# Evaluation-Driven Research in Data Science: Leveraging Cross-Field Methodologies

Bonnie J. Dorr, Peter C. Fontana, Craig S. Greenberg, Marion Le Bras, Mark Przybocki,
National Institute of Standards and Technology
{bonnie.dorr,peter.fontana,craig.greenberg,marion.lebras,mark.przybocki}@nist.gov

***Abstract*—While prior evaluation methodologies for data-science research have focused on efficient and effective teamwork on independent data science problems *within* given fields [1], this paper argues that an *enriched* notion of evaluation-driven research (EDR) supports methodologies and effective solutions to data-science problems *across* multiple fields. We adopt the view that progress in data-science research is enriched through the examination of a range of problems in many different areas (traffic, healthcare, finance, sports, etc.) and through the development of methodologies and evaluation paradigms that span diverse disciplines, domains, problems, and tasks. A number of questions arise when one considers the multiplicity of data science fields and the potential for cross-disciplinary "sharing" of methodologies, for example: the feasibility of generalizing problems, tasks, and metrics across domains; ground-truth considerations for different types of problems; issues related to data uncertainty in different fields; and the feasibility of enabling cross-field cooperation to encourage diversity of solutions. We posit that addressing the problems inherent in such questions provides a foundation for EDR across diverse fields. We ground our conclusions and insights in a brief preliminary study developed within the Information Access Division of the National Institute of Standards and Technology as a part of a new Data Science Research Program (DSRP). The DSRP focuses on this cross-disciplinary notion of EDR and includes a new Data Science Evaluation series to facilitate research collaboration, to leverage shared technology and infrastructure, and to further build and strengthen the data-science community.**

## I. Introduction

The Information Access Division (IAD) of the National Institute of Standards and Technology (NIST) has developed a new Data Science Research Program (DSRP) [2], [3]. While prior evaluation methodologies for data-science research have focused on efficient and effective teamwork on independent data-science problems *within* given fields [1], the DSRP focuses on *enriched* evaluation-driven research (EDR) to support methodologies and effective solutions to data-science problems *across* multiple fields. The motivation for this shift is the need to share common solutions and metrics, while avoiding reformulation of solutions to data-science problems in one discipline that are applicable to problems in another seemingly distinct discipline. The DSRP relies on the development of domain-independent solutions, and thus is well positioned for examination of a range of problems and solutions in many different fields, such as traffic, healthcare, finance, sports, etc.[1]

In this program, data science is viewed as the application of techniques for analysis and extraction of knowledge from potentially massive data. Data science includes notions of big data technical challenges in distributed and parallel processing, as well as considerations and insights that might arise even with smaller datasets. This paper aims to cover methodological questions for data science more broadly, thus subsuming issues inherent in big data challenges.

In this position paper, we adopt the view that progress in data-science research is enabled through this cross-field examination and through the development of methodologies for transferring knowledge of approaches, solutions, measures, and evaluation paradigms across those diverse disciplines, problems, and tasks. Taking this approach yields a significantly enriched notion of EDR beyond that of prior work through the cross-disciplinary sharing of ideas across fields and the discovery of solutions that otherwise would not have been apparent within a given field.

Toward that end, we have developed a new Data Science Evaluation (DSE) series—central to the DSRP—within which evaluations are expected to recur annually. The series consists of several *tracks*, where a track is made up of problems set in a given field of data science. Each track is planned, organized, and implemented by a "track coordinator," either a NIST data scientist or a non-NIST expert in the field of interest.

The first phase of the DSE series was completed in spring 2016: a small-scale pre-pilot evaluation was conducted by NIST that consisted of a single track with a traffic prediction use case in the automotive domain. In the fall of 2016, a pilot evaluation will take place that extends the pre-pilot evaluation track and is open to all who wish to participate. The pilot evaluation is designed to pave the way for a successful and informative full-scale evaluation encompassing multiple disciplines.

Toward that end, the goal for 2017 is to develop an inaugural evaluation, consisting of multiple evaluation tracks in different domains and use cases—championed by experts in different fields. As a first step toward this goal, the upcoming 2016 pilot evaluation encompasses these objectives:

- further develop and exercise the evaluation process at NIST in the context of data science,
- provide participants the opportunity to exercise the evaluation process prior to participating in larger-scale evaluation,

[1] This paper draws from these four domains for representative examples of data-science solutions; however, many other domains lend themselves to data-science solutions: weather, biology, law, ecology, economics, business, security, medical informatics, social sciences, humanities, and several others.

TABLE I
FOUNDATIONAL QUESTIONS FOR ENRICHED EDR

A. **Classes of Problems**: What kinds of measurable problems, techniques, and algorithms generalize across domains?
B. **Tasks**: What is an appropriate series of tasks on which data researchers in different fields may want to test their approaches, either at the task-component level or at a higher end-to-end level?
C. **Methods and Metrics**: What kinds of methods and metrics generalize to problems across domains? How effective is the generalization and what can be done to make algorithms and methods more domain-independent?
D. **Ground Truth**: Are there ground-truth considerations common to different areas of data science? Are there effective techniques to handle the lack of or limited ground truth in different both for solving problems and for evaluating methods and metrics?
E. **Data Uncertainty**: What approaches to handling gaps and inconsistencies in data are applicable to multiple domains?
F. **Community Cooperation**: Is it possible to foster cross-field cooperation for sharing of solutions?
G. **Diversity of Solutions**: Can cross-field synergies yield diverse solutions within a given field?

- serve as an archetype for the development of future evaluation tasks, datasets, and metrics,
- establish baseline performance measurements,
- identify new measurement methods and techniques that might be applied to a broad range of use cases, regardless of data type and structure.

It is the last objective above that is the central theme of this position paper. Most notably, although the pilot evaluation is set in the automotive domain, it is expected that many of the algorithms and techniques to be evaluated (as well as the evaluation approaches and metrics themselves) will generalize to other domains. This theme is fundamental to an enhanced notion of EDR where methodologies, approaches, solutions, measures, and evaluation paradigms are considered in new domains, with the expectation that many of these will be of significant utility across diverse fields.

A framework for *enriched* EDR can be characterized as a set of answers to questions such as those shown in Table I. We argue that addressing the issues inherent in such questions provides a foundation for EDR across diverse fields and thus enables research progress on many shared problems and tasks in different domains.

