Episode 1
What Makes Big Data Big?
Host: Dr. Alex LoPilato

How to cite this podcast:


This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Dr. Alex LoPilato: Welcome to the Work Science Center Podcast, brought to you by the Work Science Center of the Georgia Institute of Technology. I am your host, Alex LoPilato. You can find more about the Work Science Center at our website www.WorkScienceCenter.Gatech.edu. In today’s podcast I talk with Dr. Fred Oswald, professor of Industrial-Organizational Psychology at Rice University about big data and its influence on the study of work psychology.

Fred, I would like to thank you for agreeing to be the first ever guest on the inaugural Work Science Center Podcast. I will go ahead and start us off with our first question.

Dr. Fred Oswald: Great!

AL: What is big data? What makes big data big? I kind of want to have a definition for our listeners to go with before we start this off.

FO: Sure. Well, first off, thanks for having me on this podcast. It really is an honor. I hope people listening may glean some insights and keep their ears turning on an important topic of big data. I am so glad to be supportive of the Work Science Center. This Center promises to stimulate a lot of new ideas, bring a lot of different people together hopefully to improve our collaboration and efforts on important workplace issues. So, thanks for that.

To turn to your question, there are a lot of papers out there that define big data. You know, what is this thing called big data? Sort of like going to the zoo, what is this animal? A framework that is often used is the Vs. There can be 3 Vs, 4 Vs, maybe more. Maybe I will use that as the starting point on the definition, but only to really say that there is not a particular definition. The Vs are useful to get us thinking about data in general, whether it is big or small. I think that is important. Anyone out there should not think it is black and white and say ‘Oh I have big data, therefore I can use big data analyses’ or ‘I don’t have big data, so I can’t’ or ‘I’m not sure.’ It is really about understanding your data and applying methods that are appropriate. We will get to methods later, I guess.

To get the data themselves, certainly the V of volume matters. Data can be streaming over time, across a lot of people, or both. Another V is variety. There are tons of media through which data can flow. Maybe brain scan data will be common in the future at work. We are not there now but, we do have sensors for workers that are increasingly being used. Whether that means you are scanning your ID to get into a room and that is being sensed. Or you are being sensed for your physical presence in a room or your proximity to others. Or perhaps even...I have visited folks working at NASA and they have sensors being put on astronauts for their physical health and getting a sense of their wellbeing at work as they are doing their work. So, variety is currently an important V. The volume is important. And then, the velocity. How fast are these data coming in? And, concurrent with that perhaps is how fast should we be doing our analyses? Where should we be putting our cut into the data? And then finally, the point about veracity. People talk about veracity. I think often the folks talking about veracity may not be thinking psychologically, they are thinking about the information contained in the data, like how much static is there in the data versus the signal. What signal is in there in the noise? But, from a
psychological point of view, there is another layer which is how much of that signal is psychologically meaningful? You might be able to reliably get at the veracity. Somebody is definitely in the room. Or somebody definitely swiped their card and they are at the meeting. But from a psychological point of view, how much of that signals that an employee was conscientious? Well, that is a different sort of question but still is an important one as we try to get meaning out of our data.

AL: Yes, so it is almost like there needs to be a V for validity in a sense, for big data in the realm of psychology.

FO: Oh yeah. Exactly. More Vs.

AL: Fitting with the theme.

FO: Yeah.

AL: Well, so that leads me to a question, I, myself, have kind of been mulling over for quite some time. You know, the difference between big data and just pure dustbowl empiricism. So, if I have a giant dataset and I just happen to find that one variable is significantly related to another and I use that as a predictor without any real theory behind it is that still not necessarily a safe thing to do? Or is it?

FO: Well, so I would say that there is definitely dustbowl empiricism in big data but I would say it is of a different nature than what we have historically faced, and we need to think about it perhaps differently as a result. So historically our dustbowl empiricism, the kind we worried about, was capitalizing on chance. If somebody comes to you and says the correlation between, pick any two kinds of variables, conscientiousness and some kind of work performance, the correlation was .9. You would know from your experience that that correlation was too high, just from what you would expect. So you would look a little closer. If you found the data were based on just six data points, or rather six employees gave you data on their performance and conscientiousness, that would be a problem. You are capitalizing on chance, if anything. So it would be too high. If you had a larger dataset, then you would not be over fitting and the correlation would be lower. Right? So in the big data sense, capitalization on chance is not as much of an issue.

