An InsightaaS/Toronto Cloud Business Coalition Best Practice Report: March 2018

Problem solving: The right data for the right question

A VISION TO VALUE (V2V) BEST PRACTICE REPORT

LEAD ANALYST: MICHAEL O'NEIL

CONTRIBUTING COMMUNITY MEMBERS: Mary Allen, InsightaaS; Andre Chin, CPP; Caterina Didio-Duggan, Information Builders; Pavan Jauhal, City of Toronto; Francis Jeanson, IAMOPEN; Victor Magdic, Concept Flow; Dean McKeown, Queen's School of Business; Jan Naranowicz, Rogers; Mitchell Ogilvie, Information Builders; Kelly Piggott, Information Builders; Michael Proulx, Pride Conflict Management; Dana Saltern, BusinessOne; Glenda Schmidt, ex-OECM; Vlad Skorokhod, eclairesys; Jeffrey Veffer, Jones Lang LaSalle

March 2018

Contents

Problem solving: The right data for the right question

Identifying the challenge(s)

The report is based on the premise that problem solving is a function of obtaining 'the right data' to answer 'the right question'. Inherent in the title statement is the assumption that organizations seeking information are able to identify appropriate questions and data sets, and to use data science to develop the insights needed to deliver answers needed by the business to make informed decisions; this in turn requires expertise in structuring questions that are both important to the business and answerable with analytics, and depth in finding and using data that is clean, relevant and complete.

In some cases, this process can become circular – the 'right' question is modified to become 'the question that is answerable with available data', which may result in the entire exercise becoming misaligned with the original business problem. In other cases, appropriate questions are answered with results that are skewed by incomplete or 'dirty' data. Organizations that have established mature analytics practices have learned how to balance the analytics aspirations of business leaders with the realities of data and its uses. In the "Problem solving: the right data for the right question" working group research meeting, the V2V community dug into how this process can and should be managed to deliver best results.

Challenge 1: the right question, or the right data?

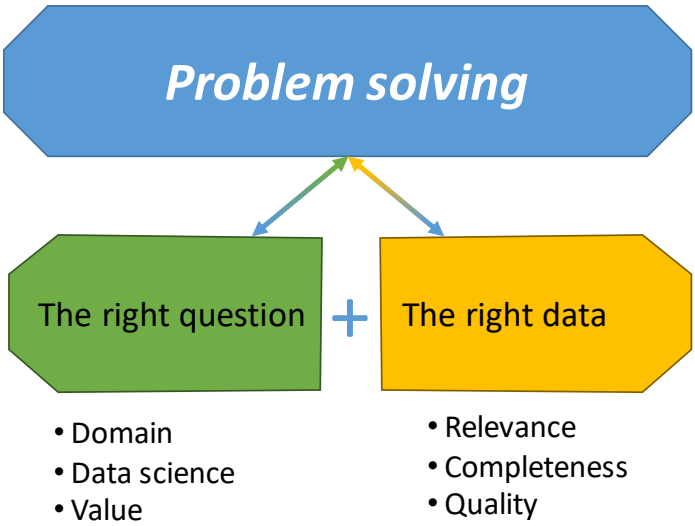
The research meeting began by asking "which is a greater challenge – getting questions defined such that they can be appropriately addressed by analytics, or assembling a data set that includes all of the data needed to address the question?"

Of 15 participants in the meeting, 12 offered opinions on the core question: three believed that data is the greater challenge, three believed the question is a greater challenge, and six believe that either or both can be the more difficult issue, depending on context: the availability of data from legacy systems, and the maturity of the organization with respect to understanding how to align what one participant flagged as *two* questions – the domain question, which reads on the business problem, and the data science question, which is a translation of the business question into syntax that can be addressed with analytics. This working group member added that "if you have both the right (data science) question and a complete data set, you can give the question to a graduate student" to solve.

The discussion in this part of the meeting focused on what drives the 'right' question or data. Through the conversation, three key elements were identified as important on each side of the balance. As Figure 1 shows, key 'right question' issues start with the *domain* – 'what do we need to know about our business in order to deliver better outcomes?' – and expand to include the *data science* question – 'how should this request be structured, to deliver the best and most reliable answer to the business?' The relationship between these two points is important, as the 'problem solving' objective can get lost in the translation. As one working group member observed that "there's a question that the business *wants* to ask, but [the question morphs] because it can't find the data to satisfy the initial request. Question A turns into question A_1 – now the right question isn't being asked any more, and you're working backwards – you stop asking the question that's needed."

