General overview, and sources and uses of **Big Data** for urban and regional analysis

Carson Farmer

@carsonfarmer

carsonfarmer.com

carson.farmer@hunter.cuny.edu

TRB Executive Committee Policy Session, January 14, 2015

I'm sure everyone in this room has heard the term 'big data' in the media, meetings, or on the street. It is certainly a buzz word at the moment, and so it seems this policy discussion is particularly timely. Indeed, the Obama administration has been thinking about Big Data Research for some time, with the Big Data Research and Development Initiative announced in 2012, to explore how big data could be used to address important problems faced by the government.

In doing a bit of 'big data' research of my own, I looked at the popularity of 'big data' as a Google search, and it seems we are approaching a point of saturation. It is clear that this is something on the minds of many people, but it is not always clear what exactly 'big data' mean.

The 4 Vs

Volume

- The scale/size of the data

Velocity

- The speed of data creation/collection

Variety

- The range of sources and types of data

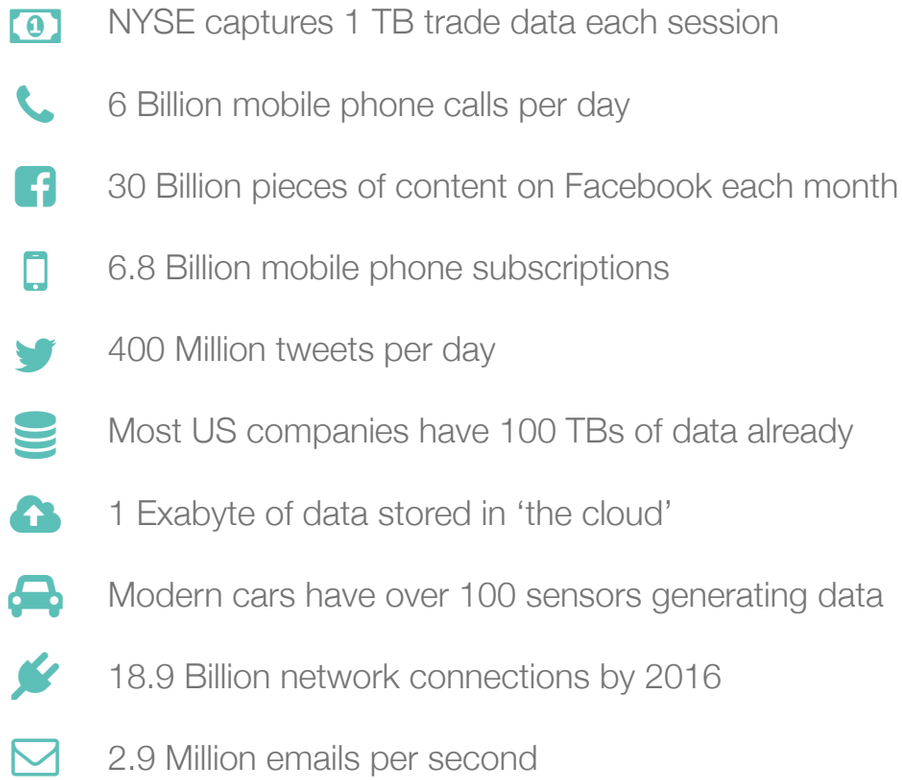
Veracity

- The uncertainly or quality of the data

Doug Laney (now with Gatner) came up with an initial three tenants of big data, the so-called 3 vs, which has helped to define the term in many people's minds. A 4th v was later added to capture the often noisy nature of big datasets. These 4 vs, volume, velocity, variety, and veracity, encompass most of what we are talking about when we talk about 'big data'.

