Okay. Let's take our seats please.

Here we go, really.

Okay. So, this brings us to the first of our segments on election and voting related work. As you know, there's been longstanding skepticism about respondents' reports of their turnout behavior, the idea that people think it's embarrassing to admit not voting, and therefore, claim to have voted when they didn't. Or maybe they usually vote and misremember accidentally that they voted this time and overlooked the fact that they uncharacteristically didn't vote this time.

And this has led researchers to an interest in replacing those self-reports with publicly available government records of that information about voters. And there is obviously a treasure trove of data available from Secretaries of State around the country. And, fortunately, we have with us I think the leading expert in this topic, Mike McDonald from George Mason.

He's quickly rose to visibility in correcting our understanding of what the actual government turnout rates were some years ago, and since then has committed himself to developing the leading sets of data and insights, I think, in this area, so I'm very happy that Mike is with us.

Thank you. So, yes, I'm a George Mason University, and also have an appointment at the Brookings Institution. I actually got my start in doing this sort of work after I graduated from Cal Tech some years ago and I worked at a political consulting firm located in Pasadena area that managed the voter registration file for the state of California, and also did redistricting work, as well, for state legislature.

So, I became immersed in doing surname matching, and figuring out all the errors in the voter registration file. Very early on, I, for the state party – Democratic party – of California, did, created I think probably one of the first micro targeting programs that state and local campaigns could use for their practices, and that was back in the mid 1990's and was deployed in a number of state legislative campaigns around the state of California.

I think I – who knows if I was one of the first people to do this, but – we, the first time I did it, we developed a list of supporters, and other responses that we've gleaned from the first time we did it and I said – and I just saved that all. And then the next time that I
saved the affidavit number on the voter registration record, and the next election cycle, I said, “Hey, we got all this data. Let's use it again and use it to mine out the supporters and volunteers.”

And it really gave those campaigns an advantage back in 1998. So, and also '98 did I think, probably one of the first early voting drives in the country, as well, through that system I programmed up.

So, I have a lot of experience with working with this data out in California, and then, since then, I've been collecting these voter files where publicly available for my own research purposes because I knew from my experience out in California that they could be informative.

And other people have been using them, too, for rather various purposes and, of course, I'm not the only one out there. So, I don't think I'm the expert. I think there are a lot of people who are good people out there doing work on this, and we're going to hear for some of them later today, as well, or tomorrow.

But, there's a lot of research out there on these voter files. So, I'm going to be talking about the voter registration files. I want to talk a little bit broadly about what people are doing just broadly with them, but also think about how they may be used for some of your purposes with the survey work that you're doing.

So, we're going to start off with voter registration files. And I'm also going to talk about the absentee ballot request files, too, because I think they're a new frontier that people haven't even really looked at yet.

So, what are they? These are the individual records that the states compile to track every voter. They're used for eligibility purposes in the general election, but also in primary elections where states have some form of closed or semi-closed primary.

And what's happened over time, and the watershed event in terms of federal legislation was in 2002 the federal government passed something called the “Help America Vote Act,” which requires all states to develop a statewide list of their voters.

Prior to then, some states had moved in that direction to modernize their election administration, but many states – and we even see this today, there's these difficulties with collecting these data – many states continue to have a very decentralized form of voter
registration. It was really the local jurisdictions, the counties, or
the townships, that were responsible for voter registration, and they
weren't sharing their data at all.

We're at a stage now where it's not just that these localities within a
state are sharing data, but now even states are sharing data and
developing regional and other even cross national databases of
voters that said they might be able to track people who are moving
from one state to another, and do some things in terms of perhaps
removing those people from the roles if they think it's appropriate
to do so.

So, there's a lot of work now that's being done with these files at
the administrative level to clean up these roles, to aggregate them
up and see what they can do with matching them with other
databases, as well.

So, what information is on these files? We've got individuals'
name, their birth date, their address – which may be associated
with a precinct, often, in the file itself that's publicly available – we
often have a registration date, or if we didn't have the date that a
person registered to vote, we may have the last time a change has
been made into the database, so a transaction record.

