Touring Dataland?
Automated Recommendations for the Big Data Traveler

William Agnew
Georgia Institute of Technology

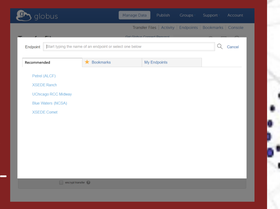
Michael Fischer
University of Wisconsin-Milwaukee

Kyle Chard (Advisor)
University of Chicago

Ian Foster (Advisor)
University of Chicago

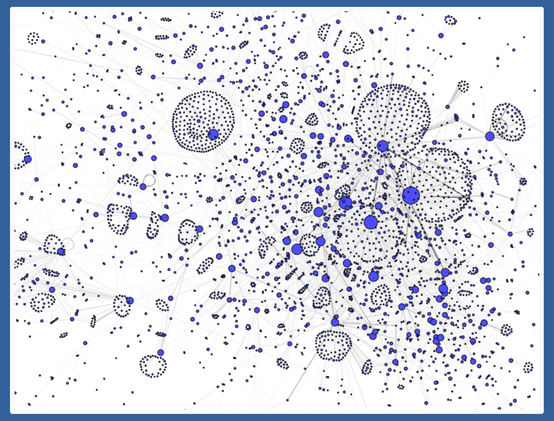
Introduction

- Big data scientists, like travelers to a new land, are faced with the daunting task of discovering which (storage) locations contain interesting attractions (i.e., research data)
- Many services, such as travel websites, provide user-specific recommendations derived from analysis of huge amounts of usage data
- We explore how recommendation approaches can be adapted and applied to big data science. In particular, we create heuristics for recommending Globus data locations

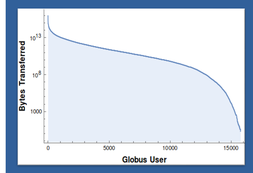


Recommendation Mockup

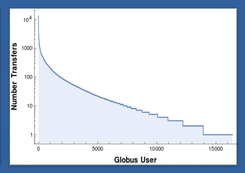
Globus



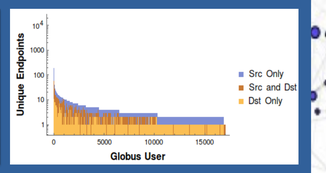
Globus [1] network. Each endpoint is a vertex, larger if endpoint is more popular. Edges between endpoints that have transferred, more visible is transfer between pair is more frequent.



Bytes Transferred per User



Transfers per User



Unique Endpoints per User

Long-tailed Usage Distributions

Heuristics

History: The most likely source (S) / destination (D) endpoint is the most recent S/D endpoint used by a user

Markov Chain: A transition matrix of the observed probabilities of using each endpoint as a S/D conditioned on a particular endpoint being previously used as a S/D

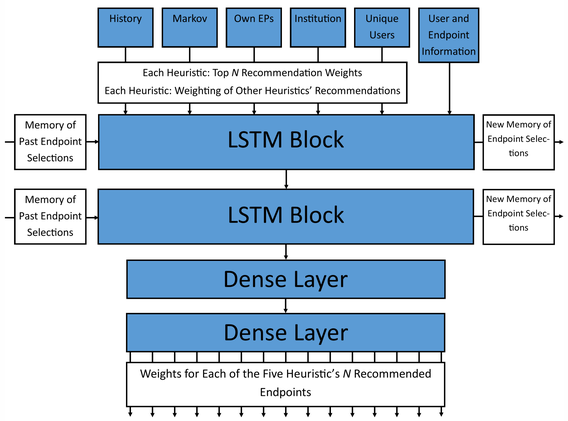
Most Unique Users: The most likely S/D endpoint is the S/D endpoint with the most unique users

Institution: The most likely S/D endpoint is the most popular endpoint at that user's institution

Endpoint Ownership: The most likely S/D endpoint is the endpoint most recently created by the user

Deep Recurrent Neural Networks

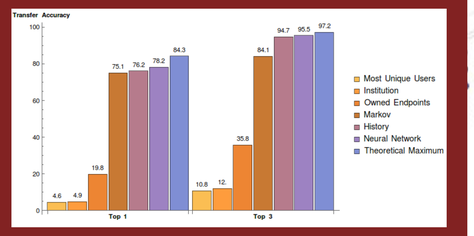
- Heuristics perform well for different classes of users
- We use a deep recurrent neural network [2] to combine heuristics by ranking the predictions of each heuristic for the series of user endpoint choices



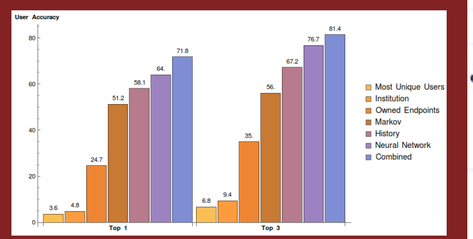
Neural Network Block. Takes as input heuristic endpoint recommendations and memory of past recommendations to user and outputs reweighted heuristic endpoint recommendations and updated recommendation memory

Results

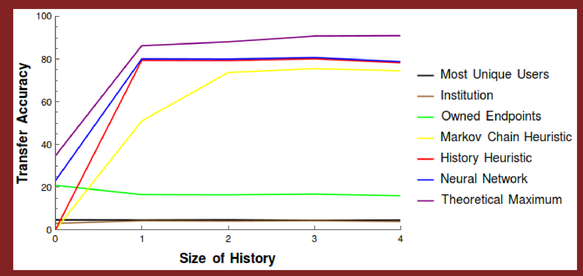
- **Transfer accuracy:** the average number of endpoints correctly predicted



- **User accuracy:** the average accuracy per user, where a user's accuracy is the fraction of that user's endpoints correctly predicted



- The neural network, which combines heuristics, outperforms all individual heuristics
- The most unique users, institution, and owned endpoints heuristics perform poorly except in cases where users have little or no transfer history



References

1. Foster, Ian. "Globus Online: Accelerating and democratizing science through cloud-based services." *IEEE Internet Computing* 15.3 (2011): 70.
2. Graves, Alex. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850* (2013).

Website

wagnew3.github.io

Acknowledgements

This work is supported in part by the National Science Foundation grant NSF-1461260 (BigDataX REU)