Session 2 – Tools of the Trade: Data Analysis and Visualization

Leader: Sean Capperis

Present: Denise Linn (Smart Chicago), Eleanor Tutt (Rise), Michael Schramm (CUPR), Zach Czczepaniak (UNC Charlotte Urban Institute), Audrey Spiegel (Atlanta Regional Commission), Natalie Young (Demography Utah), Joel Stewart (The Providence Plan), Megan Swindal (The Providence Plan), Nic Moe (Children’s Optimal Health), Sean Capperis (NYU Furman Center), Katherine Hillenbrand (Harvard Kennedy School), Alejandra Acero Murillo (IUPR – UT Dallas), Omar Rivera (Newark)

Sean Capperis – Let’s start by saying who you are and where you're from. One of the reasons I'm coming to this is to know about other tools that could be helpful. So something you use already and things that you're interested in learning about? Maybe we can do a little bit of a demo? Before we start the intros, are people cool with tweeting? [Yes.] I'm Sean Capperis at NYU Furman. I'm a very traditional analyst, SAS, Stata and ArcGIS. I'd love open source and web based tools.

Katherine Hillenbrand - I'm writing about data analysis but I don't actually do it myself.

Alejandro Acero Murillo - Used a lot Stata and Excel. I would like to learn more about SQL and I would like to know more about MATLAB?

Denise Linn- With the Smart Chicago Collaborative. I started flirting with R but I would like to learn. I'm interested in learning more about open source tools.

Omar Rivera - From the Office of Information Technology in Newark and I use ArcGIS.

Eleanor Tutt - I use a billion different tools but that's because I have a lot of neighborhood-based clients. I use QGIS, R, JavaScript libraries and I use TurfGIS

Mike Schramm - Traditional ArcGIS and SAS but would be interested in learning ETL. I spend more of my time with ETL functions cleaning stuff but are there better open source tools that can be used with a website.

Zachary Szczepaniak – With UNC Charlotte. We use the Arc suite. I'm interested in D3 for data visualization.

Audrey Spiegel - I would like to learn the programming languages and also D3 as well.

Natalie young - With with Salt Lake City. Use ArcMap and Stata, learned a little bit of Python but not really applying it. We have a SQL server and I use it for our database work.

Joel Stewart - Providence Plan. Mostly use R as my Swiss Army knife and we use the arc suite and we're starting to use tableau. Also have a secret crush on Python.

Megan Swindal - I don't use R, more of SPSS and ArcGIS and Tableau but we're starting to have internal conversations about knowing SPSS plus Rand more of that programming language. Does everyone have to be up to speed in the same approaches

Nic Moe - Austin. I'm into data visualization so my main toys are Tableau and ArcGIS. We use R quite a bit. I want to know more about the JavaScript libraries. I go to every D3 meetup so I just watch. We've tried to play with Ultrix.

Sean - So we've got a long list of things that overlap. So to just kind of look over the list, there are legacy technologies with some things that are open source and new, some things that are web based, and other things that are new. SAS Stata and SPSS in the first camp, somewhere in the middle is SQL, moving to R and JavaScript, and Tableau being kind of proprietary. Maybe we can figure out why we use some of the tools we're using? I mostly use SAS because of path dependency. Most of our data sets and processes are in SAS so I just keep using it even if there is a better tool to use. A related question is, are any of you who got out of a legacy language and go to something more open source?

Mike - So legacy programs are sort of a stewardship thing. A repeated task being done in the right order. Do some of these open source tools come with a script as well?

Sean - Do you ever do things in a new environment just because? If for example you suspect that Stata might be better?

Mike – No, we stick with the legacy program.

Eleanor - One of the reasons I use the tools is the budget. We're newer so we don’t have all the legacies. The technical debt of using legacies is high do we use free and open source tools that many universities don't.

Sean - Do people in your organizations hire people who know these technologies?

Omar - We use what you’ve got. It helps if you have someone on staff who's willing to learn. Our debt means that we're kinda stuck. It's extremely tough to adopt new things.

Sean - So adoption is tough. The people who are there to learn are actually helpful. We might be stuck with what we're doing. Growing into one of these technologies means that you need new thing. Let's go through and talk about these new technologies:

People want to learn something about SQL, why do you like it and what is it good for?

Natalie - SQL is good for manipulating data and doing things in batch. If you're working with a large set of data. If you want to mark every tract with people above a certain income. SQL is a language you can use to make relationships. The relationships are sort of set. It's a way to manipulate and add to your data,

Sean - So Oracle is a propriety version of SQL, Post Press is one of those other databases. I think personally as a SQL user, I find the language a little rough to learn but the syntax is so clear. It’s the same syntax over and over to do different things.

Megan - Can you ground it a little more?

Mike - Has anyone used her query tool in ArcMap? A simplified version of that. I think it's helpful to do a complex query with a bunch of “ands” and “ors” in there. The other thing that I really like is that there’s powerful date and time processing in there and you can tabulate things by month. But then it's so powerful. It's been used for so long

Maia - I’ve used SQL steps in SAS before. Is that the same thing? What’s the use of learning SQL if you can use the script in other programs?

Mike - A SQL based server could be a good place where you could serve data. But if it's something with just a ton of ad hoc stuff, I would not add SQL to your toolbox.

