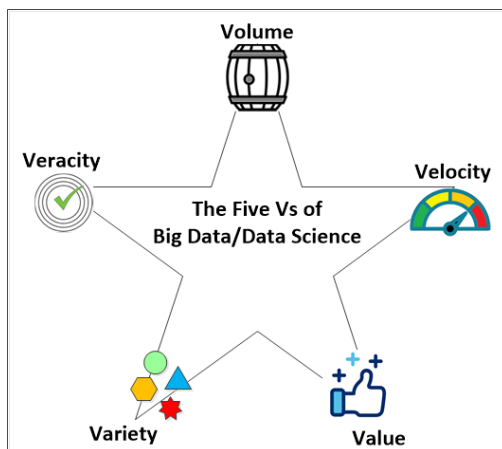# "Data-Driven" is not a Magic Elixir

By Joe Chiarella
27-May-2020

Policy makers bandy about terms like data-driven and fact-based decision-making as if they were some magic elixir that guarantees sound decisions. They are not. This article draws some distinctions around data, information, insight, cognition and decision-making. In so doing, help the general public understand that data-driven does not equal good decisions, and help the leadership/policy-makers/business-owners understand that they benefit from a more balanced process to making "data-driven decisions."

Before digging in, you may wonder about this author. I started my adult life (late 70's) in science writing code to crunch data first about geologic (water) research then later in plasma fusion research. The last 38 years of my career, I've been in the information systems and technology field. I've worked with data and information all my adult life. I'm a math and science junkie and hold two patents in data encryption. I know data and information theory. And as part of a software startup some 20 years ago, I was fortunate to learn about the science of cognition, a topic I have continued to study. I am not an academic expert in any of these sciences. I am a professional with a lot of experience, plus a few hundred relevant texts consumed.

What is Data?

Data represents a fact, but a fact is not necessarily (quantifiable) data. The number 42 represents some fact, but the number alone does not say what fact. Is it data? Strictly speaking no. Conversely, a fact is not necessarily data. For example, you can say that it is a fact that you love your spouse, but that fact is not necessarily a quantifiable piece of data. It is a fact that it could be a binary piece of data: yes or no, true or false. You love your spouse is a binary yes, so in that sense it is quantifiable data but without much variety. How much you love your spouse contains more variety but is quite hard to quantify. Do you measure it today and it is different tomorrow? And on what scale do you measure it? 0 to 100? It is going to be subjective from person to person (where the 0-100 range may vary) and, thus, the quantifiability of it is suspect (see Veracity below). We could explore this for days, but I trust this much of an example is enough to make the point: data and facts are different things.

In data science we talk about the "**Five Vs**" of data: **Volume, Velocity, Value, Variety** and **Veracity**. *[Some also include Visualization and Virality for 7 Vs.]* Do you have enough **Volume** of data? **Velocity**: how much of it do you have in each unit (of time or other unit)? What is the **Value** of the data you have? **Value** is measured lots of ways – and sometimes – often – in new situations like the one we are in – we don't know the **Value** of one piece of data versus another. That gets teased out over time using a range of well-known and understood statistical techniques for determining value. Do you have enough **Variety** in the data? Are you collecting enough different facts (for example:

temperature, viscosity, acidity, color, etc. about the fluid in a beaker of liquid)? **Variety** also means for any given fact (say temperature) are you getting enough **Variety** of values for that fact? If the temperature is always 36.4 degrees Celsius, the fact that you have a billion samples of that fact that are all the same means you have lots of **Volume** of nothing important. (unless the fact that the temperature never changes is important to know) Finally, there is **Veracity**. This means, is the data you are collecting: true? Is it verifiable? Is it CLEAR of ambiguity? Do you know the provenance of that data and is it good? There are lots of ways to measure each of the Five V's of the data itself.

As it regards Covid-19 data, Volume is lacking. We are not conducting enough tests, nor enough contact tracing. This point is well documented. However, what about the rest of the Vs? On the Velocity of data, this one is particularly hard. The needed Velocity, to some degree, is variable and dictated by the virus itself. When there is an outbreak, the needed Velocity is higher. The Value and Variety of Covid19 data is questionable and somewhat intertwined. As one example, there have been sparse reports that Vitamin D deficiency makes the human body more susceptible to a more aggressive progression of the virus. We don't know if Vit D is a valuable data element, yet. But it does represent an element of Variety of data. Some Variety, over time, will fall away as not Valuable. Vit D deficiency may be one of those data points. We don't know yet. Finally, there is Veracity. Is the data we are getting verifiable and credible?

