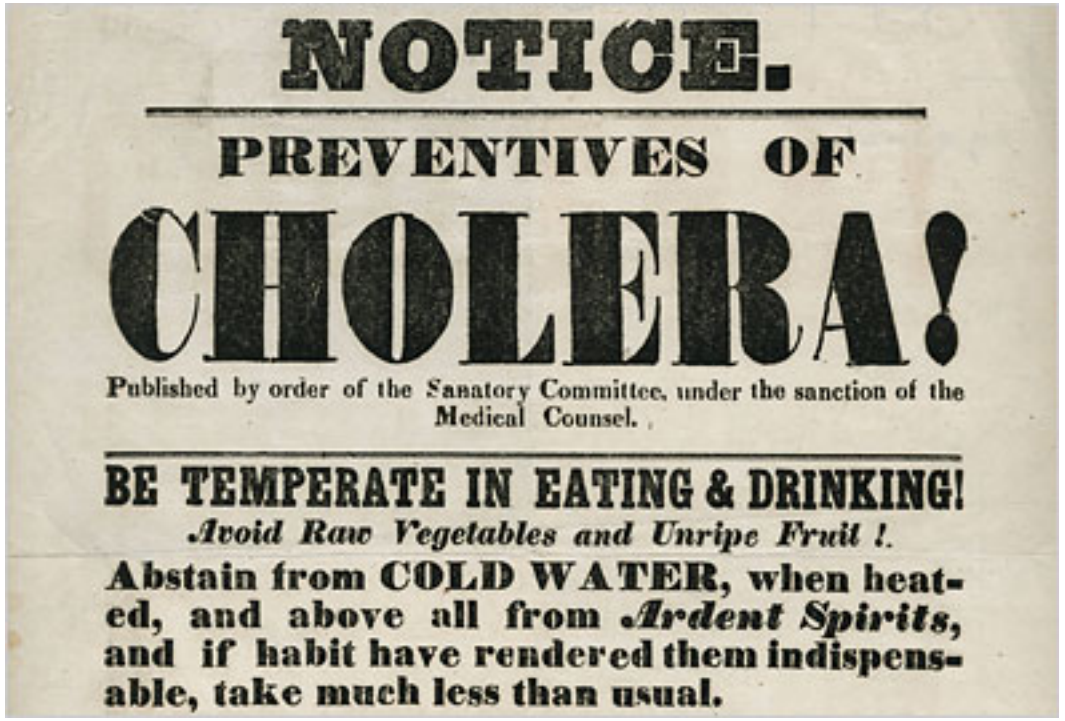
Models for secure analysis of sensitive data (CS30001)

https://ianfoster.github.io/safedata/

Snow's map of cholera deaths in Soho in the 1854 outbreak. Wellcome Library, London



NOTICE.

PREVENTIVES OF

CHOLERA!

Published by order of the Sanatory Committee, under the sanction of the Medical Counsel.

BE TEMPERATE IN EATING & DRINKING!

Avoid Raw Vegetables and Unripe Fruit !.

