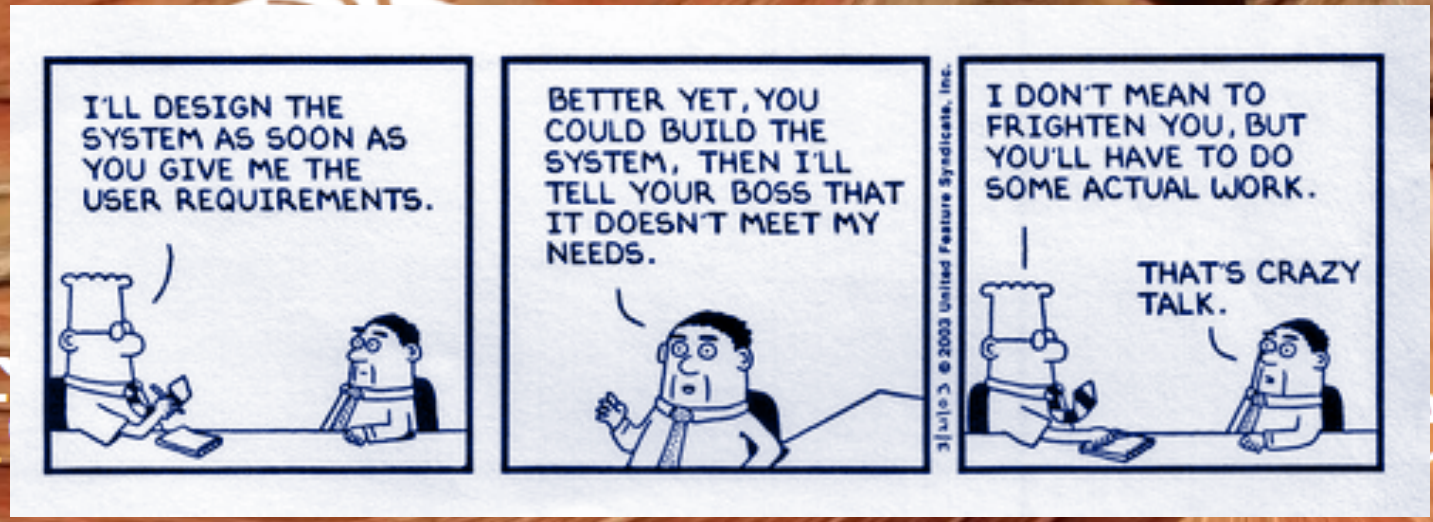
EarthCube's Use Case Collection Project

Why we did it, what we learned

Lisa Kempler, lisak@mathworks.com
Karen Stocks, kstocks@ucsd.edu

on behalf of:
EarthCube Technology and Architecture Committee's Use Case Working Group

Transf esearch



With a System of Systems (SoS) Infrastructure

earthcube.org

Why EarthCube?

https://goo.gl/images/PkRVVo

Goals of the Project

- **Make end-user geoscientists' needs visible and accessible to EarthCube technologists**
  - Get user stories from practicing geoscientists
  - Represent the range of geoscience disciplines
  - Ensure technical challenges are specific and detailed

- **Identify key barriers**

Process: Rigorous and Formal

Defined framework

- Structured Template
- 1-hour formal phone interview for completing template
  - Adhering to the template
  - Capturing everything that's remotely relevant

Use Case Template Sections: Enable fast, deep dive into the scenario details

- Use Case Name
- Contacts
- Link to primary doc
- Permissions (public)
- Objectives and Outcomes
- Science Drivers

- Actors
- Preconditions
- Critical Existing Cyberinfrastructure
- Measures of Success
- Basic Flow
- Alternate Flow
- Activity Diagram
- Major Outcome and Post Conditions
- Problems/Challenges
- References
- Notes

Template Sections (continued): Technical

Data Characteristics:

- **Data Source**
  - *Example:*
    - *Historical input data is supplied by NOAA on their publicly available data cloud.*
- **Data Format**
  - *Example:*
    - *netCDF, .csv, etc.*
- **Volume (size)**
  - *Examples:*
    - *DES: 4PB, ZTF: 1PB/yr, LSST: 7PB/yr, Simulations > 10PB in 2017*
- **Velocity (e.g., real time)**
  - *Example:*
    - *LSST: 20TB/day*
- **Variety (multiple datasets, mashup)**
  - *Examples:*
    - *1) Raw Data from sky surveys*
    - *2) Processed Image data*
    - *3) Simulation data,*
    - *4) sequence data*
- **Variability**
  - *Example:*
    - *Observations are taken nightly; supporting simulations are run throughout the year, but data can be produced sporadically depending on access to*
- **Veracity/Data Quality (accuracy, precision)**
  - *Example:*
    - *Hydrographic data uses the World Ocean Circulation Experiment*

Standards

List any standards that were followed for the cyberinfrastructure resources, even if already mentioned above. Standards most commonly apply to data, but can apply to models, metadata, etc.

