

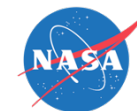
# Intelligent Systems for Scientific Discovery

**Yolanda Gil**

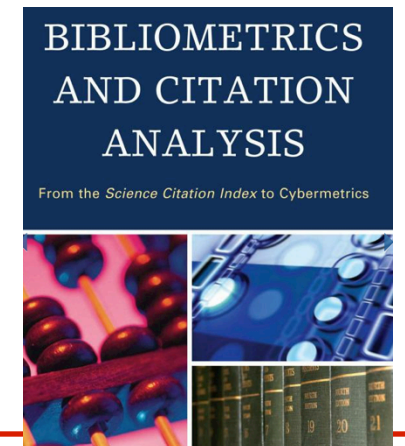
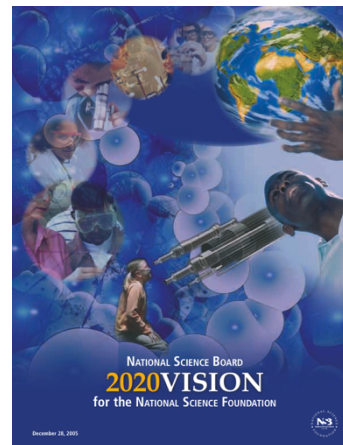
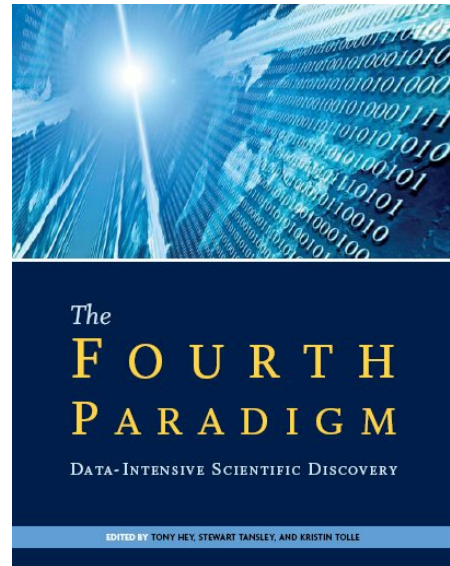
**Information Sciences Institute  
and Department of Computer Science  
University of Southern California**

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

@yolandagil  
gil@isi.edu



# Data-Intensive Computing in Science



# Artificial Intelligence and Scientific Discovery

Pittsburg Post Gazette Archives

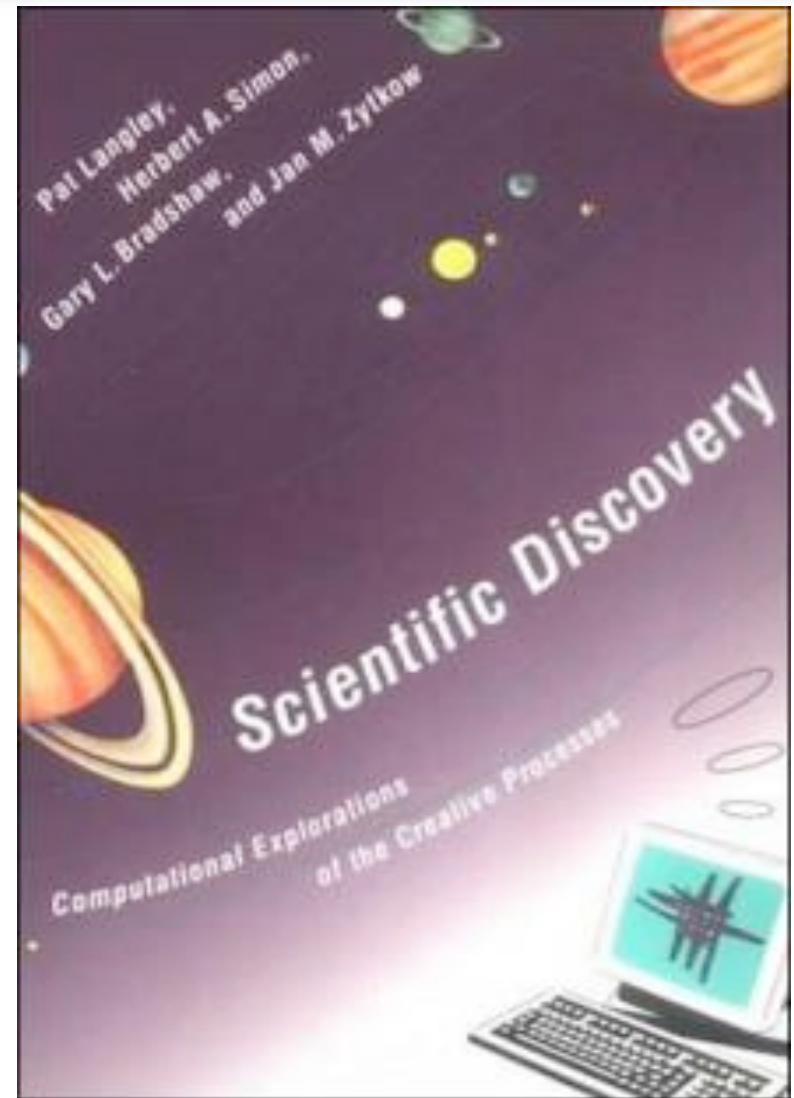




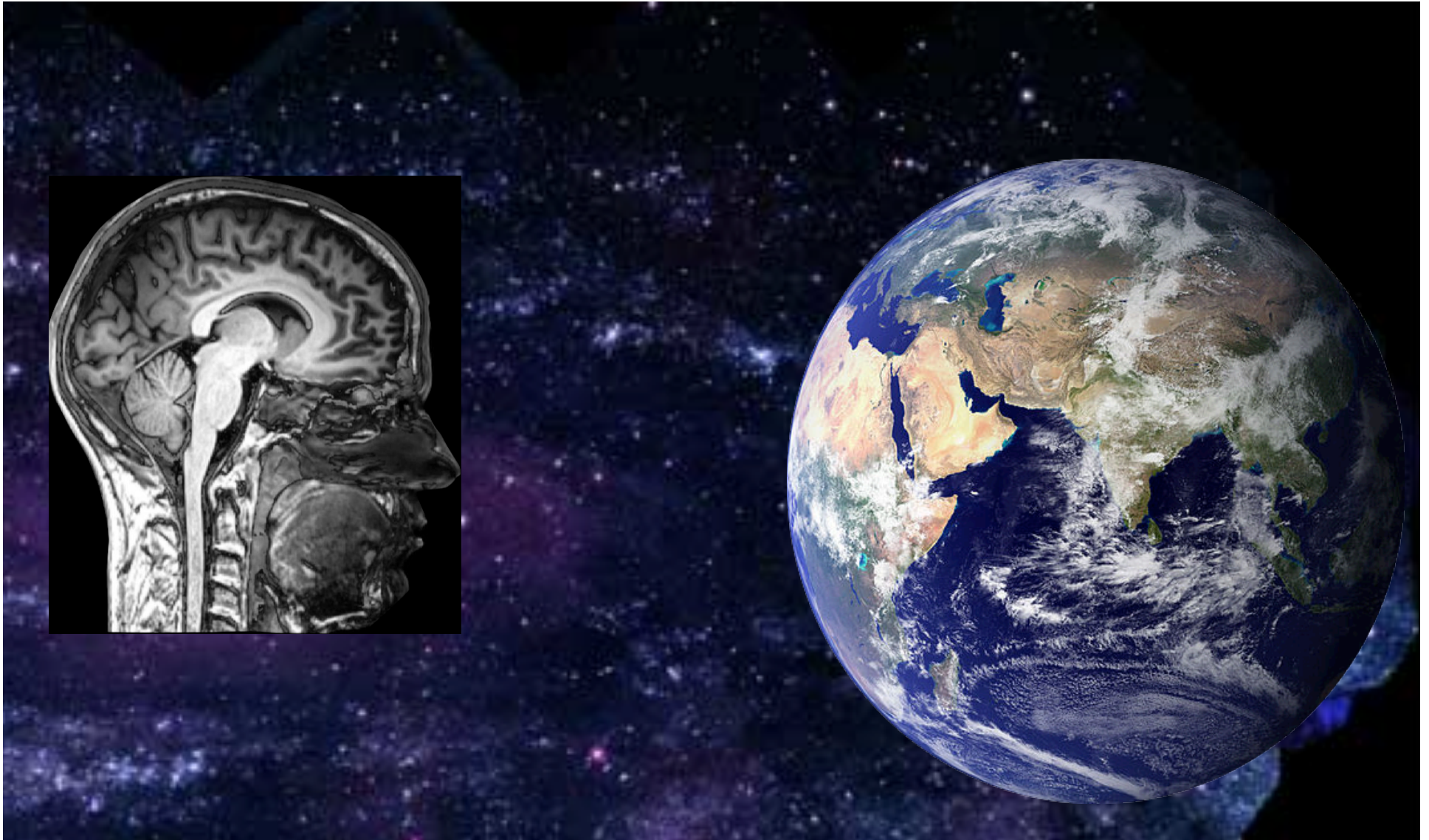
# Computational Scientific Discovery

---

- [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:MRI_brain_sagittal_section.jpg)  
[http://commons.wikimedia.org/wiki/File:Earth\\_Eastern\\_Hemisphere.jpg](http://commons.wikimedia.org/wiki/File:Earth_Eastern_Hemisphere.jpg)  
[http://www.nasa.gov/mission\\_pages/swift/bursts/uv\\_andromeda.html](http://www.nasa.gov/mission_pages/swift/bursts/uv_andromeda.html)

# AI's Coming of Age

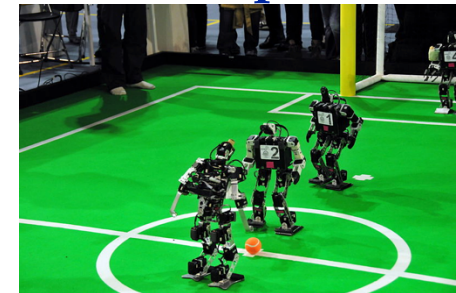
## Netflix Recommenders



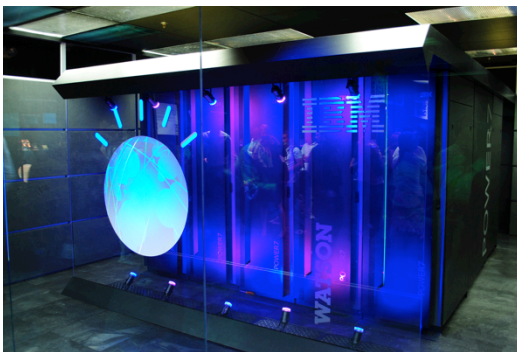
## Tesla AutoPilot



## RoboCup Soccer



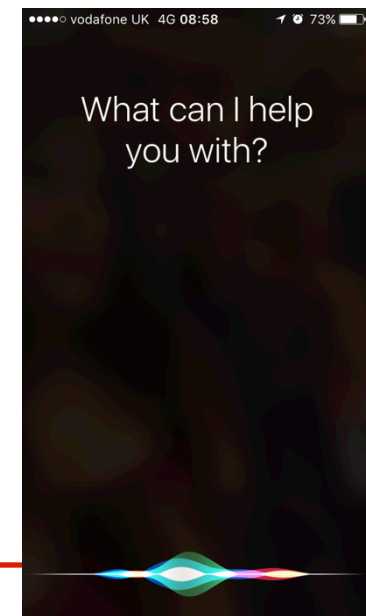
## IBM Watson



## Google Knowledge Graph



## Apple Siri



[https://en.wikipedia.org/wiki/Watson\\_\(computer\)#/media/File:IBM\\_Watson.PNG](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:Google_Knowledge_Panel.png)  
<https://commons.wikimedia.org/wiki/File:13-06-28-robocup-sindharon-005.jpg>  
<http://www.rencapreports.com/news/100482-tesla-autopilot-the-10-most-important-things-you-need-to-know>  
<https://en.wikipedia.org/wiki/Netflix#/media/File:NetflixDVD.jpg>



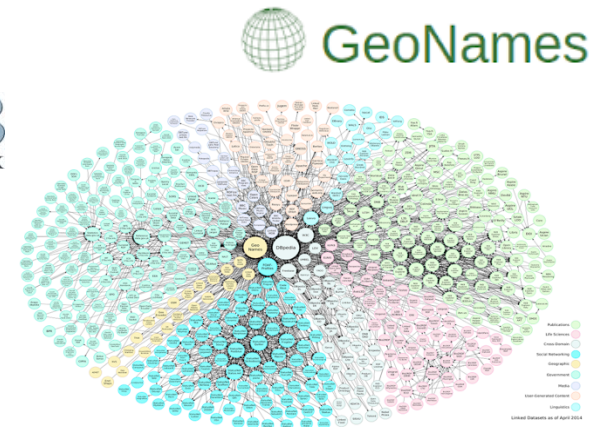
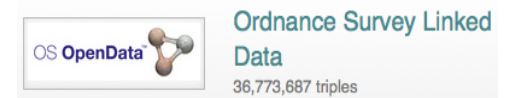
# Before There Was the Knowledge Graph...

