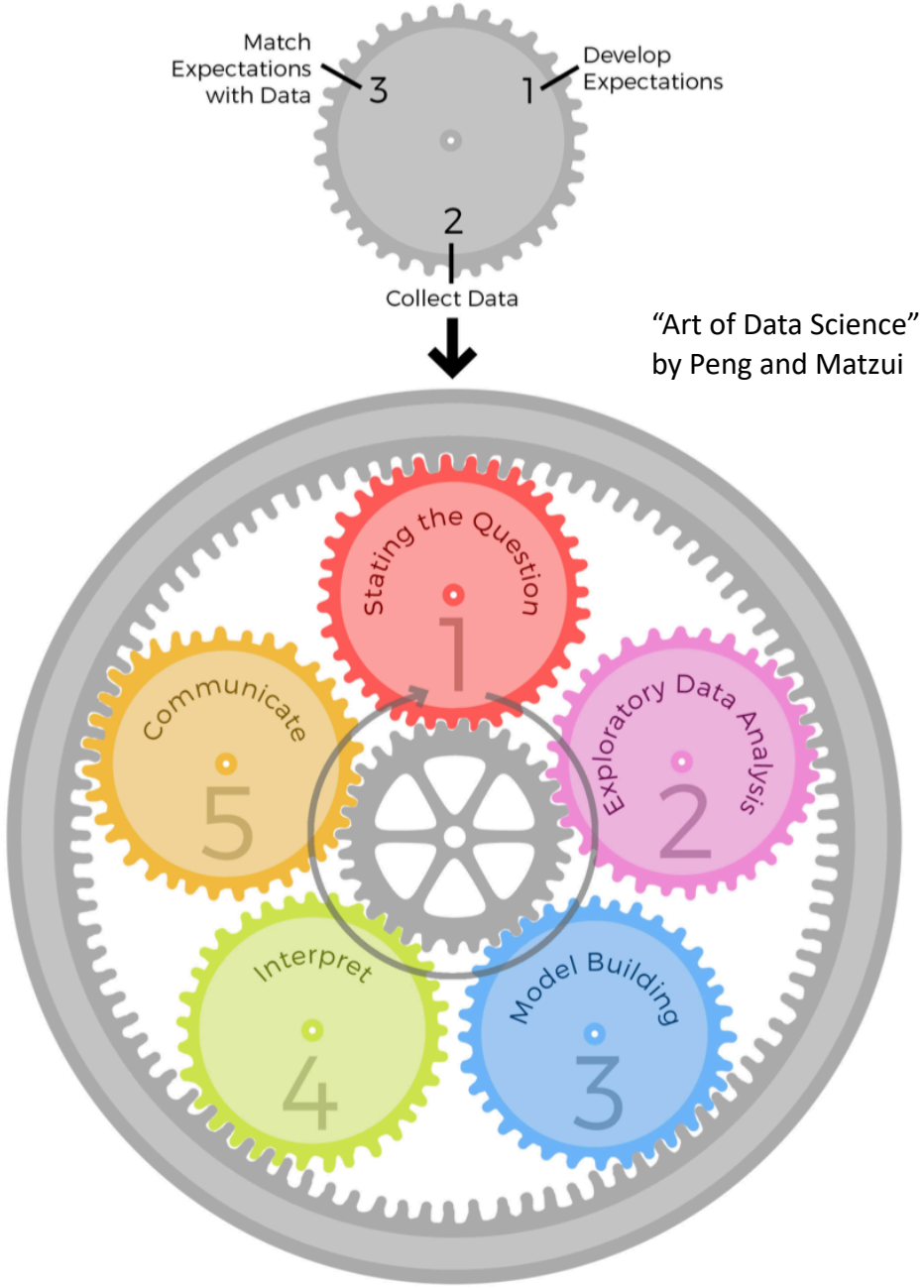
EDA

Epicycles of Analysis

1. Setting Expectations,

2. Collecting information (data), comparing the data to your expectations, and if the expectations don't match,

3. Revising your expectations or fixing the data so your data and your expectations match.



"Art of Data Science" by Peng and Matzui

Epicycles of Analysis

"Art of Data Science"
by Peng and Matzui

Types of questions:

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

Characteristics of a Good Question

"Art of Data Science"
by Peng and Matzui

"The question should be of ***interest*** to your audience, the identity of which will depend on the context and environment in which you are working with data."

"… check that the question has ***not already been answered***…."

"The question should also stem from a ***plausible*** framework."

"***Specificity*** is also an important characteristic of a good question."

"… what will happen when you translate it into a ***data problem***?"

EDA checklist

"Art of Data Science"
by Peng and Matzui

1. Formulate your question
2. Read in your data
3. Check the *"packaging"*
4. Look at the top and the bottom of your data
5. Check your "*n*"s
6. Validate with at least one external data source
7. Visualize

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

Design - Example

The general goal is to learn more about the time it takes for each of us (and our friends – even those in other cities -) to **commute to work**.

Think: how to collect data on our commute times in a systematic, but not intrusive, manner.

Possible motivations:

“When should I leave home to get to meetings on time?”

“When should I leave to arrive on time for the first meeting?”

“When is the earliest time that I can schedule the first meeting of the day?”

Problem

The general goal is to learn more about the time it takes for each of us (and our friends – even those in other cities -) to **commute to work**.

To be defined:

- Where to store the data.
- Indicating commute method.
- What are the covariates that we need to help us understand and model the commute times?

Design Questions

The general goal is to learn more about the time it takes for each of us (and our friends – even those in other cities -) to **commute to work**.

- capture “special circumstances”?
- fixed or random variation?
- estimate the average commute time for each commute method.
- what do you think the tail of the distribution looks like?
- what type of distribution?

Additional Questions

The general goal is to learn more about the time it takes for each of us (and our friends – even those in other cities -) to **commute to work**.

- Is it a fixed study designed to answer a specific question (i.e. what is the mean commute time?) within some bound of uncertainty? Or

- Is it an ongoing study where data will be continuously collected and actions will be continuously adapted as new data are collected