So You Want To Be a Data Miner?

Organizing, Struggling, and Sometimes Succeeding as a Team

Ian Mouzon
Department of Statistics
Iowa State University
The speech is on GitHub at imouzon/jsm2018

My Experiences

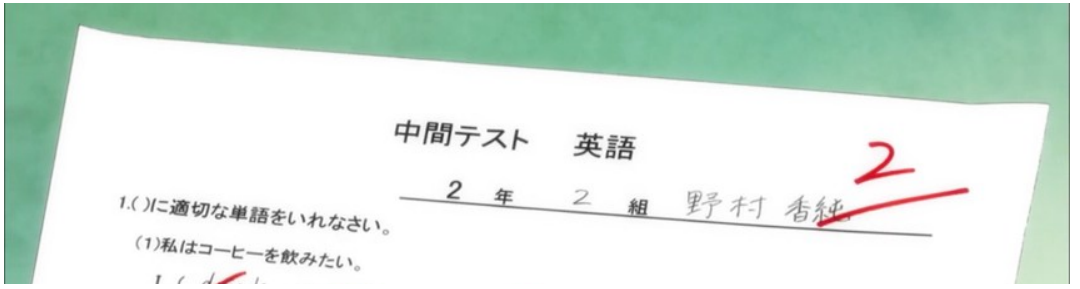
with

Big Data Competitions

Experiences

My Motivation

How I Got Started



I Was Taking Two "Big Data" Courses in the Same Semester

- My professors encouraged participation
- The team I joined that semester came in 5th Place in the Data Mining Cup

Why I Kept At It

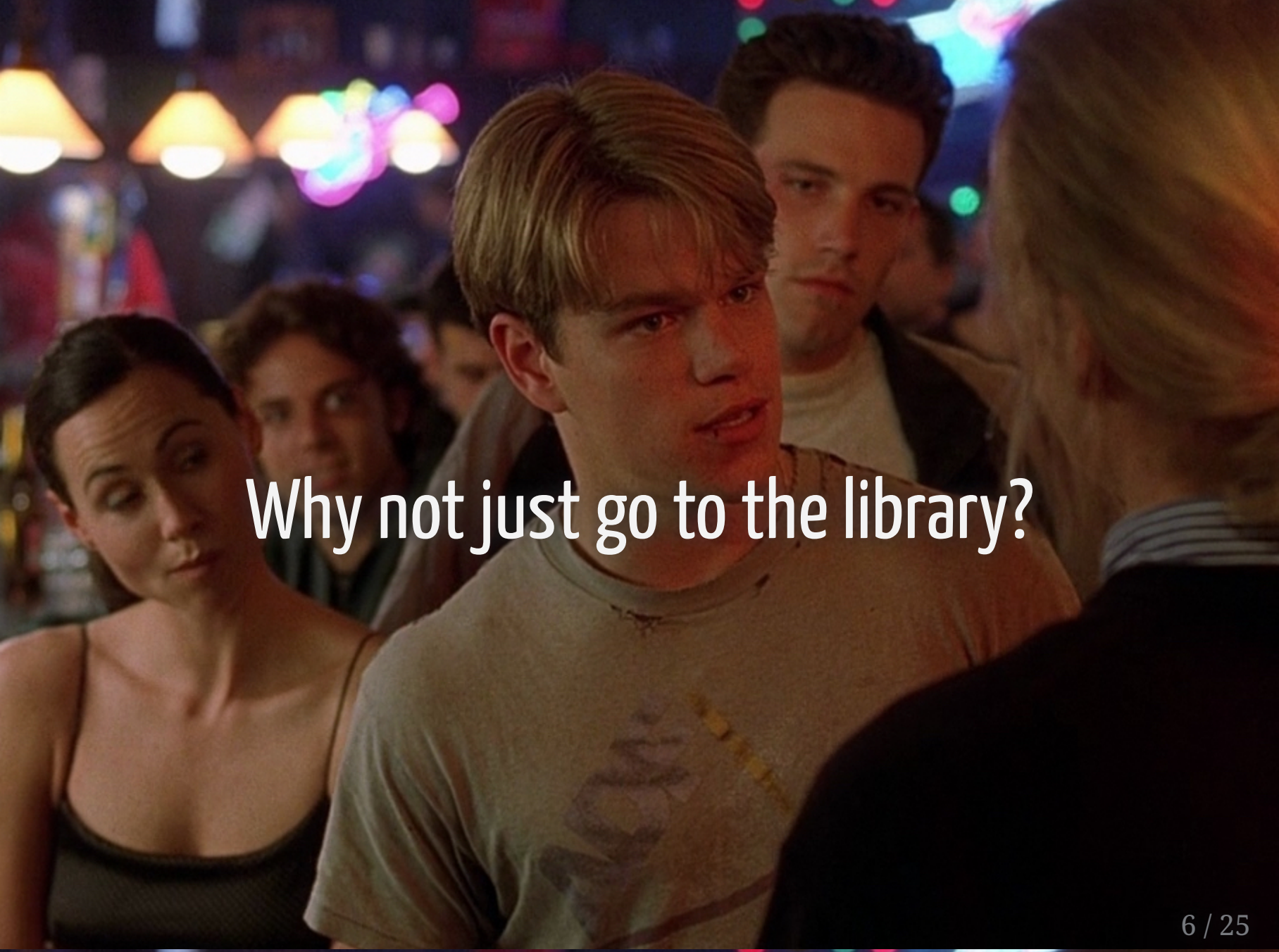
- **Positive Reinforcement**: Spiral up
- Eventually participated in about a dozen competitions over the next 3 years

