Counting Your Assets

Don’t Forget About Data!

James (Jim) Wilgenbusch

Director of Research Computing:

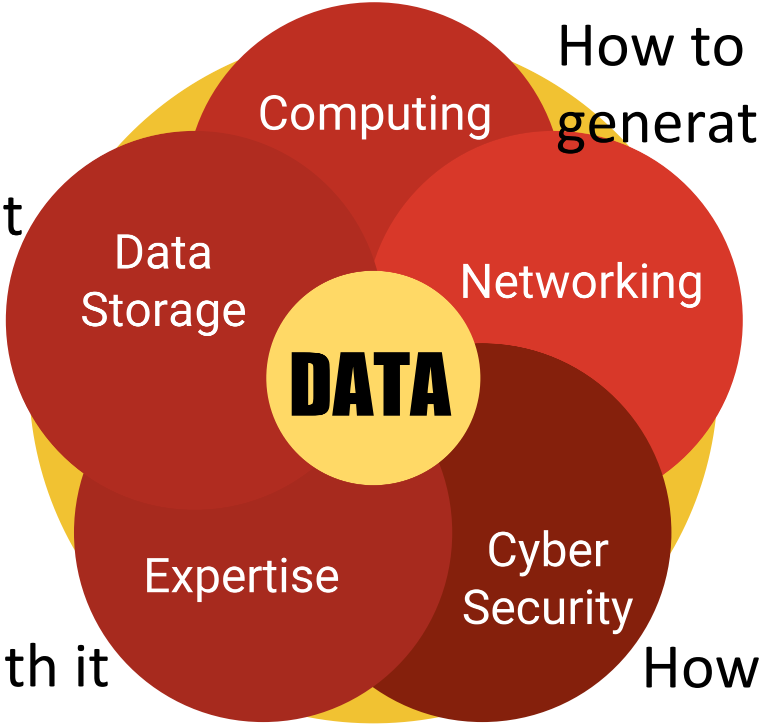
- Minnesota Supercomputing Institute
- Minnesota Informatics Institute
- U-Spatial

Office of the Vice President for Research

UNIVERSITY OF MINNESOTA

Driven to Discover®

Key Components of Research CI



How to understand/ generate it

How to retain it

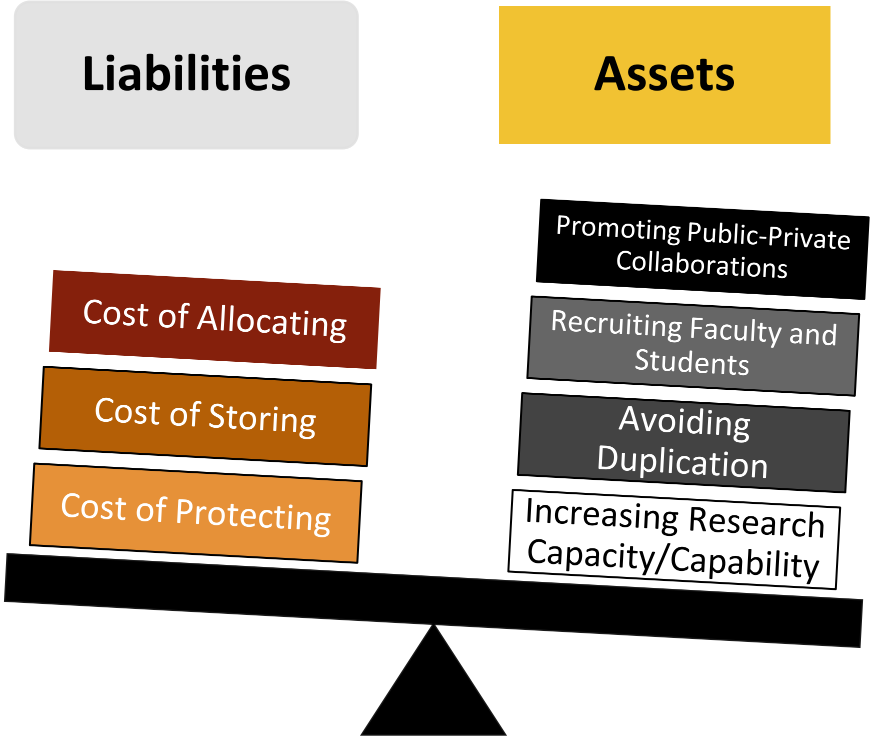
How to move it

What to do with it

How to protect it

Paradigm Shift

Data as a Liability → Data as an Asset


Liabilities
Assets
Cost of Allocating
Cost of Storing
Cost of Protecting
Promoting Public-Private Collaborations
Recruiting Faculty and Students
Avoiding Duplication
Increasing Research Capacity/Capability

Challenges are Real

- Breath of research data is enormous
  - Socioeconomic to stellar evolution
- Size of some data sets are daunting
  - 100 TB to Multi-PB data sets are common
- Data use agreements are onerous
  - Lots of time to setup and penalties for violations

Data as an Asset:

Key Ingredients

- Sustained Research Infrastructure
- Well Trained Cyberinfrastructure Professionals Cyberpractitioners

MSI Computing and Data Storage Assets

Batch High Performance Computing

- Two Supercomputers
- 25,000 CPU Cores
- 230,400 GPU CUDA Cores
- 100 TB Memory
- Infiniband Network

