USC INFORMATION SCIENCES INSTITUTE

USC Viterbi School of Engineering

Open Knowledge Networks for Geosciences, Sustainability, and Convergent Research in Natural-Human Systems

Yolanda Gil

Information Sciences Institute
and Department of Computer Science
University of Southern California
http://www.isi.edu/~gil
@yolandagil
gil@isi.edu

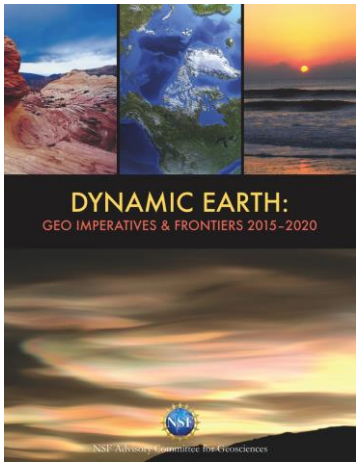
4 October 2017

Outline

1. The geosciences landscape
2. Ontologies, vocabularies, standards
3. Infrastructure
   - Data centers
   - Tools
4. Modeling
   - Model repositories
5. Model integration and convergence research
   - The need for semantics

Geosciences

- The Earth as a **system**
  - Earth (surface and subsurface)
  - Ocean
  - Atmospheric
  - Polar
  - Geospace



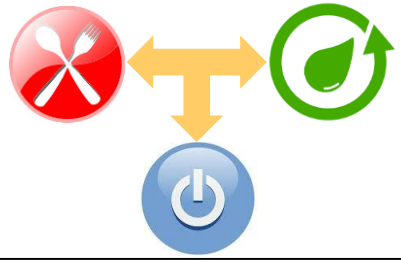
Land/Ocean Processes

Ocean-Atmosphere-Ice Ecosystems

Geo-Bio-Chem-Phys-Human Processes in Ecosystems

Urban Geosystem Science

- Natural processes/resources interact with human activities
  - Water
  - Food: Agriculture, fisheries
  - Energy: Manufacturing, infrastructure
- Ecosystems and sustainability



Introduced in Senate (05/06/2015)

Nexus of Energy and Water for Sustainability Act of 2015 or the NEWS Act of 2015

Geosciences Landscape

- Funding agencies

- Foundations
  - Sloan

- Community organizations
  - Earth Science Information Partners (ESIP)

- Scientific organizations and publishers
  - AGU, GSA, ASLO, CEDAR

- Standards organizations
  - Open Geospatial Consortium (OGC)

- Industry
  - Esri (ArcGIS)
  - Oil and gas
  - Mapping (Google, Microsoft)

- *Many groups outside US*

Outline

1. The geosciences landscape
2. Ontologies, vocabularies, standards
3. Infrastructure
   - Data centers
   - Tools
4. Modeling
   - Model repositories
5. Model integration and convergence research
   - The need for semantics

Ontologies, Vocabularies, and Standards

- General geosciences vocabularies
  - SWEET (Semantic Web for Earth and Environmental Terminology)
  - EML (Ecological Metadata Language)
  - ENVO (Environmental Ontology) in BioPortal & OBO
- Space and time
  - W3C Space and Time Ontology (builds on GeoSPARQL, KML,…)
  - Open Geospatial Consortium standards (eg, SensorML)
  - ISO 19115 (geospatial data)
- Maps:
  - Gazetteers (e.g., Geonames), USGS Geographic Names Information System (GNIS), NGA GEOnet Names Server, etc.
- Specialized ontologies:
  - WaterML, CF (Climate and Forecast) conventions, land cover,…

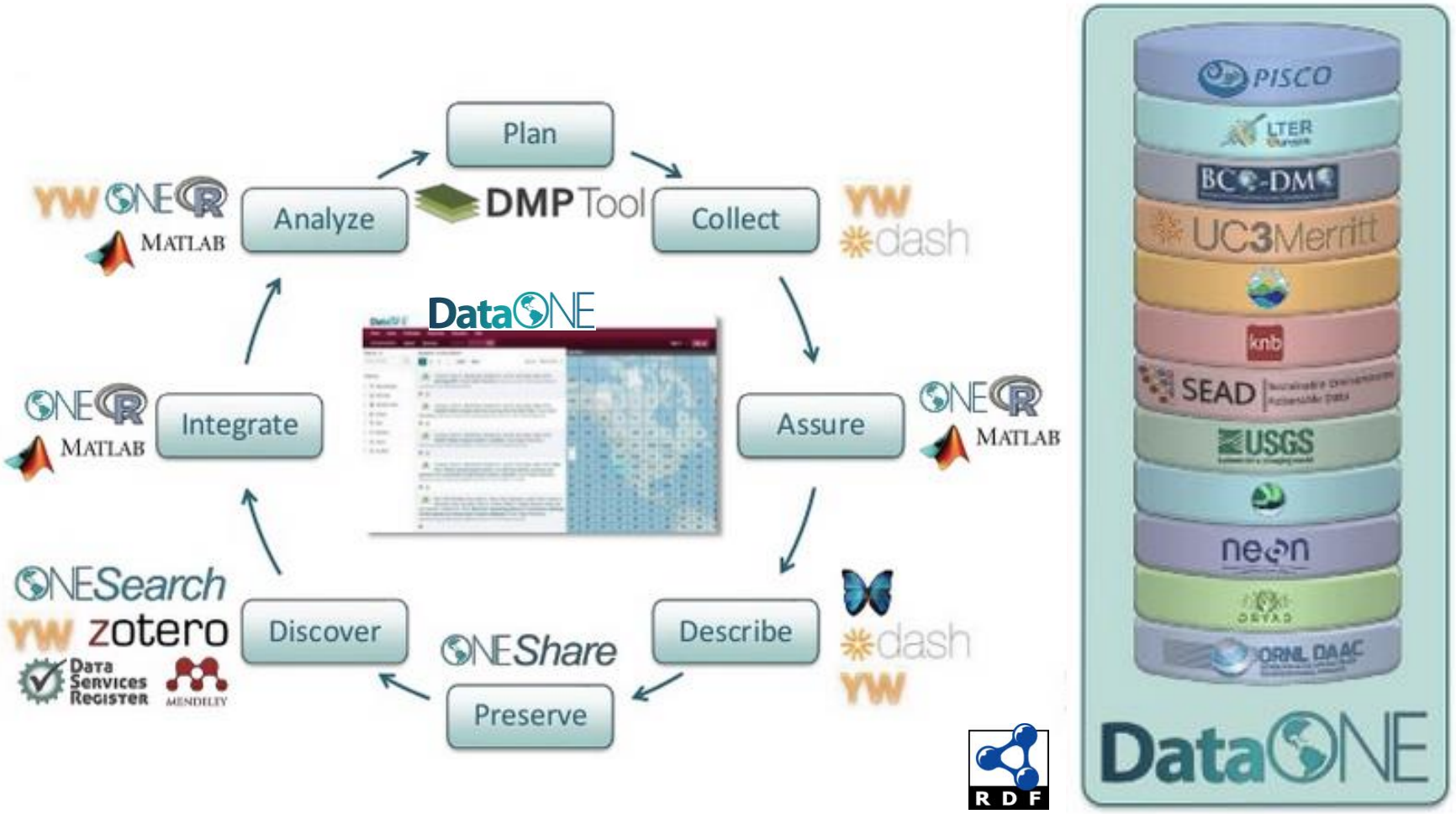
Outline

1. The geosciences landscape
2. Ontologies, vocabularies, standards
3. Infrastructure
   - Data centers
   - Tools
4. Modeling
   - Model repositories
5. Model integration and convergence research
   - The need for semantics

Infrastructure: Data Centers

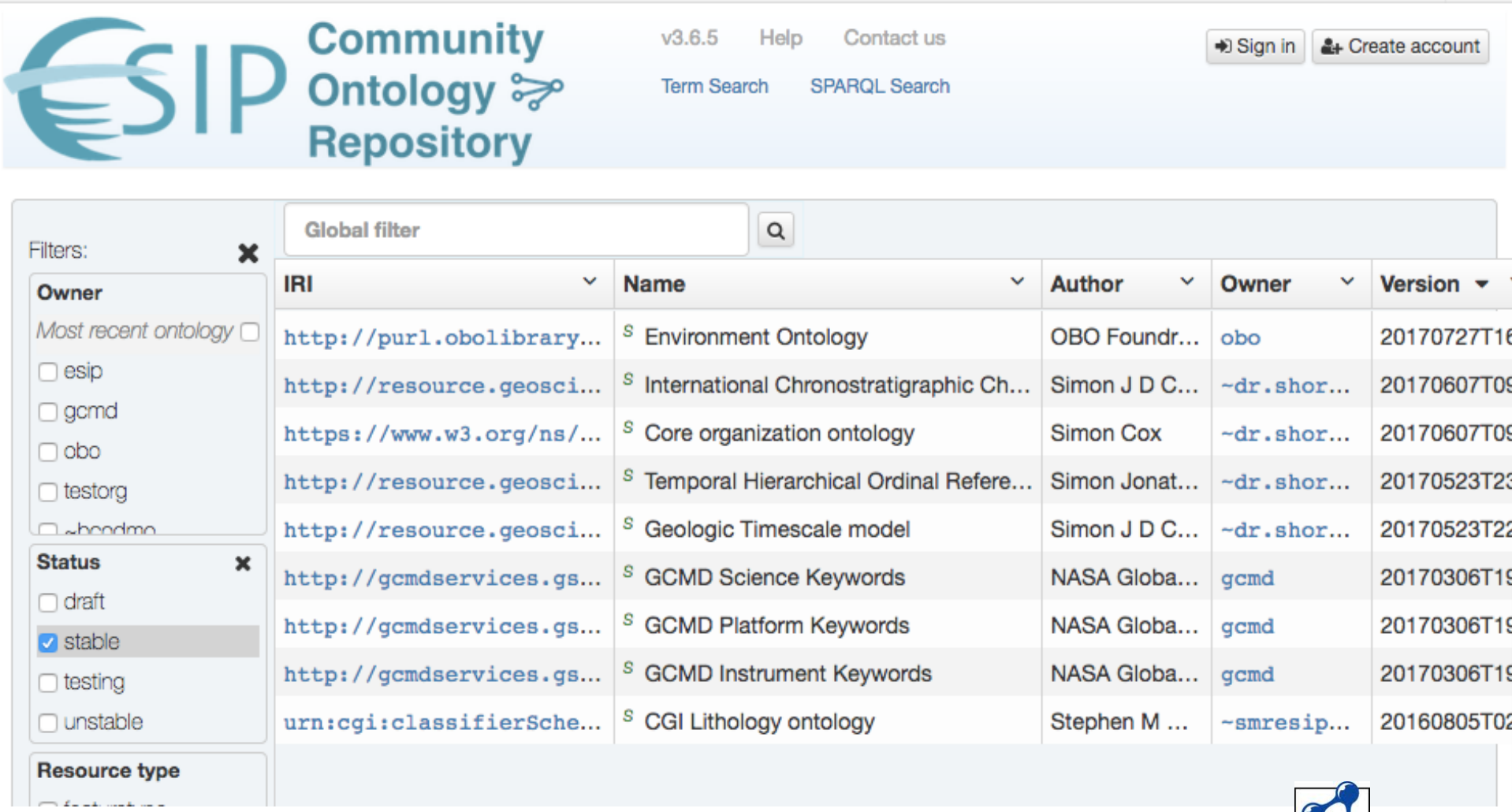
- Federal and state level
  - NASA DAACs: Remote sensing data
  - USGS and state geological surveys
- General repositories
  - Pangea
  - IGSN
- Specialized data centers with semantic APIs:
  - CUAHSI (hydrology)
  - IEDA (geology)
  - IRIS (seismology)
  - NSIDC (polar)
  - BCO-DMO (ocean)
  - Madrigal (geospace)

Infrastructure: NSF's DataONE



B. Michener, 2017, from https://www.slideshare.net/aspecht/michener-workshop-montpellier

Infrastructure:
ESIP Semantics and Ontology Working Group



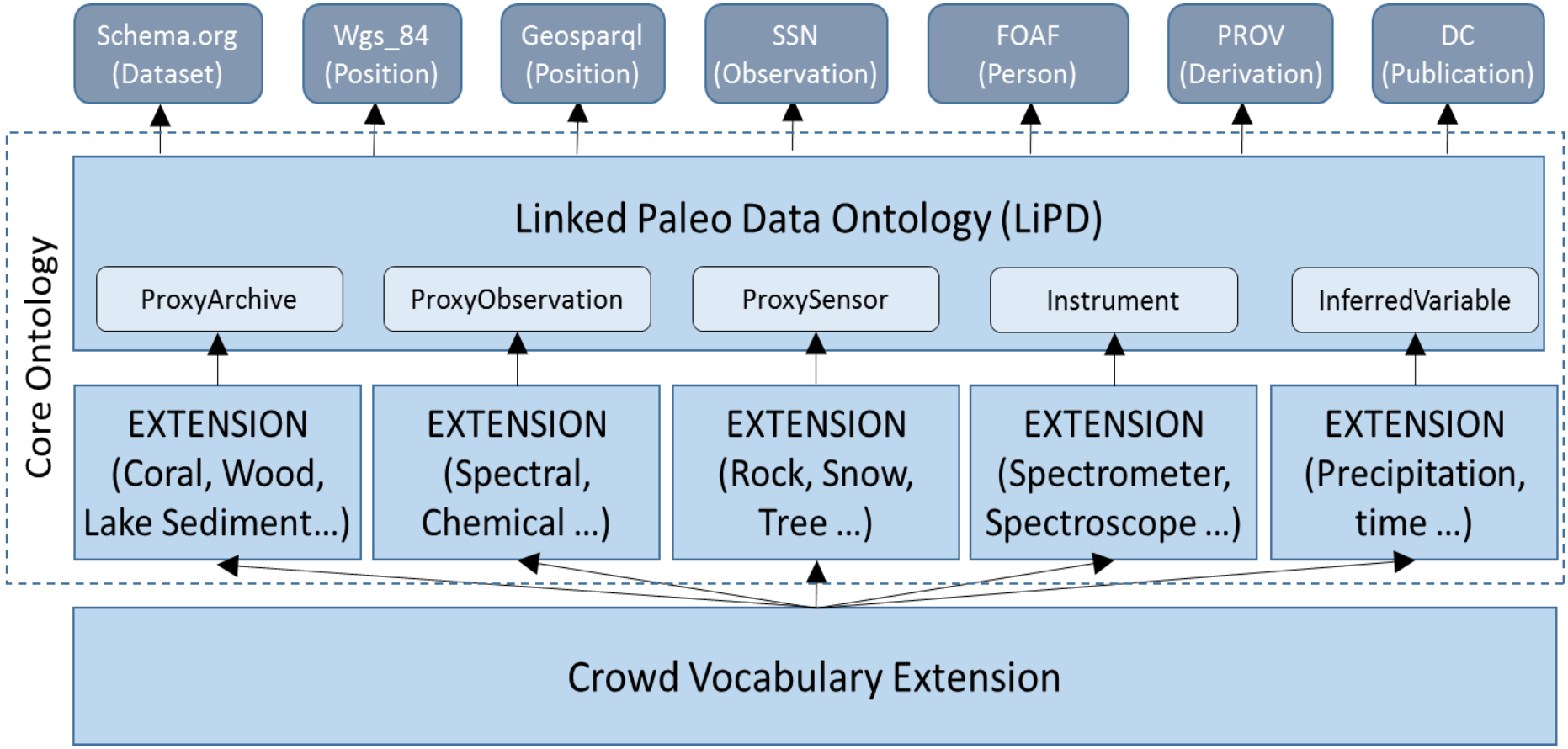
http://cor.esipfed.org/

Infrastructure: The NSF EarthCube Initiative

- Many projects use ontologies:
  - CINERGI, OntoSoft, Linked Earth, Earth System Bridge, EarthCollab, GeoDeepDive, GeoSemantics, X-DOMES, …
  - See roster at: https://www.earthcube.org/info/about/funded-projects
- Infrastructure and tools:
  - Text extraction, resource inventory, ontology inventory, data integration, model integration, mediators, semantic services, metadata crowdsourcing, …
  - See https://www.earthcube.org/tools-inventory
  - Ongoing development of integrated architecture
- Council of Data Facilities
  - Includes major data centers in geosciences
  - See https://www.earthcube.org/group/council-data-facilities

