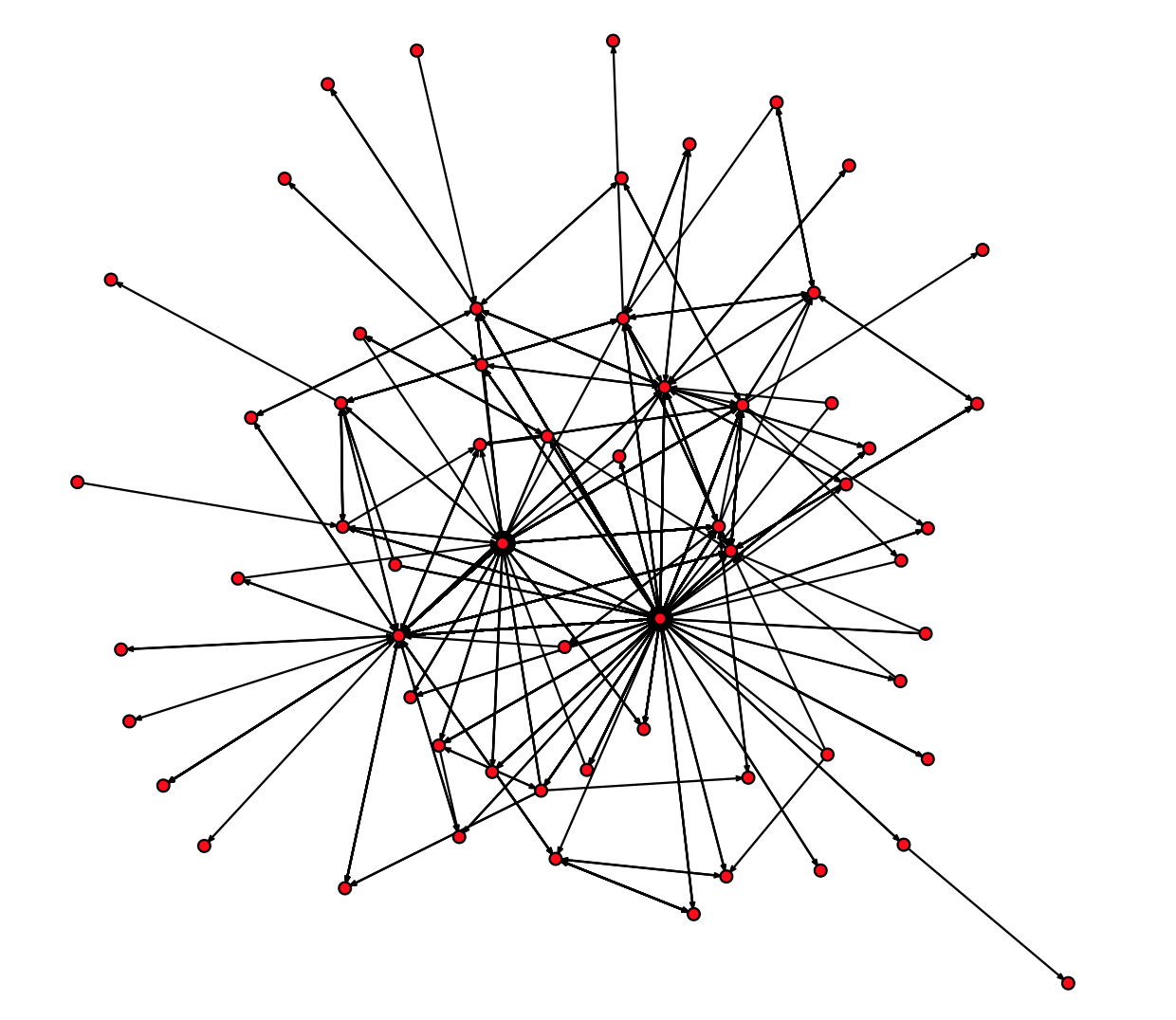
INFO 480 – Introduction to Data Science

Macintosh HD:Users:sean:SeanDropbox:Dropbox:00.  Writing:136. Github Longitudanal:data:viz:short-names_Betweenness.pdf 

# Drexel University

College of Information Science and Technology

Spring, 2013

Professor: Sean P. Goggins, Ph.D.