What I Learned by Competing

Regardless of how the team ultimately did, I always got something out of it

- I got comfortable working with real data
  - Feature engineering: taking complicated relationships in the data and getting them into a model
  - Finding ways to compromise between the model you want and the data you have
- Gain experience on the ***statistical pipeline*** instead of the ***statistical smokestack***
  - building and maintaining databases
  - writing code that other people can read and use (i.e., version control)
- I learned a lot about working with a team
  - Work with people on all sides of my skill set

But Why Compete?


Why not just go to the library?

The Problems Are Hard to Solve

These competitions provide a great opportunity to develop new ideas.

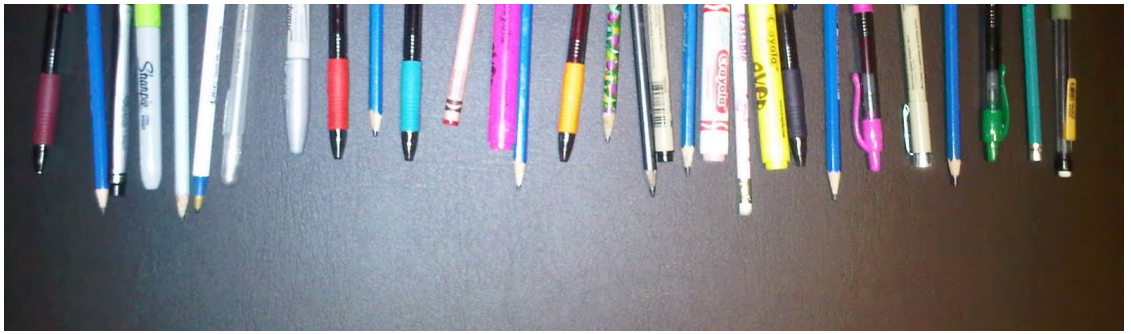
- In a lot of cases capable, intelligent, hard-working people weren't able to see a clear way through.
- If it was just a matter of tuning some stock model, it wouldn't be a competition
- Any good solution is going to have to be based, at least in part, on the creative ideas that result from frustration.

The Problems Are Hard Because of the Data

You are working with real data

- Has the data been processed and homogenized? Variables stripped of meaning? That's awful! It's also real.
- Is the data messy? Is it hard to understand what the variables even mean? That's real too.
- Is the data awkward to work with? Is it levels beyond "big", strangely structured? (shh, that's real).

The Field of Competitors is Diverse



The competitions attract people a wide range of academic and professional backgrounds

- CS people do *really* well in these competitions - there's something to learn from that.
- People with backgrounds in math fields (i.e., cryptology) or physics, too.
- You get to learn a lot about what these people do *outside* of a text book seeing their solutions.

Experiences

Opportunities

Open Problem

Open Field

Closed Period

Competitions End

Ultimately you win or lose and it's over. It's something that you get to let go of.

- It's not your research, it's not your thesis, it's not your job, you aren't stuck with it
- Your level of commitment to solving the problem can't extend beyond the deadline. Either you find the best solution or someone else does.
- The things you take away from the competition are yours to keep forever.

My Tips on What Works (and What Doesn't Work) When Organizing Teams

Take With A Grain of Salt

Meetings

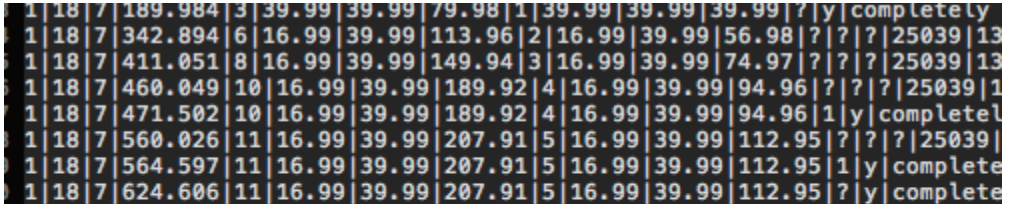
You *have* to have meetings, they need to be in-person

- People don't read emails, people don't look at each others updates
- Go for in-person meetings because that provides a good opportunity for people to meet and work together.

You *have* to have a lot of meetings

- Schedule them during a regular time
- If the competition is short, once a week isn't going to be enough - try twice a week.
- Twice a week to three times a week is good for big group meetings.
- Meet in pairs as the chance arises: getting with team members to look over problems, etc.

