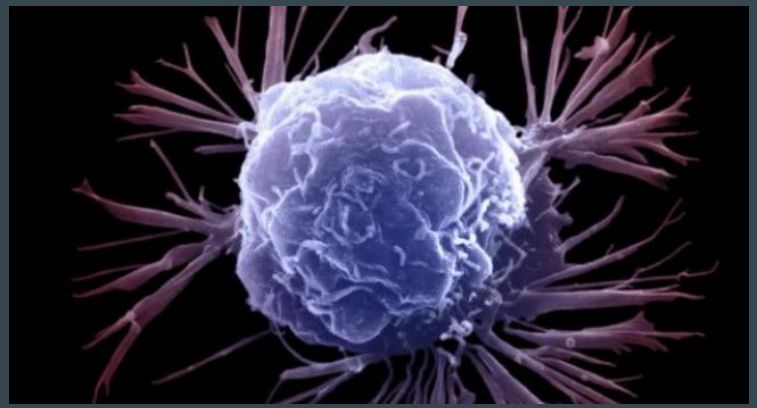
Cancer

•••

Avinash TAMBY

What Is Cancer?

- A disease developed from *genetic mutations* that lead to abnormal cell function
- Mutated cells grow and multiply
- Mutated cells invade space various parts of the body and steal resources away from normally functioning cells
- This can eventually lead to organ systems failing

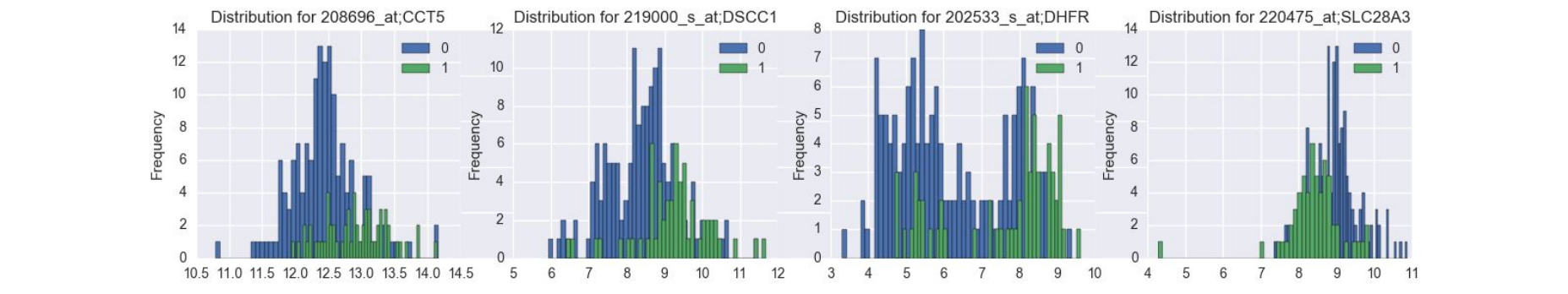


What Am I Trying To Do?

- **424** ER+ breast cancer patients
- Information about **22,215** genes *for each patient*
  - That's A LOT of genes to analyze
- I want to **predict whether or not patients in remission will relapse** solely based on their genes

Too Many Genes!

- I run a statistical test to filter out some genes
- The test (Wilcoxon Rank-Sum Test) tests whether or not Relapse patients and NoRelapse patients come from the same population
- I keep the 1,000 genes that are *least likely* to come from the same population



