USC INFORMATION SCIENCES INSTITUTE

USC Viterbi School of Engineering

Hypothesis-Driven Data Analysis of Science Repositories

Yolanda Gil

Information Sciences Institute
and Department of Computer Science

University of Southern California

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

Acknowledgments

- *WINGS contributors*: Varun Ratnakar, Daniel Garijo, Rajiv Mayani, Ricky Sethi, Hyunjoon Jo, Jihie Kim, Yan Liu, Dave Kale (USC), Ralph Bergmann (U Trier), William Cheung (HKBU), Oscar Corcho (UPM), Pedro Gonzalez & Gonzalo Castro (UCM), Paul Groth (VUA)
- *WINGS collaborators*: Chris Mattmann & Paul Ramirez & Dan Crichton & Rishi Verma (JPL), Ewa Deelman & Gaurang Mehta & Karan Vahi (USC), Sofus Macskassy (ISI), Natalia Villanueva & Ari Kassin (UTEP)
- *Organic Data Science contributors*: 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)
- *DISK contributors*: Parag Mallick, Ravali Adusumilli, Hunter Boyce (Stanford); Suzanne Pierce, John Gentle (UT Austin)
- *Biomedicine*: Parag Mallick & Ravali Adusumilli & Hunter Boyce (Stanford U.), Phil Bourne & Sarah Kinnings (UCSD), Chris Mason (Cornell); Joel Saltz & Tahsin Kurk (Emory U.); Jill Mesirov & Michael Reich (Broad); Randall Wetzel (CHLA); Shannon McWeeney & Christina Zhang (OHSU)
- *Geosciences*: Suzanne Pierce & John Gentle (UT Austin), Chris Duffy (PSU); Paul Hanson (U Wisconsin); Tom Harmon & Sandra Villamizar (U Merced); Tom Jordan & Phil Maechlin (USC); Kim Olsen (SDSU)
- *And many others!*

Community Workshops

NSF Workshop on Discovery Informatics

February 2-3, 2012

Arlington, VA

Final Workshop Report

August 31, 2012

Intelligent Systems for Geosciences

http://www.IS-GEO.org

Outline

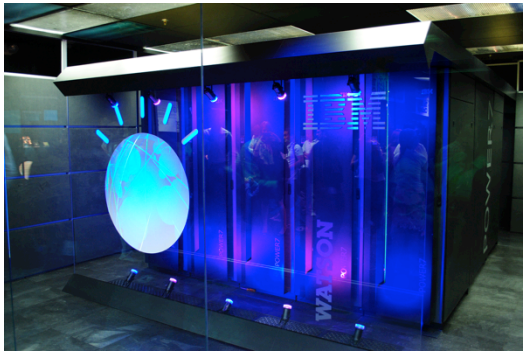
- Artificial intelligence and scientific discovery
  - The knowledge systems tradition
- Our recent work on capturing knowledge about data analysis strategies
  - Hypothesis-driven data analysis
- Representing and capturing data analysis knowledge
  - About data, software, methods, meta-analysis
- Summary of AI challenges

Outline

- Artificial intelligence and scientific discovery
  - The knowledge systems tradition
- Our recent work on capturing knowledge about data analysis strategies
  - Hypothesis-driven data analysis
- Representing and capturing data analysis knowledge
  - About data, software, methods, meta-analysis
- Summary of AI challenges

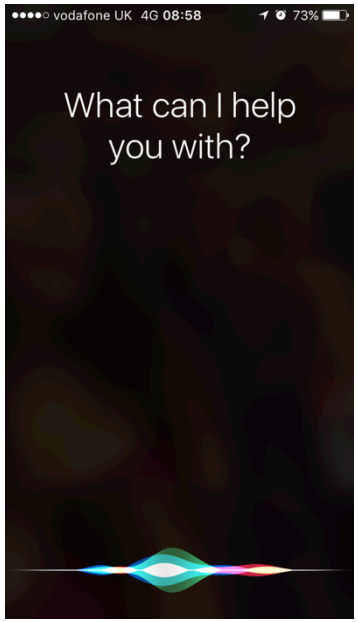
AI's Coming of Age

IBM Watson



Google Knowledge Graph

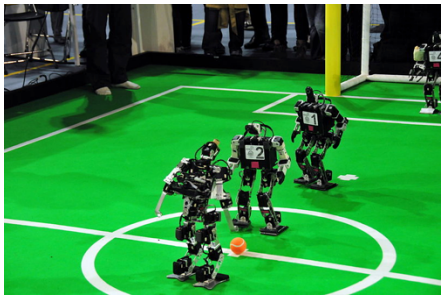
Apple Siri



Netfix Recommenders

Tesla AutoPilot

RoboCup Soccer

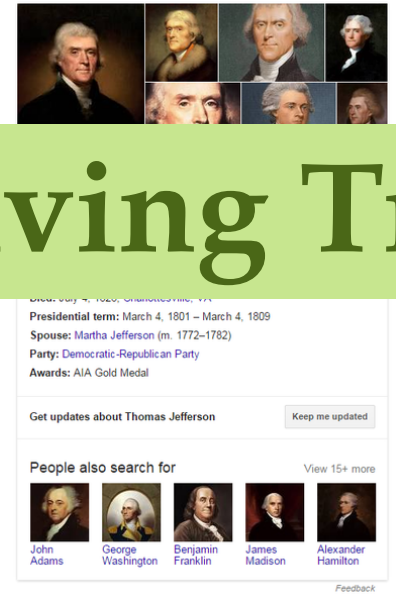


https://en.wikipedia.org/wiki/Watson_(computer)#/media/File:IBM_Watson.PNG
https://en.wikipedia.org/wiki/Siri#/media/File:SirioniOS9.png
https://commons.wikimedia.org/wiki/File:Google_Knowledge_Panel.png
https://commons.wikimedia.org/wiki/File:13-06-28-robocup-eindhoven-005.jpg
http://www.greencarreports.com/news/1100482_tesla-autopilot-the-10-most-important-things-you-need-to-know
https://en.wikipedia.org/wiki/Netflix#/media/File:NetflixDVD.jpg

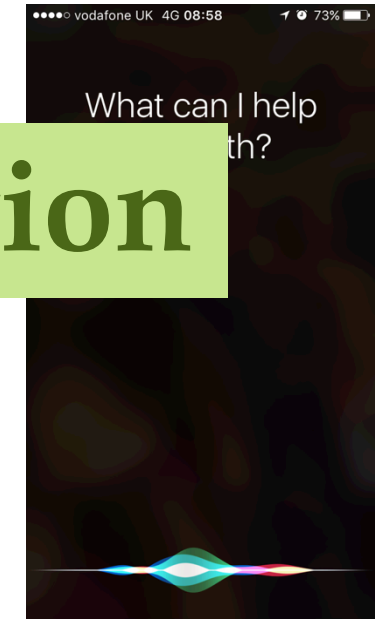
AI's Coming of Age

IBM Watson

Google Knowledge Graph



Apple Siri



Problem Solving Tradition

Recommenders

Autonomous Drivers

Robot Teams

Data Systems Tradition

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:13-06-28-robocup-eindhoven-005.jpg
http://www.greencarreports.com/news/1100482_tesla-autopilot-the-10-most-important-things-you-need-to-know
https://en.wikipedia.org/wiki/Netflix#/media/File:NetflixDVD.jpg

