Sections 55 & 58 Instructor:

Lynd D. Bacon, Ph.D., M.B.A.
lynd.bacon@northwestern.edu
+1 650 593 2198 (o)

TA: Tom Pritchard

This is the syllabus for Sections 55 and 58. These two sections are identical, except that Section 58's Sync Sessions are on Thursdays at 7 p.m. U.S. Central Time, snd Section 55's Sync Sessions are on Tuesdays at 7 p.m. Central Time.

So, let's have at it.

Course Description

This 420 course is about manipulating data to prepare it for analysis using Python tools. SQL and NoSQL technologies are referred to to some extent and in basic ways. It's expected that ElasticSearch will be the NoSQL store that we'll be using to provide access to some assignment data this term.

This course's learning goals and objectives are essentially the same as other sections taught during the Winter 2017 term. But it will likely differ from them in several respects. There is no team project. There are four main graded assignments that each student completes by him- or herself. These assignments are due every two weeks with the first due at the end of the second week of the term. Students are encouraged to work with their classmates on Canvas and in study groups. Each student is just required to submit their own work for each assignment.

The data used for assignments varies in structure and volume. It includes poorly structured and user-generated content data. The data are stored in various formats, and are about things like customer transactions, air transportation, corporate email, and customer evaluations of hospitality experiences.

Each student is expected to participate in weekly discussions on Canvas. These weekly discussions include doing simple exercises or tasks that are shared with the class. These are called "Do and Explain" (D&E) tasks. They count towards the grade in the course.

This course won't make a student a proficient Python programmer. But being at least a little familiar with Python at the beginning of the course will certainly be a help. If you are completely new to Python, and haven't coded in another language, you may want to get some exposure to Python in another course first before taking this version of 420.

Students in this course access some assignment data stored on the University's Social Social Sciences Computing Cluster (SSCC). Linux is the operating system used on the SSCC, and the data stored on it are in PostgreSQL, perhaps and one or more NoSQL, databases. Some experience with these things is a plus, but the tasks required to use the SSCC and the databases are rather basic. The vast majority of students have so far had little or no trouble with them.

In this course students are required to "peer review" their classmates' contributions to the course. Students' peer reviews count towards their final grade in the course.

Course Perspective and Pedagogical Philosophy

This isn't a course just about Python, or about SQL or NoSQL, per se. It's about getting data into a desired condition to be analyzed. It's about formulating, implementing, and testing solutions to the problem of getting data from a current state to a required state. So, code, datasets, and storage technologies aside, it's about practical data preparation problem-solving.

In the opinion of many, a good way to learn many things is by *doing* them. The famous mathematician and educator Paul Halmos once said that to do is to know, and to talk is *not* to teach. I'll add to this that a book is not by itself a course. There are readings in this course, as there are in other MSPA courses. The readings provide some reference material, and can fill in content gaps with respect to the course learning objectives. For some tasks in this course students may need to go beyond the required and recommended readings by looking elsewhere, like online. This is a key aspect of practical problem-solving, a critical skill for data scientists and for predictive anlaytics professionals.

In this course the problems to be solved consist of getting data from "State A" to "State B." They often need to be further define these problems so as to open up potential solutions, and ways to apply solutions using (usually) Python code that provides results. Students in this course are not provided with the code to do the assignments, or with complete solutions to them. General feedback, tips, and "hints" on assignments are provided, by me (your instructor), our TA, and also by your fellow 420 students as participants in this course's learning community.

FAIR WARNING: This is an "active learning" course, and is consistent with the teaching philosophy of the MSPA program. To get the most out of it, you need to engage with the material and other students consistently, and to "do the work." Every week you'll have something you need to do and to share with the class. The four major graded assignments are extensive. You'll have two weeks to complete each one. You should plan on starting on each of them as soon as they are available to you. They are almost impossible to complete successfully on the weekends they are due, at the last minute.

Here's something about active learning: https://en.wikipedia.org/wiki/Active_learning

Prerequisites

None. But some experience with Python will serve a student well. This is not an introductory Python programming course.

Course Learning Goals

These are the common 420 course goals. They are achieved by completing the course readings and doing the assignments. You will be the ultimate, and most important, judge of whether you attained them.

