Spinning Wheels

Or: Three Ideas, Minimal Execution

Idea #1: Predict TV Ratings

Final broadcast primetime ratings for Monday, September 29, 2014 (Live + Same Day) :

Time	Net	Show	18-49 Rtg/Shr	Viewers (millions)
8:00	CBS	The Big Bang Theory (8-8:31PM)	4.8/15	16.38
	NBC	The Voice	4.0/11	12.94
	FOX	Gotham	2.8/8	7.45
	ABC	Dancing with the Stars (8-10:01PM)	1.8/5	12.50
	CW	iHeartradio Music Festival Night 1 (8-10PM)	0.3/1	0.82
8:30	CBS	The Big Bang Theory - R (8:31-9:01PM)	3.1/9	12.26
9:00	CBS	Scorpion (9:01-9:59PM)	3.1/8	13.36
	FOX	Sleepy Hollow	1.7/5	5.04
10:00	NBC	The Blacklist	2.8/8	10.51
	ABC	Castle - Season Premiere (10:01-11PM)	2.2/6	10.75
	CBS	NCIS: Los Angeles - Season Premiere (9:59-11PM)	1.9/6	9.48

-

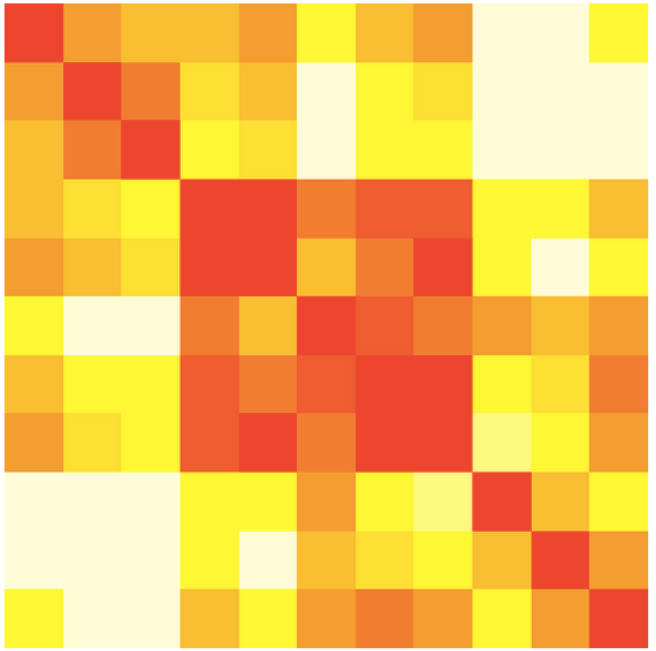
Issues with TV Rating Prediction

1) Scraping ratings is time-consuming
   a. Different types of shows are in different tables
   b. Different tables are available for different lengths of time before being deleted
2) Difficult to find generalizable show attributes (Big Bang Problem)
3) Extrapolation concerns given limited variation in feature space
   a. Same show at same time on same channel
   b. Monday Night Football ratings tell you little about how Wednesday Night Football would do

Idea #2: Grouping Observational Units with MIC

```r
clusterRsquared <- function(dataframe) {
  dissimilarity <- 1 - cor(dataframe)^2
  clustering <- hclust(as.dist(dissimilarity))
  order <- clustering$order
  oldpar <- par(no.readonly=TRUE); par(mar=c(0,0,0,0))
  image(dissimilarity[order, rev(order)], axes=FALSE)
  par(oldpar)
  return(1 - dissimilarity[order, order])
}
```

