OPEN SOCIETY FOUNDATIONS

TRANSCRIPT

# "DIGITAL ECHOES: UNDERSTANDING PATTERNS OF MASS VIOLENCE WITH DATA AND STATISTICS"

*A conversation with Patrick Ball*

*Moderator: Elizabeth Eagen*

## ANNOUNCER:

You are listening to a recording of the Open Society Foundations, working to build vibrant and tolerant democracies worldwide. Visit us at OpenSocietyFoundations.org.

## ELIZABETH EAGEN:

So welcome, everyone, to this talk on Digital Echoes: Understanding Patterns of Mass Violence with Data and Statistics. I'm Elizabeth Eagen with the information program of the Open Society Foundations. And-- this is Patrick Ball, who is the executive director of the Human Rights Data Analysis Group. Now, I hope you all had a chance to poke around and-- and click on Patrick's resume or his bio.

Because it is-- it's quite an impressive list of places that he's done his work. But-- I'll just read a few lines from it. So Patrick has spent more than 20 years conducting quantitative analysis for truth commissions, nongovernmental organizations, international criminal tribunals, and United Nations missions in El Salvador, Ethiopia, Guatemala, Haiti, South Africa, Chad, Sri Lanka, East Timor, Sierra Leone, South Africa, Kosovo, Liberia, Peru, Colombia, the Democratic Republic of Congo, and Syria.

And I wanted to remind everybody the talk is gonna be recorded. So just keep that in mind. But we will have a mic for questions. Patrick's gonna run through a brief presentation-- with-- I hear there's gonna be some pedagogy, some math. There will be some math. (LAUGHTER) So you should feel free to take a break if you need it and go grab some wine and cheese (UNINTEL PHRASE). (LAUGHTER) I'm gonna go

ahead and turn the floor over to Patrick Ball. And then I'll come back to the table when we're ready for questions. So please do have those lined up. And I will see you soon.

## PATRICK BALL:

Thank you, Elizabeth. Elizabeth anticipated the trigger warning that I was going to provide. There will be algebra. (LAUGHTER) There will be probability. And there will be taking data seriously. If any of those things upset you, perhaps you should step out. (LAUGHTER) Because my argument tonight will be that it matters very much that we get the answers right.

And it's dangerously easy to get the answers wrong when we look at data. What I will argue tonight through a series of examples is that data as it comes to us in the human rights movement is fundamentally biased. And by biased, I do not mean that it is politically connected to one or another party to a conflict but rather that the da-- that there is some data we are likely to get and there is other data we are unlikely to data.

There are blind spots in our digital picture of the world. And I'll explain what I mean by that and provide three examples. These are only three examples. My colleagues and I over the last ten years have started accumulating these examples. We have now 15 different examples.

That is: Every single country that we can test we find that the raw data tell us the wrong story. Not a story that is somehow precise. Not a story that's, well, almost good enough. But rather it tells a story that is fundamentally misleading-- with respect to one of the fundamental intellectual, and political, and historical questions that we have about that country.

That is: The statistics do exactly the wrong thing. They lead us toward false certainty unless we adjust for what we don't know. I've been at this a long time. I started this work in El Salvador in 1991—building databases for human rights groups there.

I was very excited at the time about the-- what was then very new potential of using-- with my colleagues at the nongovernment human rights commission of El Salvador, we used over 9,000 testimonies and linked those testimonies, the violence documented in those testimonies we linked to the career structures of the officers in the Salvadoran military in order to create individualized human rights dossiers on each of the officers in the Salvadoran military.

Our goal with that work was to-- demonstrate which officers had sufficient crimes alleged against them such that they should be obliged to retire—following the agreement of the peace accords in El Salvador. And it worked. The officers that we identified as the 100 worst officers in the military were an almost exact fit to the 100 officers forced to retire by the ad hoc commission-- as set up by the-- peace process in El Salvador between the government and the FMLN.

And I-- I was hooked. But I was hooked, as I'll explain, on maybe the wrong thing.

So I spent the '90s building databases for truth commissions. And while that work was very exciting and very satisfying, at the end of the '90s in two different commissions in South Africa and in Guatemala, I was asked by people workingin the commission, "Hey, how many killings-- were there?

"How many people were killed in South Africa during apartheid?" And I said, "Well, you know-- we have about 6,000 here in the database." That's not the question. The question's, "How many people were killed?" And I didn't know. So what I'm gonna talk about today is how I've been answering that question for the last 15 years.

And it comes down to this fundamental problem. It's-- it's an epistemological problem. It's-- it's-- it's a problem that all people face when we look out at the world. What we know is a function of who we talk to, 'kay? But what we don't know is a function of who we don't talk to.

And that may seem obvious. But this becomes very problematic. Because it may be the case that there's a systematic difference between what we do know and what we don't know. And so if we base conclusions on what we do know, we're making bad decisions. We are reaching the wrong conclusions.

Because there's a different story that's not included in what we-- in what we have. There's people we didn't talk to who have a different story than the people we did talk to. And it turns out that this as true for databases as it is for qualitative researchers. In fact, I argue it's worse in quantitative work because in quantitative work as soon as the graph hits the table, we forget that there are pieces missing from the graph.

And so I wanna open with an anecdote from-- work I did in the Congo in-- with the U.N. in 2010. And I was in a whole series of meetings. For those of you who speak U.N. talk, we were projection working groups. And one of the things we had to do was try to identify contexts in which specific groups of vulnerable people were at risk so that the U.N. could try to do something to protect them.

And I would sit in meetings with people who had worked in Congo for many years, people who spoke loc-- the local languages, had rich networks of connections and contacts with people in-- Congolese society, specif-- in-- you know, in-- in particular areas that we were looking at.

They understood the military history of the region. Very important. By 2010, there had been many different militias that had come and go. They knew who the local officers in the-- in the Congolese army were. They knew who were likely to be officers who might defect and join-- and-- and rejoin militias, as this happened repeatedly in Congo.

But they knew local leaders. They knew civil-- the-- the civilians who-- were at-- most at risk. And we'd have a rich, engaged, thoughtful, qualitative conversation about what needed to be done, about what the risks were, about who was at risk, about what violence had happened in the last few months.

And then someone would put a graph on the table. And it's like all the brains

drained out of people's ears and ran across the floor. (LAUGHTER) And people would say, "The graph, it goes up. And then it goes down. It goes up and down." And the conversation-- the quality-- the intellectual quality of the conversation went from a PhD seminar to *Sesame Street* in less than a second. (LAUGHTER)

And I was the only statistician in the room. And I was like, "Wait, wait, wait. The-- this graph-- this-- this graph doesn't mean anything. Come on. This just means it went down because our staff all just went on leave and they're not writing anything down this month." And they're like, "Well-- but that's the staff that was documenting the violence by the LRA.

"And so we sent the report to the U.N. Security Council saying that LRA violence has declined." And I'm like, "They went on leave. There's no decline. (LAUGHTER) We have no idea if violence is up or down." And they're like, "Well, it's already been minuted in the Security Council." That's how bad this is, people.

We're making global human rights policy on really, really weak data because we don't know what we don't know. Here's the problem. Let's think about it a little bit in a couple different angles. We collect and combine three databases. This is how most big human rights projects seem to work these days.

People build databases in different ways. Through press monitoring, through individual human rights conversations with victim communities, with other kinds of records from governments, from militaries, from police. And let's say we can combine those databases.

We can merge them together. We can determine which victims are in common across multiple databases. I work almost exclusively with homicides now. And that's the easiest of all data to integrate in this way. So we now have the world as is described in these white circles. But which world do we live in?

Do we live in the world here on the left where the white circles constitute maybe a third of the reality? Or do we live in the world over here on the right where the white circles almost completely exhaust the reality that we're looking at? And the reason this is so important is not for the magnitude.

I mean, that-- that'll get the attention of the media. You know, if you can say how many people in total have been killed. But because we probably want to compare the world on the left with the world on the right. We want to say, for example, that the Maoist guerillas of the Sendero Luminoso have committed more or fewer killings than the Peruvian army. Which is it?

Well, this is exactly in fact what we saw in Peru at the-- at the truth and reconciliation commission there where we had collected approximately equal numbers of documented killings attributed to Sendero Luminoso and attributed to the Peruvian army. But when we did some math, we found that Sendero Luminoso had kill-- committed many, many, many more killings than the army.

And I'm gonna show you that math later. Not a funny joke. Stay for the algebra. That's the good part. In all seriousness, it's the part that actually gets us out of the

limited world of the white circles and takes us to a much more global understanding but an understanding that is limited by the rules of probability.

So let's get back to the problem. We usually don't know what we don't know. And this-- the challenge that we face there is that what we don't know may and in fact likely is systematically different from what we do know. Uh-oh. This is true of almost all human rights data. Big data does not make this better.

In fact, this problem that I'm describing is called by statisticians selection bias. It is the bias introduced into data because we got that data and we didn't get this data. And my experience of big data is that big data does not ameliorate selection bias. It amplifies it.

So let me tell you why that happens. We used to in the world of human rights, in the world before cell phones all had computers in them-- if we heard about a killing, we'd report it, right? And we have the saying in our team that if a lawyer is killed at high noon in a city center, the world knows about it before dinner time. 'Kay?

And that was true in the 1980's. And that's true now. Back then we would have gotten a few reports of it. Now, we'll get hundreds, maybe thousands. But if some peasants are killed in a rural area three days' walk from a road, we didn't hear about it in the '80s. And we don't hear about it now.

This problem has not changed. And it's not-- and-- and the reason it hasn't changed is that it's not a fundamentally technical problem. It's a problem with how information moves through different kinds of social contexts and who trusts whom to share information. If you're in a tiny rural vig-- village in eastern Congo, nothing good has ever come to your village from outside.

Why would you report something horrible to outsiders who could then come and bring more bad things to your village? It just doesn't happen. So the fact that most of eastern Congo has perfectly good cell coverage now has not changed this reporting problem. So before I go much further with all this abstraction, I wanna drop into some concrete examples.

I'm gonna start with a discussion about what we know and what we still don't know about-- lethal violence in Iraq between tw-- 2003 and 2010. I'm gonna-- then gonna talk a little bit about how we've learned a little bit more about the scale of police killings in the United States. That is killings committed by police officers in the United States.

And I'll be commenting on a report-- by the U.S. bureau of Justice Statistics that was-- presented in early March this year. And then I'll give an example of some of our very first estimates about killings in Syria. So let's look at Iraq. Since-- in early 2003, a group of British researchers created a group called the Iraq Body Count.

And the Iraq Body Count has done an astonishingly meticulous and careful job of collating information about violent incidents around the worl-- excuse me from media around the world-- about violent incidents that occurred in Iraq. So when people get killed in Iraq, if it turns up in the media, people at the Iraq Body Count

create a record in their database.

And they say, "According to the sources, according to the BBC, or Al Jazeera, or the *New York Times*, or Reuters, or the AP, or whatever-- whatever their source is, we have an incident here that-- in which five people were killed at such and such a location at-- on such and such a date."

And if they get a different report that says ten people were killed in the same incident, they'll merge those two reports together and put a high l-- high number on of ten and a low number of five. And on and on they go. And they can accumulate quite a few sources. Well, starting in about 2005 and '6, colleagues of mine and I looked at that database and said, "You know, we think that not all the violence in-- in Iraq is getting covered."

