Importance of Measurement
Evaluation Driven Research

- Core technology development
- Planning
- Data
- Analysis and workshop
- Evaluation
Efficiency of Evaluation

Well designed challenge problems, datasets, and metrics **facilitate research progress**
- Reduces spin-up time and general overhead
- Provides a common framework for sharing and understanding approaches and results
- Fosters collaboration

To be effective, evaluation must be
- Goal driven
- Systematic
- Rigorous
How Evaluation Drives Progress

NIST Challenge-Problem Approach

Initial Performance (High Error)

Evaluations focus on initial barriers
- Error rates drop over time
- New barriers identified

More challenging barriers addressed
- Error rates drop
- Technology matures, new challenges possible

Usability Tests
- Applied to new data types

Engineering for speed

Improved Performance (Low Error)

Annual Evaluations
TIME
Annual Evaluations
Evaluation for ASR at NIST-IAD
Data Science Evaluation Goals

**Apply** measurement methods for data science systems
Measure the state-of-the-art and drive progress

**Research** measurement methods for data science
- General measurement/evaluation methods
- Effective use of “found” data
- Large datasets
- Workflows (component and end-to-end performance)
- Human involvement
- System benchmarking
- Mixed measurements (e.g., accuracy as a function of runtime)
# Data Science Evaluation Hurdles

<table>
<thead>
<tr>
<th>Goal</th>
<th>Hurdle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found data</td>
<td>Data licensing / rights, privacy</td>
</tr>
<tr>
<td>Workflows</td>
<td>Structure of communities</td>
</tr>
<tr>
<td>Large datasets</td>
<td>Logistical, cost</td>
</tr>
<tr>
<td>System benchmarking</td>
<td>Is difficult, requires hardware</td>
</tr>
<tr>
<td>Human involvement</td>
<td>Requires labor &amp; IRB, varied</td>
</tr>
</tbody>
</table>
Data Science Evaluation Plan

DARPA XDATA
- Identify Hurdles

Pre-pilot Evaluation
- Overcome Hurdles on Small Scale

Pilot Evaluation
- Overcome Hurdles on Large(r) Scale

Annual Evaluation Series with Multiple Tracks
- Join Measurement and Core Technology Research

Local Private Cloud
- Address benchmarking and technical challenges of running systems at NIST
Data Science Evaluation Schedule

2014 ➔ 2015 ➔ 2016 ➔ 2017

<table>
<thead>
<tr>
<th>Year</th>
<th>Phase</th>
<th>Domain</th>
<th>XDATA</th>
<th>Pre-Pilot</th>
<th>Pilot</th>
<th>Full-Scale Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>Closed</td>
<td>Multiple</td>
<td>Closed</td>
<td>Single-Track</td>
<td>Single-Track</td>
<td>Multiple Tracks</td>
</tr>
<tr>
<td></td>
<td>Domain: Traffic</td>
<td></td>
<td></td>
<td>Invitation-Only</td>
<td>Open to Everyone</td>
<td>Open to Everyone</td>
</tr>
</tbody>
</table>

The evaluation schedule includes XDATA, Pre-Pilot, Pilot, and Full-Scale Evaluation phases. Each phase has a specific domain focus and involves different access types.