Syllabus
Big Data Analytics
[ANAT 594; BCHM 551; PHRM 684; PHSL 694]
Spring, 2015

Course: ***Big Data Analytics*** [ANAT 594; BCHM 551; PHRM 684; PHSL 694]
Thursday 10:00a–12:00p
Alumni Hall, Room 138

Course Director: Dr. Christopher G. Wilson
LLUMC Room A531
cgwilson@llu.edu, x15895

Teaching Assistant: Abby Dobyns
adobyns@llu.edu, x43832

Course Coordinator: Jacqueline Brower
Mortensen Hall #203
jbrower@llu.edu, x42755

Instructors:
Dr. Traci Marin, tmarin@llu.edu
Dr. Wilson Aruni, waruni@llu.edu, x42763
Dr. Charles Wang, chwang@llu.edu
Dr. Valeri Filippov, vfilippov@llu.edu, x4856

Important Dates
April 4th–First day of class
June 12th–Project presentations

Prerequisites
Undergraduate level biology, inorganic chemistry, organic chemistry and general physics. Previous experience with computer programming is a plus but not a requirement and a course in statistics would be helpful but is not required.

Objectives
This course aims to provide graduate and medical students with a broad understanding of the new field of Big Data Analytics in the context of biomedical research. The content is focused upon developing fundamental, real-world skills for performing data analytical work.

Educational Effectiveness
Educational success will be assessed by grades earned on quizzes, project work, and in a final presentation which will assess content mastery as well as problem-solving, data analysis and communication skills.

Tests and Grading
Performance in the class will be determined by quiz scores, project evaluation, coding examples, and a final class presentation for each group. Grading is Pass/Fail.

Attendance and Make-up work
All students are expected to attend the weekly classes as this is the time we will work on our projects as a group, we will have short lectures, and there will be time for interaction with the instructor and TA. Make-up quizzes may be prepared in an ad hoc fashion by the instructor so it is best to attend each session and take regular

quizzes. Attendance is strongly recommended, and is especially critical for those elements of class in which dialog and participation are integral. Students will be responsible for all material covered in the lectures as well as any reading material assigned. Th student will be responsible for being aware of announcements made in class and obtaining materials distributed during class.

Lifelong Learning
This course, an elective for Ph. D. degrees in Anatomy, Biochemistry, Microbiology, Pharmacology and Physiology and MS degrees in Biochemistry, Microbiology, Pharmacology and Physiology, is intended to serve as a gateway into informatics-based professions based on basic and applied biomedical sciences. Such professions require continual learning. Some professional organizations that may be of interest include the American Chemical Society, the American Association for the Advancement of Science, the American Association for Cancer Research, the American Society for Microbiology, The American Physiological Society, the American Society for Biochemistry and Molecular Biology, the American Society for Pharmacology and Experimental Therapeutics, the Society for Developmental Biology, the American Association of Anatomists, and Sigma Xi. A wide variety of scientific publications, most accessible through *PubMed* (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed), are also important tools for maintaining professional currency.

Academic Integrity
The scientific enterprise is highly dependent on the integrity and reliability of each of its components. Therefore, understanding and practicing scientific and academic integrity is essential for students at each and every phase of their education. Acts of dishonesty including theft, plagiarism, giving or obtaining information in examinations or other academic exercises, or knowingly giving false information are unacceptable. With regards to this class, quizzes are the responsibility of each individual student, and by turning in such work, the student is representing that piece of work as having been completed by himself/herself. Projects are, inherently, collaborative and your group will be responsible for maintaining a balance of work for each individual and you will be collectively held to a high standard of integrity. Violations of academic integrity will generally result in a score of zero for a quiz and may endanger the grade for group projects. Violations may also be taken to the dean for further disciplinary action. These actions may include, but not be limited to, academic probation or dismissal from the program. To view the Standards of Academic Conduct Policy please visit: http://www.llu.edu/llu/handbook/6r.htm.

Flexibility
The course syllabus provides a general plan for the course; deviations may be necessary and will almost certainly occur. If it becomes necessary to alter the dates for quizzes or project presentations/check-in, the changes will be announced in class as early as possible. The course director is the final arbiter and reserves the right to make the final decision when situations not described in this syllabus arise. Students are strongly advised to contact the course director for clarification before unusual circumstances occur.

Student Learning Outcomes
Student learning outcomes have been developed at both the University and Program levels. The *Big Data Analytics Class* addresses several of these outcomes, particularly SLOs #5 (Critical Thinking) and # 1 (Communication).

University-wide Student Learning Outcomes (SLOs) and Performance Indicators
Outcome 1—Written Communication: Students demonstrate effective written communication skills in English
Outcome 2—Oral Communication: Students demonstrate effective oral communication skills in English.
Outcome 3—Quantitative Reasoning: Students demonstrate the ability to reason and develop evidence-based decisions using numerical information.
Outcome 4—Information Literacy: Students demonstrate the ability to identify, locate, evaluate, utilize, and share information.
Outcome 5—Critical Thinking: Students demonstrate critical thinking through examination of ideas and evidence before formulating an opinion or conclusion.

