INSTITUTE OF
PUBLIC ADMINISTRATION
AUSTRALIA
ACT

02 6154 9800
PO Box 4349 Kingston ACT 2604
admin@act.ipaa.org.au
www.act.ipaa.org.au

ABN 24 656 727 375

TRANSCRIPT OF EVENT

GRADUATE DATA NETWORK
DATA VISUALISATION

Simon Temby (Guest)
Data and AI Specialist
Microsoft Australia

Gavin Styles (Host)
Department of Agriculture, Water and the Environment
Graduate Data Network Member

Philippa Clark (Host)
Department of Infrastructure, Transport, Regional Development and Communications
Graduate Data Network Member

18 February 2021

Enquiries should be directed to Caroline Walsh on 0413 139 427 or at caroline.walsh@act.ipaa.org.au

GAVIN STYLES: G'day everyone. It's Gavin from the Graduate Data Network here. We're here today working with the IPAA on this wonderful little presentation with myself, joined by Phillipa Clarke, Pip, from the Graduate Data Network, and also Simon Temby, a Data and AI Specialist from Microsoft. And we're here to talk about, I guess, the what and why of data visualisation. We'll give a quick overview of the Graduate Data Network, and, over to you, Pip.

PIP CLARK: Thanks everyone for joining us. The Graduate Data Network, if you haven't heard, is a vibrant network of APS graduates who are passionate about using data to make better public policy. Last year, the network comprised seven working groups delivering data related projects. One included the Crisis Response Handbook, which use data to capture lessons from the tragedy of the summer bush fires, and other produce the cultural and linguistic diversity report, investigating representation and experiences of CLD employees within the APS.

Regardless of your current level of data knowledge, there's a place for everyone to collaborate and learn in the Graduate Data Network. Join us next month, 22nd to 26th of March, as we turn our large Graduate Data Forum into a digital event. Each day, we will be releasing new and innovating sessions, including a keynote speech from Professor Maggie Walter from the University of Tasmania on indigenous data sovereignty. So please join us then as well.

GAVIN STYLES: And that event will be running conjunction again with IPAA. They've been wonderful and been helping us out this year and last year for our major Graduate Data Forum, which suddenly had to be shifted online at the last minute. They do a wonderful job here taking care of us. All of our previous events are up on the IPAA website, including our recent panel discussion on AI and its future in the APS.

So we'll move on forward today. Simon, welcome. Thank you so much for being here. Thank you for being involved with us today.

SIMON TEMBY: No, thank you, and thank you, IPAA, and the Graduate Data Network as a whole, for the invite. I'm very excited to be here and go through data visualisations. And I already apologise if I talk too much about data. It is a passion of mine. So I will often go off into tangents. If it gets too detailed, let me know and pull me up any time.

GAVIN STYLES: Will do. We'll do our best.

SIMON TEMBY: Do your best.

PIP CLARK: We get carried away as well, sometimes.

SIMON TEMBY: I wouldn't expect anything else from the Graduate Data Network. I might just go over the agenda of what I've come here to talk about today. At a high level, the data landscape and trying to focus on more about the government, but obviously that landscape is more than just government data. What is data visualisation, and why?

Then moving on to terminologies, some terminologies that you might not be familiar with, some you may be familiar with, to help just embed what words you'll come across, so you know them if they ever come up. Types of visualisations that you will either know or be new to you. Comparisons, so comparisons is a good thing to go through, bad or good visualisations and help with that really embedding when you see a visualisation, is it going to help you or not? And then some examples, and then there'll be some useful resources at the end, just to help you continue the journey, more than anything.

I will just go straight into it, because data landscape is a big topic, and I might drone on, so let's try and keep it as short as possible. But I think overview of data landscape is... There is an ever-growing amount of data in the world. If you don't know that, there's a pretty picture down the line that I'll show you, which kind of summarises that.

Also talking about the five days of data, types of data, and how do we make sense of it all? Now, a lot of this might overwhelm, especially the picture on the data landscape slide, looking at the big data landscape. And that was a 2016 picture. They haven't updated it since then. So you're looking at all the platforms, all the analytics applications, infrastructure, that support the data landscape. And this is a worldwide picture. So there's some things that aren't used in Australia, but most things are here now. And I don't mean to overwhelm, but it's context setting.

Just move on to the next one. So data landscape. So before we dive into the data visualisations, we do need to talk data. I most always include a few sides on this as its foundational understanding that underpins literally everything government and commercial organisations do. We can see here there is a prediction that the size of the global data sphere is going to be more than 40 zettabytes for 2021, and progressing through 2025 at 175 zettabytes. Now that's-

PIP CLARK: Sorry, quick question.

SIMON TEMBY: Yes.

PIP CLARK: What is a zettabyte?

