Editor: **Daniel E. O'Leary,** University of Southern California, oleary@usc.edu

Emerging White-Collar Robotics: The Case of Watson Analytics

Daniel E. O'Leary, *University of Southern California*

Increasingly, systems are being built that might be called "white-collar robots." If you watch TV news, you likely would define a white-collar robot as something that takes jobs from white-collar human workers. That perspective seems unnecessarily limiting. So, let's define a white-collar robot as a set of computer-based capabilities that performs a task or a set of tasks historically done by white-collar workers, but without the need for human intervention. White-collar robots automate particular tasks. A human decides what task a robot will do and then chooses a robot to do that task. As in manufacturing settings, white-collar robots might be part of a process, providing information inputs or outputs to others, whether people or robots. This definition is consistent with robots used for manufacturing and does not require that the robot have "consciousness." Instead, the robot just does things that need to be done. In general, the robot is likely to be able to do some things better than the person, but the person is likely to do some things better than the robot.

What tasks might a white-collar robot do? On the basis of robots in other settings, we would probably expect a white-collar robot to perform certain tasks without the human specifying all of the details every time. For example, we might give data to such a robot and ask it to "analyze it, and I will ask you some questions later—come back to me with some suggestions as to what you think the data says." That description is exactly what is being done by some analytics software. Increasingly, even those with limited statistics or analytics knowledge can use easy and smart analytic software (robots) to analyze data. These robots are part of an effort to "leave no manager behind" in a "data-driven world" simply because those managers don't know much about analytics. One such effort, Watson Analytics (www.ibm.com/analytics/watson-analytics), provides active software that helps any user investigate data. Upload data to the cloud, and Watson Analytics will analyze the data quality, provide initial analysis, and prod you to consider different combinations of variables.

Watson Analytics is distinct from Watson Cognitive, which gained fame from capabilities used on the TV show *Jeopardy!* The latter's capabilities focus on text, natural language, and other related applications. As a result, much of the discussion of Watson's capabilities is bifurcated into two seemingly independent pieces—cognitive and analytics.

Easy and smart analytics may be of great interest to many users. To test the appeal of Watson Analytics, I had roughly 90 master's degree students use the software as part of a class introducing them to information systems and analytics. The students had limited experience with statistics and analytics, but by the end of the class, about 90 percent of them identified Watson Analytics as one of the most interesting, important, and potentially useful topics they studied.

To begin to understand some of the potential strengths and limitations of such software, first see the "Active versus Passive Software" sidebar. I will also review Watson Analytics' capabilities, analyze two sets of data in the form of case studies, and summarize some of the software's strengths and limitations.

Watson Analytics' Capabilities

Watson Analytics requires that the user (or someone helping the user) negotiate a data upload to the cloud. After the data is uploaded, Watson Analytics provides starting points for the user to analyze the data. These starting points are a sequence of questions that

1541-1672/17/$33.00 © 2017 IEEE
Published by the IEEE Computer Society

Active versus Passive Software

Historically, software largely has been passive. For example, in the case of statistical software, people input data and then choose which models to analyze the data. In contrast, Watson Analytics is active software: the software makes recommendations and does analysis without needing the user to specify the models.

Increasingly, software is becoming active, and people are taking "orders" from it. For example, it is not unusual to have direction or map software guide us when we don't know where we are going. Thus, people are beginning to expect software to make those recommendations, particularly if they don't have any particular knowledge, are unsure about the set of issues, or don't want to force themselves to think about the issues that the software can do (for example, they want to relax and not concentrate, or concentrate on something else).

Generally, active software makes assumptions as to where or what the user wants or needs. Active software is likely to make predictions about what would be useful to the user. Active software may be autonomous and function largely without the user or its inputs. Active software is likely to be goal oriented. White-collar robots generally will be active software.

Table 1. What drives satisfaction?

