

Semantic Challenges in Getting Work Done

Yolanda Gil

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

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

@yolandagil
gil@isi.edu

Keynote at the International Semantic Web Conference (ISWC)
October 21, 2014

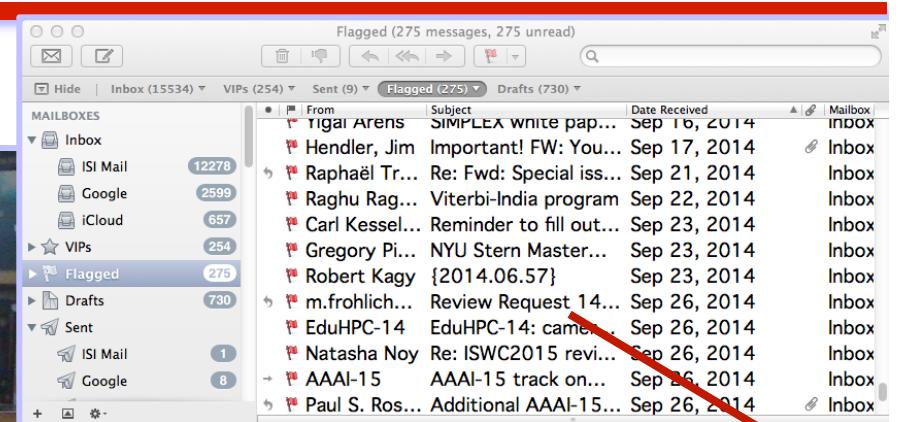
Outline

1. Managing work
 - Personal to do lists
2. Knowledge rich tasks in science
 - Semantic workflows
3. Collaborative tasks in science
 - Organic data science
4. Closing thoughts

Outline

1. Managing work
 - Personal to do lists
 2. Knowledge rich tasks in science
 - Semantic workflows
 3. Collaborative tasks in science
 - Organic data science
 4. Closing thoughts
-
- The diagram consists of four numbered sections (1, 2, 3, 4) listed vertically on the left. To the right of each section is a yellow rectangular box containing the text '2 semantic challenges'. A yellow arrow points from the right side of each section towards its corresponding yellow box.
- 1. Managing work
 - Personal to do lists
 - 2. Knowledge rich tasks in science
 - Semantic workflows
 - 3. Collaborative tasks in science
 - Organic data science
 - 4. Closing thoughts

To Dos



Email
requests

Daily to-dos

To Dos

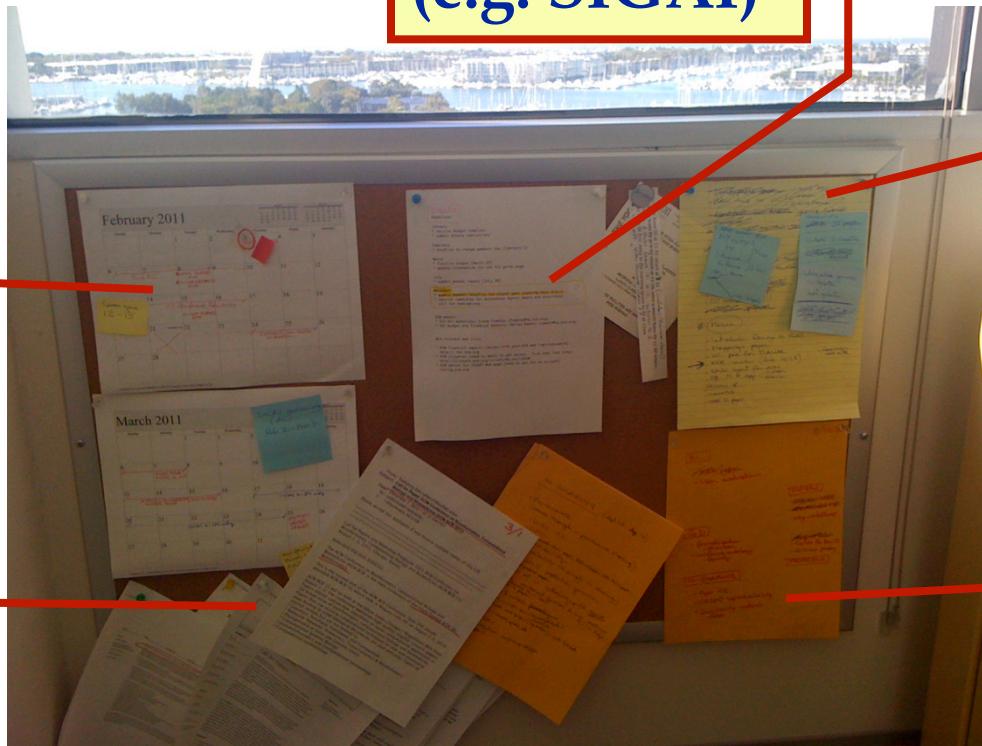
Monthly travel
and deadlines

By annual
timeline
(e.g. SIGAI)

This
week

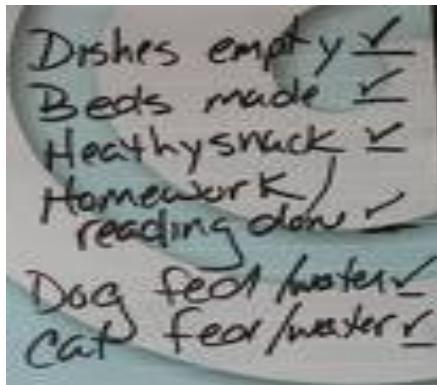
CFPs,
BAAs

By project
(next 6
months)



To Do Lists

- **To do lists are pervasive** [Kirsh 01; Norman 91]
 - Used by more than 60% of people for personal information [Jones & Thomas 97]
 - Used more than calendars, contact lists, etc.
- Prior research focused on user studies
 - [Bellotti et al 04; Dey et al 00]



Wunderlist: To-Do ...
Productivity
★★★★★ 352 Ratings
Essentials
FREE



Toodledo - To Do L...
Productivity
★★★★★ 10 Ratings
FREE

- **Opportunity for assistance**
 - **Major potential impact on productivity**

To Do List Management: Opportunities for Interpretation-based Assistance

To Do List Manager

Automate through agents

Anticipate missing
entries & sub-tasks

Group and organize

Get advice from others

Get information from Web

What Are To Do Items Like: FB app

Renew registration for car
Hotel reservation
Pay bills
Apply for financial aid
Print plane tickets
Renew BOFA card before you leave for summer

DIET!!!

Shoes? Debate... need sneakers

Be a true Christian

Mafia blog update

Think about more Facebook Money Making Ideas!

Ending of reproductive abilities

Get off my lazy arse and start achieving some stuff

Buy Air Blades Mk2 (Imperial)
Buy Ablative Shell (Imperial)
Ruff Racing Hyperblack 278 19'' 275/35 &
245/40 wrapped in NITTO 555R's
Order more AA Eneloop batteries
Return Fan to Westside via UPS
Return ugly jacket

Mettre les images des captures sur Facebook

Spinatch and Bashamel Cupcakes

Skriva kod till Simons webbsida

Watch 'arry pottaaa!

- ~1500 items collected from ~325 people
- Many are not amenable to automation
- Many could be automated fully or in part

What Are To Do Items Like: Office

- Unusual structure
 - No verb: “quarterly report to Joe”
 - Abbreviations (also typos): “Sched wed 15 ISI”
 - Questions: “How to extract data for Steve”
- Many ways to refer to the same task
 - “Meet with John about paper”, “Discuss paper with John”, ...
- Incomplete task specifications
 - “Schedule meeting with John”
- Ambiguous references out of context
 - “Meet about paper”
 - “Meet with Raytheon folks”
- Personal items
 - “Walk the dog”

- Corpus of 2400 to-do entries from users of CALO office assistant
- 77% lack a verb
- 56% missing at least one argument
- 14% could be automated by agents

Opportunities

To Do List Manager

Automate through agents

Anticipate missing
entries & sub-tasks

Group and organize

Get advice from others

Get information from Web

Agents

Colleagues

On-Line
Resources

Agents:

Beamer for CALO and Radar [Gil & Ratnakar AAAI 2008]

- Match agent capabilities to user's to dos

To Do

- Set up discussion with Bill on ISWC paper

?

Calendar Agent

- SchMtg <person> <topic> <time> <loc>

?

A2

- Action2 <x1> <x2>

?

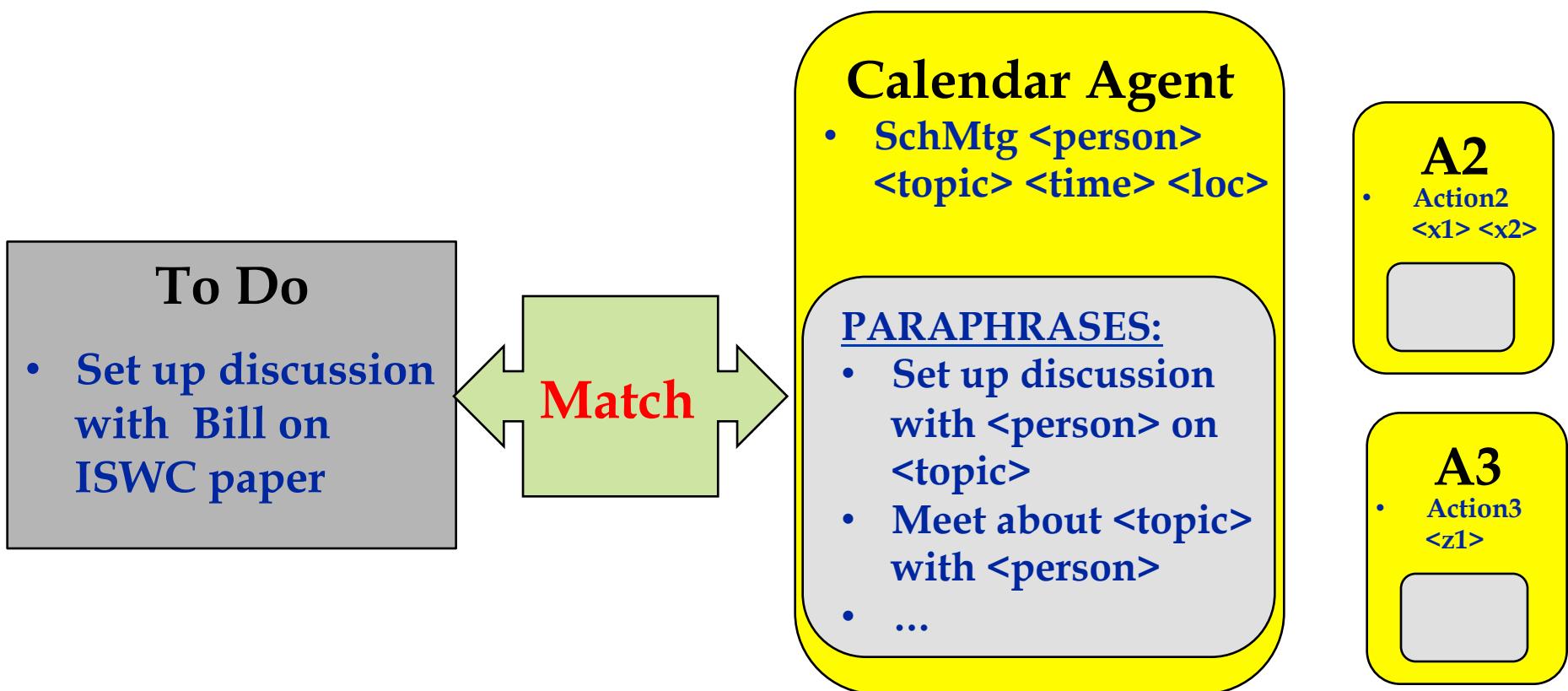
A3

- Action3 <z1>

Agents:

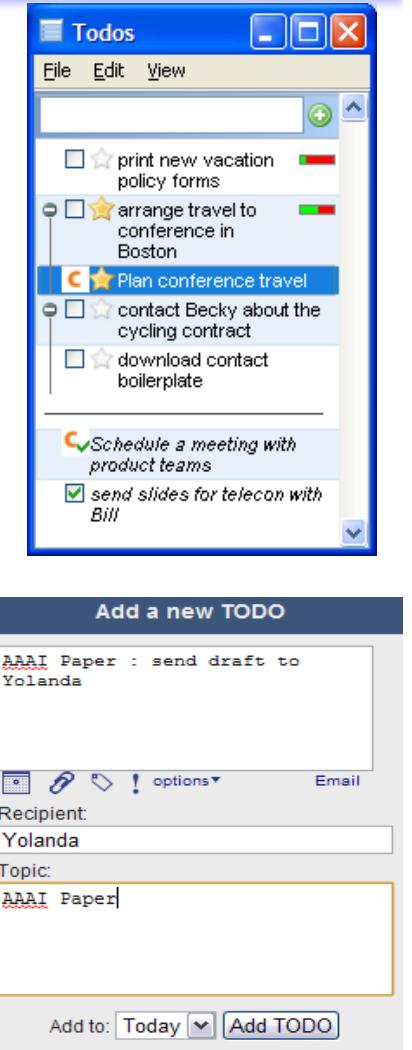
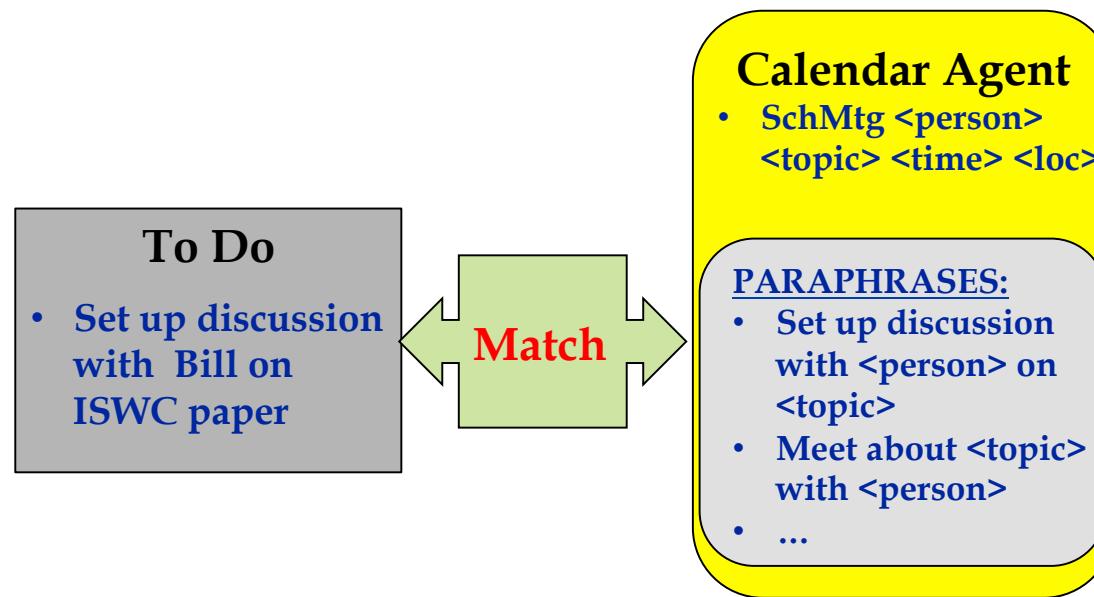
Beamer for CALO and Radar [Gil & Ratnakar AAAI 2008]

- Use paraphrase patterns of agent capabilities to match them to user's to dos



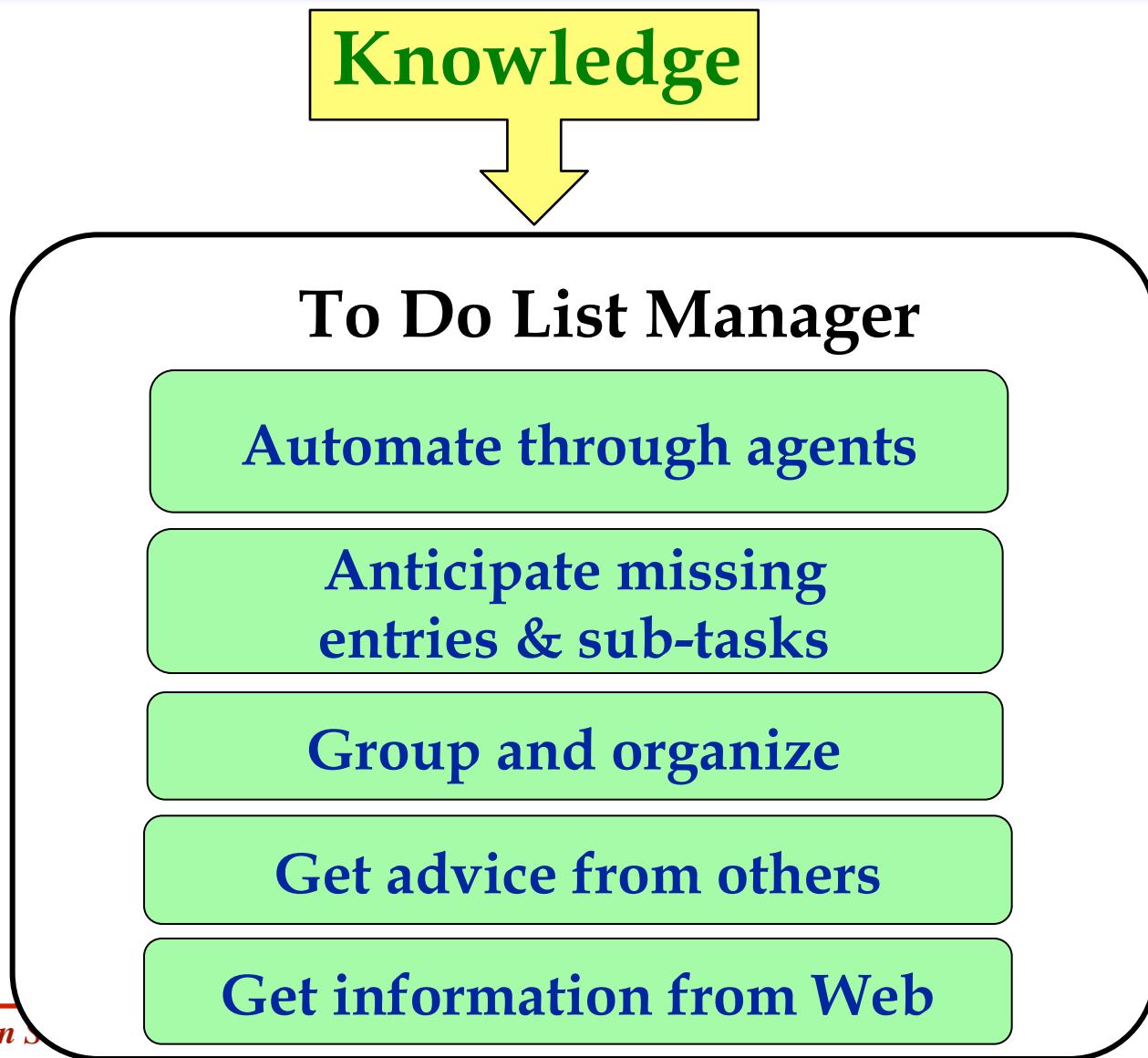
Agents: Beamer for CALO and Radar [Gil & Ratnakar AAAI 2008]

- Use paraphrase patterns of agent capabilities to match them to user's to dos



- Evaluation with CALO office assistant corpus
 - 86.7% accuracy in detecting relevance to agents
 - only 0.2 to 0.4 edits needed to set up task parameters

The Need for Semantics



Paraphrase Game [Chklovski 2005]



"this can help you"

Another Way To Say It:

this could be helpful



Already Tried

this could be of help

Hints

this could ...
Nice ...
... should ... better
... help.

Score:

0

You Can Win: 420

Common (Sense) Knowledge

[Chklovski and Gil, K-CAP 2005, AAAI 2005]

You Have Taught Me:

Learner2

A copier is also typically used to

Previously, I have been told:

duplicate AGREE DISAGREE

Now, I ask:

A copier has a piece or a part called a paper tray

(question)

duplicate a document



sonable

Example: A toothbrush has a piece or a part called a handle.

A copier is typically used to copy a document

Unreasonable question

Example: A pen is typically used to write a letter.

