NCAR’s Research Data Archive (RDA) houses data for scientific use.

- The data are organized in a large variety of diverse datasets.
- Users often want to download just a subset of one dataset.
- Subsetting requests are carried out by the supercomputer Casper using the RDA’s computing resources (time and memory).

**Goal:**
To better predict the memory and time usage of these requests, giving faster turnaround times and less wasted memory.

**INTRODUCTION**

**TOOLS & TECHNIQUES**

- Collected data on over 100,000 requests over the span of 7 months
- Extracted information from "request info" strings generated when a request is made
- Divided jobs into categories for use in a machine learning classification model
- Trained a random forest model to predict what category each request falls into
- Weighted the model’s training algorithm to pay more attention to larger memory jobs
- Made custom predictions based on the model’s outputted probabilities in order to get safer estimates for resource usage

**HOW DID WE DO?**

**Additional use of our ML model:**

**Subset Data Request 503058**

Your Subset Data request has been submitted successfully. A summary of your request will be processed soon. The estimated time that your request will be ready is 18 minutes.

- Total memory requested: 68.48 TB
- Total memory used: 2.51 TB
- Total memory wasted: 65.98 TB

**WANT TO KNOW MORE?**

- Slides
- Github

**ACKNOWLEDGMENTS**

Thank you to SiParCS, NCAR, and the NSF for supporting me in this project. A huge thank you to my mentors Riley Conroy and Brian Vanderwende for their support, their advice, and their trust. Thanks also to D.J. Gagne for his extremely valuable insights and recommendations.