IT Strategy Board

May 12, 2014

UNIVERSITY of WASHINGTON

Agenda

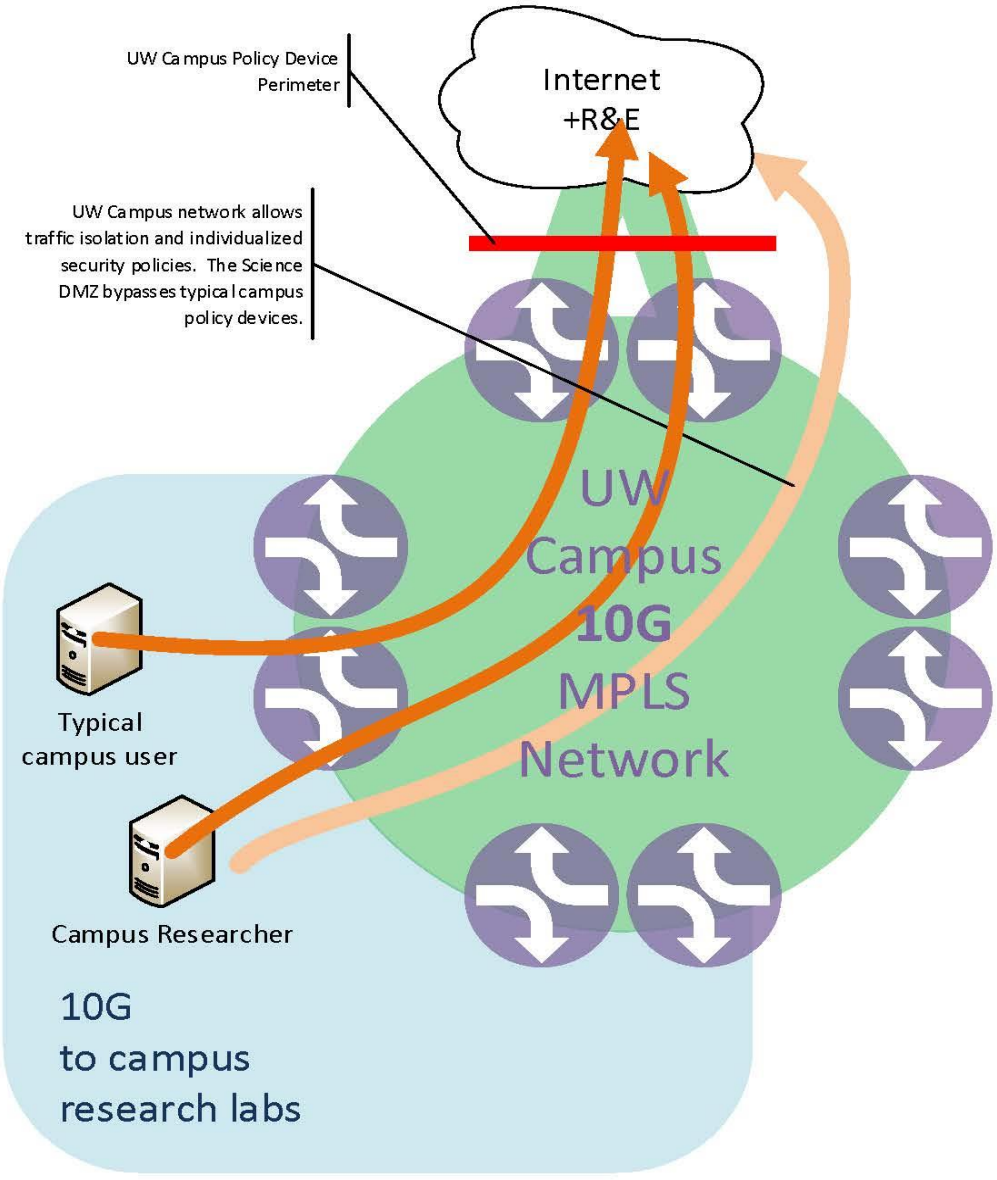
- IT Research Support
- IT Service Investment Board Portfolio Prioritization Outcomes
- Technology Recharge Fee Update
- IT Project Portfolio Executive Review

IT Research Support

Future of Networking

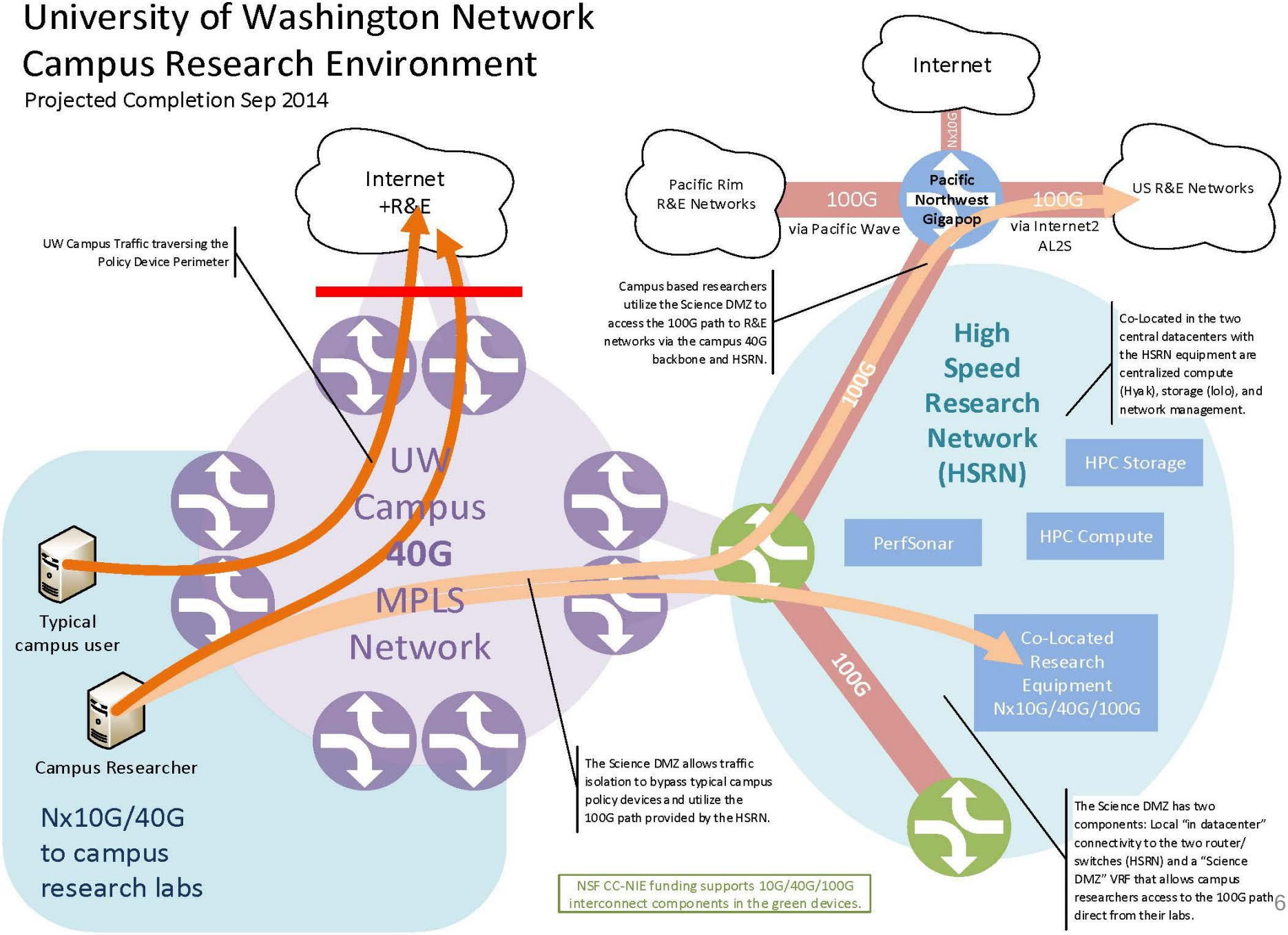
University of Washington Network
Campus Research Environment

Network circa 2012



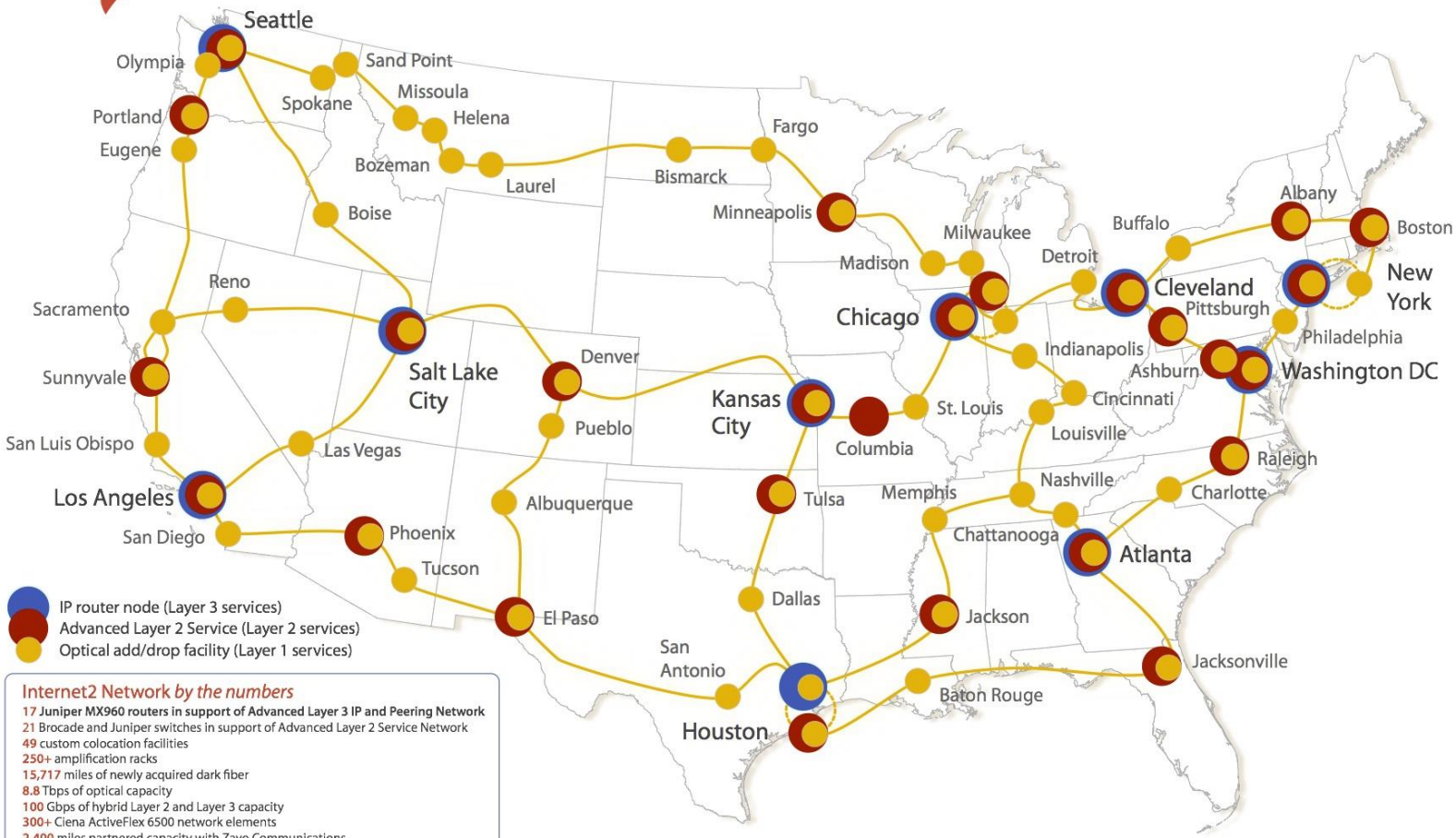
University of Washington Network
Campus Research Environment

Projected Completion Sep 2014



Internet2 Network Infrastructure Topology

July 2013



Internet2 Network *by the numbers*

- **17 Juniper MX960 routers in support of Advanced Layer 3 IP and Peering Network**
- 21 Brocade and Juniper switches in support of Advanced Layer 2 Service Network
- 49 custom colocation facilities
- 250+ amplification racks
- 15,717 miles of newly acquired dark fiber
- 8.8 Tbps of optical capacity
- 100 Gbps of hybrid Layer 2 and Layer 3 capacity
- 300+ Ciena ActiveFlex 6500 network elements
- 2,400 miles partnered capacity with Zayo Communications in support of the Northern Tier region

UNIVERSITY *of* WASHINGTON

Campus Research Environment Report Quarter/Year: Q1/2014

Program Quarterly Status								
Work-stream or Sub-project Name	**Brief Description (Include scope & projected benefits)**	**Start Date**	**Planned End Date**	**Projected End Date**	**Status Indicators – add color if applicable**			
					Overall	**Scope**	**Schedule**	**Resources**
CC-NIE Grant	Support 40G/100G interconnects to campus Science DMZ & backbone and PNWGP for research; 10G interconnects to lolo/hyak; add 10TB lolo storage capacity for researchers	Feb 2013	Aug 2014	Aug 2014	◆	◆	◆	◆
EAGER Grant	Explore Software Defined Networking & OpenFlow applications for campus; develop & test use cases	Feb 2013	Aug 2014	Aug 2014	◆	◆	◆	◆
Science DMZ	Establish high bandwidth network infrastructure outside of the campus security perimeter to support research-based Big Data transfers to/from hyak & lolo and UW partner organizations globally	Feb 2013	Mar 2014	Mar 2014	◆	◆	◆	◆
40G Campus Backbone	Two phase upgrade of campus network backbone from 10G to 40G. Phase I: 4545, UWTower and ATG routing centers; Phase 2: HSH/HSG routing center	Nov 2013	Phase 1: Jan 2013 Phase 2: Aug 2014	Phase 1: Mar 2013 Phase 2: Aug 2014	◆	◆	◆	◆

Quarterly Time Line

Work-stream or Sub-project Name	Q4 2013	Q1 2014	Q2 2014	Q3 2014	Q4 2014	Q1 2015	Q2 2015
CC-NIE Grant							
EAGER Grant							
Science DMZ							
40G Campus Backbone							

Network Virtualization and Security Implications

- We now have the ability to virtually overlay "research networks" on our physical network
  - allows for high capacity pathways to circumvent campus perimeter security
- We are seeking governance to determine appropriate levels of review and approval of requests to use this new capability.

