Fall 2017 ***Big/Small Data & Visualization*** Soc/Ling/QSS 446W

Emory University

Roberto Franzosi

Office Room No. 212 Tarbutton Hall
Email rfranzo@emory.edu

Lectures Tu-Th 4:00-5:15 Tarbutton Hall 111
Office Hours Tu–Th 12:30–2:15 or by appointment
(please, use email for contacts)

TA Craig Alder
Email calder@emory.edu

Undergraduate TA Marissa Adams
Email maadam5@emory.edu

COURSE OBJECTIVES

The course deals with new tools of data analysis and visualization, especially for text data. Many of these tools have been developed in conjunction with new technologies of data mining and extraction from large text corpora made available on the web. It is these huge amounts of (mostly textual) data that offer both humanities and social sciences new avenues of research in the form of digital humanities, and where different types of data can be pulled together on a topic and displayed on the internet in very creative ways. The course will illustrate some of the cutting-edge projects in digital humanities such as David Eltis's *Trans-Atlantic Slave Trade* website (http://www.slavevoyages.org), Hank Klibanoff's *Georgia Civil Rights Cold Cases Project* (https://scholarblogs.emory.edu/emorycoldcases.

The course will show how to use different tools of data visualization, especially network graphs dealing with relationships between objects (social actors, concepts, or just words), both static and dynamic (changing with time), and spatial maps dealing with objects in space (and time, dynamic maps) through Geographic Information System (GIS) tools. We will focus on freeware software, from *Gephi* to *Cytoscape*, *Palladio*, *Google Earth Pro*, *QGIS*, *Carto*, *TimeMapper*, *Google Fusion Tables*.

Using Natural Language Processing (NLP) tools (based on the Stanford parser *CoreNLP*) the course will show how to analyze large corpora of text and how to visualize the information extracted, through *Excel* charts or *Wordle*, *Tagcrowd*, *Voyant* displays of word frequencies and network graphs of word co-occurrences in *Gephi*. Other NLP tools and the visualization of the information they make available will also be introduced, from topic modeling (*Mallet*, *Stanford Topic Modeling Toolbox*) to *Word2Vec* (vectors representations of words, shown to capture many linguistic regularities of a corpus). The properties of such tools as Google ngrams and Bookworm will be explored. The peculiarities of dealing with words as data will be discussed

(where and how can you get corpus data? How can you convert images to text? How can you check for spelling errors, differences in documents?).

Beyond the technical aspects of data visualization, the course addresses broader questions about the impact of big data on scholarly practice. What is the relationship between macro and micro? Does it still make sense to talk about statistical outliers and their role when millions of data points (words) are now used? Are the new forms of data visualization simply descriptive? What happened to social sciences' central concern with hypothesis testing? If color, form, movement, in Kandinsky's view, are the distinctive weapons of art (and beauty), are the new visualization techniques – all based on color, shape, and movement –a game changer in the traditional ways of displaying evidence (i.e., a table of numeric estimate values)? Does this offer a rapprochement between the humanities and science, in approaches, in techniques, perhaps even in modes of writing?

Learning outcomes

By the end of term, students are expected to be able:

1. Become familiar with NLP tools
2. To deal with large bodies of text ("corpus") with NLP tools
3. Become familiar with a large number of data visualization tools
4. To make public presentations before an audience
5. To write a research report

COURSE OUTLINE

Introducing the course

Week 1: Big data and "distant reading"
Week 1: Preparing your corpus for "distant reading"

Part I: NLP (Natural Language Processing): Basic language

Week 2: NLP (Natural Language Processing): Sentence splitting, tokenization, lemmatization (lemmas and stems)
Week 2: The CoNLL table: What's in the CoNLL table?
Week 3: Word co-occurrences (KWIC)
Week 3: Basic corpus statistics

Part II: What's in your corpus? Word searches and visualization

Week 4: Searches by words' grammatical roles (POSTAG, Part Of Speech Tag)
Week 4: Searches by syntactic relations between words (DEPREL DEPendency RELations)
Week 5: Word searches in context (KWIC, Kew Word In Context)
Week 5: N-grams searches
Week 5: Word co-occurrences searches

Part III: Dissecting your corpus

Week 6: Noun density
Week 6: Verb modality: Ability, possibility, permission, and obligation
Week 6: Verb tense: past, future, gerundive
Week 6: Verb voice: Active and passive verb forms
Week 7: Sentiment Analysis: Conveying negative, neutral, and positive feelings through text
Week 8: Text readability: What grade level does a text require to be comprehensible?
Week 8: Sentence complexity: Using CLAS (Computerized Linguistic Analysis System) to measure linguistic complexity

Part IV: Topic modeling

Week 9: Using Mallet to tell us what different documents in a text corpus talk about

Part V: Visualizing words

Week 10: Word2Vec: Neural network approaches to word relations
Week 11: Word clouds

Part VI: Extracting SVOs

Week 12: Coreference resolution
Week 12: Anaphora resolution
Week 12: ClausIE and automatic SVO extraction

Part VII: Extracting and visualizing time and space

Week 13: Using NER information (Back to the CoNLL table)
Week 13: Geocoding
Week 14: Thanksgiving!

Week 15: Visualizing space
Week 15: Visualizing time and space

Part VIII: What have we learned?

Week 16: A game changer? Digital humanities (or Digital scholarship), beautiful evidence, and visual rhetoric

DEADLINES AND IMPORTANT DATES

First day of class	August 23
Fall break	October 9-10
Thanksgiving	November 22-24
Last day of class	December 5

COURSE REQUIREMENTS

The course requires students:

1. to carry out **homework**;
2. to make individual or group **presentations on specific readings;**
3. to make individual or group **presentations on specific software** and edit/write appropriate **TIPS files**; you would get **extra points if you prepare screen recordings** on how PC-ACE connects to the software you are presenting (you could use http://screencast-o-matic.com/home for this purpose).

COURSE PREREQUISITES

There are no formal prerequisites for the course, except for a general GOOD familiarity with (and lack of fears of) computers. If you do have a computer science background, of course, you will be able to do more and get more out of the course.

FULLFILLING THE WRITING REQUIREMENTS

1. The weekly assignments, by and large, consist of analyzing text corpora. These texts are mostly chosen by the individual students, except for some assigned texts (e.g., Faulkner's short story "Dry September"). Each week, students are expected to analyze their corpora using different Natural Language Processing (NLP) tools and to write up the results of their analyses, submitting their work in the form of a Word document. This document will include figures with the results of the NLP analyses (typically, screenshots of computer output) and the students' interpretations and explanations of these figures. What do the results mean? What do they tell you about the substance of the texts? What are the limits of the tools used? On average, 2 to 5 pages of writing are expected every week. But the amount of writing is expected to increase week after week as students return to the same texts using different approaches and tools, ultimately incorporating all of their analyses into one document as they approach submission of the final paper.

