Calling DIBS

Episode 72: "How Behavioural Insights & Data Science Intersect"

with Craig Hutton, Senior Behavioural Scientist with the BC Behavioural Insights Group (BC BIG)

In a great BI project, the behavioural and decision sciences provide the questions and data science provides the tools to answer those questions. As someone who is both a data scientist and a behavioural scientist, Craig Hutton is the ideal person to talk us through similarities, differences, and complementarities between these fields. Craig also shares best practices for managing, analyzing, and communicating about data.

Transcript:

KIRSTIN APPELT, HOST: Welcome to this edition of Calling DIBS. I'm your host, Kirstin Appelt, Research Director with UBC Decision Insights for Business and Society, or DIBS for short.

Today, we're calling DIBS on Craig Hutton. Craig is a Senior Behavioural Scientist with the BC Behavioural Insights Group and he's a recent graduate of the Advanced Professional Certificate in Behavioural Insights. He's a Ph.D. in psychology, and has spent quite a few years working as a data scientist, so I'm looking forward to getting nerdy and talking data with Craig. One of my favorite topics and a great person to talk about it with.

So welcome to the podcast, Craig.

CRAIG HUTTON, GUEST: Thank you very much, Kirstin, for that lovely introduction. As you mentioned, I'm happy to be with you here today and joining you from the traditional, unceded territory of the Musqueam people where I live, work, and learn.

APPELT: Wonderful. Thanks. And maybe can you just start by telling us a little bit about yourself?

HUTTON: Sure. As you mentioned, I am a Senior Behavioural Scientist with BC BIG and based out of Vancouver. Outside of work, I like to do hot yoga and go skiing and hiking when I can.

APPELT: True Vancouverite.

HUTTON: It grows on you.

APPELT: I love it. Well, I'd love to start by hearing a little bit about what brought you to behavioural insights. As long-time listeners know, I think it's so fascinating to see the either winding or direct paths people took to BI. So what was yours like?

HUTTON: Sure. I think like most BI practitioners, because it's still a pretty new field, my path had some twists and turns to it. So, I've been doing behavioural research in some capacity since I started volunteering in a psychology lab, as second year undergrad at University of Winnipeg around 15 years ago, kind of dating myself there. And then after undergrad, I did a Ph.D. in Behavioural Neuroscience at McMaster University out in

Hamilton, Ontario. And then I moved out to British Columbia for a post-doc in Neuroscience at the University of Victoria.

And that's where I started to augment my behavioural research with machine learning. So, for example, I developed a predictive model to diagnose post-concussion syndrome even years after somebody had experienced an injury. And while I was kind of working through the postdoc, I kind of decided, “Well, maybe a prof isn't exactly what I want to do”. You know, profs mostly spend their time applying for grants. They're not really in the lab that much, and I wanted to keep doing research.

So, I had this opportunity to do data science professionally for the BC government, and I did that for over four years. And while I was working as a data scientist for BC's Ministry of Social Development and Poverty Reduction, I pivoted into BI by completing the professional certificate program at UBC Sauder last year, and this really rekindled my passion for experimental psychology and helped prepare me for my current job at BC BIG.

So in some ways it feels like I've come full circle, perhaps with a few extra tools in my toolbox. And you might say that my path has been more like climbing a spiral staircase instead of following a straight line.

APPELT: Yeah, I love that. And I think, like you said, that it's doing those turns and twists where you pick up the complementary skill sets, that if you're on a linear path, you might have just the one lens. But by doing the twists and turns, you have all these different lenses, which I think makes the discipline and each person's work so much richer. So that's really cool to hear.

So, after your Ph.D. you mentioned you spent a few years more focused on the data science piece. And I think this is an area that's of real interest to folks because data is everywhere these days. But then I think there's different ways people understand what that means. So how would you define data science?

HUTTON: Sure, and as you kind of hinted at, defining data science is actually trickier than it seems. Because if you look at the job ads that are out there these days, there's a lot of variation in terms of what employers seem to expect data scientists to do that overlaps with more traditionally thought of as being associated with data or software engineering roles.

However, in research focused roles, which is where my experience is, data science is simply the practice of using data to answer research questions. So in government, data scientists are often tasked with generating insights to support evidence-based decision making, evaluate policies and optimize resource allocation.

