An open-source information model and associated cyberinfrastructure for effective environmental management and analytics

Roelof Versteeg[1], Rebecca Rubinstein[1], Reza Soltanian[2] and Doug Johnson[1]

1: Subsurface Insights 2: University of Cincinnati

2023 Global Summit
on Environmental Remediation
@REMPLEX

Acknowledgement

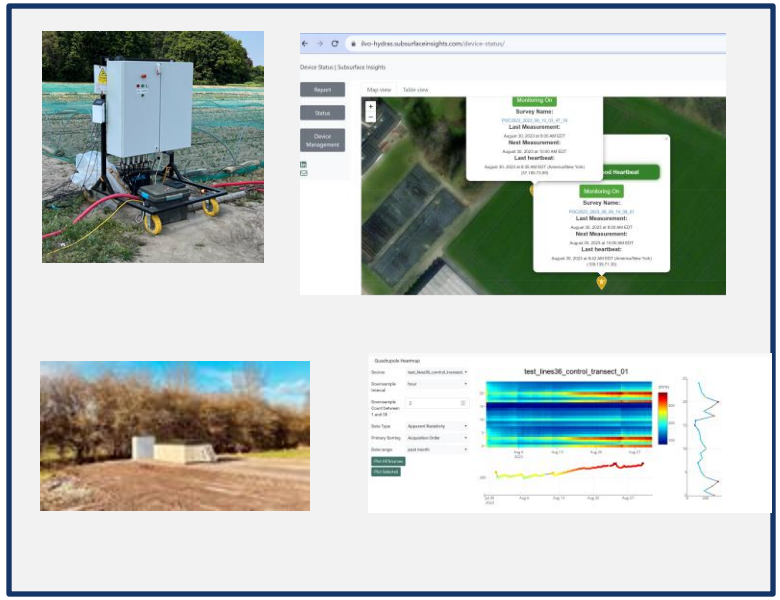
- The work presented here was funded by the DOE SBIR program under multiple SBIR awards (DE-SC-0009732, DE-SC-0018447 and DE-SC0019621)
- Examples shown include work performed by Subsurface Insights funded under the LBNL Watershed SFA and the SRNL ALTEMIS program, as well as from work done with collaborators from ILVO

Outline

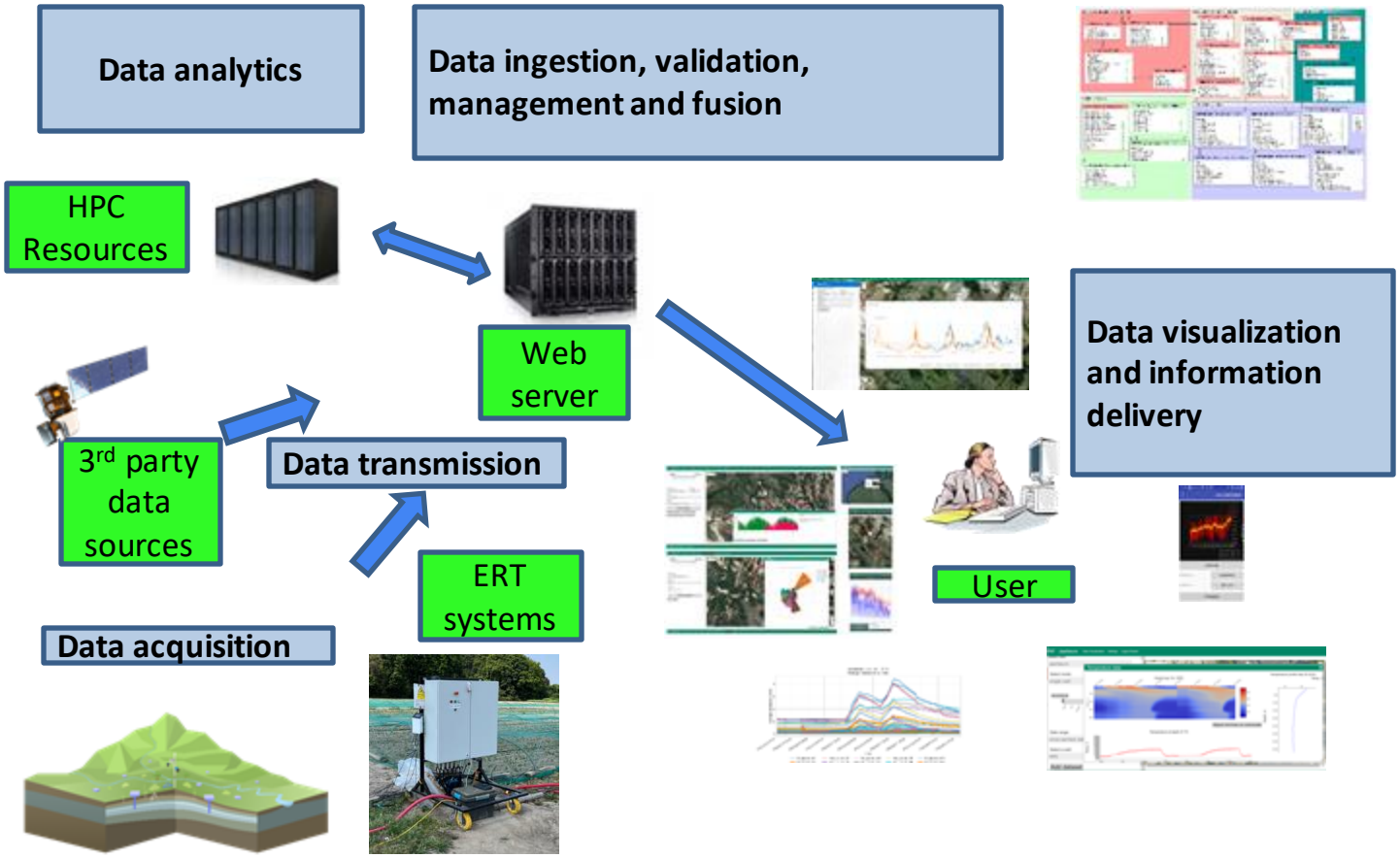
- Subsurface Insights introduction
- Setting the stage: why is data management and analytics to understand contaminated site behavior still challenging
- Solution attributes
- Solution implementation: ODMX
- Examples
- Summary