As an admin assistant, if helping with setting up a videoconference, if you need to deal with a conference time, an important activity may be: agree upon it

700,000+ statements
collected from over
3,000 users

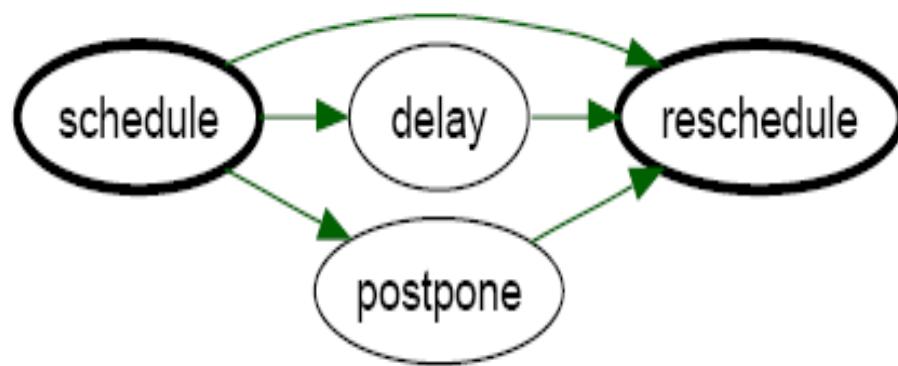
When preparing a visitor's meeting schedule, it is important that you check (a/an) room availability

AGREE DISAGREE SORT OF

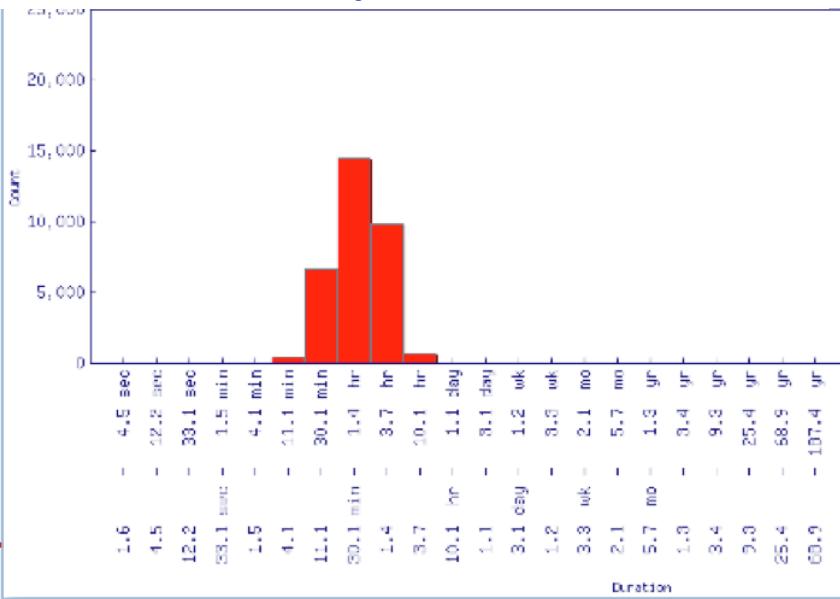
Possible problem: When attending a meeting, not having an LCD projector may cause a problem.

Possible remedy: When attending a meeting, one way to address not having an LCD projector is to locate a portable projector

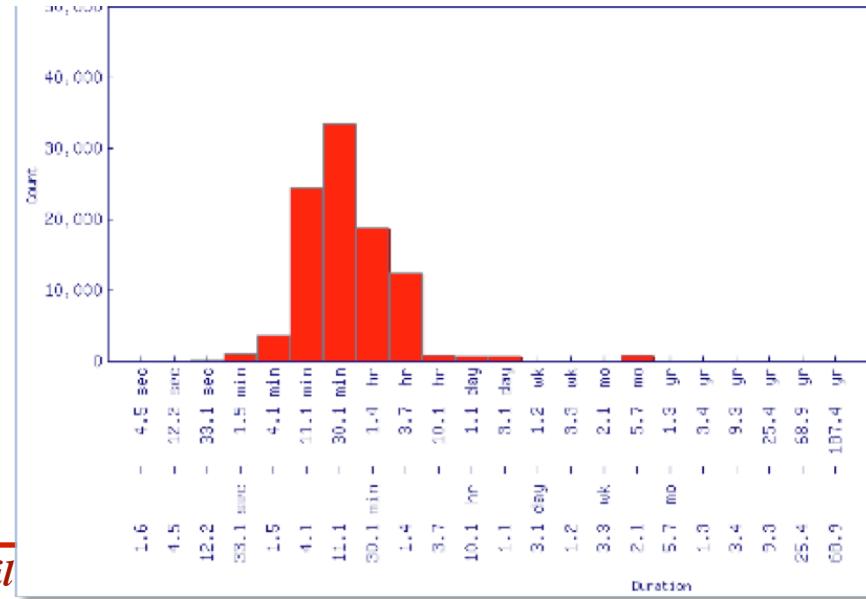
VerbOcean [Chkovski and Pantel, IJCNLP 2005]



“Lunch” – likely duration 1hr



“Presentation” – likely duration 10mins

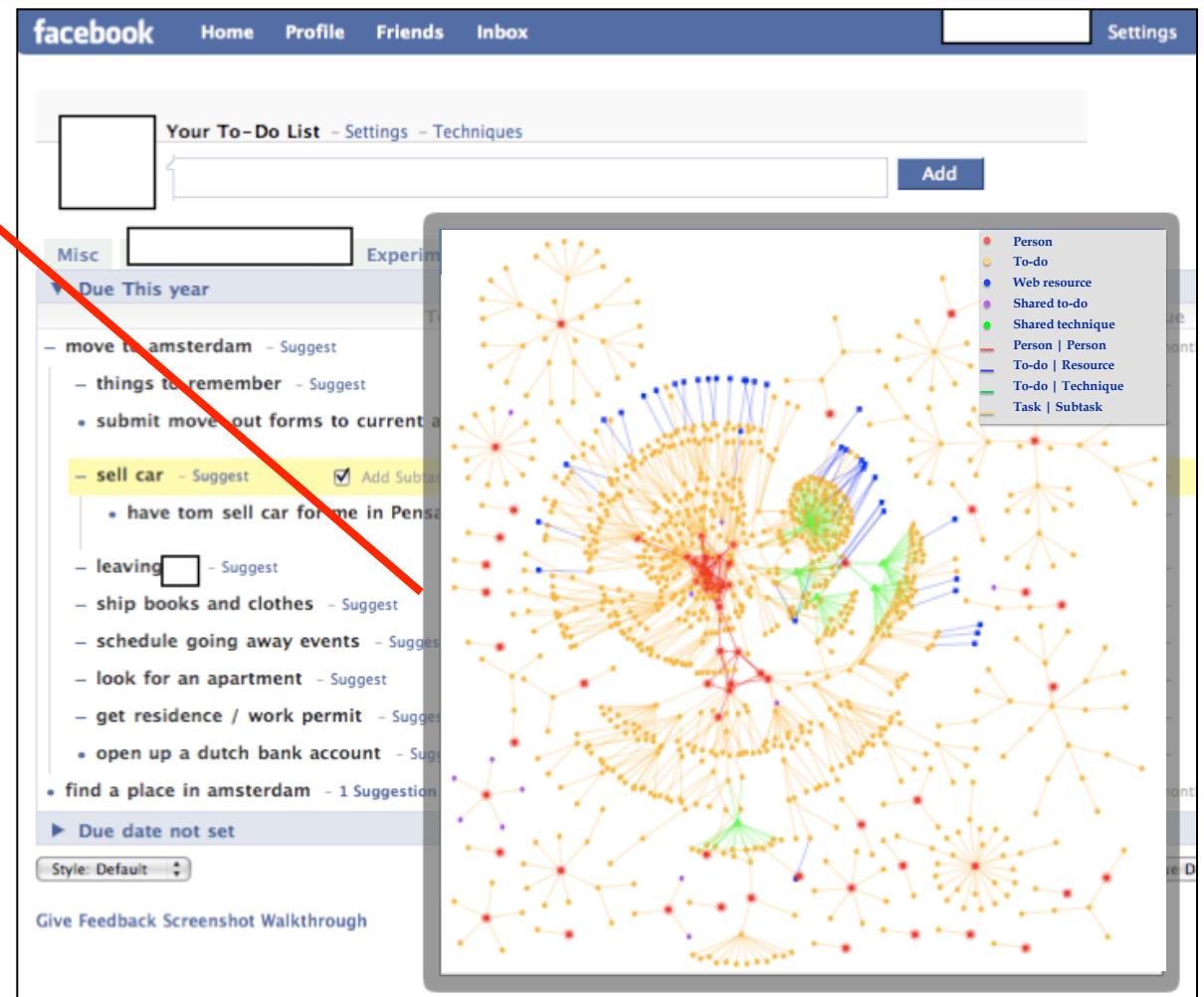


Managing To Dos through Colleagues: Social Task Networks [Groth et al 2010]

■ To Do app for FB

Social task networks

- People linked to their to-dos
- To-dos linked to their subtasks
- Tasks are linked to URIs which link to web resources



■ Open Task Repository using Linked Data Principles

Managing To Dos through On-Line Resources

[Vrandecic and Gil IUI 2011]

The screenshot shows the homepage of the **HowToDo** website. The header features the site's logo with "How" in blue and "To Do" in red/pink, followed by "Home" in a smaller blue font. Below the logo is a search bar containing the query "pack for a camping trip". To the right of the search bar are three links: "About", "API", and "Contact". A "Submit" button is located to the right of the search bar.

Below the search bar, a message reads: "[debug] Displaying 6 How-Tos that we think may match. [Tell us if they do!](#) Search powered by [Google](#).

The main content area displays six search results, each with a title, steps, source, and a brief description:

- Pack for a Camping Trip** (13 steps, Source: [wikiHow](#))
Find out where you are going, how long, what you will be staying in, etc
- Pack Light for a Camping Trip** (5 steps, Source: [wikiHow](#))
Three months before your camping trip, do a full inventory of your camping supplies
- Prepare for a Camping Trip** (8 steps, Source: [wikiHow](#))
Think about what you have to bring and what you want to bring
- Pack for a Camping Trip As a Teen Girl** (10 steps, Source: [wikiHow](#))
Think about how much luggage is appropriate
- Look Good on a Camping Trip (Girls)** (7 steps, Source: [wikiHow](#))
Make a list of things to bring! (See Things You'll Need for a list of suggestions) Make sure you take enough things but don't overdo it as you may bring too much and not find anything.
- Pack for a Week Away From Home**
Go over your list

Some Readings

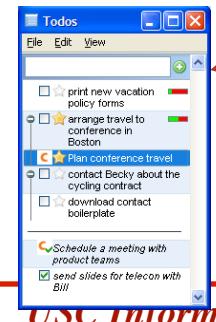
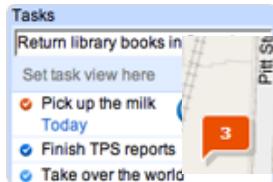
- Yolanda Gil, Varun Ratnakar, Timothy Chklovski, Paul T. Groth, Denny Vrandecic: “**Capturing Common Knowledge about Tasks: Intelligent Assistance for To Do Lists.**” ACM Transactions on Interactive Intelligent Systems, 2(3). 2012.
- Hans Chalupsky, Yolanda Gil, Craig A. Knoblock, Kristina Lerman, Jean Oh, David V. Pynadath, Thomas A. Russ, Milind Tambe: “**Electric Elves: Agent Technology for Supporting Human Organizations.**” AI Magazine 23(2): 11-24 (2002)

A Semantic Challenge: Managing Personal To Dos

To-Do list interfaces

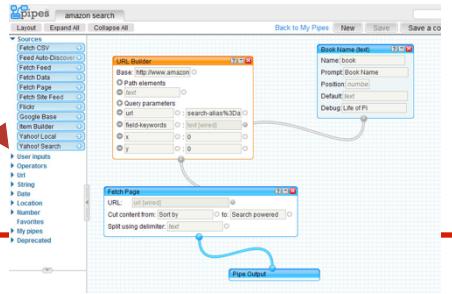
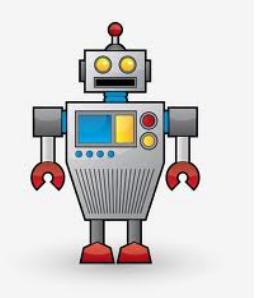
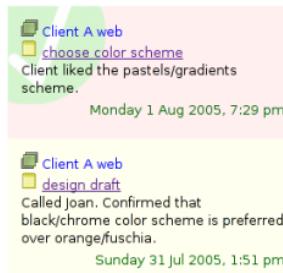
Personal
 travel buy Aruba tickets 0/1 8/2/2005
 car get oil change 0/0.5 8/2/2005
 1.5h remaining in 2 tasks »

Work
 Client A web design draft 4/8 7/29/2005
 Client A web call Mike to get specs 0/2 8/2/2005
 Client B pitch design refresh 0/1 8/2/2005
 7h remaining in 3 tasks »

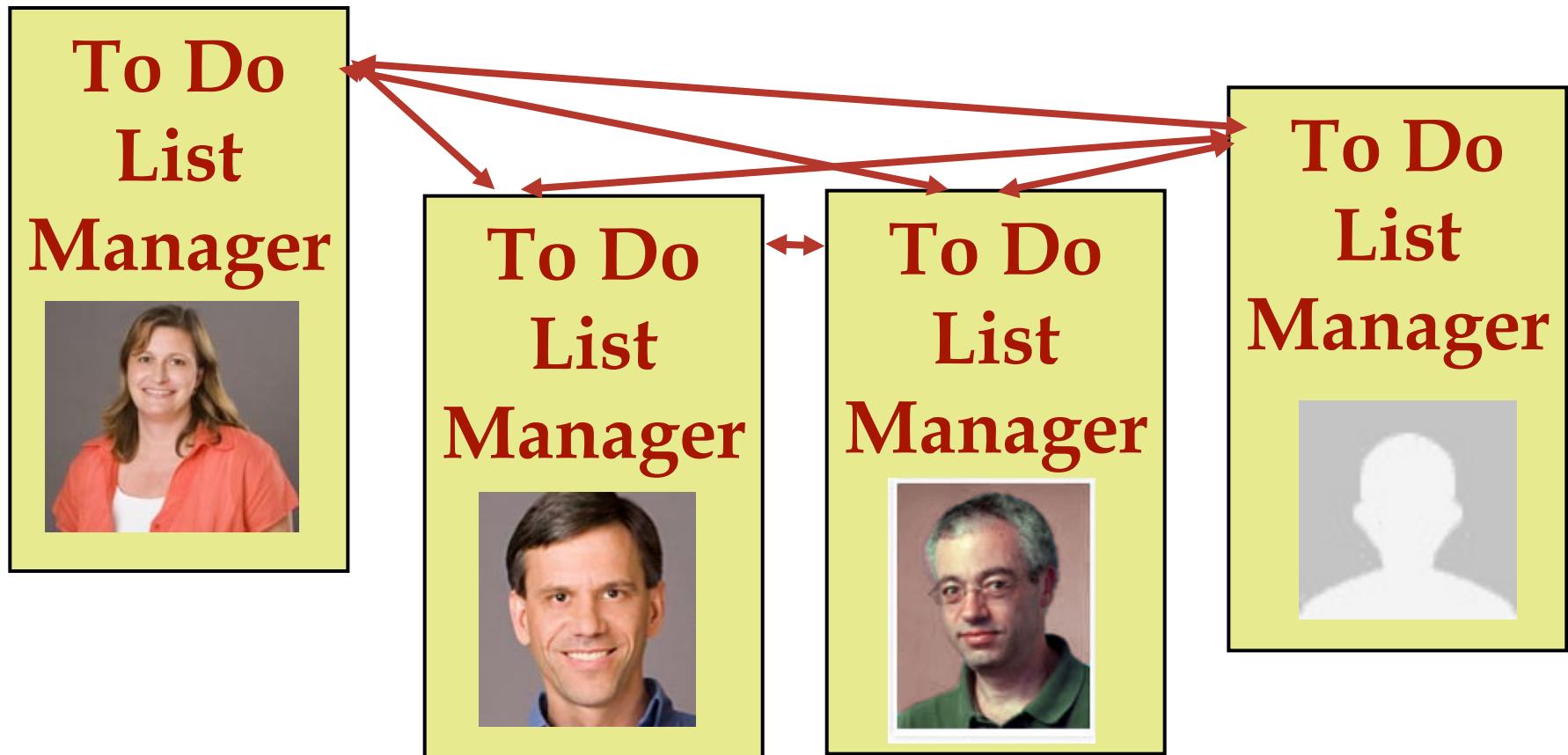


To Do List Manager

Agents/services, other people, advice web sites



A Semantic Challenge: Coordinating To Dos of Different People



Semantic Challenges in Getting Work Done

- To dos
 - Managing personal to dos
 - Managing coordinated to dos
- Knowledge rich tasks in science
- Open science

Data-Intensive Computing in Science



WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

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

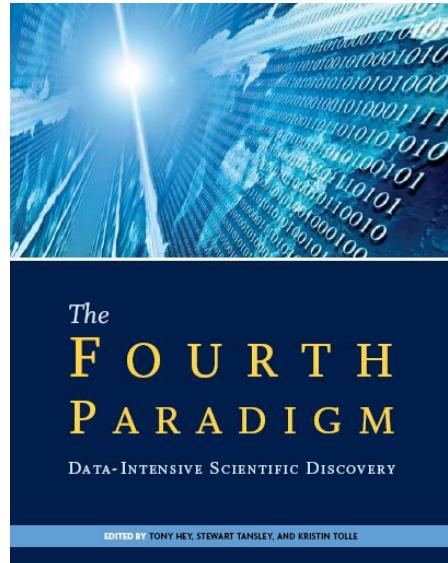
By Chris Anderson 06.23.08



: wrong, but some are



stitute



Yolanda Gil



BIBLIOMETRICS AND CITATION ANALYSIS

From the Science Citation Index to Cybermetrics



gil@isi.edu

The Bottleneck is the Process, Not the Data!

- Today: significant human bottleneck in the scientific process

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?

- Need to help machines understand the scientific research process in order to assist scientists
 - **Semantics can be a game changer**

Text Extraction in Hanalyzer (L. Hunter, U. Colorado)



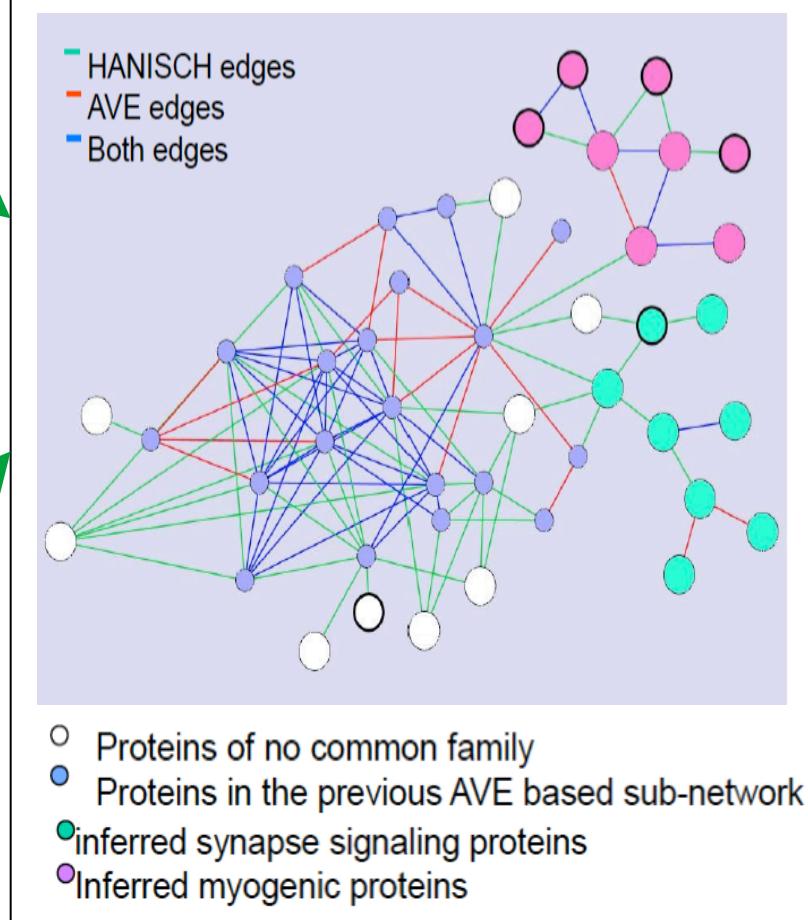
Text extraction from publications

The significance of the interaction between DAZAP1 and DAZL/DAZ remains to be defined. These proteins may act together to facilitate the expression of a set of genes in germ cells. For example, DAZAP1 could be involved in the transport of the mRNAs of the target genes of DAZL. Alternatively, DAZL and DAZAP1 may act antagonistically to regulate the timing and the level of expression. Such an antagonistic interaction between two interacting RNA-binding proteins is exemplified by the neuron-specific nuclear RNA-binding protein, Nova-1. Nova-1 regulates the alternative splicing of the pre-mRNAs encoding neuronal inhibitory glycine receptor $\alpha 2$ (GlyR $\alpha 2$) [23]. The ability of Nova-1 to activate exon selection in neurons is antagonized by a second RNA-binding protein, brPTB (brain-enriched polypyrimidine tract binding protein), which interacts with Nova-1 and inhibits its function [24]. DAZAP1 could function in a similar manner by binding to DAZL and inhibiting its function. Comparing the phenotypes of Dazl1 and Dazap1 single and double knock-out mice may provide some clues to the significance of their interaction. Dazl1 knock-out mice have already been generated and studied [6]. The spermatogenic defect in the male becomes apparent only after day 7 post partum when the germ cells are committing to meiosis (H. Cooke, personal communication). The genomic structure of Dazap1, delineated here, should facilitate the generating of Dazap1 null mutation.



Semantic integration of biomedical databases

Generation of interesting new hypotheses



Robot Scientist [King et al 2009]



Science 3 April 2009:
Vol. 324 no. 5923 pp. 85–89
DOI: 10.1126/science.1165620

REPORT

The Automation of Science

Ross D. King^{1,*}, Jem Rowland¹, Stephen G. Oliver², Michael Young³, Wayne Aubrey¹, Emma Byrne¹, Maria Liakata¹, Magdalena Markham¹, Pinar Pir², Larisa N. Soldatova¹, Andrew Sparkes¹, Kenneth E. Whelan¹, Amanda Clare¹

Science

AAAS

Intelligent Science Assistants

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?

