Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

9-2017

Combining machine-based and econometrics methods for policy analytics insights

Robert J. KAUFFMAN
Singapore Management University, rkauffman@smu.edu.sg

Kwansoo KIM
Izmir University of Economics

Sang-Yong Tom LEE
Hanyang University

Ai Phuong HOANG
Singapore Management University, aphoang.2013@phdis.smu.edu.sg

Jing REN
Singapore Management University, jing.ren.2012@phdis.smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

Part of the Databases and Information Systems Commons, E-Commerce Commons, and the Strategic Management Policy Commons

Citation

KAUFFMAN, Robert J.; KIM, Kwansoo; LEE, Sang-Yong Tom; HOANG, Ai Phuong; and REN, Jing. Combining machine-based and econometrics methods for policy analytics insights. (2017). *Electronic Commerce Research and Applications*. 25, 115-140.
Available at: https://ink.library.smu.edu.sg/sis_research/3729

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Published in Electronic Commerce Research and Applications, 2017 April, Pages 1-26
https://doi.org/10.1016/j.elerap.2017.04.004

Combining machine-based and econometrics methods for policy analytics insights

Robert J. Kauffman [a],*, Kwansoo Kim [b], Sang-Yong Tom Lee [c], Ai-Phuong Hoang [a], Jing Ren [a]

[a] *Singapore Management University, Singapore*
[b] *Izmir University of Economics, Turkey*
[c] *Hanyang University, Republic of Korea*

ABSTRACT

Computational Social Science (CSS) has become a mainstream approach in the empirical study of *policy analytics issues* in various domains of e-commerce research. This article is intended to represent recent advances that have been made for the discovery of new policy-related insights in business, consumer and social settings. The approach discussed is *fusion analytics*, which combines machine-based methods from Computer Science (CS) and explanatory empiricism involving advanced Econometrics and Statistics. It explores several efforts to conduct research inquiry in different functional areas of Electronic Commerce and Information Systems (IS), with applications that represent different functional areas of business, as well as individual consumer, social and public issues. Recent developments and shifts in the scientific study of technology-related phenomena and Social Science issues in the presence of historically-large datasets prompt new forms of research inquiry. They include blended approaches to research methodology, and more interest in the production of research results that have direct application to industry contexts. This article showcases the methods shifts and several contemporary applications. They discuss: (1) feedback effects in mobile phone-based stock trading; (2) sustainability of top-rank chart popularity of music tracks; (3) household TV viewing patterns; and (4) household sampling and purchases of video-on-demand (VoD) services. The range of applicability of the ideas goes beyond the scope of these illustrations, to include issues in public services, healthcare, product and service deployment, public opinion and elections, electronic auctions, and travel and tourism services. In fact, the coverage is as broad as for-profit and for-non-profit, private and public, and governmental and non-governmental institutions.

Keywords:
Causality
Computational Social Science
Data analytics
Econometrics
E-commerce
Empirical research
Fintech
Fusion analytics
Music popularity
Stock trading
Policy analytics
TV viewing
Video-on-demand (VoD)

1. Introduction

IT innovations, specifically high-powered computing and ubiquitous networking, have affected people's lives, and private and public organizational activities, and will continue to change our society in dramatic ways. As technology advances, increasingly abundant digital data from online music social networks, Internet-of-things (IoT)-based sensors, set-top box and handset technologies in cable TV, and mobile phones can be more thoroughly studied with the help of various data analytics methods (IDC, 2009), including Computer Science (CS), Machine Learning (ML), Statistics, and Economics.[1] These lead to new ways to understand people as individuals and in groups, organizations and their partnerships, and firms, industries and society as a whole (Carley, 2002; Anderson, 2008; Manyika et al., 2011; Agarwal and Dhar, 2014). The research inquiry approach that is discussed – *fusion analytics* – is a term that was coined in this and related research, to offer additional new ideas related to *Computational Social Science* (CSS). It builds on earlier data analytics research (IBM, 2012; Chang et al., 2014; Athey, 2015), and our past philosophy of science work related

* Corresponding author.
E-mail addresses: rkauffman@smu.edu.sg (R.J. Kauffman), kwansoo.kim@izmirekonomi.edu.tr (K. Kim), tomlee@hanyang.ac.kr (S.Y.T. Lee), aphoang.2013@phdis.smu.edu.sg (A.P. Hoang), jing.ren.2012@phdis.smu.edu.sg (J. Ren).

[1] *Machine learning* involves "data analysis that automates analytical model building, [by using] algorithms that iteratively learn from data," allowing "computers to find hidden insights without being explicitly programmed where to look" (SAS, 2017). *Econometrics* is the "application of statistical and mathematical theories in economics for the purpose of testing hypotheses and forecasting future trends. It tests economic models "through statistical trials and then compare and contrast the results against real-life. In contrast, *Statistics* "is a form of mathematical analysis that uses quantified models, representations and synopses for a given set of experimental data or real-life studies," and uses methodologies to gather, review, analyze and draw conclusions from data" (Investopedia, 2017).

to the new approaches to big data research (Kauffman and Wood, 2007).

The contexts with big data that are explored are characterized by new opportunities for creating insights with declining costs and increasing technical leverage to accomplish the work. New ways of doing data analytics open up possibilities for innovative thinking and novel contributions to scientific discovery in interdisciplinary contexts (Davenport, 2006), where it is possible to bring different bodies of knowledge and different research approaches to bear in order to reveal interesting insights. *Interdisciplinarity* is critical in this context.[2] By exploring cross-disciplinary issues, asking new questions, implementing new data collection approaches, and instantiating models that would not have been possible before, a policy analyst is able to come closer to *empirical truths* in applied settings that support large-scale data collection. This leads to increased relevance of the research and higher impacts. And yet, accomplishing such work in many organization settings is fraught with challenges, roadblocks, legal restrictions on data sharing, and very short delivery timelines when organizations are involved.

Although researchers in the past have tried to combine Computer Science, Psychology, Sociology, Regional Economics, Biostatistics, and other bodies of theory and methods knowledge to assert interdisciplinary solutions to leading problems, the perspectives they emphasized rarely were interdisciplinary.[3] This article attempts to bring some of these knowledge bases together to address *consumer, business and social issues*. These are defined and discussed in Table 1, as a way to communicate to the reader that this is not a *strict taxonomy*, but instead a *way of thinking* about the kinds of research inquiries that can be undertaken.

Big data enable new levels of sophistication for the study of settings in which technology meets consumer, business and social policy issues, facilitates social and civic empowerment, and enhances stakeholder participation in planning (Brabham, 2009). Data analytics can uncover business and social value from data, by permitting modeling, experimentation, simulation, and other scientific approaches to discover new knowledge. But it is also necessary to identify what kinds of analytics methods can be used from different disciplines, what new knowledge can be discovered regarding the issues at hand, as well as where the data will come from and how it can be collected. Some applied contexts that are appropriate can be identified based on the popular press, and government agency calls for research.[4]

Organizations have been implementing big data projects and deploying new ways of discovering market, product and consumer knowledge. They also have reported new opportunities to obtain informational advantages for the conduct of their businesses, compared to their prior reliance on traditional marketing research methods (Granados et al., 2012; Wang et al., 2016; Li et al., 2017). According to IBM (2012), every field is being changed by the large amount of data available.[5] And increasingly, there has been a desire to find consistency in the methods and inquiry approaches that are used in organizations to achieve high quality, policy-relevant information (Kenett and Shmueli, 2017).

Current technologies and others that are emerging hold out the promise to offer astonishingly rich details concerning human and social activities, contextual patterns of behavior, and the attitudes, preferences, and sentiment of different individuals and groups (Gondecha and Lieu, 2012). The related research examined how people use big data to derive valuable results. Data analytics support finding repetitive patterns and adjusting predictive models to understand how likely it is to observe various behaviors. And yet, as Shmueli (2010) has pointed out, it is important to distinguish between data analytics intended to deliver explanatory information versus predictive information. A recent trend is to use *statistical Machine Learning*, which complements traditional Econometrics (Athey, 2015). When Machine Learning is combined with Econometrics, authors often construct dependent and independent variables with the help of machine-based analytics routines and algorithms, and this sets up the use of modeling structures and error term specification. Researchers working on electronic commerce have also begun to use this approach (Lee et al., 2016; Shi et al., 2016).

This article is intended to encourage researchers, analysts and doctoral students to apply interdisciplinary fusion analytics research approaches to achieve useful new results. For example, they include pattern recognition for data and other modes of machine-based discovery to create large datasets, with explanatory analysis that yields insights into the marginal impacts of different policies through the variables that represent them. This also makes it possible to do *counterfactual impact analysis* (Mohr, 1995; Lewis, 2001; Larkey, 2015; Science and Knowledge Service, 2016). Counterfactual analysis is about simulating what the marginal effects of policy interventions may be by considering what may happen in its presence versus its absence, based on the data analytics techniques available to discover this.[6] This allows analytics to produce more managerial insights, and build powerful evidence to understand business, consumer and social policies. This also is often a matter of what researchers bring to research designs that they implement.

Section 2 discusses the interdisciplinary roots of the fusion analytics research approach, and presents a new framework to capture the essential elements of the different research approaches that have been identified. Section 3 provides additional background on the importance of data analytics, and the rise of techniques that evaluate the digital traces of human behavior in different contexts and provide various means of network analysis and information visualization. Sections 4–7 then illustrate the fusion analytics approach with applications involving large datasets that are analyzed with machine-based methods, statistics, and econometrics. Finally, Section 8 offers a concluding discussion that evaluates what has been learned in this research, and the various tactics that are required for achieving analytics success with big data and Computational Social Science research designs.

[2] It is important to point out that doing machine-based analytics with explanatory methods may yield interesting observations on patterns in the data, and relationships that have not been recognized. Adding explanatory statistics or econometrics, along with a research design for the analytics to support the discovery of causality takes the analysis to a potentially more powerful level of effectiveness and deeper insight, so it is possible for the analyst to obtain knowledge about the dataset that goes beyond correlation. Similarly, conducting statistical and econometric analysis of a big data store may fail to identify empirical regularities that are model-free, as opposed to being bound by the empirical model's design. This can be a shortcoming, and it is especially important in academic research, where the purpose is to seek theory-based causal effects explanations.

[3] Wellman (1995) has pointed out the benefits associated with using Computer Science and Artificial Intelligence (AI) to explain important relationships in Social Science, especially with agent-based approaches to artificial economies. Only in the last decade has Computer Science moved to a position of high innovation in interdisciplinary studies. Social network analysis has been especially central to all of the new scientific approaches and data analytics perspectives that we have seen (Hassan, 2009; Mayer-Schönberger and Cukier, 2013).

[4] A good starting point for identifying appropriate contexts is new research inquiry. Some of the issues that are covered include: influence of friends in social media and randomized experimental designs (Aral and Walker, 2011; Bapna and Umyarov, 2015); movie sales affected by social network interactions (Moretti, 2011); ad position auctions with consumer search (Athey and Ellison, 2011); and ranking of hotels on travel search engines by mining user-generated and crowd-sourced content (Ghose et al., 2012).

[5] For example, information about online users is necessary to optimize credit card reward programs and business partnerships and achieve high business value in customer relationships (McKinsey, 2011; Geng and Kauffman, 2017).

[6] Examples of counterfactual research questions related to our work in the first empirical research illustration in this article are: What if TV program bundling had been customized to individual households instead of just broader segments? How would cable TV services *average revenue per unit* (ARPU) have changed? No customization was done in the marketplace at the time.

Table 1
Orientation to Business, Consumer and Social Insights Data Analytics Settings.

Insights area	Definition	Representative data analytics contexts
Business	Data analytics can create a deeper understanding of the issues, situations, and contexts that offer potential of benefits. By *business insights*, analytics are involved in settings that are for-profit / not-for-profit, public / private, governmental and non-governmental, in operational, logistical, financial, human resources, strategy and industry settings. Insights can redirect mgmt's thinking, guiding future decisions for better product / service design and quality, operational performance and productivity, risk mgmt and security, and market share, profitability and performance. Applicable across business disciplines, work groups, product lines, support services, strategic business units, organizations, firms, sectors and economy; can also include transactions, processes, economic exchange, business activities, managerial issues, and the dynamics of competition.	- Stock market trade analytics - Churn rate analytics for telcos - Internet fraud detection - New product launch analytics - Resource allocation, planning - Inventory management - Faster payment and settlement - Banking product line design - Govt regulation change effects - Hotel room utilization - Emergency room services - Taxi fleet passenger pick-ups - Airline landing rights bidding
Consumer	Data analytics can also reveal a deeper understanding of consumers, customers, users, medical patients, and decision-makers.By *consumer insights*, analytics are involved in settings with individuals, and the behaviors that they demonstrate. This area of analytics enhances the understanding of the analyst through close examination of people and their interactions with products, services, businesses, cities, and government and non-business organizations, including healthcare, legal and family services. Emphasizes consumption, consumer utility, consumer informedness and satisfaction; customer and patient centricity, individual wellness; auction bidding and group-buying behavior; decisions to buy, sell, participate; everything under the general umbrella of individual welfare; and not limited to or intended to be circumscribed by consumer behavior in retailing.	- Fast-moving product turn-over - Movie box office performance - Product price change impacts - Consumer coupon redemption - Customer product returns - Customer channel management - Online, offline retail shopping - Credit card spending patterns - Personal loan defaults - Individual criminal recidivism - Home mortgage demand - Household waste recycling - Sharing-economy services
Social	Big data analytics are concerned with peer-to-peer, social network and non-network relationships, social activities and events; urban and regional programs and policies; legal, law enforcement and political issues; elections and issues referenda, community and regional planning. By *social insights*, the connection is to analytics that produce deep knowledge into activities that are of a more aggregate nature than those that typically occur in business or consumer activities. In lieu of business value, the focus may shift to outcome measures such as environmental sustainability, water quality, community wellness, growth in regional real estate value, and social value of community programs. These issues can be studied well when the promulgation of a new program, policy, law or initiative creates a natural quasi-experiment treatment for those who are affected, and a control for those who are unaffected, with randomized instead of biased participation. Examine interactions related to and social issues among people in different segments (younger / older, or race / ethnicity / income), as well as among people of different educational and social status levels, with the aim of understanding and maximizing social welfare. Data come in various sources and forms, including social network, social media, blogging, human wellness, opinions, ratings and sentiment data.	- Public social sentiment effects on candidates and elections - Social media influences in public election settings - Patterns of responses in situations of crisis involving publicly-available data - Notifications of weather events and problems to emergency services via social messaging - Online ratings and opinion analysis on consumer engagement for brands and healthcare services - Social opinions on sustainable development, linking economy, society and environment - Effects of social informedness on community recycling - Blogpost and tweet analytics to analyze public services quality

2. A fusion analytics research framework

Now the discussion turns to the technology infrastructure that supports big data analytics, as well as the use of explanatory methods from Statistics and Econometrics. The section ends with the contribution of a new framework to help others understand what are the main things considered related to the development of fusion analytics research in this area of inquiry – whether the research is conducted by industry practitioners to discover how to improve their policies, or academic researchers who are interested to deliver longer-standing theoretical and empirical truths about various problem and issue contexts.

2.1. How technology enables big data-related research inquiry

IBM (2017a) has long defined large-scale data in terms *of four V's*: *volume*, *velocity*, *variety*, and *veracity*. The fusion analytics paradigm takes advantage of technological capabilities to develop research designs that address these dimensions. *Variety* implies that many tools and approaches are needed to process data in its different forms (e.g., large cross-sections, lengthy panels). Data also exhibit *velocity*, for example, the rate of transactions at an urban ATM, the calls that arrive at a government services help center, or the number of vehicles that flow through different routes at rush hour in a crowded city – all the way up to the more rapidly streaming data from high-frequency trading (HFT) in the stock markets, and digital data on image, music and video broadcasts. They exhibit great *variety* too, such as funds transfer transactions, user sentiment in social media, and moment-to-moment geopositional updates for the cars in a large taxi fleet. Still other data represent facts, while some may relate to opinions and estimates, so data may vary in the extent of the truth, or *veracity*, in the content it delivers. Putting data together from websites, user-generated content, and sensors in automobiles and smartphones allows researchers to explain and predict individual behavior and detect trends in context also.

As a result, to make machine-based data analytics effective, big data researchers and analysts must be able to leverage *data infrastructures*. Managing a platform of services with streaming data, and trying to build useful metrics for its performance is a challenging problem. There have been efforts to model, analyze, and optimize benefits and service levels for technology infrastructures (Demirkan et al., 2008; Bardhan et al., 2010), and to determine the conditions for ideal value creation to ensue (Benaroch et al., 2010; Ma and Kauffman, 2014). Deploying adequate technology infrastructure for big data analytics in the organization is critical to support the creation of value and meaningful social insights.[7]

Other new directions that have appeared include the embedding of data analytics capabilities in cloud computing infrastructure (Cloud Standards Customer Council, 2014), as well as the rise of *cognitive computing*, which involves the embedding of human thought processes and decision-making logic into systems that can be used in practice to drive higher value (Lopez, 2016; Davenport and Krishna, 2017). Some of these capabilities for analytics have come together with new infrastructure approaches. An example is IBM's Watson cognitive computing and AI approach, and the software and services that are now available on its BlueMix cloud computing platform (IBM, 2017b).

Moreover, important opportunities exist for interdisciplinary collaboration on big data fusion analytics. Information sciences research groups in university, business, healthcare, and government settings offer bright prospects as pioneers for interdisciplinary collaboration. This is true in e-commerce and digital marketing especially, for example. Interdisciplinarity to support more effective fusion analytics should also extend to non-business and social problems, which go beyond the typical spectrum of most business school research. Big data analytics have pushed the boundaries of a number of disciplines outward based on forces that have been developing in this new environment. In industry settings, it is typical that organizations face challenges to learn about the benefits and constraints though. So it makes sense to leverage computing power that has become available through new statistics software, so new data sources contribute more useful information for improving data-driven business and organizational performance.

Data specialists today are working within an entirely new data-driven science of analytics in comparison to ten years ago. Computing power allows the creation and testing of hundreds of hypotheses, models, and simulations more quickly than ever before, and at lower cost. Algorithm use is an essential element of large-scale analytics. Machines can learn from data, and analysts can leverage the intersection of Machine Learning, Artificial Intelligence, and data processing methods that offer new ways to build datasets and study them. This will allow them to understand problems that were hidden in the presence of prior analysis methodologies (Dietterich, 2003). What is critical is whether the discovery of new Social Science knowledge, the efficacy of new theories, and more effective explanations and predictions can be achieved.

2.2. The role of explanatory econometrics building on machine-based data analytics

Traditional statistics and econometric methods generally have been undertaken separately from Computer Science methods involving machine-based data analytics. Analysts and researchers need to be aware of the role of econometric and statistical analysis that builds on the machine-based methods though. For example, the former has assumed that data observations generally are independent or grouped, as in panel data with group-wise stratification, or are linked by time (Wooldridge, 2010). Econometric modeling has been used to uncover what are the key relationships, influences, and marginal effects of the relevant variables based on carefully structured and cleaned data. Today, individuals in a social network may be interconnected in complex ways, creating new statistical challenges for determining causal relationships. The point of econometric modeling has been to uncover what are the key relationships influences and marginal effects of the relevant variables. Recent research in social networks, widely viewed as *complex adaptive systems* (Lymperopoulos and Lekakos, 2013), has sought to discover the kinds of *node-and-link dependence structures* that are present in social media and Internet advertising, peer-to-peer (P2P) lending and social crowdsourcing, and other online networks (Lazer et al., 2009). Developing methods suited to these and other network settings in e-commerce has been an ongoing challenge for statisticians[8] and econometricians as a result (Imbens et al., 2011).

Similar empirical modeling issues arise in almost every setting that is explored, and as a result, there is an ongoing need for advances in the methods that support causal explanations. The new methods that involve increasing sophistication in the study of business, consumer and social insights are a challenge for both researchers and practitioners. Nevertheless, blending Computer Science methods with econometrics has become attractive, due to the spate of recent work on statistical Machine Learning techniques that are tied ever closer to the causal relationship discovery process, and the experimental design approaches that Statistics and Econometrics offer. And, though new methods are increasingly used, knowledge about them still has not diffused very widely.

