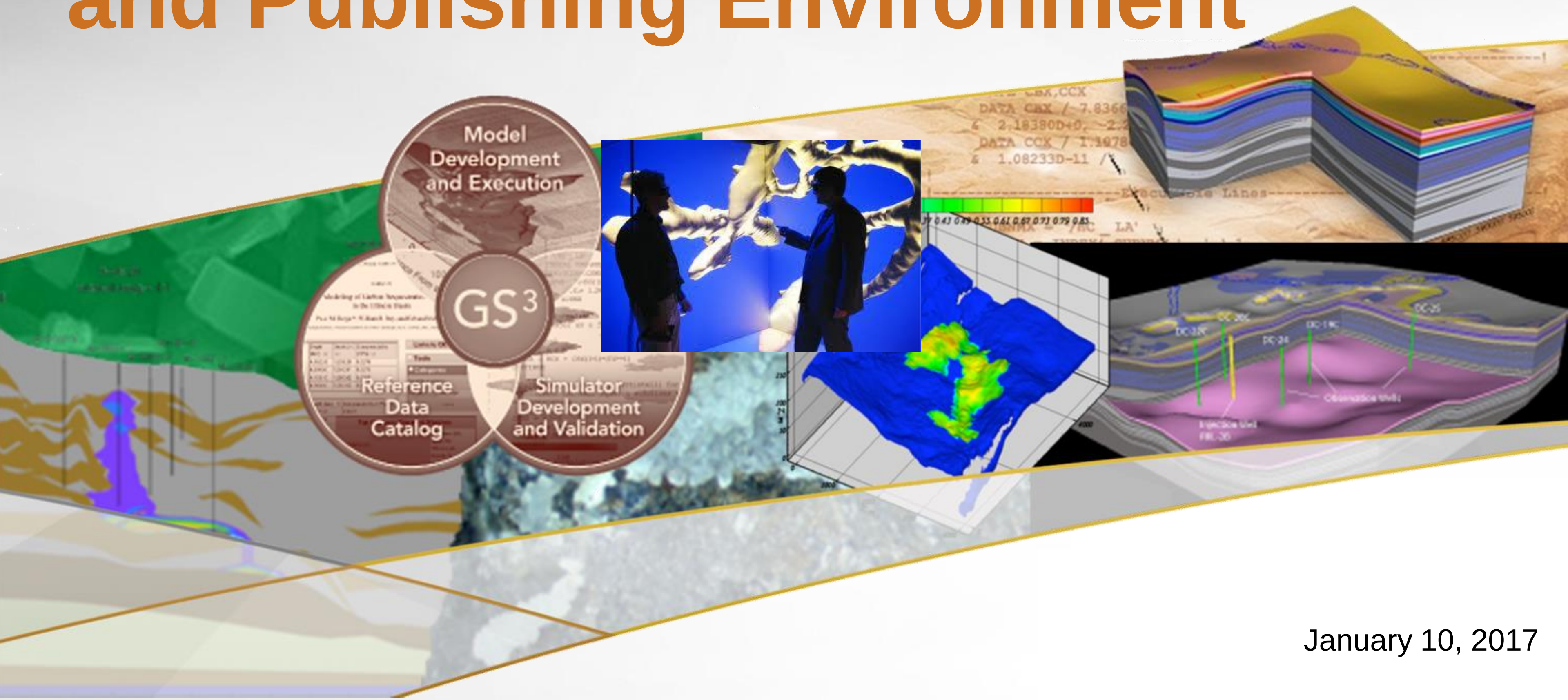
Pacific Northwest
NATIONAL LABORATORY
*Proudly Operated by* **Battelle** *Since 1965*

