The NIST IAD Data Science Research Program

Bonnie J. Dorr*, Craig S. Greenberg*, Peter Fontana*, Mark Przybocki*,
Marion Le Bras*‡, Cathryn Ploehn*, Oleg Aulov†, Martial Michel*, E. Jim Golden*, and Wo Chang*
*National Institute of Standards and Technology
†University of Maryland, Baltimore County
‡Guest Researcher
{bonnie.dorr,craig.greenberg,peter.fontana,mark.przybocki,
marion.lebras,cathryn.ploehn,oleg.aulov,martial.michel,edmond.golden,wchang}@nist.gov

***Abstract*—We examine foundational issues in data science including current challenges, basic research questions, and expected advances, as the basis for a new Data Science Research Program and evaluation series, introduced by the Information Access Division (IAD) of the National Institute of Standards and Technology (NIST) in the fall of 2015. The evaluations will facilitate research efforts, collaboration, leverage shared infrastructure, and effectively address cross-cutting challenges faced by diverse data science communities. The evaluations will have multiple research tracks championed by members of the data science community, and will enable rigorous comparison of approaches through common tasks, datasets, metrics, and shared research challenges. The tracks will measure several different data science technologies in a wide range of fields, starting with a pre-pilot. In addition to developing data science evaluation methods and metrics, it will address computing infrastructure, standards for an interoperability framework, and domain-specific examples.**

I. Introduction

Since its emergence as a uniquely identifiable field, *data science* has been of growing importance, attested to by a proliferation of government initiatives[1], research conferences[2], and academic data science initiatives and institutes[3]. As in any rapidly emerging field, there is a pressing need to explore the foundational issues underpinning data science. Indeed, the "Trends and Controversies" presented at the Data Science and Advanced Analysis conference in 2014 [2] raised a range of data science challenges, research questions, and expected advances.

A new Data Science Research Program (DSRP) introduced by the Information Access Division (IAD) of the National Institute of Standards and Technology (NIST), beginning in the fall of 2015, addresses many of the issues raised. The DSRP aims to facilitate and accelerate research progress in the field. Here, *data science* is viewed as the application of techniques for analysis and extraction of knowledge from potentially massive data. This includes notions of *big data* technical challenges in distributed and parallel processing, processing speed, and storage architectures for high *Volume* and *Velocity*, as well as the unique challenges for data visualization. The DSRP also encompasses considerations and insights that might be central even with datasets that are smaller, such as data diversity (*Variety*) and data uncertainty (*Veracity*).

The above discussion brings to light the inherent breadth of data science (DS)—spanning systems (including databases), programming languages, machine learning, statistics, and visualization, and a myriad of other disciplines, including (broadly) the natural sciences, applied sciences, and humanities. This necessary but overwhelming breadth makes clear the need to foster collaboration, provide the opportunity to coordinate research efforts, and leverage shared infrastructure across diverse communities, which are all needed in order to accelerate progress and to effectively address the present cross-cutting challenges. Several of these challenges are described in this paper.

In order to address this need, the DSRP will be developed initially[4] by means of the following four key elements:

- **Evaluation and Metrology**: Design and conduct a new international *Data Science Evaluation (DSE)* series (Section II).
- **Standards:** Leverage prior work to develop standards for data science (Section III).
- **Compute Infrastructure:** Develop an Evaluation Management System (EMS) to support compute and infrastructure needs (Section IV).
- **Community Outreach:** Build a community of interest within which data scientists can more effectively collaborate through coordination of their efforts on similar classes of problems (Section V).

A further breakdown of the elements making up this research program are outlined in Figure 1.

[1]Examples include: DARPA's announcement of the XDATA Program, NSF announcement of new Big Data solicitation of $10 million in March of 2012, NIH announced recruitment of an associate director for Data Science in 2013. Additionally, the White House appointed the first U.S. Chief Data Scientist in Feb 2015 [1]

[2]Such as: ACM's International Conference on Knowledge Discovery and Data Mining; International Conference on Big Data Analytics; IEEE's International Conference on Cloud and Big Data Computing, and International Conference on Data Science and Advanced Analytics

[3]For instance: Columbia University's announcement to create Data Sciences Institute, UC Berkeley announces first online Master of Information and Data Science degree, UMass Amherst establishes Center for Data Science, University of Michigan establishes a new data science major.

[4]As a multi-year research program, the make-up of the DSRP is expected to change and grow over time as the field and technology matures.

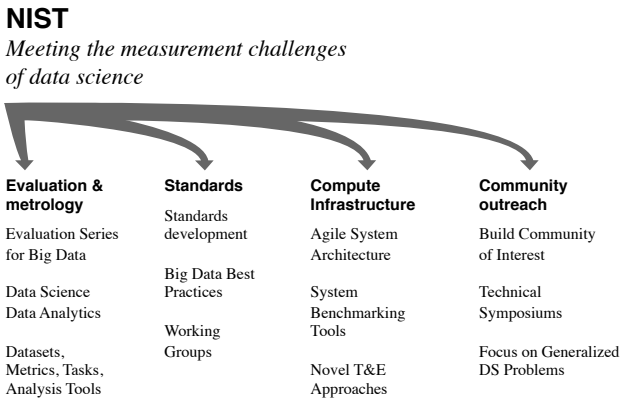

Fig. 1. NIST's role in addressing data science challenges.

This paper contains no new and novel algorithms, experiments, or results. Nor does it prescribe specific methodologies or solutions. Instead, it discusses a range of foundational data-science challenges as well as the advances necessary to drive data science forward. The contributions of this work are (1) the clear identification and examination of challenges and advances relevant to the data science community; (2) a presentation of enabling infrastructure to support research progress in data science, including the fostering of collaboration across different research communities.

