Data (corex-ad 3q)

Spring 2017

I will be teaching Data for the first time next semester.

Instructor: Azza Abouzied azza@nyu.edu Lectures: TuTh 10:25 – 11:40 @ A5 002.

Course Description: Data is everywhere. We have massive datasets keeping track of humanity's everyday minutiae from babies born to calories consumed, friends made to crimes committed. How can we use these data to make useful predictions and gain insights into ourselves and humanity in general? This course introduces the basics of *data-driven thinking* and covers topics such as wrangling, exploration, analysis, prediction and storytelling through data visualization.

Course Page: bit.ly/corexdata

Course textbook:

- John W. Foreman. **Data Smart: Using Data Science to Transform Information into Insight**. 1st. Wiley, Nov. 2013. isbn: 111866146X.

Reference textbooks:

- Charles Wheelan. **Naked Statistics: Stripping the Dread from the Data**. 1st. W. W. Norton & Company, 2014. isbn: 039334777X.
- Nate Silver. **The Signal and the Noise: Why So Many Predictions Fail–but Some Don't**. 1st. Penguin Press, 2012. isbn: 159420411X.

Getting help: Email any questions to the class piazza forum for the most immediate help from the instructor and your peers. I'm available for an hour after every class or at other times by appointment.

Learning Outcomes:

- Solve real-life problems through *data-driven thinking*
- Transform data into valuable insights, decisions and products with the help of mathematics and statistics
- Communicate effectively using the language of data-driven analysis: explain how data was collected, sampled, cleaned, transformed, analyzed and visualized
- Have a personal toolbox of data analysis techniques that can be applied to different problems including optimization, clustering, regression, and artificial intelligence
- Accurately critique analysis outcomes, predictions, journalistic and scientific presentations of data
- Understand the limitations and biases of data-driven techniques

To be an expert at data-driven thinking, you need more than a single course and you will need to drastically expand the small toolbox we will build at the end of this class. You will also need to gain lots of experience solving many problems. The most important learning outcome is that you become sufficiently excited about data-driven thinking that you apply it when suitable to your own personal or professional problems even after the course concludes.

Teaching Methodologies:

- *Lectures*: Class lectures will introduce a variety of data analysis techniques through **case studies**. Through in-class exercises, we will demonstrate and apply these techniques to real data. Lectures will provide a gentle introduction to the mathematics and statistics behind many techniques.
- *Readings*: The course textbook Foreman, *Data Smart: Using Data Science to Transform Information into Insight* is a reference book on different techniques. The course schedule lists sections (exercises) of the course textbook that you should read (attempt to complete) for each week. Required readings will include chapters from reference textbooks, research papers, news articles and blog posts. Some assigned readings are audibles. By keeping on top of the readings, you will make the best use of lecture time: you can clarify concepts you found difficult to understand and you can better participate in class discussions and exercises. Optional readings might include more advanced or related concepts, which you should refer to if you find a particular topic interesting or if it covers techniques you may need for your project.
- *Problem Sets & Project*: You will complete four problem sets in groups. There is little to *no programming* in this class. We use *spreadsheets*! We also use free, easy to use tools such as Tableau and Trifacta. Why? Because computational and data-driven thinking is not about the tools, machines or programming languages. It is about breaking down problems into simple logical processes. Anyone can engage in and benefit from data-driven thinking from a personal trainer to Facebook's CTO. Most problem sets require you to apply techniques already demonstrated in class to a different problem and data set. Through technical reports and visualizations you will improve your technical writing and learn how to effectively communicate analysis outcomes as well as the uncertainty associated with such outcomes.

Software Requirements: Install a recent Excel copy and free community versions of Tableau (for data visualization) and Trifacta (for preparing and cleaning data). Through the class, you will learn how to use these tools.

Class Deliverables:

Problem Set 1 (10%)	Visualize the effectiveness of antibiotics.
Problem Set 2 (20%)	Patchi Treats: Optimizing Chocolate Boxes for Personal Preferences and Price.
Problem Set 3 (20%)	How busy is the subway? Wrangling and visualizing NYC Subway foot traffic patterns.
Problem Set 4 (10%)	Predicting Survival: Who survives the titanic?
Project Proposal, Report & Presentation (30%)	Pick a real-life problem and apply data-driven techniques to solve the problem.
Class Participation (10%)	Attend all lectures, read assigned material before class and participate in class discussions.

Grading Policy: In general, a 90% or above is within the A range, 80%-90% is within the B range and 70%-80% is within the C range. You have 100 hours of lateness forgiveness that you can use throughout the course for any problem set or project submission deadline.

