Dr. Jeffrey Strickland is the proprietor of Simulation Educators, Inc. He has been performing modeling and simulation for over 20 years, both in the private sector and in defense. He has earned advanced degrees in mathematics, and has taught mathematics and operations research at several institutions. He has authored a number of technical books on this subject. He resides in Colorado.

This book is about really about Big Data Analytics. Yet, it is written for people who do not have a strong background in data analysis or statistics. So one might think of this as a survey of analytics for ordinary people. This is not to say the the rest us (Geeks) are extraordinary. I suppose I use ordinary because I do not care for the "Dummy" series of books. Consquently, you will not find any formulas or tutorials on tools like SAS, SPSS and R. If you are looking for this content, you need "Predictive Analytics using R".

Data science and analytics have been used as synonyms on occasion. In reality data science includes data modeling, data mining, data analysis, database architecture and so on. Analytics is what we do to make sense of the data. That is, we take data and turn it into information for business decision makers. This, our course, implies that we translate our data science jargon into English, so that ordinary people can use the results. After all, all of our work boils down to, "Is it useful to decision makers?".

Data Science and Analytics for Ordinary People is a collection of blogs I have written on LinkedIn over the past year. As I continue to perform big data analytics, I continue to discover, not only my weaknesses in communicating the information, but new insights into using the information obtained from analytics and communicating it. These are the kinds of things I blog about and are contained herein.

If you are a manager, marketer or any other ordinary person, this book is for you.

ISBN 978-1-329-28062-5
90000
9 781329 280625

ID: 16909563
www.lulu.com

Data Science and Analytics for Ordinary People

Jeffrey Strickland

lulu

Data Science and Analytics for Ordinary People

Jeffrey Strickland

Data Science and Analytics for Ordinary People

By

Jeffrey S. Strickland

Data Science and Analytics for Ordinary People

Printed in the United States of America

First Printing, 2015

ISBN 978-1-329-28062-5

Published by Lulu, Inc.

Table of Contents

Acknowledgements

The author would like to thank colleagues Adam Wright, Adam Miller, Matt Santoni. And Olaf Larson of Clarity Solution Group. Working with them over the past two years has validated the concepts presented herein.

A special thanks to Dr. Bob Simmonds, who mentored me as a senior operations research analyst.

Preface

Ordinary people included anyone who is not a Geek like myself. This book is written for ordinary people. That includes manager, marketers, technical writers, couch potatoes and so on.

Data Science and Analytics for Ordinary People is a collection of blogs I have written on LinkedIn over the past year. As I continue to perform big data analytics, I continue to discover, not only my weaknesses in communicating the information, but new insights into using the information obtained from analytics and communicating it. These are the kinds of things I blog about and are contained herein.

Data science and analytics have been been used as synonyms on occasion. In reality data science includes data modeling, data mining, data analysis, database architecture and so on. Analytics is what we do to make sense of the data. That is, we take data and turn it into information for business decision makers. This our course implies that we translate our data science jargon into English.

The book is organized into nine logical sections: Big Data, Analytics, Models, Statistical Matters, Data Science Concerns, Applications, Operations Research, Tools and Advice.

PART I – Big Data

Life before Big Data

Headlines

“

Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society.

— Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska (Dec 2008)

“

Big data: The next frontier for innovation, competition, and productivity.

— James Manyika, *et al* (May 2011)

The term “big data” was included in the most recent quarterly online update of the Oxford English Dictionary (OED). So now we have a most

authoritative definition of what recently became big news: “data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.” The term, however, may have appeared as early as 1944. [1]

Background

Predictive modeling and analytics are getting a lot of attention these days, perhaps in light of corporate America’s “discovery” of Big Data. Well, I have news for you, “big data” has been around longer than you think and corporate anyone did not “discover it”.

NASA’s Bigger Data

The Space Race and Cold War gave us big data, and NASA and DoD had the supercomputers to amass and store incredible amounts of data. Of note, on February 11th, 2000 NASA, launched the shuttle Endeavour on an 11-day mission. The mission, The Shuttle Radar Topography Mission (SRTM) was a joint project between the National Geospatial-Intelligence Agency (NGA) and the National Aeronautics and Space Administration (NASA). The objective of this project is to produce digital topographic data for 80% of the Earth’s land surface (all land areas between 60° north and 56° south latitude), with data points located every 1-arc second (approximately 30 meters) on a latitude/longitude grid. The absolute vertical accuracy of the elevation data will be 16 meters (at 90% confidence). [2]

One of the products, Global (Longitude 180 W-180 E, Latitude 90 N-90 S), including SRTM data for Longitude 180 W-180 E, Latitude 60 N-56 S region, contains 173 GB for the elevation raw data mosaic with gaps filled. Follow-on missions collected additional data and certainly “bigger data” in terms of storage (hundreds of gigabytes).

Other Bigger Data

While working at the Missile Defense Agency’s Threat Modeling Center we would build 6 degrees of freedom (DoF) threat missiles and a single missile could generate over a terabyte of data for a simulation. Most

scenarios included more than one threat missile. I think we once had a data file that was on the order of 50 terabytes.

Surveillance data from space also exceeds most people's concept of big data, and we have been collecting it since the Cold War. I once had to write an algorithm that would take the enormous amount of data and filter it (along with some data reduction) using things like value of information (VOI), or else it would over-saturate our networks and be useless.

Conclusion

The point is, "bigger data" has been around a lot longer than "big data". The growth of data corresponds with the growth in storage capacity, which corresponds to the growth in technology, and if NASA had not raced to get men on the moon, we might still be drawing histograms on graph paper.

Works Cited

1. Gil Press, "A Very Short History Of Big Data", *Forbes*, May 9, 2013
2. Strickland, J. (2011). Using Math to Defeat the Enemy: Combat Modeling for Simulation. Lulu.com. ISBN 978-1-257-83225-23

BIG DATA, small data, Horses and Unicorns

Every so often a term becomes so beloved by media that it moves from 'instructive' to 'hackneyed' to 'worthless,' and Big Data is one of those terms….

— Roger Ehrenberg

Somehow, I missed the funeral, but BIG DATA is dead. The phrase "big data" now has no value ... it's completely void of meaning. For those of us who have been around long enough, the mere mention of the phrase is awful enough to induce a big data migraine — please pass the big data Tylenol, no not strong enough, the big data Vicodin.

I didn't know it was sick…

It wasn't too long ago that Big Data was healthy as a horse. Then the horse flies got really bad. They brought disease and the horses got sick. But the media, I mean the horse flies, continued to buzz around, and the illness got worse.

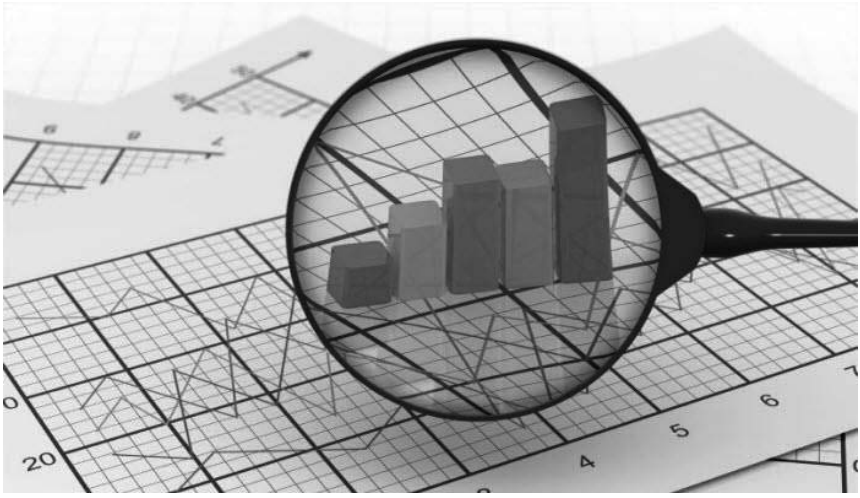
The end came quickly…

After the horse flies finished their work, the worms came and the horses got sicker. I do not have anything against venders, I mean worms, but they wouldn't let a downed horse recover. Then one day the horses died…all of them. We were not left with ponies—they had been forgotten. So it would seem that we were left no horses of any kind.

Life after the horses…

With the horses all gone, the unicorns emerged. They were smarter than the horses…like smart horses with a horn. The unicorns were so smart that they could predict the future…kind of like predictive unicorns. I hope they are real ... and the unicorn flies stay away.

Part II - Analytics



What is Business Analytics?

Business analytics (BA) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. Business analytics makes extensive use of statistical analysis, including descriptive and predictive modeling, and fact-based management to drive decision making. In recent years, prescriptive modeling has also taken a role in BA. It is therefore closely related to management science and operations research. Business analytics can answer questions like why is this happening, what if these trends continue, what will happen next (that is, predict), what is the best that can happen (that is, optimize).

Types

Business analytics is comprised of descriptive, predictive and prescriptive analytics, these are generally understood to be

descriptive modeling, predictive modeling, and prescriptive modeling.

Descriptive analytics

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive analytics provides simple summaries about the sample audience and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries may either form the basis of the initial description of the data as part of a more extensive statistical analysis, or they may be sufficient in and of themselves for a particular investigation.

Predictive analytics

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events. Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models that are in many areas, such as marketing, where they seek out subtle data patterns to answer questions about customer performance, such as fraud detection models.

Prescriptive analytics

Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen. Further, prescriptive analytics suggests decision options on how to take

advantage of a future opportunity or mitigate a future risk and shows the implication of each decision option. Prescriptive analytics can continually take in new data to re-predict and re-prescribe, thus automatically improving prediction accuracy and prescribing better decision options. Prescriptive analytics ingests hybrid data, a combination of structured (numbers, categories) and unstructured data (videos, images, sounds, texts), and business rules to predict what lies ahead and to prescribe how to take advantage of this predicted future without compromising other priorities

Applications

Although predictive analytics can be put to use in many applications, I list a few here:

- Analytical customer relationship management (CRM)
- Clinical decision support systems
- Collection analytics
- Cross-sell
- Customer retention
- Direct marketing
- Fraud detection
- Portfolio, product or economy-level prediction
- Risk management
- Underwriting

This is expansion one of my previous posts, "Predictive Analytics for Almost Everyone."

Predictive Analytics & Modeling

Predictive analytics—sometimes used synonymously with *predictive modeling*—is not synonymous with statistics, often requiring modification of functional forms and use of ad hoc procedures, making it a part of data science to some degree. It does however, encompasses a variety of statistical techniques for modeling, incorporates machine learning, and utilizes data mining to analyze current and historical facts, making predictions about future.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions. Predictive models are not restricted to business, for they are used to predict anything from the reliability of an electronic component to the success of a manned lunar landing. These model, however, are usually stochastic models that can be used in a simulation.

Predictive analytics is used in actuarial science (Conz, 2008), marketing (Fletcher, 2011), financial services (Korn, 2011), insurance, telecommunications (Barkin, 2011), retail (Das & Vidyashankar, 2006),

travel (McDonald, 2010), healthcare (Stevenson, 2011), pharmaceuticals (McKay, 2009), defense (Strickland, 2011) and other fields.

Definition

Predictive analytics is an area of *Data Science* that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown events of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs (Strickland J., 2013). The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Not Statistics

Predictive analytics uses statistical methods, but also machine learning algorithms, and heuristics. Though statistical methods are important, the Analytics professional cannot always follow the "rules of statistics to the letter." Instead, the analyst often implements what I call "modeler judgment". Unlike the statistician, the analytics professional—akin to the operations research analyst—must understand the system, business, or enterprise where the problem lies, and in the context of the business processes, rules, operating procedures, budget, and so on, make judgments about the analytical solution subject to various constraints. This requires a certain degree of creativity, and lends itself to being both a science and an art,

For example, a pure statistical model, say a logistic regression, may determine that the response is explained by 30 independent variables with a significance of 0.05. However, the analytics professional knows that 10 of the variables cannot be used subject to legal constraints imposed for say a bank product. Moreover, the analytics modeler is aware that variables with many degrees of freedom can lead to overfitting the model. Thus, in their final analysis they develop a good

model with 12 explanatory variables using modeler judgment. The regression got them near to a solution, and their intuition carried them to the end.

Additionally, the Analytics professional does not always look for a hypothesis *a priori*. Consequently, they may use a machine learning algorithm, such as *Random Forests*, which does not depend upon statistical assumptions, but instead they "learn" from the data.

Types

Generally, the term predictive analytics is used to mean predictive modeling, “scoring” data with predictive models, and forecasting. However, people are increasingly using the term to refer to related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making, but have different purposes and the statistical techniques underlying them vary.

Predictive models

Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models that are in many areas, such as marketing, where they seek out subtle data patterns to answer questions about customer performance, such as fraud detection models. Predictive models often perform calculations during live transactions, for example, to evaluate the risk or opportunity of a given customer or transaction, in order to guide a decision. With advancements in computing speed, individual agent modeling systems have become capable of simulating human behavior or reactions to given stimuli or scenarios.

The available sample units with known attributes and known performances is referred to as the “training sample.” The units in other sample, with known attributes but un-known performances, are

referred to as “out of [training] sample” units. The out of sample bear no chronological relation to the training sample units. For example, the training sample may consists of literary attributes of writings by Victorian authors, with known attribution, and the out-of sample unit may be newly found writing with unknown authorship; a predictive model may aid the attribution of the unknown author. Another example is given by analysis of blood splatter in simulated crime scenes in which the out-of sample unit is the actual blood splatter pattern from a crime scene. The out of sample unit may be from the same time as the training units, from a previous time, or from a future time.

Descriptive models

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do. Instead, descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

Decision models

Decision models describe the relationship between all the elements of a decision—the known data (including results of predictive models), the decision, and the forecast results of the decision—in order to predict the results of decisions involving many variables. These models can be used in optimization, maximizing certain outcomes while minimizing others. Decision models are generally used to develop decision logic or a set of business rules that will produce the desired action for every customer or circumstance.

Applications

Although predictive analytics can be put to use in many applications, we outline a few examples where predictive analytics has shown positive impact in recent years.

Analytical customer relationship management (CRM)

Analytical Customer Relationship Management is a frequent commercial application of Predictive Analysis. Methods of predictive analysis are applied to customer data to pursue CRM objectives, which involve constructing a holistic view of the customer no matter where their information resides in the company or the department involved. CRM uses predictive analysis in applications for marketing campaigns, sales, and customer services to name a few. These tools are required in order for a company to posture and focus their efforts effectively across the breadth of their customer base. They must analyze and understand the products in demand or have the potential for high demand, predict customers' buying habits in order to promote relevant products at multiple touch points, and proactively identify and mitigate issues that have the potential to lose customers or reduce their ability to gain new ones. Analytical Customer Relationship Management can be applied throughout the customer lifecycle (acquisition, relationship growth, retention, and win-back). Several of the application areas described below (direct marketing, cross-sell, customer retention) are part of Customer Relationship Managements.

Clinical decision support systems

Experts use predictive analysis in health care primarily to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease, and other lifetime illnesses. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care. A working definition has been proposed by Robert Hayward of the Centre for Health Evidence: "Clinical Decision Support Systems link health observations with health knowledge to influence health choices by clinicians for improved health care." (Hayward, 2004)

Collection analytics

Every portfolio has a set of delinquent customers who do not make their payments on time. The financial institution has to undertake collection activities on these customers to recover the amounts due. A lot of collection resources are wasted on customers who are difficult or impossible to recover. Predictive analytics can help optimize the allocation of collection resources by identifying the most effective collection agencies, contact strategies, legal actions and other strategies to each customer, thus significantly increasing recovery at the same time reducing collection costs.

Cross-sell

Often corporate organizations collect and maintain abundant data (e.g. customer records, sale transactions) as exploiting hidden relationships in the data can provide a competitive advantage. For an organization that offers multiple products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to efficient cross sales, or selling additional products to current customers. This directly leads to higher profitability per customer and stronger customer relationships.

Customer retention

With the number of competing services available, businesses need to focus efforts on maintaining continuous consumer satisfaction, rewarding consumer loyalty and minimizing customer attrition. Businesses tend to respond to customer attrition on a reactive basis, acting only after the customer has initiated the process to terminate service. At this stage, the chance of changing the customer's decision is almost impossible. Proper application of predictive analytics can lead to a more proactive retention strategy. By a frequent examination of a customer's past service usage, service performance, spending and other behavior patterns, predictive models can determine the likelihood of a customer terminating service sometime soon (Barkin, 2011). An intervention with lucrative offers can increase the chance of retaining the customer. Silent attrition, the behavior of a customer to slowly but steadily reduce usage, is another problem that many companies face.

Predictive analytics can also predict this behavior, so that the company can take proper actions to increase customer activity.

Direct marketing

When marketing consumer products and services, there is the challenge of keeping up with competing products and consumer behavior. Apart from identifying prospects, predictive analytics can also help to identify the most effective combination of product versions, marketing material, communication channels and timing that should be used to target a given consumer. The goal of predictive analytics is typically to lower the cost per order or cost per action.

Fraud detection

Fraud is a big problem for many businesses and can be of various types: inaccurate credit applications, fraudulent transactions (both offline and online), identity thefts and false insurance claims. These problems plague firms of all sizes in many industries. Some examples of likely victims are credit card issuers, insurance companies (Schiff, 2012), retail merchants, manufacturers, business-to-business suppliers and even services providers. A predictive model can help weed out the “bads” and reduce a business's exposure to fraud.

Predictive modeling can also be used to identify high-risk fraud candidates in business or the public sector. Mark Nigrini developed a risk-scoring method to identify audit targets. He describes the use of this approach to detect fraud in the franchisee sales reports of an international fast-food chain. Each location is scored using 10 predictors. The 10 scores are then weighted to give one final overall risk score for each location. The same scoring approach was also used to identify high-risk check kiting accounts, potentially fraudulent travel agents, and questionable vendors. A reasonably complex model was used to identify fraudulent monthly reports submitted by divisional controllers (Nigrini, 2011).

The Internal Revenue Service (IRS) of the United States also uses predictive analytics to mine tax returns and identify tax fraud (Schiff, 2012).

Recent advancements in technology have also introduced predictive behavior analysis for web fraud detection. This type of solution utilizes heuristics in order to study normal web user behavior and detect anomalies indicating fraud attempts.

Portfolio, product or economy-level prediction

Often the focus of analysis is not the consumer but the product, portfolio, firm, industry or even the economy. For example, a retailer might be interested in predicting store-level demand for inventory management purposes. Or the Federal Reserve Board might be interested in predicting the unemployment rate for the next year. These types of problems can be addressed by predictive analytics using time series techniques (see Chapter 18). They can also be addressed via machine learning approaches which transform the original time series into a feature vector space, where the learning algorithm finds patterns that have predictive power.

Risk management

When employing risk management techniques, the results are always to predict and benefit from a future scenario. The Capital asset pricing model (CAM-P) and Probabilistic Risk Assessment (PRA) examples of approaches that can extend from project to market, and from near to long term. CAP-M (Chong, Jin, & Phillips, 2013) “predicts” the best portfolio to maximize return. PRA, when combined with mini-Delphi Techniques and statistical approaches, yields accurate forecasts (Parry, 1996). @Risk is an Excel add-in used for modeling and simulating risks (Strickland, 2005). Underwriting (see below) and other business approaches identify risk management as a predictive method.

Underwriting

Many businesses have to account for risk exposure due to their different services and determine the cost needed to cover the risk. For example, auto insurance providers need to accurately determine the amount of premium to charge to cover each automobile and driver. A financial company needs to assess a borrower's potential and ability to pay before granting a loan. For a health insurance provider, predictive analytics can

analyze a few years of past medical claims data, as well as lab, pharmacy and other records where available, to predict how expensive an enrollee is likely to be in the future. Predictive analytics can help underwrite these quantities by predicting the chances of illness, default, bankruptcy, etc. Predictive analytics can streamline the process of customer acquisition by predicting the future risk behavior of a customer using application level data. Predictive analytics in the form of credit scores have reduced the amount of time it takes for loan approvals, especially in the mortgage market where lending decisions are now made in a matter of hours rather than days or even weeks. Proper predictive analytics can lead to proper pricing decisions, which can help mitigate future risk of default.

Technology and big data influences

Big data is a collection of data sets that are so large and complex that they become awkward to work with using traditional database management tools. The volume, variety and velocity of big data have introduced challenges across the board for capture, storage, search, sharing, analysis, and visualization. Examples of Big Data sources include web logs, RFID and sensor data, social networks, Internet search indexing, call detail records, military surveillance, and complex data in astronomic, biogeochemical, genomics, and atmospheric sciences. Thanks to technological advances in computer hardware—faster CPUs, cheaper memory, and MPP architectures—and new technologies such as Hadoop, MapReduce, and in-database and text analytics for processing big data, it is now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for new insights (Conz, 2008). Today, exploring big data and using predictive analytics is within reach of more organizations than ever before and new methods that are capable for handling such datasets are proposed (Ben-Gal I. Dana A., 2014).

Analytical Techniques

The approaches and techniques used to conduct predictive analytics can broadly be grouped into regression techniques and machine learning techniques. [Condensed]

Regression techniques

Regression models are the mainstay of predictive analytics.

- Linear regression model
- Ridge regression
- LASSO (Least Absolute Shrinkage and Selection Operator)
- Logic regression
- Quantile regression
- Multinomial logistic regression
- Probit regression

Classification and regression trees

- Hierarchical Optimal Discriminant Analysis (HODA)
- Classification and regression trees (CART)
- Decision trees
- Multivariate adaptive regression splines (MARS)

Machine learning techniques

Machine learning, a branch of artificial intelligence, was originally employed to develop techniques to enable computers to learn.

- Neural networks
- Multilayer Perceptron (MLP)
- Radial basis function (RBF)
- Naïve Bayes
- K-Nearest Neighbor algorithm (*k*-NN)

Criticism

There are plenty of skeptics when it comes to computers and algorithms abilities to predict the future, including Gary King, a professor from Harvard University and the director of the Institute for Quantitative Social Science. People are influenced by their environment in innumerable ways. Trying to understand what people will do next assumes that all the influential variables can be known and measured accurately. “People’s environments change even more quickly than they

themselves do. Everything from the weather to their relationship with their mother can change the way people think and act. All of those variables are unpredictable. How they will impact a person is even less predictable. If put in the exact same situation tomorrow, they may make a completely different decision. This means that a statistical prediction is only valid in sterile laboratory conditions, which suddenly isn't as useful as it seemed before." (King, 2014)

Tools

Tools change often, but SAS appears to be the industry standard, and I relay heavily on SAS Enterprise Modeler for my job. Be that as it may, I use R a great deal and find SPSS (particularly SPSS Modeler) useful for some things. Personally, I prefer R.

References

Barkin, E. (2011). *CRM + Predictive Analytics: Why It All Adds Up.* New York: Destination CRM. Retrieved 2014, from http://www.destinationcrm.com/Articles/Editorial/Magazine-Features/CRM---Predictive-Analytics-Why-It-All-Adds-Up-74700.aspx

Hayward, R. (2004). Clinical decision support tools: Do they support clinicians? *FUTURE Practice*, 66-68.

Nigrini, M. (2011). *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations.* Hoboken, NJ: John Wiley & Sons Inc.

Parry, G. (1996, November–December). The characterization of uncertainty in Probabilistic Risk Assessments of complex systems. *Reliability Engineering & System Safety, 54*(2-3), 119–1. Retrieved 2014, from http://www.sciencedirect.com/science/article/pii/S0951832096000695

Schiff, M. (2012, March 6). *BI Experts: Why Predictive Analytics Will Continue to Grow.* Renton: The Data Warehouse Institute. Retrieved 2014, from http://tdwi.org/Articles/2012/03/06/Predictive-Analytics-Growth.aspx?Page=1

Strickland, J. (2013). *Introduction toe Crime Analysis and Mapping.* Lulu.com. Retrieved from http://www.lulu.com/shop/jeffrey-strickland/introduction-to-crime-analysis-and-mapping/paperback/product-21628219.html

Strickland, J. (2013). *Introduction toe Crime Analysis and Mapping.* Lulu.com. Retrieved from http://www.lulu.com/shop/jeffrey-strickland/introduction-to-crime-analysis-and-mapping/paperback/product-21628219.html

A Dangerous Game We Play

We, the scientists who perform predictive modeling, hold a crystal ball in our hands, making concise and accurate predictions based on quality data and scientific procedures. Or, do we?

What are our assumptions?

We build models based on these assumptions:

- The data is good
- The statistical processes are failsafe
- The models are exact duplicates of the real system or phenomenon

But here is the reality:

- The data is rarely good
- We rarely implement the statistics correctly
- The models are simplistic abstractions of real stuff

Perhaps this is why Professor Box said, “all models are wrong, but some are useful.”

Can we really predict the Unpredictable?

Can we really predict the unpredictable, embodied in a time and space not yet obtained? How good is our crystal ball? Professor Box made his livelihood and reputation on predictive models, as do I. So how do we succeed with “bad” models?

Well, first the model are not “bad”. They are just models. Modeling in general is to pretend that one deals with a real thing while really working with an imitation. In operations research the imitation is a computer model of the simulated reality. A flight simulator on a PC is also a computer model of some aspects of the flight: it shows on the screen the controls and what the “pilot” (the youngster who operates it) is supposed to see from the “cockpit” (his armchair). A statistical model predicting who has the propensity to buy a certain product, or a machine learning algorithm with the same objective are also models.

What are models really?

In order to understand what it means to model a phenomena or process, we must first understand the term “model” and understand its limitation. A model is a physical, mathematical, or otherwise logical “representation” of a system, entity, phenomenon, or process. “Representation” or “imitation” are key in understanding models.

A model can also be thought of as an abstraction of the real world, or an approximation of it. If you think about the problem of modeling a human being, or just the mind of a human, you can immediately see the limitations of modeling. We can use the term “system” to encompass systems, entities, phenomenon, or process.

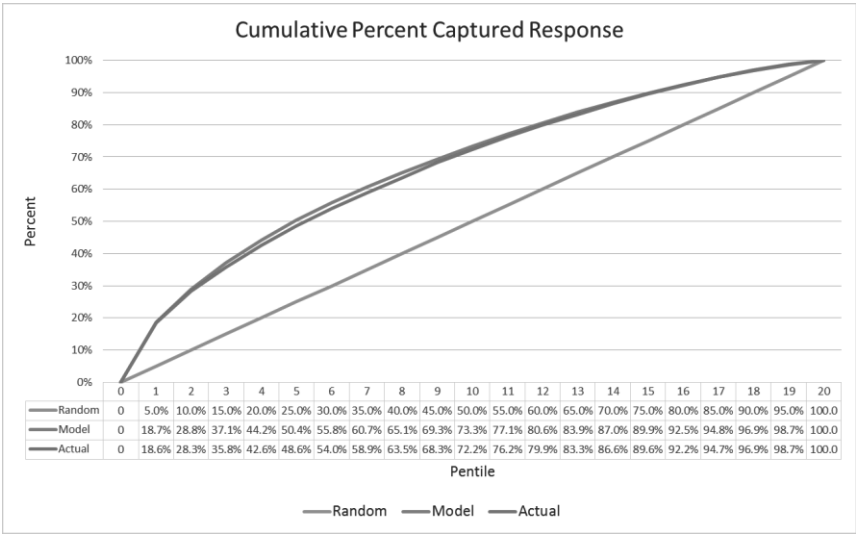
Since a model is a representation, abstraction, or approximation of the “system” being modeled, we must understand that it is not an “exact” representation, i.e., we can’t model every aspect of the system. First, we don’t know everything we need to know in order to model the system. We may not be able to define a process of the system with mathematical

precision, or with heuristic algorithms, and many of the processes may not appear logical. Second, the data that we use for a statistical model, for instance, may not be perfect (probably isn't), and the model is only as good as the data upon which it is built. Third, even if we were to know everything about the system, we may not have enough computing power to model every process, at least for complex systems, e.g., a human being, the earth's ecosystem, etc.; we may not have enough money to spend on it; and we may not have enough time to accomplish it.

Can models be "good enough"?

Even with their limitations, models are a good way to gain understanding of how a system operates. We cannot model every aspect or process of a real system (at least for most real systems). However, we can measure some aspects or processes. We cannot build a good model of the global economy. It is simply too complicated to reduce to an abstraction by simplifying assumption. But, we can build one for the American (US) economy, and another for the British economy and another for the Russian economy. Together, these separate model may provide insight to a global economy, but certainly not a very accurate one.

The question we often ask is, "How good is your model"? What we should ask is, "Is your model's performance better than no model at all?" The answer to the former is probably, "Not very good." The answer to the latter is "Probably better." The graph below represents the performance of a statistical model with predicted values, actual values and random values (or no model at all). Assuming the two curves above the diagonal line (representing no model) are statistically different from random chance, then the model would be better than having no model.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Random	0	5.0%	10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%	45.0%	50.0%	55.0%	60.0%	65.0%	70.0%	75.0%	80.0%	85.0%	90.0%	95.0%	100.0
Model	0	18.7%	28.8%	37.1%	44.2%	50.4%	55.8%	60.7%	65.1%	69.3%	73.3%	77.1%	80.6%	83.9%	87.0%	89.9%	92.5%	94.8%	96.9%	98.7%	100.0
Actual	0	18.6%	28.3%	35.8%	42.6%	48.6%	54.0%	58.9%	63.5%	68.3%	72.2%	76.2%	79.9%	83.3%	86.6%	89.6%	92.2%	94.7%	96.9%	98.7%	100.0

Did we build the “Right” Model?

Clearly a model in this instance is better than no model. But this requires qualification: is the model valid? This is different than "is the model a good one". It is not a matter of whether we built the model right, although that is equally important. No, the question is: “Did we build the right model?" If our statistical implementations were correct, then the model depicted in the figure would seem to be a good model. It captures about 70% of the audience at the 7th pentile. It is relatively smooth and is about 2.5 times better that no model at the 7th pentile. The question remains: is it the right model?

Would I use this model to predict? Yes, and I already do so. It is not a perfect model, but it is a useful one. The model addresses the client’s business case, and is used to predict the propensity of a customer to engage in action X (I cannot elaborate on the client or model any more than this). Yet, though it is used for prediction, it will still occasionally “miss the mark” and be wrong, due to inherent statistical errors and to data that may not be “perfect.” So the equation for a useful model might look like:

“good statistical procedures” + “relatively good data” – “an exact representation of reality” = “a model that is wrong but useful”

Is it still dangerous?

Thus, prediction is a dangerous game we play. We do not really know for certain what will occur in the months to come, but an educated guess—based on a model—is probably much better than a random shot in the dark. Our crystal ball may not work so well after all, but we can still *dimly* see what *might* come to pass.

Predictive Analytics: the skills and tools

Several weeks ago I posted an article called, What is Predictive Analytics, describing what it is. In the present article, I want to talk about the skills and tools that one should have to perform **predictive analytics**.

I am always at a loss in describing the skills of analytics, for there are many. I am working on another book about analytics that has a different approach than *Predictive Analytics using R*, though I am using material from three chapters. The new book is an operations research approach to analytics, covering a different set of methods, skill and tools. Combined, the two books are over 1000 pages, so perhaps you can see my dilemma. Hence, this article is going to touch the very basics.

What is Predictive Analytics?

In case you missed my previous article, this is a high level description. **Predictive analytics**—sometimes used synonymously with *predictive modeling*—is not synonymous with statistics, often requiring modification of functional forms and use of ad hoc procedures, making

it a part of data science to some degree. It does however, encompasses a variety of statistical techniques for modeling, incorporates machine learning, and utilizes data mining to analyze current and historical facts, making predictions about future. Beyond the statistical aspect lies a mathematical modeling and programming dimension, which includes linear optimization and simulation, for example. Yet analytics goes even farther by defining the business case and requirements, which are not covered here. I discussed those in How to Build a Model.

Statistical Modeling & Tools

This assumes that you already know the basics of parametric and a little bit of nonparametric statistics. If you are not familiar with these terms, then you are missing a prerequisite. However, this is gap you can fill with online courses from Coursera. Though I have never taken one, I have many colleagues who swear by them.

By statistical modeling I am referring to subject matter that would be covered beyond material in a statistics for engineering or business course(s). Here we are concerned with linear regression, logistic regression, analysis of variance (ANOVA), multivariate regression and clustering analysis, as well as goodness of fit testing, hypotheses testing, experimental design and my friends Kolmogorov and Smirnoff. Mathematical Statistics could be a plus, as it will take you into the underlying theory.

The tools one would/could use are a myriad and are often the tools our company or customer has already deployed. SAS modeling products are well-established tools of the trade. These include *SAS Statistics*, *SAS Enterprise Guide*, *SAS Enterprise Modeler*, and others. IBM made its mark on the market with the purchase of *Clementine* and its repackaging as *IBM SPSS Modeler*. There are other commercial products like *Tableau*. I have to mention *Excel* here, for it is all many will have to work with. But you have to go beyond the basics and into its data tools, statistical analysis tools and perhaps its linear programming *Solver*, plus be able to construct pivot tables, and so on.

