Computation Institute

Big process for big data

Process automation for data-driven science

Ian Foster
Computation Institute
Argonne National Laboratory & The University of Chicago

Talk at 2nd Workshop on Understanding Climate Change from Data, Minneapolis, August 6, 2012

Argonne NATIONAL LABORATORY

www.ci.anl.gov
www.ci.uchicago.edu

I'll talk about two things

- Center for Robust Decision making on Climate and Energy Policy (RDCEP) – www.rdcep.org

- Research data management as a service: Globus Online – www.globusonline.org

RDCEP – www.rdcep.org

- **Research** to improve models, confront uncertainty, and advance computational methods
- **Development** of open tools for modeling, estimation, uncertainty characterization, and related tasks
- **Education** and research experience for high-school students and teachers, undergrads, grads, and postdocs
- **Products** for decision makers and other stakeholders

Some initial projects

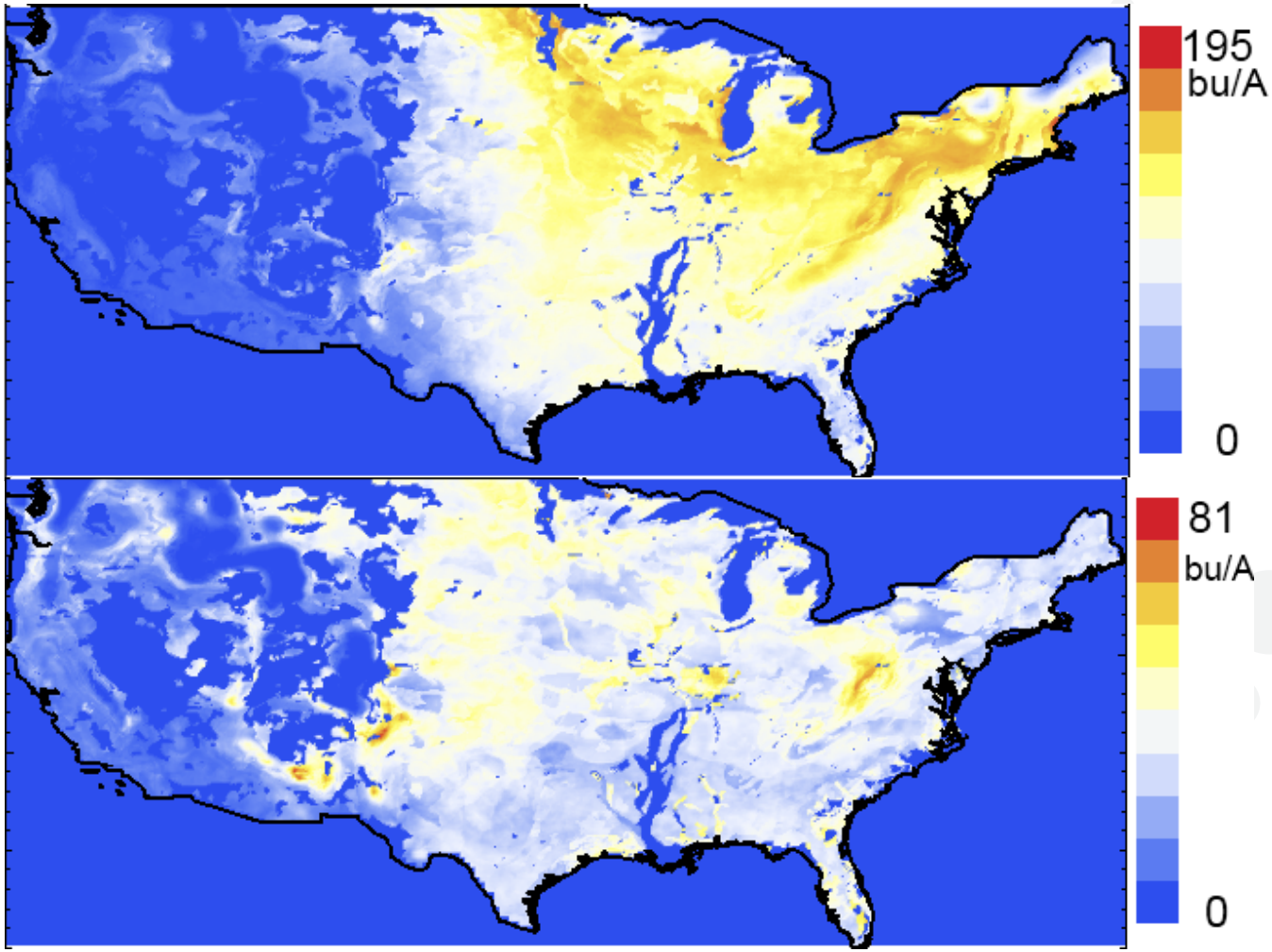
- CIM-EARTH modeling framework (Choi, Munson)
- Stochastic dynamic models (Cai, Judd, Lontzek)
- Characterizing parametric uncertainty (Elliott, Munson)
- Carbon leakage (Elliott, Foster, Kortum, Munson, Weisbach)
- Distributional impacts of climate change (Elliott, Fullerton, Rao)
- Land use decisions under uncertainty (Hertel, Judd, Steinbuks)
- Robust decision making for energy (Hansen, Munson, Sanstad)
- Climate model emulation (McInerney, Moyer, Stein)
- Uncertainty in downscaling for impacts studies (Moyer et al.)
- Partial equilibrium land use change for biofuels (Elliott et al.)
- Analysis of Illinois Renewable Portfolio Standard (Moyer et al.)
- Social Cost of Carbon studies (Moyer, Weisbach, et al.)

