PS5841

Data Science in Finance & Insurance

Front Matter

Yubo Wang

Autumn 2021

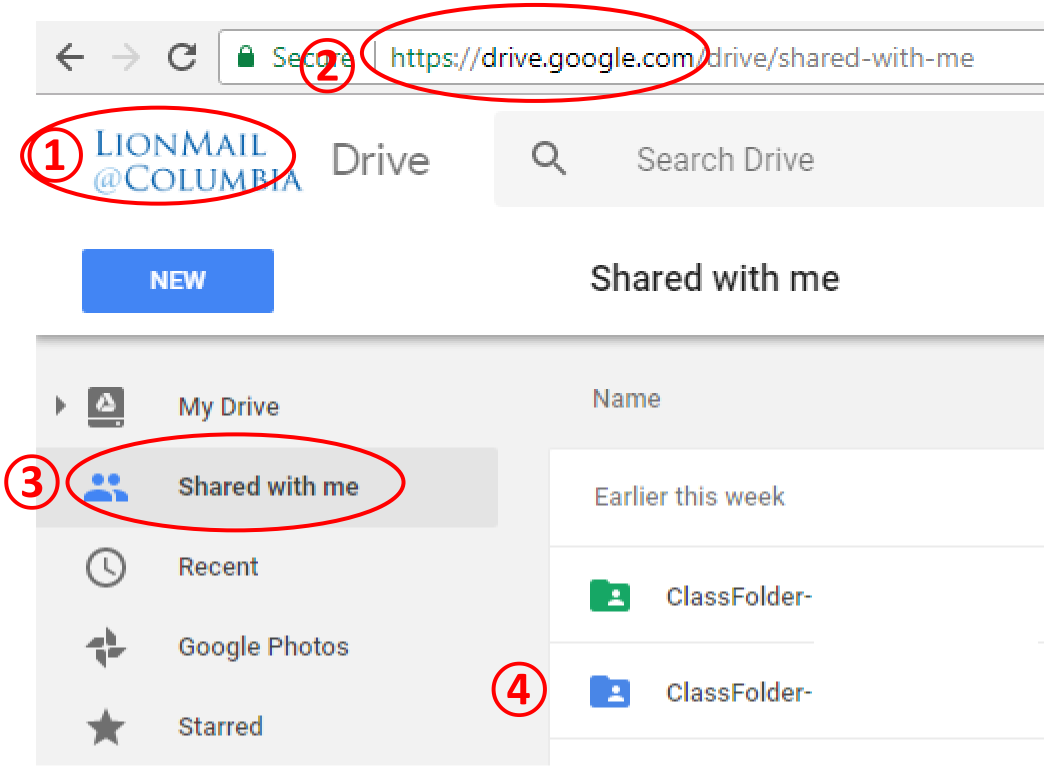
COLUMBIA
UNIVERSITY

School Stuff

- Calendar
  - First class 9/9 (Thu)
  - Midterm 10/19 (Tue)
  - No classes 11/2 (Tue) 11/25 (Thu)
  - Project 12/7, 12/9
  - Last class 12/9 (Thu)
  - Final 12/16 (Thu) 1:10pm-4pm

Class Folder

- Class Folder
  - ①Log into CU email with your UNI
  - ②Go to drive.google.com
  - ③Go to “shared with me”
  - ④Go to ClassFolder-DataSci-Fall2021


https://drive.google.com/drive/shared-with-me

- Class Folder
  ①Log into CU email with your UNI
  ②Then go to https://tinyurl.com/ds2021fall

Course Stuff

- TA
  - Weizhi Hou: wh2484
- Office hour
- Project
- Grading

COLUMBIA UNIVERSITY

Group Project (1)

