Convert the Unknowns to Knowns

- Perspectives

In the lead up to the invasion of Iraq, the then Secretary of Defense, the loquacious Mr. Donald Rumsfeld, classified the information he had or didn't have into three categories **“known knowns”;** **“known unknowns”** and **“unknown unknowns”**. He also added these prescient remarks, *“And if one looks throughout the history of our country and other free countries, it is the category of the latter (unknown unknowns) that tend to be the difficult ones”.*

Unknowns are difficult ones not just for the Department of Defense, but for any organization. They are typically the sources of surprises and risks and also the sources of missed opportunities. An organization will be well served if it can have a “crystal ball” that will let them know the unknowns. Is there a process that an organization can undertake that will help them get to the unknowns? The answer is YES and it is called Data Science.

But Data Science is an expensive process. First of all, modern organizations are mired in Data. There is the internal Data - that is Data generated through their daily operations. There is external Data - that is Data generated by other entities that is of pertinence to an organization. And there is now a new type of Data, the social media data, which is often referred to as Big Data. All these Data sources could be of relevance to an organization and when analyzed separately and sometimes together hold answers. The exercise of Data Science cannot be just another initiative within an organization, but should be a strategic exercise, that should involve the following key elements.

1. Define your strategic objectives and goals

The objective of Data Science is to find answers that will point the organization in the right way - to expanding the top line, increasing the bottom line, creating new product strategies, aligning marketing strategies and a myriad of other things. An organization will benefit the most if the data exploration can be aligned with the strategic objectives of the organization. What are your plans for the next 12 months, five years and so on? Do you need to break new grounds in product categories, or does your plans call for expanding new market segments, or are you losing market share because of market perceptions or realities of your product quality and so on. Pick the areas where you will have the biggest impact and deploy Data Science to find the unknowns in these areas to have the biggest impact in your organization.

2. Define your principal drivers of synergy

It is also important to recognize that organizations hold their Data in silos, typically by functional areas or organizational boundaries. For example, sales data is held in CRM systems, marketing data is spread across the systems that channel the outreach and campaigns, quality data is held in systems used by plants and so on. Analyzing these data in itself may give only partial truths. For example, your sales could be impacted because there is a quality issue. Analyzing just the sales data may not tell you the underlying reasons for the truths you need exposed. It is important to establish the synergy across the organization of your data islands and include in your exploration the relevant data sets. This sometimes is an iterative process (more about this in our future blogs).

3. Good answers are hard to find. Good questions are even harder.

The data science process begins with a well-posed question motivated by the organization's business needs and strategic objectives. At minimum, we are looking for a question that is (1) quantifiable, (2) empirically supportable and (3) value-driven. The analytical engine that drives data science is of course, mathematics, and therefore our questions must be posed in a way that translates easily into an appropriate mathematical framework: probability, statistics, graph theory, numerical optimization, etc. For example, instead of asking “How can we increase profitability?”, ask “What are the demographic trends for consumers of our most profitable products/services?” . The second requirement simply asserts that we can answer the question by analyzing the right kind of data.

And it doesn't necessarily have to be Big Data either. It may be effective to look at old data in new ways, or in previously unexplored combinations of silo'd data (e.g., Operations + Marketing). Finally, any question that is pursued must have a clear and direct connection to measurable business value. Although science per se is often perceived as conducting research simply for the sake of knowledge, real-world business constraints do not allow for aimless fishing expeditions. The resources committed to data science on behalf of the organization must be constantly weighed against the expected return on investment.

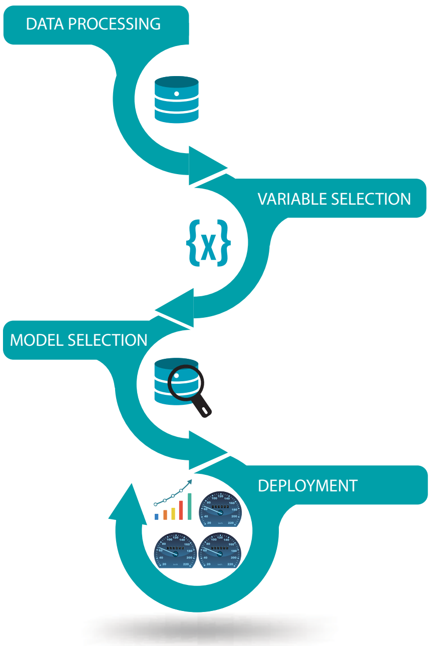
4. Inventory of Data Assets

If you are truly in pursuit of finding the “unknown unknowns”, then what data sources should be considered for investigation? The short answer is “All of them”, for the simple reason that even data entities that are already very well-understood (customer transactions, for example) may reveal previously unexpected patterns and trends when combined with other nontraditional sources. For example, an increasingly common data science practice is to create so-called “mash-ups” of otherwise unconnected data sources – a classic example being the trifecta of GPS data, calendar date and sales volume. The combination of these three dimensions allows visualization of sales quantities across geographical boundaries and over time, providing a powerful view on business activity, patterns and trends. Ultimately, you must select some subset of available data sources, and possibly commi ssion the acquisition of new data if needed.

5. Deploy Data Science

With the question, data and business motivation well in hand, the data science process may be conducted with a reasonable assurance of success. Typical tasks within the process include items such as:

- **Data cleaning, preprocessing & “munging”** – data received from the selected data sources is cleaned and converted into the desired view.
- **Variable selection** – the selected data sources may contain many thousands of variables, only a relatively small portion of which will actually be useful in answering a given question.



- **Model selection & evaluation** – evaluating the different candidate models to determine validity, robustness and performance against previously unseen data.
- **Deployment** – package and encode into a cohesive operational product: data, model and visualization.

Data Science need not be limited to large enterprises with lots of data. It can be a potent and meaningful weapon for any organization, provided it is deployed wisely and strategically.

- ***Kevin Cartier***
- ***Ganesh Iyer***