These transcriptions may contain errors, especially in spelling of names. These are unfortunate, and we regret that we do not have the resources to fix these errors. Still we believe these transcripts will be valuable to many users.

Elected Officials and Social Media

Libby Hemphill

>> It's my great pleasure to introduce Libby Hemphill. She's going to talk about more things to do with elected officials in the US, Europe, and Korea, right? And Libby has PhD and her Master's from Michigan. She's now at Illinois Institute of Technology. So, welcome to Libby and thank you for coming to speak.

>> Thanks. While you guys are getting your seats, I'll [inaudible] background about why would I present here and what are we going to do in the next hour. Actually, I'm going to do, hopefully, more listening than talking, 'cause I know you're itching to say things. I'm itching to hear them. So this might be a good opportunity for that. So, first, I just want to do--I'm going to a bookends my time with some stuff that I think you ought to know and then I'll revisit them again at the end, so quick overview. As Jenny mentioned, I started my academic work at the University of Michigan. This is a huge well-funded institution with a ton of graduate students and a lot of money. I am now at the Illinois Institute of Technology. It is a small institution with not a lot of money and I have one graduate student. And so the--I want to--you think about sort of where you are now and what you can get done now, maybe totally different from where you go next and what you can get done there just given the limitations that come with what resources are available to both human and financial. And so one of the things that I've done to try to get stuff done even though I'm at a smaller or I have fewer resources that I'm used to is that I use what I like to call "found data". And basically, it means that I scrape places and connect them with other existing data sets and see what interesting stuff happen. And today I'm going to talk mostly about Twitter. And technically, I'm not scraping Twitter 'cause I'm eating the streaming API output, so it's not scraping. But I have done a lot of scraping to try to get there 'cause that's pretty cheap. And let's see. So those are some of the things that are going to be sort of overarching themes during my talk is that you can do different things with different sets of resources and found data is one way to get work done with not very many. But it is as dangerous as it is easy to get. So we'll talk about some of the dangers of using found data. Oh, and at Michigan, I was in a high school which those of you in high schools may feel like that counts as a disciplinary home. I did when I was in one and now that I'm not, I realize that I don't have a disciplinary home and now I'm in a Department of Humanities. And so I'm a person without a disciplinary home who has moved to a place that has a really long tradition of disciplinary boundaries. The Humanities are kind of old. And so what I do and what my--the people reviewing my tenure case expect me to do is an ongoing negotiation and it means that sometimes we talk past one another and sometimes we have to sit down and we'll go, "Okay, here's what I mean when I say data, and here's what a publication looks like to me and that kind of thing." And that's been an interesting challenge and will continue to be. So that couple of years, but hopefully, we'll get it worked out. So if you have questions about, you know, what to do when you have no identity and you go to a place that does and doesn't like yours, let me know and we can talk about that. Okay. So, see if I have my things right. Okay, so yesterday afternoon at the very end, I mentioned that we're going to talk about some stuff today at 11:15 including whether or not Twitter has bias and whether Republicans are more connected. From you guys I'm interested--so, I have data about the members of the US congress, the Korean National Assembly and the European Parliament and what they're saying on Twitter and who is responding to them. So if you think about the Twitter API, it's the status says filter user stream. And I've been collecting for the US since June 2011 and for Korea and Europe since July 3rd of this year. So, given that set of data and your general interest in politics online, what are some of the other questions that you're interested about what elected officials are doing on Twitter? Yeah, Dave.

>> How do they spend their free time?

>> How do they spend their free time? Okay. [Inaudible Remark] Wait go back. Can I--at the same time? [Inaudible Remark] Huh, I hope that I could add them, but if I--also I'm a Mac user and this PowerPoint thing is very confusing for me. Can I edit my slide while I-- [Multiple Speakers]

>> I'll let you [inaudible], that's right, okay. Okay, so what do they do on free time? It just means that I'm going to have shorthand for all these things. That's great. I was supposed to teach today. This is--feels much more like teaching. Okay, what do they do in their free time, what bars do they go to that kind of thing? Jenny?

>> Actually, I want a follow up from David. And although this sounds like the kind of silly thing, I ask you to take it seriously. I, at one point worked in a computer science department, which was all men except for one other woman--woman and me and we had a little theory that a lot of decisions were made in the men's bathrooms, right after meetings. So I would like to follow up from David because I think that that's really pretty substantial question where people spend their free time and what impact that might have on decision making.

