

Intelligent Systems for Scientific Discovery

Yolanda Gil



Information Sciences Institute
and Department of Computer Science
University of Southern California

<http://www.isi.edu/~gil>

@yolandagil
gil@isi.edu



Data-Intensive Computing in Science



WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

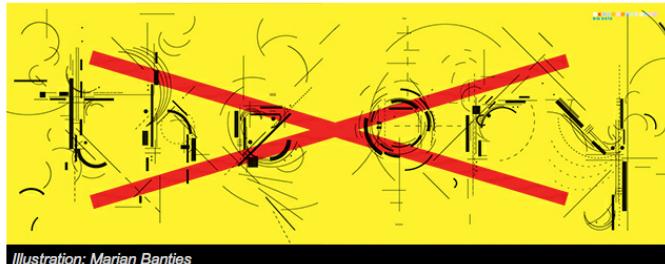
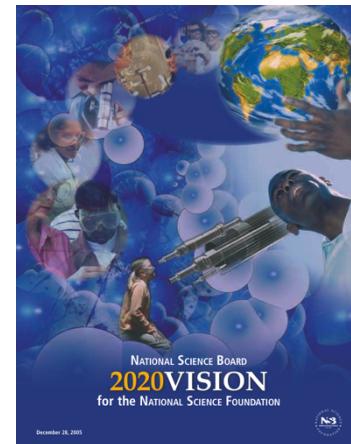
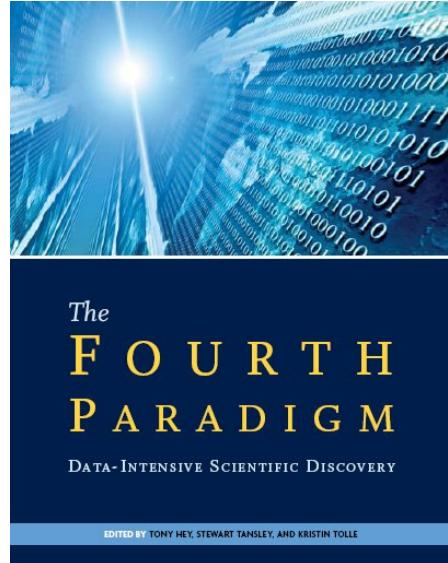


Illustration: Marian Bantjes



s Institute



Yolanda Gil



BIBLIOMETRICS AND CITATION ANALYSIS

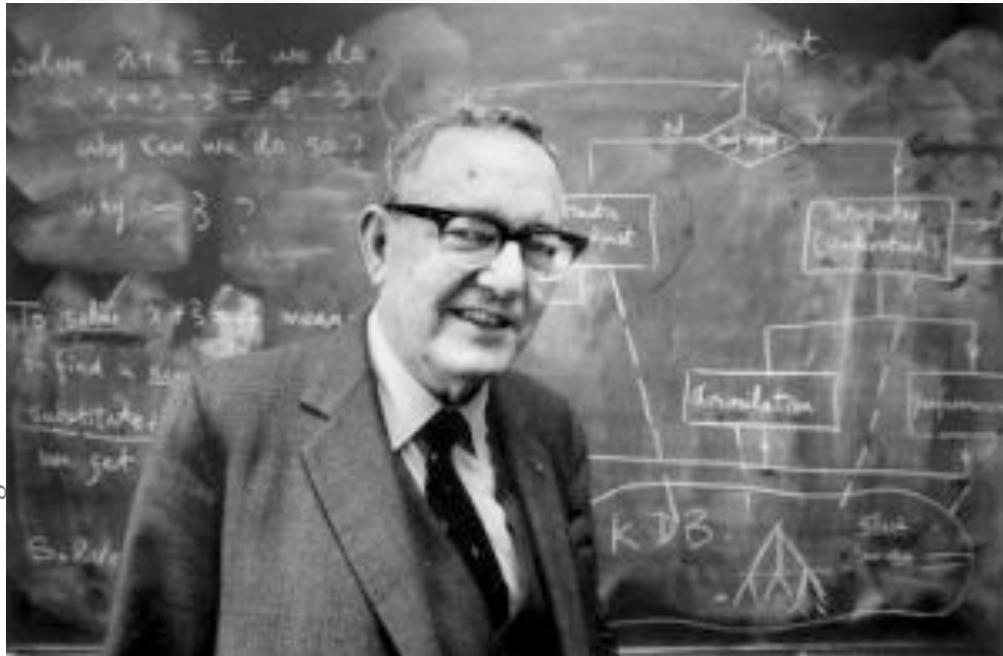
From the Science Citation Index to Cybermetrics



gil@isi.edu

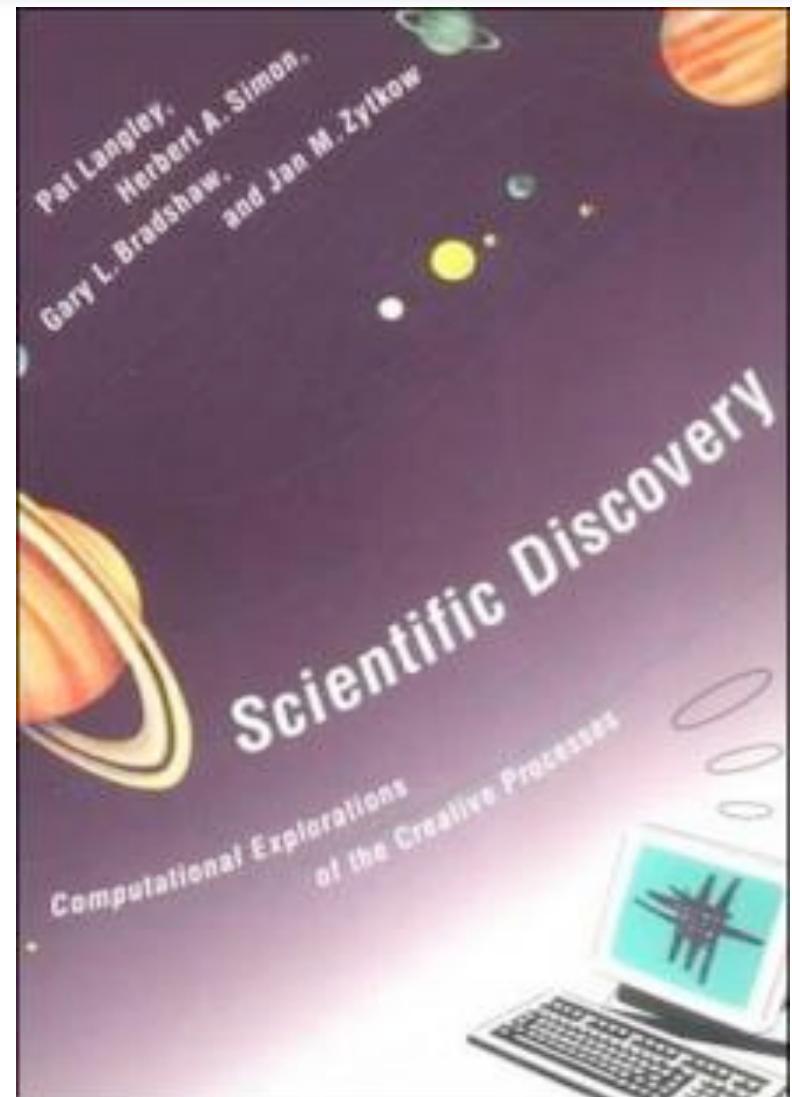
Artificial Intelligence and Scientific Discovery

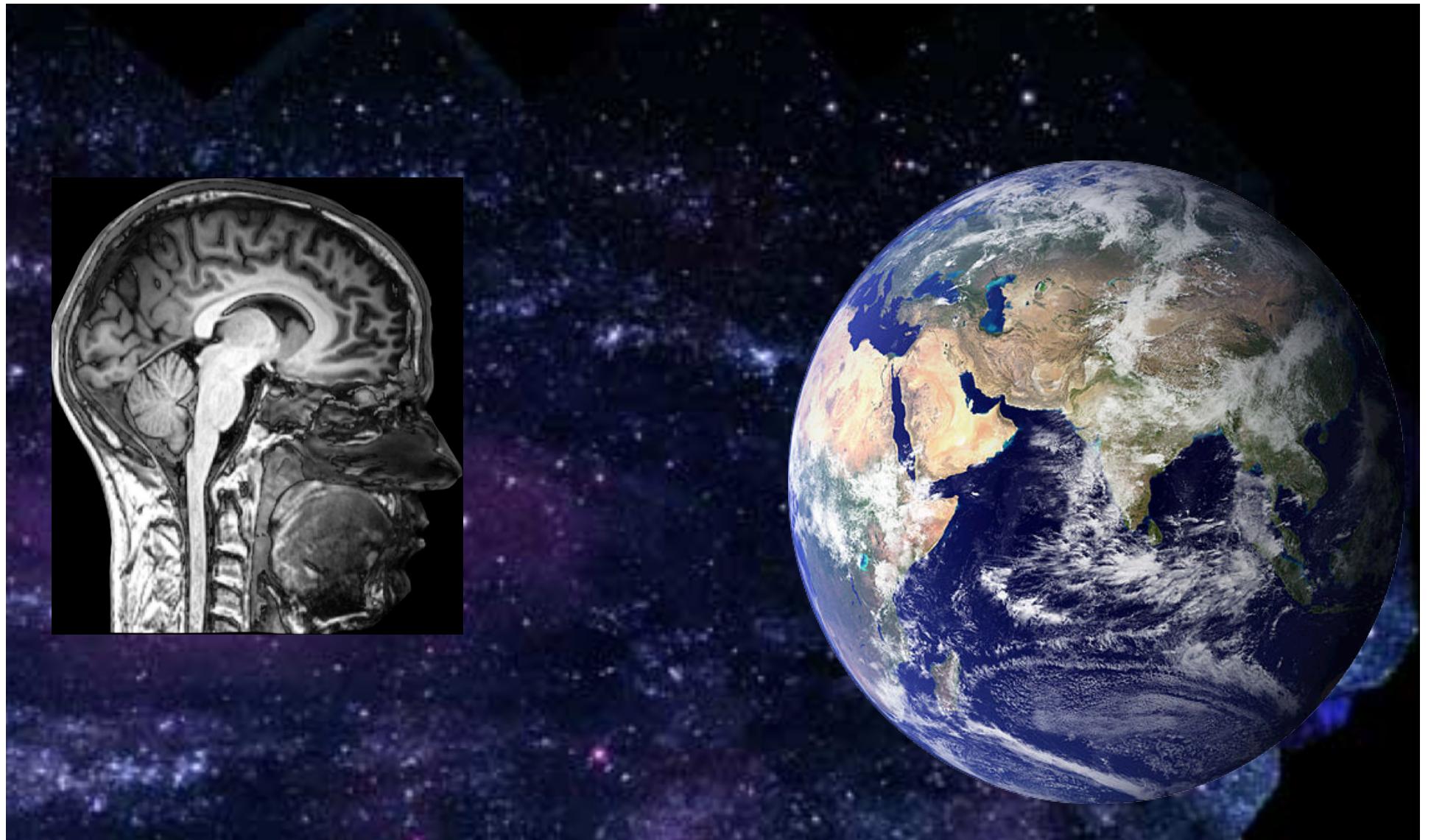
Pittsburg Post Gazette Archives



Computational Scientific Discovery

- [Lenat 1976]
- [Lindsay, Buchanan,
Feigenbaum & Lederberg 1980]
- [Langley & Simon 1981]
- [Simon et al 1983]
- [Falkenhainer 1985]
- [Langley et al 1987]
- [Kulkarni and Simon 1988]
- [Cheeseman et al 1989]
- [Zytkow et al 1990]
- [Valdes-Perez 1997]
- [Todorovski et al 2000]





http://commons.wikimedia.org/wiki/File:MRI_brain_sagittal_section.jpg
http://commons.wikimedia.org/wiki/File:Earth_Eastern_Hemisphere.jpg
http://www.nasa.gov/mission_pages/swift/bursts/uv_andromeda.html

AI's Coming of Age

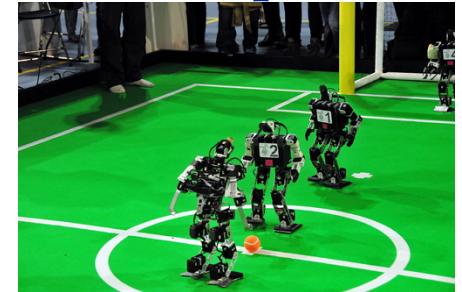
Netflix Recommenders



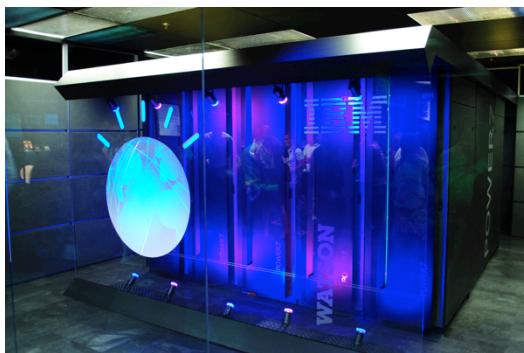
Tesla AutoPilot



RoboCup Soccer



IBM Watson



Google Knowledge Graph



Thomas Jefferson

3rd U.S. President

Thomas Jefferson was an American Founding Father, the principal author of the Declaration of Independence, and the third President of the United States. [Wikipedia](#)

Born: April 13, 1743, Shadwell, VA
Died: July 4, 1826, Charlottesville, VA
Presidential term: March 4, 1801 – March 4, 1809
Spouse: Martha Jefferson (m. 1772–1782)
Party: Democratic-Republican Party
Awards: AIA Gold Medal

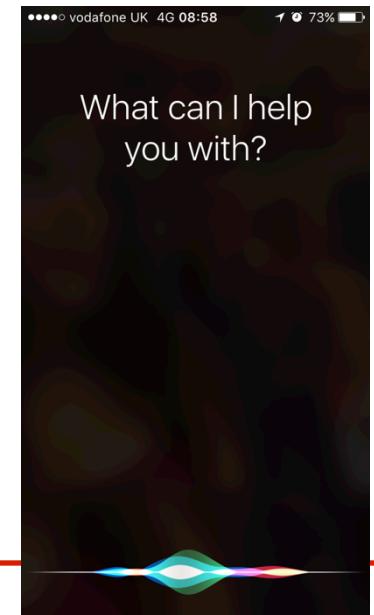
Get updates about Thomas Jefferson

People also search for



[View 15+ more](#)

Apple Siri



[https://en.wikipedia.org/wiki/Watson_\(computer\)](https://en.wikipedia.org/wiki/Watson_(computer)) #/media/File:IBM_Watson.PNG

<https://en.wikipedia.org/wiki/Siri> #/media/File:SiriOniOS9.png

https://commons.wikimedia.org/wiki/File:Google_Knowledge_Panel.png

https://commons.wikimedia.org/wiki/File:17_06_28_robo-cup-worldcup-2005.jpg

http://www.greencarreports.com/news/1000482_tesla-autopilot-the-10-most-important-things-you-need-to-know

<https://en.wikipedia.org/wiki/Netflix#/media/File:NetflixDVD.jpg>

Before There Was the Knowledge Graph...

