3,3,4,4,5,4,3,3,3,3,Satisfied
6,Male,43,Returning,Business,Business,3788,0,0,4,4,3,5,4,4,4,4,3,3,4,4,4,4,Satisfied
7,Male,43,Returning,Business,Business,1963,0,0,3,3,4,4,3,5,5,5,4,5,5,3,5,5,Satisfied
8,Female,60,Returning,Business,Business,853,0,3,3,4,3,4,4,3,4,4,4,4,3,4,3,3,Satisfied
9,Male,50,Returning,Business,
10,Female,38,Returning,Busine
11,Female,28,First-time,Busin
12,Female,27,First-time,Busin
13,Male,24,First-time,Busines
14,Male,9,Returning,Personal,
15,Male,52,Returning,Personal
16,Male,70,Returning,Personal
17,Female,48,Returning,Persona
18,Female,61,Returning,Persona
19,Female,11,Returning,Persona
20,Female,42,Returning,Personal,Economy,821,4,0,3,3,3,3,4,1,4,3,3,1,1,3,1,1,Neutral or Dissatisfied
21,Female,14,Returning,Personal,Economy,853,12,1,1,3,4,3,3,1,4,3,4,4,3,3,4,2,Neutral or Dissatisfied
22,Female,70,Returning,Personal,Economy,853,0,0,4,1,4,4,2,1,4,1,4,2,1,1,1,1,Neutral or Dissatisfied
23,Female,56,Returning,Personal,Economy,821,0,0,4,3,4,4,3,5,3,3,4,3,5,4,5,5,Neutral or Dissatisfied
```

- **Example 1: Airline-Passenger-Satisfaction Prediction**
  - ‘Satisfaction’ is our ***target variable*** and the remaining are the ***feature variables*** based on which we will predict the value of Satisfaction.

Linear Regression

- How to find the regression line $\underline{\hat{y}=\alpha x + c}$ from data points $(x_1, y_1), \ldots, (x_n, y_n)$?
- To explain/predict Y with X
- Probabilisitic model: $Y = aX + u + \varepsilon$
- Y: dependent variable; X: explanatory / independent variable.



Linear Regression: Terminology

slope parameter

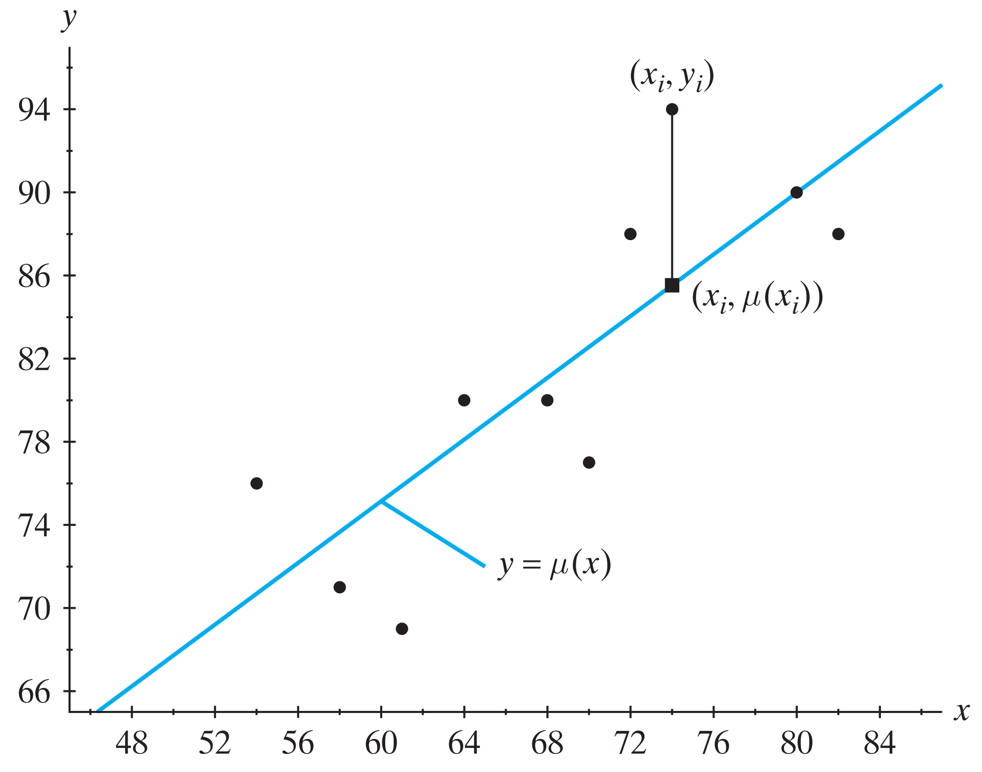
intercept

- $Y = aX + u + \varepsilon$, *where* $\varepsilon \sim N(0,\sigma^2)$

random error

independent variable / predictor

dependent variable

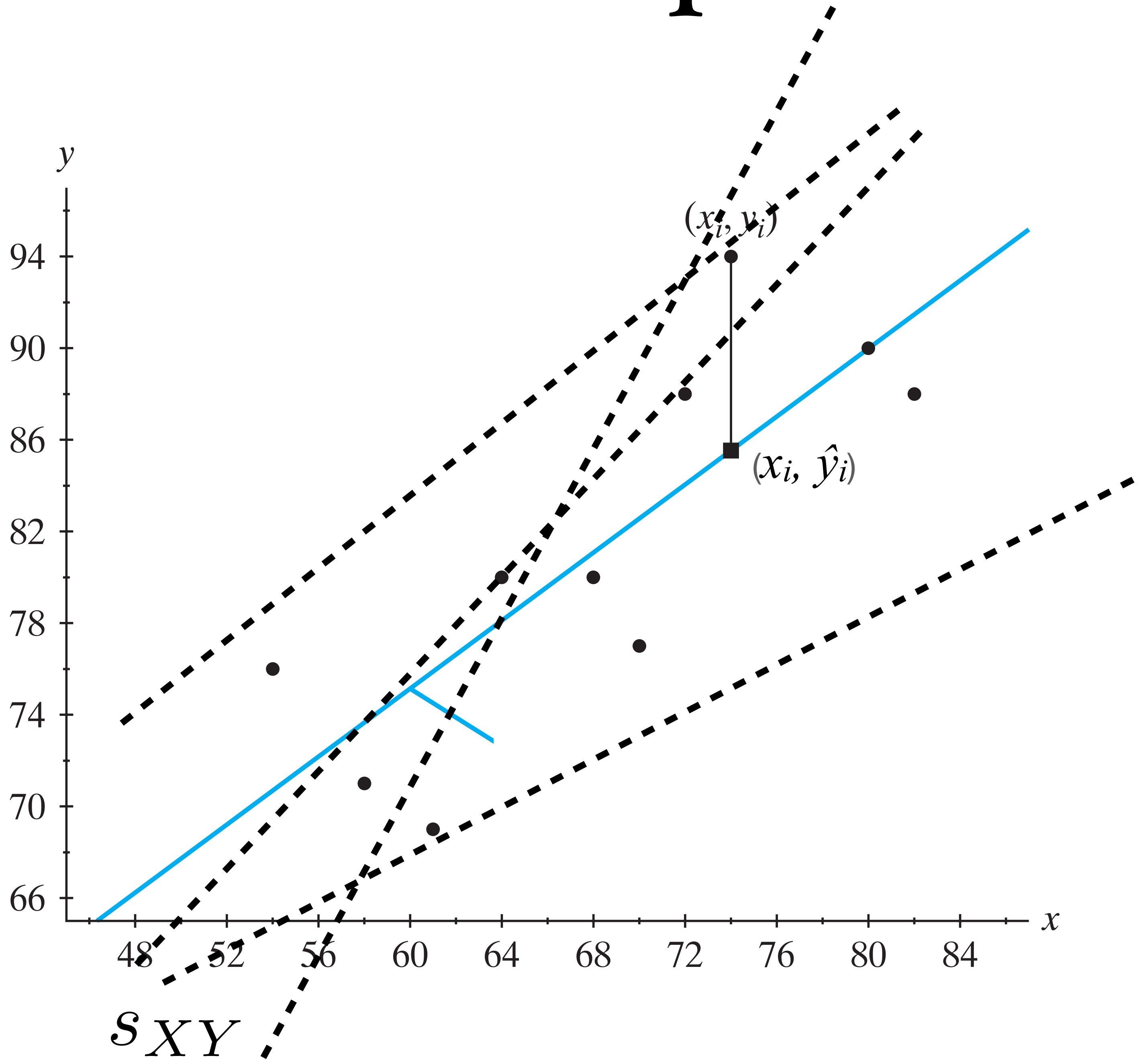


Linear Regression: Least Squares

- regression line $\hat{y}=\alpha x + c$
- Method of least squares:

$$\text{Minimize } \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

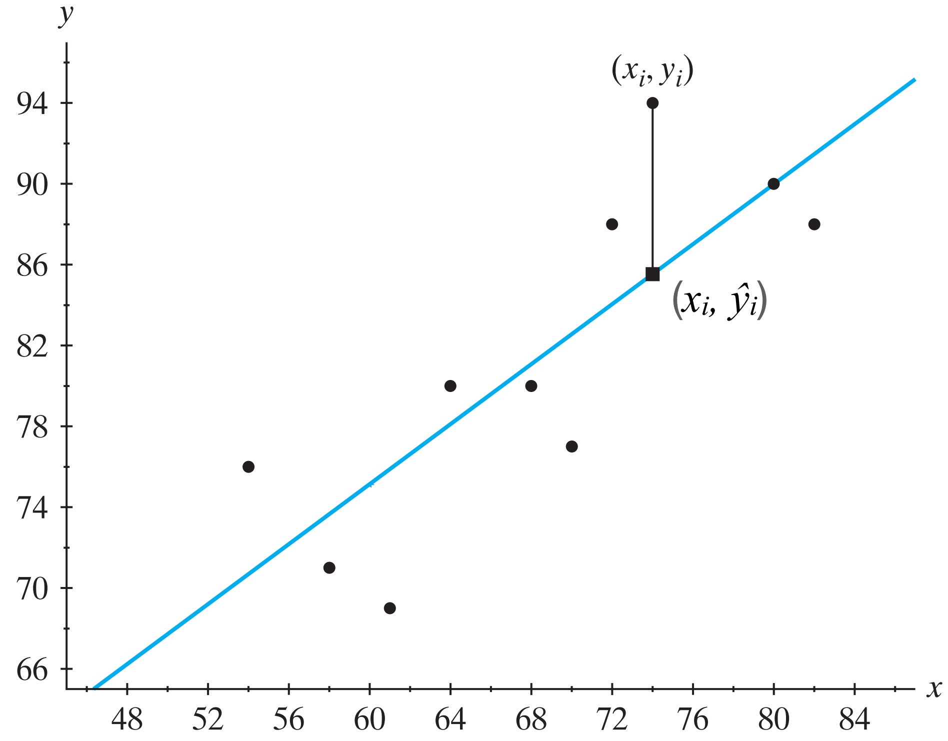
$$=\sum_{i=1}^{n}(y_i-\alpha x_i-c)^2$$



$$\alpha=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{s_{XY}}{s_X^2}$$

$$c=\bar{y}-\alpha\bar{x}$$

Linear Regression: MLE



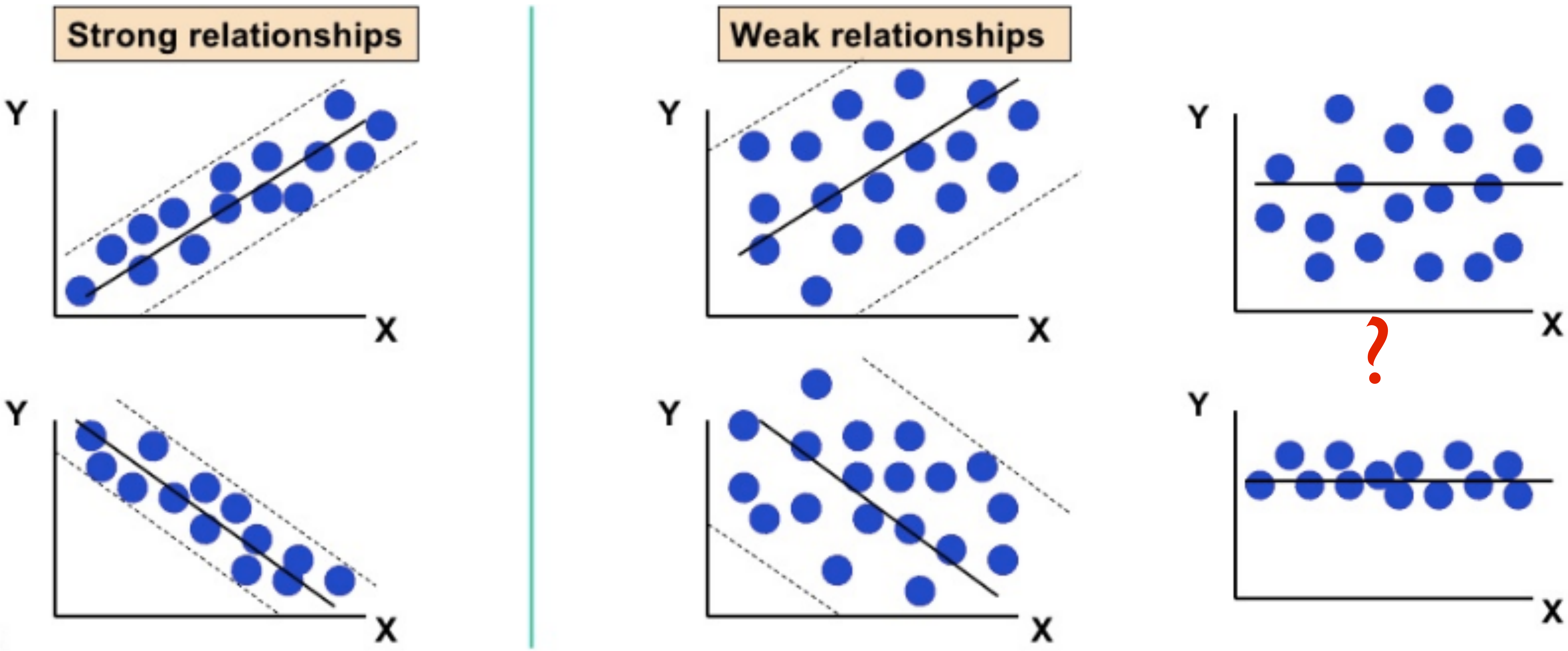
- *Y= aX + u + ε, where ε ~ N(0,σ²)*
- MLE:

$$L(a,u,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[-\frac{(y_i - ax_i - u)^2}{2\sigma^2}\Big]$$

$$\log L(a,u,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - ax_i - u)^2}{2\sigma^2}$$

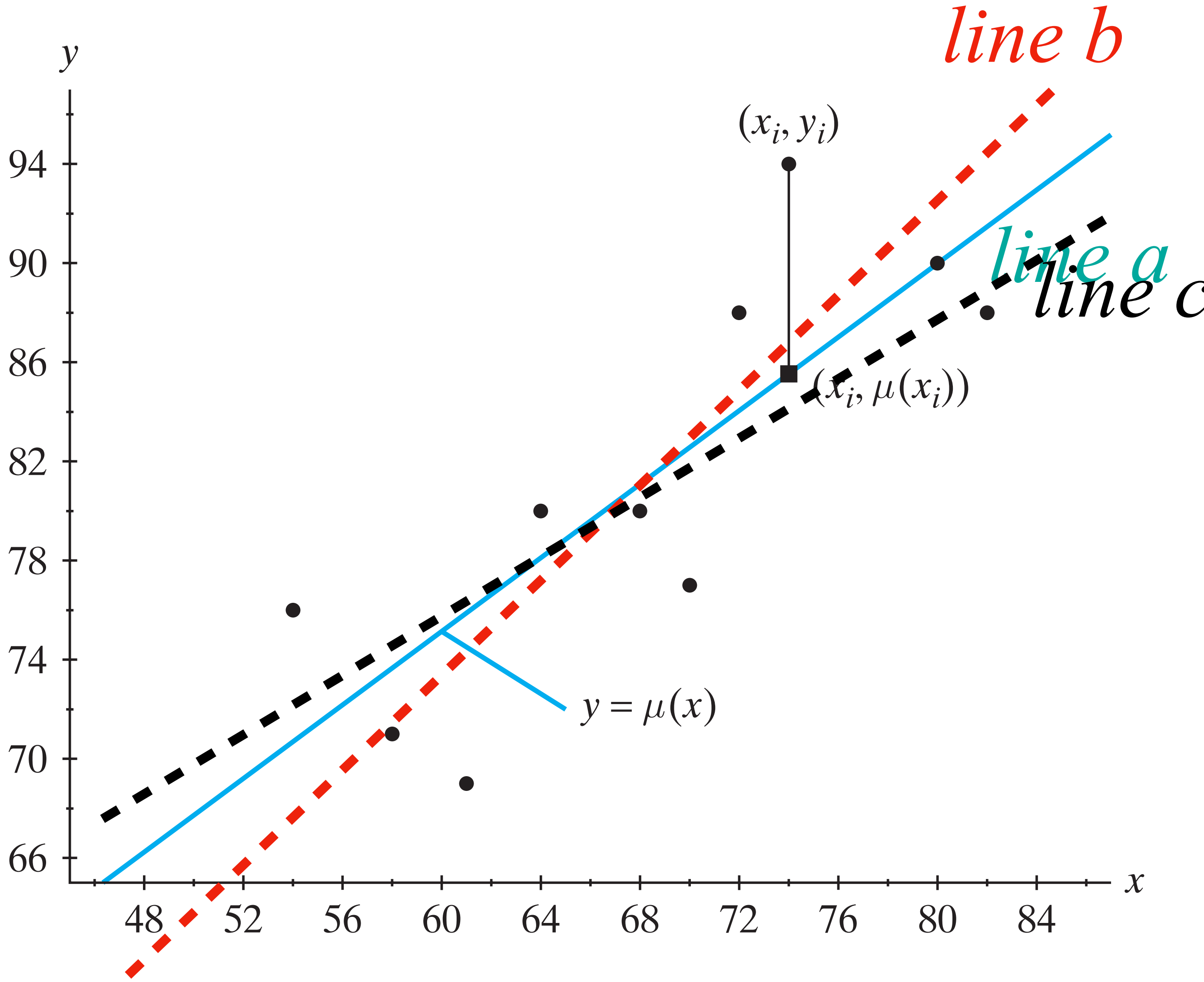
$$\hat{a} = \frac{s_{XY}}{s_X^2}, \quad \hat{u} = \bar{y} - \hat{a}\bar{x}, \quad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Linear Regression: Strong/ Weak Relations



Linear Regression: The Two Directions

- *Data generated by Y= aX + u + ε, where ε ~ N(0,σ²)*
- Regression line in the reverse direction: $\hat{x} = \beta y + c_2$
- Consider different situations...

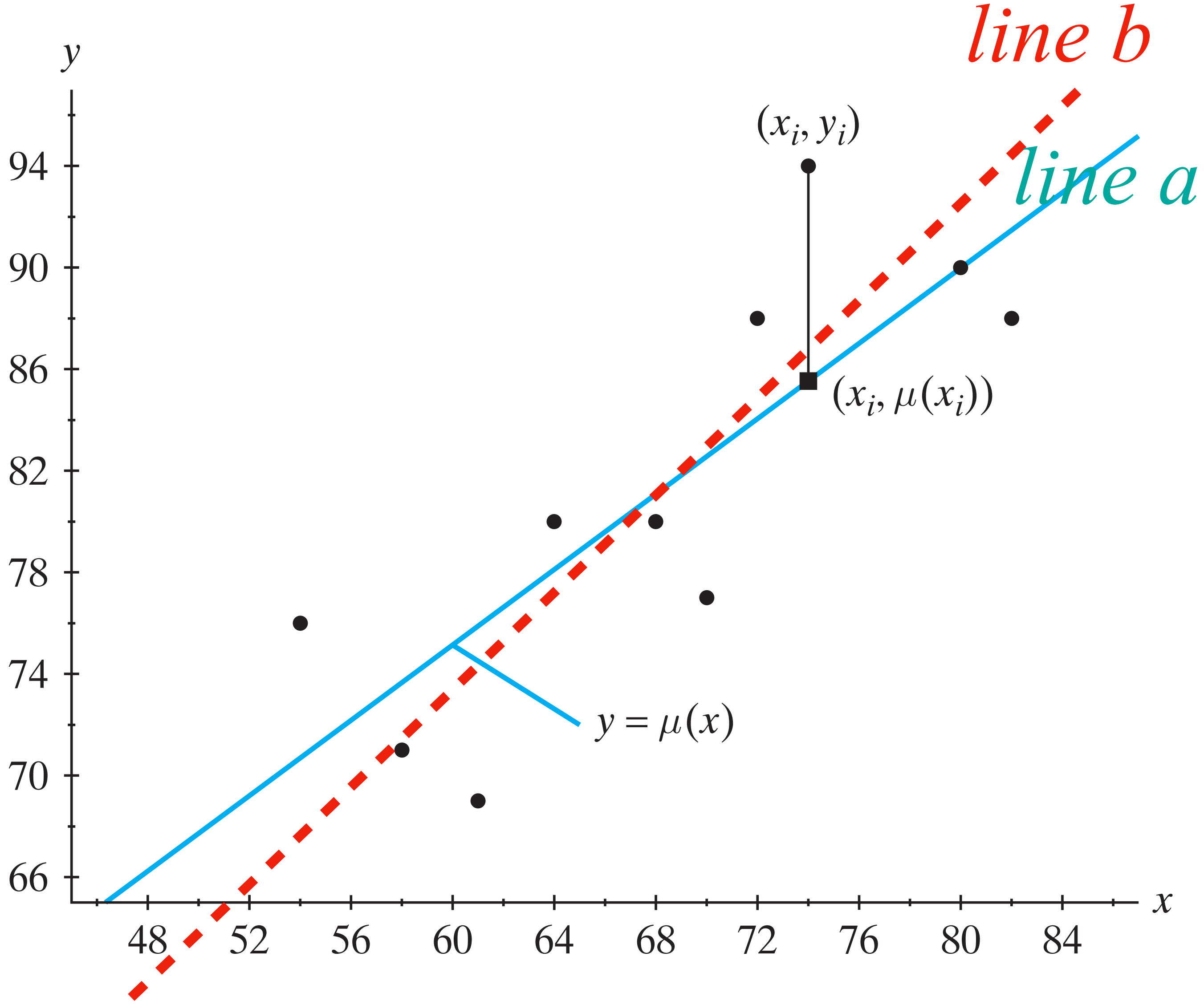


Question: 1. Interpretation of the parameter.
2. Are the regression lines from X to Y and from Y to X identical?

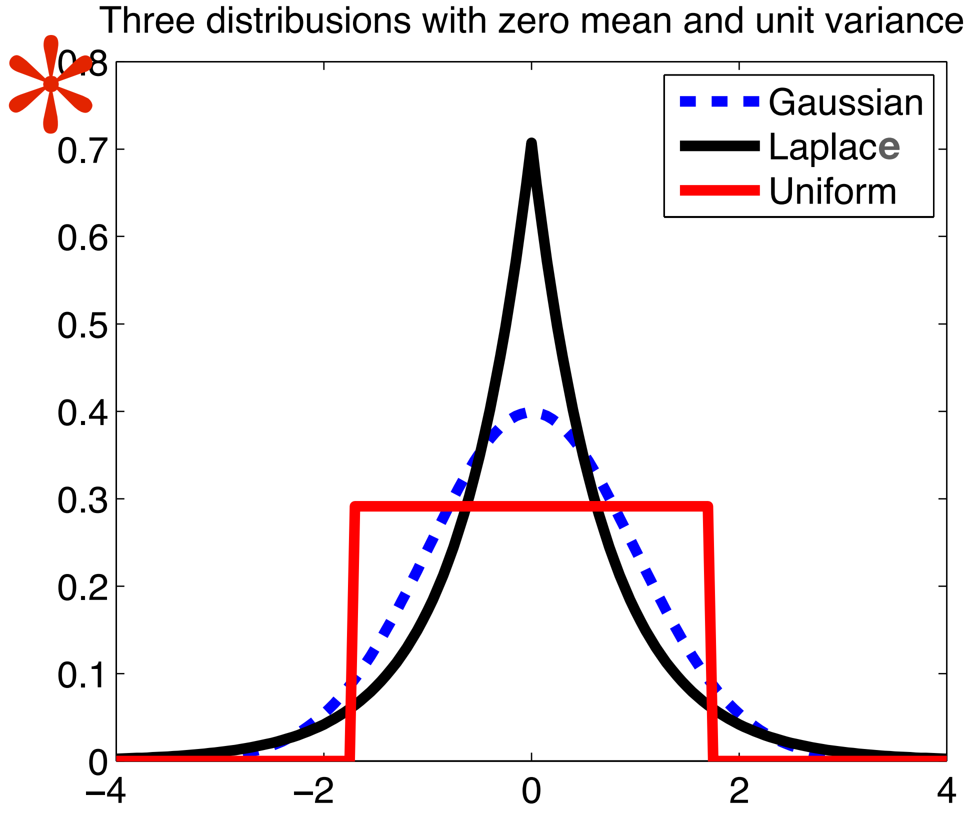
Linear Regression: The Two Directions

- *Data generated by Y= aX + u + ε, where ε ~ N(0,σ²)*
- Regression line in the reverse direction:

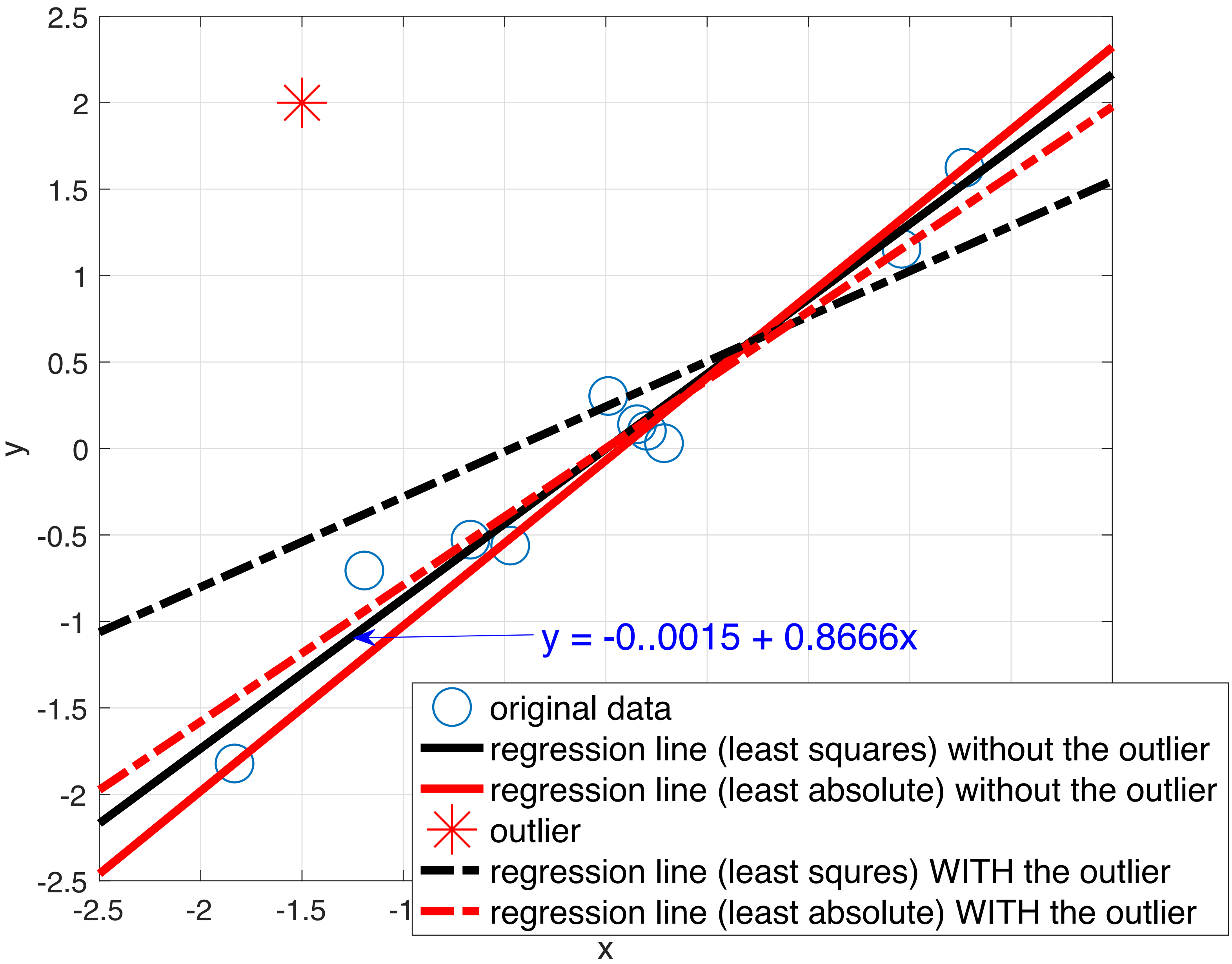
  $\hat{x} = \beta y + c_2$
- Consider different situations...



Question: 1. Interpretation of the parameter.
2. Are the regression lines from X to Y and from Y to X identical?

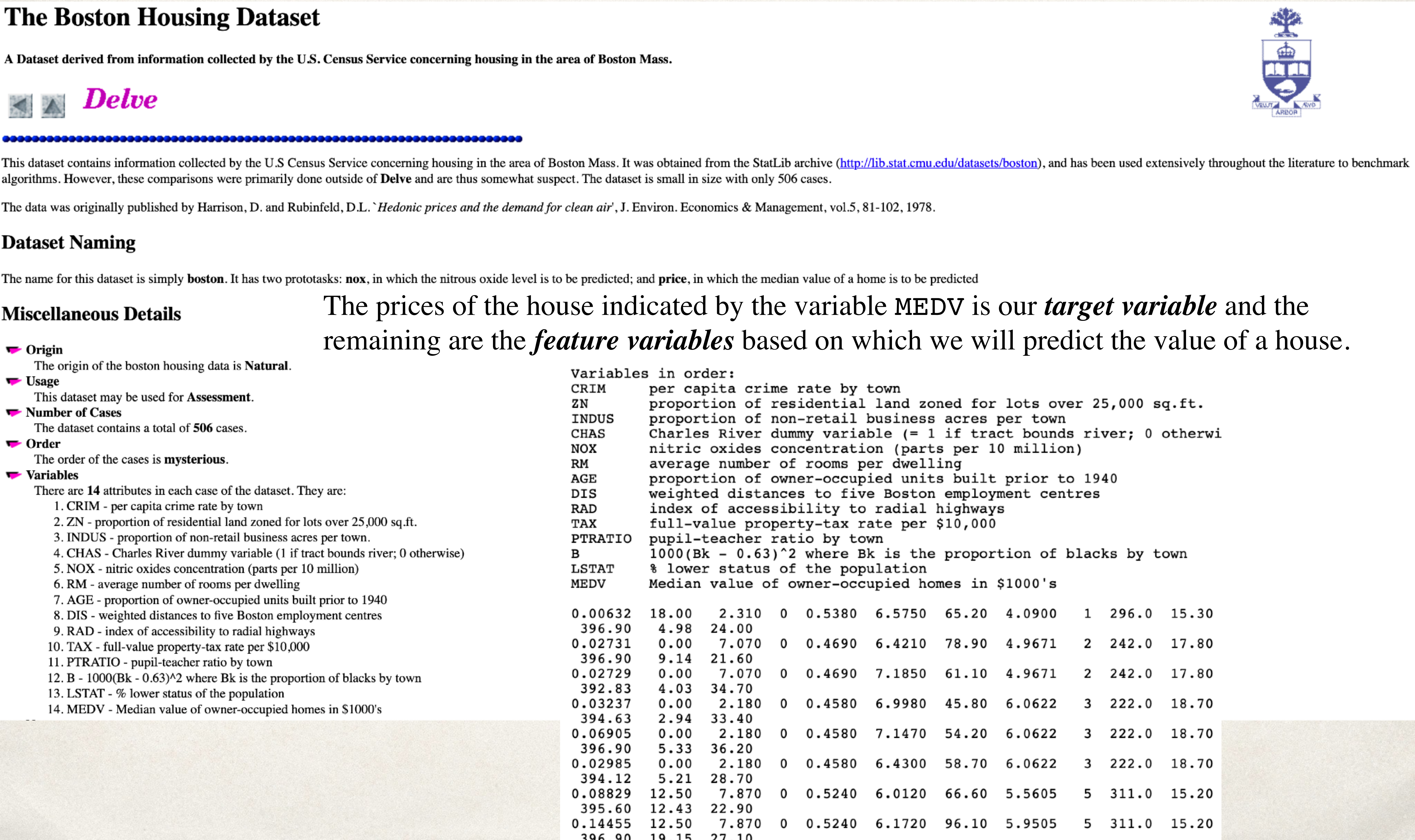


With Different Noise Distributions



- What if we use other distributions for the error in regression?
  - Laplace distribution?
  - What will you minimize?

Supervised Learning: An Example



The Boston Housing Dataset

A Dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston Mass.

Delve

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It was obtained from the StatLib archive (http://lib.stat.cmu.edu/datasets/boston), and has been used extensively throughout the literature to benchmark algorithms. However, these comparisons were primarily done outside of **Delve** and are thus somewhat suspect. The dataset is small in size with only 506 cases.

The data was originally published by Harrison, D. and Rubinfeld, D.L. `*Hedonic prices and the demand for clean air*', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Dataset Naming

The name for this dataset is simply **boston**. It has two prototasks: **nox**, in which the nitrous oxide level is to be predicted; and **price**, in which the median value of a home is to be predicted

Miscellaneous Details

- **Origin**
  The origin of the boston housing data is **Natural**.
- **Usage**
  This dataset may be used for **Assessment**.
- **Number of Cases**
  The dataset contains a total of **506** cases.
- **Order**
  The order of the cases is **mysterious**.
- **Variables**
  There are **14** attributes in each case of the dataset. They are:
  1. CRIM - per capita crime rate by town
  2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
  3. INDUS - proportion of non-retail business acres per town.
  4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
  5. NOX - nitric oxides concentration (parts per 10 million)
  6. RM - average number of rooms per dwelling
  7. AGE - proportion of owner-occupied units built prior to 1940
  8. DIS - weighted distances to five Boston employment centres
  9. RAD - index of accessibility to radial highways
  10. TAX - full-value property-tax rate per $10,000
  11. PTRATIO - pupil-teacher ratio by town
  12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
  13. LSTAT - % lower status of the population
  14. MEDV - Median value of owner-occupied homes in $1000's

The prices of the house indicated by the variable `MEDV` is our ***target variable*** and the remaining are the ***feature variables*** based on which we will predict the value of a house.

```
Variables in order:
CRIM     per capita crime rate by town
ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS    proportion of non-retail business acres per town
CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwi
NOX      nitric oxides concentration (parts per 10 million)
RM       average number of rooms per dwelling
AGE      proportion of owner-occupied units built prior to 1940
DIS      weighted distances to five Boston employment centres
RAD      index of accessibility to radial highways
TAX      full-value property-tax rate per $10,000
PTRATIO  pupil-teacher ratio by town
B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT    % lower status of the population
MEDV     Median value of owner-occupied homes in $1000's

0.00632  18.00   2.310  0  0.5380  6.5750  65.20  4.0900   1  296.0  15.30
 396.90   4.98  24.00
0.02731   0.00   7.070  0  0.4690  6.4210  78.90  4.9671   2  242.0  17.80
 396.90   9.14  21.60
0.02729   0.00   7.070  0  0.4690  7.1850  61.10  4.9671   2  242.0  17.80
 392.83   4.03  34.70
0.03237   0.00   2.180  0  0.4580  6.9980  45.80  6.0622   3  222.0  18.70
 394.63   2.94  33.40
0.06905   0.00   2.180  0  0.4580  7.1470  54.20  6.0622   3  222.0  18.70
 396.90   5.33  36.20
0.02985   0.00   2.180  0  0.4580  6.4300  58.70  6.0622   3  222.0  18.70
 394.12   5.21  28.70
0.08829  12.50   7.870  0  0.5240  6.0120  66.60  5.5605   5  311.0  15.20
 395.60  12.43  22.90
0.14455  12.50   7.870  0  0.5240  6.1720  96.10  5.9505   5  311.0  15.20
 396.90  19.15  27.10
```

Multiple Regression

- Regress Y on $\boldsymbol{X} = (X_1,X_2)^T$
- $\hat{y}=\alpha_1x_1 + \alpha_2x_2 + c$
- What if $(x_{11}, x_{12,\dots,}\, x_{1N})$ and $(x_{21}, x_{22,\dots,}\, x_{2N})$ are linearly dependent?