Independent variable sets*	Strength (%)	R-square	ANOVA statistical significance
1. Type of travel	37	0.3551	< 0.0001
2. Age and status	33	0.1884	< 0.0001
3. No. flights per annum and year of first flight	21	0.0904	< 0.0001
4. Price sensitivity and age	18	0.0496	< 0.0001
5. No. loyalty cards and age	17	0.0380	< 0.0001
6. Status	17	0.1421	< 0.0001
7. Age	14	0.0380	< 0.0001
8. Status and type of travel	Model not tested or found in Watson	0.4221	< 0.0001
9. Flights per annum and status	Model not tested or found in Watson	0.1914	< 0.0001

*The dependent variable is satisfaction.

the system has developed on the basis of the data—for example, "What drives X?" "What is a predictive model of Y?" and "What is the trend of Y and Z?"

The data format is important to Watson Analytics. Watson does not work well when there are more than two dimensions to the data, say, with row or nested headings, or both. As an example, column heading by years is appropriate, but column headings that break the data into multiple layers—such as year, month, revenue, and quantity—will not work well.[1] In particular, the system works best with data that can be placed in a classic flat file. However, unlike much statistical software, Watson can analyze variables that are categorical or numeric. Watson Analytics provides an analysis of the data quality that indicates "how ready the data is for analysis."[1] In computing the data quality for each field, Watson Analytics apparently considers several factors, including missing values, outliers, symmetry, skewness, and imbalance. The overall data quality score is an average of the data quality score of every variable in the dataset. Furthermore, the existence of a data quality score emphasizes the importance of the data quality to the user. This quality number gives the user direct information about the data's veracity.

Watson Analytics does not appear to use traditional statistical methods such as correlation and regression analyses, but it does use decision trees to determine which variables appear to influence others and to determine the "strength" of those relationships. I analyze these and other capabilities using two case studies.

Case Study 1: Customer Satisfaction

I analyzed a subset of data used by IBM to illustrate Watson Analytics to different groups.[2] The dataset had 499 randomly chosen observations related to customer satisfaction in air travel. The variables included type of travel, age, status, number of flights per annum, price sensitivity, number of loyalty cards, gender, and arrival delay.

As part of its initial analysis, Watson Analytics analyzed the data and generated 10 questions, including the following:

- What drives satisfaction?
- How are the variables of flight distance and flights per annum related?
- What is the relationship between arrival delay and departure delay?
- What is the predictive model of satisfaction?

Because my interest was in the variable "satisfaction," two of the questions that Watson asked were of direct interest. I first investigated the question, "What drives satisfaction?" Watson Analytics generated seven combinations

Table 2. Predictive model for satisfaction.

Independent variable sets*	R-square for entire sample	Sample size for model
Status, type of travel, and no. loyalty cards	0.4226	156
Status, type of travel, arrival delay, and gender	0.4483	138
Status, arrival delay, and type of travel	0.4337	91
Arrival delay and type of travel	0.3665	114

*The dependent variable is satisfaction.

of one and two variables, listed in Table 1 (that is, variable sets 1 through 7). In addition, I developed nine regression equations (for those seven sets and two additional sets of variables) to compare Watson's strength measure and a statistical measure, R-square. Table 1 summarizes the results. Unfortunately, the system did not find two of the three sets of variables listed with the largest R-square measures (variable sets 8 and 9), although they are combinations of other variables considered by Watson.

For this data, the measures of strength and R-square are statistically significantly correlated at 0.899, which is statistically significant at better than 0.006, indicating a close relationship between those two variables. However, it is clear there are some differences between the two measures for this data, particularly with variable set 5.

All of the variable sets had regression coefficients that were statistically significant at better than 0.05. However, the coefficients on "year of first flight" and "number of loyalty cards" were not statistically significant. In addition, age, number of flights per annum, and price sensitivity had negative coefficients, and thus were negatively related to satisfaction. I would expect users to be interested in knowing whether variables were positively or negatively related to satisfaction, but unfortunately, Watson Analytics did not provide that information.

I also investigated the question, "What is the predictive model of satisfaction?" (see Table 2). As part of building an overall model, Watson Analytics fits different models to different parts of the data. For this data, using decision trees, the system built four models, with corresponding subsamples of 156, 138, 91, and 114 observations. I built regression models for the entire sample on the basis of the four models and included the corresponding R-square for those models. Unlike in Table 1, the number of variables was not limited to one or two. Watson Analytics' measures of strength are not given for each of the models and corresponding subsamples, but only for the entire set of models. In this case, it was 46.1 percent, which is greater than any of the one- and two-variable models in Table 1.