Climate impacts in agriculture via gridded crop simulations with the pDSSAT model

Example: 1980-2009 rainfed, fully fertilized corn yields at 5'

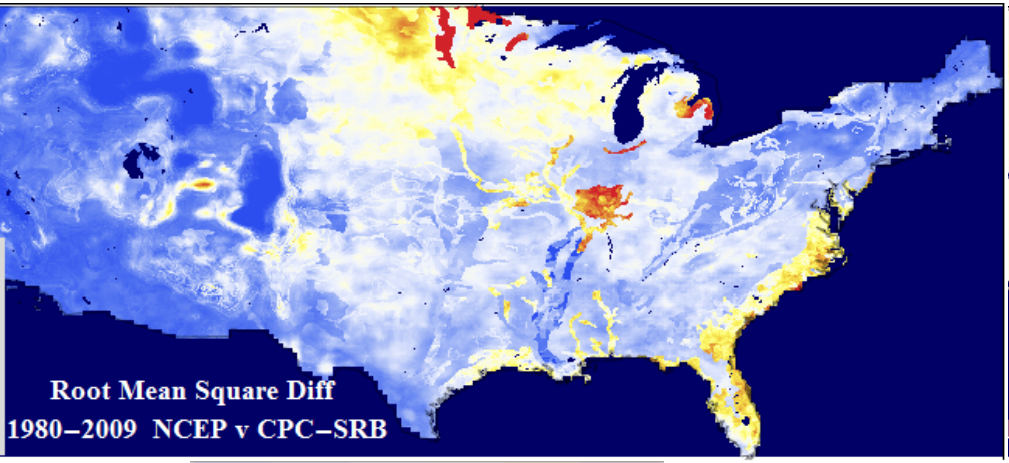
Median

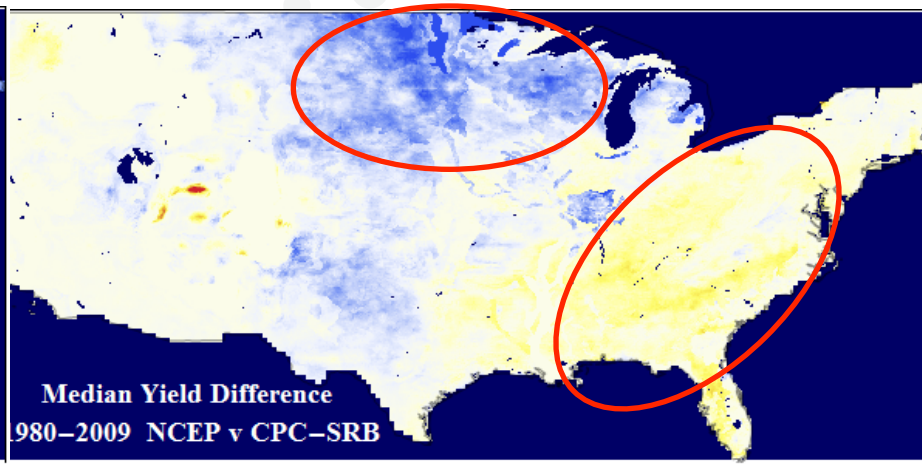
Standard deviation



Reanalysis vs. interpolated observations NCEP CFSR Temps; CPC Precip; SRB Solar

- To characterize the uncertainties that are expected from performing these simulations globally with reanalysis products, we compare the NCEP CFSR with interpolated observation products like CPC and SRB
- For rainfed maize, pure reanalysis-based simulations tend to under-predict yields across a wide swath of the US corn belt relative to simulations using observations.
- They tend to over-predict over most of the area East of the Mississippi, most notably along SE coast and throughout Florida



