**INTA 4803 JD: Digital Political Science--Spring 2018**

**Course Description & Overview**

## Introduction

*INTA 4803 JD—Digital Political Science* is a practice-focused course designed to give INTA students the knowledge and skills to effectively apply academic research methods within an increasingly digital world. The goal of the course is to help each student build a ‘personal digital toolkit’ will increase the efficiency, effectiveness and reproducibility of their academic research. As such, Digital Political Science complements the core INTA methods class (INTA 4003/6003—Empirical Methods) by focusing on how students can use software to execute the research design principles, methods and statistical analysis taught in the core class. Digital Political Science students will gain experience using digital tools to:

* Plan their research projects
* Use structured search techniques to find relevant literature and data
* Explore this literature and data using statistical and natural language approaches
* Perform common statistical analyses and
* Structure evidence into compelling arguments

The course incorporates Python, the most widely taught and easy to learn programming language in computer science, as well as a select number of other digital tools to assist in the performance of the major step in the execution of an academic research project. The course is designed as a seminar, with each class session consisting of lecture / discussion plus working sessions with the tools, and each student will also work outside of the class to apply the tools to one of his or her existing research projects.

## Background on the Course

This course trains you to be a idea translator. Let me explain.

We live in a data-rich digital world, and our lives have been made easier and more convenient by the many complex algorithms applied to these data. Unfortunately, few of us possess a basic understanding of how these algorithms work or the strengths and limitations of these data, which creates a gap in our understanding of the world and the research that we perform. The social sciences, including political science, have done little to address this gap, with most instruction on how to understand and use the data rich digital world relegated to computer science, business, and hard science departments. This handicaps political science students, as well as the advance of knowledge in the field, and it places students at a disadvantage in the job market after graduation.

This course is meant to address this gap and give you sufficient training to take advantage of the data-rich digital world, with as little pain (and computer programming esoterica) as possible. I’ve designed the class to be useful more than anything, regardless of what field each student eventually ends up in after graduation—business, think tank, government, consulting, continuing in academia, etc.  The design stems from my own experience over 20 years in business and academia, where I had to constantly figure out how to remain relevant as a strategist during the transition from a non-digital, pre-internet world to the hyper-connected, information overload world of today.  What I figured out was that the more change occurred, the greater the value of translators—people with sufficient knowledge of the technology to be able to talk to technical people and understand value and limits of what they did, but with exceptional skills in distilling the ‘so what’s’ (i.e. insights)  from the technical, situating these insights within the context of an organization’s overall challenges, and communicating with / creating bridges across people who see the world in very different ways.

I feel very strongly that: a) the digital world of information overload + ever more powerful computers and algorithms that I experienced over 20 years will continue to grow, b) we cannot know how this growth translates into what jobs are available and/or ‘hot’ in 5, 10, 15+ years, but c) translators are always going to be high value.  So, I designed the course more than anything to help students build the core technical and communication skills necessary to be good idea translators.

## Course Design

Digital Political Science is a hands-on course 'for non-techies / by a non-techie' where you will these skills and apply them to your own work. The primary objective of the class is to improve your effectiveness and efficiency in performing empirical academic research. The research domain of this course is the political science sub-field of international relations, which includes comparative politics, security studies, and international political economy / development, but the research framework and digital tools that you will learn are equally applicable to other academic domains, as well as policy development and business.

This research frame that we will use is called Structured Problem Solving & Communication (SPSC), which is a systematic approach to solving extremely difficult problems and communicating the results succinctly. SPSC originated in the 1970s at McKinsey & Company, the world's most respected business strategy consultancy, and it is based on extensive academic research on cognition and problem solving from a number of different fields. The central metaphor of the SPSC framework is the pyramid, which represents two complementary concepts. Within problem solving, the pyramid represents the concept that each step in the process should build upon the foundations laid in previous steps. This implies that the most important step in problem solving is the first, which is the definition of the problem to be solved, as the research performed and its interpretation flow directly from this definition. Within communication, the pyramid represents the concept that the your argument should be 'nested', able to be communicated at various levels of detail, and that each level should build up from an underlying fact base.

To support the execution of this framework, you will learn a set of tools that can be broadly categorized as 'digital methods'. Digital methods are computer-based tools that augment the strengths of human cognition, blending the superior data processing and pattern recognition capabilities of computers with the superior insight development, interpretation and story-telling capabilities of humans. Digital methods encompass a large number of general purpose and specialized computer programs that academic researchers use, including tools focused on statistical analysis (e.g., STATA, SPSS, R, Excel), bibliographic & research content management (e.g., Endnote, Mendeley, Zotero), natural language processing (e.g., MALLET, NVivo), GIS (e.g., ArcMap, QGIS), visualization (e.g., Tableau, Microsoft Power BI), and other areas.