There is this unfortunate tendency, however, among junior data scientists, particularly those from non-science backgrounds, to focus their training and attention on fancy cutting edge machine learning algorithms which are sometimes useful. But, you know, it's far more important to make sure that you're asking the right question, which is the one that matters to your client, and that you have the right data to answer it before you can decide how to analyze the data. And the truth is that many research questions that businesses have, can be answered without machine learning or AI, which is fine. The point is that they have questions and hopefully you can provide answers. And this is why I always try to emphasize the “science” part of “data science”.

APPELT: I think that's so important. I see that same issue cropping up, and I think that's where, to me, this idea of garbage in, garbage out. Like if you put in data that doesn't answer the question, you'll get an answer, but it doesn't address that question. It's an answer to a different question.

HUTTON: Absolutely.

APPELT: So I'd love to hear a little bit on your perspective of the similarities between data science and BI. And one thing that I think I've picked up on is that both have a lot of interdisciplinarity and combining ideas and techniques. Is that something you see?

HUTTON: Yeah, that's a great point. So, for example, the first challenge that I was faced with as a data scientist in government was to evaluate the impacts of multiple social assistance policies on the number of new income and disability assistance cases that we were seeing in BC and they gave me over 20 years' worth of monthly social assistance records to do this.

APPELT: What an embarrassment of riches!

HUTTON: I know it's a lot. I solved the challenge by borrowing a nonlinear regression method that was described in the Ecology literature called "Generalized Additive Modelling". And this enabled me to estimate the caseload impacts of multiple policies while simultaneously accounting for seasonality and other features of the time series data.

So yes, I would say that one of the common strengths of BI and data science is that both use techniques from multiple fields. And I think in large part this is because they are so interdisciplinary, which is clear even if you look at the diversity of skills and experiences among members of a single BI team. So the environment that data scientists and BI practitioners work in usually doesn't have the same boundaries and conventions that older disciplines like experimental psychology and economics do, where in those situations most professionals are working with others who have similar training and tend to favour a particular approach to data analysis, like analysis of variance in psychology or linear regression in economics.

So BI practitioners and data scientists are often, you know, faced with challenges that draw us out of our comfort zones, require us to think outside the box to address. And this encourages us to broaden our horizons and consider unconventional approaches to doing research.

APPELT: Absolutely. I think that's a really key point is just how we're driven by the questions. And so it frees us up to explore different methodologies and use the one that works and not just use the same tool like the same hammer. Not every problem is a nail kind of idea. What other similarities do you see between data science and BI?

HUTTON: Sure. One that I'd like to highlight is that both disciplines focus on client relationships and using data to make recommendations to stakeholders. So, we can talk about this a little bit more later. But communicating research and analysis insights to non-technical audiences is a key part of how both practitioners and data scientists add value to a client organization.

If you can't communicate your findings so that people who need to understand them or who can benefit from them, in the case of citizens, if you're working in government, then your work won't have much impact and you may find yourself struggling to maintain support from project sponsors or attracting new business.

APPELT: That's a great point. The insights aren't going to be impactful if you can't get them out of your head into someone else's understanding.

HUTTON: They don't share themselves.

APPELT: Even though we've tried sometimes.

Oh, yeah. And so that makes sense on similarities. What are some of the key differences?

HUTTON: Sure. So, the most obvious difference that I don't need to tell you is that BI focuses on behaviour, whereas data science is not domain specific that way. But of course, the differences don't end there.

Data science is also much more heavily influenced by software engineering practices like agile project management and organizing work into sprints. And nearly all data work is programmed, and we use version control systems like Git for record management. And here the biggest advantage of programming and why I advocate for it is that you can automate repetitive tasks and save yourself a lot of time.

As an example, a simple for loop to perform the same operation on a thousand data files saves you from having to do that operation yourself a thousand times. Another benefit of programing, data processing and analysis operations is that the process you use documents itself in the code that you write, and that code can be reused by future you and others to reproduce an analysis or audit it. Future you is very happy at the things that present you has done to save itself time later.

