Introduction to data mining

Ryan Tibshirani
Data Mining: 36-462/36-662

January 15 2013

Logistics

- Course website (syllabus, lectures slides, homeworks, etc.): `http://www.stat.cmu.edu/~ryantibs/datamining`
- We will use blackboard site for email list, grades
- 6 homeworks, top 5 will count
- 2 in-class exams, final project
- Final project in groups of 2-3, will be fun!

Prerequisites:

- Only formal one is 36-401
- Assuming you know basic probability and statistics, linear algebra, R programming (see syllabus for topics list)

Textbooks:

- Course textbook *Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani. Get it online at `http://www-bcf.usc.edu/~gareth/ISL`, use the login name: `StatLearn`, password: `book`. Not yet published, do not distribute!
- Optional, more advanced textbook: *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman. Also available online at `http://www-stat.stanford.edu/ElemStatLearn`

Course staff:

- Instructor: Ryan Tibshirani (you can call me "Ryan" or "Professor Tibshirani", please not "Professor")
- TAs: Li Liu, Cong Lu, Jack Rae, Michael Vespe

Why are you here?

- Because you love the subject, because it's required, because you eventually want to make $$$ ...
- No matter the reason, everyone can get something out of the course
- Work hard and have fun!

What is data mining?

Data mining is the science of discovering structure and making predictions in (large) data sets

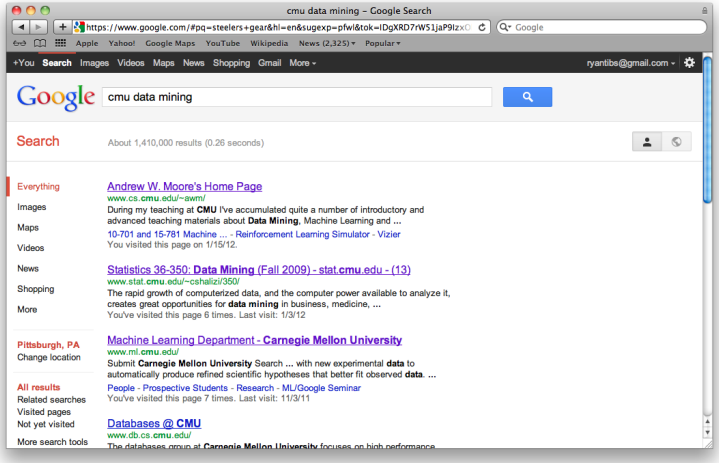
- Unsupervised learning: discovering structure

  E.g., given measurements $X_1, \ldots X_n$, learn some underlying group structure based on similarity

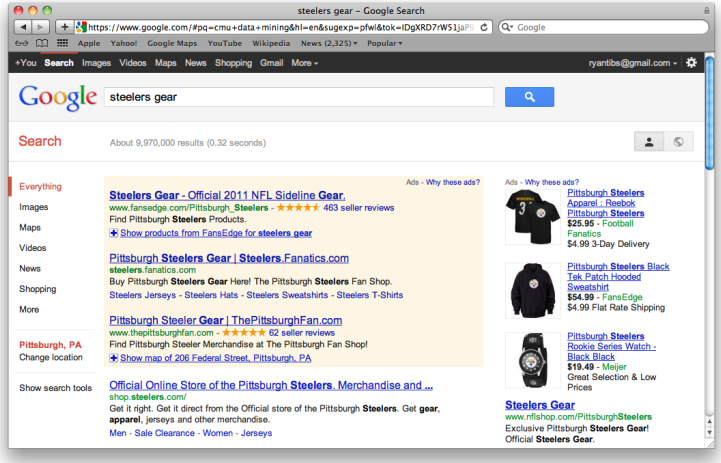
- Supervised learning: making predictions

  I.e., given measurements $(X_1, Y_1), \ldots (X_n, Y_n)$, learn a model to predict Y_i from X_i

