CrossMark

A new data science research program: evaluation, metrology, standards, and community outreach

Bonnie J. Dorr[1] · **Craig S. Greenberg**[1] · **Peter Fontana**[1] · **Mark Przybocki**[1] ·
Marion Le Bras[1] · **Cathryn Ploehn**[1] · **Oleg Aulov**[1] · **Martial Michel**[1] ·
E. Jim Golden[1] · **Wo Chang**[1]

Abstract This article examines foundational issues in data science including current challenges, basic research questions, and expected advances, as the basis for a new data science research program (DSRP) and associated data science evaluation (DSE) series, introduced by the National Institute of Standards and Technology (NIST) in the fall of 2015. The DSRP is designed to facilitate and accelerate research progress in the field of data science and consists of four components: evaluation and metrology, standards, compute infrastructure, and community outreach. A key part of the evaluation and measurement component is the DSE. The DSE series aims to address logistical and evaluation design challenges while providing rigorous measurement methods and an emphasis on generalizability rather than domain- and application-specific approaches. Toward that end, each year the DSE will consist of multiple research tracks and will encourage the application of tasks that span these tracks. The evaluations are intended to facilitate research efforts and collaboration, leverage shared infrastructure, and effectively address crosscutting challenges faced by diverse data science communities. Multiple research tracks will be championed by members of the data science community with the goal of enabling rigorous comparison of approaches through common tasks, datasets, metrics, and shared research challenges. The tracks will permit us to measure several different data science technologies in a wide range of fields and will address computing infrastructure, standards for an interoperability framework, and domain-specific examples. This article also summarizes lessons learned from the data science evaluation series pre-pilot that was held in fall of 2015.

This article extends a paper that was presented at IEEE Data Science and Advanced Analytics conference in the fall of 2015[1] and also expands upon content from a poster paper at IEEE BigData 2015 [2].

✉ Oleg Aulov
oleg.aulov@nist.gov; aulov.oleg@gmail.com

Bonnie J. Dorr
bonnie.dorr@nist.gov

[1] 100 Bureau Drive, Mail Stop 8940, Gaithersburg, MD 20899, USA

1 Introduction

Since its emergence as a uniquely identifiable field, *data science* has been of growing importance, attested to by a proliferation of government initiatives, research conferences, and academic data science initiatives and institutes. Government initiatives over the last several years include the Defense Advanced Research Projects Agency (DARPA) XDATA Program, the National Science Foundation's 2012 Big Data solicitation of $10 million, the National Institutes of Health's recruitment of an associate director for Data Science, and last year's new White House appointment of the first US Chief Data Scientist [3].

There are now many research conferences focused on data science, such as Association for Computing Machinery's

(ACM) 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.

On the academic front, data science initiatives have been emerging at a rapid rate, including Columbia University's Data Science Institute announced in July of 2012, University of California, Berkeley's announcement of the first online Master of Information and Data Science degree in 2013, a new Center for Data Science established at University of Massachusetts Amherst in 2015, and the initiation of a new Data Science major at University of Michigan in the fall of 2015.

In any rapidly emerging field, there is a pressing need to explore the foundational issues associated with that field, and data science is no exception. Indeed, the "Trends and Controversies" presented at the Data Science and Advanced Analytics conference in both 2014 and 2015 [4] raised a range of data science challenges, research questions, and expected advances [1].

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, aims to address many of these issues. The DSRP is designed to facilitate and accelerate research progress in the field of data science. Within this program, *data science* is viewed as the application of techniques for analysis (data analytics) 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—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 crosscutting challenges. Several of these challenges are described in this article.

As a multi-year research program, the DSRP is expected to change and grow over time, tracking the maturation of the field and technology. In order to address this need, the DSRP examines a set of representative classes of data science

NIST
Meeting the measurement challenges of data science
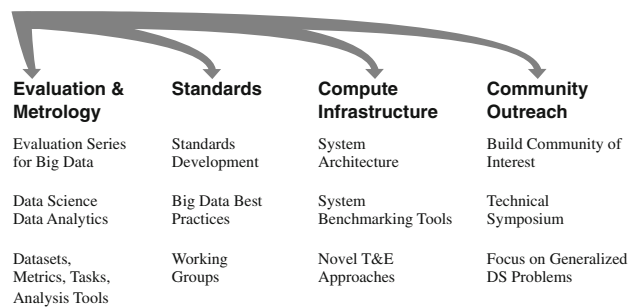
Fig. 1 NIST's role in addressing data science challenges

problems (discussed in Sect. 2) and explores different aspects of data science measurement (discussed in Sect. 3).

Four key elements, illustrated in Fig. 1, will be identified and outlined:

– *Evaluation and metrology* Design and conduct a new international *data science evaluation (DSE)* series (Sect. 4.1).
– *Standards* Leverage prior work to develop standards for data science (Sect. 4.2).
– *Compute infrastructure* Develop an evaluation management system (EMS) to inform compute and infrastructure needs (Sect. 4.3).
– *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 (Sect. 4.4).

In Sect. 5, we explain the concept of evaluation-driven research as it pertains to data science and present several examples of areas where evaluation-driven research was successful. Section 6 describes in detail NIST's data science evaluation (DSE) series (that was outlined in Sect. 4.1) that will host annual evaluations, and describes its structure and development. Section 7 describes the DSE Pre-Pilot that was held in the fall of 2015, details its tasks, datasets, and metrics, and reports on lessons learned from it.

This article is not focused on the development of novel algorithms or specific methodologies or solutions, but rather the discussion focuses on 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 (Sect. 8) and presents a summary of the next generation of data science challenges (Sect. 9). The final section delivers concluding remarks regarding NIST's role in the discipline of data science.

2 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 DSE that covers each category.

Detection Detection aims to find data of interest in a given dataset. In the traffic domain, incidents are of interest, e.g., "traffic incident detection" is an important subproblem of the traffic use case. Yang et al. [5] analyze traffic flow in order to detect traffic incidents.

Anomaly detection Anomaly detection is the identification of system states that force additional pattern classes into a model. Relatedly, outlier detection is associated with identifying potentially erroneous data items that force changes in prediction models ("influential observations"). For example, through anomaly detection in health insurance claim records, potentially fraudulent claims can be identified. 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 [6].

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 raw traffic sensor data.

Alignment Alignment refers to the process of relating different instances of the same object [7], e.g., a word with the corresponding visual object, or time stamps associated with two different time series. Data alignment is frequently used for entity resolution, which is identifying common entities among different data sources. Getoor and Machanavajjhala [8] and Christen [9], for example, describe entity resolution techniques. 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 [10]. 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 attempt to determine, for each item of interest, the type or class to which the item belongs [11]. 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 (signaling the presence of bumper-to-bumper traffic).

Regression Regression refers to the process of finding functional relationships between variables. In the traffic pilot, the posed challenge might be to model the traffic flow rate as a function of other variables.

Prediction Prediction refers to the estimation of a variable or multiple variables of interest at future times. In the pilot traffic flow prediction challenge, the participants are asked to predict traffic speed using covariates including flow volume, percentage occupancy, and training sets of past multivariate time series.

Structured prediction Structured prediction refers to tasks where the outputs are structured objects, rather than numeric values [12,13]. 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)[1]. In the traffic use case, a database of incidents and accidents could be constructed from news reports, time-stamped global positioning system (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 [14,15]. 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 subproblems using constraints that enforce global consistency. Joint inference may be used for detection and cleaning to arrive at more accurate results [16]. 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. [17]. Additional classes

---

[1] 4th workshop on automated knowledge base construction www.akbc.ws/2014/.

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, for example, for benchmarking. The next section elaborates on the range of methodologies needed for measuring technology effectiveness within the new DSRP.

3 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.

How does one measure accuracy when all the 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 real-time datasets or future data materialize (e.g., accident prediction in video). For non-predictive tasks (e.g., detection of traffic incidents), see, for example, Katariya et al.'s [18] work on active evaluation of classifiers estimates accuracy based on a small labeled set and human labeler. Some NIST evaluations (e.g., TREC [19]) 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. Within the DSRP, these and other approaches for addressing issues concerning ground truth metadata will be explored.