And in a series of discussions with the people from the Iraq Body Count, we said, "Stuff's not getting covered here." And they said, "Well, th-- maybe that's true. Maybe it's not. But nevertheless, we're covering so much of it because such an intense level of international media attention is directed at violence in Iraq during this period that our data must be good enough.

"It's good enough to do things like assess or not more men or more women are being killed, or whether the victims are elderly people, adults, or children, or whether the primary perpetrators are coalition forces of Al Qaeda in Iraq." "Wait. Isn't there another perpetrator?"

"No, not so much. Actually it doesn't appear too much. So let's not worry about it. It's not in the data." Okay. Now, we're in 2015. And if people think a little bit about what was going on ten years ago in Iraq, this should be making you very nervous. So let's look at the graph and see what we can learn.

We looked at the data. My colleague Megan Price (PH) and I looked at the data and divide-- looked at the IBC data. And we divided it into categories by size of event. So in this bar, we represent the events that have one victim. In this bar, two to five victims. In this bar, six victi-- incidents with six to 14 victims.

And here, incidents with 15 or more victims. And then the shading on each bar represents the proportion of each set of events at each size that have a certain number of sources. So over here, events of size one, you can see that almost all of them have one or two sources. Only one or two sources are documenting those events.

And a much smaller number of events of size one are documented by more sources. If we can contrast that with events of 15 or more victims, we find that none of them are documented by only one source. Only a s-- very small fraction are documented by two sources. In fact, most of the very large events are documented by 15 or more sources.

So here's a question. How many events are there with zero sources? That is: How many events are not documented at all? And where are they distributed across this event size? Anyone wanna guess? Are there more events of size one or more events

of size 15 and greater that have zero-- that are gonna have zero sources?

## MALE AUDIENCE MEMBER:

One.

## PATRICK BALL:

One. And why would that be? Why-- why does this give us that-- that intuition?

## MALE AUDIENCE MEMBER:

Because people, they can can be getting killed all over a big country. And we know there's one of out-- one out of ten of 'em get noticed.

## PATRICK BALL:

They hardly get noticed. And the way we-- we-- we realize that is because if you think about how many there-- how many events there are here of size one that only have one source, quite a few, that implies there may also be quite a few with zero sources. But over here, almost none of them. I mean, in fact, none of them have only one source.

Only a small number even have two. Most of 'em have many sources, which indicates that these are very well documented. Maybe they're almost all covered. Maybe there aren't any events or very few events that are that large that have gone unnoticed in the world's media. Okay, this is a huge problem. Okay? Because s-- small events and large events-- sorry, question?

## MALE AUDIENCE MEMBER:

What's the ratio of events with one death versus 15 deaths?

## PATRICK BALL:

Well, there's two ways to answer that question. And-- I don't know the exact answer off the top of my head for either of them. But one of the ways is: What's the ratio of the number of incidents between those two? The other is: What is the ratio of the number of victims between those two?

In both cases, there are slightly-- well, there are slightly more victims who die in events of-- of size one than there are victims who die in larger events. Okay? So

there's m-- there are relatively few large events. Okay. It's a power law (?) for those of you who follow that kinda thing. Yeah?

## MALE AUDIENCE MEMBER:

Are those independent sources? Or were they eventually, like, a wire service reporting to different-- different newspapers and them all reporting the same thing?

## PATRICK BALL:

Right. Do they play follow the leader? They tried to-- IBC reports on their website that they try not to report simple-- they-- they don't document things that are merely reprinting another-- another story. They just try to capture the original write-up on that. Okay?

So let's get back to why this is such a gigantic problem. It's a gigantic problem because small events and large events were actually two different conflicts happening at the same time. Okay? They're actually completely different. So let's go through all the ways they're different. First of all, the perpetrators are completely different. Right?

Most of the small events are committed by Shia militias that were essentially undertaking ethnic cleansing mostly in the Baghdad area whereas large events were Al Qaeda in Iraq, other insurgent groups, as well as coalition collateral damage incidents. The weapons are completely different.

The small events were almost all firearms. Large events are IEDs and air strikes or-- or firearms in massacre contexts. Small events are-- the victims are completely different. Small events are almost all adult men. These are selected, targeted killings. Large events are random selections from the population.

They include women, children, elderly people, adults-- and adult men all in roughly approximately equal proportions to their-- to-- to their proportions in the public population, the population that goes outside. And finally, the goals, as I said, are-- are-- are completely different.

So now as we watch the rise of I.S.I.S., which is very much an outgrowth of ethnic cleaning because those are Sunni groups who've been cleansed who now feel like I.S.I.S. might protect them from this-- the Shia militias, we can start asking, "Wow, by overstating the importance of large events relative to small events, first thing we did is we reinforced our initial biases, right?"

The world was really interested in Al Qaeda because this is part of the larger narrative from 9/11. So this is interesting. This was what we thought we were going to see. So naïve statistics failed to tell us the story. They failed to provide a corrective to our prejudice and our biases.

And the reason is because we didn't know what we didn't know. And what we didn't

know is that there was this v-- wildly different probability of reporting. And this is a really important idea that I wanna try to express tonight, which is that-- imagine that an event happens in the world.

What is the probability that we hear about that event? What is the probability that social knowledge is created about that event such that we actually have something we know? Well, maybe that probability varies a lot. Maybe that probability in fact depends on the kind of event.

So the probability of a large event being reported was very high. Maybe almost one. Whereas the probability of a small event being reported was very small. Maybe as small as .2. and that variability in reporting probability we're gonna see everywhere. And not only does it vary. Variability in reporting probability correlates with everything we're interested in.

That's why it's so devastating to our analytic-- our analytic capacity. Because the-- the variation in reporting probability is connected to the kinds of things we're trying to understand. So event size, right, something we really needed to understand because event size is determining the-- the way that the patterns come out on specifically the questions we're trying to ask.

Like, "Hey, what's going on? What are the events we need to pay attention to?" Yeah, except we're not seeing a whole huge fraction of them because they're hidden due to the way reporting works. So that's a big problem. I hope you're all really scared now. I hope you all feel like we don't know anything.

Because that's exactly the right motivation to do algebra. Because algebra (LAUGHTER) is actually gonna give us a really nice way to come out of this problem, okay? Algebra and a little bit of probability theory. Very basic probability theory. It turns out that the way we actually do this in the real world crazily harder than this.

There are people who write entire PhD dissertations and devote their careers in mathematical statistics to this. Put your hand up. And you can talk to her because she's done really amazing work on this stuff and terrific models that have helped us-- learn a lot more, particularly about the analysis we're doing in Syria.

Her name is Sherri Mitchell (PH). Please talk to her later if you-- if there's time. Here's the problem, okay? We're gonna look at another set of Venn diagrams. We have this circle. And statisticians by convention call a population that we're trying to look at N. That's how many people are killed, for example-- you know, by police in Texas, for example.

We don't know how big N is, okay? But we-- what we do know is that we have a list compiled by someone. And I'll talk in my next slides about some lists of police killings in the United States. We have a list with some number of people on it. We'll call that project A. And what project A has done is it's gone out into this space of killings of-- by U.S. police in Texas and it's documented some number of victims.

Not all of them but some of them. And the-- so the first question we ask is: What's the probability that a death in this population N is documented by project A? Okay?

So this is very similar to the problem of asking, "What is the probability that if I throw a coin I've thrown a head?" Okay? What is that probability?

## MALE AUDIENCE MEMBER:

Isn't it 50%?

## MALE AUDIENCE MEMBER:

Shouldn't it be 50?

## PATRICK BALL:

50%. Why is it 50%?

(OVERTALK)

## PATRICK BALL:

There's two possibilities, a head and a tail. We're looking for one of them, the head. Okay? And we're gonna draw once. So-- we're gonna draw by throwing the coin. So we've thrown the coin. And we know that we have probability-- .5. Similarly, the probability with which a death in the population N is documented by project A is A divided by N. People follow that?

Nod if that makes sense. Okay. The probability with which project B documents a given death in N is very similar. It's B over N. And here's where it gets more interesting. Remember that we can put the two projects together and determine which of the killings are in common between the two. 'Kay? That's M, the part in the center, the intersection.

And the probability that a death in N falls in M is the same as A and B. It's M over N. However, M is something else as well. M is the result of having been documented by A and B. And so this is like throwing two coins and asking, "What's the probability I've thrown two heads?" What is it?

(OFF-MIC CONVERSATION)

## PATRICK BALL:

Yeah, I hear it. Go ahead. Say it louder.

(FEMALE AUDIENCE MEMBER: UNINTEL)

## PATRICK BALL:

25%, right? It's one quarter. Why is it one quarter?

(OVERTALK)

## MALE AUDIENCE MEMBER:

--one of two possibilities.

## PATRICK BALL:

It's the first probability, half, multiplied by the second probability of half. 'Kay? Similarly, the probability that a death is in A and B is the probability that it was in A multiplied by the probability that it was in B. So M over N is equal to A over N times B over N. Now, we rearrange the terms.

We multiply through by N squared. MN equals AB. We divide by M. Hey. Hey. This is really cool. We know what N is now. We didn't know N when we started. And now we do. Well, at least we have an estimate of it. And I made four assumptions as I went through there. And I'm gonna worry in the next slide about what those are. Or at least one of them.

But you guys see the power this gives us? 'Kay? Now, we can use data that is collected by normal human rights projects, 'kay? Projects that are not statistically oriented but rather projects that have as their goal to write down the names of victims, to do-- to record historical memory, to clarify history, to build cases for prosecution, to do the stuff that human rights groups do.

But we can also do statistics with it. We can get to an answer that corrects for what we don't know. That was the problem, right? That we were-- we-- our-- our-- our data was being distorted by what we didn't know. Well, now we can correct for that. That's really powerful stuff.

And it's so powerful that even government statisticians use it. That's how powerful it is. So for some time, people in the Bureau of Justice Statistics along with everyone paying attention in America has been worrying how many-- about how many people have been killed by police in the United States. How many people are there? There's no list. 'Kay?

The F.B.I. keeps a list that they, depending on which version of their websites you look at, usually admit is a very partial list called the supplementary hom-- supplementary homicide database, which is not public but is available to Bureau of Justice Statistics researchers.

The Bureau of Justice Statistics keeps their own list called the arrest-related deaths database, which is a separate and independent list p-- compiled primarily from media sources. But they have a bunch of other inputs as well. And in March this year, they

published a report showing how those two databases interact, okay?

So they put the two databases together. They merged them just like I did in the previous slide. And they did the math that I showed you. N equals AB over M. Okay? Here is A, B, and M. And they came up with about 7,400 deaths for the period 2003 to 2009 and 2011, which is kinda weird. It's not really clear to me what happened in 2010. But there it is.

## MALE AUDIENCE MEMBER:

That number-- that is in what population?

## PATRICK BALL:

In all of the United States in those eight years.

## MALE AUDIENCE MEMBER:

The number's 7,400?

## PATRICK BALL:

Plus, minus. Yeah. Yeah, they didn't actually report the variance, which I think's kinda weird. But whatever. They did something much worse. So we're not gonna worry too much about that piece. So here's the k-- here's the key piece. You remember how when I did the math in the previous slide I said there are some assumptions?