- *Example:*
  - *netCDF data format*
  - *World Ocean Circulation Experiment (WOCE) quality assessment convention and flags*

Data Visualization and Analytics

Format for visualization

- *Example:*
  - *.vtk, .tiff, .kml, netCDF*

Software

For any important software used, describe the the important characteristics (source, language, input format, output format, CPU requirements etc)

Metadata

Provide a link to, or include any relevant metadata which can add additional detail and context the dataset(s) described above.

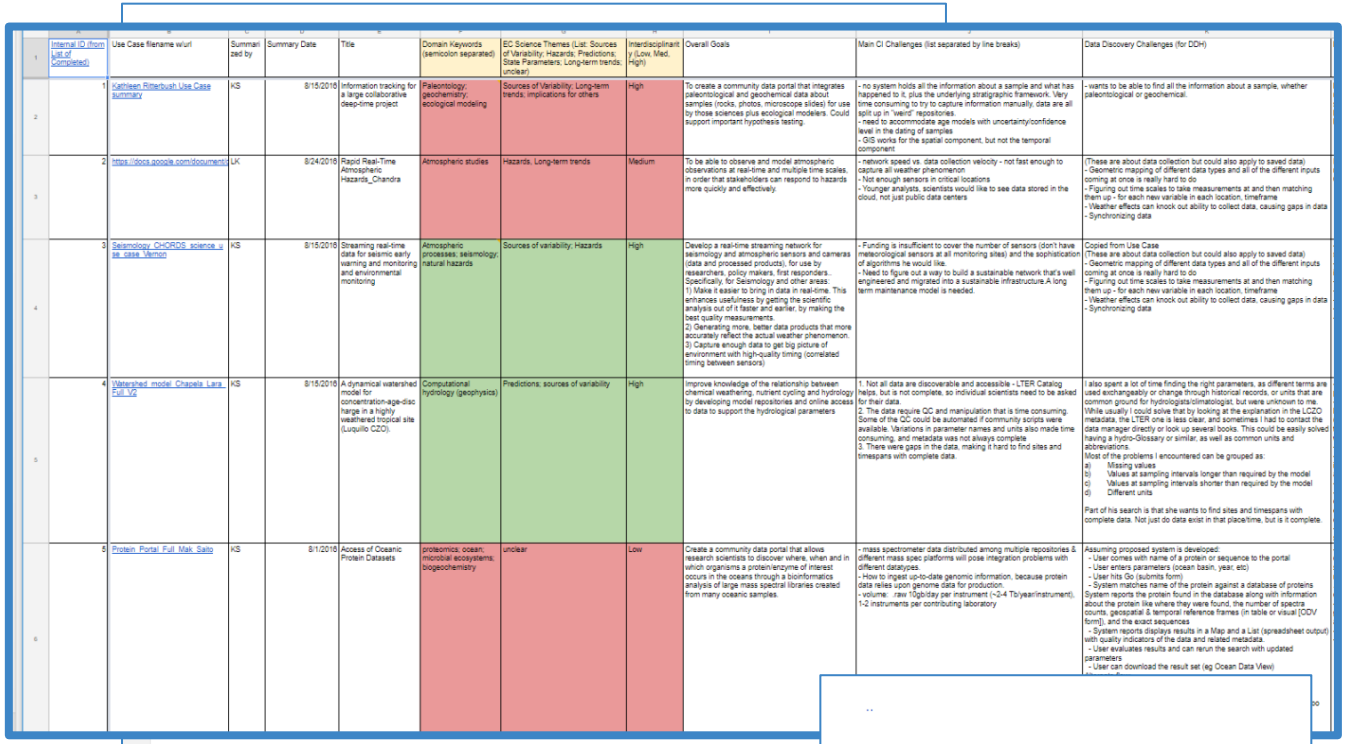
Process: Rigorous and Ad hoc

- **Defined framework**
  - Structured Template
  - Formal phone call interview setup for collecting input
- **Ad hoc realities**
  - We pulled from the pool we had access to: volunteers
    - This is a standard product management practice
    - The self-identifiers are your target audience
- **Criteria? Ex: Represent range of geoscience disciplines**
  - Best differentiator?
    - vs. institution type, computational skill, age, public vs. private, project size,
    - Ideally, you'd segment/target your market

