# Data, Decision Trees and Ensembles

## Machine Learning in Molecular Science

Prof. Michael Shirts July 23rd, 2024



#### Reminders

- All of yesterday's material is on GitHub (slides now, too!)
- Remember to work on teams during the breaks.
- Any questions now that you have thought about things more?

#### The importance of good data

- GIGO
  - Garbage in, Garbage out
- Sometimes (often?) data is at least a bit erroneous
- Need to be careful about what you include in your data set
- For LARGE data sets, especially
- Anomaly detection
  - A problem in <u>unsupervised</u> learning

#### **One Problem: Missing Data?**

- How bad is the problem?
  - Let's say we have 10,000 data points, and there are 100 features per point
  - Assume a 3% chance for each feature to be missing for every point
  - The chance that any given data point is actually complete is  $(1-0.03)^{100} = 0.048$
  - Only 4.8% of the points have all features, despite the data being 97% complete!
  - That's leaving a lot of data if we only use feature-complete data!

#### **Data imputation**

- Strategies:
  - Fill in missing data with using "around" it
  - What are the choices?
    - Mean of all other choices for that feature
    - Random selection of that feature
    - The mean of "close" data
    - Regression from other inputs
- Multiple imputation generates multiple guesses for each missing data point, which can help improve the statistics
- There are tools in pandas and scikit-learn to impute missing data.

### **Go to Notebook**

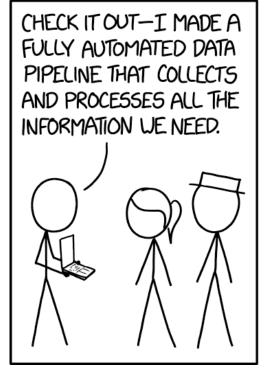
#### **Another Thing to Check for: Data Leakage**

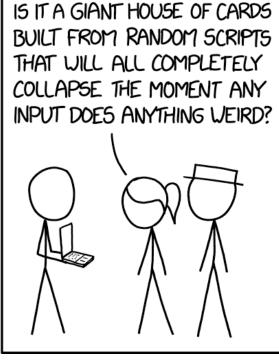
- Does training data get into the testing set?
- Are some of your features actually surrogates of your labels?
- What is going to happen with ChatGPT 8 when it ends up getting trained on outputs of ChatGPT 4 through 7?

#### Be careufi

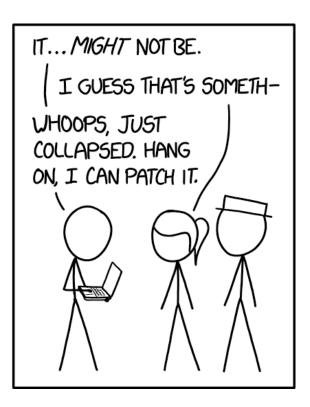
#### • Two needs:

- Automating data processing to ensure consistent treatment over the entire data set.
- Carefully human curation of the data to make sure there's nothing the automation didn't handle.









https://xkcd.com/2054/