APPELT: Yeah, that's a funny one too, because in some ways it's a very behavioural insight that later you is going to need that reminder. And I see folks, the first time they do a dataset, they name it like "Dataset.xls" or whatever, and then "Variable 1" "Variable 2". Of course later me will remember “which one was Variable 1 and which one was Variable 2”. Three months later you're like, “I don't know what project it is”. Yeah.

HUTTON: Yeah. So another thing is that data science work also teaches you to clearly annotate your code with notes that help explain what the code is doing at each step. And these notes make it a lot easier for you to figure out what it was that you did before. So you can go back, like even I can go back to analysis I did years ago and figure out exactly what I did because the code's still there and it has notes attached to it.

And beyond that, I think it's fair to say that BI in terms of like the other perspective, BI places a greater emphasis on theory and the consideration of prior research. Whereas data science projects can be somewhat more self-contained and may not require much consideration of what else is known about the phenomenon of interest.

For example, if you're a data scientist who's been tasked with estimating how many times a website was visited and developing a dashboard to track that over time, you can do that to the satisfaction of your stakeholders without understanding or discussing the behavior of the individuals who are visiting the website. And the scientist might also develop a predictive machine learning model that accurately forecasts service demand for an organization so can ensure sufficient supplies are available to meet that demand. You can do this based solely on patterns that are observed in the data.

However, without a deeper understanding of the system, you wouldn't want to try intervening on it to change the demand, which is a huge limitation of “black-box” machine learning models like neural networks. This is why we say that “correlation is not causation”. There's some great cartoons about that online you can find. And why so much of scientific research relies on randomized controlled trials to test interventions, because we care about inference and causality.

And in contrast, BI focusses on the behaviours across various actors in a system and maps the steps that they go through in attempting to achieve their goals. And it augments quantitative research approaches that dominate the behavioural and decision sciences with deeply insightful qualitative methods that are more common in user experience research. And this is a major strength of the RIDE model that we use for behaviour change at BC BIG. It's also taught at UBC's Certificate Program because developing effective BI interventions requires a causal understanding of a behavioural process and how it's impacted by the barriers that impede an actor in the pursuit of their goals. Here, qualitative research can provide us with hints as to what's really going on from the perspectives of the people involved.

So I would say that a major difference between data science and BI is that data science tends to focus more on describing and predicting as research goals, whereas BI tends to focus more on understanding, inference and intervention.

APPELT: I love how you've summed that up. That's very clear and concise, and I think you also just raise so many good points about like annotation and how that helps future you, but also can help with things like replication across people and helps with the ethics around being clear about what you're doing in the data. And to get deep into data, but if you have outliers, why you excluded them, how you excluded them all that kind of documentation is so important as the field tries to have more transparency around how data is being used. And then the idea around the multi-methods. And that goes back to the idea of the interdisciplinarity of BI using the qualitative and quantitative methods.

So a lot of rich things in there. But what I'd like to transition to is thinking about how data science can further strengthened BI, I think we've already talked a little bit about this, but I'd love to hear about the different places data science can kind of filter in and add, and maybe we can start at the beginning of the RIDE model and talk about that initial scoping and exploration phase when we're trying to understand the context of the problem. What are some ways data science can help us with understanding the problem?

HUTTON: For sure, that's a great question. So if your project partner is asking you to figure out, for example, how to increase uptake of a service and they can provide you with some relevant historical data, you can use that information to figure out what the current service utilization rate is and maybe even use some predictive modeling to estimate what it might be in the near future if nothing else changes. And this can help you establish what the base rate of the target behaviour is. And just as importantly, how much room for improvement there might be, which can help you and the project partner figure out how to proceed with the BI project or if you should proceed at all.

Knowing where things currently are and how high or low the response rate has been in the past can be tremendously helpful in guiding conversations with the client about what success might look like if you were to proceed with the project.

For example, is a 5% increase in the response rate meaningful to them? And is that degree of improvement within the range of what's been seen before, or is customer engagement already at an all time high? So, this can help you manage stakeholder expectations and avoid overpromising or setting the bar too high for yourself. I always say it's better to under promise and over deliver than to over promise and under deliver.