Today, there a multitude of open source domain tools that have become popular, including *R* and its GUI, *R-Studio*; the *S* programming package;

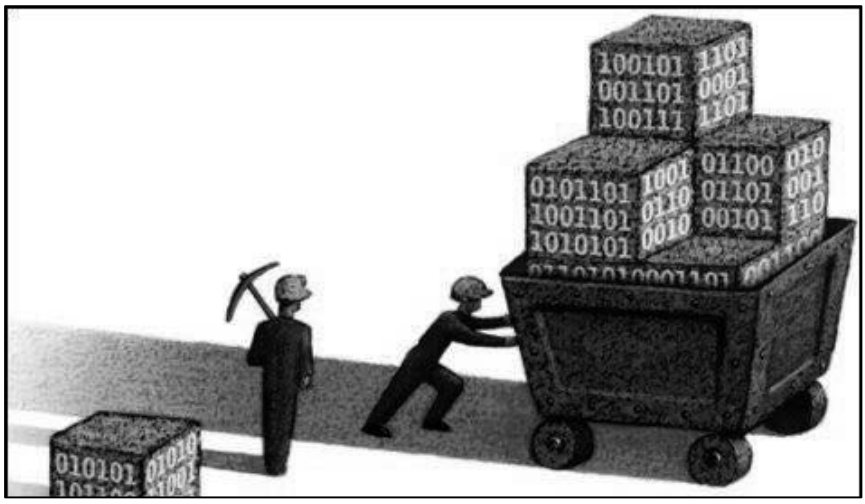
and the *Python* programming language (the most used language in 2014). *R*, for example, is every bit as good as its nemesis *SAS*, but I have yet to get it to leverage the enormous amount of data that I have with *SAS*. Part of this is due to server capacity and allocation, so I really don't know how much data *R* can handle.

Data Processing

For the forgoing methods, data is necessary and it will probably not be handed to you on a silver platter ready for consumption. It may be "dirty", in the wrong format, incomplete, or just not right. Since this is where you may spend an abundant amount of time, you need the skill at tools to process data. Even if this is a secondary task--it has not been for me--you will probably need to know Structured Query Language (SQL) and something bout the structure of databases.

If you do not have clean, complete, and reliable data to model with, you are doomed. You may have to remove inconsistencies, impute missing values, and so on. Then you have to analyze the data, perform data reduction, and integrate the data so that it is ready for use. Modeling with "bad" data results in a "bad" model!

Databases are plentiful and come in the form of *Oracle Exadata*, *Teradata*, *Microsoft SQL Server Parallel Data Warehouse*, *IBM Netezza*, and *Vertica*. The *Greenplum Database* builds on the foundations of open source database *PostgreSQL*. Or you may need to use a data platform like *Hadoop*. Also, *Excel* has the capacity to store "small amounts" of data across multiple worksheets and built in data processing tools.



What the Heck is Data Mining?

Data Mining

Generally, data mining (sometimes called knowledge discovery or data discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. It is sometimes used synonymously with data analysis or exploratory data analysis, but it is different from these, as we will see later. The information obtained from data mining can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Data Miners

Data Miners apply analytic processes designed to explore data (usually large amounts of data – typically business or market related – also known as “big data”) in search of consistent patterns or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. One goal of data mining is prediction – and predictive data mining is the most common type of data mining and one that has the most direct business

applications. However, the methods used for predictions is slightly different from the methods I use in predictive analytics, as we will see later.

The Data Mining Process

The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

Exploration

The first stage usually starts with **data preparation** which may involve **cleaning data**, **data transformations**, selecting **subsets of records** and – in case of data sets with large numbers of variables ("fields") – performing some preliminary feature selection operations to bring the number of variables to a manageable range (depending on the statistical methods which are being considered). Depending on the nature of the analytic problem, this first stage may involve processes anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods (see Exploratory Data Analysis (EDA)) in order to identify the most relevant variables. This also determines the complexity and/or the general nature of models that can be taken into account in the next stage.

Model Building and Validation

The second stage involves considering various models and choosing the best one based on their **predictive performance** (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal – many of which are based on so-called "**competitive evaluation of models**," that is, applying different models to the same data set and then comparing their performance to choose the best (I often implement these using R). These techniques – which are often considered the core of predictive data mining – include: Bagging (Voting,

Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

Deployment

The final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

Data Mining and Statistics

The concept of Data Mining is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business Data Mining (e.g., Classification Trees), but Data Mining is still **based on the conceptual principles of statistics** including the traditional **Exploratory Data Analysis** (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques.

Data Mining and Exploratory Data Analysis

There is an important general difference in the focus and purpose between Data Mining and the traditional Exploratory Data Analysis (EDA). Data Mining is more oriented towards **applications** than the basic nature of the underlying phenomena. In other words, Data Mining is relatively less concerned with identifying the specific relations between the involved variables. For example, uncovering the nature of the underlying functions or the specific types of interactive, multivariate dependencies between variables are not the main goal of Data Mining. Instead, the focus is on producing a solution that can generate useful predictions. Hence, Data Mining accepts a "**black box**" approach to data exploration or knowledge discovery. It uses not only the traditional Exploratory Data Analysis (EDA) techniques, but also such techniques as Neural Networks, which can generate valid predictions but are not capable of identifying the specific nature of the interrelations between the variables on which the predictions are based.

Data Mining and Machine Learning

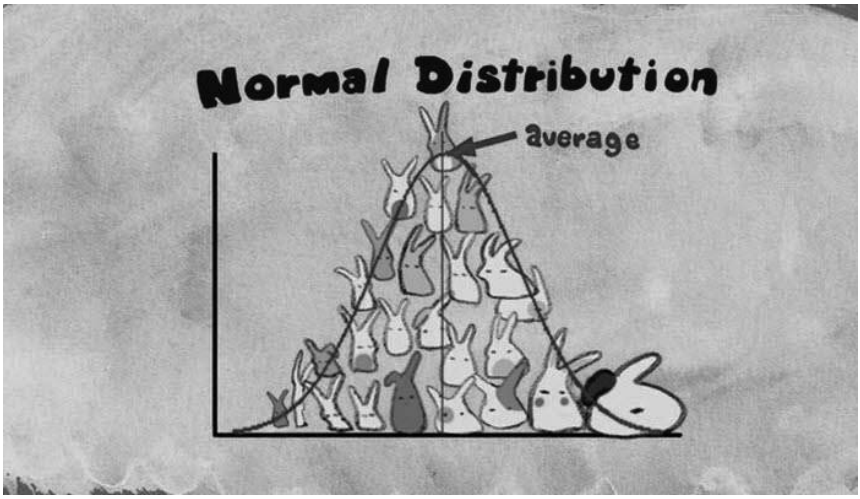
Data Mining is sometimes considered to be *"a blend of statistics, AI (artificial intelligence), and data base research"* (Pregibon, 1997, p. 8), which until very recently was not commonly recognized as a field of interest for statisticians, and was even considered by some *"a dirty word in Statistics"* (Pregibon, 1997, p. 8). Due to its applied importance, however, the field emerges as a rapidly growing and major area (also in statistics) where important theoretical advances are being made (see, for example, the recent annual International Conferences on Knowledge Discovery and Data Mining, co-hosted by the American Statistical Association).

Data Mining Techniques

For more information on Data Mining techniques, there are numerous books that review the theory and practice of data mining; the following books offer a representative sample of recent general books on data mining, representing a variety of approaches and perspectives:

- Berry, M., J., A., & Linoff, G., S., (2000). Mastering data mining. New York: Wiley.
- Edelstein, H., A. (1999). Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery & data mining. Cambridge, MA: MIT Press.
- Han, J., Kamber, M. (2000). Data mining: Concepts and Techniques. New York: Morgan-Kaufman.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning : Data mining, inference, and prediction. New York: Springer.
- Pregibon, D. (1997). Data Mining. Statistical Computing and Graphics, 7, 8.
- Weiss, S. M., & Indurkhya, N. (1997). Predictive data mining: A practical guide. New York: Morgan-Kaufman.
- Westphal, C., Blaxton, T. (1998). Data mining solutions. New York: Wiley.

- Witten, I. H., & Frank, E. (2000). Data mining. New York: Morgan-Kaufmann.



Commonly Misunderstood Analytics Terms

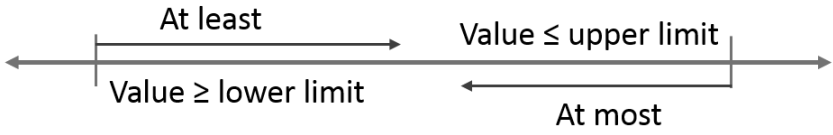
Have you ever sat in a briefing with an analyst as they describe the results of their analytics? You probably heard some of the terms described below. You may have had a statistics class in your MBA course work ten years ago, and you vaguely remember hearing the same terms there. If you are like me, you probably can spell them correctly 60% of the time, but their actual meaning escapes you. So, let's look at some commonly misunderstood terms.

"At least"

My favorite one is "at least". It is a **lower bound**. The smallest in size, amount, degree, etc. "The customer made at least one transaction" (they had not less than 1).

"At most"

"At most" is an **upper bound**. The greatest in size, amount, degree, etc. Not more than. "The customer had at most 12 internet sessions" (they had no more than 12).

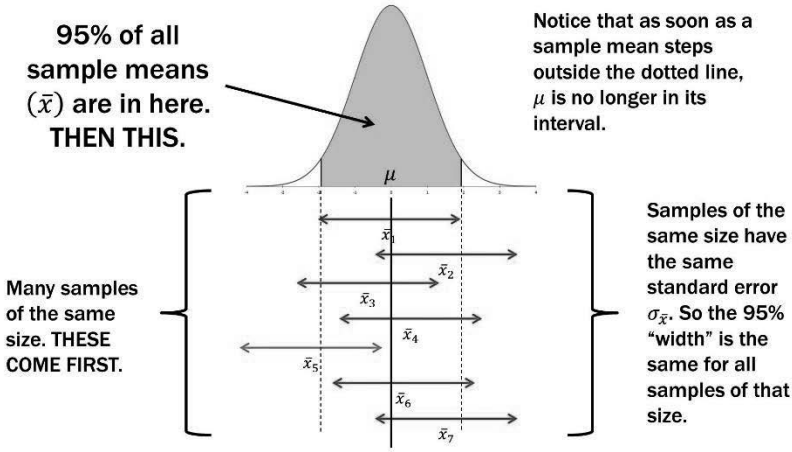


"Sample"

First, **sampling** is not a bad thing. I recently heard a corporate executive say that when you have all of the data you do not need to sample. Of course, “all of the data” is relative. If I have 15 million members in my credit union and I have complete data on every member, then I have “all the data”. But when I build a model, I will take a random sample comprised of 5% of the members (population), for example. **Random samples** are representative of the **population.** When I test the performance of my model, I will do it using another random sample from the membership (population).

"95% Confidence"

It does **not** mean 95 percent accurate! Consider a poll or survey based on a sample of the population. What a **95 percent confidence level** is saying is if the poll or survey were repeated over and over again, the results would match the results from the actual population 95 percent of the time. In the picture below $\bar{x}$ represents the arithmetic mean or average value. The subscripts of $\bar{x}$ (1, 2,..., 7) indicate mean values of seven different samples from the same population.



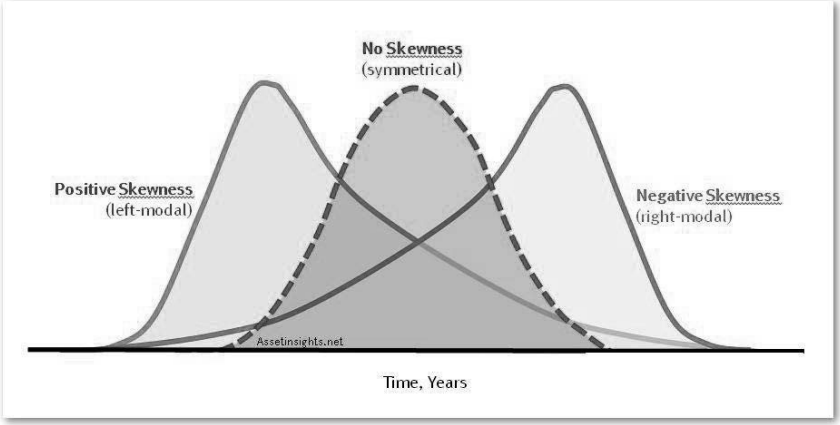
"Variance"

Variance is **not** a measure of how spread out the numbers are, as Wikipedia describes it. That would be the **standard deviation**. Variance is a measure of **dispersion**, but not in the sense of the Wikipedia definition. You can say that an observation is one standard deviation from the mean, but you cannot say that an observation is one variance from the mean. Variance is a mathematical calculation of the average of the squared differences from the Mean. We use it primarily to derive the standard deviation, the square root of variance. The equation below is the variance.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

"Skewness"

Skewness is a measure of the lack of symmetry **(asymmetry)** of a probability distribution. The direction of Skewness is usually the part that is misunderstood, for it is slightly counterintuitive. When we speak of asymmetry we say **positive Skewness** or **negative Skewness**. In this case a picture is worth a thousand words.



So, Skewness is describing the **long tail** of the distribution. A distribution with a long tail toward the right has positive Skewness. Another way to describe it is left-model, which makes it all the more counterintuitive. To make matter worse, we can describe the positive Skewness as the mean is smaller than the median and mode of the curve. To keep things straight, just focus on the direction of the longest tail.

"Central Limit Theorem"

I know this one keeps you awake at night. **The Central Limit Theorem** (**CLT)** states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

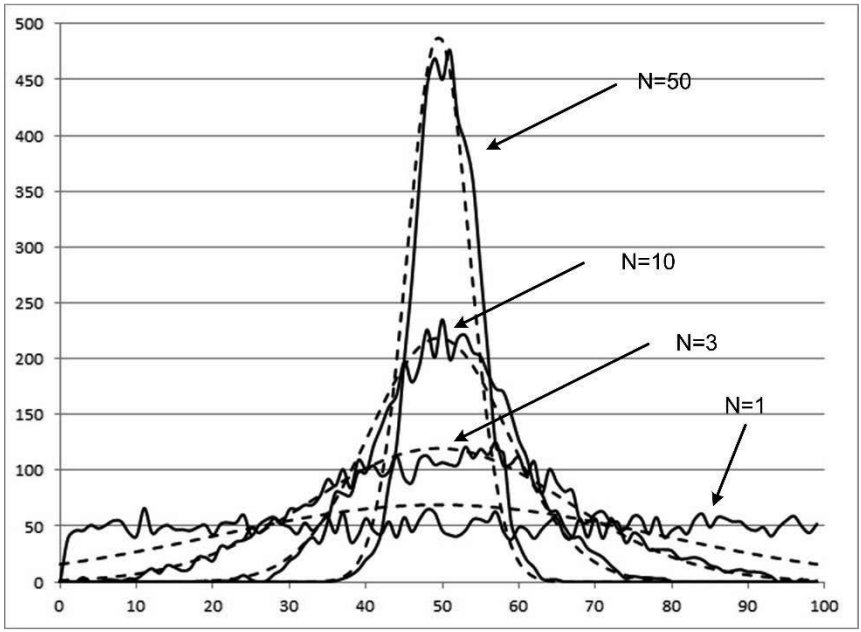
In English, the CLT tells us that if we take the **means** of the **N samples** from a population and plot the frequencies of their mean, we get a distribution that is **approximately normal**, regardless of the underlying distribution. The condition on N, the number of samples, is that it be **sufficiently large**. So how large is “large”? There are two conditions.

- Requirements for accuracy. The more closely the sampling distribution needs to resemble a normal distribution, the more sample points will be required.

- The shape of the underlying population. The more closely the original population resembles a normal distribution, the fewer sample points will be required.

In everyday use, some researchers say that a sample size of 30 is large enough when the population distribution is roughly bell-shaped. Others recommend a sample size of at least 40. But if the original population is distinctly not normal (e.g., is badly skewed, has multiple peaks, and/or has outliers), researchers like the sample size to be even larger.

One final point: the CLT does not say that the distribution of the means form a normal distribution. It say they are "approximately" normally distributed. Below is a picture of the concept. The solid lines represent the shape of the distribution of the means of samples from a population that is not normal, and the dashed lines represent a normal distribution.



Conclusion

Now you can sleep better at night, and when your analyst tells you that your confidence level in 95% due to the Central Limit Theorem, using at

least 40 samples from a positively skewed population, you will know what she is talking about, within a few standard deviations.

Part III – Models and Models and More Models

How do you build a model?

Customers and modelers, this is for you. Years ago I learned the seven steps to model building as a budding Operations Research Analyst. As I grew from sapling to tree, I realized that something was missing, but I did not really know what link was absent until I plowed into analytics. Now I am prepared to tell you, from an Operations Research (OR) perspective what I believe is a more complete set of steps. [Thanks to Yifan Gu for suggesting the following quote.]

“

If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions

—Albert Einstein

Define the problem

Defining the problem is necessary. Defining the “right” problem is absolutely critical, otherwise you're just wasting everyone's time. However, the customers you are working for may not know how to express the real problem or may not know what the problem really is. The Operations Research (OR) Analyst must ask the right questions and draw the right problem out form where it may be hiding.

Define the Business case

Once you identify the real problem, you have to help answer the question, “Is the problem worth solving (or at least worth the OR analyst’s involvement)?”. That may sound odd, but if the problem statement is not translated in to a business case, then the business does not need the problem solved badly enough to warrant the time and effort to solve it. We have to ask, “What is it worth to the business?” Increased profit? Improved production? Savings of marketing dollars?...

Define the Model Objective

It takes a well-defined problem and solid business case to ensure that we build the right model or provide the right solution. In my earlier years, I often found myself halfway through a project before realizing I was not building the right model. A model objective that answers the call of the business case is essential. It will keep you on course.

Determine the requirements

We have a well-defined problem, a business case with a supporting model objective, so now let’s go and build the model. Slow down! What are the requirements? Do you have the tools you need, the team you need, and the budget to support it? If data is to be used, do you have access to it, has it been processed, and is it in the right format? We have to determine the requirements or we may fail due to some oversight.

Gather the Data

If data is necessary, it will probably not be handed to you on a silver platter ready for consumption. It may be "dirty", in the wrong format, incomplete, of just not right. If you have fully identified the requirements, you will already know what has to be accomplished at this stage.

Process the Data

This is where you may spend an abundant amount of time. If you do not have clean, complete, and reliable data, you are doomed. You may have to remove inconsistencies, impute missing values, and so on. Then you have to analyze the data, perform data reduction, and integrate the data so that it is ready for use. Modeling with "bad" data results in a "bad" model!

Build the Model

So, you have done the hard part and it's time to have fun. Building the model is an iterative process. I doubt that anyone ever uses the first model they build. If they do, they probably should not. There is as much art as there is science in model development. Modeler judgment is as important as *p*-values, and Chi-square tests. If the model does not make sense to you—is not intuitive—then it is probably not the final model.

Interpret the Model Results

Here is where we often have the most difficulty as analysts. We understand the mathematical and statistical interpretation of results, but we have to go beyond that and translate it to the customer's language. They are not going to get the significance of the odds ratio of variable *X*. If you start thinking about that here, it will make life less painful later on.

[Thanks to David Nixon for suggesting this addition.] Whatever the model suggests probably takes a "leap of faith" within the organization. After all, what the model may have identified is a bunch of correlations

that fly in the face of conventional wisdom. The validation step is an extremely important piece of the puzzle whereby a company may have to perform a series of controlled experiments to confirm for sure that "x contributes to the explanation of y". This leap of faith extends to potentially creating the necessary data in the first place if it doesn't already exist and may involve short-term losses for the promise of a much larger long-term gain.

Validate the Model for Production

This is purely a scientific task. The model must be documented every step of the way and then stand up to the scrutiny of stress testing, peer review, and so on. This is where you find out if you really did build the right model, and you built it right.

Perform an Economic Analysis

Here is where you answer the call of the business case. You must translate the results to dollars, the economic benefit of the solution, or another metric if savings or profit are not objectives. For instance, the military OR may be saving lives instead of dollars.

Present the Results

If you bothered to interpret the results and translate them to "customer speak", this part is easy. But, you are not just presenting results, you are selling the solution. If your work and your recommendations are not implemented, you just wasted a lot of time, effort and money.

Follow-up

Too often we fix the problem and go back to our lives or onto another problem, but following up with the customer to evaluate the effectiveness of the solution you provided is just as important as the solution that you provided. It is especially important if you want to ever see that customer again.

Conclusion

So there is my view of model building. I have built just about every kind of model: linear programs, nonlinear programs, machine learning models, statistical models, Markov transition models, and so on. I have built models of unmanned aerial vehicles, space launch vehicles, communications networks, software systems, propensity to buy a product, call center operations, combat phenomena, and so on. And I have, more than I would like to admit, built the wrong model. Hopefully, this will help the modeler step out on the right foot, and help the model recipient know what to look for when asking for a modeling solution.

What are Predictive Models Anyway?

Predictive modeling does not lie solely in the domain of Big Data Analytics or Data Science. I am sure that there are a few “data scientist” who think they invented predictive modeling. However, predictive modeling has existed for a while and at least since World War II. In simple terms, a predictive model is a model with some predictive power. I will elaborate on this later.

I have been building predictive models since 1990. Doing the math, 2015 – 1990 = 25 years, I have been engaged in the predictive modeling business longer that data science has been around. My first book on the subject, "Fundamentals of Combat Modeling (2007), predates the "Data Science" of 2009 (see below).

How old is Data Science?

It is really a trick question. The term was first used in 1997 by C. F. Jeff Wu. In his inaugural lecture for the H. C. Carver Chair in Statistics at the University of Michigan, Professor Wu (currently at the Georgia Institute of Technology), calls for statistics to be renamed data science and statisticians to be renamed data scientists. That idea did not land on solid

ground, but the topic reemerges in 2001 when William S. Cleveland publishes “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.” But it is really not until 2009 that data science gains any significant following and that is also the year that Troy Sadkowsky created the data scientists group on LinkedIn as a companion to his website, datasceintists.com (which later became datascientists.net). [1]

What is Predictive Modeling?

It is not a field of statistics! Yes, we do predictive modeling in statistics, but it is really a multidisciplinary field and is based more in mathematics than in other fields. Now, if you consult the most authoritative source of factual information available to the world, *Wikipedia*, you will find an incorrect view of predictive modeling (of course, I do not believe what I said about Wikipedia). It was formed by people with too much time on their hands and too little exposure to other disciplines, such as physics and mathematics.

Predictive modeling may have begun as early as World War II in the Planning of Operation Overlord, the Normandy Invasion, but was certainly used in determining air defenses and bombing raid sizes (it may have appeared as early as 1840 [2]). Now, this is not an article about the history of operations research, so suffice it to say that the modern field of operational research arose during World War II. In the World War II era, operational research was defined as “a scientific method of providing executive departments with a quantitative basis for decisions regarding the operations under their control.”[3]

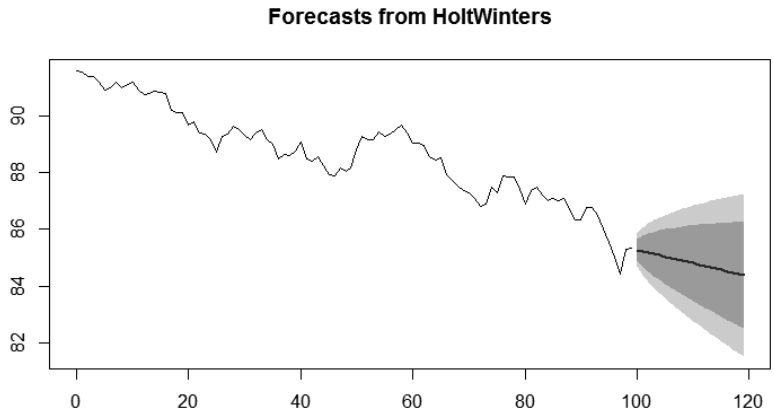
What is a Predictive Model?

The answer is easy: a model with some predictive power. I say that with caution, and use the word “some”, because more often than not, decision makers think that these model are absolute. Of course, they become very disappointed when the predictions do not occur as predicted. Rather than expand on my simplistic definition, I think some examples my help.

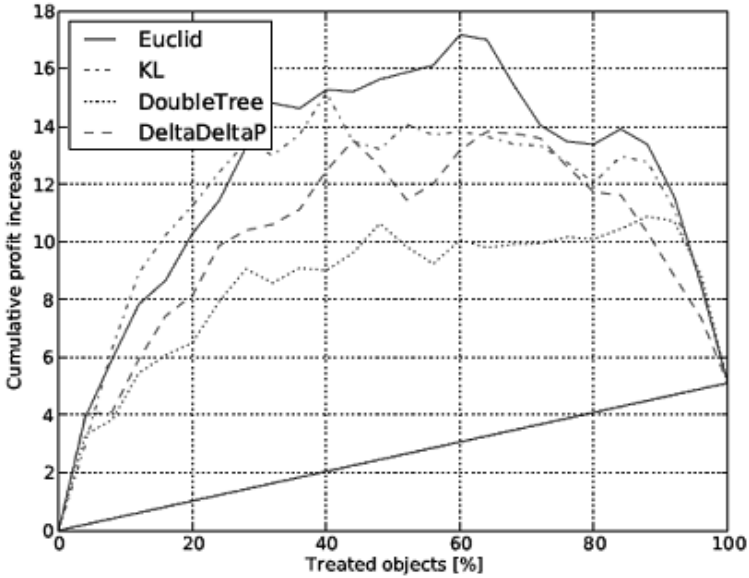
Examples of Predictive Models

The taxonomy of predictive models represented here is neither exhaustive of exclusive. In other words, there are other ways to classify predictive models, but here is one.

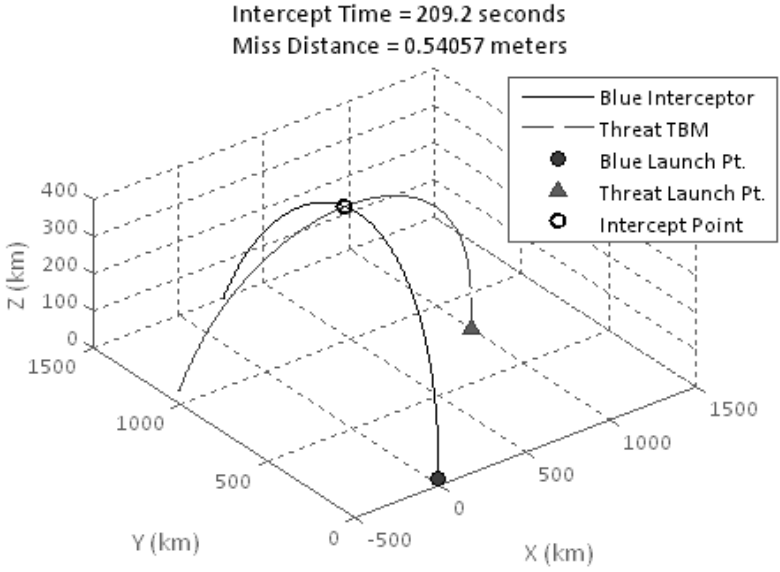
Times Series Models/Forecasting Models. This kind of model is a statistical model based on time series data. It uses "smoothing" techniques to account for things like seasonality in predicting or forecasting what may happen in the near future. These models are based on time-series data.



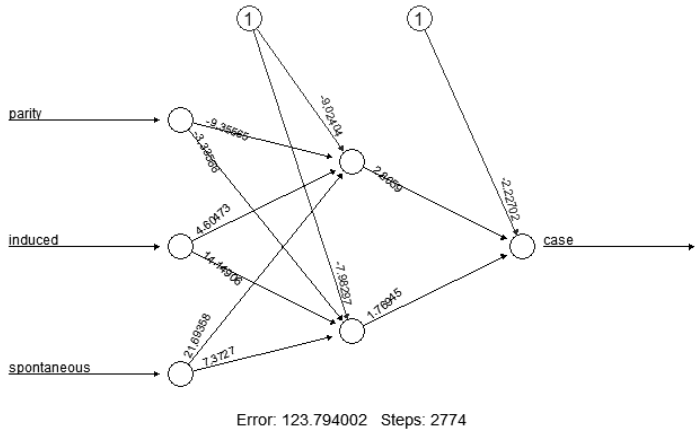
Regression Models. Time series model are technically regression models, but machine learning algorithms like auto neural networks have been employed recently in Time Series Analysis. Here I am referring to logistic regression models used in propensity modeling, and other regression models like linear regression models, robust regression models, etc. These models are based on data.



Physical models. These models are based on physical phenomena. They include 6-DoF (Degrees of Freedom) flight models, space flight models, missile models, combat attrition models (based on physical properties of munitions and equipment).

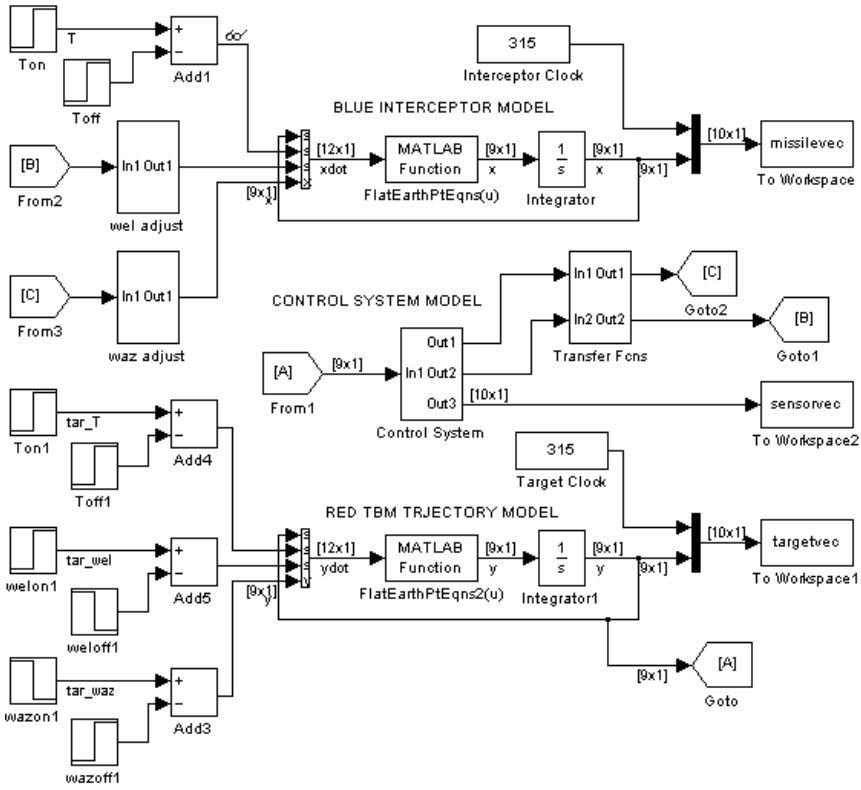


Machine Leaning Models. These include auto neural networks (ANN), support vector machines, classification trees, random forests, etc. These are based on data, but unlike statistical models, they "learn" from the data.



Weather models. These are forecasting models based on data, but the amount of data, the short interval of prediction windows and the physical phenomena involved make them much different that statistical forecasting models.

Mathematical Models. These are usually restricted to continuous time models based on differential equations or estimated using difference equations. They are often used to model very precise processes like the dynamics solid fuel rockets, or to approximate physical phenomena in the absence of actual data, like attrition coefficients approximation or direct fire effects in combat models.



Statistical Models. The first two examples, Time Series and Regression models, are statistical models. However, I list it separately because many do not realize that statistical models are mathematical models, based on mathematical statistics. Things like means and standard deviations are statistical moments, derived from mathematical moment generating functions. Every statistic in Statistics is based on a mathematical function.

$$\widetilde{w}_i = \frac{\partial \log\left(\frac{o(x)}{1-o(x)}\right)}{\partial x_i}$$

What Predictive Models have I Built?

I have built predictive models in all example categories except weather models. Models I have built include Reliability, Availability and

Maintainability (RAM) models for Unmanned Aerial Vehicle design; unspecified models involving satellites (unspecified because they are classified); unspecified missile models; combat attrition models; 6-DoF missiles models; missile defense models; propensity to purchase, propensity to engage, and share or wallet models regression models; time-series forecasting models for logistics; uplift (net-lift models) marketing models; ANN models as part of ensembles, classification trees, and random forests marketing models. I have also worked on descriptive and prescriptive models.