The next section presents related work in evaluation-driven research and generalizability across domains. Following this, Section III makes a case for concrete steps toward this generalizability and explores an enriched notion of EDR through examination of each of the questions above in turn. In Section V, we ground our conclusions and insights in a brief preliminary study developed within the DSE for evaluation-driven research aimed at strengthening the data-science community. Concluding remarks are provided in Section VI.

## II. BACKGROUND

The earliest seeds for EDR were sown in the 1970's, when George H. Heilmeier—while serving as Director of the Defense Advanced Research Projects Agency (DARPA)—developed the *Heilmeier Catechism*. (This was published two decades later (as [4]) and was subsequently reviewed by others, e.g., [5].) In this work, several questions related to innovation and novelty were posed, but also a set of evaluation-driven questions were posed along the following lines [4, pg. 15]: "What difference will it make? What are the midterm and final "exams" to check for success?"

These questions are foundational to EDR and have become the basic tenets for many evaluation campaigns across different disciplines. Moreover, these questions have driven progress in many different areas of research. EDR has been around for a long while (more on that below), it has served us well for decades, and it is not likely to go away anytime soon.

The implication is that progress in research is heavily guided by goals related to evaluation. What is often overlooked, however, is an aspect related to the cross-disciplinary (enriched) notion of EDR espoused in this paper. Interestingly, Heilmeier's 1992 article [4] included several important *lessons* that were relevant to this enriched notion, but that (at the time) received less attention than the oft-cited questions in the Heilmeier Catechism. For example, the following statement from the article is relevant [4, pg. 14]: "Approach problems from an interdisciplinary point of view. Remove the barriers to exploiting the viewpoints of other disciplines, and do not be afraid to be called naïve when venturing outside your own professional discipline."

In short, while experts in diverse fields such as those enumerated in Section I might consider their own problems to be unique, it may be the case—more often than not—that data-science solutions and metrics used in one field may be applicable to problems in another field. Before exploring this enriched notion of EDR, it is worthwhile to examine the history leading up to the point where research progress began to be driven in large part by discipline-wide evaluations.

Falling in line with Helmeier, speech researchers experienced a paradigm shift in the 1980's, where evaluations served to push research forward [6], [7]. The DARPA TIPSTER speech evaluation involved evaluation methodologies designed to support focused research, to establish momentum, to maintain continuity, and to encourage longevity, while pushing forward the state of the art. However, evaluation methodologies and metrics spanning domains and disciplines were generally unheard of back in these early days; generally each field had its own evaluation (e.g., speech recognition).

Following this paradigm shift, EDR began to serve as the basis for multiple evaluations by the National Institute for Standards and Technology (NIST), most notably in Information Retrieval, as evaluated in the Text REtrieval Conference [8]. This gave rise to several discipline-wide evaluations with well-defined tasks, data, metrics, and measurement methods (see, e.g., [3] for a more in-depth discussion). Most notably, EDR has successfully spurred research progress in automatic speaker recognition research [9], [10], machine translation [11], and optical character recognition [12].

The associated development of benchmarks has yielded the application of an evaluation methodology to different types of inputs for a given technology. For example, Word Error

Rate (WER) has been applied to evaluate effectiveness of speech recognition output for varying levels of "informality" (i.e., different genres) of speech inputs [7].[2] However, the application was always "speech transcription" and the metric was always WER. If the application domain were not speech, would this same evaluation paradigm be applicable?

Certainly within the fields of natural language processing and document processing, there are other problems to which WER has been applied: machine translation (often modified to take into account *meaning* such as Human-targeted Translation Error Rate (HTER) [13]) and optical character recognition (often applied in conjunction with character error rate [14]). This is a good start—but what about generalizing to non-language data, such as non-language medical or financial data?

We adopt the view that generalizations of approaches and measures are *often* possible across disciplines upon closer inspection. For example, in many fields, experts are focused on predicting a series of ordered consequences. When an input sequence is mapped into an output sequence for this type of prediction task—regardless of the field—a metric akin to WER could be adapted for a particular problem and, moreover, the underlying technology that implements a given solution might itself be shared across fields. In such cases, the way that the technology is both implemented and evaluated would benefit from cross-disciplinary sharing.

Another clear case of a large-scale data science evaluation that supports EDR is Kaggle (https://www.kaggle.com/)—a forum for hosting Data Science Competitions that consists of a wide range of "challenges" in multiple domains [15]. Although EDR is central to the design of these periodic evaluations, the missing aspect is that of generalizability: each competition is run (mostly) independently, with little or no synergy across the different problem areas, tasks, and and also no shared methodologies (unless by accident). Kaggle does include a means for employing metrics across domains, which is an important step toward generalizability; our aim is to further enrich the notion of EDR through cross-domain sharing of algorithms for analogous tasks.

The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) [16] is yet another example of a large-scale evaluation that is aligned with EDR. The main mission of CLEF is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure. CLEF has had a very significant scholarly impact, as evidenced by the emergence of research directions that otherwise would not have been possible and also by measures of the scholarly impact of the research fostered by benchmarking activities within CLEF.

The overarching field of study for participants in CLEF is information retrieval, but the application areas are often broad, encompassing numerous disciplines. As a forum that supports "Evaluation Campaigns" CLEF has paved the way for providing a basis for multi-disciplinary sharing that might be leveraged for an enriched notion of EDR as described in this paper. Concrete steps toward bringing about sharing of techniques and metrics—such as those discussed in the next section—might improve progress on CLEF problems both within and across evaluation campaigns.

Several other evaluation-related concepts and paradigms are related to EDR and, thus, would potentially benefit from multidisciplinary participation and enriched EDR. For example, Developmental Evaluation (DE) [17] leverages evaluation to support development and to focus on long-term, continuous improvement. This domain-independent methodology applies to many fields. Relatedly, in software engineering, the concept of a *benchmark*—a common framework for people within a discipline to discuss and compare solutions—can provide EDR benefits. Sim et al. [18] argue that benchmarking advances research by providing a setting where researchers can focus their attention on key problems. Sharing of benchmarking approaches across disciplines provides another opportunity to generalize evaluation methodologies in ways that may lead to EDR enrichment. One such framework is the use of common data sets to compare systems and algorithms. SPEC [19] and the UCI Machine learning repository [20] are two such collections of data sets.

The next section discusses concrete steps toward generalizability of concepts and paradigms related to evaluation and further develops the notion of enriched EDR.