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 [20]. 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 [21].

How does one measure data fusion? Data Fusion, as defined by the Department of Defense, is "a multilevel, multifaceted process dealing with the registration, detection, association, correlation, and combination of data and information from multiple sources to achieve [a] refined state and identity estimation, and complete and timely assessments of (the) situation" [22]. Application of data fusion is prevalent in many different areas such as vehicular traffic, geospatial information systems, business intelligence, and bioinformatics. For instance, Joshi, et al. [23] demonstrated a frugal traffic state sensing approach using data fusion of acoustic, image and sensor information. Within the DSE, data fusion is assumed to be central to data science measurement because of the key role it plays in combining data from different modalities into a unified consistent representation for analysis purposes.

How does one measure knowledge representation through Visualization of data? The visualization analytics science and technology community has developed a "VAST Challenge," run annually for the past 3 years,[2] 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 [24, 25]. Visualization paradigms are often assessed by the number of data points and the level of granularity represented [26] and by types of relationships that can be represented [27].

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 [28,29]. 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.

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

---

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

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 [30]. 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.

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 the DSE pilot 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." Two of the recent research venues that have included privacy as a central topic are SIAM[3] International Conference on Data Mining [31] and the Big Data Privacy Workshop [32]).

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)[4], which collects, creates, annotates, and distributes data, ensuring that all materials are carefully managed, with lawyers verifying copyright, licensing, and other issues. Other organizations serve a similar role as the LDC, but are geared more toward data science, such as 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 [33–35].

4 Key elements of the new data science research program

The four pillars of the new DSRP—illustrated earlier in Fig. 1—are described in more detail below.

4.1 Evaluation and metrology for data science

NIST has been conducting evaluations of data-centric technologies since the late 1980s. 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 (DSE) series (described in detail in in Sect. 6), 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, where the evaluation is the common thread that ties these communities together.

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

---

[3] Society for Industrial and Applied Mathematics.

[4] More information on LDC can be found on their Web site at: www.ldc.upenn.edu/about.

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, the Information Access Division of NIST hosted a small-scale pre-pilot evaluation in the highway traffic domain, meant to serve as a track archetype, and to surface any unexpected evaluation challenges. This track is not meant to solve any particular problem in the traffic domain, but rather to serve as an exemplar of a data science evaluation track. It consisted of heterogeneous data from traffic and weather sensors and featured data cleaning, dataset alignment, and predictive analytics tasks (as described further in Sect. 7). In 2016, NIST is following up with an open pilot evaluation in the same domain and will begin accepting track proposals for a 2017 full-scale data science evaluation.

4.2 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 [36]. 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 [37]. This framework allows data scientists to process and derive knowledge through the use of a standard interface between swappable architectural components. The following elements 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 [38]. 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 characteristics: technology, infrastructure, and vendor-agnostic capability; (2) definition of general interfaces between the reference architecture components; and (3) validation of the reference architecture by building applications through the general interfaces.

4.3 Compute infrastructure for data science research

NIST has implemented an 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 testing of different compute paradigms (software and model changes, such as testing a solution using MPI[5] and later trying it using Go[6] channels) as well as hardware accelerations in order to best assess how a given evaluation should be run.

The underlying cloud infrastructure accommodates concurrent execution of projects—such as experiments or evaluation—on shared hardware while being able to separate data access, network resources, users, and hardware accelerators (e.g., GPU[7] or Intel Xeon 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.

---

[5] Message Passing Interface.

[6] More information available at www.golang.org.

[7] Graphics Processing Unit.

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 accelerators 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, Nvidia Tesla GPUs, Intel Phi Coprocessors, high memory compute nodes, high storage data nodes) using a local OpenStack[8] 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 the local (on premise) cloud as the volume of data increases. 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[9] HTTP[10] API. All objects are stored with multiple copies and are replicated in as-unique-as-possible availability zones and/or regions.

The current test bed for the EMS has Gigabit as well as an Infiniband networks, five compute nodes with 16 cores each, 128, 192 or 256 GB of memory, and 32 or 48 TB of disk storage per node, as well as two extra compute nodes with four Nvidia Tesla C2050 and four Xeon Phi 5100, and five storage nodes with 48 TB of disk storage 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.[11]

4.4 Data science community building and 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 crosscutting 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)[12], which has been held at NIST annually since 1992. This initiative 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.

---

[8] www.openstack.org.

[9] REpresentational State Transfer software architectural style.

[10] HyperText Transfer Protocol.

[11] For more information on these technologies see: mesos.apache.org, www.docker.com, kubernetes.io, spark.apache.org, hadoop.apache.org.

[12] More information about TREC is available at trec.nist.gov.

As a starting point, in March of 2014, NIST held the first Data Science Symposium,[13] 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).

5 Evaluation-driven research as it pertains to data science

This section describes the concept of evaluation-driven research (EDR) as the motivation behind DSE series, followed by description of prior work with examples of evaluations from other domains.

5.1 Evaluation-driven research

It is often the case that researchers working on the same problem do so in relative isolation, each defining the problem in their own way, while using self-made datasets and disparate metrics for measuring their approach's success [39,40]. In such circumstances, research results are difficult to compare and the community is somewhat disjointed in its efforts; this impedes research progress. EDR is an approach to research that addresses these inefficiencies and helps accelerate progress. In order to do so, EDR provides a framework in which a neutral organizer brings researchers together by providing an evaluation, which consists of common and well-defined tasks, data, metrics, and measurement methods. This focuses effort and enables the research community to more

easily compare and contrast approaches. Moreover, EDR also provides the requisite data and measurement infrastructure for conducting research, which reduces the amount of necessary overhead.

EDR has successfully spurred research progress in numerous technologies, as described in Sect. 5.2, and this paper offers two examples in detail. Reynolds [39] tells the story of how EDR benefited automatic speaker recognition research, the task of which is to determine whether two audio recordings were spoken by the same person. In the mid-1990s, there were numerous technical approaches to speaker recognition, but the presence of multiple task formulations, datasets, and metrics made it difficult to compare approaches and to measure the state of the art. In 1996, NIST hosted the first in a series of evaluations for automatic speaker recognition technology, which focused the community on a common task, dataset, and metric. As a result, the most promising approaches to speaker recognition were quickly identified and have been steadily improved upon over time [41].

EDR has also benefited research in machine translation (MT), along with related technologies such as optical character recognition (OCR). Prior to the inception of NIST's annual MT evaluation series in 2001 (OpenMT [42,43]), evaluation was not a very powerful tool for MT research [40], because it required human judgments, making it very expensive and time consuming. Shortly after the start of OpenMT, a major transformation was brought about by the method of bilingual evaluation understudy (BLEU) [44] and other automatic evaluation paradigms, all of which relied on n-gram co-occurrence statistics. Similar approaches became central to the assessment of OCR progress, and metrics such as word error rate (WER) were used in NIST's evaluations for OCR of handwriting starting in 2010 (OpenHART [45]). These transformations in evaluation paradigms have enabled rapid optimization for experimentation with a variety of algorithms and have paved the way for human-in-the-loop measures, such as human-mediated translation edit rate (HTER) [46], that transcend purely automatic measures in evaluating effectiveness of machine translation and optical character recognition.

The process for EDR can be divided into four steps:

1. *Planning*, including defining the task and research objectives for the evaluation. It should be noted that only so many objectives can be pursued at once; it is therefore essential to choose objectives that will substantially improve the technology while being challenging yet obtainable in the near term. Receiving community input during this step is critical.
2. *Data and experiment design*, involving the development of datasets and associated tasks for experimentation. For example, in machine learning, data are typically partitioned into training, development, and evaluation

---
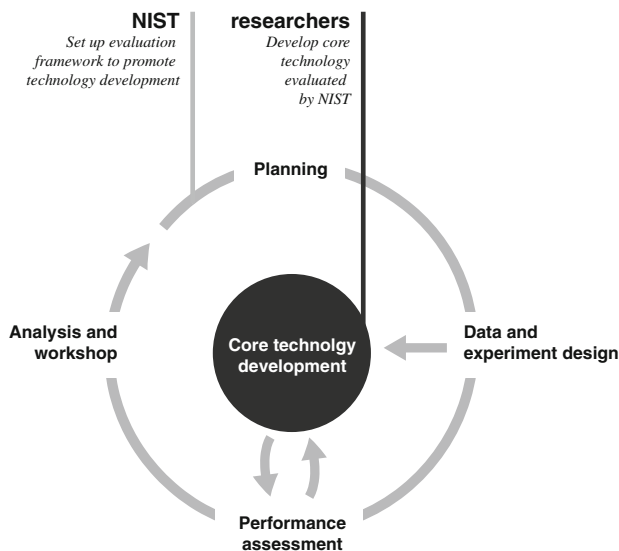
Fig. 2 The evaluation-driven research cycle

datasets, and an example of a possible experiment would be to contrast performance using different training datasets. Rigorously designing experiments and datasets is significantly easier when the data to be used were created for the evaluation (as opposed to being repurposed), though data collection design and implementation has its own challenges (for example, see [47]).