Example:

- Researcher requests High Speed Research Network (HSRN) path from a departmental computing lab to Internet at large, potentially opening lab devices to security breaches

Question: who vets these requests in light of the imputed risk/benefits and authorizes the HSRN connection, perhaps including qualifications of use?

UW-IT Campus Data-Centers

- UW Tower – Built in 2009
  - Total Space Capacity: ~9,000 sq. ft. (200+ cabinets) – currently 96% utilized
  - Total Power Capacity: ~1.5Mw – currently 36% utilized
- 4545 Data Center- Acquired in the 1970s
  - Total Space Capacity: ~12,000 sq. ft. (250+ cabinets) – currently 71% utilized
  - Total Power Capacity: ~0.65Mw – currently 45% utilized

Energy Star Certification

- 2013 Certification from U.S. Environmental Protection Agency (EPA) for UW Tower data center
- One of two university campus data centers in the country to achieve this certification
- Of 50 data centers with this certification, UW data center rank 5th in EPA scoring (95 out of possible 100)

Data Centers

- Unit data centers not designed or built to adequately support server infrastructure
- Units perceive their current server spaces as “free” (i.e., no charge to them for power, cooling, etc.) therefore no incentives for units to enact energy-saving measures
- Environmental Stewardship Committee leading effort to consolidate/virtualize servers in UW-IT managed facilities to reduce carbon emissions and meet UW and state climate action goals
- Discussion points
  - Limit the number of new decentralized data centers on campus
  - Limit upgrades/improvements to existing decentralized data centers
  - Fund UW-IT data center moves

Cyberinfrastructure Support

Overview

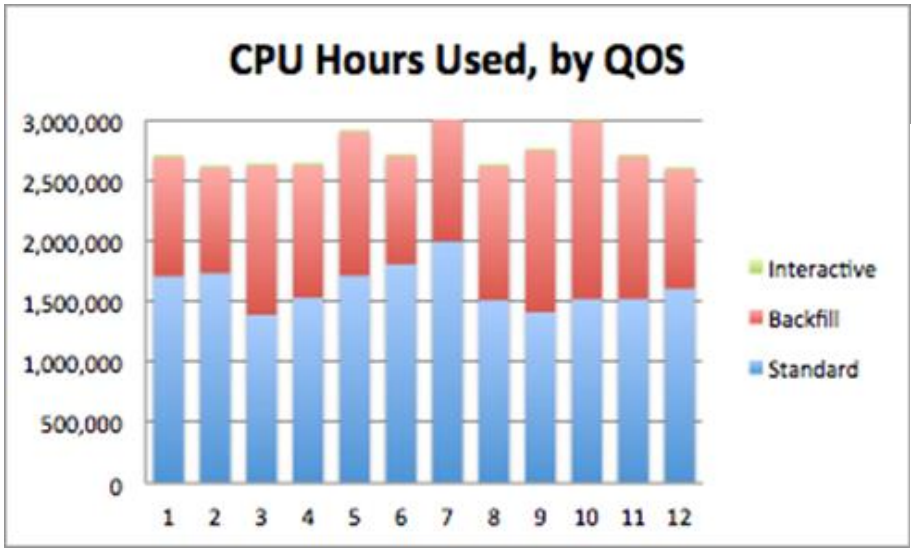
- UW-IT's Cyberinfrastructure (CI) Services
- Comparable Maturity Level
- Next Year's Plan
- Discussion Topics

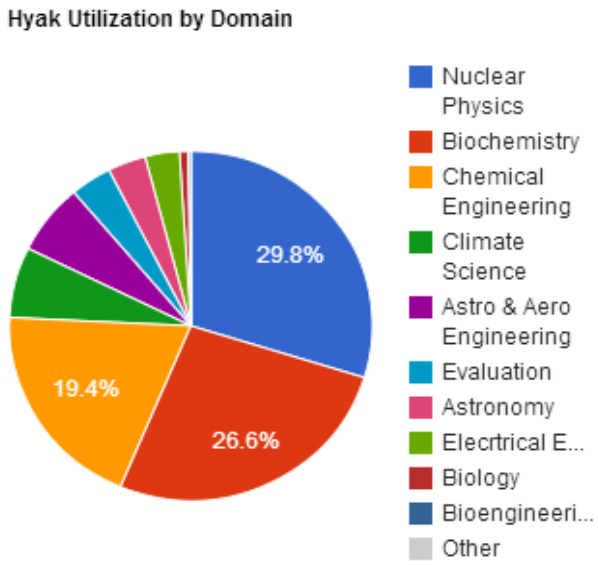
Current CI Services

UW-IT Catalog Services

- Shared Scalable Computer Cluster for Research (Hyak)
- Shared Central File System for Research Archives (lolo)
- Shared Central File System for Research Collaboration (lolo)
- Self-Managed Microsoft Azure Subscription

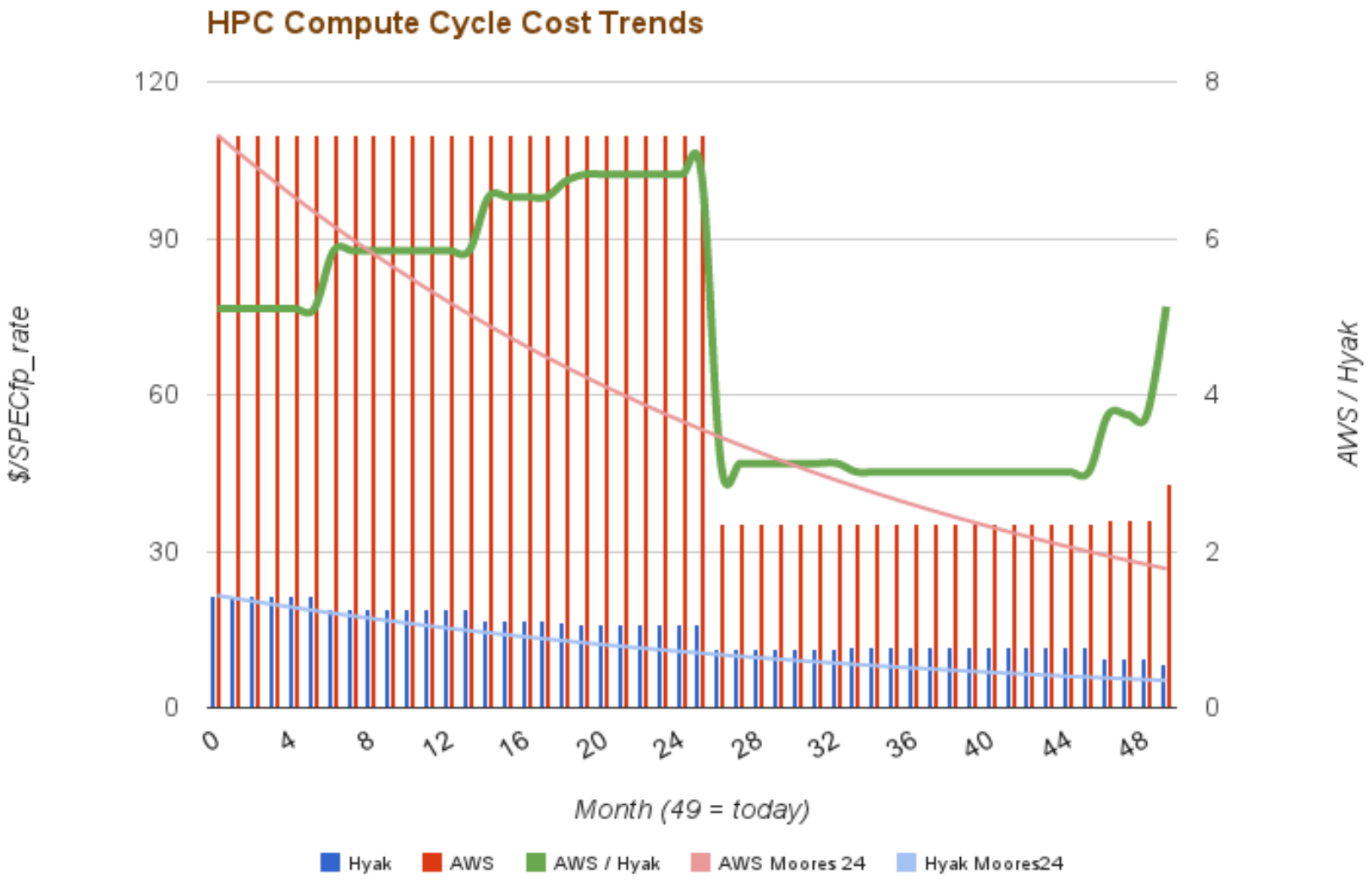
HPC Summary





Key Features:

- Zero-setup time
- Well-managed software stack
- Unused Cycles benefit other researchers

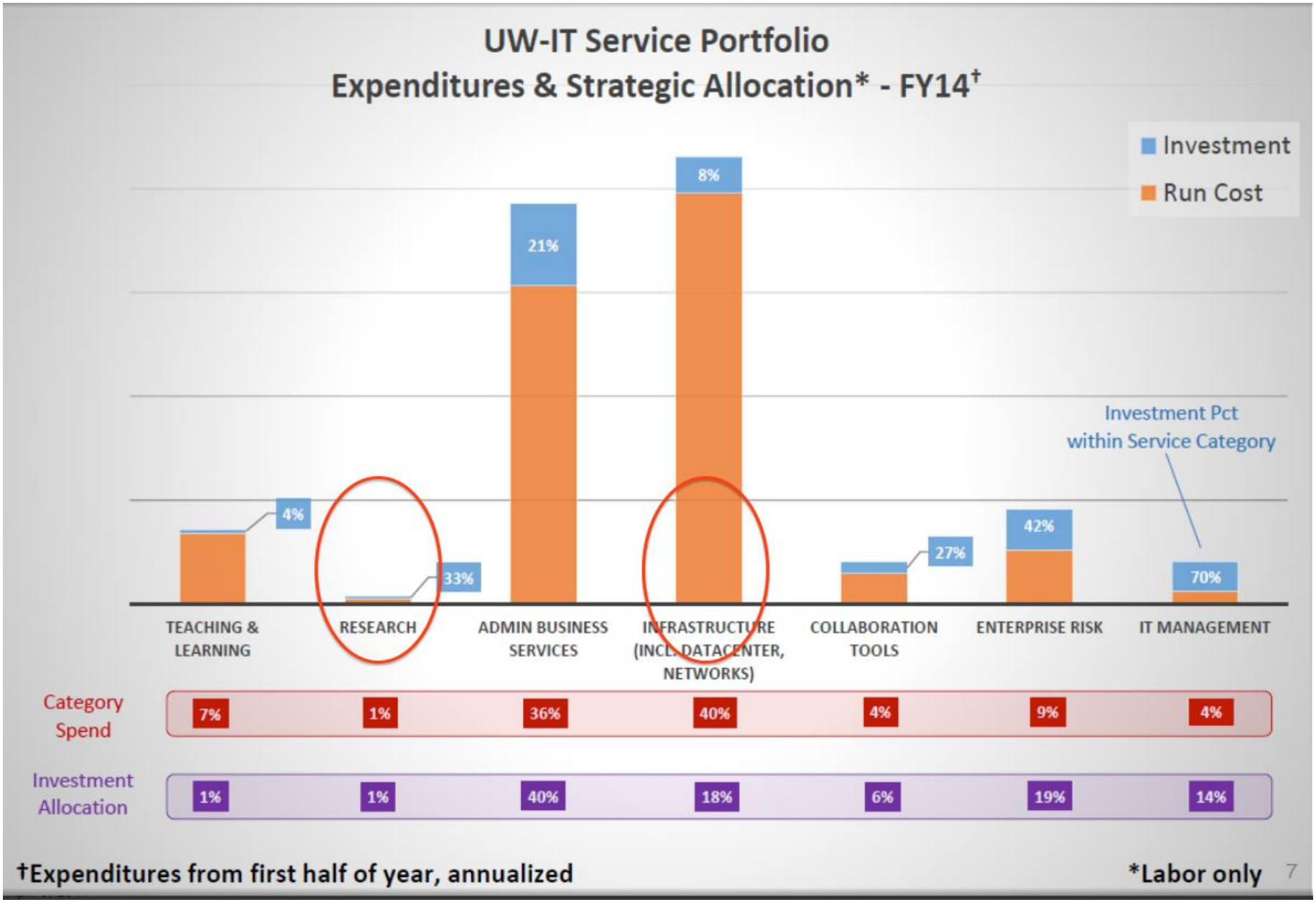


(https://sig.washington.edu/itsigs/Menu_of_equipment_options_and_prices)

Recent UW-IT Research Investments

FY14 projects include:

- Hyak Phase II Infrastructure - completed 07/2013
- High Speed Research Network (100G) - completion 09/2014
- Self-managed Microsoft Azure Tenant - completed 12/2013
- +1 FTE for Cyber-Infrastructure Research & Education Facilitator - (open)


UW-IT Service Portfolio
Expenditures & Strategic Allocation* - FY14†
Investment
Run Cost
Investment Pct within Service Category
4%
33%
21%
8%
27%
42%
70%
TEACHING & LEARNING
RESEARCH
ADMIN BUSINESS SERVICES
INFRASTRUCTURE (INCL. DATACENTER, NETWORKS)
COLLABORATION TOOLS
ENTERPRISE RISK
IT MANAGEMENT
Category Spend
7%
1%
36%
40%
4%
9%
4%
Investment Allocation
1%
1%
40%
18%
6%
19%
14%
†Expenditures from first half of year, annualized
*Labor only
7

Peer Benchmarking Report: Shared Research Computing

November 22, 2013

Subject Experts

Steve Masover, Patrick Schmitz, Chris Hoffman - IST-RIT; Harrison Dekker - Library Data Lab

Description

Includes provision for research *and* teaching of: "traditional" HPC (highly parallelized computing), Data Science methodologies & computational resources, high-powered workstations (including VMs) to support computation at a level between a typical desktop/laptop and an HPC cluster or VM array. Secure compute, storage, data transfer, and data archiving are also in scope.