A second example is that, you know, knowledge of more advanced statistical methods can enable a BI team to generate more realistic estimates of sample sizes needed to evaluate a trial. When you're doing power analyses for study designs that involve complicated situations where you might have longitudinal data or clustered observations, and then kind of a little bit related to that is the situation where you could use an unsupervised machine learning method called cluster analysis to take some existing data and information you have about that, you know, that sample or that population, to figure out if your target population is comprised of statistically distinct subgroups before you develop a BI intervention.

This approach is called customer segmentation analysis in marketing. You may have heard of that before. But yeah, that's another way that you can kind of support a BI project using data if data are available.

APPELT: Yeah, there's so much there. So I thought maybe we'd pull a few of those pieces out and talk about them a bit further. So I think you talked about base rates and how that can help you not over promise and under deliver. So I think that's clear.

But I'd love to pull apart the idea of power analyses because that can be a new topic for some folks. So what are power analyses? Why would we do them? What do they help us do in a behavioural insights project?

HUTTON: Sure. Power analysis is something that you usually do when you're planning a study and you need to know what the expected effect size might be. So this isn't something that you would necessarily do in all cases if you're trying something that's completely new, has never been tested anywhere. It would be very challenging to do it because you have to make some assumptions that an effect that has been observed before might translate to your current context.

So you would search the literature or maybe your prior research. Find examples of where a treatment had been tried, a BI intervention had been tried. Look at the effect size for that intervention and then use that effect size to estimate what the sample size might be, you know, in in your future study, to be able to evaluate if there's statistically significant difference between your treatment group and your control group.

APPELT: So then potentially we could use it, just to go over a few of the cases where it's used, you could use it to see if the sample size you have is going to be sufficient to detect a statistically significant effect. Or if you're in the position of being able to select from a large population, you could choose your sample size accordingly based on that knowledge.

HUTTON: Yeah, absolutely. And the really nice thing about this is it enables you to be as efficient as you can be in evaluating an intervention where you're not going to spend more of the organization's resources than you absolutely need to be able to evaluate this intervention.

APPELT: Yeah. And then potentially also it gives you the ability to be able to test other things. So, you know, you use the first 200 people, let's say, for the first test, and then you decide to tweak the BI intervention and you can use the next 200 people for that. It's a great point.

HUTTON: Yeah, absolutely.

APPELT: And then just picking up on the idea of segmentation a little bit. So, when you talk about subgroups, can you give any examples of subgroups that you've seen in some of your work? Of things where different interventions work differently for different groups?

HUTTON: Sure. So, this kind of speaks to what happens after a trial. After you've finished a trial, you might want to know if the impact of an intervention differed for some subgroups in the sample than others, like maybe for different age groups or for males versus females. And you would traditionally do this type of an analysis or evaluation using what's called interaction analysis or with a stratified analysis.

Well, it turns out that these kinds of interactions between a treatment variable and covariates like age or education level or gender can be more efficiently evaluated using a new machine learning algorithm developed by Susan Athey, who's an economist at Stanford University. The methods called “Causal Forest”. We don't have time to cover the methodological details now, but I actually used a Causal Forest in my capstone project last year to discover that the message we were sending to income assistance clients to try to connect them with Employment Services at WorkBC had a much larger impact on clients who had a high school diploma, but didn't really help clients who hadn't completed high school before.

So you can think of a Causal Forest as an interaction finding machine that helps you figure out if your treatment benefits some populations more than others.

APPELT: That's so interesting to see how those effects, sometimes we go in with hypotheses about subgroups, but then other times there's ones we don't expect. So that's really neat.

HUTTON: So, another technique that data science and statistics can contribute to BI is model simulations. And this allows us to explore "what if" or counterfactual scenarios. So, one of the things that we try to do in data science when trying to interpret complicated models is generate what are known as partial dependence plots. And these basically show you what happens to a model's estimates for the dependent variable or that feature you're interested in your outcome measure when you adjust one of the model's parameters at a time and hold everything else constant. Or you set those other parameters at specific values you want to explore.

So, in data science, these plots are most often used to show us what the effect of one predictor is on the dependent variable after adjusting for the other variables in the model. However, generating one of these plots is also basically like running a simulated experiment and asking the model a series of "what if" questions based on its understanding of the relationships between the variables in your data.

