Prediction study design

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

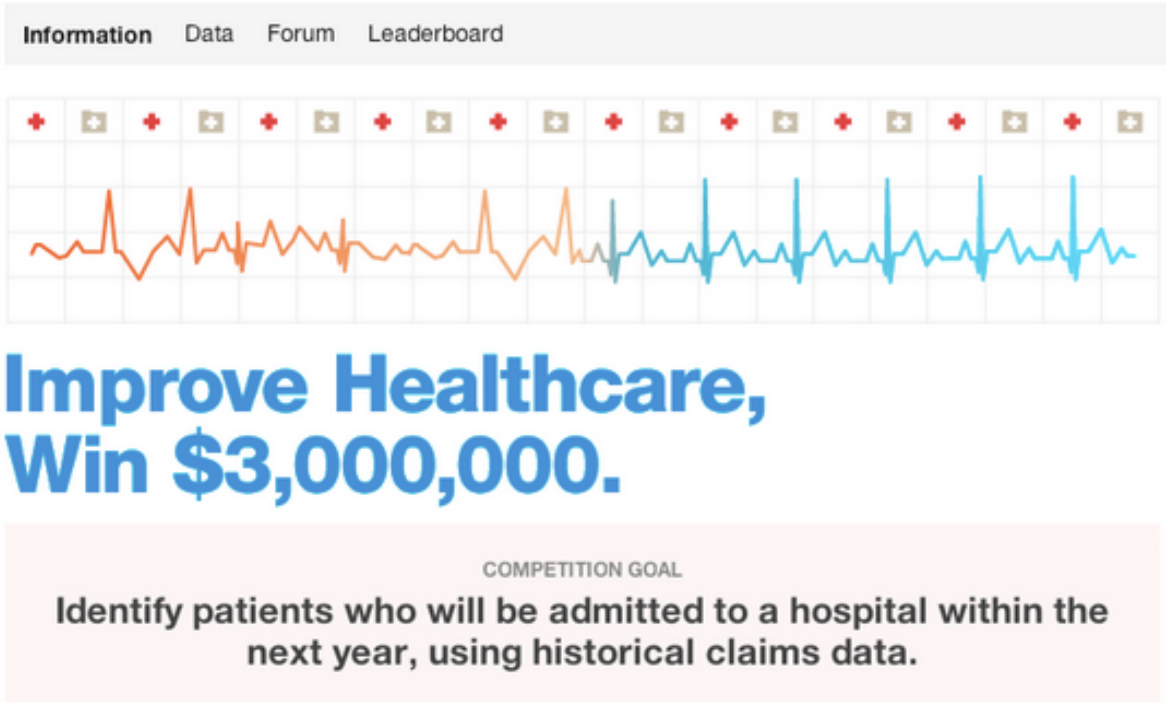
Key ideas

- Motivation
- Steps in predictive studies
- Choosing the right data
- Error measures
- Study design

Why predict? Glory!