Benchmarking Criteria

- **Coordinated program** that includes a suite of coordinated services to support computational research and teaching, including a roadmap for service evolution.
- **Support for diverse computational research techniques**, e.g., 'traditional' HPC, virtual machine arrays, and high-powered workstations (which may be virtualized); as well as data transfer and lifecycle management.
- **Training**: Availability and breadth of training.
- **Documentation**: Availability and breadth of documentation.
- **Consulting services**: Including assessment and advice on aligning research problems/needs to available computational resources; grant writing, hardware and software purchasing, and software design, tuning, and refactoring consultation.

Summary of Findings

Tier	Description	Institutions
1	• Strong across all benchmarking criteria	UC San Diego, Princeton, Northwestern
2	• Strong in most benchmarking criteria, stronger in some areas than others.	Harvard, Michigan, MIT, NYU, UCLA, Virginia
3	• Mixed assessment	Columbia, Stanford, Cornell, UW
4	• Weak assessment in most or all areas.	**Berkeley**

Draft Recommendations

Tier	Action
4 → 2	Build a comprehensive program for research computing that provides a range of services from traditional HPC to cloud VM resources to virtual workstations. Develop a community of consultants who have joint appointments in schools, colleges, centers with RIT. One time investment of approx. $1.2 million and recurring investment of up to $1.8 million.
2 → 1	Use Berkeley's strengths in innovation and partnerships with such groups as EECS/Amp Lab, D-Lab, BIDS, and science centers to grow new services in cloud-based HPC and virtual research workstations.

Description

Criteria

Findings

Recommendations

Berkeley
UNIVERSITY OF CALIFORNIA

UW Research Services	
Data Analysis: Quantitative & Qualitative	2
Data Visualization & GIS	2
Linked Open Data & Semantic Web	4
Museums, Archives, & Special Collections	2
Preservation Services	2
Research Application Dev. Support	na
Research Computing (HPC+)	3
Research Data Management	2
Survey Research Support	3

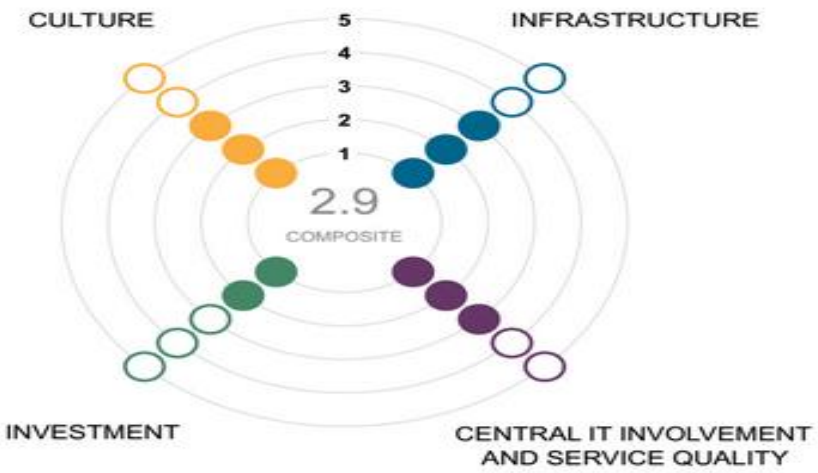
UNIVERSITY of WASHINGTON

EDUCAUSE CENTER FOR ANALYSIS AND RESEARCH

EDUCAUSE

Research Computing Maturity Index

Your Results



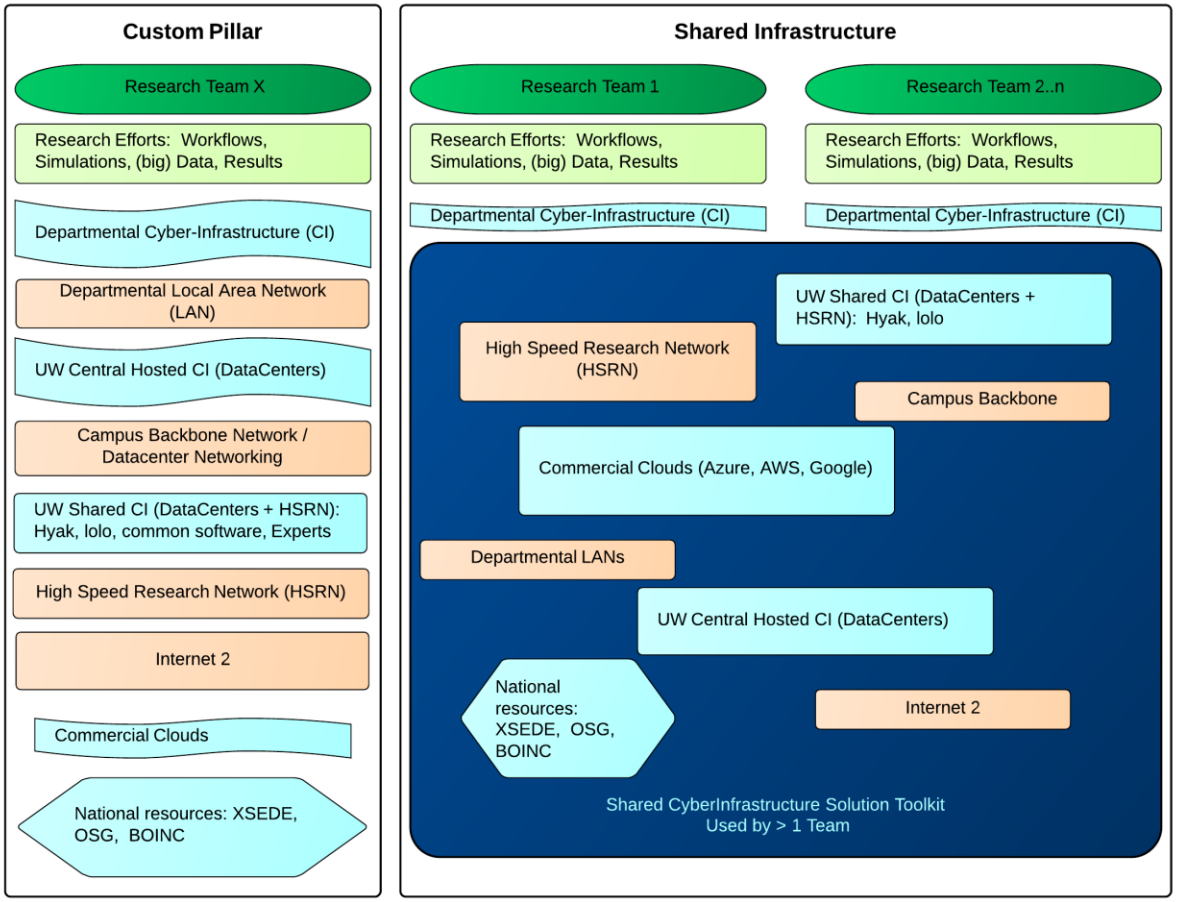
Interpreting your score:

4 Goals For Next Year

1. Develop a sustainable business plan for HPC (FY16+)
2. Outreach to Departmental IT
3. Reduce barriers for adoption of shared Cyberinfrastructure
    a. consulting
    b. equipment cost proposal
    c. annual service reviews

4. Grow a Shared Cyber-Infrastructure

- Toolkits / Software stacks [SQLShare, Mattlab, Data transfer]
- Integrated Cloud options [**Amazon AWS**, Open Science Grid, Azure]
- Toolkit Experts



Cyber-Infrastructure includes: Applications, Code Libraries, Analysis tools, UW Middleware Integrations (Identity, Groups, Job Queuing, ..), Servers and Storage (HPC+scratch, commodity, shared, archival, ...) and Experts that know how to use and maintain.

Equipment Cost Equivalence Proposal

- F&A is a significant disincentive for consolidation & cloud use
- Hyak's Condo Model won't work for other infrastructure
- Near-zero cost to remove F&A on selected services
- Suggest change applied on a service-by-service basis
- Required Approvals: CIO, Office of Research, and Office of Planning and Budgeting

Discussion Topics

- Comments on Strategic Plan for FY15
- Broaden Hyak Governance Board to CI?
- Approval of Equipment Cost Proposal

eScience Institute Initiatives

Data Science @ UW

Today

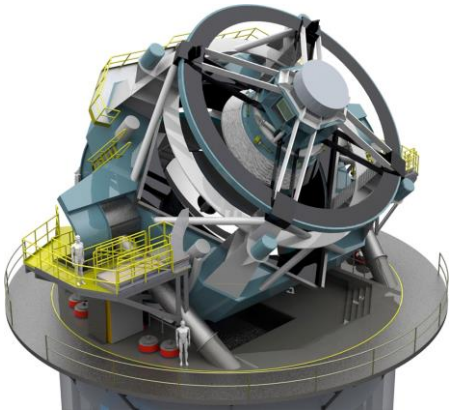
- What's all the fuss about?
- Jim Gray's "fourth paradigm": smart discovery / data-intensive discovery / eScience
- My personal story, and the story of the UW eScience Institute
- Goals and "flagship activities"
- Three science examples: survey astronomy, environmental metagenomics, neuroscience
- "The rising tide that lifts all boats"

What is data science?

Exponential improvements in technology and algorithms are enabling a revolution in discovery

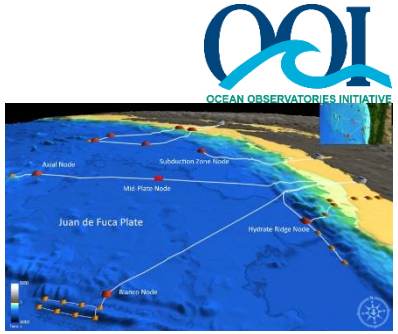
- A proliferation of sensors
- Ever more powerful models producing data that must be analyzed
- The creation of almost all information in digital form
- Dramatic cost reductions in storage
- Dramatic increases in network bandwidth
- Dramatic cost reductions and scalability improvements in computation
- Dramatic algorithmic breakthroughs in areas such as machine learning

Nearly every field of discovery is transitioning from “data poor” to “data rich”

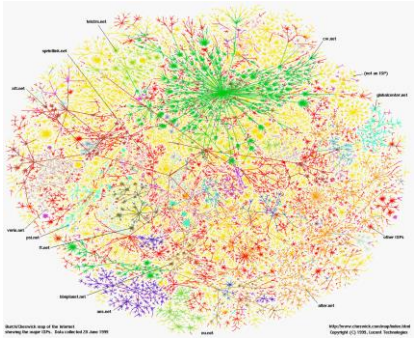


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

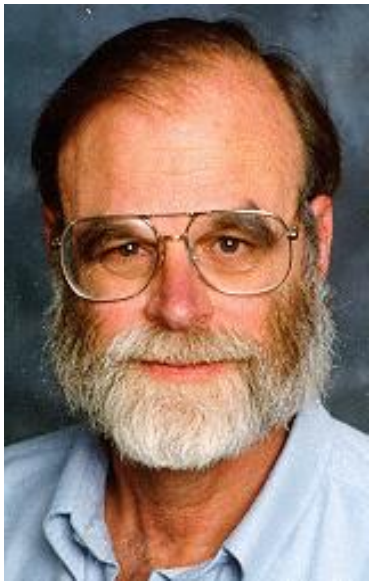
Biology: Sequencing

Economics: POS terminals

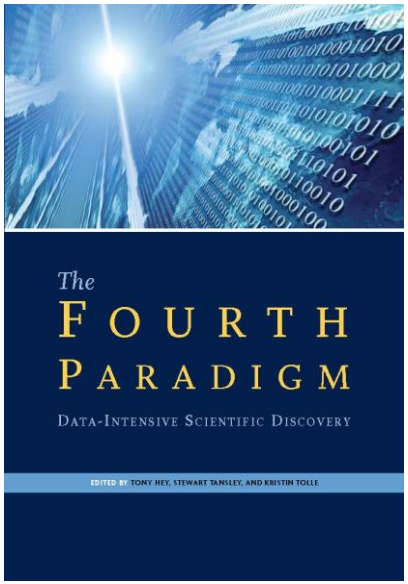
Neuroscience: EEG, fMRI

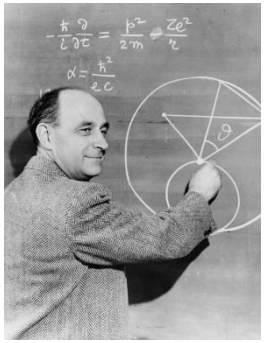
The Fourth Paradigm

1. Empirical + experimental
2. Theoretical
3. Computational
4. Data-Intensive



Jim Gray

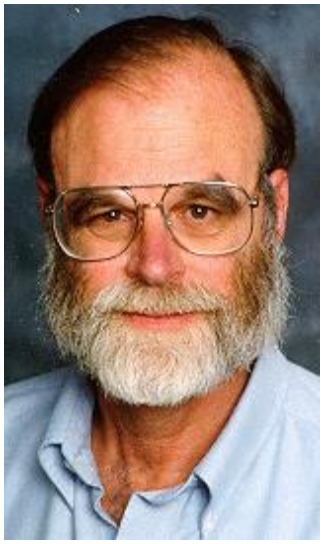




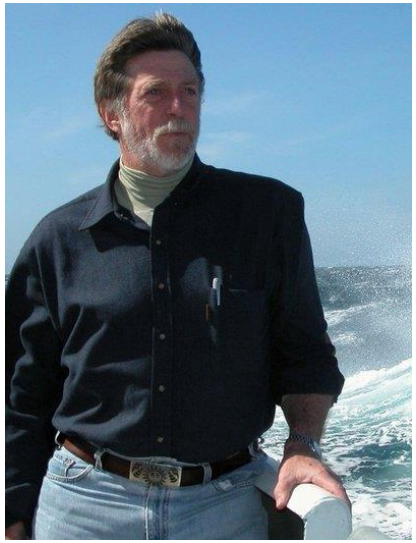
"From data to knowledge to action"