- Define key terms, concepts and issues in data management and database management systems with respect to predictive modeling
- Evaluate the constraints, limitations and structure of data through data cleansing, preparation and exploratory analysis to create an analytical database.
- Use object-oriented scripting software for data preparation.
- Transform data into actionable insights through data exploration.

Textbooks

Be sure to consider any inserted notes about the following books. Also, it's possible that newer editions of some of the following may now be available. Updates and revisions of texts in the data science area happen with great frequency.

Lubanovic, B. (2015). *Introducing Python: Modern Computing in Simple Packages.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-35936-2]

McKinney, W. (2013) *Python for Data Analysis: Agile Tools for Real-World Data.* Sebastopol, Calif O'Reilly. [ISBN-13: 978-1-449-31979-3] (Note: This is a useful book, but it's a little out of date. The most current documentation for the Pandas package can be found at http://pandas.pydata.org/)

Osborne, J. W. (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* Thousand Oaks, Calif.: Sage. [ISBN-13: 978-1-4129-8801-8]

Reference Books (Recommended. Use what fits your needs.)

Barrett, D. J. (2012) *Linux Pocket Guide* (2nd ed.). Sebastopol, Calif.: O'Reilly.
[ISBN-13: 978-1-449-31669-3]

Connolly, T. M. and Begg, C. E. (2015). *Database Systems: A Practical Approach to Design, Implementation, and Management* (6th ed.). Upper Saddle River, N.J.: Pearson. [ISBN-13: 978-0-13-294326-0] (selected chapters will be available on eReserve, but this is a useful reference)

Harrison, Matt. *Learning the Pandas Library: Python Tools for Data Munging, Data Analysis, and Visualization.* (2016). CreateSpace Independent Publishing Platform, ISBN-13: 978-1533598240. This book is available in paperback from and for Kindle on Amazon.com. It can also be ordered from www.abbotthall.bncollege.com, but it will not be returnable there. *It's a good, gentle introduction to the pandas package.*

Lutz, M. (2014). *Python Pocket Reference* (5th ed.). Sebastopol, Calif.: O'Reilly.
[ISBN-13: 978-1-449-35701-6]

Jamul, K. & Kazil, J. (2016) *Data Wrangling with Python.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-491-94881-1] (A good intro book, although not much on the Pandas package. *Be aware that if you order this through the Northwestern bookstore, it won't be returnable.*)

Paro, A. *Elasticsearch 5.X cookbook* (3rd ed.) (2017) Birmingham UK: Packt Publishing Ltd., ISBN 978-1-78646-558-0. www.packtpub.com. This book can be ordered from www.abbotthall.bncollege.com, but it will not be returnable there. It should also be available elsewhere.

Online, for the Python Beginner:

Automate the Boring Stuff with Python by Al Sweigart. Available at:

https://automatetheboringstuff.com

Also available from NoStarch Press. And see also http://inventwithpython.com

Software and Systems (Python)

No software purchases are needed for this course. The course utilizes software that is freely available for PC/Windows, Mac/OS X, and Linux systems.

What particular version or distribution of Python is the student's choice. The program suggests **Enthought's Canopy** integrated development environment (IDE) for Python, which runs on Windows, Mac, and Linux systems. A free academic version of Canopy is available from Enthought. To get a copy of canopy, go to https://www.enthought.com/products/canopy/academic and request an academic license. Available for Python 2.7x, and more recently, also for Python 3.x.

The **Anaconda Python distribution** from Continuum Analytics (https://www.continuum.io/) is another highly regarded scientific Python distribution that is available for free from Continuum. You can download it from the Continuum web site. A special license (academic or otherwise) is *not* required to use it. Note that some of the course content refers specifically to Canopy, but it is easily generalized to Anaconda, or to other Python environments. Package management may be somewhat easier in Canopy, but packages may be "fresher" (more current) in Anaconda. Anaconda is available in both Python 2.7 and Python 3.x flavors.

To learn more about why there are these two versions of Python and some ways that they differ, see:

https://dzone.com/articles/the-key-differences-between-python-2-and-python-3

If in doubt, go with Python 3 rather than 2.

