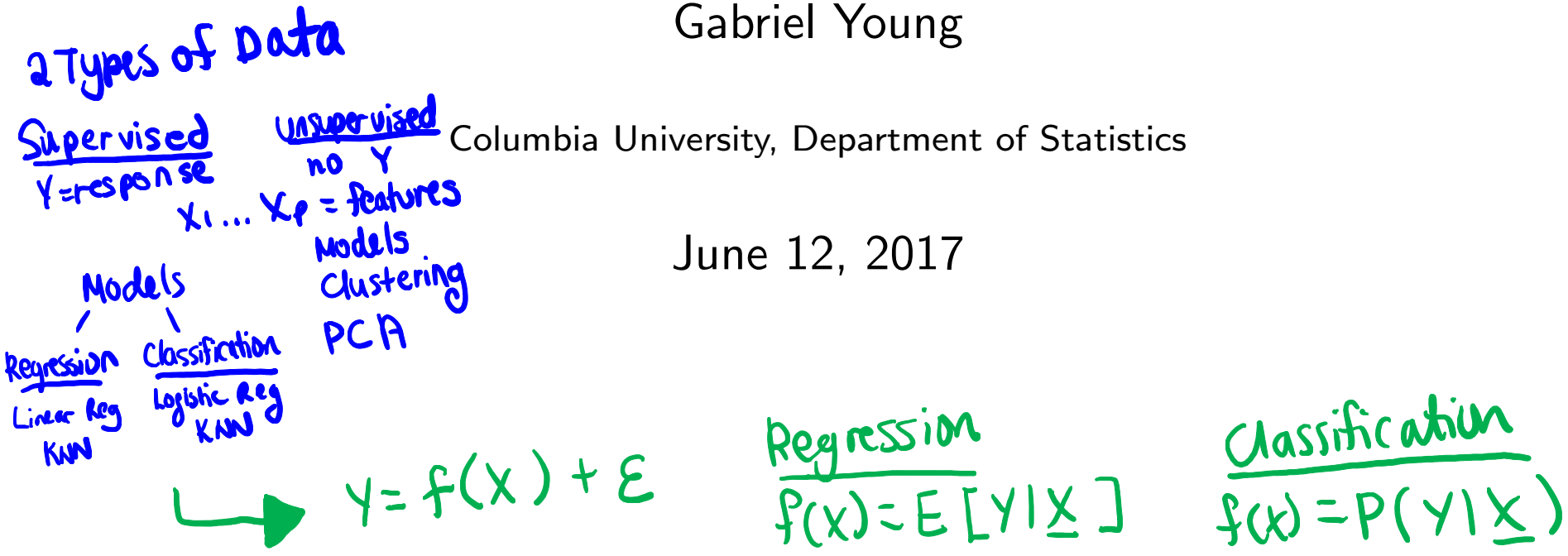
GU4206/5206
Supervised Learning

Gabriel Young

Columbia University, Department of Statistics

June 12, 2017



Dimensionality Reduction

Summary of last two classes:

- Have high dimensional continuous data
- Want to find low dimensional representation
- Principal Components Analysis (PCA) finds best K-dimensional rotation of data
- Gives a parsimonious description of the data

...but preprocessing **for what?**

Data Questions

Some questions are purely about prediction:

- Can I predict the yield of a crop next year (based on historicals, weather, ...)?
- Can I predict a patient outcome (based on condition, behavior, ...)?

Note that we naturally say things like "as a function of historicals, weather, ... "

We can wait and see what actually happens and then measure how well we did.

Data Questions

Some questions are about relationships between inputs and outputs:

- Are average crop yields increasing over time?
- Is there a negative health implication of doing z?

We need a model that tells us about the relationship between these values and explains the data well.

These patterns can be useful for directing our actions.

Data Questions

Some questions are about summarizing the data:

- What are the basic patterns that make up patients?
- Which patient types are most similar?

We need a model that summarizes the data well.

These patterns can be useful for directing further investigation.

Supervised vs. Unsupervised Learning

The questions fall into two categories: **supervised learning** and *unsupervised learning*.

Supervised Learning

- Have access to a set of p predictors $X_1, X_2, ..., X_p$ and a response Y both measured on the same n observations.
- The goal is to predict Y using $X_1, X_2, ..., X_p$ (usually by learning β parameters of a model).

Unsupervised Learning

- *Only* have access to a set of d predictors $X_1, X_2, ..., X_p$ measured on n observations.
- We are not interested in prediction, because we do not have an associated response variable Y.
- The goal is to discover interesting patterns about the measurements on the predictors $X_1, X_2, ..., X_p$.

Supervised vs. Unsupervised Learning

The questions fall into two categories: **supervised learning** and *unsupervised learning*.

Supervised learning:

- Predicting an output
- Understanding the relationship between an input and an output

Unsupervised learning:

- Summarizing the data
- Understanding underlying (hidden) factors

PCA is unsupervised learning. We will be studying supervised learning for most of UN3106.

Supervised Learning

Today we are learning about concepts that will be used throughout this course:

- Types of variables
- Estimating a function
- Prediction and inference
- Parametric and nonparametric estimation
- Interpretability

Supervised Learning

What is supervised learning? Let's start with an example.

- Let's look at the Advertising data set.
- In the data:
  - `Sales` for each of 200 products
  - Advertising budgets for each product in `TV`, `radio` and `newspaper`

If you have a new product, how should you spend your advertising money to generate the most sales? How much? In which media?

How can we use data to answer these questions?

Supervised Learning: Advertising

If you have a new product, how should you spend your advertising money to generate the most sales?

Idea: predict the level of `sales` from `TV`, `radio` and `newspaper`

Input variables (covariates, independent variables, predictors, features):

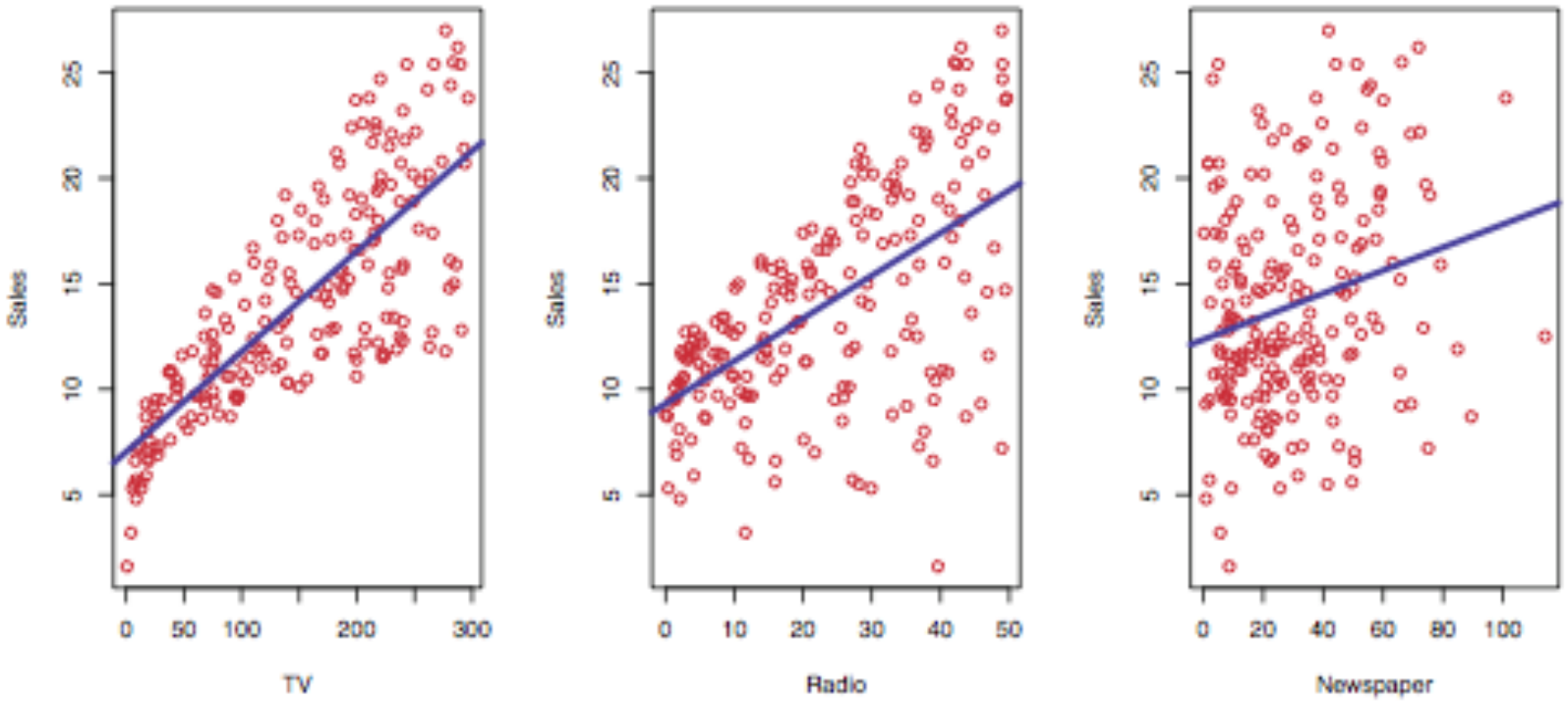
- `TV`
- `radio`
- `newpaper`

Output variable (response, dependent variable):

- `sales`

Supervised Learning: Advertising

If you have a new product, how should you spend your advertising money to generate the most sales?



Supervised Learning: Advertising

Formula:

$$\texttt{sales} = f\left(\texttt{TV}, \texttt{radio}, \texttt{newspaper}\right) + noise$$

We want to find f!

