So You Want To Be a Data Miner?

The 2015 Data Mining Cup

What to Expect When You're Predicting

A link to our github page

What Is The Data Mining Cup and Why Are We Here?

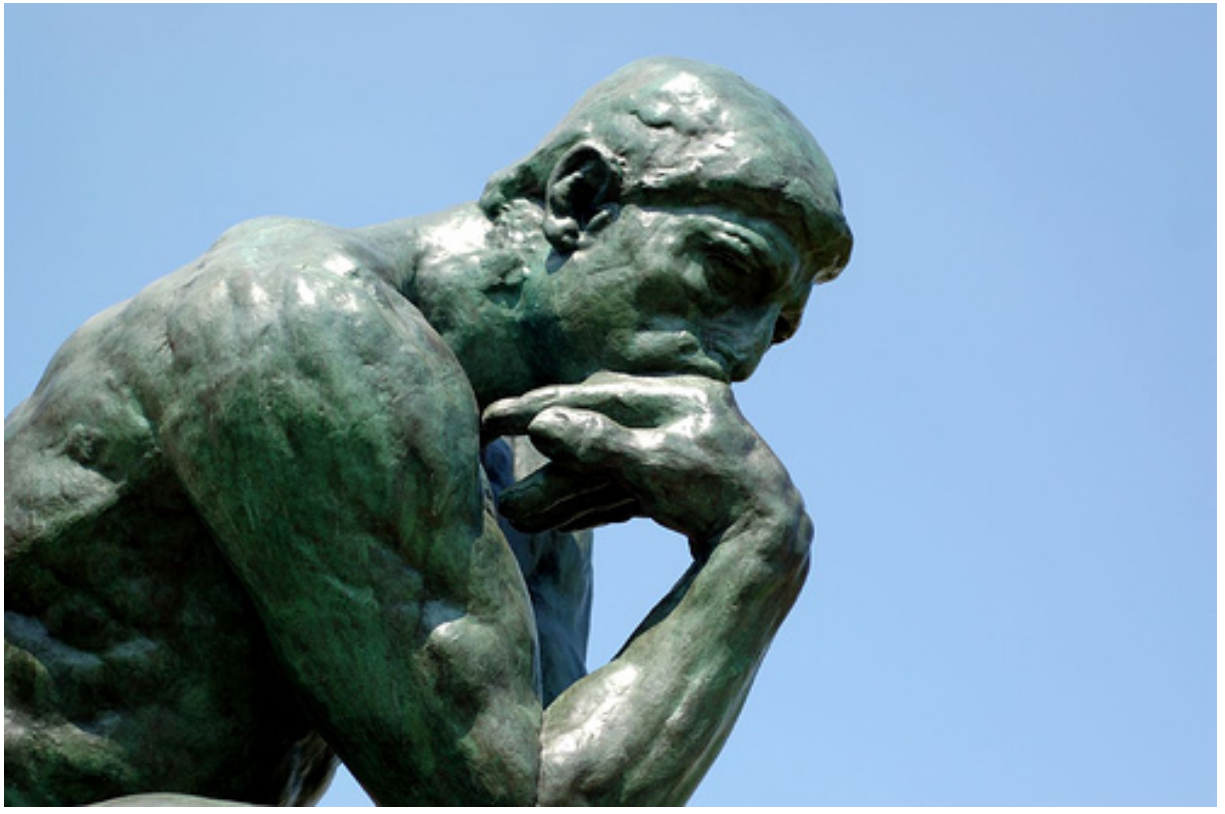
What is it?

The Data Mining Cup is a yearly competition hosted by prudsys (all lower case) a German analytics company focused on marketplace behavior. prudsys describes it this way:

> The DATA MINING CUP (DMC for short) has been inspiring students around the world to pursue intelligent data analysis since the year 2000. In 2014 over one thousand students from about 100 universities in 28 countries took part in the competition. The best teams will be invited to Berlin for the awards ceremony at the prudsys personalization summit.

