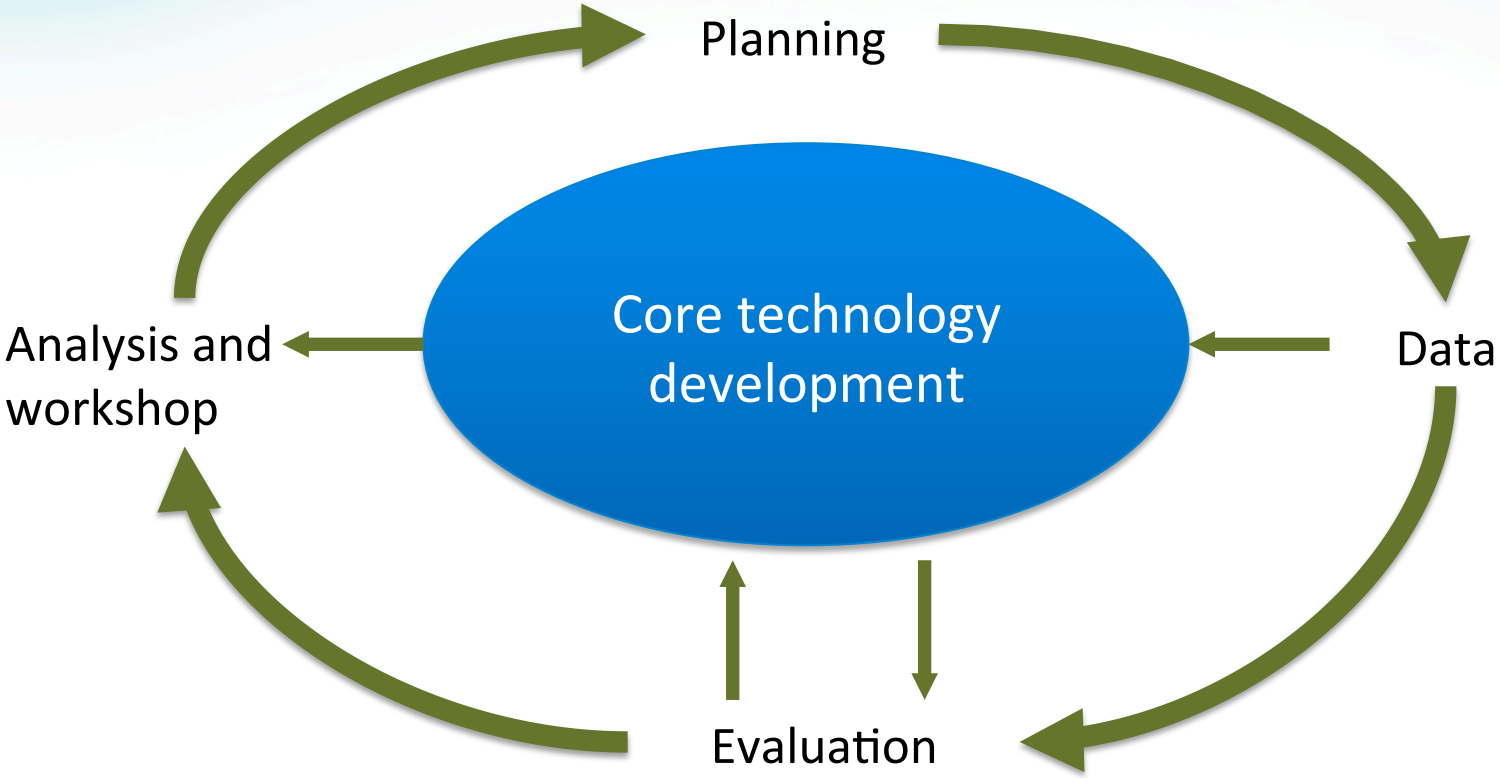
Lessons Learned from Data Science Pre-Pilot Evaluation