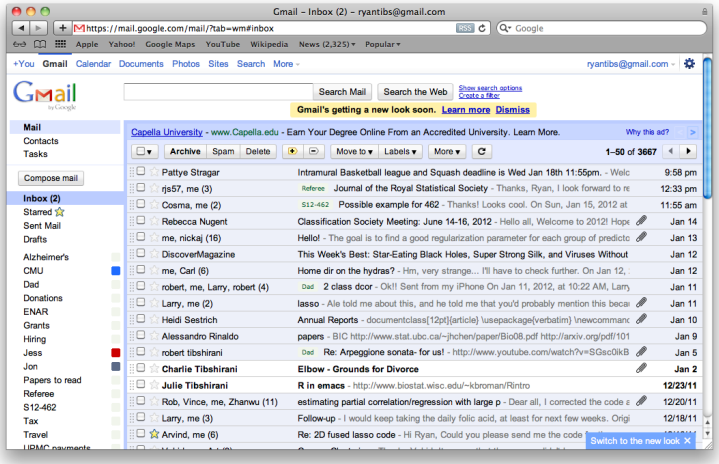
Google



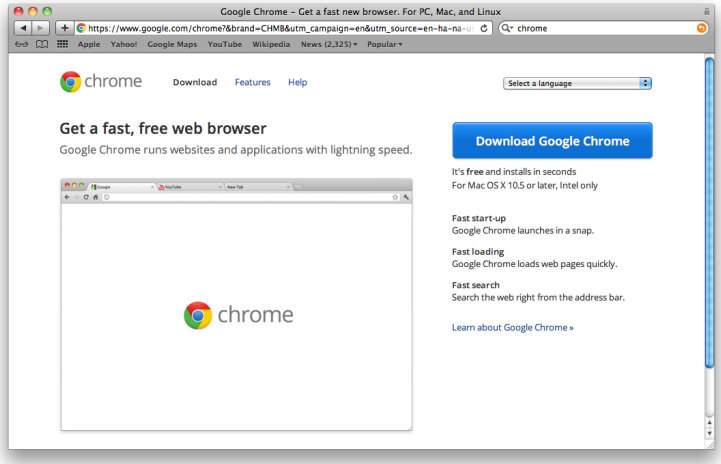
