**Description**

The digital, connected, sensor rich world is generating extraordinary amounts and variety of data (“Big Data”). CSS is an exciting new scientific perspective that incorporates new methods and models for studying human behavior from the level of neurons to collective behavior. This change in approach has already made breakthroughs possible in understanding human creativity, scientific performance, the sharing economy, human conflict, and consumer behavior.

This seminar will teach computational analysis skills. These skills include null model design and programming, and data mining for structured and unstructured data (topic models, bag of words, etc.). Students will leave the course with the technologies and intuitions needed for sophisticated independent research.

**Prerequisites**

Students must possess an understanding of how to program in Python before the course begins. Students **must have completed** **one of two options** before the course starts.

1. Have taken NICO-101 Introduction to programming for big data (P/NP or A/B/C are both allowed).
2. Register for [Datacamp](https://www.datacamp.com/) on-line and pass a set of courses at your own pace. Kellogg has negotiated access:  
   i. Go to [www.datacamp.com](http://www.datacamp.com) and create a fee account  
   ii. Go the NUIT Datacamp Access form. <https://app.smartsheet.com/b/form?EQBCT=b0e9f1016ecc4c47856f63dd1f52f62a>  
   iii. When filling out the form you will see a “Prerequisite” section that says you must first have taken one of two free introductory courses. Kellogg students may ignore that requirement and simply write “I need this for a Kellogg course” in the box labelled “Request Exception.”  
     
   This option is intended for those that learn best on their own or already know the basics of programming (in Python or another language). The following courses must be passed before the start of CSSMA:
   1. Intro to Python for Data Science
   2. Intermediate Python for Data Science
   3. Python Data Science Toolbox (Part 1)
   4. Python Data Science Toolbox (Part 2)
   5. pandas Foundations
   6. Manipulating DataFrames with pandas
   7. Importing Data in Python (Part 1)
   8. Importing Data in Python (Part 2)

**Software**

**You must have a laptop with a current version of Windows or OS X.**For Windows, you must be using at least Windows 7.For Macs, you must be using OS X 10.11 or later.

In this course we will be using the Anaconda 4.4.0 distribution with Python 3.6. It is essential that you install Python 3.6 and **not** Python 2.7.

The following videos will show you how to install Anaconda, it is an easy installation and typically has no problems (note these videos are for an old version of Anaconda, but the process is the same).

* OS X <https://www.youtube.com/watch?v=UQhOyZXHkxI>
* Windows <https://www.youtube.com/watch?v=w16iUU6IA5E>

**Course Materials**

The predominant course materials that we will use are ones that I have created and are freely available. The other course materials are primary academic literature.

**Grading**

*Assignments* (40%) Programming assignments and paper presentations.

*Participation* (10%) Code reviews and in-class discussion.

*Final Project* (50%) You will conduct an independent research project that involves the analysis of novel data. This project will commence week 4 and will be due at the end of the quarter. The deliverable will be a paper, written in academic style, that describes your research question and results. A write up of how you arrived at these conclusions (i.e. how you coded it) will comprise the methods section.

**Honor Code**

It is expected that code submitted for all assignments and projects will be original and independently written. The copying of code directly from online resources (e.g. stackoverflow) is explicitly prohibited. However, **students are allowed, and encouraged**, to help each other understand programming concepts, errors, and how to approach problems.

Week 0 — Introduction

*Session 1. What is Computational Social Science?*

What defines computational social science? Where does it owe its legacy and what does the future hold? What does this mean for the type of problems that we can now study and what are the inherent drawbacks?

*Session 2. Digital trace data* - APIs

What is digital trace data and where is it The differences between event tracking and provider APIs (Reddit, Twitter). Are there easy ways to get data from the web. Breaking down what API actually means. Defining JSON and understanding API documentation. Tokens and credentials.

*Reference:*

CA Davis, O Varol, E Ferrara, A Flammini, F Menczer. (2016) BotOrNot: A system to evaluate social bots. Proceedings of the 25th International Conference Companion on World Wide Web

Week 1 — The structure of the web

*Session 1. The structure of the web*

Defining the components of the web (HTML, CSS, JS) and the basics of how modern data is captured. The developer console and how to identify structural components of websites.

*Session 2. Processing web pages*

Defining regular expressions and understanding how to process web pages based on structural elements. Identifying individual biographies from Wikipedia.

Week 2 — Trace data on the web

*Session 1. Web Scraping*

Scraping content from the web and the difference between static and dynamic content. Downloading static content with splinter and an introduction to dynamic content scraping with selenium.

*Reference:*

Malmgren RD, Ottino JM, Amaral LAN. (2010). The role of mentorship on protégé performance. *Nature* 463, 622-626.