Each assignment is graded and comments are provided. The standards of writing are repeatedly explained in class and stressed in the comments given to students.

2. As the course deals mostly with automatic processing of texts, the issue of writing and style are implicitly at the core of the course: which verb voices are used (active or passive), which level of sentence complexity (as measured by different indices of sentence complexity), which semantic roles (e.g., agent and patient, experiencer, benefactor and beneficiary, messenger and receiver), which attributes (e.g., adjectives or adverbs) in conjunction with different nouns and verbs, which sentiments are expressed in sentences (negative, neutral or positive), which combination of description, action, and evaluation is used in the unfolding of a story. Teaching writing is then a fundamental part of analyzing writing. The pros and cons of pure automatic analyses of texts ("distant reading" through a computer) are constantly brought up, with an emphasis of a constant dialogue between distant reading and close reading.

SOFTWARE REQUIREMENTS

Several different software programs will be used in the course. The core program will be the freeware PC-ACE (Program for Computer-Assisted Coding of Events) (www.pc-ace.com**). PC-ACE setup package with NLP component will include the Stanford CoreNLP, Mallet, Word2Vec, KWIC and several other NLP and data visualization options.**

PC-ACE only works under Windows operating system. If you are using an Apple laptop, you will be expected to install Windows on your laptop. There are a couple of approaches to this, via Apple Boot Camp, Virtualbox, or Parralels. Unfortunately, many of the applications we will use are Windows based. Both Apple Boot Camp and Virtualbox are free.

Boot camp: you can download it for **free** and install it at
https://www.apple.com/support/bootcamp/ **For Boot camp to work your Mac must have an Intel processor. But Apple transitioned to Intel processors in the mid-2000s so you should be OK. But… please check your processor!**

Virtualbox: you can download it for **free** for Mac OS X from here:
http://www.oracle.com/technetwork/server-storage/virtualbox/downloads/index.html
Then follow this set of instructions.

1. Install Virtualbox you downloaded,
2. Download the Windows 10 or 8.1 ISO from Microsoft: https://www.microsoft.com/en-gb/software-download/windows10ISO
3. Open Virtualbox and select New (upper left hand corner)
4. Name the OS whatever you want and select Windows and whatever version of Windows you downloaded in step 3. Select continue.
5. Allocate 2048 - 4096 MB of RAM depending on how much RAM your Mac has and select continue.
6. Select create a virtual hard drive now and create.
7. Select VDI and continue.
8. Select dynamically allocated and continue.

9. Set the location of the virtual drive to wherever you want and allocate about 35 GB to the virtual drive. Continue.
10. Now the setup process is finished and the machine can be run.
11. Try double clicking the machine from the side bar. It will start to run and then request a bootable .iso; this is the ISO file you downloaded earlier from Microsoft's site. Point the machine to the ISO file.
12. The Windows setup process should begin. Be sure to either have a product key or a to do trial setup.

For instructions see also http://www.intowindows.com/how-to-boot-and-install-from-iso-in-virtualbox/

Parallels: you can purchase a student license for around **$40** (and volume pricing is also available for further discount) from the Emory website http://it.emory.edu/software/, then click on Academic Software (for Personal Purchases) which will redirect to http://www.academicsuperstore.com/ Parallels at http://www.academicsuperstore.com/product/search?qk_srch=parallel&x=0&y=0

Whether you choose the freeware option of Boot camp or Virtualbox (the best options!) or you purchase a license for Parallels, you will need to purchase a copy of **Microsoft Windows (Windows 10)** (volume pricing is also available for discount) from the Emory website http://it.emory.edu/software/, then click on Academic Software (for Personal Purchases) which will redirect to http://www.academicsuperstore.com/ **This is the only expense you are required to sustain since all readings will be placed on reserve.**

But… it is also possible to get a freeware copy of a Windows Beta version. Click on the main Microsoft Windows link https://insider.windows.com. Then click on "get started". On the next page you might have to create a Microsoft account if you do not have one already. Once the account is created you will be directed to the download link. Remember to download this on a blank drive. Since it is a free and beta version you will not be able to personalize the operating system once installed. However, it runs perfectly well without personalization. Microsoft office can then be downloaded from the free software available in the settings section of your email on outlook.office.com This is a free and fully trustworthy version.

Both Boot Camp/Virtualbox and Parallels Desktop allow you to run Windows on a Mac. One drawback of Boot Camp is that you have to reboot your Mac every time you want to switch between Mac and Windows. If this is not too often, it is a cheaper solution, and perhaps more reliable, than Parallels.

To prepare for installation of Boot Camp, Virtualbox, or Parallels you should have:

- A backup of all your data (Time Machine is preferred) - STS can assist with this
- The latest version of Mac OS (free upgrade available on Mac App Store)
- A copy of the latest version of Windows OS
- At least an 8GB Flash Drive

Once you have all of these, please visit the Student Technology Support (STS) desk (http://it.emory.edu/sts) on the 1st Floor of Woodruff Library. STS will provide consultation and assistance with these installations which typically take **1 – 3 hours** to complete.

Please, download the following freeware software that you will be using for textual analysis and data visualization:

1. **PC-ACE (Program for Computer-Assisted Coding of Events) (**https://pc-ace.com/download/**) the setup package with NLP components will include the Stanford CoreNLP, Mallet, Word2Vec, KWIC and several data visualization options**
2. **TACIT** http://tacit.usc.edu/download.html
3. **WordNet** https://wordnet.princeton.edu/wordnet/download/
4. **Gephi** http://gephi.org/users/download/
5. **Cytoscape** http://www.cytoscape.org/download.php
6. **Tableau** https://public.tableau.com/s/
7. **QGIS** https://www.qgis.org/en/site/forusers/download.html
8. **Google Earth Pro** http://www.google.com/earth/download/gep/agree.html
9. **Mondrian** http://www.theusrus.de/Mondrian/ (bottom of page, under downloads)
10. **OpenRefine (former Google Refine)** http://openrefine.org/download.html

Stanford CoreNLP (http://nlp.stanford.edu/software/corenlp.shtml) and **Mallet (**http://mallet.cs.umass.edu/download.php**) will be installed directly by PC-ACE.**

In addition, we will be using the following web-based data visualization software:

1. **Voyant** http://voyant-tools.org/
2. **Google Fusion Tables** https://support.google.com/fusiontables/answer/2571232
3. **Raw** http://raw.densitydesign.org/
4. **TimeMapper** http://timemapper.okfnlabs.org/
5. **Carto** https://Carto.com/
6. **Palladio** http://palladio.designhumanities.org/#/
7. **Google ngrams** https://books.google.com/ngrams
8. **Bookworm** http://bookworm.culturomics.org/
9. **Wordle** http://www.wordle.net/create
10. **TagCrowd** http://tagcrowd.com/
11. **Tagul** https://tagul.com/
12. **Tagxedo** http://www.tagxedo.com/
13. **Wordclouds** http://www.wordclouds.com/

CORPUS REQUIREMENTS

You should also download a set of texts of interest to you that you will want to analyze with NLP tools and visualize using the variety of data visualization tools illustrated in the course. These texts should be in txt format (not doc, pdf, or other since NLP tools only work with txt formats); they could be:

1. **tweets**
2. **blogs**
3. **newspaper articles**
4. **US Congress bills** (https://www.congress.gov/)
5. **US presidential speeches** (http://www.presidency.ucsb.edu/data.php)
6. **corporate/university mission statements**
7. **social science & history qualitative data**; see the US academic data depository of ICPSR of the University of Michigan (http://www.icpsr.umich.edu/index.html) or the British equivalent of the UK Data Service (https://www.ukdataservice.ac.uk/); the collection at Qualitative Data Repository (https://qdr.syr.edu/deposit), the Murray Research Archive at IQSS Harvard University* (http://murray.harvard.edu/dataverse)
8. **oral history archives**; see the list provided by the Oral History Association, (http://www.oralhistory.org/centers-and-collections/)
9. **transcribed in-depth interviews**
10. **social science journal abstracts** (http://ssrn.com/en/)
11. **NYT book reviews;** see the NYT API (http://developer.nytimes.com/docs/books_api/)
12. **song lyrics**; see, for example, the collection provided by AZlyrics (http://www.azlyrics.com/a/archive.html)
13. **books**; see the free collections at Open Library (https://openlibrary.org/) or at Hathi Trust Digital Library (https://www.hathitrust.org/); many older books are also available in Google Books (https://books.google.com/) and in other archives (e.g., The Gutenberg Project https://www.gutenberg.org/, Internet Archive https://archive.org/ , The OAIster database http://www.oclc.org/oaister.en.html)
14. **diaries & autobiographies**
15. **letters (epistolary)**
16. **folktales (e.g., Afanasiev's collection of Russian folktales analyzed by Propp)**

Make sure you check the data in your corpus.

1. **To repeat… you can only use txt-formatted files (NLP tools only work with txt files in input); use the tools found in PC-ACE to prepare your files.**
2. **Remove tables of contents, indices, weirdly formatted footnotes/endnotes, headers/footers, tables and figures. This material is not handled correctly by NLP tools.**

Web scraping. If you are obtaining your corpus from the web, you can copy and paste documents, perhaps from different websites. However, **web scraping** may provide a more efficient solution. Web scraping is the process of automatically collecting information from the World Wide Web through specialized software programs. A good, **freeware** option is **OutWit Hub**. While the full version of OutWith Hub costs around $89, the freeware option will probably serve you well. You can download it at http://www.outwit.com/products/hub/. Another good freeware option is HTTrack (https://www.httrack.com/). Scraping requires knowledge of the data structure of each website where data are taken from. Scraping will be more efficient than human copy-and-paste if the documents to be scraped are stored under the same website (so that knowledge of only one type of data structure is required); otherwise, you may be better off by copying and pasting.

When you deal with digital material, you need different tools for combining files and converting files from different formats to a TXT format (all NLP tools deal with txt files only). PC-ACE has different routines to combine Word/txt files and convert Word files to TXT. But to convert pdf files to doc or txt you will need an external program. AntConc (http://www.laurenceanthony.net/software/antconc/), one of the corpus programs supported by PC-ACE has a good conversion routine. You can also use one of the many web-based tools, such as *RTF to PDF* (https://online2pdf.com/convert-rtf2pdf). In the conversion of a pdf file to txt, the file must not contain any images or the conversion will fail. The conversion will also fail if your pdf file is an image file. You will need, first, to convert the image to OCR (optical character reader). Acrobat Pro will do that for you. Alas, not Acrobat Reader and Acrobat Pro is expensive. If you do not have Acrobat Pro, since you will only have to do this once, just go to any of the computer labs on campus and use Acrobat Pro to convert your pdf image files.

GRADING

This is an intensive computer and writing course.

Grading will be based on the following items:

1. *Participation* (10%). You are expected to attend classes regularly (attendance is enforced through a sign-up sheet) and contribute to discussion.
2. *Presentations* (5%). Students will be asked to make in-class presentations of their work.
3. *Multiple-choice tests* (15%). There will be several *impromptu* multiple-choice tests based on the required readings.
4. *Homework* (40%). You are expected to carry out bi-weekly homework that you will upload to Blackboard. Homework assignments will involve the use of specific NLP tools applied to specific corpus data (e.g., Stanford CoreNLP, Mallet, Word2Vec, KWIC, sentence length visualization); in the second part of term, homework assignments will involve the use of various network (e.g., Cytoscape, Gephi) and GIS tools (e.g., Google Earth Pro, QGIS, Carto, TimeMapper). You will need to **present screenshots** of your work **and**, especially, **interpret your results with extensive writeups**. You need to answer questions such as: what does the tool allow you to do? How does it work? What are its pros and cons? How do you interpret the results? What does each tool tell you about your data? **Each homework will be graded out of 100 points**. Make sure to include:
    a. **screenshots of your work;**
    b. **references to the readings**.

    Expect homework to take 5 or 6 hours in a combination of computer work and writing.
4. *Research paper* (30%). You will be expected to write a *final research paper* based on the analysis of corpus data of your choice. You are welcome to organize your paper in the standard format – Introduction, Literature Review, Data & Methods, Empirical Results, Conclusions, Bibliography – but you are also encouraged to experiment with creative writing (provided that all relevant information of the standard format is still provided). **You should aim to write a publishable quality paper. The paper should include plots, charts, graphs, and links to dynamic visualizations. The paper should be up to 6 thousand words in length excluding visuals.**
5. *Bonus points*. Extra points can be gained for students who want to present and discuss a software in front of the class or write a PC-ACE TIPS file or record a PC-ACE video tutorial. A TIPS file is a document originally written for PC-ACE users that provides help on a specific issue (e.g., on the use of Gephi). It is meant as basic first-time-user help on what users could do in a software they do not know. **Bonus points will only be used to help students who are borderline between final grades.**

Attendance to class is mandatory and enforced through sign-up sheets.