Artificial Intelligence and Scientific Discovery: The Problem Solving Tradition

Pittsburg Post Gazette Archives

Outline

- Artificial intelligence and scientific discovery
  - The knowledge systems tradition
- Our recent work on capturing knowledge about data analysis strategies
  - Hypothesis-driven data analysis
- Representing and capturing data analysis knowledge
  - About data, software, methods, meta-analysis
- Summary of AI challenges

DISK: Automated DIscovery of Scientific Knowledge

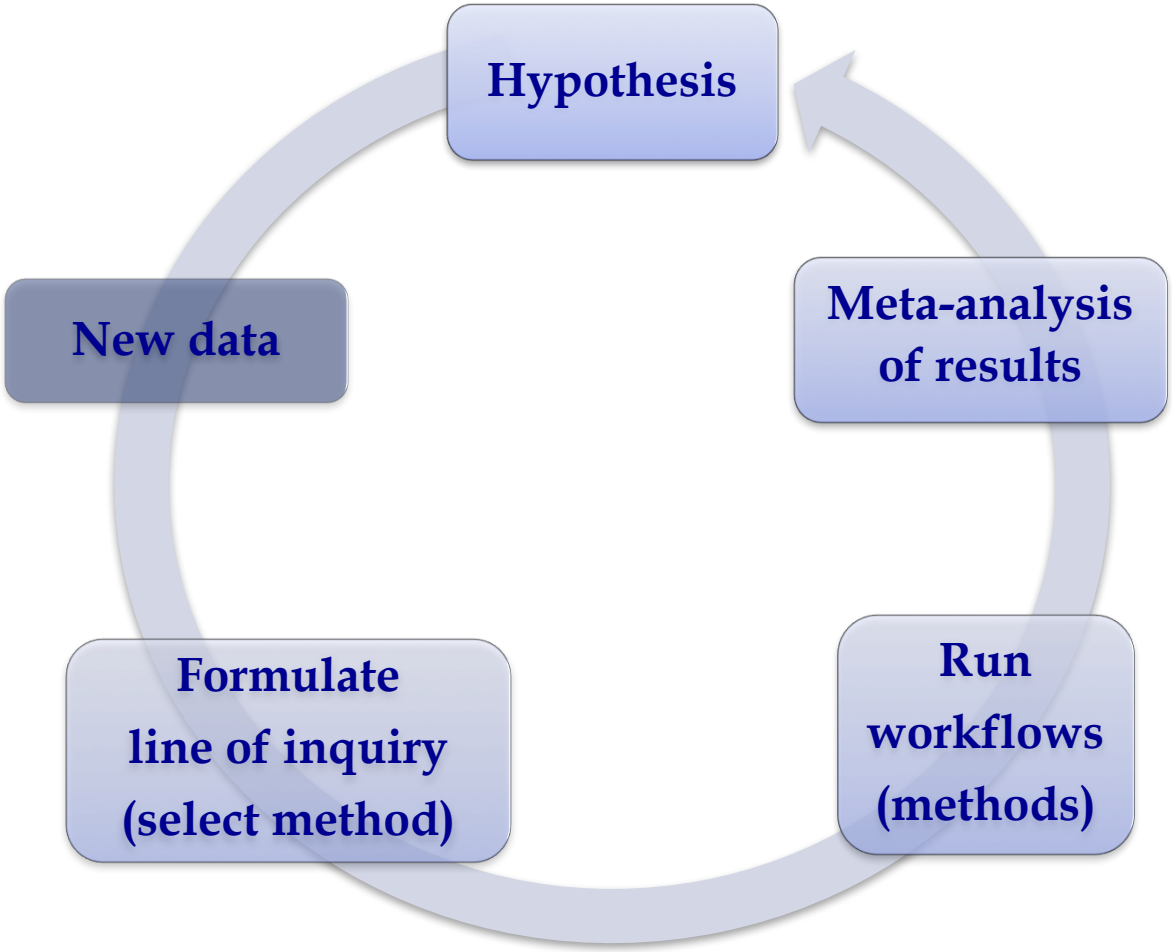
Work with P. Mallick, R. Adusumilli, H Boyce (Stanford U.)

- Long-Term Goal: Human directs automated intelligent system to explore hypotheses of interest
  - Hypothesis-driven data analysis and discovery
    - Systematic and reproducible analyses
  - Report of findings with explanations ("Friday" meeting)

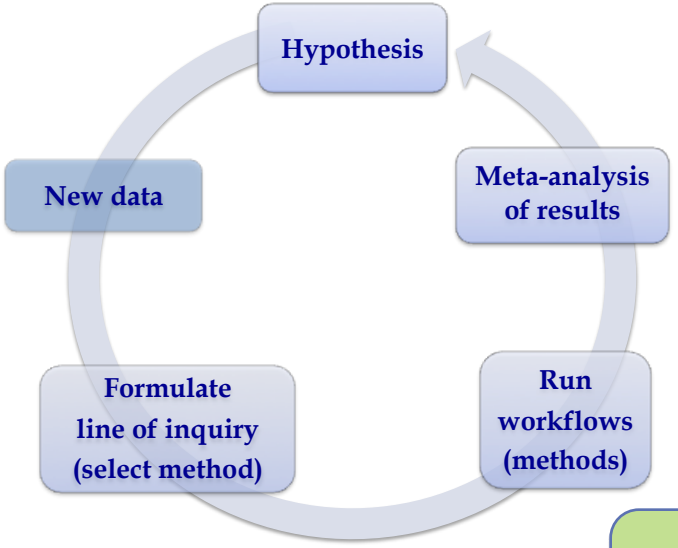
- Approach: Intelligent system that captures common data analysis strategies used by scientists in a domain
  - Build on WINGS intelligent workflow system that can adapt data analysis given the constraints of algorithmic steps

disk

Scientific Data Analysis Today: Inefficient, Incomplete, Irreproducible, One-Shot



Capturing Data Analysis Strategies through ***Lines of Inquiry***



New Hypothesis

Lines of Inquiry

Specify relevant analytic methods (workflows), type of data needed, and how to combine results

