CSE 519: Data Science

Steven Skiena
Stony Brook University

Lecture 0: Course Administration

What is Data Science?

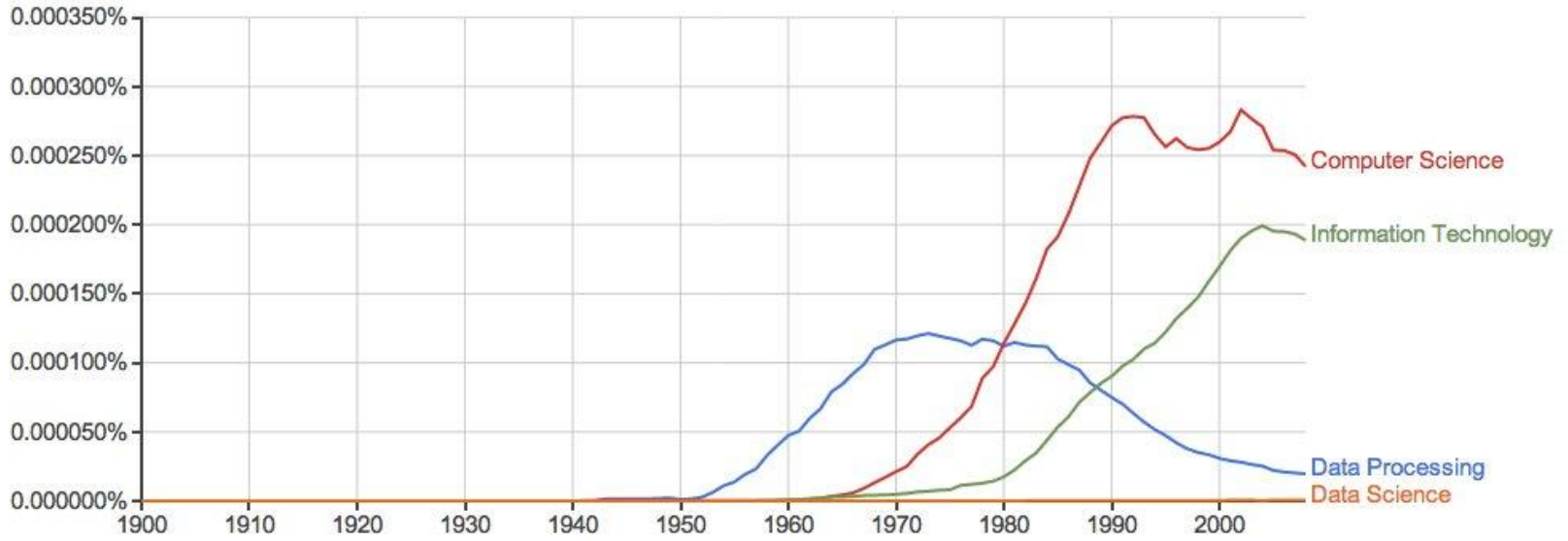
Like any emerging field, it isn’t yet well defined, but incorporates elements of:

- Exploratory Data Analysis and Visualization
- Machine Learning and Statistics
- High-Performance Computing technologies for dealing with scale.

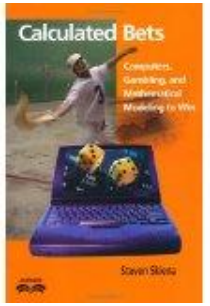
Why Data Science?

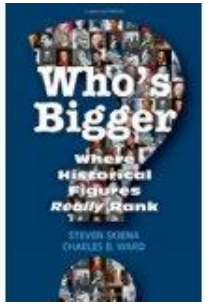
- New technology makes it possible to capture vast amounts of logging / sensor data.
- Computing advances make it possible to analyze data on ever increasing scales.
- Prominent role models (Google, Moneyball, hedge funds, Nate Silver, ...) have proven the power of modern data analytics.

Data is not new to computing...