Example, if you fit a linear regression model, the data about house prices using historical records on home characteristics like square footage, the years that they were built and sold, the neighborhood that they're in, the number of bathrooms they have, etc., you can run a simulation to ask the model what the estimated 2024 selling price of a house would be if it were a 2000 square foot home built in 2009, it has four bathrooms and is located in North Vancouver. The same logic can be applied to the behavioural processes and with enough of the right data and the right model, we can generate simulations to estimate what the impact of a particular intervention of interest might have on particular behavior of interest.

And at BC BIG, just last week, I used this approach to estimate the impact of scaling an intervention from a study sample to a population of interest. So that kind of is related to both how you might evaluate a model that you're using and learn what it understands about the data and also use it to help inform kind of how you might recommend scaling or not scaling to generate a return-on-investment estimate for your client. And that can be part of how you tell the story and communicate your results.

APPELT: Yeah, that's a great point. And I think one of the neat things about it is it's talking about how a lot of times there's this rich data that's there, especially in government context, like you said, you had 20 years' worth of records. And when you have that kind of data, there's so much you can do with it, both in the pre-early stages of the project when you're looking towards what your effects might be, and then once you have some effects, extrapolating out from there.

So I think we've talked about exploring a problem and then a bit about how once we design and test our solution, we want to see what the impact was. The other piece which we alluded to at the very beginning was the idea that the results can't just live in our head. We need to share them outside of our heads. So how can data science help us tell the story of what we did and how our solution performed?

HUTTON: Sure. A great deal of effort and data science goes into effective data visualization, which is a big part of how data scientists tell their stories with data. Because data is such a big part of data science, we rely on graphs to show stakeholders the evidence that we've gathered and the patterns we have identified. There's an excellent book I want to recommend that's available for free online called “The Fundamentals of Data Visualization” by Claus Wilke. It covers most of what you need to know in terms of how to use different esthetic elements like lines, colour and white space to make graphs, you know, attractive and effectively communicate your findings to an audience.

For example, it's almost always better to use a bar graph than a pie chart to represent counts or proportions of categorical data. And although it is an obvious the words that you use to accompany the graph in terms of the title, caption, and axis labels are at least as important as the data and aesthetic elements themselves. So, data journalist John Burn-Murdoch at the Financial Times does a great job of explaining this in some of his recorded talks you can find on YouTube where he discusses how he went about reporting on the COVID-19 pandemic. And the visual features of a graph really matter a lot, but so do the words that accompany it.

Ideally, you want the visualization to be pretty much self-contained, such that a viewer could understand it, even if you weren't there to walk them through it. And more broadly speaking, presentations and reports should focus on the project sponsors original research questions and make sure that you explain how your results relate to that question or questions.

Anything else that you've learned and thought that they might be interested in should be presented as a bonus and only if you have time or space. While you may have learned some new and exciting things about the data and it's hard to contain that excitement, sometimes you don't want to trigger cognitive overload in your audience and lose their attention in the weeds before you get to the really important thing that actually addresses their question. If these things aren't directly relevant to the client's research questions, it's often better to present those additional insights in a technical report or scientific manuscript, than a presentation or an executive summary report where, you know, the reader can really take their time processing everything at their leisure and come back to it if they need to.

APPELT: I was just going to jump in and say, I think also that those points are really interesting in the context of behavioural science, because we often think about making our BI intervention EAST: Easy, Attractive, Social, Timely.

But the data visualization also needs to be easy, attractive, not necessarily social, but timely would be good. There's a lot of data visualizations where either there's trying to be too much told at once or, like you said, the attention is paid to the numbers and not the words accompanying it.

And I really liked your point about data visualization should speak for themselves because we all know how people actually read. We think, you know, they're poring over every word in our text, but really, people are skimming for the graphs. And if you don't have them be self explanatory, they're not going to get those takeaways that you're hoping they will take.

HUTTON: Yeah, for sure. And one of the videos that I referenced that John Burn-Murdoch, where he presented, he discussed how they did some eye tracking experiments where they look to see where people are looking on a graph and they always scan like the word elements. Right. And so that's definitely an important part of how people read and interpret a graph.

So then after you've presented the statistical findings to your project partner, it's critical to explain what the findings mean in terms of managerial or practical significance. This is something I know you cover in BI108 in the program later, but for your listeners, I think it's important to cover here also. In concrete terms, you want to tell them how did the intervention affect their key performance indicator and how might implementing it or scaling it impact their business.