DATA CATALOG

Query to Retrieve Data

Analytic Workflows

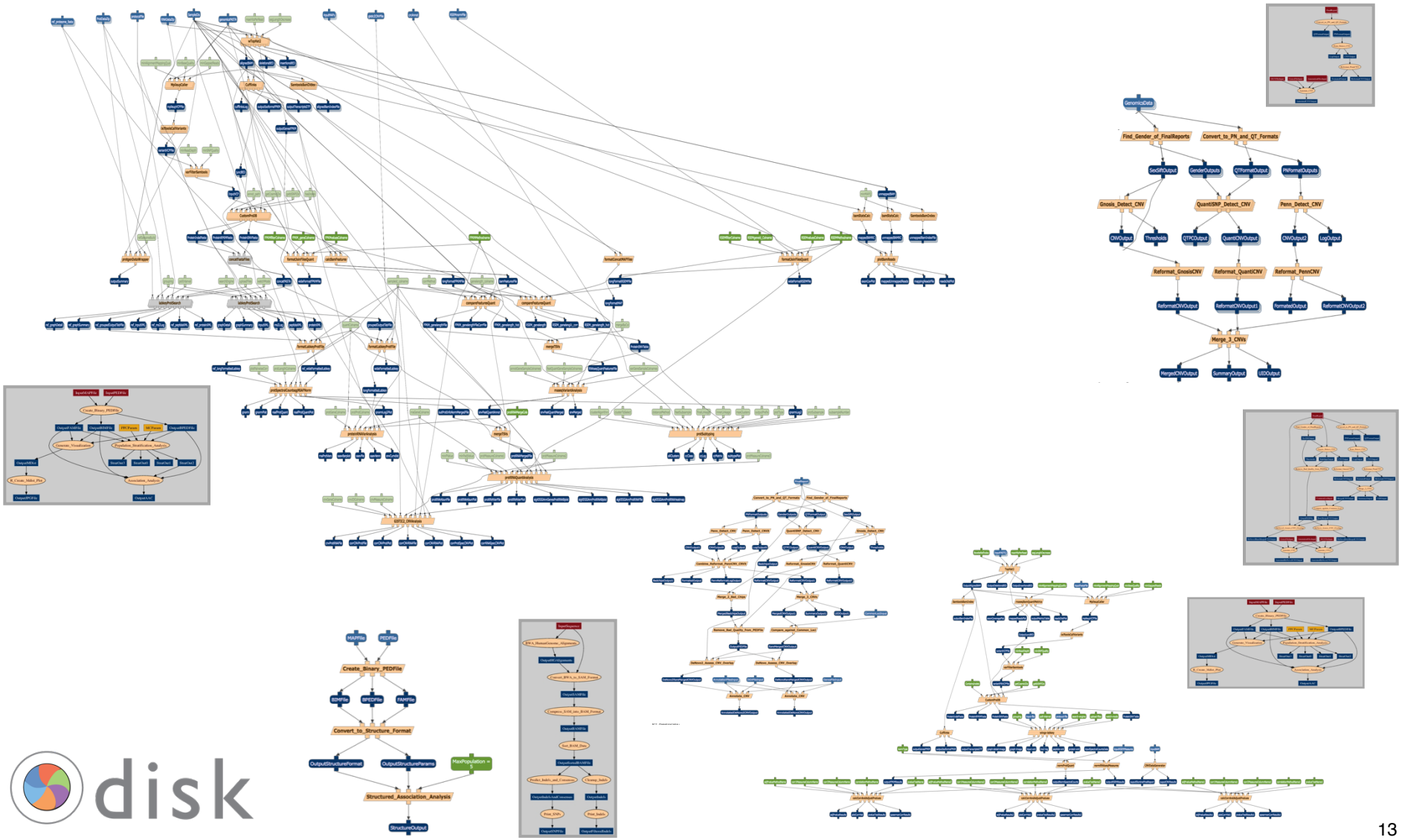
- Workflow Library

Meta-Workflows

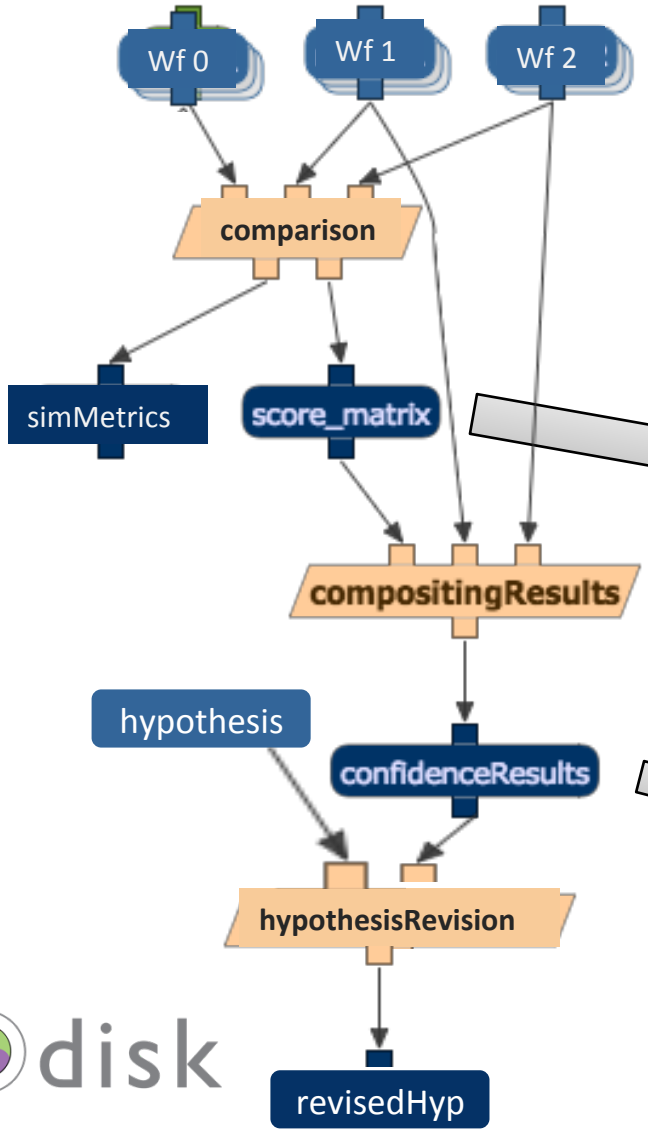
- Confidence Estimation
- Benchmarking

disk

Knowledge about Multi-Omics Data Analysis Captured in Workflows



Knowledge about Evidence Aggregation Captured in Meta-Workflows



- After running the workflows, meta-workflows analyze their results and generate a combined confidence value

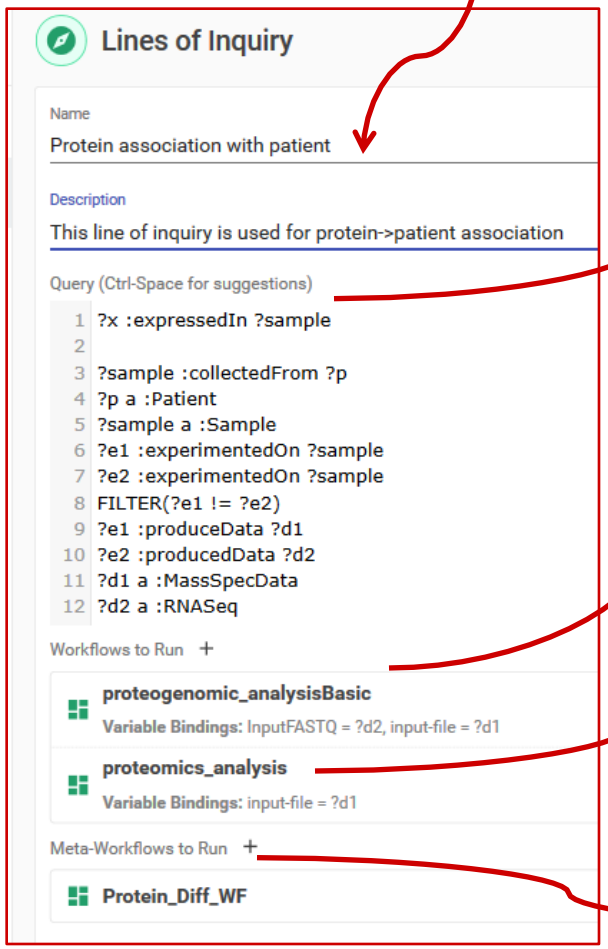
	Workflow 0	Workflow 1	Workflow 2
Workflow 0	1.00	0.96	0.72
Workflow 1	0.96	1.00	0.84
Workflow 2	0.72	0.84	1.00