Timely Analysis of Environmental Data

[Gil et al ISWC 2011]

With Tom Harmon (UC Merced), Craig Knoblock and Pedro Szekely (ISI)



California's Central Valley:
• Farming, pesticides, waste
• Water releases
• Restoration efforts



NOAA NATIONAL ATMOSPHERIC UNITED STATES

COUNTDOWN TO LAUNCH

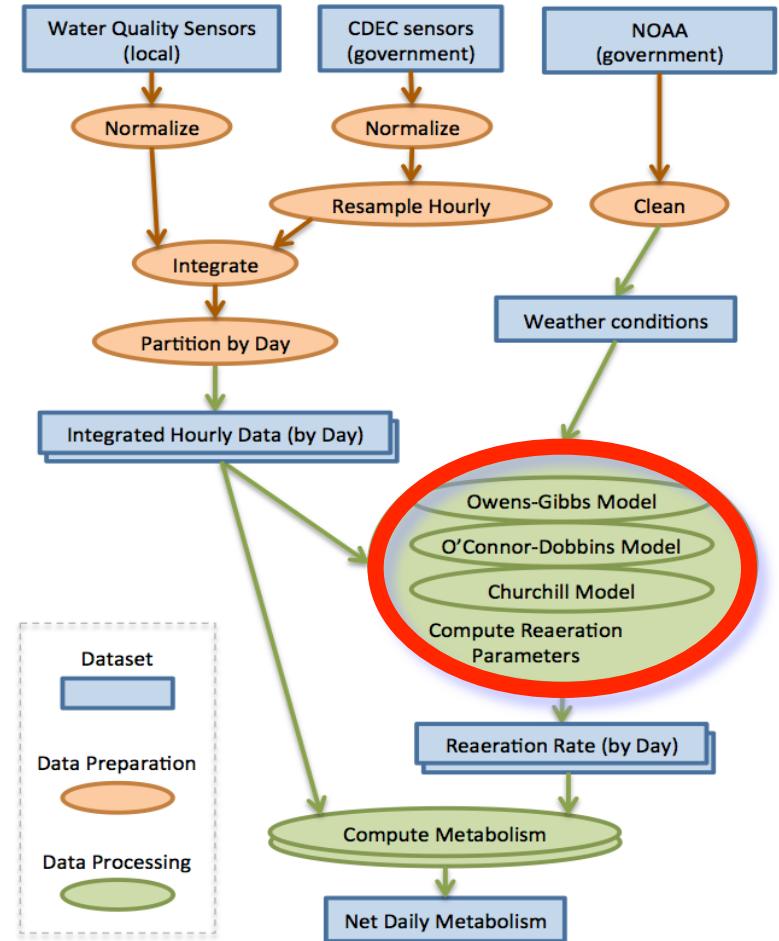
CA.GOV DEPARTMENT OF WATER RESOURCES California Data Exchange Center

MERCED RIVER NEAR STEVINSON

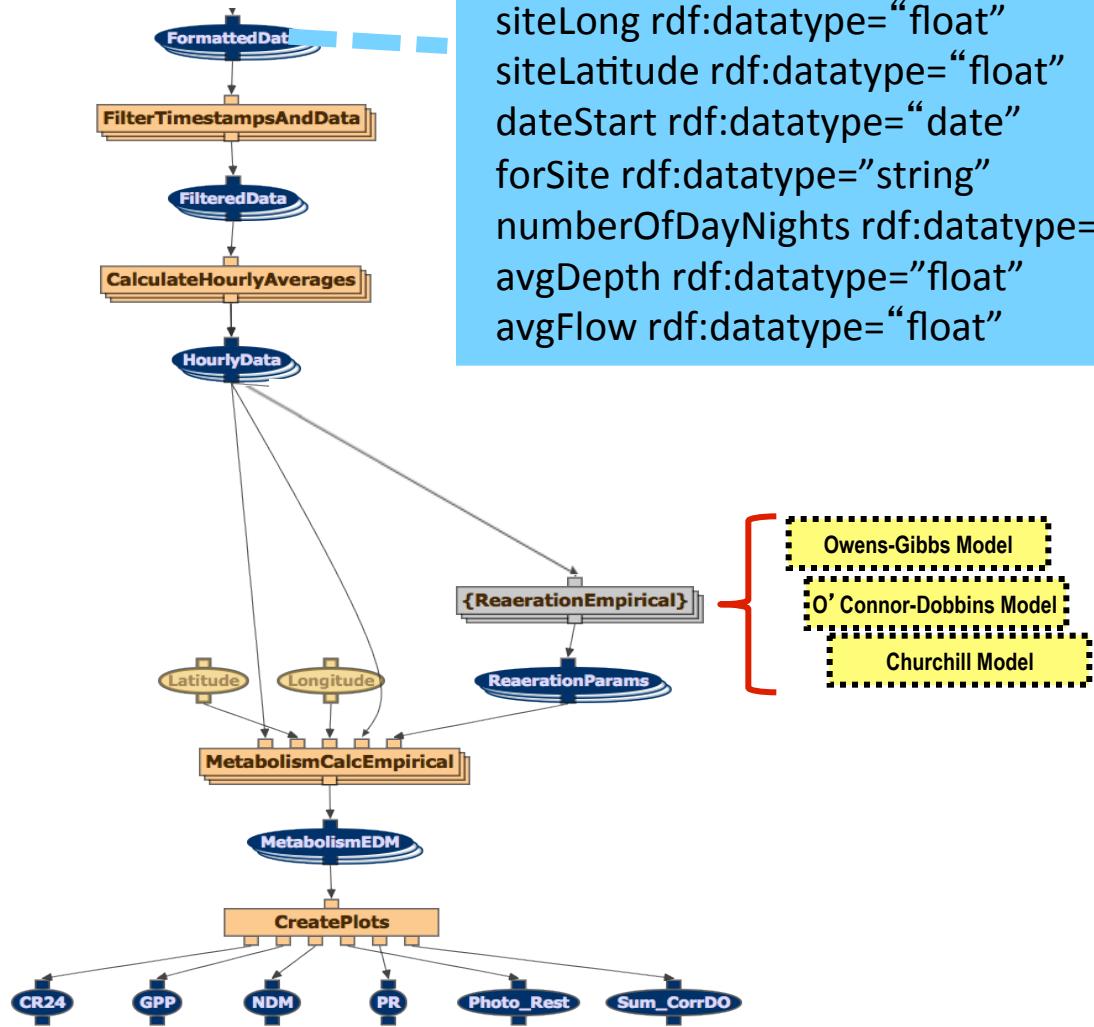
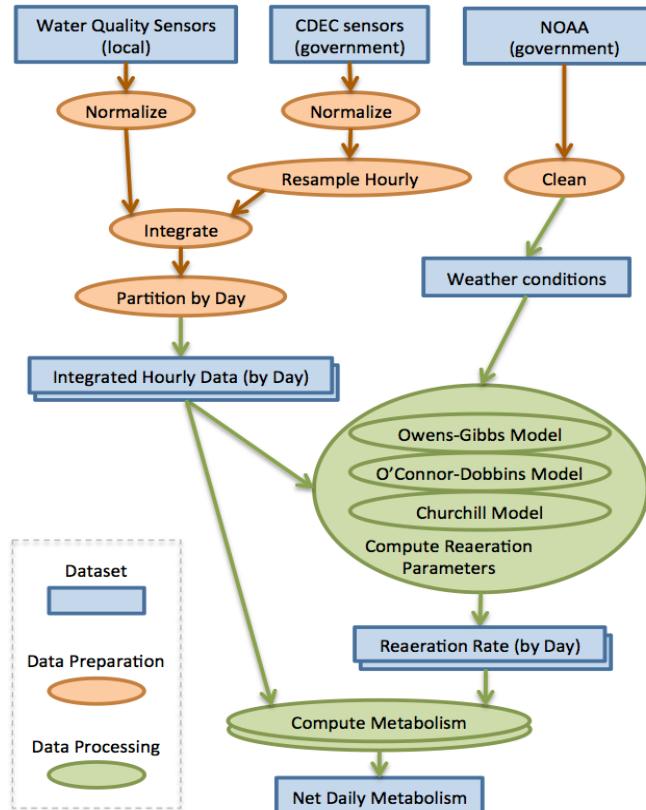
Station ID	MST	Elevation	62' ft
River Basin	MERCED R	County	MERCED
Hydrologic Area	SAN JOAQUIN RIVER	Nearby City	STEVINSON
Latitude	37.371000°N	Longitude	120.931000°W
Operator	CA Dept of Water Resources	Data Collection	

The following data types are available online. Select one of the links below to retrieve recent data.

Sensor Description	Duration	Plot	Data Collection	Data /
ELECTRICAL CONDUCTIVITY MICRO S, us/cm	(daily)	(EL COND)	COMPUTED	07/01/200
FLOW, MEAN DAILY, cfs	(daily)	(M FLOW)	COMPUTED	03/30/199
TEMPERATURE, WATER, deg f	(daily)	(TEMP W)	COMPUTED	07/01/200
BATTERY VOLTAGE, volts	(event)	(BAT VOL)	SATELLITE	02/08/200
FLOW, RIVER DISCHARGE, cfs	(event)	(FLOW)	COMPUTED	03/20/199
RIVER STAGE, feet	(event)	(RIV STG)	SATELLITE	03/20/199



A Semantic Workflow



DailySensorData

```

isa Hydrolab_Sensor_Data
siteLong rdf:datatype="float"
siteLatitude rdf:datatype="float"
dateStart rdf:datatype="date"
forSite rdf:datatype="string"
numberOfDayNights rdf:datatype="int"
avgDepth rdf:datatype="float"
avgFlow rdf:datatype="float"
  
```

Semantic Workflows in Wings

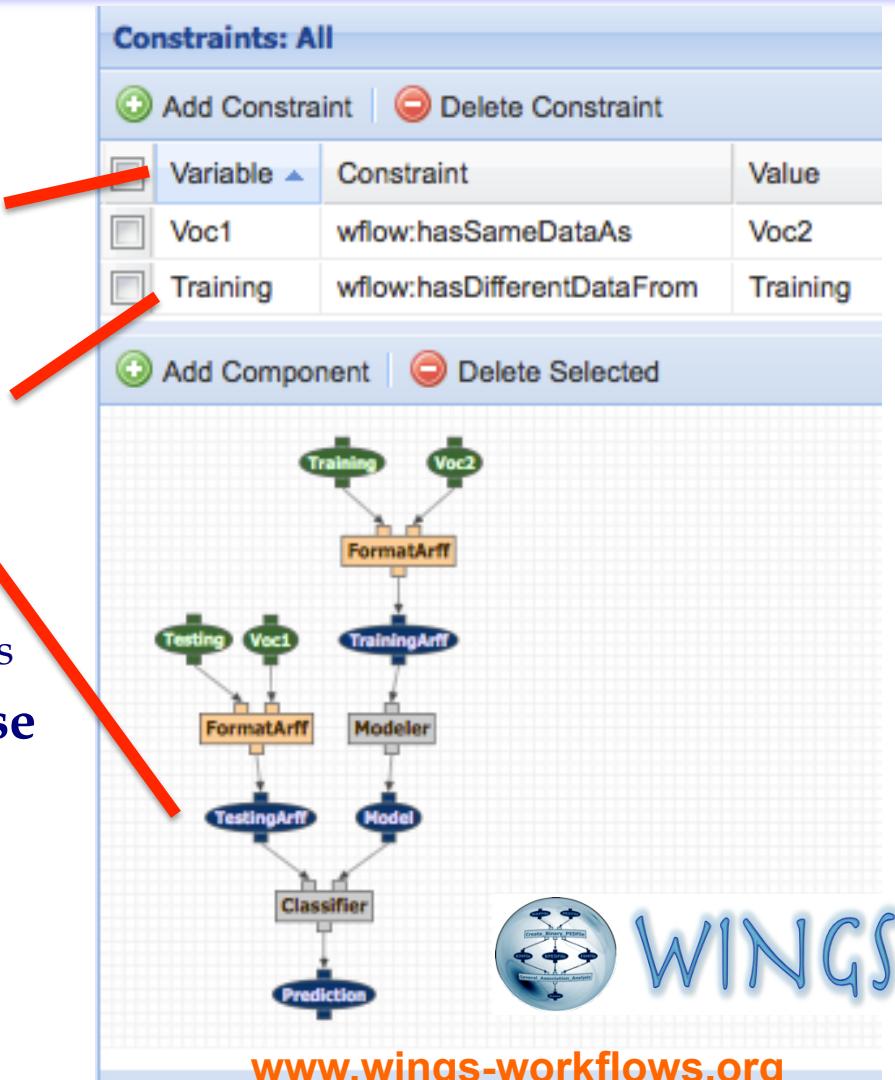
[Gil et al 10][Gil et al 09][Kim & Gil et al 08][Kim et al 06]

■ Workflows are augmented with semantic constraints

- Each workflow constituent has a **variable** associated with it
 - Workflow components, arguments, datasets
- **Constraints** are used to restrict workflow variables
- Can define **abstract classes of components**
 - Concrete components model exec. codes

■ Workflow reasoners propagate and use semantic constraints

- ## ■ Uses semantic web standards: OWL/RDF, SPARQL
- ## ■ Compilation of workflows to scalable execution infrastructure



Semantic Components in WINGS

[Gil iEMSSs 2014]

Wings Platform

Home Analysis

Components: Default

- + Add Component - Delete
- CalculateHourlyAveragedInput
- ConvertToStandardFormat
- CreateParametersFromData
- CreatePlots
- FilterTimestampsAndData
- HourlyAverage
 - Aquaflow_Hourly_Average_Inputs
- MergeMetabolismResults
- Metabolism
 - MetabolismDay
 - MetabolismCalcEmpirical
 - MetabolismCalculation
- ReaerationEmpirical
 - ReaerationCM
 - ReaerationODM
 - ReaerationOGM
- ReaerationPhysics
- ReaerationEDM

I/O			Rules	Inherited Rules	Documentation
Input Data					
Name	Type	Prefix			
InputParameters	dcdom:Hourly_Averaged_Input	-i1			
Input Parameters					
Name	Type	Prefix			
velocity	xsd:float				
depth	xsd:float				
Output Data					
Name	Type	Prefix			
K2Result	dcdom:K2_Data	-o1			

I/O			Rules	Inherited Rules	Documentation
Input Data					
Name	Type	Prefix			
InputParameters	dcdom:Hourly_Averaged_Input	-i1			
Input Parameters					
Name	Type	Prefix			
velocity	xsd:float	-p2			
slope	xsd:float	-p4			
depth	xsd:float	-p1			
flow	xsd:float	-n3			
Output Data					
Name					
K2Result					

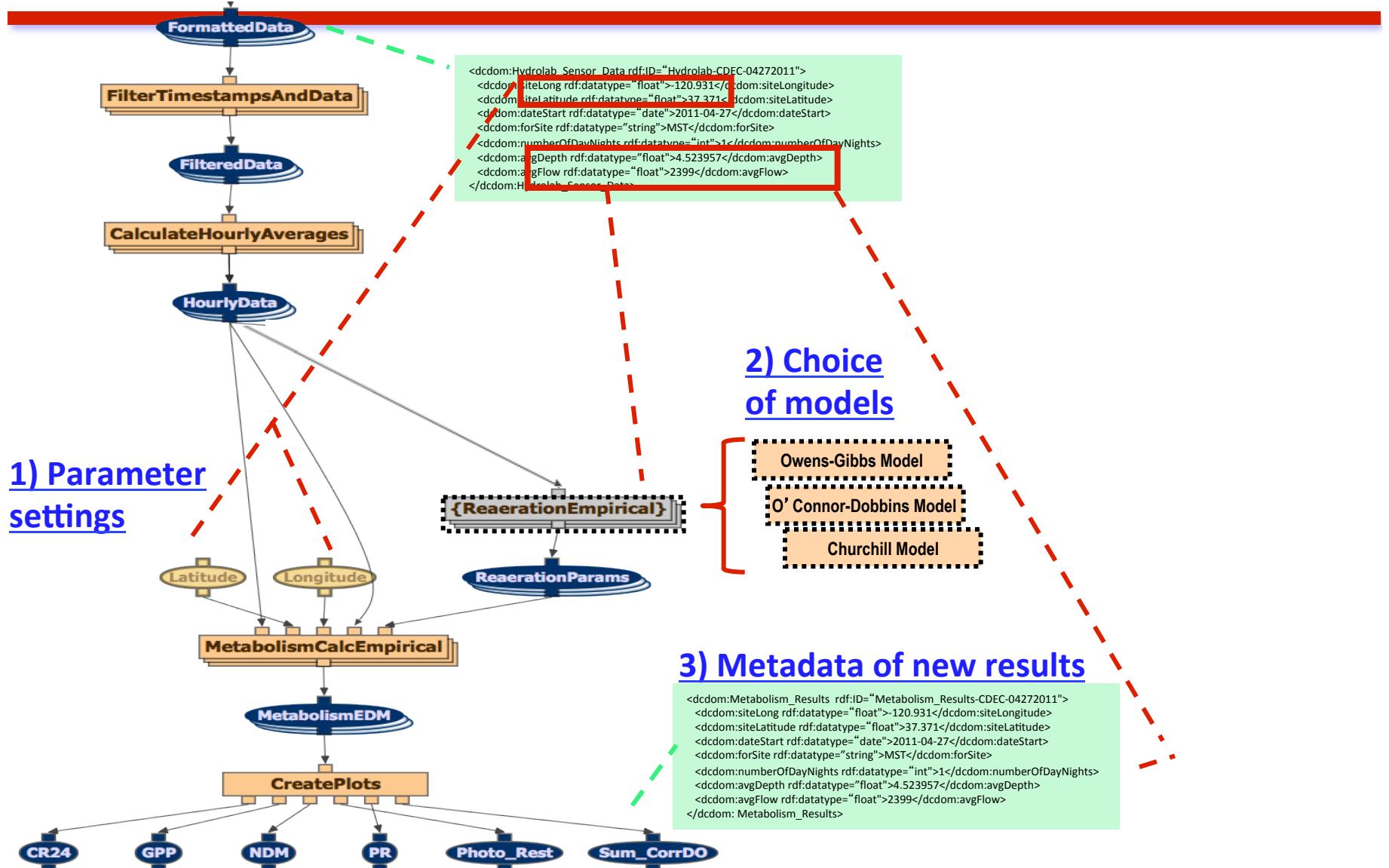
Yolanda Gil

I/O Data constraints

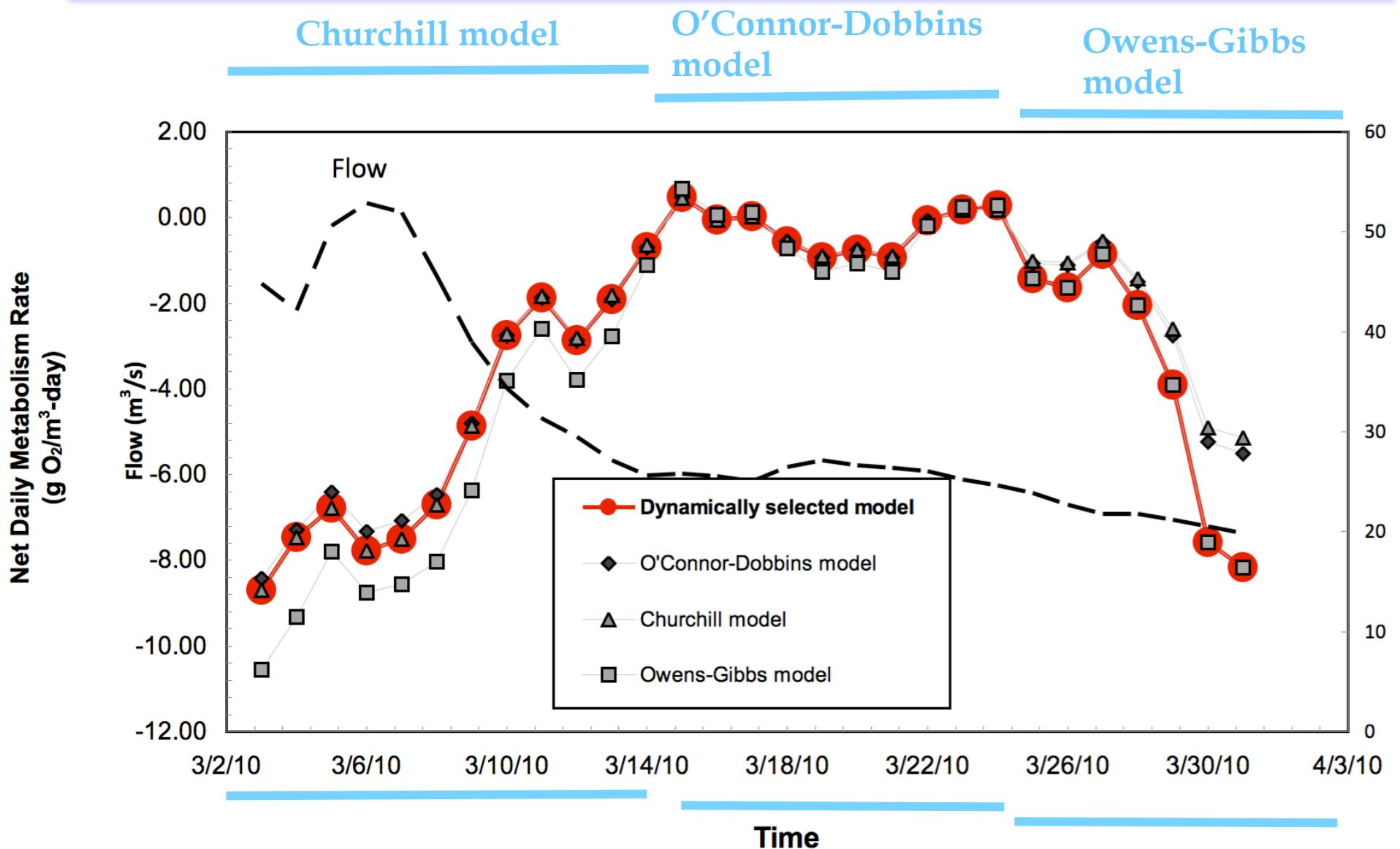
Use constraints

```
;; Depth must be over .6m
[ CMInvalidity1:
  (?c rdf:type pcdom:ReaerationCMClass)
  (?c pc:hasInput ?idv)
  (?idv pc:hasArgumentID
  'InputParameters')
  (?idv dcdom:depth ?depth)
  le(?depth '0.61')
  -> (?c pc:isValid 'true'))]
```