For example, how many people do you have to send an experimental message to for each recipient who goes on to visit that website, sign up for that new program, or take advantage of a sale on an advertised product? How much does it cost to send out each message? If you have some information about the cost of the intervention and how much time or money it might save the organization compared to the status quo, provide your stakeholders with an estimated return on investment so they can make an informed decision on what to do next.

And the last thing I'll say about communication and storytelling for now is that you should explain the limitations of your study design, analyses and findings to minimize the risk that your audience misinterprets your findings and overgeneralizes them to context in which they have not been tested. It's not the end of the world if your new intervention didn't have a statistically or practically significant effect on the outcome of interest. It's better to be honest and transparent about your findings and recommend that they try something else than it is to push an unproven intervention to scale where the stakes are much higher.

APPELT: Yeah, those are really great points. And I think going back to your point about, we often get so excited about the nuts and bolts of our data because we spent so much time in it that we forget to put it in the context of some of the constraints and decisions and compromises that had to be made and how those then impact how we can interpret those results and extrapolate from them and think about the scaling implications. So that's a really good point. I feel like we have a good handle on the different places data can play into the RIDE model.

I also think a lot about best practices because data science, data, we can think of data like a tool and tools can be used well or poorly. And we've talked on the podcast about there's intentional data fraud. There's also just sloppy practices, unintentionally bad practices. P-hacking is often, for example, just poor data management. Sometimes it's intentional, sometimes it's unintentional.

But we haven't really talked about best practices. So what are some of the best practices? And maybe let's start with data management. What should we do and what should we not do?

HUTTON: For sure. That's also a great question. And I agree with you that a lot of the time it's not intentional. It's just that, you know, maybe people don't have much experience or they haven't received a lot of training in this area. And that's normal.

Well, first of all, you should use a version control system that tracks file changes. We often use Git, but that's better than having, you know, version one, version two, version three, version final, final, whatever.

APPELT: final final v2.

HUTTON: I did that a lot when I was in grad school, and I'm glad that I switched. You should also create and maintain data dictionaries and metadata that are accessible to all members of the team, not just yourself or other experts. So, if you happen to leave the team and somebody else takes over for you, they don't have to call you up and be like, “What does this column that has like a three-letter variable name represent?”. And you want to make sure that you're doing what you can to maintain privacy of participants and the security of the data so that it can't be accessed by individuals who shouldn't be able to access it, things like that.

The second main thing is that if you're cleaning and transforming raw data prior to analysis, you should probably try to do it with code again so that you have a record of exactly what was done, how it was done, and then you can just rerun the code to update the analytical version of your dataset if your data provider sends you a new batch of files. You know, that whole idea, going back to what we said earlier about help future you save time, which you know, in BI projects I find that we often end up with multiple versions of a dataset being provisioned where we are kind of like updating things and iterating maybe on an analysis with the best data available.

In the BC government we use for data management and data project management, what they call the "Five Safes" model. And this refers to the first one is Safe People. So only authorized individuals should have access to the data. If there are other members on the team who don't necessarily need access to the data, they maybe shouldn't have access to the data.

The next one is Safe Projects. So, these data projects should have a clear benefit to the organization and the public, for government projects. They should be conducted ethically and use sound scientific methods.

The next one is Safe Data, and by this, I mean data should be de-identified to the extent possible, while enabling you to evaluate your research questions. If you have like personal address and like full name information in your data file and you don't need those for your analysis, they shouldn't be there. You should be masking that information or dropping it.

The fourth one is Safe Settings, and by this we mean that data should only be stored and accessed in secure locations. You know, password protected behind locked doors, if necessary, maybe only accessed virtually through a VPN where there are additional security measures in place.

And the last one is Safe Outputs. Now, this relates a little bit more to analysis management than data management, but the idea here is that data analysis and extracts like counts should yield insights into statistical patterns among groups of interest in the data and not focus on the characteristics of specific individuals. So generally, you want to avoid being so specific or detailed in your analysis or description of a sample that any of your study participants could be re-identified which would violate their right to privacy. Rare exception to this might be the case study reports in medical journals that have been published with the consent of the patient.