## III. An Enriched Notion of Evaluation-Driven Research for Data Science

Prior work [3] conceives of EDR for data science as a notion that is divided up into four steps that are more programmatic than foundational in nature: planning tasks and research objectives for evaluation; design of data and experiments; performance assessment; and provision of a forum to discuss evaluation outcomes. These steps are crucial to the development of a framework within which to *practice* EDR, but they do not translate directly into the fundamentals of how to *implement* various aspects pertaining to EDR. For example, these programmatic steps remain agnostic with respect to how one studies different problems, tasks, and domains in data science in a way that leverages evaluation methodologies that might potentially span multiple disciplines.

*Enriched* EDR is achieved when a solution in an existing field is applied successfully in a new field for which such a solution had not been previously imagined. As alluded to above, data science crucially spans a diverse set of disciplines and domains [21]. Leveraging the breadth of disciplines investigated by data scientists provides the basis for strengthening EDR to the advantage of each of the individual fields.

To date, EDR has driven progress with an evaluation cycle that is repeated at regular intervals and, as technology improves, the research objectives are made more challenging on each subsequent cycle. Within data science, this repeated cycle has the potential for further enrichment through the ability to

[2] In addition to level of informality, speech evaluation includes other speech-specific dimensions such as the language, noise, microphone type, and environment, each increasing the difficulty of the same challenge over time.

TABLE II
REPRESENTATIVE SET OF DOMAIN-SPECIFIC TASKS FOR FOUR DOMAINS (TRAFFIC, HEALTHCARE, FINANCE, AND SPORTS) ACROSS THREE DATA-SCIENCE PROBLEMS: ANOMALY DETECTION, PREDICTION, AND ALIGNMENT

| | | **Problem** | | |
|---|---|---|---|---|
| | | **Anomaly Detection** | **Prediction** | **Alignment** |
| **Domain** | **Traffic** | Cleaning up gaps and inconsistencies in lane detector data | Determining upcoming traffic speed using flow volume and percentage occupancy | Relating traffic events in reports to traffic video segments containing those events |
| | **Healthcare** | Detecting outliers for discovery of fraudulent claims | Predicting patients' next-day care needs from electronic patient records | Correlating patient data for patients across multiple medical databases |
| | **Finance** | Detecting potential change in direction and momentum of market | Identifying potential stock market events from sentiments in social media | Aligning past news reports with previous stock market events |
| | **Sports** | Detection of anomalies in athletic performance data (e.g., injuries, doping) | Predicting successful athletic strategies (i.e., when to pull out and/or substitute a player in a team sport) | Identifying athletes across multiple performances (i.e., games or competitions) |

leverage tasks, methodologies, and metrics across fields that might at first glance appear to be too distinct for such sharing.

## IV. DISCUSSING A SOLID FOUNDATION FOR EDR

This section highlights the potential for transferring knowledge of approaches, solutions, measures, and evaluation paradigms across those diverse disciplines, domains, problems, and tasks. Throughout this discussion, the goal is to determine what it would take to establish a solid foundation for enriched EDR and to provide data scientists easy access to methodologies, solutions, and metrics from multiple fields. The aim is to enable the discovery of new ways to use these that were not previously anticipated, thus enabling forward progress on longstanding research problems across multiple fields.

Each of the questions presented in Table I will be addressed in turn, below. We adopt the view that addressing the issues inherent in such questions provides a foundation for an enriched notion of EDR that spans a range of diverse fields.

### *A. Classes of Problems: What kinds of measurable problems, techniques, and algorithms generalize across domains?*

Identifying the classes of measurable problems, techniques, and algorithms that generalize across data-science domains is crucial for an enriched notion of EDR. In the work of [22], classes of problems in data science are outlined, with four example case studies that span these classes. The disciplinary foundation for these four studies is the same: language processing (topic modeling, emotion-word mining, and informal language analysis). The primary problem of interest in this work is classification of vocabulary usage in news and blog data.

There are, however, a number of problems in data science that span a diverse range of fields and that serve as the basis of EDR. Consider the following set of data-science problems, and their application to a representative set of domains, as shown in Table II:[3]

- Anomaly Detection: identification of items or events that do not conform to an expected pattern
- Prediction: estimation of variables of interest at future times
- Alignment: correlation of different instances of the same object

Although these three problems are generalizable across domains, the specific contexts for each of these problems varies across diverse fields—as illustrated by the range of different tasks under each of the problem headings in Table II. The next section discusses the nature of such tasks, after which we argue that—despite domain-specific contexts surrounding these tasks—there are domain-independent solutions and metrics for data-science problems that enable generalizability, thus enriching EDR.

### *B. Tasks: What is an appropriate series of tasks for which data researchers in different fields may want to test their approaches, either at the task-component level or at a higher end-to-end level?*

It is a significant undertaking to find an appropriate series of tasks for which data researchers in different fields may want to test their approaches. Among other challenges, to make progress in EDR, the tasks that are selected for an evaluation must be both technically challenging and a significant step beyond current state of the art.

A comparison of solutions is a central component of all tracks in a community-wide evaluation, and such comparisons often serve as the basis of support for future research. A wide evaluation may be used to determine whether there exist solutions that were not widely known in a given discipline at the time a problem was posed—or it may bring about new solutions that otherwise never had been considered. Scoping out the space of possible solutions (whether pre-existing or proposed anew) is critical for moving forward in ways that enable the development of approaches that are revolutionary, not evolutionary.

To support this endeavor, it is critical to ensure a low enough "barrier to admission" so that the range of tasks selected for evaluation are addressable by researchers across disciplines (hence diverse solutions), while bringing together a broader, international community of data-science researchers. This allows different approaches to be compared and the leveraging of cross-domain synergies.

[3] Additional classes of problems might be considered for examination of methodologies and EDR across domains [3]. For brevity and illustrational clarity, only four are presented here.