What is it?

Why am I here?



Only you can answer that

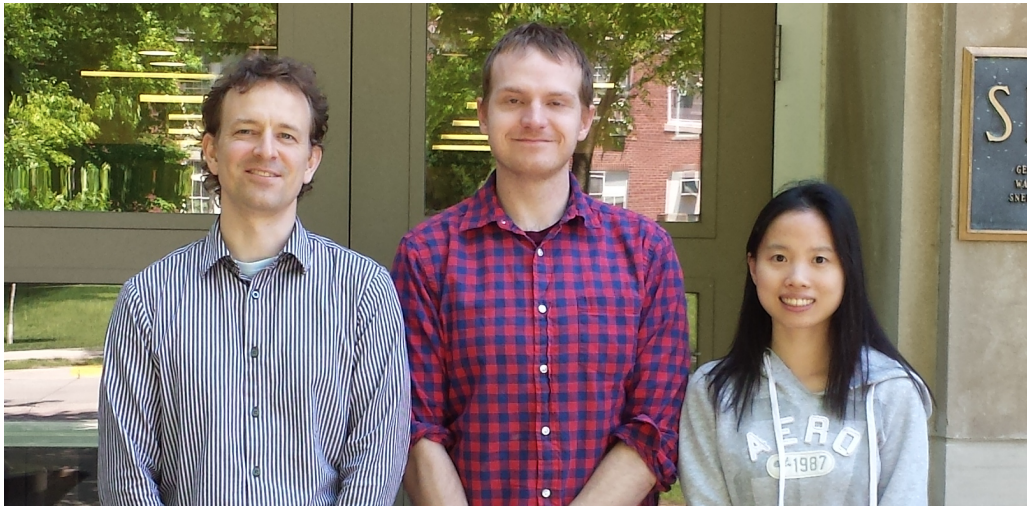
Money?

Power?

Glory?

I can tell you why I'm here

Why am I here?



In Spring 2013 I was taking STAT 602

- And 602 was really, really hard.
- So hard that I didn't even **look** at the midterm when I got it back
- But I knew that "data mining" was important and I liked it
- I just needed a chance to prove to myself that I could do it

That's when I joined the Data Mining Cup Team

And as if by magic STAT 602 started making much more sense. Without the big complicated data sets, it's hard to see where these theories and techniques matter. But working with these datasets takes up so much time, the methods become bizarre black boxes. Working at the DMC gave me:

Data Mining Experience

- A better understanding of why we need Machine Learning techniques
- A chance to apply what I was learning without any guardrails
- A chance to learn from others who had been through this before

Better Computing Skills

- The problems you have to deal with are often computational, and you get better at working through them
- I got much better at R during this competition (I learned how to use `knitr` for instance)

Collaborative Skills

- I got the chance to work hard at something because people were depending on me
- People working together on the same problem disagree - we had to resolve conflicts and move on
- Our solution is the hard work of several people, all woven together
- As the days went past, we got really close (friends for life kinda stuff)

I may not be much better at this stuff

What is it?

Why am I here?

We almost won

We came in fifth

People went to Germany

Wen Zhou

Cory Lanker

Fangfang Liu

Jia Liu

Ian Mouzon

Wei Zhang

What is it?

Why am I here?

We almost won

In 2014 we did win

What Makes the DMC Special?