SIMON TEMBY: Excellent question. I might start from the bottom and work our way up. So old school was probably, we'll say megabytes, we move up into gigabytes, we move up into terabytes, and then we move up into zettabytes. So they're basically...

GAVIN STYLES: 1,000 terabytes, 1,000 gigabytes kind of thing?

SIMON TEMBY: ... Ever scaling. Yes, yes, sure. So it's hard to grasp that in terms... If you haven't seen data on data, usually you don't see it as a single volume to hold on to. But a movie these days is about one gigabyte. So you're looking at times and times a movie on terms of data in the entire world.

And exponentially increasing, which is the interesting and more concerning part in terms of, how do we then start managing all of that, as we get more and more data about more devices are more and more people? Which is what it's going to.

Now, not all of this is government data. Actually, government data is the minority in terms of world data sizes, especially the 175 zettabytes. The most of it resides in industries like manufacturing and financial services. So mining companies, car makers, and then financials, so banks, and credit unions, etc.

Both of those industries had a financial need, though, to get a grasp of their data. Understanding how people work in the financial sector, and understanding how they take credits out, helps them advertise better, and helps them get more money that way. So their financial outcome was directly tied to understanding their data.

Manufacturing is much further on the forefront with real time data analytics than any other industry. And that's... You can see in terms of car manufacturers, a lot of robots on the floors, that are doing a lot of the manufacturing these days, and 'a lot of' is relative, but more than any other industry.

And taking a step back in terms of government data, there's not a real impulse for governments to make money off data. Especially our government. They're not there to make money, they're there to serve. And the impost on them is more of a service-based trajectory, as opposed to a financial gain. There are some total cost of ownership concepts, which we can touch on at some point. But that's key to why manufacturing and finance have kind of been on the forefront of understanding their data.

Now, it's important that data has progressively received more and more focus over the last decades, as companies start to look for better ways to engage, interact and service their customers, like manufacturing and finance. But the slower uptake in government has been notable, especially here, which is why I have a job. So thank you, Microsoft. But it's more about helping government understand what that progress is and how we get them from where they are today to improving efficiencies, outcomes or the citizen journey for Services Australia, for example.

Government departments are only on the beginning of that data journey. And it can be difficult. There are massive transformational things that will need to happen in a lot of departments for them to understand how they can scale easily and efficiently, given they are on the same exponential curve of data increase, however, it's not at the same volumes. Not 175 zettabytes by 2025. I would hope not. But it's more about them having to understand implementation, what technologies and services there are, how cloud works, and security of cloud, which is really important when we talk about our data, so citizen data, or when we only talk about intelligence data and protecting Australia and its common interests.

To date, though, however, data.gov.au have a fantastic catalog of opensource data, all purchasable data sets. They promote research and collaboration across industries as well as across government, and that's one of the DTA initiatives that has really taken off in the last couple of years, to see that increase in promotion of sharing of data and, let's say, consuming of what's out there between departments.

Now, that was a lot of time on exponential increase in data, but I'll go on to big data, which is a term that's thrown around a lot. Now big data describes volume and variety of data sets that can be used for analysis. Now, I often use the terms interchangeably depending on who I speak with, but data versus information are used as the same a lot. However, there are significant differences between the terms.

Data is usually seen as the raw facts or statistics. Zeros and ones. That's it. It's a point in time. Usually looking at data by itself doesn't provide much insight. Information on the other hand, is contextualised. So it's taking that data, providing context over the top of it, and giving clarity of that point in time data so we understand it better.

Word to the wise, and I'd take that with a grain of salt, the terms are often jumbled in conversations, and end of the day, as long as you understand what we're talking about either in this group or another group, regardless of the differentiation between the two, we'd be in a good place. And I always hold meetings where kind of clarify what we're talking about when we say data or information, to help that.

Now, into big data. Now, good question for GDN. So how many searches do we think are performed on Google every second?

PIP CLARK: Worldwide?

SIMON TEMBY: Worldwide.

PIP CLARK: Oh my goodness.

GAVIN STYLES: Probably... Every second, okay. What, seven billion people? I'm going to say like a couple million searches.

PIP CLARK: I'm going to say 100 million searches.

SIMON TEMBY: 100... That would have been exciting. You're much closer. It's about 2 million searches every second.

GAVIN STYLES: I'll take that.

PIP CLARK: You win this one, Gav.

GAVIN STYLES: I feel like that's a pretty good... I was correct order of magnitude. I'll take that.

SIMON TEMBY: I feel like you may have seen the slide beforehand and probably have a little bit of insight into the picture.

GAVIN STYLES: You're assuming I have a good memory, so I'll take that.

PIP CLARK: I think my assumption on how many people use Google is far greater than the reality.