- The ability to extract knowledge from large, heterogeneous, noisy datasets – to move "from data to knowledge to action" – lies at the heart of 21st century discovery
- To remain at the forefront, researchers *in all fields* will need access to state-of-the-art data science methodologies and tools
- These methodologies and tools will need to advance rapidly, driven by the requirements of discovery
- Data science is driven more by *intellectual infrastructure* (human capital) and *software infrastructure* (shared tools and services – digital capital) than by hardware

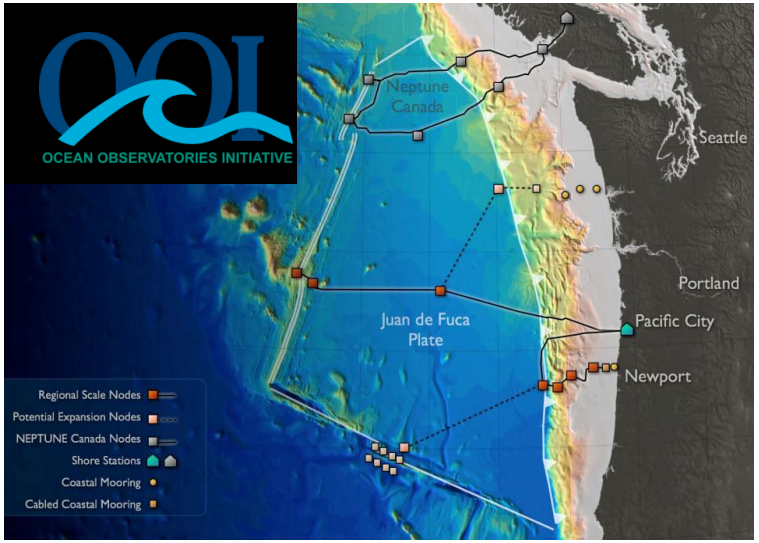
My personal story, and the story of the UW eScience Institute

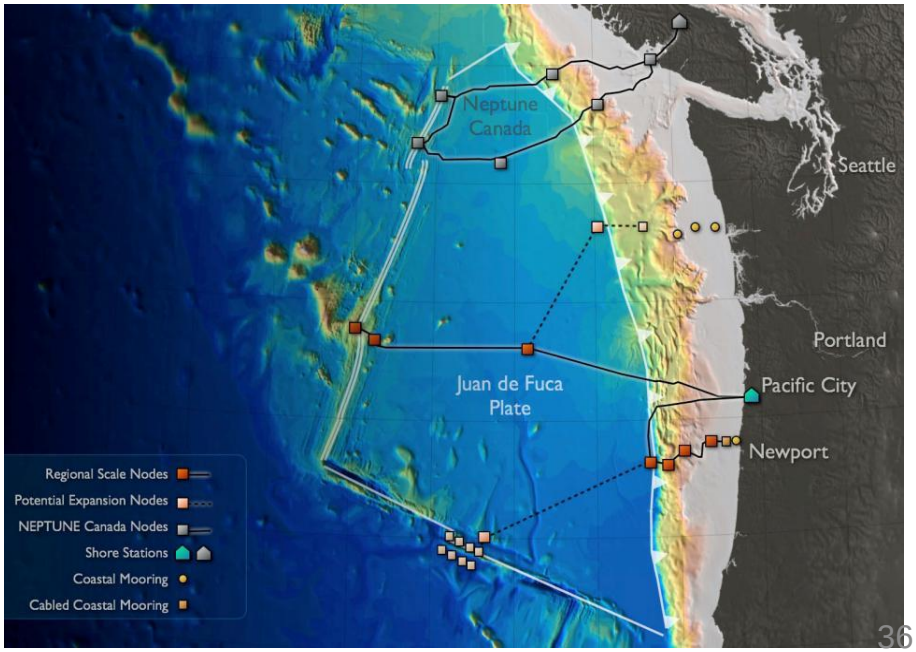


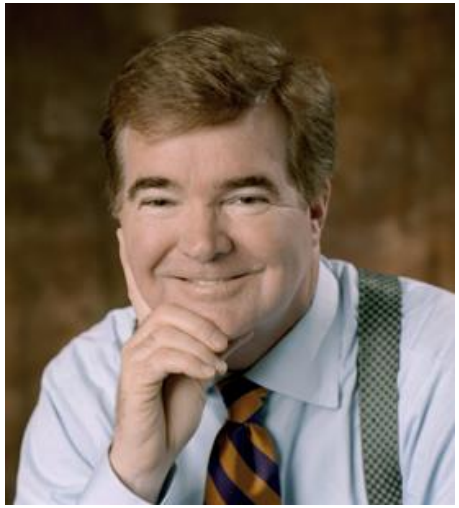
Early 1980s



Late 1990s




Neptune
Canada
Seattle
Portland
Juan de Fuca
Plate
Pacific City
Newport
Regional Scale Nodes
Potential Expansion Nodes
NEPTUNE Canada Nodes
Shore Stations
Coastal Mooring
Cabled Coastal Mooring



Mark Emmert

LSU

2004

UNIVERSITY OF WASHINGTON 1861 LVX SIT

Ed Lazowska, CSE

Tom Daniel, Biology

Werner Stuetzle, Statistics

UW eScience Institute

- *"All across our campus, the process of discovery will increasingly rely on researchers' ability to extract knowledge from vast amounts of data... In order to remain at the forefront, UW must be a leader in advancing these techniques and technologies, and in making [them] accessible to researchers in the broadest imaginable range of fields."*

History

2005

- Early discussions with Mark Emmert; survey of efforts elsewhere

2006, 2007

- Concept documents written and revised

2008

- Core funding received from legislature; eScience Institute established
  - Steering Committee established
  - Research Scientists hired
  - Research partnerships established
  - Hyak initiative launched

2012

- Control of funds moved from VP Research to UW IT (Hyak) and eScience Institute ("intellectual infrastructure")
- Emily Fox, Carlos Guestrin, Jeff Heer, Ben Taskar hired, catapulting UW into the lead in data science methodology
- Inspired by this, 4 half-faculty-positions allocated by Provost
- Led by Bill Howe, UWEO "Certificate Program in Data Science" launched

2013

- Coursera MOOC "Introduction to Data Science" created by Bill Howe
- $2.8M from National Science Foundation: IGERT to create an interdisciplinary graduate program in Data Science
- $37.8 million from Moore Foundation and Sloan Foundation to UW, Berkeley, and NYU to collaborate in the creation of "Data Science Environments"

2014

- Activities launched under Moore/Sloan initiative
  - Campus-wide rollout on February 7
  - Recruiting of research staff, administrative staff, and postdocs
  - Multiple active working groups spanning the three Moore/Sloan campuses
  - “Incubation program” launched
  - Creation of “Data Science Studio” for cross-campus collaboration
- $9.3 million from Washington Research Foundation to amplify the Moore/Sloan effort
  - Also $7.1 million to closely-related Institute for Neuroengineering, $8.0 million to Institute for Protein Design, $6.7 million to Clean Energy Institute

Faculty core team

Data science methodology


Cecilia Aragon
Human Centered Design & Engr.


Magda Balazinska
Computer Science & Engineering


Emily Fox
Statistics


Carlos Guestrin
CSE


Bill Howe
CSE


Jeff Heer
CSE


Ed Lazowska
CSE


Randy LeVeque
Applied Mathematics

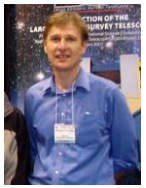

Werner Stuetzle
Statistics

Biological sciences


Tom Daniel
Biology


Bill Noble
Genome Sciences

Physical sciences


Andy Connolly
Astronomy


John Vidale
Earth & Space Sciences

Environmental sciences


Ginger Armbrust
Oceanography

Social sciences


Josh Blumenstock
iSchool


Mark Ellis
Geography


Tyler McCormick
Sociology, CSSS


Thomas Richardson
Statistics, CSSS

Faculty core team

Data science methodology


Cecilia Aragon
Human Centered Design & Engr.


Magda Balazinska
Computer Science & Engineering


Emily Fox
Statistics


Carlos Guestrin
CSE


Bill Howe
CSE


Jeff Heer
CSE


Ed Lazowska
CSE

12 Departments
5 Schools / Colleges

Biological sciences


Tom Daniel
Biology


Bill Noble
Genome Sciences

Physical Sciences


Andy Connolly
Astronomy


John Vidale
Earth & Space Sciences


Randy LeVeque
Applied Mathematics


Werner Stuetzle
Statistics

Environmental sciences


Ginger Armbrust
Oceanography

Social sciences


Josh Blumenstock
iSchool

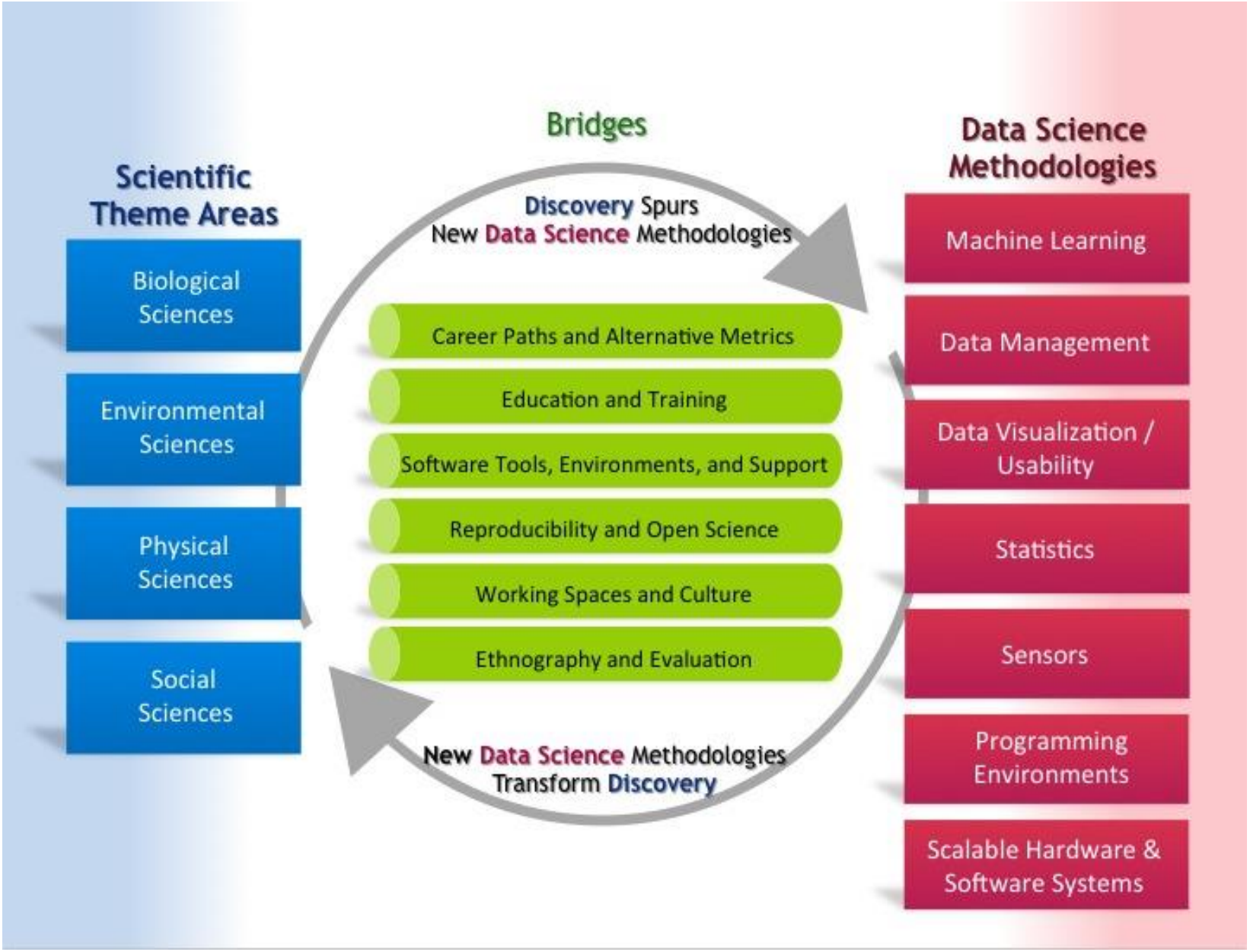

Mark Ellis
Geography


Tyler McCormick
Sociology, CSSS


Thomas Richardson
Statistics, CSSS

Goals

- *Do* breakthrough science
  - In Scientific Theme Areas
  - In Data Science Methodology areas
- *Enable* breakthrough science
  - Through new tools and methods
  - Through changing the process of discovery and driving cultural changes
- *Establish a "virtuous cycle"*

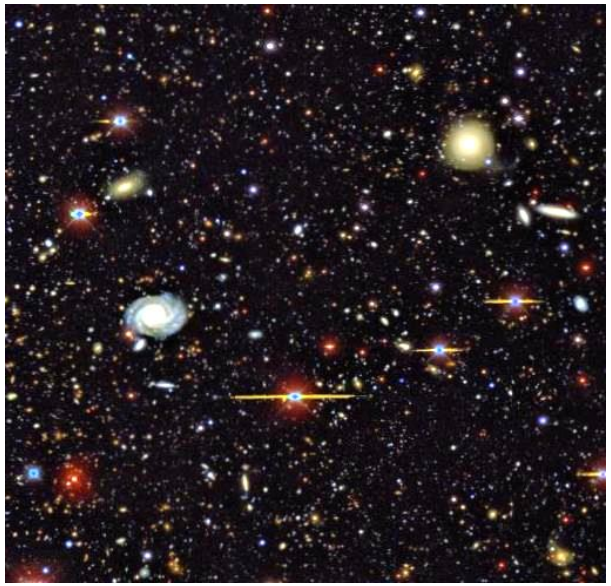

Scientific Theme Areas
Biological Sciences
Environmental Sciences
Physical Sciences
Social Sciences
Bridges
Discovery Spurs New Data Science Methodologies
Career Paths and Alternative Metrics
Education and Training
Software Tools, Environments, and Support
Reproducibility and Open Science
Working Spaces and Culture
Ethnography and Evaluation
New Data Science Methodologies Transform Discovery
Data Science Methodologies
Machine Learning
Data Management
Data Visualization / Usability
Statistics
Sensors
Programming Environments
Scalable Hardware & Software Systems