Well, let me give you a metaphor about how this logic works. And then I'll tell you what the problem is. The-- the method we're using here is kind of as if you're facing two dark rooms. You can't see into the rooms. And for some reason, you can't go into the rooms. But you'd like to know which of the rooms is larger.

So you have a handful-- your only tool here is a handful of small rubber balls that bounce very energetically. So you take the balls and you throw them. They have-- I'm sorry. The balls have a curious property, which is when they hit each other, they make a (MAKES NOISE) noise. So you throw the balls into the first room and you hear (MAKES NOISE). Collect balls, go to the second room, throw them with equal force. (MAKES NOISE) Which room's bigger?

## MALE AUDIENCE MEMBER:

The second.

## FEMALE AUDIENCE MEMBER:

The second.

## PATRICK BALL:

Right. Why?

## MALE AUDIENCE MEMBER:

Because--

(OVERTALK)

## MALE AUDIENCE MEMBER:

--if it's smaller, they're gonna hit off the wall and hit each other more often.

## PATRICK BALL:

Exactly. They're compressed in the first room, 'kay? That's very much like what we're doing here, 'kay? What we're doing here is though-- is as if we're throwing these databases into this sort of theoretical room of killings in Texas, for example, in twenty-s-- so 2006. Okay? And we're observing the databases colliding with each other.

And that's the clicking. That's M. And the more collisions we hear, the smaller we think the whole universe is. But this only works if the balls are independent of each other, if they just fly around without knowing where any other balls are. What if the balls kinda like each other? (LAUGHTER)

As they're flying by, one ball sees the other one and says, "Hi," and they reach over and hit each other. They're kinda, like, magnetic a little bit. So they're flying around the room. But when they get close, they-- (CLAPS) they hit each other. What happens then? We get too many clicks, right? We get too many (MAKES NOISE) noises.

And our inference about the size of the room is too small. That's what happened to the B.J.S. 'Kay? So the B.J.S. estimate was too low because they failed to control for the positive correlation in the probability of reporting between the ARD database and the F.B.I. database. Why would this be? What would create a positive correlation? What does that mean in non-statistics talk? Okay, remember that lawyer who got killed at high noon in a city center? What's the probability that he was reported in the newspaper?

(OFF-MIC CONVERSATION)

## PATRICK BALL:

Very large. What was the probability that the F.B.I. bothered to write his name down in the supplementary hom-- homicide database?

## MALE AUDIENCE MEMBER:

Very big.

## PATRICK BALL:

Correlated, right? Because-- th-- both of them are driven by some underlying notion of social visibility. Socially visible people are going to appear in both lists. Socially less visible people are unlikely to appear in either list, 'kay? And so those two probabilities are correlated, which creates a positive relationship between these.

And we knew this when we read that report because we've done this work in a lot of different countries. In ten different countries in fact. So my colleague-- Christian Lum (PH) with some very sage advice from Dr. Mitchell-- spent quite a bit of time thinking about pairwise list dependence, which is the statistics talk for the magnetism between the balls, 'kay?

And we looked at all our old projects, projects from Kosovo, Colombia, Guatemala, Syria, and Sierra Leone. And we said, "What's the distribution of correlations in these other projects?" Because it turns out if you have three or more databases with all that cool math I was telling you about earlier, we can actually calculate the magnetism.

And if you can calculate those correlations, you can control for it. You can include that as part of your calculation. And you can get much better estimates. The B.J.S. only had two lists. And as they said openly in their report and repeated when they were interviewed by 538.com, "Yes, there is positive relat-- there is positive correlation in these two lists. But we don't know how to control for it."

And we don't know either. Okay? But what we do know is what those distributions typically look like. We know what kinds of list dependence are likely. And in fact in particular in Colombia, we think the lists that we have for Colombia that we're using in this simulation are very much like the lists used-- by the Bureau of Justice Statistics.

So we can use the information about this magnetic attraction between the balls to correct the estimate. And what we get is not 7,400 but more like 10,000. 10,000 people killed over eight years. Twenty-- 1,250 per year on average. 1,250 people killed by police on average. Oh-- and, oh, right. We're only including70% of the jurisdictions in the United States.

It's more than that. It's more than that. It's more than 1,250. There are about 1,600 homicides in the United States every year. At minimum, there's s—something like

7.8% of all the homicides in the United States are caused by police. I think it's probably closer to ten. So let's just take that and think about for a minute what it means to say that 10% of the homicides in the United States are caused by police. Okay?

Three quarters of all homicides in the United States are committed by someone you know. If you're going to be a homicide victim, most problem outcome is that it's an acquaintance of some kind. So among the quarter of all homicides that remain, more than a third of them are police. If you are going to be killed by a stranger, the most probable outcome is that he's wearing a badge. Let's turn to Syria.

## MALE AUDIENCE MEMBER:

Can I ask you a question? So how many police are threatened? How many-- how-- (OVERTALK)

## MALE AUDIENCE MEMBER:

How many police have been killed by-- criminals?

## PATRICK BALL:

Very few. Very, very few. There is a really good list of those. The-- National Fraternal Order of Police has a website that keeps-- has really excellent data on that. And you look at police killed in the line of duty. They are, I think, the 15th-- the 14th or 15th most dangerous profession in the United States. They're not even in the top ten. And nearly all of the police deaths that happen-- at work happen as a result of car accidents.

## MALE AUDIENCE MEMBER:

Yeah, but that can't be the case listening to the police.

## PATRICK BALL:

You answered your own question. (LAUGHTER) So let's turn to Syria.

## MALE AUDIENCE MEMBER:

Excuse me.

## PATRICK BALL:

Yes?

## MALE AUDIENCE MEMBER:

How-- how many of those homicides that police commit do they know the person that they're-- that they-- that-- that they kill?

## PATRICK BALL:

Fascinating question. I don't know. We have-- I don't have anything like that precision on the data. That's a fascinating question. The narratives that I generally read about doesn't seem like they know very many. Okay? But I don't know because I have a very bad sample on that. Good question.

That would be really interesting to know. So from March 2011 to April 2014, in that period we worked with-- a whole lot of different Syrian groups. Nine different groups. We're only gonna use four of them for this example today. And during that period from-- that period, we received the names, the dates, and locations of death for about 191,000 people killed in Syria.

That's how many people we can name. Those are the people who we can-- we have two valid-- in machine learning talk, two valid name tokens, okay? Maybe a first name and a last name. But at least two names that are actually words-- that are name words, that aren't "brother of," or, "neighbor of," or anything like that.

These are actual identifying names. And-- a date of death precise to the day and a location of death precise at the least to the (UNINTEL), in many cases to the neighborhood. So we integrated all those databases. There are-- there's over 300,000—underlying records. But we integrate them to 190-- 91,000 individual deaths.

But the question is: How many do we not know about? Of course, that's the whole theme of this talk. So we're a little bit concerned about that. And, you know, we've got a few qualifiers. I'm not gonna go through them, but I'd be happy to talk them a little bit more in detail in the Q&A.

But I want us for a moment to go back to how this estimation process works. I want us to remember that we can get to an-- an estimate. And here it has a little subscript K. And I'll tell you about that in just a second. By multiplying A and B and dividing by M or by using a much more complicated model that's based on that same logic. Okay?

The K sub-- subscript indicates that we don't make one estimate. Remember when I said that th-- it's not so interesting to just grapple with magnitude? What's more important than magnitude is to compare different categories, often different periods

so you can get a pattern over time. Or estimates for different locations so you can see the pattern over space.

Is there more in the north? Is there more violence in the north? Or is there more violence in the south? Is there more violence in areas of government control? Or is there more violence in areas of insurgent control? So that K, that subscript indicates that we do it for lots of different places, 'kay?

And then we have a logic here, right? After we've got the estimate of N, which is-- wow, how much do we know? How much do we actually know about what's going on? And the way we measure that is the rate of coverage, okay? Which is the number of records we know divided by the estimate of the total number of killings that we believe to have occurred as a result of doing this math.

And here is the problem. Here is what scares us a lot. Those rates change like crazy. 'Kay? The rate of our knowledge, the coverage rate changes like crazy. So in this-- color series of maps, the really dark sections like this one are places where we know 80, 85% of the deaths. We've got most of the information.

But the really light colored-- areas in super light yellow are areas that we may know only 10 or 15% of the deaths. 'Kay? And the gray areas are places where our-- our information isn't sufficient to make an estimate. So look at the variation across there. 'Kay? Watch. Sometimes we can't make an estimate at all. Sometimes we've got almost the data. Then the data gets worse. Then it gets worse. Then it gets really, really, really bad. Then it gets really good. Up and down, up and down.

## MALE AUDIENCE MEMBER:

Is this over time?

## PATRICK BALL:

This is-- these over time, right? These are months. Each one represents a month between--

## MALE AUDIENCE MEMBER:

We can't see the headings on them.

## PATRICK BALL:

Sorry about that. June twenty-- July 2012 through June 2013, 'kay? So each map is a single month. Look at how much variation there is in our knowledge. In our knowledge. You can imagine, I hope, looking at this how distorted our conclusion would be if we drew conclusions from the raw data.

Because the raw data are just an artifact of the reporting process. Any te-- any trend we would see completely, completely dominated by the variation in our knowledge over time and space. Worse, there's no pattern here. There is no pattern. We can't predict that there's going to be a certain amount of coverage without doing the math that we showed in the previous slides.

We're stuck. So I just wanna warn again we should never be drawing important conclusions on the basis of raw observed data. Raw observed data is not a good foundation for an important conclusion. Let me tell you-- let me show you one more about how bad it is. 'Kay?

So here what we're doing in this graph is the purple regions indicate cases that we've documented by four sources, three sources, two sources, one source. And then the light blue are the total that we estimate as having not been documented at all. Now, this is Hama between December 2012 and March 2013.

And I want you in your head to draw a line across the tops of the purple. That's the observed data. And look how nicely it goes down. Wow, it's really great. Things are getting better in Hama. That's great. I—getting better. No, they're not getting better. There's a giant spike in January 2013 that's completely unobserved.

Giant spike there. Huge number of killings not documented at all. What happened? Well, we went back to the newspaper and found that in January of 2013, Syrian government took control of Hama. And they held it until February 2013 when the insurgents took it back. So what this is really telling us about is, you know, the documentation groups', the human rights groups' capacity to document.

That's what's really happening here, you know? When the government's in charge, they have a really hard time documenting very much. Whereas when the insurgents are in charge, they-- they have a much easier time documenting things, right? So it looks in some ways-- if you look at raw data for Syria, it looks like things are equally bad between the government and the-- and the insurgents. This graph suggests it's not-- that's not true at all.

## FEMALE AUDIENCE MEMBER:

But how do you know that there was a spike in deaths if they were undocumented?

## PATRICK BALL:

That-- we do the estimates. We do the estimates for each period, right? The point of the estimation is specifically that, to figure out how many are undocumented, 'kay? So it's precisely by not exactly this equation but this logic that we get to that estimate, okay? It's back to the balls.

If you imagine the balls bouncing around in that room, 'kay, in one month we hear (MAKES NOISE). And the next month, we hear (MAKES NOISE). Okay? And we're

like, "Holy cow. What's going on the next month? Nothing's clicking." There's so much more room in that-- that room is so much bigger. Okay? And there's a lotta formalization we can do-- around that that allows us to know that. Yes, please?

