

# From Big Data to Big Analytics: Automated Analytic Platforms for Data Exploration

Richard Timpone, Ipsos  
Jonathan Kroening, Ipsos  
Yongwei Yang, Google

October 26, 2018

## Abstract

As *Big Data* has altered the face of research, the same factors of *Volume*, *Velocity* and *Variety* used to define it, are changing the opportunities of analytic data exploration as well; thus, the introduction of the term *Big Analytics*. Improvement in algorithms and computing power provide the foundation to produce automated platforms that can identify patterns in analytic model results beyond simply looking at the patterns in the data itself.

Introducing the class of Automated Analysis Insight Exploration Platforms allows conducting tens and hundreds of thousands of statistical models to explore them to identify systematic changes in dynamic environments that would often be missed otherwise. These techniques are designed to extract more value out of both traditional survey as well as Big Data, and is relevant for academic, industry, governmental and NGO exploration of new insights of changing patterns of attitudes and behaviors.

This paper discusses the architecture of our Ipsos Research Insight Scout (IRIS) and then provides examples of it in action to identify insights for scientific and practical discovery in public opinion and business data. From the Ipsos Global Advisor Study we show examples from the U.S. withdrawal from the Paris Agreement and the 2016 presidential election. We then show with an example how a research project at Google is leveraging these platforms to inform business decision-making.

Jim Gray posited that the techniques and technologies of modern data-intensive scientific discovery are so different from earlier generations that it represents a new Fourth Paradigm of eScience for Scientific Discovery (Bell 2009, Hey et al. 2009). In this context, he argued for the development of better tools for data capture, curation, analysis, and visualization. While this covers the breadth of the research endeavor, much of the focus in this domain has been on the massive expansions of opportunities around Big Data that have exploded since Gray's original presentation. This paper focuses on developing a tool for the analytics in this paradigm to consider the scaling of analytic processes to match the scaling of the data itself (Goble and DeRoure 2009).

Paper prepared for presentation at the BigSurv 2018 Conference of the European Survey Research Association; Barcelona, Spain.

Rich Timpone is Senior Vice President and Head of the Ipsos Science Center at Ipsos. Jonathan Kroening is Vice President and leads the R&D team in the Ipsos Science Center at Ipsos. Yongwei Yang is a Survey Research Scientist at Google, Inc.

As exemplified in the ideas of the Fourth Paradigm, advances in the types and volumes of data that are being collected and stored, and the computing power to leverage them have fundamentally changed the face of research and brought us into a second machine age (Brynjolfsson and McAfee 2014). While there remain huge opportunities in the continuing growth of Big Data available to us, this paper focuses on opportunities to leverage automation to move from data sources (from Survey to Behavioral, machine generated and others) to the creation of statistical and analytic models in volumes that go beyond the normal ability to examine. These analytics are of many varieties (from distributional changes to time series models) and are produced at a velocity that can be overwhelming. By extending the idea of the 3 V's often used to describe and define Big Data (Beyer and Laney 2012), we see this application of automated platforms creating a comparable expansion in the form of *Big Analytics*.

While integrated into the notion of a data-intensive Fourth Paradigm of research, the impetus for considering the automated analytic platforms described in this paper was the challenge that many researchers and practitioners have had in large tracking studies where many respondents are surveyed on a regular basis on topics from global political attitudes to commercial brand preferences. Given the large volumes of information in major programs that cover multiple countries and multiple categories and types of products, some researchers and practitioners concentrate on key indicators to produce focus and help simplify the challenges.

While North Star measures are inherently valuable (c.f. Doerr 2018), focusing solely on them may create perceptions that attitudes and behaviors may be more stable over time than they actually are. In fact, beyond the new opportunities they produce, some of the dynamics pushing away from survey data to behavioral, social listening and other Big Data sources has been to provide more responsiveness to understand the dynamics that are actually occurring over time.

Rather than focus solely on specific measures or data types, the approach here is to conduct multiple models across as much data as possible to identify where changes are occurring.

The automated platforms we introduce here have a very specific purpose to identify changes in data over time. Whether we are interested in understanding the behavior of societies, markets or people, understanding what is changing allows tracking progress on specific programs, monitoring the dynamics of our data over time and across contexts, as well as identify risks and opportunities from public policy to business decision-making. While focused on practical problems, the identification of changes in broad patterns of data also can provide a data and analytic driven basis for hypothesis development to advance more fundamental scientific discovery.

Rather than traditional hypothesis driven analytic models, these analytic platforms allow for a more inductive exploration of a wide number and broad types of analytic models to identify where and when changes are occurring to drive data-intensive scientific discovery (Gannon and Reed 2009). We envision these platforms as augmenting traditional hypothesis driven research by identifying changes in vast sets of data over time more rapidly and thoroughly to reveal dynamics that could be missed otherwise and provide teams an opportunity to act on the information. Thus, these platforms augment the advances in Big Data and the types of information now being captured along with that from traditional survey research sources as well.

The remainder of this paper provides details on our approach to the idea of Automated Analytic Insight Exploration Platforms with discussion of the architecture of our implementation, the *Ipsos Research Insight Scout* (IRIS). Beyond that, we share empirical examples from public opinion and business settings to demonstrate the insights and opportunities that these new platforms provide.

### **Conceptual Overview**

Automation has fundamentally altered the progress of humanity from the industrial age through the machine and information ages. While the scientific method has led to the fundamental changes that have driven growth in human knowledge enabling these epochal advances, the expansion of Big Data and computational power have accelerated the possibilities for even greater data-intensive scientific discovery.

### **The Fourth Paradigm**

Friedman's (2016) description of what he dubs the motto of Silicon Valley, highlights the transition that has led to the era of the Fourth Paradigm:

“everything that is analog is now being digitized, everything that is being digitized is now being stored, everything that is being stored is now being analyzed by software on these more powerful computing systems, and all the learning is being immediately applied to make old things work better, to make new things possible, and to do old things in fundamentally new ways.”

The addition to Friedman's discussion of the progression is the scale of Big Data that has led to Gray's description of eScience and the Fourth Paradigm.

Each of the areas of data capture, curation, analysis and visualization, for scientific tool improvement have seen advances in the past decade. While data management in the era of Big Data is a major domain, the focus on this has often overshadowed the opportunities for the analytic tools. Even Gray's own summary of this dimension in eScience seemed more focused on expanding traditional analysis to the scale of the data describing where a “scientist analyzes database/files using data management and statistics” (Hey et. Al 2009, p xx).

While improvement in algorithms to handle the large volumes of Big Data is necessary, the ability for mining the data and the masses of models that can be conducted, highlights the opportunity for new approaches of insight and scientific discovery. As Goble and DeRoure (2009, p 144) noted “people speak of a data deluge, and data-centric science could be characterized as being about the primacy of data..., but it brings with it a method deluge.”