NIST Perspective

DSE Series Workshop
March 17-18, 2016

Peter Fontana

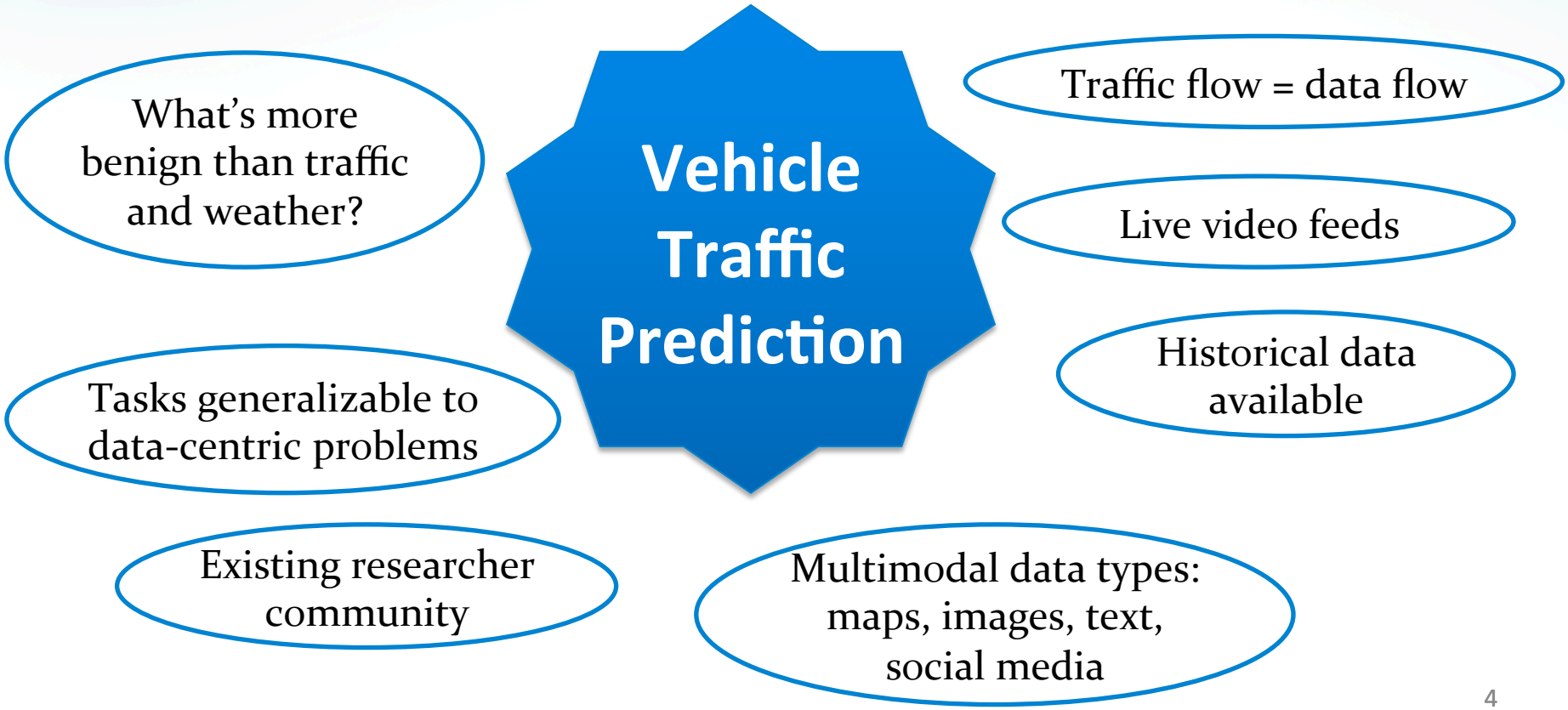
Lessons Learned Influence All Stages of the Evaluation Cycle



Overview: Aspects to Consider

- Use case
- Coverage of trials
- Baseline systems
- Metric interpretation
- Context of a metric
- Benefits of artificiality
- Barriers to entry
- Community involvement

