Predicting Scientific Data Transfer Characteristics

William Agnew, Michael Fischer, Kyle Chard, Ian Foster

Angela, welcome to your **very own Netflix homepage!**

Based on your ratings, we've filled it with personalized suggestions **JUST FOR YOU**.

The more you rate, the better we get at giving you suggestions you'll love.

Top 10 for Angela

RACHEL BILSON   TOM STURRIDGE
WAITING FOR FOREVER
"THIS ONE GRABS YOUR HEART!"

INSPIRED BY A TRUE STORY
THE FAT BOY CHRONICLES
Bullies took everything from him, except his courage.

Discovery
howstuffworks?

MEMORIAL DAY

THE KING'S SPEECH

Movies Popular on Netflix

DUHAMEL BRUCE WILLIS ROSARIO DAWSON

Disney

MARK WAHLBERG

globus

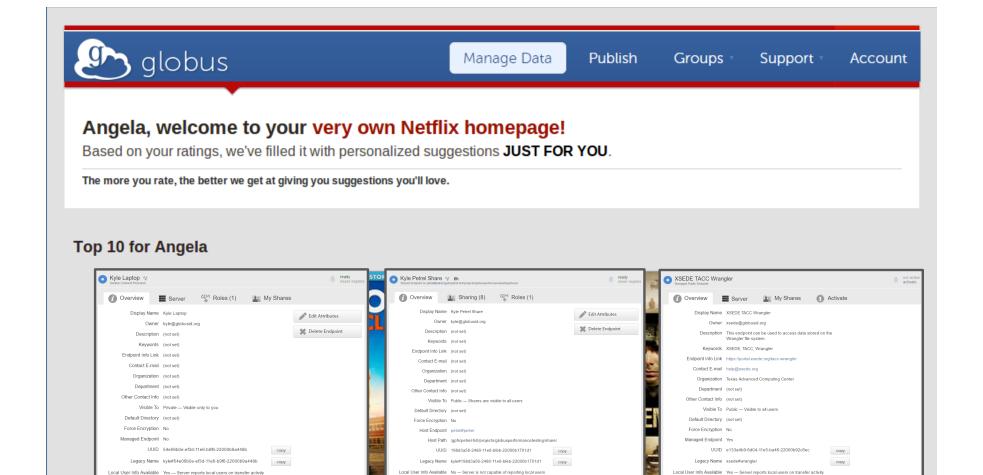Manage Data · Publish · Groups · Support · Account

Angela, welcome to your very own Netflix homepage!

Based on your ratings, we've filled it with personalized suggestions **JUST FOR YOU**.