Scan	Peptide	Peptide Prophet	Protein	Confidence
9226	R.ESALEPGPVPEAPAGGPVHAVTVVTLLEK.L	0.9958	NP_659477_rs2682123:R8P;rs1051992:L158P	0.9589
9401	M.AEYLASIFGTEK.D	0.9797	NP_659424_W95G	0.9998
7556	K.ISNPWQSPSGTLPALR.T	0.9975	NP_002446_677:AGG	0.972
8699	R.SDPVTLNVLYGPDVPTISPSK.A	0.9975	NP_002474_T100I,G239V	0.972
9614	K.LFMVDSIPK.V	0.874	NP_659403_rs7041710:L863V;rs10961700:V1502M,S1716L,A1784G;rs10961689:Q2143P	0.468
9282	M.AEYLASIFGTEKDK.V	0.802	NP_659424_W95G	0.9998
Scan	Peptide	Peptide Prophet	Protein	Confidence

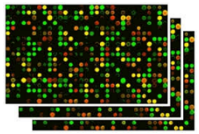
Knowledge about Data Analytic Strategies Captured in Lines of Inquiry

hypothesis

Protein PRKCDBP is expressed in samples of patient P36

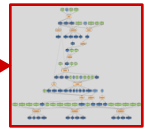


data

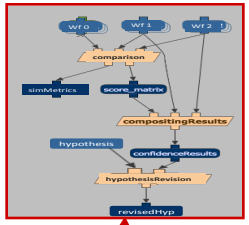


revision

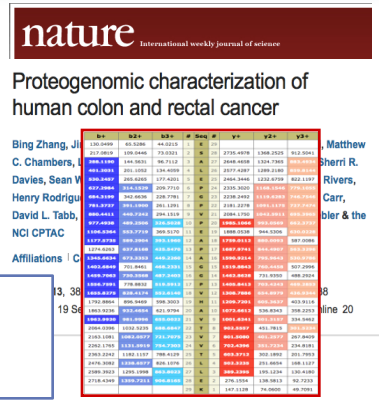
PRKCDBP mutation is expressed in P36



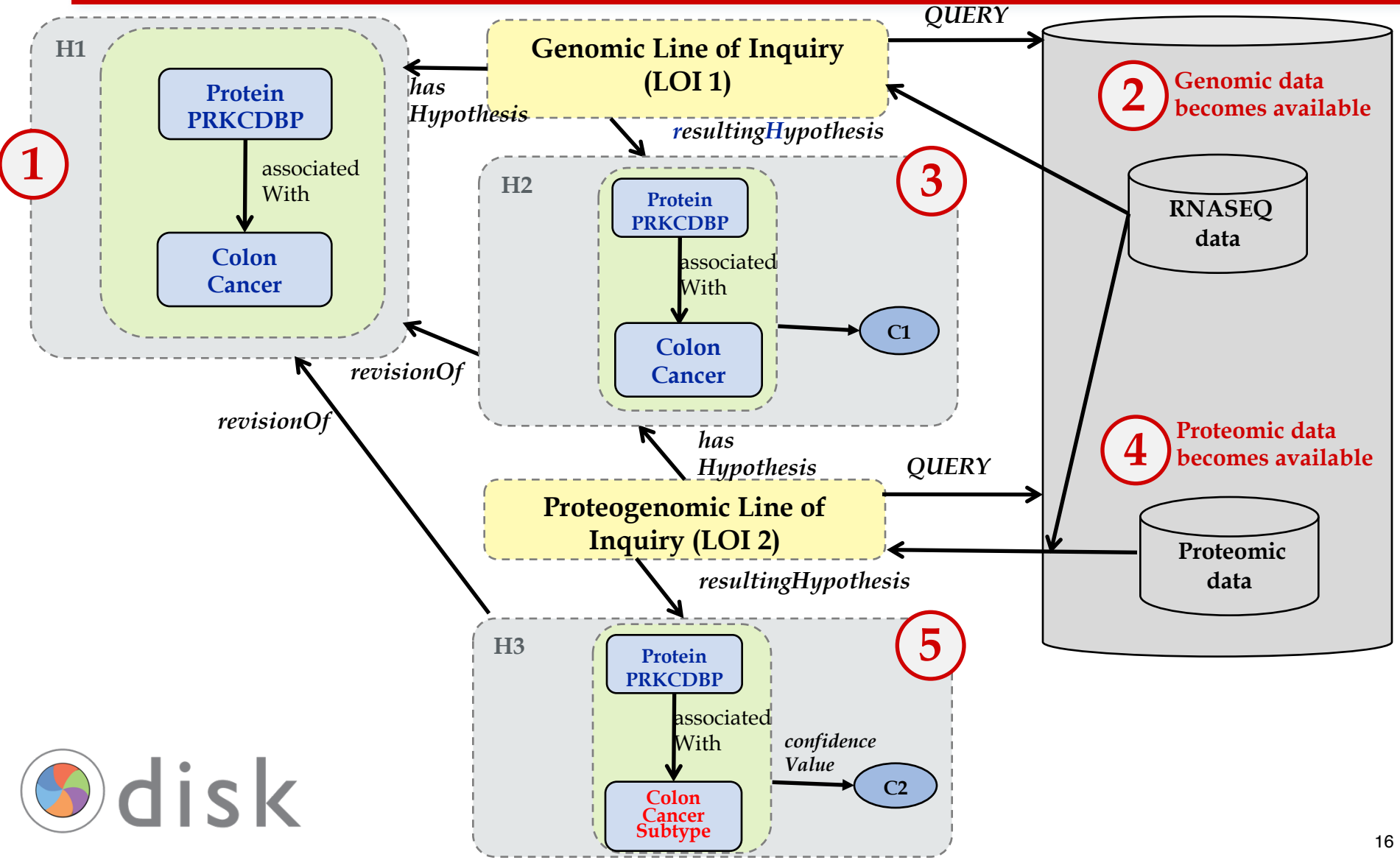
workflows



meta-workflow



Representing Hypothesis Evolution as New Data is Available and Analyzed



Addressing Problems with Current Science Data Analysis Practice

- Integrative studies of multi-source data are rare
  - Requires collaborations to cover all the specialized expertise needed, so they take years

"**Automated Hypothesis Testing with Large Scientific Data Repositories**."
Y. Gil, D. Garijo, V. Ratnakar, R. Mayani, R. Adusumilli, H. Boyce, and P. Mallick. *Proceedings of the 4th Annual Conference on Advances in Cognitive Systems (ACS)*, 2016.