>> I'm actually going to leave it as two different ones 'cause I'm going to assume this is more about how and also be a where, which gives us some idea of with whom. And my--so here's a comment about being in a place that you didn't used to be. My chair as a historian, she is especially interested in this question. She studies how social groups form and create movements, especially among women. I'm thinking about the--some of the--how social movements got started. Often it was in bars around 5 o'clock. And in the early 20th century, who was in bars around 5 o'clock? It's much like the men's bathroom. Yeah?

>> What public reactions they choose to respond to?

>> Okay.

[Pause]

If you have more, you can shout. I can-- [Inaudible Remark] Okay. [Inaudible Remark] Okay.

>> What does it mean if they [inaudible] talking point versus taking up the [inaudible]? [Inaudible Remark]

>> What? [Inaudible Remark]

>> Lobbyist?

>> And which lobbyists, special interest groups do they spend to [inaudible]? [Inaudible Remark]

[Pause]

>> How much interaction they have with people of the different [inaudible]?

[Inaudible Remarks]

>> That's a good one.

>> Okay, hang on, I could catch up [inaudible]. All right, who they follow and then you have--you fell off at the end, [inaudible] talking about who they follow. [Inaudible Remark]

>> Okay.

>> How seriously do they take to being followed?

>> And then there's a lot of oohing and I couldn't hear the--do they write their own tweets? Okay, and now I'm ready. What's next?

[Pause]

There were at least two people who were getting ready to talk. Who else?

>> What's the agenda?

>> Agenda? Okay.

>> What kind of pictures of themselves they tweet? [Laughter]

>> So agenda, I have a question. Well, we'll come back, that's fine.. Agenda, yup.

[Inaudible Remarks]

[Pause]

>> Which words do they intentionally misspell?

>> Okay.

[Laughter & Inaudible Remark]

[Pause]

>> Right, I misspelled misspell. It has both, right?

>> I have no idea [inaudible].

>> For those of you who haven't taught, this is one of the most terrifying things throughout teaching is writing in front of people. It's worse than typing in front of people. What else? [Inaudible Remark]

>> Okay.

[Inaudible Remark]

[Pause]

>> Yup?

>> I'm interested in how much they actually consume this sort of media and not just produce [inaudible]? Also how skilled they are with it, so [inaudible] I think that's especially related to the [inaudible] questions.

[Pause]

>> Do they have multiple Twitter accounts, professional and personal?

>> I'm just going to add that one over here to save myself some writing. And you guys are all friendly, right? Somebody is going to write this down for me and e-mail them to me. [Inaudible Remark] No, no, I have to write them down.

>> How many people do they follow?

>> Okay.

[Pause]

>> Is there a bilingual tweeting or is everything in English or the great Spanish population [inaudible]?

>> How spontaneous versus [inaudible]? [Inaudible Remarks]

>> And who is reading?

>> There we have it. That one is over there somewhere. But it's--oh, do they write their own, but [inaudible] or do you mean who has adapted and who hasn't?

>> Yes, I think they all have Twitter accounts by now, but yeah. They say the press secretary--

>> Yeah.

>> --the officers.

>> .They're actually--not everybody does where almost everybody is registered, but some are private, which is actually how that works. I don't really understand but--or they don't get used. But--So I'm going to leave adaption, but in the blue--yeah.

>> This is sort of been touched on, but who are people responding to the most? Is it other politicians? Is it people from their parties versus the [inaudible] party?

>> And then right in front of you was almost--yeah. [Inaudible Remark] Okay.

[Pause]

[Inaudible Remark]

>> Opinions versus voting record, okay.

[Inaudible Discussion]

A little bit of board left. Yeah. [Inaudible Remark] Okay. I'm going to generalize that to [inaudible] tweeting to offline stuff, 'cause otherwise it's going to fill up my whole board left.

[Pause]

>> I know we're talking mostly of Twitter here, but do they use other social media and how does that used compared to what they're doing [inaudible]?

[Pause]

[Inaudible Remark] Okay.

[Pause]

>> Positive versus negative, like, positive versus critical?

>> Like sentiment or yehey, rah, boo.

>> Sentiment of their post.

>> Okay.

[Pause]

[Inaudible Remark]

[Pause]

>> How many--How frequently are they retweeted and how far does that go?

[Pause]

>> I got room on my press board for one more.

[Pause]

>> Do they like it or not?

>> Do they like Twitter?

>> Yes.

>> Okay.

