



Motivation and pre-requisites

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

About this course

- This course covers the basic ideas behind machine learning/prediction
 - Study design - training vs. test sets
 - Conceptual issues - out of sample error, ROC curves
 - Practical implementation - the caret package
- What this course depends on
 - The Data Scientist's Toolbox
 - R Programming
- What would be useful
 - Exploratory analysis
 - Reporting Data and Reproducible Research
 - Regression models

Who predicts?

- Local governments -> pension payments
- Google -> whether you will click on an ad
- Amazon -> what movies you will watch
- Insurance companies -> what your risk of death is
- Johns Hopkins -> who will succeed in their programs

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!



[Sign Up](#) [About](#) [Hosting Center](#) [All Competitions](#) [Users](#) [Forums](#) [Wiki](#) [Blog](#) [Data Science Jobs](#)

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.

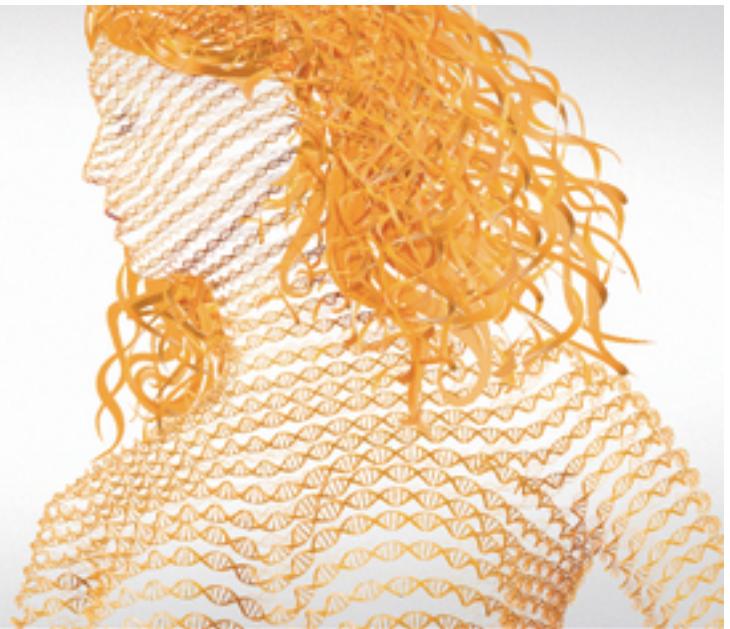
[Learn more about hosting](#)

<http://www.kaggle.com/>

Why predict? To save lives!

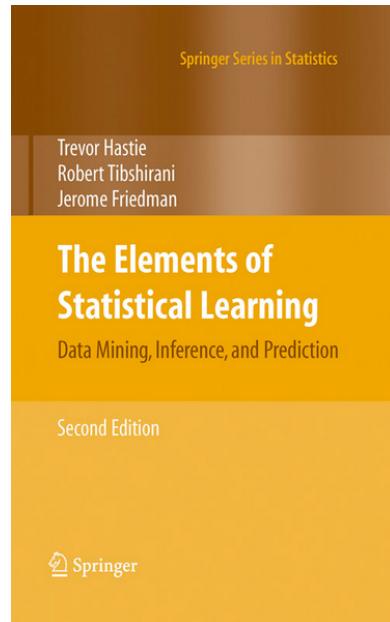
Oncotype DX® reveals
the underlying biology that
changes treatment decisions
37% of the time

Uncover the Unexpected™



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

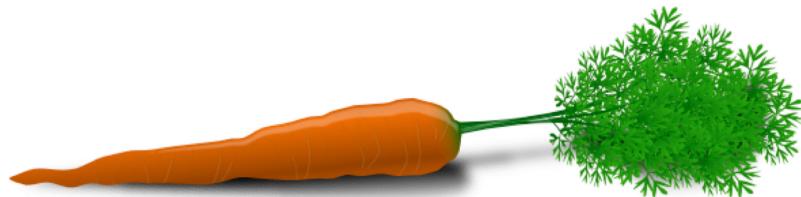
A useful (if a bit advanced) book



[The elements of statistical learning](#)

A useful package

the caret package



The **caret** package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

Links

[train Model List](#)

Topics

[Main Page](#)

[Data Sets](#)

[Visualizations](#)

[Pre-Processing](#)

<http://caret.r-forge.r-project.org/>

Machine learning (more advanced material)

Stanford Machine Learning

Andrew Ng

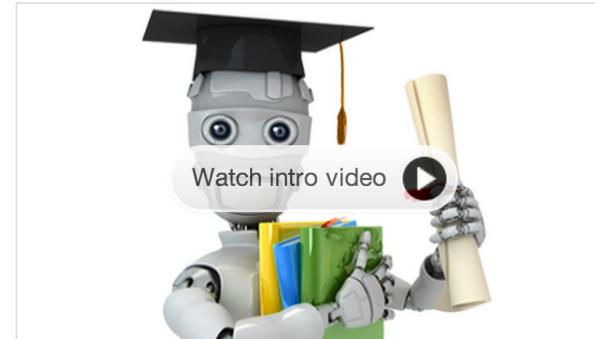
Taught In: English

Subtitles Available In: English

Sessions:

