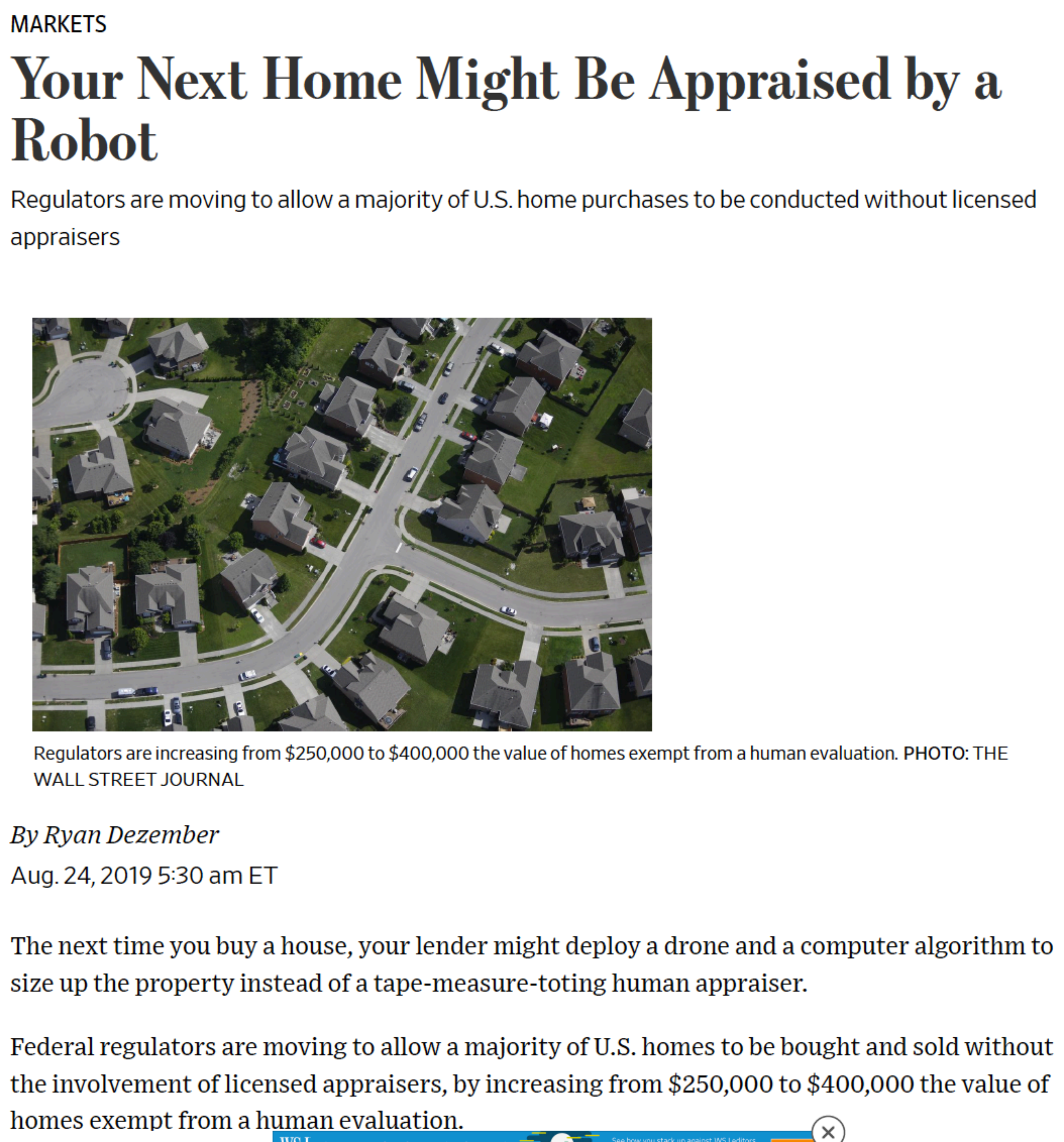
Week 2

BUS 696: Machine Learning for Managers

Prof. Jonathan Hersh

BUS 696: Week 2 Outline

1. **AI/ML in the News**
2. Data Modeling Project Cycle
3. R Coding Basics and Exploratory Data Analysis
   - Basic math in R
   - Vectors
   - Factors
   - Dplyr basics: filter(), select(), arrange(), rename(), mutate()
4. Bias Variance Tradeoff



MARKETS

Your Next Home Might Be Appraised by a Robot

Regulators are moving to allow a majority of U.S. home purchases to be conducted without licensed appraisers

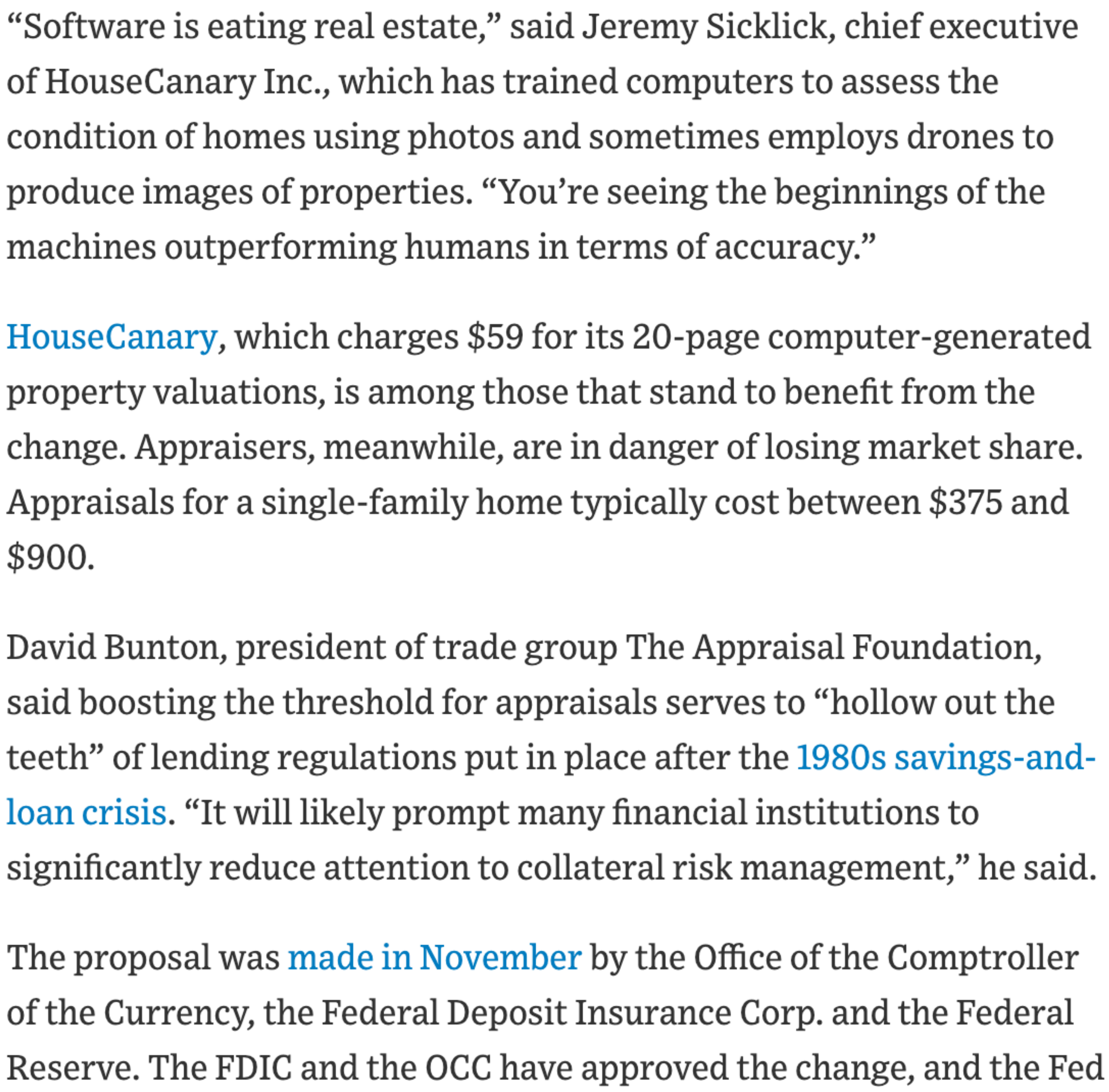
Regulators are increasing from $250,000 to $400,000 the value of homes exempt from a human evaluation. **PHOTO:** THE WALL STREET JOURNAL

By Ryan Dezember

Aug. 24, 2019 5:30 am ET

The next time you buy a house, your lender might deploy a drone and a computer algorithm to size up the property instead of a tape-measure-toting human appraiser.