- Least squares
- In matrix form

$$\mathbf{X} = \begin{array}{cc} & \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{array}{c} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \\ \mathbf{x}_N \end{array} & \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1N} & x_{2N} \end{bmatrix} \end{array}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Multiple Regression

- Regress Y on $\boldsymbol{X} = (X_1, X_2)^T$
- $\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + c$
- For simplicity, *assume all variables have zero mean*

Minimize $S_E = (\mathbf{y} - \mathbf{x}\boldsymbol{\alpha})^\intercal (\mathbf{y} - \mathbf{x}\boldsymbol{\alpha})$

$$\frac{\partial S_E}{\partial \boldsymbol{\alpha}} = 2 \cdot \mathbf{x}^\intercal (\mathbf{y} - \mathbf{x}\boldsymbol{\alpha})$$

If $\mathbf{x}^\intercal \mathbf{x}$ is invertible, setting $\frac{\partial S_E}{\partial \boldsymbol{\alpha}} = 0$

$$\Rightarrow \quad \boldsymbol{\alpha} = (\mathbf{x}^\intercal \mathbf{x})^{-1} (\mathbf{x}^\intercal \mathbf{y})$$

$$\begin{array}{cc} & \begin{array}{cc} X_1 & X_2 \end{array} \\ \mathbf{x} = & \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1N} & x_{2N} \end{bmatrix} \end{array}$$

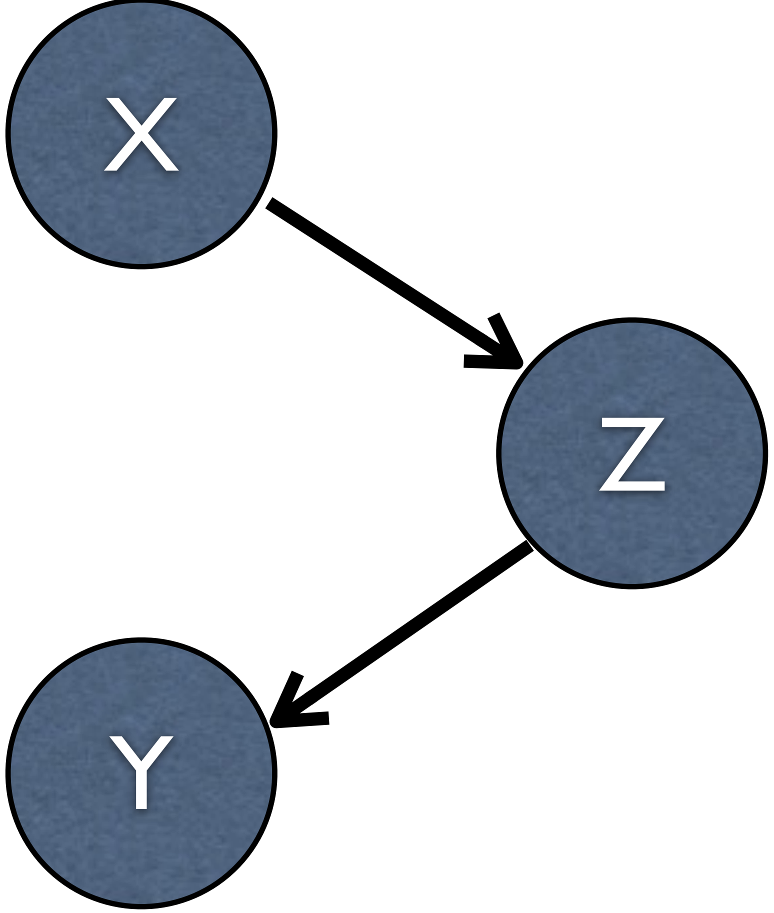
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Simple Regression vs. Multiple Regression

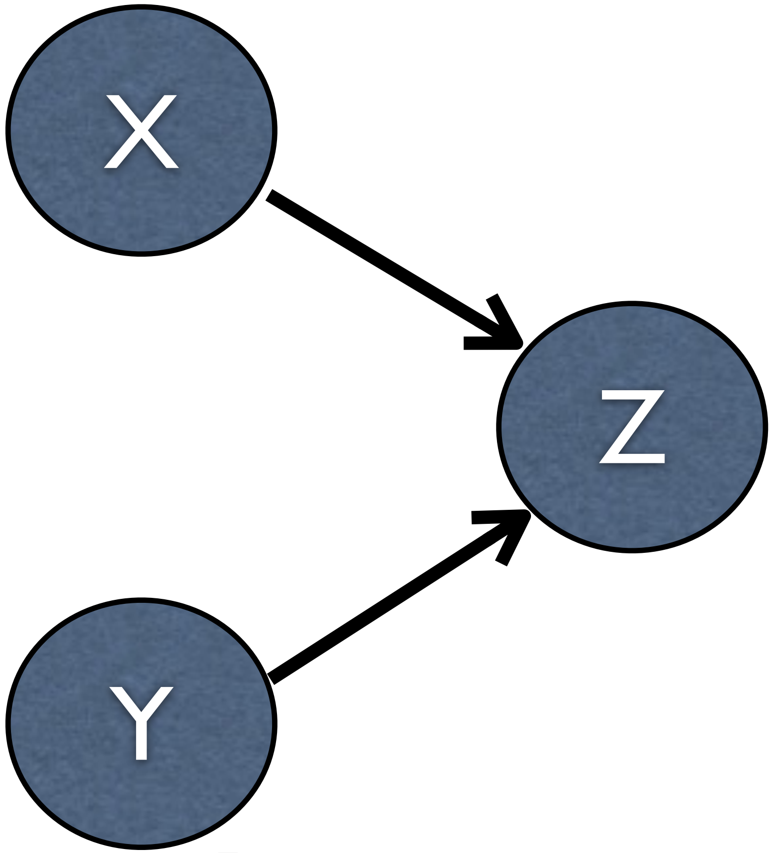
- Let's do simple regression from X to Y: $\hat{y}=\alpha x + c$
- Will α be zero?

Independence vs. conditional independence ;-) and you can see it from graph!

- Let's do regression from $(X,Z)^T$ to Y: $\hat{y}=\alpha_1 x + \alpha_2 z + c$
- Will the coefficient of x be zero?



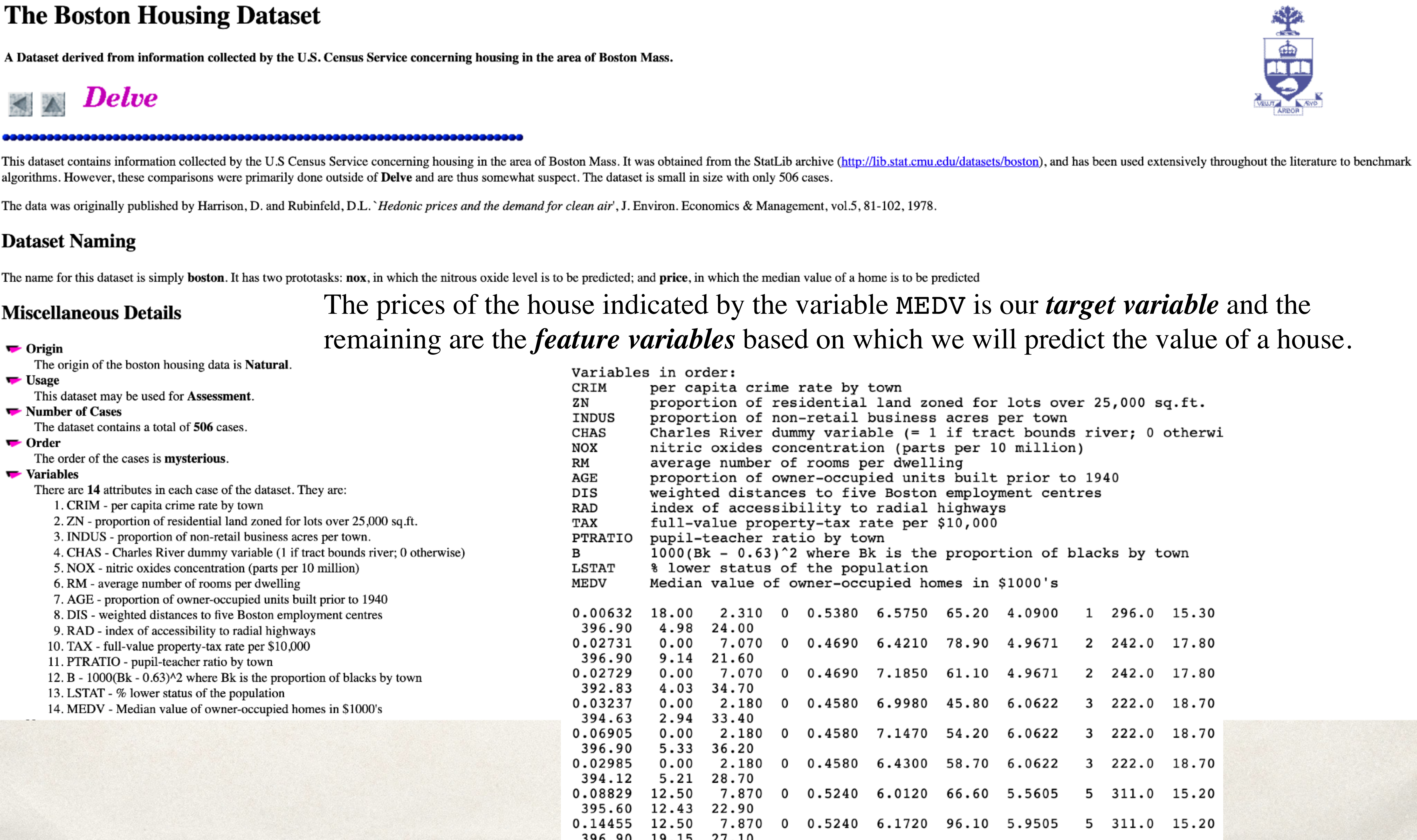
Process 1



Process 2

What is Next:
Nonlinear Regression

Supervised Learning: An Example



The Boston Housing Dataset

A Dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston Mass.

Delve

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It was obtained from the StatLib archive (http://lib.stat.cmu.edu/datasets/boston), and has been used extensively throughout the literature to benchmark algorithms. However, these comparisons were primarily done outside of **Delve** and are thus somewhat suspect. The dataset is small in size with only 506 cases.

The data was originally published by Harrison, D. and Rubinfeld, D.L. `*Hedonic prices and the demand for clean air*', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Dataset Naming

The name for this dataset is simply **boston**. It has two prototasks: **nox**, in which the nitrous oxide level is to be predicted; and **price**, in which the median value of a home is to be predicted

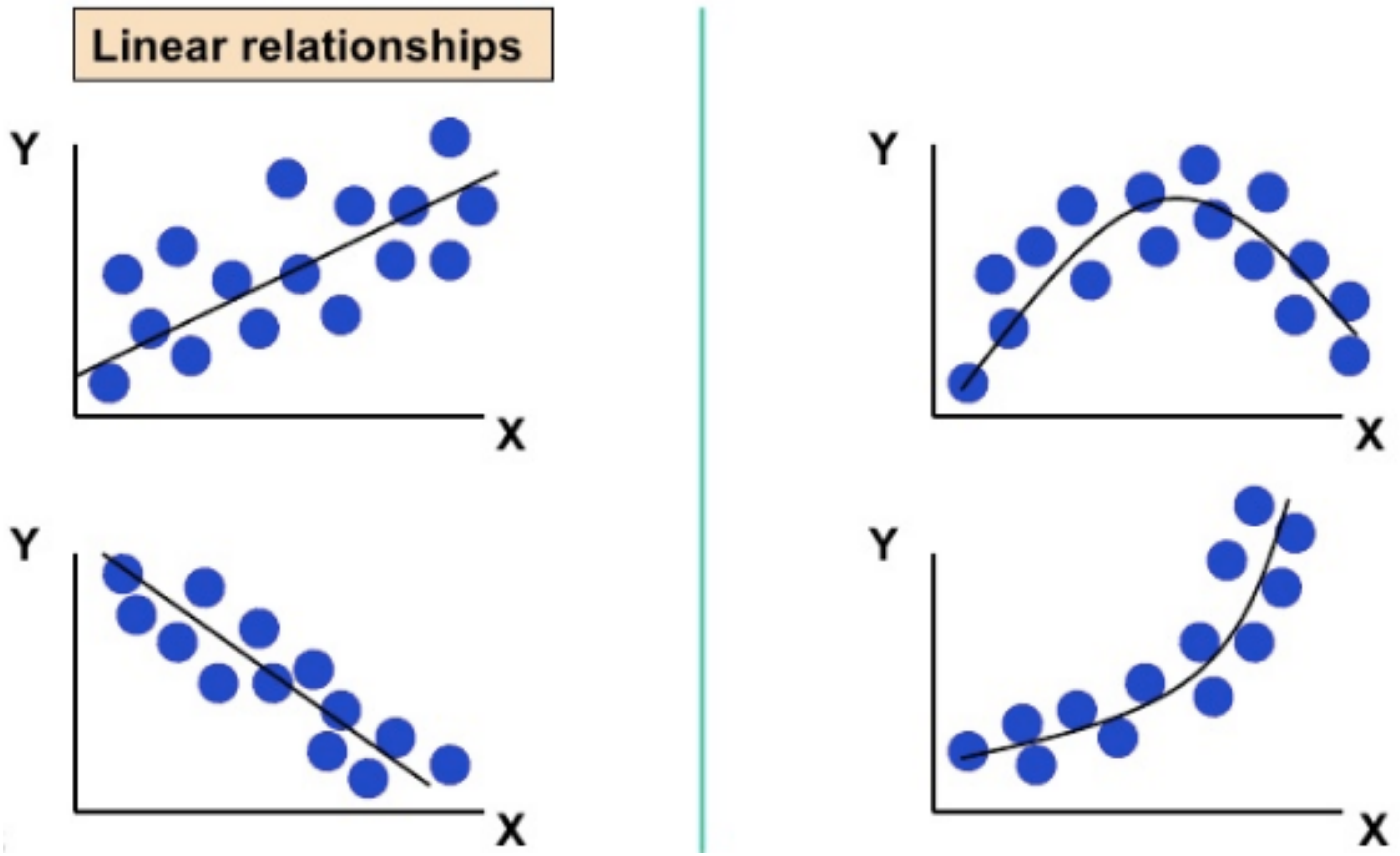
The prices of the house indicated by the variable `MEDV` is our ***target variable*** and the remaining are the ***feature variables*** based on which we will predict the value of a house.

Miscellaneous Details

- **Origin**
  The origin of the boston housing data is **Natural**.
- **Usage**
  This dataset may be used for **Assessment**.
- **Number of Cases**
  The dataset contains a total of **506** cases.
- **Order**
  The order of the cases is **mysterious**.
- **Variables**
  There are **14** attributes in each case of the dataset. They are:
  1. CRIM - per capita crime rate by town
  2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
  3. INDUS - proportion of non-retail business acres per town.
  4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
  5. NOX - nitric oxides concentration (parts per 10 million)
  6. RM - average number of rooms per dwelling
  7. AGE - proportion of owner-occupied units built prior to 1940
  8. DIS - weighted distances to five Boston employment centres
  9. RAD - index of accessibility to radial highways
  10. TAX - full-value property-tax rate per $10,000
  11. PTRATIO - pupil-teacher ratio by town
  12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
  13. LSTAT - % lower status of the population
  14. MEDV - Median value of owner-occupied homes in $1000's

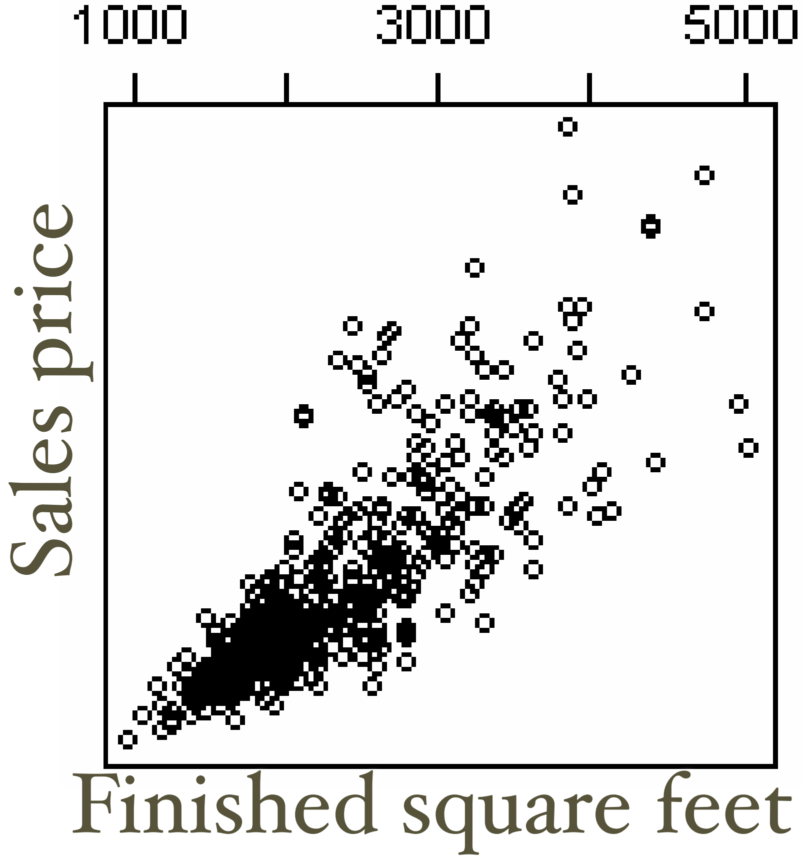
```
Variables in order:
CRIM     per capita crime rate by town
ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS    proportion of non-retail business acres per town
CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwi
NOX      nitric oxides concentration (parts per 10 million)
RM       average number of rooms per dwelling
AGE      proportion of owner-occupied units built prior to 1940
DIS      weighted distances to five Boston employment centres
RAD      index of accessibility to radial highways
TAX      full-value property-tax rate per $10,000
PTRATIO  pupil-teacher ratio by town
B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT    % lower status of the population
MEDV     Median value of owner-occupied homes in $1000's

0.00632  18.00   2.310  0  0.5380  6.5750  65.20  4.0900   1  296.0  15.30
 396.90   4.98  24.00
0.02731   0.00   7.070  0  0.4690  6.4210  78.90  4.9671   2  242.0  17.80
 396.90   9.14  21.60
0.02729   0.00   7.070  0  0.4690  7.1850  61.10  4.9671   2  242.0  17.80
 392.83   4.03  34.70
0.03237   0.00   2.180  0  0.4580  6.9980  45.80  6.0622   3  222.0  18.70
 394.63   2.94  33.40
0.06905   0.00   2.180  0  0.4580  7.1470  54.20  6.0622   3  222.0  18.70
 396.90   5.33  36.20
0.02985   0.00   2.180  0  0.4580  6.4300  58.70  6.0622   3  222.0  18.70
 394.12   5.21  28.70
0.08829  12.50   7.870  0  0.5240  6.0120  66.60  5.5605   5  311.0  15.20
 395.60  12.43  22.90
0.14455  12.50   7.870  0  0.5240  6.1720  96.10  5.9505   5  311.0  15.20
 396.90  19.15  27.10
```

Types of Relationships



Nonlinear Relationships

- $Y = a_1X + a_2X^2 + \varepsilon$
- $Y = a_1X_1 + a_2X_1X_2 + a_3X_2^2 + \varepsilon$
- $Y = a_0 \exp(a_1X + \varepsilon)$
  - What if you consider $\log Y$ and X (suppose $a_0>0$)?



