# STIA-210: Intro to Data Science

*3/22/2019*

*Data Science @ Georgetown School of Foreign Service*

The Introduction to Data Scienceis a survey course of fundamental concepts and techniques used in data science. The course is focused on evaluating and analyzing public policy, telling stories with data to make compelling and fact-based arguments.

The objective of the course is to equip students with the skills to allow data to take an active role in policy and strategy. Public policy is part of a large and sprawling social system. Parsing causality from a system of variables where everything is related requires a scalpel. This refined approach can be assembled from pre-written code and routines; but it still requires skilled assembly. We will teach an approach that leverages analytical routines that have already been written. The value of this course is in the mortar, not the bricks.

## Instructors

**Professor**: Jeff Chen (jeff.chen@georgetown.edu) is a statistician and data scientist with experience leading data initiatives in over 35 domains, working with diverse stakeholders such as firefighters, climatologists, technologists. Currently, he serves as the Chief Innovation Officer of the US Bureau of Economic Analysis (BEA) – the US agency responsible for the Gross Domestic Product (GDP). His work at BEA focuses on integrating machine learning and alternative data to improve the accuracy and timeliness of economic indicators and explore new modes of measurement. Prior to BEA, Jeff has held various technical leadership positions, most recently serving as Chief Data Scientist at the US Department of Commerce; a White House Presidential Innovation Fellow with NASA and the White House Office of Science and Technology Policy focused on data science for environmental issues; the first Director of Analytics at the NYC Fire Department where he engineered pioneering algorithms for fire prediction; and was among the first data scientists at the NYC Mayor's Office under then-Mayor Mike Bloomberg. Jeff started his career as an econometrician at an international engineering consultancy working on large scale public-private partnerships for infrastructure. On the side, he serves as a data science advisor to a Major League Soccer team and machine learning start up. Jeff holds a bachelor's in economics from Tufts University and a master's in applied statistics from Columbia University.

**Professor**: **Bryan Baird** (beb53@georgetown.edu) is a Senior Data Scientist with Civis Analytics' Government Practice. A recent graduate of Georgetown's McCourt School, he has worked as a technology policy analyst. Prior to Georgetown, worked in technology and politics, including time managing tech products at Microsoft in Seattle and CHIEF in DC. He received his Bachelor's Degree from Washington University in St. Louis, with double majors in Systems Engineering and Political Science.

## Time and location

14 classes will be held on Mondays from 6:30pm to 9:00pm:

- August 28*
- September 9, 16, 23, 30
- October 7, 21, 28
- November 4, 11, 18, 25
- December 2, 9

*Please note that for the first week only, all Monday scheduled classes will meet on Wednesday, August 28th.*

## Website

Students are expected to sign up for a Github account (https://github.com/stia-210/data-science). A e-book and accompanying notes and materials has been written for this class and will be made available via a Github web page.

## Workload and assignments

- **Assignments.** Students will be evaluated on the basis of six problem sets (60%) and one final project (40%). Late problem sets will be penalized by 10-percentage points per day. To qualify for passing the class, students will need to make submissions for all six problem sets and final project for the period when points still can be earned. All problem sets will be submitted electronically.
- **Final Projects.** The final project assignment details will be made available in late March and the final product will be due on *Monday May [TBD]*. As this class is quite hands-on, it is expected that students bring their computers to class to partake in computational activities.
- **Collaboration** is the greatest source of creativity and innovation. Thus, collaboration should not result in verbatim submissions (e.g. no copy cats). As everyone writes code following their own unique logic, the chance of identical submissions is unlikely and easily detectable. Non-unique code will be penalized.
- **Do it the hard way**. A core determinant of the success of a data scientist is being able to explain how an algorithm or analysis was constructed, not just use software. In this class, where possible, build from scratch rather than an overly convenient library. This will allow you to become more creative down the line (e.g. difference between building lego sculptures from a factory blue print versus conceiving it on your own).

## Course Outline

Data science is dependent on sound application of computer programming, mathematics/statistics, and communication. This course is thus organized into three units that dive into the fundamentals. Particular emphasis is placed on skilled assembly of empirical ideas, drawing from standard and non-standard data. The section outlines are subject to change up to one week before the lecture. Please continue to review the syllabus throughout the course.

### Section 1: Fundamentals

Data science is about designing and building data products that derive insight. This first section will focus on developing fundamental skills required to build effective products.

### Lecture 1 & 2: Preliminaries and Getting Comfortable with Programming

Where does data science fit into this world? This lecture focuses on framing and presenting data science as an active approach to societal problems using quantitative methods. This requires not only an understand of the context, but the value of combining statistical theory with advanced programming to accomplish extraordinary tasks.

**Topics**:

- Data science: What is it? What is the lay of the land?

