PS5841

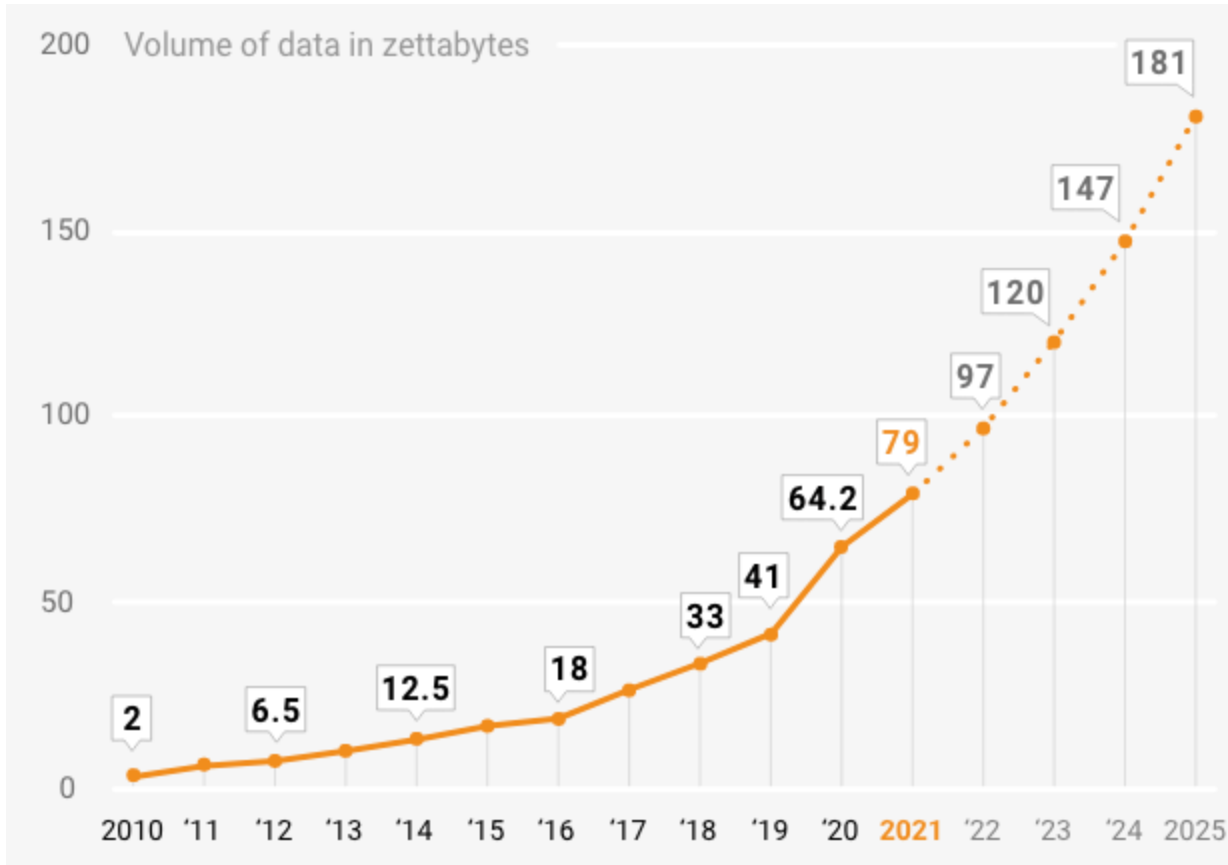# Data Science in Finance & Insurance

Front Matter

Yubo Wang

Autumn 2022

Data …

Source: Statista.com; Amount of data created, consumed and stored

… as assets

Multiple-byte units					V·T·E
Decimal		**Binary**			
Value	**Metric**	**Value**	**IEC**	**Legacy**	
1000	kB kilobyte	1024	KiB kibibyte	KB kilobyte	
1000^2	MB megabyte	1024^2	MiB mebibyte	MB megabyte	
1000^3	GB gigabyte	1024^3	GiB gibibyte	GB gigabyte	
1000^4	TB terabyte	1024^4	TiB tebibyte	TB terabyte	
1000^5	PB petabyte	1024^5	PiB pebibyte	–	
1000^6	EB exabyte	1024^6	EiB exbibyte	–	
1000^7	ZB zettabyte	1024^7	ZiB zebibyte	–	
1000^8	YB yottabyte	1024^8	YiB yobibyte	–	
Orders of magnitude of data					

