# Towards Methods for Systematic Research On Big Data

Manirupa Das, Renhao Cui, David R. Campbell, Gagan Agrawal, Rajiv Ramnath
Department of Computer Science and Engineering,
The Ohio State University
{das.65, cui.182, campbell.1760, agrawal.28, ramnath.6}@osu.edu

*Abstract*—**Big Data is characterized by the five V's - of Volume, Velocity, Variety, Veracity and Value. Research on Big Data, that is, the practice of gaining insights from it, challenges the intellectual, process, and computational limits of an enterprise. Leveraging the correct and appropriate toolset requires careful consideration of a large software ecosystem. Powerful algorithms exist, but the exploratory and often ad-hoc nature of analytic demands and a distinct lack of established processes and methodologies make it difficult for Big Data teams to set expectations or even create valid project plans. The exponential growth of data generated exceeds the capacity of humans to process it, and compels us to develop automated computing methods that require significant and expensive computing power in order to scale effectively. In this paper, we characterize data-driven practice and research and explore how we might design effective methods for systematizing such practice and research [19, 22]. Brief case studies are presented in order to ground our conclusions and insights.**

***Keywords: Data-driven research; Agile; Data Science; Methodology; Experimental Methods***

## I. Introduction and Problem Statement

"Data science" is the science of extraction of "actionable knowledge", usually from "Big Data", that is, often large volumes of unstructured or structured data generated by systems, people, sensors or devices, or personal, social and digital traces of information from people. "Unstructured data" may include web pages, blogs, news and social media, repositories of texts such as publications, internal organizational knowledge bases, emails, videos, photos and a host of user-generated content. Structured data is data whose schemas are known, and resident in identifiable repositories, such as databases. Big Data has been characterized in terms of its volume, variety, velocity, veracity and value [9], with the worth, or the *value* of Big Data and data science, being what we *do* with it.

We start by defining Data Science more precisely, as the use of statistical and machine learning techniques on big multi-structured data in a distributed computing environment to identify correlations and causal relationships, classify and predict events, identify patterns and anomalies, and infer probabilities, interest and sentiment. Data Science has been termed the science of building data *products*, i.e., software products that provide a data-supported service (such as recommendation or prediction) whose core function relies on the application of statistical or machine learning methods. The process of building data products needs to scale to deal with volume, variety and velocity while also addressing veracity (the "messiness" of data being generated), in order to create value [9].

Data science is different from database query and retrieval. While database querying asks, "What *data* satisfy this *pattern* (query)?" data-driven discovery asks, "What *patterns* satisfy this *data*?" [1]. We want interesting and robust patterns that satisfy the data, where "interesting" is something *unexpected* and *actionable* with "robustness" is a pattern expected to occur in the future [1]. Big data promises automated actionable knowledge creation and predictive models for use by both humans and computers and makes it feasible for a machine to ask and validate interesting questions humans might not consider [1]. Data-driven analytics pipelines thus often comprise the following activities: (i) Descriptive Analytics (What happened?), (ii) Diagnostic Analytics (Why did it happen?), (iii) Predictive Analytics (What will happen?) and (iv) Prescriptive Analytics (How can we make a desired effect happen?).

The first two scenarios represent traditional business intelligence (hindsight-oriented) scenarios. However, a common epistemic requirement in assessing whether new knowledge is actionable for decision-making is its predictive power, not just its ability to explain the past [1]. The final two scenarios are the ones that represent the true essence of data science activities (insight and foresight oriented), involving analysis, representation of findings, application of statistics, machine learning, domain knowledge, programming, model development and model validation to generate actionable insights. This has naturally led to the need for *data scientists* and knowledge engineers who must be equal parts analyst, computational scientist, statistician, programmer and domain expert, as new approaches and algorithms to analyze dynamic data streams and enable analytic workflows often have to be custom designed, rather than can be readily found in the software marketplace.

In general, traditional database-focused methods are not suited for knowledge discovery because they are optimized for fast access and summarization of data given what the user *knows to ask* (query), not discovery of patterns in massive swaths of data when users lack a well-formulated query [1]. In other words, knowledge discovery is more ad-hoc in nature, and a different paradigm from well-defined queries and scheduled report generation. Given this, the software and hardware architectures used for data science are very different from traditional database systems. Software applications provide *probabilistic* answers, and hardware architectures are designed for *exploration* at scale, rather than high-volume transaction processing.

In the following sections, we further explore the above concepts. We do this through the exemplars of several projects undertaken by us that use large, heterogeneous, or complex data sets for analytics and mining insights by providing ad-hoc tools for querying a dataset or to answer a larger business or research question. Based on our experiences, we then present our perspectives on characterizing data-driven research and a discussion of tools, methodologies and processes to make data-driven research more systematic. We then conclude with a discussion regarding systemization of data-driven research practices.

## II. Case Studies

### A. *Maximum Entropy Churn Prediction Using Topic Models [13]*

In this work, our objective is to explore various structured and unstructured data available within a news organization, from print and on-line properties, to gain insight into various factors affecting newspaper subscriber engagement [13]. We then use these insights, to come up with predictive models for *customer churn* using features mined from transactional databases or Web-based textual data to determine which factors most impact user engagement, using appropriate methods for each. We use any available Web data indicative either of user activity, e.g., search clicklogs, or of readership, e.g., Web news, a potential rich source of signal for the task of churn prediction, hence these sources are also included into our study to determine factors affecting churn. We hypothesize that clicklogs provide important signals related to traffic patterns and preferences of users, e.g., searches for items that: (i) cannot be easily or directly found in news or advertisements, (ii) are very related to news or advertised products, or (iii) are unrelated to news. Similarly, Web news offers almost similar content to the subscriber as print. We therefore use Web news as a source of signal to look at the impact of environmental context on consumer behavior hypothesizing that *top ranking news items* may provide valuable cues into user engagement and hence may be correlated. Our strategy was to mine these Web information sources using an unsupervised learning approach such as LDA-based topic modeling, in order to extract useful features to gauge user engagement, followed by a supervised learning using these features, for churn prediction. The dataset comprised data from various divisions of the enterprise, viz. newspaper subscriber transactional data with subscription history, viz. current status of a subscription, time-stamped transactions for start, stop, changes to or renewal of a subscription and associated memo text, and news stories, blog content and comments data from thirteen different websites. We trained a maximum-entropy based classifier on features extracted from the **TRANS** (print subscription history portion of the dataset) and **WEB-NEWS** & **WEB-CLICKLOG** (Web portion of the dataset) integrating features derived from activity, stories and sentiment from the Web, to create models of churn based on linguistic, temporal and transactional metadata features. Our experiments reveal some insightful findings for this dataset, such as *local* news is more predictive than *national* news from our Web models of churn, and that complaints beyond a certain number actually predicts an active subscriber, from our transactional models of churn [13].