WINGS Specializes Workflow Based on Characteristics of Daily Data

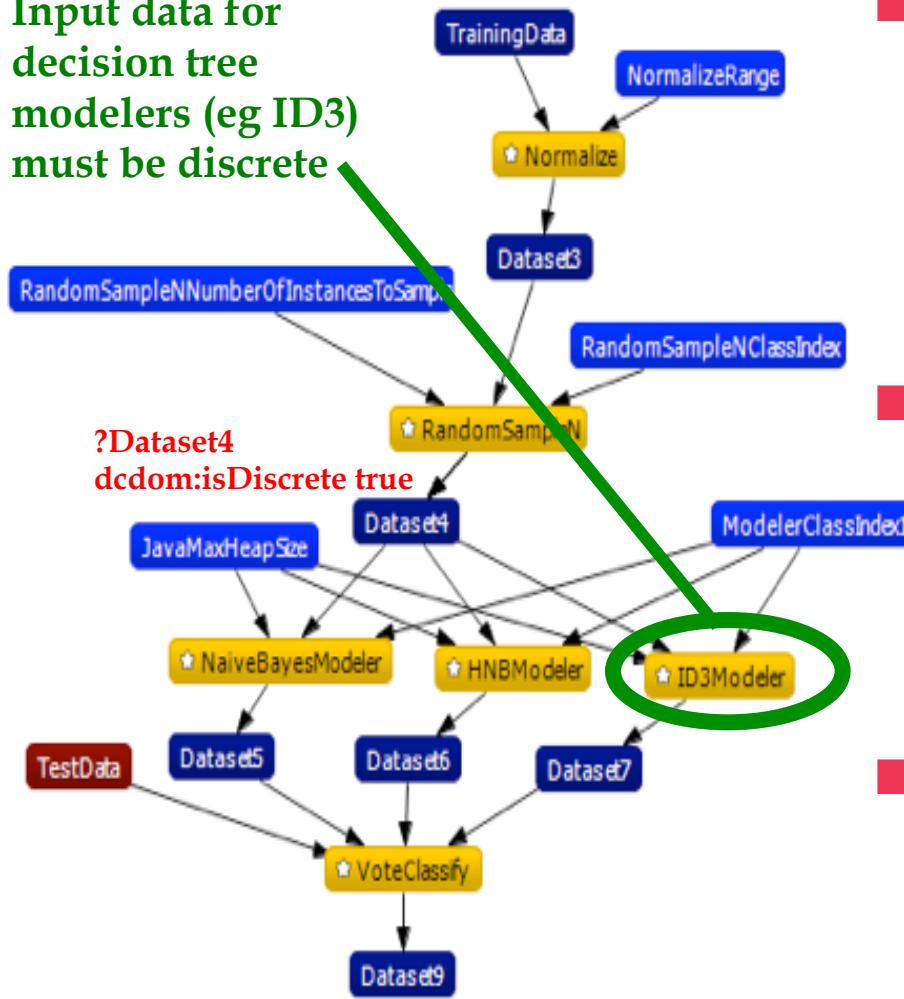


WINGS Dynamically Selects Appropriate Model Based on Daily Sensor Readings



WINGS Workflow Reasoners

Input data for decision tree modelers (eg ID3) must be discrete



■ **Key idea:** Skeletal planning, where constraints for each component are propagated through a fixed workflow structure (the skeleton)

■ **Phase 1: Goal Regression**

- Starting from final products, traverse workflow backwards
- For each node, query for constraints on inputs

■ **Phase 2: Forward Projection**

- Starting from input datasets, traverse workflow forwards
- For each node, query for constraints

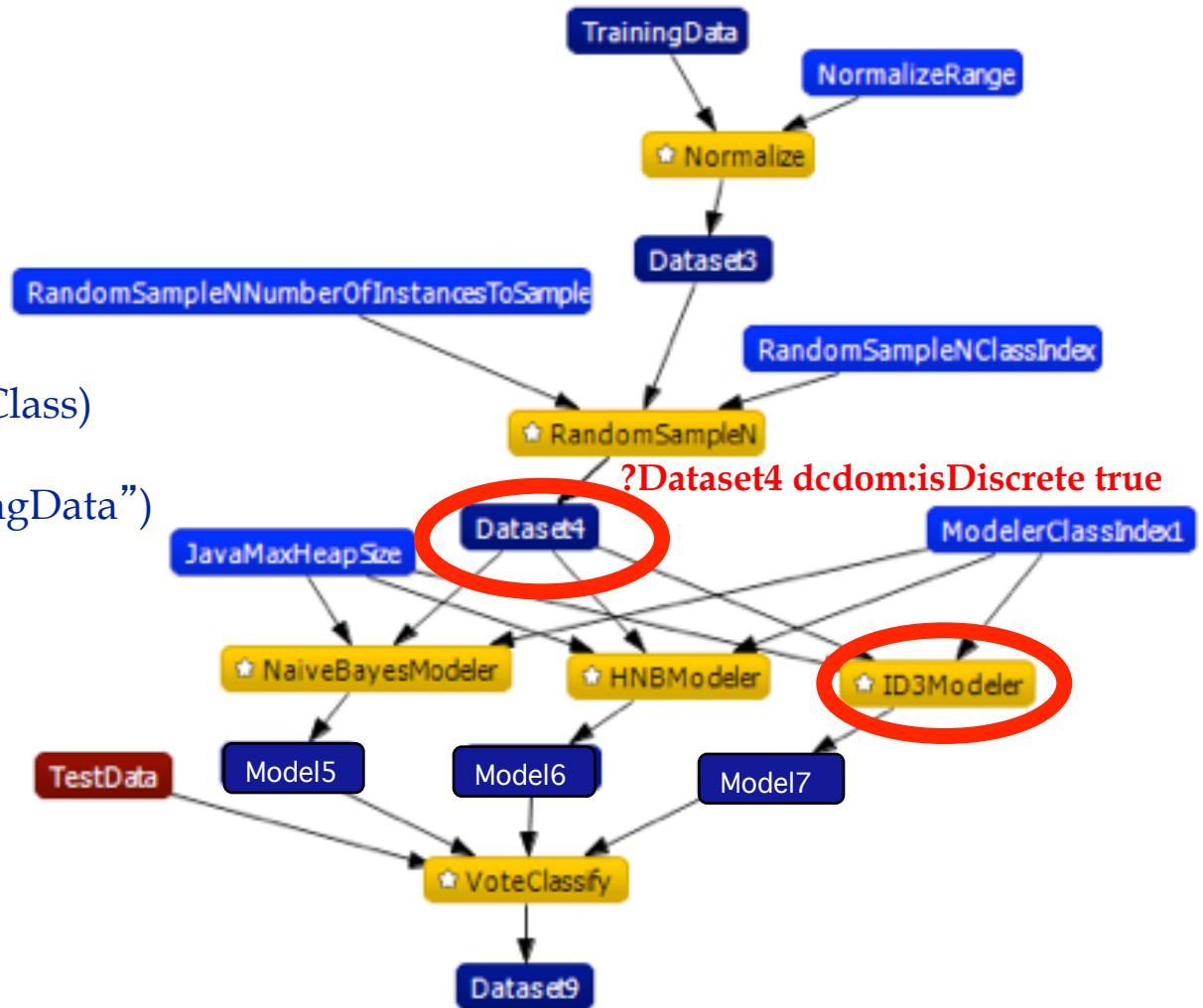
Example (Step 1 of 5)

Rule in Component Catalog:

[modelerSpecialCase2:

```
(?c rdf:type pcdom:ID3ModelerClass)
(?c pc:hasInput ?idv)
(?idv pc:hasArgumentID "trainingData")
```

```
-> (?idv dcdom:isDiscrete
"true"^^xsd:boolean)]
```



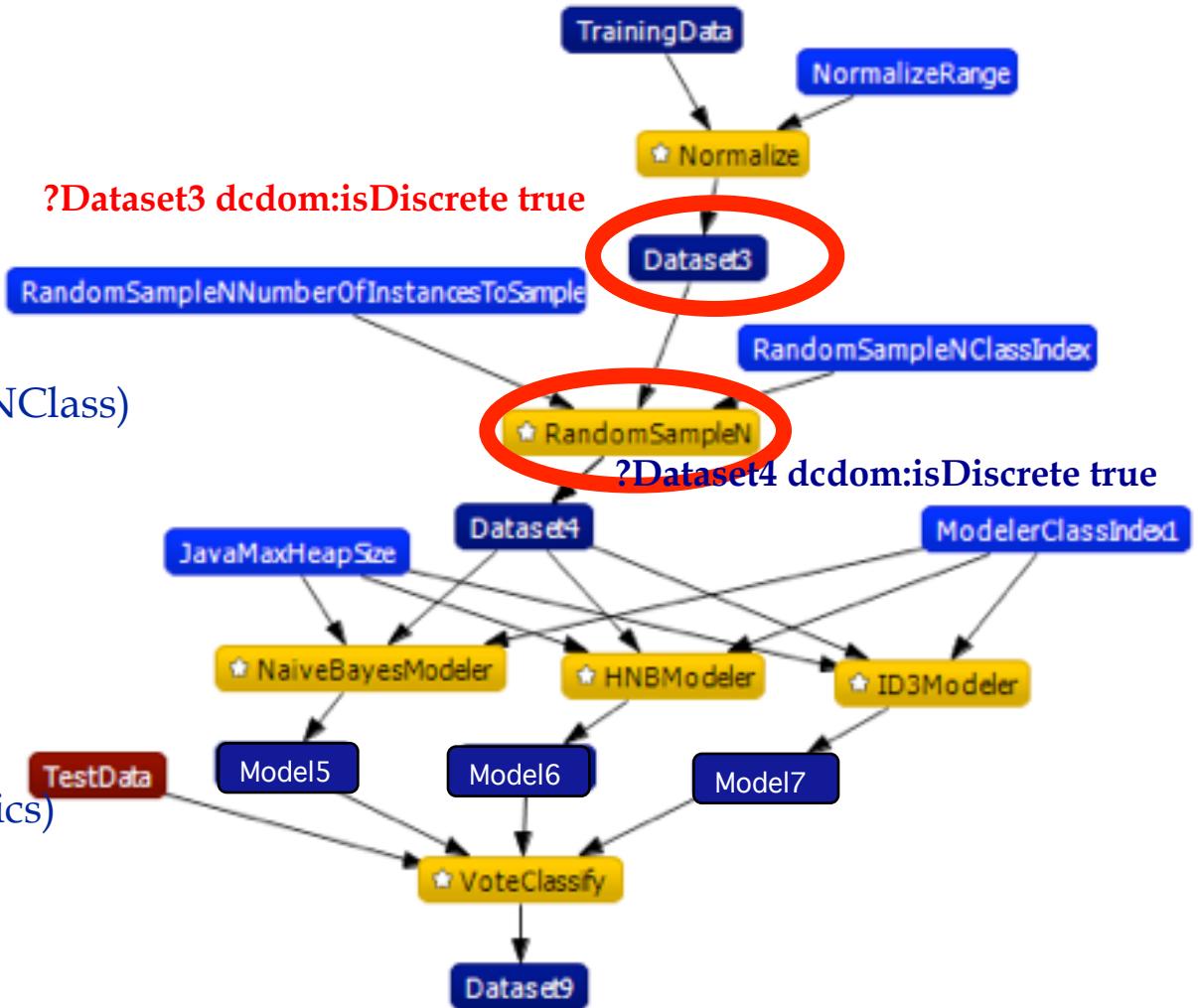
Example (Step 2 of 5)

Rule in Component Catalog:

[samplerTransfer:

```
(?c rdf:type pcdom:RandomSampleNClass)
(?c pc:hasOutput ?odv)
(?odv pc:hasArgumentID
"randomSampleNOutputData")
(?c pc:hasInput ?idv)
(?idv pc:hasArgumentID
"randomSampleNInputData")
(?odv ?p ?val)
(?p rdfs:subPropertyOf dc:hasMetrics)
```

```
-> (?idv ?p ?val)]
```

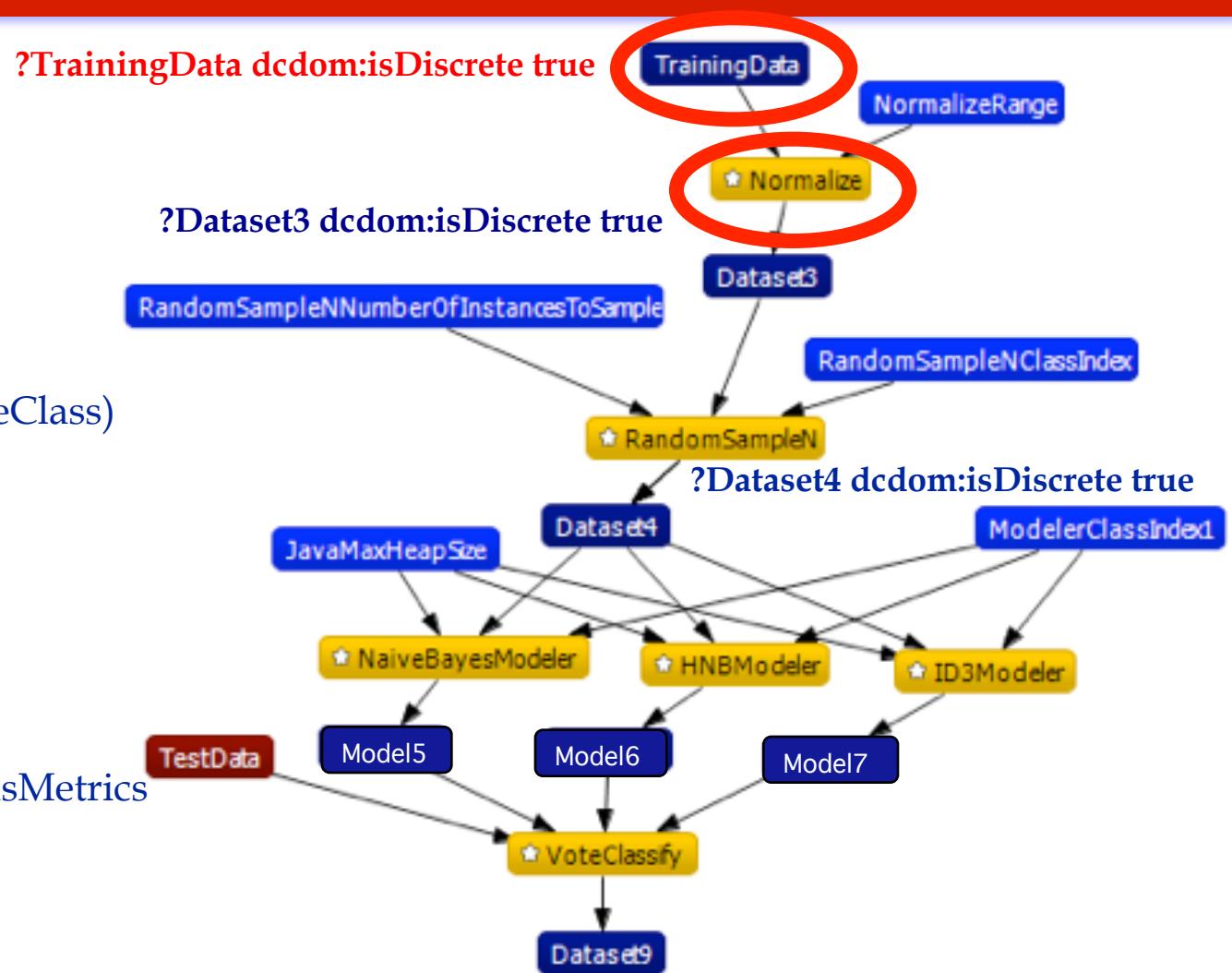


Example (Step 3 of 5)

Rule in Component Catalog:

[normalizerTransfer:

```
(?c rdf:type pcdom:NormalizeClass)
(?c pc:hasOutput ?odv)
(?odv pc:hasArgumentID
"normalizeOutputData")
(?c pc:hasInput ?idv)
(?idv pc:hasArgumentID
"normalizeInputData")
(?odv ?p ?val)
(?p rdfs:subPropertyOf dc:hasMetrics
-> (?idv ?p ?val)]
```



Example (Step 4 of 5)

Rule in Component Catalog:

[modelerTransferFwdData:

(?c rdf:type pcdom:ModelerClass)

(?c pc:hasOutput ?odv)

(?odv pc:hasArgumentID "outputModel")

(?c pc:hasInput ?idv)

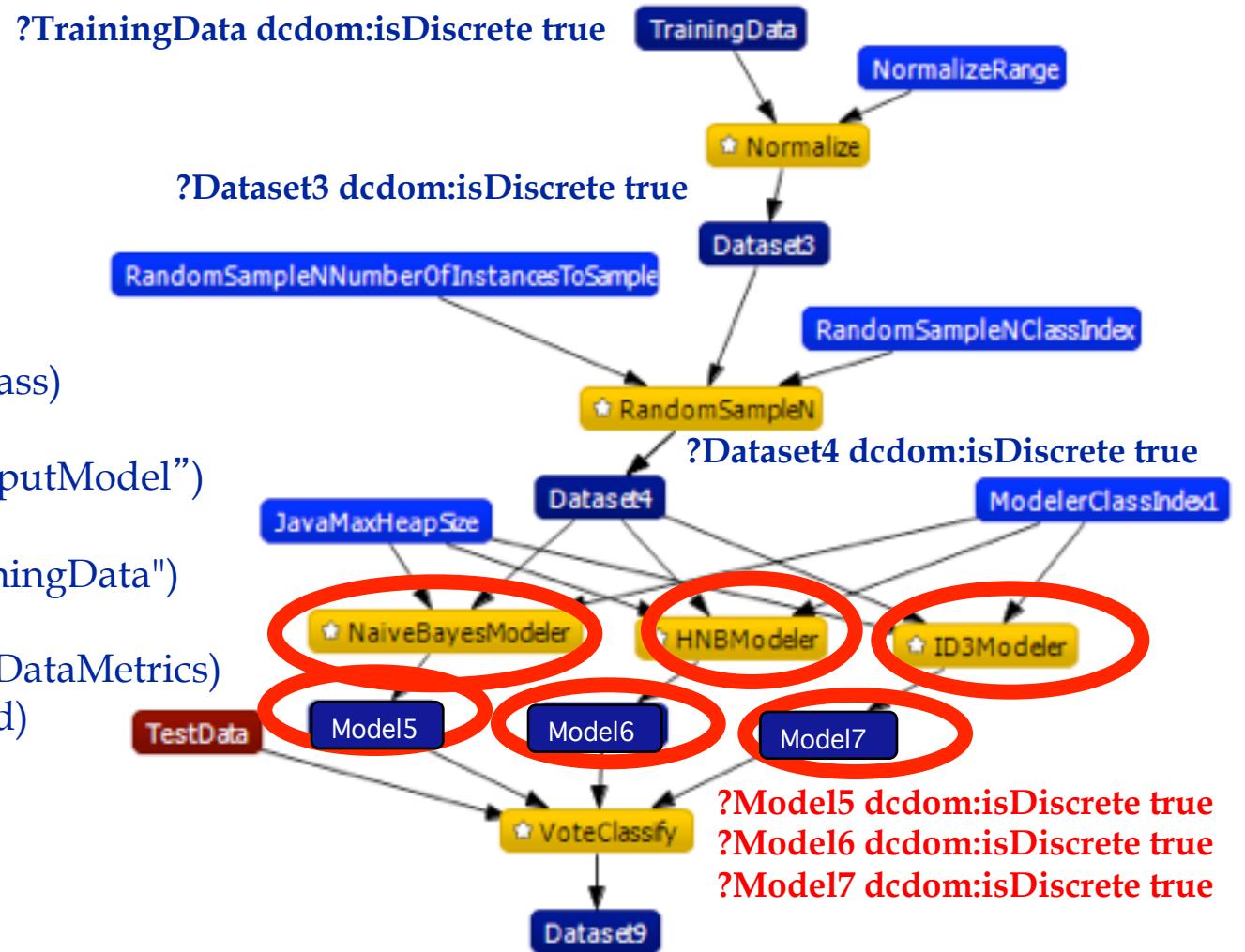
(?idv pc:hasArgumentID "trainingData")