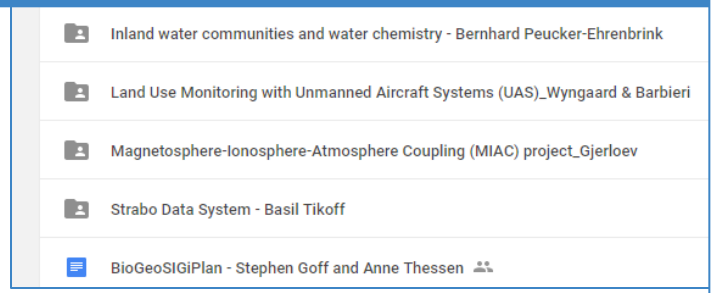
atmospheric sciences
biogeosciences
geodesy
geomagnetism
paleomagnetism and
electromagnetism
hydrology
ocean sciences
planetary sciences
seismology
space physics
aeronomy
tectonophysics
volcanology
geochemistry
petrology.[40]

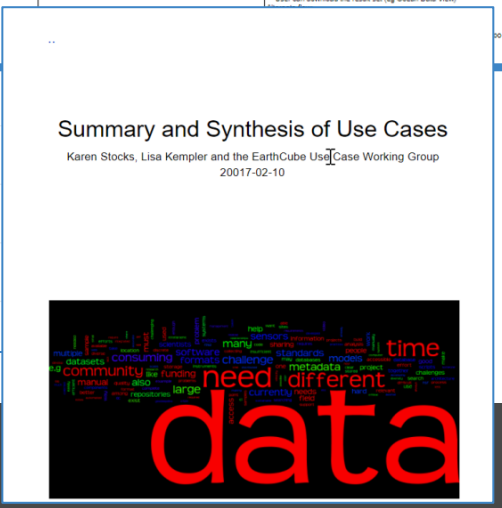
Key Outcomes of the Project

- Completed 50 use cases
- Summarized key CI challenges and technical details for all use cases
- Report, paper (in process)
- Generated engagement from EarthCube



