Knowledge-Powered Data Science for Integrated Modeling in Geosciences

Yolanda Gil

Information Sciences Institute
Department of Computer Science
Spatial Sciences Institute

University of Southern California

http://www.isi.edu/~gil
@yolandagil
gil@isi.edu

The Future is Here: Knowledge-Powered Data Science

- Focus shifted from data to **models**
  - Model characterization, reuse, and integration
- Need to **incorporate model-centered science knowledge** about phenomena and context
  - Knowledge about **physical, geological, chemical, biological, ecological, and anthropomorphic factors**
  - Knowledge about the user **goals and context**
- This would enable novel forms of reasoning, integrating, visualizing, managing, learning, and discovery with geosciences data

REVIEW ARTICLES

Intelligent Systems for Geosciences: An Essential Research Agenda

By Yolanda Gil, Suzanne A. Pierce, Hassan Babaie, Arindam Banerjee, Kirk Borne, Gary Bust, Michelle Cheatham, Imme Ebert-phoff, Carla Gomes, Mary Hill, John Horel, Leslie Hsu, Jim Kinter, Craig Knoblock, David Krum, Vipin Kumar, Pierre Lermusiaux, Yan Liu, Chris North, Victor Pankratius, Shanan Peters, Beth Plale, Allen Pope, Sai Ravela, Juan Restrepo, Aaron Ridley, Hanan Samet, Shashi Shekhar

Communications of the ACM, January 2019, Vol. 62 No. 1, Pages 76-84
10.1145/3192335

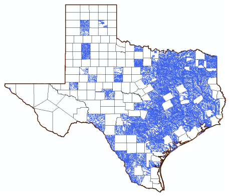
Comments

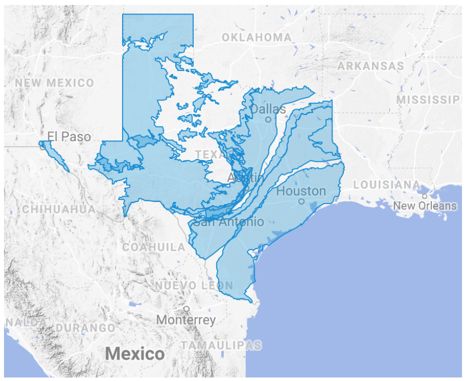
Many aspects of geosciences pose no
systems research. Geoscience data is
to be uncertain, intermittent, sparse
scale. Geosciences processes and obj
spatiotemporal boundaries. The lack
model evaluation, testing, and comp
these challenges requires breakthrou
transform intelligent systems, while
geosciences in turn. Although there l
beneficial interactions between the i
geosciences communities,[4,12] the po
research in intelligent systems for ge

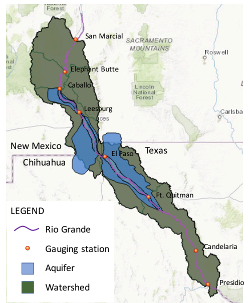
Outline

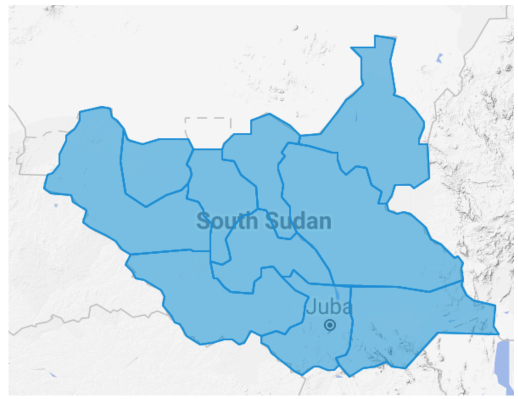
- The need for integrated modeling in geosciences
- Diversity of models across disciplines
- MINT: knowledge-powered data science for integrated modeling
- Intelligent systems for geosciences

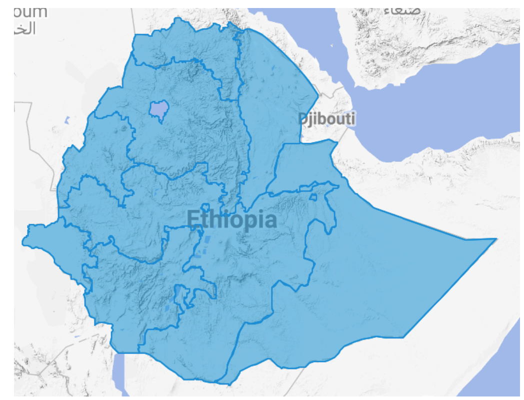
Understanding Natural-Human System Interactions: Water Use, Land Cover Changes, Food Insecurity,…

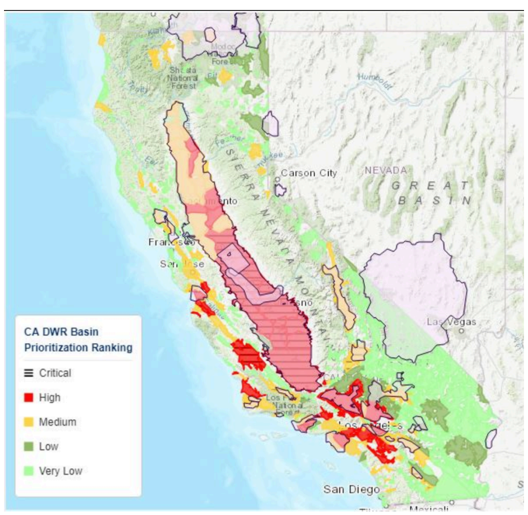




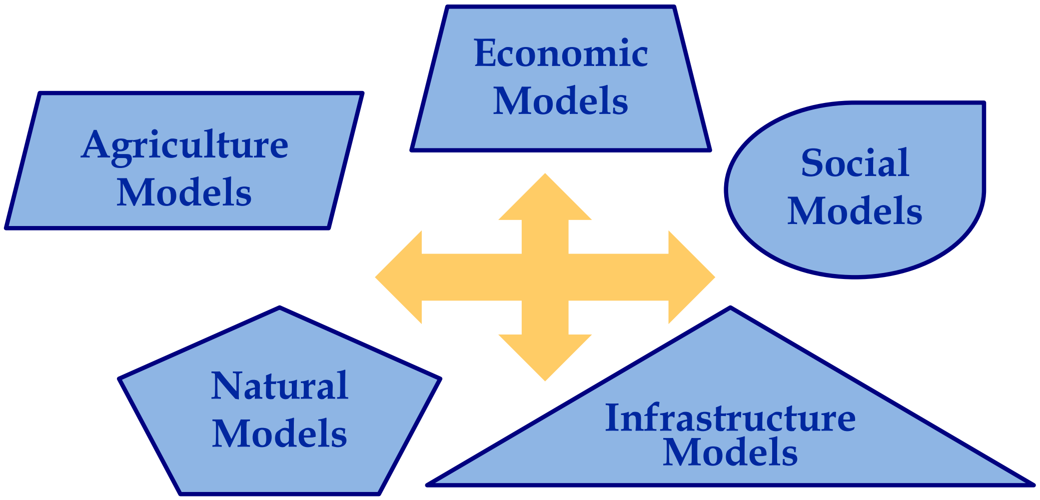








A Grand Challenge for Geospatial Data Science



Integrated Modeling

- Increased demand
- Months or years
- Currently a craft

- Integrating pairs of natural models is very hard
  - Eg, surface water models + ground water models
- Integrating natural and human models is even harder
  - Eg, agriculture + socioeconomic models of human behavior

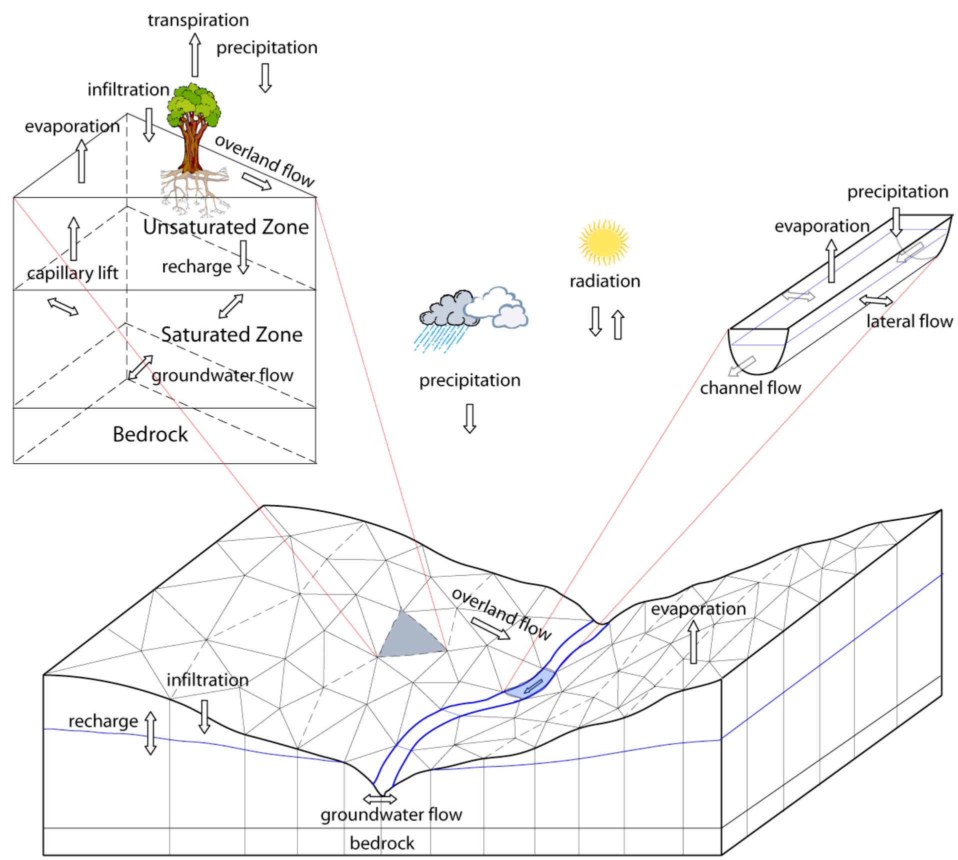
Planet Texas 2050
Credit: Deana Pennington, Cybershare Project, UT El Paso

Outline

- The need for integrated modeling in geosciences
- Diversity of models across disciplines
- MINT: knowledge-powered data science for integrated modeling
- Intelligent systems for geosciences