Subsurface Insights introduction

- Geoscience company specializing in near surface process monitoring



- Main products include
  - Autonomous electrical resistivity systems and associated cyberinfrastructure for continuous subsurface monitoring **[Contact me after my talk if interested]**
  - End to end cloud-based software framework for geoscience data ingestion, management, analytics and reporting **[THIS TALK]**



End to end cloud-based software framework for geoscience data ingestion, management, analytics and reporting: includes data acquisition, transmission, data management, analytics and reporting

Data analytics

Data ingestion, validation, management and fusion

This talk

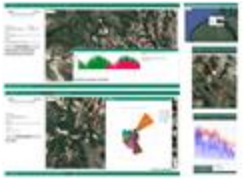
HPC Resources

Web server

Data visualization and information delivery

3rd party data sources

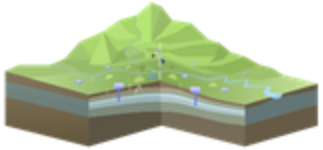
Data transmission

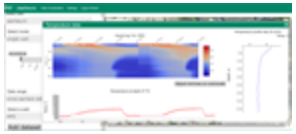


User

ERT systems

Data acquisition

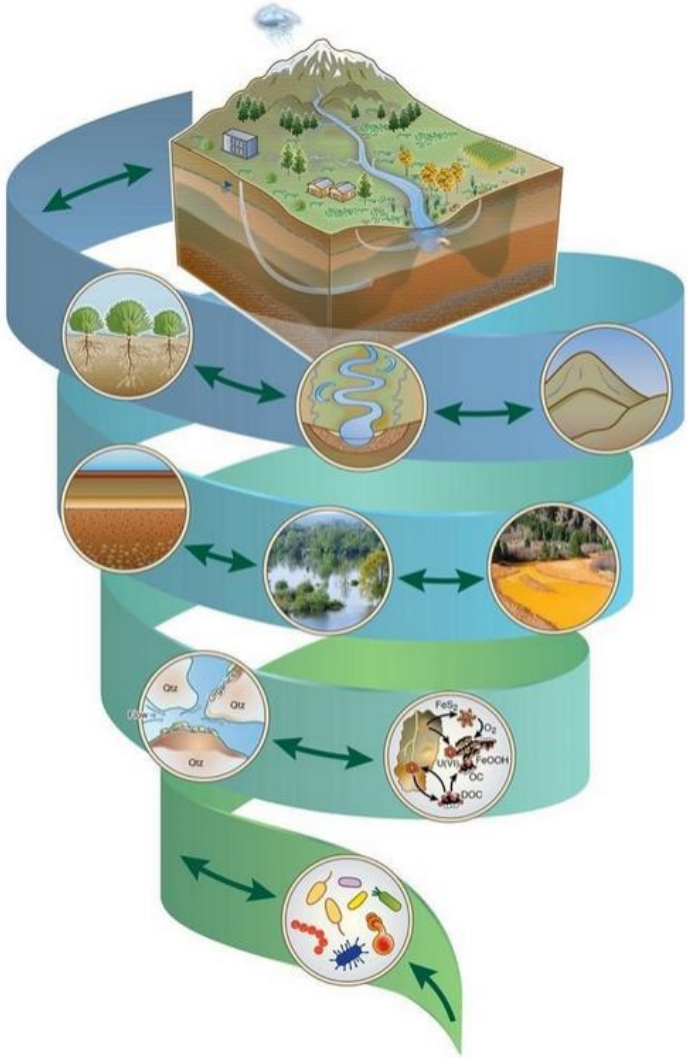


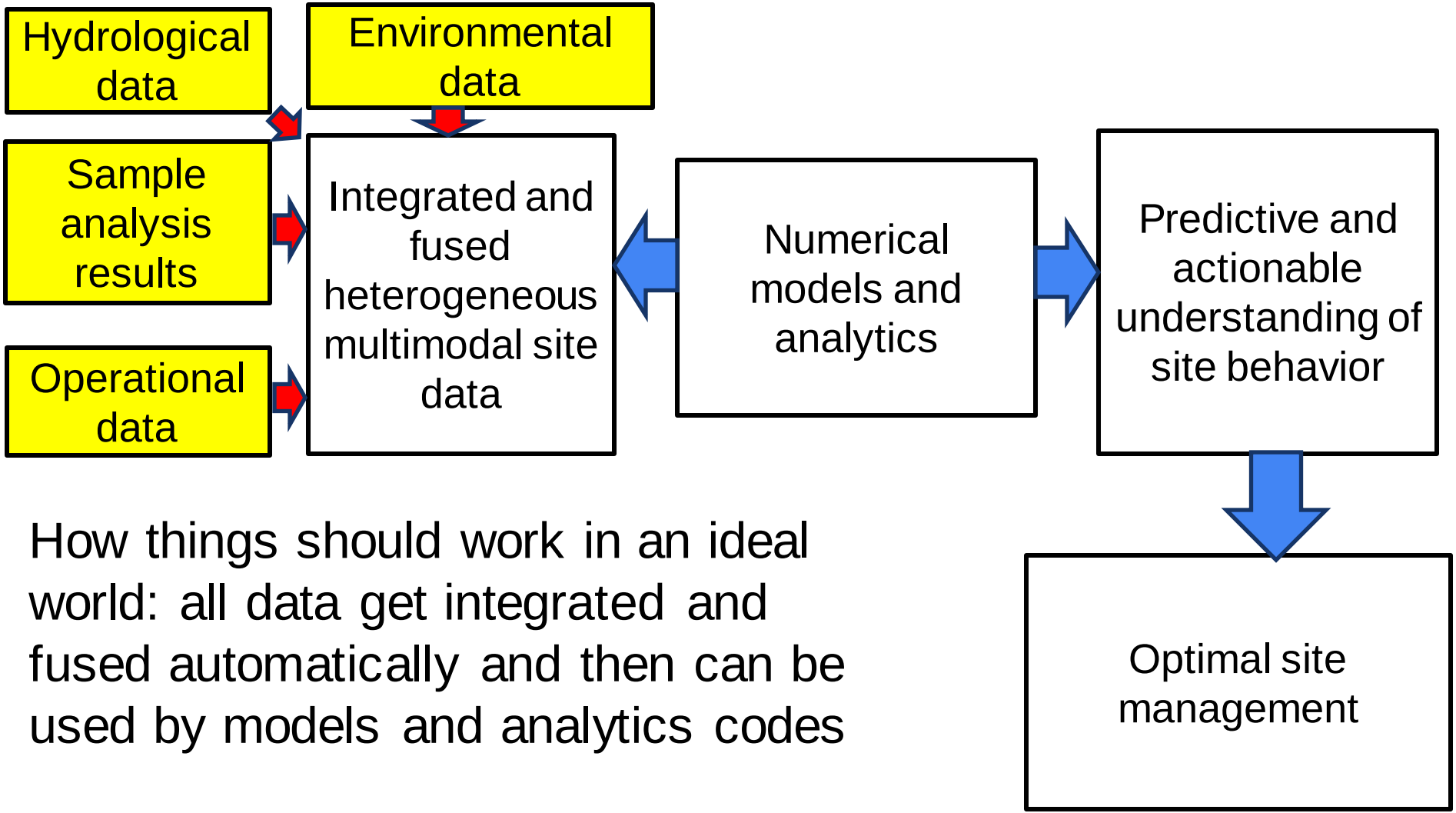


Outline

- Subsurface Insights introduction
- Setting the stage: why is data management and analytics to understand contaminated site behavior still challenging
- Solution attributes
- Solution implementation: ODMX
- Examples
- Summary

Truism: Contaminated site behavior is controlled by physical, chemical and biological natural and anthropogenic processes at multiple spatial and temporal scales





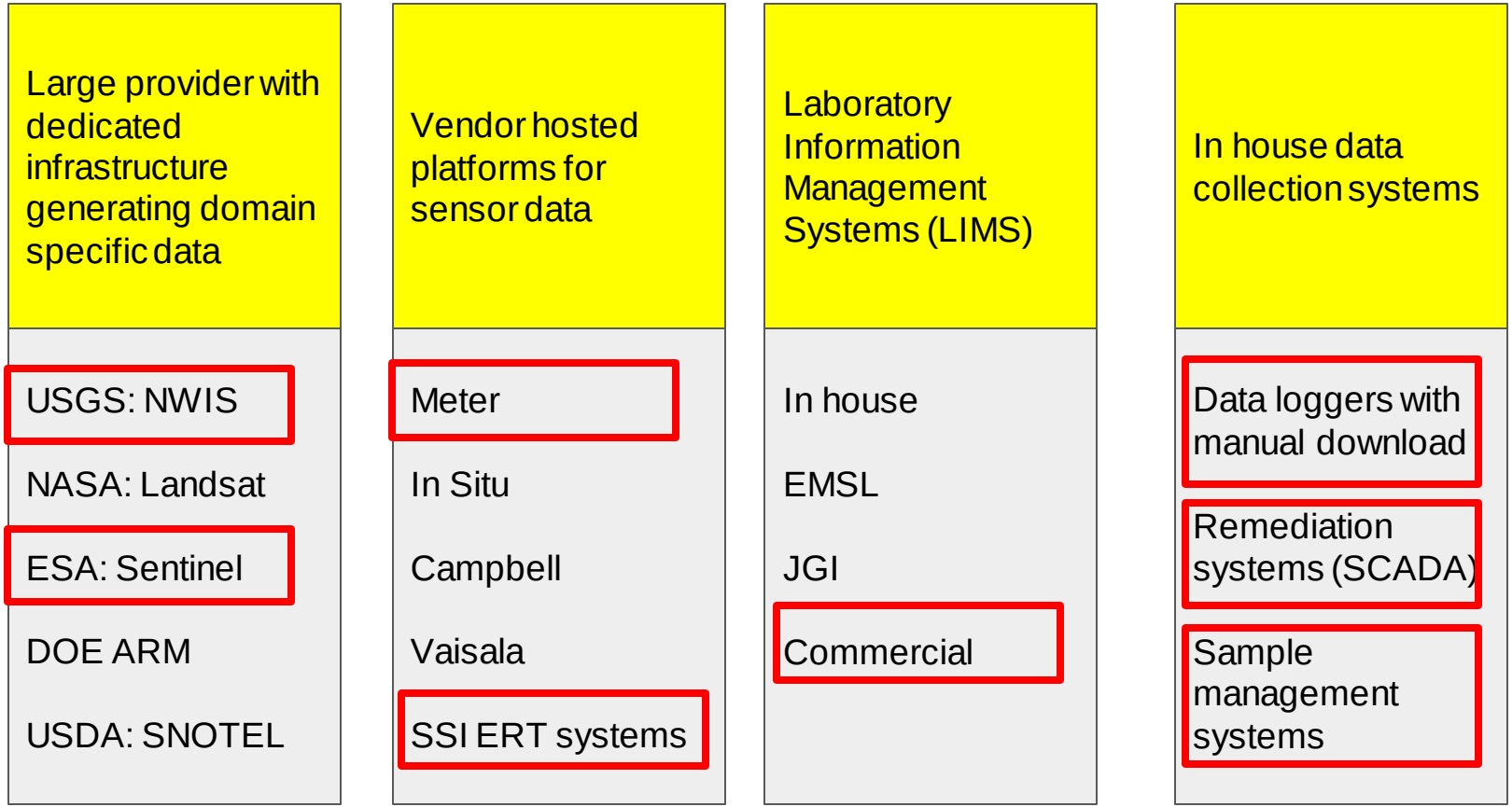
How things should work in an ideal world: all data get integrated and fused automatically and then can be used by models and analytics codes