3. *Performance assessment*, during which systems are run on the data. In some evaluations, data are sent to researchers, who run their systems locally and then submit their system output. In other evaluations, the systems themselves are submitted and are run at NIST. The latter approach is more involved, requiring an agreed upon API and ability of every system to run on a prescribed computational infrastructure, though it is better suited for evaluations using very large or sensitive datasets. Once system output is generated, both NIST and the evaluation participants analyze the experimental results.

4. *Workshop*, where the research community gathers to openly discuss evaluation outcomes, including which approaches were attempted and the degree to which they were successful, as well as other lessons learned. A crucial portion of the workshop is a discussion of future challenges and objectives, which feeds into the planning of the next evaluation. Beyond the workshop, evaluation results are published more broadly.

These four steps naturally form a cycle; in particular, the planning for an evaluation takes place in part at the workshop of the previous evaluation. See Fig. 2 for an illustration.

Progress is driven in EDR by repeating the evaluation cycle, and as technology improves, increasing the challenge of the research objectives, which are then addressed in subse-

quent evaluations (see Fig. 3). After the technology reaches a point appropriate for a given application, engineering for speed and other considerations takes place and the technology is deployed for the application. The evaluation cycle continues, driving more technological progress to enable transfer to more demanding applications. It is worth noting that NIST's roles in data-centric technology transfer are typically focused on the relatively early and late stages of the process, i.e., core technology development and standards, respectively.

5.2 Examples of work related to evaluation-driven research

There are a number of current and prior evaluations that address some of the challenges present in the new DSE series. Representative examples of prior work that addresses individual challenges are summarized below, as well as in our previous work [1]. Dorr et al. [1] discuss the DSRP, of which the DSE is a key component. A shorter preliminary discussion of the DSE that highlights the evaluation-driven research process is [2].

Evaluating Systems containing Visualizations One important challenge in data science evaluation concerns information visualization for a variety of datasets. A number of different approaches have been developed, as discussed by Bederson and Shneiderman [25], Korsar [48], Lam et al. [49], and Isenberg at al. [50]. Visualization paradigms are often assessed heuristically using best design practices as guidelines. For example, a visualization could be evaluated by the number of data points and the level of granularity represented [26] and by types of relationships that can be presented [27].

Empirical studies are also employed to measure visualization in terms of human performance. Many methods used to evaluate visualizations follow conventions set by the human–computer interaction field. The methods also depend on the goals of the evaluation, such as a focus on the visualization itself or a focus on the analytic process the visualization supports [49].

Isenberg et al. [50] assert that the visualization community borrows from two disciplines in evaluating visualizations: science evaluation and design evaluation. The scientific approach focuses on building a representational model of the problem space that is reproducible, whereas the design approach focuses on the goal of aiding the user through maximizing usability and functionality.

The visualization analytics science and technology (VAST) community has sought to develop a visualization assessment approach within the "VAST Challenge," which has run annually since 2012 [51–53]. This includes assessment of visual analytics applications for both large-scale situation analysis and cyber security. Challenges addressed in VAST include

Fig. 3 Annual evaluations address barriers in order to advance core technology
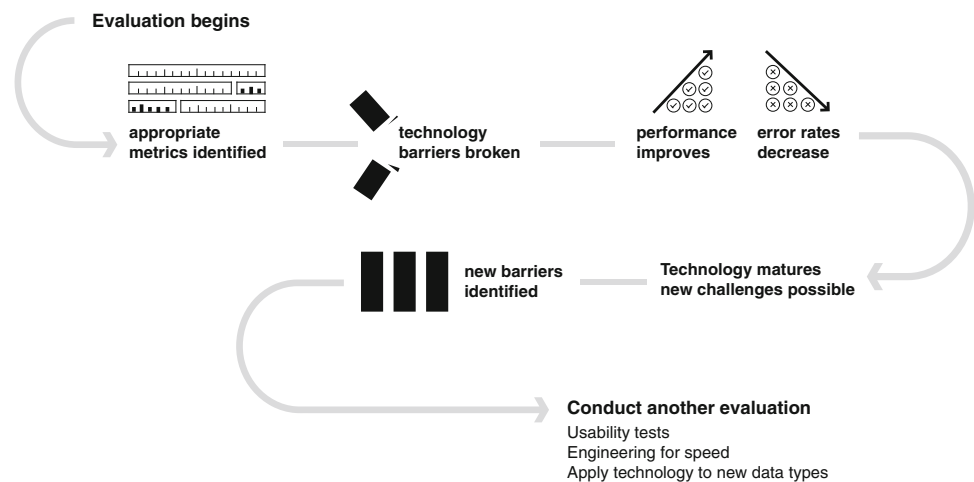
automatic graph analysis and best practices for combined and multi-modal datasets.

Evaluating Systems on Very Large DataSets A central challenge of several ongoing NIST evaluations is being able to conduct evaluations that use very large datasets, such as Text REtrieval Conference [19]. The first TREC was developed in 1992 [54] with the goal of assessing information retrieval (IR) techniques on a realistically sized test collection in a forum where all researchers were subject to the same data conditions and evaluation techniques. Researchers from both academia and industry—in large part from within the Special Interest Group on Information Retrieval (SIGIR) community and the DARPA TIPSTER program [55]—participated in the first TREC. In 2000, a special track was introduced that was devoted to video that followed this same model; this track evaluated automatic segmentation, indexing, and content-based retrieval of digital video and, within 2 years, developed into its own evaluation conference (TRECVid) [56,57], which also accommodates the use of large datasets for evaluation.

Evaluating Systems without Ground Truth Evaluation without ground truth or lacking a single ground truth is an additional challenge for the assessment of data-intensive systems. Some NIST evaluations (TREC [19]) apply accuracy measures that accommodate a lack of full truth data, often employing mediated adjudication approaches (e.g., pooling relevance assessments of participants in the evaluation to approximate recall). Other NIST evaluations (MT [42]) are faced with the lack of a single ground truth, generally applying metrics that accommodate more than one "human answer" for computing a score, e.g., BLEU [44] and TER [58]. Multiple approaches to mitigating missing ground truth data have been attempted, for example, "active evaluation," i.e., obtaining a small number of ground truth labels that are most informative in estimating a system's performance (see [18,59]), among others [60].

Evaluating Systems containing Multiple Components Another challenge to be tackled within data science evaluation is that of multi-component assessment—including the impact of one technology on the performance of a downstream technology. As further explained in Sect. 5.1 below, OpenHART applied evaluation metrics such as WER to assess the performance of handwriting OCR, but the impact of OCR's performance was also assessed within the MT application, using metrics such as BLEU and TER. In the most recent OpenHART evaluation [61], evaluation metrics were applied at the image recognition level (given the image and its text line segmentation, recognize the foreign language text in the image), the image translation level (given the image and its text line segmentation, provide an accurate and fluent translation of the foreign language text in the image), and at the text translation level (given the ground truth foreign language text in the image, provide an accurate and fluent translation of the foreign language text in the image).

Some of the data science evaluation challenges presented above have been addressed to one degree or another in various evaluation forums. However, to the best of the authors' knowledge, none have attempted to address all of the data science challenges presented in this paper—or even a large number of those challenges. The DSE series requires careful consideration of all such challenges, coupled with a rigorous framework for assessment and comparison among different approaches.

