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.

Data by Design
August 21, 2012

BEN SHNEIDERMAN, COMPUTER SCIENCE PROFESSOR, UNIVERSITY OF MARYLAND: We are proud to present Elizabeth Churchill under #tu2, please. I just love to read her work and spend time talking and listening to her. She's a fresh, original thinker who really understands in a deep way that educates me in this topic.

Thank you for being here.

ELIZABETH F. CHURCHILL, APPLIED SOCIAL SCIENTIST, INTERACTIVE TECHNOLOGY DESIGNER, RESEARCH DIRECTOR/MANAGER: Thank you so much. Thanks for having me.

I'm going to talk about a few things I've been thinking about recently. You can read the papers if you want more details but I'm going to see if I can cover a few anecdotes, thoughts, and I have a particular request. This follows on from what Ben has been saying. He's covered a lot of beliefs I'm going to try and reiterate through my own lens. The real takeaway is that as well as researching your own topics, I want to have us all think about a bigger topic, the notion of data by design and our role in designing the data that are collected as well as how they are analyzed and individualized.

[New Slide: Design/Science of Participation]

When I was invited to come along here and think about some of these concepts and technology mediated social participation, which I was involved in some of the workshops for, there are two angles to this which I'm passionate about and I'm going to talk about the second one mostly today.

The first one is this platform of mediated communication. My history, my research in the past was about communication online and thinking about the psychology of communication online, the motivations and emotions. I'm a psychologist, I care about motivations and emotions, not just actions.

I've done a lot of work on virtual environments, chat spaces and workflow systems. You can read those papers. I'm going to try and focus on the second thing. We, as a group, need to address the design of data for science very explicitly and directly. Think about where that science is going.

[New Slide: On (1) - TMSP via SMPs]

On the first point, we know there are plenty of media platforms that allow us to be aware of other's activities. There's conversation, content exchange, which are pretty good. Content storage, indexing, and search tend to be quite poor on some of these systems. I'm thinking about Facebook and Flickr. Content sharing needs to be malleable as well as stable. How do I go in and actually work on something with you, really collaborate on the co-production on something?

Coordination should be long-term as well as short-term and collaborative production should be lightweight as well as complex. Longevity for a lot of these systems is currently questionable. This picks

up on something that Ben said and I'm going to reiterate through the talk.

[New Slide: Various Images]

Here are some things we can think about in there first bucket of mediated collaboration. There's cooperative activities which are centralized, like GitHub. There's collective action which is centralized. Plenty of groups on Facebook. And then there's collective action, which is decentralized. These are some different ways that we are using current social media platforms that have been designed for recreational purposes, especially Facebook, for collective action.

These are really important for us to study and think about. There are plenty of papers coming out on those topics. Many of the people in this room said they were interested in exactly this kind of thing. I want to say that's really important. I care about it and want to talk to you about that issue and how you design centralized and how you orchestrate decentralized kinds of activities.

[New Slide: On (2) - Sciences of the Social]

The second thing is what I really want to talk about today with some examples from my own work. One is about data quality. What are all those data that are being collected? Descriptive data, predictive models over data. Observed data versus understood data. I would argue a lot of data I collected are observed, not necessarily really understood. Local versus universal. I'll keep repeating this. When you collect those data and make assertions, does that really apply to the whole of human kind or is it local to the specific context in which those data were gathered?

Again, remember I'm interested in the social sciences about people and the invariance of people's activities. I care about humans and their feelings and cognition in all contexts, not just particular contexts.

Are we reactive or proactive in what data exist? I'm currently asked -- suggest, especially from those media platforms, we're pretty reactive and not proactive. We, as scientists, are not necessarily being proactive.

Are our observations standalone or replicated? Ben talked about interventions. Are they single and standalone or comparative, replicated? Are we asking those questions in the right way? Do we understand what the science of this space really is?

Science quality is what I'm thinking about. Data stability and longevity. Are we gathering data that we can actually point to and say, "That is a really interesting invariant," "That is a really interesting observation about human kind or people in general."