SIMON TEMBY: Well, I mean, it's a key thing to show that volume. As we progress, what are these volumes and what do they look like? My favorite big data image, which isn't the one on the screen currently, but I can share it with you afterwards, shows a point in time and 60 seconds across a whole bunch of different online content or search engines.

So we look at Google. 2 million searches a second, sure. Twitter, 278,000 tweets per second. And this is data or information, depending how you talk about it. But that volume is key to understanding, globally, how big it is, and how big a part of the playing of our lives.

GAVIN STYLES: And I guess that's an interesting comparison as well. You just talked about sort of the sheer volume of data versus the volume of government data. As, I guess, governmental employees, how likely are we to have to come across and have to work with these sorts of big volumes of data?

SIMON TEMBY: I'd say more and more, because the value of this... We talk about citizen journey... Is understanding the citizens. And you don't get a lot of understanding from just the data that government would hold. They would be looking to what's happening on Twitter, or what's happening out in Google searches, for example.

As a off the cuff example, look at Google searches. If there's a lot of searches around paid parental leave forms, that would be perfect insight for Centrelink to understand, well, maybe our content is hard to find, so people are hitting up Google search more and more instead of being able to come to us and directly search on our website. So its value is there. It's sifting out what is valuable to you across 40 zettabytes or 175 zettabytes which becomes the problem.

Now, it can become quickly overwhelming, and I don't mean to overwhelm. But key to that point, how would you deal with 204 emails every minute? And sometimes even I feel like I'm getting those. Especially today. We were talking before. I have 204 emails, just today. So how do we deal with that? And that's key to understanding data visualisations, and how it underpins what we should surface and to who.

I'll move on to the five Vs of big data. So this is really how big data is explained in common terms, and takes in to consideration five key areas. Volume, velocity, verity, value and veracity. Now, in speaking about any terms, these terms are good to keep in mind as they may change and [inaudible 00:13:36] the visualisations you use. Is there a scale of data that you're trying to identify? Is there the analysis of streaming data, so velocity? Are you getting a lot of information every second? Is it a different forms of data, so you have to combine, for example, on this side, YouTube, Facebook and Twitter, to get an outcome?

I'll go through each one, just to give a quick overview, so we understand them a little bit better. Volume's pretty self explanatory, but simply put, this is just how much data. As data has increased over time, we continue to push the boundaries on how much can be stored. Definitely in my time, I've gone from talking in kilobytes to megabytes to gigabytes to terabytes. And now we're talking about zettabytes, in terms of global scale.

Organisations like Service Australia have multiple data lakes... And I'll come to data lakes later... Which store data for different business units in different forms. And whilst it's not updated every minute, the Services Australia data hauling is enormous for government organisation. That's one of the easier ones.

Velocity. I will touch on velocity, and we'll skip to verity in a second. But velocity is also self-explanatory as it relates to how much data in a minute. We're talking about Google searches. 2 million in a minute. That's your velocity that's coming in. This often related to a data stream. So we can tap into data streams, like how many Twitter comments are had every second. The New York Stock Exchange, which is a good example, one terabytes of trade information happening, and streaming that through.

It's a constant flow. So imagining the Department of Health, for example, wanted to know where everyone was traveling at any point in time to better understand COVID-19 community transmissions relevance... Besides this most likely being in conflict with the Privacy Act of knowing where everybody is every day of the week, this is a volume of data, and will continue to be generated as long as you have the stream switched on.

I will come back to verity a little bit more in the next slide because it's an important concept to understand. But in a nutshell, data can come in many forms. It's not just ones and zeros, like we say data is. It's videos, texts, pictures. Anything you can imagine, or was a piece of data at one point, is verity. Knowing how to interact with the different types of data is important, but technology is making it easier with the introduction of no code approaches... You don't have to be a full-blown developer anymore, and code and C# or Python or R... Through machine learning, which Plug is... As we continue to, let's say, evolve the machine learning algorithms, they're becoming no code as well, making anybody Joe Bloggs being able to use machine learning algorithms.

Veracity is one of my favorites, and only because it's the biggest challenge across government to date. One in three users don't trust their data, which is 30% of all people I've spoken to in the last, I'd say, 12 months, while I've been at Microsoft, basically say, "No, I believe it. I'll do it myself." Pretty much is their attitude. And it's a massive-

PIP CLARK: Is that one in three people at Microsoft, or generally-

SIMON TEMBY: Generally, in conversations.

PIP CLARK: Wow. And you guys work in the tech industry. You're part of the data collection?

SIMON TEMBY: Yep.

GAVIN STYLES: Does that tend to be a sort of garbage in, garbage out principle? People just don't trust that the data that's going in is good, or is it they just don't trust the actual algorithms and how the data's being used?