Prior Traffic Research In addition to addressing general evaluation challenges related to data science, there have been many prior applications and research initiatives focused on specific problems in the traffic domain, e.g., systems to alert users to traffic issues. Traffic alert systems generally rely on either sensor and camera data (e.g., [62,63]) or on crowdsourcing from vehicle operators (e.g., Waze [64]) and Illinois Traffic Alert System [65]. For example, Google Maps is purported to utilize cell phone sensors and historical data to

generate its traffic predictions. More recently, systems for traffic alerts have been developed that use social media input, e.g., Twitraffic in the UK [66] and research systems developed at University of Maryland, Baltimore County [67,68] and at Khalifa University [69].

These earlier efforts generally focus on one type of input, e.g., social media but not sensors, or sensors but not imagery— rather than on the combination of data from multiple sources (imagery, traffic reports, and weather data). Moreover, prior research in the traffic domain focuses on incident detection, i.e., identifying content of traffic-related tweets, not on prediction of upcoming traffic incidents and slowdowns based on historic data on local traffic and weather.

The traffic use case described in this paper brings together data diversity and data volume, while also enabling the development and assessment of technologies for *both* detection and prediction.

6 Data science evaluation series

The NIST data science evaluation (DSE) series will host annual evaluations for data analytic technologies applied to various domains and arranged into workflows that may include human-in-the-loop and visualization components. These technologies operate on some combination of large and small, structured and unstructured, and complete and incomplete data, coming from multiple sources and in various data modalities. In the context of the DSE series and in this paper, data analytic technologies refer to technologies that transform data into information. Some examples include automatic speech recognition, image classification, and video event detection technologies. A data analytic workflow consists of a combination of these technologies. These combinations could be applied in sequence, jointly or in parallel, and possibly include human-in-the-loop and visualization components. The workflow that includes a human-in-the-loop component is one where a human receives output from a system and transforms it and that transformed output becomes input to a subsequent system component [70] (it is worth noting that adding interactive visualization components to the data workflow may increase the performance of data analytic tasks [71]). Data analytic workflows have found use in network administration [72], auditing medical claims [73], stock market analysis [74], and biomedical information analysis [75], to name only a few applications and domains. The long-term goals of the DSE series are to drive data analytic technology forward, to measure the state of the art, to develop effective measurement techniques, and to encourage promising algorithmic approaches. The authors adopt the view that an evaluation capable of achieving these goals would necessarily involve the following actions:

1. Address key data analytic technology challenges. Several challenges for data analytic technologies, each with a representative set of examples, are presented in Table 3.
2. Provide a general framework that can be applied irrespective of data types or domains.
3. Apply the above framework to various domains and use cases.
4. Build interdisciplinary collaboration that enables a cross-pollination of ideas and methods, encouraging technology developers to understand the unique challenges of each domain, and exposing domain experts to a wide array of state-of-the-art approaches.

In order to create an evaluation with the above properties, several challenges need to be addressed. These challenges fall broadly into four categories:

1. *Logistical challenges* Conducting very large-scale evaluations poses several problems, including the creation and distribution of appropriate datasets, as well as the computational resources needed for systems to process the data and for the results to be analyzed. In addition, evaluation of multi-component workflows requires the coordination and timing of individual component evaluations.
2. *Evaluation design challenges* The DSE will often make use of "found" data rather than data collected in a controlled manner for the purpose of evaluation. This increases the difficulty of conducting rigorous experiments. The desire to include several domains in the evaluation gives rise to the need to design evaluations in domains without necessarily having access to expert-level domain knowledge.
3. *Measurement challenges* It will be necessary to be able to effectively measure performance in cases where hand-annotated data are difficult to obtain or no single ground truth exists. In order to measure performance of workflows and the impact of various data sources, it will be necessary to be able to measure and calibrate the propagation of error and uncertainty throughout the workflow. Measuring the effectiveness of visualization also poses unique and difficult challenges; see, for example [48–50,76,77].
4. *Organizational challenges* 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.

6.1 Evaluation structure

The DSE series will consist of regularly scheduled evaluations, expected to recur annually. Each evaluation in the series will consist of several tracks, where a track is made up

Fig. 4 Overview of the data science evaluation series
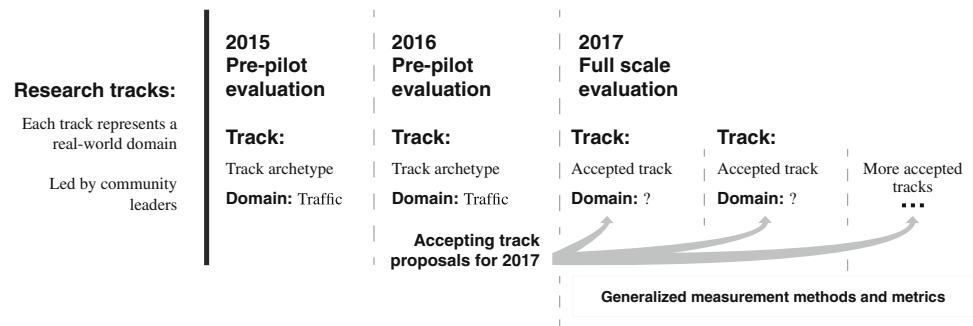
of challenge problems set in a given domain. In addition to NIST-hosted tracks, the DSE series will include community-championed tracks. NIST will solicit the community for track proposals, and those tracks included in the evaluation will be planned, organized, and implemented by a "track champion" from within the community.

The primary motivations for hosting a multi-track evaluation series are to encourage the development of technologies that are effective across different domains, to increase the breadth of domains and technologies that can be evaluated in a timely manner, and to facilitate the cross-pollination of ideas by hosting diverse communities under a single effort.

Figure 4 illustrates an evaluation structured with tracks in different domains, where each track shares some technology elements with other domains. Researchers will be encouraged to work on any or all parts of a track (i.e., multiple tasks in a single domain) as well as across tracks (i.e., a single set of tasks in multiple domains).

Including tracks championed by members of the community rather than limiting the DSE series to NIST-championed tracks offers several advantages. For example, each track can be led by experts in the track's domain and/or data modalities, more tracks can be included in a single evaluation than could possibly be run by NIST alone, and other organizations with existing or future data science challenges will be able to bring their research problems directly to the community, while benefiting from the infrastructure already put in place for the evaluation series.

It is worth noting that this approach (a multi-track evaluation with community track coordinators) has been successfully utilized in other NIST-hosted evaluation series, e.g., TREC [78], often to great effect [79].

6.2 Evaluation series development

The DSE series will be developed in three stages, as illustrated in Fig. 4. In 2015, IAD ran a pre-pilot evaluation that consisted of a single track with a traffic prediction use case (described in further detail in Sect. 7 below). IAD will undertake a pilot evaluation that will extend the pre-pilot evaluation track and be open to all who wish to participate, and an inaugural evaluation, consisting of multiple community-led

evaluation tracks in different domains and use cases. This sequence will enable NIST to immediately begin providing infrastructure for addressing data science research challenges and to rapidly evolve the evaluation to be more effective when generalized to multiple domains and tracks.

Each stage in the DSE series has a different set of objectives: The pre-pilot exercised the evaluation metrics and measurement methods as well as surface unexpected evaluation challenges. The pilot will improve upon the pre-pilot and serve as an exemplar for future DSE tracks. The inaugural evaluation will introduce multiple diverse domains and use cases. This diversity will enable NIST to begin generalizing the measurement methods and metrics in the DSE series and help ensure that the technology developed is applicable across domains and types of data. The pilot, pre-pilot, and inaugural evaluation will each conclude with a workshop, focused on examining the evaluation results and sharing approaches and lessons learned. The evaluation pre-pilot took place in the fall of 2015. In 2016, NIST will host the evaluation pilot and will begin accepting track proposals for a 2017 full-scale data science evaluation.