TABLE III
POSSIBLE DOMAIN-INDEPENDENT METHODS AND METRICS TO DATA-SCIENCE PROBLEMS. NOTE THAT THE METRICS FOR PREDICTION ASSUME A CONTEXT WHERE A CONTINUOUS VALUE (AND NOT A PROBABILITY) ARE BEING PREDICTED.

| Problems | Methods | Metrics |
|---|---|---|
| **Anomaly Detection** | Timeseries outlier detection, Statistical deviation detection | Accuracy, precision, recall, F1-Score, ROC or DET area, decision cost function, and average precision |
| **Prediction** | Multiple regression, Random forests of regression trees | Mean absolute error, Mean squared error, Root mean squared error, $R^2$ (a correlation metric). |
| **Alignment** | Substring matching, hidden markov models | Same metrics as for anomaly detection. |

A series of tasks for which data researchers in different fields can build solutions that are evaluated—both at the component level and at the end-to-end level—is important for taking EDR to the next level beyond existing approaches to evaluation. Some examples of domain-specific tasks are enumerated under each three problems shown in Table II.

What is interesting to note is that, despite the domain specificity of tasks, the solutions fall under the heading of much broader categories of problems in data science. As such, specific tasks such as "identifying errors" and "detecting change in direction" fall under the general problem of Anomaly Detection, "predicting care needs" and "identifying future stock market events" fall under the general problem of Prediction, and "relating traffic events (in reports and video)" and "identifying athletes across performances" fall under the general problem of Alignment. Thus, it is expected that similar—or possibly the same—methods and metrics would be applicable for each of these generalized problems, even in cases where the specific tasks seem to be entirely different at first glance. (This point will be discussed further in the "Methods and Metrics" section below.)

Additionally, different tasks need to be combined and evaluated in a workflow in order to better evaluate the consequences of error in one task on another. For instance, different tasks are pipelined, as one example, as five stages in the "Big Data pipeline" [23] described in Jagadish et al. [23]: data acquisition; information extraction and cleaning; data integration, aggregation, and representation; modeling and analysis. By choosing tasks that complete a workflow, such as the pipeline above, individual component-level evaluations can be combined to better evaluate the data analytic process as a whole.

### *C. Methods and Metrics: What kinds of methods and metrics generalize to problems across domains? How effective is the generalization and what can be done to make algorithms and methods more domain-independent?*

One goal of this application of EDR to data science is to develop enhanced methods and metrics that are general-purpose and can be applied to different data analytic components in a variety of domains. However, given the details of the different domains, this can be challenging.

Consider the range of domain-specific tasks associated with each of the three problems presented earlier in Table II. Some possible domain-independent methods and metrics for these tasks are summarized in Table III.

Even if the notion of *anomaly* is clearly specified for a given domain (which may not be the case), the methods associated with this notion may differ based on the context. In many cases (such as in the traffic case), anomalies are viewed as potential errors to be cleaned, yet in other cases (such as in the healthcare domain), an anomaly is considered an important point of interest; indeed, it might be the key to detecting critical information such as fraudulent claims. As such, the methods associated with anomaly detection differ depending on the use case, and the metrics would reward methods in different ways.

However, despite such distinctions, EDR is likely to be enriched by anomaly-detection techniques that work well in one domain and are successfully applied in another domain—even if the results are handled differently by downstream processes. When methods in one domain are "borrowed" for another domain, various forms of adaptation may be needed for maximum efficacy. For instance, in the traffic case the data might be viewed as time-series data where the local (timewise) points might be leveraged to assist in cleaning (Basu and Meckesheimer [24] provide one such algorithm), whereas in the healthcare case time-locality may not be available or relevant. Adaptation of existing techniques—or combinations of cross-field techniques—may yield a solution for both disciplines that might not otherwise have been considered.

As for generalization of metrics to problems across domains, consider the case of Prediction from Table II. If a continuous answer is needed, e.g., "determining upcoming traffic speed," there are a variety of metrics including mean absolute error , mean squared error, root mean squared error, and correlation metrics, such as $R^2$. If a discrete value is required for a prediction task, e.g., "predicting successful athletic strategies," there are also a variety of metrics that would be applicable across domains, including accuracy, precision, recall, F1-score, Receiver Operating Characteristic (ROC) or Detection Error Tradeoff (DET) curve, and average precision. Although it is often the case that certain metrics are traditionally used within certain disciplines or domains, there is no reason to expect that such metrics would not be applicable to multiple disciplines or domains—thus yielding insights that enable research progress in individual disciplines.

Caruana and Niculescu-Mizil [25] showed that supervised learning algorithms may vary heavily depending on the metric used, and it is often not clear which metric is most relevant for the specific domain. Even if a metric is general-purpose, the contexts in which it matters may not be. One approach to designing a general-purpose metric that can be applied in a variety of domains is to simulate (through different datasets) how the score of each metric varies in different situations. Another approach is to pick metrics that have desirable properties, such as symmetry. Xiong and Li [26] performed a study on various clustering metrics to illustrate different properties of those metrics and to single out metrics that may exhibit desirable properties in a variety of scenarios and datasets.

The idea of understanding contexts where certain methods score lower according to certain metrics is a part of EDR; likewise, the idea of understanding the strengths, limitations, and properties of metrics is a key component of EDR. Evaluations that apply different metrics to compare different methods provide insight into general-purpose methods and metrics, allowing comparisons to be made with respect to the generalizability of the metric to different contexts.

### *D. Ground Truth: Are there ground-truth considerations common to different areas of data science? Are there effective techniques to handle the lack of ground truth in different domains, both for solving problems and for evaluating methods and metrics?*

The ability to obtain and use ground truth is fundamental to measuring the effectiveness of research, both within and across domains. Although the existence of ground truth is an inescapable part of almost any data-science endeavor, identifying it may require a significant effort and is often fraught with challenges and limitations, regardless of the domain.