[Inaudible Discussion]

>> Whoo, some of those are hard to read. So I'm going to quit there only because I'm thrown out of this side of the board and I thought, well, probably like on there. But one of the things that I wanted to show you doing this is that often--and this still happens to me as a third year faculty member that I feel like I need to answer all of my research questions right now, like today. And that can't be done, right? And so it took--

what did I that take, 10 minutes maybe to come up with all of these questions? It's going to take me more than this year to get those answered and that is okay. What the problem with working with found data or with social media is that between right now and running it back over to our [inaudible] my data is going to change and that can be terrifying. Until you say, "You know what, I can't set it all and I can't set it all right now", and that's going to be okay. I'm going to find a slice that makes sense. And the slice that makes sense may differ depending on which question you want to answer, right. So even if--these questions about followers, like, who are their followers, the demographics group of followers, who do they follow? That's going to change between now and the end of the talk. And it might be that I'm entrusted in how those followers change overtime? Or maybe I really only wanted to know who do they follow on Wednesday morning, who are they following during recess, which is different from, you know--or who are they following during Christmas? A bunch of my data is from December 22nd to March 15th, which is a sort of strange time for congress is there really [inaudible] that's not changed. But that time, especially because there's major holiday and a recess, but their stuff needs to get done, and similar in some ways to August, but different because the sort of attitude of a country is different around major holidays than around summer vacation. That I may see that if I wanted to look at sentiment, whether I look in December or August or May or November, which is especially interest in this year that that might change, too. And so I need to think carefully about what question do I want to ask and for what period, because I cannot ask it for all time for everybody for all of these. Just can't be done. And that's one of the challenges when you look at something like Twitter when you jump in to studying public officials is that everybody has something that they're curious about and you can't make them all happy all the time, including all of your own selves. You know like, what I care about today might be different from what I care about, I don't know, the second week of November when I really started to fear for my rights, right? Hopefully, that won't happen, but it's possible. And, yeah, so now, we're going to look at, okay, how do I go about answering some of these questions? And when I say look at it, we're going to leave PowerPoint and we're going to go into the data and we're going to look at it, and I'll sort of walk you through how I go about answering some of these questions. So, side note, while I'm going, I don't do any of these by myself. Like I said, I only have one grad student, but I do have two other assistant professor colleagues who are really good at stuff that's different from what I'm really good at, and the three of us together are able to get a lot done. So one of them is a political scientist, his name is Matt Shapiro and he helps us address questions like how do their talking points change, what sort of framing do they use, what's their agenda, how do I know what it is. And Jahna Otterbacher--so Matt is at IIT. Jahna is joint at IIT and the University of Cyprus. Matt is an Asian political scholar, so that's how Korea ends up in our group. And Jahna is in Europe and that's how the EU ends up and we have these sorts of overarching questions about representative democracies and young governments and that sort of how do those differ from one another or look the same as. But to some, it's convenient. We have a Korean speaker who cares about Korea so we grabbed Korean, too. So we have a person in Europe who cares about Europe and so we're able grabbed European tweets. And I live in the US and care about the US, so we grabbed those as well. Jahna has backgrounds--she and I actually went to a high school together and we came out totally different. So that's how I know that I don't have an interdisciplinary home 'cause I don't look like the other products of my high school, including some that are here today. But Jahna did a lot of work in natural language processing so these questions about what language do they use, how does it compare to other places where they use language whether it's in the mass media or press events or floor speeches, those kinds of things. Jahna leads those sorts of language questions. And so the 3 of us together are able to get some things done but we're spread out quite a bit between Cyprus and Chicago. And we all have competing masters. So Matt is in a social sciences department. And what looks like production or looks like work product for him is different than for me in humanities department and Jana has two institutions with two sets of masters and what sort of counts as where--I mean master, not degree master like overlord. [Laugher] But, you know, so we have to negotiate things like author order and priority of publications based on who needs this publication this months and who can wait until next month and do we send it to political communication because that works for everybody or do we shoot for the American Journal of Political Science because Matt really needs the