- A framework: Benchmarks, Explanations, and Predictions
- Languages of data science
- Intro to programming and the R statistical programming language
- Data types and classes, including matrix, data.frame, list, and vectors
- Extracting rows, columns, and specific elements from a data frame
- Basic operations (e.g., sum, mean) on rows.

**Example application**

- Graphing photo-voltaic energy data from the National Institute of Standard and Technology's Net Zero Energy Residential Test Facility

**Warm Up Assignment (Scored but does not count)**

- Homework #0: Exercises to become familiar with `R`

## Lecture 3: Manipulation / Wrangling / Feature Engineering

The objective of this lecture is to present the most important and fundamental elements of data manipulation. These core operations include sort, merge, reshape, and collapse. We will also present loops through multiple rows or columns, and other alternatives to operate on partitions of data frames.

**Topics**

- Import and export of data
- Sort data based on column values
- Subset vectors, matrices and data frames
- Reshape data between wide and long form
- Collapse data into aggregates
- Merge two or more data sets
- Text processing and feature engineering using regular expressions

**Example application(s)**

- Parsing and conducting basic text analysis using State of the Union Speeches (2009 to 2016)
- Comparing and joining EU and UN sanction lists

**Homework #1 Assigned, due by next class**

- Convert individual police incidents into time series projections

## Lecture 4: Control Structures, Functions and Etiquette

Building upon basic data manipulation and high level analytical tasks, this session will focus on programming paradigms that are commonly relied upon when practicing data science.

**Topics**

- Control structures: `for` loops, `while` loops, `if/else` statements as well as R-specific methods such as `apply/lapply/sapply`.
- Suitable practices and etiquette such as the Google Code Style Guide.
- Writing functions that make sense and are reliable

**Homework #2 assigned, due by nextclass**

- Writing functions
- Batch Extraction of Census Housing Permits data

## Lecture 5: Exploratory Data Analysis (EDA)

The objective of this lecture is to handle missing values appropriately and script visual checks to find errors introduced in data input/output. We will also start to examine computational optimization techniques, like taking advantage of multiple cores for heavy duty operations.

**Topics**

1. Understanding data structures and idiosyncrasies
2. Statistical measures
3. Graph and visual analytics – intro to `ggplot2` library

**Example application**

- Finding health coverage patterns using the US Census American Community Survey
- Conducting analysis of missing values analysis of weather anomalies from 1880 to Present using the National Oceanographic and Atmospheric Administration's GHCN-M

**Homework #3 assigned, due by next class**

- Exploratory Data Analysis of Washington Metropolitan Area Transit Authority's train delays data

## Section 2: Building Insight

In public policy, data can be used to support evaluation of programs to understand causal mechanisms (e.g. retrospective focus) or enable the creation of data-rooted products that drive action (e.g. deployed applications). In this section, we extend data manipulation to the algorithmic intelligence necessary to make sound decisions.

## Lecture 6, 7, 8: Introduction to Supervised Learning and Prediction

Supervised learning is arguably the most relied upon class of techniques that enable causal inference but also deployed precision policy. How does changing one variable independently impact another variable? We begin to introduce basic regression analysis, correlation coefficients, ordinary least squares, and the relationship between the concepts. Note that this is a very cursory review, and the deep assumptions are not tested or expounded upon.

**Lecture objectives**

- Effect identification versus prediction – what's the difference?
- What is supervised learning?
- Prediction as a process: bias-variance trade off, model validation (e.g. train/test, k-folds cross-validation), feature selection
- Types of supervised learning: discrete problems and continuous problems
- K-Nearest Neighbors (kNN)
- Ordinary Least Squares (OLS)

**Example application**

- Prediction of missing values in satellite imagery using kNN

**Homework #4 assigned** - Write functions for cross-validation

**Homework #5 assigned** - Build a housing sales price prediction model

## Lectures 9, 10, 11: Classification techniques

Classification models are one of the workhorses of data science. Classifiers enables data-driven applications such as risk scoring, lawsuit outcome prediction, marketing lead generation, facial detection and computer vision, spam filtering, among other use cases. This session will focus on the fundamentals of classification models, types of models, and daily applications.

**Topics**

- Three common problems using classifiers
- Structure of a classification project, Target variables, Input variables, Objective function and evaluation measures, model experiment design, Cross validation versus train/validate/test, Confusion matrix, TPR, TNR, AUC
- K-Nearest Neighbors
- Linear Methods: Logistic Regression / Binomial Generalized Linear Models and General Additive Models
- Decision trees
- Ensemble and Bagging methods: Random Forests
- Boosting and other methods (time permitting)
- Appropriate uses of classification techniques, Scoring, prediction and prioritization, Propensity score matching

**Example application**

- Healthcare insurance coverage data

**Homework #6**

- Lec 7: Predict activity using smartphone accelerometer data.
- Lec 8: Hand out class project instructions, one page proposal of what you'll do due by Lec 9.