We know, in some – actually every state collects this information.
They may not share it publicly because of privacy concerns, but we
also know the method of registration. So, was this an individual
who registered at the Department of Motor Vehicles? Was it some
third party organization? Was it an online registration form? And
here's where the privacy concerns is going to come in.

In 1993, of course, we had Motor Voter and Motor Voter says that
public assistance offices also have to make these forms available,
and the states are tracking that information for compliance with
Motor Voter, and so there's concerns about someone requesting
public assistance. That information becoming publicly available,
but some states will still say whether or not a person registered at a
DMV office, for example, versus at another method.

In the states that have the closed or the semi-closed primaries, and
even some of them just have open primaries now, too, because
they have a change in law, but they keep the party registration
around. So we have party registration in some states. And then we
have vote history, and the quality of the state, or the scope of it
varies, again, on some states it's just the last party – the last
election.
Some states will have a long history of every state and local primary and general election for a decade or more within a data file. So, you get a wide variety on the sort of data that you get with vote history, but, of course, it's of interest to the people—you, in this room—because you can do vote validation with that.

And then, we get some limited demographic characteristics. Some states track gender and race, some and ethnicity. Race and ethnicity interestingly, that's an artifact of the time when we were—had—white only primaries. So, on states that have white only primaries tended to have that as a requirement on their voter registration, and they never really removed it either. So, you see some states continuing with this question and it's of interest, too, for some purposes.

And what are those purposes? Of course, we know vote validation. So, I don't think I need to go into that too much, but we can do it both at the individual level, and we can do it at the aggregate level. And Matt Berent who is not here yet, but hopefully either today or tomorrow will—oh, he is. Matt? Yes. Matt. Good to meet you finally in person.

So, Matt, of course, has been doing some of these vote validation studies for the American National Election Site. He's going to talk more about them. So, I don't want to step over too much of what Matt's going to be talking about there.

The other purpose that academics have been using these studies for quite a bit are voter mobilization studies. So, here's the Green and Gerber and their army of students they've now sent out upon the world to do these randomized experiments where there's a control group and a treatment group, and the treatment gets some sort of persuasion or mobilization message, and you can compare the effect versus relative to the control group of that message that is being delivered.

And so we've, of course, that's been a blossoming area within political science to do these sorts of things, and it's being implemented now by the campaigns, as well. So, it's one of these real world practical applications that the campaigns and the parties are really looking to academics to help them do these sorts of studies with.

Some of the things that I do, these voter registration files can be used to construct at the precinct level in states that have race—and
these, again, these were white primary states, so you're looking at places like Georgia, and South Carolina, and North Carolina, places that are covered under the voting rights act. So, when you're doing a racial block voting analysis, which is to determine the patterns of racial voting using aggregate statistics, well we can aggregate up.

We don't know how people voted individually, but we do have the precinct level election results, and we can aggregate up the race on the voter files and do ecological inference of a various forms analyses to assess out how people may have voted using other methods besides surveys.

And, of course, we can use exit polls and surveys, as well, and so it's another way to sort of cross validate what we're seeing. Do the ecological inference estimates that we're getting from these studies, if we're using voter files or if we're using census data at the precinct level, do they confirm with what we're seeing in the surveys, as well?

And, just for a second, I'll also say that, in the work that I was doing this last cycle in places like Georgia and South Carolina, doing these racial block voting analyses, there was an interesting pattern that I started to notice, which was there may be some sort of social desirability going on here, too, where on the census you get higher numbers of minorities than you get from the voter files.

And that shows up in turnout rates and other things. It's sort of a systematic thing that's observable when you're looking at the statistics that are generated by these racial block voting analyses, whether you do it using the voter files, or whether you're using the census data.

And so, some interesting things that, I think, need to be done there. And again Matt, and I know Josh will probably be talking about some of this probably tomorrow, as well, is that what information are we really getting from these administrative records? And what information are – how reliable is that information? And what sort of biases may be involved in these?

