# Data Science II
## (P8106)

Yifei Sun

Assistant Professor

Department of Biostatistics
Mailman School of Public Health
Columbia University

Spring 2024

# General Information

- Class meeting time and place

  Tuesday and Thursday, 4:00 - 5:20 PM

  Rosenfield 8th Fl Auditorium

  3/19 VEC 201

  3/21 VP&S Amp 1

  2/20 No class
- Instructor office hours: Monday, 4 - 5pm, zoom

# General Information

- Teaching Assistants

| | |
|---|---|
| Yijin Wang | yw4005 |
| Ryan Wei | rw2844 |
| Bin Yang | by2303 |
| Runze Cui | rc3521 |

- TA office hours: TBD Wed & Fri 12-1pm

# General Information

- Evaluation based on
  - Homework (30%)
  - Mid-term project (30%)
  - Final project (30%)
  - Class participation (10%)
- Course materials are available at Canvas
- References
  - "*An Introduction to Statistical Learning*" (ISL) by James et al.
  - "*Applied Predictive Modeling*" (APM) by Kuhn and Johnson
  - "*Elements of Statistical Learning*" (ESL) by Hastie et al.
  - "*Tidy Modeling with R*" (TMR) by Kuhn and Silge

# General Information

- We assume that you know/have taken
  - **Calculus and Linear Algebra**
    - Derivative and integral
    - Inner product
    - Lagrange multiplier
    - Matrix, eigenvalue decomposition/singular value decomposition
  - Data Science I
  - Biostatistical Methods I
  - Introductory level probability and statistics
- R Markdown is required for homework
- Other courses
  - More mathematical details: P9120 “Topics in Statistical Learning and Data Mining”
  - Non-biostatistical students: P8451 “Introduction to Machine Learning for Epidemiology and Public Health”

# What is Data Science?

- Data science encompasses a set of principles, algorithms and processes for extracting useful patterns from data
- Many of the elements of data science have been developed in related fields such as machine learning and data mining
- Data science is broader in scope
  - Data wrangling and databases
  - Data visualization
  - Statistics and Probability
  - Machine learning
  - Domain expertise
  - Ethics and regulation
  - ...
- Machine learning is a fundamental ingredient in the training of a modern data scientist

DS I

DS II

# Outline of the course

- In DSII, we will cover
    - Regression
    - Classification
    - Clustering, Dimension reduction
    - And their implementations in R
- 70% method/algorithm + 30% implementation

# Programming in DSII

- Every tool has its shelf life
- Different dialects/syntaxes in R
  - base R (e.g., $, [[ ]], . . .) - stable
  - Add-on packages: tidyverse (readability), data.table (fast), ...
- R packages for machine learning (e.g., glmnet, ranger, ...)
- Meta-engine (caret, parsnip)

tidymodels

# Supervised Learning

- Predictor measurements $X$
  (inputs/regressors/covariates/features/independent variables)
- Outcome measurement $Y$
  (dependent variable/response/target)
  - $Y$ is quantitative - regression
  - $Y$ takes values in a finite and unordered set - classification
- Training data: $(x_1, y_1), \ldots, (x_n, y_n)$
- Objectives:

  $\hat{y}_0$
  - Understand which inputs affect the outcome, and how
  - Accurately predict unseen cases $(x_0, y_0)$
  - Assess the quality of our predictions and inferences

# Unsupervised Learning

- No outcome $Y$
- Objectives:
  - Find groups of samples that behave similarly
  - Find features that behave similarly
  - Find linear combinations of features with the most variation
  - ...
- Can be used as a pre-processing step for supervised learning