This method deluge is a consequence of advances in analytic models and their automation as well as in computation power. Scaling to make sense of the massive amount of analytic models and outputs, as a component of *Big Analytics*, parallels the scaling and 'plethora' of data themselves in Big Data driven scientific discovery (Gannon and Reed 2009). Shifting from traditional hypothesis testing to mining of the massive amount of modeling outputs provides not only practical insights but also the foundations for generating new hypotheses. Tapping this potential in the analytic space is enabled by the expansion in the development of automated analytics.

### **Automation and Big Data**

Given the ubiquity of computers in modern research, we need to be more specific when we talk automation as this could reference anything that is assisted with these devices from analytics to data visualization to the report production that people used to do manually (Adams 2018). However, when we generally are talking automation in research, and especially when people refer to automated platforms they are most frequently discussing cumbersome, complex or repetitive processes that can be done more quickly, efficiently, accurately and cost effectively than manual processes (Fawson 2018, Simmons 2018). While all very valuable, these

advantages of automation are operational but do not necessarily create additional substantive value or expand the scientific discovery process (Phillips 2018).

The idea of Automated Analytic Insight Exploration Platforms goes beyond operational advantages to explicitly enable scientific discovery through data, and specifically statistical model intensive exploration. Moving from a few key measures of interest to leveraging the masses of data now available allows discovering extremely broad patterns of movement in our data and understanding more about them in context.

While a decade ago we were introduced to the age of the petabyte, we are already in the era of the zettabyte where over a zettabyte of information is transferred over the internet annually and the entire global datasphere is projected to grow to 163 ZB by 2025 (Reinsel, Gantz, and Rydning 2017). To get a sense of the scope of this, since one gigabyte can store 16 hours of music, a zettabyte would be the equivalent of about 2 billion years of music (Seagate blog), or to visualize this, if an 11 oz. coffee mug represented the same initial gigabyte of data, a zettabyte would be the volume of the entire Great Wall of China (Savov 2011).

While Big Data expands the limitations of traditional means of data exploration, even large survey programs can overwhelm research teams. This is most apparent with large cross-national surveys that are longitudinal in nature. In one example shown later in this paper we demonstrate how, in a survey program at Google, we conduct over 30,000 statistical models in a round of analysis to explore the questions of over time change alone and expect this to expand to over 100,000 models per period in time.

### **Insight Exploration Statistical Modeling**

The purpose of the automated analytic platforms we are discussing here are intended to identify areas of change in patterns of data over time. This entails examining many variables, often cut across different groups (countries, companies, groups of people etc.). By focusing on data changes over time, we can identify areas worthy of further exploration. Given our goals are often practical applications of the data, this can identify changes that could be missed if one only focused on high level key variables of interest.

It is worth emphasizing the fact that data can serve different purposes is relevant to the questions at hand. Doerr (2018) demonstrates the power of focusing on a few Objective and Key Result indicators for ensuring large organizations, like Google, focus on the same important issues from a management and strategy perspective. For these purposes, OKRs can be incredibly valuable for effective focus and execution. While OKRs may provide great value in a number of domains, their role in *exploratory* scientific discovery is more questionable.

A counter perspective of this, for the role in exploration, is Tenner's (2018) argument that the risk of hyper-focus is an area where efficiency challenges in Big Data can lead to missing broader patterns. While tracking OKRs is useful for monitoring progress, expanding explorations to much broader sets of data will help identify many dynamic changes potentially occurring. From a scientific discovery and practical application perspective, exploring a broad set of data will ensure the fuller context is understood which will potentially be missed by the focus on a handful of, even very important, measures.

Given this approach of testing numerous variables using many statistical tests for an analytically intensive exploration, we are shifting to exploratory data mining from traditional hypothesis testing. In the context of the Fourth Paradigm, "[w]e are in an era of data-centric

scientific research, in which hypotheses are not only tested through directed data collection and analysis but also generated by combining and mining the pool of data already available (Goble and DeRoure 2009, p 137). Thus, identifying patterns of change may identify practical opportunities directly or serve as the foundation for developing new hypotheses for examination and testing.

The explorations of the many models that we have described include a number of different types of analyses. Over two time periods, we have found that differences in means and proportions, as well as differences in distributions can provide insights. Additionally, as more time periods are captured, these explorations can be expanded with linear and non-linear trend investigations as well as alternatives to statistical analysis (parametric and non-parametric) with rule-based approaches like Control Charting. This demonstrates the flexibility of such analytic platforms in leveraging parametric, non-parametric and rule-based tests from the newest advances in Data Science to those 100 years old.

With a variety of types of models that address different temporal aspects of the data, conducted over vast numbers and types of data series, we are often generating tens of thousands of statistical models in a short period of time. This creates model results that differ from traditional models by their *Volume*, *Variety*, and *Velocity*. Thus, just as these dimensions were originally identified to define Big Data, we consider their application to these automated platforms as distinguishing these from other models as *Big Analytics*.

Given the mass of models that are produced in these data and analytic intensive discovery explorations, the risk of false positives in the statistical tests produced and examined must be addressed. In the implementation that we discuss in more detail, we leverage the False

Discovery Rate approach to balance the desire to separate the signal from the noise in the data while not throwing the baby out with the bathwater (Hastie et al. 2009, Storey 2010)

The intent of these platforms is to identify clusters or patterns for scientific and practical discovery of temporal ‘hot spots’ of change in the data. Thus, our focus on them for a deeper exploration of the data and analytic models. While this could generate specific hypotheses that could be examined in subsequent hypothesis driven analytic tests following the traditional scientific method, the nature of the platform also allows additional exploration within its framework as well. Since the platform has all of the tests, identifying one cluster of interest allows comparing to other aspects in the data as well as the substantive magnitudes.

For example, if the platform identifies a ‘hot spot’ area of change in public opinion attitudes in a particular country, the platform could identify if this is part of broader attitudinal changes within the country by examining other dimensions of public opinion, or if the views in that country are part of a broader pattern of change across regions or the globe. Beyond the contextual insights provided by these comparisons, knowing which tests are flagging a ‘hot spot’ is relevant as well. If something occurred as a change from a previous month or quarter vs. being part of a long-term trend could be relevant for determining the nature of the temporal change (whether enduring or not, abrupt and permanent or returning to the previous level, or a new normal over time) (Cook and Campbell 1979; McDowell et al. 1980).

While the exploration begins by focusing on broader patterns of significance across the vast volume of tests (at different levels of False Discovery Rate), the substantive magnitude of the changes themselves can be explored in these platforms. Thus, even if two different countries both saw their citizens’ attitudes changing significantly over time, the magnitudes can be explored directly in the platform. This is true for market research or other studies where broad

patterns may affect different companies in a category differentially. This was the case with the impact of the Cambridge Analytica scandal in the tech sector.