A third issue, *value*, was raised in the discussion as well, with one member noting that "enterprises may be swimming in data, but they need to clean, store and manage data. How does the cost relate – what is the value of the data? Are you asking something of the data that will deliver real value" to the business? This pertains broadly to the overall value of an analytics practice within a business, but it should be incorporated into evaluation of specific inquiries as well. Is the value of having the 'right answer' greater than the cost of generating the needed insight?

Figure 1. Problem solving: establishing the balance between 'right question' and 'right data'



Source: V2V/InsightaaS, 2018

Discussion of the 'right data' side of the equation also focused on three issues. The first was *relevance*: does the data available support development of the insight that the business is seeking? The second, related issue concerned *completeness*: do we have all of the data needed to complete this analysis? This is a common and potentially-pernicious issue – many inquiries are constrained by available data. One member highlighted the fact that "you don't know what you don't know. Ultimately, you want a variety of data sets, and the ability to connect them." Another member stated that "I often start with a very

small data set" because that's what the business client has; there's no real question of accessing a 'complete' set of data to answer the question, while a third noted that in "established companies with legacy systems – the business itself [is] siloed, so it was difficult to get accurate (and timely) information."

The last point from the right-hand side of Figure 1, *quality*, is bound up in the overall assessment of whether the data is 'right' for evaluation of the business problem. This was a common concern within the working group – one participant stated that "you really do have to have good data" to provide meaningful support for problem solving within the business; another added that "dirty data will return a dirty answer."

At the end of this discussion, participants focused on two words, iteration and maturity. In practice, iteration is important to connecting clean, complete and relevant data to the analysis (though one contributor cautioned that "assembling data sets can lead into correlation mining – I'd rather be looking for the data to answer the right question" than basing questions on data). Maturity, it was generally believed, is the factor that determines whether the question or the data is most likely to be an organization's primary obstacle: "Organizations that are less experienced in terms of analyzing data generally have a harder time with the formation of their questions," while those with more expertise are "usually a lot more accustomed to formulating the questions well, and it's the gathering of the data that will be the bigger challenge for them."

Correcting for 'wrong' questions

Once the groundwork for the 'problem solving' topic was complete, the group delved into each side of the right question/right data issue, starting with examination of the first part of the equation: "Can you offer an example or two from your experience of questions that were 'wrong' – in the sense that they didn't fully address a real business problem – and talk about how you managed this issue, either by reframing the question or by redirecting the (internal or external) client to a different objective?"

While the initial discussion focused on the connection between domain and data science questions, exploration of best paths to the 'right question' quickly expanded to include another issue – tactics vs. strategy. In some cases, what makes a question 'wrong' is mostly a matter of focus. One group member offered this example from a call centre which was trying to optimize their approach to collecting on overdue accounts:

> "One of the questions was 'when are we scheduling our collection agents?' It turned out that *how* we approached the calls was more important than scheduling. What we found was that the scheduling didn't really matter – what was important was the right verbiage and technique."

Another example was raised by a practitioner with experience in the public sector, looking at an organization that was "trying to implement transformation, but taking a project-based approach – everything was tactical." Ultimately, the analytics group needed to persuade senior management of the value of the analytics practice – which in turn, required a more cohesive approach to the discipline of analytics (linking tools like master data management and objectives like predictive analytics) and a better framework for applying these tools to the broader needs of the executives, rather than the specific goals of individual projects.

A third anecdote involved an organization that prioritized specific questions based on the availability of data that would provide credible answers, rather than in terms of the issues that were most pressing for the business. In this case, there were some quick-hit responses, but to questions that weren't particularly meaningful – and the business itself ended up facing a shutdown because the analytics group hadn't determined that it needed to focus on the most critical questions rather than what was readily at hand.

It could be argued that this distinction between focus on tactics vs. business outcomes should be addressed in the original question formulation stage, but it isn't always possible for the analytics professional to redirect the business user's interest in this way. It is possible, though, to infuse analytics discipline with a commitment to iteration – and to use discussion of the answer to a specific question to identify related, and potentially higher value, issues and opportunities.