Federal regulators are moving to allow a majority of U.S. homes to be bought and sold without the involvement of licensed appraisers, by increasing from $250,000 to $400,000 the value of homes exempt from a human evaluation.

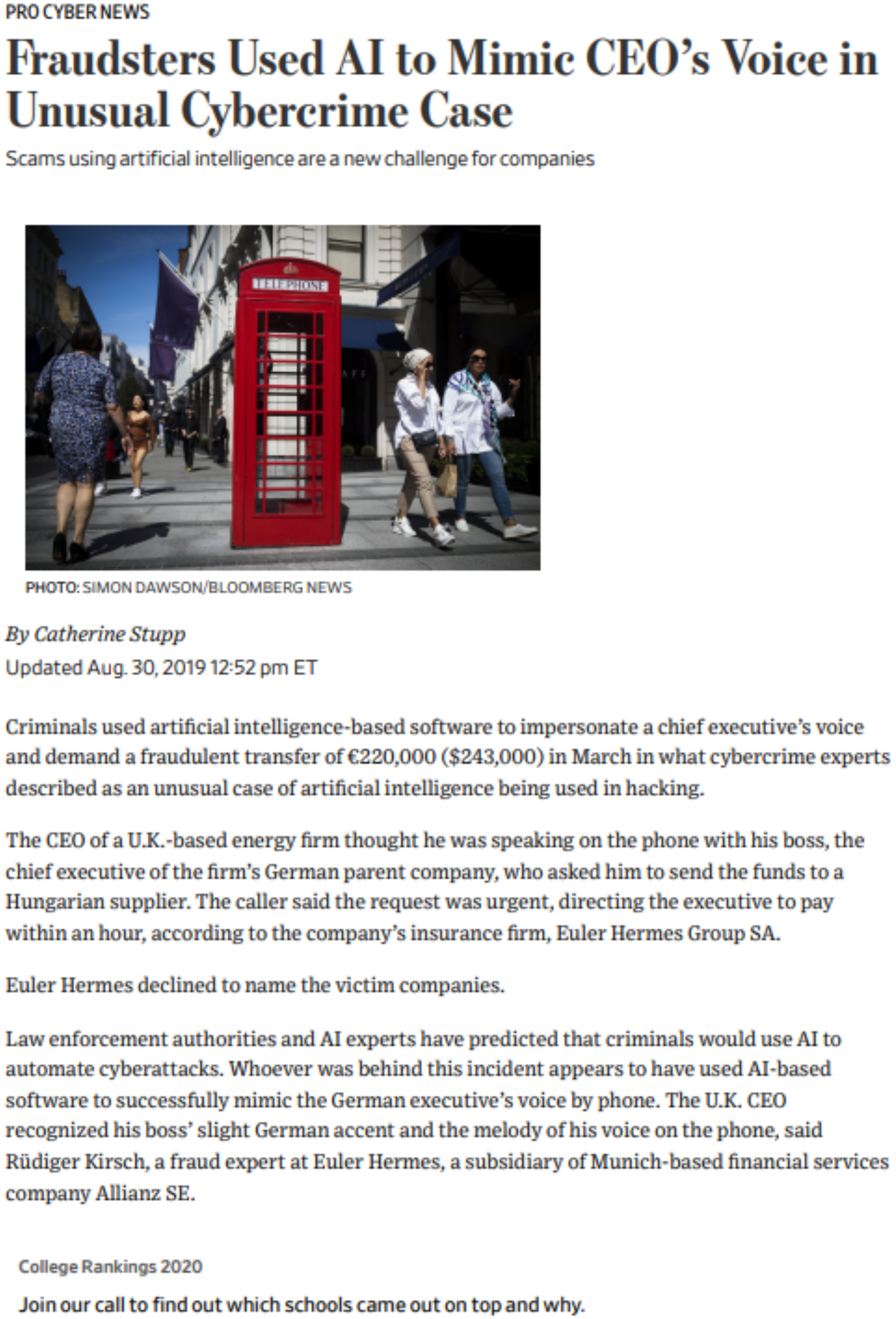


“Software is eating real estate,” said Jeremy Sicklick, chief executive of HouseCanary Inc., which has trained computers to assess the condition of homes using photos and sometimes employs drones to produce images of properties. “You’re seeing the beginnings of the machines outperforming humans in terms of accuracy.”

HouseCanary, which charges $59 for its 20-page computer-generated property valuations, is among those that stand to benefit from the change. Appraisers, meanwhile, are in danger of losing market share. Appraisals for a single-family home typically cost between $375 and $900.

David Bunton, president of trade group The Appraisal Foundation, said boosting the threshold for appraisals serves to “hollow out the teeth” of lending regulations put in place after the 1980s savings-and-loan crisis. “It will likely prompt many financial institutions to significantly reduce attention to collateral risk management,” he said.

The proposal was made in November by the Office of the Comptroller of the Currency, the Federal Deposit Insurance Corp. and the Federal Reserve. The FDIC and the OCC have approved the change, and the Fed



PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

PHOTO: SIMON DAWSON/BLOOMBERG NEWS

By Catherine Stupp

Updated Aug. 30, 2019 12:52 pm ET

Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

The CEO of a U.K.-based energy firm thought he was speaking on the phone with his boss, the chief executive of the firm's German parent company, who asked him to send the funds to a Hungarian supplier. The caller said the request was urgent, directing the executive to pay within an hour, according to the company's insurance firm, Euler Hermes Group SA.

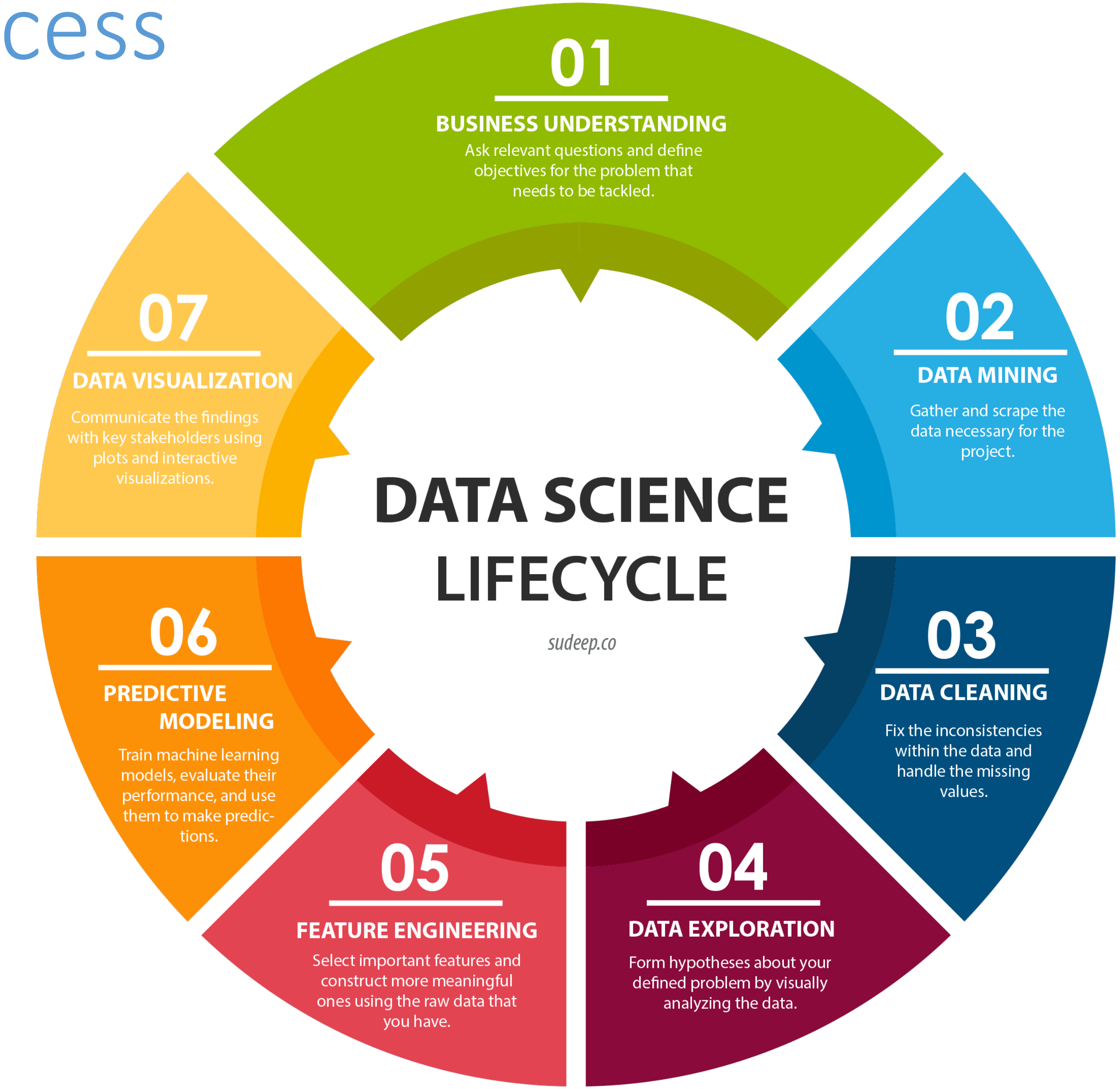
Euler Hermes declined to name the victim companies.