Students who are not satisfied with a grade received are welcome to ask for re-grading for well-motivated reasons. The result of re-grading may be a higher grade, the same grade, or a lower grade.

HONOR CODE

The Emory University honor code applies fully to this course. When you sign an exam or submit your assignments, you are pledging to the honor code. For reference, please consult: http://www.sph.emory.edu/cms/current_students/enrollment_services/honor_code.html

HOMEWORKS

For all homeworks, please, provide screenshots and extensive write-ups of your findings. Homework submitted without screenshots will receive a ZERO grade.

Homework 1 (due at the end of week 1, Sunday August 27, at midnight)

1. Provide a one-page description of your corpus detailing 1. the number of documents; The total number of words for all documents; 3. The characteristics of the corpus and the source (e.g., newspaper articles from www.chroniclingamerica.com); 4. Reasons for selecting the corpus; 5. Hunches about what to expect from an analysis of the corpus.
   Provide screenshots of successful installation of PC-ACE on your computer (opening menu form, NLP form).
2. Separately, write one-page on "distant reading". What does the concept mean? Why distant? What are the pros and cons?

Not handing in this homework will result in an F for the course.

Homework 2: **(due at the end of week 2, Sunday September 3, at midnight)**

1. Run the **Stanford CoreNLP** on your corpus to produce the CoNLL table and provide a two-page description of your results with separate screenshots of results. From the CoNLL table, compute the Sentence and KWIC table. What area all these tables?
2. Separately, write one-page on NLP. What does it mean? What set of tools come under NLP? Are all NLP tools just as accurate?

Not handing in this homework will result in an F for the course.

Homework 3: **(due at the end of week 3, Sunday September 10, at midnight)**

1. Using the basic **STATISTICAL TOOLS** of PC-ACE, analyze your corpus and provide corpus statistics. What do the numbers tell you? What are the most significant words? Do the word frequency distributions tell you anything significant trends in your corpus? *Use any visualization tool you find appropriate to display your results.* What did Moretti and Pestre (2015) get out of simple statistics?

Homework 4: **(due at the end of week 4, Sunday September 17, at midnight)**

Write a five-page report on the results of using POSTAG and DEPREL **SEARCH TOOLS** on your corpus. Make sure to define the terms POSTAG and DEPREL and to address "meaningful" questions about significant words and word relations in your corpus (e.g., which adjectives are used for which nouns). *Use any visualization tool you find appropriate to display your results.*

Homework 5: **(due at the end of week 5, Sunday September 24, at midnight)**

Write a five-page report on the results of various **SEARCH TOOLS** of PC-ACE applied to your corpus. Make sure to define such terms as KWIC, N-Grams, and word co-occurrences and, again, to address "meaningful" questions about significant words and word relations in your corpus. *Use any visualization tool you find appropriate to display your results.* What are the differences between Google Ngram Viewer and PC-ACE N-gram Viewer? Why would you want to duplicate routines?

Homework 6: **(due at the end of week 6, Sunday October 1, at midnight)**

Using PC-ACE Text Analysis tools, analyze your corpus in terms of Noun density, Verb modality, Verb tense, Verb voice. What do these terms mean? What do the numbers tell you? *Use any visualization tool you find appropriate to display your results.*

Homework 7: **(due at the end of week 7, Sunday October 8, at midnight)**

Sentiment Analysis can tell you, sentence by sentence, whether a text conveys negative, neutral, and positive feelings. Using PC-ACE Text Analysis tools run Sentiment Analysis routine via the Stanford CoreNLP. What do the results tell you? *Use any visualization tool you find appropriate to display your results.*

Homework 8: **(due at the end of week 8, Sunday October 15, at midnight)**

Using the PC-ACE Text Analysis tools, analyze your corpus for text readability and sentence complexity (CLAS).

Using the "literary" tools learned in Part III, analyze Nobel laureate William Falkner's story "Dry September" and James Murphy's story "Miracles Thicker than Fog". What do the results tell you? Do these tools *really* capture the style of these two authors?

Use any visualization tool you find appropriate to display your results.

Homework 9: **(due at the end of week 9, Sunday October 22, at midnight)**

Using the topic modeling tool **Mallet** found under the button TEXT ANALYSIS tools of PC-ACE, analyze your corpus for significant topics in your corpus. Does Mallet correctly categorize your corpus? *Use any visualization tool you find appropriate to display your results.*

Homework 10: **(due at the end of week 10, Sunday October 29, at midnight)**

Using the tool Word2Vec found under the button TEXT ANALYSIS tools of PC-ACE, analyze your corpus for significant relations among words in your corpus. What is a neural network approach to word relations? Do the results of Word2Vec confirm the results you obtained in previous analyses and based on different tools? What are stop words? Why would you include them or exclude them from analyzing your corpus? *Use any visualization tool you find appropriate to display your results.*

Homework 11: **(due at the end of week 11, Sunday November 5, at midnight)**

Using Word Clouds programs, display the words of your corpus in the various word cloud programs supported by PC-ACE. Using a photograph of yourself, take a letter or an essay you wrote and display the words in Tagxedo.

Homework 12: **(due at the end of week 12, Sunday November 12, at midnight)**

Using the SVO extractor tool found under the button TEXT ANALYSIS tools of PC-ACE, analyze your corpus to extract SVO information. Run the tool with and without anaphora and coreference resolution (via Stanford CoreNLP). What difference does it make? Finally, use the resolution GUI to resolve the 35% cases not dealt with automatically. Again… what difference does it make? Can you display the SVO information in a network graph?

Homework 13: **(due at the end of week 13, Sunday November 19, at midnight)**

Using the NER information of your corpus (or a sample batch of newspaper articles you will download), extract location information, geocode the location and map it. Use QGIS and Google Earth Pro. In QGIS, draw different types of maps (dot, dot frequency, heatmap). In Google Earth Pro display selected information in the field DESCRIPTION, with different types of information displayed in bold/italic/colors. In Google Earth Pro run a dynamic map. What kind of information do you need to draw dynamic GIS maps?

Homework 14: **THANKSGIVING BREAK. NO HOMEWORK DUE!!!**

Homework 15: **(due at the end of week 15, Sunday December 2, at midnight)**

Display the same GIS data of the previous week in Carto and TimeMapper. Do you get different type of visuals? How can you make your maps more beautiful, more vivid, following geographer Peirce Lewis's recommendations (1985)? Perhaps combining (selecting) evidence from the previous homework, can you try to write a homework that is beautiful and vivid?

