2-minute activity

What's one thing you found helpful in the course pre-read?
What's one thing you found confusing?
Type it into the Zoom chat.

Purposes

- Clarify confusing points
- Help improve the pre-read (we plan to share it publicly)

Breakout activity [20 minutes]

Problem: modeling home sale price

Data description: see doc linked in Zoom chat

What steps would you take to:

- Build an algorithmic model to predict home sale price
- Build a data model to understand whether a pool adds value to a home

Data modeling vs algorithmic modeling

Data modeling (most statisticians)

Algorithmic modeling (most of ML)

Goal: understand nature (the data)

Validate the model (yes/no) using goodness-of-fit etc.

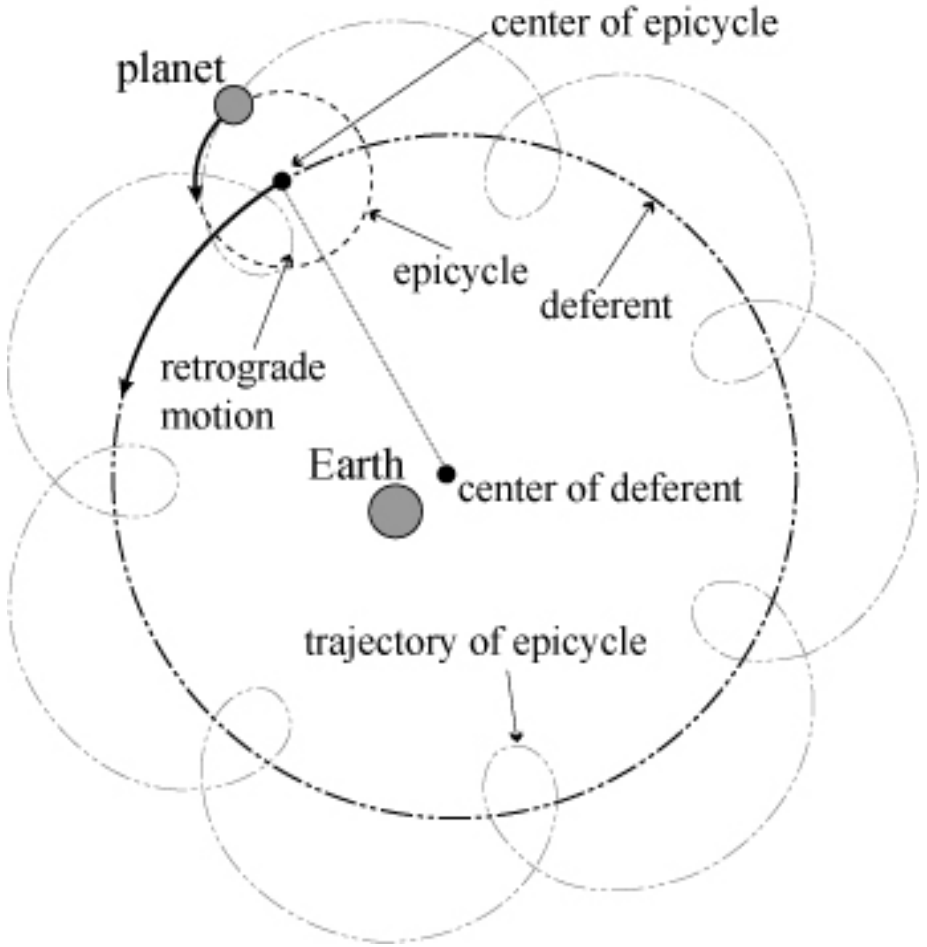
Goal: make accurate predictions

Validate predictions using hold-out set

Crude breakdown of the two cultures

Goal	Domain	Culture
Understanding	Science	Mostly data modeling
Both	Health	???
Prediction	Commerce	Mostly algorithmic modeling

Ptolemy: adventures in algorithmic modeling



The $\hat{y}$ culture and the $\hat{\beta}$ culture

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_n} x_n + \epsilon$$

Algorithmic modeling cares about errors in $\hat{y}$

Data modeling cares about biases and variances of $\widehat{\beta_i}$

Building the model: common data modeling practices

- Theory-driven variable selection
  - Variable of interest vs control variables
- Collecting new data
- Scale transformation
- …

Breiman's claims

Data modeling culture has gone wrong by over-emphasizing linear models and ignoring predictive accuracy as a measure of model usefulness [we agree]

Even when the goal is understanding, algorithmic modeling is usually superior [we don't agree]

The two cultures have been merging

Example directions

- ML for measurement
- ML for causal inference
- Interpretable and explainable ML

Breakout activity [15 minutes]

A company announces it has built a criminal risk prediction algorithm that it claims is “90% accurate”.

Imagine you’re a judge. What questions would you ask to understand what the claim means and whether it is a good estimate?

You may want to use Section 3 of the pre-read as a guide as well as Sections 3 & 4 of Hand.

Pitfalls

- Problem uncertainty
- Errors in class labels
- Researcher freedom
- Overfitting
- Drift
- Demographic biases
- Selective labels
- Other problem-specific biases
- Acting on predictions changes outcome