My Experience with Data





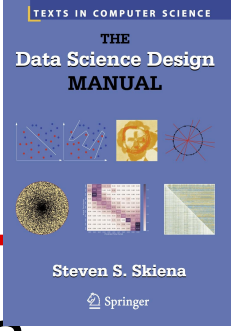
- Gambling systems in jai-alai and more
- Collaborations with biologists and social scientists
- Large-scale text analytics and NLP
- Startup companies
- Ranking historical figures

GENERAL SENTIMENT

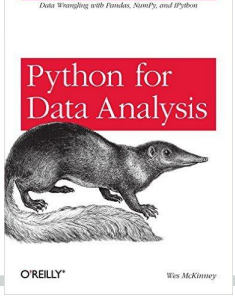
This drives what I will teach here.

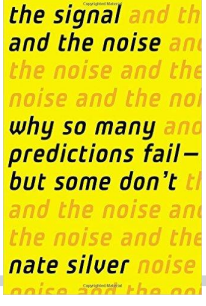
thrivemetrics

The Data Science Design Manual



- The course textbook is my book, *The Data Science Design Manual*, Springer, 2017.
- Stuff from the book is fair for quizzes/exams.
- Recommended texts include Nate Silver's "The Signal and the Noise" and "Python for Data Analysis".





Online Teaching

This is the first semester I have taught online, and beg your patience.

The way I teach relies heavily on class discussion, so I strongly encourage everyone to raise questions and discussion topics.

Recordings of the class will be available in Blackboard and (eventually) YouTube.

(Very) Distant Learning

- Some of you may be in distant time zones.
- Find project and HW partners in the same time zone.
- The final exam is the only thing you must do live this semester, but I need you to turn in assignments and quizzes by the same deadline as everyone else.

Google Classroom / Piazza

The homework assignments, and projects will be submitted by Google Classroom, so sign up as in the syllabus.

Daily quizzes are also in Google Classroom so come to class signed on.

Discussions and messages are on Piazza.

Sign up for these before the next class!

Semester Schedule (I)

8/25	L0: Course Introduction/Administration			
8/27	L1: Introduction to Data Science		1-26	(HW1 out)
9/1	L2: Mathematical Preliminaries		27-38	
9/3	L3: Python for Data Science I	PFDA		(HW1 in / HW2 out)
9/8	L5: Correlation		39-56	
9/10	L6: Assembling Data Sets	57-68		
9/15	L4: Python for Data Science II	PFDA		
9/17	L7: Data Cleaning	69-94		
9/22	L8: Scores and Rankings I	95-103		(Project out)
9/24	L8: Scores and Rankings II	104-120		(HW2 in / HW3 out)
9/29	L9: Statistical Distributions	121-134		
10/1	L10: Statistical Significance	135-154		
10/6	L11: Principles of Visualizing Data	155-169		
10/8	L12: Practice of Data Visualization	170-300		
10/13	L13: Building Models		201-212	
10/15	L14: Validating Models		213-236	(HW3 in)

Semester Schedule (II)

10/20	L15: Linear Algebra Review	237-266		(Project proposal in)
10/22	L16: Linear Regression		267-278	
10/27	L17: Gradient Descent Search/Regularization	279-288		
10/29	L18: Logistic Regression and Classification	289-302		
11/3	L19: Nearest Neighbor Methods I	303-319		
11/5	L19: Nearest Neighbor Methods II	320-329		
11/10	L20: Clustering		330-350	(Progress reports in)
11/12	L21: Introduction to Machine Learning I	351-362		
11/17	L21: Introduction to Machine Learning II	363-376		
11/19	L22: Topics in Machine Learning		377-390	

11/23-7 Thanksgiving (class cancelled)

12/1+	L23: Achieving Scale	391-418	
12/3+	L24: Human-centric Data Science	419-426	(Final reports in)

(+) denotes possible project presentations instead of lectures, if so, this material will be taught earlier

12/17 Final exam (8AM-10:45PM)

Note that the final exam has been scheduled by the university for the last possible day. Plan your winter travel accordingly, because I will not be able to give the exam earlier to any student.

Changes from the Usual

- I will probably use the last week for 4 minute video project presentations
- Thus I will probably compress / reorder some of the lecture material, but the final will cover the entire textbook.

This is Introduction to Data Science

- Some students have a substantial data science background already.
- If you know the material, take a different course…
- Read my textbook and last year's notes to gauge the level of the course.
- The "graduate student" part is the project.