Given the number of tools available, this course focuses on the use of Python, a general-purpose computer programming tool that is relatively easy to learn and has the benefit of many open source 'packages' that researchers have developed to perform specific tasks. Some of the packages that we will make use of in Python include:

* iPython / Jupyter Notebooks—An environment in which you can write and run code, see the results, and make annotations or comments in a visually compelling way. Jupyter notebooks makes learning more intuitive, and it also makes it easier to keep track of your code and comments on analysis as you go.
* Numpy/Pandas--a collection of data manipulation tools similar to excel, but with much more flexibility vis-a-vis data sources and standardization / automation of data manipulation (resulting in less effort on part of analyst and fewer errors);
* Matplotlib--a general purpose data visualization / graphing tools package that has basic and advanced visualization capabilities and is extremely customizable;
* Scikit-Learn--a package with tools supporting supervised and unsupervised machine learning, which enables researchers to perform such tasks as identifying topics within articles, classifying research observations automatically, and creating predictive models.

Although the course focuses on using Python, we will start off using components of two packaged software tools that do not require programming skills. This will enable you to apply course concepts immediately in the class, which should: a) add value to your work in other classes you are taking, and b) make it easier to learn to apply these concepts using Python later in the course. These tools are:

* Vantagepoint—A tool developed by Georgia Tech researchers to derive insights from the metadata of journal articles. Metadata include bibliographic information about the authors of a study, publication information including # of times that the article has been cited by other researchers, and text analysis of article abstracts. Vantagepoint is an excellent tool to use to quickly make sense of a body of literature about a topic, and it also serves as a good introduction to what we will be doing with Python and machine learning later in the course. Vantagepoint is free to Georgia Tech students, and installation will be covered in the first month of the class.
* nVivo—A commercial tool that is the ‘gold standard’ software for qualitative analysis in the social sciences, public health, and other fields. nVivo’s core strength is its ability to help you transform texts into quantifiable data, which you can then analyze using standard quantitative approaches. The primary method to do this is through ‘tagging’ specific words and phrases in a text, such as a journal article. nVivo has other tools that we will use, including tools to help organize texts for a literature review and a mind mapping module that will help you develop insights from your readings. nVivo is available to Tech students through mycloud.gatech.edu. We will go over how to set up and use nVivo during the first month of class.

Finally, we will spend a brief portion of the course, time permitting, using the Storymaps functionality of ArcGIS Online. Storymaps is a way to do GIS (Geographic Inforation System) mapping / analysis without having to learn the rather clunky software and very geeky information required by more comprehensive programs. It's a fantastic way to visualize data that have geographic components, like violent conflict event data, census demographic information, or voting patterns. It's also free, and it serves as a gentle introduction (and good presentation tool) for your first forays into GIS.

Some of the types of analysis that you will learn to do in this class include, with linked examples:

•***Bibliometric analysis*** —understanding how published research on IR topics has evolved over time, using metadata analysis.  Example:  [Shorter titled papers are cited more](http://www.nature.com/news/papers-with-shorter-titles-get-more-citations-1.18246).

•***Social network analysis*** —Quantifying and visualizing how ideas and/or power is distributed within groups of individuals.  Example:  [See how agent interaction in auctions in 1800 – 1820 Europe demonstrates the concentration of economic power](http://www.jeffersonbailey.com/wp-content/uploads/2014/12/figure_8.jpg).

•***Machine learning*** —Understand what it is and how to (appropriately) apply it to solve problems, such as to distill themes from large amounts of text or quantitative databases.  Examples:  [Networked Democracy lab innovation projects at USC](http://dornsife.usc.edu/labs/netdem/research/).

•***Storytelling with Maps*** —Using maps to communicate powerful stories, such as the distribution and change in the level of violence in a civil war.  Example:  [Interstate highways are linked to persistent inequalities in American cities](https://mbrunman.maps.arcgis.com/apps/Cascade/index.html?appid=9ac105b11d744d4f8ac4758c2ace3719).

Unfortunately, I cannot tell you how to do these things; you have to learn by doing. However, these things are not nearly as hard as they seem to be if you try tofigure them out on your own, or learn them from courses that are designed and delivered for people who think and dream in code. The materials that I have chosen to cover, the background information that you will read, and the exercises that you will do in your homework (yes, there is homework--sorry!) will make the process as painless as possible--but your skills will only be built by doing. Hence, the class is very hands on. I have also used data sets and articles pulled from IR / Political Science as much as possible, so that you can see the application of these tools in the context of work with which you are very familiar.