- Develop coding and database skills
- Master theoretical underpinnings and algorithms of machine learning models
- Perform data-driven decision making with applicable domain knowledge and appropriate machine learning models

School Stuff

- ## Calendar
  - First class   9/6(Tue)
  - Midterm   10/18(Tue)
  - No class   11/8(Thu) 11/24(Thu)
  - Project   12/6(Tue) 12/8(Thu)
  - Last class   12/8(Thu)
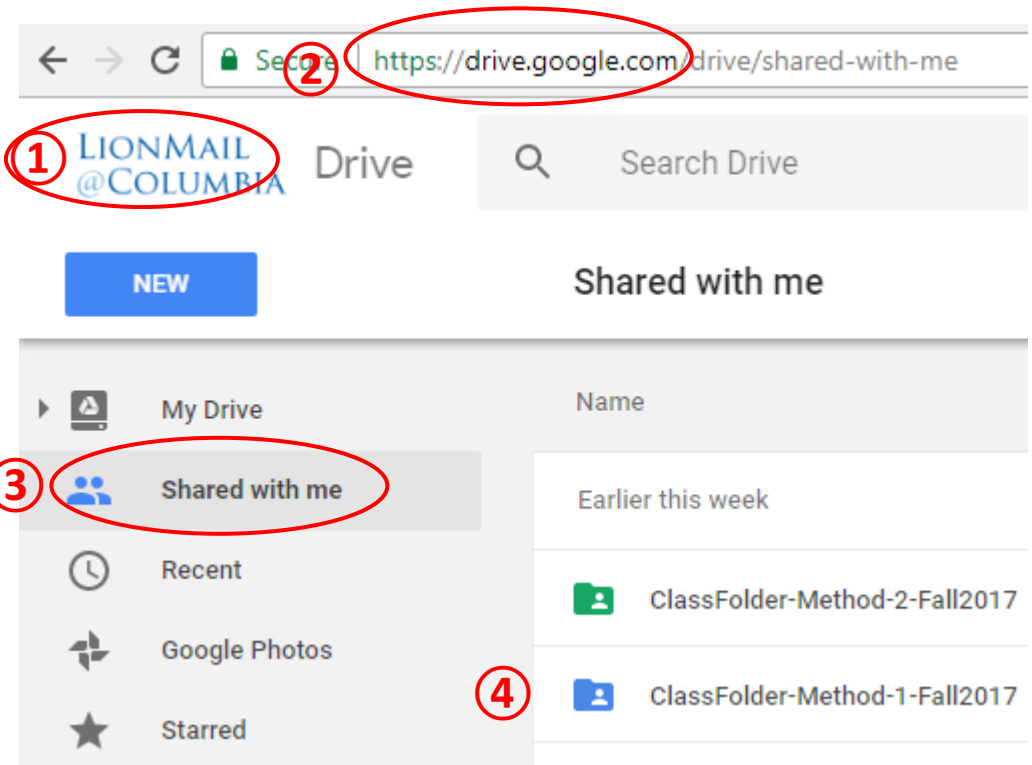 - Final   12/22   9am-12pm Math 407

- ## TA

Woonsup Kim   wk2371

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-Fall2022

    - Class Folder
      ①Log into CU email with your UNI
   ②Then go to   https://tinyurl.com/ds2022fall

Group Project (1)

- Who – around 4 people per team
- What – issues in finance or insurance
- Why – justify its merit for your audience
- How –
  - Find/Construct the relevant dataset(s)
  - 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 – last week of class

Group Project (2)