EarthCube's Linked Earth Project: Controlled Crowdsourcing for Paleoclimate Metadata

Work with D. Garijo, J. Emile-Geay, D. Khider, V. Ratnakar (USC); N. McKey (NAS)



http://wiki.linked.earth/

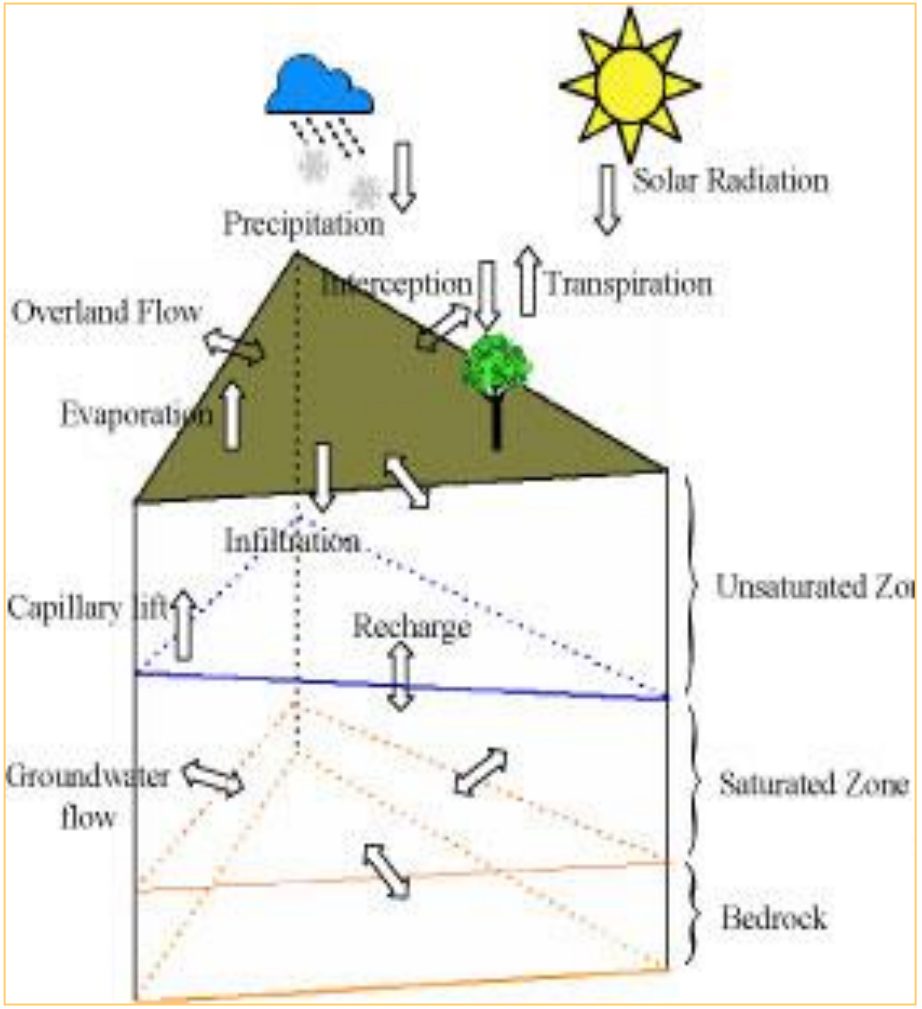
Outline

1. The geosciences landscape
2. Ontologies, vocabularies, standards
3. Infrastructure
   - Data centers
   - Tools
4. Modeling
   - Model repositories
5. Model integration and convergence research
   - The need for semantics

Modeling in Geosciences: Models of Dynamical Systems

- Historical observational data for calibration
- Forecast data for prediction
- Observational data for evaluation

Process	Governing equation/model	Original governing equations	Semi-discrete form ODEs
Channel Routing	St. Venant Equation	$\frac{\partial h}{\partial t}+\frac{\partial(uh)}{\partial x}=q$	$\left(\frac{dz_c}{dt}=P_c-\sum Q_{gc}+\sum Q_{oc}+Q_{in}-Q_{out}-E_c\right)_i$
Overland Flow	St. Venant Equation	$\frac{\partial h}{\partial t}+\frac{\partial(uh)}{\partial x}+\frac{\partial(vh)}{\partial y}=q$	$\left(\frac{\partial h}{\partial t}=P_o-I-E_o-Q_{oc}+\sum_{j=1}^{3}Q_o^{\,0}\right)_i$
Unsaturated Flow	Richard Equation	$C(\psi)\frac{\partial\psi}{\partial t}=\nabla\cdot(K(\psi)\nabla(\psi+Z))$	$\left(\frac{d\xi}{dt}=I-q^{k}-ET_s\right)_i$
Groundwater Flow	Richard Equation	$C(\psi)\frac{\partial\psi}{\partial t}=\nabla\cdot(K(\psi)\nabla(\psi+Z))$	$\left(\frac{d\zeta_g}{dt}=q^{3}+\sum_{j=1}^{3}Q_g^{\,0}-Q_t+Q_{gc}\right)_i$
Interception	Bucket Model	$\frac{dS_I}{dt}=P-E_I-P_o$	$\left(\frac{dS_I}{dt}=P-E_I-P_o\right)_i$
Snow melt	ISNOBAL	$\frac{dS_{snow}}{dt}=P-E_{snow}-\Delta w$	$\left(\frac{dS_{snow}}{dt}=P-E_{snow}-\Delta w\right)_i$
Evapotran-spiration	Penman-Monteith Method	$ET_0=\frac{\Delta(R_n-G)+\rho_a C_p\frac{(e_z-e_a)}{r_a}}{\Delta+\gamma(1+\frac{r_s}{r_a})}$	$\left(ET_0=\frac{\Delta(R_n-G)+\rho_a C_p\frac{(e_z-e_a)}{r_a}}{\Delta+\gamma(1+\frac{r_s}{r_a})}\right)_i$



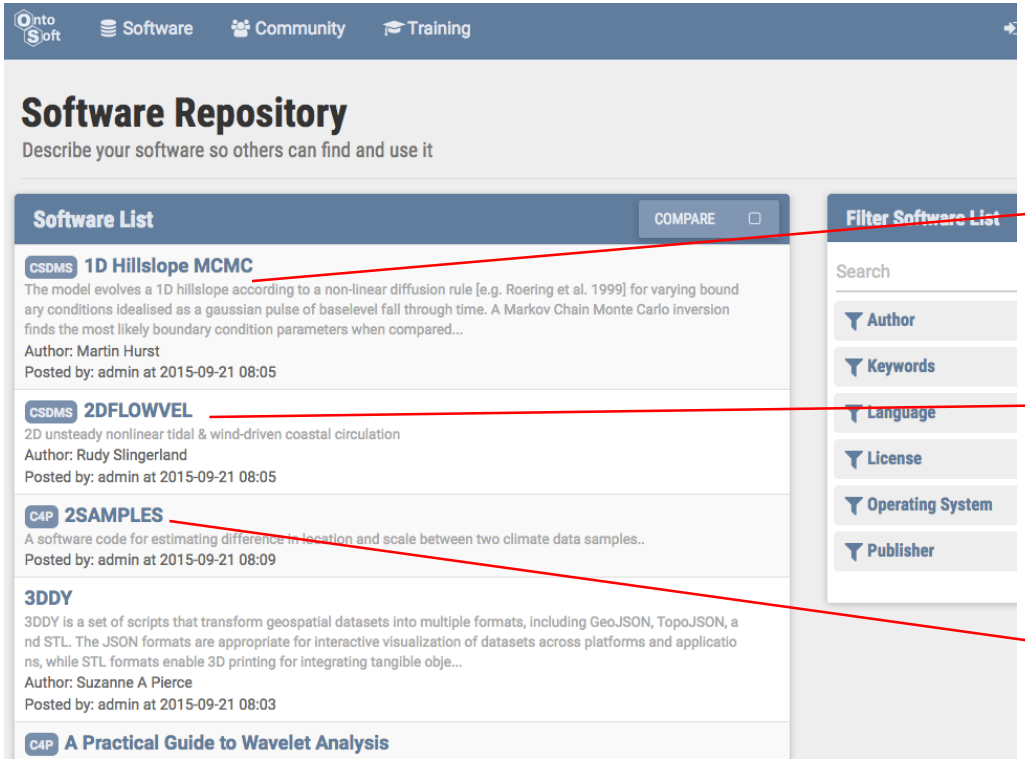
http://www.pihm.psu.edu/

Infrastructure:
Model Repositories

EarthCube's OntoSoft Project: A Software Metadata Registry [Gil et al eScience 2016]

Work with D. Garijo, J. Emile-Geay, D. Khider, V. Ratnakar (USC); N. McKey (NAS)

Searchable metadata in OntoSoft



Codes in shared software repositories

github SOCIAL CODE HOSTING

CSDMS COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

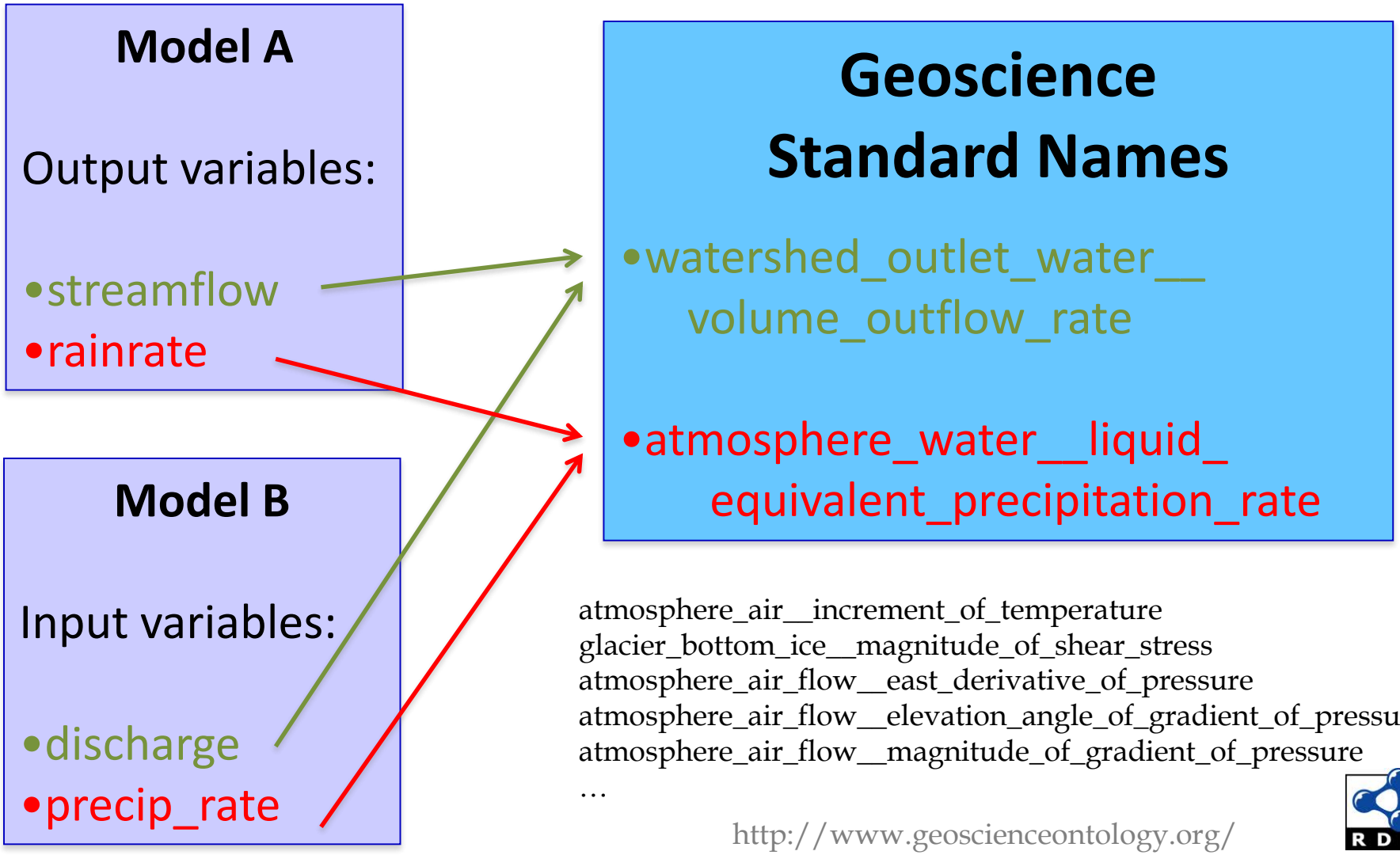
R

(Can export metadata in HTML/XML/RDF/JSON and put in code sharing site)

