Meetup 1

Data 607

Peter Kowalchuk

Welcome to the class

Introductions

What’s important in Data Science?

What’s important in Data Science?

What I found online

1. Domain Knowledge
2. **Data Collection and Cleaning**
3. Data Exploration and Visualization
4. Statistical Analysis
5. Machine Learning
6. Feature Engineering
7. Model Selection and Evaluation
8. **Programming and Tools**
9. **Data Ethics and Privacy**
10. **Communication Skills**
11. **Continuous Learning**
12. **Problem-Solving**

What’s important in Data Science?

Important, but not the focus of this course

Developing models, statistical, analytical AI, gen AI or any other

Building ML pipelines

What’s important in Data Science?

Our focus in this course

Building pipelines to acquire data

Data cleaning, eliminating the noise while minimizing the loss of useful information

Supplementing data with additional external sources of information

Transforming the data to conform to the machine learning algorithm’s requirements

Some thoughts

Mindset for this course, program, and career

If you don’t get feedback, your confidence grows much faster than your accuracy.

Philip Tetlock

The Art and Science of Prediction

Experts in their fields tend to be motivated by criticism, and see it as a sign of how well they’re progressing toward their goal.

Sue Shellenger

Former Wall Street Journal

Feedback

Collaborating in a team

- Collaboration Repositories
  - Personal GitHub
  - Class GitHub:
    - https://github.com/pkowalchuk/CUNY-Data_607-Spring_2025

