



# The Value of Machine Learning and Data Mining in Pharmaceutical Commercial Analytics

February 2018

# The Value of Machine Learning and Data Mining in Pharmaceutical Commercial Analytics

**George A. Chressanthis, Ph.D.**, Principal Scientist, Axtria Inc.

**Rajnish Kumar**, Director, Axtria Inc.

**Brian Gibbs, Ph.D.**, Principal, Axtria Inc.

## PREFACE

This white paper deals with two topics in the pharmaceutical commercial analytics (PCA) space that are receiving a great deal of attention by drug companies – machine learning (often also referred to as artificial intelligence, thus ML/AI) and data mining (DM). Rather than write two separate white papers, we have combined these topics since they are highly related to each other. Machine-Learned Pharmaceutical Commercial Analytics (MLPCA) provides more value if companies structure the expanding myriad of databases into formats that can be easily analyzed and integrated. DM techniques often can take the form of ML designed to extract valuable insights and uncover previously unknown trends or patterns in the data. Given this heightened interest and perceived value, we wish to investigate the value of MLPCA and DMPCA (Data Mining Pharmaceutical Commercial Analytics).

Before continuing further, what are the definitions of ML and DM? While there is no one universal definition, a search through the Web finds commonality to the ones noted below:

*“The defining characteristic of ML is that it allows computers to learn from data without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can access data and use it to learn for themselves.”*

The Axtria view about ML is slightly different than the common definition found on the Web:

*“ML is a **subset** of AI with a key characteristic being the ability to learn from new data without explicit programming. AI can be defined as the ability of a computer*

*to independently solve problems that they have not been explicitly programmed to address.”*

A precursor to effective ML is data mining (DM), defined through a similar Web search process as done for ML:

*“**Data mining** is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. **Data mining** tools allow enterprises to predict future trends.”*

A key differentiator between ML and DM is that there is frequently a feedback learning loop in ML that provides actionable insight without explicit coding.

As previously noted, while ML and DM are different endeavors. They are linked in that creating value through applying ML cannot occur without available, suitable, and consistently well-defined data to apply ML algorithms. This also means the ability to link databases. This final point is becoming more challenging. New databases are emerging such as those created through patient claims and electronic medical records, social media and digital channel activities, and new customer engagement channels that are becoming more pervasive. The industry is moving away from a sales rep-centric model to one with multiple touch points, and qualitative measures of sales and marketing activities. This white paper will proceed in three parts – Parts 1 and 2 cover MLPCA and DMPCA respectively, and then the final section on overall comments and conclusions about the value of MLPCA and DMPCA.

## PART 1 – MACHINE-LEARNED PHARMA COMMERCIAL ANALYTICS (MLPCA)

### Introduction

MLPCA reveals the endless opportunities in utilizing analytic techniques to address evolving complex questions. How can we make predictions based on data without explicitly programming the model to take various factors into account? A few examples of MLPCA are provided below:

1. How can we group provider accounts into meaningful segments, where members of a segment are like each and distinct from other segments?
2. What are the drivers of success for my launch brand, with how can I obtain updates to identify emerging drivers / barriers?
3. How should we tailor digital campaigns to individual customers by predicting the next best touch (timing, channel, content)?
4. What is the response to our promotional activity and marketing programs, including interaction/sequencing-effects among those programs?

The key feature of ML is that it allows for models to iteratively update prior parameters based on new observations. This key feature of ML is at the roots in statistical debates between Frequentists and Bayesians.

1. Frequentists – Planful hypothesis testing in how knowledge is generated.
  - (a) State defined question and translate into testable hypothesis.
  - (b) Collect data to test the hypothesis.
  - (c) Use statistical methods to make decisions about the hypothesis.
  - (d) Quantifying risk of error in the decision such as  $H_0$  (type I and II errors).
2. Bayesians – Updating of initial hypotheses based on new data is how knowledge is generated.
  - (a) Set initial beliefs (priors).
  - (b) Update degree of beliefs about initial hypotheses given new data (posteriors).
  - (c) Can discover something new or unexpected.

“A breakthrough in machine learning would be worth ten Microsofts.”

**Bill Gates**  
Chairman, Microsoft

The pharma analytics environment is not only data-rich, and the problems to solve are both complex and critical in determining not only brand success but also how company commercial actions ultimately affect patient health and economic outcomes. ML provides the ability to adjust modeling recommendations without extensive human-based effort based on inputs, historical evidence, and behavior. It is best used to update continuously a model with new feeds of data. The next section will look at the state and potential of MLPCA?

### State of ML and Healthcare Potential

Substantial investments are being put into machine learning, with applications and potential especially seen in health care. Below are a few factoids on investments in ML:<sup>1</sup>

1. Total annual external investment in AI was between \$8B to \$12B in 2016, with ML attracting nearly 60% of that investment.
2. While high tech, telecom and financial services are the leading adopters, the greatest increases in profit margins due to AI adoption are seen by the healthcare, financial services, and professional services industries.
3. Healthcare executives that were early AI adopters expect technologies will raise operating profit margins by 5% within three years.

The potential for beneficial ML applications in healthcare are substantial, given spending as a percent of GDP being the largest of any sector of the economy (approaching 20%) and significant need.<sup>2</sup> Below are numerous instances where ML is currently being employed for healthcare:



1. Launch evaluation and calibration. What are the underlying drivers of success for a launch brand, and how can we modify sales and marketing tactics to optimize launch success?
2. Patient treatment flow/persistence. What are the indicators of patient non-persistence that should trigger interventions that will boost compliance to therapy?
3. Dynamic target refinement. Which of our current targets for personal promotion present the greatest threat/opportunity for my brand, and how should we dynamically adjust targeting for personal promotion to achieve optimal effectiveness?
4. Dynamic message refinement. When have messages “worn in” and “worn out” for my brand for specific targets, and how should we modify message strategy for specific targets to optimize brand performance?
5. Use for seeking patterns in patient histories, medical images, improving diagnoses, forecasting the spread of diseases, customizing treatments.
  - (a) ML involved in payments and claims management, using medical histories to forecast population health risks to bring out potential savings.
  - (b) Clinicians trying to offer preventative care after using AI to forecast the spread of disease and anticipate which patients are most at risk.
6. Hospital capacity utilization improvement, including optimizing patient interactions. Using forecasts to help hospital admins schedule staff, negotiate insurance reimbursement rates, set budgets, and optimize inventory levels.
7. Helping diagnose disease and improve operations (e.g., identifying glioblastoma abnormalities in MRI and X-Ray images).
8. Treatments tailored to individual patients based on cell and molecular biology models and identifying biomarkers to suggest the best options.