The remainder of this paper describes some of the potential future breakthroughs in data science (Section VI); presents a summary of next generation of data science challenges (Section VII); categorizes different types of data science problems into explicit classes (Section VIII); discusses aspects of data science measurement (Section IX); and the final section delivers concluding remarks regarding IAD's role in the discipline of data science.

II. Evaluation and Metrology for Data Science

NIST has been conducting evaluations of data-centric technologies since the late 1980's. These evaluations cover a wide range of technologies including: automatic speech transcription, information retrieval, machine translation, speaker and language recognition, automatic fingerprint matching, image recognition, event detection from text, video, and multimedia, and automatic knowledge base construction, among many others.

Despite the stark differences among the technologies listed above, each evaluation has enabled rigorous research by sharing the following fundamental elements: (1) the use of common tasks, datasets, and metrics; (2) the presence of specific research challenges meant to drive the technology forward; (3) an infrastructure for developing effective measurement techniques and measuring the state-of-the-art; and (4) a venue for encouraging innovative algorithmic approaches. Several NIST evaluations have enjoyed substantial popularity and provided the necessary infrastructure to accelerate research progress in the corresponding core technologies.

To address several unique challenges in the burgeoning field of data science, NIST has launched the Data Science Evaluation series (DSE), to occur annually starting in the fall of 2015. These challenges stem from some combination of data characteristics (e.g., very large datasets, multi-modal datasets, data from multiple sources with varying degrees of reliability and noise) and task requirements (e.g., building of multi-component systems, enabling effective human-in-the-loop interaction, and visualization of large and complex data).

These in turn lead to various evaluation design and implementation challenges: (1) logistical aspects of conducting very large-scale evaluations, including dataset creation and distribution, and of conducting multi-component evaluations requiring coordination and timing of individual component evaluation; (2) evaluation design challenges associated with the use of "found" data rather than data collected in a controlled manner, which increases the difficulty of conducting rigorous experiments; (3) measurement challenges arising from a lack of hand-annotated data or ground truth more generally; (4) measurement and calibration of data and system uncertainty; and (5) measurement of the effectiveness of visualization. In addition, many existing research communities are formed around individual tasks, domains, or modalities—thus a multi-modal, multi-task evaluation will require the integration of multiple disparate communities.

While previous NIST evaluations have dealt with some of the challenges above, many remain unsolved. Successful data science evaluations will require addressing many of these challenges simultaneously, and in new combinations. To that end, each year of the DSE will consist of multiple research tracks—organized by domain—encouraging tasks spanning multiple tracks. In addition to one or more NIST-led tracks, community-led tracks will be included in the DSE.

As a first step, in fall of 2015, NIST will host a small scale pre-pilot evaluation in the highway traffic domain, meant to serve as a track archetype,[5] and to surface any unexpected evaluation challenges. It will consist of heterogeneous data from traffic and weather sensors and will feature data cleaning, dataset alignment, and predictive analytics tasks. In 2016, NIST will follow up with an open pilot evaluation in the same domain and will begin accepting track proposals for a 2017 full-scale data science evaluation.

III. Standards for Data Science

The design of the new DSRP leverages prior work at NIST on standards for data science, starting with those developed for big data [3]. For example, the NIST Big Data Public Working Group (NBD-PWG) developed a consensus-based, extensible interoperability framework that is vendor-neutral, technology-independent, and infrastructure-independent [4]. This framework allows data scientists to process and derive knowledge through the use of a standard interface between swappable architectural components. The following elements

[5]It's worth emphasizing the fact that this track is meant to serve as an exemplar of a data science evaluation track, not to solve any particular problem in the traffic domain.

have been formalized by the NBD-PWG—as components of a Reference Architecture ecosystem—and are expected to apply to problems in data science more generally:

- **System Orchestrator (or data scientist)**: Provides a high-level design of the dataflow between analytics tools and given datasets, computing system requirements, and monitoring system resource and performance.
- **Data Provider**: Provides an abstraction of various types of data sources (such as raw data or data previously transformed by another system) and makes them available through different functional interfaces. This includes the transfer of analytics codes to data sources for effective analytic processing.
- **Application Provider**: Provides analytics processing throughout the data lifecycle—acquisition, curation, analysis, visualization, and access—to meet requirements established by the System Orchestrator.
- **Framework Provider**: Provides one or more instances of a computing environment to support general data science tools, distributed file systems, and computing infrastructure—to meet requirements established by the Application Provider.
- **Data Consumer**: Provides an interface to receive the value output from this Reference Architecture ecosystem.
- **Security and Privacy Fabric**: Provides the security and privacy interaction to the rest of the Reference Architecture components (via the System Orchestrator) to ensure protection of data and their content.
- **Management Fabric**: Provides management interaction to other Reference Architecture components (via the System Orchestrator) with versatile system and software provisioning, resource and performance monitoring, while maintaining data quality and secure accessibility.

Recently, the NBD-PWG released working drafts of the interoperability framework for public comment [5]. These include basic definitions (concepts and vocabulary), taxonomies, use cases, reference architecture, a standards roadmap, and other elements associated with big data that are expected to apply to the space of problems in data science more generally. This framework will be released in three stages, each corresponding to a major activity relevant to the more general data science endeavor: (1) Identification of a high-level reference architecture with the following critical components: technology, infrastructure, and vendor-agnostic capability; (2) Definition of general interfaces between the reference architecture components; (3) Validation of the reference architecture by building applications through the general interfaces.

IV. Compute Infrastructure for Data Science Research

NIST has implemented an Evaluation Management System (EMS) that will serve as the infrastructure for the DSE series. EMS integrates hardware and software components for easy deployment and reconfiguration of computational needs and enables integration of compute- and data-intensive problems within a controlled cloud. In addition, EMS enables the collection of metrics on independently running instances as well as aggregation of overall performance metrics on the core problem. This design allows for test and evaluation (T&E) of different compute paradigms (software and model changes, such as testing a solution using MPI (Message Passing Interface) and later trying it using Go channels) as well as hardware accelerations in order to best assess how a given evaluation can be run.