# Velo: A Collaborative Data Management, Simulation, Analysis, and Publishing Environment

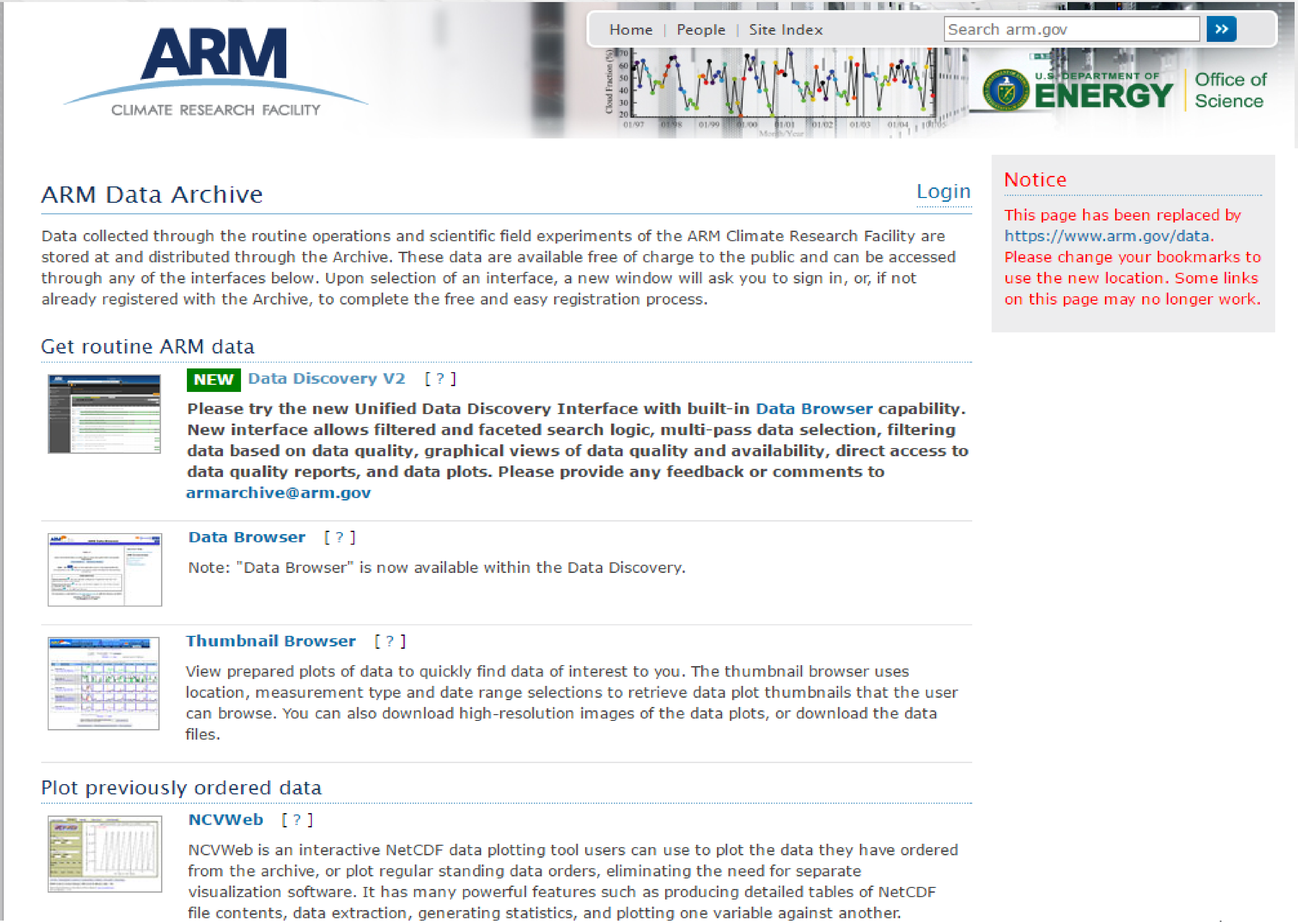

Model Development and Execution
GS³
Reference Data Catalog
Simulator Development and Validation

January 10, 2017

U.S. DEPARTMENT OF
ENERGY

# Scientific Knowledge Management

- Scientists generate massive amounts of data
  - Experiments
  - Simulations
  - Instruments
- Processing this data produces knowledge
  - Data analysis
  - Data mining
  - Machine learning
  - Visualization


ARM
CLIMATE RESEARCH FACILITY
Home | People | Site Index
Search arm.gov
U.S. DEPARTMENT OF ENERGY
Office of Science

## ARM Data Archive

Login

Data collected through the routine operations and scientific field experiments of the ARM Climate Research Facility are stored at and distributed through the Archive. These data are available free of charge to the public and can be accessed through any of the interfaces below. Upon selection of an interface, a new window will ask you to sign in, or, if not already registered with the Archive, to complete the free and easy registration process.

### Get routine ARM data

**NEW** Data Discovery V2 [ ? ]

**Please try the new Unified Data Discovery Interface with built-in Data Browser capability. New interface allows filtered and faceted search logic, multi-pass data selection, filtering data based on data quality, graphical views of data quality and availability, direct access to data quality reports, and data plots. Please provide any feedback or comments to armarchive@arm.gov**

**Data Browser** [ ? ]

Note: "Data Browser" is now available within the Data Discovery.

**Thumbnail Browser** [ ? ]

View prepared plots of data to quickly find data of interest to you. The thumbnail browser uses location, measurement type and date range selections to retrieve data plot thumbnails that the user can browse. You can also download high-resolution images of the data plots, or download the data files.

### Plot previously ordered data

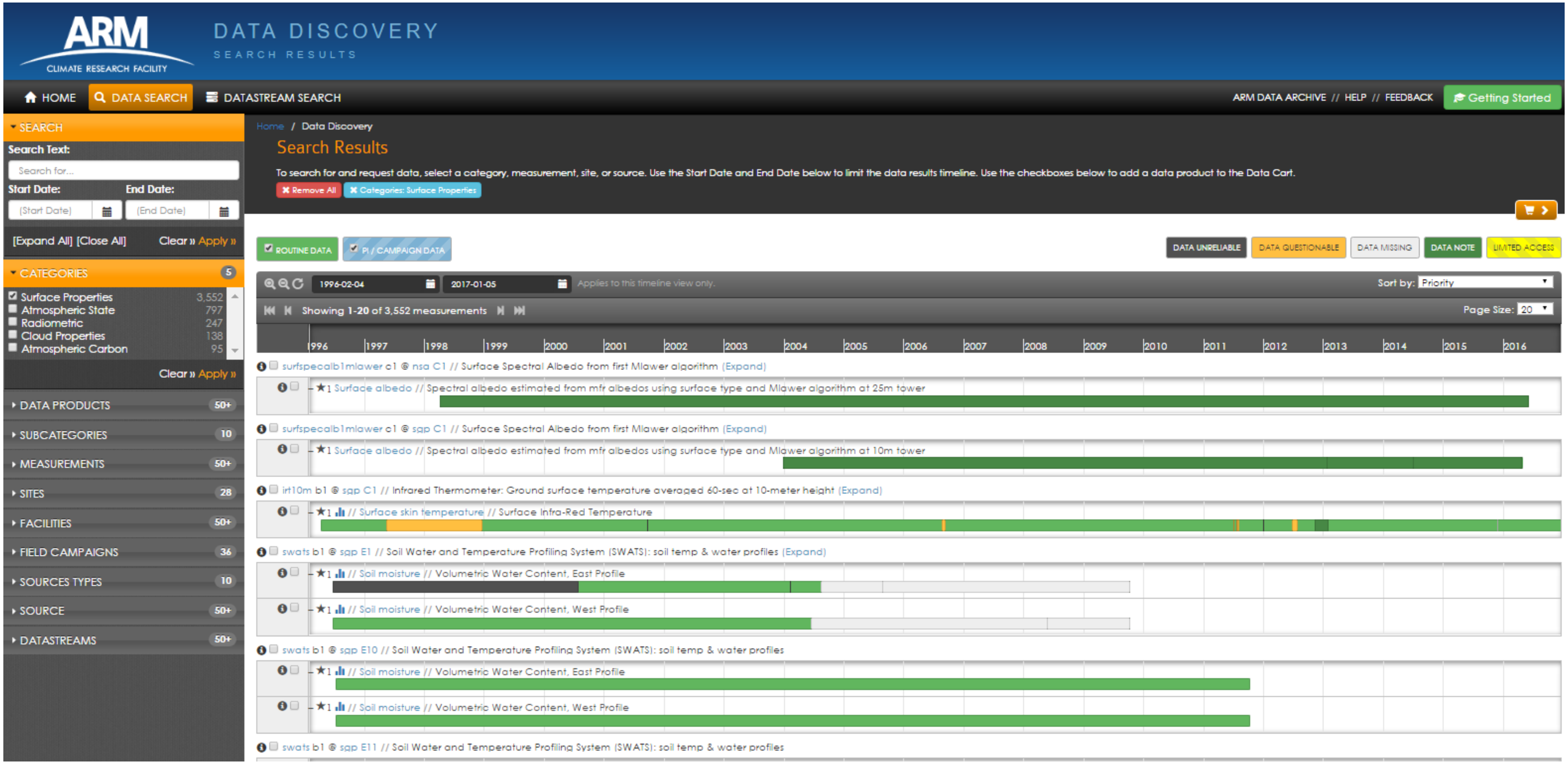
**NCVWeb** [ ? ]