## MALE AUDIENCE MEMBER:

And how did you calculate that there was correlation for-- for that (UNINTEL)?

## PATRICK BALL:

Thank you so much. What a lovely question. (LAUGHTER) So it turns out when you have more than two databases, you can use the third, fourth, fifth, and et cetera databases to in some sense take a look at th-- the pairwise estimates. And from that, you can formalize, "Well, how does it look if I use only A and B?

"How does it look if I use B-- A and C? How does it look if I do B and C?" And by looking at all those angles, you can calculate those pairwise correlations. Okay? For people who have a little bit of statistics training-- we do this in the classical way by building log linear models to predict cell counts.

And one of those cell counts is this count of unobserved deaths. So that's how it's done classically. How it's done in a Bayesian context is so much more interesting and useful. And I will leave that to the discussion afterward in which Dr. Mitchell can attend us. Thanks, Dr. Mitchell. (LAUGHTER)

I'm wrapping up here. 'Kay? Coming to the end. Look, there's only three ways to get to rigorous statistics. And if it's not rigorous, we have no business using it in human rights. People are depending on us. People are depending on us to be right. They're depending on us to be right so that we clarify history, we don't muddy it further with raw data, with naïve raw data.

They're depending on us to come up with arguments that can withstand adversarial criticism in court. That's very hard. Okay? And they're counting on us to advocate for the people who can't even be documented. If we can't name them, at least we should try to count them. But there's only three ways to get to statistics good enough for that to work.

One is you could have a perfect census. That means you have all the data. You have all the data. If you have all the data, you can do anything you want. That hardly ever happens. But I said hardly ever. I didn't say never. Sometimes it happens. In Kosovo, we now have all the data for people killed between 1998 and 2000.

In Bosnia, we have very, very close to all the data-- for the people killed between 1992 and 1995. In Northern Ireland, I think we have almost all the data on people killed in the troubles. In Israel-Palestine, we have almost the data k-- on people killed in that conflict. But those are rare.

Those are exceptions. Very rare. In some places, we have only the tiniest crumb of

the information. Congo is probably the best case of that. In other places, in-- it s-- it seems to be more common projects I work on that we have between a third and two thirds of the data. And that's El Salvador. That's-- that's Peru. That's-- well, Peru is one third.

Colombia-- Syria, Timor-Leste. Those are places where we have this somewhat smaller fraction. So we don't have perfect censuses. Another way you can get rigorous statistics is by drawing a random sample from the population and interviewing them. Now, that's hard. We don't interview very many victims of homicide for fairly obvious reasons. (LAUGHTER)

So it's actually pretty complicated to figure out what a random sample of homicides would be. How would we capture that? And as you get more deeply into that, there are epidemiological solutions which work for very specific ways. But when we're doing long retrospective projects-- as we did in Timor-Leste, it turns out to be very challenging. And the models become quite complicated.

It's hard to do. And there are many challenging technical issues. There have been a series of very well publicized-- very big errors in survey-based-- estimates of mortality in conflict. So it's-- there's a lotta moving parts. It's easy to get it very, very wrong. Finally, there's a third way, which is s-- sometimes a posterior modeling of the sampling process or-- post-stratification.

And the method that I've been showing you today is one of a family of methods of that kind. Capture-recapture is the name of the method that I've been talking to you about-- about-- with the circles, and the M, and the probability, and multiplying up to N. It's called capture-recapture.

There are others. And I wanna be really clear that that is not the only way to do it. There are a couple other methods. The one that is most commonly used is called raking. There-- there are probably others on top of that. I have explained and-- and focused on capture-recapture 'cause that's the one we do.

But I don't wanna give the impression that that's the only way to do it. But you do have to do one of these three things. If you're not doing one of these three things, if instead a data set has come to you and you've been like, "Well, I think it's probably a pretty good database. And I-- I think it's reliable. I think it's trustworthy.

"I like the people who made it. They don't seem biased to me." Okay, well, those are not the right questions, right? I mean-- the question is not whether or not the people who did it are politically unbiased. I actually prefer databases from groups that are politically biased. Why? Because they fill a niche.

They go out and they really know their community 'cause they're really focused on it. And they'll bring us data. And if there's another group that's biased in some other way, now-- now we've got some real traction. 'Kay? So political bias, not a problem. Lies are a problem. There's ways to-- to work with that actually.

But you have to do one of these three things. You have to do a census, a random sample, or-- or posterior modeling. Because if you don't, here's what happens. You

think you've got a trend that you found in your database. And what you're assuming by looking only at the raw data-- sometimes people who use raw data say, "Well, I don't wanna make assumptions."

Well, (UNINTEL). If you're going to interpret that trend as meaningful in some way, what you are actually assuming is that the proportion of the world that you captured in the first point is exactly the same as the proportion of the world you captured in the second point, which is exactly as the same as the proportion of the world you captured in the third point, and so forth.

You are assuming that the capture probability or the coverage rate is identical, is uniform across all the points that you're going to draw on your graph. And when you say it like that, people are like, "Well, I know that's not true." Well, if it's not true, then the graph is actually just telling you about the reporting process.

It's not telling you about the underlying violence that you think you're measuring. And that's a big problem. And the reason it's such a big problem and the reason I'm gonna keep thumping on the table on this stuff is because we have to be right. Okay? There are enormous things at stake here, 'kay?

And I use this slide because this story is the one that-- is the one that comes back to me so often when I think about how important it is to be right. This man's name is Edgar Fernando Garcia or was. And he was a student and labor organizer in Guatemala in the 1970's and early 1980's.

In February of 1984, he left his office and didn't come home. People-- some-- some people reported later that they saw men in civilian clothes put him in a car, an unmarked civilian car, and drive away. But that was-- that was about all we knew. So-- and we knew even that little bit because his wife spent her life trying to find the information on-- about what had happened to her husband.

She did every kind of legal appeal available in Guatemala to get information from the government. They were like-- denials, denials. But in 2006, the historical archives of the national police became available. And in the archives, there were documents which indicated pretty clearly that Mr. Garcia had been captured by the police in a s-- in a specific campaign.

And in the argument for that trial of these two men, they said, "Well, you know, hang on. We're just following orders." And the judge said, "Okay. Okay. Guilty. That's not a defense. But if you're just following orders," she said to the prosecutor, "You should go prosecute their boss, Colonel Hector Bol de la Cruz."

Well, the thing was the documents used to show that Mr. Garcia had been disappeared by the police had a lotta flaws. A lotta issues with those documents. Some were missing dates. Some were missing the stamps that the documents in the archives have showing that the document had physically as a piece of paper passed through different offices.

Others were-- were torn. Others were water stained, 'kay? So one of the questions the defense raised is: "How can you trust these documents? They-- they have gone

through and found the crappiest, most illegible documents from the archive. These documents are trash."

We showed through statistical analysis by sampling documents from throughout the archive that in fact the flaws in the documents used in this case were exactly consistent with the statistical pattern of flaws found in all the documents in the archive. We showed further that the pattern of documents used in the case was completely consistent with the larger pattern of bureaucratic flow of documents through the police archive itself, through the police bureaucracy.

We showed that this was a completely standard bureaucratic operation. And that was one piece of evidence used to convict Colonel Bol de la Cruz of the disappearance of Mr. Garcia. This matters. This stuff matters. And it matters that we're right because of the victims. This woman here is that little girl in her mother's arms up there celebrating the conviction of her father's murderers.

I think we have a duty to the victims to be right. And that means that we have to bring to bear on each of our technical s-- our technical areas the best science and technology that is available to us. And so I-- I hope you are convinced that we should not use raw data but rather we should hire statisticians (LAUGHTER) when it is time to make a really important argument. Thank you very much. (APPLAUSE)

## ELIZABETH EAGEN:

Okay. So-- if anyone-- if anyone needs to refuel—thinking about all this algebra, please go ahead. But we're gonna go ahead and use this mic in the center here. And I encourage you to come to the mic, ask your question. We have the room for about an hour. So we can go as long as we need in that time. Please go ahead. Don't be shy.

## MALE AUDIENCE MEMBER:

I'll start us off. Thanks, Patrick. That was wonderful. You had-- the X and Y axis graph. And you had a list of sources. And I just wonder if you could talk a little bit about different sources. I'm also curious whether or not crowdsourcing is an interesting thing for you. I mean, I know you've had a debate with Ushahidi about that a little bit. So truth commissions, U.N. investigations. Oh, crowdsourcing's up there. Okay.

## PATRICK BALL:

Oh yeah.

## MALE AUDIENCE MEMBER:

And you--

## PATRICK BALL:

You remember when I said that big data amplifies selection bias?

## MALE AUDIENCE MEMBER:

Yeah.

## PATRICK BALL:

That's what I was talking about. Yeah, that's-- that's really the-- the most severe version of this actually.

## MALE AUDIENCE MEMBER:

Well, I'd be curious to know if you think-- is there-- a couple different questions. One is about what-- you know, s-- a little bit of-- kind of a ranking of different kinds of sources for the purposes of your work. You know-- are the-- do you find NGO documentation to be really good?

I mean, what-- you know, what are the kinds of ranking you would do-- for the work that you're doing? But then there's a whole separate of work which has to do with narrative, testimony, trying to get a whole-- I mean, you know, stuff that isn't necessarily about the same kind of work that you're doing, which is--

## PATRICK BALL:

Absolutely--

## MALE AUDIENCE MEMBER:

--what some of these are trying to do. And then a separate question is about crowdsourcing. Is there a role for crowdsourcing in this? Is there a way to do any kinda verification around crowdsourced data that comes to you? Is there-- is there something there that's interesting in the future? Not-- as well as a critique of what's going on with crowdsourcing now.

## PATRICK BALL:

So crowdsourcing is the kind of principle case of the problem of amplifying selection bias, 'kay? And for statistical purposes, no, I don't think it's fixable, okay? But it's no worse in some sense than these other sources are in terms of producing a list of victims. The data that I'm looking for is a list of victims.

So if the crowdsourcing project is going to text in via SMS the names and maybe through the timestamp we have to assume the dates and locations of the deaths or if we're going to harvest that kind of information from YouTube videos, for example, another kind of crowdsourcing-- that's fine.

I mean, that's gonna create a list that for me is the same kind of input as these other lists. Because what I'm looking for in data is just a list. So I know that for people who are interested in case analysis, we have-- an elaborate kind of discussion around-- around corroboration, around verification, around the richness of the case.

But I'm looking for a couple of names, a date, and a location. So the only thing that I'm worried about in data is that somebody's just made it up, completely made it up. And in as much as I can figure out that it's not made up, that data source is fine with me. And I'm delighted to use it. Okay?

So I think all of these are really great sources as inputs to a-- to a s-- a process by which we adjust for what we don't know. None of them is adequate as a source by itself. Okay? And that's the key q-- the key thing. And my critique of Ushahidi is that they just flow those texts into dots of variable sizes on maps.

And the inference that we are to draw from those do-- from those maps is that this dot is big and this dot is small. And there's no dot at all over here. Well, maybe that's because everyone over here knew about Ushahidi and the people over here didn't know about Ushahidi. Maybe it's because the people over here are texting to some other service and they don't like Ushahidi.