### B. *Mining Emotion-Word Correlations in a Large Blog Corpus [11]*

Blog data presents us with unique challenges. Informal speech is less structured, and is harder to make sense of with traditional methods. Users express ideas differently in blogs e.g., emoticons, neologisms, and memes, in addition to very large data sizes. We are interested to know more about words people choose when expressing themselves in a blog environment and in identifying which, if any, correlations exist between words from specific semantically related categories drawn from a basic theory on emotion [11]. This was an exploratory study, where we try to determine whether the words people choose correlate well with these categories. If successful, we could use this information to better predict how blog entries might cluster based on emotion, leading to improved models of information retrieval for blogs, or a better understanding of theoretical models of emotion in the unstructured literary domain. We used the dataset made available by Spinn3r.com for the ICWSM 2009 data challenge, a set of 44 million blog posts spanning 62 days between August 1st and October 1st, 2008, covering some big news events such as the 2008 Olympics, both 2008 US presidential nominating conventions, and the beginnings of the financial crisis. The total size of the dataset was 142 GB uncompressed, or about 30 GB compressed. We used the Ekman (1972) model of six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. WordNet was used to expand the names of each of these categories into groups of related words. For article selection, we removed all the non-English articles and retained only articles from six website domains viz. MySpace, WordPress, LiveJournal, BlogSpot, Vox and TypePad. We wrote various programs and scripts using Java, Python, and Unix shell scripting, to carry out pre-processing, indexing, and analysis for our dataset. In particular, we wrote programs for pre-processing of the corpus for article selection using emotion words, document word frequency generation, compressed bit-vector indexing of the preprocessed data, association rule mining [6], and shell scripts for executing these on a Linux cluster. Our index achieved high compression (75x), i.e., ~400MB down from 30G, thus it was able to fit in-core, and association rule mining actually became feasible and quick [11].

### C. *Using Latent Semantic Analysis to Identify Successful Bloggers [12]*

Identifying influential bloggers in a weblog community by analyzing the blog network (Agarwal et al., 2008) is a research area of interest. In this work, we hypothesized that there may exist characteristics of language use by informal writers, such as vocabulary or word choice, that are directly associated with successful communication. Specifically, we hypothesized a relationship between the vocabulary of a blog and comment density [12]. In this study, we used latent semantic analysis (LSA) to reduce the dimensionality of a term-document matrix for each blog in a collection (where a blog is a concatenated set of blog entries). We then performed two separate experiments. First, using an unsupervised clustering approach to see if relationships to comment density naturally emerge from this analysis. The results of this approach suggest that naive clustering attempts end up clustering documents by topic and subtopic rather than communication style. Second, we attempted a supervised classification method to identify high and low comment density blogs, by using two complimentary models built through LSA. The results of this approach were above chance levels, suggesting potential future directions for this research [12].

### *D. Brand specific tweet classification with user provided topics*

Many companies take feedback information on their products from Twitter, from where a large number of tweets are collected for data analysis for the purpose of market research. In this project, we aim to label tweets that mention a certain brand or its product with predefined names. Classification on tweets becomes critical and essential with the dramatic increase in size of the data, and it is helpful for many downstream processing tasks especially human reading of the tweets, as it reduces the volume of the data and increases the concentration of the data within each label. The companies interested in feedback also define some rules containing keywords and simple logic to be able to label some tweets into certain bins of interest. However, this keyword-based solution can only cover a small portion of the vast tweet data. Therefore, starting from the brand specific data, and the simple keyword-based logic rules, we build a system that is able to label many more tweets with a certain confidence level. We utilize the Labeled Latent Dirichlet Allocation (LLDA) [20] model to assign topics/labels to each tweet. The LLDA model holds the assumption that each document/tweet is generated by using related words from some underlying topics, therefore these topics decide the idea of the document. We collect the mentioning tweets for 5 brands and build a topic model for each of them. The number of predefined labels, the ways of creating the keyword-based logic, and the amount of tweets collected for the same 6-month period differ across all brands. All these limitations make the task very domain-specific and require a data-driven subtask for each brand. Overall, we can get a precision of around 85-90% with a reasonable level of recall considering the size of the available tweet data.

Table I shows the abstracted characteristics of these projects, with a view to characterizing data-driven research as described in sections to follow.

## III. Characterizing Data-Driven research

Data-driven research differs from typical research from the standpoint of how the research goal is defined. Typical research starts from a pre-determined goal, and then collects data and validates and builds models to achieve the goal. A data-driven research project, on the other hand, starts from the *data*, and tries to reveal the pattern or information stored in the data, before establishing the goal of the research. Furthermore, data-driven research evolves the goal with more sophisticated information discovered from the data until a satisfactory conclusion is reached.