"Flagship Activities"

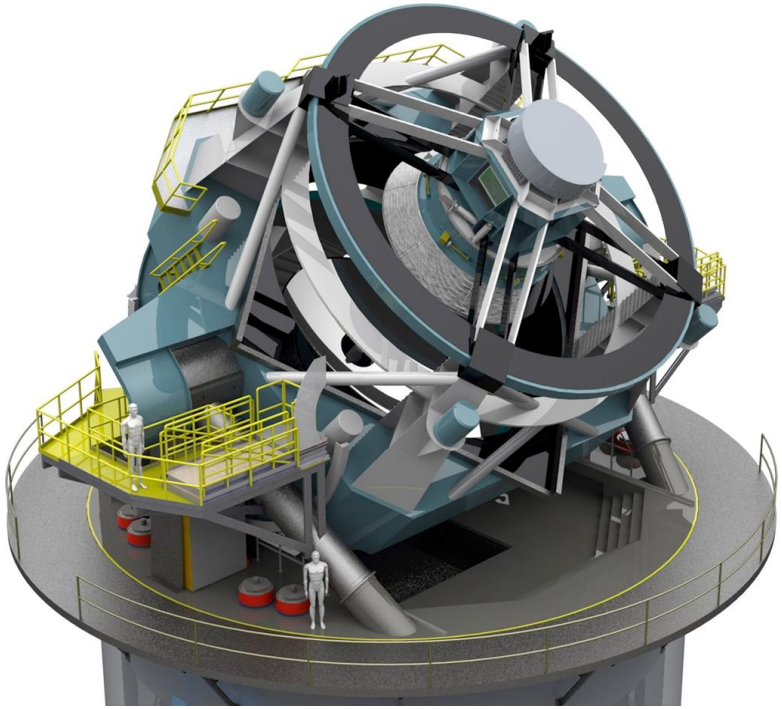
- New career paths: Establish two new roles: *Data Science Fellows* and *Data Scientists*
- Educate "Pi-shaped" students: Establish a new graduate program in data science (NSF IGERT)
- Re-create the watercooler: Establish a "Data Science Studio"
- Create scalable impact: Establish an "Incubator" seed grant program
- Establish a campus-wide community around reproducible research
- Establish a research program in "the data science of data science"
- Conduct and enable breakthrough science

AstroDB: Cosmology at Scale

Andrew Connolly (Astronomy)
Magda Balazinska (CSE)



The Large Synoptic Survey Telescope



- Survey half the sky every 3 nights (1000-fold increase in data vs. Sloan Digital Sky Survey)
- Enabled by a 3.2 Gigapixel camera with a 3.5 degree field
- 15 TB/night (100 PB over 10 years), 20 billion objects, and 20 trillion measurements

How do we do science at petabyte scale?

Science questions …

- Finding the unusual
  - Supernova, GRBs
  - Probes of Dark Energy
- Finding moving sources
  - Asteroids and comets
  - Origins of the solar system
- Mapping the Milky Way
  - Tidal streams
  - Probes of Dark Matter
- Measuring shapes of galaxies
  - Gravitational lensing
  - The nature of Dark Energy

How do we do science at petabyte scale?

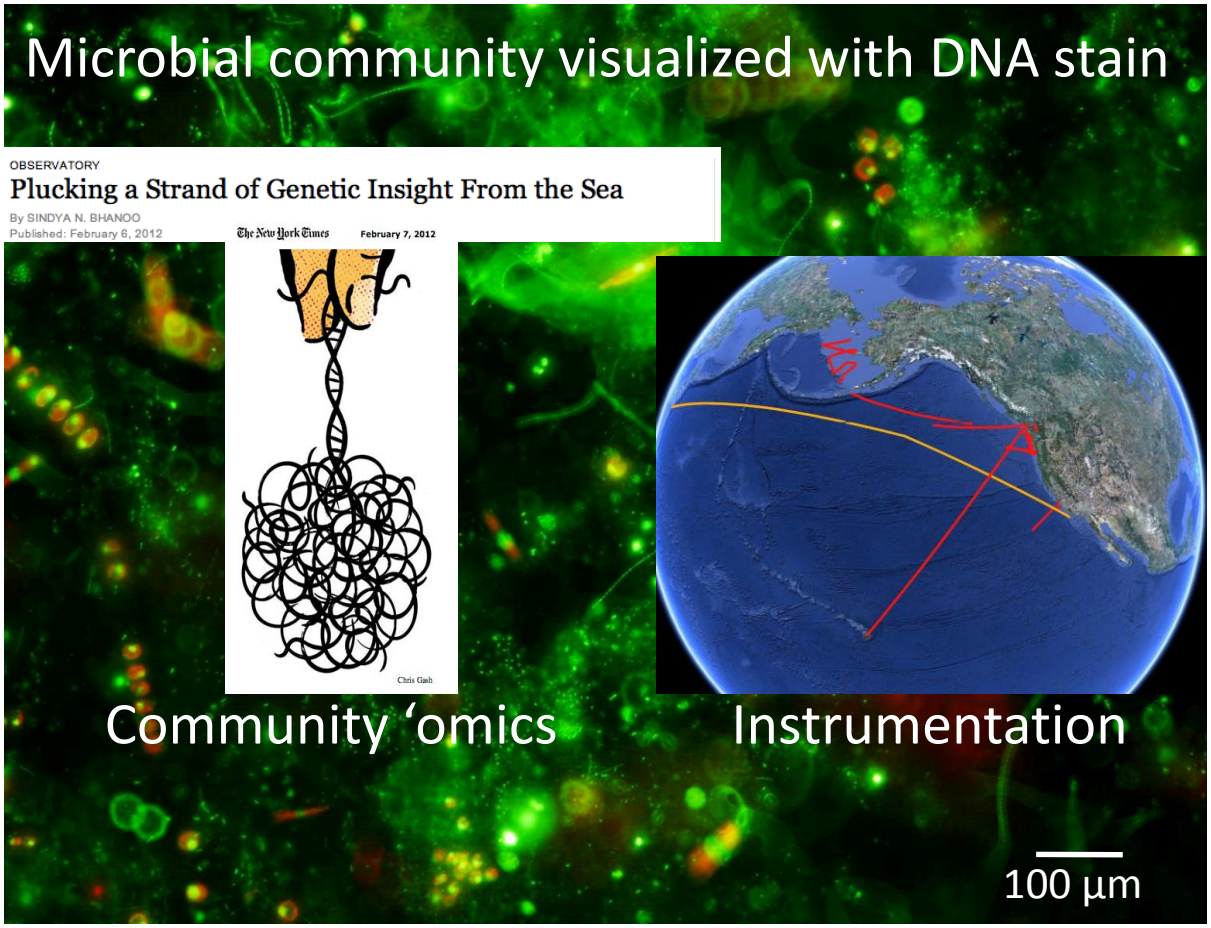
Science questions … map to computational questions

- Finding the unusual
  - Supernova, GRBs
  - Probes of Dark Energy
- Finding moving sources
  - Asteroids and comets
  - Origins of the solar system
- Mapping the Milky Way
  - Tidal streams
  - Probes of Dark Matter
- Measuring shapes of galaxies
  - Gravitational lensing
  - The nature of Dark Energy

- Finding the unusual
  - Anomaly detection
  - Density estimations
- Finding moving sources
  - Tracking algorithms
  - Kalman filters
- Mapping the Milky Way
  - Clustering techniques
  - Correlation functions
- Measuring shapes of galaxies
  - Image processing
  - Data intensive analysis

Role of microbes in marine ecosystems

Ginger Armbrust (Oceanography)
Bill Howe (Computer Science & Engineering + eScience




Challenges:
1) Integration across different data types
2) Distributed and remote labs


eScience Institute
Supporting Data-Driven Discovery In All Fields
WHO WE ARE
SQLShare: Database-as-a-Service for Science
Try SQLShare | Tutorial | Publications | Developers | How to Cite SQLShare
Python API | R API | REST API
SQLShare: Upload Data, Get Answers, Share Results
SQLShare is a database service aimed at removing the obstacles to using relational databases: installation, configuration, schema design, tuning, data ingest, and even application design. You simply upload your data and immediately start querying it.

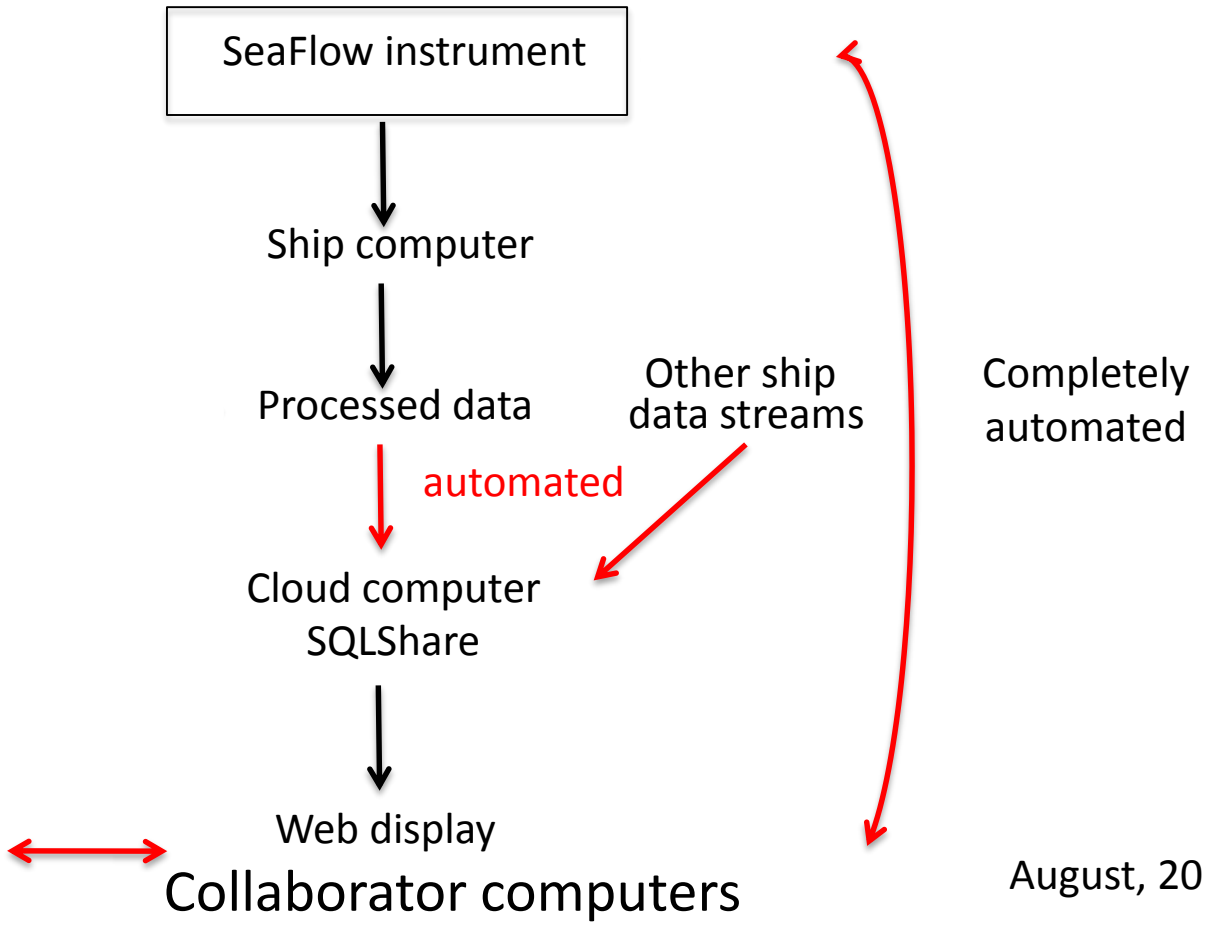
Integrating across physics, biology, and chemistry

Query across data sets in real-time
"not just faster…different!"

