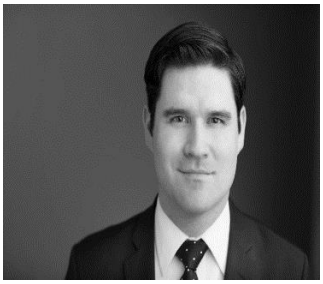
Tom Schenk, Jr.



Chief Data Officer, City of Chicago

Tom Schenk is a researcher, author, and an expert in a number of fields, including open government, data visualization, business, and research and policy in education. He is currently the Chief Data Officer at the City of Chicago, which includes overseeing Chicago's open data portal, advanced analytics team, and the City's data and business intelligence team.

He leads the strategic use of data to improve the efficiency of city operations and improve the quality of life for residents. Tom has lead the expansion of Chicago's leading open data portal, deployed predictive analytics in the City to improve data services, and has streamlined the City's data operations.

Tom wrote *Circos Data Visualization How-To*, an introductory book on using the biology data visualization libraries for use in the social sciences. He has previously served as a consultant for Institutional Effectiveness and Accountability at the Iowa Departments of Education, where he led efforts to use student-level longitudinal data to evaluation education programs, including an evaluation of Project Lead The Way and calculating rates of return for community college graduates.

He also led science, technology, engineering, and mathematics (STEM) policy in Iowa and a coauthor of Iowa's STEM roadmap. Tom was a visiting scholar with Iowa State University's Office of Community College Research and Policy where he studied graduate-student unionization. He was also a lecturer at Grand View University where he taught statistics and economics. He earned a Master's degree in economics from Iowa State University and a Bachelor's from Drake University.

Date: December 22, 2016

To: Commission on Evidence-Based Policymaking

From: Tom Schenk Jr, Chief Data Officer, City of Chicago

RE: Input on CEBP mission as defined in Public Law 114-140

Data privacy is crucial so Americans can trust the systems they rely upon for their well-being. People must have full faith in the ability to talk to their doctors, knowing that their details will not be reassociated with them again; students and parents must have faith that their grades, disciplinary records are only made to select few; and every member of the household must have faith that their responses to Census records will not be shared widely until many decades later.

Those protections should not inhibit their government from providing better, more comprehensive services. The whole sum of social sciences research has demonstrated that outcomes in health, education, well-being and other areas are highly dependent on outside factors, such as nutrition, education levels of parents and guardians, family wealth, and many other factors. Often, researchers and policymakers need to account for external factors, such as these, to help understand program effectiveness.

Enabling evidence-based, data-driven policy is crucial for governments to be more efficient and effective for Americans. This is a progression of many steps where the bipartisan coalitions in U.S. Congress and the President have made several large steps to enabling greater data sharing with the explicit goal of improving education, workforce outcomes, health, and human services. The *America COMPETES Act of 2007* and subsequent *American Recovery and Reinvestment Act* provided the targets and funding, respectively, to allow state governments to build longitudinal data systems which facilitate the ability to track students from the school system into the workforce. The recent authorization of *Workforce Innovation and Opportunity Act* (WIOA) has extended the call for linked data between education and the workforce to help job seekers.

The impact of these bills have been useful at every level of government. While heading institutional effectiveness and accountability for the Iowa Department of Education, we were able to use these longitudinal systems to conduct sophisticated analysis of the effectiveness of state-funded programs. Our team was able to calculate comprehensive rates of return to education for community colleges by each individual program for each individual college. By combining those records with state prison records, we could also determine the amount of diversion of costs by reducing the likelihood of crime. We could also determine the additional

tax revenues the state earned by those student's higher wages and lower expenses on other support programs.

Other research included sophisticated pseudo-experimental evaluations of science, technology, engineering, and mathematics (STEM) programs. These studies used sophisticated analytical techniques to follow students from middle school, through high school, and into college to judge the ability of programs to increase math and science test scores, improve high school graduation, articulation to college, and the entry into STEM college majors.

As a senior researcher with the Department of Medical Social Sciences at Northwestern Feinberg School of Medicine, we used linked patient records to understand the long-term impact of cancer interventions on not only the progression of cancer, but also the quality of life for cancer patients. And today, as Chief Data Officer for the City of Chicago, we link multiple data sources together to predict where to send food inspectors, the next outbreak of rodents, and to allocate our workforce to pro-actively

The task set for the CEBP is also enormous. If privacy is compromised or if we are unable to effectively link data to drive better decisions, it could setback progress on evidence-based policy. Based on my conversations with hundreds of residents on how governments use data, I have found that Americans are happy when this data is being used to improve lives and to make their government more efficient. However, we also know this optimism can be undermined by unethical uses of data or disclosing personal details of an individual. It is a careful balancing act that can be achieved, but takes considerable thought and attention.