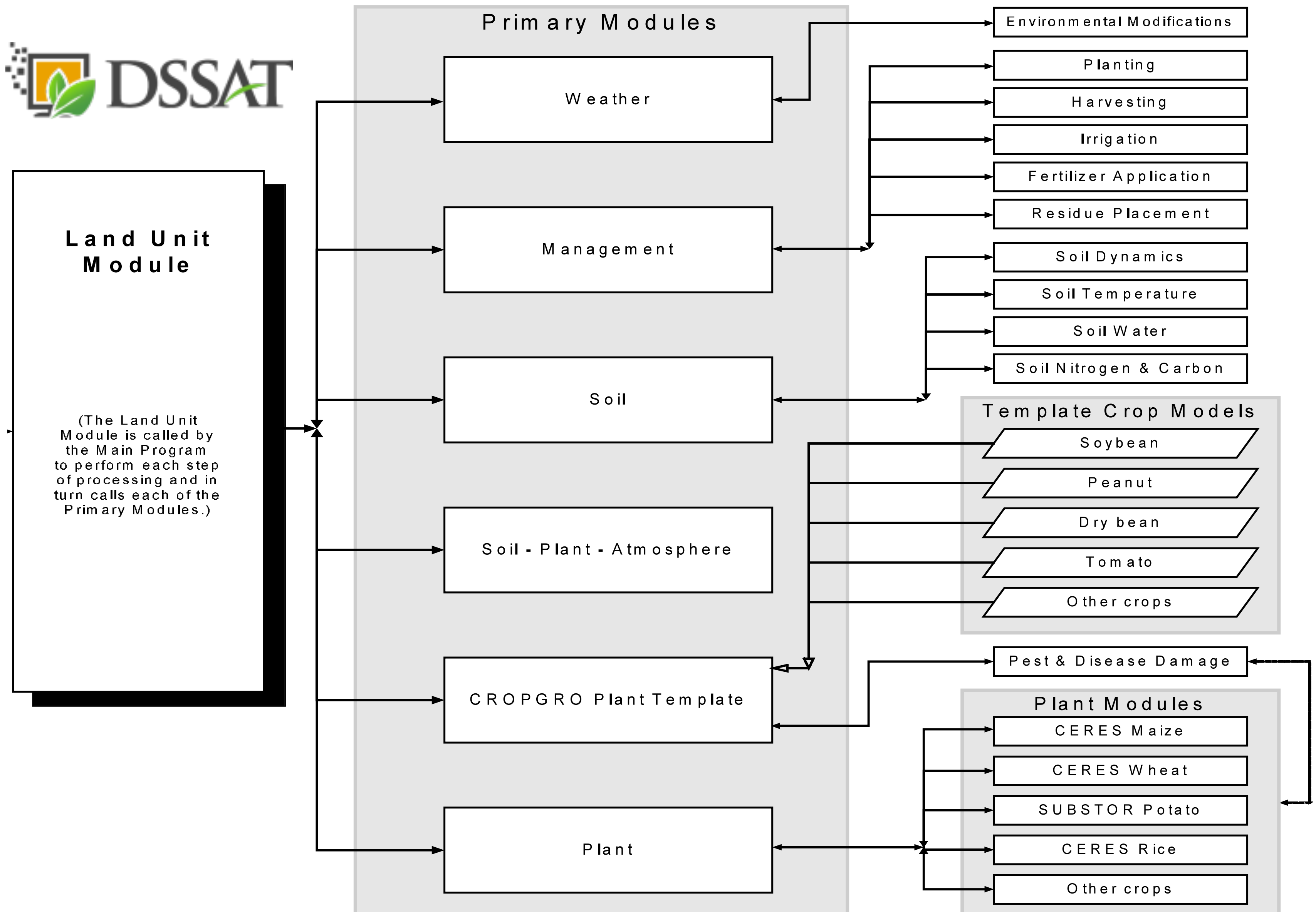
A Representative Hydrology Model

- PIHM hydrology model
  - Needs data on *essential terrestrial variables*: slope, vegetation, etc.
  - Generates discharge, flooding



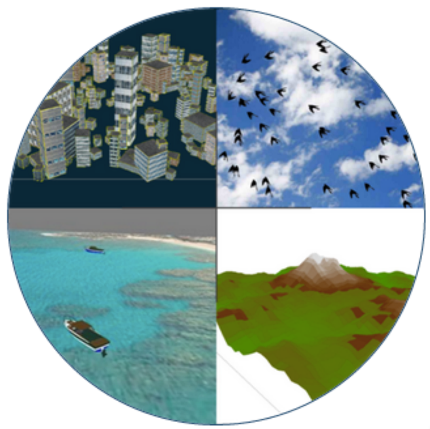
Process	Governing equation/model	Original governing equations	Semi-discrete form ODEs
Channel Routing	St. Venant Equation	$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = q$	$\left(\frac{d\zeta}{dt} = P_c - \sum Q_{gc} + \sum Q_{sc} + Q_{in} - Q_{out} - E_c\right)_i$
Overland Flow	St. Venant Equation	$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} + \frac{\partial (vh)}{\partial y} = q$	$\left(\frac{\partial h}{\partial t} = P_e - I - E_e - Q_{oc} + \sum_{j=1}^{3} Q_s^{ij}\right)_i$
Unsaturated Flow	Richard Equation	$C(\psi)\frac{\partial \psi}{\partial t} = \nabla \cdot (K(\psi)\nabla(\psi + Z))$	$\left(\frac{d\xi}{dt} = I - q^0 - ET_s\right)_i$
Groundwater Flow	Richard Equation	$C(\psi)\frac{\partial \psi}{\partial t} = \nabla \cdot (K(\psi)\nabla(\psi + Z))$	$\left(\frac{d\zeta}{dt} = q^0 + \sum_{j=1}^{3} Q_g^{ij} - Q_l + Q_{gc}\right)_i$
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	$\frac{dS_{snow}}{dt} = P - E_{snow} - \Delta w$	$\left(\frac{dS_{snow}}{dt} = P - E_{snow} - \Delta w\right)_i$
Evapotranspiration	Penman-Monteith Method	$ET_0 = \frac{\Delta(R_n - G) + \rho_a C_p \frac{(e_s - e_a)}{r_a}}{\Delta + \gamma(1 + \frac{r_s}{r_a})}$	$\left(ET_0 = \frac{\Delta(R_n - G) + \rho_a C_p \frac{(e_s - e_a)}{r_a}}{\Delta + \gamma(1 + \frac{r_s}{r_a})}\right)_i$

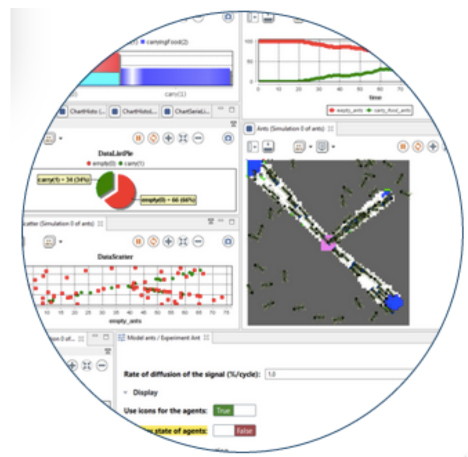
A Representative Agriculture Model

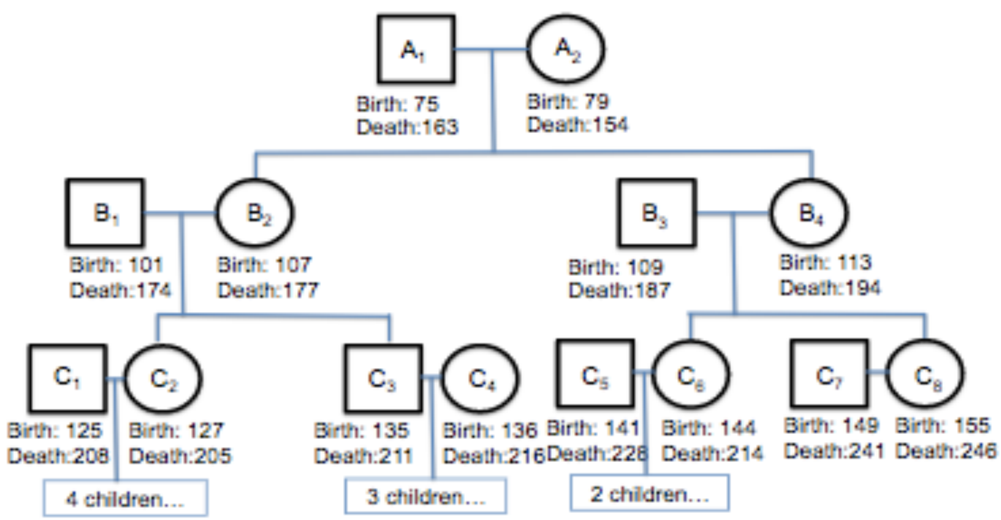


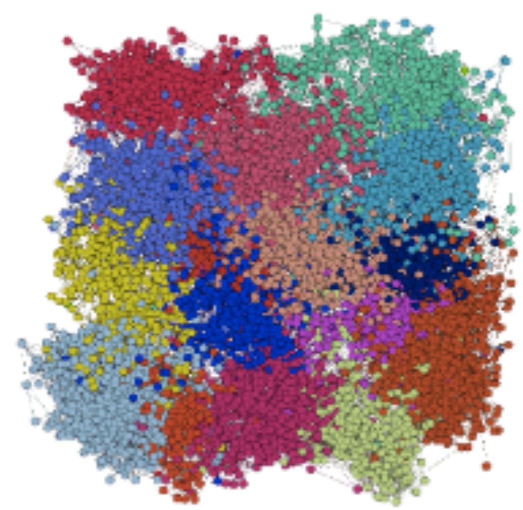
A Representative Social Model

GAMA PLATFORM









Modeling in Different Disciplines: Diversity of Approaches

Natural Models

Agriculture Models

- **Modeling methodology**
  - Empirical (from prior data)
  - Mechanistic (first principles)
  - Mixed
- **Representative models**
  - Many dimensions of diversity
- **Model variables**
  - Some standards in some domains
- **Data requirements and availability**
- **Integration approaches**
  - Very diverse

Economic Models

Social Models

Integrated Modeling: Bridging Across Disciplines

Diversity across disciplines:
- Modeling approaches
- Representative models
- Model variables
- Integration frameworks

Natural Models | Agriculture Models | Economic Models | Social Models | Infrastructure Models

Mediation at many levels

- Problem framing
- Model selection
- Variable mapping
- Data ingestion
- Runtime coordination

Model Coupling

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

How Can We Facilitate Model Integration?

Problem framing	**Very diverse scope**	*Manual definition*
Model selection	**Very diverse approaches**	*Manual selection*
Variable mapping	**Very diverse variables**	*Manual mapping*
Data ingestion	**Very diverse data needs**	*Disconnected tools*

Outline

- The need for integrated modeling in geosciences
- Diversity of models across disciplines
- MINT: knowledge-powered data science for integrated modeling
- Intelligent systems for geosciences

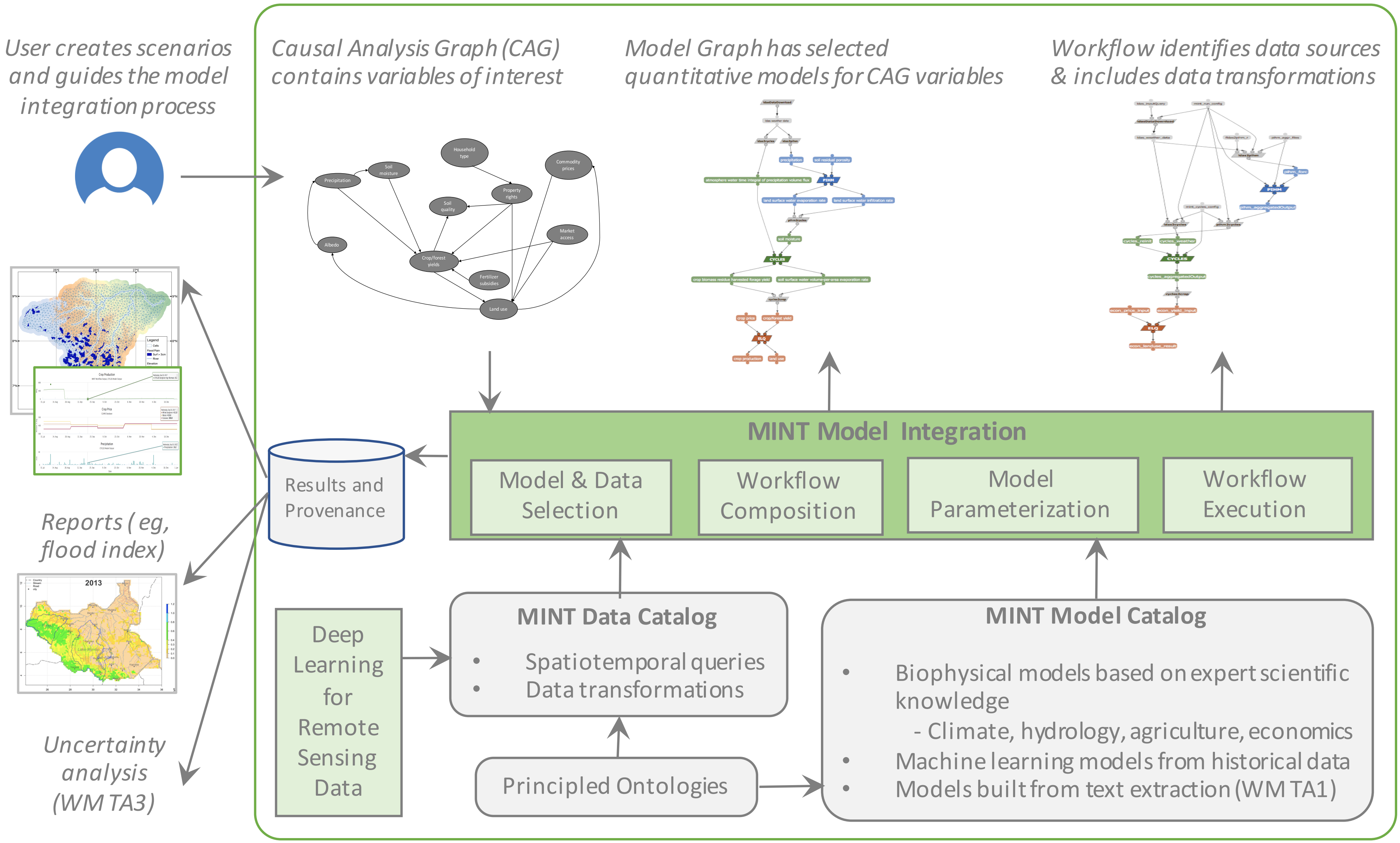
Knowledge-Powered Data Science for Integrated Modeling