Thus, the value of insight exploration platforms fits into our broader view that Big Data needs theory (Timpone 2016) by producing new hypotheses and means for valid exploration. In previous work we have argued that to identify new insights that are relevant and real, when leveraging Big Data, researchers need to consider the nature of the research questions along with statistical considerations (like sample representativeness and generalizability) and the nature of the analytics. Considering a subset of these can lead to findings that are of questionable value or accuracy as the framework in Figure 1 summarizes. While uncovering patterns in a data driven way allows identifying new areas that these dimensions can produce extremely valuable scientific insights.

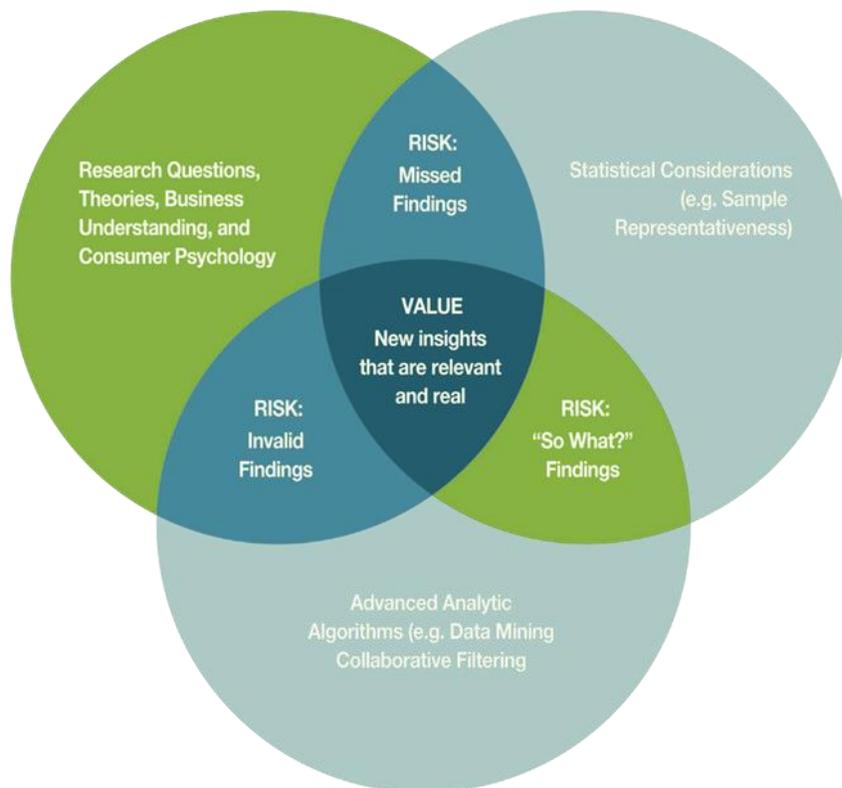


Figure 1: Big Data Research Consideration Framework

### Architecture of Automated Analytic Insight Exploration Platforms

Moving from the conceptual overview behind our ideas for these automated analytic insight exploration platforms to practical application, we transition to a more applied summary of the architecture involved. While some of this will be theoretical, as there are multiple approaches that could be leveraged, we will illustrate the principles with the platform that we have developed and leveraged for multiple programs, *IRIS- the Ipsos Research Insight Scout*. Given the exploratory nature of these platforms, we feel the idea of an insight scout is an appropriate appellation for our system.

The architecture of our Automated Analytic Insight Exploration Platform is divided into three broad phases. First is the preparation of the data and the need to specify the metadata needed for insights. Following this is the actual automated analysis of all of the data sources prepared in the first phase. Finally, is the extraction of the insights and the scientific discovery which is the purpose of the platform itself.

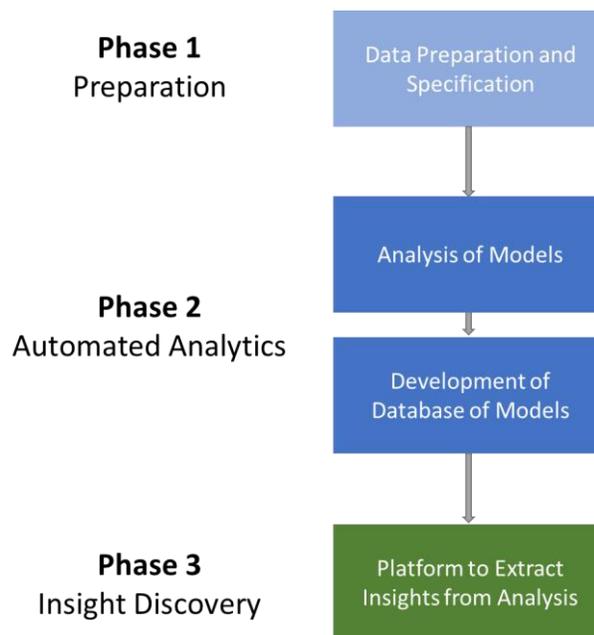


Figure 2: Phases in Automated Analytic System

**Phase 1- Preparation**

The first phase of a program requires setting up the data for analysis and developing the specifications for the metadata that will be needed to link the different data sources and variables. This stage allows the usage of different types of data including both survey and Big Data sources. We believe there is significant merit in each type of data and “do not see Big Data as a simple alternative to traditional methods like surveys. Our view is that the different types of data possess distinct merits for strategic and tactical questions. Depending on the question, these can be employed separately or in coordination” (Timpone 2016, p 9).

The substantive focus of these platforms for exploring broad numbers of data streams to identify areas of change and continuity over time make the data preparation a simpler approach than for many Big Data projects. The reason for this is that for efficiency and tractability, the many analyses conducted in phase 2 are most effectively done at the aggregate level. This is true whether analyses are done for a population of interest overall or specific subgroups of it.

For each timepoint in a data series, weekly, monthly, quarterly etc., the aggregated summary is produced. From a survey perspective this could mean that in each wave of a tracking study or each longitudinal time in a survey like the Eurobarometer, General Social Survey or American National Election Study the average or top 2 box, bottom 2 box or some other value would be calculated and used in the following phases.

For Big Data sources we would again calculate the aggregate measure for the time period and leverage that in the analysis. These could be metrics calculated from consumer generated data in social listening, summaries from sites like Google Search, output of sensor data, economic, sales, or other aggregated measures. Thus, we are able to use information ranging from economic indicators from the World Bank or U.N. (when investigating national level

questions) as well as Big Data sources directly. While some require text and other analysis to summarize, focusing on aggregates avoids issues of tractability as we move to large volumes of analyses of different types.

In addition to the raw data preparation, specification of metadata must be captured as well to allow linking the different data streams for the pattern identification of the insight ‘hot spots’ of interest. This is true within surveys as well as Big Data sources, i.e. data source, country, variable, company information. In addition to the data information, metadata will be appended in Phase 2 so that insights could be identified based on whatever aspect of analysis is of substantive interest, i.e. by country, for a specific question, like trust, for sales, or by type of analytic test conducted in Phase 2.