Homework 16: **End of term. NO HOMEWORK DUE!!!**

READINGS

Readings for the course come from books and journal articles or book chapters. All reading material has been placed on **Ereserve** and physical copies of most of the required books are on Reserve:

1. Tufte, Edward R. 2006. *Beautiful Evidence.* Cheshire, CN: Graphics Press LLC.
2. Tufte, Edward R. 2001 [1983]. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
3. Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.
4. Tufte, Edward R. 2003. *The Cognitive Style of PowerPoint*. Cheshire, CT: Graphics Press.
5. Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart.
6. Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart.
7. Bertin Jaques. 1967 (2010). *Semiology of Graphics: Diagrams, Networks, Maps.* Redlands, CA: ESRI Press.
8. Wilkinson, Leland. 1995 (2005). *The Grammar of Graphics*. Second edition. New York: Springer.
9. Yau, Nathan. 2012. *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley.
10. Munzner, Tamara. 2014. *Visualization Analysis and Design*. Boca Raton, FL: CRC Press.
11. Card, Stuart K., Jock D. Mackinlay, Ben Shneiderman (eds.). 1999. *Readings in Information Visualization: Using Vision to Think.* San Diego, CA: Academic Press.
12. Spence, Robert. 2014. *Information Visualization: An Introduction*. Third edition. New York: Springer.
13. Ware, Colin. 2012. *Information Visualization: Perception for Design.* Third edition. Waltham, MA: Elsevier.
14. Moretti, Franco. 2013. *Distant Reading*. London: Verso.
15. Moretti, Franco. 2005. *Graphs, Maps, Trees. Abstract Models for a Literary History*. London: Verso.
16. Moretti, Franco. 1998 (1997). *Atlas of the European Novel, 1800-1900.* London: Verso.
17. Forceville, Charles. 1996. *Pictorial Metaphor in Advertising*. London: Routledge.

There are some weeks with very heavy readings. And books are long... Read enough to know what they are saying. Some weeks are heavier in readings than others. Try to distribute your workload appropriately. I have separated readings in the standard format of Required and Suggested readings. However, I have placed comments in red under the Suggested readings labels. Suggested readings are only meant to provide a minimal bibliography. For the purpose of your grade, you are not expected to read them (unless, of course, you are a glutton for punishment! Although … it is also true that the more you read, the more you know… and the better you would do in your presentations and written work). You are strongly encouraged to take at least a quick look

to those readings to familiarize yourself with the basic language and arguments on specific topics.

August 24
Week 1
Introducing the course
Big data and "distant reading"
Preparing your corpus for "distant reading"

Required readings:

Moretti, Franco. 2013. *Distant Reading*. London: Verso.
Moretti, Franco. 2005. *Graphs, Maps, Trees. Abstract Models for a Literary History.* London: Verso.
Kirschenbaum, Matthew G. 2009. "The Remaking of Reading: Data Mining and the Digital Humanities." Talk given at the 2009 National Science Foundation Symposium on the Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation.
Kirschenbaum, Matthew G. 2011. "Digital Humanities Archive Fever." Plenary lecture at the Digital Humanities Summer Institute at the University of Victoria, June 2011. August 22, 2011 at 9:56 PM · https://vimeo.com/28006483
Jockers, Matthew L. and David Mimno. 2013. "Significant Themes in 19th-Century Literature," *Poetics*, Vol. 41, No. 6, pp. 750-769.
Pannapacker, William. 2009. "The MLA and the Digital Humanities." *The Chronicle of Higher Education*, December 28, 2009.

Video. Talk by Nello Cristianini, "The Big-Data Revolution and its Impact on Science and Society." https://www.youtube.com/watch?v=PzicexAmycA **(some words of caution on the big-data revolution**…)

August 29-September 14
Weeks 2-4
Part I: NLP (Natural Language Processing): Basic language

August 29-31
Week 2
NLP (Natural Language Processing): Sentence splitting, tokenization, lemmatization (lemmas and stems)
The CoNLL table: What's in the CoNLL table?
Word co-occurrences (KWIC)

Software: PC-ACE & Stanford CoreNLP

Required readings:

Top 20 free software for Text Analysis, Text Mining, Text Analytics
http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/

Franzosi, Roberto. NLP TIPS files.

Video. Talk by Nello Cristianini on Big Data ("Patterns in Media Content)
https://www.youtube.com/watch?v=mmWRNRPb0W0

September 5-7
Week 3:
Word co-occurrences (KWIC)
Basic corpus statistics

Video. Talk by Nello Cristianini on visualizing millions of tweets ("Mood Changes of UK during 2009-2012") https://www.youtube.com/watch?v=gG5ZH2JfqIU
Video. Talk by Nello Cristianini on visualizing millions of newspaper articles ("Key Actors in US Presidential Elections 2012 – primaries")
https://www.youtube.com/watch?v=ptH5FkKSvvU

Take a quick look at some of these readings. Familiarize yourself with what the ready availability of digital newspaper archives would allow you to do/and how.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences (PNAS), PNAS,* Published online January 9, 2017 E457–E465.
Sudhahar, Saatviga, Giuseppe A. Veltri, and Nello Cristianini. 2015. "Automated Analysis of the US Presidential Elections Using Big Data and Network Analysis." *Big Data & Society*, Vol. 2, No. 1, pp. 1–28. DOI: 10.1177/2053951715572916.
Mohr, John, Robin Wagner-Pacifici, Ronald L. Breiger, Petko Bogdanov. 2013. "Graphing the Grammar of Motives in National Security Strategies - Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics," *Poetics*, Vol. 41, No. 6, pp. 670-700.
Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data," *Theory and Society*, Vol. 43, No. 3, pp. 465-482.
Seguin, Charles. 2015 web download. "Scraping Historical Newspaper Archives: The Transformation of Public Lynching Inquiry." Sociological Theory 10:164–93.
Light, Ryan. 2014. "From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses." *Social Currents*, Vol. 1, No. 2, pp. 111–129.
Light, Ryan and Jeanine Cunningham. 2016. "Oracles of Peace: Topic Modeling, Cultural Opportunity, and the Nobel Peace Prize, 1902–2012." *Mobilization: An International Quarterly*, Vol. 21, No. 1, pp. 43–64.
He, Qin. 1999. "Knowledge Discovery through Co-word Analysis." *Library Trends* 48:133–59.
Discourse in the US." http://badhessian.org/2014/01/scraping-historical-newspaper-archives-the-transformation-of-public-lynching-discourse-in-the-us/Snowsill, Tristan, Ilias Flaounas, Tijl De Bie, and Nello Cristianini. 2010. "Detecting Events in a Million New York Times Articles," *Lecture Notes in Computer Science*, pp. 615-618.
Zervanou, Kalliopi, Marten Düring, Iris Hendrickx, and Antal van den Bosch. 2014. "Documenting Social Unrest: Detecting Strikes in Historical Daily Newspapers," *Lecture Notes in Computer Science*, pp. 120-133.