“Machine learning is the hot new thing.”

**John Hennessy**  
*President, Stanford University*

“...science is not just about **predictions**; it’s also about **explanation** and **understanding**.”

**Pedro Domingos**  
*from The Master Algorithm:  
 How the Quest for the Ultimate Learning  
 Machine Will Remake Our World (2015)*

- 9. Virtual agents as primary patient touchpoints – speech, and image recognition technologies combined with ML used to conduct patient consultations, prescribe drugs, make diagnoses.
- 10. J&J in partnership with SAP used machine learning to estimate customer demand, inventory levels, and product mix.
- 11. CareSkore, a population health management company that leverages machine learning for improved patient care, applies a predictive analytics platform, uses machine learning to determine the likelihood of a patient readmittance to the hospital.

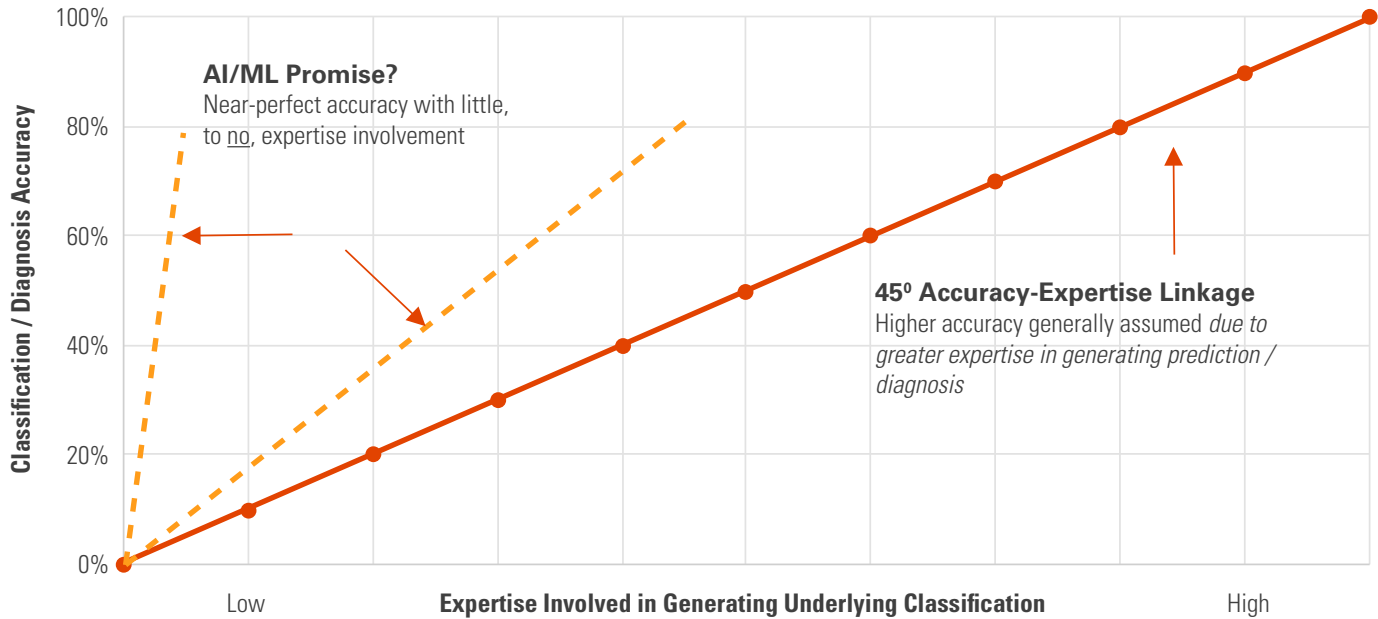
The key takeaways here are the potential for ML/AI capability growth within the pharmaceutical and healthcare sector to affect key outcomes metrics is substantial, that ample funds are being made available for expanding/investing these technologies, and that perceptions are high there will be a strong return on investment and an improvement in operations with their implementation.

**Interest in MLPCA**

Why all the buzz over ML? The core underlying motif for thinking about ML is the so-called 45-degree accuracy-expertise relationship (see **Figure 1**). The dominant world view on many issues has been that classification accuracy increases in lock-step with underlying expertise involved in generating the classification (e.g., MDs, PhD economists, etc are worth the investment). ML offers to promise higher or equal levels of accuracy at lower investment in expertise, or maybe no expertise at all (moving the so-called 45-degree angle closer to 90 degrees).

However, as the next quote notes, can ML be expected to deliver accuracy on its own without the prerequisite expertise to provide the explanation and understanding underling the prediction/diagnosis? The view here at Axtria is that such accompanying expertise is essential to achieve both accuracy and insight.

**FIGURE 1: 45° Classification/Diagnosis Accuracy vs. Expertise Involved Generation**



**Note:** This figure is not drawn to scale.



“

Doctors can be replaced by software – 80% of them can. I’d much rather have a good machine learning system diagnose my disease than the median or average doctor.

**Vinod Khosla**

*Indian-born American engineer, businessman, and billionaire*

”

Is ML more about hype and less about substance? Or, is ML at the peak of the technology “hype cycle” as suggested by the Gartner research organization?<sup>3</sup> There is much promise here but maybe more hype to improve commercial outcomes in the life sciences. Put another way, is ML simply old wine in new bottles?

What interest is there in MLPCA as measured by Google? A check on Google Scholar results, putting in 4 ML phrases which measures the degree of interest in each topic, reveals the following breakdowns captured since 2016 (going back even further would see a dramatic rise in absolute results).