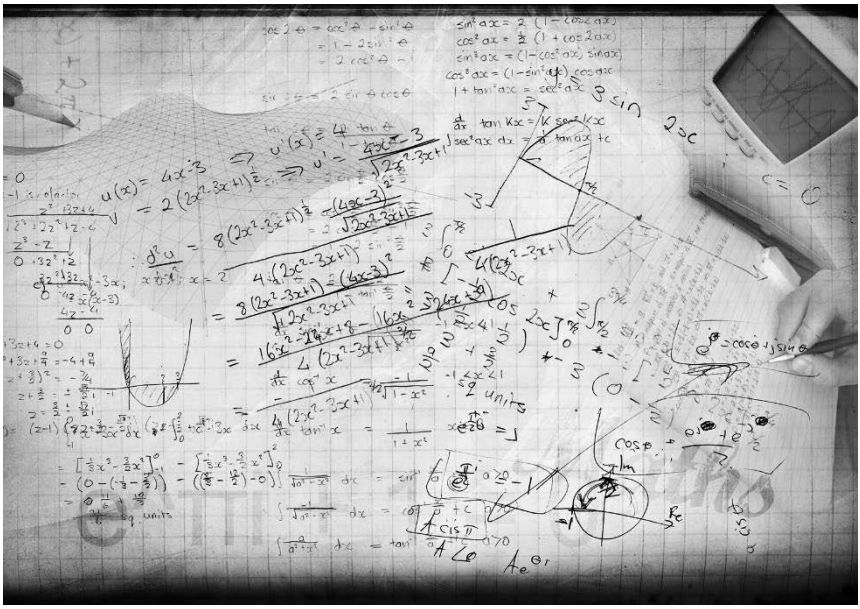
Models I have consulted on include the NASA Ares I Crew Launch Vehicle Reliability and Launch Availability; The Extended Range Multi-Purpose (ERMP) Unmanned Aerial Vehicle RAM Model, The Future Combat Systems (FCS) C4ISR family of models; FCS Logistic Decision Support System Test-Bed Model; Unspecified models (unspecified because they are classified).

References

Press, G. “A Very Short History Of Data Science”, Forbes, May 28, 2013 @ 7:09 AM, Retrieved 05-29-2015.

P. W. Bridgman, The Logic of Modern Physics, The MacMillan Company, New York, 1927.

Operational Research in the British Army 1939–1945, October 1947, Report C67/3/4/48, UK National Archives file WO291/1301. Quoted on the dust-jacket of: Morse, Philip M, and Kimball, George E, Methods of Operations Research, 1st Edition Revised, pub MIT Press & J Wiley, 5th printing, 1954



Mathematical Modeling

Again, there are prerequisites like differential and integral calculus and linear algebra. Multivariate calculus is a plus, particularly if you'll be doing models involving differential equations and nonlinear optimization. The skills you need to acquire beyond the basics include mathematical programming--linear, integer, mixed, and nonlinear. Goal programming, game theory, Markov chains, and queuing theory, to name a few, may be required. Mathematical studies in real and complex analysis, and linear vector spaces, as well as abstract algebraic concepts like group, fields and rings, can reveal the foundational theory.

Simulation modeling, including Monte Carlo, discrete and continuous time, plus discrete event simulation can be applied in analytics--I have not seen this as common practice in business analytics, but it certainly has its place. These models may rely heavily upon queuing theory, Markov chains, inventory theory and network theory.

The corporate mainstay is the powerhouse combination of *MATLAB* and *Simulink*. *MATrix LABoratory* or *MATLAB* (that is why it is spelled with all caps!). Other noteworthy commercial products include *Mathematica*

and *Analytica*. *Otave* is an open-source mathematical modeling tool that reads MATLAB code and there are add-on GUI environments (like *R-studio* for *R*) floating around in hyperspace. I recently discovered the power of *Scilab* and the world of modules (packages) that are available for this open-source gem.

For simulation, Simulink works "on top of" MATLAB functions/code for a variety of simulation models. I wrote the book "Missile Flight Simulation", using MATLAB and Simulink. ExtendSim is an excellent tool for discrete event simulation and the subject of my book "Discrete Event Simulation using ExtendSim". In Scilab, I have used Xcos for discrete event simulation and Quapro for linear programming. Both are featured in my next book.

There is a general analytics tool that I do not know much about yet. *BOARD,* in its newest release, boasts a predictive analytics capability. I will be speaking on predictive analytics at the *BOARD* User Conference during April 13th-14th in San Diego. Again, I would be remiss not to mention *Excel*, and particularly the *Solver* add-in for mathematical programming. Another 3rd-party add-in to consider to @Risk.

Conclusion

If you aspire to become an analytics consultant or scientist, you have a lot of open-source tools, free training and online tutorials at your fingertips. If you are already working in analytics, you can easily specialize in predictive analytics. If you are already working in predictive analytics, you have what you need to become an expert. All of the tools will either work with your PC's native processing power or through a virtual machine, for example, when using *Hadoop,* or remote server.

If You Build It They Will Buy...

Most everyone knows that using a propensity model to predict customers' likelihood of buying, engaging, leaving and so on, produces a larger target audience. That is too bad... for what "everyone knows" is actually incorrect.

A Field of Dreams

In the movie *File of Dreams*, actor Kevin Costner (whom people often confuse me with), plays an Iowa corn farmer struggling to pay the mortgage on his homestead. He is out farming one day and hears a voice, like a loud whisper in the wind:

If you build it they will come.

And in the end he does build it and they do come...

Models are not like that. In fact, they reduce the size of the customer population that has a high propensity to do whatever it is that you are modeling. If you heard a voice in this context it would say:

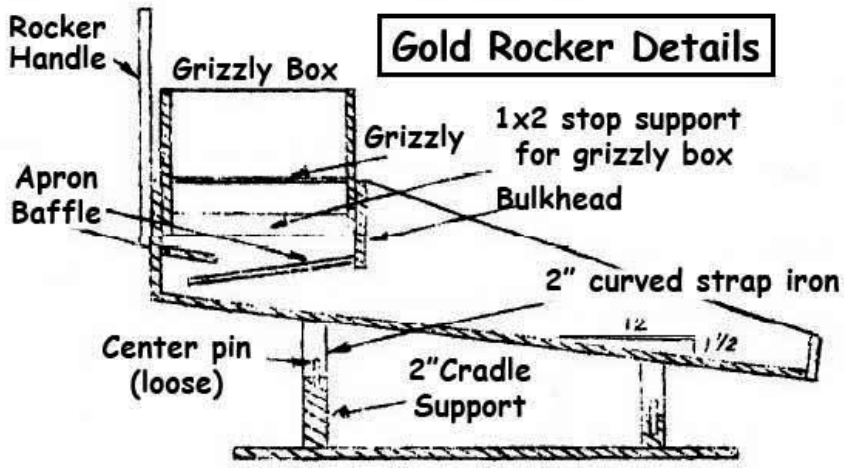
If you build it, fewer will come.

Wow, that is really bad! So why do we use models?

Actually, it is not bad—that is what we want models to do. But you must think I am outside my mind for saying so. However, modeling the propensity of customers to buy, engage, etc. is like prospecting for gold. There is a lot of rock and sand that has to be filtered out to get the "pure" gold dust.

Rocker Boxes

A rocker box (also known as a cradle) is a gold mining implement for separating alluvial placer gold from sand and gravel which was used in placer mining in the 19th century. It consists of a high-sided box, which is open on one end and on top, and was placed on rockers.

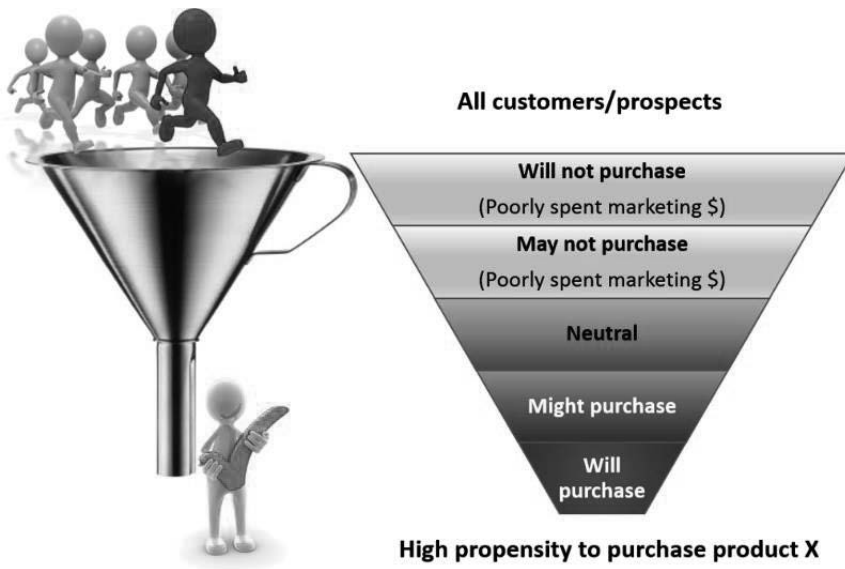


The inside bottom of the box is lined fles (usually glass of plastic) and a usually a carpet similar to a sluice. On top of the box is a classifier sieve or grizzly (usually with half-inch or quarter-inch openings) which screens-out larger pieces of rock and other material, allowing only finer sand and gravel through. Between the sieve and the lower sluice section is a baffle, which acts as another trap for fine gold and also ensures that the aggregate material being processed is evenly distributed before it enters the sluice section. It sits at an angle and points towards the closed back of the box. Traditionally, the baffle consisted of a flexible apron or "blanket riffle" made of canvas or a similar material, which had a sag of

about an inch and a half in the center, to act as a collection pocket for fine gold. Later rockers (including most modern ones) dispensed with the flexible apron and used a pair of solid wood or metal baffle boards. These are sometimes covered with carpet to trap fine gold. The entire device sits on rockers at a slight gradient, which allows it to be rocked side to side.

Propensity Models

A propensity model is like a rocker box: what comes out with a high propensity score is your gold dust customers—the ones that are rich with propensity to do whatever it is that you are modeling. The debris scores low in the model. If you have 10,000,000 customers, chances are that only 10,000 (or fewer) have a high propensity to buy product X. But those customer are your gold dust. (A friend, Steve Cartwright, just posted an article on microtargeting.) I also use the analogy of filtered funnel, where the top has a large opening and the bottom a much smaller one, and where my "gold dust" customers are small enough to pass through the tiny funnel opening at the bottom. Many go in, but few come out.



It is incumbent on us modelers to manage the expectations of our customers. Though quite a few already understand this model phenomenon, we must establish with the customer what to anticipate as an outcome at the binning of a model project, even before we think about data, model functional form and so on.

So, the voice that said, "If you build it they will come", was not talking to me (even though I strongly resemble Kevin Costner). Instead, my voice said, "If you build it fewer, but richer will come."

Uplift Models in Plain English, sort of...

I have used the term "Uplift" when speaking of class of predictive models, but now I want to provide a plain English description. Before I do that , however, do not get your hopes up. Uplift models are hard to construct and even harder to maintain--they are not a "cure all" solution to marketing problems.

Uplift or netlift refers to what marketing analysts call "incremental lift" or "incremental response". It is the lift (response) you get from touching people with a campaign message, that would not have responded without it--the "Persuadables"..

Uplift models require data from a controlled experiment, that is, a campaign design with a treated group and a control group. If you use direct mail as the media for your campaign, your treated group would receive the direct mail and you control group would not. Among the customers in the control group are those who will purchase what you are promoting without receiving the campaign message. These are called the "buy anyways" and are a very important group in uplift modeling (some of these are in the treated group as well). They are the customers that you do not need to spend marketing dollars on, and the difference in the responses from these individuals and the responses from the treated group represents the incremental response. So, the

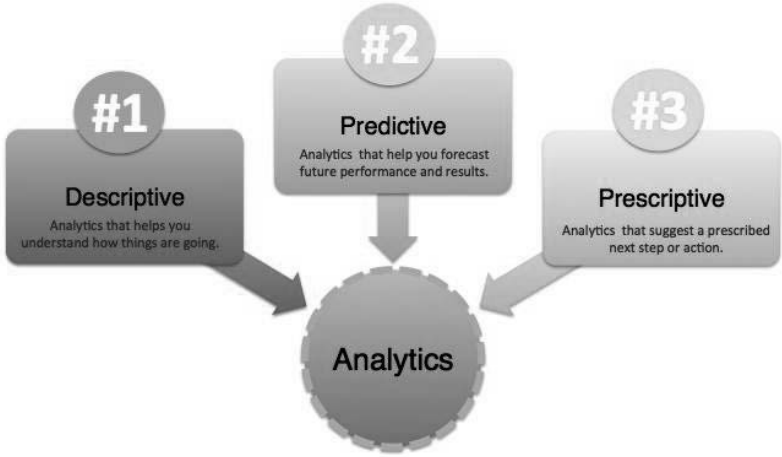
netlift is equal to responses from the treated group minus responses from the control group during the campaign acquisition window, roughly.

Now, what makes it hard to maintain an uplift model? The model is super-sensitive to any changes in the campaign upon which it was built. In other words, if you alter the campaign message too much--and "too much" is relative--then the uplift model may be a poor predictor. Thus the campaign message needs to remain fairly consistent over time. If you cannot do this, then an uplift model is probably not your solution.

Now for a point about data collection. You do not necessarily have to track who is responding to the campaign. You only need to see who purchases the product you are marketing during the acquisition window. Then for both the treated and control group, if a purchase is made during the campaign acquisition window it counts as a response. If your customer database is set up so that you track products on a regular basis--daily, weekly, monthly, and so on--then if the acquisition window is 60 days, you merely look at the product count before the 60-day window and the product count after the 60-day window, and measure the difference. Without going into the technical detail, the modeler or marketing analyst usually does this, and only one response is counted regardless of the number of products purchased. The non-technical reason for this is that we are measuring "response to a campaign," not "number of products purchased.

Finally, in order to validate an uplift model, you have to have multiple campaigns with that "consistent" campaign message I talked about. Again, without going into the technical detail, the model is used to score the other campaigns and the performance results are compared, using netlift (usually by deciles or 10 groups) as the metric.

If you are eager for the technical details, or if you are suffering from insomnia and wish to sleep, download my free e-Book entitled "Predictive Modeling and Analytics". Uplift modeling is found in Chapter 17. Of course you are always welcome to purchase a copy at Lulu.com or Amazon.com, thus contributing to my gas money.



What's the Difference?

I have been writing a little bit about analytics and models. Here is the differences between three types of models I have discussed that may need clarification.

Descriptive models

These models describe what has already occurred and in analytics they are usually statistical. They help the business know how thing are going. Clustering models might fall into this category.

Predictive models

Predictive models tell us what will probably happen in the future as a result of something that has already happened. They help the business forecast future behavior and results. These may be comprised of statistical models, like regression, or machine learning models, like neural networks. "Propensity to buy" models fall into this category.

Prescriptive models

These models are like descriptive-predictive models. They go an extra step and tell us why and help the business prescribe a next course of action. That is, they not only tell us what probably will happen, but also why it may happen. Form here we can look at what's next, for if we know the why, we can better know how to mold things for the future.

Analytics

According to The Institute for Operations Research and the Management Science (INFORMS) Analytics is defined as **the scientific process of transforming data into insight for making better decisions.** To perform good analytics, companies should employ all three types of modeling in order to cover the entire spectrum of their business.

Terms

I use words like “probably” and “may” because there is always a bit of uncurtaining in predictive and prescriptive models. You usually hear about this in phrases like, “95 percent confidence”, or perhaps in terms like, “Type II error.”

What is a Propensity Model?

Predictive analytics was a topic of one of my recent posts, and I have been often asked how specifically marketers can use predictions to develop more profitable relations with their customers. Yesterday, I talked specifically about predictive modeling. There are three types of predictive models marketers should know about, but I will only talk about the first one in this article:

- Propensity models (predictions)
- Clustering models (segments)
- Collaborative filtering (recommendations)

Propensity models are what most people think of when they hear "predictive analytics". Propensity models make predictions about a customer's future behavior. With propensity models you can anticipate a customers' future behavior. However, keep in mind that even propensity models are abstractions and do not necessarily predict absolute true behavior (see, "What is Predictive Modeling?"). I'll go through six examples of propensity models to explain the concept.

Model 1: Predicted customer lifetime value

CLV (Customer Lifetime Value) is a prediction of all the value a business will derive from their entire relationship with a customer. The Pareto Principle states that, for many events, roughly 80% of the effects come from 20% of the causes. When applied to e-commerce, this means that 80% of your revenue can be attributed to 20% of your customers. While the exact percentages may not be 80/20, it is still the case that some customers are worth a whole lot more than others, and identifying your "All-Star" customers can be extremely valuable to your business. Algorithms can predict how much a customer will spend with you long before customers themselves realizes this.

At the moment a customer makes their first purchase you may know a lot more than just their initial transaction record: you may have email and web engagement data for example, as well as demographic and geographic information. By comparing a customer to many others who came before them, you can predict with a high degree of accuracy their future lifetime value. This information is extremely valuable as it allows you to make value based marketing decisions. For example, it makes sense to invest more in those acquisition channels and campaigns that produce customers with the highest predicted lifetime value.

Model 2: Predicted share of wallet

Predicted share of wallet refers to the amount of the customer's total spending that a business captures in the products and services that it offers. Increasing the share of a customer's wallet a company receives is often a cheaper way of boosting revenue than increasing market share. For example if a customer spends $100 with you on groceries, is this 10% or 90% of their grocery spending for a given year? Knowing this allows you to see where future revenue potential is within your existing customer base and to design campaigns to capture this revenue.

Model 3: Propensity to engage

A propensity to engage model predicts the likelihood that a person will engage in some activity, like unethical behavior or post purchases. For

example, a propensity to engage model can predict how likely it is that a customer will click on your email links. Armed with this information you can decide not to send an email to a certain “low likelihood to click” segment.

Model 4: Propensity to unsubscribe

A propensity to unsubscribe model tells you which customers not to touch: if there are high value customers you are at risk of losing to unsubscribe, you need to find other ways to reaching out to them that are not by email. For example, you can predict how likely it is that a customer will unsubscribe from your email list at any given point in time. Armed with this information you can optimize email frequency. For “high likelihood to unsubscribe” segments, you should decrease send frequency; whereas for “low likelihood to unsubscribe” segments, you can increase email send frequency. You could also decide to use different channels (like direct mail or LinkedIn) to reach out to “high likelihood to unsubscribe” customers.

Model 5: Propensity to buy

The propensity to buy model tells you which customers are ready to make their purchase, so you can find who to target. Moreover, once you know who is ready and who is not helps you provide the right aggression in your offer. Those that are likely to buy won’t need high discounts (You can stop cannibalizing your margin) while customers who are not likely to buy may need a more aggressive offer, thereby bringing you incremental revenue.

For example, a “propensity to buy a new vehicle” model built with only data the automotive manufacturer has in their database can be used to predict percent of sales. By incorporating demographic and lifestyle data from third parties, the accuracy of that model can be improved. That is, if the first model predicts 50% sales in the top five deciles (there are ten deciles), then the latter could improve the result to 70% in the top five deciles.

Model 6: Propensity to churn

Companies often rely on customer service agents to "save" customers who call to say they are taking their business elsewhere. But by this time, it is often too late to save the relationship. The propensity to churn model tells you which active customers are at risk, so you know which high value, at risk customers to put on your watch list and reach out. Armed with this information, you may be able to save those customers with preemptive marketing programs designed to retain them.

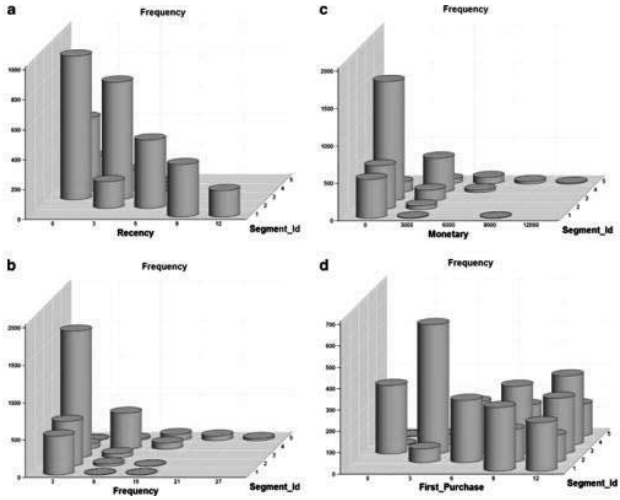
Often propensity models can be combined to make campaign decisions. For example, you may want to do an aggressive customer win back campaign for customers who have both a high likelihood to unsubscribe and a high predicted lifetime value.

Conclusion

Predictive analytics models are great, but they are ultimately useless unless you can actually tie them to your day-to-day marketing campaigns. This leads me to the first rule of predictive analytics:

Always make sure that your predictive analytics platform is directly integrated with your marketing execution systems such as your email service provider, web site, call center or Point of Sell (POS) system.

It is better to start with just one model that you use in day-to-day marketing campaigns than to have 10 models without the data being actionable in the hands of marketers.



What Are Clustering models?

Clustering is the predictive analytics term for customer segmentation. Clustering, like classification, is used to segment the data. Unlike classification, clustering models segment data into groups that were not previously defined. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them.

With clustering you let the algorithms, rather than the marketers, create customer segments. Think of clustering as auto-segmentation. Algorithms are able to segment customers based on many more variables than a human being ever could. It's not unusual for two clusters to be different on 30 customer dimensions or more. In this article I will talk about three different types of predictive clustering models.

Predictive model 1: Behavioral clustering

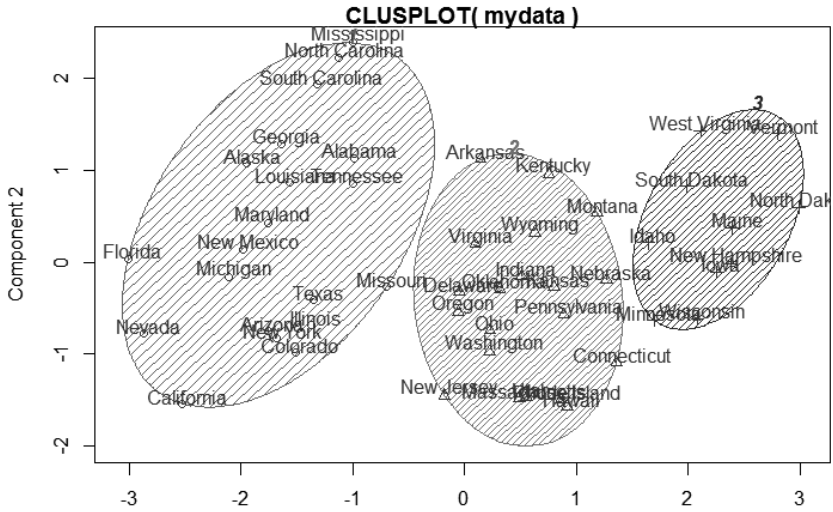
Behavioral clustering informs you how people behave while purchasing. Do they use the web site or the call center? Are they discount addicts? How frequently do they buy? How much do they spend? How much time will go buy before they purchase again? This algorithm helps set the right tone while contacting the customer. For instance, customers that buy

frequently but with low sized orders might react well to offers like 'Earn double rewards points when you spend $100 or more.

Behavioral clustering can also informs us on other behaviors, such as crime and is used in performing crime analysis. In the example below there are three crime clusters (only the top ten are shown in the table):

	Murder	Assault	UrbanPop	Rape	classif
Alabama	13.2	236	58	21.2	1
Alaska	10.0	263	48	44.5	1
Arizona	8.1	294	80	31.0	1
Arkansas	8.8	190	50	19.5	2
California	9.0	276	91	40.6	1
Colorado	7.9	204	78	38.7	1
Connecticut	3.3	110	77	11.1	2
Delaware	5.9	238	72	15.8	2
Florida	15.4	335	80	31.9	1
Georgia	17.4	211	60	25.8	1

Graphically, the clusters appear as follow:



Cluster plot created in R-Studio

Predictive model 2: Product based clustering (also called category based clustering)

Product based clustering algorithms discover what different groupings of products people buy from. See the example below of a category (or product) based segment or cluster. You can see people in one customer segment ONLY buy Pinot Noir, whereas those in another customer segment buy different types of Varietal products, such as Champagne, Chardonnay, Pinot Grigio and Prosecco – but never Cabernet Sauvignon, Malbec or Espumante. This is useful information when deciding which product offers or email content to send to each of these customer segments.

Offer #	Offer date	Product	Minimur	Discoun	Origin	Past Peak	1	2	3	4
1	January	Malbec	72	56	France	FALSE	0	0	4	6
2	January	Pinot Noir	72	17	France	FALSE	4	0	4	2
3	February	Espumante	144	32	Oregon	TRUE	0	0	2	4
4	February	Champagne	72	48	France	TRUE	0	0	7	5
5	February	Cabernet Sa	144	44	New Zealar	TRUE	0	0	2	2
6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
7	March	Prosecco	6	40	Australia	TRUE	0	12	4	3
8	March	Espumante	6	45	South Africa	FALSE	0	11	6	3
9	April	Chardonnay	144	57	Chile	FALSE	0	0	7	3
10	April	Prosecco	72	52	California	FALSE	0	0	5	2
11	May	Champagne	72	85	France	FALSE	0	0	7	6
12	May	Prosecco	72	83	Australia	FALSE	0	0	3	2
13	May	Merlot	6	43	Chile	FALSE	0	6	0	0
14	June	Merlot	72	64	Chile	FALSE	0	0	5	4
15	June	Cabernet Sa	144	19	Italy	FALSE	0	0	2	4
16	June	Merlot	72	88	California	FALSE	0	0	5	0
17	July	Pinot Noir	12	47	Germany	FALSE	7	0	0	0
18	July	Espumante	6	50	Oregon	FALSE	0	11	2	1
19	July	Champagne	12	66	Germany	FALSE	0	0	2	3
20	August	Cabernet Sa	72	82	Italy	FALSE	0	0	4	2
21	August	Champagne	12	50	California	FALSE	0	0	2	2
22	August	Champagne	72	63	France	FALSE	0	0	0	21
23	September	Chardonnay	144	39	South Africa	FALSE	0	0	3	2
24	September	Pinot Noir	6	34	Italy	FALSE	12	0	0	0
25	October	Cabernet Sa	72	59	Oregon	TRUE	0	0	3	3
26	October	Pinot Noir	144	83	Australia	FALSE	8	0	5	2
27	October	Champagne	72	88	New Zealar	FALSE	0	0	6	3
28	November	Cabernet Sa	12	56	France	TRUE	0	0	4	2
29	November	Pinot Grigic	6	87	France	FALSE	0	15	2	0
30	December	Malbec	6	54	France	FALSE	0	16	2	4
31	December	Champagne	72	89	France	FALSE	0	0	10	7
32	December	Cabernet Sa	72	45	Germany	TRUE	0	0	3	1

Predictive model 3: Brand based clustering

Brand-based clustering, on the other hand, focuses on the brand of items customers purchase. Marketers can use this information to project what other brands those customers are likely to buy. Customers are then ordered according to Nike, Adidas, Under Armour, etc. Now you know what specific brands to pitch to certain customers. When a brand releases new products – you know who is likely to be interested.

What is a Simulation Model?

The question is actually misleading. A simulation can be comprised of one model or many interacting models. For instance we might take a statistical model and simulate it using random variates. In this case we are simulating one model. But suppose we want to simulate the effectiveness of a ballistic missile defense system. This system is comprised of models for sensors, interceptors, guidance systems, command and control, and so on. So in this instance we are simulating a system of models.

Perhaps I should just answer the question, "What is simulation?" **Simulation** is the process of executing a model over time (or another quantity) to see how a system performs. Yet this very answer is misleading as well. For a single run of a model does not generally yield reliable results. This is due to a number of inherent challenges of simulation, which I will not address here, but suffice it to say that we generally want to run a model over and over and see how it behaves on average.

There are several types of simulation models, some of which you may have heard.

Monte Carlo Simulation Model

Monte Carlo simulation is used to estimate stochastic processes where either the underlying probability distributions are unknown or difficult to calculate by exact computations. Monte Carlo simulations sample probability distributions for each variable to produce hundreds or thousands of possible outcomes. The results are analyzed to get probabilities of different outcomes occurring. For example, a comparison of a spreadsheet cost construction model run using traditional “what if” scenarios, and then run again with Monte Carlo simulation and Triangular probability distributions may show that the Monte Carlo analysis has a narrower range than the “what if” analysis. This is because the “what if” analysis gives equal weight to all scenarios, while Monte Carlo method hardly samples in the very low probability regions. The samples in such regions are called "rare events". Monte Carlo simulations may be used to:

- To make inventory-ordering decisions in the face of uncertain demands
- To price bids for contracts when competitors’ bids are uncertain
- To insure the consistent quality of manufactured goods when the sources of defects are uncertain
- To schedule and manage large projects when the durations of individual tasks are uncertain.
- To make investment decisions in sometimes volatile markets

Dynamic Simulation Models

Dynamic Simulation models include **continuous system simulation** and **discrete system simulation**. If time plays a substantive role, these are called continuous time models. Continuous systems simulations are typically solved using systems of differential equations. Discrete systems simulations are usually in the form of discrete time model or discrete event models.

Discrete Time Simulation Models

In Dynamic Time simulation, the models are iterated of fixed time increments. For example, we may want to experiment to the operation of a car wash over time and increment the model every ten minutes to observe the frequency that it is used through the working day. We might make decisions about its operation based on the frequency of use in peak hours. Or we might want to see how an interceptor missile performs over time as it attempts to destroy a theater ballistic missile, making observations every second. We can use these simulations to:

- To make inventory-ordering decisions in the face of uncertain demands
- To study the reliability, maintainability, and availability of machinery and systems
- To study the waiting times of customers is a certain service process

Discrete Event Simulation Models

With a Discrete Event simulation time plays a role, but the model is not iterated base on time. Rather, the model is iterated when an event occurs. Events are generally classified as arrivals, services, and departures (from queuing theory). The time between events is not constant, but are drawn from probability distributions. For example, we may know that the interarrival times of customers at a bank are exponentially distributed. Then, we would use random numbers to generate random arrival times. Likewise, we would have random service times. If the service times are significantly greater than the interarrival times, we might need to hire more bank tellers, for example. In the case of the ballistic missile intercept, we might want to observe its behavior every time its guidance system causes a change in course, rather than every second. We can use these simulations to study:

- Manufacturing plants
- Communications networks
- Transportation systems
- Combat operations
- Other systems that involve service, queuing, and processing

Simulation Architectures

Simulation architectures exist for running mixed time simulations, such as discrete time and discrete events simulations. One of the earliest architectures is known as Distributed Interactive Simulation or DIS. The High Level Architecture (HLA) is another example.

Conclusion

We use simulations to experiment with systems and perform analyses without having to use the actual systems. For example, it would be very disruptive to a call center if we performed experiments with it during its normal operating hours, which in many cases is 24-7. However, if we can construct a reasonable model of the call center, then we can use it to simulate call center operations and perform our experiments using the simulation.

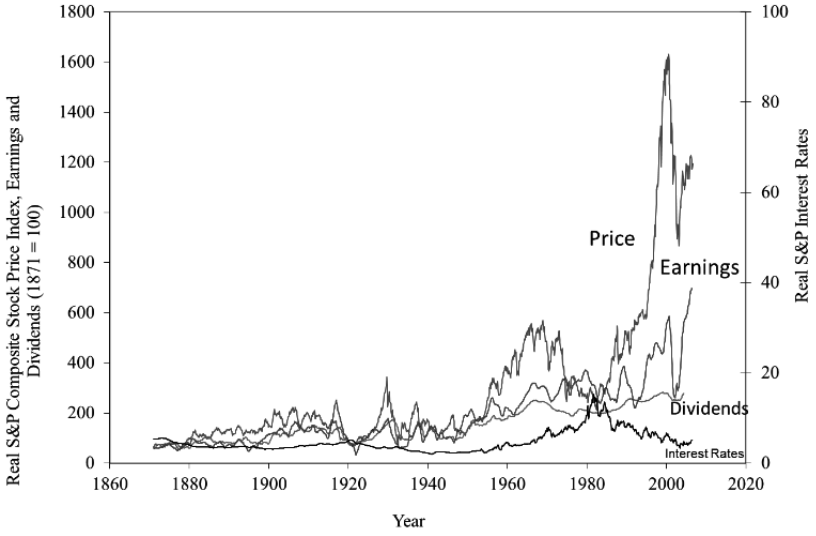
What are Stochastic Models

Is there a chance?

In my previous post on simulation, I used the term stochastic. What does that mean? This article is about the meaning of “stochastic” and its counterpart, “deterministic”.

Stochastic

Stochastic is synonymous with “random.” The word is of Greek origin and means “pertaining to chance” (Parzen 1962, p. 7). It is used to indicate that a particular subject is seen from point of view of randomness. Any system or process that must be analyzed using probability theory is stochastic at least in part. Stochastic systems and processes play a fundamental role in mathematical models of phenomena in many fields of science, engineering, and economics. Familiar examples of processes modeled as stochastic (or stochastic time series) include stock market and exchange rate fluctuations, signals such as speech, audio and video, medical data such as a patient's EKG, EEG, blood pressure or temperature, and random movement such as Brownian motion or random walks. Stochastic models are sometimes referred to as “probabilistic models”.