clustering

PCA

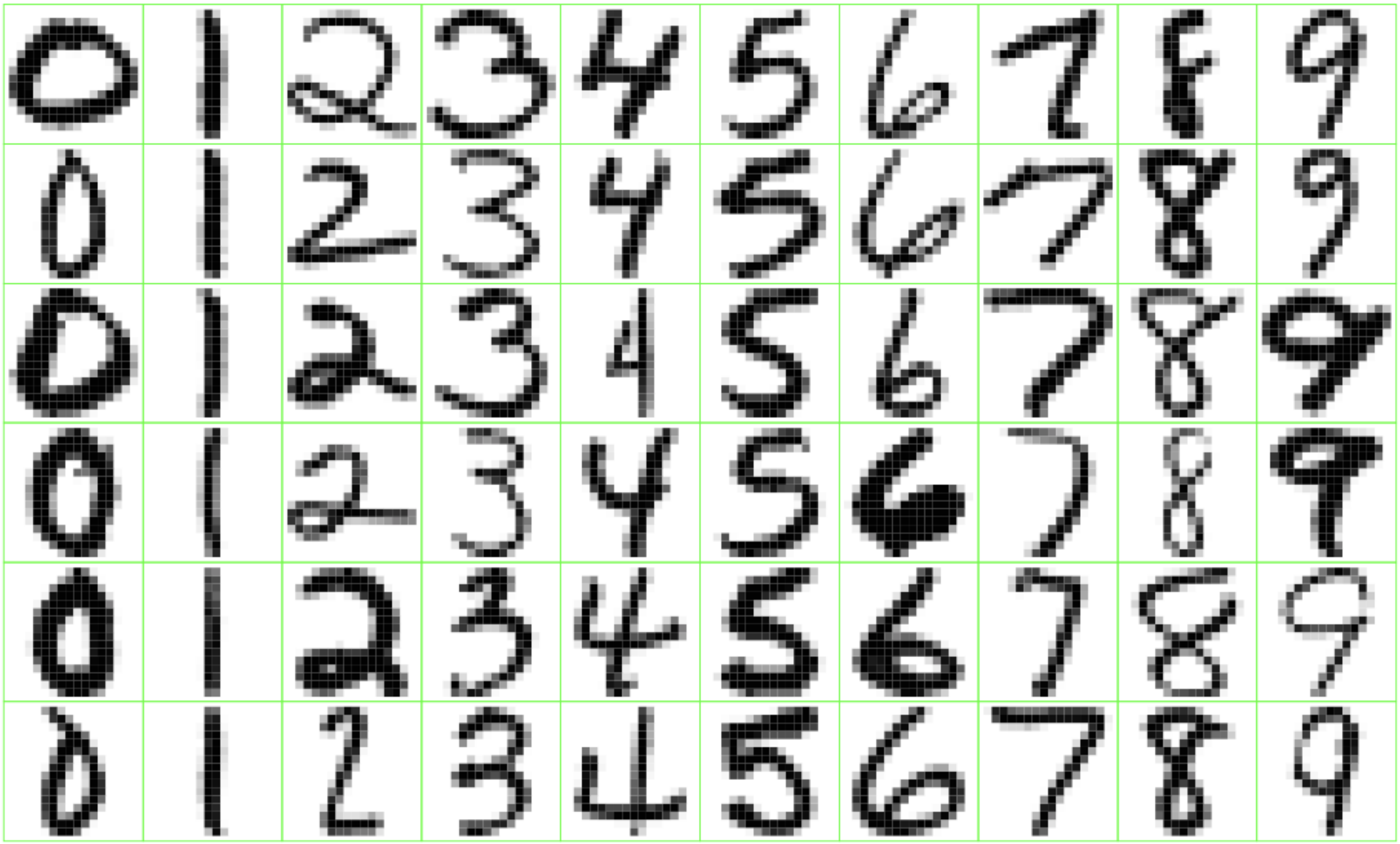
# Example $n=60$

Identify the numbers in a handwritten zip code

$X = (X_1, \cdots, X_{256})$

16

16

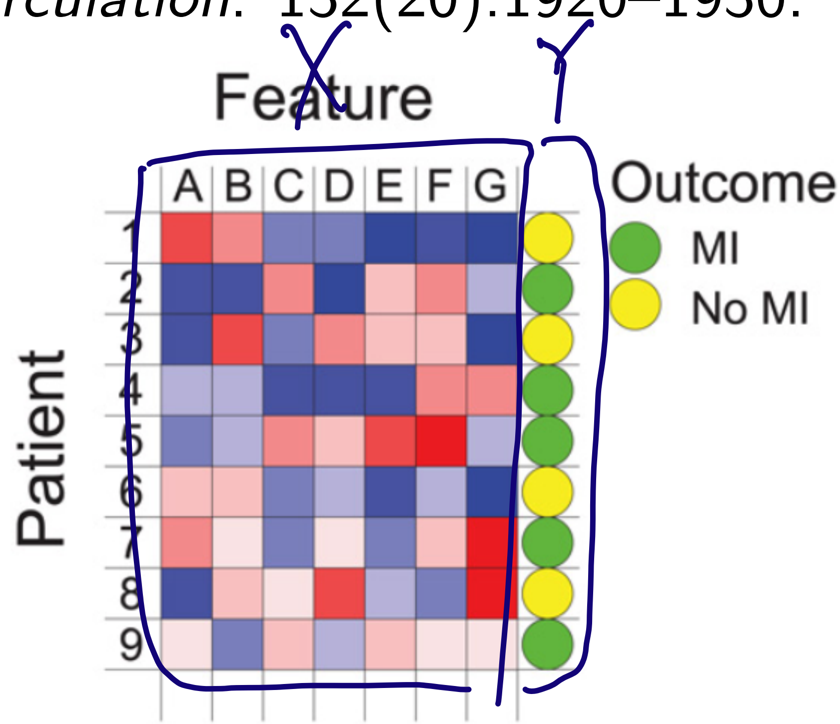


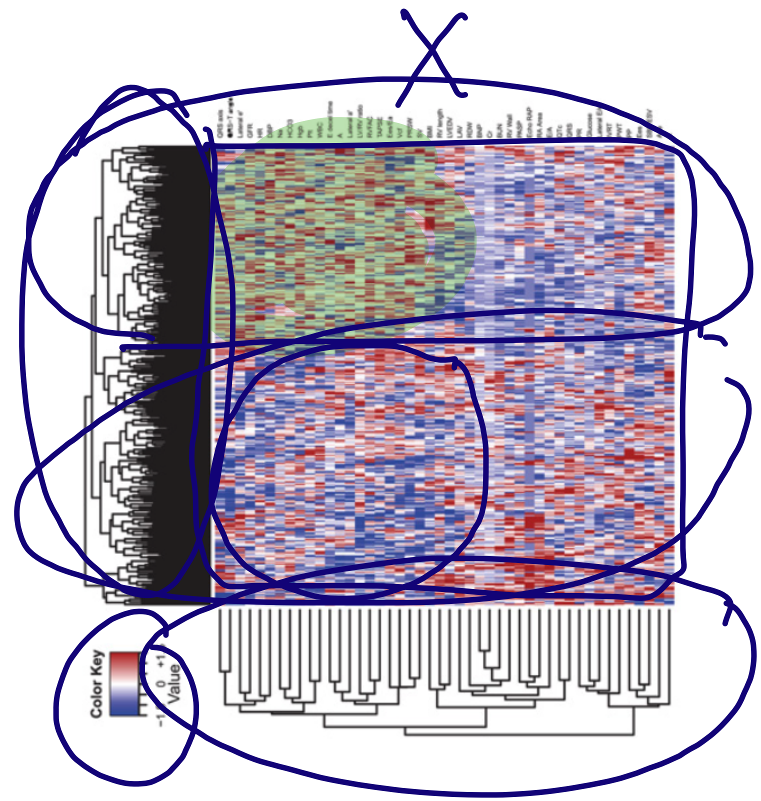
$Y$ 0 1 2 ... 9

Supervised? Unsupervised?

# Real world applications: medical science

- Supervised learning $Y \in \{0, 1\}$
  - Predict disease outcome or risk score $p(Y=1 \mid X)$
- Unsupervised learning
  - Identify disease subtypes
- "Machine Learning in Medicine."
  *Circulation*. 132(20):1920–1930.



# Notation

- Quantitative response $Y$
- $p$ different predictors, $X = (X_1, X_2, \ldots, X_p)$.
- Now we write our model as

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

  - $f$ represents the systematic information $E[\varepsilon \mid X] = 0$
  - $\epsilon$ is a zero-mean error term and independent of $X$
- Statistical learning refers to approaches for estimating $f$

# Why estimate $f$?

- Information: To extract some information about how the response variables are associated with the input variables.
  - Understand which components of $X$ are important in explaining $Y$
  - Understand how each component $X_j$ of $X$ affects $Y$
- Prediction
  - Make predictions of $Y$ at new points $X = x$

# The regression function

- What is a good prediction of $Y$ at any point $X = x$?
- $f(x) = E[Y \mid X = x]$

$\hat{f}(x)$

## The regression function $f(x)$

$E(Y - c)^2$

$c = EY$

- $f(x) = E(Y|X = x)$ is the *optimal* predictor of $Y$ with regard to the **mean-squared prediction error**
  i.e., $f(x) = E(Y|X = x)$ is the function that minimized $E[(Y - g(X))^2|X = x]$ over all functions $g$ at all points $X = x$
- $\epsilon = Y - f(x)$ is the *irreducible* error
  i.e., even if we knew $f(x)$, we would still make error in prediction, since at each $X = x$ there is typically a distribution of possible $Y$ values
- How do we estimate $f$?

