Agenda

- Disclaimer..
- What format does my data need to be for the Machine Learning frameworks?
- Transforming Data
- Options/Alternatives
- Moving Forward..
How do I get data into PowerAI?

ONE DOES NOT SIMPLY
GET DATA INTO POWERAI
What format does my data need to be in for Machine Learning?

- Labeled data
  - Metadata (this is a dog, cat, etc.)
  - Sub folders named appropriately (/dog /cat)
  - CSV. (prevalent)
- TensorFlow
  - TFRecords
- Caffe
  - Blobs
## Transforming Data

<table>
<thead>
<tr>
<th>ANIMAL</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>🦆️</td>
<td>?</td>
</tr>
<tr>
<td>🐿️</td>
<td>?</td>
</tr>
<tr>
<td>🦆️</td>
<td>?</td>
</tr>
<tr>
<td>🐨️</td>
<td>?</td>
</tr>
<tr>
<td>🐨️</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANIMAL</th>
<th>DUCK?</th>
<th>BEAVER?</th>
<th>WALRUS?</th>
</tr>
</thead>
<tbody>
<tr>
<td>🦆️</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>🐿️</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>🦆️</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>🐨️</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>🐨️</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Transforming Data (null)
Transforming Data

Total number of examples

Training Set

Test Set
Transforming Data

Tumor Proliferation Assessment – mitosis detection
Images from electron-microscope
Size of image - 70K * 60K

Data Transformation

<table>
<thead>
<tr>
<th>Framework</th>
<th>Format</th>
<th>Input Size (Faster R-CNN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe</td>
<td>LMDB</td>
<td>1K*1K</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>TensorRecord</td>
<td>1K*1K</td>
</tr>
</tbody>
</table>

Data Distribution among training, validation and testing

Data Shuffle
Options/Alternatives

Distributed Cross Validation

Model #1 Training
Model #2 Training
Model #3 Training

Best Model
DSX

• Access data from local files
  – Local CSV data.
  – Loaded and transformed to DSX “Asset”
  – Drag and drop/file explorer
• Access HDFS data
  – Cloudera/Hortonworks/BigInsights
• Access RDB Data
  – Scala, Python, R APIs
  – Db2, Netezza, Informix, Oracle, Mongo
• Remote data?