

An initiative of the Computation Institute of the University of Chicago and Argonne National Laboratory



Computation & Data-Enabled Urban Design, Planning, and Operation

November 6, 2013
ULI Fall Meeting
Chicago, Illinois

Charlie Catlett, Senior Computer Scientist, Argonne National Laboratory
Senior Fellow, Computation Institute of the University of Chicago and Argonne National
Laboratory
Director, UrbanCCD
catlett@anl.gov

www.urbanccd.org

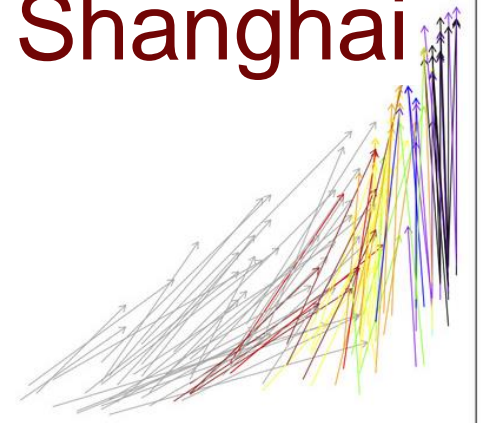
Chicago



Delhi



Shanghai



→
1999-2000 Land
Area Increase

↑
1999-2000 Building
Height Increase

Cities in India and Southeast Asia are Growing at Unprecedented Rates

Frolking S, T Milliman, KC Seto, MA Friedl. 2013. A global fingerprint of macro-scale change in urban 2-D and 3-D structure from 1999 to 2009, Environ. Res. Lett.

An aerial photograph showing a large, rectangular development site on the Chicago Lakeside. The site is bordered by a dense residential neighborhood to the west and a large body of water to the east. The site itself is divided into several sections, including a large rectangular area with a light-colored, sandy or dirt ground, and a central area with a winding canal or waterway. The surrounding cityscape is visible in the background, with the Chicago skyline and Lake Michigan in the distance.

Chicago Lakeside Development
Chicago, Illinois
2013

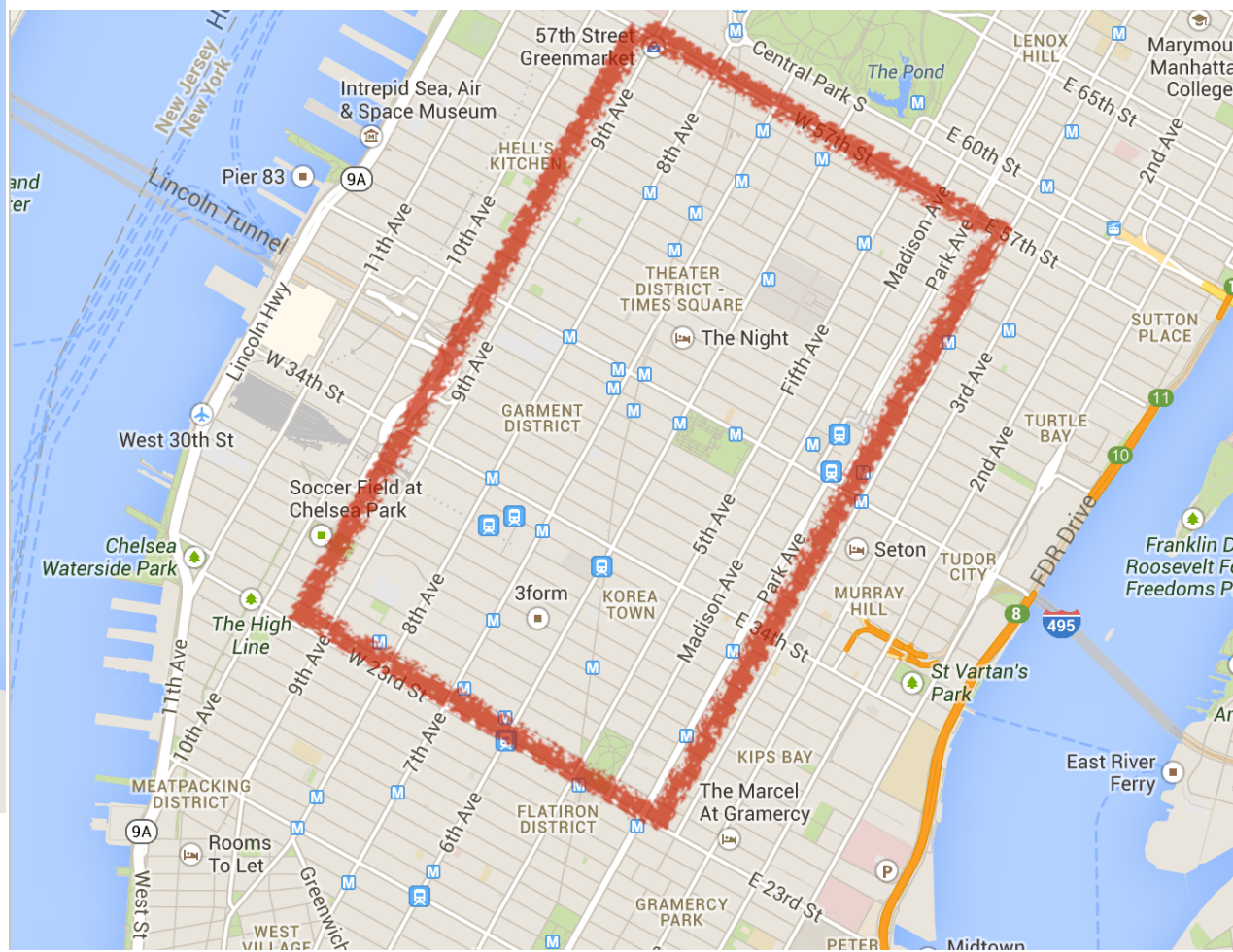
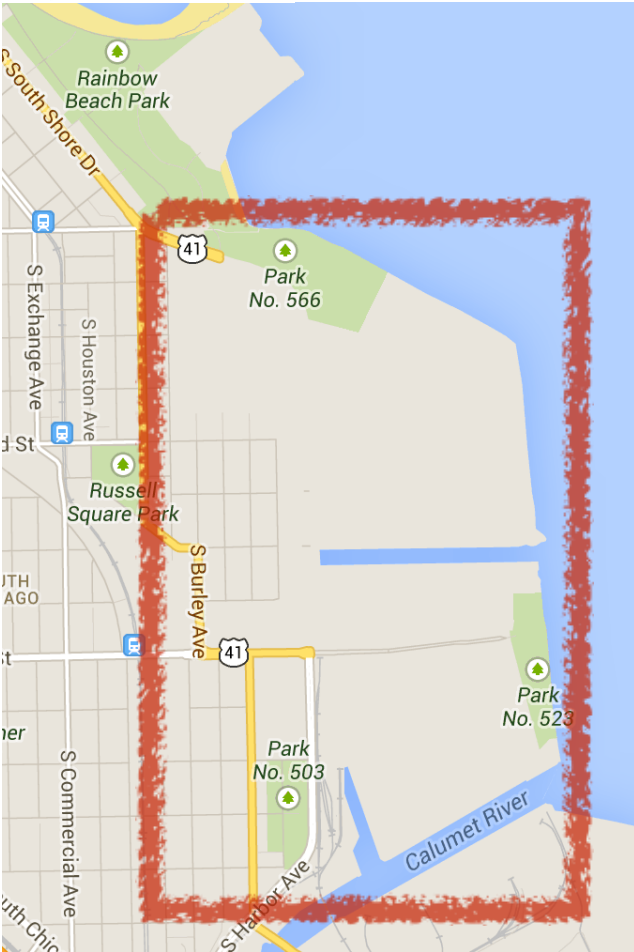