In general:

$X = (X_1, X_2, \ldots, X_p)$	inputs
Y	output
$Y = f(X) + \epsilon$	relationship

Supervised Learning: Regression and Classification

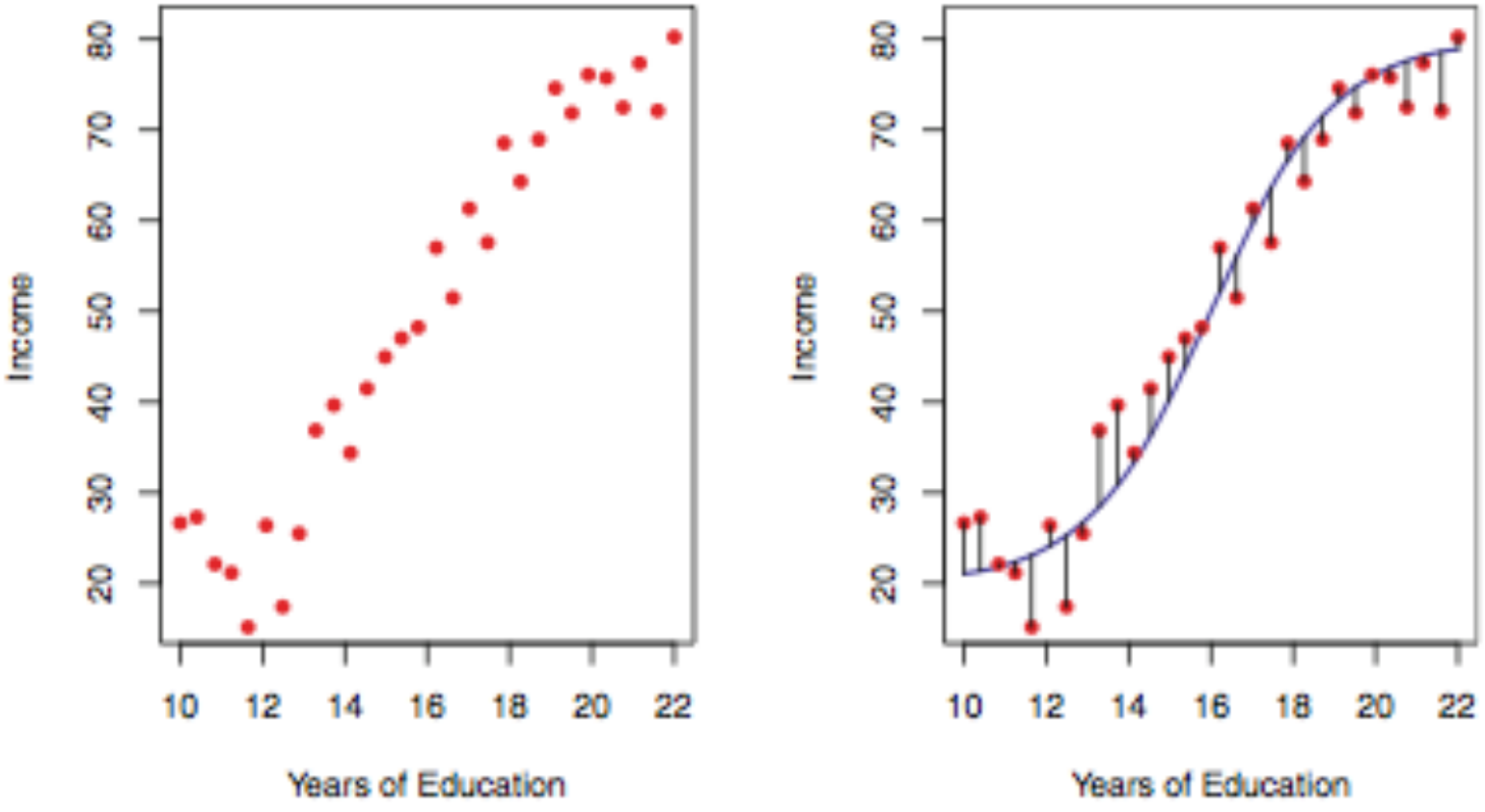
$$
\begin{aligned}
X &= (X_1, X_2, \ldots, X_p) && \text{inputs} \\
Y &&& \text{output} \\
Y &= f(X) + \epsilon && \text{relationship}
\end{aligned}
$$

We are interested in studying f in two settings:

- **Regression:** Y has continuous values, like \$81,200 or 72.
- **Classification:** Y has categorical values, like low/medium/high or red/green/blue

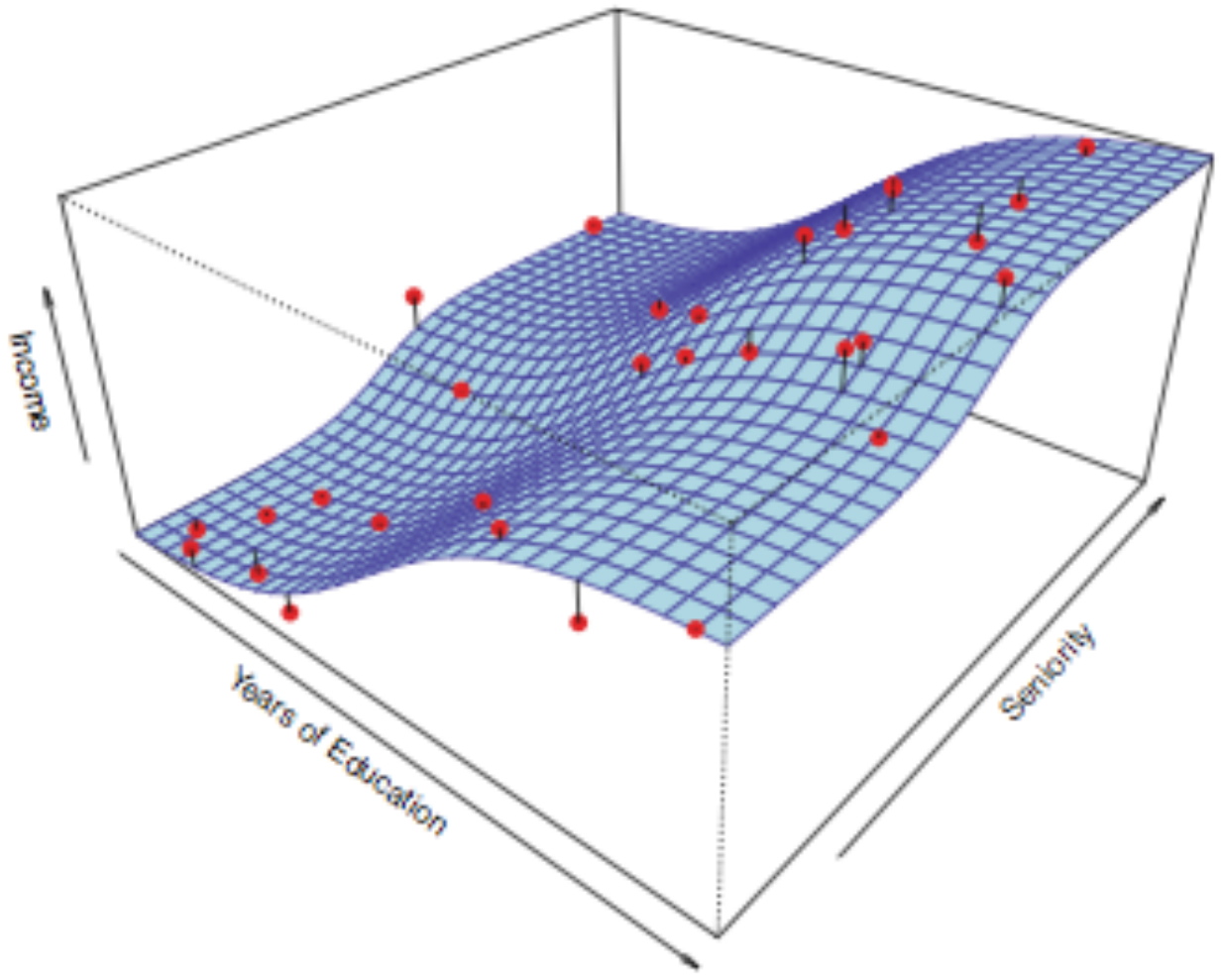
Supervised Learning: Income

Another example: we want to predict the annual income (`income`) of an individual based on years of education (`years of education`).



Supervised Learning: Income

Let's include seniority (`seniority`) as well.



Prediction and Inference

Why estimate f?

- Prediction
- Inference

Prediction:

- We have a new product with a set advertising budget `TV`, `radio` and `newspaper`). What will its sales be?
- Alice has 16 years of education and 0 years of seniority. What will her income be?

Goal: accurately estimate output for new inputs.

Prediction

In general, prediction is a two-step process:

1. Use data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to estimate f with $\hat{f}$
2. Feed new input X through $\hat{f}$ to get estimated output:

$$\hat{Y} = \hat{f}(X)$$

How accurate can we make $\hat{Y}$?

$$Y - \hat{Y} = \epsilon + f(X) - \hat{f}(X)$$

- irreducible error: ϵ
- reducible error: $f(X) - \hat{f}(X)$

Prediction

Generally, we measure our success by the expected **mean squared error** (MSE):

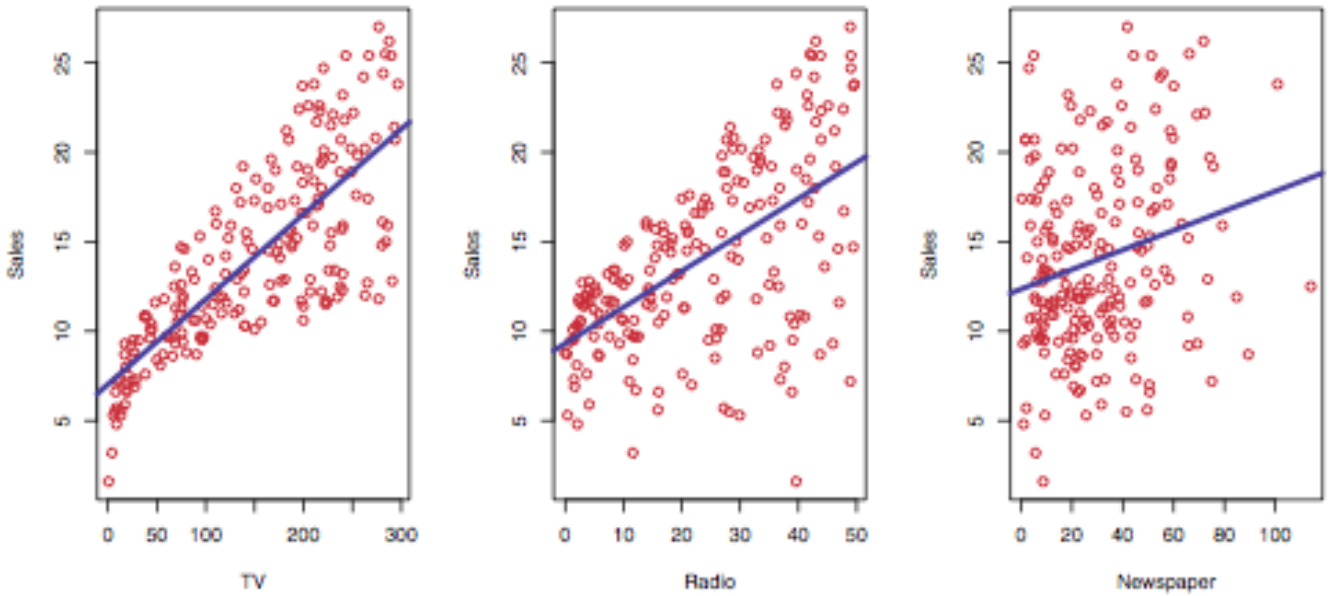
$$\mathbb{E}(Y - \hat{Y})^2$$

Fix both X and $\hat{f}$. What are the reducible and irreducible errors with the MSE?