**Table 1. Projects and Characteristics**

| **Projects** | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| Scientific Discipline/ Industrial Domain | Media, Publishing | Data Mining, Computational Linguistics | Artificial Intelligence Computational Linguistics | Market Research |
| Mostly Structured Data | Yes | No | No | No |
| Mostly unstructured Data | Yes | Yes | Yes | Yes |
| Hypothesis Testing | Yes | No | Yes | No |
| Hypothesis Generation | Yes | Yes | Yes | No |
| Internet-based | Yes | Yes | Yes | Yes |
| Scale | TB | GB | GB | TB |
| Distributed elements | Yes | Yes | Yes | Yes |
| Computationally intensive Data Preparation | Yes | Yes | Yes | Yes |
| Computationally intensive Execution | Yes | Yes | Yes | Yes |
| In-memory execution | No | Yes | No | No |
| Parallelizable code | Yes | Yes | Yes | Yes |
| Non-traditional analysis | LDA-based Topic modeling | Association Rule Mining | Latent Semantic Analysis | LLDA-based Topic modeling |
| Ad-hoc data product | Yes | Yes | Yes | No |

In this sense data-driven research is atypical in that we may not have a clear purpose and outcome defined at the very outset, but evolve this in an iterative fashion.

E.g., we notice from Table 1 that we were generating hypotheses in more cases than we were testing hypotheses. Further, most of our projects made use of more unstructured, rather than structured relational data, thus introducing a necessary step of *Information Extraction and Cleaning*. They all used user-generated Internet-based content in some fashion; all involved both computationally intensive data preparation and execution phases; and distributed computational elements including offline or ad-hoc batch processing. All of them involved developing an algorithm for non-traditional analysis or resulted in an ad-hoc data product.

We underline here that despite the undefined and evolutionary aspects of data-driven research, it still builds on the established paradigms of traditional research, where at a certain point after research goals have been sufficiently honed in and agreed upon, standard approaches are applied. Thus, bearing in mind this notion, that in data-driven research, *starting from the data* itself, we aim to define what our research objectives should be, ***following are certain key considerations we must take into account*** to characterize our own primary research activities to better evaluate which aspects of our research fit within a data-driven model:

### A. Clarity About Purpose ([4], Ch. 5)

Purpose is the controlling force in traditional research. Decisions about design, measurement, analysis, and reporting all flow from purpose. Therefore, the first step in a research process is getting clarity about purpose. The centrality of purpose in making methods decisions becomes evident from examining alternative purposes along a continuum from theory to action. The five different types of research along this continuum are:

*1-Basic research*: Contributes to fundamental knowledge and theory

*2-Applied research*: Illuminates a societal concern or problem in search for solutions

*3-Summative evaluation*: Determines if a solution (policy or program) works

*4-Formative evaluation*: Improves a policy or program as it is being implemented

*5-Action research*: Understands and solve a problem as quickly as possible.

Standards for judging quality vary among these five different types of research. Expectations and audiences are different, as are reporting and dissemination approaches. Different purposes lead to different ways of conceptualizing problems, different designs, different types of data gathering, and ways of publicizing and disseminating findings.

### B. Methods considerations: Contrasting Qualitative and Quantitative Approaches and Outcomes ([4], [8])

*“The key to making good forecasts is weighing quantitative and qualitative information appropriately*” – Nate Silver, 2012, [4], suggests that thinking about design alternatives and methods choices leads directly to consideration of the relative strengths and weaknesses of qualitative and quantitative data, where some questions naturally lend themselves to numerical answers, while some don’t. Quantitative methods require the use of standardized measures so that varying perspectives and experiences of people can be fit into a limited number of predetermined response categories to which numbers are assigned, thus facilitating comparison and statistical aggregation of the data, giving a broad, generalizable set of findings presented succinctly and parsimoniously. By contrast, qualitative methods facilitate study of issues in depth and detail, i.e., they typically produce a wealth of detailed information about a much smaller number of people and cases. This increases the depth of understanding of the cases and situations studied, but reduces generalizability [4]).

Quantitative research [8], is defined as *“Explaining phenomena by collecting numerical data that are analyzed using mathematically based methods (in particular statistics)”.* – Aliaga and Gunderson (2000). Explaining phenomena is a key element of all research, be it quantitative or qualitative. When we set out to do some research, we are always looking to explain something. E.g., in Education, this could be questions like ‘Why do teachers leave teaching?’ or ‘What factors influence pupil achievement?’ In case of quantitative research, the specificity lies in collecting numerical data. In order to be able to use mathematically based methods, our data have to be in numerical form. This is not the case for qualitative research. Qualitative data is not necessarily numerical, and therefore cannot be analyzed by using statistics alone, while quantitative research is about collecting numerical data to explain a particular phenomenon [8]. Many researchers, both quantitative and qualitative, take a pragmatist approach to research, using different methods depending on the research question they are trying to answer. In some cases, this will lead them to quantitative research, as, for example, when they need to give a quantitative answer to a question, need to generalize findings to a population, or are looking to test a theory mathematically; in other cases, they will employ qualitative methods. In many cases, a mixed-methods approach combining quantitative and qualitative methods will be the most appropriate [8]. The four main types of research questions that quantitative research is particularly suited to are:

1. Questions demanding a quantitative answer, e.g., ‘How many students choose to study Science?’

2. Trends or numerical change can likewise accurately be studied only by using quantitative methods, e.g., Are the numbers of students in our university rising or falling?

3. Wanting to find out about the state of something or other, we often want to explain phenomena, e.g., what factors predict the recruitment of Math teachers?

4. The final activity for which quantitative research is especially suited is the testing of hypotheses [8]

*“A hypothesis is a tentative explanation that accounts for a set of facts and can be tested by further investigation.”* E.g., one hypothesis we might want to test is that poverty causes low achievement. Quantitative researchers design studies that allow us to test these hypotheses. We collect the relevant data (e.g., parental income and school achievement) and use statistical techniques to decide whether or not to reject or provisionally accept the hypothesis. Accepting a

hypothesis is always provisional; as new data may emerge that reject it later on. Problems one and two above are called 'descriptive'. We are merely trying to describe a situation. Three and four are 'inferential'. We are trying to explain something rather than just describe it [8].