Sean - So we used to have a SQL server, we asked way to much of it. I don't know a ton about it and now we're rebuilding it from SAS and using SQL script. We use that server for a database that we have a facsimile of, so we take their data and the primary key and he tables are indexed. That's two of the benefits of SQL. But that takes a ton of time and energy and setup costs.

Mike - If you’re wanting to serve data through a web portal, that's when you want a SQL data portal. But if you're just creating an integrated data set, the processing time for SAS is not that efficient, you could get that performance that Sean is talking about.

Sean - There was some talk of R and that's great because it's free.

Joel - There's so much strength in the libraries you can access. I use it to clean up data, to do exploratory data analysis, to do visualizations as well until we started transitioning to Tableau.

Sean - How did you come to use R?

Joel - R Commander or R Studio as an undergrad. But it was for an introductory kind of lesson.

Eleanor - If you're interested in trying out R, download R Studio.

Joel - R Studio is an interactive environment, it'll show you visualizations and you can access documentation help.

Sean - I've also seen people use it for making thematic maps?

Natalie - You can get an actual map out of it which is really cool.

Mike - The SAS ArcGIS paradigm is awful,

Natalie - We wanted to join the 2000 census blocks for the 2010 tracts. That's a lot of blocks because we're doing the whole state. We could do it programmatically through R.

Audrey - In terms of visualizations what made you change from R to tableau?

Joel - Institutional interest, and it was helpful because everyone was using it.

Megan - if you could choose, would you choose R because it has more power?

Joel - The cost of it is tough, it would take more time in R and in Tableau it would be faster.

Sean - What's the elevator speech for tableau?

Joel - sSuper user friendly. I'll just do it real quick in R and work with it in Tableau.

Nic - You can't do anything with the data in tableau. It's this iterative process because Tableau isn't always happy about the data structure.

Joel - Do you use a free version?

Nic - Yeah, structurally it's not any better in the paid version but it has a lot more in the data security. The free version used to be that you only had a little bit of space and you couldn't hide anything. And you can't hide anything. We need to aggregate it first so Tableau didn't always visualize what we want.

Zachary - Now we can connect to my SQL database with the paid version, and on the user side, the way you set up dashboards within our activity is incredibly useful. Now we're looking at something and changing the kind of visualization that's harder to implement with other software,

Sean - Talking about d3?

Eleanor - I use it. I don't use it well. I was using Protovis which is like the pre D3. But it's great, it can handle all your different projections. I use it for network diagrams and like measuring collaborations, and there are tons of examples. If you use Github, you can make a gist and view it in real-time on something called blocks. A gist in Github, instead of starting a whole repository of code, it's like a Post-it note.

Nic - If anyone wants to play around, you can do it at [codepen.io](http://codepen.io). You can search by the language, you can do what's already been built, and in real-time you can play with the data. And a cool way to learn how people think through it.

Sean - When you work with D3, are you working with the platform itself or like high charts?

Eleanor - I grab the libraries and grab people's examples and tweak them. I have seen a few charting libraries.

Megan - As we're building out this micro-site, we were thinking around playing with D3 but the charting libraries was much easier. We have students who come with no coding at all and the charting libraries are way easier to learn with.

Mike - Rather than having a big enterprise system, I think the way to go is topical visualization tools. You can have a student look at something like that. Food stamps over lead data or something like that.

Sean - Do you have a problem with people who create or have institutional knowledge and the leave?

Mike - Right now we're hiring for a new programmer. At a university it's hard to find someone who can stay for a long time.

Eleanor - Tell people to comment the code and be explicit about that. Probably an opportunity for our network too. Are there code snippets or analysis that we can work together?

Mike - My halfhearted suggestion—I think we should just have a whole event dedicated toward tech and tools and stealing ideas from each other. But something more applied, like half a day on R or D3 or Tableau so these major things we’re talking about can happen. So that sort of happened in Pittsburgh with the half day with Bob but if we dedicate to some sort of tool, but that's another expense and another meeting,

Nic - It could be a great sponsorship opportunity for those companies. In Austin or New York or something, so a group like this could be really helpful. It could be a really cool opportunity. This is about stealing each other's ideas.

Sean - Maybe these companies can use it as research experience,

Eleanor - Would be cool to build in a hack day and we could build tools that we could all use

Nic - Even having, like maybe an internal Google doc? Like what expertise each group has, but you guys are doing a lot of similar things.

Sean - So on the bio pages on the website, you can put down a bunch of topical expertise. We should put down tech expertise. Or even like data sets too. Turf sounds cools, I've heard people talk about Turf.

Eleanor - I'll just pitch a few. JS.geo is kind of the cutting edge stuff. Turf is only like a year old and allows people to draw an arbitrary shape and grab census centroids and calculate the population of that neighborhood. The trick is that it's all in the browser so if your data set is too large, then it'll be too slow. It's built by Mapbox and they're working on something that allows your data to be in the browser—it's called vector tiles. If you add Turf, you're going to have just drastically better visualization tools.

Megan: The evolution of the tech tools that's out there for NNIP tools has just changed. The pace of change is going to be so quick.

Mike - The magic word API has not appeared in this conversation. Your open data portal has just become the thing. We usually build an app for some topical tool.

Sean - My colleague at NYU is a big Python user for data analysis.

Mike - I'm using it for app building not so much analysis.

Sean - I use it for text analysis.