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}\left[f(X) + \epsilon - \hat{f}(X)\right]^2 \\
&= \mathbb{E}\left[(f(X) - \hat{f}(X))^2 + 2\epsilon(f(X) - \hat{f}(X)) + \epsilon^2\right] \\
&= (f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\mathbb{E}[\epsilon] + \mathbb{E}[\epsilon^2] \\
&= (f(X) - \hat{f}(X))^2 + \mathrm{Var}(\epsilon)
\end{aligned}$$

Inference

If you have a new product, how should you spend your advertising money to generate the most sales?



Inference Questions:

Which media contribute to sales? Which gives the biggest boost?
If I spend more on advertising, how much should sales increase?

Inference

Inference

We want to learn about relationships between inputs and outputs:

- How will increasing one input affect the output?
- Is a specific combination of inputs associated with an increase in the output?

Can you think of some inference questions? Prediction questions? What is the difference between the two?

Fitting f

Data:

- Suppose we have n observations, $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$.

We now have:

- A question that I want to answer (**prediction** or **inference**)
- Data

Estimation:

- To answer my question, I need to estimate the relationship

$$Y = f(X) + \epsilon.$$

- How do we find $\hat{f}$ using $(x_1, y_1), \ldots, (x_n, y_n)$?

Fitting f

How do I find $\hat{f}$ using $(x_1, y_1), \ldots, (x_n, y_n)$?

1. select a statistical model
2. select the model parameters using the data

What types of statistical models are there?

- **Parametric:** described by a finite number of parameters, say

$$\beta_1, \beta_2, \ldots, \beta_d$$

- **Non-parametric:** not described by a finite number of parameters

Parametric Models

Parametric Models

A **parametric model** is a statistical model described by a finite number of parameters. Examples include:

- a Gaussian distribution ($N(\mu, \sigma^2)$)
- a Bernoulli distribution (Bern(p))
- a *linear model*

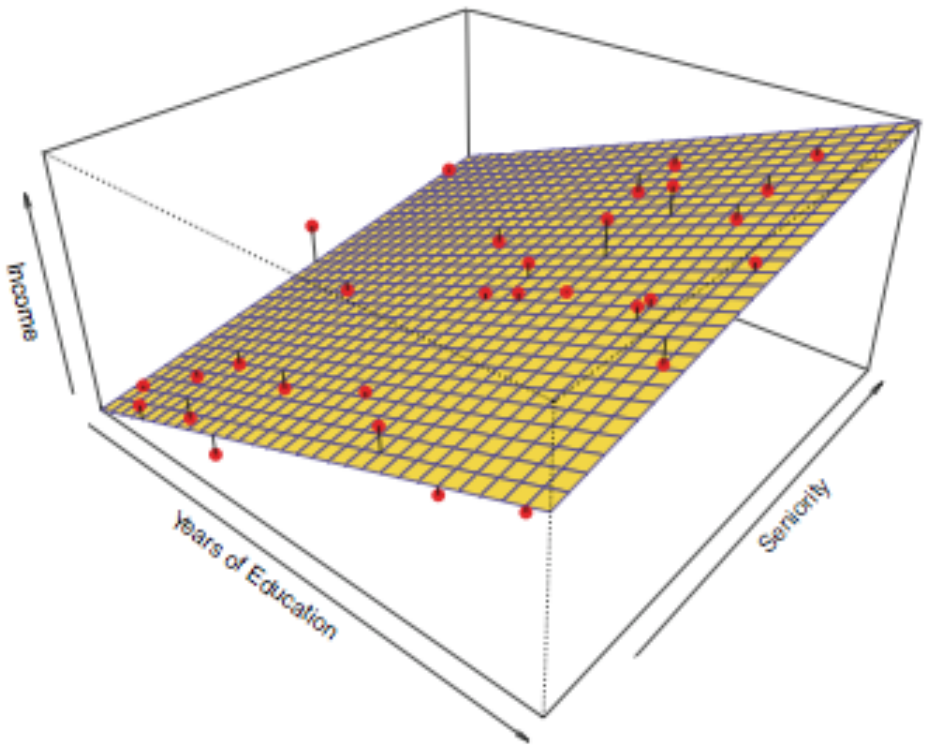
$$Y = \beta_0 + \beta_1 X_d + \cdots + \beta_d X_d + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Parameters

- What are the parameters of the Gaussian?
- What are the parameters of a linear model?

Parametric Models

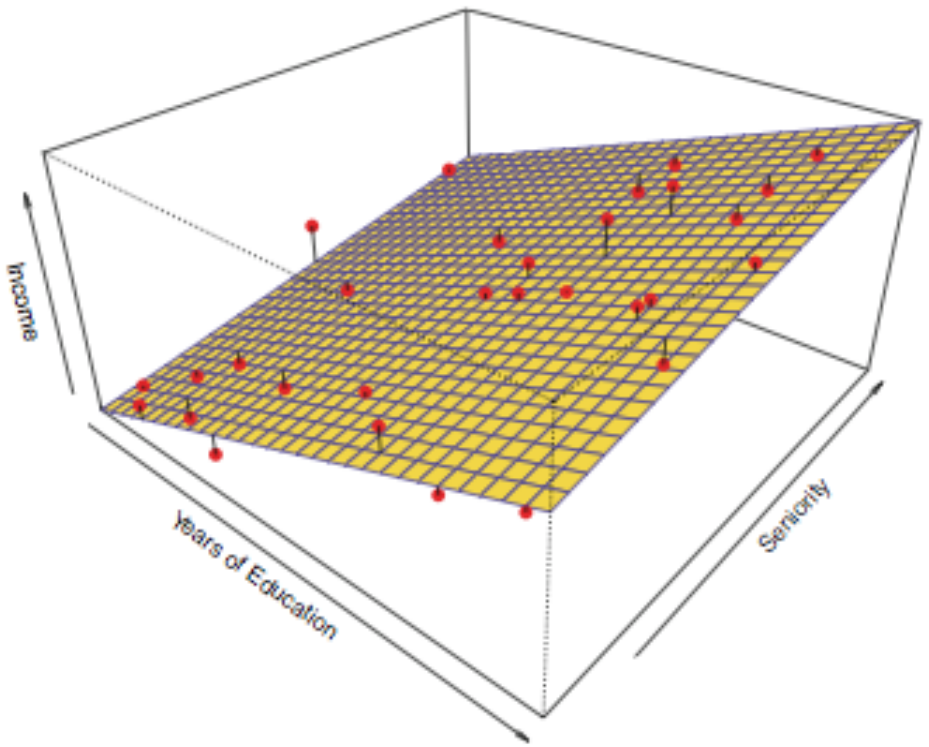
$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{years of education} + \beta_2 \times \texttt{seniority}$$



What does this model say about the structure of f?

Parametric Models

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{years of education} + \beta_2 \times \texttt{seniority}$$



Is this model good for prediction? What can it tell us for inference?

Parametric Models

Parametric Models:

- Few parameters (when is this good?)
- Well-described interactions between inputs, parameters and output (when is this useful?)
- Limited flexibility (desirable or undesirable?)

Nonparametric Models

Nonparametric Models:

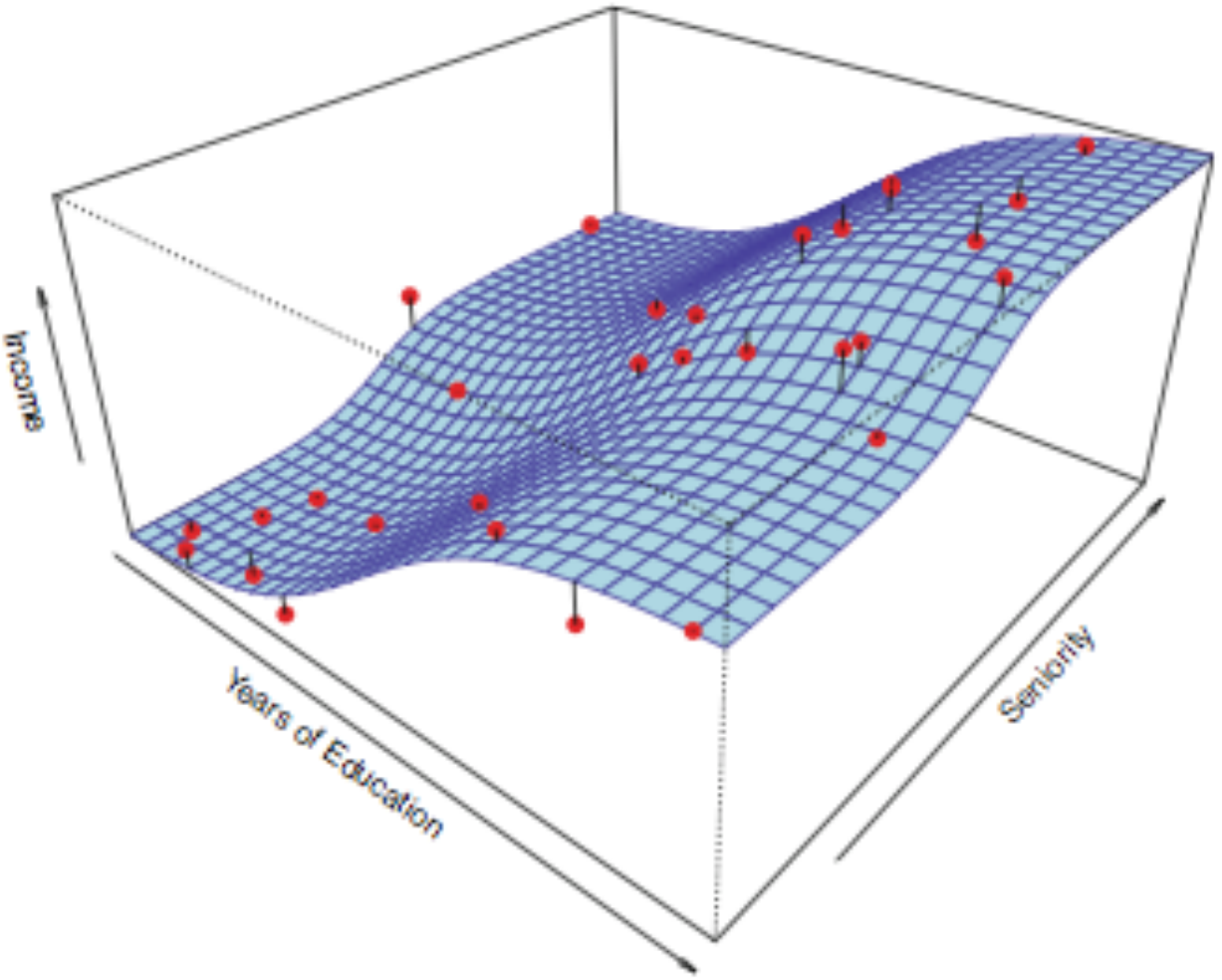
- Nonparametric models are not described by a finite number of parameters.
- So, what does that mean?
- Nonparametric models assume less about the population.
- In the model

$$Y = f(X) + \epsilon,$$

we let the data decide what the function f looks like.