While quantitative methods are good at answering the above four types of questions, there may be other questions that are not well suited to quantitative methods, e.g.,

1. When we want to *explore a problem in depth*. Quantitative research is good at providing information in breadth, from a large number of units, but when we want to explore a problem or concept in depth, quantitative methods can be too shallow. To really get under the skin of a phenomenon, we will need to go for ethnographic methods, interviews, in-depth case studies and other qualitative techniques ([8], Ch. 1).
2. Quantitative research is well suited for the *testing of theories and hypotheses*. What quantitative methods cannot do very well is develop hypotheses and theories.
3. If the issues to be studied are *particularly complex,* an *in-depth qualitative study* (a case study, for example) is more likely to pick up on this, than a quantitative study.
4. Finally, while quantitative methods are best for looking at cause and effect (or causality), qualitative methods are more suited to looking at the meaning of particular events or circumstances ([8], Ch. 1).

If we want to look at both breadth and depth, or at both causality and meaning, it is best to use a so-called mixed-methods design, in which we use both quantitative (e.g., a questionnaire) and qualitative (e.g., a number of case studies) methods. Mixed-methods research is a flexible approach, where the research design is determined by what we want to find out, rather than by any predetermined epistemological position. In mixed-methods research, qualitative or quantitative components can predominate, or both can have equal status [8].

### *C. Type and Availability of Data ([9], Ch. 2; [6], Ch. 1)*

Business Analytics, data-driven research and science and Big Data work with structured data (data located in a fixed field within a defined record or file, e.g., in a relational database or spreadsheet) and unstructured data (websites, text files or documents such as email, blogs and social media posts, images, videos, slideshows etc.) [9]. Internal data accounts for everything a business currently has or can access, e.g., Customer feedback, Sales data, Employee or customer survey data, CCTV video data, Transactional data, Customer record data, Stock control data, HR data [9]. External data is the infinite array of information that exists outside a business. External data is either public or private. Examples of external data include: Weather data, Government data such as census data, Twitter data, Social media profile data, Google Trends, or Google Maps [9]. Most human and computer based activities leave a digital trace (or data) that can be collected and analyzed to provide insights. Data is now being mined from our: activities, conversations, photos and video, sensors, the Internet of Things [9]. Our activities leave a data trail when we go online so that web servers log what we are searching for, what websites we visit, and what we share or like or buy, how much we paid for it, when it was delivered, and often what we thought of the product or service [9]. Social media is creating unfathomable amounts of data: e.g., More than a billion tweets are sent every 48 hours; every minute, 293,000 status updates are posted on Facebook; and every second, two new members join LinkedIn [9]. An estimated 571 new websites are created every minute of the day. Every minute, Tumblr owners publish approximately 27,778 new blog posts and 3 million new blogs come online every month [9]. Thus, data can be structured or unstructured and may be in various forms such as recorded, individual daily traces, time-varying, distributed, graph-based, or from the social Web, sensors and the Internet ([9], Ch. 2).

*Heterogeneous and Complex Data ([6], Ch. 1):* Rapid advances in data collection and storage technology have enabled organizations in research and industry, to accumulate vast amounts of data. Often, traditional analysis techniques cannot be used due to the massive size of the dataset, or sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even though the dataset is small. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical, or even more complex data objects. Examples of non-traditional data types include collections of Web pages containing semi-structured text and hyperlinks, DNA data with sequential and 3D structure, and climate data that consists of time series measurements (temperature, pressure, etc.) at various locations on Earth's surface. Exploring and mining such complex data objects needs to take into consideration relationships in the data, e.g., temporal and spatial auto-correlation, graph connectivity, parent-child relationships between elements in semi-structured text and XML documents. In the scientific communities such as ocean modeling or weather forecasting self-describing file formats such as netCDF or HDF are well known.

*Data Ownership and Distribution ([6], Ch. 1):* Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead the data is geographically distributed among resources belonging to multiple entities, which requires implementation of distributed algorithms that take into account: (1) how to reduce the amount of communication needed to perform distributed computation, (2) how to effectively consolidate results obtained from different sources, and (3) how to address data security issues.

### *D. Type of Experiments ([9], Ch. 3)*

Numerous classifications of experiments exist in scientific literature. According to the primary goal of an experiment, there exist ***testing*** experiments (the empirical verification of a certain hypothesis) and ***search*** experiments (the acquisition of necessary empirical data for constructing or refining a conjecture or idea). Based on the character and diversity of the means and conditions of an experiment, it is possible to separate out ***direct*** experiments (the means are directly used to study an object), ***model*** experiments (an object's model is used in place of the object), ***field*** experiments (in natural conditions, e.g., in space), and ***laboratory*** experiments (in artificial conditions). As described earlier, we can also consider ***qualitative*** and ***quantitative*** experiments (depending on different results of an experiment) where, a qualitative experiment may be conducted for identifying the impact of certain factors on an analyzed process (without establishing a precise quantitative relationship between characteristic parameters), while, to guarantee exact values of essential parameters influencing the behavior of a studied object, one should organize a quantitative experiment. According to the character of experiment's strategy, we can distinguish among: 1. experiments implemented by the trial-and-error technique; 2. experiments based on a closed algorithm; 3. experiments involving the "black box'' technique, leading from conclusions by the knowledge of a function to cognizing the object's structure; 4.experiments using the "open box,'' enabling the design of a sample with given functions (based on the knowledge of structure). In many cases, when addressing direct experiments or with material models in real systems is impossible, computer ***simulation*** experiments dramatically simplify the research process (they serve to "reproduce'' different situations by developing a model of a studied system). ***Retrospection*** is a look in the past, a review of the past events. Retrospection research aims to study the state of an object and its development trends historically. ***Forecasting*** is a special scientific study of concrete development prospects of an object. Naturally, each branch of scientific knowledge possesses well-established traditions in treating and applying research methods ([7], Ch. 3).

