



URBAN  
CENTER FOR  
COMPUTATION  
AND DATA

### Charlie Catlett, Director

Senior Fellow, Computation Institute of the University of Chicago and Argonne National Laboratory

Senior Computer Scientist, Argonne National Laboratory  
catlett@anl.gov

Presentation to the National Academy of Sciences  
Committee on National Statistics  
February 6, 2014



MacArthur  
Foundation



## Rapid Urbanization in Developing Economies



Landsat images of the Pearl River Delta in 1980 and 2005, illustrating the impact of urbanization on the planet.



Between now and 2020, the Guangdong province will invest \$229B in 202 ongoing and 258 new transport infrastructure projects to create a single 50M person city.

In 2025:

**70%**

of Chinese people will live in cities with 1M or more people.

And by 2030...

**221**

Chinese cities will have 1M or more people.

China will add

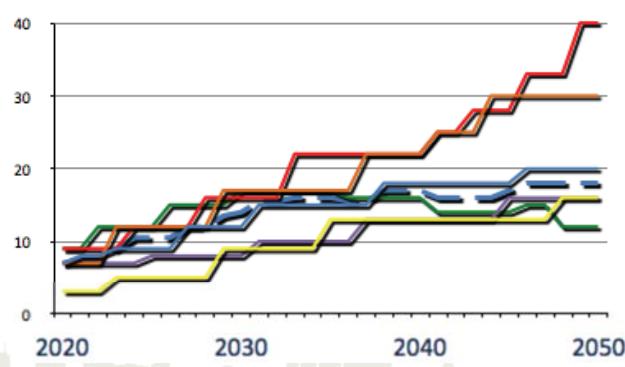
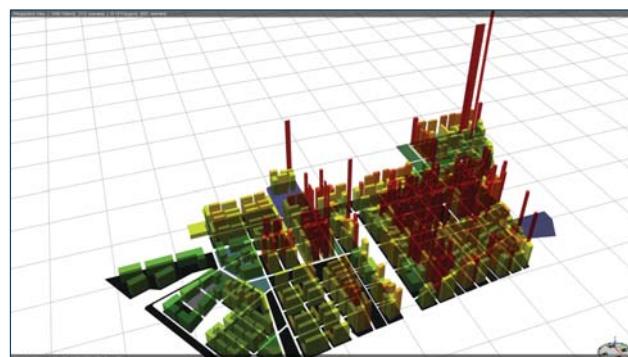
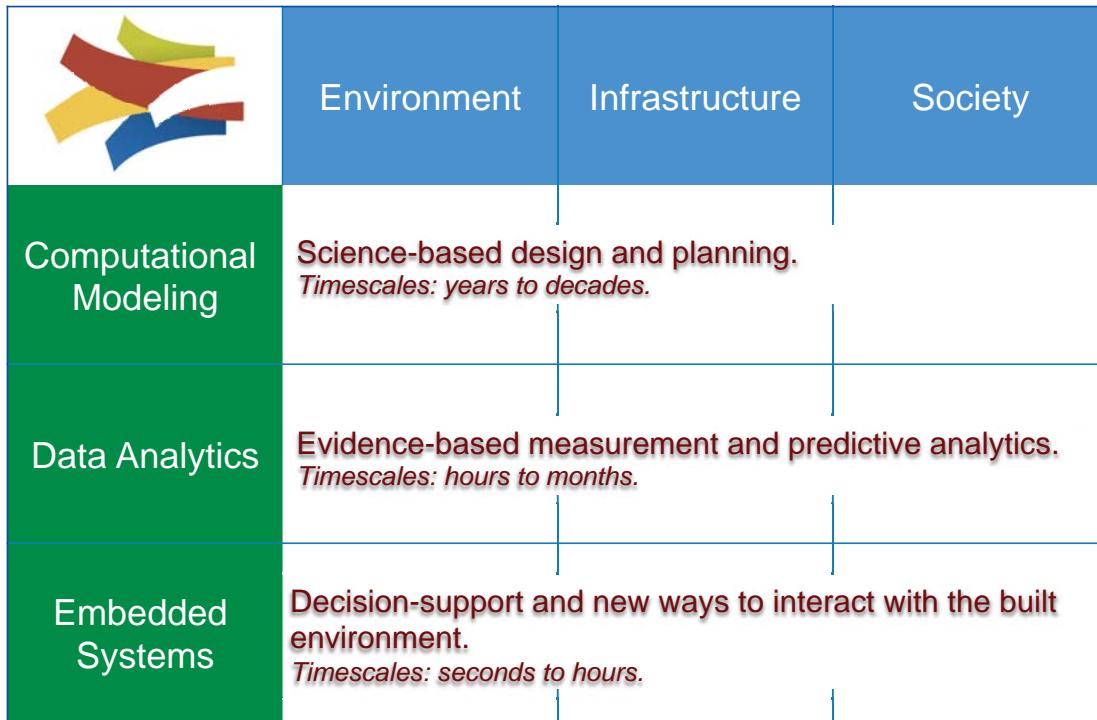
**400**