SIMON TEMBY: It's both. There is definitely a garbage in, garbage out problem, and whilst departments try to fix it, once you've lost trust, it's really hard to gain that trust back. I've been involved in projects where we're doing data analysis and trying to understand what along that process and where the data goes, and kind of data lineage, for example. And the one that comes to mind.. Did one, it was a 500 step process for this organization to go from data into an outcome, which was basically a 500 page report. It was a ridiculous size.

Along that process, we were identifying parts where different business units would do data analytics in Excel. So basic analytics and risk profiling matrices. They would hand the outcome over to the next team. The next team would be like, "No, we do it again." And they were like, "Why are you doing it again?" "Because they sometimes give us wrong answers."

And we found across that chain of 500 steps, it was happening multiple times. No feedback loop. It was just disparate and unmanaged. And it came down to that trust of the other teams, trust in the outcomes they have been provided. Even though they had a central unit that was talking to doing... Their whole job was the analytics. And I would trust them to do the analytics, because there are people, and they're much smarter than I am, doing analytics and data. However, the business units just didn't.

And this leads to one of the bigger problems in government, which is mostly around process efficiencies around data. And having to re transform the data every time takes time out of the entire process to get to an outcome. The duplication of effort, wasting time with multiple resources and extending the process time frame out. Imagine what could happen if we actually trusted the data in that example.

And finally, value. So value in the sense of this image indicated financial value, but in government, there's no focus on financial gain from the services I've spoken about earlier. So how do we define value for government data? And it's conversation I've had across many government departments. And there are several different schools of thought. And honestly, it generally comes down to impact. What impact does the business need or want to make? And thus, the value in the data can be aligned to to helping provide that impact.

In a past life, I was contracted to provide a department cost savings, and this was accomplished through data analytics, attributing to approximately millions... I'd say $90 million worth of savings for a single group and area in terms of their spending. And that was just done through a better understanding of what the data was, their inputs, and the impact that would help with that outcome. Importantly, though, value is derived from business. The old adage, someone's trash is another's treasure, comes into play. Not all data provides value for everybody.

Now, that was a lot. But like I said, data is my thing. And the five Vs of big data is a definite need. Now, we also want to talk types of data and get back to verity. Before, are there any questions? I know that was a lot.

GAVIN STYLES: I think you did a really good job of explaining it. I feel like I took a lot in.

PIP CLARK: Yep.

SIMON TEMBY: Okay.

GAVIN STYLES: I'm worried about how early we are in the talk and how much more information I'll be able to take in, but for now we're good.

SIMON TEMBY: You should have asked for more comedy.

PIP CLARK: How essential are the five Vs of big data? Is a concept across pretty much anyone who works with data?

SIMON TEMBY: This is what kind of defines big data. That there are a few other Vs. Sometimes there'll be less Vs, sometimes there'll be more Vs. Commonly, it's the five. But I've seen less talked about in organisations, and I've also seen an extra one. I forget what the extra one is. I was always kind of... It's a stretch. It kind of can fall into other things. So it's not essential, but it's how we describe big data. Big data, as a term, is used through those five Vs.

Let's move on to the types of data, because then I'll try and whiz through this, because it can get a little bit overwhelming again. But we talk about, there are many types of data. The range of different data types seem to be continually evolving as well. We wouldn't have thought of so much geospatial data way back in the day.

But it's commonly understood that they fall into either quantitative data, so that's numbers, or qualitative data, and we're just going to say it's not numbers. In the big bad world of data visualisations and data analytics, data types start to unravel, with the mentioning of timestamp data, or dark data, or spatio temporal data types. For the non physicists in the crowd, spatio temporal is data that relates to both space and time. Not that I'm talking to physicists.

GAVIN STYLES: We might be. We never know.

SIMON TEMBY: Well, don't ask me any questions about spatio temporal type, because I haven't had to deal with that just yet.

GAVIN STYLES: I'll make sure I keep it on the down low.

SIMON TEMBY: Well, ignoring the Einstein-like data types, the most open about my experience is structured, unstructured, and semi structured data types. This is less a type of data, and more a form, a structure of the data. Knowing these is important to understand how it can be used in visualisations and what contexts it brings, or what might be needed to aid in presenting relevant visualisations.

Now, we'll touch on structured and structured, because it is the most commonly referred to when we get into these meetings around data and what we're going to do with it. Unstructured or semi structured data have different meanings depending on their context. In the context of relational databases... Me getting technical... Unstructured data cannot be stored in predictably ordered columns and rows like you would an Excel spreadsheet. One type of unstructured data is typically stored in blob, so binary large objects, and catch all data types available in the most relational database management systems.

Unstructured data may also refer to irregular or randomly repeated column patterns that vary from row to row, or files of natural language that do not have detailed metadata, which is becoming more and more of a thing.