- Data resources are constantly growing
  - Scientists rarely re-evaluate evidence for hypotheses as new data becomes available

"**Towards Continuous Scientific Data Analysis and Hypothesis Evolution**."
Y. Gil, D. Garijo, V. Ratnakar, R. Mayani, R. Adusumilli, H. Boyce, A. Srivastava and P. Mallick. *To appear in the Proceedings of the 31st Annual Conference of the Association for the Advancement of Artificial Intelligence (AAAI), 2017.*

Representation Challenges in Hypothesis-Driven Discovery from Scientific Data Repositories

- A domain-independent framework that works across scientific disciplines
- Hypothesis representation and evolution
- Meta-reasoning strategies to select and prioritize analyses
- Aggregation of evidence from multiple types of data/observations
- Generation of explanations from provenance records
- Learning to improve from user feedback
- Identifying "interestingness" in results
- Designing new data analysis methods
- Incorporating science knowledge and theories to guide hypothesis formation

Outline

- Artificial intelligence and scientific discovery
  - The knowledge systems tradition
- Our recent work on capturing knowledge about data analysis strategies
  - Hypothesis-driven data analysis
- Representing and capturing data analysis knowledge
  - About data, software, methods, meta-analysis
- Summary of AI challenges

Representing and Capturing Scientific Knowledge

Data

Software

Workflows

Meta-Workflows

Provenance

Knowledge about Data: Linked Earth

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

Palmyra Atoll [edit]

WIKIDATA

Main page
Community portal
Project chat
Create a new item
Item by title
Recent changes
Random item
Help
Donate

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]

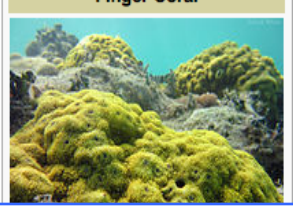
- [x] http://dbpedia.org/resource/Porites
- [add source]

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



Web of Data

RDF

[add source]

Porites — indica <sameAs>

hide — dbpedia

class	Coral
classis	Coral
familia	Poritidae
family	Poritidae
genus	Porites
kingdom	Animal
name	Finger Coral
order	Fungiina
order	Scleractinia
ordo	Scleractinia
phylum	Cnidaria

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

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

[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]	Measurement		
	MeasurementMaterial		
	MeasurementStandard		
	MeasurementUnits		
[x]	Reference		
[x]	Species		

AI opportunities:
- collection
- normalization
- organization

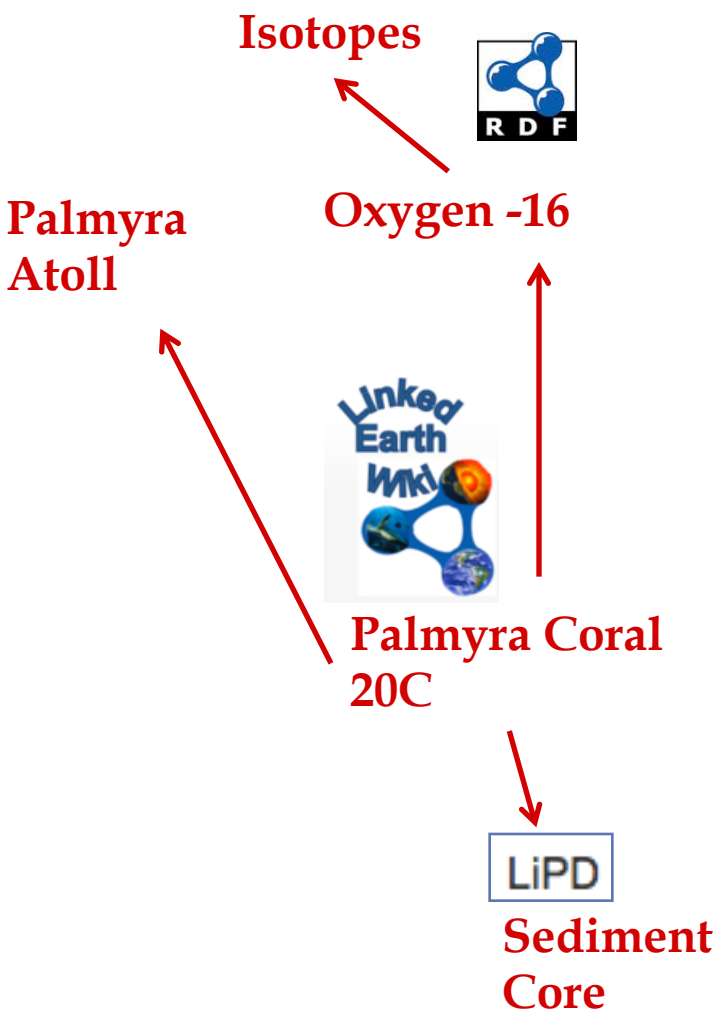
Credits

Users who have contributed to this Page:

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

RDF

Linked Data and Entities: Semantic Web Objects as RDF + URIs



Knowledge about Software: OntoSoft

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

Software Community Training

Compare Software

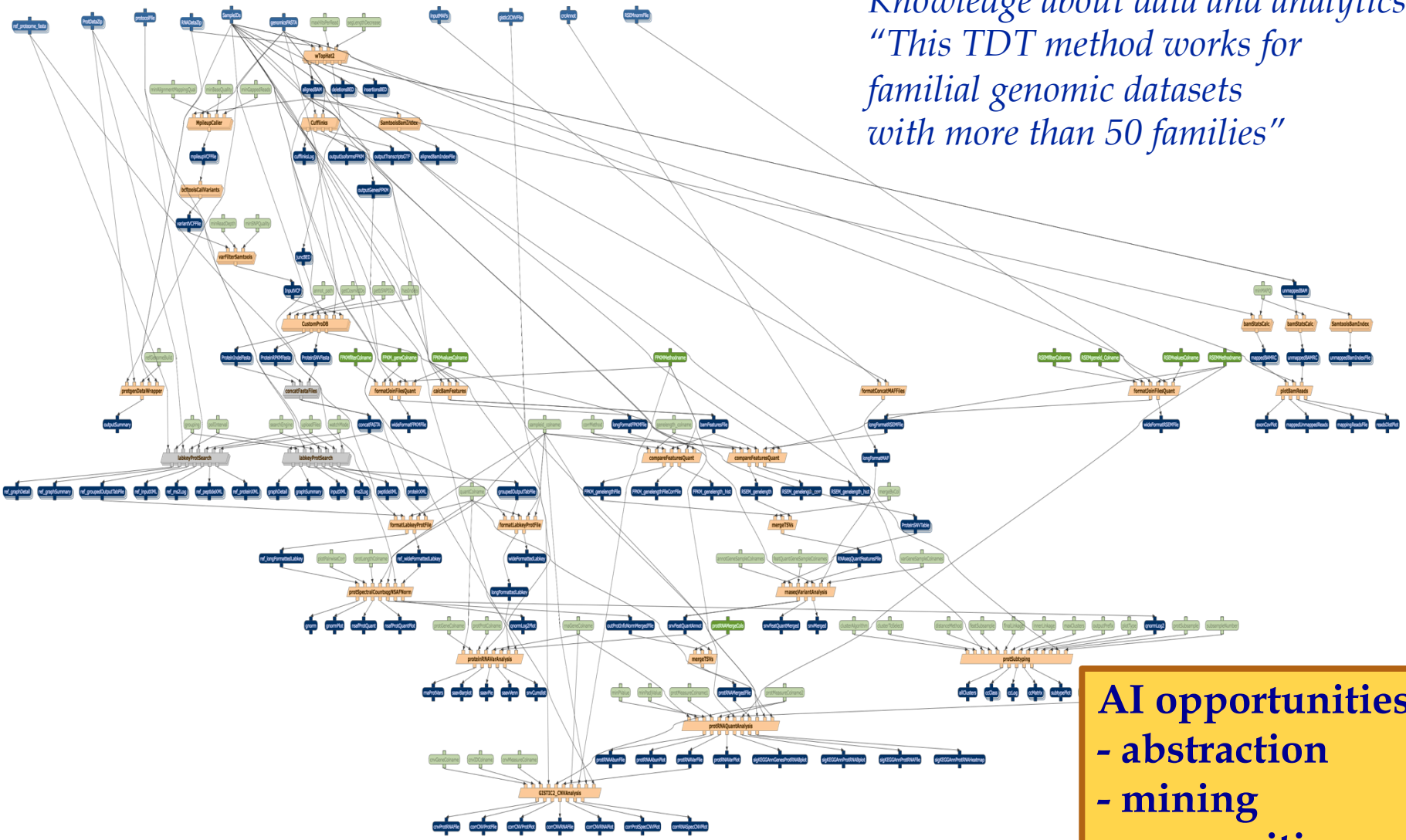
DrEICH algorithm, PIHM, PIHMgis, TauDEM, WBMsed