Maybe it's because. Maybe it's because. Maybe it's because. And once you start the "Maybe it's because" path, yeah, you're-- you're kinda done, you know? There is no way we are going to get back to a rigorous statistical analysis because we're stuck speculating about, "Maybe it's because."

So we need some sort of rigorous process. And by rigorous, I mean some kind of math that takes us from a series of inputs to an estimate. And there are-- as I said, there are a variety of approaches. I've presented one tonight. I always like NGO documentation. But that's more about my heart than about my-- my work.

It's just because I love the way NGOs engage with people. And I think that NGO documentation creates all sorts of secondary value like narrative, like closure for a lotta the victims, like a sense that something can be done. So it's not that I prefer NGO documentation because I think it's statistically better but because I think it produces a lot of other secondary value.

And I worry in particular about crowdsourcing because I think it's so-- well, it's not

just easily misused. It's generally misused in the first instance by comparing dots on a map. And then we're stuck in this debate about, "Well, we have something. Look, we have a map with dots on it."

And I'm like, "Well, it's not really reliable. We don't really know if any of that's true. In fact, everything we know suggests it's not true." Go, "But it's-- got dots. And it's a map." (LAUGHTER) And the conversation kinda dies there. And it has died there for years, as I think you know, which is why you set me up with that question. (LAUGHTER) So-- yeah, I think-- and is that that all the--

## MALE AUDIENCE MEMBER:

Yeah.

## PATRICK BALL:

Yeah?

## JESSE:

Hi. Thanks a lot. Fascinating-- presentation. Thank you. Very energetic, too. Really appreciated that. (LAUGHTER) I'm Jesse (PH). I'm from Global Youth Connect. And we run human rights programs in-- post-violence countries including Bosnia and Rwanda. I was fascinated by your statistic about Bosnia.

And I'll take a look at that. I-- I don't know about other people in the room. But I-- I think it's pretty common that there's a big controversy about the data related to Rwanda. And I hadn't heard that in your list. And I was curious if you have done any reflection on that data. I know there's a lot of-- question about whether or not the government of Rwanda or other people who are saying that there were a million Tutsis-- Tutsis killed are actually covering up many other deaths or whether that-- counter-argument is actually hogwash.

And-- the BBC came out with a documentary called *The Untold Story* which brought out a lot of these statistics. And then Rwanda shut down the BBC again and, you know, said this is hogwash. So I'm just curious if you could point me in the right direction. I will look forward to reading and watching some of these things with-- your-- presentation in mind.

## PATRICK BALL:

So two political scientists, Al Stam and Christian Davenport, who were cited extensively in the BBC piece, you could read their work. The Rwandan government calls them holocaust deniers. That seem like a misreading of their work to me. It

seems like they're reading-- they're-- they're-- they're not denying holocaust.

They're s-- or-- or genocide. They're saying that there was an-- a very complicated pattern of-- of killing of civilians by all the armed groups but asymmetrically. So I'll leave it there. I'll leave you to look up Al Stam and Christian Davenport are their names.

And if you Google Christian, you'll find that he and I have been co-authors many times. And so perhaps I have a little bit of connection to that (LAUGHTER) work. But I haven't actually done any math with them or really looked at their data very much. So I can't answer much more than that. So--

## PETER XAVIER:

Hi. My name is Peter Xavier (PH). I'm a 1L at N.Y.U. I am-- I come from the business t-- and tech world. And so I'm a bit dismayed, very dismayed to see-- or to hear about the-- lack of rigorous mathematical and statistical-- knowledge within human-- human rights organizations. At least that's how it seems from the presentation.

So I'm curious. If you were to suggest-- ways for someone like myself to tech up in statistics, algebra-- I took stats-- years ago, did just well enough to get through it. I don't pretend to be, you know, a professional in-- in that field. But I-- I-- I do wanna kinda build those skills. And so if you were looking at-- policy, law students-- what would you suggest that they-- or professionals, that they learn?

## PATRICK BALL:

Well, one thing is they could come be our intern. And so I'll give you (LAUGHTER) my card, 'kay?

## MALE AUDIENCE MEMBER:

I'm assuming these are unpaid internships? (LAUGHTER)

## PATRICK BALL:

Let me talk to my donor. (LAUGHTER) That's one thing. But I think it's-- it's very-- I-- I am not critical of the human rights community for not having very much of this expertise. It makes a lot of sense to me. Because the human rights community has largely been driven by-- by lawyers because it's primarily a legal field.

And-- and we've picked up a few other disciplines along the way. You know, we've gotten historians, and archivists, and forensic anthropologists, and now satellite imagery analysts, and journalists, and psychologists. And-- you know, and now we're

just starting-- we're picking up statisticians.

I don't think that this is a problem that is in any way unique to the human rights space. We see the craving for evidence-based or data-based argument throughout all of the world right now. All-- everybody is craving it. And I think that it's actually industry that is primarily to blame. Because the criticisms that I am raising here don't apply to most industry contexts.

If you're Amazon, you know every single click that every one of your customers makes. You know every minute they spend on every single page of every single product they look at. You know which ones they buy. You know which ones they don't buy, et cetera. Every little piece of the data is available to you.

And so your analytics can be as complicated and sophisticated as you want. And you'll be right. And that power has leaked out from industry and is affecting the rest of the world. Unfortunately, most of the rest of the world, we don't have perfect data.

In most of the rest of the world, we have in fact really problematic data, very challenging data even when it's produced by automated or semi-automated means like crowdsourcing, which is produced by hundreds or thousands of people sending text messages in to a central point, okay?

Furthermore, we've gotten very, very excited in the world about the notion of big data, okay? We think that-- we all have in our heads somehow that the error of-- of some kind of statistical calculation gets smaller the more data we use to make that calculation. And that isn't fundamentally wrong, okay?

At the base of statistics, we have the logic that if we chose the sample at random-- and we usually forget that part. But if we chose the sample at random, the larger the sample, the smaller the error. That's true. If we chose the sample at random. Remember that part? We didn't ever choose any of these samples at random.

And as a consequence, getting larger, and larger, and larger samples is amplifying the bias rather than ameliorating it. Because we're getting better, and better, and better at covering the places that are visible and coverable. And we're not really getting any better at all in the places that are aren't. And I-- I mean, I-- I'm asserting that. But I've shown three examples.

And I can point you to some articles where we've written, you know, another six or eight examples. We have tons of examples of this now in-- in-- in the human rights context. But I think it's true across the social space. Across almost all social statistics, the kinds of analyses that are done on-- statisticians call this-- all this data, we call this convenience data, which does not mean it's convenient to collect.

It means it's not randomly selected. And stuff that's not randomly selected I think all has this kind of problem that it is produced in some way, and there's a reason it got produced, and there's a reason other information did not get produced. Okay? And those reasons very often correlated with what we're trying to figure out. So not a unique problem to human rights by any means. It's just that's what gets me all head up. (LAUGHTER)

## ELIZABETH EAGEN:

So I'm gonna take moderator p-- sorry. One second. Take moderator privilege and inject some gender balance to our question and answer. I-- I wondered—following up on that question, I wonder if you could talk a little bit more about the kinds of barriers to getting greater understanding of your work.

I mean, clearly there's math. But also I feel like we have-- I feel like we have two worlds operating in parallel. We have a narrative world, a qualitative world, a legal world. And we have the work that you're trying to do. So I think talking through a little bit of-- whether it's about honing the right questions that can be answered by what you do or if there's some other angle you should be discussing. I wonder if you could sorta speak to that.

## PATRICK BALL:

Well, certainly honing the questions we're asking is-- if-- if we wanna build an integrated project in which we use information from lotsa different fields, let's kinda talk through what that would look like, right? So we're-- we're gonna start. You know, it's gonna be a big human rights project. Let's say it's a truth commission.

Or let's say it's-- it's-- it's a prosecution of a human rights case. We're gonna start with lawyers, right? That's where it's gonna start. Because they're gonna tell us what the legal framework is about why this is a human rights problem rather than some other kinda problem.

We probably wanna bring in other disciplines, too. And we'll bring the historians in. But we don't ask the lawyers about the history. And we don't ask the historians about the legal framework. We have to figure out, "Okay, how does the historical context situate the legal framework?"

And then we'll bring the forensic anthropologists. And we don't ask them about the legal framework either. We say, "Hey, what's the cause of death that you can infer from the pattern of human remains that you exhume?" So with each additional discipline, we have a challenge of figuring out what's the question this discipline can answer that substantiates the rest of the argument. 'Kay?

And statistics is no different. I think that when-- in-- in-- in an argument, in a larger argument of what we're doing with the statistics is simply trying to figure out the magnitude, we're not really using the statistics very-- in a very intelligent way. Magnitude's not an especially interesting question un-- you know, unless you're trying to get-- the media to write something about.

You know, but-- but other than that, if what you have is an analytic question, you need a comparison. If you're making an argument about genocide, for example, the question that you want to ask is: What is the m-- mortality rate of the group that's the target of genocide? And what is the mortality rate of the group that you think is not the target of genocide?

Those rates should be really different. Because if they're not, the statistics do not substantiate the argument that there is genocide. That doesn't mean there wasn't. I just means the statistics aren't going in that direction. And there may be good reasons for that. But that's something to think about.

But if there is a really big differential, then you can conclude that the statistical argument is consistent with the-- with the hypothesis there was genocide. And that was in a nutshell the argument that I presented in the prosecution of General Jose Efrain Rios Montt in Guatemala for genocide in 2013.

There was a giant differential in the mortality rates between the indigenous community and the non-indigenous community. In fact, if you were an indigenous person, the probability that you were killed the army was eight t-- times greater than the probability that your non-indigenous neighbor was killed by the army in the same period in the th-- same three counties where the case focused.

So that's differential. That's the comparison. And that's the kind of argument that we should be looking for in-- to use statistics in a larger human rights project. Like, what is-- what is the trend? What is the pattern? What is the comparison? What is the kind of things that we can look at next to each other?

I'm-- I'm getting more and more interested as we get deeper and deeper in our work on Syria to think about these patterns over time that relate to the movement of the conflict across-- across the country. As areas change control, as they've been doing constantly for the last four years, how do the killing rates change? Okay?

And that's especially interesting because the-- the-- the denominators here are getting smaller and smaller. People are leaving. People are moving out. Millions and millions of people have left Syria. And the number-- the-- the raw number of people being killed continues to be very, very high.

So in fact the rates are skyrocketing if you think about it in a p-- in proportional terms. So this is very interesting. And I think that what you're-- you-- you-- you need to bring in this contextual knowledge to be able to frame the question appropriately. Let me mis-- deliberately misunderstand the second part of your question, (LAUGHTER) if I may-- and say "How can we do this better?"

And I think it's by having more statisticians working in this field. That's hard because-- first of all, it's really hard because there's this giant sucking power of money that pulls statisticians into the private sector. (MAKES NOISE) (LAUGHTER) And the human rights community has a really hard time competing with that.

So that's one of the being challenges. The other challenge is that we actually do some fairly s-- s-- specific kinds of statistics. And statisticians are as-- one statistician is as different from another as a contract lawyer is from someone who prosecutes war crimes. Right?

So there-- these-- it's a big field. And there is a lot of different kinds of things people do. Our vision in the Human Rights Data Analysis Group-- for the future is to-- bring in people who have some-- ideally some graduate level preparation in math, and

statistics, and programming and to teach them to become first rate analysts and programmers so that they can process data and make estimates in what we think is a really responsible way to support the human rights community and then spin them out.