We really don't know yet. We're just at the very frontier now of starting to really seriously look at the administrative data. The data that we're getting from the consumer databases, and trying to understand what the reliability of that data so that if we're going to use it for various purposes, what is it actually telling us? So here's one instance of where I think there needs to be more study.
Other uses, but most recently in all the litigation that we had over the last year or so over photo identification laws within the states, people were using the administrative records and matching against, say, DMV records to get an estimate of the number of people who didn't have photo identification within states like Pennsylvania.

But, again, that's a very challenged work, especially with the matching, which is something that Matt will be talking about in a few minutes, as well.

And then we've got voter turnout, which is some of the studies I've done, as well, which is to look at the aggregate numbers that we can see in terms of race and age and that we can gather from the voter files and validate that against the current population survey and the exit polls. And when you sort of triangulate that way, you can see rather clearly that the exit polls had this age bias in them.

And, of course, Warren and Joe did their analyses back in 2004 also saying, “Look, we know we have an age bias, and we've got to do some things to mitigate that problem.” And they, to their credit, of course, made that very transparent about it and worked to solve those issues. So, here's another instance of where we can validate that yes, we do indeed see a problem, and it does need some sort of solution.

Other examples, and this is just sort of stuff that I've been doing, and other people have been doing, but we can aggregate people up by household, and that might be useful information. Again, it could be helpful for some of the things that we've been talking about here. If individuals are going out and looking at households, you can, when you look at the people at an address, you can see, well is that a group quarters of some sort? Are there a large number of people registered at that address? Is there a family situation?

And that's what I've been looking at. You can look at parent pairs and a child pair, registered at the same address, and you can actually see that, much like those earlier studies that people have been doing on the panel studies where, do the parents have the same – two Democrats, is their child a Democrat? You can actually see that with party registration, as well.

I've been looking at something called pre-registration, which allows young people to register to vote as young as 16. What effect does it have on turnout? In litigation I've been involved with
the timing of the voter registration, because, again, we have that
date of when voter registration's happening. And then, I have a
graduate student, of course, Matt Thornburg, who's going to be on
the market, so there's some people not here. He's on the market
now. But, he's looking at party idea and its congruence with party
registration.

And boy, I would love to get some good data on that. I mean, we
have some validation studies of party registration, but there's not a
lot out there, and much of what I was looking at in terms of the
eye early vote this time around we have party registration statistics, but
nobody was asking in any of their surveys, Gary or others, it's
always self-identified partisanship. It's not which party are you
registered with?

And a lot of misinterpretation, I think, of those early voting data as
a consequence of people not really understanding that someone not
registered with a political party is not the same thing – it's similar,
but not the same thing as someone who's a political independent
when they answer a survey. So, I'd love to see more validation on
all that.

And this gets us to some of the problems here. So we have errors
in the records, and these are easily to observe. You get, on the
birth date – not surprisingly, this is data entry effort – you've got
millions of records, you've got local election officials across the
country doing this data effort – not surprising that you find people
from the future who have been voting in our elections, or people
thousands of years old.

But, it becomes more pernicious when you start looking at the
addresses, and you see four digit ZIP codes that have been entered
into the databases, and it's – for example back in 2008, there was
this talk of doing a re-vote in Florida over the primary, the Hillary
Clinton campaign wanted to do that. And I pointed out, well, just
look at it.

The errors, actually, are correlated highly with race, because that's
another state where we have race being reported. So, you find
these errors more prevalently along African Americans. So, if
you're going to rerun an election by mail, which is what they
wanted to do in Florida, you're going to have some problems there
of actually getting these ballots out to the people who would be
qualified. And that pretty much put a kibosh on the whole thing,
because then there were going to be voting rights concerns and
everything else.
So, these errors are important, and they can affect some of the matching that happens with databases as you look across different databases, there are biases in these errors, as well, that aren't really well known except – or well-studied – but it's something, again, that, as we move forward doing this sort of matching, and what election officials are doing right now with matching records of individuals with lists of citizens, and trying to purge off the non-citizens off the voter files, we already know that there are problems with that. It's deeply challenged.