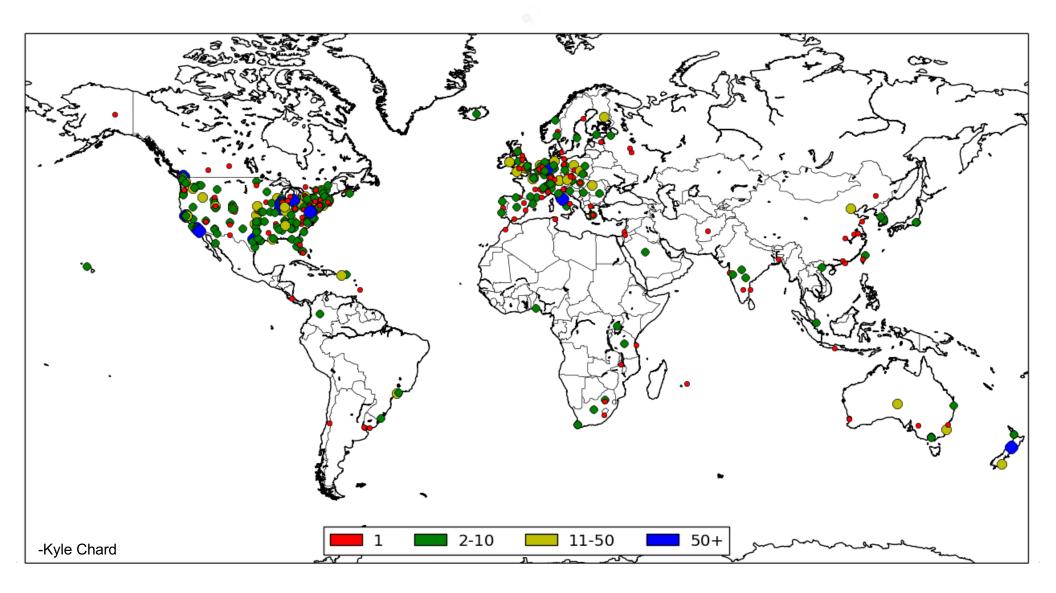The more you rate, the better we get at giving you suggestions you'll love.
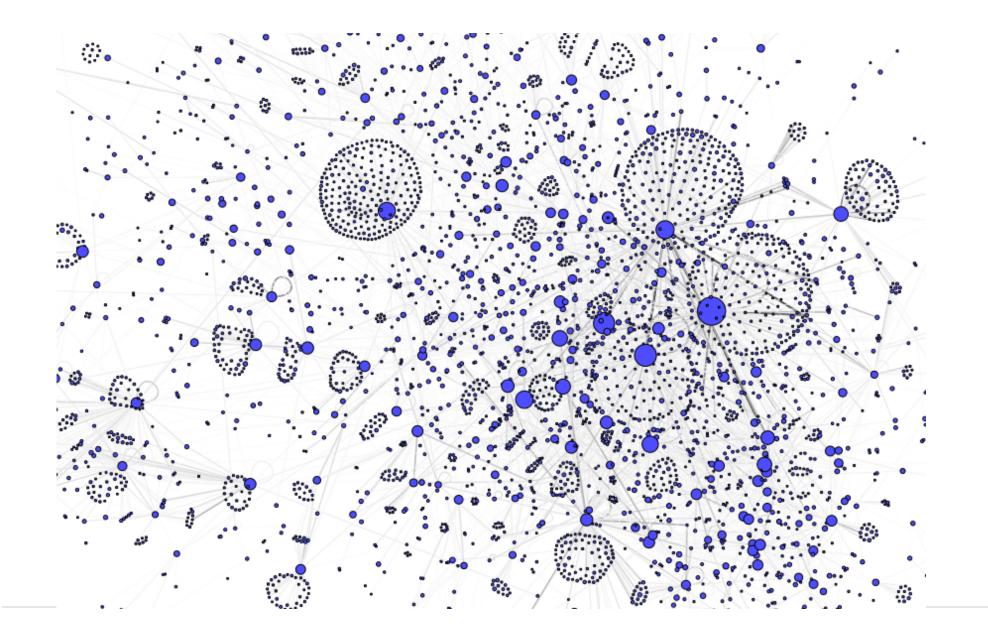
Top 10 for Angela

Kyle Laptop
Globus Connect Personal
ready never expires

Overview	Server	Roles (1)	My Shares

Display Name	Kyle Laptop
Owner	kyle@globusid.org
Description	(not set)
Keywords	(not set)
Endpoint Info Link	(not set)
Contact E-mail	(not set)
Organization	(not set)
Department	(not set)
Other Contact Info	(not set)
Visible To	Private — Visible only to you
Default Directory	(not set)
Force Encryption	No
Managed Endpoint	No
UUID	54e06b0e-ef3d-11e6-b9f6-22000b9a448b
Legacy Name	kyle#54e06b0e-ef3d-11e6-b9f6-22000b9a448b
Local User Info Available	Yes — Server reports local users on transfer activity

Edit Attributes · Delete Endpoint

Kyle Petrel Share
Shared Endpoint on petrel#petrel /gpfs/petrel-fs0/projects/globusperformancetesting/share/
ready never expires