Both Canopy and Anaconda support the use of the **Jupyter Notebook** (http://www.jupyter.org), which is how assignments are provided in this course. Jupyter was previously called iPython Notebook. It is a web browwer application that can combine text, runable code from various languages including Python, R, and Julia, models, and visualizations.

Some good books for learning more about Python include the following:

Beazley, D. (2009). *Python Essential Reference* (4th ed.). Upper Saddle River, N.J.: Pearson/Addison-Wesley

.

Beazley, D. & Jones, B. K. (2013). *Python Cookbook* (3rd ed.). Sebastopol, Calif.: O'Reilly.
Chun, W. J. (2007). *Core Python Programming* (2nd ed.). Upper Saddle River, N.J.: Prentice Hall.

Gift, N. and Jones, J. M. (2008). *Python for Unix and Linux System Administrators: Efficient Problem Solving with Python.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-0-596-51582-9]

Hellmann, D. (2011). *The Python Standard Library by Example.* Upper Saddle River, N.J.: Pearson/Addison-Wesley. [ISBN-13: 978-0-321-76734-9]

A useful overview of the world of Python, *The Hitchhiker's Guide to Python* by Kenneth Reitz, is available online at <http://docs.python-guide.org/en/latest/>.

There's an abundance of useful Python documentation, coding examples and advice online. If you have a problem to figure out, search for it online. Chances are someone else has encountered it before you.

There is *a lot* of assigned reading in this course. A good strategy for approaching it is to start with the readings that are appropriate for your level of experience with Python, with Linux, and with the database systems used. Something that might "lighten the load" somewhat is to share covering the assigned and recommended readings content amongst members of a study group.

Software and Systems (Linux)

This course uses Northwestern's Social Sciences Computing Cluster (SSCC) servers for accessing and manipulating some data. These servers are running the Linux operating system. They may be accessed from your computer over a VPN connection. If you don't have a VPN client on the computer you'll be using for this course, you'll need to get one and install it. See: http://www.it.northwestern.edu/oncampus/vpn/index.html

The SSCC is a cluster (two, actually) of servers in Evanston, Illinois with a wide variety of statistical software. It serves as a research facility for faculty and graduate students in the social sciences and business. The School of Professional Studies (SPS) has joined in the support of this facility so that its graduate students (especially those in Predictive Analytics) have access to the extensive software that it provides. General information about the SSCC is available at http://sscc.northwestern.edu.

As an MSPA student, you should have an SSCC user account. If you do not yet have one, go to http://www.it.northwestern.edu/research/user-services/sscc/overview.html
and use the link to request an account.

SSCC accounts are *not* associated with individual courses or instructors. Your SSCC account is tied to your network identity. SPS Predictive Analytics has its own log-in host: dornick. Files can be uploaded and downloaded using software tools employing secure file transfer protocol (SFTP). Filezilla, available for PC/Windows, Mac/OSX, and Linux computers, is a free public-domain tool for file transfer. Another file transfer tool is "CyberDuck." Additional information about file transfers is available at http://www.it.northwestern.edu/research/sscc/filetransfer.html.

Using the SSCC requires using the Linux operating system. The SSCC's servers are running the Red Hat Linux distribution and current stable versions of PostgreSQL and other database systems.

In addition to the *Linux Pocket Guide* used as a reference book, there are many sources for learning more about Linux. Here is a good introduction to the Linux operating system:
Ward, D. (2015). *How Linux Works: What Every Superuser Should Know* (2nd ed.). San Francisco: No Starch Press. [ISBN-13: 978-1-59327-567-6]

Software: sqlite

Some of the exercises and assignments in this course may use **sqlite**. **sqlite** is a widely used, public domain, serverless SQL database engine that you can interact with from Python, from other computing environments and languages, and from the command line. It's a small, easy to install and to use, engine that's suitable for many kinds of applications. Linux and Mac OS X distributions almost always include it by default. Windows, usually not. You can download it from http://www.sqlite.org.

A Word About Code Used in this Course, and What's Required for Assignments
Some example code is available in readings and online. Some may also be provided with respect to particular problems and tasks. It is the responsibility of each student to produce, test, and use the code needed to do their assignments. The code is a key part of a student's graded work product.