We need to understand the terms of service for the places from which we collect those data. We need to understand what those content are and who has social responsibility around those data. Is it us as a group or the organizations that post those data?

I'm also going to argue -- I'm going to reinforce something that's been said already. We really should be getting designers, statisticians, computer scientist, data scientists and social scientists. This multi-disciplinarity is really important because if we're starting to actually make assertions about the social and data sciences, we should have folks talking together.

I'm asking everyone to do, whatever bucket you put yourself in in any one of these categories is think of yourself as an epistemologist or an ontologist. I want you to think about where your epistemology comes from. What are the roots and core beliefs of that and of data online right now? What are you finding? What are you not finding? To what extent are you really getting the data you want to answer the questions you have, truthfully, honestly, the core things to care about? And to what extent that you're using the data that just happen to be available?

I'm going to keep saying that over and over again.

[New Slide: Focus on (2)]

And I'm happy for someone to tell me I'm completely wrong and it's all been solved.

(LAUGHTER)

I read this years ago, 2010 or something. It's when started to think about data science. This is a quite nice piece. O'Reilly, of course, has been at the front writing in an accessible way, interesting things that are going on in the technology sphere. I would recommend looking at that. I went back through it the other day and was interested in what differentiates data science from statistics is that it's a holistic approach. I really like this, "gathering data, massaging it into a tractable form."

How many of us have done that?

(LAUGHTER)

"Making it tell its story, and presenting that story to others," as if the story somehow resides in the data that we have massaged.

I think that's a really interesting assertion.

"The first step of any data analysis project is 'data conditioning' or 'getting data into a state where it's usable'."

I'm an experimental psychologist. Conditioning it means something very specific to me. It means pellets and rats. I was wondering in what way I was an agent in conditioning the data, just like I was an agent in conditioning my rats years ago. I want you to think about that.

[New Slide: On Data Science]

The problem with big data is when the size of the data itself becomes part of the problem. We all know that, right? It's going to get bigger.

This, I found interesting. You may not know what's important until after you've analyzed the data. Then, do you go back and do it all again? Do you have access still? Just a question.

Data scientists come up with new ways to view the problem. Here's a lot of data, what can you make from it? Some really interesting notions of agency and mergence and the role of the data science in the

production of what questions can be answered as well as what questions will be answered.

The future belongs to companies who figure out how to collect and use data successfully. What about us as scientists? Does the future belong to us in any sense? It should.

[New Slide: Business Logic is not Science Logic]

I told Ben this morning, I'm going to be an activist.

The bottom line is business logic is not science logic, OK. There are reasons for that.

[New Slide: Social Media and the Big Data Explosion]

Here's an interesting one, a nice article. "Social Media and the Big Data Explosion." This is really aimed at marketers, CEOs, et cetera. Without revenue, chatter is just chatter.

I'm sorry, people. My chatter happens to be my human relationships and my deep connections with other people. My chatter is not just chatter. It is the way in which my friendship is solidified with you. There are things I share online with you -- that YouTube video is a proxy for my social relationship to you. The thing, that video, all of the metadata around it, my comments and the fact that I shared it with you? That's about my relationship, it's a place holder for my relationship with you. It's not just chatter. I'm not saying one is right and one is wrong. I'm saying these are very different approaches to thinking about it.

As a psychologist who cares about people, I want that chatter to matter. That was an accident.

(LAUGHTER)

I'm very proud of it but it was an accident.

UNIDENTIFIED MALE: Chatters matters.

CHURCHILL: Yes, chatter does matter.

[New Slide: Data - the "This is the dataset" Problem]

There's the data set problem. Where you're looking might be what's available to you and it might be your epistemological particular viewpoint, OK?

[New Slide: Image of Colander with Text]

The other piece of this is I care about design: interface design, interaction design. That interface that you're looking at right now is sort of like a sieve. All of the things you're thinking. Think about all about the top, the information, think about that as emotion and communication needs and requirements and desires, hopes, and fears. All goes through a button, potentially, right?