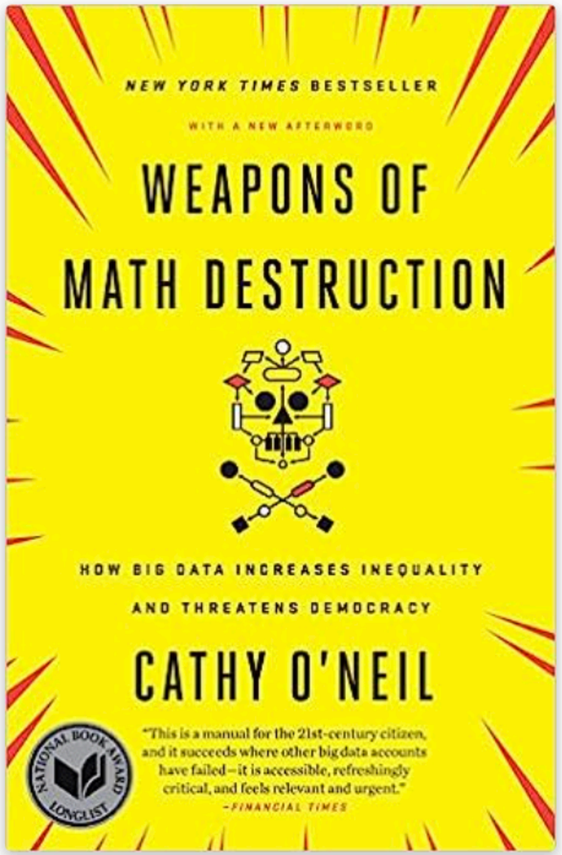
Course Overview

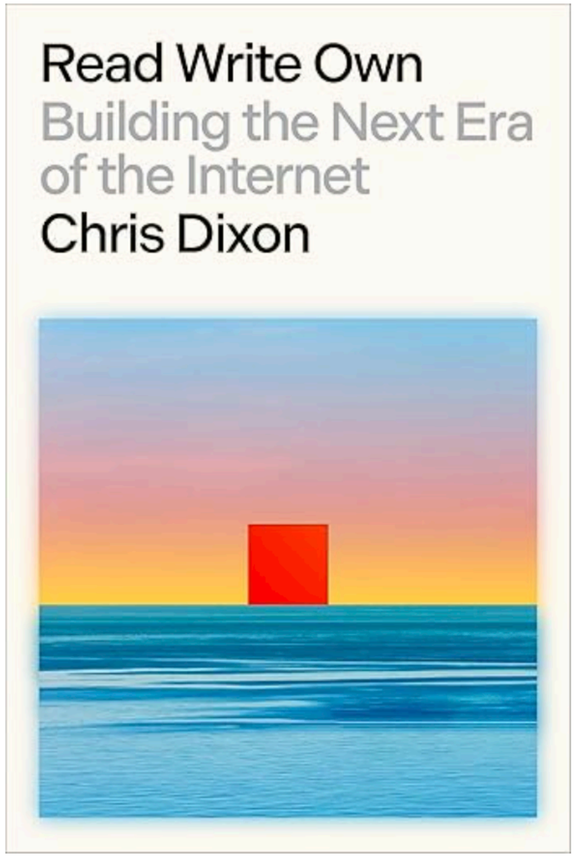
Syllabus and Schedule

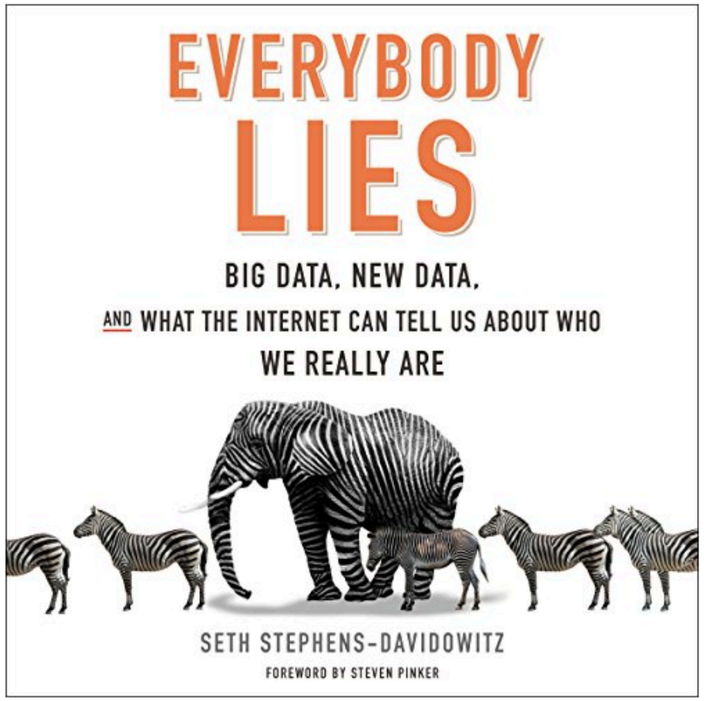
Unit	Topic	Core Readings	Assign	Projs.	Final Proj. Prop.	Final Proj.	Final Proj. Press.	Discuss	Data Sc.i in Cox.	Tidy Verse
Wk. 1	Data in R	R for Data Science (2e), https://r4ds.hadley.nz/ chapters 1, 2, 3, 7,14,15,27,28	50					20		
Wk. 2	Data in SQL	R for Data Science (2e), https://r4ds.hadley.nz/ chapter 8, 9, 10, 23	50							
Wk. 3	Data structures in SQL	R for Data Science (2e), https://r4ds.hadley.nz/ chapters 14, 15, 16, 17, 18, 19	50							
Wk.4	Data structures in R	R for Data Science (2e), https://r4ds.hadley.nz/ chapters 11, 12, 13, 20		90						
Wk. 5	Exploratory Data Analysis	R for Data Science (2e), https://r4ds.hadley.nz/ chapters 4, 5, 6	50					40		
Wk. 6	Data Transformations	Feature Engineering and Selection, http://www.feat.engineering/ chapter 1		90						
Wk. 7	Data formats	R for Data Science (2e), https://r4ds.hadley.nz/ chapter 25	50	20						
Wk. 8	Data ownership	R for Data Science (2e), https://r4ds.hadley.nz/ chapter 26		70						
Wk. 9	Working with REST APIs	httr quickstart vignette,: https://cran.r-project.org/web/packages/httr/vignettes/quickstart.html	50					40		25
Wk. 10	Natural Language Processing	Text Mining w/ R, https://www.tidytextmining.com/, ch 1-4	50							
Wk. 11	Recommender Systems	Mining Massive Datasets, http://www.mmds.org/, ch 9	50					40		
Wk. 12	Data Architectures	Selected readings from web			20					
	Thanksgiving									
Wk. 13	Cloud Technologies	No Readings		90						15
Wk. 14	Parallel Computing	No Readings							50	
Wk. 15	Final Presentation	No Readings				150	30			
		Total Points	300	270	20	150	30	140	50	40
		Total Percents	30%	27%	2%	15%	3%	14%	5%	4%