September 12-21
Weeks 4-5
Part II: What's in your corpus? Word searches and visualization

September 12-14
Week 4
Searches by words' grammatical roles (POSTAG, Part Of Speech Tag)
Searches by syntactic relations between words (DEPREL DEPendency RELations)

Software: PC-ACE & Stanford CoreNLP

Required readings:

Franzosi, Roberto. NLP TIPS files.

September 19-21
Week 5
N-grams searches
Word co-occurrences searches

Software: PC-ACE & Stanford CoreNLP, Google ngrams

Required readings:

Franzosi, Roberto. NLP TIPS files.

Become familiar with the basic language of culturomics!

Michel, Jean-Baptiste and Erez Lieberman Aiden. 2011. "What we learned from 5 million books". https://www.ted.com/talks/what_we_learned_from_5_million_books?language=en

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, 14 January 2011, Vol. 331, pp. 176-182.

Leetaru, Kalev H. 2011. "Culturomics 2.0: Forecasting Large–scale Human Behavior Using Global News Media Tone in Time and Space." *First Monday,* Vol. 16, No. 9 (on-line journal).

Nunberg, Geoffrey. 2009. "Google's Book Search: A Disaster for Scholars." *The Chronicle of Higher Education,* August 31, 2009.

Schwartz, Tim. 2011. "Culturomics Periodicals Gauge Culture's Pulse." *Science,* Vol. 332, 1 April 2011, p. 35-36.

September 26-October 26
Weeks 6-8
Part III: Dissecting your corpus

September 26-28
Week 6
Noun density
Verb modality: Ability, possibility, permission, and obligation
Verb tense: past, future, gerundive
Verb voice: Active and passive verb forms

Software: PC-ACE & Stanford CoreNLP

Required readings:

Franzosi, Roberto. NLP TIPS files.
Moretti, Franco and Dominique Pestre. 2015. "BANKSPEAK: The Language of World Bank Reports." *New Left Review*, Vol. 92, pp. 75-99.

October 3-5
Week 7
Sentiment analysis

Required readings:

Video. Talk by Min Song on Sentiment Analysis. https://www.coursera.org/learn/text-mining-analytics/lecture/5RwtX/5-6-how-to-do-sentiment-analysis-with-sentiwordnet

You can download SentiWordNet at http://sentiwordnet.isti.cnr.it/

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche. Pisa, IT.
Bradley, Margaret M. and Peter J. Lang. 1999. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. NIMH Center for the Study of Emotion and Attention. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
Dodds, Peter Sheridan and Christopher M. Danforth. 2010. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents." *Journal of Happiness Studies*, Vol. 11, pp. 441–456.
Esuli, Andrea and Fabrizio Sebastiani. 2006. "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining." In: pp. 417–422. *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, Genova, IT.
Hills, Thomas, Eugenio Proto, and Daniel Sgroi. 2015. "Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books." *IZA (Forschungsinstitut zur Zukunft der Arbeit/Institute for the Study of Labor)*, Discussion Paper No. 9195, pp. 1-25.
Nguyen, Thin, Dinh Phung, Brett Adams, Truyen Tran, and Svetha Venkatesh. 2010. "Classification and Pattern Discovery of Mood in Weblogs." In: pp. 283–290, M. J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, Berlin: Springer-Verlag.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes". *EPJ Data Science*, Vol. 5, No. 31, pp. 1-12.

Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas." 2013. *Behavior Research Methods*. Advance Online Publication. DOI: 10.3758/s13428-012-0314-x. [PubMed]

Fall break October 9-10

October 10-12
Week 8

Text readability: What grade level does a text require to be comprehensible?
Sentence complexity: Using CLAS (Computerized Linguistic Analysis System) to measure linguistic complexity

Required readings:

https://readable.io/

http://www.readabilityformulas.com/gunning-fog-readability-formula.php

Pakhomov, Serguei, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. 2011. "Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing". *Behavior Research Methods*, Vol. 43, No. 1, pp. 136–144.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2013. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas". *Behavior Research Methods*, Vol. 46, pp. 904–911.

Hills, Thomas T. and James S. Adelman. 2015. "Recent Evolution of Learnability in American English from 1800 to 2000." *Cognition*, Vol. 143, pp. 87–92.

October 17-19
Week 9
Part IV: Topic modeling
Week 9

Using Mallet to tell us what different documents in a text corpus talk about

Software: Mallet (in PC-ACE)

Required readings:

Franzosi, Roberto. NLP TIPS files.

Graham, Shawn, Scott Weingart and Ian Milligan. 2012. *Getting Started with Topic Modeling and MALLET*. The Programming Historian. Document available on the web at http://programminghistorian.org/lessons/topic-modeling-and-mallet

Flaounas, Ilias, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini, 2013 "Research Methods in the Age of Digital Journalism: Massive-

scale Automated Analysis of News: Content Topics, Style and Gender," *Digital Journalism*, Vol. 1, No. 1, pp. 102–116.

@

There are some great readings in this 2013 special issue of *Poetics*. Take a quick look at these articles and dive deeper in the ones that go to the heart of your interests.

McFarland, Daniel A. and Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D. Manning, Daniel Jurafsky. 2013. "Differentiating Language Usage through Topic Models," *Poetics*, Vol. 41, No. 6, pp. 607-625.

DiMaggio, Paul, Manish Nag, and David Ble. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding," *Poetics*, Vol. 41, No. 6, pp. 570-606.

Miller, Ian Matthew. 2013. "Rebellion, Crime and Violence in Qing China, 1722–1911: A Topic Modeling Approach," *Poetics*, Vol. 41, No. 6, pp. 626-649.

Marshall, Emily A. 2013. "Defining Population Problems: Using Topic Models for Cross-national Comparison of Disciplinary Development," *Poetics*, Vol. 41, No. 6, pp. 701-724.

Tangherlini, Timothy R. and Peter Leonard. 2013. "Trawling in the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities Research," *Poetics*, Vol. 41, No. 6, pp. 725-749.

