A4834: Datamining the City

Culture, Urbanism, and Web 2.0

Instructor: Danil Nagy (dn2216@columbia.edu)
Meeting time: Thursdays, 3:00pm-5:00pm in 300 Buell North

"Cities have the capability of providing something for everybody, only because, and only when, they are created by everybody."

- *Jane Jacobs, "The Death and Life of Great American Cities"*

"Only by moving from cities to models and back again can we develop an appropriate understanding of their dynamics."

- *Michael Batty, "Cities and Complexity"*

"Telling the future, when it comes right down to it, is not solely a human yearning. It is the fundamental nature of any organism, and perhaps any complex system. Telling the future is what organisms are for."

- *Kevin Kelly, "Out of Control"*

Context:

The concurrent growth of computational capacity and new data sources has revolutionized the theory and practice of urbanism in the last few decades. In particular, the evolution of the internet as a seemingly endless store of personal, granular information has opened up completely new forms of data that can be used to measure the city. At the same time, advances in computing have created new models that can match the complexity of city systems, and translate this data into tools for understanding the city.

In 1961, Jane Jacobs cited cities as the example *par excellence* of organized complexity[1], driven by dynamic forces and a huge amount of variables we barely understand. However, while they are complex, cities are not random, and thus can theoretically be modeled and predicted. According to complexity theory, cities exist on 'the edge of chaos', far from equilibrium but able to be understood and modeled within restricted parts of space and time. While such predictive models were mere dreams at the time of Jacobs' writing, new sources of urban data, along with rapid advances in computation, are finally making such systems possible.

[1] Jane Jacobs, *The Death and Life of Great American Cities (1961)*

Scope:

This seminar will focus on developing strategies for datamining large datasets from the web and processing them spatially to derive new knowledge about the city. Lectures will provide students with a theoretical and historical basis for this kind of research, as well as training in the specific tools that will be used. The class will take a hands-on workshop approach to teach practical skills in basic programming, web scraping, big data, GIS, and visualization. The main tools will be *Python* and *QGIS*.

The context of the research will be the Pearl River Delta in China, the world's largest megalopolis containing at least 60 million people. The research will focus on urban issues that have been difficult to research using traditional data and tools, including migration, informal housing, and grey market economies.

Expectations:

Students will be provided with an initial set of data, but will be expected to gather and process additional data sources to supplement their thesis. The class will be divided into groups to produce 4-6 projects, with each group expected to fulfill the following requirements:

- Generate a proposal or hypothesis for research based in the Pearl River Delta
- Create custom scripts to collect, organize, and process this data
- Spatially analyze this data to create a model within a GIS
- Create provocative visualizations that demonstrate some conclusion about the chosen issue.

In addition to the group project, students will be graded based on their attendance, participation, and completion of several small workshop assignments according to the following rubric:

Class Participation	10%
Class Blog/Readings	10%
Individual Exercises (3)	30%
Final Group Project	50%

Schedule:

Session A – Introduction to Python, web scraping, data gathering and analysis

Week 1	Big data, micro-informatics, and the Pearl River Delta
Week 2	Python for basic data processing
Week 3	Python for data gathering, web scraping and API
Week 4	Basic QGIS for visualization
Week 5	Advanced QGIS I – processing and spatial statistics
Week 6	Advanced QGIS II – analysis using heat maps, interpolation, and time

Session B – Modeling, Machine Learning, and Advanced Visualization

Week 7	Graph theory for urban network analysis
Week 8	Introduction to Machine Learning: classification and regression
Week 9	Advanced Machine learning: unsupervised learning, clustering and trees
Week 10	Interactive visualization in Tilemill and Mapbox
Week 11	Advanced visualization, animation, d3
Week 12	Final Review, presentations

** Since the class will be based around a group research project, students are highly encouraged to register for both sessions A and B*