## Google Knowledge Graph (2012)



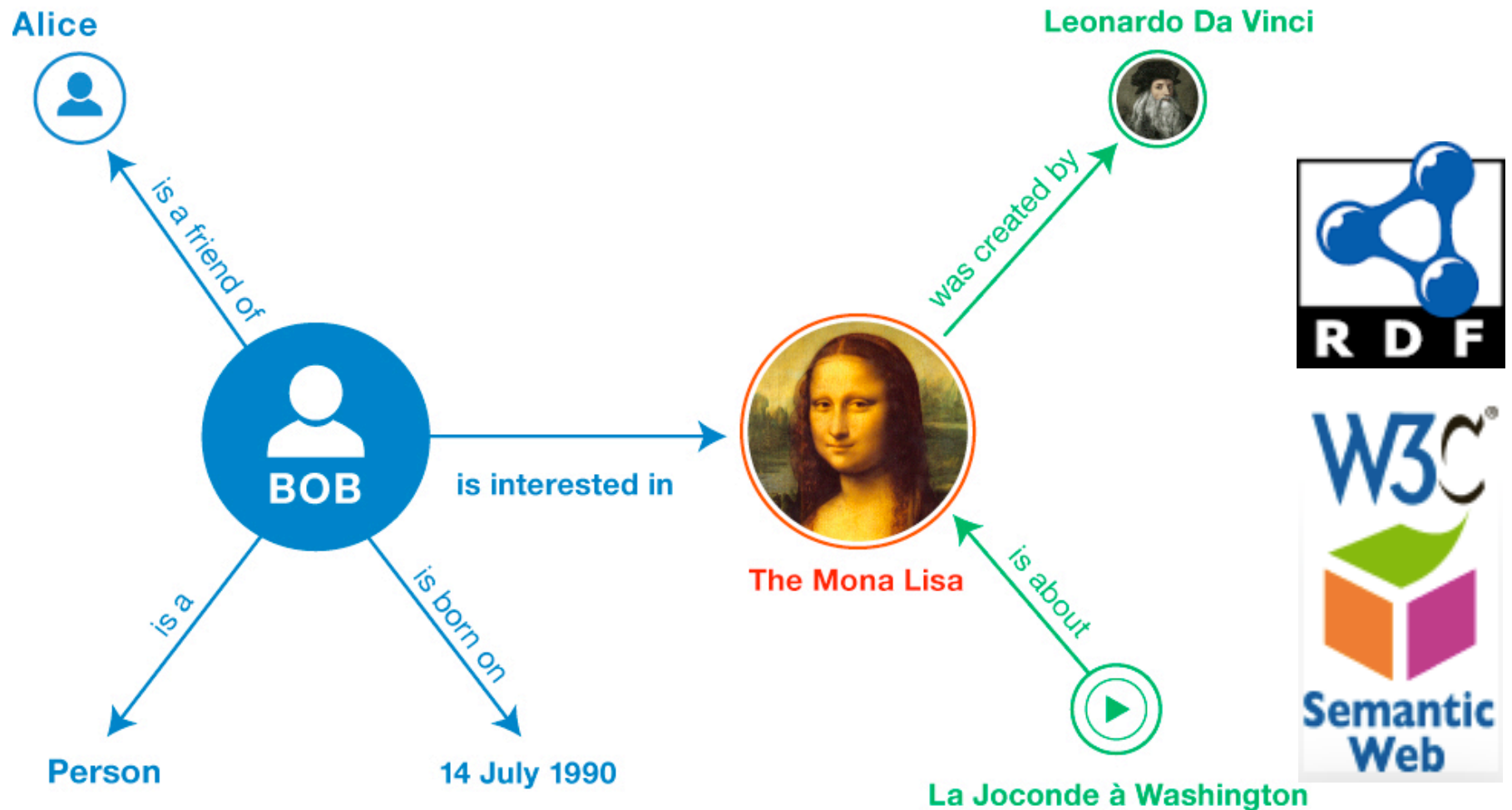
A screenshot of the Google Knowledge Graph for Thomas Jefferson. It features a grid of eight portrait images of Jefferson at the top. Below the images is the name "Thomas Jefferson" and his title "3rd U.S. President". A short biographical paragraph follows, mentioning his role as a Founding Father and author of the Declaration of Independence. Key facts are listed: 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. There is a "Keep me updated" button and a "People also search for" section with portraits of John Adams, George Washington, Benjamin Franklin, James Madison, and Alexander Hamilton.

## Linked Data (2007)



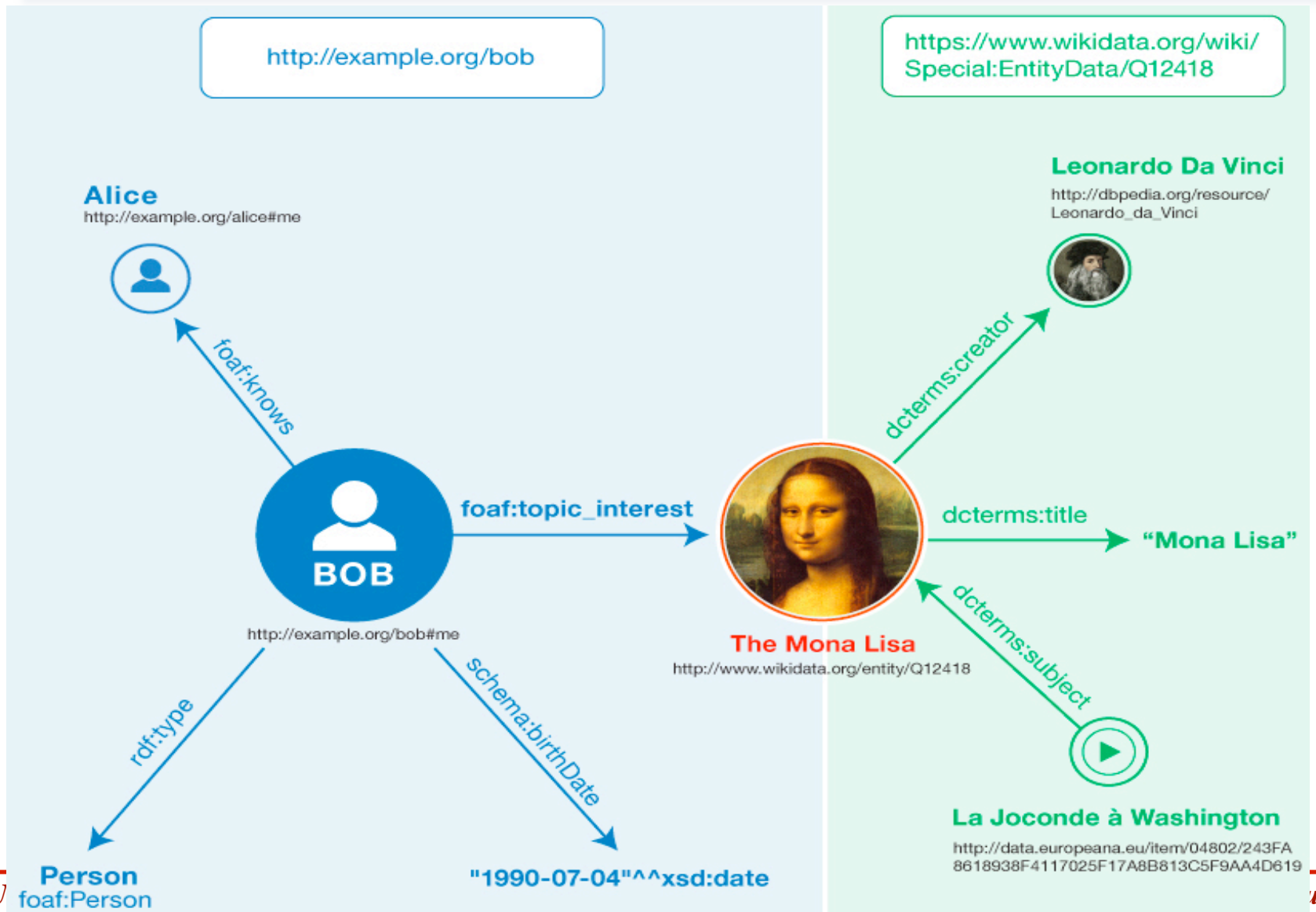


# Giving Meaning to Hyperlinks on the Web



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

# The Semantic Web



# Data and Ontologies on the Semantic Web

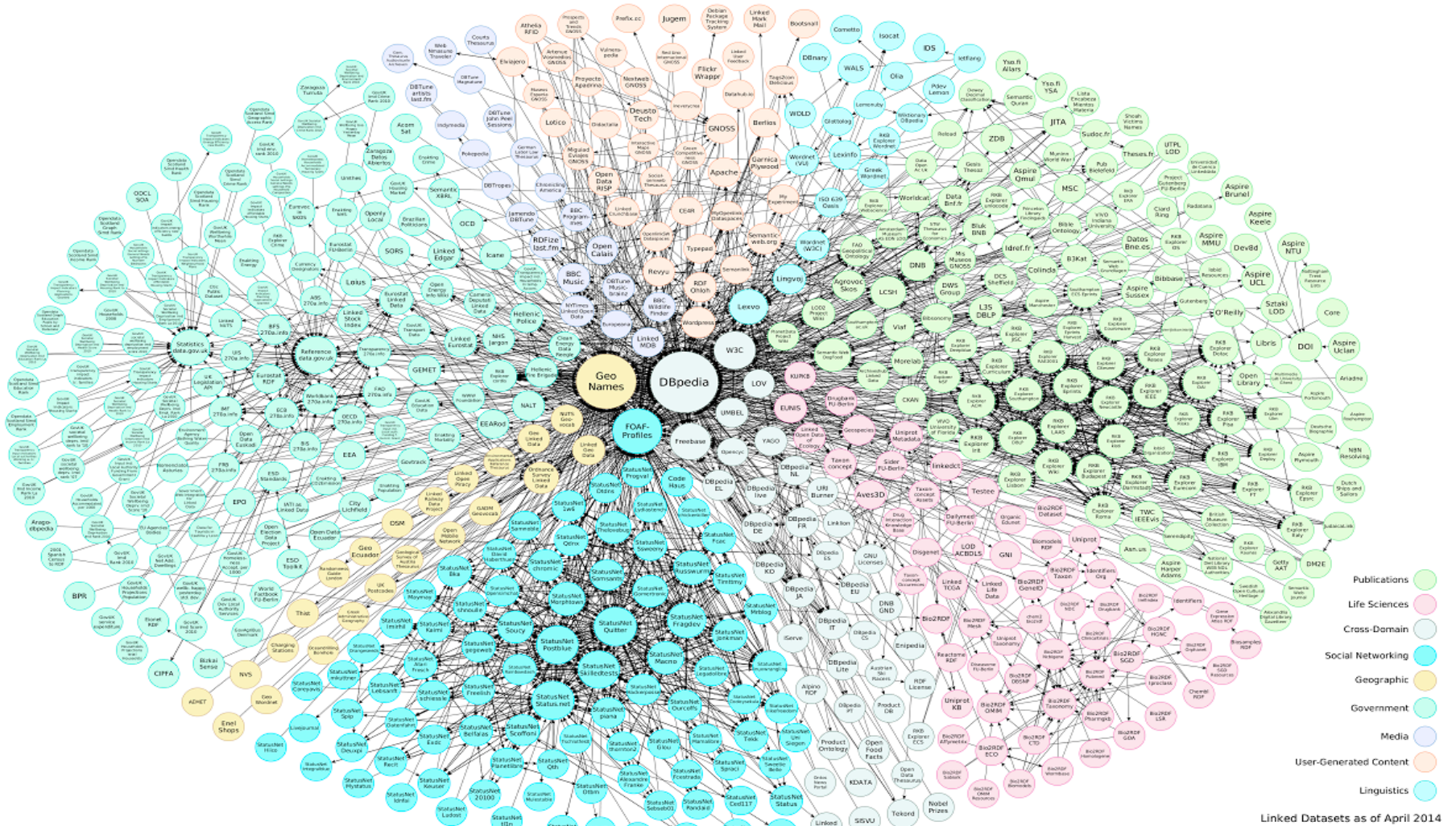
---

<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**

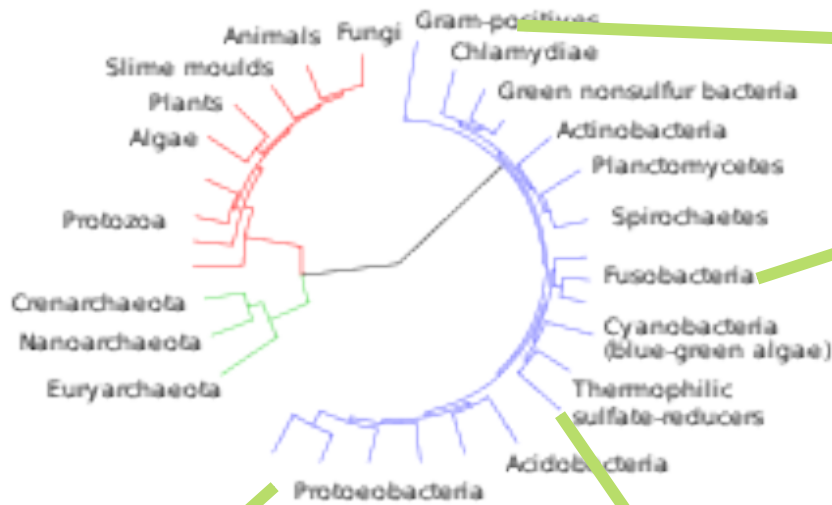
**FOAF: from 27% to 59%**

**DC: from 31% to 56%**

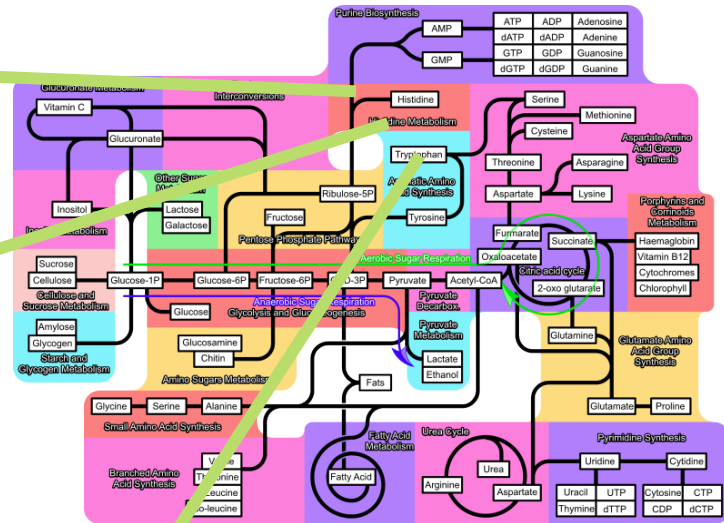
<http://lod-cloud.net>  
<http://stats.lod2.eu>