October 24-November 2
Weeks 10-11
Part V: Visualizing words
October 24-26
Week 10
Word2Vec: Neural network approaches to word relations

Software: Word2Vec (in PC-ACE)

October 31-November 2
Week 11
Word clouds

Software: Bookworm, Wordle, TagCrowd, Tagul and Taxedo (Tagul and Tagxedo allow to draw word clouds in specific shapes)

Required readings:

The 8 Best Free Word Cloud Creation Tools for Teachers: http://elearningindustry.com/the-8-best-free-word-cloud-creation-tools-for-teachers

Nine free on-line word clouds generators: http://www.smashingapps.com/2011/12/15/nine-excellent-yet-free-online-word-cloud-generators.html

Video. Ted Talk by Erez Lieberman Aiden and Jean-Baptiste Michel, 2011, "A picture is worth 500 billion words". http://tedxtalks.ted.com/video/TEDxBoston-Erez-Lieberman-Aid-2

Goldstone, Andrew and Ted Underwood. 2014. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History*, Vol. 45, No. 3, pp. 359-384.

November 7-9
Week 12
Part VI: Extracting SVOs
Week 12

Coreference resolution
Anaphora resolution
ClausIE and automatic SVO extraction

Required readings:

Del Corro, Luciano and Rainer Gemulla. 2013. "ClausIE: Clause-Based Open Information Extraction." Proceeding WWW '13 *Proceedings of the 22nd international conference on World Wide Web*, pp. 355-366, Rio de Janeiro, Brazil – May 13-17, 2013.

Computer scientists are coming closer to finding automated solutions to extracting the "who, what, when, where, why, and how" of narrative. It will not be long before they will put social scientists out of their miseries of manual coding!

Sudhahar, Saatviga, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. 2015. "Network Analysis of Narrative Content in Large Corpora," *Natural Language Engineering*, Vol. 21, No. 1, pp. 81-112.

Sudhahar, Saatviga and Nello Cristianini. 2013. "Automated Analysis of Narrative Content for Digital Humanities," *International Journal of Advanced Computer Science*, Vol. 3, No. 9, Pp. 440-447.

Sudhahar, Saatviga, Thomas Lansdall-Welfare, Ilias Flaounas, and Nello Cristianini. 2012. "Quantitative Narrative Analysis of US Elections in International News Media." The Internet, Policy & Politics Conferences, Oxford Internet Institute, University of Oxford. http://ipp.oii.ox.ac.uk/2012/programme-2012/track-a-politics/panel-5a-topics-memes-and-sentiment/saatviga-sudhahar-thomas-lansdall

Finlayason, Mark Alan. 2012. *Learning Narrative Structure from Annotated Folktales.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Lendvai1, Piroska, Thierry Declerck, Sándor Darányi, Pablo Gervás, Raquel Hervás, Scott Malec, and Federico Peinado. 2010. "Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case," Proceedings of the Seventh conference on International Language Resources and Evaluation, European Language Resources Association (ELRA).

Scott Malec, Sándor Darányi, Trevor Cohen, and Dominic Widdows. [no date]. "Landing Propp in Interaction Space: First Steps toward Scalable Open Domain Narrative Analysis with Predication-based Semantic Indexing."

November 14-23

Weeks 13-15
Part VII: Extracting and visualizing time and space
November 14-16
Week 13
Using NER information (Back to the CoNLL table)
Geocoding

Software: Carto, Google Earth Pro, QGIS, Tableau, TimeMapper, GeoNames, OpenStreetMap

Required readings:

Franzosi, Roberto. Geocoding TIPS files.

Graham, Mark and Taylor Shelton. 2013. "Geography and the Future of Big Data, Big Data and the Future of Geography." *Dialogues in Human Geography*, Vol. 3, No. 3, pp. 255–261.
Gregory, Ian, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. 2015. "Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research." *International Journal of Humanities and Arts Computing*, Vol. 9, No. 1, pp. 1–14.
Jessop, Martyn. 2008. "The Inhibition of Geographical Information in Digital Humanities Scholarship." *Literary and Linguistic Computing*, Vol. 23, No. 1, pp. 39-50.
Kitchin, Rob. 2013. "Big Data and Human Geography: Opportunities, Challenges and Risks." *Dialogues in Human Geography,* Vol. 3, No. 3, p. 262–267.

Suggested readings:

Basso, Keith H. 1988. "'Speaking with Names': Language and Landscape among the Western Apache." *Cultural Anthropology,* Vol. 3, No.2, pp. 99-130.
Rosenberg, Daniel and Anthony Grafton. 2010. *Cartographies of Time*. New York, Princeton Architectural Press.
Massey, Doreen. 2005. *For Space.* Thousand Oaks, CA: Sage.

Check out some cool mapping sites

http://www.radicalcartography.net/
http://selfiecity.net/
http://www.floatingsheep.org/

http://dsl.richmond.edu/
http://photogrammar.yale.edu/
http://atlas.lib.uiowa.edu

Software: Carto, Google Earth Pro, QGIS, Tableau, TimeMapper, GeoNames, OpenStreetMap

Required readings:

Franzosi, Roberto. GIS TIPS files.
Yuan, May. 2010. "Mapping Text". In: pp. 109-123, David J. Bodenhamer, John Corrigan, and Trevor M. Harris (eds.), *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington, IN: Indiana University Press.
Yuan, May. 2014. "Temporal GIS for Historical Research." In: pp. 45-55, A. Crespo Solana (ed.), *Spatio-Temporal Narratives: Historical GIS and the Study of Global Trading Networks*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

November 22-24. THANKSGIVING BREAK.

November 21
Week 14

No readings!!!

November 28-30
Week 15
Visualizing space
Visualizing time and space

Required readings:

Corrigan, John. 2010. "Qualitative GIS and Emergent Semantics". In: pp. 76-88, David J. Bodenhamer, John Corrigan, and Trevor M. Harris (eds.), The Spatial Humanities: GIS and the Future of Humanities Scholarship. Bloomington, IN: Indiana University Press.
Lewis, Peirce. 1985. "Beyond Description." *Annals of the Association of American Geographers*, Vol. 75, No. 4, pp. 465-478.

December 5
Part VIII: What have we learned?

December 5
Week 16: A game changer? Digital humanities (or Digital scholarship), beautiful evidence, and visual rhetoric

Required readings:

Anderson, Chris. 2008. "The end of theory: The data deluge makes the scientific method obsolete." *Wired Magazine*, Vol. 16, No. 7,
Available at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
Franzosi, Roberto. 2015. "Of Stories and Beautiful Things: Digital Scholarship, Method, and the Nature of Evidence." Unpublished manuscript.
Gold, Matthew K. (ed.). 2012. *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Healy, Kieran and James Moody. 2014. "Data Visualization in Sociology," *Annual Reviews of Sociology*, Vol. 4, pp. 105–28.
Kirschenbaum, Matthew G. 2012. "What is Digital Humanities and What's it Doing in English Departments?" In: pp. 3-11, Matthew K. Gold (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.
Liu, Alan Y. 2012. "The State of the Digital Humanities: A Report and a Critique." *Arts and Humanities in Higher Education*, Vol. 11, Nos. 1-2, pp. 8-41.
Liu, Alan Y. 2013. "The Meaning of the Digital Humanities." *PMLA*, Vol. 128, No. 2, pp. 409-423.