## Course Flow / Timeline

The course is organized into a series of modules that build on each other and that follows the sequence of iterative steps used in 'structured problem solving'. SPC is well suited to a variety of highly intellectual domains including engineering, business strategy, and political science

## Course Timeline

The breakdown for the class is as follows. Note that there may be some slight modifications to this timeline as we progress through the course.

| **Week # (Dates)** | **Module** | **Topics** | **Description** |
| --- | --- | --- | --- |
| **W1 (1/11)** | ***Introduction: What / Why Digital IR?*** | Course Overview, Research Pyramids, and ‘Guerrilla Reading’ Journal Articles | Introduction to the course, including core ideas behind digital political science, efficient research design & structured problem solving and communication. We’ll also review what I call ‘guerilla reading’—an approach to quickly identify the thesis and arguments in complex journal articles, and something that you will find immediate value from in your other courses. |
| **W2 (1/18)** | ***Digital Analysis of Qualitative Data--Literature Review*** | Digital Literature Review I: Effective Search Strategies and Metadata Analysis | In this session, we will cover the foundation for all digital methods-- how to efficiently search for and distill core concepts from literature. We will use Vantagepoint to perform basic analysis of a large-ish body of literature and find interesting insights about its contents.  **Homework #1 due.** |
| **W3 (1/25)** |  | Digital Literature Review II: Identifying Themes in a Body of Research | We will continue our module on literature review by using Vantagepoint to identify topics in the literature, identify the first use of key concepts, and describe the uptake of these concepts over time by different researchers.  **Homework #2 due.** |
| **W4 (2/1)** |  | Digital Literature Review III: Mind Maps / Network Maps | Our final session on digital literature review will focus on two types of non-geographic mapping. Mind maps are conceptual models of how different ideas relate to each other, and they are extremely useful to structure and learn new topics. Network maps, or relationship maps, are visual depictions of how  **Homework #3 due.** |
| **W5 (2/8)** | ***Python Crash***  ***Course for Digital Political Science*** | Python Crash Course I—Numerical Python | In this session, we will learn the basics of Python by doing simple analytical tasks with numerical data, specifically data exploration. This includes some work on data visualizations that are useful for data exploration.  **Mini-quiz A (in class / 15 minutes—don’t stress)**  **Homework #4 due.** |
| **W6 (2/15)** |  | Python Crash Course II—Basic Inferential Stats & Visualizations with Python | In this session, we will continue learning the ‘bare bones’ of Python by using it to perform inferential statistical tasks, such as calculating correlations and multivariate regressions. We will also explore visualizations in more detail as related to quantitative analysis. **Homework #5 due.** |
| **W7 (2/22)** |  | Data Gathering & Web Scraping | This will be our first ‘heavy’ application of Python, which we will use to efficiently grab data from a political science data source using its API (Application Program Interface). We will also cover the basics of extracting structured data from static files like PDFs. **Homework #6 due.** |
| **W8 (3/1)** | ***Identifying Patterns & Evidence (Python / SPSC)*** | Exploratory Data Analysis I— Understanding the Data & Finding Interesting Patterns | Exploratory data analysis is the process of getting to know the strengths & limitations of the data that you have and identifying interesting patterns. This is an important area of the research process that does not get sufficient attention, and Python can help you do it efficiently.  **Mini-quiz B (in class / 15 minutes—don’t stress)**  **Homework #7 due.** |
| **W9 (3/8)** |  | Inferential Analysis I— ANOVA, Correlation, and Regression (MVR, Binary) | This session will cover how to use Python to walk through the steps of causal statistical analysis, including the very important step of examining patterns in residuals (i.e., variance in the data not explained by the causal analysis).  **Homework #8 due.** |
| **W10 (3/15)** |  | Machine Learning I: Natural Language Processing | Our second heavy use of Python will be with a hot topic that has applicability for academic research—natural language processing (NLP). NLP is the process of converting text into data objects that you can manipulate (count, order, etc.) using Python, and it is useful for ‘remote reading’—using the computer’s capabilities to extract core ideas and concepts from a body of academic literature.  **Homework #9 due.** |
| ***-- (3/22)*** |  |  | ***Spring Break--No Class*** |
| **W11 (3/29)** |  | Machine Learning II: Trees & Random Forests | The last topic we will cover with Python will be machine learning, which we touched on in the NLP session. Machine learning is an approach to distill patterns of relationships between large numbers of variables in big data sets. These techniques power many day-to-day interactions we each have with the internet, such as the ‘You might also like...’ feature on Amazon.com.  **Homework #10 due.** |
| **W12 (4/5)** | ***Pyramids for Insight Development & Communications (SPSC)*** | Finding Your Point— Fact Based Insight Development | The most important part of any communication is having a point. In this session, we will focus on how to use the SPSC pyramid principles to organize the results of your analysis from the ‘bottom up’ into pyramids of insight.  **Mini-Quiz C (15 minutes / in-class. Don’t stress).** |
| **W13 (4/12)** |  | Making Your Point— Pyramid Communications | This session will focus on the ‘top down’ application of the SPSC pyramid concept to effectively communicate your research findings in a fact-based way. This top-down structuring is the most effective way to communicate in most situations, and it is applicable to written, verbal and visual communications. |
| **W14 (4/19)** | ***Wrap-up and Next Steps*** | Course Wrap Up & Presentations | In the last class session, individuals will verbally present a 2-3-minute executive summary of the research project that they worked on during the term. We will also have a group activity to build a communication pyramid that encapsulates the key concepts covered in the course. |
| **Finals (4/30)** | ***End of Term Deliverables*** | Presentations and classmate feedback | Individual presentations and oral feedback by all class members. **Attendance mandatory for all of time period.** |
| **End of Finals (5/2)** |  | Paper Due | Paper due to T-Square by 11:59PM 5/2 |