Big data analytics, the core of those analytics is not to capitalize on chance, it is to develop models in part of your data and then verify or test in a new set of data. So, to the extent you find relationships holding in fresh data, based on models you developed elsewhere, then those relationships are cross-validated or otherwise robust. So what does dustbowl mean in the context of big data? What it means is your relationships are robust. They just may be mysterious. So in other words, you might find these small effects or maybe even small effects that cumulatively add up, but you may not know why they are there. You may not even know where they came from. Some big data models, without getting into the technical details, they involve all the variables in a dataset in a way where you are not even sure which variables are responsible, and
yet you get improved prediction. So, that sort of dustbowl or black box notion is still in big data. But, what is off the table is capitalizing on chance. You are not capitalizing on chance so much given the methods that are there. It is a different type of dustbowl we are dealing with. And we are struggling with this not just in psychology.

DARPA, the US Government, is funding a big initiative through DARPA, the military agency. DARPA stands for Defense Advanced Research Projects Agency. But they are funding a project where they are trying to understand what is behind big data algorithms. So in other words, the algorithms achieve the “what” of prediction. So they improve in our language, $R^2$. They improve, they reduce mean squared error. But they do not get at the “why.” How did it get there? Why is this prediction evident? So they are spending a lot of money across over 10 projects on dealing with how do we extract meaning from big data prediction. And that basically gets at your question, Alex, about dustbowl. How do we figure out, when there is dustbowl, why it exists? What’s the meaning behind it?

AlphaGo is another example where, you may be familiar with this, where a machine learning program defeated high level human experts at the game of Go. And that is a relatively contained problem, but it is a very complex one. So what we have used are so called deep learning algorithms and the advances being made are not only in terms of prediction that the machine wins in these games, but also the meaning. What does intuition mean to these algorithms? What does it mean when an algorithm takes a certain approach, and can humans learn from that? So these experts at Go are actually trying to learn, trying to extract their own meaning from these algorithms that may not confess to what their strategies are. They are trying to figure that out. So we have a lot of work to do on the organizational side to extract meaning from big data analytics, whether it is a dustbowl kind of problem or not.

AL: OK so is it fair to kind of summarize that with the dataset we have traditionally been used to, there is a lot more uncertainty in the parameter estimates or the strengths of the relationships as opposed to just the pure existence of a theoretical relationship? Whereas big data, it seems the uncertainty of the parameter estimates has gone down quite a bit, but now there is more uncertainty surrounding why these empirical relationship exist in the first place?

FO: Well, I think the uncertainty of big data modelling is captured better. So, I would say that we are more certain, I hate to be confusing here, but we are more certain about our uncertainty. Right? Because we have actually tried to gain support for our models with external data. So historically, researchers have been sensitive to this problem, and I would say that I-O psychologists have been especially sensitive.

They have written papers and thought about so called cross-validation. So, a regression model gets developed in a sample or in one part of a sample, and then a fresh set of data gets put through the model after it is developed, after the coefficients are established, to see how good predictions are in a fresh set of data. And, what happens naturally is that $R^2$ has to go down because it was optimized for that particular original sample. It has to go down in new samples, mathematically, it just has to. The question is how much? Psychologists have done cross-
validation in the past, but not frequently. Even in our top journals today, how often do you see cross-validation? Or, you can also take a formula approach and kind of estimate what would the prediction be in new samples. We do not do it as often as we should.

By contrast, so called big data analytics, which covers a lot of different methods, does this all the time, inherently as part of the analytic approach. So, again to answer your question, we know the uncertainty better. We never know uncertainty precisely. We never know the future. We never know how models will work in future data. But, these analytics put the data through more torture test to try and get at robust relationships.

AL: OK. So there is clearly power in big data. There is a lot to be learned from it. How have I-O psychologists applied big data in either research or practice?

FO: So, my area is more focused on personnel selection and testing. HR types of application. I cannot pretend to know the wide world of I-O. I can speak to some areas outside of mine. Some things that have come up recently have been, for example, in the data mining arenas. Considering how resumes can be mined for their themes within the text that is provided within resumes and that can be used for more structured types of predictions than relying on human intuition. Getting people to read the resumes and kind of guess at, they are educated guesses, but you get hiring managers looking at resumes to make decisions. The question is, can you go beyond that and mine the text? Learn from a body of resumes algorithmically to make better decisions about individual resumes?