Problem framing	Structured frameworks for scenario scoping
Model selection	Semantic descriptions of models and assumptions
Variable mapping	Ontologies of variables and relations
Data ingestion	Knowledge-guided information extraction and integration

MINT: Model INTegration [Gil et al iEMSs'18]

http://mint-project.info

Collaboration with Daniel Garijo, Deborah Khider, Craig Knoblock, Ewa Deelman, Rafael Ferreira (USC/ISI), Vipin Kumar (UM), Scott Peckham (CU), Chris Duffy & Armen Kemanian (PSU), Kelly Cobourn (VT), Suzanne Pierce (UT)



Knowledge-Powered Data Science for Integrated Modeling in MINT

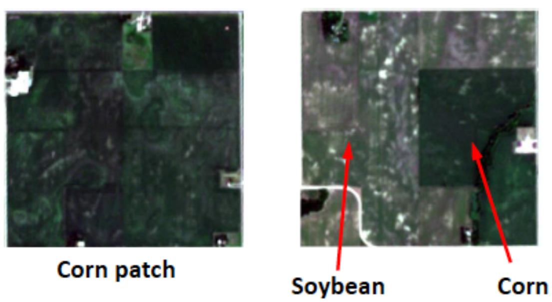
Problem framing	Structured frameworks for scenario scoping
Model selection	Semantic descriptions of models and assumptions
Variable mapping	Ontologies of variables and relations
Data ingestion	Knowledge-guided information extraction and integration

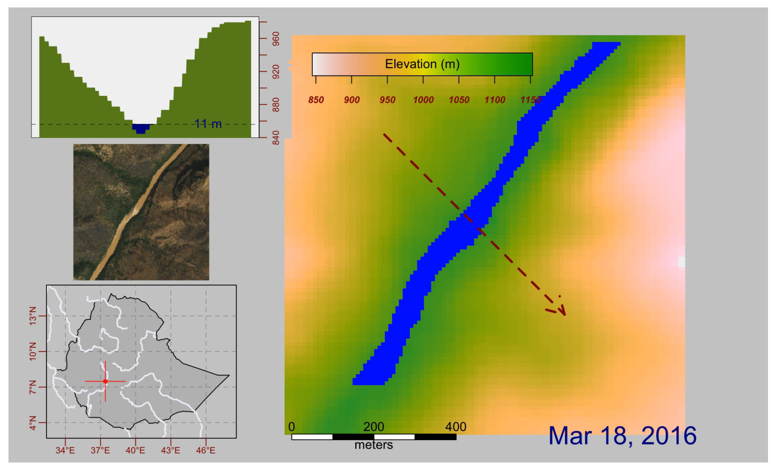
Knowledge-Powered Data Science for Integrated Modeling in MINT

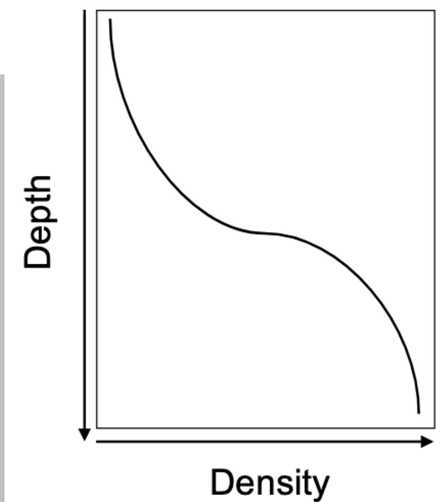
Problem framing	Structured frameworks for scenario scoping
Model selection	Semantic descriptions of models and assumptions
Variable mapping	Ontologies of variables and relations
Data ingestion	Knowledge-guided information extraction and integration

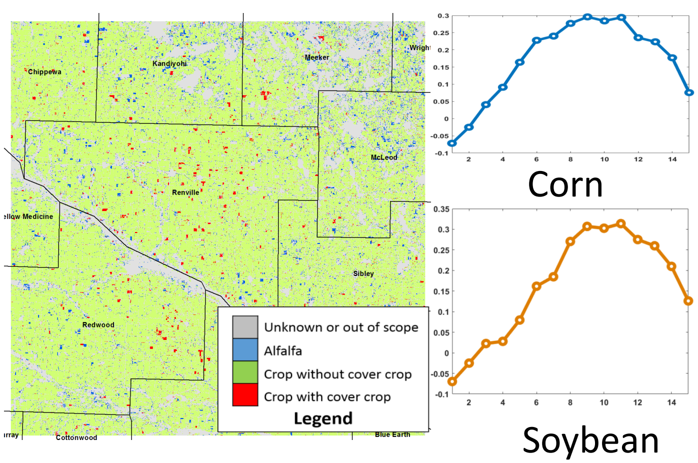
TGDS: Theory-Guided Data Science [Kumar et al 2017]

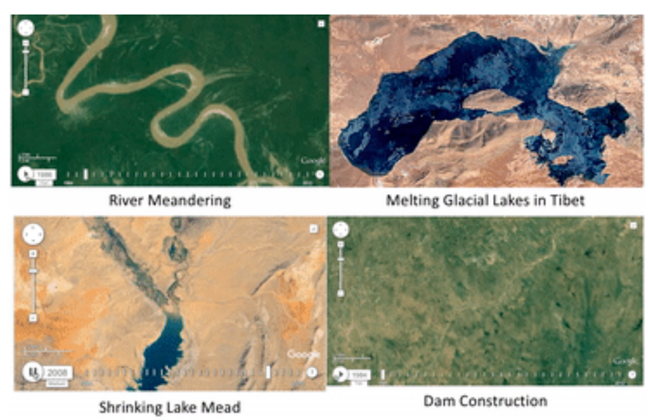
- Machine learning to generate physically consistent models
- Physics laws guide deep learning

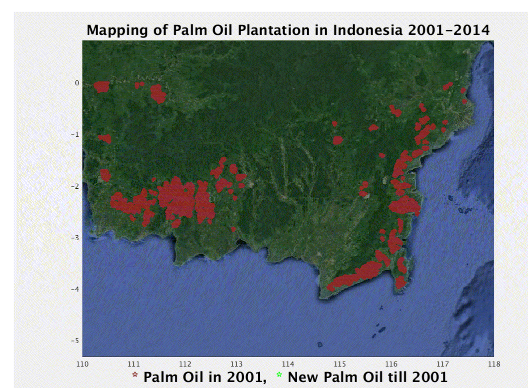




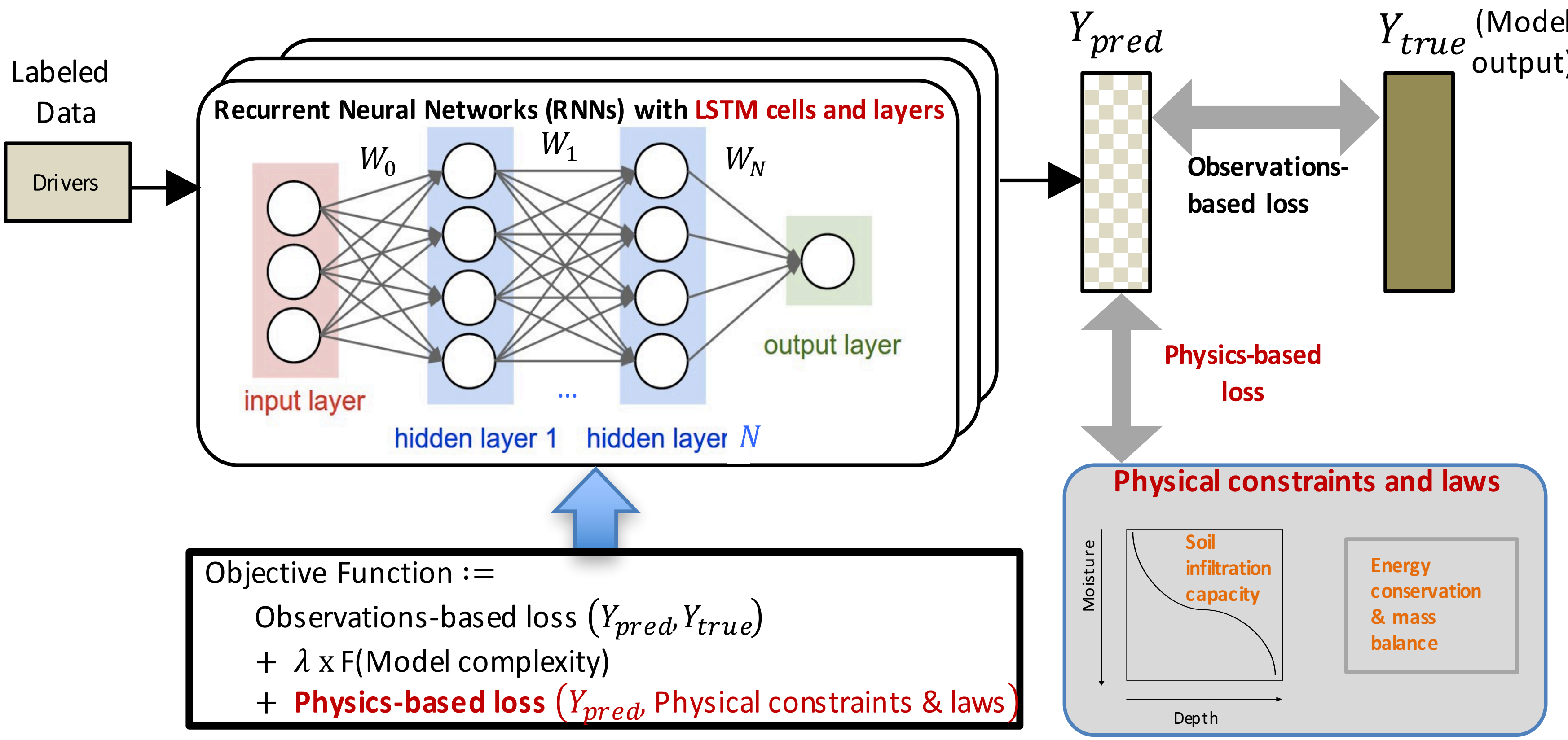


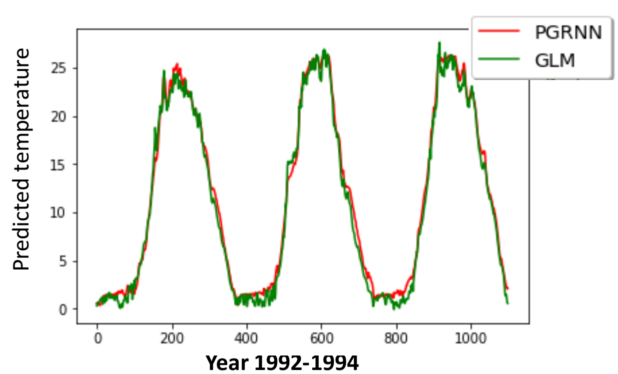






Physics-Guided Neural Networks (PGNNs) [Karpatne et al 2017]





Karpatne et al., "Physics-guided neural networks (PGNN): An Application in Lake Temperature Modeling," arXiv: 1710.11431, 2017.