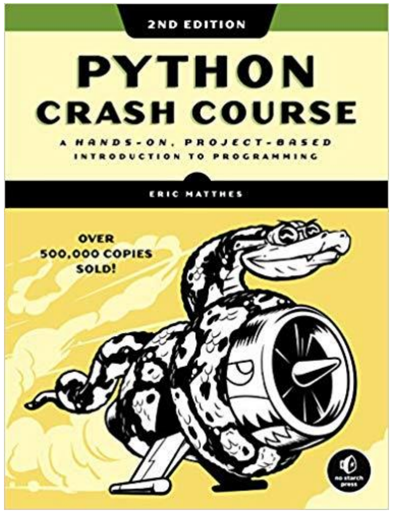
- Who – minimum 3 and maximum 4 people per team
  - Get to know your peers
  - Build on each other's strengths
- What – issues in finance or insurance
- Why – justify its merit for you and your audience
- How –
  - Find/Construct the relevant data set
  - Apply the tools and approaches discussed in the course to appropriately analyze the data to shed light on your questions
  - Educate the class with your informative and lively presentation!
  - Writeup
- When – see the next page

Group Project (2)

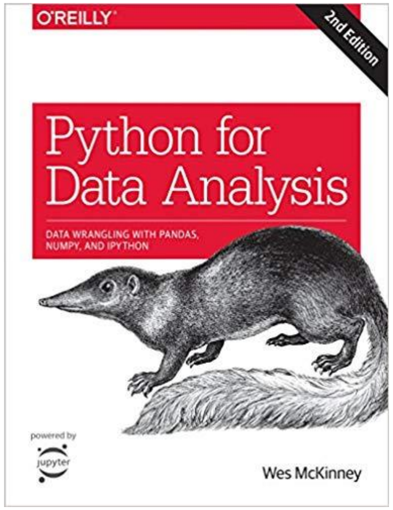
- Keep the dates
  - Project proposal due week 8 (10/28)
  - Draft writeup due week 12 (11/28)
  - Project presentation week 14 (12/07, 12/09)
  - Final writeup due at Final (TBA)

Book Stuff

- Matthes, *Python Crash Course, 2nd ed*, No Starch Press.


2ND EDITION
PYTHON
CRASH COURSE
A HANDS-ON, PROJECT-BASED INTRODUCTION TO PROGRAMMING
ERIC MATTHES
OVER 500,000 COPIES SOLD!
no starch press

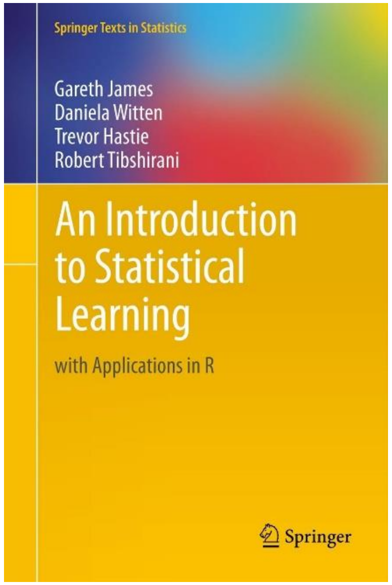
- McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython, 2nd ed.*, O’Reilly Media.



Book Stuff (2)

- James, Witten, Hastie & Tibshirani, *An Introduction to Statistical Learning, with Applications in R,* Springer.

- Goodfellow, Bengio and Courville, *Deep Learning,* MIT Press.



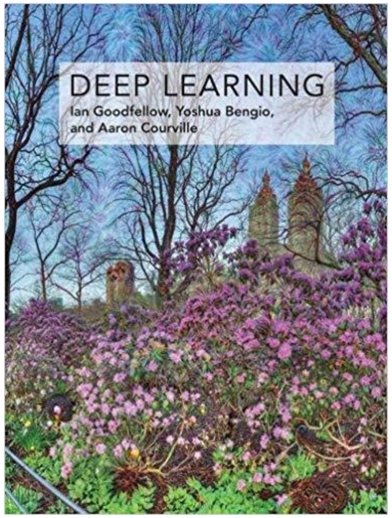
Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