Oct 14th 2013 (10 weeks long) ▾

Learn for Free



3,794

12k

14k

Tweet

+1

Like

<https://www.coursera.org/course/ml>

Even more resources

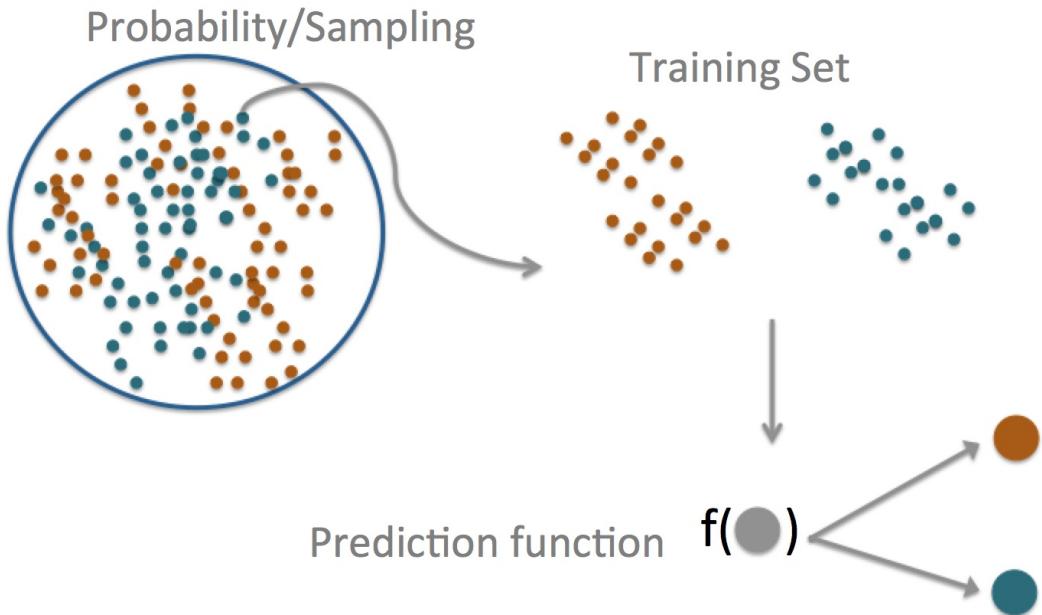
- [List of machine learning resources on Quora](#)
- [List of machine learning resources from Science](#)
- [Advanced notes from MIT open courseware](#)
- [Advanced notes from CMU](#)
- [Kaggle - machine learning competitions](#)



What is prediction?

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

The central dogma of prediction



What can go wrong

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,2*} Ryan Kennedy,^{1,3,4} Gary King,³ Alessandro Vespignani^{5,6,3}

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (3, 4), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can



Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

run ever since, with a few changes announced in October 2013 (10, 15).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

<http://www.sciencemag.org/content/343/6176/1203.full.pdf>

Components of a predictor

question -> input data -> features -> algorithm -> parameters -> evaluation

SPAM Example

question -> input data -> features -> algorithm -> parameters -> evaluation

Start with a general question

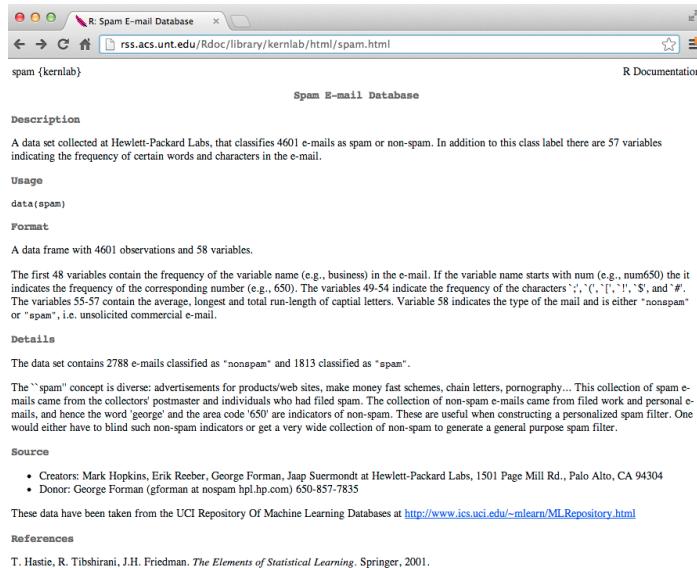
Can I automatically detect emails that are SPAM that are not?

Make it concrete

Can I use quantitative characteristics of the emails to classify them as SPAM/HAM?

SPAM Example

question -> **input data** -> features -> algorithm -> parameters -> evaluation



The screenshot shows a web browser window with the title "R: Spam E-mail Database". The URL in the address bar is "rss.acs.unt.edu/Rdoc/library/kernlab/html/spam.html". The page content is the R documentation for the "spam" dataset, which is part of the "kernlab" library. The documentation includes sections for "Description", "Usage", "Format", "Details", "Source", and "References". It describes the dataset as containing 4601 e-mails, 57 variables, and 48 frequency variables. It also provides details about the "spam" concept and its creators.

R Documentation

spam {kernlab}

Spam E-mail Database

Description

A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail.

Usage

```
data(spam)
```

Format

A data frame with 4601 observations and 58 variables.

The first 48 variables contain the frequency of the variable name (e.g., business) in the e-mail. If the variable name starts with num (e.g., num650) the it indicates the frequency of the corresponding number (e.g., 650). The variables 49-54 indicate the frequency of the characters ':', ',', '.', '!', '\$, and '#'. The variables 55-57 contain the average, longest and total run-length of capital letters. Variable 58 indicates the type of the mail and is either "nonspam" or "spam", i.e. unsolicited commercial e-mail.

Details

The data set contains 2788 e-mails classified as "nonspam" and 1813 classified as "spam".

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography... This collection of spam e-mails came from the collectors' postmaster and individuals who had filed spam. The collection of non-spam e-mails came from filed work and personal e-mails, and hence the word "george" and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

Source

- Creators: Mark Hopkins, Erik Reber, George Forman, Jaap Suermondt at Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304
- Donor: George Forman (gforman@nosspam.hpl.hp.com) 650-857-7835

These data have been taken from the UCI Repository Of Machine Learning Databases at <http://www.ics.uci.edu/~mlearn/MLRepository.html>

References

T. Hastie, R. Tibshirani, J.H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

<http://rss.acs.unt.edu/Rdoc/library/kernlab/html/spam.html>

SPAM Example

question -> input data -> **features** -> algorithm -> parameters -> evaluation

Dear Jeff,

Can you send me your address so I can send you the invitation?

