BST 219
Core Principles of Data Science

Lecture 23: Introduction to Machine Learning
November 19, 2024

Recipe of the Day!

Mexican Street Corn (Elote)

~ Coding in a coffee shop is fun ~

Agenda

- Announcements
  - Lab this week!
  - Homework #4 released, see Canvas announcement (last homework assignment!)
  - Final projects have been assigned (see Canvas announcement) and TFs will be assigned later today
- Continue introduction to Machine Learning

Beginner's guide to machine learning in R

"Statistics"	**"Machine learning"**
Typical goal: Explanation	Typical goal: Prediction
Does X have an effect on Y?	What best predicts Y?
Example: Does a low-carb diet lead to a reduced risk of heart attack?	Example: Given various clinical parameters, how can we use them to predict heart attacks?
Task: Develop research design based on a theory about the data-generating process to identify the causal effect (via a randomized experiment, or an observational study with statistical control variables). Don't try out various model specifications until you get your desired result (better: pre-register your hypothesized model).	Task: Try out and tune many different algorithms in order to maximize predictive accuracy in new and unseen test datasets. A theory about the true data-generating process is useful but not strictly necessary, and often not available (think of, e.g., image recognition).
Parameters of interest: Causal effect size, p-value.	Parameters of interest: Accuracy (%), precision/recall, sensitivity/specificity, ...
DON'T: Throw all kinds of variables into the model which might mask/bias your obtained effect (e.g., "spurious correlation", "collider bias").	Use whatever features are available and prove to be useful in predicting the outcome.
Use all the data to calculate your effect of interest. After all, your sample was probably designed to be representative (e.g. a random sample) of a population.	DON'T: Use all data to train a model. Always reserve subsets for validation/testing in order to avoid overfitting.

Special Announcements

- Lecture on **11/26** will be held via **Zoom**
  - Will be on a special topic that will be stand alone and not part of an assignment
- Heather's office hour on **11/26** will be **Zoom only**
- The TFs will not hold office hours the week of Thanksgiving (next week)

The Machine Learning Workflow

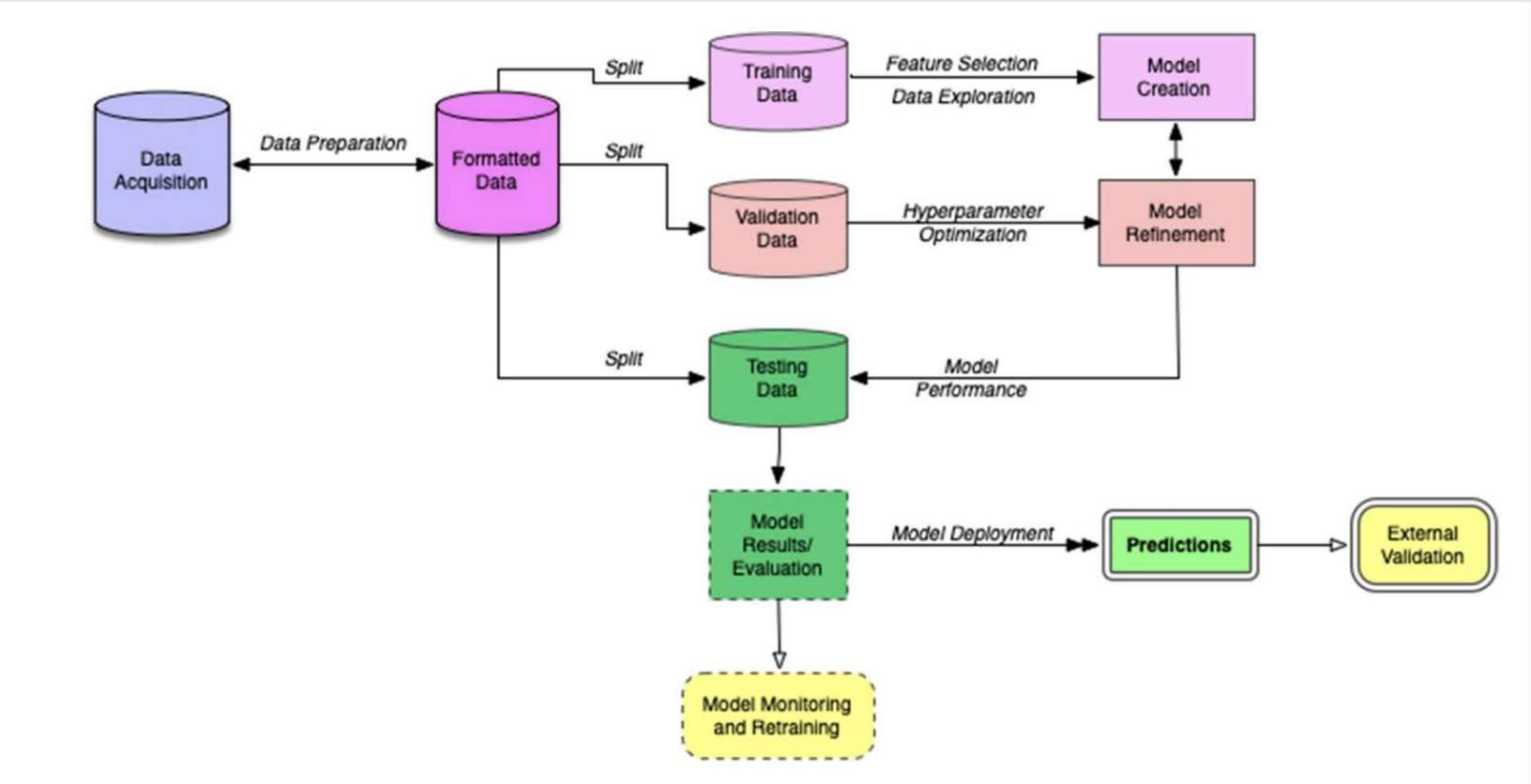

Image source

Train, Validation, and Test Sets