I have been tracking this data carefully for two months now. I've seen cumulative infections and cumulative deaths both decline from one day to the next.

I can see the cumulative infection rate fall IF a test result is later confirmed as a false positive (the test said the person was infected but was later proven to be false). But it is just as likely that a test results in a false negative (they have Covid19 but the test says they don't). So, hard to know why the confirmed case counts go down. Then there is the fact that the Covid19 tests themselves have varying levels of reliability is a concern for data gathering. Unless each test result (data) is accompanied by the name of test that was used and then double-checked by a second test to confirm the first – I'm not sure how we can know, ultimately, the veracity of any test result.

I cannot fathom how a cumulative death count can go down. People don't rise from the dead, and certainly not by the dozens. This makes me question both the data reporting activity and the data itself. Ergo, overall, as it regards the Five V's of big data – all five are questionable.

Data is lacking. Then there is what we do with it.

It is good that we are collecting data. And it is appropriate that our decision-making be "guided" by data. But it is clear that the Five Vs of Covid19 data are lacking. What is more concerning to me is what appears to be a lack of understanding of what to do with that (poor) data.

What is Information?

The word "information" derives from the root "inform" – which emerges from the Latin "informare" which ultimately means "formation of the mind". The mind seeks meaning more than it seeks data. That is to say – it seeks understanding. Data can lead to understanding but is not understanding in itself.

Data by itself is useless; until it is turned into information. More importantly: until it is turned into insight and/or meaning. A gallon of gasoline (data by metaphor) is useless without an internal combustion engine (or other process) to turn that potential energy into horsepower (meaning) to move

you down the road. Data needs an engine (some process) to turn it into information and more, like insight. This is key: not all data engines (processes) are created equal.

Take the current press practice of reporting that there are X number of cases of Covid19 and Y (un)associated deaths, today. This is data, but it is not information. For example, taken in isolation, what is the value in knowing the <u>fact</u> that there are 12,345 cases today? Alone it doesn't mean much. But, if you have that piece of data for each of the last 29 days, you can combine that data with the dimension of time – and plot the rate of increase or decrease over that time dimension. Now you have information. But it took combining one dimension: cases with another dimension: time, to really turn that data into information. Information, in a sense, is about CONCLUSIONS derived from historical data. It tells you "what was" or at best "what is."

That is a simple example. Let's build on it a bit. Let us pretend that we've been tracking confirmed infections every day for the last 47 days and we've also been tracking hospitalizations, recoveries and fatalities. More specifically, we've been plotting these four time-series data on a standard line chart with dates on the x-axis and quantity on the y-axis. By tracking spikes in the time dimension for each line – can we find a regular periodicity of a spike on one line and a spike on another? If the hospitalization spike is always 17 days (+/- a day or so) after the infection spike – that association – that correlation is information derived from the raw data. It is in those associations and correlations where information is found. The data itself, again, is the fuel only. We find information (meaning) in the process of how we arrange, assemble, and associate the data.

But if your data is graded a "D," then you must consider those associations speculative, not conclusive. Beware of making decisions on speculative (or ambiguous) information.

Decision-making is best when it emerges from conclusive information (when it is available – and it isn't always). But people process information differently due to variations in cognition and situation.

Cognition's role in decision-making:

The world of cognition might be loosely divided into three levels. There are cognitive "perceptors" – which is how you and I gather information from the inner and outer worlds we live in. Then there are cognitive "processors" – which is how you and I process what we perceive. Finally, there are cognitive "controls" which guide and direct our "processors" by directing what (and how) we focus. That is, to what (and how) we put our attention.

Two of these cognitive controls are relevant to this article. One is called "**Category Width** (Narrow vs Wide)" and the other is **Field** (Dependent vs Independent).

The Narrow Category Width people tend to be more discriminating in how they categorize information. These people put fewer things into a category – preferring to exclude. These thinkers tend to have more categories with fewer things in them. The Wide are the opposite; they tend to be more inclusive or less discriminating in creating categories of information. They have fewer categories with more things in them.

The Field Dependent vs Field Independent continuum is also very crucial to decision-making. The Field Independent (FI) thinker tends to be more internally focused where the Field Dependent (FD) is more externally focused. Here, internal vs external has more to do with the center of attention rather than

"introvert vs extrovert". Said another way, the FI thinker prefers to limit their "field of focus" to the leaf on the tree (internal) while the Field Dependent (FD) thinker prefers to take in the whole forest (external) into their "field of focus".

These two cognitive controls have some "interaction" but are not correlated. You can be FI and Narrow or FI and Wide or FD and Narrow or FD and Wide.