At the source of many recurring problems associated with data-science evaluation—regardless of domain—is the absence of the "right" data against which to evaluate.[4] For example, there may be no way to collect the data in the first place, or the data collected may not be varied enough to represent the range of cases that actually arise, or the data collected "after the fact" may be unrealistic in some way that is a mismatch with the actual situation. More concretely, the following are the difficulties that are likely to occur in many different areas of data science:

- The "answer" to the problem could be unknown, i.e., *ground truth is completely missing*. This could be because humans cannot know the answer—the information does not exist (e.g., predicting a system output that will compel a user to take a particular action). Alternatively, if the information does theoretically exist, it may by design or otherwise never be recorded (e.g., grouping bitcoin wallets). Having no "answer" for comparison of system outputs means no comparative analysis of accuracy can be made. In such cases, determining whether a system actually addresses the problem it set out to solve becomes a significant challenge. However, should the problem fall into a well defined category of data science, mapping the system to another domain may give an estimation of how well the system addresses the given problem. For example, if ground truth is unavailable to evaluate patient matching algorithms (matching different occurrences of the same patient across data sources, e.g. hospitals, GPs, labs etc., where there can be many data-entry mistakes), researchers may try to evaluate the same algorithm in the domain of athlete matching, where information is less sensitive and therefore more available.[5]
- Ground truth does exist but may have significant limitations (e.g., only 1% of the data are labeled). This could be for multiple reasons, often due to the expense of gathering ground truth, whether in terms of resources, machine time, or person hours. Evaluating on limited ground truth introduces biases that hinder the ability of researchers to correctly assess the accuracy of their systems. For example, in a classification problem, or some classes could be missing, the spread of classes could be unrepresentative of the main dataset. Techniques to limit these biases have been developed. For example, Katariya et al.'s [27] work on active evaluation of classifiers across multiple domains estimates accuracy by utilizing a (usually human) labeler to construct a small (adaptive) labeled set, which is then used as limited ground truth.
- Ground truth does exist but is either partial or unreliable. In the case of partial ground truth, the answers may be available only under specific conditions or they may represent only part of the relevant information. In the case of unreliable ground truth, the answers may be associated with low-confidence output, e.g., results obtained from crowd-sourcing approaches such as "turking".[6] Being able to use partial or unreliable answers is sometimes the only way to evaluate systems for a given problem, given the lack of complete and/or accurate "answers." Previous NIST evaluations (e.g., [8]) 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).

Therefore with respect to ground truth, we make two claims: (1) The more standardized the tasks, measurement methods, and metrics, the easier it is to adapt techniques across problems (see the above example in the case of absent ground truth); (2) Techniques that are developed in one domain to address incompleteness of ground truth can be adapted to other domains.

### *E. Data Uncertainty: What approaches to handling gaps and inconsistencies in data are applicable to multiple domains?*

Uncertainty arises in every domain. It comes from measurement error and noise that add variance and, possibly,

[4] In this paper, Ground Truth is referred to in the context of the evaluation (i.e. "test") side, but it is worth noting that methods that address the lack of, or limited, Ground Truth may also significantly aid training purposes.

[5] Of course, mapping a problem to a different domain gives rise to its own set of problems (e.g., there may be a different distribution of errors for athlete matching than for patient matching), and this may impact effectiveness of systems as well.

[6] Using frameworks like Amazon Mechanical Turk [28], researchers can outsource the labeling of their data. This method is often considered unreliable as workers are paid by the amount of data labeled and are anonymous.

TABLE IV
EVALUATION METHODOLOGY: DOMAIN-INDEPENDENT METRICS TO BE APPLIED IN THE PILOT EVALUATION (FALL 2016)

| **Data Science Problem (Task)** | **Measures of Success (Metric)** |
|---|---|
| Cleaning$_{Det}$ and Alignment | Decision Cost Function ($DCF$) representing a linear combination of the miss and false alarm rates at a threshold $\tau$. The overall performance metric is the minimum $DCF$ value obtained considering all $\tau$: $\min_{\tau}\left(DCF(\tau)\right)$, where $DCF(\tau) = c_{miss} * P_{target} * \frac{\lvert\text{misses}(\tau)\rvert}{\lvert\text{target trials}\rvert} + c_{fa} * (1 - P_{target}) * \frac{\lvert\text{false alarms}(\tau)\rvert}{\lvert\text{non-target trials}\rvert}$ |
| Cleaning$_{Cor}$ | Mean Absolute Error ($MAE$), where $n$ is the number of trials, and for each trial $i$: $\widehat{v_i}$ is the estimated data value and $v_i$ is the correct data value. The overall performance metric is: $MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert\widehat{v_i} - v_i\rvert$ |
| Prediction | Root Mean Squared Error ($RMSE$), where $\mathcal{E}$ is the set of event types, $n$ is the number of trials, and for each trial $i$, $\widehat{e_i}$ is the predicted count of events of type $e$ and $e_i$ is the true count of events of that type. The overall performance metric is the average of each trial's $RMSE$: $\frac{1}{n}\sum_{i=1}^{n} RMSE(i)$, where $RMSE(i) = \sqrt{\frac{1}{\lvert\mathcal{E}\rvert}\sum_{e\in\mathcal{E}}\left(\widehat{e_i} - e_i\right)^2}$ |

bias to the measured value. To understand the nature of uncertainty in a more general context, consider, for example, multiple measurements of an object by hand with a ruler: it is expected that slightly different lengths would be obtained for each measurement. The resulting distribution of measurement values provides a representation of the uncertainty.

Despite uncertainty's impact on measurement, it is common to ignore uncertainty at one or more stages of the data-to-decision pipeline. In the worst case, uncertainty cascades, causing the output to have no positive bearing on the decision.

Such concerns can be addressed by overcoming several challenges associated with the handling of uncertainty, namely how to represent it, how to communicate it, how to measure it, and how to calibrate its measurement—all at both the component and pipeline levels. We adopt the view that a cross-field analysis supports the development of methodologies for measurement and handling of uncertainty across all domains.

Advances in measuring uncertainty in computer simulations has impacted meteorology [29], medicine [30], and many other domains [31]. Similarly, progress in handling uncertainty in, for example, clustering could impact every domain in which challenges naturally modeled as clustering problems arise, such as image segmentation (where image pixels are grouped by the object the represent) and entity resolution (where mentions of entities in text are grouped by the real-life entity to which they refer).

Thus the potential impact of progress in handling uncertainty in data science is exceptionally broad and its importance will continue to rise as system performance improves to the point where uncertainty is large relative to the performance error[7] and as data science is more commonly used to discover knowledge and make decisions of great consequence.

### *F. Community Cooperation: Is it possible to foster cross-field cooperation for sharing of solutions?*

Even in the days of TIPSTER [6], evaluation frameworks and their accompanying infrastructure were designed to foster collaboration and progress. However, such programs were focused on problems within one discipline (speech recognition in the case of TIPSTER), in contrast to the data science arena, which is expected to bring together vastly different disciplines.