### *E. Type of Analytics ([6], Ch. 1)*

Knowledge discovery tasks are divided into two major categories:

a. **Predictive tasks:** The objective of these tasks is to predict the value of a particular attribute based of the values of other attributes. In predictive modeling, we build a model of the target variable as a function of the explanatory variables. There are two types of predictive modeling tasks: 1. ***Classification*** – used for predicting a category e.g., whether a Web user will make a purchase or not (discrete target variable, i.e., binary-valued) [8], 2. ***Regression*** – used for predicting a value, e.g., predicting stock prices (continuous target variable) [8], and a third, 3. ***Recommendation*** – used for predicting a preference (item-based, content-based, collaborative filtering) [10]

b. **Descriptive tasks:** The objective here is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in the data. Descriptive tasks are often exploratory in nature and frequently require post-processing techniques to validate and explain the results. ***Cluster analysis*** seeks to find groups of closely related observations that belong to the same cluster. ***Anomaly detection*** identifies observations that are significantly different from the rest of the data, known as anomalies or outliers. ***Association Analysis*** is used to discover patterns that describe strongly associated features in the data. The discovered patterns are typically represented in the form of implication rules or feature subsets. Because of the exponential size of its search space, the goal of association analysis is to extract the most interesting patterns in an efficient manner, e.g., finding groups of genes that have related functionality or understanding the relationships between the elements of the Earth's climate system ([6], Ch. 1).

*Desired Response Times:* Depending on nature of desired output of the research activity, e.g., whether data processing/analytic job submission must be done in batch, or interactively, i.e., either incremental or real-time updates or responses are desired, would have an impact on various aspects of the research be it experimental design or participant engagement, hence this is a major consideration in data-driven research design.

*Non-traditional Analysis* ([6], Ch. 1)*:* The traditional statistical approach is based on a hypothesize-and-test paradigm, i.e., a hypothesis is proposed, an experiment is designed to gather the data, and the data is then analyzed with respect to the hypothesis. This process is extremely labor-intensive. Current data analysis tasks often require generation and evaluation of thousands of hypotheses, and consequently, the development of some techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Furthermore, the data sets analyzed are typically not the result of a carefully designed experiment and often represent opportunistic samples of data, rather than random samples. Also, the data sets frequently involve non-traditional types of data and data distributions.

### *F. Infrastructure considerations*

Infrastructure and resource considerations such as mode of data acquisition, storage platform, available compute and backup resources, project members with technical and domain expertise, availability of technical and system support personnel are important supporting factors in the execution of a data-driven research project. Other considerations relating to implementation are scalability, dimensionality and traditional methods.

*Scalability ([6], Ch. 1)*: Because of advancements in data generation, data sets with sizes of terabytes and even petabytes are nowadays very common. Thus many data mining algorithms have to employ special search strategies to handle exponential search problems. Scalability may require the implementation of novel data structures to access individual records in an efficient manner, e.g., *out-of-core* algorithms may be required when processing of datasets cannot fit *in main memory*. Scalability can also be improved by using *sampling* or developing *parallel* and *distributed* algorithms

*High dimensionality ([6], Ch. 1)*: It is now common to encounter data sets with hundreds or thousands of attributes instead of just a handful, e.g., in bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality, e.g., measurements of temperature at various locations, or temperature measurements taken repeatedly for an extended period, where the number of dimensions increases in proportion to number of measurements taken.

Data-driven methods employing e.g., data mining and information retrieval techniques, draw upon ideas from: (1) sampling, estimation, hypothesis testing from statistics (2) search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning, (3) optimization, evolutionary computing, information theory, signal processing and visualization. Traditional areas of computing play key supporting roles. In particular, database systems are needed to provide support for efficient storage, indexing and query processing. Techniques from high performance (parallel) computing, and distributed techniques are also very important in addressing not only the massive size of certain data sets but also when the data cannot all be gathered, stored or processed at one location ([6], Ch. 1).

## IV. TOOLS, METHODS AND PROCESSES

### A. *Computing Platforms*

The selection of tools for data analytics is roughly analogous to choosing the architecture for an Agile project. There is an enormous variety of frameworks for approaching big data analytics, and the choice of which one to use requires more careful thought than to use Hadoop. These frameworks are developed by companies such as Google, Yahoo, Facebook, or Amazon to solve data problems at scale on distributed computing clusters running on commodity hardware in their datacenters. The primary goal of such frameworks is to maintain roughly linear growth in their ability to process data as an organization adds machines to a distributed computation, and avoid or mitigate bottlenecks that can limit the throughput of a large system. Each framework is designed around a core set of requirements that force the developers to choose trade-offs in terms of consistency, partitioning, and availability as stated by Eric Brewer's CAP theorem. [27]

Google introduced two of the foundational technologies for big data analytics, in the form of the Google File System [28] and the MapReduce algorithm [21]. GFS was introduced as a means to reliably store and access large datasets across a cluster of machines in a datacenter. To efficiently process these distributed datasets, Dean & Ghemawat introduced the MapReduce algorithm [21], which borrows the functional programming concepts of map and reduce to apply them to a distributed dataset to solve embarrassingly parallel problems.

For an organization seeking to methodically choose the appropriate architecture for their dataset, this introduces two important questions about which tool is appropriate for the job. The first is whether the dataset is so large that it needs to be stored on a cluster of machines in the first place, and the second question is whether the analysis of data will be suited to parallel computing or whether it requires some other approach. In the case that the answer to these questions is yes, the most popular tool for use is Hadoop. Originally developed by Yahoo, Hadoop is a framework written in Java for running applications on large clusters of commodity hardware and incorporates features similar to those of the Google File System and of MapReduce. Hadoop implements the MapReduce algorithm to run on the Hadoop Distributed File System, HDFS, which is a highly fault-tolerant distributed file system and like Hadoop designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications that have large data sets (In the range of terabytes to zetabytes).

While it is a common recommendation to use Hadoop for analytics, it's not always the case that it is the most appropriate tool for the job. The Hadoop Distributed File System (HDFS) is originally designed to run on commodity hardware and for embarrassingly parallel applications. Using Hadoop introduces some distinct disadvantages but also a lot of computational overhead which must be carefully balanced with the goals of a group. Hadoop is designed to handle large files that can be disassembled into blocks, so that when a dataset is loaded into Hadoop, the blocks must be reliably replicated across the HDFS, which incurs significant disk and network costs. As pointed out by Rowstron & Narayanan et al. [29], sometimes those engaged in data analytics are better off choosing singular, more powerful computers, or employing other alternatives when the volume does not warrant the use of Hadoop.