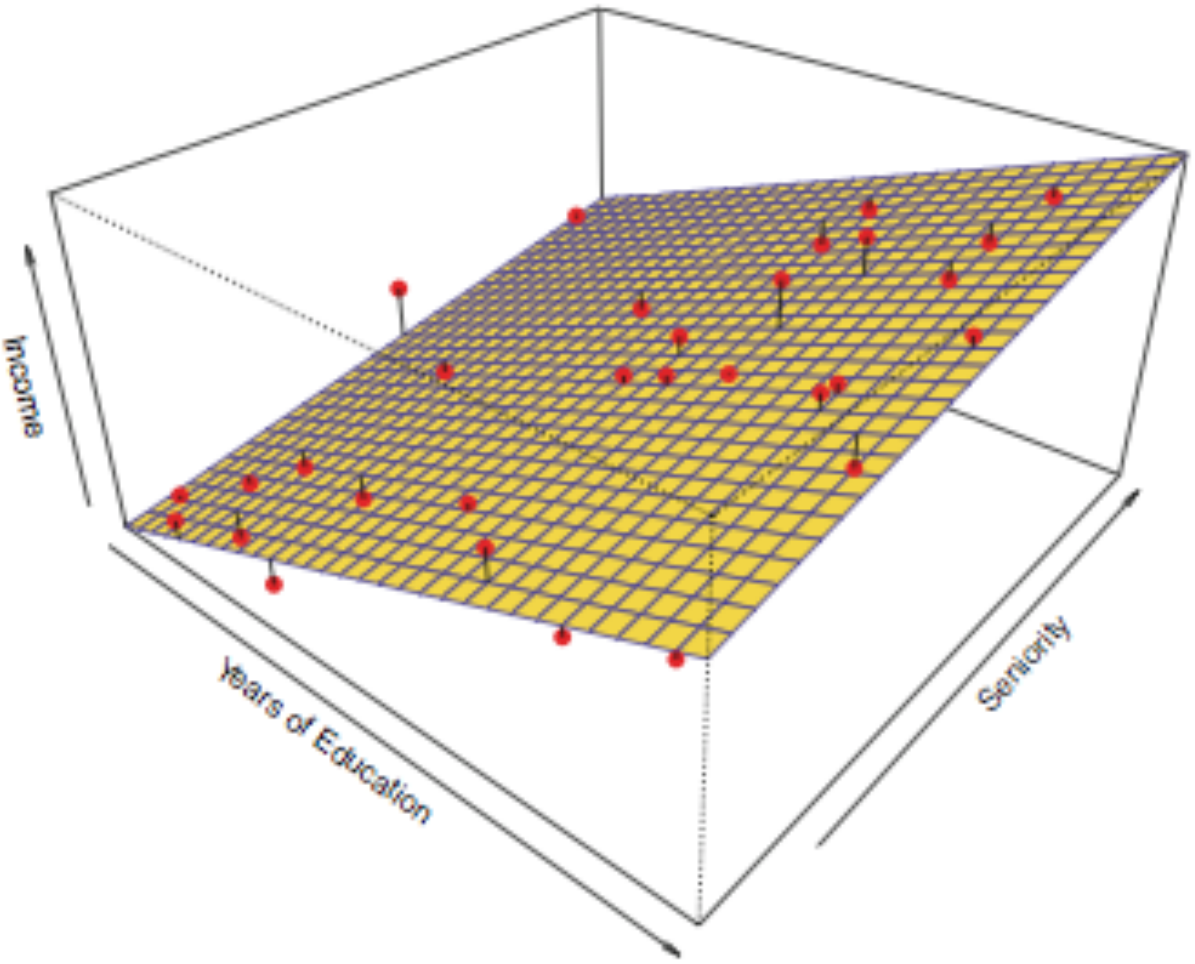
Non-Parametric Model

$$\texttt{income} \approx f(\texttt{years of education}, \texttt{seniority})$$



Parametric Model

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{years of education} + \beta_2 \times \texttt{seniority}$$



Problem Types (Generally)

	Continuous	**Categorical**
Supervised	Regression Parametric LS, MLE,... Nonparametric kNN,..	Classification Parametric logistic,... Nonparametric kNN,..
Unsupervised	Dimension Reduction PCA,..	Clustering k-means,..

$k-$Nearest Neighbors (kNN)

Non-Parametric Method: kNN

- Let's introduce another non-parametric method: $k-$nearest neighbors (kNN)
- We will discuss KNN in more detail next lecture .

Idea: average the values of the k closest observations

$$\hat{Y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

where $N_k(x)$ is the set of observations with the k smallest distances to the query point x.

Questions

Why is this a nonparametric model? Does it have parameters?

Is It a Good Day to Go for a Run?

k-Nearest Neighbors: Classification Example

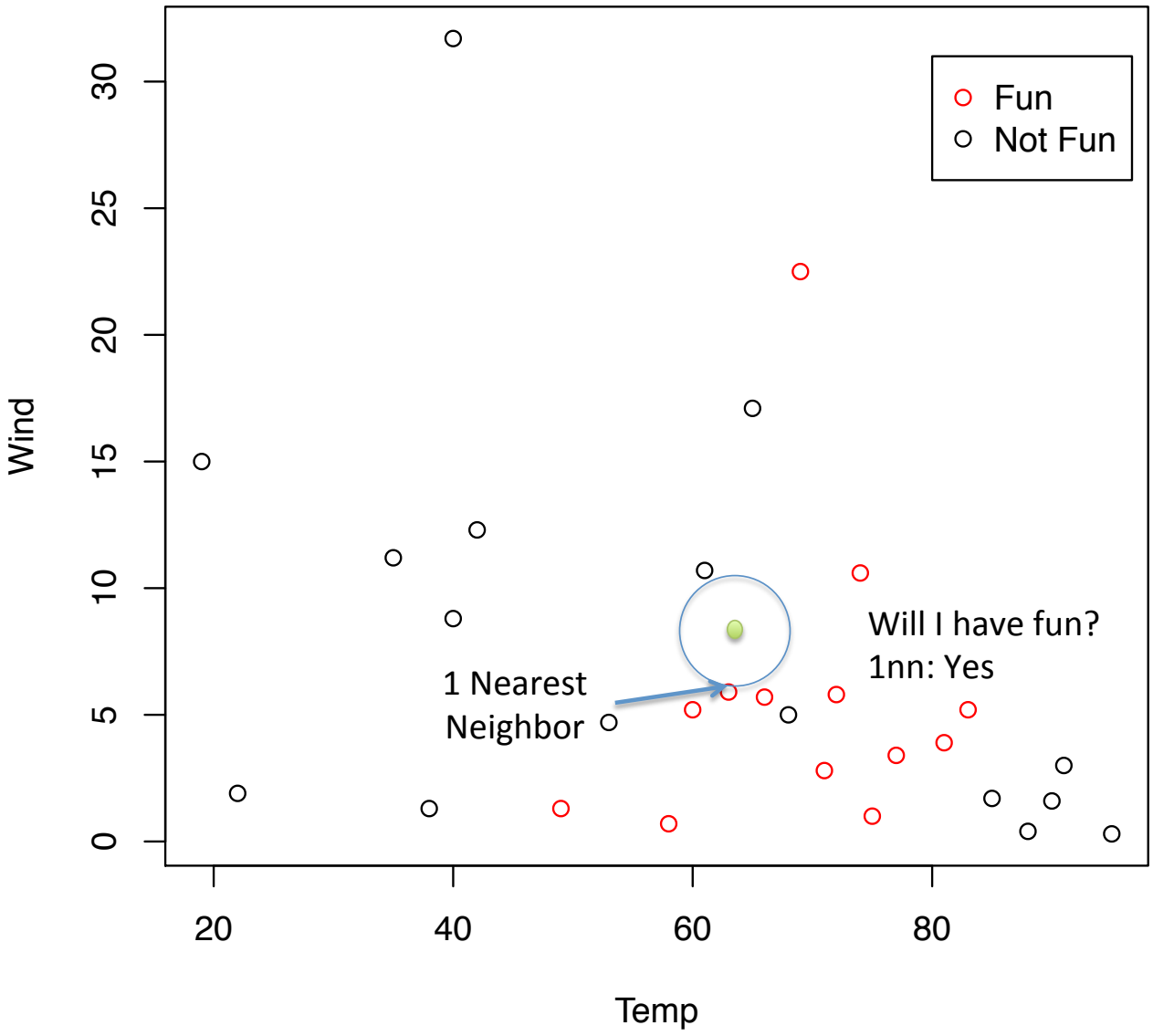
I have data on my past running. I recorded the temperature and whether the run was fun:

- Temperature (degrees F)
- Wind Speed (mph)
- Fun (yes, no)