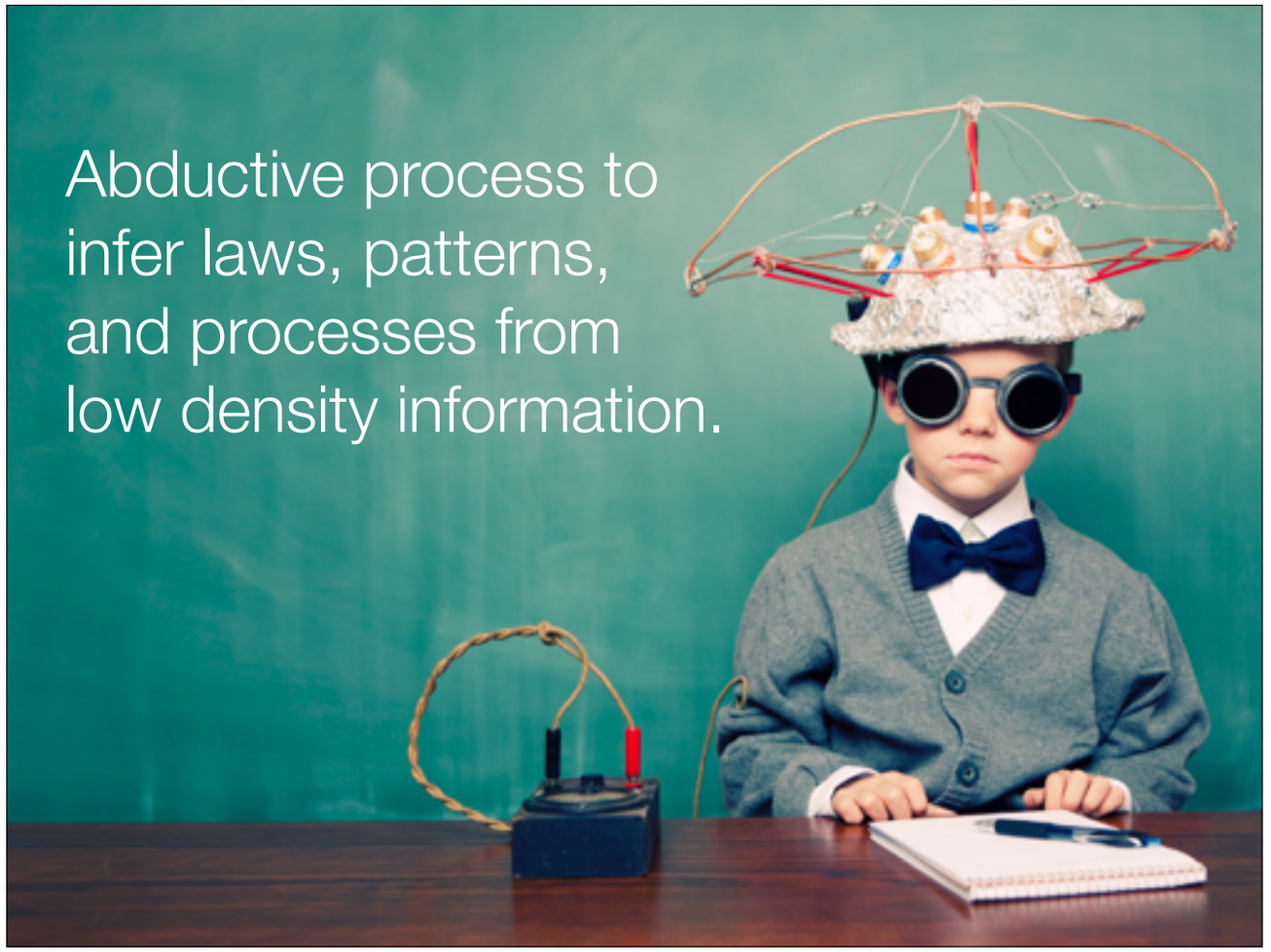
To give you an example of where this is all coming from, think about the amount of data that is currently being generated on a daily basis.

As of last year, about 2.3 zettabytes (2.3×10^21) of data were created every day. That's well over 3 Million times the entire Library of Congress collection produced, every day!



Examples of ‘big data’ are plenty, and showcase not only the volume and velocity of data, but also the variety (from videos to text to images to voice to sensor measurements). All the while, the quality of the data continues to be a serious issue. In part because much many of the numbers are based on incomplete data, or derived from unrepresentative or self-selected samples.