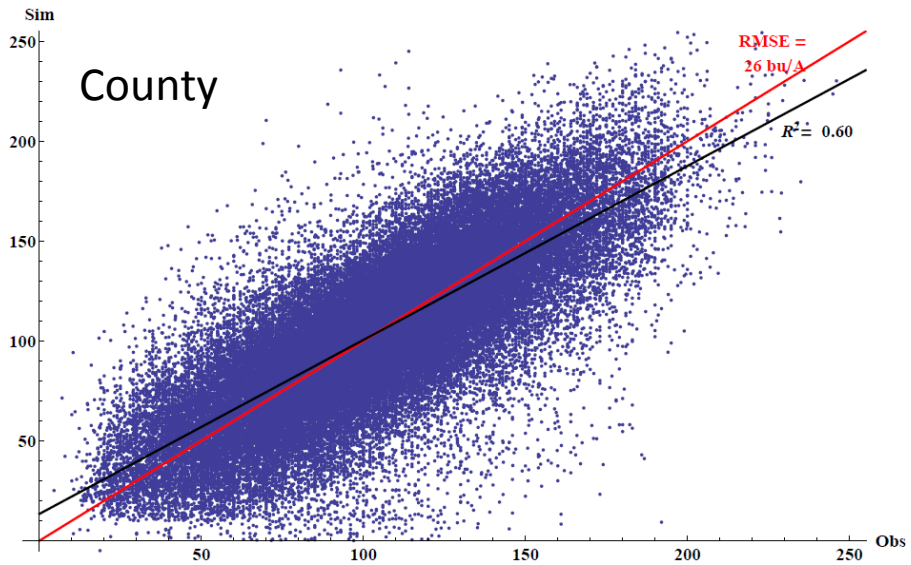


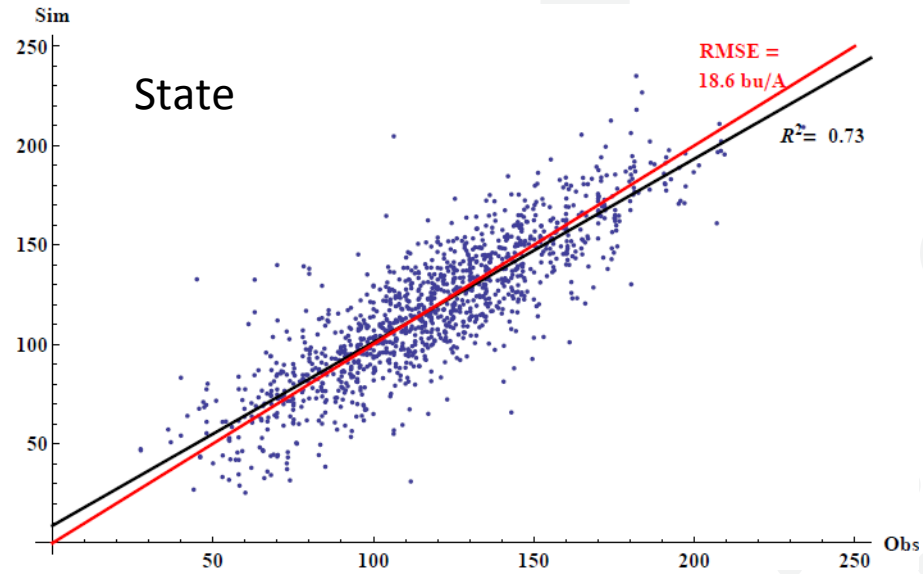
Bu/acre 0 45 90

Bu/acre -60 0 60

Aggregation and Validation

- Simulated corn yields aggregated to county and state level and plotted against USDA survey data.
  - 1980-2009; 60423 data points from 2373 counties; 1225 from 41 states.
  - RMSE and R^2 between simulated and observed aggregate yields:
    - County: 26 bu/A (25% of sample mean), R^2 = 0.60
    - State: 18.6 bu/A (15.7% of sample mean), R^2 = 0.73
    - National: 13.6 bu/A (10.8% of sample mean)





