Jon Krosnick: Okay, we will continue on this roller coaster of adventure of linking survey data to other data and we’re very fortunate after yesterday’s discussion to have Bob Blaemire with us from Catalist and I tried to brief Bob a little bit before about what our discussion has been about with regard to Catalist style data and Catalist data so that he has a sense of the kinds of issues people are looking forward to hearing about. And so I’m looking forward to hearing what he has to say.

Bob Blaemire: Thank you Jon. Jon gave me these warnings about the previous conversation. It reminds me of an old story about Daniel and the lion’s den. You know where he’s thrown in the lion’s den and he’s surrounded by all these lions and he falls to his knees and clasps his hands and prays the Lord, you know please protect me. And nothing’s happening. He looks to his left and looks to his right on each side of him there’s a lion on its knees with its paws clasped. And he says, “Thank you Lord for delivering me.” And the one lion says “I don't know what you're doing buddy but we're saying grace.”

You know I would tell voter file jokes but there’s very hard to find a voter file joke that is clean. So you know the small world thing I have to say, Josh Pasek just having spoken, Josh’s brother and my son went to college together at University of Michigan so I find it fascinating to walk in here and have Josh speaking. Let me give you a little bit of background on who I am, where I come from, what Catalist is, what we do, how we build data and clearly you know we can have a give and take if you like about how we work with academics and surveys versus real data and so on.

I came to Washington when I was 18 years old go to go George Washington University and right away went down, first day in Washington after my parents left I went down to the Senate to volunteer to work for my state’s United States Senator Burch Bayh of Indiana and ended up eventually getting a job there and staying with him for the next 13 years until we lost in 1980 to Dan Quayle who you might remember. Although there were other factors that went on that year I think that were more important than Dan Quayle. But during those 13 years I spent a lot of time learning about what did the Senate offer us in terms of services and one of the things I learned early on when I was very young was that there was a computer services division that compiled computer databases for members of Congress or members of the Senate to do franked mail. And I said well great, what does the Bayh office have? And he says, “You have one list of 13,000 names that you mail once a year at Christmastime.
And Bayh that point had been a senator for nine years. And I can
tell you that in the next nine years we built a list that ended up
having 2.8 million households in Indiana with 250 identification
codes and for two years in a row we’re the biggest mailer in the
United States Senate which a lot of people would argue was a big
waste of money. For me it was a great badge of pride because not
only did we learn tons and tons about building a database for
constituents but also about the communication process about what
worked, what didn’t work, you know inviting people to town hall
meetings, do they come, it was fascinating. The first time we ever
did that. We sent a frank mailing out to the geography around
where a town hall meeting was going to be there and went out with
him for it and there were probably four times the number of people
that would fit in the room there and many of them had the letter in
their hand; very satisfying experience.

So we lost the election in 1980 and couple years later, knowing
that I had to get a real job I ended up getting hired by a firm that
wanted to build voter lists to match them to the AFL CIO
membership for the AFL to find out who was and was not
registered. And so the first question came, what do you know
about voter files and all I knew about voter files is that we didn’t
have them in Indiana, that we had to use commercial databases
where we did not know in our ’80 campaign whether people were
registered or not. We knew certain commercial aspects of them
and we did geodemographic targeting using clear task cluster
technology if you’re familiar with that but we didn’t know a lot of
the things we would like to know about them in politics. You
know number one is are they registered.

So I didn’t know a lot but spent a lot of time learning, calling
around the country, finding out what was computerized and
happened to be in a good time in this particular history because the
country was starting to become computerized. And as over a
period of time in continuing this investigation on what was
available in computer form, more and more was every year. And
looking back I realize when we were doing this nobody was doing
this. This is kind of the beginning of a business which has gotten
quite a bit larger now and we didn’t know that we were inventing
something. But I ended up working in this business with another
company for about eight years and then started my own company,
again, building voter file databases mostly for Democratic state
parties and Democratic campaigns and becoming the vendor
providing data to Democrats through the good offices of the party
which was a good thing all the way around because it brought the
party back into candidate services. It helped candidates have access to computerized voter data that they had not had access to before and it helped vendors reach candidates we could otherwise never reach. And that business grew.

I had my company for 17 years and then merged with Catalist in 2007. Catalist was formed by a combination of progressive organizations after the 2004 America Coming Together project. If you're familiar with that, this was an effort by the progressive community to try to coordinate its activities so basically the country would be covered better, people wouldn't be all going after the same voters and you wouldn't have huge gaps of geography omitted. And at the end of the 2004 campaign when of course we did not win the presidency, a lot of the organizations got back together with Harold Ickes who had been the Deputy Chief of Staff for Bill Clinton and is now and still the president of Catalist. The organizations got together and the common theme was a complaint about data that data wasn’t available in a consistent, reliable form that all the organizations could use.