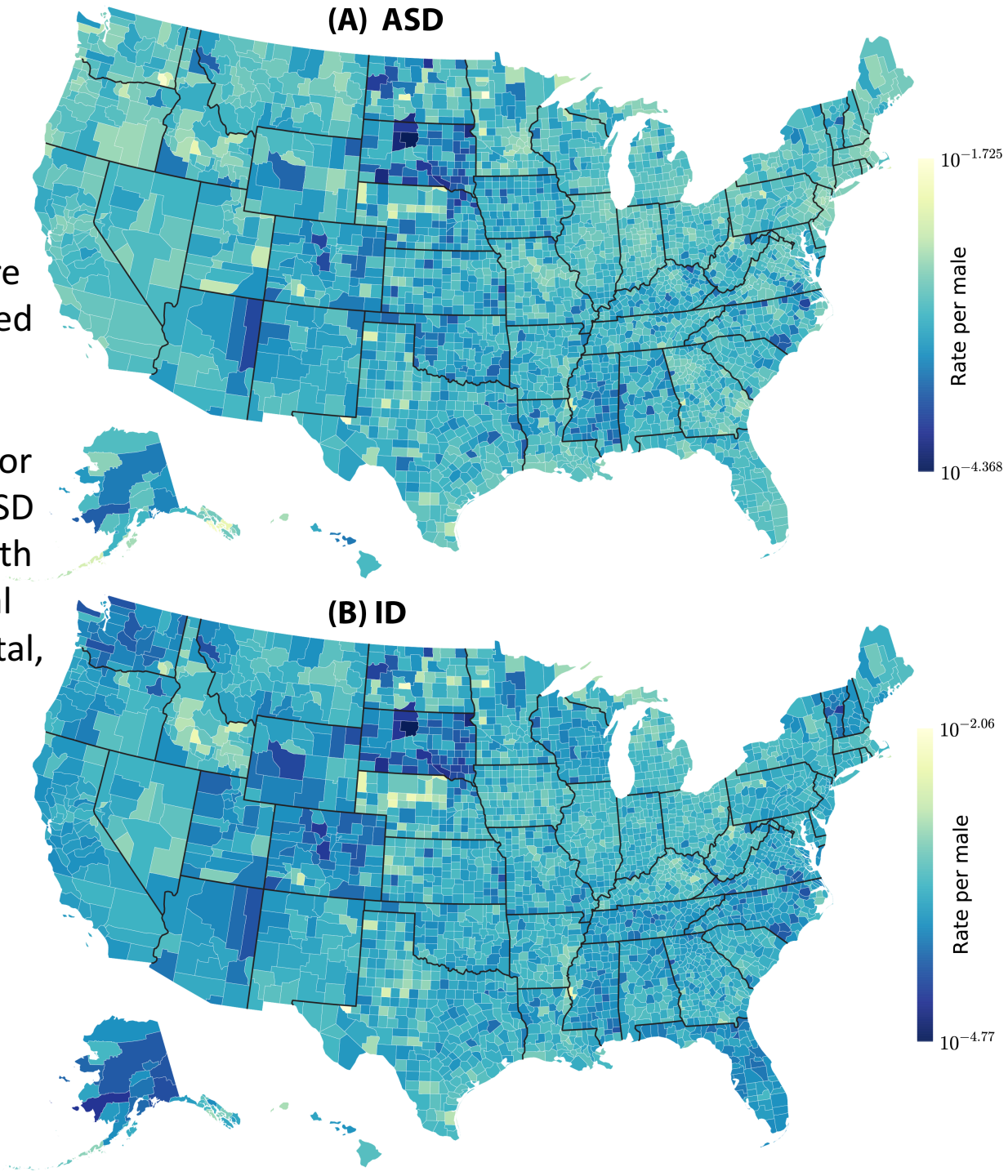
Abstain from COLD WATER, when heated, and above all from *Ardent Spirits*, and if habit have rendered them indispensable, take much less than usual.

https://en.wikipedia.org/wiki/Cholera_outbreaks_and_pandemics

Environmental and State-Level Regulatory Factors Affect the Incidence of Autism and Intellectual Disability
Andrey Rzhetsky et al., PLoS Comp Biol, 2014
http://dx.doi.org/10.1371/journal.pcbi.1003518.g002

Disease clusters are defined as geographically compact areas where a particular disease, such as a cancer, shows a significantly increased rate. It is presently unclear how common such clusters are for neurodevelopmental maladies, such as autism spectrum disorders (ASD) and intellectual disability (ID). In this study, examining data for one third of the whole US population, the authors show that (1) ASD and ID display strong clustering across US counties; (2) counties with high ASD rates also appear to have high ID rates, and (3) the spatial variation of both phenotypes appears to be driven by environmental, and, to a lesser extent, economic incentives at the state level.

The data was derived from a very large insurance claims database containing nearly 100 million patients in the United States, which was augmented with census data to introduce additional county-level covariates that captured socioeconomic, demographic, and environmental effects.