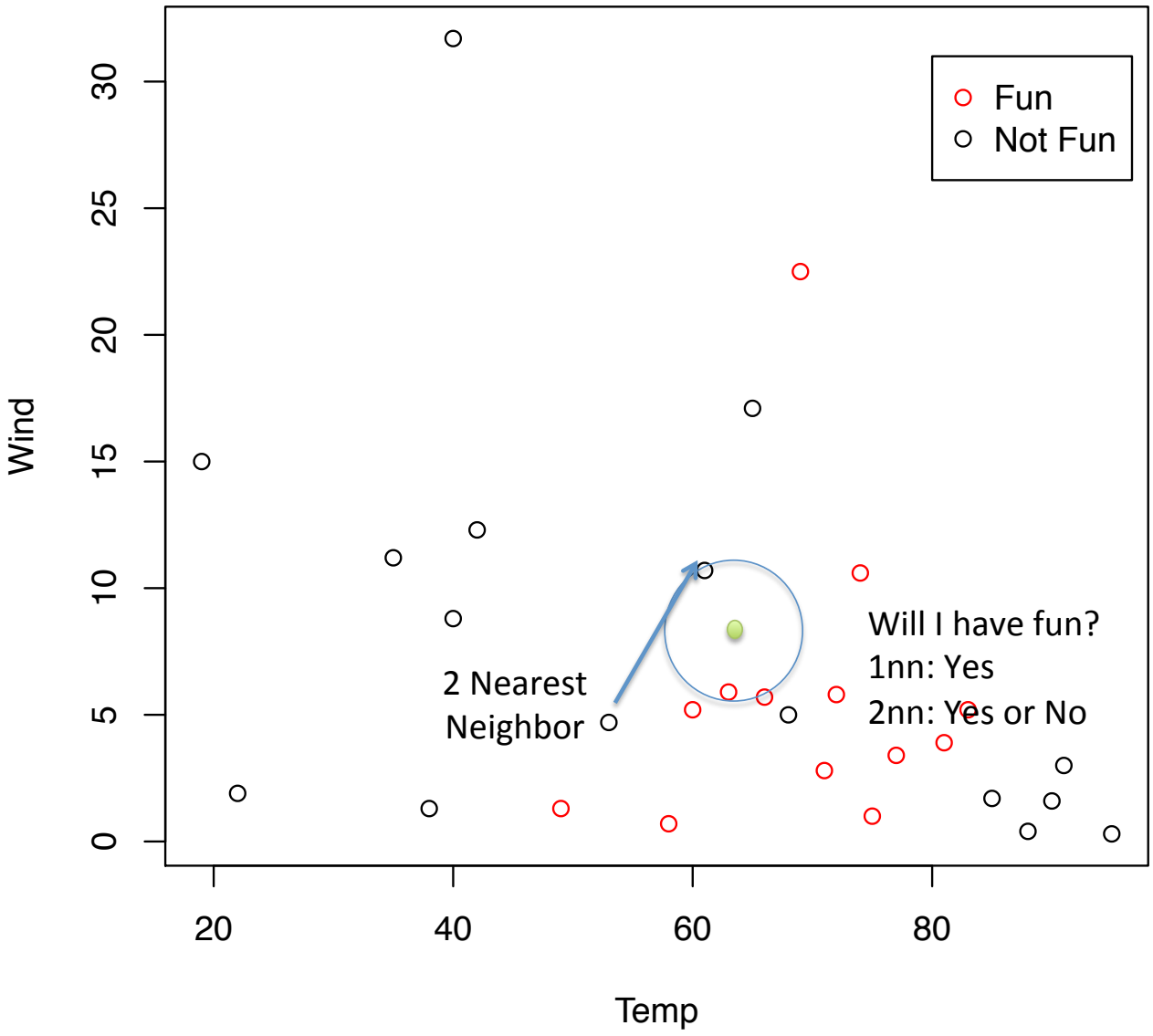
Goal

- It is now 65 degrees and the wind is 9 mph. (**input**)
- Will my run be fun? (**prediction**)