And that goes to the next date point, the next bullet point, which is that when you do this matching on a broad scale – and this is just a very simple simulation that I did and published. Not surprisingly, when you've got millions of records, and you're matching people, you're going to find two people with the exact same name and birth date.

So I'm talking about year and month and day. It just happens when you have millions of people, it's that birth date problem, except it's now, instead of just within two people sharing the same month and day, it extends out to the year, and when you have even a small classroom of 30 people, you're almost guaranteed to have a match. Now you're talking about a million people. So, it happens.

And so, these election officials have been out there sending out notices to people this election cycle saying, “Hey, you're not a citizen,” even though they are a citizen, and it's just because of a bad match. So, on these matching, it leads to both false positives and false negatives.

And then, again, I mentioned social desirability issues with records of race. This gets, actually to me to something I would just love for someone to do, NSF, somebody, I don't know, it's got to be done. If we're really going to understand the vote, the worth of vote validation, someone's got to do something which I call a pre-validation study.

And that is take the voter file after the election, and go out and validate – instead of doing the backwards thing which we do, which is to validate the survey with the administrative records, take the administrative records where we have the record of voting, and go out and ask people whether or not they voted or not.

We got to check the reliability of the administrative records themselves, because a long time ago, when ANES was doing this,
that's one of the findings that they had, and it's one of the reasons why they stopped doing the vote validation was they weren't convinced that the records themselves were reliable.

So, there's this siren of looking at – the siren call of saying “Oh, we've got all this wonderful data. It's easy to collect now because of the Help America Vote Act has now created these databases, and the list vendors like Catalyst and Aristotle have compiled all of this data for us and they'll sell it to us so we don't even have to do that data collection anymore.”

Yeah. They've done it all, but you're just assuming that that data is reliable data, and I think it's a very strong assumption, and we need to really poke at that, and, again, going back to the social desirability of race. I think that there's a lot of error there that's not just random error that people are from data entry are – I think there's probably some systematic biases in these data that we really don't appreciate yet. So, I want to just put that out there. I know we're begging for money from NSF. Not really, but yes.

Absentee ballot requests. So, there are these other files that are out there which very few people are even aware of or have looked at. And these are on the individual level of records that track the disposition of the absentee ballot requests.

So what are these? These are linked to the voter registration file, so you have a ID, unique ID, and it links back to the voter registration files. And what the states are collect – they have to do this. The election administrators need to make sure that someone doesn't vote twice. So, they are collecting all this information. It's part of their reporting systems.

Any person that requests a ballot, that's information that's tracked by these local election officials. The reason for that request is often in there so we know if this is an elderly person in a nursing home. We know if it's an overseas military or domestic military, an overseas civilian, a domestic civilian. We know all of that information. That's something that the election officials are tracking.

And we know which address they're sending it to, so we know which other state. Is it another country? We know that information. We know the date that the ballot's been sent out, returned, and we know the disposition. Well, we actually know the mode. And that's kind of interesting, as well, because now, for the
first time in this election, we had electronic voting put inside New Jersey and that emergency issue in New Jersey.

But, the military, and overseas civilians were voting electronically this time around. Little known fact. And 70 percent in North Carolina had voted electronically in this last election, so there were several thousand people who were voting by e-mail in this last election. But we also know, besides electronically, mail and person, whether or not a person voted early in either mode.

And then, we finally know reasons that the ballots were being rejected. So there's a lot of wealth of information here on these ballot requests, and so far, they're being used to do things like estimate the overseas population.

Rand did a study on this using the voter files and trying to figure out how many overseas civilians there are, and they used this as one data point to try and find that information, something that the census bureau ten years ago said, “Eh, we're not – impossible to do.”