Accuracy improves as we aggregate from county to state to national level

I'll talk about two things

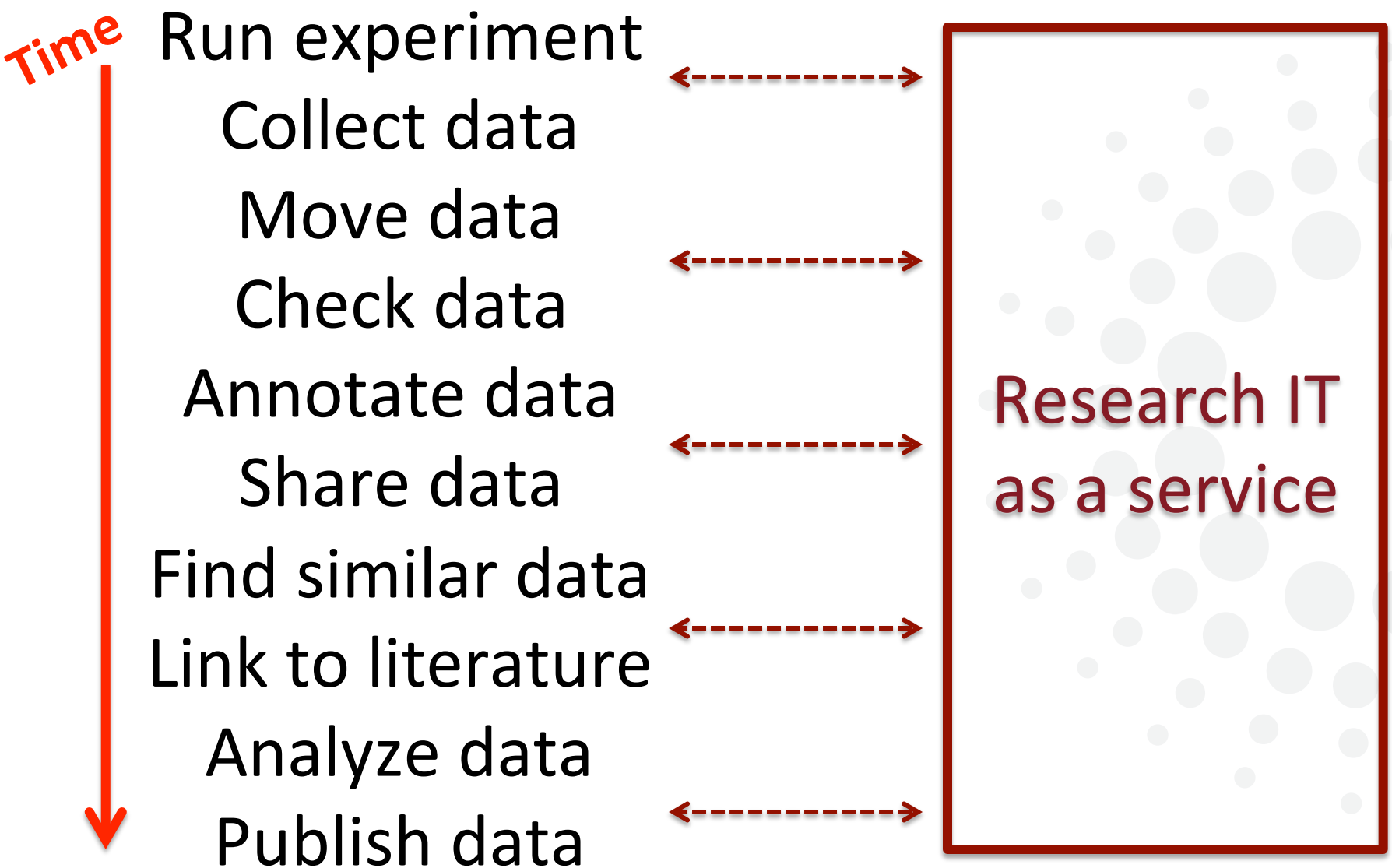
- Center for Robust Decision making on Climate and Energy Policy – www.rdcep.org

- **Research data management as a service: Globus Online – www.globusonline.org**