The underlying cloud infrastructure accommodates concurrent execution of projects—such as experiments or evaluation—on a shared hardware while being able to separate data access, network resources, users and hardware accelerators (e.g., GPU or Phi). Applications within a given project communicate with one another and access data shared with a specific user and application.

This infrastructure supports the integration of distributed as well as parallelized computations, thus providing a flexible hardware architecture for running projects on the system. Performance metrics for individual applications, their data, network and memory usages are aggregated in order to compute per-application metrics as well as global project metrics. This enables direct comparisons between different algorithmic approaches for a given project and supports studies of hardware accelerations or comparisons of compute paradigms.

The initial emphasis of the EMS is to support NIST evaluations, leveraging a private cloud infrastructure for easy deployment. To facilitate this process, a model for abstracting the complexity of inter-evaluation components (such as ingestion, validation, scoring, report generation, and return of results to participants) enables reproducibility of given problems on different compute architectures. As the model is enhanced, encrypted point-to-point communication will be integrated to protect intellectual property and sensitive data used by the infrastructure.

NIST has integrated hardware resources within a private cloud testbed (Gigabit and Infiniband networks, Tesla GPUs, Intel Phi Coprocessors, high memory compute nodes, high storage data nodes) using a local OpenStack deployment. OpenStack is open source and provides several core components that support an expandable cloud solution:

- **Computing Engine:** Deploys and manages virtual machines and other instances to handle computing tasks
- **Network Controller:** Enables fast and managed network communications
- **Storage System:** Stores objects and files (using OpenStack) and a block storage component for user control when data access speed is essential
- **Identity Services:** Provides user management
- **Image Services:** Uses virtual copies of hard disks for deployment of new virtual machine instances
- **Telemetry Services:** Keeps a verifiable count of each user's system
- **Orchestration Component:** Supports the definition of cloud resource specifications and enables the management of infrastructure for cloud services

- **Front End:** Provides a quick glance at components running on the cloud and creates new instances
- **Application Programming Interface (API):** Enables extension of core components

Since OpenStack provides block and object storage based on commodity hardware solutions, it is possible to easily add new storage components to our local cloud as the volume of data increase. Also, OpenStack can be deployed between multiples sites where each site has its own OpenStack and storage can be configured as a single shared pool or separate pools. The use of OpenStack Swift gives access to streamed data, be it local or remote via an industry-standard RESTful HTTP API. All objects are stored with multiple copies and are replicated in as-unique-as-possible availability zones and/or regions.

Our current test bed for the EMS has Gigabit as well as an Infiniband networks, 5 compute nodes with 16 cores each, 128GB, 192 GB or 256 GB of memory, and 32 TB or 48 TB of disk per node, as well as 2 extra computes nodes with 4 Tesla C2050 and 4 Phi 5100, and 5 storage nodes with 48TB of disk per node.

This cloud infrastructure allows NIST to integrate and use different technologies, such as Apache MESOS, Docker, or Google Kubernetes Containers. It also enables the use of other compute engines such as Apache Spark or Apache Hadoop.

V. Data Science Community Building & Outreach

Because data science spans a wide range of very diverse fields (biology, forensics, finance, public health monitoring, etc.), the tendency is for researchers to work independently, often applying similar, but independently developed, data-processing tools and re-solving problems that span multiple data domains. The result of this mode of operation is an overall reduction in efficiency, delayed progress, and a lack of knowledge about cross-cutting synergies and best practices for many common classes of problems.

To address issues with this siloed approach to algorithm development, NIST aims to build a community of interest within which it is expected that many of the questions posed in the sections below will be addressed. Technical symposia with a focus on generalized problems in data science are expected outcomes of this aspect of NIST's work. Within a shared community, data scientists can more effectively collaborate, coordinating their efforts on similar classes of problems.

There are already several successful examples of existing NIST programs, within which community-wide mechanisms are in place (such as symposia) for technology development, assessment, and cross-community discussion. One such example is the Text Retrieval Conference (TREC) [6], which has been held at NIST annually since 1992. This research program includes an evaluation series where researchers are able to share ideas and to compare their approaches with those of other community members by participating in shared tasks defined within tracks.

[6]http://trec.nist.gov

As a starting point, in March of 2014, NIST held the first Data Science Symposium [7], at which data scientists had the opportunity to discuss data science benchmarking and performance measurement, datasets and use cases for data science research, and challenges and gaps in data science technologies. There were over 700 registrants from the data science community—spanning multiple fields—with several dozen paper and poster presentations and breakout groups on topics related to data science, e.g., human-computer interaction, manufacturing, and meta-data.

It was at this symposium that many of the challenges and expected breakthroughs presented below were brought to the fore, and researchers in a range of different fields began to discuss best practices for development and assessment of data science algorithms. The next symposium for the DSRP will be held at NIST in winter of 2016, where researchers participating in the traffic pre-pilot will have the opportunity to evaluate the effectiveness of their algorithms on traffic incident detection and traffic prediction tasks.

It is expected that the new DSRP will leverage lessons learned in the initial pre-pilot to move forward effectively on a range of issues that carry across different domains (e.g., biology vs. finance), across different modalities (e.g., video data vs. structured reports), and for commonly occurring data-related tasks (e.g., anomaly detection and data cleaning).

