



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



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

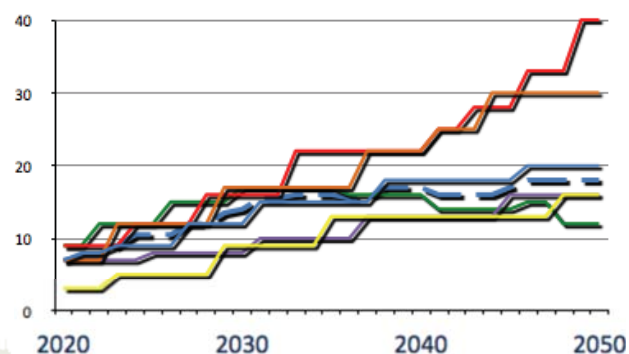
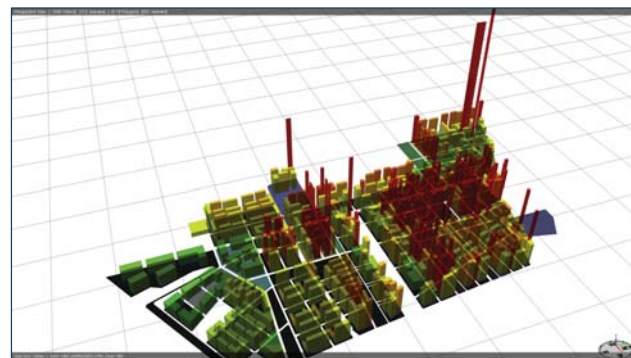
million city dwellers

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

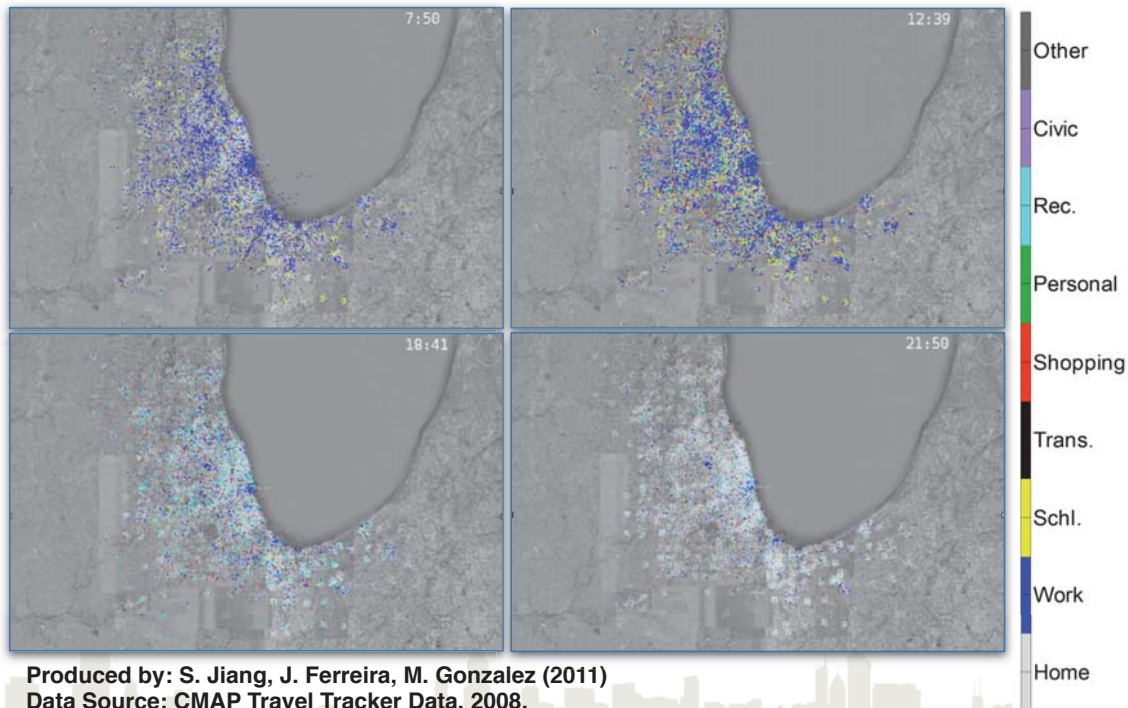


Source: Foreign Policy Magazine, Sep/Oct 2010, "Megacities," Richard Dobbs (McKinsey Global Institute)