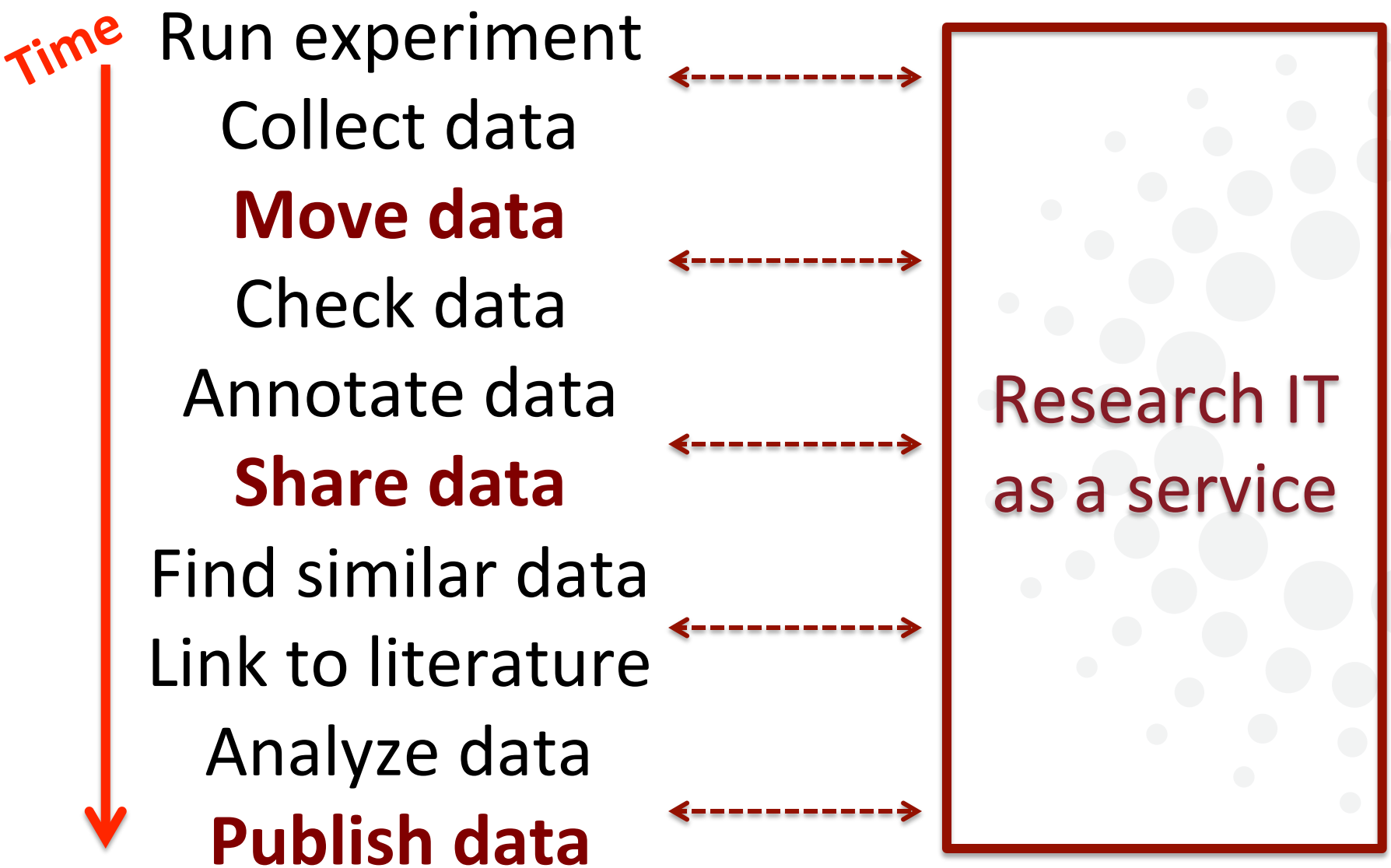
The climate data problem

- Climate data (observations + simulations) are highly distributed
- Climate data are big and getting bigger
- Mundane tasks associated with preparing for and performing analyses take more & more time