Thanks,

Ben

SPAM Example

question -> input data -> **features** -> algorithm -> parameters -> evaluation

Dear Jeff,

Can **you**

send me your address so I can send **you** the invitation?

Thanks,

Ben

Frequency of **you** = $2/17 = 0.118$

SPAM Example

question -> input data -> **features** -> algorithm -> parameters -> evaluation

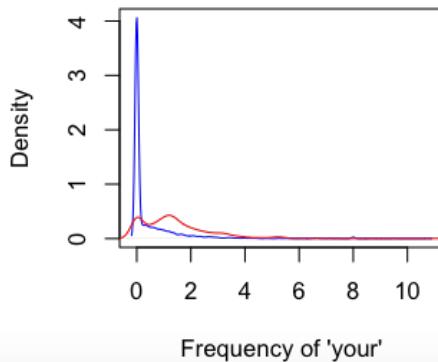
```
library(kernlab)
data(spam)
head(spam)
```

	make	address	all	num3d	our	over	remove	internet	order	mail	receive	will	people	report	addresses	
1	0.00	0.64	0.64	0	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	
2	0.21	0.28	0.50	0	0.14	0.28	0.21	0.07	0.00	0.94	0.21	0.79	0.65	0.21	0.14	
3	0.06	0.00	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0.00	1.75	
4	0.00	0.00	0.00	0	0.63	0.00	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0.00	0.00	
5	0.00	0.00	0.00	0	0.63	0.00	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0.00	0.00	
6	0.00	0.00	0.00	0	1.85	0.00	0.00	1.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	free	business	email	you	credit	your	font	num000	money	hp	hpl	george	num650	lab	labs	telnet
1	0.32	0.00	1.29	1.93	0.00	0.96	0	0.00	0.00	0	0	0	0	0	0	0
2	0.14	0.07	0.28	3.47	0.00	1.59	0	0.43	0.43	0	0	0	0	0	0	0
3	0.06	0.06	1.03	1.36	0.32	0.51	0	1.16	0.06	0	0	0	0	0	0	0
4	0.31	0.00	0.00	3.18	0.00	0.31	0	0.00	0.00	0	0	0	0	0	0	0

SPAM Example

question -> input data -> features -> **algorithm** -> parameters -> evaluation

```
plot(density(spam$your[spam$type=="nonspam"]),
      col="blue",main="",xlab="Frequency of 'your'")
lines(density(spam$your[spam$type=="spam"]),col="red")
```



SPAM Example

question -> input data -> features -> **algorithm** -> parameters -> evaluation

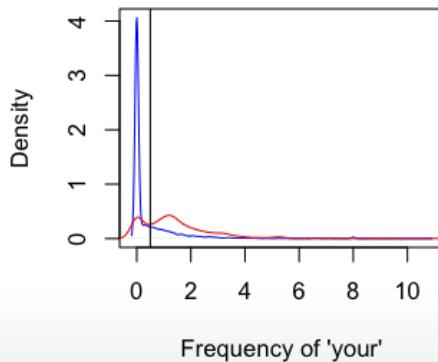
Our algorithm

- Find a value C.
- **frequency of 'your' > C** predict "spam"

SPAM Example

question -> input data -> features -> algorithm -> **parameters** -> evaluation

```
plot(density(spam$your[spam$type=="nonspam"]),
      col="blue",main="",xlab="Frequency of 'your'")
lines(density(spam$your[spam$type=="spam"]),col="red")
abline(v=0.5,col="black")
```



SPAM Example

question -> input data -> features -> algorithm -> parameters -> **evaluation**

```
prediction <- ifelse(spam$your > 0.5, "spam", "nonspam")
table(prediction,spam$type)/length(spam$type)
```

```
prediction nonspam    spam
nonspam   0.4590 0.1017
spam      0.1469 0.2923
```

Accuracy≈ 0.459 + 0.292 = 0.751



Relative importance of steps

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

Relative order of importance

question > data > features > algorithms

An important point

“ The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.”

John Tukey

Garbage in = Garbage out

question -> **input data** -> features -> algorithm -> parameters -> evaluation

1. May be easy (movie ratings -> new movie ratings)
2. May be harder (gene expression data -> disease)
3. Depends on what is a "good prediction".
4. Often more data > better models
5. The most important step!

Features matter!

question -> input data -> **features** -> algorithm -> parameters -> evaluation

Properties of good features

- Lead to data compression
- Retain relevant information
- Are created based on expert application knowledge

Common mistakes

- Trying to automate feature selection
- Not paying attention to data-specific quirks
- Throwing away information unnecessarily

May be automated with care

question -> input data -> **features** -> algorithm -> parameters -> evaluation



<http://arxiv.org/pdf/1112.6209v5.pdf>

Algorithms matter less than you'd think

question -> input data -> features -> **algorithm** -> parameters -> evaluation

TABLE 1

Performance of linear discriminant analysis and the best result we found on ten randomly selected data sets

Data set	Best method e.r.	Lindisc e.r.	Default rule	Prop linear
Segmentation	0.0140	0.083	0.760	0.907
Pima	0.1979	0.221	0.350	0.848
House-votes16	0.0270	0.046	0.386	0.948
Vehicle	0.1450	0.216	0.750	0.883
Satimage	0.0850	0.160	0.758	0.889
Heart Cleveland	0.1410	0.141	0.560	1.000
Splice	0.0330	0.057	0.475	0.945
Waveform21	0.0035	0.004	0.667	0.999
Led7	0.2650	0.265	0.900	1.000
Breast Wisconsin	0.0260	0.038	0.345	0.963

<http://arxiv.org/pdf/math/0606441.pdf>

Issues to consider

The “Best” Machine Learning Method

Interpretable

Simple

Accurate

Fast
(to train and test)

Scalable

<http://strata.oreilly.com/2013/09/gaining-access-to-the-best-machine-learning-methods.html>

Prediction is about accuracy tradeoffs

- Interpretability versus accuracy
- Speed versus accuracy
- Simplicity versus accuracy
- Scalability versus accuracy