Saving Teens: Using a Policy Discontinuity to Estimate the Effects of Medicaid Eligibility*

Laura R. Wherry[†]
University of California, Los Angeles

Bruce D. Meyer[††]
University of Chicago and NBER

March 31, 2015
http://harris.uchicago.edu/sites/default/files/PDDraft_033115.pdf

Abstract

This paper uses a policy discontinuity to identify the immediate and longer-term effects of public health insurance coverage during childhood. Our identification strategy exploits a unique feature of several early Medicaid expansions that extended eligibility only to children born after September 30, 1983. This feature resulted in a large discontinuity in the lifetime years of Medicaid eligibility of children at this birthdate cutoff. Those with family incomes at or just below the poverty line had close to five more years of eligibility if they were born just after the cutoff than if they were born just before. We use this discontinuity in eligibility to measure the impact of public health insurance on mortality by following cohorts of children born on either side of this cutoff from childhood through early adulthood. We distinguish between deaths due to internal and external causes and

“The study of the labor market outcomes of ex-offenders, welfare recipients, and veterans. These data are complex—there are multiple datasets from multiple sources, in different formats, and are transactional and dynamic in nature. They are sensitive—in addition to social security number, name and data of birth, there are detailed information about their offense, disability, and sources of income. As with many new data of quite low quality, there are substantive name variations and thus identifiers are not unique. Yet, high-quality research requires that data are accurately linked and analyses are replicable. Poor links can lead to systematic selection bias resulting from non-matches and incorrect attribution of characteristics or outcomes for either false positive or false negative links. Analytical decisions need to be reviewed for robustness. And once a solid foundation is laid, it should be reused to avoid waste of time and resources and to permit rich additional analyses.”

Public Law 114–140
114th Congress

An Act

To establish the Commission on Evidence-Based Policymaking, and for other purposes.

Mar. 30, 2016
[H.R. 1831]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

Evidence-Based Policymaking Commission Act of 2016.

SECTION 1. SHORT TITLE.

This Act may be cited as the "Evidence-Based Policymaking Commission Act of 2016".

SEC. 2. ESTABLISHMENT.

There is established in the executive branch a commission to be known as the "Commission on Evidence-Based Policymaking" (in this Act referred to as the "Commission").

SEC. 4. DUTIES OF THE COMMISSION.

(a) STUDY OF DATA.—The Commission shall conduct a comprehensive study of the data inventory, data infrastructure, database security, and statistical protocols related to Federal policy-making and the agencies responsible for maintaining that data to—

(1) determine the optimal arrangement for which administrative data on Federal programs and tax expenditures, survey data, and related statistical data series may be integrated and made available to facilitate program evaluation, continuous improvement, policy-relevant research, and cost-benefit analyses by qualified researchers and institutions while weighing how integration might lead to the intentional or unintentional access, breach, or release of personally-identifiable information or records;

(2) make recommendations on how data infrastructure, database security, and statistical protocols should be modified to best fulfill the objectives identified in paragraph (1); and

(3) make recommendations on how best to incorporate outcomes measurement, institutionalize randomized controlled trials, and rigorous impact analysis into program design.

(b) CLEARINGHOUSE.—In undertaking the study required by subsection (a), the Commission shall—

(1) consider whether a clearinghouse for program and survey data should be established and how to create such a clearinghouse; and

(2) evaluate—

(A) what administrative data and survey data are relevant for program evaluation and Federal policy-making and should be included in a potential clearinghouse;

(B) which survey data the administrative data identified in subparagraph (A) may be linked to, in addition to linkages across administrative data series, including the effect such linkages may have on the security of those data;

(C) what are the legal and administrative barriers to including or linking these data series;

(D) what data-sharing infrastructure should be used to facilitate data merging and access for research purposes;

(E) how a clearinghouse could be self-funded;

(F) which types of researchers, officials, and institutions should have access to data and what the qualifications of the researchers, officials, and institutions should be;

(G) what limitations should be placed on the use of data provided;

(H) how to protect information and ensure individual privacy and confidentiality;

(I) how data and results of research can be used to inform program administrators and policymakers to improve program design;

(J) what incentives may facilitate interagency sharing of information to improve programmatic effectiveness and enhance data accuracy and comprehensiveness; and

(K) how individuals whose data are used should be notified of its usages.