VI. Where are the Important Future Breakthroughs?

To support the DSRP, a significant effort will be put toward investigation of the basic premises underlying data science, including big data, as well as a focus on the types of future breakthroughs that are expected. Four V's are often cited to illustrate the challenges and the need for breakthroughs in this field: Volume, Velocity, Variety, and Veracity.[8] The earliest formulation by Douglas Laney [8] included only the first three, briefly summarized below:

- Volume: Vast amounts of data generated from multiple sources, often too large to store or analyze using traditional database approaches.
- Velocity: Speed at which the massive data are being produced and collected, making it a challenge for real-time processing.
- Variety: Diverse and potentially incompatible data types

[7]http://www.nist.gov/itl/iad/data-science-symposium-2014.cfm

[8]A fifth V that has been proposed is Value [6], i.e., the degree to which the worth of the data is conveyed. Providing a means to visualize data can increase understandability and accessibility in ways that would otherwise be impossible, thus clarifying the underlying value of the data. In the scope of this paper Value is considered to cross-cut several data science challenges, most notably a sixth V proposed by McNulty [7] (Visualization), which we address separately as a next generation challenge.

and formats coming from multiple sources.[9]

Veracity is a fourth V, attributed to researchers at IBM [9]:

- Veracity: Quality and trustworthiness of data, given the variety of sources and degree of accuracy.

Of these four V's, the first (Volume) and second (Velocity) are critical for processing of big data. These are important aspects of the DSRP, both for the initial traffic use case where (ultimately) traffic monitoring may lead to realtime data sets (including issues of latency) and for new tracks involving very large data that one might find, e.g., in the biological domain. The third (Variety) and fourth (Veracity) encompass a wide range of next generation challenges within which algorithmic breakthroughs are critical for the advancement of data science, as will be described in the section below.

Variety, frequently referred to as *heterogeneity* [10], [11], is central to building web-scale systems for tasks such as entity resolution [12], [13]. Data diversity is a consideration for all sizes of data, not just large datasets. Indeed, a critical area of measurement science within the new DSE series is that of measuring the ability of an algorithm to analyze, assimilate, adapt, and/or fuse diverse data types.

Veracity is also a critical challenge faced by many data scientists, as the algorithms they develop are expected to apply to a wide range of diverse inputs, including data that are errorful, noisy, and inconsistent across different inputs.

The emergence of data science and the challenges associated with the four V's above are accompanied by technological progress leading to:[10]

- Massively scalable processing and storage
- Pay-as-you-go processing and storage
- Flexible schema on read vs. schema on write[11]
- Easier integration of data retrieval and analysis
- Well-supported interfaces between various domain specific languages/tools.
- Open source ecosystem during innovation[12]
- Increased bandwidth, network access, speed, and reduced latency.

The ability of data-science algorithms to address the four V's—and the provision of a methodology for assessment corresponding to challenges within these—is critical now more than ever before in light of changes such as those above.

[9]Variability is a seventh V that has been proposed [7]—distinct from the notion of Variety. The former refers to the degree to which the meaning behind data can change according to time and context; the latter refers to the degree to which data formats differ from each other, according to the domain and level of formality (e.g., structured vs. unstructured). We consider Variability to be a challenge to be addressed in different ways across domains rather than a challenge that might be more broadly addressed by techniques that carry across different areas of data science.

[10]This list of areas in which technological progress has been made is an augmented version of those presented recently by Franklin [14].

[11]Flexible schema on read is an approach that allows data to be parsed at read time, rather than requiring pre-formatting prior to loading the data. Schema on write refers to prescriptive data modeling where database schema are statically defined prior to reading the data.

[12]An "ecosystem" of service providers combined with open source development allows easier sharing of applications, cross-sector use of the same components (smart homes, city services, etc.), and exchange and re-use of applications and components.

VII. Next Generation Data Science Challenges

Several areas of data science merit an extended, in-depth study, requiring input from the research community, and aligned with next generation challenges. Table I presents some key challenges, each with a representative set of examples. The table also presents a set of traffic-related use cases, in line with the focus of the pre-pilot study mentioned in Section II. These key challenges are described in more detail below.

Provenance. Where does each piece of data come from and is that data still up to date [24]? In the context of database systems and queries, provenance refers to the ability to determine the origin of the data, or which tuples of which original databases were used (and how they were used) to produce each tuple in subsequent databases resulting from database queries [25], [26]. More generally, data provenance involves being able to determine where the data came from and the processes through which this data was derived from its original sources [27].

Data heterogeneity. How does one process data from multiple, large heterogeneous datasets? Data heterogeneity refers to different representations of the same real-world object. The differences may be structural or semantic [16].

Real time and predictive analytics. How can trends be identified and distinguished from random fluctuation in order to provide a calibrated forecast of future values. How can this be executed in real time [28]? Further, is it possible to effectively trade-off between execution time and accuracy? Predictive analytics refers to the extraction of information from data to estimate future outcomes and trends.

Knowledge assimilation and reasoning from data. How might algorithms reason with data, e.g., inferring causality [24], [29]? Knowledge assimilation and reasoning refers to understanding new data in terms of information available in an already-existing dataset, and applying the necessary processing to reflect the expert's view of the domain.

Big data replicability. How is reproducibility of data science experiments ensured, especially given that the truth may be hard to find among millions of data points where there is lots of room for error [19]? Big data replicability refers to the ability to repeat results across studies where the same research question is being asked on the same dataset.

Visualization of data. How might one visually represent data, whether in a raw form or after post-processing by any number of algorithms? Visualization refers to use of visual metaphors (boxes, groups, lines, etc.) that serve as building blocks for displaying complex data representations (e.g., spatiotemporal network analysis [30]), each with their own constraints in the amount and type of data to be displayed [31]. The integration of visualization into data science activities aids in the analysis of vast volumes of information [32], may increase efficiency [33], and may reduce user errors [20].

Data uncertainty. How might one handle quality issues due to untrustworthy or inaccurate data? Data uncertainty refers to gaps in knowledge due to inconsistency, incompleteness, ambiguities, and model approximations.