So that is one arena. Another one is extracting themes from social media, which brings up issues of privacy and ethics, of course. But the question is, can you learn about the person, and therefore the employee from various forms of social media? That has come up. There are also applications where you happen to have very large employee datasets where you have applicant data and then you may have a whole wealth of criterion data, much more than just the annual performance review, where you can look at these relationships in large datasets from a big data perspective. Maybe mine some predictions that you would not have gotten from more traditional methods. I will say, there in the data mining arena, one point I did not make about dustbowl is that you could worry about missing by being robust. You could overlook more specific relationships that could be there, that gathering targeted data would end up figuring out. So in other words, I guess the point is, just because you find robust relationships in big data, does not mean…you should still try to improve your data sets, use good data, not just big data, to get at more refined relationships. No dataset contains all of the relationships that may be of interest that might actually be there if you had good data in your data set. What I am trying to say is that, sometimes we think of big data as a definitive approach, or we lean that way, and we should not. We should always approach data as a scientist, as a psychologist, with healthy skepticism, and with our own expertise in interpreting what comes out of those analyses with the particular data we happen to have.

AL: So there is still room for psychometrics. No fear of that field being wiped out by big data just yet?
FO: I do not think so. I addressed that exact issue at the last SIOP conference. I contrasted methods in the big data arena for clustering items, with the more traditional psychometric modeling approaches, whether that is confirmatory factor analysis, or Chronbach’s alpha, our old friend. And basically, if you develop good measures the signal comes out either way. The psychometric approach has the benefits of extracting meaning in ways that big data does not know, necessarily, that you have developed a set items in a systematic and meaningful way. It just treats it as undifferentiated data, where other data might be in there with those items. So you do clustering and it does not necessarily know. Whereas a more confirmatory approach with psychometrics will tell you more about the nature of items and item refinement. So, no. That is not going away. I think it becomes, I do not know if I would say more important, but it becomes a more distinctive tool indeed in order to ensure that, again going back to good data. To ensure you have good data going into big data through well-developed measures that are informed by our more confirmatory approaches or even exploratory, through exploratory factor analysis. But, more systematic approaches to measure development is an important supplement to big data. A critical supplement to big data in some cases.

AL: I completely agree. So kind of going from there, what you said about using big data from a selection perspective, a lot of what I have read about big data, organizations seems to use it pretty heavily for making better predictions for who to hire, who is going to quit and so forth. However, the Work Science Center is pretty interested and focused on the experience of working for the worker. So I was wondering if you have any ideas of how big data might change the day to day experience of working for people? One thing I have kind of thought about, is if you could real time monitor people’s stress. I do not know if that would violate basic. If the employee would be uncomfortable recognizing that they are being monitored like that. Or that there might be a benefit either in the health sector or the health of trying to intervene when it looks like the employee is getting stressed out in some way. Or, even just from a continual learning perspective, if you could measure how day to day the number of errors employees are making and so forth is going to help them develop as a worker. I wonder if you have any thoughts on this?

FO: Yeah. Well, I think that that is interesting. As the Work Science Center monitors the aging workforce, and brings together experts in the field that are thinking about aging and work, having their, keeping their eye on the radar for new technology for older workers, and applying their research sensibilities to those data will become, I think, increasingly important. So, for example, think about the Fitbit. Where we have monitored our fitness over time. I think going to the privacy issue is the idea that these things change over time, and they change over settings in terms of their acceptability. So, sometimes the Fitbit might, if you only have one worker using a Fitbit because it is new to the organization somehow. That is a different story than if everybody has heard of Fitbit and half of them are using it and the other half is looking at them and thinking well, maybe I should. Or there is a strong organizational policy in place where you just kind of accept it and everybody is doing it. Those sorts of factors are important to some psychologists.
Big Data: Interview with Dr. Fred Oswald

March 27, 2018
© Copyright 2018

and need to be considered as contextual information around the big data you might be collecting from sensors. Right? To know, what is the context around which you are collecting your data?

You know, for the aging workforce, there is an interest in real time wellbeing and stress, trying to intervene, and improve a workforce setting. Right? Whether you are young or old, there is always an interest in wellbeing and engagement and conversely less withdrawal, less attrition, less stress. The issues will not go away. The technologies will keep improving in terms of sensing. I think what we do with those data will continue to be challenging. When to intervene? How to intervene? That will always be the case.