Automated Data Transformations [Knoblock et al 2018]

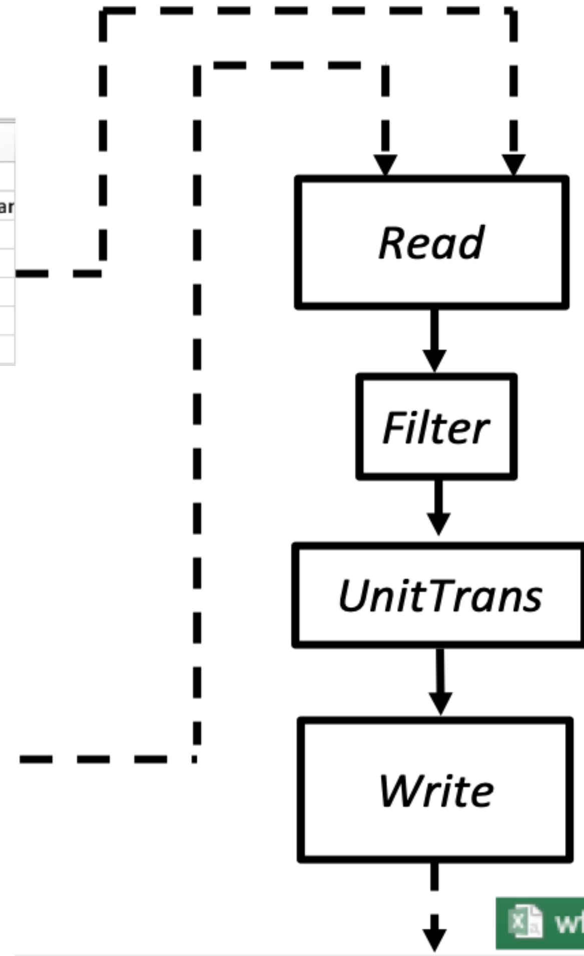
wfp_food_prices_ethiopia

	A	B	C	D	E	F	G
1	date	cmname	unit	category	price	currency	country
2	#date	#item+name	#item+unit	#item+type	#value	#currency	#country+nar
3	7/15/05	Sorghum - Wł	100 KG	cereals and tu	238	ETB	Ethiopia
4	8/15/05	Sorghum - Wł	100 KG	cereals and tu	250	ETB	Ethiopia
5	9/15/05	Sorghum - Wł	100 KG	cereals and tu	248	ETB	Ethiopia
6	10/15/05	Sorghum - Wł	100 KG	cereals and tu	233	ETB	Ethiopia
7	11/15/05	Sorghum - Wł	100 KG	cereals and tu	252	ETB	Ethiopia

wfp_food_prices

D-REPR yml file

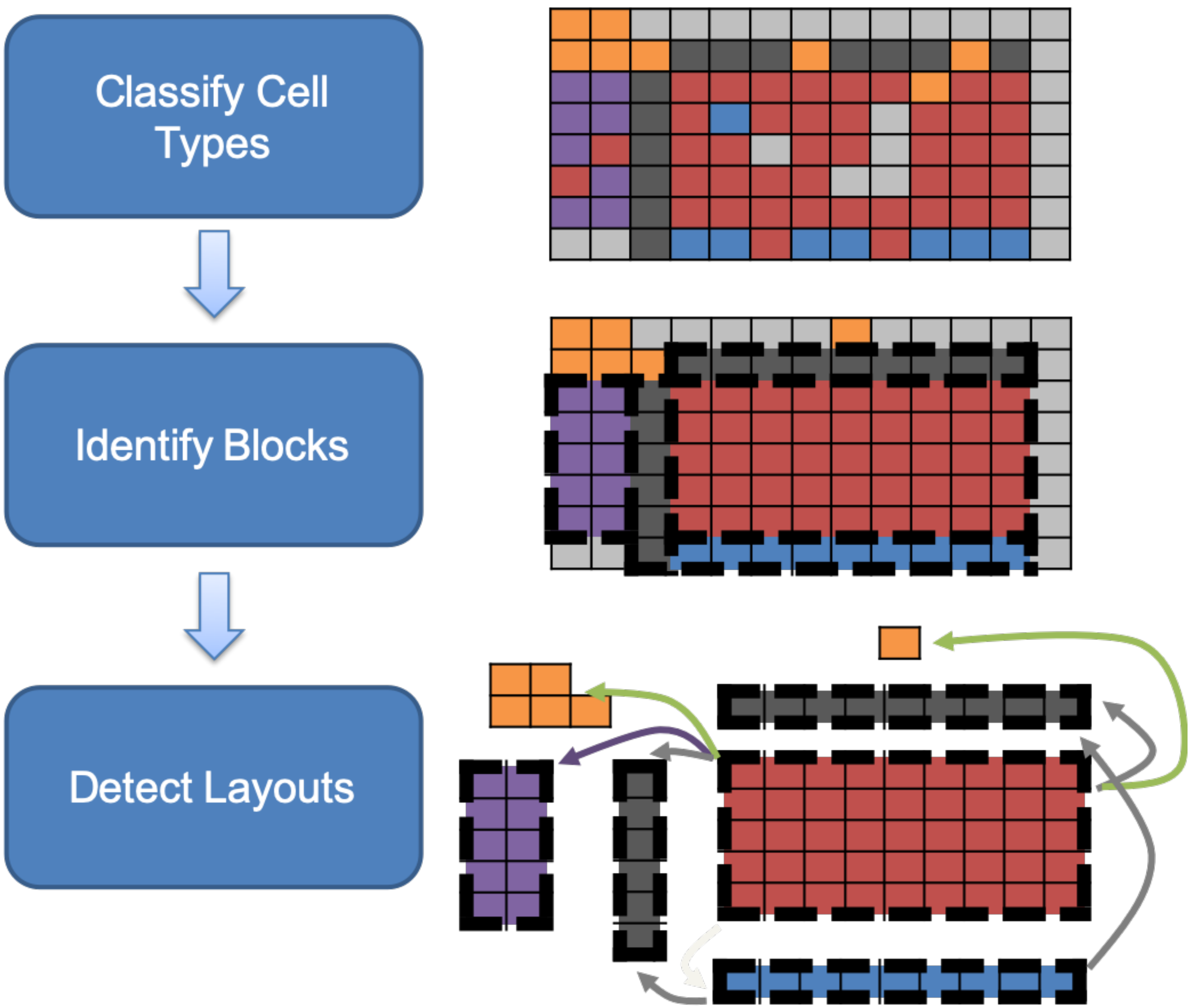
```yaml
version: "1"
resources:
  default:
    type: csv
    delimiter: \t
variables:
  date: [2.., 0]
  crop: [2.., 1]
alignments:
  - type: dimension
    value: date:0 <-> crop:0
semantic_model:
  data_nodes:
    date: qb:Observation:1--sdmx-
    crop: qb:Observation:1--dcat-
    grain_yield: qb:Observation:1
```



wfp_food_prices_ethiopia.modified

	A	B	C	D	E
1	sdmx-dimen:	sdmx-attribute:refArea	sdmx-at	dcat-dimension:thing	dcat:measure_1_value
2	7/15/05	Ethiopia - Addis Ababa	$/kg	Sorghum - Wholesale	0.082134514
3	8/15/05	Ethiopia - Addis Ababa	$/kg	Sorghum - Wholesale	0.08627575
4	9/15/05	Ethiopia - Addis Ababa	$/kg	Sorghum - Wholesale	0.085585544
5	10/15/05	Ethiopia - Addis Ababa	$/kg	Sorghum - Wholesale	0.080408999
6	11/15/05	Ethiopia - Addis Ababa	$/kg	Sorghum - Wholesale	0.086965956
7	12/15/05	Ethiopia - Addis Ababa	$/kg	Sorghum - Wholesale	0.087311059

Table Understanding [Vu et al WWW'19]



Knowledge-Powered Data Science for Integrated Modeling in MINT

Problem framing — Structured frameworks for scenario scoping

Model selection — Semantic descriptions of models and assumptions

Variable mapping — Ontologies of variables and relations

Data ingestion — Knowledge-guided information extraction and integration

Ontologies for Modeling Variables [Peckham & Stoica 2019]

Object: Soil

Processes (related to verbs)**:**
Aeration, Bioturbation, Creep, Drainage, Erosion, Fertilization, Formation, Gelifluction, Movement, Slumping, Solifluction, Tilling, Weathering

Process Quantities:
creep_speed
erosion_rate
fertilization_time
tilling_depth

State Quantities:
age
alkalinity
brooks-corey_c_parameter
clay_volume_fraction
depth (to bedrock)
hydraulic_conductivity
organic_matter_volume_fraction
porosity
pressure_head
sand_volume_fraction
saturated_hydraulic_conductivity
water_volume_fraction

Properties:
color
fertility

- Property names are nominalizations of adjectives.
- Quantity is a subclass of Property. (Numerical, often has units.)
- A pairing of an object with a quantity is a **variable**.

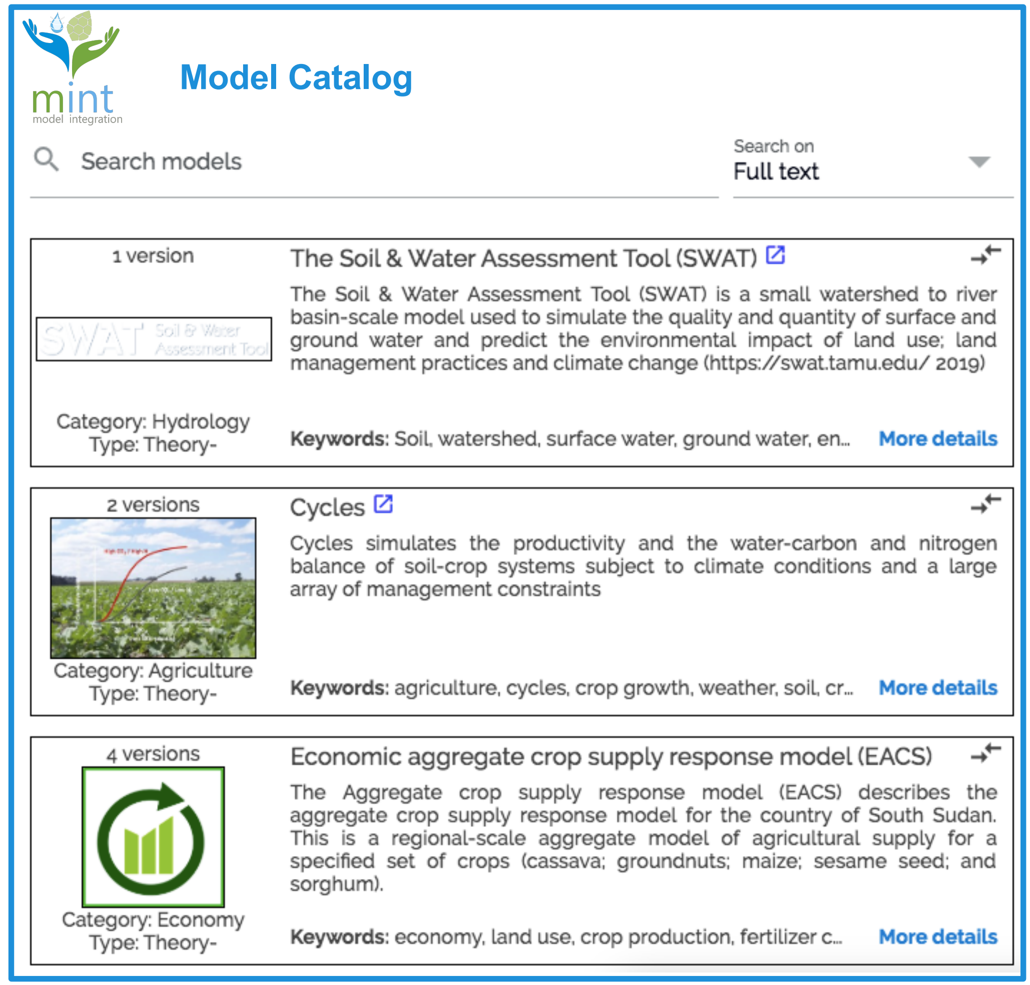
- An index or scale numerically quantifies a property (eg, drought index, flood index)
- Index is a subclass of Quantity