Moreover, the efficacy of explanatory methods must be built on their application in public and private organizations that have highly knowledgeable staff, who know where to acquire and how to work with big datasets. There are a growing number of such datasets in healthcare (Fagella, 2016), education (IBM, 2017c), public transportation (Nemschoff, 2014), and the government domain (Data.gov), among others. Industry practitioners and government researchers in these areas need access to readily implemented technical, behavioral, managerial and economic approaches, as well as scalable data analytics systems and infrastructures (Hu et al., 2014). They also must understand the power of diverse data sources, but often lack the technical capabilities for powerful machine-based, econometrics and statistical analytics, to take advantage of big data in their environments (Ernst and Young, 2014).

The art of conducting empirical research involving decision-making and policy issues is to recognize that understanding the *context is key to obtaining meaningful results*. In the 1990s, Marketing issues were the object of intense interdisciplinary research work, but did not include Computer Science as a front-end methodology to the explanatory statistical and econometric analysis methods that were used. In the 2000s and more recently though, the Marketing discipline has been moving more rapidly to embrace the new machine-based methods for its research and data analytics contexts.

Marketing researchers now more routinely rely on classification and recommendation algorithms from CS, such as the *latent Dirichlet allocation* (LDA) *topic model*, and blend their use with advanced statistical and econometric analysis (cf. Tirunallai and Tellis, 2014). From channel choices to price-setting and recommendations, music bundling and online reviews, and social networks, interdisciplinary researchers have conducted highly impactful data

[7] The technical issues related to data capture, streaming, archiving, and parallel computing are central to CS methods. For example, the Hadoop family of technologies and services were developed for the handling of data-at-scale in Google, cloud computing, and other technical environments. Data scientists have been trying to detect patterns in data as a way of advancing knowledge. This paradigm includes new algorithmic, computational, and pattern recognition tools to produce value from available data (Bell et al., 2009; Hey et al., 2009).

[8] This is based on personal discussion with Professor Steven Fienberg of Carnegie Mellon University (now deceased), when he was Co-Director of the SMU-CMU Living Analytics Research Centre (LARC) (2011 to 2016). His ideas and research have had a salient influence on the perspectives, methods, projects, and research inquiries that were undertook in sponsored projects.

analytics research that offers specific findings that are relevant for practitioners, as well as more lasting results that have become a part of academic knowledge and interdisciplinary theory (Elberse, 2010; Chellappa et al., 2011).[9] Researchers are continuing to become even more proficient in the use of innovative research designs, Machine Learning techniques, and econometric methods, as new opportunities with big data open up.

2.3. A framework for computational social science fusion analytics

Various machine-based methods, such as *data mining*, *natural language processing*, *machine learning* and *statistical learning*, have been applied in different research inquiries to identify patterns and hidden aspects of large datasets. They are limited in their abilities to help an analyst or a researcher draw causal inferences. This affects the *out-of-sample prediction power* that can be obtained with them. For example, the application of data mining to social network posts and tweets supports learning about topics that others are interested in. The public may be tracking a variety of issues at one time; yet knowing *what* – without knowing *how, why* and *what's next* – may not produce useful or meaningful insights.

The usage of machine-based methods from CS for the collection and analysis of data in our fusion analytics methods process deserves additional discussion. What commonalities do these methods share that have been leveraged in the research studies that are used as illustrations in this article? They allow learning to occur in the data collection and data refinement process so that more *value-focused* and *insight-creating research questions* can be evaluated with advanced statistical and econometric models. Researchers employ different methods to extract appropriate data that allow them to address causality, whether through quasi-experiment designs or deep and advanced econometric modelling. The data outputs from these machine-based methods are rich, structured and tailor-made to support explanatory empiricism for causal inference. This is an iterative process, with a strong foundation of contextual knowledge, and various theoretical perspectives, when the methods are used for producing knowledge from academic research. Fig. 1 depicts the *analytics fusion research approach*.

The elements that characterize the fusion analytics research approach offered in this article include: (1) the *purpose* of the data analytics work; (2) the *machine-based methods* that are used; (3) the output obtained by the researcher or analyst obtains (either harvesting and construction of a dataset, or different kinds of patterns, recommendations, and learning model information); (4) *the statistical and econometric explanatory methods* that are to be used; and (5) the *policy insights* that emerge to address issues in the public and private domains, related to aggregated or disaggregated data, with many different kinds of outcomes and value produced. Through this approach, it is possible to extract valuable business, consumer and social insights that are useful for policy-makers for the contexts of interest.[10]

3. Policy analysis

3.1. Data analytics background

Business intelligence became a popular term in the business and IT communities in the 1990s. By the late 2000s though, the term *data analytics* replaced it (Davenport, 2006). More recently, *big data analytics* has become popular, with datasets and analytical techniques for large and complex applications, from location sensors to social media data (Chen et al., 2012). Data analytics rely on new ways of accomplishing data extraction, and contemporary analysis methods (Chaudhuri et al., 2011; Watson and Wixom, 2007).

An important issue is whether the economic benefits of organizational and societal decisions are realized. What is needed is methodology for how large-scale data can be used to understand the impacts of different policies, and what the economic consequences of the related decisions are. So it is important to identify performance variables (as *impact factors*), their coefficients (as *marginal contributions*), and derive normative implications the data permit an analyst to craft effective policies – whether the context is public or private, or involves disaggregate or aggregate activities.

In addition to the underlying data processing technologies and analytical techniques, data analytics include process-centric practices and methods that can be applied to high-impact applications, such as e-commerce, market intelligence, e-government, healthcare, and security (Cohen et al., 2009). Performance management often uses scorecards and dashboards that are helpful to analyze a variety of relevant metrics. In addition to established reporting functions, data mining and statistical analysis support associational analysis, data segmentation, clustering, classification, and anomaly detection.

3.2. The policy analytics perspective

Policy analytics evaluates and legitimizes strategies and tactics for public and private organizations – small and large firms, cities, states, regions and even countries. Its findings are often based on agreed-upon outcomes to measure, such as equity, economic efficiency, social acceptability, legality, and the proxies that support their evaluation. They enable an analyst to establish recommendations or assessments by processing information from relevant data using appropriate tools to make important inferences and discover useful insights. In this kind of process for problem formulation, evaluation and selection of policies, it is necessary to consider the interests and preferences, the applicable priorities, and the values of a diverse set of stakeholders (Brickland, 2001).[11] As De Marchi et al. (2016) have stated, the analytics themselves are most useful for *evidence-based policy-making*, which "*helps people make well informed decisions about policies, programs and projects by putting the best available evidence from research at the heart of policy development and implementation*" (Davies, 1999). And yet, the decisions that can arise out of such processes are complicated with stakeholder differences of opinion, bureaucratic requirements, tight deadlines for completion, and sometimes overwhelming political considerations. And for researchers, organizational access, cooperation, relationship sustainability, and completion can be hard too.

[9] Many different sources of data have been used, and there have been substantive advances in the collection and analysis of web data since the earliest efforts were made to write down the requirements for the persistent operation of data-collecting software agents in the present of possible system aborts and telecom disruptions (cf. Kauffman et al., 2000).

[10] There are many approaches that have been developed to the collection of data-at-scale to support asking new research questions about patterns, issues, observations and relationships. Researchers can choose any method to collect, extract, structure, refine and apply machine-based analysis to the relevant data, as readers will see from the research presented in this article. The objective is to create a relevant and integrated dataset, and to identify different kinds of empirical facts, such as patterns, classifications, valences on sentiments and opinions, and other kinds of useful information. Then, a researcher can apply statistics and econometrics to discover new ways to explain what has been observed, on the basis of variables that aid in explanation.

[11] *Decision theory* was developed to support managerial understanding regarding a rational individual's consistency and effectiveness in making choices under uncertainty, mostly through the use of *expected value analysis*. *Policy analysis* for firm decision-making is different. It often needs modifications, flexibilities and different theoretical thinking to deal with problems of group consensus, ill-defined objectives, and disparate information sources (Bunn, 1977).

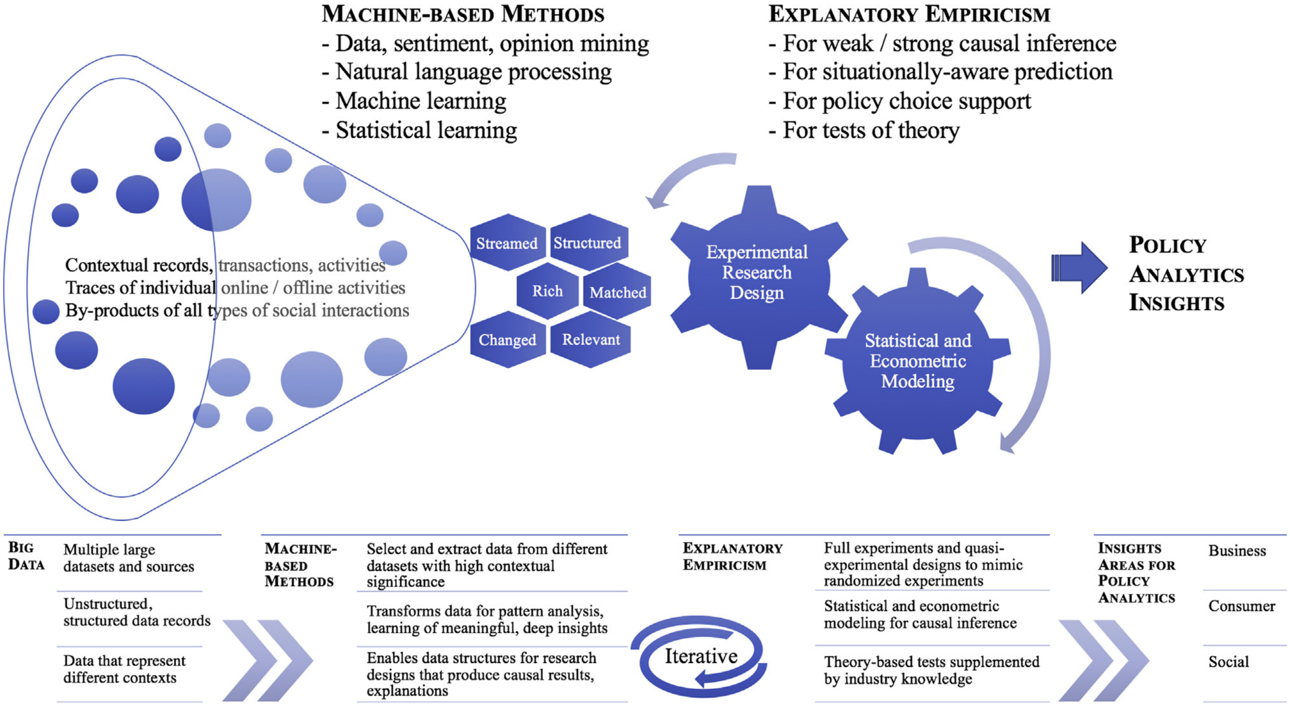

Machine-based Methods
- Data, sentiment, opinion mining
- Natural language processing
- Machine learning
- Statistical learning
Explanatory Empiricism
- For weak / strong causal inference
- For situationally-aware prediction
- For policy choice support
- For tests of theory
Contextual records, transactions, activities
Traces of individual online / offline activities
By-products of all types of social interactions
Streamed
Structured
Rich
Matched
Changed
Relevant
Experimental Research Design
Statistical and Econometric Modeling
Policy Analytics Insights
Big Data
Multiple large datasets and sources
Unstructured, structured data records
Data that represent different contexts
Machine-based Methods
Select and extract data from different datasets with high contextual significance
Transforms data for pattern analysis, learning of meaningful, deep insights
Enables data structures for research designs that produce causal results, explanations
Iterative
Explanatory Empiricism
Full experiments and quasi-experimental designs to mimic randomized experiments
Statistical and econometric modeling for causal inference
Theory-based tests supplemented by industry knowledge
Insights Areas for Policy Analytics
Business
Consumer
Social

Fig. 1. A fusion analytics research process for computational social science research.

Forecasting methods need to be integrated into decision-analytic frameworks, if they are to provide useful evidence for decisions. This is to ensure that there is recognition of how to answer the important *prospective question* of "What should be done going forward?" versus the *retrospective question* of "What should have been done in the past?" A consequence of adopting a decision-theoretic approach to forecasting is the need to develop ways of synthesizing the results of a set of available prediction methods. Such an approach to support forecasting is different from conventional selection methods. In the policy analysis process, whose values should be invoked? Are the interests of certain groups more important than those of others? How can the preferences of different stakeholders be effectively gauged? Big data and related methods in the area of social sentiment analysis have the potential to be useful for answering these kinds of questions. This is true for consumer preferences and organizational planning as well – whether they are related to books, movies, or luxury products, and products to redesign, services to revamp, and policies to adjust.

To understand how to gauge the importance of uncertainty, policy analysts need to create appropriate analytic contexts around the data they use and the issues they wish to study. One way to achieve this is through *data integration*, the technological process of combining different sources of data using technical infrastructure for big data operations to create more useful information for different policy context (Delen and Demirkan, 2013). This has traditionally been true with methods such as *Delphi sessions*, but today it is more applicable to settings in which there are social comments appended to geospatial location data, such as in Twitter and FourSquare, and a myriad of law enforcement-related data analytics applications that pertain to terrorism and crime prediction (Stroud, 2014). They include: the U.S. Department of Homeland Security's Future Attribute Screening Technology Project (FAST), a potential terrorist diagnosis application; and Rutgers University's Risk Terrain Modeling Diagnostics (RTM Dx). The latter is an application deployed to police agencies in different states of the U.S. that:

> "… *uses geolocation and crime data to measure the spatial correlation between where the crimes have occurred in relation to different features of the environment such as nightclubs or bars. With that, officers can measure correlations between various sites and crime rates and then decide which … correlations are worth monitoring and pursuing.*" (Mor, 2014)

3.3. The rise of digital traces, and the business and social value of data analytics

A valuable insight must be unique and suggestive of actions that can be taken to improve decisions about all kinds of individual, organizational, healthcare and societal activities.[12] As consumers become more involved in activities in the digital economy and e-commerce, and are subject to many forms of social sensing, they leave many *digital traces* of their behavior, from TV viewing habits, to opinions on products and services, to mobile phone use and texting behavior, as well as their ATM, PC and phone banking transactions. These are among the most powerful sources of data for insight creation, and are driving changes and transformation in marketing, logistics and information security research.

Other techniques that take advantage of the digital traces of people, phones, automobiles, tourist and shoppers have grown popular as well. *Information visualization* (Aigner et al., 2008) and *network analysis* (Lazer et al., 2009) have a long history in the natural sciences. However, their impact has increased, as they have overlapped with the emergence of big data in producing *business value and social value* for public and private sector organizations.[13]

[12] Understanding the guilt that consumers feel when they eat something they like, and then combining it with the brand experience of a product, can lead to insights about "pleasure that justifies the guilt," and makes the consumer willing to pay more. Such insights are critical, and can be handed over to a firm's creative team to supplant the guesswork that has dominated (Sen, 2003).

[13] Social media analytics involve a three-stage process of (1) *capturing underlying patterns*, (2) *understanding what they represent and may mean*, and then (3) *presenting them to others*, so it is possible *to identify policy actions*.

These techniques go beyond text analytics to include opinion mining, sentiment analysis, topic modeling, social network analysis, trend analysis, and visual analytics. They can be used in businesses to realize value in all phases of a product or service life cycle, for changing market tastes, advertising campaign effectiveness, responding to operational problems or crises, and creating new ways to get competitive intelligence. And they often are used in public organizations to gauge how well informed citizens are about issues, the perceived effectiveness of government policies and emergency services, and how environmental issues are being managed.

In social issue-related applications, Del Guidice et al. (2015) studied emerging economies. They covered practices and tools for emerging markets, social media, statistics, peer opinions, buzz and viral marketing, and word-of-mouth. Harrysson et al. (2014) outlined opportunities related to consumer behavior in social media, which blends social and consumer insights research inquiry. They suggested ways to: create dispersed networks to achieve a deep understanding of the business; equip employees to browse blogs to create followers and match their interests to in-store retail offerings; and use insights for marketing, sales, product design and customer support. Again, it must be stressed that such research may span multiple areas of the business, consumer and social insights spectrum, as opposed to matching just one of them.

3.4. Value creation, the data analytics skills shortage, and organizational alignment

Data analytics have improved over the past few years, giving public and private organizations of different types greater value and more meaningful insights. Today, massive databases require a mix of automated analysis techniques and human effort to give users cost-effective and deep insights about the activity on their websites – in the law enforcement context, for example, as well as about the characteristics of a country's tourists and visitors, in the context of national-level hospitality and tourism services planning. In addition, with millions of click-streaming records generated every day, aggregated to transaction record summaries for credit card users, public transportation users, drivers on toll roads, and financial market traders, there is a need for automated techniques to find meaningful patterns in the data, and translate them into relevant knowledge. To make decisions based on the large datasets that are collected, data analysts must be savvy in the use of high-tech applications for extracting information that blend data analysis with task, context, and time-specific knowledge.

And yet most modern economies do not have enough skilled people to populate the kinds of job roles and deliver the necessary skills that are in such high demand. For example, the U.S. is forecasted to have a gap of 50% to 60% between lower supply and higher demand for data analytics skills in 2018, according to a 2011 report by McKinsey (Manyika et al., 2011). Organizations are investing more in new data analytics technologies, and gleaning insights through data collection and analytics methods that require knowledgeable professionals to drive their planning, implementation and ongoing management. Farris (2010) pointed out the contrast between traditional metrics, such as market share, sales force performance, rebates, market reach, and revenue production. He also noted the importance of digital economy metrics – for web campaigns, e-commerce opportunities, and leading indicators of digital financial performance.

By broadening the uses of analytics in organizational processes, the solutions can enrich and go beyond customer, patient and citizen-centric applications to support sales and marketing, supply chain visibility, pricing and access, workforce management, and healthcare and public services better than before. Finally, value-maximizing analytics solutions have to produce results that are actionable, with ways to measure the effects of the changes that occur. The challenges of big datasets continue to involve their collection, validation, integrity, and security. Such issues will continue to arise with increased use of big data for policy-making in the information society (Mayer-Schönberger and Cukier, 2013).

In Sections 4–7, overviews and methods commentaries are offered on research projects that illustrate Computational Social Science fusion analytics methods from the framework in Section 2. They leverage Computer Science and Machine Learning methods to help collect, extract, integrate, structure and classify the data, and then combine this work with explanatory Econometrics and Statistics methods to extract new business, consumer and social insights. (For a summary of the studies according to the set of dimensions that are considered, see Appendix A, Table A1.)

4. Mobile phone-based stock trading

Next, an illustration of machine-based analysis of social sentiment for mobile phone users who trade stock is presented. Text analytics for social sentiment analysis were combined with econometric methods, including *feasible generalized least squares* (FGLS), Granger causality based on *panel vector autoregression* (PVAR), and *kernel regularized least squares* (KRLS) building on other traditional estimation methods. They helped to validate a *general conjecture* (an assertion that some phenomena have a basis in theory, but where none is specified yet) regarding whether mobile phone-based stock-trading is subject to *social sentiment effects* that cause unusual feedback trading. (See Appendix A, Table A1; it offers an overview of this study in terms of the fusion analytics research framework, that will guide the reader.)

4.1. Context, data acquisition, and machine-based data analytics

4.1.1. Context and data

Social media sentiment affects market trading volume for equities, which may be biased in systematic ways, in the mobile phone channel.[14] Using data from Korea, evidence was obtained that suggested there were different patterns of trading – especially positive and negative feedback trading. They appeared in the short run, but disappeared over time as mobile traders became more informed (Kim et al., 2016). The results support alternative conclusions about the trading of securities via mobile phones: either as a *smart channel* for beneficial exchange, or possibly as a *noisy channel* with low information and few benefits for investors.