*Session 2. Web Crawlers*

Defining a web crawler and the differences between crawling for search and research. Sampling procedures and design, as well as understanding what “load” means and how targets change that. Instagram as an example of the value of crawling versus the walled garden.

Week 3 — Structuring unstructured data

*Session 1. Text processing*

Natural Language Toolkit as the key. Processing, stop words, tokenization and n-grams.

Parsing text in complex documents. Understanding word usage and frequency.

*Reading:*

Gerlach M, Shi H, and Amaral LAN. (2019). A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence: (1) 606-612.*

*Session 2. Information theory and change*

How can we understand the evolution of a written system over time? How can we account for the intricacies of language and still quantify the complexity of textual usage?

*Reading:*

*Murdock J, Allen C, DeDeo S.* (2016) Exploration and exploitation of Victorian science in Darwin’s reading notebooks. Cognition (159) 117-126.

Week 4 — Sentiment and emotion

*Session 1. Basics of sentiment analysis*

Where does positivity or negativity of words come from? How do we use this to understand the emotion of writers and characters.

*Reading:*

Dodds PS , Harris KD, Kloumann IM, Bliss CA, Danforth CM, and. (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. https://arxiv.org/abs/1101.5120

*Session 2. Learned sentiment*

How do we learn sentiment for a specific corpus? Bayes rule and naïve bayes as a classifier. Negation, sarcasm, and the limits of quantitative analysis (and why it typically will wash out).

*Reading*

R. Gonzalez-Ibanez, S. Muresan, and N. Wacholder. (2011). Identifying sarcasm in Twitter: a closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: 581-586.

Week 5 — Mapping document clusters

*Session 1. Topic maps*

How do we categorize text and documents? What are higher level groupings that we can use to convey meaning. Introduction to Latent Dirichlet Allocation and its associated hyper-parameters.

*Reference:*

Blei DM, Ng AY, Jordan MI. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research: (3)993-1022.

*Session 2. Dirichlet priors and unsupervised dangers*

How do we evaluate and test the performance of an unsupervised solution? Understanding the dangers in text analytics.

*Reference:*

A. Lancichinetti, M. I. Sirer, J. X. Wang, D. Acuna, K. Kording, and L.A.N. Amaral. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X* 5, 011007. [**arXiv:1310.4546**](https://arxiv.org/abs/1310.4546)

Week 6 — Text vectors and concept universality

*Session 1. Vectorization of text*

What is the underpinning of word2vec and doc2vec? How can they be used and how do they differ from the groups that we may identify through LDA?

*Reference:*

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. (2013). Distributed representations of words and phrases and their compositionality.

*Session 2. Concept universality*

How far does word2vec and doc2vec go in terms of concept similarity? What are the dangers and how should these be handled when making inferences?

*Reference:*

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Week 7 — Networks

*Session 1. Networks*

Defining what is a network. Nodes, degrees, and traversal. Weighted vs. unweighted and bipartite networks. Movement patterns of Divvy riders.

*Session 2. Information Networks*

How do we analyze networks of information? What do these connections say about human behavior? Integrating metadata to expand analyses.

*References:*

Evans JA, Foster JG. (2011) Metaknowledge. *Science*. 331, 6018:721-725.

Week 8 — Social Networks

*Session 1. Social Networks*

On-line social networks and their use to understand changing human behavior. Small worlds, path lengths, clustering, and power laws. Contagion and homophily. Attempts to control for selection bias.   
*References:*

D.J. Watts and S.H. Strogatz. (1998). Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440-442.

N.A. Christakis and J.H. Fowler. (2007) The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* **357**, 370-379.

*Session 2. Social systems over time*

How does time influence behavior? How do we deal with this – defining multiple approaches: static, identifying the appropriate unit of time, and temporal statistics.

*References:*

J. Saramaki and E. Moro. From seconds to months: multi-scale dynamics of mobile telephone calls. https://arxiv.org/pdf/1504.01479.pdf

Week 9 — Communities and Significance

*Session 1. Community detection*

What are the basics of identifying groups in a network? What do these groups mean?

What are the different methods? Modularity vs. Information theoretics vs. Statistical fit.

M Rosvall and CT Bergstrom. (2008). Maps of random walks on complex networks reveal community structure. *PNAS* 105(4):1118-1123.

A. Lancichinetti, F. Radicchi, J.J. Ramasco and S. Fortunato. (2011) Finding statistically significant communities in networks. *PLoS ONE* **6**, e18961.

*Session 2. Null models, bootstrapping, and the significance of position*

What does a null model truly mean? What is bootstrapping? What are we truly estimating when we bootstrap? How do we test significance?