## Course Materials

The following texts are required for the course, and they will also serve as a solid foundation for your digital political science library in the future:

* Kane, Frank (2017). Hands-On Data Science and Python Machine Learning. Birmingham- Mumbai: Packt. Perhaps the best practical orientation to using Python to do things vs. to learn how to code. Downloads for data for book exercises [here](https://www.packtpub.com/books/content/support/29713) (note--you must register your email to get code). Price / availability: eBook [at publisher](https://www.packtpub.com/big-data-and-business-intelligence/hands-data-science-and-python-machine-learning)--$5.00 ($39.99 e-book + print), or at Amazon (Kindle $25.99/ $30.99 paperback).
* The Vantagepoint, Inc. (2015). VantagePoint Users Guide. Norcross, GA: Search Technology, Inc. Somewhat technical guide to the software, but a good reference that will be useful. Price: $0. Available online at:   
  [https://www.thevantagepoint.com/\_Analyst\_Guide\_Online\_/Users\_Guide/VP Users Guide.pdf](https://www.thevantagepoint.com/_Analyst_Guide_Online_/Users_Guide/VP%20Users%20Guide.pdf)
* Vothihong, Phuong, Czygan, Martin et al (2016). Python: End-to-End Data Analysis Learning Path. Birmingham-Mumbai: Packt. Don’t be scared of this book; we are not going to use all of it, and if you can get the eBook for $5 it is well worth acquiring. A comprehensive view on how to do most (all?) data science stuff with Python, in a step-by-step ‘cookbook’ fashion. Price / availability: eBook only, [at publisher](https://www.packtpub.com/big-data-and-business-intelligence/python-end-end-data-analysis) ($5.00) or at Amazon (Kindle only, $57.99).

In addition to thse materials, I will assign supplemental readings as we as background, examples, or instructions for specific analyses. These will be available on Canvas.

Additional books you may find useful:

* Andres, Ana (2009). Measuring Academic Research: How to Undertake a Bibliometric Study. Oxford, UK: Woodhead Publishing Limited. Detailed review of why bibliometric analysis is useful and how it is performed, written with all of the citations and footnotes that you would expect of academics.
* Bruce, Andrew and Bruce, Peter (2017). Practical Statistics for Data Scientists. Sebastopol, California: O'Reilley Media Inc. A fine, concise overview of the most useful statistical methods used in real-world research. This is a nice book in that it combines brief overviews of the use—and limitations—of different statistical methods, then shows examples of how you can do this in Python and/or R. Much more approachable and useful than either a general statistics reference textbook or the manual for a specific piece of statistical software like SPSS or SAS.
* Downey, Allen B. (2015) Think Stats: Exploratory Data Analysis. Sabastapol, CA: O'Reilly Media. A very good book on a neglected aspect of research, exploratory analysis.
* Magallanes-Reyes, Jose Manuel (2017). Introduction to Data Science for Social and Policy Research: Collecting and Organizing Data with R and Python. Cambridge, UK: Cambridge University Press. Good, if technical, book on how to access every conceivable type of data from websites, either through APIs (the web for data) or by web-scraping.
* Wolf, Nicholas H. and Silver, Christina (2018). Qualitative Analysis Using NVivo: The Five Level QDA Method. New York, NY: Routledge. Good introduction to both the process of qualitative analysis of text as well as the performance of this analysis using NVivo. Price: $33.91 Kindle, $33.75 paperback at Amazon.