We don't intend to retain them. We don't wanna be the locus of this work. But rather, we wanna be a training network so that after we've worked with people who have been 1Ls and then interns with us for some time-- they go and work (LAUGHTER) in other human rights projects.

In international NGOs, in donors, in U.N. missions-- in tribunals, in whatever that project is so that th-- there's an in-house voice. My experience is that it's way easier for people to listen to an in-house statistician than it is to get advice from someone outside. So that's really the model that I think is gonna work.

## MARK LATONERO:

Hi. Mark Latonero on Data in Society Research Institute. Patrick, great presentation. I really appreciate your last comment on-- we need to be right. And I also wanna ask kind of how right we actually need to be. And what I mean by that is-- you know, how right do-- do we need to be in terms of-- a situation like real time decision making on a current human rights abuse?

For-- for instance, the c-- the crowdsourcing-- could be the crowdsourcing is-- is not to estimate all the acts of abuse, or-- or-- or killing, or homicide in an area? It could just be to report even one incident that someone with operational capacity is going to go to that incident and-- you know, act in real time? And so maybe you can shed some light on, you know, the decision making process as we move from sort of retrospective-- human rights abuses to sorta towards the real time.

## PATRICK BALL:

Sure. I think that's a really good question. I appreciate it, Mark. Let me start off by saying-- let me go back to the slide that we were looking at before with the list of all the different kinds of information. Or not all but many kinds of information about-- human rights violence that I've looked at over the years.

And there's a lotta different kinds of data there. All of that data is interesting if our job is case analysis. If what we wanna know is more ab-- about the existence of a case, "Oh my gosh. There is a killing here. We need to deal," or we wanna learn more about the case, we can use any information we like.

We don't have a concern about representivity, which in, you know, one word is fundamentally the concern I'm raising here. We don't have that concern. We have concern about existence, the existence of knowledge about that specific thing. So if what we wanna know is, "Is there a homicide?" any of these methods will get us there. Okay?

If what we wanna know is, "Where's the hot spot of homicides where we deploy resources?" all of these will fail egregiously. That is a terrible use of these techniques. And I think that crowdsourcing has been-- very, very problematic in this real time decision process that you proposed.

And I'd be happy to-- maybe afterwards we can talk through a couple of examples of-- of cases where I think that's been very, very bad. But I'll just flag first off Haiti. That was a very, very problematic use. So I'm-- and we-- I can talk through some of our analysis on that if you'd like later. But if you're looking for case data, have at it. Absolutely. Use everything you can get.

## ENRICO BERTINI:

Hi. Thanks a lot for your talk. It's really fascinating. I am Enrico Bertini. I am a assistant professor from N.Y.U. Polytechnic School of Engineering. I don't know much about-- human rights. But I work with Meg, (LAUGHTER) who you-- you know. And-- and we are trying to work together on some of these things.

I have a question maybe a little bit (UNINTEL). I don't know. So I totally agree with you that when we-- so there is a problem when you communicate information using only exclusively raw data. And raw data is clearly biased, right? But maybe there is another side of the coin.

So whenever you have an estimate, you also have uncertainty on these estimate. And people und-- as-- as well as people don't understand bias very well, I guess people don't uncertainty very well. And I'm just curious to hear from you what's your-- what's your opinion on that.

## PATRICK BALL:

It's a gigantic problem. (LAUGHTER) You're totally right. This is huge. Estimates do have uncertainty. So what-- what-- what that means is that-- if you think about when you've seen-- a survey result, like a public opinion survey, and someone says, "Well, 40% of people agree with this policy. And there's a margin of error of plus or minus 3%." Okay?

What's being said there is that if we were to repeat the data collection 100 times, then in 95% of those recollections the estimate of people who agree with this policy would fall between-- 37 and 43%. Well, that's a little cumbersome. And it's really hard to get your head around. Fortunately, th-- new statistical approaches allow us to make much, much better statements of uncertainty.

And we can talk about that a little bit. And we're just-- we're-- our new-- our new round of estimates are all fully Bayesian. So we're-- we've made that move now. But this is, I think, our last ever (UNINTEL) set of estimates. And so we present our graphs as much as we can with error whiskers on them. And then we talk in the narrative-- about what the error means, okay, around the estimate.

Now, is that adequate? It's hard to say. In other contexts, I don't think that caveats America useful. So when we sometimes say, "But isn't it true that all we have to do to get out of all this-- I agree with you, Patrick. Everything you've said is certainly the case. But what if we just said that these are the statistics according to testimonies given to the truth commission?"

Well, that's what I did for all through the '90s. All of our statistical arguments s-- have that caveat. But what we found is that nobody remembers that caveat later. Even us. Okay? And I saw a couple weeks ago a presentation by-- a big shot big data guy who said the same thing at the beginning of his talk.

But by the end of his talk-- it was a 10-minute talk. By the end of his talk, he was generalizing from his conclusion to the whole world. Nobody can remember those caveats. So the caveats don't work at all. So I'm not actually sure if my interpretation of error works all that well.

Because I don't know that people carry it forward either. But by saying it's an estimate, we kind of constantly evoke the importance of the measurement of uncertainty. For people-- who are not full-on stats geeks here, many statisticians consider statistics to be the study of uncertainty, okay? That that's how many statisticians understand our job.

Because anyone can come up with an estimate. The rigorous part is putting the error bounds around it. The rigorous mathematical part that requires all this elaborate mathematics is saying how uncertain that estimate is. And some of these estimates are very uncertain. And the uncertainty may be asymmetric.

And these are very asymmetric kinds of uncertainty. We're much less certain about how high the estimate might be than how low it is. We know it's almost c-- certainly not lower than that. But wow. It could be way up there. You can see how asymmetric they are. That's-- that's typical f-- in this kind of statistical approach.

So big problem. Don't know that I have great-- approaches. Now that we're doing Bayesian stuff, we have a lot, lot, lot more capacity to express uncertainty in graphical form because--

## ENRICO BERTINI:

Thank--

## PATRICK BALL:

--posterior distribution is much easier to represent.

## ENRICO BERTINI:

Thank you.

## PATRICK BALL:

So--

## ELANA BEISER:

Hi. My name is Elana Beiser. I'm from the Committee to Protect Journalists. And I wanted to ask you-- most of your talk has focused on killings, murders, homicides, very black and white-- someone's either dead or they're not dead. At CPJ, we track attacks, imprisonments, killings of-- of journalists.

But we're very focused on the motivation. So we might know that someone was killed. But what we're trying to determine is: Were they killed specifically because of their journalistic work? Or were they just in the wrong place at the wrong time and (UNINTEL PHRASE)? And I wonder if you can address that kind of subjective analysis at all-- its unique problems, or how it fits into--

## PATRICK BALL:

Well, it's-- it's not unique, okay? You'll notice I didn't say anything about-- when I was talking about police homicides, I didn't say, "J-- these are unjustified," or, "These are justified," "These are—according to the-- each department's procedures," or they're not. I-- I-- I-- when we look at Syria, I didn't say, "These are combatants. These are civilians."

I didn't say that. And the reason is that all of those k-- kinds of determinations are very difficult to make at scale. You can make them-- for each case and maybe even we could get what a statistician would call inter-rater reliability. We-- we-- we could get agreement among five or six people who independently decide on each case reading the available material whether or not this is an A or a B case, whether it's not killed because he's a journalist or killed because-- for some other reason.

You know, because someone wants to steal his car. That's-- that's-- that-- th-- even if we got that l-- high level of inter-subj-- subjective agreement, it's not clear that that agreement is the same as the agreement we would get in Guatemala, or in Nigeria, or in South Africa.

I-- I mean, it's-- it's just-- it's too complicated. And when we come to legal determinations about someone's civilian or combatant status, for example, no way do I believe that those decisions made hundreds of thousands at a time are reliable and consistent across that whole process.

So we stick to things that are accountable precisely because we want to be able to make statements that we can defend. Even though those statements are not as interesting or as useful as statements would be if we could qualify them with, "These are people killed because they were journalists." That would be a much more interesting and useful statement. But we don't feel comfortable making it. It's just

too hard. And-- and it's too-- too mushy.

## ELANA BEISER:

Can I ask a follow-up question?

## PATRICK BALL:

Sure.

## ELANA BEISER:

So one of the things that we try to do as C.P.J. is focus on emblematic cases-- you know, and push for, for example, prosecutions in-- emblematic where-- (UNINTEL PHRASE) in-- in Russia, that kind of thing. You--

## PATRICK BALL:

Absolutely.

## ELANA BEISER:

Any thought on--

## PATRICK BALL:

That's great.

## ELANA BEISER:

I mean-- it's a safe--

(OVERTALK)

## ELANA BEISER:

--thing to do, I guess--

(OVERTALK)

## PATRICK BALL:

Absolutely safe. And, look, I want nothing I've said tonight to be construed as a criticism of qualitative reasoning. I think that qualitative reasoning gets us out of many of these traps. Because we rest the legitimacy of our conclusion on the reputation of the researcher. We say, "Oh wow. This is-- a report from someone that I think really knows a ton about this field.

"She knows the economic history. She knows the military history. She knows all these different people involved. She knows what's going on here. So she can paper over the people she didn't talk to with judgment, with knowledge that comes from other contextual areas." Yeah, contextual knowledge, not so easy to include in statistics.

And before the-- before very recently, we couldn't include it at all. Now, we can include it in tiny, little, very, very specific narrow ways. But even so, not-- not really. That's not really the job of statistics to do. So I think that qualitative research is what I advocate in place of this. We need much more qualitative research that is self-consciously qualitative and less bad statistics. Bad statistics are worse than no statistics. And that's something we have to convince even the Security Council, that thinks it can create statistics by resolution.

## MINKY WORDEN:

Hi. Minky Worden from Human Rights Watch. And not a stats geek or whatever it was that you just said. (LAUGHTER) But we at Human Rights Watch have been longtime believers in your message. And-- I think you have an important-- point that is being heard. But let me-- ask you on the other side.

With everything that-- with all of the caveats, and cautions, and the scare that you put into human rights groups who feel the urgent need for a graphic of some kind to make their human rights report sail out into the world and find a home on the front pages of the world's newspapers, how-- how can you empower people to use statistics without fear? And what t-- tools are you and HRDAG or-- whatever your acronym is, what tools are you bringing to the table to help us do what we need to do there?

## PATRICK BALL:

I'm not really bringing tools in a software or mathematical sense. But what we are bringing is people that we can place in your organizations to do it with you, which we've done for you. (LAUGHTER) So-- so--

## MINKY WORDEN:

But how do we clone you? I mean--

## PATRICK BALL:

Pardon me.

## MINKY WORDEN:

How do we clone you?

## ELIZABETH EAGEN:

There need to be, you know--

## PATRICK BALL:

We're-- we're working on it. I mean, we bring in (LAUGHTER) interns. We train people. And then we wanna spin them out. That's-- that's what I think the answer is. The answer is not mechanical. It is not-- the statistics is not technology. This is-- science is not technology. These are often conflated.