Plot of S&P Composite Real Price Index, Earnings, Dividends, and Interest Rates using data from
http://irrationalexuberance.com/shiller_downloads/ie_data.xls

Deterministic

A deterministic system is one whose resulting behavior is entirely determined by its initial state and inputs, and which is not random or stochastic. Processes or projects having only one outcome are said to be deterministic their outcome is 'pre-determined.' A deterministic algorithm, for example, if given the same input information will always produce the same output information. An example of a deterministic system is the physical laws that are described by differential equations, even though the state of the system at a given point in time may be difficult to describe explicitly. An example of a deterministic model is a logistic regression model of customers' propensity to buy life insurance products. A stochastic system is one that is non-deterministic.

References

Parzen, E. Stochastic Processes. Oakland CA: Holden Day, p. 7, 1962.

What Are Neural Networks?

Are you neural?

What I really meant to say is, “Are your models neural?” Should they be? Could they be? I have built some crazy models, but nothing like a neural network or auto-neural network (ANN). Maybe we should ask “Is you modeler neural?” Be that he may be, here is a layman's explanation...hopefully (just remember I am a modeler and cut me some slack).

We have been building these since the advent of computers…wait…my alter ego does not agree… really? It appears that the first artificial neuron was produced in 1943 by the neurophysiologist Warren McCulloch and the logician Walter Pits. But the technology available at that time did not allow them to do too much. So, this happened before computers? I thought this was a machine learning algorithm… Oh. It seems as though ANN were first described as an information processing paradigm that was inspired by the way biological nervous systems, such as the brain, process information.

The key element of this paradigm was the novel structure of the information processing system. It was composed of a large number of highly interconnected processing elements (neurons) working in unison

to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well.

There was an initial surge in ANNs with support from neuroscience. In 1958 Frank Rosenblatt stirred considerable interest in ANNs when he designed and developed the Perceptron. The Perceptron had three layers with the middle layer known as the association layer. This system could learn to connect or associate a given input to a random output unit. Other development emerged during this period but by 1969, Minsky and Papert wrote a book in which they generalized the limitations of single layer Perceptron to multilayered systems. In the book they said: "...our intuitive judgment that the extension (to multilayer systems) is sterile", which resulted in funding for these projects drying up and a period of “frustration” was incurred.

There was a re-emergence of interest and research in the late 1970s and early 1980s. However, in the 1990s, neural networks were overtaken in popularity in machine learning by support vector machines and other, much simpler methods such as linear classifiers. Yet the neurons would not dies and between 2009 and 2012, work at the Swiss AI Lab IDSIA produced award winning neural networks in competitions in pattern recognition and machine learning.

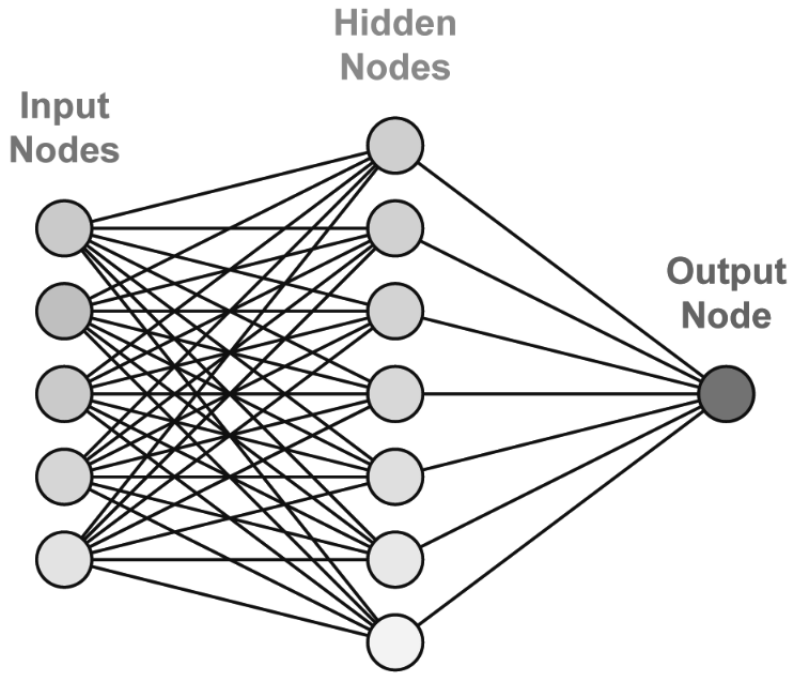
At any rate, An ANN is typically defined by three types of parameters:

1. The interconnection pattern between the different layers of neurons
2. The learning process for updating the weights of the interconnections
3. The activation function that converts a neuron's weighted input to its output activation

Without going into the mathematical detail (my book, “Predictive Modeling and Analytics”, does that), the main attraction in neural networks is the possibility of learning. Given a specific task to solve, and a class of functions, learning means using a set of observations to find optimal function (or optimal solution) in this set of functions, or a

function that solves the task in some optimal sense. The simple ANN can consists of three layers as depicted below. Within the hidden layer, learning occurs. In very simple pseudo-code, where ***f*** is an input node ***h*** is a hidden node and ***f**** is an output node, this might look like:

```
input f
DO apply learning rule to h
     DO until h stops learning (in an optimal sense)
          apply learning rule to h
          .
          .
          .
          h stops learning
     OD
OD
apply activation function
output f*
```



Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and

detect trends that are too complex to be noticed by either humans or other computer techniques. A trained ANN can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions. ANNs are universal approximators, and they work best if the system you are using them to model has a high tolerance to error. You would not want to use a neural network to balance checkbook! However they work very well for:

- capturing associations or discovering regularities within a set of patterns;
- where the volume, number of variables or diversity of the data is very great;
- the relationships between variables are vaguely understood; or,
- the relationships are difficult to describe adequately with conventional approaches.

Another instance where they might be useful is when the available response data is sparse. I have also found them to be useful in ensemble models along with logistic regression models. The tendency is to make a good preforming logistic regression model perform even better as a component model of the ensemble. In other words, it makes something that already tastes good taste even better.

Before using an ANN model, you should have (or if you are not the modeler, the modeler should have) an understanding of the mathematically complexity of ANNs. To some degree, they “black box” models; that is a model in which you cannot see the internal workings. In other words, using our pseudo-code, we cannot see how h is learning.

References

http://www.kurzweilai.net/how-bio-inspired-deep-learning-keeps-winning-competitions 2012 Kurzweil AI Interview with Jürgen Schmidhuber on the eight competitions won by his Deep Learning team 2009–2012

McCulloch, Warren; Walter Pitts (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics 5 (4): 115–133. doi:10.1007/BF02478259

Minsky, M.; S. Papert (1969). An Introduction to Computational Geometry. MIT Press. ISBN 0-262-63022-2.

Rosenblatt, F. (1958). "The Perceptron: A Probalistic Model For Information Storage And Organization In The Brain". Psychological Review 65 (6): 386–408. doi:10.1037/h0042519

Strickland, J. (2014). “Predictive Modeling and Analytics”. Lulu Inc. ISBN 978-1-312-37544-4

What is Discrete Event Simulation?

I do not go to the bank very much anymore. My handy iPhone app allows me to deposit checks online and the ATM gives me fast cash. But, have you ever waited in line for service at a bank on your lunch hour? Enter *Discrete Event Simulation*...

Discrete Event Simulation (**DES**) is the process of codifying the behavior of a complex system as an ordered sequence of well-defined events. In this context, an event comprises a specific change in the system's state at a specific point in time (arrival at the bank, service by a teller, etc.). Rather than stepping based on a time increment, like every second, DES advances based on events—events that may or may not be equally spaced in time.

Common applications of DES include:

- stress testing (a form of deliberately intense or thorough testing used to determine the stability of a given system or entity)
- evaluating potential financial investments
- studying call center operations
- modeling procedures and processes in various industries, such as manufacturing and healthcare
- studies that support system design

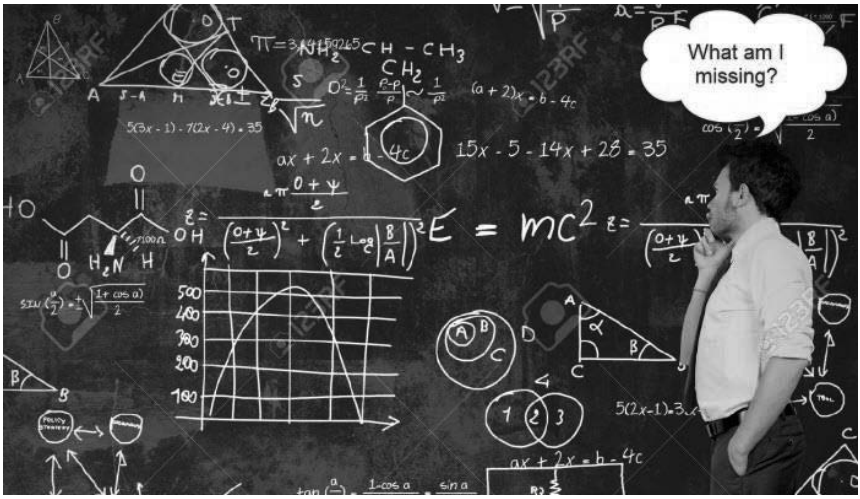
- studying reliability, availability and maintainability (RAM)
- just about anything that involves queuing (arrivals, waiting time, service time)

As an example of a situation that lends itself to DES, consider the operations of a call center. When a customer calls, you have an arrival event. If they have to wait for a representative they enter a queue until such time as their call is answered, which begins a service event. If their call was rerouted upon entering the system--the initial arrival event--the rerouting creates a new arrival event for the call center to which it was transferred, and so on. One can study where bottle-necks occur, rerouting efficiency, waiting times, balking (customers deciding not to join the queue if it is too long), reneging (customers leave the queue if they have waited too long for service), etc.

An effective DES process must include, at a minimum, the following characteristics:

- Predetermined starting and ending points, which can be discrete events or instants in time (arrivals and departures, for instance).
- A method of keeping track of the time that has elapsed since the process began (waiting time, for instance).
- A list of discrete events that have occurred since the process began (begin service, for instance).
- A list of discrete events pending or expected (if such events are known) until the process is expected to end.
- A graphical, statistical, or tabular record of the function for which DES is currently engaged (plot of waiting times and service times, for instance).

DES is commonly used to monitor and predict the behavior of investments; the stock market is a classic example. DES can also help administrators predict how a network will behave under extraordinary conditions, such as the Internet or claims call center during a major disaster, like a wildfire or earthquake.



What is Predictive Analytics Missing?

Simulation! Oh sure, there are pockets of analysts using simulation in predictive analytics and pockets of companies that use the results, but the use of simulation to predict behavior (e.g., propensity to do something) is not widespread.

I am taking a break today from the “how to” and “X reasons why” articles and offer instead a little editorial contemplation.

Several industries rely heavily upon simulation models: Aeronautics, Aerospace, Defense, Transportation, Medical and Maritime for example. So, why do these industries use simulation, while the Financial and Insurance industries do not?

Advantages of using simulation

Simulation provides a method for checking your understanding of the world around you and helps you produce better results, faster. It is an important tool you can use to:

- Predict the course and results of certain actions

- Understand why observed events occur
- Identify problem areas before implementation
- Explore the effects of modifications
- Confirm that all variables are known
- Evaluate ideas and identify inefficiencies
- Gain insight and stimulate creative thinking
- Communicate the integrity and feasibility of your plans

What are the Insurance and Financial industries missing?

I see, read about, and create statistical models for predictive analytics in this realm. It is amazing there are so many people who will accept a predictive model based on internal or third party data. Even if they realize how bad their data really is, they're still locked into statistical models. I know of at least two instances (two different companies) where their whole approach to predictive analytics is based upon time-series analysis—nothing else is considered.

So, your financial institution may be using time-series analysis solely to predict propensity, deploying the model and operating based on its results. Wow! What faith it takes to do something like that! There is very little understanding of why certain behaviors or events occur and very little insight into potential problems before implementation. And the data is blindly trusted—"it must be good since we collected it!"

What could they do differently?

Perhaps there is a fear of the unknown, and perhaps the unknown is anything outside of the realm of logistic regression or time-series analysis. This is of course what I do, so I am talking about impacting my livelihood. Moreover, I know, as George Box would echo if her were still with us, that my models are wrong! In order to get the closer to "right" (that is like trying to reach infinity), I know I need simulation. Simulate to ensure all the variables are known. Simulate to know the variables are reasonably distributed and experiment with alternatives. Simulate to gain insight into a particular propensity phenomenon. Simulate to…

We are doing just fine

The response to this might be that they already understand these phenomena well and they already know what variables are important. They probably did at one time. However, I know for sure my spending behavior, engaging behavior, borrowing behavior and so on, has been altered drastically by a struggling economy. “Oh, but our economy is healthy,” you might say. So then, why do products and services cost more and I get paid less. Ten years ago, a Senior Operations Research Analyst was worth $X thousand, but today he or she is worth $(X – Y) thousand, but that energy bill keeps rising without bound. Should we not even test the idea that the world has changed and what we used to know is now just a mystery? However, we are doing just fine, aren’t we?

The One-eye Man in the Kingdom of the Blind

You may have read a post by this title or at least seen the quote that this comes from, but for those of you who have no idea what I am talking about, here is an explanation. Kurt Vonnegut, author of Player Piano, said this in his book:

“Almost nobody’s competent, Paul. It’s enough to make you cry to see how bad most people are at their jobs. If you can do a half-assed job of anything, you’re a one-eyed man in the kingdom of the blind.”

The Operations Research analyst should have one good eye and hence lead the blind to safety. I argued before about how OR analysts must have a holistic approach to problem solving, and that includes simulations. “Oh, but I have not delved into simulation before”, one might say. My response is, “what are you waiting for?” LPs, IPs, and MILPs (if you are an OR analyst you know what these are) are not going to solve every problem we face, but it seems to be the focus in many OR educational programs. Get out of your comfort zone and get busy being a holistic problem solver.

Break the rules!

Every OR analyst should have a little bit of rebel inside them. Our approaches must be novel in many cases and the rules that say regression models must only be broken if we are to be worth our weight in salt (since we are not on a gold standard). When our models are not producing much lift (or net lift), something may be broken. It could be the data or the preconceived notion of customer propensity. Simulation can help find the root causes, for you can simulate a system or phenomenon without ever touching the real, operational system.

A little at a time

That financial industry might not be ready to pour the weight of their analytics resources into simulation, and that is understandable. So, perhaps start with a pilot study, get some results and show them how simulation can help them. Imagine where we might be if the financial industry had been doing this kind of "What if" analysis in the early part of this century. Would we have seen the events of 2008-ish coming?

Will it crash?

We do not fly launch vehicles in space without simulation, nor launch test missiles without simulation. It is too expensive—in people and material—to launch and fail. In our early years of space exploration, we experienced many crashes before the first human was launched into space, and we have had some failures since then. But a lot of time, effort and money go into simulation for these programs. We used simulation before executing Operation Iraqi Freedom, and one senior officer was quoted saying that we had fought that battle over and over in simulation, prior to deploying, which made the fight look easy (even though there is no "easy" in warfare). Imagine how many lives were saved by simulation. Will it crash? Of course it will, but to simulate the crash is much better than experiencing the crash in real time.

Conclusion

I have thus far failed to impress my financial institution customer with the importance of simulation. Technically, I am a failure as an OR analyst. Sure, I have given my customer the product they asked for, but sometimes the customer does not know what they really need. It is incumbent on us to help them to ask the right questions, develop the best business cases, and use the correct solution method—and that may be simulation.

Part IV – Statistical Maters

Why you should care about Statistics

I read a recent post that used the old adage, *"Lies, More Lies, and Statistics."* The author claims to have a few statistic courses and was apparently influenced by a professor who has never done anything that was "real". The author claimed that you can lie with statistics. He is wrong!

> *There are three types of lies — lies, damn lies, and statistics.*
>
> — Benjamin Disraeli

Who is Benjamin Disraeli?

Benjamin Disraeli was one of the great British politicians of the nineteenth century, Disraeli served twice as Tory Prime Minister (1868 and 1874 - 1880) and was also a prominent figure in opposition. He is most famous today for the bitter hatred between himself and his

political rival William Gladstone. He enjoyed the favor of Queen Victoria, who shared his dislike of Gladstone. Benjamin was not a statistician!

Can we lie statistics?

The average of a sample is a statistics, as is the shape of the data's distribution. Statistics tell you what is in the data—it describes the data and uses data to predict future events. Statistics is based on mathematical theory. If we have data comprised of the ages 12, 13, 11, and 10, the expression (12+13+11+10)/4 = 11.5 is the mathematical average of the ages. The calculation is absolute, based on the data. That is why "real" statistician are required to study mathematical statistics. The mathematics does not lie, so the statistics do not either.

"

I couldn't claim that I was smarter than sixty-five other guys--but the average of sixty-five other guys, certainly!

— Richard P. Feynman, Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character

Can we lie with data?

However, you can manipulate the data. You can collect the data that only supports your position or hypothesis. You can select specific data from a database that supports your position or hypothesis. You can make up data that supports your position or hypothesis. The statistics will still tell you what is in your data, and it will do so using the mathematical foundations that it uses for all data. Suppose we do not want the average age to be 11.5 (perhaps decimals scare us). We could replace the actual age "10" with "12" (even though the true age is 10). Then we would have (12+13+11+12)/4 = 12. Age 12 is the mathematical average of the ages. The problem is not the statistics, it is the data!

"

Outside of the killings, DC has one of the lowest crime rates in the country.

— Marion Barry

Can we lie with statistical models?

We modelers say that our models are only as good as the data they are built upon. We try to select data that is predictive, without the biases of our positions or theories. Sometimes the phenomena we are trying to predict, say the propensity to shed a product, is just not there. It could be that our intuition is wrong. It could be that the data is wrong. It could be that we do not have all the pertinent data. Hence, we are often left with the conclusion that we need more data, better data, or we cannot build a predictive model. Again, there is nothing wrong with statistical methods, say linear regression. The problem lies with the data: not enough, too much, missing data, and so on.

“

One of the first things taught in introductory statistics textbooks is that correlation is not causation. It is also one of the first things forgotten.

— Thomas Sowell, The Vision of the Anointed: Self-Congratulation as a Basis for Social Policy

Can we be confident in models?

No matter how hard we try to be unbiased, the data does not always allow us this luxury. Often, biases are built into the data itself, through sampling, collection and entry errors. However, statistics give us the ability to check for biases, correlation between variables, collinearity, etc. If it were not for statistics, we could not use data for anything—we could not turn data into information. Averages, variances, distributions, and so on are all statistics. Having 10 million test scores is useless, if you cannot describe the distribution of the scores, for example.

“

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

— Sir Ronald Fisher

Conclusion

Perhaps we should change the old adage to, *“Lies, More Lies, and Manipulated Data.”*

Why is Analysis like Hiking?

I was in the Rocky Mountains last week, near Buena Vista. I went on several hikes along the Colorado Trail and over some of the roughest terrain I have ever hiked, due to the rainfall and resulting washouts. Afterward, I went to my customer's site for a few days. I missed the hiking, but as I reflected on it the reason was not just because I was at work again. It was because it was like doing analytics.

Planning

Hiking requires extensive planning. My longest hike was 16 kilometers. Part of the hike (toward the end) was on the Colorado Trail. The beginning was up a set of switchbacks that went from about 2740 meter to 3050 meters. At the top, I had to go cross-country until picking up the main trail. During this first portion I would probably not see another human being. I would also be exposed to the intense sun. Since I was hiking alone, I needed to let someone know my route and the time I expected to return. If I were incapacitated on the isolated portion of my hike, it might be days before I was found. I also had to plan my load,

which needed to provide protection from the sun and mosquitos, protection from potential rainfall, water, and so on.

The descent would be on the Colorado Trail using a set of switchback. This portion of the trail has loose rocks and gravel with a steep descent. Thus, I had to plan for a means of stabilizing myself and my 50 lb. load, as well and a means for recovery should I slip. In all, I planed my route and communicated it to someone. I also planned a load that included one meal, a means for staying dry and warm, trekking poles, 3.0 liters of water, a means for cutting timber, rope for recovery, sunscreen and so on. The planning included contingencies for storms and staying on the ridge overnight.

Planning an analytics project is much the same. You have to consider the path that your analytic solution will take you down, where you might go astray, different contingencies, bottlenecks, time, tools required and how you would finish.

Execution

Executing the hike started with walking up a dirt road for half a mile to the base of the switchbacks, ascending to the top of the ridge, going north to the Colorado Trail, the west along the top of the ridge to the descent, and back to a dirt road for a 1.5 mile walk to the basecamp.

Although I was familiar with the route, rainfall in May had altered he appearance of the terrain and the path to the Colorado Trail was difficult to follow, so I had to dead-reckon until I came upon a cairn (a small pile of rock used to mark a trail). Once I found the trail, there was a gradual ascent to about 3050 meters. Here I stop for lunch and rested my legs that felt like jelly. The descent down the Colorado Tail would require strong legs and concentration. Often going down is much more difficult than going up, especially if you are carrying a load. Fortunately, I did not encounter any rain or storms and made it back to camp an hour early.

Project execution is similar. Often the first part of a project is not exciting, like walking up a dirt road. You have to define requirements with the client and stakeholder, acquire data, clean the data, explore the data, and so on. Once you start implementing your analytic approach,

you start climbing over rough terrain, attempting to a solution at the apex of your approach. Once you get your solution, you start your decent of post-processing and evaluating your solution. Finally you have to interpret and explain your solution on your way toward project completion.

Relating your Story

Back at basecamp, I wanted to relate my experience of hiking the Colorado Trail to numerous people, some of which were neither hikers nor outdoorsmen. I recognized the fact that some of the people I want to relate my experience to had taken the route by horseback a few years back. This knowledge helped by explain what I experience on the train from a ground perspective. I could talk about key intersections, the steepest switchback and so on, because I understood the context that they saw it in. I also talked about feet and miles instead of the metric measure on my map, using their language. For instance about 10,000 feet, instead of 3050 meters. I did not use terms like *trekking poles*, *alpine zone*, *bivouac*, *bushwhack*, *cairn*, or *gendarme*. Instead I used *hiking poles*, *above tree-line*, *temporary camp*, *hiking off-trail*, *small pile of rocks*, and *steep-sided rock formation along a ridge*. Finally, I allowed them to fell the weight of my load (backpack).

Relating the analysis story is similar. It almost requires a story teller who either has never hiked or has hiked and also travel by horseback. The storyteller has to be in the culture of their listeners and in the proper context. Obviously, we would not use terms like mean and type I error, but use average and false positive. We may believe that everyone know what a “mean” value is, but many in fact do not. We have to allow the stakeholder to “experience” our analytic journey within their culture and context.

Conclusion

The two words 'information' and 'communication' are often used interchangeably, but they signify quite different things. Information is giving out; communication is getting through.

—Sydney J. Harris

If we are not ensuring that our information is getting through to the client, we not only fail to relate our analysis journey, we also fail to deliver our analytic solution, for delivery requires both deployment implementation. Thus, as we hike, or climb, or ride, we also analyze.

12 Ways Not to Please Your Customer

I guess that "250.5 ways NOT to please your customer", seemed a bit daunting. Thus, I am going to cut it to the top twelve—the first six are additions to the original 6 (of 250.5). If you do not have a direct customer, then interpret customer as boss or supervisor. Also, if your spouse is your boss—as mine is, but I am not implying she is bossy—then interpret customer as spouse or significant other, which could be your Great Dane. I had intended to add a few every week, since there are 1א (pronounced aleph-1 and you'll have to read my post on Jessica Rabbit and the Number System) ways too NOT please your customer. Now I think I will just write twelve and most others are variations. I have committed most of the taboo actions at one time or another, so this is based on lessons learned by a slow learner.

1. Do not get to work after your customer, unless your customer goes to work a 2:30 AM, or has scolded you several times for being too early. I always want to be available to my customer, especially if they come in early and want to get thing accomplished be people typically show up for work. Sometimes when they know you are

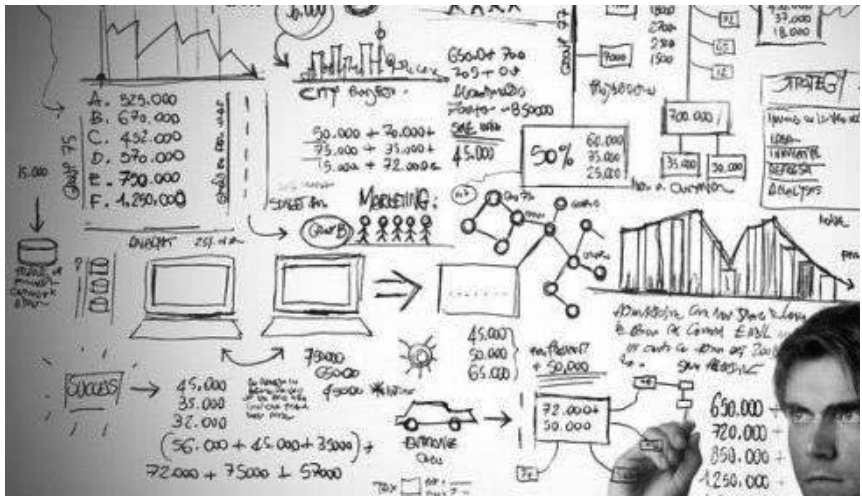
going to be there for them, they turn to you as someone they can depend upon.

2. Do not go home before your customer, unless they are a hardcore workaholic. At least check with your customer before you leave for the day, to make sure they have no pressing deadlines, or something just came up that they need help with. While others are having a beer at Bubba's Bar, you will be gaining the trust of your customer. However, do this within reason if you have a family. Your family is more important than going the extra mile for your customer.
3. Do not blindly follow the instruction of your customer. Wait, will that not displease them? Not if you do it selectively and with tact. You have the responsibility to help you customer do the right thing and make them successful. It is not your job to shine. It is your job to make sure them shine. If you want glory, change careers—become an actor, a singer, a professional quarterback, etc. Continued business and a paycheck is your reward. Ask questions that will cause the customer to consider, or reconsider courses of action, and do so with great humility. "Mam, I know you are probably right about this, but have you considered X?" In the end when the customer says execute, do so as I it was your own idea, unless ethics are involved.
4. Do not let your customer fail due to lack of planning. Make sure they have (or you have for them) considered redundancy and contingencies. Remember Charlie Beckwith, always take an extra helicopter (look up Operation Eagle Claw). Sub-rules: (1) Always use pencil—things change often, and (2) Do things faster, harder, and smarter
5. Never work on company business or personal stuff on the customer's time. Even if you are not charging, if this occurs during normal work hours, this could appear as unethical—perception is everything, and this is a quick way to lose trust, and perhaps your contract. Sometimes your contract will specify things that you can do on the customer's time, like company ethics training. If in doubt ask—questions do not cost as much as breaches of integrity.
6. Do not withhold information from your customer (unless it is proprietary and you have signed an agreement saying that you will not). This is not the same as lying, but it is related, and will destroy trust just as quickly as lying. Don't use information as power. Give them information to leverage as power.

7. Do not tell them what they can't do. Our job is to help do what they want to do, to achieve what they want to achieve. We have to help them fulfill their vision. That does not mean that we cannot advise them about risks and issues—we need to be diligent in doing this—but we have to also help them mitigate the risks and find a solution to get there. If it is NASA—well, at least before the next lunar exploration program was cancelled—then we have to help them reach the moon.
8. Do not tell them that something they want is out of scope. At least do not do this directly and without discussing scope with your contract representative. Going to the moon with a new launch platform (like the Ares I) is probably not one of those "out of scope" tasks. However, if they want you to build the Starship Enterprise, at least without Mr. Scott on your team, and maybe a Vulcan, then it might be out-of-scope. Often scope can be changed to help the customer get to here they want to go.
9. Do not tell them they are wrong. Your customer is never wrong—and if your boss is your spouse and your spouse is your wife, she is never wrong, unless you have a really nice doghouse in the back yard. Your customer's idea, desire, command, etc., may be misguided (out of alignment with their vision), and our job is to also help them align tasks, project and programs with their vision and to provide a roadmap to get there. Now, they may be wrong, but if you help guide them, they will come to this realization themselves, usually.
10. Do not—never ever ever—lie to your customer. In fact, never lie to anyone. It is perfectly okay to say that you cannot reveal information due to its security classification of proprietary nature, but never lie. If you did not cross a "t" or dot an "i", then own up to it and do it. This will, at a minimum, fulfill your contract obligation to complete a task, but it may also say to your customer, "this person can be trusted, because they always tell me the truth." Now, you may be incompetent and truthful at the same time, at the incompetence mat get your fired, but at least you maintain your integrity. The author of the Piano Player [Kurt Vonnegut] said this: "Ninety percent of people are incompetent at what they do. You would be surprised at how really bad they are. If you are marginally competent, you may go far." Thus, if you a marginally competent, non-liar, you at least have a chance of pleasing your customer.

11. Do not berate your customer behind their back—or to their side, or over them, or in a parallel universe. They WILL “hear it through the grapevine.” I have been a supervisor/manager and as such suffered the natural phenomenon of beratization—add that to your dictionary—but sooner or later, I always here about it, even if I have to sneak around and eavesdrop. Do not think your customer will not do the same. Nothing good can come from it!
12. Do not mention your customer on social media, even if it is in praise, unless it is part of a task on your contract to do so. Without belaboring the point, nothing good can come from it!

So, twelve makes a dozen—maybe I will do the more later—so perhaps we should title the post “Half-a-Dozen Ways NOT to Please Your Customer”, but 250.5 just sounds better to me. Moreover, I have changed my mind about half of a way not to please your customer:



Analytics and Statistics: Is there a difference?

Is there a difference between **statistics** and **analytics**? In previous posts, I have claimed that there is a difference. Here, I will attempt to explain my reasoning for approaching this conclusion.

What is Statistics?

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data [1]. In applying statistics to, e.g., a scientific, industrial, or societal problem, it is necessary to begin with a population or process to be studied. Populations can be diverse topics such as “all persons living in a country” or “every atom composing a crystal”. It deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments [1].

Two main statistical methodologies are used in data analysis: descriptive statistics, which summarizes data from a sample using indexes such as the mean or standard deviation, and inferential statistics, which draws conclusions from data that are subject to random variation (e.g., observational errors, sampling variation) [2]. Descriptive statistics are most often concerned with two sets of properties of a distribution

(sample or population): central tendency (or location) seeks to characterize the distribution's central or typical value, while dispersion (or variability) characterizes the extent to which members of the distribution depart from its center and each other. Inferences on mathematical statistics are made under the framework of probability theory, which deals with the analysis of random phenomena. To make an inference upon unknown quantities, one or more estimators are evaluated using the sample.

Standard statistical procedure involve the development of a **null hypothesis**, a general statement or default position that there is no relationship between two quantities. Rejecting or disproving the null hypothesis is a central task in the modern practice of science, and gives a precise sense in which a claim is capable of being proven false. What statisticians call an alternative hypothesis is simply a hypothesis that contradicts the null hypothesis. Working from a null hypothesis two basic forms of error are recognized: Type I errors (null hypothesis is falsely rejected giving a “false positive”) and Type II errors (null hypothesis fails to be rejected and an actual difference between populations is missed giving a “false negative”). A critical region is the set of values of the estimator that leads to refuting the null hypothesis. The probability of type I error is therefore the probability that the estimator belongs to the critical region given that null hypothesis is true (statistical significance) and the probability of type II error is the probability that the estimator doesn’t belong to the critical region given that the alternative hypothesis is true. The statistical power of a test is the probability that it correctly rejects the null hypothesis when the null hypothesis is false. Multiple problems have come to be associated with this framework: ranging from obtaining a sufficient sample size to specifying an adequate null hypothesis.

What is Analytics?

Actually, analytics and statistics share a common thread: they both use statistical procedures and analyses. However, unlike statistics, the analytics scientist often deals with analyses where there is no assumed null hypothesis, and subsequently employs machine learning algorithms in the analyses. This is an important distinction in understanding my point of view. For instance, I very seldom approach predictive modeling

with a null hypothesis. On the other hand, I do perform statistical analyses with a null hypothesis at the forefront—there is a subtle but distinct difference.

I will not go into great detail here, since my last post described Predictive Analytics, but I will summarize a description of analytics. **Analytics** is the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, machine learning, computer programming and operations research to quantify performance or predictions. Analytics often favors data visualization to communicate insight [3].

Firms may commonly apply analytics to business data, to describe, predict, and improve business performance. Specifically, arenas within analytics include predictive analytics, enterprise decision management, retail analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modeling, web analytics, sales force sizing and optimization, price and promotion modeling, predictive science, credit risk analysis, and fraud analytics. Since analytics can require extensive computation, the algorithms and software used for analytics harness the most current methods in computer science, statistics, and mathematics [4].

I emphasized the use of optimization in analytics applications because it is a method used in operations research, not in statistics. Optimization may be performed either mathematically or heuristically. The heuristic approach often “violates” the principals of mathematics (in terms of proving them out) in arriving at a “close enough” solution. As a student years ago, I worked under a department head that said, “This is really good stuff, but it is not mathematics.” Moreover, many of the applications of analytics are not stochastic—not based on probability.