million city dwellers

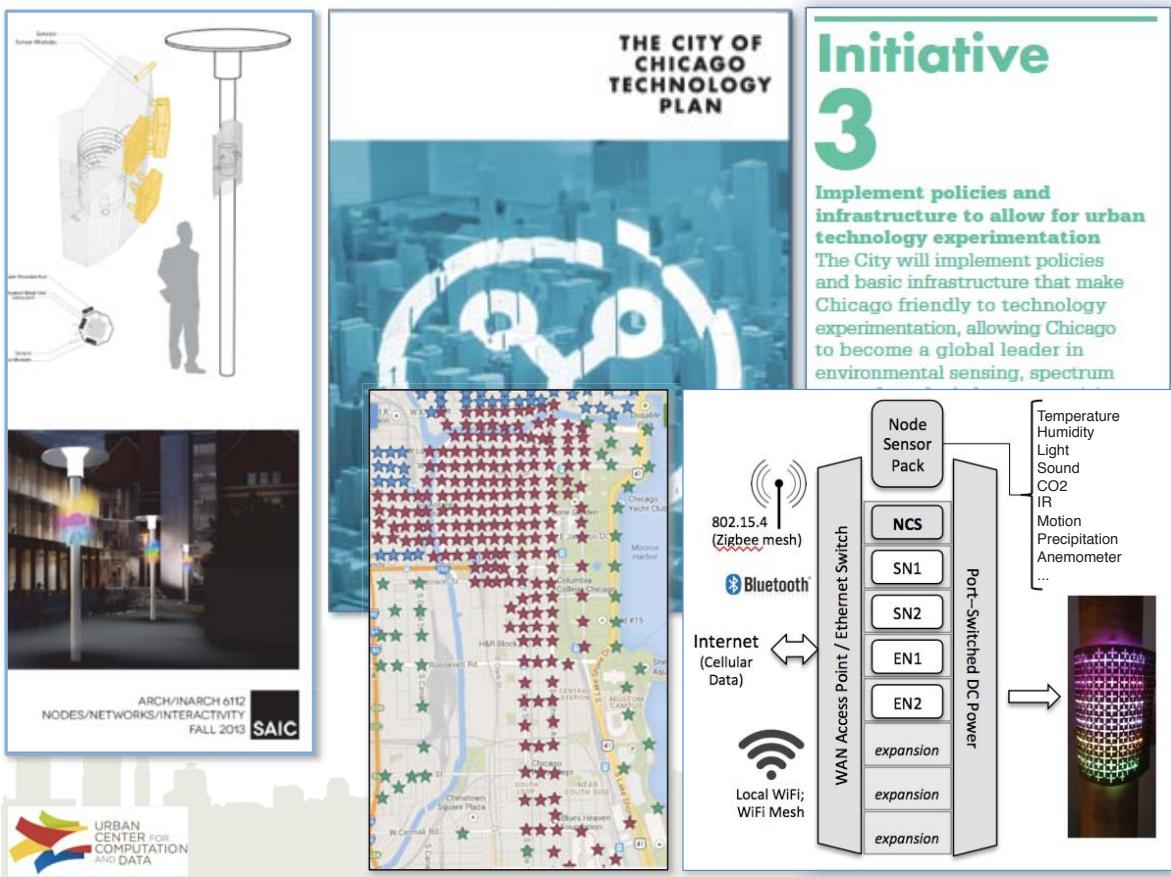
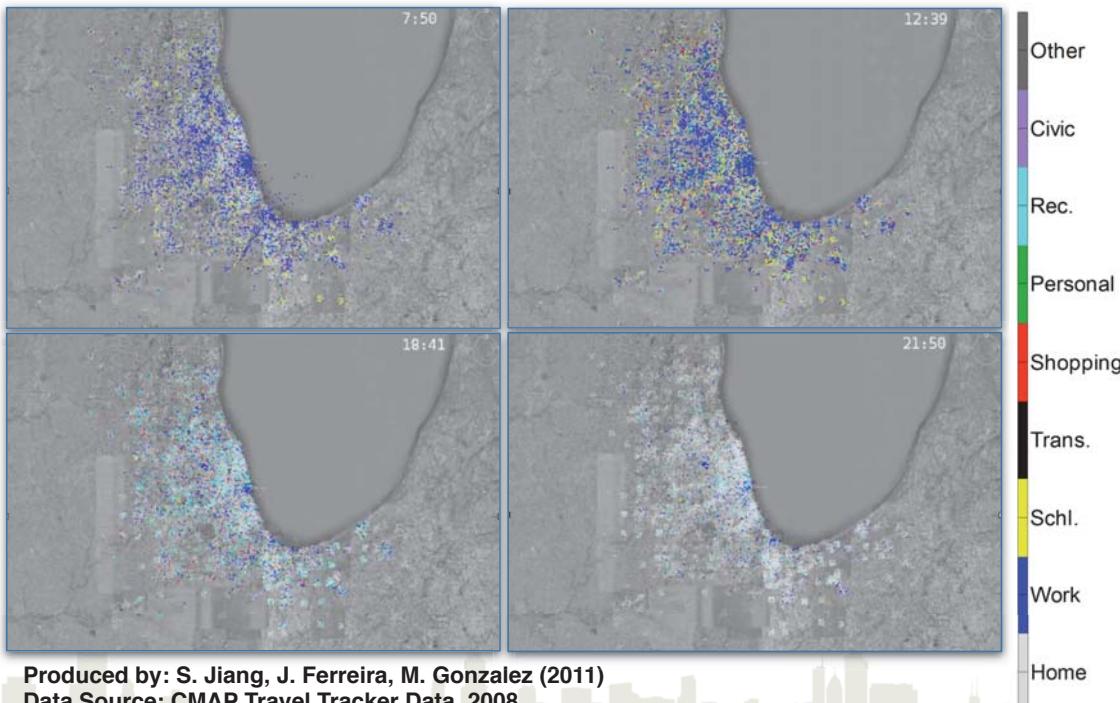
....requiring the construction of one New York City every year for several decades