Office: Rush 405b

Phone: 215 948 2729

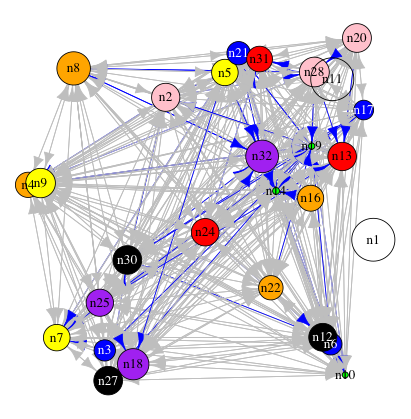
Email: [spg46@drexel.edu](mailto:spg46@drexel.edu)

Mailbox Location: Third floor of RUSH Hall

Instructor Websites: <http://www.groupinformatics.org> <http://www.socialmediainformatics.org> <http://www.seangoggins.net>

Note: Feel free to email me with questions at any time.

Course Website – <http://www.seangoggins.net/info480>



Version

Two of

Syllabus

# Course Description

## Catalog Description

Introduces students to “Data Science”. Develops skills required for data intensive work. Students will deliver data science products and services through analysis, data transformation and data access techniques. The assignments will involve web programming, statistics, and the ability to manipulate data sets with code, following examples provided.

## Pre-Requisites

Sophomore standing. Students will develop the necessary programming skills for the course by using worked examples, peer review and materials prepared by the instructor.

## Curriculum Role

Course is an elective in the undergraduate BSIS, and BSIT programs. It is taken by students interested in designing and delivering analysis of “Big Data” from public, social media and organizational sources. It can also be taken as an elective by undergraduate students interested in acquiring a basic knowledge of the *practice of big data analysis* and the social implications for privacy in society.

## Course Rationale

This course is offered to prepare students to lead the design and configuration of data analysis in the workplace using contemporary tools. Such tools enable unprecedented mechanisms for making sense of organizational behavior and enhancing organizational performance by making invisible activity more visible. The value of this course for students is the experience they will gain reading about, discussing and designing data analysis applications using real data. Such data analysis is increasingly designed, configured and implemented by non-technical end users. This course complements INFO 215 with hands on application of and reflection on the principles described as *social impacts of computing* in INFO 215.

## Course Outcomes

Upon successful completion of this course, a student will be able to:

* Apply their knowledge of an organizational context and indications of data quality to process data prior to analysis into a usable analytical form.
* Identify ways to present demographic, interaction, and other types of data in a manner that tells a visual story.
* Design data products to support business or social objectives.
* Produce observations about the relationship between data, its processing and visualizations and a specific set of institutionally relevant questions.
* Evaluate observations, visualizations and experiments with data to iteratively develop an increasingly useful and relevant analytical toolset for a particular set of data.
* Identify the privacy implications of research and analysis of data from social media and other public sources where individual identity is revealed.

## Required Materials

All required readings are listed in the course schedule. Each reading will be made available through the Hagerty library. Where materials are open source or open published, they will be available on the course website.

You will also need to purchase a copy of the following textbooks:

* Janert, PK. (2011) *Data Analysis with Open Source Tools*. New York: O’Reilly.
* Vaidhyanathan, S. (2005) *The Anarchist in the Library: How the Clash Between Freedom and Control is Hacking the Real World and Crashing the System.* New York: Basic Books.

## Course Content

Principal topics and the approximate number of weeks devoted to each are:

• Introduction to Data Science (1)

• Data Preparation (1)

• Data Preparation Tools (2)

• Data Presentation & Tools (2)

• Building Data Products (1)

• Data Analysis and Observations (1)

• Conducting Data Experiments (1)

• Final Project Preparation and Presentation (1)

## Presentation

Teaching methods will include hands on activities and discussion. Lecture will be limited to the minimum necessary to contextualize class activities with course readings and learning outcomes.

## Assessment

Evaluation will be based on successful completion of each course section’s evaluative assignment. Data collection, analysis, written papers, projects and class participation will contribute to the final grade.

## Course Materials

*Note: Course materials will change frequently, as this is an evolving field of inquiry and application*.

Required:

* Janert, PK. (2011) *Data Analysis with Open Source Tools*. New York: O’Reilly.
* Vaidhyanathan, S. (2005) *The Anarchist in the Library: How the Clash Between Freedom and Control is Hacking the Real World and Crashing the System.* New York: Basic Books.
* Additional readings as assigned.