	Internal ID (from List of Completed)	Use Case filename w/url	Summarized by	Summary Date	Title	Domain Keywords (semicolon separated)	EC Science Themes (List: Sources of Variability; Hazards; Predictions; State Parameters; Long-term trends; unclear)	Interdisciplinarity (Low, Med, High)	Overall Goals	Main CI Challenges (list separated by line breaks)	Data Discovery Challenges (for DDH)
2	1	Kathleen Ritterbush Use Case summary	KS	8/15/2016	Information tracking for a large collaborative deep-time project	Paleontology; geochemistry; ecological modeling	Sources of Variability; Long-term trends; implications for others	High	To create a community data portal that integrates paleontological and geochemical data about samples (rocks, photos, microscope slides) for use by those sciences plus ecological modelers. Could support important hypothesis testing.	- no system holds all the information about a sample and what has happened to it, plus the underlying stratigraphic framework. Very time consuming to try to capture information manually, data are all split up in "weird" repositories. - need to accommodate age models with uncertainty/confidence level in the dating of samples - GIS works for the spatial component, but not the temporal component	- wants to be able to find all the information about a sample, whether paleontological or geochemical.
3	2	https://docs.google.com/document/d	LK	8/24/2016	Rapid Real-Time Atmospheric Hazards_Chandra	Atmospheric studies	Hazards, Long-term trends	Medium	To be able to observe and model atmospheric observations at real-time and multiple time scales, in order that stakeholders can respond to hazards more quickly and effectively.	- network speed vs. data collection velocity - not fast enough to capture all weather phenomenon - Not enough sensors in critical locations - Younger analysts, scientists would like to see data stored in the cloud, not just public data centers	(These are about data collection but could also apply to saved data) - Geometric mapping of different data types and all of the different inputs coming at once is really hard to do - Figuring out time scales to take measurements at and then matching them up - for each new variable in each location, timeframe - Weather effects can knock out ability to collect data, causing gaps in data - Synchronizing data
4	3	Seismology_CHORDS_science_use_case_Vernon	KS	8/15/2016	Streaming real-time data for seismic early warning and monitoring and environmental monitoring	Atmospheric processes; seismology; natural hazards	Sources of variability; Hazards	High	Develop a real-time streaming network for seismology and atmospheric sensors and cameras (data and processed products), for use by researchers, policy makers, first responders.. Specifically, for Seismology and other areas: 1) Make it easier to bring in data in real-time. This enhances usefulness by getting the scientific analysis out of it faster and earlier, by making the best quality measurements 2) Generating more, better data products that more accurately reflect the actual weather phenomenon. 3) Capture enough data to get big picture of environment with high-quality timing (correlated timing between sensors)	- Funding is insufficient to cover the number of sensors (don't have meteorological sensors at all monitoring sites) and the sophistication of algorithms he would like. - Need to figure out a way to build a sustainable network that's well engineered and migrated into a sustainable infrastructure.A long term maintenance model is needed.	Copied from Use Case (These are about data collection but could also apply to saved data) - Geometric mapping of different data types and all of the different inputs coming at once is really hard to do - Figuring out time scales to take measurements at and then matching them up - for each new variable in each location, timeframe - Weather effects can knock out ability to collect data, causing gaps in data - Synchronizing data
5	4	Watershed_model_Chapela_Lara_Full_V2	KS	8/15/2016	A dynamical watershed model for concentration-age-discharge in a highly weathered tropical site (Luquillo CZO).	Computational hydrology (geophysics)	Predictions; sources of variability	High	Improve knowledge of the relationship between chemical weathering, nutrient cycling and hydrology by developing model repositories and online access to data to support the hydrological parameters	1. Not all data are discoverable and accessible - LTER Catalog helps, but is not complete, so individual scientists need to be asked for their data. 2. The data require QC and manipulation that is time consuming. Some of the QC could be automated if community scripts were available. Variations in parameter names and units also made time consuming, and metadata was not always complete 3. There were gaps in the data, making it hard to find sites and timespans with complete data.	I also spent a lot of time finding the right parameters, as different terms are used exchangeably or change through historical records, or units that are common ground for hydrologists/climatologist, but were unknown to me. While usually I could solve that by looking at the explanation in the LCZO metadata, the LTER one is less clear, and sometimes I had to contact the data manager directly or look up several books. This could be easily solved by having a hydro-Glossary or similar, as well as common units and abbreviations. Most of the problems I encountered can be grouped as: a) Missing values b) Values at sampling intervals longer than required by the model c) Values at sampling intervals shorter than required by the model d) Different units Part of his search is that she wants to find sites and timespans with complete data. Not just do data exist in that place/time, but is it complete.
6	5	Protein_Portal_Full_Mak_Saito	KS	8/1/2016	Access of Oceanic Protein Datasets	proteomics; ocean; microbial ecosystems; biogeochemistry	unclear	Low	Create a community data portal that allows research scientists to discover where, when and in which organisms a protein/enzyme of interest occurs in the oceans through a bioinformatics analysis of large mass spectral libraries created from many oceanic samples.	- mass spectrometer data distributed among multiple repositories & different mass spec platforms will pose integration problems with different datatypes. - How to ingest up-to-date genomic information, because protein data relies upon genome data for production. - volume: ~raw 10gb/day per instrument (~2-4 Tb/year/instrument), 1-2 instruments per contributing laboratory	Assuming proposed system is developed: - User comes with name of a protein or sequence to the portal - User enters parameters (ocean basin, year, etc) - User hits Go (submits form) - System matches name of the protein against a database of proteins System reports the protein found in the database along with information about the protein like where they were found, the number of spectra counts, geospatial & temporal reference frames (in table or visual [ODV form]), and the exact sequences - System reports displays results in a Map and a List (spreadsheet output) with quality indicators of the data and related metadata. - User evaluates results and can rerun the search with updated parameters - User can download the result set (eg Ocean Data View)





Results – CI Challenges

Word cloud created from the summary of cyberinfrastructure challenges extracted from each use case. The size of each word is directly proportional to the number of uses of that word. Numbers and common words are not included.