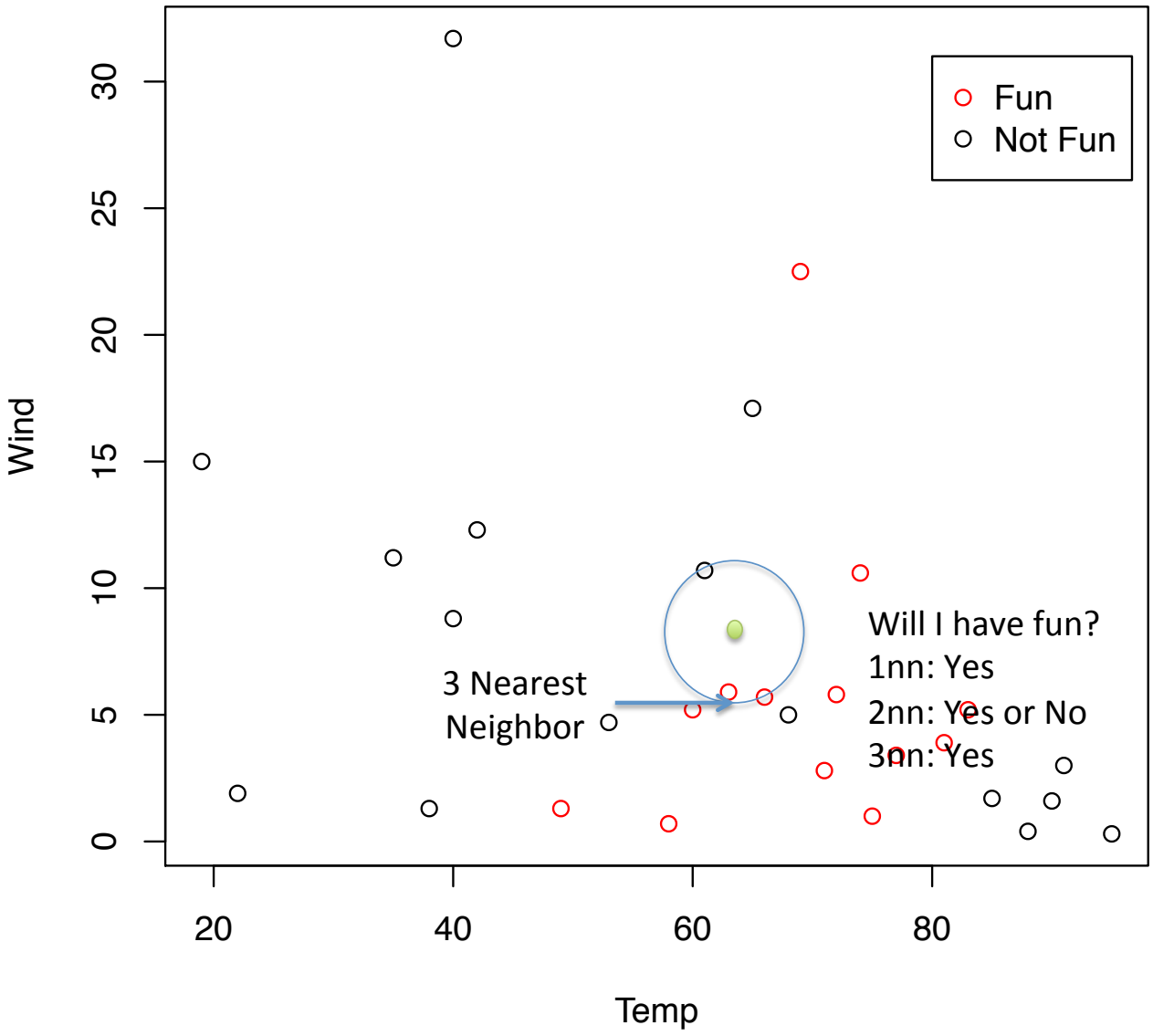
k-Nearest Neighbors: Classification



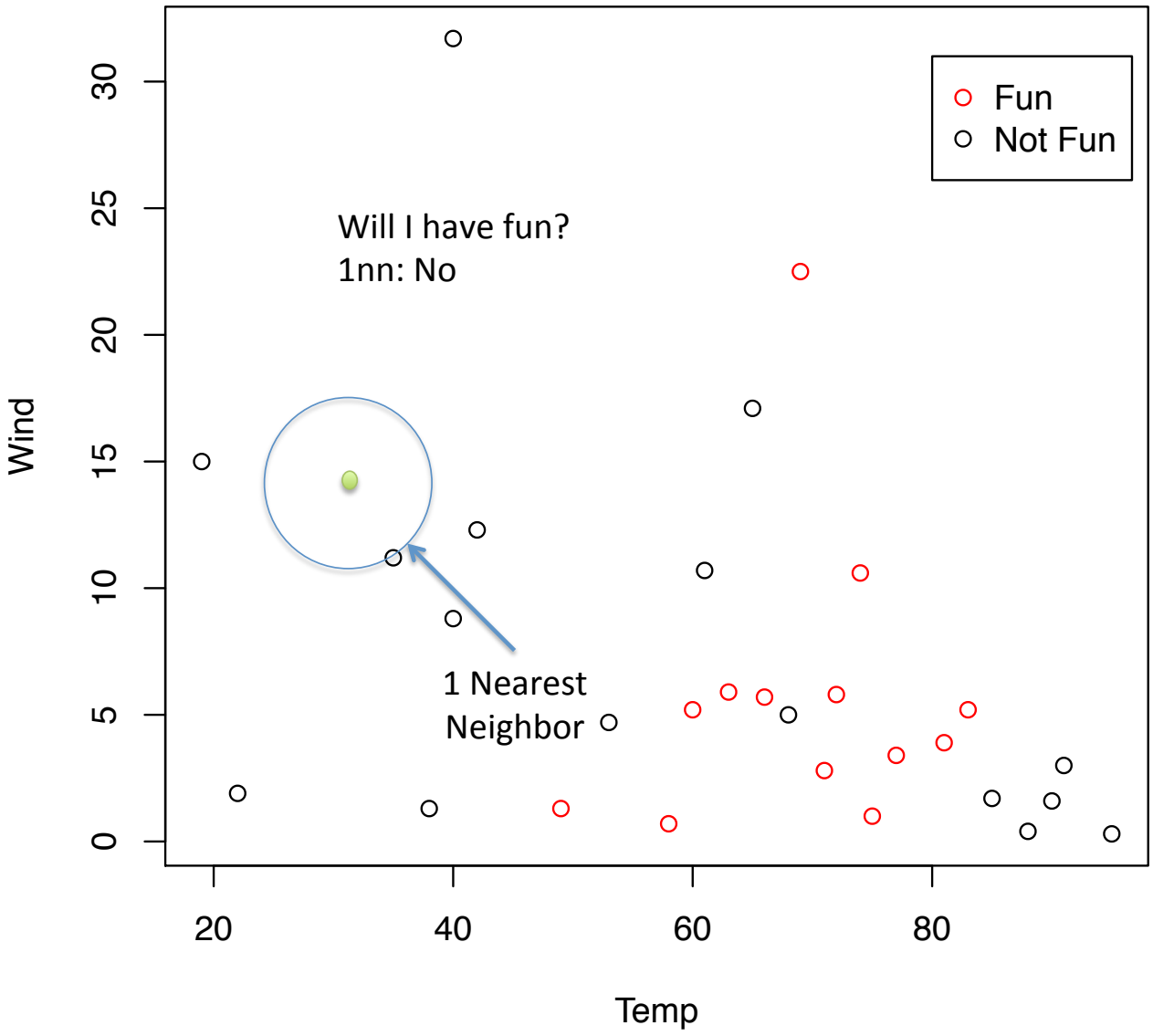
k-Nearest Neighbors: Classification



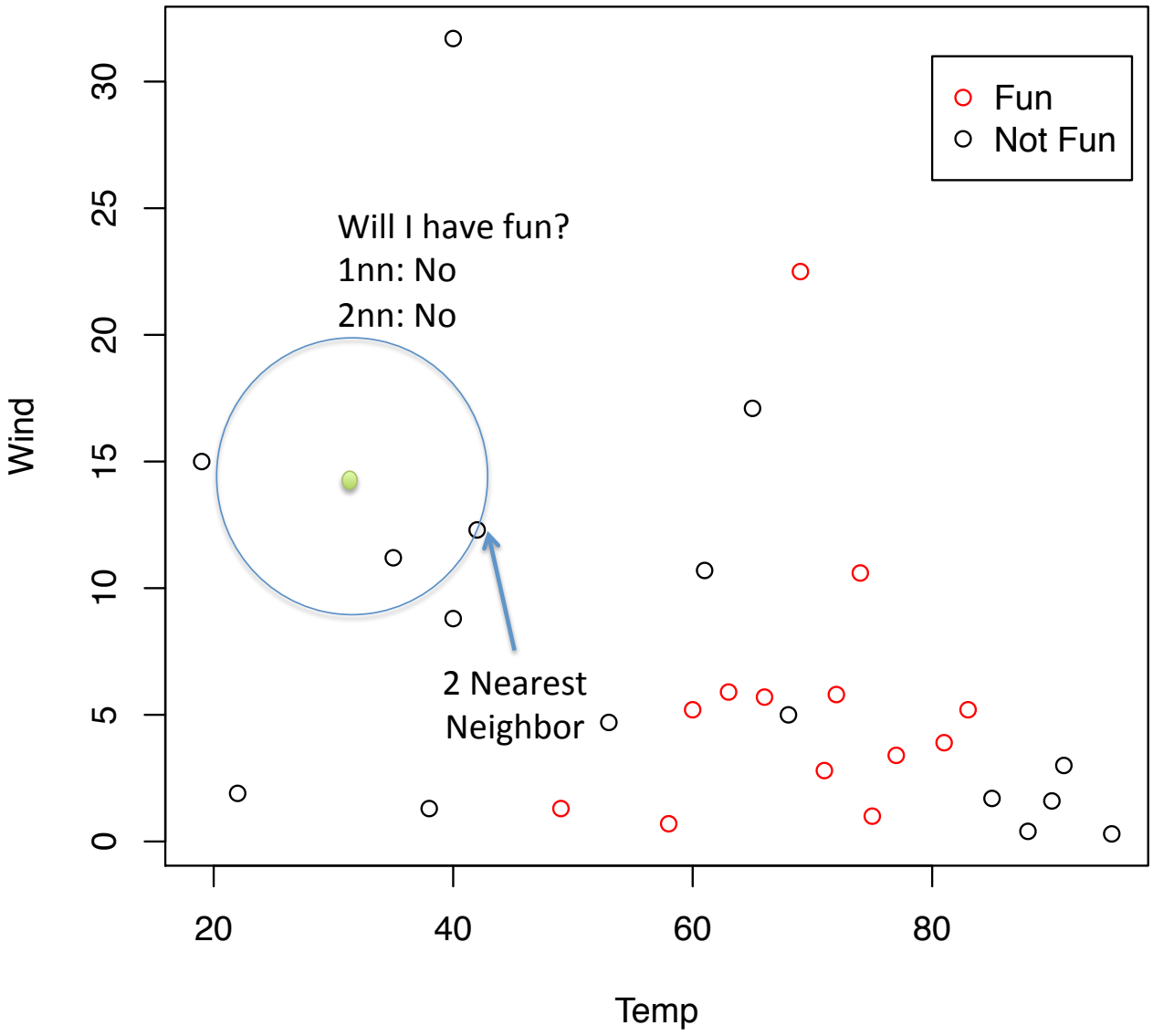
k-Nearest Neighbors: Classification



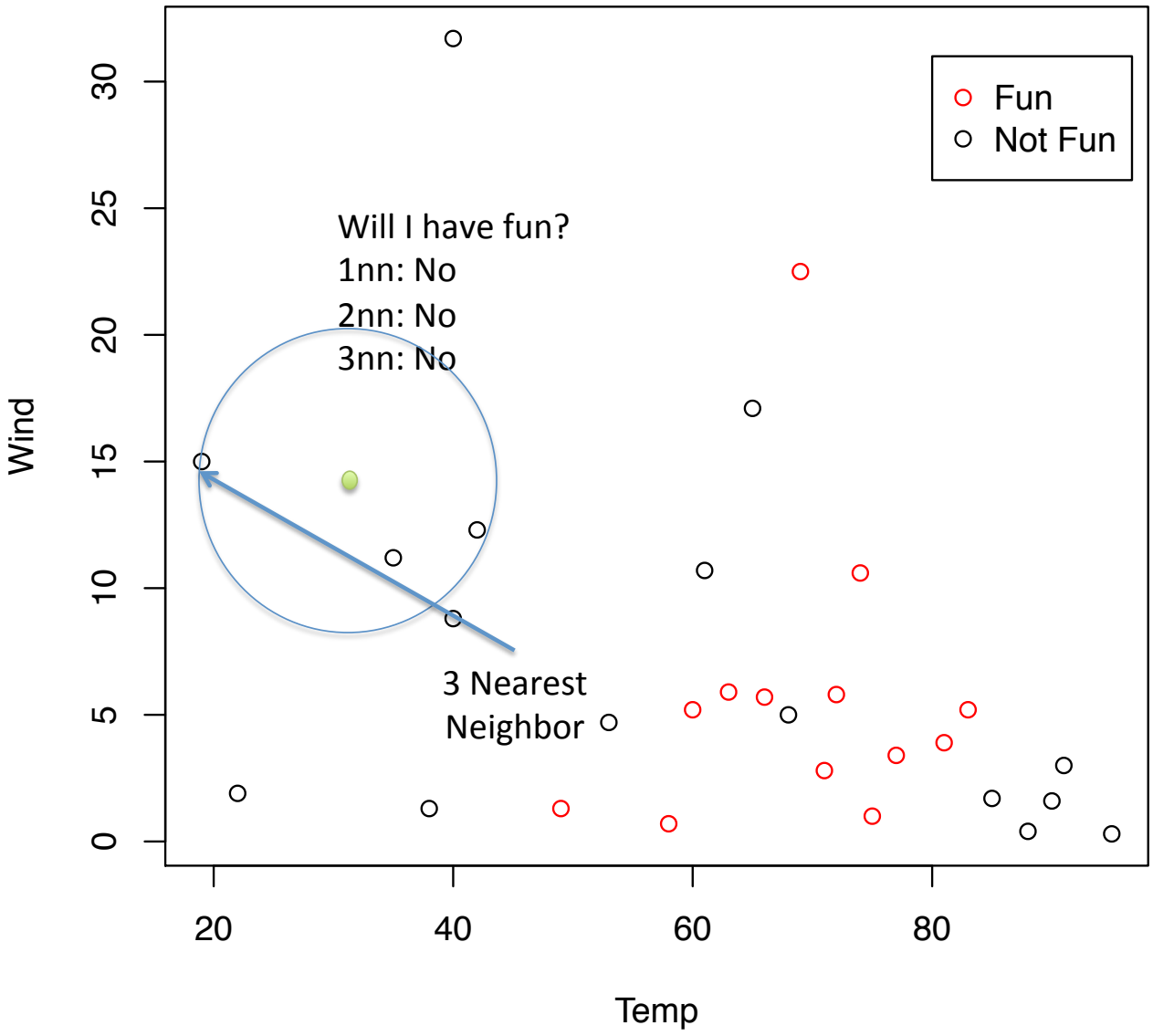
k-Nearest Neighbors: Classification



k-Nearest Neighbors: Classification



