The NIST IAD Data Science Evaluation Series: Part of the NIST Information Access Division Data Science Research Program

Bonnie J. Dorr*, Craig S. Greenberg*, Peter Fontana*, Mark Przybocki*,
Marion Le Bras*†, Cathryn Ploehn*, Oleg Aulov*, and Wo Chang*
*National Institute of Standards and Technology
†Guest Researcher
{bonnie.dorr,craig.greenberg,peter.fontana,mark.przybocki,
marion.lebras,cathryn.ploehn,oleg.aulov,wchang}@nist.gov

***Abstract*—The Information Access Division (IAD) of the National Institute of Standards and Technology (NIST) launched a new Data Science Research Program (DSRP) in the fall of 2015. This research program focuses on evaluation-driven research and will establish 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. The evaluation series will consist of a pre-pilot to be launched in the fall of 2015, a pilot evaluation to be launched in 2016, and a full-scale multiple-track evaluation in 2017. In addition to these evaluations, this new research program aims to address several infrastructure challenges and to encourage easier group collaboration.**

I. Summary

The Information Access Division (IAD) of the National Institute of Standards and Technology (NIST) is launching a new Data Science Research Program (DSRP) in the Fall of 2015. NIST's mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology. Through this research program, NIST aims to provide a framework for the research community to examine a range of different algorithms and methodologies in data science (DS) and to address current challenges and breakthroughs in data science. The DSRP focuses on building domain-independent solutions, i.e., those that can solve a variety of data science challenges across different data domains. The components of the DSRP are summarized in Figure 1. These four key components are:

- **Evaluation and Metrology**: Design and conduct a new international *Data Science Evaluation (DSE)* series.
- **Standards:** Leverage prior work to develop standards for data science.
- **Compute Infrastructure:** Develop an Evaluation Management System (EMS) to support compute and infrastructure needs including test and evaluation (T&E) of different compute paradigms
- **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.

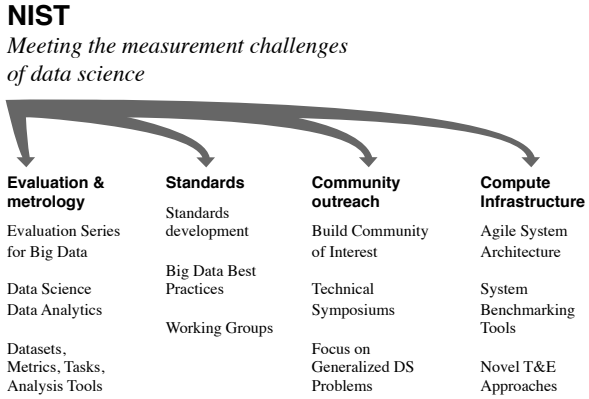

Fig. 1. A summary of the NIST Data Science Research Program. Figure is from [1].

Dorr et al. [1] present more information on this research program, including background and additional citations.

One critical component of the DSRP is the DSE. 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 of challenge problems set in a given domain. In addition to evaluator-hosted tracks, the DSE series will include community-championed tracks. Track proposals will be solicited from the community, and each track included in the evaluation will be planned, organized, and implemented by a "track champion" from within the community.

The DSE series will be developed in three stages: a pre-pilot evaluation that will consist of a single track with a traffic prediction use case, a pilot evaluation that will extend the pre-pilot evaluation-track and will be open to all who wish to participate, and an inaugural evaluation that will consist of multiple community-led evaluation tracks in different domains and use cases. This sequence will enable immediate deployment of a new infrastructure for addressing data science research challenges. This infrastructure will be leveraged for

rapid development and evolution of the DSE series and will effectively enable generalizations to multiple domains and tracks.

II. Evaluation-Driven Research

The core of the DSE is to leverage the framework of evaluation-driven research and to apply it to the area of data science.