Big Data Storage & Analysis

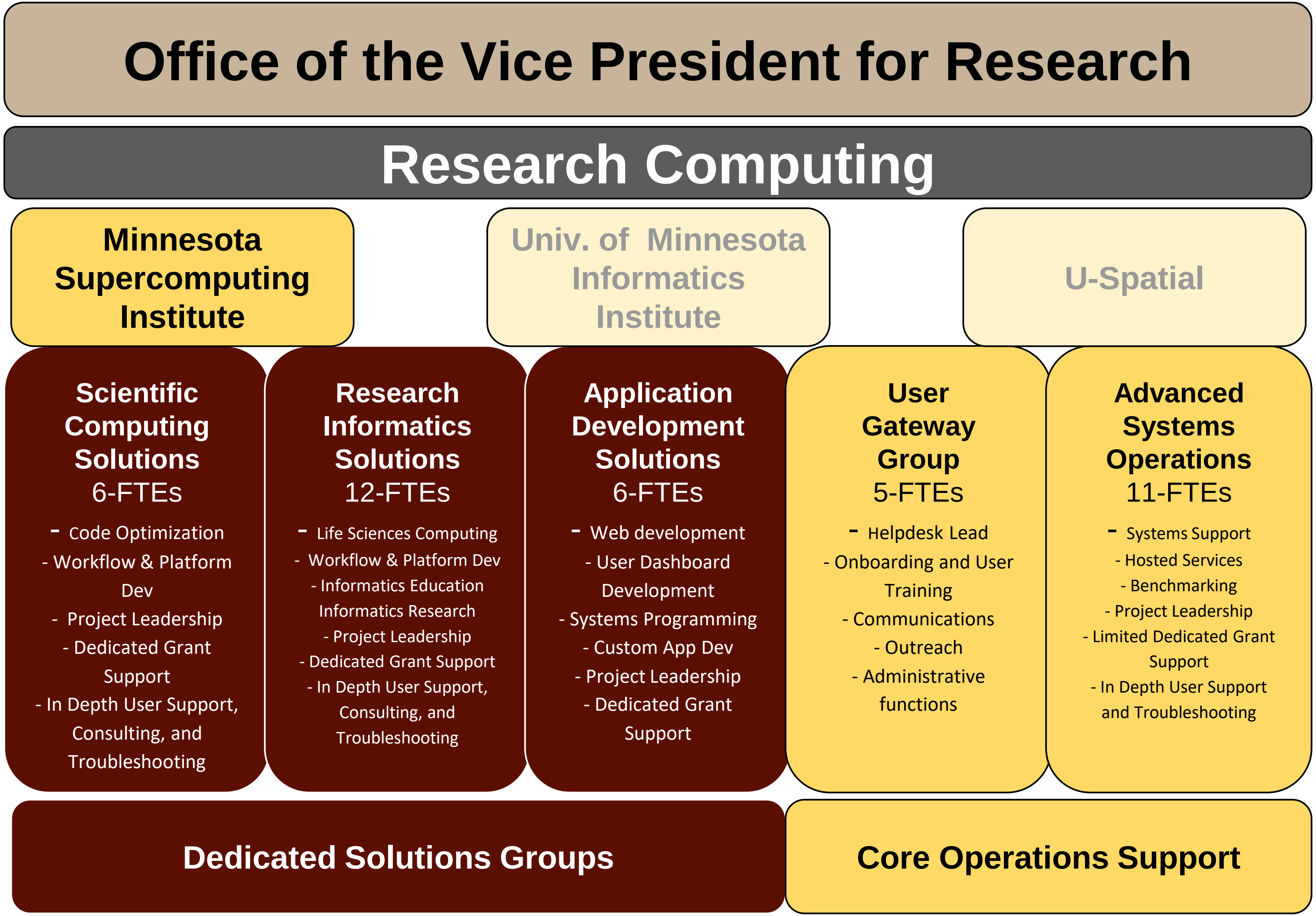
- 6 PB Primary High Performance
- 3 PB Second Tier
- 30 PB Archive Tape Library

Interactive & Cloud Computing

- Citrix VDI for Windows
- DCS Nice for Linux Desktops
- OpenStack for Secure Cloud
- 100 Gbps Campus Research Network
- Regional & National Optical Networks

Web Portals & Databases

- Galaxy for Multi-omics
- Jupyter Hub
- Custom Interfaces & Applications



Office of the Vice President for Research

Research Computing

Minnesota Supercomputing Institute

Univ. of Minnesota Informatics Institute

U-Spatial

Scientific Computing Solutions
6-FTEs

- Code Optimization
- Workflow & Platform Dev
- Project Leadership
- Dedicated Grant Support
- In Depth User Support, Consulting, and Troubleshooting

Research Informatics Solutions
12-FTEs

- Life Sciences Computing
- Workflow & Platform Dev
- Informatics Education Informatics Research
- Project Leadership
- Dedicated Grant Support
- In Depth User Support, Consulting, and Troubleshooting

Application Development Solutions
6-FTEs

- Web development
- User Dashboard Development
- Systems Programming
- Custom App Dev
- Project Leadership
- Dedicated Grant Support

User Gateway Group
5-FTEs

- Helpdesk Lead
- Onboarding and User Training
- Communications
- Outreach
- Administrative functions