Source: Foreign Policy Magazine, Sep/Oct 2010, "Megacities," Richard Dobbs (McKinsey Global Institute)



## Chicago Human Activity Patterns: Weekday





Citizen engagement and communication

## Instrument

Usage, management, access policies/processes

Instrument training materials/activities

Programming and Development Tools

## Platform

Device provisioning, diagnostics, management

Shared devices (configurations, capabilities)

Device install and management specifications

Enclosures, Placement, Density, Power, Internet...

**City of Chicago Data Portal**

[Home](#) [About](#) [Help](#) [Developers](#) [Terms of Use](#) [City of Chicago](#) [Sign Up](#) [Sign In](#)



**Energy Usage**  
An unprecedented dataset containing electrical and gas energy usage throughout Chicago in 2010. Data is available by Census block and month.



**Alternative Fuel Locations**  
List of locations in Chicago where alternative vehicle fuels are available.



**Crimes - 2001 to present**  
Review reported incidents of crime that occurred in Chicago from 2001 to present.



**Chicago Transit Authority Datasets**  
Explore CTA data, including bus routes, 'L' lines, ridership information and fare card sales outlet locations.

Search

**View Types**

- [Datasets](#)
- [Charts](#)
- [Maps](#)
- [Calendars](#)
- [Filtered Views](#)
- [External Datasets](#)
- [Files and Documents](#)
- [Forms](#)
- [APIs](#)

**Categories**

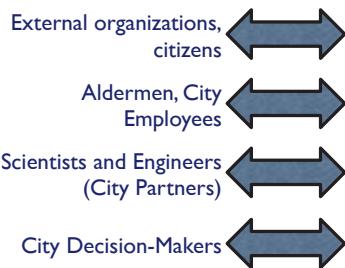
- [Administration & Finance](#)
- [Buildings](#)
- [Community & Economic Development](#)
- [Education](#)

**Search & Browse Datasets and Views**

Name	Popularity
<input checked="" type="checkbox"/> <a href="#">Elevation Benchmarks</a> Buildings, benchmarks, gis	4,662 views
1. The following dataset includes "Active Benchmarks," which are provided to facilitate the identification of	
<input checked="" type="checkbox"/> <a href="#">Performance Metrics - Innovation &amp; Technology - Site Availability</a> Administration & Finance	360 views
2. performance metrics, technology The website availability metrics below are derived from an automated monitor that sends a request	
<input checked="" type="checkbox"/> <a href="#">Relocated Vehicles</a> Transportation, vehicles, streets	2,954 views
3. This dataset presents current and former locations of vehicles that have been relocated by the City of	
<input checked="" type="checkbox"/> <a href="#">Towed Vehicles</a> Transportation, vehicles, streets	5,744 views
4. This dataset displays location for vehicles that have been towed and impounded by the City of Chicago	
<input checked="" type="checkbox"/> <a href="#">Chicago Traffic Tracker - Congestion Estimates by Segments</a> Transportation, traffic, sustainability	1,499 views
5. This dataset contains the current estimated speed for about 1250 segments covering 300 miles of	
<input checked="" type="checkbox"/> <a href="#">Chicago Traffic Tracker - Congestion Estimates by Regions</a> Transportation, traffic, sustainability	1,545 views
6. This dataset contains the current estimated congestion for the 29 traffic regions. For a detailed	