### **Phase 2- Automated Analytics**

Not surprisingly, the heart of our Automated Analytic Insight Exploration Platform is the automated analytics that are conducted on the data. Given the setup described above, it should be clear that the analytics themselves can be fairly straightforward. The current *IRIS* platform includes wave over wave tests (tests of distribution change as well as mean and proportion differences) and longitudinal tests (linear trends, non-linear trends, and control charts). That said, moving to conducting tens of thousands, projecting to expand to more than 100 thousand in each time period in the future, make the need and nature of the automation obvious.

Many of the analyses are straightforward so we will not go into difference of means and proportions and even the linear and non-linear regression functions to identify significant changes over time (across 2 or more time points as appropriate). These can be applied using the

methods from standard statistical references (Greene 2003; Gujarati 2003; Wonnacott and Wonnacott 1990).

While also a basic technique, taught in first introductory statistics classes, we did want to discuss the distribution tests briefly though. This is less because of any technical aspect but rather to highlight the value of the insights which may be less apparent in these examinations. In some cases, the distribution test will be similar to the difference in means, but not always.

Figure 3 shows three different ways that distributions of a variable could change over time. In the first case a distribution shifts as values shift lower (or higher) on a scale used. In this case changes of means and distributions will both reflect similar patterns.

The other examples in Figure 3 though may not show up in mean differences but are meaningful insights producing different types of discovery. In one case we see a shift in kurtosis as values either flatten out or become more concentrated (Gujarati 2003). In the final case, the distribution could become polarized or bimodal. Either of these cases may not be identified by a focus on central tendencies alone and so this is an important addition to both the mean tests but also the tests, like the longitudinal ones, that are based on aggregated values for each timeframe. Given that some distributions change more slowly, this test is valuable not only at 2 consecutive time points but potentially over longer periods as well (say over a year or more).

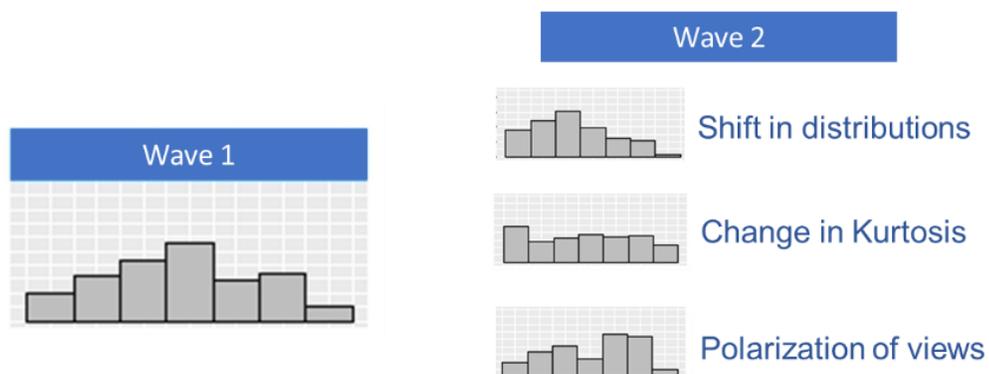


Figure 3: Potential Shifts in Distribution Across 2 Timepoints

This is not simply a theoretic consideration. The polarization of attitudes is a substantial aspect being seen in political views in a number of countries around the globe including the United States as well as across Europe. This growing tribalism in views is creating divergence from the traditional single central tendencies we are seeing in many sets of attitudes (Young 2018).

Figure 4 shows a practical example of this type of distributional shift. This shows the ideological views of members of the United States Senate (upper legislative body) over time. The image on the left shows the average ideology over a period of nearly 70 years. Across this timeframe, the mean ideology is far more stable than the dynamic political views would seem to show (Poole and Rosenthal voteview.com data compiled by brookings.edu). The right 2 charts however show that from the 1970s through earlier this decade the ideology of individual members became more polarized (disaggregated voteview.com data compiled by Pew Research). This is not only true among elites, but the population at large as well. It also is not limited to politics as trust in institutions, companies and other areas have shifted more fundamentally over time as well (Young 2018).

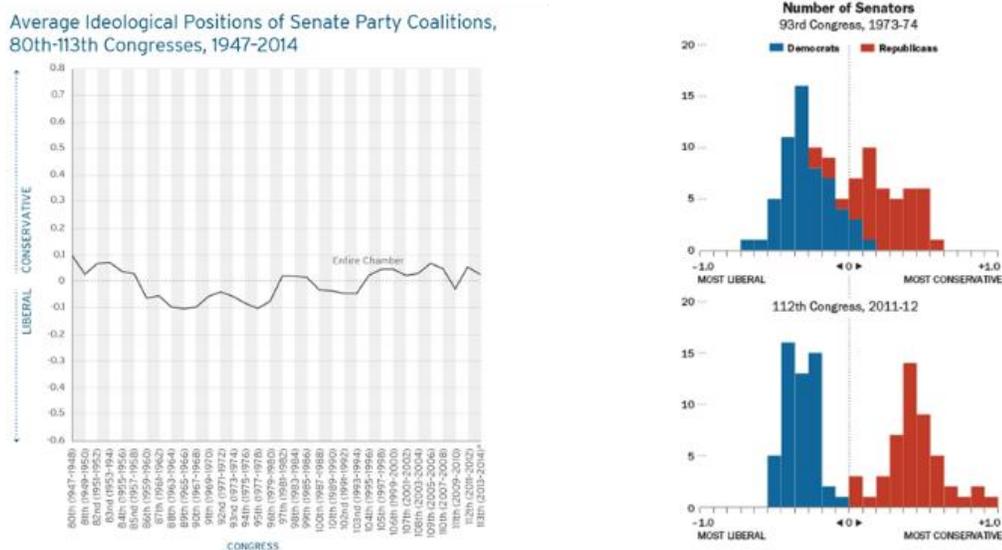


Figure 4: Ideological Views of U.S. Senate Over Time

At an initial level, we conduct the various tests at a 95% confidence level, but the need for multiple test correction is obvious in this case. From a data mining perspective addressing the risk of false positives is absolutely necessary. Rather than select a specific level with family Wise Error rate corrections (which are extremely conservative) or False Discovery Rate (Hastie et al. 2009; Storey 2010), we calculate a q-value score for each test in IRIS. This is the minimum FDR that can be attained when the test is considered significant. In other words, if that test is flagged as significant, its associated q-value is the proportion of false positives one would expect to find in all flagged tests with a lower q-value than that test. This value allows moving from using a fixed cut-off for examining the robustness of the results. In this case, we leverage a slider which allows going from 95% confidence to more restrictive levels to determine what level of balance is desired for the substance and noise elimination. This allows identifying those patterns that are observed that are robust as a fundamental feature in the platform itself.