TABLE I
NEXT GENERATION CHALLENGES IN THE FIELD OF DATA SCIENCE

Challenge	Relevant Questions	Examples	Traffic Use Case
Provenance	Where did the raw data originate? Is it current? What processes were applied through which the data was derived from its original sources?	A genome sequence dataset may be recreated from raw data and the provenance records associated with genomic annotations [15].	The time of a traffic accident may be determined from traffic incident reports and provenance records associated with video data that has been cleaned and aligned with the reports.
Data Heterogeneity	How to use data from multiple large heterogeneous datasets?	A publisher may be represented either as a publication-producing entity, or as an attribute of a publication [16].	A vehicle may be represented visually in video data and descriptively in an incident report.
Predictive Analytics	How can trends be identified and distinguished from random fluctuation in order to provide a calibrated forecast of future value?	Stock market events may be forecasted from sentiments expressed in social media [17].	Future traffic patterns may be guessed from weather, imagery, and historical traffic data.
Knowledge Assimilation	How might algorithms understand new data, e.g., inferring causality from the data or accommodating real-time inference retraction?	Fraudulent activity may be inferred from (potentially altered) digital and physical data representations of known entities and events [18].	A traffic accident may be detected from the position of two cars in a video clip.
Big Data Replicability	How to reproduce experimental findings given that truth may be hard to find, consistently?	Using the same (usually massive) genomic dataset in two different studies to find genetic contributions to a particular disease may yield different results [19].	Using historical data from weather reports, traffic incident data, and traffic video data to detect an incident may yield different results.
Visualization of Knowledge	How to visually represent knowledge for decision making?	Intrusion detection systems often utilize dashboards to reflect network status and to alert security administrators of suspicious activity [20].	Visualization may be used to communicate traffic flow and accidents.
Data Uncertainty	How to handle gaps in knowledge due to the potential for untrustworthy or inaccurate data?	In RFID (radio-frequency identification) Data Management, raw antenna readings are frequently missed or tags are read by two adjacent antennas [21].	Uncertainty may arise from the lack of data available from points that occur between traffic detectors.
Mitigating Error propagation	How can algorithms mitigate cascading of error through data processing steps?	In Geographic Information Systems (GISs) inaccuracies may propagate and cascade from one layer to another, resulting in an erroneous solution to the GIS problem [22].	Errors associated with cleaning of traffic incident reports may propagate to incident detection and traffic prediction tasks.
Managing Privacy and Security	How to manage data and develop algorithms in the face of privacy and concerns/policies?	Model checking to verify that HIPAA (the federal Health Insurance Portability and Accountability Act) is being followed [23].	— (Privacy and security are not a focus in the traffic domain given the minimally restricted, or unrestricted, nature of traffic and weather data.)

Propagation and cascading of error: How might algorithms be written to mitigate propagation and cascading of error(s)? Error propagation and cascading refers to situations where one error leads to another or where a solution is skewed when imprecise or inaccurate information is combined into multiple layers [22].

Data privacy and security. How does one manage data and develop algorithms for processing data in the face of privacy and security concerns? Data privacy and security refers to the challenge of providing effective approaches for secure management of distributed data and data sharing, including those that may contain personally-identifiable information (PII). Detection of PII for anonymization purposes [34] and structural diversification for protecting privacy [35] are particularly important problems to be addressed. Other critical areas include management of access, sharing and distributability (e.g., data specific tools, metadata).

These are important challenges that cut across multiple areas of data science. There may be common algorithmic approaches and evaluation metrics associated with each of these challenges. Community input garnered within the DSRP will bring forth new insights to address cross-cutting issues pertaining to the data itself and measures associated with approaches to processing data.

The next section presents a set of representative classes of data science problems, setting the stage for defining measures to assess data science technologies within the DSRP.

VIII. CLASSES OF DATA SCIENCE PROBLEMS

This section examines several classes of problems for which techniques might be developed and evaluated across different domains, and defines representative classes of problems accompanied by examples from the planned use case of traffic incident detection and prediction, although the problem classes are broader than this single use case. Different categories of algorithms and techniques in data science will be examined, with an eye toward building an assessment methodology for the DSRP that covers each category.

Detection. Detection aims to find data of interest (often an anomaly or outlier—see Anomaly Detection below) in a given dataset. In the traffic domain, incidents are of interest, e.g., "traffic incident detection" is an important sub-problem of the traffic use case. Yang, Kalpakis, and Biem [36] analyze traffic flow in order to detect traffic incidents.

Anomaly detection. Anomaly detection is the identification of previously unseen system states that force additional pattern classes into a model. Relatedly, outlier detection is associated with identifying erroneous data items that force changes in

prediction models ("influential observations"). In the traffic case, an incident may be seen as an anomaly relative to data representing free-flowing traffic. Detection of incidents in traffic data with incident and non-incident data may also be seen as system state identification and estimation.

Cleaning. Cleaning refers to the elimination of errors, omissions, and inconsistencies in data or across datasets. In the traffic use case, cleaning might involve the identification and elimination of errors in dirty traffic detector data.

Alignment. Alignment refers to the process of relating different instances of the same object [37], e.g., a word with the corresponding visual object, or time stamps associated with two different time series.[13] In the traffic use case, this might involve aligning traffic camera video and traffic incident reports.

Data Fusion. Fusion refers to the integration of different representations of the same real-world object, encoded (typically) in a well-defined knowledge base of entity types [11]. In the traffic use case, fusion might be applied to bring together a video representation of a vehicle with a description of the same vehicle in an incident report.

Identification and classification. Identification and classification attempts to determine, for each item of interest, the type or class to which the item belongs [39]. In the traffic use case, the type of incident may be critical, e.g., slipping off the road, or stopping for an extended period of time (as in bumper-to-bumper traffic).

Regression. Regression refers to the process of finding functional relationships between variables. In our pilot traffic flow prediction challenge, we wish to predict traffic speed using covariates including flow volume, percentage occupancy, and training sets of past multivariate time series.

Prediction. Prediction refers to the estimation of a variable or multiple variables of interest at future times. In our traffic pilot, we might pose the challenge of predicting traffic flow rate as a function of other variables.