NCVWeb is an interactive NetCDF data plotting tool users can use to plot the data they have ordered from the archive, or plot regular standing data orders, eliminating the need for separate visualization software. It has many powerful features such as producing detailed tables of NetCDF file contents, data extraction, generating statistics, and plotting one variable against another.

### Notice

This page has been replaced by https://www.arm.gov/data. Please change your bookmarks to use the new location. Some links on this page may no longer work.

# http://www.archive.arm.gov/discovery/#v/results/s/fcat::sfcprop


ARM
CLIMATE RESEARCH FACILITY
DATA DISCOVERY
SEARCH RESULTS
HOME
DATA SEARCH
DATASTREAM SEARCH
ARM DATA ARCHIVE // HELP // FEEDBACK
Getting Started
SEARCH
Search Text:
Search for...
Start Date:
End Date:
(Start Date)
(End Date)
[Expand All] [Close All]
Clear » Apply »
CATEGORIES 5
Clear » Apply »
DATA PRODUCTS 50+
SUBCATEGORIES 10
MEASUREMENTS 50+
SITES 28
FACILITIES 50+
FIELD CAMPAIGNS 36
SOURCES TYPES 10
SOURCE 50+
DATASTREAMS 50+
Home / Data Discovery
✖ Remove All
✖ Categories: Surface Properties
☑ ROUTINE DATA
☑ PI / CAMPAIGN DATA
DATA UNRELIABLE
DATA QUESTIONABLE
DATA MISSING
DATA NOTE
LIMITED ACCESS
Applies to this timeline view only.
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

☑ Surface Properties 3,552
☐ Atmospheric State 797
☐ Radiometric 247
☐ Cloud Properties 138
☐ Atmospheric Carbon 95

## Search Results

To search for and request data, select a category, measurement, site, or source. Use the Start Date and End Date below to limit the data results timeline. Use the checkboxes below to add a data product to the Data Cart.

1996-02-04 2017-01-05

Sort by: Priority

Showing 1-20 of 3,552 measurements

Page Size: 20

☐ surfspecalb1mlawer c1 @ nsa C1 // Surface Spectral Albedo from first Mlawer algorithm (Expand)
- ☐ ★1 Surface albedo // Spectral albedo estimated from mfr albedos using surface type and Mlawer algorithm at 25m tower

☐ surfspecalb1mlawer c1 @ sgp C1 // Surface Spectral Albedo from first Mlawer algorithm (Expand)
- ☐ ★1 Surface albedo // Spectral albedo estimated from mfr albedos using surface type and Mlawer algorithm at 10m tower

☐ irt10m b1 @ sgp C1 // Infrared Thermometer: Ground surface temperature averaged 60-sec at 10-meter height (Expand)
- ☐ ★1 // Surface skin temperature // Surface Infra-Red Temperature

☐ swats b1 @ sgp E1 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles (Expand)
- ☐ ★1 // Soil moisture // Volumetric Water Content, East Profile
- ☐ ★1 // Soil moisture // Volumetric Water Content, West Profile

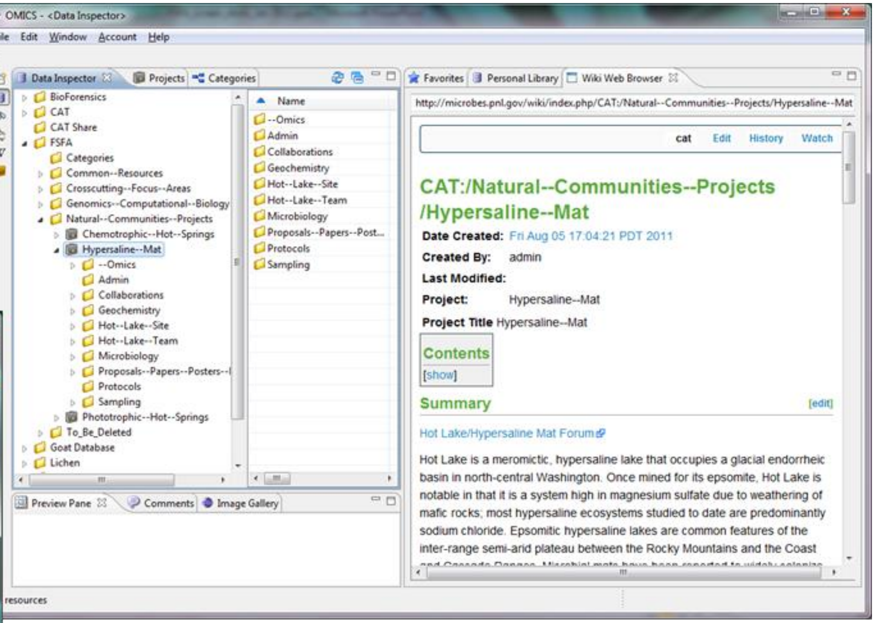
☐ swats b1 @ sgp E10 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles
- ☐ ★1 // Soil moisture // Volumetric Water Content, East Profile
- ☐ ★1 // Soil moisture // Volumetric Water Content, West Profile

☐ swats b1 @ sgp E11 // Soil Water and Temperature Profiling System (SWATS): soil temp & water profiles