II. Clearinghouse

One task laid out for CEBP is to "consider whether a clearinghouse for program and survey data should be established and how to create such a clearinghouse". Single clearinghouses of data are often too big and too complicated of a task to be done well. We all have ideals of a single large repositories of all data that can be used by every researcher. However, we must take rational first steps and focus on "quick wins". Attempts at creating websites for everyone often leads to creating something too confusing and ultimately not useful for anyone.

Instead, a two-pronged approach should be taken for the Commission to meet its target by creating smaller, linked hubs and to create a service to allow for *ad hoc* data matching.

A. Setup smaller "hubs" as initial steps to clearinghouse

In order to provide a tractable mission for CEBP, the organization must focus on data matching problems to help answer specific domains of policy questions. For instance, congress has passed several pieces of legislation focused on linking education and workforce outcomes. There are

clear outcomes that can be measured through these linked records. The initial hubs should focus on the highest priority questions, incorporating the most critical sources of data.

Additional hubs should be created to explore topics such as recidivism, health care, impacts of poverty, and other areas deemed of high importance. As new legislation is passed, Congress and the respective departments can continue to integrate data into these hubs so outcomes under the legislation can be effectively tracked and new research questions can be explored.

These hubs should continue to grow, but with a goal of eventually becoming interoperable themselves. This approach can help ensure that researchers get access to valuable data earlier while continuing to progress

B. A service to link data and remove unique identifiers should be provided for researchers

Researchers need access to a variety of linked data but while balancing privacy. One way to achieve this is to create a centralized service which empower teams whose responsibility is to match personally identifiable data for the purpose of research and then discard personally identifiable data before it is provided to researchers.

Of course, other data besides unique identifiers can compromise privacy. A unique combination of data—such as gender combined with age, race, and ethnicity—could allow someone to tie seemly anonymous data back to individuals. For instance, there may be a lone 32 year-old female immigrant in a ZIP code, so the release of that data could undermine privacy. Further steps on imputation, sometimes called “hot deck imputation”, are needed to mask personally identifiable data.

The United Kingdom's Data Services—operated by the UK's Economic and Social Research Council (ESRC), an organization akin to the United States' National Science Foundation—provides similar services. The State of Florida’s Integrated Network for Data Exchange and Retrieval (FINDER) has also set a benchmark for such services.

There should also be efforts to capitalize on the existing open data movement. Federal government agencies and hundreds of state and local governments have launched open data portals which make data easily accessible without any barriers. Linked data—at its most summary level—would be a great addition to open data portals.

III. Barriers to data exchange

A. Data needs to be exchanged between levels of government

Policymakers are sometimes surprised to find their respective government does not have all of the data relevant to their operations. In contrast to wide belief, the federal government does not have access to all of the data available for local and state agencies. There are significant and legal barriers that prohibit governments and agencies to sharing data between levels of government.

Congress and government agencies can take steps to facilitate data sharing between localities and states, between states, and between Federal governments and other government agencies. For instance, WIOA requires the tracking of educational outcomes even though states are unable to track those outcomes—such as wages and state certifications—in nearby states. Legislation should encourage, if not require, these data sharing exchanges to facilitate research studies.

B. Guidance on existing privacy laws

Some of the limited ability to share data is simply caused by misunderstanding privacy laws, perceiving or misinterpretation an inability to share data when it is possible. There will always be value in providing further guidance for governments to clarify permissible data sharing between departments and governments.

IV. Limitations on data usage

These new ideas must also evolve our notions of ethical uses of data. Currently, legislation exists on whether some can or cannot *have* data. A future clearinghouse, hubs, and robust linked data needs to consider how data is allowed to be *used*.

Further guidance will need to clarify ethical and unethical uses of data, which should extend to whether researchers would be permitted to access and use linked data. Violating these terms should remove the researcher's ability to be able to access linked data in the future.

V. Making research relevant and useful for policymakers

Research must be relevant to policymakers that answers their immediate questions and also be able to foresee other valuable research questions. A crucial aspect of relevant research is geographical relevance. National studies can indicate important trends, but often are not representative of each state or city or every participant. Research should be prioritized if it provides geographic breakdowns, breakdowns by subgroups, and other subsets deemed important for programs.

These summary statistics are incredibly useful for policymakers and should be frequently published. When data systems to talk with each other, agencies should still issue the reports that

take technical details and summarize high-level statistics. These statistics should also be downloadable.

Researchers should be encouraged to improve their communication channels for policymakers. They should rely on recent developments in data visualization research and issue short-form abstracts to quickly communicate findings and become less reliant on long-form journal articles and reports.

VI. Rigorous impact analysis

Randomized control studies have been considered the “gold standard” of evidence-based policy. However, these studies are not always possible because of logistical constraints, ethical concerns, or would be too narrow to represent overall performance of a program. Meanwhile, other sophisticated techniques, such as pseudo-experimental methods like propensity score matching, difference-in-difference, and other methods can be used to measure causal impact of programs. The framework on evaluation should consider the “gold standard”, but also a “silver” and “bronze” standard of rigor.