NNIPCamp Columbus, June 19, 2013

**Session 1: Communicating Complex Data Issues**

Led by Tim Bray, Dallas

Notes by Rob Pitingolo

How do we keep questionable data from becoming fact (Bray)?

* Example: someone in Dallas said 9% of HS grads are college ready, now it’s repeated and floated despite uncertain source. (Bray)
* Example 2: how do we communicate confidence intervals on ACS (and other) data? (Bray)
* Example 3: what do we do when a survey changes methodology and the numbers change? (Bray)

Let’s get good examples of when data was unreliable (Milwaukee).

* Example: Buffalo’s poverty rate was really high then really low then really high again. Was the middle data point an unreliable data point? (Michael Barndt)

Idea: don’t release jpeg maps, all maps are in pdf format with commentary attached.

Idea: only release data when it’s statistically significant. Even when graphs have error bars people misinterpret it anyway so we have to make the judgment.

There’s an ArcMap extension to incorporate MOEs into maps (author of the extension works at GWU, name of the extension not known). (Laura McKieran, San Antonio)

Decision makers don’t always know how to use data. We’ve got an obligation to push them to help them (Kingsley, UI)

Simplify uncertainty by labeling estimates good/ok/bad instead of using MOEs. People don’t read methodology sections or endnotes, so you have to make it explicit. A little red stop sign is a powerful visual (Green circle = .12 Coefficient of Variance (CV) and below; yellow square = .12 - .4 CV; red stop sign = .4 & above CV ) (Bray)

Question: Has anyone ever called out someone for using data ‘badly’?

* Lots of behind the scenes conversations about questionable data reporting
* It’s ok to be the bad guy. Best to shut down bad data before it spreads and you can’t anymore.
* Being the bad guy works better if you offer an alternative.
  + Example: the data isn’t reliable at this small geography, but you can use a bigger geography to get a better estimate

Community people often need a basic explanation of ‘what data is’ especially if it’s not ‘theirs’.

User education around data is important – teaching people to ask the right questions about the data. Offer alternatives to bad data – incubate a new data source (McKieran)

Have a short presentation with some simple concepts (Bray)

* What is a rate vs. a number (Bray)
* What happens when sample size is small (Bray) – bus crash that never happened
* What is noise (Bray) - e.g. weighing yourself multiple times during the day

San Antonio does a data literacy training.

Tim Bray has example slides that he can post on the NNIP websites.

Statistics is not created equally and context matters.

* Stats for epidemiologists is not the same as stats for social science
* We can learn things from hard scientists

Politics often drives decision making, providing evidence to drive decision is a big win (Millea, Austin)

Uses pals from universities to form a technical advisory committee for projects to review them. Compromise between pure and useful is necessary. (Millea)

Some of the data work is a balancing act- more than technically correct (Bray)

Use ACS data more as background than the focus of analysis (Millea).

What is something is “statistically significant” but based on a small sample size? (Bray)

* Example: Teen unemployment 25% in the south, 75% in the north, based on 60 interviews. Can we do anything with this? (Bray)

Cities should have a “data ombudsman” maybe who writes for the local paper or something (Kingsley)

* Idea: the guy on the local news who reads health dept reports and makes a big deal about restaurants with violations, can we do that for data?

Crime data has problems, inconsistent reporting ,etc. Media has started to learn more about the data and whether changes are real or based on a formula or something. (Janikowski, Memphis)

Help others to be the hero on data collection. We worked with Mayor’s Initiative on Infant Mortality (mayor and staff) to get them to report 3 year rolling averages. But weren’t successful on getting the right indicators presented through the Teen pregnancy Initiative. Sometimes not a lack of understanding but politics. (Katie Pritchard, Milwaukee)

Teen pregnancy rate is a tough one to understand. The percentage of teens who give birth is different from the percentage of births to teens – lots of people don’t get this (Bray)

Texas pulled past BRFSS (behavioral risk factor surveillance survey) data from their website because there was change in methodology and they didn’t want people making comparisons (Bray).

Similar issues with changes in the way population calculated (due to Census/ACS) and used in the Uniform Crime Report (UCR) rates (Janikowski).

We need education to help people realize where these numbers come from.

Race is not constant, in the sense that the choices aren’t the same on every survey, need to know how different databases code race.

Once a number goes into the newspaper it becomes gospel truth even if wasn’t the right number to be using (Bray)

The dirty truth behind the Kids Count report is that the rankings are often misleading but it’s a perfect media hook. A new feature is the rank but a graph to show if how states are “clumped up” for an indicator to address this concern. (Guy, AECF)

People have trouble understanding disproportionality, whether people are over or under represented (Millea).

Infographics are getting big.

* They’re nice because you can annotate the graphs better
* There are some really bad infographics floating around out there, so be careful
* Like them because you can put the words in the jpeg itself (and not detached from the graphic)
* Need more community of practice around infographics

Poverty is a good example of a messy indicator.

* Explanation: we used to know how many people are poor. Now we know how many people are poor, averaged over 5 years, only if the person has a phone and answers the phone, etc. etc.

There are lots of goofy incentives to watch out for in data collection.

* Doctors code a ‘broken arm’ as just a broken arm, even if it’s the result of a domestic violence crime, because doctors don’t want to go to court and deal with it.

Private sector has a lot of data we can’t access.

* We need to partner with them now
* Private data is fraught with issues which are a new frontier that will have to be dealt with.

Good big data article in Foreign Affairs. <http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data>

The risk is that people want to stop using data at all.

* Example: “if the data has all these problems, let’s just throw it out and use anecdotes”
* Need a balance between a disclaimer and data issue fatigue

Data training is fine but the audience is key (Barndt)

* Training police is different from training the public to use crime data (Barndt)
* Things concern the police that don’t concern the public and vice versa (Barndt)
* The public goes nuts when crime numbers fluctuate, the police expects fluctuations (Barndt)
* If this isn’t dealt with properly the public starts thinking that the cops are cooking the books because the data is changing all the time (Barndt)

Sometimes data changes for causal reasons (like better reporting vs. a change in incidence)

* STD detections go up because more people are getting tested
* From a health perspective, this is great, but it can freak out the public

Suggestion: put the cautionary stuff **up front** so that it’s not too late after a bad number gets out. Semantics are important.

Narratives help describe graphics, but if they’re too long people just skip it and go to the pictures.

Put stuff in the graphs (like a double break to indicate a break in series due to changes in methodology) that make people ask what’s going on.

Putting data out there can create side effects.

Report context around data and initiatives (Janikowski).

Closing thoughts: Uptown in Dallas is seen as a success story because economic development saved a bad neighborhood (less crime). Problem is that the neighborhood may be ‘saved’ but new people live there now. (Bray)