I was providing a lot of the data to ACT but a lot of the organizations within ACT would not use our data because they were 501(c)(3)’s and they wouldn’t use political party data which most of mine was owned by political parties. So there was a – the constant struggle and nobody had the whole country. So Harold put together a bunch of investors that built Catalist and started building the whole country to make it available to the progressive community. So we sit in the center of a lot of progressive organizations, AFL CIO, League of Conservation Voters, Planned Parenthood, SCIU, a number of organizations like that almost like a public utility where they have online access to the data, they pull data out of our system and they put response data, ID’s that they collect in the phone and in person and by mail back into the database. We then can take those ID’s and all the data that we have and use that soup to create models and the increasing level of predictability of the models has become sort of the coolest new thing in politics and I think that’s something that sorta has given us the edge in this particular business.

Are these databases perfect? No. Are they better now than they were when I got started? Ooh, they sure are. You know I used to build the main voter file for instance from 516 sources, towns all over the state. One of the hardest things you ever want to do. We sent out letters to every town asking to send their voter file for the Democratic Party and right away 300 of them come in and those last couple hundred are killers. You know you get a message one
of them says call our town board between 2:00 and 3:00 of the first Tuesday of every month and of course the line’s busy. You can’t reach them. So it was very hard.

And increasingly over those years, particularly after the Gore, Bush Florida debacle and Congress passed the Help America Vote Act, more and more states have become statewide databases and they’ve gotten better and better. And the progress in this area has gotten better and I can just tell you that the error rate on companies like ours has also gone down. So while they’re not perfect, they certainly are better than anything that has ever been available in this particular business.

So let me show you a little bit of a PowerPoint both about Catalist and about how we build the data and maybe some applications about how it’s used if you're interested in that and then I’m happy to open up to questions.

[PowerPoint video Start]

All organizing starts with lists, lists of the people you need to reach. Before Catalist voter and prospect lists were built and rebuilt and pieced together from multiple sources at great expense. When each campaign ended the data would be left to grow old, get lost or get tossed and targeting was too broad. Sawed-off shotguns were being used when we needed heat-seeking missiles. So five years ago a national database was built of all adult Americans, owned by a trust and only available to progressive organizations. It aggregates the billions of data points that progressive groups of all kinds collect.

This knowledge asset, Catalist has revolutionized targeting. The likelihood of an individual to vote, to donate, to take action or be reached through certain media can now be predicted, letting progressive groups zero in on the most receptive targets and stretch their resources. Women’s Voices lowered their costs by 40 percent per new voter registration and increased registration applications nine-fold. Planned Parenthood found supporters in neighborhoods they didn’t even know they had supporters in and mobilized them to build good will where they plan to open new clinics. The league of conservation voters used Catalist to gain a 30 percent lift in fundraising response and a 24 percent higher average gift in small dollar prospecting programs. Great progress.

But now it’s time to take it to the next level. First just as financial institutions are building consumer information about us every time
we swipe a credit card; we are now harnessing civic engagement information to help organizations raise more small dollar money. Next, we’re continuing to innovate by making it easier for new software apps, particularly for social and mobile to tap directly into catalyst data. Finally, our voter data is now being used to target online advertising in an area where corporate cash is drowning out progressives on broadcast. Data driven online advertising may be the lifeline that ensures progressive candidates and not for profits will still be heard.

[PowerPoint video End]

Bob Blaemire: I have a tendency to talk fast and please tell me if I’m talking too fast but that is not my hand moving there. Okay, so this is how we build the database basically. The component parts, we start with about 184 million voter file records in every state. Included in that is over a billion items of vote history or ballots cast. Another 80 million unregistered persons 18 and above that come from a commercial database. And then we have specialty lists matched in things like aviation employees, doctors and nurses, farmers, teachers, hunters and fishermen. Then we have the ID’s these response data that came from America Coming Together in 2004 and the Kerry ID’s from 2004. This is the basis. You think I’m going too fast, all right.

Audience: [Question]

Bob Blaemire: Okay, which part, the whole thing?

Audience: [Question]

Bob Blaemire: Okay, well basically we’ll start with the voter data. The voter comes in most cases, all but two states we get from Secretaries of State. That’s not true; all but three states we get them from Secretaries of State. When I started with Catalist in 2007 there were only 28 states that were statewide. There’s now – all of them are statewide but a couple of them are very restricted on who can get them and we cannot so we have to get them by county. And we build the voter file multiple times per year.