7 Data science evaluation pre-pilot

The data science evaluation pre-pilot is the first stage in the development of the DSE series; its primary goal is to develop and exercise the evaluation process in the context of data science. The pre-pilot consisted of data and tasks set in the traffic domain—a domain chosen due to its relevance to everyday life of the general public and its accessibility and availability of large amounts of public data. It is important to note that the pre-pilot is not meant to solve any particular problem in the traffic domain. The objective is for the developed measurement methods and techniques to apply to additional use cases, regardless of the domain and data characteristics.

The tasks included in the evaluation pre-pilot consist of data cleaning, data alignment, and predictive analytics. These tasks were chosen to highlight a variety of data science challenges. Table 1 lists some of the key data science challenges from Table 3 and describes how these may appear in the traffic prediction context.

For logistical reasons, only a subset of these challenges was addressed in the pre-pilot, namely data heterogeneity,

Table 1 Data science evaluation challenges for traffic use case

Challenge	Traffic use case
Provenance	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	A vehicle may be represented visually in video data and descriptively in an incident report
Predictive analytics	Future traffic patterns may be guessed from weather, imagery, and historical traffic data
Knowledge assimilation	A traffic accident may be detected from the position of two cars in a video clip
Big data replicability	Using historical data from weather reports, traffic incident data, and traffic video data to detect an incident may yield different results
Visualization of information	Visualization may be used to communicate traffic flow and accidents
Data uncertainty	Uncertainty may arise from the lack of data available from points that occur between traffic detectors
Mitigating error propagation	Errors associated with cleaning of traffic incident reports may propagate to incident detection and traffic prediction tasks

predictive analytics, and data uncertainty.[14] The handling of large datasets, cross-track interaction, and the visualization of information will be introduced during the pilot or inaugural evaluation.

7.1 Data

Several datasets were made available as part of the pre-pilot. These datasets come from multiple sources, many of which are public, and consist of different data modalities, including numeric data, numeric data with short textual fields, and video data. See Table 2 for more details.

Although the tasks focus on traffic prediction, the available data are not restricted solely to traffic information; weather, US Census data, and video data are also included. This gives participants the ability to integrate rich and diverse data in order to address the evaluation tasks.

The original source data are used as ground truth for the evaluation pre-pilot. In order to support the data cleaning task, errors were introduced to these data. Test data typically consist of a subset of the data, either hiding key records or key fields of specific records.

7.2 Tasks

There are four tasks in the pre-pilot evaluation set in the traffic use case. Each task is listed below with the full task name followed by a one-word abbreviation.

1. *Cleaning* finding and eliminating errors in dirty traffic detector data.
2. *Alignment* relating different representations of the same traffic event in video and numeric data to find events of interest.
3. *Prediction* estimating the likelihood of various traffic related events.
4. *Forecasting* estimating future values in a time series of traffic flows.

In order to create a data analytic workflow, the tasks are ordered such that the cleaning task could be completed first; the output of the cleaning task would then be used as input to the other tasks. This order, if followed, allows NIST to measure the influence of the cleanliness of the traffic flow data on prediction accuracy.

When describing this workflow, the term "dirty data" is used to refer to the traffic detector data participants are being asked to clean; "cleaned data" to refer to the output traffic flow data from the cleaning task; and "truth data" to refer to the ground truth data, which is the correct answer to the cleaning task. This workflow is shown in the "Flow" section of Fig. 5 and is broken into two phases:

– *Phase 1* Participants were given the dirty data. They were asked to submit system outputs from the four tasks using the dirty data as input. Additionally, participants were asked to submit the results of the alignment, incident, and flow tasks using the cleaned traffic detector data.
– *Phase 2* Participants following the previously described pipelining were given the cleaning task truth data and were asked to run the same systems for the alignment, incident, and flow tasks, using the dirty data input with the truth data.

---

[14] The "Managing Privacy and Security" challenge from Table 3 is omitted from Table 1. Due to the minimally restricted (or unrestricted) nature of traffic and weather data, this is not a focus of the traffic-related pre-pilot but is expected to be addressed for domains involving personally identifiable information (e.g., health IT).

Table 2 Summary of available datasets

Data type	Data subset	Description
Lane detector	Lane detector inventory	List of all traffic lane detectors as a comma-separated values (CSV) file. Each detector is uniquely identified by its `lane_id` value, and each detector inventory gives the location of the detector (in decimal latitude and longitude coordinates), the source organization for the measurements of those detectors, and other relevant information
	Lane detector measurements	Measurements from traffic sensors in locations in the DC Metro area and the Baltimore area. Traffic sensors are placed on both directions of the highways, in each lane. Lane and zone (multiple lanes of the same road going in the same direction) data are provided. The measurements are the following: (1) *Volume* the number of cars to have passed through the lane detector since the last measurement; (2) *Speed* the average vehicle speed since the last measurement; (3) *Occupancy* the average percent of time a vehicle was in front of the detector since the last measurement; and (4) *Quality* a data quality field
Traffic events	Traffic event instances	A traffic event is defined as a situation that involves traffic which includes accidents, construction, lane closures, hazardous weather, and special events. Each traffic event listing contains the following fields (among others): (1) Description; (2) Location, both in formatted text (the intersection) and in decimal latitude and longitude; (3) Times the event was created, confirmed and closed; and (4) The type and subtype of the traffic event, labeled with the fields `event_type` and `event_subtype`
Traffic camera video	Camera inventory	A list of all traffic cameras with their locations, described both in text (the intersection) and in decimal latitude and longitude
	Camera video feeds	Consecutive 15-min video segments from traffic cameras in Maryland with start times. The traffic cameras may be remotely operated by humans, who can rotate the camera and zoom, which happens when the human operator chooses to look at a traffic situation. Some cameras may have watermarks indicating the direction the camera is facing (E for east, SW for southwest, etc.), or the current time
U.S. Census	2010 US Census	Publicly available information including population counts; age, income, and occupation demographics; and household demographics in summary files and public use microdata sample (PUMS)
	American community survey (ACS)	A more frequent survey providing statistics on transportation and commutes, such as the average commute length, the percentage of people who carpool, and the percentage of people who use public transportation. There are 1-, 3-, and 5-year surveys as summary Files and PUMS, like the US Census Data
OpenStreetMap	[No subset]	Map data from from OpenStreetMap, describing the road network in the DC-MD-VA area as well as locations including airports, public transportation stations, and buildings that host large events. These maps also support lookup by latitude and longitude coordinates
Weather	Integrated surface (ISD)	A dataset of measurements from weather stations in the DC-MD-VA area with a variable number of measurements. Measurements include station information, temperature, air pressure, weather condition, precipitation, and other elements. The ISD set is quality controlled. The quality control does not state that it is free of errors or missing data, only that others have looked at it to try to improve the quality of the data. Lott [80] discusses the quality control process that is used in the ISD to check for formatting errors and outliers
	Severe weather data inventory (SWDI)	A compilation of many types of severe weather, including storms, hail, tornados, and lighting strikes

Participants were encouraged to submit system output for multiple tasks, but the pipelining of data from one task into another was not required. It was also anticipated that many participants would perform additional data cleaning beyond that required for the cleaning tasks; such additional data processing was welcome and would be described when submitting a system.