k-Nearest Neighbors: Classification



Congressional Approval

k-Nearest Neighbors: Regression Example

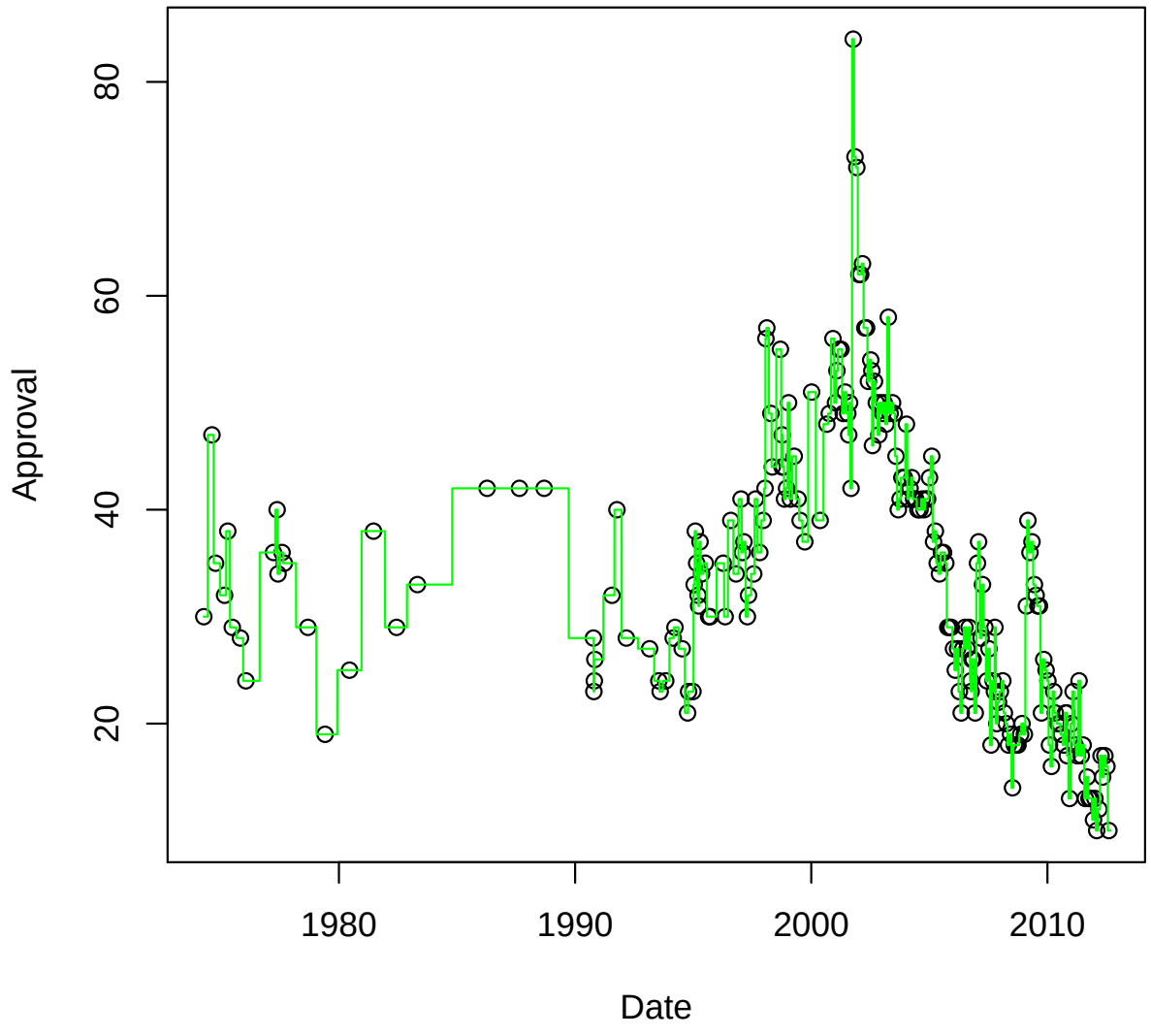
Consider a data set on congressional approval ratings recorded from 1974 to 2012.

Goals

- Estimate f based the data that predicts the **average** approval rating as a function of time.
- What was the approval in 1985?

