
CSE 519: Data Science
Steven Skiena
Stony Brook University

Lecture 0: Course Administration

What is Data Science?

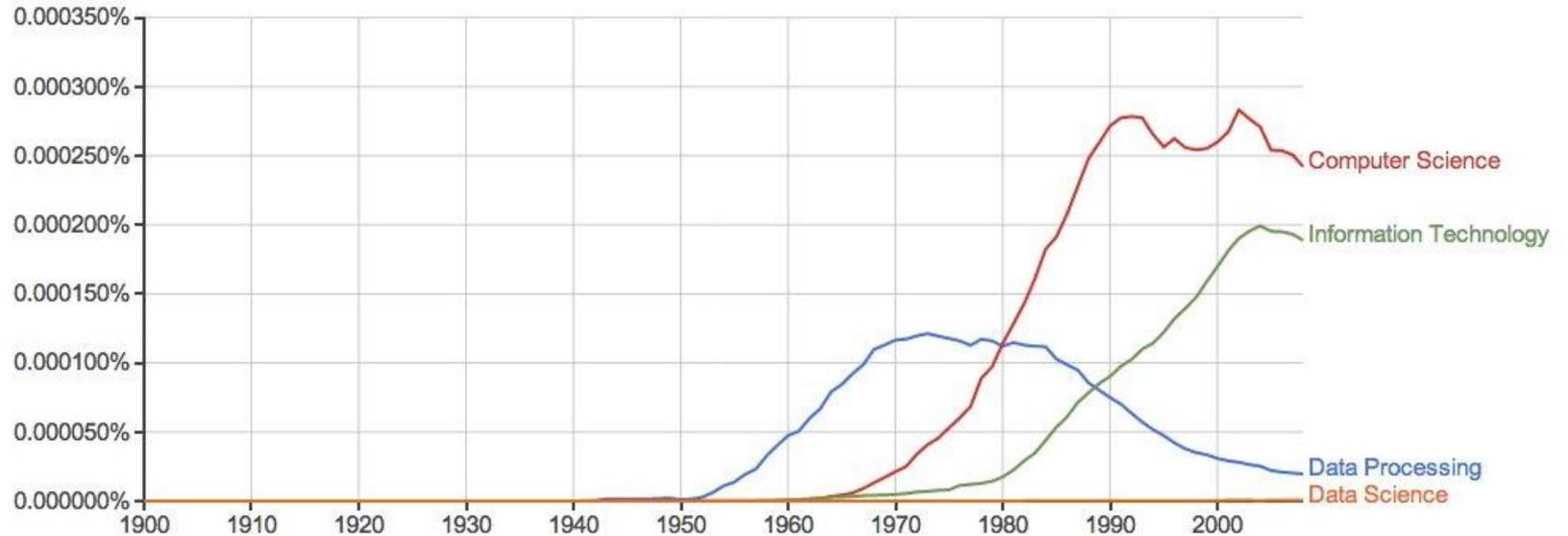
Like any emerging field, it isn't yet well defined, but incorporates elements of:

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

Why Data Science?

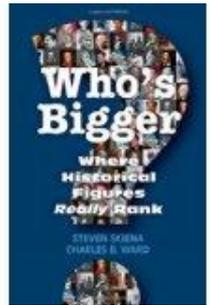
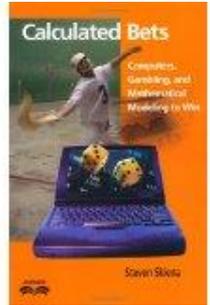
- New technology makes it possible to capture vast amounts of logging / sensor data.
 - Computing advances make it possible to analyze data on ever increasing scales.
 - Prominent role models (Google, Moneyball, hedge funds, Nate Silver, ...) have proven the power of modern data analytics.
-

Data is not new to computing...



My Experience with Data

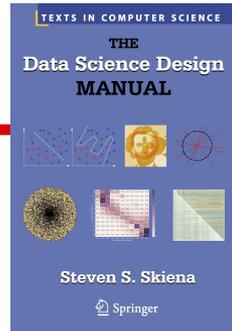
- Gambling systems in jai-alai and more
- Collaborations with biologists and social scientists
- Large-scale text analytics and NLP
- Startup companies
- Ranking historical figures



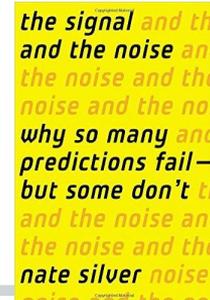
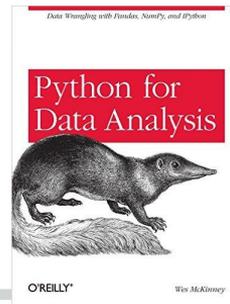
This drives what I will teach here.