# Interlinking Scientific Knowledge

## Taxonomical



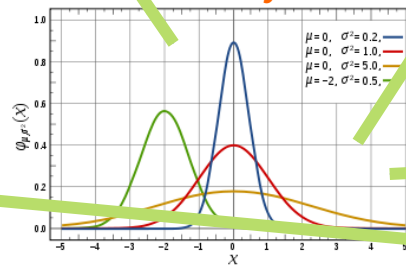
## Networks



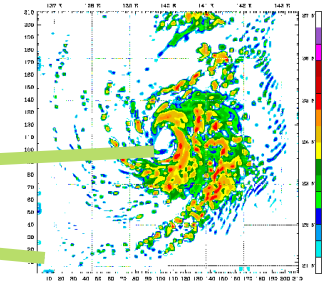
## Mathematical

$$E_r = \sqrt{(m_0c^2)^2 + (pc)^2}$$

## Bayesian



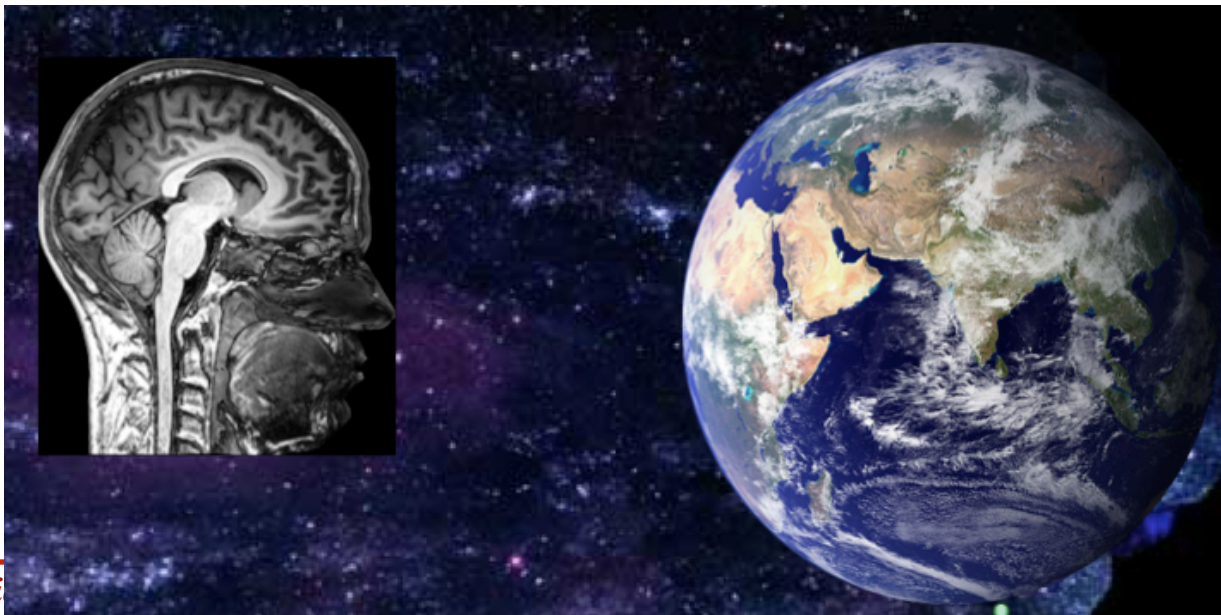
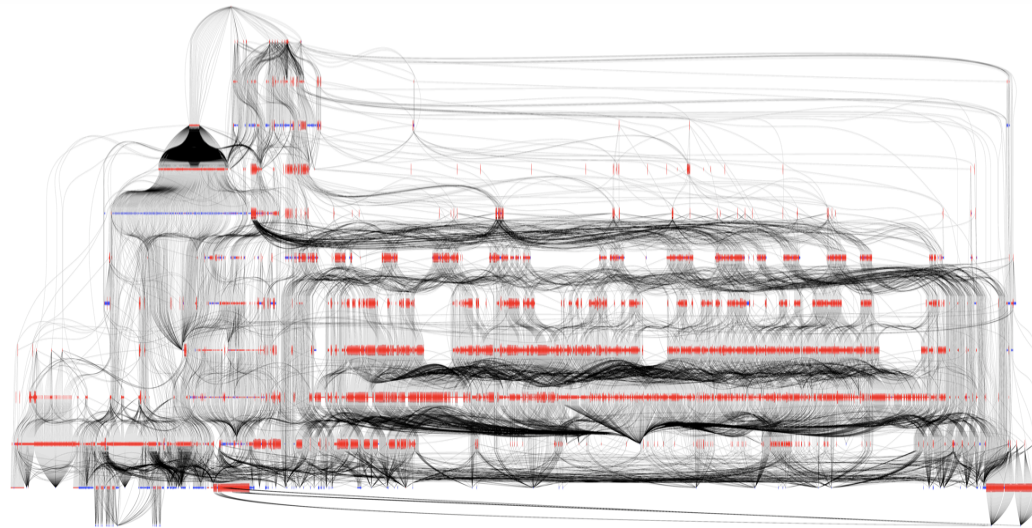
## Simulations





# 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<sup>®</sup> PROV

OPMW Workflow repository

Meta-Workflows

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]

### Palmyra coral 20C

Data

• DOWNLOAD

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

Structured Properties

[x] SiteName	Palmyra	(By Julien)
[x] Archive	Coral	(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 Nick)
[x] MeasurementMaterial		
[x] MeasurementStandard		
[x] MeasurementUnits		
[x] Reference		
[x] Species		

Credits

Users who have contributed to this Page:

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

### Porites

Structured Properties


add fact

[x] Property:Name	Topic:Finger Coral	[hide]
	<ul style="list-style-type: none"> <li>[x] <a href="http://dbpedia.org/resource/Porites">http://dbpedia.org/resource/Porites</a></li> <li>[add source]</li> </ul>	

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



Web of Data

[add source]

Porites	<sameAs>	dbpedia
hide		
class		Coral
classis		Coral
familia		Poritidae
family		Poritidae
genus		Porites
kingdom		Animal
name		Finger Coral
order		Funglina
order		Scleractinia
ordo		Scleractinia
phylum		Cnidaria

### Geochemistry datasets

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

```

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

**AI opportunities:**

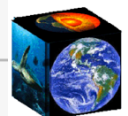
- collection
- normalization
- organization



```

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

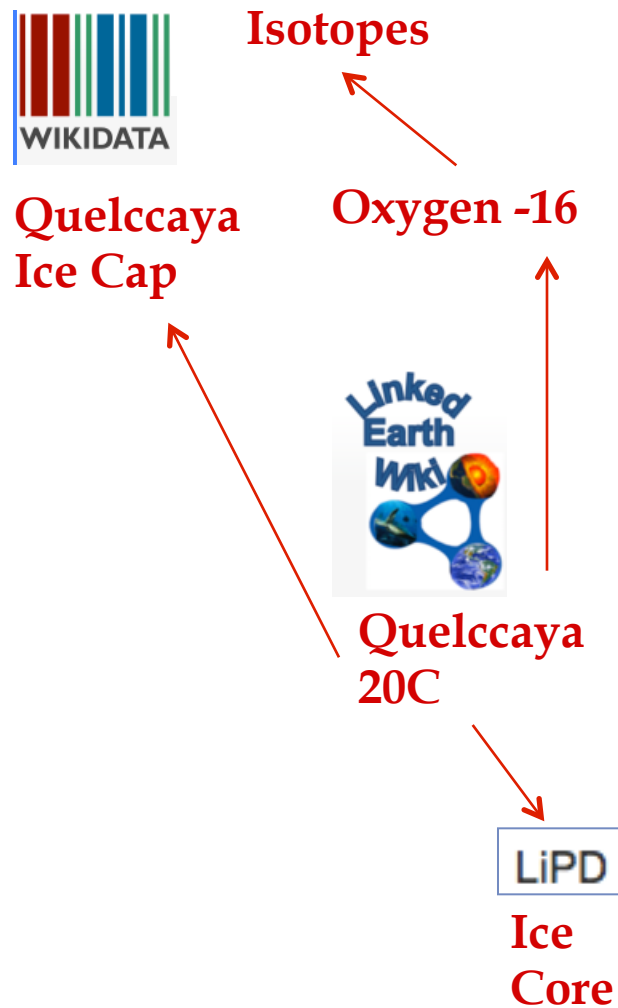
Yolo





# Linked Data and Linked Knowledge

---



# Capturing Scientific Knowledge

---

Data



Software



Provenance

W3C<sup>®</sup> PROV

OPMW Workflow repository

Meta-Workflows

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

---

**Locate**  
unique description

---



**Experiment**  
run with other data

---

**Compose**  
run with other software

---

**Cite**  
scientific publications

---

**Trust**  
quality and ratings

---

**Relate**  
domain knowledge

---

**Access**  
download

---

**Install**  
execution requirements

---

**Run**  
testing execution

---



# Knowledge About Software: Physical Variables and Assumptions



The screenshot shows the Software Ontology Editor (SOE) interface. On the left is a project tree with the following structure:

- Software
  - 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


The main window displays the 'Describe Software' tab for the 'PIHM' project. It includes a 'Standard Names' table with the following data:

Object	Quantity	Operators
<input type="checkbox"/> Object	Quantity	Operators
<input type="checkbox"/> air	relative_humidity	
<input type="checkbox"/> air	temperature	
<input type="checkbox"/> air_water_vapor	partial_pressure	
<input type="checkbox"/> atmosphere_water	precipitation_rate	
<input type="checkbox"/> atmosphere_water_vapor	partial_pressure	
<input type="checkbox"/> ground_water_table	depth	
<input type="checkbox"/> land_snow	melt_rate	
<input checked="" type="checkbox"/> land_surface	None	



# OntoSoft: Comparing Software Implementations



 <a href="#">Software</a> <a href="#">Community</a> <a href="#">Training</a>				
<h2>Compare Software</h2> <p>DrEICH algorithm, PIHM, PIHMgis, TauDEM, WBMsed</p>				
PIHM	PIHMgis	DrEICH	TauDEM	WBMsed
<b>What are domain specific keywords for this software ? (eg: hydrology, climate)</b>				
Geomorphology, Hydrological, Bedrock channel ero-	Basins, Continental	Basins, GIS	Hydrologically corrected DEM, Watershed	Sediment flux, Global model, Hydrological model
<b>What Operating Systems can the software run on ?</b>				
Unix Linux	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Linux
<b>Is there any test data available for the software ?</b>				
<b>Test Data Location:</b> <a href="http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full">http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full</a> <b>Test Data Description:</b> Two test DEMs are included in the repository,	<b>Test Data Location:</b> <a href="http://sourceforge.net/projects/pihmmodel/">http://sourceforge.net/projects/pihmmodel/</a> <b>Test Data Description:</b> Upper Juniata River 875 km <sup>2</sup> : see: <a href="http://source-">http://source-</a>		<b>Test Data Location:</b> <a href="http://csdms.colorado.edu/wiki/Model:TauDEM#Testing">http://csdms.colorado.edu/wiki/Model:TauDEM#Testing</a> <b>Test Data Description:</b> The Logan River DEM is a small test dataset useful	<b>Test Data Location:</b> <a href="http://csdms.colorado.edu/wiki/Model:WBMsed#Testing">http://csdms.colorado.edu/wiki/Model:WBMsed#Testing</a> <b>Test Data Description:</b> Extensive input dataset is available on the CSDMS

# OntoSoft: Publishing Software Metadata as RDF



The screenshot shows the OntoSoft web interface for the PIHM software. The header includes the OntoSoft logo and a menu icon. The main content area displays the software name "PIHM" and the author "[Christopher Duffy]". A navigation bar contains three buttons: "HTML", "RDF/XML", and "JSON", which are circled in red. To the right of these buttons is a "RATE" button. Below the navigation bar, the "Identify" section is visible, followed by the "Locate" section with the subtitle "Unique description". Two search prompts are shown: "What is the software called?" and "What is a short description for this software?". The first prompt has a result: "PIHM". The second prompt has a result: "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".