Overview	Sharing (8)	Roles (1)

Display Name	Kyle Petrel Share
Owner	kyle@globusid.org
Description	(not set)
Keywords	(not set)
Endpoint Info Link	(not set)
Contact E-mail	(not set)
Organization	(not set)
Department	(not set)
Other Contact Info	(not set)
Visible To	Public — Shares are visible to all users
Default Directory	(not set)
Force Encryption	No
Host Endpoint	petrel#petrel
Host Path	/gpfs/petrel-fs0/projects/globusperformancetesting/share/
UUID	198d3a56-2480-11e6-bfeb-22000b1701d1
Legacy Name	kyle#198d3a56-2480-11e6-bfeb-22000b1701d1
Local User Info Available	No — Server is not capable of reporting local users

Edit Attributes · Delete Endpoint

XSEDE TACC Wrangler
Managed Public Endpoint
not active activate

Overview	Server	My Shares	Activate

Display Name	XSEDE TACC Wrangler
Owner	xsede@globusid.org
Description	This endpoint can be used to access data stored on the Wrangler file system.
Keywords	XSEDE, TACC, Wrangler
Endpoint Info Link	https://portal.xsede.org/tacc-wrangler
Contact E-mail	help@xsede.org
Organization	Texas Advanced Computing Center
Department	(not set)
Other Contact Info	(not set)
Visible To	Public — Visible to all users
Default Directory	(not set)
Force Encryption	No
Managed Endpoint	Yes
UUID	e133a4b9-6d04-11e5-ba46-22000b92c6ec
Legacy Name	xsede#wrangler
Local User Info Available	Yes — Server reports local users on transfer activity

Movies Popular on Netflix

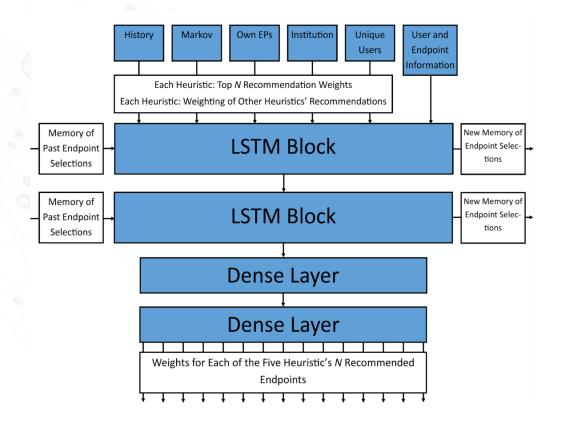|						
🟥	1	🟩	2-10	🟨	11-50	🟦	50+

-Kyle Chard

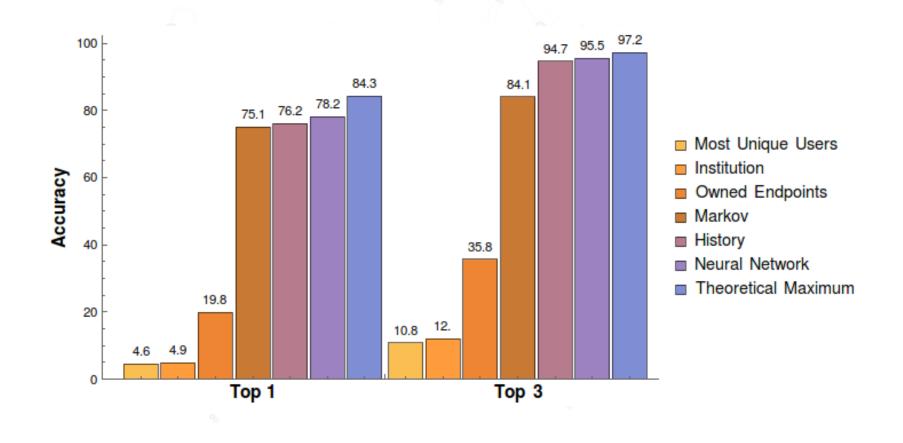# Long Tailed Distributions

Endpoint Prediction Heuristics

- **History**: The most likely source (S) / destination (D) endpoint as 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 for a user is the most popular endpoint at that user's institution
- **Endpoint Ownership**: The most likely endpoint is the most recent endpoint the user created