Google Knowledge Graph (2012)



Thomas Jefferson

3rd U.S. President

Thomas Jefferson was an American Founding Father, the principal author of the Declaration of Independence, and the third President of the United States. [Wikipedia](#)

Born: April 13, 1743, Shadwell, VA

Died: July 4, 1826, Charlottesville, VA

Presidential term: March 4, 1801 – March 4, 1809

Spouse: Martha Jefferson (m. 1772–1782)

Party: Democratic-Republican Party

Awards: AIA Gold Medal

Get updates about Thomas Jefferson

Keep me updated

People also search for



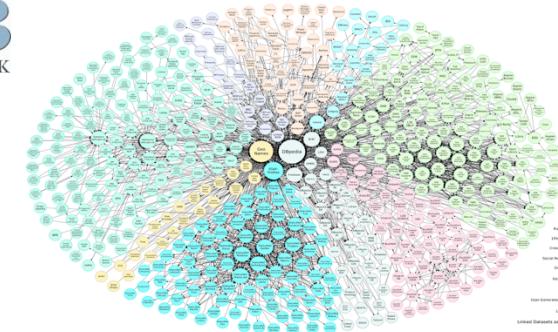
Feedback

Linked Data (2007)

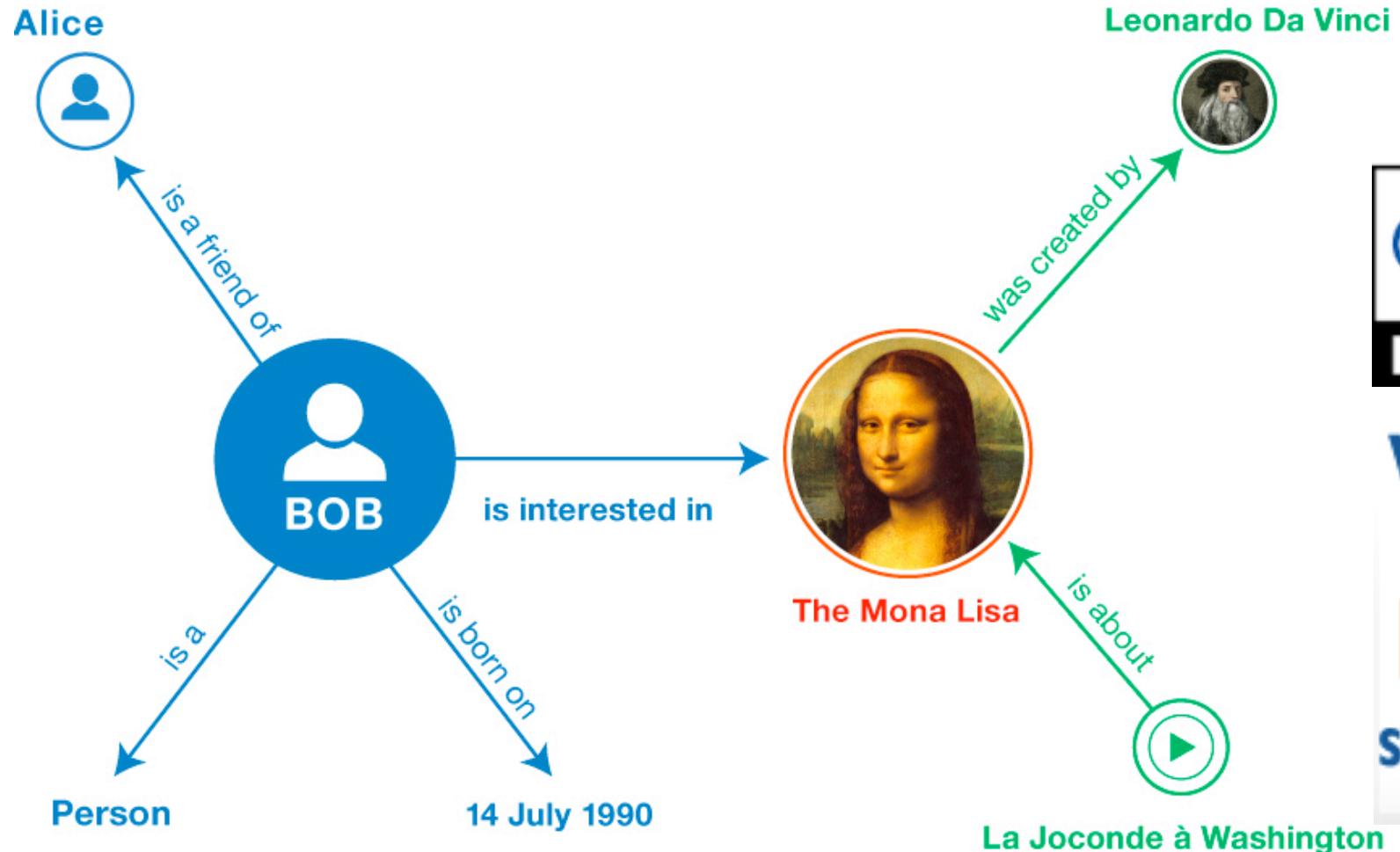


Ordnance Survey Linked Data

36,773,687 triples

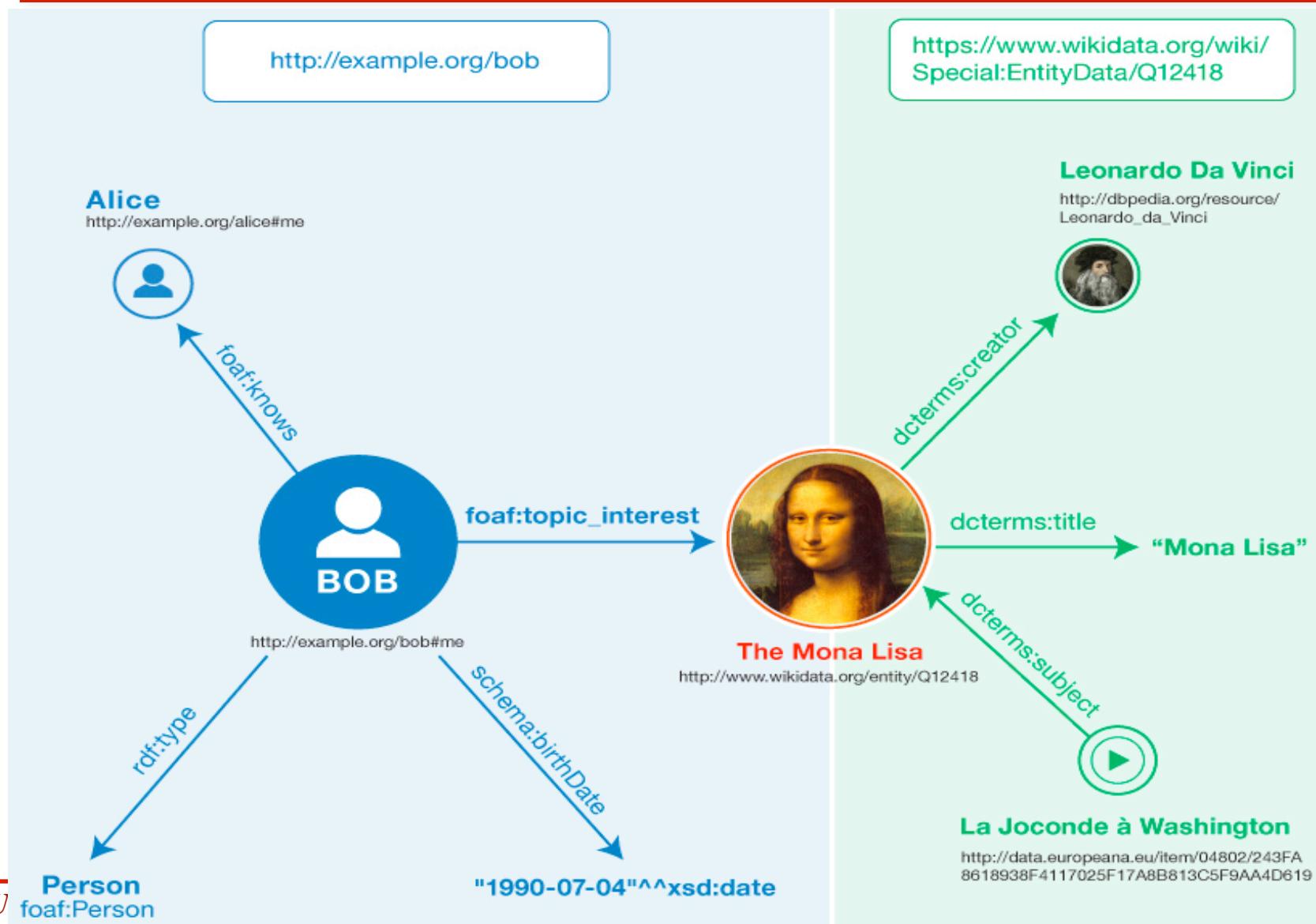


Giving Meaning to Hyperlinks on the Web



<http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>

The Semantic Web

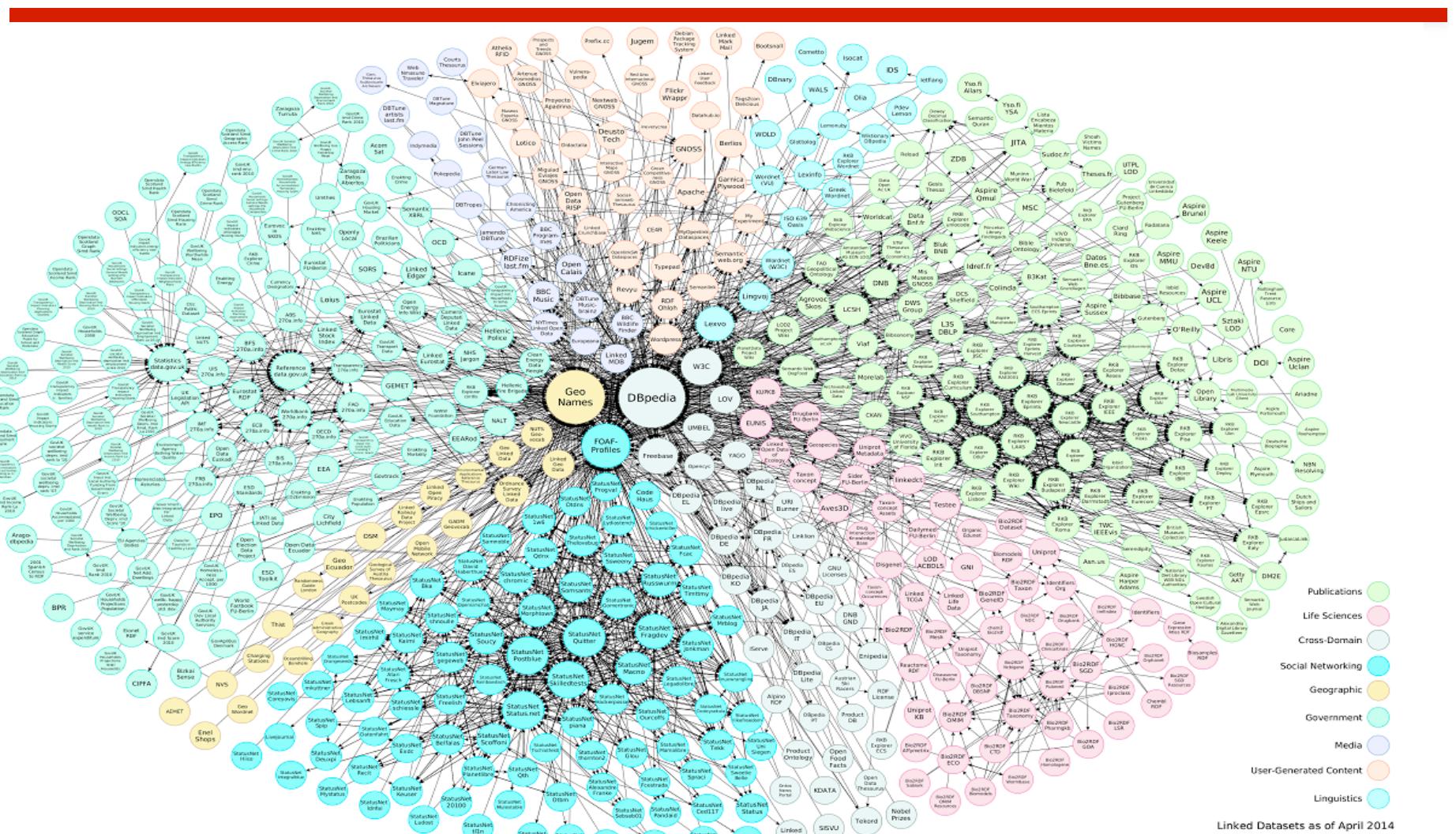


Data and Ontologies on the Semantic Web

```
<Bob> <is a> <person>.  
<Bob> <is a friend of> <Alice>.  
<Bob> <is born on> <the 4th of July 1990>.  
<Bob> <is interested in> <the Mona Lisa>.  
<the Mona Lisa> <was created by> <Leonardo da Vinci>.  
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>.
```

```
<Person> <type> <Class>  
<is a friend of> <type> <Property>  
<is a friend of> <domain> <Person>  
<is a friend of> <range> <Person>  
<is a good friend of> <subPropertyOf> <is a friend of>
```

Interlinked Data and Ontologies in the Semantic Web



"Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net/>"

Interlinked Data and Ontologies on the Web



	2007	2011	2015
Datasets	294	571	3426
Triples	2B	31B	85B
Cross-refs	2M	500M	

74% of datasets in a weakly connected component

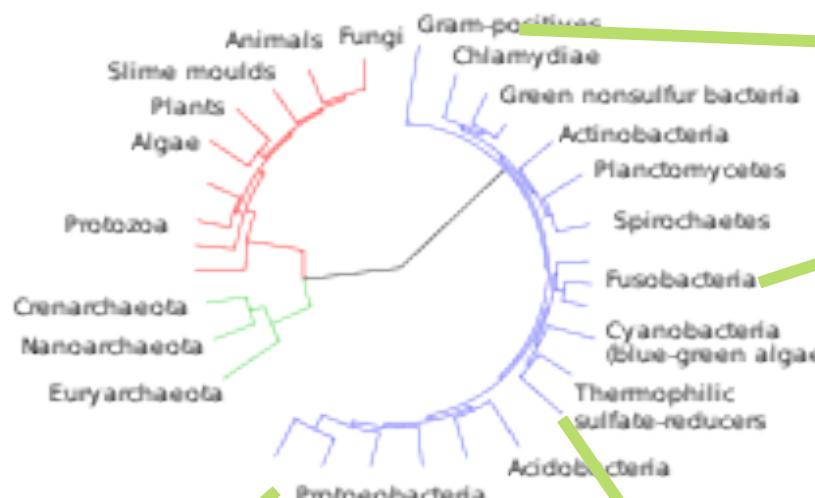
FOAF: from 27% to 59%

DC: from 31% to 56%

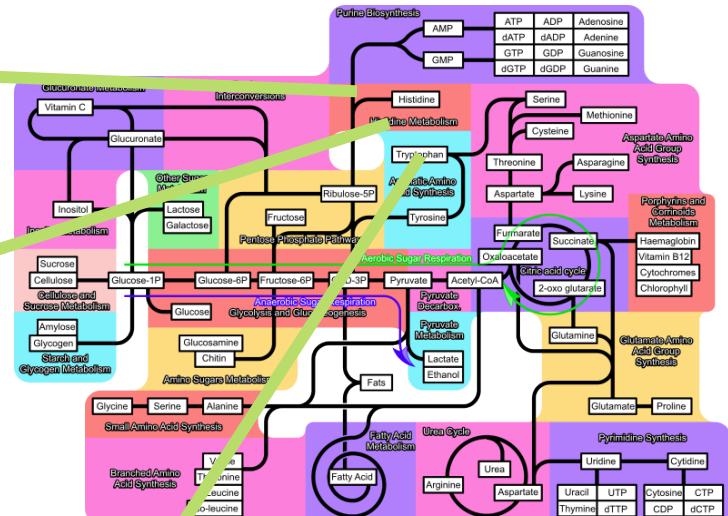
<http://lod-cloud.net>
<http://stats.lod2.eu>