Most financial markets should have basic similarities when it comes to trading, irrespective of the trading channel. Korean market stock trade volumes were collected from data previously culled from a mobile stock-trading platform. Social sentiment postings, such as on Twitter and via blogs, also were obtained for May to September 2012. The data cover about 251 firms that were discussed in Korean social media. Two groups of equities were listed on the Korean Exchange (KRX) at that time: the 125 largest firms, and another 126 firms of somewhat smaller size.[15] The companies

[14] In Finance, the technical term used when the trader does not know much about what is going on in the market, or about a specific equity issue that is being bought or sold, is an *uninformed noise trader*. According to De Long et al., 1990, stock investors also can be *informed value traders*, rationally anticipating asset value. Uninformed traders react to changing sentiment in the market though, and this may cause persistent mispricing of stock-related bids and offers. Traders who use mobile phones for stock transactions may view social signals as relevant information but it makes sense to view them as being relatively uninformed. After all, sentiment is the *noise of the market-at-large*.

[15] This classification helped to discover more nuanced results for the impacts of social sentiment on observed trading volumes.

were divided into *IT and non-IT firms*, and into *index categories* of the *Korea Composite Stock Price Index* (KOSPI), as well as according to definitions from the *Korean Securities Dealers Automated Quotations* (KOSDAQ).

4.1.2. Lexicon and learning-based algorithms for text mining

Progress has been made in social sentiment tracking techniques that extract indicators of sentiment from social media content, particularly from large-scale Twitter and blog posts. *Social metrics analytics* from DaumSoft (www.daum-soft.com) were employed in this research for text message mining, reflecting the language, sentiment, and opinions of social media users.[16] Its social metrics implement *contextual social analysis.* The text mining algorithms not only extract and enumerate the data that are processed, they also show where each piece of sentiment data originated, and why it was presented to the analyst.[17,18]

Two types of algorithms are implemented in DaumSoft's software: lexicon-based and learning-based algorithms. The *lexicon-based algorithm* uses a dictionary of words to perform entity-level sentiment analysis. Its words are tagged with their semantic polarity and used to calculate an *overall text document polarity score*. This gives high precision but low recall. Lexicon-based techniques with large dictionaries enable good results, but require using a lexicon, which may not be available for a language or domain.

In contrast, a *learning-based algorithm* creates a model by training the classifier with a labeled training dataset. Thus, it was necessary to first gather a dataset with examples for positive, negative, and neutral classes in the stock-trading domain. Next was to extract the *features and words* from the examples, and then train the algorithm based on them. This approach delivers good results, but requires obtaining appropriate large-scale datasets and then making sure the algorithm learns to analyze sentiment from social media content, such as tweets and status updates. On a procedural level, the algorithm takes a *string of words*, and returns a *positive*, *negative*, or *neutral* sentiment rating: a *sentiment valence score* useful for recognizing the social mood of the public. To do this, it identifies the *meaning of words* in their *social context*, different from *social network analysis monitoring*, as other algorithms do. The learning-based algorithm also provides a compound reading on overall sentiment for a set of strings.

Positive or negative *mood words* were identified for the stock-trading domain in terms of their frequency of occurrence. To support the reader's intuition, *positive sentiment* was represented by the values of variables that showed a positive response toward a firm or its stock. For example, sentiment related to profit, improvements, innovativeness, new concepts, and about 200 other words were related to business progress. *Negative sentiment* was classified through variables that showed a negative response or opposition to something in a firm's business, operations or strategy. Examples include sentiment on the likelihood of a recession, faulty products, or lagging profitability, and 184 other words associated with business impediments. Table 2 shows the descriptive statistics for the panel dataset.[19]

Table 2
Descriptive statistics.

Variable	Mean	Std. Dev.	Min.	Max.
Frequency	198.75	778.70	0	31,224
Positive	52.73	182.57	0	4,623
Negative	18.95	171.70	0	18,928

Note: There are 17,068 observation in this dataset, for each of the descriptive variables, *Frequency*, *Positive* and *Negative*.

4.2. Empirical modeling

The empirical modeling and estimation process that was used involves a number of steps. First, *feasible generalized least squares* (FGLS), used to resolve error term issues that characterize this and other similar settings (Amemiya, 1985; Kmenta, 1986), was applied to estimate stock-trading volume. Second, *panel vector autoregression* (PVAR) was used for capturing the impact of exogenous changes in endogenous variables on the other variables in a PVAR system of equations. In this applied setting, an endogenous relationship between stock trading volumes and social sentiment was hypothesized to exist. This led to the use of a Granger causality test for the estimation of the PVAR system (Abrigo and Love, 2015). The estimates derived from PVAR are seldom interpreted in isolation (Canova and Ciccarelli, 2013).

Third, *kernel regularized least squares* (KRLS) was used in place of *generalized least squares* (GLS), a method which imposes stringent assumptions that sometimes are not accurate for data in real-world settings. One such assumption is that the marginal effects of the explanatory variables are constant (Sekhon, 2009). GLS further assumes that observations with similar values of their variables should have similar marginal impacts on the dependent variable of a model on average. This reduces misspecification, and avoids the need for users to guess the functional form to be used (Amemiya, 1985).

KRLS, in contrast, is a recent advance with respect to GLS's handling of functional form issues. KRLS uses *regularization*, which emphasizes a prior preference for the estimation of a smoother functional form over a more erratic functional form for the model (Hainmueller and Hazlett, 2013). As a result, KRLS is able to minimize over-fitting the model by reducing the variance and fragility of the estimates, and diminishing the influence of inappropriate points. Similar to GLS, it also is suitable when the functional form of the model is unknown.

The following mathematical model was specified to account for unobserved fixed effects related to each firm i involved and the period of time t.

[16] DaumSoft is an Internet portal that analyzes data on social media and issues. It is known for mining user interest-related words for marketing strategy in various business settings.

[17] For example, the social metrics are able to analyze the text of stories in social media for the word "iPhone," what related topics are being talked about (e.g., new products, performance, or problems, and what topics they represent. This yields a *topic flow diagram* for the word "iPhone" rather than the number of messages or tweets about it.

[18] For additional information on opinion mining for the stock-trading context, the interested reader should see Das and Chen (2007), which covers: web data scrapers and helper applications for message parsing and data handling; data pre-processing and classification algorithms, multiple data-related sources (stock data, data dictionary and word lexicon, and grammar information), as well as classified messages and statistical summation.

[19] The blog posts and tweets were not separated; instead, they were combined all together. The statistical estimates suggest that there was a good fit for the data based on standard model selection criteria. This was done via the χ^2 distribution, and a check for multicollinearity with *variance inflation factors* (VIF) and *condition indices* (CI) due to use of multi-period panel dataset rather than a one-period cross-sectional dataset for this. These diagnostics suggested the data were appropriate for our analysis.

$$\begin{aligned}
StockTradingVolume_{it} = \beta_0 &+ \beta_{Industry-Level}(\text{Dummy variables}: \ Size_t, \ Industry_t, \ Market_t) \\
&+ \beta_{Industry-Level}(\text{Normal returns}: \ MarketTradingVolume_t) \\
&+ \beta_{Firm-Level}(\text{Common factors}: \ Frequency_{it}, \ Cumulative_{it}) \\
&+ \beta_{Firm-Level}(\text{Social sentiment}: \ Pos_{it}, \ Neg_{it}, \ Cumul_{it}) \\
&+ u_i(\text{Unobserved fixed effects for firm } i) \\
&+ \mu_t(\text{Unobserved fixed effects for time } t) \\
&+ \varepsilon_{it}(\text{Errors}) \text{ with} \varepsilon_{it} = \rho\varepsilon_{i,t-1} + \phi_{it} \text{ and } \phi_{it} \sim N(0, \ \sigma^2)
\end{aligned}$$

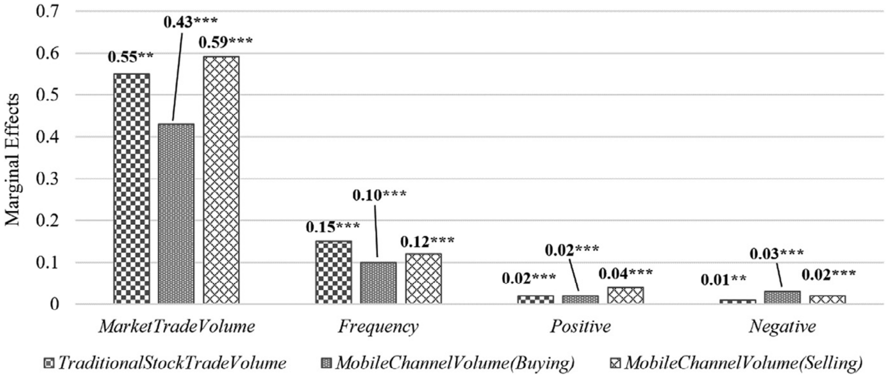


Fig. 2. Stock trade volume estimation results.

Table 3
Granger causality for panel data vector autoregression results.

Variables		*Traditional channel stock trade volume*	*Mobile channel trade volume (buying)*	*Mobile channel trade volume (selling)*
Volume	← *All*	15.18^{*}	36.06^{***}	34.18^{***}
Frequency	← *Volume*	28.01^{***}	5.62	27.13^{***}
Positive	← *Volume*	27.79^{***}	4.76	21.12^{***}
Negative	← *Volume*	9.26^{**}	0.74	15.50^{***}

Note: 251 firms. Variables represent % changes. Signif.: $^{*} = p < 0.1$; $^{**} = p < 0.05$; $^{***} = p < 0.01$.

4.3. Explanatory empirical results

First obtained were results on the impact of social media sentiment on stock-trading volume based on the FGLS model, as shown in Fig. 2. The results suggest that traders who transacted with mobile phones seemed to have been more easily swayed by social media sentiment. In contrast, the influence of social sentiment at the market level probably affected stock-trading in the traditional channel more.

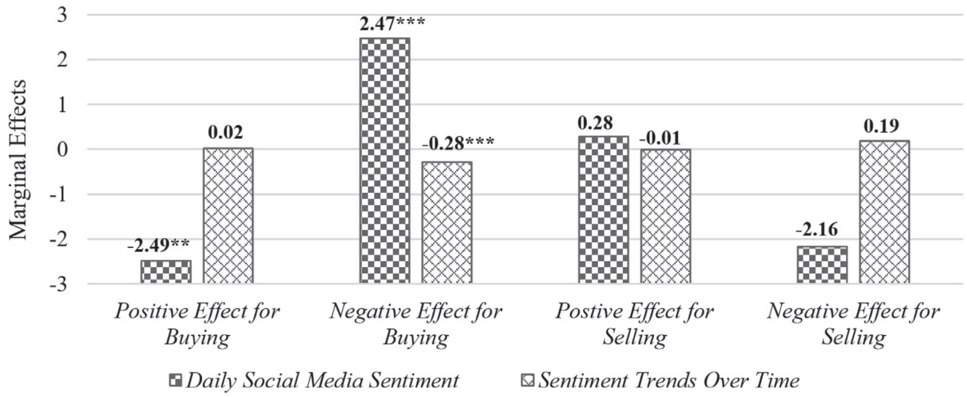
The results that were obtained show the direction of causality based on a PVAR-based Granger causality test, as shown in Table 3.

Social sentiment appears to have been a driver of causality, as one may expect. Social sentiment played a role in driving the observed stock-buying behavior of mobile phone-based traders, but there was less of an effect from social sentiment with respect to stock-selling behavior.[20] Next obtained were the KRLS data regularities results, as shown in Fig. 3.

The graphical presentation of the coefficient estimates shows that mobile phone-channel traders did more negative feedback trading in response to daily social sentiment for buying ($\beta = 2.47$, $p < 0.01$), though this effect was not significant for selling stocks ($\beta = -2.16$, $p > 0.10$). They tended to buy stocks when negative sentiment increased, and appear to have sold stocks when positive sentiment increased. This behavior of mobile-phone channel traders seems to go against the typical positive feedback trading strategy that has been often observed. The effects of sentiment trends that develop over time – one week of accumulated sentiment in this research – did not produce results with coefficient values that were very positive or negative though.

What makes this fusion analytics work relevant in the policy analytics context is that the results describe unusual feedback trading due to the inappropriate processing of social sentiment and opinions on the value of stocks and companies. This may be problematic from a regulatory point of view. The common wisdom is that no specific channel – whether making trades through a human broker by phone or in person, or via an e-trading software system on a home computer near to the market or distant from it – should be disadvantaged in the completion of fair trades for the investors who initiate them. Yet if buying and selling stock in the mobile phone-based channel leads to over-reacting investors due to the rapid dissemination of social sentiment, such uninformed

[20] Our utilization of measures for buying and selling behavior separately is related to individual trader actions, and how individuals may be differently influenced by social sentiment, depending on what they want to do. Market volume, in contrast, measures trades made in the entire market, so that for every bid that is made for stock, there must also be an offer to sell the stock. And so there is this tie-in between them.



Note. The word *Effect* under the four pairs of bars indicates the effects of positive or negative sentiment.

Fig. 3. KRLS results for social sentiment in the mobile phone-based channel.

trading behavior and the systematic losses that occur may become the subject of regulatory and governmental scrutiny.

5. Music semantics and the duration of music track popularity

Achieving a deeper understanding of how music popularity works in social environment has been of significant interest in industry and academia over the years. The research described in this section combines Machine Learning methods for analyzing music semantics, including an LDA topic model (Blei et al., 2003), with classification and prediction. Music semantics, Machine Learning methods, and explanatory empirics enable a new way to model, understand and interpret the top-chart ranking popularity of music tracks, compared to prior research. (Appendix A, Table A1, again, summarizes the fusion analytics research framework dimensions, and it will support the reader's understanding of why this research is a representative illustration.)

5.1. Context, data acquisition and observations on the dependent variables

Social networks, as a new medium for music distribution, have changed listener behavior dramatically by easing communication and music sharing among listeners. Music is a *durable information good* that can bring value and utility to the music industry and listeners, and social media to enhance the effects. Even one hit song can lead to the rise of a new music superstar. And classic songs can prompt people to remember singers years after their hit songs were released. This has led to more business value in the music industry, in spite of the impacts of artist-led music production, easier availability of online tracks and videos, and the transformation of the industry's value chain (Bockstedt et al., 2005). As a result, large and small music labels have paid increasing attention to how music can be promoted by social networks. Beyond CD sales, they now have partnerships with YouTube and Baidu, for example, to boost their digital sales through enhanced promotion (IFPI, 2012, 2013, 2015). By emphasizing precursors of high popularity, it is possible to discover digital traces of how a song became popular, and predict music superstars. Data were collected using extensive machine-base methods to make this possible, as shown in Fig. 4.

Multiple CS studies have examined how to gauge the importance of music acoustics and the content of the lyrics (Dhanaraj and Logan, 2005; Ni et al., 2011; Lee and Lee, 2015). Social Science, in contrast, has paid more attention to artist and album popularity, and non-musical factors, including Artist Reputation and superstardom, major label association, date of track release, P2P sharing activities, and social media buzz (Hamlen, 1991; Bhattacharjee et al., 2007; Grace et al., 2008; Strobl and Tucker, 2000). The present research supplements these factors with more fine-grained music semantics factors to provide fuller information about the drivers of the duration of music popularity in Last.fm.

Music in social networks often has considerable staying power when it achieves popularity over time and appeals to audience tastes. Assessment of the sustainability of music popularity online can be accomplished by modeling the *number of weeks duration* that a music track stays on the *top-chart ranking list* before dropping off. It is also possible to analyze variables and higher-level constructs that determine the popularity of music online.[21] The empirical modeling approach views the duration of a track's popularity on Last.fm in terms of three events: *track release*; *rise to top-rank chart listing*; and then *chart drop-off*. The last two measures in this process reflect the performance of a specific music track. *Duration* is the time that ensues from when the track reaches top-rank until it drops off. And *Time2TopRank* is the time from release to the track's first ranking among top-ranked songs. These measures represent how quickly and how long a track appeals to audience tastes.

For the empirical analysis of track popularity-related top-rank chart duration, 421 left-censored and 108 right-censored tracks were removed. Overall, 3,881 tracks by 477 music artists were used in the analysis. Fig. 5 depicts the empirical regularities. The bottom right shows a log distribution of music track top-rank chart list duration. Compared with the whole dataset, when censored data were removed, a smoother Gaussian distribution emerged, with a skewness of 0.42, and a kurtosis of 2.45.

The music social network under investigation, Last.fm, made it possible to collect many variables for higher-level constructs that include possible determinants of music popularity. For example, controls were included to capture differences in the music marketplace and the developing sophistication of Last.fm as a music social network. Beyond the duration popularity analysis, this research

[21] This was possible through the collection of data from Last.fm's Weekly Listening Chart, where social network members report what they listen to. This longitudinal study employs data from February 2005 to May 2015 on the top-150 music tracks and week-by-week rankings on the chart. The data cover 532 weeks and 12+ million music tracks. Few music tracks made it the top-150 chart though: just 4,410 tracks or 0.04% of the total: again, an instance of a "data needle in a haystack" (Chang et al., 2014).

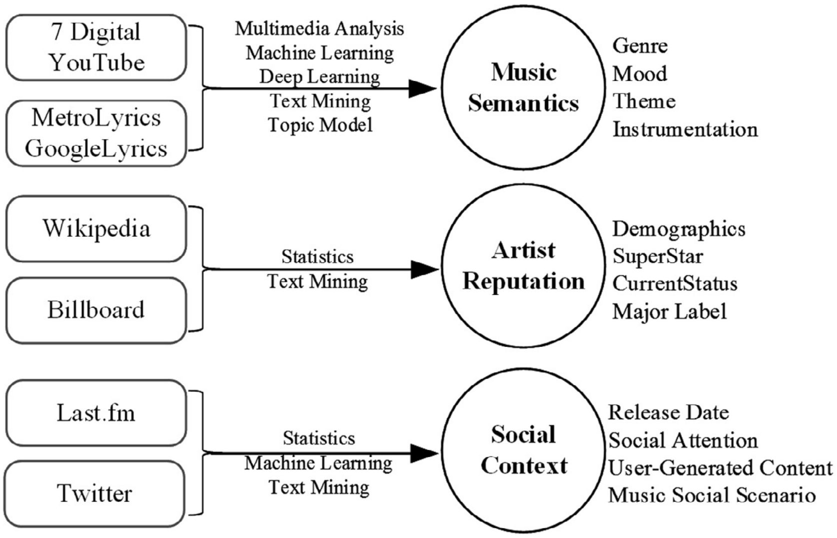
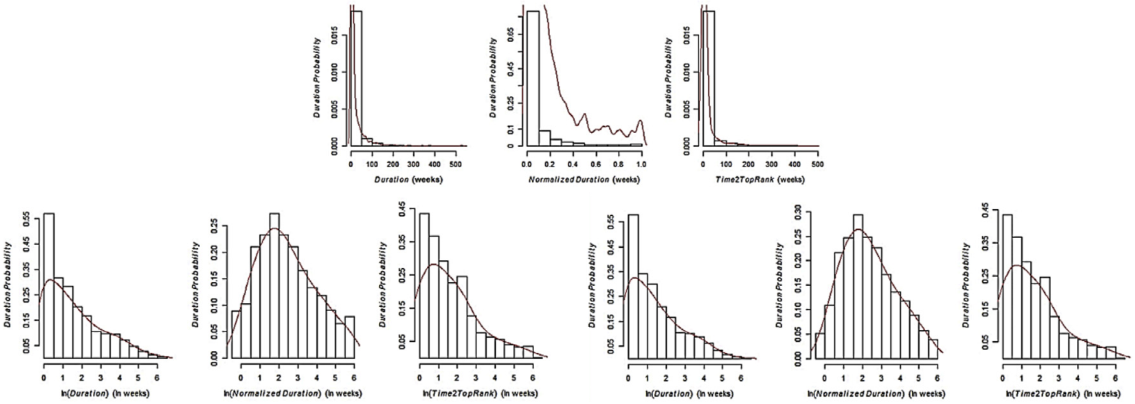


Fig. 4. Machine-based data collection/extraction methods for the Last.fm project.



Note. The top row shows music track popularity *Duration* and *Time2TopRank* for tracks in weeks. The left bottom row shows probability densities for ln(•) for the entire dataset (obs. = 4,410), and the right illustrates the similarity of the distributions after censored data were removed (obs. = 3,881). This is a robustness check that ensures data appropriateness.