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 mod
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	
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://sourceforge.net/projects/pihmmodel/ **Test Data Description:** Upper Juniata River 875 km^2: see: http://source-		**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/Model:WBMsed#Testing **Test Data Description:** Extensive input dataset is available on the CSDMS

AI opportunities:
- functional descriptions
- organization
- linking to data

Knowledge About Analytic Methods: WINGS

Knowledge about data and analytics: "This TDT method works for familial genomic datasets with more than 50 families"

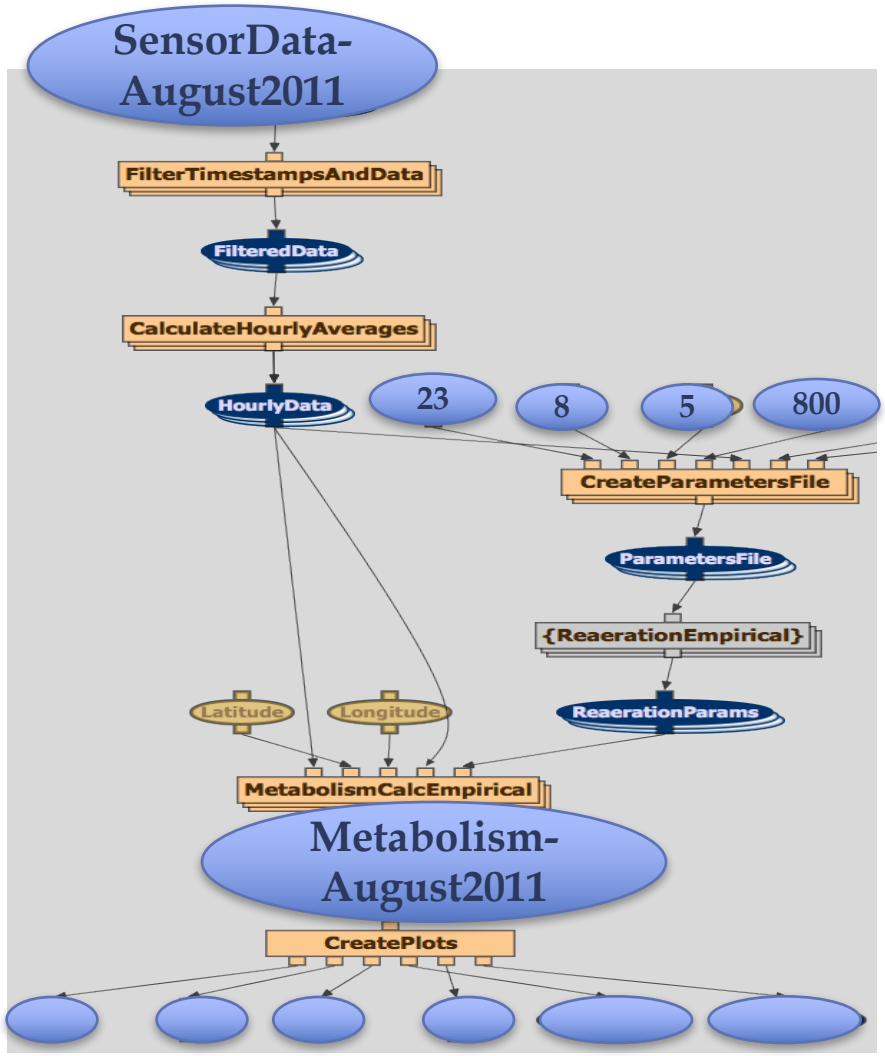


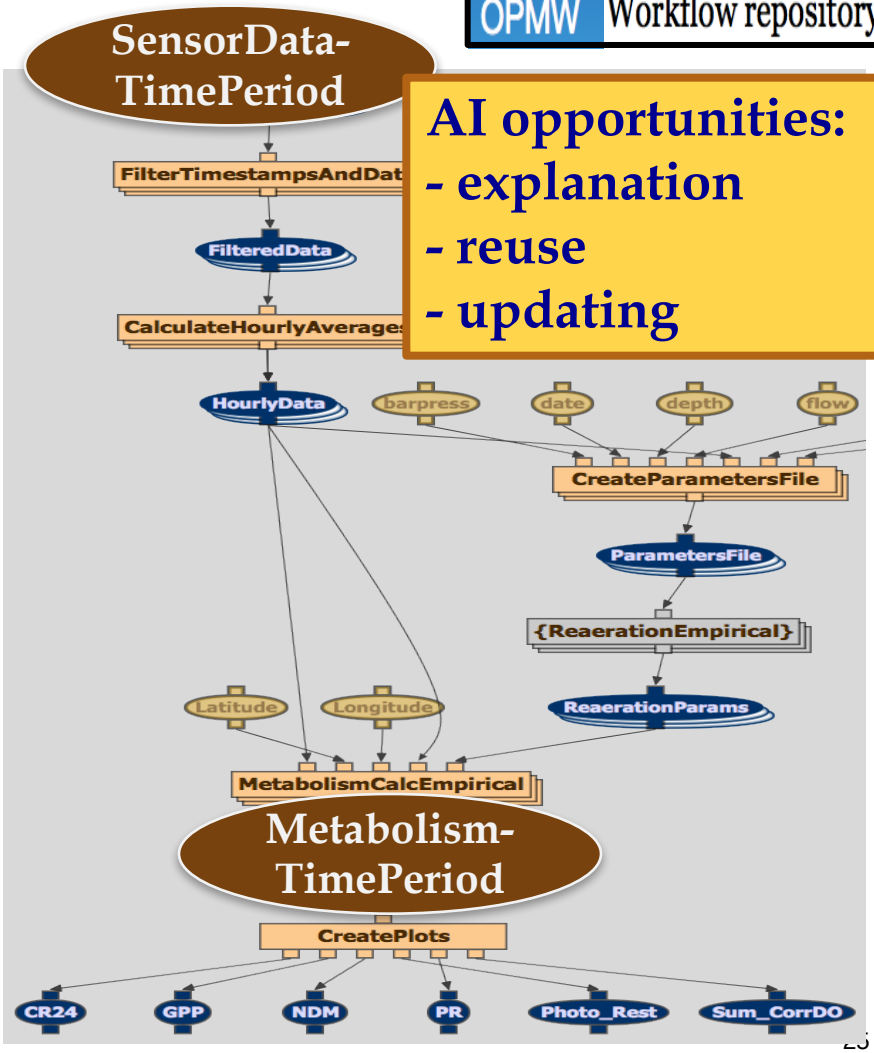
AI opportunities:
- **abstraction**
- **mining**
- **composition**