Interlinking Scientific Knowledge

Taxonomical

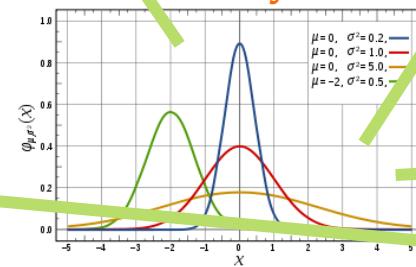


Networks

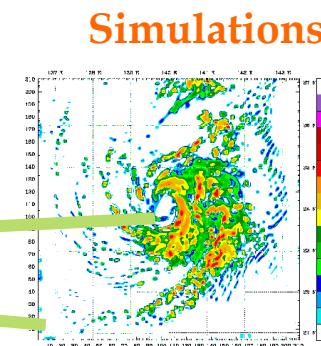


Mathematical

$$E_r = \sqrt{(m_0 c^2)^2 + (pc)^2}$$

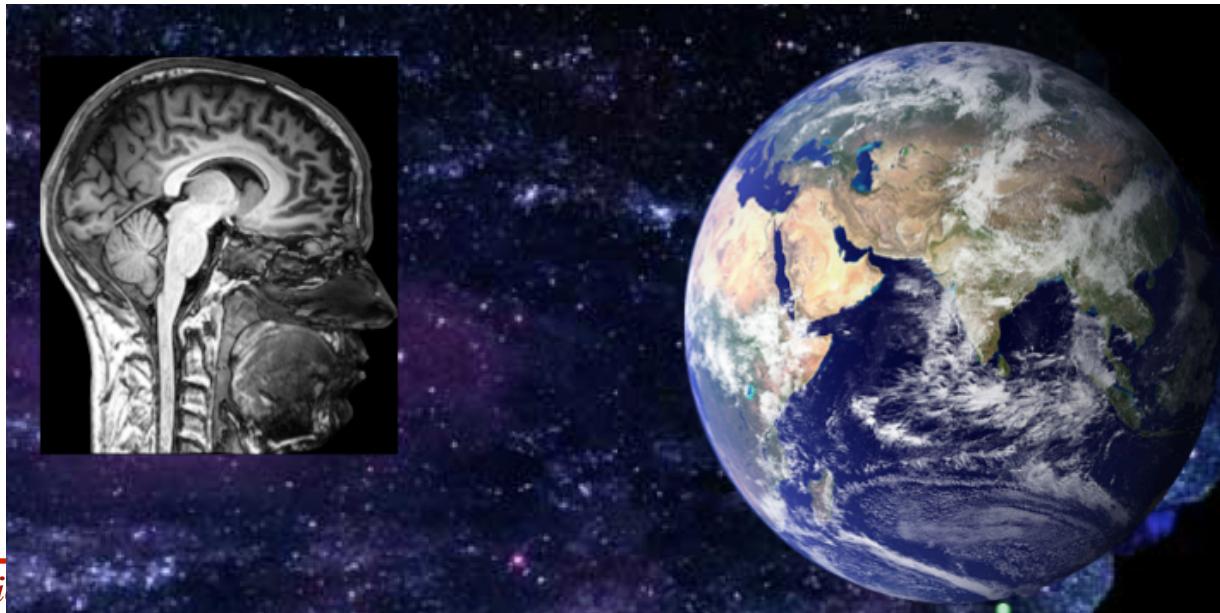
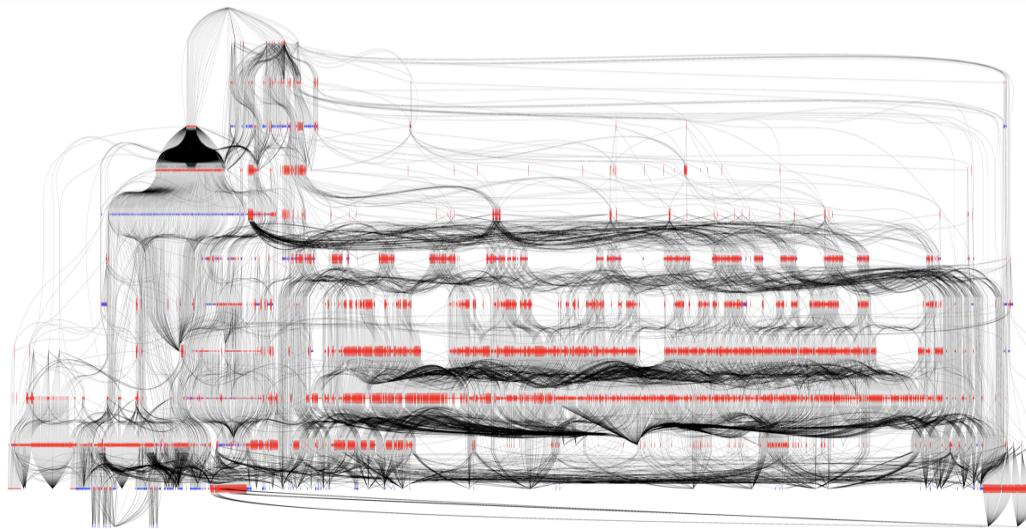


Bayesian



Simulations

Complexity of Scientific Endeavors



Focus: Intelligent Systems for Data Analysis

What is the state of the art?

What is a good problem to work on?

What is a good experiment to design?

What data should be collected?

What is the best way to analyze the data?

What are the implications of the experiments?

What are appropriate revisions of current models?

What to focus on next?

Capturing Scientific Knowledge

Data



Software



Provenance

W3C® PROV

OPMW Workflow repository

Workflows



Meta-Workflows



Knowledge about Data: Linked Earth Wiki



Work with Julien-Emile Geay of USC and Nick McKay of NAU

Palmyra Atoll [edit]

Structured Properties

main type (GND)	geographical feature	[edit]
is in the administrative unit	United States Minor Outlying Islands	[edit]

Porites

Structured Properties

add fact [x] Property:Name Topic:Finger Coral [hide]

- [x] [http://dbpedia.org/resource/Porites](#)
- [add source]

Wikipedia Entry go to original Wikipedia article

Porites is a genus of stony coral; they are SPS (Small Polyp Stony) corals. They are characterised by a finger-like morphology. Members of this genus have widely spaced

Finger Coral

Geochemistry datasets

	Archive	Interpretation	MeasurementMaterial	MeasurementStandard	MeasurementUnits
Lake Bosumtwi	LakeSediments	Lake Level	Authigenic Calcite	VPDB	Permil
Quelccaya	IceCore		Ice	VSMOW	Permil
Palmyra coral 20C	Coral	SST,SSS	Skeletal aragonite	VPDB	Permil

Palmyra coral 20C

Data

- **DOWNLOAD**

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

Structured Properties

<input checked="" type="checkbox"/>	SiteName	<input checked="" type="checkbox"/>	Palmyra	(By Julien)
<input checked="" type="checkbox"/>	Archive	<input checked="" type="checkbox"/>	Coral	(By Julien)
<input checked="" type="checkbox"/>	Domain(s)	<input checked="" type="checkbox"/>	Climate,geochemistry	(By Julien)
<input checked="" type="checkbox"/>	Forward model	<input checked="" type="checkbox"/>	10.1029/2011GL048224	(By Julien)
<input checked="" type="checkbox"/>	Genus	<input checked="" type="checkbox"/>	Porites	(By Julien)
<input checked="" type="checkbox"/>	Interpretation	<input checked="" type="checkbox"/>	SST,SSS	(By Nick)
<input checked="" type="checkbox"/>	Measurement			
<input checked="" type="checkbox"/>	MeasurementMaterial			
<input checked="" type="checkbox"/>	MeasurementStandard			
<input checked="" type="checkbox"/>	MeasurementUnits			
<input checked="" type="checkbox"/>	Reference			
<input checked="" type="checkbox"/>	Species			

Credits

Users who have contributed to this Page:

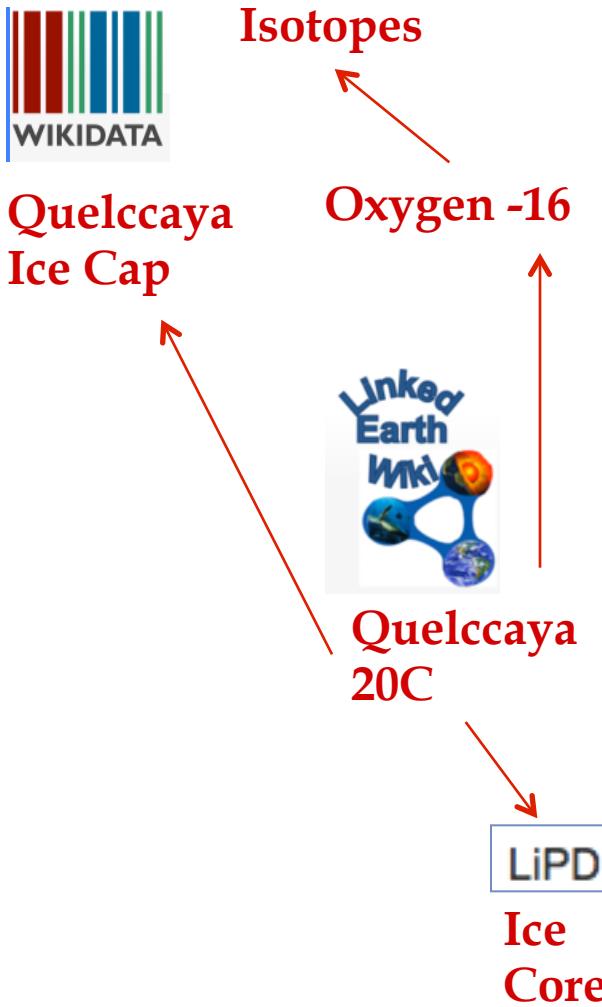
- **Julien** (43 Edits)
- **Nick** (34 Edits)

You

{{#ask: [[Is a:dataset]] | ?Domain=geochemistry | ?Archive | ?MeasurementMaterial | ?MeasurementStandard | ?MeasurementUnits}}

AI opportunities:
 - collection
 - normalization
 - organization

Linked Data and Linked Knowledge



Capturing Scientific Knowledge

Data



Software



Provenance

W3C® PROV

OPMW Workflow repository

Workflows



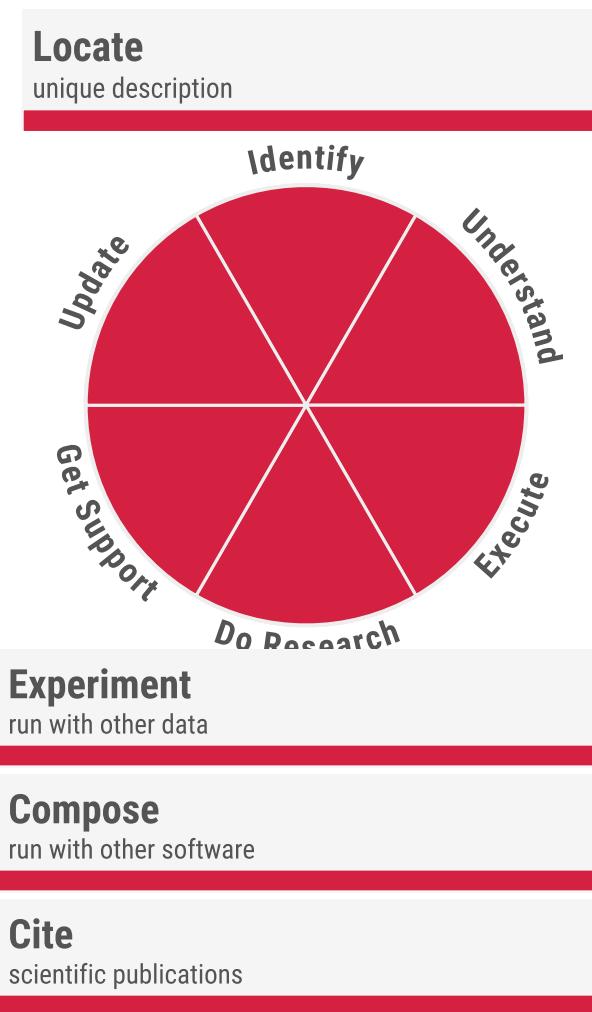
Meta-Workflows



Knowledge about Software: OntoSoft

Work with C. Duffy of PSU, C. Mattmann of JPL, S. Peckham of CU, and E. Robinson of ESIP

Contribute evolution
Track versions
Discuss support and community



Trust quality and ratings
Relate domain knowledge
Access download
Install execution requirements
Run testing execution

Knowledge About Software: Physical Variables and Assumptions



Software

Add ▾ Rename Delete

- App
- SoftwareComponent
 - DataProcessingComponent
 - ForceAnalysis
 - mklcmat.m
 - ModelComponent
 - ReaerationModels
 - ReaerationModels-Empirical
 - BatchReaerationCM
 - ReaerationCM
 - ReaerationODM
 - ReaerationOGM
 - ReaerationModels-Physics
 - ReaerationEDM
 - VisualizationComponent
 - PlotK2
 - plotlcprofiles.m
- SoftwarePackage
 - ModelPackage
 - PIHM
 - TopoFlow
- VisualizationPackage

Describe Software PIHM

Make suggestions Save

I/O

Assumptions

Standard Names

Summary

Cor

IN Inputs

+ Add -

Identifier

ProjectNa

MeshFile

AttFile

GeoFile

RivFile

ForcFile

IbcFile

Init

Calib

CF CSDMS

Standard Names

+ Add Standard Name - Delete

Object	Quantity	Operators
air	relative_humidity	
air	temperature	
air_water_vapor	partial_pressure	
atmosphere_water	precipitation_rate	
atmosphere_water_vapor	partial_pressure	
ground_water_table	depth	
land_snow	melt_rate	
land_surface	None	



C S D M S

COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

OntoSoft: Comparing Software Implementations



Compare Software				
PIHM	PIHMgis	DrEICH	TauDEM	WBMsed
What are domain specific keywords for this software ? (eg: hydrology, climate)				
Geomorphology, Hydrological, Bedrock channel ero-	Basins, Continental	Basins, GIS	Hydrologically corrected DEM, Watershed	Sediment flux, Global model, Hydrological model
What Operating Systems can the software run on ?				
Unix Linux	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Linux
Is there any test data available for the software ?				
Test Data Location: http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full Test Data Description: Two test DEMs are included in the repository,	Test Data Location: http://source-forg.../projects/pihmmodel/ Test Data Description: Upper Juniata River 875 km^2: see: http://source-forg.../projects/pihmmodel/		Test Data Location: http://csdms.colorado.edu/wiki/Model:TauDEM#Testing Test Data Description: The Logan River DEM is a small test dataset useful	Test Data Location: http://csdms.colorado.edu/wiki/Mod...:WBMsed#Testing Test Data Description: Extensive input dataset is available on the CSDMS

OntoSoft: Publishing Software Metadata as RDF



Screenshot of the PIHM software metadata page on the OntoSoft platform.

The top navigation bar includes the OntoSoft logo, a menu icon, and a circular icon with a red and blue gradient.

The main content area shows the software title "PIHM" and author "[Christopher Duffy]". Below the title is a button bar with "HTML", "RDF/XML", and "JSON" options, with "RDF/XML" being highlighted. A red oval surrounds this button bar.

The "Identify" section contains the "Locate" sub-section with the text "Unique description".