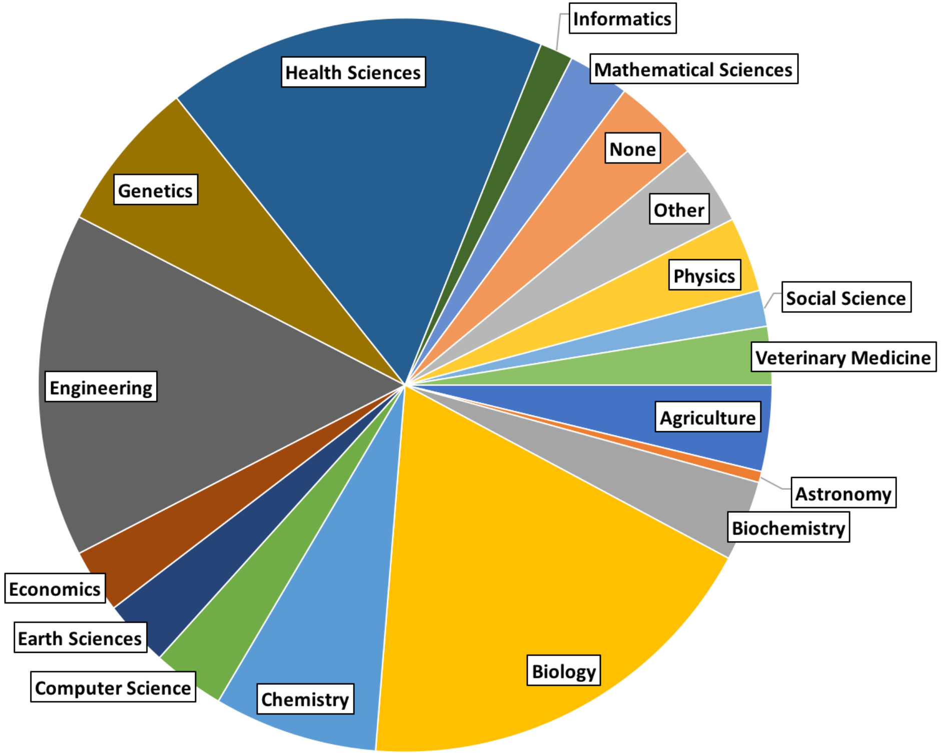
Advanced Systems Operations
11-FTEs

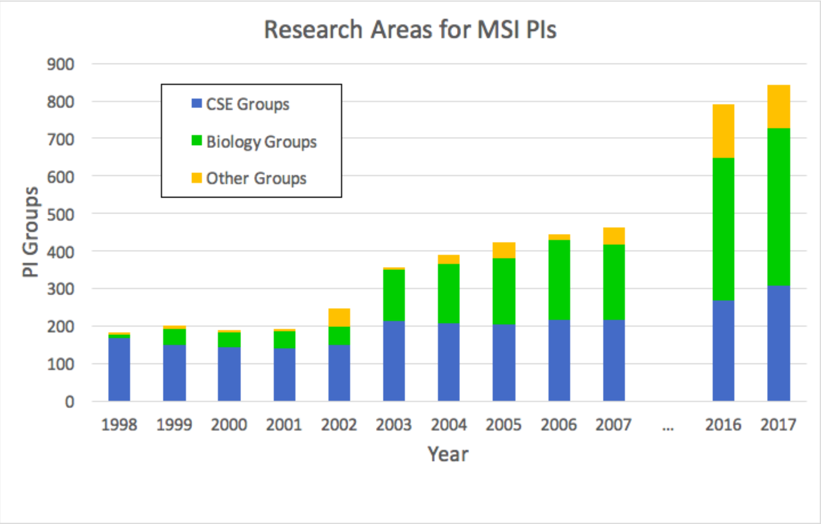
- Systems Support
- Hosted Services
- Benchmarking
- Project Leadership
- Limited Dedicated Grant Support
- In Depth User Support and Troubleshooting

Dedicated Solutions Groups

Core Operations Support

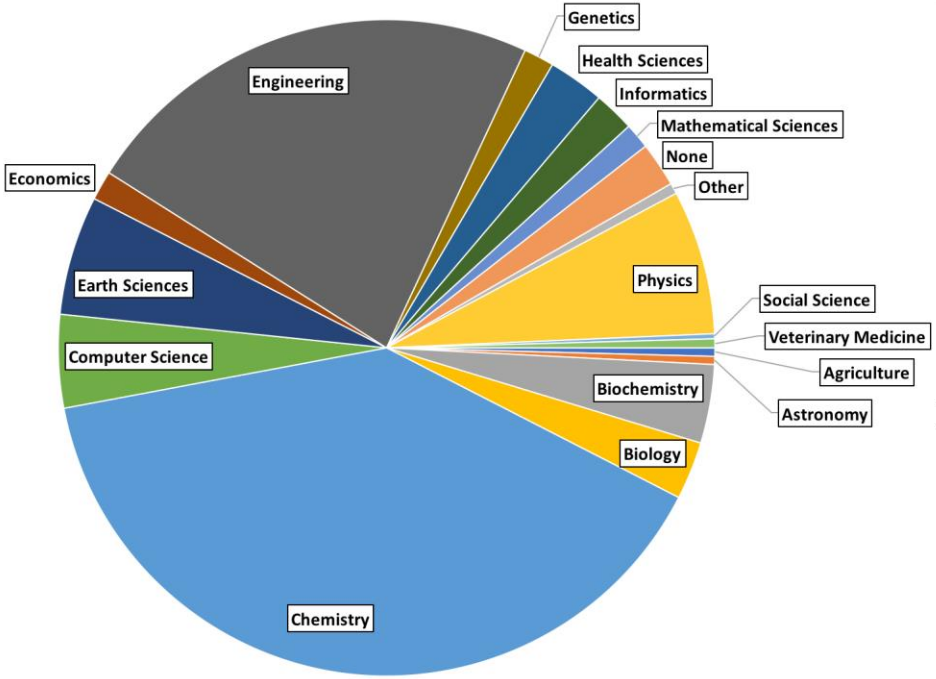
In 2018: 888 User Groups, 4,555 Active users