linear or nonlinear?

Polynomial Regression

- *m*-th order polynomial regression $\hat{y} = f(x; \boldsymbol{\alpha})$ is given by

$$f(x; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + ... + \alpha_m x^m$$

- If *m* is larger than 1, it is *nonlinear* in *x*,
- but *linear* in $\boldsymbol{\alpha}$
- How to estimate $\boldsymbol{\alpha}$?

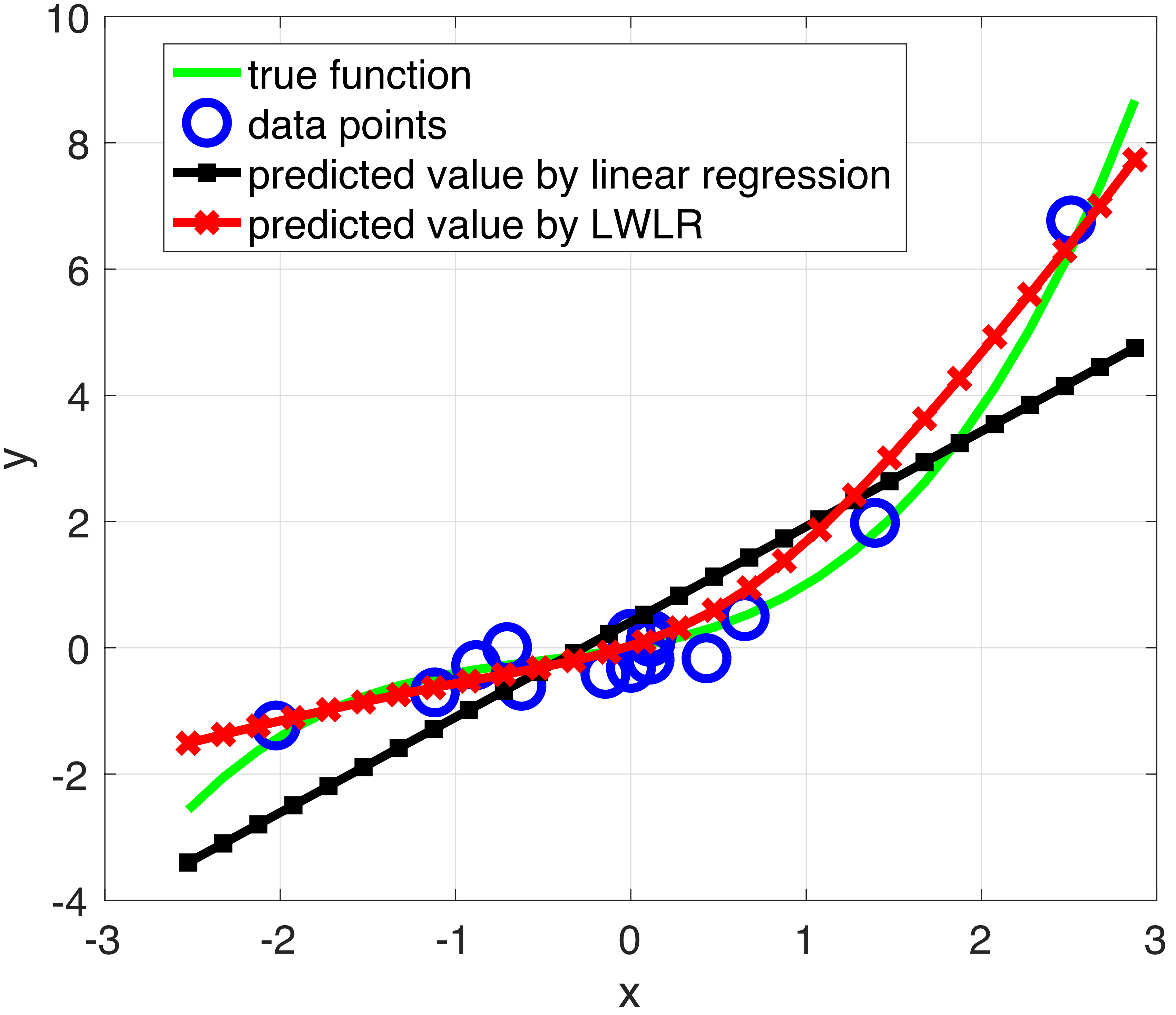
Additive Models

- More generally, predictions can be based on a linear combination of a set of basis functions $\{\phi_1(\mathbf{x}), ..., \phi_m(\mathbf{x})\}$:

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 \phi_1(\mathbf{x}) + ... + \alpha_m \phi_m(\mathbf{x})$$

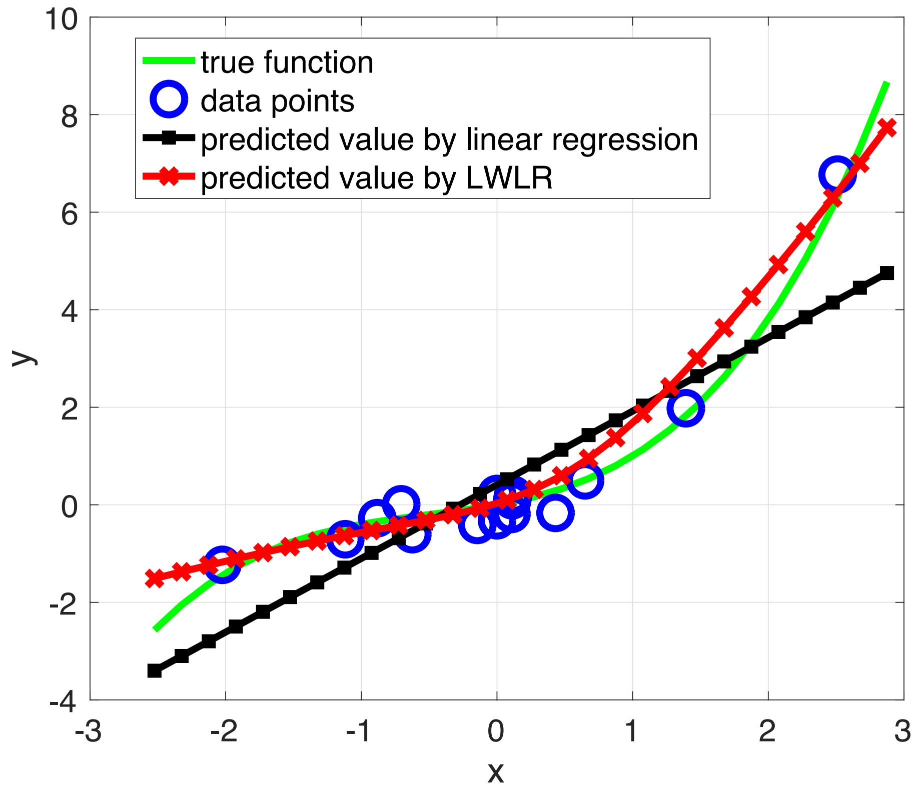
- Examples…
- How to estimate **α**?

Locally Weighted Linear Regression



(see the demo with 'LocallyWeightedLinearRegression.m')

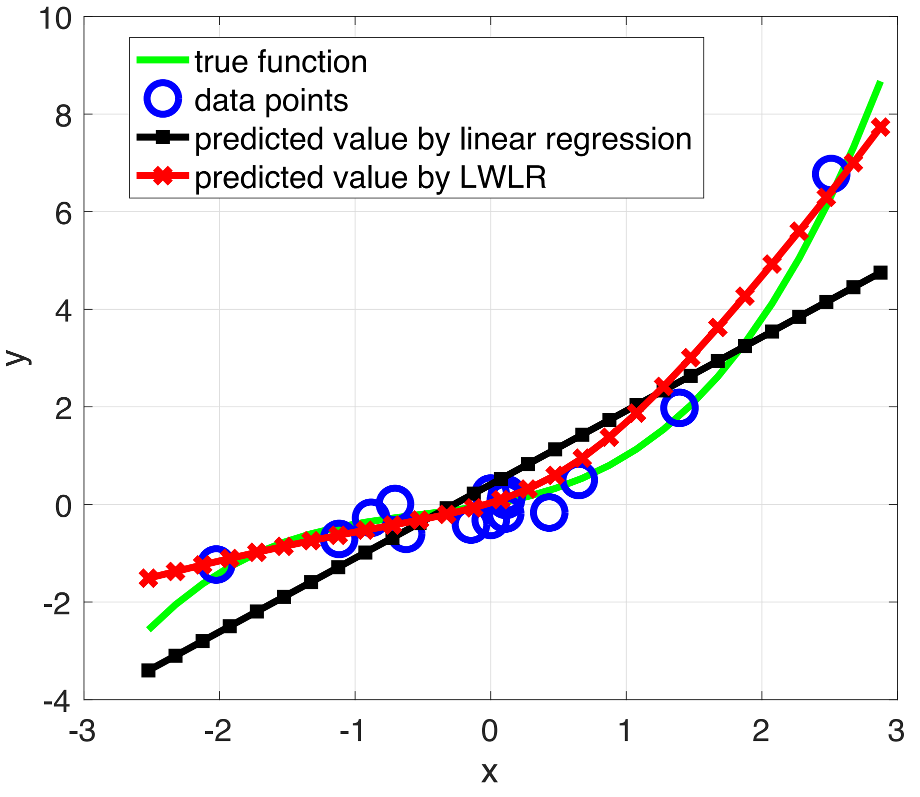
Locally Weighted Linear Regression



- Linear regression
  - Find parameter α to minimize squared error $\sum_{i=1}^{n} (y_i - \alpha^\intercal \mathbf{x}_i)^2$
  - Predicted value is $\hat{y} = \alpha^T \mathbf{x}$
- Locally weighted linear regression
  - Find parameter α to minimize *weighted* error $\sum_{i=1}^{n} w_i (y_i - \alpha^\intercal \mathbf{x}_i)^2$
  - Predicted value is $\hat{y} = \alpha^T \mathbf{x}$
  - w_i: non-negative valued weights; a typical choice is

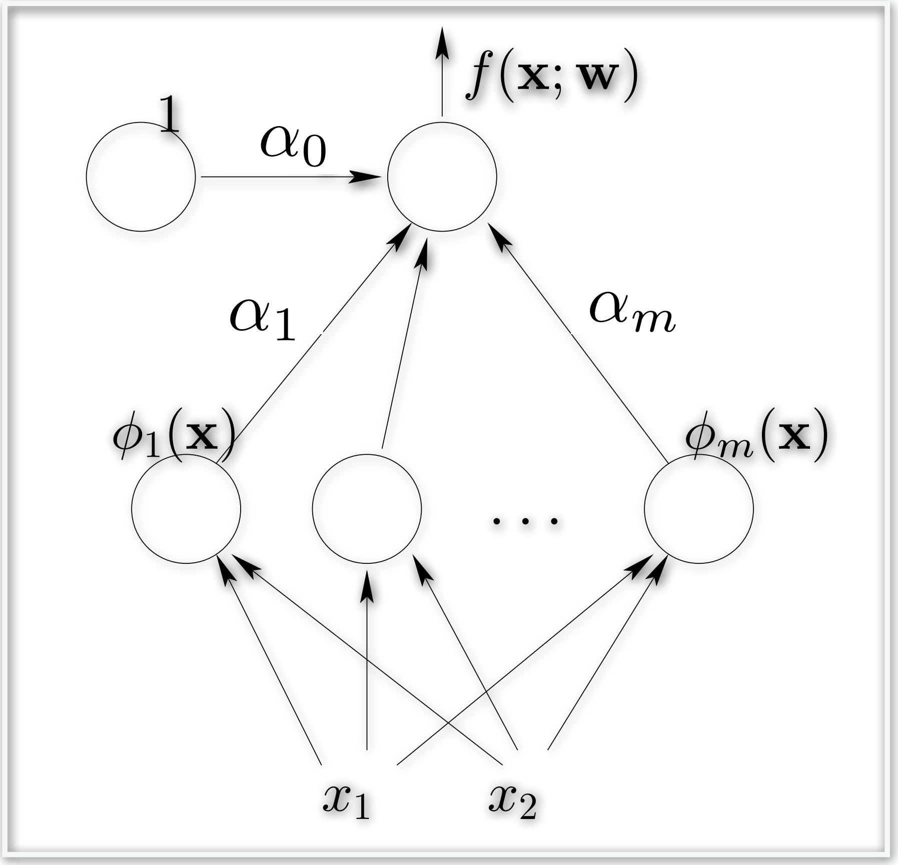
$$w_i = e^{-\frac{||\mathbf{x}_i - \mathbf{x}||^2}{2\tau^2}}$$

Parametric vs. Nonparametric Models



- Linear regression algorithm: *parametric*
  - has a fixed, finite number of parameters
  - Once they are learned and stored, we no longer need the training data for future predictions
- Locally weighted linear regression
  - For making prediction, we need to keep the entire training set around
  - *Nonparametric*: the amount of stuff we need to keep in order to represent the model grows with the size of the training set

With Radial Basis Functions

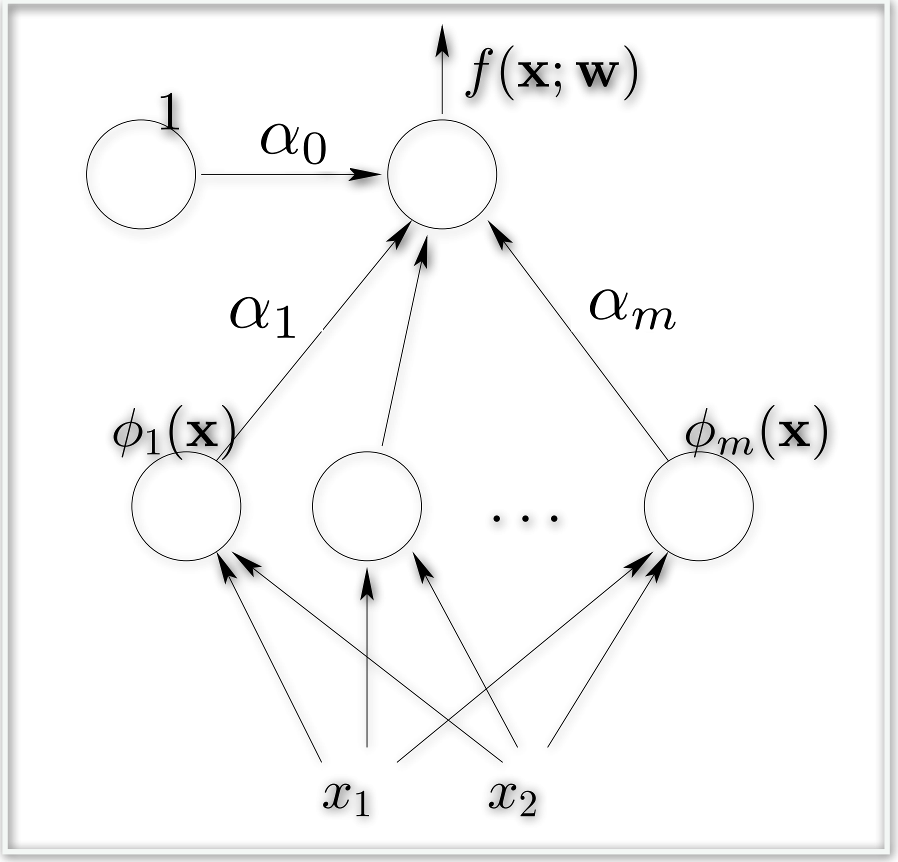


$$f(\mathbf{x};\boldsymbol{\alpha}) = \alpha_0 + \alpha_1\phi_1(\mathbf{x}) + ... + \alpha_m\phi_m(\mathbf{x})$$

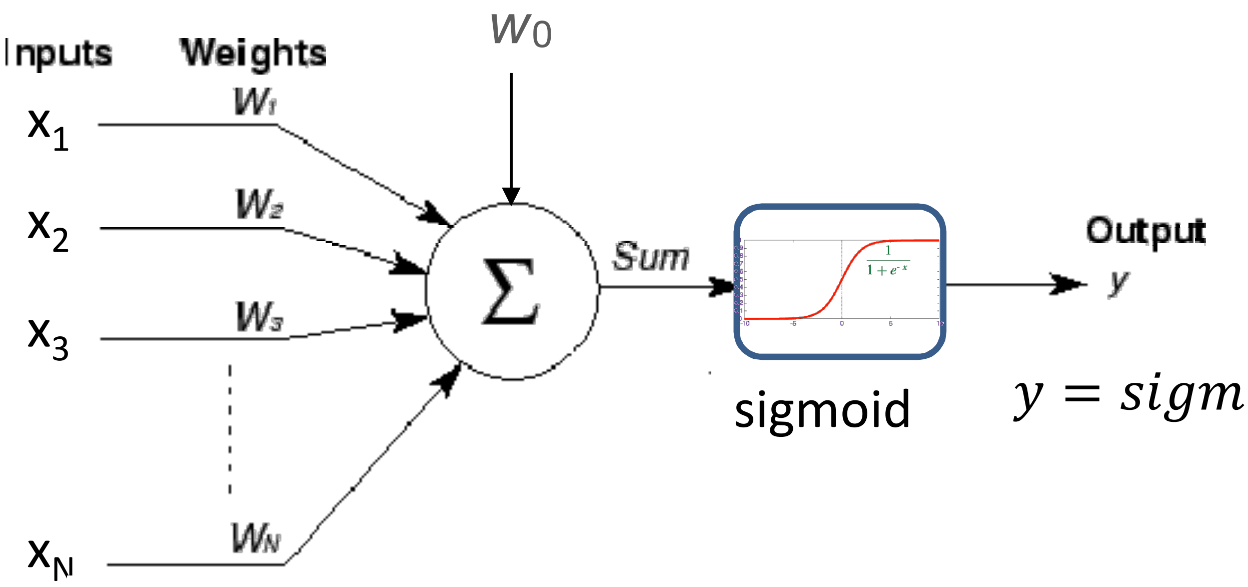
- Can also make predictions by gauging the similarity of new examples to “prototypes”, with “radial basis functions” measuring the similar to a “prototype”:

$$\phi_k(\mathbf{x}) = e^{-\frac{||\mathbf{x}-\mathbf{x}_k||^2}{2\tau^2}}$$

- Training data points themselves could serve as prototypes

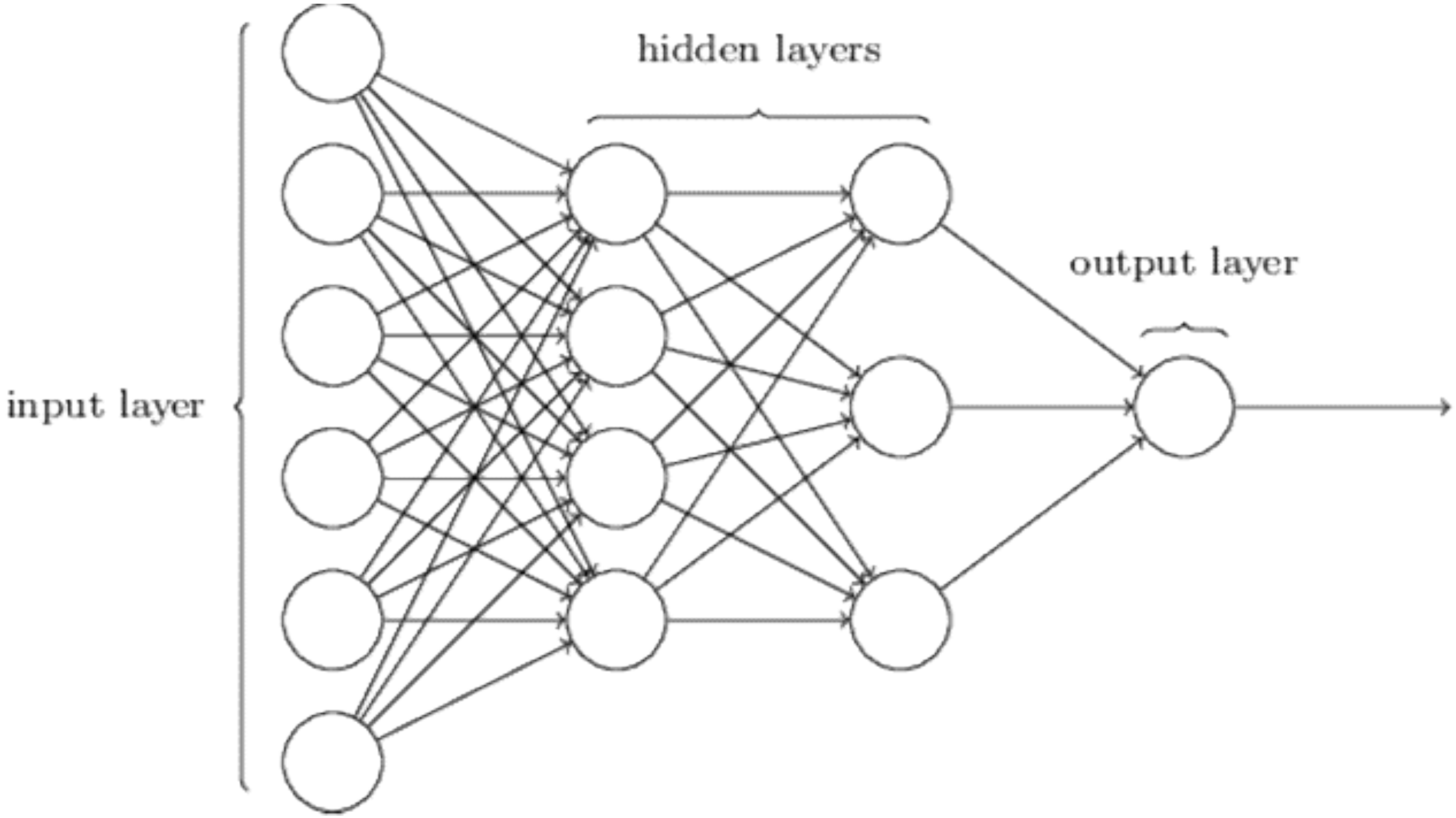


Neural Networks



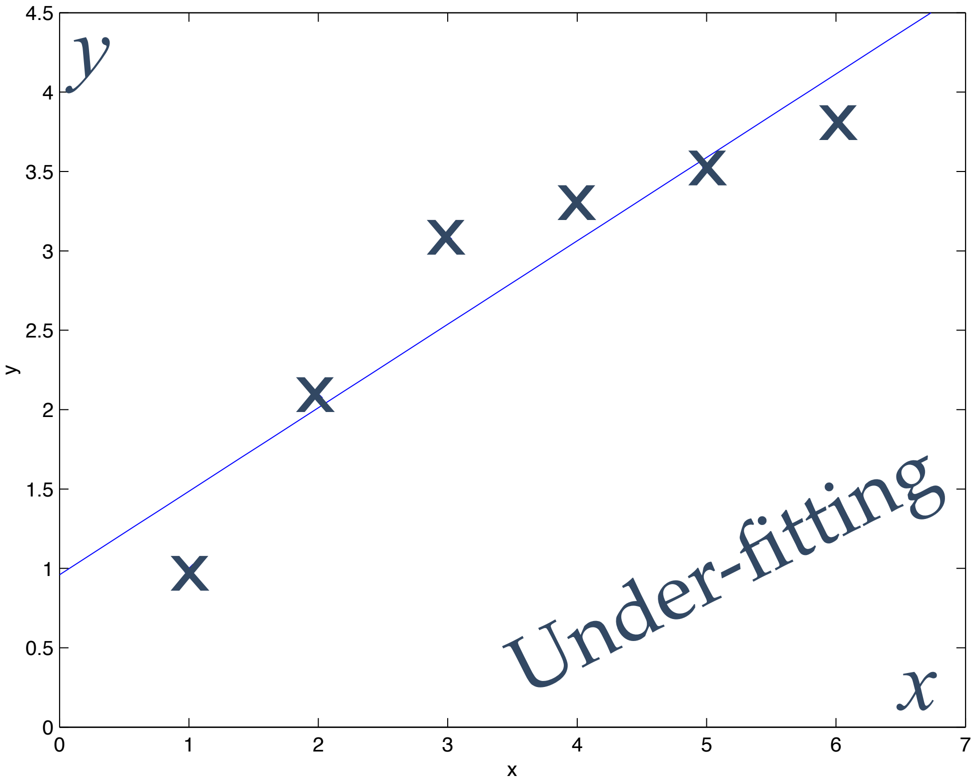
$$y = sigmoid(\sum_i w_i x_i - b)$$

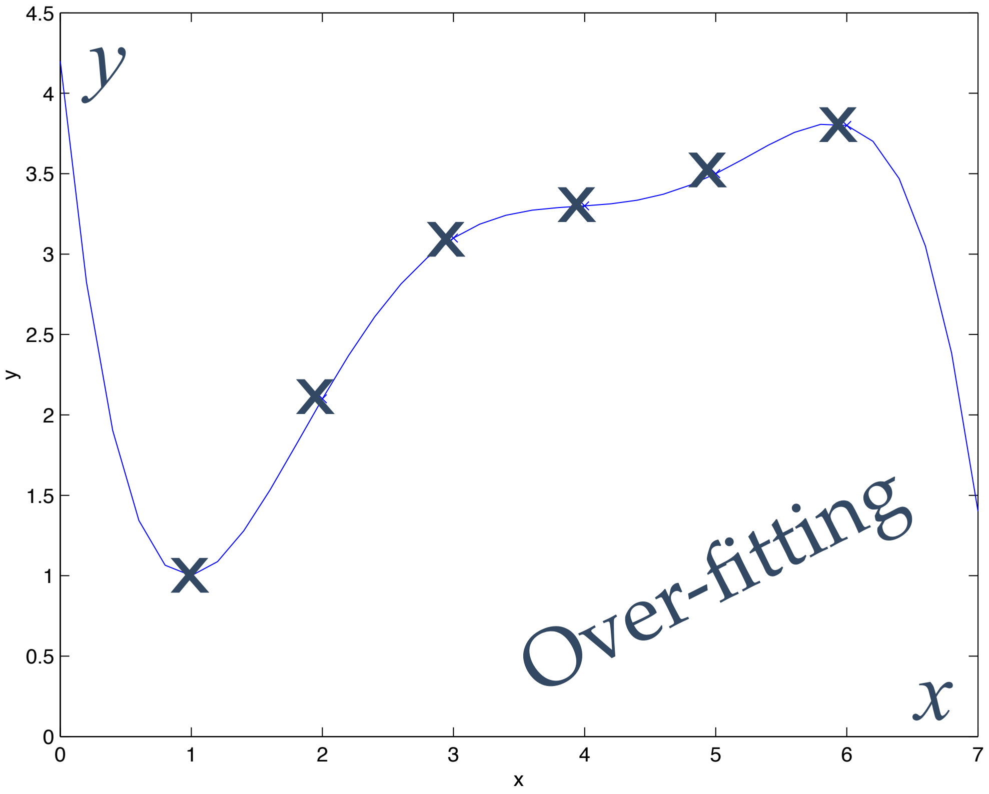
- In neural networks the basis functions themselves have *adjustable* parameter
- and it can have multiple layers

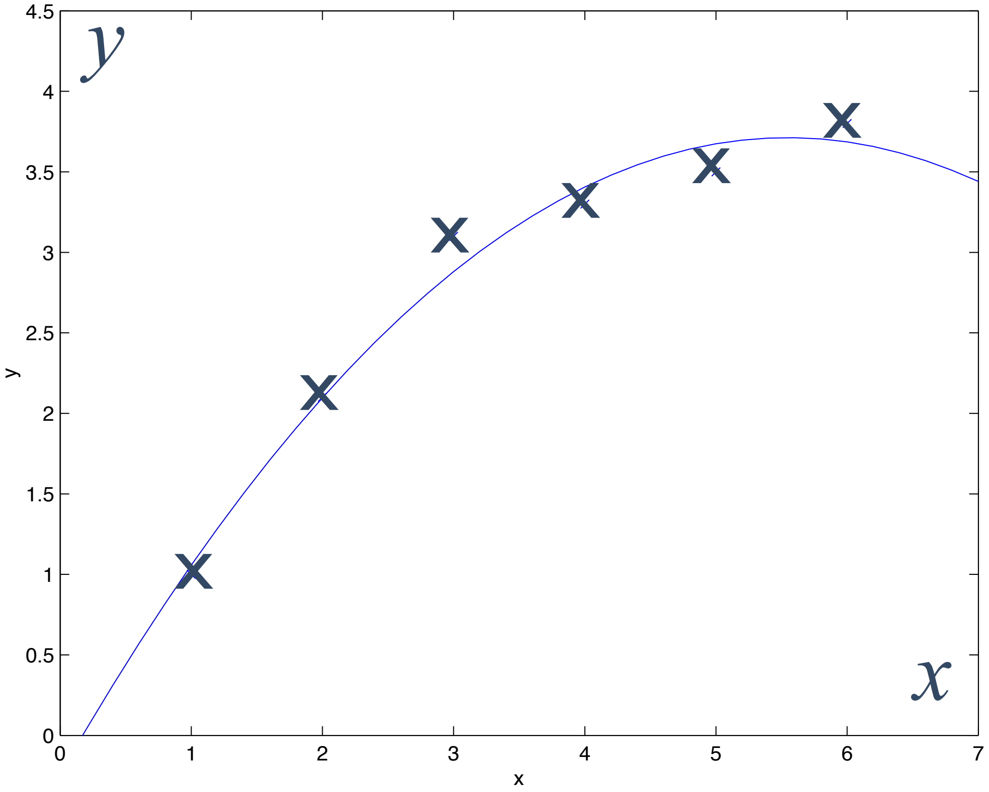


Basic Ideas of Model Selection

Avoiding under-fitting and over-fitting



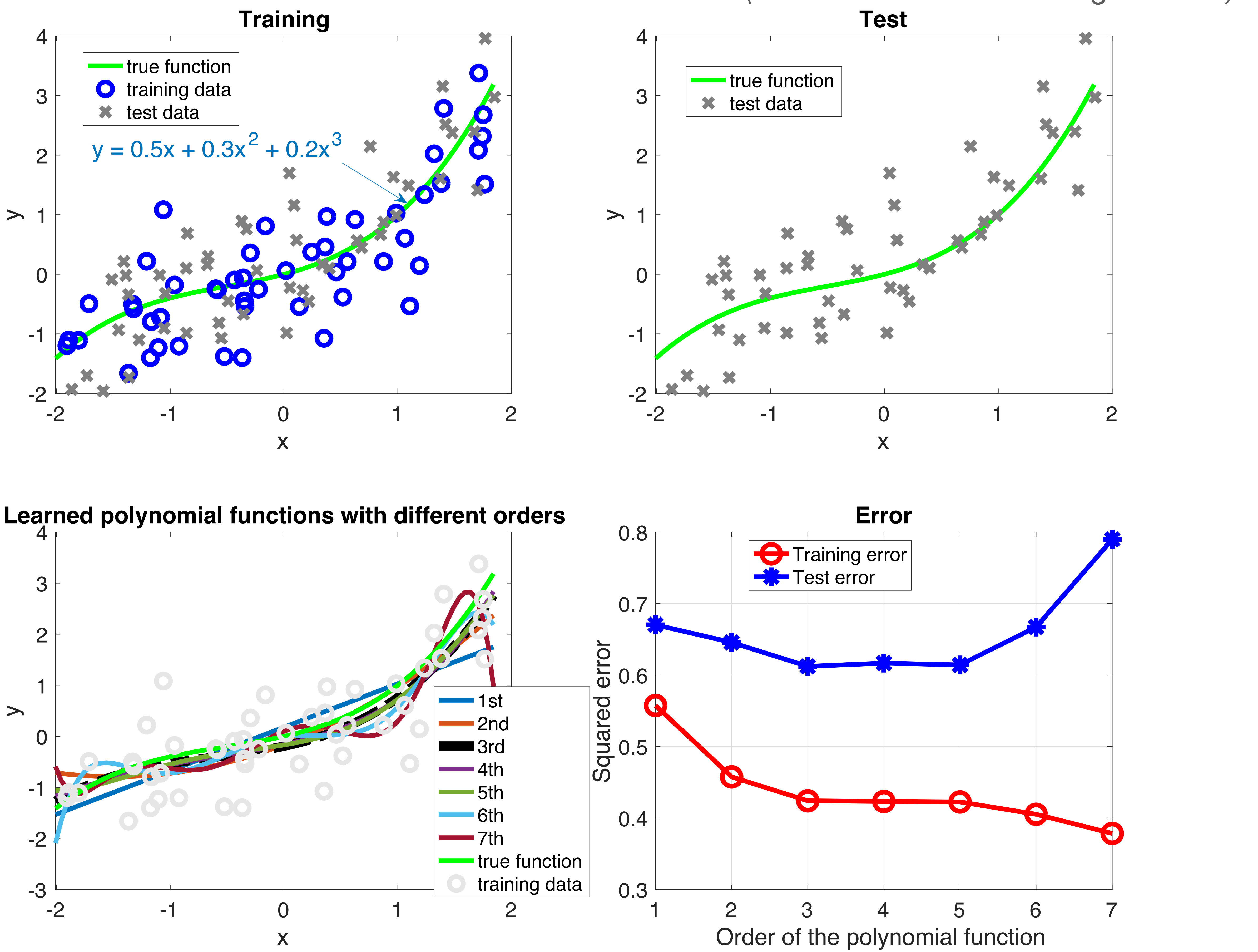




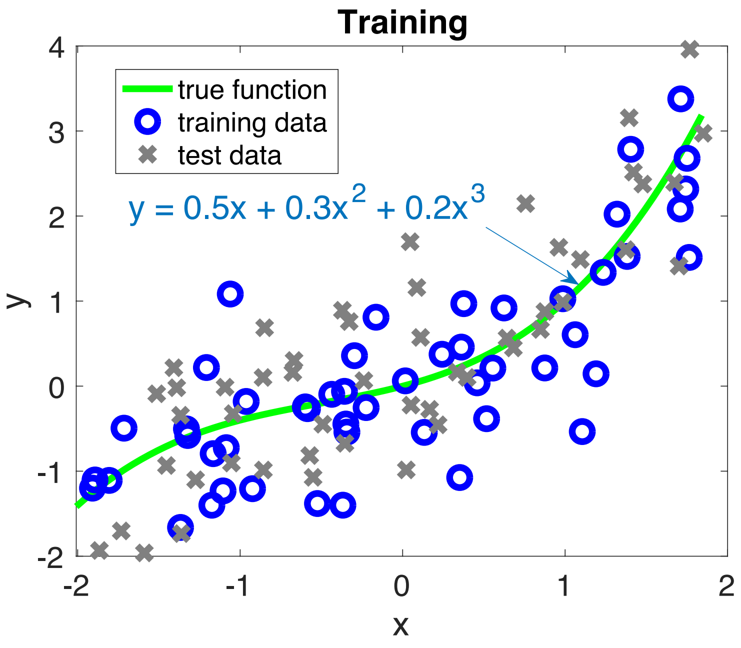
Why?

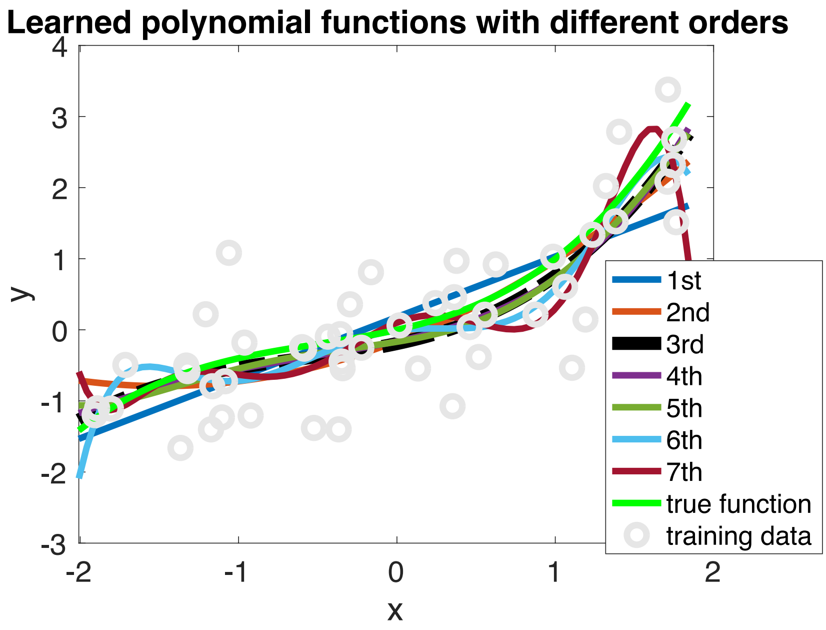
Machine Learning Cares about the Performance on New, Unseen Data

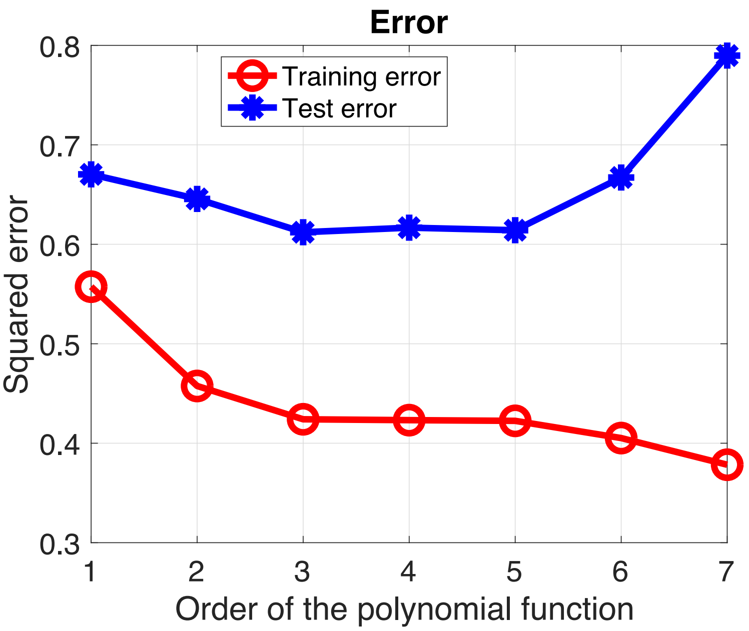
(see the demo with ‘trainingvstest.m’)



Machine Learning Cares about the Performance on New, Unseen Data


Training


Learned polynomial functions with different orders

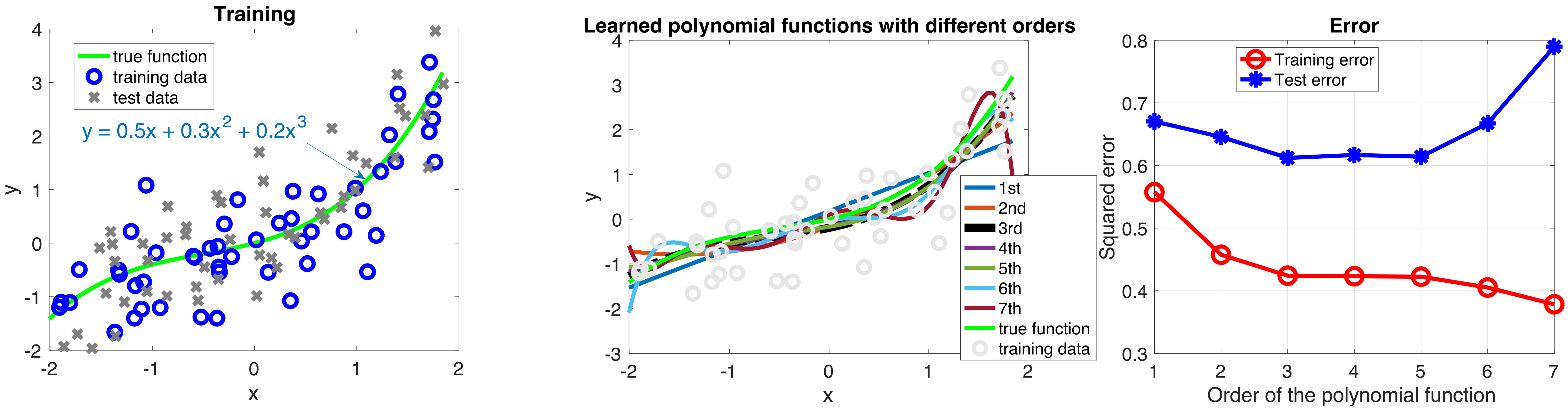

Error

- Machine learning problems (e.g., regression) are typically ill-posed: the observed data is finite and does not uniquely determine the classification or regression function.

- In order to find a unique solution, and learn something useful, we must make assumptions

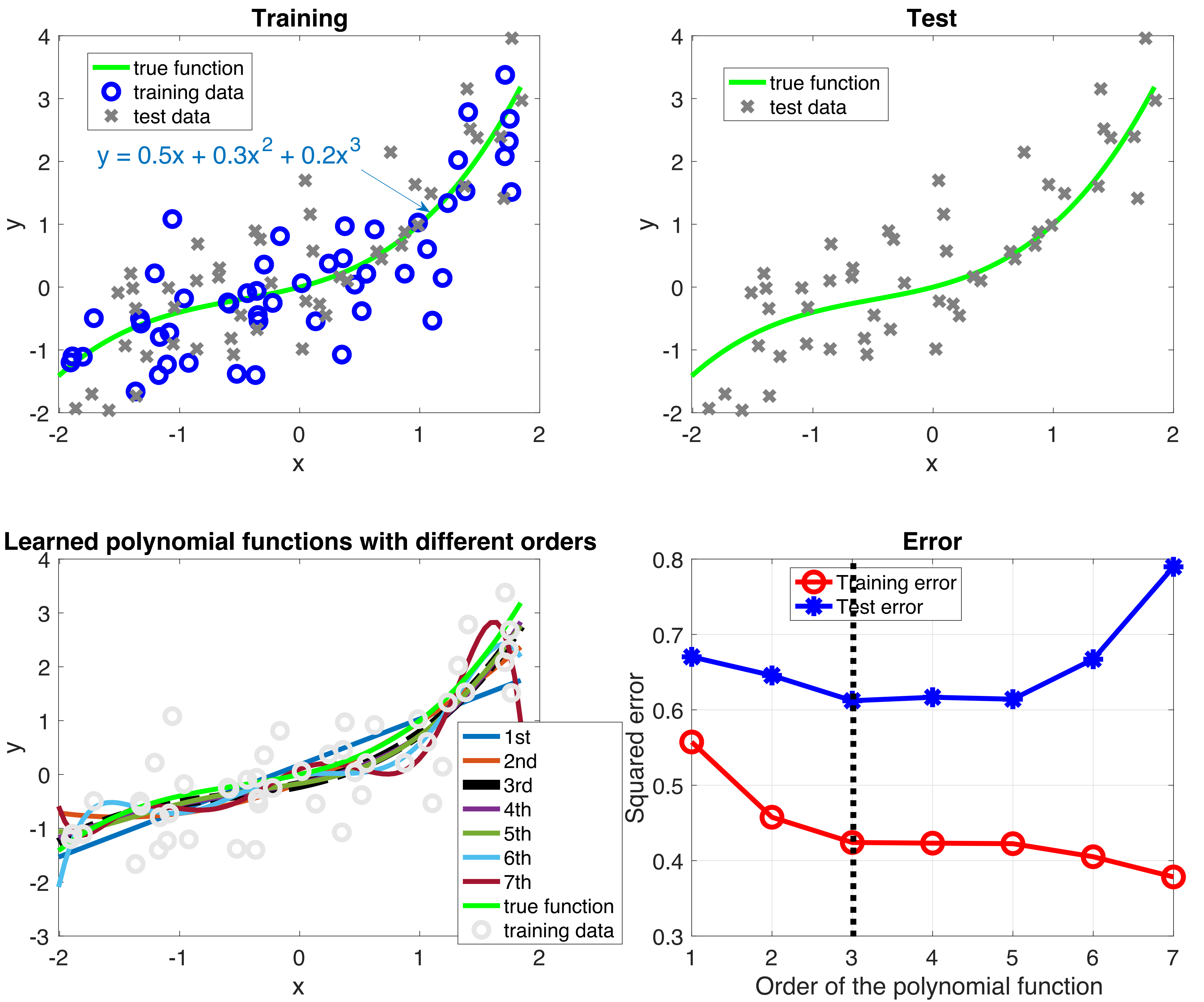
- The goal of ML is not to replicate the training data, but to predict unseen data well, i.e., to generalize well.

Machine Learning Cares about the Performance on New, Unseen Data

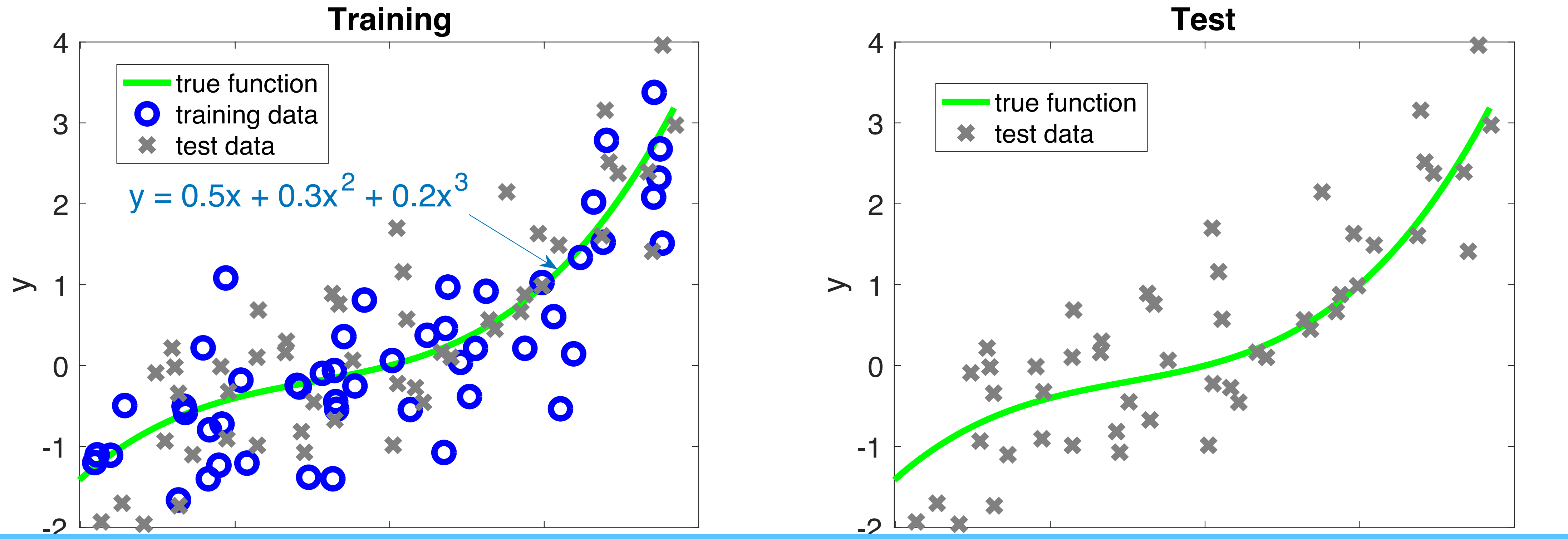


- The goal of ML: predict unseen data well, i.e., to generalize well
- Training error $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}))^2$ vs. test error $E_{(X,Y)\sim P}(Y - f(X; \hat{\boldsymbol{\alpha}}))^2$
- If the class of functions is
  - less complex: underfitting (e.g., fitting a line in the above example)
  - more complex: overfitting (e.g., fitting a 7th order polynomial)

Machine Learning Cares about the Performance on New, Unseen Data



Bias-Variance Tradeoff: A Rough Picture



Assume that data were generated according to $y = f^*(x) + \varepsilon$, where the noise ε has zero mean and variance σ^2.

We aim to find function $\hat{y} = f(x; D)$, where D denote the training dataset, to approximate the true function $f^*(x)$. Its expected error on an unseen sample x is

$$E_D\left[(y - f(x;D))^2\right] = \left(\mathrm{Bias}_D[f(x;D)]\right)^2 + \mathrm{Var}[f(x;D)] + \sigma^2,$$

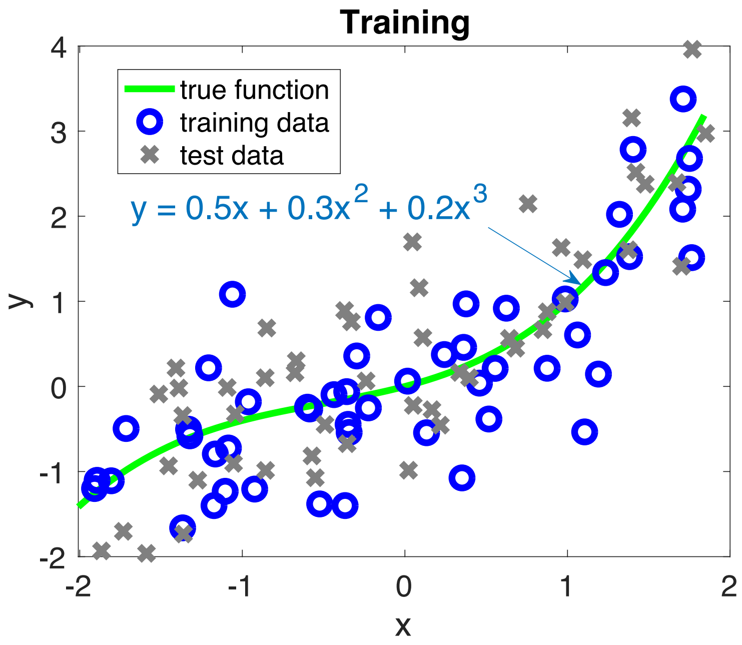
where

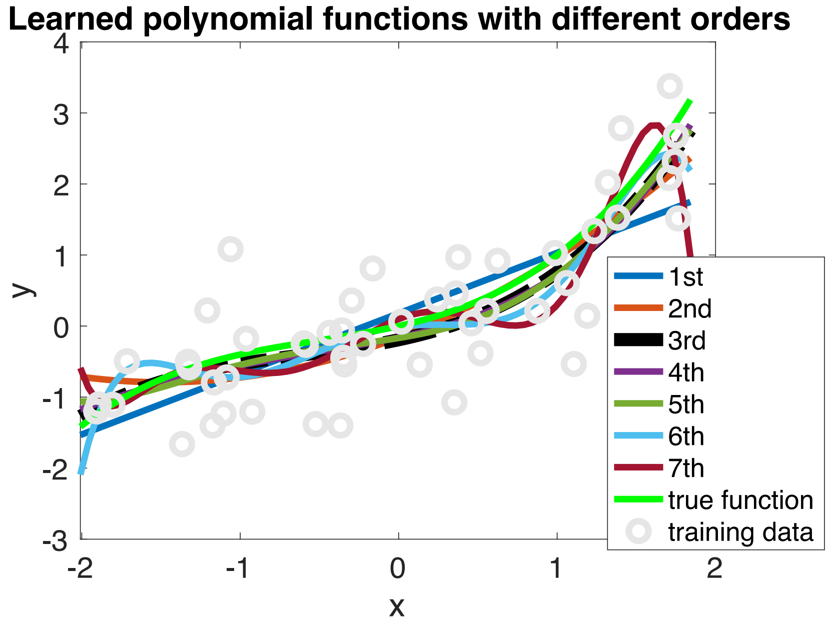
$$\mathrm{Bias}_D[f(x;D)] = E_D[f(x;D)] - f^*(x),$$

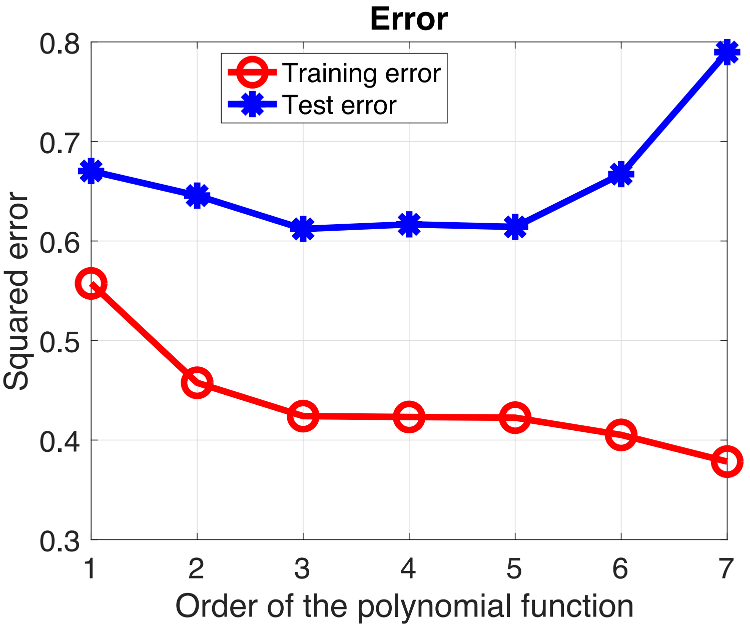
and

$$\mathrm{Var}[f(x;D)] = E_D\left[\left(f(x;D) - E_D[f(x;D)]\right)^2\right].$$

Model Selection

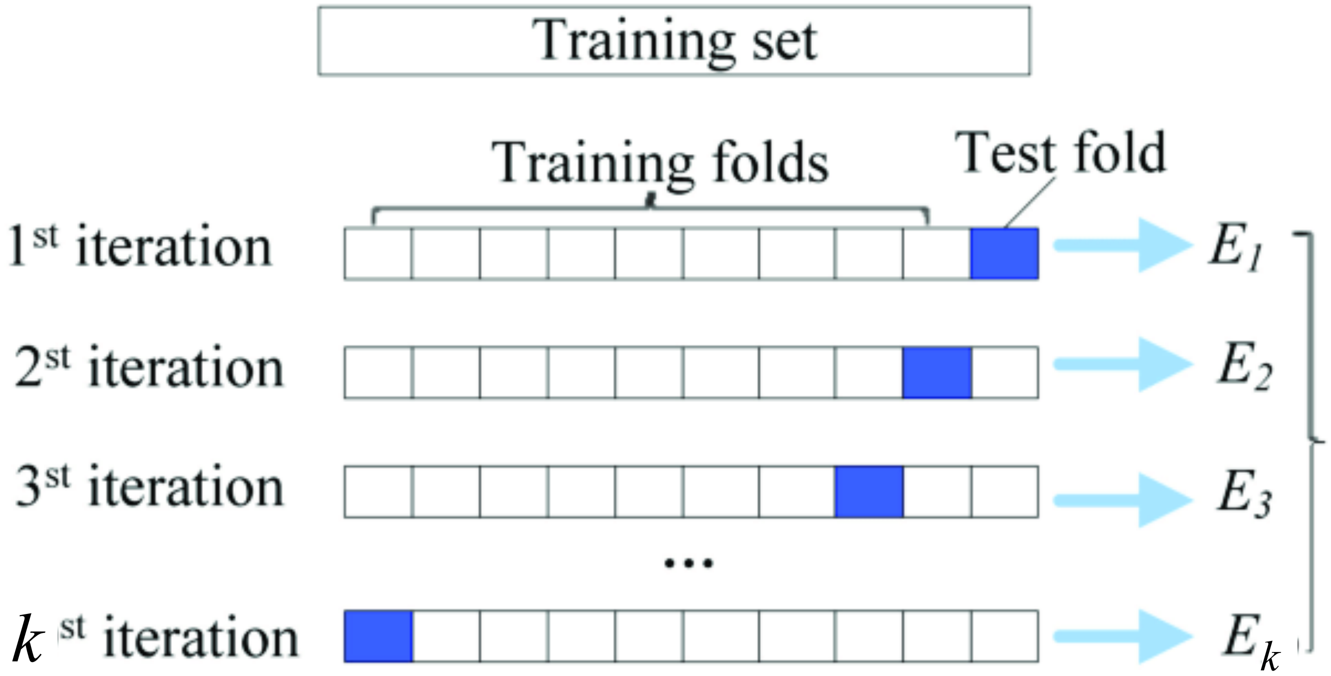






- *Cross validation* allows us to estimate the generalization error based on training examples alone…
- *Information Criterion* takes into account by the training error and the model complexity…

Cross Validation



Training set

Training folds

Test fold

1st iteration

2st iteration

3st iteration

...

k st iteration

E_1

E_2

E_3

E_k

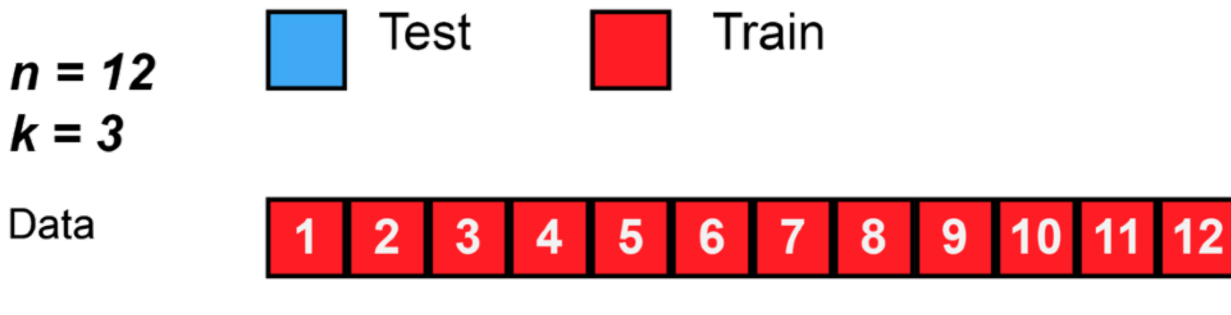
$$E = \frac{1}{k}\sum_{i=1}^{k} E_i$$

(Note: E may be squared error for regression.)

- *Cross validation* allows us to estimate the generalization error based on training examples alone

- *k*-fold cross validation

- Leave-one-out cross-validation treats each training example in turn as a test example (*k*=*n*).