(?idv ?p ?val)

(?p rdfs:subPropertyOf dc:hasDataMetrics)

notEqual(?p dcdm:isSampled)

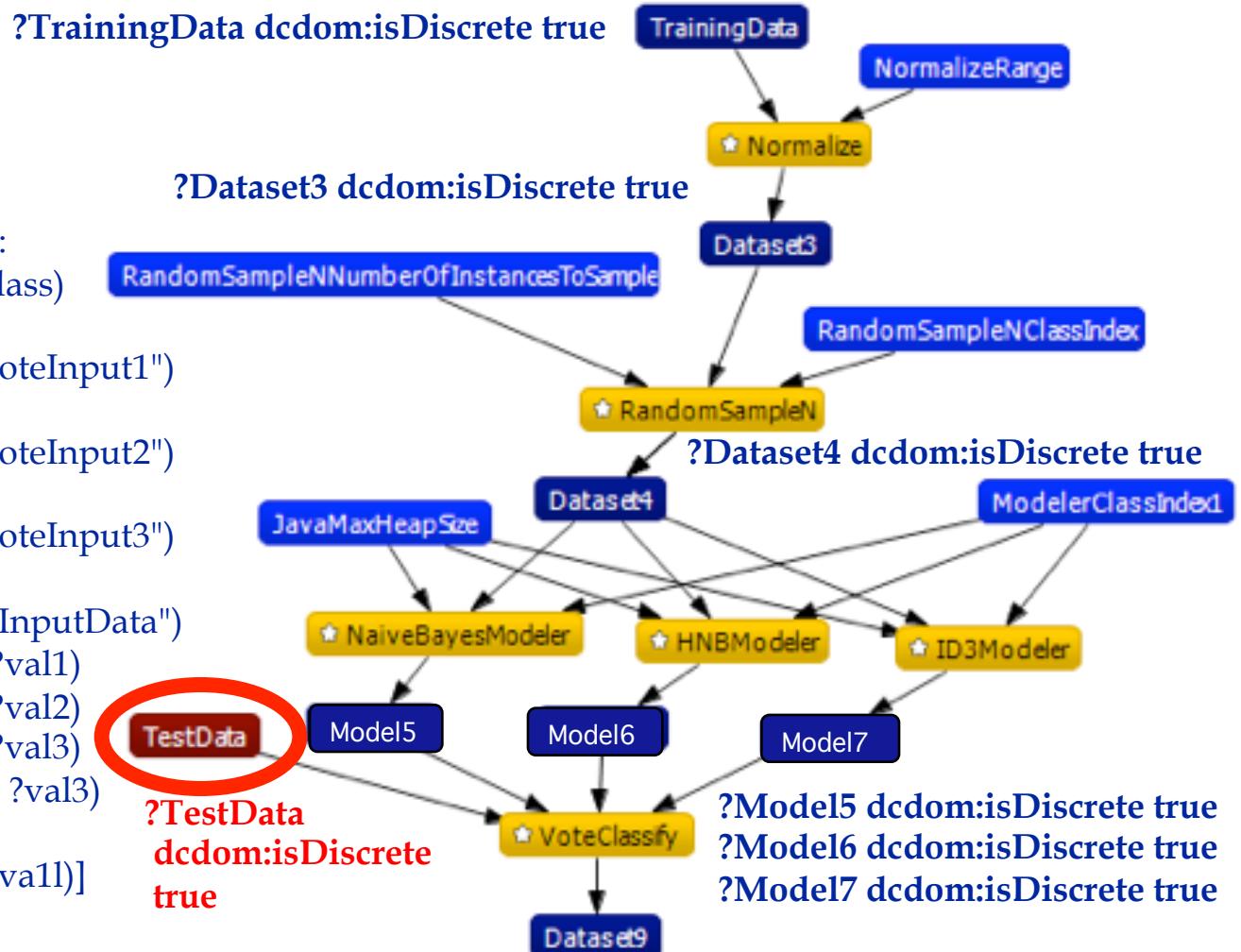
-> (?odv ?p ?val)]



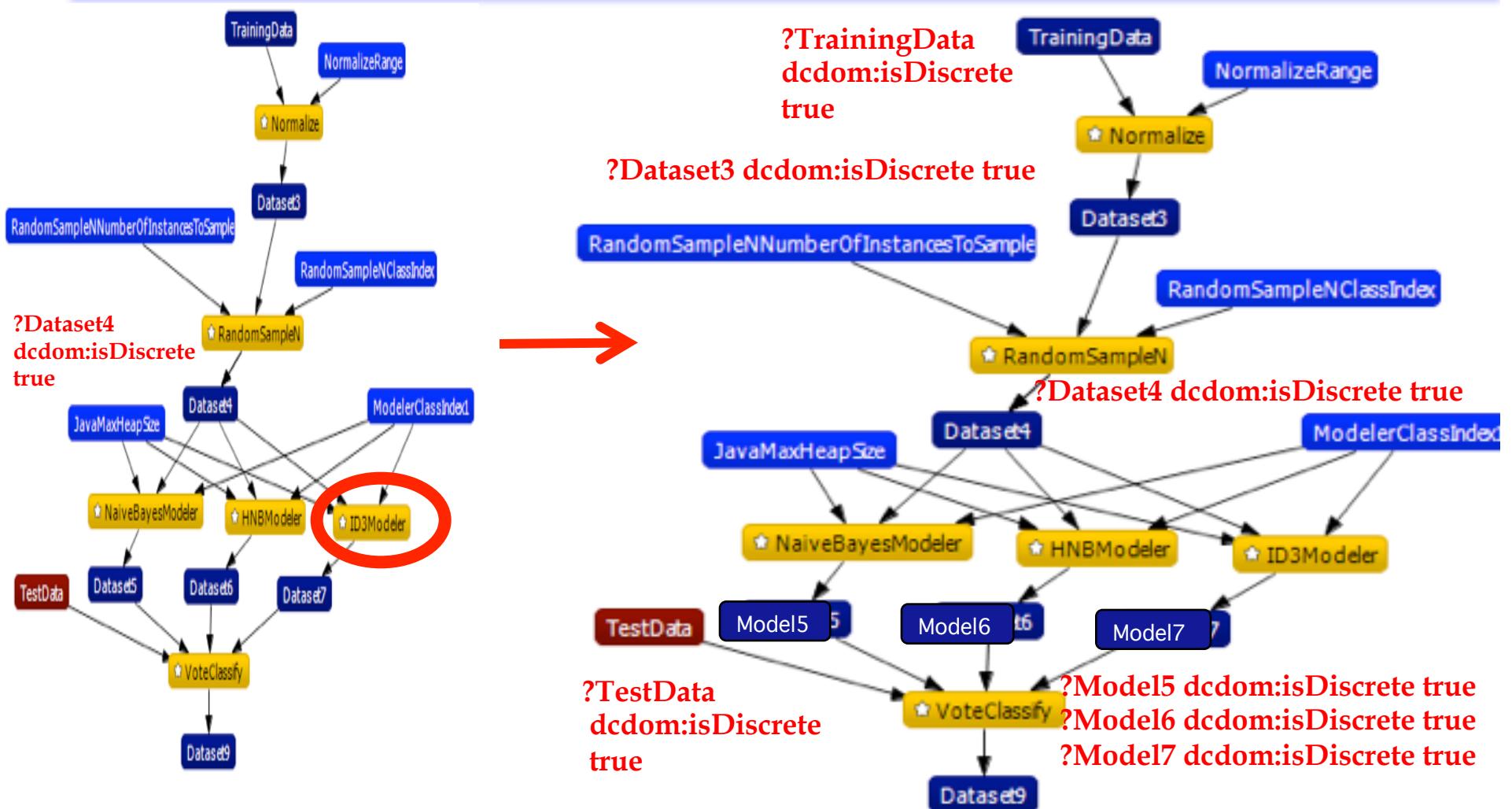
Example (Step 5 of 5)

Rule in Component Catalog:

```
[voteClassifierTransferDataFwd10:
 (?c rdf:type pcdom:VoteClassifierClass)
 (?c pc:hasInput ?idvmodel1)
 (?idvmodel1 pc:hasArgumentID "voteInput1")
 (?c pc:hasInput ?idvmodel2)
 (?idvmodel2 pc:hasArgumentID "voteInput2")
 (?c pc:hasInput ?idvmodel3)
 (?idvmodel3 pc:hasArgumentID "voteInput3")
 (?c pc:hasInput ?idvdata)
 (?idvdata pc:hasArgumentID "voteInputData")
  (?idvmodel1 dcdom:isDiscrete ?val1)
  (?idvmodel2 dcdom:isDiscrete ?val2)
  (?idvmodel3 dcdom:isDiscrete ?val3)
  equal(?val1, ?val2), equal(?val2, ?val3)
 -> (?idvdata dcdom:isDiscrete ?val1)]
```

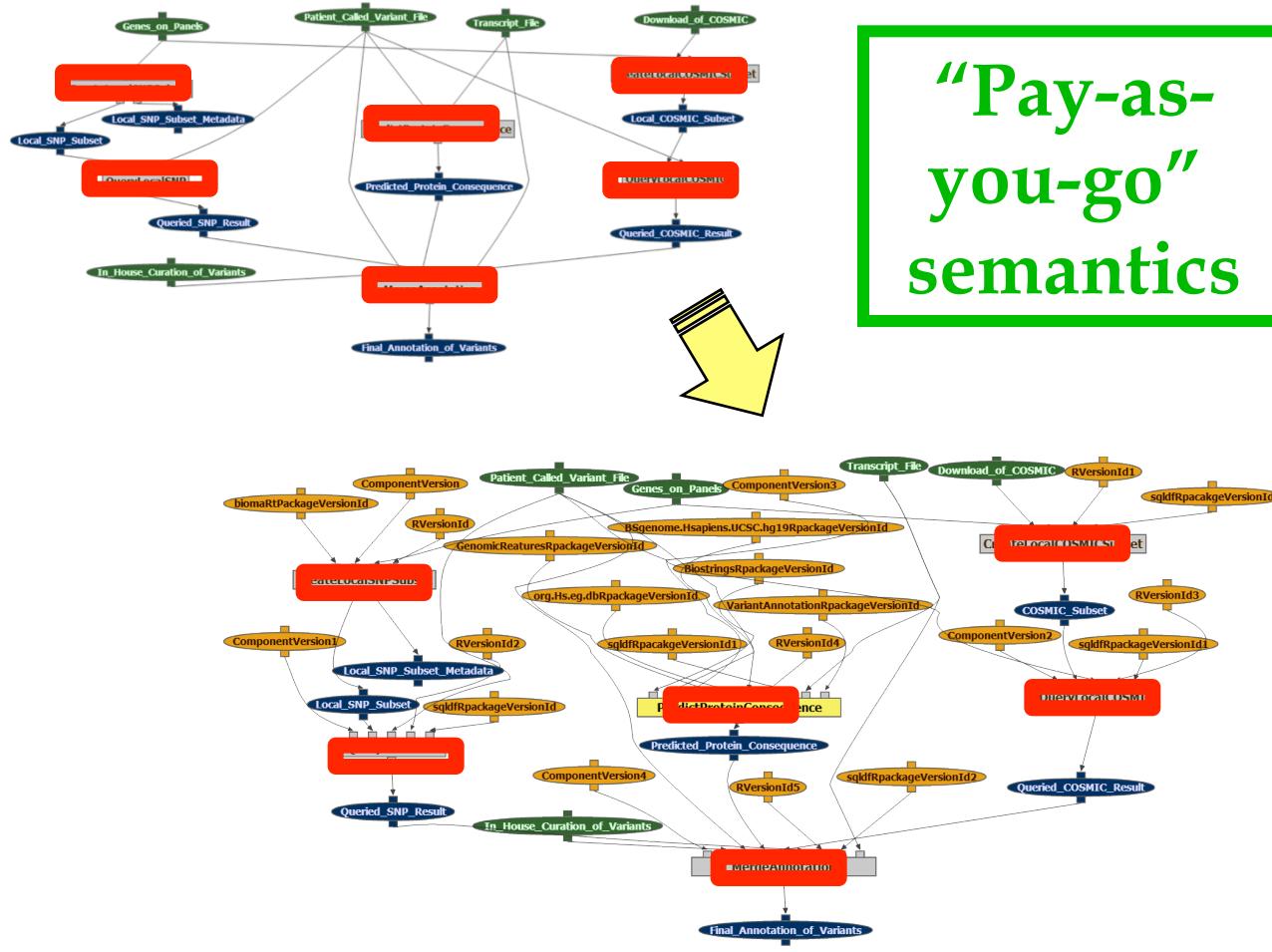


WINGS Workflow Reasoners: Result



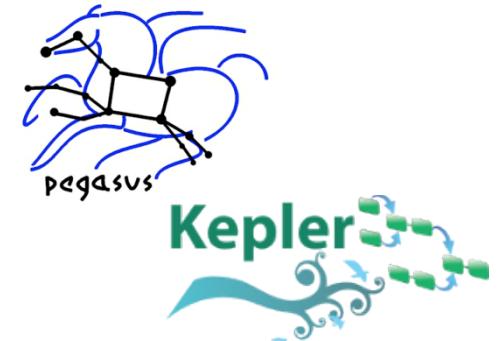
WINGS Automatic Workflow Generation Algorithm [Gil et al JETAI 2011]

*Work with P. Gonzalez (UCM) and Jihie Kim (ISI)
Workflows with S. McWeeney & C. Zhang (OHSU)*



Workflows

- Workflow systems
 - [Goble et al 2007]
 - [Ludaescher et al 2007]
 - [Freire et al 2008]
 - [Mattmann et al 2007]
 - [Mesirov et al 2009]
 - [Dinov et al 2009]
- Workflow representations
 - [Moreau et al 2010]
 - [IBM/MSR 2002]



Semantic Process Models

- Composition from first principles
 - [McIlraith & Son KR 2002] [Sohrabi et al ISWC 2006] [Sohrabi & McIlraith ISWC 2009] [Sohrabi & McIlraith ISWC 2010]
 - [McDermott AIPS 2002]
 - [Kuter et al ISWC 2004] [Sirin et al JWS 2005] [Kuter et al JWS 2005] [Lin et al ESWC 2008]
 - [Lecue ISWC 2009]
 - [Calvanese et al IEEE 2008]
 - [Bertolli et al ICAPS 2009]
 - [Li et al ISSC 2011]
- Representations
 - [Burstein et al ISWC 2002] [Martin et al ISWC 2007]
 - [Domingue & Fensel IEEE IS 2008] [Dietze et al IJWSR 2011] [Dietze et al ESWC 2009]
 - [Fensel et al 2011] [Vitvar et al ESWC 2008] [Roman et al AO 2005]

Semantic Descriptions of Software Components in Geosciences

Work with C. Duffy (PSU), S. Peckham (CU), C. Mattmann (JPL), J. Howison (UT)

The screenshot shows the Turbosoft Portal interface. On the left, a sidebar titled "Software" lists various software components and packages:

- 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

In the center, a "Describe Software" dialog is open for the "PIHM" component. It includes tabs for "I/O", "Assumptions", "Standard Names", "Summary", and "Comments". The "I/O" tab is selected, showing "Inputs" and "Outputs".

	Identifier	Type
<input type="checkbox"/>	ProjectName	ASCII
<input type="checkbox"/>	MeshFile	ASCII
<input type="checkbox"/>	AttFile	ASCII
<input type="checkbox"/>	GeoFile	ASCII
<input type="checkbox"/>	RivFile	ShapeFile
<input type="checkbox"/>	ForcFile	ASCII
<input type="checkbox"/>	IbcFile	ASCII
<input type="checkbox"/>	Init	ASCII
<input type="checkbox"/>	Calib	ASCII

CSDMS Standard Names [Peckham iEMSs 2014]

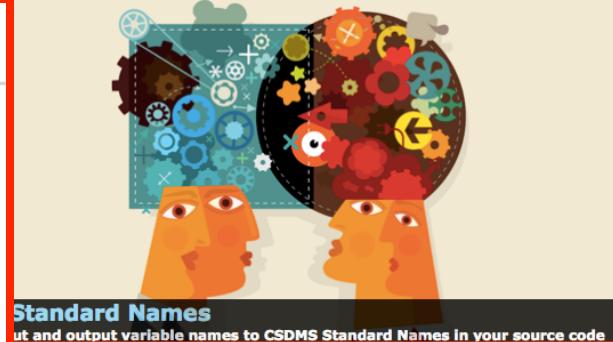
<http://csdms.colorado.edu/>



COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

Objects

acetic_acid
air
air_ball
air_carbon_dioxide
air_helium_plume
air_to_fuel
air_visible_light
air_water
air_water_vapor
airfoil
airfoil_enclosing_curve
airplane
airplane_wing
alaskan_black_bear_brain_to_body
alaskan_black_bear_head
aluminum
atmosphere
atmosphere_aerosol
atmosphere_carbon_dioxide
atmosphere_cloud
atmosphere_clox_as_chlorine



Standard Names
Put and output variable names to CSDMS Standard Names in your source code

Quantities

absorbed_shortwave_irradiation_flux
actual_bond_angle
affinity
age
air_dried_pressure_head
albedo
altitude
amplitude
area
area_fraction
aspect_angle
aspect_ratio
avogadro_constant
backscattered_shortwave_irradiation_f

Operators

anomaly
area_time_integral
azimuth_angle
curl
downward_component
downward_eastward_component
downward_northward_component
eastward_component
eastward_downward_component
gradient
increment
laplacian
lower_limit
magnitude
max
max_over_grid
max_over_grid_and_time
max_over_increment
max_over_time
mean_over_domain
mean_over_increment
min
min_over_grid
min_over_increment
normal_component



Home Software File Types Community

Software Describe Software PIHM

Add Rename Delete Make suggestions Save

CF CSDMS

Standard Names

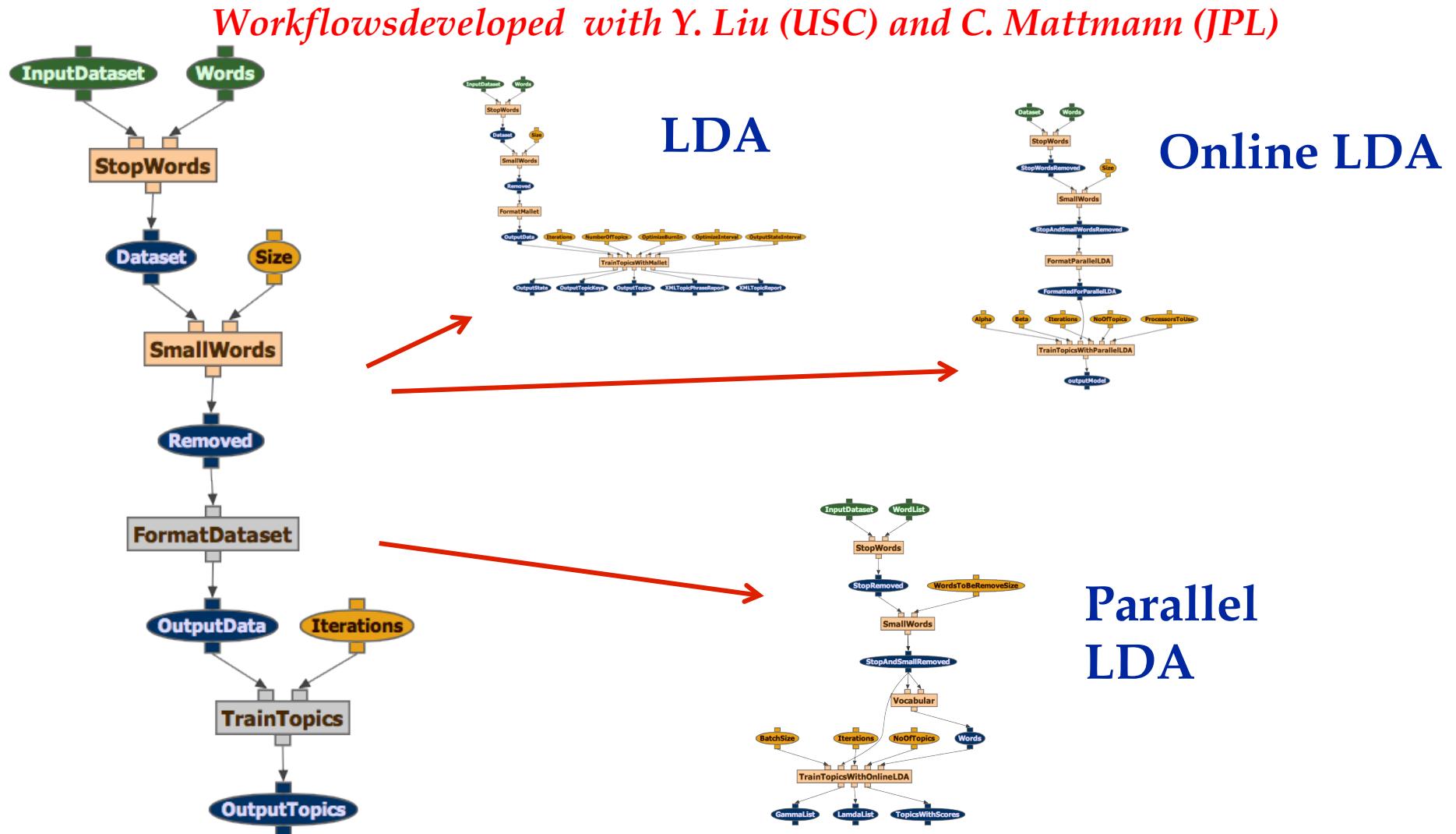
+ Add Standard Name - Delete

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	
	albedo	
	aspect_angle	
	backscattered_shortwave_irradiation_flux	
	brutsaert emissivity canopy_factor	
	brutsaert emissivity cloud_factor	
	...	