We will be working with data

Ethics, privacy and responsible use

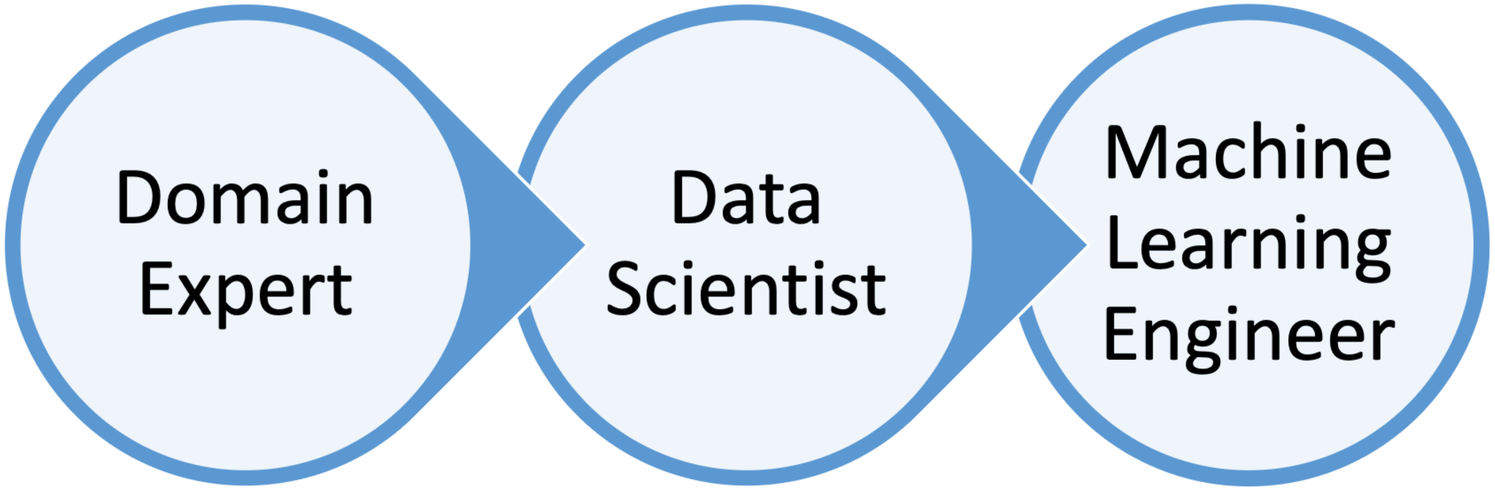






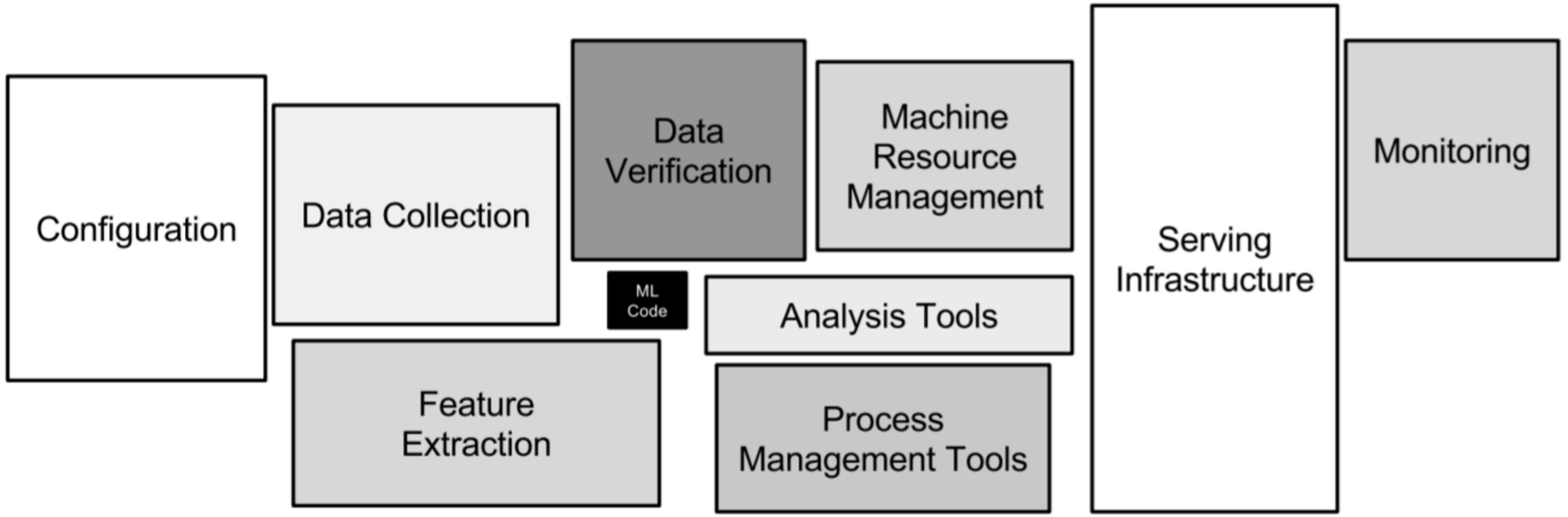