Reality

- different units and parameter names for same parameters (airtemp_c, tempair_F)
- timestamps in different time zones and formats
- different data formats and data density
- different meaning and “value” of different datasets

Site data is heterogeneous in type, provenance and meaning (“how data relates to processes”)

- **Process Drivers**: precipitation, solar radiation, regional groundwater gradients, contaminant releases, remedial efforts
- **Process Context:** Elevation, Soil type, Geographic location
- **Process manifestation:** Vegetation, soil microbiology, soil moisture, contaminant concentrations



Example: Data heterogeneity in type, provenance and meaning

The data management challenge:

integrate the multiple inter-related datasets that provide complementary views of the same system so that we can use this data effectively for process understanding and site operation

Note: data management is an enabler

Three main requirements

1. Data needs to be normalized
   - Same parameter collected by two groups can have different names and/or different units
   - Data can be reported in different coordinate systems
   - Data can be reported in different time zones
   - Data organization and structure differs between different providers

   ➔ Normalize parameter names, units, coordinate systems and timestamps

2. Data needs to be fused

3. Data needs to be findable, accessible and usable by analytics codes

What is data fusion* and why do we need it

- **Data fusion**: an approach for combining and interconnecting multi modal data by providing relations between different data sets
- For some data existing relations are sufficient
- For some data we need to define additional relations

* In this context - there are other definitions of data fusion

Age

Salary

Blood pressure

Zip code

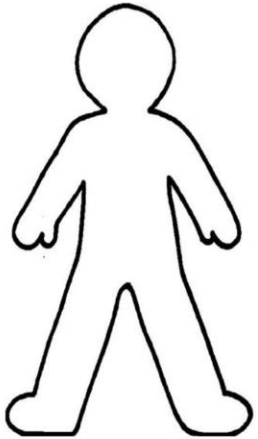
Cholesterol

Weight

Education

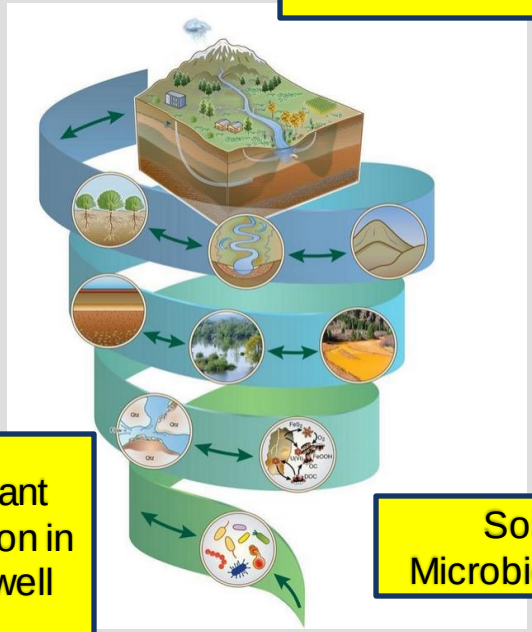
Height

Education



Implicit relation: all data connect to a person and time – may be sufficient

Precipitation



Contaminant concentration in sampling well

Soil Microbiology

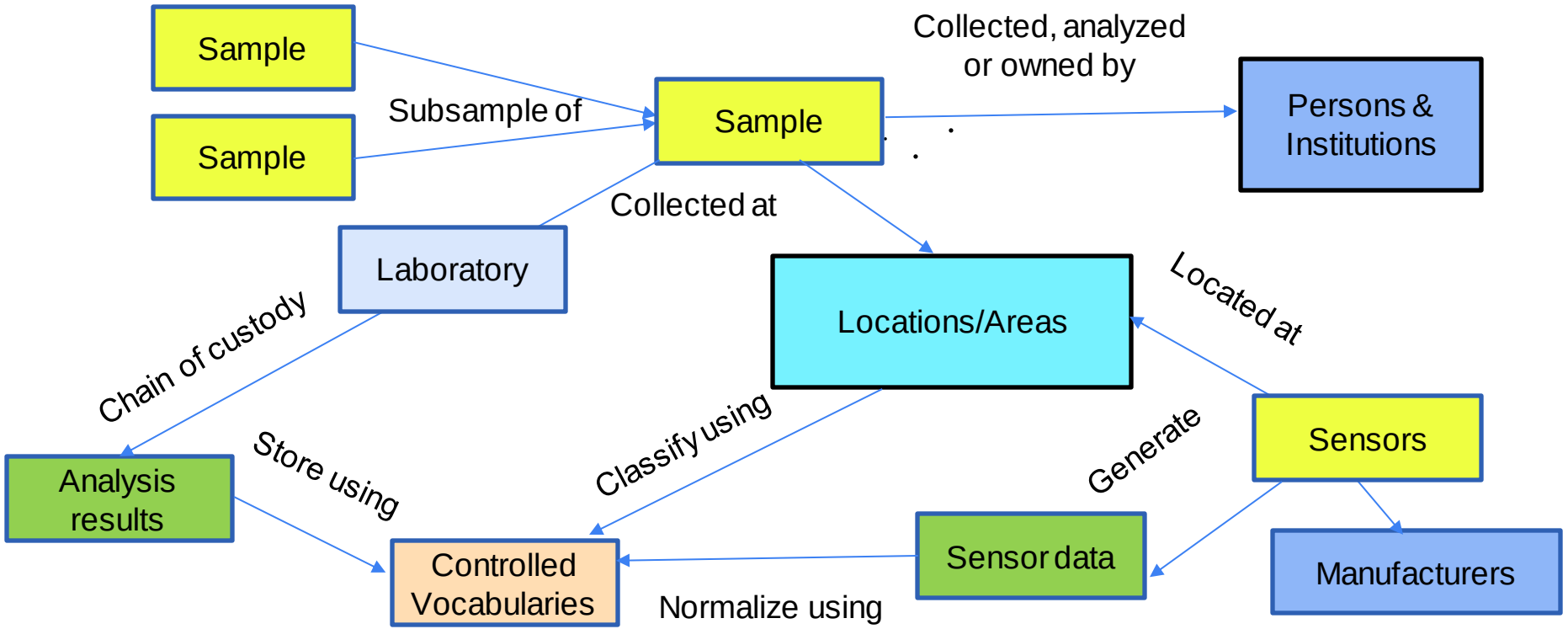
Implicit relation: time, but need to define additional relations