[New Slide: Interface Elements]

Here's an example I like to give. A few of you have seen this before. So, what do I know? On this side we have the western toilet. On this side we have a Japanese toilet. We have different interfaces that give you choices about what you do in that most private of rooms.

I'm a data scientist and I want to know what you do in that most private of rooms.

(LAUGHTER)

I have a very different picture of you from this one versus this one. Those are interfaces that lead to data that lead to conclusions and questions and understandings. I want to put these images in your head so next time you go online you think about how that rich limbic system and frontal lobe and motor cortex when it comes to the connect and your sensors and devices get streamed into particular things that are recorded or are not recorded.

[New Slide: Image of Elizabeth's Facebook Page]

Here's my Facebook page. What about me would you find out from this? One thing is this very controversial button: "Like." I press "Like" when I like it, when I hate it, when I think it's funny, when I just want to reach out and just touch you to let you know I read it. What's "Like" mean? I'm not the first person to say any of this stuff, I just want to reinforce it in this room so we can talk about it more critically as we go forward this week.

[New Slide: Dating]

I've done a lot of work on dating. Anyone in this room dating? Ever been on a dating site? Yes. I always ask and I get very few numbers and I never believe them.

(LAUGHTER)

[Images of Two Different Dating Websites]

I was saying the data get collected, remember the light? If you are on eHarmony, there are all kinds of questions that get asked of you about your commitment to marriage. That's because there's a particular belief system behind that tool, which is called a dating site, just like the other dating site over here, which has a different set of values.

On one of these, the thing you will upload that people will look at the most is the picture of your face. On another one of these, the thing that will get looked at the most before you get selected is not a picture of your face.

(LAUGHTER)

These are all aspects of human emotionality and sexuality but they are partitioned in interesting ways according to a particular business logic. As a social scientist, if I have access to these data sets, what am I finding out and what am I missing? And how am I putting the parameters around my conclusions? How am I phrasing them?

[New Slide: Profile Creation]

Anyone ever filled one of these out? It's a dating site but I've bet you've done profiles. The categories of what matters for the matching algorithms are pre-specified by people who may or not be designers, psychologists and most certainly aren't social scientists.

Does not mean we can't learning something interesting. What it means is we have to be careful about our conclusions and the quality of data for purpose for us.

[New Slide: Image]

Huge numbers of excellent comments on what you should do for your profile. Even if you're not doing a checkbox, there are norms of how you represent yourself if you actually want to be successful online and get that date.

I could, in fact, and I do think that walking along the beach at sunset holding hands is profoundly dull. But, if you put it up on one of these sites, you're probably more successful because that's how we code in our culture what constitutes a romantic person who has meaningful intentions toward another.

I think that's important to know because actually over here, we have the top 10 Indian dating sites. I don't personally know what looks like a really good profile there. Elizabeth Goodman, who's a history student at UC Berkeley and I did some comparative studies of different countries and dating sites and there are different emphases on what constitutes a marriageable, datable or engageable person. I think that's interesting.

[New Slide: Image of Links to Research Articles]

Great research out there. Many of these papers are really excellent and make great observations about things like homophily, how like you are someone else and how likely you are pick someone who looks like yourself. Lying. Lying is actually when you keep changing that profile. Elizabeth Goodman and I found that people often change the profile because they just like to have more results to choose from, not because they're necessarily lying to the other person. They are in a deep conversation with the algorithms, not necessarily the people they're hoping to date. They are in a deep conversation with a set of algorithms that will represent them through algorithmic match to other people.

We only found that out because we weren't doing the data science, we were doing a different kind of data science, qualitative data analysis. We went out and sat with people. We talked with two people and watched them build profiles and watched them horrified and mortified by lack of results and too many results and the wrong results. We actually went and watched people get ready to go out and we interviewed them when they came back from dates, assuming the date was not really successful.