Two input fields are present:

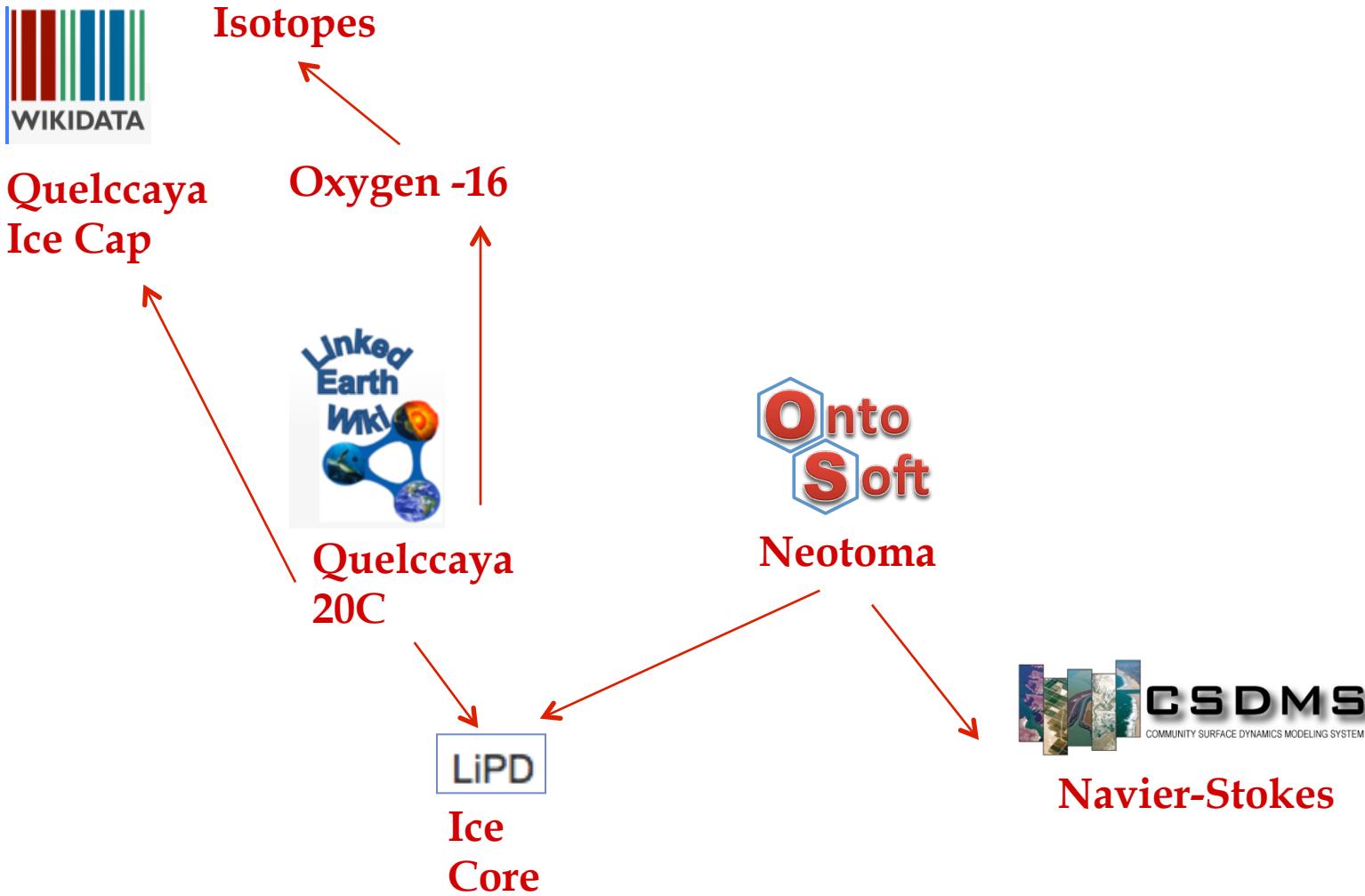
- "What is the software called ?" with the option "PIHM" selected.
- "What is a short description for this software ?" with the text "PIHM is a multiprocess, multi-scale hydrologic model where the major hydrological processes are fully coupled using the semi-discrete finite volume method. PIHM is a physical model for surface and groundwater flow and solute transport.".

A yellow callout box on the right lists "AI opportunities":

- functional desc.
- organization
- linking to data

Page footer: "US" and page number "3".

Linked Data and Linked Knowledge



Capturing Scientific Knowledge

Data



Software



Provenance

W3C® PROV

OPMW Workflow repository

Workflows



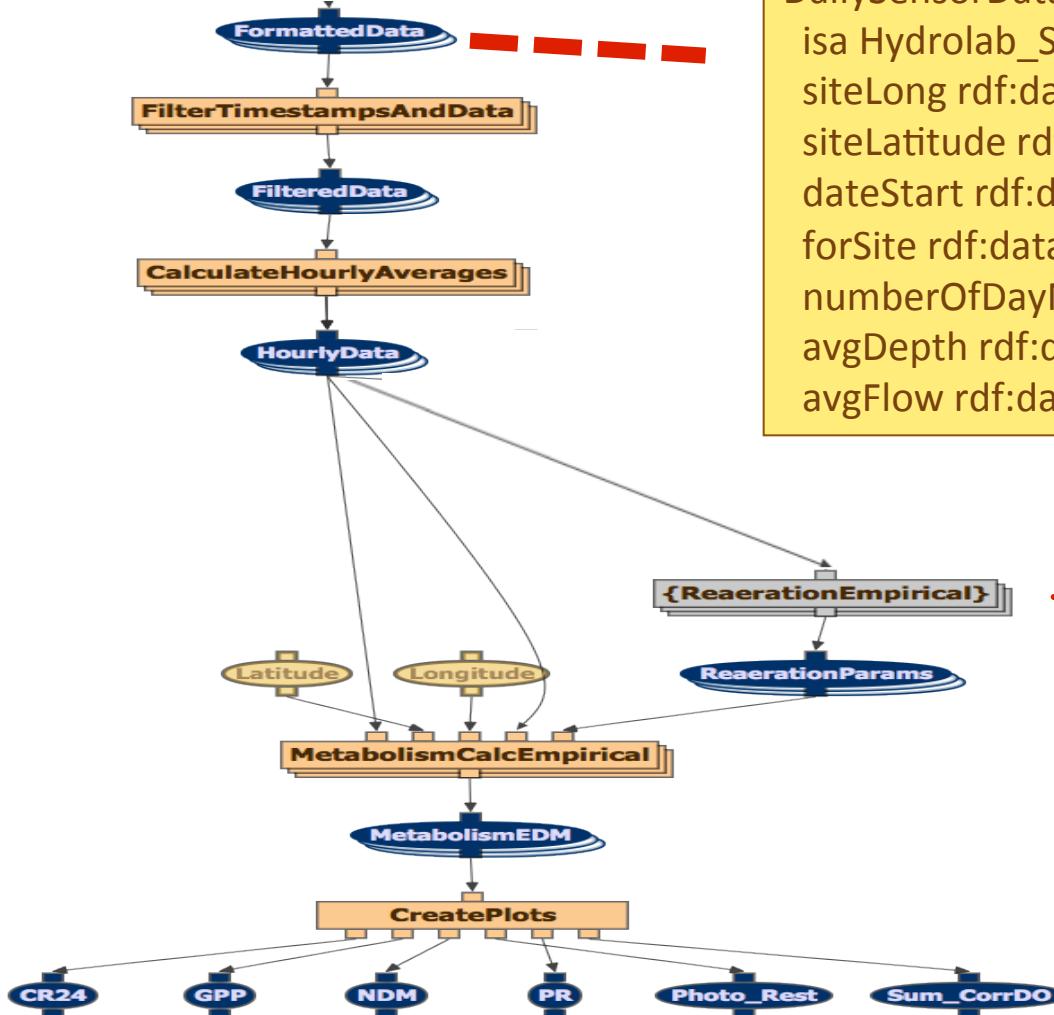
Meta-Workflows



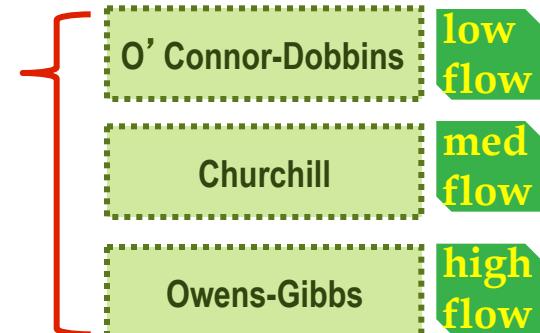
Knowledge about Data Analysis: WINGS



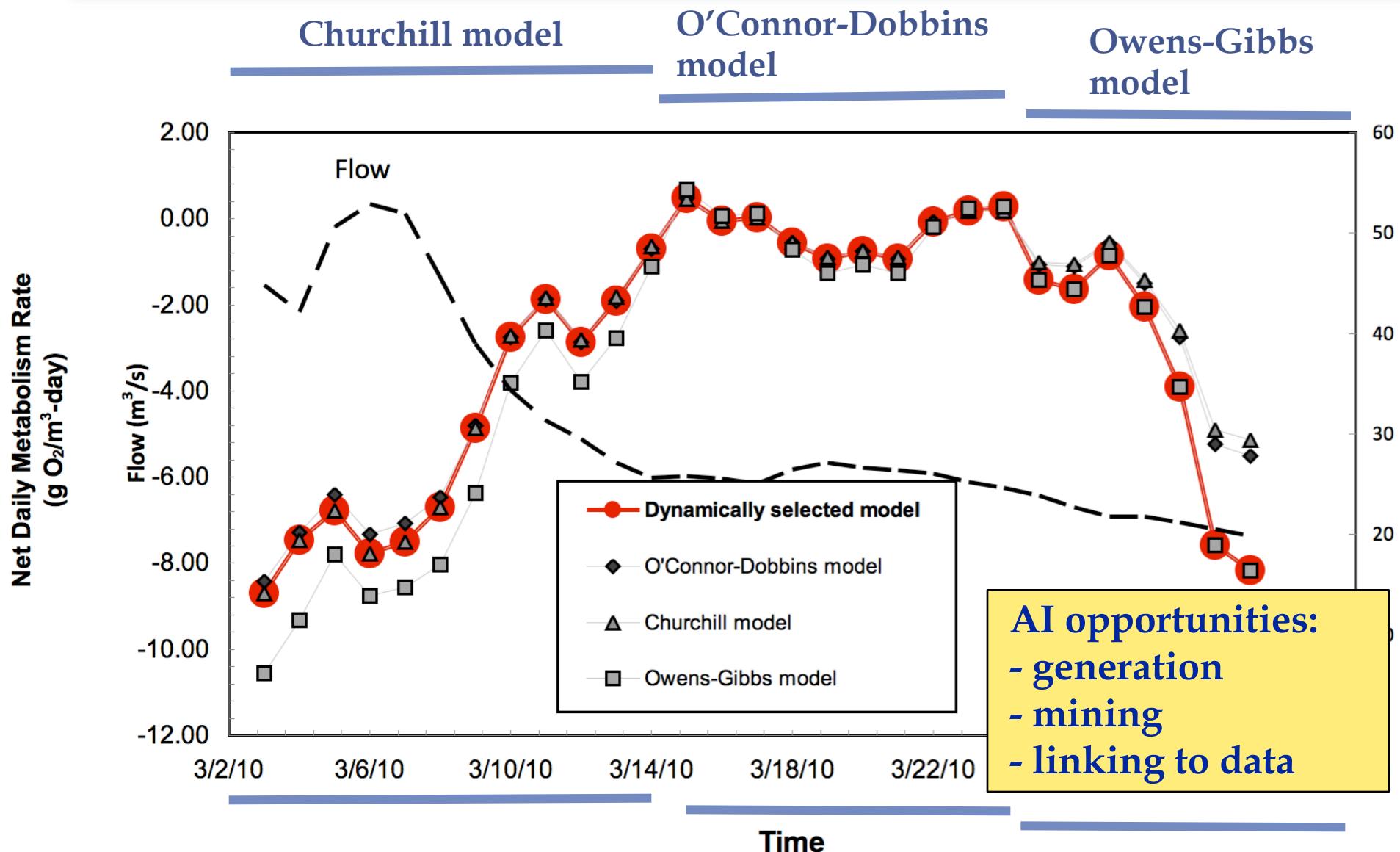
Work with V. Ratnakar (USC)



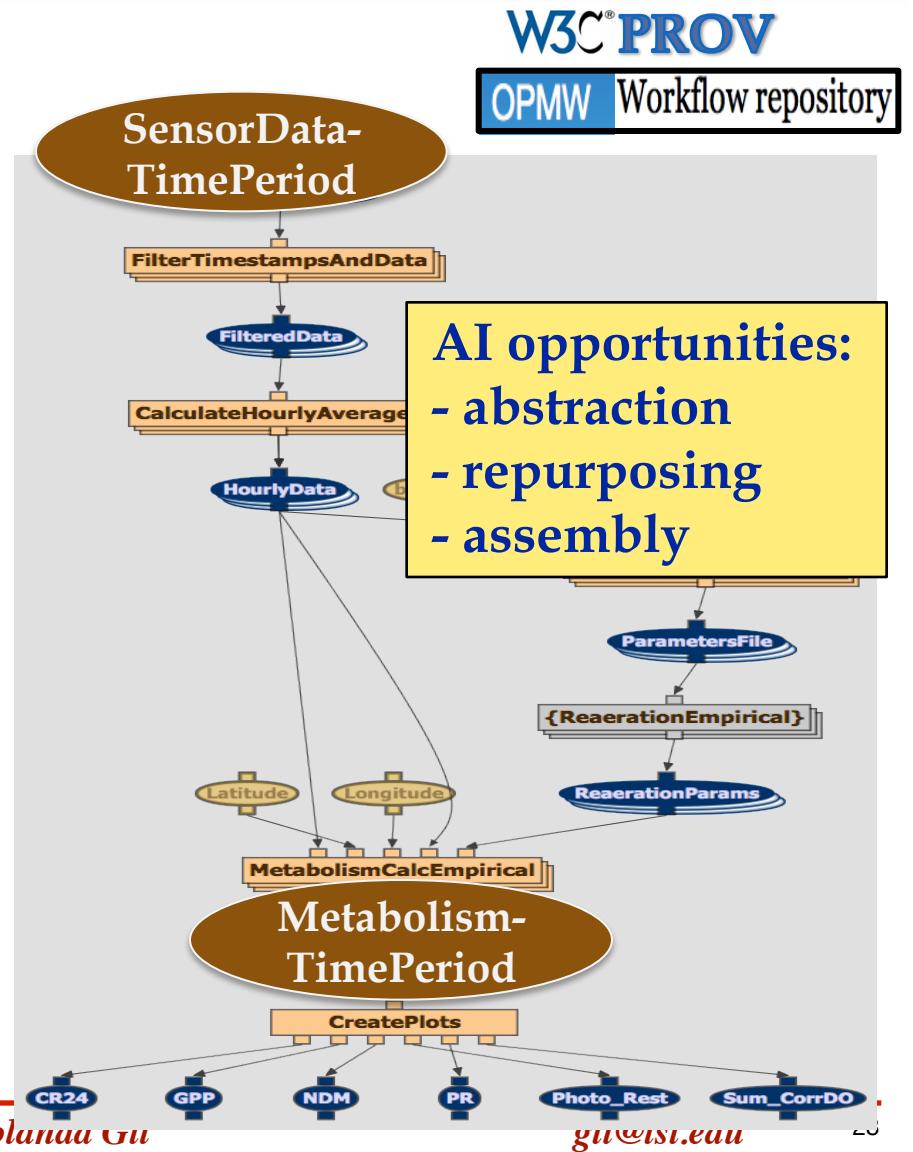
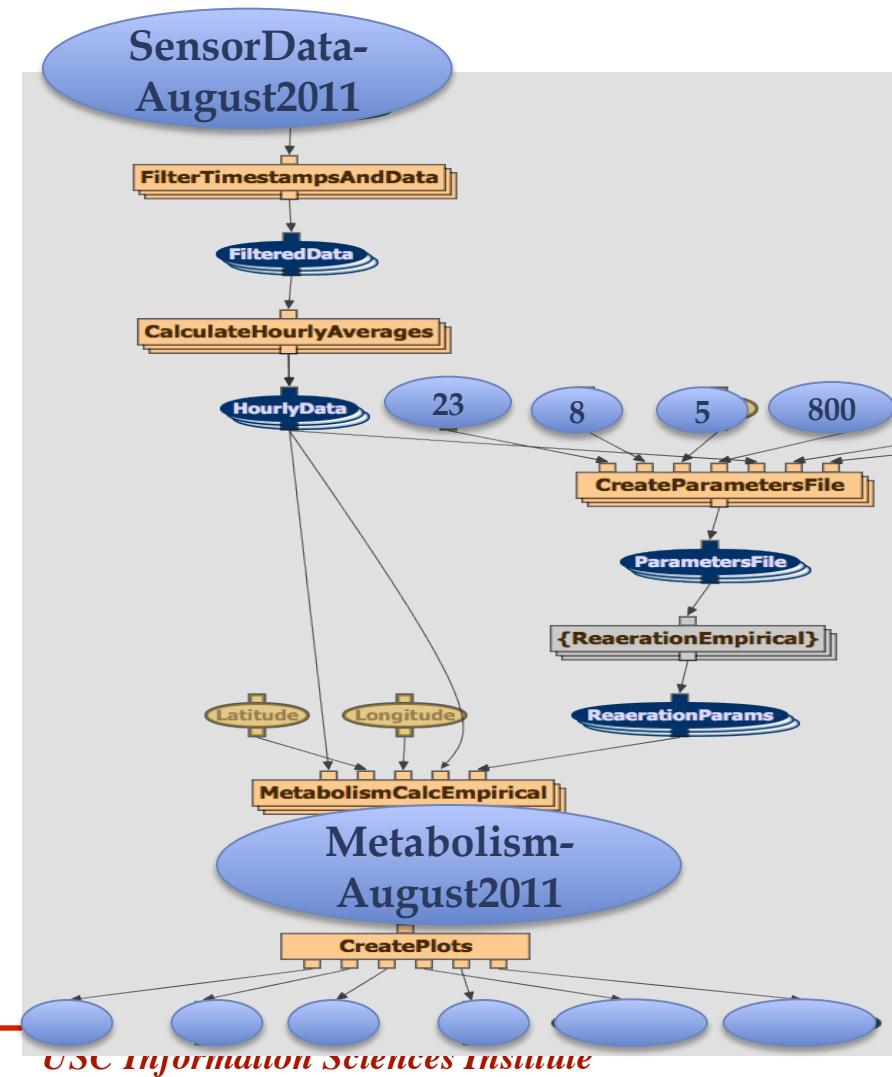
DailySensorData
isa Hydrolab_Sensor_Data
siteLong rdf:datatype="long"
siteLatitude rdf:datatype="lat"
dateStart rdf:datatype="date"
forSite rdf:datatype="site"
numberOfDayNights rdf:datatype="int"
avgDepth rdf:datatype="depth"
avgFlow rdf:datatype="flow"