Relations can be defined at the data management and at the analytics level

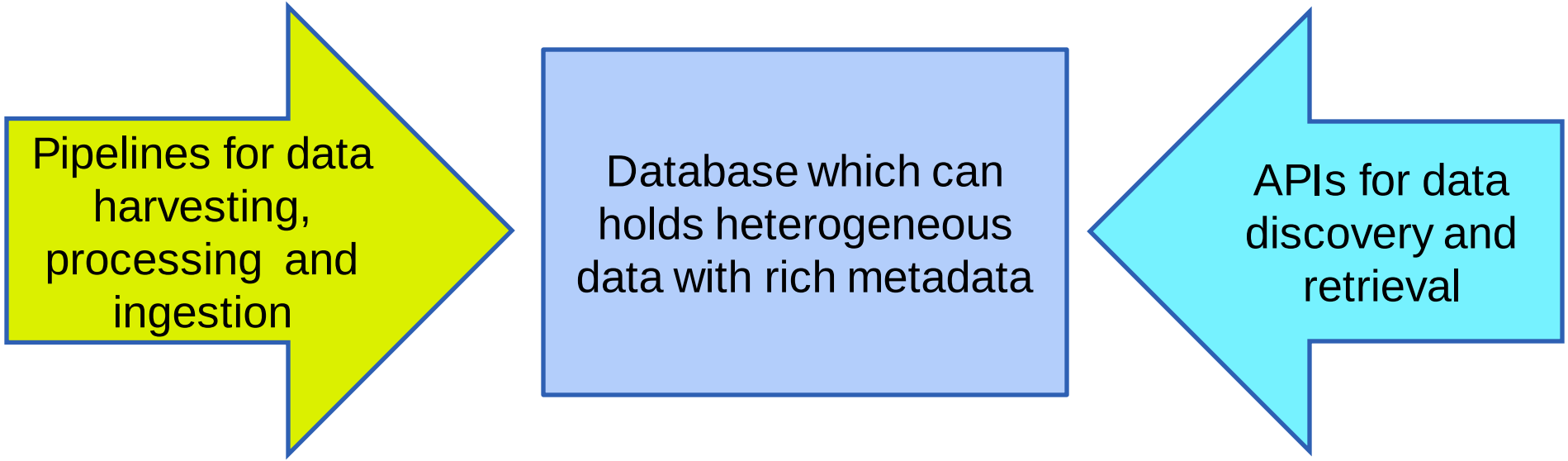
Outline

- Subsurface Insights introduction
- Setting the stage: why is data management and analytics to understand contaminated site behavior still challenging
- Solution attributes
- Solution implementation: ODMX
- Examples
- Summary

Attribute 1 of solution: represent complex relationships and attributes, support heterogeneous data types and ways to fuse data



Attribute 2 of solution: effective ways to ingest and extract data and allow for data discovery



Attribute 3 of solution: easy to deploy and use, flexible and extensible

Robust software tools allowing users to write their own parsers

Extensible (e.g. project specific Controlled Vocabularies)

Cloud deployable

Two part solution which has these attributes: ODMX

Open source

ODMX Information Model

ODMX Cyberinfrastructure

Based on powerful NSF funded ODM2 Information Model (Horsburgh, 2016)

Developed under SBIR funding. Deployed and used by e.g. LBNL SFA and SRNL ALTEMIS