(LAUGHTER)

[New Slide: Anxiety, Self-Reflection, Identity ...]

We found incredible amounts anxiety, issues of self-identity, self-reflection. We found that people were going out and buying clothes and changing the way they look and doing all kinds of things.

Most of those papers don't talk about that. They allude to it but don't talk about the psychology of that

incredible anxiety. There are, of course, exceptions. There's a lovely book here -- again to Ben's point of interdisciplinary. This is not an experimental psychology or a data science book, it's a book about the marketplace of dating, about the ways in which we learn to construct ourselves, put ourselves on the dating marketplace. Some really interesting insights in that book about the phenomenology of being on a dating site.

[New Slide: Flickr]

That was just one example. I'm trying to get you to think differently about what the data are that are collected.

Flickr. Who has a Flickr around? Flickr is why I went to Yahoo. I love Flickr.

[New Slide: Elizabeth's Flickr site]

That's me on Flickr.

[New Slide: Elizabeth's Flickr Stats/Data]

Those are my data, my stats. I'm not particularly viewed. I have these little spikes. Talk about anxiety, this makes me profoundly sad about myself. Oh, nobody likes me.

You're about to see a map and I want you to think about the privacy settings on Flick. This was work done with Tony Lamb, done several years ago.

[New Slide: Map]

The privacy settings on Flick allow you to make things private, only to your friends, only certain network, or completely public. One of the things we did was had a look at what percentage of people's pictures were public versus private and we put it on a map. Red means most of them are private, green means most of them are public.

A bunch of you have seen this before but I want to say that we were only playing around with the data asking questions that came out of interviews around data privacy because the predominate assumption was that people don't really change their privacy settings that much or it's pretty general, because nobody had mapped or asked the question about location. This is self-reported location.

By actually having a social science sensibility around cultural difference allowed us to go in and visualize this. It didn't answer any questions but what it did was lead to a program of research using discourse analysis and interviews to find more about how people felt about private versus public.

The reason we did interviews and discourse and content analysis -- the folks in Scandinavia, the discourse around what it means to be private wasn't just, I'm afraid you're going to see my stuff or rip me off or come down to my house. It was also around politeness. People don't need to have to see all of my stuff.

Whereas, over here, when we talked to people in San Francisco and other areas, they said, if you find my stuff that's your problem. You get to sift through my stuff.

There was a social responsibility about what is appropriate to communicate to the public and to your social group or not. I thought that was an interesting result. There's a lot more on that that I will share if anyone wants to ask.

[New Slide: Flickr: The Commons]

How many people know about the Flickr Commons? Great.

I'm going from the personal, private and emotional to the cultural and bigger picture of --

(OFF-MIKE QUESTION)

CHURCHILL: It's an old data set from about 2006.

(OFF-MIKE QUESTION)

CHURCHILL: It was a sparse data problem. In 2006, there were not that many people from China not India who were actually online on Flickr and so we have a sample taken from an old data set. The specific details of the numbers is not what I want you take away. I want you to take away that only by surfacing and applying that method of interrogating the data in a particular way, did we actually start to see something interesting that led to a set of very qualitative research questions.

It was a good question, though, so thank you for letting me clarify that.

The Flickr Commons. Photos on Flickr, I'm going to get to the social responsibility point, rather than the personal psychology point, which is the Library of Congress, the Powerhouse Museum, the Smithsonian, the New York Public Library, Cornell University Library, and others are putting up pictures that have only been held in archives, up onto Flickr for people to share and look at. This is a super-exciting, amazingly exciting initiative. It's been around for a few years.

[New Slide: A Commons Sample]

Please go and look at it. There are some fantastic pictures. This is really what Ben was talking around.