The decision-tree models are based on specific values of the particular variables (for example, number of loyalty cards >2). Thus, using Watson, it is not clear whether the variables are positively or negatively related to satisfaction. Furthermore, in the predictive model, each submodel is for only subsets of the variables. As a result, if actions must be taken for the entire dataset, it is unclear which model to use. However, the prediction model does generate what ultimately seem to be some sets of variables that result in better regression models for the data, and could clearly be used to guide further analysis. Finally, those different models offer different potential "global" strategies.

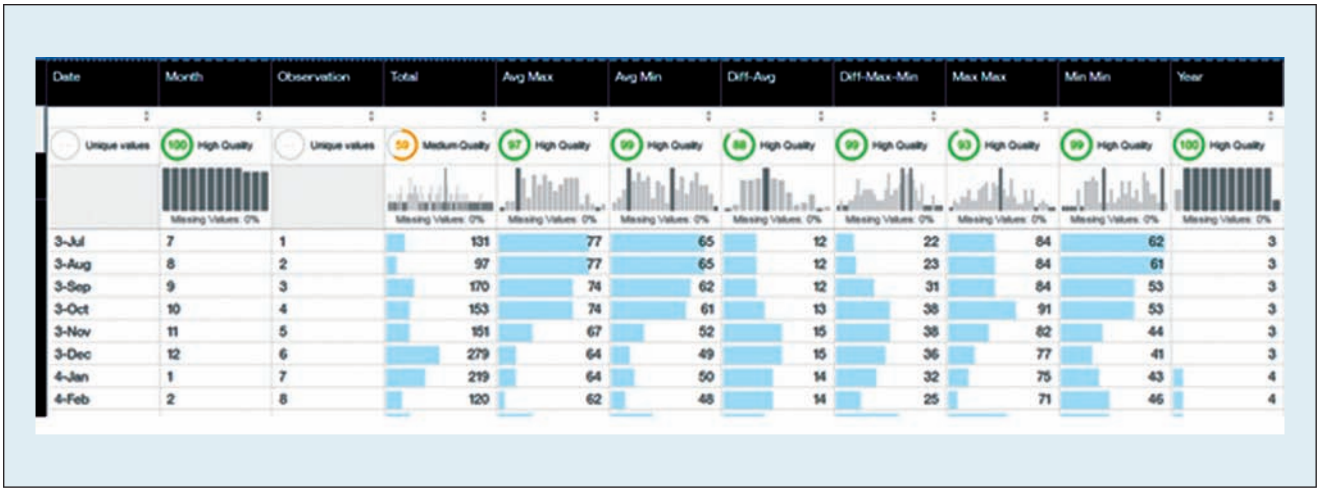
Case Study 2: Water Leak Data

As a second case study, I uploaded data from a study of the impact of temperature on the total number of water leaks in Los Angeles.[3] The data included the observation number, year, month, total number of leaks in the month (total), and several monthly temperature measures: minimum (min), maximum (max), average (avg), difference between the maximum and minimum (diff-max-min), and difference between the monthly average maximum and minimum (diff-avg). Figure 1 summarizes some of the data, along with data quality and reliability measures on each variable. In addition, the system provides visualizations of the data, first in an overview of the data as a whole, and then for individual observations.

Watson Analytics generates a data reliability measure that can help users consider the data's quality before they make assertions about the data. The system notes that one of the data elements is a "unique value" and does not assess it (it is an index value). The data for the number of the month is given a score of 100, whereas "total" is given a score of 59. Although users don't directly receive the reasons for the data's low reliability, they could conceivably use the ratings to massage the data to generate a higher reliability rating, which could be appropriate and could drive different decisions or data. For example, the user could remove the more extreme observations.

After the data was input into Watson Analytics, the system came back with the following (starter) questions:

- What drives diff-avg?
- What is the predictive model for diff-avg?
- How do the variables of diff-avg compare by month?
- How are the values of diff-avg and total related?
- What are the most common values of month?