App Software Data Model Visualization SoftwarePackage ModelPackage PIHM TopoFlow VisualizationPackage

Benefits of Semantic Workflows:

1) Automatic Workflow Elaboration [Gil et al WORKS'13]



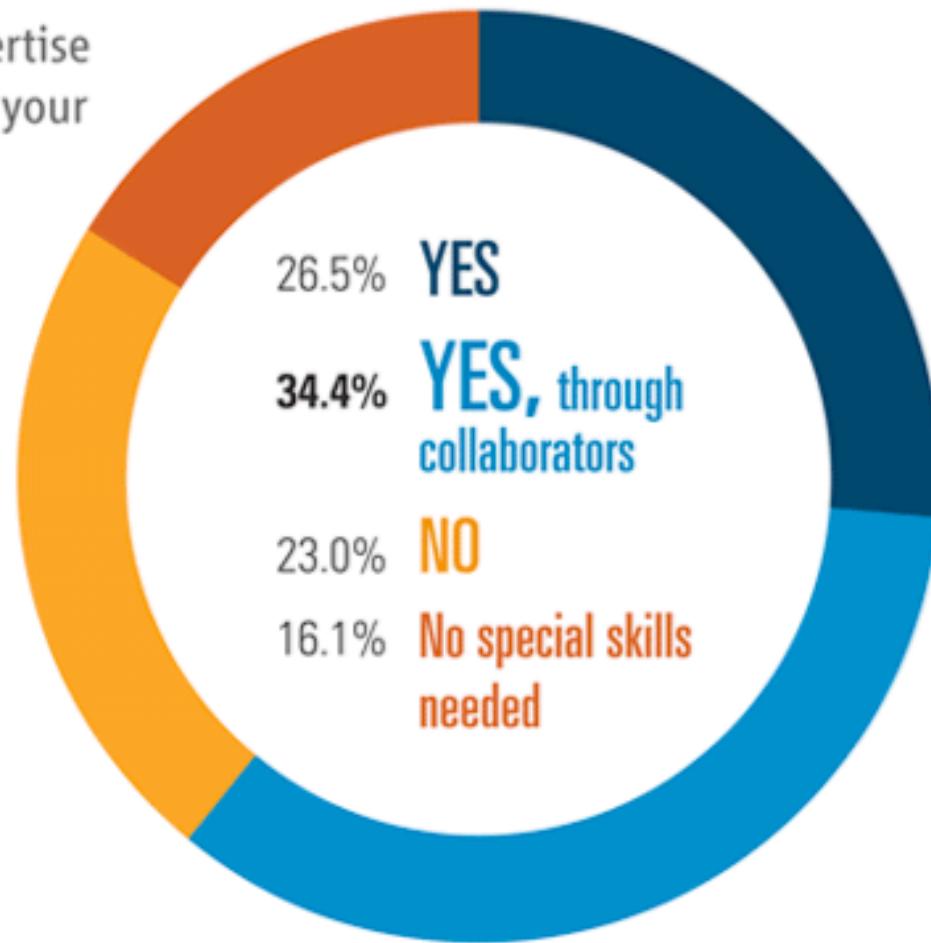
Benefits of Semantic Workflows:

2) Access to Data Analytics Expertise

Science, Dec 2011

Do you have the necessary expertise
in your lab or group to analyze your
data in the way you want?

“ The next few years
[particularly in medicine]
**the volume of data we
need to analyze will
expand exponentially.** ”

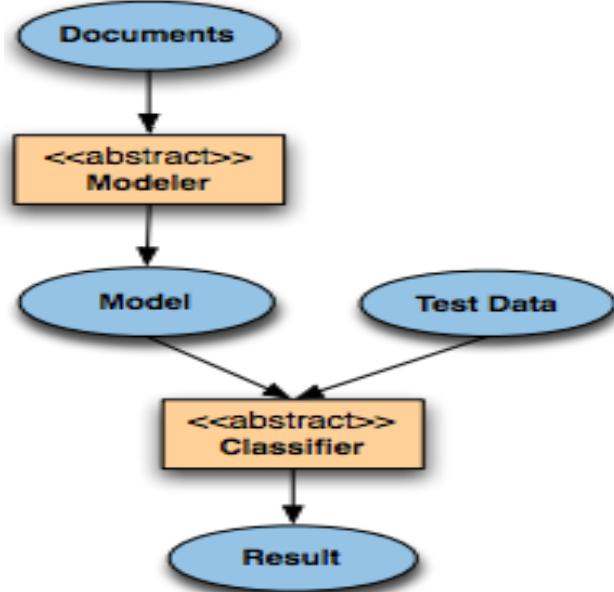


Capturing Expertise through Workflows

[Hauder et al e-Science 2011]

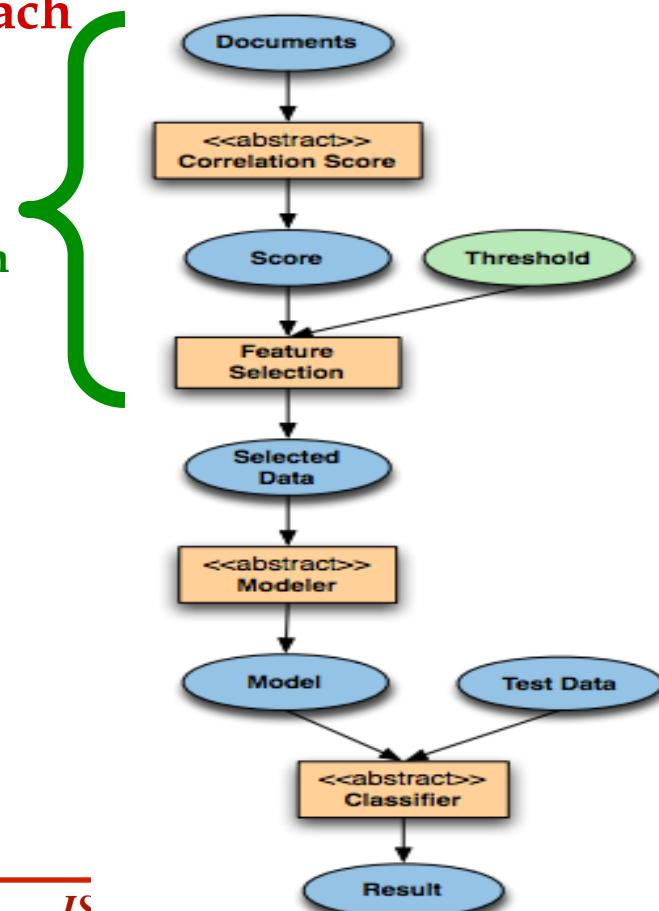
Workflows for text analytics, joint work with Yan Liu (USC) and Mattheus Hauder (TUM)

Naïve Approach



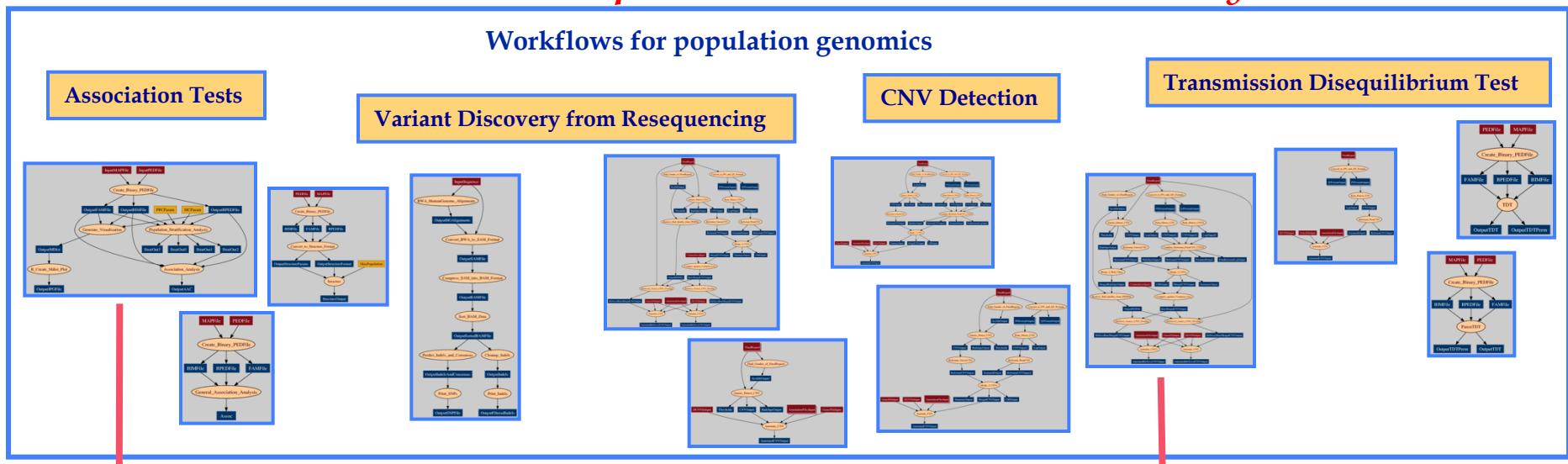
Expert Approach

Feature selection



Capturing Expertise [Gil et al 2012]

Work with Christopher Mason (Cornell University)

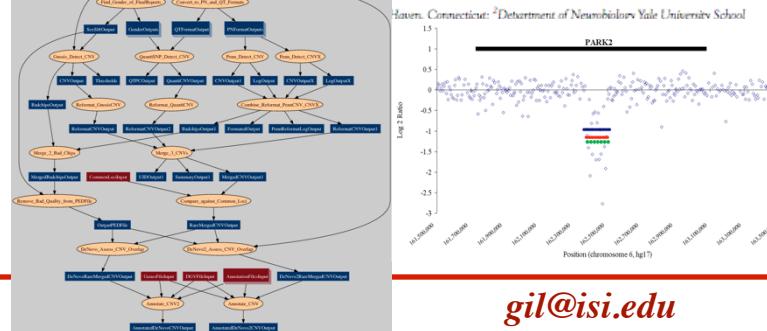
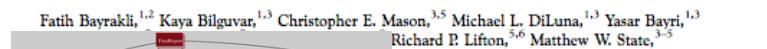
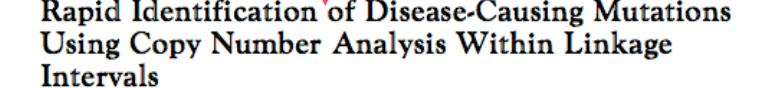
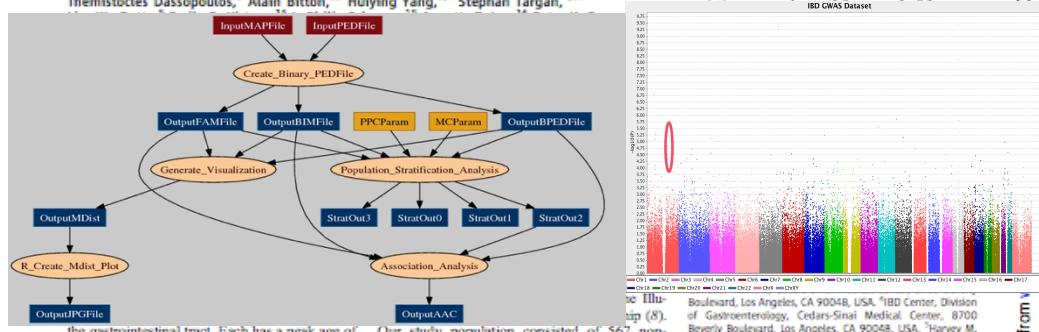


A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr,^{1,2} Kent D. Taylor,^{3,4} Steven R. Brant,^{5,6} John D. Rioux,^{7,8} Mark S. Silverberg,⁹ Mark J. Daly,^{1,10} A. Hillary Steinhardt,⁷ Clara Abraham,¹¹ Miguel Regueiro,^{1,2} Anne Griffiths,¹² Themistocles Dassopoulos,⁵ Alain Bitton,¹³ Huizing Yano,^{3,4} Stephan Tarsoian,^{4,5,6}

dicate. In addition to Arg381Gln, nine other markers in *IL23R* and in the intergenic region between *IL23R* and the adjacent IL-12 receptor, beta-2 gene (*IL12RB2*), had association P -values < 0.0001 in the non-Jewish, ideal CD case-control cohort (Table 1 and Table S1a).

We next tested for association of *IL23R* markers in an independent ileal CD case-control cohort, consisting of 401 patients and 433 controls, all of Jewish ancestry (8). Significant as-

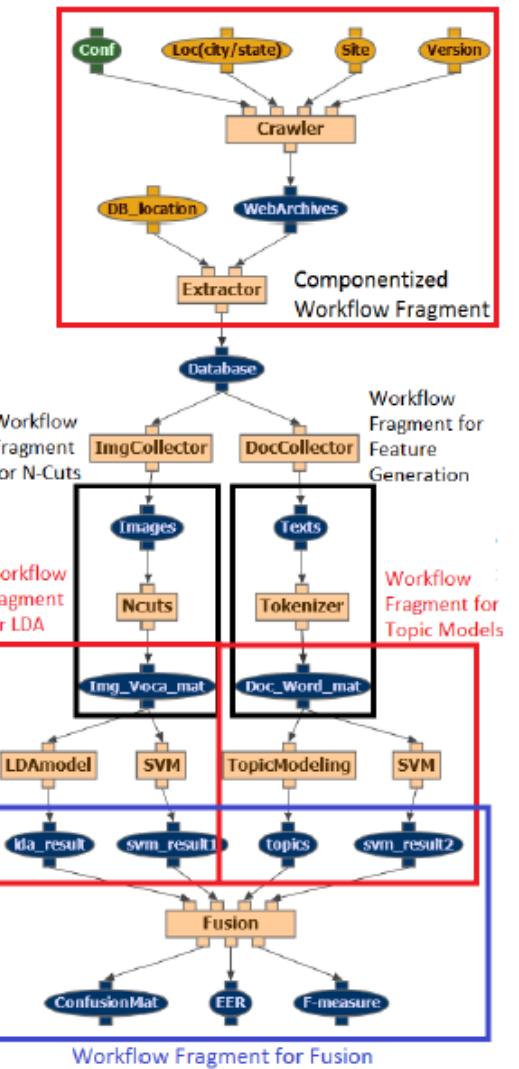
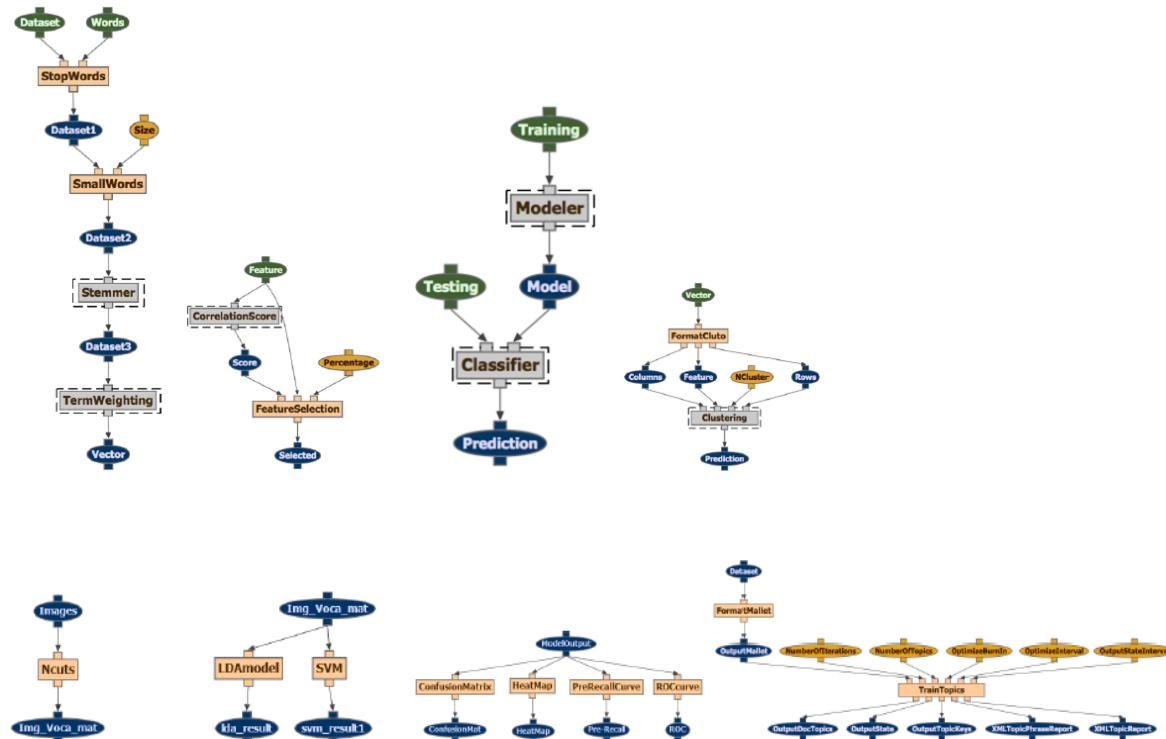


gil@isi.edu

Benefits of Semantic Workflows:

3) Saving Time Through Reuse [Sethi et al MM'13]

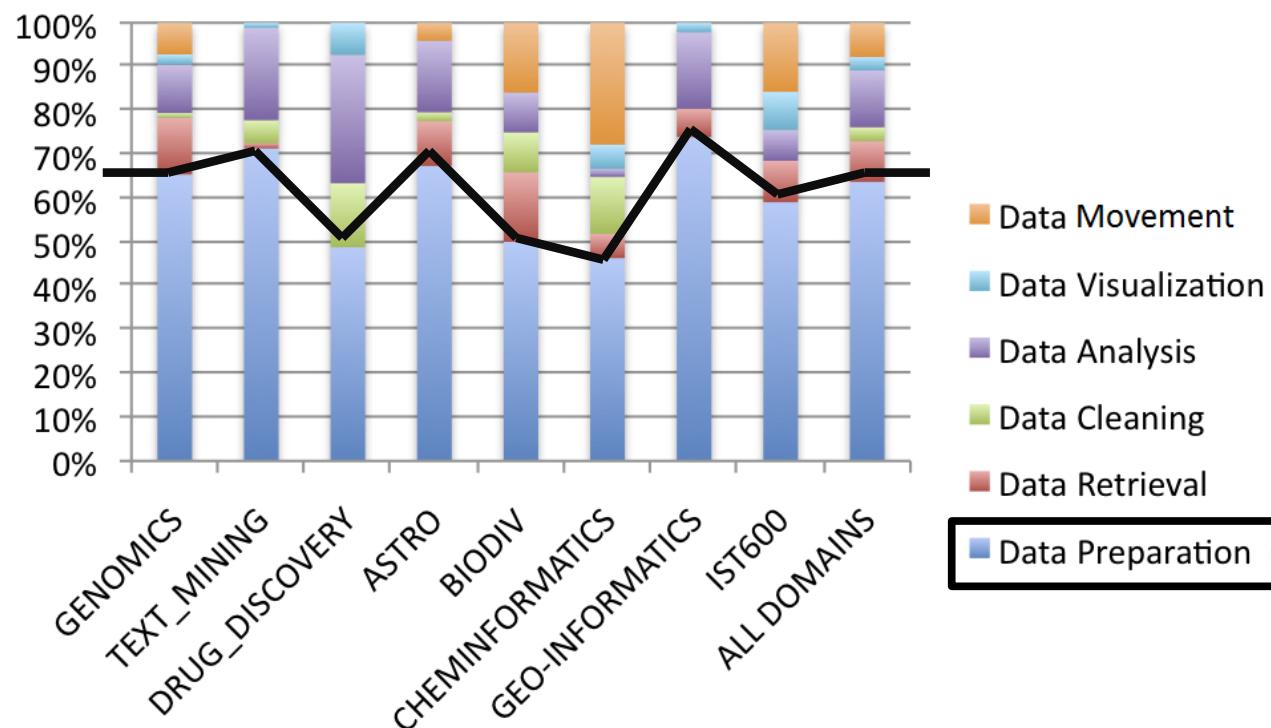
Work with Ricky Sethi and Hyujoon Jo of USC



Saving Time through Reuse [Garijo et al FGCS'13]

Work with D. Garijo and O. Corcho (UPM), P. Alper, K. Belhajjame, and C. Goble (UM)