### B. *Methodologies ([3], Ch. 1)*

Agile software development is a group of software development methods in which solutions evolve through collaboration between self-organizing, cross-functional teams [30]. It promotes adaptive planning, evolutionary development, early delivery, continuous improvement, and encourages rapid and flexible response to change. These lightweight methods included: from 1994, unified process and dynamic systems development method (DSDM); from 1995, SCRUM; from 1996, crystal clear and extreme programming (aka "XP"); and from 1997, adaptive software development and feature-driven development. Although these originated before the publication of the Agile Manifesto in 2001, they are now collectively referred to as Agile methods; and often abbreviated loosely as Agile [30].

Agile Analytics [3] is a recommended development style, not a prescriptive methodology, where the dynamics of each project within each organization require practices that can be tailored appropriately to the environment, the primary objective being a high-quality, high-value, working analytic system. The following characteristics serve this goal:

*1-Iterative, incremental, evolutionary*: Foremost, Agile is an iterative, incremental, and evolutionary style of development. Work is in short iterations that are generally one to three weeks long, and never more than four weeks. The system is built in small increments or "chunks" of user-valued functionality and we evolve the working system by adapting to frequent user feedback [3].

*2-Value-driven development*: The goal of each development iteration is the production of user-valued features. While data and computer scientists appreciate the difficulty of complex data architectures, elegant data models, efficient ETL scripts etc., users may not. Users of analytic systems care more about the presentation of and access to information that helps them either solve a business problem or make better business decisions. Thus every iteration must ideally produce at least one new user-valued feature [3].

*3-Production quality*: Each newly developed feature must be fully tested and debugged during the development iteration. Agile development is not about building hollow prototypes; it is about incrementally evolving to the right solution with the best architectural underpinnings [3].

*4-Barely sufficient processes*: Agile falls somewhere between just enough structure and just enough flexibility. If a data dictionary is deemed important for use by future developers, then perhaps a digital image of a whiteboard table or a simple spreadsheet table may suffice [3].

*5-Automation, automation, automation*: The only way to be truly Agile is to automate as many routine processes as possible. Test automation is perhaps the most critical. Agile Analytics teams seek to automate any process that is done repeatedly, and thus, focus on developing user features [3].

*6-Collaboration*: Agile business intelligence acknowledges that there is a broader project community that shares responsibility for project success, which includes the sub-communities of users, business owners, stakeholders, executive sponsors, technical experts, project managers, and others. Frequent collaboration between the technical and user communities is critical to success [3].

*7-Self-organizing, self-managing teams*: Hire the best people, give them the tools and support they need, then stand aside and allow them to be successful. This is a key shift in the Agile project management style compared to traditional project management [3]. Thus the Agile project manager's role is to enable team members to work their magic and to facilitate a high degree of collaboration with users and other members of the project community [3].

### C. Challenges in data-driven research

In the process of data-driven research, clear and informative discoveries are critical and can have a big influence on the tasks to follow. However, unstructured data brings many barriers in doing this. Most difficulties with unstructured data are introduced by the informal and non-standard ways of creating the data. Therefore, it is important to also highlight certain challenges that may be presented in the face of structured versus unstructured data, whilst attempting to adhere to Agile processes in a data-driven computational environment.

1. *Informal use of language and creation of new words*: Many text-processing models rely on the use of word or tokens with a specified dictionary. Therefore, multiple user generated word variations and new-created words greatly affect classic dictionary based methods.
2. *Noise and redundant information*: Data-driven research requires collecting a wide range of data to discover and build a solid goal, so it is normal to collect a huge amount of data although not all of it is needed eventually. However, with the enormous data size, it is really easy to include a considerable portion of noise or redundant data. Such noise and redundancy makes it harder to look for a clear pattern of the data, therefore affecting the whole planning process. Moreover, noisy data is harmful to most machine learning models as well.
3. *Inadaptable methodology*: The huge difference between structured and unstructured data may create a big drop in performance or research goals when utilizing a methodology built for one type of data, with another. This challenge is very likely to happen where a lot of state-of-art models are built on structured data that cannot provide the same solution on unstructured data. Then the conversion of the methods, is yet another data-driven process where old methods are changed to be adaptive to the new unstructured data by either modifying the method given the new data, or training the old model on the new data.
4. *High-frequency data generation*: Unstructured data is generated at a very high speed, and the most useful data is very likely to be recently generated. Therefore the timeliness of the unstructured data becomes another challenge for data-driven research. The performance of a model trained on a certain collection of data may start to fade when the training data becomes old and the inference data is new. This challenge would either trigger more frequent data collection, or a model that is less sensitive to the timeliness of the data, would be needed.

## V. Discussion: How to make Data-Driven Research more Systematic

The core of data-driven research is the *data*, but it should start from a question of what to get from the research and the data. The final output of the research decides the process of how to handle the data and what to aim at the data. The output could be a real value prediction, a class label assignment, a statistical analysis, trend analysis, reports or even a method comparison and evaluation based on the data. Aiming for a systematic process for data-driven research, we believe that the data should be the focal point along each of the following *Agile analytic* steps for data-driven planning and execution of the research pipeline:

1. *Information extraction and cleaning* – Big data makes it easier to extract a huge amount of data for a certain purpose; however, it also increases the variation and noise contained in the extracted data, especially in data from social media platforms. In this case, some data or some attributes of the data would need to be eliminated regardless of the purpose of the research.
2. *Preliminary data analysis* – To find an appropriate target and starting point of the research, we need to reveal some patterns or information from the data. Simple methods can be applied as the first step of data analysis. Some unsupervised models such as clustering are easy to start with to find underlying patterns or to narrow down to specific data domain for further analysis.
3. *Research goal or Hypothesis generation* – Research goal is the key part of a successful research project, and it needs to be well designed to fulfill the requirement of the academic, business, or industrial purpose. The observation from the preliminary data analysis can provide some suggestions towards defining a good research goal.