ODMX Information Model

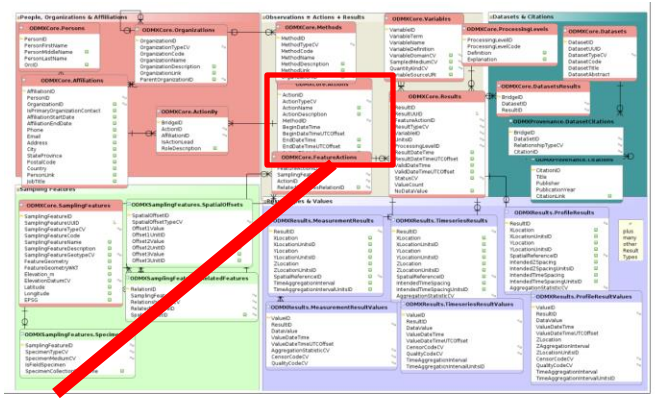


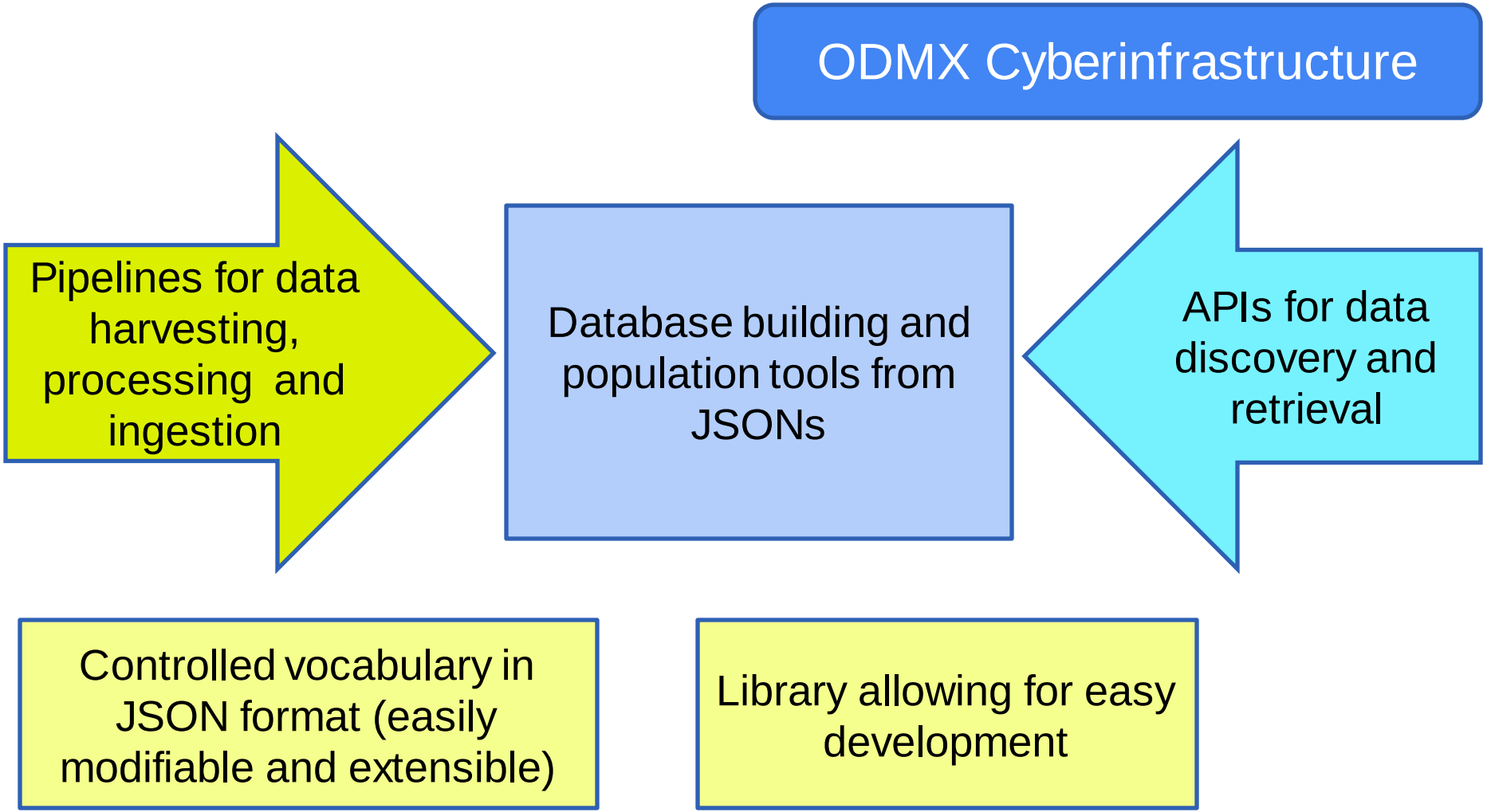
Table: Actions

Describes actions (e.g., observation, sample collection, sample analysis, field visits, field activities, etc.)

Columns

PK	Name	Data Type	NULL	Auto	Default	Comment	FK
🔑	ActionID	Integer		✔	nextval('"ODMXCore"."Actions_ActionID_seq"'::regclass)	Primariy key identifier	
	ActionTypeCV	VarChar()				CV term describing the type of action (e.g., observation, sample collection, sample analysis)	✔
	ActionName	VarChar()	✔			A name for an Action or activity.	
	ActionDescription	VarChar()	✔			Text that describes the action	
	MethodID	Integer				Foreign key identifier for the method used to complete the action	✔
	BeginDateTime	DateTime				The date/time at which the action began	
	BeginDateTimeUTCOffset	Integer				The UTCOffset for the BeginDateTime	
	EndDateTime	DateTime	✔			The date/time at which the action ended	
	EndDateTimeUTCOffset	Integer	✔			The UTCOffset of for the EndDateTime	
	ActionFileLink	VarChar()	✔			A URL or path to a file created by or used by the Action, such as instrument output or configuration.	

- 196 tables
- Can represent complex relationships, hold data, trace provenance
- Flexible extensions
- Includes model for microbiological data
- Substantially enhanced and improved from ODM2


ODMX Cyberinfrastructure
Pipelines for data harvesting, processing and ingestion
Database building and population tools from JSONs
APIs for data discovery and retrieval
Controlled vocabulary in JSON format (easily modifiable and extensible)
Library allowing for easy development

Key aspect of data model: sampling features

- Sampling Features defined in the Open Geospatial Consortium Observations, Measurement and Samples standard (Cox 2010).

- *“geospatial or physical entities on which or at which observations are made”*.

- All data in ODMX is associated with a Sampling Feature.

Sampling feature example

- Site
- Instrumented borehole
- Sample (specimen) and sub-specimens

Data is associated with sampling features, and sampling features are associated with each other

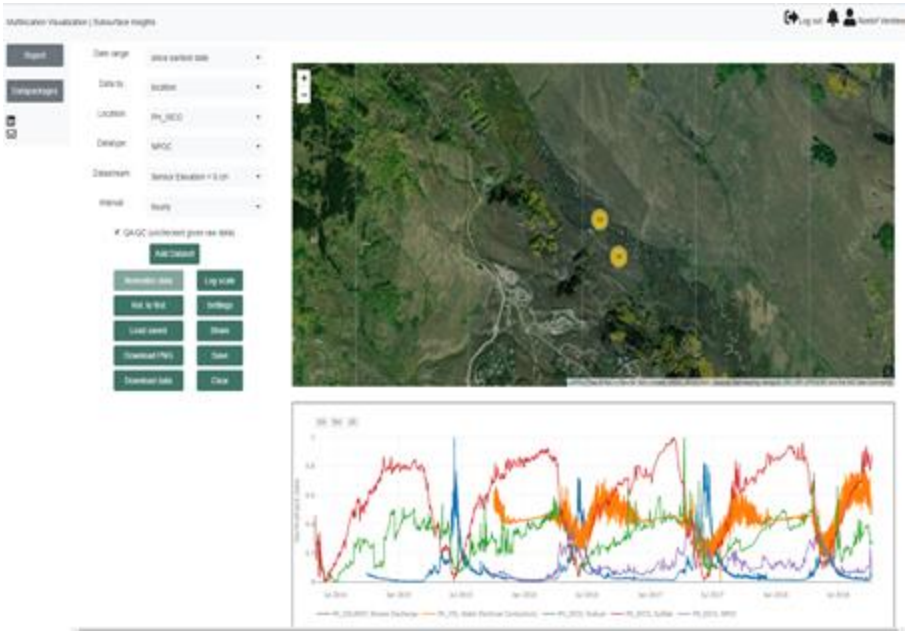
Sub-specimens **are a child of** Specimens, which **was collected at** an Instrumented Borehole, which **is contained in** a Site