	Environment	Infrastructure	Society
Computational Modeling	Science-based design and planning. <i>Timescales: years to decades.</i>		
Data Analytics	Evidence-based measurement and predictive analytics. <i>Timescales: hours to months.</i>		
Embedded Systems	Decision-support and new ways to interact with the built environment. <i>Timescales: seconds to hours.</i>		



Chicago Human Activity Patterns: Weekday

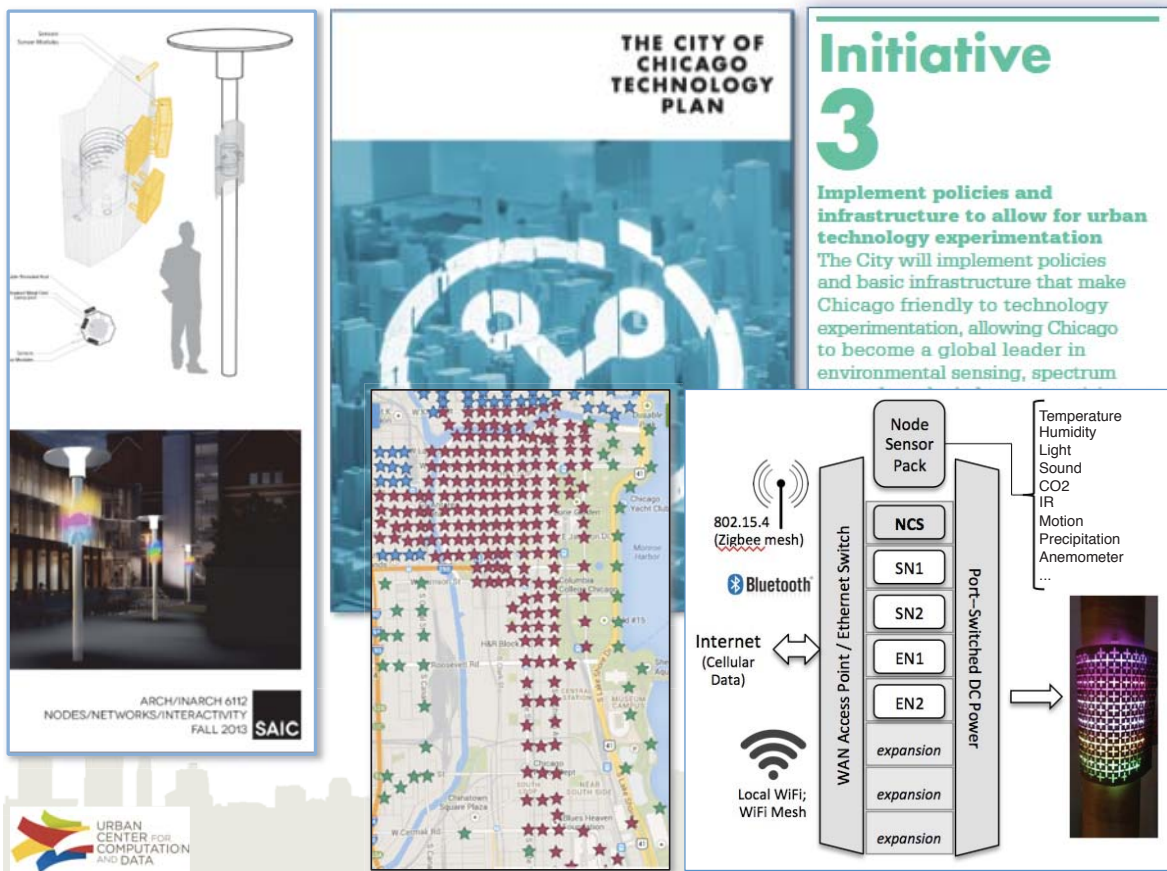


Produced by: S. Jiang, J. Ferreira, M. Gonzalez (2011)

Data Source: CMAP Travel Tracker Data, 2008.

Reference: Jiang, S., J. Ferreira, and M. González. 2012.