# Parametric models

The linear model is an important example of a parametric model:

$$f_L(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$$

- A *linear* model is specified in terms of $p+1$ parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ $\hat{f}_L(x)$
- We estimate the parameters by fitting the model to training data
- Although it is *almost never correct*, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$

# Nonparametric methods

$Y \quad X \in \{0, 1\}$

$\hat{E}[Y | X=0] \quad \hat{E}[Y | X=1]$

- No explicit assumptions of the functional form of $f$
- Typically we have few if any data points with $X = x$ exactly
- Relax the definition and let

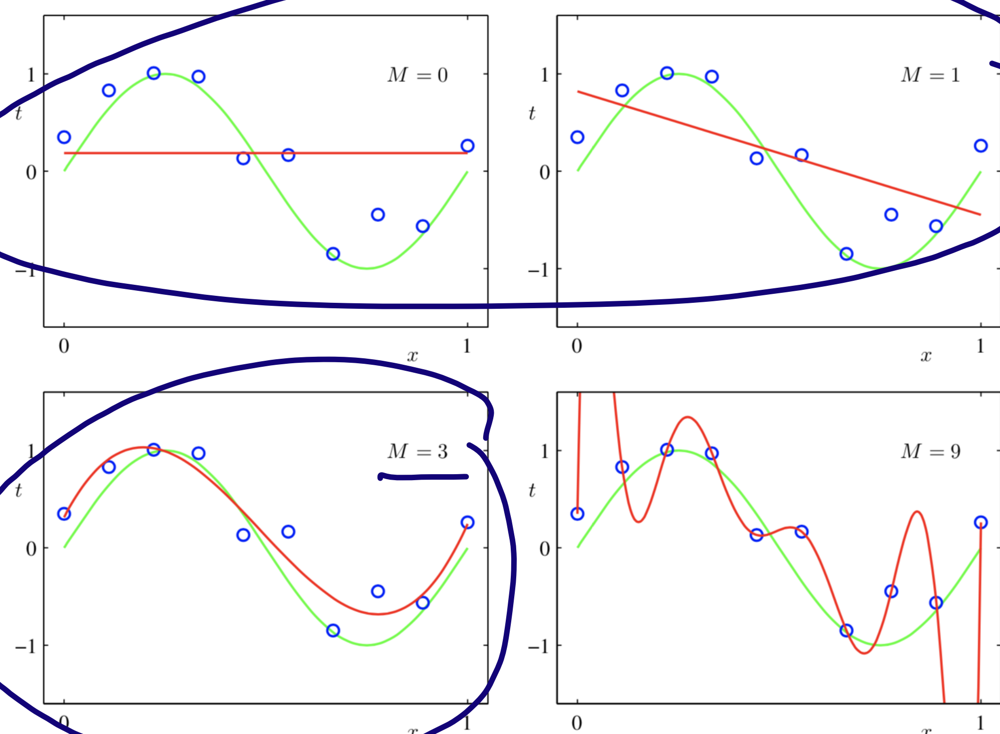
$$\widehat{f}(x) = Ave(Y | X \in \mathcal{N}(x))$$

where $\mathcal{N}$ is some *neighborhood* of $x$

- Advantage: Can be used to fit a wider range of possible shapes for $f$
- Disadvantage: A large number of observations is needed

# Simulated example

- $f(x) = \sin(2\pi x)$ is the green curve
- Blue points are simulated from the model $Y = f(X) + \epsilon$
- Red curves are polynomial functions of orders $M$ fitted to the data $f_M(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_M X^M$



# Overfitting

- A low degree of freedom leads to underfitting
- With an extremely high degree of flexibility, the model does its best to account for every single data point
- Cannot generalize well to new data

# Assessing model accuracy

Suppose we fit a model $\widehat{f}(x)$ to the **training** data
$\mathrm{Tr} = \{(x_i, y_i), i = 1, \ldots, n\}$