Fig. 5. Raw and logarithmic distributions of popularity duration in weeks for music tracks.

also considered popularity patterns based on *Duration* and *Time2TopRank*.

5.2. Machine-based methods for data collection and extraction

Data collected from multiple sources were extracted and structured into three high-level constructs: *Music Semantics*, *Artist Reputation* and *Social Context*, as shown in Table 4.

5.2.1. Music Semantics

Representing the perceptual fundamentals of music tracks that attract listeners is a challenge. There are two parts to consider: *acoustic content* and the *lyrics*. People characterize music content through the use of well-known music semantic topics: *Mood*, *Genre*, *Theme*, and *Instrumentation*. The measures associated with these semantic concepts reflect how listeners perceive a music track's content. To learn the semantics of music tracks, Machine

Table 4
Music semantic topics and variables acquired with machine-based analytics.

Semantic topics	Variables acquired
Genre (18)	*Rock, Alternative, Indie, Pop, HipHop, Rap, Electronic, Metal, Folk, Soul, Blues, Country, R&B, Punk, Classic, Jazz, Experimental, Reggae*
Instrumentation (12)	*Cello, Guitar, Drumkit, Violin, Piano, Tuba, Flute, Clarinet, Saxophone, Trombone, Trumpet, Snare*
Mood (5)	*Passionate, Lively, Brooding, Humorous, Intense*
Theme (5)	*Life, Dance, Passion* (We, like, dance, young, live, good, sweet, dream) *In Love, Relationships* (You, love, like, baby, wanna, need, girl, feel) *Soul* (Eyes, heart, soul, fall, cold, dark, blue, blood, left) *Sad Life, Love* (Back, alone, long, over, wrong, lost, leave, remember) *Anger, Hostility* (Like, fuck, shit, rock, bitch, fucking, hit, damn)

Learning methods were applied to extract the key descriptive features from 30-s music clips and the lyrics of 4,410 tracks.[22] Through their analysis, a 167-dimensional *acoustic feature vector* was obtained, including *Tempo*, *Rhythm*, *Timbre* and other features.[23]

Further Machine Learning techniques were implemented, including the *support vector machine* (SVM) *method* (Basak et al., 2007). This is a discrimination algorithm that classifies a subset of data introduced to it, creating a separating multidimensional hyperplane in the process, to categorize the other remaining data. The theory behind this method is that it provides a means to construct a linear decision boundary between different classes of data, such that the margin or distance between them is maximized (Kecman et al., 2005).[24]

SVM learned the information content of three subconstructs to represent acoustic content for this research using a low-level *acoustic feature vector*, in terms of the variables associated with *Genre*, *Mood* and *Instrumentation* (Cheng and Shen, 2016).[25] The lyrics of songs were examined as text documents, and LDA was used to build *topic models* to learn the *semantic topics* in the 4,410-track dataset. LDA exhibited effective performance for classifying topics in the text based on *document-word-topic* relationships it identified. The topic model was run by varying the number of topics from 3 to 15. The process identified 5 topics as providing the best summary, with LDA hyperparameters for the higher-level characteristics of $\alpha = 2.0$, and $\beta = 0.1$, which were established after 3,000 iterations. The last row in Table 4 shows the semantic topics that emerged, with representative single or multiple word descriptors. Around 65% of the tracks in the dataset were about *Love and Life,* a *semantic empirical regularity* in this dataset, representing topics 1, 2, and 4. By combining acoustic and lyrics factors, a 40-dimensional higher-level *Music Semantics* construct was obtained for each music track.

5.2.2. Artist Reputation

The popularity duration of a track depends on who performs it, and famous artists attract large audiences. How to best measure *Artist Reputation* is open to debate though. One approach is to leverage news about the artists and their performances at the music awards shows. Another relevant tag is their labels. Major labels (with 20 sub-labels in the dataset) have more resources to produce and promote high-quality tracks. Additional reputation control variables from Wikipedia's and Billboard's charts were collected to identify whether an artist was a vocal performer or associated with a major label, and what was an artist's historical and current reputation in the market.

5.2.3. Social Context

Last.fm had 59.2 million users in August 2015. Users can "tag," "like," and "shout" about (comment on) tracks and artists. Social comments capture what people are interested in. Artists typically attract a group of followers over time, even when they are not that famous. The dataset also has listener information and *ReleaseDate*, *TrackFirstTopRank*, and *#TrackComments* from Last.fm in the first four weeks since a track was released, and *#TopRankTracks* released before a given new track was released. Taken together, these things led to a 54-dimension music construct vector for each track.

5.3. Explanatory modeling

Music track popularity duration is modeled with explanatory variables for the higher-level constructs to explain when top-rank chart list drop-off occurs. A *hazard function* specifies the duration until time t when an event of interest happens. A starting point is the *proportional hazard* (PH) *model*:

$$\lambda(t|\mathbf{X_i}) = \lambda_o(t)exp(X_1\beta_1^{PH} + X_2\beta_2^{PH} + \cdots) = \lambda_o(t)exp(\mathbf{X_i}\cdot\mathbf{B})$$

$\lambda_0(t)$ is the *baseline hazard* or likelihood that a track drops off the top-rank chart. $\mathbf{X}_i$ are explanatory variables for a track i, ($i = 1, 2,\ldots$) and β_i^{PH} are parameters to be estimated to gauge whether there are modifications to the hazard rate of drop-off beyond the baseline (Kleinbaum and Klein, 2006).

The *Weibull hazard function*, $\lambda(t)$, follows a monotonic curve, $\lambda(t) = \lambda z t^{z-1}$. There is a scale parameter λ, and z is a shape parameter, which allows the hazard function to be steeply declining, constant, or a function that increases at an accelerating rate. This fits many applied situations. Other distributions are non-monotonic. For example, the log-*logistic hazard function* $\lambda(t)$ follows a non-monotonic curve, $\lambda(t) = \lambda z t^{z-1}/(1 + \lambda t^z)$. It slowly decreases after reaching a peak (Hayat et al., 2010). This mimics the dynamics of situations that involve an initially increasing and later decreasing hazard rate, as with the diagnosis and treatment of leukemia and cancer. A *log-normal hazard function*, in contrast with a log-logistic function, is based on a normal distribution, with a positive, skewed distribution with a lower mean time-to-event and higher variance. This is used in Finance so that simulated price observations for equities that are less than the mean are not so extreme. It has also been applied in medicine to understand the occurrence of chest pain and the subsequent onset of heart disease (Hussain et al., 2014).

These three hazard functions were useful for assessing the duration of music track popularity. A simpler *ordinary least squares* (OLS) linear model also was used (Bhattacharjee et al., 2007). Such a linear model with a log-transformation yields an approximation to the more refined hazard models as a baseline for estimation, with both fixed and time-varying variables for each track:

$$\lambda(t) = f(\lambda,\ z,\ t;\ \mathbf{X}_{Music}\mathbf{B}^{OLS}_{Music},\ \mathbf{X}_{Artist}\mathbf{B}^{OLS}_{Artist},\ \mathbf{X}_{Social}\mathbf{B}^{OLS}_{Social})$$

5.4. Results

5.4.1. Duration model estimation

The results for the Last.fm data show that *music semantics*, *artist reputation*, and *social context* all drive the duration of popularity, but with different impacts, as in Fig. 6.

Genre and *Theme* were the two most important music semantics subconstructs. In the last 10 years, *Pop* genre music tracks had a 59.9% ($p < 0.01$) longer popularity duration (positive impact),[26]

[22] Using ML methods to learn music semantics is unaffected by data censoring problems. Acoustic content and lyrics exhibit track-to-track variation, but the commonality is the time-invariant nature of any music track.

[23] Each music clip was broken down into fifteen 2-s segments, and a 167-dimensional *acoustic feature vector* for each segment was extracted. This procedure yielded 66,150 vectors in total.

[24] The problem of *data sparseness* often arises with support vector machine-based learning. Training the learning algorithm is made more difficult due to the lack of a large enough number of instances for individual users (Cha et al., 2009; Li et al., 2015).

[25] 100 labeled clips for each semantic dimension were selected. For each clip, a 15-segment acoustic feature vector was selected. An SVM model supported learning and testing for the 52,500 labeled segments, and then was used to label the acoustic data for 4,410 acoustic tracks. Finally, construct filtering was applied to combine the labels of the 15×35 labeled clip vector into 1×35 construct representing the acoustic semantics.

[26] This result follows since the dependent variable is in log form while the explanatory variable is not. Comparing *Popular* and *non-Popular* music, the difference is $1 - e^{0.47} = 59.9\%$ (Bhattacharjee et al., 2007). The other results are computed similarly.

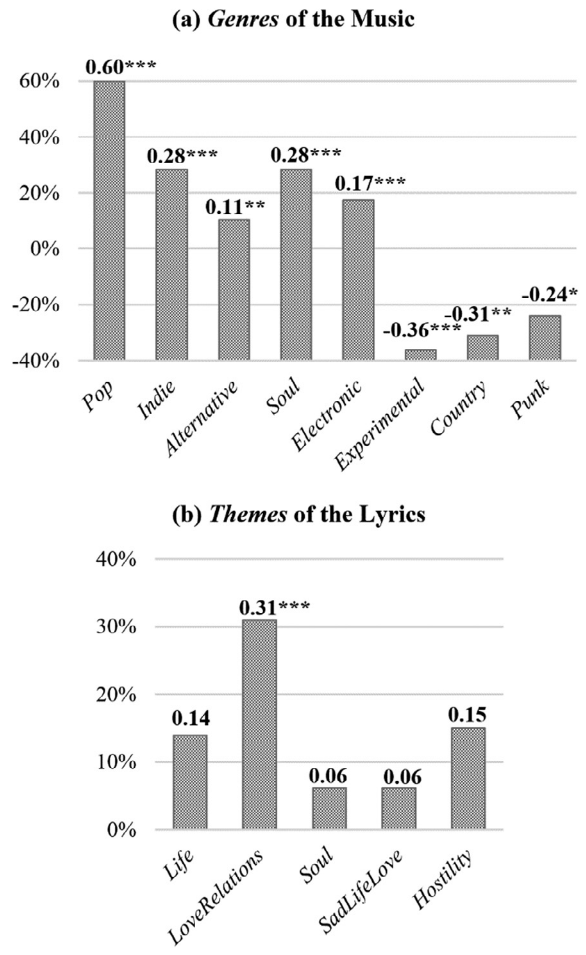


Fig. 6. Estimation results for music track top-rank chart popularity duration model.

while *Experimental* music tracks had 36.0% ($p < 0.01$) shorter popularity duration (negative impact). Tracks representing the *Life*, *Love*, and *Relationships* themes were popular longer. For example, music with the theme of *LoveRelations* had a 31.0% ($p < 0.01$) longer popularity duration (positive impact). This suggests why the music of artists such as Beyoncé and Adele has enjoyed such high popularity in recent years. For *Instrumentation*, *Piano* and *Guitar* were less successful; they had negative popularity duration effects. In the *Social Context*, tracks released during Christmas time in North America and Europe had a greater chance for longer popularity, but there was no guarantee amid the competition.

When a track first rose to the top-rank chart is important, and the higher the first rank, the longer should be its duration. An artist's previously top-ranked tracks before a new track was released demonstrated the significant effect of the growth of the artist's social network: every artist has their own unique group of followers. If listeners were already fans of an artist, they tended to adopt that artist's next album ($p < 0.01$). Among the control variables for *Artist Reputation*, *Major Label* turned out to not be significant. However, tracks from the same album were typically tied to an album's performance as a *co-integrating pair*, and correlated over time in their popularity durations, even if they were not identical ($p < 0.01$). In contrast, artist *Current Reputation* had a positive effect ($p < 0.01$), while *Historical Reputation* had a negative impact ($p < 0.01$). This suggests that listeners paid more attention to the recent tracks of famous music artists instead of famous older or less active artists.

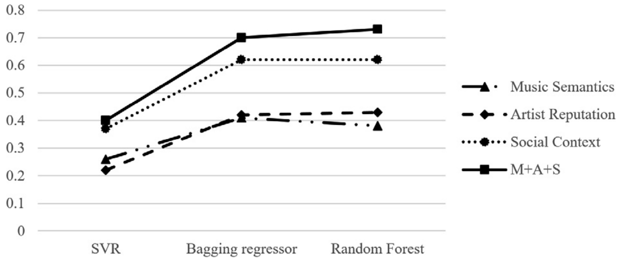


Fig. 7. Music popularity duration prediction: SVM, bagging, random forest.

5.4.2. Machine-based popularity prediction

The duration estimation yielded information about the significant explanatory variables that can also be used to predict music track popularity. The predictions were assessed based on *support vector regression* (SVR), the bagging regressor, and the random forest procedures in statistical Machine Learning (SciKit Learn, 2017).[27] (see Fig. 7.) 80% of the data observations were selected for 10-fold cross-validation, and then Machine Learning methods were applied to the remaining 20%.

There was a high correlation between predicted and actual music popularity durations based on the random forest procedure, which achieved a near-maximum for *Music Semantics* and *Social Context* ($\rho = 0.72$, $p < 0.01$). An additional 1% increment to the maximum with *Artist Reputation* included was not meaningful ($\rho = 0.73$, $p < 0.01$). This information on the 70%+ correlation is helpful for a record label or independent producer to assess future performance, by improving the match of music tracks to market outcomes. This may clarify how much money marketers ought to be willing to spend to improve the likelihood that a track will have longer popularity, since popularity is correlated with future sales.[28]

5.4.3. Popularity sustainability pattern analysis

The discussion on this research closes with results that were obtained through the use of two measures of music track popularity – *Duration* and *Time2TopRank* – as shown in Table 5. This further indicates the high relevance of this kind of approach to extract deep knowledge about popularity patterns, as in Table 6. Through unsupervised analysis, music tracks were classified into six popularity patterns, all based on Machine Learning approaches.

This study represents policy analysis work and fusion analytics in the sense that estimates of the staying power of a music artist's popularity and the top chart sustainability of much tracks are things that music labels and artist management services actively view as worthy of financial investment, advertising and other kinds of outreach. These can support the growth of popularity and a stronger audience following in the market. A basic rule-of-thumb is that data analytics have potentially high information value. The payoff in terms of marginal expenditure for promotion should be roughly equal to marginal revenue in music and related sales – the $MC = MR$ relation in microeconomics – to maximize the associated value. Decisions on this can be tied to data analytics insights,

[27] *Bagging* is short for *bootstrap aggregating*, a form of model averaging in machine learning that combines the classifications of different training datasets to smoothe discrimination and avoid unnecessary errors. Its roots are in decision tree-style learning algorithms, especially *B-trees*. The *random forest* procedure builds on bagging by randomizing feature selection learned through iterative execution of the algorithm. The reader should see Breiman (1994) and James et al., 2013 for additional details.

[28] This perspective is not without controversy though. Many music artists object that record labels' interests in maximizing profit are at odds with an artist's interest in artistic creation and individual expression. An example is David Bowie's album, "Blackstar." *Artist Reputation* was key in driving popularity for music produced just prior to Bowie's early passing in January 2016 – and even more so after that event occurred.

Table 5
Duration and *Time2TopRank* in quartiles for whole dataset and without censored data.

Dataset (All Obs.)	Min	Max	Mean (SD)	1st quartile	Median	3rd quartile
Duration	1	532	17.9 (47.2)	1	3	11
Time2TopRank	1	473	20.1 (53.7)	2	4	10
Dataset (without Censored Obs.)						
Duration	1	504	13.1 (31.6)	1	3	9
Time2TopRank	1	395	11.4 (27.7)	2	3	9

Notes: Obs.: February 2005 to May 2015. Full dataset: 4,410 obs; dataset after censored obs. were removed: 3,881 obs.

Table 6
Music track popularity patterns discovered in this research.

Pattern name	Description
Flash-in-the-Pan, Short Popularity	An artist's tracks stayed in the top-rank chart list for less than three weeks before dropping off. 56% of 3,881 tracks belonged to this pattern, and 60% attracted attention during the month since they released. The effect did not persist though.
Overnight Sensation, Long Popularity	Tracks in this pattern become truly popular. They attracted attention after they were release, and stayed in the top-rank chart list for a long time.
Slower Rise, Long Popularity	These tracks stayed "under the radar" for quite a while, but eventually emerged on the top-rank chart list and attract more listeners with high sustainability.
Normal Rise, Long Popularity	This pattern of popularity duration occurred when a track needed a normal amount of time to rise to top-rank, but still had a lengthy period of popularity.
Faster Rise, Average Popularity	This pattern was common, and such tracks achieved average top-rank chart list survival duration, but they reached the top-ranking more quickly.
Slower Rise, Average Popularity	This was similar to the prior pattern, only the artists' tracks took a longer time to achieve a position in the top-rank chart list.

possible value-relevant actions that carry out a strategy, and subsequent evaluation that must be done (Kauffman, 2014a).

6. Preferences in household TV viewing

Do households exhibit concentrated viewing preferences for the TV channels and programs genres they watch? The context, data and analytics for answering this question are discussed next. Some of them are common to this and the research presented in Section 7, especially the demographics of the households, the viewing data source, and the number of channels the household subscribed to. (The reader is again encouraged to scan the fusion analytics research framework. See Appendix A, Table A1.)

6.1. Context, data acquisition, and machine-based data analytics

Chang et al. (2014) collected data from a large digital entertainment firm, which broadcasted over 170 cable TV channels to several hundred thousand households in its operating territory. Three kinds of anonymized household data were obtained: (1) demographics and residence information; (2) bundle subscriptions for accessible channels; and (3) set-top box data that tracked TV channel viewing behavior. The research design used a *random sampling approach* with iteration for 10,000 cable TV-subscribing households employed for one month of data in 2011.[29] Observations with missing values or mismatches between the subscriptions and household viewing choices were eliminated, leading to a sample of 4,720 households, and essentially all of their TV viewing data, including moment-to-moment streaming data.

The dependent variable, *ViewingConcentration*,[30] is a 0–1 proxy variable for the diversity of household-level viewing patterns. It employs the *Gini coefficient* (World Bank, 2013), adapted to gauge differences in household viewing times across the TV channels the household viewed.[31]

Also explored were household TV viewing patterns in terms of a number of main effects variables, and a set of controls. *#SubscribedChannels* captures how many channels the household subscribed to. Their TV viewing choices were constrained by their subscription decisions. The more channels a household was able to view, the more likely were household members to be able to find what they liked. They also were able to watch many different channels, and seek variety in their viewing experience. With multiple household members, the sum of their viewing – an expression of *family-level unitary preferences* – resulted in varied viewing patterns. *ViewingTime* captures how long households spent on TV viewing within the observation period. More time gave viewers more chances to watch different channels. *PreferenceClusters* use machine-based dummies from cluster analysis to represent viewer content preferences.

For the control variables, individual household demographics did not support very meaningful implications for household viewing patterns. So the only demographic variables considered were those that reflect household information. *SubscriberAge* was for

[29] Limited computer memory and run-time capacity make it difficult for very large datasets to be processed and analyzed directly using statistics software (Cohen et al., 2009).

[30] The *sigmoid function* to perform the logit estimation was considered, with $0 \leq Viewing\ \ Concentration = f\ \ (Variables) = g\ \ (Marginal\ \ Effect \times Variable) \leq 1$. This function corresponds to $g(Marginal\ Effect \times Variable) = 1/[1 + e^{-(Marginal\ Effect\ \times\ Variable)}]$. In this case, if the function g was greater than or equal to 0.5, the dependent variable was estimated as 1 and 0 otherwise. The Gini coefficient was obtained by computation. If it was more than 0.5, then *ViewingConcentration* was 1, otherwise 0.

[31] Ordinary least squares estimation usually imposes two conditions when proportional variables are estimated: (1) the *conditional-expectation function* must be nonlinear since it maps onto a bounded interval; and (2) its variance must be heteroskedastic, since the variance will approach zero as the mean approaches 0 or 1 (Kieschnick and McCullough, 2003).

those who actually subscribed to cable TV services. It provides useful information on family household demographics, such as younger people, or middle-aged people, or elderly people, who all have different preferences and patterns of TV viewing. The control variable, *#Rooms* in a family's residence, offered a way to control for family income, since larger and wealthier families typically lived in larger residences in the study area during the observation period.[32] All pair-wise correlations between variables were less than 30%, so there were no estimation issues.