Aligning questions with evidence limits

Another means of correcting for unfocused or misdirect questions is to elevate business understanding of what can be asked of and answered by the data held by the organization. A participant remembered working with business leaders "to expose them to what the data warehouse contained" and finding that the domain experts "understanding what data was available helped them to ask better questions."

This can be a particularly thorny issue when dealing with survey data, an issue that was raised by several working group members. There are times when the question is not well enough understood by the respondents to represent real answers to a key question, and others where filtering the sample to eliminate non-relevant responses leaves the N values too small for credible analysis.

Other data sources may also fail to fully represent the information needed to make a good decision. For example, referral engines on websites are helpful where they can direct a potential customer to a product that is likely to be of interest – but these engines fail if they *don't* make relevant recommendations, and the recommendations themselves rely on having a rich body of data showing preferences for different types of consumers. If the evidence isn't adequate, the initiative may need to be scaled back until additional data can be added to the analysis.

Other examples of these types of mismatches pertain to the degree of certainty that the business user wants, vs. the type of response that the data will support. One contributor noted that "clients often ask for deterministic, predictive models – but the limits of evidence often will support only a probabilistic model. Business professionals don't like probabilistic – [for example] sales people want to know which deals will close, not the probability of a close!"

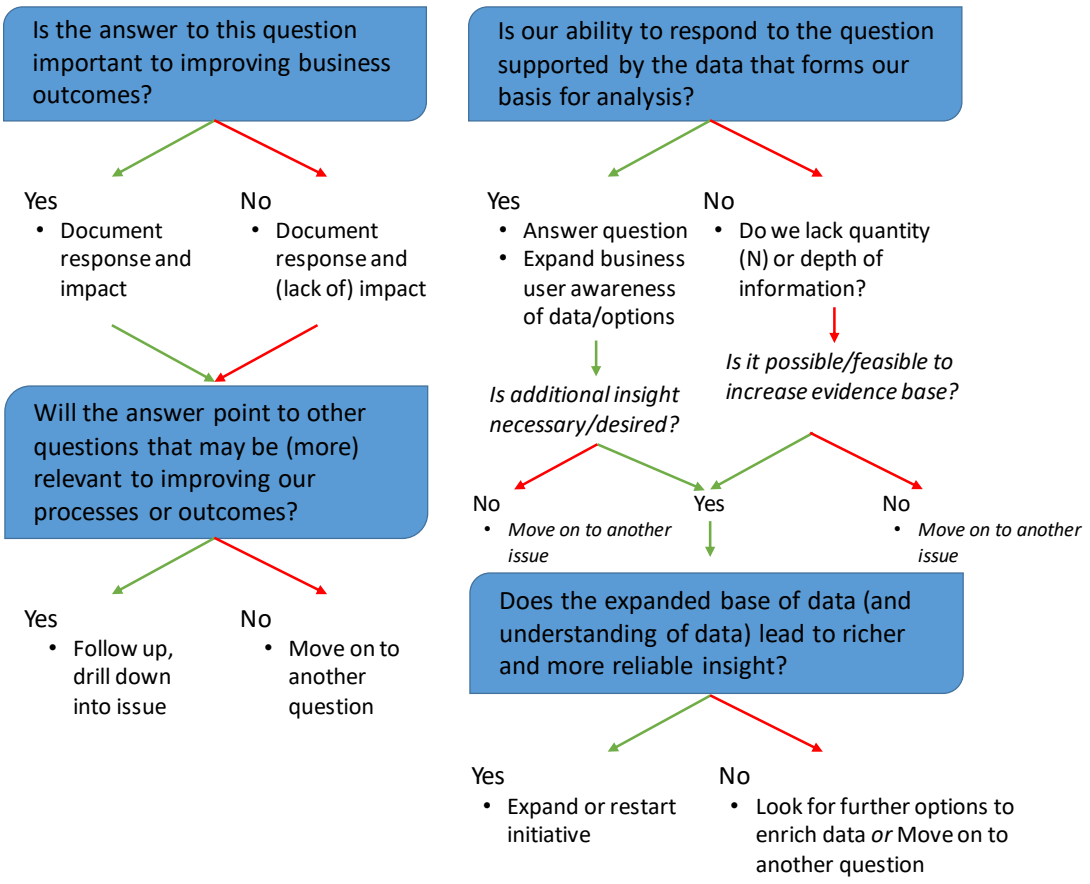
In these examples, the basic problem isn't one of focus; instead, the key issue is that the question is too ambitious for the supporting data. In cases of these types, the analytics professional needs to redirect the business person to focus on the benefits that *can* be gained through analytics. One group member identified this as a need to build data maturity, so that business people develop an "understanding of the real value of the data.' The member added that by working collaboratively with business-side clients, the analytics professional may be able to bring attention to "structural areas that are being overlooked that you can see with data…[the result is] not deterministic, but probabilistically, [this type of analysis can] indicate something that you need to fix."