Moreover, in many of the early, single-discipline evaluations (such as *machine translation* [11]), the goal was to find solutions that achieved *as close to* human performance as possible. It could be argued that this goal is entirely supplanted by a very different one in the field of data science, which is to *surpass* human performance (as data sizes are often far too large for humans to achieve the same degree of speed or to manage the cognitive load and fatigue that lead to human error) while *not reinventing* the wheel (as the solution to a data-science problem in one field may have already been found in another). Both goals are difficult to achieve, but they are quite different. For the latter, it is clear that a more concerted effort is needed to encourage cross-fertilization, with novel paradigms that foster collaboration in some very new ways.

Data science problems can be viewed through a much broader lens than that of many prior evaluations. To ensure that researchers in one field have access to knowledge about problems and solutions in another field, it is important that the *evaluation methodology itself* be designed to span multiple domains. In so doing, problems may, in fact, be shared across entirely independent tracks within a data-science evaluation series, and the evaluation framework thus enables accessibility to solutions and metrics that span multiple domains.

At first it may seem difficult—or impossible—for such a large of researchers who study such a vast array of disciplines (such as those from Table II) to come together in a way that enables transfer of knowledge of approaches, solutions, measures, and evaluation paradigms. However, it has become clear that this is absolutely essential with large data sets across many domains, and the potential for sharing common solutions to managing such large sets.

Researchers are willing, more now than ever before, to openly discuss their successes and failures in the context of data-science evaluation. As community-wide evaluation fora (e.g., NIST, Kaggle, and CLEF) and multi-week research workshops (e.g., Frederick Jelinek Memorial Workshops on Speech, Language and Computer Vision [32]) have become increasingly grounded in multi-disciplinary problems, opportunities have emerged to leverage the openness and community spirit that is necessary for an enriched notion of EDR. This type of enrichment was elusive in the early days of EDR, when researchers and companies were expected to compete for status or top ratings—academics climbing the tenure ladder, researchers battling it out for recognition through awards and

[7] which, incidentally, is when one might most want to use a system and to understand the uncertainty of a given output.

funding, and commercial entities holding their "secret sauce" under lock and key.

While some of these competitive aspects are still present nowadays, many communities of researchers are now willing to share ideas, approaches, and lessons learned—and to collaborate in ways that yield better results than would be achieved individually. The time is ripe to take advantage of this forward motion and apply this collaborative spirit not just *within* research communities, but *across* them. The field of data science is uniquely positioned, with its inherent multidisciplinarity, to provide the basis for an enriched EDR at a macro level that has not previously been attained.

A concrete example of an evaluation methodology that can enhance both within- and across-discipline cooperation is the provision of a solid evaluation infrastructure for *Evaluation-as-a-Service* (EaaS) [33], which allows participants to submit systems (through a software container) directly to the evaluator who runs the systems on the evaluator's computing resources and (sometimes hidden) data. Although EaaS sometimes prevents data sharing, it broadens the range of disciplines within which problems can be studied by enabling evaluations even in contexts where the data are sensitive, private, or cannot be shared for other reasons. This paradigm is a direct contrast to those where participants run their methods on their computing resources and then submit results to the evaluator for scoring. A strong case can be made for EaaS as an evaluation methodology that scales to the size and complexity of data found in many fields of data science (e.g., medical data sets). Such an approach is an initial step toward the notion of enriched EDR presented in this paper.

### *G. Diversity of Solutions: Can cross-field synergies yield diverse solutions within a given field?*

For any field, independent solutions may yield rather large gains when combined, if the independent solutions are based on approaches that are diverse enough to yield significant complementarity. Researchers who espouse "hybrid approaches" rely heavily on very distinct sub-approaches such that the combination yields the highest performance gains. Many examples are seen in the field of language, where statistical approaches are combined with symbolic approaches [34] or in tasks related to autonomy and coactive design where automatic techniques are combined with human-in-the-loop guidance [35].

Data science is inherently diverse due to the degree of multidisciplinarity. The solutions expected from data-science researchers are likely to be applicable across different domains—yet be distinct enough to be complementary, such that combinations of approaches may be generated and tested within the evaluation paradigm described in this paper. For example, dynamic programming algorithms used for efficient alignment of data sequences have potential to be applied to alignment tasks in both the traffic and healthcare domains.

An even more compelling result—leading to enriched EDR—is one where an algorithm from one field is combined with that of another field to yield improvements over either one alone. For example, *hidden markov models*, which have been used in conjunction with dynamic programming for alignment problems in biological and healthcare domains for years (e.g., protein sequence alignment [36]), might be used in conjunction with dynamic programming in the traffic domain to improve performance of systems that align traffic video segments to traffic incident reports.

An important enabler of diversity and cross-field synergies (and thus diverse solutions) is global accessibility of evaluation resources and infrastructure. One approach that successfully involves geographically diverse participants is that of Kaggle [15], for which participants invest hundreds of hours in exploring the potential solution space, as mentioned above in Section II. Kaggle competitions include a large number of well-developed tutorials that have been designed to lower the barrier to admission, and solutions to Kaggle competitions are compared within specific problems.

The degree to which cross-domain synergies are leveraged for solutions across different data-science problems (e.g., development of alignment techniques for problems in both finance and sports) is expected to be higher within an evaluation framework that supports a unified effort that covers many different areas of data science simultaneously. Toward that end, we have conducted a case study based on an upcoming pilot evaluation of traffic incident detection and prediction. The goal is to solidify a framework that is designed to support enriched EDR for multi-track data-science evaluation in future years.

## V. A Case Study: NIST's Experience with Traffic Incident Detection and Prediction

In formulating the problems to be addressed for the upcoming (Fall 2016) Pilot evaluation, the NIST team has adhered to the guiding principle underlying the Data Science Evaluation series that the tasks to be evaluated would span multiple fields, such that researchers would be able to test their approaches across different evaluation tracks (domains)—thus giving rise to an enriched EDR and accelerated research progress.

The initial tasks designed for this Pilot Evaluation fall within the realm of the three data-science problems below, with specific application to the traffic domain as an example of each case.

1) **Cleaning.** Detecting and correcting errors in traffic lane detector data flow values.
2) **Alignment.** Matching traffic events with traffic video segments containing those events.
3) **Prediction.** Inferring the number and types of traffic incidents in an upcoming time interval based on historical data.

A fourth data-science problem related to prediction is also included in the Pilot evaluation but not highlighted in this paper is *forecasting*, which produces a timeseries for the predicted values. Like the other three, this task has an analogous standing in other domains such as patient matching in healthcare records, financial trending, and sports.