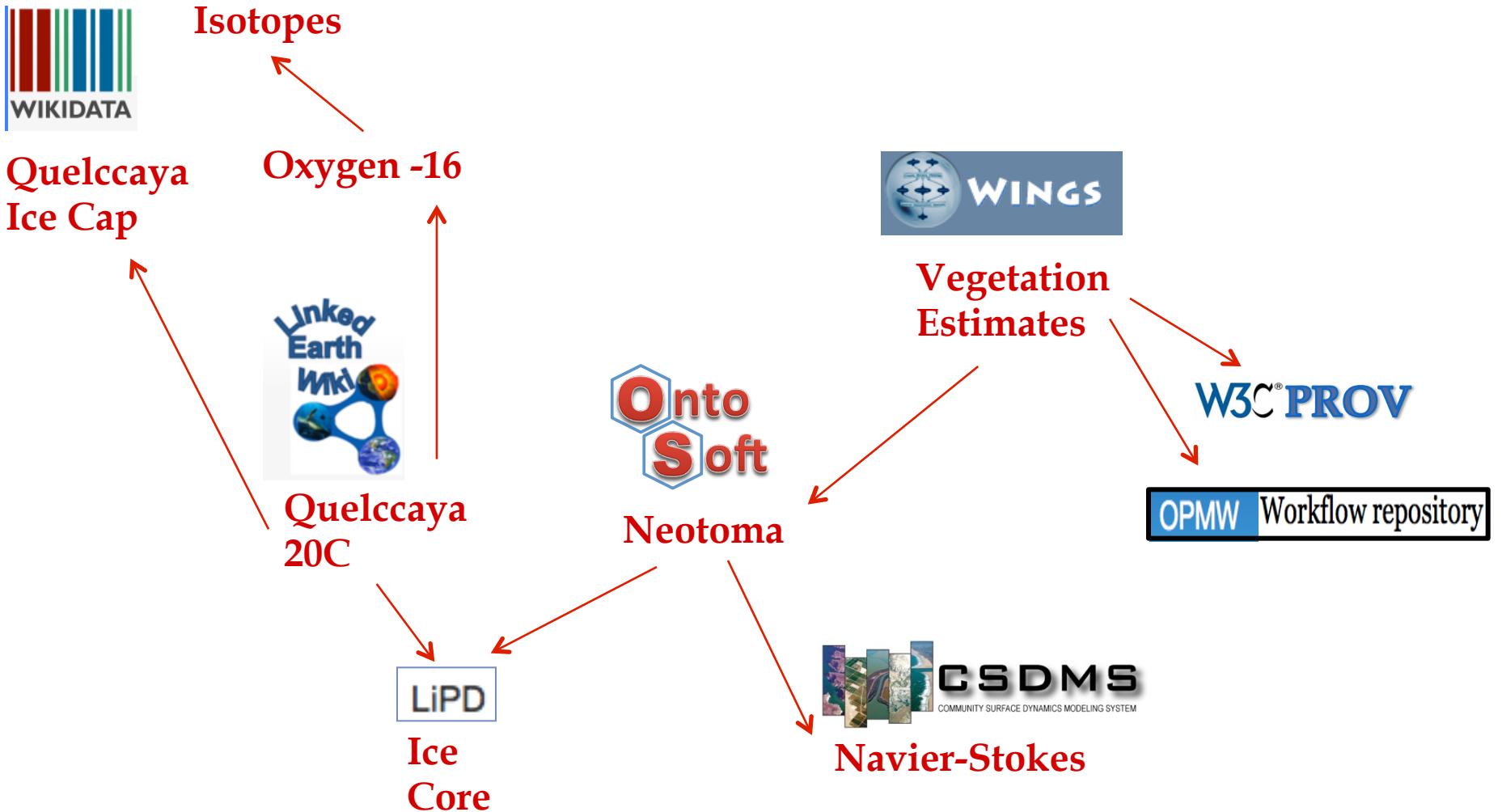
WINGS Dynamically Customizes the Workflow Based on Daily Sensor Readings



Describing Execution (Provenance) vs General Method (Workflow)



Linked Data and Linked Knowledge



Capturing Scientific Knowledge

Data



Software



Provenance

W3C® PROV

OPMW Workflow repository

Workflows



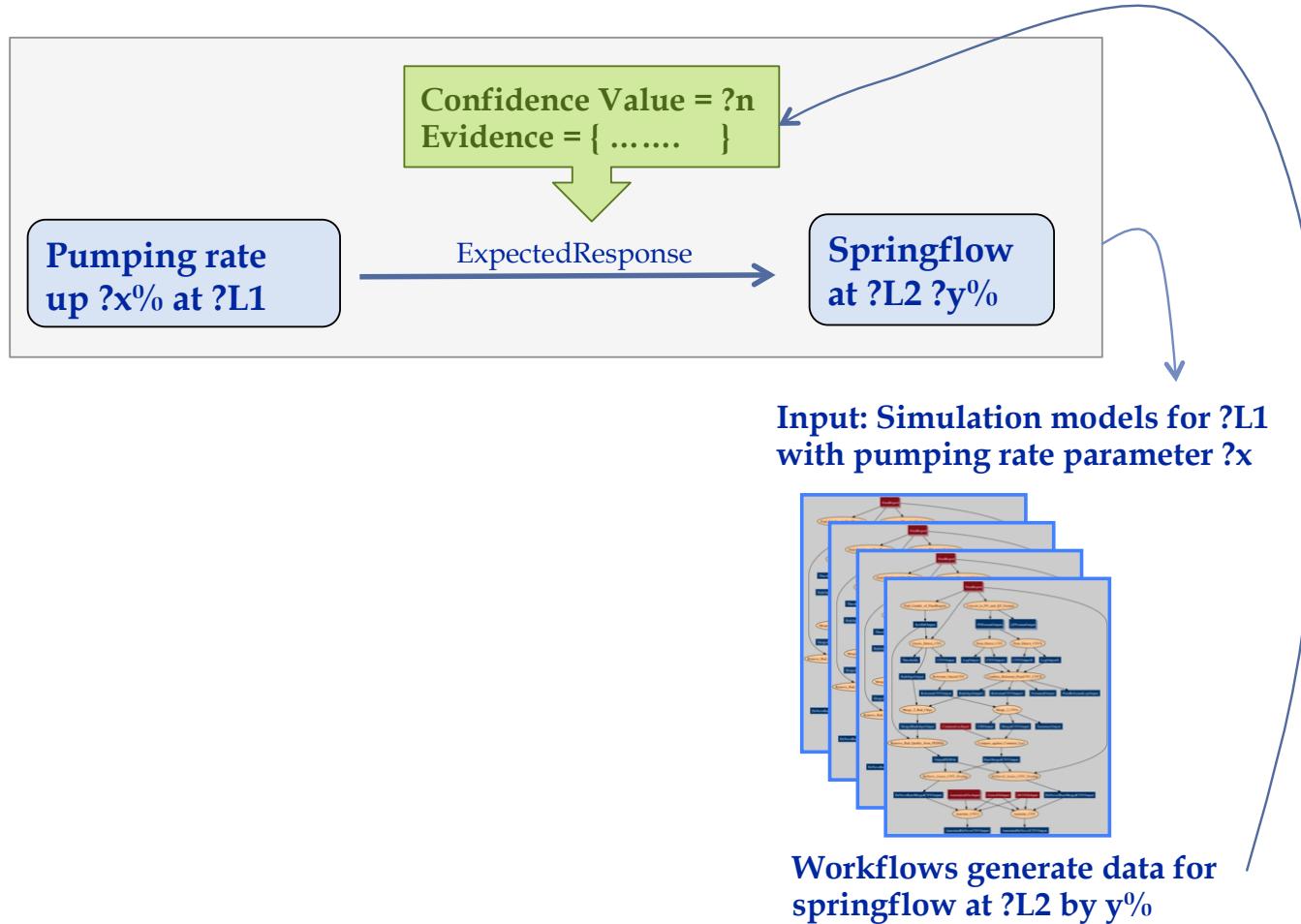
Meta-Workflows



Knowledge about Meta-Processes: DISK



Work with P. Mallick (Stanford U) and S. Pierce (UT Austin)



DISK: Hypotheses

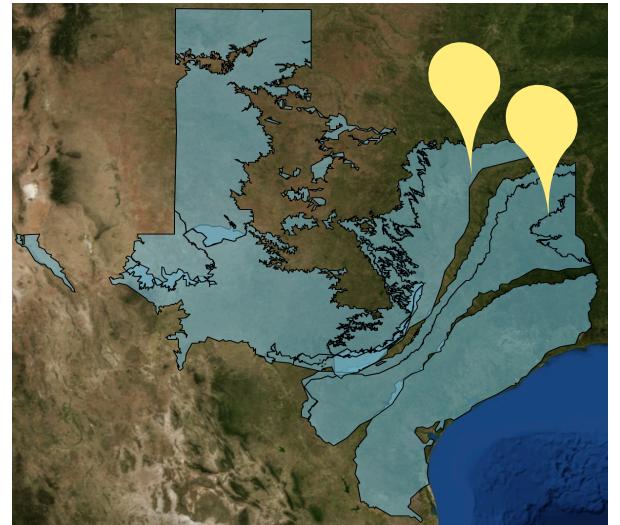


Pumping rate up
10% at Kemp

Expected Response

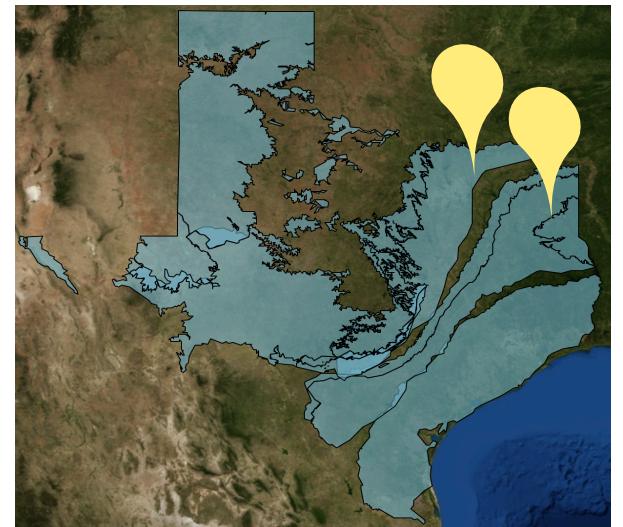
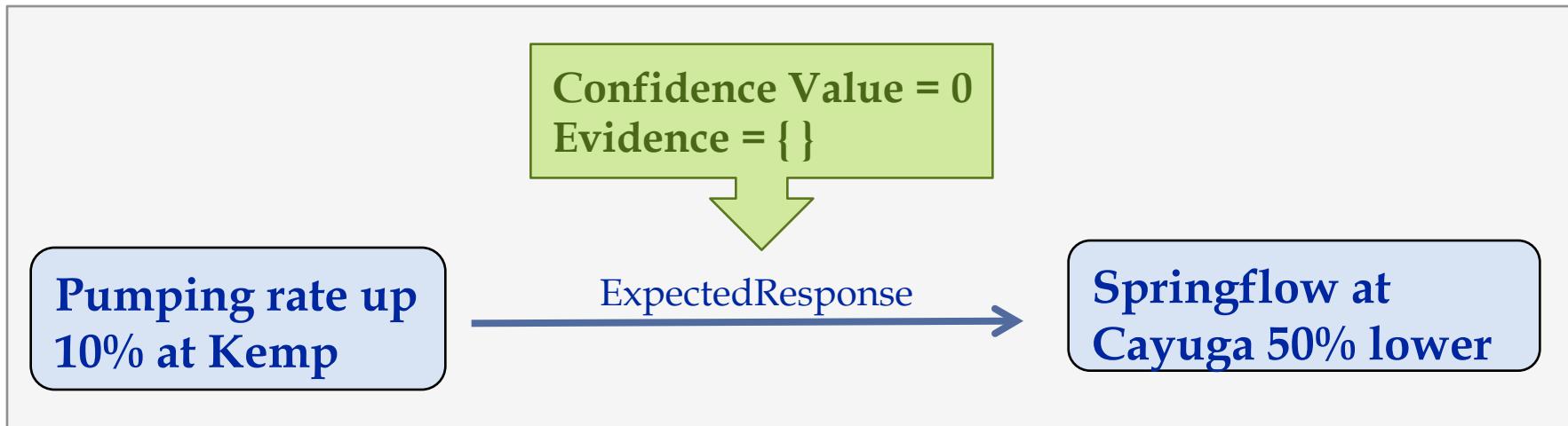
Springflow at
Cayuga 50% lower

