15-388/688 - Practical Data Science: The future of data science

J. Zico Kolter
Carnegie Mellon University
Fall 2016

Outline

Loose ends

The "future" of data science

Data science at CMU

Q&A

Final thoughts on the course

Outline

Loose ends

The "future" of data science

Data science at CMU

Q&A

Final thoughts on the course

Outline

Loose ends

The "future" of data science

Data science at CMU

Q&A

Final thoughts on the course

The “future” of data science

Technological trends are extremely difficult to predict

Example: I honestly don’t know what’s going to happen with the recent surge in Artificial Intelligence (and I work in AI)

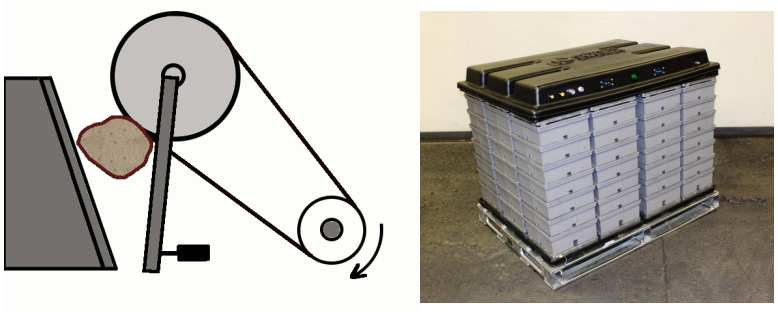
But I’m pretty confident in this prediction: data science (by one name or another) is here to stay

Data science for _____

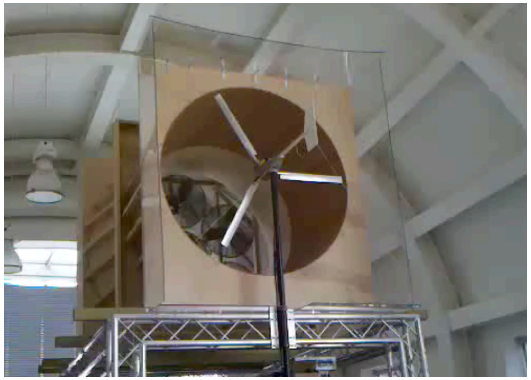
Hard to find a field that isn't at least trying to develop a "data-driven" component to it

Examples I've personally worked with at least tangentially: energy systems, building management, wind power, material science, chemical engineering, aerospace, robotics, fluid dynamics, industrial manufacturing, fraud detection, weather forecasting

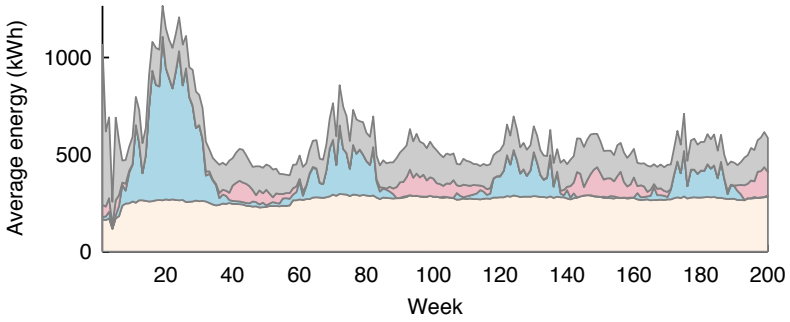
Some of my own research: data science for energy systems



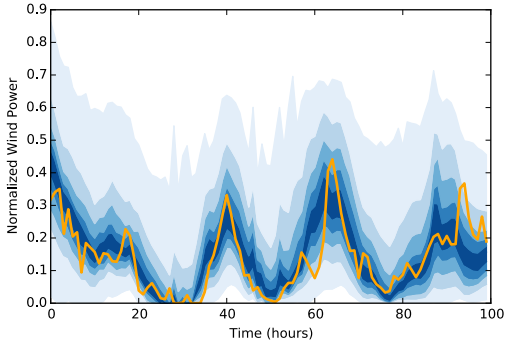
Demand side management for cement plants augmented with energy storage



Learning control for wind turbines



Energy disaggregation from low and high frequency data



Probabilistic forecasting for renewable energy systems

Big(ger) data

I've already mentioned that I'm typically skeptical of need for big data in many data science problems

But big data is here to stay, especially with advent of the "internet of things" and automated data collection

Evolution of the big data pipeline

Privacy in data science

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

Antonio Bolfo/Reportage for The New York Times

AOL apologizes for release of user search data

Search log information originally intended for use on new research site; company calls data posting a mistake.

Tech Industry

by **Dawn Kawamoto**
August 9, 2006 5:38 AM PDT

AOL apologized on Monday for releasing search log data on subscribers that had been intended for use with the company's newly launched research site.
The randomly selected data, which focused on 658,000 subscribers and posted 10 days ago, was among the tools intended for use on the recently launched AOL Research site. But the Internet giant has since removed the search logs from public view.