RDF

http://www.ontosoft.org/

Standard Names for Model Variables [Peckham iEMSs 2014]



Transparency and Reproducibility: Third Global Climate Assessment through GCIS [Tilmes 2014]

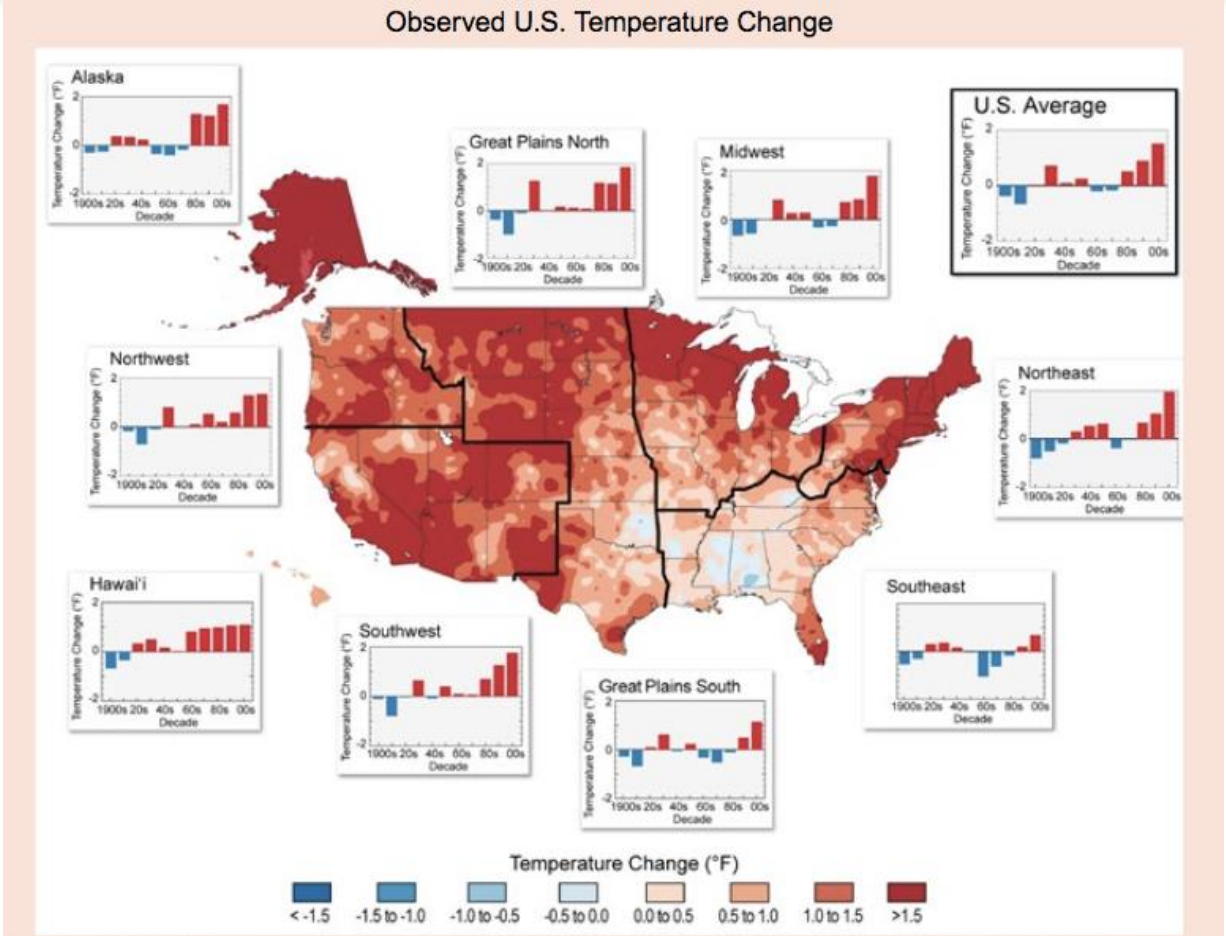


Figure 2.7. The colors on the map show temperature changes over the past 22 years (1991-2012) compared to the 1901-1960 average, and compared to the 1951-1980 average for Alaska and Hawai'i. The bars on the graphs show the average temperature changes by decade for 1901-2012 (relative to the 1901-1960 average) for each region. The far right bar in each graph (2000s decade) includes 2011 and 2012. The period from 2001 to 2012 was warmer than any previous decade in every region. (Figure source: NOAA NCDC / CICS-NC).

http://nca2014.globalchange.gov/downloads

W3C® PROV

Data
+ Models
+ Software
+ Workflow

RDF

scientificpaperofthefuture.org/gpf

Modern Paper

Text:
Narrative of the method, some data is in tables, figures/plots, and the software used is mentioned

Data:
Include data as supplementary materials and pointers to data repositories

Open Science

Sharing:
Deposit data and software (and provenance/workflow) in publicly shared repositories

Open licenses:
Open source licenses for data and software (and provenance/workflow)

Metadata:
Structured descriptions of the characteristics of data and software (and provenance/workflow)

Reproducible Publication

Software:
For data preparation, data analysis, and visualization

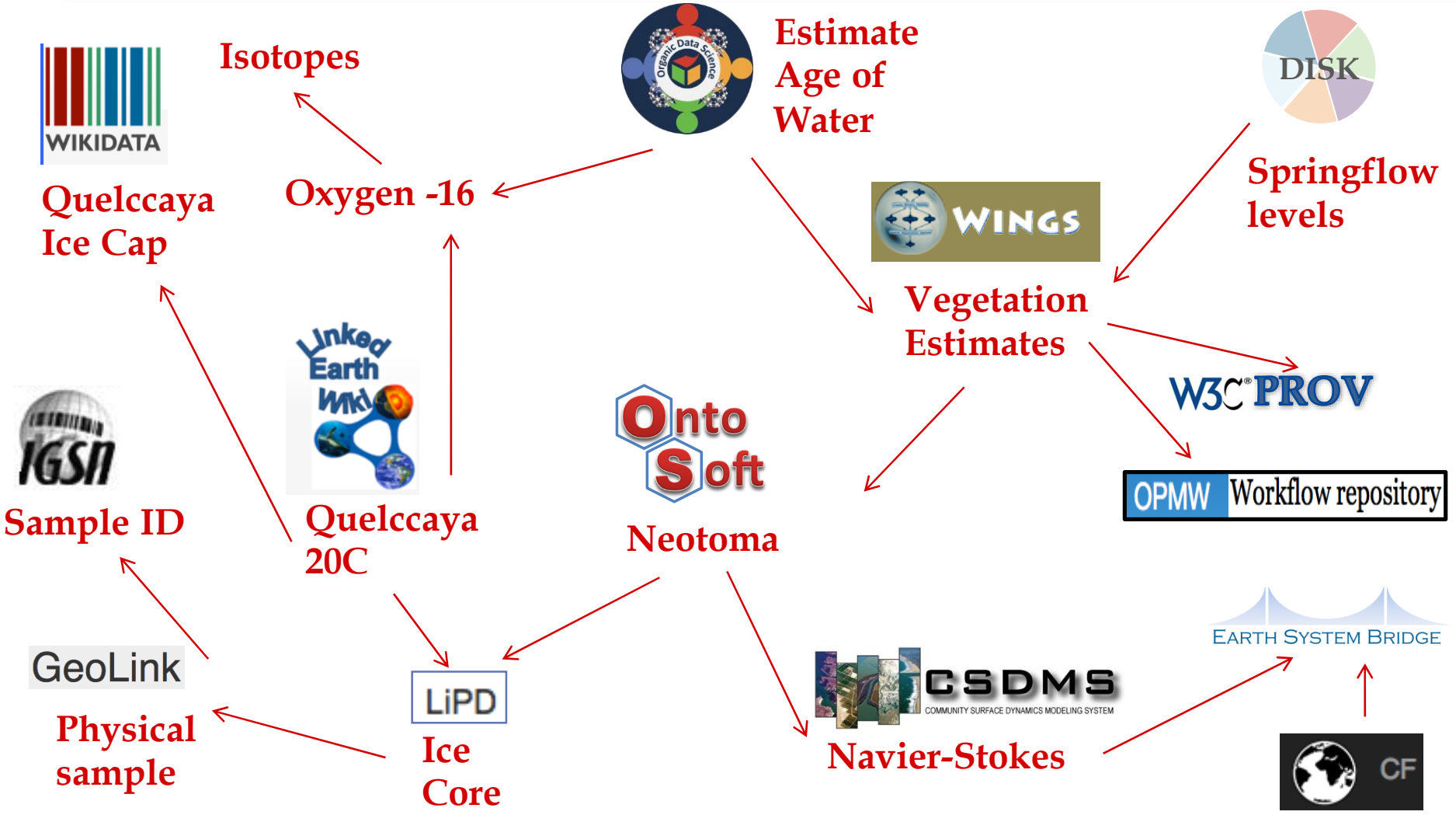
Provenance and methods:
Workflow/scripts specifying dataflow, codes, configuration files, parameter settings, and runtime dependencies

Digital Scholarship

Persistent identifiers:
For data, software, and authors (and provenance/workflow)

Citations:
Citations for data and software (and provenance/workflow)

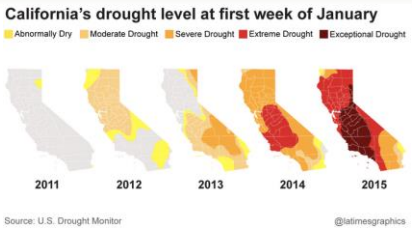
Linked Data and Knowledge in Geosciences: Data + Models + Software + Workflows

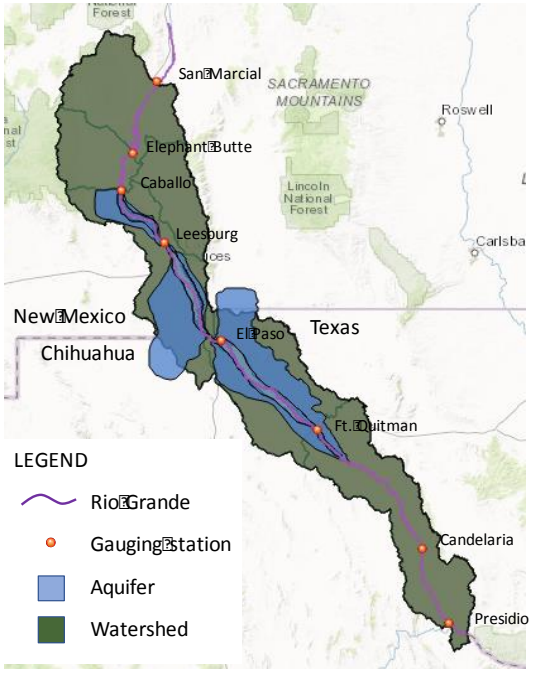


Outline

1. The geosciences landscape
2. Ontologies, vocabularies, standards
3. Infrastructure
   - Data centers
   - Tools
4. Modeling
   - Model repositories
5. Model integration and convergence research
   - The need for semantics

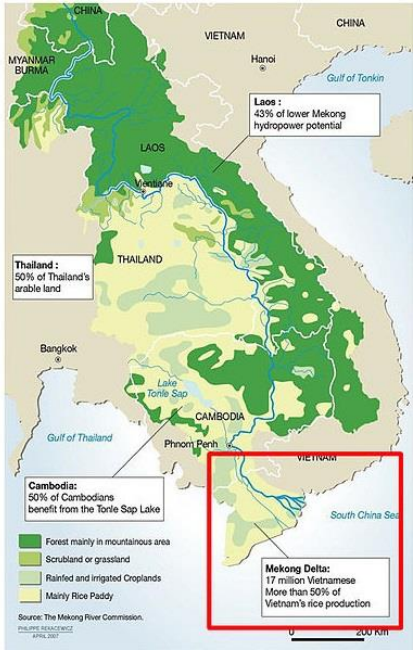
Model Integration Is Needed to Understand Water Use, Land Cover Changes, Food Insecurity,…





Credit: Deana Pennington, Cybershare Project, UT El Paso

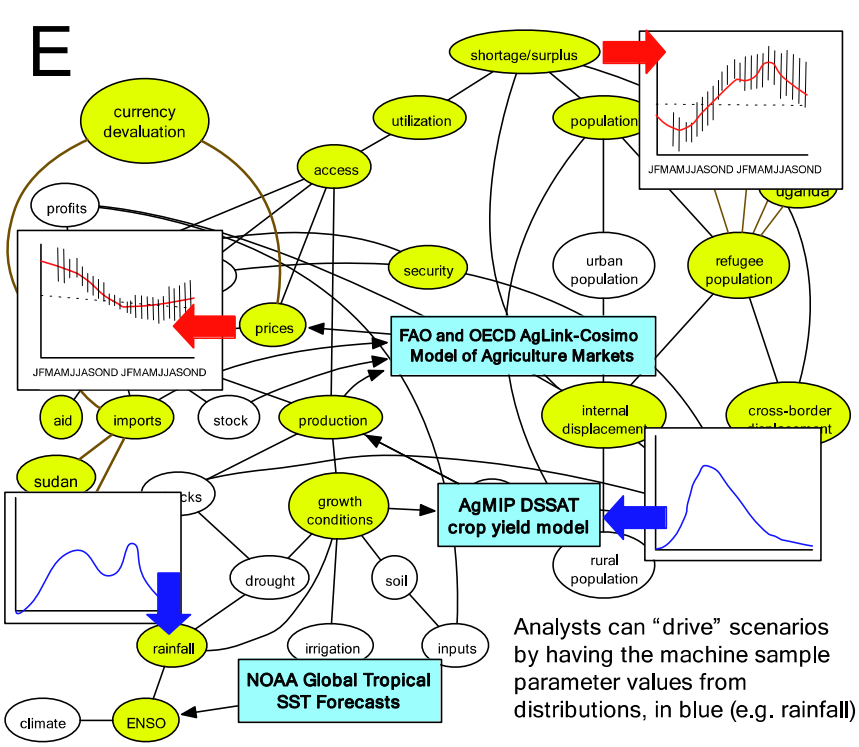
Pecan crops have greatest value but are high water users. Economic value of agriculture is much less than industrial uses, but first in time/first in right in U.S. precludes water allocations to these uses; Mexico has reallocated all surface water to industry