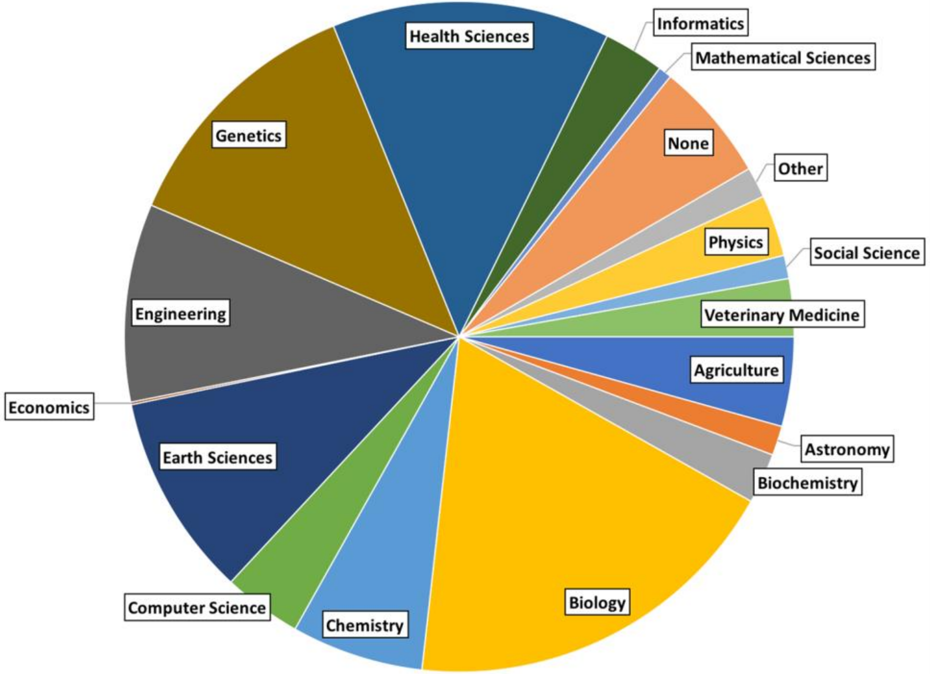


Biggest increasing in Life Sciences

Resource Utilization by Group

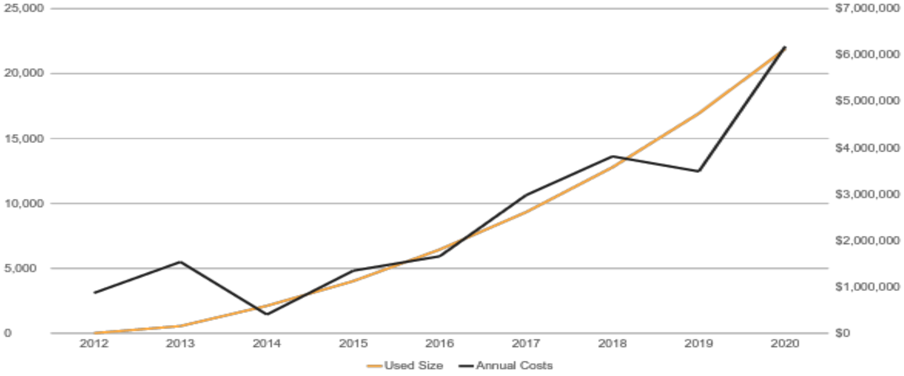


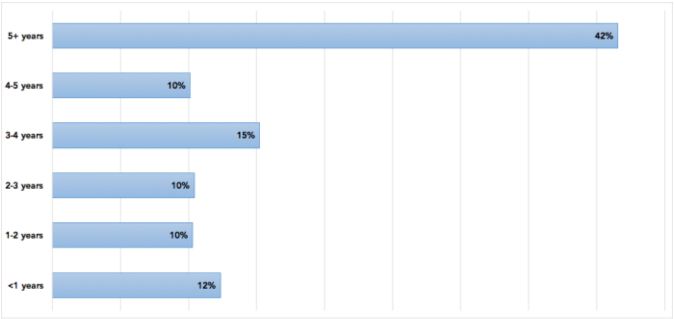
CPU Hours
150 Million Total



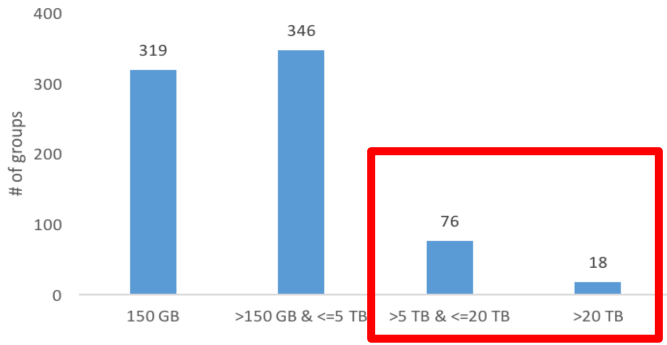
HPC Storage
1.5 PetaBytes

What’s the Problem?