### **Phase 3- Insight Discovery**

Once the database of all of the statistical tests with the metadata is captured and stored, mining it for substantive insights is the heart of the discovery process. This can be done globally, across all models run, or targeted to identify those specific to a subset of data to facilitate discovery within specific countries, for specific companies or targeting specific topics. In addition to different levels of discovery, the methods that can be applied to this stage in the architecture framework vary as well. This insight generation can be done in increasingly sophisticated ways from smart summarization to rule based, expert system, insight identification to machine learning and deep learning algorithms.

The most common use case of the IRIS platform is to analyze patterns when a new round of data is collected (weekly, monthly, quarterly etc.). In this case, the data stream patterns for the

tests are examined for clusters of change. While this is the most common use case, in our public opinion example in the next section we show both that approach as well as one where mining analytic results across time periods can also be useful for practical insights and scientific discovery.

From a summarization perspective, *IRIS* has a dynamic interface that allows focusing on specific topics of interest. In our public opinion data for instance, we can focus on a specific domain of issues, such as terrorism or economic concerns, and see when and where citizenry around the globe was exhibiting significant changes in their views. Likewise, in market research data, we could focus on areas of concern and opportunity for a specific brand across the globe, attitudes for a category within a specific market, views of a specific topic, such as trust as we leverage the platform for different types of insight generation.

Beyond the platform's use for targeted exploration, the totality of the models conducted could be summarized across all of the combinations of the dimensions of metadata captured in Phase 2 of the platform. In *IRIS*, we currently look at all combinations of up to 7 dimensions of metadata such as country, platform of data, brand, product, etc. Since not all combinations are mutually exclusive, we have the ability to examine the simplified unique summary as well as the full combinatoric landscape (to address the fact that deeper cuts and combinations are likely to see more variation than top level summaries such as by country or brand alone (across all countries)).

While such smart summarization is richly valuable, the potential to extend machine driven exploration will add an additional level for more regular exploration and reporting. We envision rule-based approaches, which would be the equivalent of older school expert systems where the platform acts as an automated research assistant expert to be an extension as rules by

teams are formalized (Benfer et al. 1991; Taber and Timpone 1994, 1996). Beyond this, full deep learning models are being explored to identify those ‘hot spot’ areas of interest. The main challenge to this is the development of training sets for the models to learn from. It is for this reason that we think considering tools from earlier in the evolution of the discipline of Artificial Intelligence are still worth considering today (see Dormehl 2017 for a brief history of the discipline of AI).

While more and more of the process and insight identification will be automated in the future, from the perspectives of Scientific Discovery and practical insight identification, we envision this to be a domain of *Human + Machine* rather than a simple transition from one to the other (Daugherty and Wilson 2018). In addition to training, people will continue to play critical roles in these domains even as machine learning and AI moves to generating hypotheses and insights. In this way we see these platforms identifying new insights and serving as Intelligence Amplification tools in the research process (Kasparov 2017).

### ***IRIS* in Action- Empirical Case Studies of Automated Platform**

Moving from the conceptual overview and discussion of the architecture of these automated analytic platforms to a few practical examples allows us to demonstrate how such automation moves beyond operational efficiencies to the ability to identify new insights. In this way, they can leverage a variety of data sources and analytics to play a new role in scientific and insight discovery.

The following two sections provide examples from public opinion and business settings. The first is from the Ipsos Global Advisor survey of global attitudes, and the second is an

example from the work being done with Google to understand changing business risks and opportunities.

### **Public Opinion Data- Ipsos Global Advisor**

The Ipsos Global Advisor program is a 28 country, monthly survey-based research program that provides views of societal attitudes of interest to media outlets, clients and academics. This includes attitudes across a number of domains including financial, social, environmental, health and terrorism. The following two examples leverage Ipsos Global Advisor data from March of 2010 through September of 2017, and automated analytics were conducted across all of the substantive questions in the survey for each time period. Here we see that the data has been aggregated for different periods from year over year to quarter over quarter, month over month and specific points in the year.

In the first example, we examine the platform as described demonstrating the most common use case where it would be examined at a specific point in time. This would be as if we had just acquired the data and were considering implications given changes in attitudes (in the case of the data examined here). In June of 2017, President Trump announced his intention to withdraw the United States from the Paris agreement of the United Nations Framework Convention on Climate Change. Here our question was whether following this event, if and where changes toward environmental issues may have been observed.

To address this exploratory type of question, we specifically look at the results of the analyses to see in what substantive categories attitudes saw movement in Q3 of 2017 as well as whether it was consistent across the globe or varied by region. For this example, we examined

quarter over quarter changes with the False Discovery Rate for the statistical tests set at 5%. Results are shown in Figure 5.

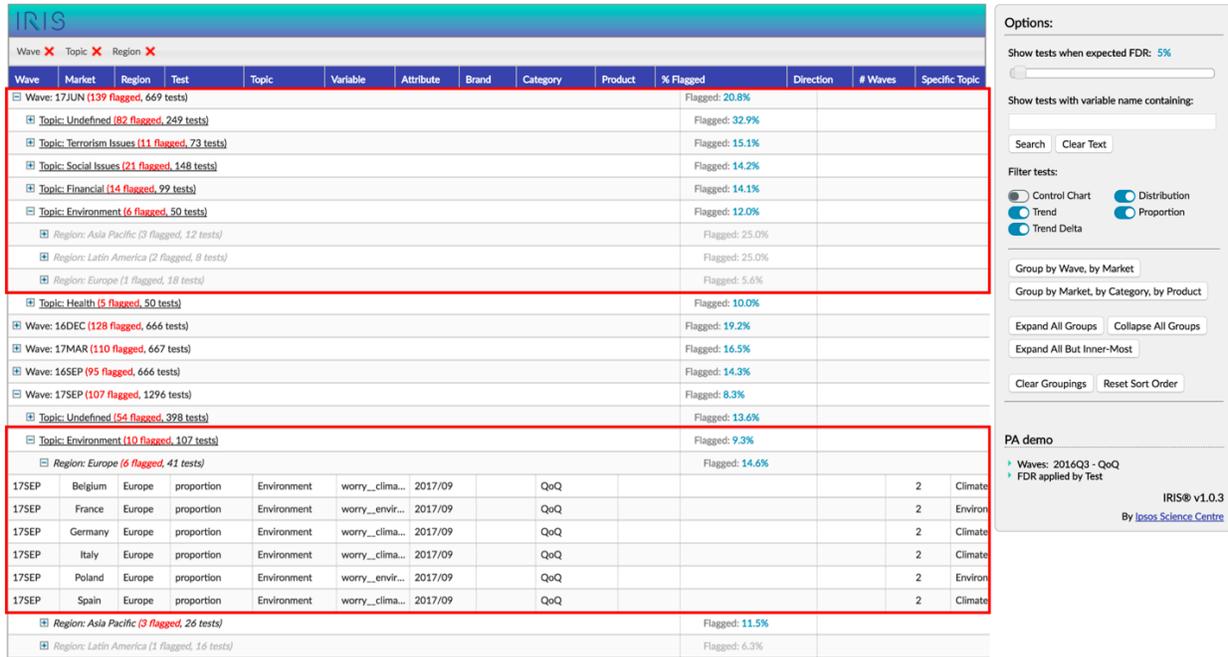


Figure 5: Changing Attitudes Toward Environmental Issues in 2017

While we are focusing on the Q3 period, it is interesting to see how the clusters changed from the previous quarter. In Q2, the proportion of tests around environmental attitudes was lower than the proportion of changes in views globally around terrorism, social and financial issues. In Q3 after the United States’ announcement, the proportion of changes around environmental issues was higher than all of these other categories.