For this case study, we successfully arrived at an evaluation methodology that included measures of success. We apply this methodology for the Pilot evaluation using the specific metrics and tasks shown in Table IV to the traffic domain, where each metric in the table is the metric used to measure a system's performance on the specific data analytic task. For example, misses and false alarms (for Cleaning$_{Det}$ and Prediction) are defined in terms of traffic events, data values (for Cleaning$_{Cor}$) are defined in terms of traffic flow, and event types (for Alignment) are defined in terms of a range of different types of traffic incidents.

Note that the *detection* task associated with Cleaning—a variant of the "Anomaly Detection" problem—is evaluated by the same metric that is used for Alignment (the Decision Cost Function). However, there is a second task associated with Cleaning, *correction*, that is evaluated by an entirely different metric: Mean Absolute Error (MAE). Prediction is evaluated using yet another metric, Root Mean Squared Error (RMSE).

We further note that there are subtle differences between the metric for the detection task (for the Cleaning problem) and the metric for the Alignment task. The inclusion of the parameters $c_{miss}$, $c_{fa}$, and $P_{target}$ allow the metric to be customized so that it is applicable to a variety of scenarios. For the traffic use case, the parameter values selected for this decision cost function are based on two different scenarios: for detection, $c_{miss} = 1$, $c_{fa} = 1$, and $P_{target} = 0.0312$; for alignment, $c_s = 1$, $c_{fa} = 100$, and $P_{target} = 0.5$.

This *within-domain* parameterizability provides a level of flexibility for the specification of metrics that also enables applicability *across* domains, in support of enriched EDR. For example, the evaluation of anomaly detection in financial trending would be analogous to that of the detection task for data cleaning in the traffic domain, where misses and false alarms in the Decision Cost Function would be re-mapped from the typical traffic events (such as an accident) to financial events (such as a major change in stock price). Similarly, patient matching in the healthcare domain could be viewed as an Alignment task, where the problem of correlating patient data for patients across multiple medical databases is treated analogously to matching traffic events with traffic video segments containing those events; the same Decision Cost Function metric used in the traffic domain would then be applicable to the healthcare domain for this problem.

Data uncertainty received limited attention in the pilot, though arose in the form of noise and error present in the traffic sensors. These were handled by deleting obviously errorful measurement values, and otherwise treating the sensor output as accurate. In this sense, the cleaning task is focused on uncertain data. Incorporating uncertainty into system output for NIST to measure is a likely focus of future work in data uncertainty.

Challenges with respect to ground truth in the pilot were addressed by carefully designing tasks, "removing data“, and through selective annotation. It is worth noting that this is an evaluation problem, not entirely dissimilar to the problem of unsupervised (or semi-supervised or active) learning.

The pilot evaluation will prototype the EaaS concept described earlier [33] by testing the concept of accepting systems as submissions. Although not previously referred to as EaaS, prior NIST evaluations (including the Fingerprint Vendor Test Evaluation) have utilized this framework of accepting algorithms as submissions and running them internally to evaluate them. The DSE pilot includes this in-house running and evaluation of algorithms—an approach that brings the experiment to the data (rather than the other way around) and that enables additional evaluation features including system performance benchmarking.

Finally, it is already the case that cross-field cooperation has played a central role in the development of the data-science evaluation series, with an enriched EDR at the heart of its design. In the very first workshop on this new evaluation series [37], attendees expressed interest in forming new tracks. Two proposals were presented in the areas of plant identification and predictive security analytics. It is expected that there will be 2–3 new tracks within the next year, on problems analogous to those shown in Table II, and that research will progress effectively and efficiently due to the sharing of algorithms and their combinations.

## VI. Conclusion

We have argued that an enriched notion of *Evaluation-Driven Research* (EDR) supports methodologies and effective solutions to data-science problems across multiple fields. We have provided a methodology and evaluation design within which progress in data-science research is enabled through access to techniques that are applicable to, and valuable for, problems in different disciplines.

This paper espouses the view that, to ensure success of this enriched EDR paradigm, it is important to examine challenges associated with cross-field generalizations, and to invest effort and time in adapting work that has already been done by researchers in different fields.

We have grounded our conclusions and insights in a brief preliminary study as a part of a new Data Science Research Program (DSRP). We have defined a set of goals within this program that enables the building and strengthening of different areas of evaluation within data science. As a part of this program, we have built a new Data Science Evaluation (DSE) series in which the principles spelled out in this position paper are currently being leveraged for a multi-track evaluation series with broad coverage of problems in diverse fields.

Most notably, we have made the case that, through cross-field generalizations of tasks, measurement methods, and metrics: (1) solutions in existing fields may be applied successfully in a field for which such solutions had not been previously imagined; (2) techniques that are developed in one domain for understanding and representing uncertainty and addressing ground-truth considerations can be applied to another domain for which such techniques have not yet been discovered; and (3) cross-domain synergies may be leveraged in an evaluation framework that supports a unified effort that covers many different areas of data science simultaneously.

## Disclaimer

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

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

## References

[1] J. S. Saltz, "The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 2066–2071.

[2] B. Dorr, C. Greenberg, P. Fontana, M. Przybocki, M. Le Bras, C. Ploehn, O. Aulov, M. Michel, E. J. Golden, and W. Chang, "The nist iad data science research program," in *Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2015, pp. 1–10.

[3] B. J. Dorr, C. S. Greenberg, P. Fontana, M. Przybocki, M. Le Bras, C. Ploehn, O. Aulov, M. Michel, E. J. Golden, and W. Chang, "A new data science research program," *International Journal of Data Science and Analytics*, vol. 1, no. 3, 2016.

[4] G. Heilmeier, "Some reflections on innovation and invention," in *Founders Award Lecture, National Academy of Engineering*, Washington, D.C., 1992. [Online]. Available: https://www.isi.edu/~johnh/TEACHING/CS651/ARCHIVE/Heilmeier92a.pdf

[5] J. Shapiro, "George h. heilmeier," *IEEE Spectrum*, vol. 31, no. 6, pp. 56–59, 1994. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=284787

[6] J. D. Prange, "Evaluation Driven Research: The Foundation of the TIPSTER Text Program," in *Proceedings of a Workshop on Held at Vienna, Virginia: May 6-8, 1996*, ser. TIPSTER '96. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996, pp. 13–22. [Online]. Available: http://dx.doi.org/10.3115/1119018.1119022

[7] D. S. Pallett, "A look at NIST's benchmark ASR tests: past, present, and future," in *ASUR 2003: IEEE Workshop on Automatic Speech Recognition and Understanding, 2003*. IEEE, 2003, pp. 483–488. [Online]. Available: http://itl.nist.gov/iad/mig/tests/rt/ASRhistory/pdf/NIST_benchmark_ASRtests_2003.pdf

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

[9] D. Reynolds, "Speaker and language recognition: A guided safari," 1 2008, keynote speech at Odyssey 2008. [Accessed: 2015 07 15].