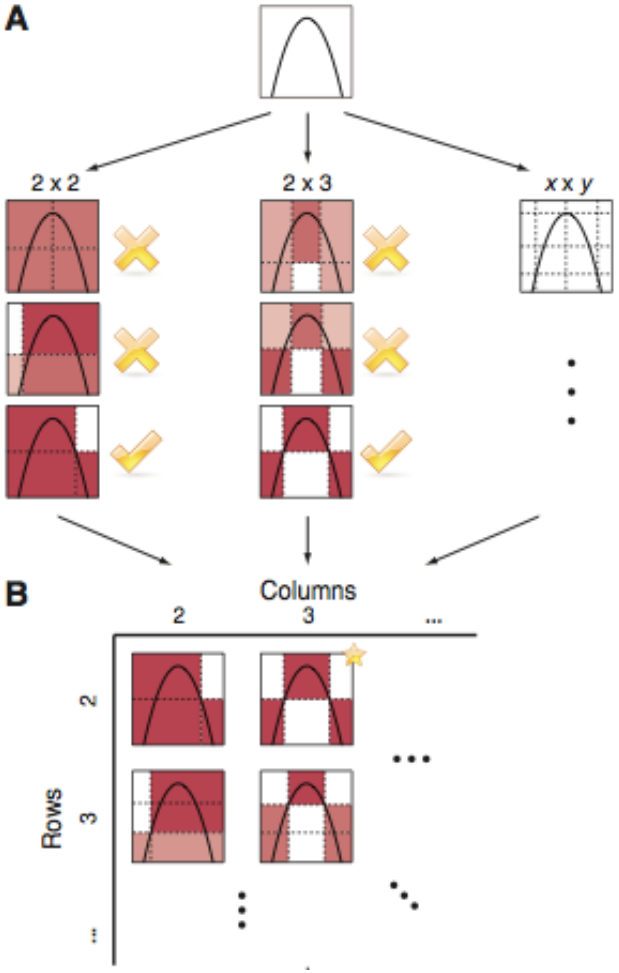
	drat	am	gear	mpg	wt	hp	cyl	disp	carb	qsec	vs
drat	1.00	0.51	0.49	0.46	0.51	0.20	0.49	0.50	0.01	0.01	0.19
am	0.51	1.00	0.63	0.36	0.48	0.06	0.27	0.35	0.00	0.05	0.03
gear	0.49	0.63	1.00	0.23	0.34	0.02	0.24	0.31	0.08	0.05	0.04
mpg	0.46	0.36	0.23	1.00	0.75	0.60	0.73	0.72	0.30	0.18	0.44
wt	0.51	0.48	0.34	0.75	1.00	0.43	0.61	0.79	0.18	0.03	0.31
hp	0.20	0.06	0.02	0.60	0.43	1.00	0.69	0.63	0.56	0.50	0.52
cyl	0.49	0.27	0.24	0.73	0.61	0.69	1.00	0.81	0.28	0.35	0.66
disp	0.50	0.35	0.31	0.72	0.79	0.63	0.81	1.00	0.16	0.19	0.50
carb	0.01	0.00	0.08	0.30	0.18	0.56	0.28	0.16	1.00	0.43	0.32
qsec	0.01	0.05	0.05	0.18	0.03	0.50	0.35	0.19	0.43	1.00	0.55
vs	0.19	0.03	0.04	0.44	0.31	0.52	0.66	0.50	0.32	0.55	1.00



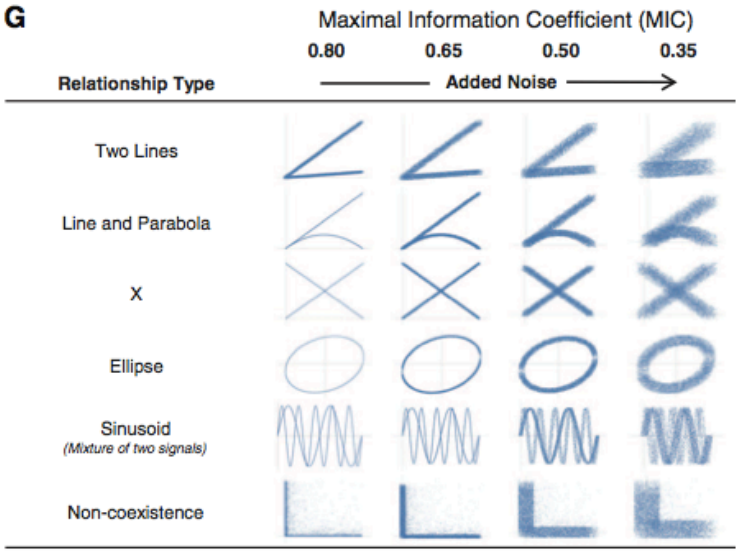
Concept: Replace R^2 with MIC to pick up non-linear relationships

Motivating Example: Identify proteins/hormones/ions in a feedback loop in biological systems

Maximal Information Criteria



Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76



Issues with MIC Grouping

1) Finding suitable data
   a. http://yeastmine.yeastgenome.org/ does not have longitudinal measures (or I could not find them)
   b. Longitudinal stock prices maybe out there somewhere (pandas.io.web.get_data_yahoo ?)
2) Not obvious how to pluck out strongly related groups and discard isolated variables in an automated manner
   a. Hierarchical clustering thing Aaron did?
   b. Graph partitioning algorithm?