Beyond the amount of changes seen regarding environmental issues, Figure 5 also shows how dramatically this varied by region. In this case we see that 14.6% of the tests regarding environmental concerns changed in Q3 of 2017 in Europe, followed by 11.5% in Asia Pacific, and 6.3% in Latin America. After adjusting for robustness with the FDR in this case none of the environmental variables at all were flagged in North America. Since the purpose of the platform is to identify clusters of movement, when there are fewer than 20 tests considered, the interface

greys out the results to not be overly influenced by small numbers of tests. This can occur by drilling down many levels of metadata depending on the nature of the data captured and organized.

Again, beyond the quantity of changing variables, the substance matters. For instance, of the 14.6% of tests that were flagged in Europe in this case, all of them show an increase in concerns about climate change and environmental threats. While identifying clusters is informative, along with statistical significance is the importance of substantive significance and the magnitudes of changes. Figure 6, shows one of the simple tests that was flagged in the environmental analyses and this is the proportion of people in France who felt that environmental threats were one of the three most important issues they were worried about for their country. This proportion jumped by almost 5.5ppt from 10.5% ranking this as a top 3 issue before the U.S. pulled out of the Paris agreement to 15.9% afterward.

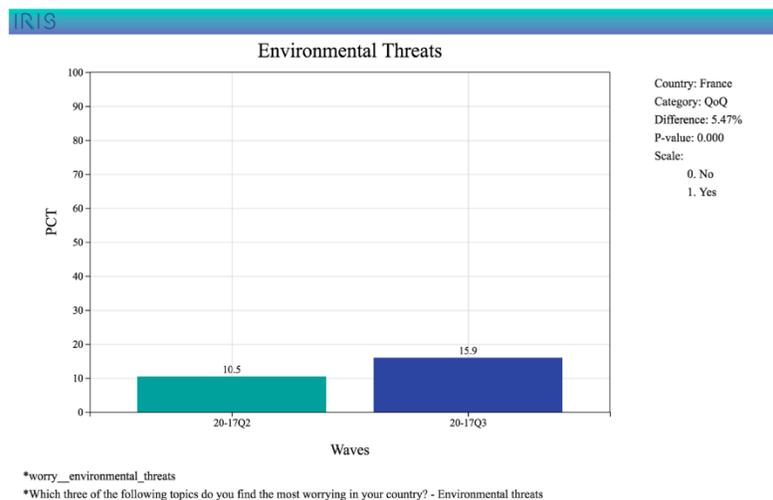


Figure 6: France 2017 Increase in Worrying About Environmental Threats

In the previous example, we leveraged the platform as it is commonly used by examining the patterns of change seen at a specific point in time. In the example that follows here, we show how including the timing of the tests as another dimension in the scientific discovery process can

be a source of new insights as well. Across all of the countries with the False Discovery Rate set at 5%, 16.9% of the quarter over quarter tests were identified as statistically significant. Figure 7 presents a screenshot of the dynamic summary of the interface of the wave over wave tests (proportions and distributions). This shows significant variation across countries with the largest amount of change over time seen in Russia with 30.5% of the tests flagged as significant, while the least was in Israel with 7.2%.

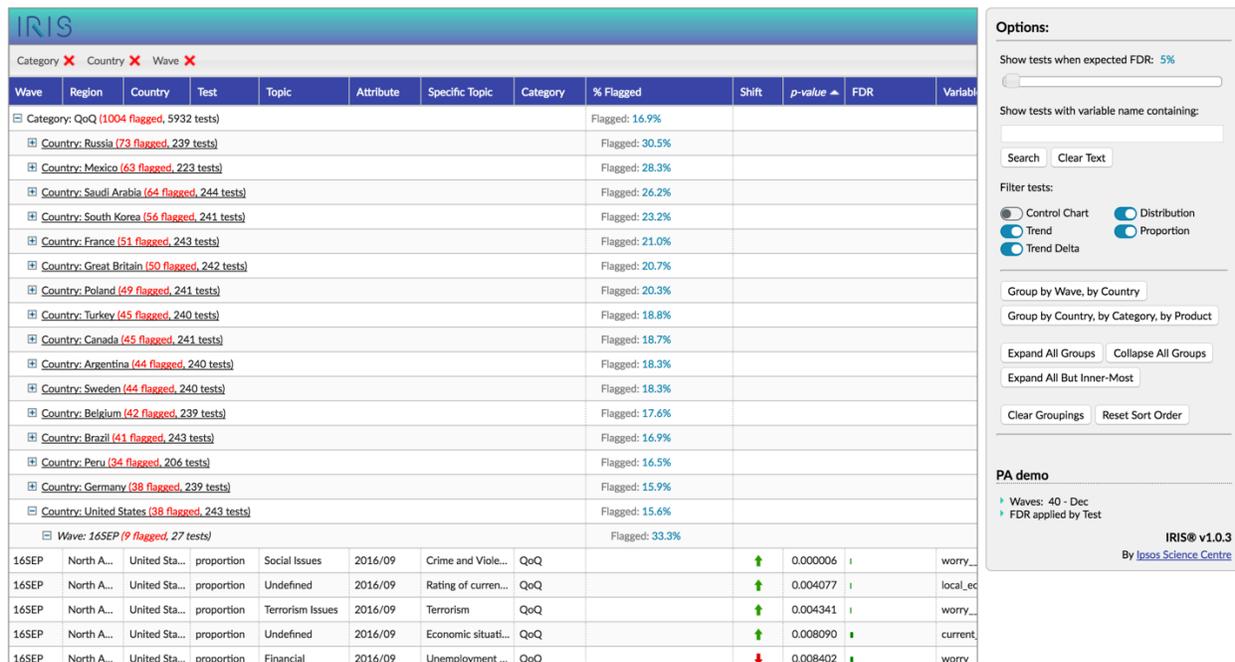


Figure 7: Changing Public Opinion Across Countries Over Time

In this exploration we see that the United States falls in the middle third of countries relative to how much significant movement is observed. But this also varies over time. The amount of movement peaked in the United States during this time in the 3<sup>rd</sup> quarter of 2016 with a third (33.3%) of the statistical tests significant even after setting the False Discovery Rate at 5%. This compares to the first quarter of 2016 where only 3.7% of the tests were significant.