A few insights can be taken away from **Table 1**. First, many results were generated for each phrase since just 2016. Naturally, a reduction in results appeared as the phrasing becomes narrower in focus. Second, while the percentage of 7.8% for pharmaceutical ML may appear small, it is large for focus on just one specific industry area given the many industries and academic areas that apply and study

ML concepts. Also, upon further reflection, this is also not surprising. Healthcare and pharmaceuticals operate in very data-rich and big-data environments. Thus, there are boundless opportunities to use ML to uncover new insights with significant social value in not only improving business efficiency and effectiveness but also health and economic outcomes important to all key stakeholders in the healthcare system. Third, pharmaceutical *analytics* ML generated over one-quarter (26.5%) of results relative to pharmaceutical ML. This is a natural consequence of the growth of decision science tools being applied to ever-growing and available databases to solve increasingly complex business and scientific/clinical problems. Lastly, almost half (47.2%) of pharmaceutical analytics ML is in the commercial area, an indication of the business applications seen through ML.

### How Does MLPCA Work?

Advocates note ML adapts to real-world complexities and provides data-driven insights. Knowledge engineering (KE) explicitly encapsulates and integrates knowledge into a model to make predictions and drive decisions. Given input data (information plus answers), ML then utilizes relationships, patterns, dependencies, and hidden structures by applying algorithms and techniques to derive output (e.g., optimum model). Data with billions of records and thousands of attributes can be utilized when making predictions. Dependencies among features and hidden structures in the data can be identified, allowing for more robust decision-making, e.g., determining the number of calls for each high-value target based on various characteristics available in the data.

**TABLE 1: Google Scholar Results (since 2016) for Machine Learning (ML), Pharmaceutical ML, and Pharmaceutical Analytics ML, Pharmaceutical Commercial Analytics ML, and Percentage Relative to a Baseline Figure<sup>a</sup>**

| Phrase                                 | Results (#) | Percentage (%) <sup>b</sup> |
|--|-------------|-----------------------------|
| Machine Learning (ML)                  | 222,000     |                             |
| Pharmaceutical ML                      | 17,300      | (1) 7.8%                    |
| Pharmaceutical Analytics ML            | 4,580       | (2) 26.5%                   |
| Pharmaceutical Commercial Analytics ML | 2,160       | (3) 47.2%                   |

**Notes:** <sup>a</sup> Google Scholar search done on 5 October 2017. <sup>b</sup> Number in parentheses represents the phrase for the baseline figure of results, with the % being relative to the preceding number directly above.

Here are two examples in the process how ML can be used to answer key pharma business questions:

**1. Prediction of occurrence of Parkinson’s disease.**

Use of ML algorithms help predict the probability of Parkinson’s disease through various choice attributes using alternative estimation techniques (i.e., logistic regression, artificial neural networks, and support vector machines). Evaluate model performance on accuracy & false positives / negatives by method employed.

**2. Next best action in marketing.** Provide an alternative to the traditional push-marketing strategy that has grown increasing obsolete with evolving customer behavior, creating an environment where customers demand the control of the flow of information. Use of ML techniques can answer the following questions:

- (a) Channel Evaluation - Are our promotional investments working well together, and should we consider changing coordination across channels?
- (b) Customer Preference - Which channel does a customer prefer to receive promotion, or, which is most effective for this customer?

“ I fear that AI may replace humans altogether. The development of full artificial intelligence could spell the end of the human race. ”

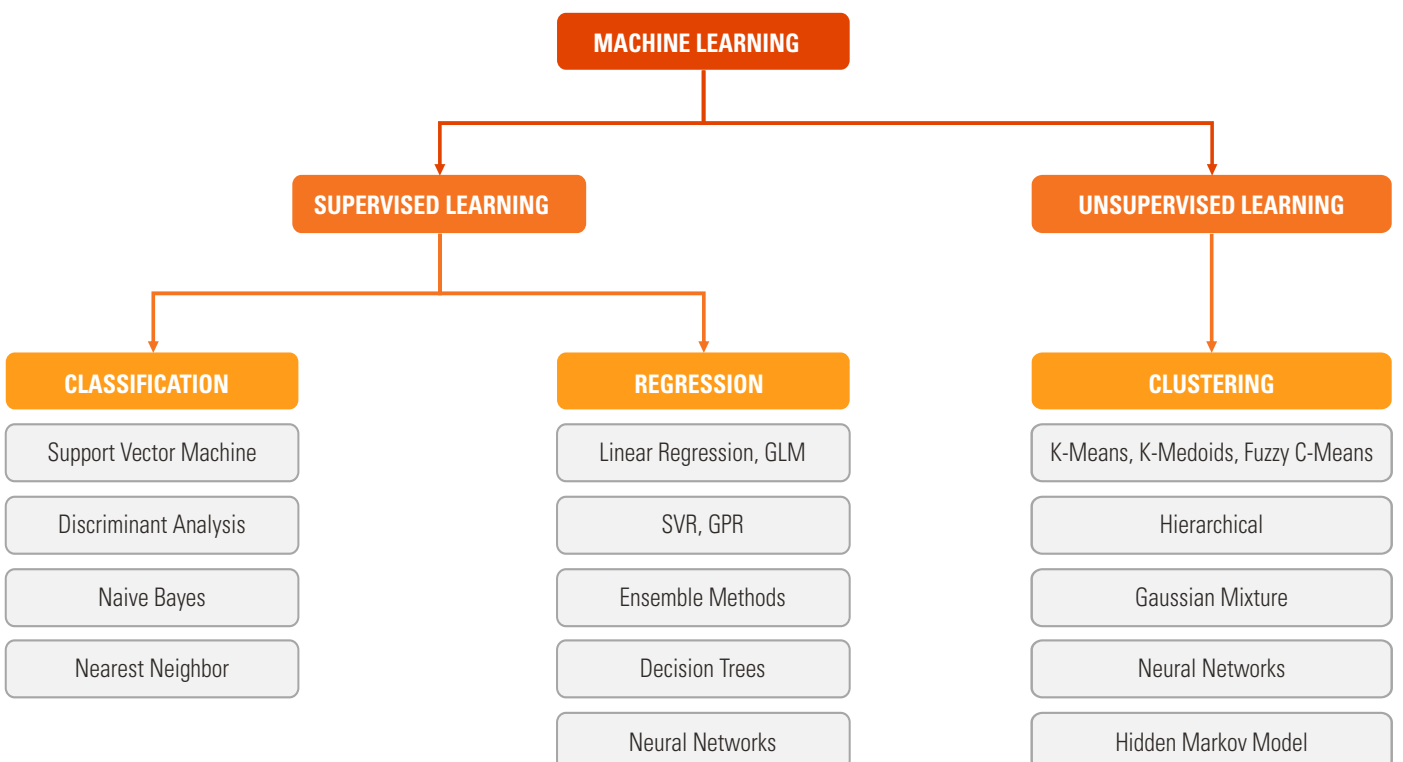
**Stephen Hawking**  
Theoretical physicist

- (c) Message Evaluation - How can we modify the content of our marketing efforts (particularly details) to enhance their effectiveness?