Conclusion

If you accept the descriptions of statistics and analytics that I have presented here, then there is an obvious difference between the two disciplines, thought they do share some things. This is not meant to belittle the work of the statistician. The statistician is very professional,

understands their discipline extremely well and provides valuable, focused analyses. While the statistician is a “focused” of “specialized” discipline, analytics is much more general—the analytics professional is to some degree a “jack of all trades”. Often, the analytics professional will turn to the statistics professional for help with analyses where statistical hypotheses are paramount.

References

Dodge, Y. (2006) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9

Lund Research Ltd. "Descriptive and Inferential Statistics". https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php. Retrieved 2015-01-26.

Strickland, J. (2014) Predictive Modeling and Analytics. Lulu. ISBN 978-1-312-37544-4.

Strickland, J. (2015) *Predictive Analytics using R*. Lulu. ISBN 978-1-312-84101-7.

Are Statisticians a Dying Breed?

“

“Facts are stubborn things, but statistics are pliable.”

— Mark Twain

I am not a statistician, though I perform statistical analyses almost every day; I am an Operations Research Analyst. I have seen a lot of discussion regarding the title of this article, and as someone who does not have a comprehensive background in statistics, I would say loudly, “I need statisticians”.

“

“Every time I sit with our general manager at a baseball game, and there's number-cruncher and statistician guy - I'm sitting around - they start talking about stuff, and I say, 'What's that? I've never heard of that one before.'”

— George Brett

People have been espousing the idea that statisticians have no role in analytics and data science. Almost all of what I do in predictive analytics is statistical in nature. One of the most important statistical procedure is

design of experiments, both for collecting the data and for performing the analysis. Statisticians are very good at this.

“

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”

— Ronald Fisher

Statisticians do more than regression analysis, hypothesis testing and analysis of variance (ANOVA). More often than not, statisticians are cleaning data, performing data reduction and other prepossessing functions. Today, the statistician does more than they have in the past and their service is crucial to developing unbiased information to decision makers.

“

“Definition of a Statistician: A man who believes figures don't lie, but admits than under analysis some of them won't stand up either.”

— Evan Esar

Can statisticians do everything that is required in analytics and data science? For the most part, probably not, but neither can I. I am sure there are a few super-humans out there that can, but these fields require multi-discipline teams in order to provide a holistic approach. Depending on the analyses at hand, teams should be comprised of statisticians, psychologists, marketing professional, data architects, operations research analysts, and so on.

“

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

– John Tukey

We need to change the conversation from who is not needed to who is needed and why. The fray at hand is counter-productive.

“

“I couldn't claim that I was smarter than sixty-five other guys--but the average of sixty-five other guys, certainly!”

— Richard P. Feynman, Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character

Statistics is Obsolete

“

Experts often possess more data than judgment.

— Colin Powell

Hey, we no longer need **statistics!** Cancel all college and university statistics programs…fire all the statistics faculty…shut down SAS…stop wasting time on R and Python…the end has come…there is a new beginning…ALL DATA.

Why do we no longer need Statistics?

“

Being deeply knowledgeable on one subject narrows one's focus and increases confidence, but it also blurs dissenting views until they are no longer visible, thereby transforming data collection into bias confirmation and morphing self-deception into self-assurance.

— Michael Shermer

I am sure you have heard of BIG DATA, but have you heard of ALL DATA? Probably not, because I think I may just invented the term. Apparently

there are some major players who actually believe that when they have "all of the data", they no longer have to take samples, and that statistics only applies to samples. Since we do not need to take samples anymore, then we do not need statistics anymore. By the way, Statisticians, we do not need you anymore. Go find a nice job as a greeter at Walmart, not that there is anything wrong with Walmart greeting—I kind of like the special touch.

What's wrong with this Picture?

"

Not everything that can be counted counts, and not everything that counts can be counted.

— Albert Einstein

First, the CEO's, CTO's, and other O's that are espousing this view absolutely have no earthy idea (or one from a parallel universe) of what they are talking about. I would call them naïve, but they are really somewhere between "so full of themselves" and just plain oblivious to reality! I think their issue is their lack of understanding statistics, statistics anxiety, or a combination of the two. But, ALL DATA will take care of that, since it will not require any statistics.

Second, ALL DATA does not equal PERFECT DATA. I deal with colossal data sets. In my experience roughly 80% of the data is useless when building a model of some phenomenon, like the propensity to buy product X or the probability that the O-ring on component Y of the latest space launch vehicle will fail. Moreover, regarding the 80% of useless data, about half of that is really bad data. Of the twenty percent that is useable, only about 10% of it actually contains strong predictors. And this stuff is pretty easy to figure out using Weight of Evidence (WOE) and Information Value (IV) algorithms. But never fear, ALL DATA has morphed into ALL-PERFECT DATA.

Third, I think it is great that we have ALL-PERFECT DATA, but how are you going to process it? I know, you'll just push a button get a prediction. And inside that "black box" there are no statistics involved. There is no regressing, no averaging, no transforming, no imputing, on so on. The

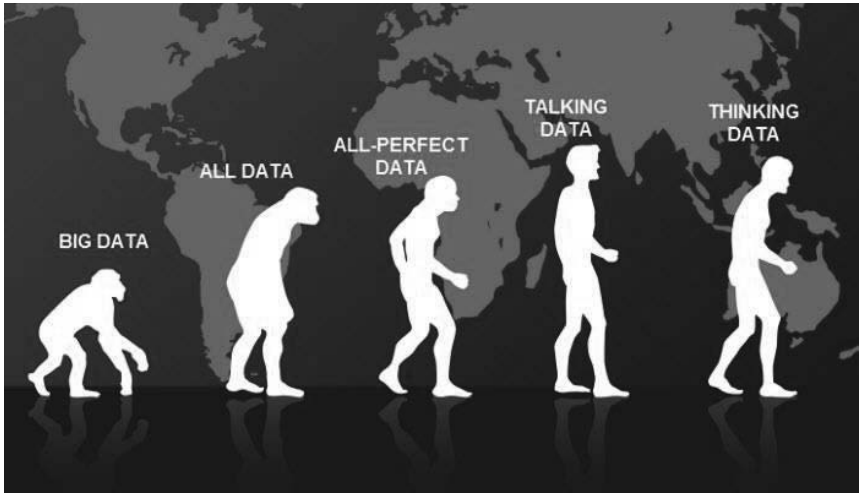
data is just speaking for itself. Oh wait, we have gone from ALL-PERFECT DATA to something much better, TALKING DATA.

“

One person's data is another person's noise.

— K.C. Cole

Fourth, this implies we do not have to analyze the results. I suppose that is due to our new TALKING DATA. It speaks the results itself! No more histogram, means, variances, Skewness, etc., because we do not care. Our TALKING DATA will tell us what decisions to make. Wait a minute…if our TALKING DATA is making decisions, then it is really THINKING DATA. So now we do not need any decision makers! Fire all the CEO’s, CTO’s, and other O’s. Wait another minute! Who is going to implement the decisions made by our THINKING DATA? I know, it’s the MAGICAL DATA (probably in the form of unicorns). Wait another minute, if we have MAGICAL DATA, we do not need anyone else, for magic can do it all. Layoff everyone. No one ever needs to work again, thanks to MIGICAL DATA!



Fifth, I do not really have a fifth so I am going to ask my BIG, ALL, PERFECT, TALKING, MAGICAL DATA for its opinion. What was that? Our UNICORN said, "all those people who think we do not need statistics are not learned!" Now, do not get angry with me. I did not say that. It came for the UNICORN you created…

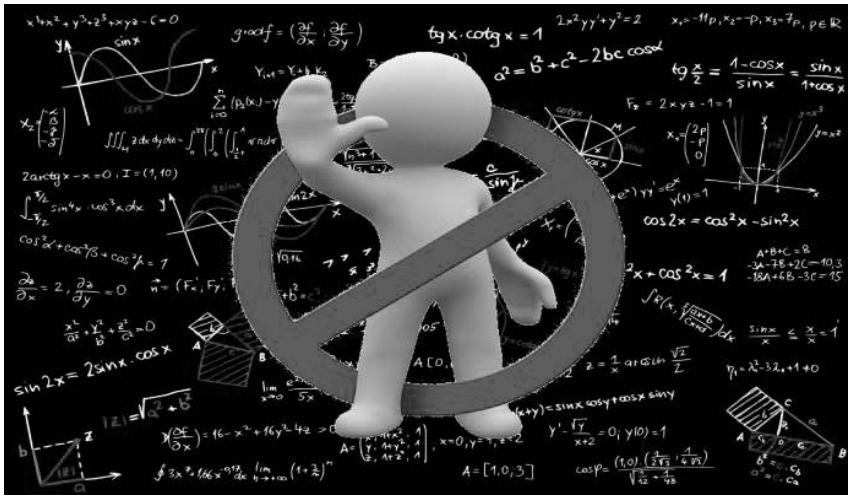
How do we fix the Picture?

“

Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.

— Aaron Levenstein

That is an easy one, fire all the brainless CEO’s, CTO’s and other O’s who wouldn’t know a statistic if it slapped them in the face. But that would not work. Who would create the UNICORN?



Math, Physics and Chemistry are Obsolete

I mentioned in “Statistics are Obsolete” that there are some pretty influential people who reject statistics as the underlying science of data analysis with respect to big data, or in their words ALL DATA (specifically, “We have all of the data.”). I think the essence of the issue for those people is that simply do not understand statistics, or for that matter any science that has “formulas”, like math, physics and chemistry.

The ALL DATA proponents seem to cluster with those who have softer quantitative degrees or degrees in management, psychology, and so on. Of course, cluster analysis is obsolete, as are principle component analysis, regression analysis and so on (and probably any kind of analysis that involves using the scientific method). So just ignore the clustering.

If the ALL DATA people prevail, then I suppose Math, Physics and Chemistry are obsolete, as well. Therefore, let’s be proactive and shut-down all the mathematics, physics and chemistry programs. All you math, physics and chemistry majors, you should switch majors to anything that does not require mathematics. All you tenured professors out there, well, you just lost your tenure, but I hear MacDonald’s is hiring.

Parents, stop teaching your kids how to count. I know it's cute, but it's also futile in light of ALL DATA. We do not need anything numerical anymore. In fact, we do not need anyone to think anymore. These absolutely brilliant ALL DATA supporters and their THINKING DATA will think for us (they are not really brilliant, just cleaver).

Bummer, ten years of formal mathematics, physics and chemistry education is down the drain. Does that call for a toga party or a road trip?

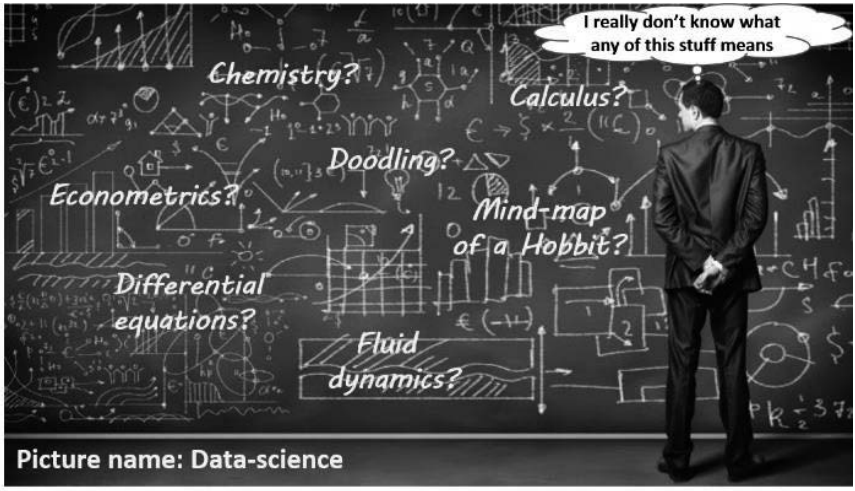
So I believe I have narrowed down the requirements for one work in Data Science or Big Data Analytics with our new ALL DATA:

- Obtain a High School Diploma with at least a 2.0 grade point average
- Take at least one *Coursera* class in anything except math, physics, chemistry or statistics
- Obtain a bachelor's degree in Under–Water Weaving or equivalent
- Score at least 25 on your SATs
- Have at least one year experience doing just about anything

If you exceed these requirements, then you are over-qualified. After all, with the possession of ALL DATA you just need a chimpanzee to push a button on a black box and predict anything you care to predict, because it is in the data.

Could I interest you in some ocean-front property in Arizona?

Part V – Data Science Concerns



Data Scientists are Dead, Long Live Data Science!

I know, that is pretty confusing. What we have, however, is a distorted, all-over-the-place, inadequate, confusing, and too broad a definition of "data scientist". I do not know where we went wrong, but we have the equivalent of "bridge scientist" if you happen to work on a bridge, whether you design it, build it, resurface it, and who knows, maybe just drive across it.

Call it what it is

I think that life would be better in all things data if we just called thing what they are, like calling mathematicians by their function: number theorist, math teacher, mathematical modeler, algebraist, etc. How about *datatician?* Will that cause confusion with statistician?

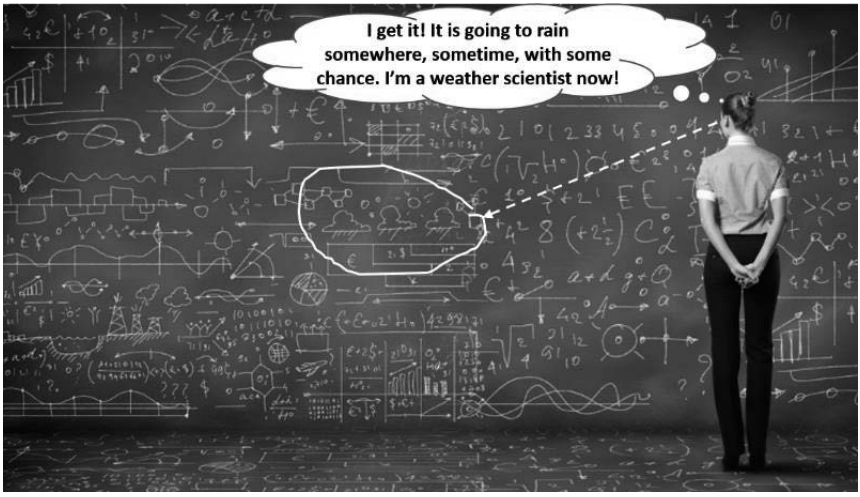
Mimic Mathematical Sciences?

Maybe “data science” is okay and the problem is “data scientist”. We have “mathematical sciences” (note the plural), but few mathematicians would call themselves a “mathematical scientist”, even though that sounds pretty cool. The problem is it doesn’t describe what one does in the field of mathematics. Algebraist, however, says that one works in the mathematical area of algebraic structures. The mathematical sciences might looks like this (I borrowed these from the National Science Foundation):

- Algebra
- Number Theory
- Analysis
- Applied Mathematics
- Combinatorics
- Computational Mathematics
- Foundations
- Geometric Analysis
- Mathematical Biology
- Probability
- Statistics
- Topology

This is not an all-inclusive list. Note that statistics falls under mathematical sciences at the NSF. And “real” statisticians know from the mathematical statistics courses why this might be the case.

A Data Science Taxonomy?



Perhaps under "data science", we eliminate the term "data scientist" and call out titles by what we do:

- Data architect
- Data processor
- Data miner
- Data modeler
- Data analyst
- Data explorer
- Database administrator
- Database developer
- Algorithm developer
- etc.

Of course, the guru of data analysis and data modeling is the "statistician". But perhaps statistician can be divided in like manner, as the mathematician: regression analyst, exploratory data analyst, etc.

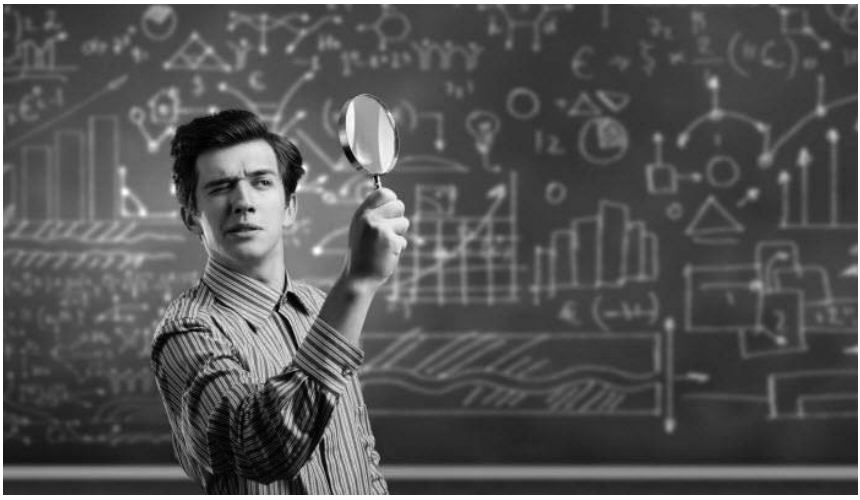
Fun with Machine Learning

What about machine learning?

- Machine learning architect
- Neural networker
- Classification tree farmer
- Random forest ranger
- (now I am just being silly)

Conclusion

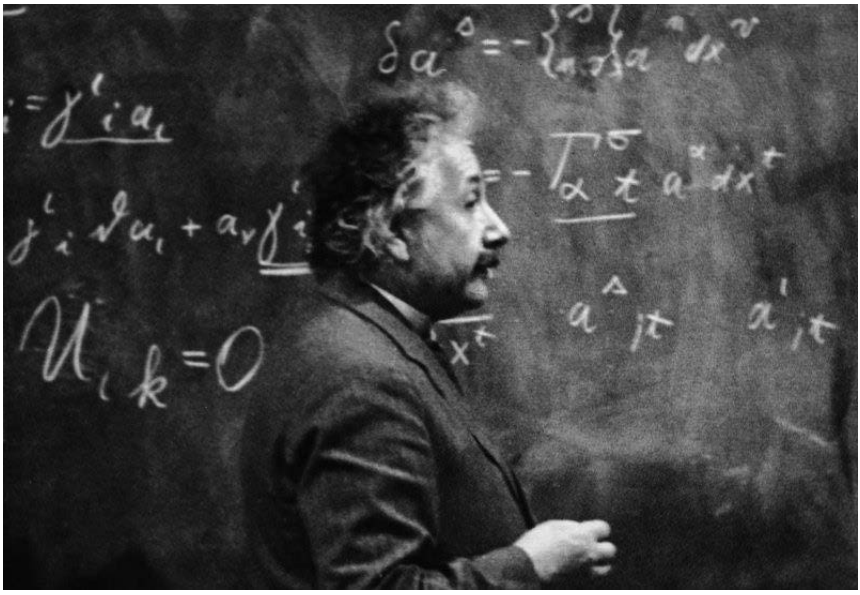
Somehow, “data scientists does not cut it for me, and there are many in the industry and the field that have similar issues. I have read and ben told that many organization will no longer hire data scientist, partly because there is no standard for what a data scientist is or does. I do not think we will reach a standard, so let’s label a person who works with data by what they actually do. Then hire them accordingly.



So you think you are a Data Scientist?

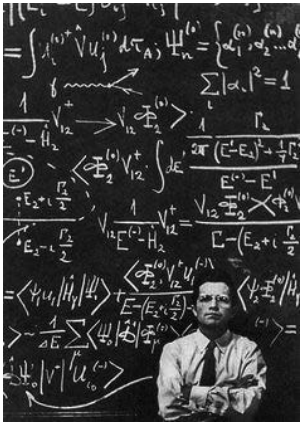
Many are writing articles about the skills required to be a data scientist. I disagree with most of them. Before you can be a data scientist, you must qualify to be a **scientist.** A degree in management does not qualify you to be a Chemical Scientist or Mathematical Scientist, nor does it qualify you to be a Data Scientist. A degree in Management Science or Operations Research could qualify you, depending on the course content.

This article represents my opinion and will hopefully add substance to an ongoing dialog regarding what constitutes a data scientist. Also, there is nothing derogatory about being a data analyst, data processor, data architect, database administrator, business analyst, marketing analyst, analytics consultant and so on, but doing these jobs do not make you a data **scientist.** Neither does a few courses from *Coursera.*



This guy might qualify to be a data scientist

In general, **scientists** perform research toward a more comprehensive understanding of nature, including physical, mathematical and social realms. Scientists are also distinct from engineers, those who design, build and maintain devices for particular situations. When science is performed with a goal toward practical utility, it is called applied science. An applied scientist may not be designing something in particular, but rather is conducting research with the aim of developing new technologies and practical methods.



If you don't understand these formulas, you might not be a data scientist.

The following are requirements for scientists from a General Services Administration (GSA) Schedule related to machine learning (*www.**gsa**.gov/*), for instance (this is the closest I could find for data science). Many who call themselves Scientists (i.e., Data Scientists) are

not qualified to be an Assistant Scientist. Though I may be qualified to be a Principal Data Scientist, I do not use the title. Instead, I use Big Data Analytics Consultant, being careful not to misrepresent myself.

Assistant Scientist I/II

- BS in mathematics, science, engineering or equivalent
- One to three years research experience or equivalent course/project work
- Strong foundation in object-oriented analysis and design
- Knowledge of and experience with a variety of software languages and platforms
- Demonstrated research skills
- Demonstrated ability to understand appropriate algorithms and concepts
- Demonstrated ability to write clear, grammatically-correct documents
- Demonstrated ability to effectively communicate and work with other members of project team
- Strong analytical skills
- Strong fit with Company goals and values (work ethic, attitude, interest in work)

Scientist I/II

- MS (or BS plus three years of experience) in mathematics, science, engineering or equivalent
- Two to five years research experience or equivalent course/project work
- Demonstrated object-oriented analysis and design skills
- Knowledge of and experience with a variety of software languages and platforms
- Proven research skills
- Proven ability to understand appropriate algorithms and concepts
- Demonstrated ability to employ a hierarchical writing process to generate clear, grammatically correct documents
- Demonstrated ability to provide technical guidance to interns and other more junior team members

- Strong analytical skills
- Strong fit with Company goals and values (work ethic, attitude, interest in work)

Senior Scientist I/II

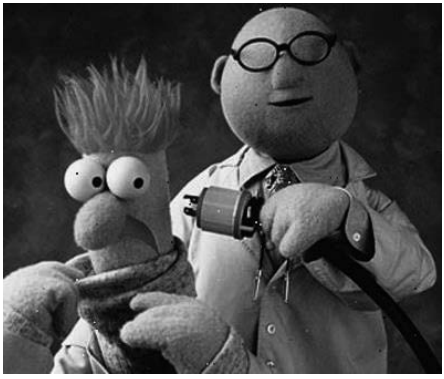
- PhD (or MS plus four years of experience) in mathematics, science, engineering or equivalent
- Three to six years research experience or equivalent course/project work
- Strong object-oriented analysis and design skills
- Knowledge of and experience with a variety of software languages and platforms
- Strong research skills proven by publication record or equivalent references from the research community
- In-depth experience with appropriate technologies
- Strong ability to employ a hierarchical writing process to generate clear, grammatically-correct documents
- Two years of experience leading teams
- Demonstrated program management skills
- Strong analytical skills
- Strong fit with Company goals and values (work ethic, attitude, interest in work)

Principal Scientist I/II

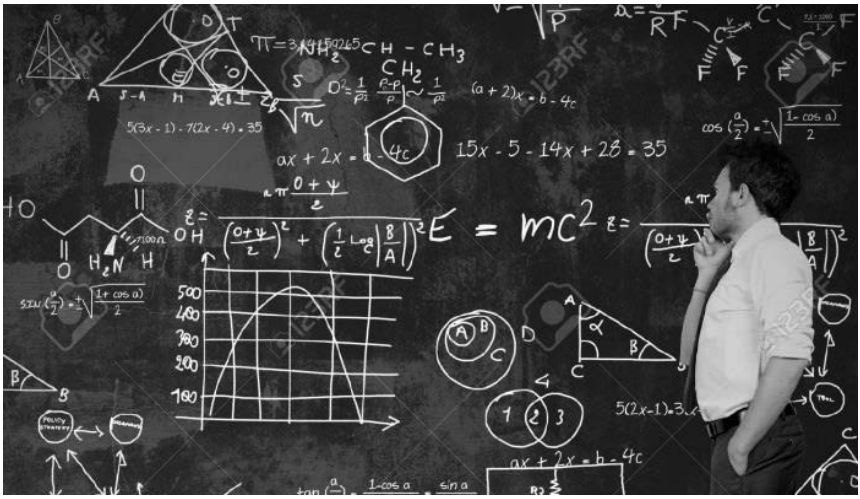
- PhD (or MS plus five-years of experience) in mathematics, science, engineering or equivalent
- Four plus years research experience or equivalent course/project work Strong object-oriented analysis and design skills
- Knowledge of and experience with a variety of software languages and platforms
- Superior research skills proven by extensive publication record or equivalent references from the research community
- Expert knowledge of appropriate technologies
- Demonstrated ability to write winning proposals
- Three years of experience managing teams
- Strong analytical skills

- Strong fit with Company goals and values (work ethic, attitude, interest in work)

Regarding the education requirements for mathematics, science, engineering or equivalent, the disciplines for "equivalent" are not mentioned. However, in my experience these would generally include statistics.



No wonder companies who are hiring data scientist are being disappointed in their performance and contribution. Too many people are watering-down the requirements. As a community of professionals, we need to define the role of data scientist in terms of requirements, rather than trying to be all inclusive for people who work with data. Moreover, the requirements might possible vary from country to country, so we should look at common requirements. For instance a Master's degree in Economics or Informatics in Europe in not necessarily the same as one in the United States. What really matters in the end is the scientist's role in turning data into information and information into "useful" information. Oh, and you should have one of those white lab coats...wireframe glasses are optional.



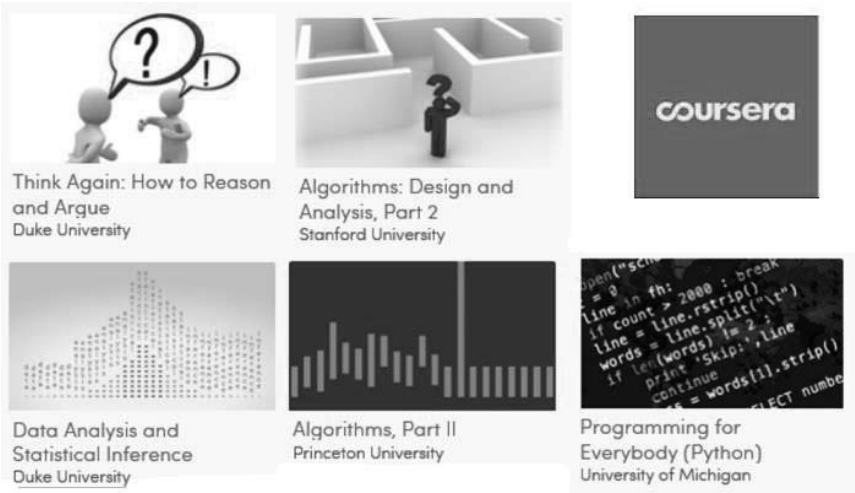
5 Signs that you might be a Data Scientist

Why are people so down on Data Scientists? Probably because they don't know what one looks like. There are a lot of folks out there calling themselves Data Scientist just because they had some introductory statistic courses, and maybe they can make really cool charts in Excel. All that makes you is someone who had some introductory statistic courses and can make some really cool charts in Excel. So what might a real Data Scientist look like? Here are some signs that you might "really" be a Data Scientist.

1. **You are a scientist.** In order to be a data scientist, you must first be a scientist. The Government Services Administration (GSA) defines a scientist as a subject matter expert who: “Provides technical knowledge and analysis of highly specialized applications and operational environment, high-level functional systems analysis, design, integration, documentation and implementation advice on exceptionally complex problems that require graduate level knowledge of the subject matter for effective implementation.” Thus, you might write charts diagrams and formulas on a chalk board and contemplate them for hours.

2. **You have a graduate degree in a science** or related field, i.e., Analytics, Predictive Modeling, Risk Analysis, Applied Statistics, Database Architecture, etc. Most occupations where the title "Scientist" is used, requires a doctorate degree. I am taking the liberty to soften the requirement to a **Masters of Science** degree. Sorry, a MBA does not make the cut. Alternatively, you professional certification like Certified Analytics Professional (CAP).
3. **See 1 and 2.** Moreover, you include some or all or the following skills in your profile: Data Scientist, Analytics Scientist, SPSS Modeler, SAS Programming, Regression Analysis, Statistical Modeling, Logistics Regression, Machine Learning, SQL, Python, etc.
4. **See 1 and 2.** Additionally, your boss or customer listens to you and takes action when you present your analysis and recommendations. Moreover, profit was made or money was saved (or for non-profits some measure of performance was achieved) based on your work.
5. **See 1 and 2.** Not only can you crunch numbers with the best of them, you can logically present the results of your analysis in plain language, or at least your customer's language. Additionally, you make contributions to the body of knowledge through scientific experimentation or studies, resulting in blogs, articles, papers, or presentations.

The *real scientists* worked very hard to become one; they did not just attach a label to their title. That does not mean that only scientists work with data. Beside data scientists, other people who work with data may include data analysts, data consultants, data architects, data miners, analytics consultants, and so on. However, to be a data scientist you need the prerequisites outlined in 1 and 2. For hiring managers, project managers, human resource professional, etc., when you hire a data scientist, make sure you actually hire a **scientist** who works with data. Otherwise, do not call them "scientists". You might be less disappointed in their work.



How can I be a Data Scientist?

Several aspiring data scientist and analytics scientist have asked me how to achieve their knowledge and training goals. Though there are many ways to accomplish this, below are two options offered through Coursera.

Johns Hopkins University – Data Science

- The Data Scientist's Toolbox
- R Programming
- Getting and Cleaning Data
- Exploratory Data Analysis
- Reproducible Research
- Statistical Inference
- Regression Models
- Practical Machine Learning
- Developing Data Products
- Data Science Capstone

University of Illinois at Urbana-Champaign – Data Mining

- Pattern Discovery in Data Mining
- Text Retrieval and Search Engines

- Cluster Analysis in Data Mining
- Text Mining and Analytics
- Data Visualization
- Data Mining Capstone

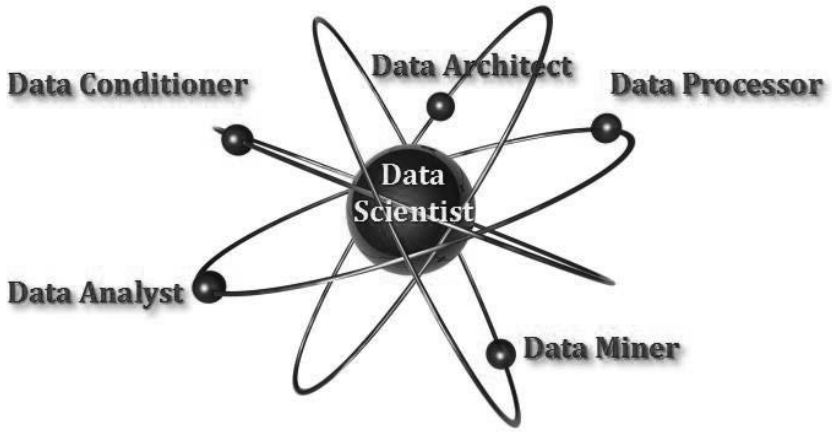
SAS and Hadoop for Learning

If you need SAS for training or self-study, you can get the university addition through the link below. You do not have to be student or faculty, but you cannot use it for work.

- http://www.sas.com/en_us/software/university-edition.html

If you need to use hadoop for training try:

- http://hortonworks.com/products/hortonworks-sandbox/#install



Why you might not want to be a Data Scientist

Data Scientist was one of the world's most popular technical jobs, at least in 2014. From what I have been told, of all the major countries, it was second only in the United States. Universities everywhere are creating, or have already created, graduate programs in Data Science. Data is becoming more and more available and in large quantities. Recently, President Obama appointed a US Chief Data Scientist (Dr. DJ Patil). Data science is at center stage!

So, why would you not want to be a data scientist? The following reflects my opinion only, and should not be construed as advice, merely something to consider.

Saturation

We are putting more and more data scientists into the field every day. Personally, I am asked several times a month to consider filling a data scientist role here or there. Today, there are ample opportunities for data scientists. It may be different tomorrow. If you are thinking about becoming a data scientist is some way, shape or form, especially if you have just started college or will start soon, consider the idea that the job

market may be saturated by the time you are qualified. Ask yourself if there are alternative courses of study and course of action.

Internet of Things (IoT)

I think the jury is still out deliberating guilt or innocence. The Internet of Things is a difficult concept to define precisely. In fact, there are many different groups that have defined the term, although its initial use has been attributed to Kevin Ashton, an expert on digital innovation. Each definition shares the idea that the first version of the Internet was about data created by people, while the next version is about data created by things. In 1999, Ashton said it best in this quote from an article in the RFID Journal:

If we had computers that knew everything there was to know about things - using data they gathered without any help from us - we would be able to track and count everything, and greatly reduce waste, loss and cost. We would know when things needed replacing, repairing or recalling, and whether they were fresh or past their best.