So, now I know the census bureau has accuracy as the primary goal, so looking at these numbers, we're not going to get an accurate number, but at least we can get something. So, that’s what Rand was trying to do for the federal voting assistance program.

And something that I'm going to be doing for the federal voting assistance program is analyze patterns of vote failures, because we have that information, the military voters, where they – what reason were their ballots being rejected – are they being rejected at a higher rate? Are they not being returned? What's the rate of return? We're going to try and figure that out so we might be able to better serve our military and overseas civilians in the very difficult things that they have to do in order to vote.

And then finally, and this is something that came up this election cycle. I brought it up with Mark Blumenthal and he did a little post in Pollster about it. The error in the early vote. [Laughter] I mean, we had some of these surveys out there estimating implausibly large early vote percentages.

I think it's – we could actually do some sort of vote validation on the early vote, as well, because we can ask people that question, and then we can validate the mode in which the individual voted, the time of which they voted, if we wanted to go even further down
into it. And so, no one's done that yet, but it's certainly something that's possible to do with these data, and hopefully someone will take a stab at it.

Finish up with data availability. So, we know about HAVA. I already talked about that. However, for our purposes, there are -- many states make these things publicly available. In fact, you can get the -- North Carolina just puts it on an FTP site, and you can just go and download all of this data. It's a really wonderful resource.

But other states restrict access to either the voter registration file or the vote history, or the absentee ballot request files to the political campaigns. And so, we, as people on the outside, unless we're affiliated with a campaign somehow in our work, we can't get access.

In fact, Florida, for example, wouldn't even let the federal voting assistance program have access to their absentee files so we can better serve military. So, I think they're going to change that law in Florida. I don't think any elected official wants to be on the wrong side of the military. But, we'll see.

So, if you're doing this on a state by state basis, certainly in some states, the data are very much available and accessible, and you can get them and you can do analyses of them. But, if you're going to do it in a national sort of way, then you have to turn back to the list vendors.

And even these list vendors are constrained by some of the state laws for sharing data. So, for example, Virginia, in the Polimetrics surveys, is not validated because the state of Virginia forbids that dissemination of data outside of campaign purposes. So, even using, turning to the commercial list vendors, you still can't do the full vote validation that you might wish to do unless they are going to break the law somehow. I know Bob Blaemire isn’t here yet, but he'll be talking about it tomorrow. He may tell you something different.

And I know that there's some people that are very high on this. I know Stephen Ansolabehere really loves Catalyst. They've been doing the vote validation with the Catalyst data. I am not as convinced. When I was helping Jon and Matt do their vote validation, I looked at the reliability of the data that Catalyst had been collecting versus the data that I've been collecting.
I came to the conclusion, look, Catalyst has a different purpose than doing research. That's the bottom line. Their purpose is serving their clients, which is to have the most up to date voter registration list. It's not to collect, for example, the absolute snapshot of the electorate on election day, because some of those people by the time a couple months from now, they've moved. And so for Catalyst purposes, they don't care about getting the best data on vote history, because it's not useful to them.

Now maybe if there's demand for it among the survey community or others for that data, maybe we'll see them try and collect that data, but I was able – I had that purpose in mind, and so I was collecting the data as soon as I possibly could get it. And some of these states are doing real time purging of their data. So, the data's gone. You're never going to get it again.

And so, when I was comparing my vote history with Catalyst's, I had more voters than Catalyst did. That was by a few thousand votes, but still there was a real difference there. When we looked to Catalyst to do some sort of vote validation for the federal voting assistance program, they told us, “Well, we don't care about the overseas civilians and the military. That's not – you can't reach out to those people. So, we're collecting data for our clients, which do voter contacts.” And so, they don't care about the overseas civilian population. They get purged from their files.

And I can go on about this. The data is also being cleaned by individuals who are going out and doing the contacts, and if they observe somebody that's not at the address anymore, that gets reported back. Well, look, Catalyst is a Democratic leaning organization, by and large, and so the data, the contacts that you're getting are for contacts that are on the Democratic side. They're cleaning the Democratic side of the electorate.