Knowledge About Provenance: W3C PROV

W3C PROV

OPMW Workflow repository

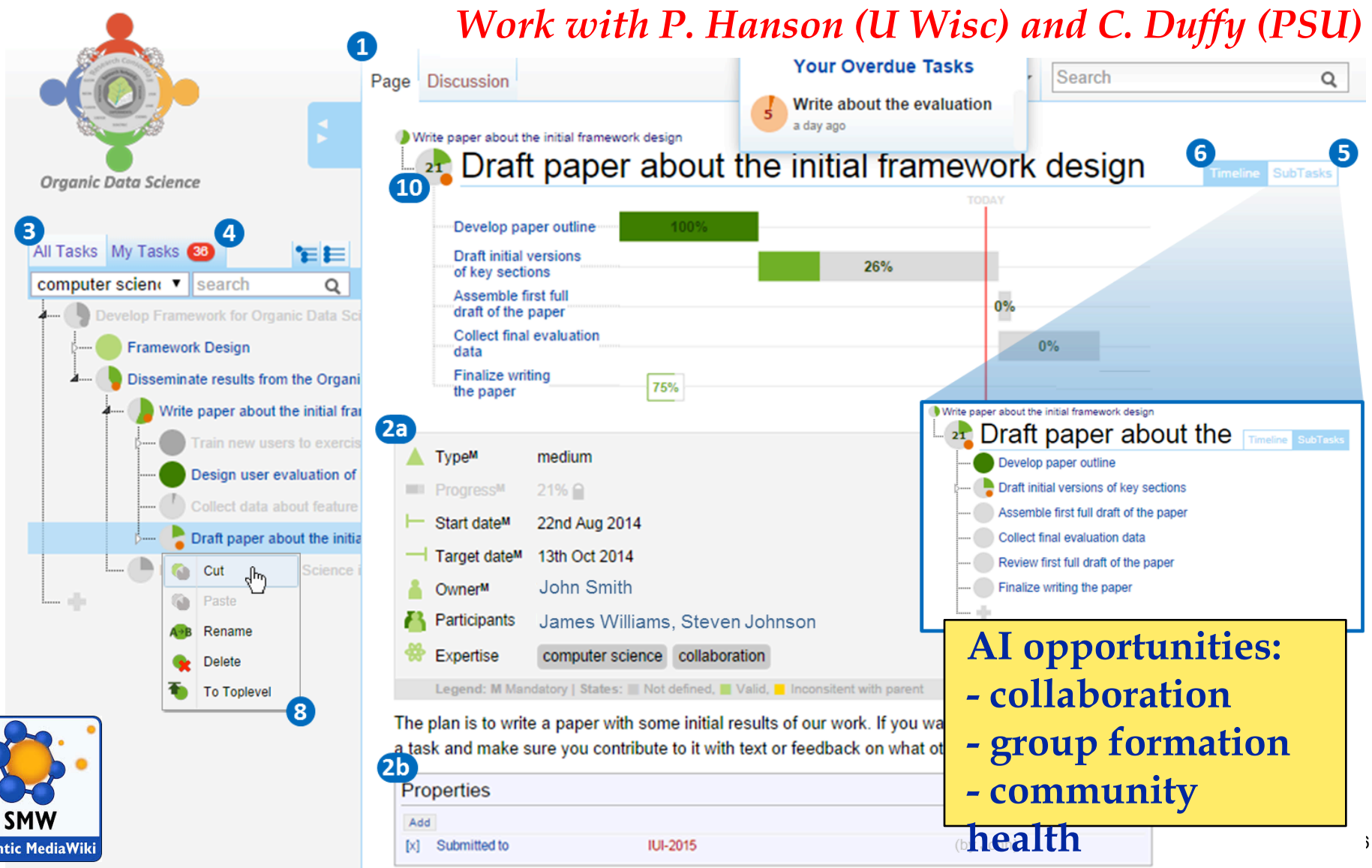




AI opportunities:
- explanation
- reuse
- updating

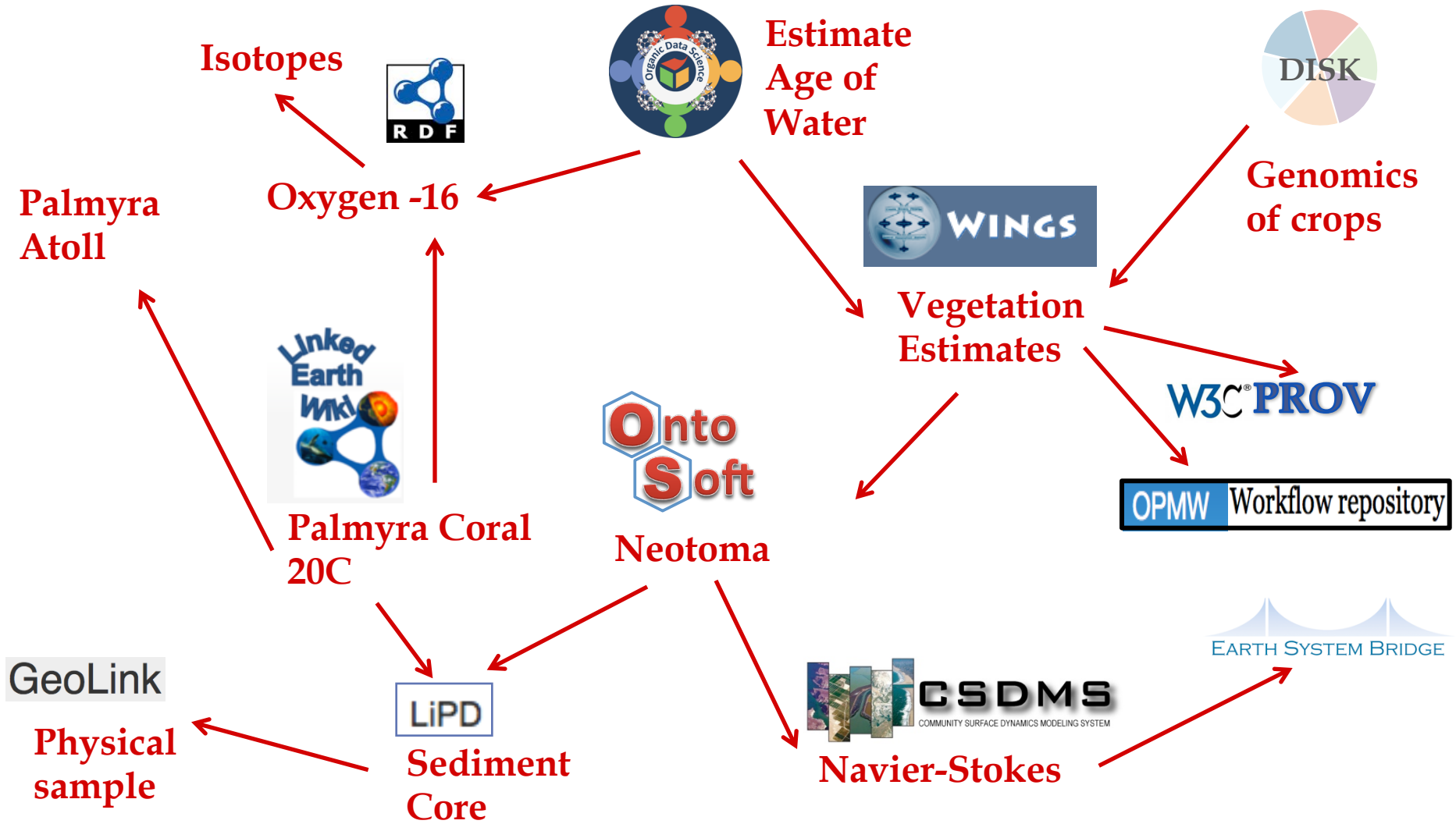
Knowledge about Meta-Processes: Organic Data Science