While it is not entirely surprising that attitudinal change was broadest in the United States approaching the 2016 presidential election, the fact that this movement was continuing through

the third quarter demonstrates the fluidity of these attitudes in a number of domains. While a number of dimensions have been identified as key determinants of individual voters' election choice, it is interesting that a majority of the attitudes we see having moved significantly in advance of the election were regarding economic issues. One specific element that is highlighted through these two examples is that the initial approach examining clusters of movement is focused on those that we have confidence are robust movements. That said, as we saw in the environmental example, the direction and magnitude of movement is critically valuable as well. That information can be captured in Phase 2 of the process and in the case of *IRIS*, both the directionality of the effects (where appropriate) and the magnitude itself are incorporated in the platform.

Note that magnitudes are comparable within tests. The difference in proportions or means vs. the slope of a linear trend line should not be directly compared to each other, but we can examine the clusters of movement by magnitude as well so that we not only examine those changes that are significant but also see which are larger substantively as well. In our business applications this becomes a critical aspect as we often see movement by multiple brands in a category, but these magnitudes often differ which creates very different risks and opportunities.

### **Business Strategy Data- Google's Market Research Example**

While Ipsos's focus is on the total understanding of societies, markets and people and these automated platforms have potential for broad scientific discovery, much of our focus is built supporting the business needs of clients. As part of this support, Ipsos partners with a Google market research team to understand the business environment and opportunities of various brands and products.

As part of this partnership, each month the *IRIS* platform has been leveraged to identify movements in trending data from large-scale, multi-national consumer surveys. This is an important expansion over typical use of consumer survey data that often focuses on just a few high-level key performance indicators. The *IRIS* platform allows identification of patterns of change that could be easily missed with unsystematic or manual investigation – patterns that, if missed, could adversely affect a company’s ability to serve customers and meet business objectives.

While much of the data is proprietary, the timing of the program has allowed a natural experiment. In March of 2018, privacy issues were raised in light of Cambridge Analytica’s use of Facebook user’s data (Cadwalladr 2018, Rosenberg 2018). This allowed leveraging the *IRIS* platform to explicitly view how brand perceptions toward technology companies were changing in response to such an event.

In February of 2018, for instance, the percentage of questions that were flagged by the platform for Facebook was at the average level of all brands. This was at the 95% confidence level (which we use here to illustrate the multi-test correction in the post March analysis). If we change the False Discovery rate level to 10%, none of the 816 different tests calculated on Facebook attitudes globally remained significant.

Following the revelations of Cambridge Analytica in March and the negative backlash against Facebook, the story is significantly different. First, the proportion of tests that were flagged at the 95% confidence level jumped to 22.5% in April for Facebook. Additionally, not all markets saw this much increase. The U.K. saw the largest jump and the U.S., Italy, Canada, Germany and Australia also saw increases but at a lesser magnitude.

While the clusters of results again demonstrate the value of leveraging patterns identified in the data exploration of the multitude of statistical models conducted, as in the Public Opinion example, substantive relationships help clarify the fuller story. Not surprisingly, the directions of changes were uniformly negative, both when looking at immediate changes between two time points as well as with trend tests with longitudinal data. While the specific values of the analyses cannot be shared, Figures 8 and 9 show the outputs produced for measures of brand support in two different countries. These demonstrate how in addition to identifying the robust patterns of change, examining the detail helps to identify the magnitude of risks and opportunities in changing market conditions.

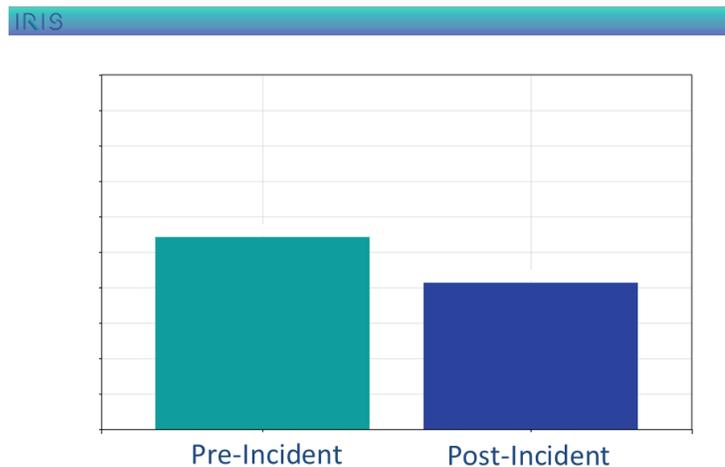


Figure 8: Changing Brand Support in Wave Over Wave Tests

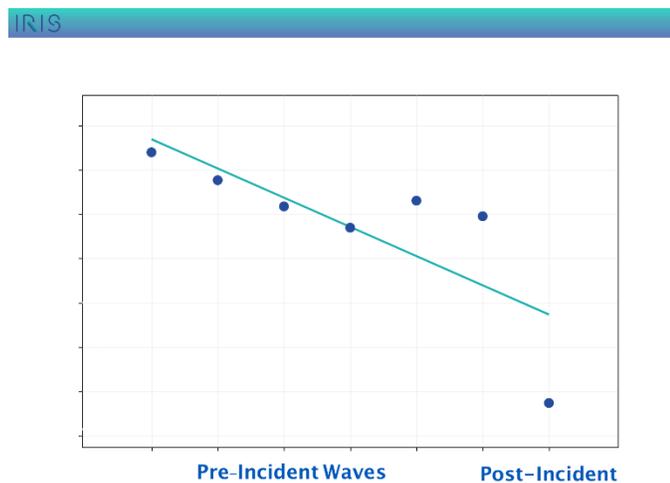


Figure 9: Changing Brand Support in Trend Analysis

While this is now a historical exploration, similar to the Ipsos Global Advisor examples, at the time, this was live information that allowed a timely pulse of a dynamically moving set of attitudes. This also allowed seeing how other tech companies were affected in the moment even without involvement in the issue and with very different privacy policies. In fact, the *IRIS* platform allowed month by month tracking to see which views returned to previous levels, which returned to new levels and which were more enduring with the abrupt change.

Beyond crises, business markets are extremely dynamic and this provides a brief illustration how specific areas for individual or sets of products or category movement in far flung areas of the globe can be identified in a timely way to inform teams that can help spotlight areas where needs and views can be served better and where competitive opportunities and risks can be addressed in ways that allow more nimble competitive action.

### **Concluding Comments**

From the perspective of scientific and actionable insight discovery, the advances in data availability, storage and computing power have created a change to a Fourth Paradigm of research. These advances enable the development of Automated Analytic Insight Exploration Platforms which fall in the domain of leveraging analytic advances for eScience. Our creation of this type of framework allows leveraging many different types of data, from traditional Survey to social listening and other Big Data sources to create large volumes of statistical models of a variety of forms and produced at a velocity that is fundamentally different in kind from traditional hypothesis driven models. The application of the traditional 3 V's in this case led us to dub these aspects of data-intensive scientific discovery, *Big Analytics*.