Law enforcement authorities and AI experts have predicted that criminals would use AI to automate cyberattacks. Whoever was behind this incident appears to have used AI-based software to successfully mimic the German executive's voice by phone. The U.K. CEO recognized his boss' slight German accent and the melody of his voice on the phone, said Rüdiger Kirsch, a fraud expert at Euler Hermes, a subsidiary of Munich-based financial services company Allianz SE.

BUS 696: Week 2 Outline

1. AI/ML in the News
2. **Data Modeling Project Cycle**
3. R Coding Basics and Exploratory Data Analysis
    - Basic math in R
    - Vectors
    - Factors
    - Dplyr basics: filter(), select(), arrange(), rename(), mutate()
4. Bias Variance Tradeoff

Data Modeling Process



DATA SCIENCE LIFECYCLE

01 BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING

Gather and scrape the data necessary for the project.

03 DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING

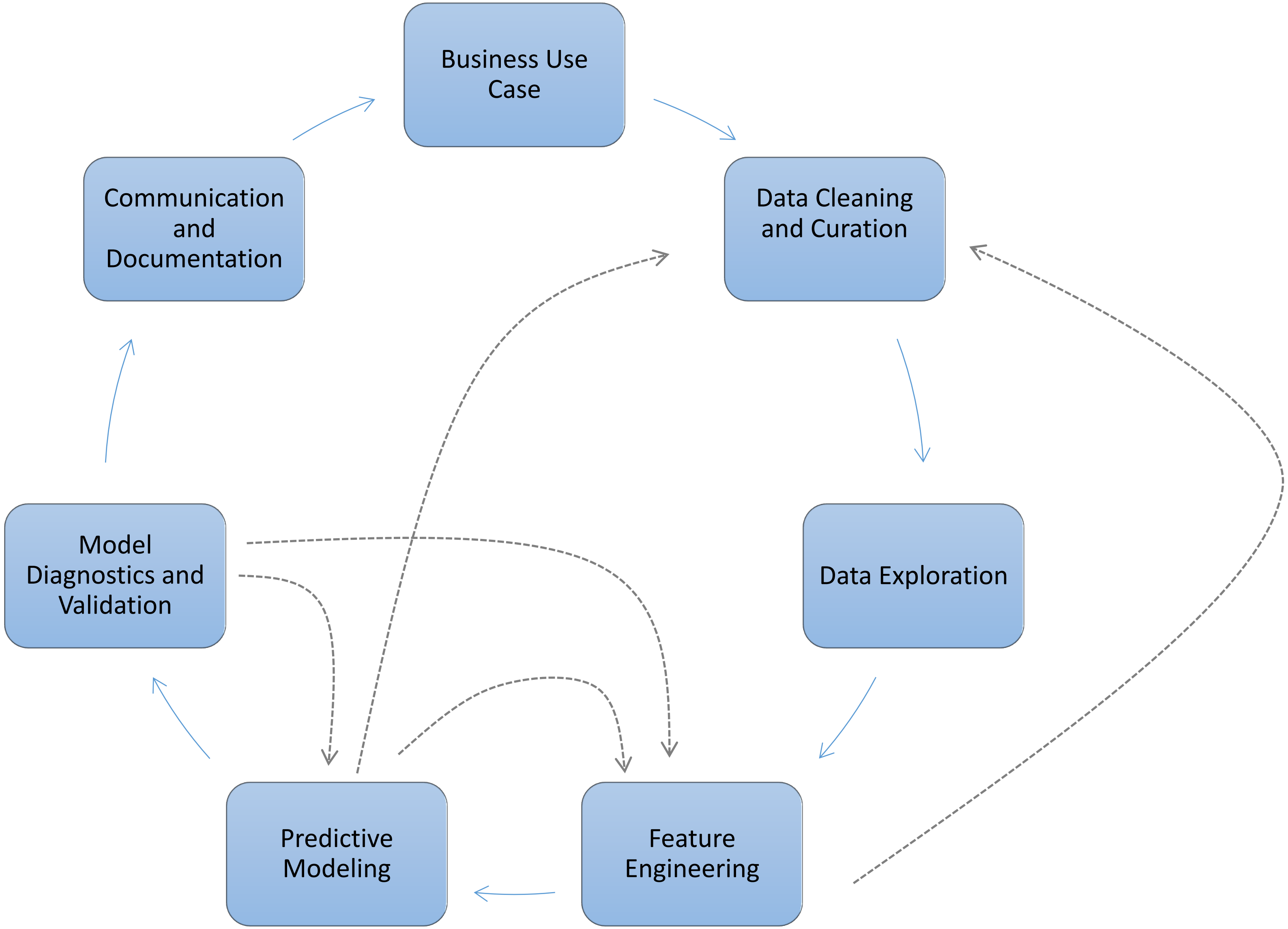
Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION

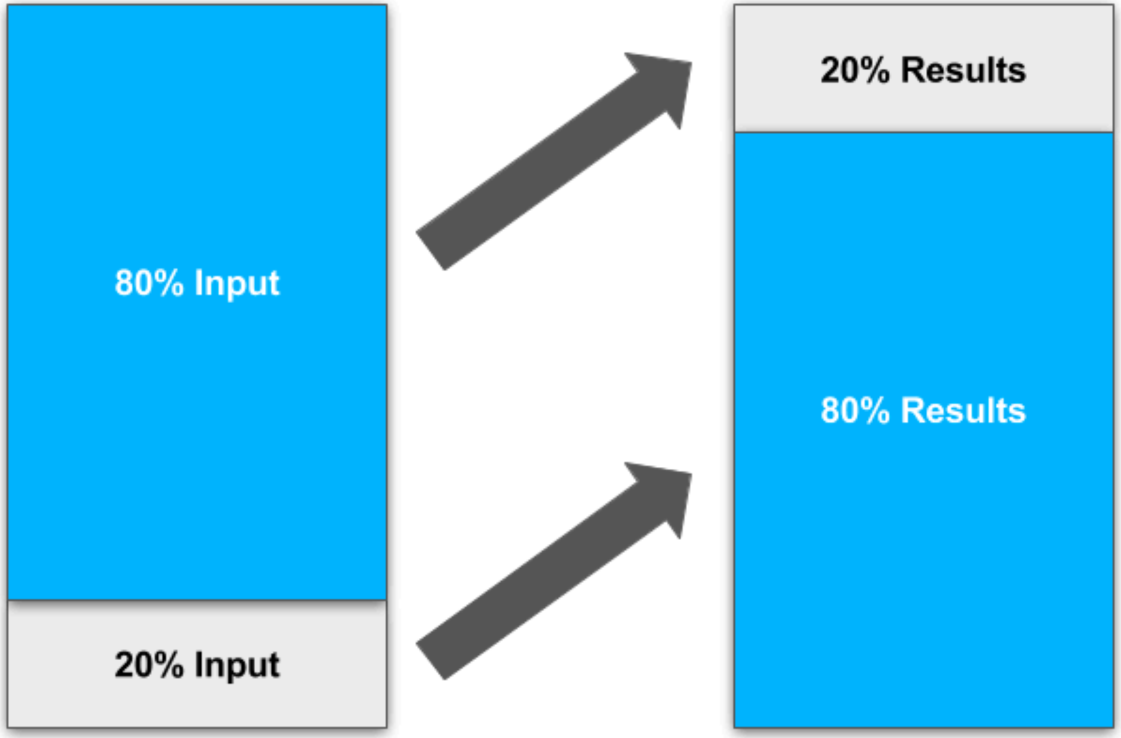
Communicate the findings with key stakeholders using plots and interactive visualizations.