## Course Management

Course Coordinator: Sean P. Goggins

# Student Responsibilities

The standard amount of work time expected for a Drexel undergraduate course is three hours of class time per week **plus** nine hours of reading and other coursework per week. This means that you should expect to spend an average of 12 hours per week working on this class.

Since meeting deadlines is an important professional responsibility, grades on late work will be lowered one full letter. If students make prior arrangements with the instructor lateness may be cleared in advance. As a professional, the ability to work successfully with your colleagues will be vital to your career. Thus, class participation and interaction with your fellow students are important components of this course.

## Accommodating Documented Disabilities

Any student with a documented disability who needs accommodations should contact the Office of Disability Services (ODS) at 215-895-2506, or visit their website at http://www.drexel.edu/ods/index.html. Students must register with the ODS and receive an Accommodation Certification Form prior to receiving accommodations. All ODS communication is strictly confidential.

## Academic Integrity

Finally, any incidence of plagiarism or other academic dishonesty will result in an automatic **F** for the course. Links to Drexel’s official policies are below:

<http://www.drexel.edu/provost/policies/academic_dishonesty.asp>

http://www.drexel.edu/studentlife/judicial/honesty.html

# Course Schedule

The over riding teaching objective is student mastery of the material, and this may not follow the linear path articulated in this document.

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Date** | **Online Activity** | **In Class Activity**  *(Assignments In Brackets)* |
| Week 1 – Introduction | April 2, 2013 |  | Opening lecture, sample data, review syllabus, Tools overview |
| Week 2 – Data Prep | April 9, 2013 | GitHub | No class session |
| Week 3 – Data Prep Tools | April 16, 2013 | GitHub | Single/Multiple Variable Distribution |
| Week 4 – Sharing Data Prep Tools | April 23, 2013 | GitHub | Multivariate Graphical Data Analysis ***(A1)*** |
| Week 5 – Data Presentation Tools | April 30, 2013 | GitHub | Data Analysis Session ***(A2)*** |
| Week 6 – Sharing Data Prep Results and Tools | May 7, 2013 | GitHub | Probability Estimation |
| Week 7 – Building Data Products | May 14, 2013 | GitHub | Classical Statistics and Clustering ***(A3)*** |
| Week 8 – Data Analysis and Observations for Social Insight | May 21, 2013 | GitHub | Guest Lecture from a local data scientist ***(A4)*** |
| Week 9 – Conducting Data Experiments | May 28, 2013 | GitHub | Financial Calculations and Modeling ***(A5)*** |
| Week 10 – Final Projects | June 4, 2013 | GitHub | Student Presentations ***(A6)*** |
| Exam Week | June 14, 2013 | Github | No class session ***(A7)*** |

Assignments are listed in **BOLD** text. Readings are not bolded.

## Week 1: Introduction to Data Science

General introduction and explanation of syllabus. Example data science activity using Github data. Demonstration of data collection and management, analysis and visualization lifecycle. Privacy implications of analyzing public data discussed. The social value of analyzing “big data”: How we learn about social and organizational performance and structure using “big data”.

**Readings Due:**

1. Mascaro, C., Magee, R., and Goggins, S. 2012. Not Just a Wink and a Smile: An Analysis of User-Defined Success in Online Dating. iConference 2012.
2. Howison, J., Wiggins, A., and Crowston, K. 2012. Validity Issues in the Use of Social Network Analysis with Digital Trace Data. Journal of the Association of Information Systems. 12(2), 2.
3. Goggins, S. P., Mascaro, C., and Valetto, G. 2013. Group Informatics: A Methodological Approach and Ontology for Understanding Socio-Technical Groups. JASIS&T. Accepted.

## Week 2: Data Preparation

Preparing Data: Finding raw, public or organizational data and defining what questions can be answered using the organization’s data. I will provide you with sets of data and specific analyses desired. You will need to analyze the data and decide how it needs to be “reshaped” in order to perform the desired analysis. We will then discuss these analyses on GitHub; I will provide similar examples and ask you to develop and prepare the data further. You can use any software tool you are familiar with, but I will be providing examples and help with Python. Choose your own tools and you are “on your own” to some extent. ☺ … but tools are covered more fully in week 3. I do encourage you to go through some Python tutorials during week’s 1 & 2 --- <http://docs.python.org/2/tutorial/> (we will be using Python 2.7.3 in this course; Python 3 is another discussion.)

