What in the world were you thinking???

written by DaltonRuer | September 18, 2016

I can tell you that as the father of two daughters, the grandfather of 7 and a 20 year veteran coach/instructor for thousands of adolescent female athletes I've probably said "What in the world were you thinking" at least a thousand times. You know what I mean … children so often do things that just completely defy all logic or known thought processes.

The irony is that as adults we say this mostly in gest as we roll our eyes. All the while knowing full well the problem wasn't what they thought but the fact that they didn't think. They simply allowed themselves to be distracted by something else.

Two years ago I began this blogging journey and I've greatly enjoyed every minute of research, every post and every conversation that was sparked about Data Visualization topics. But I have to be honest watching the battle of hype versus hope unravel right before my eyes on the Data Science and Big Data fronts has kind of driven me crazy. So as this blogging journey is about me, I find that I need to begin at least

intermixing what I'm learning and feel about Data Science and Big Data in with my posts on Data Visualization.

The American Recovery and Reinvestment Act of 2009 pushed $20 Billion into data producing factories in the form of EHR systems. Unlike the common myth data storage isn't cheap. You need bigger data centers, with more racks of disks, which require more power, which require more cooling, which requires more backups, more network bandwidth both internally and externally for redundancy and require more staff to manage the infrastructure. Ugh!

Not really sure what they were thinking. To my knowledge real factories don't produce goods that can't be consumed. Yet here many of you sit 7 years later with data centers full of unused 0's and 1's. Producing them at a frantic pace, but doing nothing with them. Because the push was to collect data but there was no plan on how to utilize the data.

Data Science

Over the past several years I have spent a great many hours consuming free training about Data Science via Coursera. Why would I read "Data Science for dummies" when geniuses like Roger Peng and Jeffrey Leek of Johns Hopkins are teaching Data Science courses. Free courses! Free courses that I can take from the comfort of my own sofa I should add. When they

recently authored Executive Data Science – A guide to training and managing the best Data Scientists, I figured I could afford to pay for their book since I had already MOOChed off of their expertise so much. I bring up their book because they had a profound concept that you may want to write in permanent ink on your monitor … "The key word in data science is not data, it is science. Data science is only useful when the data are used to answer a question. That is the science part of the equation."

No wonder these guys are professors at Johns Hopkins. Seriously as I start this series on Big Data and Data Science I wanted to ensure that we are all on the same footing. As I refer to the term "Data Science" it's always, always, always going to be in regards to applying science to data to answer some business question.

Data Science, like anything new, has been greatly over hyped for sure. Many businesses jumped in with both feet and lots of money praying that they would magically uncover a "Beer and Diapers" or "predicting pregnancy" story of their own that would help their company make a billion dollars in the following quarter. What in the world were they thinking? Data science isn't black magic that you just conjure up answers with … it's science. It follows scientific principles. It takes discipline.

Unfortunately due to some of the expected failures due to a lack of using reasoning, many, many more are sitting on the sidelines watching their business lose money hand over fist ignoring the fact that data science is available. They don't understand how data science works so they simply ignore it instead. What in the world are they thinking?

[Tweet "many more are sitting on the sidelines watching their business lose money hand over fist ignoring the fact that data science is available."]

Is data science for everyone? Of course not. But tucking your head in the sand while other companies use it as a competitive asset just isn't a good business practice. You want to separate "hype" from "hope" so you know if it is right for you then start with "What is the question I am trying to answer with data." Follow that up with "Do I even have the data I need to answer it?" If the answer to both is yes, then allow the science to lead you to the answers that is hidden in your data.

Big Data

One of the reasons for so many dashed hopes and dreams is that some organizations starting building massive data lakes thinking the more data I have the better the answers I can get. They had no business questions in mind they just figured if they assembled enough data files together on disk drives that problems would somehow solve themselves. Quite simply they ignored the science and focused on the data. I don't want you to make the same mistake.

If you are going to undertake anything new like Data Science or Big Data you have to understand that major changes like this require organizational change as well. They aren't just a technical matter. If you are going to go with a Big Data solution then for goodness sakes please start by following sound advice like that found in Benjamin Bowen's book titled Hadoop Operations. He makes it clear that organizations must combine three facets of strategy: Technical, Organizational and Cultural.

The difficulty for many who have succeeded in Analytics but are afraid to jump into Big Data is the simple fact that it's hard for many to truly understand what Big Data really is. I can't blame someone for not wanting to invest in something that they can't understand. At least "science" is a word that people can relate to and that's why Peng/Leek focused on their phrase immediately as they began their book. It gives you a point of reference.