Data without a human touch? I do not know, nor does my crystal ball show me, a glimpse of the impact that IoT could have, but it is something to consider.

Imposters

I am sure I will get in trouble here, but it would seem to me that before you can be a data scientist, you have to first be a Scientist. The terms Scientist, Engineer, Analyst, and so on, have specific meanings. For instance, the Government Services Administration defines a scientist as a subject matter expert who:

- Provides technical knowledge and analysis of highly specialized applications and operational environment, high-level functional systems analysis, design, integration, documentation and implementation advice on exceptionally complex problems that require graduate level knowledge of the subject matter for effective implementation.

- Applies principles, methods and knowledge of the functional area of capability to specific task order requirements, advanced mathematical principles and methods to exceptionally difficult and narrowly defined technical problems in engineering and other scientific applications to arrive at automated solutions.

They define an Engineer a subject matter expert who:

- Provides technical knowledge and analysis of highly specialized applications and operational environment, high-level functional systems analysis, design, integration, documentation, and implementation advice on exceptionally complex problems that necessitate high-level knowledge of the subject matter for effective implementation.
- Participates as needed in all phases of software development with emphasis on the planning, analysis, modeling, simulation, testing, integration, documentation and presentation phases.

They list the following occupational specialties separately:

- Data Architect
- Data Communication Manager
- Data Warehousing Administrator
- Data Warehousing Analyst
- Data Warehousing Programmer
- Data/Configuration Management Specialist
- Database Analyst/Programmer
- Database Manager/Administrator

If you are a hiring manager, watch out for imposters. If you are an aspiring data scientist, make sure you are qualified before you make your claim. Imposters have the potential for damaging the reputation and expectations of the occupation.

Alternatives

There are really some cool alternatives to becoming a data scientist and still being involved with data science. I do not call myself a data scientist. Instead, I claim to be a "predictive analytics scientist". However, by prior

training and occupation, I am really an Operations Research Analyst (or Scientist). Operations Research (OR) has been around since World War II and OR analysts did things like plan the Normandy invasion (Operation Overlord). The International Forum for Operations Research and Management Science (INFORMS) now sponsors the Certified Analytics Professional (CAP) program. Many universities offer degree program or concentrations in OR. ORs work for cruise lines, airlines, the entertainment industry, the wholesale and retail sales industries, the trucking and train industries, the automotive industry, manufacturing industry, and the departments of Defense, Transportation and others. In addition to data "stuff", they perform optimization, predictive modeling, inventory control, and simulation.

OR Analyst (along with other disciplines like Statistician) may be a viable alternative to Data Scientist.

Count the Costs

Whatever you consider doing, evaluate the potential cast and outcome. If you do pursue becoming a "real" data scientist, and all of my "points to consider" actually come to pass, you will still be a highly skill individual. It might not be a great leap to acquire some additional knowledge and become an OR Analyst or an Actuary, for instance. However, if you are unwilling or unable to achieve the requirements for becoming a scientists, then perhaps data analysts or data miner would be appropriate fields. Just think before you leap...

I am an Analyst, really!

Many people say they are analysts, as I do. But are we really analysts? Do we perform analysis? That is, do we take problems and break them into smaller pieces to solve, and then develop an aggregate solution. Analyst is a term for someone who uses an analytic approach to problem solving. The opposite problem solving approach is intuition.

Analysts come is a variety of flavors: intelligence analysts, operations research analysts, risk analysts, text analysts, and so on. But all analysts have this in common: they use the analytic approach, breaking large problems down into smaller ones and solving them. Someone who uses the alternative approach is not an analyst, unless intuition is used in conjunction with the analytic approach.

Definition: an *analyst* is an individual who uses the *analytic approach* to *problem solving*, in any discipline.

I am confronted regularly with "fake analysts", who typically do not realize they are posing as analysts in title only. If an employer needs an analyst, then they should seek out the real ones, and save those who are not for other positions.

I have often written about data scientists and what actions and skills they manifest most. However, not all Data Scientist are analysts--they may not be engaged in problem solving. At any rate, it should be clear that Analytics is strongly correlated to the problem solving process.

So, who are the real analysts, the ones we may be looking to hire, who approach problem-solving by disaggregating problems into their smaller parts? These are people who have specific training and experience, either qualitative or quantitative, in the science and logic of a field that many are calling Analytics.

Analytics is often used synonymously with quantitative modeling, like statistical modeling, mathematical programming, simulation, and so on. However, qualitative modeling such as text analysis, is just as important in Analytics as its quantitative cousin.

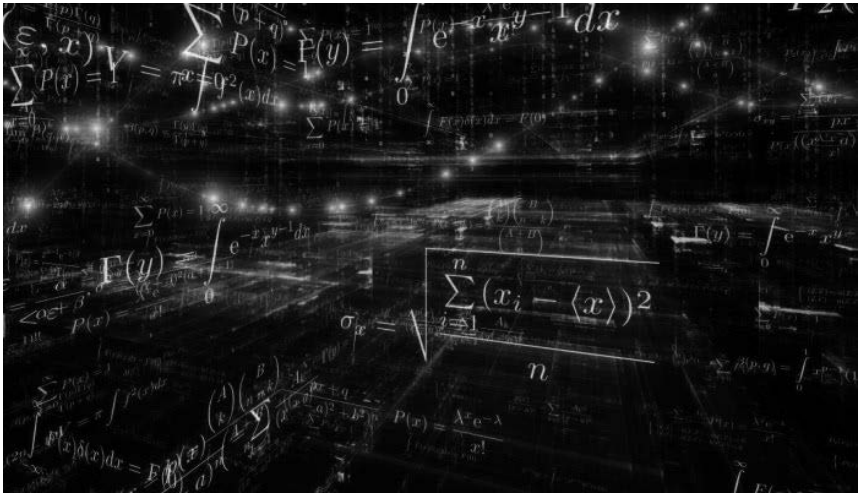
There does exist, however, the situation where the intuitive problem solvers call themselves analyst, and this may not necessarily be the case. There is a need for this, but not in Analytics.

On the other hand, some who are statisticians or mathematicians call themselves analysts when "real problem solving" is not their forte. For instance, describing or characterizing data is important, but it is not Analytics if it is not performed as part of the problem solving process.

You could face a situation where you have a sociologist and a statistician both claiming to be analysts. Suppose the latter routinely characterizes data and then passes that information to someone else for problem solving. We could argue that the statistician is not the analysts.

With the sociologist, we might be confronted with someone who takes social problems, say unemployment, breaks them into their component parts, and then solves each piece though non-quantitative means. We could argue that this person is an analyst.

Before you start throwing analyst or analysis on your resume, ask yourself if you are an integral part of problem solving possess using the analytic method. If you are not, do not call yourself an analyst, rather describe what you actually do.



All Things Data

Data Scientist, Analytic Scientist, Statistician, Operation Research Analyst, Predictive Modeler, … They are all very different… or are they?

We often focus on differences to delineate our profession and what we can bring to the analytical table to feast upon. Yes there are differences, but let's try something novel. Let's focus on the similarities. What do we all have in common? Data!

Is it a surprise that the Administration has appointed a Chief Data Scientist (congratulation Dr. Patil), or is it more surprising that it took so long? Everyone else has a chief data scientist or a "chief someone who uses data to do good stuff with". After all, it doesn't really matter that you have a bunch of data, or that you have a fancy title. What matters is whether you can turn that data into useful information that is used and results in making the business, the industry, the government and/or the country better off tomorrow than it was yesterday.

Ask not what your data can do for you; ask what you can do with your data. After all, to turn that data into useful information, you have to scientisize it, or statisticize it, or researchacize it, or modelize it (all of

these are new words now). You might even have to analyze it, or synthesize it, perform some other –izes upon it. And is that not what statisticians, modelers, analysts, scientist who work with data do from day to day?

So, instead of "how to become a data scientist" and "statisticians are not data scientists", why not "we are all a bunch of professionals trying to turn tons of data into useful, applicable and actionable information and this is how you can join us"?

I do not label myself as a data scientist, but as an analytic scientist. But day-to-day I live in data, trying to make sense of it, trying to manipulate it, cleanse it, impute it, transform it and model it. Is the statistician doing anything much different? Or is the operations researcher, or the data scientist? No, we are all Hadooping, and SASing, and R-ing, and Tableauing, and Pythoning, and other –ings every day to a boatload (a very big boat like Royal Caribbean's Oasis of the Seas) of data. I work with all of these disciplines every day and I see us all doing the same stuff, maybe from a different angle. And who's to say what the best angle is, except all of them combined. After all, it does take eight 45% angles to make a circle, at least in this solar system.

Like Alan Turing, who had to get his machine to determine millions of different combinations of Enigma settings to help in the fight against Nazi Germany, so we too have to sift through millions of combinations of variables to find useful information that will aid in our fight against stupidity. That's right, we have been stupid for too long. Regardless of what CNN or FOX News tells you, we are not doing just fine. Look around your neighborhood, or mine. We need to set aside our differences and join our collective intellects to make this enormous amount of data that is available to us, work for us and make us better tomorrow than we are today.

If you are expecting our new Chief Data Scientist to solve all of our ills, don't hold your breath. The Department of Education didn't work, other than give jobs to people who have very little talent for getting things done; the House and the Senate isn't doing anything but fighting a standoff and keeping the status quo; the IRS isn't doing anything other than than taking your hard earned dollars and redistributing them. Let's just fire them all and give Dr. Patil an army of "professionals who do

great stuff with data." Well, that's not going to happen, but still the answer may be in All Things Data.

Python plus R equals Data Scientist?

I have read a number of articles and posts that claim that if you had courses in linear algebra and multivariate calculus, and know *R* and *Python*, then you are a data scientist. That sounds really good. I have courses in embryology, histology, and physiology, so I must be an OBGYN doctor—I also stayed in a *Holiday Inn Express* last night.

Other things I have read indicate that data scientists are carrying a blemish which causes them not to be needed. No wonder! It appears as if I give $10 to Vincent 'Vinnie' Antonelli [Steve Martin in My Blue Heaven], then he can make me a data scientist. I have talked to recruiters who have been disappointed in the actual qualifications of some who label themselves a data scientist.

Are these people real? Two college courses plus *R* and *Python*? Although my view may not be popular, a monkey can learn to use tools. Tools do not make you anything other than a tool user. A deeper understanding of data structure, data analysis, data modeling, data … is what one needs

to be Data Scientist. Conceptual understanding, not tools. If you have the former, you can pick up the latter in no time.

However, I will go on to say that conceptual understanding is not enough. When I see the data scientist depicted in cartoons or on TV, I see a reclusive geek with no social skills and a propensity to immerse themselves in data at the expense of all else. But that is not a data scientist.

The term “scientist” refers to someone who is an expert in their science and uses the scientific method. As a profession the scientist of today is widely recognized. Scientists include theoreticians who mainly develop new models to explain existing data and predict new results, and experimentalists who mainly test models by making measurements — though in practice the division between these activities is not clear-cut, and many scientists perform both tasks.

Then what makes a data scientist? I would say that coursework leading to a degree, conceptual understanding, a sprinkle of research, a dash of new development, an occasional bath, plus “effectiveness” makes a data scientist. When a client uses the product produced by the data scientist, whether it be a database, a data architecture, a data model, or data analysis, then we have “effectiveness”, and they person who did all of this is a data scientist.

If you are a pilot of a Boeing 757 and you cannot land the aircraft (it may be a skill you never acquired), then you are not a pilot. You are just an airline employee with wings on your chest. Landing the plane is pretty important, and “landing” your solution with the big data customer is pretty important. When you talk as a data scientist, then clients should listen and take action based on what you tell them.

This is where we are missing the boat. We can ‘python-ate’ or ‘statistic-ate’ [two new verbs] all day long, but at the end of the day when the client is no longer listening, then what are we?

Why your client might not be listening

What is the Problem?

To effectively communicate, we must realize that we are all different in the way we perceive the world and use this understanding as a guide to our communication with others.

—Anthony Robbins

Have you ever worked your fingers to the bone developing a model or other analytic solution to a problem for your client or department, and then observed that your solution was never used? Is the boss or customer listening to your advice? If the answers are Yes and No, respectively, then there is a problem, but it is a common problem. There a number of root causes for this, the the most common is communication.

Data Analyst Jargon

As big data analysts, we have our own language. This is sometimes called jargon. We understand it, but other do not. Others may include

managers, hiring managers, intelligence analysts, chief technology officer, etc. When we talk they are probably nodding their head and do not have a clue what we are saying.

The two words 'information' and 'communication' are often used interchangeably, but they signify quite different things. Information is giving out; communication is getting through.

—Sydney J. Harris

Client Interaction

There is a group on YouTube called Studio C (If you have not watched them before, you are missing out on some good comedy). I watched one yesterday called Poke Face, https://www.youtube.com/watch?v=XQ6_GdODuww. Each player is showing their poker face and we get to hear each one of their thoughts. One guy does not know the game at all—not the rules, the poker chips, etc.—and when it is time to show his hand, he is showing UNO cards. However, he kept his poker face the whole time. This is your client. Just because you are an outstanding analyst, providing excellent analytic solutions, does not guarantee that people will listen to you. You have to speak their language.

Language is Important

Don't use words too big for the subject. Don't say infinitely when you mean very; otherwise you'll have no word left when you want to talk about something really infinite.

— C.S. Lewis

There was a certain Christian missionary who was trying to reach a certain native tribe in a remote part of the world with his message. The tribe killed him. Later, the tribe was converted to Christianity, and it was not until then they understood what they had done was wrong. There had been a communication gap of several layers. One was language, one was culture, and another was belief system. The tribal chief is your client.

You have to speak his language, understand his culture, and recognize his belief system.

A Real Example - Sort of

In the land of Gibberish, the man who makes sense, the man who speaks clearly, clearly speaks nonsense."

— Jarod Kintz, This Book Has No Title

I said this is a real example, but it is really a real example with the client using their jargon with the analyst, rather than the other way around. Recently, a colleague asked for some advice for an upcoming interview. The telephone interview was with the CTO and this is what was taken away from that interview. In answering what kind of projects would be expected, the CTO said that the following could be expected

1. Message testing
2. Spent-time model of customers that are at risk and at grow
3. Atomization of models/algorithms/etc.

He is what I draw out of the CTO's description:

1. Message testing probably involve constructing an acquisition model, like a propensity to buy or engage. These models are typically logistic regression models. Once you build a model, say in SAS, you can deploy it and make operational on a main frame for example. The the client could run the model on a regular basis. The output would be a score of customers' propensity to buy or engage. Using the score file, they would market/message the top deciles. This way they are spending marketing dollars on those whom the message would have the most likelihood of a response.
2. Customers at risk is also a propensity model: the propensity to shed. This is also a logistic regression model and the same principles from above apply, except the scores would be high for people who are at risk.
3. Automation could take a number forms, but the most common is model deployments as I described above. For example, when

I build a model in SAS Enterprise Miner, I take the score code and wrap it in an execution macro and then give that code to a guy who deploys it for production on a mainframe. The same guy runs the model each month along with about 20 other models during the first week of the month. The process is automated and the scores are sent to a marketing directory for access. He also runs a model performance report on each model, which is partially automated. An emerging process is real-time modeling. The same models would be used, but they might be written in Python in a Hadoop environment and the data would be refresh a regular short time periods. This could be used to direct traffic to call center. For example, if a caller has a high score for propensity to shed a mortgage, that customer could be directed to the mortgage company and provided some incentive not to shed their mortgage.

Conclusion

The ability to simplify means to eliminate the unnecessary so that the necessary may speak.

—Hans Hofmann

The CTO is speaking a different language. To understand the CTO's language you must first understand his culture, belief system and context. Not knowing these, I am guessing at my interpretation above. However, if you interview for a job you should understand these things first. The same is true for our communication with clients. We must understand their culture, belief system—their corporate belief system, not religious—the relative context, and then their language they understand. If we do not do these things first, our analysis results land on deaf ears and we fail. Our solutions are neither understood nor implemented.

Part VI - Applications

Big Data Analytics and Human Resources?

The ability to capture and analyze big data is recognized by many enterprises as an accelerator. It has enabled many enterprises to increase revenues by better understanding, more accurately target customers and cut costs through improved business processes.

Big data also has attracted the attention of enterprise managers and their human resource (HR) managers. Many believe that they can now analyze mountains of structured and unstructured data to answer important questions regarding workforce productivity, the impact of training programs on enterprise performance, predictors of workforce attrition, and how to identify potential leaders. For the purpose of this article, HR will refer to the enterprise function and human resource will refer to the humans who are resources for the enterprise.

Hogwash!

The HUMAN in HR barely exists as it is. Now I hear that HR is going to identify leaders through big data analytics (I am trying not to laugh too hard). You would think that one would need to know what defines a leader, but with all the "junk" articles I read on leadership, most people including HR do not know what leadership is! Many use management and leadership synonymously, and they are mistaken.

What is Analytics anyway?

Analytics is slightly better than an educated guess and much better than a random choice. I am an analytics scientist, making a living from performing analytics, but I declare that is not a "fool-proof" science. Good big data analytics is based on the quality data. However, big data does not imply good data. In fact, most of the so-called big data I have worked with is very poor data. Quantity does not equal quality, so stop saying that big data is the solution to every problem that arises, including HR problems.

Root Causes.

I am not picking on HR. This is a common problem across disciplines, and I know many competent, caring and professional HR personnel. Having stated that, I often use a quote from Kurt Vonnegut, author of Player Piano:

"Almost nobody's competent, Paul. It's enough to make you cry to see how bad most people are at their jobs. If you can do a half-assed job of anything, you're a one-eyed man in the kingdom of the blind."

Analytics does not make incompetent people more competent—incompetent people performing analytics produces "garbage"! If you want to improve HR, then put the HUMAN back into HR—transform HR personnel into those competent, caring and professional ones I know and work with.

Analytics also does not make bad products better, poor services better, and so on. It may provide insight into why products and services are perceived as being poor, but the only way to fix these problems is to "make a better product" and "provide better services".

Where did the Human go?

I remember a time when HR was different than it is today. I have not performed analytics on HR, so I would be guessing as to what happened to it, how it became less human. So, for now I will guess. Human resources implies managing humans who are resources for your enterprise. There was also a time when enterprises (in the US) where flourishing, having fewer constraints in terms of capital resources, regulations, and so on. Somewhere around 2007 that changed. Human resources (the resources, not HR) became surplus in an over-saturated market. Instead of struggling to find good talent to fill a growing enterprise, HR seems to have been forced into constraining resources. What was already in the pipeline was a resource, and everyone else was a potential element that would break the pipeline. So, guarding the enterprise became more important than managing its resources, and HR became more involved in the decision making of enterprises than should have been allowed. The service-oriented HR personnel became constrainers, and the HUMAN disappeared. Now we treat human resources (the humans who are the resource) as dollars or paper and pencils.

Now, with the perception that HR is "broken" or that something in the enterprise related to human resources is diminished, we aim to fix it with analytics… Perhaps we should find the HUMAN and put it back in HR first.

What can Analytics do?

I have mentioned what analytics cannot do. Here are some ways that analytics could help the enterprise and its HR functions:

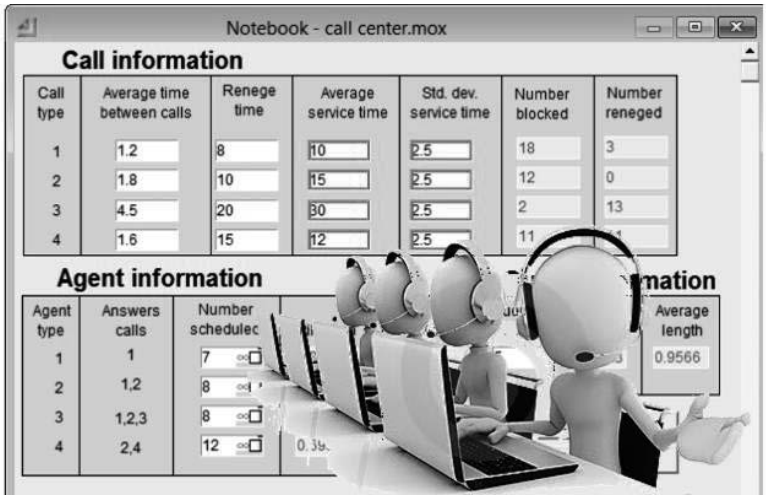
- Recruitment cost per hire
- New hire failure factors
- Employee turnaround rates
- E-learning abandonment rates

- Bonus compensation rates

In spite of the fact that I have been downplaying the need for analytics in HR, when we know these things kinds of things, we can actually put the HUMAN back in HR. I know, it is a paradox. The point is do not use big data analytics to fix HR, rather, use it to enhance HR functionality. If an enterprise feels it needs to fix its HR, then it should fix its HR before adding big data analytics to the sauce. If HR is broken, adding analytics is just going to produce more of the same sauce, and you will not like its flavor.

Conclusion

HR is an important function for the enterprise, unless you replace humans completely with machines, robots and androids. Improving the enterprise's HR function is a worthy undertaking and big data analytics may help improve its functionality, but only if it is already functional!



Call Center Analytics: What's Missing?

What is the Problem?

My customer frequently asks if this model or that model can be used to direct call center traffic. Usually the models are projecting an acquisition or engagement window with too much of a gap to perform this function, and this is hardly part of their use case as acquisition models. However, I can perceive a model that does perform this function. Yet a model is not entirely what they need.

Models seem to be the magic “fix-all” in the Financial Services and Insurance Industry (FSI). Customers think that models will fix their product, marketing strategy and call center. The reality is: anything that pertains to analytics seems to be a model to them. That is an article for another day. For this one, I want to talk about the call center problem, namely, directing calls more efficiently.

A model might be the answer they are looking for. However, you cannot just build a model and leap into its use blindly, without testing. The best way to test in this situation is using discrete event simulation.

What is Discrete Event Simulation?

Discrete Event Simulation (**DES**) is the process of codifying the behavior of a complex system as an ordered sequence of well-defined events. In this context, an event comprises a specific change in the system's state at a specific point in time (arrival at the bank, service by a teller, etc.). Rather than stepping based on a time increment, like every second, DES advances based on events—events that may or may not be equally spaced in time.

Common applications of DES include:

- stress testing (a form of deliberately intense or thorough testing used to determine the stability of a given system or entity)
- evaluating potential financial investments
- studying call center operations
- modeling procedures and processes in various industries, such as manufacturing and healthcare
- studies that support system design
- studying reliability, availability and maintainability (RAM)
- just about anything that involves queuing (arrivals, waiting time, service time)

Can We Simulate the Call Center?

One of the most applicable situations that lends itself to DES is the operations of a call center. When a customer calls, you have an arrival event. If they have to wait for a representative they enter a queue until such time as their call is answered, which begins a service event. If their call was rerouted upon entering the system–the initial arrival event–the rerouting creates a new arrival event for the call center to which it was transferred, and so on. One can study where bottle-necks occur, rerouting efficiency, waiting times, balking (customers deciding not to join the queue if it is too long), reneging (customers leave the queue if they have waited too long for service), etc.

An effective DES process must include, at a minimum, the following characteristics:

- Predetermined starting and ending points, which can be discrete events or instants in time (arrivals and departures, for instance).
- A method of keeping track of the time that has elapsed since the process began (waiting time, for instance).
- A list of discrete events that have occurred since the process began (begin service, for instance).
- A list of discrete events pending or expected (if such events are known) until the process is expected to end.
- A graphical, statistical, or tabular record of the function for which DES is currently engaged (plot of waiting times and service times, for instance).

Using DES, we can test the model with multiple parameter, or test multiple models, without disrupting the operations of a call center. Once we determine which model (or which parameters) work the best, we can then use that particular model to direct call, for instance. Continued monitoring and used of the DES for simultaneous parallel testing can be used to maintain and improve efficiency.

Win, Lose or Draw…Is this a Game?

What economists call ***game theory*** psychologists call the ***theory of social situations***, which is an accurate description of what *game theory* is about. Although game theory is relevant to parlor games such as poker or bridge, most research in game theory focuses on how groups of people interact. There are two main branches of *game theory*: ***cooperative*** and ***non-cooperative*** game theory. Non-cooperative game theory deals largely with how intelligent individuals interact with one another in an effort to achieve their own goals. That is the branch of game theory I will discuss here.

In addition to *game theory*, economic theory has three other main branches: ***decision theory***, ***general equilibrium theory*** and ***mechanism design th****eory*. All are closely connected to game theory.

Decision Theory

Decision theory can be viewed as a theory of one person games, or a game of a single player against nature. The focus is on preferences and the formation of beliefs. The most widely used form of decision theory argues that preferences among risky alternatives can be described by

the maximization of the expected value of a numerical utility function, where utility may depend on a number of things, but in situations of interest to economists often depends on money income. Probability theory is heavily used in order to represent the uncertainty of outcomes, and Bayes Law is frequently used to model the way in which new information is used to revise beliefs. Decision theory is often used in the form of decision analysis, which shows how best to acquire information before making a decision.

General Equilibrium Theory

General equilibrium theory can be viewed as a specialized branch of game theory that deals with trade and production, and typically with a relatively large number of individual consumers and producers. It is widely used in the macroeconomic analysis of broad-based economic policies such as monetary or tax policy, in finance to analyze stock markets, to study interest and exchange rates and other prices. In recent years, political economy has emerged as a combination of general equilibrium theory and game theory in which the private sector of the economy is modeled by general equilibrium theory, while voting behavior and the incentive of governments is analyzed using game theory. Issues studied include tax policy, trade policy, and the role of international trade agreements such as the European Union.

Mechanism Design Theory

Mechanism design theory differs from game theory in that game theory takes the rules of the game as given, while mechanism design theory asks about the consequences of different types of rules. Naturally this relies heavily on game theory. Questions addressed by mechanism design theory include the design of compensation and wage agreements that effectively spread risk while maintaining incentives, and the design of auctions to maximize revenue, or achieve other goals.

Example

One way to describe a game is by listing the players (or individuals) participating in the game, and for each player, listing the alternative choices (called actions or strategies) available to that player. In the case of a two-player game, the actions of the first player from the rows, and the actions of the second player the columns, of a matrix. The entries in the matrix are two numbers representing the utility or payoff to the first and second player respectively.

The Prisoner's Dilemma (PD): The problem can be traced back to von Neumann and Morgenstern [von Neumann, 1944] and, of course, John Nash [Nash, 1953], you and a friend have committed a crime and have been caught. You are being held in separate cells so that you cannot communicate with each other. You are both offered a deal by the police and you have to decide what to do independently. Essentially the deal is this.

- If you confess and your partner denies taking part in the crime, you go free and your partner goes to prison for ten years.
- If your partner confesses and you deny participating in the crime, you go to prison for ten years and your partner goes free.
- If you both confess you will serve six years each.
- If you both deny taking part in the crime, you both go to prison for six months.

What will you do? The game can be represented by the following matrix of payoffs

		Friend	
		not confess	confess
You	not confess	5,5	-4,10
	confess	10,-4	1,1

Note that higher numbers are better (more utility). If neither of you confesses, you both go free, and split the proceeds of their crime which we represent by 5 units of utility for each suspect. However, if one of you confesses and the other does not, the one who confesses testifies

against the other in exchange for going free and gets the entire 10 units of utility, while the one who did not confess goes to prison and which results in the low utility of -4. If both of you confess, then both are given a reduced term, but both are convicted, which we represent by giving each 1 unit of utility: this is called ***mutual defection***.

Mutual defection is a ***Nash equilibrium*** of PD [it is a crucial concept in non-cooperative games, and it won Nash the Nobel Prize in economics in 1994]. Informally, a set of strategies is a Nash equilibrium if no player can do better by unilaterally changing his or her strategy. Mutual defection is the unique Nash equilibrium of PD, which means that it is the only stable solution to this game. In real world scenarios, however, a Nash equilibrium is not necessarily played. Some conditions to guarantee that the Nash equilibrium is played are:

1. The players aim to maximize their own payoffs.
2. The players know the Nash equilibrium strategy of all players.
3. The players believe that a deviation in their own strategy will not cause deviations by any other players.
4. There is *common knowledge* that all players know these conditions.

This game has fascinated game theorists for a variety of reasons. First, it is a simple representation of a variety of important situations. For example, instead of confess/not confess we could label the strategies "contribute to the common good" or "behave selfishly." This captures a variety of situations economists describe as public goods problems. An example is the construction of a bridge. It is best for everyone if the bridge is built, but best for each individual if someone else builds the bridge. This is sometimes referred to in economics as an externality. Similarly this game could describe the alternative of two firms competing in the same market, and instead of confess/not confess we could label the strategies "set a high price" and "set a low price." Naturally it is best for both firms if they both set high prices, but best for each individual firm to set a low price while the opposition sets a high price.

A second feature of this game, is that it is self-evident how an intelligent individual should behave. No matter what a suspect believes his partner is going to do, it is always best to confess. If the partner in the other cell

is not confessing, it is possible to get 10 instead of 5. If the partner in the other cell is confessing, it is possible to get 1 instead of -4. Yet the pursuit of individually sensible behavior results in each player getting only 1 unit of utility, much less than the 5 units each that they would get if neither confessed. This conflict between the pursuit of individual goals and the common good is at the heart of many game theoretic problems.

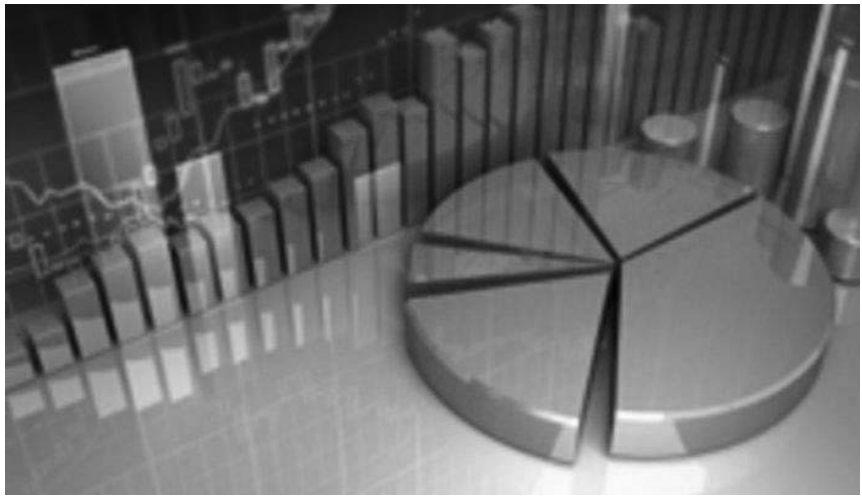
A third feature of this game is that it changes in a very significant way if the game is repeated, or if the players will interact with each other again in the future. Suppose for example that after this game is over, and the suspects either are freed or are released from jail they will commit another crime and the game will be played again. In this case in the first period the suspects may reason that they should not confess because if they do not their partner will not confess in the second game. Strictly speaking, this conclusion is not valid, since in the second game both suspects will confess no matter what happened in the first game. However, repetition opens up the possibility of being rewarded or punished in the future for current behavior, and game theorists have provided a number of theories to explain the obvious intuition that if the game is repeated often enough, the suspects ought to cooperate.

References

Nash J. Two-Person Cooperative Games. *Econometrica* 21: 128-140, 1953.

von Neumann J., Morgenstern O. *Theory of Games and Economic Behavior*. Princeton NJ: Princeton Univ. Press, 1944.

Part VII – The Power of Operations Research

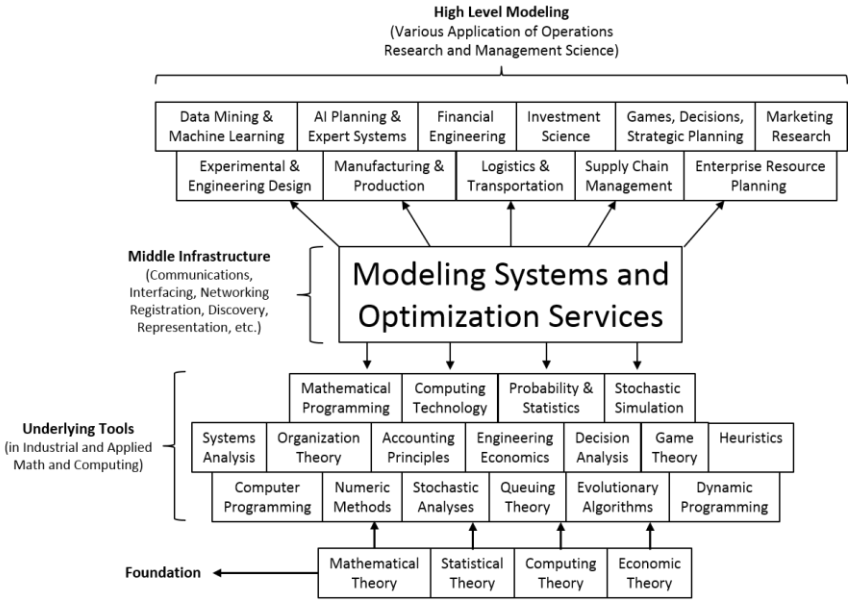


What is Operations Research?