US Steel Plant
Chicago, Illinois
1945

Environment / Infrastructure / People



Chicago Lakeside Development
Chicago, Illinois
2040



Design



Analyze

Independent
consultant studies

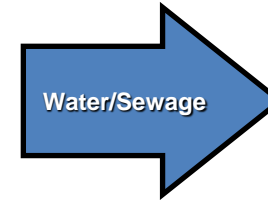
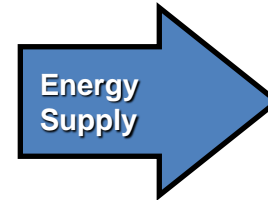


weeks to months

Spreadsheets

*History-based
models*

Plan



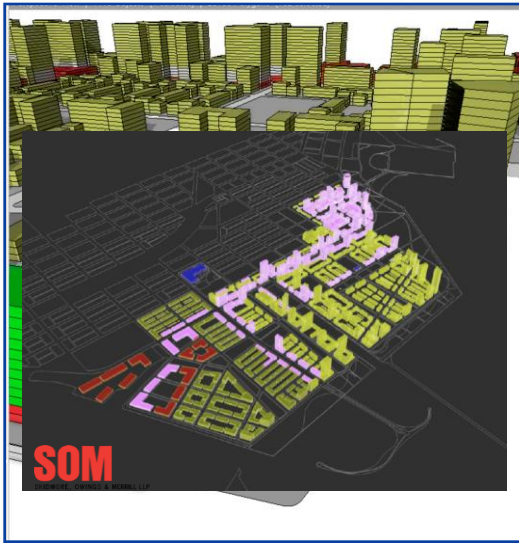
Decide

Pictures

Charts

Reports

Design



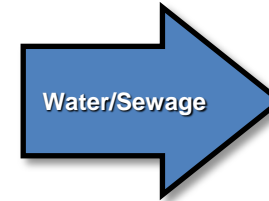
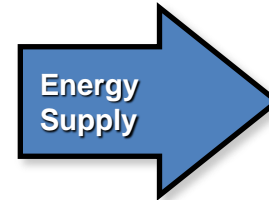
Analyze