Unfortunately Big Data is an entirely different beast. I wish I could write something profound like "The most important word in Big Data is big" or "The most important words in Big Data" is data. To help you focus. But the truth is the most important word in "Big Data" is neither big, nor data. The most important word to describe it is actually a set of the 3 words: Volume, Veracity and Variety. However the hard part even for the Qlik Dork to explain is that none of them alone explain the concept and you need to refer to them in combination and here is why:

> Volume – Just because your organization has Gajigbytes of data doesn't mean you need to turn to Big Data. Relational database systems, especially Teradata, can be grown to be as large as you will ever need so it's not just volume that forces the issue.
>
> Velocity – Simply means the speed with which the data is coming. There are all sorts of interfaces that handle rapidly moving data traffic so again, that alone doesn't constitute a need for Big Data.
>
> Variety – In the context of the Big Data field it is most often used to refer to the differences between structured and unstructured data. Unstructured data would be things like documents, videos, sound recordings etc. Don't let me shock you when I say this but "I was storing those things into SQL Server 20 years ago as BLOB's (binary large objects.)" So guess what, again this "variety" by itself

isn’t what big data is about.

So what then is Big Data? It is a combination of all 3 of those things and oh by the way you also need to include business components like time and money. Big Data is centered around the fact that you can use commodity hardware including much cheaper disks than you would typically use for large Storage Area Network (SAN) disk infrastructure. The reason that it is typically considered “faster” in terms of storage is that it doesn’t deal with transactions and rows it simply deals with big old blocks of data so massive files are a breeze to store. The fact that it is block/file oriented means it doesn’t really matter what you throw at it. A stack of CSV or XLS or XML files, a bunch of streaming video or HL7 or sound no problem. You throw and go.

So you can store a wide variety of data, quicker and at less cost than you would using a traditional RDBMS type system. Bonus is also the time savings because nobody in IT really needs to be involved in the process once the infrastructure is put in place. You can have data available and within no time your analysts or your data scientists can begin consuming the data. No requirements documents. No prioritization process. No planning meetings. Very little overhead. And oh by the way it allows the business to actually own the process of solving the problems that they business has. Crazy concept I know.

Examples

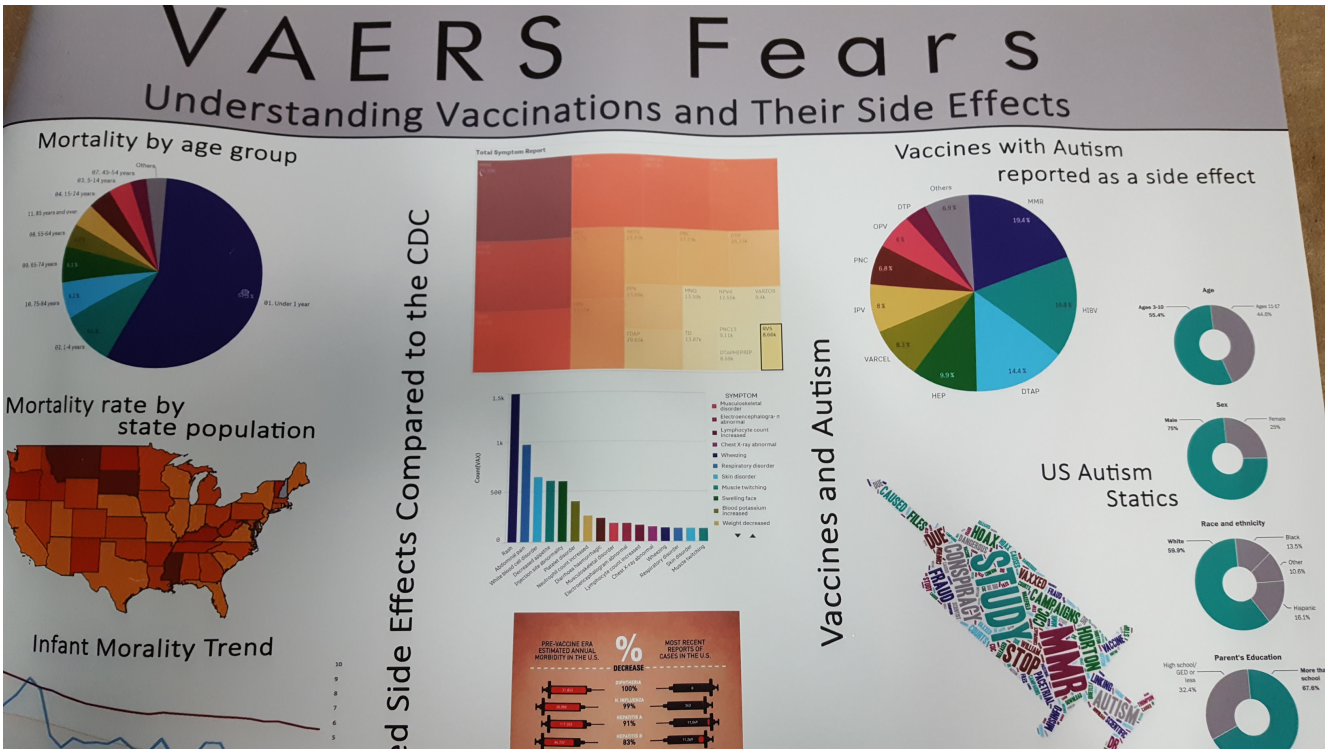
Enough of my musing, let’s just get down to a few practical examples.