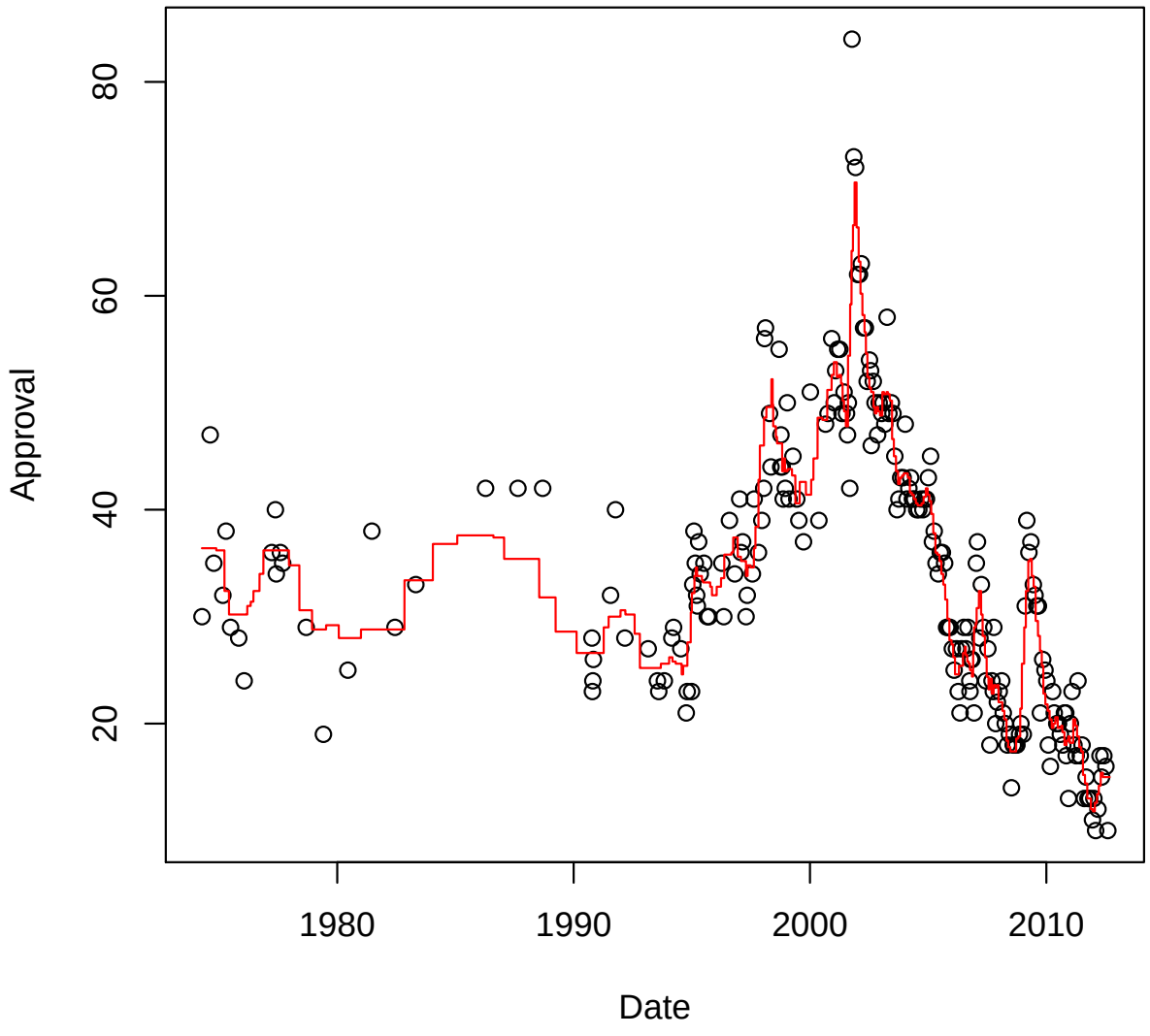
k-Nearest Neighbors: Regression

1-Nearest Neighbor: Congressional Approval



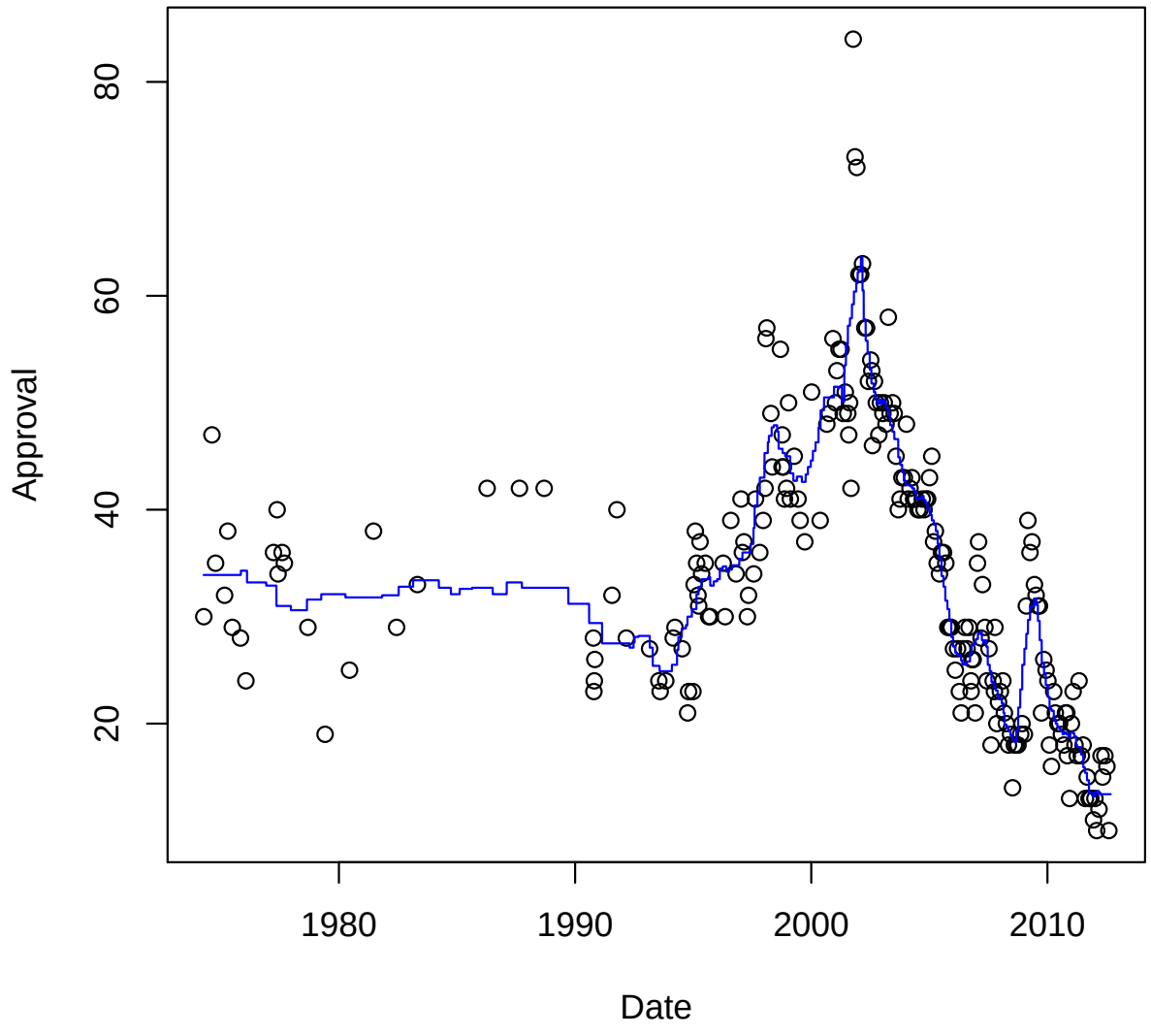
k-Nearest Neighbors: Regression

5-Nearest Neighbors: Congressional Approval



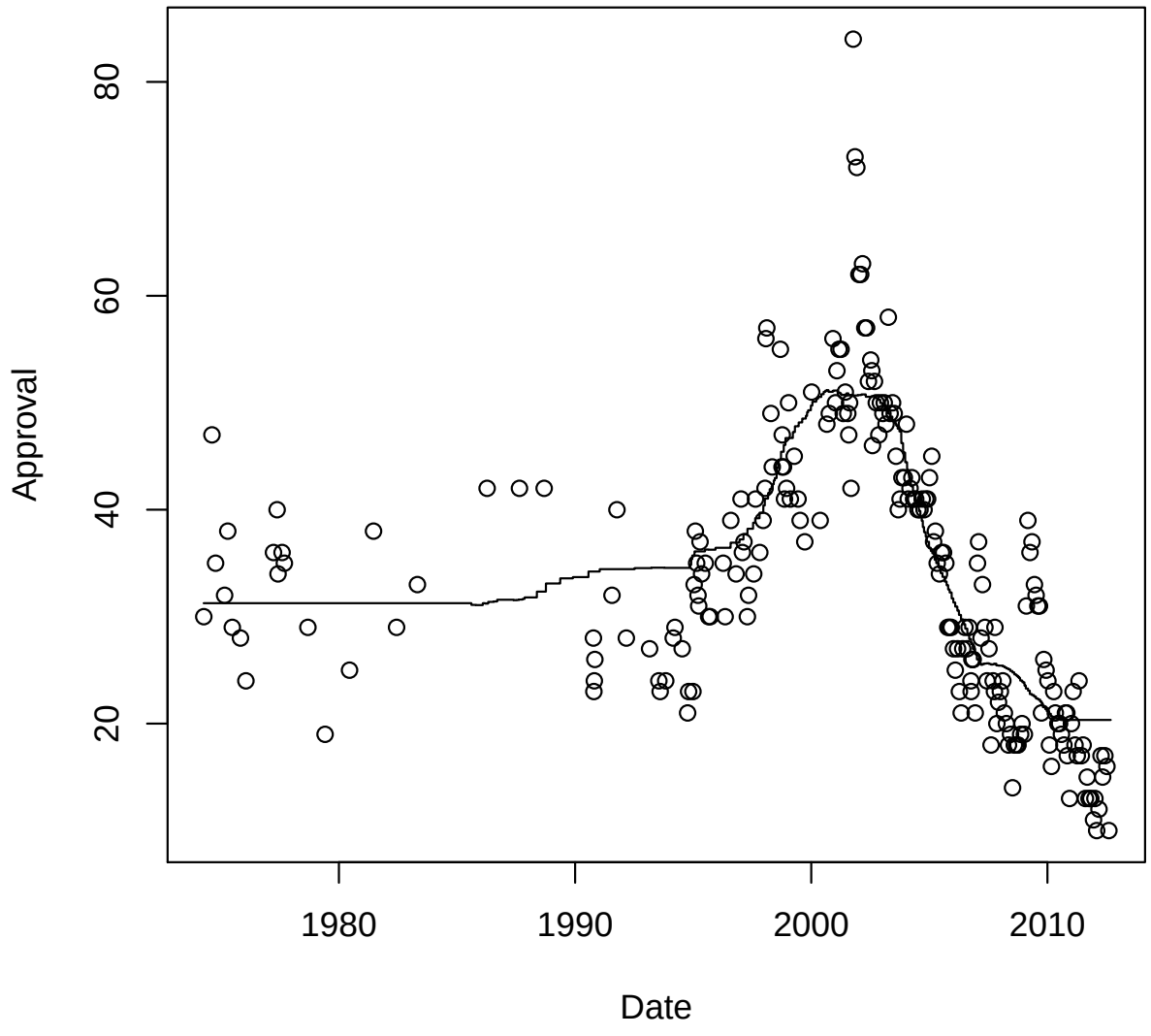
k-Nearest Neighbors: Regression

10-Nearest Neighbors: Congressional Approval



k-Nearest Neighbors: Regression

50-Nearest Neighbors: Congressional Approval

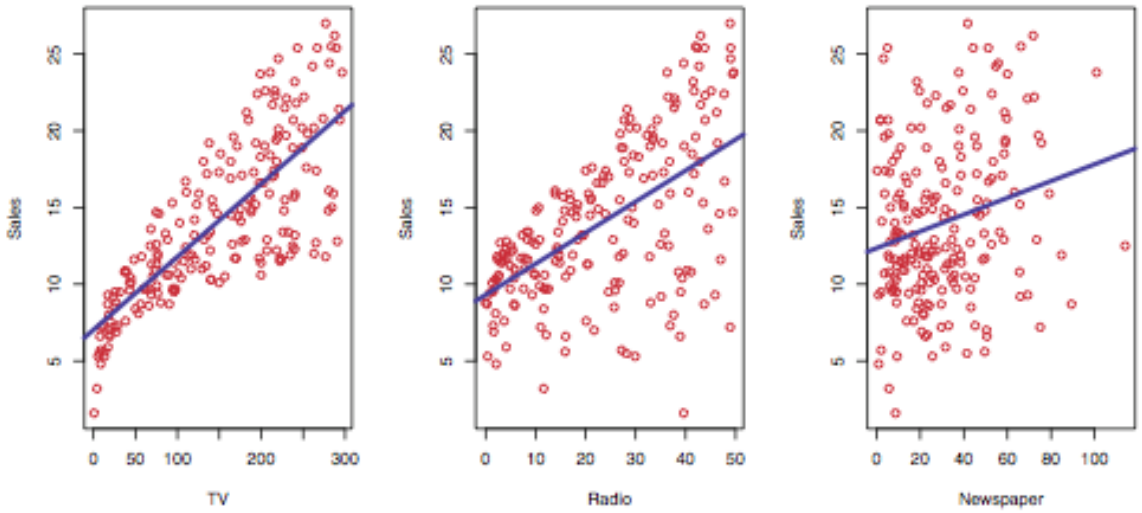


k-Nearest Neighbors: Regression

R Code

```
## Write kNN reg funciton
kNN.Regression <- function(x,data=data,k=5) {

  # Define distances
  D <- sqrt((x-data[,2])^2)

  # Find k-closest time points to x
  K.closest.ratings <- order(D)[1:k]

  # Return the mean of the k-corresponding response values (approval ratings)
  return(mean(data[K.closest.ratings,1]))

}

## Plot
K=10
x <- seq(from=min(data[,2]),to=max(data[,2]),by=.05)
plot(data[,2],data[,1],xlab="Date",ylab="Approval",main=paste("k =",K))
lines(x,sapply(x,kNN.Regression,data=data,k=K),col="purple")
```

Nonparametric vs Parametric Models



Let's go back to `Advertising`. Suppose that we fit it with a linear model and kNN.

- Which model will produce a more accurate prediction? with a lot of data? With a little data?
- Which model will tell us about which media will produce the best return?
- Which model will be faster to evaluate when we have a lot of data (n is large)?

Interpretability

Flexibility

Nonparametric models are generally more *flexible* than parametric models.

Why would we ever want a more restrictive model?

- Prevent overfitting
- Data compression
- **Interpretability:** model parameters often mean something

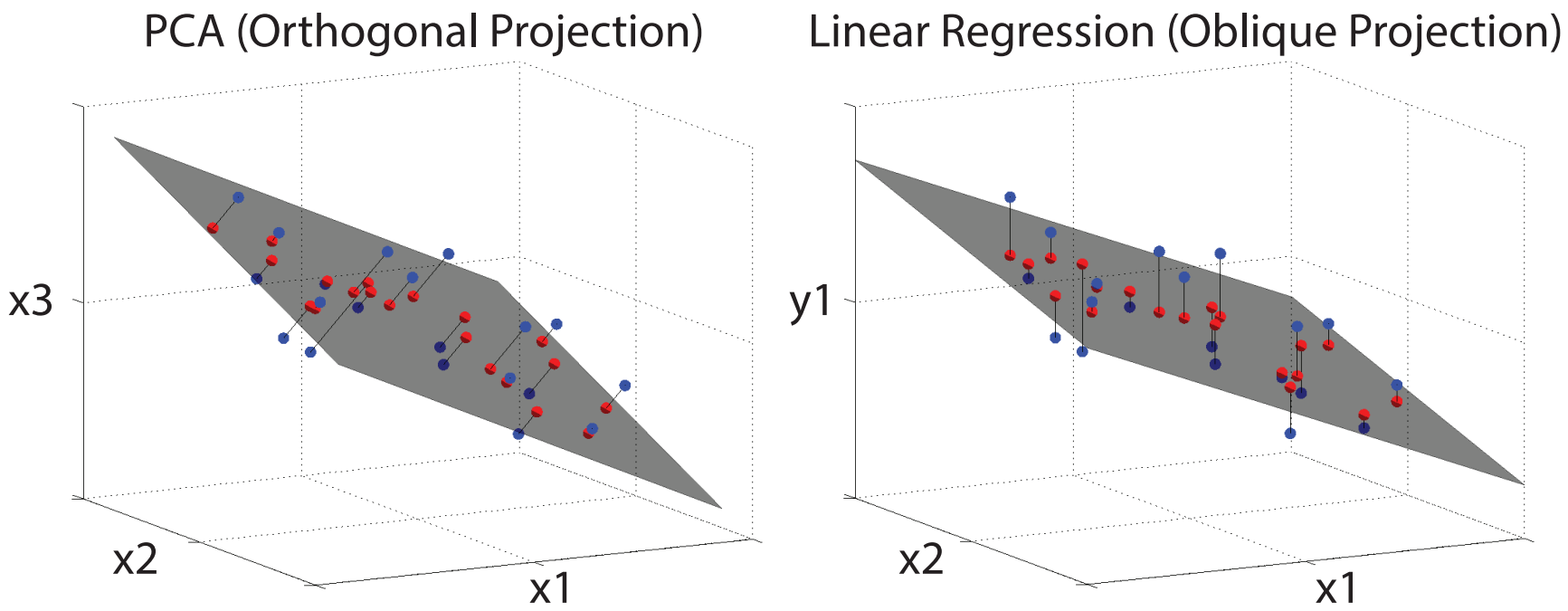
Often, more interpretable models are less flexible and vice versa.

Big Picture Concepts

- Inference vs. Prediction
- Parametric vs. Nonparametric
- Supervised vs. Unsupervised

Big Picture: Supervised vs Unsupervised

- Unsupervised learning seeks explanatory factors
- Supervised learning asserts explanatory factors



Next Time

We will talk more about:

- Classification
- Decomposing predictive error
- Evaluating predictive error