[New Slide: Images and Text]

This might be history rather than science. It's not the Hubble. It's the Hubble of the history of social sharing. The amount of comments that people are putting on these pictures. This is really crowd sourcing. It's everything from phatic emotional, "My gosh, that's a fantastic picture," to "That's a picture of my grandmother and she was a riveter and her name was blah, blah, blah, blah, blah." And, "That's a picture of the house where I grew up in 19-whatever."

It's absolutely fantastic history data set, as well as a data set around social collaboration and storytelling and the telling of place and time.

[New Slide: Skeleton Photo]

Here's one of the examples that's used frequently. I'm going to put a paper up. I really recommend you reading this paper because it makes some nice points on the notion of social media and what constitutes data.

[New Slide: Skeleton Photo with Notes]

These are notes. Remember the interface elements allows you to put notes on? This is one of most notated images in The Commons. If you read it's everything from, "Cool," to "That's a (INAUDIBLE)" and technical information there.

Notes are allowed and they were clearly designed to put one or two on. Actually trying to read these is a design issue. It's quite hard trying to read them, pick them out and work out the metadata.

When were they put? By who? Is this a person that I should trust?

Those data are amazing.

[New Slide: Museum and the Web 2011]

If you get a chance to read this paper, please have a look at it because they talk about the data and metadata and how you can and cannot extract certain things to tell certain stories about the telling of the history of these pictures. They've got super smart people working on this problem. They also don't even really know how to do the evaluation of, is this an engaged audience or not, because people are putting links to these on their personal sites and so the data of what the influence and impact is of this image are spread across so many different places and sources it feels like it's impossible to really collate what does this mean to a culture or an individual.

They do a really nice job in this paper of talking about some of these details.

[New Slide: Data Longevity]

I wanted to bring up this last point about data longevity. They say that all Flick Commons members, the other qualitative measure we value highly is the sheer inventiveness of Flickr members who engage with the photographs.

What are the metrics of inventiveness? Are they available through the data that we collect through Flickr API or not?

Currently, Cornell saves links to examples of reuse on delicious and displays them as a feed on its website.

There are other places where these are being collected, other platforms. Remember decentralized versus centralized?

[New Slide: Article - Yahoo Sells Delicious to YouTube Founders]

I want to say Delicious changed during the time between that research and now. I want to put up here that Yahoo was really responsible for giving us a transition period to move, to take our accounts.

I was just not paying attention and I missed it. I lost all my delicious links. My personal responsibility, I should have notionally known that I had a very narrow window to do it and I should have taken action to do it. I wonder how many other people forgot in the window? I don't spend my whole life monitoring. It's like gardening trying to keep track of all of these things.

[New Slide: Why all the pros are leaving Flickr for 500px]

What about Flickr now? Professional photographers are starting to leave because the site, the interface is not being updated and invested in.

[New Slide: Flickr is Dead]

People are claiming that Flickr is dead. I love Flickr. I would do anything to save Flickr.

[New Slide: Important Information Regarding the Shutdown of AOL Hometown, Journals (blogs) and KW FTP]

This is not the only one of these examples. Do you remember Hometown? Gone.

[New Slide: Hometown Image]

People like me missed the deadline for getting their stuff out. Gone.

[New Slide: What a Shambles Text Box]

What a shambles. That's an emotion. Who's charting the metric of this emotion? Shouldn't that be part of our social psychological agenda of the psychology, the Hubble telescope of what it means to be online?

[New Slide: Sorry, New GeoCities Accounts are No Longer Available]

GeoCities. Has this person moved to GeoCities, right? Gone.

[New Slide: Status of Google Wave]

It's not just Yahoo, just so we're all really clear. Ben mentioned this earlier.

[New Slide: Twitter Handcuffs Client Apps with New API Changes]

And, APIs. Twitter just announced this a few days ago. Right. What are we going to do about it?

[New Slide: Business Logic is not Science Logic]

Business logic is not science logic. There are reasons why these things are closing. There's a bottom line a company needs to make. Again, I'm not trying to make a moral judgment, I'm trying to make a science judgment around the ethics and who is participating in the conversation and how.