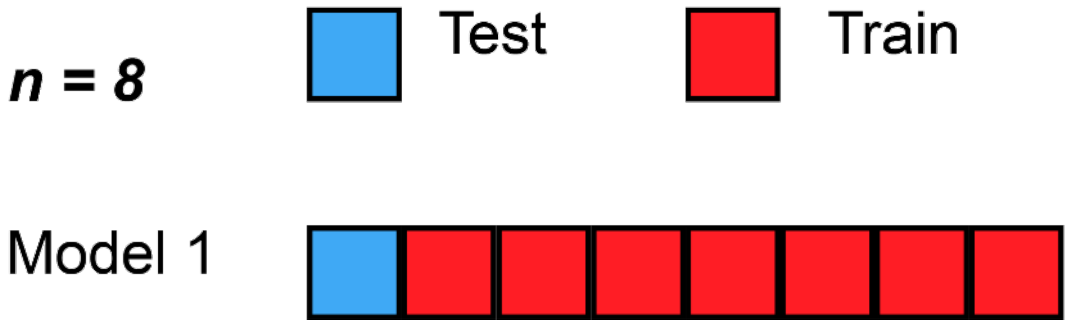


n = 12
k = 3

🟦 Test 🟥 Train

Data

1 2 3 4 5 6 7 8 9 10 11 12

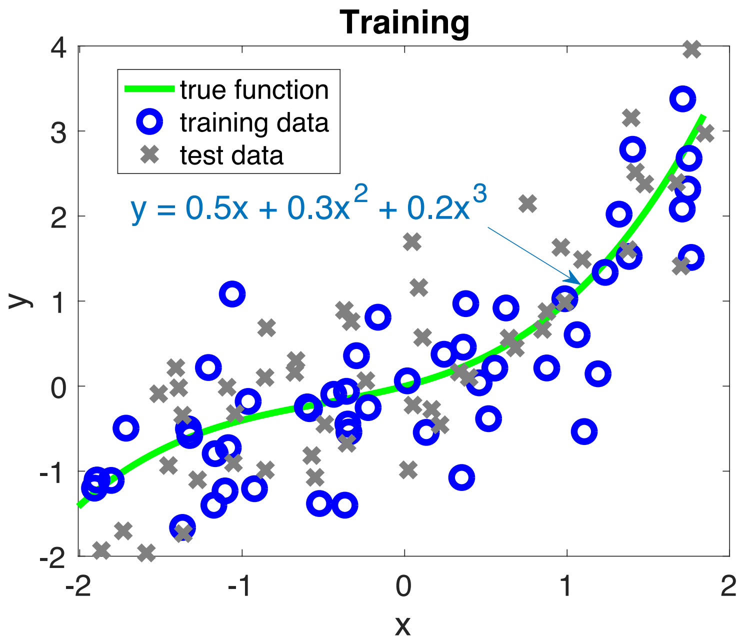


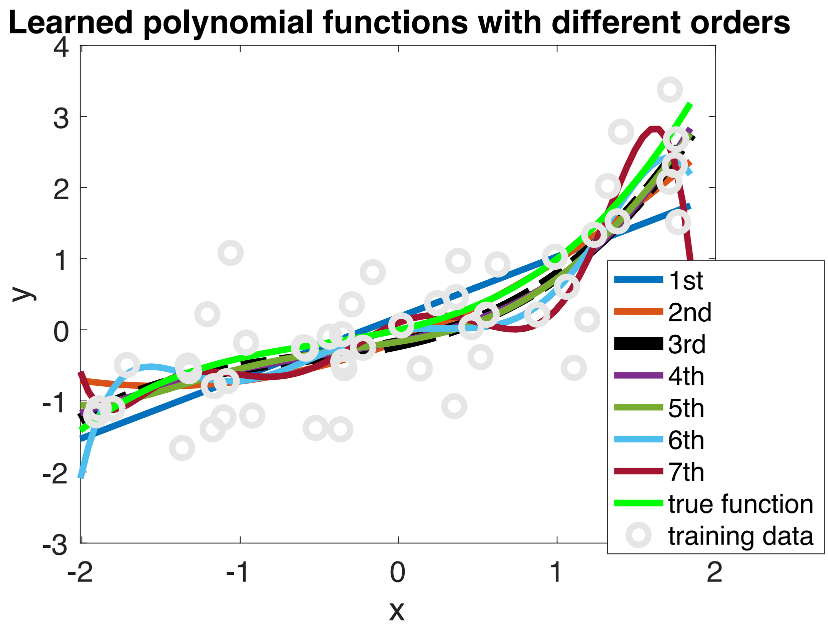
n = 8

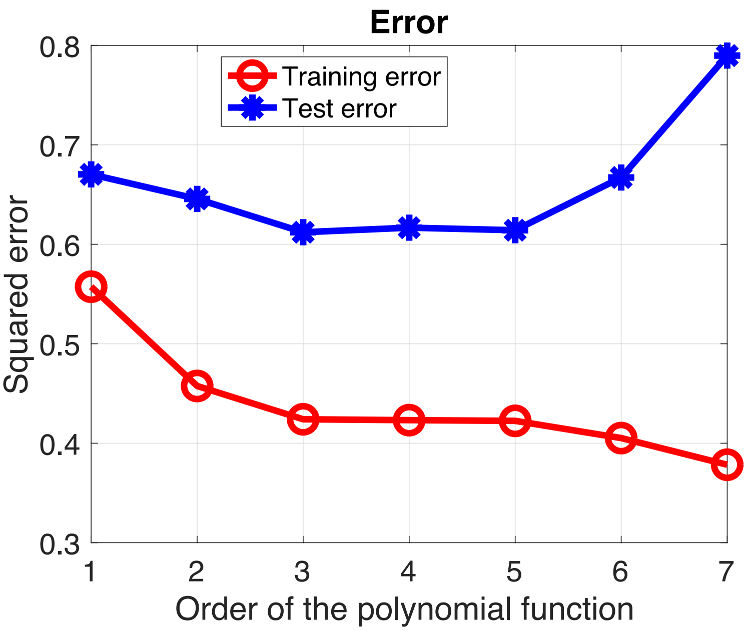
🟦 Test 🟥 Train

Model 1

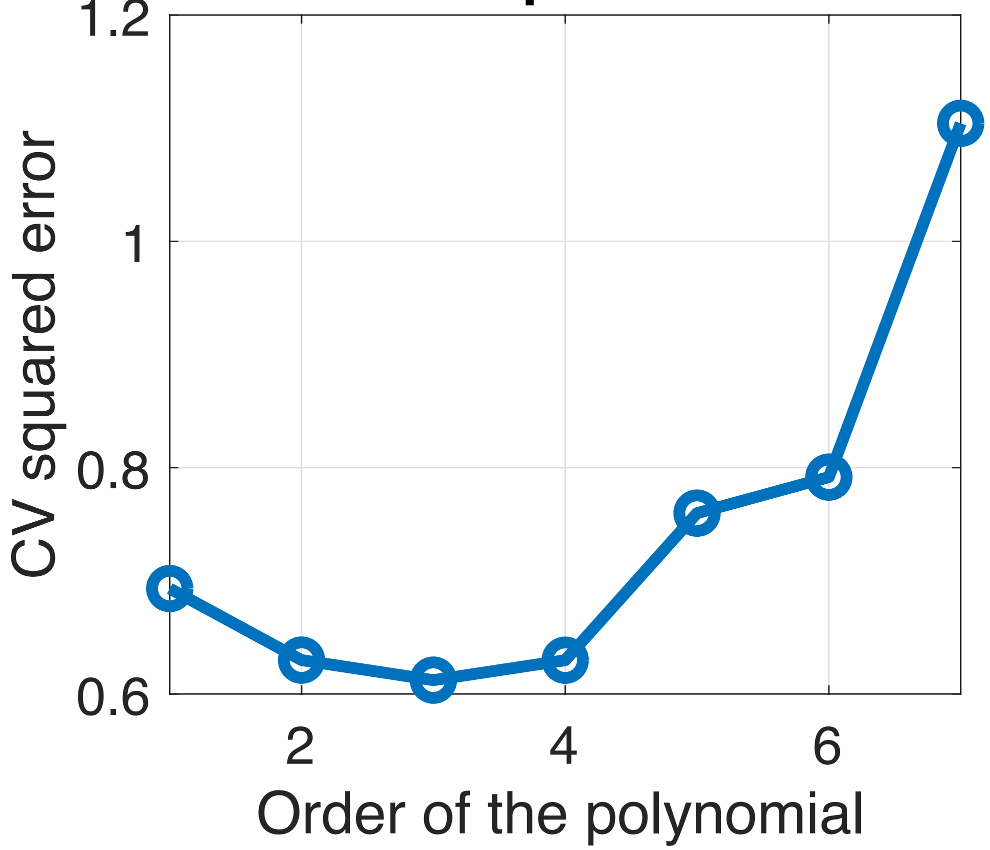
Cross Validation: Illustration



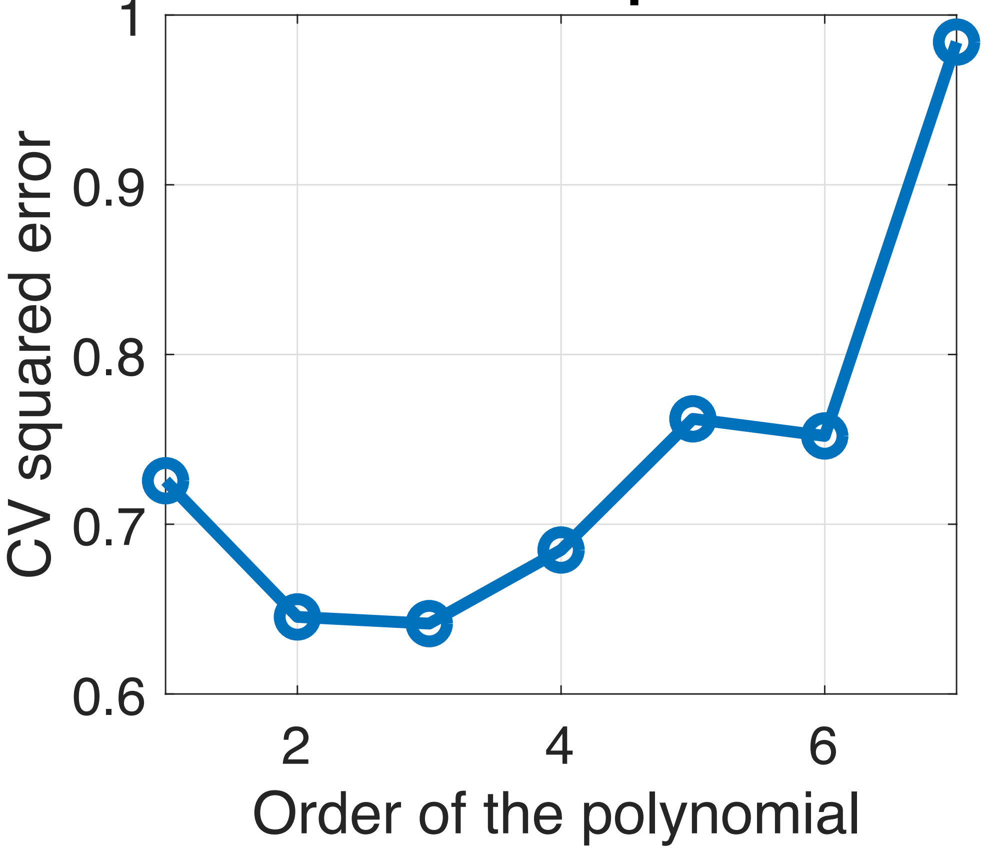




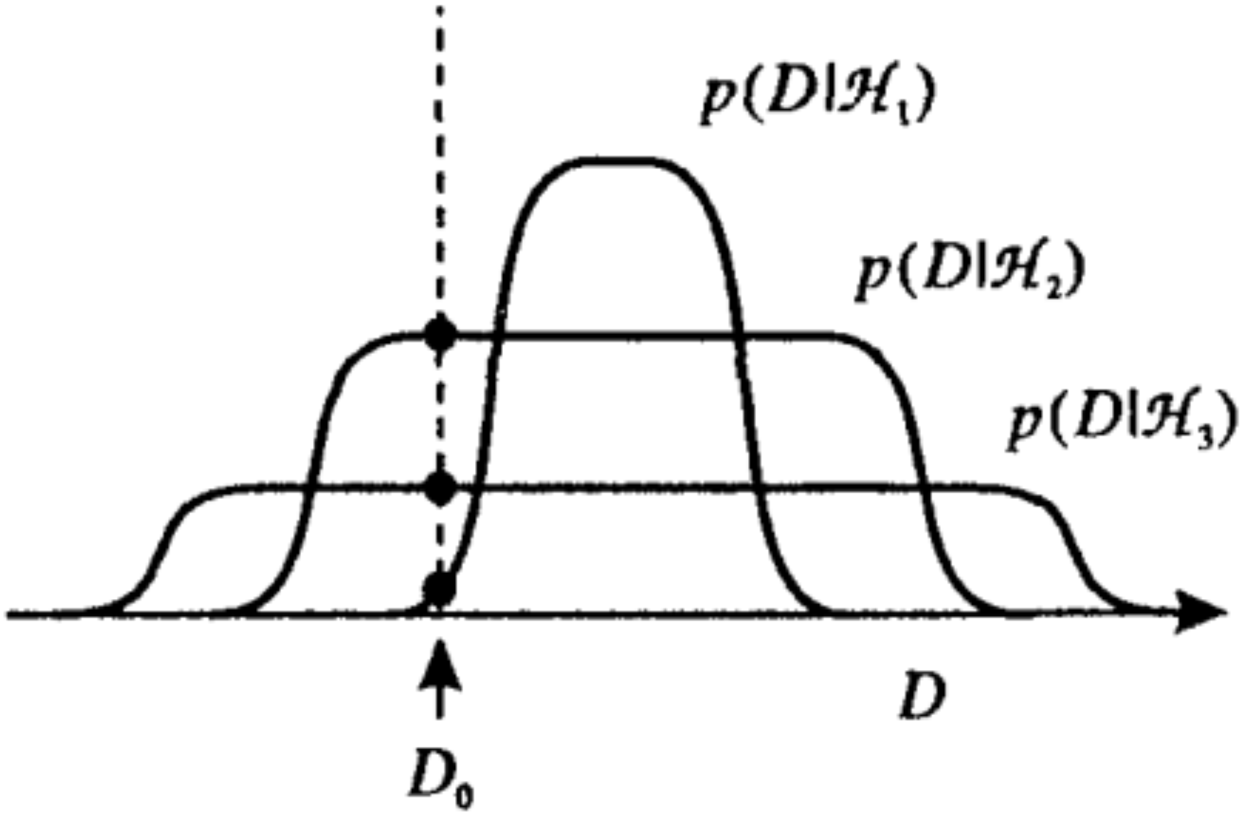
Cross-validated squared error with K=5



Leave-one-out squared error



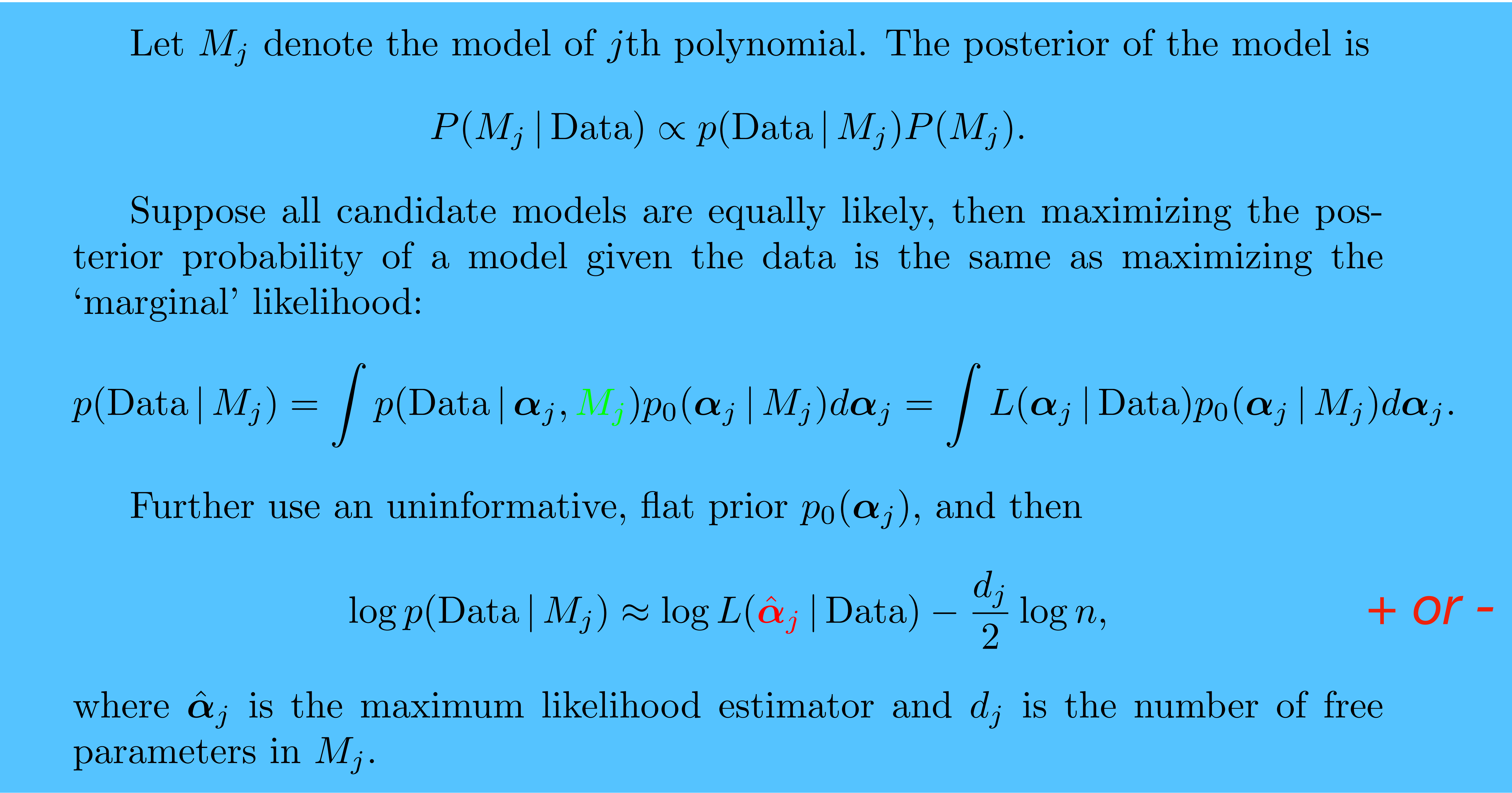
Complex Models Not Necessarily Good



Schematic example of three models, $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$, which have successively greater complexity, showing the probability (known as the *evidence*) of different data sets D given each model $\mathcal{H}_i$. We see that more complex models can describe a greater range of data sets. Note, however, that the distributions are normalized. Thus, when a particular data set D_0 is observed, the model $\mathcal{H}_2$ has a greater evidence than either the simpler model $\mathcal{H}_1$ or the more complex model $\mathcal{H}_3$.

Bayesian Information Criterion (BIC)

- *Cross validation* also known as the Schwarz Criterion after Gideon Schwarz (1978)



Let M_j denote the model of jth polynomial. The posterior of the model is

$$P(M_j \,|\, \text{Data}) \propto p(\text{Data} \,|\, M_j) P(M_j).$$

Suppose all candidate models are equally likely, then maximizing the posterior probability of a model given the data is the same as maximizing the 'marginal' likelihood:

$$p(\text{Data} \,|\, M_j) = \int p(\text{Data} \,|\, \boldsymbol{\alpha}_j, M_j) p_0(\boldsymbol{\alpha}_j \,|\, M_j) d\boldsymbol{\alpha}_j = \int L(\boldsymbol{\alpha}_j \,|\, \text{Data}) p_0(\boldsymbol{\alpha}_j \,|\, M_j) d\boldsymbol{\alpha}_j.$$

Further use an uninformative, flat prior $p_0(\boldsymbol{\alpha}_j)$, and then

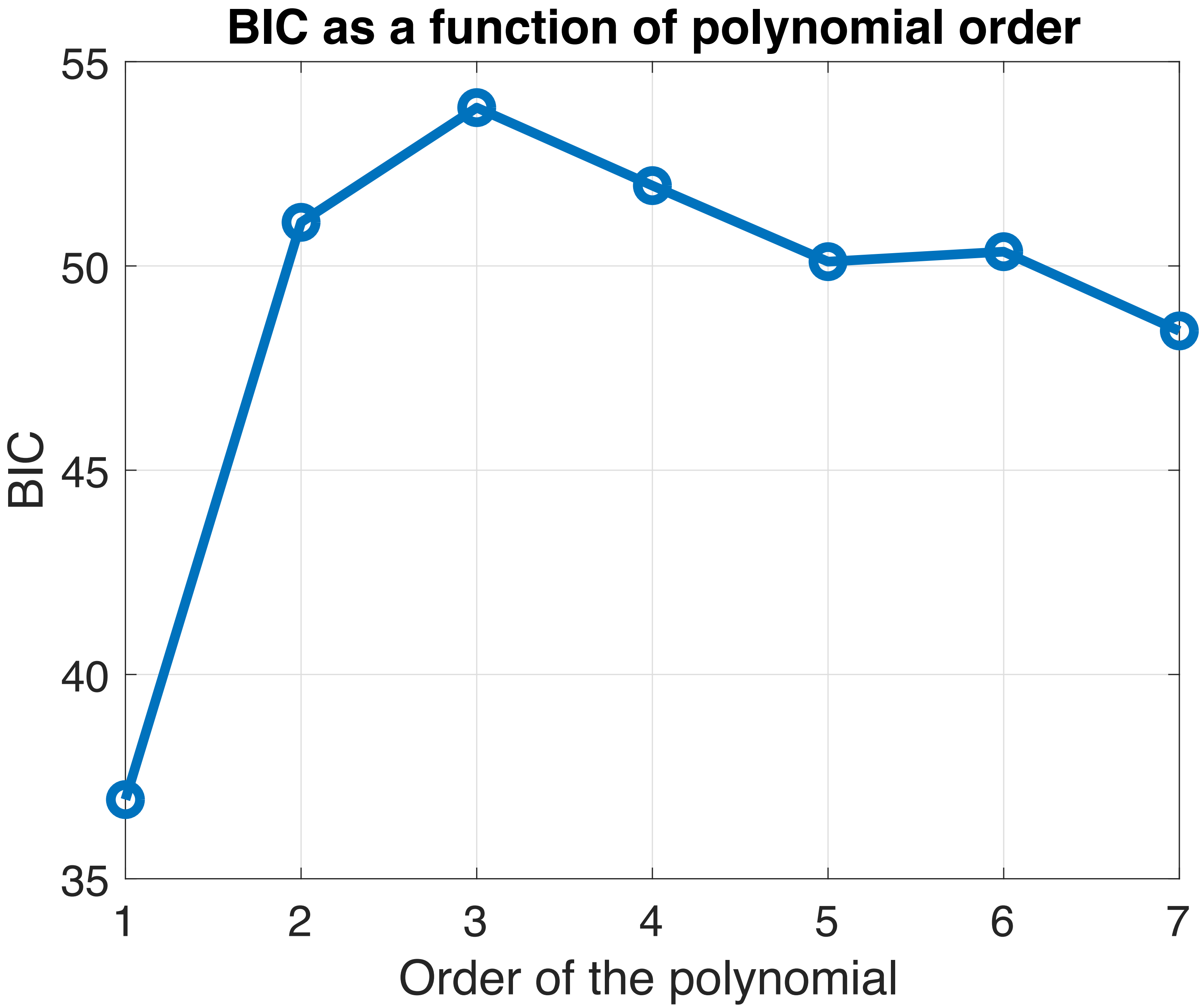
$$\log p(\text{Data} \,|\, M_j) \approx \log L(\hat{\boldsymbol{\alpha}}_j \,|\, \text{Data}) - \frac{d_j}{2} \log n,$$

+ or -

where $\hat{\boldsymbol{\alpha}}_j$ is the maximum likelihood estimator and d_j is the number of free parameters in M_j.

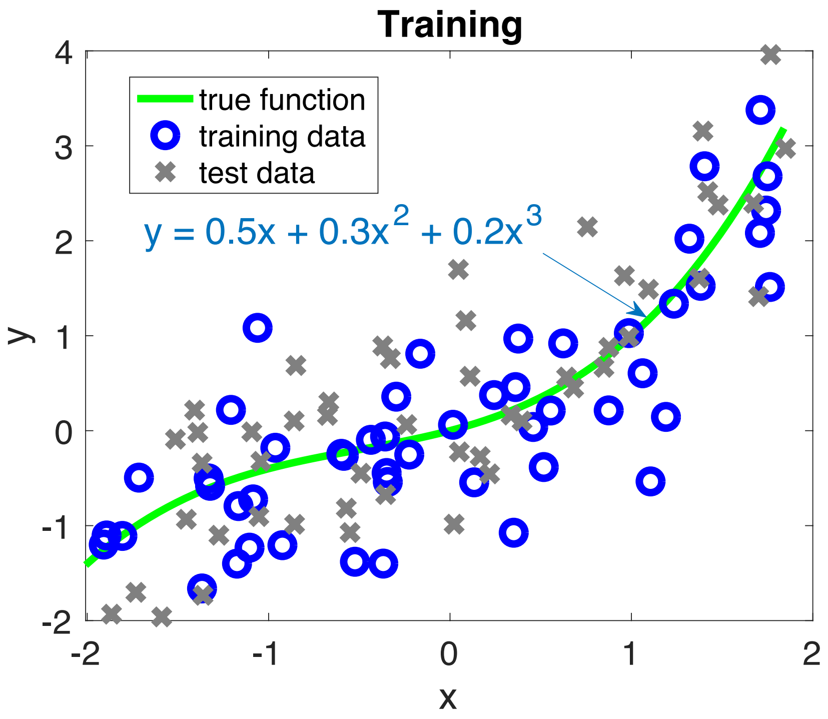
Bayesian Information Criterion (BIC)

$$\log p(\mathrm{Data} \mid M_j) \approx \log L(\hat{\boldsymbol{\alpha}}_j \mid \mathrm{Data}) - \frac{d_j}{2} \log n$$



Summary: Basic Ideas of Model Selection

- Why prefer simple models?
- How simple is simple enough?
- The simplest model and the most probable model
    - Help find causal model?
- Methods for model selection
    - Cross validation
    - Information criteria
    - …



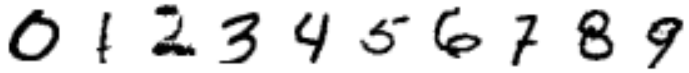
Supervised Learning: Examples

- Regression: the target variable to be predicted is continuous
  - Predict the price of a car from its mileage.
  - Navigating a car: angle of the steering.
- Classification:
  - Face recognition (difficult because of the complex variability in the data: pose and illumination in a face image, occlusions, glasses/beard/make-up/etc.)

Training

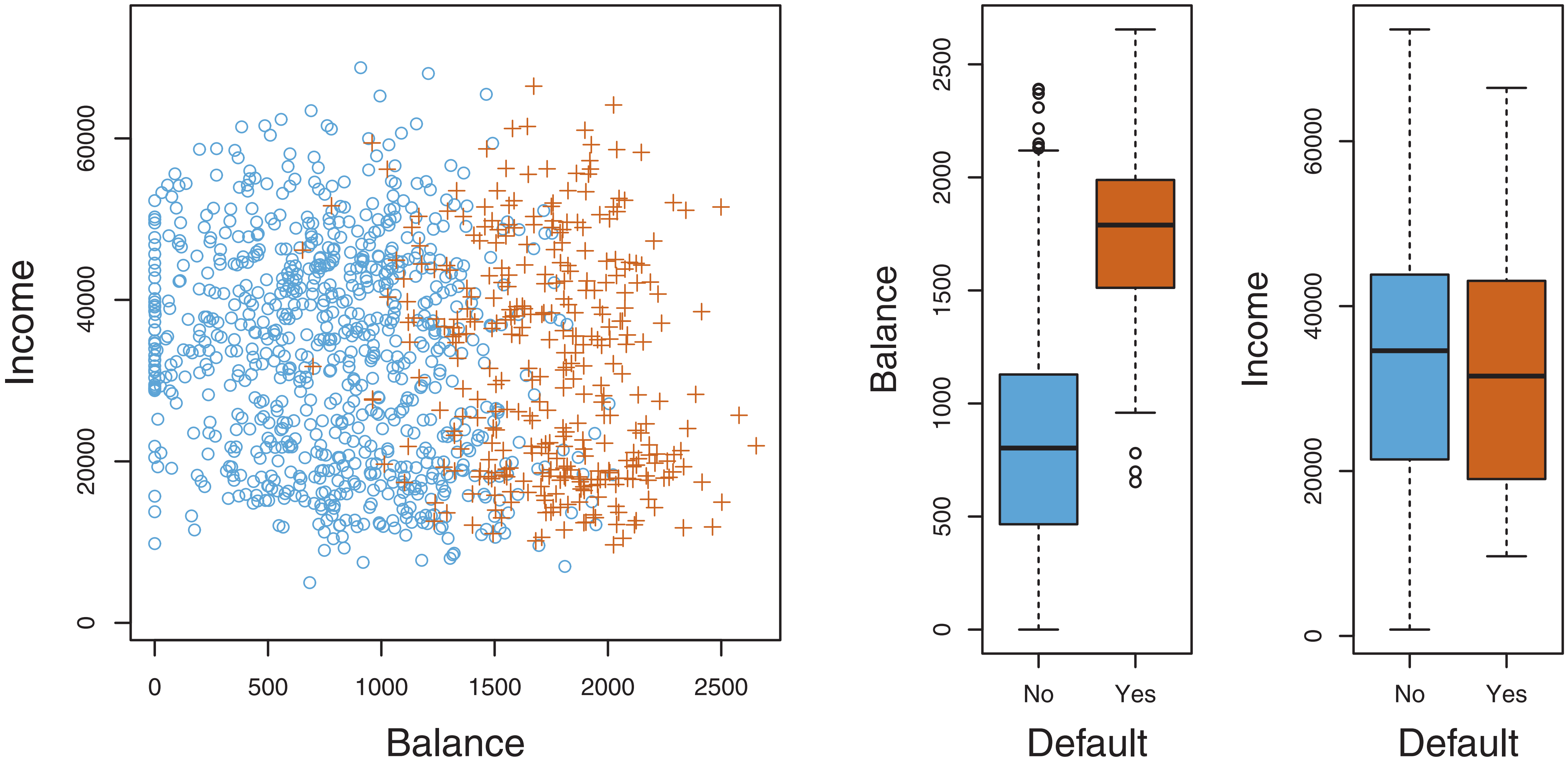
Test

  - Optical character recognition (with different styles…)



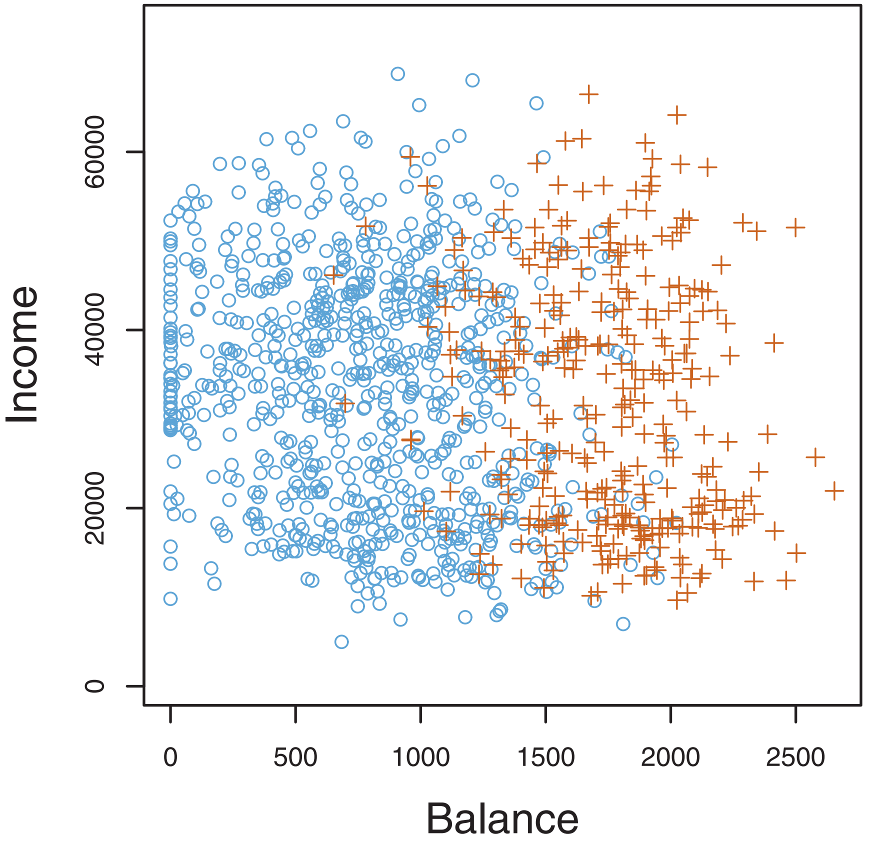
  - Medical diagnosis
  - Credit scoring: classify customers into high- and low-risk, based on their income and savings, using data about past loans (whether they were paid or not)

Classification: Example



The `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

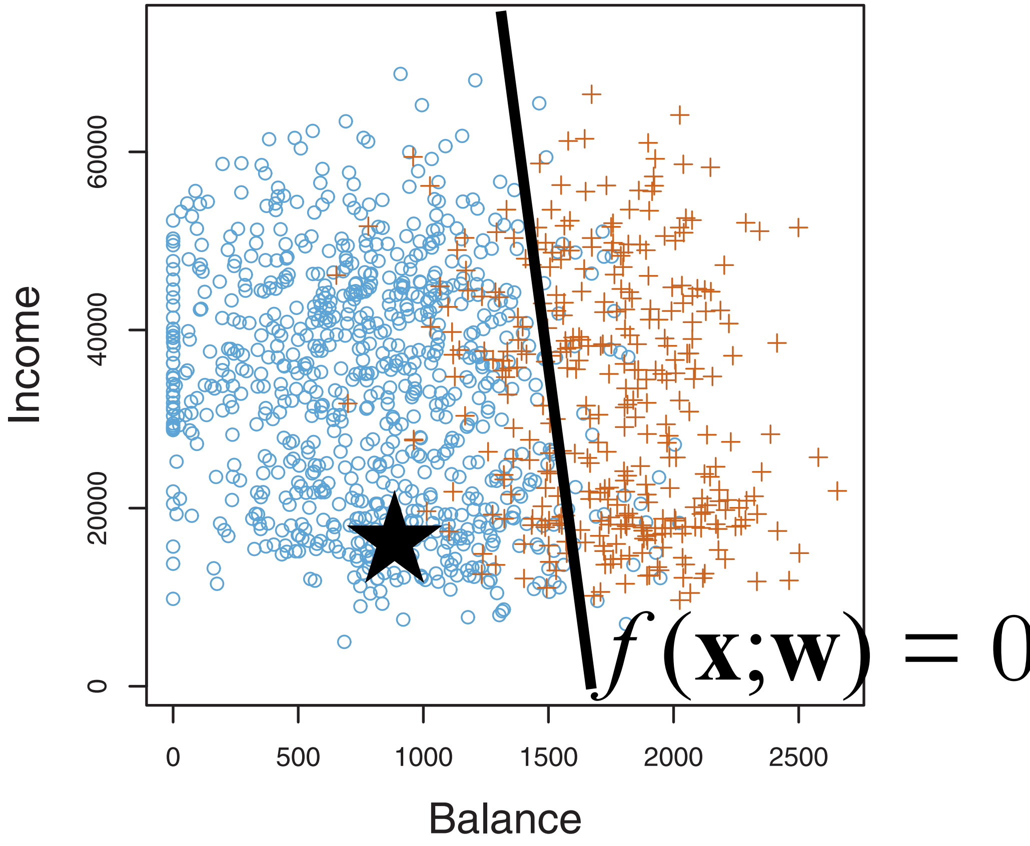
Classification



- We are given a training set of labeled examples (positive and negative)and want to learn a classifier that we can use to predict *unseen* examples, or to understand the data.

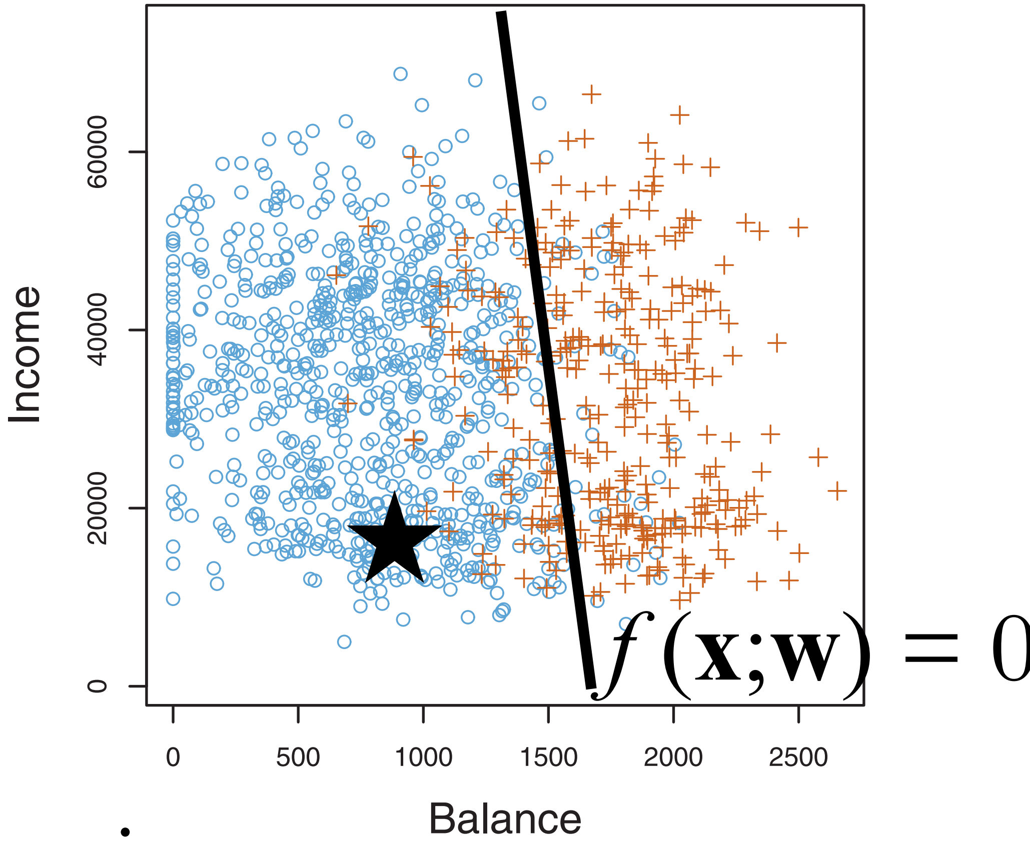
- Training set: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i$ is the *i*th input vector and $y_i \in \{0,1\}$its class label.

Classification



- We are given a training set of labeled examples (positive and negative) and want to learn a classifier that we can use to predict *unseen* examples, or to understand the data.
- Training set: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i$ is the ith input vector and $y_i \in \{0,1\}$its class label.
- How? A natural way is to find a decision boundary! $f(\mathbf{x};\mathbf{w}) = 0$

Classification with Linear Regression?