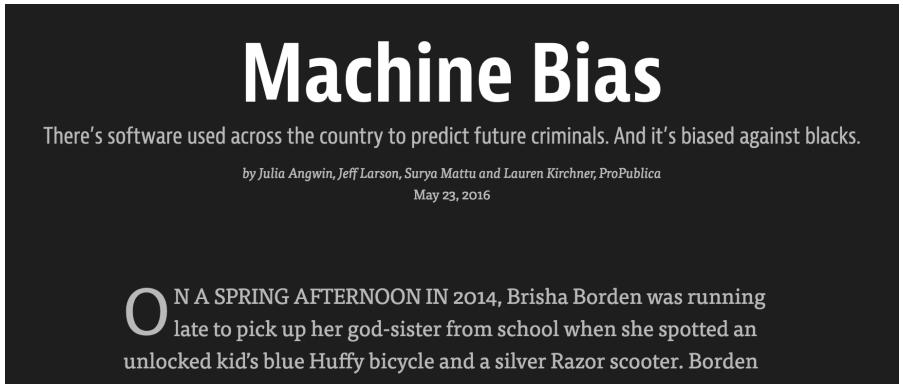
"This was a screw-up, and we're angry and upset about it. It was an innocent enough attempt to reach out to the academic community with new research tools, but it was obviously not appropriately vetted, and if it had been, it would have been stopped in an instant," AOL, a unit of Time Warner, said in a statement. "Although there was no personally identifiable data linked to these accounts, we're absolutely not defending this. It was a mistake, and we apologize. We've launched an internal investigation into what happened, and we are taking steps to ensure that this type of thing never happens again."

Although AOL had used identification numbers rather than names or user IDs when listing the search logs, that did not quell concerns of privacy advocates, who said that anyone among the 658,000 could easily be identified based on the searches each individual conducted.

As more data on users become available, the opportunities to connect this data and reveal private information is growing

An active research area in how we preserve user privacy while still attaining benefits of aggregate analysis

Bias in data science



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Machine learning and other inference algorithms make predictions based upon past training data

If the training data suffers from bias, there is a good chance the resulting algorithms will suffer from the same bias

Outline

Loose ends

The "future" of data science

Data science at CMU

Q&A

Final thoughts on the course

Additional courses to look into

CMU is an amazing place, and there are a huge number of courses available to those who want to pursue data science in more depth

To name a few (absolutely not exhaustive): 10-601/10-701 (Machine Learning), 36-402 (Advanced Data Analysis), 05-839 (Interactive Data Science), 10-605 (Machine Learning with Big Data Sets), 15-826 (Multimedia Databases and Data Mining), 15-780/15-781 (Artificial Intelligence), 11-641 (Machine Learning for Text Mining), 10-807 (Deep Learning)

What is academic data science?

"Data science" is not really an area of academic research…

Data science work comes up most often in the content of applied research in other fields, you can be a vastly stronger researcher in your area of interest if you are familiar with these techniques

The academic work in the area typically involves:

1. Fundamental reserach in machine learning or statistics (with data-science-like applications)

2. Methods in "automating" data science, e.g. "Automatic Statistician" (http://www.automaticstatistician.com)

Getting involved in data science research

Find an applied area you are interested in, find a faculty advisor in the area, start using the techniques you've learned in this class

Anecdotally, most researchers will be interested in how data science and machine learning techniques can be applied to their domains, but you will need to spend *substantial* time learning the domain itself

Follow on research course

I've had several people ask me about the possibility of independent study projects related to a topic in data science

I am going to launch an official course in Spring 2016 for data science research (15-388 or 688 will be a hard prerequisite)

Goal will be something like the class project, but on a larger scale, with the goal of producing a tangible data set and paper on the analysis

12 units, occasional meetings throughout semester, course number will be posted on Piazza once it is in the registrar (hopefully this week)

Outline

Loose ends

The "future" of data science

Data science at CMU

Q&A

Final thoughts on the course

Q&A

Questions from Piazza (and from class, if you have them)

What you've studied in this course

Data processing: web scraping and APIs, relational data and databases, data visualization, matrices and linear algebra, graphs and networks, free text, geospatial data (if you read the tutorial)

"Classical" learning methods: linear regression, linear classification, nonlinear methods using feature transformations, overfitting and cross validation, regularization, probability and statistics, maximum likelihood estimation, naïve Bayes, hypothesis testing

Other learning methods: decision trees and boosting, clustering and dimensionality reduction, mixtures of Gaussians, expectation maximization, recommender systems, deep learning, probabilistic models

Other: big data and MapReduce, debugging data science

Piazza statistics

1961 total posts
9244 total contributions
1465 instructors' responses
953 students' responses
33 min avg. response time

Top student answerers:
Abhik Mondal 34
Hao Huang 31
Ahmet Emre Unal 30
Sushain K. Cherivirala 27
Cheng Wang 27

The TAs on Piazza

Dhivya Eswaran
261 answers
470 contributions

Eric Wong
1112 answers
2115 contributions

“We need a session at the last lecture to thank the TA team … their support is one of the best I have seen so far! … Agree. The most painful TA Job ever! ... The students are so much while the TA team size is so small. They did really a good job! … Best TA performance I've ever seen … Absolutely awesome. Most efficient TA team I've seen so ... I can feel that the TAs in this course are making extraordinary effort in making the homeworks and answering the questions. … I love Eric … Eric Wong is a living legend … Amazing job by both TAs.”

Final Thoughts

Putting together a first-time course with 220 students (as of current roster), >30 remote students, 2 TAs, and virtually all new course material is perhaps a bit crazy…

Overall, this has been a wonderful experience, and I hope you've managed to learn a great deal throughout the course

Thanks for participating in the course, and thanks to the TAs for handling an insane course load

I hope to stay in touch with as many of you as possible, as you go forward in your careers