Outline

- About Me
- Motivation
- Outlier Detection
- Outlier Attributes
- Emailing Data
- Limitations
- Acknowledgements
About Me

- 4th year undergraduate computer science student at University of Nebraska - Lincoln
- Interest in this topic stemmed from taking an introductory course in deep learning
- Hobbies: soccer, skateboarding, running, late night drives and tunes, travelling
Motivation

- Researchers rely on stability on OSG sites
- Accounting data is stored on ElasticSearch database
- Monitoring accounting data can determine failed sites
- No current way to determine partially failed sites
  - Partially failed defined as sites that have not shut down, but might not be functioning as they are intended
Motivation

- Develop a way to monitor and detect partially failed sites
- Send information to system administrator to look at more closely
- Allow for better use of resources and faster job completion
- How do we process and view trends through all this data?
What is an Outlier?

- Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

- For accounting data, an outlier is an unexpected change in behavior.
Detecting Outliers?

- How are we detecting them?
  - We look at the last year of usage, to determine if the last few weeks are an outlier
- Split dataset
  - Training (1 year long)
  - Testing (Last 3 weeks)
- Use **isolation forest** to determine outliers
  - Scalability
  - Fast detection with less memory
Outlier Attributes

● Currently, we only look for changes in CoreHours per VO per Site
● We could add changes in CPU time (which may indicate changes in IO patterns)
● We could look at changes in user behavior:
  ○ Was someone submitting jobs, then suddenly stop?
  ○ Did someone suddenly start submitting jobs?
  ○ Did a user’s jobs suddenly become much longer on average?

<table>
<thead>
<tr>
<th>#</th>
<th>4.789</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>5</td>
</tr>
<tr>
<td>#</td>
<td>194</td>
</tr>
<tr>
<td>#</td>
<td>117</td>
</tr>
<tr>
<td>#</td>
<td>77</td>
</tr>
</tbody>
</table>
Examining Data

- Usage starts in April
- Large spike in September
Examining Data Cont...

- Arizona State
- Dip in early September
  - New compute entry point added on 8/18
- Older cluster and new one worked simultaneously
- New CE broke some way
  - Usage spiked
- CE operators looking into this
Emails

- Emails are automated to be sent out at the beginning of every week showing data from previous week
- Email contains a .png attachment of graphs
- VO names listed
- Run from a python script
- Deployed using Docker container
Limitations

- Only last 3 weeks can be determined to be an outlier
- False positives
Acknowledgements

- Derek Weitzel
- Carrie Brown
- Tim Cartwright
Thank you!
Questions?