hours to days



Plan

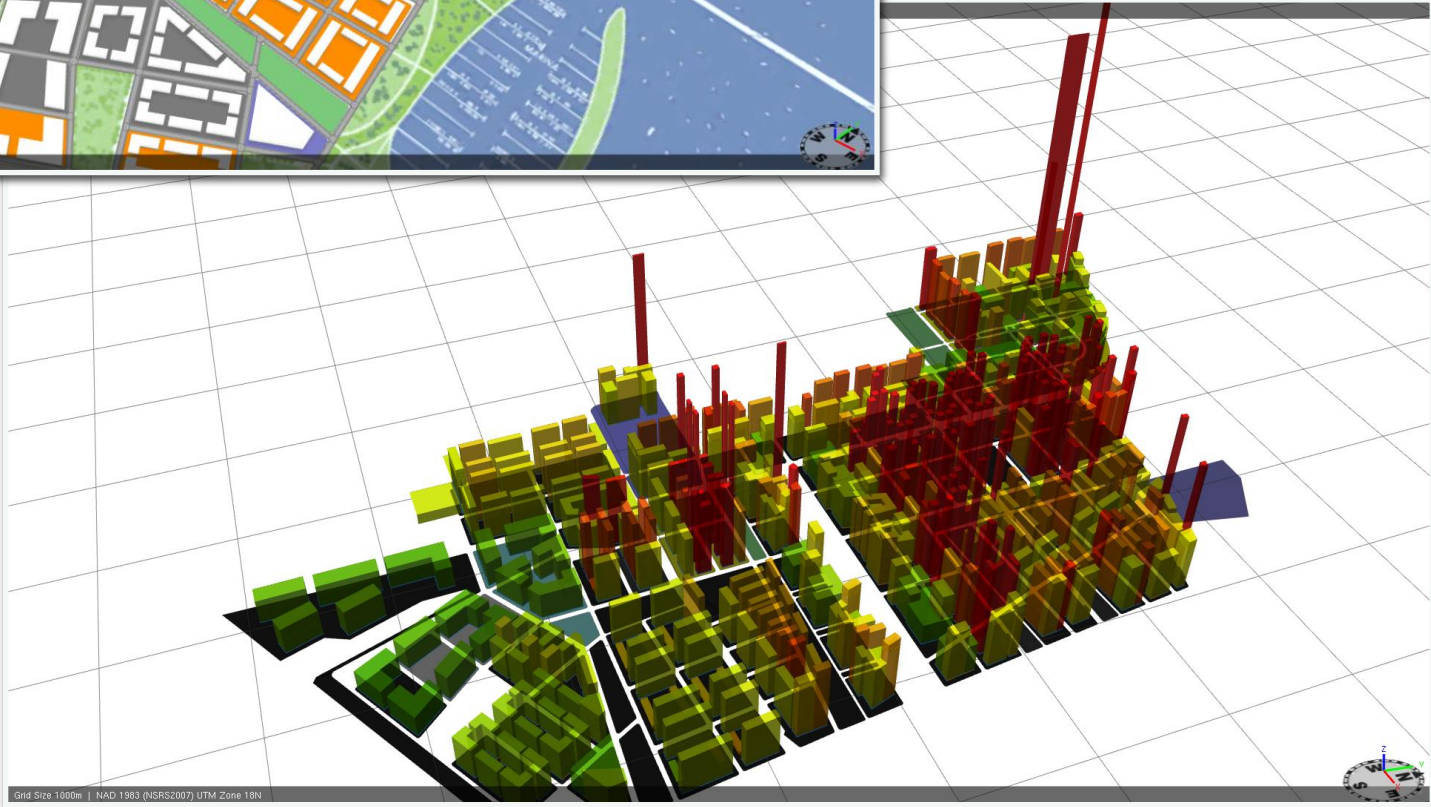


Decide

Perspective View | 43988 Objects (1 selected) | 87473 Polygons (1 selected)



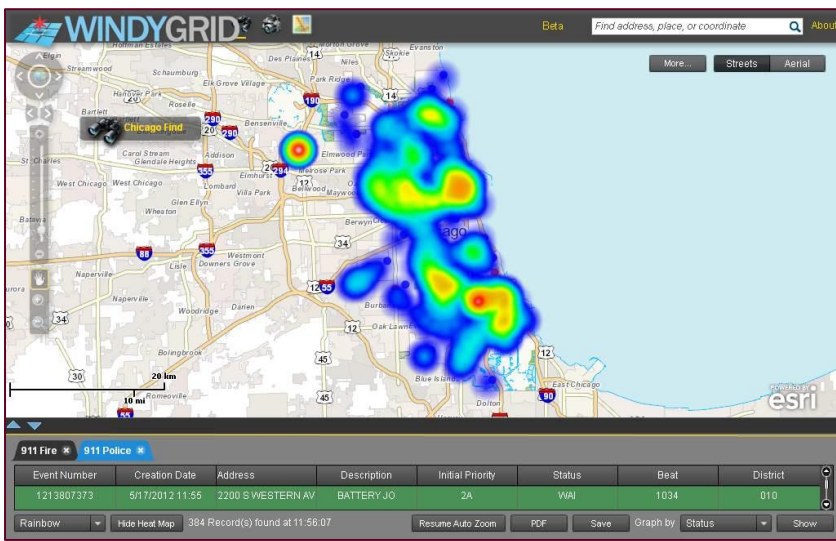
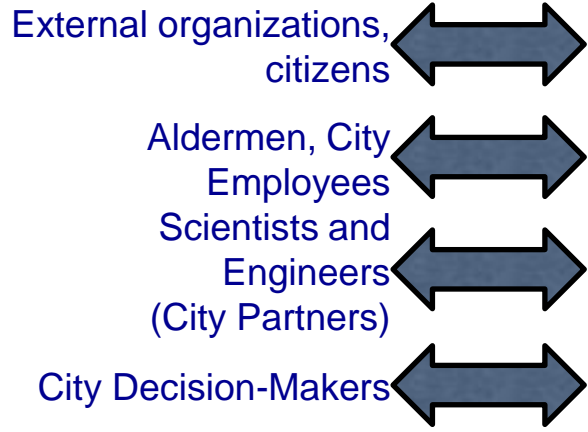
Grid Size 1000ft | NAD 1983 StatePlane Illinois East FIPS 1201 (US Feet)



Grid Size 1000m | NAD 1983 (NAD83) UTM Zone 18N



Data Access, Authorization, Privacy

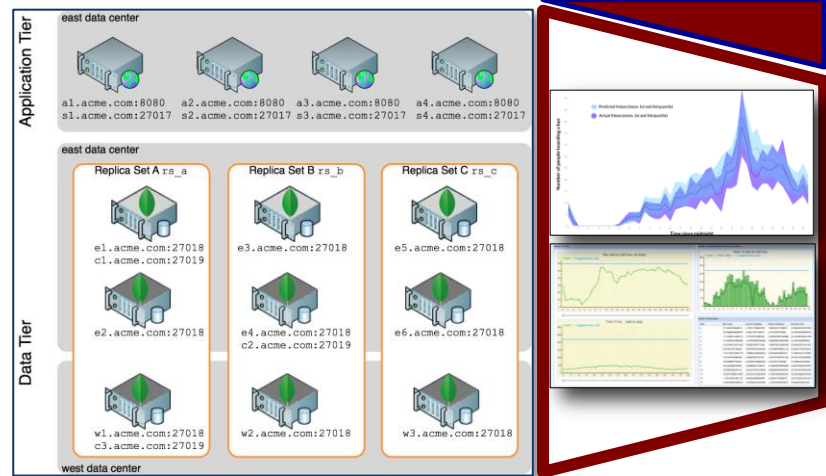


Visual Interaction, Mapping, Analysis Tools

Application Programming Interfaces (API)