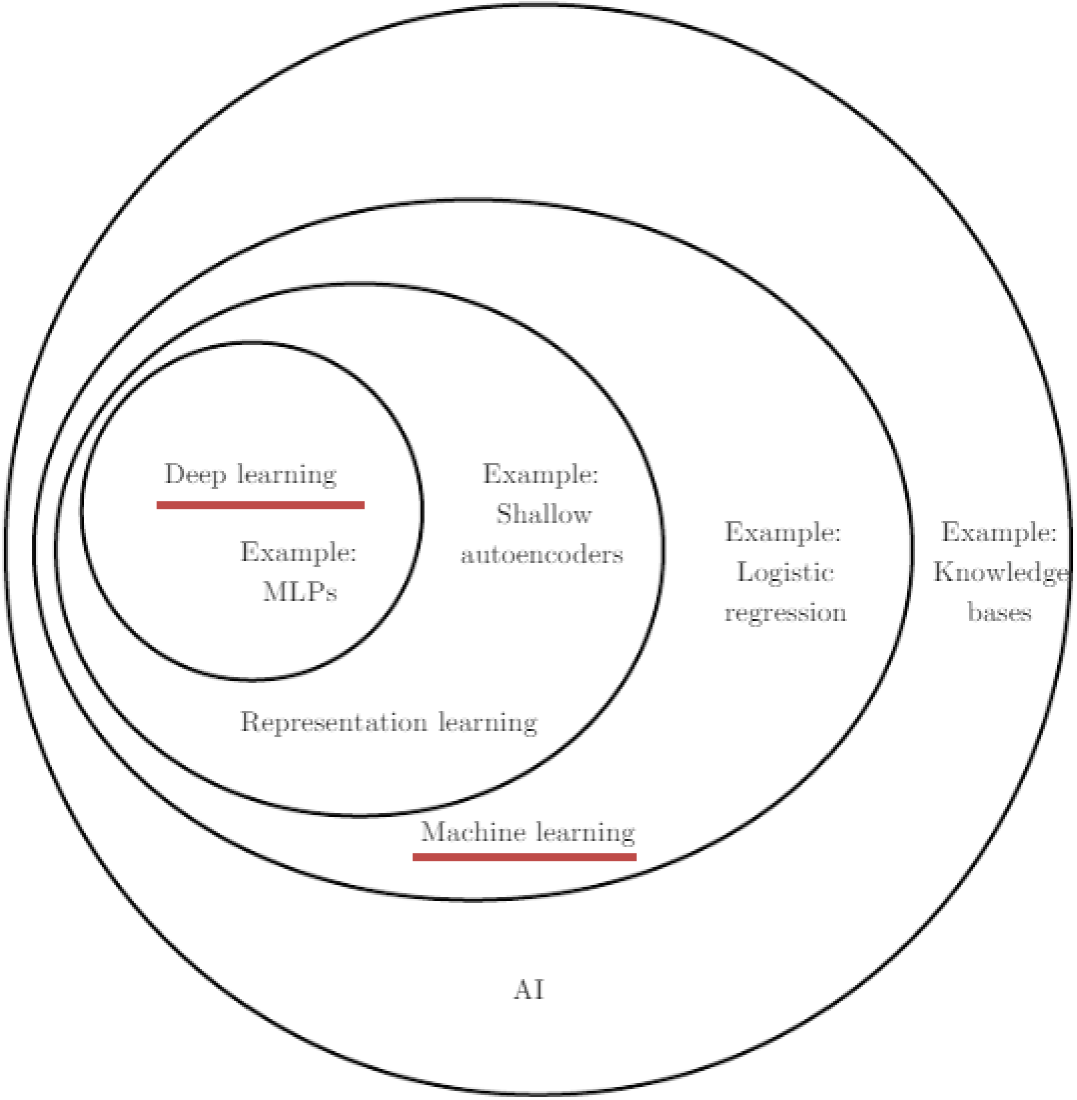


COLUMBIA UNIVERSITY

Computing Environment

- Tools
  - Python, and virtual environments
  - R
  - Spreadsheets
- Modes
  - Terminal
  - Editor and IDLE(e.g. spyder)
  - Jupyter-notebook