Data Science

Dissecting the discipline



Data Acquisition and Management

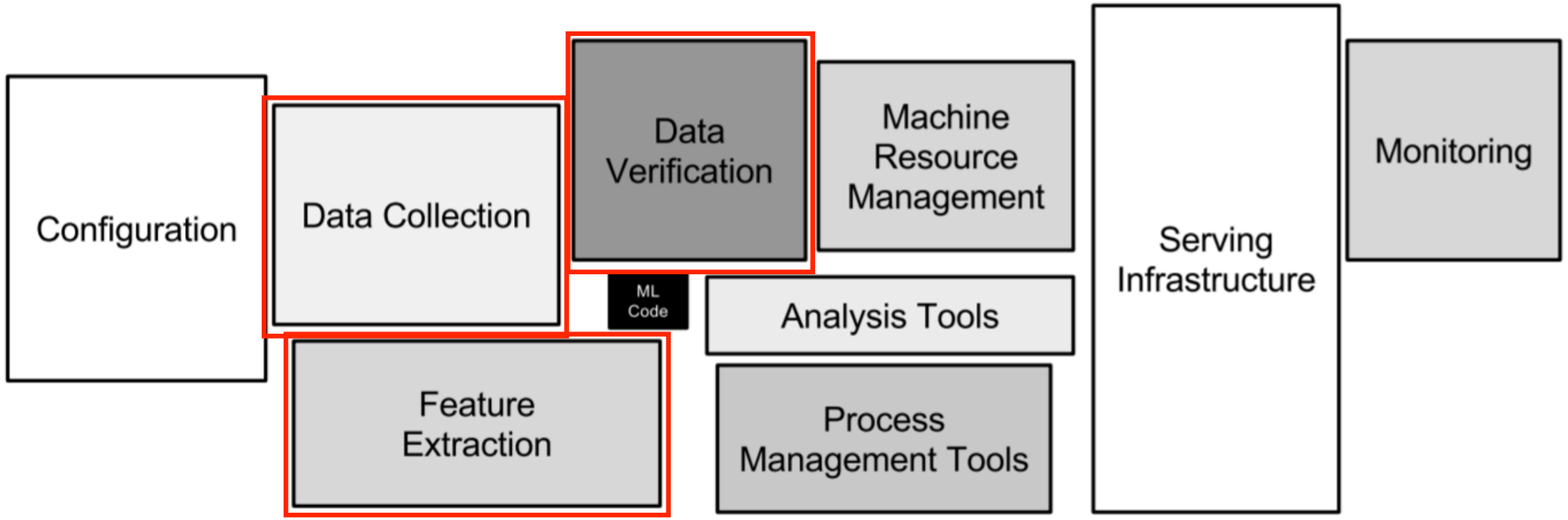
Where does it fit?



