R Basics + Exploring data with R

IMT 573A - Data Science 1 - Theoretical Foundations

6-Oct-2020 (Week 2, Day 1)

Our Zoom class sessions will be recorded.

Today's Topics

1. Review
2. R basics - demo
3. Exploratory data analysis - key steps
4. Lab

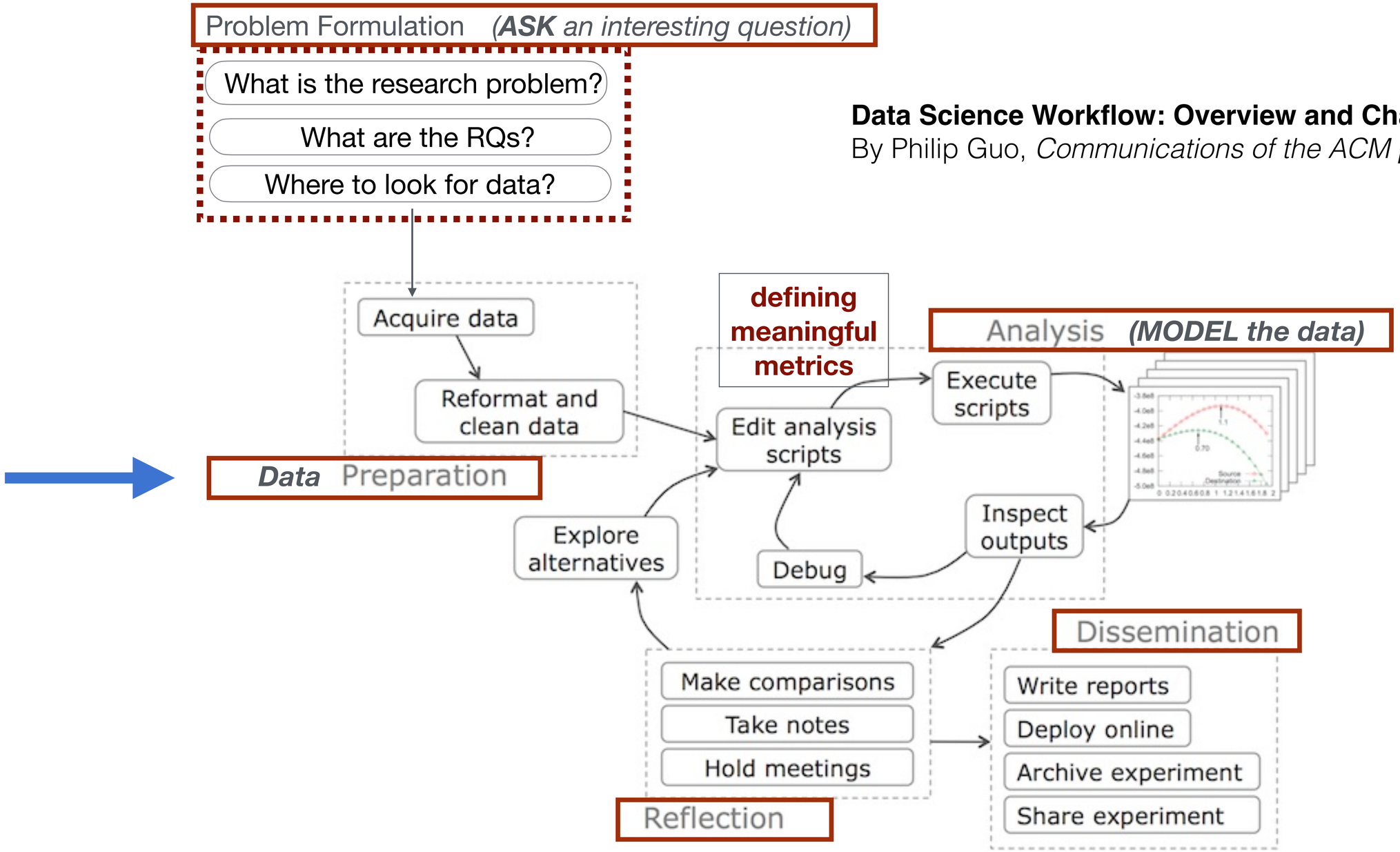
Questions? If you have questions, feel free to unmute and ask or type them on the Zoom chat box.

Learning Objectives:

- *R and RStudio*
- *Basic R: variables, functions, loops*
- *Vectorized operations*
- *Indexing*
- *Basics of data programming: data frames, variables*
- *loading and saving data*

Data Science Workflow

Data Science Workflow: Overview and Challenges
By Philip Guo, *Communications of the ACM* ***[Required Reading]***



Exploring Data: A checklist

1. Interpret your data
2. Formulate your question
3. Read in your data
4. Examine your data, look at the top and bottom of your data, look at structure
5. Tidy data (Data cleaning)
6. Try the easy solution first
7. Challenge your solution, validate with external data
8. Follow up with new interesting questions/directions

Think about your data! Think about your question!