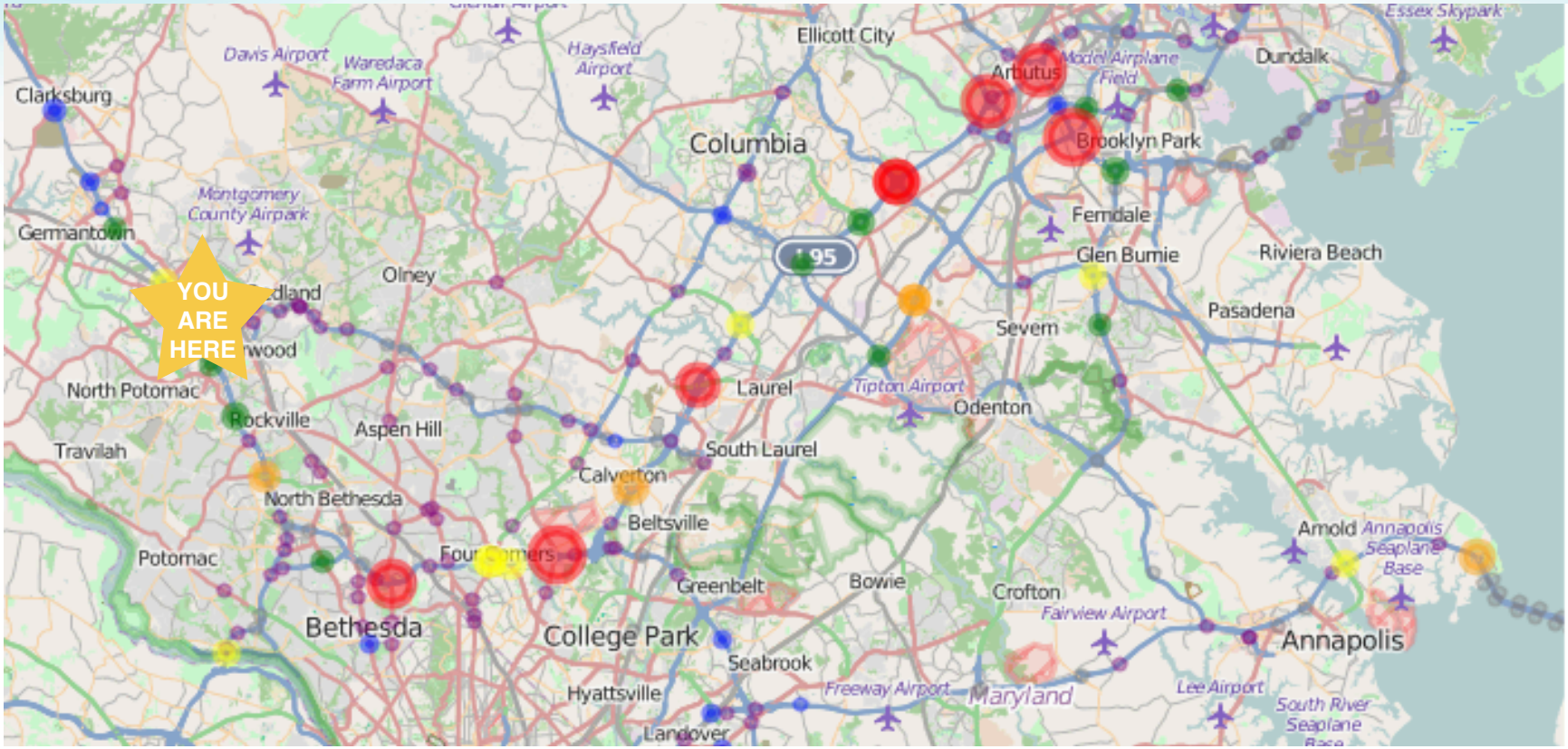
Choose a use case with desirable properties



Select trials that allow for an analysis of variance

- Design the experiment so the **"In what conditions do systems have a worse score"** becomes easy to analyze.
- Utilize **experimental design**, analysis of variance, factor analysis, to choose the trials so that statistical analysis can be performed
- **Domain expertise** helps in selecting factors