Knowledge-Powered Data Science for Integrated Modeling in MINT

Problem framing	Structured frameworks for scenario scoping
Model selection	Semantic descriptions of models and assumptions
Variable mapping	Ontologies of variables and relations
Data ingestion	Knowledge-guided information extraction and integration

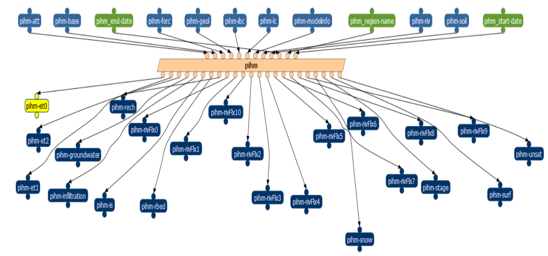
Knowledge-Rich Catalogs of Models [Garijo et al eScience 2019]

1. Model invocation
2. Data formats
3. Model variables
4. Constraints
5. Adjustable parameters
6. Interventions



1) Model Invocation

- Configuration: selected processes
- Setup: calibrated for specific areas



Cycles

Cycles simulates the productivity and the water,carbon and nitrogen balance of soil-crop systems subject to climate conditions and a large array of management constraints. Overall the model is set up to be daily. Some processes such as water balance are temporally nested (subdaily)

- **Authors:** Armen Kemanian
- **Publisher:** The Pennsylvania State University

Creation date: 2016
Category: Agriculture
Model type: Theory Guided

Select a configuration
Cycles configuration (v0.9.4) exposing weed fraction and fertilizer rate

Select a configuration setup
Cycles calibrated model (v0.9.4) for the Pongo region with planting dates. Weather file can be c

Overview | Parameters and Files | Variables | Assumptions | Compatible Software | Technical Information

2) Data Formats

Creation date: 2016
Category: Agriculture
Model type: Theory Guided

Cycles

Cycles simulates the productivity and the water,carbon and nitrogen balance of soil-crop systems subject to climate conditions and a large array of management constraints. Overall the model is set up to be daily. Some processes such as water balance are temporally nested (subdaily)

- **Authors:** Armen Kemanian
- **Publisher:** The Pennsylvania State University

Select a configuration
Cycles configuration (v0.9.4) exposing weed fraction and fertilizer rate

Select a configuration setup
Cycles calibrated model (v0.9.4) for the Pongo region with planting dates. Weather file can be c

Overview | Parameters and Files | Variables | Assumptions | Compatible Software | Technical Information

Output files

Name	Description
cycles_outputs	Cycles season configuration file
cycles_crop	Cycles crop output file
cycles_nitrogen	Nitrogen file. Results in this file are for the sum of all layers in the soil profile, i
cycles_water	Cycles water file
cycles_weatherOutput	Cycles weather output file
cycles_season	The season.dat file provides information about each crop harvest.
cycles_soilProfile	Results in this file are for the sum of all layers in the soil profile, including surf
cycles_som	Cycles annual SOM file. This file provides annualized measurements of the ca ratio by soil layer. One column will be created for each layer in the soil profile heading indicates LAYER 1..x
cycles_summary	The summary file provides a summarized output of total C inputs over the du annual rates for N cycling processes.

3) Model Variables

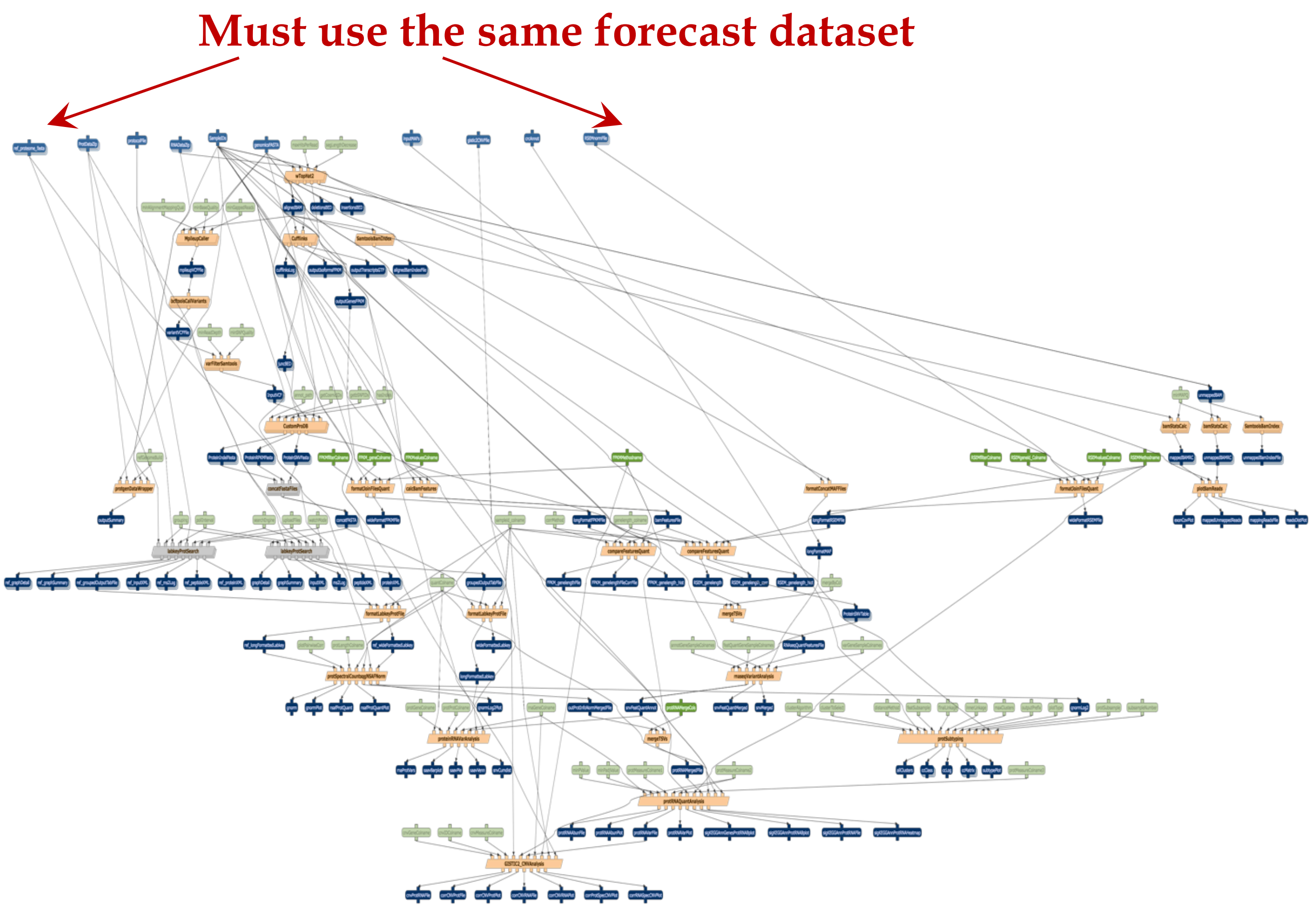
- Ontology of standard scientific names [Peckham iEMSs 2014; Peckham & Stoica 2019]
  - Eg SSN: `watershed_outlet_water__volume_outflow_rate` is more precise than "streamflow" or "discharge"

cycles_soil   Cycles soil description file. Soil files typically have a suffix of .soil, but any naming convention can be used as long as it matches the soil file na

Label	Long Name	Description	Standard Name	Units
RV	rock volume	Rock volume expressed as volume over total volume	soil_rock__volume_fraction	m3 m-3
DZ	soil layer thickness	Soil layer thickness	soil_layer__thickness	m
SLOPE	slope of the field	Average slope of field of interest	land_surface__slope	m m-1
LAYER	soil layer number	Soil layer number from 1 to an integer that is user defined (or database defined)		
SOM	soil organic matter	Soil organic matter per unit of non-rock soil expressed as percentage over total mass	soil_matter~organic__mass_fraction	kg kg-1 x-100
SILT	silt percentage	Silt mass per unit of soil mass (no rocks) and expressed as a percentage	soil~no-rock_silt__mass_fraction	%
BD	bulk density	Soil mass dry and wihtout rock divided by the sampled volume	soil~no-rock~dry__mass-per-volume_density	Mg m-3
CLAY	clay percentage	clay particle size fraction size fraction of each soil layer in %.	soil_clay_particle__volume_fraction	%

4) Constraints and Preconditions

- Daily rainfall data vs monthly rainfall data



5) Drivers and Adjustable Parameters

- Drivers are typically input data files (eg weather forecast)
- Adjustable parameters are those useful to explore what-if scenarios

Parameter	Description	Relevant for intervention	Value on setup
`start_year`	**Year when the simulation started**		2000 (default)
`end_year`	**Year when the simulation ended**		2017 (default)
`crop_name`	**Name of the crop to run the simulation for. Accepted values are: Maize, Sorghum, Peanut**		Maize (default)
`start_planting_day` The range is from 1 to 365	**Day of the year for the start of the planting window**	Planting Windows	100 (default)
`end_planting_day` The range is from 1 to 365	**Day of the year for the end of the planting window**	Planting Windows	149 (default)
`fertilizer_rate` The range is from 0 to 1250	**Mass of nitrogen fertilizer added each year (kg/ha)**		0 (default)
`weed_fraction` The range is from 0 to 1	**Areal fraction of weed**	Weed Control	0 (default)
`use_forcing`	**Use forcing data from a hydrology model (when available)**		FALSE

6) Interventions

- Associated with specific input parameters

Parameters:

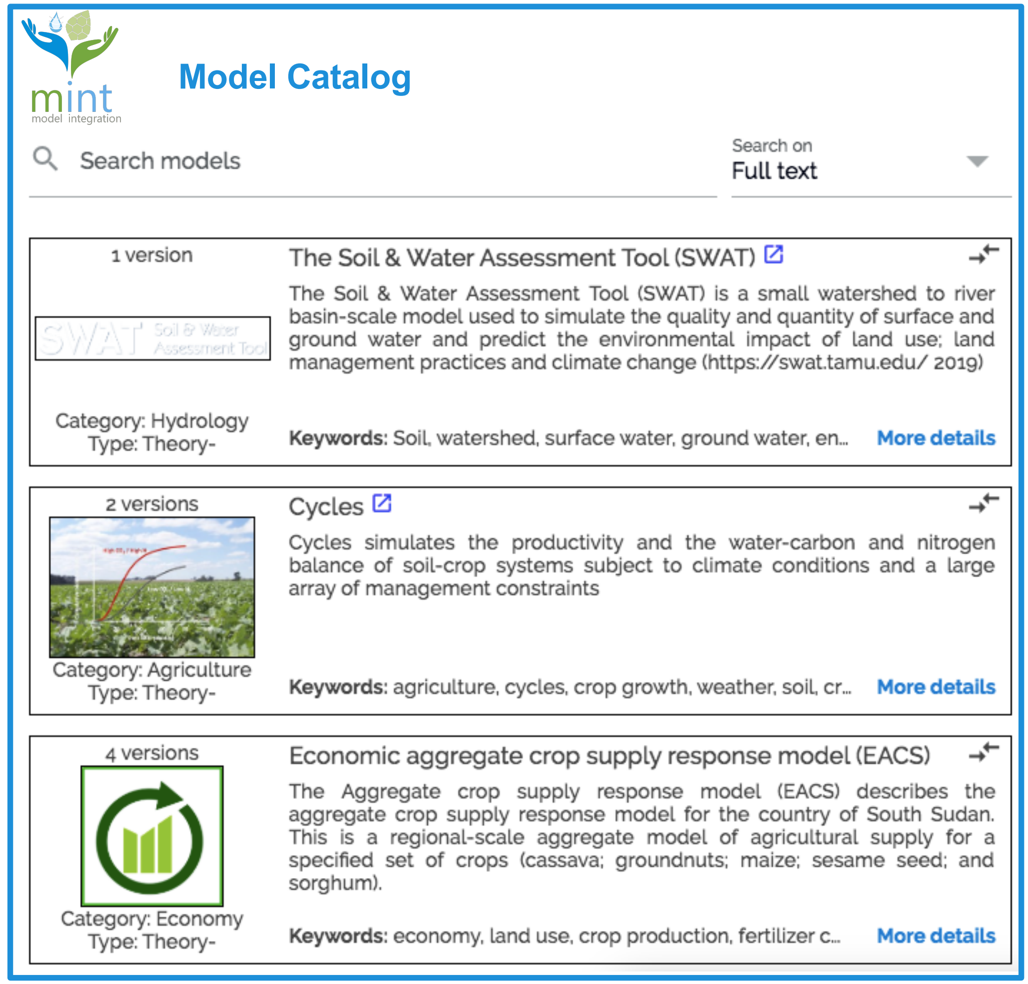
Parameter	Description	Relevant for intervention	Value on setup
`start_year`	**Year when the simulation started**		2000 (default)
`end_year`	**Year when the simulation ended**		2017 (default)
`crop_name`	**Name of the crop to run the simulation for. Accepted ... Peanut**		Maize (default)
`start_planting_day` The range is from 1 to 365	**Day of the year for the start of the planting window**	Planting Windows	100 (default)
`end_planting_day` The range is from 1 to 365	**Day of the year for the end of the planting window**	Planting Windows	149 (default)
`fertilizer_rate` The range is from 0 to 1250	**Mass of nitrogen fertilizer added each year (kg/ha)**		0 (default)
`weed_fraction` The range is from 0 to 1	**Areal fraction of weed**	Weed Control	0 (default)
`use_forcing`	**Use forcing data from a hydrology model (when available)**		FALSE