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

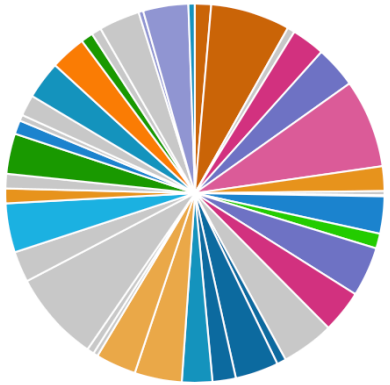
About the GDC | About the Data | Access Data

Any user accessing GDC open data must adhere to the NIH Genomic Data Sharing (GDS) Policy which indicates that investigators who download unrestricted-access data from NIH-designated data repositories should:

- Not attempt to identify individual human research participants from whom the data were obtained.

The Next Generation Cancer Knowledge Network

Case Distribution by Disease Type



Data Availability Summary

Programs	2
Projects	39
Disease Types	38
Cases	14551

The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

More about the GDC

High Quality Data Sharing Enables Precision Medicine

The GDC obtains validated datasets from NCI programs in which the strategies for tissue collection couples quantity with

The **GDC Data Portal** provides a platform for efficiently querying and downloading high quality and complete data. The GDC also provides a **GDC Data Transfer Tool** and a **GDC API** for programmatic access.

More about Accessing Data

Submit Data

The GDC provides tools to guide data submission including the **GDC Data Submission Portal**, a web-based tool for submitting clinical, biospecimen and small volumes of molecular data as well as the **GDC Data Transfer Tool**, a client-based tool for submitting large, high volume molecular data. A secure **GDC API** is also available for batch data submissions.