Scope



COLUMBIA UNIVERSITY

Coverage

- Supervised learning
  - Regression
  - Classification
- Unsupervised learning
  - Dimension reduction
  - Clustering
- ML methods
  - Regularization
  - Dimension reduction
  - Ensemble learning

COLUMBIA UNIVERSITY

Supervised Learning

- Supervised learning
  - learn to predict Y from X, in essence by estimating $p(Y|X)$
  - Regression: predict quantitative values from input

$$f: R^p \to R$$

  - Classification: specify to which of the categories input belongs

$$f: R^n \to \{1, \dots, k\}$$

Unsupervised Learning

- Unsupervised learning
  - Learn about X, in essence by estimating $p(X)$

COLUMBIA UNIVERSITY

Important Pieces

- Training Set
- Model
- Model Space
- Validation Set
- Test Set

Bias & Variance

- Data generating scheme

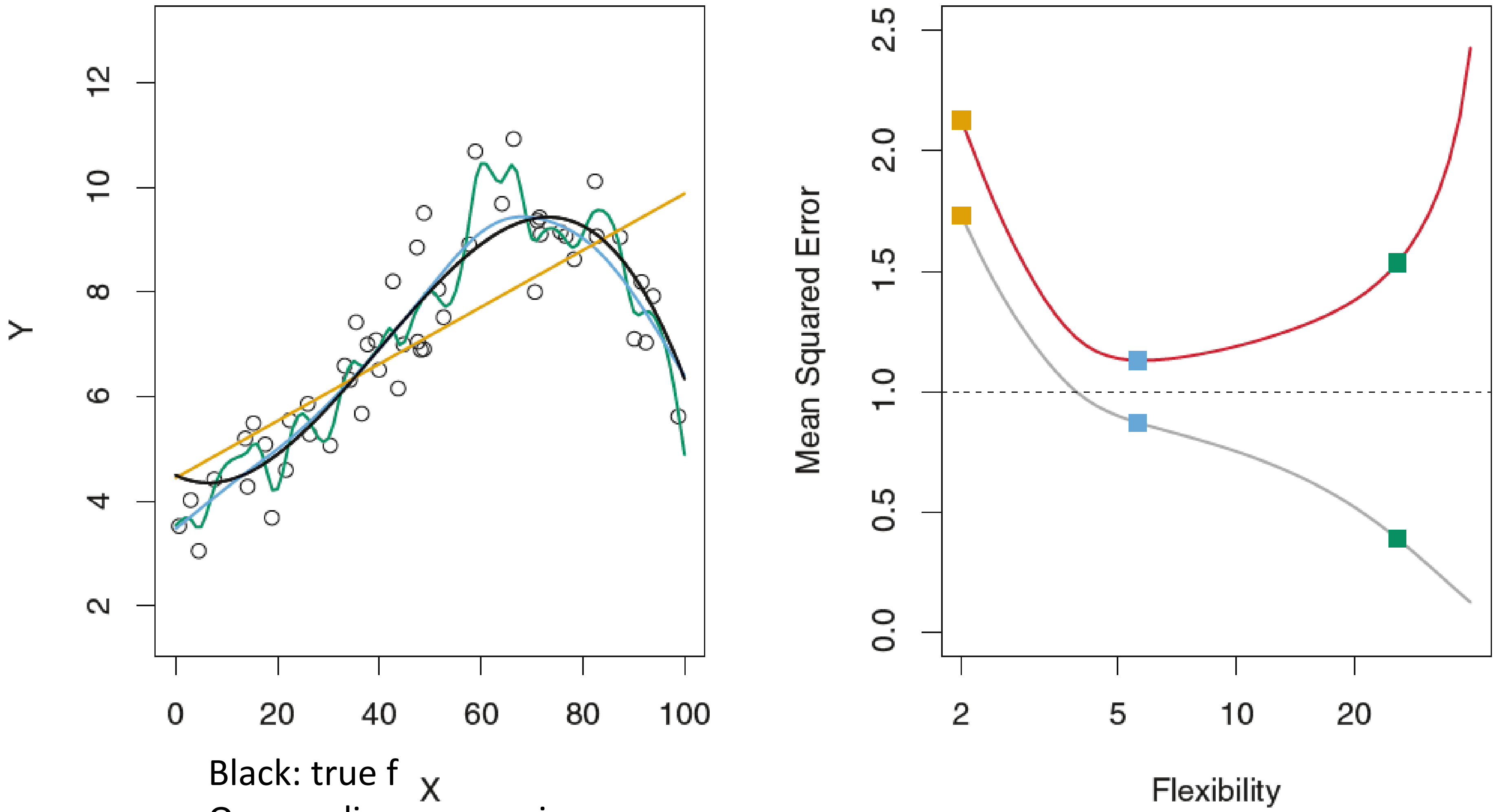
$$y = f(x) + \epsilon$$

- Let (x_0, y_0) be an observation point in the test set, and τ the training set space

$$E_\tau[(y_0 - \hat{y}_0)^2]$$
$$= \sigma_\epsilon^2 + [E_\tau(\hat{y}_0) - f(x_0)]^2 + E_\tau\{[\hat{y}_0 - E_\tau(\hat{y}_0)]^2\}$$
$$= \sigma_\epsilon^2 + bias_\tau^2(\hat{y}_0) + Var_\tau(\hat{y}_0)$$

where $\hat{y}_0 = \hat{f}(x_0)$

An Example



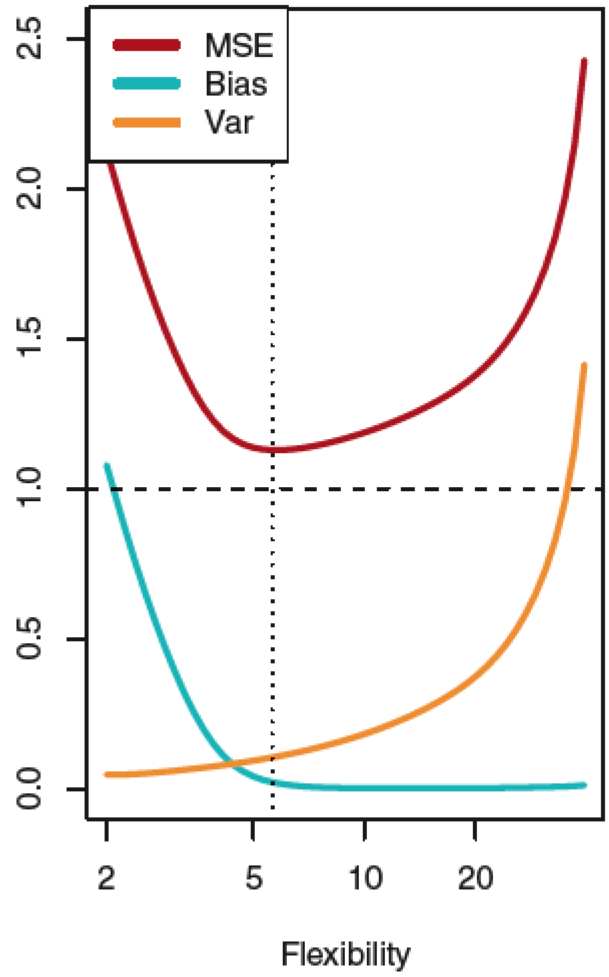
Black: true f

Orange: liner regression

Blue: less flexible smoothing spline

Green: more flexible smoothing spline

Trade-Off Between Bias vs Variance



That was