These two cognitive controls have immense power in directing the way an individual processes data and information.

The FD that is a Wide categorizer is all about understanding how everything is connected and interacts – and by "everything" I mean – a WIDE view of their field of focus. People of this nature tend to gather lots of data and loosely categorize them into fewer groups of more stuff and then seek to find how all that stuff interrelates. These people require more time to reach decisions and their decisions take a lot into consideration.

Two important cognitive controls		Category Width	
		Narrow	**Wide**
Field	**Independent**	Focus is on this leaf of a tree or that paw print in the mud. Seeks to understand the meaning of each particular focus, aware that it is part of a larger whole, but doesn't want to be distracted by that bigger picture. Within whatever "Field" of focus: More groups of fewer items. Might group grasses separately from moss or shrub separate from tree.	Focus is on this leaf of a tree or that paw print in the mud. Seeks to understand the meaning of each particular focus, aware that it is part of a larger whole, but doesn't want to be distracted by that bigger picture. Within whatever "Field" of focus: More items in fewer groups. Might group all green plants in one group, separate from another group of all wildlife.
	Dependent	Focuses on the Forest as a system. Cares about the interrelationships between all the foliage and inhabitants of the forest. Is aware that there are greater details in each element of the forest, but doesn't want to be distracted by those details. Within whatever "Field" of focus: More groups of fewer items. Might group grasses separately from moss or shrub separate from tree.	Focuses on the Forest as a system. Cares about the interrelationships between all the foliage and inhabitants of the forest. Is aware that there are greater details in each element of the forest, but doesn't want to be distracted by those details. Within whatever "Field" of focus: More items in fewer groups. Might group all green plants in one group, separate from another group of all wildlife.

Another important Cognitive Control is "Cognitive Tempo" which also has two ends of a continuum: Reflective and Impulsive. These are relatively self-explanatory. They influence how quickly a person will reach a conclusion – independent of how they categorize or control their field of focus. The two extremes if all three controls are considered are: Narrow, Independent and Impulsive (FAST and TACTICAL) vs Wide, Dependent and Reflective (SLOW and STRATEGIC). Each end, and each position between them, has value – just different values.

At the other extreme is the FI that is a Narrow categorizer. These thinkers are more likely to drill down on the narrowly defined collection of data or information they have and tend to approach decision-making with a mindset of "don't distract me with that stuff – it isn't relevant to me at this moment."

In between are the FD/Narrow or FI/Wide people.

I like to say if you need to go fast and have stomach for risk – you want the decision-maker to be FI/Narrow. They'll move quickly. However, they may discover that down the road they've created problems for themselves because they failed to take enough into account.

In contrast, your FD/Wide folks are capable of deep understanding and insight, but not quickly. And sometimes, they can get tied up in trying to understand the deeper meaning found in the connective tissue of the data/information they are processing. This can cost time and limit risk-taking.

It is important to hasten to add that cognition in general, and these specific cognitive controls in particular, are complex. There is danger in the oversimplification of them. A risk I take here for brevity.
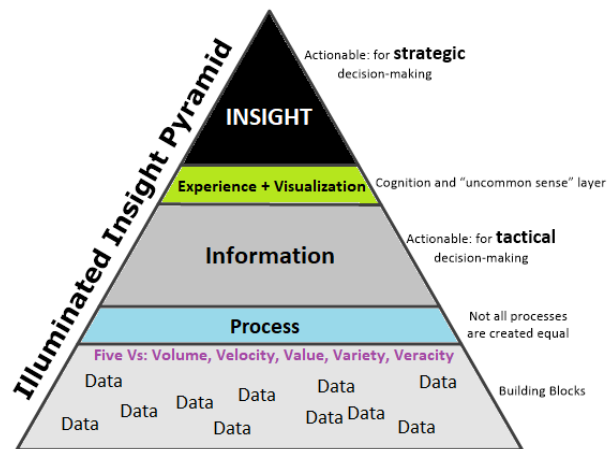
Putting it all together:

Beyond information there is insight. Dictionaries define insight with varying language around "deep understanding." I prefer the definition that adds "of the relationship between cause and effect in a given context." Consequently, this cause and effect understanding can lead to "future sight" – the ability to predict where the context will go. Thus, information is about "what was" whereas Insight is about "what might be." Insight, some say, is a gift and cannot be learned. I prefer to think that everyone is capable of insight, but it may come easier and be more profound for some.

I call this combination of solid data, with a strong process for turning it into information, harnessed by someone(s) capable of seeing the cause and effect relationships ("uncommon sense") between the information and events: "illuminated insight."