The sheer volume and velocity of data is slated to increase to over 43 Trillion Gbs (or 40 zettabytes) by 2020. These numbers may increase significantly as we see more and more developments in the Internet of things. But ‘big data’ is about more than just the data…



… it is also about what the data represents, and how best to try to understand it.

Many are wondering if ‘big data’ is anything more than hype? It is certainly a buzz word at the moment, but there are some clear differences between big data and more traditional analysis of large datasets.

We recognize it’s “big” but is there anything unique about data this time around?

Key differences

Unstructured

- Email, blogs, social-media, sensors (~90%)…

High resolution

- High granularity and large samples

Real-time

- Collection and analysis ‘on the fly’

Interactions

- Opinions, preferences, behaviors, sharing

The answer is likely **yes**, and these differences derive largely from the types of data to which we have increasing access: Unstructured, high-resolution data arriving in ‘real-time’, with a heavy focus on human interactions.

Joe Public is the primary data creator.

Users of various enterprise and public services, products, websites, etc.

Individuals who are shopping, working, learning, and making decisions in real-time, all the time.

Much of the data being produced has a spatial (or geographic) component. Estimates range from %50 to %80 of data that is used in decision-making has a geographic component. This is particularly useful in the context of transportation policies and issues because it allows us to use ‘non-traditional’ forms of data to conduct our geographical analyses.