Process automation for science



Process automation for science



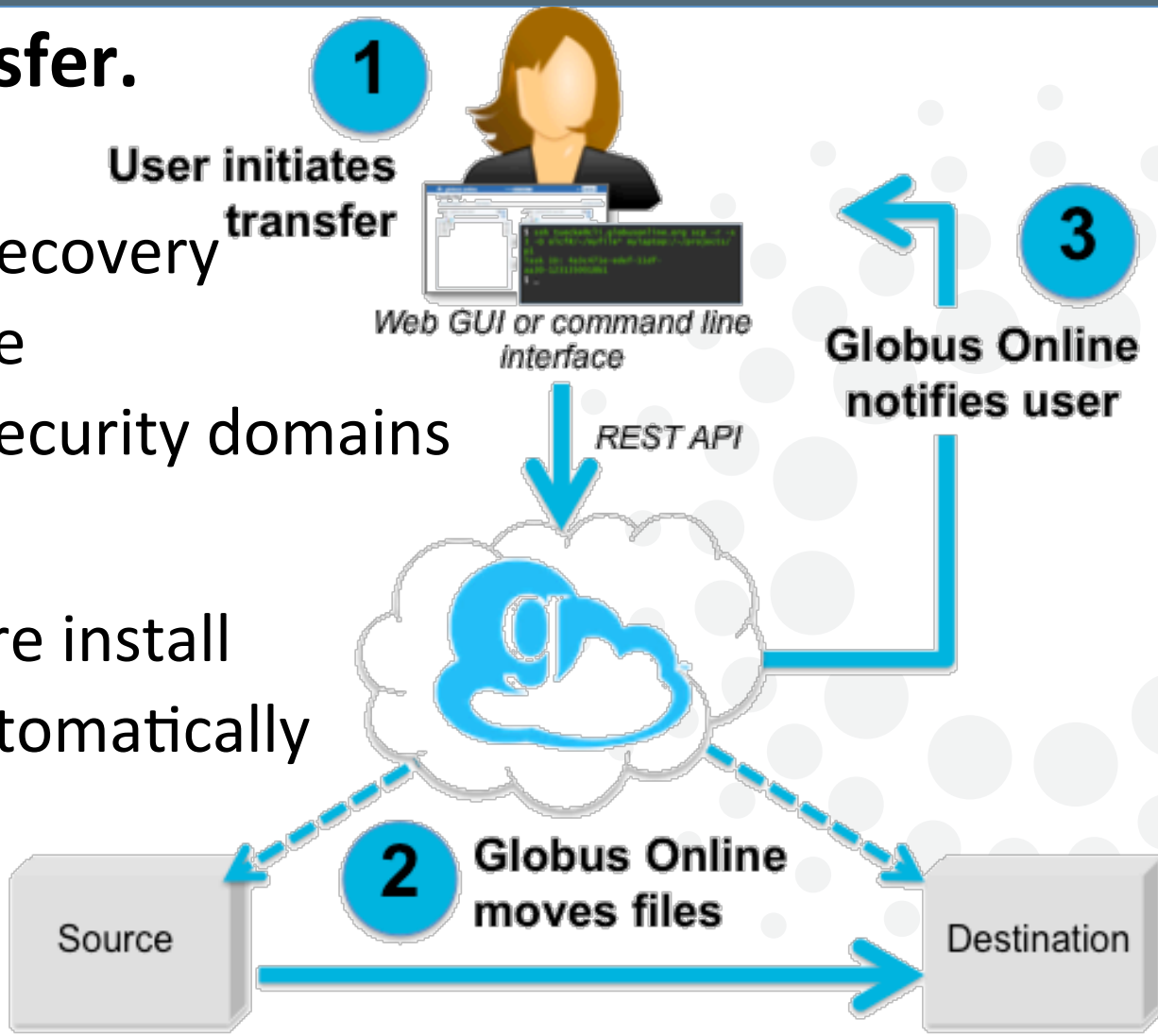
Eppur si muove ("and yet it moves") …

- Galileo said this about the **earth** (the earth is not the center of the universe)

- Same observation holds about **data**, which is often in the "wrong place" for various reasons

Globus Online: Data transfer as SaaS

- **Reliable file transfer.**
  - Fire-and-forget
  - Automatic fault recovery
  - High performance
  - Across multiple security domains
- **No IT required.**
  - No client software install
  - New features automatically available
  - Consolidated support and troubleshooting



Works with existing GridFTP servers; also Globus Connect

globus online

Manage Data | News & Events | About | Support | Log In | Sign Up

Reliable, high-performance, secure file transfer.

Move files fast. No IT required.

WATCH A VIDEO

Globus Online in a nutshell

6,194,694,928 MB TRANSFERRED

Why Use Globus Online?
See how easy file transfer can be

For HPC Resource Owners
Enable Globus Online for your users

For Developers
Integrate with Globus Online

globus online

Manage Data | Support | ian

start transfer | view transfer activity | manage endpoints | dashboard

Transfer Files

Get Globus Connect

Turn your computer into an endpoint.

Endpoint isimipgo#us-archive Go

Path /~/data/ Go

select all | none up one folder refresh list

GFDL-ESM2M	Folder
HadGEM2-ES	Folder
IPSL-CM5A-LR	Folder
MIROC-ESM-CHEM	Folder
NorESM1-M	Folder
RCP_CO2_Data	Folder
SSP	Folder
abstracts	Folder
global.mean	Folder
monthly.mean	Folder
test_neill_2Mb.txt	1.95MB
upload_area_mkdirs_slim.sh	2.01kB

Endpoint xsede#trestles Go

Path /~/Data/ Go

select all | none up one folder refresh list

No files available to display

t isimipgo#us-archive

more options

Label This Transfer

This will be displayed in your transfer activity

globus online

Manage Data | Support | ian

start transfer | view transfer activity | manage endpoints | dashboard

Transfer Files

Get Globus Connect
Turn your computer into an endpoint.

Endpoint isimipgo#us-archive Go

Path /~/data/ Go

Endpoint xsede#trestles Go

Path /~/Data/ Go

select all | none up one folder refresh list

GFDL-ESM2M	*Folder*
HadGEM2-ES	*Folder*
IPSL-CM5A-LR	*Folder*
MIROC-ESM-CHEM	*Folder*
NorESM1-M	*Folder*
RCP_CO2_Data	*Folder*
SSP	*Folder*
abstracts	*Folder*
global.mean	*Folder*
monthly.mean	*Folder*
test_neill_2Mb.txt	*1.95MB*
upload_area_mkdirs_slim.sh	*2.01kB*

select all | none up one folder refresh list

No files available to display

more options

Label This Transfer

This will be displayed in your transfer activity.

globus online

Manage Data | Support | ian

start transfer | view transfer activity | manage endpoints | dashboard

Transfer Request Submitted Successfully. Task ID:01fab246-dfe8-11e1-bf56-1231380b8963

Transfer Files

Get Globus Connect

Turn your computer into an endpoint.

Endpoint isimipgo#us-archive Go

Path /~/data/ Go

Endpoint xsede#trestles

Path /~/Data/

select all | none up one folder refresh list

GFDL-ESM2M	*Folder*
HadGEM2-ES	*Folder*
IPSL-CM5A-LR	*Folder*
MIROC-ESM-CHEM	*Folder*
NorESM1-M	*Folder*
RCP_CO2_Data	*Folder*
SSP	*Folder*
abstracts	*Folder*
global.mean	*Folder*
monthly.mean	*Folder*
test_neill_2Mb.txt	*1.95MB*
upload_area_mkdirs_slim.sh	*2.01kB*