Search



Ads

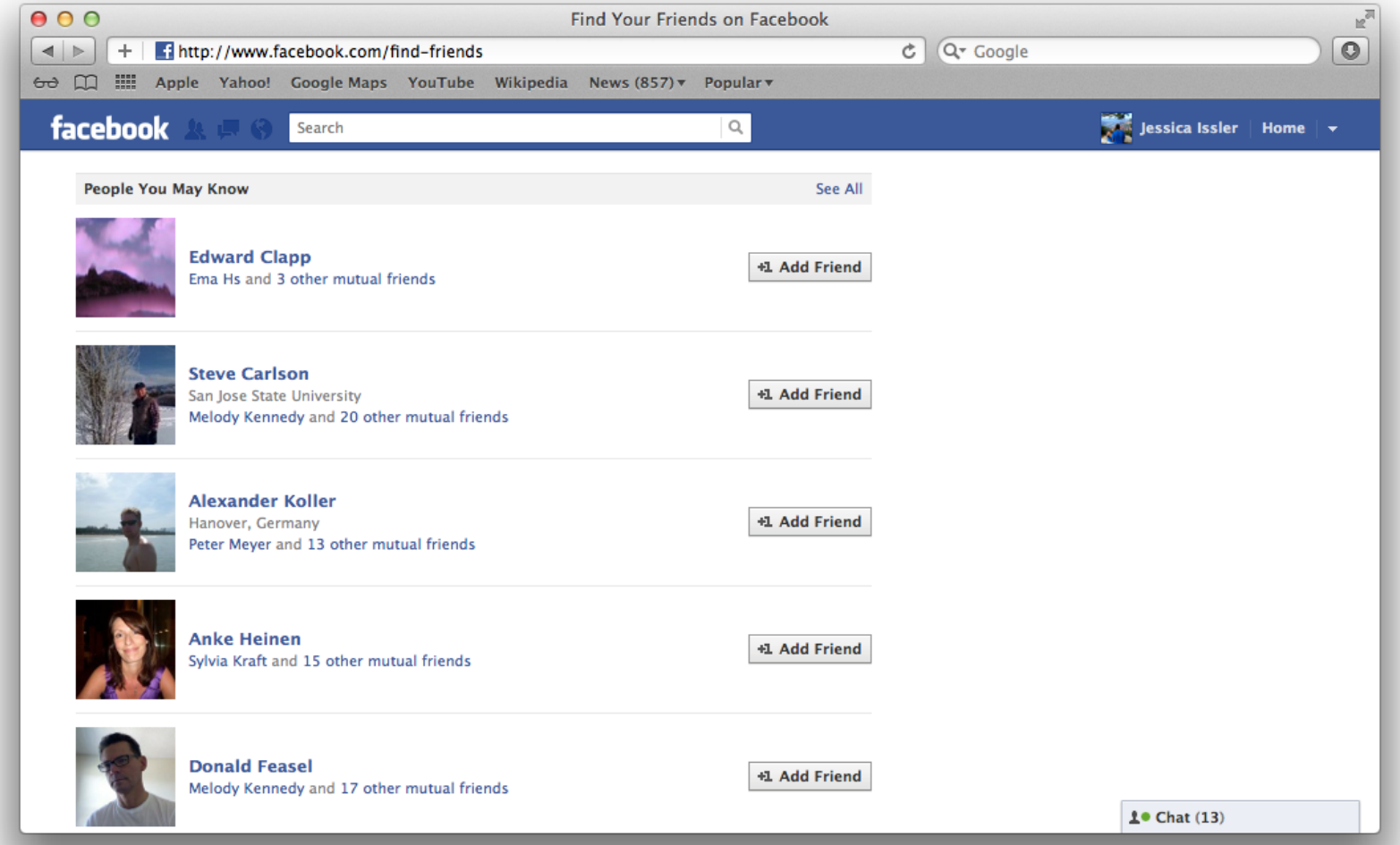


Gmail