- **Keep the dates**
  - Project proposal due week 8 (10/27)
  - Draft writeup due week 12 (11/29)
  - Project presentation week 14 (12/6, 12/8)
  - Final documents due at Final
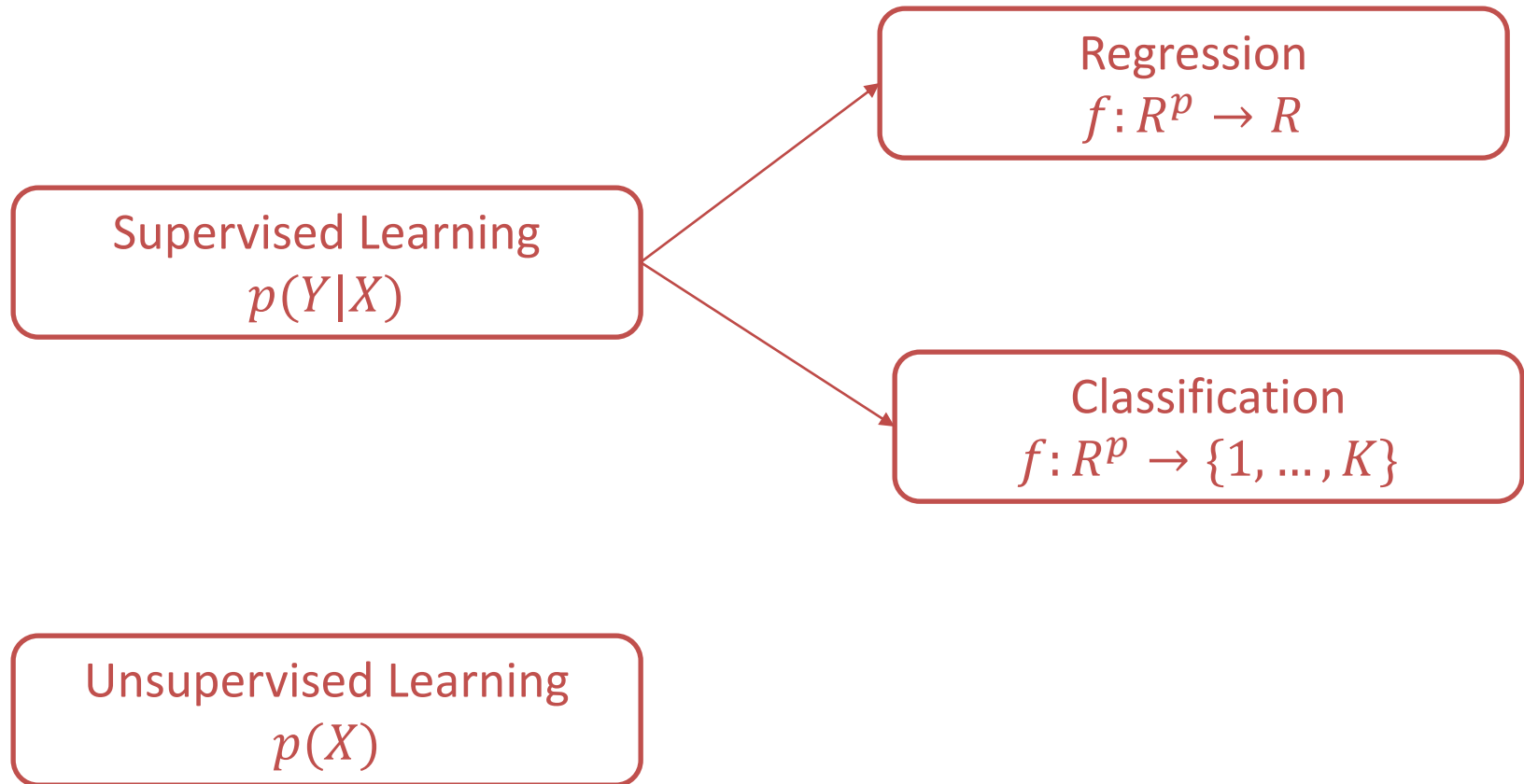   - Slides
    - Codes
    - Writeup
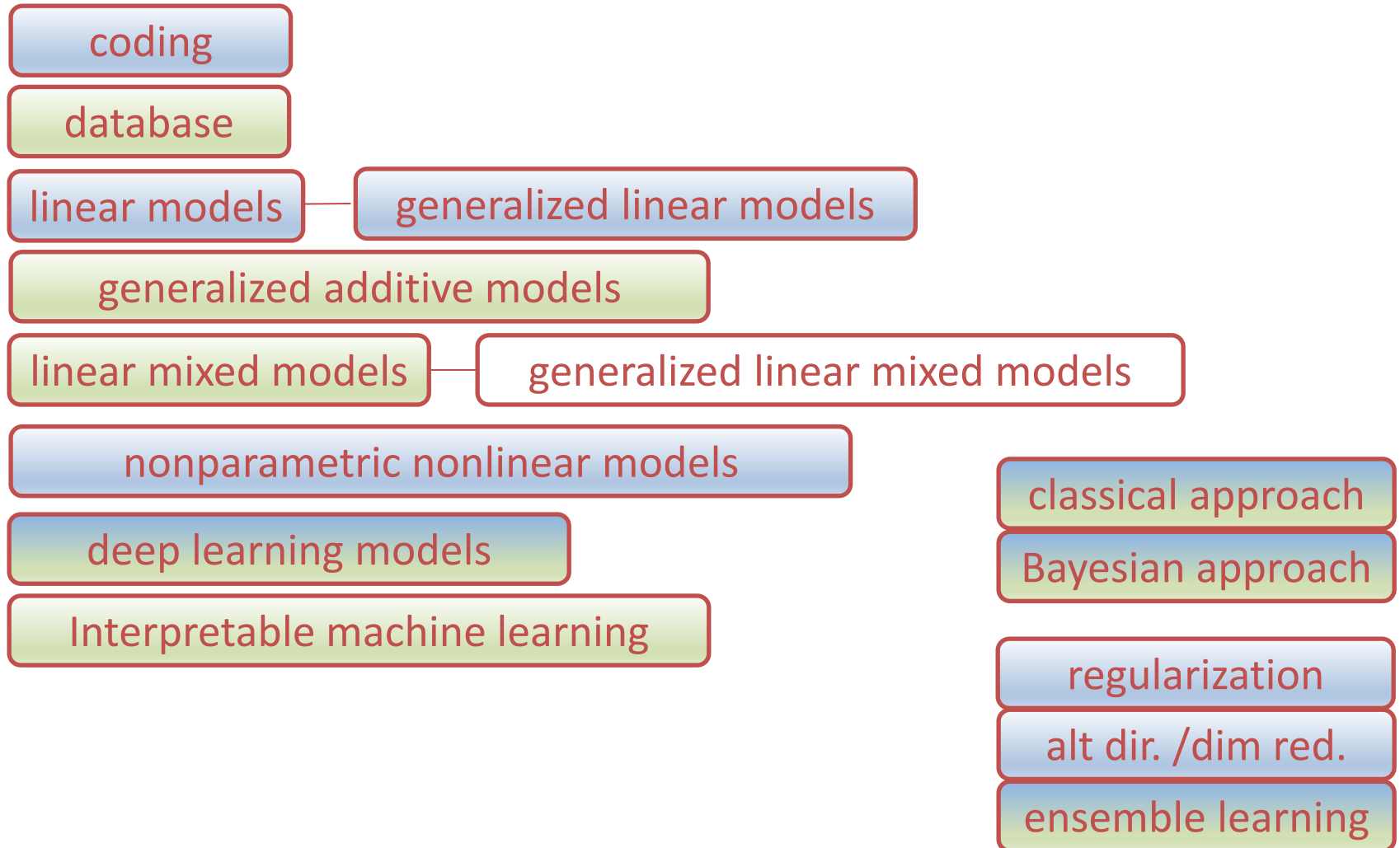
Book Stuff
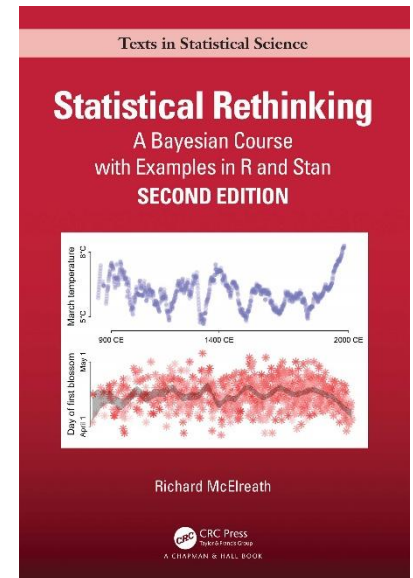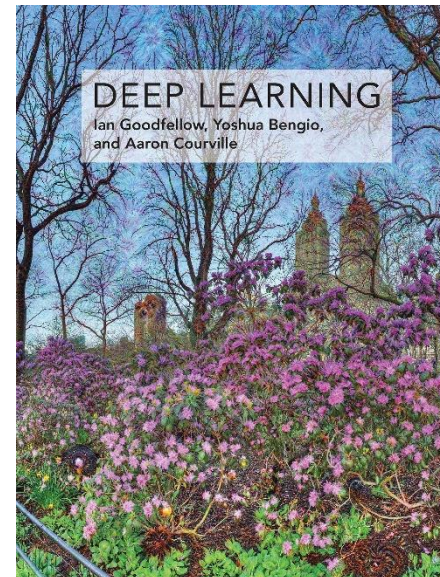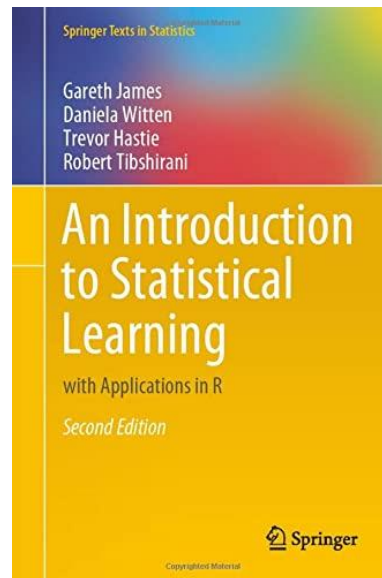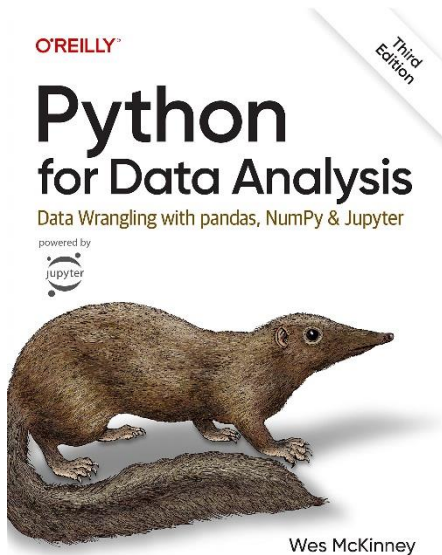
Computing

- Tools
  - Python in virtual environment(s)
  - R
  - Spreadsheet
- Modes
  - Terminal
  - Editor
  - Jupyter

- Open Source Python Packages
  - Numpy, Pandas
  - Scipy
  - Matplotlib, Plotnine
  - Scikit-learn, Statsmodels
  - Tensorflow/Keras

Select Learning Schemes

Supervised Learning
$p(Y|X)$

Regression
$f: R^p \rightarrow R$

Classification
$f: R^p \rightarrow \{1, \ldots, K\}$

Unsupervised Learning
$p(X)$

Course Map (DS & ADS)

coding

database

linear models — generalized linear models

generalized additive models

linear mixed models — generalized linear mixed models

nonparametric nonlinear models

deep learning models

Interpretable machine learning

classical approach

Bayesian approach

regularization

alt dir. /dim red.

ensemble learning

That was

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK