White Paper

Connecting the Experts with the Data Scientists

by Gerhard Pilcher

Vice President and Senior Scientist

September 2014

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

"Can Machines Think?" was the cover of Time magazine a generation ago. The impetus was a face-off between world chess champion Gary Kasparov and IBM's "Deep Blue" chess machine. The first match was won by us humans, but algorithms and hardware won out in the end. The spectacle occasioned wild speculation. "Of course computers can think, after all, humans are just computers made of meat!" said an MIT Computer Science professor. How can anyone who works with both humans and computers, believe that? My colleague, Dr. John Elder, likes to say "He's from MIT; maybe he's never worked with humans!"

The unscientific claim ignores the power of human vision and the brain's ability to connect the context and semantics of information — something a computer does poorly if at all. Of course, humans are unacceptably slow at repetitively looking through millions of records that contain hundreds of features and accurately make repeated calculations. Humans also bring along some baggage called biases that can distort our observations about information, something computers lack. The focus of this article is to illustrate how using the complementary strengths of humans and computers can lead to developing better analytical models and directing qualified action on the results.

Impatience with Business and Data Understanding

It is often said that "sixty to seventy percent of the time building an analytic model is spent cleaning the data." Perhaps from a data scientist's point of view, everything that happens prior to the fun part of fitting multiple models to the data is lumped into the category of data cleaning. We (analytic professionals) can be whiners when the data doesn't perfectly suit our modeling needs! But as I've examined this phenomenon through the lens of experience, I find that the bulk of that quoted sixty to seventy percent in time is actually spent refining the business question(s) and gaining a deeper understanding of the raw data. Data modeling contests (see Kaggle.com) are fun because the hard work of defining the problem and preparing the data has been completed before the contest is posted.

All data is messy to a certain extent but I think much of the perception of messy is about these harder tasks of refining the business question(s) and earning enough subject matter knowledge to gain a competent understanding of the data elements and the processes used to create the data. I use the adjective "competent" deliberately. It is defined as "having suitable or sufficient skill, knowledge, experience, etc., for some purpose; properly qualified." An analytic professional must have sufficient knowledge and experience with the data for the purpose of building a properly qualified model. It's easier to make assumptions about the meaning of a data element based on the name or label associated with the data than to dig into data dictionaries or to track down and interview subject matter experts. I've witnessed well-intentioned people making assumptions about the meaning or source of data elements based on the name or label associated with the data only to discover later that their assumptions were in the opposite direction of the truth.

In virtually every project with which I've been associated, there has been an invisible hand pressuring the project to shortcut the process of business and data understanding. From the perspective of a data scientist, I think we might fall into the trap of thinking that we're not doing our job (what we've been trained to do) until we are in the model fitting process, or maybe we are simply less interested in the hard work of understanding the business and data. Whatever the reason, there is an underlying sense of urgency to get to the model fitting process at the cost of potentially short-cutting the essential step of gaining a competent understanding of the business and data. Not only are analytic professionals impatient to get into the model fitting process, business owners and management drive some of the impatience by applying pressure to see early results (typically cloaked as "having to justify the investment made in project costs"). Management may also fall into the trap of resisting the extra work required to coordinate the availability of subject matter expertise and the uninterrupted time needed to think through refinements to the business questions driving the analysis.

Analytic professionals and management with more experience in analytical projects have gained an appreciation for the value of refining the business questions and realize that analytical methods applied in a vacuum — absent subject matter knowledge — can result in models that may be technically "accurate" but fail to provide insight or predictions that are useful to the business. At the heart of success is someone who has taken the time and has the skill to manage communication between groups with different knowledge sets and lexicons. There is no "magic" in the mathematics of fitting models but there is some art in managing the communication path between business experts and analytic professionals.

I'm not advocating that projects be allowed to go willy-nilly for as long as anyone remains interested, but instead that we direct our evaluation of progress early in the project on how complete the knowledge transfer is between business and data experts. The idea of competent understanding provides a rough benchmark of the subject matter knowledge the data scientist should have when beginning the process of fitting models and generating analytic results.

Once fitted models begin producing reliable analytic results — as measured purely from a data science perspective — then subject matter experts are again needed to qualify the results from a business perspective. In other words, bring the human brain's ability to connect the context and semantics of information into the interpretation of the results.

Using Experts to Qualify Model Results

In the famous example of Anscombe's quartet (below), there are four series which all have the same exact statistical relationships and data descriptions (shown in the red box). But if you plot the four data series {(X,Y1), {(X,Y2), {(X,Y3), {(X4,Y4)} the human eye can instantly recognize major differences.

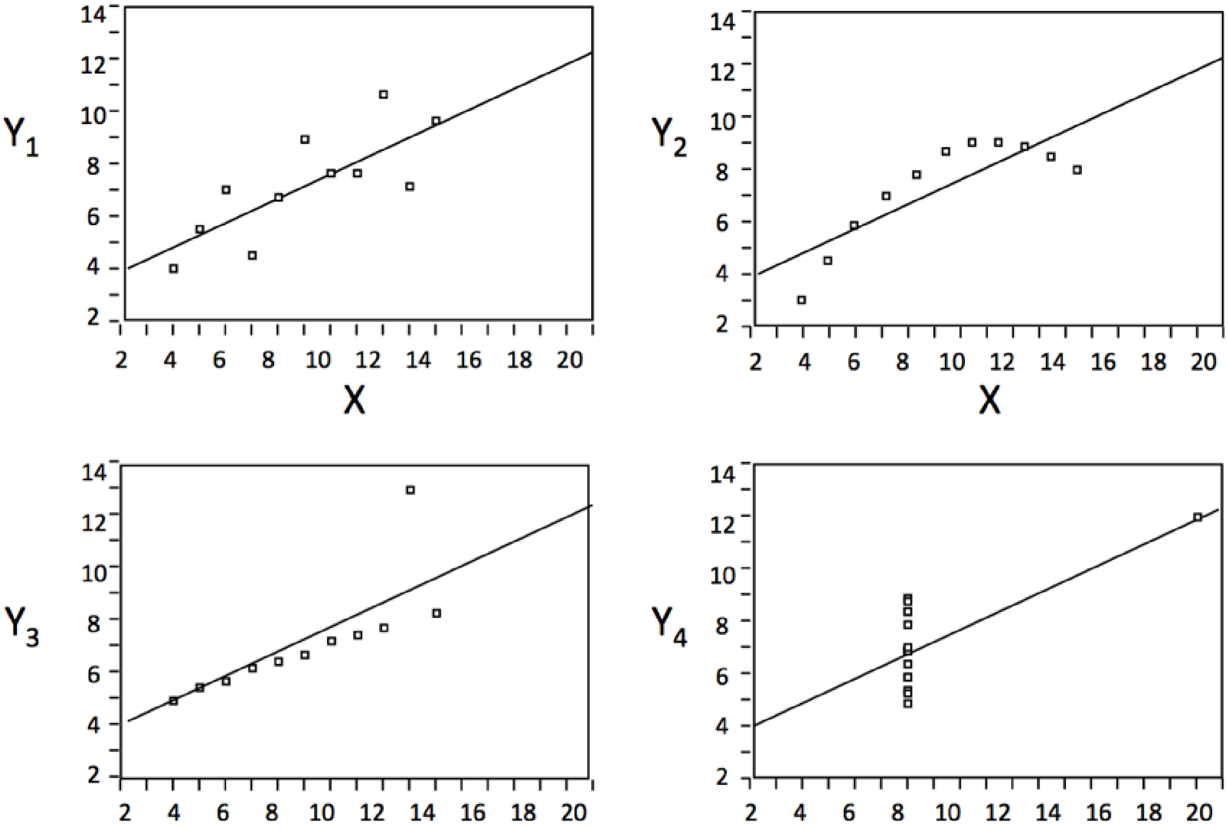
X	Y_1	Y_2	Y_3	X_4	Y_4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