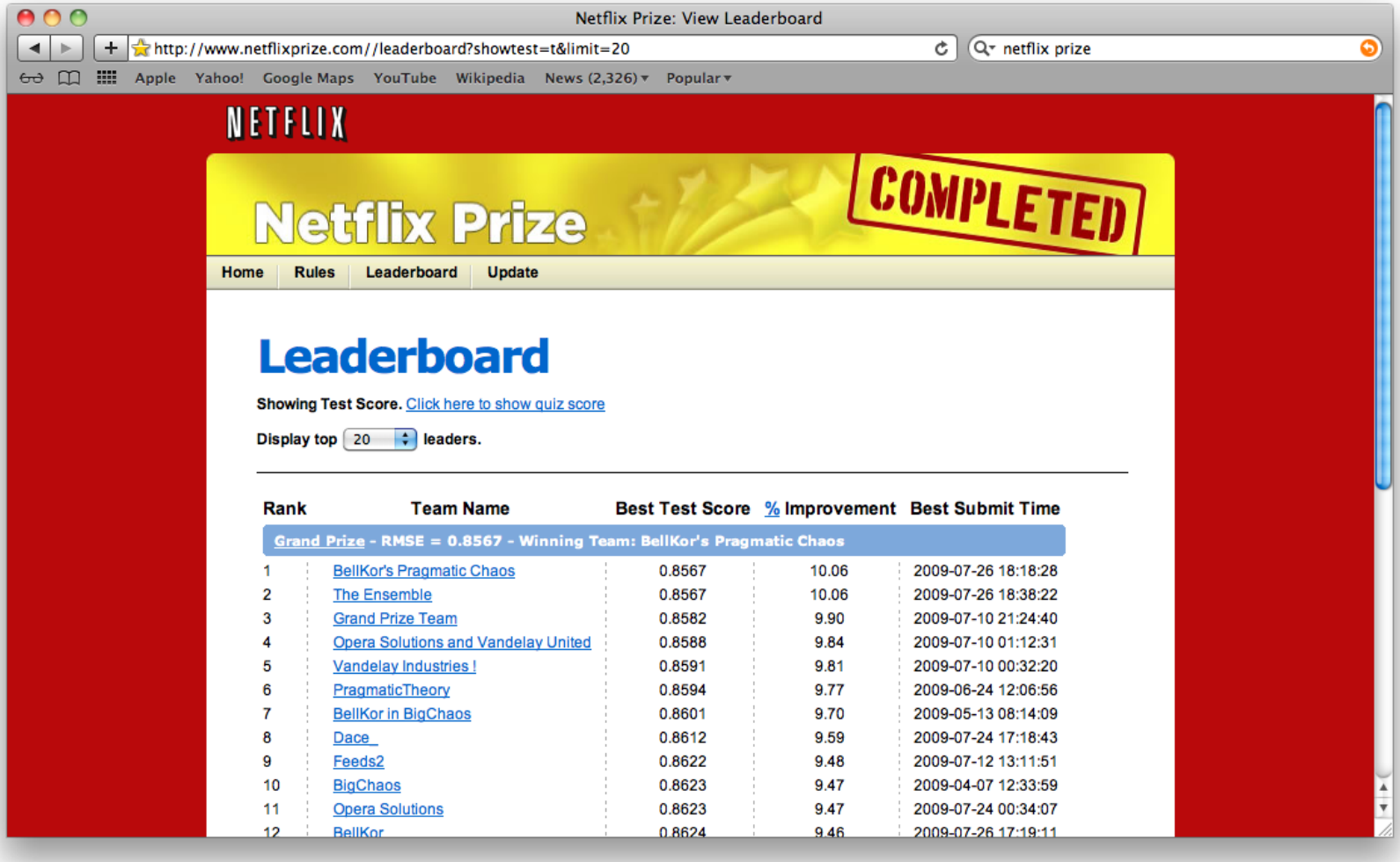
Chrome

Facebook



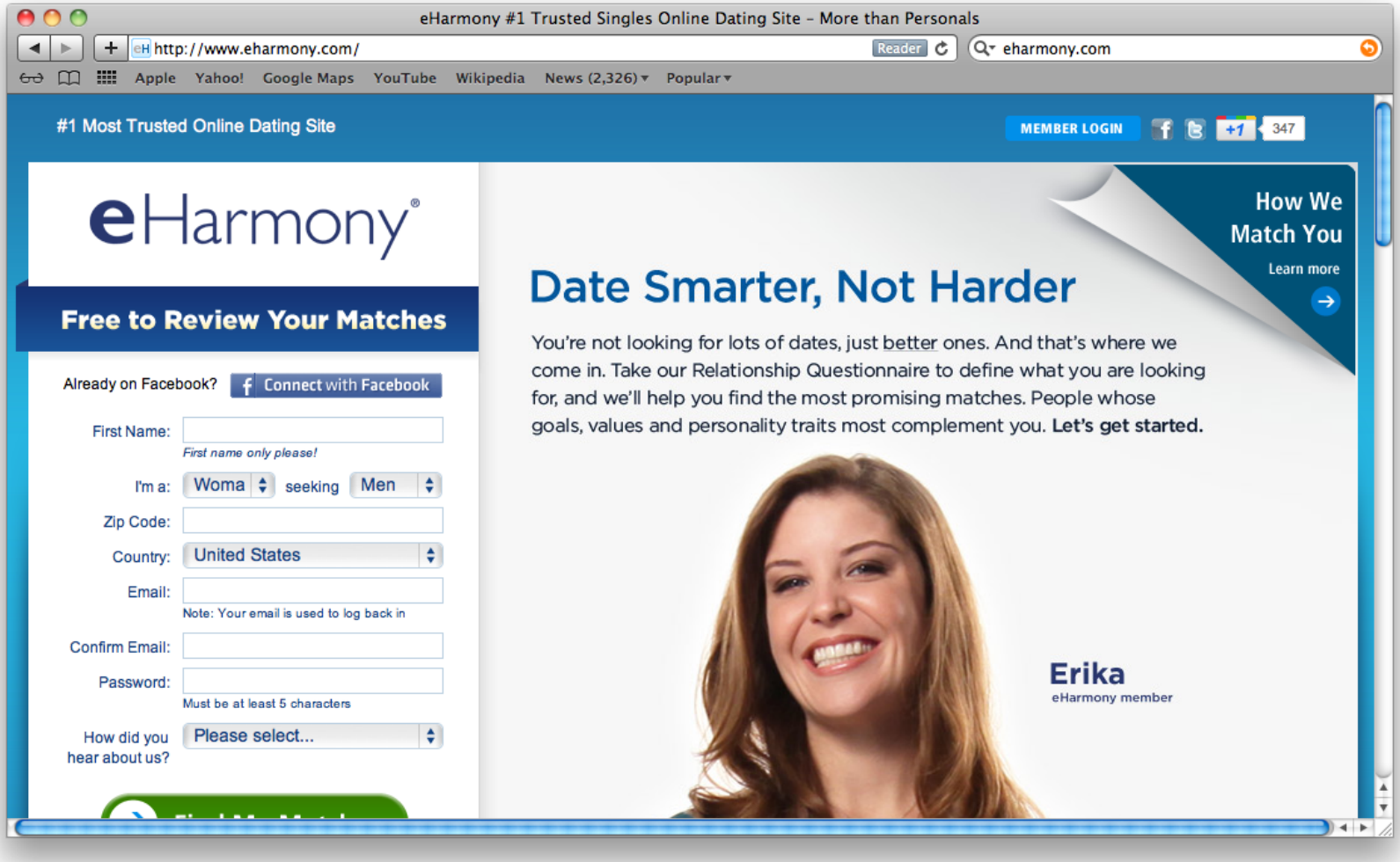