Instructor Style Disclaimer

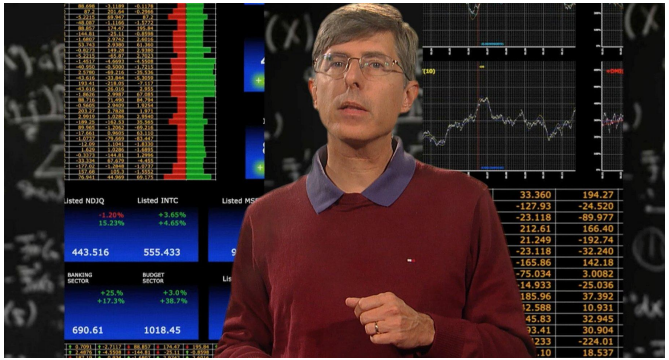
- I try to make lectures fun through jokes and analogies, but always fear saying something that may offend someone in the class.
- I am particularly fearful of teaching online, as I will miss feedback mechanisms I am used to in the classroom.
- I want everyone to feel comfortable in my classroom.
- If anything I say bothers you, please come by and tell me so. I will apologize, and then do my best to understand the issue to avoid doing so again.

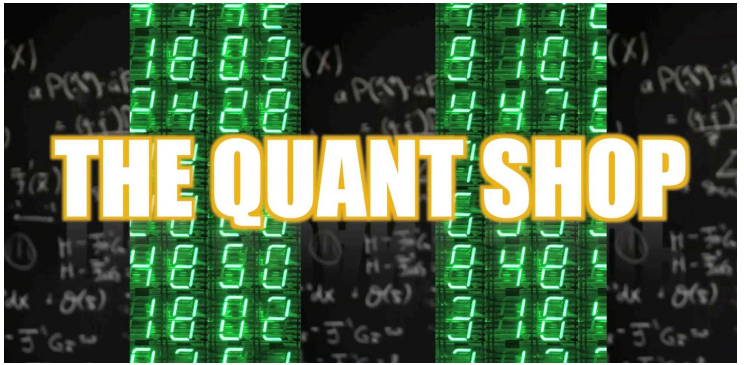
Course Project

- This will be a small group project, where each team takes on a particular forecasting challenge and builds a predictive model.
- Each team will start from scratch, including finding/constructing the relevant data sets.
- The smaller class size should let us offer a broader and more flexible set of projects

Reality TV (www.quant-shop.com)

In Fall 2014, each group's course project was professionally edited for public viewing.





Quant Shop: Episodes

1. Finding Miss Universe
2. Modeling the Movies
3. Winning the Baby Pool
4. The Art of the Auction
5. White Christmas
6. Predicting the Playoffs
7. The Ghoul Pool
8. Playing the Market

The projects will be used as ongoing examples, so start watching at www.quant-shop.com.

Each program runs 30 minutes.

Piazza Behavior

The 519 students in 2018 behaved badly online

Minimize anonymous postings: I have changed settings so no one is anonymous to me.

Any posting/peer grading not maintaining a positive, professional tone earns a strike costing a substantial % of your semester grade.

2019 was much better: no strikes given.

note

202 views

A suggestion

Dear Professor and TAs,

Considering how disrespectful people have been over the last couple of days, Please consider removing this 'Anonymous to everyone' feature from Piazza (like some other course instructors have done). It should always be visible to the instructors who is saying what. If done so, People will behave in their limits in future classes. Peace

exam other logistics

An 'Anonymous' student suggesting to remove the anonymous to everyone feature.

P.S - even I second that the students have been disrespectful to the point beyond tolerance limit

Anonymous 8 months ago I knew such kind of stupid comments would come. I said to remove "anonymous to everyone", not "anonymous to classmates" Idiot.

Anonymous 8 months ago Hahahaha.. 2nd Irony
Now aren't you being disrespectful?

How can you justify your comment over other's? I'm curious

Anonymous 8 months ago I don't need to, not to people like you. Clearly, you are proving me right. You are an IDIOT! Not going to entertain more of this by the way. Peace

This Year We Will Do Better

note

203 views

Dear Peers!