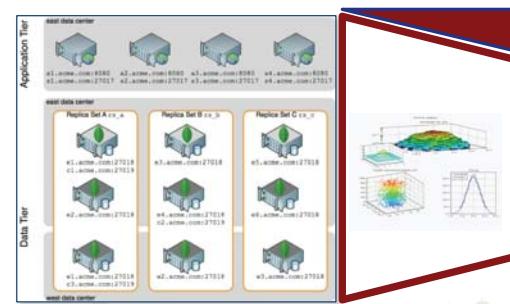
## Data Access, Authorization, Privacy



## Visual Interaction, Mapping, Analysis Tools

## Scalable Data Handling

(pre-processing, cleaning, merging, de-duplication, access...)



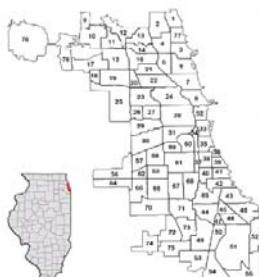
## Application Programming Interfaces

## Automated Continuous Data Analytics

## Data Sources



## Background: Crime Prediction in Chicago



Since 2009, we have been working with the Chicago Police Department (CPD) to predict and prevent emerging clusters of violent crime.

Our new crime prediction methods have been incorporated into our **CrimeScan** software, run twice a day by CPD and used operationally for deployment of patrols.

**Carnegie Mellon University**  
**HeinzCollege**



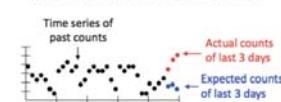
### CrimeScan: Cluster Detection

We aggregate daily counts for each leading indicator at the block level, and search for clusters of nearby blocks with recent counts that are significantly higher than expected.

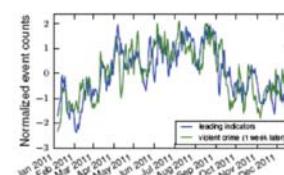
Imagine moving a circular window around the city, allowing the center, radius, and temporal duration to vary.



Is there any spatial window and duration  $T$  such that counts have been significantly higher than expected for the last  $T$  days?



### Results: Exploratory Analysis



Considering all subsets of census tracts within each of the 77 neighborhoods of Chicago, 28 different potential predictors, and a 1-week lag, we found a correlation of  $r = .786$  between violent crime and a subset of 12 leading indicators, for 10 census tracts in the West Englewood neighborhood.

Total run time for all 77 neighborhoods was **2.1 hours**.



Daniel B. Neill (Carnegie Mellon University - Event and Pattern Detection Laboratory),  
Brett Goldstein (fmr Chicago CIO/CDO; Fellow, Harris School of Public Policy, Computation Institute, UChicago)

Garbage Cart Black Maintenance	Inspection	No Building Permit & Construction Violations	Restaurant Complaint	Animal Abandoned Water Quality
Sanitation Code Violation	Building	Missing Lid/Grade	Unwanted Animal	Sewer Odor/Bad Odor
Building Violation	Sewer Cave In	Block Party Request	Blue Recycling Cart	Dead Bird
Stray Animal	Inspection	Recycling Pick Up	Animal Bite	Animal In Trap
Dead Animal Pick-Up	Garbage Pickup	Nuisance Animals	Dumpster Task Force	Animal Business
Sewer Cleaning	Debris Removal	Animal - Inhumane	Inspection	Animal Fighting
	Fly Dumping	Treatment	Demolition Inspection	



## Rodent Infestation



- Liquor license (decreased likelihood)
- Low-risk v. high-risk gas station (both less likely than non-gas station)
- Tobacco license
- Neighborhood
- Common names (e.g., Tom's market) in store name