Program Student Learning Outcomes
1. Students will demonstrate a broad knowledge of the biomedical sciences.
2. Students will demonstrate subject mastery in data handling, signal theory, and data visualization.
3. Students will interpret the current literature in their chosen discipline as it informs data analytics practice.
4. Students will make original contributions to the body of biomedical knowledge.
5. Students will demonstrate an understanding of the principles of scientific and professional ethics in regard to data handling—particularly in regard to data security and privacy.

Americans with Disabilities Act (ADA) Policy
If you are an individual with a certifiable disability and need to make a request for reasonable accommodation to fully participate in this class, please visit the Dean's Office of the School of Medicine. To view the Disability Accommodation Policy please go to: http://www.llu.edu/llu/handbook/6e.htm. Students with learning difficulties requesting modifications to the standard testing outlined in this syllabus must submit written approval for the requested accommodations to the course director a minimum of 1 week prior to the first examination.

Protected Health Information
The purpose of the Protected Health Information (PHI) policy is to provide guidance and establish clear expectations for students regarding the appropriate access to and use of PHI during course studies and related program activities. Under the Health Insurance Portability and Accountability Act (HIPAA), patient health information is protected. For further information, please go to: http://www.llu.edu/llu/students/documents/phi-guidelines.pdf.

Course material by week

I. Foundations 1 (3/30 – 4/3)
- A. Using text editors for writing/reading code (emacs/vim/geany/gedit)
- B. Using the *IPython* notebook for data exploration
- C. Using version control software to track changes (we will use *git*)
- D. Using the class Discussions on Canvas

II. Foundations 2 (4/6 – 4/10)
- A. *iPython/NumPy/SciPy* for data analysis
  1. Basic commands, magic commands
  2. Using the *IPython* notebook for reproducible data analysis and sharing
- B. Loading and evaluating data with *IPython*
  1. *Pandas* — The Python Data Analysis Library
  2. *PyEEG* — Loading EEG data from EDF+ files into *IPython* with *PyEEG*
  3. *PyHDF* — Loading Hierarchical Data Format files into *IPython* with *PyHDF*
- C. Data visualization tools accessible within Python
  1. *Matplotlib* for 2D visualization
  2. *Mayavi* for 3D visualization
- D. Basic signal theory

III. Analysis methods (4/13 –4/17)
- A. Basic signal theory overview
- B. Time-series data analysis
  1. Points, histograms, and bars (advanced use of *Matplotlib*)
- C. Non-linear analyses
  1. Dynamical systems analyses of data variability
     (Poincaré maps, deterministic and non-deterministic complexity)
  2. Information theory measures (entropy) of complexity using *PyEnt*
- D. Spectral measures
  1. Moving into the frequency domain from the time-domain
  2. Fast Fourier Transforms, time-varying spectrum (*SciPy/NumPy*)

3. Wavelets (*PyWavelets*)

IV. Sequence data (4/20 – 4/24)

1. Overview/foundations of sequence data
2. Exploring sequence data — Using *R/Bioconductor*
3. Understanding the differences between mRNA-Seq, gene-array, proteomics, and deep-sequencing
4. Visualizing data from gene/RNA arrays (heat maps, volcano plots, etc.)
5. Loading *R/Bioconductor* output into *IPython*

V. Large data set storage and retrieval (4/27 –5/29)

A. Basics of relational databases
B. To SQL or NOSQL?
   1. SQL-based relational databases (*Physio-MIMI* as an example case)
   2. NOSQL basics (*MongoDB*)
   3. *MongoDB* and *Hadoop/MapReduce* as Big Data platforms
C. Cloud storage vs. local NAS/servers and computing clusters
   1. Using parallel computational tools to speed your work in *IPython* (interfacing with *Hadoop/Pig/MapReduce*)
E. Adding data and analyses to biomedical databases in near real-time (*MongoDB*)
F. Meta-data and dynamic visualization (putting it all together to provide INFORMATION!)
   1. *PhysioMIMI* as a web-based interactive query platform (SQL-based)
   2. Ontology in the world of databases
   3. Using XML for tagging events and defining data relationships
   4. Securing patient data using *REDCap*

VI. Data integrity and security (6/1 – 6/5)

A. The *Health Insurance Portability and Accountability Act* (HIPAA) and what it means for data management
B. Data security best practices
C. De-identifying patient data (some of this will be taught in section I) (*PyEEG, PyHDF*)
D. Making data available to the public—Implications for data transparency and large-scale data mining

VII. Project Presentations (6/8 – 6/12)