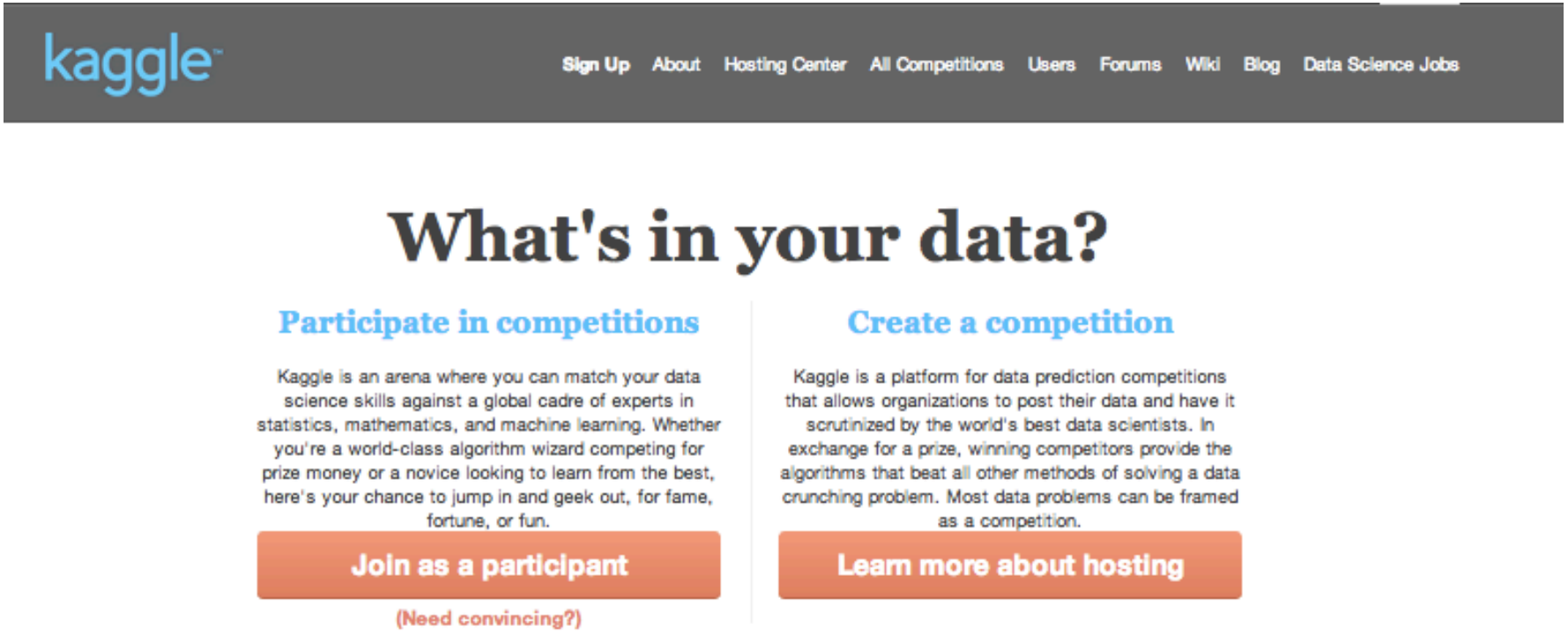
http://www.zimbio.com/photos/Chris+Volinsky

Why predict? Riches!



http://www.heritagehealthprize.com/c/hhp

Why predict? For sport!



http://www.kaggle.com/

5/15

Why predict? To save lives!

http://www.oncotypedx.com/en-US/Home

6/15

Steps in building a prediction

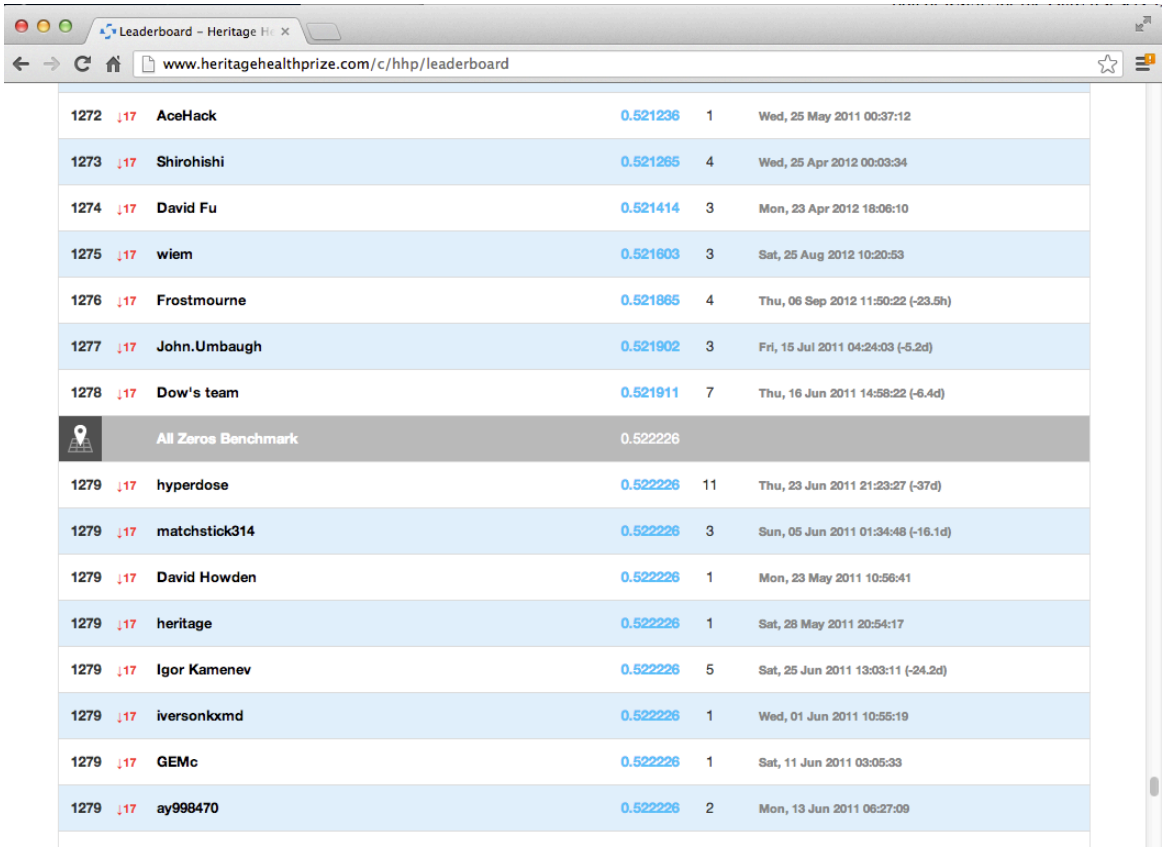
1. Find the right data
2. Define your error rate
3. Split data into:
    - Training
    - Testing
    - Validation (optional)
4. On the training set pick features
5. On the training set pick prediction function
6. On the training set cross-validate
7. If no validation - apply 1x to test set
8. If validation - apply to test set and refine
9. If validation - apply 1x to validation