Automate Continuous Data Analytics

Scalable Data Management

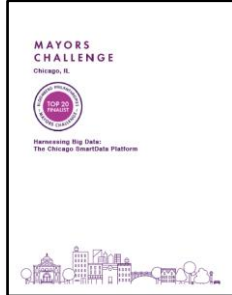


Data Sources



City Databases Sensor Nets Video GPS Social Media

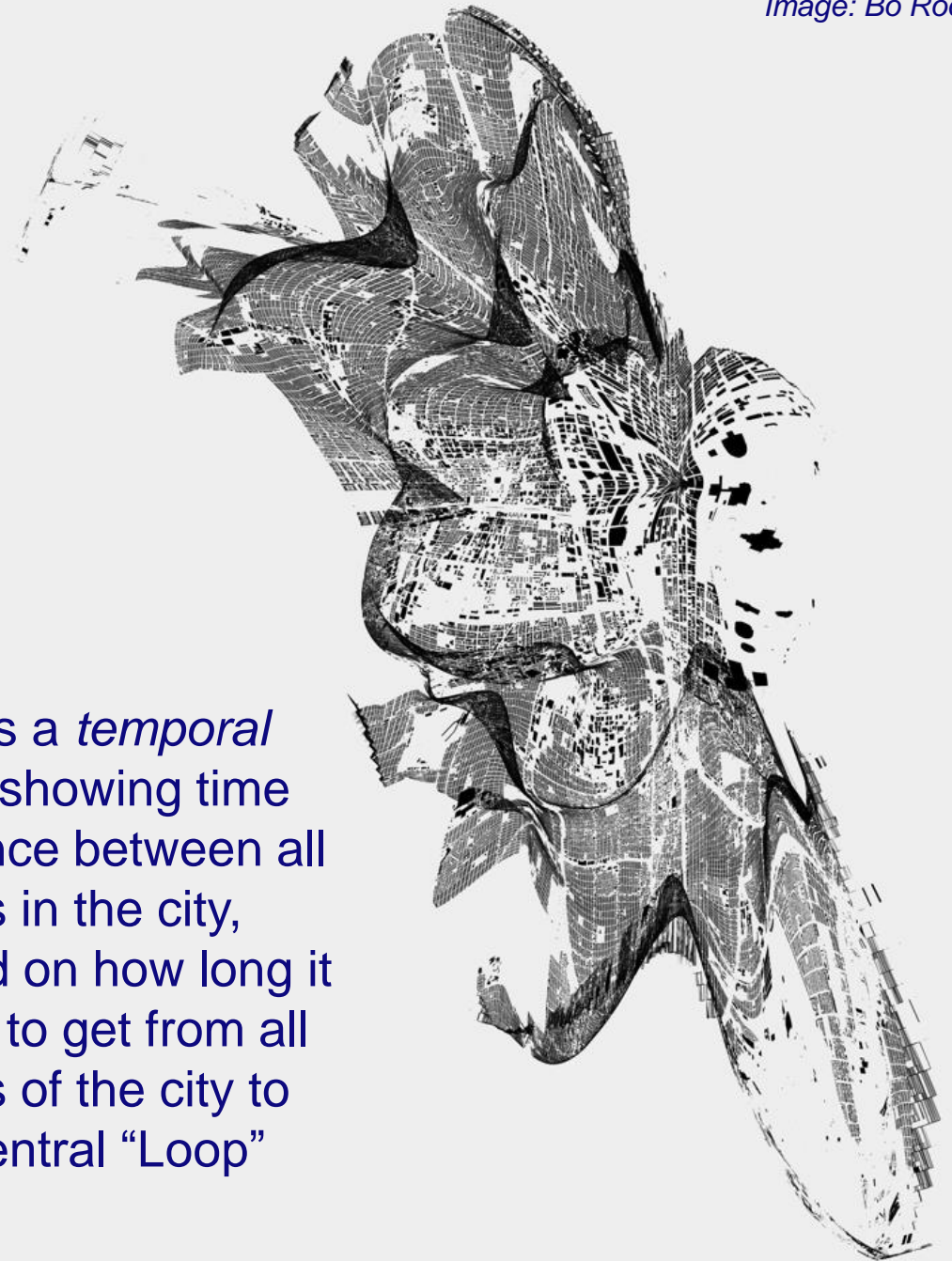
Initiative 14: Increase and improve City data.



More powerful data and tools such as transportation transit estimates can enable scientists to look at cities in new ways.

A common map shows *spatial* distance between all points in the city.

This is a *temporal* map, showing time distance between all points in the city, based on how long it takes to get from all points of the city to the central “Loop” area.



Stanford University
University of California-Irvine
Arizona State University
University of Texas-Austin
University of Wisconsin-Madison
University of Texas-Austin
Instituto Tecnológico Autónomo de México

University of Chicago
University of Illinois-Chicago
University of Michigan
University of Alabama
Georgia Institute of Technology
Carnegie Mellon University
Cornell University
Israel Institute of Technology

University of Maryland
City University of New York
Columbia University
Yale University
Harvard University
Massachusetts Institute of Technology
University of Cambridge



Redshift Data Base
(copy of CTA's DB)