Interpretability matters

```
if total cholesterol ≥160 and smoke then 10 year CHD risk ≥ 5%
else if smoke and systolic blood pressure≥140 then 10 year CHD risk ≥
5%
else 10 year CHD risk < 5%
```

<http://www.cs.cornell.edu/~chenhao/pub/mldg-0815.pdf>

Scalability matters



Innovation

by Mike Masnick

Fri, Apr 13th 2012
12:07am

5

Why Netflix Never Implemented The Algorithm That Won The Netflix \$1 Million Challenge

from the *times-change dept*

You probably recall all the excitement that went around when a group **finally won** the big Netflix \$1 million prize in 2009, improving Netflix's recommendation algorithm by 10%. But what you might *not* know, is that **Netflix never implemented that solution itself**. Netflix recently put up a blog post **discussing some of the details of its recommendation system**, which (as an aside) explains why the winning entry never was used. First, they note that they *did* make use of an earlier bit of code that came out of the contest:

<http://www.techdirt.com/blog/innovation/articles/20120409/03412518422/>

<http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html>



In sample and out of sample error

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

In sample versus out of sample

In Sample Error: The error rate you get on the same data set you used to build your predictor. Sometimes called resubstitution error.

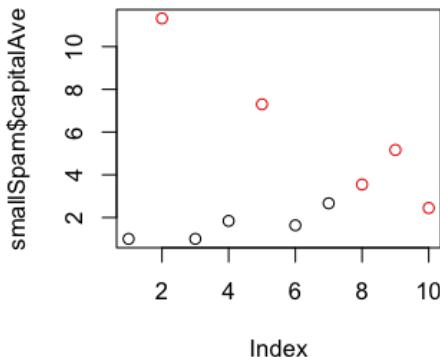
Out of Sample Error: The error rate you get on a new data set. Sometimes called generalization error.

Key ideas

1. Out of sample error is what you care about
2. In sample error < out of sample error
3. The reason is overfitting
 - Matching your algorithm to the data you have

In sample versus out of sample errors

```
library(kernlab); data(spam); set.seed(333)
smallSpam <- spam[sample(dim(spam)[1],size=10),]
spamLabel <- (smallSpam$type=="spam")*1 + 1
plot(smallSpam$capitalAve,col=spamLabel)
```



Prediction rule 1

- capitalAve > 2.7 = "spam"
- capitalAve < 2.40 = "nonspam"
- capitalAve between 2.40 and 2.45 = "spam"
- capitalAve between 2.45 and 2.7 = "nonspam"

Apply Rule 1 to smallSpam

```
rule1 <- function(x){  
  prediction <- rep(NA,length(x))  
  prediction[x > 2.7] <- "spam"  
  prediction[x < 2.40] <- "nonspam"  
  prediction[(x >= 2.40 & x <= 2.45)] <- "spam"  
  prediction[(x > 2.45 & x <= 2.70)] <- "nonspam"  
  return(prediction)  
}  
  
table(rule1(smallSpam$capitalAve),smallSpam$type)
```

	nonspam	spam
nonspam	5	0
spam	0	5

Prediction rule 2

- $\text{capitalAve} > 2.40 = \text{"spam"}$
- $\text{capitalAve} \leq 2.40 = \text{"nonspam"}$

Apply Rule 2 to smallSpam

```
rule2 <- function(x){  
  prediction <- rep(NA,length(x))  
  prediction[x > 2.8] <- "spam"  
  prediction[x <= 2.8] <- "nonspam"  
  return(prediction)  
}  
table(rule2(smallSpam$capitalAve),smallSpam$type)
```

	nonspam	spam
nonspam	5	1
spam	0	4

Apply to complete spam data

```
table(rule1(spam$capitalAve),spam$type)
```

	nonspam	spam
nonspam	2141	588
spam	647	1225

```
table(rule2(spam$capitalAve),spam$type)
```

	nonspam	spam
nonspam	2224	642
spam	564	1171

```
mean(rule1(spam$capitalAve)==spam$type)
```

Look at accuracy

```
sum(rule1(spam$capitalAve)==spam$type)
```

```
[1] 3366
```

```
sum(rule2(spam$capitalAve)==spam$type)
```

```
[1] 3395
```

What's going on?

Overfitting

- Data have two parts
 - Signal
 - Noise
- The goal of a predictor is to find signal
- You can always design a perfect in-sample predictor
- You capture both signal + noise when you do that
- Predictor won't perform as well on new samples

<http://en.wikipedia.org/wiki/Overfitting>



Prediction study design

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Prediction study design

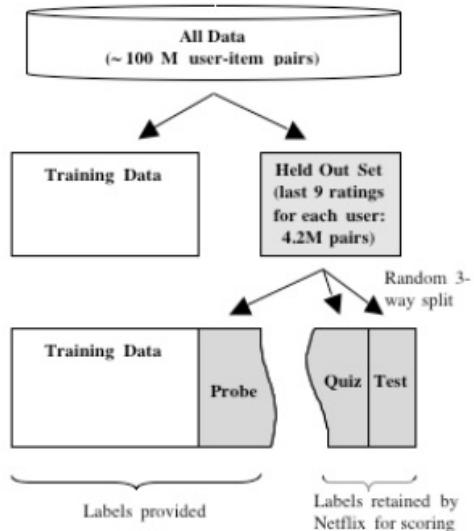
1. Define your error rate
2. Split data into:
 - Training, Testing, Validation (optional)
3. On the training set pick features
 - Use cross-validation
4. On the training set pick prediction function
 - Use cross-validation
5. If no validation
 - Apply 1x to test set
6. If validation
 - Apply to test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