$\rho_{xy} = 0.85$

$y_{LS} = 3 + 0.5x$

MSE = 1.25

$R^2 = 0.67$

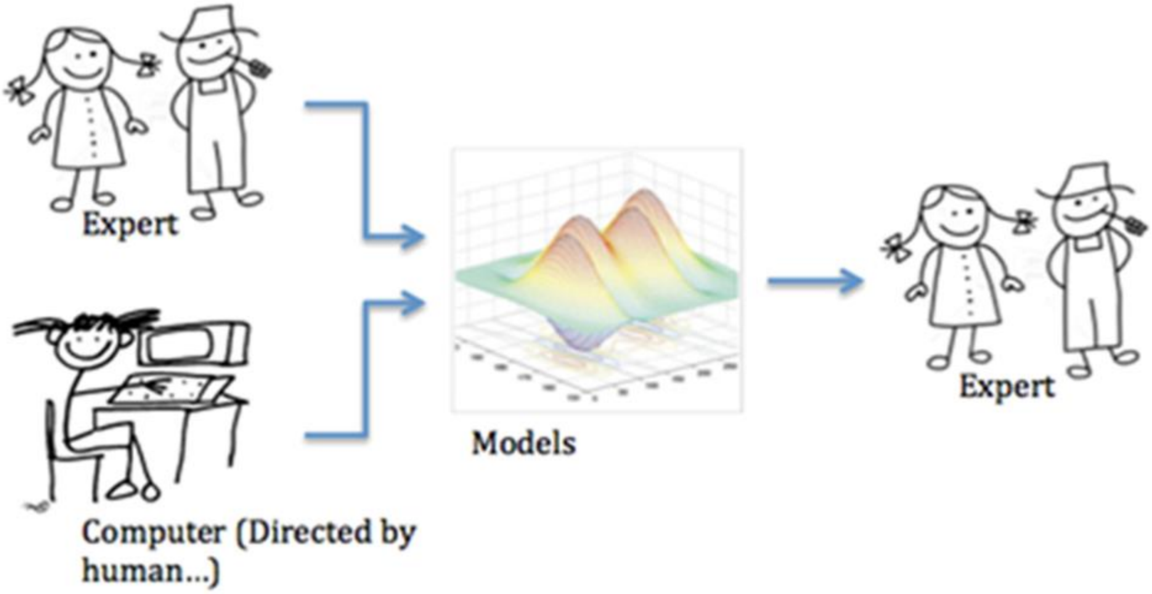


In our experience the highest return on investment is achieved when projects find a way to marry the knowledge of their subject matter experts with sound data models. In his book "The Signal and the Noise", Nate Silver points out how important humans are in qualifying analytical results. He details several different examples from multiple disciplines, from baseball scouting to weather forecasting by the National Weather Service. The latter is particularly compelling because the NWS measure the value of human participation by maintaining two sets of forecasting records: one of model performance alone and another that accounts for how much value humans contribute to model output in the published forecast. Based on NWS numbers, human input improves precipitation forecasts by 25% and temperature forecasts by 10% over the models alone.

Experts are crucial to the development of models but even more so to the use of model results as demonstrated by the NWS example. Computers are great at decision support when all of the decision parameters are within the known (i.e. programmed) domain. Consider flight computers in commercial aircraft. Modern flight computers are capable of directing aircraft from takeoff to landing but I suspect none of us would be comfortable with an empty cockpit! Pilots guard our safety in unexpected circumstances such as when that USAir flight made an emergency landing on the Hudson River in New York after both engines failed. Flight computers are programmed to land aircraft on runways equipped with guidance systems that can provide course and altitude corrections. I'm certain there are not any guidance systems along the Hudson River!

Experts and Success

Successful projects begin and end with business matter subject experts contributing to the definition, understanding, and qualification of analytic models. A secondary benefit to expert involvement in the modeling phase of the project is a better understanding of the analytic methods and hence a higher level of confidence in the model results. The "magic" in analytics might boil down to the ability to create a mutually supportive communication channel that connects the knowledge sets and lexicons of subject matter experts and data scientists.



About the Author

Vice President and Senior Scientist Gerhard enjoys predictive analytics and data mining, especially related to the areas of Fraud Detection, Financial Risk Management, and Health Care outcomes using various analytical methods, working with people, leading change, and timely management of complex projects. His work experience spans both private and government sectors including international experience.

Gerhard teaches at Georgetown University as an adjunct faculty member in the Math and Statistics Masters degree program. He also is an instructor for the three day SAS Business Knowledge Series course "Data Mining: Principles and Best Practices" and been invited to teach at international conferences.

Gerhard currently serves on the Institute for Advanced Analytics Advisory Board and George Washington University Masters in Science in Business Analytics Advisory Board.

Gerhard has extensive industry experience in government oversight, financial, construction and telecommunication industries both as a business owner and executive. He is a recognized expert in three dimensional roadway modeling and automated machine guidance using Global Positioning Satellite systems and has presented to various agencies including the Transportation Research Board. In his role as Chief Technology Officer and VP of Engineering for Pulse Communications, Gerhard directed the design of early digital subscriber line systems (internet over the telephone line) and was a member of the international forum defining the standards for DSL implementation. Prior to Pulse Communications he was Director of Operations for Bell Northern Research leading the design and delivery of hardware and software for large scale telephony switching and fiber optic systems.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673