Interventions that force specific target planting windows can be expressed in this model as a start or end planting date

Knowledge-Rich Catalogs of Models [Garijo et al eScience 2019]

1. Model invocation
2. Data formats
3. Model variables
4. Constraints
5. Adjustable parameters
6. Interventions

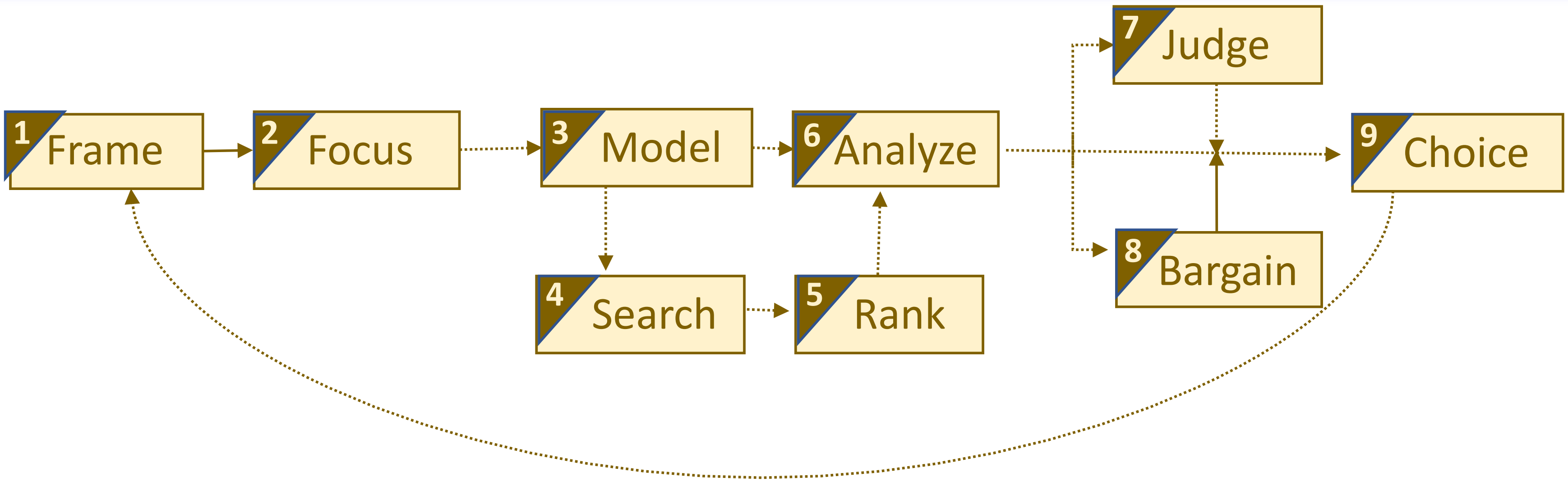
and much more!!



Knowledge-Powered Data Science for Integrated Modeling in MINT

Problem framing	Structured frameworks for scenario scoping
Model selection	Semantic descriptions of models and assumptions
Variable mapping	Ontologies of variables and relations
Data ingestion	Knowledge-guided information extraction and integration

Decision Making Process
[Pierce et al 2019]

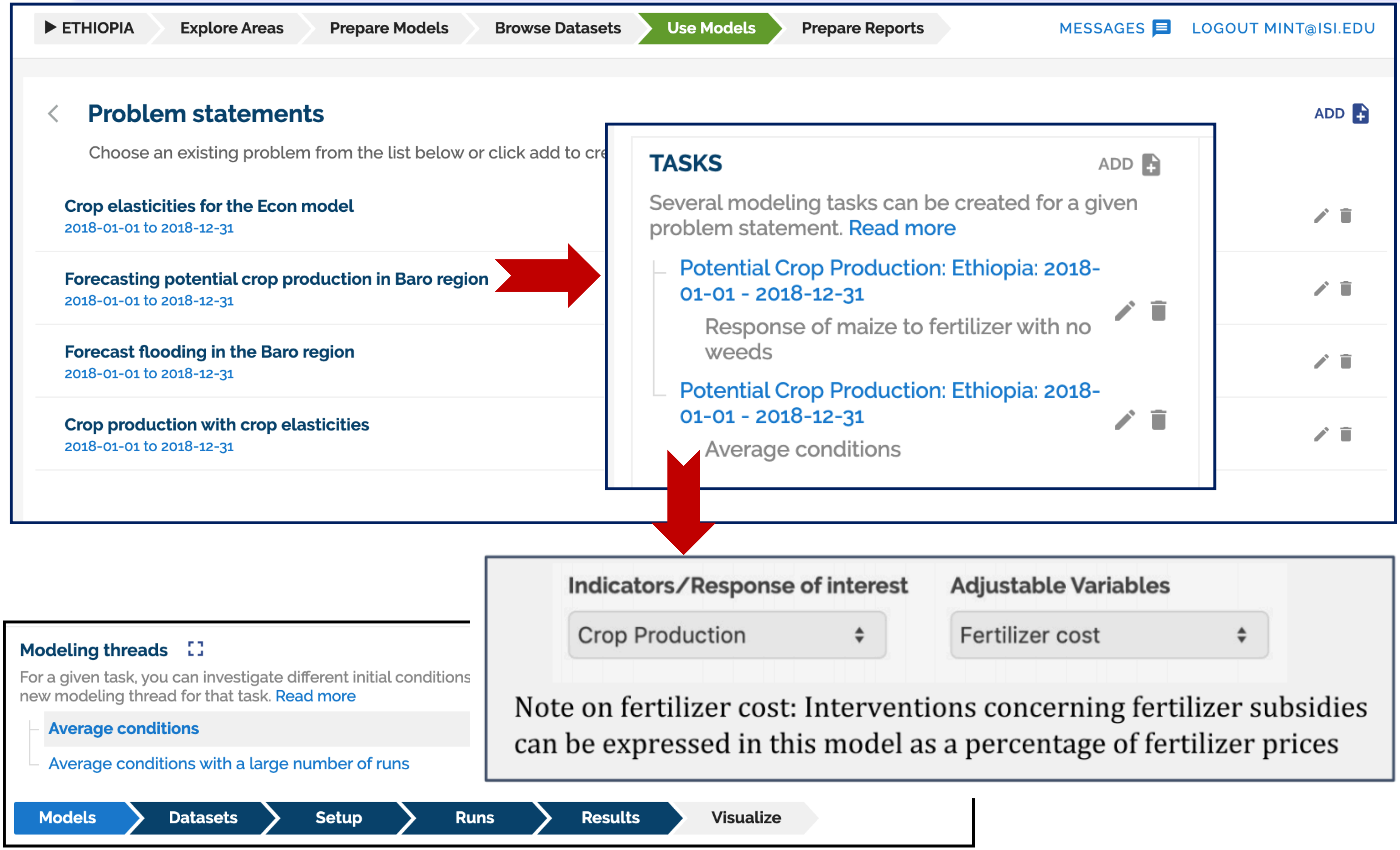


DECISION MAKING STEPS:

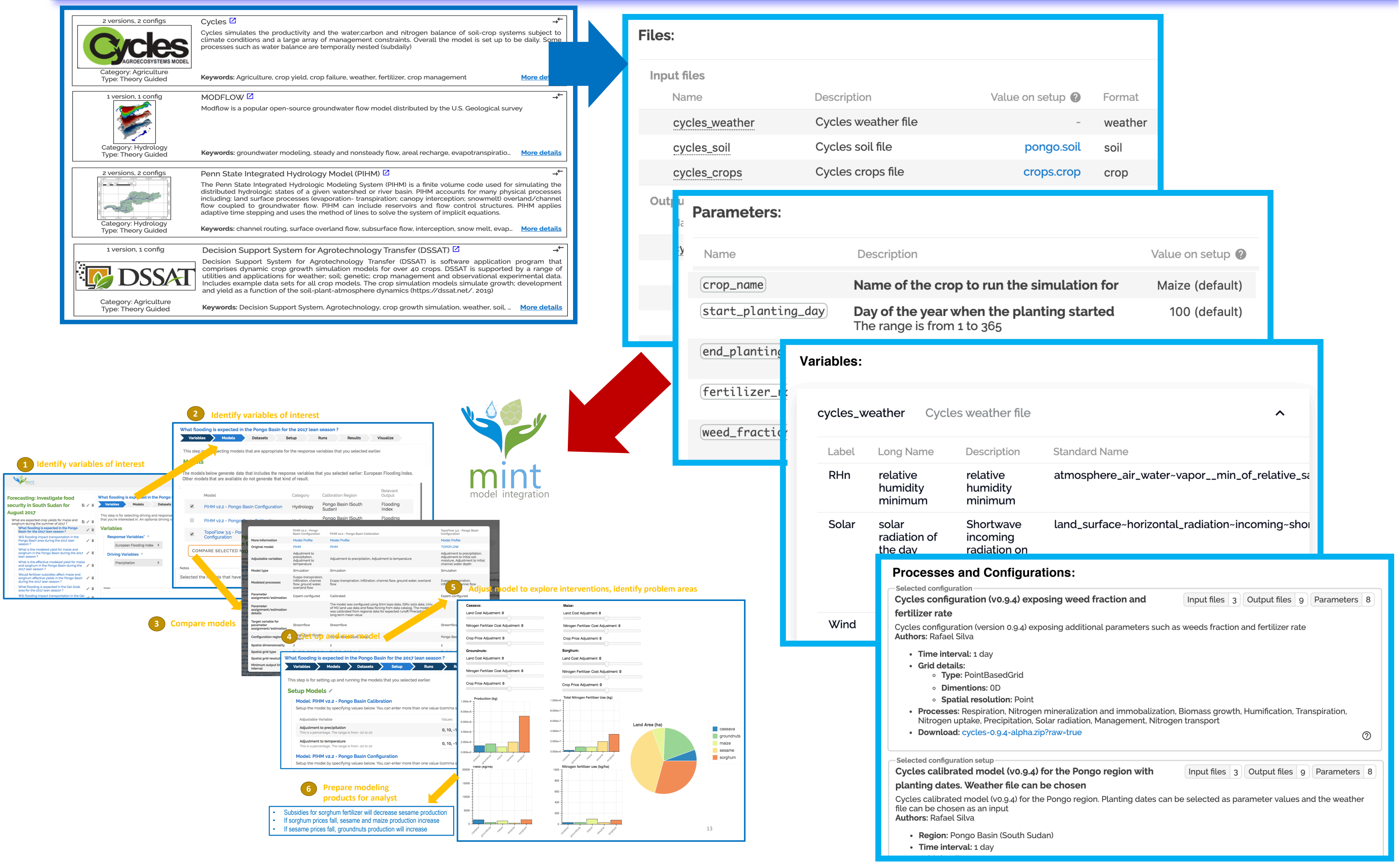
1. **Frame: Recognize and identify key issues**
2. **Focus: Define the context, key variables, reference behaviors, and decision problem**
3. **Model: Construct/Run Analysis Pipeline – run base case**
4. **Search: Explore solution space, consider interventions**
5. **Rank/Filter: Select interesting solutions**
6. **Analyze: Create summary of solutions explored, assess uncertainty**
7. **Judge: Adjudicate possible actions**
8. **Bargain: Consider tradeoffs**
9. **Choice: Make recommendation, take action, converge on decisions**