Every state is at least twice and in the election year we did several states, oh man, 20 times. We were doing Colorado like every couple days and the last month. Vote history again comes generally with the voter data. Sometimes it’s separate, it depends on the state. Infogroup is our vendor, you mentioned Infogroup, one of the large list houses in the country where we get the
unregistered people and our unregistered people are basically made up of three types of persons, those that when we match our file to their file, they give us back people who are 18 and above who don’t match.

We then retain people who were on previous versions of the voter file but not the current one. We call them dropped voters. And then we also have, if someone has filed a change of address and moved into another state, they’re available to that other state as an unregistered person in that state though they may be registered somewhere else, someone we refer to as a multiple person.

Then we go out when clients want different kinds of specialty lists that we can get, we go out and see where we can get publically available license files like gun owner licenses or hunter licenses or teachers or farmers. There are some states you can get these very, very easily, inexpensively, there’s other states you can’t get them at all. I remember getting – when I had my company, getting a list of gun owners in Ohio for $10.00 and other places, you know you can’t get lists almost of any kind for anything. So essentially very few of these are 100 percent national but they all become – they're all large enough that they give us the kind of data we need in order to be able to model in some of these areas.

The ACT ID’s and the Kerry ID’s are in a whole range or party preference, issue preference and candidate preference that were taken throughout 2003 and 4. Membership records, these are all of our organizations, one of the things they do when they become a Catalist client is they take their membership and they match it in. So we have an indication of which of the voters or the unregistered people are members of a given organization. And so there are 54 million people that are a member of one or more of the organizations that are on the database.

Then we have, and most of this is from the 2008 campaign and this PowerPoint refers to data mostly from that last presidential. This one’s a little recent for us to have any data –

*Audience:* I’m sorry, can I just ask, when you said 54 million people, is it 54 million members and fewer than 50 –

*Bob Blaemire:* It’s probably 54 members match with the voter file. It’s probably fewer people. That’s a good point. You guys strive for much more accuracy as those of us in the voter world strive for as I’m sure you’ll agree. 266 million contacts to almost 126 million individuals so we know who each organization is contacting for
different purposes throughout the campaign in 2008 and what responses came back. And then 264 models that had been built by our clients and ourselves scoring 2.5 billion records. In other words everybody multiple times.

So the traditional voter file gives us the name, the contact information, the address, phone, gender, age and sometimes includes party. There are 20 states that do not have party registration. Vote history, and again one of the things about the Catalist database as opposed to if you went out and did it right now yourself, we would have a lot more history 'cause we built it over time. We keep everything we have and we keep building on top of it so our history for most states goes back before 2000.

Race, there are seven southern states that provide race on the voter file which again gives us a huge dataset to be able to model race. Now this is again, looking back at Lancaster Pennsylvania, done in 2008 and looking at the traditional targeting where we would take precinct level targeting first where everybody in a precinct performs a certain way, that aggregate data used to be all that was available or precinct targeting, you know everyone was a target or they are not. And the voter file helps us exclude Republicans and 65 percent Democratic precincts as one example or the voter file could also identify Democrats in heavily Republican precincts.

And then we go to the enhanced voter file which is clearly where we are now, where we have all the basic contact information and then commercial data. And the commercial data we get through Infogroup which is hundreds of fields of commercial data items that are used for analytics purposes and are not necessarily available for clients for contact purposes. Census data which as you know is right now is very fresh but part of this, it was getting old. Historical ID’s not only from our own clients but many of their clients that they collect data from as well. And then the specialty data, those particular kinds of lists that I was talking about, hunters and fishermen and so on.

So the enhanced voter file, again Lancaster County, identifies registered Republicans and then modeling predicts which Republicans strongly oppose the Iraq war for example. And modeling also identifies republicans who strongly support choice.

**Audience:** [Question]

**Bob Blaemire:** It comes from modeling, having the issue responses. I mean one of our clients for instance is a polling consortium which is a
consortium as it sounds like, of pollsters that getting agreement to take their responses on a regular basis and upload them to our database. So we end up having thousands and thousands and thousands every week of actual responses on issues in a variety of categories. So when you get a dataset that’s big enough and you think you can build a model from it, we go to work.

Now our database is comprised of 280 million voting age Americans with 700 fields of information so it’s a lot of data. And it’s a lot of data that you can create these models from. And I have to tell ya, I look over the years of doing this crazy business. And in many respects we’ve always tried to predict what we’re going to do. We try to say is a Democrat going to be more likely to behave like a Democrat than a Republican? Does an older person care more about Social Security than a young person?

