A Step-by-Step Plan for Implementing Advanced Analytics
Retail / Wholesale / Light Manufacturing

William Vorhies, Chief Data Scientist

October 2015
# Table of Contents

**Overview** ................................................................. 1

**Who is this report for** .................................................. 1

**Investment: are you big enough to benefit?** .......................... 2

**Aren't all businesses different?** ...................................... 2

**Whose problem is this to solve?** ...................................... 4

**The most important start-up consideration** .......................... 4

**Phase 1: Predictive Analytics** ........................................ 6

**Step 1. Predictive Modeling for Basic Scoring and Classification Models:** 6

**Step 2. Predictive Modeling with Appended Demographic and Contextual Data:** 7

**Step 3: Uplift Modeling:** .................................................. 8

**Step 4. Market Basket or Affinity Analysis:** .......................... 8

**Phase 2: Geospatial Analytics** ......................................... 10

**Step 1: Store Performance:** ............................................. 10

**Step 2: Store Location** .................................................... 11

**Phase 3: NoSQL (Hadoop) Based Analysis of Varied Data Types and Sources** ................................................ 12

**Step 1: Text Analytics and Natural Language Processing** .......... 12

**Step 2: Click Stream Analytics:** ......................................... 13

**Step 3: Geophysical Spatial Analysis (GSA):** ......................... 14

**Phase 4: Forecasting, Optimization, Simulation** .................. 15

**Phase 5: Advanced Techniques** ....................................... 16

**About Data-Magnum** ..................................................... 20
OVERVIEW

WHO IS THIS REPORT FOR

Advanced Analytics programs have been adopted by essentially all of the largest corporations, governments, not-for-profits, and universities worldwide. Every week we can read at length about Predictive Analytics, Big Data, Deep Learning, Image Processing, and the dozen other disciplines that make up Advanced Analytics.

Certainly not all of the adopters have embraced all of these techniques. Many are just emerging. Many have applications in smaller niche markets or processes. But the fact remains that the largest and most successful organizations have established Centers of Excellence for Advanced Analytics and are actively competing on and profiting from these investments.

The surprising fact however, is that of all companies, Gartner Research estimates that fewer than 20% have even begun similar programs. Why? There are no doubt many reasons and chief among them would be:

- Not knowing which core techniques of advanced analytics represent sound investments.
- Not understanding how to organize and staff for advanced analytics.
- Not being equipped to estimate either the cost or potential ROI from these investments.
- In short, simply not knowing where to start.

This report is written for the tens-of-thousands of businesses who are large enough to benefit but who remain in the 80% of non-adopters. It presents a simple Step-be-Step plan for getting started with Advanced Analytics, establishing your initial Advanced Analytics group, and investing in those skills, techniques, and projects that have the highest payoff with least investment and risk. In short:

How to Get Started?
INVESTMENT: ARE YOU BIG ENOUGH TO BENEFIT?

What’s the smallest company that can benefit from advanced analytics? We illustrated in a recent article that even a company with only a few thousand customers could benefit from some types of predictive analytics for only a few thousand dollars (read it here). Realistically though, the cost threshold for a real program starts around $750,000 to $1 Million. That’s enough to fund a team of four or five including data scientists and supporting IT staff and analysts. You may need only one or two outside hires to get underway but by the end of the first year a dedicated staff of three to five should be your expectation.

You make the call. Our estimate is that you need revenue in excess of $50 Million and probably more like $150 Million to make this commitment. However, we also estimate that you’ll be recapturing multiples of your $1 Million investment in revenue and margin within two years.

If you are a $250 Million business or larger and you are not actively engaged with Advanced Analytics, you are at a significant competitive disadvantage.

AREN’T ALL BUSINESSES DIFFERENT?

To be sure there are industries with specialized needs. If you are in a highly specialized business like insurance underwriting, aerospace manufacturing, big-web-user all digital companies, or operating a hospital, the sequence of the plan we present may not be exactly right.

However, there are a very large number of companies whose business models incorporate combinations of retail, wholesale, ecommerce, and light manufacturing for whom these steps will be almost universally appropriate.

To make this as simple as possible for the largest number of companies, we present this Step-by-Step Plan for implementing advanced analytics, to get the biggest return with the smallest reasonable investment and least risk.
In laying out this plan, we kept in mind the hypothetical XYZ Corp. that has these characteristics.