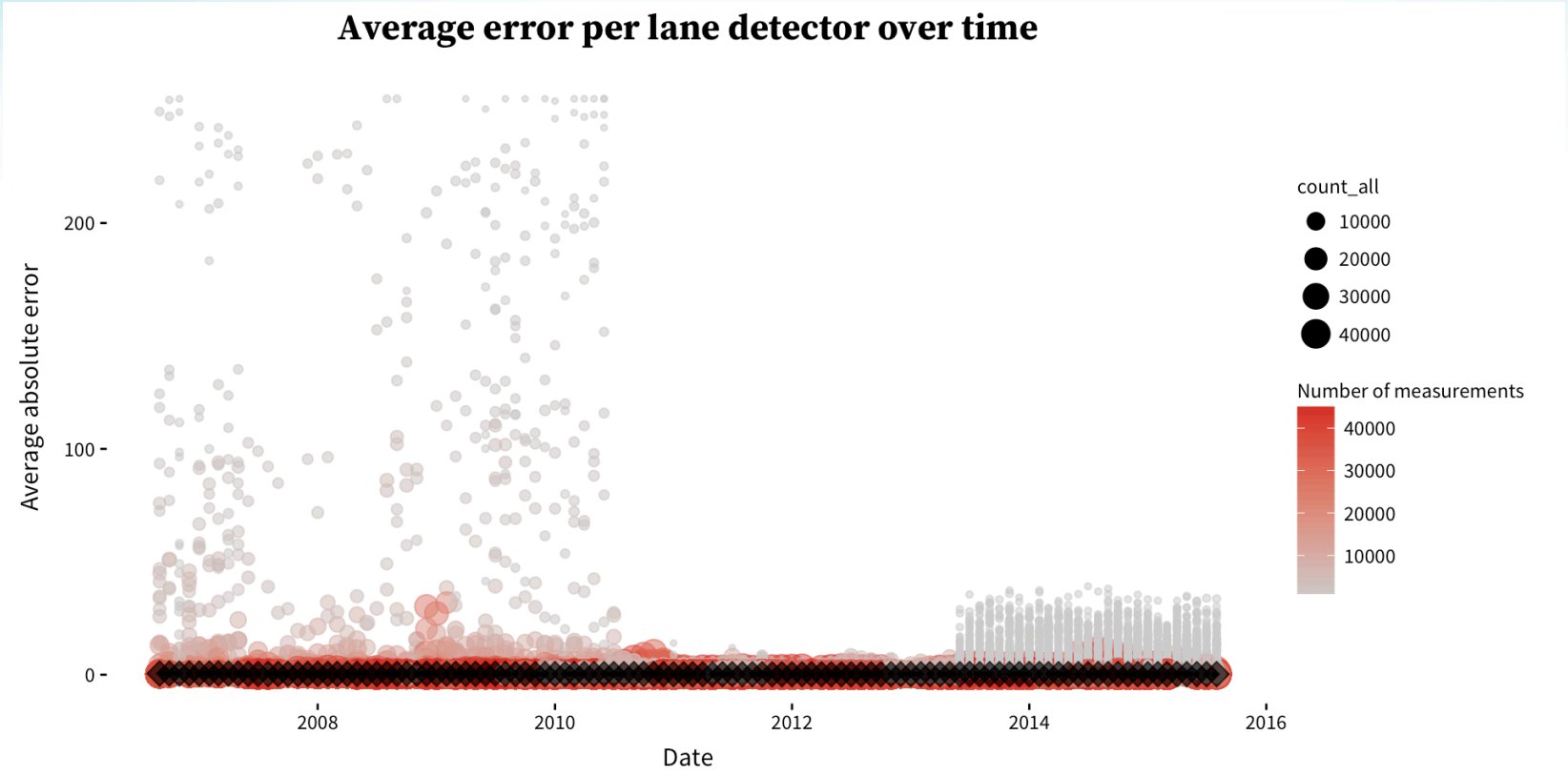
Select trials that allow for an analysis of variance



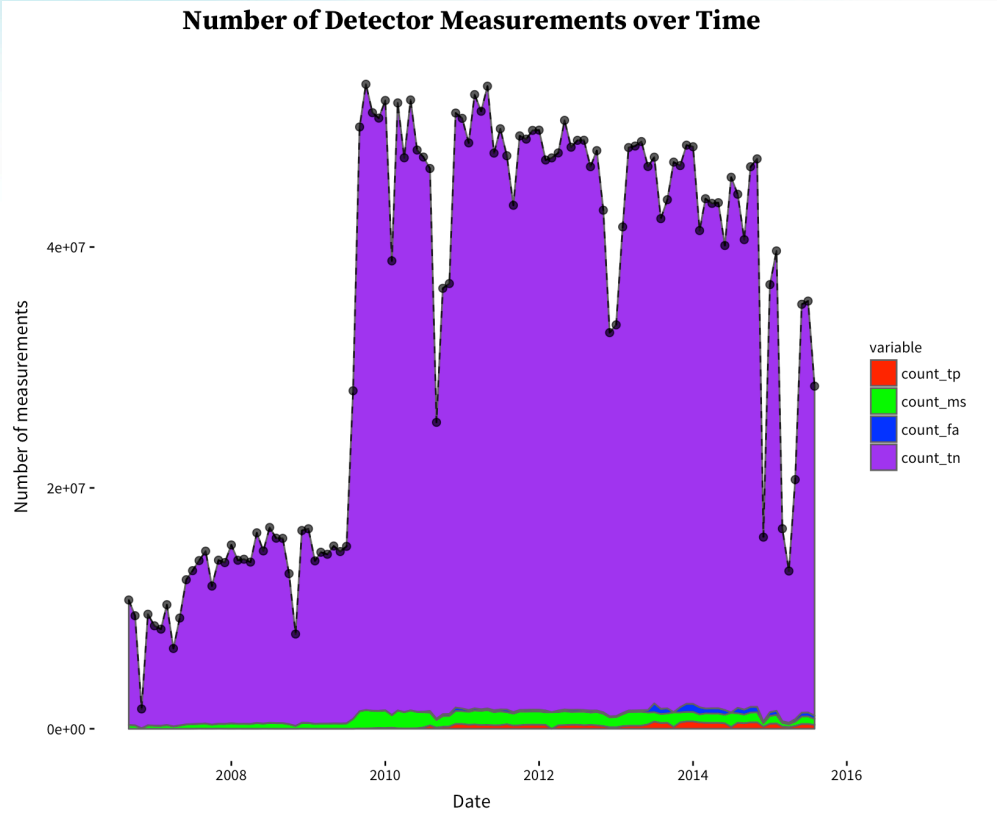
Develop baseline systems to improve task design