You know, I think we all agree that there are lots of things, assumptions we can make about voters and the more we know about a voter the better we can communicate with them because the better likelihood we have to make accurate assumptions about what they’re likely to feel about. We’re not dealing again with perfection, we’re dealing with likelihoods. We want to make sure that if we’re – you know I’ve joked many times that if you try to get my wife’s attention by starting a conversation about the Washington Redskins; you’ve lost her at the outset. But I’m paying attention. If you try to get a young person to pay attention to your campaign by telling them about the future of Social Security, they don’t care.

There are certain things that may be obvious to us on a little bit of data but it can become more nuance, more detailed as we have more data and we have a lot of data. You know I often use an example that if we bump into each other on the street, what do we talk about? The weather? We don’t know anything about each other. I can perceive your likely gender, your likely age group, your likely race but I don't know anything else about you. Oh I find out that you're meeting at this meeting, I have something to talk about. I find out you're at a baseball park where I am, we can talk about baseball. The more I know about you the better I can communicate with you and the better chance I have of persuading you to do something and that’s what we’re about.

So our motives are far different than yours in terms of we want to get the data as good as we can get it for practitioners, for people that need to use it for political communication. This data is restricted by law as well. We cannot use it for commercial
purposes so you cannot use it to sell products. The only clients we have that are for-profit clients are consultants and pollsters who agree that it is only being used by a permissible entity, that they have to tell us who it is. And I’ll give you an example, I’ve had to turn down campaigns like a casino gambling campaign because the actual client in this case was Harrah’s Casino and if it was a coalition of let’s say nonprofit organizations that got together to promote or oppose a casino, that would be different. We could do that but we can’t do it if it’s indirectly commercial.

We have different restrictions we have to deal with in different states and this is not an easy process. Building these files is difficult and it’s fraught with error. I like to believe that one of the advantages that I’ve been able to bring to Catalist having done this as long as I have is helping us do this better. And hopefully we do. I would like to give you some examples of something I think is pretty cool because for many years I’ve always recommended this to candidates and sold it to candidates ’cause I believe it makes sense. I believe it works. And I’ve always believed that but now we have really sort of more empirical proof that it does work.

We can compare people that were contacted to those that weren’t contacted to see if there’s a difference in turnout and support. So we create models and these models, you know they’re tools, they don’t predict anything. They help us target better on an individual level. They improve our efficiencies in selecting universes and to expand and contract our universes and to tailor what messages we use. And the models we have are things like national partisanship.

We have a database that has everybody in the country who’s registered by party and everybody in the country who votes in a partisan primary and tens of millions of people who have expressed a party preference on the telephone. That’s a huge data set to create a partisan model. It was created in 2007 or early 2008 and it has been revalidated and re-tweaked I should say or re-jiggered every year since. It’s probably the most popular, most widely used among models. Second to that is our voting propensity models which we’ve done every year. That’s the one that’s easiest to validate, the likelihood to vote because we can see after the election whether they did or not.

Did we predict it and did they? You can tweak it very well and that is also used extensively but then we have more esoteric models you might say like likelihood of going to church at least once a week, likelihood of being a hunter, likelihood of being a gun owner, likelihood to support Obama in 2012 against Romney.
A series of models like that including an ideology model which allows us very well to slice and dice different partisan portions of the electorate, find Republicans for instance that have progressive views, find Republicans that are pro-choice or believe that gay marriage is not a sin. It allows our clients to do a better job of slicing and dicing the electorate.

In 2008 again, 90 organizations used our data to make 335 million contacts, phone, mail and at the door to 126 million individuals with another 7.4 million unique voter registration applications. This map indicates where the contact was more Democratic than 2004 or more Republican and you can see the contact, at least among our clients, was far more Democratic oriented.

This is a time lapse view of the data being pulled out and put into our system throughout the 2007-8 period. This is the Iowa caucus, New Hampshire primary, Nevada caucus, Florida primary, Super Tuesday. John McCain becomes presumptive nominee, the Obama race speech in Philadelphia, Pennsylvania primary, West Virginia primary, Hillary Clinton concedes, the Obama speech in Berlin, Democratic convention, Fanny Mae and Lehman Brothers collapse, John McCain suspends his campaign, the DOW falls 800 points, the last presidential debate, Obama’s prime time TV address and the election. And if you see, talking about the early motivation why we, Catalist was created and you look at the logic of the contact over time, you see Illinois is an island of non contact. That makes sense. Why would you spend a lot of money contacting Illinois when Obama not only is Illinois normally a blue state but the senator from Illinois was the candidate? Now you see a lot of contact around Chicago.