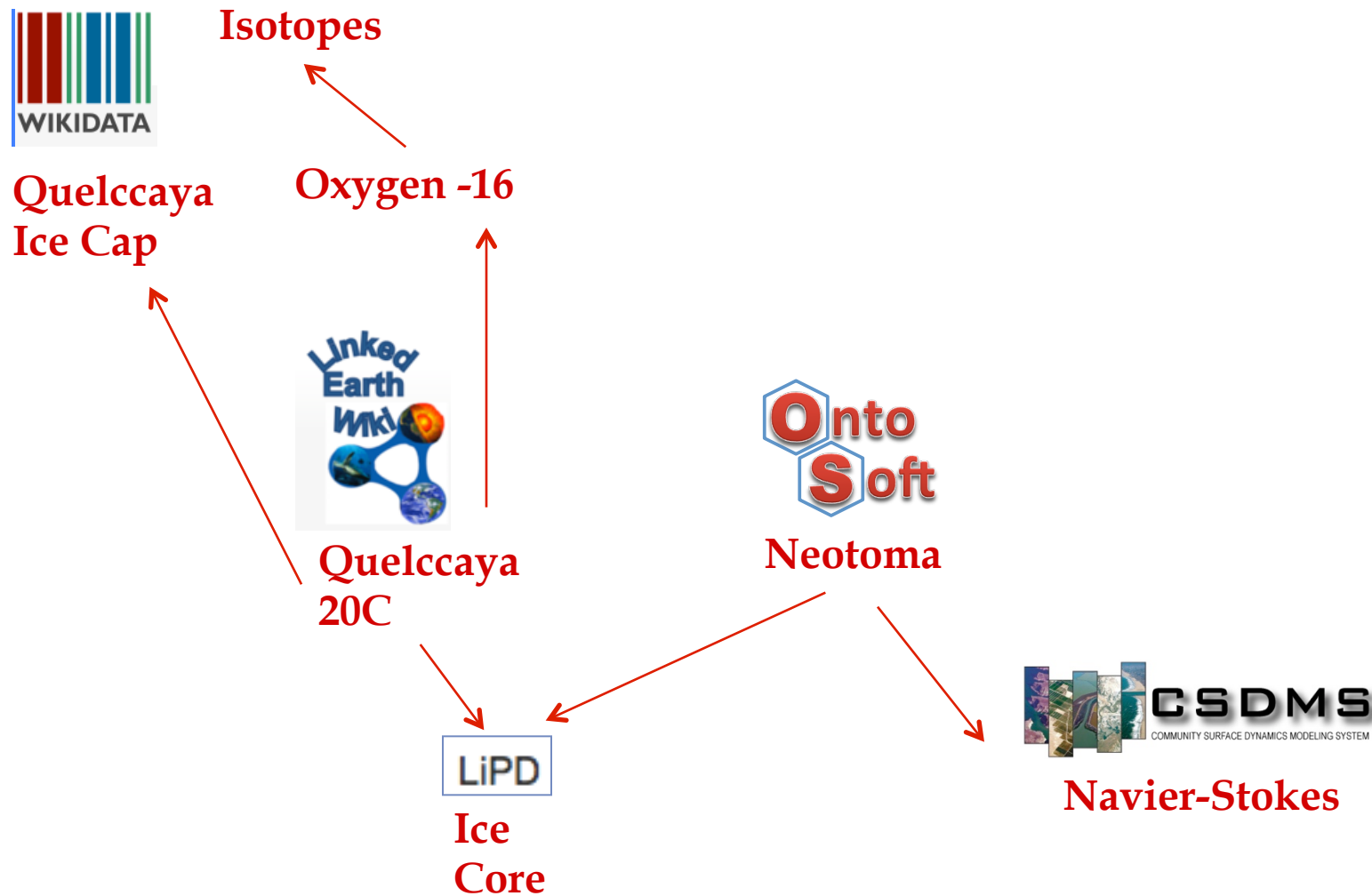
**AI opportunities:**

- functional desc.
- organization
- linking to data



# Linked Data and Linked Knowledge

---



# Capturing Scientific Knowledge

---

Data



Software



Provenance

W3C<sup>®</sup> PROV

OPMW Workflow repository

Meta-Workflows

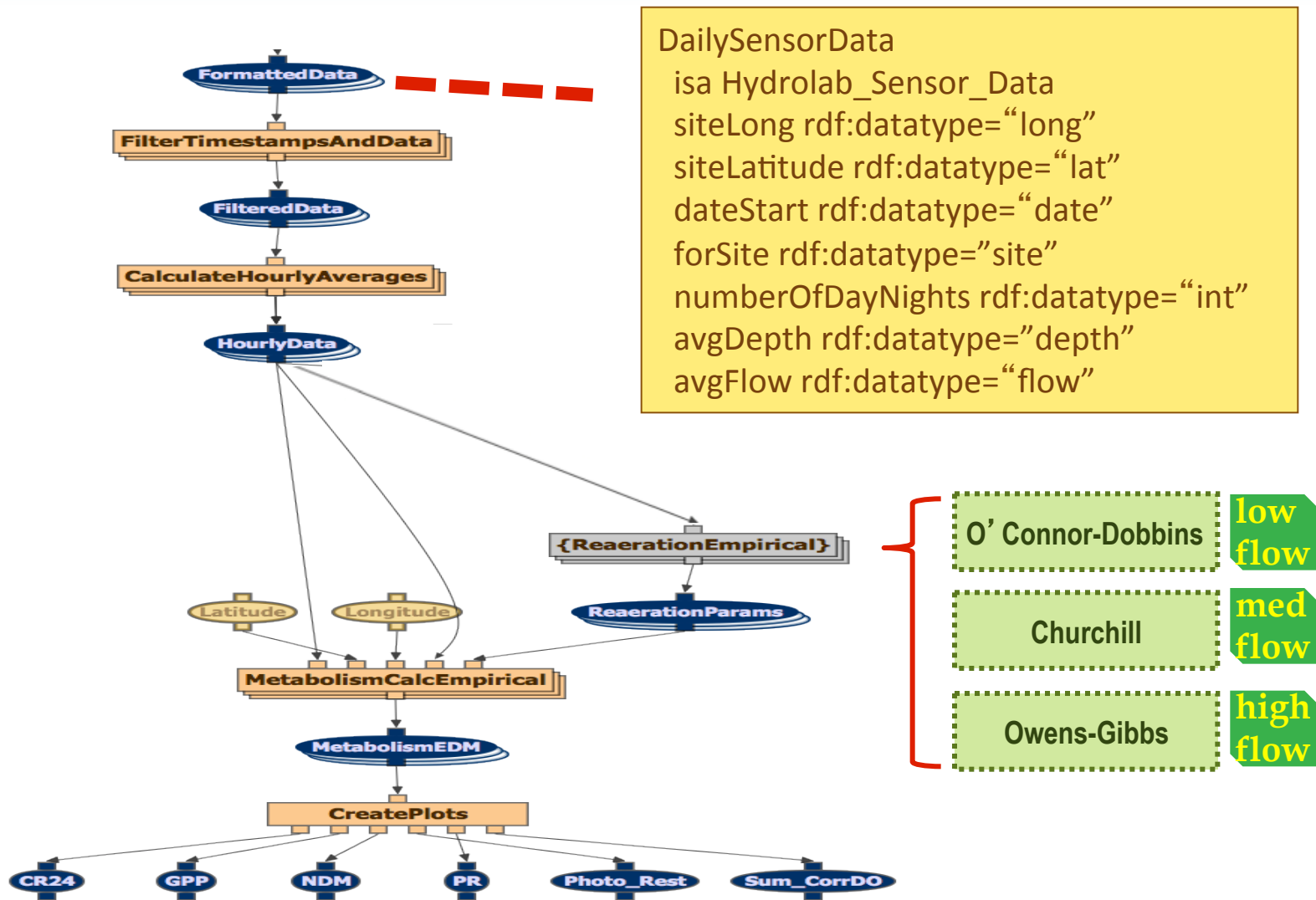
Workflows



# Knowledge about Data Analysis: WINGS

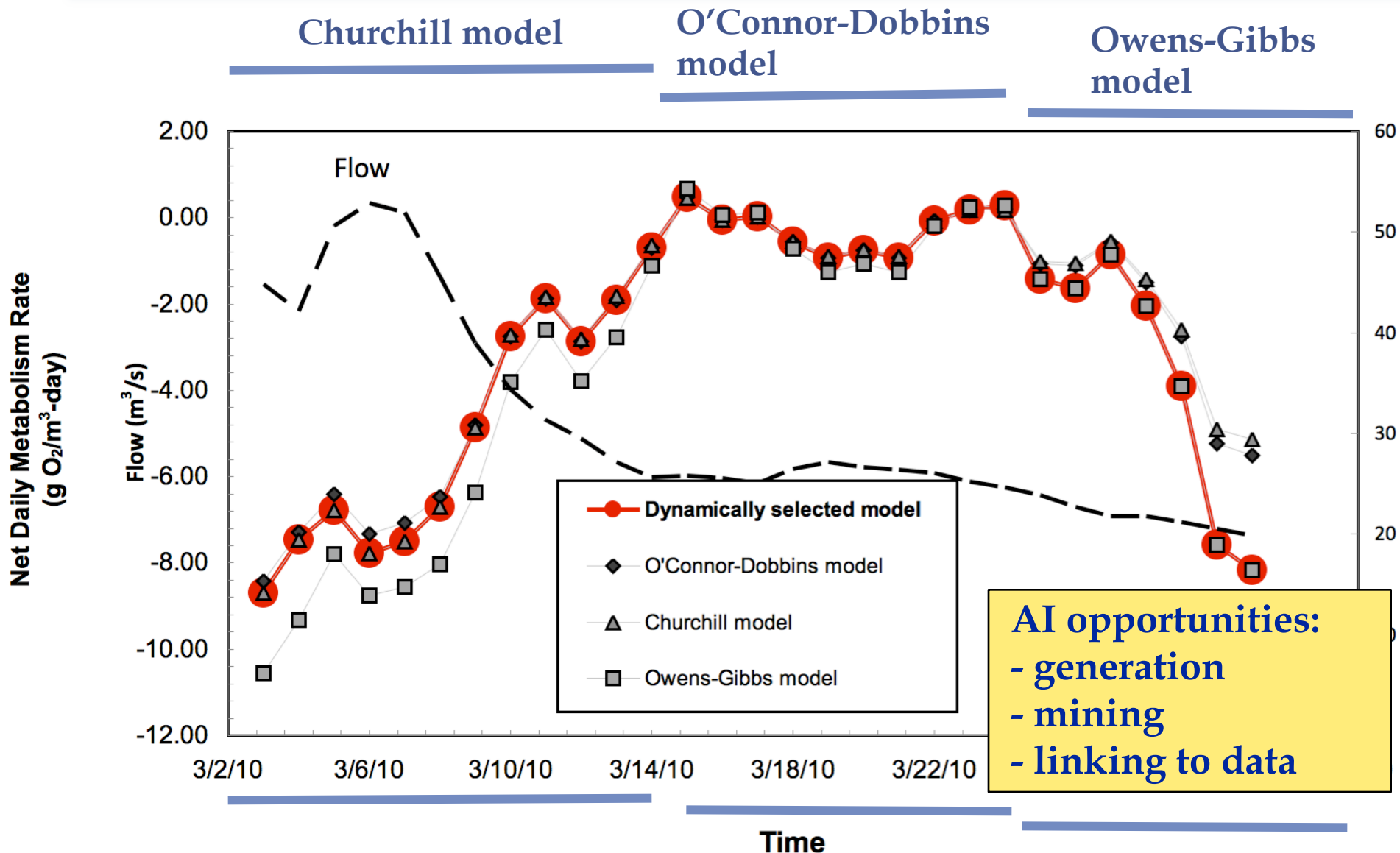


*Work with V. Ratnakar (USC)*





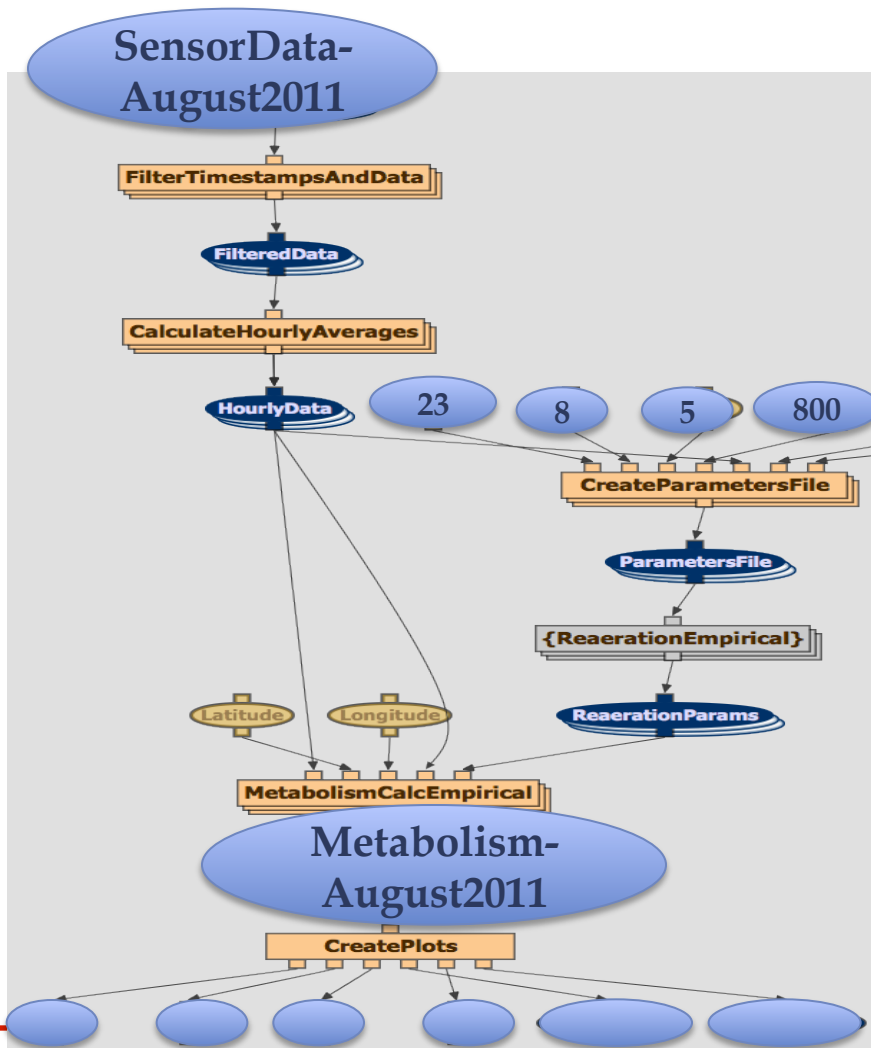
# WINGS Dynamically Customizes the Workflow Based on Daily Sensor Readings



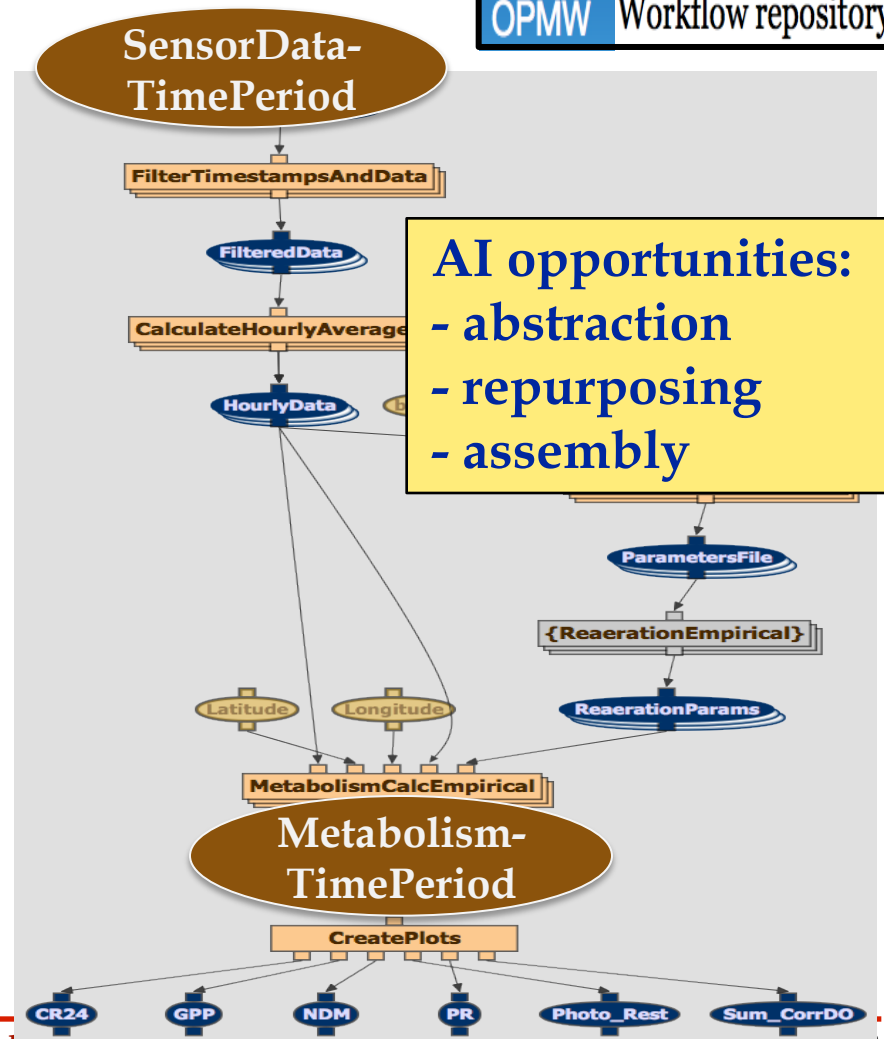
# Describing Execution (Provenance) vs General Method (Workflow)