Structured prediction. Structured prediction refers to tasks where the outputs are structured objects, rather than numeric values [40], [41]. This is a desirable technique when one wishes to classify a variable in terms of a more complicated structure than producing discrete or real-number values. In the traffic domain an example might be producing a complete road network where only some of the roads are observed.

Knowledge base construction. Knowledge base construction refers to the construction of a database that has a predefined schema, based on any number of diverse inputs. Researchers have developed many tools and techniques for Automated Knowledge Base Construction (AKBC)[14]. In the traffic use case a database of incidents and accidents could be constructed from news reports, time-stamped GPS coordinates, historical traffic data, imagery, etc.

Density estimation. Density estimation refers to the production of a probability density (or distribution function), rather than a label or a value [42], [43]. In the traffic use case, this might involve giving a probability distribution of accidents happening over a given time interval.

Joint inference. Joint inference refers to joint optimization of predictors for different sub-problems using constraints that enforce global consistency. Joint inference may be used for detection and cleaning to arrive at more accurate results [44]. In the traffic use case, weather conditions may act as a constraint on traffic incident detection outcomes, while at the same time, traffic incident detection may act as a constraint on weather conditions during time periods where weather data may not be available.

Other classes of problems. Data science problems may involve ranking, clustering, and transcription (alternatively called "structured prediction" as defined above). Several of these are described by Bengio et al. [45]. Additional classes of problems rely on algorithms and techniques that apply to raw data at an earlier "preprocessing" stage.

Given the broad scope of the classes of problems above, a number of different data processing algorithms and techniques may be employed for which an evaluation methodology is essential, e.g., for benchmarking. The next section elaborates on the range of methodologies needed for measuring technology effectiveness within the new DSRP.

IX. Methodologies for Measuring Effectiveness of Data Science Technologies

This section examines a range of different questions for the development of assessment methodologies, divided broadly into three categories: (1) aspects of data science measurement; (2) how to pursue data science without compromising privacy; and (3) how to preserve and distribute data and software. These questions set the stage for the new DSRP, addressing some of the most critical issues and areas of inquiry for data science.

A. Aspects of Data Science Measurement

1) How does one measure accuracy when all truth data are not annotated fully?: Ground truth may be prohibitively expensive or laborious to obtain in cases where human-labeled data are needed. In some cases, ground truth may be entirely "unobtainable," where the true answer is not known. For most predictive tasks, ground truth data become available when realtime datasets or future data materialize (e.g., accident prediction in video). For non-predictive tasks (e.g., detection of traffic incidents), Katariya et al.'s [46] work on active evaluation of classifiers estimates accuracy based on a small labeled set and human labeler. Some NIST evaluations (TREC, [47]) apply accuracy measures that accommodate the lack of full truth data, often employing mediated adjudication approaches (e.g., pooling relevance assessments of participants in the evaluation to approximate recall). Another potential approach is to use simulated data as a proxy for ground truth.

[13]Data alignment is frequently used for entity resolution, which is identifying common entities among different data sources. Getoor and Machanavajjhala [12] and Christen [38] are two works that describe entity resolution techniques.

[14]4th Workshop on Automated Knowledge Base Construction http://www.akbc.ws/2014/

Within the DSRP, these and other approaches for addressing issues concerning ground-truth metadata will be explored.

2) How does one measure data assimilation?: Data assimilation—a process by which observations are incorporated into a computer model of a real system—addresses the problem of not knowing the initial conditions of a given system [48]. Using current and past limited available observations and short range forecasts, data assimilation analyzes the data to estimate the background of the observed system and produces the best estimate of the initial conditions of the forecast system. The better the estimate, the more accurate the forecast [49].

While assimilation and fusion are similar in nature, there are differences: assimilation refers to modeling observations of the same objects (in situ, remotely, etc.) from sensors of different types, whereas fusion refers to bringing together different datasets to arrive at a result or response. Within the DSRP, both assimilation and fusion are assumed to be central to data science measurement.

3) How does one measure knowledge representation through Visualization of data?: The Visualization Analytics Science & Technology community has developed a "VAST Challenge," run annually for the past 3 years[15], for assessment of visual analytics applications for both large scale situation analysis and cyber security. Topics of importance for the DSRP include automatic graph analysis and best practices for combined and multimodal datasets. Several different approaches to developing and assessing information visualization for very large datasets have been implemented [50], [51]. Visualization paradigms are often assessed by the number of data points and the level of granularity represented [52] and by types of relationships that can be represented [31].

4) How does one develop sound benchmark measurements that satisfactorily convey system performance?: Sound benchmarking requires the integration of a variety of research areas: the mathematics of designing good benchmark metrics, the systems research of implementing monitors that collect the data with minimal overhead, and the understanding of the field in choosing representative workflows to measure the performance of different computer systems [53], [54]. As computer systems change and needs change, the desired workflows need to be changed. Within the DSRP, the use of program-agnostic metrics and software performance monitors that can run on a variety of hardware architectures will enable the application of benchmark metrics and monitors in future workflows on different software and hardware architectures.

5) How does one measure the effectiveness of each data characteristic for making decisions or formulating knowledge?: Principal Component Analysis and other dimensionality reduction techniques give some indication of the dimensions of variation present in the data. Various feature selection approaches may be applied to better understand the contribution of data characteristics for decision making and knowledge formulation [55]. As a clarifying example, in the traffic domain within the DSRP, a task would be to determine how much lane detector, weather, and accident data contribute to the ability to perform the overall tasks of traffic incident detection and traffic prediction.

[15]The latest (2015) VAST Challenge information can be found at: http://vacommunity.org/VAST+Challenge+2015

B. How does one pursue data science without compromising privacy?

Collection and sharing strategies are needed so that researchers are able to run experiments on the same data, with minimal barriers. For example, the traffic and weather data in our pilot DSE evaluation are open and easily distributable. However, the DSRP will include a wide range of domains (multiple tracks) and thus will need to keep track of what can and cannot be shared and under what conditions. Personally Identifiable Information (PII) or, by fusion, merging multiple datasets that bring PII into the composite result, cannot be shared. In cases where PII data are needed, it is important to determine the feasibility of *data construction*—but the scale may not be as large as it would be for "data in the wild." Recent conferences that have included privacy as a central topic, e.g., SIAM International Conference on Data Mining [56] and some that have focused entirely on this issue (e.g., the Big Data Privacy Workshop [57]).