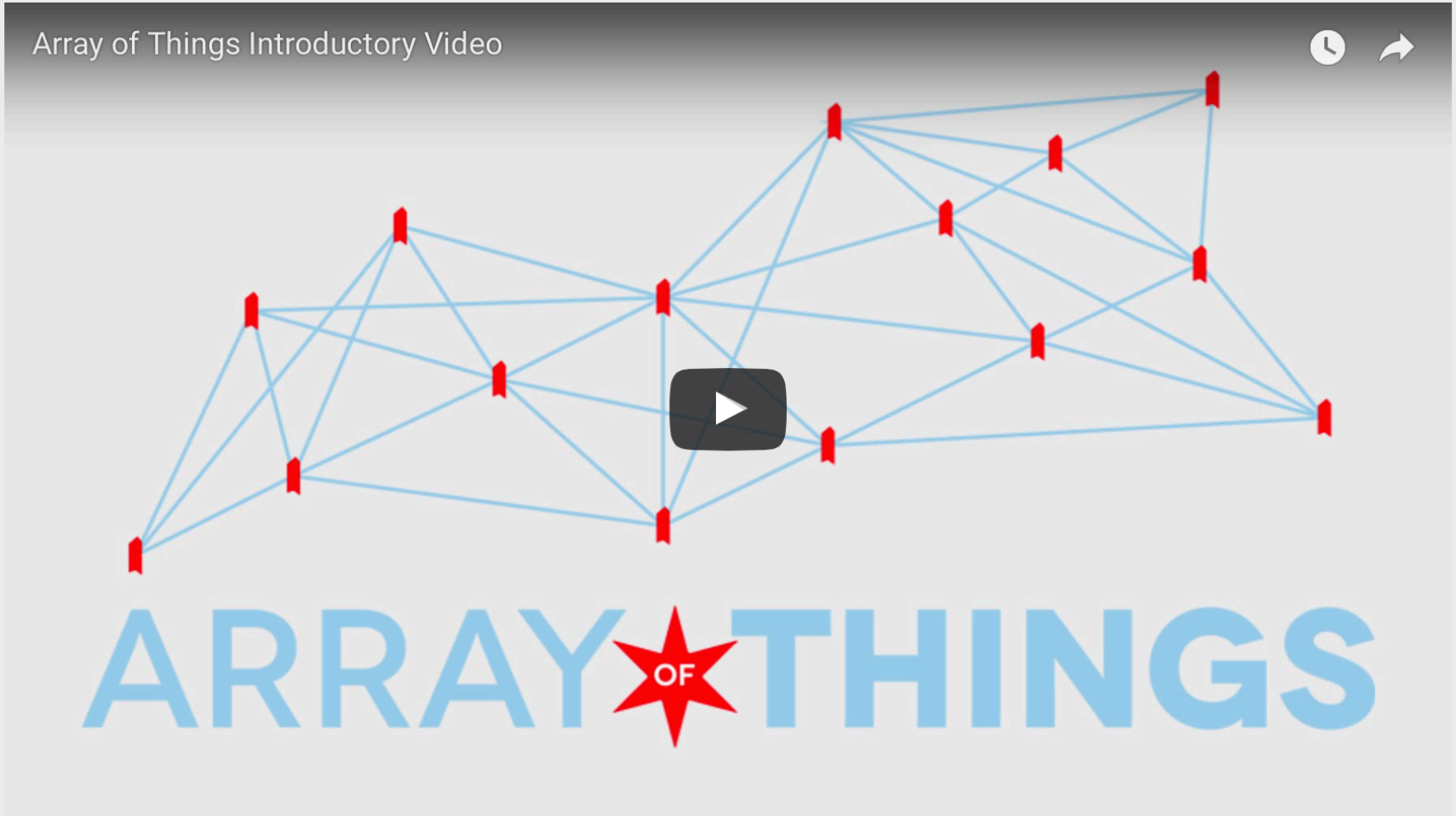
More about Submitting Data

RYAN SINGEL SECURITY 03.12.10 2:48 PM

NETFLIX CANCELS RECOMMENDATION CONTEST AFTER PRIVACY LAWSUIT

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine.

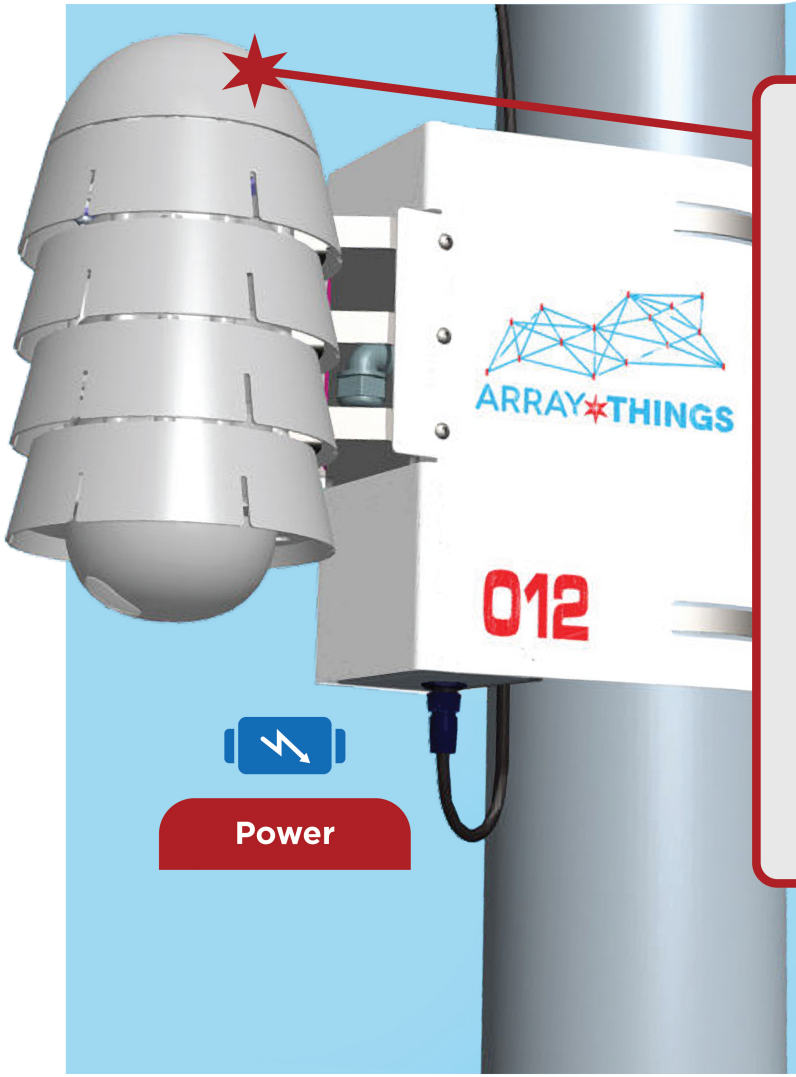
Friday's announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to expand the amount of information it gave to researchers in hopes that its recommendation system — a key part of Netflix's customer retention strategy — would get even better. That was then followed with a warning by prominent data privacy lawyers that the new dataset was easily de-anonymized.