In this case tens of thousands to a hundred thousand and more models can be produced to identify where clusters of variables are moving and drill down further to identify specific items to see the magnitude of changes within the broader patterns. An important aspect of this is that these are exploratory and show us the patterns within the data. Similar to Big Data this allows identifying things that are changing. While there are a lot of diagnostics that exist in the platform with the ability to compare patterns, consider the types of tests that are moving and examine the specifics of the individual models, including the magnitudes, these still focus more on *What* we are observing but does not necessarily address the *Why*.

From both a practical and theoretic level both the *What* and the *Why* are important elements. While the platform may provide enough guidance for some action, it is often producing questions and pointing to areas for deeper exploration. In the domain of Big Data, some have taken the stronger position that sometimes “knowing *what*, not *why*, is good enough” (Mayer-Schönberger and Cukier 2013). Our view, as summarized in Figure 1, is that this is sometimes not enough (see also Pearl and Mackenzie 2018). Just as in Big Data, we feel this is the case for observationally driven summaries of the automated analytics we presented here. While the deeper diagnostics available in the platform may allow confident decision making, we recognize that this will also play a role in identifying new hypotheses in the Scientific Discovery process.

The examples we shared do not adequately demonstrate this point, as in each case an intervention or event can be seen as part of the key dynamics, from the withdrawal of the United States from the Paris Agreement, the context of the 2016 U.S. presidential election or the Cambridge Analytica scandal. In most cases of scientific and insight discovery that these platforms facilitate, the more common situation is for patterns and clusters of change to be seen,

then to be examined for robustness and meaningful magnitudes and explored further to be understood more fully. It is this iterative process that produces deep understanding into dynamic environments that facilitates deep understanding which can be leveraged to advance scientific exploration or take practical action from policymaking to business decisions by extending the analytic tools in the new era of data-intensive eScience.

## References

- Adams, S. (2018). Start Your Engines: Choosing the Right Vehicle for Your Automation Adventure. *GRIT Report*.
- Bell, G. (2009). "Forward" in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, ed. T. Hey, S. Tansley and K. Tolle. Redmond, WA.: Microsoft Research.
- Benfer, R.A., E.E. Brent Jr., and L. Furbee. (1991). *Expert Systems*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-077. Thousand Oaks, CA: Sage.
- Beyer, M.A. and D. Laney. (2012). "The Importance of 'Big Data': A Definition." G00235055. Stamford, CT: Gartner.
- Brynjolfsson, E.B. and A. McAfee (2014). *The Second Machine Age*. New York and London: W.W. Norton & Co.
- Cadwalladr, C. and E. Graham-Harrison. (2018). "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach." 17 March 2018, theguardian.com (<https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>)
- Cook, T.D., and D.T. Campbell. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Co.
- Daugherty, P.R. and H.J. Wilson. (2018). *Human + Machine: Reimagining Work in the Age of AI*. Boston: Harvard Business Review Press.
- Doerr, J. (2018). *Measure What Matters: How Google, Bono, and the Gates Foundation Rock the World with OKRs*. New York: Portfolio/Penguin.
- Dormehl, L. (2017) *Thinking Machines: The Quest for Artificial Intelligence—and Where It's Taking Us Next*. New York: TarcherPerigree.

- Fawson, B. (2018). Something Old, Something New: The Evolution of Automation in Research- and Why It's a Net Gain. *GRIT Report*
- Friedman, T.L. (2016). *Thank You for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations*. New York: Farrar, Straus and Giroux.
- Gannon, D. and D. Reed. (2009). "Parallelism and the Cloud" in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, ed. T. Hey, S. Tansley and K. Tolle. Redmond, WA.: Microsoft Research.
- Goble, C. and D. DeRoure. (2009). "The Impact of Workflow Tools on Data-Centric Research" in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, ed. T. Hey, S. Tansley and K. Tolle. Redmond, WA.: Microsoft Research.
- Greene, W.H. (2003). *Econometric Analysis, fifth edition*. Upper Saddle River: Prentice Hall.
- Gujarati, D.N. (2003). *Basic Econometrics, fourth edition*. Boston: McGraw Hill.
- Hastie, T., R. Tibshirani, and J. Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second edition*. New York: Springer.
- Hey, T., Tansley S., and K. Tolle, eds. (2009). "Jim Gray on eScience: A Transformed Scientific Method" in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, ed. T. Hey, S. Tansley and K. Tolle. Redmond, WA.: Microsoft Research.
- Kasparov, Garry. (2017). *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. New York: PublicAffairs.
- Mayer-Schönberger and K. Cukier. (2013). *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. Boston, New York: Eamon Dolan: Houghton Mifflin Harcourt.

- McDowall, D., R. McCleary, E.E. Meidinger, and R.A. Hay Jr. (1980). *Interrupted Time Series Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-021. Thousand Oaks, CA: Sage.
- Pearl, J. and D. Mackenzie. (2018) *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.
- Phillips, S. (2018). The Future of Marketing Automation: Faster and Cheaper Isn't Enough. *GRIT Report*
- Reinsel, D., J. Gantz and J. Rydning (2017). Data Age 2025: The Evolution of Data to Life-Critical. IDC White Paper.
- Rosenberg, M., N. Confessore and C. Cadwalladr. (2018) "How Trump Consultants Exploited the Facebook Data of Millions". 17 March 2018, nytimes.com.  
(<https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>)
- Savov, V. (2011). "Visualized: A Zettabyte", engadget.com 29 June 2011;  
(<https://www.engadget.com/2011/06/29/visualized-a-zettabyte/>)
- Seagate Blog. "2016: The Year of the Zettabyte" – accessed 10 October 2018;  
(<https://blog.seagate.com/business/2016-the-year-of-the-zettabyte/>)
- Simmons, Z. (2018). Breaking Down Barriers to Reach the Consumer and Transform Your Business. *GRIT Report*.
- Storey, J.D. (2010). "False Discovery Rate", Princeton University Working Paper.
- Taber, C.S. and R.J. Timpone. (1994) "The Policy Arguer: The Architecture of an Expert System." *Social Science Computer Review* 12:1-25.

Taber, C.S. and R.J. Timpone. (1996). *Computational Modeling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-113. Thousand Oaks, CA: Sage.

Tenner, E. (2018). *The Efficiency Paradox: What Big Data Can't Do*. New York: Alfred A. Knopf.

Timpone, R. (2016) Big Data: A Guided Tour. Ipsos Views #3.

Wonnacott, T.H. and R.J. Wonnacott. (1990). *Introductory Statistics for Business and Economics, fourth edition*. New York: John Wiley & Sons.

Young, C. (2018) Our Age of Uncertainty. Ipsos Point of View.