People you may know

Netflix



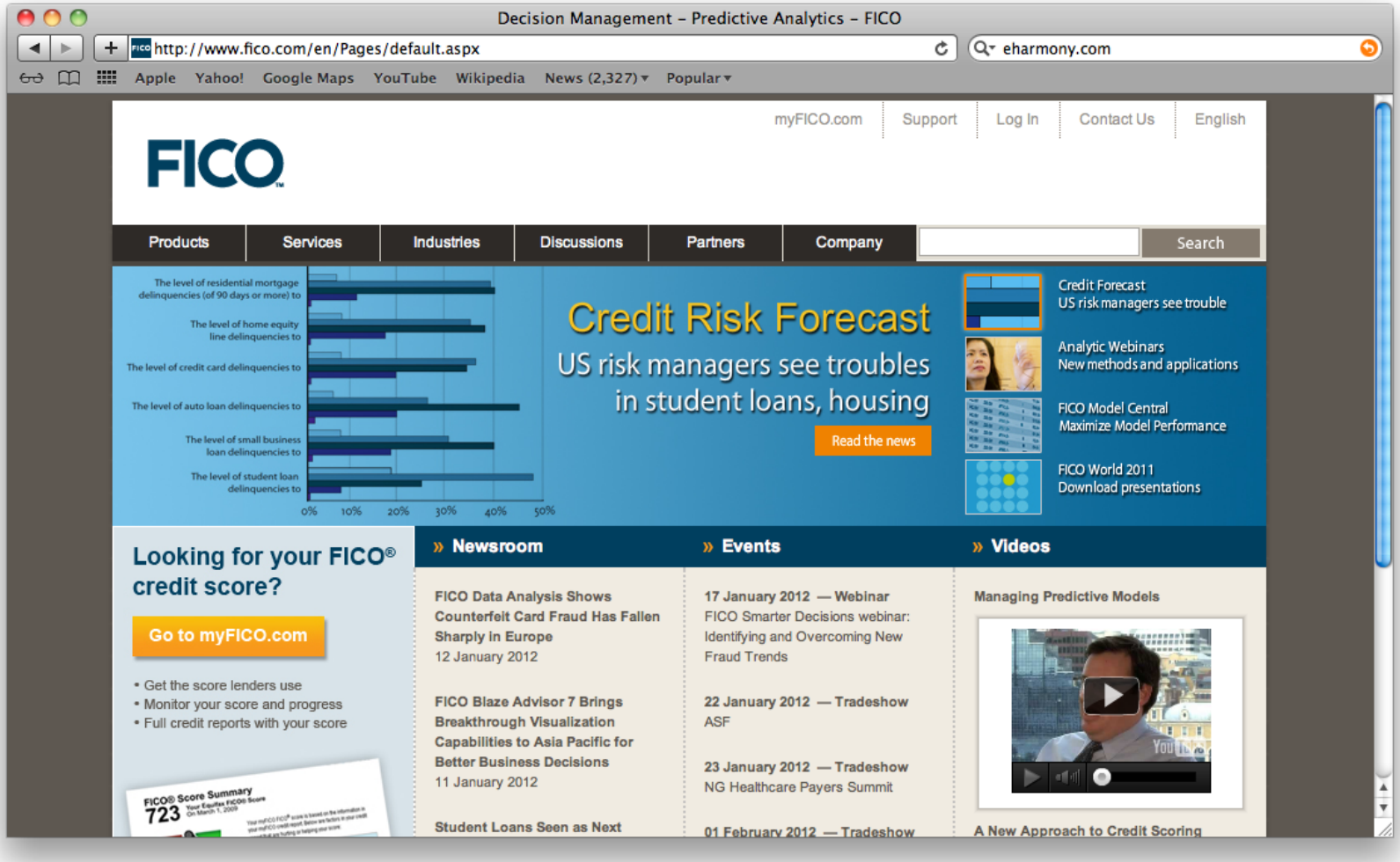
$1M prize!