PIP CLARK: Could you give us an example of what might be an unstructured data set, like what you might...

SIMON TEMBY: Images is a good one. So images, without metadata, you don't know what they are just by looking at them. You can't really store them in a Excel spreadsheet. And the easiest way to understand that... Could you put this data into an Excel spreadsheet? Would it make sense? Yep. Then you're looking at structured data. If you couldn't, then you're looking at unstructured data. That's the easiest way to think about the two separate things.

Semi structured would possibly look at images but have metadata that says this image has a truck in it, or it's a red truck. And then you have semi structured data. So that's the easiest way to put it.

PIP CLARK: I think that's good. It gives a nice visual for everyone listening to this podcast to understand what types of data might this unstructured blob be.

SIMON TEMBY: I will probably touch on Excel later, but it is a good way. And everybody uses Excel, to probably the bane of my existence, but it is a good way to help people understand that-

GAVIN STYLES: Come on. I thought you were Microsoft here. Come on.

SIMON TEMBY: Yeah, I am Microsoft. Go Excel. Use Excel all the time. Thank you. So just quickly, many of these data types... And I'll go into this... Like emails, word processing text files, PDFs, PPTs, so PowerPoints like today's, image files, video files, conform to a standard that offers the possibility of metadata.

Touching on metadata, metadata can include information such as author, time and creation, like trucks or colors. What's in the image. And this can be stored in a relational database. Therefore, it may be more accurate to talk about this as semi structured. Semi structured documents or semi structured data. But no real consensus seems to have been reached. It really does come down to those, what are we talking about when we say semi structured in the conversation, and just getting that out in the open when we're having these meetings to understand... Get everybody on the same page.

And unstructured data can also simply be the knowledge that business users have about future business trends. Business forecasting naturally aligns with the BI systems, like we're talking about today, because business users think of their business in aggregated terms. Capturing the business knowledge that may only exists in the minds of business users provides some of the most important data points for a complete business intelligence solution.

Now, I will just flick through a few commonly termed data types. So as mentioned, data types range wisely. But this picture here helps give a good way to understand where it would fit, adding to the already mentioned, you might come across, data types, like machine data, open data, real time data. Operational data is a useful one, and then high dimensional data.

So data visualisations. Finally got the underpinnings of data. We can talk about what I actually am here to talk about today. I found this image on the web and I thought it was hilarious. I do have a daughter myself, so I relate to that way too closely. Anxiety level reaches, peaks when you go through 'how do you make'... Heart rate's pumping... And then, 'bread.' Dodged a bullet.

PIP CLARK: So just for our listeners, to describe the graph, the y axis is anxiety level and the x axis is daughter's sentences or daughters words. So the anxiety level increases with each word, and then rapidly decreases with the final word, 'bread'.

SIMON TEMBY: I had it this weekend. I kid you not. "Daddy, how do you make..." Bread. Okay, cool. We're good. We'll move on. So the body's data visualisation and why. Now, the next slide... I did steal this from CGS Solutions blog, which is 15 Statistics That Prove the Power of Data Visualisations. So all the credit goes to them. But it's the best summary of the impact of imagery I've come across.

Two key points here is that 93% of communication in today's day and age is non-verbal, meaning our brains are more accustomed to understanding imagery faster. So what is data visualisation? Data visualisation is the graphical representation of information and data. And that's a very scientific term, I'm sure. Most commonly, we'd all we all have interacted with graphs or charts through schooling, and Excel. Cough, cough. But data visualisation can also be extended to maps or just images when represented correctly.

Now, given the statistics we're talking about here... So I'll add to those. The human brain can process an image in just 13 milliseconds, which is an incredible amount of time, as opposed to having to listen through a whole sentence. 50% of the brain is active in visual processing. Human brains process visual 60,000 times faster than they do text, which I think is important for data visualisations. And we are exposed to five times more information today than we were in 1986, which I thought was a fantastic statistic, which is why I threw it up there.

Given all that, it becomes pretty obvious why to use data visualisations as opposed to Excel spreadsheets these days. As big data gets more and more visibility, visualisations is an increasingly key tool to make sense of trillions of rows of data generated every day. Data visualisation helps to tell stories by curating data into a form easy to understand, highlighting the trends and outliers. A good visualisation tells a story, removing the noise from data and highlighting the useful information.

However, it's not simply as easy as drawing up a graph to make it look better, or slapping on the info part of an infographic. Effective data visualisation is a delicate balancing act between form and function. The plainest graph could be too boring to catch any notice or to make it tell a powerful point. The most stunning visualisation could utterly fail at conveying the right message, or it could speak volumes.