# Great Site, but ….

- Expensive to build
  - Millions of dollars
- Expensive to run/maintain
  - Own servers, disks, administration, data center
- Expensive to evolve and scale
  - More and more data
    - We're talking petabytes …
  - More and more features needed
  - More and more to support

# Bioinformatics Example


OMICS - <Data Inspector>
File Edit Window Account Help
Data Inspector
Projects
Categories
BioForensics
CAT
CAT Share
FSFA
Categories
Common--Resources
Crosscutting--Focus--Areas
Genomics--Computational--Biology
Natural--Communities--Projects
Chemotrophic--Hot--Springs
Hypersaline--Mat
--Omics
Admin
Collaborations
Geochemistry
Hot--Lake--Site
Hot--Lake--Team
Microbiology
Proposals--Papers--Posters--I
Protocols
Sampling
Phototrophic--Hot--Springs
To_Be_Deleted
Goat Database
Lichen
Name
--Omics
Admin
Collaborations
Geochemistry
Hot--Lake--Site
Hot--Lake--Team
Microbiology
Proposals--Papers--Post...
Protocols
Sampling
Preview Pane
Comments
Image Gallery
Favorites
Personal Library
Wiki Web Browser
http://microbes.pnl.gov/wiki/index.php/CAT:/Natural--Communities--Projects/Hypersaline--Mat
cat
Edit
History
Watch
resources

## CAT:/Natural--Communities--Projects/Hypersaline--Mat

**Date Created:** Fri Aug 05 17:04:21 PDT 2011

**Created By:** admin

**Last Modified:**

**Project:** Hypersaline--Mat

**Project Title** Hypersaline--Mat

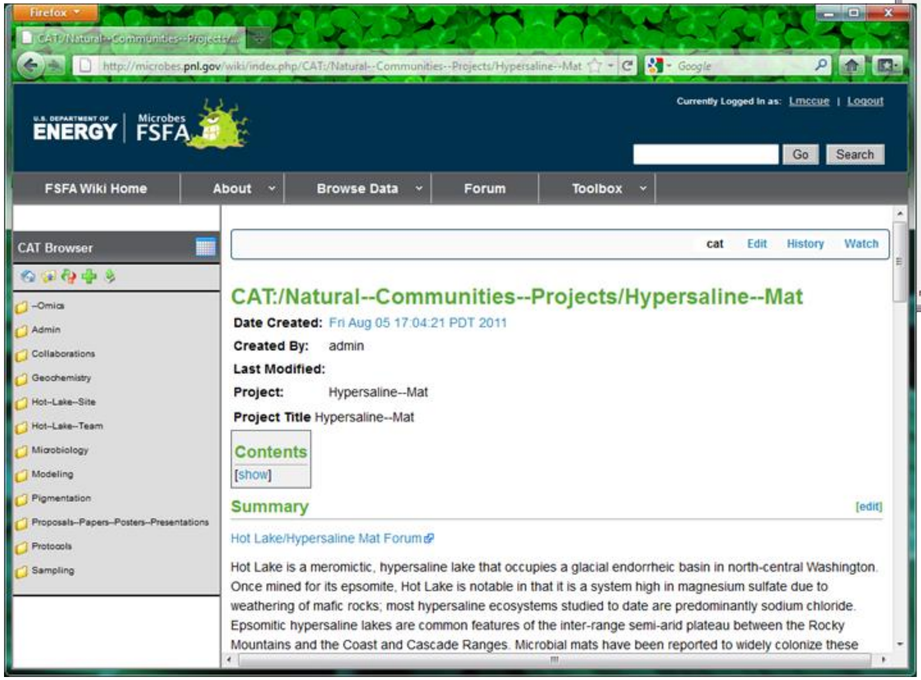
Contents
[show]

### Summary

[edit]

Hot Lake/Hypersaline Mat Forum

Hot Lake is a meromictic, hypersaline lake that occupies a glacial endorrheic basin in north-central Washington. Once mined for its epsomite, Hot Lake is notable in that it is a system high in magnesium sulfate due to weathering of mafic rocks; most hypersaline ecosystems studied to date are predominantly sodium chloride. Epsomitic hypersaline lakes are common features of the inter-range semi-arid plateau between the Rocky Mountains and the Coast


Firefox
CAT:/Natural--Communities--Projects/...
http://microbes.pnl.gov/wiki/index.php/CAT:/Natural--Communities--Projects/Hypersaline--Mat
Google
U.S. DEPARTMENT OF ENERGY
Microbes FSFA
Currently Logged In as: Lmccue | Logout
Go
Search
FSFA Wiki Home
About
Browse Data
Forum
Toolbox
CAT Browser
--Omics
Admin
Collaborations
Geochemistry
Hot--Lake--Site
Hot--Lake--Team
Microbiology
Modeling
Pigmentation
Proposals--Papers--Posters--Presentations
Protocols
Sampling
cat
Edit
History
Watch

CAT:/Natural--Communities--Projects/Hypersaline--Mat

**Date Created:** Fri Aug 05 17:04:21 PDT 2011

**Created By:** admin

**Last Modified:**

**Project:** Hypersaline--Mat

**Project Title** Hypersaline--Mat

Contents
[show]

Summary

[edit]

Hot Lake/Hypersaline Mat Forum

Hot Lake is a meromictic, hypersaline lake that occupies a glacial endorrheic basin in north-central Washington. Once mined for its epsomite, Hot Lake is notable in that it is a system high in magnesium sulfate due to weathering of mafic rocks; most hypersaline ecosystems studied to date are predominantly sodium chloride. Epsomitic hypersaline lakes are common features of the inter-range semi-arid plateau between the Rocky Mountains and the Coast and Cascade Ranges. Microbial mats have been reported to widely colonize these

# Implications

- Thousands of different scientific research projects
- All with:
  - ‘similar’ data management and analysis needs
  - Limited budgets
  - Scaling and growth requirements

# So what do scientist do?

- Use wikis
- Use simple web site
- Use Sharepoint
- Use primitive tools that haven’t changes for 20 years!