Core ML techniques can be classified into *supervised* versus *unsupervised* contingent on existence of a known Y for learning (see **Figure 2**). Note that many view Reinforcement Learning as a separate component of ML, but we view this as integral to the entire ML concept.

Supervised learning is the development of an algorithm linking input variables (Xs) to an explicit known output variable (Y) known as *labelled* data. The Y can be continuous or dichotomous, and is much more frequently dichotomous in ML with the commercial use of the algorithm generally

**FIGURE 2: Supervised vs. Unsupervised Machine Learning**



“

AI has quickly demonstrated its potential to be a game-changer in many areas of health care.

**Arlene Weintraub**

*Trend 9, "A Smart Computer Will "Treat" You" from 10 Health Care Trends That Will Affect You from U.S. News & World Report, 2018 edition Best Hospitals*

”

allowing for accurate classification of future events. Typical examples of applied supervised learning for **classification** purposes in the life sciences are the following:

1. HCP choice to put a new patient on drug therapy.
2. HCP choice to put a new patient on specific drug.
3. HCP response to a specific call to action from digital campaign.

4. Patient response to a patient support enrollment initiative.
5. Patient script fulfilment (fill vs. abandon).
6. Patient drug persistence months after initiation (on drug, not on drug).
7. Patient drug adherence months after initiation.
8. Patient *transition* along disease progression stages.

Unsupervised learning is the development of an algorithm linking input variables (Xs) for classification purposes where Y is unknown (*unlabeled* data). A typical example of applied unsupervised learning in life science is the identification of HCP/patient/account segments, or the identification of atypical HCP behaviors and interactions that trigger specific pre-planned action.

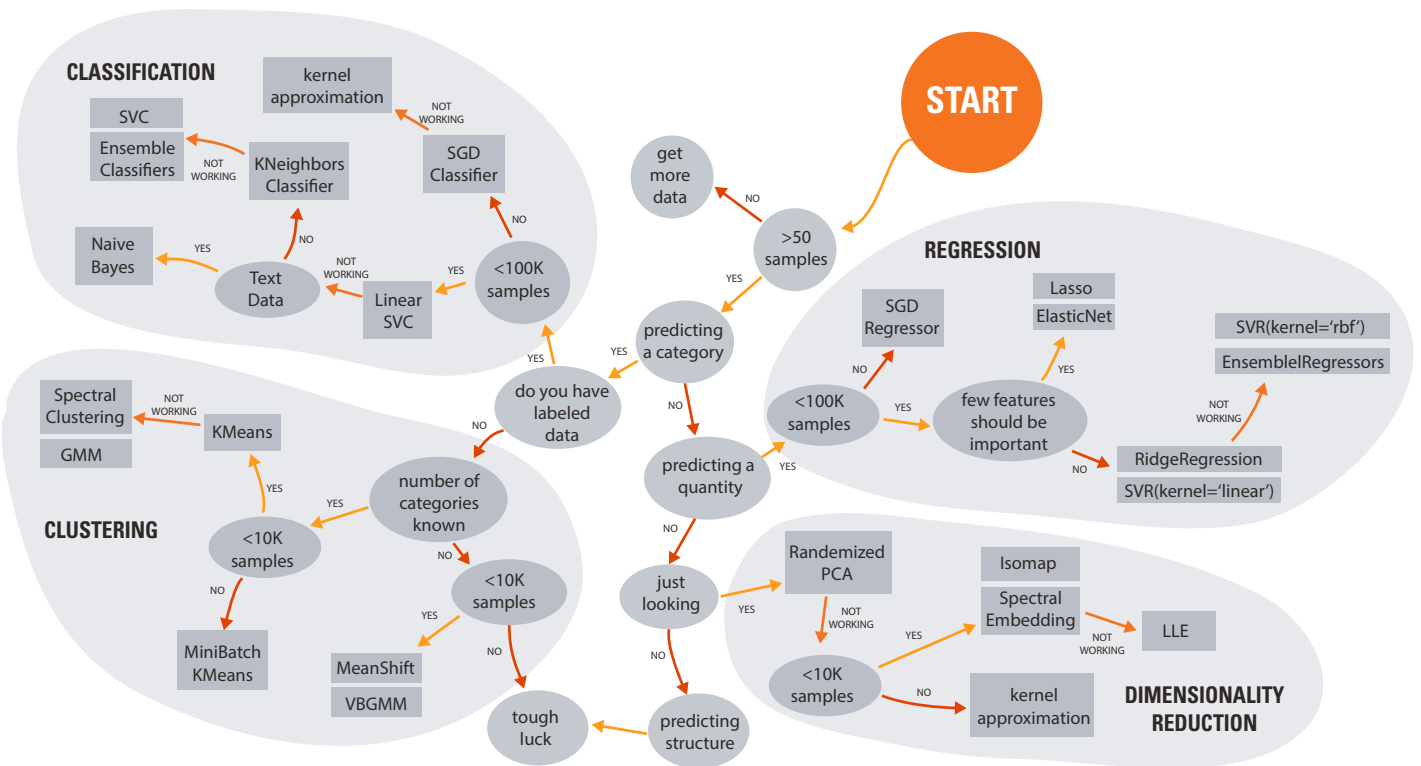
Machine learning draws on a wide range of analytical techniques (see **Figure 3**). Matching the right techniques to a given business problem is part science and part art, and is often an iterative process that combines inductive and deductive analyses.