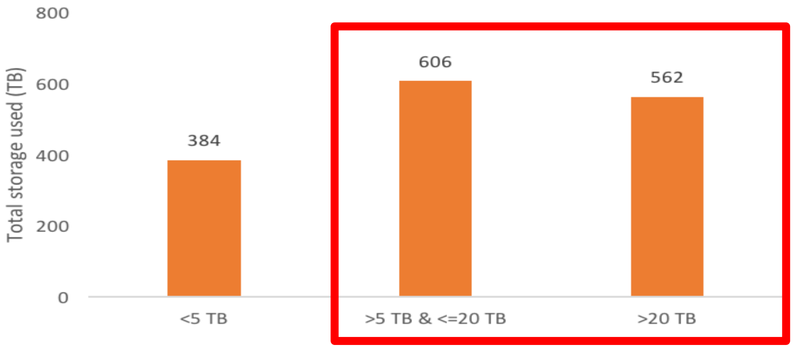




Number of Groups by Allocation Tier



Total Storage Used by Allocation Tier



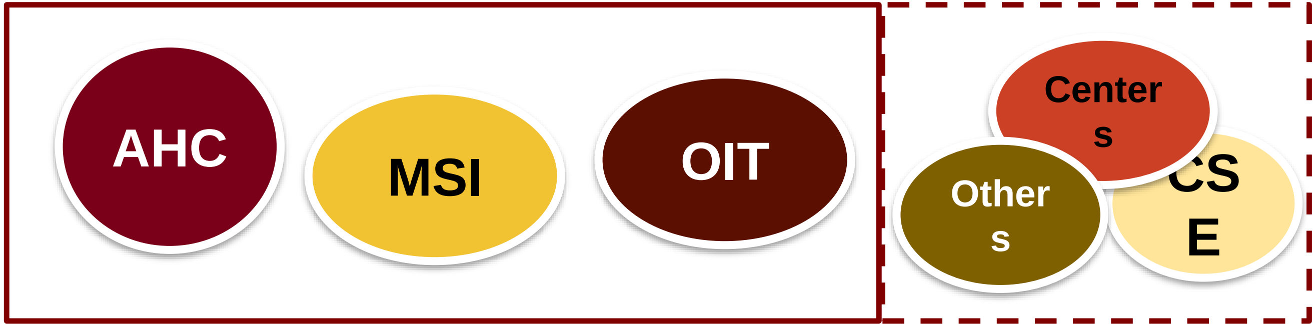
Data as an Asset:

Key Ingredients

- Sustained Research Infrastructure
- Well Trained Cyberinfrastructure Professionals Cyberpractitioners
- **Good Storage Governance that Spans Institutional Reporting Lines**

It Doesn’t Have to Be Scary

Assemble Stakeholders



AHC

MSI

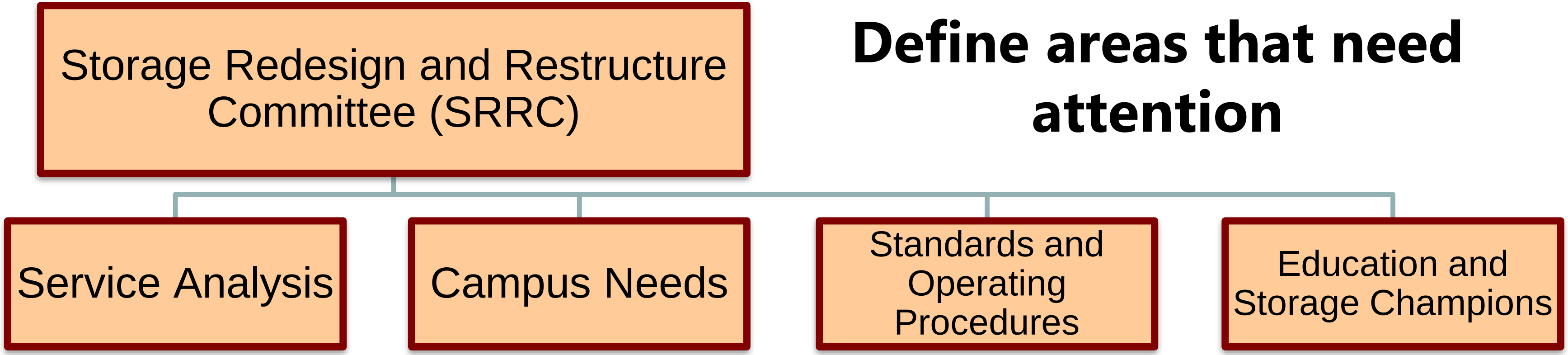
OIT

Center
s

CS
E

Other
s

Define areas that need attention



Storage Redesign and Restructure Committee (SRRC)