Workflow

Exploration

- Where to look and what questions to ask

Analysis

- Asking and answering questions

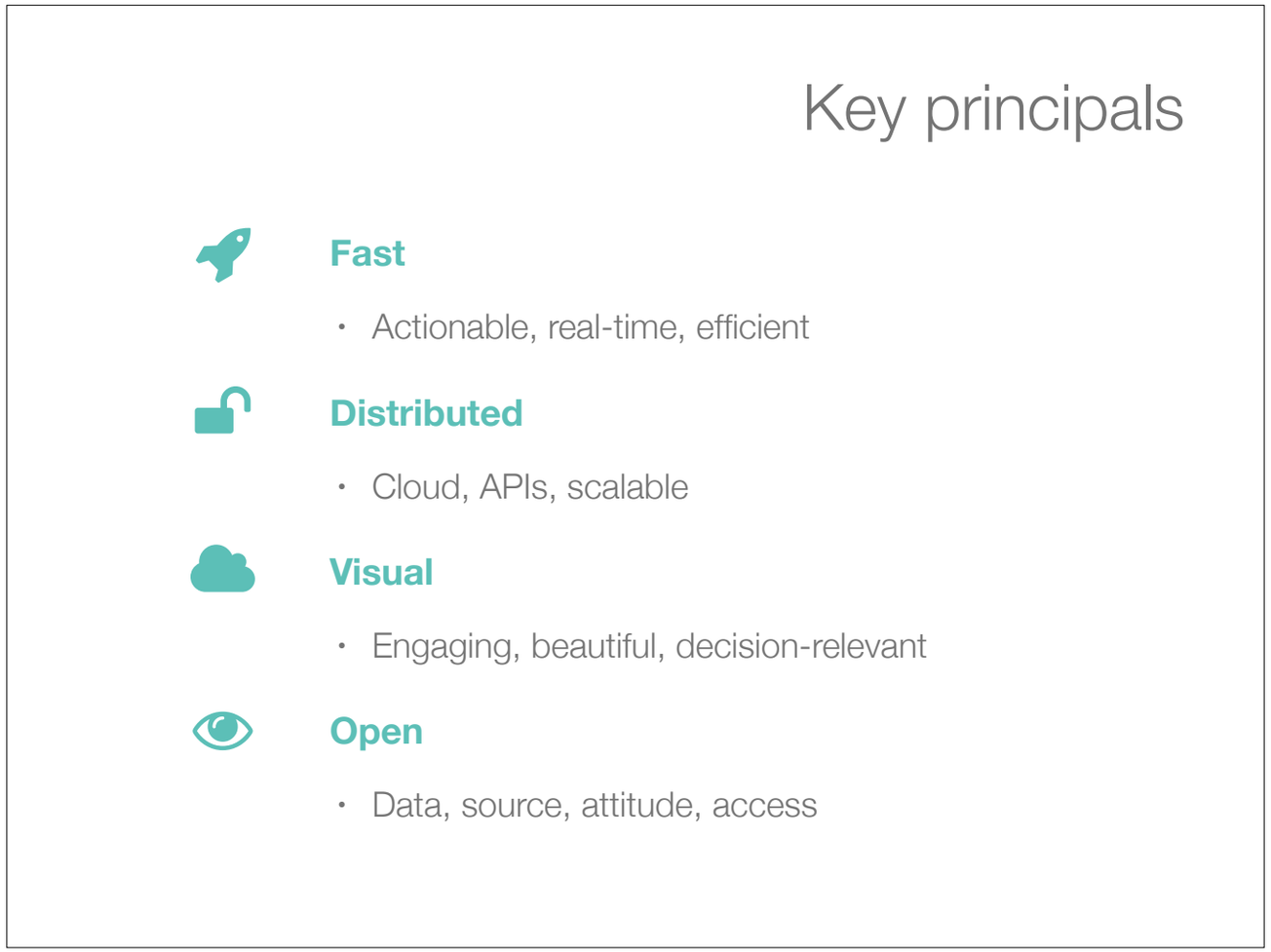
Presentation

- Creating actionable insights

Feedback

- Iterative process with regular updates/changes

In most contexts (i.e., government, academia, the private sector), 'big data' involves exploration of data in order to identify key questions to explore in depth. Building on this, a more in-depth