**FIGURE 3: Machine Learning Algorithms**





**FIGURE 4: Machine Learning Toolkit Roadmap**



A ML toolkit roadmap provides guidance on which analytic techniques are appropriate for the question and data (see **Figure 4**):

1. Regression and Classification – Regression approaches area appropriate for labelled data with continuous values; classification approaches for labelled data and non-continuous values.
2. Clustering and Dimensionality Reduction – Clustering approaches are appropriate for predicting a category from *unlabelled* data; Dimensionality Reduction for predicting continuous value from *unlabelled* data.

Finally, contrasting with ML is Knowledge Engineering (KE). KE is the codification of knowledge generated by experts and/or expert-driven analytics *driving directed pre-planned action*. Commercial examples in the life sciences of KE are the following:

1. Targeting and Timing - Sophisticated analytics indicates that the 10,000 HCPS are high value targets (HVTs) for personal prometom based on potential value, institutional constraints, behavioral predispositions and likelihood to

engage. The pre-planned action: KE indicates 48 details/year are optimal for Tier 1 targets; 24/year for Tier 2, and 12/year for Tier 3. *KE does not inform when those interactions should occur.*

2. Channel Communication Strategy - Sophisticated analytics indicates channel preference for each HVT with 30% preferring details only, 15% open to personal (P) and non-personal (NPP) with preference for P, another 15% open to P and NPP with a preference for NPP; 10% NPP only and 30% unlikely to engage through any channels. The pre-planned action: HVT-specific sales and marketing teams campaign design.
3. Message Communication Strategy - Sophisticated analytics reveals personality segments for each HVT with relatively unique preferences for content about treatment(s), coverage for those treatments, patient-treatments are most appropriate for, and communication style preferences. The pre-planned action: HVT-specific content and communication preferences for campaign design.

ML provides updates and refinements to KE pre-planned commercial actions based on new data. Using the preceding (1) – (3) examples, ML calibration provides the following actions:

1. Targeting and Timing - Algorithms applied on real-time data on HVT behaviors (testing, diagnoses, referrals, treatments), interaction with promotional efforts, constraints and other factors to identify when targeting should be adjusted. Calibrated action: ML informs when those interactions should occur, and whether baseline marketing strategy and tactics should be adjusted.
2. Channel Communication Strategy - Algorithms applied on real-time data on HVT interaction with, and response to, promotional efforts via specific channels. Calibrated action: ML informs whether baseline channel marketing strategy and tactics should be adjusted, and how.
3. Message Communication Strategy - Sophisticated analytics reveals personality segments for each HVT with relatively unique preferences for content about treatment(s), coverage for those treatments, patient-treatments are most appropriate for, and communication style preferences. Calibrated action: KE informs HVT-specific content and communication preferences for campaign design.

In summary, MLPCA can be useful in numerous ways to aid in decision-making:

1. Allows for continuous generation of output from a broad selection of models to be applied as new data is entered. The models run can range from the simplistic (e.g., logistic regression) to the more complex (e.g., structural models, neural networks).
2. Generates prediction outcomes in a highly efficient manner. The decision-maker can test for the sensitivity of prediction outcomes based on a range of methodologies employed.
3. Produces prediction results with accuracy without human intervention and potential bias that they can impose on the solution(s).

4. Creates more value from pharma big data by uncovering potential insights that can then be applied in inferential statistical models to determine the cause behind changes in the dependent variable. There is no limit to the data size and number of data fields to be analyzed through ML, which is especially beneficial given the data-rich environment in healthcare and pharmaceutical commercial analytics.

### Limitations of MLPCA and Conclusions

What then are the limitations of employing ML in pharma commercial analytics? A search through the Web finds a few common themes:

1. There is a specialized training dimension to each application employed on the data.
2. Learning must generally be supervised (i.e., have a clearly defined outcome/output to model on).
3. Challenges around learning from the analysis.
4. The belief is that big-data is required, when the future of the pharma environment is more about how to work with smaller datasets. The pharma industry trend to focus on specialty medicines, especially targeted personalized medicines, as seen in oncology, means working with smaller patient and physician-level databases will be the ever-growing challenge. Thus, deeper ML will become more challenging given limited data issues.
5. ML models are fine for prediction and correlation, but less well-suited for understanding causation. Only through linking an appropriate theoretical model with an empirical model design can there be a causal understanding behind variations in the dependent variable.
6. Insights from ML are only as good as the databases employed, the ability to link different datasets, the algorithms programmed, and how to select which method produces the best result. Human bias will enter here on all these noted steps.
7. Problems of model overfitting or underfitting can plague ML.

8. Errors or mistakes using ML in healthcare and pharma commercial analytics means potentially lives at stake. How can this risk be minimized?
9. There is a “black box” problem with certain algorithms, which presents issues for explaining results and in understanding why & how the model underperforms if there are environment changes that affect ML outcomes.

The above limitations do not mean that MLPCA should not be employed. What they do mean is to understand what benefits can come from ML while recognizing its limitations, and taking steps to mitigate the risks. Enormous amounts of data available are not being utilized to help inform and make data-driven risk-mitigated decisions. ML presents endless possibilities to finally capitalize on the insights that are hidden in untouched data. As marketing science professionals in the pharmaceutical industry, we should be able to anticipate an outcome and adjust our recommendation(s) accordingly without extensive human-based effort based on inputs, historical evidence, and behavior resulting in less biased information to make more robust business decisions.

## PART 2 – DATA MINING PHARMACEUTICAL COMMERCIAL ANALYTICS (DMPCA)

### Introduction

A few examples of data mining pharmaceutical commercial analytics (DMPCA) reveals the endless opportunities in utilizing this process:

1. Looking at physician-level prescriber data and associated factors over time to predict what and when physicians will likely switch to another brand or generic drug.
2. Viewing APLD (anonymized patient-level data) and associated factors over time to predict what patients will engage in generic drug switch-backs to a branded drug after making a brand-to-generic substitution or having started on generic drug therapy.
3. Analyzing claims data and all associated fields usually contained in such databases to predict what patients will likely benefit from drug therapy intervention.
4. Investigating physician-level prescribing data over time and putting together a list of physician attributes that will predict the greatest impact from detailing.