33 groundwater
models for Texas



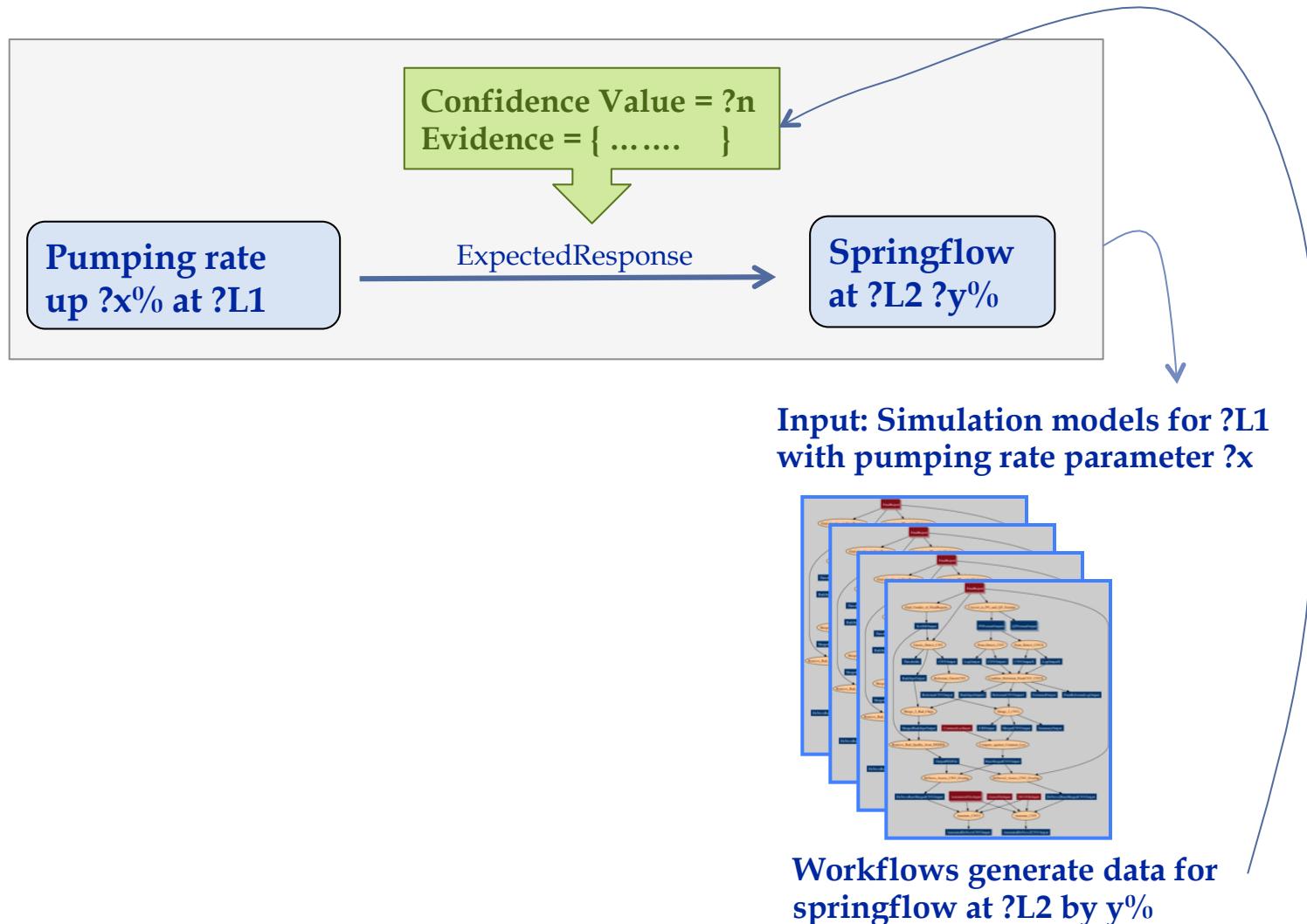


DISK: Hypotheses

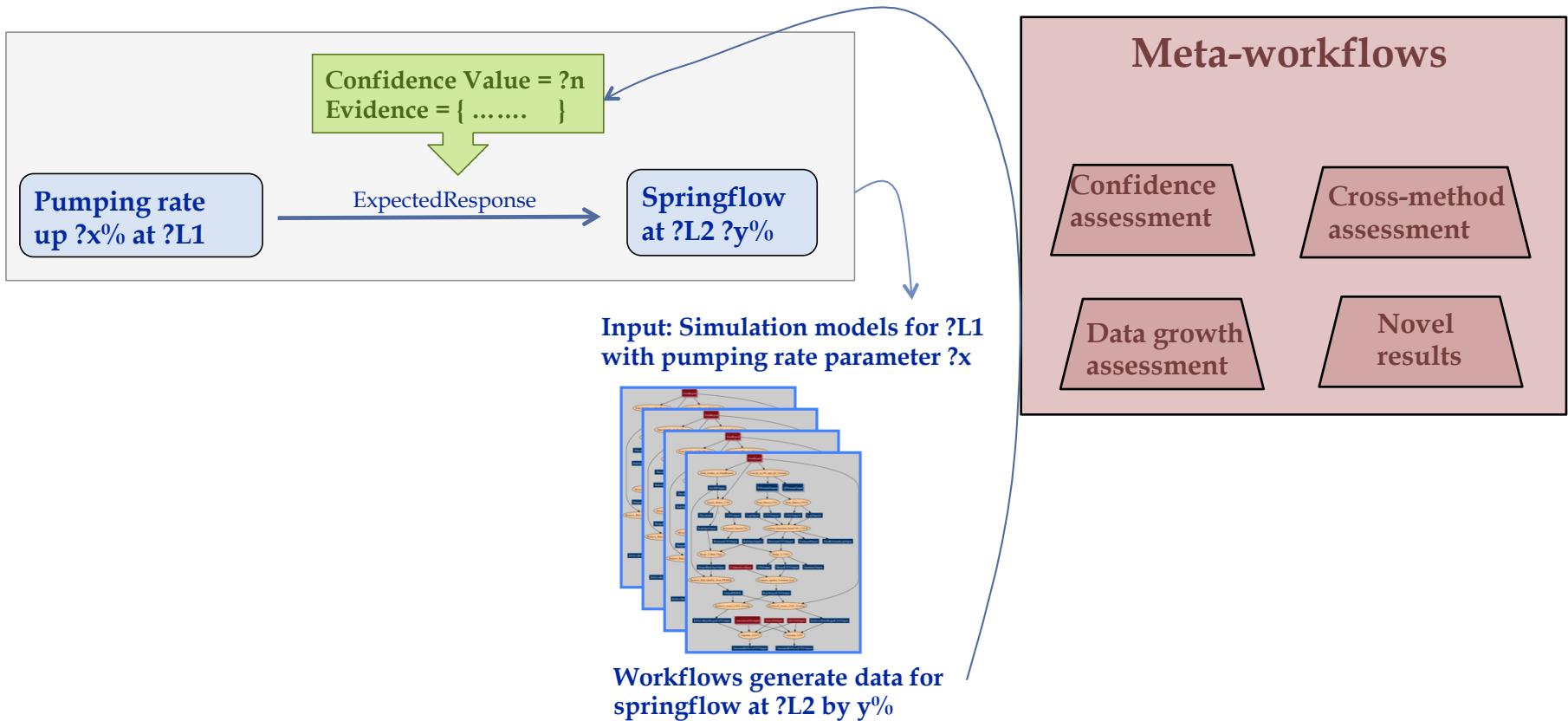




DISK: Lines of Inquiry



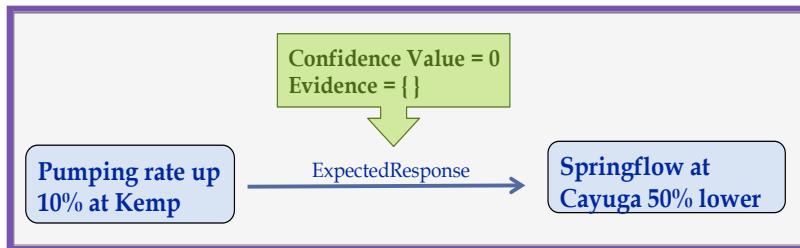
DISK: Lines of Inquiry



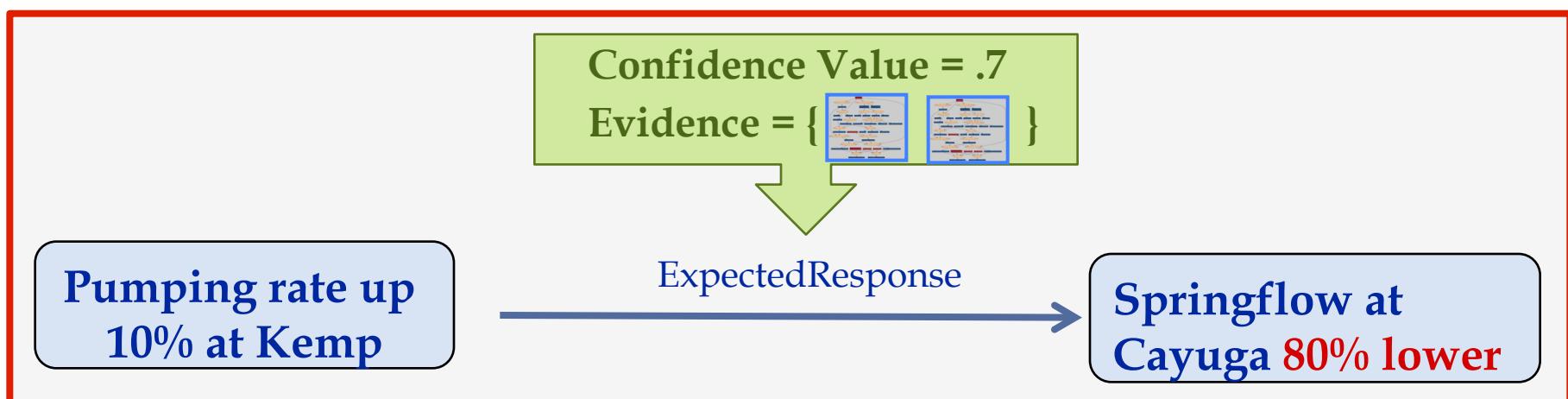
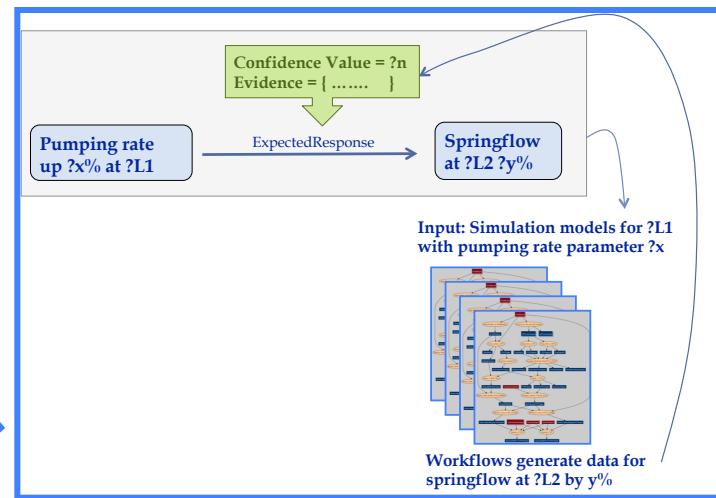
DISK: Matching Hypotheses Against Lines of Inquiry



Hypotheses



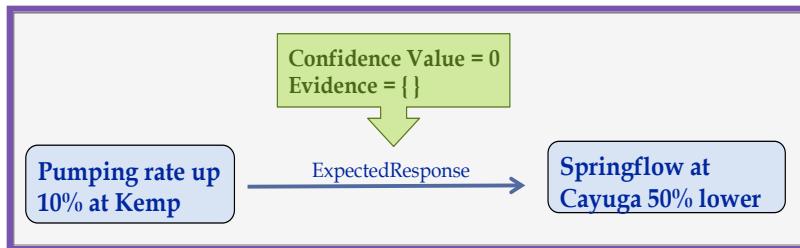
Lines of Inquiry



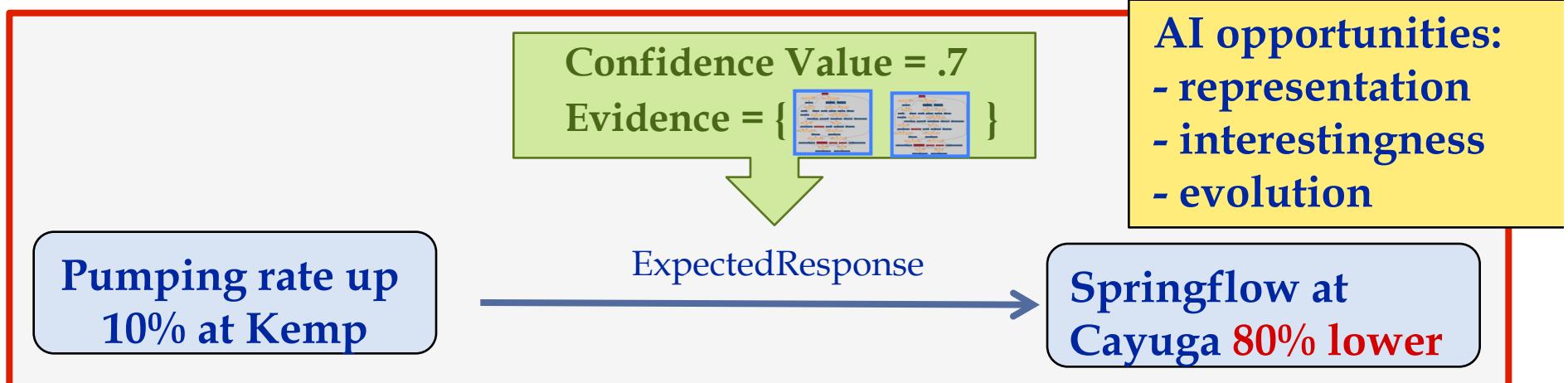
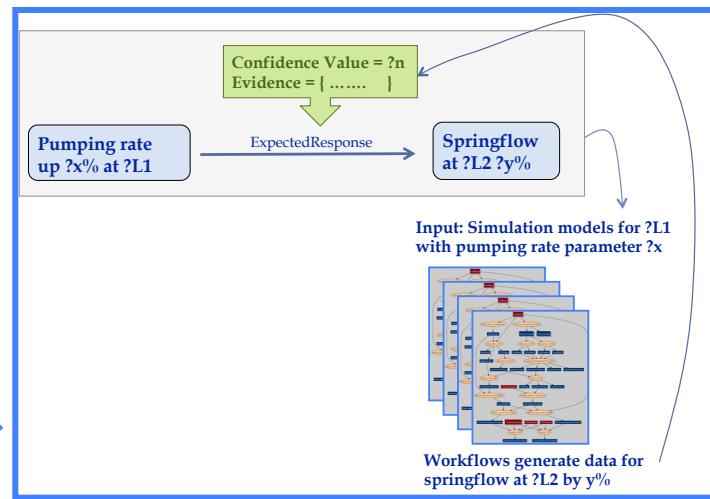
DISK: Matching Hypotheses Against Lines of Inquiry



Hypotheses



Lines of Inquiry



Knowledge about Meta-Processes: Organic Data Science

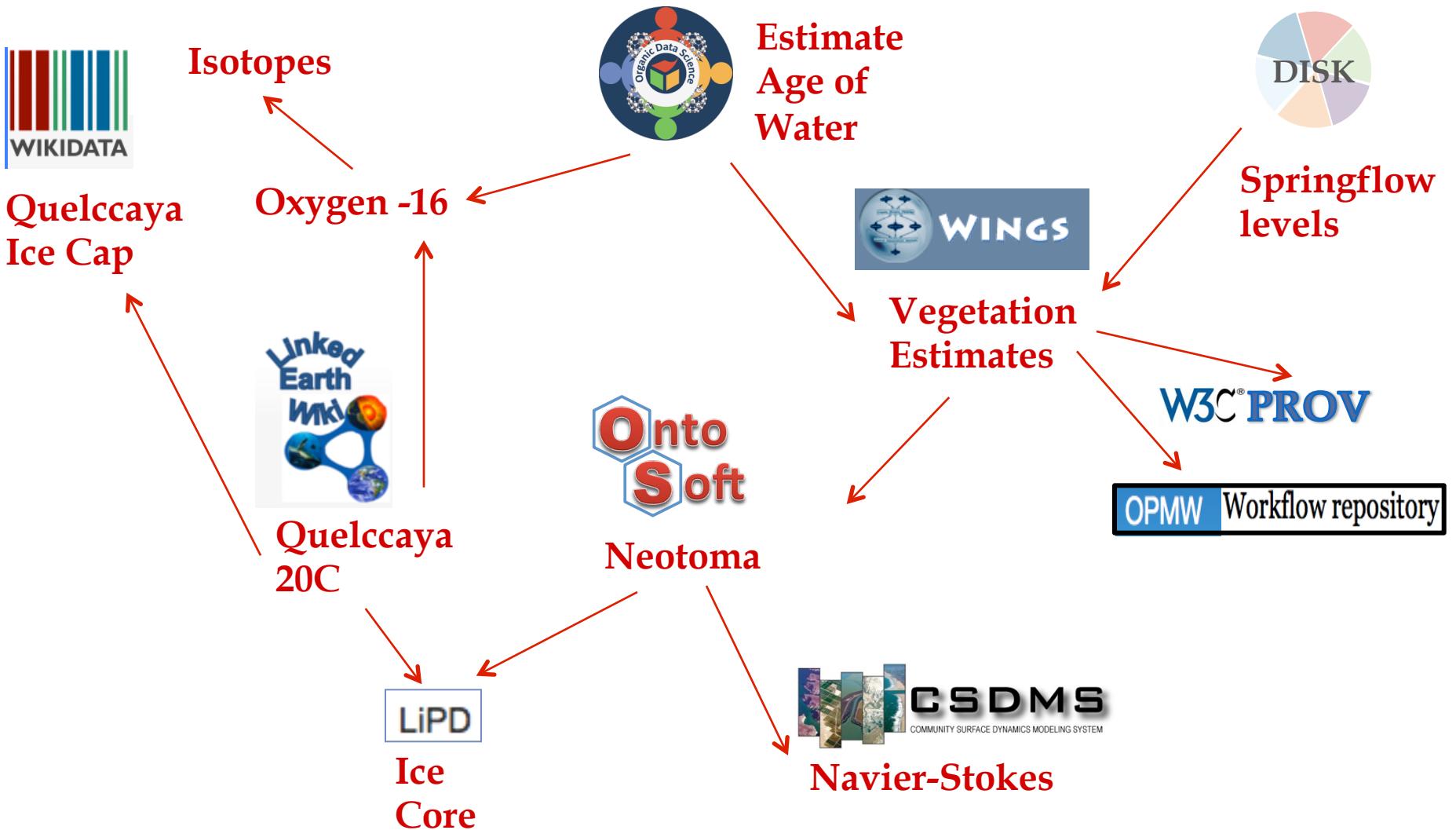


Work with P. Hanson (U Wisc) and C. Duffy (PSU)