Suggested readings:

Moretti, Franco. 1998 (1997). *Atlas of the European Novel, 1800-1900*. London: Verso.
Tufte, Edward R. 2006. *Beautiful Evidence*. Cheshire, CN: Graphics Press LLC.
Tukey, John W. 1969. "Analyzing Data: Sanctification or Detective Work?" *American Psychologist*, Vol. 24, No. 2, pp. 83-91.
Tukey, John W. 1980. "We Need Both Exploratory and Confirmatory." *The American Statistician*, Vol. 34, No. 1, pp. 23-25.
Wainer, Howard. 1984. "How to Display Data Badly," *American Statistician*, Vol. 38, No. 2, pp. 137–47.

On visual rhetoric:

Required readings:

Kostelnick, Charles. 2007. "The Visual Rhetoric of Data Displays: The Conundrum of Clarity," *IEEE Transactions on Professional Communications*, Vol. 50, No. 4, pp. 280–94.
McQuarrie, Edward F. and David Glen Mick. 1996. "Figures of Rhetoric in Advertising Language." *The Journal of Consumer Research*, Vol. 22, No. 4, pp. 424-38.

Suggested readings:

"Ad-writers are some of the most skilled rhetoricians in our society." (Edward P.J. Corbett and Robert J. Connors) Whatever else data visualization does… hopefully, it contributes to creating persuasive evidence. And if it is persuasive, it is rhetorical, rhetoric being the art of persuasion.

Tom, Gail and Anmarie Eves. 1999. "The Use of Rhetorical Devices in Advertising." *Journal of Advertising Research*, Vol. 39, July-August, pp. 39-43.
Forceville, Charles. 1996. *Pictorial Metaphor in Advertising*. London: Routledge.
Dyer, Gillian. 1988[1982]. "Chapter 8. The Rhetoric of Advertising", In: pp. 127-150, *Advertising as Communication*. Oxford: Routledge.
Leigh, James H. 1994. "The Use of Figures of Speech in Print Ad Headlines." *Journal of Advertising*, Vol. 23, No. 2, pp. 17-33.
McQuarrie, Edward F. and David Glen Mick. 1999. "Visual Rhetoric in Advertising: Text-Interpretive, Experimental, and Reader-Response Analyses." *The Journal of Consumer Research*, Vol. 26, No. 1 pp. 37-54.

Scott, Linda M. 1994. "Images in Advertising: The Need for a Theory of Visual Rhetoric." *The Journal of Consumer Research*, Vol. 21, No. 2, pp. 252-73.
Bush, Alan J. and Gregory W. Boller. 1991. "Rethinking the Role of Television Advertising during Health Crises: A Rhetorical Analysis of the Federal AIDS Campaigns." *Journal of Advertising*, Vol. 20, No. 1, pp. 28-37.
Barnard, Malcolm. 2005. "Metaphor/metonymy/synechdoche". In" pp. 50-54, *Graphic Design as Communication*. Abingdon, UK: Routledge.

Suggested readings:

Tufte has been a leading scholar on data visualization. Bertin, Cleveland, and Wilkinson are "classical" readings on data visualization. Some of the other readings, Yau in particular, represent the current state of the art on data visualization.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.
Tufte, Edward R. 2003. *The Cognitive Style of PowerPoint*. Cheshire, CT: Graphics Press.
Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart.
Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart.
Bertin Jaques. 1967 (2010). *Semiology of Graphics: Diagrams, Networks, Maps.* Redlands, CA: ESRI Press.
Wilkinson, Leland. 1995 (2005). *The Grammar of Graphics*. Second edition. New York: Springer.
Yau, Nathan. 2012. *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley.
Munzner, Tamara. 2014. *Visualization Analysis and Design*. Boca Raton, FL: CRC Press.
Card, Stuart K., Jock D. Mackinlay, Ben Shneiderman (eds.). 1999. *Readings in Information Visualization: Using Vision to Think.* San Diego, CA: Academic Press.
Spence, Robert. 2014. *Information Visualization: An Introduction*. Third edition. New York: Springer.
Ware, Colin. 2012. *Information Visualization: Perception for Design.* Third edition. Waltham, MA: Elsevier.
Cleveland, William S. and Robert Mcgill. 1984. "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, Vol. 79, No. 388, pp. 807-22.
Funkhouser, H. Gray. 1937. "Historical Development of the Graphical Representation of Statistical Data," *Osiris*, Vol. 3, pp. 269-404.
Kosslyn, Stephen M. 1987. "Understanding Charts and Graphs." DTIC unpublished document.
McGill, Robert, John W. Tukey and Wayne A. Larsen. 1978. "Variations of Box Plots." *The American Statistician*, Vol. 32, No. 1, pp. 12–16.
Wickham, Hadley and Lisa Stryjewski. 2011. "40 Years of Boxplots." Unpublished manuscript.
Tufte, Edward R. 2001 [1983]. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
Anscombe, Francis J. 1973. "Graphs in statistical analysis." *American Statistician*, Vol. 27, pp. 17–21.
Friendly, Michael and Daniel Denis. 2005. "The Early Origins and Development of the Scatterplot." *Journal of the History of the Behavioral Sciences*, Vol. 41, No. 2, pp. 103–130.

Marshall, Alfred. 1885. "On the Graphic Method of Statistics," *Journal of the Statistical Society of London*, Jubilee Volume (Jun. 22 - 24, 1885), pp. 251-260.
Keynes, John M. 1938. "Review of H.G. Funkhouser, Historical Development of the Graphical Representation of Statistical Data." *Economic Journal*, Vol. 48, No. 190, pp. 281–82.
Spence, Ian. 2005. "No Humble Pie: The Origins and Usage of a Statistical Chart," *Journal of Educational and Behavioral Statistics*, Vol. 30, No. 4, pp. 353–368.

Digital humanities websites: Trans-Atlantic Slave Trade (http://www.slavevoyages.org) by David Eltis, Georgia Civil Rights Cold Cases (https://scholarblogs.emory.edu/emorycoldcases) by Hank Klibanoff

The Digital Scholarship Lab at the University of Richmond, http://dsl.richmond.edu/
The Yale photographic site http://photogrammar.yale.edu/ for the visualization of some 170,000 photographs from 1935 to 1945 created by the United States Farm Security Administration and Office of War Information (FSA-OWI).

Atlas of Early Printing at the University of Iowa, http://atlas.lib.uiowa.edu