## Lecture 12: Unsupervised learning

No, this is not an independent study session. Unsupervised learning techniques such as clustering help to identify recognizable patterns when no labels (e.g. outcomes, dependent variables) are available. In sales and recruitment offices, customer segmentation may use current customer data, then use clustering techniques to identify k-number of distinct customer profiles. In resourceful law firms, data scientists may develop topic modeling algorithms to automatically tag and cluster hundreds of thousands of documents for improved search. This session will focus on clustering methodologies that are commonly employed in applied research.

**Topics**

- Defining clustering and distance
- Structure of unsupervised learning project, Input variables, optimization methods
- Statistical assumptions and mechanics, risks/strengths, implementation, sanity checks, non-technical explanation of a few techniques
- K-means clustering
- Hierarchical clustering
- Stability testing using silhouette coefficients

**Example application**

- Univariate clustering application: k-means
- Multivariate clustering application: Customer segmentation using NYC 311

# Section 3: Beyond Algorithms

Beyond the data preparation and modeling, the presentation layer is the glue that will allow a data science project to stick with target audiences. Often times, presentation is graphical and relies upon a rich ecosystem of visualization, web services, and interactive applications to communicate pertinent issues.

## Lecture 13: Data products

Data science presents organizations with the opportunity to allow data to play an active, action-oriented role in daily operations. Newspapers such as the New York Times and the Washington Post are increasingly relying upon visual narratives to illustrate the point. Tech giants like Facebook and Amazon rely on recommendation engines to drive sales. Political canvassing operations, direct-to-consumer marketers, and infrastructure engineers rely on prioritization models to increase their hit rates in the field. For data science to drive value, data products must be grounded in an audience and a well-defined need.

**Topics**

- Visual narratives – reports with interactive graphical elements (e.g. policy reports, data journalism, maps)
- Recommendations – well-packaged lists of items that users may be interested in (e.g. eCommerce shopping, search)
- Prioritization exercises – rank ordered lists of things that an operation should act upon (e.g. targeted marketing list, inspection lists)

**Example application(s)**

- Build a recommendation engine for consumer purchases
- Process spatial data (GIS) and produce an interactive web map

## Lecture 14: SQL and Python

Data science, statistics, and machine learning are agnostic of languages. That being said, there are a host of techniques and technologies that data scientists rely upon to be effective. In the increasingly more competitive labor market, knowing more than one language is to one's advantage. In addition to R, which is more of the research data scientist's tool, Python is a pseudo-code ("p-code") language that is relied upon for build full-scale web application. To extract data from databases, Structured Querying Language is also relied upon.

**Topics**

- Understand how to write SQL queries
- Intro to Python in iPython Notebook (Jupyter)

**Pre-Class Activity**

- Download and install Anaconda – a pre-packaged suite of Python-based tools

# Academic Resource Center/Disability'Support

If you believe you have a'disability, then you should contact the Academic Resource Center (arc@georgetown.edu) for further information. The Center is located in the Leavey Center, Suite 335 (202-687-8354). The Academic Resource Center is the campus office responsible for reviewing documentation provided by students with disabilities and for determining reasonable accommodations in accordance with the Americans with Disabilities Act (ASA) and University policies. For more information, go to http://academicsupport.georgetown.edu/disability/.

## Important Academic Policies and Academic Integrity

McCourt School students are expected to uphold the academic policies set forth by Georgetown University and the Graduate School of Arts and Sciences. Students should therefore familiarize themselves with all the rules, regulations, and procedures relevant to their pursuit of a Graduate School degree. The policies are located at:'http://grad.georgetown.edu/academics/policies/

## Provosts Policy Accommodating Students Religious Observances

Georgetown University promotes respect for all religions. Any student who is unable to attend classes or to participate in any examination, presentation, or assignment on a given day because of the observance of a major religious holiday (see below) or related travel shall be excused and provided with the opportunity to make up, without unreasonable burden, any work that has been missed for this reason and shall not in any other way be penalized for the absence or rescheduled work. Students will remain responsible for all assigned work. Students should notify professors in writing at the beginning of the semester of religious observances that conflict with their classes. The Office of the Provost, in consultation with Campus Ministry and the Registrar, will publish, before classes begin for a given term, a list of major religious holidays likely to affect Georgetown students. The Provost and the Main Campus Executive Faculty encourage faculty to accommodate students whose bona fide religious observances in other ways impede normal participation in a course. Students who cannot be accommodated should discuss the matter with an advising dean.

## Statement on Sexual Misconduct

Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. However, university policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

Jen Schweer, MA, LPC
Associate Director
Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley
Trauma Specialist
Counseling and Psychiatric Services (CAPS)
(202) 687-6985
els54@georgetown.edu

More information about campus resources and reporting sexual misconduct can be found at http:// sexualassault.georgetown.edu.