number 1 right now. And those kinds of conversations depend in part on what conversation is your data allowing you to have with other researchers and that helps you determine which is the right outlet for me. And in some cases it comes down to, okay, who needs this more right now. And that's a real problem for us because all three of us will be looking for [inaudible]. And so, some of those negotiations that don't always come up when you're talking about your research. But, okay, so let's--those are all those questions. So, here, my point with this slide is not for you to read it, but to see--these are the questions that we are actively trying to answer that I could write down in under 2 minutes when I gave myself time last night, right. So, you guys came up with all these questions, we came up with all those, like [inaudible] I'm going to be busy for a while and I appreciate your input and I welcome colleagues and coauthors on any of these projects who are interested in this kind of stuff. But--yeah, so those are some of the things that I wonder. And I'm going to leave my technology toolkit up for a second, but I'm going to do another--sort of walk back to say if it's easy for you guys to come up with lots of questions today and maybe what public officials do and maybe social media even aren't really what you're interested in. For me, as part of negotiating my belonging in my new department, I used to study big science and science collaborations and then I got to IIT and they we're like, yeah, we already know how to do that. You should probably study something else, like, "Hmm, okay, curious that you hired me after that job talk, but I'll do what I can." And I have an interest in social media. In fact, when I was doing a post doc [inaudible], what, 3-1/2 years ago now, Mark and I had a call. We said, "Hey, we should study congress. We should look into that." I was like, "Yup, first I got to get a job", [inaudible]. So when I got a job and now I can look at congress, but the real reason that I started this study was that last summer, I was following a bunch of--I was using TweetDeck, so I'll show, you know, for those of you who don't know TweetDeck. It looks like this, lots of columns, you can change what shows up, et cetera. And [inaudible] Desk I have these two monitors and I'm watching my columns come in and one of my columns was congress and one of my columns was a bunch of celebrities that I like to follow 'cause I am obsessed with television. And there was one Friday afternoon that I was like, "Wait a minute, did my columns move? I don't think I moved those." And it was the day that New York passed marriage [inaudible]. Something good? Oh, good job Noah [phonetic]. [laughter] But--So what I noticed was I was confused about which column was which because one of them was all about politics and one of them was all about television and that was really normal, but they were reversed. So on the day that New York passed marriage equality, which for me is a very--like, kind of matters, celebrities were talking about how great it was that New York was allowing people the same rights as others and politicians were talking about what Sunday Morning new show you could see them on. And I was like, wait a minute, I had this backwards and I want to know why and I want to know if this happens all the time or if it's just Friday afternoons in June when my rights run on. And so this morning, sorry guys, down it goes, I took a screen shot--well, let's see, of--yeah okay, so here is the House of Representatives' list right when break started, right. So now it's Wednesday morning and they're--as far as I know, theirs is no major legislation happening in any of the states today that I need to worry about, but you can see sometimes they talk about issues down here with [inaudible] and sometimes and they talk about what's happening in my family, [inaudible] school. And then there's a bunch of when you can see me on TV. And at the same time, I'd go--we have Aimee Broneman [phonetic] was very active last night and this morning, I guess 'cause normally there's a bunch of private practice [inaudible] that have a lot of things to say. But Aimee Broneman's tweets are about the Ryan Romney ticket and how afraid she is, right. So, it's not just Friday's in June. There is something happening here between sort of what is political and what is celebrity. And once I get done with all of these, that's where I'm going to go, right. So, like, what, 4 years maybe. And that's not--honestly it will probably take that long 'cause it takes a long time to answer a lot of question. So, you're thinking about where do I find the research project, right? Or where do I find a research question? It might be that you're just looking at something online and going "Hey, that's curious. Why is it like that? I wonder if anybody can tell me why it's like that." And so, I went to find out, can anybody tell me why politicians are telling me about television and celebrities are talking about politics. And, no, but I did find some early work like [inaudible] talks about, okay, what is congress doing on Twitter? I was like, "Huh, okay." So, that's a good place to get started. And now, there's room for me. There's this place that I'm curios about that I bet other people are and looks like they are. But