6.2. Machine-based data analytics methods

TV programs were labeled in multiple genres based on their contents (Creeber et al., 2001). To set up the econometrics work, *cluster analysis* was used in this research. It constructs meaningful sub-groups of individuals or objects (Haaijer et al., 1998) related to a classification problem, and can be used as an effective way to identify customer preferences (Yankelovich and Meer, 2006). To identify different viewing preference patterns, the *k-means algorithm* was used to classify households into a number of clusters by minimizing the sum of intra-cluster distances, while simultaneously maximizing the sum of inter-cluster distances among the points that were members of a cluster (Tan et al., 2005).[33] Two statistical indexes were adopted that were useful for evaluating the cluster quality results for the *k*-means analysis:[34] the *Davies-Bouldin index* and *silhouette values* (Rousseeuw, 1987). They represented their main program contents and differed somewhat from the research sponsor's definitions. The number of clusters was varied for $k = 2$ to 20, and the analysis was run for each value of k 100 times.[35] This yielded average Davies-Bouldin index and silhouette values. The optimal silhouette value occurred for nine clusters.[36] Fig. 8 shows the patterns for *household-level proportion of viewing time spent on different genres* (*TimeProportion*). Household channel viewing times were combined to get genre viewing time distributions.

The shape of each centroid was determined in relation to the average percentage of time spent for each genre by households within the cluster relative to their time spent over all of the clusters. The first 8 clusters showed that households had strong preferences for 1 of the 8 genres. For example, households belonging to Cluster 2 on average spent over 70% of their viewing time on *Drama* channels, with only 30% spent on channels of other genres. Most households could be classified via patterns that suggested preferences for specific genres, with lesser preferences for the rest. The last, Cluster 9, showed balanced preferences for multiple genres.[37] To capture the diversity of households, 9 categories for *ClusterPreference* were identified, with a base case and 8 dummies for the econometric model.

6.3. Econometric modeling

A *limited dependent variable model* [38] was used to estimate whether households exhibit *concentrated viewing preferences*, and whether their preference clusters mattered. It has the following form:

$$\begin{aligned}\mathit{ViewingConcentration} = \\ &\exp(\beta_0 + \beta_1 \#\mathit{SubscribedChannels} + \beta_2 \mathit{ViewingTime} \\ &\quad + \beta_3 \mathit{PreferenceClusters} + \beta_4 \#\mathit{Rooms} + \beta_5 \mathit{SubscriberAge})\end{aligned}$$

6.3.1 Limited dependent variable estimation

The logit model, which instantiates an *average value-based estimation*, was done two ways. The first way involved the *beta distribution regression* for the dependent variable. It is used to estimate dependent variables that represent proportions (Ferrari and Cribari-Neto, 2004). The second way involved a *quasi-likelihood function-based regression* to obtain the model's parameters. The latter one is more efficient when the data are sufficiently large to justify the asymptotic arguments underlying the quasi-likelihood approach (Kieschnick and McCullough, 2003).

6.3.2. Quantile regression

Since other modeling issues could have harmed the results, *quantile regression* was implemented to handle this possibility (Yu and Moyeed, 2001). Modeling the quantiles of a dependent variable's probability distribution as a function of a set of independent variables is equivalent to knowing the entire conditional distribution of the dependent variable at different quantiles, in contrast to average value-based analysis.[39] This is a useful refinement to ensure robustness when households are in different quantiles for demographics, channel subscriptions, and other TV viewing-related behavior.

6.4. Explanatory Econometrics results

Although baseline analysis showed that *#SubscribedChannels* did not affect *ViewingConcentration*, as seen in Fig. 9, quantile regression revealed other details that indicate the usefulness of the explanatory model-based explanation. The beta distribution and quasi-likelihood function-based regression results are not presented due to limited space; instead, our focus is on the quantile regression results.

ViewingConcentration was affected by *#SubscribedChannels*, but only for households with very high or very low levels of *ViewingConcentration* – and in the opposite direction. This is like a *cancelling effect* that explains why there was no significant relationship

[32] A reviewer pointed out that there may also be a difference between public and private housing. In our dataset, there is only one kind of public housing. There may be another nuance to consider though: a possible interaction effect between *DwellType* and *Rooms*. We found no evidence of a meaningful difference for this beyond the single-variable effects.

[33] *Euclidean distance* is intuitive relative to the outputs generated by the algorithm. It forms each cluster as a *hypersphere of arbitrary shape* positioned around a *centroid vector* in higher-order dimensional space, based on the dimensions it uses to form the clusters. A *centroid vector* in a cluster is a mean set of values for all the observations that belong to it. The observations around the centroid show more or less variation in the values of their descriptive dimensions relative to the centroid vector.

[34] A classic algorithm, *k-means clustering with Euclidean distance*, was employed. See Bishop (1995) for additional details.

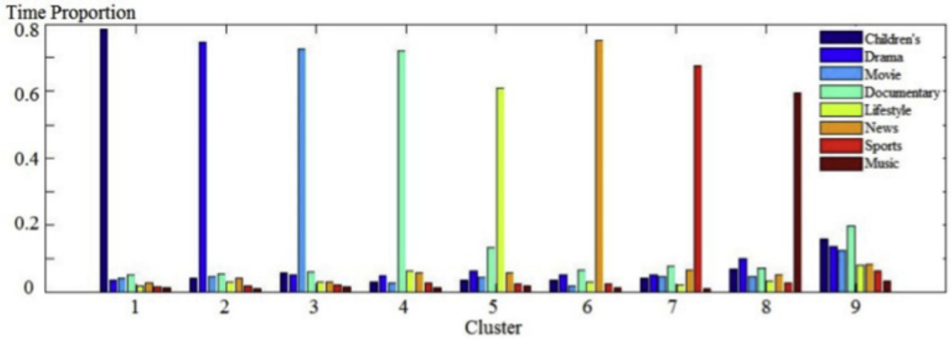
[35] Two limitations affect how a *k-means algorithm* establishes clusters: *k* must be determined with analysis (Ray and Turi, 1999); and unstable clusters can result from *randomly-selected centroids* used to jump-start analysis of the data (Ben-David et al., 2007).

[36] There was no number of clusters *k* that optimized both measures. It was determined that 9 clusters were appropriate for the 8-genre setting: 8 clusters with preferences for each of the individual genres, and 1 cluster with mixed preferences.

[37] The clusters were labeled with terms that are common to digital entertainment firms that operate around the world: *Children*, *Drama*, *Movies*, *Documentary*, *Lifestyle*, *News*, *Sports*, *Music*, and *Mixed*.

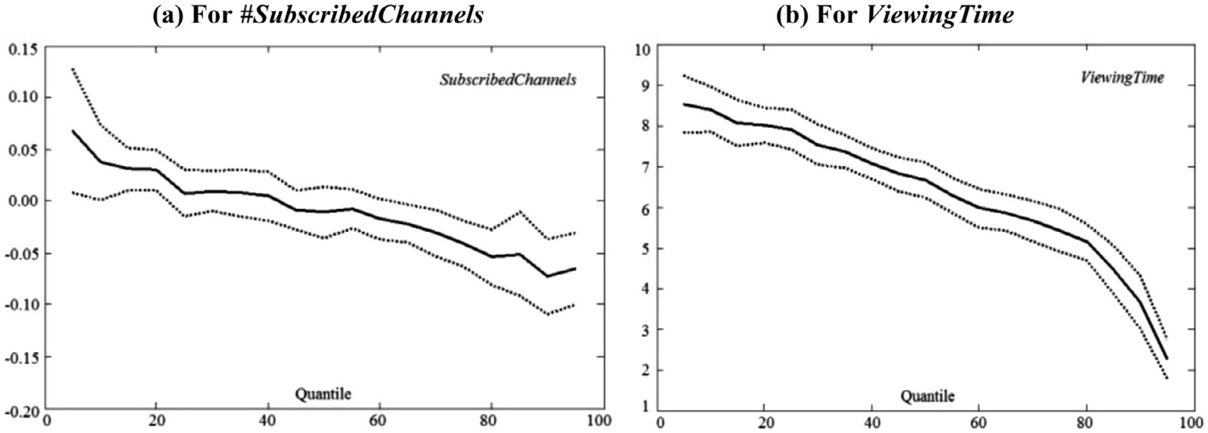
[38] A second model was estimated using a similar functional form and explanatory variables for *ViewingEfficiency*, the fraction of available channels that a household watched for at least thirty minutes during the study period. This did not have a linear relationship with *ViewingTime*, so a piecewise spline regression was used. This allowed for changes in slope of the relationships between the independent and dependent variables, with the restriction that the regression line was continuous (Marsh and Cormier, 2001).

[39] The main effects, *#SubscribedChannels* and *ViewingTime*, also were estimated at five intervals of the dependent variable, *ViewingConcentraion*. The *beta distribution* and *quasi-likelihood function-based estimation* were used.



Note. The *x*-axis represents clusters and the *y*-axis is the *Average % of Total Viewing Time*, which is shown as *TimeProportion.* The height of each bar is the proportion of a household's time spent on each genre for those genres identified as being in the cluster.

Fig. 8. TV viewership cluster centroid shapes.


(a) For #*SubscribedChannels*

(b) For *ViewingTime*

Fig. 9. Quantile regression results.

in the average value-based regression results for the base case model. So the use of quantile regression was justified.

The effect of *ViewingTime* on *ViewingConcentration* decreased from $\beta = 8.524$ (SE = 0.354, $p < 0.01$) at the 5% quantile, to $\beta = 2.259$ (SE = 0.250, $p < 0.01$) at the 95% quantile. This showed that, for the study dataset at least, households with higher *ViewingConcentration* were less sensitive to having more *ViewingTime*: for them, an increase in *ViewingTime* was associated with a smaller increase in *ViewingConcentration* than for other households. Finally, regarding the *PreferenceClusters* variable, households in the *Children* cluster had the highest *ViewingConcentration* at $\beta = 0.070$ (SE = 0.027, $p < 0.05$), whereas households in the *Mixed* cluster had the lowest with $\beta = -0.150$ (SE = 0.030, $p < 0.01$). This information from the data analytics was useful for management to think through the subscription packages offered at the time, relative to the different levels of market demand present.

This study offers another appropriate example of fusion analytics methods to support policy analysis because it can help a digital entertainment provider to more deeply understand the nature of demand for the different programming genres that it offers, along with the differences that are observed in household preferences, and the effects of viewing time on the relative concentration of household viewing across the content that is offered. As with other explanatory empirical methods, those used here provide the marginal effects associated with the different variables that were studied, as an evidence-based approach for the construction of different programming bundles and content offerings. The deep household-level insights are the basis in this context for improved services.

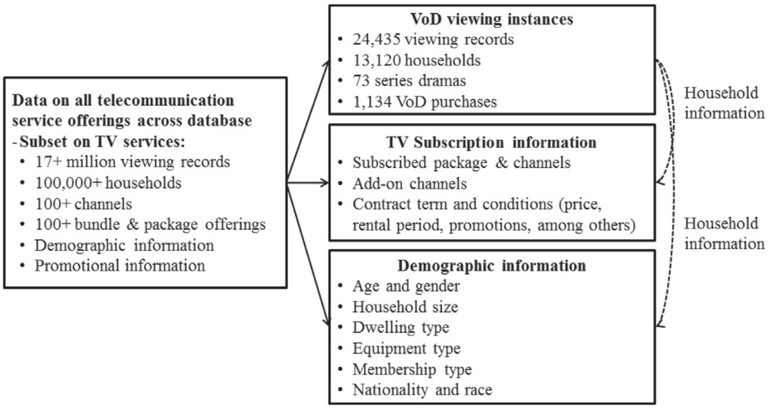
7. Household content sampling and video-on-demand purchases

The next application of fusion analytics was designed to obtain business insights from household-level *video-on-demand* (VoD) viewing activities in data-at-scale from a digital entertainment firms' subscribers. Advances in digital distribution technologies have made it possible for digital entertainment content providers to repackage and sell their content on a stand-alone, on-demand basis.[40] (See Appendix A, Table A1 for an overview of the applicable fusion analytics research characteristics.)

7.1. Context, data acquisition, and machine-based data analytics

In sponsored research that was undertaken during the past several years, Hoang and Kauffman (2016) collected 17+ million household cable TV viewing records for a one-month period in 2011 from a large service provider. The authors focused on viewing

[40] VoD service is an important source of revenue for digital entertainment and telecom firms. The VoD market is expected to reach US$45.25 billion in 2018, at a cumulative annual growth rate (CAGR) of 16.5% (Lafayette, 2014, Newswire, 2013).



Note. The solid arrows in the middle of the figure indicate the three categories of data that are related to telecom service offerings for digital entertainment. The dashed arrows on the right of the figure indicate that the TV subscription information and the demographic information on households were machine-matched to the instances of VoD viewing of series dramas, both for episode sampling and for VoD series purchasing.

Fig. 10. Overview of the data and variables for vod sampling and purchases.

records captured via its household-level TV set-top boxes that enabled extraction of VoD sampling, purchases and viewing for series dramas offered in the market. A *series* consists of multiple episodes that are bundled, sold with an attractive discount (and margin), to create strong demand from niche viewers (Dennis and Gray, 2013). To stimulate demand for their digital entertainment goods, firms offer first episodes as free samples. Then, households are able to purchase individual episodes or the entire series at a discount. Of interest was to study the effectiveness of the service provider's sampling strategy for TV VoDs, based on a vast amount of moment-to-moment streaming data and other household-level data. Fig. 10 presents the machine-based approach to the acquisition of the data, and elimination of duplicate records.

Altogether, 73 series dramas were offered during the study period, accounting for 24,435 free-episode sample viewing records, and 1,134 VoD purchases across 13,120 households. The dataset allowed observations of how households sample and purchase VoDs, given their subscription packages. The main dependent variable, *VoDPurchases,* represents the number of series a household purchased. The main explanatory variable, *FreeEpisodes*, refers to the number of VoDs that the household sampled at least once.

Digital content was sometimes consumed such that one type of programming hindered the consumption of other types. The explanatory variable, *SubscribedChannels*, captures how many channels the household subscribed to. Households with many alternatives may not be very interested in additional VoD content. Their current subscription packages reveal a household's expected level of utility from TV viewing and its willingness-to-pay. Another variable, *AddOnChannels*, refers to additional channels that a household could choose among. When households already have add-on channels that matched their viewing preferences, other content is less desirable for them. Nevertheless, these households actually expressed a higher willingness-to-pay for special content, and were more likely to purchase more VoD dramas if they triggered their viewing interest, representing demand for variety.

Also considered were several household-level demographic variables that may have influenced households' likelihood to purchase VoDs. The variable, *HouseholdSize*, was intended to reflect higher demand for TV services due to the number of household members. Also, *DwellingType* identified a household as occupying a house with its own land or an apartment in a large building, and was intended to reveal household spending power. After closer examination though, the demographic variables ended up yielding little additional explanatory power for VoD purchases. Moreover, eliminating them permitted the authors to retain a larger dataset for estimation, since not all of the households provided their demographic information during their digital entertainment services sign-ups. The dataset for estimation comprised 7,932 households with complete subscription information and set-top box data.

7.2. Machine-based methods for data collection

Even with help from the research sponsor's management team, the dataset was somewhat less than ideal. It was neither as long, deep or rich a time-series as was hoped for in terms of the variables for study. Some were not accessible, due to national data privacy regulations on customer-related personally-identifiable information that prohibited data sharing across multiple industries, including digital entertainment. Others were less sensitive but too costly to retrieve. The innovation in data collection involved machine-based simultaneous extraction and matching of qualifying data from different large-scale databases. The data extraction technique was developed based on understanding how different business processes were linked for service providers in multiple areas of the *triple-play business model* (phone, Internet and TV services). This required a smart algorithm and machine-based data crawler.

An important aspect of empirical research with consumer data is to obtain as deep an understanding of consumer behavior as the data will allow. As a result, data were extracted from multiple sources to bring together subscription and demographic information on households to acquire a sufficient number of observations for the empirical modeling analysis. With the large number of missing values for the demographic variables though, machine-based data collection was implemented to make choices about usable *data rectangles* (Techatassanasoontorn and Kauffman, 2014). The idea of a data rectangle is to support decisions on the rows and columns of data tables to run. For example, it is possible to use many rows of observations but fewer columns of variables, which produces a *tall but narrow data rectangle* in the sense of an SQL data table, and is most effective for including the largest number of data records. It also is possible to consider more columns of variables but fewer rows of data, which produces a *shorter but wider* rectangle, with fewer records but more variables to study to address each of the study's research objectives. The latter approach often makes it hard to establish statistical significance in empirical models that employ data stratification (e.g., fixed effects, time-wise effects, and so on), and so a taller but narrower data rectangle may be preferred by the analyst. See Fig. 11 for additional details about the machine-based methods used to construct the data rectangles for explanatory estimation.

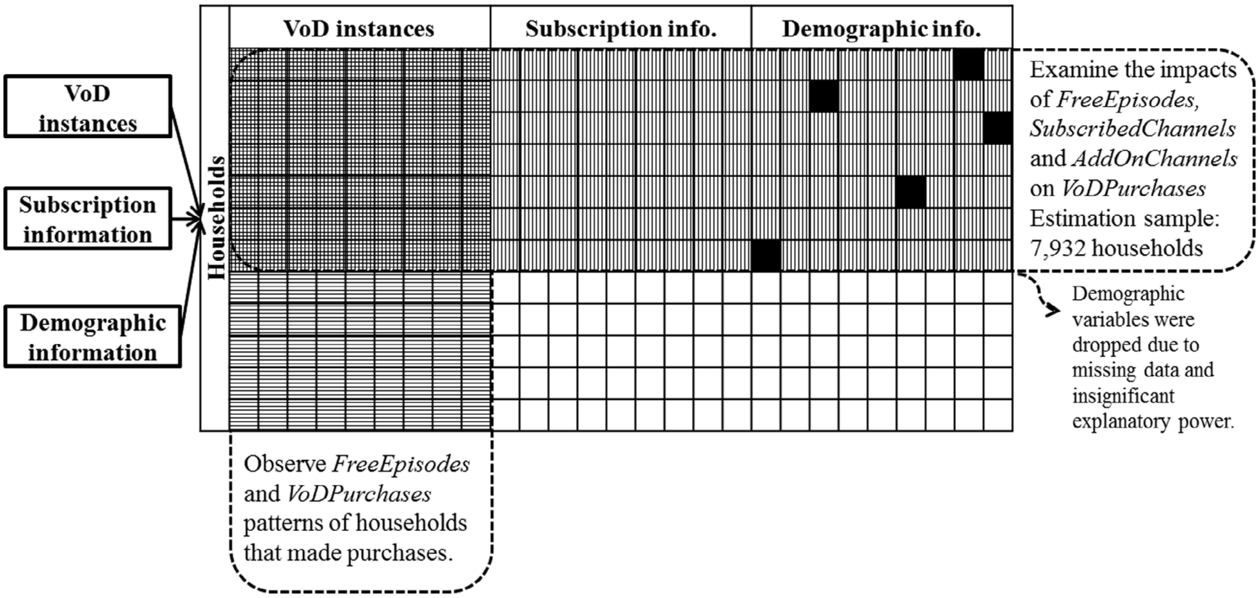
To observe the impact of sampling strategies, the research employs a shorter and wider data rectangle involving household subscription and demographic information. In data rectangle terms, this means that there were fewer rows of observations were used from the data table, while more variables for the observations were included. Altogether, there were 7,932 households with complete subscription information in this short data rectangle, out of 13,120 households in the original, taller rectangle of the full dataset. To explore the viewing and purchase patterns of the households, however, the viewing and purchase records of 13,120 households (a taller rectangle with more observations, and a narrower one with fewer variables) were used. Balancing the height and width of the data rectangles, as a basis for providing meaningful results, made different datasets usable for the empirical analysis. Table 7 includes the descriptive statistics for the dataset involving 7,932 households.