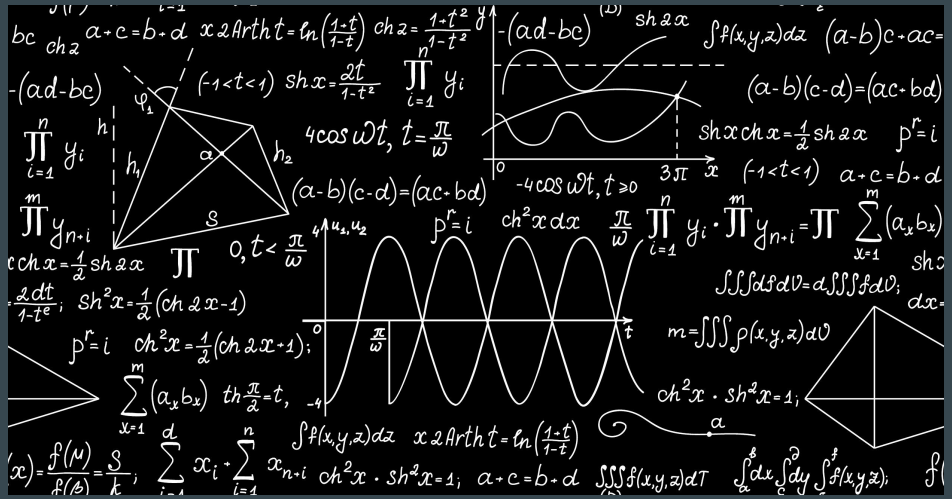
Modeling

- Split 424 patients into a training group and a testing group of equal sizes
- I *train* my models on the training set *only!*
- I *test* how well my models perform on the testing set
- I built 3 models:
  1. Random Forest Classifier
  2. Gradient Boosted Decision Tree Classifier (XGBoost)
  3. Top Scoring Pair Classifier
- Random Forests and Gradient Boosted Trees are classic, highly respected algorithms across data science. Some may even say that XGBoost is a free lunch*...

*Note: XGBoost is not a free lunch

Problems with Classic Machine Learning Algorithms

- Results sensitive to preprocessing
- Many parameters to tune
- Black-box decision-making



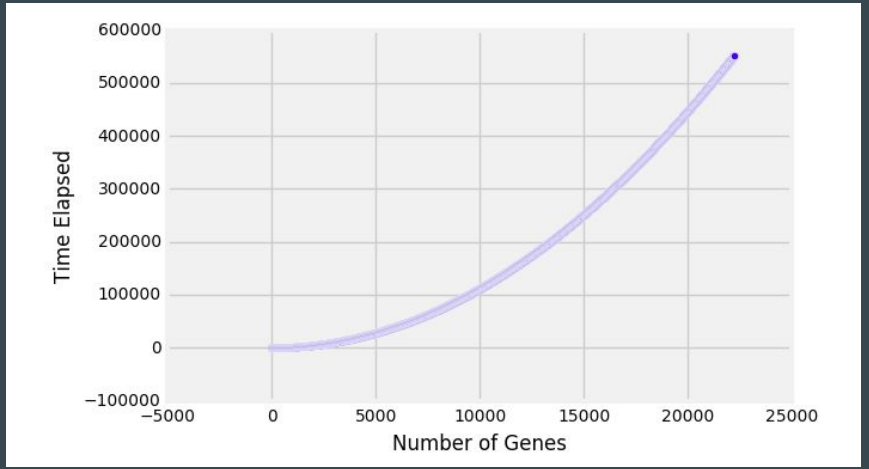
Top Scoring Pair

- TSP attempts to differentiate between Relapse and NoRelapse patients by finding pairs of genes whose expression levels typically invert from one class to the other
- The pair that achieves the largest score relative to a measure of discrimination votes for the class which makes the observed ordering within that pair the most likely

- Pros
  - Invariant to preprocessing
  - No parameters to tune
  - Transparent decision-making
- Cons
  - Computational complexity
  - Overfitting
    - Is this pair *actually* important, or was it just just due to chance?

Computational Complexity

- 100 genes - 7 seconds
- 200 genes - 34 seconds
- 2,500 genes - 1hr 11min
- 5,000 genes - 7hrs 46min
  - 12,497,500 pairs of genes (what I used)
- 10,000 genes - 1 day, 4 hrs
- 22,215 genes - **6 days 7hrs**
  - 246,742,005 pairs of genes



How Did I Do?

- Random Forest
  - Accuracy: 67%
- XGBoost
  - Accuracy: 68%
- TSP (using only 5,000 genes of the 22,215)
  - Accuracy: 54% :(
    - There are 247 million possible gene pairs. Since I only looked at 5,000 genes, I only looked at 12.5 million of the 248 million possible pairs, so I only looked at **5%** of the total possible gene pairs
- **Does accuracy tell the whole story?**

Evaluation (XGBoost)

	predicted_Relapse	predicted_NoRelapse
Relapse	21	54
NoRelapse	13	124

- How many *Relapse* patients did I classify correctly (Sensitivity)? 21 / 75 = **28%**
  - That's not good. I'm telling a lot relapsing patients that they didn't relapse
- How many *NoRelapse* patients did I classify correctly (Specificity)? 124 / 137 = **91%**
  - That's not good either. I'm telling 9% of patients who didn't relapse that they did!

Conclusion

- This is cool
  - Analyzing genes to help cancer patients
- This is hard
  - Biostatisticians do a lot of research in this field
- This is fun
  - Math *and* science!