The working group added two observations to this conclusion:

- Involve – meaningfully – stakeholders with different perspectives. In some cases, this will help the analytics professional to re-frame the question to better address business requirements with existing data. In others, it might indicate a flaw in the analytics approach; as one contributor observed, "if the CFO is pushing back [on a question or an answer] you may have to change your problem-solving approach.
    - This said, as a follow-up point, the group warned against trying to accommodate too may different perspectives. One member called a situation where the call for questions was "opened up to many members of the business community, asking 'what is it that you want to see?' The company ultimately asked too many questions – and unfortunately, that led to a misguided project – and these things do cost money; they need to be managed in such a way that you get best value."
- Make sure that there is a clear path from insight to action. In many cases, if the question needs to be corrected, the best approach to building consensus on a new approach is to get all stakeholders to agree on the business outcomes that they are pursuing, and to use that as a way to open discussion on the question(s) that will best support the activities needed to realize the goal.

Figure 2 presents a decision tree-style perspective on the correcting for the wrong question, looking both at issues of focus and of insufficient evidence.

Figure 2. Paths to 'correcting for the wrong question'



Source: V2V/InsightaaS, 2018

'The right data': Quality and completeness

The third part of the research discussion focused on the 'right data' side of the equation. It started with a statement – "it's possible to use the wrong data to answer a question – because it's incomplete, or because the data is mismatched from the question, or because it doesn't read on the underlying business objective; similarly, it's possible to use analytics correctly but to arrive at a wrong answer, if the evidence used isn't complete enough to provide accurate guidance" – which led to a question for the group: "how do you make sure that the evidence used in your analytics provides a complete basis for delivering the right answer to the right question?"

The working group, like the analytics community as a whole, found this a difficult question. One member offered that "I don't think knowing if you have the right data is black and white – this is a really gray area."

A group member summarized this initial stage of the discussion by observing that the issue can be framed as "what does it take to convince the audience you're working with?" He then expanded on the first question: "it comes down to – do you have a really sound question in the first place? Does answer map to intuition/experience? Is the science, math and statistics credible?"