7.3 Metrics

Each task was scored with specified metrics, as appropriate for each task. For each task, participants were provided with scores that measured the discrepancy between the system outputs and the ground truth data. Additionally, when appropriate, Detection Error Tradeoff (DET) curves were

Table 3 Next-generation challenges in the field of data science

Challenge	Relevant questions	Examples
Provenance	Where did the raw data originate? Is it current? What processes were applied through which the data were derived from its original sources?	A genome sequence dataset may be recreated from raw data and the provenance records associated with genomic annotations [89]
Data heterogeneity	How does one integrate data from multiple large heterogeneous datasets with varying degrees of noise? What is the relative value of a given dataset for a particular analytic task?	A publisher may be represented either as a publication-producing entity or as an attribute of a publication [90], and these representations originate from multiple sources
Predictive analytics	How can trends be identified and distinguished from random fluctuation in order to provide a calibrated forecast of future values?	Stock market events may be forecasted from sentiments expressed in social media [91]
Knowledge assimilation	How might algorithms understand new data, e.g., inferring causality from the data?	Fraudulent activity may be inferred from (potentially altered) digital and physical data representations of known entities and events [92]
Big data replicability	How does one consistently reproduce experimental findings given that truth may be hard to find?	Using the same (usually massive) genomic dataset in two different studies to find genetic contributions to a particular disease may yield different results [93]
Visualization of information	What is the most effective way to visualize information for decision making by a potentially diverse set of people?	Cybersecurity systems can utilize dashboards to reflect network status and to alert security administrators of suspicious activity [26]
Data uncertainty	How does one handle gaps in knowledge due to the potential for untrustworthy or inaccurate data?	In radio frequency identification (RFID) data management, raw antenna readings are frequently missed or tags are read by two adjacent antennas [94]
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 [95]
Managing privacy and security	How does one 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 [96]

Fig. 5 The pre-pilot evaluation task workflow
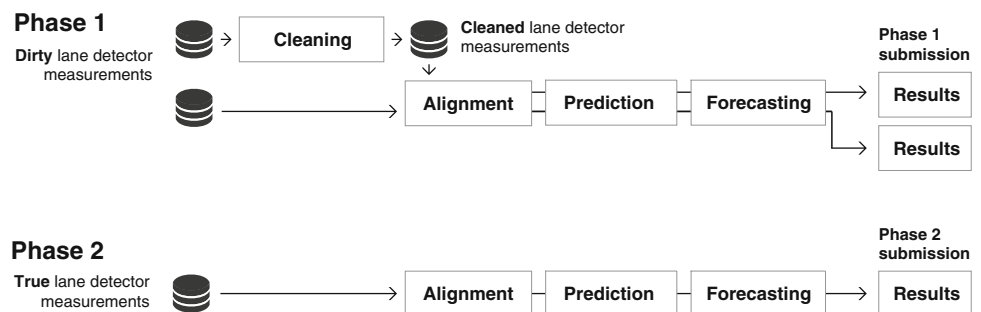
used to illustrate the tradeoff between misses and false alarms. See Martin et al. [81] for more information on DET curves.

In addition, by comparing the results of systems based on whether the dirty data, cleaned data, or truth data (as discussed in Sect. 7.2) was used as input, the differences in the metric values for the same systems on the same tasks give a way of measuring error propagation from the cleaning task to the other tasks. NIST plans to expand this initial measuring of error propagation in the pilot and the inaugural evaluation.

7.4 Lessons learned from the data science evaluation series pre-pilot

In fall 2015, NIST hosted a small-scale pre-pilot evaluation in the highway traffic domain. The motivation for having a small-scale pre-pilot was multi-fold: to establish the evaluation series with a methodology and metrics to form and test tasks, to receive feedback from participants before the launch of the larger-scale pilot, as well as to promote participation in the pilot. One of the main goals was to exercise the mechanics of the evaluation, later to be the possible foun-

dation of a traffic track. While obtaining results was not the focus of the pre-pilot, achieving them serves to demonstrate that the evaluation tasks that were created with the methodology described in this paper are now tried and tested and can be successful in applying evaluation-driven research to the field of data science. A number of lessons were learned that will enable springboarding to the next level with the pilot evaluation and full evaluation. A small number of participants were invited, and the end result was a total of 10 submissions across all tasks.

In the area of outreach, even with very targeted solicitations, the pre-pilot participants were brought in by invitation only, to overcome the start-up challenges of recruiting participants into a newly inaugurated series. Many potential participants showed interest early on, but many of these were focused on the traffic prediction track rather than on the topic of data science in general. It has become increasingly clear through this experience that having multiple challenge problems in multiple domains will provide more opportunities to attract participants.

Regarding algorithmic and metrology lessons learned, the evaluation resulted in several participants focusing on the specific domain of *traffic* rather than on the development of algorithms that are more generalized. The focused interest in traffic prediction yields two important insights. First, in order to encourage people to participate, it is necessary to have challenge problems and data sets of interest to the community. Second, in order to evaluate data science more broadly, tracks in multiple domains must be developed and evaluated. Doing so provides the basis for evaluation approaches that are likely to be generalized to new domains, as opposed to those that are focused on a particular challenge area or that involve only a subset of the data science community at large.

Finally, it was very clear from this initial pre-pilot that infrastructure will be a key component of both the upcoming pilot and later full-scale evaluations. The EMS that will serve as the core for the DSE series will be leveraged for easier application of algorithms to data science problems without requiring such significant overhead of engineering entire systems.

8 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. Although a fifth V has been proposed as Value [82] (i.e., the degree to which the worth of the data is conveyed), providing a means to visualize data can increase understandability and acces-

sibility 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 crosscut several data science challenges, most notably a sixth V proposed by McNulty [83] (Visualization), which is addressed separately as a next-generation challenge. The earliest formulation by Douglas Laney [84] 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 and formats coming from multiple sources.

Veracity is a fourth V, attributed to researchers at IBM [85]:

– 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 real-time datasets (including issues of latency) and for new tracks involving very large data that one might find, for example, 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* [10,86], is central to building Web-scale systems for tasks such as entity resolution [8,87]. 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. A seventh V that has been proposed is Variability [83] and is 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). In the scope of DSRP, Variability is considered 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. The emergence of data science and the challenges associated with the four V's above are accompanied by technological progress leading to:

– Massively scalable processing and storage.
– Pay-as-you-go processing and storage.
– Flexible schema on read vs. schema on write[15]
– Easier integration of data retrieval and analysis.
– Well-supported interfaces between various domain-specific languages/tools.
– Open-source ecosystem during innovation[16]
– Increased bandwidth, network access, speed, and reduced latency.

This list of areas in which technological progress has been made is an augmented version of those presented recently by Franklin [88]. 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 Future work: 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 3 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 Sect. 4.1. 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 [97]? 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 [98,99]. More generally, data provenance involves being able to determine where the data came from and the

processes through which these data were derived from its original sources [100].

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 [90].

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 [101]? Further, is it possible to have an effective tradeoff 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 [97,102]? 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 are lots of room for error [93]? 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 [103]), each with their own constraints in the amount and type of data to be displayed [27]. The integration of visualization into data science activities aids in the analysis of vast volumes of information [104], may increase efficiency [105], and may reduce user errors [106].

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.

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 [95].

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 [107] and

---

[15] 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.

[16] 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 reuse of applications and components.

structural diversification for protecting privacy [108] 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 crosscutting issues pertaining to the data itself and measures associated with approaches to processing data.

10 Concluding remarks: NIST's role for data science

This paper lays out the foundation of NIST's newly formed data science research program and describes much of NIST's proposed 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.

NIST'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.

NIST also 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.

The DSE series that was described in detail in this paper seeks to address challenges in data science evaluation, and like those faced by the technologies supported by the evaluation, these challenges are numerous and varied. The pre-pilot, pilot, and inaugural evaluation will serve as first steps; however, a successful DSE series must go further. Future plans for the DSE series include the introduction of complex and dynamic workflows, which will assist in quantifying the propagation of error and uncertainty and understanding the effect of reordering components in a workflow; the inclusion of data that must be protected, for example, due to privacy concerns; system run time and resource utilization benchmarking, which will assist in quantifying the trade-off between system run time and accuracy.

While the challenges are substantial, we believe that overcoming them has enormous potential for impact. To the same or greater extent as other technologies have benefited from evaluation-driven research, there is opportunity for data science research to benefit from the DSE series.

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.