**Readings Due:**

1. Part One of “Data Analysis with Open Source Tools”.
2. Chapters 1-3 of “The Anarchist in the Library: How the Clash Between Freedom and Control is Hacking the Real World and Crashing the System”.

## Week 3: Data Preparation Tools

Preparing data using open source tools. Traditional and more advanced, statistical data analysis tools will be used.

**Readings and Assignments Due:**

1. Sawyer, S. and Crowston, K. 2004. Information systems in organizations and society: Speculating on the next 25 years of research. Information systems research. 35-52.
2. Chapter 4 of “The Anarchist in the Library: How the Clash Between Freedom and Control is Hacking the Real World and Crashing the System”.

## Week 4: Sharing Data Preparation Tools

Students will become familiar with the use of open source software configuration management tools for sharing software developed to prepare open data. The specific environments used may change as open source software evolves. Github, Google Code and Source Forge are prominent SCM exemplars.

**Readings and Assignments Due:**

1. **Prepared Data #1 (Transform an instructor supplied data set into an analyzable form using open source tools)**
2. Read chapters 1-3 of “Pro Git” <http://git-scm.com/book> (Or other, repository specific orientation as selected by the instructor and course coordinator)
3. Part Two of “Data Analysis with Open Source Tools”.

## Week 5: Data Presentation Tools

This session will focus on sharing what we have learned so far, and be focused on lab interactions.

**Readings and Assignments Due:**

1. **Software Sharing #1 (Share scripts produced in week 3 using an open source software configuration management tool)**.
2. Part Three of “Data Analysis with Open Source Tools”.

## Week 6: Sharing Data Preparation Results and Tools

**Readings and Assignments Due:**

Presentation involves sharing data with other people in a way that is visually insightful. Students will be asked to bring an example of a visualization of data from a website or news organization, and make a short presentation about what makes the visualization insightful.

1. **Data Visualization Example Presentation**
2. Chapters 4-7 of “The Anarchist in the Library: How the Clash Between Freedom and Control is Hacking the Real World and Crashing the System”.

## Week 7: Building Data Products

The full lifecycle of data preparation and presentation of results. Students will use sample data provided by the instructor.

Planning for a data product that is derived from open data students collect, manage, transform, analyze and visualize.

**Readings Due:**

1. **Data Product One**
2. Part Four of “Data Analysis with Open Source Tools”.
3. Chapters 8 and 9 of “The Anarchist in the Library: How the Clash Between Freedom and Control is Hacking the Real World and Crashing the System”.

## Week 8: Data Analysis and Observations for Social Insight

Data analysis as an exploratory process. Types of exploratory analysis for social information.

**Readings Due:**

1. Goggins, S., Valetto, G., Mascaro, C., and Blincoe, K. 2012. Creating A model of the Dynamics of Socio-Technical Groups using Electronic Trace Data. User Modeling and User-Adapted Interaction: The Journal of Personalization Research. Accepted.
2. Chapters 10-12 of “The Anarchist in the Library: How the Clash Between Freedom and Control is Hacking the Real World and Crashing the System”.

## Week 9: Conducting Data Experiments

Students present an exploratory analysis of the data set chosen for their final projects and discuss the analysis in small, in class groups.

**Assignments Due:**

1. **Exploratory analysis of final project data and final project proposal**

## Week 10: Final Project Preparation and Presentation

Present final projects, and evaluate the resulting data products.

**Assignments Due:**

1. **Final project presentation**

## Finals Week: Open Sourcing Data Products and Tools

Students will refine their final projects based on feedback received from their peers and the instructor. The finished public data sets, transformation scripts, data profiles, visualization and explanation of analyses will be deployed in an open source repository.