7.3. Econometric modeling and methods

The research employed a baseline model to represent the connection between the number of *VoDPurchases* and other factors: the number of *FreeEpisodes*, *SubscribedChannels* and *AddOnChannels*. The dependent variable of interest, as was noted earlier, is *VoDPurchases*, which yielded:

$$\begin{aligned} \mathit{VoDPurchases} = f(&\mathit{FreeEpisodes}, \mathit{SubscribedChannels}, \\ &\mathit{AddOnChannels}; \mathit{Controls}; \mathit{ErrorTerm}) \end{aligned}$$

If a household did not purchase any dramas, then *VoDPurchases* was left-censored at 0. Due to time and budget constraints, some households did not make many purchases; and the observed maximum was just 7 dramas. Thus, the VoD purchase count was either 0 or a small positive number. Count data models of different forms were used to estimate the explanatory model for the dataset. *Count data models* restrict the dependent variable to non-negative integer values and take into account the relationship between the mean and variance of the distribution that is used to characterize the dependent variable.



Examine the impacts of *FreeEpisodes*, *SubscribedChannels* and *AddOnChannels* on *VoDPurchases* Estimation sample: 7,932 households

Demographic variables were dropped due to missing data and insignificant explanatory power.

Observe *FreeEpisodes* and *VoDPurchases* patterns of households that made purchases.

Note. A challenge in this research was not only to acquire the appropriate "data needles in a large digital data haystack," but also to cross-compare the data to ensure that the identified data records had complete household demographics, household subscription information, and VoD set-top box viewing details for free episodes, paid episodes, and paid series. A further challenge involved handling errors and anomalies in the various very large databases, and this required algorithmic record deduplication, to ensure the integrity of the data in support of explanatory econometric analysis.

Fig. 11. Machine-based data rectangle construction: acquisition and record de-duplication.

Table 7
Overview of the VoD sampling and purchase dataset.

Variables	Descriptive statistics					Correlations			
	Min	Med	Mean	Max	SD	1	2	3	4
VoDPurchases	0	0	0.11	7	0.43	1.000			
FreeEpisodes	0	1	2.18	26	2.16	0.146	1.000		
SubscribedChannels	0	14	14.70	41	6.25	0.104	-0.028	1.000	
AddOnChannels	0	2	3.16	25	3.04	0.122	-0.014	0.648	1.000

Note: Obs.: 7,932. The least-correlated variables are *FreeEpisodes* and *AddOnChannels* (−0.014), and the most correlated ones are *SubscribedChannels* and *AddOnChannels* (0.648).

7.3.1. *Poisson regression*

In this discrete regression model for count data, the events to be estimated are independent of one another (Kauffman et al., 2012). Its advantage is its use of the Poisson distribution, which does not restrict the values of the dependent or independent variables. The underlying distribution of the dependent variable must exhibit equal means and variances, which sometimes does not match up with real-world considerations. This is used as a *baseline* for the count data models.

7.3.2. *Negative binomial model*

The dataset has characteristics that do not match the standard Poisson model very well. The dependent variable matrix is sparse, which is common in purchase conversion research settings in Marketing. 91.3% of households did not purchase any dramas, a larger proportion than a normal distribution might suggest. When the *conditional variance* of the dependent variable exceeds the *conditional mean*, the estimated values of the parameters will tend to be greater than predicted, a sign of *overdispersion*; and the standard errors of the parameters estimated will be underestimated (Cameron and Trivedi, 1990). *Negative binomial regression* generalizes Poisson regression, with an extra parameter to model overdispersed data. Its confidence intervals are narrower than a Poisson model's.

7.3.3. *Zero-inflated negative binomial and hurdle models*

Poisson regression also assumes that the 0 s and non-0 s come from the same *data-generating process* (Cragg, 1971). This does not hold true in this setting though. *Zero-inflated and hurdle models* relax this assumption (Hu et al., 2012). The *hurdle model* assumes there is a Bernoulli probability that governs the binary outcome for the count variable having a 0 or *positive realization*. Once the hurdle is crossed, and a positive realization occurs, the conditional distribution of the positive outcomes can be represented by a *truncated-at-zero count data model* (Gurmu and Trivedi, 1996; Jain et al., 1995). With zero-inflated models though, the response variable is a mixture of the Bernoulli and Poisson distributions. Model choice is based on knowledge of what causes the excess 0s.

Relative to the price of the drama series offerings, which can be expensive in comparison to other digital entertainment services, households must decide what they are willing to pay for. A *no-purchase decision* may result from two different processes. If a household does not have time or money to watch a whole series, they will not purchase it regardless of whether they watched a free episode. Yet, if a household has time and money, then its decision-making process can be represented as a *count process* influenced by the variable *FreeEpisodes*. The expected count for different values of *k* is:

$$\begin{aligned} E(VoDPurchases = k) &= Prob(HouseholdWithConstraints) \cdot 0 \\ &\quad + Prob(HouseholdWithoutConstraints) \cdot \\ &\quad E(y = k | HouseholdWithoutConstraints) \end{aligned}$$

To account for the excess 0 s from the two different processes, a *zero-inflated negative binomial* (ZINB) *model* (Greene, 2007) was used. Its specification has two parts: a *logit model* and a *negative binomial model*. The logit part models the probability of excess 0 s independently; the probability of *VoDPurchases* = 0, if a household has neither budget nor time.[41]

7.4. *Explanatory Econometrics results*

7.4.1. *Poisson model estimation*

The coefficients of the number of *FreeEpisodes, SubscribedChannels* and *AddOnChannels* were positive and significant in the Poisson estimation. The coefficient for *FreeEpisodes* was 0.137, so the marginal value of ln(*VoDPurchases*) for an additional free episode was 0.137. If a household watched one more free episode, its *incidence rate ratio* would increase by a factor of 1.147. This suggests the household would have purchased the full VoD series drama 14.7% more of the time. Likewise, the marginal impacts on value of ln(*VoDPurchases*) for an additional channel and an add-on channel in a household increased by 0.026 and 0.075, respectively, as shown in Table 8.

7.4.2. *Negative binomial and ZINB model estimation*

Next, the ZINB model was used to account for the excess 0 s that come from the different processes. The probability of 0 s is modeled independently. The coefficients from the negative binomial regression matched some of the effects that were expected to be present, and were similar to those of the Poisson regression. The binomial regression's parameters were significant and larger though.

In the count part of the ZINB regression, the impact of *FreeEpisodes* on VoD purchases was positive and significant. The expected change in log (*VoDPurchases*) for one additional free episode was 0.194. So an additional free episode was associated with a 21% increase in VoD series purchases. This impact of a free VoD episode was stronger compared to the results from the Poisson and negative binomial models. Yet the coefficients for *SubscribedChannels* and *AddOnChannels* were not significant. The logit part of the regression models the excess 0s independently. The log odds of excess 0s decreased by 0.277 for every add-on channel that a household had. Table 8 again shows the results for these two models.

7.4.3. *Causality*

To strengthen the claim about the causal relationship between free-episode sampling and the likelihood of a household to make a purchase, the next analysis involves observations about the patterns of household VoD episode sampling, conditional on at least one purchase occurring in the household. For this analysis, 3,652 observations were used for 823 households that purchased 1,130 VoDs in the period. Fig. 12 shows the average number of free-episode samples and VoD purchases by weekday and the portion stimulated by free-episode sampling.

Similar sampling and purchasing patterns, supporting the positive relationship between *FreeEpisodes* sampling and *VoDPurchases*, were observed. The sampling peaked on Friday, followed by a peak in purchasing on Saturday. Viewers showed less interest in VoDs after the weekend, when sampling and purchasing decreased. The patterns tell how households actually sampled and purchased VoDs.

The high amount of free-episode sampling without evidence of a concomitant amount of VoD purchases raises a question for service providers. How should they improve the *sampling-to-sales conversion rate* to support their VoD programming sales, in the presence of so much household interest? The *conversion rate* is the ratio of the number of VoD purchases divided by the number of free episodes that were sampled. A purchase stimulated by a free episode occurs when a household purchases a VoD that it previously sampled. Consistent with the finding of the positive impact of free-episode sampling on *VoDPurchases*, it also was learned that 50%+ of purchases were associated with episode sampling.

Some households made their purchases without sampling. They may have been influenced by outside information, as in other

[41] The associated probability density function (Lawal, 2012) is:

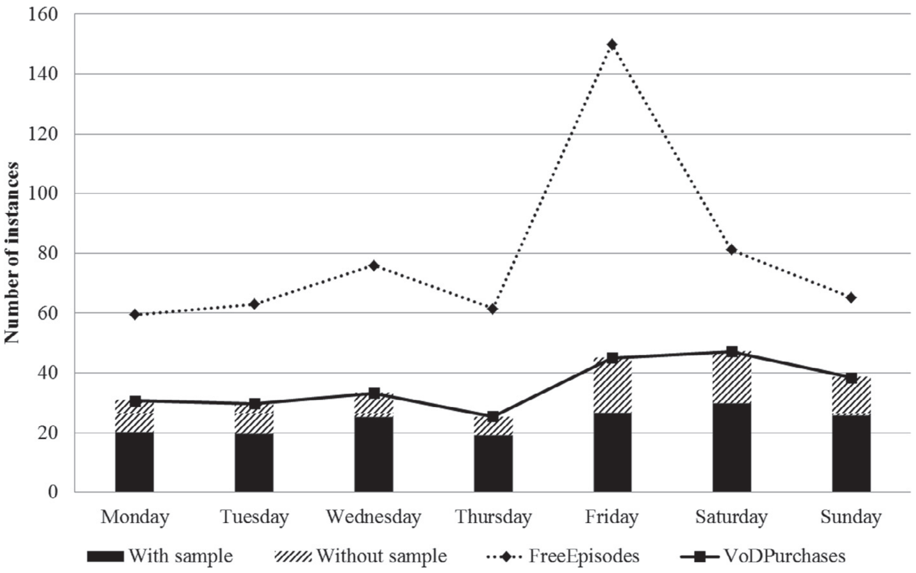
$$\Pr(Y_i = y_i) = \begin{cases} \Phi + (1-\Phi)(1+k\mu_i)^{-k^{-1}}, & \textit{if } y_i = 0 \\ (1-\Phi)\frac{\Gamma(y_i+k^{-1})}{y_i!\,\Gamma(k^{-1})}\frac{(k\mu_i)^{y_i}}{(1+k\mu_i)^{y_i+k^{-1}}}, & \textit{if } y_i > 0 \end{cases}$$

with E(y) = μ_i (1 – ϕ); and Var(Y_i) = μ_i (1 – ϕ) (1 + $k\mu_i$ + μ_i), where μ_i and ϕ depend on the covariates, and $k \geq 0$ is scalar. When ϕ or k is greater than 0, *over-dispersion* occurs. When ϕ = 1, the equation reduces to a negative binomial model, and when k = 0, it becomes a zero-inflated Poisson model.

Table 8
Comparison of the results of the three econometric models.

Dependent variables	Poisson		Negative binomial		ZINB			
					Count		Logit	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Intercept	-3.23***	(0.094)	-3.39***	(0.018)	-2.28***	(0.298)	1.19**	(0.437)
FreeEpisodes	0.14***	(0.008)	0.17***	(0.014)	0.19***	(0.021)	0.03	(0.030)
SubscribedChannels	0.03***	(0.007)	0.03***	(0.008)	0.01	(0.012)	-0.04	(0.024)
AddOnChannels	0.08***	(0.012)	0.08***	(0.015)	0.02	(0.022)	-0.28***	(0.075)
Ln(θ)					-0.35	(0.260)		

Notes: Obs.: 7,932. For ZINB, the two columns reflect the parts of the estimation: count data and logit parts. Signif.: $^{***}p < 0.01$, ** $p < 0.05$, * $p < 0.10$; $\theta = 0.703$.



Note. *Instances* refers to the de-duplicated set-top box sampling or purchasing records of households.

Fig. 12. Samples of *FreeEpisodes* and *VoDPurchases* by Day of the Week.

digital entertainment research conducted in Portugal (de Matos et al., 2016). Establishing an understanding of the relative impacts of sampling strategies and outside sources of information enable entertainment providers to design and implement more effective marketing strategies.

The econometric models suggest that sampling of *FreeEpisodes* had a positive impact on the number of VoD series purchases. Households that subscribed to premium packages, in terms of the number of their *SubscribedChannels*, and the purchases they made of *AddOnChannels* also showed their interest in VoD offerings. These households apparently obtained a higher level of utility from TV viewing, and demonstrated a correspondingly higher willingness-to-pay for such programming. Also, *VoDPurchases* were not constrained by a household's subscription. After accounting for the majority of *no-purchase* decisions in the study sample, the evidence showed that the more *AddOnChannels* a household had, the more likely that it had time and money to view VoD series.

This study demonstrates the key elements of a fusion analytics-based research investigation for Computational Social Science. Its use of machine-based methods primarily focused on the construction of a dataset for the study of VoD sampling and purchases, from 17+ million digital data records acquired from a TV set-top box-supported data repository. The methods used also permitted the distillation of the large number of records so much more detailed explanatory empirical analysis using count data model choices that took into account the household decision processes at work for VoD sampling and purchases.

8. Concluding discussion

This article has presented research involving Computational Social Science (CSS) fusion analytics, in an effort to illustrate some new ways that interesting and insightful research is being done related to E-Commerce and other kinds of Social Science research. Fusion analytics encompass multiple methods from CS, Statistics, Econometrics, and other disciplines that study business, consumer and social insights issues to improve data analytics performance in the presence of big data and enhanced computing power. The research work discussed supports research design innovation and more effective inquiry, through approaches that come ever closer to discovering the empirical truths in data in ways that emphasize causality. This is in line with the paradigm shift that enables study of Social Science issues with unprecedented control and insights (Agarwal and Dhar, 2014; Chang et al., 2014; Kenett and Shmueli, 2017).

8.1. Assessment of the applications

The research presented in this article is a by-product of long-term efforts involving organizational sponsors in two cases: VoD sampling and purchasing, and recovery of household patterns of preference for cable TV viewing. In the two other cases, it was possible to achieve access to enormous amounts of publicly-available data for the phenomena of interest: mobile phone-based stock trading, and music track popularity sustainability. In each of the research settings, the researchers were careful to articulate research questions up front that required thought to be given to the purpose for which data collection was undertaken, and what kinds of details were required to set up an interesting, insightful and in-depth assessment to yield meaningful insights. Various machine-based techniques were applied to harvest observations in unique ways. The methods choices involved what to use to acquire, filter, structure, and classify the data to produce datasets to support unique research inquiries. Also considered were the preliminary outputs that were to be captured. They included identification and classification of: different kinds of social sentiment that are relevant to mobile phone-based stock trading; multiple genres of TV programming to create a foundation for defining how household viewers consume digital entertainment; patterns of music and artist growth, and sustainability of social popularity; and instances of household consumption of VoD series dramas for which free sampling occurred with VoD series episodes in advance of the purchase. A generalized description of the fusion analytics research process, and key characteristics associated with it for different research settings are provided.

Two of the research projects that benefited from funding were interesting in terms of how their fusion analytics research processes were carried out, and the quality of insights they yielded for research and practice. In the case of household-level viewing of TV programs, surprising policy insights were created related to the extent to which different households watched more concentrated sets of programming around the TV show genres they targeted, and the potential usefulness that such insights support for the design of household-level *genre-focused*, *customized programming services*. For the same organization, the in-depth analysis of drama series sampling and VoD purchases yielded useful information on the value of series drama episodes. In follow-up work on that business process, additional insights have been obtained for instances where households first consume a free-episode sample, but may sample another paid episode or two later, suggesting that different pathways for sampling are appropriate before a final purchase. And further, the length of a VoD series drama may also influence household purchase behavior. In each of the research settings that were explored, useful insights for business strategy, service design, and a deeper understanding of stock-trading, digital entertainment-seeking, and music-focused consumers emerged that all go beyond the norm in traditional research inquiries that involve machine-based methods alone or explanatory econometric methods separately without their more powerful combination.

8.2. Translating fusion analytics findings into policy guidance and practitioner action

An important capability in this kind of research inquiry is that it should be possible for an analyst to translate fusion analytics methods and findings for practice in an actionable way to achieve business, consumer or social value. Academic researchers, in contrast, often require a theoretical perspective to create a strong and scientifically acceptable basis for in-depth explanation – often based on the use of theory, a key aspect of Social Science – in comparison to describing some phenomenon in terms of what the data tell an analyst that the person does not already know. This has often (but not always) been the case with Computer Science research, which is much more methods and algorithm-focused.

Balancing these contrasting perspectives on research inquiry, the 2011 Nobel Laureate in Economics, Christopher Sims (1980), described an early strategy for empirical research in Macroeconomics. He suggested to not create too strong a basis with explanatory theory for empirical models used to understand the complex relationships in macroeconomic developments and the policies that respond to how they arise. The view he suggested was to "let the data speak." His point, based on our reading of his work nearly 40 years later, is that too much effort directed to formulating theory-based models at the outset of an empirical inquiry may diminish the capacity of an analyst to acquire a richer and more nuanced understanding of a phenomenon under consideration.[42] This is pragmatic advice for academic researchers who wish to be effective in collaboration with industry practitioners, who may dismiss theory as being too superficial to be applicable.

This perspective is useful. Why? When big data analytics are implemented in corporate and organizational environments, they lead to data-supported actions – De Marchi et al.'s (2016) call for *evidence-based policy-making* – and possibly theory-motivated decisions too, if a theoretical perspective has been adopted. This enables policy changes that are *normative*, and based on what can be suggested as an appropriate course of action – even if it is different than what has come before.

The *policy analytics process* should not just be a patterned sequence of steps that have arisen in practice without organizational, managerial and strategic justification, as suggested by past-president of the American Management Association and organizational theorist, Andrew Van de Ven (1992). Moreover, another Nobel Prize-winning economist, Herbert Simon, has reminded us that "*human-beings, viewed as behaving systems are quite simple. The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves*" (Simon, 1996, p. 53). Indeed, the complexity of individuals in public and private organizations is multiplied several times by the nature of the complex systems (Weaver, 1948) that are present within them. Further complexity arises because organizations exist in proximity, competition and cooperation with one another, and within cities, economies, and regions that also contribute to it. And when the influence of technologies, processes, regulations and laws present are considered, the result is what Kelly (1995) has referred to as the "*new biology of machines, social systems, and the economic world*" – difficult to truly understand and interpret.

The return on business, consumer and social insights associated with data analytics is the

> *"value of information for the actions it prompts in the presence of new information minus the value of a decision, policy or action made in the absence of it – adjusted for the data acquisition analytics and processing costs"*

(Kauffman, 2014a). This value of information perspective is true, for example, for the analysis of household patterns in cable TV viewing. The findings that were obtained in the research suggest it may be possible for digital entertainment providers to capture the digital traces of household TV program viewing – even over a relatively brief period of time (Sisario, 2016; Maheshwari, 2017). This kind of hyperdifferentiated approach to delivering services though *resonance marketing* (Clemons et al., 2005; Granados et al., 2011) supports a streaming-data and knowledge-based service design, to maximize a household customer's utility for the set of programs in the digital entertainment packages that they purchase. The same is true for the potential use of the findings related to the vagaries of trader behavior in the mobile phone-based stock trading channel.

[42] A classic example is found in a book on competing economic, behavioral, institutional, and contract-based theories of product price rigidity and flexibility by Blinder et al. (1998). A more recent article on the price dynamics in Internet-based product and service retailing in e-commerce by Kauffman and Lee (2010) updated this view for the digital economy, stressing the richness associated with multi-level theoretical perspectives.

This channel has not been regulated heretofore by any securities exchange authority in any country to the authors' knowledge, specifically related to social media information diffusion – though there have been pre-regulations comments and agency notes. Inappropriately shared social sentiment and stock rumors in the financial markets – such as the *pump-and-dump strategies* of fraudsters, and inappropriate foreign exchange (FX)-related social sentiment – have been widely recognized as a source of deception and quality-of-market problems (U.S. Securities and Exchange Commission, 2013; Roberts, 2016).

