Prediction study design

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

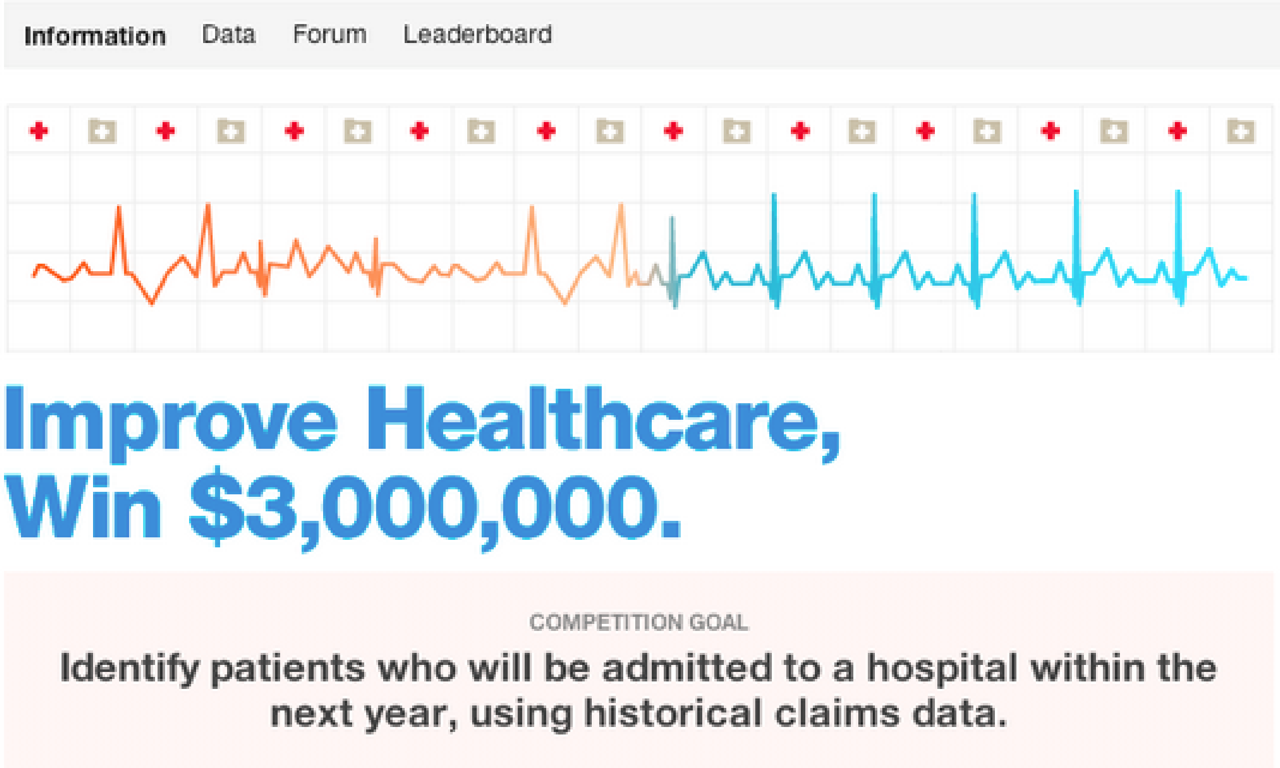
Key ideas

- 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!



