**Data 101 – Pilot Class on data literacy (198:142)**

**Mission**: To teach data literacy

**Goal:** Get students excited about how being “data literate” can change their careers and make them more successful. How to win an argument using data? How to use data to make more money or to win an election? Here we will teach them how to visualize patterns in data using very easy to use software. In the process they will learn basic statistics, find out about performance challenges and need for more sophisticated stat techniques as well as for programming. This class should make them interested and motivated to take more advanced classes in statistics and programming with data.

The term “big data” is the buzzword of the decade, but it’s not just a buzzword. The amount of data is expected to double every year, and data will affect nearly every profession. The New York Times predicts massive openings in jobs on data science:  “A report last year by the [McKinsey Global Institute](http://www.mckinsey.com/Features/Big_Data), the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with “deep analytical” expertise and 1.5 million more data-literate managers, whether retrained or hired.”

This is why data literacy ranks as a key skill, equivalent to the ability to read and write. It’s a skill which requires a combination of statistics, database management, and data extraction and cleaning. Some programming is a must. It will allow you to find data, extract it, analyze it and draw statistically valid conclusions which are important and actionable to your line of business, whatever that business is.

Everyone from liberal arts majors through the sciences should be data literate. What does this mean? It means to be able to understand where to find data, how it is formatted, how to extract it, and how to reason from it. This requires some basic computer science and certainly some basic statistics. But data literacy also means awareness of major issues regarding the impact of data on society. Recent controversies regarding the NSA work are a good example of impact I am talking about. A month ago *Nature p*ublished a paper showing that 4 locations are sufficient to identify a user with 95% confidence, this was based on the massive data set from one of the cellular providers. Many users are puzzled how Google knows and target ads so accurately, as if it was watching your every action on internet. In fact it does. Each such issues has its pros and cons.

A data literate university graduate should be aware of these societal pros and cons since they will likely face them.

***COURT OF DATA: The Teaching Method***

There are many classes teaching “Data science” which are available online and taught at many universities today. There are MOOC courses as well. But I do not believe that anyone teaches it the way I plan to.

*How to make class interesting, hands on and useful?*

The core of the proposed “Data” class will be centered on the ***Court of Data***. Modeled after Court TV (I will be Judge Judy☺) and Harvard MBA Case analysis, I will hear cases presented by students from the **data puzzles**

**PLAN (**still under construction)

1. Week 1-2: – **Data Buzz:** **Pros and cons**

The role of data science in society today – based on several sources starting with great book by political blogger and statistician Nate Silver

<http://en.wikipedia.org/wiki/The_Signal_and_the_Noise>

and some amazing successes he had in election prediction and other areas

<http://en.wikipedia.org/wiki/Nate_Silver>

New York Times data graphs - some of great data visualizations which data department at New York Times publishes from time to time some plus material from classic books on statistical data visualization

<http://en.wikipedia.org/wiki/Edward_Tufte>

How much data is collected about you now? And how much will be collected in the future and how? Facebook, mobile devices, Google, gmail etc etc. I want to discuss positives and negatives of this.

Dave Eggers in his recent book *“The Circle*” addresses fiction the Orwellian world where Facebook like company (The Circle) makes privacy totally obsolete (“privacy is a theft” is one of the slogans in the Circle) I will talk about it and have discussion in class. This book has created a lot of discussion and controversy and it was just published in the Fall of this year.

<http://www.amazon.com/The-Circle-Dave-Eggers/dp/0385351399>

<http://www.cnn.com/2013/10/24/opinion/galant-eggers-the-circle/>

Thus, rather than start with math/stats or tedious intro to unix like programming (R), I would like to get students intrigued and excited about “court of data”.

The result of this first two weeks should be “Isn’t that cool?” I would like to do that!

1. **Week 3: How one can be wronged by data.**

One has to be careful with data but not jumping to false conclusions on the basis of sample which it too small or ignoring hidden variables which are causal and instead presenting random associations. I will present a number of famous examples of wrong conclusions from data, led by famous Simpson’s paradox. This will motivate students about data science – it easy to be misled by data.

<http://en.wikipedia.org/wiki/Simpson's_paradox>

3..Week 4-5: **“You gotta see it” . Visualize!**

I will show some of the MIT Media Lab incredible displays of data visualizations. The site http://www.manyeyes.com/software/analytics/manyeyes/site represents some social network on data analysis and visualizations

Every student will “own” a data set. We will use and teach SAS Institute’s JMP software. It is a simple menu driven GUI that can apply any state techniques to any data set and show the results in graphical form. Most importantly we will introduce data visualization techniques such as [scatter plots](http://en.wikipedia.org/wiki/Scatter_plot), [histograms](http://en.wikipedia.org/wiki/Histogram), [probability plots](http://en.wikipedia.org/wiki/Probability_plot), [spaghetti plots](http://en.wikipedia.org/wiki/Spaghetti_plot), residual plots, [box plots](http://en.wikipedia.org/wiki/Box_plot), block plots and [biplots](http://en.wikipedia.org/wiki/Biplot" \o "Biplot), clusters, and some other techniques summarized below

<http://en.wikipedia.org/wiki/List_of_graphical_methods>.

We will also introduce core probability concepts, averages, variances, normal distribution and techniques such as regression, basic sampling and prediction as we go.

1. Weeks 6-12: **The COURT OF DATA**

Students will present their findings in class and there will be critical discussions which follow their presentations.

We will introduce basic stat analysis but through examples. In “data driven form” . The philosophy will be “Used something? Explain how it works” and I will be there to help them go through it.

1. Last two weeks**: Datathon (Hackathon to find real data patterns)** Everyone will be given the same data set and everyone will try to find “something interesting there” and back it up statistically. The data set will most likely be synthetic, i.e. I will embed there patterns to be found by students. Students will compete for the Data Science Idol title.