- What **about the problem** results in this system having a worse score?
- What computational infrastructure is needed to **complete** these tasks (barrier to entry)
- What computational infrastructure is needed to **benchmark** these tasks (analysis)

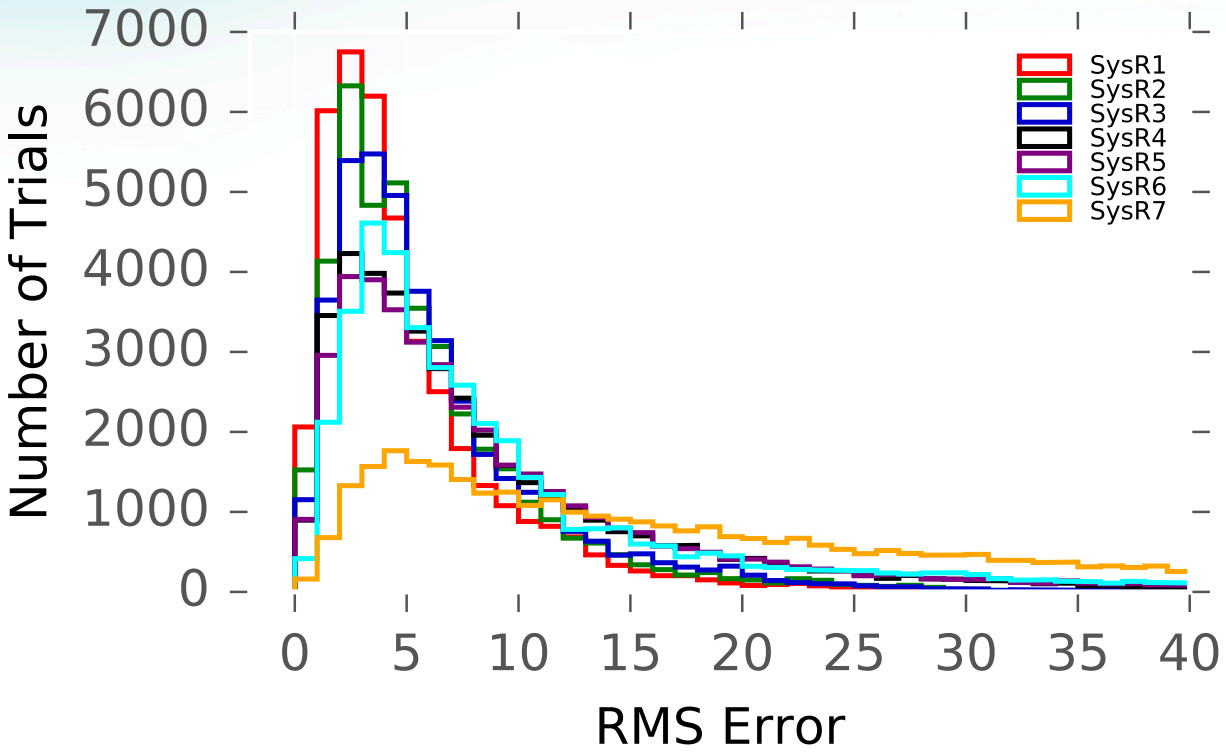
Baseline systems aid in analysis and in evaluation design