- Compute the average squared prediction error over $\mathrm{Tr}$

$$\mathrm{MSE}_{\mathrm{Tr}} = \mathrm{Ave}_{i \in \mathrm{Tr}}[y_i - \underbrace{\widehat{f}(x_i)}_{\hat{y}_i}]^2$$

- Can we use $\mathrm{MSE}_{\mathrm{Tr}}$?
- Instead, we should, if possible, compute it using fresh **test** data $\mathrm{Te} = \{(x_i^*, y_i^*), i = 1, \ldots, m\}$:

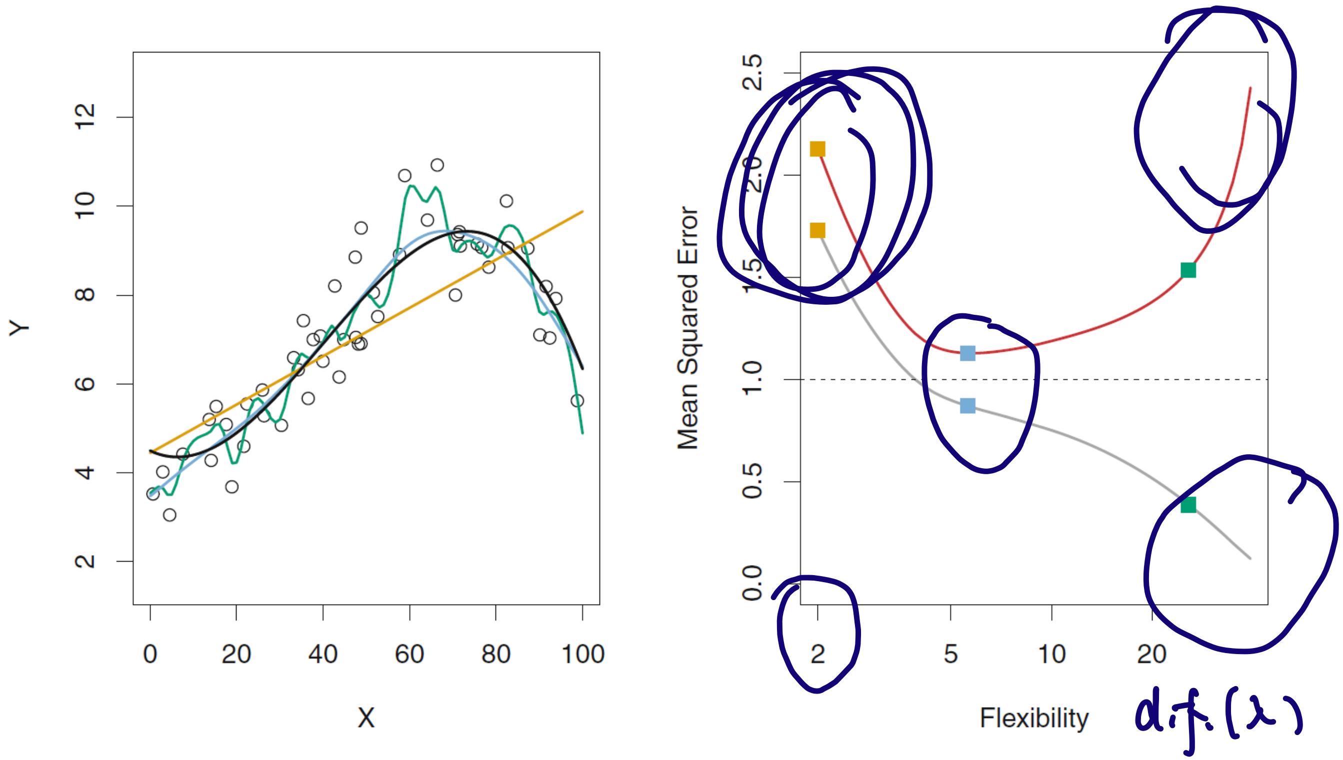
$$\mathrm{MSE}_{\mathrm{Te}} = \mathrm{Ave}_{i \in \mathrm{Te}}[y_i^* - \widehat{f}(x_i^*)]^2$$