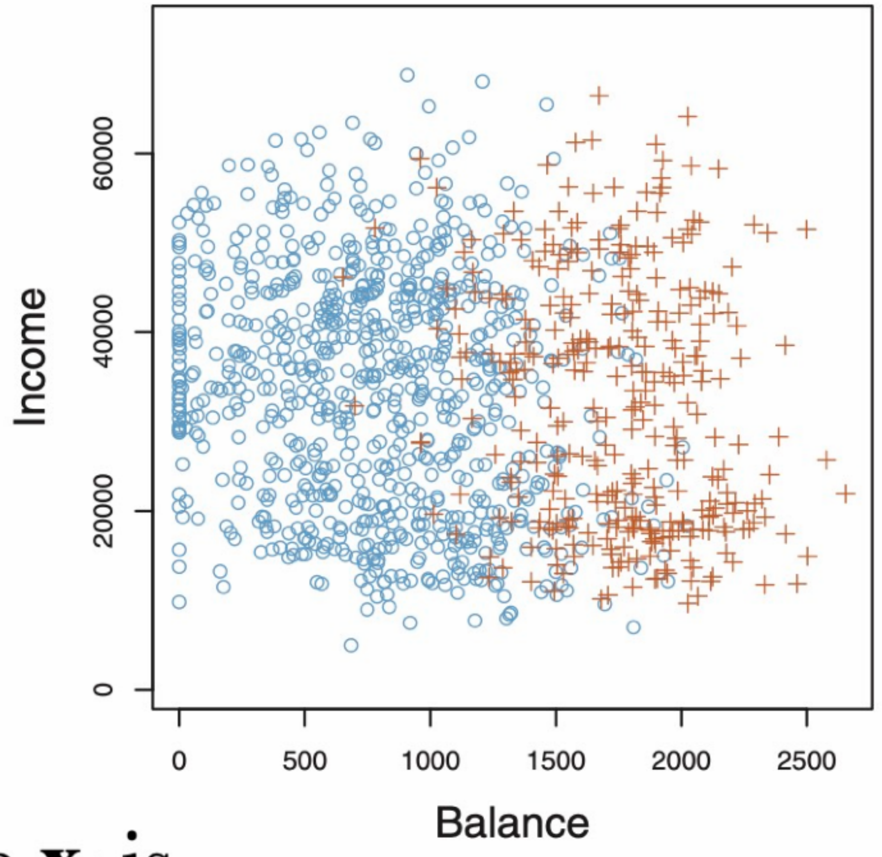


- We have training set: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i$ is the ith input vector and $y_i \in \{0,1\}$its class label

- How about using linear regression? $\hat{y} = l(\mathbf{x};\mathbf{w}) = \mathrm{w}_0 + \mathbf{w}_1^{\mathrm{T}}\mathbf{x}$ for prediction then classify a new (test) example according to
    - label = 1 if $l(\mathbf{x};\mathbf{w})>0.5$, and label = 0 otherwise

- Not a good idea… See why…

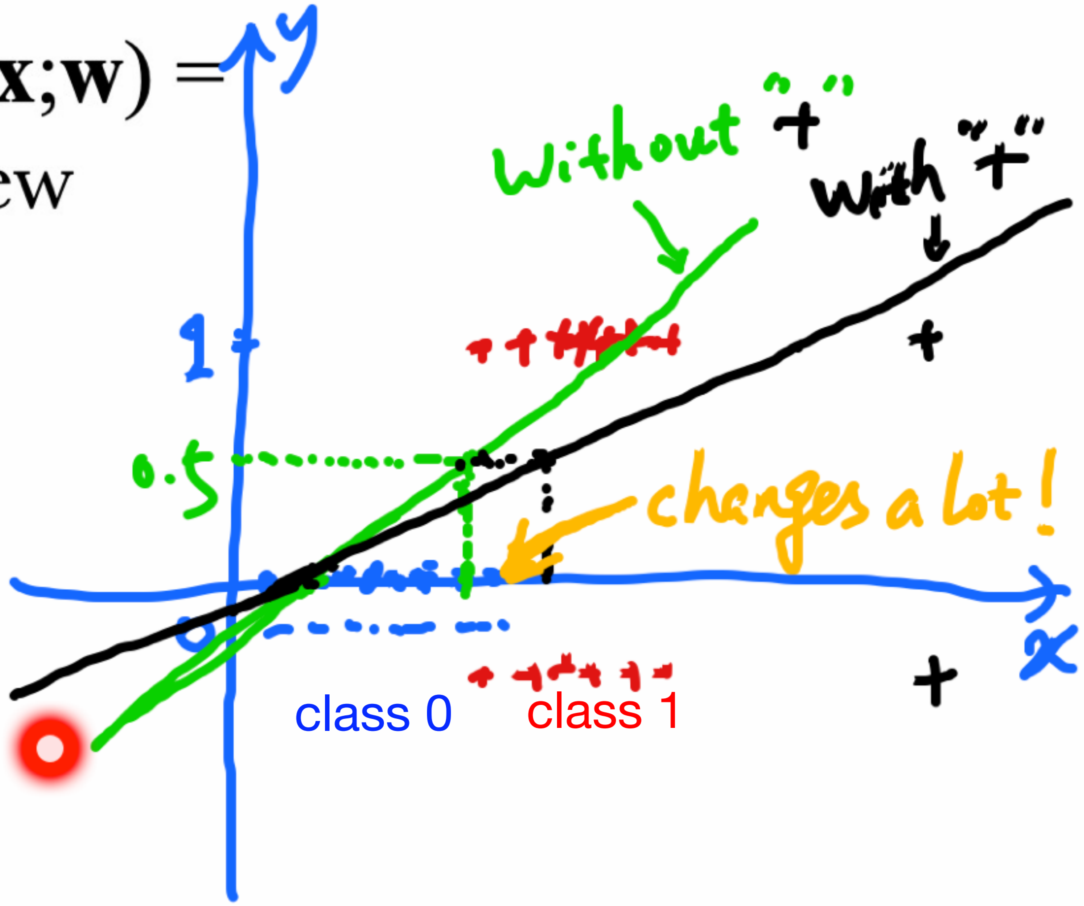
Classification with Linear Regression?



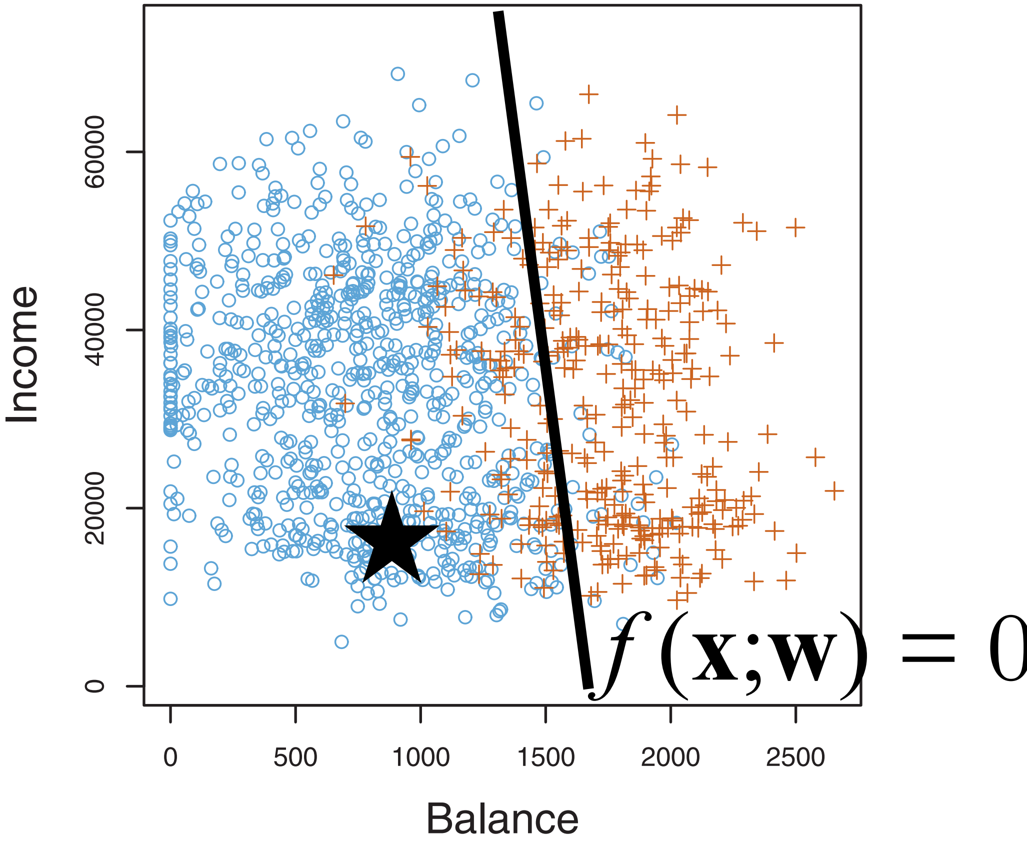
- We have training set: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i$ is the ith input vector and $y_i \in \{0,1\}$its class label

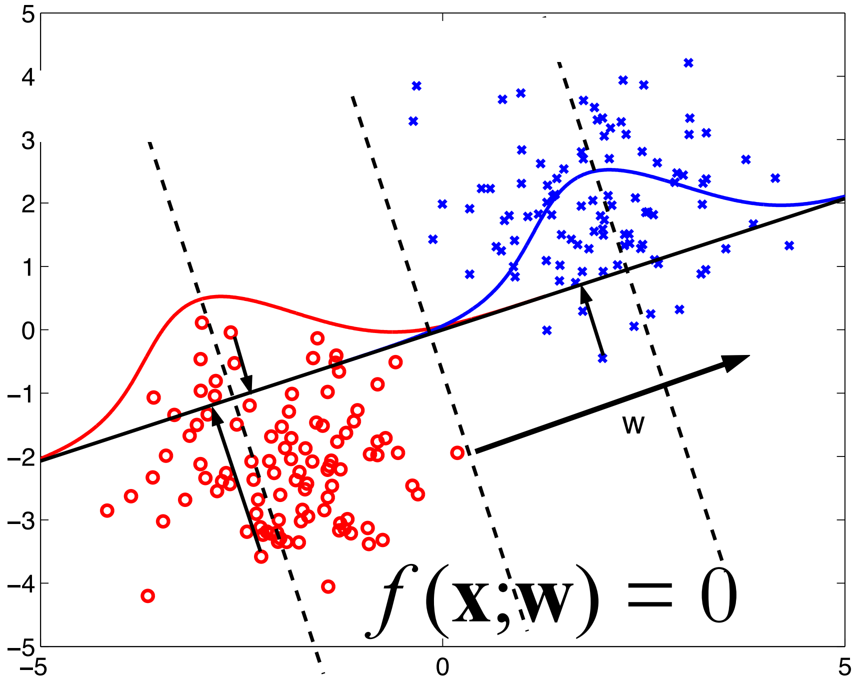
- How about using linear regression? $\hat{y} = l(\mathbf{x};\mathbf{w}) = \mathrm{w}_0 + \mathbf{w}_1^{\mathrm{T}}\mathbf{x}$ for prediction then classify a new (test) example according to
    - label = 1 if $l(\mathbf{x};\mathbf{w})>0.5$, and label = 0 otherwise

- Not a good idea… See why…



Projections of the Data



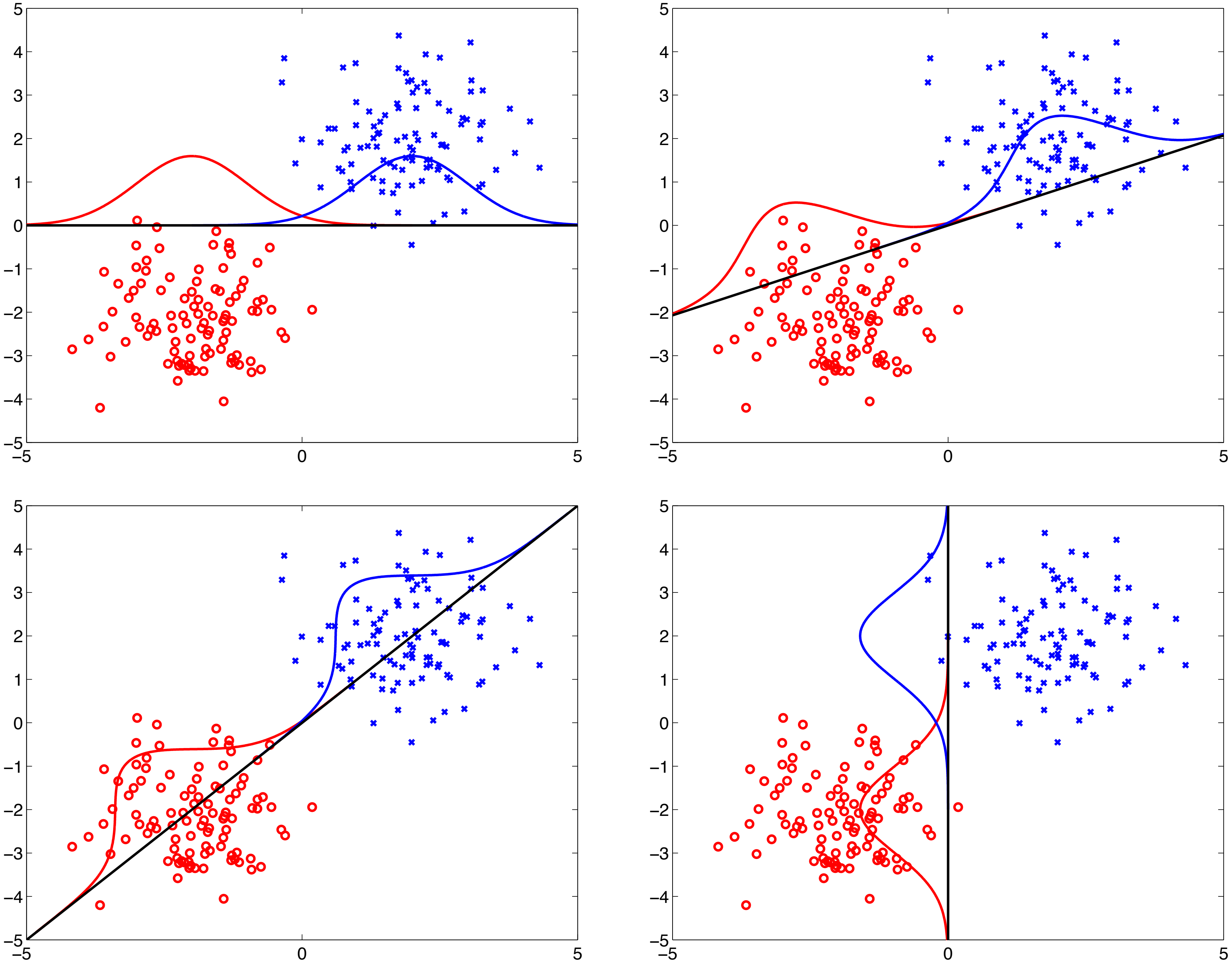


- A linear function $f(\mathbf{x};\mathbf{w}) = w_0 + \mathbf{w}_1^T\mathbf{x}$ projects each point $\mathbf{x}= (x_1, x_2)^T$ to a line parallel to $\mathbf{w}_1$.

- Let's see how well the projected points, determined by $\mathbf{w}_1$, are separated across the classes.

Effect of Different Projects

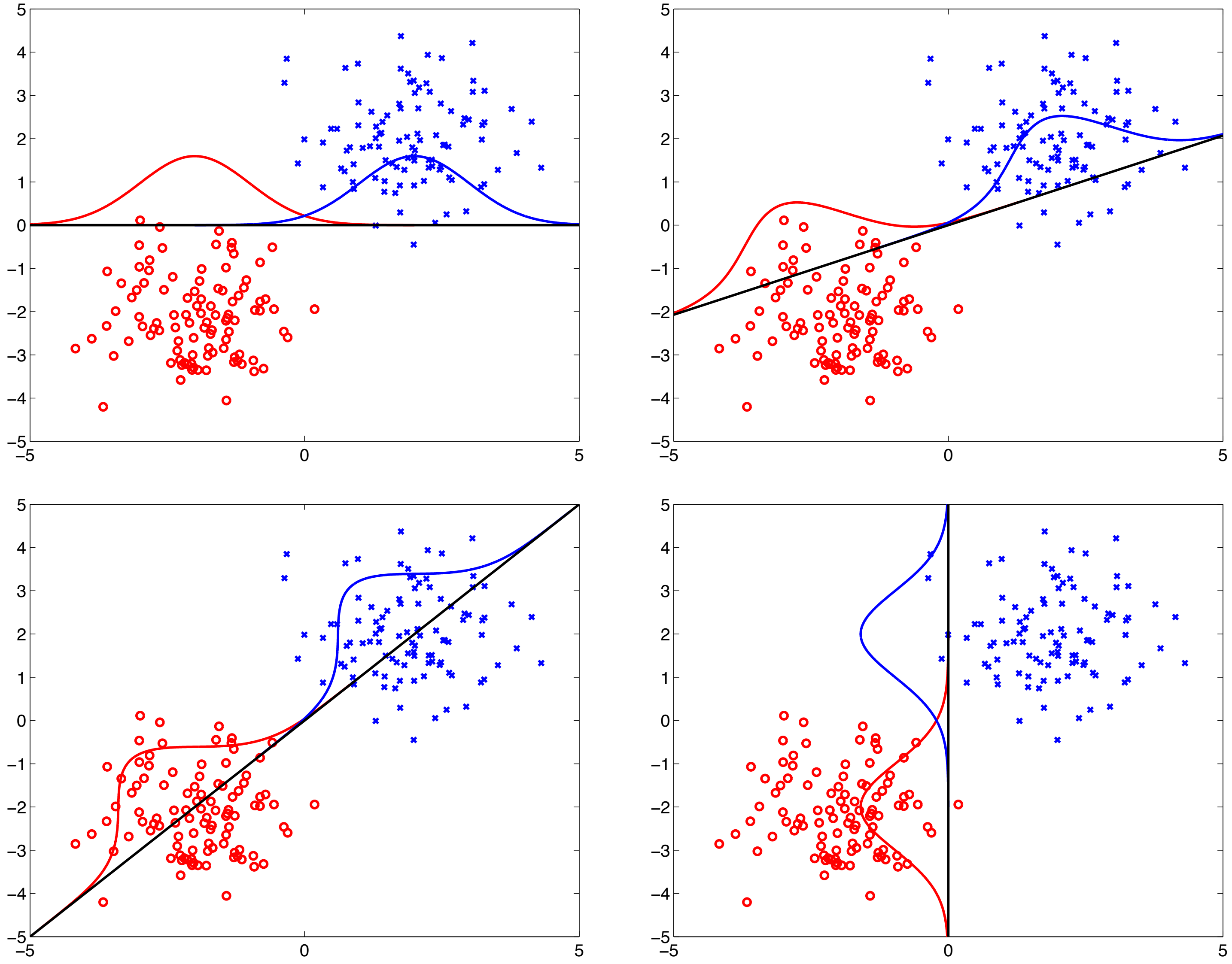
- With different $\mathbf{w}_1$:



(Partly adapted from Tommi Jaakkola's slides)

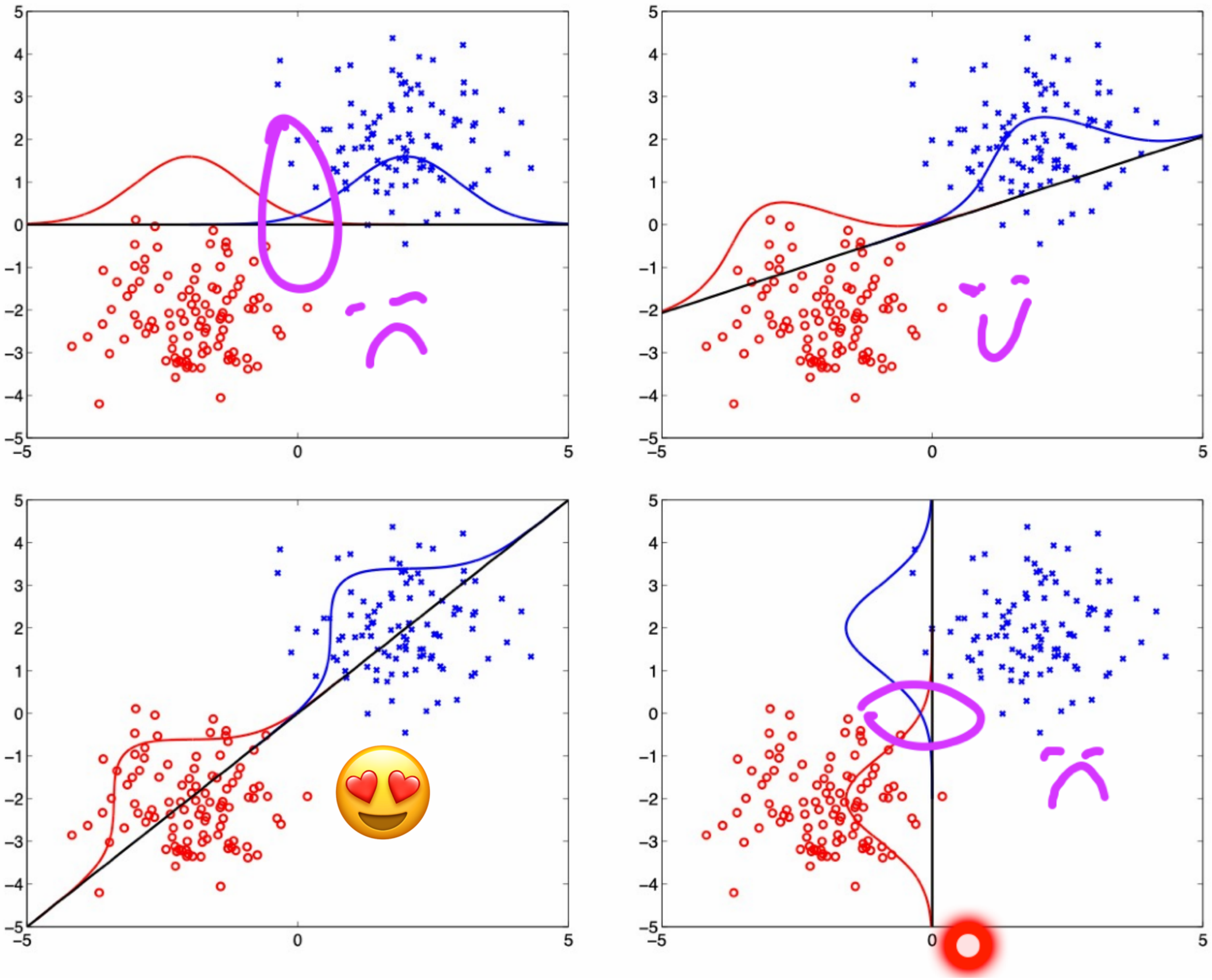
Finding the Optimal Projection

- Find $\mathbf{w}_1$ to maximize the "separation" of the projected points across classes
- Quantify the separation (overlap) in terms of means and variances of the resulting 1-dimensional class distributions

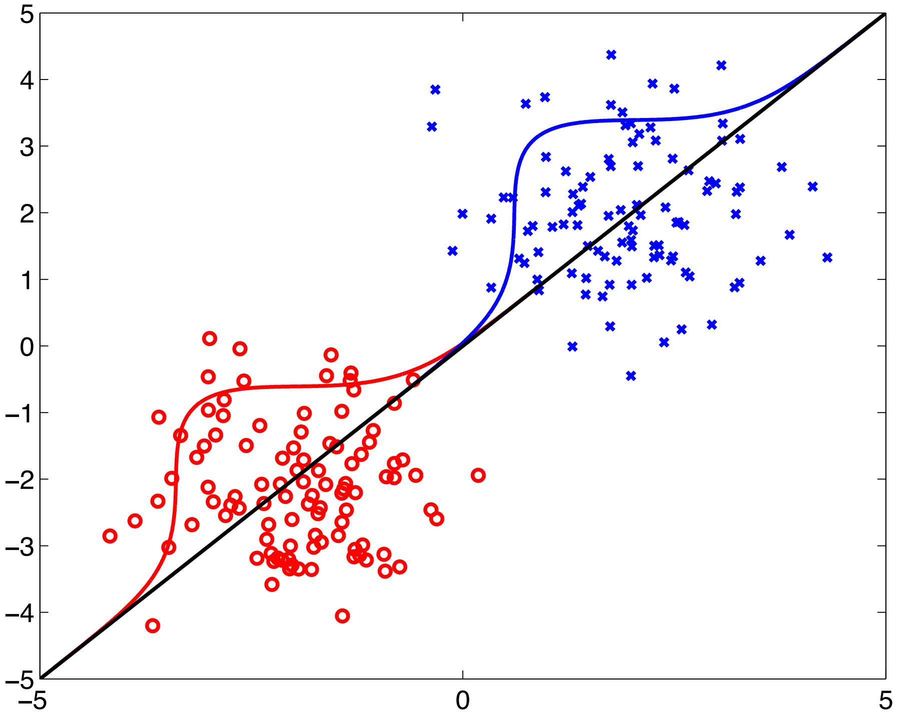


Finding the Optimal Projection

- Find $\mathbf{w}_1$ to maximize the “separation” of the projected points across classes
- Quantify the separation (overlap) in terms of means and variances of the resulting 1-dimensional class distributions

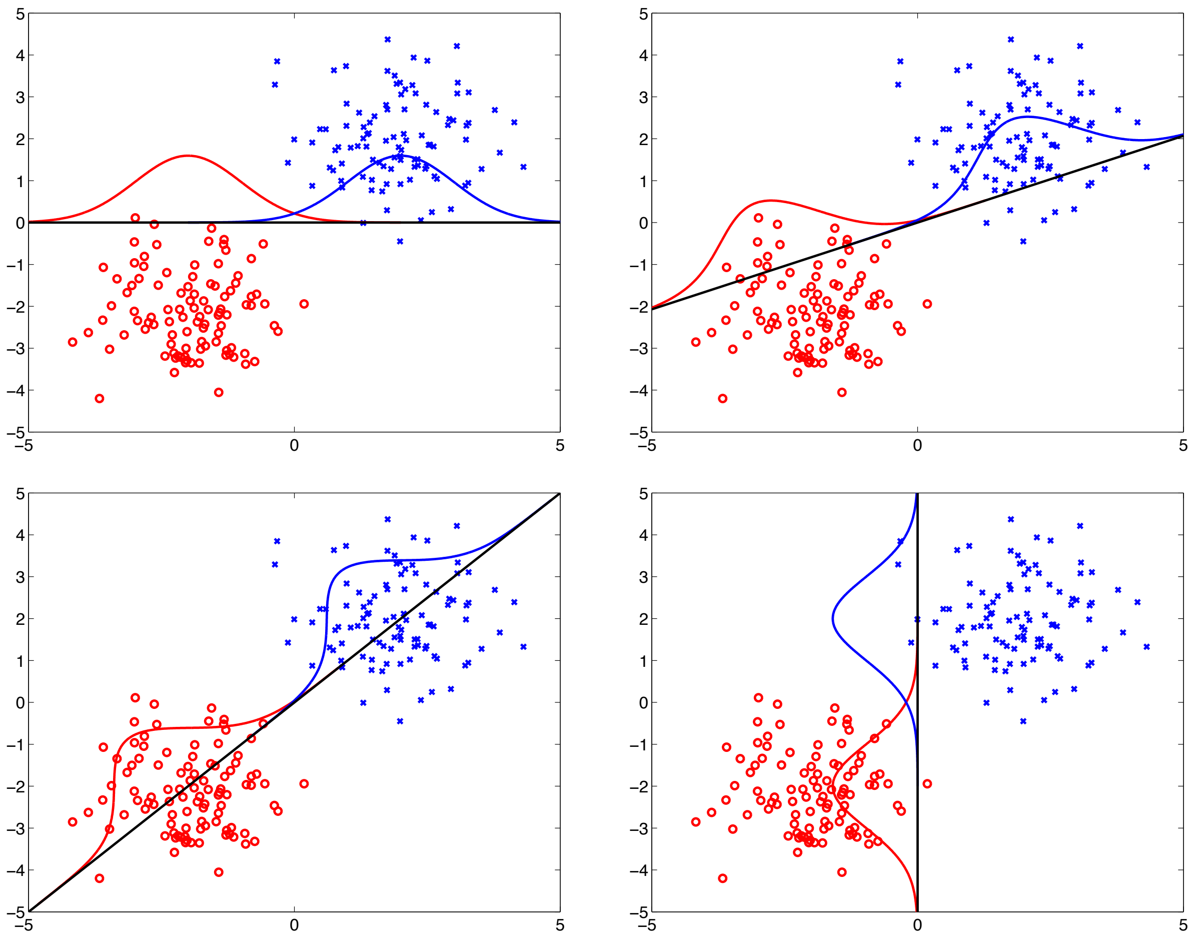


Before & After the Projection



- For original **x** (*d*-dimensional):
  - class 0: n_0 samples, mean μ_0, covariance Σ_0
  - class 1: n_1 samples, mean μ_1, covariance Σ_1
- Projected class descriptions (1-dimension):
  - class 0: n_0 samples, mean $\mathbf{w}_1^T\mu_0$, variance $\mathbf{w}_1^T\Sigma_0\mathbf{w}_1$
  - class 1: n_1 samples, mean $\mathbf{w}_1^T\mu_1$, variance $\mathbf{w}_1^T\Sigma_1\mathbf{w}_1$

Fisher Linear Discriminant



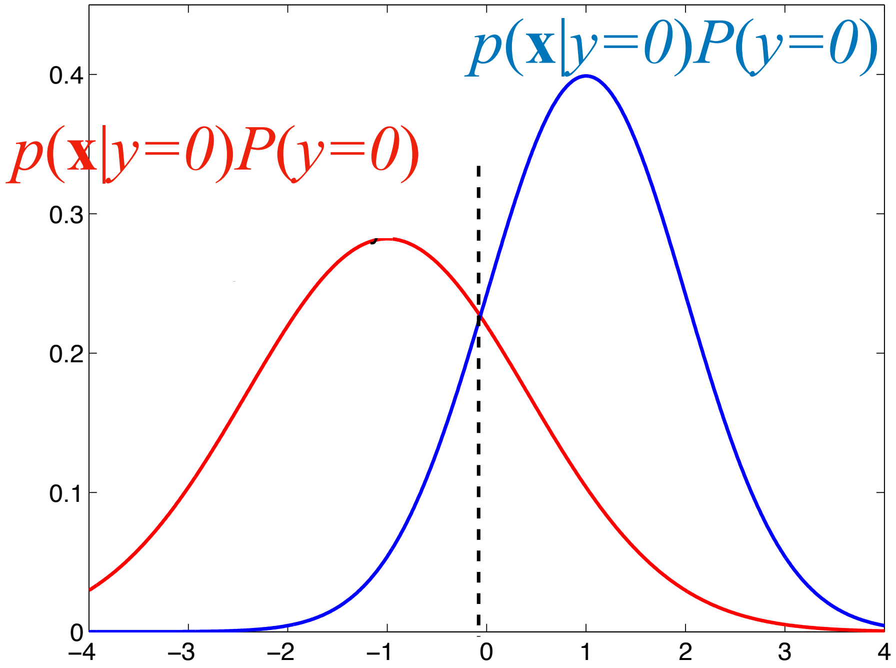
- Objective: Finding projection $\mathbf{w}_1$ to maximize

$$J_{Fisher}(\mathbf{w}) = \frac{(\text{Separation of projected means})^2}{\text{Sum of within class variances}} = \frac{(\mu_1^T \mathbf{w}_1 - \mu_0^T \mathbf{w}_1)^2}{n_1 \mathbf{w}_1^T \Sigma_1 \mathbf{w}_1 + n_0 \mathbf{w}_1^T \Sigma_0 \mathbf{w}_1}$$

- The solution is $\mathbf{w}_1 \propto (n_1\Sigma_1 + n_0\Sigma_0)^{-1}(\mu_1 - \mu_0)$
  - Theoretically optimal for two normal populations with equal covariances $\Sigma_0 = \Sigma_1$

Simple Decision Theory

- Suppose we know the class-conditional densities $p(\mathbf{x}|y)$ for y = 0, 1 as well as the overall class frequencies $P(y)$.
- How do we decide which class a new example $\mathbf{x}^{\text{new}}$ belongs to in order to minimize the overall probability of error?



The minimum probability of error decisions are given by

$$
\begin{aligned}
y^{new} &= \arg\max_{y=0,1} \; P(y|\mathbf{x}^{new}) \\
&= \arg\max_{y=0,1} \; \{p(\mathbf{x}^{new}|y)P(y)\}
\end{aligned}
$$

Known as Bayes classifier.

Logistic Regression

- For binary classification problems, we can write these decisions as

$$y = 1 \text{ if } \log \tfrac{P(y=1\,|\,\mathbf{x})}{P(y=0\,|\,\mathbf{x})} > 0, \text{ and } y = 0 \text{ otherwise.}$$

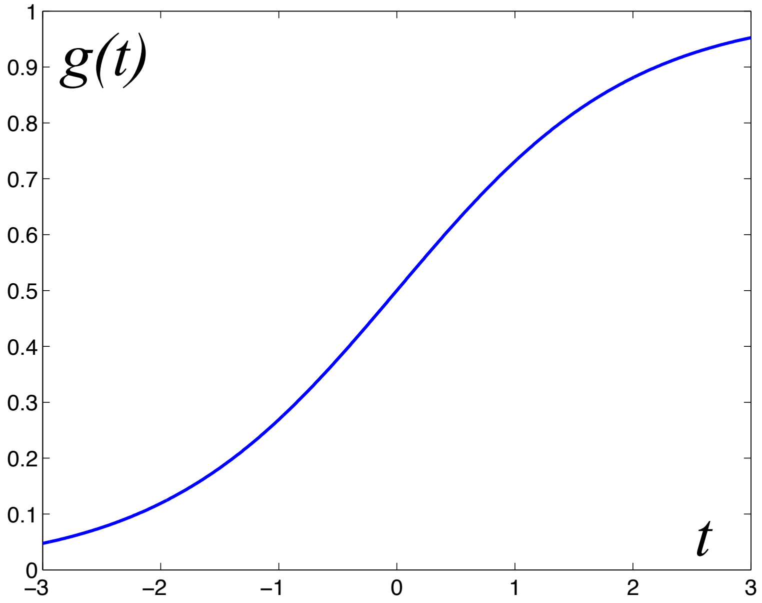
- We generally don't know $P(y|x)$, but we can parametrize the log-odds with a liner function:

$$\log \frac{P(y=1\,|\,\mathbf{x})}{P(y=0\,|\,\mathbf{x})} = f(\mathbf{x};\mathbf{w}) = w_0 + \mathbf{w}_1^\intercal \mathbf{x}.$$

- It gives rise to the logistic model

$$P(y=1\,|\,\mathbf{x}) = g(w_0 + \mathbf{w}_1^\intercal \mathbf{x}),$$

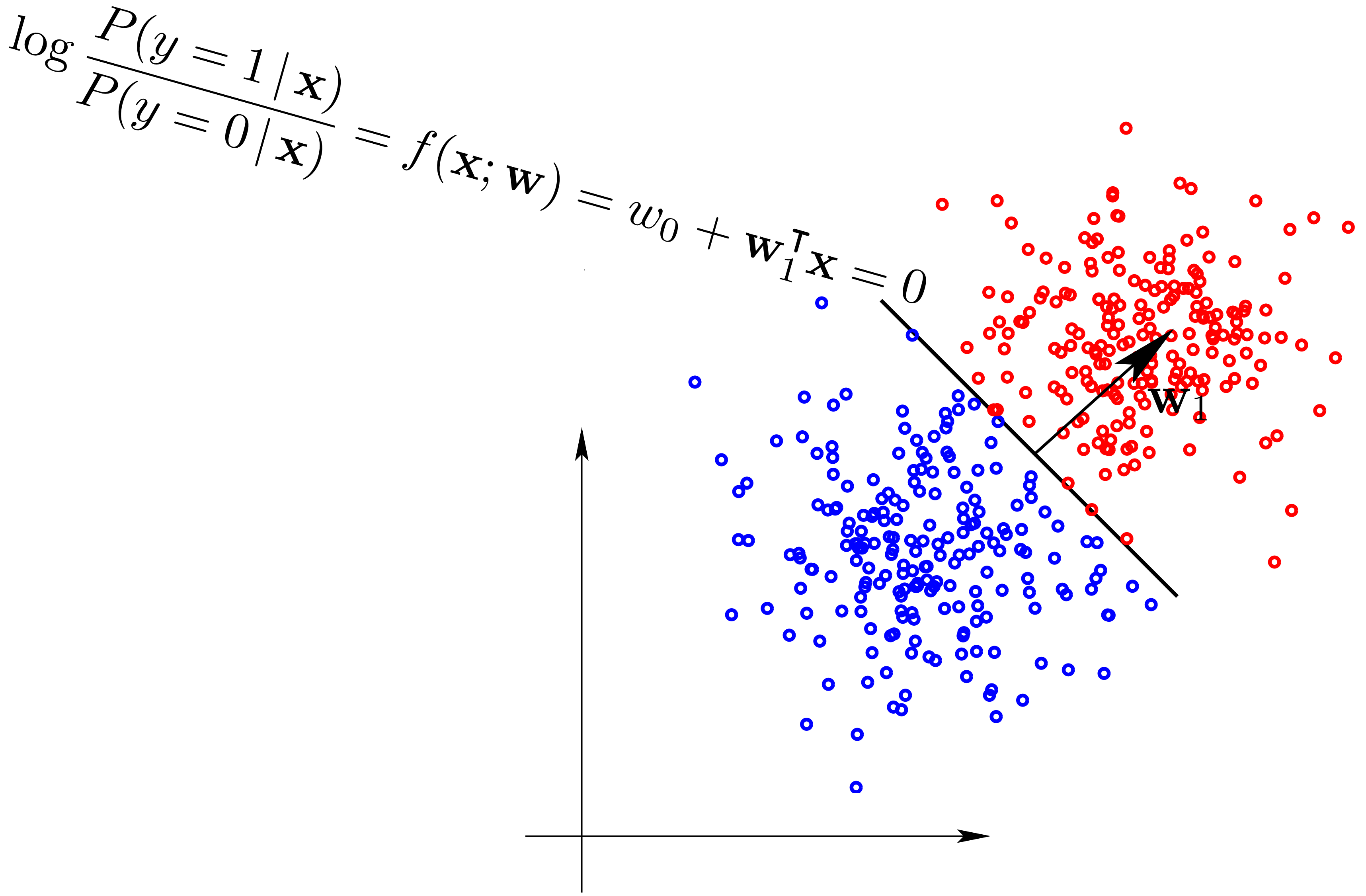
where $g(t) = \frac{1}{1+e^{-t}}$ is a logistic "squashing function" that turns linear functions into probabilities.