Two paths to clarity

Domain understanding

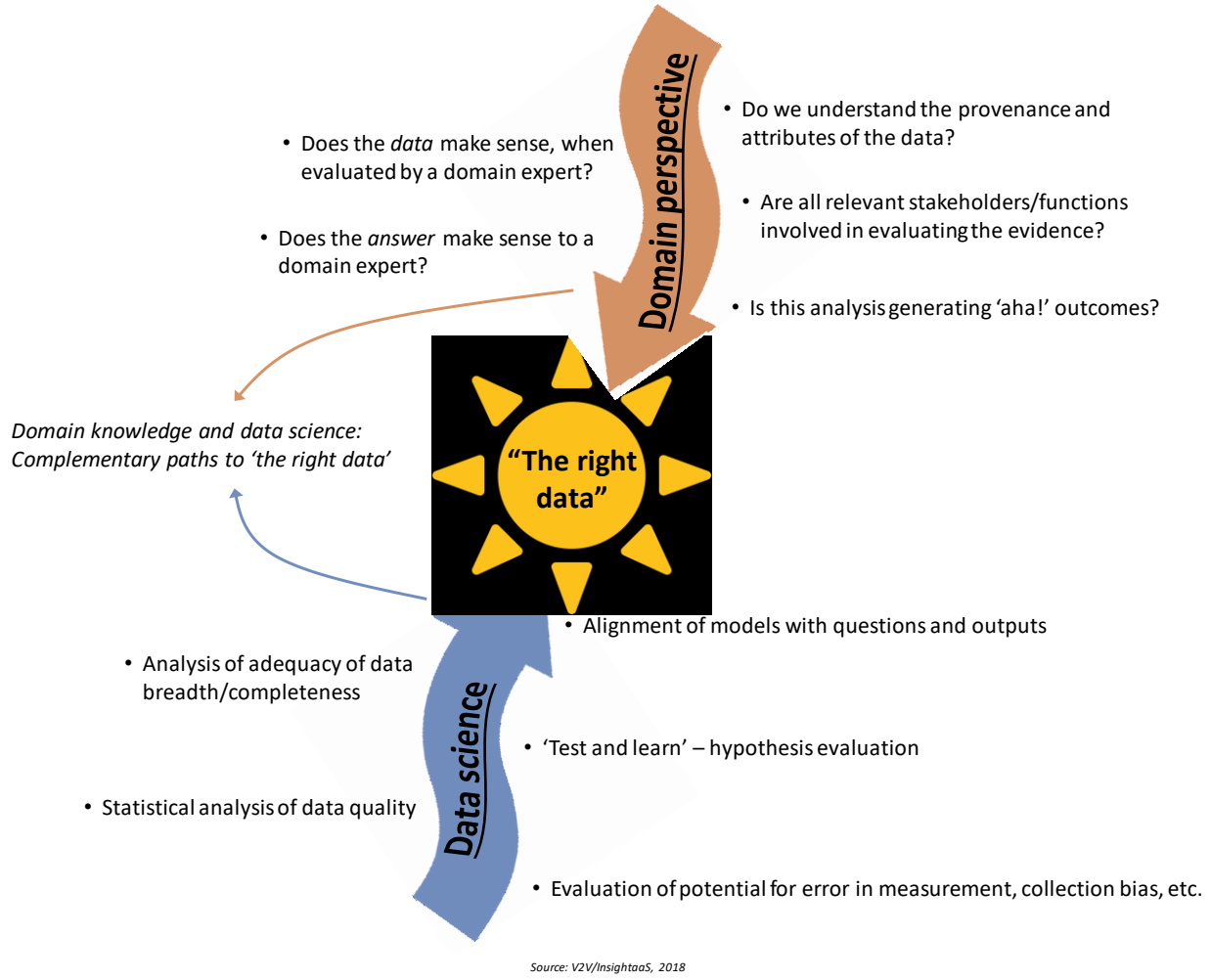
This observation led to a group discussion on the two issues of mapping to domain expert understanding and data science. One response offered by a contributor elicited laughter from the group: "How do I know when I have the right data? When it gives me the answer that I want!"

Beyond the mirth, though, this response was offered in one form or another by many members of the group. When the laughter died down, the working group member added that a domain expert "should have a feel for what the answer should be…if I see that the data and my feeling is totally off, I need to ask more questions, or go to the source and find where the mismatch comes from."

The issue of whether an answer 'makes sense' begs a related question – 'to whom?' Several group members stated that having stakeholders representing multiple points of view is important to establishing the validity of both the data and the answer(s) that it creates.

Before leaving this issue, the working group explored one additional topic: whether using data to support 'an answer' was actually a sufficient outcome to an analytics initiative. One observation along these lines held that "you can't use data to prove a point; all you can do is to use the data to improve the reliability of the answer…you really do need to have some instinct and some knowledge and some contextual understanding that works with the data." A successful data-driven investigation, the member explained, doesn't end with an answer to a question, but rather, an "'aha!' moment."

Figure 3. Complementary paths to 'the right data'



Science and quality

As is shown in Figure 3, the other path to establishing that the analytics initiative is predicated on 'the right data' is application of data science – building a practitioner-level understanding of data quality (including completeness) and of the 'fit' between models and objectives. One participant observed:

> "A perfect data set would describe the truth in a non-ambiguous manner – but these types of data sets just don't exist in business domains. There is always measurement error, human error, opinions instead of fact, things like that. This contributes both to data being incomplete because of missing signals, or incorrect because someone made a mistake, or the measurement system didn't work very well, or because our expert is biased towards some type of belief instead of giving us a pure processed view of the data...the point here is that we have to accept [data set flaws] and learn how to deal with it."

