Mlbase: distributed machine learning system
Adapted slides from mlbase.org
Lay of the Land

- **Matlab, R**
  - Easy (Resembles math, limited set up)
  - Sufficient for prototyping / writing papers
  - Ad-hoc, non-scalable scripts
  - Loss of translation upon re-implementation

- **MLI**

- **MLlib**
  - Scalable and (sometimes) fast
  - Existing open-source libraries
  - Difficult to set up, extend

- **Mahout**

- **GraphLab, VW**

Ease of use vs Performance, Scalability
Spark: cluster computing system designed for iterative computation
MLlib: low-level ML library in Spark
MLI: API / platform for feature extraction and algorithm development
  - Platform independent
# MLI Ease of Use

## Logistic Regression

<table>
<thead>
<tr>
<th>System</th>
<th>Lines of Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matlab</td>
<td>11</td>
</tr>
<tr>
<td>Vowpal Wabbit</td>
<td>721</td>
</tr>
<tr>
<td>MLI</td>
<td>55</td>
</tr>
</tbody>
</table>

## Alternating Least Squares

<table>
<thead>
<tr>
<th>System</th>
<th>Lines of Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matlab</td>
<td>20</td>
</tr>
<tr>
<td>Mahout</td>
<td>865</td>
</tr>
<tr>
<td>GraphLab</td>
<td>383</td>
</tr>
<tr>
<td>MLI</td>
<td>32</td>
</tr>
</tbody>
</table>
Typical Data Analysis Workflow

1. **Obtain / Load Raw Data**
   - Spark, MLI

2. **Data Exploration**
   - Spark, [MLI]

3. **Feature Extraction**
   - MLI

4. **Learning**
   - MLI, MLlib

5. **Evaluation**
   - MLI

6. **Deployment**
   - Scala