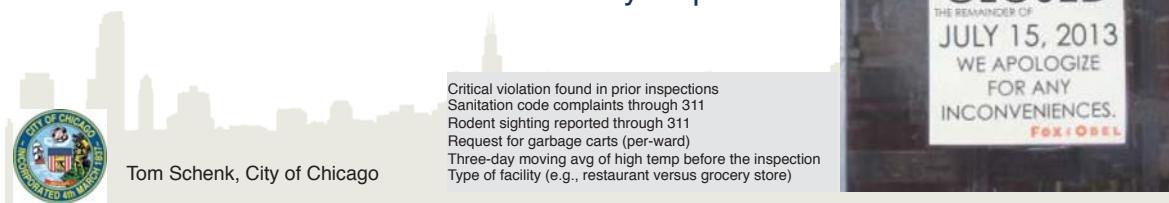
## Black Market Cigarettes



Critical violation found in prior inspections  
Sanitation code complaints through 311  
Rodent sighting reported through 311  
Request for garbage carts (per-ward)  
Three-day moving avg of high temp before the inspection  
Type of facility (e.g., restaurant versus grocery store)

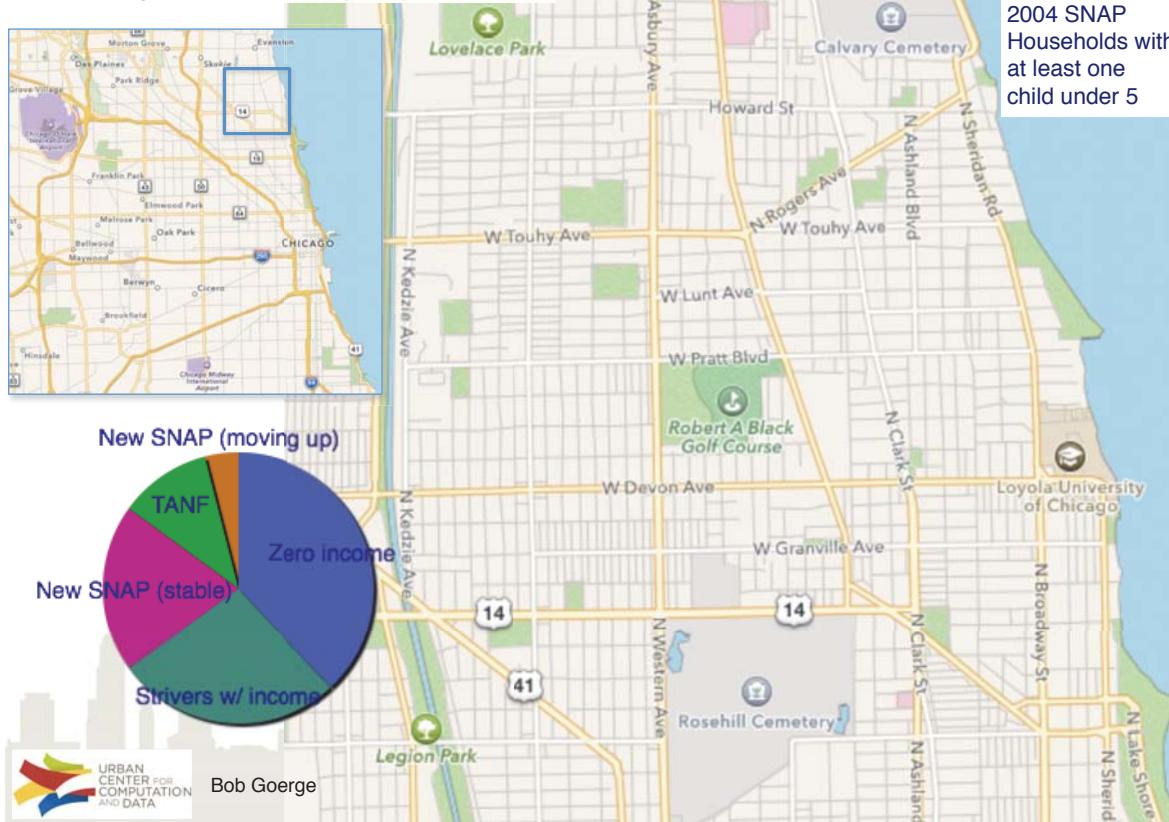


## Food Safety Inspections



Chapin Hall at the University of Chicago

Policy research that benefits children, families, and their communities





### Negative Impacts

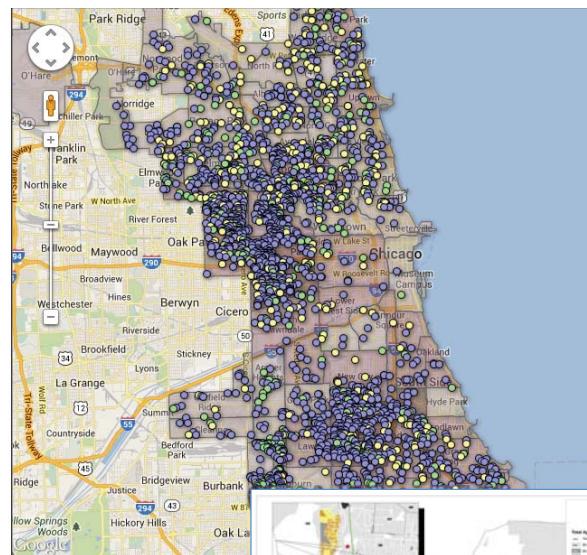
-  **Lowers property value**
-  **Crime**
-  **Lost taxes**