DO
THINGS
THAT
~~SCALE~~
DON'T
SCALE
Paul Graham

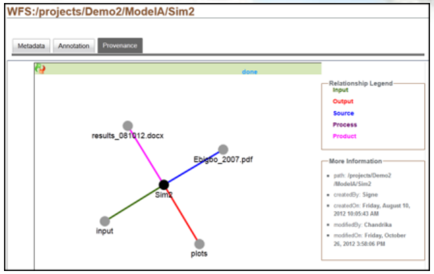
# What about Content Management Technologies?

In Progress
34% complete

# Velo

- We built Velo 7 years ago at PNNL
    - Underlying data management 10+ years ago
- Used by many science projects
- Built on Alfresco
- Works well but has:
    - scalability challenges
    - Enhanced incrementally through small projects over several years – no clean redesign
    - Not easy for a new science project to deploy
        - Software/hardware configuration

# That’s where you come in …



WFS:/projects/Demo2/ModelA/Sim2

Metadata
Annotation
Provenance
done
results_081012.docx
Ebigbo_2007.pdf
Sim2
input
plots
Relationship Legend
Input
Output
Source
Process
Product
More Information

- path: /projects/Demo2/ModelA/Sim2
- createdBy: Signe
- createdOn: Friday, August 10, 2012 10:05:43 AM
- modifiedBy: Chandrika
- modifiedOn: Friday, October 26, 2012 3:58:06 PM

Web Services for the Real World
RESTful
Web Services
O'REILLY

amazon
webservices™

Collaboration
Data access
Visualization

Security
Upload/download
Navigation
Site configuration

Data storage
Search
Analysis
Metadata

designed by freepik.com

# Before we go in to the details

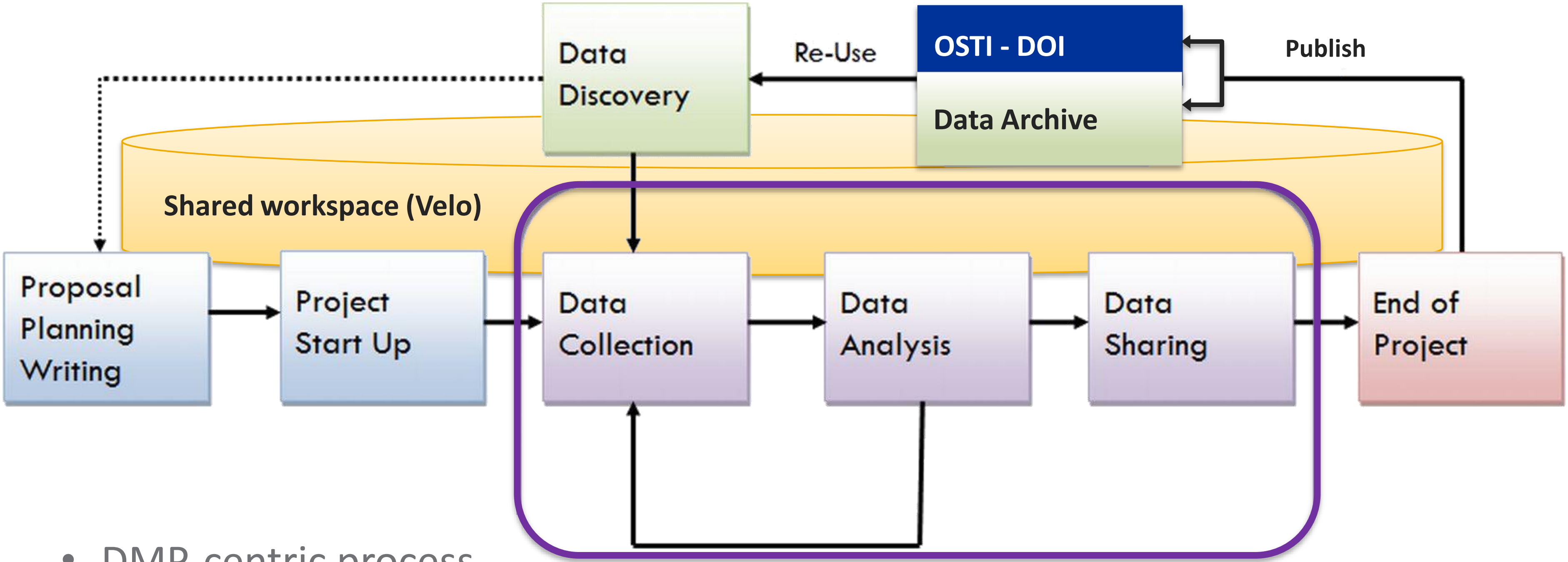
- Expect to be confused!!!!
- Ask questions!!!
- Don’t be afraid to be creative
- You won’t get all the requirements ‘on a plate’
  - You will have to discover some ….
  - You can create some ….
- Technology selections are pretty open
  - AWS based
  - Data storage/database?
  - Web technologies?
  - Languages?