sudeep.co

A More Realistic Data Project Life-Cycle Model


Business Use Case
Data Cleaning and Curation
Data Exploration
Feature Engineering
Predictive Modeling
Model Diagnostics and Validation
Communication and Documentation

80/20 Rule for Results vs Effort



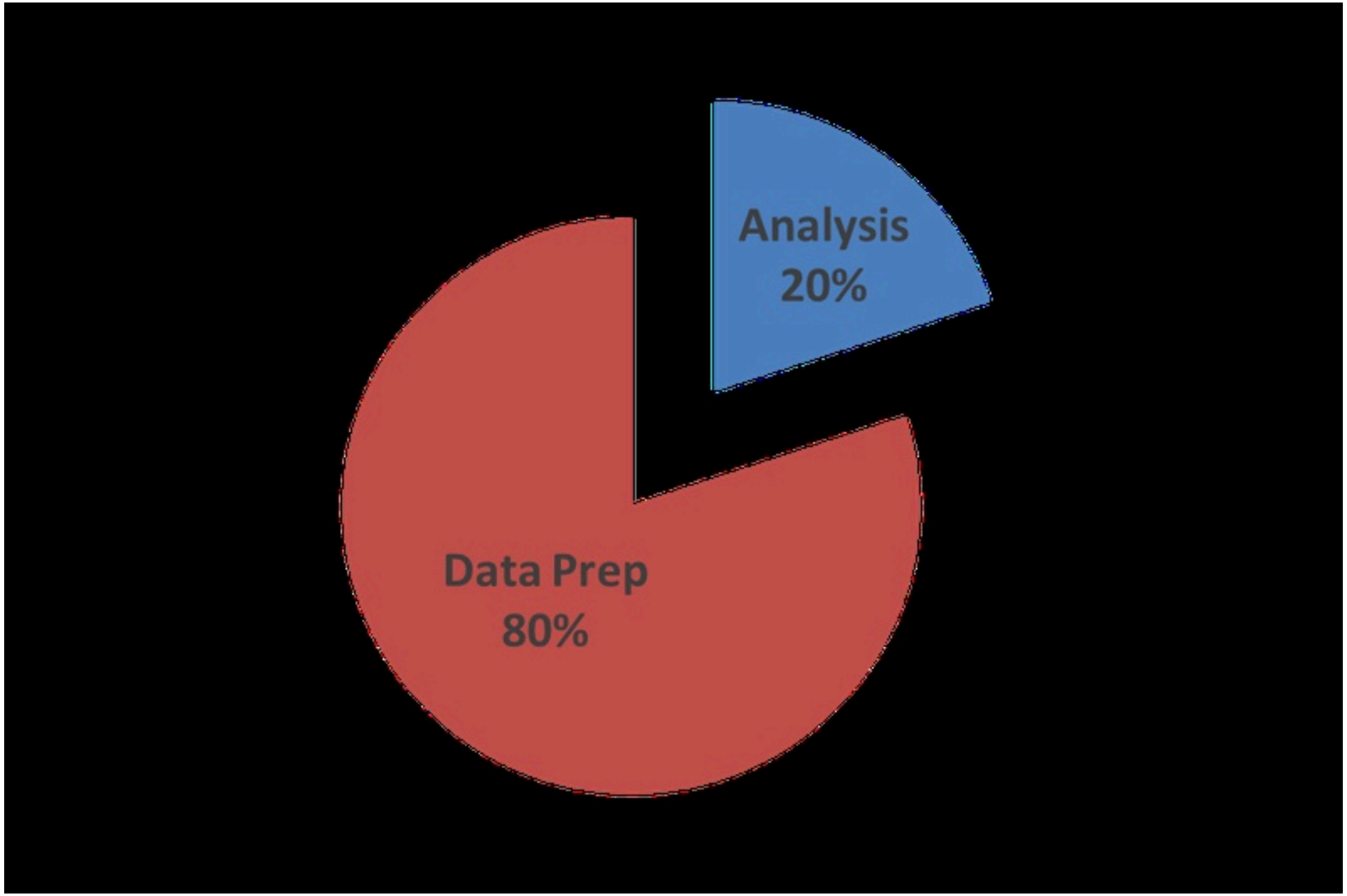
80% Input

20% Input

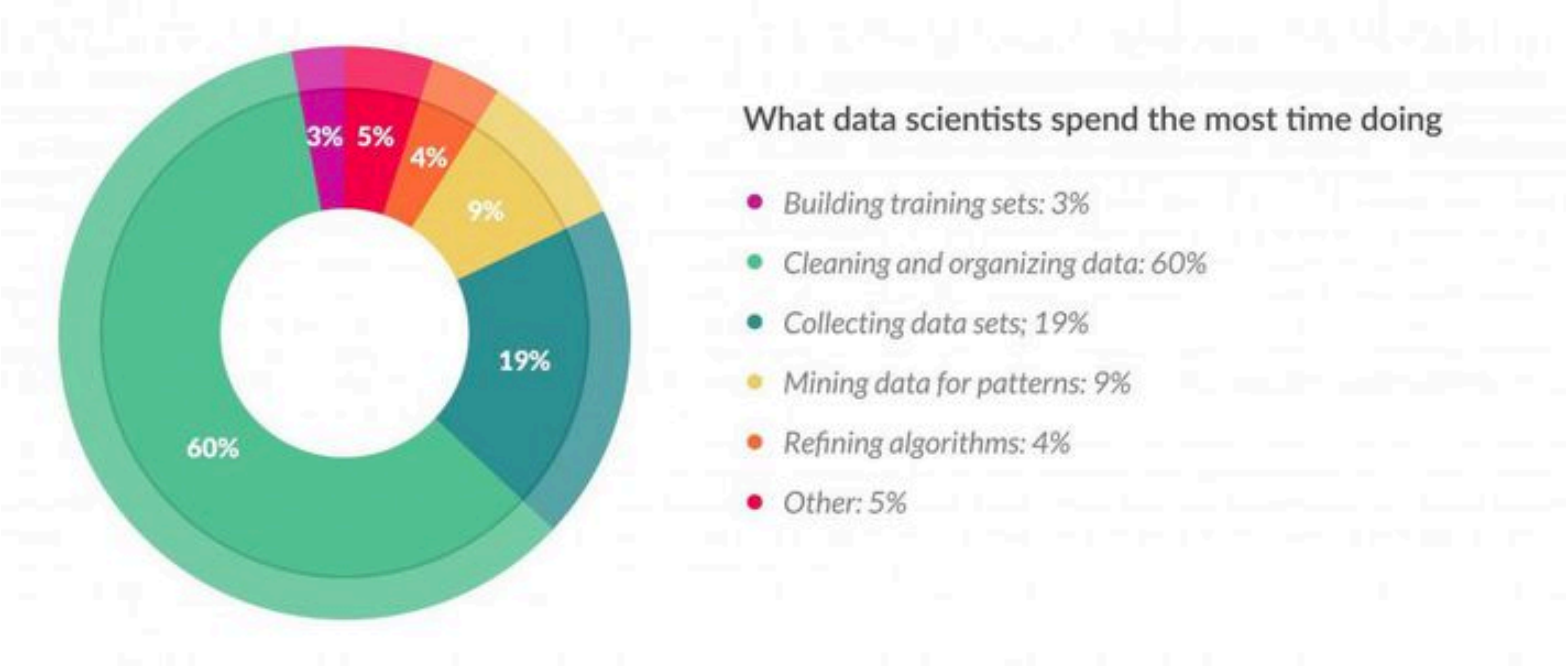
20% Results

80% Results

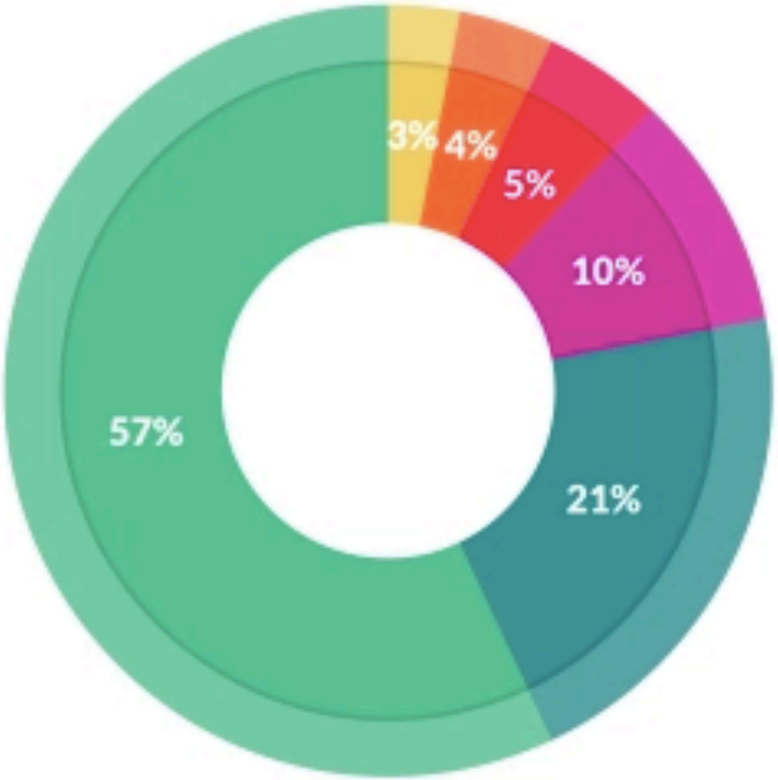
80/20 Rule for Data Science



How do Data Scientists Spend Their Time



Least Enjoyable Aspect of Data Science



What's the least enjoyable part of data science?

- ● *Building training sets: 10%*
- ● *Cleaning and organizing data: 57%*
- ● *Collecting data sets: 21%*
- ● *Mining data for patterns: 3%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

BUS 696: Week 2 Outline

1. AI/ML in the News
2. Data Modeling Project Cycle
3. **R Coding Basics and Exploratory Data Analysis**
    - **Basic math in R**
    - **Vectors**
    - **Factors**
    - **Dplyr basics: filter(), select(), arrange(), rename(), mutate()**
4. Bias Variance Tradeoff

BUS 696: Week 2 Outline

1. AI/ML in the News
2. Data Modeling Project Cycle
3. R Coding Basics and Exploratory Data Analysis
   - Basic math in R
   - Vectors
   - Factors