[Clustering Daily Patterns of Human Activities in the City.](#) *Data Mining and Knowledge Discovery*. Volume 25, Number 3, Pages 478-510





Instrument

Platform

Citizen engagement and communication

Usage, management, access policies/processes

Instrument training materials/activities

Programming and Development Tools

Device provisioning, diagnostics, management

Shared devices (configurations, capabilities)

Device install and management specifications

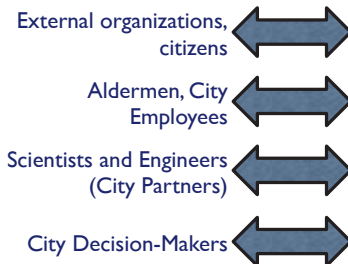
Enclosures, Placement, Density, Power, Internet...



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



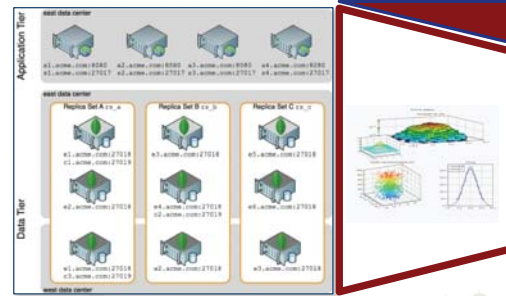
Data Access, Authorization, Privacy



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



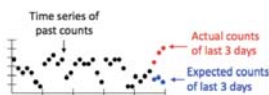
CrimeScan: Cluster Detection

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.

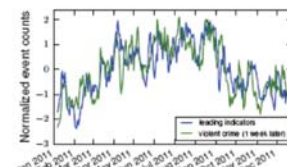
Imagine moving a circular window around the city, allowing the center, radius, and temporal duration to vary.



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

Food Safety Inspections



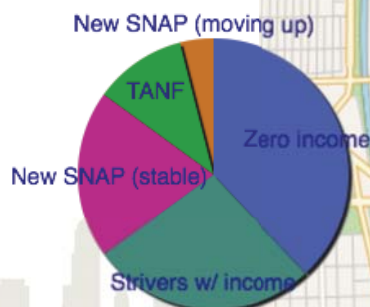
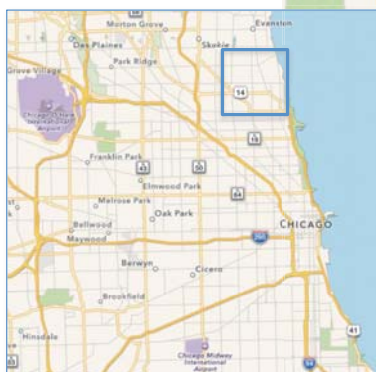
Tom Schenk, City of Chicago

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)