**Readings and Assignments Due:**

1. **Open source distribution of final data products**

# Assignments

1. **Prepared Data #1 (Transform an instructor supplied data set into an analyzable form using open source tools)**
   1. Instructor supplies a public data set for analysis, along with a set of questions or insights that are obtainable from the data set
   2. Students use the data and what they have learned in the first weeks of the course to prepare the data for analysis
   3. Included in the assignment is a 500 word explanation of how the preparation of the data alters the original data and a description of how the questions can be answered using the data.
2. **Software Sharing #1 (Share scripts produced in week 3 using an open source software configuration management tool)**.
   1. Students will refine and then share their scripts with other students
   2. Included in the assignment is a 500 word explanation of how their script could be improved, optimized and adapted to other data of a similar type.
   3. The “read me” file distributed with the script will explain to another user how to apply the script to the data distributed in assignment one. This will include specific, technical specifications.
3. **Data Visualization Example Presentation** 
   1. Students will select a visualization from the public media explaining an event, social, political or financial phenomena of their choosing.
   2. The presentation will include a powerpoint slide or poster scale copy of the visualization referenced.
   3. The explanation presented by students will include:
      1. A description of what interests of theirs drew their attention to this visualization
      2. A description of the source data the visualization is drawn from
      3. Two critical observations about how the source data might have been visualized to tell a different story
      4. A characterization of the visualization as informative or persuasive, and a description of how the visualization influenced their knowledge or position.
4. **Data Product One**
   1. Students will apply the transformed data and scripts from assignments one and two to develop a visualization of the data.
   2. The assignment will include a full “data product” package, with refinements of the scripts from assignment one and two, and a “readme” file that describes how the visualization can be generated.
5. **Exploratory analysis of final project data and Final Project Proposal**
   1. Students will select a set of publicly available data from a list provided by the instructor. Candidate data sources include Facebook, Twitter, Government Open Data and other data that is in the public domain.
   2. Students will conduct an exploratory analysis of the data, including building scripts for analysis, and preliminary results
   3. Based on the outcomes of the exploratory analysis, students will prepare a final project proposal including the following areas:
      1. Description of data set
      2. Description of proposed analysis, including questions to “ask” of the data
      3. A characterization of the data project as persuasive or informative
      4. A plan for when the specific steps in the final project will be completed
      5. At least one hypothesis about what they expect to learn from the data analysis
6. **Final project presentation**
   1. Students will present the visualizations and “findings” from their work on the final project
   2. Presentations will include
      1. Three alternate visualizations of the data
      2. Explanations of each visualization
      3. A story explaining how the data analysis presented explains the underlying data, creates insight and is persuasive.
   3. Students will peer review each other’s presentations and be asked to rank them.
7. **Open source distribution of final data products**
   1. Students will package their final projects including:
      1. Links or copies of the open data used in the project
      2. Descriptions of the data structure
      3. Transformation scripts and descriptions of how to use them
      4. Visusalization scripts or web pages and descriptions of how to use them
      5. A 500 word description of what the package is for, what it does and in what other data contexts it might be useful
   2. Students will upload these projects to an open source code sharing site specified by the instructor

# Assessment

* (10%) Prepared Data #1 (Transform an instructor supplied data set into an analyzable form using open source tools)
* (15%) Software Sharing #1 (Share scripts produced in week 3 using an open source software configuration management tool).
* (10%) Data Visualization Example Presentation
* (10%) Data Product One
* (10%) Exploratory analysis of final project data and Final Project Proposal
* (10%) Final project presentation
* (20%) Open source distribution of final data products
* (15%) Class Participation

# Grades

Each assignment will have a total of 100 points. Grades correspond with points in the following manner:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Grade*** | **A+** | **A** | **A-** | **B+** | **B** | **B-** | **C+** | **C** | **C-** | **D+** | **D** | **F** |
| *% Range* | 96 - 100 | 90 - 95.95 | 88 - 89.95 | 85 - 87.95 | 79 - 84.95 | 76 - 78.95 | 73 - 75.95 | 63 - 72.95 | 60 - 62.95 | 57 - 59.95 | 44 - 56.95 | 0 - 43.95 |

# For each assignment, the following is required

I certify that:

·         This paper/project/exam is entirely my own work.

·         I have not quoted the words of any other person from a printed source or a website without indicating what has been quoted and providing an appropriate citation.

·         I have not submitted this paper / project to satisfy the requirements of any other course.

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date       \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 

# Classroom/Lecture Recording

Lectures and class discussions may be recorded and/or streamed and rebroadcast for educational purposes only.Relevant University Policies

Academic Dishonesty Policy: http://www.drexel.edu/provost/policies/academic\_dishonesty.asp

Academic Integrity Policy: http://www.drexel.edu/studentlife/judicial/honesty.html

Course Drop Policy: http://www.drexel.edu/provost/policies/course\_drop.asp

**(NOTE:** Any changes to this syllabus will be communicated to all registered students via email.)