In many cases, the group opined, it is difficult to persuade business colleagues that data is not precise, accurate and complete. One member said that business colleagues "think that having MORE data adds greater validity to predictive capabilities, but that's not necessarily the case…just because you have more, it's not improving the predictive quality of what has already been assessed." This focus on

aggregating *more* rather than *meaningful* data can have a significant negative effect on decision quality; one working group member recalled a situation where a firm "collected data on every single thing but didn't know what they wanted to find out – too much data, not enough understanding of what the data *should* look like – leads to death by a thousand cuts – we don't know if we have the right data, we don't know if the data is clean, but we start to make decisions based on [the evidence]."

One method of using data science to break through to the right data is an iterative 'test and learn' approach. It can be described in this way: "if you end up with results that you weren't expecting, then you get back into the team (a diverse team of different people from different areas - people who know the data, people who know the business), and start re-evaluating your hypothesis, working through the characteristics of the business problem [and the data], figuring out where you could do better. That's where you can identify bad data, or incompleteness/missing fields or elements. etc. – keep moving until you get the result that you want or understand why you're not, and maybe you're going down the wrong path." Ultimately, the practitioner is responsible for dealing with data sets in the context of business demands – and this will likely entail understanding both what the answer should look like and what the data should look like.

Best practice advice

The key objective of V2V – and all of the communities managed by InsightaaS – is identification of best practices that can be used to advance the use of technology to drive better business outcomes. Asked to provide their input on best practices in applying the right data to the right question, group members focused on the following eight issues:

The right data is more of a journey than a snapshot

- Continuous improvement of data quality is very important (especially in smaller organizations, understanding how to continue to update both quantity and quality of data).
- Understand data provenance, and the transformations and operations done on data. Knowing the source and what's happened to data is key – some data scientists won't touch data without this info. Also, data sets created by joining data and bringing in data from other sources is another area to look at – it's not always an easy task to accomplish, and the resulting data needs to be reviewed before it is used.
- Success requires collaboration and implementation of a robust governance program – data dictionary, data lineage… – and having that information shared with the business community, I think, is crucial to supporting a strong analytics department.
- There is real merit in defining data quality parameters – establishing a numerical measurement scale for data quality, so that you avoid subjective interpretations. Once we've got that, we can apply metrics, and if we can measure quality regularly, we stand a better chance of having better data – putting in a data quality firewall, if you will, with the right checks and balances to ensure that data is being entered into the system with acceptable quality in the first place. On top of that, look for ways of incorporating data quality into business functions and processes, so it's not an afterthought but actually an integral part of processes…shift towards 'doing it right' the first time.

Teamwork counts!

- Collaboration between business and IT is critical. I (as a practitioner) can gain access to any type of data that a business user might want – but I'm not necessarily going to know whether that's the data that the business user requires.

Don't forget the 'science' part of data science

- This is where 'real statistics' comes into play. When you're looking at statistics around your data and your analysis, and you see (for example) a low P-stat value, then you know that there's probably going to be a good relationship between two variables…you have to go back to some of those basics.

Don't overstate the certainty associated with a statistical inference

- One of the things that I see/have seen is that people are excited by the potential that data offers, and they have a tendency to draw conclusions that maybe they shouldn't. It's a number – so (it's assumed that) numbers tell the truth, so if you use a number, then you're telling the truth. I see that in marketing…and I'm sure it permeates elsewhere. What I would really stress is, if you don't really understand what the numbers mean, consult with your experts, and make sure that the way that you're using the information is responsible.

Be cautious of systems that make repeated use of a single data set

- These kinds of questions get into play with AI and cognitive analytics. When you're building a system that's going to make decisions for you without supervision, you have to make sure that what you're implementing is a close to perfect as possible. But – no model is perfect. So you have to ask, how good is this model, and how long is it going to be good for us? How good is the data, and how long is it going to be good for us? When do we need to make a change?

Don't focus solely on the task at hand

- Data should be positioned as a strategic enabler, illustrating opportunities to create a new business or a new product offering. Many organizations are just focused on the issues/problems at hand from a tactical perspective. There is a need to take a more strategic view.
- Analytics initiatives deliver current outcomes from a business perspective, but also should be looking at answers that can't be found in what exists in the systems (except from connotations).
- Strategy vs. tactics – each requires different data. Strategic data is going to be different than tactical data – so is what you infer from either data set.