W3C<sup>®</sup> PROV

OPMW Workflow repository



USC Information Sciences Institute

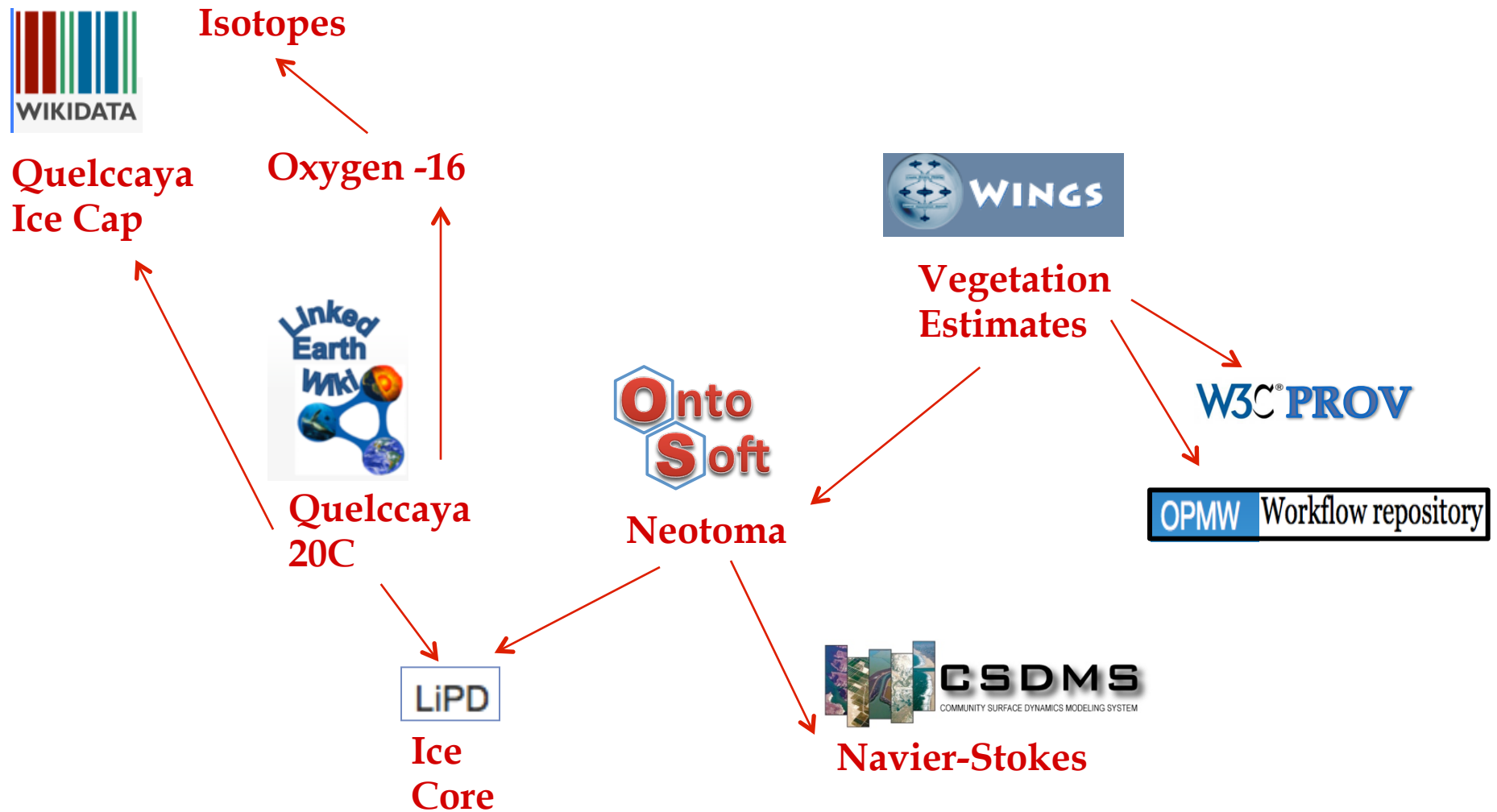


Yolanda Gu

gu@isi.edu

AI opportunities:  
 - abstraction  
 - repurposing  
 - assembly

# Linked Data and Linked Knowledge



# Capturing Scientific Knowledge

---

Data



Software



Provenance

W3C<sup>®</sup> PROV

OPMW Workflow repository

Meta-Workflows

Workflows

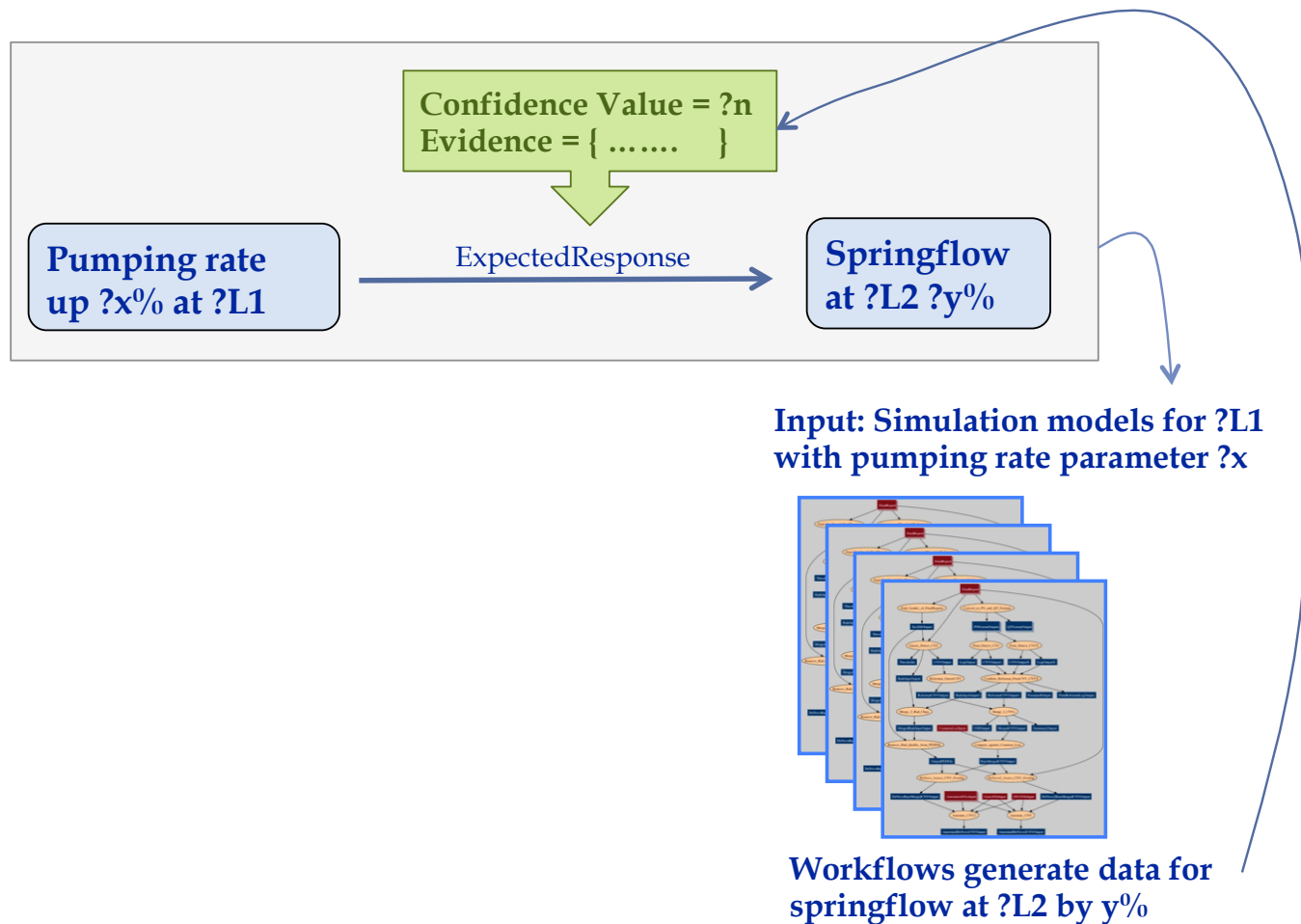




# Knowledge about Meta-Processes: DISK



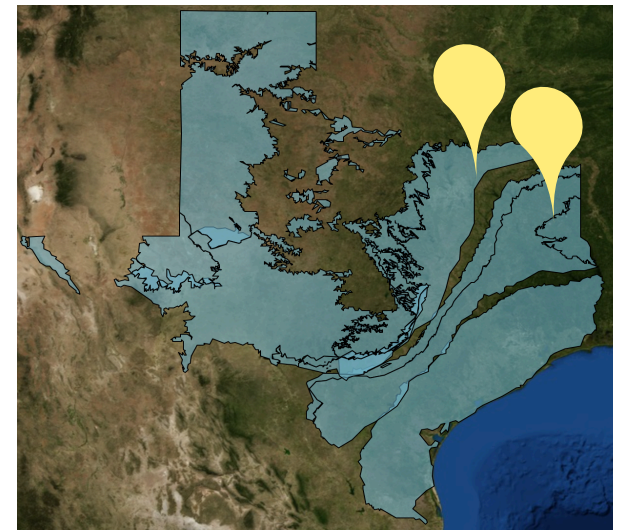
*Work with P. Mallick (Stanford U) and S. Pierce (UT Austin)*



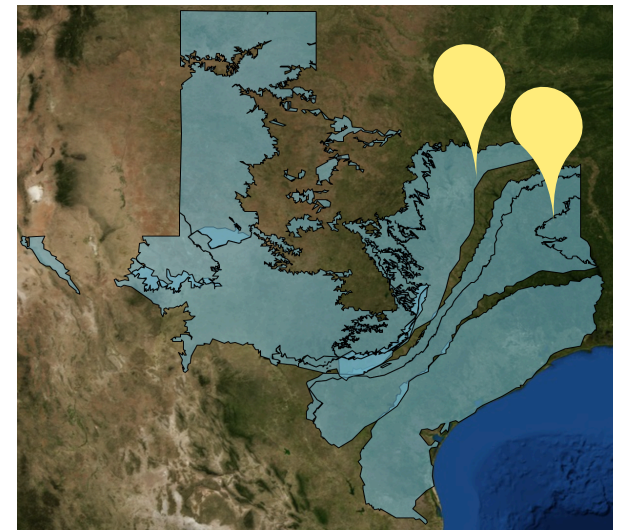
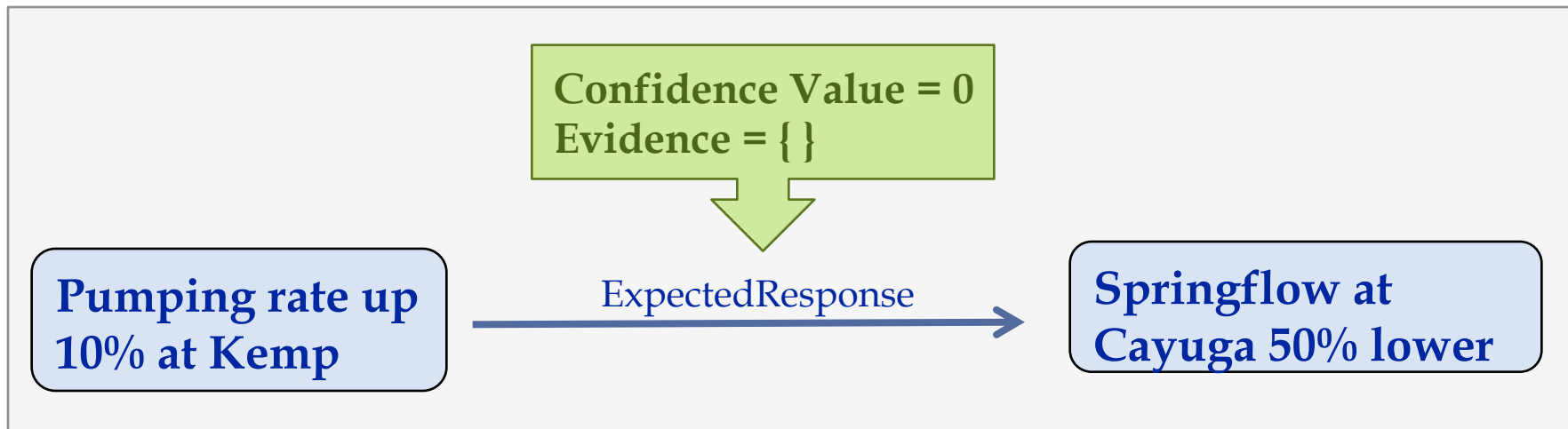
# DISK: Hypotheses