# PNNL - Introductions

- Gary Black
- Carina Lansing
- Ellen Porter
- Bibi Raju
- Zoe Guillen

# Vision

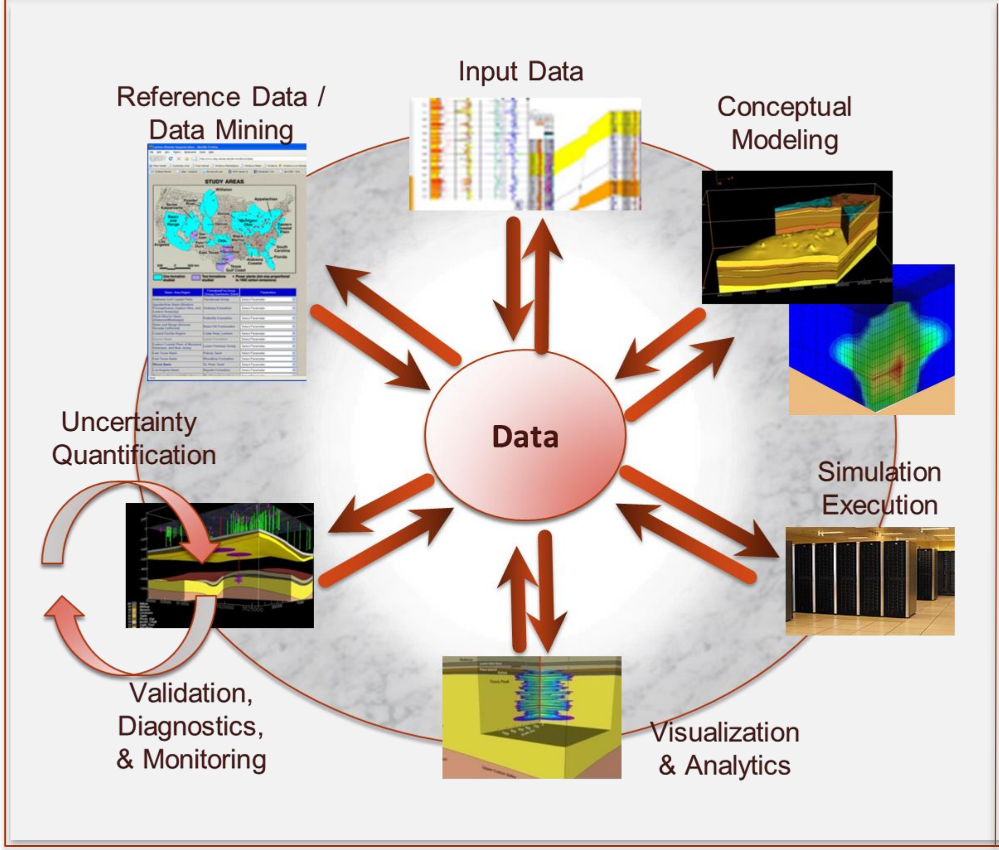
## Full Lifecycle Support for Research Data

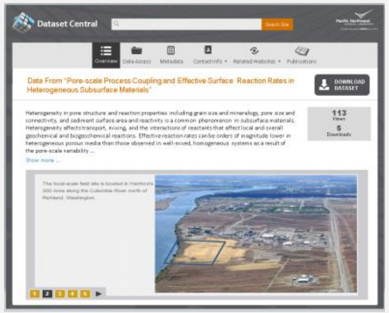

Data Discovery
Re-Use
OSTI - DOI
Data Archive
Publish
Shared workspace (Velo)
Proposal Planning Writing
Project Start Up
Data Collection
Data Analysis
Data Sharing
End of Project

- DMP-centric process.
- Shared workspace for all project documents and research data.
- Seamless integration with PNNL systems from project proposal through publication.
- Long term:  facilitate data discovery and reuse.

# Velo Scope

Do your work... → Publish... → Discovery & Reuse


Input Data
Conceptual Modeling
Simulation Execution
Visualization & Analytics
Validation, Diagnostics, & Monitoring
Uncertainty Quantification
Reference Data / Data Mining
Data


Dataset Central
Search Site
Overview
Data Access
Metadata
Contact Info
Related Websites
Publications
DOWNLOAD DATASET
113 Views
5 Downloads
1 2 3 4 5

Data From "Pore-scale Process Coupling and Effective Surface Reaction Rates in Heterogeneous Subsurface Materials"

Heterogeneity in pore structure and reaction properties including grain size and mineralogy, pore size and connectivity, and sediment surface area and reactivity is a common phenomenon in subsurface materials. Heterogeneity affects transport, mixing, and the interactions of reactants that affect local and overall geochemical and biogeochemical reactions. Effective reaction rates can be orders of magnitude lower in heterogeneous porous media than those observed in well-mixed, homogeneous systems as a result of the pore-scale variability ...

Show more ...

The local-scale field site is located in Hanford's 300 Area along the Columbia River north of Richland, Washington.

DATA.GOV

osti.gov

Google

ENERGY.GOV

ResearchGate

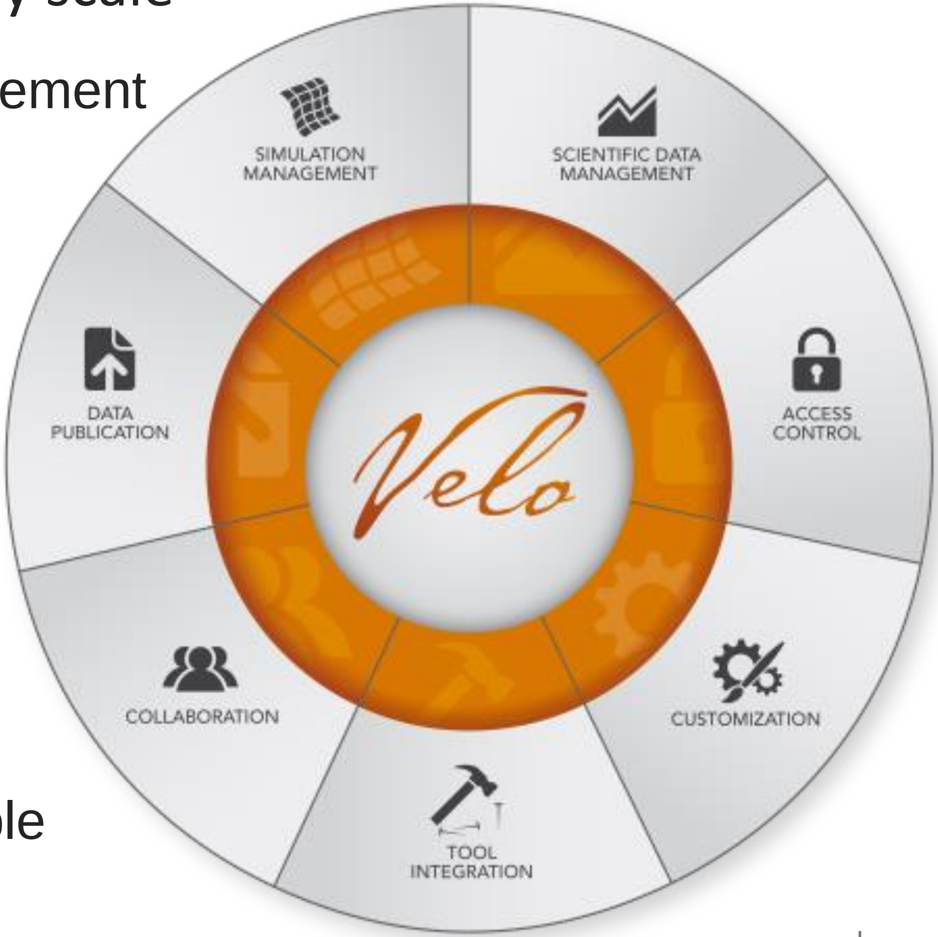
nanoHUB

NGDS
National Geothermal Data System

↔ Velo

# Key Features

- Full life cycle, integrated support for modeling, simulation, analysis, and visualization at any scale
- Data/metadata/relationship management
- Tool Integration
  - Connect tools to data
  - Chain tools together
- Simulation management, including UQ ensembles
- Collaboration across multi-disciplinary teams
- Provenance – who did what, when
- Open, Extensible, and Customizable