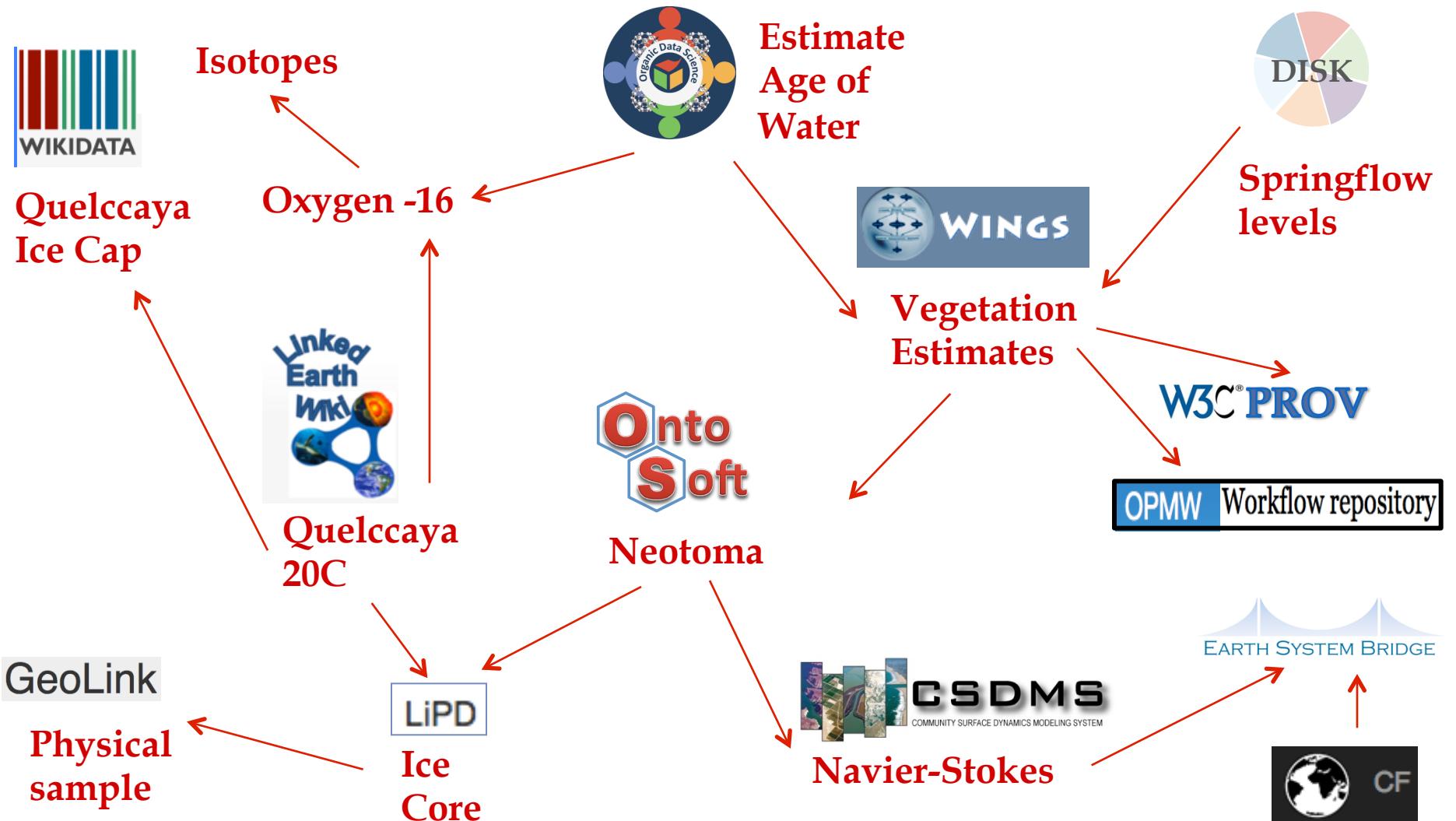
The screenshot displays a complex web-based application for managing research projects, specifically focusing on the 'Organic Data Science' framework. The interface includes:

- Header:** Shows the 'Organic Data Science' logo.
- Left Sidebar:** A navigation bar with 'All Tasks' (38), 'My Tasks' (36), a search bar ('computer science'), and a dropdown menu. It also shows a tree view of tasks under 'Write paper about the initial framework design'.
- Central Area:**
 - Task List:** A list of tasks for the 'Draft paper about the initial framework design' task, including 'Develop paper outline' (100%), 'Draft initial versions of key sections' (26%), 'Assemble first full draft of the paper' (0%), 'Collect final evaluation data' (0%), and 'Finalize writing the paper' (75%).
 - Properties Panel:** Shows detailed properties for the selected task:
 - Type: medium
 - Progress: 21%
 - Start date: 22nd Aug 2014
 - Target date: 13th Oct 2014
 - Owner: John Smith
 - Participants: James Williams, Steven Johnson
 - Expertise: computer science, collaboration
 - Text Annotation:** A note at the bottom states: 'The plan is to write a paper with some initial results of our work. If you want to contribute to a task and make sure you contribute to it with text or feedback on what other people have written.'- Right Sidebar:** A timeline and sub-task view for the same task, showing a similar list of sub-tasks and their progress.
- Bottom Right Callout:** A yellow box contains the text: 'AI opportunities:
 - collaboration
 - group formation
 - community health'
- Bottom Left Logo:** The Semantic MediaWiki logo (SMW).

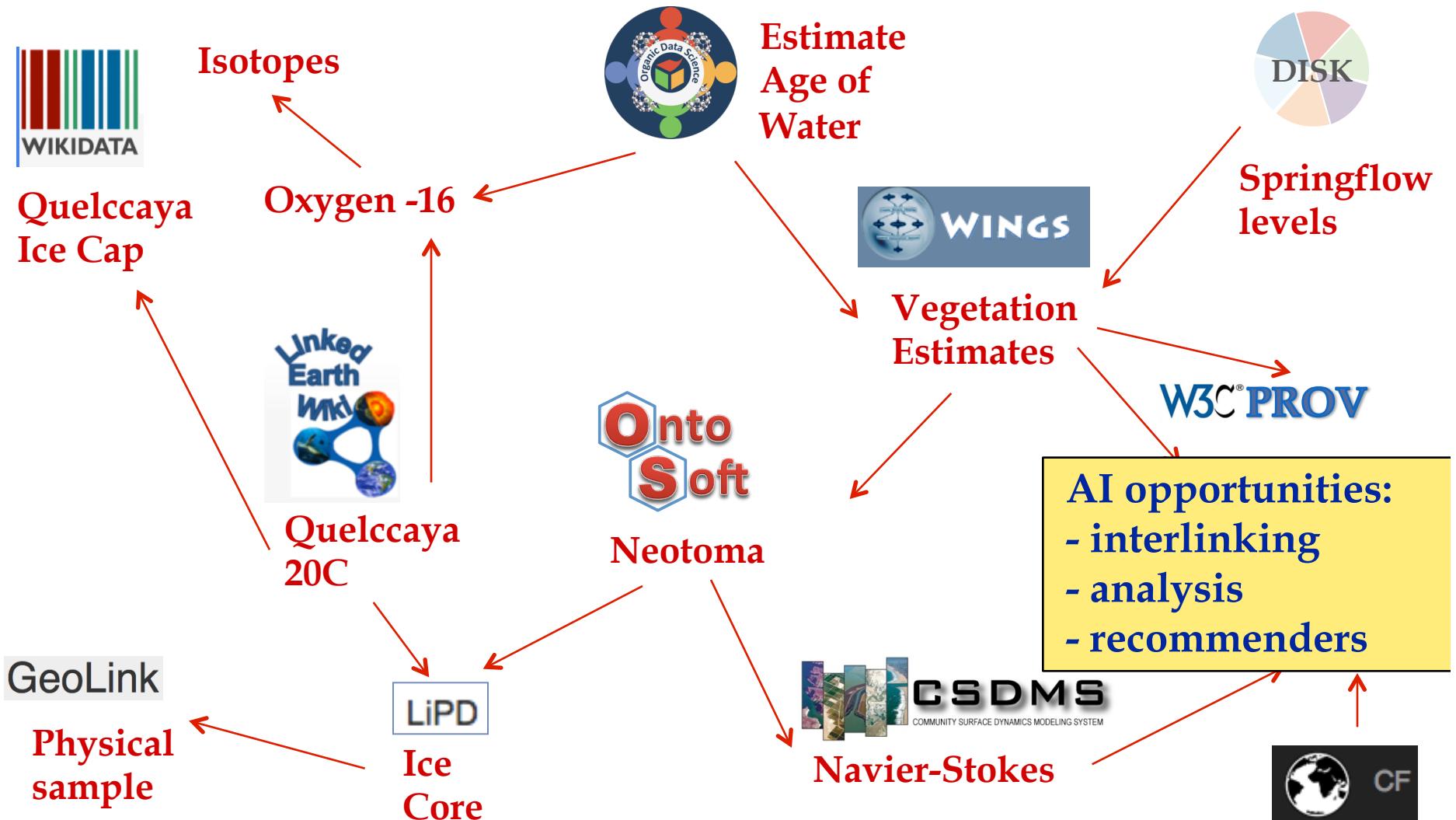
Linked Data and Linked Knowledge



Linked Data and Linked Knowledge



Linked Data and Linked Knowledge



Capturing Scientific Knowledge

Data



Software



Provenance

W3C® PROV

OPMW Workflow repository

Workflows



Meta-Workflows



Focus: Intelligent Science Assistants for Data Analysis

What is the state of the art?

What is a good problem to work on?

What is a good experiment to design?

What data should be collected?

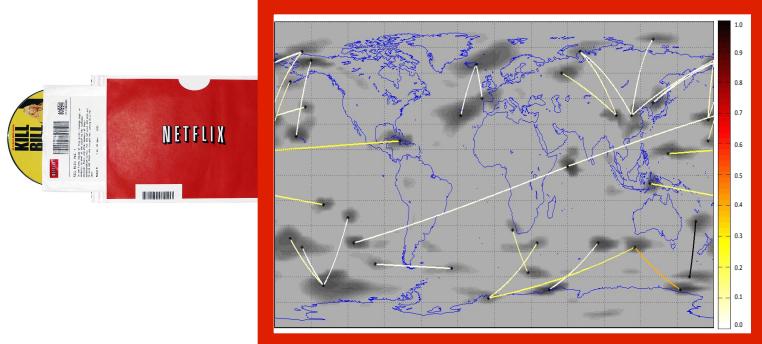
What is the best way to analyze the data?

What are the implications of the experiments?

What are appropriate revisions of current models?

AI Technologies: Use in Science

Netflix Recommenders



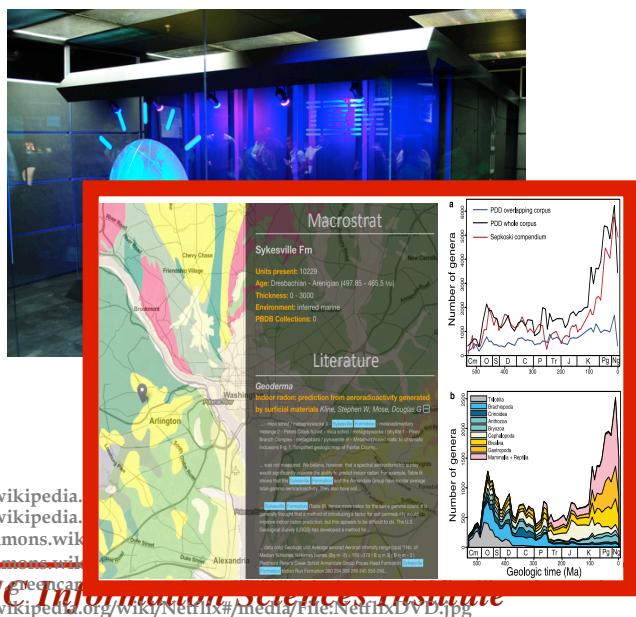
Tesla AutoPilot



RoboCup Soccer



IBM Watson



Google Knowledge Graph



Thomas Jefferson

3rd U.S. President
Thomas Jefferson was an American Founding Father, the principal author of the Declaration of Independence, and the third President of the United States. [Wikipedia](#)

Born: April 13, 1743, Shadwell, VA
Died: July 4, 1826, Charlottesville, VA
Presidential term: March 4, 1801 – March 3, 1809
Spouse: Martha Jefferson (m. 1772–1782)
Party: Democratic-Republican Party
Awards: AIA Gold Medal



Get updates about Thomas Jefferson

Keep me updated

People also search for

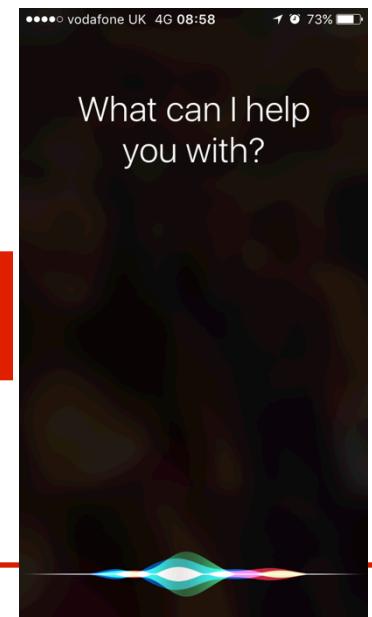


View 15+ more

[things-you-need-to-know](#)

[Feedback](#)

Apple Siri



https://en.wikipedia.org/wiki/Thomas_Jefferson
https://en.wikipedia.org/wiki/Geologic_time
<https://commons.wikimedia.org/w/index.php?title=File:Macrostrat%20Screenshot%202013-05-10%2014-45-10.jpg>

https://commons.wikimedia.org/w/index.php?title=File:IBM_Watson_at_Science_Museum_NY_2013-05-10_14-45-10.jpg
<http://www.greencar.com/2013/05/tesla-autopilot-is-a-game-changer/>
<https://en.wikipedia.org/wiki/Netflix#/media/File:NetflixDVD.jpg>



A Research Agenda for Intelligent Systems in Geosciences (<http://www.is-geo.org>)