Service Analysis

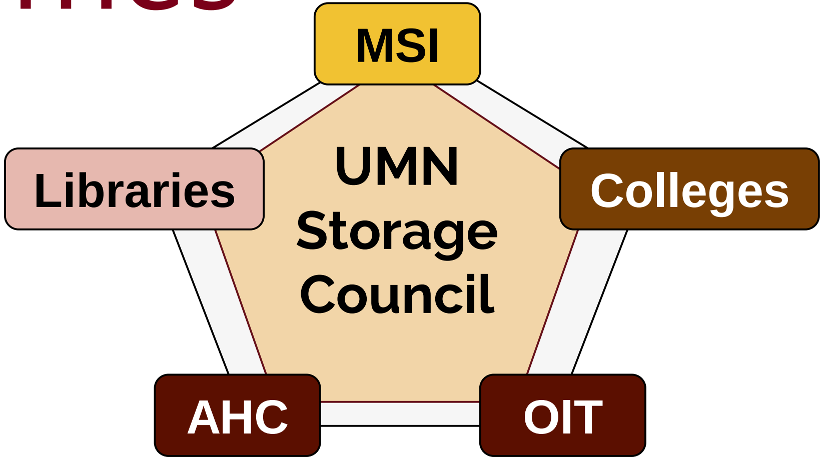
Campus Needs

Standards and Operating Procedures

Education and Storage Champions

Outcomes

Establish a University-wide Storage Council

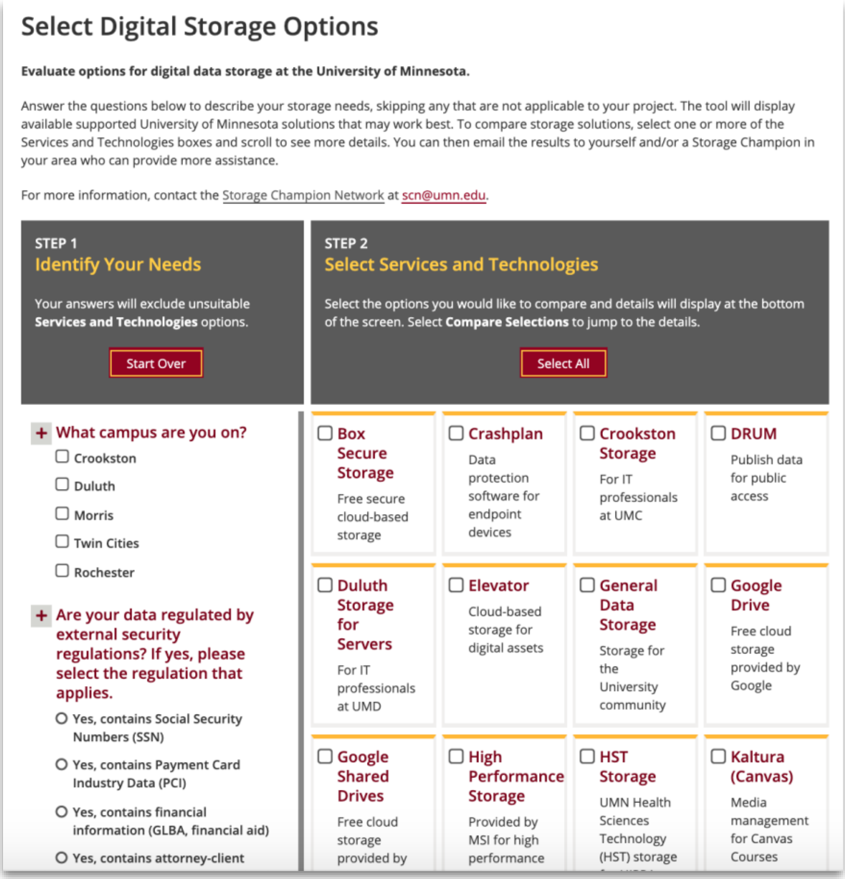


Council is Charged to:

- Develop a Storage Champion Program
- Enhancing Website Infrastructure
- Enhancing User Training and Onboarding
- Collaborating and Sharing Internal Knowledge Articles
- Promoting Marketing and Communication

Outcomes

Storage Champions Network
Kickoff meeting
Septemeber 19, 2019
z.umn.edu/scn



Select Digital Storage Options

Evaluate options for digital data storage at the University of Minnesota.

Answer the questions below to describe your storage needs, skipping any that are not applicable to your project. The tool will display available supported University of Minnesota solutions that may work best. To compare storage solutions, select one or more of the Services and Technologies boxes and scroll to see more details. You can then email the results to yourself and/or a Storage Champion in your area who can provide more assistance.

For more information, contact the Storage Champion Network at scn@umn.edu.

STEP 1
Identify Your Needs

Your answers will exclude unsuitable **Services and Technologies** options.

STEP 2
Select Services and Technologies

Select the options you would like to compare and details will display at the bottom of the screen. Select **Compare Selections** to jump to the details.

+ What campus are you on?

□ Crookston
□ Duluth
□ Morris
□ Twin Cities
□ Rochester

+ Are your data regulated by external security regulations? If yes, please select the regulation that applies.

○ Yes, contains Social Security Numbers (SSN)
○ Yes, contains Payment Card Industry Data (PCI)
○ Yes, contains financial information (GLBA, financial aid)
○ Yes, contains attorney-client

□ Box Secure Storage
Free secure cloud-based storage

□ Crashplan
Data protection software for endpoint devices

□ Crookston Storage
For IT professionals at UMC

□ DRUM
Publish data for public access

□ Duluth Storage for Servers
For IT professionals at UMD

□ Elevator
Cloud-based storage for digital assets

□ General Data Storage
Storage for the University community

□ Google Drive
Free cloud storage provided by Google

□ Google Shared Drives
Free cloud storage provided by

□ High Performance Storage
Provided by MSI for high performance

□ HST Storage
UMN Health Sciences Technology (HST) storage

□ Kaltura (Canvas)
Media management for Canvas Courses

Storage Selection Tool
z.umn.edu/storage-selection-tool

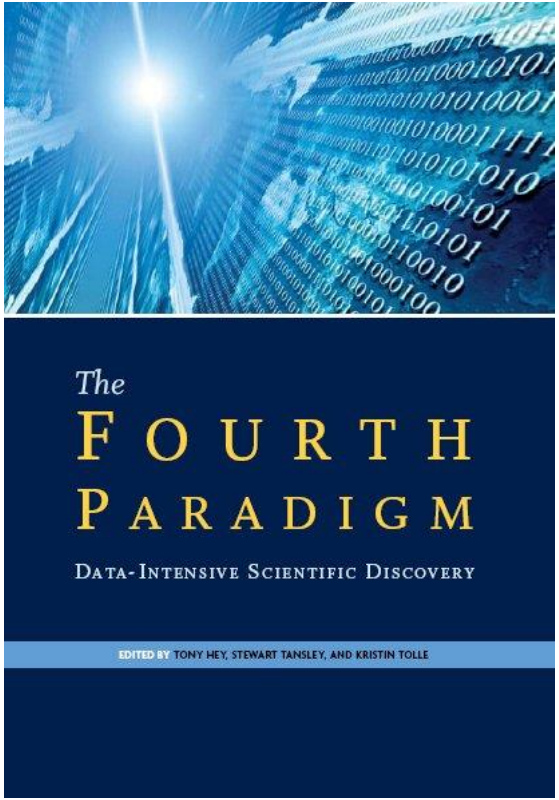
Data as an Asset:

Key Ingredients

- Sustained Research Infrastructure
- Well Trained Cyberinfrastructure Professionals -- AKA Cyberpractitioners
- Good Storage Governance that Spans Institutional Reporting Lines
- **Tools to Make Data Interoperable and to Facilitate Analyses and Sharing**

“Today, the tools for capturing data both at the mega-scale and at the milli-scale are just dreadful.”