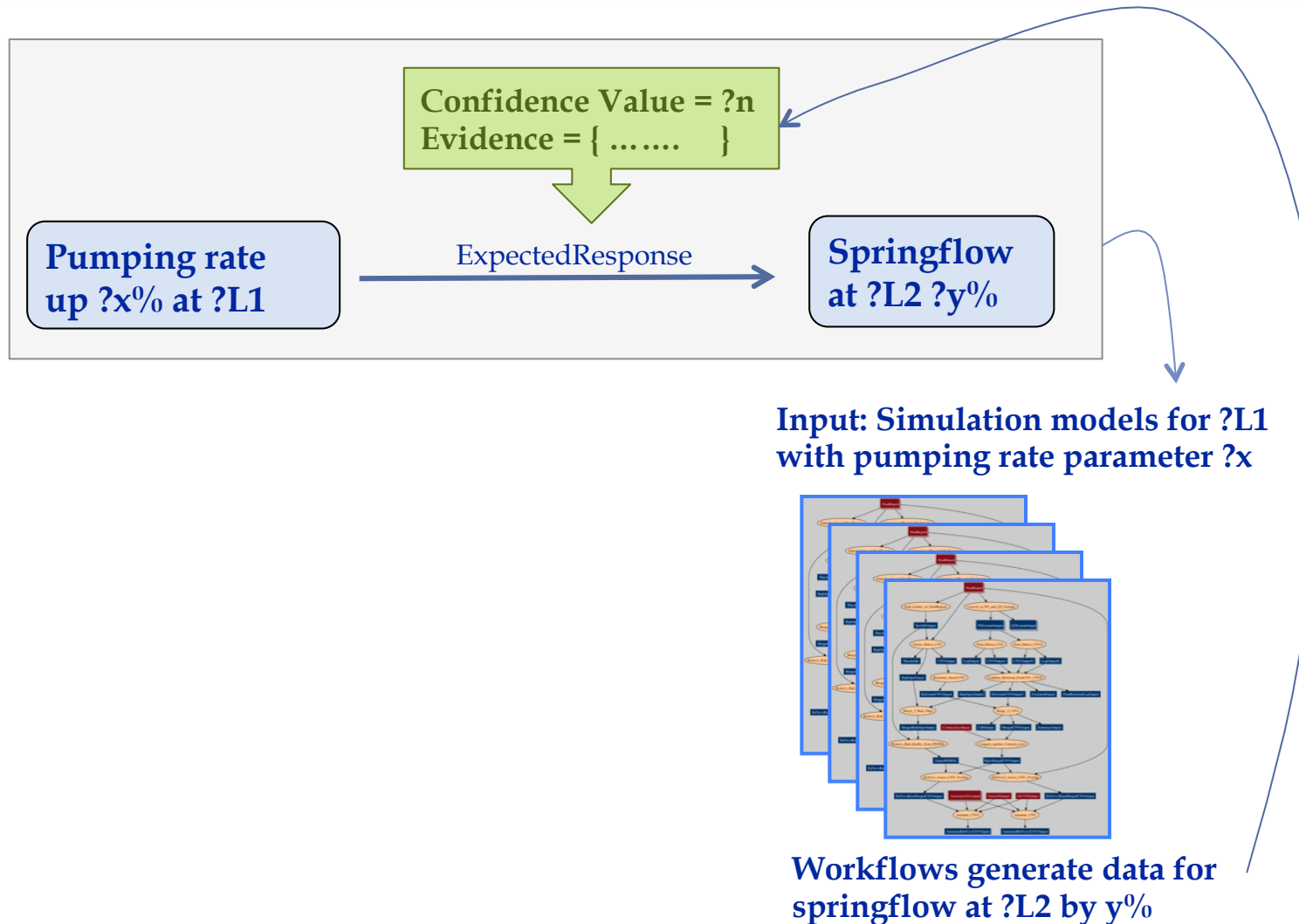
**33 groundwater  
models for Texas**



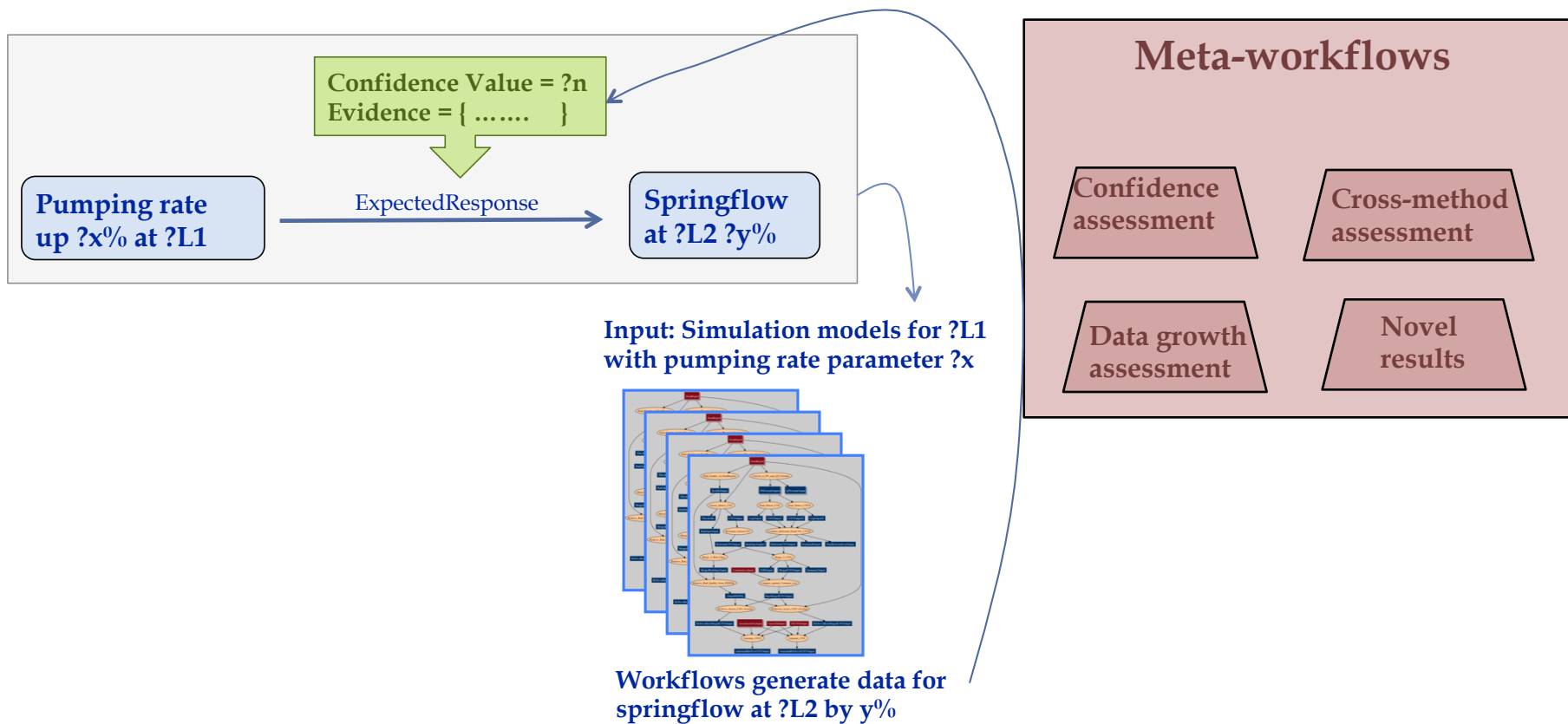
# DISK: Hypotheses



# DISK: Lines of Inquiry



# DISK: Lines of Inquiry

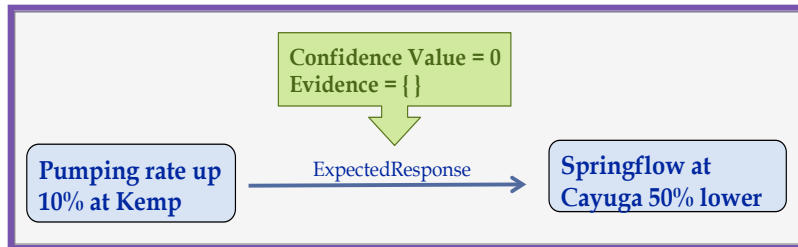




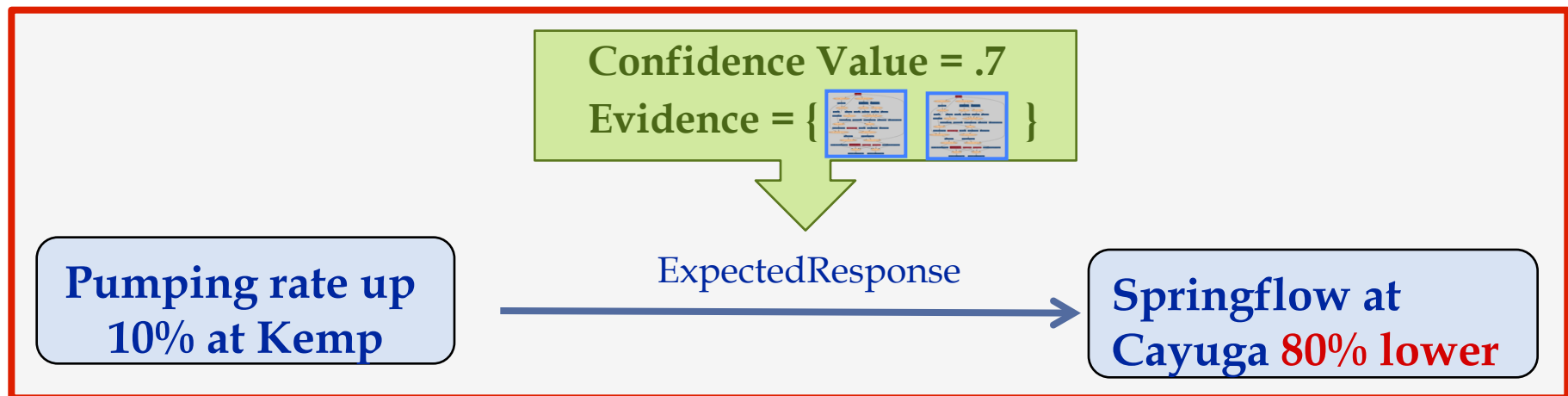
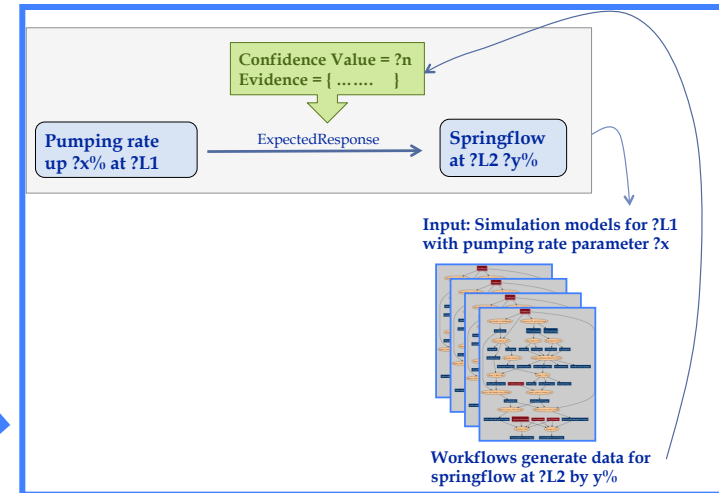


# DISK: Matching Hypotheses Against Lines of Inquiry

## Hypotheses



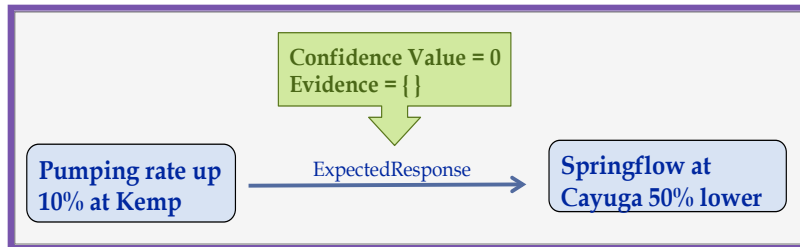
## Lines of Inquiry



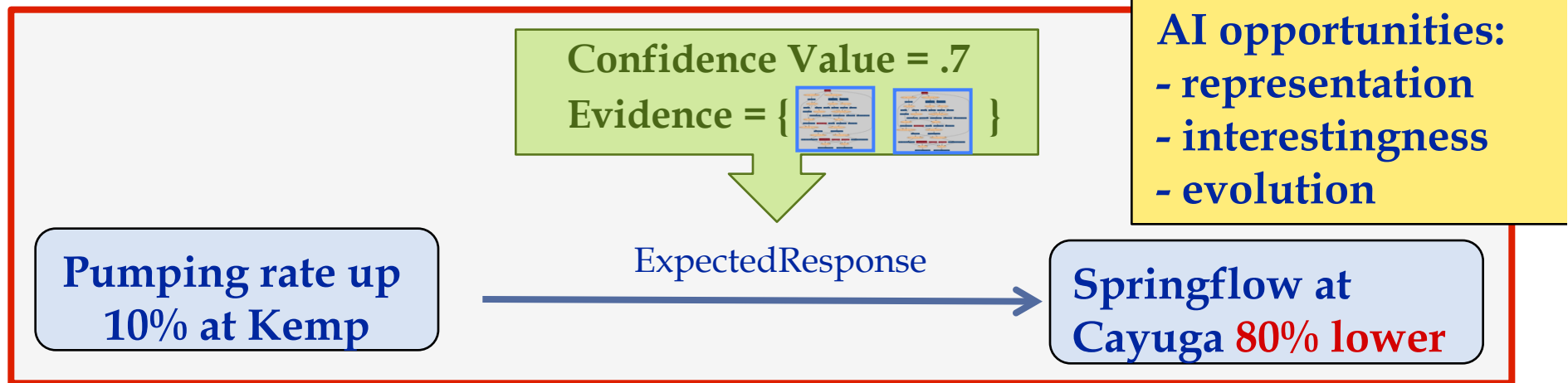
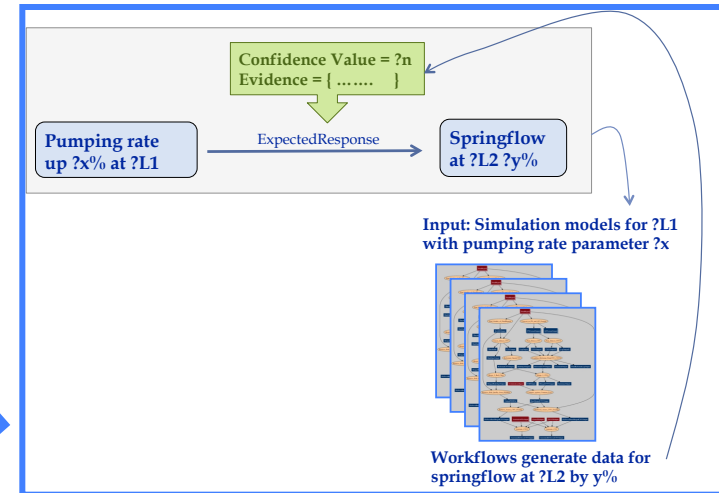


# DISK: Matching Hypotheses Against Lines of Inquiry

## Hypotheses



## Lines of Inquiry



# Knowledge about Meta-Processes: Organic Data Science



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

**1** Page Discussion

**2** Your Overdue Tasks

**3** Write paper about the initial framework design

**4** Write about the evaluation a day ago

**5** Draft paper about the initial framework design

**6** Timeline SubTasks

**7** Develop paper outline 100%

**8** Draft initial versions of key sections 26%

**9** Assemble first full draft of the paper 0%

**10** Collect final evaluation data 0%

**11** Finalize writing the paper 75%

**12** Write paper about the initial framework design

**13** Draft paper about the

**14** Develop paper outline

**15** Draft initial versions of key sections

**16** Assemble first full draft of the paper

**17** Collect final evaluation data

**18** Review first full draft of the paper

**19** Finalize writing the paper

**20** Type<sup>M</sup> medium

**21** Progress<sup>M</sup> 21%

**22** Start date<sup>M</sup> 22nd Aug 2014

**23** Target date<sup>M</sup> 13th Oct 2014

**24** Owner<sup>M</sup> John Smith

**25** Participants James Williams, Steven Johnson

**26** Expertise computer science collaboration

**27** Legend: M Mandatory | States: Not defined, Valid, Inconsistent with parent

**28** The plan is to write a paper with some initial results of our work. If you wa a task and make sure you contribute to it with text or feedback on what ot