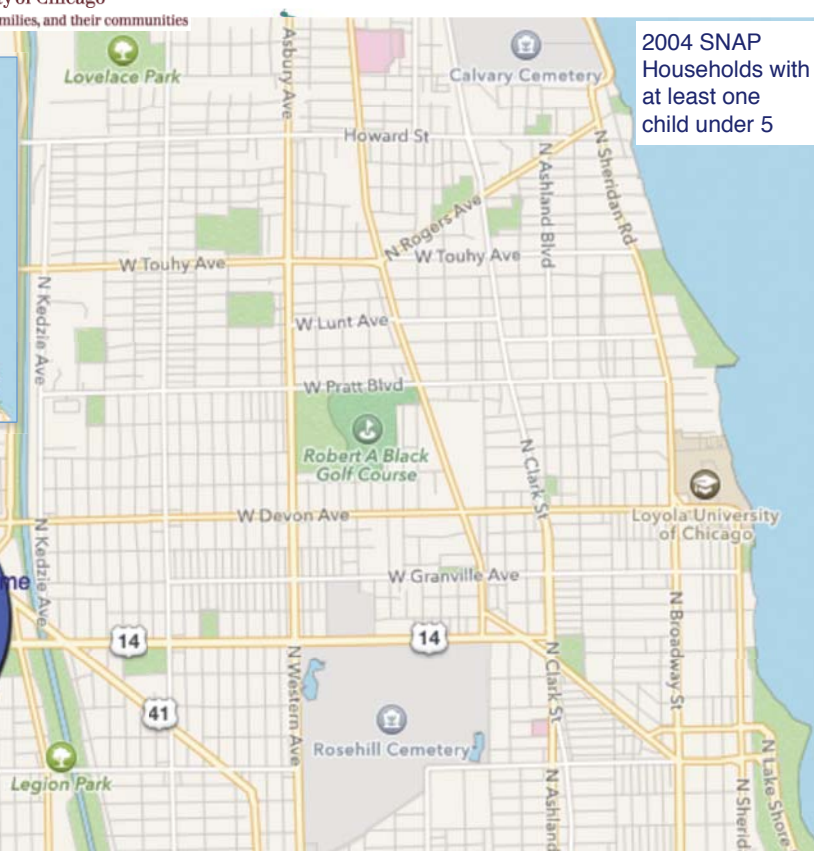
ChapinHall at the University of Chicago

Policy research that benefits children, families, and their communities






Bob Goerge

2004 SNAP
Households with
at least one
child under 5








Negative Impacts

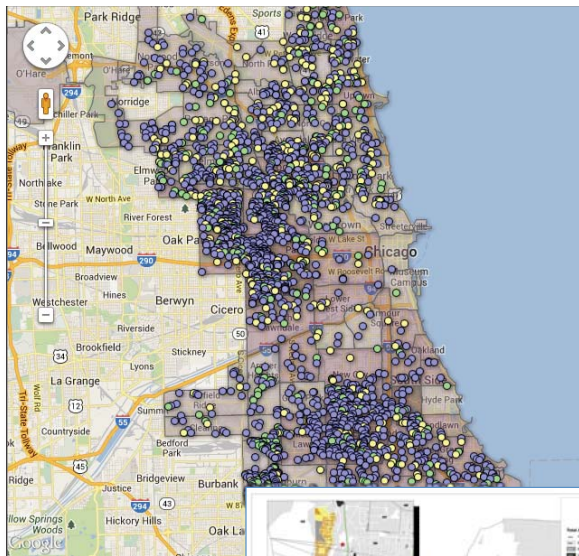
-  Lowers property value
-  Crime
-  Lost taxes

Legal Obstacles

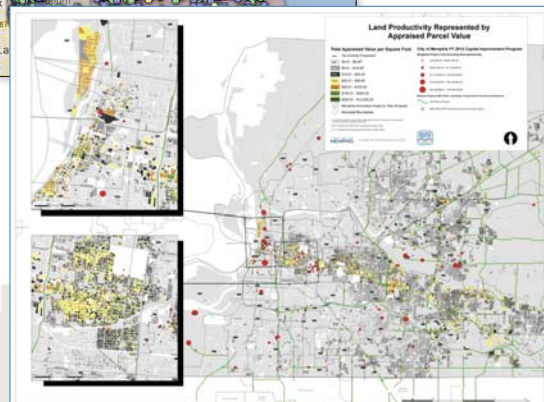
-  Unknown owner
-  Unpaid bills
-  Contaminated land