Source: D. Sculley et.al., “Hidden Technical Debt in Machine Learning Systems,” NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems, Pages 2503-2511, Dec 7, 2015. https://papers.nips.cc/paper/5656-hidden-technical-debt-in- machine-learning-systems.pdf.

Data Acquisition and Management

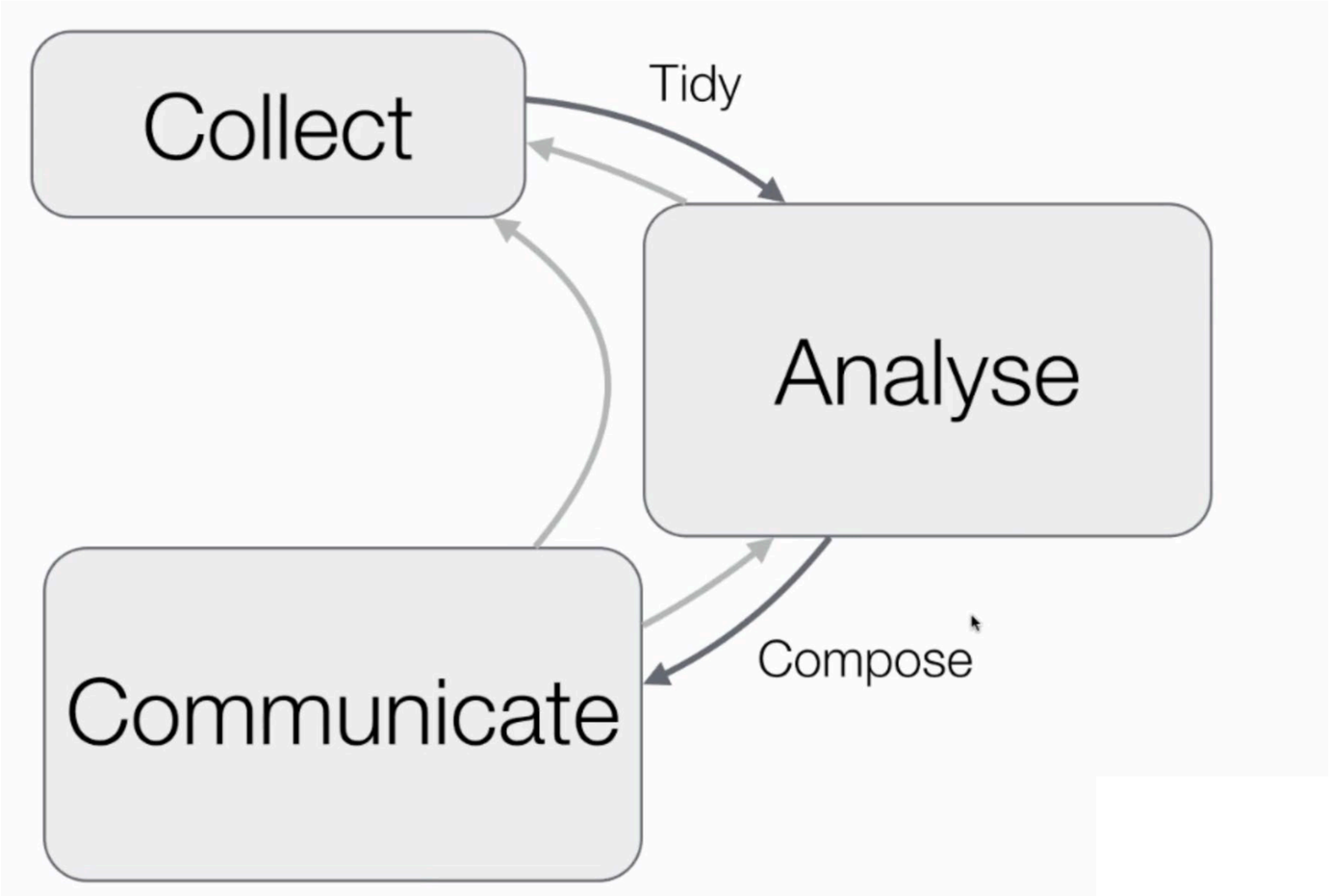
Where does it fit?