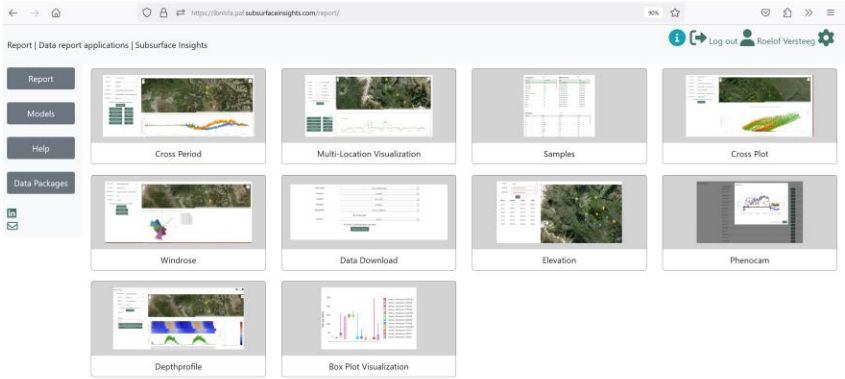
Sampling features attributes

- Sampling Features can be hierarchical
- Sampling feature relations allow us to add additional level of information which provides relations between data
- Relations expressed through extensible controlled vocabulary, e.g.
  - Is Contained in
  - Is Sub Specimen of
  - Is Child of
  - Is Derived from
  - Is Part of
  - Is Related to
  - Was collected At

- **Relations between sampling features provide a natural way to fuse data**

Examples
LBNL SFA Site





Report

Models

Help

Data Packages

Location	# of samples
ER-INY1	269
ER-MCP5	4
ER-MDP1	1
ER-MDP2	1
ER-MDP3	1
ER-MMT1	124
ER-PHF0	2184
ER-PLM1	252
ER-PLM11	18
ER-PLM12	17

Specimens search

Name	Date
ER-PHF0_2014_05_02_12_00_00_collected	May 2, 2014
ER-PHF0_2014_05_03_12_00_00_collected	May 3, 2014
ER-PHF0_2014_05_04_12_00_00_collected	May 4, 2014
ER-PHF0_2014_05_05_12_00_00_collected	May 5, 2014
ER-PHF0_2014_05_06_12_00_00_collected	May 6, 2014
ER-PHF0_2014_05_07_12_00_00_collected	May 7, 2014
ER-PHF0_2014_05_08_12_00_00_collected	May 8, 2014
ER-PHF0_2014_05_09_12_00_00_collected	May 9, 2014
ER-PHF0_2014_05_10_12_00_00_collected	May 10, 2014

Samples at ER-PHF0 on 5 4 2014 search

Name	Description
ER-PHF0_2014_05_04_12_00_00_icpms_subspecimen	Subspecimen of ER-PHF0_2014_05_04_12_00_00_collected
ER-PHF0_2014_05_04_12_00_00_anion_subspecimen	Subspecimen of ER-PHF0_2014_05_04_12_00_00_collected

Sample ER-PHF0_2014_05_04_12_00_00_icpms_subspecimen search

Analyte	Value	Std. Dev.	Units
Li	0.746999	0.05900559	ppb
B	3.389706	0.11921479	ppb
Na	1378.756579	35.58669468	ppb
Mg	7217.630019292	140.380532850635	ppb
Al	22.809252	0.52407244	ppb
Si	2087.004965	46.76930719	ppb
K	477.832012	1.43885408	ppb

https://lbnlsfa.paf.subsurfaceinsights.com/crossperiod/ 90%

Cross period | Subsurface Insights

Log out Roelof Versteeg

Report

Models

Help

Data Packages

Start: 09 / 26 / 2016

Data by: type

Datatype: Air Temperature

Location: ER-KLP1 (Kettle_Ponds)