C. How does one preserve data and software used for data science?

In the field of Natural Language Processing, researchers rely heavily on the University of Pennsylvania's Linguistic Data Consortium (LDC), which collects, creates, annotates, and distributes data, ensuring that all materials are carefully managed, with lawyers verifying copyright and other issues (e.g., licensing). Other organizations serve a similar role as the LDC, but are geared toward more data science, e.g., Research Data Alliance and Data.gov. In addition, NIST is working on data preservation and archival (i.e., keeping bits around forever) and tracing the history of data [58]–[60].

X. Concluding Remarks: IAD's Role for Data Science

This paper lays out the foundation of IAD's newly formed Data Science Research Program and describes IAD's role in the future of the data science discipline. Classes of data science problems and next generation data science challenges as well as areas of important future breakthroughs are discussed. An overview of evaluation and metrology, standards, computing infrastructure needs, and methodologies for measuring effectiveness of data science technologies is presented.

IAD's role for meeting the measurement challenges for data science has four primary facets. These include developing measures for assessment, establishing standards, forming working groups consisting of researchers in the community, and deploying a sound framework for evaluating technology effectiveness.

In addition, IAD aims to build a community of interest within which it is expected that many of the questions posed

in this paper will be addressed. Technical symposia with a focus on generalized problems in data science are expected outcomes of this aspect of NIST's work.

Additionally, it is expected that agile system architectures, system benchmarking tools, and novel approaches will emerge from the development of technologies that are evaluated in the DSE series.

Finally, the DSE series will be organized each year by NIST, in coordination with the data science research community, for the assessment of technologies for big data and data science analytics. NIST will serve the community in providing relevant datasets, metrics, tasks, protocols, and analysis tools.

DISCLAIMER:

These results are not to be construed or represented as endorsements of any participants system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

REFERENCES

[1] M. Smith, "The White House names Dr. D.J. Patil as the first U.S. chief data scientist," Feb 2015. [Online]. Available: https://www.whitehouse.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist

[2] L. Cao, H. Motoda, G. Karypis, and B. Boethals, "DSAA trends and controversies," in *International Conference on Data Science and Advanced Analytics (DSAA)*, Shanghai, 2014.

[3] W. Chang, "1st ISO/IEC JTC 1 study group on big data meeting." [Online]. Available: http://jtc1bigdatasg.nist.gov/

[4] ——, "NIST special publication 1500-6 information technology laboratory: DRAFT NIST big data interoperability framework: Volume 6, reference architecture."

[5] ——, "NIST big data public working group (NBD-PWG) request for public comment," 2015. [Online]. Available: http://bigdatawg.nist.gov/V1_output_docs.php

[6] B. Marr, "Why only one of the 5 Vs of big data really matters," 2015. [Online]. Available: http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters

[7] E. McNulty, "Understanding big data," *Dataconomy*, 2014. [Online]. Available: dataconomy.com/seven-vs-big-data/

[8] D. Laney, "3D data management: Controlling data volume, velocity, variety," Feb. 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

[9] IBM, "The four V's of big data," 2013. [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[10] C. A. Knoblock and P. Szekely, "Exploiting semantics for big data integration," *AI Magazine*, vol. 36, no. 1, pp. 25–38, 2015.

[11] J. Sleeman, T. Finin, and A. Joshi, "Entity type recognition for heterogeneous semantic graphs," in *2013 AAAI Fall Symposium Series*, 2013.

[12] L. Getoor and A. Machanavajjhala, "Entity resolution: theory, practice & open challenges," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2018–2019, 2012.

[13] J. Pujara, H. Miao, L. Getoor, and W. W. Cohen, "Using semantics & statistics to turn data into knowledge," *AI Magazine*, vol. 36, no. 1, pp. 65–74, 2015.

[14] M. Franklin, "Big data and data science: Some hype but real opportunities," University of Florida, Mar. 2015. [Online]. Available: https://www.cise.ufl.edu/content/uf-informatics-institute-inaugural-symposium

[15] S. S. Morrison, R. Pyzh, M. S. Jeon, C. Amaro, F. J. Roig, C. Baker-Austin, J. D. Oliver, and C. J. Gibas, "Impact of analytic provenance in genome analysis," *BMC Genomics*, vol. 15, no. Suppl 8: S1, 2014.

[16] D. George, "Understanding structural and semantic heterogeneity in the context of database schema integration," *Journal of the Department of Computing*, no. 4, 2005.

[17] A. Mittal and A. Goel, "Stock prediction using twitter sentiment analysis," 2011.

[18] D. Doermann, "Visual media forensics: Knowing when seeing is believing," University of Florida, mar 2015. [Online]. Available: https://www.cise.ufl.edu/content/uf-informatics-institute-inaugural-symposium

[19] T. H. Saey, "Big data studies come with replication challenges," *Science News*, vol. 187, no. 3, pp. 22–27, 2015.

[20] S. Few, *Information Dashboard Design: Displaying Data for At-a-glance Monitoring*. Analytics Press, 2013.

[21] D. Suciu, D. Olteanu, C. Ré, and C. Koch, "Probabilistic databases," *Synthesis Lectures on Data Management*, vol. 3, no. 2, pp. 1–180, 2011.

[22] K. E. Foote, "The Geographer's craft project," 2015. [Online]. Available: http://www.colorado.edu/geography/gcraft/contents.html

[23] A. Datta, "Privacy through accountability: A computer science perspective," in *Distributed Computing and Internet Technology*. Springer, 2014, pp. 43–49.