Evaluation

Students complete individual exercises, "GrExercises," that count towards their course grade. They also participate in weekly on-line discussions. Their participation is graded, and is also evaluated by other students at the end of the term in a simple, peer-review process.

Student deliverables that count towards the course grade, which is based on a total of 600 possible points, are as follows. Assignments are due and discussion board participation must be completed at the end of different weeks of the term, as follows:

- Week 2. GrEx 1 (worth 110 points)
- Week 4. GrEx 2 (110 points)
- Week 6. GrEx 3 (110 points)
- Week 8. GrEx 4 (110 points)
- Week 9. Peer reviews of course participation (60 points)
- Weeks 1–10 Discussion Board Participation, including the D&E tasks (100 points)

Total possible points, excluding any extra credit points: **600**

An extra credit assignment may be made available for completion during the ninth week of the course.

Grading Scale

A	= 95%–100% (570–600 pts.)
A-	= 90%–94% (540–569 pts.)
B+	= 87%–89% (522–539 pts.)
B	= 83%–86% (498–521 pts.)
B-	= 80%–82% (480–497 pts.)
C+	= 77%–79% (462–479 pts.)
C	= 73%–76% (438–461 pts.)
C-	= 70%–72% (420–437 pts.)
F	= 0%–69% (0–419 pts.)

Discussion Board Participation

The purpose of the discussion boards is to allow students to freely exchange ideas in their role as members of this course's learning community. Active and frequent participation is required. Frequency is important, as is content. All posted work should be original. Post Web links to the work of others (credit where credit is due, you know); do not post images or files obtained from the others. Please remember to cite sources. Discussion board activity is graded week-by-week, with the end of each week being Sunday at 11:55 p.m. U.S. Central Time.

Discussion grading is based on participation and student-to-student interaction. During the 9th week of the course, each student's contributions to discussions are evaluated by his or her peers. These evaluations count towards the course grade as indicated above.

Question-and-answer (Q and A) forums have been set up for questions about course procedures, individual assignments (these are the “Huddles”), and general Python and programming issues. Direct questions for me (your instructor) and our TA about assignments, using software, etc., should be posted to the forums, and *not* emailed directly to us. But do email me directly about personal issues at lynd.bacon@northwestern.edu.

Attendance

This course doesn’t meet at a particular time each week. All course goals, weekly learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), participation in threaded discussion boards is required, graded and also peer-reviewed, and paramount to success in this class. Real-time participation in any scheduled online synchronous (“Sync”) meetings is optional. While your attendance is highly encouraged, it is not required and you will not be graded on your attendance or participation in these sessions. *Students are expected to review session recordings*, however, as they may include content that aids in the successful completion of the course requirements. Students are responsible for what is covered in this course's Sync sessions.

Two (2) Sync sessions are on this course's calendar. The sessions for this course usually run from 1 to 1.5 hours in length. Real-time participation in them is optional. They are recorded for subsequent viewing.

Late Work

All assignments must be submitted in Canvas before the assigned due date and time. Assignments turned after their submission deadlines will lose 25% of the possible points for each 24 hours late. That is, an assignment turned in on the day after it is due will only be worth up to 75% of full credit. An assignment turned in on the second day after it is due will be worth at most 50% of full credit. On the third day, only 25% of full credit. The exceptions are Exercise 4 (the last assignment, which is due at the end of the 8th week), and the peer review ratings (due at the end of the 9th week). No credit will be given for these if they are turned in late.

Students can request a one day, penalty free extension for up to two (2) assignments, with the exception of for Exercise 4 and the peer review ratings. Extensions must be requested at least 24 hours before an assignment is due (that is, by 11:55 p.m. U.S. Central Time on Saturday, the day before) by emailing me directly at lynd.bacon@northwestern.edu.

In the case of an unforeseen event like a death in the immediate family, a hospitalization, or a housing catastrophe resulting from a natural disaster, the student should email me directly at lynd.bacon@northwestern.edu as soon as possible.