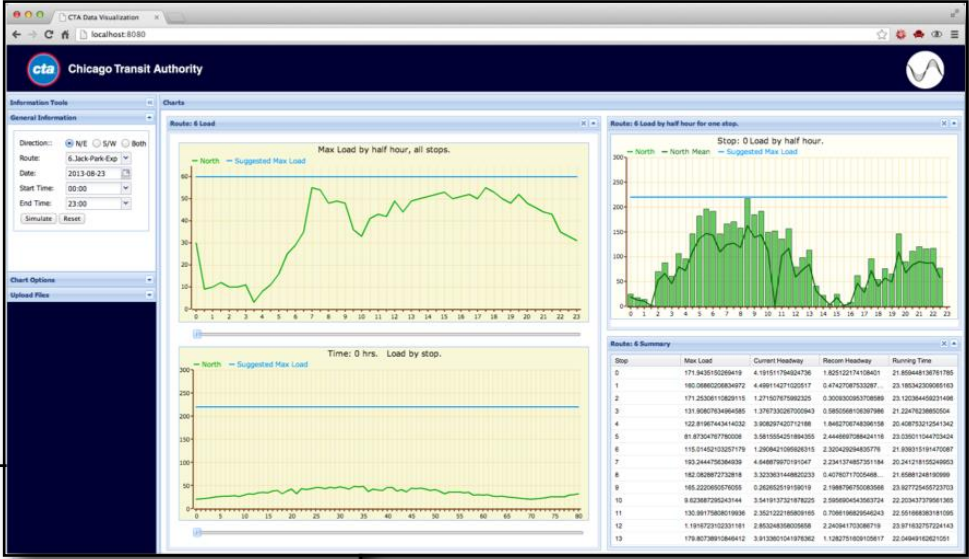
Model



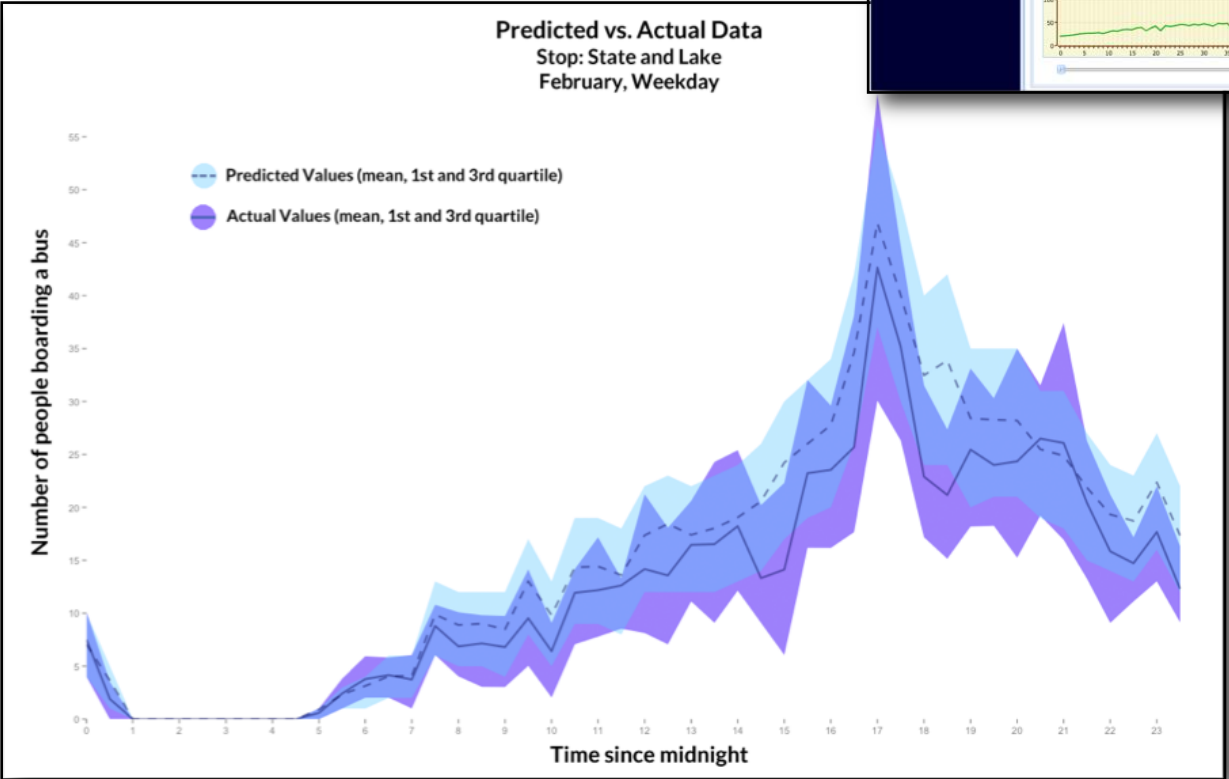
Simulation



Webapp



Predicted vs. Actual Data
Stop: State and Lake
February, Weekday



Jordan Bates, Andrés Akle Carranza, Walter Dempsey, David Sekora, Brandon Willard



Example: City of Chicago Proactive Intervention

Create "**Neighborhood Health Index**" - identify and quantify neighborhoods w.r.t. education, economics, public safety, and other domains. Look for leading indicators to neighborhood decline or revitalization.

Research Capabilities

Analytics, machine learning, and other techniques, guided by social, economic, education, health, and related research areas, to examine urban data as a means to define and measure neighborhoods and related functions.

Impacts on City Challenges

Early detection of at-risk neighborhoods w.r.t. crime, education, sustainable energy use, economics, employment, and other factors, enabling preventative vs reactive intervention.

Predict locations of abandoned buildings/vacant lots.

Predict restaurant failures based on food service reports, social media, and other data.