# Example I on training and test MSE

small $\lambda$ large

smoothing splines

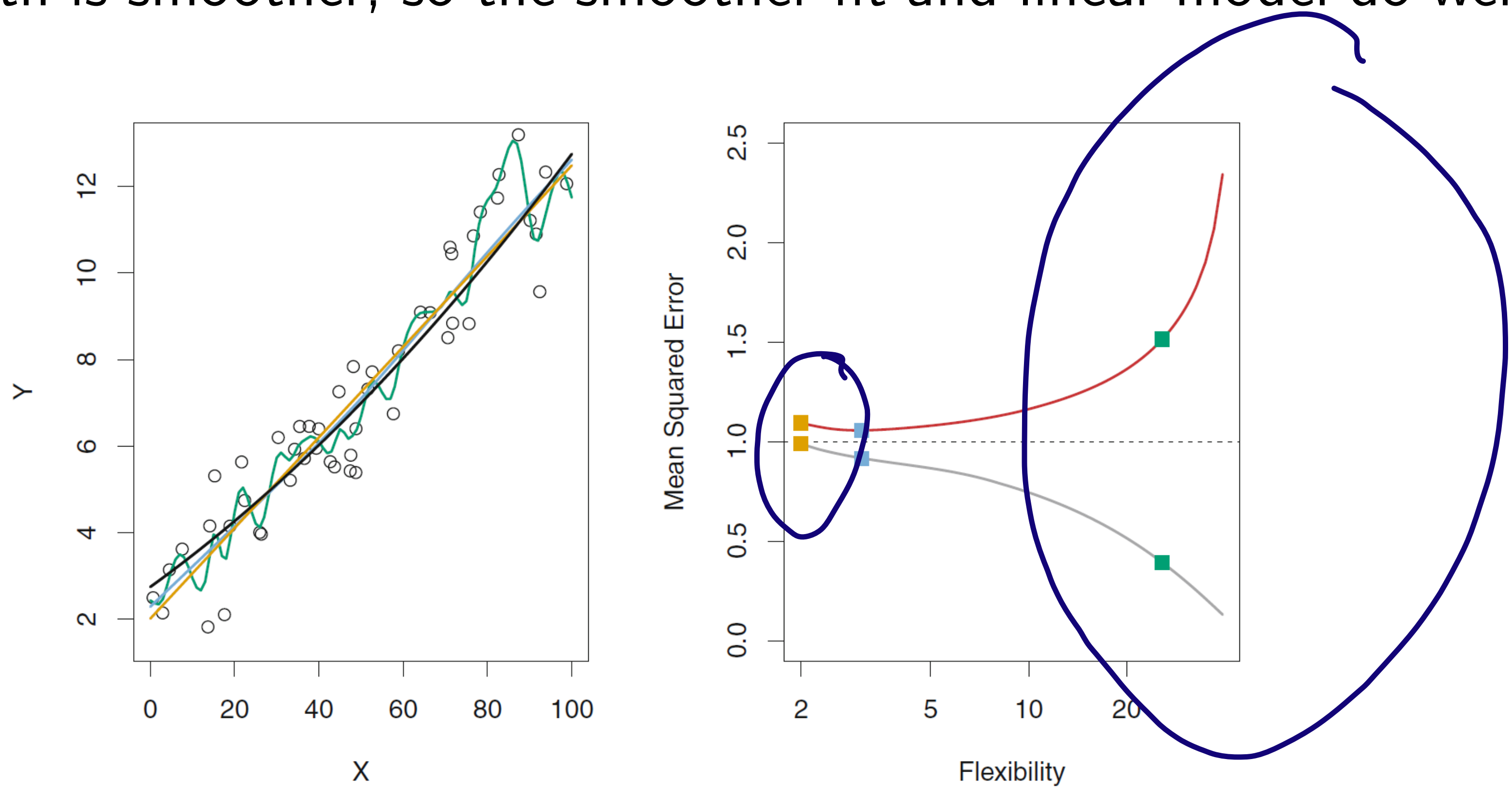
- Left: Black curve is truth
- Right: Red curve on right is $\mathrm{MSE}_{\mathrm{Te}}$, grey curve is $\mathrm{MSE}_{\mathrm{Tr}}$
- Orange, blue and green curves/squares correspond to fits of different flexibility




[ISL] Figure 2.9

# Example II on training and test MSE

The truth is smoother, so the smoother fit and linear model do well




[ISL] Figure 2.10

# A question for you

- High/low
- A model that underfits the data will have high training error and high testing error
- A model that overfits the data will have low training error and high testing error