Iterate to refine/reframe/refocus modeling goals

Knowledge-Driven Problem Framing: Indicators and Adjustable Variables

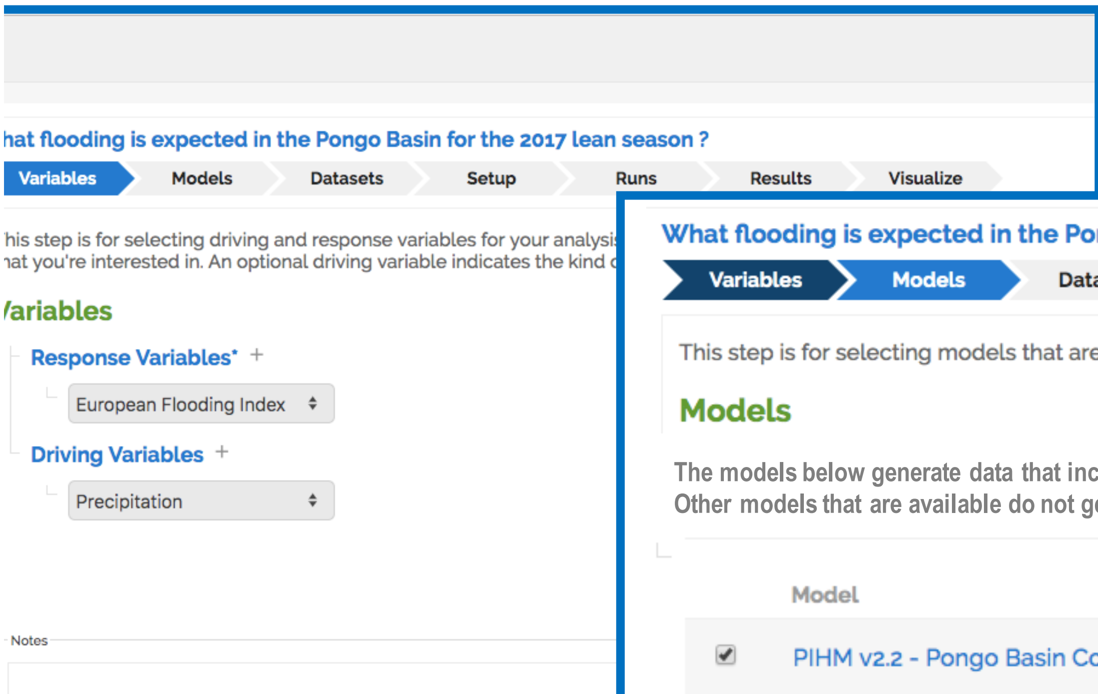


Forecasting and Interventions

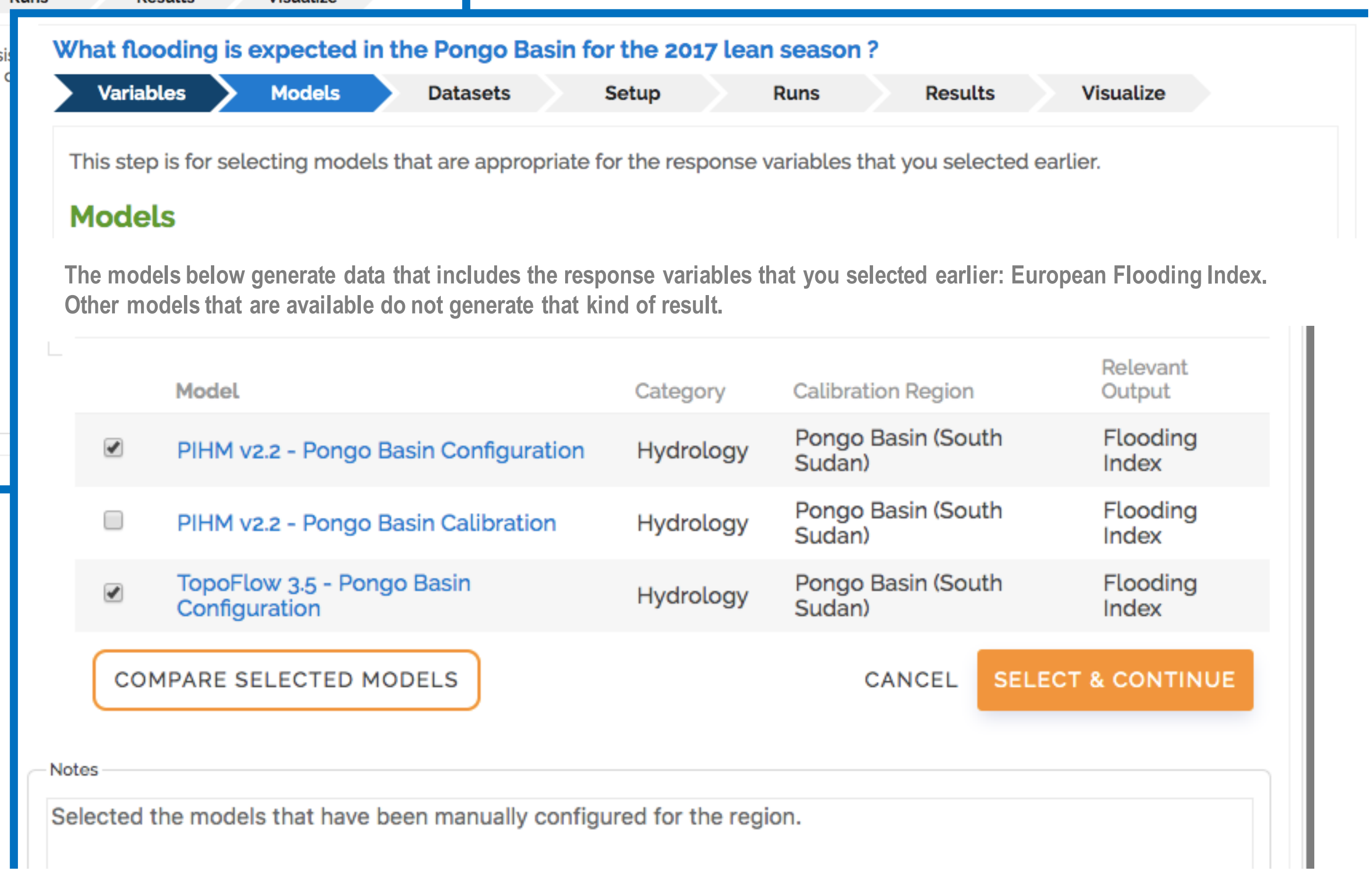


MINT User Interaction

❶ Identify results of interest

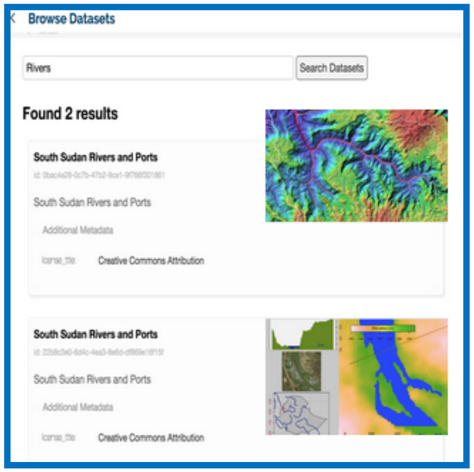


❷ Identify relevant models



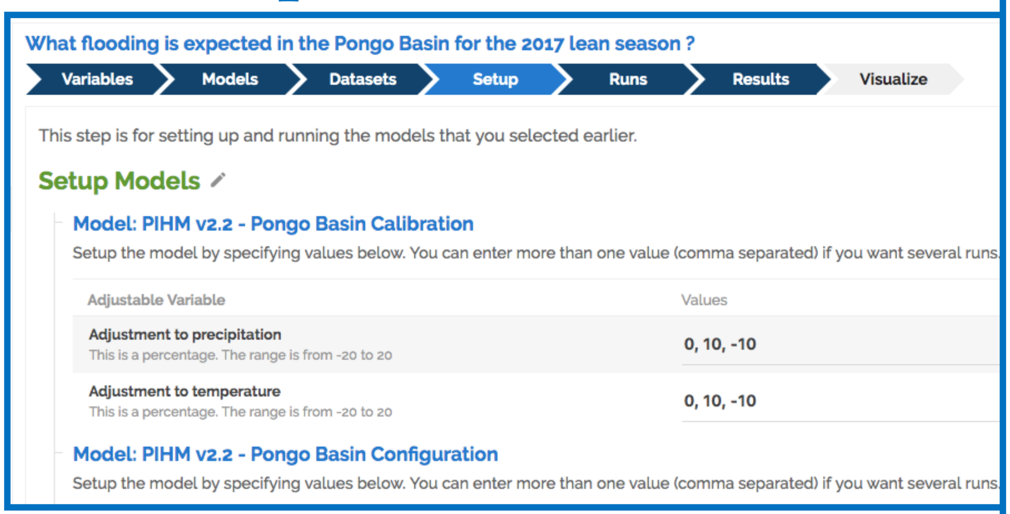
MINT User Interaction (Cont'd)