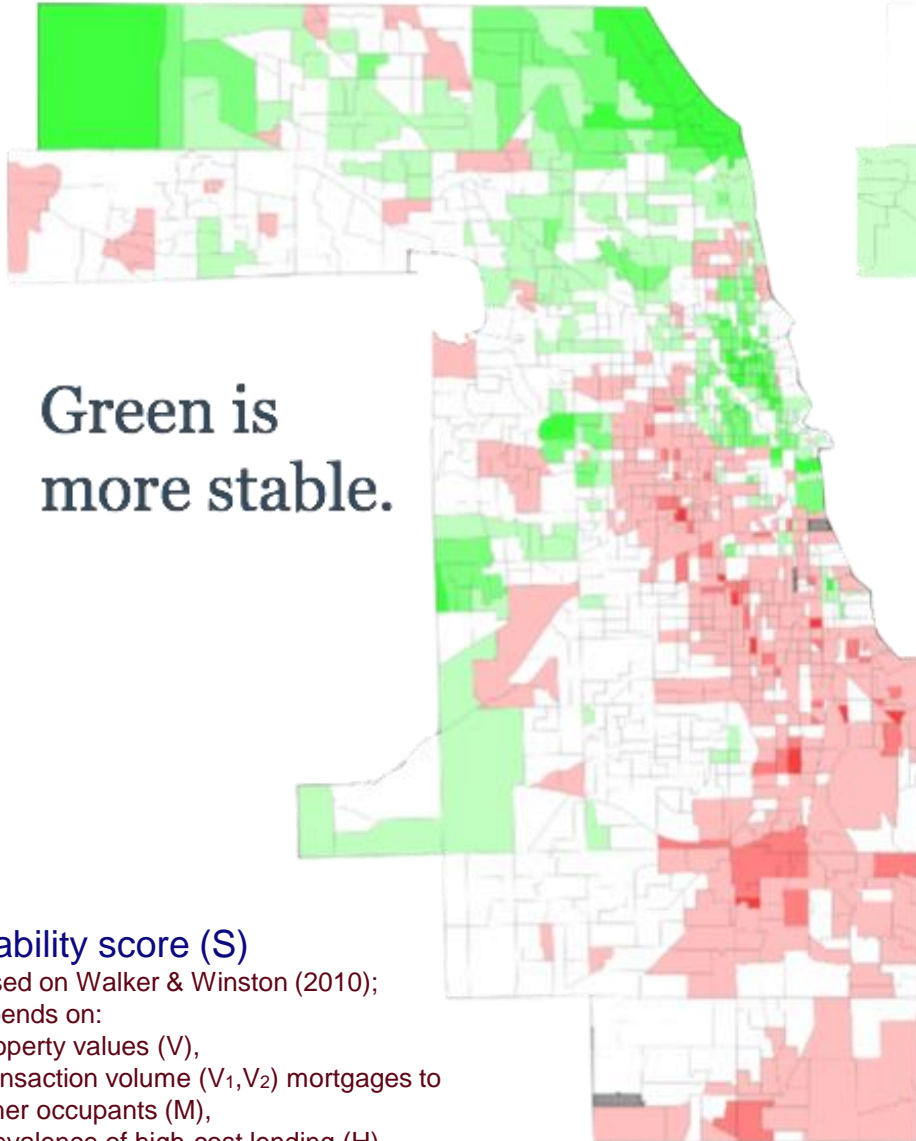
Estimate economic health ("micro-GNP") of neighborhoods and sub-neighborhoods.

Stability score

$$S = 0.6V + 0.3V_1 + 0.1V_2 + 0.2M - 0.4H$$

$$A = \int_{P/4}^{\infty} f(I) dI$$

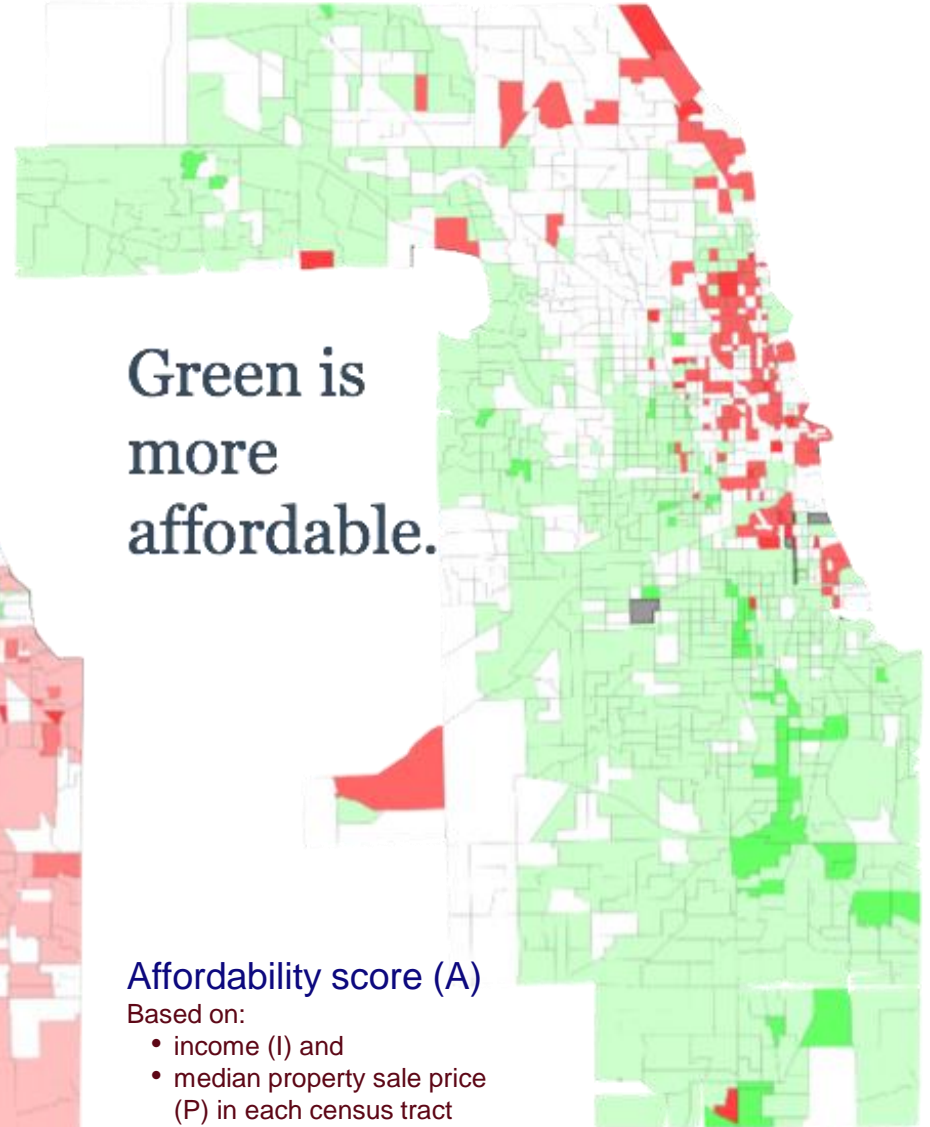
Affordability score



Stability score (S)

Based on Walker & Winston (2010); depends on:

- property values (V),
- transaction volume (V_1, V_2) mortgages to owner occupants (M),
- prevalence of high-cost lending (H).