Source: D. Sculley et.al., “Hidden Technical Debt in Machine Learning Systems,” NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems, Pages 2503-2511, Dec 7, 2015. https://papers.nips.cc/paper/5656-hidden-technical-debt-in- machine-learning-systems.pdf.

Data Acquisition and Management

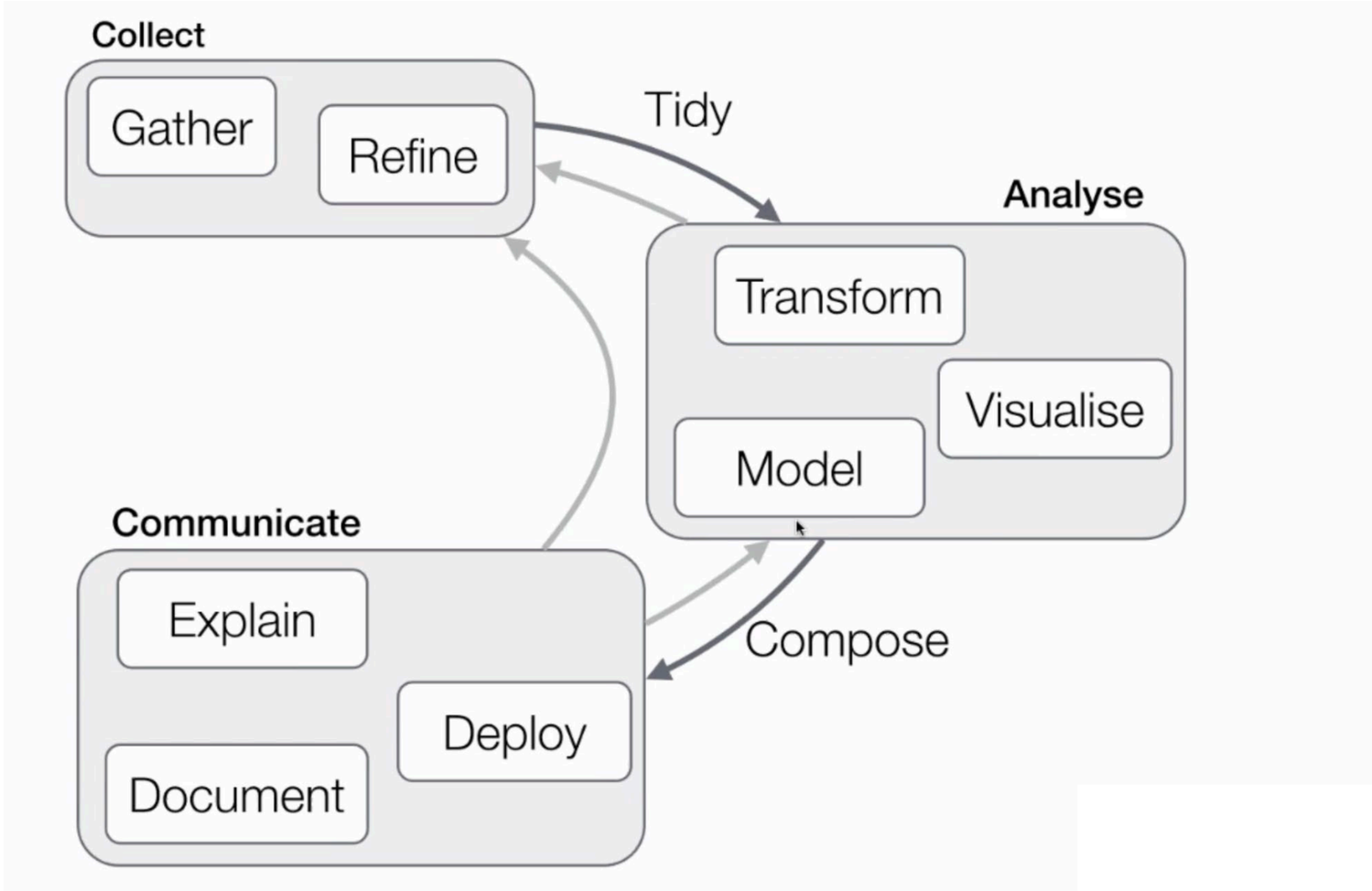
What we will do in this course



Presentation by Hadley Wickham on “The Grammar and Graphics of Data Science”, https://www.rstudio.com/resources/webinars/archives/