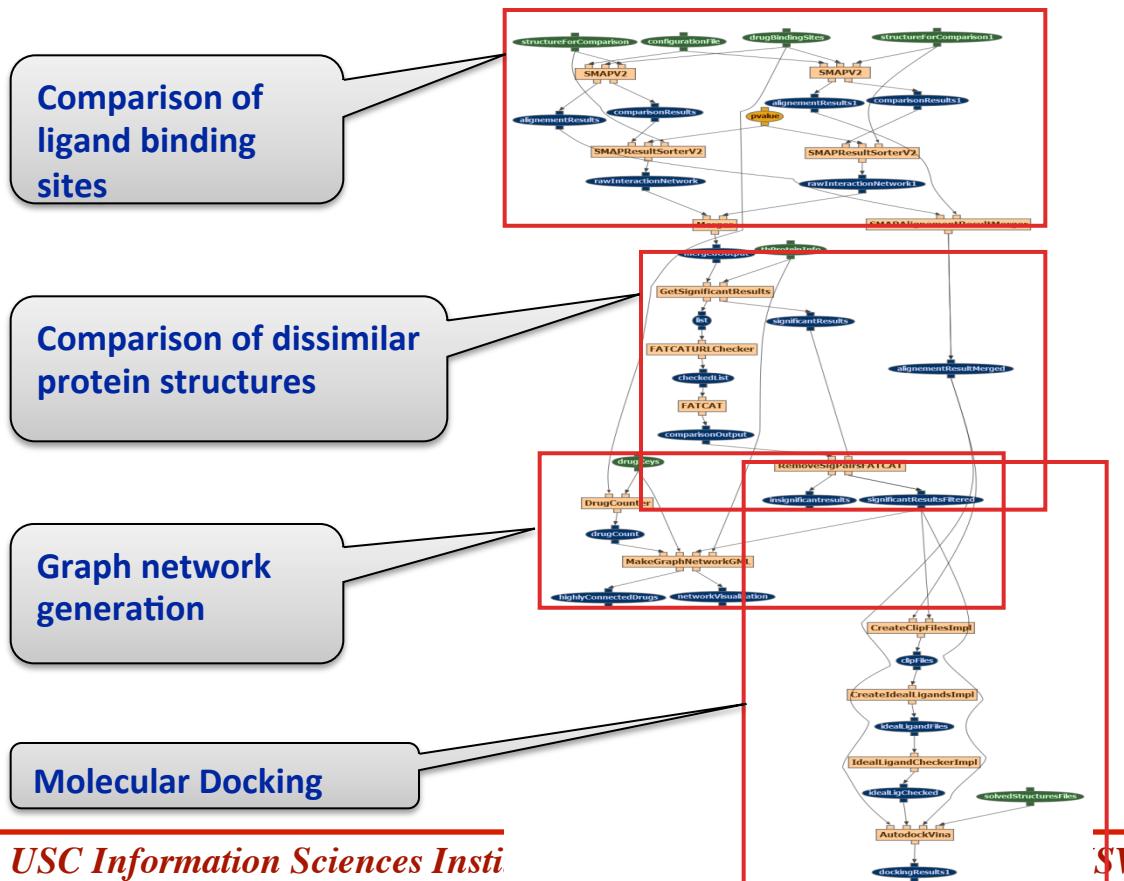
- “Scientists and engineers spend more than 60% of their time just preparing the data for model input or data-model comparison” (NASA A40)



Measuring Time Savings with “Reproducibility Maps” [Garijo et al PLOS CB12]

Work with D. Garijo of UPM and P. Bourne of UCSD

- 2 months of effort in reproducing published method (in PLoS' 10)
 - Authors expertise was required



Comparison of Ligand Binding Sites

SMAP1	SMAP2	SMAP Result Sorter1	SMAP Result Sorter2	Merger	Align Result Merger	Minimal
SMAP1	SMAP2	SMAP Result Sorter1	SMAP Result Sorter2	Merger	Align Result Merger	Novice
						Author

Comparison of dissimilar protein structures:

GetSignificant Results	FATCAT URLChecker	FATCAT	Remove Significant Pairs	Minimal
GetSignificant Results	FATCAT URLChecker	FATCAT	Remove Significant Pairs	Novice
GetSignificant Results	FATCAT URLChecker	FATCAT	Remove Significant Pairs	Author

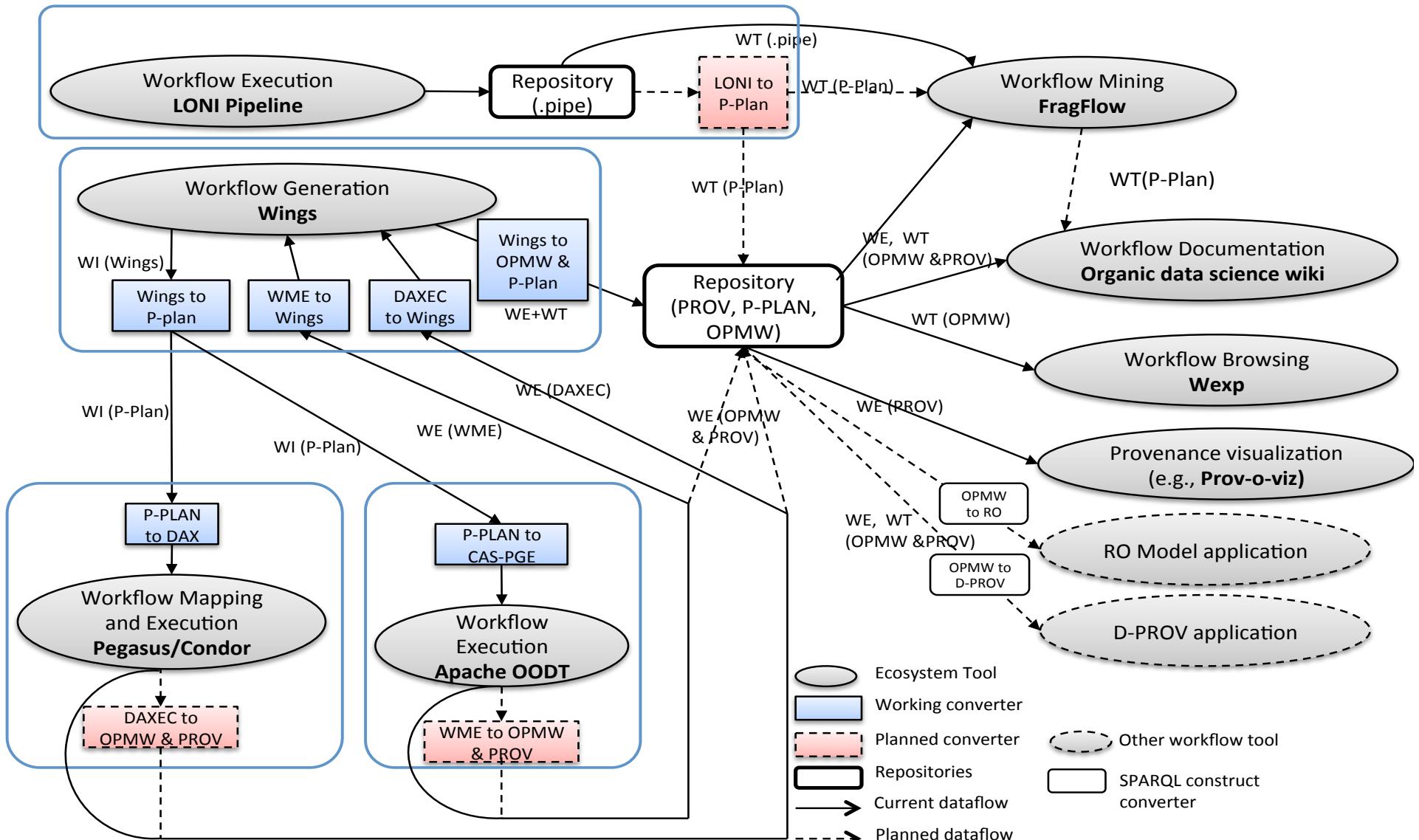
Docking

CreateClip Files	Createideal Ligands	IdealLigand Checker	Autodock Vina	Minimal Novice
CreateClip Files	Createideal Ligands	IdealLigand Checker	Autodock Vina	Author

Benefits of Semantic Workflows:

4) Interoperability in a Workflow Ecosystem [Garijo et al 2014]

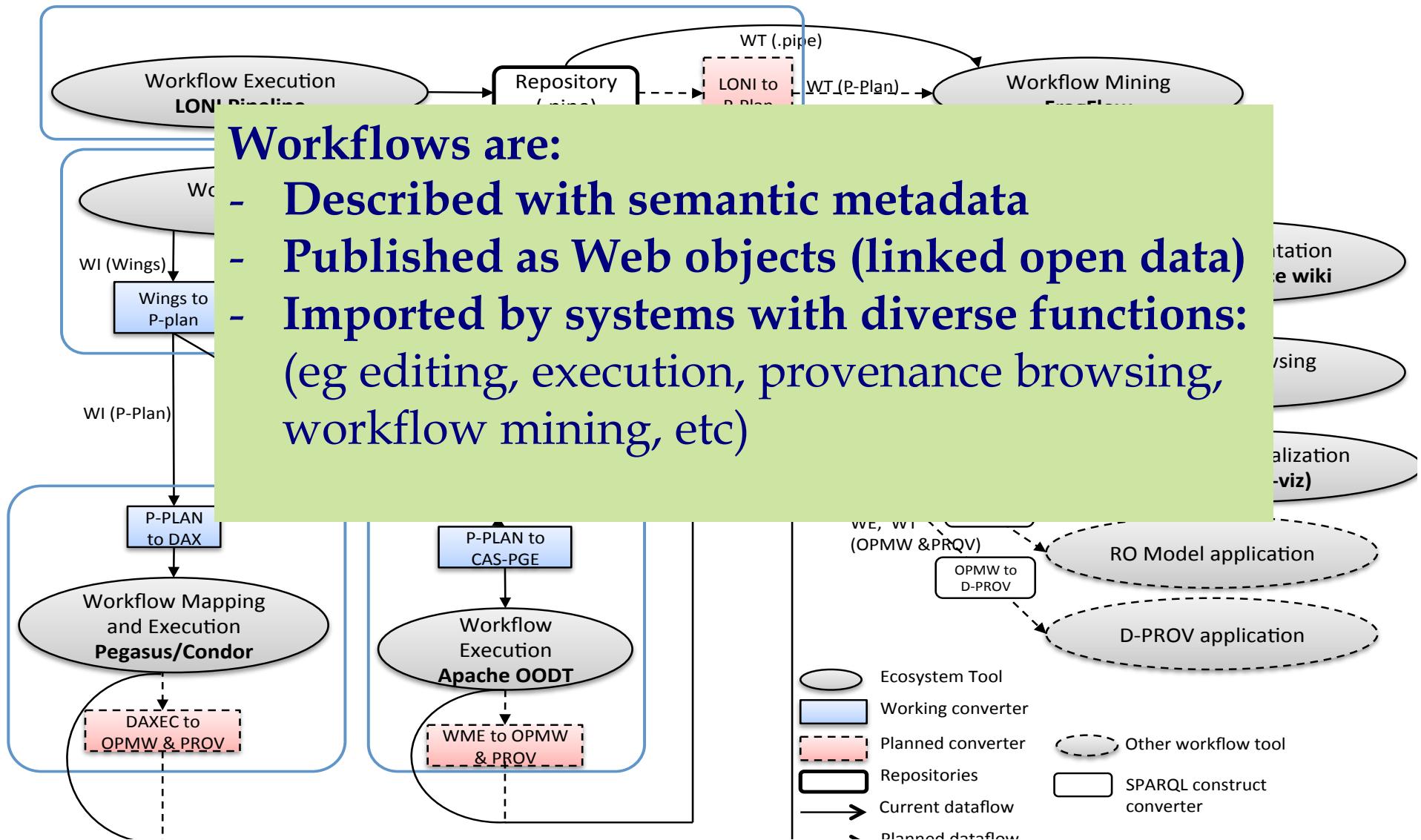
Work with D. Garijo and O. Corcho of UPM



Benefits of Semantic Workflows:

4) Interoperability in a Workflow Ecosystem [Garijo et al 2014]

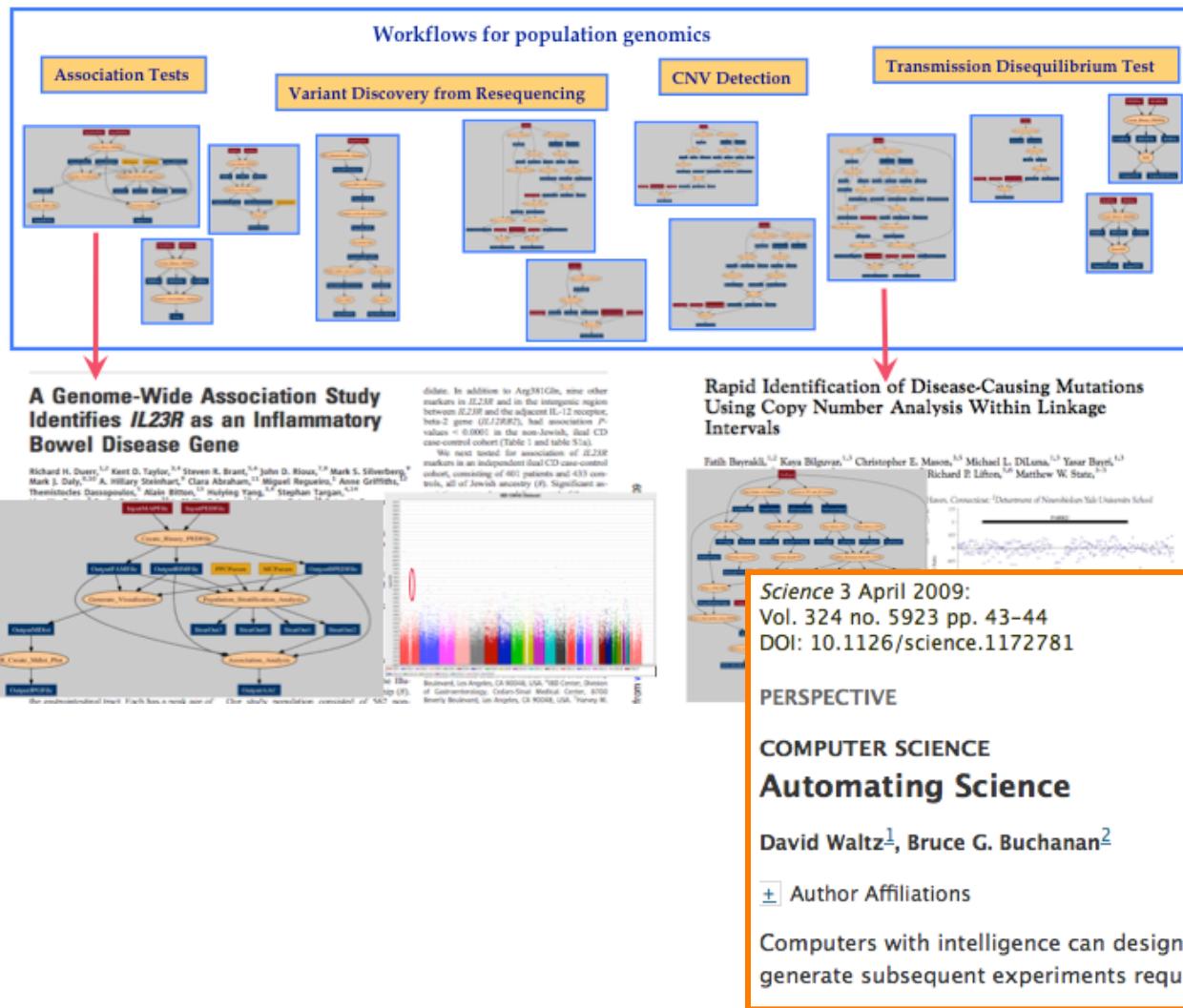
Work with D. Garijo and O. Corcho of UPM



Some Readings

- Yolanda Gil: “Intelligent Workflow Systems and Provenance-Aware Software.” Proceedings of the Seventh International Congress on Environmental Modeling and Software (iEMSS), San Diego, CA, 2014.
- Yolanda Gil: “From Data to Knowledge to Discoveries: Artificial Intelligence and Scientific Workflows.” *Scientific Programming* 17(3), 2009.
- Ewa Deelman, Chris Duffy, Yolanda Gil, Suresh Marru, Marlon Pierce, and Gerry Wiener: “EarthCube Report on a Workflows Roadmap for the Geosciences.” National Science Foundation, Arlington, VA. 2012.

A Semantic Challenge: Automatic Paper Generator



Capture knowledge about analytic methods

- Run workflows in existing data repositories
 - Report new findings

Science

A Semantic Challenge: A Web of Semantic Workflows/Processes

Assist people to:

- Share
- Copy
- Reuse
- Adapt
- Remix
- Update
- Certify
- Review
- ...

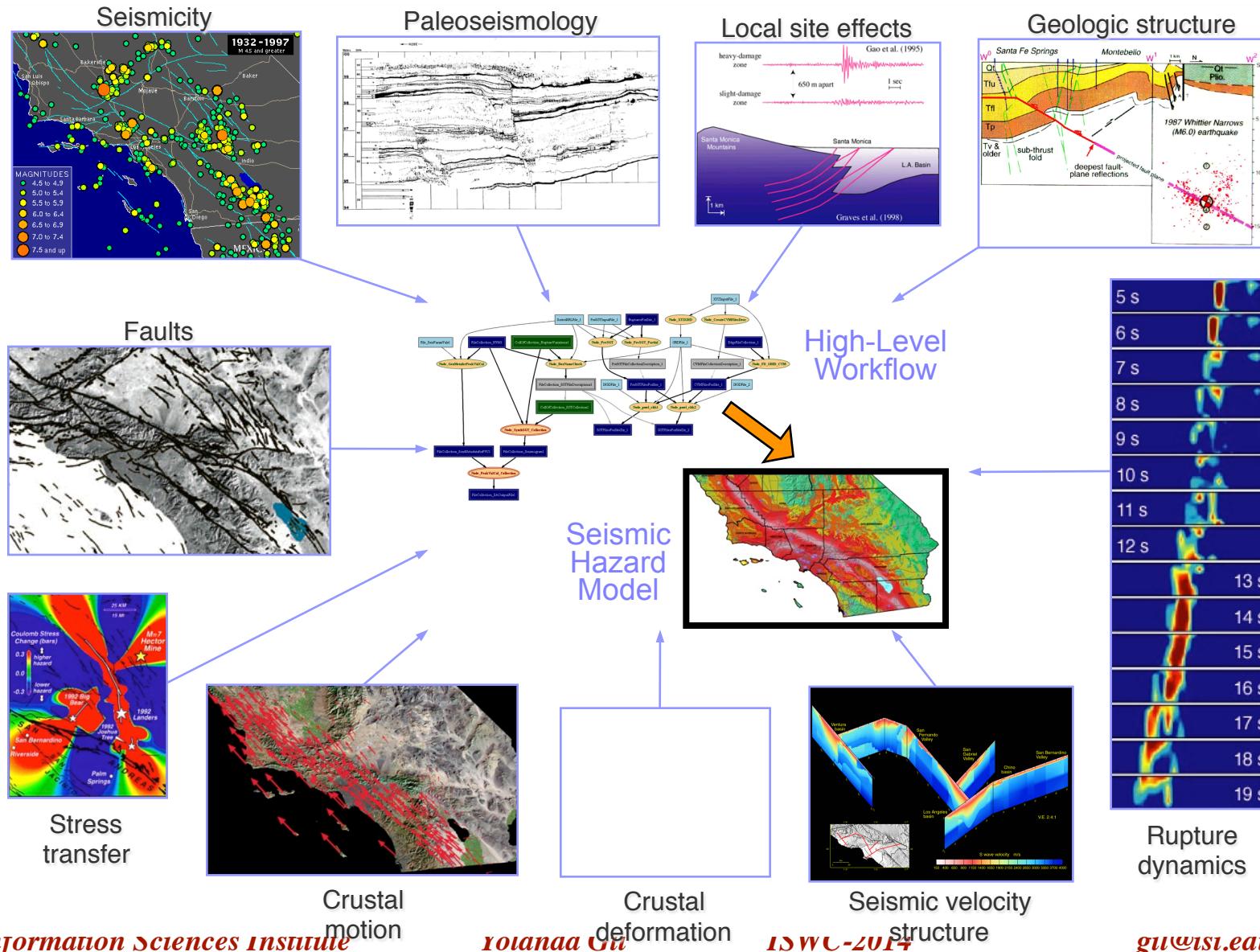
“Pay-as-you-go”
semantics

Semantic Challenges in Getting Work Done

- To dos
 - Managing personal to dos
 - Managing coordinated to dos
- Knowledge rich tasks in science
 - Automatic paper generator
 - A Web of semantic workflows/processes
- Open science