The process for evaluation-driven research can be divided into four steps:

1) *Planning.*, Planning includes 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 but reachable in the near term. Receiving community input during this step is critical.
2) *Data and experiment design.*, The experiment design involves developing datasets and associated tasks for experimentation. For example, in machine learning, data are typically partitioned into training, development, and evaluation datasets. An example of a possible experiment is to contrast performance using different training datasets. Rigorously designing experiments and datasets is significantly easier when the data to be used was created for the evaluation (as opposed to being repurposed), though data collection design and implementation has its own challenges (for example see [2]).
3) *Performance assessment.* After the experiment is designed, the performances of the systems are evaluated. In this stage, systems are trained on the training data and run on the test data. In some evaluations, the data is sent to researchers, who run their systems locally and then submit their systems' outputs. In other evaluations, the systems themselves are submitted and then are run by the evaluator. The latter approach is more involved and requires an agreed upon API and ability for every system to run on a prescribed computational infrastructure, though is better suited for evaluations using very large or sensitive datasets. Once system outputs are generated, the experimental results are analyzed.
4) *Workshop.* After the performance assessment, a workshop is held. At this workshop, the research community gathers to openly discuss research in the context of a shared evaluation, 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, wherein the planning for an evaluation takes place, in part, at the workshop of the previous evaluation. See Figure 2 for an illustration.

Progress is driven in evaluation-driven research by repeating the evaluation cycle and, as technology improves, increasing the challenge of the research objectives, which are then addressed in subsequent evaluations. 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 the evaluator's roles in data-centric technology transfer are typically focused on the relatively early and late stages of the process, i.e., core technology research and standards, respectively.

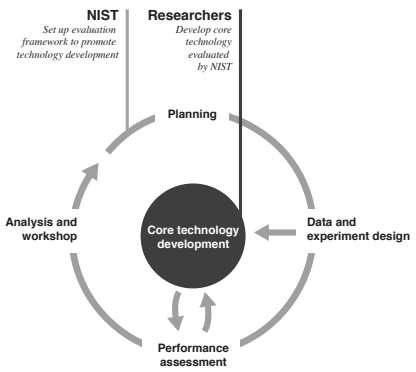

Fig. 2. Overview of the evaluation-driven research cycle.

III. Evaluation Timeline

The evaluation pre-pilot will take place in the fall of 2015. In 2016, an evaluation pilot will be conducted and track proposals will be accepted for a 2017 full-scale data science evaluation. A summary of the DSE is in Figure 3.

Details about the pre-pilot, which is currently underway, are provided in Figures 4, 5, and 6. The data and tasks for the pre-pilot are set in the traffic domain—a domain chosen due to its relevance to everyday life of the general public and due to the accessibility and availability of large amounts of public data associated with this domain. It is important to note that, although the pre-pilot focuses on 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.

Lessons learned from the pre-pilot will be leveraged for development of a larger-scale pilot evaluation, which will still be in the traffic prediction domain. After the pilot, a multi-track full-scale evaluation will be conducted—the first full evaluation in the series.

IV. Conclusion

In summary, the goals of the Data Science Research Program and the Data Science Evaluation Series are:

- to further build and strengthen the data science community,
- to address infrastructure challenges, and
- to provide standards to facilitate group collaboration.

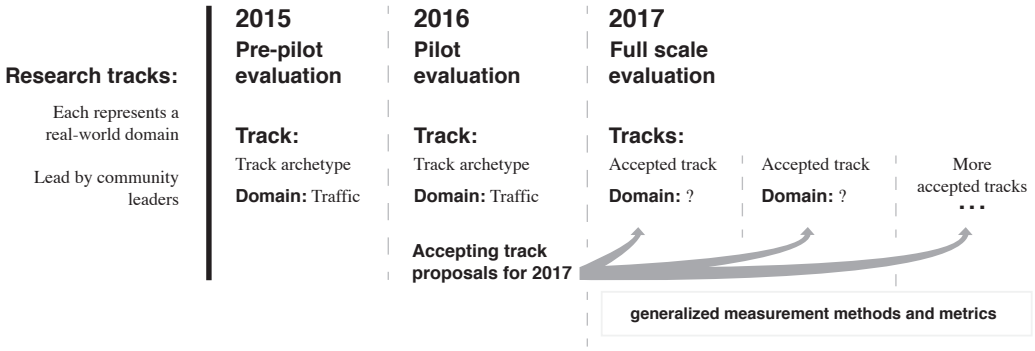

Fig. 3. Overview of the data science evaluation series.