*Audience:* Fundraising.

*Bob Blaemire:* Fundraising, California same thing, an island of non contact. Fundraising in San Francisco, fundraising in L.A. Even Texas you see fundraising in Dallas. New York, contact around the city for fundraising but not the rest of the state. So there was a logic. This coordination did seem to work. Then we looked at the state of North Carolina, the first state that would provide us a database with vote history uploaded right after the ’08 election.

So we do this graph to look at the sporadic vote or also known as surge voters, the voter that would vote in a presidential general election but not in a off year election. In 2006 or 2010. And we see the turnout of the sporadic voters and the Democratic sporadic voters overall compared to the turnout of the sporadic and the
Democratic sporadic when they’d been contacted by one of the 90 organizations working with our data. The turnout went up. Similarly in Ohio, Ohio is not a party registration state though they do vote in partisan primaries. It’s tough to target in Ohio because only about half of the people vote in primaries. I know 'cause I was the vendor of the Ohio Democratic Party, the Kerry campaign and the coordinated campaign in 2004 and I argued tooth and nail unsuccessfully with them about how they should target in Ohio in 2004.

But let’s look at 20008. The partisan model left to right, Republican to Democrat. The turnout model, low turnout to high turnout. The heat map showing how these models appear. Now if we’re targeting we want to go after the low turnout Democrats and the high turnout moderates using these models. I think you might agree with that logic. And then we look at what actually happened, the heat map of the actual contact with green being contact or red being noncontact and you can see that there was more contact in the areas where there should have been in Ohio in 2008.

Again the tools and the coordination allow for a more effective campaign and Obama won in Ohio compared to 2004 where you can see the contact was all over the map and heavily focused on the Democrats and many cases the Democrats who were already going to vote which was a waste of money. So again I think it’s just a graphic presentation of how this stuff can be made to work. Similarly North Carolina, looking at the turnout model across the bottom and the actual turnout on the bell curve, shows you that the turnout actually tracked the model fairly closely with one exception, this area where the predictor said they wouldn’t vote but they did. So you might guess who, what kind of voters those are. Basically Black voters in North Carolina voted for Black candidate where they had not previously voted.

Again this shows the progressive footprint of voter registration contacts and ID’s where you can also see those islands of noncontact. And it’s fascinating when you talk to different people in the states and they’ll readily confirm or – and agree that this was exactly what did or didn’t happen in their state. First time I showed this was in Tennessee. And the whole reason to have me out there was to talk about the failure in the 2008 campaign to mount an effective campaign while the whole country was going Democrat and they were electing Republicans across the board. Similarly we use these models for advocacy where traditionally when you’re advocating an issue we might have a mindset that says well I’m from Indiana. You're not going to advocate pro choice in
Indiana. That would be wrong because there are pro choice people and pro environment people and pro healthcare people everywhere.

The question is by modeling can you find them on an individual level instead of these typical geographic ways we might do it. I’ll just run these at the same time and show you how we zero in on actual people who are healthcare reform activists used extensively in the 2009 healthcare debate or environmental activists. Here again Nebraska, not a big liberal state as you know and the environmental activists there is in Arkansas. And the reality is that the organizations that are pushing issues, environmental issues and healthcare issues could use this data to find people everywhere to pressure members of Congress, everywhere on these issues. So you may not, you know there may not be a preponderance or a 50 plus one percent view on an issue in that state but you can still create pressure from voters to the members of Congress on these issues.

Similarly we use this in fundraising. We’ve created a progressive donor model that basically has been able to show organizations that if they use this they can spend less money per donor and net more by using the progressive donor model. We showed one Democratic organization that if they took their previous two million prospects that they mailed to, ran the donor model against it and cut off the bottom scoring half million pieces, huge savings, it would’ve only cut their donations by 16 percent. So essentially this stuff works. Is it the level of perfection that you want? I’m sure it’s not. Is it better than it used to be? It sure is. I can tell ya every year it gets better, every year it can be more precise.

The way we have academics using this around the country is we have an academic subscription where we have a number of universities, Stanford being one, that have access to our data. They have online access to the data to run unlimited counts, queries, cross tabs of information. We have a number of colleges that send us survey data to match. You know we can argue whether or not that’s worth doing. I happen to think it is. I think that having a lot of information appended to that survey or starting with that information for your survey makes sense.