Affordability score (A)

Based on:

- income (I) and
- median property sale price (P) in each census tract

Optimizing Public Transit Schedules to Reduce Crowding

The Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship 2013
 Jordan Bates (Arizona State University), Andrés Akle Carranza (ITAM, México), Walter Dempsey (University of Chicago), David Sekora (University of Chicago), Brandon Willard (University of Chicago)

Motivation

Crowding on buses is an acute transit issue that frustrates riders. The **Chicago Transit Authority (CTA)** is aware of the problem, and has launched a **crowding reduction initiative** to tackle the problem by reallocating bus service where it's needed most.

The simulation of passenger boardings and alightings incorporates time-dependent covariates:

- Hour:** Dependent half-hour categorical variables to help capture changes in passenger flow over the course of a day.
- Month:** These indicators allow us to incorporate seasonal trends
- Week/weekend:** This indicator allows us to capture both schedule and ridership changes from week to weekend.

Evaluation

It is crucial for the various elements that compose the simulation to accurately reflect the ridership data. Our model predicts well on a MSE basis. Not only do we need to accurately predict the mean levels of ridership, but also the inherent variability.

Data and Present Solution

To understand why buses get crowded, the CTA collects volumes of data of:

- GPS bus location: allowing the agency to see how well the buses adhere to the proposed schedule.
- Passenger counts: allowing the agency to monitor



Evaluating the Efficiency and Effectiveness of Garbage Pickup in Chicago

The Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship 2013
 Jonathan Auerbach (Columbia University), Matt Gee (University of Chicago), Michael Cassio (University of Chicago), Sarah Evans (Stanford University), Andrea Fernández Conde (Instituto Tecnológico Autónomo de México), Zach Seeskin (Northwestern University), Vidhur Vohra (Georgia Institute of Technology)

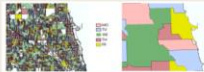
Summary

The Department of Streets and Sanitation recently changed the way they pick up Chicago's garbage. We are performing an exploratory data analysis to identify how garbage truck behavior changed as a result and testing whether trucks pickup more trash faster.

Our strategy is to divide the work a truck completes throughout a day into tasks and determine how efficiently those tasks are completed. We are currently testing hypotheses generated from this analysis.

Context

- There were three major improvements to Chicago's garbage collection system over the study period (2010-2013):
 - Change from ward-based to grid-based trash collection
 - Gradual expansion of the Blue Cart Recycling Program
 - Implementation of new trucks and technology



This is an example of a system-wide improvement to the weekly trash collection schedule.

Visualization Tool



Data

We observe trucks completing a variety of tasks as indicated by the symbols below:

- Garbage Pickup in Alley
- Refueling Gas at Fuel Sites
- Dumping Trash at Dump Sites

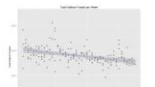
We have various GIS datasets and data on missed pickup and code violations.

We are creating a tool for the Department to aid in this analysis. It allows us to visualize the order and location of tasks completed by a truck in a day.

Below is a description of our analysis and how it fits into the work day of a garbage truck.

At the Fuel Site...

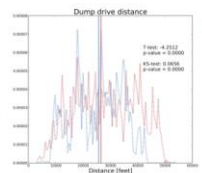
...we observe how far each truck has traveled since it last fueled and how many gallons it fueled. We also know how many alleys each truck has visited between fuelings and how many tons of trash it moved.



We are investigating how many alleys a truck visits per mile and how much trash it moves per gallon. An improvement to the program should enable trucks to do more with less resources.

From Alley To Dump...

...we are looking at how far trucks drive. Dump sites are generally located far from population centers and driving to a dump site constitutes a large portion of the distance a truck drives in a day.

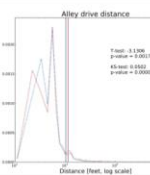


We are measuring the distance a truck drives between a dump and the last alley in the series of alleys driven. We are also looking at the distance between the dump and the next alley a truck visits.



From Alley to Alley...

...we are measuring the distance between consecutive alley locations in the work schedule of a truck. We will then look at changes in the distribution of these distances to see whether improvements to the system reduced "lag time". A reduction in lag time would suggest that a truck's work schedule is more efficient.



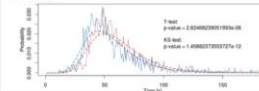
Flagging Outliers

In order to flag trucks that exhibit unusual behavior, we implemented a method that compares the lag time of each truck to the lag time of its peers. We broke down the work trucks into "legs". For each "leg" we generated a distribution. We graded a truck's efficiency by assigning it a percentile score for each leg and averaging the scores.

We will look at whether truck grades rose after service improvements were implemented.

In the Alleys...

...we are investigating whether there was a change in service time. After normalizing the time spent in alleys by the number of service buildings, we will estimate the change and its statistical significance. We will also evaluate the likelihood of future service performances by fitting a log-normal distribution.



This work was done during the Eric & Wendy Schmidt Data Science for Social Good Fellowship at the University of Chicago.

Understanding and Improving Energy Consumption in Commercial Buildings

The Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship 2013
 Scott Alfeld, Andrea Fernández Conde, Camelia Simoiu, Brandon Willard

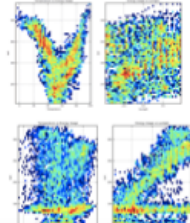
The Problem

Estimating the financial and energy savings of an energy-efficiency building retrofit proves to be a challenging task for commercial buildings.

No two buildings are alike and thus the potential energy savings vary substantially by property, so the return on investment of fixing up property is highly uncertain.

Data Set

- hourly interval energy consumption (kwh) for 6,000+ commercial buildings
- corresponding hourly temperature
- NAICS code (type of business)
- Latitude and longitude
- Federal and state holidays
- Position of the sun



How sensitive is a building to temperature and sunlight?