“If you mine the data hard enough, you can also find messages from God. [Dogbert]”

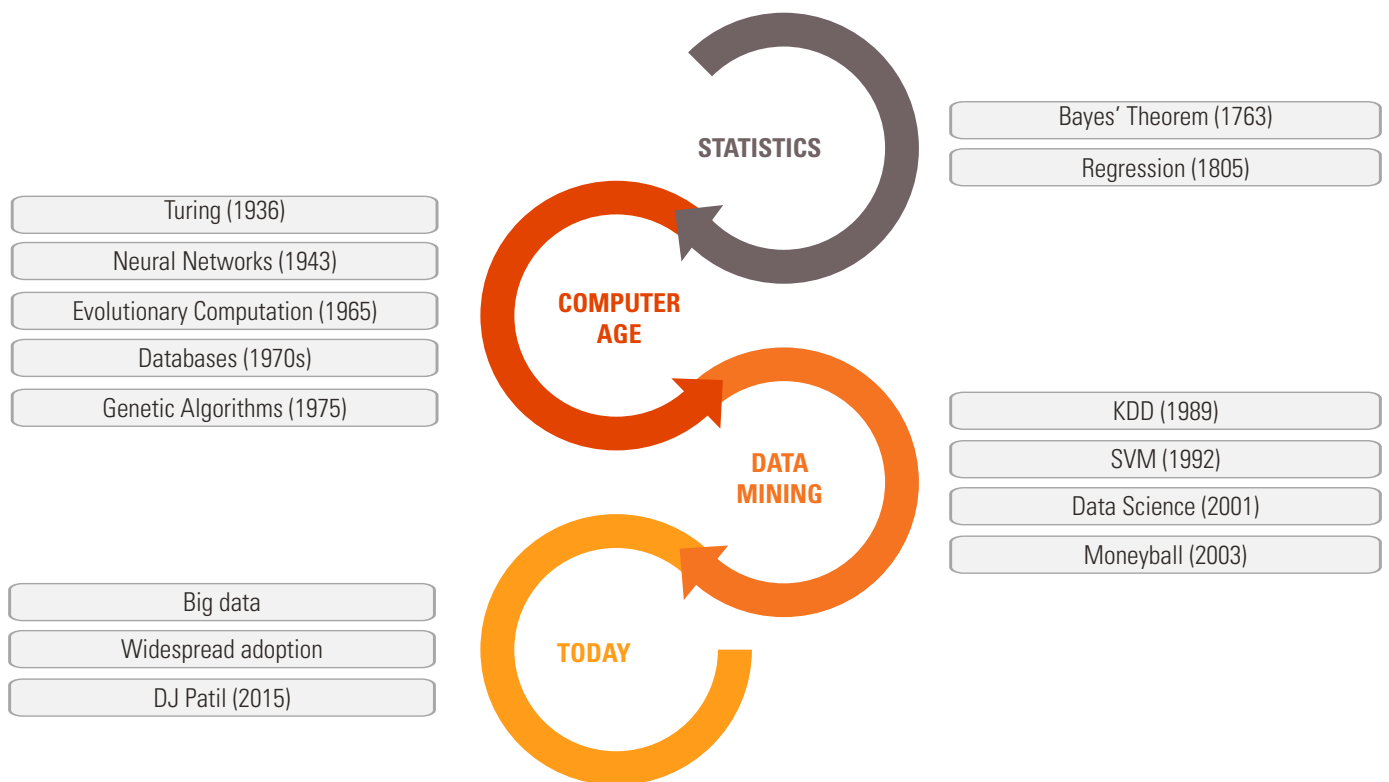
**Scott Adams**  
Creator of the Dilbert comic strip and author

These are just a few of the endless examples that proponents will highlight regarding the benefits from using DMPCA. The question this part will address is simply what is the value of DMPCA?

DM is currently a topic of discussion in all industries, but especially in healthcare given the extensive array of databases available and pressing questions to address. Interest in DM, and especially in healthcare, is not new. There is a long history regarding the origins of data mining, with more recent developments still approaching 30 years old (see **Figure 5**).<sup>4</sup>

Perhaps the renewed buzz about DM in healthcare stems from the combination of large and complex databases, technological developments that can rapidly and accurately analyze this data, and business/social/individual needs that are unique to healthcare. Discussions among people about the value of DM are likely to bring about as many opinions as the number of individuals! While the discussion here is on DMPCA, it is well-understood the benefits of DM can include its use in the scientific/clinical areas too. For example, using DM-models to predict what projects in early the development stages have the best opportunity to reach the clinical trial stage. The pharma commercial analytics environment is extremely data-rich and becoming increasingly more granular. This means the data sizes are getting much larger, thus the problem of finding those needles in the hay stack. The problems to solve are also becoming complex and addressing critical issues that ultimately affect patient health and economic outcomes, not to mention important business questions.

**FIGURE 5: The History of Data Mining**



Source: Exastax. The history of data mining. Exastax, published online 20 January 2017, available at <https://www.exastax.com/big-data/the-history-of-data-mining/>.

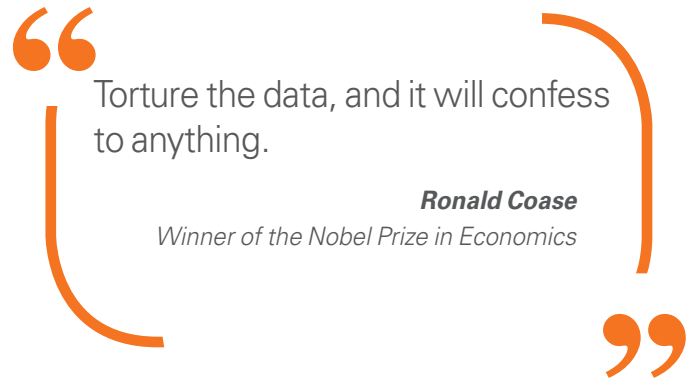
## Interest in DMPCA

What is the interest in DMPCA as measured by Google? A check on Google Scholar results, putting in 4 DM phrases which measures the degree of interest in each topic, reveals the following breakdowns captured since 2016 (going back even further would see a dramatic rise in absolute results).