Study Groups and Tutors

Many students find working on course assignments and content with others to be helpful. Students in this course are encouraged to participate in study groups. Each student must submit their own work on assignments and for other individual course requirements. Students taking this course who have been rather new to programming in general or to Python have sometimes benefitted from the help of a tutor. The MPSA program does not provide tutors or information about tutors, so if you want to avail yourself of one, it'll be up to you find one. You might try posting to one of the MSPA LinkedIn groups to see if you can get suggestions or recommendations from fellow students or from alumni.

Succeeding in 420

Here are some tips for success in this course. First, turn in all required work on time. Works submitted late is discounted rapidly the later it is turned in. Second, participate early and often in each and every discussion. Third, help others as opportunities arise. Fourth, ask questions about assignments and exercises on Canvas (and not by emailing me directly) so that others can benefit by our consideration of them. Fifth, be sure to address every question or issue in each assignment that you submit. Sixth, consider forming or joining a study group. This should be particularly important if you are new to Python or to using the Linux operating system.

Engaging early and often in the weekly discussions will benefit both you and your classmates. Don't underestimate the importance of the discussions and the D&E tasks to how much you benefit from taking 420. The sooner you participate each week, the greater the benefit will be. Also, don't forget that your engagement with other students counts towards your course grade, and it will be evaluated by them. (And you will evaluate theirs.)

Other Processes and Policies

Please refer to your SPS student handbook at <www.sps.northwestern.edu/grad/information/handbook.cfm> for additional course and program processes and policies, and information about academic integrity.

## Course Schedule

Important Note: Changes may occur to the syllabus at the instructor's discretion. When changes are made, students will be notified via an announcement in Canvas or in email.

Week 1: Introducing Software and Systems (mo 06/19 – su 06/25)

Learning Objectives

After this week the student will be able to:

- Define the role of an operating system.
- Identify properties of open-source scripting language.
- Access remote computers, analytics and database servers.
- Install an integrated development environment (IDE) for editing and executing programs/scripts.

Course Content

Textbook Readings

Lubanovic, B. (2015):

Chapter 1: A Taste of Py (pages 1–14) and Chapter 2: Py Ingredients: Numbers, Strings, and Variables (pages 15–39)

McKinney, W. (2013):

Chapter 1: Preliminaries (pages 1–6) [Ignore the discussion about installation because you are (probably) using Enthought Canopy or Continuum Anaconda.] and Chapter 3: IPython: An Interactive Computing and Development Environment (pages 45–78)

Course Reserves

Connolly, T. M. and Begg, C. E. (2015). *Database Systems: A Practical Approach to Design, Implementation, and Management* (6th ed.). Upper Saddle River, N.J.: Pearson. [ISBN-13: 978-0-13-294326-0] Chapter 1: Introduction to Databases (pages 3–33)

Gift, N. and Jones, J. M. (2008). *Python for Unix and Linux System Administrators: Efficient Problem Solving with Python.* Sebastopol, Calif.: O'Reilly. (Chapter 2: IPython, pages 21–69.) [ISBN-13: 978-0-596-51582-9]

Ward, D. (2015). *How Linux Works: What Every Superuser Should Know* (2nd ed.). San Francisco: No Starch Press. [ISBN-13: 978-1-59327-567-6]
Chapter 1: The Big Picture (pages 1–10) and
Chapter 2: Basic Commands and Directory Hierarchy (pages 11–43)

Software Start-ups

Download and install Enthought Canopy and/or Continuum Anaconda. Download and install Filezilla or CyberDuck.

If you don't already have sqlite on your computer, download it from http://www.sqlite.org and install it.

Prepare personal computer for working with the Social Sciences Computing Cluster (SSCC). You'll find the note in the file sscc_sps_tutorial_1_2_3_4_5.pdf provided in Resources on Canvas to be (somewhat) helpful for some of the assignments. (This document is a little gnarly and hard to work through, but you'll need it for at least one of the graded assignments.) You should also be able to find on Canvas a “SSCC Cheat Sheet” pdf document which is a somewhat kinder and gentler introduction to accessing the SSCC. Additional content regarding using the databases on the SSCC will likely be provided during the term.

Download and install a VPN client if you don’t already have one. See if you can use it to connect to the SSCC using the guidance provided in the tutorial or Cheat Sheet document, above.