eHarmony



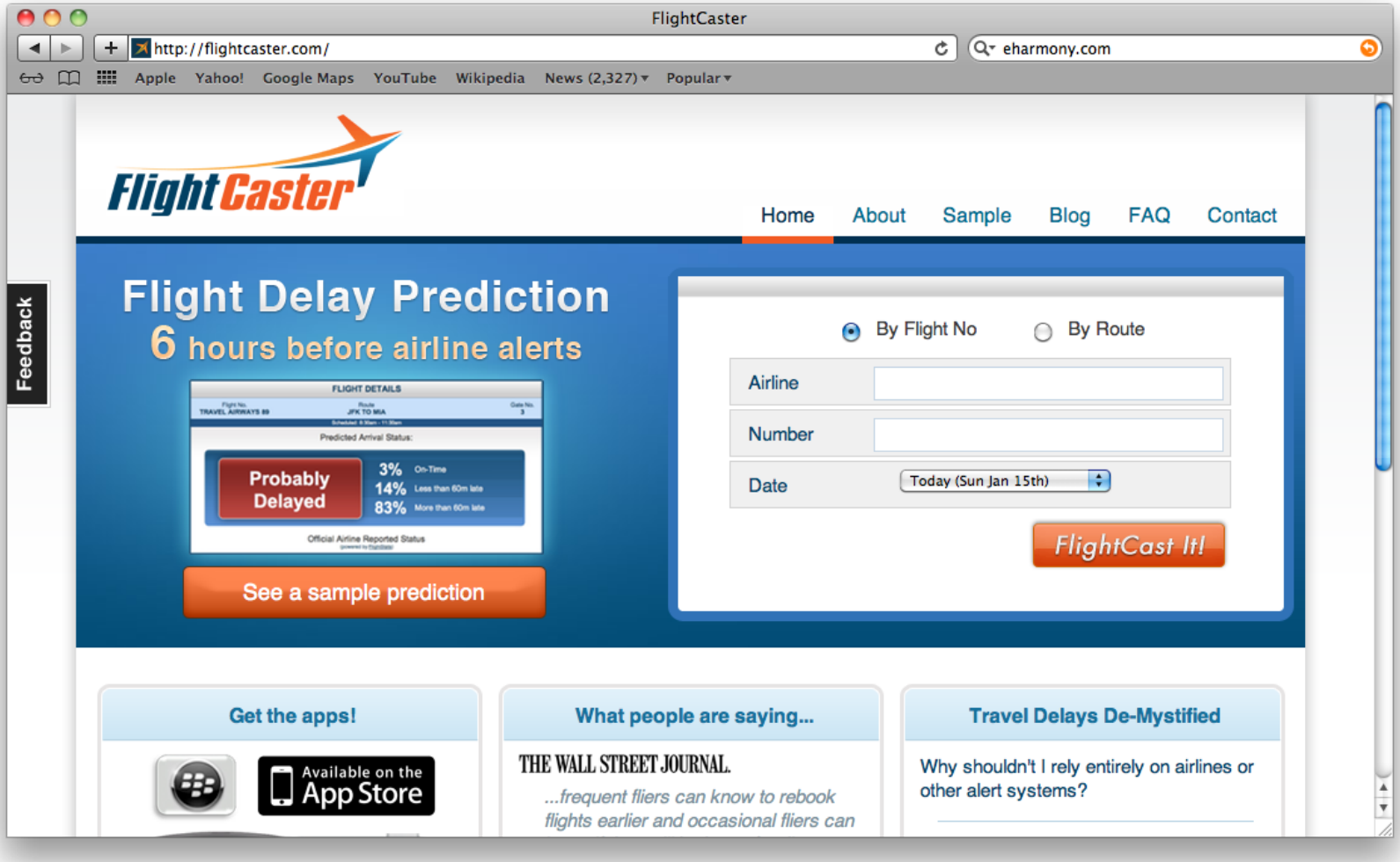
Falling in love with statistics

FICO



An algorithm that could cause a lot of grief

FlightCaster

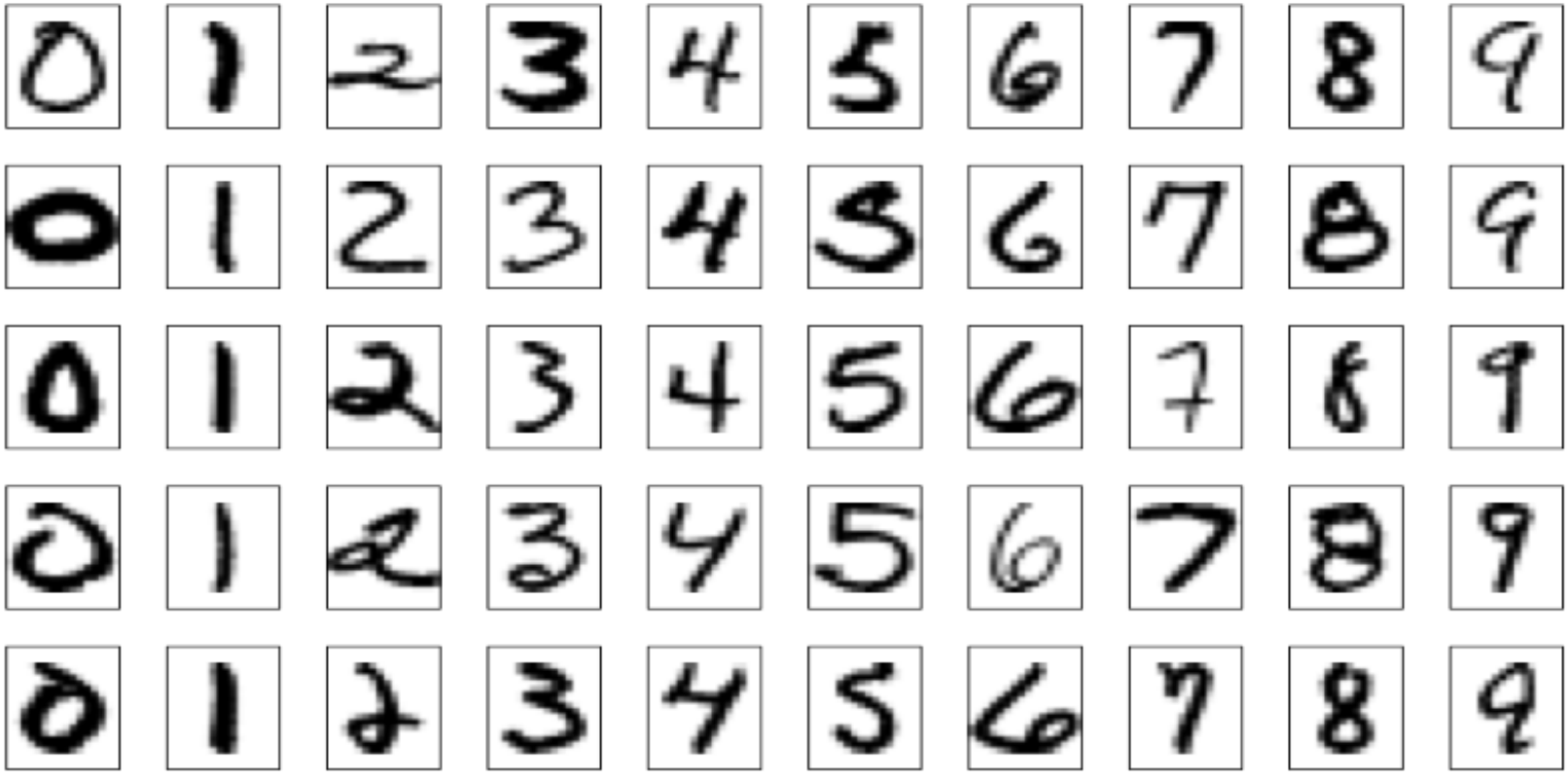


Apparently it's even used by airlines themselves

IBM's Watson