Idea #3: Cascades Thwarting Wisdom of Crowds

Alinea Like Page
November 13, 2012 ·

We are updating our database structure for diner history -- and did a data dump. So for fun -- and perhaps a prize or two -- guess how many last names we have dining notes on at Alinea.

-- Tie Breaker... how many start with the letter Z (and yes, FB fans, Zuckerberg is in there, apparently 'in town to film Oprah').

This is May 4, 2005 until today.

Like · Comment · Share

142 people like this.

View previous comments 50 of 1,182

Joe Monett 170,000 6000 z's
November 14, 2012 at 1:39pm · Like

Daniel Gutenplan 78,886 and 546
November 14, 2012 at 2:06pm · Like

Ashley Cathcart 34,555 names & 100 Zs
November 14, 2012 at 2:11pm · Like

Jim Perkins 38,142 total and 265 z's
November 14, 2012 at 2:15pm · Like

Leigha Kemmett 36,000 names, 600 Zs
November 14, 2012 at 2:24pm · Like

Irene Rozenberg 210,000 names, and 10,000 z's
November 14, 2012 at 2:38pm · Like

Pat Russ 6,160 18Z
November 14, 2012 at 2:58pm · Like

Adit Kumar 6,700 names, 25 Zs
November 14, 2012 at 3:07pm · Like

Barry McCardel 10,203 names, 300 Z's
November 14, 2012 at 3:09pm · Like

Bill Moylan 7488 202
November 14, 2012 at 3:11pm · Like

Drew Mosley 279,000 and 587 z's
November 14, 2012 at 3:12pm · Like

Average Guess of Number of Last Names: 98,335

Correct Answer: 23,980

Guaranteed Smooth Sailing

datascope analytics
data-driven consulting and design

Wisdom of the crowd?

November 15, 2012

Alinea, the Chicago restaurant known for combining food, science and art, recently asked people on their Facebook page to guess how many last names have made reservations at the restaurant since it opened on May 4, 2005. I decided to take stab at the question. Nerd that I am, though, I also looked at other people's guesses - which turned out to be more interesting than answering the question itself.

Our guess

First, I estimated the total number of tables seated at Alinea since opening day. If Alinea has 20 tables (number of tables), and about 2 parties get seated at each table a night, and the restaurant is open 250 nights a year, and it's been open for 7.5 years, that works out to 75,000 table seatings. You can only get a table at Alinea if you get a reservation and presumably they add your name to their "dining notes" database when you reserve a

"There's a lot of correlation between a person's guess and what others guessed in the comments right before them (on Facebook, you can click to see the 50 previous comments before yours)…"

The Guarantee Wasn't 100%

```
> cor(data[0:7])^2
              guess         lag1         lag2         lag3         lag4         lag5         lag6
guess 1.000000e+00 4.239768e-05 3.588500e-04 4.445246e-04 9.954928e-03 4.495994e-06 1.493469e-05
lag1  4.239768e-05 1.000000e+00 4.160080e-05 3.631596e-04 4.428680e-04 9.982531e-03 4.685900e-06
lag2  3.588500e-04 4.160080e-05 1.000000e+00 4.053078e-05 3.652549e-04 4.397010e-04 9.988184e-03
lag3  4.445246e-04 3.631596e-04 4.053078e-05 1.000000e+00 3.935641e-05 3.687548e-04 4.368959e-04
lag4  9.954928e-03 4.428680e-04 3.652549e-04 3.935641e-05 1.000000e+00 3.815051e-05 3.687980e-04
lag5  4.495994e-06 9.982531e-03 4.397010e-04 3.687548e-04 3.815051e-05 1.000000e+00 3.688599e-05
lag6  1.493469e-05 4.685900e-06 9.988184e-03 4.368959e-04 3.687980e-04 3.688599e-05 1.000000e+00
```

Random Forests rarely improve the MSE relative to the MSE when predicting the mean guess always

Correlation between guess and average of 50 previous guesses is a whopping 0.0006197145

Potential Issues with Alinea Data

1) Constructed lagged variables in a wrong way
2) No “there” there: perhaps previous guesses don’t influence a person’s guess
3) Heterogeneous effects: perhaps different people respond to different prior guesses differently in a way that leads to no overall effect of lagged guesses