   - Dplyr basics: filter(), select(), arrange(), rename(), mutate()
4. **Bias Variance Tradeoff**

Recall: $\mathbf{X}$, x_i, $\mathbf{x}_j$ and $\mathbf{y}$ for Data

$\mathbf{X}$ $n\times p$

x_i $1\times p$

$\mathbf{x}_j$ $n\times 1$

$\mathbf{y}$ $n\times 1$

$i = 1, \ldots, n$

$j = 1, \ldots, p$

Admit	GRE	GPA
0	380	3.61
1	660	3.67
1	800	4.00
0	520	2.93
1	760	3.00

Matrix Multiplication

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{d} a_{ik}\, b_{kj}$$

$$\begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 5 \\ 3 & 10 \end{bmatrix}$$

Random Variable: Variable whose possible values are outcomes of some random process

$X \sim$roll of dice

$$X \in \{1,2,3,4,5,6\}$$

$$p_1, \ldots, p_6 = \left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$$

Expected Value of R.v.: long run average value

$$\mathbb{E}[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

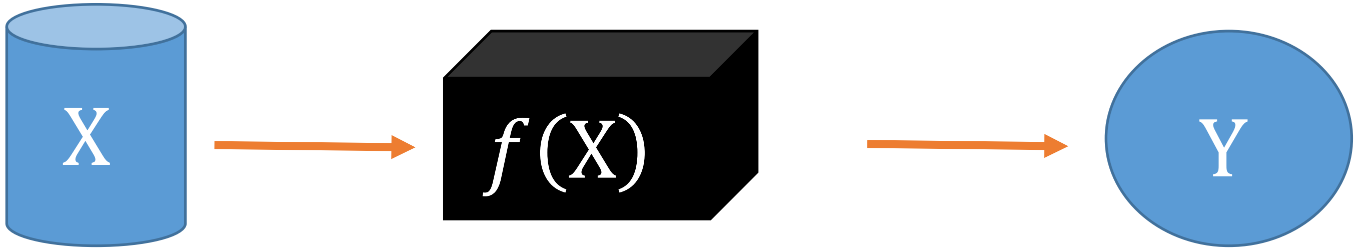
$\mathrm{X} \in \{1,2,3,4,5,6\}$

$$p_1, \ldots, p_6 = \left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$$

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

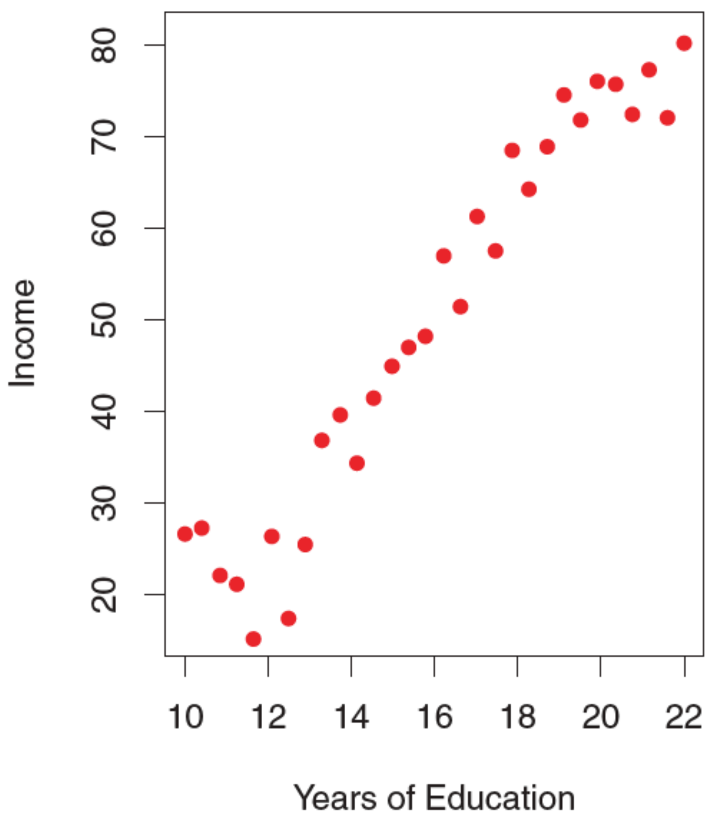
$f(\mathrm{X})$: our predicted out given inputs

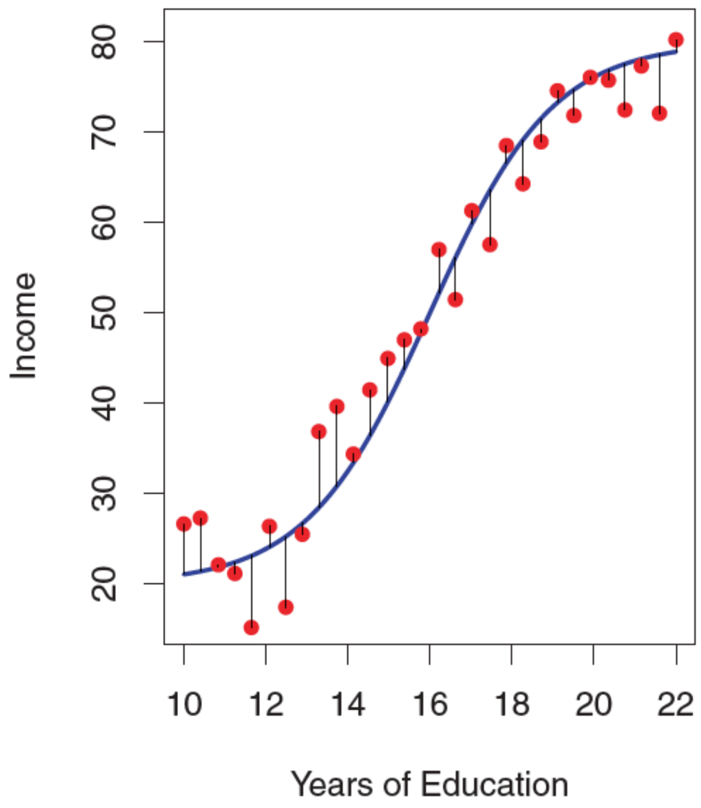
$$\mathrm{Y} = f(\mathrm{X}) + \epsilon$$