The data and the visuals need to work together, which is key. And there's an art to combining great analysis with great storytelling. If we take this data table here, that's on slide nine, this data is the amount of digital camera and mobile phone sales spanning from the 1950s to the 2019. The original data set is six pages long, with nine columns. Not super large data set, but it's beefy enough. Six pages worth of Excel spreadsheet. Fantastic.

Now, how long would you take to identify which year the most amount of purchase occurred for cameras, or how many smartphones were sold in 2015? And if we look at the table itself, it would take forever, and honestly, I didn't bother looking at the table and searching. You're just overwhelmed.

And yes, you could use Excel and some fancy filters or functions to get the answer, but the point is, you'd have to re transform the data every time you thought of a new question. So it's not just those questions. Every time you have a new question, you're in Excel going, filter, sort, etc.

What's easier, do we think? And we look to graphical representations or data visualisation. So how long do you think it would take to get the same answers at the line graph? And this is an easy demonstration of a power of visualisation. And you look at here, how many cameras were sold in 2019, and 15 million. You don't have to do a sort by, and grouping, and aggregation. You see by looking at the graph. And even without pointing out the 15 million cameras down the bottom, it's easy to read when we sold 80 million. So I don't have a better example than that.

GAVIN STYLES: No, I think that's wonderful. And I mean, you even said about it... I sort of thought about it and was going, "In the time it would have taken me to find where the most smartphones were sold, I've actually probably answered six or seven questions that I've come up with since then from just looking at the graph."

SIMON TEMBY: Yep.

PIP CLARK: I also really like that you started with a point of, you're asking questions about the data. So there are certain answers you want to find, which probably also helps you make a visualisation that's way more useful, because rather than including all this extraneous extra detail, it's, "What are my actual questions, and what's the story I want to tell here?"

SIMON TEMBY: Definitely. And an often lost art is the five whys. And segueing a little bit, we often come across, Power BI for us reports, and people are like, "How do I make this better?" And we're often going, "Why? Why have you put that title there?"

I actually had one the other day, literally this week, where they showed me what they had done, and the first thing I noticed was they had just a number blog, which had 19.2457%. And I went, "Why do we need three decimal places for the percentage on this graph?" And they actually had a very good reason, but it's one of those things which clutters up the data. I don't need to round the three decimal places usually in government. They had a very specific use case, so I kind of let it slide, even though inside I was like, "It still looks terrible." But they had a reason to.

But the whys is definitely important. And that organisation was actually asking for help on how they guide their business users to better questioning what they need in their visualisations instead of just throwing up pie charts, for example, which are a bane of the visualisation community.

PIP CLARK: Yes, if you take anything away from this presentation, please steer clear of pie charts.

SIMON TEMBY: Yeah, I think I skipped on that later on. Pie charts, boo. We'll move on to business intelligence, and we'll just touch on this. Business intelligence is a level up from data visualisation, and kind of what we're talking about here. According to Forrester Research, business intelligence is a set of methodologies processes and architectures, and technologies, that transform raw data into meaningful and useful information, to use to enable more effective strategic tactical operational insights and decision making. Deep breath.

Commonly, we view business intelligence as the end-to-end process, which is what that's just describing. And it's end to end of gathering, storing and making sense of data for better business outcomes. So like I said earlier, better impact and alignment to what business needs.

The important buzzwords here are insights and decision making. Within government, there is a big way to increase processes, efficiencies and output, like I mentioned earlier, but further is for the processes to result in better outcomes or decisions.

Organisations have been living in a data swamp mindset for so long. Now data swamp is viewed as a data store with no structural system behind it. So just chucked everything into a central location, not thinking about how we might need to get it out later.

I lost myself there. Now, remembering the five Vs. Imagine that you want to know how many surgical masks you had to order for a hospital. Now, to get an accurate number, you'd need to know certain information. How many masks were used? What for? Surgeries, emergencies, normative care? And then how many occurrences of those happened within a month? Do we need a buffer for it? If so, what should the buffer stock number be?

Now, where would we find this information? Worst case scenario, it's going to be in 15 different systems, ten disparate data stores. And that's probably close to what some organisations are experiencing to date, trying to get answers. And they all have different access requirements, or they have teams that manage those data sets, or teams that manage the data sets. Suddenly, what seems like an easy question to ask becomes an arduous process. This is where business intelligence comes into play.

I'll push on. And I'll push on straight into, importantly, just data visualisation terms, real quickly. Now, I've put ten in the slide deck, which I'm hoping will go out to the people listening to this. And we've covered some today, but things that should be aware of...

Data warehousing, gap analysis, data mining metadata, which we've also talking about, but there's also Hadoop, which is an interesting one, analytics types and behavioral analytics, which is also important.