3 Find and transform datasets

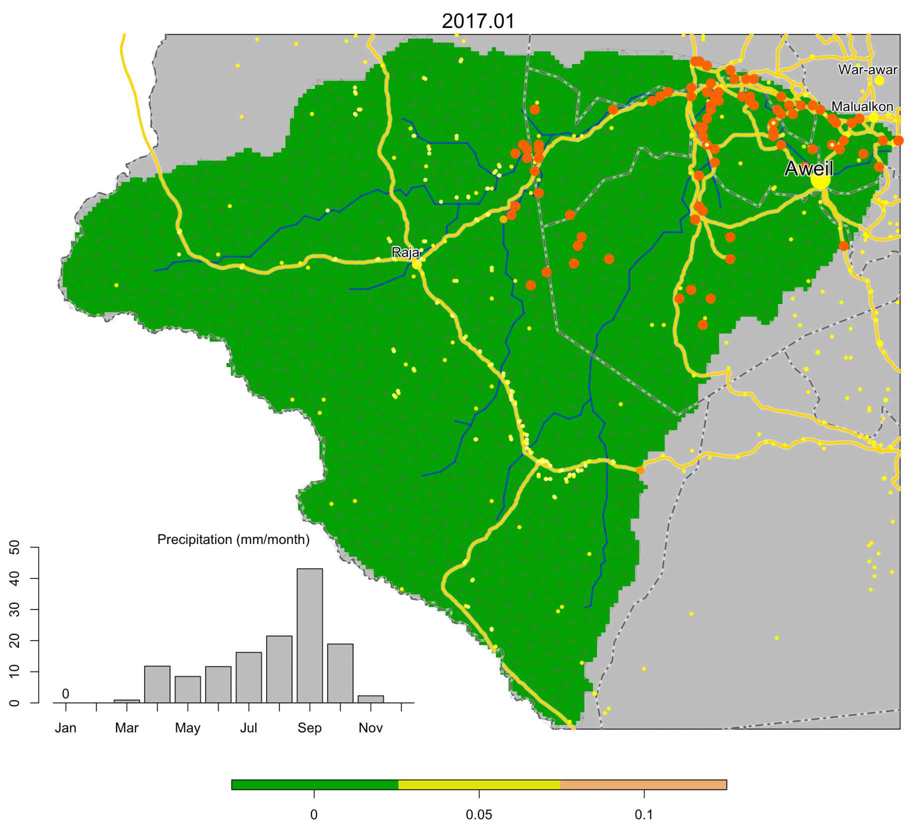


5 Adjust model to explore interventions, identify problem areas

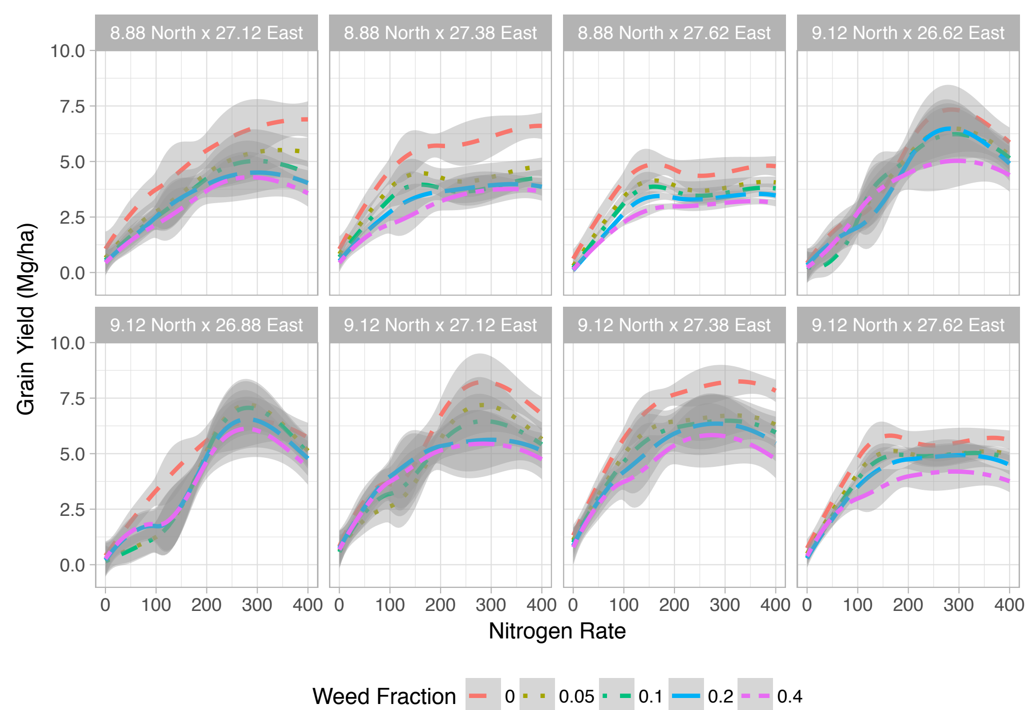
4 Set up and run models



Future Work in MINT



- Causal models
- Forecasting
- Interventions
- Uncertainty



Knowledge-Powered Data Science for Integrated Modeling in MINT

- **Knowledge-guided machine learning**
- **Knowledge-rich catalogs of models and data**
- **Knowledge-driven problem framing**

Problem framing	**Structured frameworks for scenario scoping**
Model selection	**Semantic descriptions of models and assumptions**
Variable mapping	**Ontologies of variables and relations**
Data ingestion	**Knowledge-guided information extraction and integration**

Outline

- The need for integrated modeling in geosciences
- Diversity of models across disciplines
- MINT: knowledge-powered data science for integrated modeling
- Intelligent systems for geosciences

REVIEW ARTICLES

Intelligent Systems for Geosciences: An Essential Research Agenda

By Yolanda Gil, Suzanne A. Pierce, Hassan Babaie, Arindam Banerjee, Kirk Borne, Gary Bust, Michelle Cheatham, Imme Ebert-phoff, Carla Gomes, Mary Hill, John Horel, Leslie Hsu, Jim Kinter, Craig Knoblock, David Krum, Vipin Kumar, Pierre Lermusiaux, Yan Liu, Chris North, Victor Pankratius, Shanan Peters, Beth Plale, Allen Pope, Sai Ravela, Juan Restrepo, Aaron Ridley, Hanan Samet, Shashi Shekhar

Communications of the ACM, January 2019, Vol. 62 No. 1, Pages 76-84
10.1145/3192335

Comments

VIEW AS: | SHARE:

Many aspects of geosciences pose no
systems research. Geoscience data is
to be uncertain, intermittent, sparse
scale. Geosciences processes and obj
spatiotemporal boundaries. The lack
model evaluation, testing, and comp
these challenges requires breakthrou
transform intelligent systems, while
geosciences in turn. Although there l
beneficial interactions between the i
geosciences communities,[4,12] the po
research in intelligent systems for ge

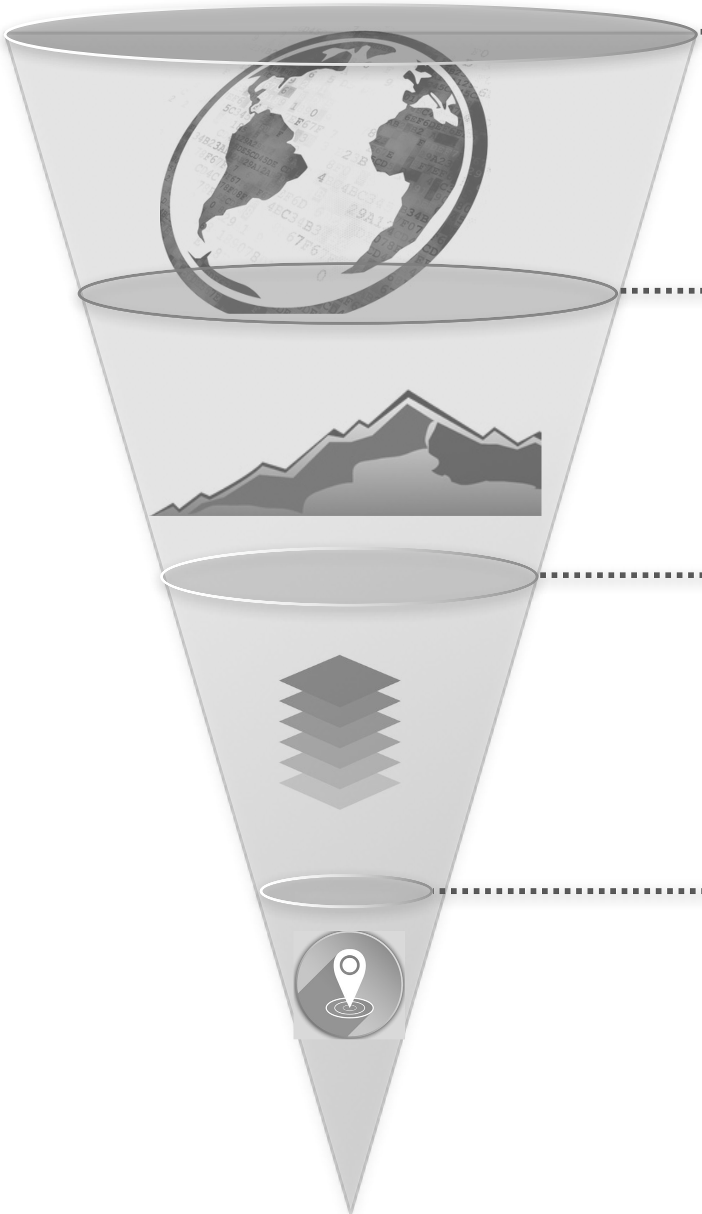
Report of the 2015 NSF Workshop on Intelligent Systems for Geosciences

March 26-27, 2015
Arlington VA

http://www.IS-GEO.org

Co-Chairs:
Yolanda Gil (gil@isi.edu)
Suzanne Pierce (suzpierce@jsg.utexas.edu)

AI Opportunities Across Scales



Interactive Workspaces

- Interactive model building
- Synthesis studies
- Specify high-level research questions

Theory-Guided Learning

- Large but minuscule datasets
- Extreme events
- Multiple scales

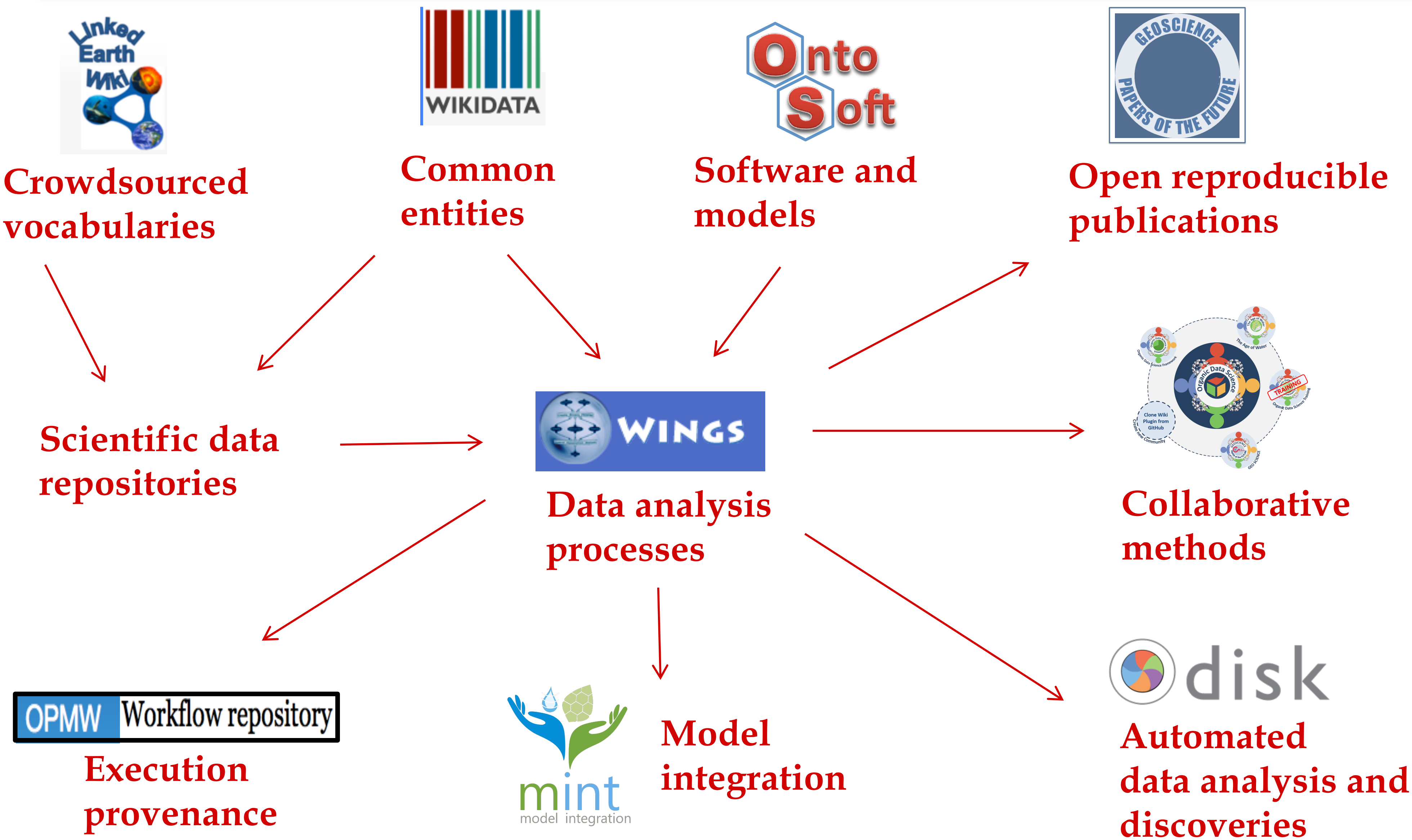
Trusted Threads

- Many locations & data types
- Georeferenced & curated data
- Interlinked information

Model-Driven Sensing

- Selective data collection
- Inaccessible locations
- Optimize experimental design

Open Knowledge Networks



The Future is Here: Knowledge-Powered Data Science

- Focus shifted from data to **models**
  - Model characterization, reuse, and integration
- Need to **incorporate model-centered science knowledge** about phenomena and context
  - Knowledge about **physical, geological, chemical, biological, ecological, and anthropomorphic factors**
  - Knowledge about the user **goals and context**
- This would enable novel forms of reasoning, integrating, visualizing, managing, learning, and discovery with geosciences data

Continue the conversation at http://www.IS-GEO.org

A 20-Year Community Roadmap for Artificial Intelligence Research in the US

Yolanda Gil, USC, AAAI President
Bart Selman, Cornell, AAAI President-Elect

https://arxiv.org/abs/1908.02624

CCC
Computing Community Consortium
Catalyst

AAAI
Association for the Advancement of Artificial Intelligence