Data Access / Heterogeneity Challenges

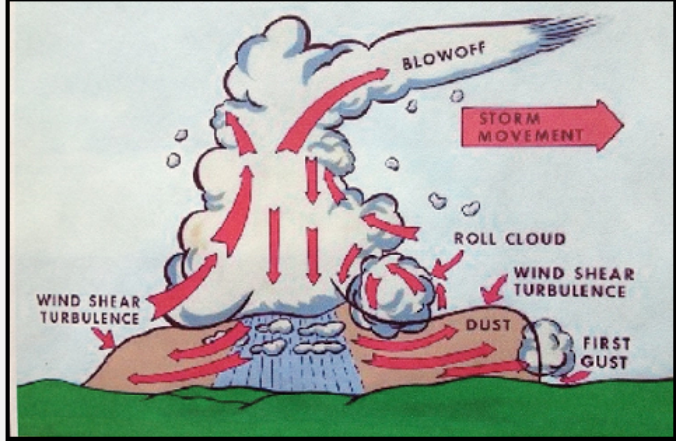
75%	Data Access/availability	
	28%	Data not online
	18%	Data in multiple online sources
	12%	Hard to search for desired data in online source
	12%	Important relationships between data in multiple sources missing
	8%	Hard to find/access data in publications
	8%	Sharing data is difficult/lacks incentives

32%	Data variety, diversity, and heterogeneity issues	
	24%	Data format diversity
	8%	Semantic variability
	12%	Integrating different data types (discrete vs continuous, sensor vs 4D model, etc.)

Other Data Challenges

18%	Total data volume
16%	Needed data does not exist (e.g. not enough sensors, or gaps in the data)
14%	Insufficient or uncertain data quality
14%	Insufficient metadata

Sample User Scenario



wind graphic

- **Specialized research goal**
  - Planetary wind turbulence and impact on % sand flux
- **Data not available yet or not calibrated right**
  - Multiple data sources, structures
  - Different hardware instruments
  - No community data storage site
- **Lots of time required to sync up data**
  - Requested guidelines about where and how
- **Wrote code for data acq, processing, analysis**
  - Saved MATLAB code on Github (17 contribs in 2017)
  - Data included

Question: What should EarthCube build for him?

Counted Category Types

- Formats
- Standards
- Software

“Standards”: Users Didn’t Connect with the Question

- **25 total standards mentioned**
  - 5 - OGC: A consortium, not 1 standard
  - 2 - DOI, EML, GPS, iGSN, netCDF
  - 1 - all others
- **26% - none**

Software Usage: Diversity + Long-Tail of Tools

- **155 mentions of 92 distinct S/W tools**
  - ❏ **33% use MATLAB**
  - ❏ **20% use "In-house code" ⇒ scripts, apps created by the scientist's team**
  - ❏ **20% use Excel**
  - ❏ **average = 2 tools/user**

Software Packages
(3+ mentions shown)

Software	# of Use Cases
MATLAB	17
In-house code	10
Excel	9
ArcGIS	7
R	5
Adobe Illustrator	4
Python	3
Google Earth Engine	3
IRIS/DMC tools	3

Summary:
Four Steps to End-User Partnering

Interested end-user communities self-select

- Corollary: If you can't find people, you haven't found a market
  (The really outspoken visionaries are not your market)

Create the opportunity – Formalize the process

- Really listening is hard
- They'll engage. People love to talk about their work

Figure out what they need, not what they asked for

- Latent need? Technology disruption? (ex: Cheap Gas vs. EV)
- Using the results to justify your direction is backwards
  - Usability (or UX) testing is different, later in the process

Responding takes creativity

- Functional requirements, design are steps beyond data gathering

Thank you!

- **Resources**
  - **Use Case Template: https://goo.gl/o5TqOB**
  - **Use Case Summary Matrix: https://goo.gl/z084X4**
  - **Summary and Synthesis of Use Cases: https://goo.gl/ERhCIS**
  - **Folder of completed use cases: https://goo.gl/56ij3u**
  - **Use Case Working Group Home Page: https://earthcube.org/group/use-cases-wg**
- **EarthCube Use Case Working Group Contacts**
  - Lisa Kempler, **lisak@mathworks.com**, Co-chair
  - Karen Stocks, **kstocks@ucsd.edu**, Chair

Data Formats: Lots of Variation, Popular = Common

- 25 mentioned
  - rest are singletons
- Often > 1 mentioned per use case

Format	# of Use Cases
CSV	14
netCDF	10
MATLAB .mat	6
Excel	5
.txt	4
ArcGIS/ESRI shapefiles	4
jpeg	3
tiff	3
.tsv	2
SEED	2

Conclusions: Data Challenges and Lone Researcher Phenomenon

- **Data access is the primary concern**
  - Data access is 1st workflow step but infuses the rest
  - Do they know of existing resources? Does it matter?
- **Some convergence on formats and software, but long-tailed is the most obvious feature**
  - Standards are diverse or not standards

⇒ Any system supporting these scientists needs to accommodate diversity