## Course Website--Canvas

We will be using Canvas, Georgia Tech's new learning management system (LMS) for course communications, distribution of materials, assignments (materials distribution and for turning in), and for betwen-class comments. I will also be using Canvas for grading and to administer quizzes. Canvas (eventually) will replace T-Square, which I will not expect you to check after the first class meeting. We will go over the Canvas functionality in the second class meeting.

## Computer Equipment and Software

Digital methods are by definition methods that are computer-enabled, and half of each class will be lab-type work. **Therefore, you will need to bring a computer to class**. You can use what operating system you prefer--Mac, Windows (I use both), or Linux. Python itself is open source and available for any of these three operating systems, and its functionality is identical on each. However, Vantagepoint is only available on Windows, and the Mac version of NVivo has only parts of the functionality of the Windoes version. To make the use of these programs consistent for everyone, we will access these programs through the virtual lab for Tech on the web (www.mycloud.gatech.edu). These programs should be pre-installed on your virtual machine, which we will check in the first class. You may prefer to also do your Python work on the virtual machine, so that all of your course-related work is in one place and available whenever you have internet access, but that is your call. Personally, I have Parallels desktop on my Macs so that I can run Windows programs and Mac programs without rebooting my machines. VantagePoint is a hassle to install locally, however, so I use mycloud to use that program. I use a thumb drive to keep most of my files, in addition to Dropbox for more permanent / accesible storage.

I also highly suggest--both for this class as well as more generally to manage your work--that you learn to go completely digital. The two most important, but underutilized, tools I think are critical to this are a bibliographic management software program and a note taking program. Bibliographic management programs include Mendeley (which I use), EndNote, Zotero and others. Their ostensible purpose is to help you manage references for papers, and they are indeed worth the investment in time and effort for this feature alone. However, they have much greater functionality and they can serve as the 'one source' for all of your journal article files, a highlighting / note taking solution for these articles, and a good basic search tool for your information. We will go over some of these uses in the first class, and if you haven't done so already please start using one of these. A complementary program for your completely digital life is a notebook tool, such as OneNote (Microsoft) or Evernote (which I use). These programs are 'digital scrap books / notebooks', and you can do just about anything that you could think of with paper in them--take notes, take screenshots of web pages, etc., scan in paper documents etc. In addition to ridding yourself of a bunch of junky looking paper scraps (which maybe you don't lose but I do), these tools also have powerful search open search features that makes it easy to find virtually anything. They index all of your text, which means that you don't have to create a lot of files / folders to stay organized. This means that you can search for, say, for your Ga Tech ID # if you don't quite remember it (again, maybe this is just me) by simply typing in 'Tech ID' 'ID' or 'Tech', instead of having to remember what folder you put it in (or where your wallet is).

An additional program you may consider is a password manager, which becomes increasingly critical as you use more and more sites and software and as more and more evil trolls attempt to steal your identity. I use LastPass, but there are others including the one that is native to your browsers. The advantage of LastPass is that it works as an add in to all of your browsers, which becomes important once you realize that you probably want to use several different browsers since they all have different strengths and weaknesses, as well as significantly different speeds. I use Chrome primarily because it is integrated tightly with several Google apps that I use including Google Maps and Google Developer (they have lots of really cool new toys including ways to access their distributed computing capabilities--amazing how much faster it is for really data intensive work like GIS analysis), with Firefox and Safari tied for second.

## Determination of Final Grade

|  |  |  |
| --- | --- | --- |
| **Component** | **Weight** | **Comments** |
| Class participation | 20% | Includes attendance, in class assignments, contribution to class discussion that includes evidence of having done readings, etc. |
| Graded homework assignments | 20% (10 @ 2% each) | To be discussed. Pass / fail, graded on effort more than ‘right’ answer. |
| Flash quizzes | 20% (4 @ 2.5% each) | Short, relatively basic quizzes to test your comprehension of key concepts as we go through the class. |
| Research project output (presentation + supporting files) | 30% | 1 page executive summary--20%; 10 - 15 slide powerpoint deck with speaker notes--50%; oral presentation--30% |
| Course concept map | 10% | 1 page visual representation of what you learned in the class and how interconnected |