Course Schedule:

Week	Lectures, Readings, Case Studies, Assignments
	The importance of seeing your data
1	**Overview & Why a deeper analysis of data matters?** *Case Study:* Looking at UC Berkeley's 1973 Graduate Admissions Data. Was UC Berkeley sexist? *Reading:* Wikipedia, *Simpson's Paradox* *Reading:* Tufte, *Visual and Statistical Thinking: Displays of Evidence for Making Decisions* *Optional Reading:* Pearl, "Comment: Understanding Simpson's Paradox"
2	**Visualization** *Reading:* Cleveland and McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods" *Reading:* Tufte, *The Visual Display of Quantitative Information Paperback*, Data Ink & Chart Junk *Optional Reading:* Norman, *Things That Make Us Smart: Defending Human Attributes In The Age Of The Machine*, Chp 3. The Power of Representation *Optional Reading:* Zhang and Norman, *The Representation of Numbers* *Optional Reading:* Segel and Heer, "Narrative Visualization: Telling Stories with Data" *Assignment:* Visualization Design: Visualize the Effectiveness of Antibiotics (1 week) *Project:* Team Formation for Class Projects
3	**Probability, Descriptive Statistics & Hypothesis testing** *Reading:* Wheelan, *Naked Statistics: Stripping the Dread from the Data*, Chp 2, 3, 8 *Reading:* Foreman, *Data Smart: Using Data Science to Transform Information into Insight*, Chp 1 *Optional Reading:* RadioLab, *A Very Lucky Wind* *Optional Reading:* Life, *No Coincidence, No Story!*
4	**Outliers & Liars** *Case Study:* Hadlum v. Hadlum: Who is the father? *Reading:* Foreman, *Data Smart: Using Data Science to Transform Information into Insight*, Chp 9 *Reading:* Varian, "Letters to the Editor", Benford's Law
5	**Data Wrangling** *Case Study:* Where not to eat? Wrangling SFO's restaurant inspection data. *Assignment:* Wrangling & Visualizing: NYC's Subway Foot Traffic Patterns (1 week) *Project:* Proposal (1-2 page summary)
	Going deeper into the data
6	**Optimization: The grunt work of data analysis** *Reading:* Foreman, *Data Smart: Using Data Science to Transform Information into Insight*, Chp 4 *Assignment:* Data Collection and Problem Solving: Patchi Treats — The best chocolate box that you can afford! (2 weeks)
7	**A or B** *Case Study:* A data-driven presidential campaign *Reading:* Rush, *Optimization at the Obama campaign: a/b testing*

8 **Clusters & Profiles: Similarity Measures & Nearest Neighbors**

Case Study: Spam or Not?

Reading: Foreman, *Data Smart: Using Data Science to Transform Information into Insight*, Chp 2, 5

Project: Update 1 (2-3 pages project report)

9 **Predicition**

Case Study: Are you Pregnant?

Reading: Silver, *The Signal and the Noise: Why So Many Predictions Fail–but Some Don't*, Chp 1. A Catastrophic Failure of Prediction

Reading: Backstrom and Kleinberg, "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook"

Reading: Duhigg, "How Companies Learn Your Secrets"

Reading: Duhigg, "What Does Your Credit-Card Company Know About You?"

10 **Prediction: Regression and Other Techniques**

Reading: Foreman, *Data Smart: Using Data Science to Transform Information into Insight*, Chp 6

Assignment: Prediction: The Titanic — Who will survive? (1 week)

11 **Bayesian Techniques**

Reading: Foreman, *Data Smart: Using Data Science to Transform Information into Insight*, Chp 3

Reading: Silver, *The Signal and the Noise: Why So Many Predictions Fail–but Some Don't*, Chp 8. Less and Less and Less Wrong

Know your limits: don't go too deep

12 **Unknowns and Hard Problems**

Reading: Silver, *The Signal and the Noise: Why So Many Predictions Fail–but Some Don't*, Chp 13. What you don't know can hurt you

Project: Update 2 (2-3 pages project report)

13 **Biases and the ethics of data-driven decision-making**

Case Study: Predicting Criminality

Reading: Angwin et al., *Machine Bias: Investigating the algorithms that control our lives*

Reading: Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*

Optional Reading: Kay, Matuszek, and Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations"

Optional Reading: Amit Datta, Tschantz, and Anupam Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination"

Optional Reading: Crawford, *Artificial Intelligence's White Guy Problem*

Optional Reading: Dwork et al., "Fairness Through Awareness"

Where to next?

14 **What more can you do? (How little do we know?)**

Project: Presentations & Final Report (4-6 pages)

Additional Course Readings:

- Julia Angwin et al. *Machine Bias: Investigating the algorithms that control our lives*. May 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- Jeff Larson et al. *How We Analyzed the COMPAS Recidivism Algorithm*. May 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
- Wikipedia. *Simpson's Paradox*. 2016. URL: http://bit.ly/1OHFSOk.