[24] A. Meliou, W. Gatterbauer, and D. Suciu, "Bringing provenance to its full potential using causal reasoning." in *Theory and Practice of Provenance*, 2011.

[25] P. Buneman, S. Khanna, and W.-C. Tan, "Data provenance: Some basic issues," in *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, ser. Lecture Notes in Computer Science, S. Kapoor and S. Prasad, Eds. Springer Berlin Heidelberg, 2000, vol. 1974, pp. 87–93.

[26] L. C. James Cheney and W.-C. Tan, "Provenance in databases: Why, how, and where," *Foundations and Trends in Databases*, vol. 1, no. 4, pp. 379–474, 2007.

[27] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *SIGMOD Rec.*, vol. 34, no. 3, pp. 31–36, Sep. 2005.

[28] S. Finlay, *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*. Palgrave Macmillan, 2014.

[29] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[30] J. Gelernter and K. M. Carley, "Spatiotemporal network analysis and visualization," *International Journal of Applied Geospatial Research*, vol. 6, no. 2, pp. 77–97, 2015.

[31] I. Meirelles, *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport Publishers, 2013.

[32] D. A. Keim, "Information visualization and visual data mining," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, no. 1, pp. 1–8, 2002.

[33] U. Fayyad, A. Wierse, and G. Grinstein, *Information Visualization in Data Mining and Knowledge Discovery*, ser. The Morgan Kaufmann series in data management systems. Morgan Kaufmann, 2002.

[34] C. Li, C. Aggarwal, and J. Wang, "On anonymization of multi-graphs," in *Proceedings of the 2011 SIAM International Conference on Data Mining*, ser. Proceedings. Society for Industrial and Applied Mathematics, Apr. 2011, pp. 711–722.

[35] C.-H. Tai, S. Y. Philip, D.-N. Yang, and M.-S. Chen, "Structural diversity for privacy in publishing social networks," in *Proceedings of the 2011 SIAM International Conference on Data Mining*, B. Liu, H. Liu, C. Clifton, T. Washio, and C. Kamath, Eds. Philadelphia, PA: Society for Industrial and Applied Mathematics, Apr. 2011, pp. 35–46.

[36] S. Yang, K. Kalpakis, and A. Biem, "Detecting road traffic events by coupling multiple timeseries with a nonparametric bayesian method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1936–1946, Oct. 2014.

[37] R. Fagin, L. Haas, M. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis, *Conceptual Modeling: Foundations and Applications*. Springer, 2009.

[38] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, ser. Data-Centric

Systems and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[39] M. Jeevan, "Fundamental methods of data science: Classification, regression and similarity matching," Jan. 2015. [Online]. Available: http://www.kdnuggets.com/2015/01/fundamental-methods-data-science-classification-regression-similarity-matching.html

[40] G. H. Bakir, T. Hofmann, B. Schlkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan, Eds., *Predicting Structured Data (Neural Information Processing)*. The MIT Press, 2007.

[41] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[42] E. Fix and J. Hodges, J. L., "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review / Revue Internationale de Statistique*, no. 3, pp. pp. 238–247, 1957.

[43] B. W. Silverman and M. C. Jones, "An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951)," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. pp. 233–238, 1989.

[44] C. Mayfield, J. Neville, and S. Prabhakar, "A statistical method for integrated data cleaning and imputation," Purdue University, Tech. Rep. 09-008, Sep. 2009.

[45] Y. Bengio, I. J. Goodfellow, and A. Courville, "Deep learning," 2015. [Online]. Available: http://www.iro.umontreal.ca/bengioy/dlbook

[46] N. Katariya, A. Iyer, and S. Sarawagi, "Active evaluation of classifiers on large datasets," in *2013 IEEE 13th International Conference on Data Mining*, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2012, pp. 329–338.

[47] "Text retrieval conference," 2014. [Online]. Available: http://trec.nist.gov

[48] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, 1st ed. New York: Cambridge University Press, Dec. 2002.

[49] O. Talagrand, "Assimilation of observations: An introduction," *Meteorological Society of Japan Series 2*, vol. 75, pp. 81–99, 1997.

[50] C. Ware, *Information Visualization, Third Edition: Perception for Design*, 3rd ed. Waltham, MA: Morgan Kaufmann, Jun. 2012.

[51] B. B. Bederson and B. Shneiderman, *The Craft of Information Visualization: Readings and Reflections*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[52] R. Marty, *Applied security visualization*. Addison-Wesley Upper Saddle River, 2009.

[53] R. Jain, *The Art Of Computer Systems Performance Analysis: Techniques For Experimental Design, Measurement*. John Wiley & Sons, 1991.

[54] J. C. De Kergommeaux, E. Maillet, and J. Vincent, "Monitoring parallel programs for performance tuning in cluster environments," *Parallel Program Development for Cluster Computing: Methodology, Tools and Integrated Environments" book, P. Kacsuk and JC Cunha eds*, 2001.

[55] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[56] M. Zaki, Z. Obradovic, P. N. Tan, A. Banerjee, C. Kamath, and S. Parthasarathy, Eds., *Proceedings of the 2014 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2014.

[57] "Big data privacy workshop: Advancing the state of the art in technology and practice," 2014. [Online]. Available: http://web.mit.edu/bigdata-priv/

[58] W. Allasia, W. Bailer, S. Gordea, and W. Chang, "A novel metadata standard for multimedia preservation," *Proceedings of iPres*, 2014.

[59] W. Chang, "Preliminary digital preservation interoperability framework (dpif) results," in *Archiving Conference*, vol. 2010, no. 1. Society for Imaging Science and Technology, 2010, pp. 202–202.

[60] ——, "Advanced digital image preservation data management architecture," in *Archiving Conference*, vol. 2009, no. 1. Society for Imaging Science and Technology, 2009, pp. 178–182.