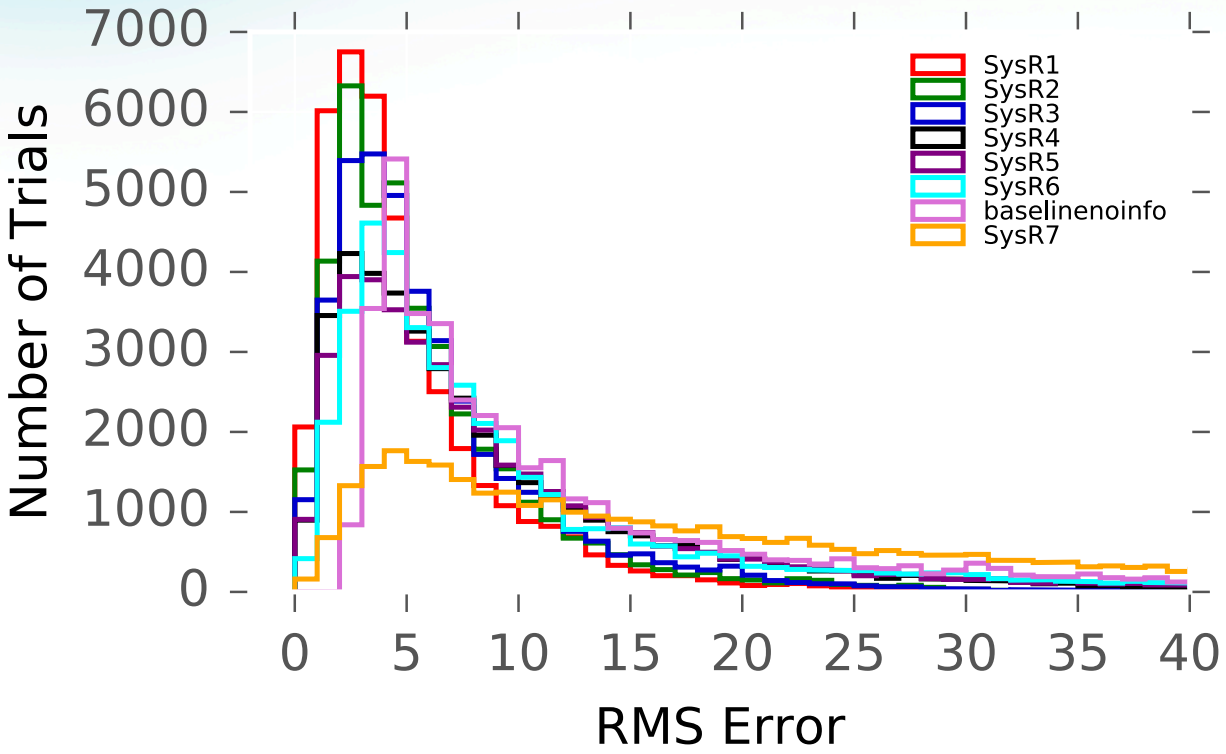
Baseline systems aid in analysis and in evaluation design



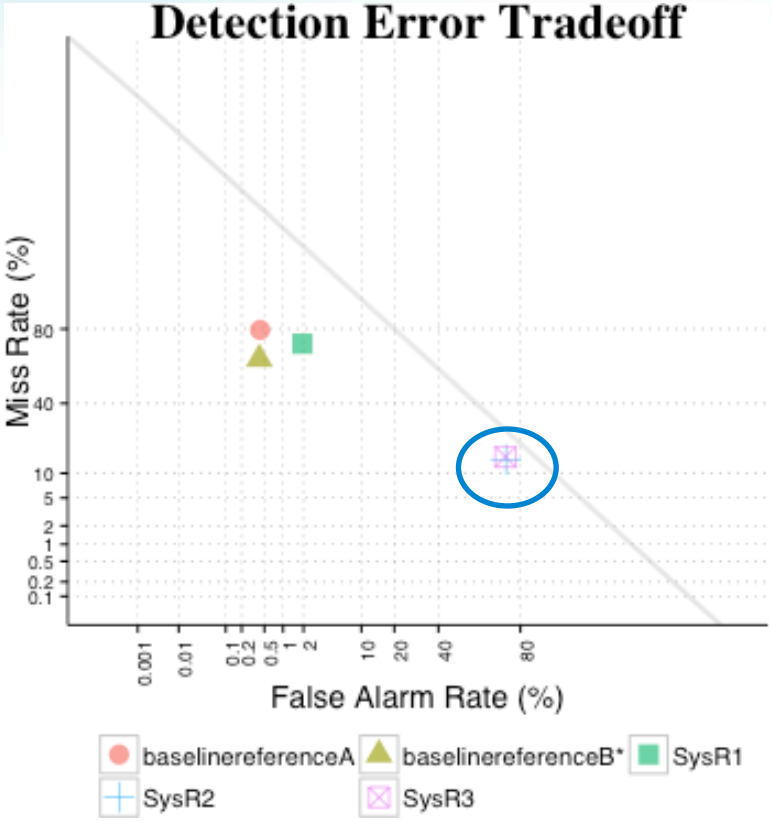
Evaluate known systems to add meaning to scores