Kaggle issues

Kaggle Gets Most Things Right

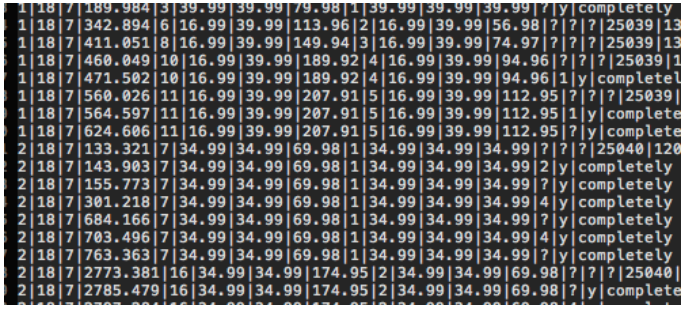
- Kaggle attracts lots of teams
- Large data sets
- Big prizes

But they also get a lot of things wrong

- In Kaggle you have get anywhere from 5 to 10 test predictions a day (the "leaderboard")
- Data in Kaggle competitions is often big without being complicated
- The data is often processed and feels generic (often it's just a black box)

Kaggle issues

DMC more realistic



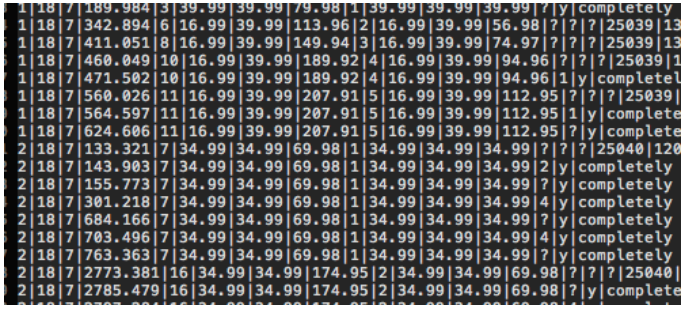
The DMC data is real

- DMC data is real data
- It Hasn't been overly processed, needs cleaning
- It isn't given to you in the format that you are "supposed" to use, it's given to you in the format it was collected
- You know what each record means, there are no "f145", "f513" columns

Kaggle issues

DMC more realistic

DMC data good



The DMC data is interesting

Both the 2013 and 2014 data sets were incredibly deep.

- There were time dependencies, repeat customers, and other complicated interactions
- Since we could approach the data from multiple angles, we could create a much larger and more unique set of features than anyone could have given us
- Big and complex data is much more fun than just "large memory allocation" data

```
#the 2013 data in the github repo dmc2014/dmc2013/
d.class = read.table('./data/transact_class.txt',header=TRUE,sep="|")
d.train = read.table('./data/transact_train.txt',header=TRUE,sep="|")
```

sessionNo	startHour	startWeekday	duration	cCount	cMinPrice	cMaxPrice	cSumPrice	bCount	bMinPrice	bMaxPrice	bSumPrice
1	6	5	0.00	1	59.99	59.99	59.99	1	59.99	59.99	59.99
1	6	5	11.94	1	59.99	59.99	59.99	1	59.99	59.99	59.99
1	6	5	39.89	1	59.99	59.99	59.99	1	59.99	59.99	59.99
2	6	5	0.00	0	?	?	?	0	?	?	?
2	6	5	15.63	0	?	?	?	0	?	?	?
2	6	5	26.23	0	?	?	?	0	?	?	?
2	6	5	71.20	0	?	?	?	0	?	?	?
2	6	5	94.47	0	?	?	?	0	?	?	?

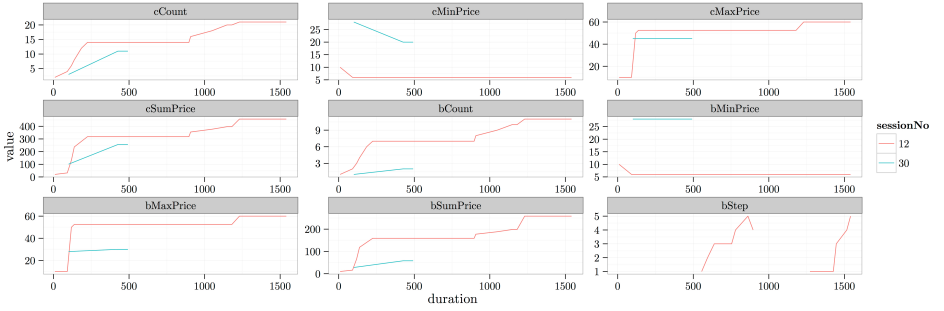
bStep	onlineStatus	availability	customerNo	maxVal	customerScore	accountLifetime	payments	age	address	lastOrder	order
?	?	?	1	600	70	21	1	43	1	49	y
2	y	completely orderable	1	600	70	21	1	43	1	49	y
?	y	completely orderable	1	600	70	21	1	43	1	49	y
2	y	completely orderable	?	?	?	?	?	?	?	?	y
?	y	completely orderable	?	?	?	?	?	?	?	?	y
4	y	completely orderable	?	?	?	?	?	?	?	?	y
4	y	completely orderable	?	?	?	?	?	?	?	?	y
?	y	completely orderable	?	?	?	?	?	?	?	?	y