Everybody Should Work with the Raw Data

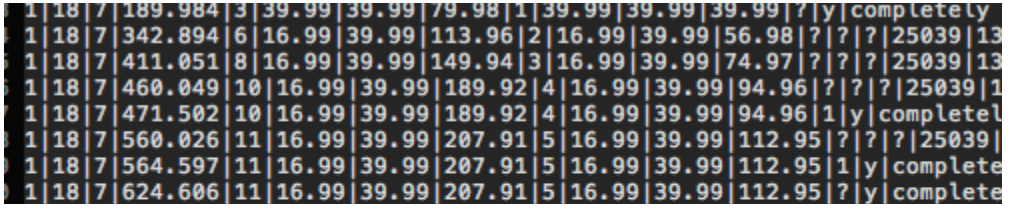


Essentially, the mistake comes when we identify disparge some work as "cleaning" then it's not statistics anymore. The justifications not to are something like this:

> Everyone doesn't need to clean the data, the end result will be exactly the same anyway

- First, where do you draw the line between "*cleaning*" and "*changing*"? I bet we might have different lines.
- Even obvious cases (i.e., an item color is given as `brwon`) provide information about the data as a whole.
- This is much more important in cases where the data structure is complex and awkward (in which case the desire to work with a simpler set is stronger)

Everybody Should Work with the Raw Data



Note: this does not mean that you can never have preprocessed/clean data sets to work with

- As you might imagine, you will mainly work with data that has been processed in some way.
- Ultimately, though, there should be a general awareness of how the preprocessed set differs from the original.
- Why? Your models are built on data manipulations that start with the raw data.
- Knowing how the raw data is structured gives your team the ability to know imagine exactly how they can extract a structure that they can model.

What Actually Happens

It may seem like a way to save time and effort at first, but it just doesn't work in the long run with these competitions.

Route A (Tyranny)

- The problem becomes alien to the people who didn't work with the raw data
- Feature being created quickly outpace their ability to stay active contributors
- Eventually, all they can do is run models on datasets
- They are reduced to being servers running code (fun!)

Route B (Revolution)

- The person "controlling" the data is overwhelmed by the responsibility
- Can't keep up with demands, becomes a bottleneck
- Resentment builds up until the team falls apart

Time Only Matters Sometimes

(No one at Texas A&M was panicking about the clock here)

There are more important things to do at the start of the competition than worry about time.

- The clock running out in a data mining competition is usually merciful instead of stressful.
- The only time it is stressful is when you have a good solution.

Time Only Matters Sometimes

Don't Worry About Repeating Others Work

- No team I have ever been a part of was undone by people doing the same thing
- The first 100 things everyone does are the same anyway (this is good)
- That's OK - you can't manage your way out of that.
- *Bad Idea*: Team Member A does plots for variable X, Team Member B does plots for variable Y
- *Worse Idea*: Team Member A does all the plots and spends a lot of time making them presentable, Team Member B looks at them for 5 minutes. B eventually leaves the team two weeks later because they "got busy with something else".

Time Only Matters Sometimes

Because It Isn't Really Repeated Work

- Analogy: if you go to Australia and show me pictures of your trip, you didn't save me the effort of the trip.
- Even if it's mechanically identical, at some point we'll split off. That's when we both are starting to make progress.

Prioritize Everyone Understanding the Core Problems

- It's not intuitive, but this means that during the early stages of the competition, you should not pay much attention to people doing what looks like the same thing.
- It may seem like a way to save time and effort at first, but it just doesn't work in the long run with these competitions.

Don't Delegate

Do Work Together, Do Volunteer

Delegating doesn't work

- My general impression is that people do and probably should resent this
- Also: who decides what to delegate? Who knows that much?

If you have an idea but aren't sure how to implement it:

- Bring it up at the team meetings (so vital)
- See if someone doesn't know how to do it (work together)
- If someone has an idea you could help on, do so

Timeline

How the Competition Progresses

Timeline

The Early Game

- Exploring the data
- Comprehending the fundamental challenges
- Getting a sense of the structure
- Simple predictions

The Middle Game

- Building a feature matrix
- First steps in prediction

The Late Game

- Better predictions
- Combining predictions

Conclusion

Conclusion

Data mining/Big Data competitions are uniquely rewarding

- They provide an opportunity to learn a set of experiences that are hard to find elsewhere
- They require no long term commitment but you might develop some skills or habits that persist (forever)
- Constraints and difficulty force creativity in multiple areas

Conclusion

The nature of the competitions reward specific habits

- Frequent meetings keep people in the loop as the competition progresses/changes
- Early focus should be on getting everyone working as close to the raw data as is feasible
- Keep an eye on the deadline, but don't let it lead you to block creative ideas
- Keep the team as flat as possible, outside of organizational tasks, for most of the competition.
- If you have a strong solution as the deadline looms, flip things over: everyone falls in line and does whatever the team leader says

Thank You Very Much

Any Questions or Feedback?

imouzon@iastate.edu