Logistic Regression for Classification: Illustration

- For binary classification problems, we can write these decisions as

$$y = 1 \text{ if } \log \frac{P(y{=}1 \,|\, \mathbf{x})}{P(y{=}0 \,|\, \mathbf{x})} > 0, \text{ and } y = 0 \text{ otherwise.}$$

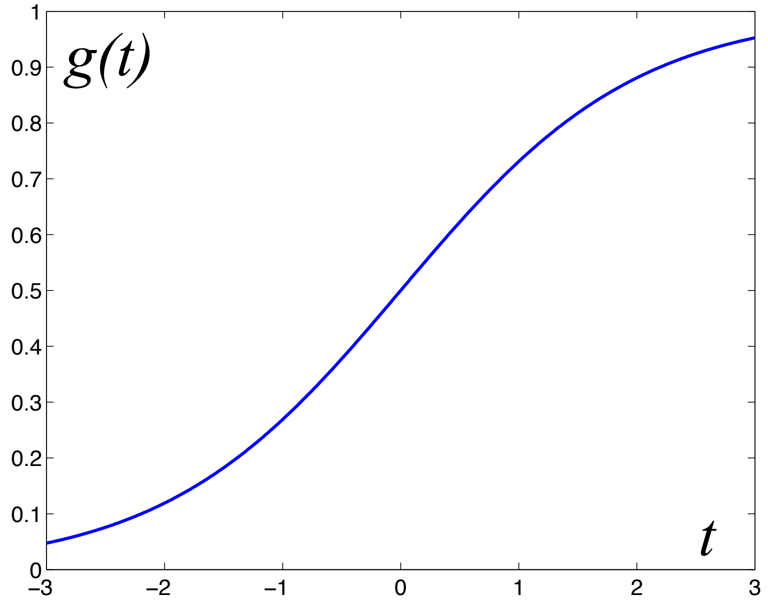


Parameter Estimation for Logistic Regression

- As with regression models, we can fit the logistic models using maximum (conditional) log-likelihood:

$$\log L(\mathbf{w};\mathrm{Data}) = \sum_{i=1}^{n} \log P(y_i \,|\, \mathbf{x}_i, \mathbf{w}), \quad \text{where}$$

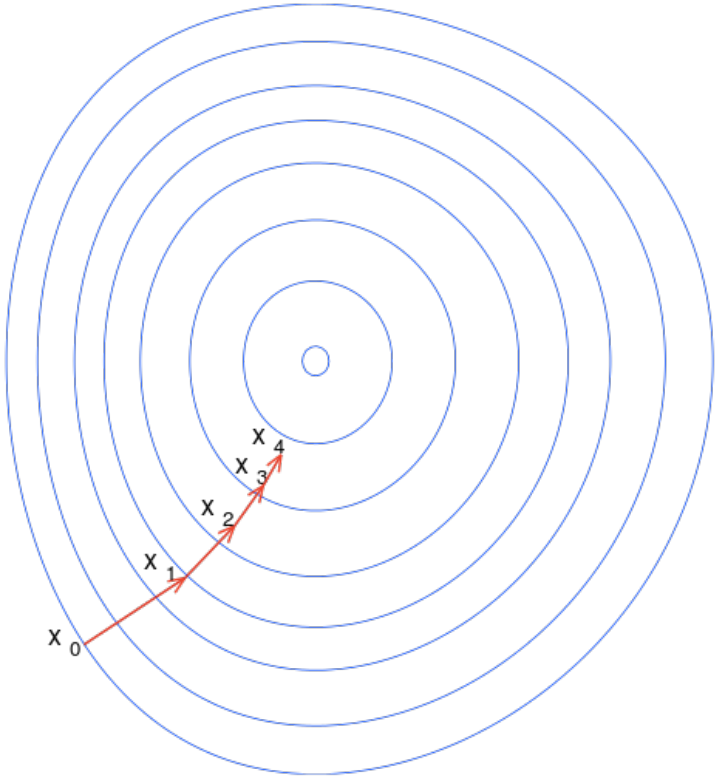
$$P(y=1\,|\,\mathbf{x},\mathbf{w}) = g(w_0 + \mathbf{w}_1^{\intercal}\mathbf{x}) = \frac{1}{1+e^{-(w_0+\mathbf{w}_1^{\intercal}\mathbf{x})}}.$$



- No closed-form solution (like optimization of parameters in neural networks—remember?)

- By gradient-based method (or more advanced numerical methods)

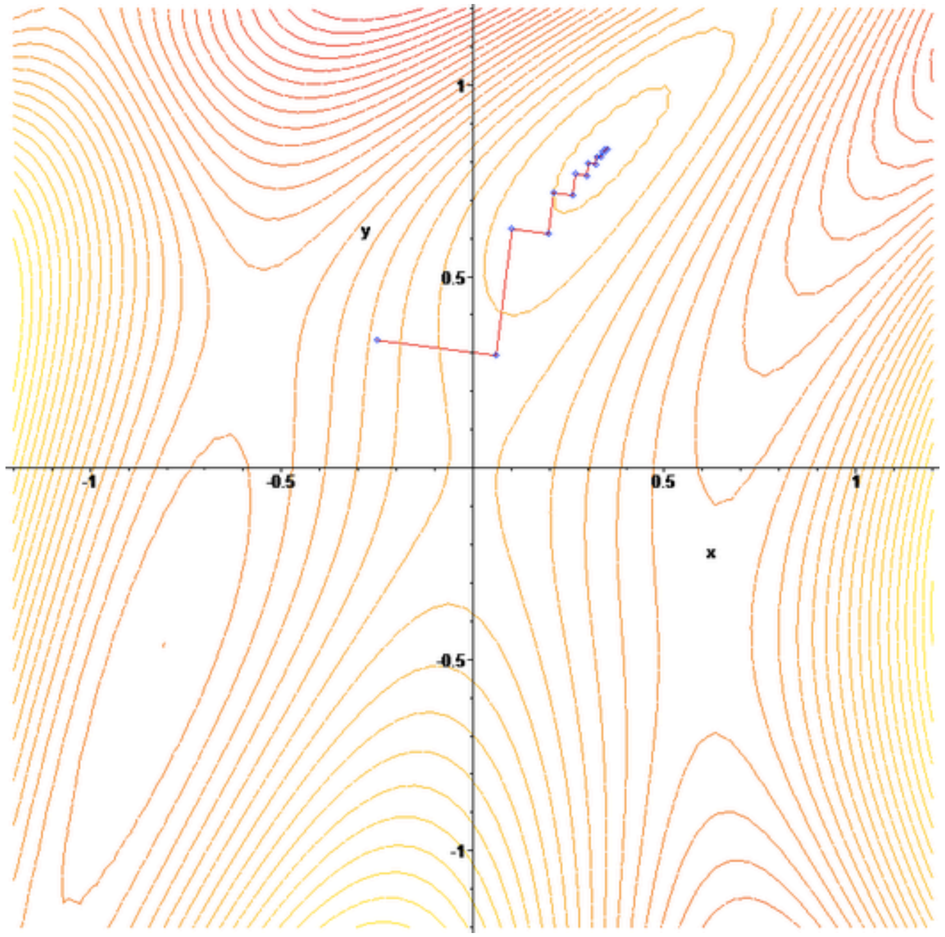
Elementary Optimization: Gradient Method

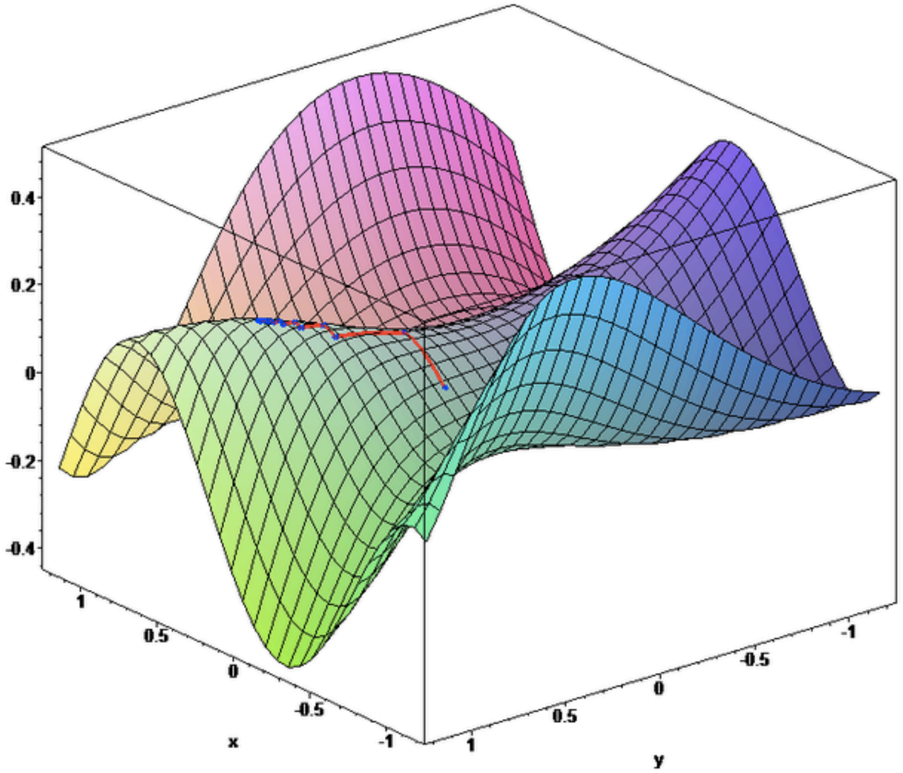


- A general approach to $\min\limits_{x\in\mathbb{R}^n} f(x)$:

$$x^{new} = x^{old} - \gamma \bigtriangledown f(x^{old})$$

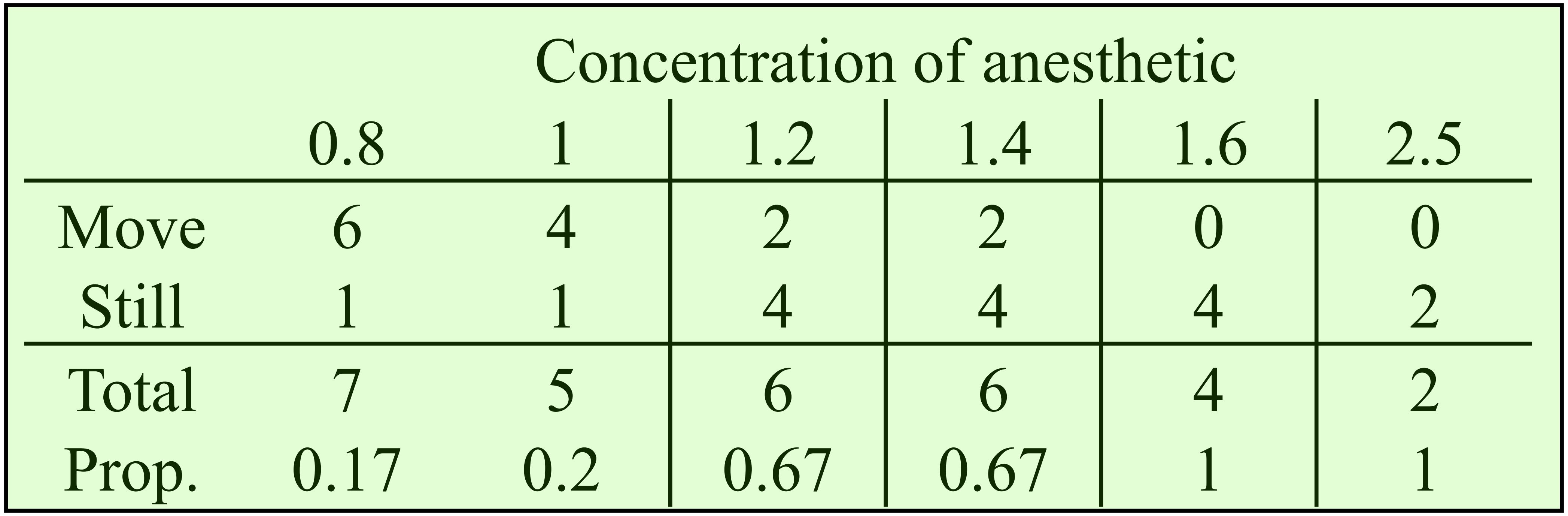
- Example: min $f(x_1, x_2) = \sin\left(\frac{1}{2}x_1^2 - \frac{1}{4}x_2^2 + 3\right)\cos(2x_1 + 1 - e^{x_2}).$





A nice demo is given on https://www.benfrederickson.com/numerical-optimization/

Logistic Regression: Example



	Concentration of anesthetic					
	0.8	1	1.2	1.4	1.6	2.5
Move	6	4	2	2	0	0
Still	1	1	4	4	4	2
Total	7	5	6	6	4	2
Prop.	0.17	0.2	0.67	0.67	1	1

- Maximize the Likelihood function
- Fitted model:

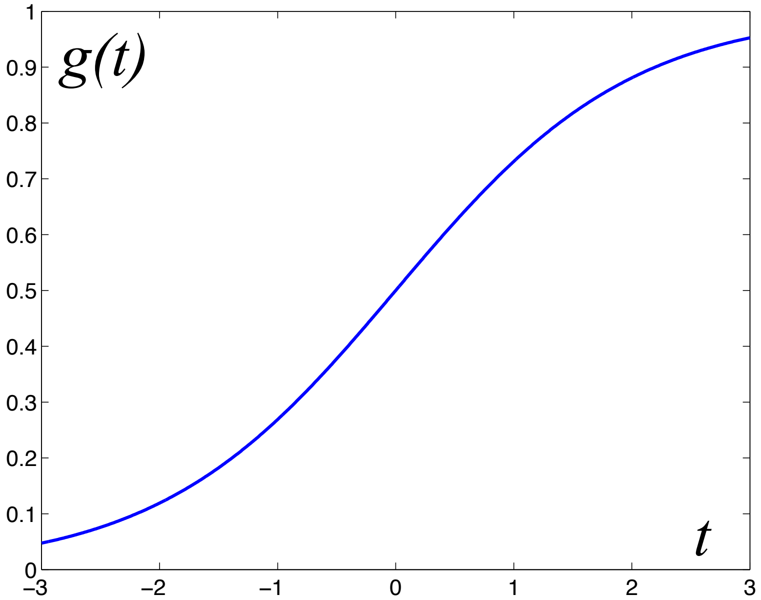
$$P(y = 1 \,|\, x) = \frac{1}{1 + e^{-(-6.47 + 5.57x)}}.$$

Extended Logistic Regression

- As with regression models, we can fit the logistic models using maximum (conditional) log-likelihood:

$$L(\mathbf{w};\mathrm{Data}) = \sum_{i=1}^{n} \log P(y_i \,|\, \mathbf{x}_i, \mathbf{w}), \quad \text{where}$$

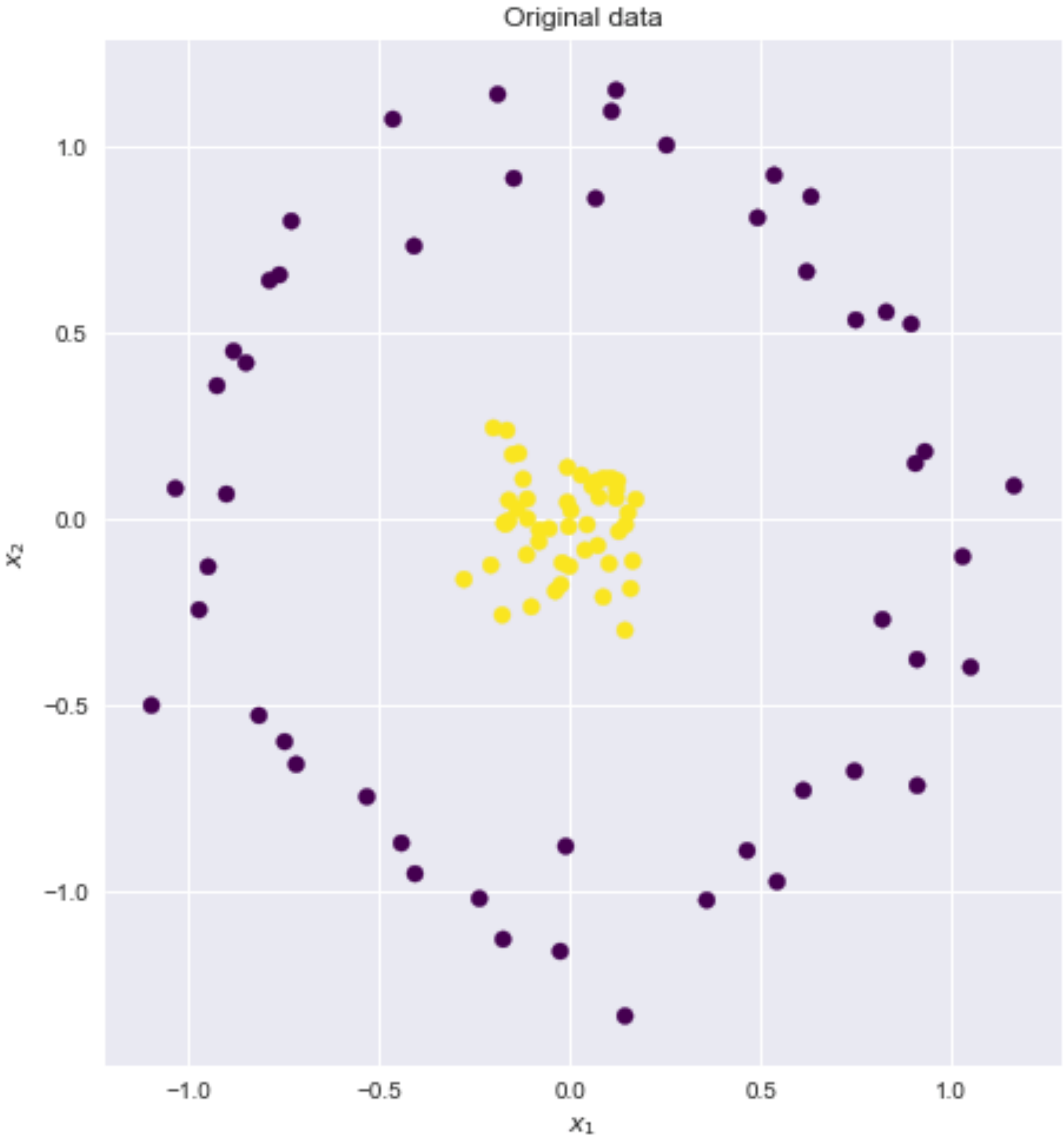
$$P(y=1\,|\,\mathbf{x},\mathbf{w}) = g(w_0 + \mathbf{w}_1^\intercal \mathbf{x}) = \frac{1}{1+e^{-(w_0+\mathbf{w}_1^\intercal \mathbf{x})}}.$$



For instance, we can use additive models instead of the linear model:

$$f(\mathbf{x};\mathbf{w}) = w_0 + w_1\phi_1(\mathbf{x}) + \ldots + w_M\phi_M(\mathbf{x}).$$

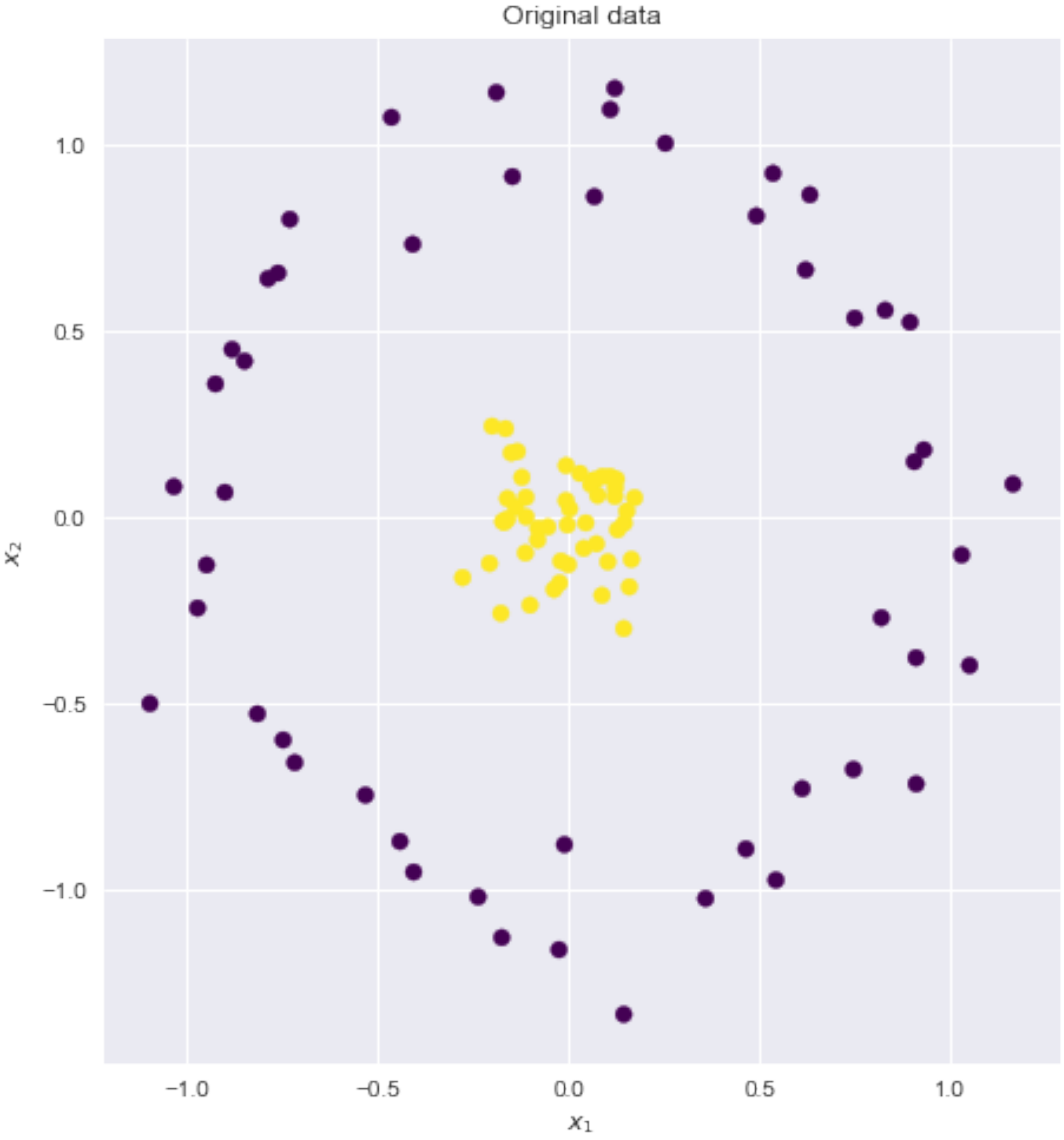
Kernel Methods for Classification

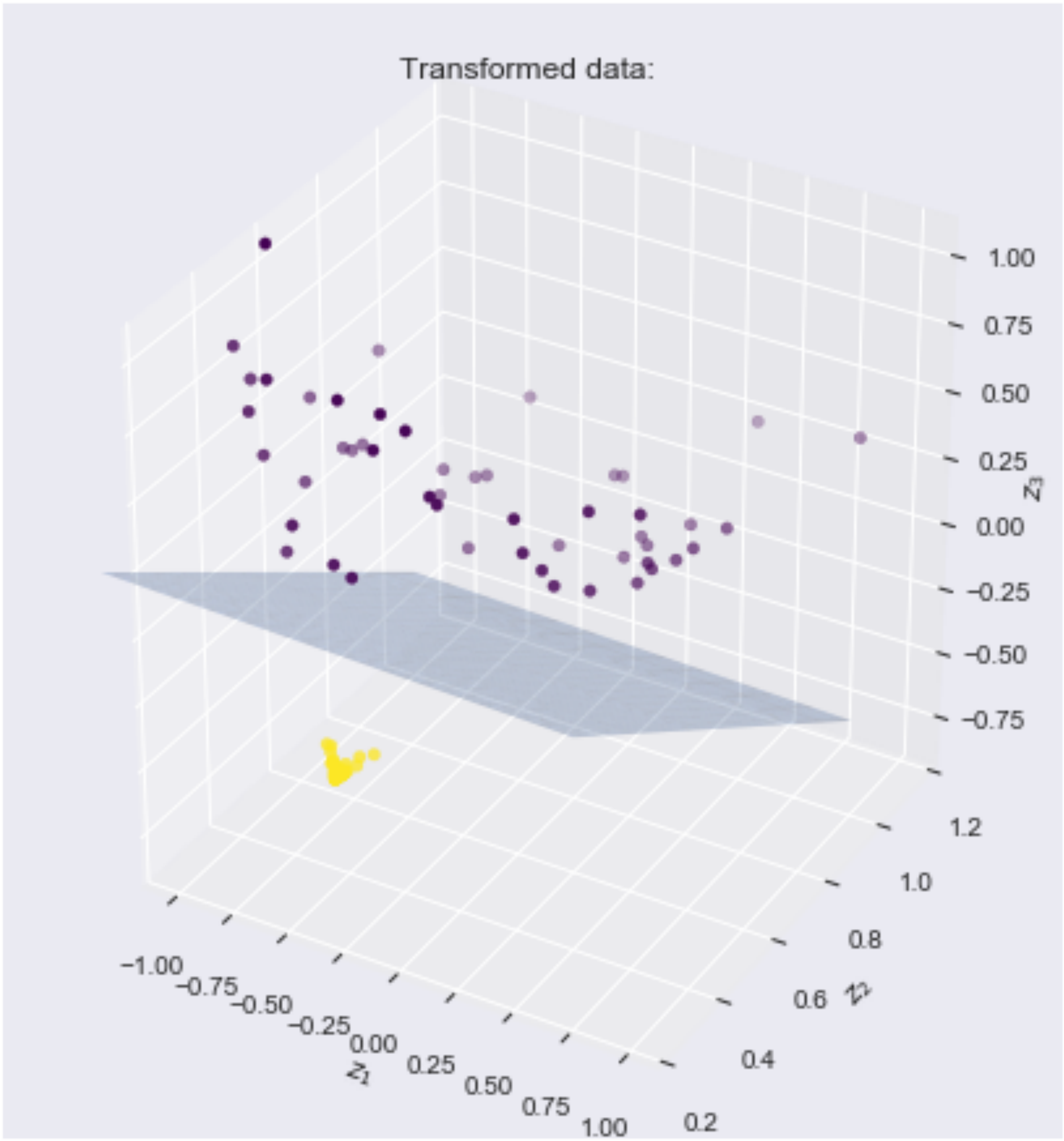


The two classes are not linearly separable. :-(

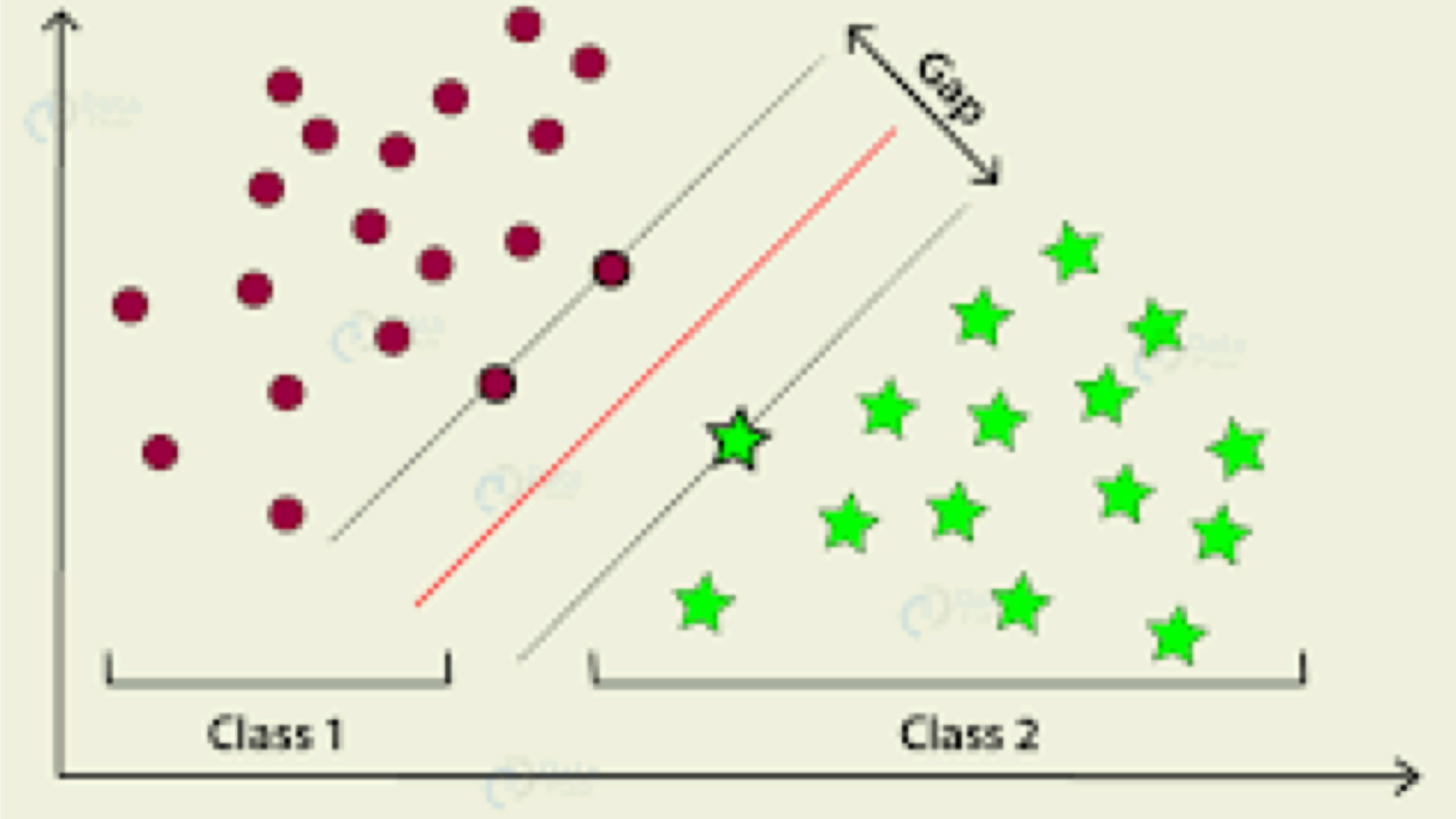
Kernel Methods for Classification

What if you use a polynomial kernel with *p*=2 ?

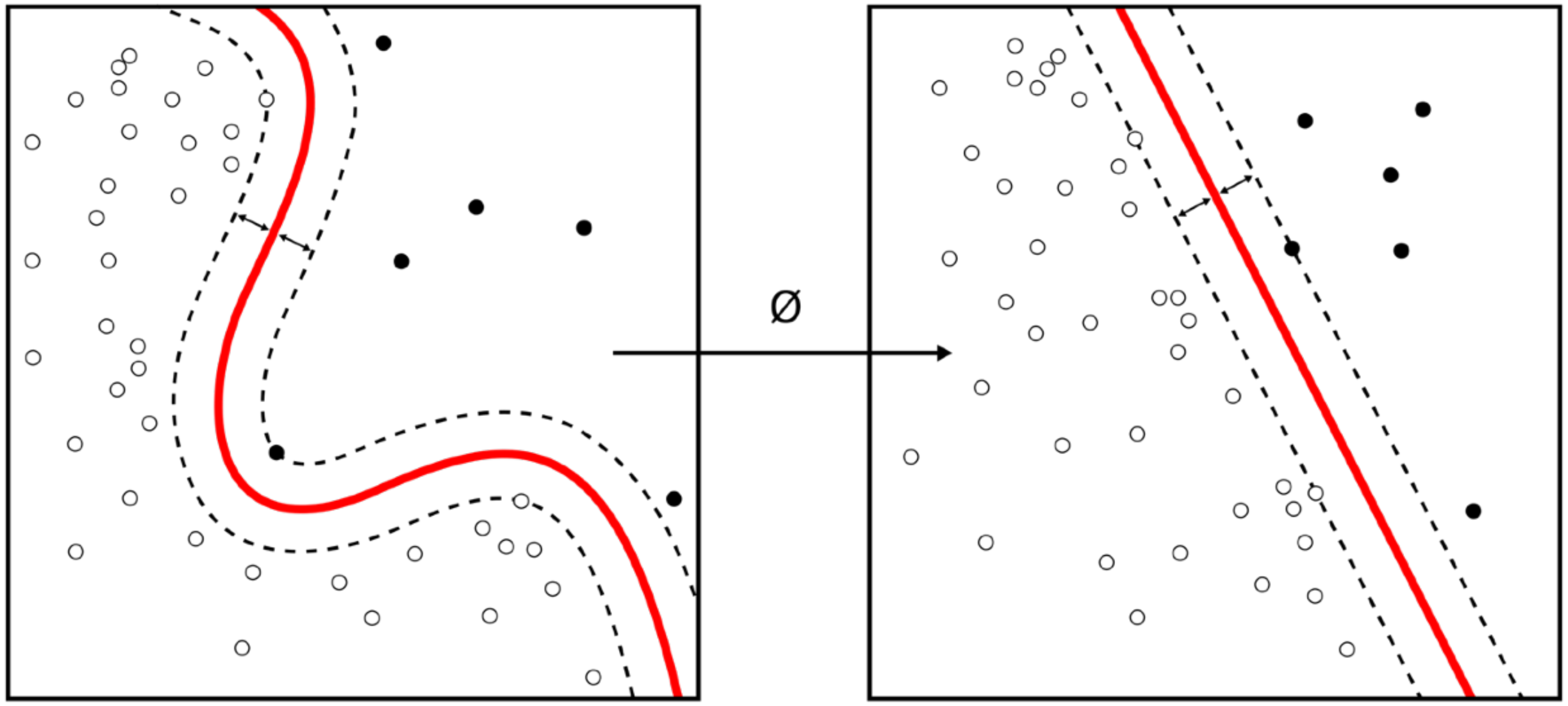




Support Vector Machines

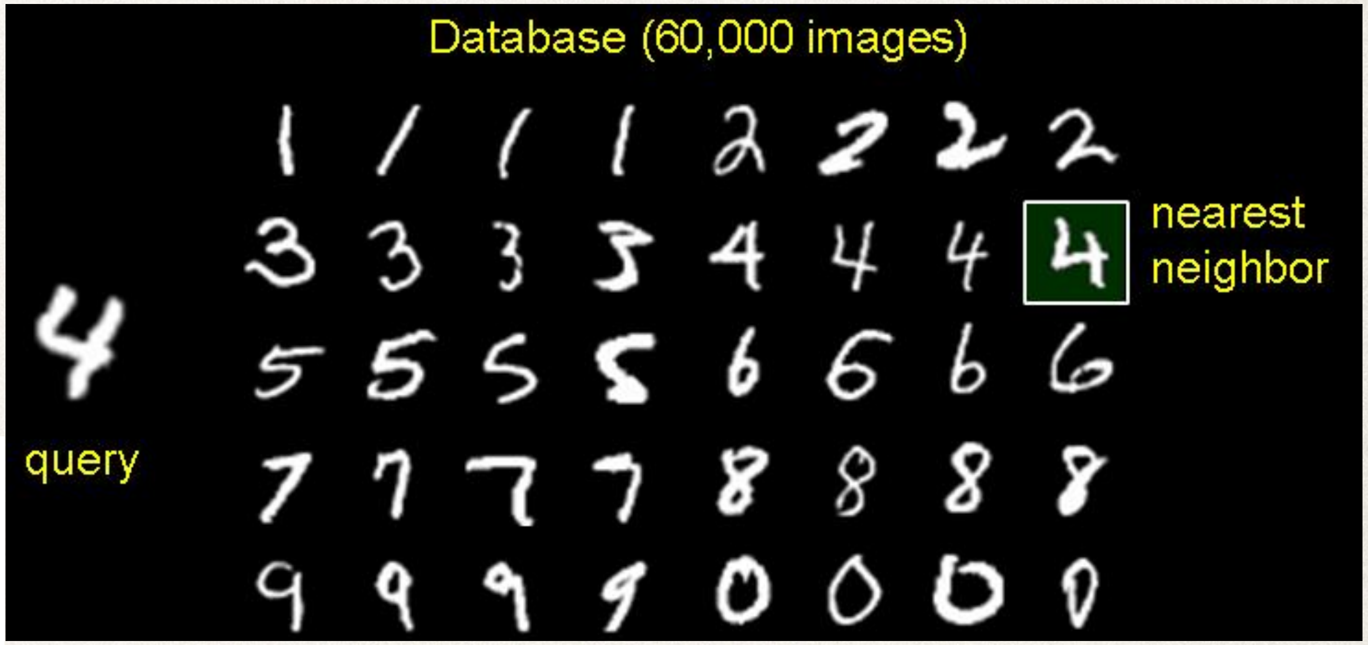


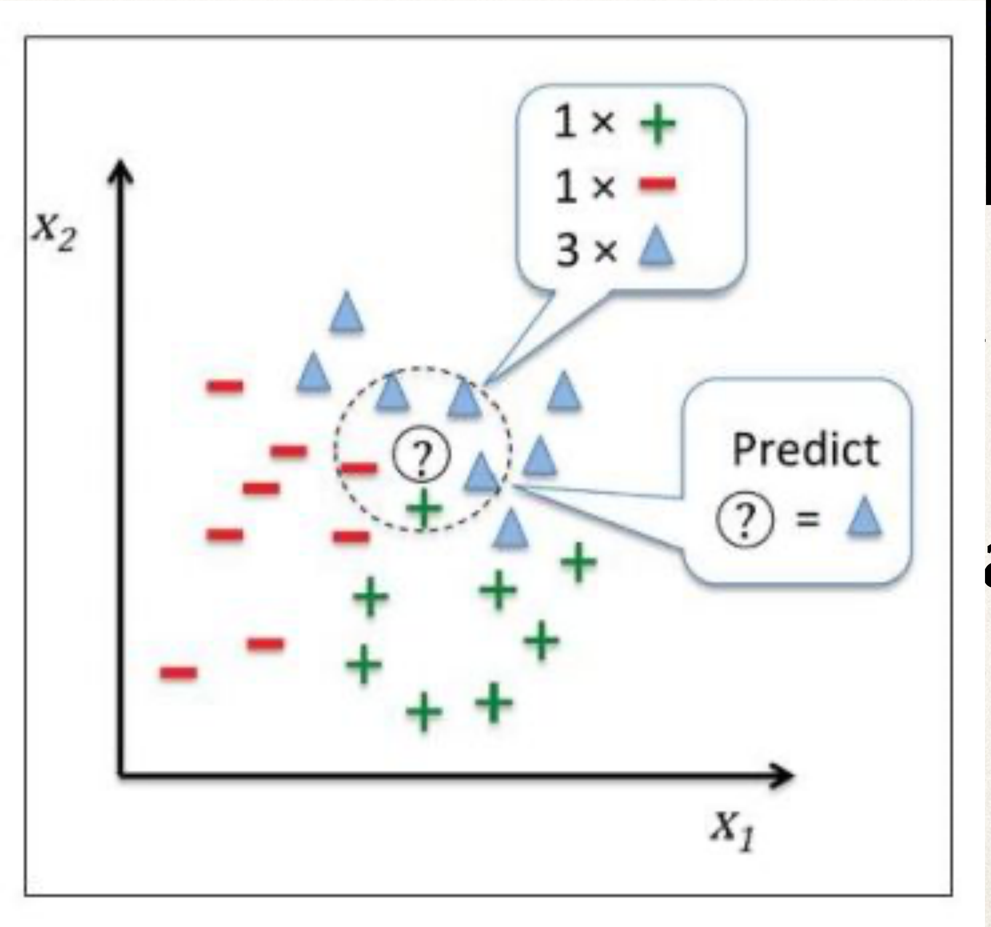
Can perform nonlinear classification with kernel trick.