Those currently responsible for decision-making around Covid19 are naturally a mix of FD/FI and Wide/Narrow (and Reflective/Impulsive) – operating in a field of data that is fluid and scores poorly on the Five V's test.

It is thus natural that we are going to have a wide range of decisions that may even be in contradiction. It doesn't necessarily mean that opposing views are a result of good vs. bad intentions or smart vs. stupid. Differing decisions may just be a consequence of differing ways that leaders gather, discriminate and process data and information for meaning.

With that said, my observation is that there are two problems here that are hampering our response to the pandemic. First, the data are poor and, second, the decision-making is leaning toward the FI/Narrow (read: uninformed, incomplete and short-term = tactical). To clarify how I got to that conclusion would more than double the length of this article, so will have to remain out of scope.

The question is: what is to be done about it?

First problem: the data quality is poor. Efforts are already underway to improve the quality of the data. The answer here is relatively easy to grasp but hard to do. My advice is to make sure, however you must, that you are getting sufficient Volume, Velocity, Value, Variety and Veracity. Hire experts in information theory and/or data science if you don't already have them – and lean on them more if you do. But, in addition, work with good data architects to organize that data and information to enhance public transparency, access and visualization; all three are currently lacking. In doing so, you will increase buy-in from the general public and, maybe, crowd-source an insight or two. Think about it.

Second problem: what to do about the decision-making vis-à-vis the cognitive nature of the decision-maker? In this case, the answer is harder to grasp and even harder to do. Leaders should better understand their own cognitive nature and seek to find help that can balance that nature.

It seems to me, as previously mentioned, that a lot of the decision-making I am observing is based on data rather than information and insight. It also appears that the decision-making is decidedly tactical.

There is a time for tactical for sure; but there is always a time for strategic. Strategic thinking might not have avoided this pandemic, or it might have. Either way, as I like to say "An ounce of insight is worth a pound of information (or a ton of data)." We need more insight and that means more Field Dependent thinking. If leadership isn't wired that way – they need to supplement with people who are. FDs are the statistical minority in our culture. It also means more pattern recognition. It means more visualizations not some number ("The total number of Covid19 cases today is ###.") spewed day after day. Policy-makers are in such a rush to act (understandably – and I'm not saying that they shouldn't) – but some time and focus must be reserved for seeing the bigger picture and asking the deeper questions.

Let me cite one relevant and concrete example. As the pandemic ravaged community after community, we learned quickly that the elderly were the most vulnerable to its ravages. We understood that fact, but we didn't "illuminate" the data around it soon enough. Weeks before it became a "thing" to focus on testing in Long Term Care Facilities – I released an analysis of two demographics: nursing home versus general public. That analysis, or "illuminated insight," clearly showed that nursing homes were tinderboxes where the least spark would ignite the whole place. It showed that over 70% of all deaths were coming from these communities which, tiny as they are, were producing about 20% of all the cases. While the overall Covid19 numbers were rising – the vast majority of those numbers were coming from these small populations. NOT breaking them out, carefully, was masking the progression of the pandemic making it appear more pervasive in the general public than it was in reality.

Leaders failed to see the deeper import of that. The analysis showed that the likelihood of someone carrying Covid19 INTO these tinderboxes was about 1 in 419, but the likelihood of someone carrying Covid19 OUT was 1 in 8. In other words, it was fifty times more likely that the virus was flowing out than in. (Though I hasten to add that there is no empirical data to prove that viral flow because we still are not collecting that specific data – or if we are (contact tracing) – it is not being shared with the public so I can illuminate it.)

These locations were literal Covid19 incubators tucked in the larger community that were likely "spewing" the virus into the larger community – while all the "data-driven decision-making" was focusing on containing the spread in the general public. It was counter-intuitive that we should focus on these tiny contained communities… until the raw data, properly processed into information and then illuminated – became insight. Now, weeks and weeks later, policy-makers are finally focusing on these communities.

Why did it take so long? It would be an oversimplification to say Field Independent, Category Width Narrow, Cognitive Tempo Impulsive thinking fueled by data lacking in all Five Vs – but it also wouldn't be too far from the truth. Balancing that kind of thinking with the "complementary" thinking of Field Dependent, Wide, Reflective thinking – may have brought such insight sooner – and avoided unnecessary deaths, illness, and economic impact.

It is my hope that, at least, this article has helped you understand that "data-driven decision-making" is not a magic-elixir of correct decisions (particularly when the data is questionable). At most, I hope it has helped our leaders understand that there is more to it than that and we would all benefit from considering other (literal) ways of thinking about it.