A few insights can be taken away from **Table 2**. First, many results were generated for each phrase since 2016. Naturally, a reduction in results appeared as the phrasing becomes narrower in focus. Second, while the percentage of 15.8% for pharmaceutical DM may appear small, it is quite large for focus on one specific area given the many industries and academic areas that apply and study DM. However, with further reflection, this is also not surprising. Healthcare and pharmaceuticals operate in very data-rich and big-data environments. Thus, there are boundless opportunities to use DM to uncover new insights with significant social value in not only improving business efficiency and effectiveness but also health and economic outcomes important to all key stakeholders in the healthcare system. Third, pharmaceutical *analytics* DM generated one-quarter of results relative to pharmaceutical DM. This is a natural consequence of the growth of decision science tools being applied to ever-increasing and available databases to solve increasingly complex business and scientific/clinical problems. Lastly, pharmaceutical commercial analytics DM generated almost one-half (48.1%) of all pharmaceutical analytics DM. This is one example of the high interest and potential value seen in DM in the pharma commercial analytics space.

## Where is DMPCA Best Utilized?

So where is DMPCA best utilized to add business value? The preceding and following quotes accurately reflect where that value is best located (and not).



1. Looking at the data from many different perspectives is always worthwhile as a first step to better understand movements in your objective variable and observe potential reasons for its variations. DMPCA can be used to find the association between two or more variables, which can lead to an understanding of variations in your objective variable. Hidden patterns in the dataset can be uncovered and used to identify strong associations between your objective variables with other variables or patterns of variables.
2. A DM exercise may uncover ideas that were previously unknown to the researcher. For instance, anomaly detection can lead to new insights by identifying unusual patterns that do not conform to expected behavior. These anomalies can take the form of point anomalies when the anomaly is a single instance of data, contextual anomalies when the anomaly is context specific, or collective anomalies when a set of data instances collectively help in detecting anomalies. By identifying these anomalies, we can separate true signals from noise as data changes over time, allowing us to uncover anomalous behaviors or patterns that would have otherwise gone unnoticed. Using DM for anomaly detection greatly simplifies analyses by surfacing the

**TABLE 2: Google Scholar Results (since 2016) for Data Mining (DM), Pharmaceutical DM, and Pharmaceutical Analytics DM, Pharmaceutical Commercial Analytics DM, and Percentage Relative to a Baseline Figure<sup>a</sup>**

| Phrase                                 | Results (#) | Percentage (%) <sup>b</sup> |
|--|-------------|-----------------------------|
| Data Mining                            | 139,000     |                             |
| Pharmaceutical Data Mining             | 21,900      | (1) 15.8%                   |
| Pharmaceutical Analytics Data Mining   | 5,590       | (2) 25.5%                   |
| Pharmaceutical Commercial Analytics DM | 2,760       | (3) 48.1%                   |

**Notes:** <sup>a</sup> Google Scholar search done on 2 October 2017. <sup>b</sup> Number in parentheses represents the phrase for the baseline figure of results, with the % being relative to the preceding number directly above.



“

Some of the best theorizing comes after collecting data because then you become aware of another reality.

**Robert J. Shiller**

*Winner of the Nobel Prize in Economics*

”

most relevant trends, exposing previously unknown ideas and freeing up valuable time to find actionable insights.

3. The above quote appropriately acknowledges how DM can bring about better theoretical model development by improving one's understanding of the environment surrounding the dependent variable. There are often vast amounts of data available – so much data that it cannot all be used in the model. Unless you have the ability and time to spend copious amounts of time going through every variable, much of the data as well as any insights it might hold goes unused and largely unknown. DM can be used to simplify and summarize the data in a way that we can easily understand, allowing us to make conclusions about specific cases based on the patterns we have observed. Looking at a summarized and simplified view of the data allows us to incorporate the appropriate data into the model and fully understand the environment and circumstances surrounding the dependent variable.
4. DM is best used as a prediction tool. Suppose a response variable is to be regressed against many covariates. Using all covariates in your model would likely lead to severe multicollinearity or difficulty in identifying your regression coefficients. In addition, your standard errors will be unacceptably large, and predictions may be very inaccurate. In this situation, you need to reduce the number of covariates in your model to a smaller set that appropriately represent the full list of covariates. The econometrics method of principal components (PC) can be used to take a set of variables selected through a DM exercise and collapsed into a set of “super variables”.

5. DM-generated models used in conjunction with the application of dynamic tools can be powerful for prediction as the opening examples indicate, though with a key limitation: DM cannot be used to build a causal model. There are advantages of using a causal model over a DM model due to the inherent differences in hypotheses and outputs of these two techniques. A more thorough explanation of the limitations of DM is noted in the next paragraph.

Further, applying data mining to unstructured data can yield numerous valuable insights, such as:

1. Understanding opinion/awareness metrics from social media.
2. Mining an open payment database to see patterns through the following:
  - (a) Identifying KOLs.
  - (b) Prioritizing certain therapeutic areas based on market opportunity.
  - (c) Competitor targeting.
  - (d) Understanding speaker program HCP selection and spend.
  - (e) Benchmark spending for various HCP targeting, such as, spend volume by type of HCP promotion (meals, etc.).
  - (f) Geographic coverage of targeting.

### Limitations of DMPCA and Conclusions

Where is using DM not appropriate? Unfortunately, DM-generated models are often confused with models needed to explain why or what causes variations in the dependent variable. DM is not a substitute for causal model building based on theoretical foundation which is the only way to address the “why” question. How does the practitioner answer the question about “what causes or why the dependent variable changes, and what management-control mechanisms can be utilized to affect that change? Only through establishing a strong linkage between a theoretical foundation (to establish the causal relationships), which in turn drives the empirical model specification, and then





through the right estimation technique, can provide this confidence through statistical inference analysis. While the empirical statistical analysis will still yield associations, when affirmed model coefficients are linked to a strong theoretical foundation, the model results can speak more strongly about suggesting causality. Making significant business decisions using this approach will yield more robust solutions than those generated through a DM process. Lastly, when using DM-generated models for prediction, the practitioner will not really know what factors are really at work affecting the prediction. Causal-based models have the added advantage when used for prediction in that you have greater confidence in measuring the predictive change in the dependent variable by factors under your control in the model. A pharma example would be how changes in macroeconomic trends at the MSA-level (metropolitan statistical area) affect local drug demand. A causal-based model is developed showing how MSA-level macroeconomic and management control variables affect drug demand. A company cannot change the course of MSA-level macroeconomic trends. But by understanding the degree of effect these variables have on drug demand, the practitioner can develop counter-measures by using management control mechanisms at the MSA level to mitigate any effects as defined in the model. Such business

applications could not be done using a DM-generated prediction model.