Jim Gray, 2007


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

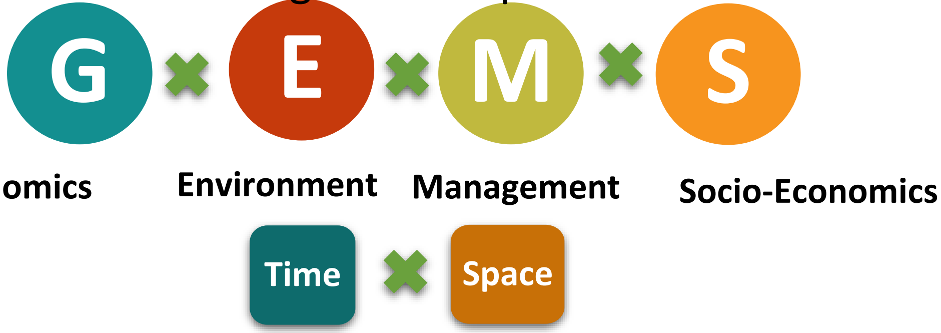
The Challenge

GEMS

Data-Driven Agricultural Innovation

GENETICS • ENVIRONMENT • MANAGEMENT • SOCIOECONOMICS

A novel data sharing and analysis platform to enable public-private research collaborations for innovation in agricultural production and other domain areas.



UNIVERSITY OF MINNESOTA

CFANS
COLLEGE OF FOOD, AGRICULTURAL AND NATURAL RESOURCE SCIENCES

MINNESOTA SUPERCOMPUTING INSTITUTE

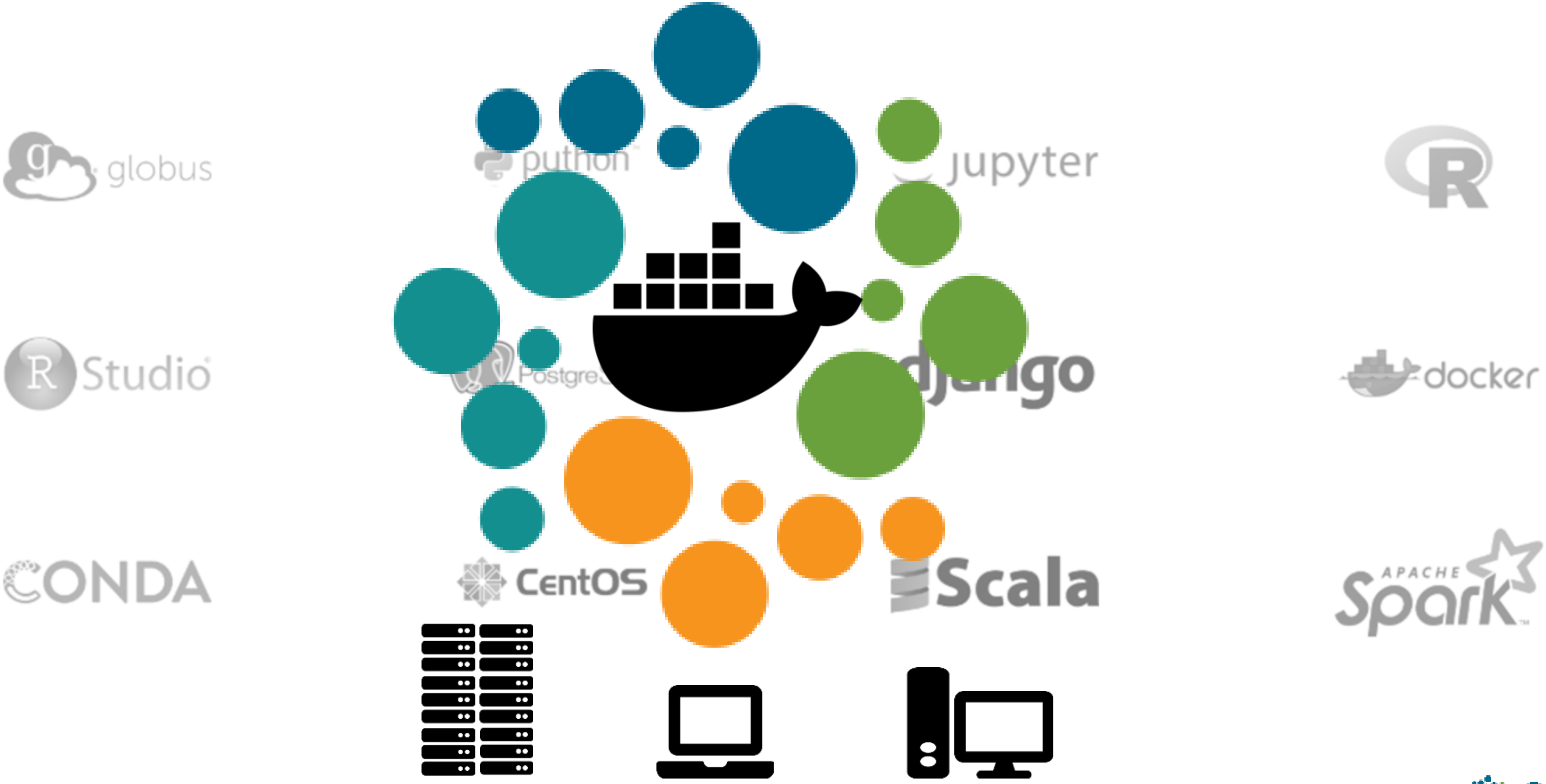
Our Specific Contributions

GEMShare TM

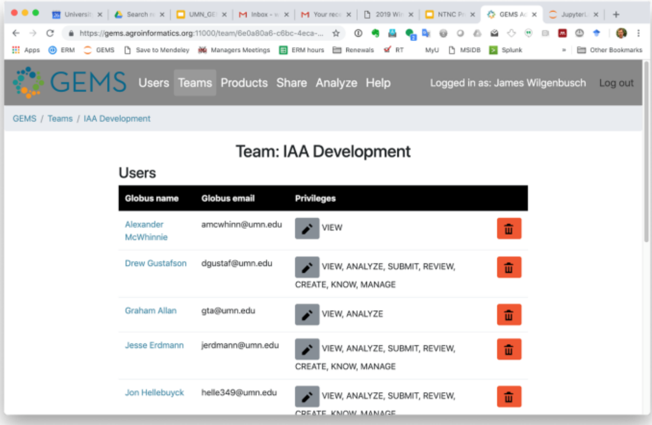
- Smart sharing -- Enables data providers to control who sees what, and when
- Data Versioning – Ensures reproducibility and ability to roll back from changes
- Supports -- open, private, and pooled data
- Beyond data -- Enables sharing of tools and workflows too