Know the benchmarks

Leaderboard – Heritage Health Prize					
#	User	Score	Rank	Date	Comments
1272	AceHack	0.521236	1	Wed, 25 May 2011 00:37:12	
1273	Shirohishi	0.521265	4	Wed, 25 Apr 2012 00:03:34	
1274	David Fu	0.521414	3	Mon, 23 Apr 2012 18:06:10	
1275	wiem	0.521603	3	Sat, 25 Aug 2012 10:20:53	
1276	Frostmourne	0.521865	4	Thu, 06 Sep 2012 11:50:22 (-23.9h)	
1277	John.Umbaugh	0.521902	3	Fri, 15 Jul 2011 04:24:03 (-5.2d)	
1278	Dow's team	0.521911	7	Thu, 16 Jun 2011 14:58:22 (-6.4d)	
<hr/>					
All Zeros Benchmark					
1279	hyperdose	0.522226	11	Thu, 23 Jun 2011 21:23:27 (-37d)	
1279	matchstick314	0.522226	3	Sun, 06 Jun 2011 01:34:48 (-16.1d)	
1279	David Howden	0.522226	1	Mon, 23 May 2011 10:56:41	
1279	heritage	0.522226	1	Sat, 26 May 2011 20:54:17	
1279	Igor Kamenev	0.522226	5	Sat, 25 Jun 2011 13:03:11 (-24.2d)	
1279	iversonkxmd	0.522226	1	Wed, 01 Jun 2011 10:55:19	
1279	GEMc	0.522226	1	Sat, 11 Jun 2011 03:06:33	
1279	ay998470	0.522226	2	Mon, 13 Jun 2011 06:27:09	

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

Study design



<http://www2.research.att.com/~volinsky/papers/ASASStatComp.pdf>

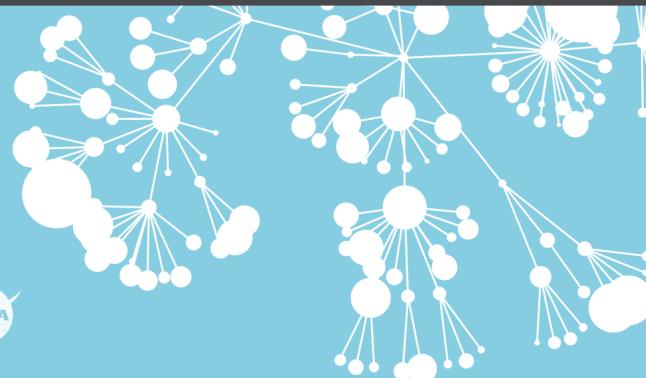
Used by the professionals

kaggle

Customer Solutions Competitions Community ▾

Sign Up Login

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



facebook. GE MasterCard. MERCK NASA

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

<http://www.kaggle.com/>

Avoid small sample sizes

- Suppose you are predicting a binary outcome
 - Diseased/healthy
 - Click on ad/not click on ad
- One classifier is flipping a coin
- 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

Rules of thumb for prediction study design

- If you have a large sample size
 - 60% training
 - 20% test
 - 20% validation
- If you have a medium sample size
 - 60% training
 - 40% testing
- 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 (called [backtesting](#) in finance)
- 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



Types of errors

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Basic terms

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.

Key quantities

		DISEASE	
		+	-
TEST	+	TP	FP
	-	FN	TN

Sensitivity

→ $\Pr(\text{positive test} \mid \text{disease})$

Specificity

→ $\Pr(\text{negative test} \mid \text{no disease})$

Positive Predictive Value

→ $\Pr(\text{disease} \mid \text{positive test})$

Negative Predictive Value

→ $\Pr(\text{no disease} \mid \text{negative test})$

Accuracy

→ $\Pr(\text{correct outcome})$

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

<http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/>

Key quantities as fractions

		DISEASE	
		+	-
TEST	+	TP	FP
	-	FN	TN

Sensitivity

→ $TP / (TP+FN)$

Specificity

→ $TN / (FP+TN)$

Positive Predictive Value

→ $TP / (TP+FP)$

Negative Predictive Value

→ $TN / (FN+TN)$

Accuracy

→ $(TP+TN) / (TP+FP+FN+TN)$

<http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/>

Screening tests

Assume that some disease has a 0.1% prevalence in the population. Assume we have a test kit for that disease that works with 99% sensitivity and 99% specificity. What is the probability of a person having the disease **given the test result is positive**, if we randomly select a subject from

- ▶ the general population?
- ▶ a high risk sub-population with 10% disease prevalence?

General population

		DISEASE	
		+	-
TEST	+	99	999
	-	1	98901

<http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/>

General population as fractions

		DISEASE	
		+	-
		+	99
TEST	+	99	999
	-	1	98901

Sensitivity

$$\rightarrow 99 / (99+1) = 99\%$$

Specificity

$$\rightarrow 98901 / (999+98901) = 99\%$$

Positive Predictive Value

$$\rightarrow 99 / (99+999) \approx 9\%$$

Negative Predictive Value

$$\rightarrow 98901 / (1+98901) > 99.9\%$$

Accuracy

$$\rightarrow (99+98901) / 100000 = 99\%$$

<http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/>

At risk subpopulation

		DISEASE	
		+	-
TEST	+	9900	900
	-	100	89100

<http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/>

At risk subpopulation as fraction

		DISEASE	
		+	-
TEST	+	9900	900
	-	100	89100

Sensitivity $\rightarrow 9900 / (9900+100) = 99\%$

Specificity $\rightarrow 89100 / (900+89100) = 99\%$

Positive Predictive Value $\rightarrow 9900 / (9900+900) \approx 92\%$

Negative Predictive Value $\rightarrow 89100 / (100+89100) \approx 99.9\%$

Accuracy $\rightarrow (9900+89100) / 100000 = 99\%$

Key public health issue

Vast Study Casts Doubts on Value of Mammograms

By GINA KOLATA FEB. 11, 2014



One of the largest and most meticulous studies of mammography ever done, involving 90,000 women and lasting a quarter-century, has added powerful new doubts about the value of the screening test for women of any age.