[New Slide: Design/Science of Participation]

That's what I've been focusing on, the second point.

[New Slide: Reflections of Requirements]

I'm also focusing on the requirements for science. I'm really pretty old. I had a training in science that was too positivistic and reductive back in the day when we actually cared what a sample and population were and we didn't just say, eh, the bigger the number, it doesn't really matter.

(LAUGHTER)

I still don't quite believe that.

Stability, the existence of content in an accessible form and hopefully the same format over time would be lovely. Is it possible?

What happens if we don't have it? What's the new science? Science for me was about consistency, recoding the data the same way. Remember I said in the beginning, data science is about massaging?

Reproducibility. I want to point out these were actually points taken not only from experimental science but when I was moving to content analysis and science and analytic approaches. Everyone was telling me if I moved from quantitative to qualitative I was going to have to be more rigorous but no one would believe me. We really thought about consistency, reproducibility, accuracy, validity and the notion of what constitutes proof. We were really careful.

How many people were trained on Atlas TI and other things where you were so careful about your codes? God, help us. We're going to shove them out to give them to everybody to make sure the inter-rater reliability was there. Where's that? Where are we going to have that?

Generalizability. I really like the interventions. How are those really going to be compared?

Maybe it's all been solved. If it has, can everyone in the room tell me because I need to know.

[New Slide: On (2) = Sciences of the Social]

All right. The point I made at the beginning. We need to have everybody, the epistemologists, ontologists in the room thinking about this.

[New Slide: Questions?]

Now, have I depressed you all?

(LAUGHTER)

No, because this is our chance. We have to come together, think about this, and think about how we're going to address this going forward.

Back to Ben's point: as you're doing your specific, detailed research, think about how we can actually learn the lessons. Esther (ph) is going to be talking. She did a wonderful book on things that weren't reproducible and didn't go according to plan, and things that were hard in doing qualitative and quantitative methods. We want to be sharing these things.

What's easy, what's hard? Where is the massaging occurring? Where are the data sets disappearing? I want to hear about them. I think if we can hear about that, we can start to provide some of the collective substrate behind going forward to get the Hubble telescope model stable, and at least discussable.

All right. Lampoon me. Thank you.

(CLAPPING)

SHNEIDERMAN: We're going to take some time for questions.

UNIDENTIFIED MALE: (OFF MIKE)

CHURCHILL: I'll still feel the love even if you're shouting at me.

QUESTION: You talked about data and how you can get stable (ph) data and you say that business logic is separate from (OFF-MIKE). We, as scientists are dependent on those businesses to give us (OFF-MIKE) data because we are in the business (OFF-MIKE). It's always difficult to get data from companies, and you know that.

CHURCHILL: I know that.

UNIDENTIFIED MALE: What are your thoughts on that going forward?

CHURCHILL: I don't know and I think that's what I would like to have as a discussion in this group. I know that many companies actually want to do the right thing but they don't necessarily have the time or resources to do it effectively and there are piecemeal efforts like internships, they're certainly not piecemeal but you know what I mean, and data sharing agreements and visiting scholars and so forth.

I think companies are really trying hard to do the right thing by the data and they really want to hear from scientists about how to collect what data in a computationally effective way. I think what we need to do is actually make use of Ben's (OFF-MIKE) to help us going forward. How do we make requests that are really rooted in a collective science agenda, not just a personal and (OFF-MIKE) one? How do we think about things like privacy going forward so we can protect the company's requirements and policies on individuals, as well as answer our (OFF-MIKE).

(BACKGROUND CHATTER)

I think you just asked a question and really beautifully summarized well (ph) and what I would like us to do is think about how we can share stories to take us forward as a constructive, collective agenda, not a set of individual research results and (OFF-MIKE).