### Legal Obstacles

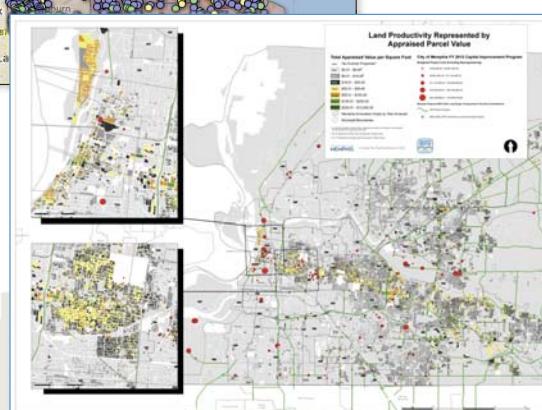
-  **Unknown owner**
-  **Unpaid bills**
-  **Contaminated land**

### Guiding Indicators

-  **Housing stability**
-  **Affordability**
-  **Vacancy**

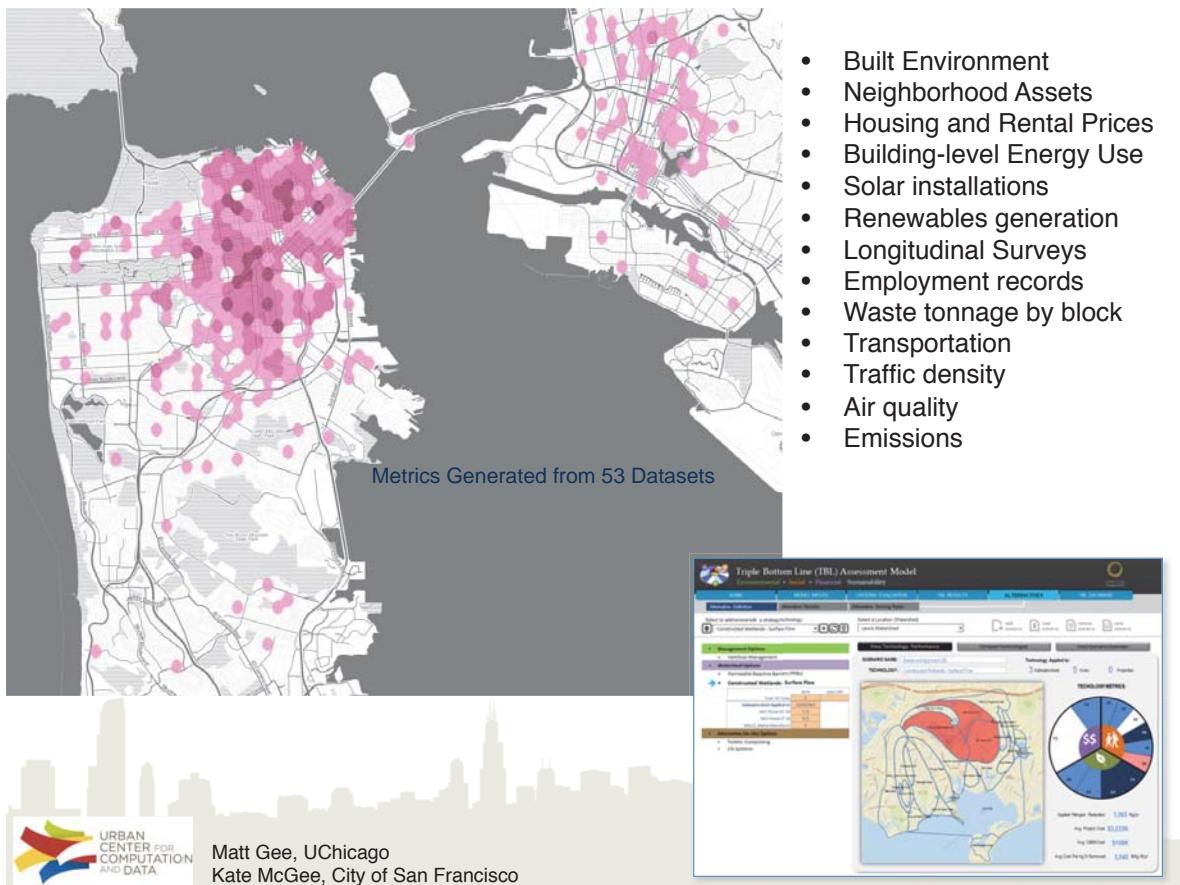


**City of  
Memphis**





Matt Gee, UChicago  
Kate McGee, City of San Francisco



**Improving Communities through Data-Driven Land Banks**

Sophia Alice, Evan Missula, Skyler Whorton, and Tom Plagge

**Summary**

The foreclosure crisis led to an explosion of abandoned properties in Cook County—properties that can destabilize neighborhoods, depress tax revenues, and contribute to rising crime rates. The Cook County Land Bank is a new agency tasked with putting them back to use.

**The Problem**

The Land Bank has several tools at its disposal to redvelop abandoned buildings. They can clear title, forgive back taxes, combine parcels, and hold land taxes until a developer is found. The problem is that since there are over 10,000 vacant residential addresses in Cook County (and 100,000 HUC/USPS), the Land Bank must be selective in its acquisitions.

**Property Scores**

We are measuring the health of neighborhood real estate markets along several dimensions, including stability and affordability. The stability score ( $S$ ) is based on Walker & Associates' property value index ( $V$ ), transaction volume ( $V_{tr}$ ), and the prevalence of high-cost lending ( $H$ ). The affordability score ( $A$ ) is based on the median property sale price ( $P$ ) in each census tract.

**Community Scores**

We are incorporating the scores and maps we developed into a Django web application with a PostgreSQL database and PostGIS extensions. The purpose of the application is to provide an API that it can be leveraged with the Land Bank's inspection and inventory systems.

**Web Application**

We are incorporating the scores and maps we developed into a Django web application with a PostgreSQL database and PostGIS extensions. The purpose of the application is to provide an API that it can be leveraged with the Land Bank's inspection and inventory systems.

**Conclusions**

The Cook County Land Bank plans to effect the problem of vacant residential property in a sensible, data-driven way. Having all of the relevant information in one place will save the staff time, and summarizing the information in a way that is easy to understand and use will help make the bank's decisions clear and transparent. A systematic approach to property acquisition and a better understanding of the community it serves will effect the surrounding property values.