Aristotle's probably cleaning. I don't know much about them, because I don't work with them as much, but there are some academics who do. But they're probably cleaning the Republican side. So, there are going to be some biases in here that are in these data, and I don't think anybody really knows. And so, this comes back to my last plea, again, because these data are being used for so many different purposes now.

And Catalyst will sell academics access to their list for $10,000.00. I think it's creating a – I can't afford it. I'm at George Mason University. I don't have a lot of money. So, I can't afford access to Catalyst. It's creating a resource divide within the profession of
people who can get access to Catalyst and people who can't, and it's – we don't know the biases of those data. There are people who are using it say it's great, they love it, but they're not – we don't know. We don't know those biases.

So, what we really need to do with this pre-registration survey – or pre-validation survey, we need to collect up these lists. We need to let a lot of academics take a look at them and, again, as a plea for funding is that let's collect this up. Let's disseminate as much information as we can so that you can use it, perhaps, for your purposes, and other academics can take a look at it and look at its validity, as well.

So, that's my plea. At the end, of course. Any questions? Yeah, Gary.

*Gary Langer:* I'll try not to talk too much. This is Gary Langer. I wasn't aware of this. My head is exploding at the concept that academics are using the Catalyst list for their research.

*Michael McDonald:* Yes.

*Gary Langer:* Catalyst is a partisan political firm founded by Harold Ickes, Jr., the former Deputy Chief of Staff in the Clinton administration, with the purpose of developing lists of voter contacts to drive turn out the vote mobilization campaigns.

The sources of its data are ill disclosed and, in fact, generally non-disclosed, and are produced for partisan purposes. I, through a client, looked into using Catalyst for a survey of Hispanics. They offered me a list of more Hispanics in the United States than there are.

*Response:* [Laughter]

*Gary Langer:* And when I asked, “Why?” They said, “Well, there were three strata here. We have the high probability, the middle probability, and the low probability – lower probability – segments of the list.” And I said, “Great. What are the probabilities?” They had no idea. So, there was no way to compute the coverage, utterly no way to compute the coverage or the non-coverage of the list compared to the actual population of Hispanics, nor was there any source information on where the data in this list came from.
So, any sort of list based research has to start with the integrity of the list, and that has to start with a really careful parsing of how the heck it was assembled, and by whom, and for what purpose.

**Michael McDonald:** And that's one of the reasons why we, when we were looking at Catalyst to do this for the ANES, we walked away, because we – it was a black box, and they weren't willing to let us look into the black box. What little we could see when we looked in the black box, we got the same – we came to the same conclusion, is that there were some biases in there and we're not now comfortable with working with these data. So, I appreciate that comment very much.

Josh?

**Josh Pasek:** Yeah, Josh Pasek, University of Michigan. I'm going to harp on this a little bit more in my talk tomorrow, but I think this really is a big issue, that when we want to think about in very general terms adding on additional data to the surveys we're collecting to various different data sources, if we're not in academe putting together those linkages, and we're relying on these commercial firms that by and large, aren't opening up the process, I think this is a huge issue for thinking about the quality of these data, what they might do, what biases we're going to end up with, et cetera.

And so, we need, I think, to really, to think about how we can ensure data quality, either in a world where the data sources that are most easily available are ones where we can't get access to the data. And so, this game is going to be one where we either have to create something ourselves, or we're going to have to try to find some other way of dealing with this problem, but I don't know what that might be.

**Michael McDonald:** Again, just to underscore what I was saying earlier, there are a lot of people using these data now. And they're buying into Catalyst. I mean, Catalyst has made this data available and people are using – I think they're even using NSF funds, in some cases to purchase Catalyst data. So, it's – yeah. And, [laughter] we'll pick up the pieces of your brain [laughter] after the talk, Gary.

So, I think we really need to get a handle on this before we go much further down this path. [inaudible] Yeah. [inaudible] Sorry.