1. $500 Million in sales and 1,000 employees from mostly the same line of business.
2. Multiple channels including their own brick and mortar stores, a wholesale channel to retailers, and some B2C digital ecommerce (typical light manufacturing or wholesaler/retailer profile).
3. Some of the product is manufactured in the US but mostly overseas with longer lead times, and they have about 1,000 SKUs.
4. Their products encourage repeat or additional purchases so they want to have an on-going relationship with their customers.
5. At least some of their products are sufficiently complex that they have a pretty good size call center for customer service that can double for in-bound and out-bound sales when not helping current customers.

XYZ Corp. has no Advanced Analytics program. However, if you ask senior management about how they value and use data they will all profess to find high value and great usage. In fact though, XYZ is relying on its fairly modern integrated transactional and BI system with the ability to drill down into historical data. They have some ad hoc reporting, some visualization, and even perhaps a dashboard. They may use modern management applications like SalesForce.com. They also have a large team of data analysts who use Excel and other special programs to answer management’s questions.

What they actually have is a view of data that is almost entirely historical and rearward looking and where forecasting and the problem solving approach based on MSPITR (most senior person in the room).
**WHOSE PROBLEM IS THIS TO SOLVE?**

There is quite a bit of discussion around this point. Many companies new to Advance Analytics assume that their CIO must lead the way and many CIOs can make good leaders for these efforts. The logic is that since the CIO is responsible for safekeeping the data this must be the place to start. This requires your CIO to think not in the usual terms of controlling costs but instead, that your Advanced Analytics Group is a revenue and margin generating activity.

Since many of the opportunities to be addressed lie in the field of customer behavior, senior sales and marketing executives also often make good champions for getting these projects underway.

Whoever is nominated or recognizes the opportunity and volunteers, they must be sufficiently senior to launch a program of this financial scope with potentially company-wide impacts.

The fact is that it is unlikely there is anyone in senior management who has a good grasp of the Big Data, predictive analytic and machine learning techniques and project requirements to independently formulate a plan. Your first hire or consultant will be with a senior Data Scientist who has the skills and experience to formulate and manage these projects, and equally as important, to communicate the benefits, risks, and requirements to senior management.

**THE MOST IMPORTANT START-UP CONSIDERATION**

Let’s take the point of view of the newly engaged Chief Data Scientist. It is unlikely that he will get much more than high level guidance from his executive sponsor. The marching order that this plan assumes is to get fast returns for the lowest reasonable investment with the least risk.

A common business point of view would be to make this opportunity driven. That is, to do the ‘good consultant’ thing, interviewing LOB and process leaders to build a portfolio of opportunities. That’s not necessarily bad but the guidance is to get the most out of those four or five initial Advanced Analytic team members. No sense getting the LOB and process leaders all fired up only to find they may have to wait until year 2 or 3 before
their projects rise to the top.

If you hired well, your Chief Data Scientist will have seen enough real-world projects to know how Advanced Analytics will help in each of these different areas. The real constraint is how to get quick wins with smallest reasonable investment.

Here’s what we suggest is the first real constraint. The list of all possible opportunities may require a Natural Language Processing (NLP) specialist, an optimization specialist, a big data architect, and a number of others.

This Step-by-Step plan with a starting team of only four or five focuses on one set of core skills that will be self-contained and successful.

Instead of going opportunity by opportunity, we suggest you prioritize skill by data science skill.
PHASE 1: PREDICTIVE ANALYTICS

What is the analytic skill set your team should start with to return that biggest bang for the buck? In our experience, start with predictive modeling.

**STEP 1. PREDICTIVE MODELING FOR BASIC SCORING AND CLASSIFICATION MODELS:**

Predictive Modeling is the ability to build scoring and classification models including clustering and segmentation. Here we are focusing on ‘response modeling’ though these same skills can be used for anomaly or fraud detection and a wide variety of forecasting or values estimating.

Keep in mind our premise. Start with the smallest viable team and focus on excellent performance in a single analytic skill set. Your initial team will likely start with an ‘analyst’ who knows the current transactional and BI data. The second member will be a junior data scientist who can do the data-prep for modeling. The third person will be an experienced predictive modeler. The fourth is you as project lead. And the fifth, by the end of the first year will be a model manager responsible for guiding implementation of the models in production systems and monitoring for model refreshes.