APPELT: Yeah, there was so many good points in there. And I think the idea of thinking about safety through those different lenses is really an important one because I think we often focus on one or two of them and not all of them at once.

So, unsurprisingly, I like having models to help us remember the different components. And so you started to talk a little bit about best practices for data analysis. What are some other best practices besides making sure we're not accidentally re-identifying people?

HUTTON: Sure. Beyond maintaining privacy, for your participants, the next thing I'll say is that you shouldn't jump right into evaluating your treatment. Before you do any inferential analysis, you should start by looking at your data, and by that I mean like actually looking at the raw values and then exploring it with descriptive statistics and graphs like histograms. You're not going to find any problems there might be in your data if you don't look at it. And by this point you should have already clearly stated your hypotheses if you had any.

And you should evaluate your assumptions about the data, any statistical methods that you might be using. With respect to hypotheses, these should be specified and established well before you collect the data and ideally pre-registered or at least shared with a project partner prior to analysis, when you're developing your study plan with them.

And regarding assumptions, if you're doing a T-test, is your outcome variable normally distributed? Is the relationship you're modeling with that linear regression actually linear? Checking assumptions is just as important when you're exploring a data set the first time. Does this variable called date actually contain dates and do all the values use the same format? Chances are they don't, by the way.

Data entry errors and unstandardized formatting are common, particularly if humans enter the data into the Excel file that you received. Did the data provider use a special 999 code to represent missing values in a column where every other value ranges from 1 to 10? Failing to fix issues like that can seriously bias any analysis you might do afterwards.

APPELT: Yeah, and I think beyond the human element as often in BI we have-- we're pulling data from across sources and so did Department A and Department B use the same keys? And then I also just on that point about descriptive statistics, just wanted to add that when you're checking if the data makes sense, again, it could be things like, “Oh, they had 999's for empty data”, but also just looking for like if you collected things like age and income which are usually positively correlated, are those positively correlated here? If not, was there an entry error? Was it that people were not understanding the questions? Was it that people were not paying attention?

Because if you find those flags, like you said early on, then it helps you understand if the data is actually valid data to be analyzing or if it's data where something went wrong, whether it was on the recording or the provision of the data or what. And you want to find that out as soon as possible. You don't want to run your tests, find significant results, celebrate and tell everyone and then say, “Oh, actually the data is not valid, sorry”.

HUTTON: Yeah, you definitely don't want to end up in that situation. The last thing that I'll say with respect to this for now, is that there's no single best way to analyze all kinds of data. In data science this is called the “No Free Lunch” theorem. There are no shortcuts here. The right analysis approach for a particular data set should depend on the structure of the data set, your research question and your subject matter knowledge about the phenomenon of interest, and not whichever method yields the lowest p-values or how proficient you are at doing a particular test.

Ask for help if you need help. I've seen too many, way too many students in neuroscience labs analyzing everything with T-test because that's the only test that they know and that they're comfortable using. So don't be afraid to get out of your comfort zone and try a new approach, if the data are messier than you expect them to be and you suspect the assumptions of your favorite statistical tests are no longer valid. It's okay to ask for help and to look things up online. I still do that all the time.

APPELT: Absolutely. Yeah. And I think we always make the point too that the more you do data analysis, the more you learn. I don't think there's a "Oh, I know all of the analysis" is that's not a thing that happens there's always more analyses to learn and like you said you want to use the right method for the data and the questions you're asking because using a different method will answer a different question. And if you're not careful, you won't realize that you asked a potentially slightly different or very different question than you intended to be asking.

HUTTON: Yeah. Your project sponsor probably won't appreciate it if you come back to them with a report that answers the wrong questions.

APPELT: Yeah. And then sometimes you don't even realize that you answered a slightly different question. So, you also want to just be careful that there's not an accidental bait and switch around what you've asked of the data. And maybe drawing on that point a little more and moving in to best practices for data communication and visualization, what do you recommend there?

HUTTON: Sure. I mean, beyond what I've mentioned already, practice is the number one thing. So try your presentation out on other members of your team who haven't contributed to the analysis. Or, you know, if your project partner's okay with it, maybe try it on like a family member or friend to see if they can understand it. Can you explain your findings to your supervisor? If not, maybe your client won't get them either.

