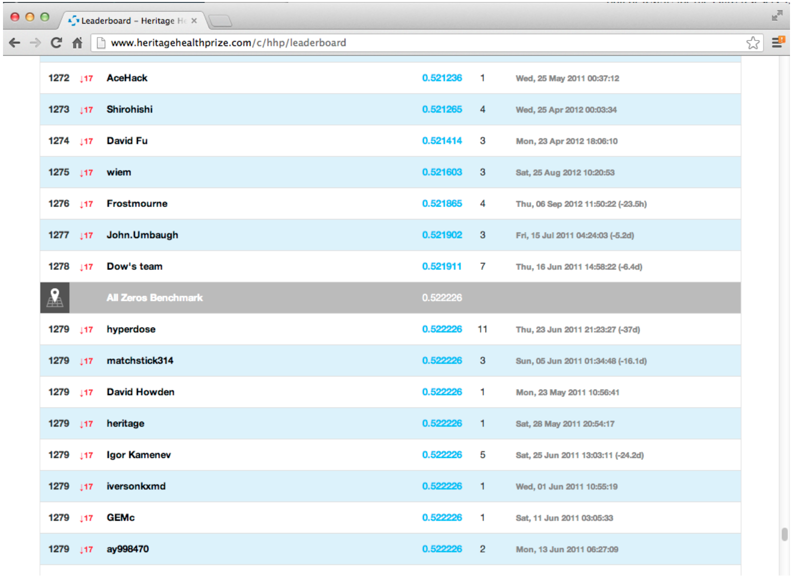
Prediction study design

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Prediction study design

1. Define your error rate
2. Split data into:
   - Training, Testing, Validation (optional)
3. On the training set pick features
   - Use cross-validation
4. On the training set pick prediction function
   - Use cross-validation
5. If no validation
   - Apply 1x to test set
6. If validation
   - Apply to test set and refine
   - Apply 1x to validation

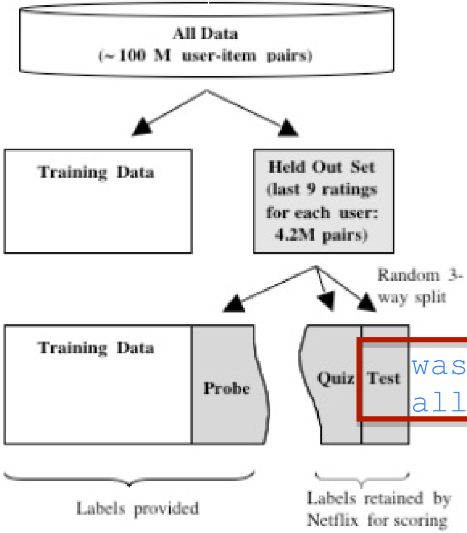
Know the benchmarks



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

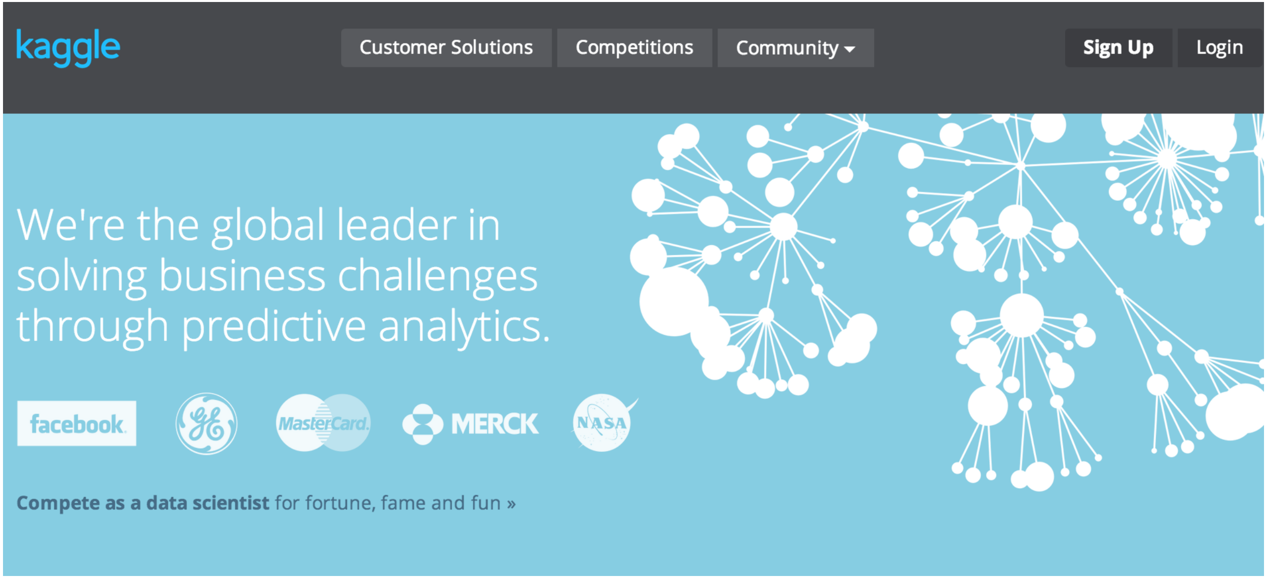
Study design


All Data
(~ 100 M user-item pairs)
Training Data
Held Out Set (last 9 ratings for each user: 4.2M pairs)
Random 3-way split
Training Data
Probe
Quiz
Test
Labels provided
Labels retained by Netflix for scoring

was used only once to test all models!

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

Used by the professionals



We're the global leader in solving business challenges through predictive analytics.

Compete as a data scientist for fortune, fame and fun »

http://www.kaggle.com/

Avoid small sample sizes

when splitting the samples into train/test/validation sets.

- Suppose you are predicting a binary outcome
  - Diseased/healthy
  - Click on ad/not click on ad
- One classifier is flipping a coin …sehr einfacher Algorithmus
- Probability of perfect classification is approximately:
  - $\left(\frac{1}{2}\right)^{\text{sample size}}$
  - n = 1 flipping coin 50% chance of 100% accuracy
  - n = 2 flipping coin 25% chance of 100% accuracy
  - n = 10 flipping coin 0.10% chance of 100% accuracy

-> mit genuegend grossen Training-Sets koennen wir sicherstellen, dass wir nicht einfach zu Zufall gute Accuracy bekommen.

Rules of thumb for prediction study design

- If you have a large sample size
  - 60% training
  - 20% test
  - 20% validation

was sagte Ng?

- If you have a medium sample size
  - 60% training
  - 40% testing

wie viel ist large/medium/small ???!

- If you have a small sample size
  - Do cross validation
  - Report caveat of small sample size

Some principles to remember

- Set the test/validation set aside and *don't look at it*
- In general *randomly* sample training and test
- Your data sets must reflect structure of the problem
  - If predictions evolve with time split train/test in time chunks (calledbacktesting in finance)

`independence… (?)`

- All subsets should reflect as much diversity as possible
  - Random assignment does this
  - You can also try to balance by features - but this is tricky