- Lars Backstrom and Jon Kleinberg. **Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook**. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing*. CSCW '14. Baltimore, Maryland, USA: ACM, 2014, pp. 831–841. ISBN: 978-1-4503-2540-0. URL: http://doi.acm.org/10.1145/2531602.2531642.
- Charles Duhigg. **How Companies Learn Your Secrets**. In: *The New York Times Magazine* (Feb. 2012). URL: http://nyti.ms/18LN5uz.
- Kyle Rush. *Optimization at the Obama campaign: a/b testing*. [Online, accessed 08-Nov-2016]. Dec. 2012. URL: http://kylerush.net/blog/optimization-at-the-obama-campaign-ab-testing/.
- Charles Duhigg. **What Does Your Credit-Card Company Know About You?** In: *The New York Times Magazine* (May 2009). URL: http://nyti.ms/1Lwmum3.
- Edward R. Tufte. **The Visual Display of Quantitative Information Paperback**. 2nd. Graphics Press, May 2001. ISBN: 1930824130.
- Edward R. Tufte. **Visual and Statistical Thinking: Displays of Evidence for Making Decisions**. Graphics Press, Apr. 1997. ISBN: 0961392134. URL: http://www.cc.gatech.edu/~stasko/7450/Papers/shuttle.pdf.
- William S. Cleveland and Robert McGill. **Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods**. In: *Journal of the American Statistical Association* 79.387 (1984), pp. 531–554. ISSN: 01621459. URL: http://www.jstor.org/stable/2288400.
- Hal R. Varian. **Letters to the Editor**. In: *The American Statistician* 26.3 (1972), pp. 62–66. ISSN: 00031305. URL: http://www.jstor.org/stable/2682871.

Optional

- RadioLab. *A Very Lucky Wind*. URL: http://www.radiolab.org/story/91686-a-very-lucky-wind/.
- Jiajie Zhang and Donald A. Norman. *The Representation of Numbers*. URL: http://bit.ly/2fPXAkU.
- Kate Crawford. *Artificial Intelligence's White Guy Problem*. [Online, accessed 08-Nov-2016]. June 2016. URL: http://nyti.ms/28VgTst.
- Lauren Kirchner Jeff Larson Surya Mattu and Julia Angwin. *How We Analyzed the COMPAS Recidivism Algorithm*. May 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
- Wikipedia. *Tay (bot)*. [Online; accessed 08-Nov-2016]. 2016. URL: https://en.wikipedia.org/wiki/Tay_(bot).
- WildML. *Deep Learning for ChatBots*. [Online, accessed 08-Nov-2016]. Apr. 2016. URL: http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/.
- Amit Datta, Michael Carl Tschantz, and Anupam Datta. **Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination**. In: *Proceedings on Privacy Enhancing Technologies* 2015 (1 Apr. 2015), pp. 92–112. ISSN: 2299-0984. URL: http://dx.doi.org/10.1515/popets-2015-0007.
- Matthew Kay, Cynthia Matuszek, and Sean A. Munson. **Unequal Representation and Gender Stereotypes in Image Search Results for Occupations**. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: ACM, 2015, pp. 3819–3828. ISBN: 978-1-4503-3145-6. URL: http://doi.acm.org/10.1145/2702123.2702520.
- Judea Pearl. **Comment: Understanding Simpson's Paradox**. In: *The American Statistician* 68.1 (2014), pp. 8–13. URL: http://dx.doi.org/10.1080/00031305.2014.876829.
- This American Life. *No Coincidence, No Story!* Mar. 2013. URL: https://www.thisamericanlife.org/radio-archives/episode/489/no-coincidence-no-story.
- Cynthia Dwork et al. **Fairness Through Awareness**. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: ACM, 2012, pp. 214–226. ISBN: 978-1-4503-1115-1. URL: http://doi.acm.org/10.1145/2090236.2090255.
- Edward Segel and Jeffrey Heer. **Narrative Visualization: Telling Stories with Data**. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (Nov. 2010), pp. 1139–1148. ISSN: 1077-2626. URL: http://dx.doi.org/10.1109/TVCG.2010.179.
- Donald A. Norman. **Things That Make Us Smart: Defending Human Attributes In The Age Of The Machine**. Basic Books, Apr. 1994. ISBN: 0201626950. URL: http://bit.ly/2gatpJc.

Academic Integrity: As set forth in NYU Abu Dhabi's Academic Integrity Policy, the relationship between students and faculty at NYU Abu Dhabi is defined by a shared commitment to academic excellence and is grounded in an expectation of fairness, honesty, and respect, which are essential to maintaining the integrity of the community. Every student who enrolls and everyone who accepts an appointment as a member of the faculty or staff at NYU Abu Dhabi agrees to abide by the expectation of academic honesty.

The full policies and procedures relating to Academic Integrity may be found on the NYUAD Student Portal.