ϵ = “epsilon” (unexplained portion)

Example: education and income





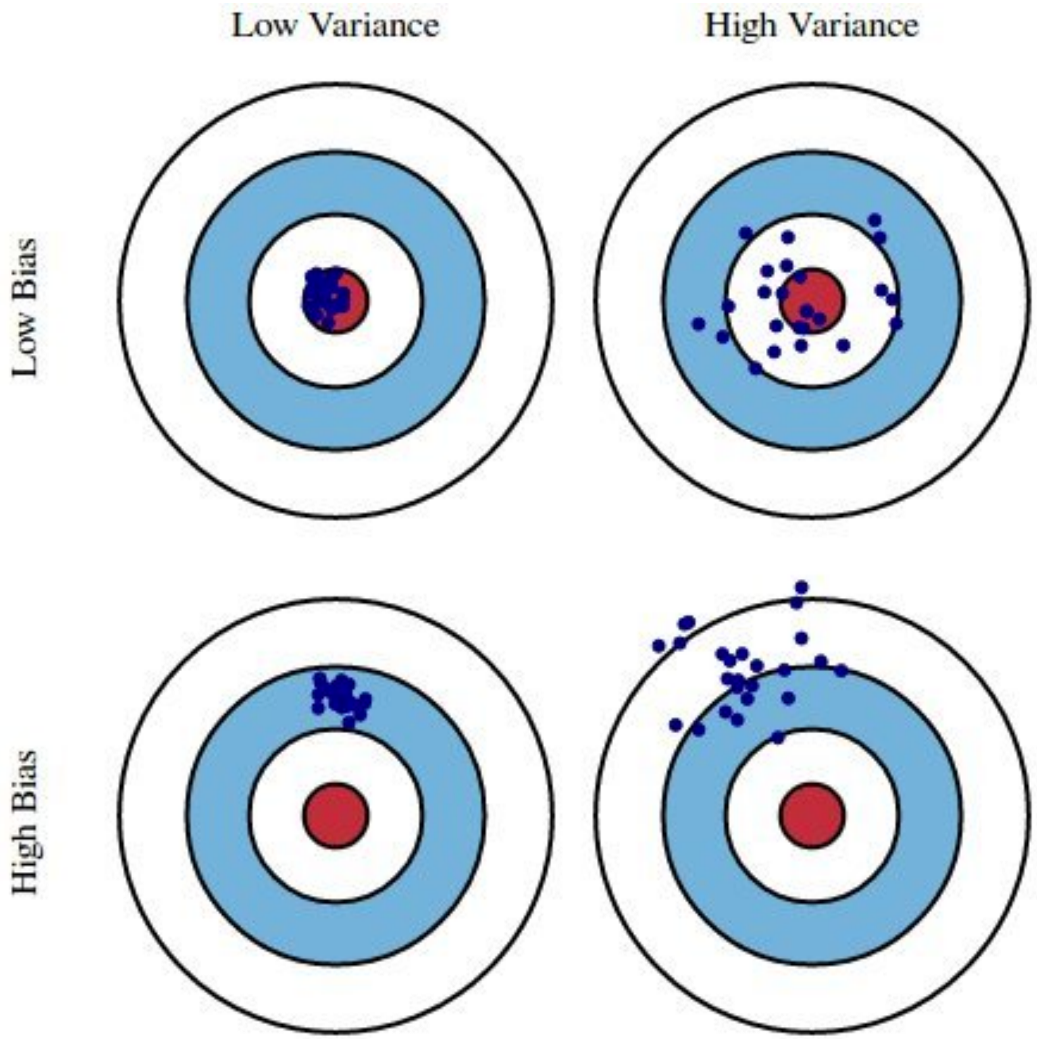
“Estimating” $f(\mathrm{X})$

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

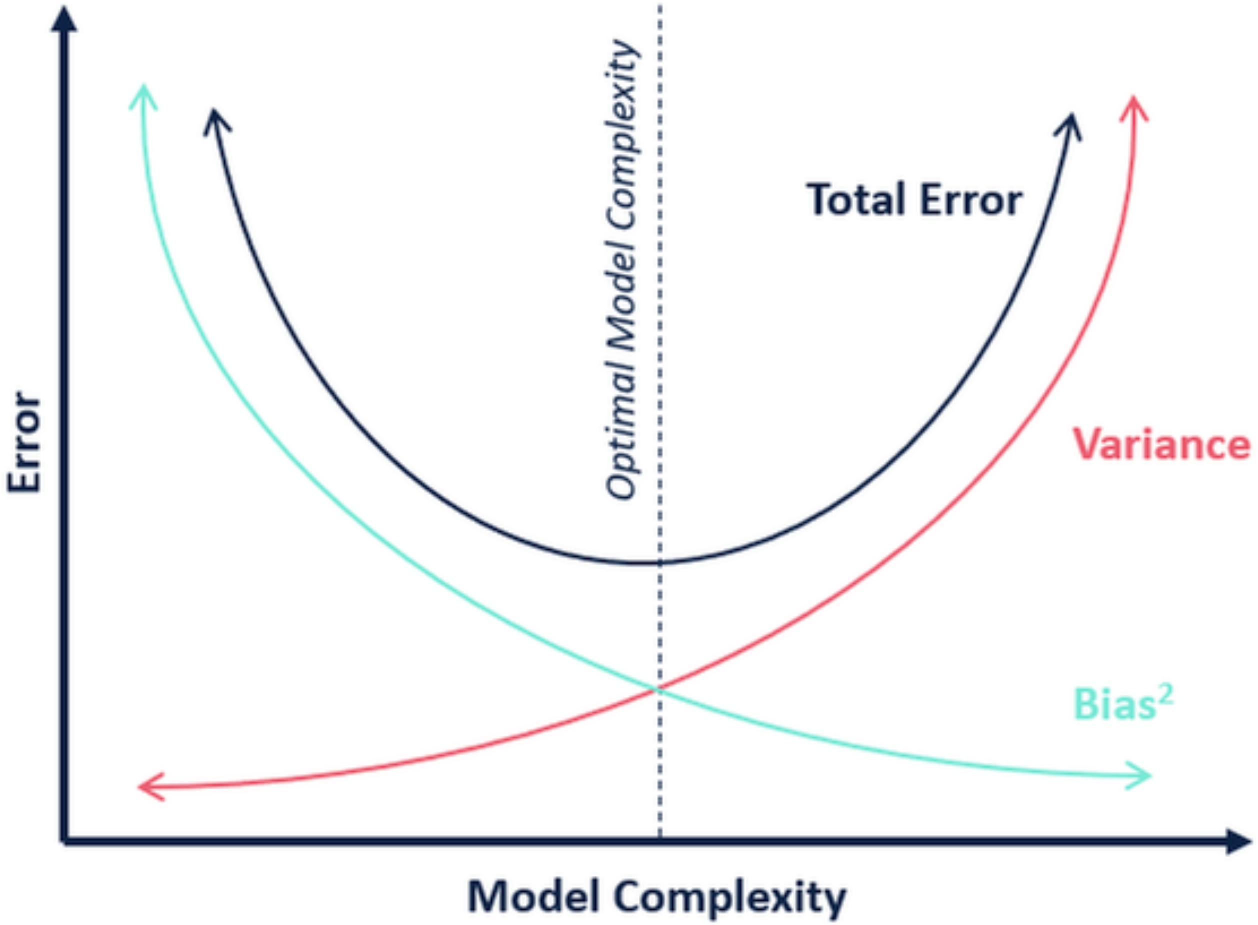
How do we know how $f(\mathrm{X})$ should transform

$\mathbf{x}_1, \dots, \mathbf{x}_p$ to generate our guess for Y?