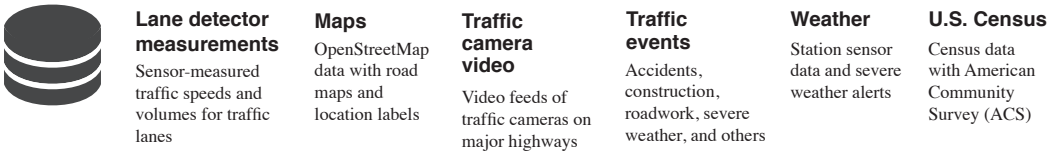

Fig. 4. Summary of the data available for use in the pre-pilot.

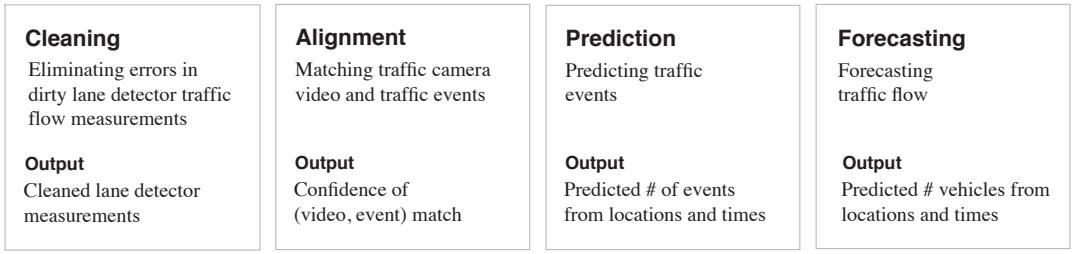

Fig. 5. Summary description of the four tasks in the pre-pilot.

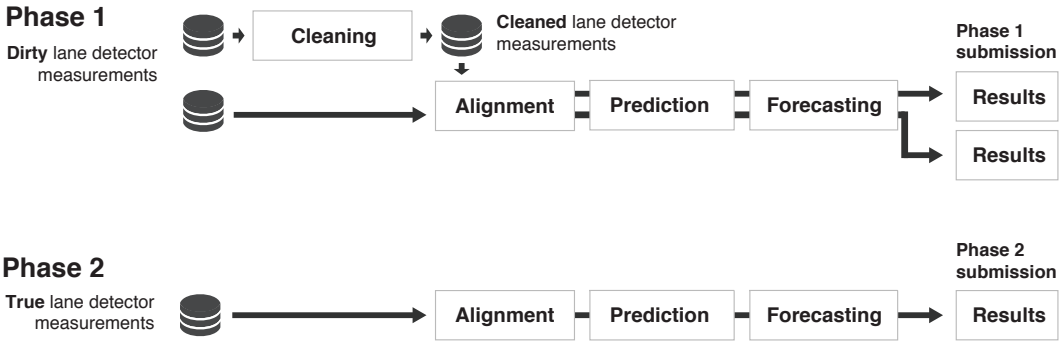

Fig. 6. Summary of the evaluation flow of the Pre-Pilot evaluation. In phase one, participants submit two sets of results for the alignment, prediction, and forecasting tasks: one submission using the original dirty traffic lane detector data, and a second using the cleaned traffic detector data, which is the output of the cleaning task.

REFERENCES

[1] B. J. Dorr, C. S. Greenberg, P. Fontana, M. Przybocki, M. Le Bras, C. Ploehn, O. Aulov, M. Michel, E. Golden, and W. Chang, "The NIST data science initiative," in *To appear in the proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, October 2015.

[2] K. Gallagher, A. Stanley, D. Shearer, and L. V. Klerman, "Challenges in data collection, analysis, and distribution of information in community coalition demonstration projects," *Journal of Adolescent Health*, vol. 37, no. 3, Supplement, pp. S53–S60, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1054139X05002508