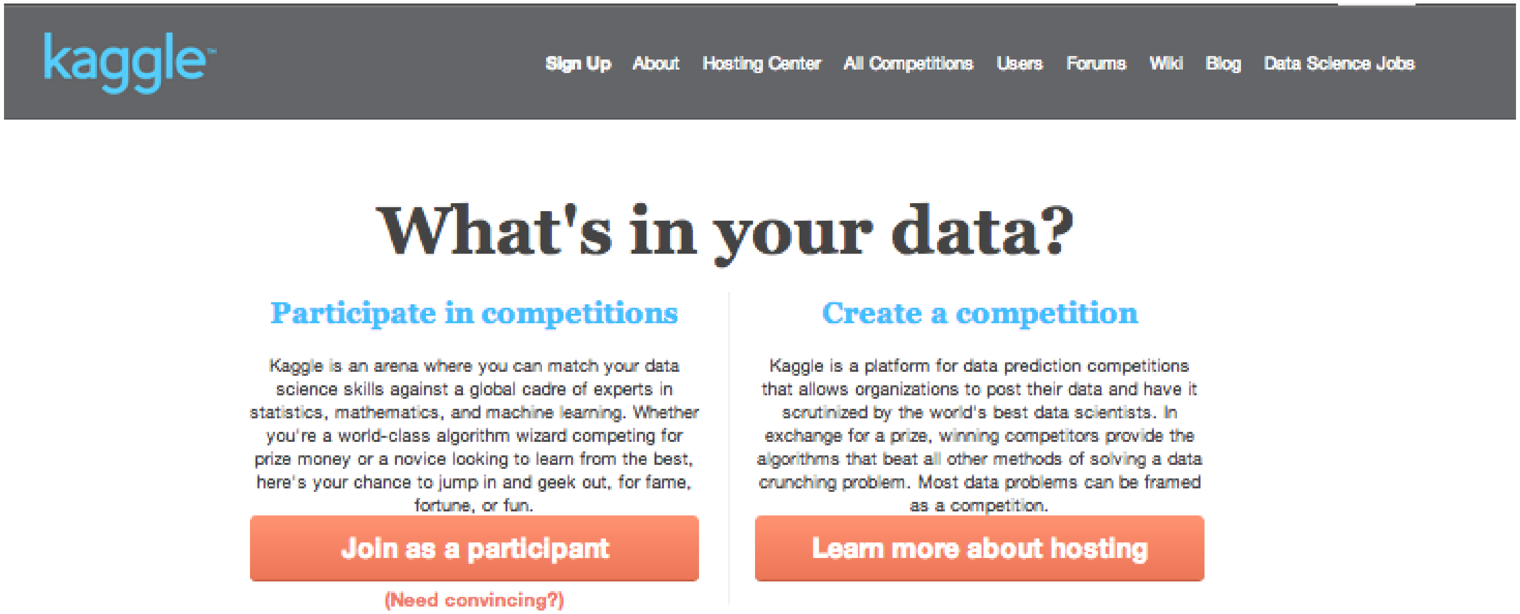
Improve Healthcare,
Win $3,000,000.

COMPETITION GOAL

Identify patients who will be admitted to a hospital within the next year, using historical claims data.

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

Why predict? For sport!



What's in your data?

Participate in competitions

Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.

Join as a participant

(Need convincing?)

Create a competition

Kaggle is a platform for data prediction competitions that allows organizations to post their data and have it scrutinized by the world's best data scientists. In exchange for a prize, winning competitors provide the algorithms that beat all other methods of solving a data crunching problem. Most data problems can be framed as a competition.

Learn more about hosting

http://www.kaggle.com/

Why predict? To save lives!

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

Steps in building a prediction

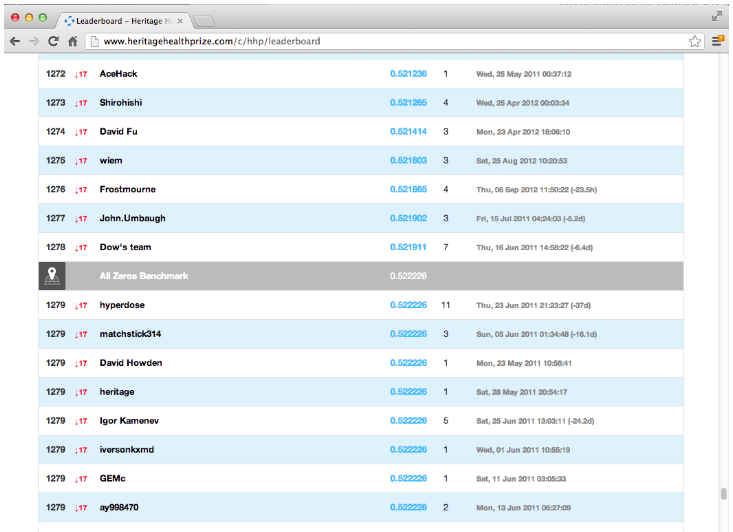
1. Find the right data
2. Define your error rate
3. Split data into:
    - Training
    - Testing
    - Validation (optional)
4. On the training set pick features
5. On the training set pick prediction function
6. On the training set cross-validate
7. If no validation - apply 1x to test set
8. If validation - 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 $\left(\frac{1}{2}\right)^{test\ set\ sample\ size}$