Never in all the last 8 course piazzas, in past 1 year, saw such non-viable, malevolent and hateful comments.

Speaking from **experience**, to the 0.5-1 marks that people are losing their minds for- look for the long term goals of taking any course! You are either going to work in industry or higher education. You need not put ur gpas on resume. You will be tested on the practical applications of your knowledge, your eagerness to learn and charisma in personality. Use piazza to broaden your knowledge rather than going hostile for marks. Marks are a way to gauge that learning. Even Professor couldn't tolerate it further.

This class has gone to far greater heights of disrespect or perhaps the criteria bar of student intake in college has significantly lowered.

other

edit · good note | 19

Updated 8 months ago by Anonymous

Academic Grading

- 45% of grade is from your group project, split between proposal, progress, and final reports, presentation and peer grading.
- There will be a final exam worth 25% of the grade, and daily quizzes worth 8%.
- The remaining 22% comes from three HW assignments before the project.

Warning: 519 Grades are Subjective

Because the HW/projects favor open-ended work, evaluation cannot be made deterministic. We will do our best to be fair and consistent.

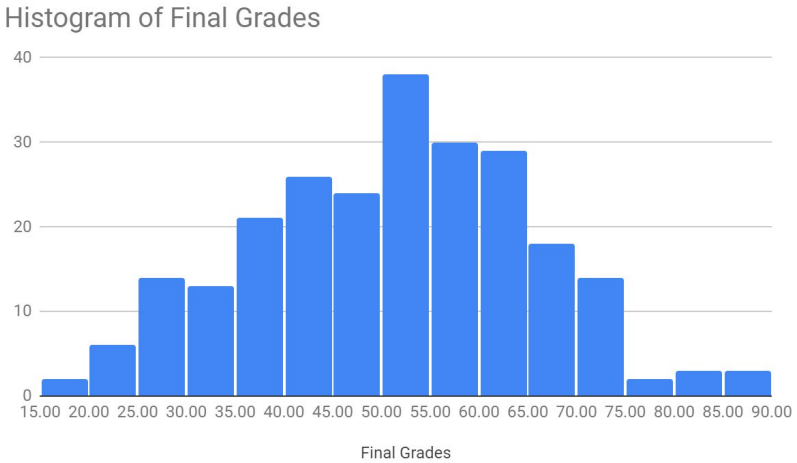
If this bothers you, take a different class.

We use comparative grading and have multiple scores and assignments to reduce variance.

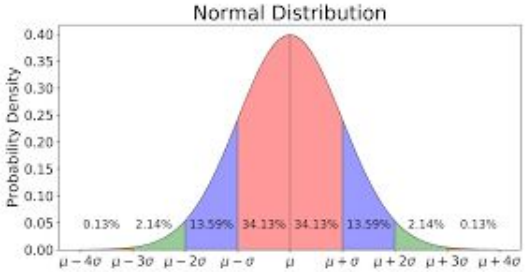
CSE 519 Grades Are Consistent

Observe positive correlations between each grade component with every other grade component.

	HW	Quiz	Peer Review	Project	Finals
HW	1	0.47300	0.32639	0.35557	0.34665
Quiz	0.47300	1	0.41466	0.26892	0.29629
Peer Review	0.32639	0.41466	1	0.31078	0.29675
Project	0.35557	0.26892	0.31078	1	0.45885
Finals	0.34665	0.29629	0.29675	0.45885	1



Regrades are Biased



Grading errors are presumably unbiased, and hence correct themselves over time.

But regrade complaints are a function of how aggressive the student is, and favor some students over others.

This is why we will not be considering regrades or changes to HW/project/quiz/exam grades.

The No Complaint Grading Policy

Each time a student asks/questions/comments about assignment fairness or a grade (for any reason other than an addition mistake) to me or on Piazza they will receive a strike.

The No Extension Grading Policy

Each time a student asks for an extension on a an assignment they will receive a strike.

The only exceptions are hospitalizations or family emergencies of two weeks or more.

I Don’t Mean be Too Grumpy

I am looking forward to teaching you this semester, in this very special time.