SIMULATION MANAGEMENT
SCIENTIFIC DATA MANAGEMENT
ACCESS CONTROL
CUSTOMIZATION
TOOL INTEGRATION
COLLABORATION
DATA PUBLICATION
Velo

# Key Benefits

- Facilitates reproducibility & discovery
  - Metadata is way harder to collect after the project is completed.
- Focused on integration – don't reinvent the wheel
- Improved organization
  - Easier for users to track where their different data parts are in a project
- Improved productivity
  - Less time spent in data orchestration tasks = more time for science
  - Low barrier of entry for HPC systems

Pacific Northwest
NATIONAL LABORATORY

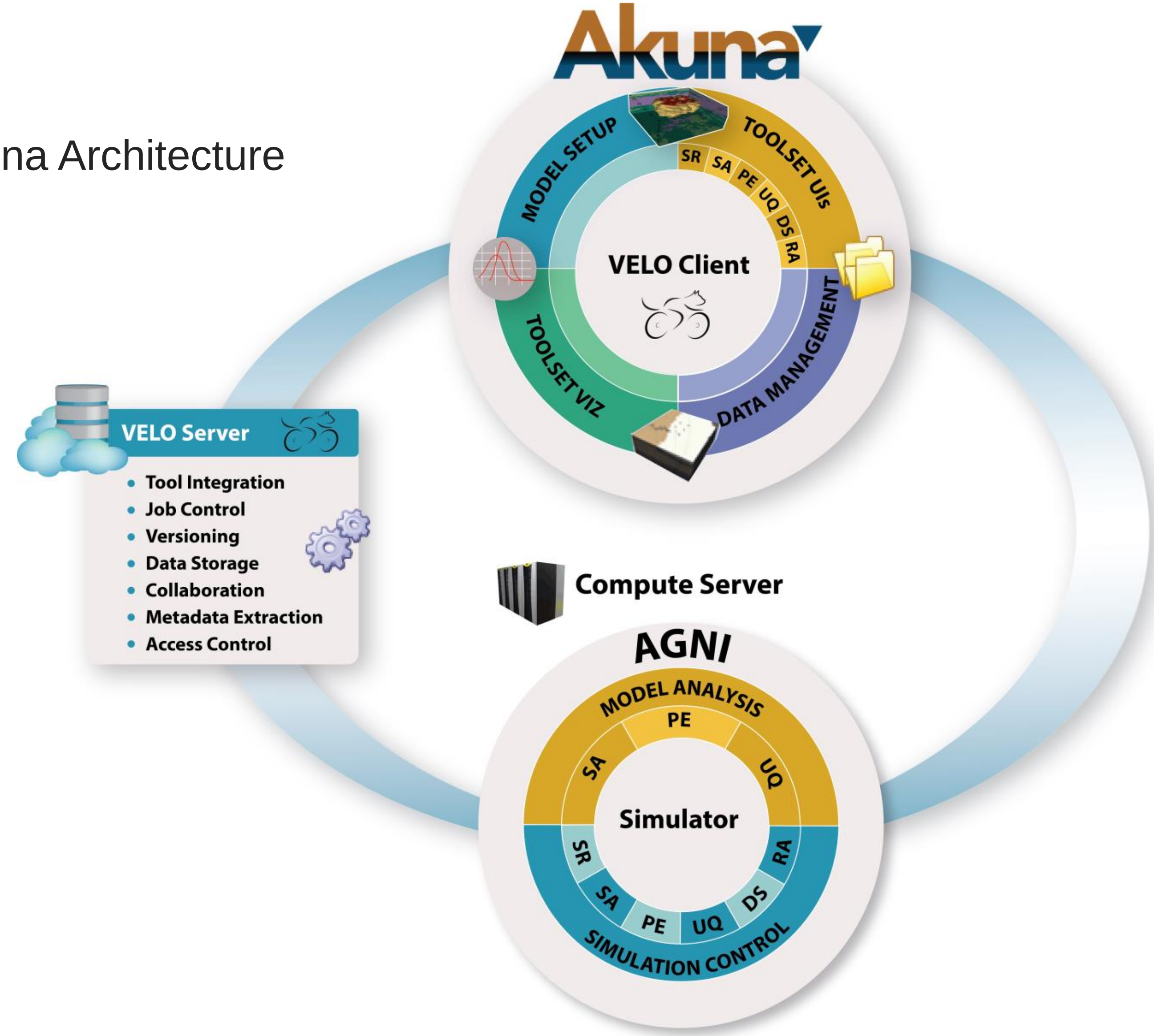
*Proudly Operated by* **Battelle** *Since 1965*

# Velo Use Case: Akuna

U.S. DEPARTMENT OF
ENERGY

# Velo Use Case: Akuna

- Akuna Architecture



Akuna

VELO Client

MODEL SETUP

TOOLSET UIs

SR SA PE UQ DS RA

DATA MANAGEMENT

TOOLSET VIZ

**VELO Server**

- Tool Integration
- Job Control
- Versioning
- Data Storage
- Collaboration
- Metadata Extraction
- Access Control