It found that the death rates from breast cancer and from all causes were the same in women who got mammograms and those who did not. And the screening had harms: One in five cancers found with mammography and treated was not a threat to the woman's health and did not need treatment such as chemotherapy, surgery or radiation.

[The study](#), published Tuesday in The British Medical Journal, is one of the few rigorous evaluations of mammograms in decades.

<http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/>



Nearly 75 percent of American women 40 and over say they had a mammogram in the past year. Damian Dovarganes/Associated Press

Key public health issue

The New York Times

Business Day

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS

Search Global DealBook Markets Economy Energy Media Technology

Looser Guidelines Issued on Prostate Screening

By ANDREW POLLACK

Published: May 3, 2013

In a major shift, the American Urological Association has pulled back its strong support of [prostate cancer](#) screening, saying that the testing should be considered primarily by men aged 55 to 69.

The association had staunchly defended the benefits of screening men with the prostate test, even after a government advisory committee, the United States Preventive Services Task Force, said in 2011 that healthy men should not be screened because far more men would be harmed by unnecessary prostate cancer treatments than would be saved from death.

 FACEBOOK

 TWITTER

 GOOGLE+

 SAVE

 EMAIL

 SHARE

 PRINT

 REPRINTS

For continuous data

Mean squared error (MSE):

$$\frac{1}{n} \sum_{i=1}^n (\text{Prediction}_i - \text{Truth}_i)^2$$

Root mean squared error (RMSE):


$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\text{Prediction}_i - \text{Truth}_i)^2}$$

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](#)



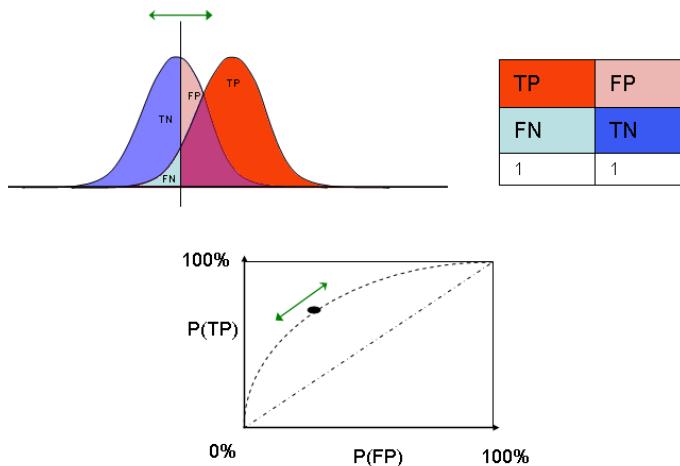
ROC curves

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Why a curve?

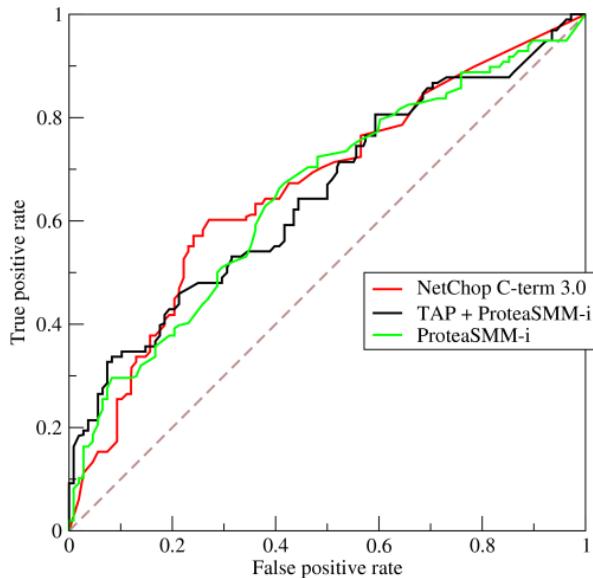
- In binary classification you are predicting one of two categories
 - Alive/dead
 - Click on ad/don't click
- But your predictions are often quantitative
 - Probability of being alive
 - Prediction on a scale from 1 to 10
- The *cutoff* you choose gives different results

ROC curves



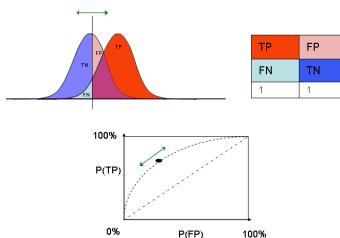
http://en.wikipedia.org/wiki/Receiver_operating_characteristic

An example



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

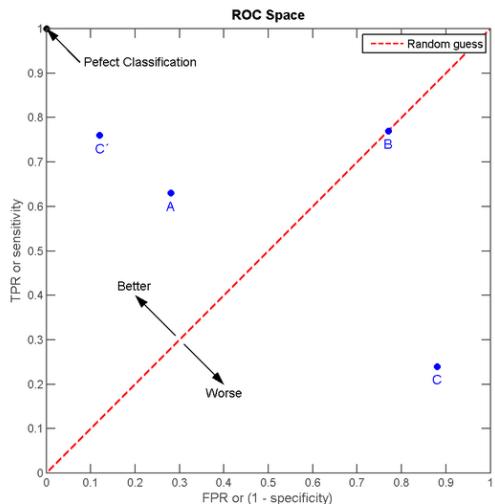
Area under the curve



- AUC = 0.5: random guessing
- AUC = 1: perfect classifier
- In general AUC of above 0.8 considered "good"

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

What is good?



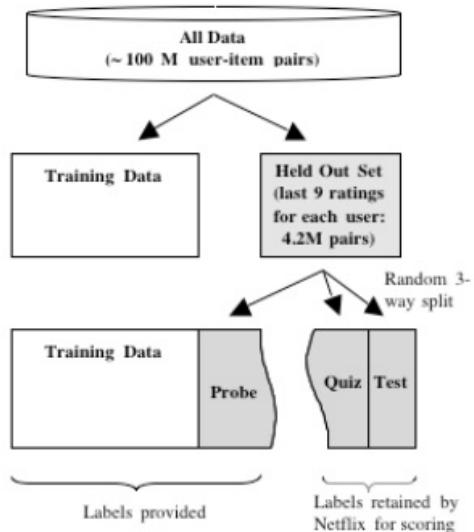
http://en.wikipedia.org/wiki/Receiver_operating_characteristic



Cross validation

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Study design



<http://www2.research.att.com/~volinsky/papers/ASASStatComp.pdf>

Key idea

1. Accuracy on the training set (resubstitution accuracy) is optimistic
2. A better estimate comes from an independent set (test set accuracy)
3. But we can't use the test set when building the model or it becomes part of the training set
4. So we estimate the test set accuracy with the training set.