To the left are four heat maps illustrating a building with high sensitivity to temperature and low sensitivity to sunlight (top), and low sensitivity to temperature and high sensitivity to sunlight (bottom).

Bike Share: A Balancing Act

The Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship 2013
 Walter Dempsey (University of Chicago), Adam Fishman (Yale University), Jette Henderson (UT Austin), Breanna Miller (University of Michigan), Hunter Owens (University of Chicago), Juan-Pablo Velez (University of Chicago), Vidhur Vohra (Georgia Institute of Technology)

Summary

We are improving how urban bike share programs work by predicting how full or empty their stations will be in the future so that they can more effectively address the problem of rebalancing bikes across the bike share stations.

The Problem

Bike share programs have been shown to reduce traffic and congestion in many cities. Because of commuting patterns, bikes tend to pile up downtown in the morning and in residential areas in the afternoon. This imbalance can make using bike share difficult and the program ineffective because riders cannot take out bikes from empty stations or drop off their bikes at full stations.



To address this problem, bike share operators rebalance by dispatching trucks to reallocate bikes from full stations to empty ones. Dispatchers can only see the **current number of bikes** at each station, not how many will be there in an hour or two.

By analyzing weather and bike share station trends, we predict how many bikes are likely to be at each station in the future. Using our predictions, dispatchers can proactively adjust the distribution of bikes before stations are actually empty or full, diffusing problems that could hinder riders.

- Our work will lead to:
 - Bike stations being empty or full less often
 - Improved experience of bike share users
 - Increased mobility for city residents

Data

Our data set includes station-level reports of the number of bikes and empty slots available every minute, as well as weather data observed on an hourly

basis. The following is a sample of the station data aggregated to fifteen minute intervals:

Time	Bikes Available	Slots Available	Rain (cm)	Temp (°F)
10/6/10 21:30	3	13	none	58
10/6/10 21:45	3	12	none	58
10/6/10 22:00	4	11	none	56
10/6/10 22:15	4	11	none	56
10/6/10 22:30	11	4	0.1	56

Methodology

We use three methods to model the number of bikes at a station: binary logistic, ordinal logistic, and Poisson point process. Each station is modeled separately. All models use the same predictor variables: time, day of week, and temperature. Additionally, the logistic methods use an autoregressive structure, using three previous time points as predictors.

The **binary logistic regression** models p , the probability that each available slot will be full.

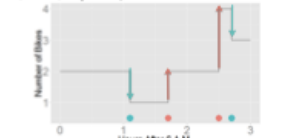
$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \beta_2 \text{time} + \beta_3 \text{weekend} + \beta_4 \text{temperature})}}$$

An **ordinal logistic regression** uses the same regressors as the binary logistic but instead of predicting probability values, it predicts the breaks in a logit function that determine probabilities for ordered states.

$$P(\text{number of slots full} \geq j) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X - \beta_2 j)}}$$

Where X contains the predictor variables, and β_1, β_2 is the point above which j slots are predicted to be full in the logistic distribution pdf.

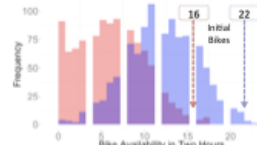
The **Poisson regression** models the rates of bikes arriving and leaving the station. Because we assume bikes arriving and leaving the station are independent, these rates are modeled separately. The graph below illustrates bikes arriving and leaving from a station, with green (red) dots at the times when bikes left (arrived) respectively.



Preliminary Evaluation

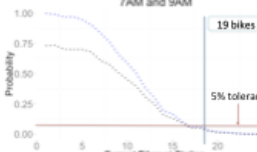
With the estimated rate parameters, we assume that the change in bikes over an interval can be given by Poisson point process. By simulating the stochastic process from an initial bike availability, we can infer distribution of bike availability several hours into the future.

Distribution of Bikes at 9AM at a Station Conditional on Current Bikes



Not only can we infer conditional distributions, we can also assess the probability of the bike station becoming empty or full at any time in the interval if any initial value of bikes. Then, for any tolerance level, we can suggest an appropriate initial bike choice.

Probability that Station is Empty between 7AM and 9AM

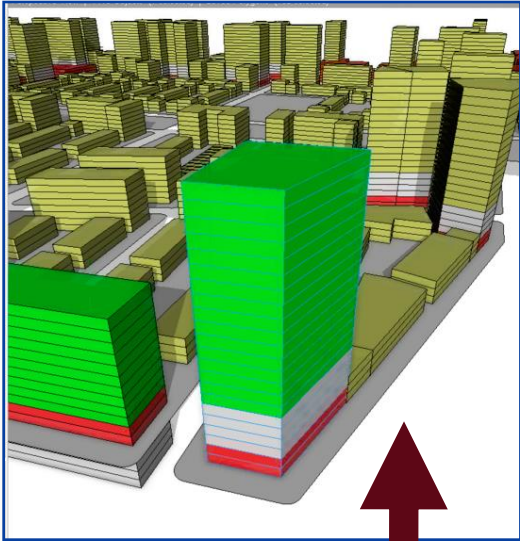


Future Work

- Add more weather features to models
- Evaluate our models more completely using out-of-sample validation with a rolling training window, using a minimum of one year as training set
- Extend this work to other cities
- Use transactional data to incorporate state of nearby stations into prediction
- Combine modeling methods in an ensemble to improve results
- Deliver user-friendly prediction tool to bike share operators

This work was done during the Eric & Wendy Schmidt Data Science for Social Good Fellowship at the University of Chicago.

Design



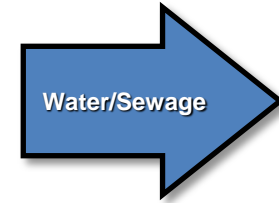
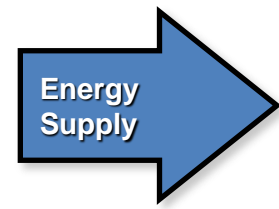
Analyze



hours to days



Plan



Decide