Array of Things has officially launched in the City of Chicago! Read about the first nodes installed on Chicago intersections, watch our new video, and read coverage of the launch from USA Today and CNN.

We have published the final governance and privacy policies for the Array of Things, as well as responses to public feedback and an engagement report from Smart Chicago. We thank the public for their valuable input.

ARRAY OF THINGS ARCHITECTURE



Power

Node Components

Environmental Sensors

Air temperature, Humidity, Barometric Pressure, Vibration, Sound Intensity, Magnetometer

Air Quality Sensors

Nitrogen Dioxide, Ozone, Carbon Monoxide, Hydrogen Sulfide, Sulfer Dioxide

Light & Infrared Sensors

Light intensity, infrared (CLOUD COVER; SURFACE TEMPERATURE), camera, vehicle and pedestrian traffic. Images processed in-situ and discarded.

Linux Node Controllers

CPU CPU

Image Processing Computer & System Health Manager and Control/Communications Computer

Node Power Manager

Node health monitoring and resilience functions

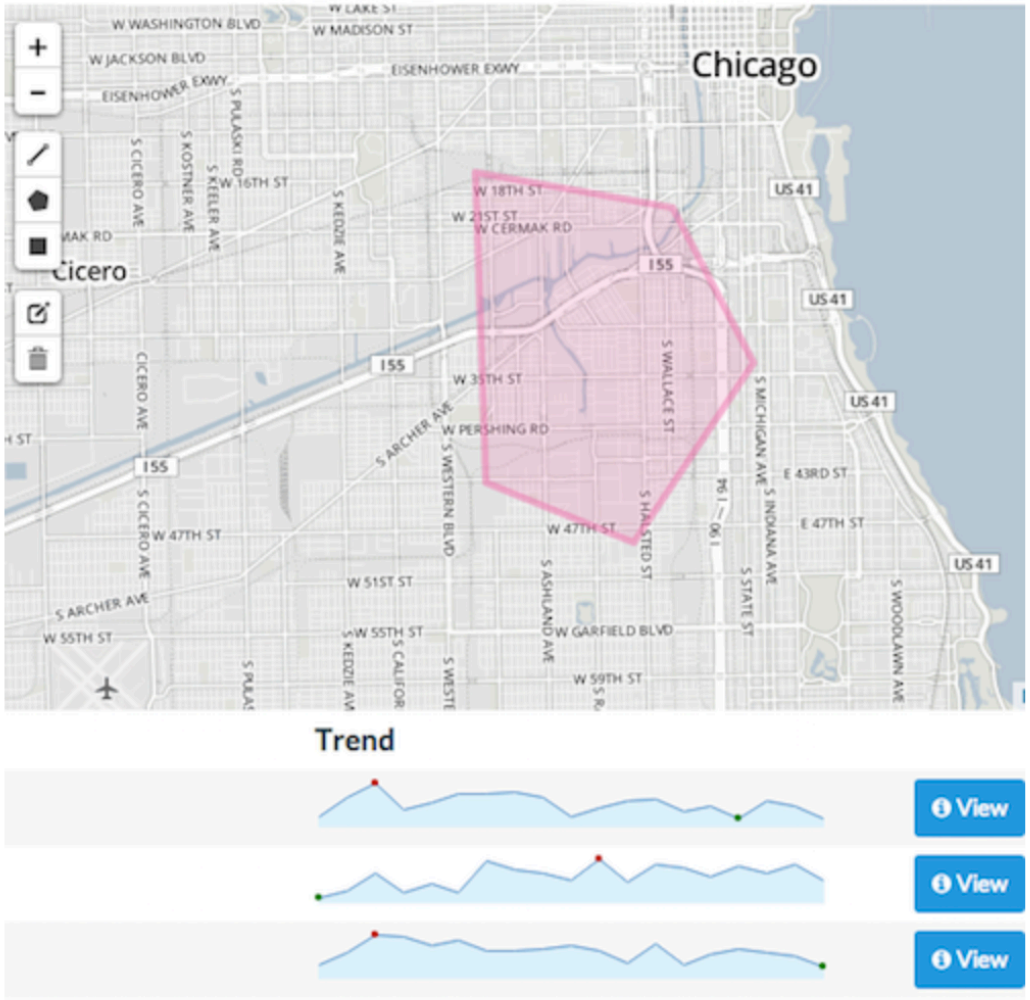
Argonne NATIONAL LABORATORY

Argonne Server

Plenario, Open Data Portals, Dashboards, and Apps

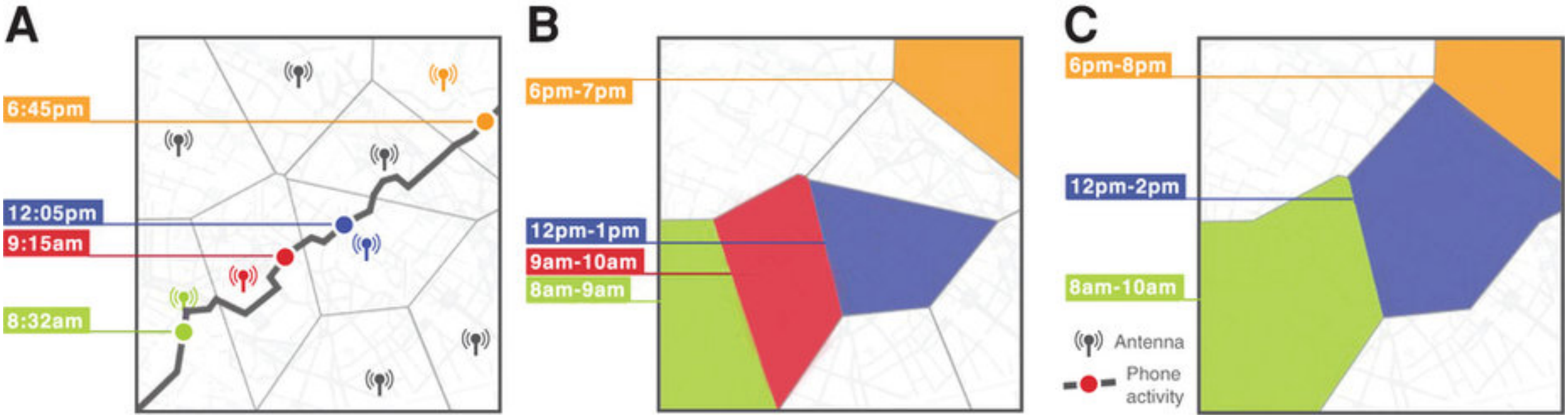
URBAN CENTER FOR COMPUTATION AND DATA

What Data Will be Available?



Data collected by AoT will be open, free, and available to the public. The nodes will transmit data to a secure central database server at Argonne National Laboratory. Data will then be published openly to allow individuals, organizations, researchers, engineers and scientists to study urban environments, develop new data analysis tools and applications, and inform urban planning. Raw data will also be posted to the City of Chicago's open data network and Plenario, a web-based portal that supports open data search, exploration, and downloading with open datasets from Chicago and around the world.

In addition, software, hardware, parts, and specifications will also be published as open source, to encourage participation and oversight from the developer community and public. You can view the architecture and current sensor list of the nodes here. Full specifications will be available soon at our Github page, when the initial node design is finalized.



(A) Trace of an anonymized mobile phone user during a day.The dots represent the times and locations where the user made or received a call. Every time the user has such an interaction, the closest antenna that routes the call is recorded. (B) The same user's trace as recorded in a mobility database. The Voronoi lattice, represented by the grey lines, are an approximation of the antennas reception areas, the most precise location information available to us. The user's interaction times are here recorded with a precision of one hour. (C) The same individual's trace when we lower the resolution of our dataset through spatial and temporal aggregation. Antennas are aggregated in clusters of size two and their associated regions are merged. The user's interaction are recorded with a precision of two hours. Such spatial and temporal aggregation render the 8:32 am and 9:15 am interactions indistinguishable.

doi:10.1038/srep01376