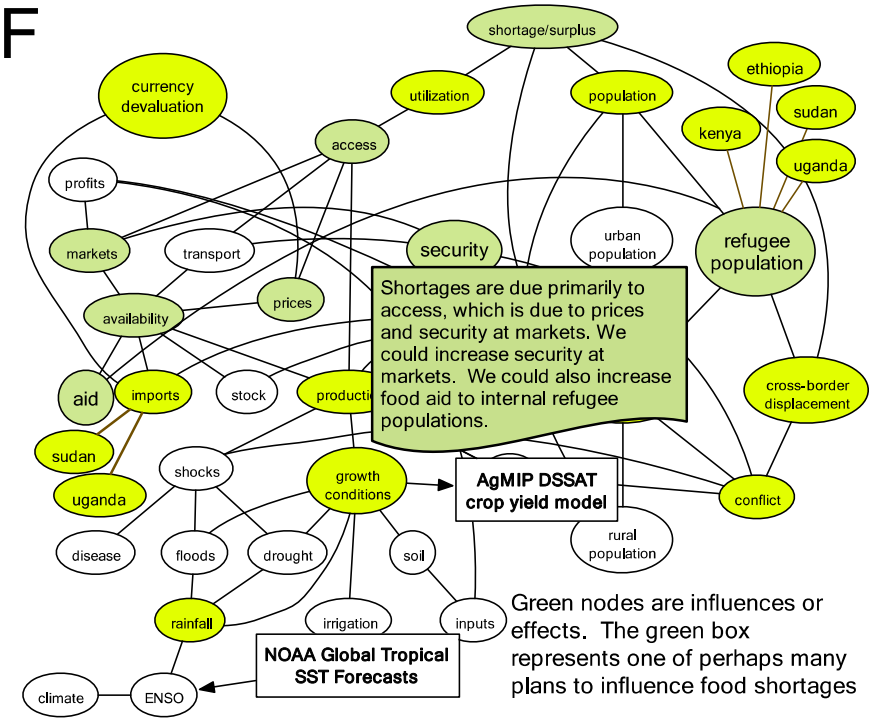


https://news.mongabay.com/2016/10/vietnam-sweats-bullets-as-china-laos-dam-the-mekong/

Extends through extends through Tibet, South China, Thailand, Laos, Myanmar, Cambodia, and Vietnam. More than 70 dams are planned in several nations. Recorded deeper droughts and bigger floods than ever before. 2M tons of fish and 500,000 tons of other aquatic animals. Forest cover has decreased from 73 percent in 1973 to 63 percent in 1993. Rice in Cambodia…

New DARPA World Modelers Program

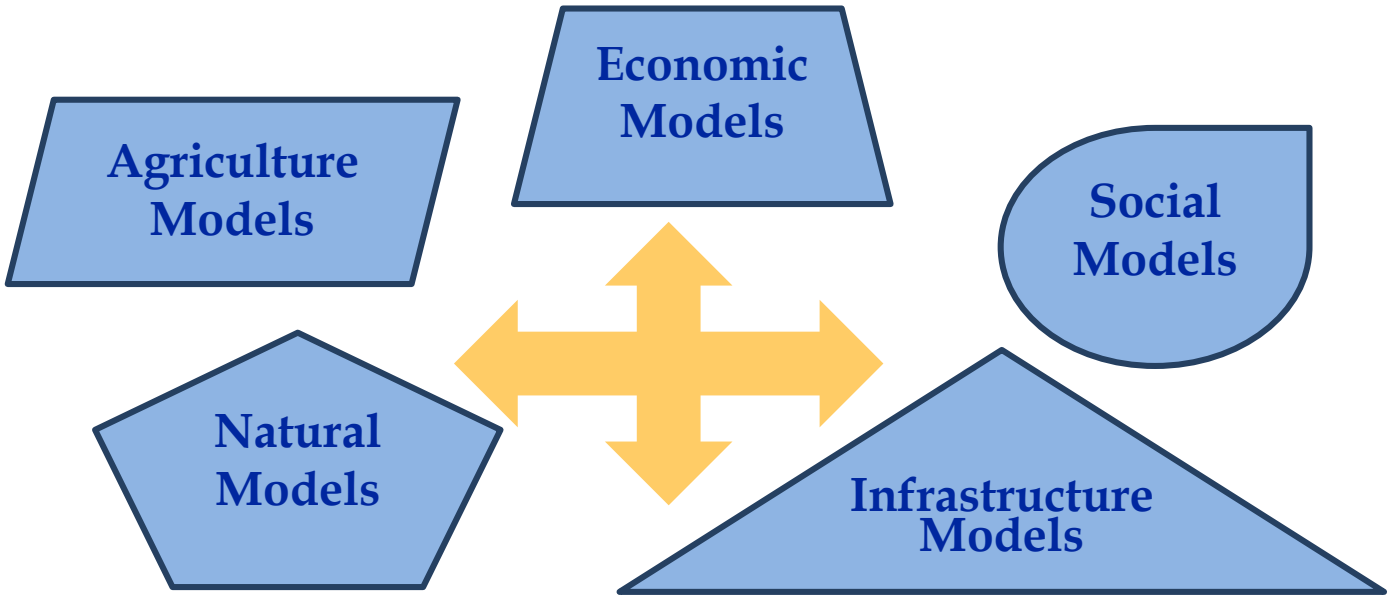




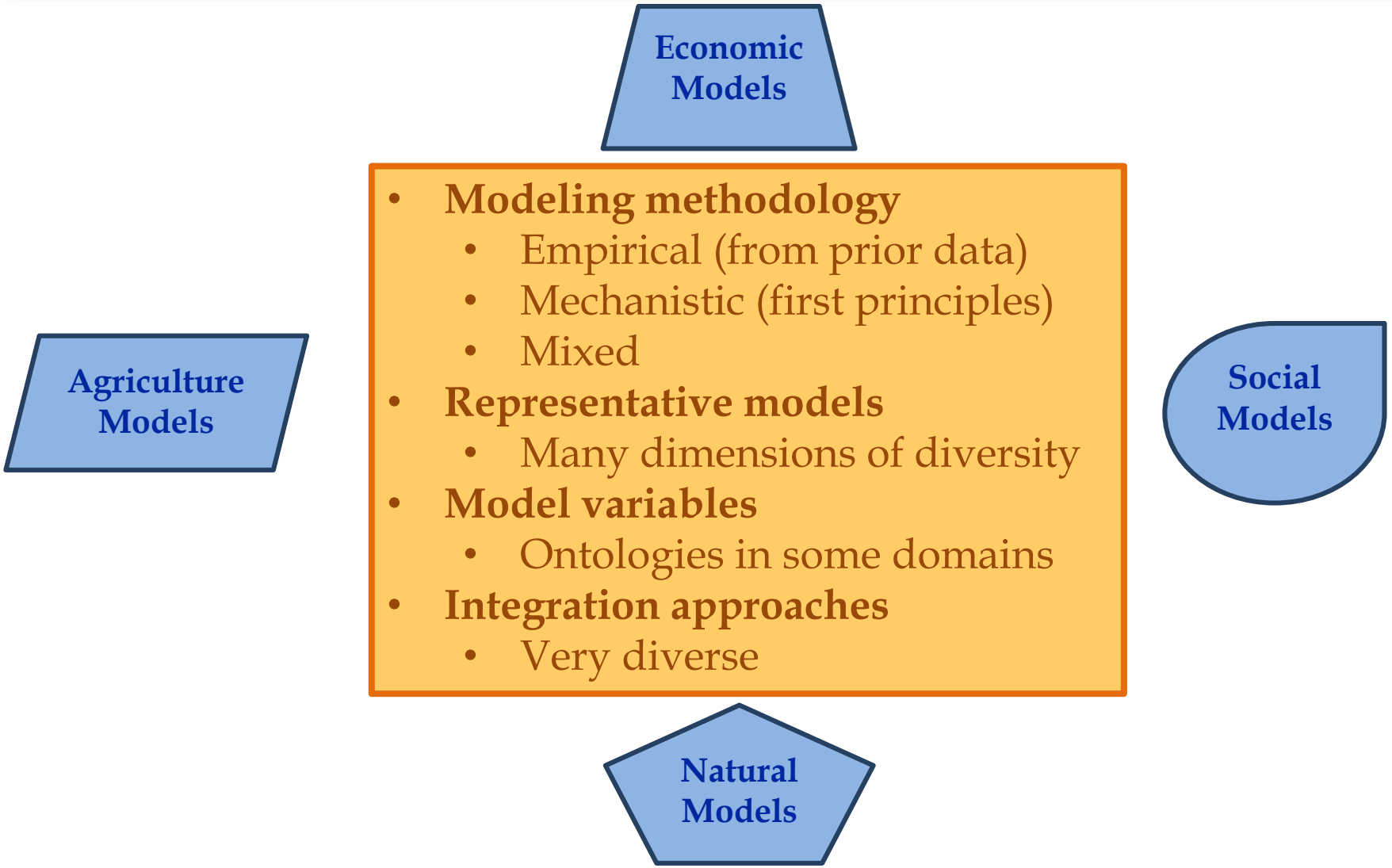
"World Modelers aims to develop technologies to facilitate analyses that are comprehensive, targeted, causal, quantitative, probabilistic, and timely enough to recommend specific actions that could avert crises."

New DARPA World Modelers Program: Semantic Challenges in Model Integration

- **A challenging aspect is mapping model variables**
  - **Standard ontologies needed to describe diverse models**

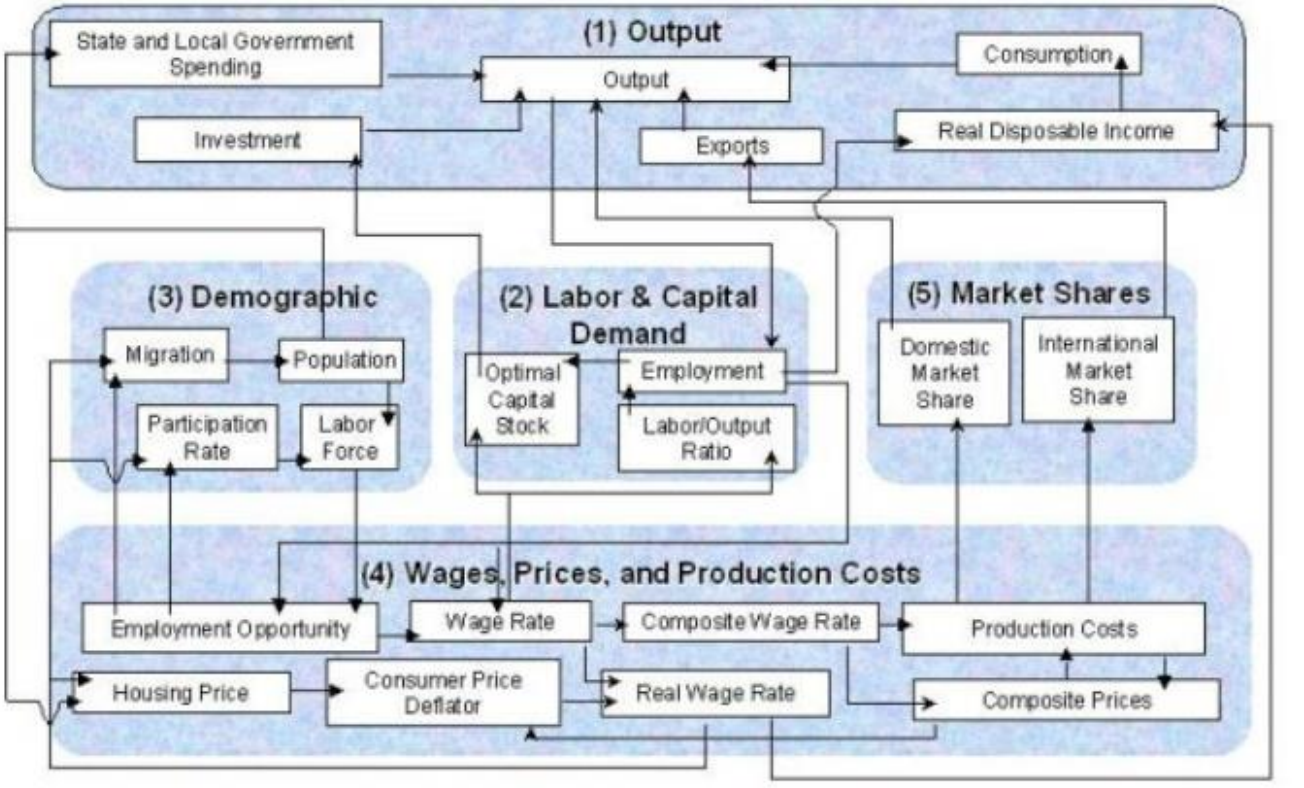


New DARPA World Modeler Program: Model INTegration (MINT) Project [Gil et al 2017]

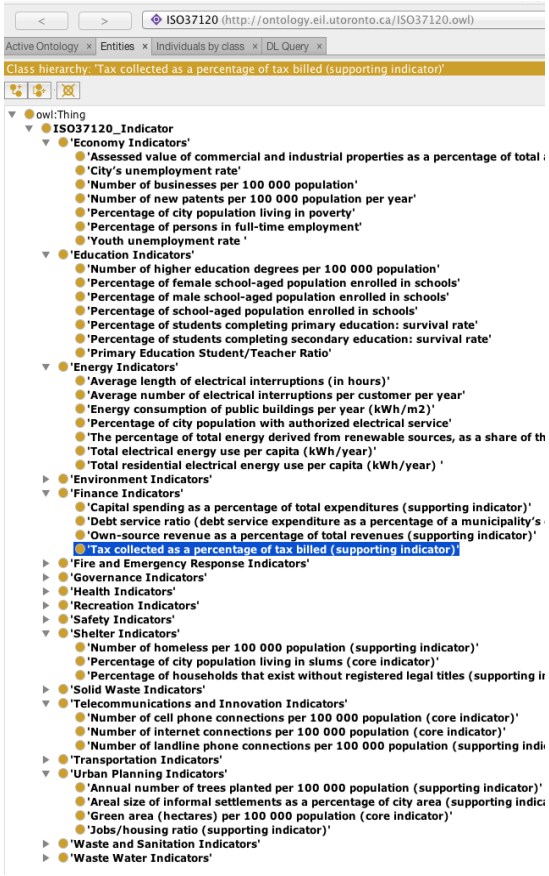


Economic Models: Very Different from Natural Models, Difficult to Reuse and Integrate

REMI



City Indicators (ISO-37120)