I can--that's a space I can work in. It turns out it's also a space that my institution really likes and that they're happy for me to investigate. And so that makes my life and work easier, too. Yeah. But it means that I have to follow Twitter all day everyday which might sounds fun but it's really not, like it's just really not. And I have to learn a lot of things about our government and our officials that I really was happy not knowing, honestly. But, yeah, it sometimes you'll find things that are disturbing or that John McCain plays games on his iPhone and then challenges you to beat him, the games on his iPhone. I'd be like this person ran for President and that's--what he want--you know, and--I don't know. It's just--I was more surprised by that from him than I would have been from like a junior representative or from one of my graduate students or even my family members or something that--I don't know, I expected more serious persona but maybe that's a part of the--I want to have a beer with him way that reelect--I don't know. All right. Anyway, so I can close those, okay. So, what does it take to get to the point where I can start answering some of these questions? So, Mark talked last night about Note Excel and I definitely do use Note Excel but you'll see that I use a lot of other stuff, too. And the stuff on the left is about getting and sorting data. You'll see that Note Excel for me is an analysis tool. It's not a getting tool in part because of what Twitter limits are and in part because I have enough programming background that I want to wrote my own but not enough that I really can. And so I sort of roll what I can and then I buy it from somebody else. So I grab the data myself and then I get help analyzing it. And the order of magnitude that I'm dealing with Excel can't handle. Turns out neither can UCINET. I was getting a bunch of stock overflow errors this morning which I don't know how often you guys get those. But I feel really proud of myself. But I'm like, yeah, my data is too big for RAM. And then I remember that the only reason I have this PC is because I needed Note Excel and UCINET and so it's not really, like, pimped out, right. It was--I can spend up to 500 dollars without ask--without going through the institution. I can get reimbursed for that, so this laptop cost 499 dollars and 99 cents. And it was at the time the most that I could get for that much money. So, it's going to run in to stock overflow issues. But, anyway, so the way that I use to get data is to grab directly from--see if I go--here we go. Okay, so I'm going to back and forth among these three sides--slides, excuse me. So, when I put the mine in quotes, 'cause none of these data is mine; it's yours and every other Twitter users. I just borrowed it from Twitter. And then I never give it to anyone ever 'cause then Twitter would sue me and my life would really go down the toilet. But I want to give it to you. So if you say we're working on a paper together, then I can't, so. Anyway, so when I say mine, I have data about US congress and then these little bullet points, those are kind of mine because they're--I've done something to the raw data to make it different somehow, whether it's that I pulled out just mentions or I've done some sort of processing. But, as Brian knows, I've been thinking a lot about what it means to share my data and I'm thinking about, you know, does sharing my data mean giving you the raw JSON that I grabbed from Twitter or does it mean giving you the Note Excel workbook that has the mentioning each other data or does it mean that I just tell you, hey, I did some of that, you should look into it? And that--you know, data has a life cycle. And sort of how much processing I do on that data impacts what kinds of questions you'll be able to answer from it, all right, especially if you're entrusted in what they misspelled, right, if I've corrected their spelling, then you can answer that question. And that might have been a thing that I did in order to make them more readable. Now that the Twitter API rules have changed again this week, I mean, I don't do much processing to tweets except to, you know, figure out if they're spam or not. I don't actually change any of the language, but now I really won't 'cause then Twitter will get really upset with me. And then, you know, do I share the code that I wrote in order to do that processing? And if I do, do I give it to you in Ruby or Python or PHP or do I ask you which one you want it in? And the reason that I have to write in Ruby, Python, and PHP, well, and Pearl even, 'cause Jahna loves Pearl. And if I want to do something in languagy that I need to work with Jahna and so I got to either write it in Pearl or accept written code from her in Pearl 'cause that's where she's comfortable. But my graduate students, they're taking Ruby while they're on campus, so they want to do stuff in Ruby. Everybody, like they all want to work in Ruby, so I tried to meet them there. But I have a background in ASP and PHP and where does that fit? So, it depends on sort of who needs it, what's the fastest way to get it done and the cheapest? And, yeah, we take sort of software development approach to, like a sort of fast, agile approach saying, "You know what, good enough is fine. Just ship it." So, as soon as we get code that

gets the data that we need in the format that we need then we're done. But it means that we don't go back and do a lot of good documenting or storage or anything like that. But we do then put it up on Get Help for anybody who's brave enough to use our good enough code with the caveat that it is good enough. It's not good, but it is good enough. It will translate the data. But the other reason is that a lot of these data sets use two mode data, so people to hashtags or people to links, rather than one mode data people to each other. And you guys wondered about followers, followers would be another one. But congress has so many followers that it breaks every computer I have except my giant Amazon Instance. And so that we only update every once in a while, 'cause followers there's just too many of them that I can't take the follower data with me and show it to you, but that's another one mode network that--Note Excel is really good with one mode networks of a certain size. Most of my networks are two mode and much bigger than that. But I did write--well, I paid somebody else to write--oops, where did it go--to write a macro that would convert a two mode network into a one mode network. It's very popular in Korea right now. Get a lot of questions from Korea. So if you have two mode network, two mode data of a small size, there is a macro that will help you get one mode data so that Note Excel can work with it. And you can visualize two mode data in Note Excel, but you can't really run any analysis 'cause they all just be wrong. But if you're just looking to draw a graph which sometimes you are and it is the best tool for that by far 'cause it takes about 30 seconds not 2-1/2 hours. Yeah. So--And then I also put this sort of getting and storing. Like I said, I work with two other professors and one other graduate student, and we have to share data. So, you might think, "Oh, I work by myself. I don't have to worry about data sharing." But you do have to worry about data sharing 'cause you might have to share it with somebody [inaudible] with. You definitely have to share it with yourself next week. And keeping track of what you did and how you got there--see, I'll show you a--you have to think about where am I going to put it, what is it going to say, what's the process that I went through to get that data. So last night, when I was working on making this link graph for you guys which will open up in a second--ooh, so that's the SQL statement that I used to pull the data from my SQL database. Where is the--is there a keyboard shortcut for maximize?