References

1. Dorr, B.J., Greenberg, C.S., Fontana, P., Przybocki, M., Le Bras, M., Ploehn, C., Aulov, O., Michel, M., Golden, E.J., Chang, W.: The NIST data science initiative, In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10. IEEE (2015)
2. Dorr, B., Greenberg, C., Fontana, P., Przybocki, M., Le Bras, M., Ploehn, C., Aulov, O., Chang, W.: The NIST IAD data science evaluation series: part of the NIST information access division data science research program. In: Proceedings of IEEE BigData 2015, pp. 2575–2577. IEEE, Santa Clara, CA (2015)
3. Smith, M.: The White House names Dr. D.J. Patil as the first U.S. chief data scientist. www.whitehouse.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist (2015)
4. Cao, L., Motoda, H., Karypis, G., Boethals, B.: DSAA trends and controversies. In: International Conference on Data Science and Advanced Analytics (DSAA). Shanghai (2014)
5. Yang, S., Kalpakis, K., Biem, A.: Detecting road traffic events by coupling multiple timeseries with a nonparametric bayesian method. IEEE Trans. Intell. Transp. Syst. **15**(5), 1936 (2014). doi:10.1109/TITS.2014.2305334
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. (CSUR) **41**(3), 15 (2009)
7. Fagin, R., Haas, L., Hernández, M., Miller, R.J., Popa, L., Velegrakis, Y.: Conceptual Modeling: Foundations and Applications. Springer, New York (2009)
8. Getoor, L., Machanavajjhala, A.: Entity resolution: theory, practice & open challenges. Proc. VLDB Endow. **5**(12), 2018 (2012)
9. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Data-Centric Systems and Applications. Springer, Berlin (2012)
10. Sleeman, J., Finin, T., Joshi, A.: Entity type recognition for heterogeneous semantic graphs. In: 2013 AAAI Fall Symposium Series (2013)
11. Jeevan, M.: Fundamental methods of data science: Classification, regression and similarity matching. http://www.kdnuggets.com/2015/01/fundamental-methods-data-science-classification-regression-similarity-matching.html (2015)
12. Bakir, G.N., Hofmann, T., Schlkopf, B., Smola, A.J., Taskar, B., Vishwanathan, S.V.N. (eds.): Predicting Structured Data (Neural Information Processing). The MIT Press, Cambridge (2007)
13. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)

14. Fix, E., Hodges, J.J.L.: Discriminatory analysis. Nonparametric discrimination: consistency properties. Stat. Rev. Rev. Int. Stat. **57**(3), 238 (1989)

15. Silverman, B.W., Jones, M.C.: An important contribution to nonparametric discriminant analysis and density estimation: commentary on Fix and Hodges (1951). Int. Stat. Rev. Rev. Int. Stat. **57**(3), 233 (1989)

16. Mayfield, C., Neville, J., Prabhakar, S.: A statistical method for integrated data cleaning and imputation. Technical Report 09-008, Purdue University (2009)

17. Bengio, Y., Goodfellow, I.J., Courville A.: Deep learning. http://www.iro.umontreal.ca/bengioy/dlbook (2015)

18. Katariya, N., Iyer, A., Sarawagi, S.: Active evaluation of classifiers on large datasets. In: 2013 IEEE 13th International Conference on Data Mining, vol. 0 , pp. 329–338. IEEE Computer Society, Los Alamitos, CA, USA (2012). doi:10.1109/ICDM.2012.161

19. Text retrieval conference. http://trec.nist.gov (2014)

20. Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability, 1st edn. Cambridge University Press, New York (2002)

21. Talagrand, O.: Assimilation of observations: an introduction. Meteorol Soc Jpn Ser **2**(75), 81 (1997)

22. Waltz, E., Llinas, J. et al.: Multisensor data fusion, vol. 685. Artech house Boston (1990)

23. Joshi, V., Rajamani, N., Katsuki, T., Prathapaneni, N., Subramaniam, L.V.: Information fusion based learning for frugal traffic state sensing. IJCAI. Citeseer (2013)

24. Ware, C.: Information Visualization, Third Edition: Perception for Design, 3rd edn. Morgan Kaufmann, Waltham (2012)

25. Bederson, B.B., Shneiderman, B.: The Craft of Information Visualization: Readings and Reflections. Morgan Kaufmann Publishers Inc., San Francisco (2003)

26. José Cardoso, C., Kacsuk, P.: Parallel Program development for cluster computing: methodology, tools and integrated environments. Vol. 5. Nova Publishers, Commack, NY, USA (2001)

27. Meirelles, I.: Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations. Rockport Publishers, Beverly (2013)

28. Jain, R.: The Art Of Computer Systems Performance Analysis: Techniques For Experimental Design, Measurement. Wiley, Hoboken (1991)

29. De Kergommeaux, J.C., Maillet, E., Vincent, J.: Monitoring parallel programs for performance tuning in cluster environments. In: Kacsuk, P., Cunha, J.C. (eds.) Parallel Program Development for Cluster Computing: Methodology, Tools and Integrated Environments (2001)

30. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157 (2003)

31. Zaki, M., Obradovic, Z., Tan, P.N., Banerjee, A., Kamath, C., Parthasarathy S. (eds.): In: Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics (2014)

32. Big data privacy workshop: Advancing the state of the art in technology and practice. http://web.mit.edu/bigdata-priv/ (2014)

33. Allasia, W., Bailer, W., Gordea, S., Chang, W.: A novel metadata standard for multimedia preservation. In: Proceedings of iPres (2014)

34. Chang, W.: Preliminary digital preservation interoperability framework (dpif) results. In: Archiving Conference, vol. 2010, pp. 202–202. Society for Imaging, Science and Technology (2010)

35. Chang, W.: Advanced digital image preservation data management architecture. In: Archiving Conference, vol. 2009, pp. 178–182 Society for Imaging, Science and Technology (2009)

36. Chang, W.: 1st ISO/IEC JTC 1 study group on big data meeting. http://jtc1bigdatasg.nist.gov/

37. Chang, W.: NIST special publication 1500-6 information technology laboratory: DRAFT NIST big data interoperability framework: volume 6, reference architecture. NIST, Gaithersburg, MD (2015)

38. Chang, W.: NIST big data public working group (NBDPWG) request for public comment. http://bigdatawg.nist.gov/V1_output_docs.php (2015)

39. Reynolds, D.: Speaker and language recognition: a guided safari. Keynote Speech Odyssey (2008). Accessed 15 Sept 2015

40. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138–145 (2002)

41. Przybocki, M., Martin, A.: NIST speaker recognition evaluation chronicles. Comput. Speech Lang. **20**(23), 15 (2006)

42. NIST open machine translation evaluation. http://nist.gov/itl/iad/mig/openmt15.cfm (2015)

43. NIST open machine translation evaluation. http://www.itl.nist.gov/iad/mig/tests/mt/ (2001)

44. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Association for Computational Linguistics (2002)

45. NIST open handwriting recognition and translation evaluation (OpenHaRT). http://www.nist.gov/itl/iad/mig/hart.cfm (2010)

46. Dorr, B.J., Olive, J., McCary, J., Christanson, C.: Chapter 5: machine translation evaluation and optimization. In: Olive, J., Christianson, C., McCary, J. (eds.) Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation, Springer, New York, pp. 745–843 (2011)

47. Gallagher, K., Stanley, A., Shearer, D., Klerman, L.V.: Challenges in data collection, analysis, and distribution of information in community coalition demonstration projects. J. Adolesc. Health **37**(3), S53 (2005)

48. Korsar, R., Healey, C., Interrante, V., Laidlaw, D., Ware, C.: Thoughts on user studies: why, how, and when. Comput. Graph. Appl. **23**(4), 20 (2003)

49. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: seven scenarios. IEEE Trans. Vis. Comput. Graph. **18**(9), 1520 (2012)

50. Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., Moller, T.: A systematic review on the practice of evaluating visualization. IEEE Trans. Vis. Comput. Graph. **19**(12), 2818 (2013)

51. VAST challenge 2012. http://vacommunity.org/VAST+Challenge+2012 (2012)

52. VAST challenge 2013. http://vacommunity.org/VAST+Challenge+2013 (2013)

53. VAST challenge 2014. http://vacommunity.org/VAST+Challenge+2013 (2014)

54. Harman, D.: Overview of the first text retrieval conference. In: Proceedings of the 16th Annual International ACM/SIGIR Conference, pp. 36–48 (1993)

55. Harman, D.: The darpa tipster project. SIGIR Forum **26**(2), 26 (1993)

56. NIST TRECVid surveillance event detection evaluation. http://nist.gov/itl/iad/mig/sed.cfm (2015)

57. NIST TRECVid multimedia event detection evaluation. http://nist.gov/itl/iad/mig/med.cfm (2015)

58. Snover, M., Dorr, B.J., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas. http://www.cs.umd.edu/~snover/tercom/ (2006)

59. Sawade, C., Landwehr, N., Bickel, S., Scheffer, T.: Active risk estimation. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 951–958 (2010)