analysis often follows to generate insights and understanding. From here, results are presented to decision- and policy-makers, and feedbacks are generated to continue the cycle all over again.



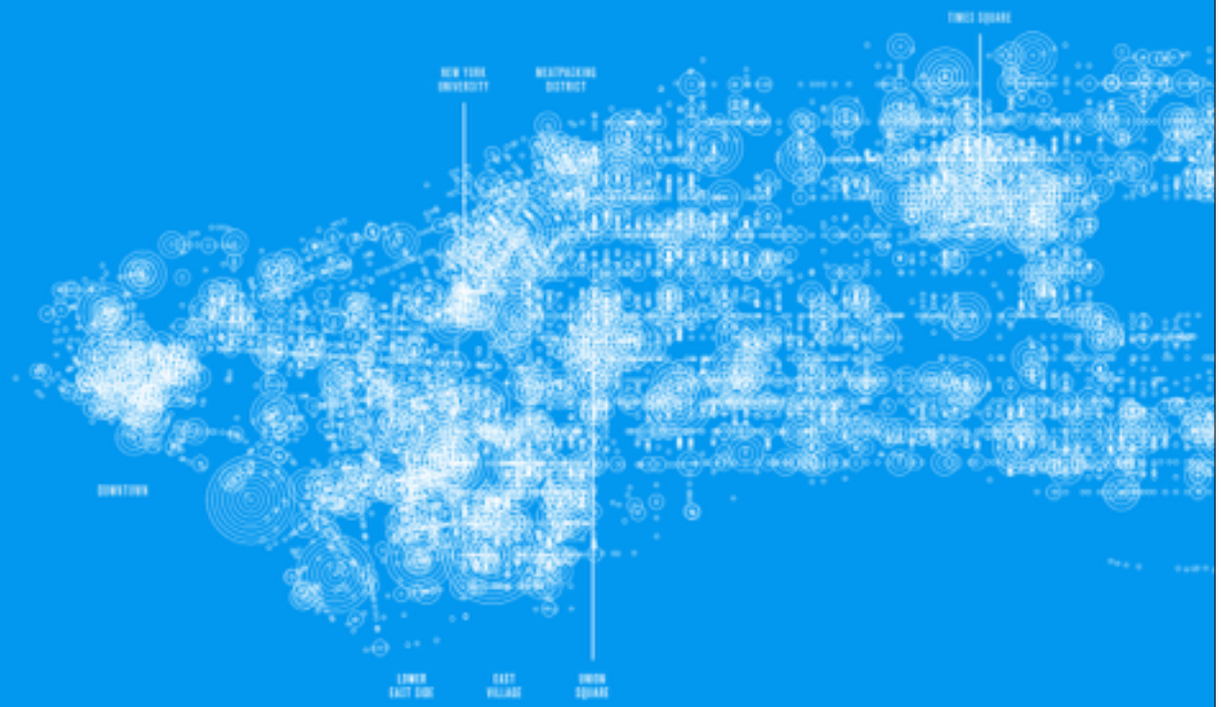
This whole process relies on several key principals that many ‘big data’ workflows are based upon. Here, I’m infusing some of my own opinions and observations, but most ‘big data’ folks will likely agree to some extent.