>> Double click the title bar

>> Which one is the title bar? This guy? [Inaudible Remark] Little higher.

>> Double click.

>> Okay, thank you. [Inaudible Remark] Well, we got it done. Okay, so there is the SQL statement that I had to make to pull data back out. So until July--yeah, right? [Inaudible Remark] There're 387 or something, and then I got to export my data to Excel and I remind myself double check it, double check it, double check it, in part because Twitter allows you to put line breaks in your tweets. If I could change one thing about Twitter, it would be the terms of service. If I could change two things, it would be to stop letting users mess up my data. And when you guys put up line breaks in your tweets, you make life difficult because Excel on the Mac and Excel on the PC and Pearl, Python, Ruby, all these guys totally disagree on what it means, so start a new line. You really wish they'd agree on what that means right now, but no, slash R, slash N, space, all it took--it's just irritating, so just don't put them in there. You know, it's 140 characters. You really want them to have spaces? But, anyway, so double check it. And then I resolves the shorten URL. So I've been storing the short URLs. And in order to do that I had to go get somebody else's code, 'cause I don't want to write my own resolver. Once you get somebody else's code, but then I had to alter it to let it take an array of all of those links, so 25,000 roughly links from December to March that members of congress shared they wanted to know. Chances are they're pointing at the same press release but with different beat links. At the time it was beat URL, so I had to resolve them and then add those [inaudible] URLs to Excel and then I had to copy those columns into a text file so that UCI Net could read it so that I could UCINET to export a matrix that Note Excel could import, right. This is why it takes a long time to answer each of these questions. And you've got another question that I don't have data that I can use, right. I'm still munging it. I'm still making it

work in a format for me. And then import it and then last Excel and [inaudible] look up. There's a life changing function. It lets me keep one canonical data set that is congress and all of the attributes that I have about them. So their sex, their office, their party, their district, how long they've been there, when they started Twitter, what their Twitter followers were on X day. All that stuff is sort in one place and then I just view look up to stick it in each of my Note Excel workbooks. And then I double check it again. Somewhere in here I'd probably be double check it, too, but UCINET is a mystery to me. It makes things happen and it spits out plain text then it works. And then because if I resort that canonical data set ever, it will break my view lookups. You do a select all, copy paste values. This is clutch or your data will tell you that [inaudible] is from Florida. What--Wait, what? And it just even mismatch. So once you do that, you end up with--would you guys rather do hashtags or links? If we're looking at what are people talking about? Are they talking about the same thing?

>> Hashtags.

>> Hashtags. [Inaudible Remark] Okay. So this same process only for hashtags. And this is when I use this store stuff in a normalized MySQL database. I don't do that anymore because the retrieval time. So normally, databases are optimized for input not for retrieval and most of my work requires getting it back out and not putting it in. And so it's managed to break every server that I have access to. And so now, I just grab the raw JSONs, stick it in [inaudible] and worry about it later. Like I said earlier, your data is going to change everyday anyway, so I have to solve that problem today. I just need to store it. So, I'm storing it and we'll purse it later in whatever we agree upon, Pearl, PHP, Ruby, something, maybe Java. There's a guy just applied for a job who uses Java. I don't really know what that will do for us, but--okay, so here we have the one word projection of hashtag affiliation. So, if they have--do those words make sense? Mostly not? Okay. What do you think we're going to see what works on this.

>> We're going to see party affiliation pointing--clustered around shared hashtags.

>> Okay, so we're going to see blue group and a red group, but not really--kind of. So I have colored the vertices using IF statements instead of [inaudible] because I want them to be red and blue. And so the red, blue, and yellow and then there also a square or a disc based on their sex, male or female, which--in our current congress I'm pretty certain that the male/female, men/women matching is 100 percent but that doesn't mean I refer to them as man/woman if the data I have is male/female. So we started talking about gender earlier. Male/female are sex terms, men/women are a gender terms. Let's try to use the term that we mean when we say the category. So in here it's sex terms. And then I set up the--I did use auto fill to get tool tips, right. So I've concanated representative name, party affiliation, and state in order to figure that out, so, once we get there. How would you like me to play with it to see if we ever get those groups that we expected to see? Yeah, Mark.

>> Move to [inaudible] and layout.

>> Okay.

>> And lay it out again.

>> Do you want to adjust how many times it runs that?