**Tom Smith:** Yeah. Tom Smith. I've worked with a large number of auxiliary data sets, including Catalyst, I'm not going to speak in favor or against them specifically, but here's my point: Most of these – the
The general rule is they're all black boxes, whether it's Catalyst or whether it's all kinds of other things.

But there is one glimmer of good news. I've found that if you engage them, many of these database providers, in deliberation and discussion, you can find out a lot more. Some of them are keeping this information private for proprietary reasons. Some of them are perfectly willing and says, “Oh, well no one's ever asked us how we did this.” And you can get a lot more out of them, but you really have to pursue it.

The initial thing you get is “Yeah, we collected this data. It represents voters, or motorists, or something.” And, you know, the top line is, basically almost no information at all. But, I've found in more than half the cases, you would get much more detailed information if you engage the auxiliary data providers.

Michael McDonald: I bet you it's highly correlated with how much money they're making off of that data – [Laughter]

Response: [Laughter]

Michael McDonald: – whether or not they're going to provide that information. Yeah.

Richard Freeman: Yeah. I'm not familiar with this literature, so you'll have to excuse my naiveté, but has anyone ever – you mentioned aggregating up to the household level, because you have the address information and the voter registration records – has anybody ever looked at the prevalence of voting or anything at the household level linking that to a given survey and then seeing whether or not for that particular survey sampling frame if that kind of voter activity is predictive of response behavior in a survey? Have there been a lot of studies like that? Or has anybody ever tried doing that?

Michael McDonald: To my knowledge, no.

Richard Freeman: Hmm.

Michael McDonald: No one's done that study.

Richard Freeman: Because it seems like voter behavior might be predictive of engagement in survey requests. So...

Michael McDonald: Yeah. Where people have done the – but it's not in a survey context. It's in the voter mobilization and Green and Gerber literature. There have been people who have studied and found
that a contact within a household stimulates other people within the household to vote, as well. Yeah.

Richard Freeman: Has anybody tried to reverse engineer the mysterious lists? There's got to be someplace it comes from, and I can imagine the nice little project of someone trying to figure out, and then going and saying, “Hey, is this what you did?” Et cetera.

Michael McDonald: We have the canonical data. It's the data that's coming from the election officials. And so, I could do a very crude analysis of just comparing the topline number – the number of voters – in their file to some of the states that I had collected and I could see that they weren't the same, and I knew why that was happening, because they'd collected their data at a later point in time. So, I've done it that way.

Now, I suggested because Pew is another organization, not the Pew survey folks, but the Pew research, the Pew Make Voting Work side of things, has also been buying up Catalyst data. And I said, “You know what you guys need to do is you need to do this validation,” exactly what you're saying, which is let's see what the difference is between what Catalyst is giving you and what we're getting from the election officials.

Because they're using this information now to say what the accuracy of the election officials data is. That's what Steve did a paper on that for Pew. This was the one that generated a bit of controversy about the number of non-citizens on the rolls. So, and Secretary of State Gessler out in Colorado loved it, and so they were using it as evidence, and actually brought some folks in to take a look at their voter roll based on that project.

And so, now we're [laughter] – let's stop for a second here and take a look at the accuracy of Catalyst and what they're doing. And I think Josh is going to, tomorrow, give you some pause on that, as well. Because the other thing – and I don't want to get, step too much on what Josh is doing here – is they're taking those files, and they're matching them with the consumer databases and they're supplementing that data quite a bit with other data that's out there.

And they're doing some cleaning. They're doing some purging. They're adding data to it. And that's where you could get what Gary was talking about, which is they're probably, I don't know, 200 million Hispanics in the country, the way in which they can
supplement these data somehow. And we don't know a lot about that, so that's certainly doing this through, at least what we can tell from the outside since we're not privy to their black box, can we reverse engineer and see what's going on there? I think that's kind of an important thing to so.