And then, the second thing is that you should try to really get to know your audience and tailor your communications to them. Either BI practitioners and specialists will want more methodological details than executives or members of the public.

For example, if I'm presenting the results of a study that evaluated the impact of a BI intervention on enrollment in a retirement savings program, and I analyzed the data with a logistic regression model, I might present the treatment effect as an odds ratio and 95% confidence interval to a specialist audience, whereas I might present the effect as the difference in the enrollment rate or total enrollment count for the treatment group versus the control group to an executive audience.

APPELT: I think that's a really good point and kind of brings us back to what we were talking about a few minutes ago. It's about segmentation. We often talk about segmentation when we're talking about a BI solution and how it might work differently for different groups. The same is true of communication.

When we're communicating out our results, the executive who spends, you know, 2 minutes on a topic isn't going to want the same level of detail as someone who has a data science background and really wants to know the exact models run. So, it's really important that you're able to communicate the key pieces to all the different audiences. And it's not that you're hiding anything, but you are making sure the right information gets to the right people and then the other layers of information are available if they want to go deeper.

HUTTON: For sure.

APPELT: So, any final data tips, any best practices or questions around best practices that I haven't asked you?

HUTTON: Yeah. The last thing that I kind of want to mention with respect to data has to do with relationships and specifically the relationship with the client's data team and data stewards. This is almost as important as a relationship with the project sponsor at the client organization themselves.

So, the advice is to connect with them, with that team, and build those relationships as soon as you can on BI projects that will involve using data. I know it's not always easy to tell when you're scoping if you're going to end up needing data, but it never hurts to kind of reach out and just connect with those individuals.

They're the ones who can help you figure out what data are available, which steps you might need to go through to access them. They're also usually the people you'll need to talk to if you have any questions about the data. Like “What did that column actually represent? How did you collect this?”, you know “Are there any known issues in the past that might impact my interpretation of what this is?”. They're the experts for the data that you want to have a good relationship with.

APPELT: Yeah, I think that's really important. And I don't know which situation I've been in more often, the one where other members of the client organization say, “Oh, we definitely collect that data”. And then the data team says, “No, we don't”. Or the project team says, “Oh, no, we don't have that data”. And then the data team says, “Oh, actually we have that and so much more”.

So either way, it might be a good answer or a bad answer, but you want to have that question answered before you get too far in. And I think also, like you said, the data team is usually a wealth of resources and I find that often they're not looped in. So if you loop them in, they're really excited to be able to tell someone about the data they've been collecting. So you often get a lot of buy in from that group, I think.

HUTTON: Yeah. If you want to nerd out about your, you know, research methods and the data they have, they're the best people to talk to for sure.

APPELT: Yeah, when we were talking about segmentation, they're the ones who want the details and want the nitty gritty. Awesome. Well, then I'll ask my traditional last question, which is just do you have a message for our new BI practitioners in training?

HUTTON: Sure. And I spent a little bit of time thinking about this. And I think the best advice I can give you is that science is hard and research rarely goes as planned. So don't be too discouraged if a study doesn't yield statistically significant results. Every study tells you something, and sometimes learning what doesn't work can be as useful to a client as learning what does work. Maybe you'll save them millions of dollars that would have been spent implementing something that doesn't work the way that they assumed it would.

APPELT: Yeah, I think that's always a really good message and it really is true when you, not only like you said, you potentially save them from running something that wouldn't have worked, but also often the behavioural insight's statistical significance is a small portion of all the learnings that come out of a project. So, whether or not you have a statistical significance, the project is never a failure. It's just telling you different pieces of information than you might have expected.

HUTTON: Definitely.

APPELT: Well, thank you very much. Data is always the underlying story and I love getting to do a deep dive on it. So, this has been really fun and I love the way that you combine your data science expertise and your passion for BI. And I'm excited to see what all you contribute to the field with your willingness to explore different methods. Where I think some of us get stuck in our ruts, you're really willing to try out different methods, and I think that's hugely valuable. So thank you for sharing your unique perspective with us.

HUTTON: Thank you for hosting me.

APPELT: And thanks to our listeners for joining another episode of Calling DIBS.