The Data Science Design Manual



- The course textbook is my new book, published by Springer-Verlag, 2017.
- I will appreciate feedback / error correction.
- Recommended texts include Nate Silver's "The Signal and the Noise" and "Python for Data Analysis".



Semester Schedule (I)

8/28	L0: Course Introduction/Administration		
8/30	L1: Introduction to Data Science	1-26	(HW1 out)
9/4	Labor Day (classes cancelled)		
9/6	L2: Mathematical Preliminaries	27-38	
9/11	L5: Correlation	39-56	(HW1 in / HW2 out)
9/13	L6: Assembling Data Sets	57-68	
9/18	L7: Data Cleaning	69-94	
9/20*	L3/4: Python for Data Science	PFDA	
9/25	L8: Scores and Rankings I	95-103	(HW2 in / HW3 out)
9/27	L8: Scores and Rankings II	104-120	(Project out)
10/2*	L9: Statistical Distributions	121-134	
10/4*	L10: Statistical Significance	135-154	
10/9	L11: Principles of Visualizing Data	155-169	
10/11	L12: Practice of Data Visualization	170-300	(HW3 in)
10/16	L13: Building Models	201-212	
10/18	L14: Validating Models	213-236	

Semester Schedule (II)

10/23	SLACK CLASS		(Project proposal in)
10/25	L15: Linear Algebra Review	237-266	
10/30	L16: Linear Regression	267-278	
11/1	L17: Gradient Descent Search/Regularization	279-288	
11/6	L18: Logistic Regression and Classification	289-302	
11/8	L19: Nearest Neighbor Methods I	303-319	
11/13	L19: Nearest Neighbor Methods II	320-329	
11/15	L20: Clustering	330-350	(Progress reports in)
11/20	L21: Introduction to Machine Learning I	351-362	
11/22	Thanksgiving (class cancelled)		
11/27	L21: Introduction to Machine Learning II	363-376	
11/29	L22: Topics in Machine Learning	377-390	
12/4	L23: Achieving Scale	391-418	
12/6	L24: Human-centric Data Science	419-426	(Final reports in)
12/12	Final exam (8:30-11PM)		

Course Project

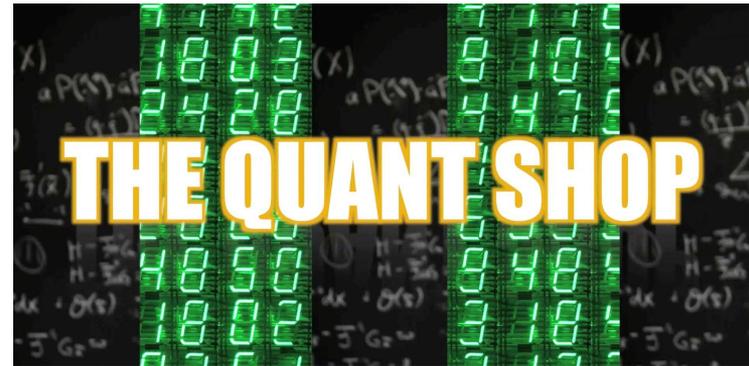
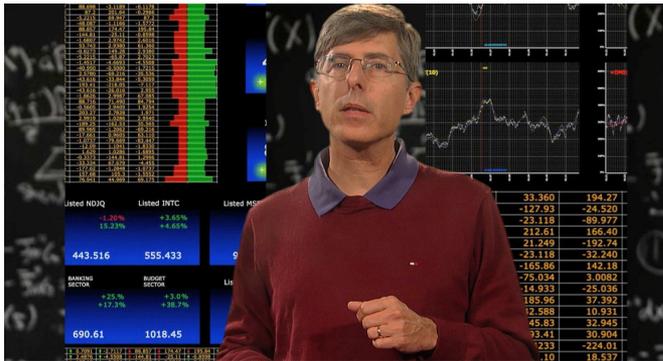
- The work in this course will primarily be a group project, where each team takes on a particular forecasting challenge and builds a predictive model.
 - Each team will start from scratch, including finding/constructing the relevant data sets.
-

Grading

- Total of 50% of your grade based on your group project, split between proposal, progress, and final reports.
 - There will be a final exam worth 30% of the grade, based on exercises from the text.
 - The remaining 20% comes from three HW assignments before the project.
-

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.

Registration Survey

- Fill out the course registration questionnaire, to help me finalize registration.
 - You will be given a very brief screen test when you walk out the door, as you turn in your survey.
 - I will fix registration decisions before the next class.
-