Many people probably never heard the phrase "*Operations Research*" used. Operations Research (OR), or operational research in the U.K, is a discipline that deals with the application of advanced analytical methods to help make better decisions. The terms management science and analytics are sometimes used as synonyms for operations research. Yet, in my experience OR extends far beyond either. The figure shows a hierarchy of operations research activities, and I'll let you decide if they are also performed in analytics.

- Data Mining and Machine Learning
- Artificial Intelligence and Expert Systems
- Financial Engineering
- Games, Decision, and Strategic Planning
- Marketing Research
- Investment Science

- Experimental and Engineering Design
- Manufacturing and Production
- Logistics and Transportation
- Supply Chain Management
- Enterprise Resource Planning



Modeling Systems and Optimization Services is an interface part that bridges OR modeling with OR tools. When implemented smoothly, it is the part that is not noticed by modelers or users.

Where do you find them?

Operation Research Analysts work in many industries, including maritime, space operations, defense, airlines, train lines, financial service, entertainment and many more. Wherever operation occur, operations research analysts are usually there. The following lists several key functional areas for operations research analysts.

- Communications
- Interfaces
- Networks

- Scheduling
- Routing
- Manpower
- Modeling

What are their tools?

Underlying Tools is the level that is typically regarded as what uniquely defines Operations Research.

- Mathematical Programming
- Computing Technology
- Probability and Statistics
- Stochastic Simulation
- Systems Analysis
- Organization Theory
- Accounting Principles
- Engineering Economics
- Decision Analysis
- Game Theory
- Heuristics
- Computer Programming
- Numeric Methods
- Stochastic Analysis
- Queuing Theory
- Evolutionary Algorithms
- Dynamic Programming

What are they built upon?

Foundations upon which OR are built include:

- Mathematical Theory
- Statistical Theory
- Computing Theory
- Economic Theory

What is their history?

Operational Research was born during the early year of WWII and matured rapidly. One of its primary function was the planning of Operation Overlord or the Normandy Invasion. It has its foundations in mathematics, computing and economic theories, on which basic tools in optimization and simulation are built. Today OR's are employed by airlines, train lines, logistic systems, delivery systems (e.g., FedEx), defense systems, military, oil companies, insurance companies, financial institutions, manufacturing, marketing and many more.

Getting the Question Right

As an analysts and modeler I have a variety of customers with a variety of problems in business. We usually frame their “question” as a business case. The question of course is something they require an answer for it usually stems from a problem or a desire to have more Share of Wallet or other metric. What is interesting here is though they know they have a question that needs to be answered, they often do not know how to state their question.

In "Making Analysis Relevant", Mr. Vince Roske reiterated a recipe provided by John D. (Dave) Robinson, MG USA, Ret, when he was the Director of Joint Staffs J-8. The recipe is provided below:

- What’s the question?
- What’s the “real” question?
- What do the final slides look like?
- What do I already know?
- How do I get the remaining information that I need?

So, this idea of what is the real question is not new and it is not unique to business.

As an analyst, we have to draw that real question out through dialogue with our customers. Doing this is often as much an art as it is a science. If we do not do it, we stand a good chance delivering a well-built solution that happens to be the wrong one. I have done this at least twice in the last three years, knowing that getting the question right was paramount. In spite of my failures I have found some keys to help with formulating the business case.

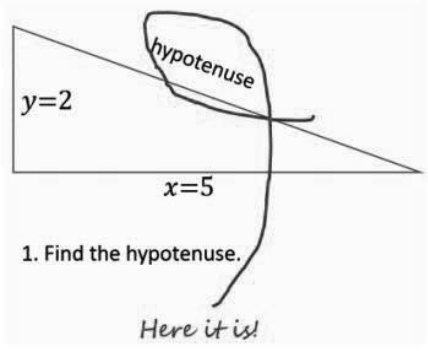
1. **Tie it to a metric.** Even if currency is not at stake, you have to have a metric to focus on. Otherwise you can fail to see the forest for the trees. As your project rolls along you will forget what you are building to without a target. It can be Return on Investment, Share of Wallet, Key Performance Parameter, or something else, but you need a metric.
2. **Do not mention the solution.** Do not state your analytic solution method in the business case. The customer may want a model and a model may be the right analytic solution for answering the question, but most customers do not understand what models do and don't do. The business case should focus on the question/problem, not the solution. If you allow it to, the customer will start building your model for you, without the expertise to do so. I have seen this happen.
3. **Use timing.** Guide the customer to include timing in their business case, i.e., achieve metric A with X months, detect behavior B at least Y months in advance, and so on. With a timing "device" you can help the customer realize the temporal nature of an analytic solution. On one hand, the solutions do not work forever, but many believe otherwise. On the other hand, solutions apply only to events and behaviors in a specified time period. Behavior farther out from the phenomenon or event are quite different from behavior closer to it.
4. **Tie it to the business unit.** That might sound like a no-brainer, but I have seen business units responsible for web activity or call center activity want to tie the business case to a product. If you are trying to improve call center operations, you are trying to improve call center operations, period. Now, you are probably doing so with product sales as a downstream goal, but you cannot put the cart before the horse. Fix your call center problem and then ask a new question.

I am sure there is a lot more that can be said about this subject, but this is what I have learned over time and I am just one analyst. The fact remains that if we do not get the “real question”, then we may not provide the “right solution.”

There are other interesting pieces of Roske’s Recipe I did not address here. Perhaps another article?

Reference

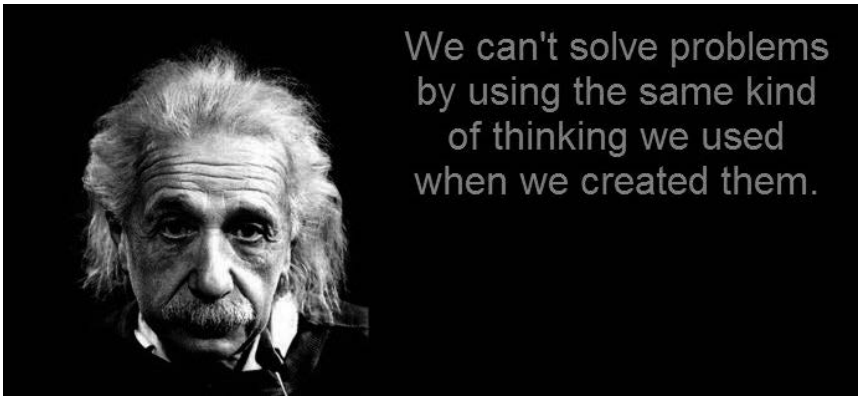
Mr Vincent P. Roske, Jr., “Making Analysis Relevant,” PHALANX Volume 31, Number 1, March 1998



Holistic Analysis and Operations Research

Introduction

I am not really sure what holistic analysis is, so I will define it. Our English word comes from the Greek ὅλος (*holos*, meaning "all, whole or entire"). Reductionism may be viewed as the complement of holism. Reductionism analyzes a complex system by subdividing or reduction to more fundamental parts. For businesses, knowledge and know-how, know-who, know-what and know-why are part of the whole business economics. Having a holistic view keeps us from missing the forest due to the trees.



Operations Research (OR), or **operational research** in the U.K, is a discipline that deals with the application of advanced analytical methods to help make better decisions. The terms management science and analytics are sometimes used as synonyms for operations research. Yet, in my experience OR extends far beyond either. The list below is a collection of operations research activities – I'll let you decide if they are also performed in analytics.

- Data Mining and Machine Learning
- Artificial Intelligence and Expert Systems
- Financial Engineering
- Games, Decision, and Strategic Planning
- Marketing Research
- Investment Science
- Experimental and Engineering Design
- Manufacturing and Production
- Logistics and Transportation
- Supply Chain Management
- Enterprise Resource Planning

History of Operations Research

Operational Research was born during the early year of WWII and matured rapidly. One of its primary functions was the planning of Operation Overlord or the Normandy Invasion. It has its foundations in mathematics, computing and economic theories, on which basic tools in optimization and simulation are built. Today OR's are employed by

airlines, train lines, logistic systems, delivery systems (e.g., FedEx), defense systems, military, oil companies, insurance companies, financial institutions, manufacturing, marketing and many more.

What do Operations Research Analysts do?

The Operations Research Analyst is a jack-of-all-trades (one guy's opinion), or at least that has been my experience. Some tend to specialize in a particular area, like mathematical optimization, but I think this is a mistake. One can find people who specialize in a particular methodology or discipline, but they would be challenged to find a good Operations Research analyst without a holistic view of the problem space.

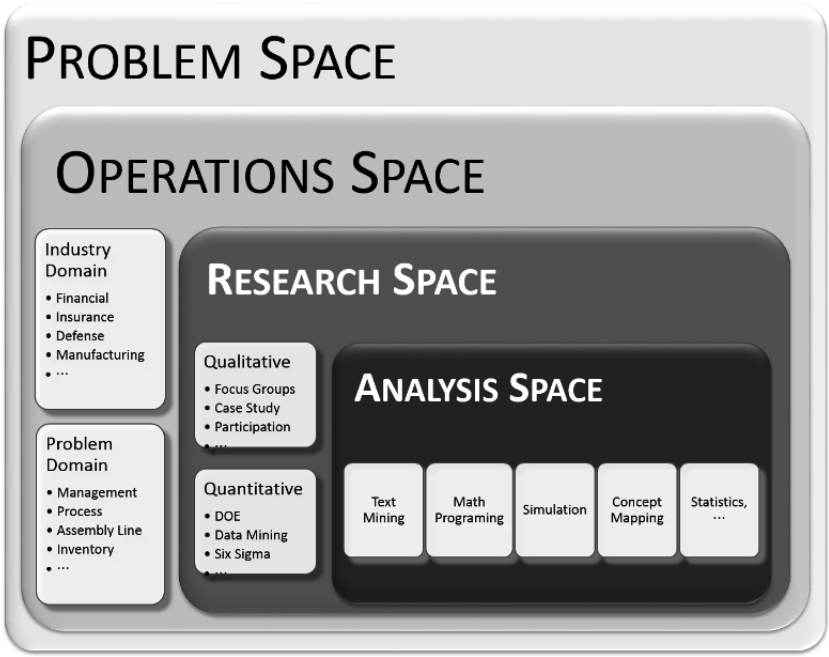


Figure 1. The OR Problem Space (I made this up yesterday)

An OR's view of the problem space is really what defines them and describes what they do. The list above displayed some of the activities that ORs engage in, but not without a holistic view of the problem space. Figure 1 depicts the entire problem space. Mathematically, we could look at it like this:

({(Analysis Space)⊂Research Space}⊂Operations Space)⊂Problem Space

The OR Analyst must enter the problem space with the following in mind: (1) the potential operational domains, (2) the types of research that may be used, and (3) the types of analyses that may be appropriate. If one goes in having done nothing more than math programming for 10 years, that analyst is NOT an operations research analyst—they are just a math programmer.

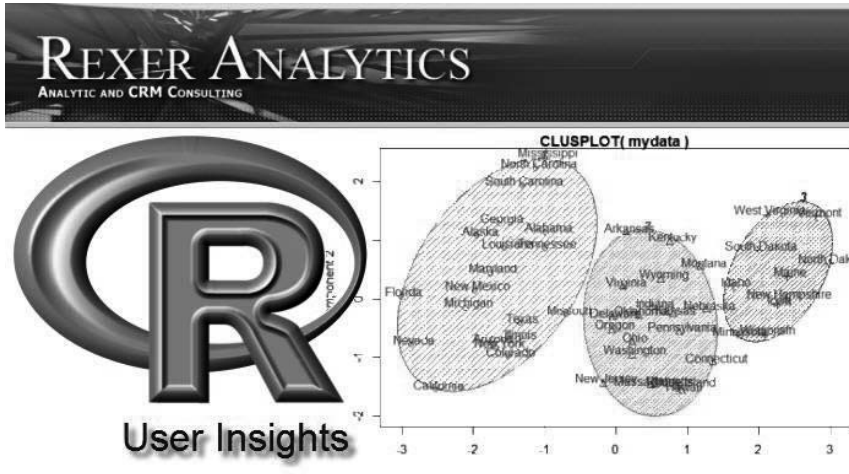
Operations research analysts provide this holistic view, which then allows for the definition of the right problem within any domain, and application of the most appropriate research methodology, using the most appropriate analyses. You cannot build a house with just a screwdriver, unless you are MacGyver[1].

If we look at the historical context of OR, as we discussed earlier, we should be able to ascertain that anything short of a holistic point of view may have resulted in operational chaos much worse than missed dropped zones. Operation Overlord—the most complex operation ever executed—could have easily failed.

Reference

[1] MacGyver is an American action-adventure television series created by Lee David Zlotoff. Henry Winkler and John Rich were the executive producers. The show follows secret agent Angus MacGyver, played by Richard Dean Anderson, who works as a troubleshooter for the fictional Phoenix Foundation in Los Angeles and as an agent for a fictional United States government agency, the Department of External Services (DXS). Resourceful and possessed of an encyclopedic knowledge of the physical sciences, he solves complex problems with everyday materials he finds at hand, along with his ever-present duct tape and Swiss Army knife.

Part VIII - Tools



What are R Users Saying?

Partial Results of the 2011 Survey, Copyright © 2011 Rexer Analytics

The following is a excerpt from 5th Annual Survey (2011) conducted by Rexer Analytics. This is a partial list of insights from R users. The full survey results can be found by following the survey link above. The following is entirely the work of Rexer Analytics. The surveys they conduct contain more categories, such as *Overcoming Data Mining Challenges*, *Analytic Success Measurement* and *The Positive Impact of Data Mining*. I have included the more substantial comments. The details of the most previous years (2012, 2013 and 2014) are not yet available.

"Over the years, respondents to the 2007-2011 Data Miner Surveys have shown increasing use of R. In the 5th Annual Survey (2011) we asked R users to tell us more about their use of R. The question asked, “If you use R, please tell us more about your use of R. For example, tell us why you have chosen to use R, why you use the R interface you identified in the previous question, the pros and cons of R, or tell us how you use R in conjunction with other tools.” 225 R users shared information about

their use of R. They provided an enormous wealth of useful and detailed information. Below are the verbatim comments they shared."

- Best variety of algorithms available, biggest mindshare in online data mining community, free/open source. Previous con of 32-bit in-memory data limitation removed with release of 64-bit. Still suffers some compared to other solutions in handling big data. Stat-ET plugin gives best IDE for R, using widely used Eclipse framework, and is available for free. RapidMiner add-in that supports R is extremely useful, and increasingly used as well.
- The main reason for selecting R, back in the late 1990's, was that it had the algorithms I needed readily available, and it was free of charge. After getting familiar with R, I have never seen a reason to exchange it for anything else. R is open-source, and runs on many different system architectures, so distributing code and results is easy. Compared to some latest commercial software I've evaluated, R is sluggish for certain tasks, and can't handle very large datasets (mainly because I do not have a 64-bit machine to work with). On top of that, to be really productive with R, one needs to learn other languages, e.g., SQL, but that's just how things are. Besides, knowledge of those other languages is needed anyway.
- I've migrated to R as my primary platform. The fact that it's free, robust, comprehensive, extensible, and open. More than offsets its shortcomings – and those are disappearing over time anyway. I'm using Revolution's IDE because it improves my productivity. R has a significant learning curve, but once mastered, it is very elegant. R's shortcomings are its in-memory architecture and immature data manipulation operations (i.e. lack of a built-in SQL engine and inability to handle very large datasets). Nevertheless, I use R for all analytic and modeling tasks once I get my data properly prepared. I usually import flat files in CSV format, but I am now using Revolution's XDF format more frequently.
- Why I use R: Initial adoption was due to the availability of the randomForest package, which was the best option for random forest research, as rated on speed, modification possibility, model information and tweaking. Since adopting R, I have further derived satisfaction from the availability of open source R packages for the newest developments in statistical learning (e.g. conditional inference forests, boosting algorithms, etc.). I use the R command line interface, as well as scripting and function creation, as this gives

me maximum modification ability. I generally stay away from GUIs, as I find them generally restrictive to model development, application and analysis. I use R in conjunction with Matlab, through the statconnDCOM connection software, primarily due to the very slow performance of the TreeBagger algorithm in Matlab. Pros of R: Availability of state-of-the-art algorithms, extensive control of model development, access to model information, relatively easy access from Matlab. Cons of R: Lack of friendly editor (such as Matlab's Mlint; although I am considering TinnR, and have tried Revolution R); less detailed support than Matlab.

- I employ R-Extension in the Rapid Miner interface. I use R for its graphing and data mining capability. I use R for its ability to scrape data from different sources (FTP, ODBC) and implement frequent and automated tasks Pros: -Highly customizable -Great potential for growth and improved efficiencies -Fantastic selection of packages – versatility -Growing number of video tutorials and blogs where users are happy to share their code. CONS: -x64 not fully integrated yet -Steep learning curve -Limited number of coding and output examples in package vignettes -Difficult to setup properly (Rprofile.site examples are scarce online) -Memory constraints with i386 32-bit -Output – reporting design limitations suitable for the business environment.
- I have about 12 years experience with R. It's free, the available libraries are robust and efficient. I mostly work with the command line, but I am moving towards R-Studio because it's available both as a desktop application and a browser-based client-server tool set. I occasionally use Rcmdr. The only other tool I use as heavily as R is Perl – again, lots of experience with it, it's free, and there are thousands of available library packages.
- The analytics department at my company is very new. Hence we haven't yet decided which analytics-tool we'll be using. R is a great interim solution as it is free and compatible with a reasonable amount of commercial analytics tools. The reason I use the standard R GUI and script editor is simply because I haven't invested the time in trying out different GUI's, and I haven't really had any reason to. The advantage of using R (at least compared to SAS, which was the analytics-tool at my old job) is mainly that you have much more control over your data and your algorithms. The main problem with R is that it can't really handle the data sets we need to analyze.

Furthermore, your scripts can tend to be a bit messy if you are not sure what kind of analysis or models you are going to use.

- I use R for the diversity of its algorithms, packages. I use Emacs for other tasks and it's a natural to use it to run R, and Splus for that matter. I usually do data preparation in Splus, if the technique I want to use is available in Splus I will do all the analysis in Splus. Otherwise I'll export the data to R, do the analysis in R, export results to Splus where I'll prepare tables and graphs for presentations of the model(s). The main drawback to R, in my opinion, is that R loads in live memory all the work space it is linked to which is a big waste of time and memory and makes it difficult to use R in a multi-users environment where typical projects consist of several very large data sets.
- We continue to evaluate R. As yet it doesn't offer the ease of use and ability to deploy models that are required for use by our internationally distributed modeling team. "System" maintenance of R is too high a requirement at the moment and the enormous flexibility and range of tools it offers is offset by data handling limitations (on 32 bit systems) and difficulty of standardizing quick deployment solutions into our environment. But we expect to continue evaluation and training on R and other open source tools. We do, for instance, make extensive use of open source ETL tools.
- I use R extensively for a variety of tasks and I find the R GUI the most flexible way to use it. On occasion I've used JGR and Deducer, but I've generally found it more convenient to use the GUI. R's strengths are its support network and the range of packages available for it and its weaknesses are its ability to handle very large datasets and, on occasion, its speed. More recently, with large or broad datasets I've been using tools such as Tiberius or Eureqa to identify important variables and then building models in based on the identified variables.
- I'm using R in order to know why R is "buzzing" in analytical areas and to discover some new algorithms. R has many problems with big data, and I don't really believe that Revolution can effectively support that. R language is not mature for production, but really efficient for research: for my personal researches, I also use SAS/IML programming (which is for me the real equivalent for R, not SAS/STAT). I'm not against R, it's a perfect tool to learn statistics, but I'm not really for data mining: don't forget that many techniques

used in data mining comes from Operational Research, in convergence with statistics. Good language, but not really conceived for professional efficiency.

- R is used for the whole data loading process (importing, cleaning, profiling, data preparation), the model building as well as for creating graphical results through other technologies like Python. We use it also using the PL/R procedural language to do in-database analytics & plotting.
- Utilize R heavily for survey research sampling and analysis and political data mining. The R TextMate bundle is fantastic although RStudio is quickly becoming a favorite as well. Use heavily in conjunction with MySQL databases.
- I use R in conjunction with Matlab mostly, programming my personalized algorithms in Matlab and using R for running statistical test, ROC curves, and other simple statistical models. I do this since I feel more comfortable with this setting. R is a very good tool for statistical analysis basically because there are many packages covering most of statistical activities, but still I find Matlab more easy to code in.
- I greatly prefer to use R and do use it when working on more "research" type projects versus models that will be reviewed and put into production. Because of the type of work that I do, SAS is the main software that everyone is familiar with and the most popular one that has a strong license. Our organization needs to be able to share across departments and with vendors and governmental divisions. We can't use R as much as I would like because of how open the software is – good for sharing code, bad for ensuring regulators that data is safe.
- Main reasons for use: 1) Strong and flexible programming language, making it very flexible. 2) No cost, allowing me to also having it on my personal computer so that I can test things at home that I later use at work. I use RODBC to get the data from server and let the server do some of the data manipulation, but control it from within R. Have also started to use RExcel with the goal as using that as a method to deploy models to analysts more familiar with Excel than R.
- Personally I find R easier to use than SAS, mostly because I am not constrained in getting where I want to go. SAS has a canned approach. I see using GUI's as a "sign of weakness" and as preventing understanding the language at its core. I have not found

Rattle to be particularly helpful. I have also tried JGR and Sciviews and found I could not surmount the installation learning curve. Their documentation did not produce a working environment for me.

Why You Might Use SAS

I write a good bit of content about using open-source tools for analytics and operations research. However, my workhorse happens to be SAS. Actually, I use SAS Enterprise Guide (EG) and SAS Enterprise Miner (EM).

When to use open-source

I have argued for the use of open-source tools, but they have their place. If you have a limited budget, open-source is a good path to journey down. Also, if you are teaching or are a student at a university, open-source seems like a logical option.

How I use SAS

I perform predictive modeling in the Financial and Insurance (FSI) industry. My method requires me to use SAS EG to retrieve variables from multiple data sets (between 10 and 20 sometimes), resulting in about 2000 variables. Once these variables are merged into one data set (for up to 11 million customers), I run an information value algorithm to determine which variables have the most predictive power for the response variable. After eliminating variables that cannot be used for marketing bank products (fair lending acts and so on), I import between 150 and 300 variables into SAS EM. In EM, I perform data partitioning,

data imputations, and data transformations prior to running a model, like logistic regression. When I get an adequate model, I take the scoring code from EM and port it to EG, wrapped in a macro. I run the macro in EG to measure model performance. When I get a model that performs well enough against challenger models, I develop model production code in EG, which includes the EM scoring code.

Why I use SAS

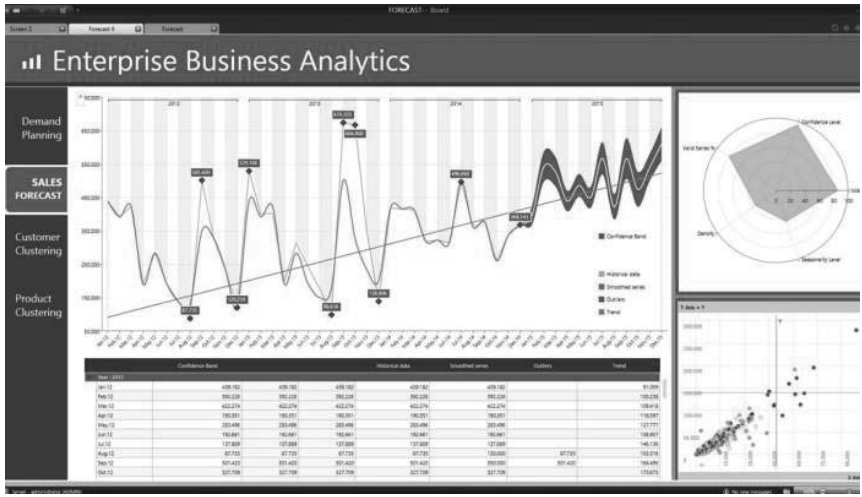
SAS EG and EM are trusted by the FSI that I service. I actually began using SAS in 1990 on a mainframe computer. It has been the statistical software of choice by many industries for many years. With external model governance and validation, the models I have developed in SAS EG and EM have been stress tested and proven valid for use by the FSI. This does not imply that models created with open-source tools are any less valid, as I will discuss next.

Open-source and SAS

Once I build a predictive model in SAS, I attempt the build the same model in R, for example. The models in R are similar enough to show that the models in SAS are indeed valid models. For instance, I constructed an uplift model in SAS using logistic regression and an uplift model for the same acquisition situation using random forest in R. The overall net lift was identical, although the distribution among deciles was slightly different. Anyone can use SAS EG for learning, regardless of your status (non-student, student, professor, professional, etc.) at http://www.sas.com/en_us/software/university-edition.html.

Conclusion

I will not argue that either SAS or open-source tools are better than the other. Instead I will state that they each have their use in various situations. Yet, SAS is tried and true…and I will continue to use both.



What is BOARD BEAM ?

Is it a beam supporting a roof? Is it a board made from Canadian timber? Is it capable of leaping a tall building in a single bound? No, it is a tool for performing analytics! BOARD is the company and software platform and BEAM is a software module: BOARD Enterprise Analytics Modeling (BEAM).

Disclaimer

I am not a spokes-person, customer, or user of BOARD BEAM, but I think that it is worthy enough of our attention to write about it here.

What is it?

You could probably compare BOARD BEAM most closely with Tableau, but I will not do so here. I'll just tell you what I have seen with BOARD BEAM. At first glance, it integrates advanced and predictive analytics with business intelligence and performance management. It's capabilities range from analytics reporting to predictive analytics using time series models. It is also easy to use. I have spoken with project managers who use it for powerful business insights and take action, and

data analyst who use it for reporting medical procedure information in healthcare to insurance providers.

BOARD BEAM was conceived and developed in collaboration with IDSIA, a Swiss research institute recognized amongst the top 10 world-players in the Artificial Intelligence space. It is an answer for two key business needs. First, the necessity to embed more advanced analytical and predictive capabilities into companies' business processes ensuring the agility necessary to operate at the rapid pace of today's businesses. Second, is the need to do this in a simpler way, allowing business users to take the lead without having to depend on an army of experts as necessary with the traditional data-mining and analytics products to manually build analytical data models. Of course, that could potentially leave me unemployed!

What does it do?

BOARD BEAM provides a solution that covers many analytical areas through three different modules: Predictive Analytics, Clustering and Analytical Functions.

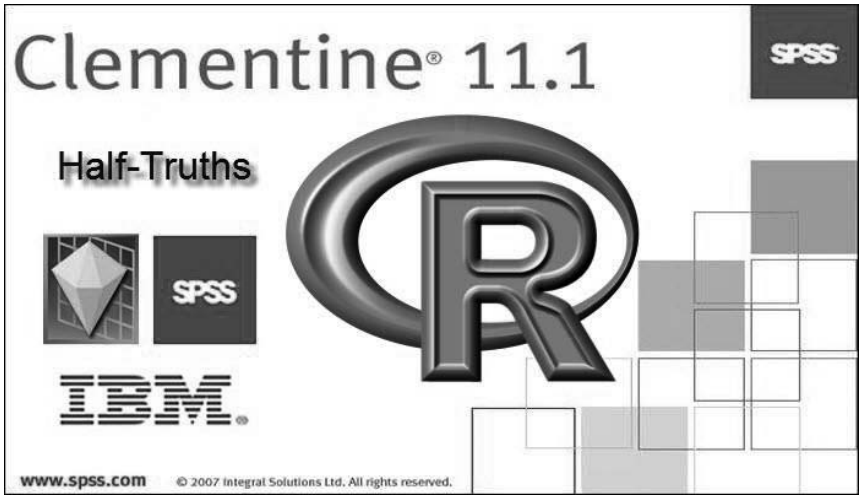
BOARD BEAM possesses automated predictive modelling to execute extremely accurate forecasts in a fraction of the time compared to traditional analytic solutions. Using time-series models for smooth, intermittent and discontinuous Time-series data.

You can automatically group your customers, products etc. into clusters and immediately use them as analysis dimensions in your Business Intelligence environment. Using k-means as its basis, clustering can be performed with a few clicks of the mouse. There is no coding involved, and if you can spell "K-means" correctly, you can probably perform it as well.

It provides Out-of-the-Box statistical functions to increase the value of your business analyses, reports, and dashboards and to build new powerful analytical models. These functions range from the traditional min/max, average, standard deviation, to algorithms specifically designed for business analysis such as frequency, recency, dormancy and nascency.

How can I get it?

You can test-drive BOARD BEAM by visiting http://www.board.com/us/.



R, SPSS Modeler and Half-Truths

The Ad

"Please join us on May 26th, 2015 at 9am PDT for our latest **Data Science Central** Webinar Event: **7 Reasons to combine SPSS Statistics and R** sponsored by **IBM**."

The Claim

The ad does on to claim: “According to the Rexer survey,* R is the analytic software of choice for data scientists, business analysts, and data miners in the corporate world. Despite R's popularity, adoption of R has lagged due to a few limitations like:”

- Deployment - Using R to integrate predictive outputs into an operational environment can be difficult.
- User Interface - R does not have a modern graphical user interface, which makes it difficult for those who are not R programmers to use it.
- Learning curve - R is not easy to learn for everyone. Not everyone is a programmer.

- Data Complexity - R does not easily connect to databases natively.
- Output - Production of publish-ready output is difficult.
- Performance & Scalability - R can very quickly consume all available memory.
- Collaboration - R makes sharing work among an analyst team difficult, especially when team members do not have the same level of R knowledge.
- Enterprise security - The security of the packages that you download is not assured.

The Survey

I posted partial results of Rexer survey yesterday on bicorner.com and today on LinkedIn. I have read the entire survey of about 225 responses. I would not draw the same conclusions entirely, unless I worked for IBM.

IMB's Motive?

It is of course in IBM's best interest to draw unsupported conclusions to support their product, *SPSS Modeler*. However, what they do not tell you is that *SPSS Modeler* is not that user friendly either. *Modeler* alone lacks functionality that could be provided by SPSS Stats, but you have to pay extra for that. Consequently, building a model for measuring the information value of variables is cumbersome in *SPSS Modeler*, but is quite easy in *SAS Enterprise Guide*.

How is SPSS Modeler?

Admittedly, *SPSS Modeler* is a nice tool and I do use it from time to time; however, I prefer *SAS Enterprise Miner* (EM) over it. *SAS EM* is more intuitive to use and the available documentation is much better than what is available for SPSS Modeler. Formerly *Clementine*, IBM did not improve documentation when they acquired the tool.

Half-Truths

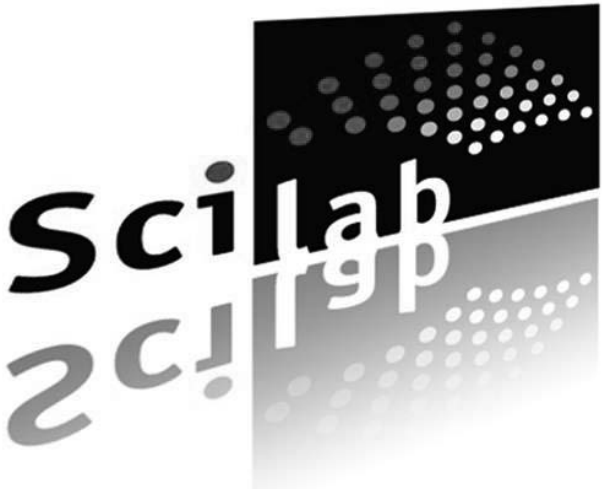
What bothers me about this is the half-truths that are espoused by the ad (and to me half of a truth is the equivalent of a lie). You could read the results of the survey and use them to justify purchasing a red Mustang convertible just as easily. This is not to say that there are not some fundamental issues with relying on R solely for your analytics. But, some of their "bold" conclusions are not accurate. For instance, "R is not easy to learn for everyone. Not everyone is a programmer," is true; however, to do anything other than very basic modeling with *SPSS Modeler*, you have to write scripts, and the scripting language is more difficult that R's programming language. Plus the documentation on the scripting language is almost non-existent.

Open-Source versus Commercial

Moreover, the statement, "R makes sharing work among an analyst team difficult, especially when team members do not have the same level of R knowledge," is also partially true, with the added especially. But the same is true for *SPSS Modeler*, "especially when team members do not have the same level of" *SPSS Modeler* "knowledge." Without their qualification, "especially", the whole idea behind open-source is about sharing. But IBM certainly does not want to share the market with R. So as R gains momentum in the analytic community, why not put forth the idea that R is good but *SPSS Modeler* makes it better (perhaps the opposite is more truthful).

Conclusion

Nevertheless, it will probably be a good webinar. It is just lame that you have to tell half-truths to advertise it. Finally, if you want *SPSS Modeler* to be better, include *SPSS Stats* functionality for the modeling blocks included in *Modeler,* and work on you documentation, a lot!