Bias-Variance Tradeoff

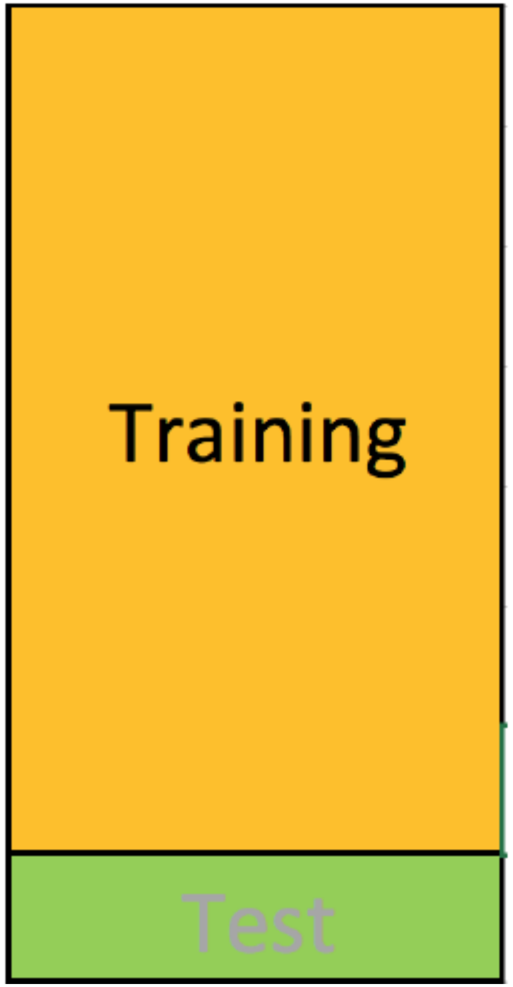


Bias-Variance Tradeoff



Testing/Training Split

Training set: (observation-wise) subset of data used to develop models

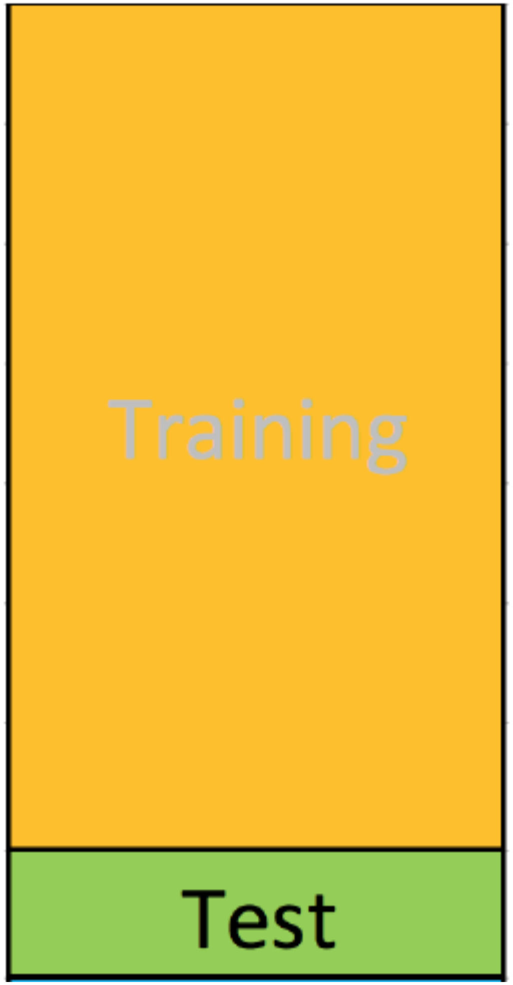


Testing/Training Split

Training set: (observation-wise) subset of data used to develop models

Test set: subset of data used during intermediate stages to “tune” model parameters

Rule of thumb 75% training 25% test -ish



Training

Test

Bias and Variance (conceptually)

Bias: **Tendency of an in-sample statistic to over or under estimate the statistic in the *population***

Variance: **Tendency to noisily estimate a statistic**.

E.g., sensitivity to small fluctuations in the training dataset.

Assessing Model Accuracy

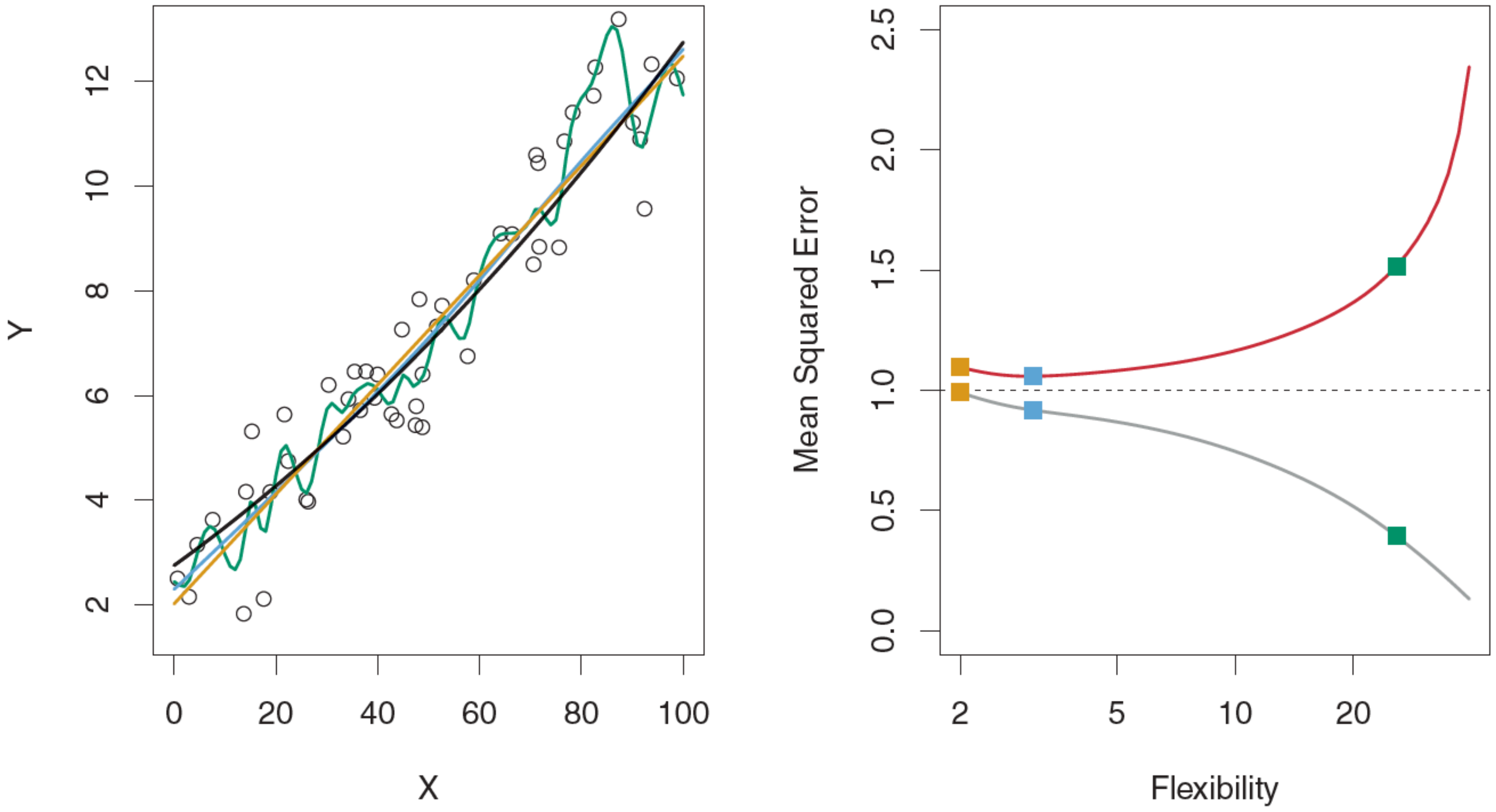
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