The pollsters that subscribe to this kind of data, and we have 13 of the most prominent Democratic pollsters are subscribers; they want to make sure they’re only sampling people and polling people that they want to poll. Secondly their questionnaires aren’t nearly as long because they know a lot of the information about that voter before they talk to them. They don’t need to ask the preliminary
questions which make the cost of the survey so high. So they may spend more money to get the right sample and less money implementing the poll. But it’s been very popular, it’s used all over and they take advantage of the models as well.

So that’s basically my spiel and this is my contact information if you’re interested in any particular questions. But I happen to think this is a huge database, it’s hugely valuable. I can’t tell you that it’s going to solve all the concerns that the academic community wants for your particular kind of studies but there’s an awful lot of information on it and there’s an awful lot of things you can do with it. For the purposes that we build it, is there a bias because we’re progressive? There’s a bias in who we work with. The data would be built exactly the same if we were doing this for both sides or for the other side.

You try to do the best job of building a comprehensive database with as much information as possible so the clients can make the best possible decisions. You know there are lots of adages we throw out in this business saying you know you pick your cherries where the cherries are. I’ve always thought that was relevant because this helps you pick the right cherries and maybe know which ones are sour and which ones work. You know there are a lot of things that this allows in political communication in the nonprofit and the academic world. A lot of things that you can do with this that weren’t able to be done before. You know whether you’re looking at the impact of redistricting, looking at old districts, new districts, you want to look at the impacts of race on elections or turnout or race by income or race by gender.

I mean we run – we have people running cross tabs on this data all the time. Yesterday I actually got – any time somebody calls about campaign work after the election it’s kind of annoying given what we just went through. But I had somebody call yesterday and they needed to run some counts because there’s a runoff in one of the congressional districts in Louisiana and I was pleased by the kind of questions she was asking. I was running cross tabs on ideology by party and ideology by turnout and partisanship by turnout to let her figure out which kind of universe she wanted because what happens in politics generally is people have a budget that governs the decision making process. I have enough I can afford 10,000 pieces of mail or 50,000 phone calls. So you want to take the best 50,000 households or the best 10,000, whatever that number is. And this data helps you whittle those universes down to get to that universe. Yes sir.
Cliff Young: Let’s just say – Cliff Young. Let’s say you're giving talking points or a series of questions to a naïve consumer, let’s say I’m the naïve consumer of databases. And I have to choose between three vendors and you're one of the vendors. What are the, what’s quality in what you do? If I wanted to evaluate the quality of what you have to offer versus let’s say your competitors, what would be those talking points, those questions I should ask? What should I look for? You talked about lift, you talked about other things and I’d like to hear from you what –

Bob Blaemire: As one who does the marketing I can pretty well tell you what – at least what I try because the quality has been a huge part of my life for the last 30 years is trying to get it better. One of the reasons you want to do that as a practical matter, people yell at you less. The better the data is the easier your life is. The better you know, you only want to be yelled at for mistakes so often. A lot of it’s anecdotal; you know people that are pleased, people that find the things that work.

In comparing vendors you want to know things like how recent the files are. You want to know the percentage of the phone match. You want to know how often they’re updated. You want references. I mean this business references are huge. Who’s using it and who’s pleased with it? I’m getting anecdotal stuff all the time. I can tell you that early in this business I used to always use ethnic surname dictionaries to identify likely ethnicity. And I can argue that ethnicity is not as important as it used to be earlier in our republic.

You know there aren't really that many Italian communities where everybody votes for the Italian anymore. But there’s a lot of it that is valuable and particularly in the soup for modeling a lot of it’s valuable. And I can tell you that you know, the limitations that I’ll discuss, I mean Shaquille O’Neal is not Irish. Right? Robert E. Lee was not an Asian American. There are lots of limitations of ethnic data. There intermarriage, there are – I had a candidate running for Congress in Florida, a Cuban American, Rosario Kennedy married to the mayor of Miami at the time, David Kennedy.

So there are huge limitations in using that data but again it’s better than not having it at all. It may not be the perfection that you want but in terms of having it and not having it, I’d rather have it. I’ll have people that’ll call and say, “This stuff’s right on, this stuff’s accurate.” People that took our unregistered for instance that are scored with the partisan model so they could try to register people.
that were likely Democrats coming back to me and saying it worked. These people were Democrats.

So glad you had that model. So it’s anecdotal, it’s very hard to make a strong quality argument without knowing what the other database looks like. I used to often defeat some of my competitors by knowing them and I would often say that the client, “Ask them what their most recent registration date is on their file.” And if mine was two weeks ago, that’s pretty current data.