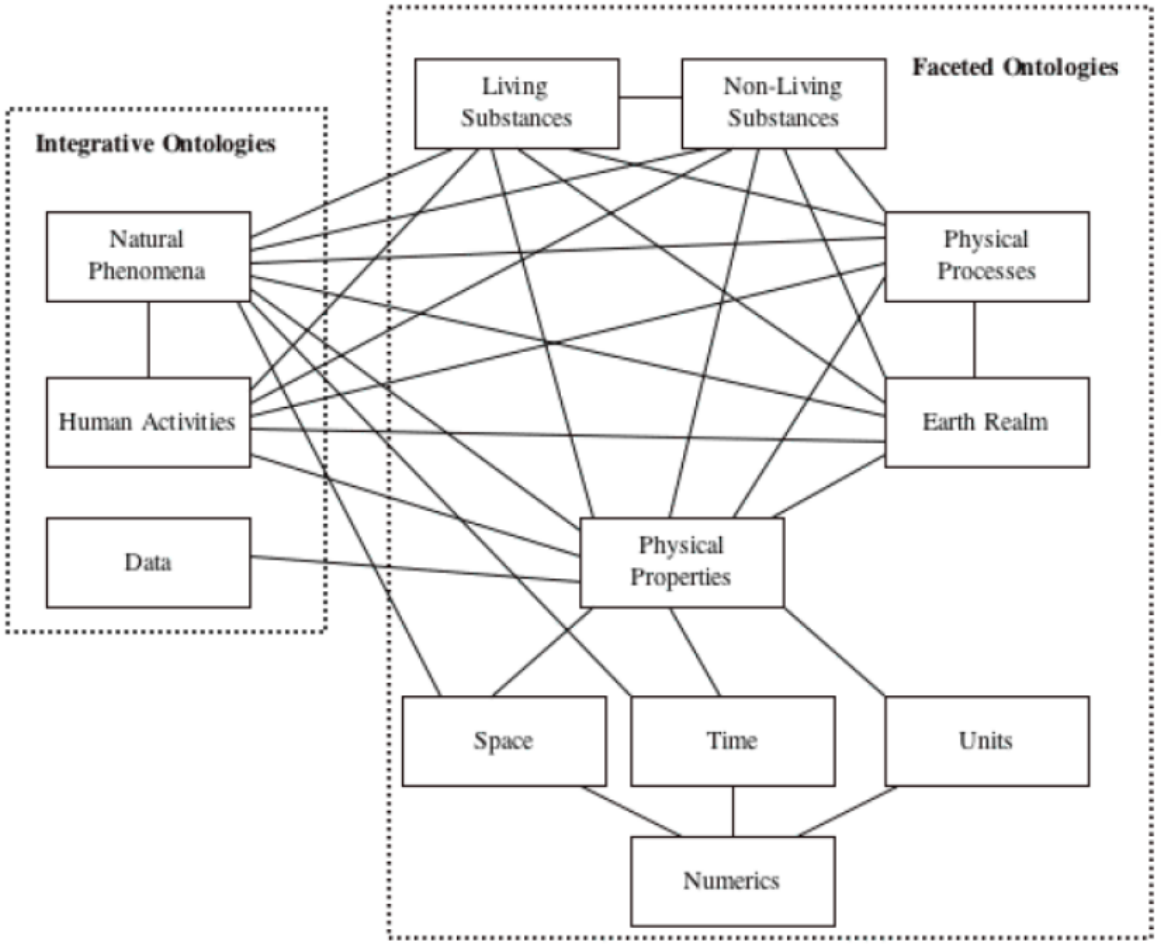


Summary

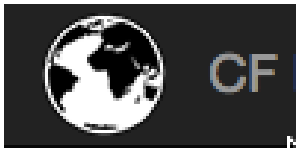
1. The geosciences landscape
   - Initial focus could be the NSF CISE-GEO EarthCube initiative
2. Ontologies, vocabularies, standards
   - ESIP Community Ontology Repository
3. Infrastructure
   - Data centers generally speak RDF
   - Tools developed in many EarthCube projects
4. Modeling
   - Model repositories
5. Model integration
   - Model reuse and integration requires semantics
   - Model integration is at the heart of convergent research in geosciences with great societal impact

Ontologies and Vocabularies: Examples

SWEET Ontologies [Raskin and Pan 2005]



https://doi.org/10.1016/j.cageo.2004.12.004

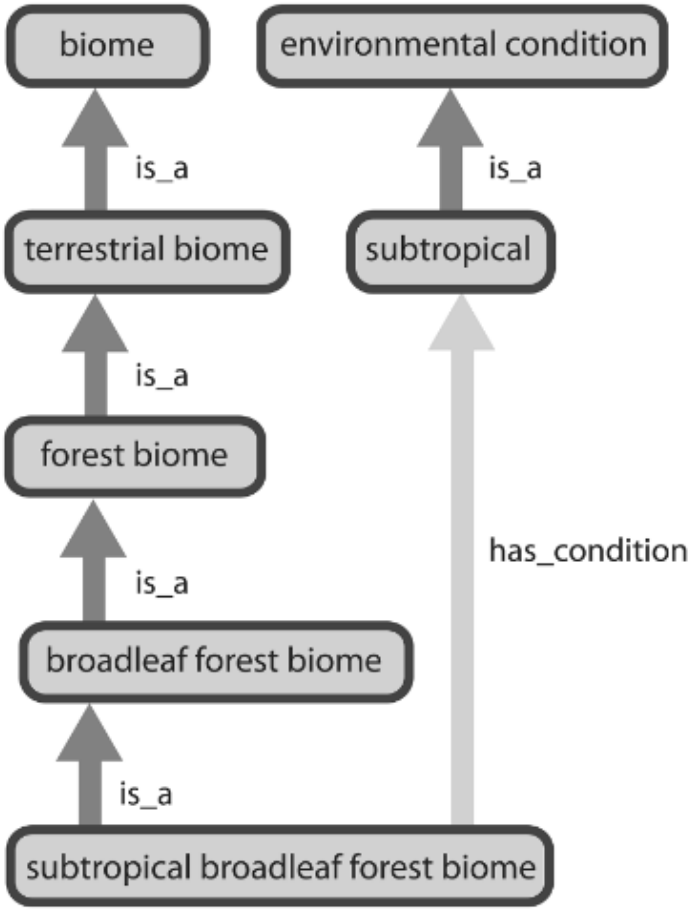


Name: change_over_time_in_surface_snow_amount

Description: The surface called "surface" means the lower boundary of the atmosphere. "change_over_time_in_X" means change in a quantity X over a time-interval, which should be defined by the bounds of the time coordinate. "Amount" means mass per unit area. Surface amount refers to the amount on the ground, excluding that on the plant or vegetation canopy.

Canonical units: kg m-2

ENVO [Buttigieg et al 2013]



https://doi.org/10.1186/2041-1480-4-43

EarthCube's Linked Earth Project:
Creating New Metadata Properties as Needed

Lake Bosumtwi Sediments Dataset

Data

- **DOWNLOAD**

From: http://www.organicdatapublishing.org/index.php/Lake_Bosumtwi_Sediments_D

Structured Properties

Add

[x]	Archive	LakeSediments	(By
[x]	Doi	10.1126/science.1166352	(By
[x]	Domain(s)	Paleolimnology, Geochemis…	(By
[x]	Interpretation	Lake Level	(By
[x]	Is a	Paleoclimate dataset	(By
[x]	Measurement	Delta18O	(By
[x]	MeasurementMaterial	Authigenic Calcite	(By
[x]	MeasurementStandard	VPDB	(By
[x]	MeasurementUnits	Permil	(By
[x]	Reference	Shanahan et al. 2009	(By
[x]	Resolution	5	(By
[x]	ResolutionUnits	Years	(By
[x]	SiteLatitude	6.5	(By
[x]	SiteLatitudeUnits	Degrees North	(By
[x]	SiteLongitude	358.58	(By
[x]	SiteLongitudeUnits	Degrees East	(By
[x]	SiteName	Bosumtwi	(By
[x]	TimeUnits	Years CE	(By
[x]	WhoAnalysed	Tim Shanahan	(By

Palmyra coral 20C

Data

- **DOWNLOAD**

From: http://www.ncdc.noaa.gov/paleo/metadata/noaa-coral-1865.html

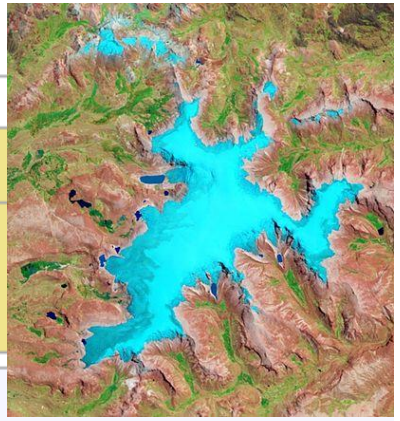
Structured Properties

Add

[x]	Archive	Coral	(By Julien)
[x]	Doi	10.1029/2001GL012919	(By Julien)
[x]	Domain(s)	Climate,geochemistry	(By Julien)
[x]	Forward model	10.1029/2011GL048224	(By Julien)
[x]	Genus	Porites	(By Julien)
[x]	Interpretation	SST,SSS	(By Julien)
[x]	Is a	Paleoclimate dataset	(By Gil)
[x]	Measurement	Delta18O	(By Julien)
[x]	MeasurementMaterial	Skeletal aragonite	(By Julien)
[x]	MeasurementStandard	VPDB	(By Julien)
[x]	MeasurementUnits	Permil	(By Julien)
[x]	Reference	Cobb et al, 2001	(By Julien)
[x]	Resolution	0.0833	(By Julien)
[x]	ResolutionUnits	Years	(By Julien)
[x]	SiteLatitude	5.883	(By Julien)
[x]	SiteLatitudeUnits	Degrees North	(By Julien)
[x]	SiteLongitude	197.917	(By Julien)
[x]	SiteLongitudeUnits	Degrees East	(By Julien)
[x]	SiteName	Palmyra	(By Julien)
[x]	Species	Blue porites	(By Anonymous)
[x]	Species	Lutea	(By Julien)

EarthCube's Linked Earth Project:
Promoting Property Normalization and Standards

Quelccaya Ice Core Dataset



Data

- **DOWNLOAD**
- **Data Types**

Structured Properties

Add

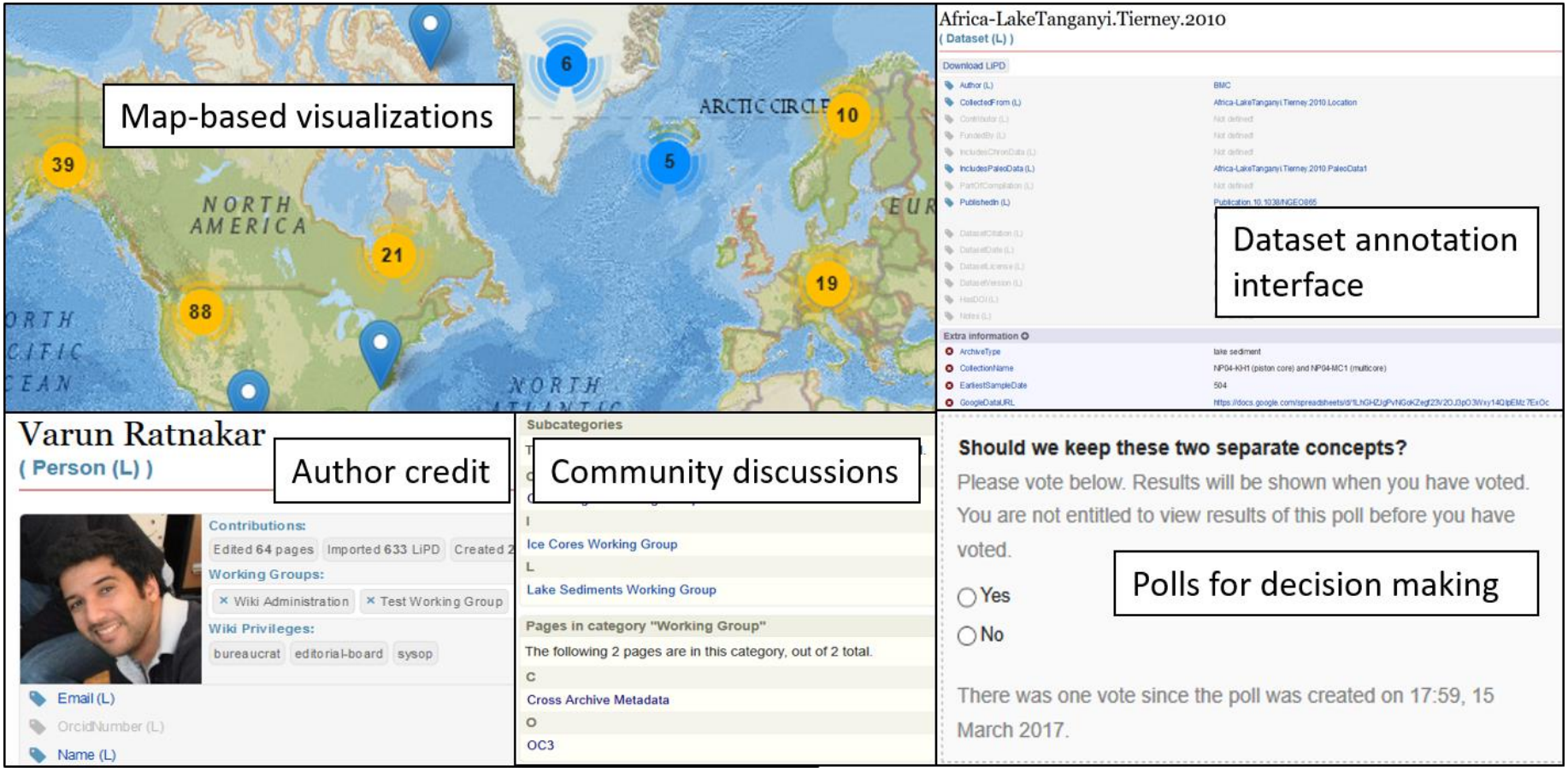
[x]	Archive	IceCore	(By Julien)
[x]	Doi	10.1073/pnas.0603900103	(By Julien)
[x]	Domain(s)	Climate,Glaciology,Geoche…	(By Julien)
[x]	Is a	Paleoclimate dataset	(By Gil)
[x]	Measu		

Measurement
MeasurementMaterial
MeasurementStandard
MeasurementUnits

EarthCube's Linked Earth Project: Connecting to Other Ontologies/Data

- NCBITaxon ontology: porites http://bioportal.bioontology.org/ontologies/NCBITAXON?p=classes&conceptid=http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FNCBITAXON%2F46719