Supervised Learning Algorithms

- **Nearest-neighbor**
- Dec
- Lin on
- Neu arning
- …

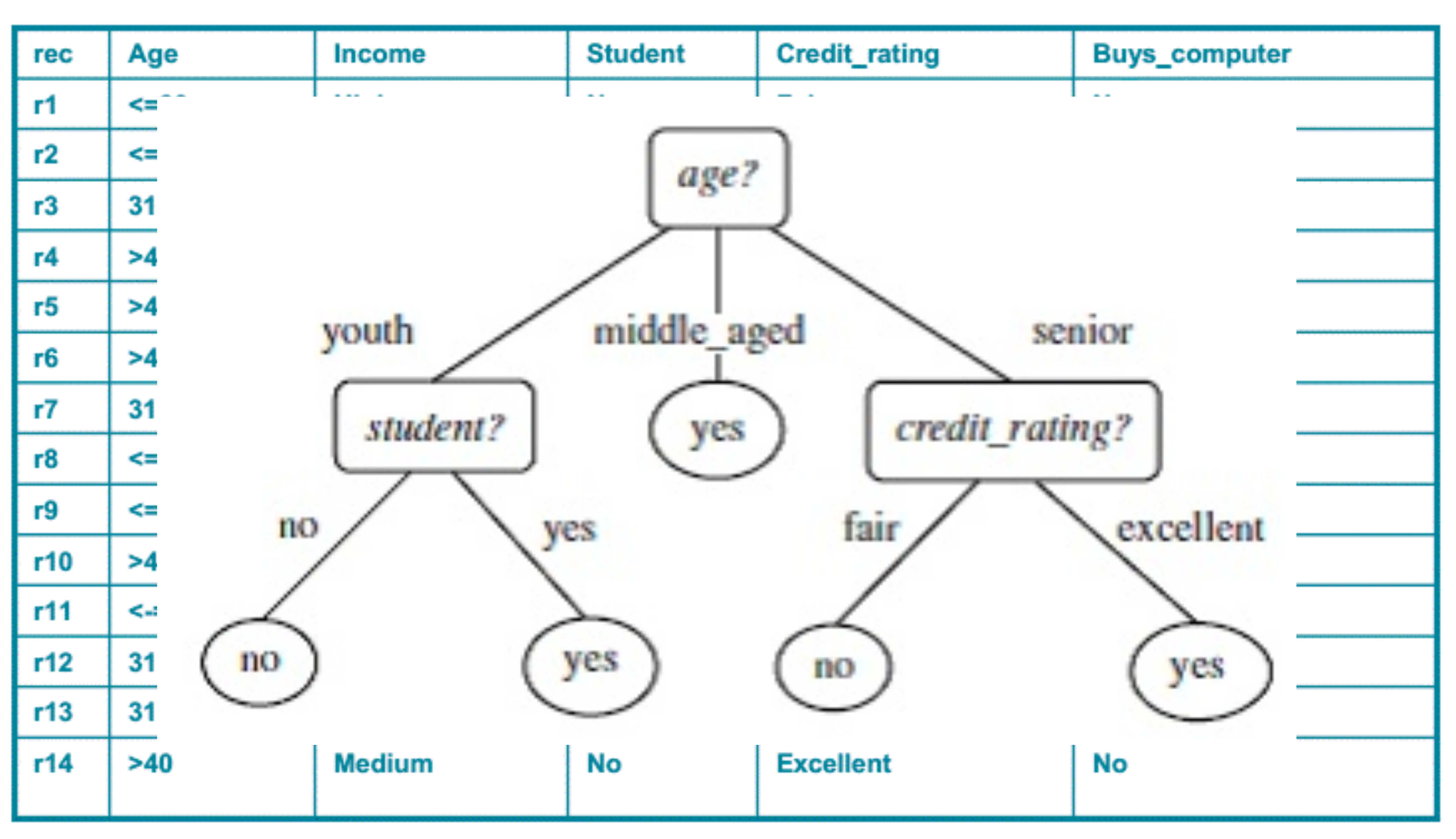




- Choose the number of k and a distance metric (k = 5 is common)
- Find k-nearest neighbors of the sample that you want to classify
- Assign the class label by majority vote

Supervised Learning Algorithms

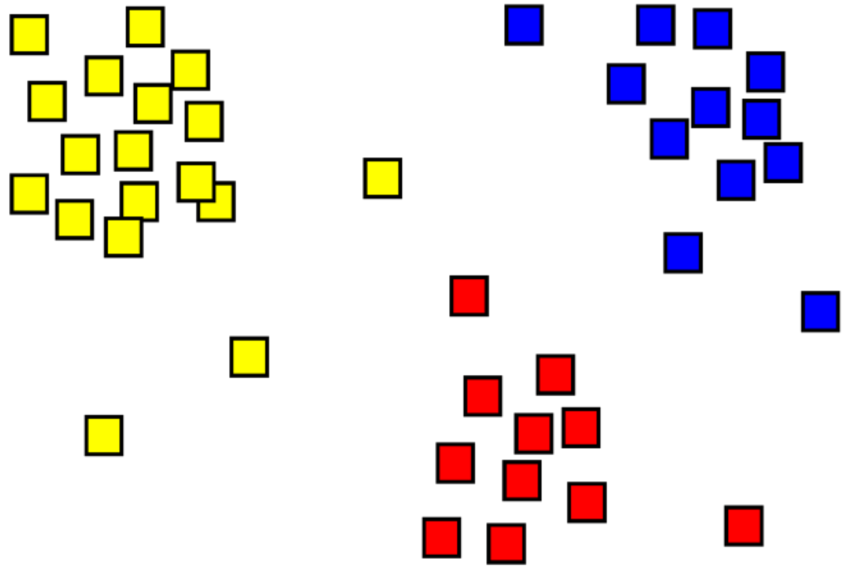
- Nea
- **Dec**
- Line
- Neu
- …



rec	Age	Income	Student	Credit_rating	Buys_computer
r1	<=				
r2	<=				
r3	31				
r4	>4				
r5	>4				
r6	>4				
r7	31				
r8	<=				
r9	<=				
r10	>4				
r11	<				
r12	31				
r13	31				
r14	>40	Medium	No	Excellent	No

Summary: From Regression to Classification

- Classification
  - Fisher linear discriminant, Bayes classifier, logistic regression, decision trees, nearest neighbors, SVM, Kernel methods…
  - What if we are not given y?

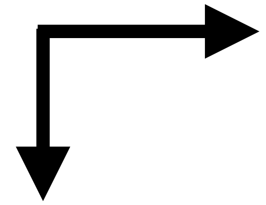


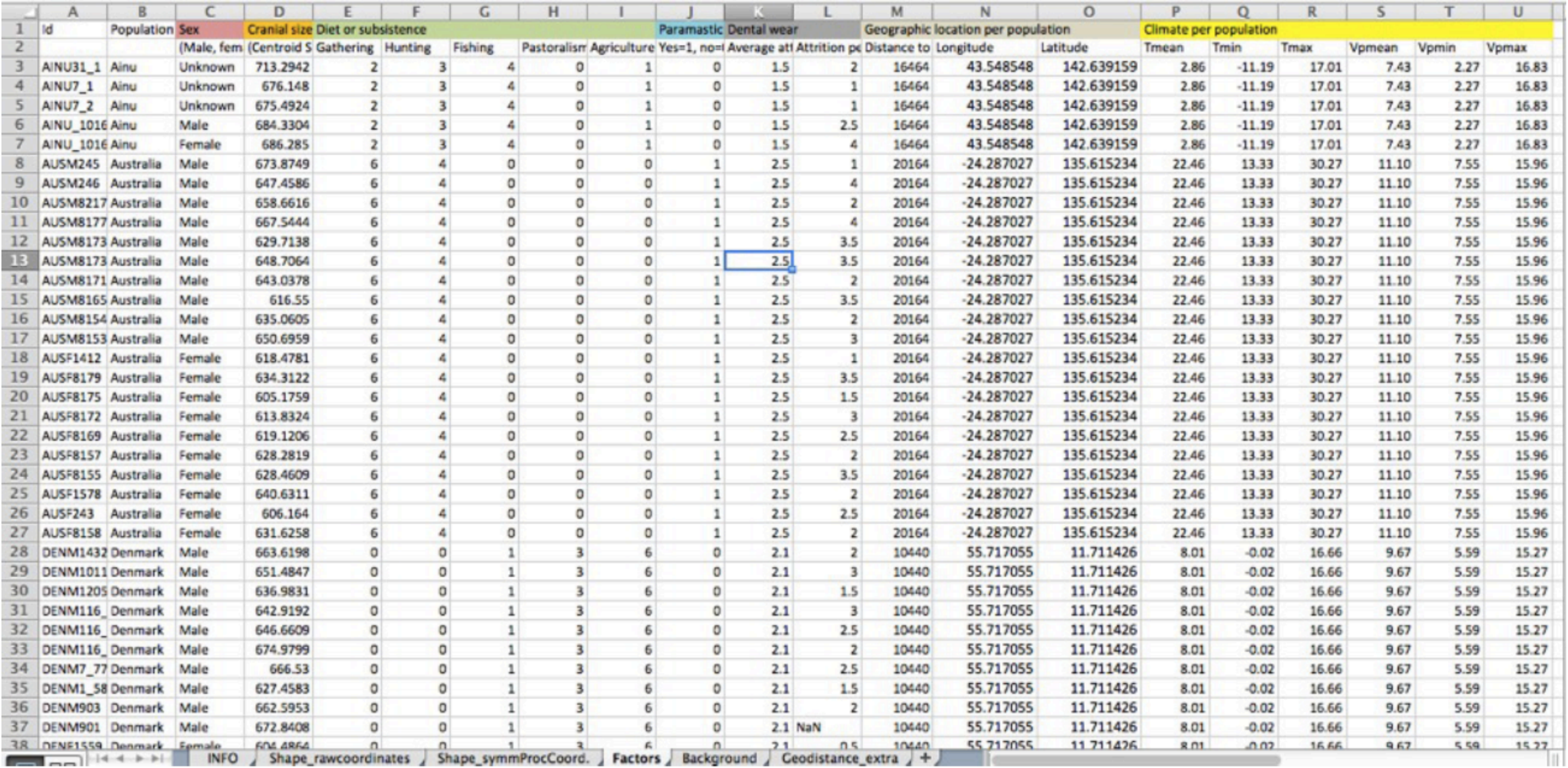
Outline

- Supervised learning
  - From linear regression to nonlinear methods
    - Properties of regression
  - From parametric models to nonparametric models
  - Model selection: Why? What? How?
  - Classification
- Unsupervised learning
  - Clustering ↓
  - Dimensionality reduction… →

Two Ways of Finding *Simpler* Data Representations

- Fewer "data points" vs. fewer dimensions (#variables)?



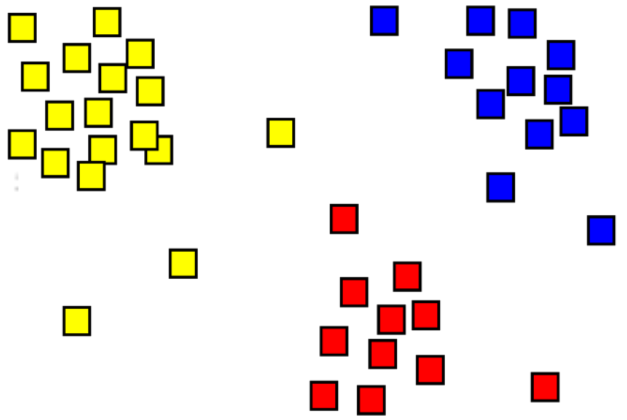


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	Population	Sex	Cranial size	Diet or subsistence					Paramastic	Dental wear		Geographic location per population			Climate per population					
2			(Male, fem	(Centroid S	Gathering	Hunting	Fishing	Pastoralism	Agriculture	Yes=1, no=	Average att	Attrition pe	Distance to	Longitude	Latitude	Tmean	Tmin	Tmax	Vpmean	Vpmin	Vpmax
3	AINU31_1	Ainu	Unknown	713.2942	2	3	4	0	1	0	1.5	2	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
4	AINU7_1	Ainu	Unknown	676.148	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
5	AINU7_2	Ainu	Unknown	675.4924	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
6	AINU_1016	Ainu	Male	684.3304	2	3	4	0	1	0	1.5	2.5	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
7	AINU_1016	Ainu	Female	686.285	2	3	4	0	1	0	1.5	4	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
8	AUSM245	Australia	Male	673.8749	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
9	AUSM246	Australia	Male	647.4586	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
10	AUSM8217	Australia	Male	658.6616	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
11	AUSM8177	Australia	Male	667.5444	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
12	AUSM8173	Australia	Male	629.7138	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
13	AUSM8173	Australia	Male	648.7064	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
14	AUSM8171	Australia	Male	643.0378	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
15	AUSM8165	Australia	Male	616.55	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
16	AUSM8154	Australia	Male	635.0605	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
17	AUSM8153	Australia	Male	650.6959	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
18	AUSF1412	Australia	Female	618.4781	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
19	AUSF8179	Australia	Female	634.3122	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
20	AUSF8175	Australia	Female	605.1759	6	4	0	0	0	1	2.5	1.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
21	AUSF8172	Australia	Female	613.8324	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
22	AUSF8169	Australia	Female	619.1206	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
23	AUSF8157	Australia	Female	628.2819	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
24	AUSF8155	Australia	Female	628.4609	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
25	AUSF1578	Australia	Female	640.6311	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
26	AUSF243	Australia	Female	606.164	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
27	AUSF8158	Australia	Female	631.6258	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
28	DENM1432	Denmark	Male	663.6198	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
29	DENM1011	Denmark	Male	651.4847	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
30	DENM1205	Denmark	Male	636.9831	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
31	DENM116_	Denmark	Male	642.9192	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
32	DENM116_	Denmark	Male	646.6609	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
33	DENM116_	Denmark	Male	674.9799	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
34	DENM7_77	Denmark	Male	666.53	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
35	DENM1_58	Denmark	Male	627.4583	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
36	DENM903	Denmark	Male	662.5953	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
37	DENM901	Denmark	Male	672.8408	0	0	1	3	6	0	2.1	NaN	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
38	DENF1559	Denmark	Female	604.4864	0	0	1	3	6	0	2.1	0.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27

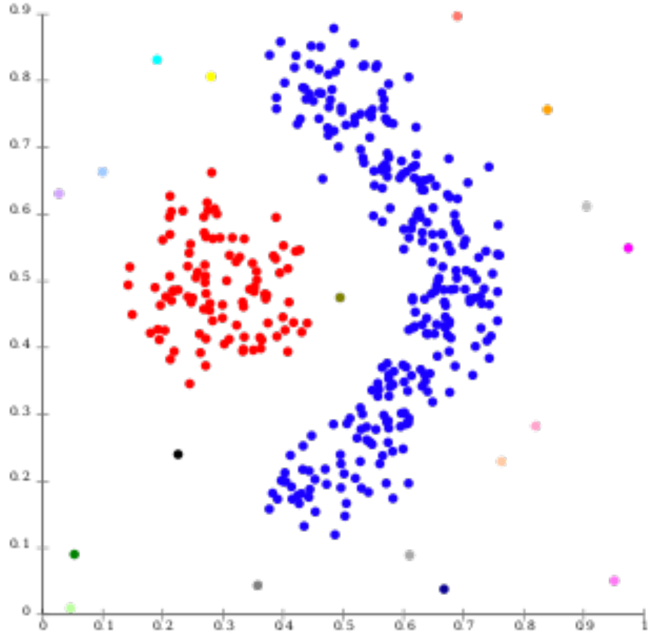
Unsupervised Learning

- Draw inferences from datasets consisting of input data without labeled responses
- Visualization, understanding…
- Clustering
  - Centroid-based clustering
  - Distribution-based clustering
  - Connectivity-based clustering…

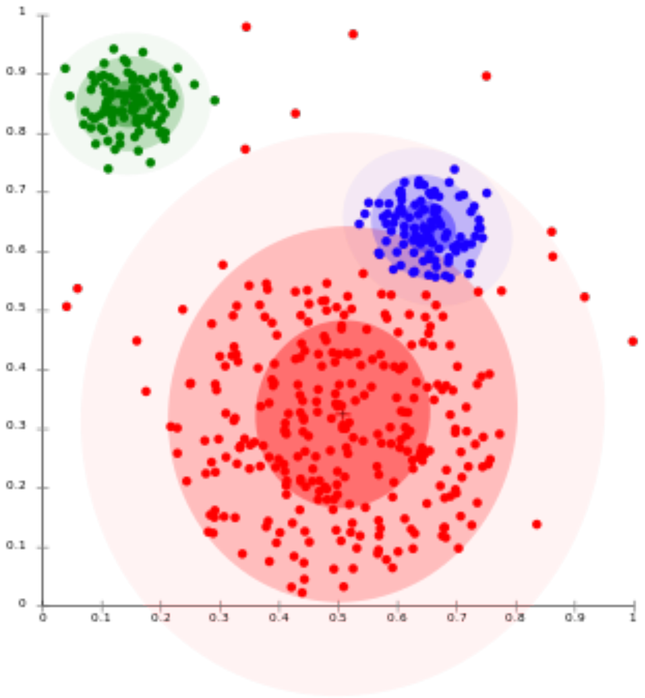
$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$



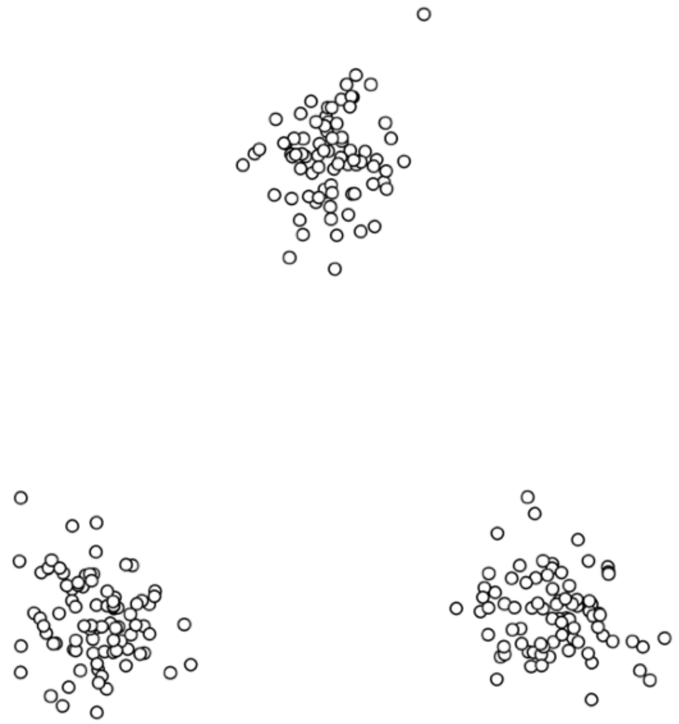
Single-linkage clustering example



Gaussian mixture model clustering example



K-Means Clustering

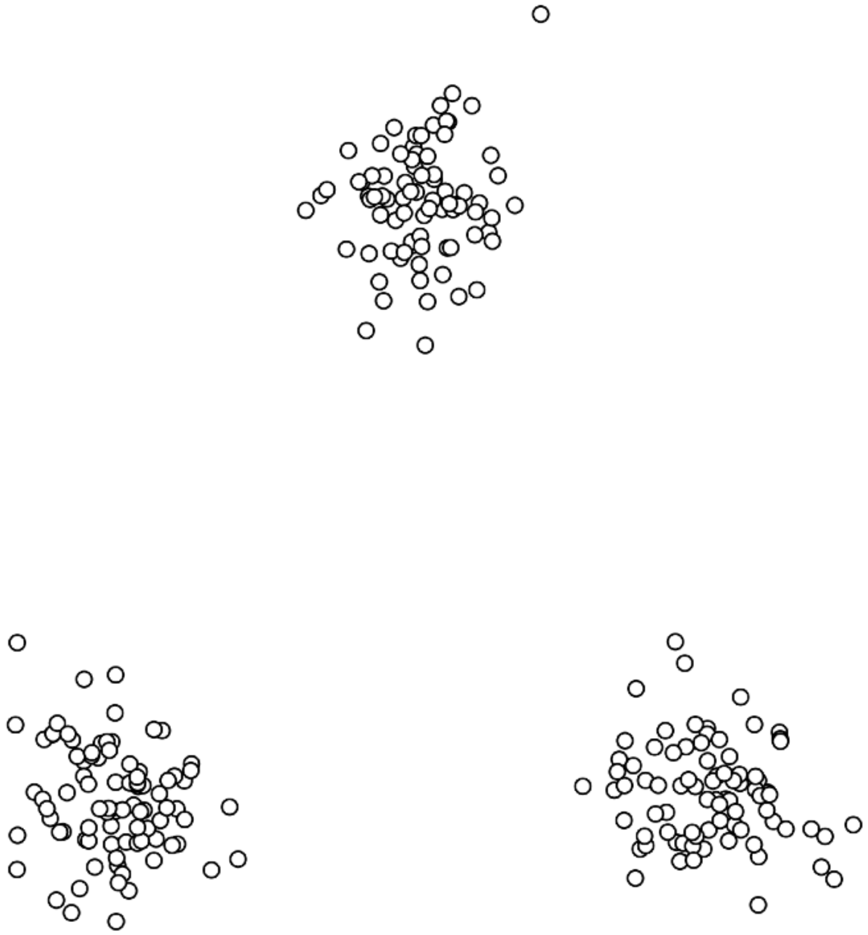


- Goal: Partition *n* observations into *K* clusters in which each observation belongs to the cluster with the nearest mean (or center), serving as a prototype of the cluster (as a method of vector quantization in signal processing).
- A bit history (https://en.wikipedia.org/wiki/K-means_clustering#History): The term "K-means" was first used by James MacQueen in 1967, though the idea goes back to Hugo Steinhaus in 1956. The standard algorithm was first proposed by Stuart Lloyd of Bell Labs in 1957 as a technique for pulse-code modulation, although it was not published as a journal article until 1982. In 1965, Edward W. Forgy published essentially the same method, which is why it is sometimes referred to as the Lloyd–Forgy algorithm.
- Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), where each observation is a d-dimensional real vector, K-means clustering aims to partition the *n* observations into *K* ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, ..., S_K\}$ so as to minimize the within-cluster sum of squares:

$$\arg\min_{\mathbf{S}} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} ||\mathbf{x}_i - \boldsymbol{\mu}_k||^2, \quad \text{where} \quad \boldsymbol{\mu}_k = \frac{1}{|S_k|} \sum_{\mathbf{x}_i \in S_k} \mathbf{x}_i.$$

K-Means Clustering: Procedure

$$\arg\min_{\mathbf{S}} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} ||\mathbf{x}_i - \boldsymbol{\mu}_k||^2, \quad \text{where} \quad \boldsymbol{\mu}_k = \frac{1}{|S_k|} \sum_{\mathbf{x}_i \in S_k} \mathbf{x}_i.$$



- Think of it as an alternating optimization procedure (by alternating optimizations over the individual subsets of parameters)

- Start with some guess of the means of the clusters

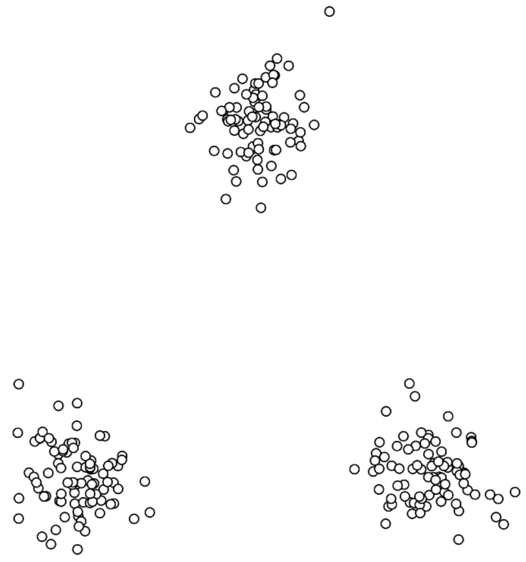
- Refer to https://www.naftaliharris.com/blog/visualizing-k-means-clustering/ to see how we can update the partitioning and means iteratively

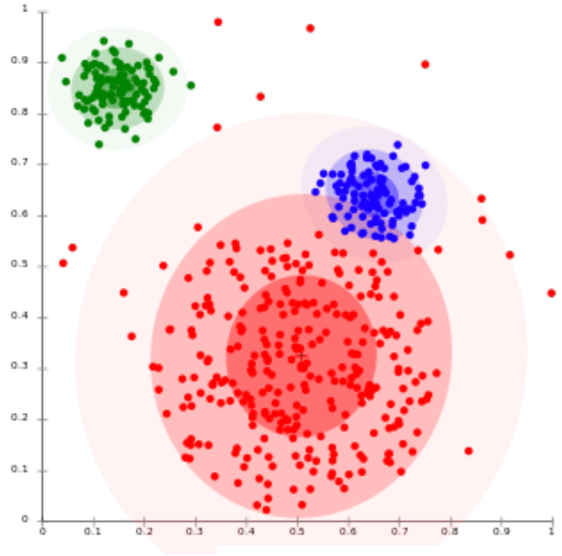
K-Means Clustering: Procedure

$$\arg\min_{\mathbf{S}} \sum_{k=1}^{K}\sum_{\mathbf{x}_i\in S_k} ||\mathbf{x}_i-\boldsymbol{\mu}_k||^2, \quad \text{where} \quad \boldsymbol{\mu}_k = \frac{1}{|S_k|}\sum_{\mathbf{x}_i\in S_k}\mathbf{x}_i.$$

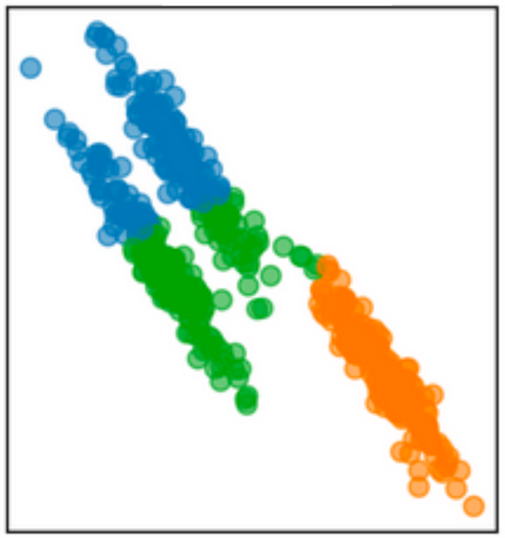
- Procedure: Given an initial set of K means $\mu_1^{(1)},\ldots,\mu_K^{(1)}$, the algorithm proceeds by alternating between two steps, until convergence:
  - **Assignment** step: Assign each observation to the cluster with the nearest mean (i.e., with the least squared Euclidean distance)
  - **Update** step: Recalculate means (centroids) μ_k for observations assigned to each cluster.
- Problems: local minima, strong assumptions…
- How to make it more flexible?

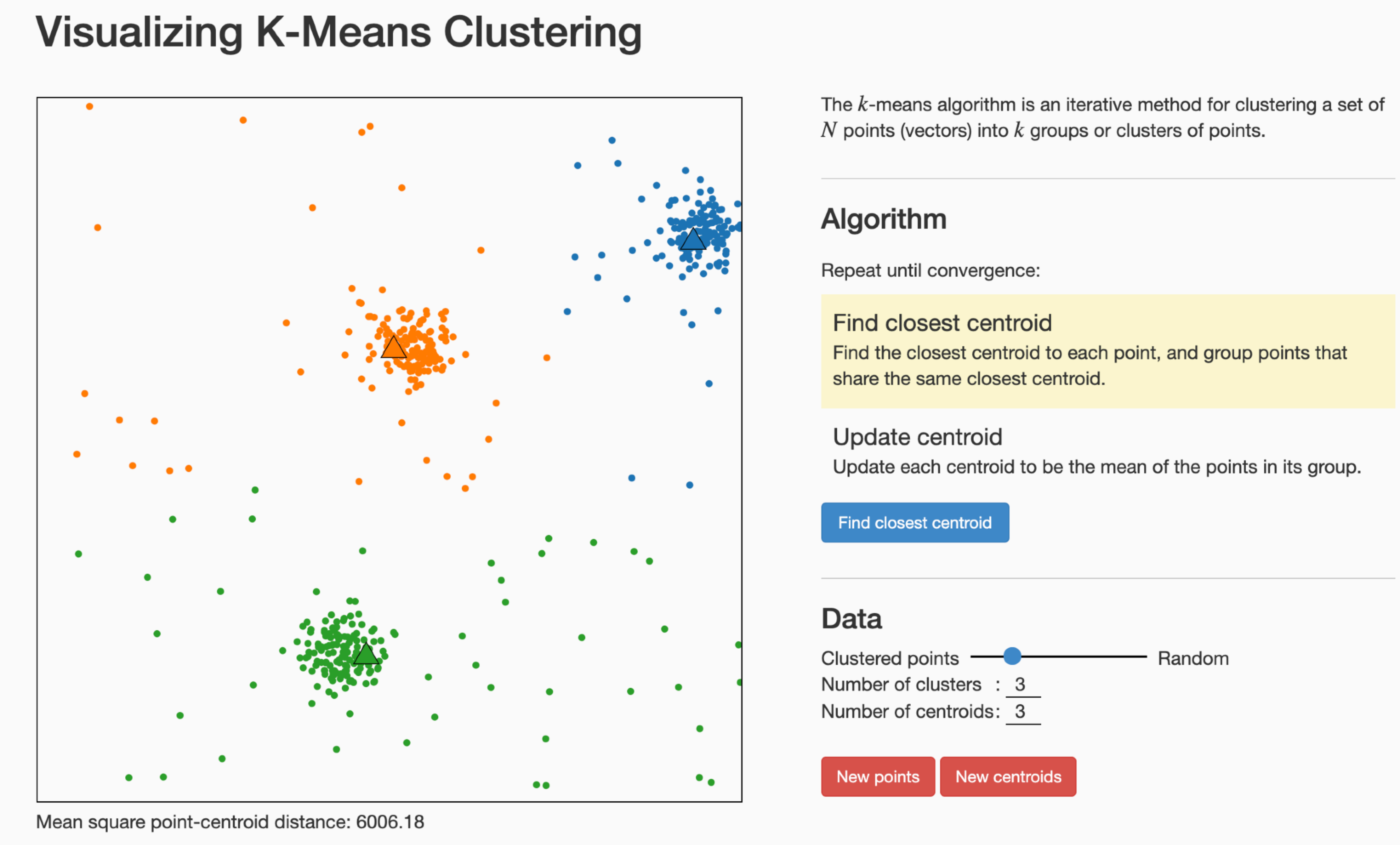
Winers take all!





Issues…





Visualizing K-Means Clustering

Mean square point-centroid distance: 6006.18

The k-means algorithm is an iterative method for clustering a set of N points (vectors) into k groups or clusters of points.

Algorithm

Repeat until convergence:

Find closest centroid

Find the closest centroid to each point, and group points that share the same closest centroid.

Update centroid

Update each centroid to be the mean of the points in its group.