Dan Halperin,
Research Scientist, eScience Institute

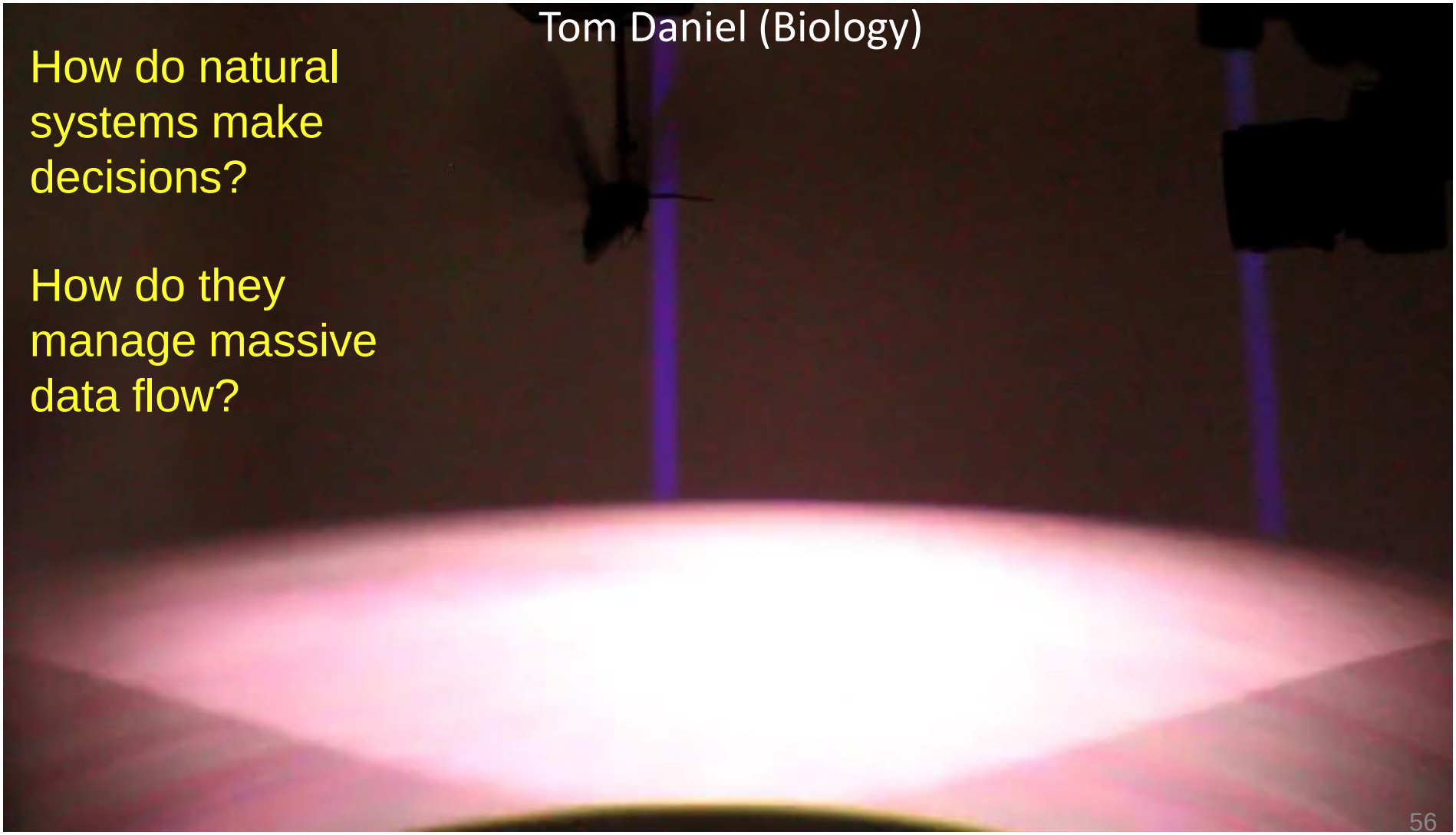
Konstantin Weitz
Graduate student, CSE

Connecting across distributed labs



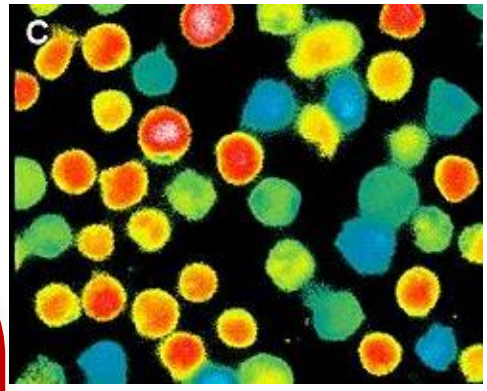
August, 2013

Devices + Neuroscience + Data Science



What features do animals extract to solve problems?

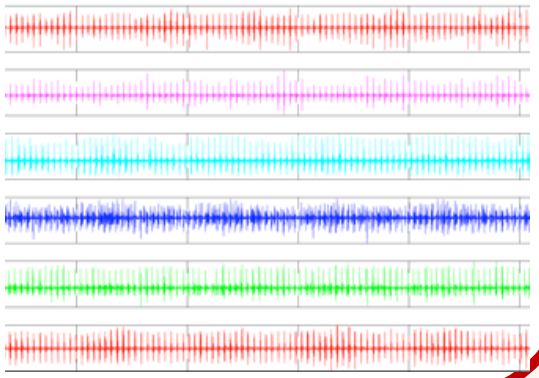
Neural activity



How is information synthesized to drive decisions?

Complex environments

Motor activity



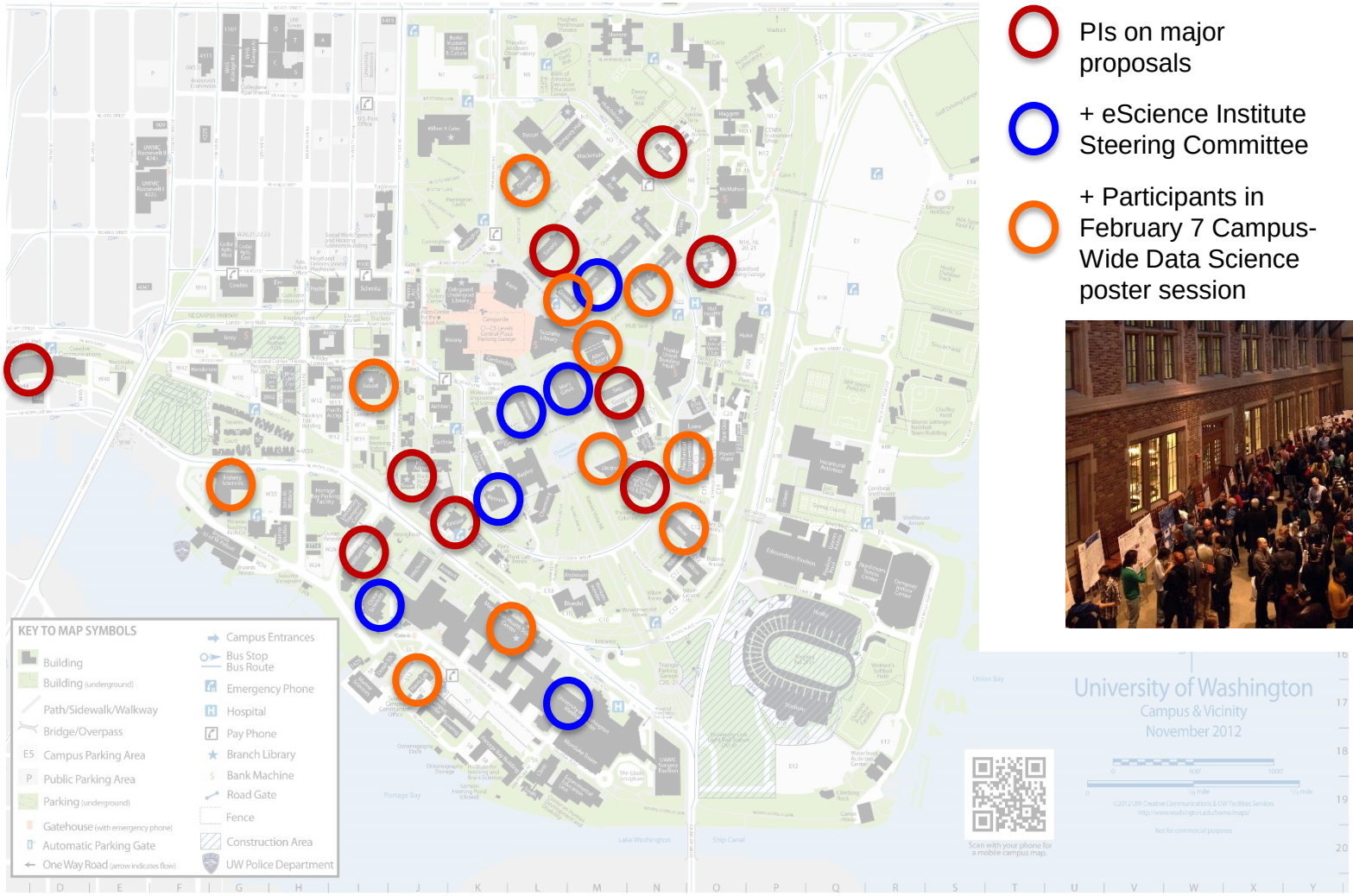
Behavioral output

How does action affect subsequent sensation?

How do muscles work together to perform actions?

These scientists are involved because their science can only succeed if there is a major cultural shift within universities and a major change in the way we approach discovery

Data science: The rising tide that lifts all boats



PIs on major proposals

+ eScience Institute Steering Committee

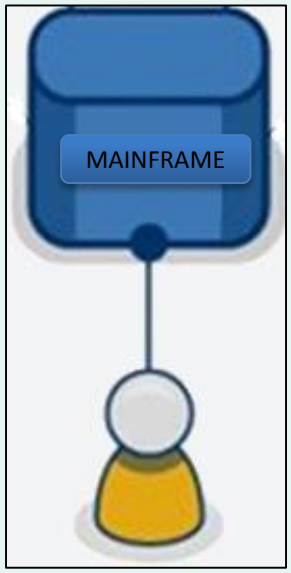
+ Participants in February 7 Campus-Wide Data Science poster session

We're at the dawn of a revolutionary new era of discovery and of learning

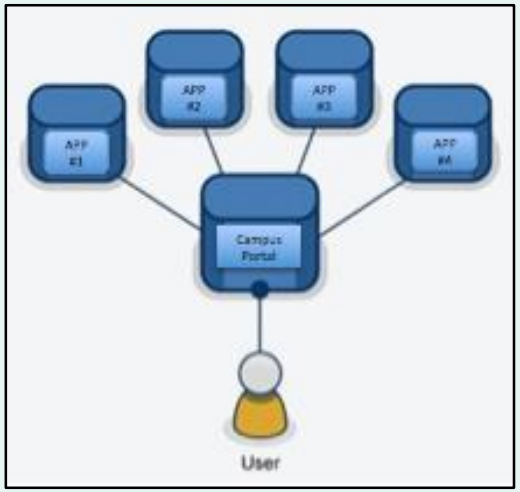
TIER Collaboration

A Short History: Identity Management 1960-2012

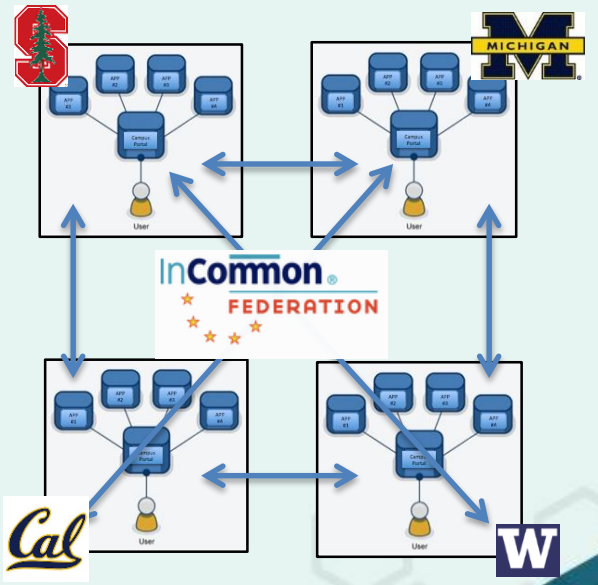
In the beginning, there were individual accounts for individual systems – the "Mainframe Model"



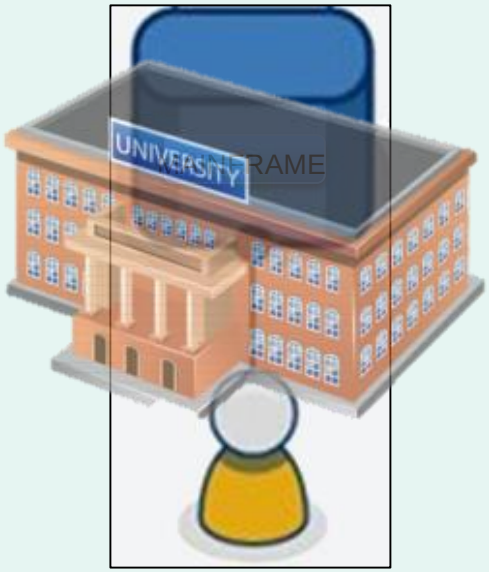
Client Server then Intranet broke that. So we invented "Single Sign-On"

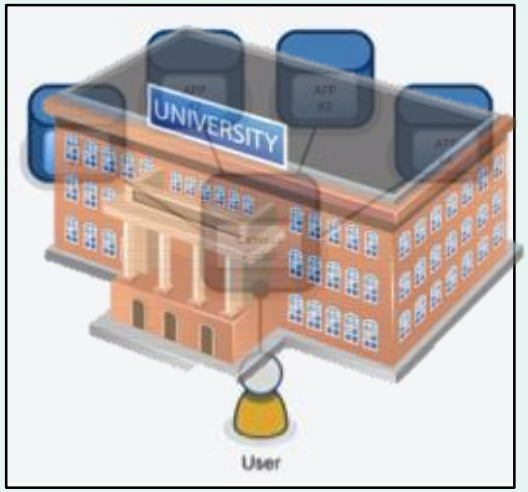


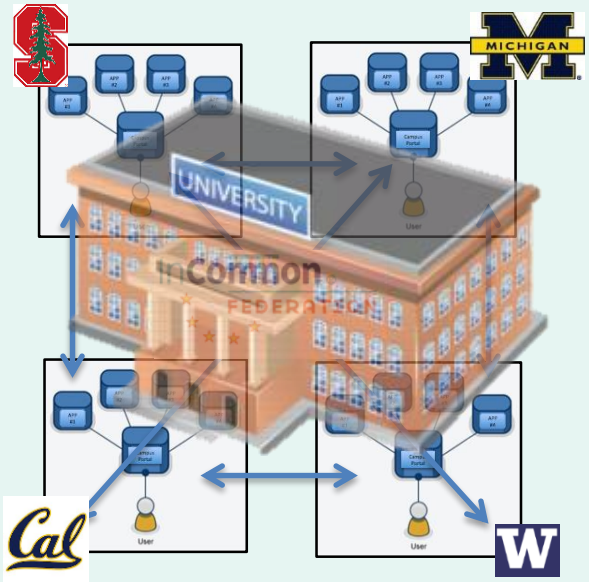
Then Federation was established to apply SSO across multiple campuses