An Analytics Best Kept Secret

I have written about R. I have told you about Octave. Little did I know that I had not found the motherlode. Thanks to my friend and colleague Dan Pompea, I now have discovered SciLab.

If MATLAB has any advantage over this open-source gem, I have yet to find it (at least for use in operations research and analytics). Not even Simulink is worthy of mention in the same breath. I have used MATLAB extensively. It was the foundation of my bestselling book, Missile Flight Simulation. Now I must turn my back to this industry-standard workhouse and its ferocious price tag to this newcomer and it's incredibly lost cost: a donation if you choose.

I have yet to explore its vast capabilities but what I read is remarkable. However, I have carefully stressed its modeling and simulation features, and I say, "Farewell poor Simulink, I knew you so well." I am so sorry for your loss, dear Octave, for you were becoming a shining star. But now there is a new kid in town, whose glimmer is brighter than your own.

To R, I say, "Work hard in your niche or you will be overcome." Though SciLab boasts of it statistical modeling capabilities, I have thus far not found the equalizer. Thus my dear R, you still have life abundantly.

To ExtendSim, I say, "I enjoyed our time together, and the role you played in my second best-seller, Discrete Event Simulation using ExtendSim." To Arena Simulation, which I introduced to Army Operations Research ten year ago, I say, "Farewell trusted companion. I hope your future is bright."

To SciLab, I say "Welcome, jewel of the Earth."

What is Linear Programming?

Linear programming (LP) is a tool for solving optimization problems. In 1947, George Dantzig developed an efficient method, the simplex algorithm, for solving linear programming problems (also called LP). Since the development of the simplex algorithm, LP has been used to solve optimization problems in industries as diverse as banking, education, forestry, petroleum, and trucking. In a survey of Fortune 500 firms, 85% of the respondents said they had used linear programming. As a measure of the importance of linear programming in operations research, approximately 70% of this book will be devoted to linear programming and related optimization techniques.

Optimization simply means to find the optimal of best as possible solution to a problem. These are often framed as maximums or minimums. For example, maximize profit and minimize costs.

What Is a Linear Programming Problem?

Here I will introduce linear programming and define important terms that are used to describe linear programming (optimization) problems.

Algorithm. Optimization modeling**.**

Step 1. Find the **decision variables**, i.e. find out what are the variables whose values you can choose.

Step 2. Find the **objective function**, i.e. find out how your objective to be minimized or maximized depends on the decision variables.

Step 3. Find the **constraints**, i.e. find out the (in) equalities that the decision variables must satisfy.(Don't forget the possible sign constraints!)

This is not the best example in the world, because we can solve this without a linear program. But it is one that makes illustration of the concept simple. Let's say we want to maximize *Profit* on a certain product. *Profit* would be our decision variable. In simplest terms, *profit* is a function of *revenue* minus *costs*, so we could write the objective function as

$$\text{Maximize: } Profit = Revenue - Cost$$

Now, do we ever have anything constraining revenue and cost? What about our budgets? So, say we cannot spend any more than $500 on advertising the product and it cost $20 to make each product, and for simplicity, those are our only costs. Let's also say that we can only produce 250 of this product, which would constrain revenue. Let's further say that we can only charge $40 for each product.

Our constraints are:

- Marketing cost must be less than or equal to $500
- Production cost equal $20 per item time 250 items
- Revenue is less than or equal to $40 time 250 items.
- All of the values are positive numbers or zero

So, we write this in mathematical shorthand as, let P = Profit, R = Revenue and C = Costs, (s.t. is shorthand for subject to, which identifies the constraints) then

$$\max P = R - C$$

s.t.:

1. *C*(marketing) <= $500
2. *C*(production) = $20*(250)
3. R <= $40*(250)
4. C >= 0 and R >= 0

We would then code this into a software program like Excel, LINDO, MATLAB, etc. and let the software solve the LP. Though there are many other technical details associated with arriving at a solution, this is basically what linear programing is all about. We could make this problem more interesting by allowing the number of items produce vary between zero and its upper bound of 250.

MATLAB Code

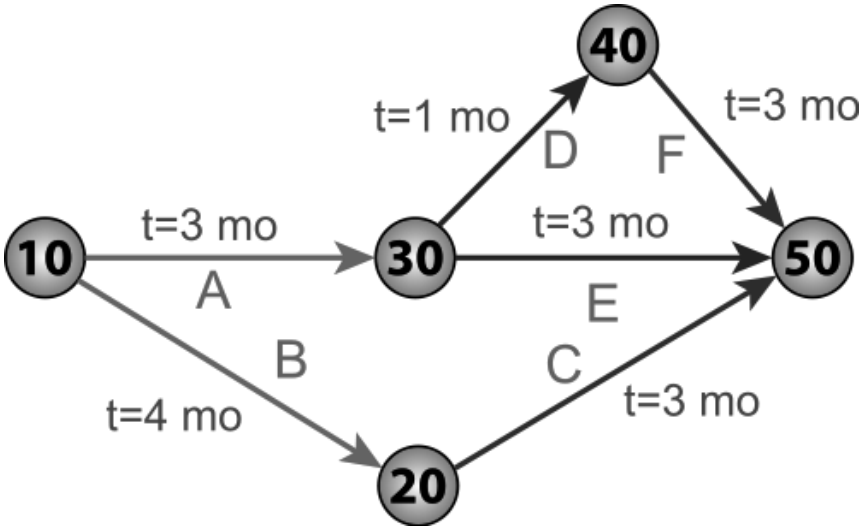
This is an LP with six variables and seven constraints. The code consist of defining three matrices, A, b and c. The code that executes the LP is the line at the bottom. It happens to be a minimize LP. **glpk** is the function that executes the LP, the zeros are lower bounds of the variables, "L" indicates the constrains have lower bounds, "C" indicates the variable are continuous, and "1" tells the program to minimize. The ***c*** matrix represents the objective function, While ***A*** and ***b*** constitute the constraints. This also runs in an open source program called Octave.

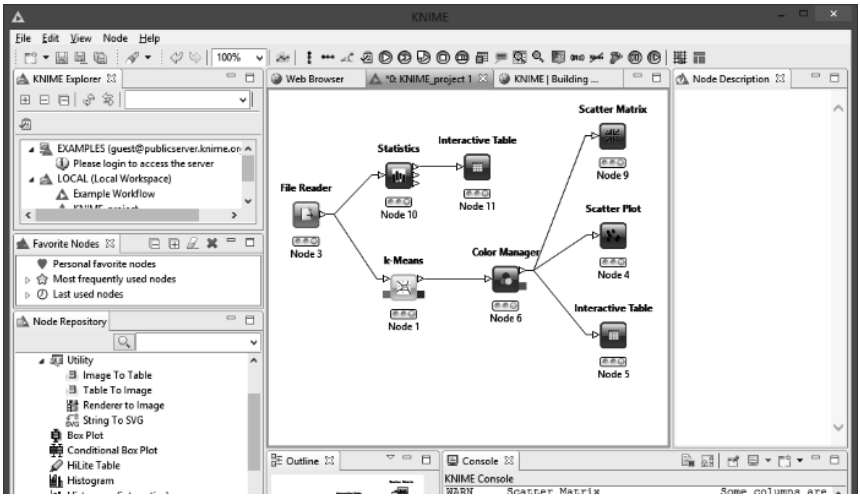
```
c=[-1, 0, 0, 0, 0, 1];
A=[-1, 0, 1, 0, 0, 0;
-1, 1, 0, 0, 0, 0;
0, 0, -1, 0, 1, 0;
0, 0, -1, 1, 0, 0;
0, 0, 0, -1, 1, 0;
0, 0, 0, 0, -1, 1;
0, -1, 1, 0, 0, 0];
b=[6; 9; 8; 7; 10; 12; 0];
[x_min,z_min,status,extra]=glpk(c,A,b,[0;0;0;0;0;0;0],[
],"LLLLLLL","CCCCCC",1)
```

Conclusion

LPs are used for optimizing schedules, networks, transpiration systems, delivery systems, manufacturing, investing, and so on.

These will often be comprised of many variables and constraints. LPs can be used to optimize a project schedule in Project Management. *Operations research analysts* are specifically trained to formulate and solve these kinds of problems.





What is KNIME?

Summary

KNIME, pronounced "naim", is a modular data exploration platform that allows the user to visually create data flows (referred to here as workflows). One key behind the success of KNIME is its inherent modular workflow approach, which documents and stores the analysis process in the order it was conceived and implemented, while ensuring that intermediate results are always available.

Core KNIME features include:

- Scalability through sophisticated data handling (intelligent automatic caching of data in the background while maximizing throughput performance)
- High, simple extensibility via a well-defined API for plugin extensions
- Intuitive user interface
- Import/export of workflows (for exchanging with other KNIME users)
- Parallel execution on multi-core systems
- Command line version for "headless" batch executions

Available KNIME modules cover a vast range of functionality, such as:

- I/O: retrieves data from files or data bases
- Data Manipulation: pre-processes your input data with filtering, group-by, pivoting, binning, normalization, aggregation, joining, sampling, partitioning, etc.
- Views: visualize data and results through several interactive views, allowing for interactive data exploration
- Highlighting: ensures highlighted data points in one view are also immediately highlighted in all other views
- Mining: uses state-of-the-art data mining algorithms like clustering, rule induction, decision tree, association rules, naïve Bayes, neural networks, support vector machines, etc. to better understand your data

You can check out the complete node documentation for a comprehensive list of nodes and detailed descriptions at http://www.knime.org/files/node-documentation/index.html.

Supported Operating Systems

- Windows - 32bit (regularly tested on XP and Vista)
- Windows - 64bit (regularly tested on Vista and verified to work under Windows 7)
- Linux - 32bit (regularly tested on RHEL4/5, OpenSUSE 10.2/10.3/11.0, amongst others)
- Linux - 64bit (regularly tested on RHEL4/5, OpenSUSE 10.2/10.3/11.0, amongst others)
- Mac OSX - 64bit Intel-based architecture with Java 1.6

Building a Simple Workflow

A workflow is like a Diagram is SAS Enterprise Miner or a Stream in SPSS Modeler. Figure 1 shows the KNIME workbench windows, including the workflow.

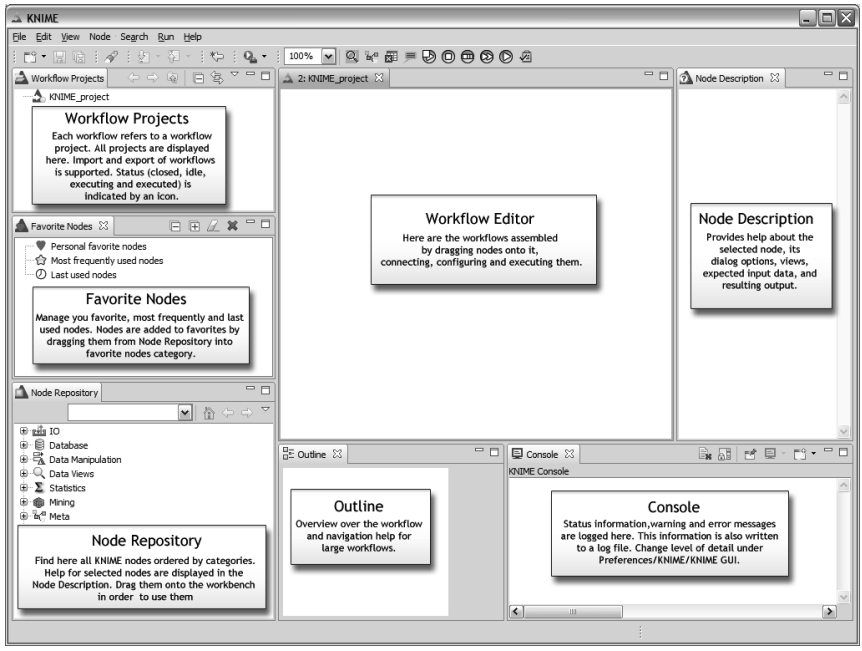


Figure 1. The KNIME workbench (GUI)

I am going to walk through the process of building a small, simple workflow so that you can get an idea of the environment. I will read in data from a comma delimited (CSV) file, assign color to it, cluster the data and display the data in a table and a scatter plot. I will also display the data with a Scatterplot Matrix. After I execute this flow we will examine the data model that has been built. We assume you have just started KNIME with an empty workflow.

Adding Nodes

KNIME has a Node Repository containing all the functions used in creating a workflow (see Figure 2). I am using the IO, HiLite (Highlight), Mining and Statistics modules. To import my data I expand the “IO” and the contained “Read” category as depicted below (left picture in Figure 2) and drag & drop the File Reader icon to the Workflow Editor window. The next node for now will be the k-Means clustering algorithm. I expand the "Mining" category followed by the "Clustering" category, and then drag the K-Means node to the flow (picture on the right in Figure 2).

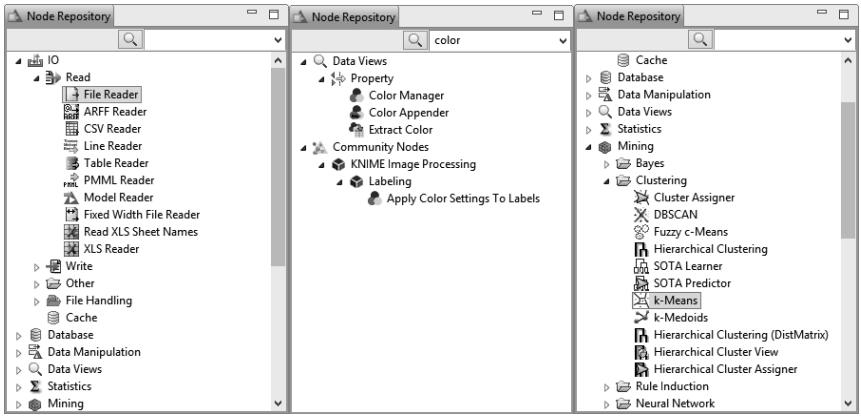


Figure 2. Where to find workflow nodes in the Node repository

Since I do not always remember were various nodes are found, I use the search box of the Node Repository and I enter “color” and press Enter. This limits the nodes shown to the ones with “color” in their name (see Figure 2 in the middle). I then drag the Color Manger node to the workflow (this node will define the color in the data views later). To see all nodes in the repository again, I press the ESC or Backspace key in the search field of the Node Repository. Now, I drag the Interactive Table and the Scatter Plot from the "Data Views" category to the Workflow Editor and position it to the right of the Color Manager node.

Connecting Nodes

Now I need to connect the nodes in order to get the data flowing. I click an output port and drag the connection to an appropriate input port. The complete flow is pictured in Figure 3.

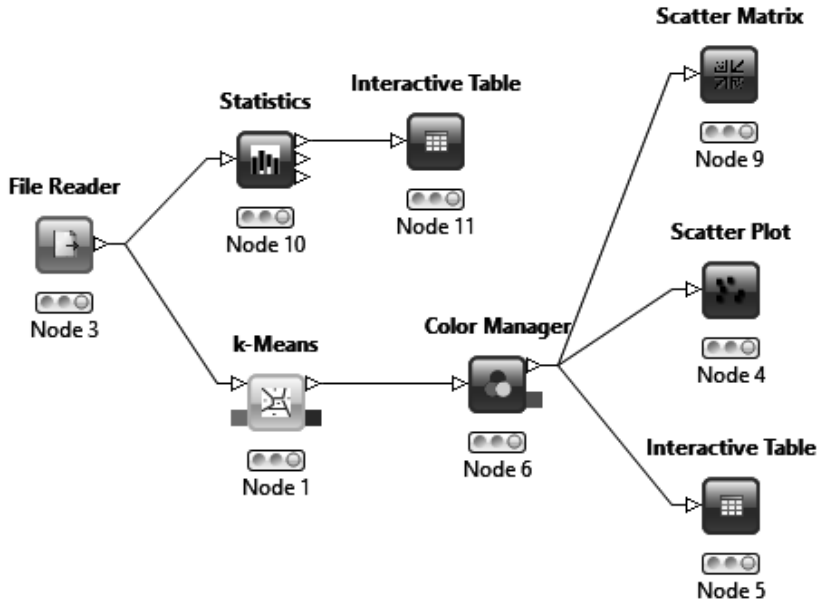


Figure 3. Final workflow diagram

My nodes will not show a green status, as long as they are not configured and executed.

Configuring Nodes

Fully connected nodes with a red status icon need to be configured. I start with the File Reader, right-click it and select “Configure” from the menu. I navigate to the directory where my data is located. I select the "iris.csv" file from this location. The File Reader's preview table shows a sample of the data (see Figure 4).

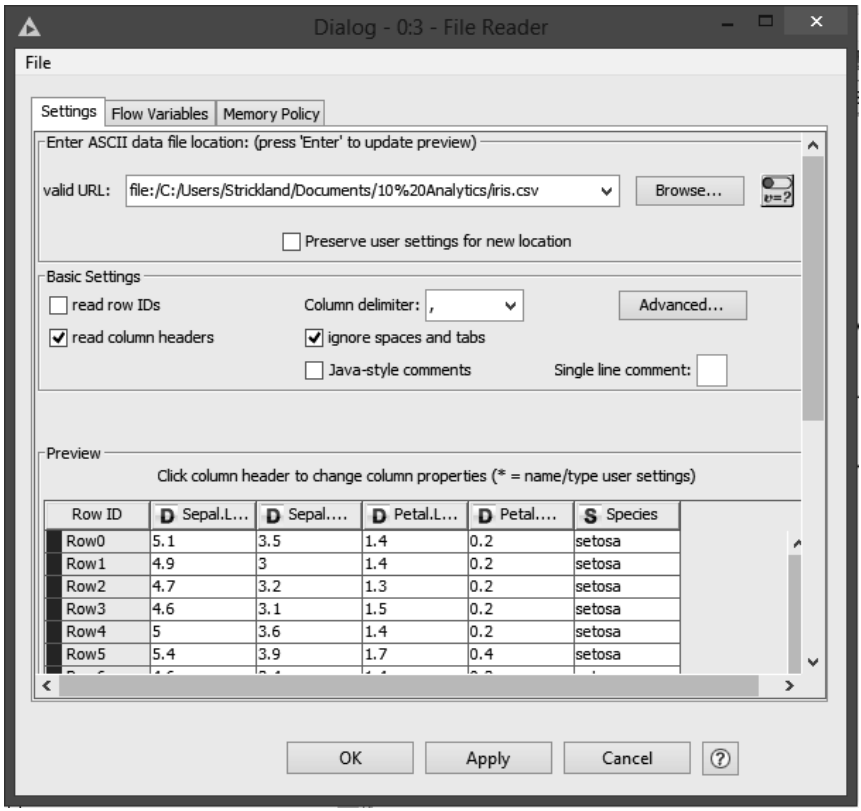


Figure 4. Configuring the File Reader node

I will skip the remaining node configurations here.

When I execute the workflow for k-Means I actually execute the Scatter Plot node. To view the node, I right-click on the Scatter Plot node and select View: Scatter Plot. The scatter plot showing clusters is Figure 5.

Scatterplot Matrix

There are some problems related to the scatterplot. First, if several data points share the same coordinate it is not visible to the user. This problem is also known as over-plotting. Second, a scatterplot can always use only two dimensions. If the data has more dimensions (attributes) the user must constantly switch the displayed dimensions in order to understand the data.

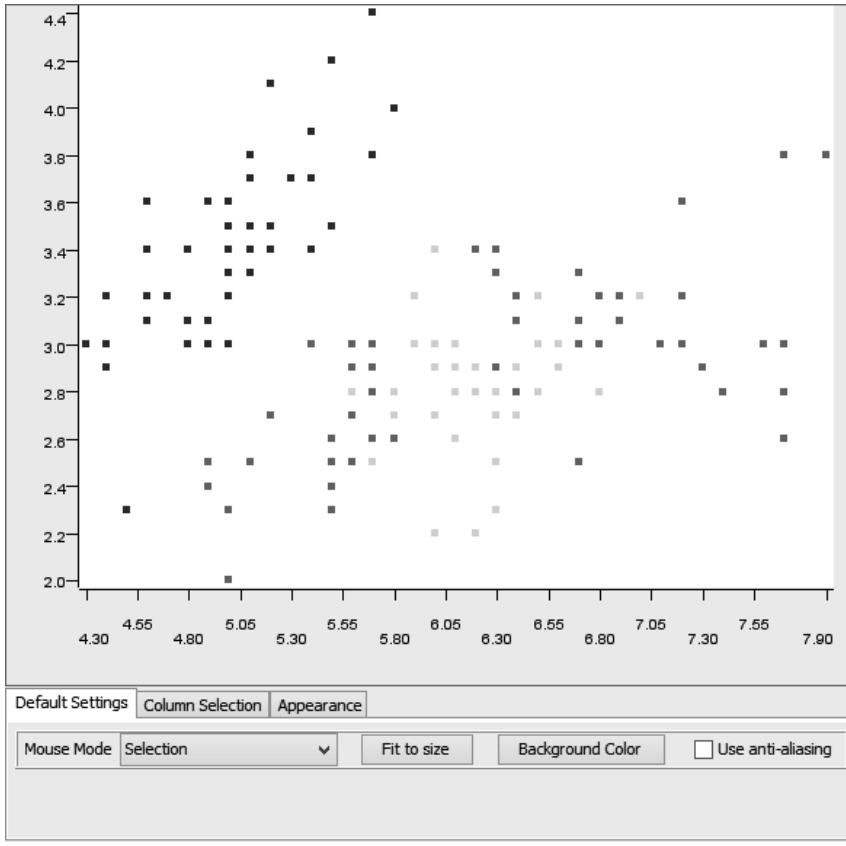


Figure 5. Iris dataset cluster scatterplot.

The use of a scatter matrix overcomes the second problem (but is also restricted to low dimensional data). All dimensions of the dataset are plotted as a matrix, where each attribute is plotted against each other. Figure 6 shows the scatterplot matrix for the iris dataset (compare it with R in Figure 7).

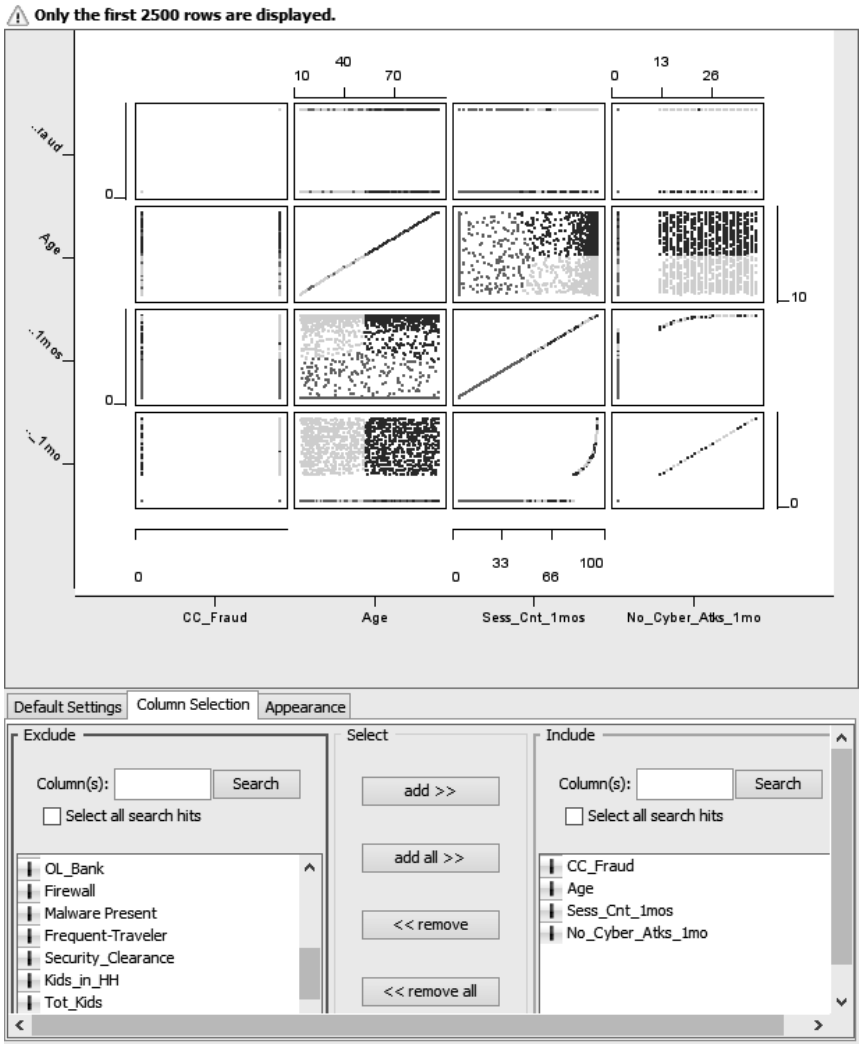


Figure 6. Iris dataset scatterplot matrix

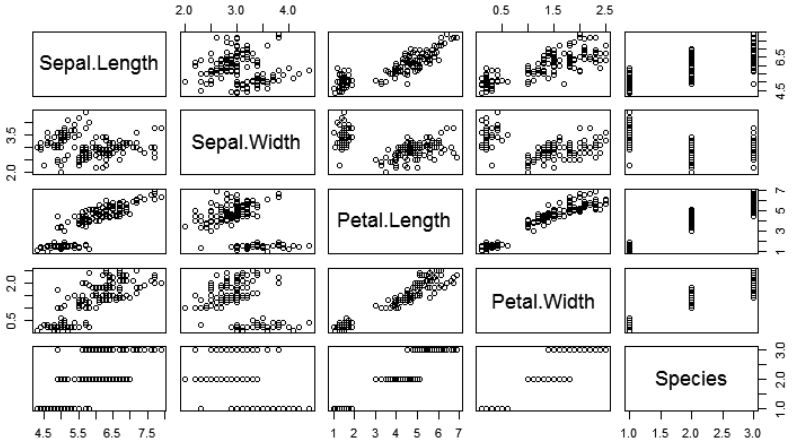


Figure 7. Iris dataset plot in R

Additional Packages

KNIME supports other programs including R and MATLAB. Figure 8 shows the R nodes in the Nodes Repository.

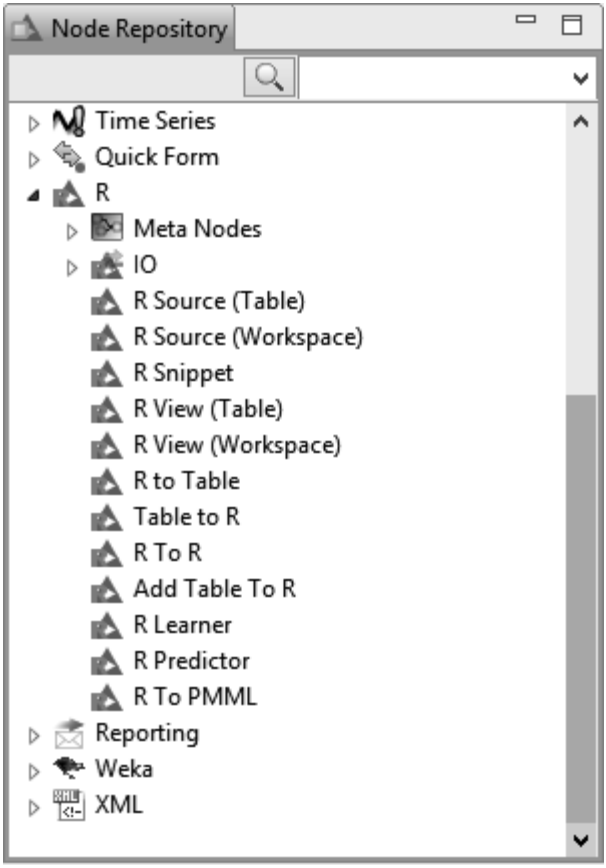


Figure 8. R nodes in the Nodes Repository

Conclusion

Though I have not worked with KNIME as much as I would like, my first impression is a favorable one. To download KNIME to you MAC or PC, use this link http://knime.org/download.

Part IX - Advice

Seven Analytics Lessons Learned

I have now been a Senior Analytics Consultant with ***Clarity Solution Group*** for three years (from 22 May 2012). This was a career change of sorts, moving from modeling, simulation and analysis in aerospace and defense industry, to predictive modeling in the FSI industry. Here are some things I have learned (not the ones on the chalk board):

1. When dealing with "big data" (#big#data), about 2/3 of you project time is spent getting access to the data, getting the right data, preprocessing the data, and exploratory data analysis, prior to modeling building.
2. Presentation of the results are as important as the analytics performed. If you cannot convey the results clearly and succinctly, you're not adding value for your customer. In particular, your results must be translated to economic value or a similar metric.
3. Everything benefits from peer reviews, including the 2/3 work performed before modeling (See #1). You should develop a template for conducting peer reviews so that they are thorough and consistent.

4. You should attempt to validate everything. This includes any requirements, data provided (response data, etc.), macros provided, etc. That is to say, not just your model.
5. Listen carefully to the customer and help them express their “real” problem. Often the customer knows they have a problem but either cannot express it or they do not identify the root problem.
6. Ensure that the variables in a model are intuitive and limit them to the most predictive variable. Rule of thumb: not more than ten variables.
7. Document your work as you progress. Either keep a notebook for your project or record it electronically. I use a spreadsheet with a tab for each task or sub-task. This may be used later for a writing development document or providing information for validation.

I am certain that I could provide more, but seven sounds good. In addition to the lessons learned, it has been fun to step back from being a technical director to a modeler/analyst. It has also been great to work for a company that values its employees and treats them with respect. Ii has also been a pleasure sharing knowledge (non-proprietary) with you, my peers.

About the picture: The lessons learned on the picture are seven of ten staff rules I learned from the executive officer when I was a personnel officer in 3rd Squadron 2nd Armored Cavalry Regiment. Note: Charlie Beckwith commanded Desert One during Operation Eagle Claw (Google it).

Here's one more the thing…advice for young people

In this article I want to add to "Here's the thing…more advice for young analysts", based on questions I continue to receive through LinkedIn messages and e-mail. I also want to extend this beyond advice to analysts, and include anyone who may have an ear to hear.

Here's the thing…Everyone wants to be noticed, to stand-out from the crowd. One of the best ways to do this is blogging, like posting articles here on LinkedIn. It also helps to have your own blog. But keep it professional. If analytics is your area of expertise, write about analytics. If it is human resources, then write about that. Yet, only write what you know about, or if you are just stimulating conversation and feedback, then be honest. People will see right through a fake.

Here's the thing…Not everyone can write. If you are one of these, then read and comment on other people's articles/blogs. Better yet, comment and share both on LinkedIn and Twitter. Every LinkedIn post you read has an icon (under the title) for LinkedIn, Twitter, Google+ and Facebook. Sharing helps the author get more exposure, and the person who shares. Also, reading more may also make you a better writer.

Here's the thing…There are two ways to do things: right and over. Even the best of us make mistakes. I can make a mistake in the blink of an eye, even though I have been at my profession for over 20 years. So when I do not get it right, I own up to my mistake and do it (whatever task I was trying to perform) OVER! I suppose that leads to some advice for not so young people: allow your young employees to make mistakes and have them do it over. This leads to learning and better workers.

Here's the thing…I have to repeat this one: some jobs require a Ph.D., like a professor at an accredited university. (You can teach without a Ph.D., just not as a professor—I know a really great guy who has been doing this for years.) Most jobs do not require a Ph.D. However, some people feel like they will never be of much worth without it. And here is the crux of the matter:

"If you are not god enough without it, you will never be good enough with it."

Here's the thing…Some of the very best Data Scientists and Analytic Professionals that I know do not have Ph.D.'s, and hardly anyone addresses me a Doctor (usually just "Hey you!").

Here's the thing…This one is also worth repeating: there are many online courses and tutorials that can provide you with the training you need without getting another degree. For young analysts, I refer you to another bicorner.com article I wrote recently: Online Education in Analytics and Data Science. In general, Coursera offers many courses with flexible scheduling at no cost. The key here is to become a "lifelong learner", constantly seeking self-improvement. Oh, and it shows a little initiative, not waiting on the company to train you. Most of them do not offer much training anyway.

Here's the thing…Know who you are competing with and (here's the shocker) help them! This can back-fire, but a good supervisor will take note of the one who is helping their colleagues and might say, "I see leadership potential there". There are two ways of doing this (and they are not "right" and "over"): with fanfare and without it. One way to openly do this is searching for people on LinkedIn who have the "title" or "professional headline" that you are interested in. Study profiles to see what these people are doing and have done and connect with them.

Oh, and I mean CONNECT with them, not just add someone to your network. In fact, go back to my second "Here's the thing..." and share what other people are doing. At work, I would do it with less fanfare--not that I would hide it, but I wouldn't go to extremes in making sure my help was noticed.

Here's the thing…and this one is difficult and counter-intuitive: Be selfless! That's right, I did not say selfish, rather SELF-LESS. I think that literally means "without self", so as an adjective describes the one who acts without self in the forefront. Face it, we are all a little selfish--it is human nature. But, the ability to suppress our nature and act out of regard for others first is also a characteristic that will not go unnoticed by most people.

Index