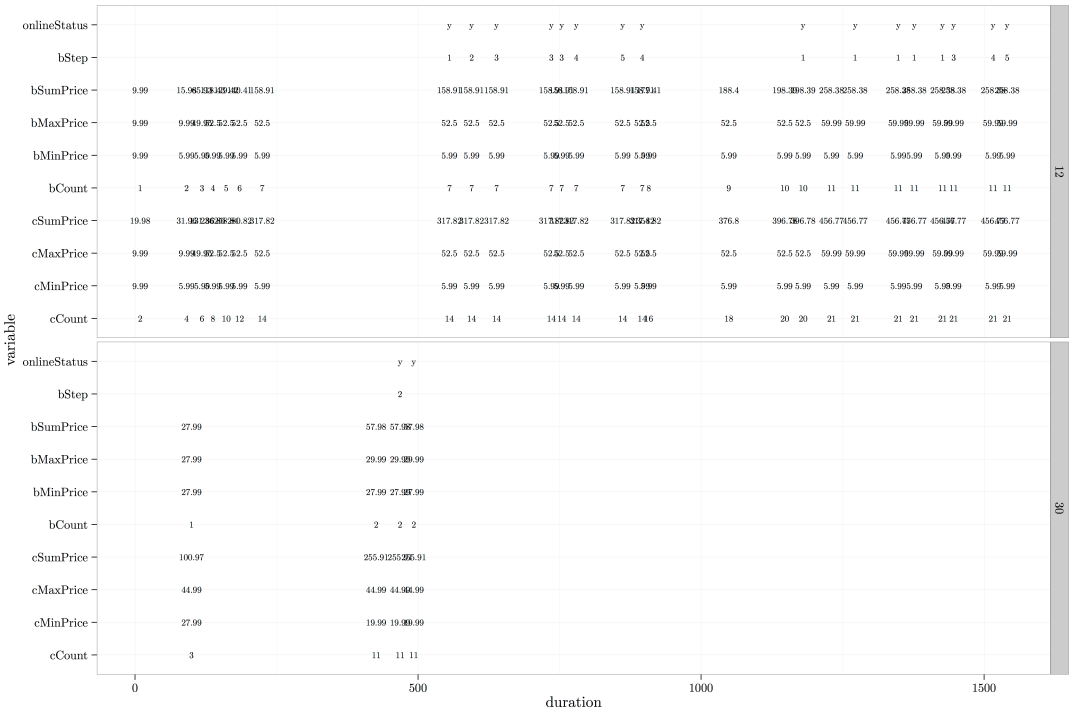
Kaggle issues

DMC more realistic

DMC data good

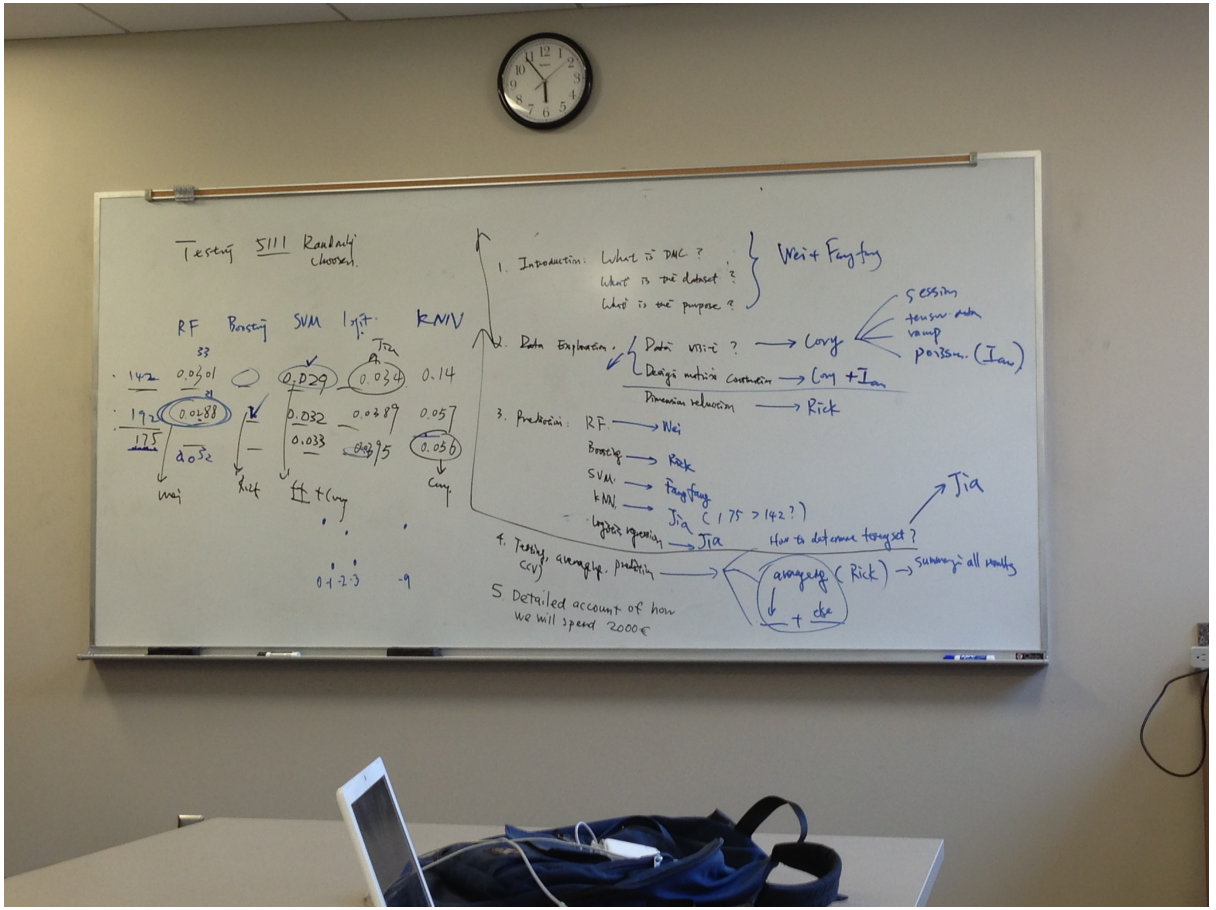
It matters how we handle time





What is Participating in the Data Mining Cup Like?

It's a lot of fun



- You get to work with smart people
- On something you are all interested in
- And you get to be creative

It's a lot of fun

It's a lot of work

The comptition only lasts five weeks

- It's pretty fast paced
- It takes a little bit of commitment
- It can take a backseat to other things going on in your life

It's a lot of fun

It's a lot of work

It's a third thing

I think the first 100 or so things we think about data are usually the same

- How many rows, columns, etc.
- Are these variables counts or classes?
- Is there a relationship between this variable and that variable?

and so on.

But after that, we start to think about different things

- You start to bring your own third things to the problem
- Things that only you can do
- Things that could make the difference between winning and losing