Prioritization matters

- Look strategically – best practice should be aligned with company strategy/direction. We can lose a lot of time sifting through data or answering questions that we don't need to answer – issues that are not necessarily relevant…
- Be cautious of requests for "vanity metrics" – data points that look good but which do not have much impact on decisions – vs. data that is actionable.

There is a need to iterate

- If your analysis seems clear, if you've identified a business problem and the analysis provides you with what you feel are the correct insights to move forward and you're confident, with the stats backing you up, then you're good to go. But if you hit a problem – for example, trying to

relate two variables that shouldn't be related, then you have to be realistic – stop, and try to figure out what's going on. You need to understand what you're trying to accomplish, and whether you've accomplished it.

- When you first get the business question and the supporting data set – your client isn't asking whether you want a good or bad data set – it's usually the best possible and the only data set they have (and they always believe that it's good). After the first iteration, the first attempts to come up with and test a hypothesis, you have a pretty good idea of what the recommendations should be, what kind of changes would be desirable in data policies. Whether it's possible to implement these changes may be a different question…

Final thoughts: applying metrics

At the end of the conversation, the group was asked if there were any metrics or milestones that could be added to the best practice guidance. The consensus regarding metrics was clear:

- There's no 'one size fits all' set of metrics that can be applied to all problem solving analytics initiatives. In some cases, metrics focus on outcomes (ROI, reach, click-throughs, etc.); in others, the metrics focus on the data (timeliness, accuracy/precision, completeness). The right metrics are determined by both the business question and the evidence needed to address the question.
- That said, metrics should be designed to "minimize the risk of answering the *wrong* question" – for example, they should be selected to reduce tendency towards confirmation bias.
- Regardless of what metrics are selected, they should be specified in advance of the initiative. The metrics themselves will likely change from initiative to initiative – but "up front is more important than a specific number."

About V2V

Vision to Value: The Economics of Data (V2V) is a Best Practices group based in Toronto. It was launched in October 2017 with a steering committee meeting that developed an agenda designed to support business professionals who want to capitalise on "the economics of data' in their business processes, customer interactions, management decision making and other facets of their operations.

The steering committee's guidance resulted in V2V adopting an agenda that spans twelve topics organized under three main headers:

Analytics in a business context

- Developing the analytics business case; identifying metrics
- Problem solving: the right data for the right question
- Monetizing data: identifying (and capturing) ROI on analytics
- Introducing analytics to the organization: where and how to begin

Establishing parameters for analytics operation

- Data classification to support standards, governance, privacy and security
- Data integration: expanding the evidence base with metadata and matching
- Data veracity: developing data trust
- Building data architectures: new models for new information sources

Analytics process change

- Value stream mapping: working back from end objective to project plan
- Changing culture and behaviour: socializing analytics success
- Responding to user needs: role-based self-service engagement
- Data visualization: democratizing data access and decision support

V2V members meet regularly by teleconference to produce documents in these series, and hosts Meetups to discuss research findings with the broader V2V community.

Sponsoring member

V2V wishes to thank Information Builders for sponsoring the dialogue on the business value of analytics and data.

Contributors to this document

We would like to acknowledge the following contributing community members, whose input shaped this document.

Caterina Didio-Duggan
Area Marketing Director, Information Builders

Caterina is an experienced B2B marketing leader who creatively uses multiple channels of marketing - email, direct mail, social media, events, promotions - to deliver a compelling message to drive executable leads and results.

Dana Saltern
Director, BusinessOne Corporation

Dana is an executive advisor with 20+ years of transformation experience in both the public and private sectors, driving strategic initiatives by planning and implementing solutions to increase engagement, improve performance and foster innovation.

Dean McKeown
Associate Director, Master of Management Analytics, Queen's University

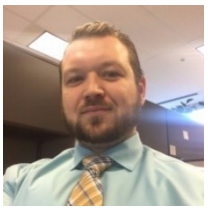
Dean has held many roles in the post-secondary education sector, and has served as a trustee on several non-for profit boards. Dean's research interests include big data analytics, governance, privacy and public policy.