8.3. Issues that still need to be addressed to make fusion analytics more effective

8.3.1. Fundamental and applied science involving organizational sponsors

In sponsored research of the sort that has been discussed, it is important to keep in mind that *fundamental science* and *applied science* go together. Effort has been devoted to making innovations with theory-based thinking, and pioneering new ways to work with the data. They have come in response to requests for assistance in business and other kinds of organizations, including collaborative work to produce solutions to targeted problems. Taken together, the organizational and social settings that were studied have provided opportunities to work on solving "hard" problems – even though there was considerable immediacy for the delivery of just-in-time research results. They also have supported work on underlying problems that require advances in Social Science research design and theory, and Computer Science data analytics advances, including streaming data and datasets that are never quite complete, but that grow and change over time.

Although the undertaking and management of such organizational relationships and the technical work involved with the construction of clean and usable large datasets are a heavy load, the by-products of a sustained effort with this kind of research are highly beneficial. They support the research team's acquisition of relatively deep domain knowledge and new expertise, great opportunities to interact with knowledgeable industry professionals, insights for securing additional research funding and attracting scientific help, and the growth of a reputation for advancing the depth of the research inquiry in specific scientific areas. This also yields opportunities to discuss the idea of *experimenting with experimentation* with industry partners – trying out different ways to construct experiments in the organizational contexts that are intended to create causality-focused research designs. The research experiences that have been discussed revealed some of the critical issues for achieving business and scientific ROI in industry and academic collaborations on the problems discussed in this article (Kauffman, 2014b).

8.3.2. Choosing research designs and analytics approaches for causality

Another issue to be considered is how to select an appropriate research design and analytics approach, when Machine Learning, Statistics and Econometrics are applied in different settings and contexts. A conceptual assessment suggests it is necessary to evaluate how a given fusion analytics-based research design will yield sufficient power to approximate answers to the research questions that are specified. This is more a matter of *accuracy* than *precision* though, since the overall goal is to reveal the empirical facts and findings that come close to the truth of a setting, and not necessarily to achieve repeatable measurements that are close to one another or achieve some level of decimal point precision. More interesting with recent methods and designs is how *knowledge about causality* is discovered, and how the results can be leveraged to create value.

8.3.3. Issues that arise along the way with causality-focused analytics

Obtaining evidence of causality is an interesting aspect of data analytics, but it never has been easy to accomplish – whether due to lack of data access, conditions that do not support causal tests, or a lack of awareness on the part of researchers as to what is possible in methodological terms. There is widespread awareness of the greater effectiveness of laboratory and field experiments designed to include treatments and controls with the randomization of subjects, as discussed for lab and social experiments in Economics (Friedman and Sunder, 1994), and in other Social Science fields (Rubin, 1978).

Many issues have been recognized that make it difficult to achieve causal explanations though. For example, there are problems with regression when it only produces associational but not causal explanations in different problems and settings (Rutter, 2007; Constantine, 2012). Also well recognized now are the limitations of new methods in Statistics in the presence of ineffective research designs with variables that have statistical confounds in their relationships involving the dependent variable and other independent variables (Freedman, 2010). There have been efforts over time to define the *language of causality* and reconcile the different perspectives of Computer Science and Social Science (Pearl, 2009a,b). Researchers have also sought to codify methods for *quasi-experimental and natural-experimental* approaches, where the setting does not permit assignment of treatment and control groups, as it does in fully-controlled lab-based experiments – a gold standard for discovering causality (Cook and Campbell, 1979; Meyer, 1995; Shadish et al., 2002).

In other research relevant for big data analytics in E-Commerce, one is reminded of how important it is to take advantage of methods that involve the construction of *matched samples* through the use of *propensity score matching* (PSM), as a basis for building causal tests for statistical inference (Rubin, 2006; Sekhon, 2009). And yet, recent research by King and Nielsen (2016) cautions us that PSM's main weakness arises through its mimicry of a *completely randomized experiment*, as opposed to a *fully-blocked experiment* (e.g., with matched pairs). They suggest that other matching approaches, for example, the *Mahalanobis distance metric* (MDM) (Mahalanobis, 1936) and *coarsened exact matching* (CEM) (Iacus et al., 2011) are worthwhile to consider in some empirical contexts, where the characteristics of a quasi-experiment experimental setting require different ways to establish causality.

A final example in the application of explanatory approaches for data analytics is the use of *difference-in-differences* (DiD) *models*. They are useful to analyze real-world settings that offer quasi-experimental structure to study a treatment and a control for a matched population of participants, defined for the period before and after a treatment has been applied (Angrist and Pischke, 2008). The classic studies are associated with Labor Economics. For example, Card and Krueger (1994) studied how changes in minimum wages affected bordering states in the U.S. and failed to match expectations from theory: higher wages should have led to lesser employment in the fast food industry in a state, but were not observed. In addition, Ashenfelter and Card (1985) explored publicly-funded training programs, and the differences in prior and post-participating earnings histories of people who did or did not participate. Bertrand et al. (2004) have pointed out that difficulties may arise in the application of DiD models, when the *interventions* in a setting are endogenous, how different groups that are used are similar to control groups that might be used if the research were fully-experimental, and how serial correlations in time-wise data may result in inconsistent estimation model standard errors. So what is viewed as a strong empirical modeling choice for explanation in data analytics also may have problems that diminish its ability to produce knowledge that extracts the truth that actually exists in a setting.

8.3.4. The perceived value of data analytics

Choices about method selection are made more complex due to organizational perceptions about what is worthwhile to study in depth. Why? Beyond the existing skill base of data analytics knowledge referred to earlier, the additional costs of carrying out the analytics work often fall short of the expected net value of having management commit to a new a substantial data analytics project. There may be different growth opportunities for the creation of value with new business ventures, products and services in the presence of different *organizational value disciplines* and *customer intimacy*, for example (Treacy and Wiersema, 1993). Also, different kinds of IT to aid in transforming a business process so it can become more *information-intensive* are expensive to bring into a public or private organization, and thus better able to drive sustainable advantage for profitability (Davern and Kauffman, 2000). So data analytics for strategy and operational insights need to be crafted in ways that recognize that there is no single benchmark and no "silver bullet" for defining value. Instead, it requires a value discipline that is able to change over time in an organization to match changes in its customers, competition, marketplace, and the economy.

Overall, the issues that have been discussed should come as no surprise, at least from the point of view of the persistence of the importance of how IT investments and digital economy activities create many different kinds of value wherever strategies are formulated and new technologies are used (Strassmann et al., 1988; Kauffman and Walden, 2001). In business and international trade, medicine and science, education and government, and other domains, the problem of the value of IT has never gone away. This is made plain by a recent forecast for IT spending worldwide to reach USD 2.7 trillion, with industries such as healthcare, manufacturing and financial services leading the way (IDC, 2016). With such a large quantum of spending, assessing the marginal impacts of technology-related investments has achieved an even more intense level of importance today. And the increasingly data-intensive areas – social media, social commerce, and social network analytics (Zhou et al., 2013); the Internet of things, sensor networks, and *sociophysical analytics* (Misra et al., 2014); and the new approaches associated with the fintech revolution (Dietz et al., 2016), and cognitive systems (Davenport and Krishna, 2017) – are creating ever-stronger forces for making technology yield more value than before. Moreover, Brynjolfsson and McAfee (2011) and Brynjolfsson (2013) have reminded us that the new impetus in many settings – whether with autonomous vehicles, sensor analytics, and other emerging technologies – is to "race with the machines," and take advantage of their capabilities in conjunction with how people need to work to create value.

8.4. Final thoughts

In closing, academic researchers and their industry partners must figure out some new ways to make fusion analytics work in practice. Data schema and definitions, reference data standards, federated data models, and big-data sourcing-hubs and sharing programs within and across organizations, industries and government agencies, all need to be given more attention. In addition, a major area for potential innovation involves the creation of new models for industry and public sector data-sharing with universities and research organizations. Although heretofore, the primary model was for research organizations to acquire data directly from their organizational sponsors, the corporate, legal and regulatory environment associated with this *inside-out data-sharing model* has become increasing difficult in many countries in recent years. This has been true in Singapore for researchers at LARC, for example, where some of the research was conducted. The Personal Data Protection Act (PDPA) of 2012 established:

> "*a data protection law that comprises various rules governing the collection, use, disclosure and care of personal data. It recognises both the rights of individuals to protect their personal data, including rights of access and correction, and the needs of organisations to collect, use or disclose personal data for legitimate and reasonable purposes*" (Personal Data Protection Commission Singapore, 2016).

This suggests the paramount importance of *data-sharing protocols and partnerships* that maintain strong information security, and manage the risk that organizations face in the event that data are not effectively managed, mishandled or hacked (Singapore Management University, 2017). Another alternative business model may be useful to update the manner in which such data analytics-focused industry-university collaboration occurs: with an *outside-in data-sharing model*. The essential idea of this approach is that no data should leave the organization that owns it. Instead, this kind of model may be viewed as a variation of what is done in scientific laboratories involving hazardous microbiological organisms and radioactive materials. A scientist can go into the lab environment, but is never permitted to actually "touch the hazardous material" – the data in this case – or to take away any part of it. Instead, there need to be contractual relationships in place that balance safe-handling of the data and the related results from the viewpoint of industry-side intellectual property, while affording the academic side reasonable ability to port out results that can be disguised and approved for sharing in scientific conferences, journal publications and books. This further opens up the need for methods research on safe-access shared-data analytics, secure porting algorithms, and software applications that allow the benefits to go beyond the costs. There is much more work to be done in this area, to support the realization of the Computational Social Science fusion analytics vision that has been shared.

Acknowledgments

This research was supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the Infocomm Development Authority. Readers should recognize that non-disclosure agreements prevent us from sharing the organization identities and names of key informants, as well as details of the data and the findings that our sponsors view as privileged information. A related version of this work (Kim et al., 2017) was presented at the 50th Anniversary Hawaii International Conference on Systems Science (HICSS) in January 2017, held in Waikoloa, Hawaii, in the Minitrack on 'Integrating Business Operations, IT and Consumer Behavior' in the Organizational Systems and Technologies Track. We are grateful to a number of individuals for their comments and interest in this research: Avi Seidmann, Jennifer Zhang, Yabing Jiang, Huaxia Rui, Rajiv Dewan, and several anonymous reviewers of the conference submission. We benefitted from the input at different times of other colleagues: Ee-Peng Lim, Steven Miller, Steve Fienberg, Hoong Chuin Lau, Archan Misra, Alexis Tsoukiàs, and Ramayya Krishnan, as well as Pulak Ghosh, Bing Tian Dai, Zhuolun Li, Gwangjae Jung, Peiran Zhang, Myung-Rae Chang, and YoungOk Kwon, when they were at SMU's Living Analytics Research Centre (LARC). We also appreciated suggestions from Kustini Lim-Wavde, Ryan Sougstad, Bin Wang, Jessica Chen, Ursula Higgins, Morad Benyoucef, Tuan Anh Hoang, Jiali Du, Zhiyuan Gao, David Phang, Zhaoxia Wang, Qiuhong Wang, and Chris Yang. The authors are solely responsible for any errors and omissions.

Appendix. A. Fusion analytics examples

Table A1 offers a summary of the research based on: (1) the purpose for which the data are intended to be analyzed; (2) the machine-based methodologies that have been applied; (3) the kinds of outputs that result from the application of the machine-based methods; (4) the econometric models and estimation methods that are used to delve deeper into the relevant data to create useful information; and (5) the policy insights that emerge from the sequence of steps in the research.

Table A1
Four studies that illustrate fusion analytics methods in this article.

Authors	Purpose	Machine methods	Machine output	Statistical and econometrics methods	Policy insights
Kim et al. (2016)	Gauge social sentiment for mobile phone stock trades	Text analytics for social sentiment identification; DaumSoft software package	Extracted indicators on social sentiment for stock investments	Feasible generalized least squares (FGLS); Granger causality, panel vector autoregression (PVAR); kernel regularized least squares (KRLS)	How sentiment drives trade herding; how mobile trader behavior is different for social sentiment stimuli
Chang et al. (2012)	Acquire household TV viewing preferences	*k*-means clustering; supervised/ unsupervised learning; cluster centroids; Davies-Bouldin indices; silhouette values	TV viewing cluster centroids obtained; households characterized by composite viewing preferences	Limited dependent variable model; quantile regression	How household viewing preferences vary by channel, program genre
Ren and Kauffman (2017)	Capture rich semantics for music tracks	Text mining; topic model; SVM, SVR; baggingRegressor; random forest	High-level music constructs; low-level music features	Duration model; hazard function assumptions; statistical patterns	How music track top-rank popularity in Last.fm works
Hoang and Kauffman (2016)	Identify VoD sampling, purchases	Data rectangle construction; dataset matching	Viewing records for sampling, purchases	Count data models; statistical patterns	How episode sampling affects VoD series purchases

References

Abrigo, M.R.M., Love, I., 2015. Estimation of panel vector autoregression in Stata: A package of programs. In: Proc. 2015 Intl. Panel Data Conf., Central Eur. Univ., Budapest, Hungary, June 29–30.

Agarwal, R., Dhar, V., 2014. Big data, data science, and analytics: the opportunity and challenge for IS research. Inf. Syst. Res. 25 (3), 443–448.

Aigner, W., Miksch, S., Muller, W., Schumann, H., Tominski, C., 2008. Visual methods for analyzing time-oriented data. IEEE Trans. Vis. Comput. Graphics 14 (1), 47–60.

Amemiya, T., 1985. Generalized least squares theory. Chap. 6 in Advances in Econometrics, Harvard, Boston, MA.

Anderson, C., 2008. The end of theory: The data deluge makes scientific method obsolete. Wired.

Angrist, J.D., Pischke, J.S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, New York, NY.

Aral, S., Walker, D., 2011. Identifying social influence in networks using randomized experiments. IEEE Intell. Syst. 26 (5), 91–96.

Ashenfelter, O., Card, D., 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. Rev. Econ. Stat. 67 (4), 648–660.

Athey, S., 2015. Machine learning and causal inference for policy evaluation. Proc. 21st ACM SIGKDD IntL. Conf. Knowl. Disc. Data Mgmt.. ACM Press, New York, NY, pp. 5–6.

Athey, S., Ellison, G., 2011. Position auctions with consumer search. Q. J. Econ. 126 (3), 1213–1270.

Bapna, R., Umyarov, A., 2015. Do your online friends make you pay? a randomized field experiment on peer influence in online social networks. Manage. Sci. 61 (8), 1902–1920.

Bardhan, I., Demirkan, D., Kannan, P.K., Kauffman, R.J., Sougstad, R., 2010. An interdisciplinary perspective on IT service management and service science. J. Manage. Inf. Syst. 26 (4), 13–64.

Basak, D., Pal, S., Patranabis, D.C., 2007. Support vector regression. Neural Inf. Process. 11 (10), 203–224.

Bell, G., Hey, T., Szalay, A., 2009. Beyond the data deluge. Science 423, 1297–1298.

Benaroch, M., Dai, Q., Kauffman, R.J., 2010. Should we go our own way? Analyzing backsourcing flexibility in IT service outsourcing contracts. J. Manage. Inf. Syst. 26 (4), 317–358.

Ben-David, S., Pál, D., Simon, H., 2007. Usability of k-Means Clustering. In: Bshouty, N., Gentile, C. (Eds.), Learning Theory. Springer, Berlin, pp. 20–34.

Bertrand, M., Dufflo, E., Mullainathan, S., 2004. How much should we trust in difference-in-differences estimates? Q. J. Econ. 119 (1), 249–275.

Bhattacharjee, S., Gopal, R.D., Lertwachara, K., Marsden, J.R., Telang, R., 2007. The effect of digital sharing technologies on music markets: a survival analysis of albums on ranking charts. Manage. Sci. 53 (9), 1359–1374.

Bishop, C., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Blinder, A.S., Canetti, E.R.D., Lebow, D.E., Rudd, J.E., 1998. Asking about Prices: A New Approach to Understanding Price Stickiness. Russell Sage Foundation, New York, NY.

Bockstedt, J.C., Kauffman, R.J., Riggins, F.J., 2005. The move to artist-led online music distribution: a theory-based assessment and prospects for structural changes in the digital music market. Int. J. Electron. Commun. 10 (3), 7–38.

Brabham, D., 2009. Crowdsourcing public participation process for planning projects. Plan. Theor. 8 (3), 242–262.

Breiman, L., 1994. Bagging predictors. Technical report 421. Stat. Dept., Univ. Calif., Berkeley, CA.

Brickland, T., 2001. Introduction to the Policy Process: Theory, Concepts and Methods for Policy Making. Sharpe, Armonk, NY.

Brynjolfsson, E., 2013. The key to growth: Race with the machines. TED Talk.

Brynjolfsson, E., McAfee, A., 2011. Race Against The Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy. Digital Frontier Press, Lexington, MA.

Bunn, D., 1977. Policy analytic implications for a theory of prediction and decision. Pol. Sci. 8 (2), 125–134.

Cameron, A.C., Trivedi, P.K., 1990. Regression-based tests for overdispersion in Poisson model. J. Econometrics 46 (3), 347–364.

Canova, F., Ciccarelli, M., 2013. Panel vector autoregressive models: a survey. In: Fomby, T., Killian, L., Murphy, A. (Eds.), VAR Models in Macroeconomics. Emerald, Bingley, UK, pp. 205–246.

Card, D., Krueger, A.B., 1994. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. Am. Econ. Rev. 84 (4), 772–793.

Carley, K.M., 2002. Computational organization science: a new frontier. Proc. Natl. Acad. Sci. 99 (3), 7257–7262.

Cha, M., Mislove, A., Gummadi, K.P., 2009. A measurement-driven analysis of information propagation in the Flickr social network. Proc. 18th Intl. Conf. on Worldwide Web. ACM Press, New York, NY, pp. 721–730.

Chang, R.M., Kauffman, R.J., Son, I., 2012. Consumer micro-behavior and TV viewership patterns: data analytics for the two-way set-top box. In: Bichler, M., Kauffman, R.J., Lau, H.C., Yang, Y.P. (Eds.), Proc. 14th Intl. Conf. Elec. Comm.. ACM Press, New York, NY, pp. 272–273.

Chang, R.M., Kauffman, R.J., Kwon, Y.O., 2014. Understanding the paradigm shift to computational social science in the presence of big data. Decis. Support Syst. 63, 67–80.

Chaudhuri, S., Dayal, U., Narasayya, V., 2011. An overview of business intelligence technology. Commun. ACM 54 (8), 88–98.

Chellappa, R.K., Sin, R.G., Siddarth, S., 2011. Price formats as a source of price dispersion: a study of online and offline prices in the domestic U.S. airline markets. Inf. Syst. Res. 22 (1), 83–98.

Chen, H., Chiang, R., Storey, V.C., 2012. Business intelligence and analytics: from big data to impact. MIS Q. 36 (4), 1165–1188.

Cheng, Z., Shen, J., 2016. On effective location-aware music recommendation. ACM Trans. Inf. Syst. 34 (2), 13.

Clemons, E.K., Spitler, R., Gu, B., Markopoulos, P., 2005. Information, hyperdifferentiation, and delight: the value of being different. In: Bradley, S., Austin, R. (Eds.), The Broadband Explosion: Leading Thinkers on the Promise of a Truly Interactive World. Harvard Business School Press, pp. 137–164.

Cloud Standards Customer Council, 2014. Deploying big data analytics applications in the cloud: Roadmap for success. Needham, MA, May.

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C., 2009. MAD skills: New analysis practices for big data. In 2009 Proc. Very Large Data Bases Endowment, Lyon, France, 2(2), pp. 1481–1492.

Constantine, N.A., 2012. Regression analysis and causal inference: cause for concern? Perspect. Sex. Reprod. Health 44 (2), 134–137.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin, Boston, MA.

Cragg, J.G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39 (5), 829–844.

Creeber, G., Miller, T., Tulloch, J. (Eds.), 2001. Television Genre Book. British Film Inst, London, UK.

Das, S.R., Chen, M.Y., 2007. Yahoo! for Amazon: sentiment extraction from small talk on the Web. Manage. Sci. 53 (9), 1375–1388.