Why start here? Cross Sell and Up Sell to Current Customers: XYZ has an on-going relationship with its customers. They are not just one-and-done shoppers. The easiest and least expensive sale you can make is cross sell or up sell to your current customers.
**Where to implement?** This will vary somewhat based on business circumstances but the XYX call center should be an obvious early target. Based on the profile of the caller using existing transactional data for recency, frequency, dollar value, and products purchased, you can build models to determine next most likely product for upsell and cross sell and display appropriate scripts for the CSR in real time while the customer is already on the phone. Similarly you could also model for potential defection of existing customers and use the same approach to promote loyalty and diminish churn.

**New Customer Acquisition:** Depending on the method of outreach used for the acquisition of new customers, for example catalog mailings, these same techniques will create obvious wins there compared to any other intuitive or non-data science technique.

**STEP 2. PREDICTIVE MODELING WITH APPENDED DEMOGRAPHIC AND CONTEXTUAL DATA:**

By the middle of the first year you should be adding demographic append data to the customer records. You can purchase this from any number of data brokers or major information houses like Experian. Demographic append data will make your models significantly more accurate and therefore profitable. Be sure to consider the cost and level of effort required to maintain this new data accurately.

Context data can have to do with the method of purchase (in store or on line, credit card type); it can be geographic, time of year, time of day, or any variety of other variables. Alongside demographic append data, context data will also increase model accuracy. We could have said click stream data from the ecommerce site but we’re intentionally holding off on that one because we would need to add skills to our team or use outside consultants.

When you first start out and don’t have this append data you will necessarily be building ‘good enough’ models. But model accuracy can make big differences in campaign ROI and should be an important focus of your activities. See our earlier article “the value of accuracy in predictive modeling”.

www.Data-Magnum.com
© 2015 all rights reserved
**STEP 3: UPLIFT MODELING:**

This variation on straight response modeling is only valuable if you have a sufficiently large number of customers since it requires modeling both the target and control groups. It prevents you from spending promotional dollars on those who would have purchased even if they were not contacted. The technique is easy to implement if your team firmly understands predictive response modeling, but would not be economic or useful unless you had more than about 100,000 customers. It would be most useful for single-item promotions but less so for broadcast or multi-offer or catalog mailings. See more about uplift modeling in our article “There is Something New Under the Sun”. However, if your circumstances allow, it can increase promotional ROI 4X to 10X.

Note that between steps 1, 2, and 3 you have probably already saved enough money to cover the start-up expense of your advanced analytics group.

We would expect that it would require between 12 and 24 months to reach this point.

**STEP 4. MARKET BASKET OR AFFINITY ANALYSIS:**

This technique can be performed easily by a team already grounded in predictive analytics. In brief, these techniques seek to identify pairs (or triads or baskets) of specific products where the purchase of one is strongly associated with the purchase of the other. Read more about this technique in our article “Affinity Analysis”.
Where to implement?

**Brick and mortar stores:** Market Basket / Affinity Analysis is used to guide merchandise location so the products with strong affinity are typically placed together. Similarly, the strongest pull-through products are often located in such a way to cause the shopper to pass by other types of merchandise.

**Web sites:** Similar to brick and mortar stores, market basket and affinity analysis guides the positioning of products on related pages to increase click-through and sales.

**Pricing and Discounting:** Knowing that two product sets are already strongly associated reduces the need to discount both. This prevents unnecessary discounting and increases gross margin.
**PHASE 2: GEOSPATIAL ANALYTICS**

Remember that we’re being guided by getting the largest, fastest ROI for our efforts. While there are many directions you could go after mastering basic Predictive Analytics, the question is what will provide the biggest return that our current team can provide with the addition of the least additional skills.

The concepts in geospatial analytics (GSA) are very similar to the uplift models in predictive analytics but add the variables of location of an event or customer in space, and potentially within time. Some addition to the team’s skills or the capabilities of the analytic packages you use will be required but they are not disruptively large. It is possible that your primary ‘model jockey’ will have this skill or you may need to add a specialist or upgrade the qualifications of your team.

The most basic uses of geospatial analysis will warm the heart of any CFO or CMO.