Cross-validation

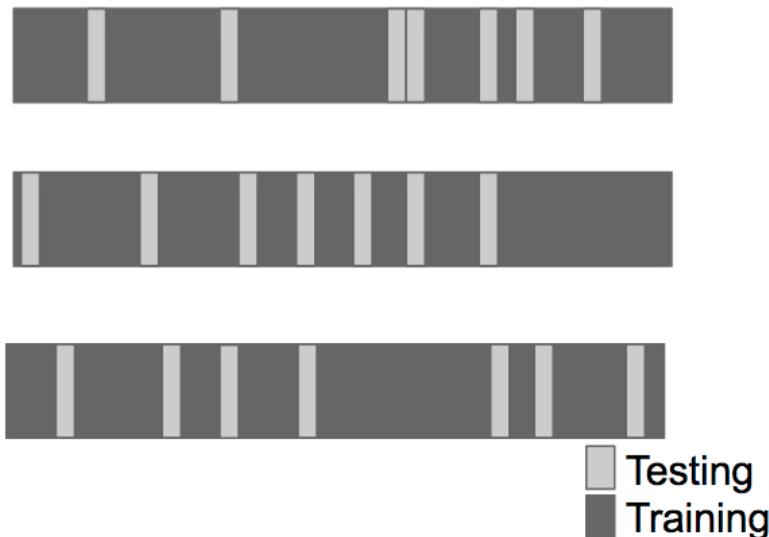
Approach:

1. Use the training set
2. Split it into training/test sets
3. Build a model on the training set
4. Evaluate on the test set
5. Repeat and average the estimated errors

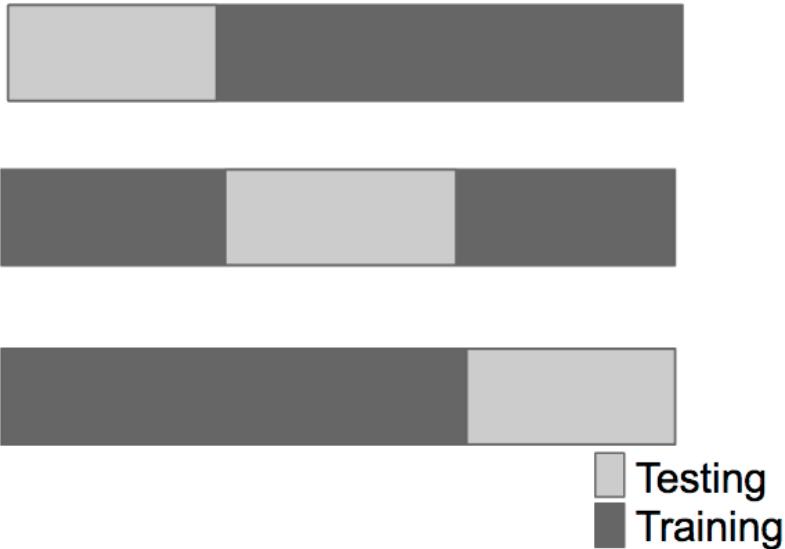
Used for:

1. Picking variables to include in a model
2. Picking the type of prediction function to use
3. Picking the parameters in the prediction function
4. Comparing different predictors

Random subsampling



K-fold



Leave one out



Considerations

- For time series data data must be used in "chunks"
- For k-fold cross validation
 - Larger k = less bias, more variance
 - Smaller k = more bias, less variance
- Random sampling must be done *without replacement*
- Random sampling with replacement is the *bootstrap*
 - Underestimates of the error
 - Can be corrected, but it is complicated ([0.632 Bootstrap](#))
- If you cross-validate to pick predictors estimate you must estimate errors on independent data.



What data should you use?

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

A successful predictor



fivethirtyeight.com

Polling data

Home GALLUP® Search Gallup.com

HOME POLITICS ECONOMY WELL-BEING WORLD GALLUP ANALYTICS

HOT TOPICS: Healthcare Law Guns U.S. Government Shutdown Iran Syria Russia U.S. Leadership Approval Race Relations T.

One in Four U.S. Uninsured Plan to Remain That Way



December 3, 2013

Twenty-eight percent of uninsured Americans say they are more likely to pay the fine for not having health insurance than to obtain insurance, as required by the healthcare law. Politics appear to be a major factor in that decision.

U.S. Economic Confidence Rises in November

December 3, 2013

U.S. Economic Confidence Index, Monthly Averages



Date	Index Value
Jan '12	-22
May '12	-26
Sep '12	-27
Jan '13	-16
May '13	-7
Sep '13	-13
Dec '13	-19
Jan '14	-25
May '14	-35

Inside Strategic Consulting

<http://www.gallup.com/>

Weighting the data

FiveThirtyEight

6.06.2010

Pollster Ratings v4.0: Methodology

by [Nate Silver](#)

Rating pollsters is at the core of FiveThirtyEight's mission, and forms the backbone of our forecasting models. But, it has been two years since we [last revised our ratings](#). Here, at last, is an update. We have both substantially increased the amount of data that we are evaluating, and significantly refined our methodology.