**29** Properties

**30** Add

**31** Submitted to IUI-2015 (by John)

**32** AI opportunities:  
- collaboration  
- group formation  
- community health

**33** Organic Data Science

**34** All Tasks My Tasks 38

**35** computer sci search

**36** Develop Framework for Organic Data Sci

**37** Framework Design

**38** Disseminate results from the Organi

**39** Write paper about the initial fra

**40** Train new users to exercis

**41** Design user evaluation of

**42** Collect data about feature

**43** Draft paper about the initia

**44** Cut

**45** Paste

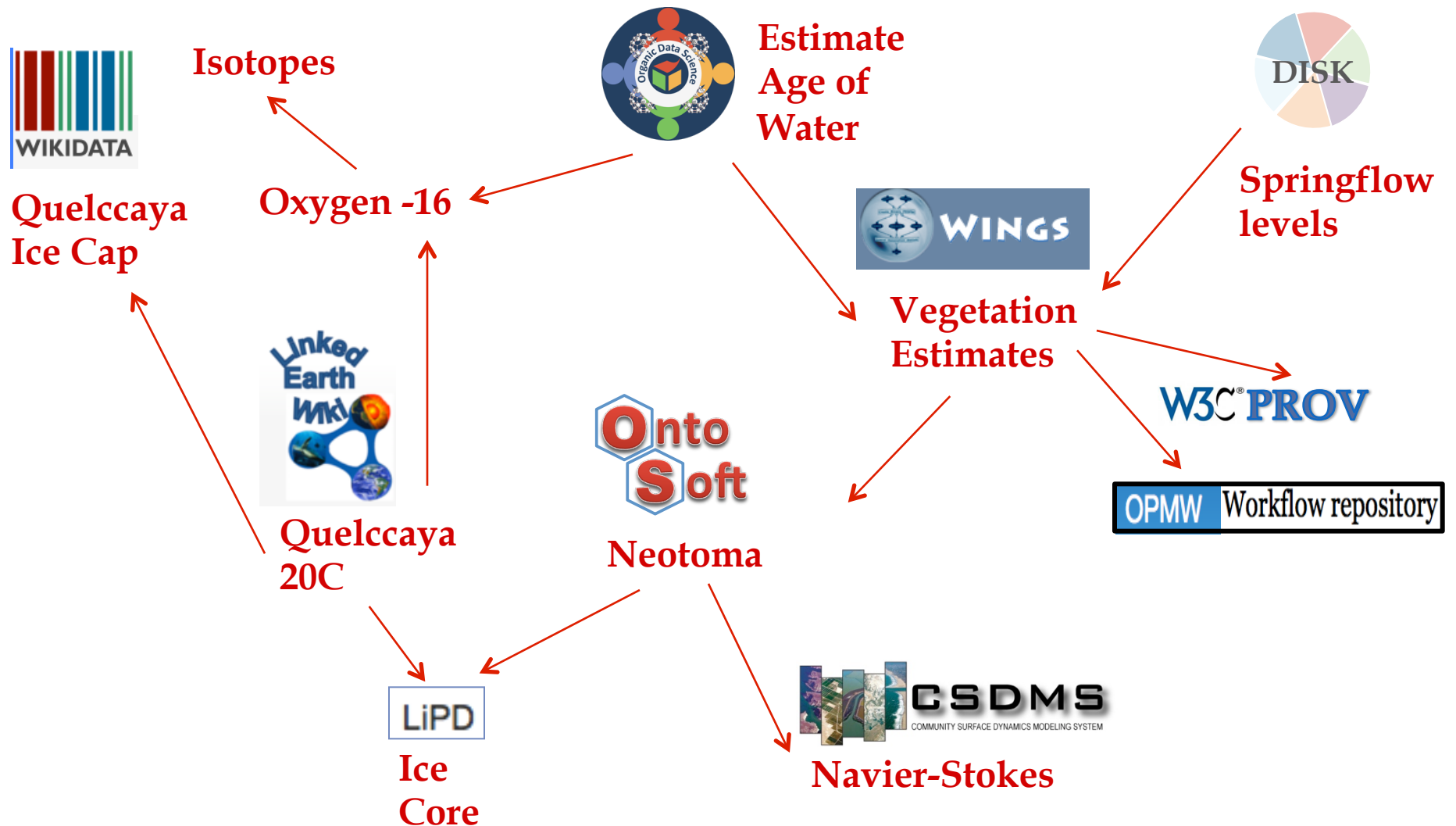
**46** Rename

**47** Delete

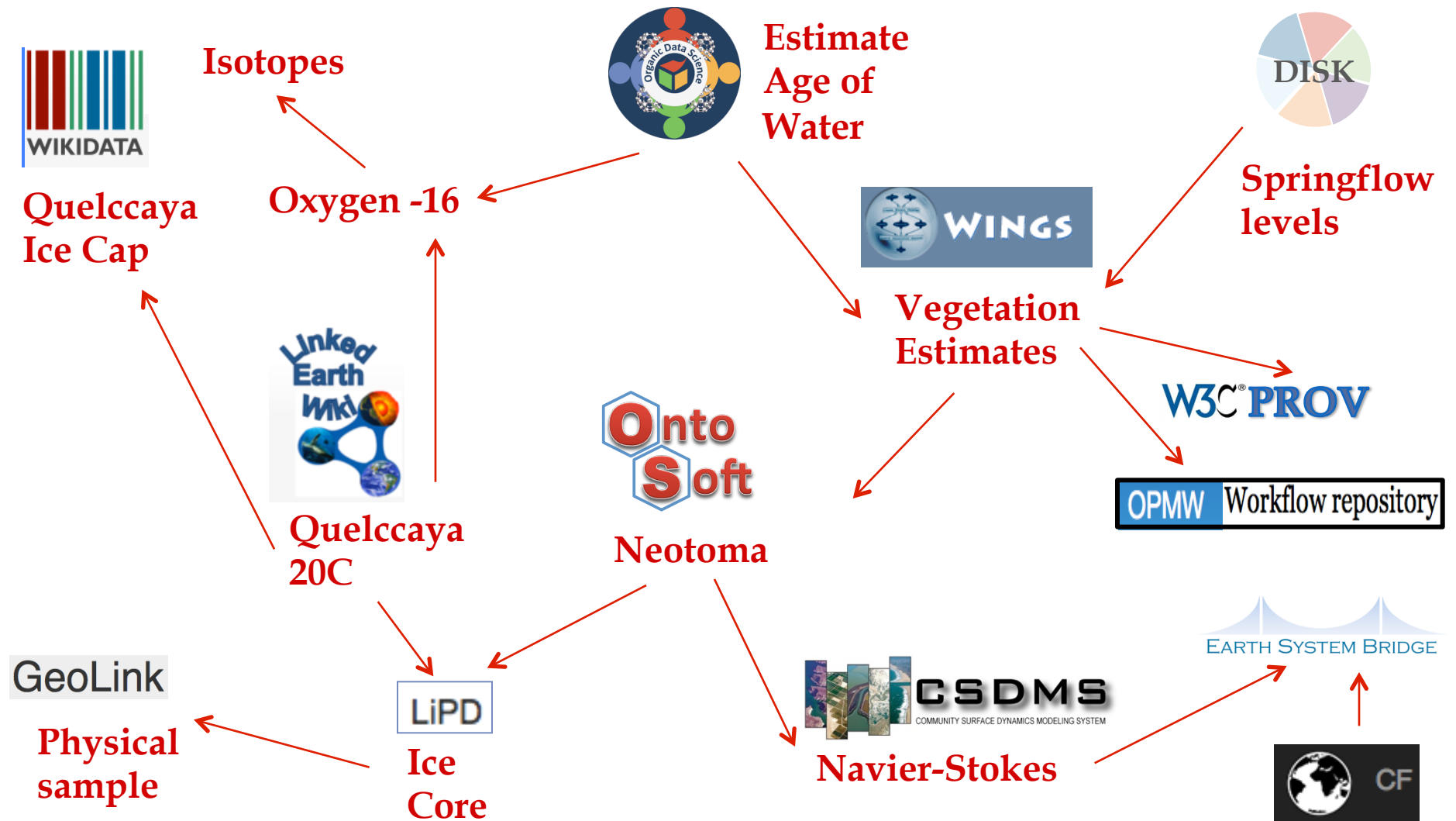
**48** To Toplevel

**49** SMW Semantic MediaWiki

# Linked Data and Linked Knowledge

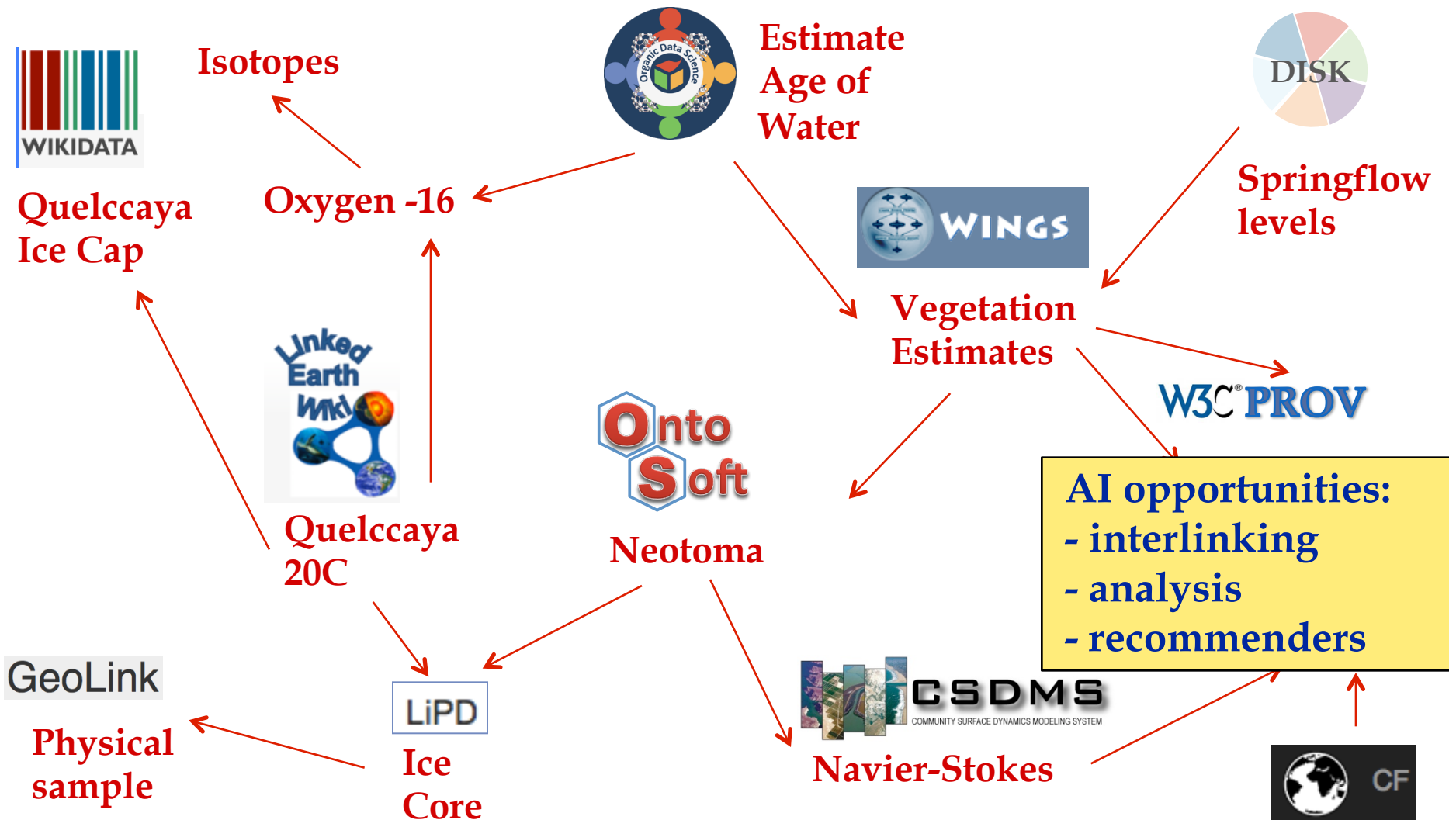


# Linked Data and Linked Knowledge





# Linked Data and Linked Knowledge



# Capturing Scientific Knowledge

---

Data



Software



Provenance

W3C<sup>®</sup> PROV

OPMW Workflow repository

Meta-Workflows

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?

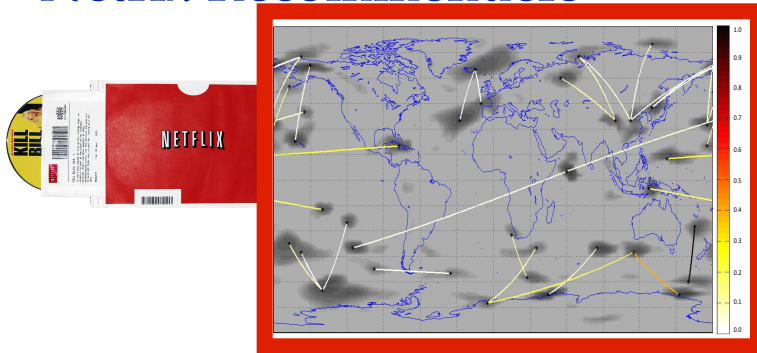
What is the best way to analyze the data?

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



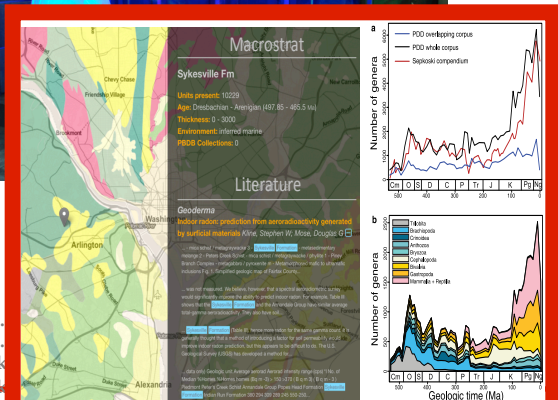
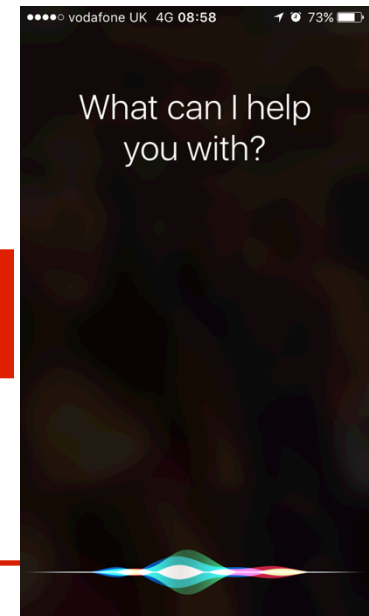
## IBM Watson