**STEP 1: STORE PERFORMANCE:**

If two stores are producing the same revenue and the same gross margin are they therefore performing the same? Absolutely not. It depends on the accessible customer pool in which they are located. The core concept here is to model the characteristics of the in-store customer and to compare them to the characteristics of the surrounding population. It would not be unusual to find that one store is twice as productive at attracting in-store sales as another clearly indicating unequal performance.
GSA allows you to draw geographic boundaries around store locations which can be as simple as Euclidean shapes or as complex as drive-time boundaries. It is also possible to map your in-store customers onto the map to further refine these boundaries.

There are challenges in data gathering to capture the addresses of in-store customers but there are a number of strategies for doing so. This can be as simple as asking for their address, working from mailed catalogue or flyer lists, or reverse engineering data from credit card sales.

It is likely you will need some specialized software as well as the skills to operate them. These costs are a fraction of one FTE so not strategically challenging.

**STEP 2: STORE LOCATION**

Similar in concept to step 1, GSA uses the historical patterns of similar stores to look for appropriate concentrations of potential customers around available retail sites. The potential value of a physical location can be directly predicted by this analysis.

There are more sophisticated applications of GSA but they require the use of a NoSQL database and we don’t get to that level of investment until the next step.
PHASE 3: NOQL (HADOOP) BASED ANALYSIS OF VARIED DATA TYPES AND SOURCES

It’s not particularly expensive to acquire a NoSQL DB and the associated analytic platforms. You can rent the NoSQL DB on AWS or set up your own for a hardware/software cost in the low to mid five figures. But you will need to significantly expand the skills on your team to include DBA and systems architects for NoSQL. If you’re using an analytics platform like SAS, SPSS, or Alteryx, or have good R skills on your team you’re ready to move forward.

The analytic techniques associated with being able to use Big Data (here focusing on a variety of data types) are almost too varied. You know, too many choices is the mother of indecision. We’ll attempt to keep to the game plan and prioritize these by potential payoff.

STEP 1: TEXT ANALYTICS AND NATURAL LANGUAGE PROCESSING

When most people read these phrases they immediately think of Sentiment Analysis and monitoring social media. It is true that controlling the customer experience does require monitoring and responding to negative social media and learning from the good things folks may be tweeting. However we believe the most powerful application is actually in dynamic pricing.

Dynamic Pricing: Using NLP you can design programs to scrape the web for your competitors’ real time competitive pricing information. More practically you can hire specialty firms to do this for you, then store and analyze the output in your NoSQL DB. It is possible to highly automate this whole process but it may be more practical to use some human labor and a definition of ‘real time’ that is closer to overnight or weekly than down to the
It should be obvious that the ability to respond to competitors’ pricing within a very short time frame can increase sales and market share. If you want to be as sophisticated as Amazon you can even vary pricing based on how and what time of day the customer is accessing your web site. You can also create intentional differences in pricing strategies between channels such as the web and in-store. The analysis also allows you to keep an eye on what your wholesale channel is doing with pricing through their independent outlets.

Dynamic pricing can have immediate and substantial economic impact.

**STEP 2: CLICK STREAM ANALYTICS:**

Large numbers of specialized procedures exist for tracking and improving the customers’ experience on your web site, ranging from simple A/B testing through sophisticated presentation of alternate offers based on path. All require the capture and analysis of click stream data. You will need to significantly expand the skills of your team or hire an outside firm to help you with this.

You will need to do a cost-benefit analysis on all these potential projects and any company should commit to a B2C Marketing Automation and Management System as a focus for central control for all these potential activities. (Note that Marketing Automation and Management Systems centralize control over all messages in all channels, not just Content Management of your web site).

While you are working through all those options, we suggest you start here: **Recommenders.**

Recommenders are those displays on our web site that suggest what we would like to watch, what else we would like to buy, and even who we might want to date based on our
previous browsing and purchasing behavior. This is straightforward cross-sell and upsell and that’s why we give this priority.

A few years back it was necessary to work with NoSQL Graph DBs to build recommenders and many of the big players today still use Graph DBs. However, Recommenders can now be built using simple columnar NoSQL DBs (MapR, one of the big-three distributors of Hadoop for example says they can build recommenders in real time during the customer’s web visit). Even companies like SAS provide the capability to build recommenders without resorting to Graph DBs.