Now, I've just mentioned Hadoop. Now, Hadoop's just a programming framework that supports the processing of large data sets like big data. And it's in a distributing computing environment. Now, if you want to know more about that, there's plenty on Wikipedia, but I would definitely get to know those terms.

GAVIN STYLES: The one term I'd actually really like to flag is sort of something that I had in my mind already, reading your slides beforehand, and it's actually come up in your talk already. You've used the term data lake, you've used the term data swamp, and now there's data warehouse. I hear data warehouse and data lake almost interchangeably. Is there a real difference between the two? Is it just a different terminology?

SIMON TEMBY: So data lake is usually size. They usually have a central data lake, which might be a single data warehouse. There might be multiple data warehouses, or multiple databases to form the data lake. Again, it's one of those terms where it depends on the organisation, how they've progressed down that path.

Data swamp is definitely in the unstructured domain. It's a... Yeah, we'll just pump everything in there and we'll deal with it later. Data lakes and data warehouses are usually a little bit more regimented in terms of what they pump and why, for use later on.

PIP CLARK: So the data swamp is my hard drive of random films, whereas a data warehouse would be like, my budget or like my shopping list, whereas I've really structured that.

SIMON TEMBY: Your data swamp would be everything, and then your data warehouse would be the structured, what I'm going to use later. So I mean, the data swamp would be your entire hard drive. Everything that's ever sitting on your hard drive would kind of look like a data swamp. Because there's structured things on it, there's non structured things on it. Your data warehouse would be like your My Documents folder. All organized and I go, "My financial information is in that folder." For example.

GAVIN STYLES: Assuming a lot of my My Documents folder here.

SIMON TEMBY: It should be all in the cloud anyway, shouldn't it?

PIP CLARK: Should it? I don't know. One in three people don't trust the cloud.

SIMON TEMBY: Now, it'd be good to just talk about the tools now. I know I mentioned Power BI earlier, but skipping through to the types of tools. There are more tools in the market than organisations know what to do with. The three that I hear most commonly are Power BI, obviously as a Microsoft rep, but there's also Tableau and then Qlik.

All three are used heavily across all government organisations, and I come across them all the time. But definitely, given the Microsoft Stack and 365, Power BI is having a little bit more usage these days. It used to be sold as Excel on steroids, pretty much was its sales piece internally. But it's progressed far more than that, doing analytics, and can incorporate R now into Power BI if you need to.

Importantly, we'll move onto data visualisations. So there are multiple types and variants of data visualisation options. And how do you know what to use? Now this slide here, slide 18, is a great little reference guide for where you should go. What would you like to know about the first question, and coming back to the five why's. Not just doing it because I need to. What's the question I'm trying to answer and what would I like to show? Am I trying to show composition and something changing over time? And then into, is there are a few periods or many periods? And then we end up at either a stacked bar chart or a stacked area chart.

Now, this is one take on it. There's many takes on this, but I think it's a good reference for those that haven't had the exposure to types of visualisations and how to use them. It's a good place to start. Obviously, some questions can be answered outside of this little reference chart, so don't take it as gospel. But definitely use it if you don't have that confidence in what type of data visualisation to use.

And coming back down to... I'll go through a few of these... The important bar charts. And bar charts and line charts are stock standard go to. They appear everywhere. I'm sure you know that. And they both great at representing information quickly. And they're easy to read as long as you don't over complicate it.

There are great comparisons among bar, or comparison over time, for line graph. Now a key to these is usually keep it simple. A lot of lines and a line graph is not going to be useful to anybody. Makes it hard to read and ascertain important information.

Go back to the slide earlier, where it was just two lines for cameras and mobile phones. Perfect. Gives us that comparison over time real easily. Started adding in multiple lines, different vendors... You just get lost information. So keep the bar and line charts simple.

Pie charts. Now, I don't like to touch on pie charts but we will touch on pie charts because they're a contentious visualisation. And this is just generally because they're used heavily in circumstances that don't actually help the reader understand. If you have to use a pie chart, and I say this with a pain in my heart, just keep it to as little sections as possible. Make sure they always equal 100% or just don't use them.

This is a great example of what's easier to read and ascertain. Is it easier to understand what's next after tigers in terms of this chart? Or is it easier to do it in the bar graph? Further, it's not constantly... Sorry, I lost my position on this.... You're not constantly darting from the pie chart to the legend. That's the one of the biggest pains of pie charts, you having to go "Oh, cool, I can see what part is massive, which is the blue. Who is that?" Whereas for the bar graph, it's really... Tigers at the top, those represent. Just avoid them if you can.

Maps are an important one, and I will just show this one because I think it's a great slide to go through. This is where the wild things grow. It's a Tableau public map chart, and it's helping understand bioluminescence occurrences across the east coast of Australia. It's visually stunning, uses different colors to identify what types of species and where they can be found. And whilst it is very pretty, and I like it a lot, for business use, I'm not sure I personally would use that type of thing.