1272	↓17	AceHack	0.521236	1	Wed, 25 May 2011 00:37:12
1273	↓17	Shirohishi	0.521265	4	Wed, 25 Apr 2012 00:03:34
1274	↓17	David Fu	0.521414	3	Mon, 23 Apr 2012 18:06:10
1275	↓17	wiem	0.521603	3	Sat, 25 Aug 2012 10:20:53
1276	↓17	Frostmourne	0.521865	4	Thu, 06 Sep 2012 11:50:22 (-23.5h)
1277	↓17	John.Umbaugh	0.521902	3	Fri, 15 Jul 2011 04:24:03 (-5.2d)
1278	↓17	Dow's team	0.521911	7	Thu, 16 Jun 2011 14:58:22 (-6.4d)
		All Zeros Benchmark	0.522226		
1279	↓17	hyperdose	0.522226	11	Thu, 23 Jun 2011 21:23:27 (-37d)
1279	↓17	matchstick314	0.522226	3	Sun, 05 Jun 2011 01:34:48 (-16.1d)
1279	↓17	David Howden	0.522226	1	Mon, 23 May 2011 10:56:41
1279	↓17	heritage	0.522226	1	Sat, 28 May 2011 20:54:17
1279	↓17	Igor Kamenev	0.522226	5	Sat, 25 Jun 2011 13:03:11 (-24.2d)
1279	↓17	iversonkxmd	0.522226	1	Wed, 01 Jun 2011 10:55:19
1279	↓17	GEMc	0.522226	1	Sat, 11 Jun 2011 03:05:33
1279	↓17	ay998470	0.522226	2	Mon, 13 Jun 2011 06:27:09

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

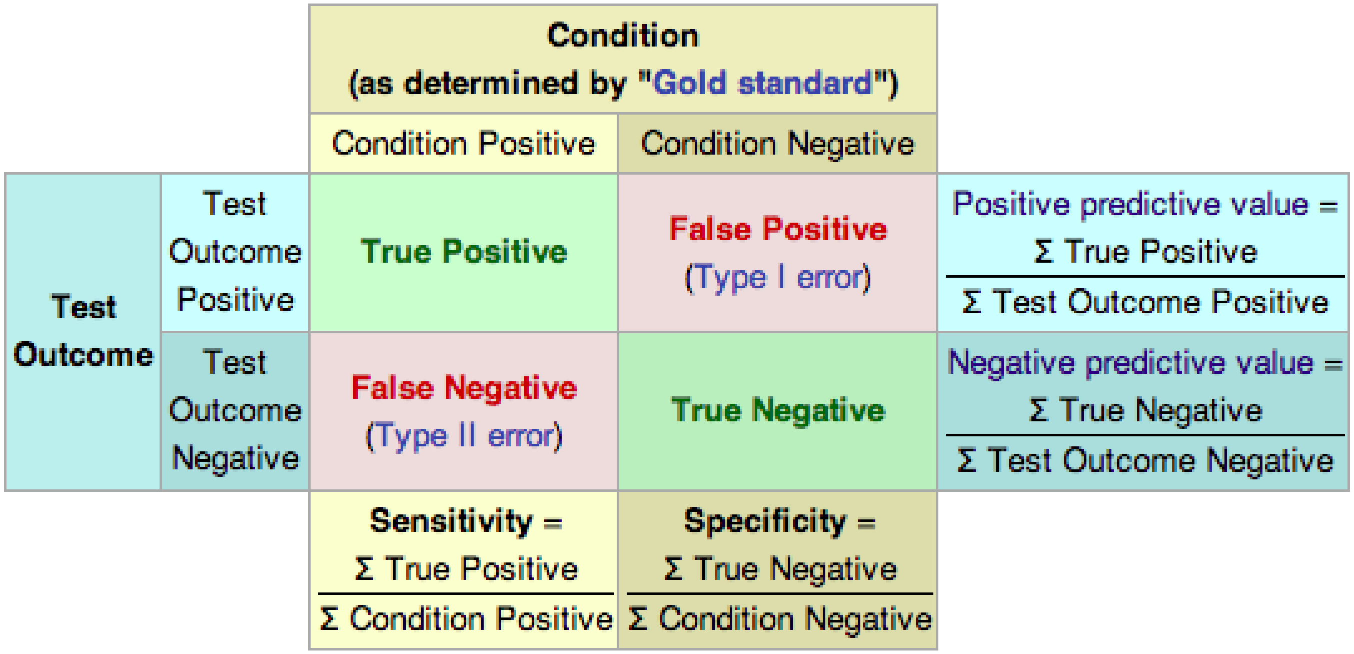
False positive= Healthy people incorrectly identified as sick

True negative = Healthy people correctly identified as healthy

False negative = Sick people incorrectly identified as healthy.