Date	Month	Observation	Total	Avg Max	Avg Min	Diff-Avg	Diff-Max-Min	Max Max	Min Min	Year
Unique values	100 High Quality	Unique values	59 Medium Quality	97 High Quality	99 High Quality	88 High Quality	99 High Quality	93 High Quality	99 High Quality	100 High Quality
	Missing Values: 0%		Missing Values: 0%	Missing Values: 0%	Missing Values: 0%	Missing Values: 0%	Missing Values: 0%	Missing Values: 0%	Missing Values: 0%	Missing Values: 0%
3-Jul	7	1	131	77	65	12	22	84	62	3
3-Aug	8	2	97	77	65	12	23	84	61	3
3-Sep	9	3	170	74	62	12	31	84	53	3
3-Oct	10	4	153	74	61	13	38	91	53	3
3-Nov	11	5	151	67	52	15	38	82	44	3
3-Dec	12	6	279	64	49	15	36	77	41	3
4-Jan	1	7	219	64	50	14	32	75	43	4
4-Feb	2	8	120	62	48	14	25	71	46	4

Figure 1. Water leak data quality.

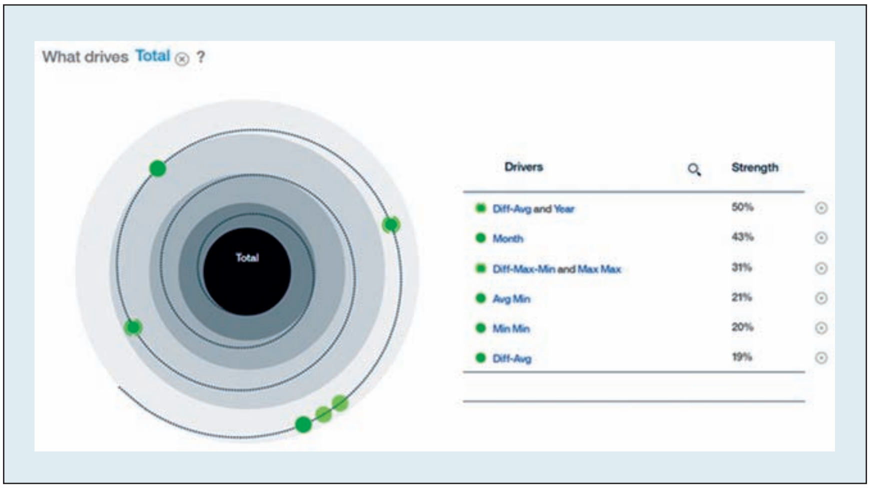


Figure 2. Initial Watson analysis of "total" (the number of leaks).

Unlike in the first case study, in which the system found our particular variable of interest (satisfaction), in this case, limited attention is given to "total" (that is, the number of water leaks) and the relationships with temperature. Instead, these questions show primary interest in diff-avg, rather than our concern for the total. Some of Watson Analytics' questions would probably rarely be asked by a human investigator (for example, "What are the most common values of month?"). This illustrates that Watson Analytics apparently does not employ any semantic understanding of the variables. As a result, its findings do not consider the semantic reasonableness of the relationships generated. This can result in apparent surprise findings, both interesting and uninteresting.

If the system does not generate the questions that the user wants answered, the user can generate new questions. Thus, I asked the system, "What drives total?" and "What is the predictive model of total?" In response to the first question, the system created an interesting visualization of a spiral graph (see Figure 2). As in the first case study, the system provided multiple sets of one and two variables in response to the question. The variables are listed by a measure of "strength."

Table 3 summarizes the Watson strength measures and the R-square values from the regression models built with the corresponding set of variables. Unlike in the first case study, strength and R-square are not statistically significantly correlated for this data.

I also analyzed the findings associated with the question, "What is a predictive model of total?" Watson Analytics developed a decision-tree model with a predictive strength of 10.9 percent that employed only a single variable diff-avg. In previous work,[3] I built regression models with the month and year as control variables. Given those two control variables, the best three-variable model was with diff-average, a model with an R-square of 0.3078, and not found by Watson. Whether or not the control variable approach was appropriate, it does not seem that Watson automatically recognizes the notion of a control variable. In addition, I also found an overall best model of year, month, min-min, max-max, and diff-avg, with an R-square of 0.3513, which was different than Watson's predictive model.[3]

Unfortunately, Watson Analytics does not provide information as to whether the variables are positively or negatively related to the number of leaks as part of the prediction.