A combination of many things, including data mining

Handwritten postal codes

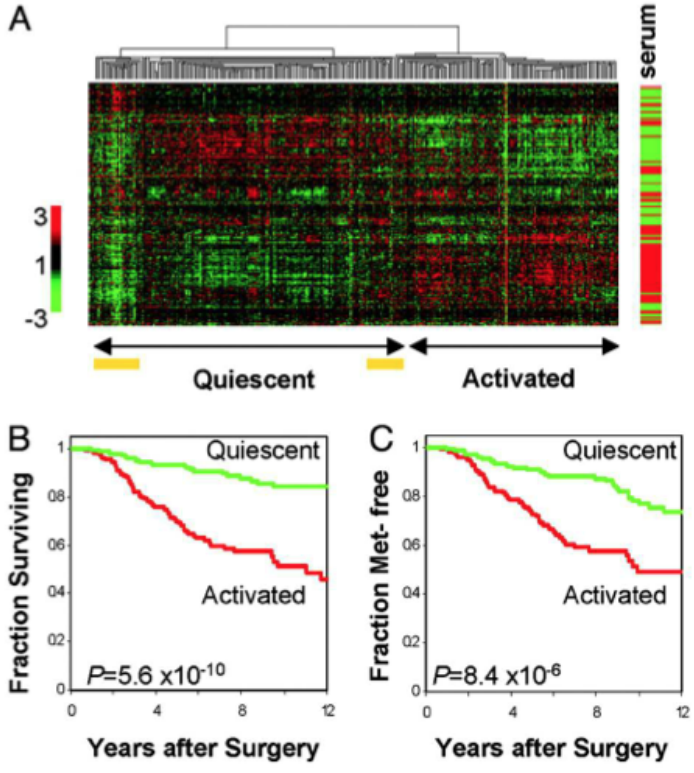


(From ESL p. 404)

We could have robot mailmen someday

Subtypes of breast cancer

Fig. 1. Performance of a "wound response" gene expression signature in predicting breast cancer progression