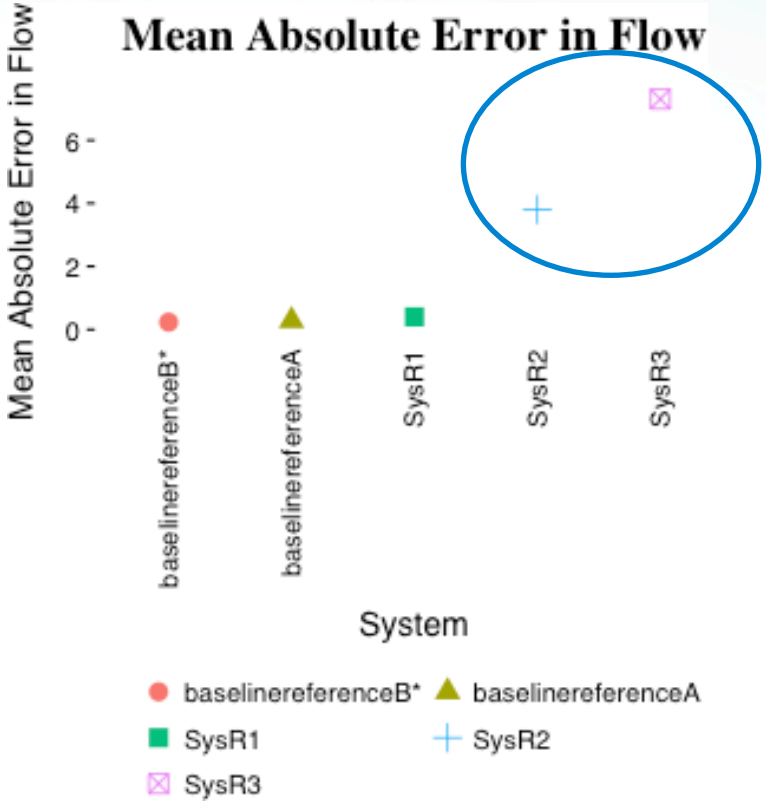


Evaluate known systems to add meaning to scores



Design metrics to treat all approaches fairly





Introduce artificiality when advantageous

- Real data and real problems have many desirable properties
- **Cleaning Task Error Priors:** Unrealistic yet increase external validity and reduce system bias
- **Pre-Processed Video:** can separate alignment from domain research, reducing the barrier of entry
- **Synthesized Flow Errors:** synthetic data allows for holding some factors constant and changing others

Reduce barriers to entry for participants

- Help others **bring processing to the data**
- **Dry runs** iron-out submission and evaluation logistics
- **Leverage technology** developed for other programs within NIST

Involve the community in evaluation design

- The evaluation provides a **framework to discuss data science problems**.
- **Evaluation Design:** get community input to make the evaluation inclusive, interesting, and fair
- **Example: participants' infrastructures:** 150 GB may or may not be a "toy problem"

Involve the community in evaluation design

- **Community-Designed Tracks**
  - Led by **track coordinators** within the community
  - Data and problems of community interest
  - Different domains make evaluation inclusive

Conclusion: Lessons Learned

- Choose a use case with desirable properties
- Select trials that allow for an analysis of variance
- Develop baseline systems to improve task design
- Evaluate known systems to add meaning to scores
- Design metrics to treat all approaches fairly
- Introduce artificiality when advantageous
- Reduce barriers to entry for participants
- Involve the community in evaluation design

Lessons Learned by the Pre-Pilot Participants

datascience@nist.gov