Since these lead directly to additional sales and because the technology is relatively straightforward, we suggest you start here.

**STEP 3: GEOPHYSICAL SPATIAL ANALYSIS (GSA):**

Using click stream data and other signals from mobile devices it is now possible to establish a geo-fence around your potential customer not only down to the location of your store but down to where they are standing in the aisle and what merchandise is immediately adjacent to their location.

The implication is pretty straightforward. Send the customer a promotional offer on their mobile device relating to merchandise that is within arm’s reach, or at least inside your store to encourage them to act now.

The largest retailers are actively experimenting with this and reporting some pretty positive results. However, we are now reaching the point in our recommendations where the technology is quite cutting edge and may not be appropriate for all. This technique for example requires significant investment in sensors, support, and analytic talent.
These three analytic techniques have been around for a long time and were born in the discipline of Operations Research. Some pretty sophisticated forecasting can be accomplished with the skills of Predictive Modeling you mastered in Phase 1. Optimization and Simulation use elements of data science combined with mathematics and statistics to select the single best answer to a complex problem.

**Forecasting** is clearly not simply eye-balling an extension of the demand line from its current position. Companies like SAS provide very extensive forecasting packages incorporating all the latest thinking and procedures in projecting outcomes and incorporating what-if scenarios.

**Optimization** is a type of prescriptive analytics that finds a "best" solution from a set of "feasible" solutions, using a mathematical algorithm that maximizes or minimizes a specified objective function subject to constraints. Common approaches to solve this problem include linear programming, integer programming, stochastic programming and constraint programming.

**Simulation** is a predictive analytics approach that involves building a model to imitate a system or process. The goal is to study the behavior of how it works or run "what if" scenarios around a real-world process. The two most common approaches are "Monte Carlo" and "discrete event" simulations. Simulation and optimization are among the most computationally challenging disciplines of advanced analytics and the skills requirement is substantial.

Your CFO or Chief Supply Chain Officer are likely to be the biggest fans of these techniques. And while they can increase revenue and profit, they are frequently thought of as controlling risk or used for forecasting materials lead times in complex manufacturing and supply chain scenarios.
PHASE 5: ADVANCED TECHNIQUES

You could easily spend five or more years mastering Phases 1, 2, 3, and scratching the surface of Phase 4. In the process, producing a lot of incremental revenue and margin. If you follow the press however, you are bombarded by exotic new developments and capabilities all designed to tempt you to jump in at the bleeding edge. Don’t do it. Follow the outline of good quality, cost effective, and proven blocking and tackling laid out here.

Gartner Research is best known for tracking new and emerging technologies and tracking their progress along their “hype cycle”. The following list is drawn from the most current version of their report for Advanced Analytics and Data Science. These are in no particular order.

REAL TIME / IN MEMORY / STREAM PROCESSING / INTERNET OF THINGS

We listed these together because they are frequently combined together in very confusing ways in the press. Real Time is valuable for some but does not necessarily require in memory or stream processing. Note that the first project in Phase 1 had you create customized scripts for real time up sell and cross sell guidance to CSRs dealing with customers and no particularly fancy technology was required. As SSD memory becomes increasingly less expensive you may find yourself having access to in-memory analytics in the normal process of IT upgrades, for example what SAP is currently doing by changing out its normal DBs with in-memory HANA DBs. If you have a sensor-heavy environment or the need to intervene in your customer’s behavior while it is happening there could be value here, but in most cases batch analytics combined with click stream detection will accomplish very much the same thing.
NATURAL LANGUAGE GENERATION (NLG) COMBINES NATURAL-LANGUAGE PROCESSING (NLP)

In Phase 3 you used NLP as the basis for a dynamic pricing program. You may also have moved into monitoring social media to better control your customer’s experience. What’s on the hype curve however is more exotic.

Whereas NLP is focused on deriving analytic insights from text data, NLG is used to synthesize text, written content by combining analytic output with contextualized narratives. NLG is also being used for operational/regulatory report automation (e.g. threat assessments).

Very early in market adoption and penetration, NLG is already being used to reduce the time and cost to conduct repeatable analysis and writing reports on data. These tend to be for required operational and regulatory report automation, in financial services (earnings reports), government (benefits statements or weather forecasts) and in advertising (personalized messages).