Ensemble Methods:
Using Recurrent Neural Network

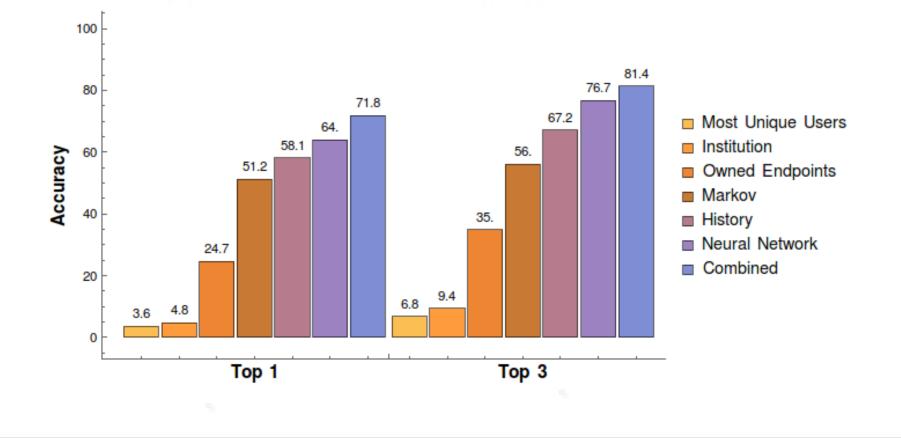- Combine multiple heuristics into a better heuristic
  - Simple example: output the top recommendation of each heuristic
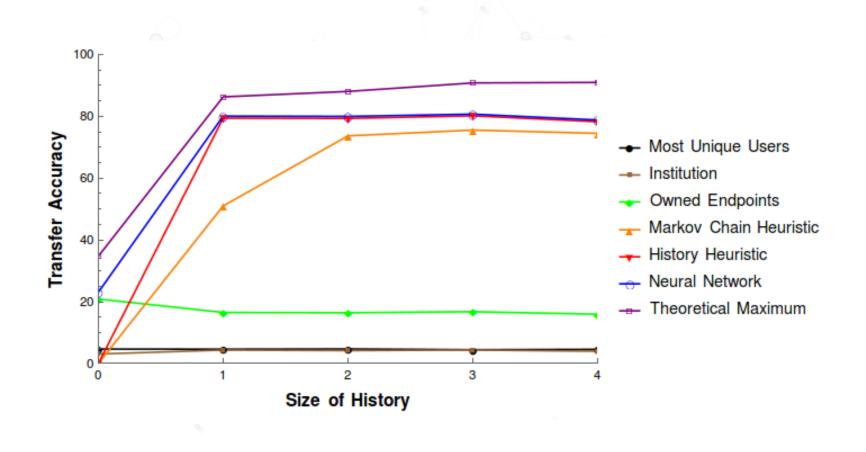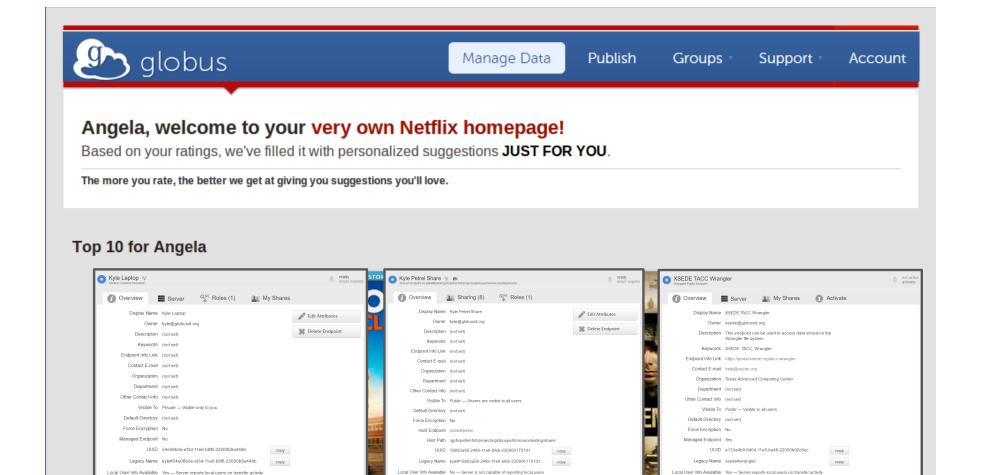
Results: Transfer Accuracy