Peng, Roger D. Exploratory Data Analysis with R

Exploring Data: Example

Air pollution data from the US Environmental Protection Agency (EPA)

https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw

Data on Canvas

Hourly Data

Criteria Gases

Year	Ozone (44201)	SO2 (42401)	CO (42101)	NO2 (42602)
2020	hourly_44201_2020.zip 791,370 Rows 5,856 KB As of 2020-05-19	hourly_42401_2020.zip 431,530 Rows 2,901 KB As of 2020-05-19	hourly_42101_2020.zip 181,190 Rows 1,438 KB As of 2020-05-19	hourly_42602_2020.zip 288,225 Rows 2,356 KB As of 2020-05-19
2019	hourly_44201_2019.zip 9,152,376 Rows 68,914 KB As of 2020-05-19	hourly_42401_2019.zip 3,879,123 Rows 25,763 KB As of 2020-05-19	hourly_42101_2019.zip 2,163,948 Rows 16,217 KB As of 2020-05-19	hourly_42602_2019.zip 3,565,238 Rows 28,527 KB As of 2020-05-19
2018	hourly_44201_2018.zip 9,474,271 Rows 71,281 KB As of 2020-05-19	hourly_42401_2018.zip 3,865,278 Rows 25,801 KB As of 2020-05-19	hourly_42101_2018.zip 2,278,236 Rows 16,965 KB As of 2020-05-19	hourly_42602_2018.zip 3,547,695 Rows 28,008 KB As of 2020-05-19

Interpret Data

- Acquiring domain knowledge: To understand the context of the data
- Understanding the data schema

Interpret Data

Gathering domain knowledge requires outside research (not just looking at the data files or running programs and building models on the data)

- Acquiring domain knowledge:

What's the problem domain of the data?

What are the topics that are relevant to the problem domain domain of the data?

Ask questions about the data

https://data.seattle.gov/Permitting/Land-Use-Permits/ht3q-kdvx

Seattle | Open Data Program | TechTalk Blog | Public Records Requests | Other City Data | Sign In

Land Use Permits

Land Use permits that are in progress or that have been issued in Seattle.

Find in this Dataset | More Views | Filter | Visualize | Export | Discuss | Embed | About

PermitNum	PermitClass	PermitClassMapped	PermitTypeMapped	PermitTypeDesc	Description	Housingl
3009387-LU	Multifamily	Residential	Master Use Permit		Land use application to adjust the bo...	
3020870-EG	Multifamily	Residential	Early Design Guidance	Streamlined Design Re...	Early Design Guidance for: Land use ...	
3018857-LU	Commercial	Non-Residential	Master Use Permit		Streamlined Design Review for a four...	
3022144-LU	Single Family/Duplex	Residential	Master Use Permit		Land Use Application to subdivide on...	
3006054-LU	Single Family/Duplex	Residential	Master Use Permit		CANCELED PER APPLICANT'S REQUE...	
3034240-LU	Industrial	Non-Residential	Master Use Permit		Proposed New Building - Commercia...	
3017309-LU	Single Family/Duplex	Residential	Master Use Permit		Shoreline Substantial Development ...	
3019783-LU	Single Family/Duplex	Residential	Master Use Permit		Land Use Application to subdivide on...	
3026890-EG	Multifamily	Residential	Early Design Guidance	Streamlined Design Re...	Early Design Guidance for: Streamlin...	
3026219-LU	Multifamily	Residential	Master Use Permit		Land use application to adjust the bo...	
3032146-EG	Commercial	Non-Residential	Early Design Guidance	Design Review	Design Review Early Design Guidanc...	

< Previous Next >

Showing rows 1 to 100 out of 19,660

Preview of land use permits data from the City of Seattle

Domain knowledge: *land use permits*

Interpret Data

- Understanding the data schema: *What is represented by the rows and columns of the data (data schema)? What's the context for those values?*

GUIDING QUESTIONS:

- What meta-data is available for the dataset? (Data about the data, e.g., how big is the data, summary of the dataset, etc.)
- Who created the data set? Where does it come from? Questions of bias, provenance, or other subtleties about the data can surface.
- What features the dataset have? (Understanding the columns)
- Do you understand all the terms or jargon associated with the data?

Let's answer these questions with the Seattle land use permit dataset (breakout rooms)

https://data.seattle.gov/Permitting/Land-Use-Permits/ht3q-kdvx

Interpret Data ~~> Using data to Answer Questions

As a data scientist, you will be responsible for translating from various domain questions to specific observations & features in your data set.

The Seattle Times

Search

Real Estate

Log In | Subscribe

Business | Local Business | Local News | Real Estate

As affordable housing shrinks in Seattle, permitting delays keep apartment projects in limbo for months

Dec. 27, 2019 at 6:01 am | *Updated Dec. 27, 2019 at 9:13 pm*

Are there permit delays in Seattle?

What are the worst instances of permit delays?

R basics

Code in R studio
Lab 1: Upload on Canvas