EarthCube's Linked Earth Project: Social Aspects of Vocabulary Crowdsourcing



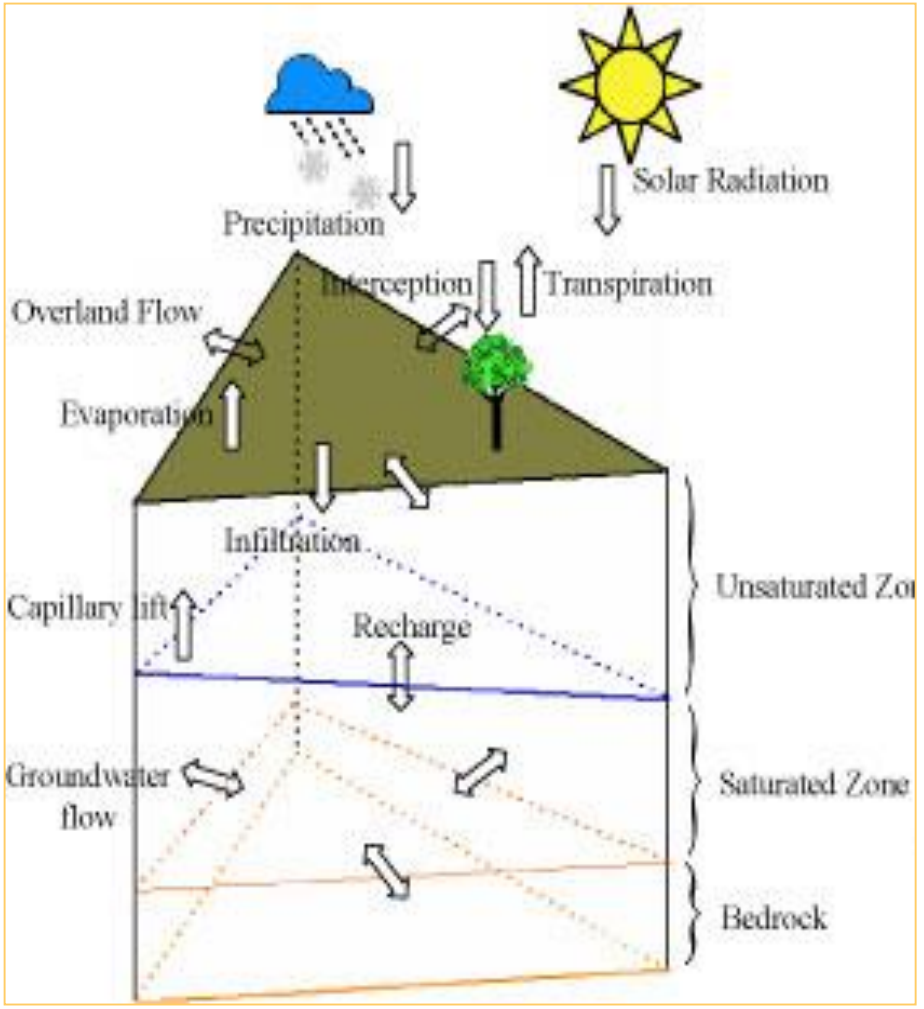
Modeling in Geosciences: Models of Dynamical Systems

- A **simulation model** of a dynamical system captures the relationships and dependencies between a set of variables used to describe it
- Models of dynamical systems are framed in time and space
- Output variables depend on the input variables, internal state variables, exogenous variables, and random variables
- Models can have parameters that can be adjusted to fit the empirical observations taken on the system being modeled
- Models are used to make predictions about hypothetical configurations and future states of the target system

Example:
The PIHM Hydrology Model [Duffy et al 2015]

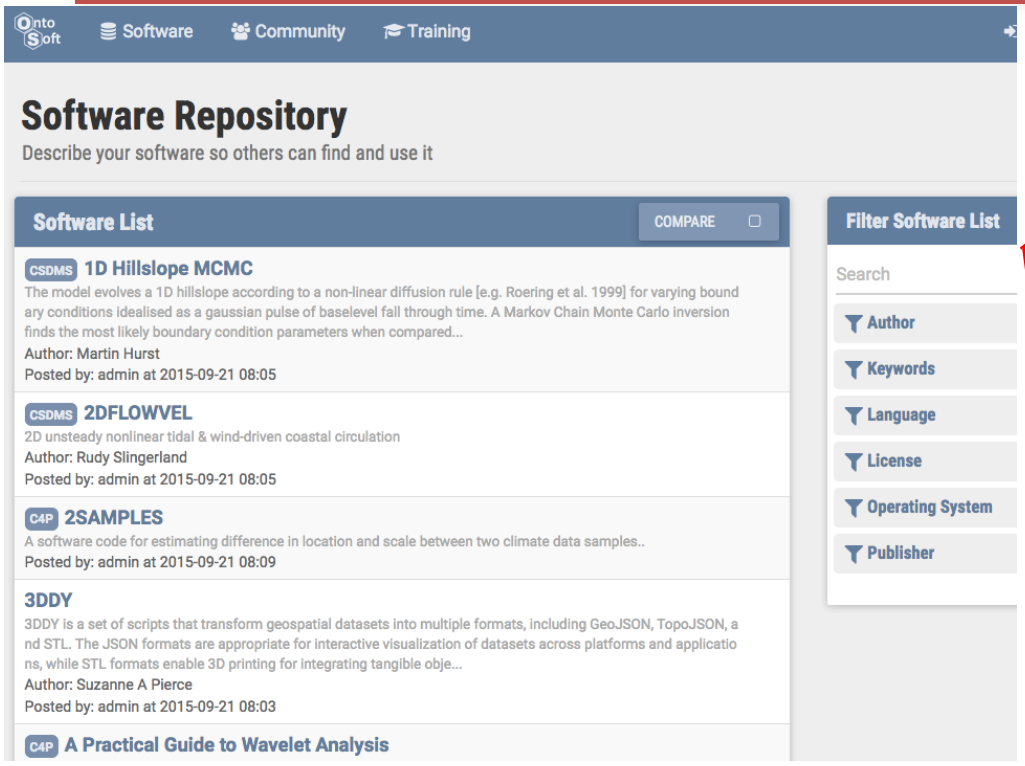
- Historical observational data for calibration
- Forecast data for prediction
- Observational data for evaluation

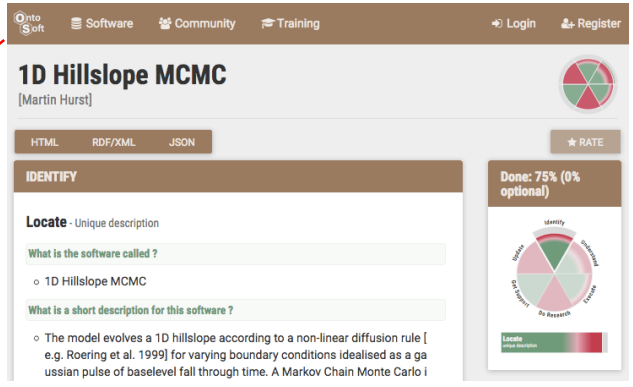
Process	Governing equation/model	Original governing equations	Semi-discrete form ODEs
Channel Routing	St. Venant Equation	$\frac{\partial h}{\partial t}+\frac{\partial(uh)}{\partial x}=q$	$\left(\frac{dz_c}{dt}=P_c-\sum Q_{gc}+\sum Q_{oc}+Q_{in}-Q_{out}-E_c\right)_i$
Overland Flow	St. Venant Equation	$\frac{\partial h}{\partial t}+\frac{\partial(uh)}{\partial x}+\frac{\partial(vh)}{\partial y}=q$	$\left(\frac{\partial h}{\partial t}=P_o-I-E_o-Q_{oc}+\sum_{j=1}^{3}Q_o^{\,0}\right)_i$
Unsaturated Flow	Richard Equation	$C(\psi)\frac{\partial\psi}{\partial t}=\nabla\cdot(K(\psi)\nabla(\psi+Z))$	$\left(\frac{d\xi}{dt}=I-q^{k}-ET_s\right)_i$
Groundwater Flow	Richard Equation	$C(\psi)\frac{\partial\psi}{\partial t}=\nabla\cdot(K(\psi)\nabla(\psi+Z))$	$\left(\frac{d\zeta}{dt}=q^{3}+\sum_{j=1}^{3}Q_g^{\,0}-Q_t+Q_{gc}\right)_i$
Interception	Bucket Model	$\frac{dS_I}{dt}=P-E_I-P_o$	$\left(\frac{dS_I}{dt}=P-E_I-P_o\right)_i$
Snow melt	ISNOBAL	$\frac{dS_{snow}}{dt}=P-E_{snow}-\Delta w$	$\left(\frac{dS_{snow}}{dt}=P-E_{snow}-\Delta w\right)_i$
Evapotran-spiration	Penman-Monteith Method	$ET_0=\frac{\Delta(R_n-G)+\rho_a C_p\frac{(e_s-e_a)}{r_a}}{\Delta+\gamma(1+\frac{r_s}{r_a})}$	$\left(ET_0=\frac{\Delta(R_n-G)+\rho_a C_p\frac{(e_s-e_a)}{r_a}}{\Delta+\gamma(1+\frac{r_s}{r_a})}\right)_i$

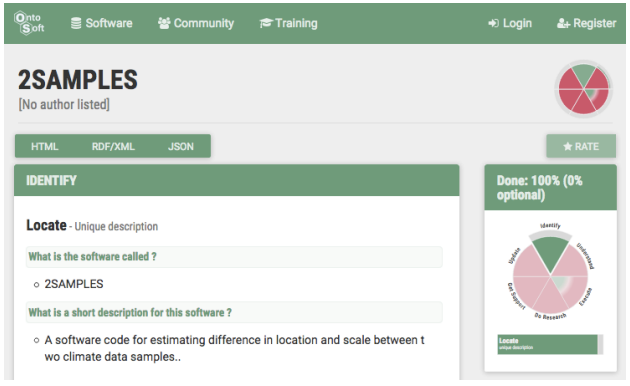


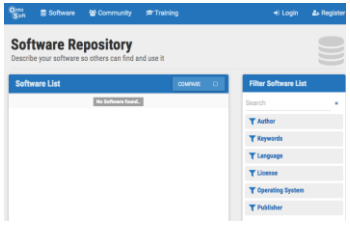
http://www.pihm.psu.edu/

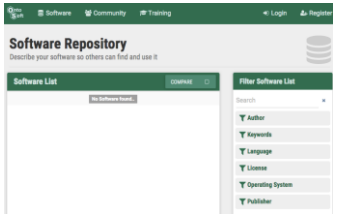
EarthCube's OntoSoft Project: Distributed Architecture for Software Registries