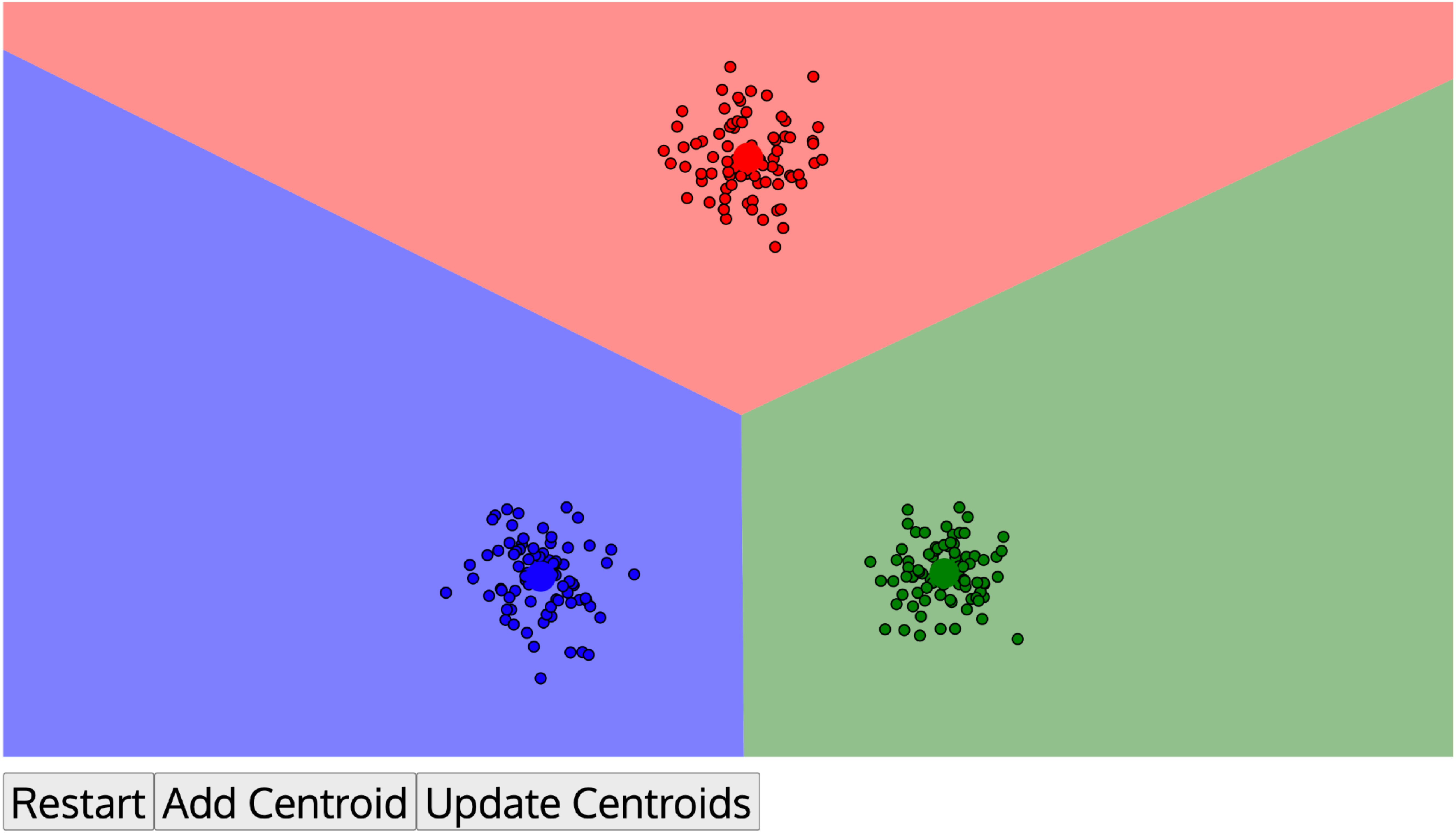
Data

Clustered points ● Random

Number of clusters : 3

Number of centroids: 3

https://stanford.edu/class/engr108/visualizations/kmeans/kmeans.html



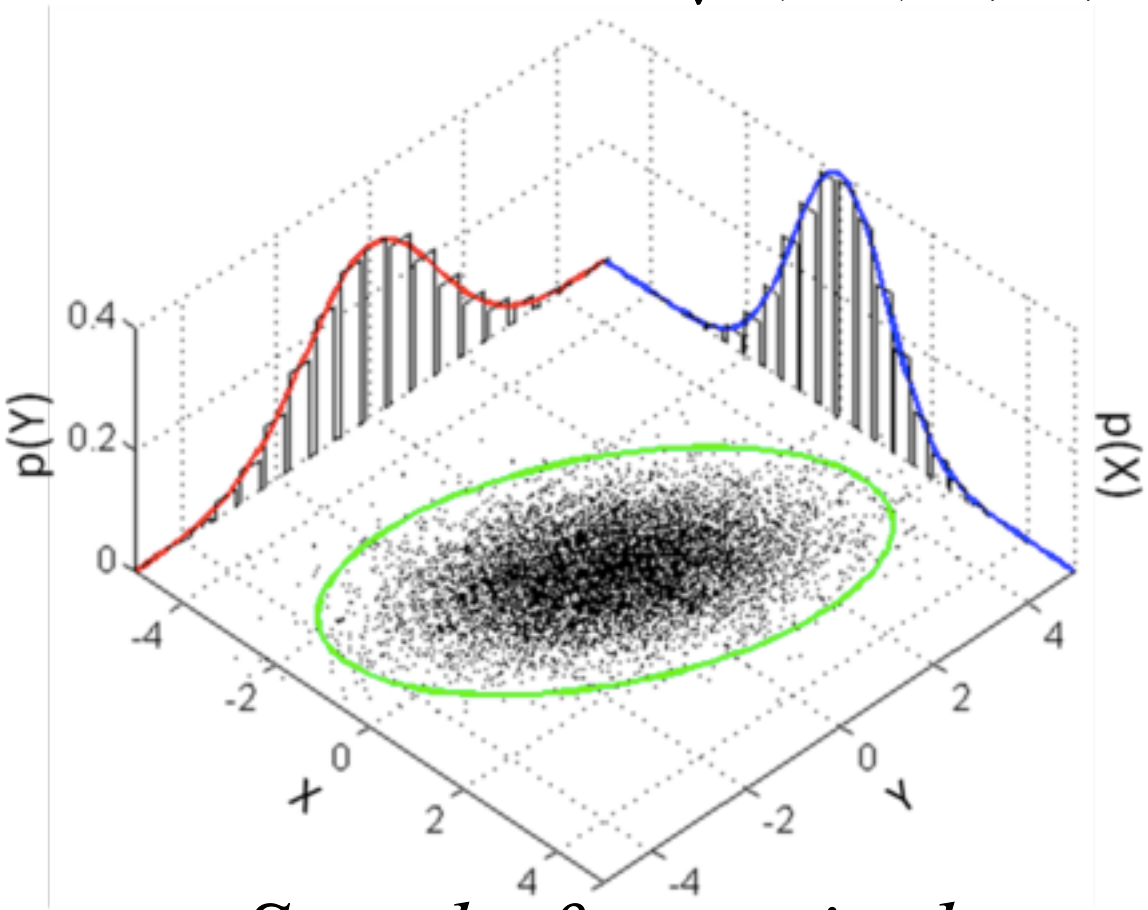
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Multivariate Normal Distribution

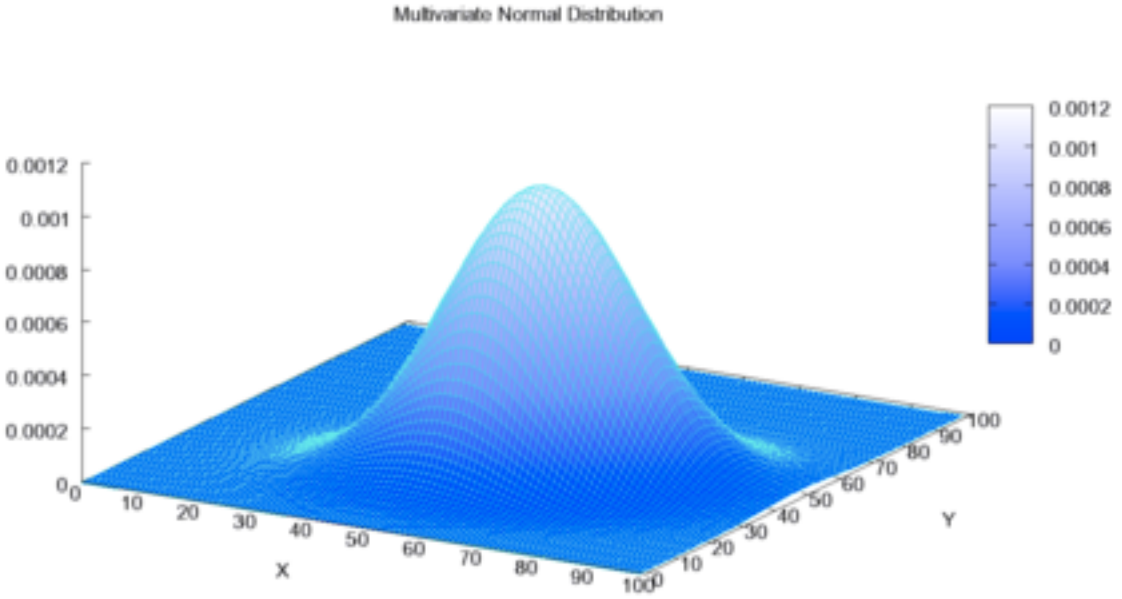
$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots \\ Cov(X_2, X_1) & Var(X_2) & \cdots \\ \vdots & \vdots & \end{bmatrix}$$

- PDF for d-dimensional point **x**, specified by mean **μ** and covariance matrix :

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\intercal} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$



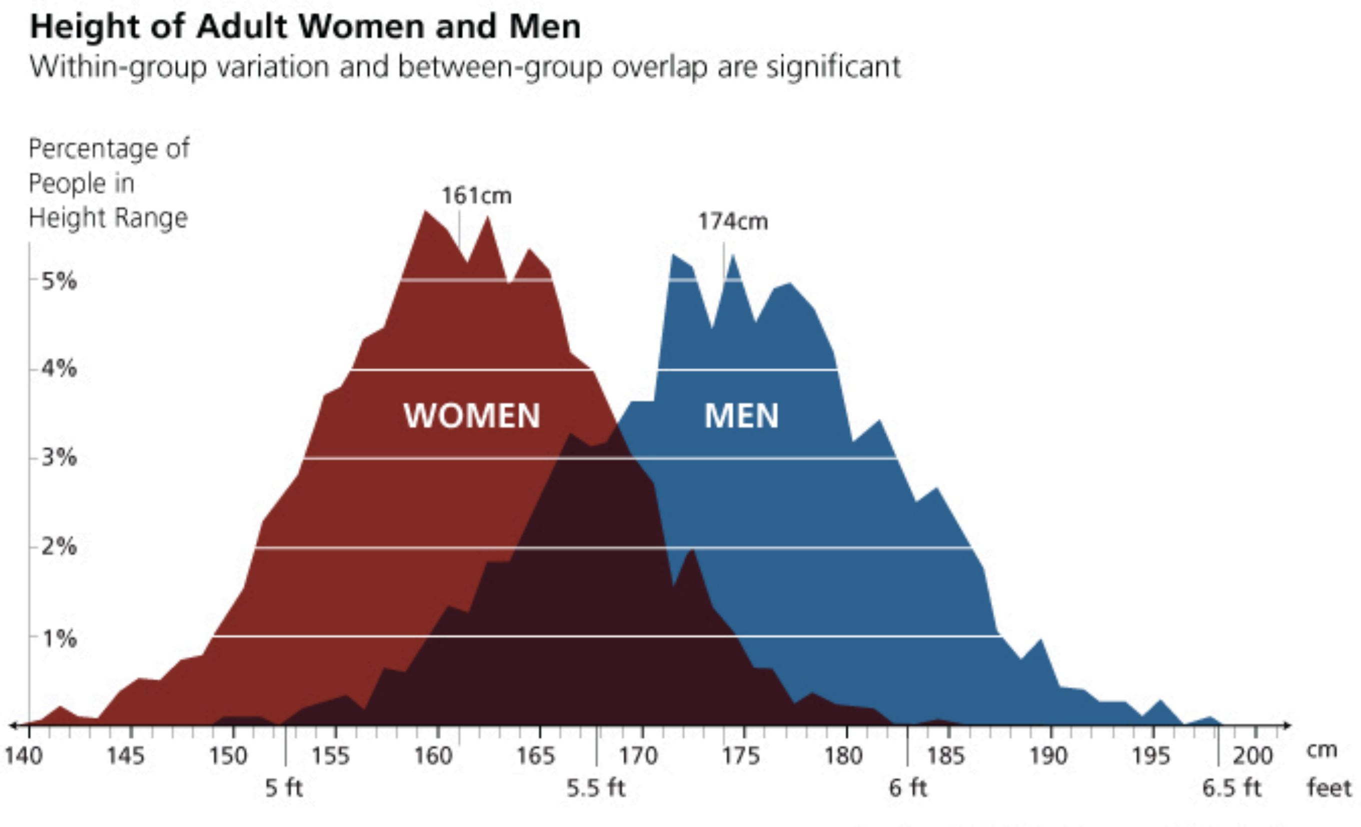
Sample & marginal



pdf

Remember This Way of Specifying Causal Mechanisms?

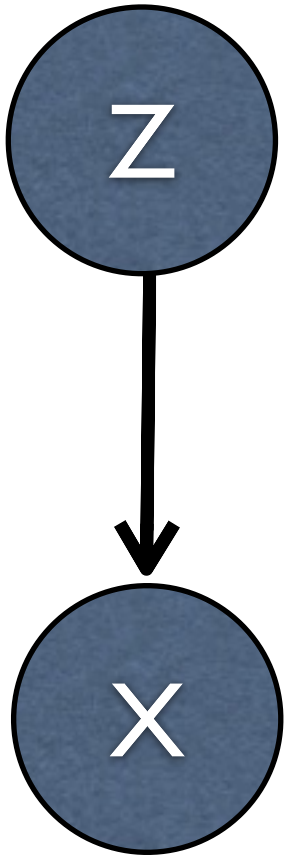
- Let X be sex and Y be height



Height of Adult Women and Men

Within-group variation and between-group overlap are significant

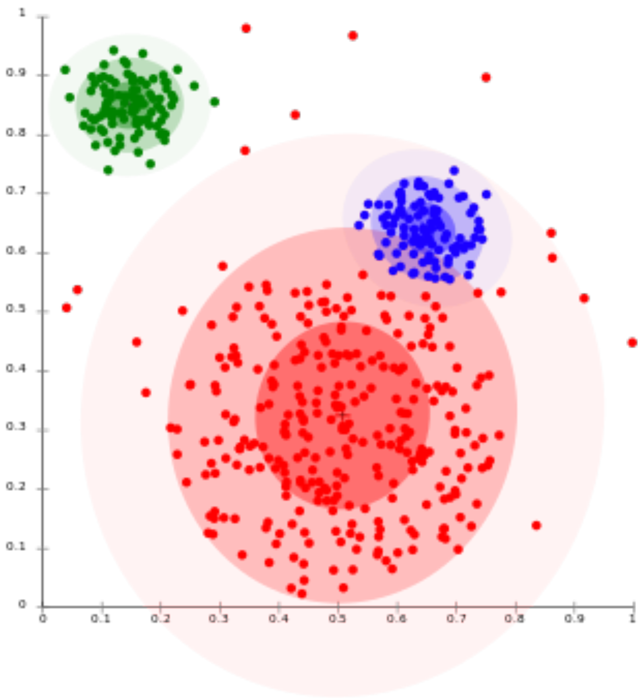
Data from U.S. CDC, adults ages 18-86 in 2007



- How can you specify the causal mechanism?
- How about using a conditional Gaussian model?

GMM for Clustering

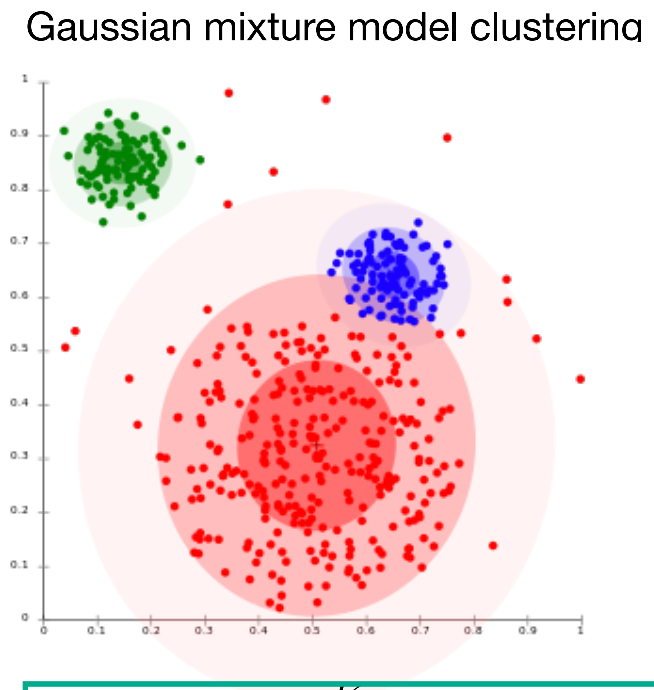
- Z: which cluster the observation is from. $P(Z=k)=\pi_k$
- $p(\boldsymbol{X}=\mathbf{x}|Z=k) = N(\mathbf{x};\, \mu_k,\, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\intercal \Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$
- What is the distribution of $\boldsymbol{X}$?
    - Distinguish between Gaussian Mixture Model (GMM) and *the sum of Gaussian variables*
- Fit the GMM to the data, and then $P(Z=k|\boldsymbol{X}=\mathbf{x}_i)$, known as the posterior/membership prob. of Z, provides a *soft* way of clustering the ith data point
- How to estimated the parameters?



Gaussian mixture model clustering example

$$p(\mathbf{X}=\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x};\boldsymbol{\mu}_k,\Sigma_k)$$

GMM for Clustering: Parameter Estimation


Gaussian mixture model clustering

$$p(\mathbf{X} = \mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k)$$

- Z: which cluster the observation is from. $P(Z=k)=\pi_k$
- $p(\boldsymbol{X}=\mathbf{x}|Z=k) = N(\mathbf{x};\, \mu_k,\, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\intercal \Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$
- How to estimated the parameters **Θ**? Maximum likelihood estimation. Gradient-based?
- Expectation-Maximization (EM) algorithm: A general technique for finding maximum likelihood estimators with latent variables (Z)
    - E step: Estimate the latent variable Z with the posterior

$$h_{ki} := P(Z = k|\mathbf{X} = \mathbf{x}_i) = \frac{\pi_k N(\mathbf{x}_i; \mu_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l N(\mathbf{x}_i; \mu_l, \Sigma_l)}$$

    - M step: update parameters **Θ**

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n} h_{ki}; \quad \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} h_{ki}\mathbf{x}_i}{\sum_{i=1}^{n} h_{ki}}; \quad \Sigma_k = \frac{\sum_{i=1}^{n} h_{ki}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\intercal}{\sum_{i=1}^{n} h_{ki}}.$$

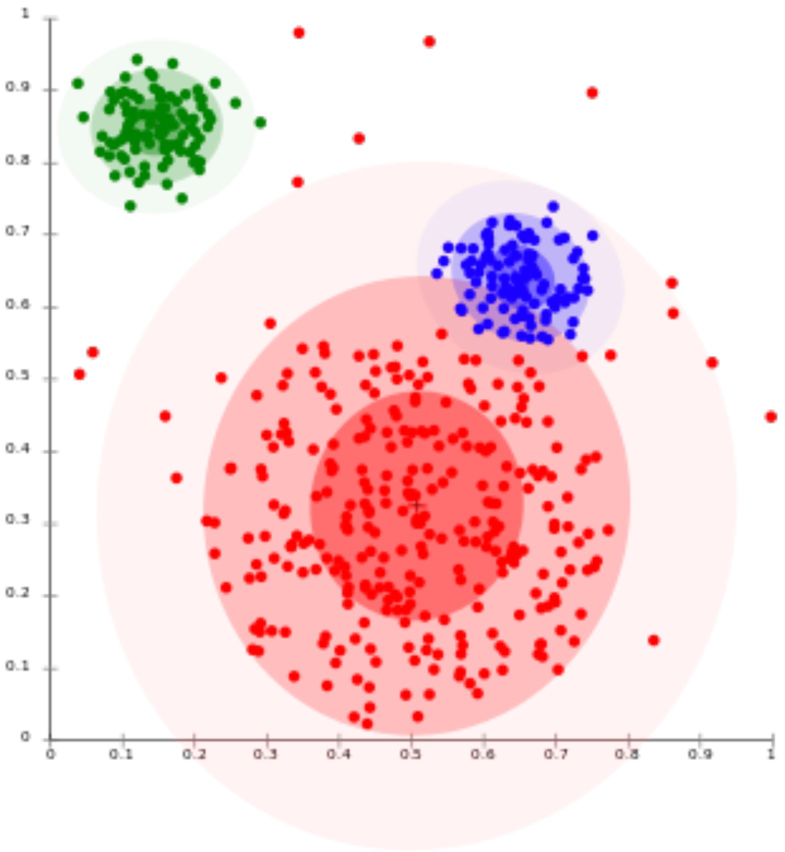
*

GMM for Clustering: A Demo

- What does the decision boundary looks like?

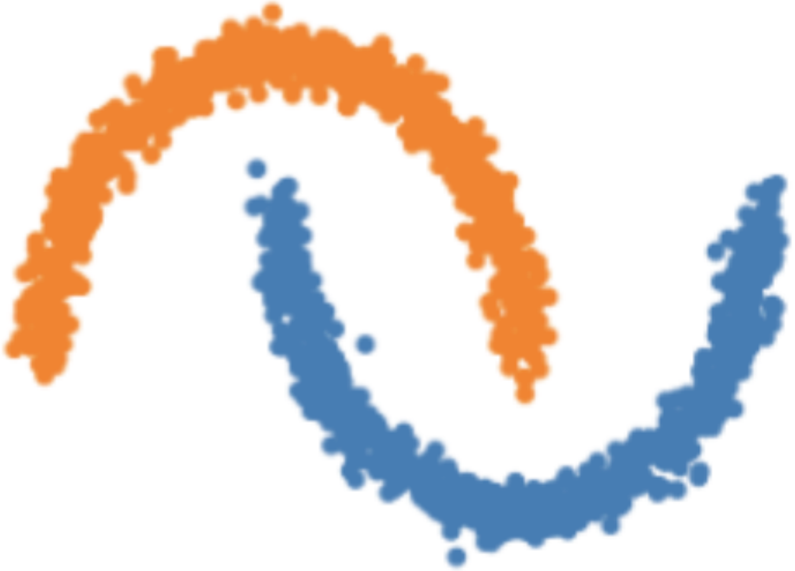
- Demo: https://lukapopijac.github.io/gaussian-mixture-model/

Gaussian mixture model clustering example



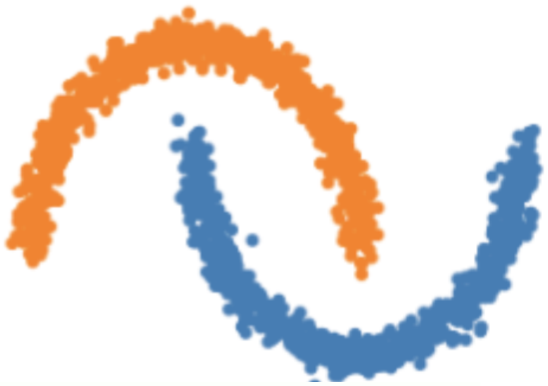
What If We Have Such Data…

Issues…



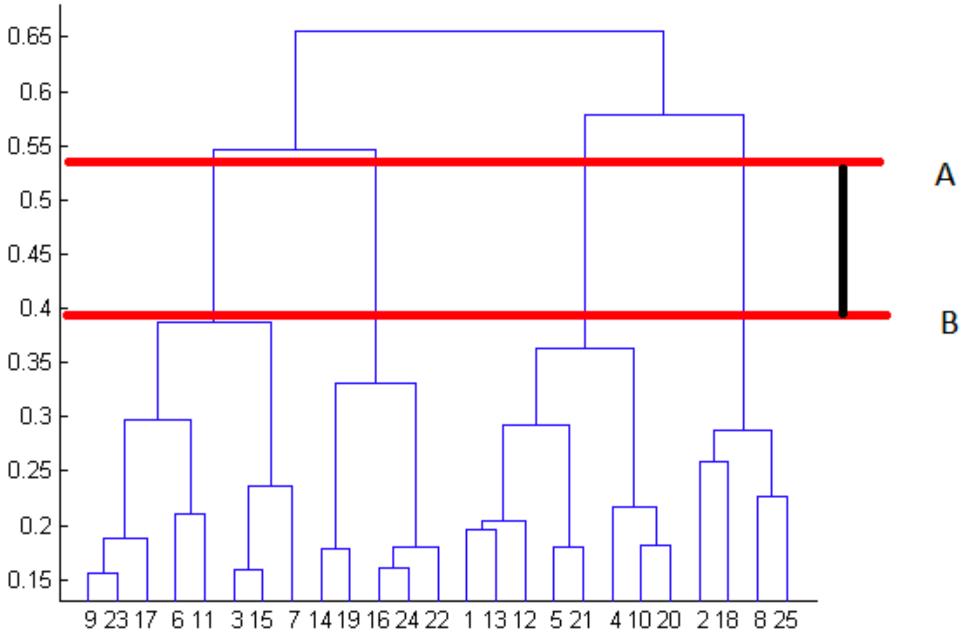
Agglomerative Clustering

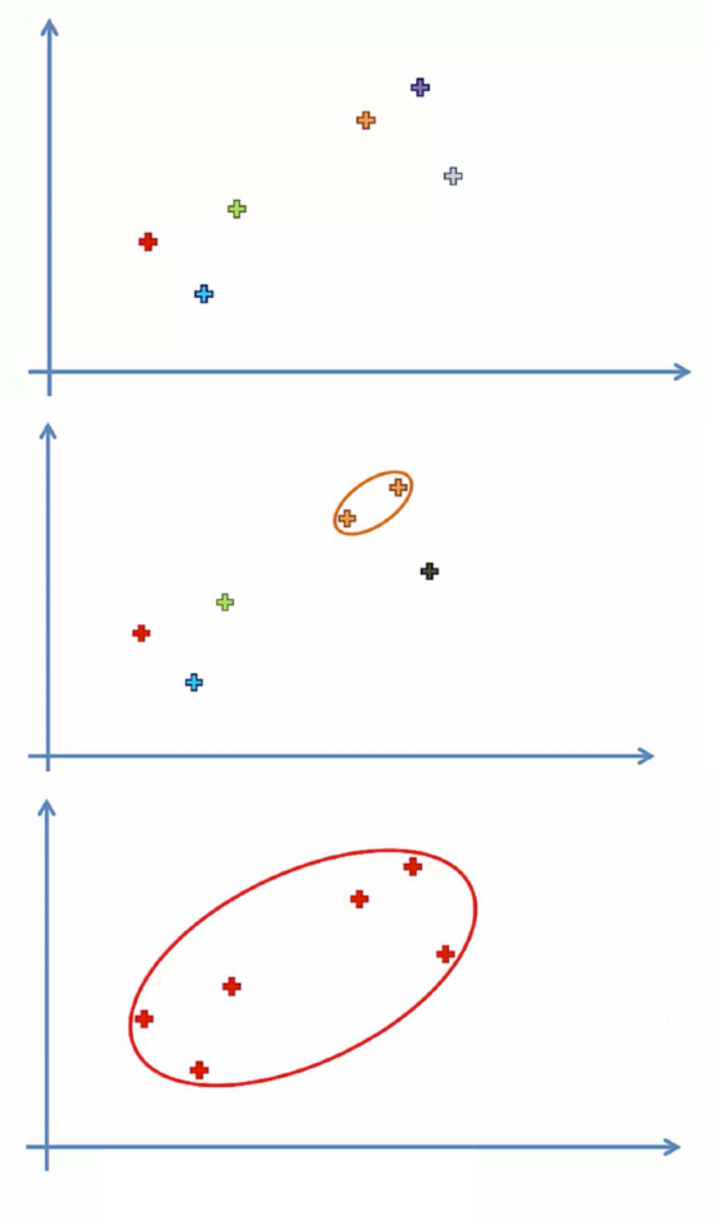
Issues…



- Bottom-up manner

1. Make each data point a cluster
2. Take the two closest clusters and make them one cluster
3. Repeat step 2 until there is only one cluster

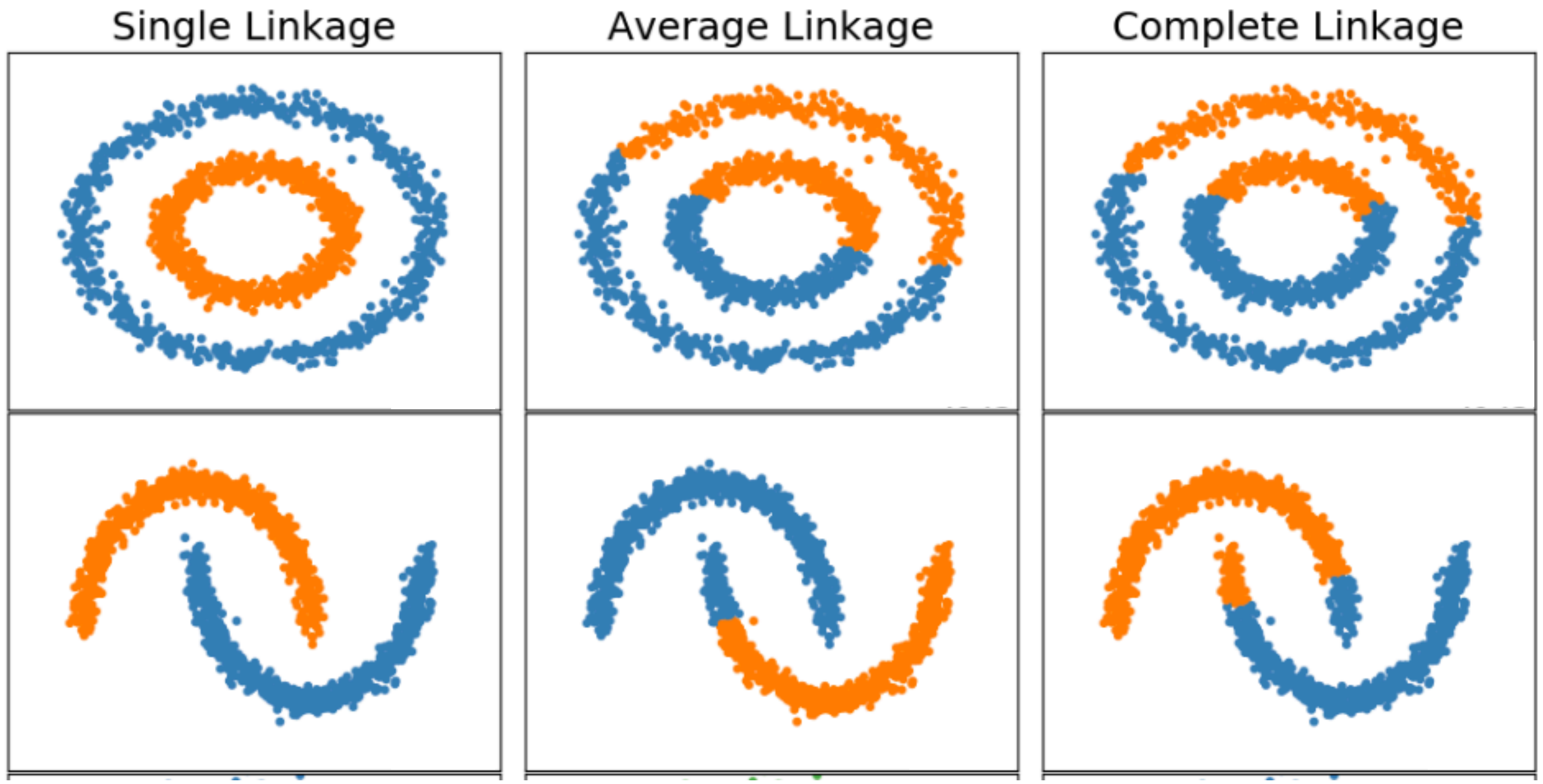




Did you see why it is a method of hierarchical clustering?

Agglomerative Clustering: Linkage Criteria

2. Take the two **closest clusters** and make them one cluster

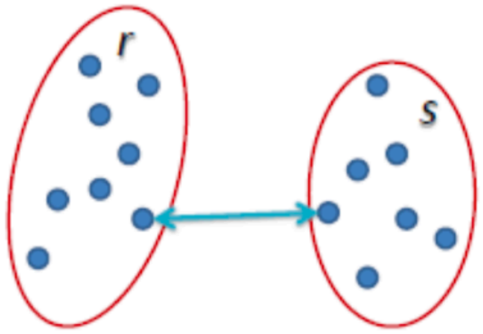


Single Linkage

Average Linkage

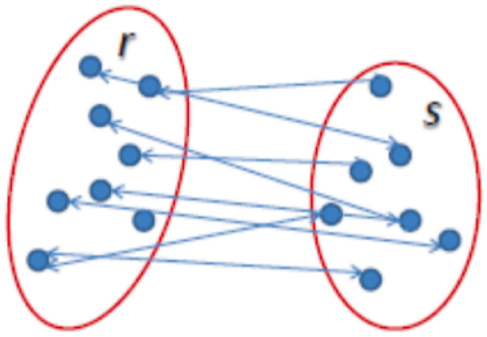
Complete Linkage

Single linkage



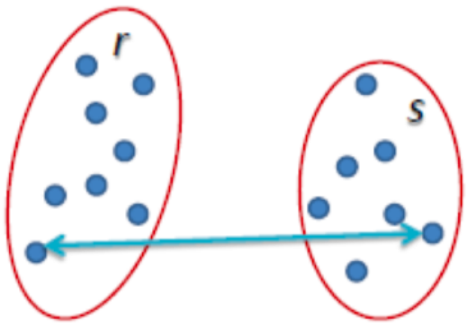
$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Average linkage



$$L(r,s) = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Complete linkage

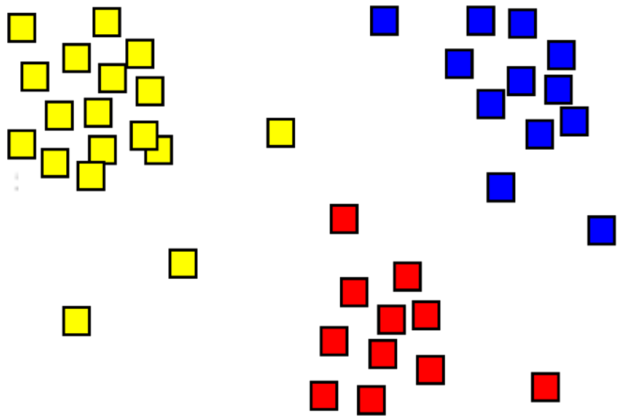


$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

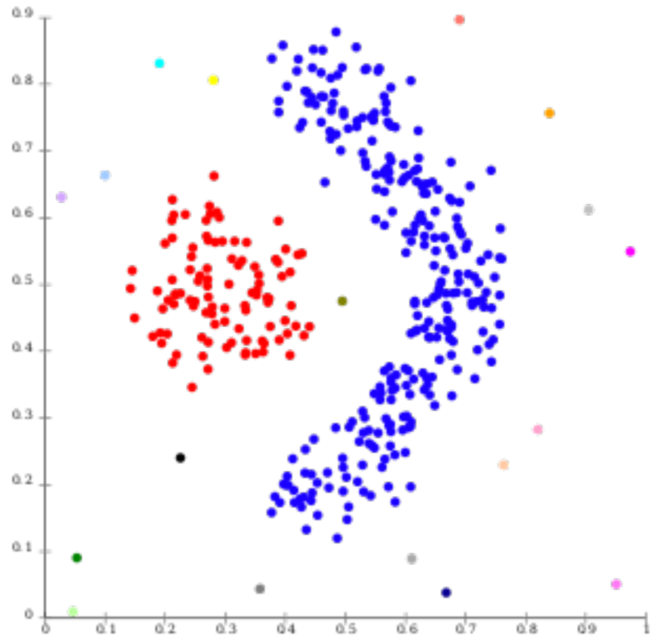
Unsupervised Learning 1: Summary

- Draw inferences from datasets consisting of input data without labeled responses
- Visualization, understanding…
- Clustering
  - Centroid-based clustering
  - Distribution-based clustering
  - Connectivity-based clustering…

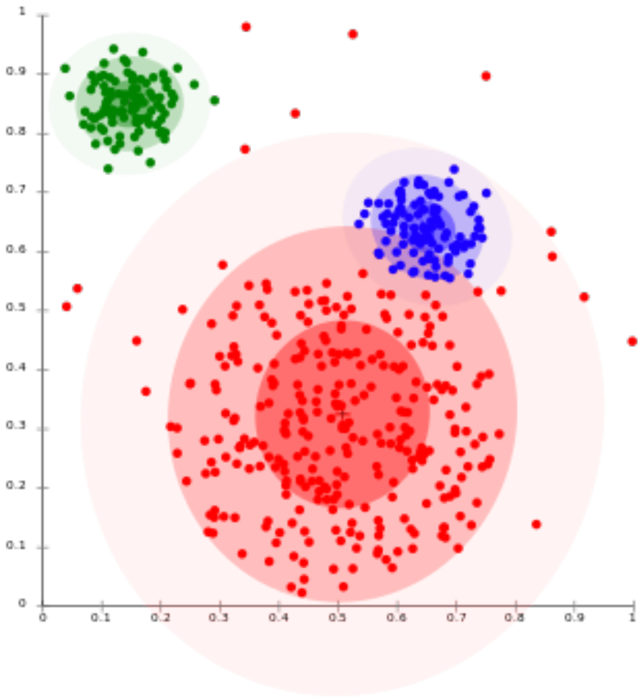
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$



Single-linkage clustering example

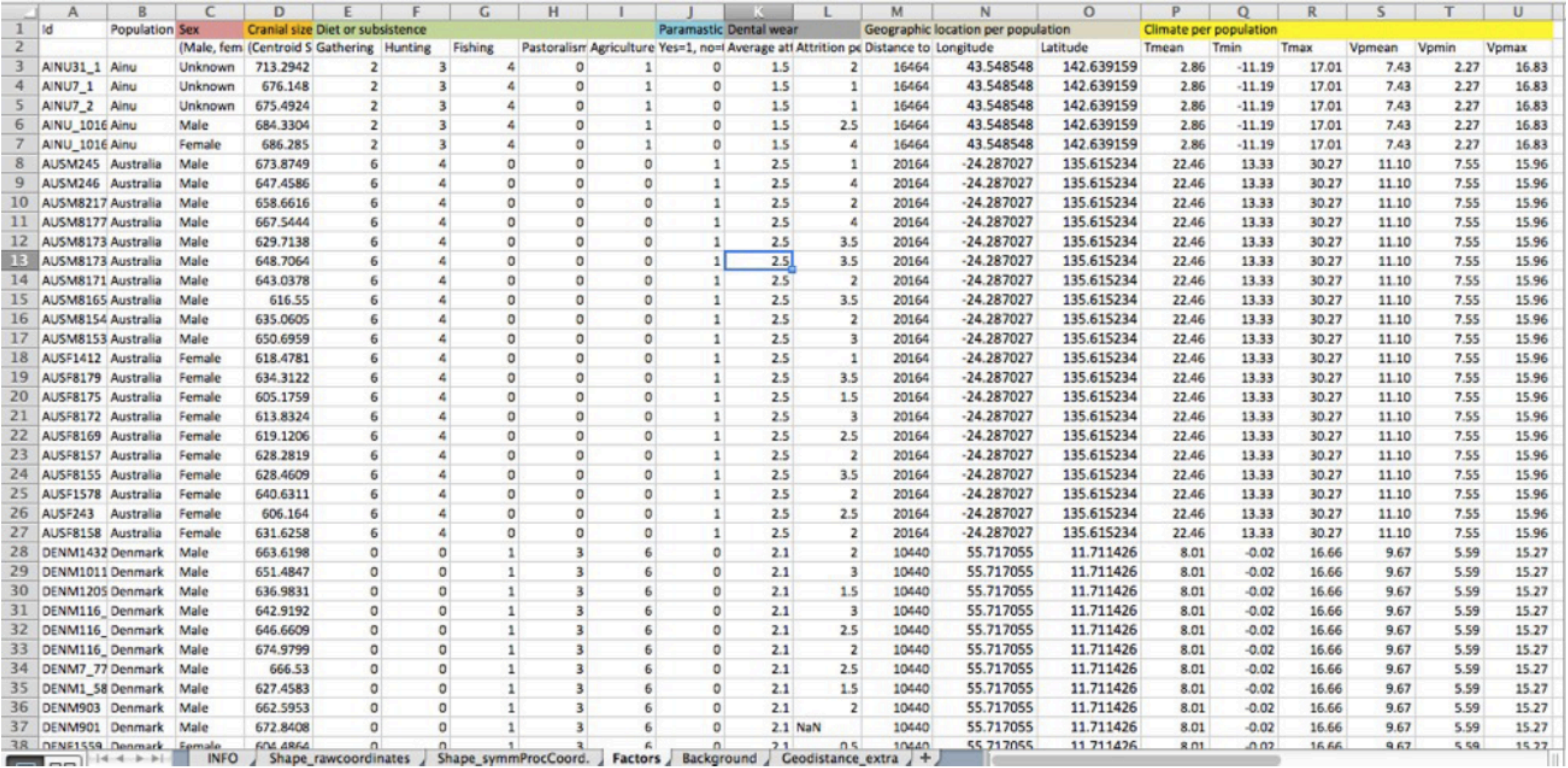


Gaussian mixture model clustering example



Two Ways of Finding Simpler Data Representations

- Fewer “data points” vs. ***fewer dimensions (#variables)***?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	Population	Sex	Cranial size	Diet or subsistence					Paramastic	Dental wear		Geographic location per population			Climate per population					
2			(Male, fem	(Centroid S	Gathering	Hunting	Fishing	Pastoralism	Agriculture	Yes=1, no=	Average att	Attrition pe	Distance to	Longitude	Latitude	Tmean	Tmin	Tmax	Vpmean	Vpmin	Vpmax
3	AINU31_1	Ainu	Unknown	713.2942	2	3	4	0	1	0	1.5	2	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
4	AINU7_1	Ainu	Unknown	676.148	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
5	AINU7_2	Ainu	Unknown	675.4924	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
6	AINU_1016	Ainu	Male	684.3304	2	3	4	0	1	0	1.5	2.5	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
7	AINU_1016	Ainu	Female	686.285	2	3	4	0	1	0	1.5	4	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
8	AUSM245	Australia	Male	673.8749	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
9	AUSM246	Australia	Male	647.4586	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
10	AUSM8217	Australia	Male	658.6616	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
11	AUSM8177	Australia	Male	667.5444	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
12	AUSM8173	Australia	Male	629.7138	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
13	AUSM8173	Australia	Male	648.7064	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
14	AUSM8171	Australia	Male	643.0378	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
15	AUSM8165	Australia	Male	616.55	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
16	AUSM8154	Australia	Male	635.0605	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
17	AUSM8153	Australia	Male	650.6959	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
18	AUSF1412	Australia	Female	618.4781	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
19	AUSF8179	Australia	Female	634.3122	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
20	AUSF8175	Australia	Female	605.1759	6	4	0	0	0	1	2.5	1.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
21	AUSF8172	Australia	Female	613.8324	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
22	AUSF8169	Australia	Female	619.1206	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
23	AUSF8157	Australia	Female	628.2819	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
24	AUSF8155	Australia	Female	628.4609	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
25	AUSF1578	Australia	Female	640.6311	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
26	AUSF243	Australia	Female	606.164	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
27	AUSF8158	Australia	Female	631.6258	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
28	DENM1432	Denmark	Male	663.6198	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
29	DENM1011	Denmark	Male	651.4847	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
30	DENM1205	Denmark	Male	636.9831	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
31	DENM116_	Denmark	Male	642.9192	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
32	DENM116_	Denmark	Male	646.6609	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
33	DENM116_	Denmark	Male	674.9799	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
34	DENM7_77	Denmark	Male	666.53	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
35	DENM1_58	Denmark	Male	627.4583	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
36	DENM903	Denmark	Male	662.5953	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
37	DENM901	Denmark	Male	672.8408	0	0	1	3	6	0	2.1	NaN	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
38	DENF1559	Denmark	Female	604.4864	0	0	1	3	6	0	2.1	0.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27

Next: multivariate analysis & its connection to causal analysis