Datastream: Sensor Elevation = 2 meter, Instr

Interval: daily

Add Dataset | Advanced settings

Log scale

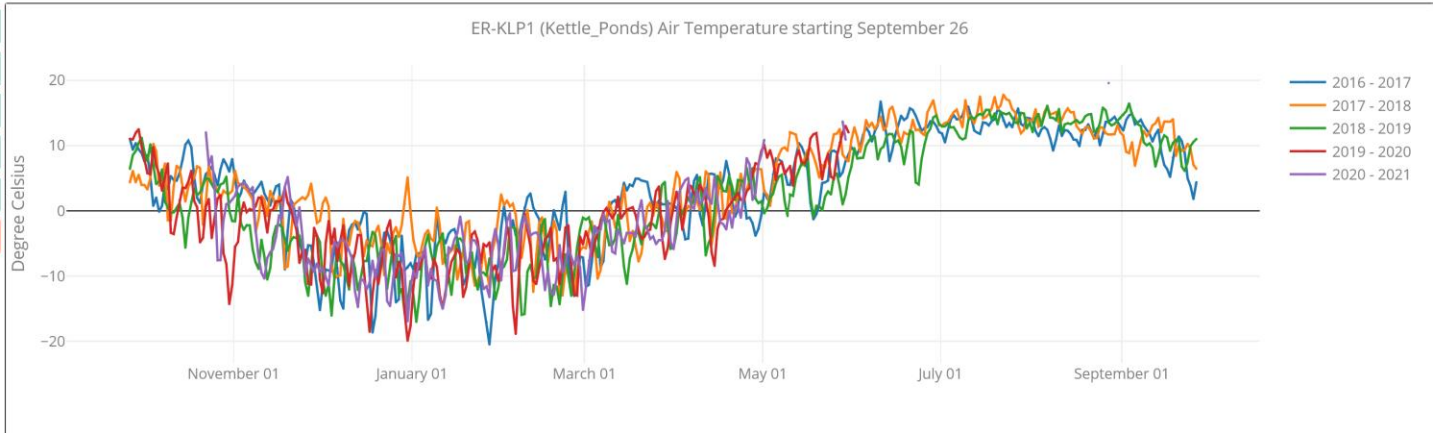
Manual scale

Download PNG

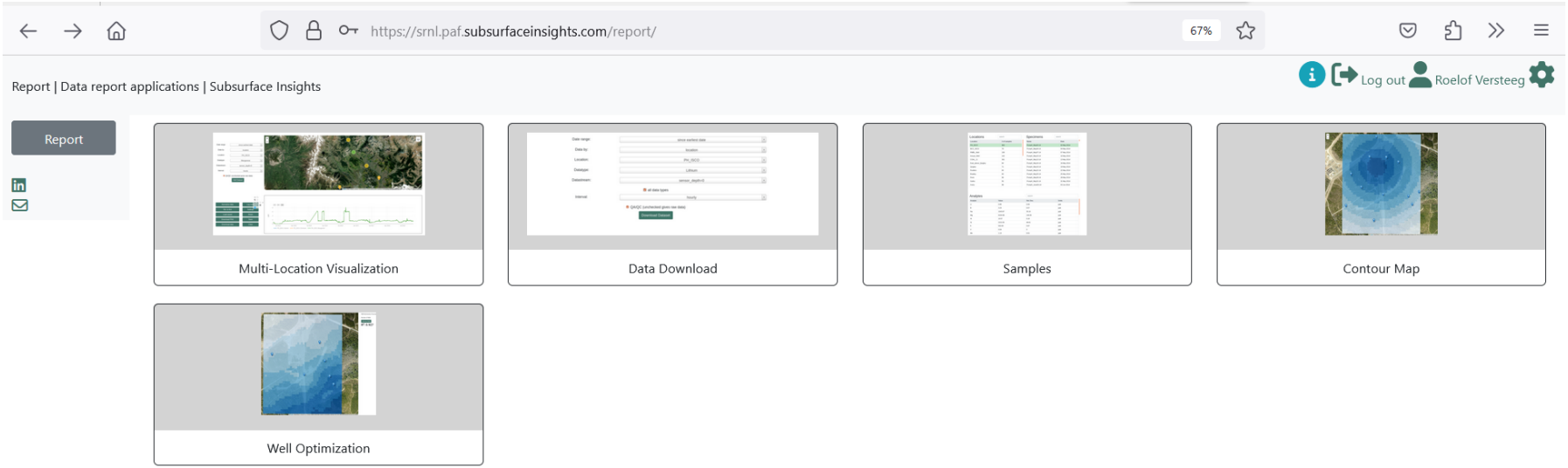
Download data

Toggle legend

Clear



ALTEMIS project (Savannah River)



https://srnl.paf.subsurfaceinsights.com/multilocation/?graph_uuid=623a87df-a9ad-b27a-c608-821e1c8f7569 80%

Multi-Location Visualization | Subsurface Insights

Log out Roelof Versteeg

Report

Filter By: Datatypes

Datatype: pH

Location: FSB126D

Datastream: Sensor Elevation = 0 meter, Instrument

Interval: daily

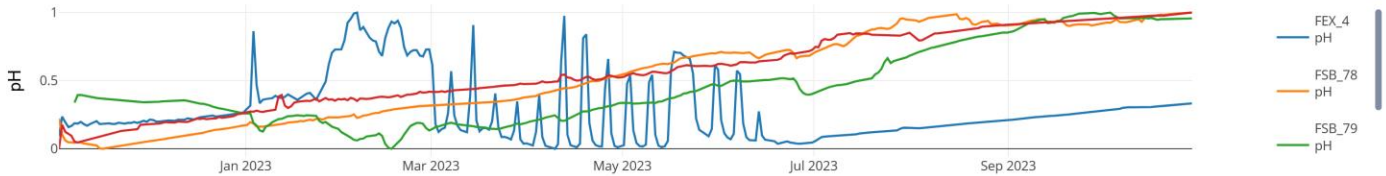
Date range: since earliest date

Add Dataset

Advanced Settings

Manage Graphs

Revert | Rel. to First | Linear Scale | Manual Scale | Download PNG | Download Data | Hide Legend | Clear



Integration with analytics codes

- Analytics codes can query ODMX through python API – allows for data and relation discovery and retrieval
- ODMX and associated databases are used by Subsurface Insights analytics codes
- Example is the use by the ALTEMIS project of ODMX for machine learning using PyLEnM (Meray et al, 2022)
- Same capability used by electrical resistivity processing codes

Summary

- Effective environmental datamanagement requires
  - Normalizing multi-modal data
  - Fusing multi-modal data
  - Ensuring that data can be easily used by analytics codes

- ODMX provides a robust and extensible solution which meets these requirements

https://github.com/subsurfaceinsights/odmx

https://odmx.org

Questions?

Roelof Versteeg
603 443 2202
Roelof.Versteeg@subsurfaceinsight.com
www.subsurfaceinsights.com

ODMX additional details:

- Github repo has example on how to build from scratch
- Includes automatically generated library which matches information model
  - All functions and attributes accessible through autocomplete
  - Includes type checking
  - Allows users to quickly write their own parsers