Weekly Discussions on Canvas

Participating by both posting and responding to other students' comments is required. You must do so before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

None.

Don't delay on installing the software you need, and on working through the readings for this week that are the most useful for you.

Sync Session

Section 55: Tues 06/20 7:00 p.m. U.S. Central Time (a Blue Jeans webinar invitation will be sent)

Section 58: Thur 06/22 7:00 p.m. U.S. Central Time (a Blue Jeans webinar invitation will be sent)

Week 2: Working with Files (mo 06/26 – su 07/02)

Learning Objectives

After this week the student will be able to:

- Distinguish among file formats: plain text, comma-delimited text, JSON, XML.
- Access files within hierarchical structures on remote and local systems.
- Read and write files on remote and local systems.
- Transform from one file format to another on remote and local systems.

Course Content

Textbook Reading

Lubanovic, B. (2015):

> Chapter 3: Py Filling: Lists, Tuples, Dictionaries, and Sets (pages 41–67) and Chapter 8: Data Has to Go Somewhere (pages 173–193 up to section on relational databases)

McKinney, W. (2013):

> Chapter 5: Getting Started with pandas (pages 111–154) and Chapter 6: Data Loading, Storage, and File Formats (pages 155–174 up to section on interacting with databases)

Course Reserves

Q. Ethan McCallan. (2012). *Bad Data Handbook: Mapping the World of Data Problems.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-32188-8]

> Chapter 12 (pages 151–162) When Databases Attack: A Guide to When to Stick to Files.

Weekly Discussions on Canvas

Don't forget! Your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

GrExercise 1 is due Sunday, July 2, at 11:55 p.m. U.S. Central Time

Sync Session

None

Week 3: Understanding Relational Databases (mo 07/03 – su 07/09)

Learning Objectives

After this week the student will be able to:

- Define relational database terms: table, record, schema, index, key, view, normalize.
- Describe the functions that a database management system should provide.
- Explain why relational database systems are well suited for transaction processing.
- Access data from a relational database using an administrative shell.

Course Content

Textbook Reading

Lubanovic, B. (2015):

> Chapter 4: Py Crust: Code Structures (pages 69–107), Chapter 8: Data Has to Go Somewhere (pages 193–216), and Chapter 10: Systems (pages 241–259)

McKinney, W. (2013):

> Chapter 6: Data Loading, Storage, and File Formats (pages 174–176)

Course Reserves

Connolly, T. M. and Begg, C. E. (2015). *Database Systems: A Practical Approach to Design, Implementation, and Management* (6th ed.). Upper Saddle River, N.J.: Pearson. [ISBN-13: 978-0-13-294326-0]

> Chapter 4: The Relational Model (pages 101–118) and Chapter 22: Transaction Management (pages 619–623 defining transactions and their properties)

Course Reserves (Recommended)

Obe, R. and Hsu, L. (2012). *PostgreSQL Up and Running: A Practical Guide to the Advanced Open Source Database.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-32633-3]

> Preface (pages ix–xii) and Chapter 1: The Basics (pages 1–8)

Worsley, J. C. and Drake, J. D. (2002). *Practical PostgreSQL.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-156592846-6]

> Chapter 3: Understanding SQL (pages 33–89)

Weekly Discussions on Canvas

You know the drill: Your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

None.

Sync Session

None.

Week 4: Accessing and Manipulating Relational Data (mo 07/10 – su 07/16)

Learning Objectives

After this week the student will be able to:

- Access data from a relational database using an object-oriented scripting language.
- Define structured query language (SQL) and explain its importance in working with databases.
- Select data from a relational database based on particular criteria.
- Join database tables to form views and tables.