Davenport, T.H., 2006. Competing on analytics. Harv. Bus. Rev. 84 (1), 98–107.

Davenport, T., Krishna, D., 2017. The changing world of technology in financial services. Deloitte University Press, New York, NY, January, p. 27.

Davern, M., Kauffman, R.J., 2000. Discovering potential and realizing value from information technology investments. J. Manage. Inf. Syst. 16 (4), 121–144.

Davies, P.T., 1999. What is evidence-based education? Br. J. Educ. Stud. 47 (2), 108–121.

De Long, J.B., Shleifer, A., Summers, L., Waldmann, R., 1990. Noise traders' risk in financial markets. J. Pol. Econ. 98 (4), 703–738.

De Marchi, G., Lucertini, G., Tsoukiàs, A., 2016. From evidence based policy making to policy analytics. Ann. Oper. Res. 236, 15–38.

de Matos, M.G., Ferreira, P., Smith, M.D., Telang, R., 2016. Culling the herd: using real world randomized experiments to measure social bias with known costly goods. Manage. Sci. 62 (9), 2563–2580.

Del Guidice, M., Della Peruta, M., Carayannis, E., 2015. Social Media and Emerging Economies: Technology, Cultural and Economic Implications. Springer, Berlin, Germany.

Delen, D., Demirkan, H., 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in the cloud. Decis. Support Syst. 55 (1), 412–471.

Demirkan, H., Kauffman, R.J., Vayghan, J., Fill, H.G., Karagiannis, D., Maglio, P., 2008. Service-oriented technology and management: perspectives on research and practice for the coming decade. Electron. Commun. Res. Appl. 7 (4), 356–376.

Dennis, D.M., Gray, D.M., 2013. An episode-by-episode examination: what drives television-viewer behavior – digging down into audience satisfaction with television dramas. J. Ad. Res. 53 (2), 166–174.

Dhanaraj, R., Logan, B., 2005. Automatic prediction of hit songs. In: Proc. 2005 Intl. Soc. Music Info, Retr., London, UK, September 11–15, pp. 488–491.

Dietterich, T.G., 2003. Machine learning. In: Nadel, L. (Ed.), Encyclopedia of Cognitive Science. Macmillan, London, UK.

Dietz, M., Khanna, S., Olanrewaju, T., Rajgopal, K., 2016. Cutting through the noise around financial technology. Financial Services, McKinsey and Co., New York, NY, February.

Elberse, A., 2010. Bye-bye bundles: the unbundling of music in digital channels. J. Mark. 74 (3), 107–123.

Ernst and Young, 2014. Big data: changing the way businesses compete and operate. Insights on Governance, Risk and Compliance, London, UK.

Fagella, D., 2016. Where healthcare's big data actually come from. TechEmergence, San Francisco, CA, p. 8.

Farris, P.W., 2010. Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. Pearson, Upper Saddle River, NJ.

Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modeling rates and proportions. J. Appl. Stat. 31 (7), 799–815.

Freedman, D.A., 2010. Statistical Models and Causal Inference: A Dialogue with the Social Sciences. Cambridge Univ. Press, Cambridge, UK.

Friedman, D., Sunder, S., 1994. Experimental Methods: A Primer for Social Scientists. Cambridge Univ. Press, Cambridge, UK.

Geng, D., Kauffman, R.J., 2017. Decomposing the impact of credit card promotions on customer behavior and merchant performance. In: Bui, T., Sprague, R. (Eds.), Proc. 50th Hawaii Intl. Conf. Sys. Sci. IEEE Comp. Soc. Press, Washington, DC.

Ghose, A., Ipeirotis, P.G., Li, B., 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. Mark. Sci. 31 (3), 493–520.

Gondecha, P., Lieu, H., 2012. Mining social media: a brief introduction. In: Mirchandani, P.B., Smith, J.C., Greenberg, H.J. (Eds.), Tutorials in Operations Research: New Directions in Informatics, Optimization, Logistics, and Production. Institute for Mgmt. Sci., U. Maryland, Baltimore County, Catonsville, MD.

Grace, J., Gruhl, D., Haas, K., Nagarajan, M., Robson, C., Sahoo, N., 2008. Artist ranking through analysis of on-line community comments. Proc. 17th ACM Intl. World Wide Web Conf.. ACM Press, New York, NY.

Granados, N.F., Kauffman, R.J., Lai, H., Lin, H., 2011. Decommoditization, resonance marketing, and information technology: an empirical study of air travel services amid channel conflict. J. Manage. Inf. Syst. 28 (2), 39–74.

Granados, N.F., Gupta, A., Kauffman, R.J., 2012. Online and offline demand and price elasticities: evidence from the air travel industry. Inf. Syst. Res. 23 (1), 164–181.

Greene, W., 2007. Functional form and heterogeneity in models for count data. Found. Trends Econometrics 1 (2), 113–128.

Gurmu, S., Trivedi, P.K., 1996. Excess zeroes in count models for recreational trips. J. Bus. Econ. Stat. 14 (4), 469–477.

Haaijer, R., Wedel, M., Vriens, M., Wansbeek, T., 1998. Utility covariances and context effects in conjoint MNP models. Mark. Sci. 17 (3), 236–252.

Hainmueller, J., Hazlett, C., 2013. Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. Pol. Anal. 22 (2), 143–168.

Hamlen Jr., W.A., 1991. Superstardom in popular music: empirical evidence. Rev. Econ. Stat. 73 (4), 729–733.

Harrysson, M., Metayer, E., Sarrazin, H., 2014. The Strength of Weak Signals. McKinsey Qtrly..

Hassan, N., 2009. Using social network analysis to measure IT-enabled business performance. Inf. Technol. Manage. 26 (1), 61–76.

Hayat, E.A., Suner, A., Burak, U., Dursun, Ö., Orman, M.N., Kitapçioğlu, G., 2010. Comparison of five survival models: breast cancer registry data from Ege University Cancer Research Center. Turkiye Klinikleri J. Med. Sci. 30 (5), 1665–1674.

Hey, T., Tansley, S., Tolle, K. (Eds.), 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, WA.

Hoang, A.P., Kauffman, R.J., 2016. Experience me! The impact of content sampling strategies on the marketing of digital entertainment goods. In: Bui, T., Sprague, R. (Eds.), 2016 Hawaii Intl. Conf. on Sys. Sci.. Comp. Soc. Press, Washington, DC.

Hu, M.C., Pavlicova, M., Nunes, E.V., 2012. Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. Am. J. Drug Alcohol. Abuse 37 (5), 367–375.

Hu, H., Wen, Y., Chua, T.S., Li, X., 2014. Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2, 652–687.

Hussain, M., Khan, N., Uddin, M., 2014. Log-normal duration model is the best fitted model for duration from chest pain to coronary artery disease diagnosis: an outcome of retrospective cross sectional study. Pakistan J. Stat. Oper. Res. 10 (4), 369–379.

Iacus, S.M., King, G., Porro, G., 2011. Multivariate matching methods that are monotonic imbalance bounding. J. Am. Stat. Assoc. 106 (493), 345–361.

IBM, 2012. Analytics: The real-world use of big data. Institute for Business Value, New York, NY.

IBM, 2017a. The four V's of big data. Infographics and animations, Big Data & Anal. Hub, Armonk, NY.

IBM, 2017b. Solve real problems: Build with infrastructure, Watson, software, and services on the BlueMix cloud platform. Armonk, NY.

IBM, 2017c. Education. Big Data & Analytics Hub, Armonk, NYIBM, 2017c. Education. Big Data & Analytics Hub, Armonk, NY.

IDC, 2009. Digital data to double every 18 months. Framingham, MA, May 18.

IDC, 2016. Worldwide IT spending forecast to reach $2.7 Trillion in 2020 led by financial services, manufacturing, and healthcare, according to IDC. Framingham, MA, August 29.

IFPI, 2015. Digital music report: Recording industry in numbers. Intl. Fed. Phono. Ind, London, UK.

Imbens, G., Barrios, T., Diamond, R., Kolesar, M., 2011. Clustering, spatial correlations and randomization inference. Mimeo, Harvard University, Boston, MA.

Investopedia, 2017. Econometrics. Available at: www.investopedia.com/terms/e/econometrics.asp.

Jain, D., Mahajan, V., Muller, E., 1995. An approach for determining optimal product sampling for the diffusion of a new product. J. Prof. Innovation Manage. 12 (2), 124–135.

James, G., Witten, D., Hastie, T., Tubshirani, R., 2013. An Introduction to Statistical Learning: With Applications to R. G. In: Fienberg, Casella S., Olkin, I. (Eds.). Springer, Berlin/Heidelberg, Germany.

Kauffman, R.J., 2014a. Digital canaries in an urban data mine. Asian Manage. Insights 1 (1), 50–57.

Kauffman, R.J, 2014b. Achieving scientific and business returns on investments in fundamental and applied social science research. Seminar presentation, Agency for Science, Technology and Research (A∗STAR), Singapore, May.

Kauffman, R.J., Lee, D.W., 2010. A multi-level theory approach to understanding price rigidity in Internet retailing. J. Assoc. Inf. Syst. 11 (6), 303–338.

Kauffman, R.J., Walden, E.A., 2001. Economics and electronic commerce: survey and directions for research. Int. J. Electron. Commun. 5 (4), 4–115.

Kauffman, R.J., Wood, C.A., 2007. Revolutionary research strategies for e-business: a philosophy of science view in the age of the Internet. In: Kauffman, R.J., Tallon, P.A. (Eds.), Economics, Information Systems and Electronic Commerce: Empirical Advances, Advances in Management Information Systems Series. M. E. Sharpe, Armonk, NY.

Kauffman, R.J., March, S.T., Wood, C.A., 2000. Design principles for long-lived Internet agents. Intl J. Intell. Syst. Acc. Finance Manage. 9 (4), 217–236.

Kauffman, R., Techatassanasoontorn, A., Wang, B., 2012. Event history, spatial analysis and count data methods for empirical research in information systems. Inf. Technol. Manage. 13 (3), 115–147.

Kauffman, R.J., Kim, K., Lee, S.Y., 2017. Computational social science fusion analytics: Combining machine-based methods with explanatory empiricism. In: Sprague, R., Bui, T. (Eds.), Proc. 50th Hawaii Intl. Conf. Sys. Sci., Hawaii, HI. IEEE Comp. Soc. Press, Washington, DC.

Kecman, V., Huang, T.M., Vogt, M., 2005. Iterative single data algorithm for training kernel machines from huge data sets: Theory and performance. In: Wang, L. (Ed.), Support Vector Machines: Theory and Applications. Springer, Berlin, Germany, pp. 255–274.

Kelly, K., 1995. Out of Control: The New Biology of Machines, Social Systems and the Economic World. Perseus Books, New York, NY.

Kenett, R., Shmueli, G., 2017. Information Quality: The Potential of Data and Analytics to Generate Knowledge. Wiley, New York, NY.

Kieschnick, R., McCullough, B.D., 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Stat. Model. 3 (3), 193–213.

Kim, K. Lee, S.Y., Kauffman, R.J., 2016. How do traders react to social media sentiment in the mobile channel? In: Intl. Symp. Smart Fin., Shenzhen, China, May.

King, G., Nielsen, R., 2016. Why propensity scores should not be used for matching. Working paper, Institute for Quantitative Social Science, Harvard University, December 2016.

Kleinbaum, D.G., Klein, M., 2006. Survival Analysis: A Self-Learning Text. Springer, Berlin, Germany.

Kmenta, J., 1986. Generalized linear regression model and its applications. Elements of Econometrics,. MacMillan, New York, NY.

Lafayette, J., 2014. Threat becomes profit center as TV leverages technology. BroadcastingCable.com, 144, 16.

Larkey, P.D., 2015. Evaluating Public Programs: The Impact of General Revenue Sharing on Municipal Government. Princeton Univ. Press, Princeton, NJ.

Lawal, B.H., 2012. Zero-inflated count regression models with applications to some examples. Qual. Quant. 46 (1), 19–38.

Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., van Alstyne, M., 2009. Life in the network: the coming age of computational social science. Science 323 (5915), 721–723.

Lee, J., Lee, J.S., 2015. Predicting music popularity patterns based on musical complexity and early stage popularity. Proc. 3rd Workshop on Speech, Lang. Aud. in Multimed.. ACM Press, New York, NY.

Lee, G., Qiu, L., Whinston, A., 2016. A friend like me: Modeling network formation in a location-based social network. SSRN Paper 2769696.

Lewis, D., 2001. Causation as influence. J. Philos. 97 (4), 182–197.

Li, X., Wang, H., Gu, B., Ling, X., 2015. Data sparseness in linear SVM. Proc. 24th Intl. Conf. Artif. Intell., Buenos Aires, Argentina. ACM Press, New York, NY.

Li, Z., Kauffman, R.J., Dai, B., 2017. Can I see beyond what you can see? Blending machine learning and econometrics to discover household TV viewing preferences. In: Bui, T., Sprague, R. (Eds.), Proc. 50th Hawaii Intl. Conf. Sys. Sci.. IEEE Comp. Soc. Press, Washington, DC.

Lopez, M., 2016. IBM speaks on why cognitive is a business imperative. Forbes, November 15.

Lymperopoulos, I., Lekakos, G., 2013. Analysis of social network dynamics with models from the theory of complex adaptive systems. In: Douligeris, C., Polemi, N., Karantjias, A., Lamersdorf, W. (Eds.), Collaborative, Trusted and Privacy-Aware e/m-Services, IFIP Adv. Info. Comm. Tech, vol. 399. Springer, Berlin/Heidelberg, Germany.

Ma, D., Kauffman, R.J., 2014. Competition between software-as-a-service vendors. IEEE Trans. Eng. Manage. 61 (4), 717–729.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2 (1), 49–55.

Maheshwari, S., 2017. For marketers, TV Sets are an Invaluable Pair of Eyes. New York Times.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey and Co., New York, NY, May.

Marsh, L.C., Cormier, D.R., 2001. Spline Regression Models. Sage, Thousand Oaks, CA.

Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin, New York, NY.

McKinsey, 2011. Creating value through credit card partnerships in Latin America. New York, NY.

Meyer, B., 1995. Natural and quasi-natural experiments in economics. J. Bus. Econ. Stats. 13 (2), 151–162.

Misra, A., Jayarajah, K., Nayak, S., Prasetyo, P., Lim, E.P., 2014. Socio-Physical Analytics: Challenges and Opportunities. Proc. 2014 Workshop on Physical Analytics, Bretton Woods, NH. ACM, New York, NY.

Mohr, L.C., 1995. Impact Analysis for Program Evaluation. Sage, Thousand Oaks, CA.

Mor, Y., 2014. Big data and law enforcement: Was 'Minority Report' right? Wired.

Moretti, E., 2011. Social learning and peer effects in consumption: evidence from movie sales. Rev. Econ. Stud. 78 (1), 356–393.

Nemschoff, M., 2014. Why the transportation industry is getting on board with big data and Hadoop. MapR, San Jose, CA, p. 28.

Newswire, P.R., 2013. Video on demand market worth $45.25 billion by 2018. December 6.

Ni, Y., Santos-Rodriguez, R., McVicar, M., De Bie, T., 2011. Hit song science once again a science? Proc. 4th Intl. Workshop on Mach. Learn. and Music, Sierra Nevada, Spain, December 17.

Pearl, J., 2009a. Understanding propensity scores. Causality: Models, Reasoning, and Inference,. Cambridge University Press, New York, NY.

Pearl, J., 2009b. Causal inference in statistics: an overview. Stat. Surv. 3, 96–146.

Personal Data Protection Commission Singapore, 2016. Personal Data Protection Act of 2012: Overview. February.

Ray, S., Turi, R.H., 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. Proc. 4th Intl. ConfAdv. in Pattern Recog. Dig. Tech., 137–143

Ren, J., Kauffman, R.J., 2017. Understanding music track popularity in social networks: A fusion analytics approach. Working paper, School of Information Systems, Singapore Management University, 2017.

Roberts, J.J., 2016. Meet the latest scary form of social media fraud. Fortune, November 11.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comp. Appl. Math. 20, 53–65.

Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Ann. Stat. 6 (1), 34–58.

Rubin, D.B., 2006. Matched Sampling for Causal Effects. Cambridge University Press, New York, NY.

Rutter, M., 2007. Proceeding from observed correlation to causal inference: the use of natural experiments. Perspect. Psychol. Sci. 2 (4), 377–395.

SAS, 2017. Machine Learning: What It is and Why It Matters. Cary, NC.

Science and Knowledge Service, 2016. Counterfactual impact evaluation. Joint Science Research Centre, European Commission, p. 7.

SciKit Learn, 2017. Sklearn.ensemble.BaggingRegressor. Scikit-learn.org.

Sekhon, J., 2009. Opiates for the matches: matching methods for causal inference. Ann. Rev. Econ. Pol. 12, 487–508.

Sen, A.G., 2003. Consumer insights and creativity. Indian Inst. Manage. Bangalore Manage. Rev. 15 (3), 124–126.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA.

Shi, Z., Lee, G.M., Whinston, A.B., 2016. Towards a better measure of business proximity: topic modeling for industry intelligence. MIS Q. 40 (4), 1035–1056.

Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25 (3), 289–310.

Simon, H.A., 1996. The sciences of the artificial. MIT Press, Cambridge, MA.

Sims, C., 1980. Macroeconomics and reality. Econometrica 48 (1), 1–48.

Singapore Management University. SMU's Professor Robert Deng conferred AXA Chaired Professorship of Cybersecurity. 2017. Press release on presentation and discussion roundtable, Singapore, April 17.

Sisario, B., 2016. Nielsen Acquires Gracenote, Highlighting the Value of Data. New York Times.

Strassmann, P., Berger, P., Swanson, E.B., Kriebel, C.H., Kauffman, R.J., 1988. Measuring Business Value of Information Technologies. ICIT Press, Washington, DC.

Strobl, E.A., Tucker, C., 2000. The dynamics of chart success in the U.K. pre-recorded popular music industry. J. Cult. Econ. 24 (2), 113–134.

Stroud, M., 2014. The minority report: Chicago's new police computer predicts crimes, but is it racist? Verge.

Tan, T.N., Michael, S., Kumar, V., 2005. Introduction to Data Mining. Addison-Wesley Longman, Boston, MA.

Techatassanasoontorn, A.A., Kauffman, R.J., 2014. Examining the growth of digital wireless phone technology: a take-off theory analysis. Decis. Support Syst. 58, 53–57.

Tirunallai, S., Tellis, G., 2014. Mining marketing meaning from online chatter: strategic brand analysis of big data using latent Dirichlet allocation. J. Mark. Res. 51, 463–479.

Treacy, M., Wiersema, F., 1993. Customer intimacy and other value disciplines. Harv. Bus. Rev. 71 (1), 84–93.

U.S. Securities and Exchange Commission, 2013. Updated investor alert: Social media and investing. Washington, DC, November 5.

Van de Ven, A.H., 1992. Suggestions for studying strategy process: a research note. Strateg. Manage. J. 13 (SI), 169–188.

Wang, Y., Lewis, M., Cryder, C., Sprigg, J., 2016. Enduring effects of goal achievement and failure within customer loyalty programs: a large-scale field experiment. Mark. Sci. 35 (4), 565–575.

Watson, H.J., Wixom, B.H., 2007. The current state of business intelligence. IEEE Comput. 40 (9), 96–99.

Weaver, W., 1948. Science and complexity. Am. Sci. 36 (4), 536–544.

Wellman, M.P., 1995. The economic approach to artificial intelligence. ACM Comp. Surv.. ACM Press, New York.

Wooldridge, J.W., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.

World Bank, 2013. Gini index. Washington, DC

Yankelovich, D., Meer, D., 2006. Rediscovering market segmentation. Harv. Bus. Rev. 84 (2), 122–131.

Yu, K., Moyeed, R.A., 2001. Bayesian quantile regression. Stat. Prob. Lett. 54 (4), 437–447.

Zhou, L., Zhang, P., Zimmermann, H.D., 2013. Social commerce research: an integrated view. Electron. Commun. Res. Appl. 12 (2), 61–68.