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

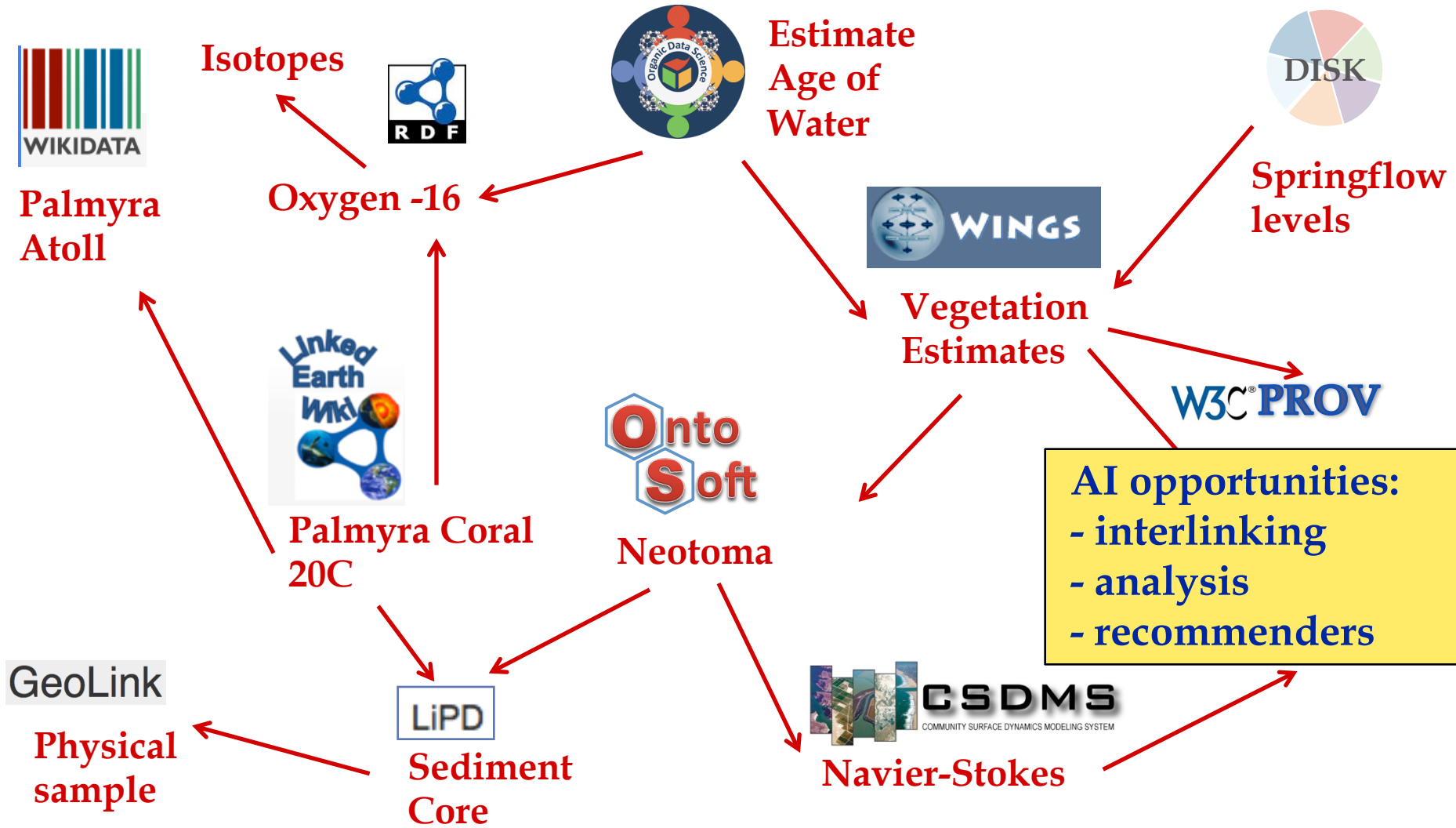


AI opportunities:
- collaboration
- group formation
- community health

Capture and Interlink Scientific Knowledge



Linked Data and Linked Knowledge



Scientific Paper of the Future

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)

Outline

- Artificial intelligence and scientific discovery
  - The knowledge systems tradition
- Our recent work on capturing knowledge about data analysis strategies
  - Hypothesis-driven data analysis
- Representing and capturing data analysis knowledge
  - About data, software, methods, meta-analysis
- Summary of AI challenges

Artificial Intelligence and Scientific Discovery: Challenges for Knowledge Systems

Knowledge Representation Challenges

Metadata:
- collection
- normalization
- organization

Software:
- functional descrs.
- organization
- linking to data

Methods:
- abstraction
- mining
- composition

Provenance:
- explanation
- reuse
- updating

Meta-analysis:
- collaboration
- group formation
- community health

Problem Solving Challenges

Hypothesis-driven discovery:
- A domain-independent framework across scientific disciplines
- Hypothesis representation and evolution
- Capturing analytic knowledge
- Meta-reasoning strategies to select and prioritize analyses
- Aggregation of evidence from multiple types of data/observations
- Generation of explanations from provenance records
- Learning to improve from user feedback
- Identifying "interestingness" in results
- Designing new data analysis methods
- Incorporating science models to guide hypothesis formation