Assessing Model Accuracy

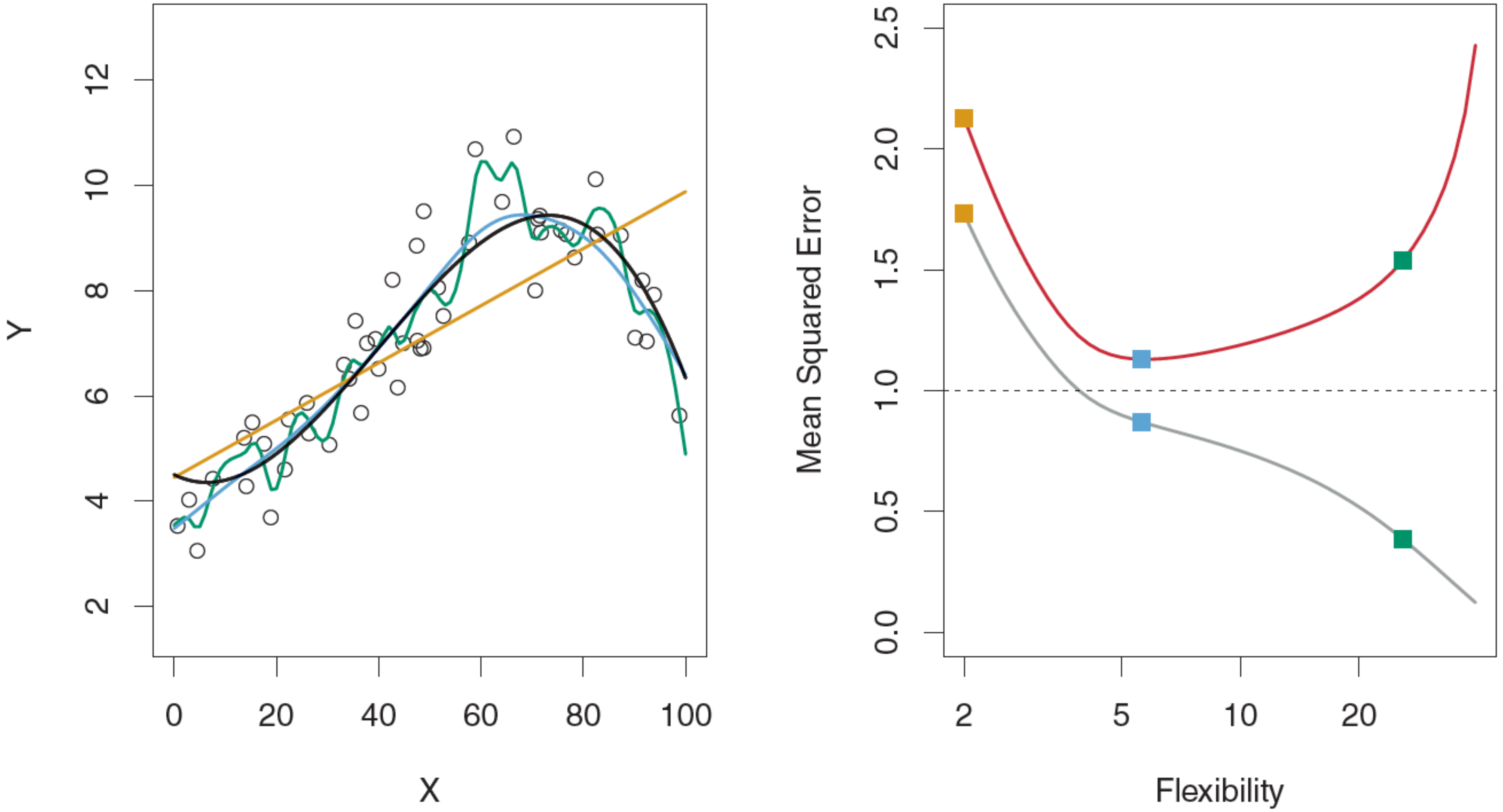
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

y_i	$\widehat{y}_i$	$y_i - \widehat{y}_i$	$(y_i - \widehat{y}_i)^2$
5	5		
6	7		
9	8		
10	1		
14	13		

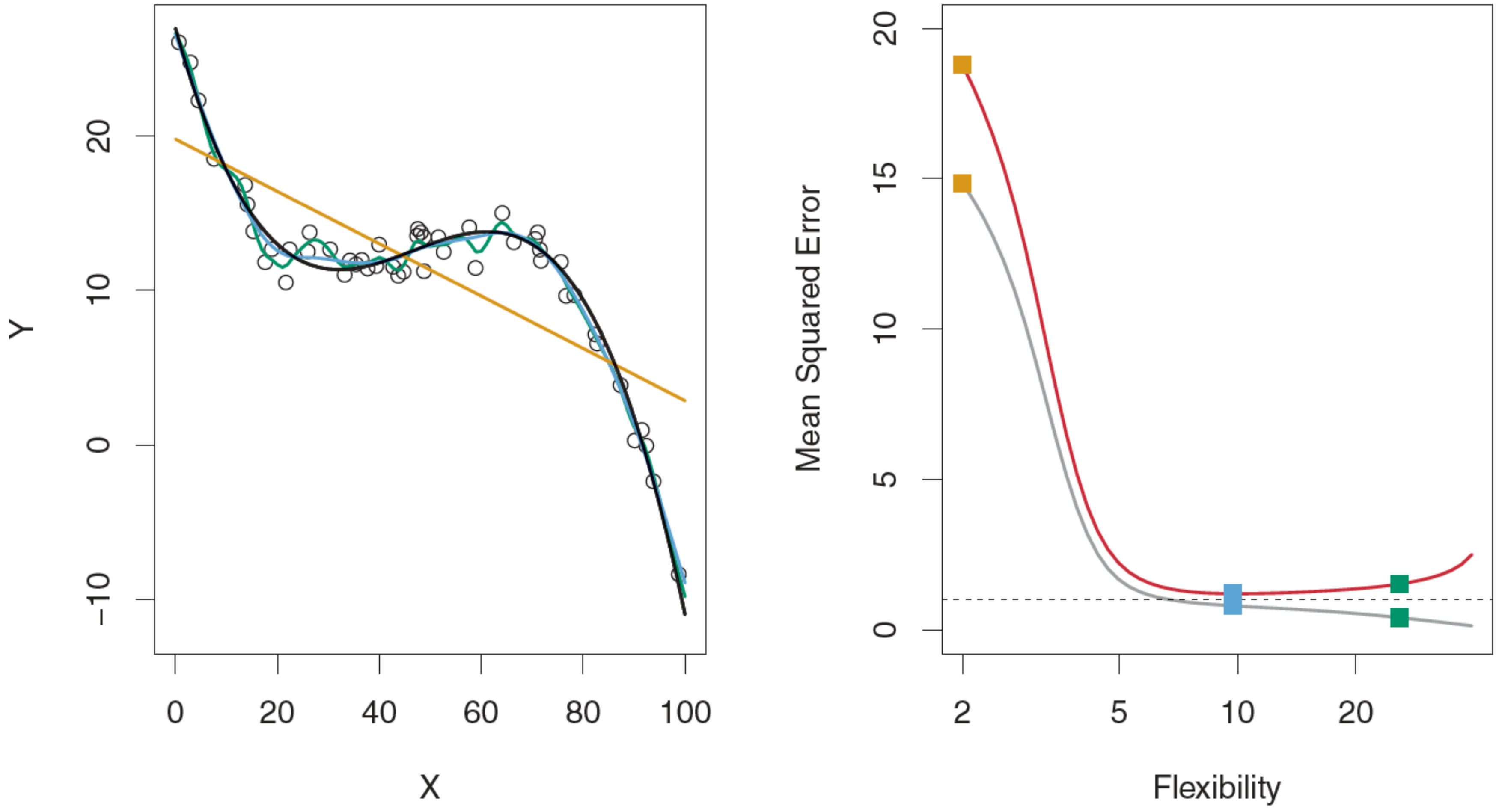
Example: Overfitting (DGP Linear)



Example: Overfitting (DGP Slightly Complicated)



Example: Overfitting (DGP Very Complicated)


Y
X
Mean Squared Error
Flexibility