They are completely different. You cannot build a piece of technology that will do this in-- in a trivial way so that people can just pour the data in and get an answer. That's not going to happen. And if people promise it to you, the name--

(OVERTALK)

## PATRICK BALL:

--for what they're offering is snake oil. (LAUGHTER) Okay? And they may not know that themselves. But that's very, very dangerous. And I-- I-- I just will put in this. I'm concerned about the entire field that's called data science now because it strikes me that data science is a field of very skillful programmers who are not at all worried about probability.

It is a field that does nev-- just never uses the word sampling. And if you don't know how the data was sampled from the population, your visualizations, no matter how beautiful, are misleading. So I give these scares. The tool I offer is, "Hey, send me someone with some basic-- programming, math, and statistics preparation." And Megan, and Christiane (PH), and I will train her or him into a first rate analyst for you.

## MINKY WORDEN:

Thank you.

## FEMALE AUDIENCE MEMBER:

I mean, aren't you suggesting organizational behavior change? Like, really--

## PATRICK BALL:

Yes--

## FEMALE AUDIENCE MEMBER:

--what you're saying is that there is a choice to be made in allocation of scare resources in a human rights organization to pay for a statistician and not another lawyer.

## FEMALE AUDIENCE MEMBER:

Yeah.

## PATRICK BALL:

Either that or-- (LAUGHTER) either that or don't use statistics, right? I mean, that's the choice. I mean, listen, here's-- here's a metaphor, people. And I'm sorry. I don't mean to b-- here's a metaphor, okay? Let's say that in your village you find a mass grave. Do you say, "Hey, I'm gonna call all my friends to come over with shovels and we're just gonna dig it all up. We'll get it done in an afternoon?"

No, you call the forensic anthropologist, right? You call the people who studied the-- h-- how you understand-- the-- the signs of violence on human remains in order to exhume the data in a way that preserves its evidentiary value. If you want to submit a case to your country's supreme court, you don't s-- you don't say to yourself, "You know, my daughter's a really good writer.

"You know, and she's just finished eighth grade. And she can write like crazy. So I'm just gonna have her write it all up." No, you hire a lawyer. You hire the best lawyer you can find. Why do we think statistics is easier somehow or more mechanical than forensic anthropology or the law?

Why do we think that? Is it because statistics is so trivially easy? I don't think so. So it's not. If groups wanna do statistics, they've gotta statisticians. They've gotta hire really serious ones and ideally really statisticians that are very, very sensitive to exactly the kinds of problems that we have in human rights data.

## MALE AUDIENCE MEMBER:

I think it's a very interesting presentation. And the work you do is very good in the human rights field. Now, my interest is a little bit-- how do we-- first, people make their decisions based upon evolutionary biology, psychology, emotion. And too many people in this country reject science outright.

And many more wanna cherry pick science to favor whatever views they favor-- already. So how do we help convince people what you're doing is good for what they're doing? And, you know, as far as the scope of hu-- of-- social problems we have, whether it's denial of global warming, or what's going on in the world, or-- or my comment earlier about-- the police-- just police alone.

Until you are the victim of police misconduct, the police are there to protect you. And you fall for their-- union's propaganda about how danger it-- it is and all the cops that are killed by violent felons. But it seems to me what we need and that also we have the ability in this country with FOIA-- Freedom of Inf-- Freedom of Information laws.

And now, everybody has a iPhone. So there is the data that's available at least to the N.S.A. How do we convince or look at the data to see, you know, how many people are killed by cops? How many people threaten the cops and kill the cops? And-- and give them information where they can say, "You know what? Looks like cops are acting in-- in general-- it's too many homicides vis-à-vis what's going on." So I guess my point is: How do you sell what you're doing to the population and maybe get the population to see there's value in what you're doing to all sorts of different things, police misconduct being one?

## PATRICK BALL:

I'm a simple country statistician. (LAUGHTER) I don't know. I mean, we very self-consciously position ourselves as non-advocates. And I'm just not really very good at advocacy. I don't know how to-- I don't know how to do that. I don't know how to advocate on behalf of science. I would leave that to my very competent friends at the American Association for the Advancement of Science.

I think they do a really good job. Maybe not good enough. But I think they-- they labor mightily in this area. I teach a little kids summer camp for my friends' kids where we do little scientific experiments that lead to statistical understandings for nine-year-olds. You know, and I-- it's a blast. And what I find is when we do these little experiments, the kids are s-- kinda sold on the scientific method.

And so I think good scientific training is really super valuable. But I don't know how to scale that to, you know, hundreds of millions of people. I'm not-- but I-- I know how to do it, you know, for 12 kids or something. Honestly, I'm sorry. I'm j--

## MALE AUDIENCE MEMBER:

Well, may-- let me-- let me-- let me clarify. People are more afraid of a snake or a spider than of a car although scientifically or statistically you're much greater at risk of dying in an automobile accident than from a s-- a spider or a snake. So how do we as people or you, people who are into numbers, help convince people you've gotta be worried about the car, not the snake?

## PATRICK BALL:

Yeah. People are way more worried about terrorists than they are about police, for example. And the difference is about a factor of ten in your probability of being killed. So-- look, I don't-- I don't-- I-- I don't have an answer to that either. I'm really sorry. I-- I'm just doing this thing, man. (LAUGHTER) (UNINTEL PHRASE). This is what I got. That's it. Sorry. I mean-- but, I mean, I-- I-- it's a really important question. I just don't know the answer.

## MALE AUDIENCE MEMBER:

Hi. I'm a data sciences student at Columbia University.

## PATRICK BALL:

Right on, man. Tell me about your probability training.

## MALE AUDIENCE MEMBER:

Quite a bit actually. I can't say--

## PATRICK BALL:

Really?

## MALE AUDIENCE MEMBER:

--I did very well in the class. But they give us at least 30--

(OVERTALK)

## PATRICK BALL:

You're worrying about sampling a lot?

## MALE AUDIENCE MEMBER:

Very much so--

## PATRICK BALL:

Sampling bias?

## MALE AUDIENCE MEMBER:

Yes.

## PATRICK BALL:

Make estimates and adjust the data before you do a visualization?

## MALE AUDIENCE MEMBER:

I can't speak specifically to the techniques. I've only done the introductory course so far. (LAUGHTER) But there's two more courses to go.

## PATRICK BALL:

Good. Well, I'm so glad to hear that. If only they did that at Cal. (LAUGHTER)

## MALE AUDIENCE MEMBER:

So-- my question is: Do you see any application of machine learning techniques to-- the work that you do?

## PATRICK BALL:

Yeah, tons. Tons. We do a lot of machine learning. But I have all kinds of worries about it, too. My worries about machine learning. So when we do that, that M in the middle, if we have 300, 350,000 records, that's primarily a machine learning problem. So d-- duplication database duplication is our primary application in machine

learning.

So I do a lot of programming in that. I like machine learning a lot. I'm really, really worried about it in a lotta spaces. In particular around the use of machine learning for policing. Because the models we're training we're training with data that is information known to the police.

So what we're doing is training the models to do exactly what the police did in the last round, which isn't all that useful but isn't all that harmful by itself. The problem is that we think we're predicting crime. We're doing no such thing. We're predicting policing in a very literal way.

Because it's a circular process by which crime known to police is used to train the data, train the models, which then tell the police what to do. So the police go and do what the model says. And, by the way, they may look at model, be like, "Hey, that's great. That's just what we're gonna do," like that's a good thing. And then circularity. So I'm delighted to hear that there is probability in that. I assure you that is not what most data science programs do. Yeah. So-- but you go to a good one. Yay alma mater. I went to Columbia. (LAUGHTER)

## MORRIS:

Hi. My name is Morris (PH). I'm actually also from Human Rights Watch. And I have a bit of a technical question. You mentioned that-- one of your biggest problems-- is lies in your lists. Is there a way you normalize that-- through statistical analysis? Or do you have to go through manually to sort that kind of thing about.

## PATRICK BALL:

Well, I wouldn't know a lie from a truth. I mean, Mohammed A, Mohammed B, both killed in Hama in December 2013. Maybe one's a lie and maybe one's not. I don't know. What I do know though-- what I do know is that if I'm asking a substantive question like, "Was there a giant rise in killings in-- in January of 2013 in Hama?" that what I can do is make different kinds of assumptions about the structure and pattern of those lies, and delete the records systematically, and re-estimate, and thereby simulate the impact of the lies.

What if I drop 2,000 records from the first-- from data source one? And I only drop the records that only appear in data source one, right? So they're like this. They're like the left side of the circle there. They only appear in data source one but not in data source two.

And I say, "Well, I think data source one may be telling us porky pies." So we just drop a whole lotta records from that and we see, "Does that change the output in a way that is substantively meaningful? Does that give us a different answer?" If it does, then we conclude that our result is sensitive to the-- the possibility that there are a lot of lies in data source one.

We have a very specific conclusion. Now, we don't-- again, we still don't know if there are lies or not in data source one. We're gonna have to rely on some external source of information to figure out what the rate of lies might be. All data has some kind of lies in it. Most of them are inadvertent.

Most of them are errors, not lies. But errors and lies for us are kinda the same thing. If we report a man as dead-- in a conflict but really he died when a building fell on him and he-- you know, or he had a heart attack, well, maybe he's not really dead in the conflict. But he's in our list.

So what's the impact of his presence on our result? And another way to think about that is: What if he's not dead at all? He's actually in a refugee camp in Lebanon. But everybody thought he was dead. And they didn't dig up the building 'cause they weren't running away from more fighting. 'Kay?

Totally reasonable. But actually he got out and he's in Lebanon. Well, it turns out one of the things that I so admire about these amazing, amazing human rights groups in Syria is that they keep correcting these databases. There is an elaborate and very, very, very detailed process of revision.

And we know this because we keep getting updates from them. And the data just keeps changing. So we look-- we are like, "That record's gone now. Where'd it go?" You know, and we-- we now know that they're editing some out. They're changing some of the fields. And so there's this constant churn of records, which is exactly what we would hope for if the groups are updating their-- their-- their databases-- as they get new information. So I think-- I think the Syrian groups are absolutely terrific on this score.

## MORRIS:

Do you think that's unique to human rights-related data, that-- there's such a in-- inaccuracy in a lot of collection and that it has to be updated that much as opposed to other more normal data?

## PATRICK BALL:

I mean, sure. I mean-- you're collecting data in the middle of a conflict. Bombs are falling. You're hungry. You don't have so much access to the area. Maybe the people who give you the report don't know the person who they're reporting about very well. There's plenty of reasons, plenty of completely legitimate reasons that the information would be inaccurate on the first draft and get better over time as people continue researching it.

Yes, every reason in the world I think maybe particularly for human rights data. But I can imagine other contexts, humanitarian catastrophes, for example, where your first draft of information is-- is-- is often pretty vague. Yeah. Yeah, we've got one and then--

## FEMALE AUDIENCE MEMBER:

I'll wait.

## PATRICK BALL:

One, and then two, and then Meg.

## MALE AUDIENCE MEMBER:

Say it into the microphone.

## AMELIA WOLF:

Hi. My name's Amelia Wolf with the Council on Foreign Relations. And I work mostly in conflict prevention, which I know is kind of a touchy topic with statistics. And it seems like a lot of this analysis will be useful for post-conflict or holding people accountable. So is there any applicability or might there be in the future for prevention, or I know forecasting, or anything like that?