Moreover, the research goal does not only cover what idea or result the research needs to output, but also what type and quality of the data the output needs to be.

4. *Research data design* – With the setup of the research goal, more work can proceed on the extracted data. Extended from the preliminary processing of the data, the data can be further trimmed and built into certain structures. Further analysis can then be applied on this data given the goal of the research, to ensure the representativeness and consistency of the data, which is not guaranteed by simply including a large volume of data. These analyses can be seen as an enhanced version of step 2, with more understanding about the data and the research, so that an improved research goal may be generated as in step 3.

5. *Model and feature selection* – Models are the tools to convert collected data to the type of desired output of the research goal, and features convert the collected data to a certain form that the model can handle. Models and features are closely related and serve together to affect the research output. The selection of model relies on both the project goal and collected data, while selection of features relies on the model and the data. Within this step, the main task is to build a good system to meet the goals, so that unnecessary change of the research goal is prevented unless requirements change.

6. *Output evaluation* – The evaluation of the output of the project should be geared towards the research goal, but taking into consideration the input data. A high quality output is always demanding, but to meet the specific research goal with the limitations of the model and data is a more reasonable target in terms of evaluation. Within an Agile process, the output evaluation can be done in iterative enhancements, starting at step 3. Figure 1 illustrates all of these steps in more detail.

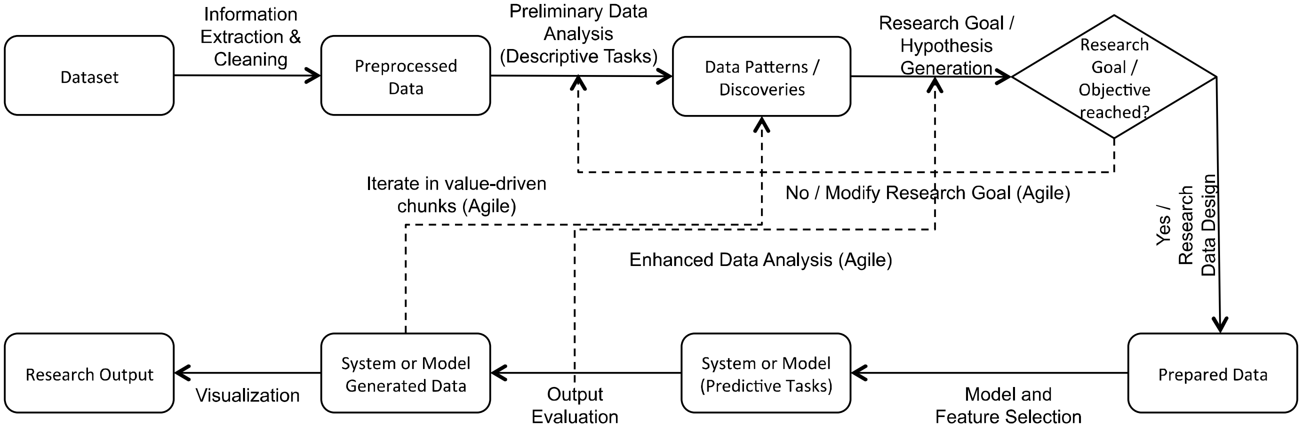
Dataset
Information Extraction & Cleaning
Preprocessed Data
Preliminary Data Analysis (Descriptive Tasks)
Data Patterns / Discoveries
Research Goal / Hypothesis Generation
Research Goal / Objective reached?
Iterate in value-driven chunks (Agile)
No / Modify Research Goal (Agile)
Yes / Research Data Design
Enhanced Data Analysis (Agile)
Research Output
Visualization
System or Model Generated Data
Output Evaluation
System or Model (Predictive Tasks)
Model and Feature Selection
Prepared Data

**Figure 1: A Process for Systematic Data-driven Research**

7. *Visualization* – As the ability and opportunity to analyze more and more data has grown, it is impossible to keep on top of it without meaningful data visualization, hence the need to find ways to communicate and report the results effectively [9]. Infographics – a hybrid of 'information' and 'graphics' – is a one-page visual representation intended to express a lot of information, data or knowledge quickly and clearly [9]. An infographic of a detailed report or data analysis or survey, for example, can instantly convey the message by using a combination of headlines, graphics and narrative to tell the whole story through a one-page visual map [9].

8. *Iterate in value-drive chunks (Agile)* – Any or all of the above steps may be repeated in an iterative fashion, till the desired results or level of performance is achieved. A feasible approach may be to execute above workflow with a small subset of the dataset, identify issues early on, and only when satisfactory, expand to the entire dataset, and expend the needed capacity, to process the entire dataset.

Besides iterative Agile planning and execution, there are other aspects that can make data-driven research more systematic and consistent, such as a generalized dataset and standardized data processing. A *generalized dataset* tries to expand the chance that a single data extraction step can work for multiple projects and purposes. This does not necessarily mean increasing the volume of the data, but enlarging the *coverage* of the data, thus including more attributes of the data. Although different research tasks may examine different aspects of the data, the initial planning step can start from the same generalized dataset and the differences can be handled by the subsequent workflow. *Standardized data processing* focuses on some *common process* of the data that can be abstracted to make it reusable for multiple projects. Given a big data dataset, especially a generalized one, many process or mapping functions are required before the data can be used. Many among these aim to solve various common problems across similar datasets, such as filtering out noise data, or fixing the usage of certain language. Standardization collects these common solutions and builds tools for similar types of data-driven research.

## VI. Conclusion

Our motivation in this paper was to explore the systematization of data-driven research practices, stemming from the realization that, all too often, it is the *volume* aspect of Big Data that receives the most attention (to an extent that not even Terabytes, but only data in the scale of Petabytes or more, is considered *Big Data*). This may be expected given that this trend and its supporting technologies first grew out of Web companies such as Google, Yahoo, Facebook and Twitter among many others, who naturally had at their disposal large volumes of data in order to serve Web users of the entire globe.