[10] M. Przybocki and A. Martin, "NIST speaker recognition evaluation chronicles," *Computer Speech & Language*, vol. 20, no. 23, pp. 15–22, Apr. 2006.

[11] "NIST open machine translation evaluation," 2015. [Online]. Available: http://nist.gov/itl/iad/mig/openmt15.cfm

[12] "NIST Open Handwriting Recognition and Translation Evaluation (OpenHaRT)," 2010. [Online]. Available: http://www.nist.gov/itl/iad/mig/hart.cfm

[13] M. Snover, N. Madnani, B. J. Dorr, and R. Schwartz, "Fluency, adequacy, or hter?: Exploring different human judgments with a tunable mt metric," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, ser. StatMT '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 259–268. [Online]. Available: http://dl.acm.org/citation.cfm?id=1626431.1626480

[14] O. Kolak, W. Byrne, and P. Resnik, "A generative probabilistic ocr model for nlp applications," in *Proceedings of the HLT-NAACL*, 2003, pp. 55–62. [Online]. Available: http://www.aclweb.org/anthology/N03-1018

[15] R. Metz, "Startup turns data crunching into a high-stakes sport," *MIT Technology Review*, vol. Spring, 2012. [Online]. Available: https://www.technologyreview.com/s/426796/startup-turns-data-crunching-into-a-high-stakes-sport/

[16] T. Tsikrika, B. Larsen, H. Müller, S. Endrullis, and E. Rahm, "The scholarly impact of clef (2000–2009)," in *Lecture Notes in Computer Science: Information Access Evaluation. Multilinguality, Multimodality, and Visualization (Volume 8138)*, 2009, vol. 8138.

[17] M. Q. Patton, "Developmental evaluation," *Evaluation Practice*, vol. 15, no. 3, pp. 311–319, Oct. 1994.

[18] S. E. Sim, S. Easterbrook, and R. C. Holt, "Using Benchmarking to Advance Research: A Challenge to Software Engineering," in *Proceedings of the 25th International Conference on Software Engineering*, ser. ICSE '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 74–83. [Online]. Available: http://dl.acm.org/citation.cfm?id=776816.776826

[19] "Standard Performance Evaluation Corporation (SPEC)," Website, 2016. [Online]. Available: https://www.spec.org/

[20] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[21] J. Leek, "The key word in data science is not data, it is science." *Simply Statistics*, 2013. [Online]. Available: http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

[22] M. Das, R. Cui, D. R. Campbell, and R. R. Gagan Agrawal, "Towards methods for systematic research on big data," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 2072–2081.

[23] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big Data and Its Technical Challenges," *Communications of the ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2611567

[24] S. Basu and M. Meckesheimer, "Automatic outlier detection for time series: an application to sensor data," *Knowledge and Information Systems*, vol. 11, no. 2, pp. 137–154, Aug. 2006.

[25] R. Caruana and A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 161–168. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143865

[26] H. Xiong and Z. Li, "Clustering Validation Measures," in *Data Clustering: Algorithms and Applications*, C. C. Aggarwal and C. K. Reddy, Eds. CRC Press, 2013, ch. 23, pp. 571–605.

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

[28] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, pp. 3–5, 2011.

[29] J. M. Murphy, D. M. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, "Quantification of modelling uncertainties in a large ensemble of climate change simulations," *Nature*, vol. 430, no. 7001, pp. 768–772, 2004.

[30] R. S. Dittus, S. D. Roberts, and J. R. Wilson, "Quantifying uncertainty in medical decisions," *Journal of the American College of Cardiology*, vol. 14, no. 3, pp. A23–A28, 1989.

[31] J. Oden, T. Belytschko, J. Fish, T. Hughes, C. Johnson, D. Keyes, A. Laub, L. Petzold, D. Srolovitz, and S. Yip, "Simulation-based engineering science: Revolutionizing engineering science through simulation–report of the national science foundation blue ribbon panel on simulation-based engineering science, february 2006."

[32] "Frederick Jelinek Memorial Workshops on Speech, Language and Computer Vision," 2016. [Online]. Available: http://www.clsp.jhu.edu/workshops/16-workshop/frederick-jelinek-memorial-workshops-on-speech-language-and-computer-vision/

[33] A. Hanbury, H. Müller, K. Balog, T. Brodt, G. V. Cormack, I. Eggel, T. Gollub, F. Hopfgartner, J. Kalpathy-Cramer, N. Kando, A. Krithara, J. Lin, S. Mercer, and M. Potthast, "Evaluation-as-a-Service: Overview and Outlook," *arXiv:1512.07454 [cs]*, Dec. 2015, arXiv: 1512.07454. [Online]. Available: http://arxiv.org/abs/1512.07454

[34] N. Habash, B. Dorr, and D. Traum, "Hybrid Natural Language Generation from Lexical Conceptual Structures," *Machine Translation*, vol. 18, pp. 81–127, 2003. [Online]. Available: http://ict.usc.edu/pubs/Hybrid%20Natural%20Language%20Generation%20from%20Lexical%20Conceptual%20Structures.pdf

[35] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, B. van Riemsdijk, and M. Sierhuis, "The fundamental principle of coactive design: Interdependence must shape autonomy," *Lecture Notes in Computer Science: Coordination, Organizations, Institutions, and Norms in Agent Systems VI*, vol. 6541, pp. 172–191, 2010.

[36] J. Lyons, K. K. Paliwal, A. Dehzangi, R. Heffernan, T. Tsunoda, and A. Sharma, "Protein fold recognition using hmm-hmm alignment and dynamic programming," *Theoretical Biology*, vol. 393, pp. 67–74, 2016.

[37] "NIST IAD Data Science Evaluation Workshop," Mar. 2016. [Online]. Available: http://www.nist.gov/itl/iad/mig/dseworkshop.cfm