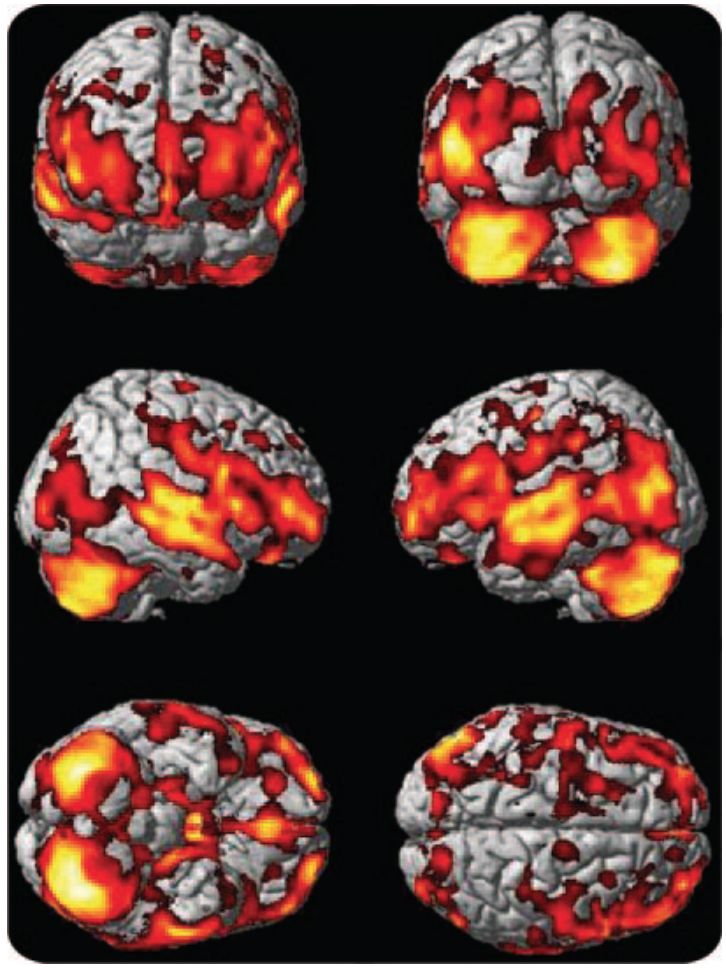


Chang, Howard Y. et al. (2005) Proc. Natl. Acad. Sci. USA 102, 3738-3743

PNAS

Subtypes of breastcancer based on wound response

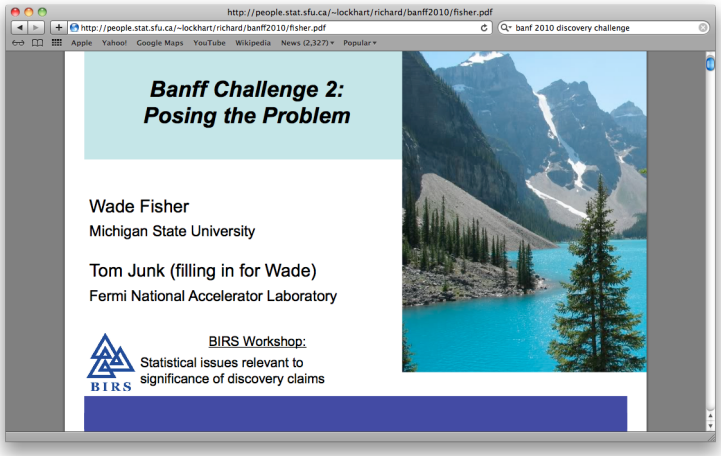
Predicting Alzheimer's disease

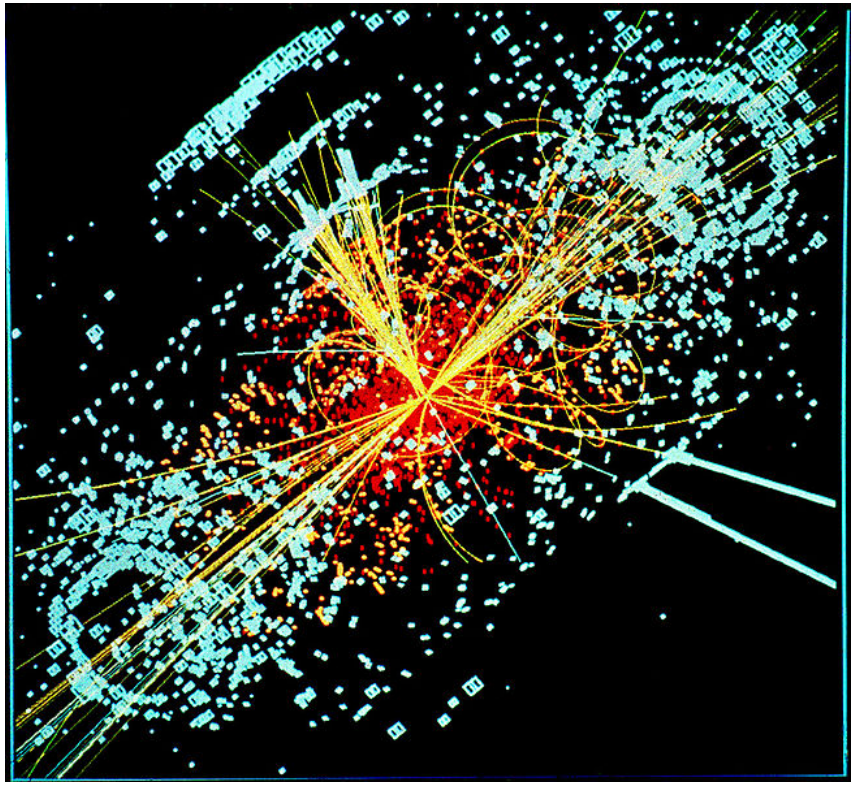


(From Raji et al. (2009), "Age, Alzheimer's disease, and brain structure")

Can we predict Alzheimer's disease years in advance?

Banff 2010 challenge





Find the Higgs boson particle and win a Nobel prize!

Will it be found by a statistician?

What to expect

Expect both applied and theoretical perspectives ... not just a course where we open up R and go from there, we also rigorously investigate the topics we learn

Why? Because success in data mining comes from a synergy between practice and theory

- You can't always open up R, download a package, and get a reasonable answer
- Real data is messy and always presents new complications
- Understanding why and how things work is a necessary precursor to figuring out what to do

Reoccuring themes

Exact approach versus approximation: often when we can't do something exactly, we'll settle for an approximation. Can perform well, and scales well computationally to work for large problems

Bias-variance tradeoff: nearly every modeling decision is a tradeoff between bias and variance. Higher model complexity means lower bias and higher variance

Interpretability versus predictive performance: there is also usually a tradeoff between a model that is interpretable and one that predicts well under general circumstances

There's not a universal recipe book

Unfortunately, there's no universal recipe book for when and in what situations you should apply certain data mining methods

Statistics doesn't work like that. Sometimes there's a clear approach; sometimes there is a good amount of uncertainty in what route should be taken. That's what makes it so hard, and so fun

This is true even at the expert level (and there are even larger philosophical disagreements spanning whole classes of problems)

The best you can do is try to understand the problem, understand the proposed methods and what assumptions they are making, and find some way to evaluate their performances

Next time: information retrieval

	but	cool	dude	party	michelangelo	raphael	rude	...
doc 1	19	0	0	0	4	24	0	
doc 2	8	1	0	0	7	45	1	
doc 3	7	0	4	3	77	23	0	
doc 4	2	0	0	0	4	11	0	
doc 5	17	0	0	0	9	6	0	
doc 6	36	0	0	0	17	101	0	
doc 7	10	0	0	0	159	2	0	
doc 8	2	0	0	0	0	0	0	
query	1	1	1	1	1	1	1	