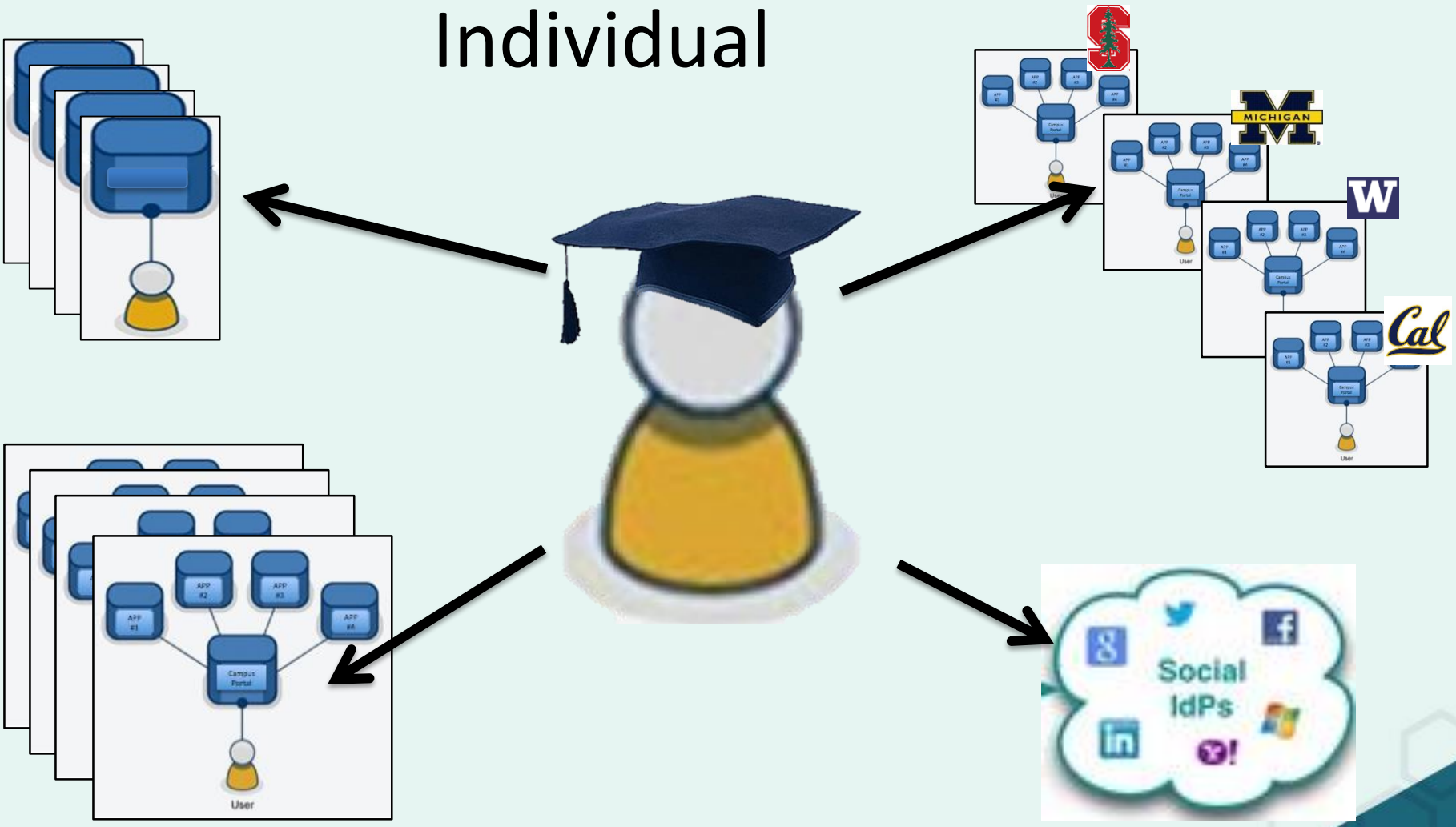
What did all of these have in common? Enterprise Design as the Core





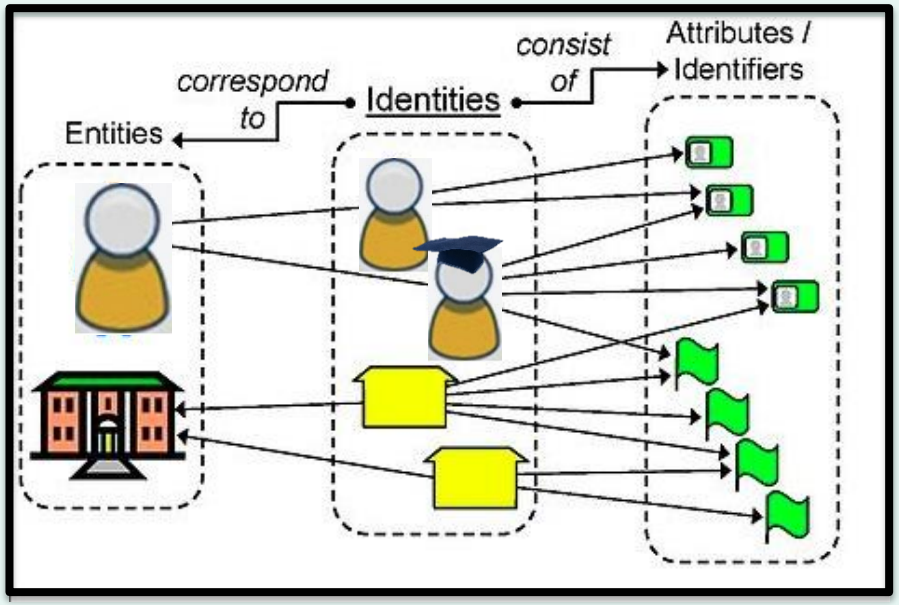


Center on Individual

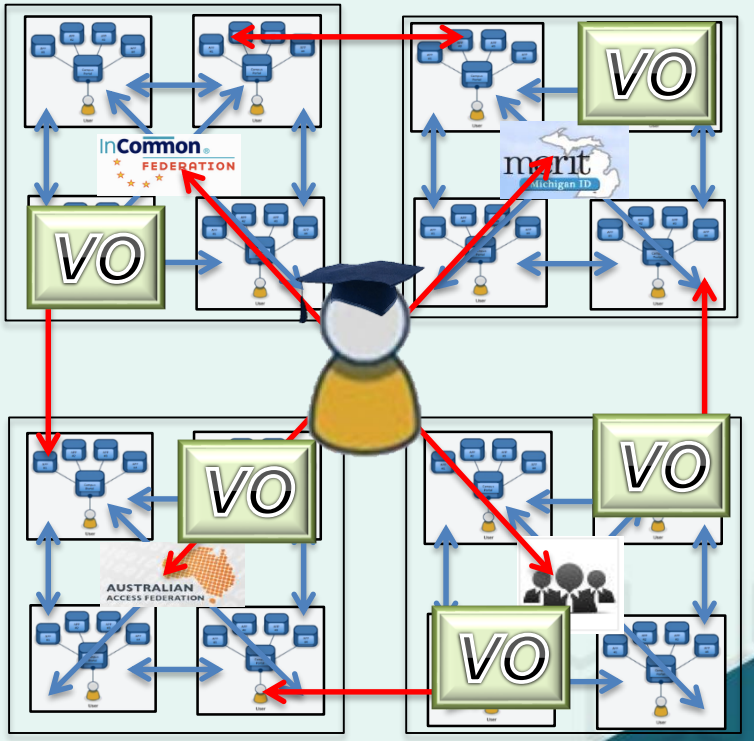


A Needed Future: Trust and Identity 2014-2024

Empower “Individual Opt In” and Require Standards plus “Commitment of Participation” for Release of key Institutional Attributes



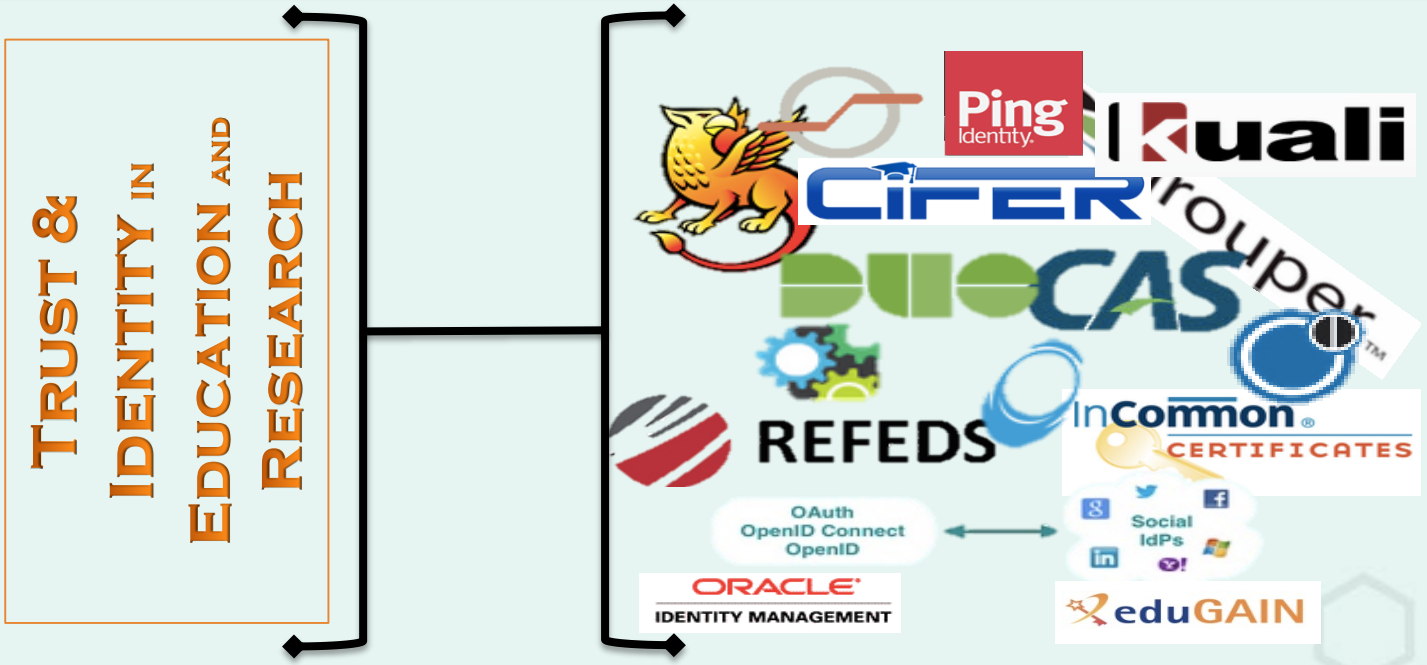
Adjusted Design Point for all Trust and Identity Activities from “Enterprise” To “Virtual Organization



[65]

A Needed Future: Trust and Identity 2014-2024

Aligned Comprehensive Governance (and Strategy) for all higher education middleware and services by TIER



Four Parallel Work Streams to Deliver TIER

TRUST & IDENTITY IN EDUCATION AND RESEARCH

Governance

The Trust and Identity governance will have responsibility for the larger higher education community and incorporate all community offerings and architectural standard decisions. Klara J. Chair

Architecture

All architecture for middleware, API and service integration for trust and identity will be mapped and coordinated through here for review and approval by the steering committee. Steve Zoppi Lead

Campus

Campus to work internally to implement adoption of standards as identified above, with specific requirements for participation in an updated federation structure around attributes. Campus CIO's Lead

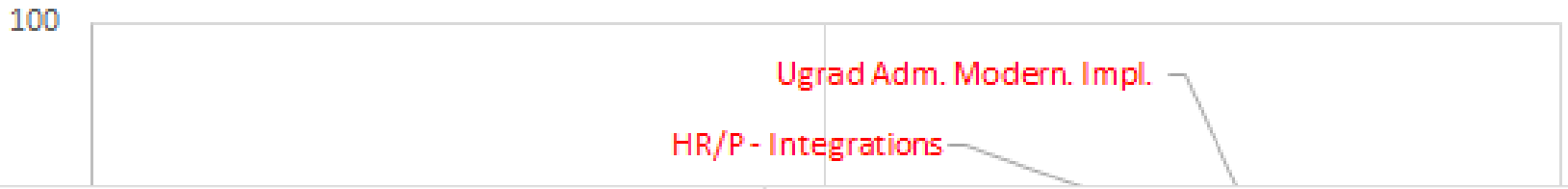
Develop and Deploy

All products will be adopted based on above architecture, with objective to develop an integrate suite ultimately leading to Identity as a Service. Some individual projects needed here. S. Zoppi Lead

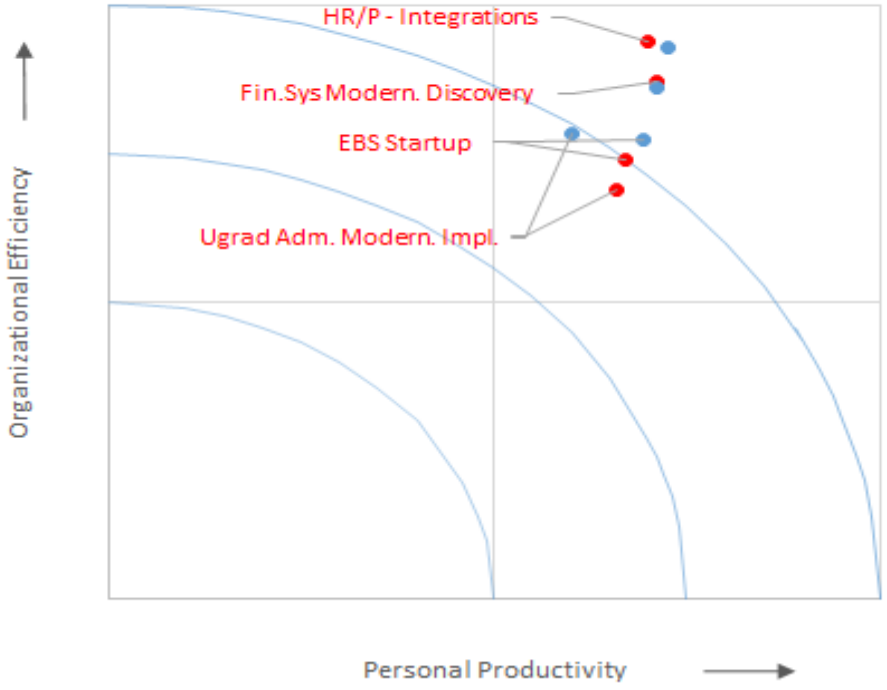
IT Service Investment Board Portfolio Prioritization Outcomes