## PATRICK BALL:

There are some-- there are some interesting forecasting projects, okay? You know the Correlates of War Project? You probably know. Yeah. I know Phil Schrodt. You know, we've-- we are on panels together and stuff. And, look, forecasting is an interesting and hard problem.

But it's also in curious ways-- and Phil has con-- it's Phil who's convinced me of this. It's actually easier than estimating magnitudes. Here's why. When you're doing a forecast, you've got a little variables at play, right? You're gonna say, "We're gonna predict the presence or absence of conflict in all the countries in the world or in some subset of countries next year."

And the way we're gonna do that is we know whether or not there was-- you know, we have a yes or no on conflict in all those countries for some large number of past years. 'Kay? So that's the dependent variable. In statistics logic, we're predicting the presence or absence of conflict.

And it turns out you do need to know that variable exactly. You need to know correctly, and completely, and exactly whether or not there was conflict, this-- whatever-- or whatever your dependent variable is, in order to make forecasts that aren't completely garbage. However, you have a lot of other variables in the model that are the things you're going to use to make the prediction.

You know, the change in GDP from last year, the unemployment rate-- the presence

or absence of conflicts in neighboring regions, the-- you know, the presence of some kind of ideological or religious-- you know, extremism. Et cetera, et cetera, et cetera. You're gonna have a whole set of these-- these variables.

And I-- I know that the Correlates of War people have gone through literally hundreds of variables fishing to find the ones that best predict the-- the outcome, the dependent variable. Turns out doesn't matter how bad the-- the independent variables are. The independent variables can be completely crap.

They can be totally biased. They can be actually meaningless or uninterpretable. It doesn't matter. If they help you make a good prediction, you're good to go. 'Kay? Now, that's really, really interesting to me. That's really, really interesting to me. The second observation I would make about conflict forecasting is that it's not actually very good. All right?

The best models-- really only can predict the cases that are very obvious. And the obvious case is no conflict next year. Because no matter what subset of countries you look at, the overwhelming majority of them are going to have no conflict in the subsequent period.

So your prediction-- if you just say, "No conflict," you're right, like, 80% of the time. Right? And that's not a very interesting model. But you get a really, really impressive score having no independent variables. Right? So if you add a few independent variables, you can get a little better.

But honestly, the best models don't do that much better from obvious in my experience. You know, there-- and we really-- we don't really know how this stuff works. These guys have proclaimed victory. And Phil and I kinda fought about this at a conference in Wisconsin last November.

They were like, "We got it. Look at this. We're gettin'-- we're-- we can account for, like, 83% of the variability in conflict next year." And I'm like, "Yeah, because 80% of it's no," you know? (LAUGHTER) It's like you s-- it was easy, right? You got all the easy cases. And so you bulked up your success rate.

But you can't get past that threshold. You're stuck there. And I think you're stuck there because it's contextual. It's not structural. I just don't believe it. So I don't really believe that-- that we're ever gonna get early warning or conflict prediction. I just think, you know, we've-- we've got, what, 20 years of work on this stuff now. And we're really not doing any better than we did, like, after the first two years. So dead in the water.

## ERNIE DRUCKER:

Hi. My name's Ernie Drucker. I'm-- a public health epidemiologist that works in criminal justice now in the anthropology department at John Jay College (UNINTEL PHRASE). And one of the things we've been very concerned about is the police killings. Every-- you know, the students think about this.

The faculty, the average person thinks about it. And you ran very casually before through a set of numbers when I asked you about the number of police killings in the United States in a year. And if you go to-- and if you go to Google and ask that question, the first ten hits you get will give you three or four different estimates.

## PATRICK BALL:

Of police numbers?

## ERNIE DRUCKER:

Of-- of how many police--

## PATRICK BALL:

Of course--

## ERNIE DRUCKER:

--killings were there. Different-- different sources. And they--

## PATRICK BALL:

Absolutely.

## ERNIE DRUCKER:

And they all seem to be authoritative. Some of them are even governmental ranging from 456 to 1,104 for 2013 or '14. I forgot the year. That's a pretty big discrepancy. Almost-- two and a half times as big. And 1,000 means two-- it means 20 every week.

(OVERTALK)

## ERNIE DRUCKER:

And we've had eight of these stories that have dominated the news in the country for the last year. And correctly so. And the pictures on the newspaper of that windshield with all the bullets in it. And imagine what the pictures of the victims look like if they (UNINTEL PHRASE). But as they should be like (UNINTEL) was.

Because when you realize the level of the violence, and-- violence in the sh-- title of your session here. And that's what a lotta this is about. This is a level of violence that

people aren't used to thinking about. So when you say you have a method of estimating the unknown portion of that denominator of police killings-- and-- I mean, the big lesson-- I taught epidemiology at Columbia and-- was about the denominator.

That's what public health is about. What is that N? What is that parent population of the real number of people? You know, within 10 or 20% is good enough for most purposes of policy. But when you take that number up to 19,000 possibly for five years-- that was the estimate you had in your text there.

## PATRICK BALL:

No, no, no. 10,000 over eight years is the--

## ERNIE DRUCKER:

And-- oh, I'm sorry. It was eight years. Okay.

## PATRICK BALL:

Yeah, yeah. It's eight years.

## ERNIE DRUCKER:

But then, you know, you pick up on the fact that--

## PATRICK BALL:

Here we go--

(OVERTALK)

## ERNIE DRUCKER:

--that thir-- that-- that 30% of the possible cases that weren't--

(OVERTALK)

## ERNIE DRUCKER:

--they weren't surveilled. So now they are. When you add those stats in, you add another 30% on. And that figure of having30% of all-- homicides of someone not by someone you know, by strangers-- take the remainder and 30% of those might, you

know-- possibly be due to police action, that blows my mind.

And that's a very, very powerful fact, idea. That what you-- you talked about, not-- not-- you know, not just being a humble country statistician. But you're sitting on a very hot piece of information there that is very hard to refute actually once you walk through it the way you do.

And the facts support it without having to know the details. And public health as a science is based on not the best data in the world but having so much of it you can overcome its-- its-- its inadequacies and get the big picture. And the big picture is the level of violence being perpetrated by police on black men in the United States such that in every concentrated poor black population 95% of those men have been in prison at some point.

D.C.-- Washington, D.C., the six areas of New York City that fill the prisons here. So everyone knows about this in those communities. And they know the level of prevalence, if you will, of-- of violence-- police violence. Stop and frisk showed that very clearly-- even though it made the epidemiologists crazy. 'Cause it was-- you know, something was tiny and getting smaller all the time. And therefore people wanna say, "Oh, that means it's going away." No. It's being sampled differently. It's all sampling.

## PATRICK BALL:

Right.

## ERNIE DRUCKER:

Anyway, I-- I think you-- you have a very powerful tool there from a human rights point of view that-- (UNINTEL) your name and your organization's name. But to use that for advocacy is-- is a very-- is an important job that you-- that you're welcome to participate in. And I know you do. (LAUGHTER) You-- well, it doesn't cheapen your-- your findings or your methods. It really doesn't.

## PATRICK BALL:

Pardon me?

## ERNIE DRUCKER:

I-- I'm saying it d-- d-- you know, to take the-- you know, to take the evidence that you uncovered of-- of a gross underestimation of the level of police violence in this country-- is the answer to-- to the-- you know, what-- what's the retort of the police? "Oh, well, if you look, there are many, many more killings of black men by-- by other

black men than there are by police."

## PATRICK BALL:

I-- I don't know-- I don't know what they're gonna say.

## ERNIE DRUCKER:

This-- well, they give you those. It's-- it's lies and statistics. You know the-- the you know-- you know the book. So I-- I think there's a lot of room, you know? I-- I-- I think you-- you can ask for more of the lay audience about understanding scale, for example. This is five every week we're talking about. And you see what happens with eight of those cases in the paper--

(OVERTALK)

## PATRICK BALL:

--actually yes.

## ERNIE DRUCKER:

(UNINTEL PHRASE) more. (LAUGHTER) So I just wanted to be encouraging around that work--

## PATRICK BALL:

Thank you. Our-- our job--

(OVERTALK)

## ERNIE DRUCKER:

The technical part is very important. And it's very much appreciated.

## PATRICK BALL:

Our goal is to have other groups that are advocacy groups use this. So we work with-- I've spent the week in New York meeting with my advocacy partners, among my other meetings, explaining a lotta this stuff to them and saying, "Hey, why don't you use this in your work? Because that's-- you're good at it.

"You've got an audience. You know how to speak to that audience. You know how to

do this." I don't. And honestly, I don't have time and even really interest in learning those skills. I wanna do more machine learnin'. I wanna do more statistics. I wanna learn how to build really cool Bayesian models from-- from Shira (PH), you know?

I have a lot of things I wanna do. And I can't do that. We are three full-time professionals in my group. That's it. Okay? So we are really, really saturated. And I totally appreciate how important this stuff is. That's why we're doingit. But I can't do the other pieces. So we wanna hand it off.

We don't wanna compete with our partners. We wanna support them. We often say our-- in our best case, we're a footnote in somebody's report but we're a damn good footnote. (LAUGHTER) And it's a footnote that's right. It's gonna stick. They're never gonna be discredited on that footnote. But we wanna be a footnote in somebody else's report.

## MALE AUDIENCE MEMBER:

You can be a headline. (LAUGHTER)

## ELIZABETH EAGEN:

I think we can take one more question. Meg, were you on the list for--

## MEG:

No, it's okay.

## MALE AUDIENCE MEMBER:

Hi. I'm just a 90-year-old who never went to your camp. So if I wanna (LAUGHTER) say-- you know, to teach older people, you know, to know about statistics and how interesting-- do you have some kind of resource that we could go to? That's the first one. You know, to make-- statistic more interesting instead of, like, big textbooks and, you know, professor whatnot, right?

And then the second question is that-- how do we use statisticians? You know, how do we use your professional so we could, you know, have-- instead of, like-- you know, (UNINTEL) is correct-- we-- we could-- you know-- be able to-- you know, maximize the resource, and the time, and-- and without gettin'-- hogwash or without, you know, going to—barking up the wrong trees and so on like that?

So I would like to have some kind of resources that-- us laymen that-- like, we do, you know, or general public to say, "Hey, statistics is cool." Second is: Hey, you a statistician. I know how to use you. And I know, you know, ask you right questions or whatnot.

## PATRICK BALL:

Well, in this I'm very fortunate to be a member of the American Statistical Association. It does a huge amount of work specifically on those two questions, on the education of laypeople around statistical questions and on how can you learn more about statistics and get contact with a statistician for nonprofit projects in particular.

There's a network called Statisticians Without Borders that will do-- eng-- will engage with nonprofit projects on a voluntary basis to do deep statistical thinking about specific problems. Not specific to human rights. Specific to the range of issues that affect-- human wellbeing.

So tho-- that's-- those are the two key answers. If you want more-- if you want some links into that, you could just go their website, AmStat.org, A-M-S-T-A-T.O-R-G, and-- and-- and look for educational materials. They've got tons there. And they-- there is an enormous effort at the American Statistical Association to do more-- outreach to non-statisticians and explain how cool statistics are.

## MALE AUDIENCE MEMBER:

Okay. Thank you.

## PATRICK BALL:

Thanks. Thanks-- thanks everybody-- everybody very much for coming. (APPLAUSE)

* * *END OF TRANSCRIPT* * *