Course Content

Textbook Readings

Lubanovic, B. (2015):

Chapter 5: Py Boxes: Modules, Packages, and Programs (pages 109–122)

Course Reserves (Recommended)

Hellmann, D. (2011). *The Python Standard Library by Example.* Upper Saddle River, N.J.: Pearson/Addison-Wesley. [ISBN-13: 978-0-321-76734-9]

Chapter 7: Data Persistence and Exchange (pages 333–420)

Worsley, J. C. and Drake, J. D. (2002). *Practical PostgreSQL.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-156592846-6]

Chapter 4: Using SQL with PostgreSQL (pages 91–155)

Weekly Discussions on Canvas

You know the drill: Your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

GrExercise 2 is due Sunday, July 16, at 11:55 p.m. U.S. Central Time

Sync Session

None

Week 5: Moving Beyond Relational Databases (mo 07/17 – su 07/23)

Learning Objectives

After this week the student will be able to:

- List reasons for moving beyond relational database systems to NoSQL systems.
- Explain what is meant by a distributed file system and MapReduce.
- Access data from a NoSQL database using an administrative shell.
- Access data from a NoSQL database using an object-oriented scripting language.

Course Content

Textbook Reading

Lubanovic, B. (2015):

Chapter 6: Oh, Oh: Objects and Classes (pages 123–143)

Course Reserves

Gheorghe, R., Hinman, M.L., and Russo, R. (2016). *Elasticsearch in Action.* Shetler Island, N.Y.: Manning. [ISBN-13: 978-1617291623]:

Chapter 1: Introducing Elasticsearch (pages 3–19)
Chapter 2: Diving into Functionality (pages 20–52)

Chodorow, K. (2013). *MongoDB: The Definitive Guide* (2nd ed). Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-34468-9]

Chapter 1: Introduction (pages 3–5),
Chapter 2: Getting Started (pages 7–28), and
Chapter 4: Querying (pages 53–77)

Franks, B. (2012). *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics,* New York: Wiley. [ISBN-13: 978-1-118-20878-6]

Chapter 1: What Is Big Data and Why Does It Matter (pages 3–27) and
Chapter 4: The Evolution of Analytic Scalability (pages 87–119)

Course Reserves (Recommended)

Dean, J. and Ghemaway, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM,* 51:1, 107–113.

Rajaraman, A. and Ullman, J. D. (2012). *Mining of Massive Datasets.* Cambridge UK: Cambridge University Press. [ISBN-13: 978-1-107-01535-7] Chapter 2: Large-Scale File Systems and Map-Reduce (pages 18–52)

Weekly Discussions on Canvas

Your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

None.

Sync Session

None.

Week 6: Accessing and Manipulating Text Data (mo 07/24 – su 07/30)

Learning Objectives

After this week the student will be able to:

- Access and manipulate unstructured and semi-structured text data files.
- Access and manipulate text data within a relational database system.
- Access and manipulate text data within a NoSQL database system.
- Parse text data with regular expressions in an object-oriented scripting language.

Course Content

Textbook Reading

Lubanovic, B. (2015);

Chapter 7: Mangle Data Like a Pro (pages 145–171)

Course Reserves

Levy, J. (2011). Bad Data Lurking in Plain Text. In Q. Ethan McCallan, ed.,*Bad Data Handbook: Mapping the World of Data Problems.* Sebastopol, Calif.: O'Reilly. [ISBN-13: 978-1-449-32188-8]

Chapter 4 (pages 53–68)

Course Reserves (Recommended)

Hellmann, D. (2011). *The Python Standard Library by Example.* Upper Saddle River, N.J.: Pearson/Addison-Wesley. [ISBN-13: 978-0-321-76734-9] Chapter 1: Text (pages 3–68)

Weekly Discussions on Canvas

Don't let up, now, because it counts towards your course grade. Your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

GrExercise 3 is due Sunday, July 30, at 11:55 p.m. U.S. Central Time

Sync Session

Section 55: Tues 07/25, beginning at 7:00 p.m. U.S. Central Time

Section 58: Thur 07/27, beginning at 7:00 p.m. U.S. Central Time

Week 7: Selecting and Sampling Data (mo 07/31 – su 08/06)

Learning Objectives

After this week the student will be able to:

- Define sampling terms: target population, sampling frame, sample, and representative sample.
- Distinguish among alternative forms of sampling, including random sampling, stratified sampling, and cluster sampling.
- Select data from databases following a sampling scheme.
- Address problems of under-coverage and sampling bias.