- Teaching & Learning
  - —Academic Explorer
  - —MyHusky Experience – Implementation
  - —Curriculum Management - Build Out
- Administration/Business Systems
  - —Seattle Undergraduate Admissions Modernization
  - —HR/Payroll – Core Implementation and Integrations
  - —Enterprise Business Services Program - Startup
  - —Financial System Modernization: Discovery
- Research
  - —Storage, Consulting & Tools for Researchers
- Collaboration
  - —Network-based Collaboration Apps

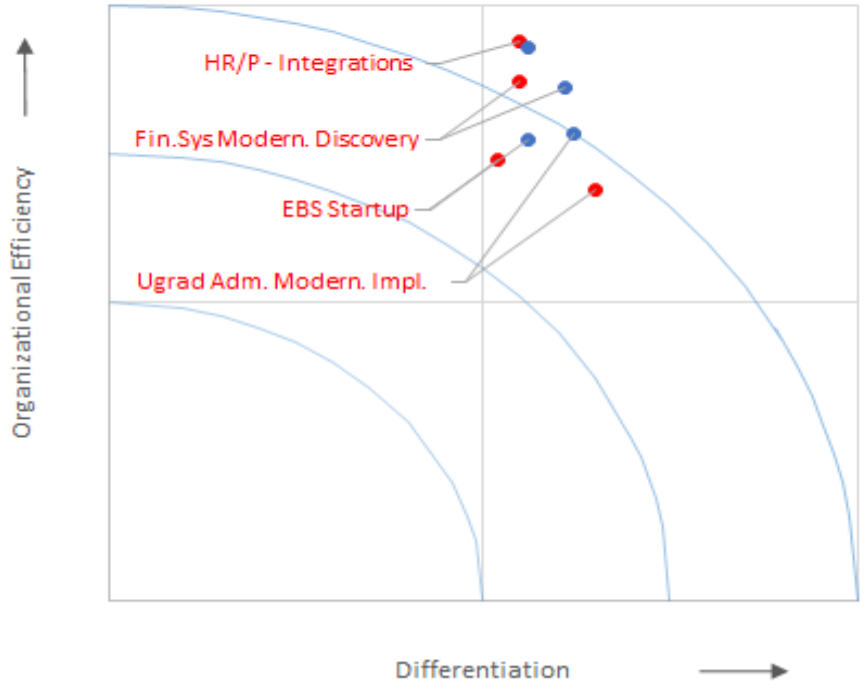
Admin / Business
Basic Metrics



Admin / Business
Who Benefits?

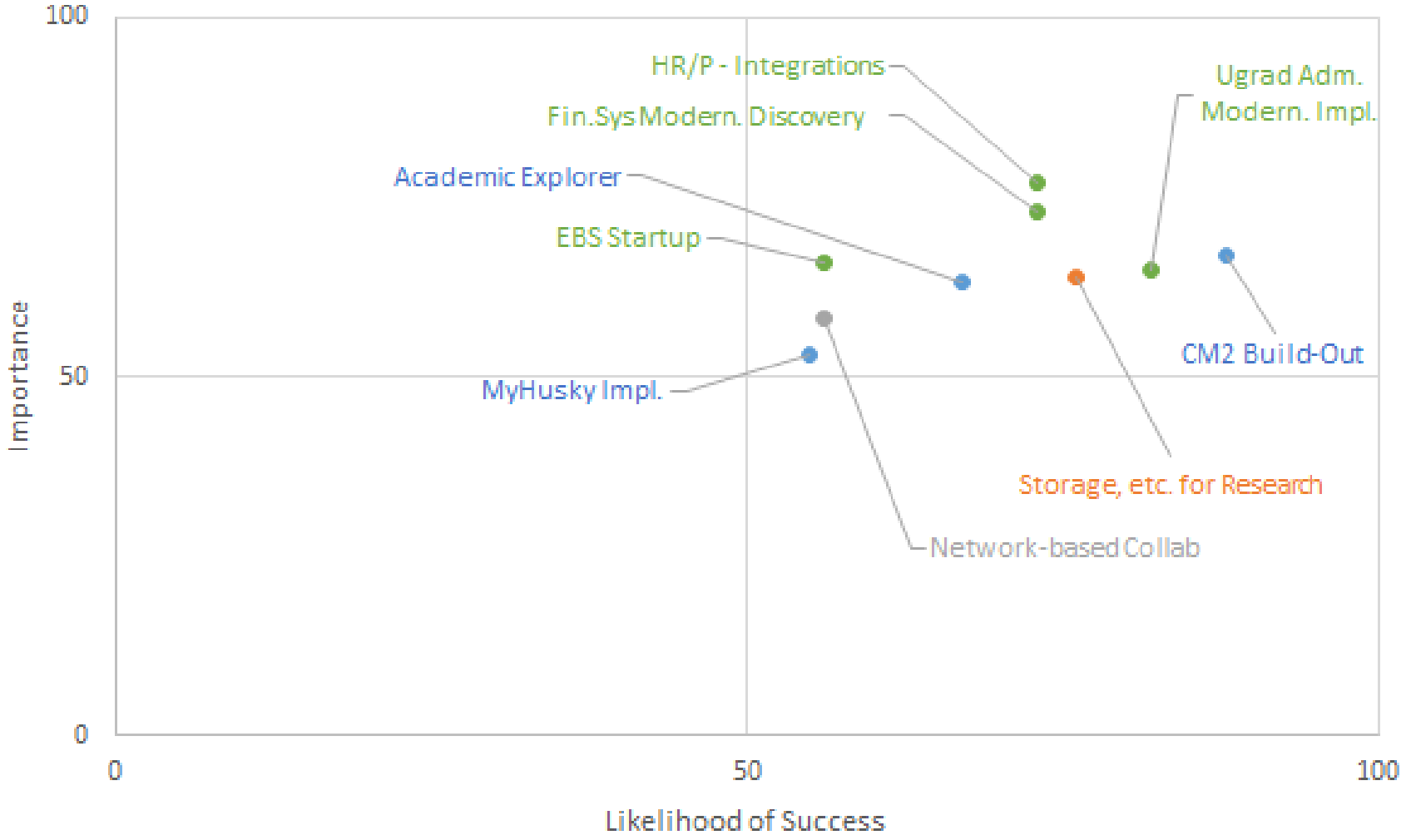


Admin / Business
Impact - Visible or Hidden?



SIB Business Cases

Basic Metrics

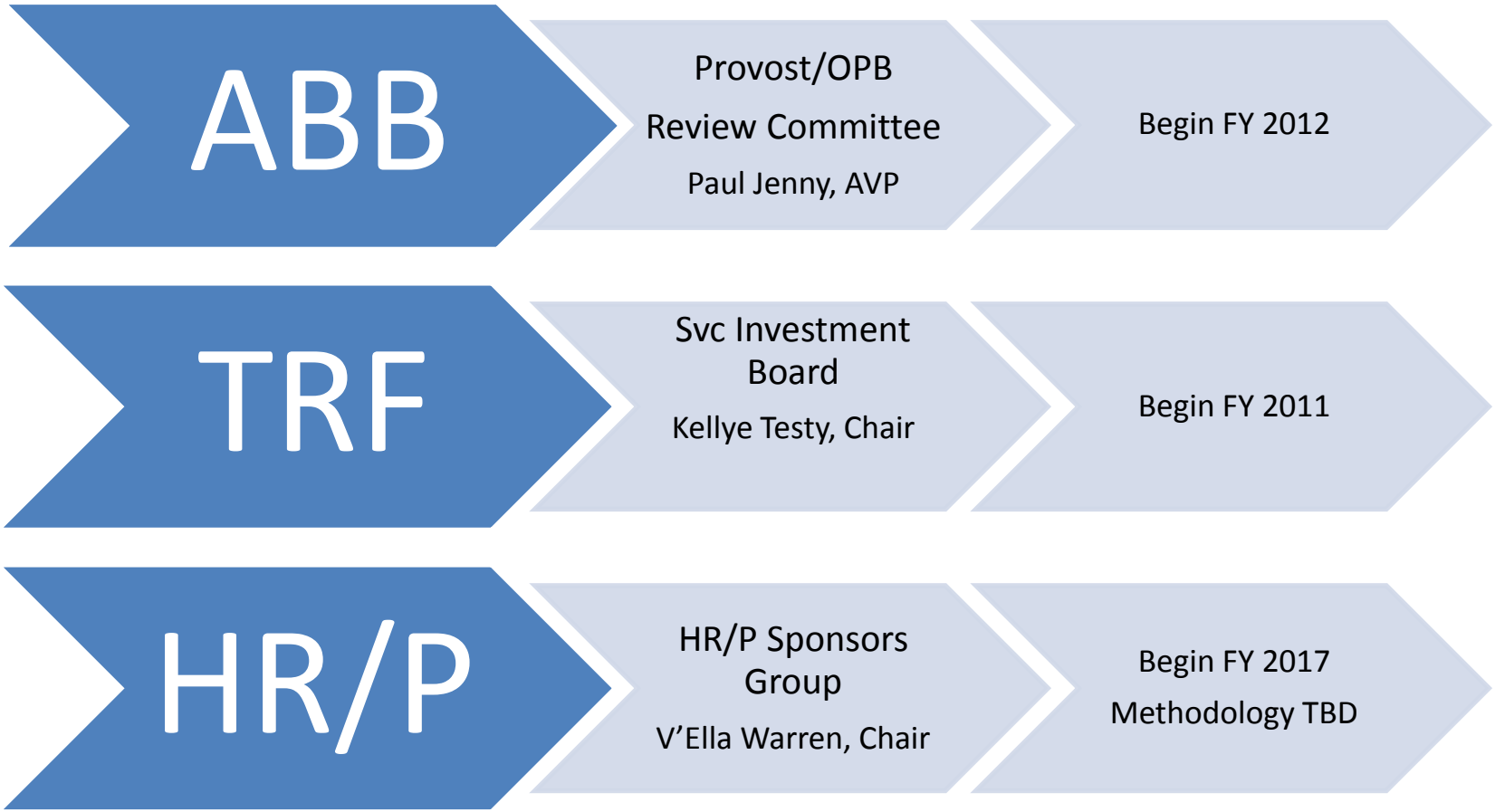


UW-IT Portfolio Prioritization Process Outcomes

- Hold the following projects
  - —MyHusky Experience
  - —Enterprise Business Services Program
  - —Network-based Collaboration Apps
- Use prioritization process outcomes to guide UW-IT FY 2015 project resource allocations
  - —Focus resources on high scoring projects when conflicts arise
  - —Identify other projects to slow down or hold

TRF Update

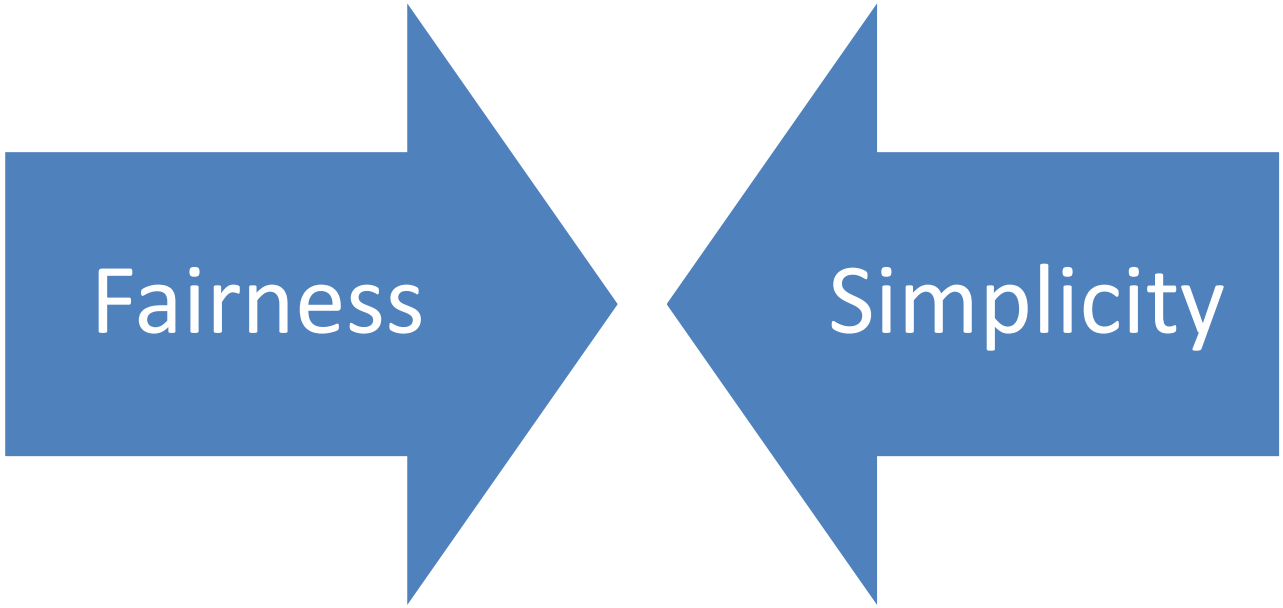
Parallel Processes



TRF Advisory Committee Timeline

- Spring >> Review Principles
  Discuss Methodology/Criteria
- Summer >> Develop Proposed UW-IT Budget
  Update Services
  Data Modeling for TRF
- Fall >> Discuss/Validate Outcomes
  Review with Service Investment Board

Conflicting Principles



TRF Advisory Committee feels we are close to the right balance and current methodology is “equally unfair”.

TRF Advisory Committee Outcomes – 4/7/14

- Focus on opportunities to reduce complexity and improve transparency
- Maintain current treatment of IT costs for students
- Explore alignment of TRF with current UW-IT organization and services
  - **Assess fiscal impact**
- Coordinate with HR/P Cost Allocation committee on per capita methodology

IT Project Portfolio Executive Review

Questions & Discussion