GEMSToolsTM is an ever-expanding suite of web-based and command-line analytical tools designed to:

- Cleanup messy (meta-)data
- Intelligently impute missing data
- Enable data interoperability
- Apply advanced analytic methods to genomic, environmental, management and socio-economic data


globus
python
jupyter
R
Studio
PostgreSQL
django
docker
CONDA
CentOS
Scala
APACHE Spark

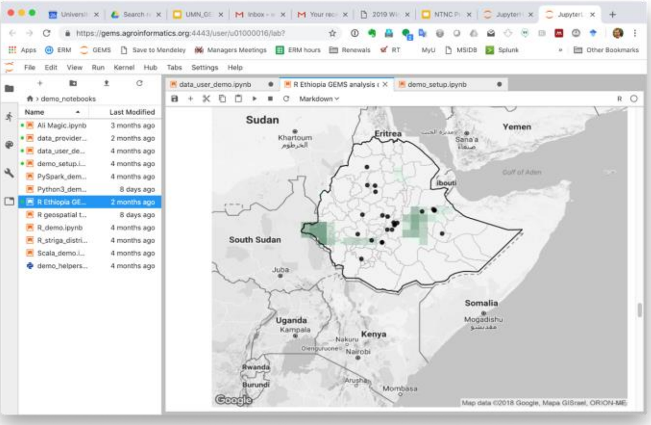
GEMS in Action


https://gems.agroinformatics.org:11000/team/6e0a80a6-c6bc-4eca-...
GEMS
Users
Teams
Products
Share
Analyze
Help
Logged in as: James Wilgenbusch
Log out
GEMS / Teams / IAA Development

Team: IAA Development

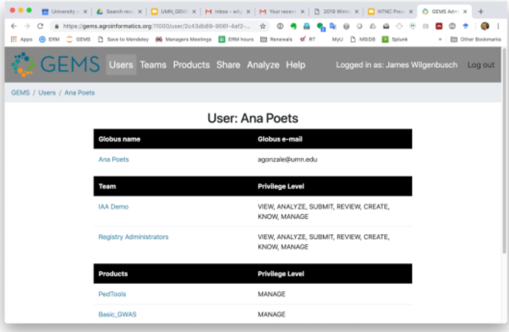
Users

Globus name	Globus email	Privileges
Alexander McWhinnie	amcwhinn@umn.edu	VIEW
Drew Gustafson	dgustaf@umn.edu	VIEW, ANALYZE, SUBMIT, REVIEW, CREATE, KNOW, MANAGE
Graham Allan	gta@umn.edu	VIEW, ANALYZE
Jesse Erdmann	jerdmann@umn.edu	VIEW, ANALYZE, SUBMIT, REVIEW, CREATE, KNOW, MANAGE
Jon Hellebuyck	helle349@umn.edu	VIEW, ANALYZE, SUBMIT, REVIEW, CREATE, KNOW, MANAGE



demo_notebooks

Name	Last Modified
Ali Magic.ipynb	3 months ago
data_provider...	2 months ago
data_user_de...	4 months ago
demo_setup.i...	4 months ago
PySpark_dem...	4 months ago
Python3_dem...	8 days ago
R Ethiopia GE...	2 months ago
R geospatial t...	8 days ago
R_demo.ipynb	4 months ago
R_striga_distri...	4 months ago
Scala_demo.i...	4 months ago
demo_helpers...	4 months ago


https://gems.agroinformatics.org:11000/user/2c43db89-966f-4af2-...
GEMS
Users
Teams
Products
Share
Analyze
Help
Logged in as: James Wilgenbusch
Log out
GEMS / Users / Ana Poets

User: Ana Poets

Globus name	Globus e-mail
Ana Poets	agonzale@umn.edu

Team	Privilege Level
IAA Demo	VIEW, ANALYZE, SUBMIT, REVIEW, CREATE, KNOW, MANAGE
Registry Administrators	VIEW, ANALYZE, SUBMIT, REVIEW, CREATE, KNOW, MANAGE

Products	Privilege Level
PedTools	MANAGE
Basic_GWAS	MANAGE

https://agroinformatics.org/

GEMS^{TM}: Enabling External Partnerships Through the International AgroInformatics Alliance (IAA)

- CGIAR Big Data Platform
- Embrapa, Brazil
- Pepsico
- G2F (Genomes to Fields)
- Diversity Arrays Technology (DArT/KDDART)
- CIAT (cassava, edible beans, forages, rice,)
- University of Adelaide
- Oat Global
- Stellenbosch University, South Africa
- CIMMYT (corn, wheat, socio-economics, genetic resources, IT)

- GRDC (Grains Research Development Corporation), Australia
- MN Department of Agriculture
- PPIRC, Phenotyping and Imaging Center, Canada
- CIP (potatoes, sweet potatoes)
- ICRISAT (sorghum, millet, chickpeas, groundnut)

Thank You

Questions?

jwilgenb@umn.edu