60. Zafarani, R., Liu, H.: Evaluation without ground truth in social media research. Commun. ACM **58**(6), 54 (2015)

61. NIST open handwriting recognition and translation evaluation (OpenHaRT). http://www.nist.gov/itl/iad/mig/hart.cfm (2013)

62. Yang, S., Kalpakis, K.: Detecting road traffic events by coupling multiple timeseries with a nonparametric bayesian method. IEEE Trans. Intell. Transp. Syst. **15**, 1936 (2014)

63. Yang, S., Kalpakis, K., Biem, A.: Spatio-temporal coupled bayesian robust principal component analysis for road traffic event detection. In: 16th International IEEE Conference on Intelligent Transportation Systems-(ITSC), pp. 392–398. IEEE (2013)

64. Waze. https://www.waze.com/ (2015)

65. Illinois traffic alert system. http://www.iltrafficalert.com/ (2015)

66. Twitraffic in uk. http://twitraffic.co.uk/ (2015)

67. Wibisono, A., Sina, I., Ihsannuddin, M.A., Hafizh, A., Hardjono, B., Nurhadiyatna, A., Jatmiko, W., Mursanto, D.P.: Traffic intelligent system architecture based on social media information. In: International Conference on Advanced Computer Science and Information Systems (ICACSIS). Depok, Indonesia (2012)

68. Sakaki, T., Matsuo, Y., Yanagihara, T., Chandrasiri, N.P., Nawa, K.: Real-time event extraction for driving information from social sensors. In: Proceedings of the IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems. Bangkok, Thailand (2012)

69. Wang, D., Al-Rubaie, A., Davies, J., Clarke, S.S.: Traffic intelligent system architecture based on social media information. In: IEEE Symposium on Evolving and Autonomous Learning Systems (EALS). Orlando, FL (2014)

70. Liu, J., Wilson, A., Gunning, D.: Workflow-based human-in-the-loop data analytics. In: Proceedings of the 2014 Workshop on Human Centered Big Data Research, p. 49. ACM (2014)

71. Chen, M., Floridi, L., Borgo, R.: What is visualization really for? In: Floridi, L., Phyllis, I. (eds.) The Philosophy of Information Quality, vol. 358 pp. 75–93. Springer, Cham (ZG), Switzerland (2014)

72. Kidder, K.L., Haring, J.M., Bishop, R.J., Trent, J.D., Pham, L.D.: System for automated workflow in a network management and operations system. US Patent 6,445,774 (2002)

73. The power of combining big data analytics with business process workflow. CGI Whitepaper, Montreal, Quebec, Canada (2013)

74. Rostoker, C., Wagner, A., Hoos, H.: A parallel workflow for real-time correlation and clustering of high-frequency stock market data. In: Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International, pp. 1–10. IEEE (2007)

75. Yu, H., Qingwei, X., Bin, H., Jianyong, W.: An integrative software system for biomedical information analysis workflow. In: BioMedical Information Engineering, 2009. FBIE 2009. International Conference on Future, pp. 61–64. IEEE (2009)

76. Bederson, B.B., Shneiderman, B.: The Craft of Information Visualization: Readings and Reflections. Morgan Kaufmann, Burlington (2003)

77. Ware, C.: Information Visualization: Perception for Design. Elsevier, Amsterdam (2012)

78. Text REtrieval Conference. http://trec.nist.gov (2015)

79. Tassey, G., Rowe, B.R., Wood, D.W., Link, A.N., Simoni, D.A.: Economic impact assessment of NIST's Text REtrieval conference (TREC) program, Report prepared for National Institute of Technology (NIST) (2010)

80. Lott, J.N.: The quality control of the integrated surface hourly database. In: 84th American Meteorological Society Annual Meeting, vol. 7.8. American Meterological Society, Seattle, WA. http://www1.ncdc.noaa.gov/pub/data/inventories/ish-qc.pdf (2004)

81. Martin, A.F., Doddington, G.R., Kamm, T., Ordowski, M., Przybocki, M.A.: The DET curve in assessment of detection task performance. In: Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997. Rhodes, Greece. http://www.isca-speech.org/archive/eurospeech_1997/e97_1895.html (1997)

82. Marr, B.: Why only one of the 5 Vs of big data really matters. http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters (2015)

83. McNulty, E.: Understanding big data, dataconomy. http://dataconomy.com/seven-vs-big-data/ (2014)

84. Laney, D.: 3D data management: Controlling data volume, velocity, variety. http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/ (2001)

85. IBM. The four V's of big data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data (2013)

86. Knoblock, C.A., Szekely, P.: Exploiting semantics for big data integration. AI Mag. **36**(1), 25 (2015)

87. Pujara, J., Miao, H., Getoor, L., Cohen, W.W.: Using semantics & statistics to turn data into knowledge. AI Mag. **36**(1), 65 (2015)

88. Franklin, M.: Big data and data science: some hype but real opportunities. https://www.cise.ufl.edu/content/uf-informatics-institute-inaugural-symposium (2015)

89. Morrison, S.S., Pyzh, R., Jeon, M.S., Amaro, C., Roig, F.J., Baker-Austin, C., Oliver, J.D., Gibas, C.J.: Impact of analytic provenance in genome analysis. BMC Genomics **15**(Suppl 8: S1), 1 (2014)

90. George, D.: Understanding structural and semantic heterogeneity in the context of database schema integration. J. Dep. Comput. **4**, 29 (2005)

91. Mittal, A., Goel, A.: Stock prediction using twitter sentiment analysis Standford University, CS229. http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf (2012)

92. Doermann, D.: Visual media forensics: knowing when seeing is believing. https://www.cise.ufl.edu/content/uf-informatics-institute-inaugural-symposium (2015)

93. Saey, T.H.: Big data studies come with replication challenges. Sci. News **187**(3), 22 (2015)

94. Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic databases. Synth. Lect. Data Manag. **3**(2), 1 (2011)

95. Foote, K.E.: The geographer's craft project. http://www.colorado.edu/geography/gcraft/contents.html (2015)

96. Datta, A.: Privacy through accountability: A computer science perspective. In: International Conference on Distributed Computing and Internet Technology, pp. 43–49. Springer, Bhubaneswar, India (2014)

97. Meliou, A., Gatterbauer, W., Suciu, D.: Bringing provenance to its full potential using causal reasoning. TaPP, Crete, Greece (2011)

98. Buneman, P., Khanna, S., Tan, W.C.: Data provenance: some basic issues. In: Kapoor, S., Prasad S. (eds.) FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science. Lecture Notes in Computer Science, vol. 1974, pp. 87–93. Springer, Berlin (2000). doi:10.1007/3-540-44450-5_6

99. James Cheney, L.C., Tan, W.C.: Provenance in databases: why, how, and where. Found. Trends Databases **1**(4), 379 (2007). doi:10.1561/1900000006

100. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. SIGMOD Rec. **34**(3), 31 (2005). doi:10.1145/1084805.1084812

101. Finlay, S.: Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods. Palgrave Macmillan, London (2014)

102. Pearl, J.: Causal inference in statistics: an overview. Stat. Surv. **3**, 96 (2009). doi:10.1214/09-SS057

103. Gelernter, J., Carley, K.M.: Spatiotemporal network analysis and visualization. Int. J. Appl. Geospatial Res. **6**(2), 77 (2015). doi:10.4018/ijagr.2015040105

104. Keim, D.A.: Information visualization and visual data mining. IEEE Trans. Vis. Comput. Graph. **8**(1), 1 (2002). doi:10.1109/2945.981847

105. Fayyad, U., Wierse, A., Grinstein, G.: Information Visualization in Data Mining and Knowledge Discovery. The Morgan Kaufmann series in data management systems (Morgan Kaufmann, 2002)

106. Few, S.: Information Dashboard Design: Displaying Data for At-a-glance Monitoring. Analytics Press, Burlingame (2013)

107. Li, C., Aggarwal, C., Wang, J.: On anonymization of multi-graphs. In: Proceedings of the 2011 SIAM International Conference on Data Mining, Proceedings, pp. 711–722. Society for Industrial and Applied Mathematics (2011)

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