And then, again, going back, there will always be the cultural issue. What is the work experience like once you have these sensors in place? Is it a good thing or is it a bad thing? Or it just different? Eventually we take our technologies for granted. And some new technologies will, like anything, be taken for granted. So all of that is interesting and important. Hopefully the center can take advantage of that.

**AL:** Yeah, that is interesting. One thing I have talked about with Dr. Kanfer, the director of The Center is differences. So older workers have not grown up with social media, Facebook and so on. So they are more used to privacy at least when it comes to the digital world. Whereas younger workers, you have workers in the labor force who do not know before Facebook.

**FO:** Who do not want to know before Facebook.

**AL:** It is interesting to think that monitoring individuals at work, for younger workers it might just be the same thing different day. Privacy, at least in the digital realm, is not something that is there for them so much. It is not something that is on their radar. Whereas older workers, you might get a lot more pushback.

**FO:** Right we also do not know, or it is evolving, the state is evolving where we are beginning to learn, what is it like for workers who have never known anything other than a world with the internet to be working over long spans of time. The Work Science Center can kind of monitor this and track what does the need for privacy mean in a modern age and how will that change over time as media comes in different forms. I know, to my niece and nephew for example, they are 15 and 13, and they just do not know any different. The world has the internet, Facebook is actually passé and it is more about Instagram and texting. I do not know, I think there are workplace implications already that I am sure I am unaware of.

**AL:** It is incredible how fast moving it is, honestly.

**FO:** It is a real opportunity for the Work Science Center to think about research. I would hazard a guess that practice is really ahead of research in terms of trying to guess and create what the workplace needs to improve worker engagement and productivity and so on. And, the practice will benefit from the research side or the academic side in taking systematic scientific approaches to the tools and the data they are collecting.
AL: That is an interesting point. I have been wondering recently if I-O researchers have kind of missed the boat, or at least missed the first boat on big data. Whereas practitioners have maybe by virtue of organizations being quick to adopt it, seem to be much more involved in the big data landscape. Or, at least the ones I have come across.

FO: Yeah, I think that is right. I have that sense as well, Alex. Part of me wonders, I am just, every generation experiences this idea that practice is just ahead of science in terms of what is going on in terms of the workplace. I think in the big data arena, maybe that gap is more pronounced, or it sort of has a distinct flavor that is worth looking at. Where the practice side is...to get big data to work in any application you have many disciplines having to work together, one would think. And that sort of collaboration and technology is very different from the way academics are approaching big data as an issue or something of research interest. So there needs to be more partnerships in order to make our science relevant. Again, that has always been true in the past, but that certainly is true for big data.

AL: I agree entirely. I think we are actually coming close to the time we have for this podcast. Is there anything else you would like to say about the role big data could play in I-O psychology or work science in particular?

FO: Well, the data are not going away. You do not need me saying that, but we know that. Analytics are not going away, they just get more sophisticated. The toolbox is huge. I think research is always evolving to what analytics are going to work when, for what data. And, often the data we collect, how should I put this, the data cannot live up to the models in big data. What I could say as a parting thought is, big data will continue to speak very strongly in the coming years and even decades in terms of work relevant topics. Certainly, technology will continue to progress in the workplace, and the Work Science Center can monitor how technology affects the work experience and vice versa to some extent. Also, work is a more fluid notion with the availability of technology allowing workers to work remotely and flip across different jobs and the career boundaries are more fluid. I think that is an interesting place where big data comes into play in examining the experience of the workers, organizations, and even the workforce at a broad economic level. So, I think there are a lot of avenues for big data being relevant in the workplace and relevant to research, and a lot exciting avenues for the Work Science Center, so I think that the future is always interesting.

AL: Well, Fred, that is about all the time we have. I would like to thank you again for agreeing to be a guest. This has been an incredibly interesting conversation and I hope to have you on again soon.

FO: Great, thanks very much, Alex. I really appreciate being invited to this podcast and I am very excited for the Work Science Center and the directions in which it is headed. I would encourage anyone that is listening to go to the Work Science Center website and check out the information that is there at WorkScienceCenter.gatech.edu and get involved.

AL: Thank you so much.

FO: Thanks a lot, take care.