>> No, that's only [inaudible].

>> Oh, okay.

>> You can do it that way.

>> This was 25 times by the way. [Inaudible Remark] I thought maybe so. I also--I think I'm two versions behind. I think I'm running 2-19.

>> Oh, well, you know.

>> But, yeah.

>> 2-21 has a cup holder. [laughter] Okay, so the next thing you might do is go into Note Excel right now and maybe we should group. Now let's click Note Excel.

>> So we have groups.

>> Okay.

>> I just took off the properties, but we can put them back in.

>> Okay, 'cause then you might go into the layout menu over in the graph and--

>> No, first, strong opinion. [Laughter]

>> Yeah, but--

>> Strong opinion here.

>> Yes.

>> And [inaudible] we should all be thankful for those. But in this case what we can see here is this really spherical structure that on lower part is blue and on the top part is red and that's a core. Around that is something [inaudible]. Now, if you group [inaudible].

>> Yeah, I was--

>> If you group [inaudible], you're going to see these Republicans, these are Democrats, and you're going to assume spatially because they'll be spatially arranged in separate places that they're very separate 'cause they're talking about very different things. But you could see that there's actually more than [inaudible] there's a spherical core of Twitters, some of them are blue, some of them are red, and around that is another sort of periphery of Twitters, some of them are blue and some of them are red. Within that core there is clear arrangement of blue and red, but to artificially impose this segmentation there is perhaps to artificially apply a certain separation that's not necessary in the data. So there are some red tweets down below and scattered upon the blue. You're going to rip them out of there and say, no, you're going to Republican box. There are some blue that are tweeting in the red. We're going to rip them [inaudible].

>> Not necessarily though if we don't group by attribute. We don't say group Republicans versus Democrats. Let's say let's group by connectivity then we're actually going to see essentially the boundary crossers.

>> Right, so it depends on--

>> Did you run the cluster?

>> Connectivity as well.

>> Yes, I ran the clusters first.

>> All right, so group by cluster, please.

>> And then--

>> Students notice even the [inaudible] quarrel-- [Laughter]

>> Yeah I'm really-- [Multiple Speakers] How do you want me to do this?

>> So what I would have you do is go--you actually have to go all the way up to the top to the Note Excel ribbon.

>> Yeah.

>> Okay.

>> All the way up, up, up, up, up. Click. Go to groups.

>> Groups.

>> And what we want is to group by cluster.

>> There you go.

>> And then you can pick. Just click.

>> Click it.

>> And we're going to take--we'll start, let's say, [inaudible].

>> Oh, it shouldn't--that's how it came up with these guys.

>> Okay. In--

>> You did?

>> Yeah-- [Multiple Speakers]

>> Groups Republicans, Democrats [inaudible]?

>> Nah, those are not groups. Those are node the attributes color.

>> Oh, well, okay, then we're all right.

>> Yeah. And then I just--

>> Okay, so now go to the groups.

>> Yeah, it's going to run that again.

>> Do the layout.

>> So we'll see if it comes up with the same groups.

>> Click and see if you have focus.

>> No.

>> No? I don't know what it's doing.

>> It's running that algorithm with that.

>> Okay.

>> But while we wait for that though, timeout for learning balls. So, the other thing that I wanted you guys to get, right, is that answering these questions is really, really messy even when you know what you're doing. And that this is--you know, it can seem like, "Oh, they're telling Libby how to do data analysis and she must feel weird about that." And, no, I'm like, awesome. Tell me what you are curious about and then when I get home I'll run it again and I'll see if we get different results based on how I do these visual properties, right. The assumptions that I make about what my data tells you differ based on what I see, absolutely. And the assumptions that you make based on the data that I let you see will also differ. And so it's important for me to look at the data more than once with more than one group of people and more than one set of settings. So I can start to see, okay, where am I just seeing what I thought I was going to see? Where am I seeing something that I never thought about? And that playing with your data is a really important part of doing that analysis process. And I think it's not done yet. Now, it's really unhappy with me, Mark.

>> It should be showing your progress. I think it may have lost focus or--

>> All right, let's just kill it. There we go. All right.

>> Yeah, yeah.

>> Okay, so now, I have groups and they're placed in them. I just removed the visual properties from the groups.

>> Okay, so now, go into--

>> 'Cause groups always overrides my own settings and they're--

>> There's a setting for that.

>> Yeah. I didn't know where it was. This is like 2 o'clock so.