Guiding Indicators

-  Housing stability
-  Affordability
-  Vacancy



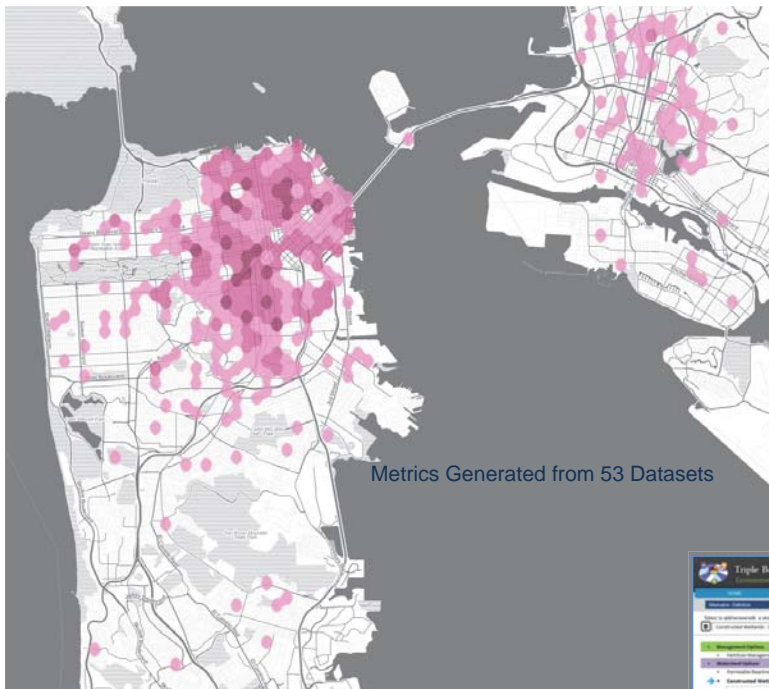
City of
Memphis



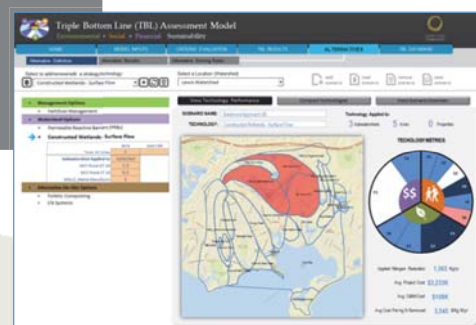
Equitable Development Community Development Habitat and Ecosystems Function Water, Wastewater, and Stormwater Energy Materials Management and Waste Health and Well-Being Access and Mobility



Matt Gee, UChicago
Kate McGee, City of San Francisco



- Built Environment
- Neighborhood Assets
- Housing and Rental Prices
- Building-level Energy Use
- Solar installations
- Renewables generation
- Longitudinal Surveys
- Employment records
- Waste tonnage by block
- Transportation
- Traffic density
- Air quality
- Emissions



Matt Gee, UChicago
Kate McGee, City of San Francisco

Improving Communities through Data-Driven Land Banks
Sophia Alice, Evan Misshula, Skyler Whorton, and Tom Plagge

Summary
The foundation exists to an explosion of abandoned properties in Cook County—properties that are deteriorating, blighted, and often vacant. The Cook County Land Bank is a new agency tasked with getting these back to use.

The Problem
The Land Bank has several tools at its disposal to address abandoned buildings. They can clear lots, begin back taxes, combine parcels, and hold land tax-exempt until demand for the property returns. However, since there are over 100,000 vacant residential addresses in Cook County (per US HUD/CSPR), the Land Bank must be selective in its acquisitions.

Community 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 Tract & Writings (2000), and depends on property value (V), vacancy rate (V), and the percentage of high-cost housing (H). The affordability score (A) is based on income (I) and median property sale price (P) in each census tract.

$$S = 0.6V + 0.3V_1 + 0.1V_2$$

$$A = \frac{I}{P}$$

Web Application
We are incorporating the scores and maps we developed into a Chicago web application with a PostgreSQL database and PostGIS extension. The parcel data will also be available via an API so that it can be kept consistent with the Land Bank's inspection and inventory systems.

Conclusions
The Cook County Land Bank plans to attack the problem of abandoned buildings in a timely, data-driven way. Having all of the relevant information in one place will save the staff time, and summarizing the information in a handful of meaningful, digestible scores will help make the bank's decisions clear and transparent. A systematic approach to property acquisition will also allow the agency to evaluate the impact of the strategies it pursues.

Future Work
The application and algorithms we provide to the Land Bank will be principles, intended primarily to guide the bank's strategic decisions. As the agency moves into operation, the indicators and scores can be calibrated against real results, and can be revised to reflect changing strategies or market conditions.

Property Scores
We are also developing scores for individual properties based upon their relative values to their neighbors and their economic impact on the community. The former is based on 311 and crime reports. For the latter, we are using a hedonic pricing model that takes into account property and neighborhood characteristics. Based on historical data, we can estimate the percentage by which a demolition, vacancy, or demolition in a given community will affect the surrounding property values. The preliminary model indicates that, controlling for the base demographic and economic characteristics of its community, each demolition that occurs within 1/4 mile of a property has approximately a -5% effect on its price.