Vaccinations and Side Effects

This week I met two of the most wonderful young Data Scientists. Liam Watson and Misti Vogt just graduated from Cal State Fullerton and delivered a presentation at the Teradata Conference in Atlanta, Georgia on a phenomenal use for data

regarding the side effects of vaccinations. In the coming weeks I will be presenting their research and application, but I wanted to quickly plant a seed regarding their work that I think makes an excellent pitch for those of you who may be on the fence about proceeding with Data Science or Big Data.

Much of the "science" of what they did revolved around data that parents completed to report side effects after getting their child vaccinated. The form, like so many in the healthcare and other industries is a typical check this box for this condition, check that box for that condition … Other (Please type in) kind of thing. The check boxes would be considered structured data. The "other" would certainly be considered unstructured 0's and 1's that get manufactured in our EHR factories and left to accumulate dust.



If these two used Static Reporting they would have had no choice but to simply ignore the "other" category and count up how many of A, B, C, D or E were checked. But let's face it if these two were ordinary I wouldn't be talking about them. Instead they chose the path of using Data Science (which says you can't leave data behind just because it doesn't fit your

simple report query model and isn’t clean) and they needed to use Big Data because it provides them with so many wonderful text analytics functions.

What they uncovered was that White Blood Cell Disorder which came from the hand input “Other” text box was the third highest side effect. To me that’s like gold. It’s a discovery that quite simply would be overlooked in a traditional environment because it didn’t fit the “we can only deal with structured data mold.”

There is a lot of time and effort expended in tracking physicians and beating them over the heads if they don’t sign off on documentation in a timely manner. I certainly understand that without their signature the organization doesn’t get paid. But I can’t help but wonder what gold may be lying in the textual notes that physicians dictate daily. Don’t believe your organization is ready for Data Science and Big Data to mine for that gold? Not sure what you are thinking.

Zika

I recently recorded a video showcasing a stunning use of Data Science and Big Data that was created by two of Qlik’s partners, Bardess Group and Cloudera. The application demonstrate the impact that accumulating data quickly from a wide variety of sources like weather, flights, mosquito populations, suspected and reported Zika infections and supply chain data could have when brought to bear on a problem like Zika.

Right now most organizations are still struggling to understand their own costs and understand their own clinical variances. Move to a population health model? Unthinkable for them as they can’t produce the static reports nor consume them fast enough to understand their own patients, let alone begin consuming data from payers, the census bureau etc.

As you watch the video and you hear the variety of data sources involved in the Zika demo, imagine the time and energy that would have to go into a project to do the same thing in a traditional way. As much as I "like" the work they've done to help with the Zika virus issue (and the work is continuing with aid agencies and hospitals), I "love, love, love" the use case it makes for the healthcare world that we need to embrace Data Science and Big Data not run from it because neither fits our current working models.

Summary

Blaise Pascal, the 17th century mathematician, once wrote "People almost invariably arrive at their beliefs not on the basis of proof, but on the basis of what they find attractive." We have science that can help us find truth in data and yet we continue to perpetuate treatment plans based on myths and heresay.

[Tweet "We know our current organizational structures are failing to keep pace with the onslaught of changes and the amounts of data we are generating."]

We know our current organizational structures are failing to keep pace with the onslaught of changes and the amounts of data we are generating. But instead of changing to grow cultures that are more data fluent organizations are converting employees to 2×2 cubes so that they can "collaborate" more. No more data is being consumed but at least the status quo is maintained and employees now get to hear endless conversations with spouses and children.

Would I be wrong if I guessed that your organization has a backlog of hundreds of reports, while the previous 10,000 are seldom even if read? What if I guessed that the morale of the report writers is at an all time low because new requests are far outpacing their ability to generate them?

In his book Big Data for Executives author David Macfie puts it pretty eloquently “In a traditional system the data is always getting to you after the event. With Data Science/Big Data the goal is to get the information into your hands before the event occurs.” Put simply static reporting and traditional processes simply aren’t designed to handle the crisis of overrun data centers. I’m not sure what in the world organizations are thinking that are doubling down on static reports.

To be honest I’m not entirely sure what in the world I was thinking taking so long to write this as my thoughts have been bubbling up for so long. If you have yet to actually begin researching or are among those burying your head in the sand and ignoring Data Science and Big Data then you know what is coming … What in the world are you thinking?