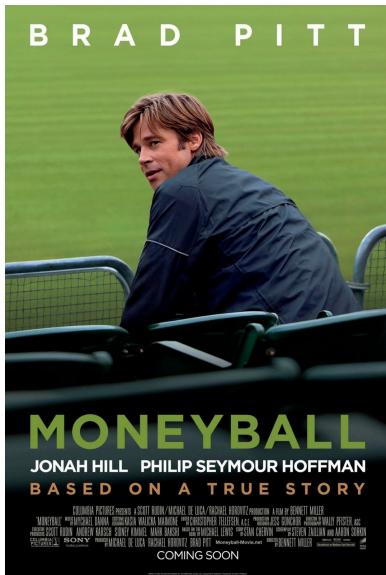
<http://www.fivethirtyeight.com/2010/06/pollster-ratings-v40-methodology.html>

Key idea

To predict X use data related to X

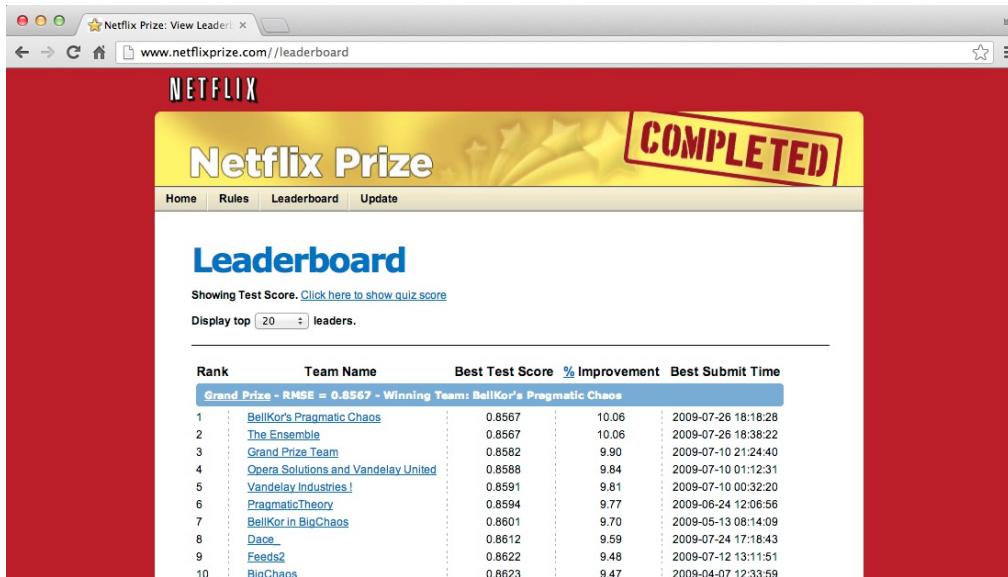
Key idea

To predict player performance use data about player performance



Key idea

To predict movie preferences use data about movie preferences



The screenshot shows a web browser window for the Netflix Prize Leaderboard at www.netflixprize.com//leaderboard. The page has a red header with the Netflix logo and a large yellow banner that says "COMPLETED". Below the banner, there's a navigation bar with links for Home, Rules, Leaderboard, and Update. The main section is titled "Leaderboard" and displays a table of top teams. A note above the table says "Showing Test Score. Click here to show quiz score". The table has columns for Rank, Team Name, Best Test Score, % Improvement, and Best Submit Time. The winning team, "BellKor's Pragmatic Chaos", is highlighted in a blue row with a RMSE of 0.8567.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8562	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8568	9.84	2009-07-10 01:12:31
5	Vandelay Industries!	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor In BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59

Key idea

To predict hospitalizations use data about hospitalizations

Information Data Forum Leaderboard



**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.

Not a hard rule

To predict flu outbreaks use Google searches



<http://www.google.org/flutrends/>

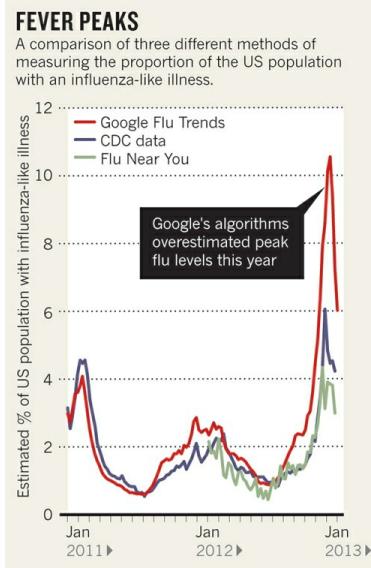
Looser connection = harder prediction

Oncotype DX® reveals
the underlying biology that
changes treatment decisions
37% of the time

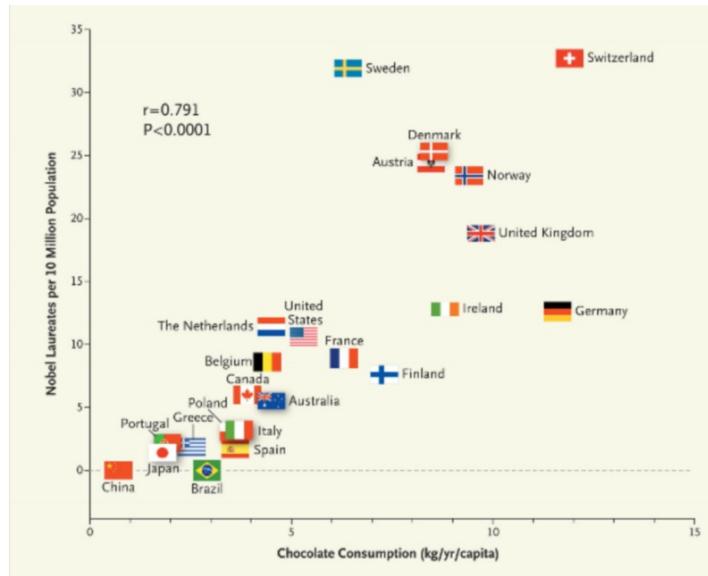
Uncover the Unexpected™



Data properties matter



Unrelated data is the most common mistake



<http://www.nejm.org/doi/full/10.1056/NEJMoa1211064>