Compute Server

AGNI

MODEL ANALYSIS

PE

SA

UQ

Simulator

SR

SA

PE

UQ

DS

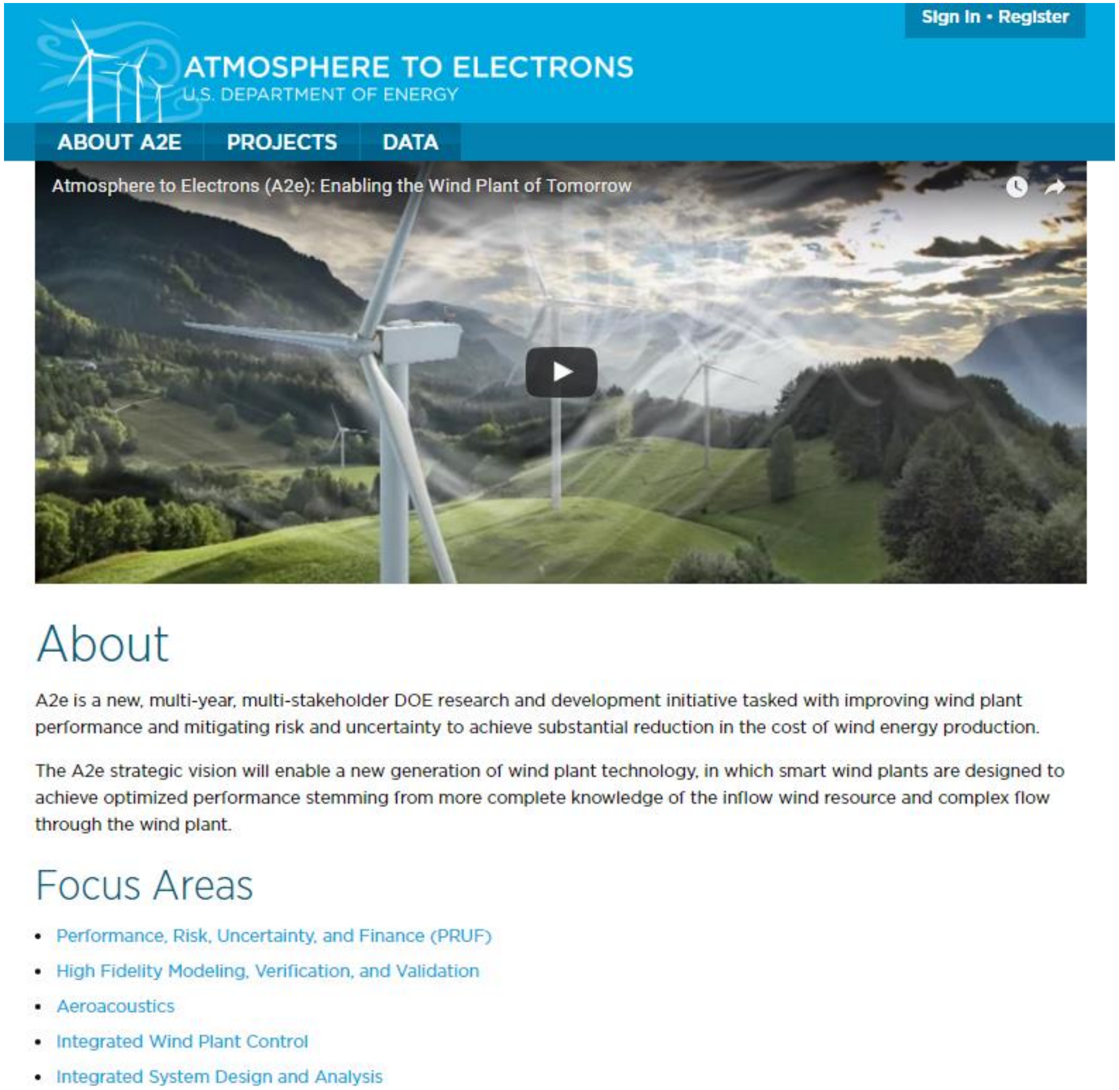
RA

SIMULATION CONTROL

# What Next …

- Lot’s of questions ….
- Target is Week 4 for defining first sprint
  - Scope
  - Requirements
  - Teams
  - Technologies
- Need to form teams around subsystems, eg:
  - API – reliable, scalable, high performance
  - Data storage - both science data and metadata
  - Web UI – site creation, user definition, data/metadata browsing and access
  - Security – role based definition and  access
- And all this needs to leverage AWS capabilities …

# A2E Use Case


Sign In • Register
ATMOSPHERE TO ELECTRONS
U.S. DEPARTMENT OF ENERGY
ABOUT A2E
PROJECTS
DATA
Atmosphere to Electrons (A2e): Enabling the Wind Plant of Tomorrow

## About

A2e is a new, multi-year, multi-stakeholder DOE research and development initiative tasked with improving wind plant performance and mitigating risk and uncertainty to achieve substantial reduction in the cost of wind energy production.

The A2e strategic vision will enable a new generation of wind plant technology, in which smart wind plants are designed to achieve optimized performance stemming from more complete knowledge of the inflow wind resource and complex flow through the wind plant.

## Focus Areas

- Performance, Risk, Uncertainty, and Finance (PRUF)
- High Fidelity Modeling, Verification, and Validation
- Aeroacoustics
- Integrated Wind Plant Control
- Integrated System Design and Analysis

# The end game?

- Some possibilities:
- End of class presentation/seminar to PNNL scientists (and other institutions both in Boston and Seattle)
- Code open source on git
- Technical report/paper with whole class as contributors
- Youtube demos of the new Velo technology
- Performance and scalability studies to demonstrate improvements over existing Velo platform and explain what is possible with AWS
- Examples of complex analysis and visualizations of science data that we get from A2E or other projects
- ??????

# Logistics

- [https://github.com/gortonator/CS6510-Advanced-Software-Development/blob/master/docs/CS6510-syllabus.md](https://github.com/gortonator/CS6510-Advanced-Software-Development/blob/master/docs/CS6510-syllabus.md)
  - Includes recommended reading for each week
  - Important to adopt tooling for agile processes
- Basically:
  - Weeks 1-3: investigate problem, technologies, possible solutions, form teams, share ideas, ask questions
  - Week 4: define requirements, spring 1 targets
  - Week 7: Report on Sprint 1 outcomes, specify Sprint 2 plans
  - Week 10: Report on Sprint 2 outcomes, specify Sprint 3 plans
  - Week 11 – code reviews/retrospectives
  - Week 14: The Big Deal ☺