And I know some of the vendors aren't getting every state. They’ll say they have every state but when you ask them about Mississippi you find out it’s ten years old and there’s a reason for that. I mean Mississippi is the third world of data processing. Yes sir. I’m not sure I fully answered your question. I think maybe to a degree.

I’m Scott Keeter. I have a question about when you will have updated turnout data in your database for the whole country from this election.

Can’t tell you for sure. I could tell you what we did in 2010. We had it done by March of 11. You know you're not done – you're only as fast as the slowest state. I can tell you that again like everything else in this business it’s gotten better. The absentee early vote process which you know is virtually nonexistent before 2008.

I mean absentee voting has happened forever but absentee databases were not available forever. They’re always available possibly on the file after you got it but not on a daily or regular basis. That started in the 2008 cycle. This year, this last three months we’ve been getting 36 states every day and updating it and being able to produce reports online, which got a lot of press actually, we’re getting shot at by a number of Republican organizations the other day saying that we were predicting far too many Democratic early voters in Ohio and if you saw the reports the other night, they were accurate.

Their early voting was very Democratic but the short answer is the states are getting a lot better at compiling the data quickly and accurately and making it available. Some of them go online with the counties. North Carolina is example. The reason we got North Carolina, they are on real-time with the counties so when the country makes an update the state has it.
There are other states that are on batch process with the counties so they accumulate it and you know, other states, it depends on what kind of budget they have in their county or state election board. This is a human process. I mean one of the problems of quality of data, an example I’ve often used is let’s say I get a voter file that I know 50 percent of the people voted. I get the voter file and 25 percent show up as having voted. Now does that mean I’m missing half of them? I have 25 percent that are wrong or a combination of the two? I can guess but I don't know.

We are very much at the mercy of a human process that involves different rules and different budget practices by states but so far it’s the best available to be able to put a database together of this nature for the purposes that we do it. But I think probably March at the latest, possibly February we’ll be all done. I know I actually have had in the last about three weeks I’ve had a couple of different academic and GFK contacting me about how quickly they could do a match to find out who voted. Yes sir.

Bob Belli: Bob Belli, first of all I noticed that blue dot over Lincoln Nebraska. The question –

Bob Blaemire: I used to be the vendor the Nebraska Democratic Party and I tell ya some of my coldest memories in the world are being in Nebraska in the wintertime.

Bob Belli: It’s not so cold anymore with global warming.

Bob Blaemire: Bob Kerry was my client.

Bob Belli: I was really intrigued by this notion that you have a database concerning sporadic voters. That database could be very, very rich in terms of I think Jon would be interested in this as well, trying to model cognitive processes in terms of people remembering their past behavior. I am – one of the things when I thought about how wonderful it would be to have that kind of a record is problems that would be associated with tracking people who have moved and I’m just sort of wondering how well this database may be able to do that.

Bob Blaemire: Well it’s a tough – that’s a good question and it’s a very hard one because I personally have come to the view that moving without filing a change of address for the post office should be a capital offense and I would be far better off. We are regularly updating national change of address data. We change your mailing address when we have a new address for you. We don’t change your
registration address until the county or state election boards give us a registered address. I mean theoretically what you want is if I change my registration from Maryland to Virginia, once I become a registered voter in Virginia my Maryland voter record goes away even though all that data accumulates. That’s why I often have people with party ID in states that don’t have party ID. But we try.

We know that only a percentage of the people that move actually fill out a change of address. Because we can’t identify everyone who has died and everyone who has moved and the death index from the Social Security administration is what we use to identify deceased. Now the problem with that is the death index if you’ve looked at it, it comes the name, the last known ZIP code, the birth date and the death date. Well the only relevant pieces of information there for us tell us what state they’re in, at least what state they died in, their date of birth and their name and address. So we match the name, not their address, their name, their date of birth and the state to encode someone as deceased. We know we can’t get everybody. Similarly with moves.

We know not everybody files a change of address so as a result what we did was we created a model called deadwood, the people that we think are not there, that the model says are probably not there. Again it’s for practitioners. Don’t mail or phone people who are likely deadwood. You want to worry about wasting money, you know guy goes into the doctor and says, “Doc it hurts when I do this.” He says, “Don’t do this.” It hurts when you spend money on people that are dead and moved and so we try to help people not do that. It’s very hard.