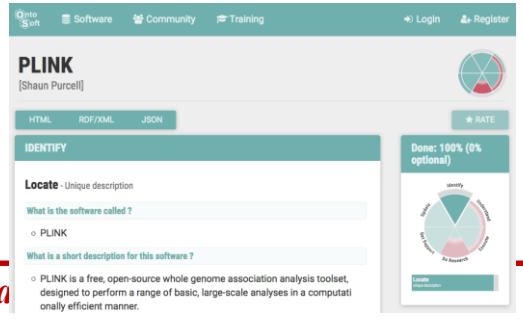




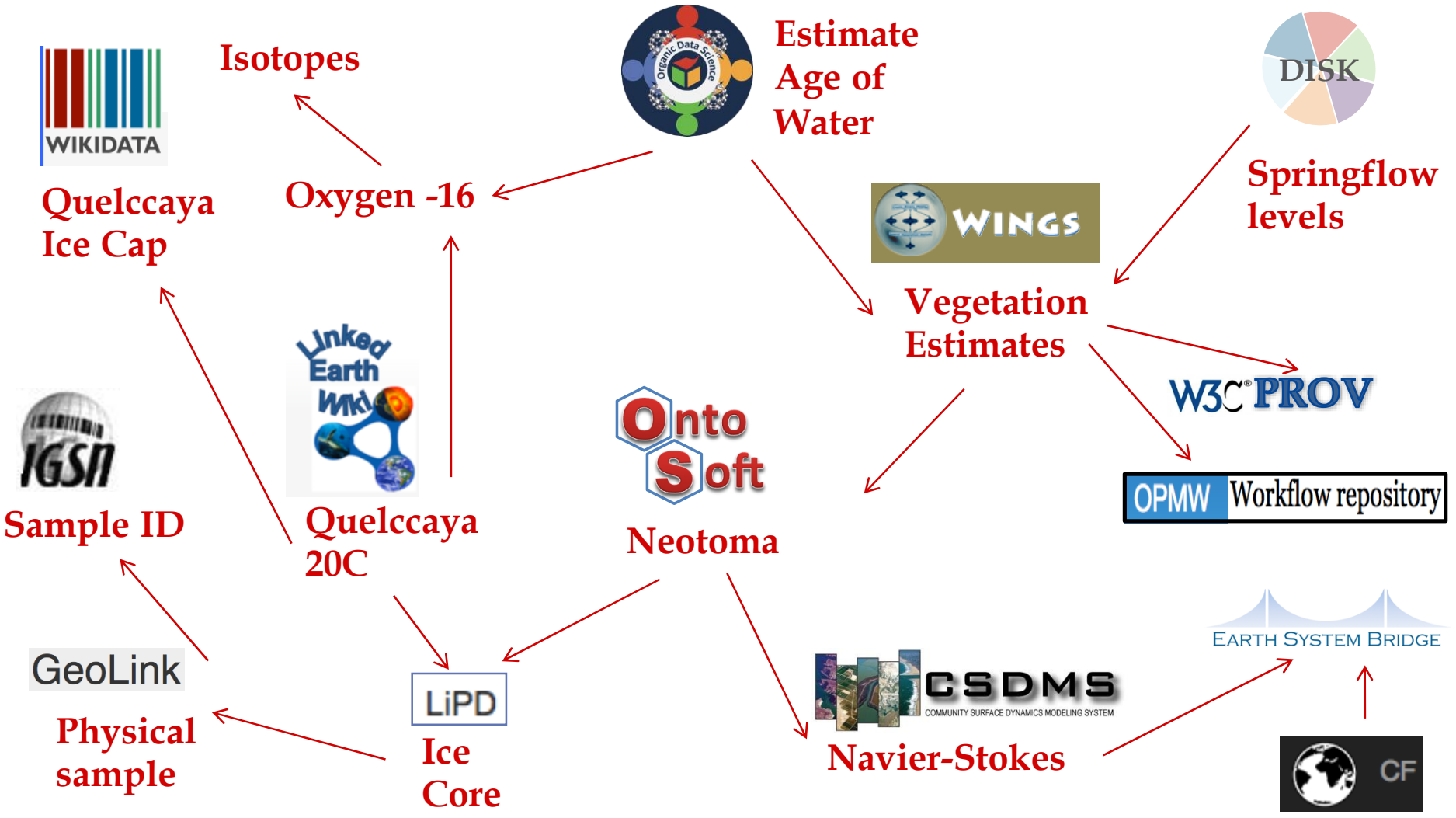








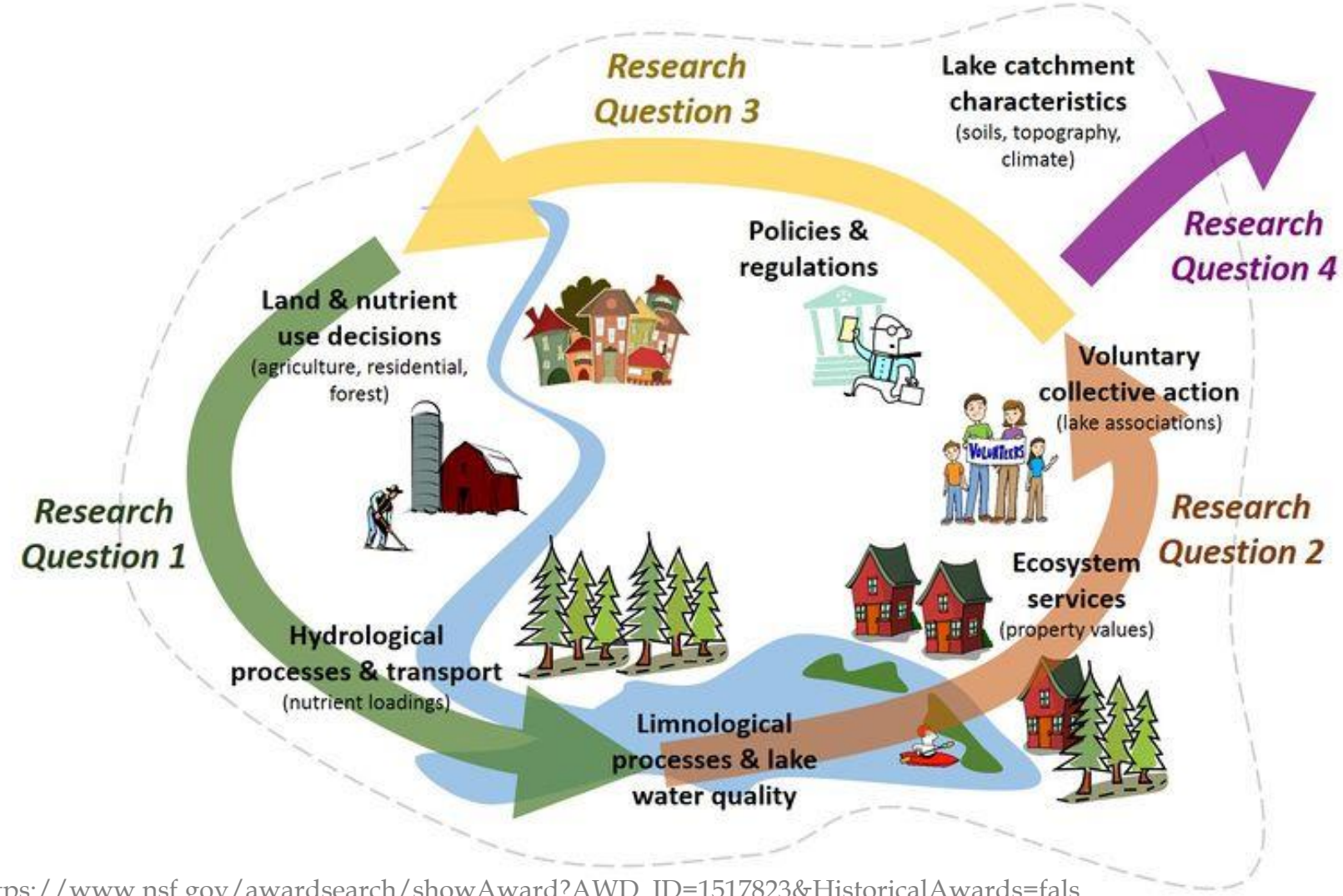
Linked Science Data and Knowledge: Data + Models + Software + Workflows



Modeling in Geosciences: Models of Dynamical Systems

- A **simulation model** of a dynamical system captures the relationships and dependencies between a set of variables used to describe it
- Models of dynamical systems are framed in time and space
- Output variables depend on the input variables, internal state variables, exogenous variables, and random variables
- Models can have parameters that can be adjusted to fit the empirical observations taken on the system being modeled
- Models are used to make predictions about hypothetical configurations and future states of the target system

Coupling Natural and Human Systems [Cobourn, Duffy, Hanson, et al 2016]



https://www.nsf.gov/awardsearch/showAward?AWD_ID=1517823&HistoricalAwards=fals

Model Integration: Diversity of Strategies

Model Blending

Interleaved Execution

Sliced execution by time tic: first one model, then another, in a round-robin way

eg: CSDMS

Implicit Interleaving

Collection of equations that are designed to be solved together, then a solver runs them.

eg: CGE economic models

Code Merging

MPI code to implement all models

eg: earthquake simulations

Shared Memory

Models share a R/W memory

eg: Synthetic Information

Model Combination

Result Chaining

The result of a model is input to another model, as in a workflow

eg: pSIMS, CEMSA

Output Comparison

Results from several models (or the same model) are aggregated (eg, an ensemble)

eg: regional weather prediction

Output Analysis

Same model is run with many configurations or parameter values, to do parameter estimation, sensitivity analysis, or uncertainty quantification

Model Distribution

Code Parallelization

The model is implemented as parallel code (eg to process each grid cell separately)

Interleaved Behavior

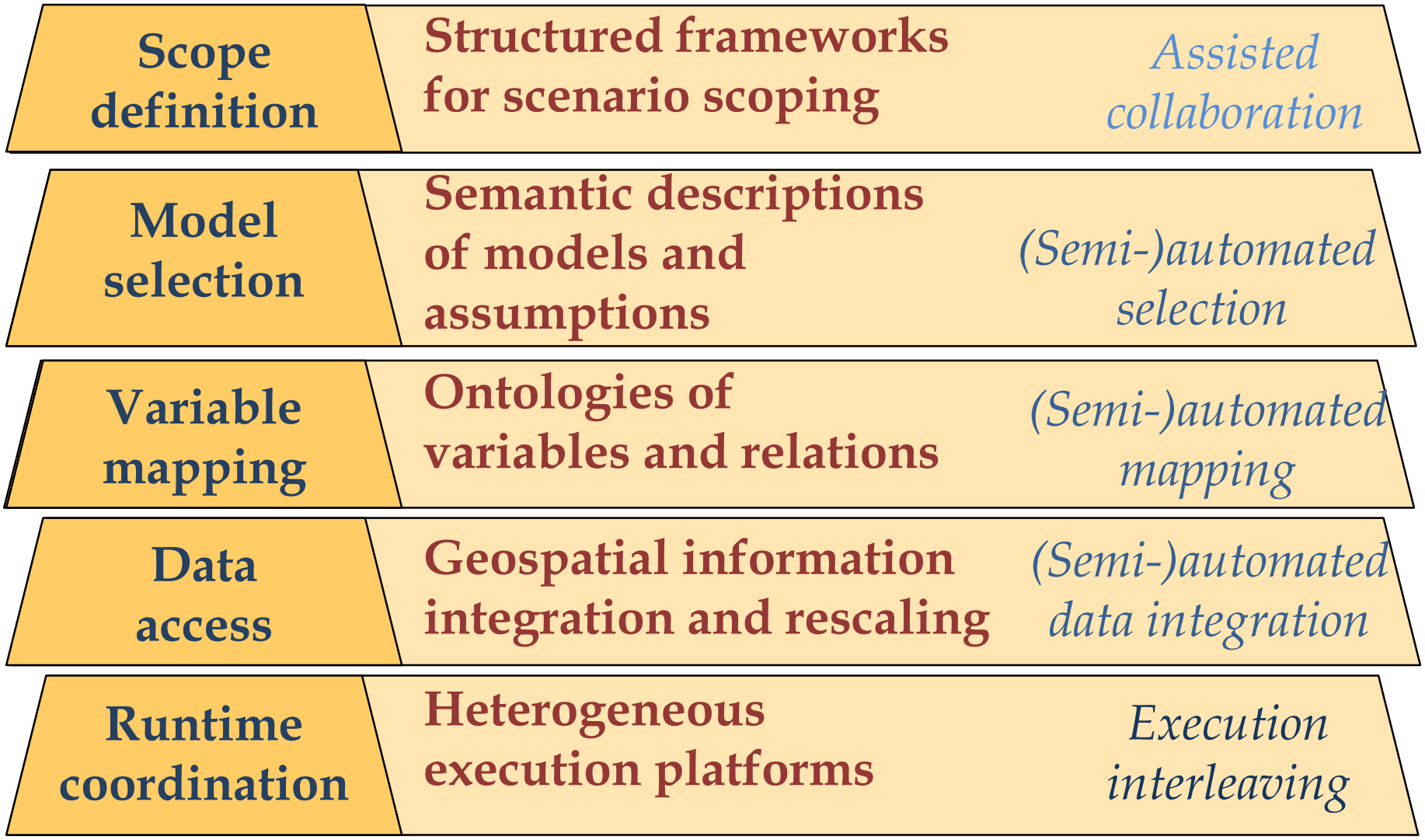
Individual agents proceed based on information made available to their simulation environment