GRAPH ANALYSIS

Enthusiasts of Graph DBs like to say that they can do anything that row or column based DBs can do plus a lot more. They’re no doubt correct. The challenge is that working with Graph DBs requires a whole new way of thinking about data and will likely require one or more specialists to be added to your team. Plus many of the high payoff opportunities like recommenders and fraud detection can now be accomplished with more traditional NoSQL DB types.

Graph DBs explore relationships between nodes (entities of interest that may be organizations, people, products, or transactions) and focus on the strength and number of connections with other nodes. They are indeed good at things like route optimization, market basket analysis, fraud detection, social network analysis, CRM optimization, supply chain monitoring, and load balancing.
**DEEP LEARNING**

Although there is a lot of press devoted to deep learning, the number of commercial applications is still quite small, mostly in the area of image detection and recognition. Deep learning is a form of unsupervised modeling which is true machine learning, and is based on neural net algorithms. Let this one mature quite a bit before diving in.

**SPEECH ANALYTICS**

Audio mining/speech analytics embrace keyword, phonetic or transcription technologies to extract insights from prerecorded and, more recently, real-time voice streams. This insight can then be used to classify calls, trigger alerts/workflows, and drive operational and employee performance across the enterprise. See our previous comments on determining whether real-time analytics is something you really need.

**VIDEO ANALYTICS**

In its simplest form, video analytics has been used in manufacturing process control for many years to automatically spot and take action on special conditions on the assembly line, like poorly filled bottles in a bottling line. What’s on the hype curve however is video analytics to evaluate human behavior. This works primarily in real-time video streams to develop context-appropriate automated responses and to guide consumer buying behavior.

**EMOTION DETECTION/RECOGNITION**

Put Audio and Video analytics together and focus them close-up on the human interaction. Detects context sensitive cues in real time streams based on facial expressions, gestures, posture, tone, vocabulary, respiration and skin physiology (temperature and clamminess). Understanding what a person is communicating involves deciphering that individual's modulation scheme. This requires knowledge of social and cultural mores and cues, as well as a familiarity with the individual. Emotion recognition will lead to optimal computing only if the system is able to consider all of these factors.
**Prescriptive Analytics**

Simply put, "prescriptive analytics" describes a set of analytical capabilities that specify a preferred course of action. We have been skeptical about breaking this out as a separate capability since it is based on predictive models with the application of some simple optimization mathematics. We’ve been doing that for a long time without resorting to creating a new category. See our earlier article “Prescriptive Analytics”.

**Machine Learning**

This is another ‘category’ that is very loosely used in the press. We like many others have used ‘machine learning’ to describe the combination of traditional supervised (predictive modeling), unsupervised (clustering), and new NoSQL analytic techniques. Gartner would have us use this phrase in a more tightly controlled way meaning only techniques that are unsupervised (the machine can learn without reference to known examples). This would then encompass very well used clustering and segmentation models along with deep learning, plus the thrust into citizen data science assisted by analytic platforms that operate without the help of expert data scientists.
ABOUT DATA-MAGNUM

Data-Magnum is a Big Data and Predictive Analytics consulting and technology integration company focused across industries and across business processes on the selection and successful implementation of new solutions in Big Data and Predictive Analytics.

With deep understanding of data science, predictive analytics, and machine learning Data-Magnum can organize and execute profitable data discovery projects in organizations of all sizes.

If your organization would benefit from the new technologies of Big Data and NoSQL or NewSQL databases, Data-Magnum can provide the information, education, and assessment necessary for the planning and successful implementation of Big Data projects ranging from the earliest discussions of how Big Data can benefit your company through project execution and maintenance.

Data Magnum is equipped to implement your selected solution or structure and implement pilot projects. We can also provide the talented staff you may need to augment your team and can provide the real-world training to up-skill your existing staff.

About the author: Bill Vorhies is President and Chief Data Scientist at Data-Magnum and a founding member of the American Institute of Big Data Professionals (AIBDP). Bill has practiced as a data scientist and commercial predictive modeler since 2001. His experience in predictive analytics covers ecommerce, scientific and engineering applications, financial markets forecasting, financial services and insurance, manufacturing, sports, healthcare, and the public sector. He is also Editorial Director of DataScienceCentral.com and is a graduate of Princeton University.

He can be reached at:

Bill@Data-Magnum.com

www.Data-Magnum.com