principles of urban informatics/ CUSP/ Fall 2016/ Syllabus

This course will teach you the basis of data driven urban research. You will acquire basic computational skills, basic knowledge of statistical analysis, error analysis, good practises for handling data and big-data, and communication and visualization skills. After this class you should be able to formulate a question relevant to Urban Science, find an appropriate data to answer the question, prepare and analyze the data, get an answer, to whichever confidence level, and communicate your answer, and your confidence level in the answer. *At times you will have a hard time figuring out the solutions to problems. Remember that we admitted you because we believed you would have a positive influence on the class, and that being at CUSP can fulfill your potential as an Urban Scientist. Don't worry about how much you already know, especially do not compare it to what other students know. You may have the wrong perception of your skills, and of the skills of your classmates, and your strengths and the strengths of your background may lie in another set of skills, just as important for an Urban Scientist. We are here to help you develop the skills you do not yet have and strengthen the skills you already have. You are here because we want you to be here and believe in your potential.*

This course should serve as the basis for all the following classes, and your future projects. The course will be organized in a modular fashion, with guest lectures.

The instructors is: Dr. **Federica Bianco** (fb55@nyu.edu).
Office hours are Tu/F 2-4 PM (or by appointment).
The Lab TAs are:

Himanshu Kumawat (Tue) hk1953@nyu.edu
office hours: Tue/Wed 3-5
Saubreen Syedmajeed (Thu) ss9570@nyu.edu
office hours: Mon/Thu 2-4

Books:

The primary textbooks are:

- computing and coding: **Beginning Python Visualization, 2009**
- data analysis: **Statistics in a nutshell, S. Boslaugh, O'Reilly Media**
- visualizations: **Visualizations Analysis and Design, T. Munzer, 2014**

In addition, we will use the

- **Python for Data Analysis, W. McKinney, O'Reilly Media**
- **Interactive Data Visualization, S. Murray, O'Reilly Media**

Copies of these books are available at CUSP.

Each week you will attend one lecture and one lab session. **Attendance in lecture and lab is mandatory. Generally you cannot switch between the lab or class sections**, because they may be on different schedules and be different for different tracks. If you must, please consult me and the labs instructors to arrange a make-up session.

Grades are based on

- **5% pre-class questions**
- **10% labs performance and participation**
- **25% homeworks**
- **25% midterm**
- **35% final**

Weekly assignments will be handed out at the end of the class, and will be due before the first class of the following week. Please come to class on time: at the beginning of each class you will be handed a sheet of "Pre-class Questions" to be answered **before each lecture and before each lab**. You will have until 5 minutes into the class to answer them. The later you arrive at the class, or lab, theless time you will have to answer the questions. This will affect your homework grade. as described above. The questions will cover the material in the previous classes, and the reading assignments.

Late homeworks will not be accepted. A single 2-day exception is allowed throughout the semester, explicitly declare that you are going to use it, and do use it wisely. The lowest grade in the first half of the course (before midterm), and the lowest grade in the second half will be disregarded in assigning you a final grade. If you fail to turn in an assignment that will be a 0, and (likely) the lowest grade. This means you will lose the chance to disregard your worst performance. Homework will be exclusively received through github.

Homework projects **must be turned in as iPython notebooks by checking them into your github account in the PUI2016_<netID>/HW<hwnumber>_<netID>/** repo (unless otherwise stated). We encourage you to work in groups of up to 5 people, but as a collaborative project where different group members lead different aspects of the work. A statement to that extent **must be included in the README** (a

la). Midterm and Final will include aspects of the work developed in the homework sessions. Failing to actively participate in the homework will result in not being able to get the Midterm and Final done.

For the midterm and the final you are responsible for material in the labs, the reading, and the homework. **In preparing for the exams, use the homeworks as a guide to which material I believe is essential.**

Lecture and reading schedule (subjects to change as needed!):

1. Sept week 1 (9/6 & 9/8):
   Lecture: philosophy and good practices of data science: the flow chart of a data-driven project from idea to divulgation, the concepts of falsifiability, reproducibility, open science, the importance of version control, iPython Notebooks
   - Lab: command line tools. github repositories, setting up your environment, Python vs iPython, and iPython notebooks

2. Sept week 2: (09/20 & 09/22)
   Lecture: acquiring and preparing data (CSV, TSV, downloadable ascii files, basic SQL) merging data from different files, plotting histograms and scatter plots, data types incl ordinal, continuous, categorical data
   Lab: read and clean data, Citibikes, Pluto, Census, introduction to data structures (dictionaries, lists, arrays), style guides

3. Sept week 3: (9/13 & 9/15)
   Lecture: Introduction to the statistics, why everything is gaussian (...or not), bias, basic distributions, moments, Hypothesis testing (chi-square, z-test, p-value).
   Lab: basic statistics on Pluto, Census, Citibikes data, moment extraction, deviations from gaussianity/poissonity, histograms, proper binning.

4. Sept week 4: (09/28 & 10/1)
   Lecture: PDF/CDF, data dredging, sample of 1, correlation vs causality, error analysis, testing models (KS, anderson darling, KL divergence), basic plotting.
   Lab: hypothesis testing

5. Oct week 1: (10/04 & 10/6)
   Lecture: Likelihood, OLS, WLS, basic bayesian concepts (maybe), missing data, small data, Unstructured data, API.
   Lab: goodness of fit, geopandas

6. Oct week 2: (10/11 & 10/13)
Lecture: (time)-series techniques: smoothing, detrending, stationary, non-stationary, homeo- & hetero-scedastic noise, vectorization

7. Oct week 4: (10/18 & 10/20)
Lecture: Visualizations. Communication through visualizations, history, significance, good and bad visualization examples, what have we learnt since the 1800s?
Lab: a viz

Oct week 4 (10/25 & 10/17): Midterm

Nov week 1: (11/1 & 11/3) Guest Lecture: Prof. Julia Lane
Lecture: Data Hygiene
Lab: Data Hygiene

8. Nov week 2: (11/8 & 11/10)
Lecture: Spatial+Temporal Urban data (multidimensional data), clustering.
Lab: ZBP data

Nov week 3: (11/15 & 11/17) Guest Lecture
GeoPandas

9. Nov week 4/Dec week 1:(11/22 & 12/8)
Lecture: Spatial interpolation - Kriging

Nov week 4/Dec week 1: (11/29 & 12/1): Guest Lecture Prof Huy Vo
Databases: manipulation of large data

10. Dec week 1/2: (W 12/06 M 12/15)
Lecture: Spatial simulations: MonteCarlo Methods

The Final exam is cumulative: you are responsible for ALL OF THE MATERIAL.

Missing Topics:

- Graph Theory
- Audio vs Video data

- multiprocessing/multithreading
- Visualizations w D3
- GIS
- D3+GIS visualizations
- optimization techniques (gradient descent, annealing...)
- web/mobile coding
- designing experiments