Robotics and Sensing

Model-Driven Sensing

Optimizing collection
Unanticipated uses
Active sampling
Crowdsourcing
Virtual sensing

Machine Learning

Theory-Guided Learning

Incorporating knowledge
Combining simulation
Modeling extremes
Evaluation methodologies
Active learning

Information Integration

Trusted Threads

Distributed repositories
Threaded resources
Recommender systems
Trust and provenance
Literature extraction

Knowledge Representation & Capture

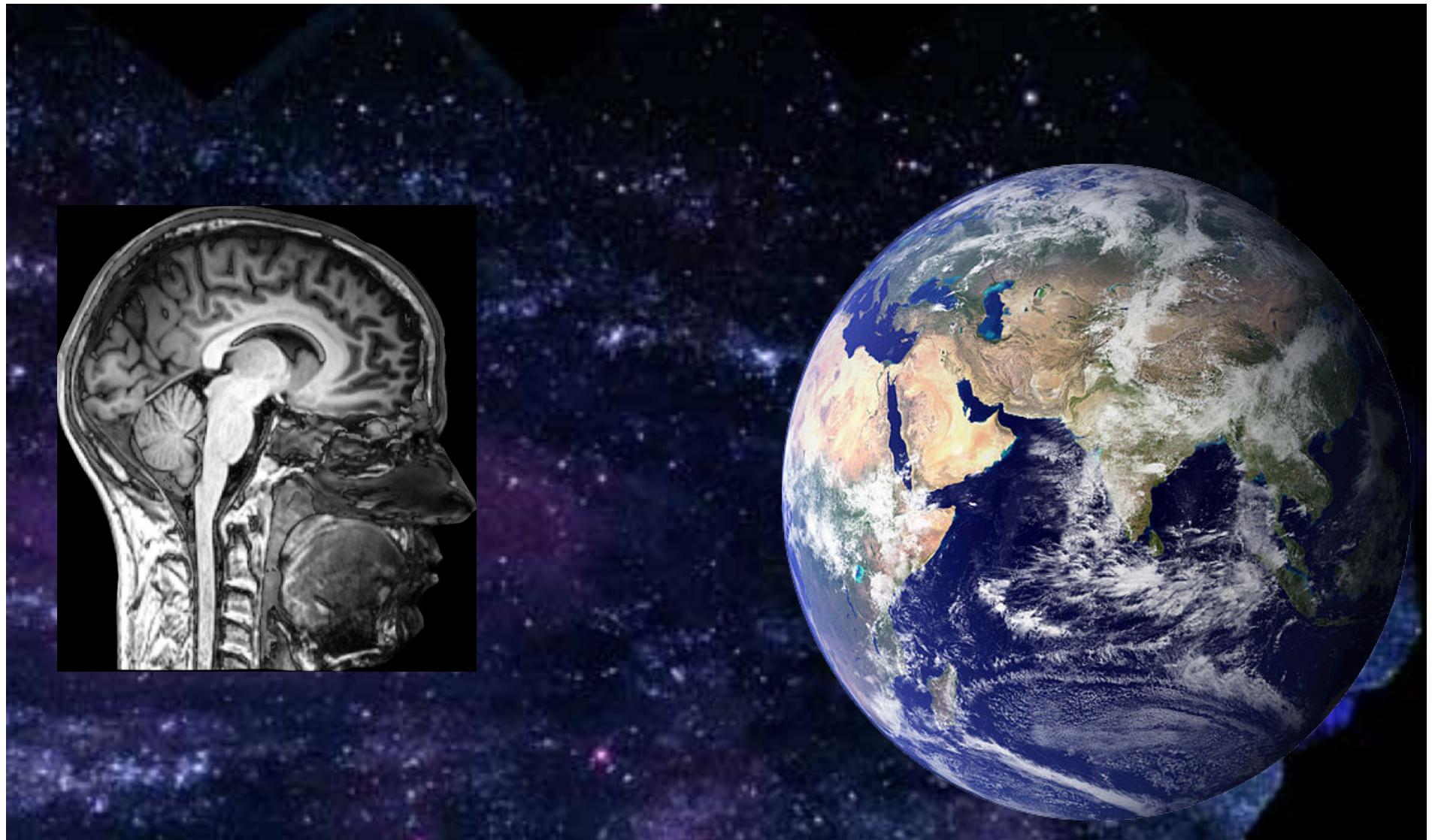
Knowledge Maps

Scientific metadata
Spatio-temporal processes
Interoperation and diversity
Assisted authoring
Automated extraction

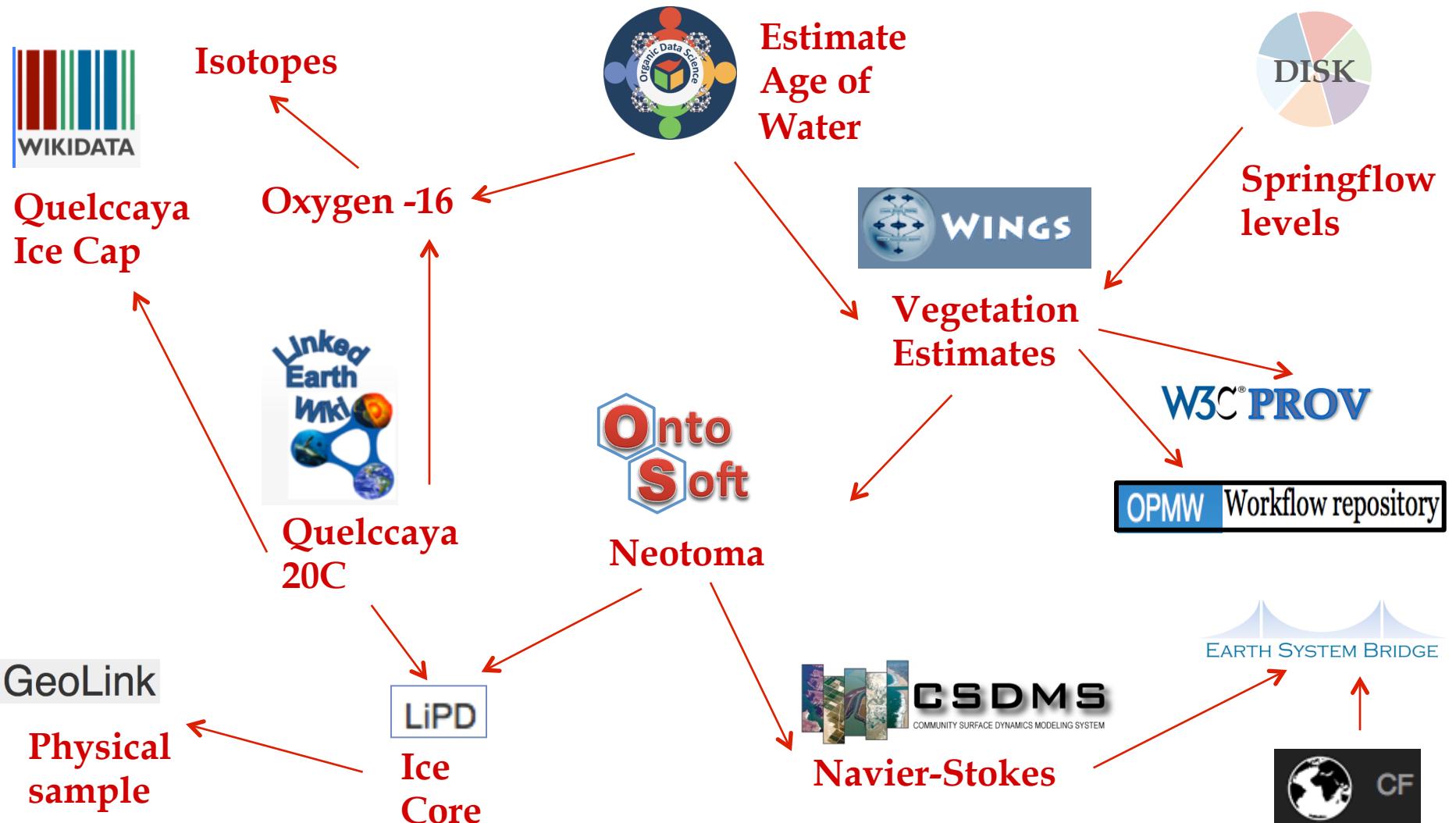
Intelligent User Interfaces

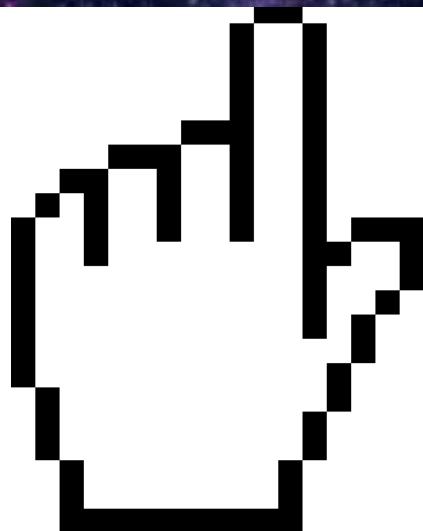
Interactive Analytics

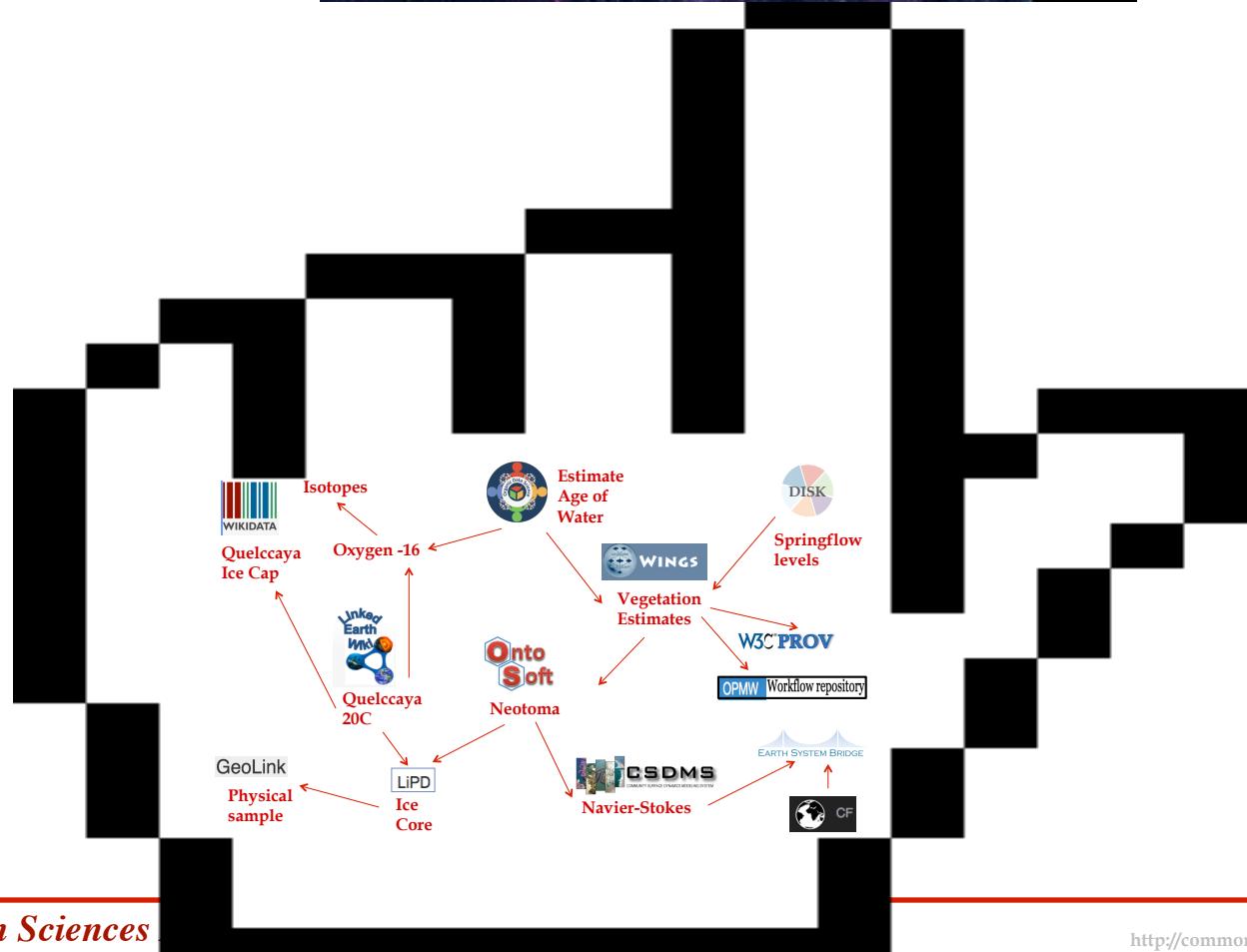
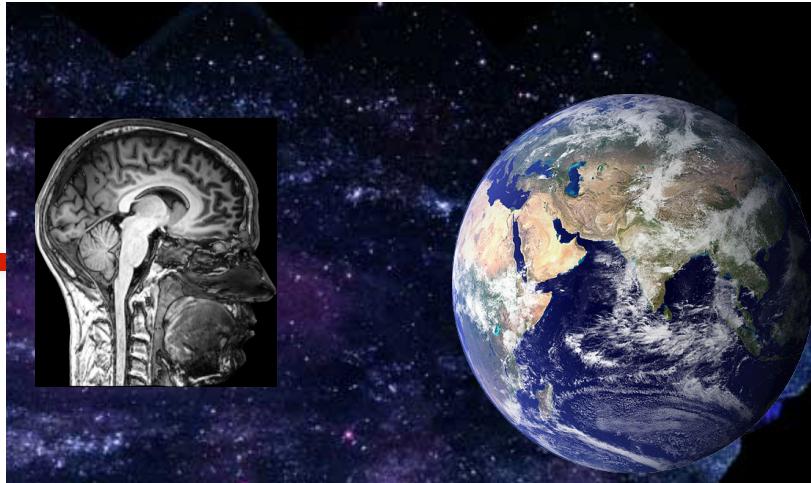
Visualization-rich processes
Automated visualizations
Immersive visualizations
Interactive model building
Spatio-temporal interfaces
Collaboration and assistance



Capture and Interlink Scientific Knowledge







Thank you!



<http://www.isi.edu/~gil>

<http://www.ontosoft.org>

<http://www.wings-workflows.org>

<http://www.organicdatascience.org>

- *Wings contributors:* Varun Ratnakar, Ricky Sethi, Hyunjoon Jo, Jihie Kim, Yan Liu, Dave Kale (USC), Ralph Bergmann (U Trier), William Cheung (HKBU), Daniel Garijo and Oscar Corcho (UPM), Pedro Gonzalez & Gonzalo Castro (UCM), Paul Groth (VUA)
- *Wings collaborators:* Chris Mattmann (JPL), Paul Ramirez (JPL), Dan Crichton (JPL), Rishi Verma (JPL), Ewa Deelman & Gaurang Mehta & Karan Vahi (USC), Sofus Macskassy (ISI), Natalia Villanueva & Ari Kassin (UTEP)
- *Organic Data Science:* Felix Michel and Matheus Hauder (TUM), Varun Ratnakar (ISI), Chris Duffy (PSU), Paul Hanson, Hilary Dugan, Craig Snortheim (U Wisconsin), Jordan Read (USGS), Neda Jahanshad (USC), Julien Emile-Geay (USC), Nick McKay (NAU)
- *Biomedical workflows:* Phil Bourne & Sarah Kinnings (UCSD), Parag Mallick (Stanford U.) Chris Mason (Cornell), Joel Saltz & Tahsin Kurk (Emory U.), Jill Mesirov & Michael Reich (Broad), Randall Wetzel (CHLA), Shannon McWeeney & Christina Zhang (OHSU)
- *Geosciences workflows:* Chris Duffy (PSU), Paul Hanson (U Wisconsin), Tom Harmon & Sandra Villamizar (U Merced), Tom Jordan & Phil Maechlin (USC), Kim Olsen (SDSU)
- *And many others!*