However, in most cases, an organization rarely gets data from, or needs to provide a service to the entire globe, but rather to a much more limited audience. But this does not mean that they do not have sufficient data from which they can derive value. An organization still needs to run efficiently and provide the best possible service to its customers, or produce the best possible research in its community, which can be done by harnessing systematic processes to better channel data, in order to provide *data-driven value*. Thus, we believe and emphasize, that in today's competitive research and business landscape, it is in fact the *heterogeneity* (*variety*), the *speed* at which data is being *received or generated* (*velocity*) and the *inconsistency* and *incompleteness* (*veracity*) of the data, which are the most cross-cutting aspects of Big Data, as not only many of the technical challenges presented by Big Data (as outlined by Jagadish et al. [22]), reside in exactly these three dimensions, but also because these aspects touch organizations of nearly every type and size. E.g., often, high volume and high velocity data such as that acquired by atmospheric sensors, may already be uniform in the number and types of dimensions, rendering this data relatively easier to process via embarrassingly parallel computation, compared to data which may be orders of magnitude smaller in size and relatively static, such as the archived transactional or email data of an organization, where it is much harder to determine important dimensions a priori, and to specify what exactly is to be extracted and how. Thus, it is our belief that every organization can and should have its teams take advantage of not just algorithms, but also processes and best practices, surrounding Big Data, and that these processes and practices need to be defined.

We illustrate the above argument through our case studies where we demonstrate that regardless of the *scale* of the datasets in these projects, (often not exceeding a few hundred GB or TB, or performed only on a relatively static subset of the Web), we still observe most of the technical and process challenges of Big Data [19, 22, 24], given the nature of the task we are trying to perform, or the research question we are attempting to answer, or the algorithm we are trying to run on a particular dataset. Thus, drawing from our experiences in conducting data-driven research projects, we provide a fairly comprehensive overview of research methods and key considerations in characterizing data-driven research, including types of available data and experiments, and using the same, we recommend a process for performing systematic research on Big Data, akin to Agile methodology for software development.

## VII. References

[1] Dhar, Vasant. "Data science and prediction." Communications of the ACM 56.12 (2013): 64-73.

[2] Brown, B., and J. Sikes. "Minding your digital business." *McKinsey Global Survey Results* (2012): 1-9.

[3] Collier, Ken. *Agile analytics: A value-driven approach to business intelligence and data warehousing*. Addison-Wesley, 2011.

[4] Patton, Michael Quinn. *Qualitative evaluation and research methods*. 4th Edition. SAGE Publications, inc, 2015.

[5] Csikszentmihalyi, Mihaly. *Handbook of research methods for studying daily life*. Eds. Matthias R. Mehl, and Tamlin S. Conner. Guilford Publications, 2013.

[6] Pang-Ning, Tan, Michael Steinbach, and Vipin Kumar. "Introduction to data mining." *Library of Congress*. 2006.

[7] Novikov, Alexander M., and Dmitry A. Novikov. *Research methodology: From philosophy of science to research design*. Vol. 2. CRC Press, 2013.

[8] Muijs, Daniel. "Introduction to Quantitative Research." *Doing Quantitative Research in Education with SPSS*.: Sage, 2010. Print.

[9] Marr, Bernard. "Big Data: Using SMART Big Data Analytics and Metrics To Make Better Decisions and Improve Performance." (2015).

[10] Segaran, Toby. Programming collective intelligence: building smart web 2.0 applications. " O'Reilly Media, Inc.", 2007.

[11] Annatala Wolf and Manirupa Das, "Mining Emotion-Word Correlations in a Large Blog Corpus", 17 pp. OSU-CISRC-9/15-TR15. 2009.

[12] Aron Price, Manirupa Das and Annatala Wolf, "Using Latent Semantic Analysis to Identify Successful Bloggers", 4 pp. OSU-CISRC-9/15-TR16. 2009.

[13] Das, Manirupa, et al. "TopChurn: Maximum Entropy ChurnPrediction Using Topic Models Over Heterogeneous Signals." Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2015.

[14] Melnik, Sergey, et al. "Dremel: interactive analysis of web-scale datasets." Proceedings of the VLDB Endowment 3.1-2 (2010): 330-339.

[15] Flynn, Michael J., et al. "Moving from petaflops to petadata." Communications of the ACM 56.5 (2013): 39-42.

[16] Cattaneo, Riccardo, et al. "Runtime adaptation on dataflow HPC platforms." Adaptive Hardware and Systems (AHS), 2013 NASA/ESA Conference on. IEEE, 2013.

[17] Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the Web." (1999).

[19] Saltz, Jeffrey S. "The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness". Proc. of IEEE BigData 2015, First Annual Workshop on Methodologies and Tools to Improve Big Data Projects, Santa Clara. IEEE, 2015.

[20] Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora."

[21] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.

[22] Jagadish, H. V., et al. "Big data and its technical challenges." Communications of the ACM 57.7 (2014): 86-94.

[23] Anderson, Chris. "The end of theory: The data deluge makes the scientific method obsolete." (2008): 16-07.

[24] Pavlo, Andrew, et al. "A comparison of approaches to large-scale data analysis." Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009.

[25] Franklin, Matthew. "The Berkeley Data Analytics Stack: Present and future." Big Data, 2013 IEEE International Conference on. IEEE, 2013.

[26] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

[27] Eric A. Brewer. Towards robust distributed systems. (Invited Talk) Principles of Distributed Computing, Portland, Oregon, July 2000

[28] S. Ghemawat, H. Gobioff, S. Leung. "The Google file system," In Proc. of ACM Symposium on Operating Systems Principles, Lake George, NY, Oct 2003, pp 29–43.

[29] Rowston, Narayanan, et. al. "Nobody ever got fired for using Hadoop on a cluster." (2012) HotCDP.

[30] "Agile software development". *Wikipedia*. Wikipedia, 9 September 2015