Course Content

Textbook Reading

McKinney, W. (2013):

Chapter 7: Data Wrangling: Clean, Transform, Merge, Reshape (pages 177–217)

Osborne, J. W. (2013):

Chapter 1: Why Data Cleaning Is Important: Debunking the Myth of Robustness (pages 1–16), Chapter 2: Power and Planning for Data Collection: Debunking the Myth of Adequate Power (pages 19–41), Chapter 3: Being True to the Target Population: Debunking the Myth of Representativeness (pages 43–69), and Chapter 4: Using Large Data Sets with Probability Sampling Frameworks: Debunking the Myth of Equality (pages 71–83)

Course Reserves (Recommended)

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology* (2nd ed). New York: Wiley. [ISBN-13: 978-0-470-46546-2]

Chapter 3: Target Populations, Sampling, Frames, and Coverage Error (pages 69–95),

Weekly Discussions on Canvas

As you surely know by now, your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

None.

Sync Session

None.

Week 8: Cleaning Data (mo 08/07 – su 08/13)

Learning Objectives

After this week the student will be able to:

- Identify bad data problems.
- Clean and update data items using an object-oriented scripting language.
- Screen data for potential problems, identifying outliers and miscoded data using an object-oriented scripting language.
- Address problems of missing data in surveys and databases.

Course Content

Textbook Reading

McKinney, W. (2013):

> Chapter 7: Data Wrangling: Clean, Transform, Merge, Reshape (pages 177–217) [Repeat of reading from Week 7.]

Osborne, J. W. (2013):

> Chapter 5: Screening Data for Potential Problems: Debunking the Myth of Perfect Data (pages 87–104), Chapter 6: Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness (pages 105–138), Chapter 7: Extreme and Influential Observations: Debunking the Myth of Equality (pages 139–168), Chapter 12: The Special Challenge of Cleaning Repeated Measures Data: Lots of Pits in Which to Fall (pages 253–259)

Course Reserves (Recommended)

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology* (2nd ed). New York: Wiley. [ISBN-13: 978-0-470-46546-2]

> Chapter 10: Postcollection Processing of Survey Data (pages 329–367)

Weekly Discussions on Canvas

There's no doubt about it! Your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

GrExercise 4 due Sunday, August 13, at 11:55 p.m. U.S. Central

Sync Session

None

Week 9: Transforming and Organizing Data (mo 08/14 – su 08/20)

Learning Objectives

After this week the student will be able to:

- Distinguish among cross-sectional, temporal/time series, spatial, panel, and spatio-temporal data.
- Perform data aggregation within an object-oriented scripting language.
- Recode and transform data fields using an object-oriented scripting language.
- Aggregate, group and reorganize data using an object-oriented scripting language.

Course Content

Textbook Readings

McKinney, W. (2013):

Chapter 9: Data Aggregation and Group Operations (pages 251–288) and
Chapter 10: Time Series (pages 289–328)

Osborne, J. W. (2013):

Chapter 8: Improving the Normality of Variables through Box-Cox Transformations: Debunking the Myth of Distributional Irrelevance (pages 169–190) and
Chapter 11: Why Dichotomizing Continuous Variables is Rarely a Good Practice: Debunking the Myth of Categorization (pages 231–252)

Weekly Discussions on Canvas

Verily we say on to you, your participation in both posting and responding to other students' comments is graded. You must participate before Sunday at 11:55 p.m. U.S. Central Time. Your required participation may include having to post code or results for discussion purposes.

Assignment

Your Peer Review Ratings are due Sunday, August 20 at 11:55 p.m. U.S. Central Time. Your ratings cannot be turned in late if they are to count towards your course grade. The reason is that your ratings are used to calculate other students' peer review scores for their course letter grades.

Sync Session

None

Week 10: Review (mo 08/21 – su 08/27)

Learning Objectives

No new learning objectives are introduced in this week. (Hooray!)

Course Content
None.

Weekly Discussions on Canvas

Same old, same old? Each week you've been required to participate in our discussions on Canvas. Your participation in both posting and responding to other students' comments has been graded. Participation graded through the end of the week, Sunday at 11:55 p.m.

No D&E this week. But your input to this week's discussion(s) is particularly important. And, it's worth something towards your course grade.

Assignment
None.

Sync Session
None