(BACKGROUND CHATTER)

QUESTION: You talked about, do we have the right data, are we looking at the right data, is it the only available data? Sometimes it feels like we're talking about this mythical good data that exists, where the data is in the same format over time, we get a complete picture of all the data, we have access to all the data we want. I feel like a lot of the kind of science that we do is basing our results on that mythical good data, instead of the actuality, which is every big data set I've ever seen is kind of crazy, terrible and really traumatic to look at and go, what am I going to do with that.

My question is, do you feel we have to reorient ourselves to look at what data actually looks at now as oppose to the idea of it?

CHURCHILL: I would argue exactly the opposite and argue that people are playing with the messy data and claiming results that actually make us assume those data aren't quite so messy. They're not revealing the everyday practices of (OFF-MIKE) and that's the agenda I want to take forward, to reveal the practices. Not because those practices are bad and not because there's not ideal interface but because of the layers of accountability and there is a rhetorical way in which we can interrogate the results and understand where the latitude is and where we can do different things, including making recommendations back to corporations about how they can collect cleaner data, which will result in a business result for them, as well as a science result for us.

QUESTION: Don't you think that there's incentive problems in the sciences in the sense that we're motivated, when we're trying to get our (OFF-MIKE) published, we're incentivized to say, look, I've solved the world's problems and (OFF-MIKE).

Is there a way around that?

CHURCHILL: I don't know. I hope there is. When I was a graduate student, we wanted to have a Journal of Nonsignificant Results.

(LAUGHTER)

CHURCHILL: Esther's book, which is really looking at the problematics of certain kinds of research is one step in that direction and this kind of sharing is the way we start to get the messiness of science -- and all of its glorious messiness -- forward to maybe start addressing that institutional requirement, as well.

I don't know. It could be as simple as at the end of our papers, we always get to write what was hard. Some journals like this, some don't. Some conferences like it, some don't. When I was writing experimental psychology journal papers, and I always wanted to apply it. If you do implicit learning, this is what you find, this is how you imply it to a design situation. I was told to take that messy bit out.

What I did was left experimental psychology and became a design person. I wonder if there's a way we could collectively get the whole community moving forward to talking differently about this.

SCHNEIDERMAN: Let me just add to that and remind you that Zeynep Tufekci and Lee Rainie will be speaking Thursday at the Brookings, are both addressing this issue. Zeynep's current paper about what's wrong with big data has provoked discussion.

Jimmy Lin, Jenny's faculty member who spent two years at sabbatical at Twitter really responded about

how the people there are really attending to this. He claims they're devoted and careful and trying to make these careful assumptions, careful refinements to the data so that it can be used.

The pressure of people in this community is extremely important. You do need to speak up about the problems you face and the difficulties that make it hard to run science. I might say at the same time, this is a substantive enough issues to provoke a research paper about how to clean up social data.

I read the very powerful arguments about the climate science Data -- Richard Muller of Berkeley turned around. He was a climate skeptic and now he's a climate supporter. They wrote five papers on this, one of them was entirely on how they cleaned the data from 35,000 observation sites over 250 years. A whole paper just about data cleaning. We're not the only ones with data problems and there's important to make it a national priority so they we work with reasonable data.

CHURCHILL: Agreed.

Thank you very much.

SCHNEIDERMAN: Another question? We'll take one more.

CHURCHILL: I don't want to make anyone late.

QUESTION: When we think about big data I keep thinking whether the current narrative (ph) is just about one (OFF-MIKE) of the data in the sense that actually if we're doing qualitative research -- (OFF-MIKE) and what not, actually we could utilize the technologies to do a lot of detailed lobbying (ph) -- just thinking about what our (OFF-MIKE) on each of these sites where (OFF-MIKE) but we could start thinking about -- (OFF-MIKE).

CHURCHILL: Absolutely. Some of the work I've done with Ayman Sharma at Yahoo -- he and I were looking at how you blog social media systems for experience mining and for ethnographic insights, what logging do you do?

I think your question provokes an even bigger agenda around that and I'd love to discuss what that might look like in the next few days.

Great.

SCHNEIDERMAN: All right. Thank you.

(APPLAUSE)

END