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

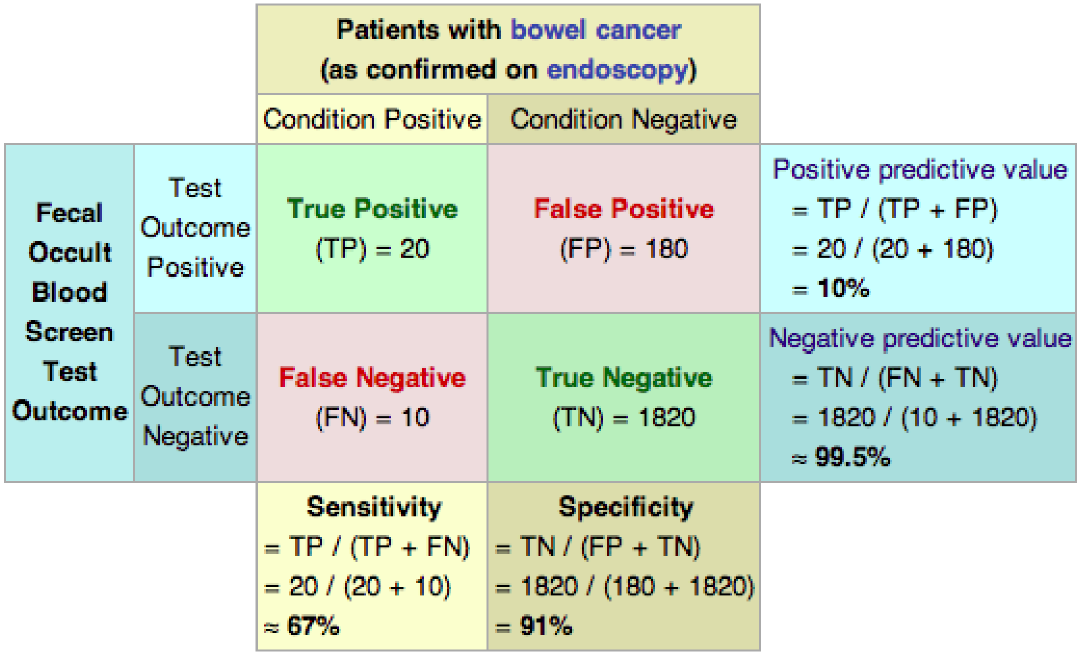
Define your error rate



		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

Why your choice matters



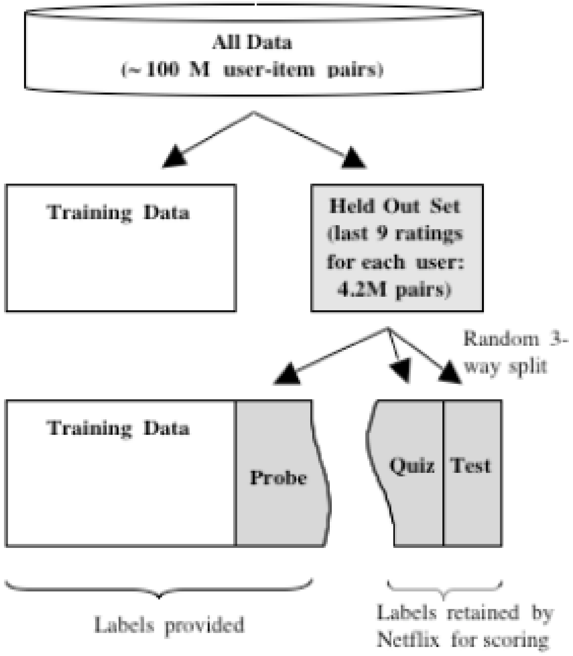
		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome 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

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
7. Predictive value of a positive (precision)
   - When you are screeing and prevelance is low

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