Ex: agent-based frameworks

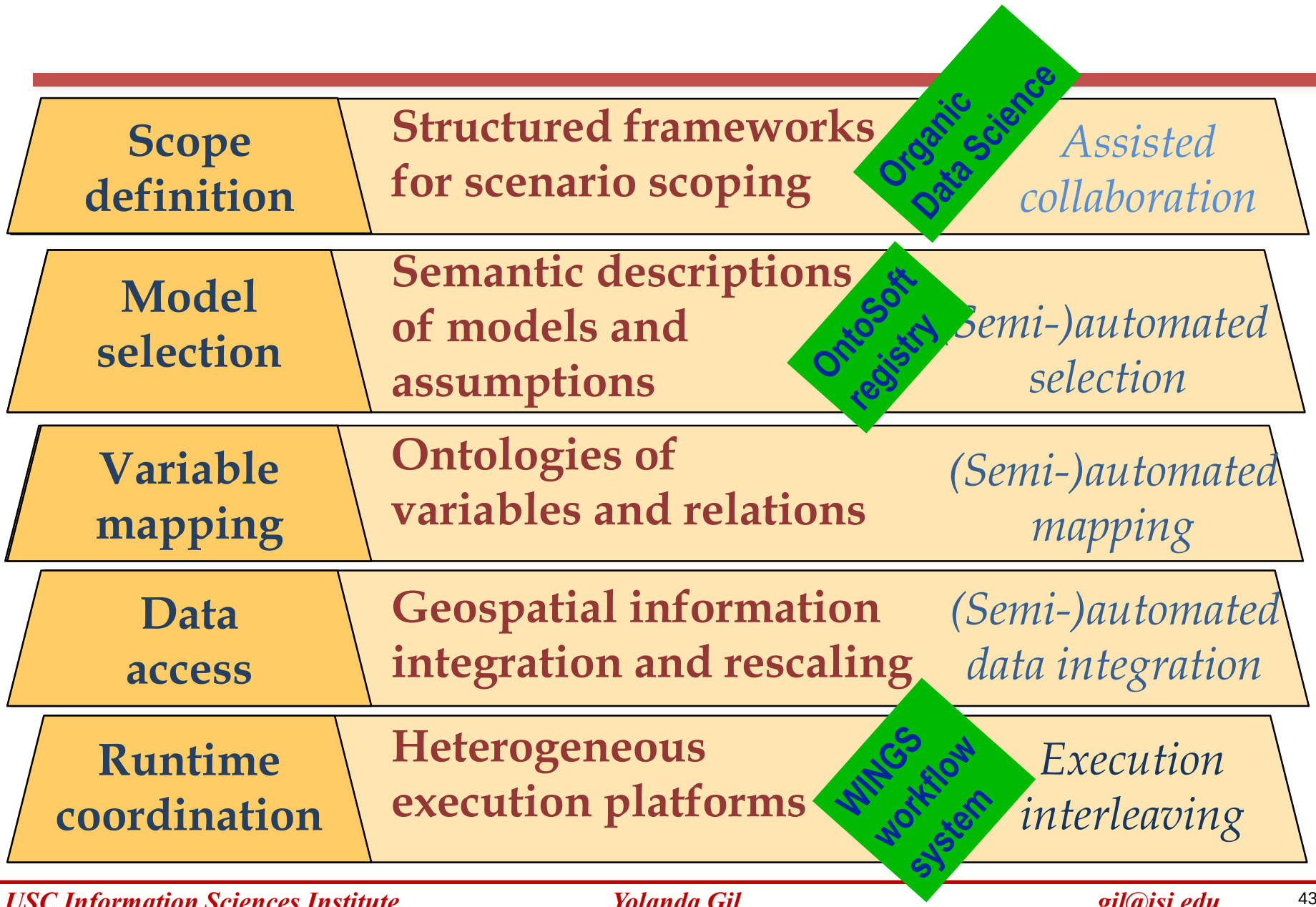
Integrated Behavior

Agents are given several behavior models that determine their actions

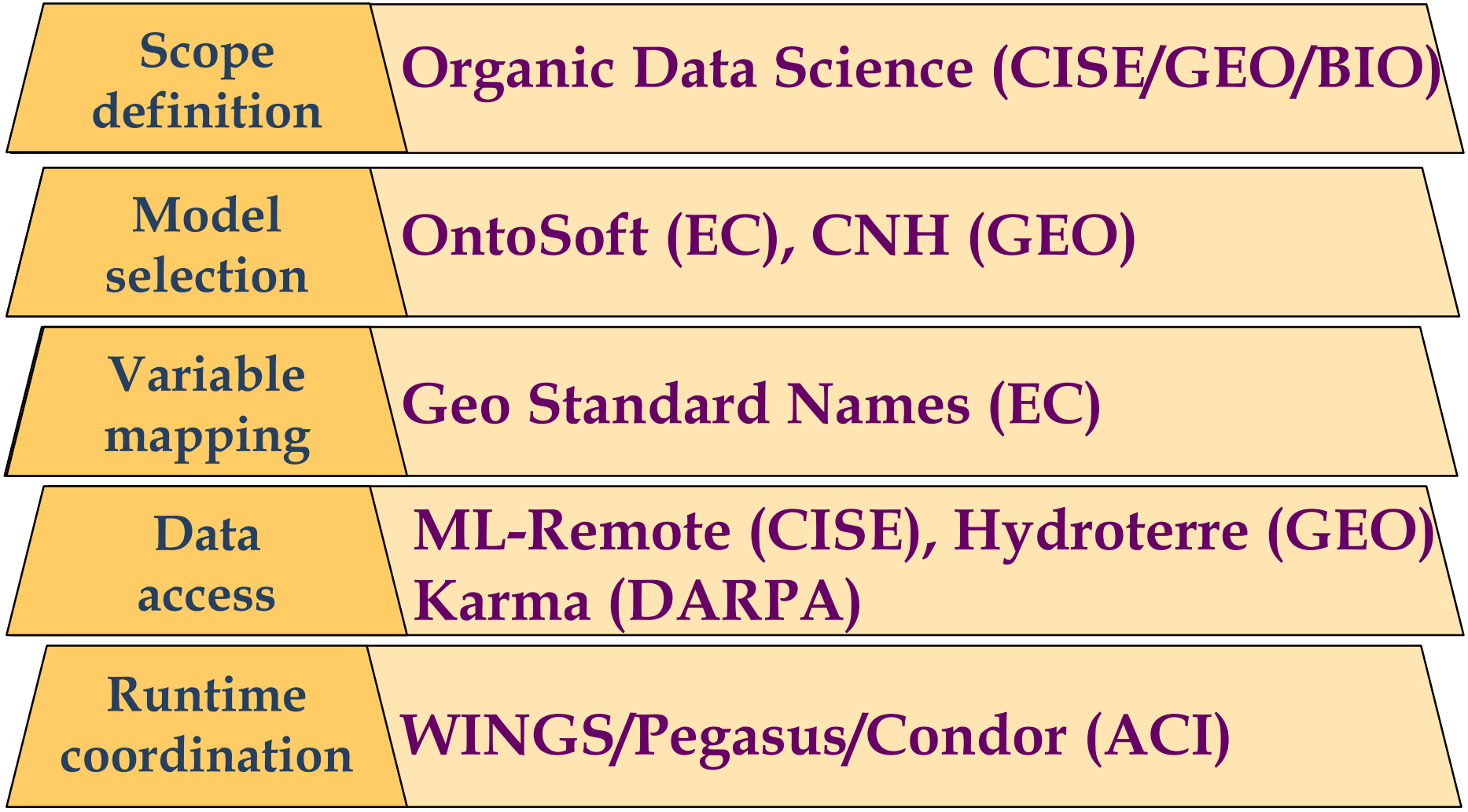
A Research Agenda for Model Integration



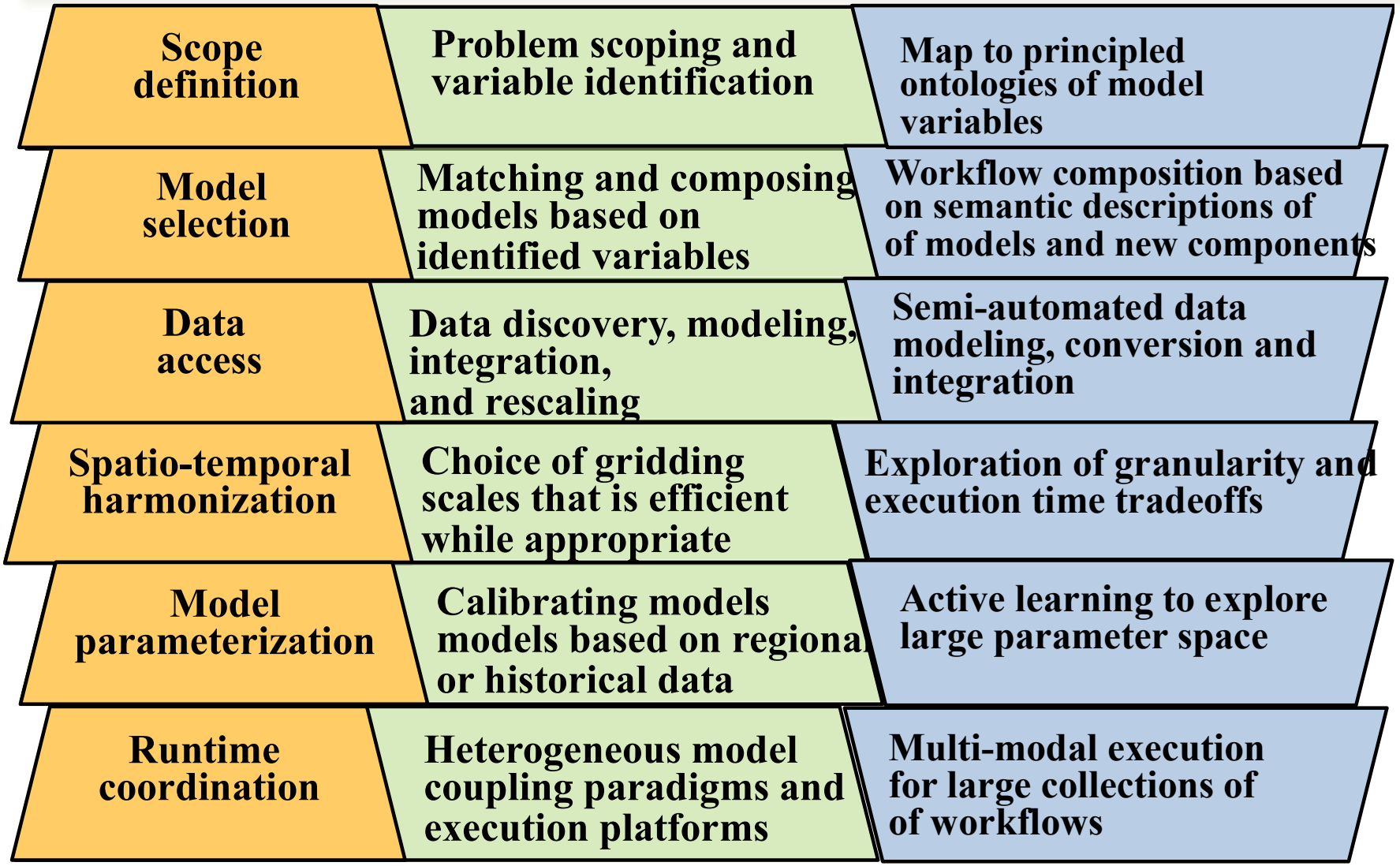
A Research Agenda for Model Integration



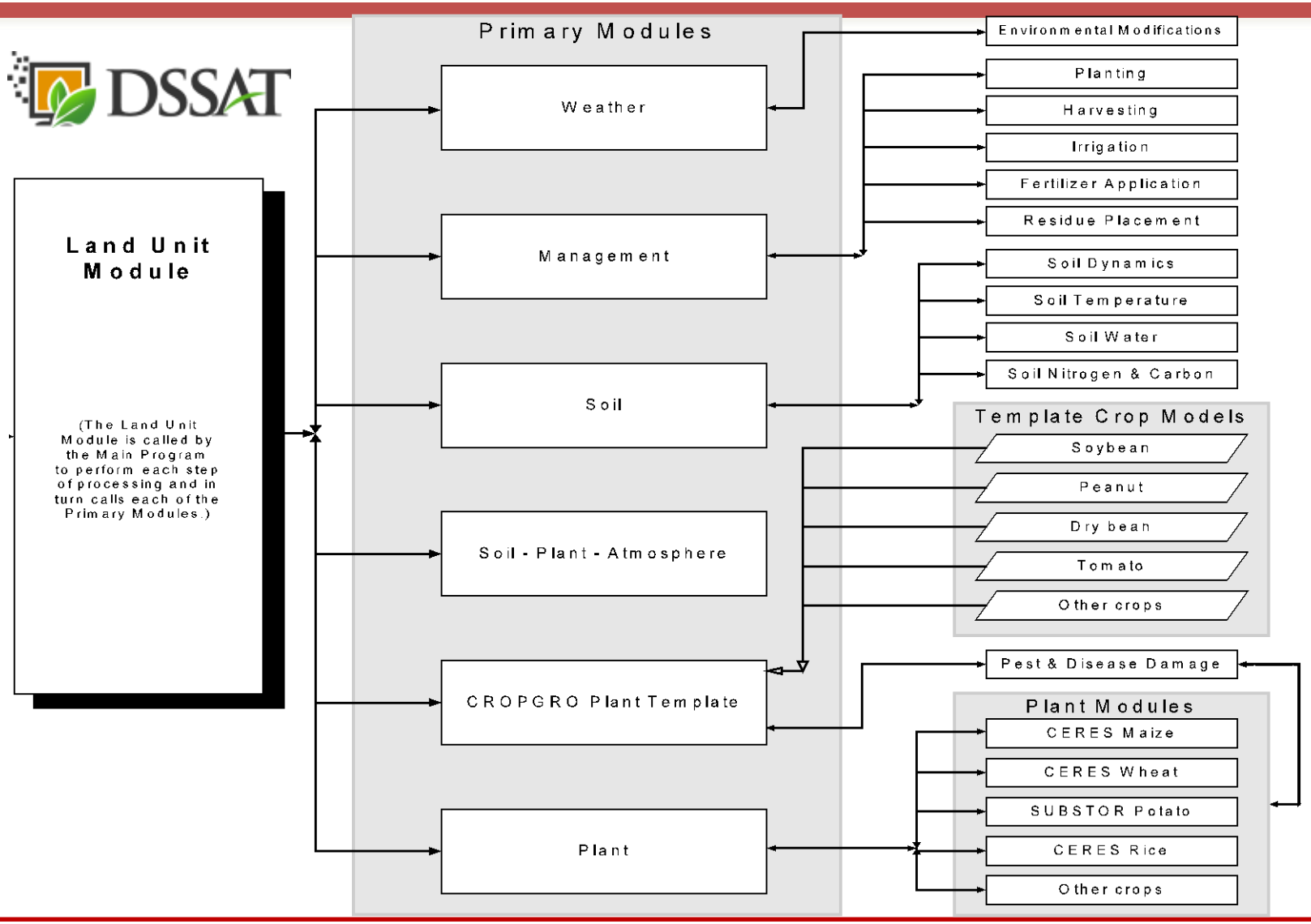
Core Building Blocks



New Project: MINT (Model INTegration)



Agriculture Models: Representative Model

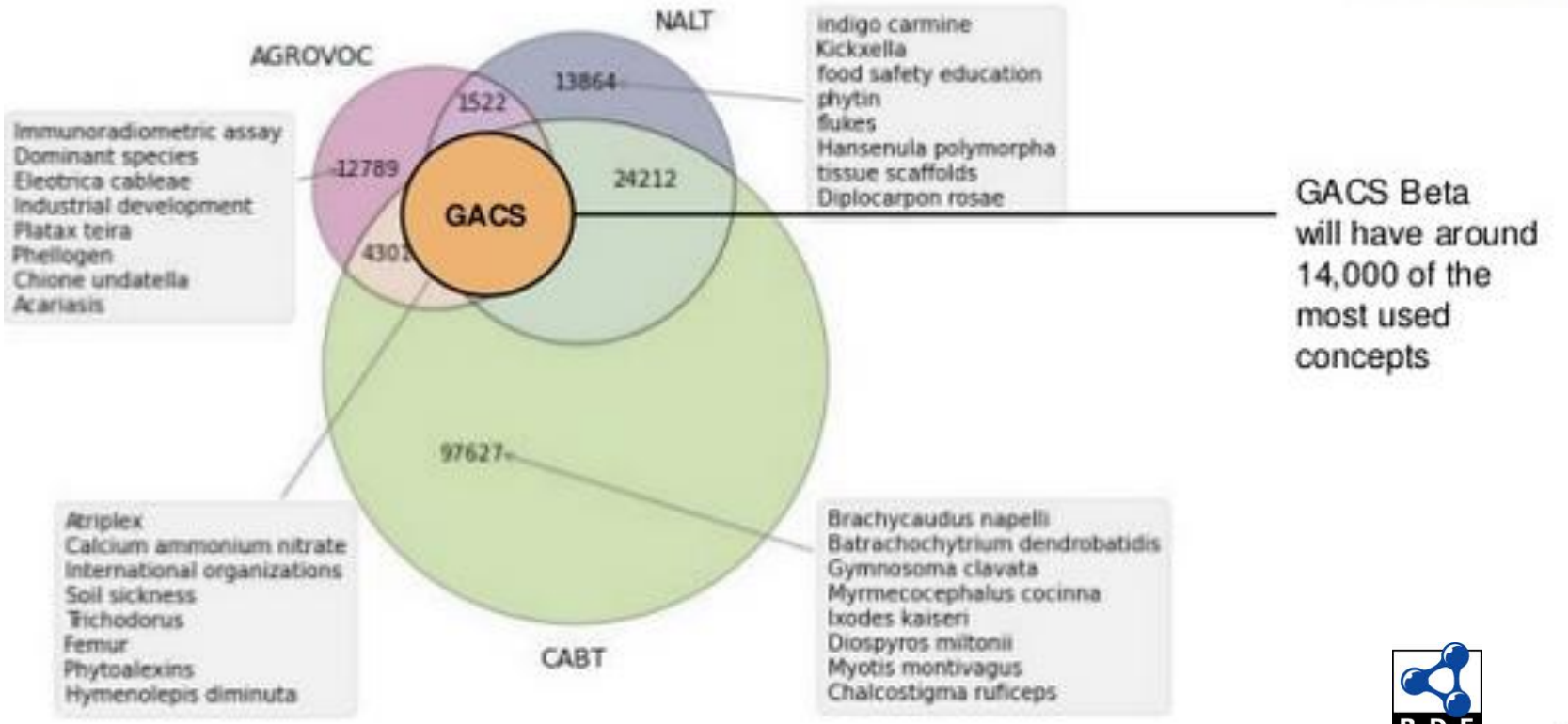


Agriculture Models: Ontologies of Model Variables

AGROVOC
Multilingual Agricultural Thesaurus
Food and Agriculture Organization of the United Nations

USDA United States Department of Agriculture
National Agricultural Library
Thesaurus (NALT)

www.cabi.org



GACS Beta will have around 14,000 of the most used concepts

RDF

http://www.slideshare.net/CIARD_/gacs-for-rdapresentedbycynthiaparrmarch2015

Social Models: Representative Models

GAMA PLATFORM