>> [Inaudible] let's touch on that briefly and then let's go up to the group's menu and then go to the bottom of that. You'll see that it has group options. And one of the groups options is whose display attributes trump the others, the vertex attributes or the group attributes. So if you get to choose--

>> I see.

>> Groups are shades with one--shades or colors, which should be.

>> Okay.

>> So you would get what you want. So you cancel that and then go back into where it says [inaudible] scale. Down, down, down. That's it. And drop, drop, and then go to layout options. And now go and I think it's off screen.

>> Yeah.

>> There we go. And then click the second [inaudible].

>> Layout each grassroots in its own box.

>> And then click that checkbox down there.

>> Ben, do you want me to hide things for now? [Inaudible Remark]

>> Use the grid layout, [inaudible] so it's okay. It doesn't matter. It doesn't matter.

>> You want them showing, okay.

>> [Inaudible] And then you have to--probably after the layout, I'll lay it out again.

>> Layout again.

>> Refresh. [Inaudible Remarks]

>> Layout again only refreshes--so there's two layers in Note Excel. There's what's in the spreadsheet and what's being displayed and you can filter it. You can do all sorts of things with display layer but not the data layer. Refresh graph means go back to the spreadsheet, bring the data back in and rerun with the graph. Layout again does says use the data as it is and just lay it out again.

>> Okay, so refresh didn't work. So I'll try to lay out again, right, 'cause unless they're all one box--

>> Unless they are all one box, yes, it could be.

>> But it tells me that I have three groups. So it should group them, right, okay.

>> We did rerun clustering so it may not yet really clustered.

>> All right, well, we'll try it again. We have questions? I will continue to multitask. But, so first [inaudible] what we did. So first I wanted you to see that there are lots of questions that you can ask just from one set of data and it's okay for it to take a really long time to do those. That how we go about answering these

questions depends on the tools that we're using and who's in the room with us while we do them. We have to make negotiations around our sort of institutional identities and the stuff they were curious about. Research questions can come from Friday afternoons watching Twitter. These are some of your takeaways. So, questions. And I'll keep--Just fire them out 'cause I'm going to have to look back and forth.

>> [Inaudible] how you are navigating the sample [inaudible]. So can we just talk about why it differs [inaudible]?

>> Sure. So, I use--well, up until July 3rd of this year I used the PHP firehose library, but I gave it a set of IDs based on all the people that we could confirm as members of Twitter in March. So, we do members of Twitter updates once a month, but since we are moving to the Mongo Python collector, we quit updating the PHP one in March but we used the firehose library. Is that the question that you mean or?

>> Well, I'm just wondering are you using the actual 100 percent firehose to get that, all the tweets are using one of the--

>> No, you pass fire--well, you pass the API filter. That's the Twitter term is filter, and then you give it a set of user IDs. So I send filter statuses, user IDs where all the user IDs are the members of congress I can confirm or the members of the national assembly I can confirm or EU parliament I can confirm.

>> Oh, so you're going [inaudible]?

>> Yeah, streaming API. So we gather the IDs ourselves and pass them to the streaming API. If I went by--so I also do this for the Chicago City Council and everybody on the Chicago City Council is a democrat so it makes for an interesting sort of national experiment about what happens when there aren't party politics. But I don't [inaudible] profile because then I'd end up with stuff like Chicago's mayor, who is not Rahm Emanuel even though he claims he is and now he has a book deal, but he's not the mayor. So, it doesn't matter. He registered the Twitter ID before Rahm did and Rahm offered him tens of thousands of dollars. Do you want to stay with class [inaudible]?

>> Yeah, please.

>> Okay. But, so we have to be careful about profiles, so we look him up by hand. Undergrads love that job. [Laughter]

>> [Inaudible] my question was, you're kind of making a big deal about [inaudible] alternative characters that show up in tweets particular return. And one of the projects that I'm working on more specifically reflecting English and Arabic tweets.

>> Okay.

>> And what's very interesting is in alternative scripts, particular Arabic, because that's written in a different direction, the use of [inaudible] return is really important to make the tweets readable in the web. And so the [inaudible] return character is actually kind of important because you can see then people who are making sort of stylistic choices. So it's not just that, it's about the content. When you think about what's in a tweet, there's very much a style issue about the tweet and how a tweet is going to be presented to a user and the issue here is in alternative scripts you see that. So if you want to work in foreign language--

>> Yeah, we're getting this in Korean, too. It's a very good point.

>> Yeah, but Korean is still [inaudible]

>> But the [inaudible] returns matter.

>> Yeah.

>> So, yes. And now that--I mean, so but the stylistic choices--and Twitter and I apparently disagree on this as of this week that the device that you use to read--