In closing, the use of DMPCA can be a very powerful process that can yield tremendous benefits for a drug company. The opportunities are endless. The key however is to understand when DM is best utilized and the benefits it can derive, while also realizing its limitations. Listen to the data and be wiser for it. Use DM-generated models for predictive power, not for statistical inference. Don't use data mining as a substitute for more formal causal-based model building and the thinking that goes behind it. This part ends with the following quote and the importance of understanding what the data is trying to say to you.

“Data will talk to you if you're willing to listen.”

**Jim Bergeson**  
Division VP, Customer Engagement  
at BI WORLDWIDE

### **PART 3 – OVERALL CONCLUSIONS ABOUT THE FUTURE AND VALUE OF MLPCA AND DMPCA**

It is widely believed that ML and DM are right on the cusp of becoming among the most important breakthroughs that the healthcare industry has seen in years. Applications of these techniques for improvements in drug discovery have already occurred,<sup>5</sup> given increases in R&D risks and costs.<sup>6</sup> However, wide-spread applications in pharmaceutical commercial analytics have come more recently. Pharma companies are now looking to these techniques to identify additional sources of value, especially as performance-based payer contracts and treatment paradigms decisions by providers are increasingly focused on health & economic outcome measurements. Pharma companies have responded by gaining access to claims data and electronic medical records to see how management control variables and various types of interventions from payers, providers, patients, and public policy can affect outcome measurements. The expansion of these large databases, coupled with expanded sales and marketing channels (especially through digital and social media) and associated metrics, have generated substantial opportunities for value to be extracted through data mining and machine learning techniques. Further,

increasing complexities of the pharma environment also mean that machine learning can provide earlier anticipation of and faster responses to new threats (e.g., biosimilar market entry, changes in payer drug formulary, or predictions about changes in physician decision drug choices) and opportunities.

Predictive analytics suggestion engines and chat bots will be an integral part of a brand sales and marketing toolkit enabling them to maximize the use of data and truly bring information and insight at the point of sales. Beyond sales and marketing, machine learning and data mining together present an opportunity for companies to assist in the development of a new commercial model design for a rapidly evolving pharma environment. The challenge for companies is to ensure they are properly used to deliver value-added insight in an increasingly complex and turbulent pharma environment. Despite the inherent challenges that accompany the applications of ML and DM, the emergence of these capabilities will lead to a complete transformation of pharmaceutical commercial analytics, introduce the ability to gain new insights and value from new and existing drugs faster than ever before, and realize positive effects from drug utilization on both health and economic outcomes.

# References

1. Columbus L. McKinsey's state of machine learning and AI, 2017. *Forbes*, published online 9 July 2017, available at <https://www.forbes.com/sites/louiscolumbus/2017/07/09/mckinseys-state-of-machine-learning-and-ai-2017/#6d5b5ca675b6> (accessed 23 January 2018).
2. Bughin J, Hazan E, Ramaswamy S, et al. Artificial intelligence: the next digital frontier? McKinsey Global Institute, June 2017.
3. Gartner, Inc. Gartner's 2016 hype cycle for emerging technologies identifies three key trends that organizations must track to gain competitive advantage. *Gartner, Inc.*, published online 16 August 2016, available at <https://www.gartner.com/newsroom/id/3412017> (accessed 23 January 2018).
4. Exastax. The history of data mining. *Exastax*, published online 20 January 2017, available at <https://www.exastax.com/big-data/the-history-of-data-mining/> (accessed 23 January 2018).
5. Balakin K and Ekins S. *Pharmaceutical data mining: approaches and applications for drug discovery*. Hoboken, NJ: Wiley, December 2009.
6. DiMasi J, Grabowski H and Hansen R. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics* 2016; 47: 20-33.



**George A. Chressanthis, Ph.D.**

Principal Scientist, Axtria Inc.  
300 Connell Drive, Suite 5000  
Berkeley Heights, NJ 07922  
E: [george.chressanthis@axtria.com](mailto:george.chressanthis@axtria.com)



**Rajnish Kumar**

Director, Axtria Inc.  
300 Connell Drive, Suite 5000  
Berkeley Heights, NJ 07922  
E: [rajnish.kumar@axtria.com](mailto:rajnish.kumar@axtria.com)



**Brian Gibbs, Ph.D.**

Principal, Axtria Inc.  
300 Connell Drive, Suite 5000  
Berkeley Heights, NJ 07922  
E: [brian.gibbs@axtria.com](mailto:brian.gibbs@axtria.com)

**Contact Us**

+1-877-9AXTRIA  
[info@axtria.com](mailto:info@axtria.com)

**Disclaimer**

Axtria® understands the compliance requirements behind personalization and we do not work with any personally identifiable data that can identify an end-customer of a business.


We have the strictest data security guidelines in place as we work with businesses to improve the experience for their customers.

 [www.axtria.com](http://www.axtria.com)

 [facebook.com/AxtriaInc/](https://facebook.com/AxtriaInc/)

 [info@axtria.com](mailto:info@axtria.com)

 [linkedin.com/company/axtria](https://linkedin.com/company/axtria)

 [@AxtriaConnect](https://twitter.com/AxtriaConnect)

Founded in 2009, Axtria® is a Big Data Analytics company which combines industry knowledge, analytics and technology to help clients make better data-driven decisions. Our data analytics and software platforms support sales, marketing, and risk management operations in the life sciences, finance, retail, and technology industries. We serve clients with a high-touch on-site and onshore presence, leveraged by a global delivery platform that focuses on reducing the total cost of ownership with efficient execution, innovation, and virtualization.

For more information, visit [www.axtria.com](http://www.axtria.com)

Follow Axtria on Twitter, Facebook and LinkedIn

Copyright © Axtria Inc. 2018. All Right Reserved