Data Acquisition and Management

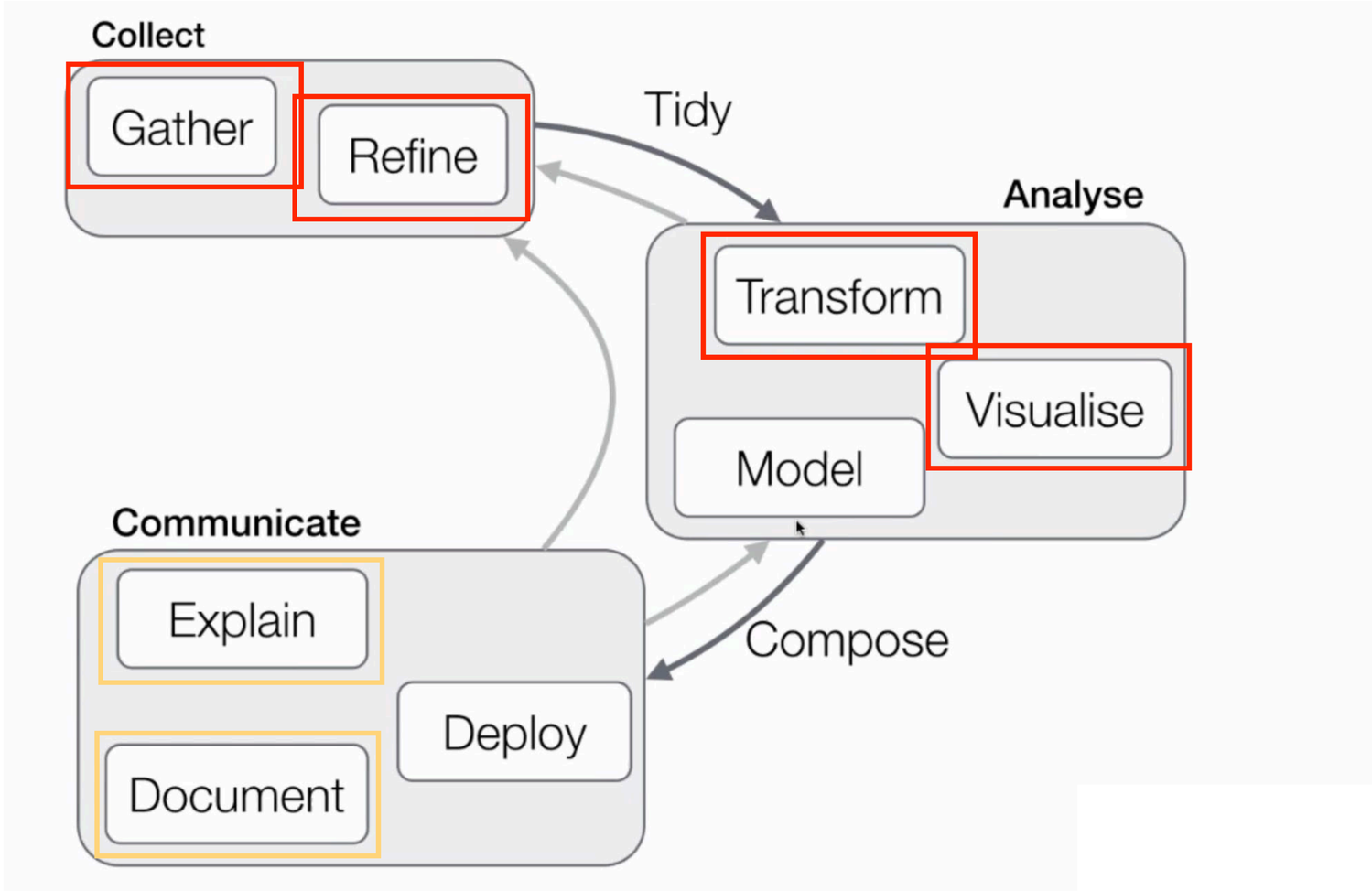
What we will do in this course



Presentation by Hadley Wickham on “The Grammar and Graphics of Data Science”, https://www.rstudio.com/resources/webinars/archives/

Data Acquisition and Management

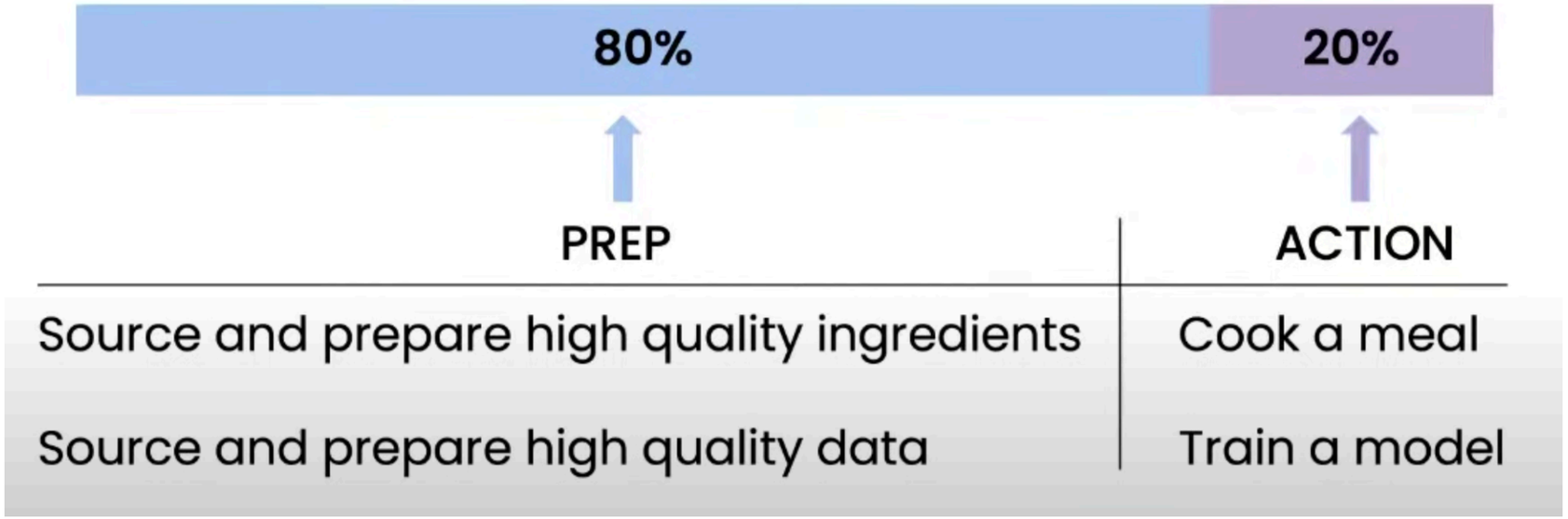
What we will do in this course



Presentation by Hadley Wickham on “The Grammar and Graphics of Data Science”, https://www.rstudio.com/resources/webinars/archives/

Is Data Acquisition and Management Important?

Data is Food for AI



Ng,Andrew."ChatwithAndrewonMLOps:FromModel-centric to Data-centric AI." YouTube, uploaded by Deep Learning AI, 24 Mar. 2021
https://www.youtube.com/watch?v=06-AZXmwHjo

Welcome to the course

Lets look at an example and set the expectation for all deliverables