Table 3. What drives the total number of leaks?

Independent variable sets*	Strength (%)	R-square	ANOVA statistical significance
Diff-avg and year	50	0.3007	< 0.0001
Month	43	0.0115	< 0.0001
Diff-max-min and max-max	31	0.1308	< 0.0001
Avg-min	21	0.1089	< 0.0001
Min-min	20	0.1307	< 0.0001
Diff-avg	19	0.1623	< 0.0001

*The dependent variable is "total."

The importance of the relationship between the variables is illustrated by the potential concern of the original analysis, in which the city of Los Angeles claimed that the number of leaks was decreasing over time. This would result in a negative coefficient on the year variable, and users would be interested in that finding, because it indicates that the number of leaks is decreasing over time, as the Department of Water and Power suggested.

Case Studies' Strengths and Limitations

In this short investigation, my analysis has focused on two questions asked by Watson Analytics: "What drives X?" and "What is the predictive model of X?"

This analysis revealed several strengths of Watson Analytics that probably apply to other similar analytics software as well. First, such software can provide analytical analysis to virtually anyone who can upload data to the cloud. Second, the system clearly can generate interesting questions. In some cases, analysis seems likely to bring some surprising relationships to the user's attention. Third, the system considers the data quality, a variable that is often ignored. Fourth, increased use of analytics through systems such as Watson Analytics is likely to impact a manager's need to understand such analysis, and could lead to more detailed statistical analysis of a broader range of issues. Fifth, Watson is particularly effective in those settings in which some discrete variables have a threshold value(s) that needs to be managed.

However, there are also limitations. Watson Analytics likely frames the data for the user, which could lead the user to form initial conclusions that are not appropriate under greater scrutiny. Furthermore, Watson Analytics does not focus on how independent variables drive the dependent variable, either positively or negatively. As a result, although users might understand which variables predict or drive, they may not know in which direction this occurs.

Other limitations include the apparent unavailability of classical statistical tests, an inability to designate control variables, and an inability to choose particular statistical methods. In addition, the modeling approach limits the number of variables (or at least the user's perspective) used in estimation to one or two variables. Although this approach does make it easy to understand the resulting prediction as to what drives a variable, it potentially ignores larger sets of variables. Furthermore, Watson prediction models could be developed on the basis of multiple variables or not, as both case studies showed. From a statistical perspective, Watson Analytics does not seem to account for multicollinearity or endogeneity.

Finally, Watson Analytics does not seem to employ semantic understanding. Perhaps the user could be asked to provide information about the variables to facilitate the analysis. Alternatively, the system could use natural language and try to understand what the variables represent, providing a glimpse at a potential theoretical structure. For example, a person seeing a variable labeled "month" or "year" would likely infer knowledge about that data. As another example, in an analysis of another dataset, Watson Analytics found that "total time" was related to the pair "start time" and "finish time," providing limited insight to the users.

In the movie *The Incredibles*, the villain Syndrome hired Mr. Incredible (a superhero) to unknowingly design an almost unbeatable foe. As Mr. Incredible found limitations in Syndrome's robot creations, Syndrome would eliminate the limitations and come up with a more formidable and, ultimately, a seemingly indestructible robot. I expect that Watson Analytics (and other smart and easy analytics) will take a similar evolution as those robots, evolving from its current incarnation to increasingly more capable versions as people find strengths and limitations associated with the software. I expect that Watson Analytics will begin to embed additional analytics and statistics knowledge. Furthermore, I expect some users will begin to integrate semantic knowledge and domain expertise to further build the capabilities of Watson Analytics to provide users with a domain-guided analytics experience.

References

1. *Introduction to IBM Watson Analytics Data Loading and Data Quality*, IBM, Mar. 2016.
2. "IBM Watson Analytics Workshop Reference Guide," IBM, May 2016.
3. D.E. O'Leary, "Modeling Los Angeles Water Leaks Using Different Measures of Temperature," *Advances in Business and Management Forecasting*, K.D. Lawrence and R.K. Klimberg, eds., vol. 11, 2016, pp. 187–207.

Daniel E. O'Leary is a professor in the Marshall School of Business at the University of Southern California. Contact him at oleary@usc.edu.