Francis Jeanson
Data Science and AI entrepreneur

Francis is the founder of IAMOPEN, a start-up focused on helping organizations to share and discover valuable data. Francis has a PhD in Cognitive Science; before IAMOPEN, he led the informatics and analytics initiative at the Ontario Brain Institute.

Glenda Schmidt
EMBA candidate; former Senior Business Relationship Manager, OECM

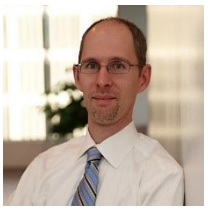
Glenda works at understanding complex issues, capitalizing on connections, and collaborating with teams. She has professional experience in strategic and tactical planning, analytics, process development and other key business requirements.



Jan Naranowicz

Reporting Analyst and WFM Leader, Rogers Communications

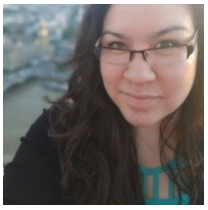
Within Rogers, Jan has a leadership role for support of the Customer Activations, Migrations, and Onboarding departments. He is responsible for daily SLA targets for all departments, and for each department's capacity plan and monthly budget.



Jeffrey Veffer

Senior Consultant, IoT and Analytics, Jones Lang LaSalle

Jeffrey manages IoT product delivery as well as business/partnership development at JLL. Prior to JLL, Jeffrey was Director, Internet of Things and Business Intelligence for Brookfield Global Integrated Solutions.



Kelly Piggott

Marketing Analyst, Information Builders

Kelly is a marketing ninja with an eye for detail and an ear for ideas. She works with the Information Builders team to build their personal and professional brands, human-to-human. She holds a BSc in Computer Science and Logic from U of T.

Mary Allen

Chief Content Officer, InsightaaS

Co-founder of InsightaaS, Mary has devoted two decades to understanding and communicating key trends shaping Canadian and global IT markets. She has authored hundreds of reports, articles and analyses on advanced technologies.

Michael Proulx

Founder, Pride Conflict Risk Management

Michael is an enthusiastic and collaborative professional with proven experience in academic and corporate environments, well-versed in capital/investment strategies, risk management assessment and analysis, client relationship management.

Mitchell Ogilvie
Senior Solutions Architect, Information Builders

Mitchell specializes in the data integration, data quality and master data management space. He has deep experience in helping customers assess, scope and deliver enterprise information management solutions across different verticals.

Pavan Jauhal
City of Toronto

Pavan is a Business Analyst for the City of Toronto. She has more than a decade of experience in IT, focused primarily on banking projects dealing with compliance and reporting. Making informed decisions from data analytics is my passion!

Victor Magdic
Managing Director, Concept Flow

Victor holds a law degree from Osgoode Hall Law School and is trained in numerous IT methodologies and disciplines. He has a strong focus on the financial services industry, cloud and business blockchain.

Vlad Skorokhod
Principal and Director of Data Science, eclairesys

Vlad has a Ph.D. engineering, M.Sc. in experimental physics, and is a co-inventor of 51 US patents. With a background in experimental science and system engineering, he now applies scientific principles and data science to complex business problems.

Lead analyst

Michael O'Neil is Canada's leading IT industry analyst. Through the course of his career, he has he has helped executives at leading buy-side and sell-side organizations to capitalize on new technologies and accompanying business opportunities. O'Neil has authored hundreds of reports and whitepapers, and two acclaimed management books: *Building Cloud Value: A Guide to Best Practice, 2016* (with InsightaaS partner Mary Allen) and *The Death of Core Competency: A management guide to cloud computing and the zero-friction management future* (InsightaaS Press, 2014).

About InsightaaS

Dedicated to exploring "the 'why' in enterprise technology," InsightaaS was founded by Mary Allen and Michael O'Neil in 2013. The company operates Canada's deepest IT content website and provides strategic consulting and channel management guidance to leading firms in Canada, the US and abroad.

In 2015, InsightaaS launched the Toronto Cloud Business Coalition, a community dedicated to the co-creation of Best Practice guidance designed to accelerate adoption and use of cloud in Canada. The tremendous success of the group has spawned three additional communities – IoT Coalition Canada, Canadian Analytics Business Community and V2V: The Economics of Data, plus the CIA-Plus meetup community. These groups continue to help Canadian businesses to capture value from advanced technology.