## Google Knowledge Graph



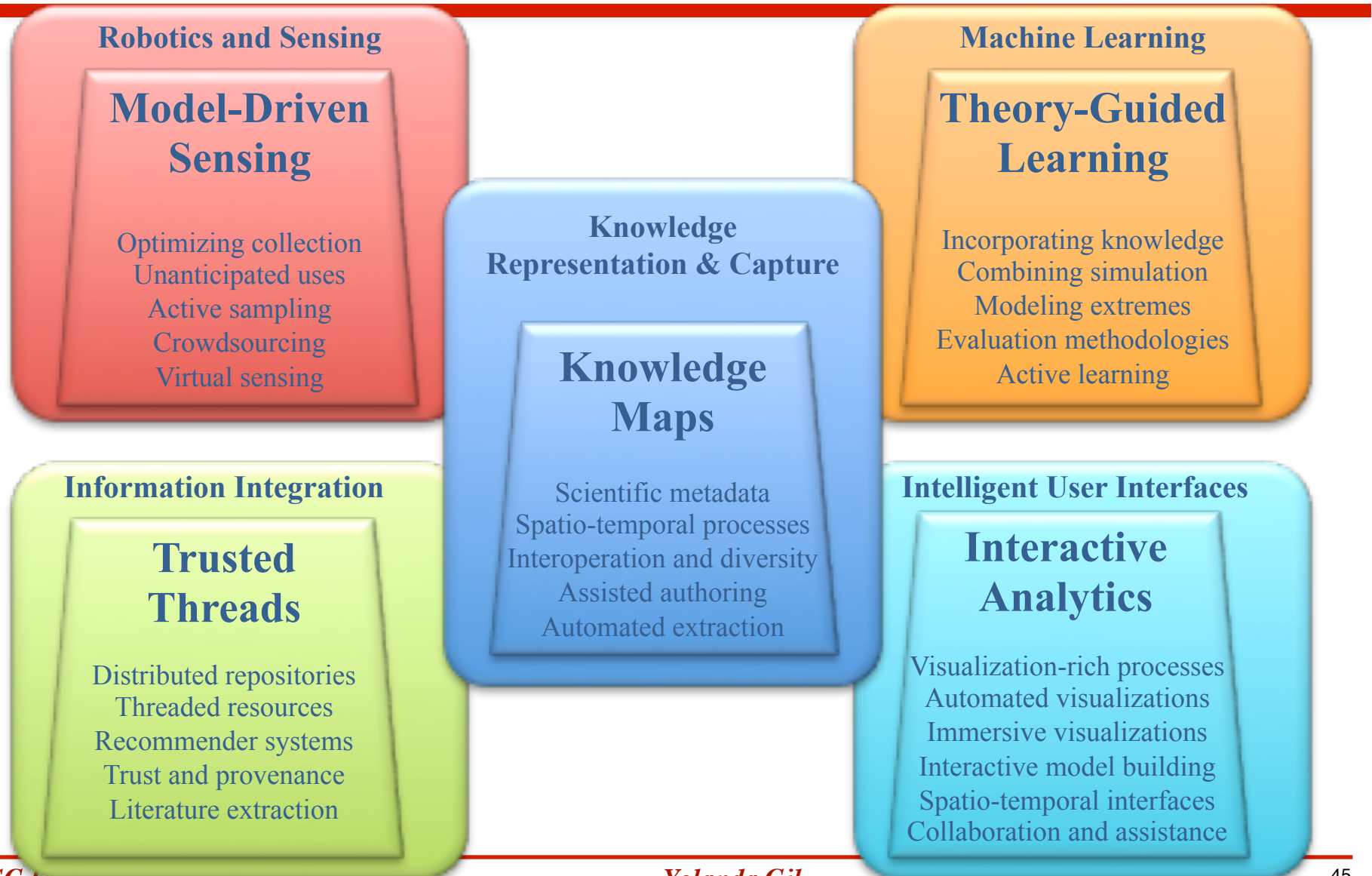
## Apple Siri



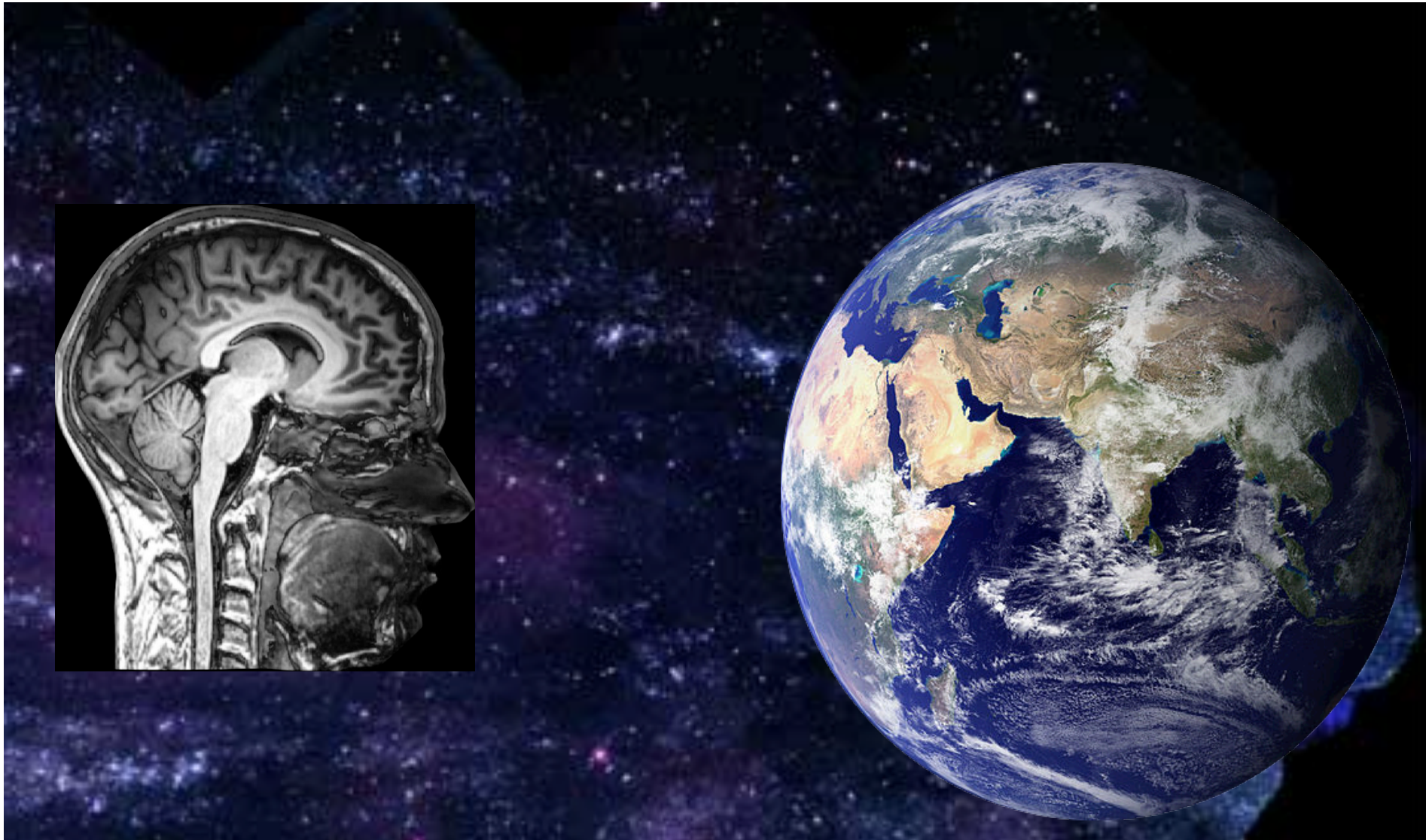
<https://en.wikipedia.org/wiki/Netflix#Science>  
<https://en.wikipedia.org/wiki/Netflix#Science>  
<https://commons.wikimedia.org/wiki/File:NetflixDVD.jpg>  
<https://www.needcar.org>  
<https://en.wikipedia.org/wiki/Netflix#Science>



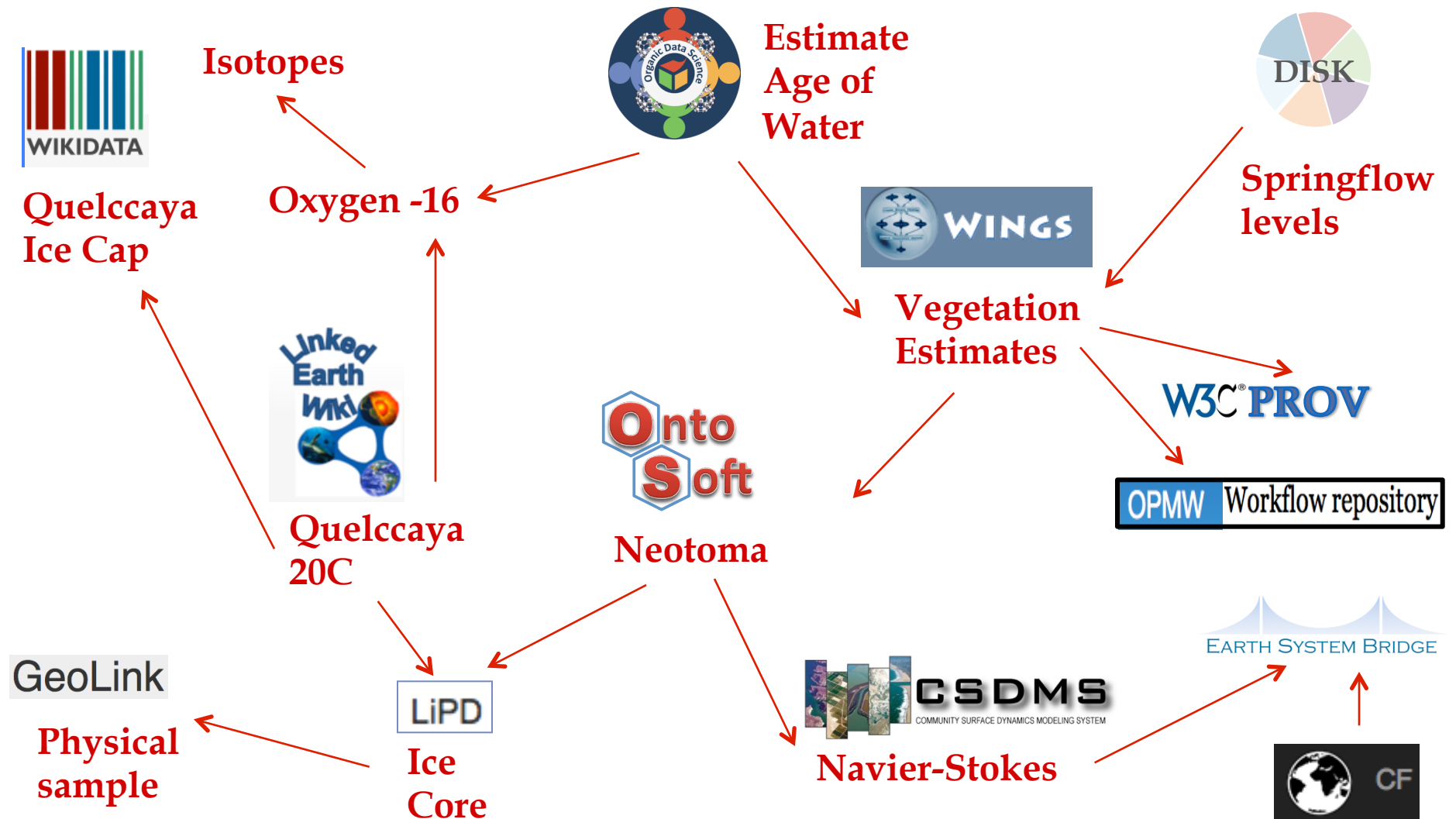
# A Research Agenda for Intelligent Systems in Geosciences (<http://www.is-geo.org>)



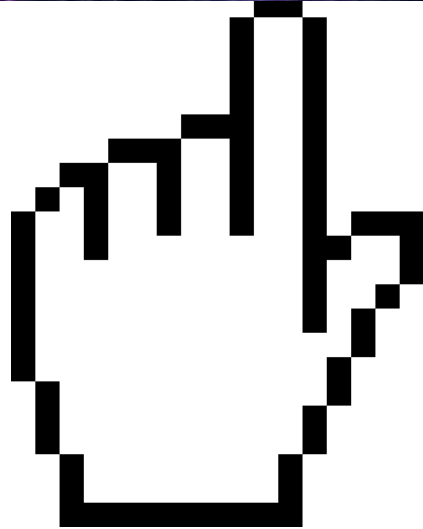
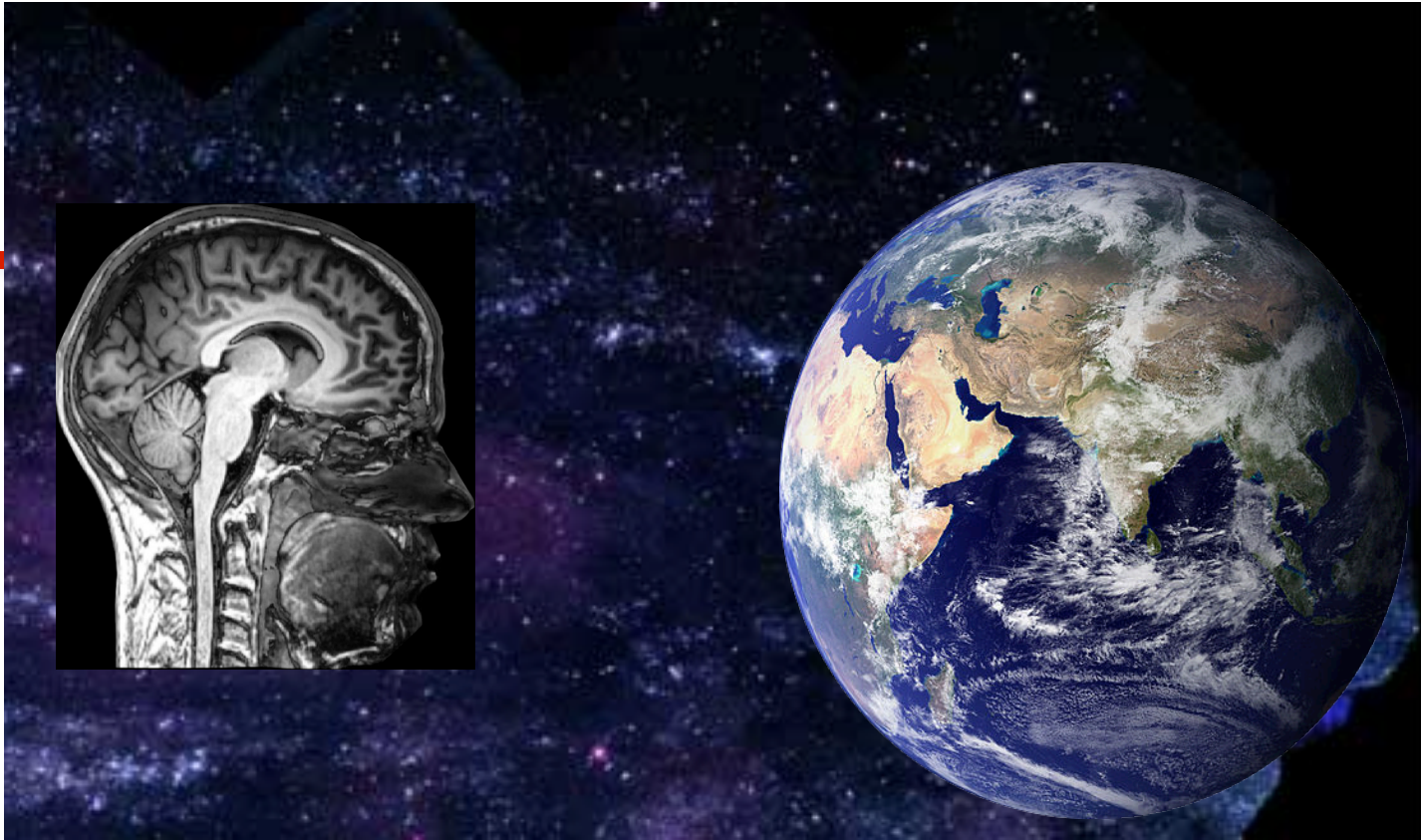


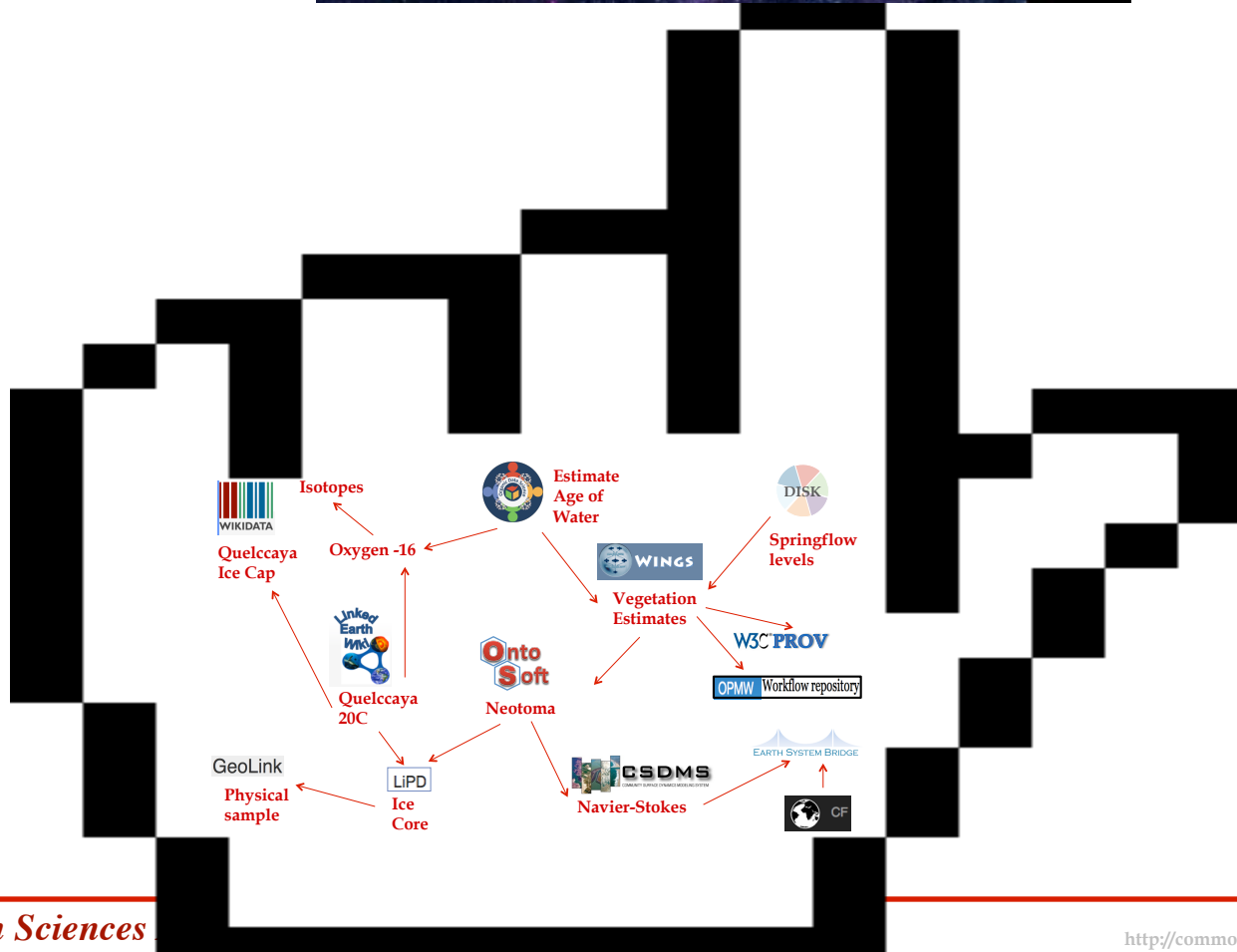


# Capture and Interlink Scientific Knowledge









# 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!*