select all | none up one folder refresh list

No files available to display

Transfer Activity

Cancel

1 of 6

View 25 Records

	Status	Label	Task Progress	Completion Time	Request Time
		Task Id:5cbcf6bc-dfe3-11e1-bf56-1231380b8963	197 / 1026		08/06/2012 11:25 AM
		Task Id:1fe43dd4-df1d-11e1-bf56-1231380b8963	4362 / 4362	08/06/2012 03:34 AM	08/05/2012 11:46 AM
		Task Id:d3618606-df1c-11e1-bf56-1231380b8963	1 / 1	08/05/2012 11:45 AM	08/05/2012 11:44 AM
		Task Id:b3c3dba0-defe-11e1-bf56-1231380b8963	0 / 1 / 1	08/05/2012 11:42 AM	08/05/2012 08:08 AM
		Task Id:a37697c4-defe-11e1-bf56-1231380b8963	624 / 624	08/05/2012 10:08 AM	08/05/2012 08:08 AM
		Task Id:d370fd7c-deac-11e1-bf56-1231380b8963	3738 / 3738	08/05/2012 07:51 AM	08/04/2012 10:22 PM
		Task Id:278af9f2-dea9-11e1-bf56-1231380b8963	624 / 624	08/05/2012 12:12 AM	08/04/2012 09:56 PM
		Task Id:6d8aa43c-dcfd-11e1-bf56-1231380b8963	299 / 717 / 1016	08/02/2012 07:00 PM	08/02/2012 06:54 PM
		Task Id:c150b78e-dcfb-11e1-bf56-1231380b8963	96 / 528 / 624	08/02/2012 06:49 PM	08/02/2012 06:42 PM
		Task Id:e84f6718-dcf6-11e1-bf56-1231380b8963	11 / 11	08/02/2012 06:08 PM	08/02/2012 06:08 PM
		Task Id:784d8008-dcf6-11e1-bf56-1231380b8963	11 / 11	08/02/2012 06:04 PM	08/02/2012 06:04 PM
		Task Id:33f4206a-dcf6-11e1-bf56-1231380b8963	11 / 11	08/02/2012 06:03 PM	08/02/2012 06:02 PM
		Task Id:60950aae-dcf5-11e1-bf56-1231380b8963	9 / 2 / 11	08/02/2012 06:01 PM	08/02/2012 05:57 PM
		Task Id:a4273bfc-dcf0-11e1-bf56-1231380b8963	11 / 11	08/02/2012 09:13 PM	08/02/2012 05:23 PM
		Task Id:77993940-d2c4-11e1-bf56-1231380b8963	1339 / 1339	07/21/2012 06:18 PM	07/20/2012 06:41 PM
		Task Id:bff65a14-d135-11e1-bf56-1231380b8963	1556 / 1556	07/19/2012 06:27 PM	07/18/2012 07:07 PM
		Sync test	1078 / 1078	07/18/2012 10:17 AM	07/18/2012 10:16 AM
		Task Id:be22047c-d0ea-11e1-bf56-1231380b8963	1077 / 1077	07/18/2012 10:13 AM	07/18/2012 10:11 AM
		Task Id:d927feec-d0e6-11e1-bf56-1231380b8963	1 / 1	07/18/2012 10:14 AM	07/18/2012 09:44 AM
		Task Id:19b4e54a-d0da-11e1-bf56-1231380b8963	1 / 1	07/18/2012 08:12 AM	07/18/2012 08:11 AM

Transfer Activity

Cancel

1 of 6

	Status	Label	Task Progress	Completion Time
		Task Id:5c		
		Task Id:1f		
		Task Id:d3		
		Task Id:b3		
		Task Id:a3		
		Task Id:d3		
		Task Id:27		
		Task Id:6c		
		Task Id:c1		
		Task Id:e8		
		Task Id:78		
		Task Id:33		
		Task Id:60		
		Task Id:a4273bfc-dcf0-11e1-bf56-1231380b8963	11 / 11	08/02/2012 09:13 PM
		Task Id:77993940-d2c4-11e1-bf56-1231380b8963	1339 / 1339	07/21/2012 06:18 PM
		Task Id:bff65a14-d135-11e1-bf56-1231380b8963	1556 / 1556	07/19/2012 06:27 PM
		Sync test	1078 / 1078	07/18/2012 10:17 AM
		Task Id:be22047c-d0ea-11e1-bf56-1231380b8963	1077 / 1077	07/18/2012 10:13 AM

Transfer Details

Task ID: 5cbcf6bc-dfe3-11e1-bf56-1231380b8963

Status	ACTIVE	Request Time	08/06/2012 11:25 AM
Origin	isimipgo#us-archive	Deadline	08/07/2012 11:25 AM
Destination	xsede#trestles	Completion Time	
User	ian	Bytes Transferred	18583448810
Directories	10	Files	3731