Find the right data

1. In some cases it is easy (movie ratings -> new movie ratings)
2. It may be harder (gene expression data -> disease)
3. Depends strongly on the definition of "good prediction".
4. Often more data > better models
5. Know the bench mark
6. You need to start with raw data for predictions - processing is often cross-sample.

Know the benchmarks

Probability of perfect classification is approximately $\left(\frac{1}{2}\right)^{test\ set\ sample\ size}$



http://www.heritagehealthprize.com/c/hhp/leaderboard

Defining true/false positives

In general, **Positive** = identified and **negative** = rejected. Therefore:

True positive = correctly identified

False positive = incorrectly identified

True negative = correctly rejected

False negative = incorrectly rejected

Medical testing example:

True positive = Sick people correctly diagnosed as sick

False positive= Healthy people incorrectly identified as sick

True negative = Healthy people correctly identified as healthy

False negative = Sick people incorrectly identified as healthy.

http://en.wikipedia.org/wiki/Sensitivity_and_specificity

10/15

Define your error rate

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

http://en.wikipedia.org/wiki/Sensitivity_and_specificity

11/15

Why your choice matters

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

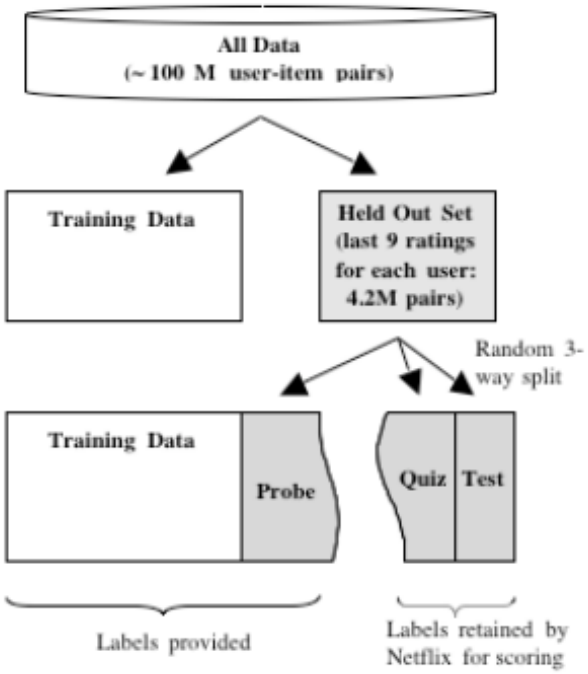
http://en.wikipedia.org/wiki/Sensitivity_and_specificity

12/15

Common error measures

1. Mean squared error (or root mean squared error)
   - Continuous data, sensitive to outliers
2. Median absolute deviation
   - Continuous data, often more robust
3. Sensitivity (recall)
   - If you want few missed positives
4. Specificity
   - If you want few negatives called positives
5. Accuracy
   - Weights false positives/negatives equally
6. Concordance
   - One example is kappa

Study design



http://www2.research.att.com/~volinsky/papers/ASAStatComp.pdf

Key issues and further resources

Issues:

1. Accuracy
2. Overfitting
3. Interpretability
4. Computational speed

Resources:

1. Practical machine learning
2. Elements of statistical learning
3. Coursera machine learning
4. Machine learning for hackers