**Future Work**

The application and algorithms we provide to the Land Bank will be prototypes, intended primarily to guide the land bank's work. As the Land Bank begins to use the application, the indicators and scores can be calibrated against real results, and can be refined to reflect the impact of the strategies it pursues.

**Mapping**

Because of the power of the Land Bank to combine adjacent parcels, it will be useful for staff to be able to see potential acquisitions on a map along with other nearby existing and possible future buildings. We have produced an

**Mapping**

We've incorporated 2011 and crime report data from the City of Chicago, property data from HUC/USPS, lending data from HMDA, and census data from the census.

**Data**

We used housing market data provided by the Institute of Housing Studies at DePaul University and Cook County. Data from the 2010 census, HUC/USPS property assessments, developers and property parcel boundaries throughout the county, tagged by geography and property type.

**Methodology**

We've incorporated 2011 and crime report data from the City of Chicago, property data from HUC/USPS, lending data from HMDA, and census data from the census.

**Conclusion**

This work was done during the Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship 2013.

**SCHMIDT FAMILY FOUNDATION**

**Bits**

JULY 25, 2013 7:47 AM [Comment](#)

**A Summer of Data Hacking Social Problems**

By STEVE LOHR

**Facebook** **Twitter** **Google+** **Save** **E-mail** **Share** **Print**

The idea, Rayid Ghani recalled, grew out of his experience speaking to computer science students at elite schools like Carnegie Mellon, Stanford and the University of Chicago. President Obama had just won his re-election bid last fall. And Mr. Ghani, chief scientist for the campaign, was on a kind of explanatory victory tour, describing how cutting-edge data analysis and computing tools gave its side an edge.

For Mr. Ghani, the Obama campaign demonstrated how those tools could be used to influence people in fields beyond the well-known commercial ones, like search, social networks and online advertising. And beyond politics, he would tell the students, were a host of social challenges in health care, education and urban development where their skills could be put to good use, working with nonprofits, civic

**Rayid Ghani**, chief scientist for President Obama's re-election campaign.



Where We Work A project by DataMade and Data Science for Social Good

Where do Chicagoans live and work?