Mapping
More use of the power of the Land Bank is to consider 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

SCHMIDT FAMILY FOUNDATION

Bits
JULY 25, 2013, 7:47 AM | Comment

A Summer of Data Hacking Social Problems

By STEVE LOHR

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

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



THE HARRIS SCHOOL
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

Computing Institute

URBAN CENTER FOR COMPUTATION AND DATA

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

Example Federal data source: LODES
(US Census Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics)

Where do Chicagoans live and work?
Select an area on the map to find out where people commute to and from work.
Read more »

☒ Inbound workers
☐ Outbound workers

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

Leaflet | Map data © 2011 OpenStreetMap contributors, Imagery © Mapbox

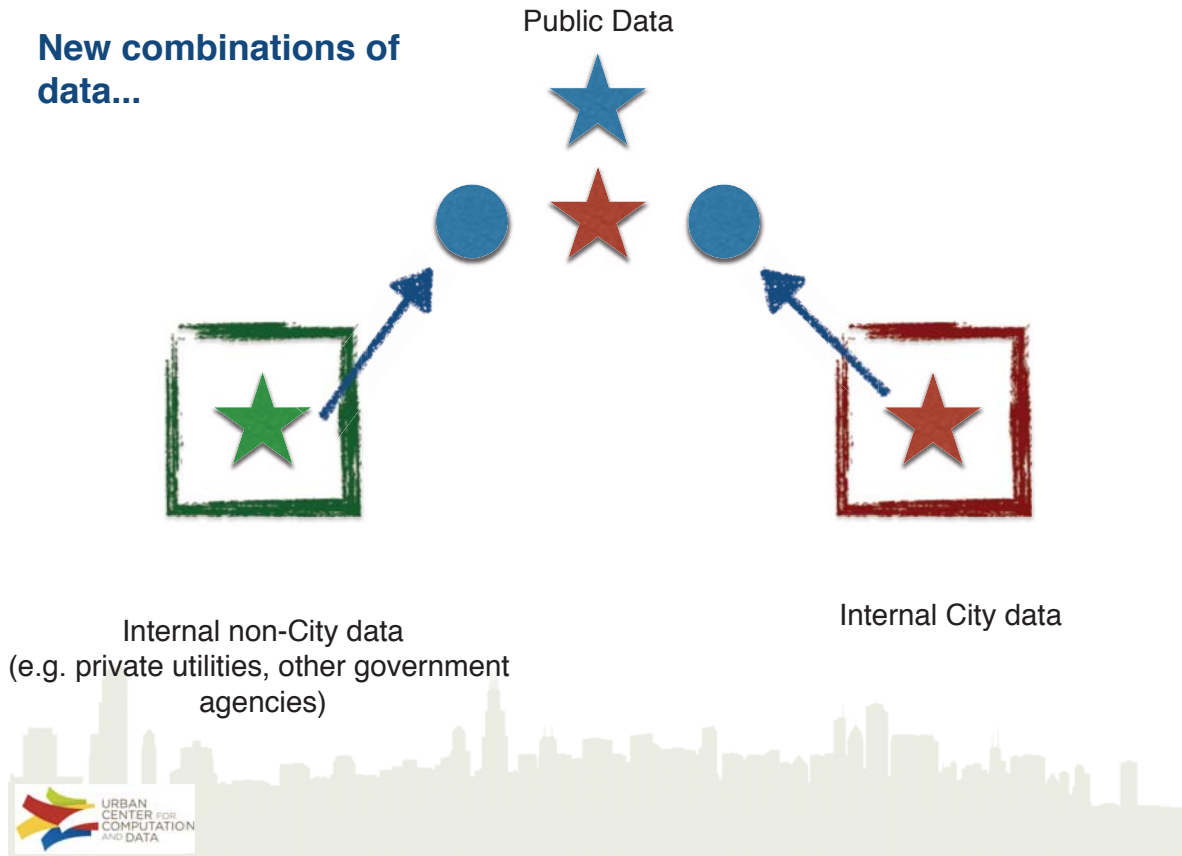
OnTheMap
OnTheMap is a web application that allows users to explore and analyze data from various sources, including the US Census Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES). The application features a map interface where users can select specific geographic areas to view detailed data and statistics.

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.