We don’t throw away anything though. Our matching database that we match against includes every address we’ve ever had for you, any phone number we’ve ever had for you, any e-mail we’ve ever had for you, all the data, it’s huge. We entered into a matching competition last year called the Miter Challenge which is international, and involves government contractors, IBM, a lot of companies on matching technology and matching technique and we came in second. We do a really good job of matching. It’s one of the things I have to tell you having been one who has sold matching service for years and years, I’m glad I wasn’t competing with Catalist when I was doing this ’cause it’s really good.

We have very high match rates. And the accuracy is very high. It’s not perfect and I really wish we knew every time somebody moved or every time somebody died. I always say to somebody, “If you have a big database, I guarantee you there are dead people
on it and I guarantee you there are more tomorrow.” It’s kind of the way it works. Yes sir.

**Audience:**

I want to follow-up a little more about these kinds of data quality issues. You're probably in a terrific position to comment some on the inconsistencies between databases. So you said earlier you buy from the commercial database and presumably you're going to get a lot of records that have information on characteristics that you already have information on from some of the other databases and you might have race ethnicity and you also have it from the voter registration in some states. Can you give us some flavor of the nature of these problems and how you deal with them?

**Bob Blaemire:**

Yeah it’s tough. There’s no one rule that fits all that’s for sure. We know for instance when you come from a database like Experian or Axiom or Infogroup there are a lot of people that fill out these warranty cards and they put their pet’s name on it or their infant’s name on it and periodically we get some bad press because an organization will be working with will be doing voter registration and we’ll send a piece of mail to somebody’s fish. That actually just happened few months ago in Virginia from a voter participation center that we work with sent some mail and got press on one piece of mail out of three million they sent because it went to somebody’s fish. Now I’m not sure the fish wasn’t registered myself. The jury’s still out on that. But there are data quality issues clearly.

We’re able to see in looking at Infogroup’s data for instance what data fields are reliable and which aren't. We find for instance the birth dates on a voter list are far more reliable. People tend to tell the truth when they fill out their voter registration. It doesn’t mean that the data entry process has been perfect but they tend to tell the truth if they want to vote. You know try to vote and have different information on your – than what’s on the list. It varies.

One of the things you try to do is you try to standardize everything the same way, by running through address standardization software you clean up a lot of the addressing issues. But I mean I’ve looked at databases where you see somebody’s record come in and the first name is Goerge. Now is that George spelled wrong? Probably. You know you just know that’s there. Yet on these commercial databases you see people’s names coming in that have four letter words in their middle name.

Now you have to assume that’s probably not accurate. It varies. It varies a lot. But again we’re dealing with probabilities, not
perfection. You try to get better, you try to find out what’s not workable, you try to make whatever is not as good as you want to be to make it better so it’s constantly tweaking it.

We have, you know we have a really good department of analytics people that not only build the models but spend a lot of time on the data quality issues, a lot of time on that. We have measurements what the state for instance reports on vote history and then what we get, how close is it? In fact our release notes that we produce on every state if you're online you want to look information on the state we have what we call release notes and it shows you what the state reports and what we have. And if they're not close we usually will go back to them. Sometimes they just say sorry.

_Audience:_ But in your merged database, do you keep the inconsistencies? So if I’m an academic and I pay to access your data, or do you – I mean what happens? So you have two different ages from two different sources –

_Bob Blaemire:_ We would not replace the vote filed age with theirs. If somebody doesn’t have age on a voter file and we have it on a Infogroup file we would pass it over. Again it’s better than nothing. We know that their date of birth data is not as accurate. A lot of times it’ll come in with just month and year for instance.

_Audience:_ But do you leave the age from Info U.S. in your database?

_Bob Blaemire:_ Um hmm.

_Audience:_ You do so the inconsistencies are in the database.

_Bob Blaemire:_ No, no we don’t – we don’t keep two dates of birth for somebody for instance. We don’t take the date of birth if we already have it.

_Audience:_ Is that true for all variables or you have a different race in two different sources?

_Bob Blaemire:_ I can’t tell you for sure. I would have to look – I mean 700 fields is a lot of stuff. I would have to tell ya and I don't know, I mean their database doesn’t have a lot of the fields we have for instance and a lot of fields that for instance Infogroup may have, they may have gotten from us. They shouldn’t have but I mean political party is something you only get from a voter list. Data registration you only get from a voter list. Precinct, and we try to keep previous precincts and the current precincts.
I tell you, redistricting, one of the huge pains in the ass in this business, redistricting. It ought to be outlawed. We shouldn’t do it. It was very, very hard. We’re all done? I’m sorry folks. But seriously if you have other questions, send me an e-mail, I’ll be happy to try to do what I can to answer it and if I don’t know the answer I’ll make one up. Let me get rid of this. Bingo.