Results: User Accuracy

Cold Start: Predictions for New Users

globus

Manage Data   Publish   Groups   Support   Account

Angela, welcome to your **very own Netflix homepage!**
Based on your ratings, we've filled it with personalized suggestions **JUST FOR YOU**.

The more you rate, the better we get at giving you suggestions you'll love.

Top 10 for Angela

Kyle Laptop
Globus Connect Personal
ready
never expires

Overview | Server | Roles (1) | My Shares

Display Name	Kyle Laptop
Owner	kyle@globusid.org
Description	(not set)
Keywords	(not set)
Endpoint Info Link	(not set)
Contact E-mail	(not set)
Organization	(not set)
Department	(not set)
Other Contact Info	(not set)
Visible To	Private — Visible only to you
Default Directory	(not set)
Force Encryption	No
Managed Endpoint	No
UUID	54e06b0e-ef3d-11e6-b9f6-22000b9a448b copy
Legacy Name	kyle#54e06b0e-ef3d-11e6-b9f6-22000b9a448b copy
Local User Info Available	Yes — Server reports local users on transfer activity

Edit Attributes
Delete Endpoint

Kyle Petrel Share
Shared Endpoint on petrel#petrel /gpfs/petrel-fs0/projects/globusperformancetesting/share/
ready
never expires

Overview | Sharing (8) | Roles (1)

Display Name	Kyle Petrel Share
Owner	kyle@globusid.org
Description	(not set)
Keywords	(not set)
Endpoint Info Link	(not set)
Contact E-mail	(not set)
Organization	(not set)
Department	(not set)
Other Contact Info	(not set)
Visible To	Public — Shares are visible to all users
Default Directory	(not set)
Force Encryption	No
Host Endpoint	petrel#petrel
Host Path	/gpfs/petrel-fs0/projects/globusperformancetesting/share/
UUID	198d3a56-2480-11e6-bfeb-22000b1701d1 copy
Legacy Name	kyle#198d3a56-2480-11e6-bfeb-22000b1701d1 copy
Local User Info Available	No — Server is not capable of reporting local users

Edit Attributes
Delete Endpoint

XSEDE TACC Wrangler
Managed Public Endpoint
not active
activate

Overview | Server | My Shares | Activate

Display Name	XSEDE TACC Wrangler
Owner	xsede@globusid.org
Description	This endpoint can be used to access data stored on the Wrangler file system.
Keywords	XSEDE, TACC, Wrangler
Endpoint Info Link	https://portal.xsede.org/tacc-wrangler
Contact E-mail	help@xsede.org
Organization	Texas Advanced Computing Center
Department	(not set)
Other Contact Info	(not set)
Visible To	Public — Visible to all users
Default Directory	(not set)
Force Encryption	No
Managed Endpoint	Yes
UUID	e133a4b9-6d04-11e5-ba46-22000b92c6ec copy
Legacy Name	xsede#wrangler copy
Local User Info Available	Yes — Server reports local users on transfer activity

Movies Popular on Netflix