Task Statistics

Pending	3342	Succeeded	399
Cancelled	0	Expired	0
Failed	0	Retrying	0

View Event Log

Globus Online Notification

To: Ian Foster

Task a37697c4-defe-11e1-bf56-1231380b8963: SUCCEEDED

=== Task Details ===
Task ID : a37697c4-defe-11e1-bf56-1231380b8963
Task Type : TRANSFER
Parent Task ID : n/a
Status : SUCCEEDED
Request Time : 2012-08-05 13:08:24Z
Deadline : 2012-08-08 14:10:32Z
Completion Time : 2012-08-05 15:08:31Z
Total Tasks : 624
Tasks Successful : 624
Tasks Expired : 0
Tasks Canceled : 0
Tasks Failed : 0
Tasks Pending : 0
Tasks Retrying : 0
Command : API 0.10 GO
Label : n/a
Data Encryption : No
Checksum Verification: No
Sync Level : 0
Files : 623
Files Skipped : 0
Directories : 1
Bytes Transferred : 440658434897
Bytes Checksummed : 0
MBits/sec : 489.213
Faults : 1

2 hours

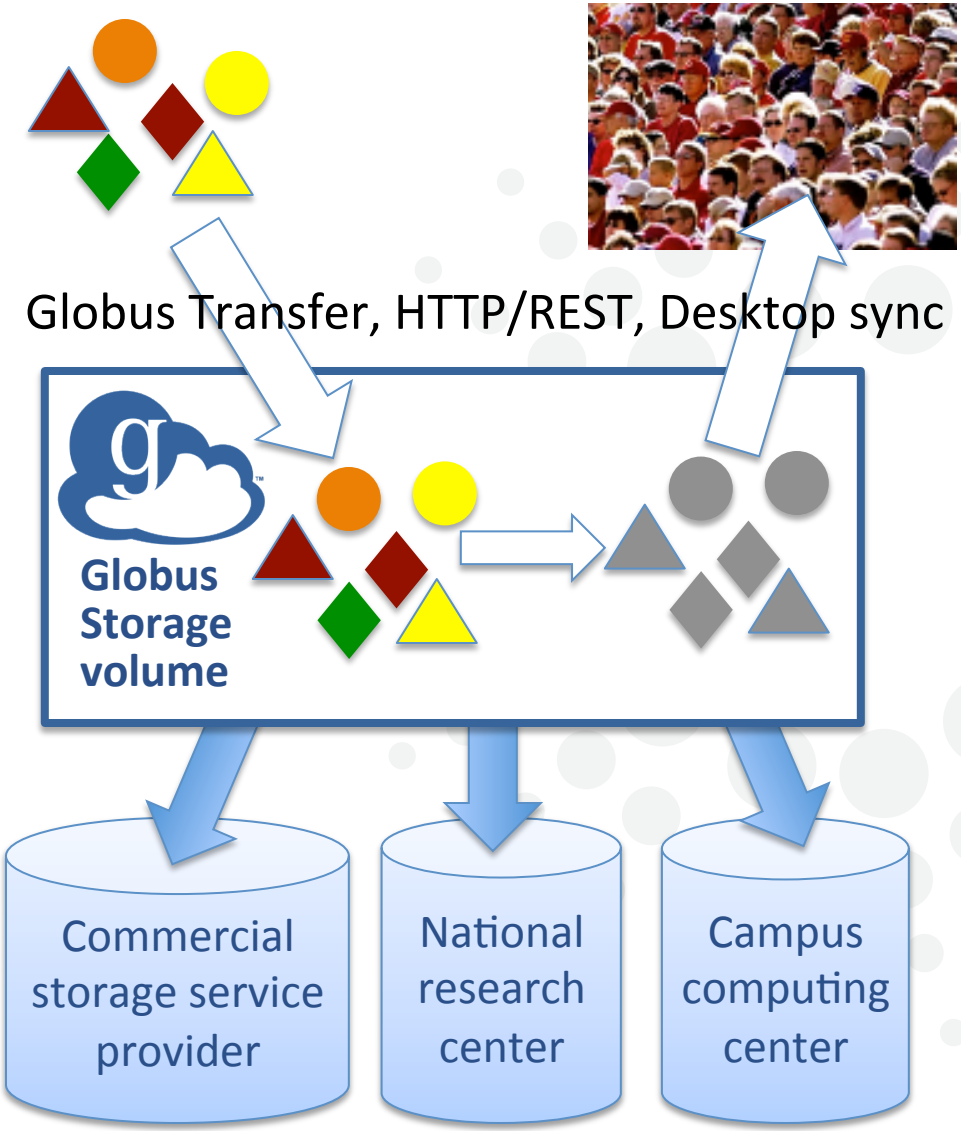
440 GB @ 60 MB/sec

Globus Transfer to date

- In 18 months
  - 5,000 users
  - 6 PB moved
  - 500M files
  - 99.9% uptime
- Broad adoption
  - Experimental facilities
  - Supercomputers
  - Campuses
  - Individuals
  - Projects

Globus Storage: For when you want to …

- **Place** your data where you want
- **Access** it from anywhere via different protocols
- Update it, version it, and take snapshots
- **Share** versions with who you want
- **Synchronize** among locations



Computation Institute

Thank you!

rdcep.org

globusonline.org

foster@anl.gov
foster@uchicago.edu