These key principals are that ‘big data’ analysis has to be fast, is often distributed, is highly visual, and in many cases, relies on ‘openness’. Speed is an obvious one, and so is scalability, particularly as the term ‘cloud computing’ is about as pervasive as the term ‘big data’. Visual, while intuitive, is a less obvious requirement, but is an important aspect of the “analysis -> presentation -> feedback” workflow. Openness is probably the most ‘controversial’ of my 4 principals or tenets, but I use it in a fairly ‘open’ way.

By open, I mean that ‘big data’ analysts must have access to tools, technologies, and data that they may not have had in the past. Certainly open data is important here, and an open attitude towards alternative methodologies (this is a hard requirement), but also open source, because this means analysts have access to a wider range of affordable tools. Even proprietary systems incorporate open APIs and open source components to facilitate access.

The most well-known ‘big data’ engine, ‘hadoop’ is open source and in a recent poll, the most popular analytics software packages for ‘big data’ are ‘R’ and ‘RapidMiner’, both open source software packages. Furthermore, the most ‘obvious’ example of ‘big data sources’ comes from various open (or public) APIs across the web.

Big **Civic Data**



311 noise complaints in NYC | source: http://www.karlsluis.com

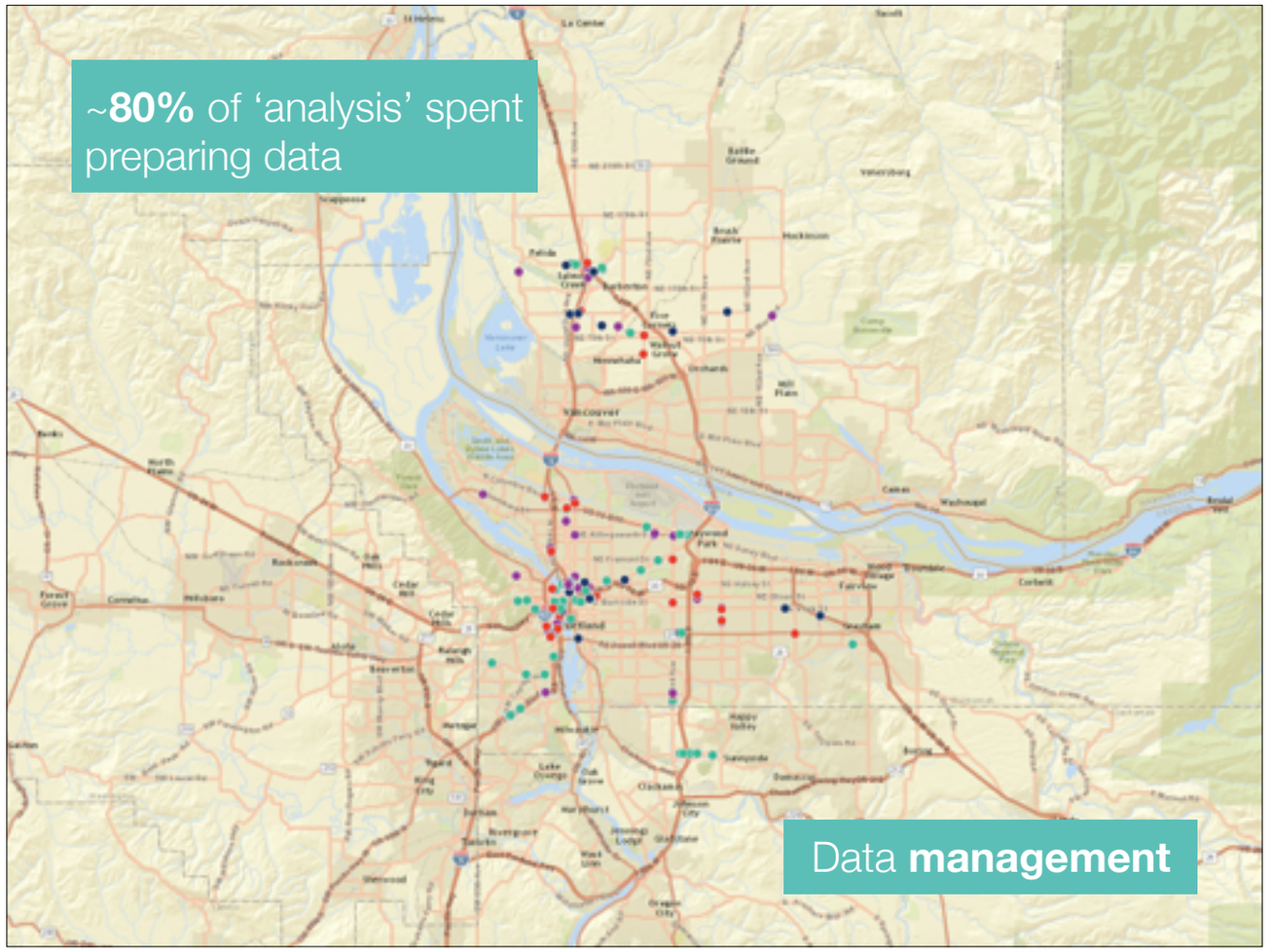
In the context of what I have been tasked to talk about today, there are many excellent examples of sources and uses of 'big data' in urban and regional analysis.

Sources of 'big data' include many of the social media data sources mentioned previously, as well as the increasing amounts of publicly available data that are produced by governments and cities.

Examples include: Traffic counts, 311 calls, Bluetooth sensors, Smartphone survey data, Mobile phone tower intensities, Civic data (via portals), Taxi and bikeshare data, Police and infraction reports, Parking violations, Census data, and much much more.

With the increasing VARIETY of data sources available, and the increasing ACCESSIBILITY of more traditional data sources, 'big data' analysis promises to tackle 'bigger' and 'boarder' problems.

This beautiful map comes from designer Karl Sluis, who mapped 311 noise complaints in Manhattan. This is more 'art' than 'science', but it illustrates nicely the types of 'non-traditional' data sources available to urban planners and policy makers as part of the 'data-driven' revolution.



A recent study has revealed that researchers spend about 80% of their time ‘playing’ with their data in order to make it usable. A major part of the ‘big data’ challenge is therefore, data management and integration.

An illustrative example of this type of work is given by the regional transportation data archive for the Portland, OR-Vancouver, WA metropolitan region, Portal. This system integrates multiple data sources, to provide a level of data integration required for ‘big data’ research. This is just one example of the increasing number of data portals and infrastructures available for ‘big data’ research.

An example of 'big data' USAGE comes from researchers at my own institution, CUNY, who have looked at both bikeshare systems and taxicab movements in the NYC region to examine social justice issues.

Many cities have taxis, but the question is, what are they doing? In NYC we have 12,533 Yellow Cabs, running about 500,000+ trips per day. With over 40 months of data, this equates to well over 600 Million records!

By combing through all of these taxi pickups and drop-offs, researchers were able to identify areas with excessive or limited taxi activity. Their findings revealed that about 50% of trips start and end south of 59th street in Manhattan, and that 80% of all trips begin in Manhattan. Higher income residential areas were better served than lower income areas, and the average trip length is just over 10 minutes.