It was a little bit hard to ascertain what I would be asking. What am I asking this? Coming down to those questions, what am I trying to get out of this? And in our government context, it's generally not... For pretty understanding of what's happening, excellent. For answering the question of business, not so much.

There are maps that do help, which is an example of better business intelligence, which is the next slide on 22. And this aligns different colors to event amounts and displays by state for America. And that easily is much more recognisable and understandable at a quick glance, which overarchingly is where we want to drive to for business intelligence and visualisations, is quick glance, reading these and understanding outcomes.

So putting this all together, which is the next slide on 23, this is a Power BI dashboard representing sugarcane harvesting globally. Now, as you can see... And it's an interactive one. I think it is online and available for others... But a bar chart shows the different productivity per country, a table has the hectares and tons. And there's an interesting infographic that describe in relative terms the volume of area, so the harvest. You can see the soccer pitches down the right hand corner.

GAVIN STYLES: Soccer pitches and whales. Two really great touch points.

SIMON TEMBY: Look, you got to talk about data volume in a way that people understand, which this probably did it better than I did earlier. What I will say about this is, it is a great dashboard. However, I even feel that that's a little bit messy. There's too much. It's a bit overwhelming to understand.

I will show you though, and this is at the top level, at the next slide, if we narrow in at Australia, it filters it all down. And that's a little bit more understandable than, it's half a soccer pitch of tons generated, and it's about to whales that Australia produces.

PIP CLARK: Sorry. It is quite funny, describing sugarcane harvests in terms of whales.

SIMON TEMBY: Again, got to communicate the image as best as we can. What I will just wrap up with... There are a lot of different ways to do business visualisations, and I think we've gone from the data underpinnings, which can over complicate it and overwhelm, through to what is possibly good and what is bad. Cough. Don't use pie charts. Cough.

But the key thing that I think everybody should be taking away is those five whys. When developing these visualisation dashboards, we need to be very aware of why we are choosing that? Why should we have a bar chart? What question are we trying to answer? And then how do we best represent that?

And using the terminologies in conjunction with some of these quick reference guides that are on the slides, like 18 where we have the comparisons. what would you like to show? I'd always be asking that question. And then the five why's is always a why, but until you get to the fifth why, you usually don't get a correct answer. I assure you. So always ask more than one why. It needs to be like an annoying four year old, when you're trying to do data visualisations. "Why?" "Because I've got told." "But why did you get told? So that follows that chain.

Most of all, always try and compare what you've done to what somebody else has done as well, and reiterations. Don't be afraid to get it wrong the first time. The perfect example of a Power BI dashboard that gets the right answer is multiple people have looked at this and gone, "What were you doing, and what question am I looking at?" And when you get to a point where they no longer ask you why about what you've done, you know you've probably hit the jackpot in terms of a dashboard.

I wouldn't be Microsoft if I didn't plug for Microsoft resources. There are a lot of Power BI training that they do open, and it's freely online so if you ever have any questions and you want to understand how Power BI works, hunt down all the available stuff. There is also training that government is involved in that they give for free. So always look those up. But alternatively, I know Tableau and Qlik do similar things.

GAVIN STYLES: I come from an R and Python background. I'm more than happy to plug Power BI. It's been fantastic and so much easier to use.

SIMON TEMBY: Oh, thank gosh.

PIP CLARK: I don't come from a data background at all and I've even used Power BI.

SIMON TEMBY: Excellent. Look, if there are any questions feel free to reach out. I know you guys have my email addresses but I'm more than happy to answer and take questions about visualisations or do this again, and connect with you guys, or the IPAA on anything around data visualisations, moving forward.

GAVIN STYLES: Wonderful. Well thanks so much Simon. It's been really great. I mean I guess we did dig a lot more into the why and the what, but that's okay it was still in the title, I think we did really well there was a lot of really useful information here. I'd love to say a massive thank you to the IPAA for being involved with this, as well, as always just doing a wonderful job with our events, so please do look up the rest of the Graduate data networks events on the IPAA websites, reach out to us for graduate work if you are interested in getting more resources are getting involved in the amazing things we're doing as Pip said at the start, we've got our Data Week Graduate Data Forum online version, coming up soon. That's coming up March 22 to 26th so keep an eye on the IPAA website and hopefully your own APS network communications.

SIMON TEMBY: Hope to see you there.

GAVIN STYLES: And we'll talk to you all soon. Thank you so much again Simon for being involved in absolutely wonderful job was so great to have you on board, and thanks again to the IPAA and to you Pip it for being here with me.

PIP CLARK: Thanks Gav.

GAVIN STYLES: Thank you.