Select an area on the map to find out where people commute to and from for work.

[Read more »](#)

Inbound workers  
 Outbound workers

**Example Federal data source: LODES**

(US Census Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics)

**Tract 17031839100**

664,745 inbound workers  
2,499 connected tracts

Top 5 tracts	# workers
17031330100	5,110
17031081800	3,845
17031320100	3,221
17031280100	3,029
17031839100	2,634

**OnTheMap**

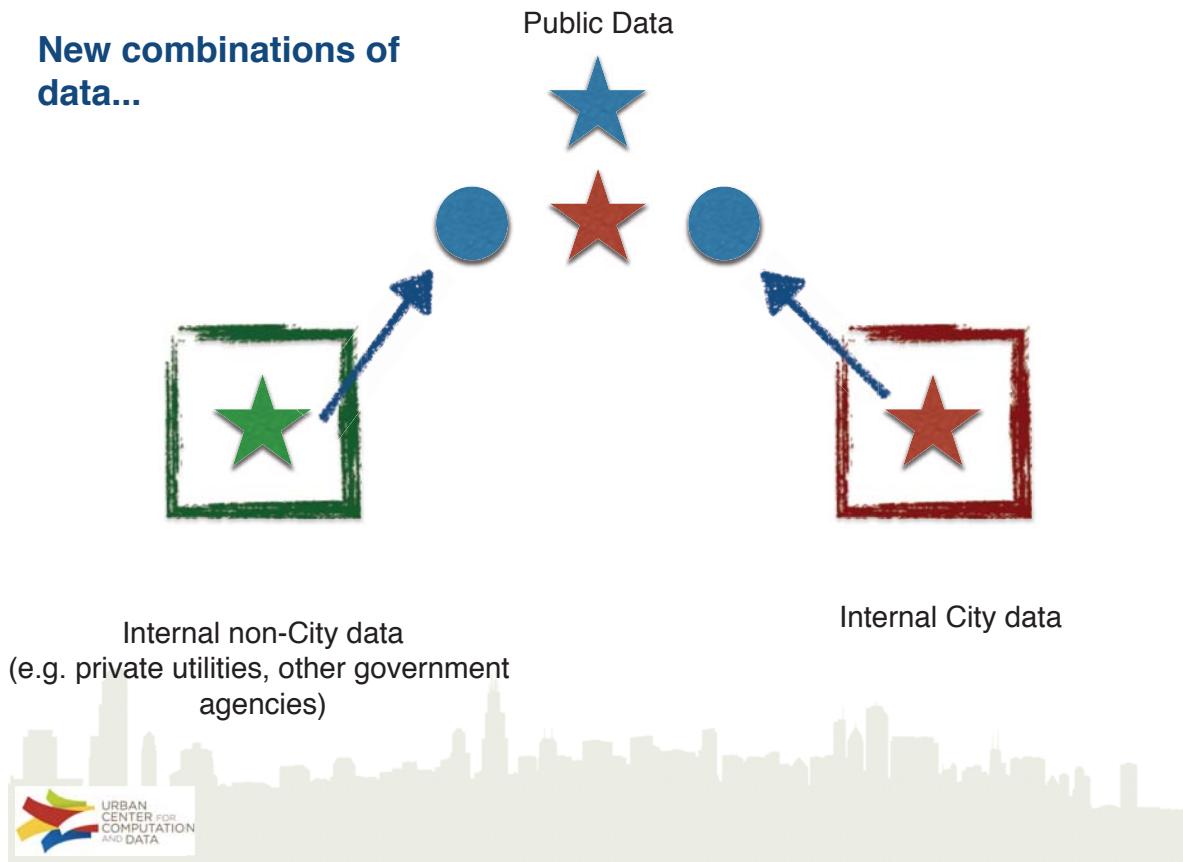
Leaflet | Map data © 2011 OpenStreetMap

**No API - must download individual (huge) flat files or use online tool ("OnTheMap").**  
**Significant effort to enable geographic queries/navigation**  
**Data is 1-2 yrs old. Ideally could use local data to fill gaps.**



Matt Gee, UChicago  
Derek Eder, DataMade

New combinations of data...



## Summary

Cities are creating (open) data resources

Many critical city decisions also rely on county, state, federal data

Spatial and temporal resolution are key (along with timely data)

## How Can the Federal Government Help?

### ACCESS

Modern Data Infrastructure *with APIs*

### PREPARATION

Open Source Tools for Common Tasks

### INCENTIVES

Strategically chosen standards and requirements

### ACCELERATION

Fund city/academic exemplars to provide replicable/sharable capabilities and to assist other (esp. smaller) cities.