They were also able to identify many out of area trips, with over 360,000 trips starting in New Jersey! Additionally, they were able to identify different driver behavior patterns; with some drivings optimizing their fare to driving ratio quite well, and others much less so. These types of findings may provide useful information for planners and taxi commissions seeking to optimize driver efficiencies and other taxi-related questions.

Additionally, bikeshare systems have been used to identify accessibility and demographic issues. Researchers were able to show that, based on usage patterns, the fee structure of NYC's bikeshare program is may differentially impacts different types of users (e.g., mean vs women). One suggested impact was a bias against how women tend to use the system. Using the bikeshare data, they were able to devise an alternative payment system that improves gender payment balances through a credit-based pricing. Ultimately, these bike share systems use city resources (expensive parking spaces), so some level of equity is likely desired. {for details on this work, please see the relevant paper: https://dl.dropboxusercontent.com/u/35674979/CFP/proceedings/bduic2014_submission_57.pdf}

Both of these analyses were based on large datasets, with extremely high granularity (individual trips), high spatial and temporal resolution, and the analyses used a combination of visual and statistical techniques.

My own research group, in collaboration with colleagues at 52 Degrees North in Germany, are developing a mobile app and associated data infrastructure for collecting continuous environmental data in near real-time, using existing vehicle-based technologies. Here we use a vehicles on-board computer to estimate CO2 outputs, which then provides a broad picture of CO2 across an urban enlivenment.

I won't dwell on this application, particularly because it is a bit self-serving, but this is just one example where 'app-based' research is contributing to 'big data' analytics. We are now in the process of analyzing much of the data collected in Munster Germany, and are using this data to develop models to estimate CO2 outputs in New York State. {Please contact me for details on this projects and its outputs}

The proceedings from a recent NSF sponsored workshop on ‘big data’ and urban informatics also provides some excellent examples of the use of ‘big data’ for tackling important urban issues.

Example themes included: Analytics of User-Generated Content, Data behind Urban Big Data, Big Data for Urban Plan-Making, Changing Perspectives with Big/Open Data, Urban Knowledge Discovery, Health and Well-Being, Urban Data Management, Livability & Sustainability, Insights into Social Equity, Emergencies and Crisis Informatics, Urban Knowledge Discovery, as well as an entire session theme devoted to urban knowledge discovery for Transportation.

Additionally, several sessions and panels on Big Data for Urban and Regional Analysis has been scheduled for the next American Association of Geographers AGM in Chicago in April. This promises to bring together a number of key researchers and policy-makers in the area of ‘big data’ to discuss the potentials and pitfalls of ‘big data’ for urban and regional analysis.

Issues with Big Data

Speaking of pitfalls, before we declare that ‘big data’ is a panacea for everything, there are some key issues that we need to keep in mind during our discussions. Many of these issues are specific to ‘big data’, but some are general to quantitative research more broadly.

Correlation is not Causation

- Good at detecting correlations, not meaning

‘Big Data hubris’

- A supplement, not a replacement

Abuse, manipulation, and robustness

- Grading systems and Google Flu trends

Heterogeneity and non-stationarity

- Good at finding trends, not differences

Multiple comparisons

- Eventually, significant results will arise

Over quantification

- Unjustified appearance of exactitude

Representativeness

- Not a random sample, often self-selected

Validation

- Difficult to assess quality and find bugs

Privacy

- Probably the biggest policy issue

The hype

- The end of theory? Not likely!

Summary

New forms of data

- Confluence of many ‘non-traditional’ data sources

New way of doing research

- Computational, data-driven process

New questions and situations explored

- Real-time analysis for real-time results

New issues to be discussed

- Privacy, representativeness, utility

Carson Farmer
@carsonfarmer
carsonfarmer.com
carson.farmer@hunter.cuny.edu

Hunter College, CUNY
695 Park Avenue,
New York NY, 10065

carsilab.org