Frauke Kreuter: Frauke Kreuter. Just a quick question. Two. One and a half, maybe. I was wondering about the recency of these, or the timeliness of these data. I mean, I know very little about voting registration since, in Germany, you're automatically registered, and I'm not allowed to register here, so I've never paid attention to that. But, I wonder if there's a lot of disparity among the records within a state on how recent they are and how accurate they might be?

And, related to that, the question: Could they be in any way or form supplement screening efforts that people do when they try to find people a household, or get a sense of household compositions? And that leads me to the question, the thing is like, can these individual lists be compiled and identify a household, or not? Or is it just like whoever lives at that address?

And so if you can speak to timeliness and household composition, that would be great.

Michael McDonald: Yeah. So the timeliness issue's easier for me to get a handle on. So, we know that it's – and this varies among the localities and the states that are out there. So, some states, we get a really great snapshot of the election, if that's our goal, because they take that snapshot of their voter file at that point in time and you can request that file if you wish it.

Other states are doing transactions into their files in real time, and some of them keep all those transactions. They have a whole – I think all of them are probably keeping a whole transaction history available, but it's not something you can publicly purchase from the state. You'd have to ask them to do something special for you to get that full transaction list and see all the changes.

And then, these transactions are not just address changes, but also name changes and party changes. So there's a lot of things that are happening within that voter file that can be revealed by understanding a little bit better those transactions that are happening.

But if your purpose is to get the most recent file from the state, that's – they can take that list and they can pull it off at a point in
time and give that to you. But, also understand that there are different types of election administration out there. There are some states that we call top down states, where the election administration is centralized at the state level.

And there are others that are bottom up, where its localities are primarily responsible and they're feeding data into the statewide system as mandated by HAVA. And in those states, you could get delays in the transmission. So maybe it's only like on a once a week basis, or once a month basis, especially when there's not an election going on. So you may have, in those states, and unfortunately, you have to really look at each individual state to know this, what the timing is.

And I'm sure that those practices change by the month, but you could certainly see a delay in those numbers and, in fact, for example, I was working with a reporter down at the Miami Herald, Marc Caputo, about the Florida early vote, and he was using a voter file.

And he said, “I have a voter file I know is out of date. I know I have 100,000 fewer records in here than it should be because we could see the state produce that new data from the time of the beginning of the month to the registration deadline. Because we know from registration timing, that's when everybody's registering.” But the state wasn't providing that last file to him, so he was kind of stuck and was doing analyses knowing that he was incomplete analyses for that reason.

On the household composition, the research I've been doing on it so far – first of all, I can tell you the campaigns certainly look at party reg, and they know something about the household composition. They know that some of their target rich environments are all people registered with one party or the other party, depending on which party you favor. And so, certainly, they're doing it.

I do – and that's really actually my experience in doing that work led me to say, “Hey, you know you can actually do this in party transmission?” Reanalyze those things. And all I can say for that is that, well one, if you can find the pairing of people's names and a child with the same name, so two parents of about the same age and a child the same age, it works. You can get party registration transmission.
It looks a heck of a lot like that self-identified partisan transmission that we've seen in the survey literature. So, it's kind of cool, and you can actually – because you've got millions of records here you're dealing with – you can actually look at some, like, do two Libertarians have Libertarian children? Yeah. They do, actually. So... [laughter]

You can really do some things that you can't do with the survey research. But, beyond that, I don't have a good feel for it. We know that people move around a lot. You know that there are people who are still residing at the address and, by federal law, unless an individual has notified an election official that they have moved, they must remain on the voter file for two years, or, excuse me, two consecutive federal general elections.

And so, you get a lot of what we call “dead wood.” And you can see multiple people registered at an address, especially apartments, right? And then, with apartments, we often don't even have the apartment number, too. So, now you're dealing with that issue, in terms of the data quality issues.

So, in some cases they'll say, yeah, if you, maybe, depending on what your purpose for the data, you could devise a scheme to get some information about household that might help you out, but in other contexts, I would be very cautious about just broadly applying some rules to say that this is what a household is, and all households that I can observe on the voter file are these sorts of households.