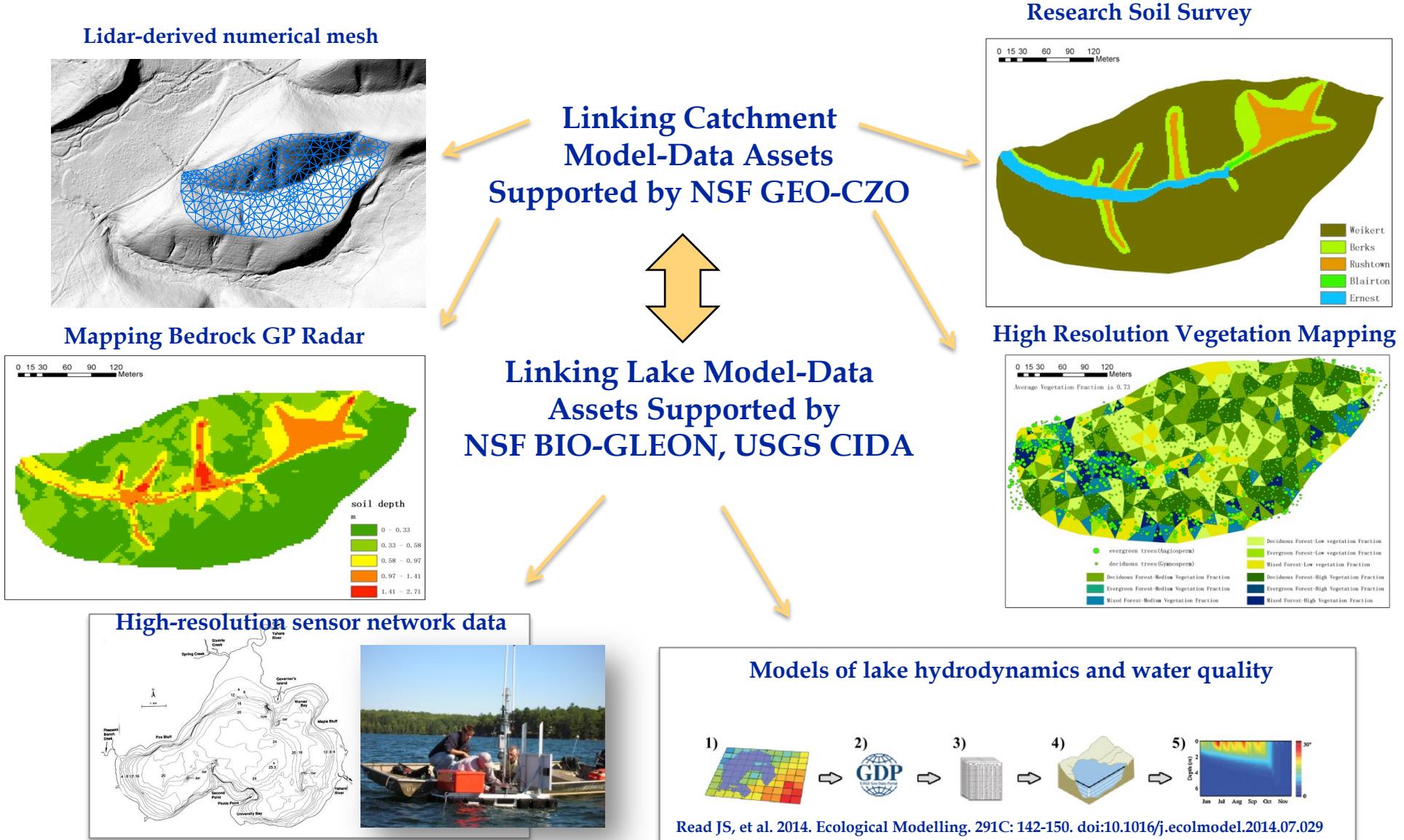
Collaboration to Develop Workflows [Gil et al 2007]

Slide from T. Jordan of USC and SCEC



Understanding the “Age of Water”

Work with P. Hanson (UWisc), C. Duffy (PSU), and J. Read (USGS)



A New Kind of Collaborative Platform

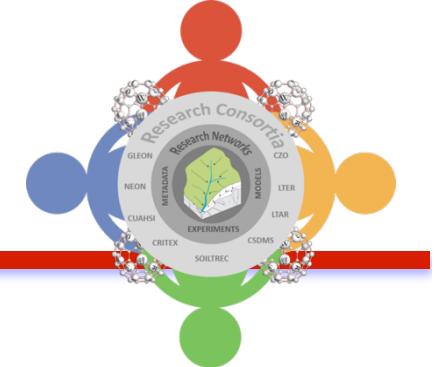
■ Taxonomy of Science Communities [Bos et al 2007]

Shared Instruments	NEON
Community Data Systems	PDB
Open Community Contribution Systems	Zooniverse
Virtual Communities of Practice	GLEON
Virtual Learning Communities	VIVO
Distributed Research Centers	ENCODE
Community Infrastructure Projects	CSDMS

- ## ■ Need a platform to support science collaborations that require:
- Significant organization and coordination
 - Maintaining a community over the longer term
 - Growing the community based on unanticipated needs

Organic Data Science

Work with F. Michel and M. Hauder of TUM



- **Organic data science** is a novel approach to on-line scientific collaboration that supports:
 - **Self-organization of communities** by enabling any user to specify and decompose tasks
 - **On-line community support** by incorporating social sciences principles and best practices
 - **An open science process** by capturing new kinds of metadata about the collaboration that give necessary context to newcomers

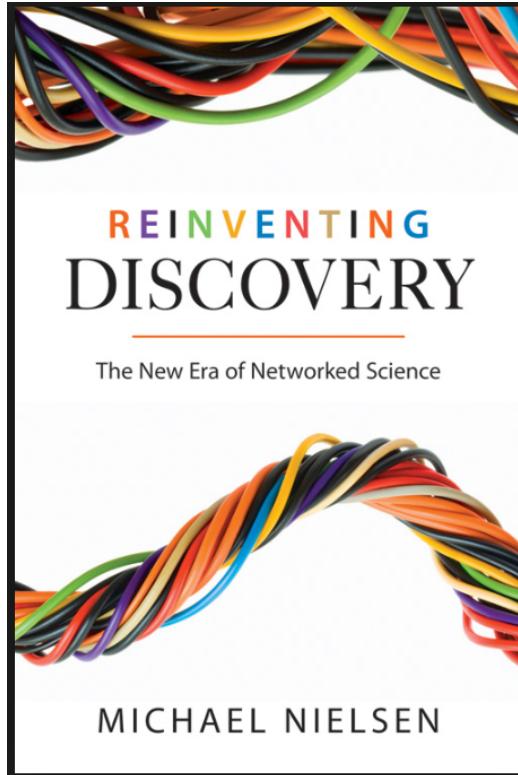
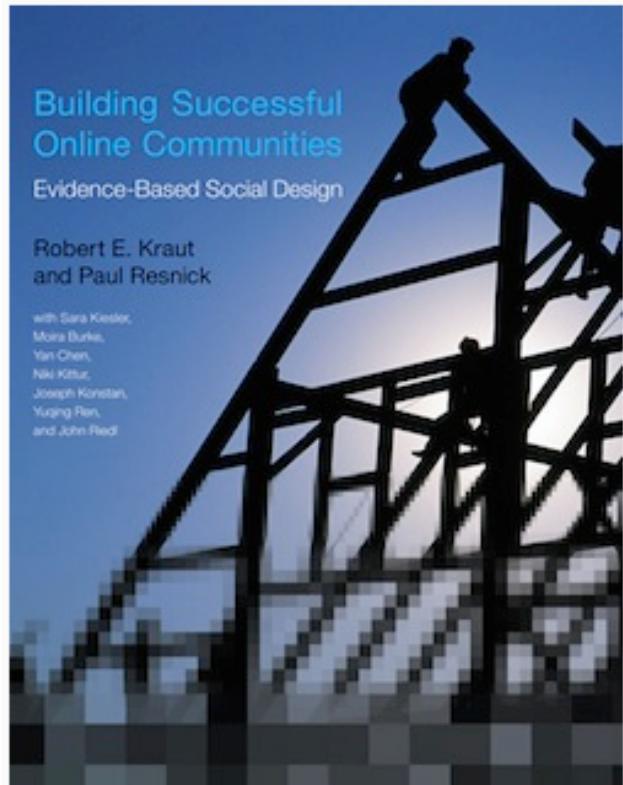
Self-Organization through Task Decomposition

o Implement the lake model for North Temperate Lakes

- Document the GLM Lake Model
- Set up the GLM lake model for North Temperate Lakes
 - Gather bathymetric and elevation data for North Temperate Lakes
 - Obtain meteorological data for North Temperate Lakes
 - Obtain initial conditions data for North Temperate lakes
 - Obtain initial water level
 - Obtain initial water temperature data
 - Obtain initial nutrient amount
 - Obtain initial phytoplankton concentration
 - Gather stream flow data for each major inflow
 - Collect daily volumetric data for each major inflow
 - Obtain water temperature data for each major inflow
 - Obtain or estimate nutrient and carbon concentration data for each major inflow

- Many tasks involved
- Necessary data resides in different repositories
- Different people understand different kinds of data
 - Where it is
 - How to use it
- If other data needed, unclear who has it

Social Principles for Online Communities



WIKIPEDIA
The Free Encyclopedia



The polymath blog

January 20, 2014

Two polymath (of a sort) proposed projects
Main Page | discussion

Main Page

This is the wiki for [polymath](#).
Gowers' blog post [Is it](#)
Many polymath projects
A Polymath logo is being developed.
The wiki is currently k

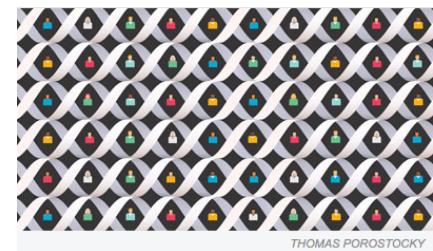
I. A polymath proposal: Convex varieties.

One of the interesting questions regarding the polymath endeavor is how to use it to develop a theory/new area.

My idea is to have a few open-ended projects ("hubs") that are simply a finite set of points in a well-developed area of mathematics that has not been studied much. I suspect that for such a project the first discussions will be very slow, but then things will pick up.

In general (but perhaps more so for an open-ended project), it is scale than existing ones but perhaps less intensive, and that perhaps more fun.

II. A polymath-of-a-sort project: Riemann Hypothesis



Social Principles: Some Examples

- **Starting communities**, e.g.:
 - Organize content, people, and activities into subspaces
 - Inactive tasks should have “expected active times”
- **Encouraging contributions**, e.g.:
 - Simple tasks with challenging goals are easier to comply with
 - Publicize that others have complied with requests
- **Encouraging commitment**, e.g.:
 - Interdependent tasks increase commitments and reduce conflict
- **Dealing with newcomers**, e.g.:
 - Design common learning experiences for newcomers
 - Provide sandboxes while they are learning

Opening Science: Polymath [Nielsen, Gowers 09]

PolyMath

navigation

- [Main Page](#)
- [Community portal](#)
- [Current events](#)
- [Recent changes](#)
- [Random page](#)
- [Help](#)

search

toolbox

- [What links here](#)
- [Related changes](#)
- [Special pages](#)
- [Printable version](#)
- [Permanent link](#)

Main Page

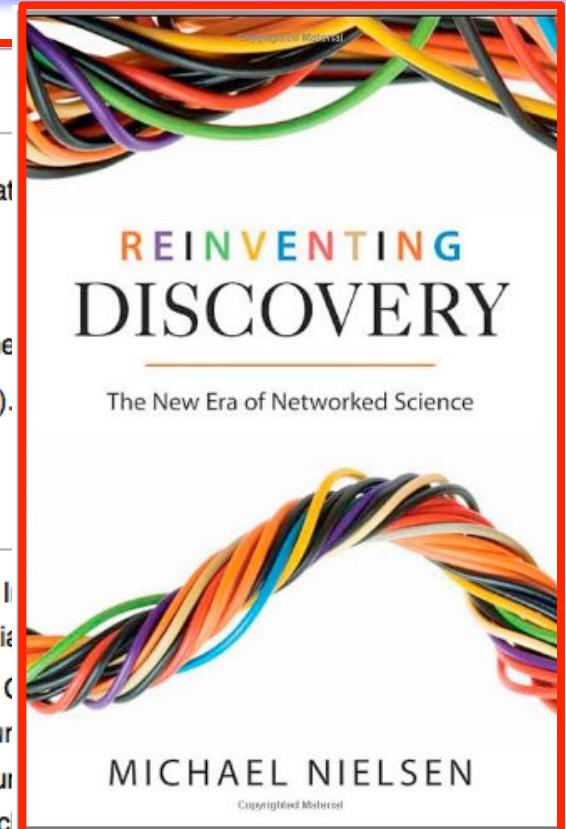
This is the wiki for *polymath* projects - massively collaborative online mathematics. Is massively collaborative mathematics possible? [↗](#)

Many polymath projects will be proposed, planned, and run at [This Blog ↗](#). A Polymath logo is being trialled. If you have more suggestions, please add them.

The wiki is currently locked down due to a major influx of spam (July 29, 2013). I will do my best to reply quickly.

Existing polymath projects

- [Polymath1](#): New proofs and bounds for the density Hales-Jewett theorem. Initiated July 12, 2009; launched Aug 12, 2009; solved Aug 18, 2009.
- [Polymath2](#): Must an "explicitly defined" Banach space contain c_0 or l_p ? Initiated Sept 12, 2009; launched Oct 1, 2009; still open.
- [Mini-polymath1](#): Solving Problem 6 of the 2009 International Mathematical Olympiad. Proposed July 17, 2009; launched July 20, 2009; solved July 24, 2009.
- [Polymath3](#): The polynomial Hirsch conjecture. Proposed July 17, 2009; launched July 20, 2009; solved July 24, 2009.
- [Polymath4](#): A deterministic way to find primes. Proposed July 27, 2009; launched July 28, 2009; solved Aug 12, 2009.
- [Polymath5](#): The Erdős discrepancy problem. Proposed Jan 10, 2010; launched Jan 12, 2010; solved Feb 12, 2012.
- [Mini-polymath2](#): Solving Problem 5 the 2010 International Mathematical Olympiad. Proposed July 17, 2010; launched July 20, 2010; solved July 24, 2010.
- [Polymath6](#): Improving the bounds for Roth's theorem. Proposed Feb 5, 2011.
- [Mini-polymath3](#): Solving a problem from the 2011 International Mathematical Olympiad. Proposed June 10, 2011; launched June 12, 2011; solved June 16, 2011.
- [Mini-polymath4](#): Solving a problem from the 2012 International Mathematical Olympiad. Proposed June 10, 2012; launched June 12, 2012; solved June 16, 2012.
- [Polymath7](#): Establishing the Hot Spots conjecture for acute-angle triangles. Proposed Dec 10, 2012; launched Dec 12, 2012; solved Dec 16, 2012.
- [Polymath8](#): Improving the bounds for small gaps between primes. Proposed Jan 12, 2013; launched Jan 14, 2013; solved Feb 12, 2014.



Organic Data Science



=> Task-oriented self-organizing on-line communities for open collaboration in science

- ***Organic data science*** is a novel approach to on-line scientific collaboration that supports:
 - **Self-organization of communities** by enabling any user to specify and decompose tasks
 - **On-line community support** by incorporating social sciences principles and best practices
 - **An open science process** by capturing new kinds of metadata about the collaboration that give necessary context to newcomers

Self-Organization through Dynamic Task Decomposition

The screenshot illustrates a web-based task management system designed for self-organization through dynamic task decomposition. The interface includes a sidebar for navigation and search, a main workspace for viewing tasks, and a detailed view for editing task properties.

1 Page Discussion

2a Type^M: medium
Progress^M: 21%
Start date^M: 22nd Aug 2014
Target date^M: 13th Oct 2014
Owner^M: John Smith
Participants: James Williams, Steven Johnson
Expertise: computer science, collaboration

2b Properties
Submitted to: IUI-2015 (by John)

3 All Tasks My Tasks (38) **4** search computer science **5** Timeline SubTasks

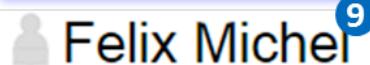
6 Your Overdue Tasks **7** Write about the evaluation a day ago **8** Cut **9** **10** Draft paper about the initial framework design

11 Develop paper outline (100%) **12** Draft initial versions of key sections (26%) **13** Assemble first full draft of the paper **14** Collect final evaluation data **15** Review first full draft of the paper **16** Finalize writing the paper (75%)

17 The plan is to write a paper with some initial results of our work. If you want to be a co-author, add yourself as a participant in a task and make sure you contribute to it with text or feedback on what other people write.

18 Semantic MediaWiki

Organic Data Science: Contributors



Felix Michel 9

Collaboration 15 Computer science 16 Software engineering 0

Current Tasks 9

 Develop social principles to guide collective problem solving 
Will be completed in 10 months

 Write paper about the initial framework design 
Will be completed in 13 days

 Disseminate results from the Organic Data Science framework 
Will be completed in a year

 Draft paper about the initial framework design 
Will be completed in 13 days

 Write related work 
Will be completed in a day

Future Tasks 2

 Finalize writing the paper 
Starts in 9 days

 Collect final evaluation data 
Starts in a day

Completed Tasks 25

 Framework Design 
Completed 10 days ago

 Develop paper outline 
Completed 18 days ago

 Comparison with Other Collaborative Platforms 
Completed a month ago

US

CDEC WEATHER 2010 03 02

Data

Data

- [DOWNLOAD](#)
- [Data Types](#)
 - Daily Sensor Data

Structured Properties

Add

[x]	Barpress	760	(By Admin)
[x]	Depth	1.0214570760727	(By Admin)
[x]	Flow	1550.6185302734	(By Admin)
[x]	ForSite	SMN	(By Admin)
[x]	HasSize	8316	(By Admin)
[x]	SiteLatitude	37.347213745117	(By Admin)
[x]	SiteLongitude	-120.97618103027	(By Admin)
[x]	Slope	0.000099999997473788	(By Admin)
[x]	Velocity	0.65311223268509	(By Admin)

- Used as Input in the following Workflows:

- AF NTM Execution 2 March 2012 to 8 March 2012
- AF EDM Execution 2 March 2012 to 8 March 2012
- AF EM Execution 2 March 2012 to 8 March 2012

Models

AQUAFLOW NTM NIGHTTIMEMODEL

Component

- Input
 - PrevDayHourlyData
 - InputHourlyData
- Param
 - Longitude
 - Latitude
- Output
 - MeanNightDO
 - OutputNTMParams
- Used in the following Templates:
 - AQUAFLOW NTM

Structured Properties

Add

[x]	InputDataVariable	PrevDayHourlyData	(By Anonymous)
[x]	InputDataVariable	InputHourlyData	(By Anonymous)
[x]	InputParamVariable	Longitude	(By Anonymous)
[x]	InputParamVariable	Latitude	(By Anonymous)
[x]	OutputDataVariable	MeanNightDO	(By Anonymous)
[x]	OutputDataVariable	OutputNTMParams	(By Anonymous)

Workflows

AQUAFLOW EDM

Workflow

- [x] AQUAFLOW EDM

Processes

- CALCULATEHOURLYAVERAGES
- FILTERTIMESTAMPSANDDATA
- CONVERTTOSTANDARDFORMAT
- REAERATIONEDM
- CREATEPARAMETERSFILE
- METABOLISMCALCEMPirical
- CREATEPLOTS

Data Variables

- HOURLYDATA
- FILTEREDDATA
- FORMATTEDDATA
- DAILYDATA
- REAERATIONPARAMS
- PARAMETERSFILE
- METABOLISMEDM
- NDM
- PR
- CR24
- PHOTO REST
- GPP
- SUM CORRDO

Parameter Variables

- DATE
- SLOPE
- DEPTH
- FLOW
- BARPRESS
- VELOCITY
- LONGITUDE
- LATITUDE

Workflow Executions

- AF_EDM_Execution_2_March_2014
- AF_EDM_Execution_2_March_2014

Contributor

WATER

Contributor

- WATER

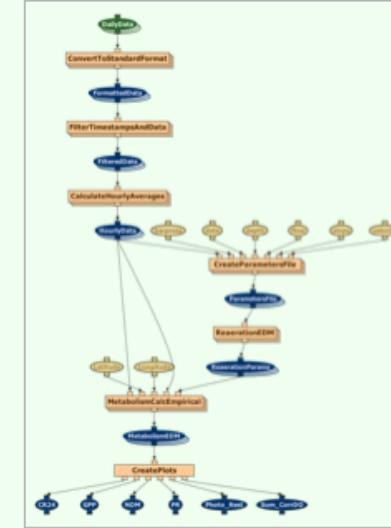
Workflow Created In

- wings.isi.edu

Template File

- AquaFlow EDM.owl

Workflow Template Image



Structured Properties

Add

Credits

Users who have contributed to this Page:

Category: Workflow

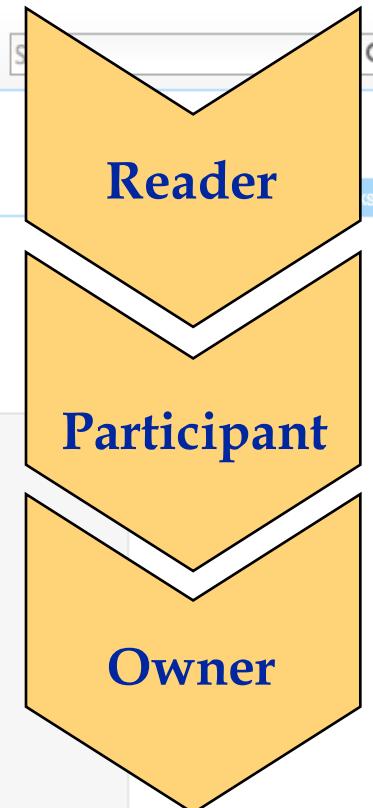
Training Newcomers

The screenshot shows the Organic Data Science wiki interface. At the top left is a logo with a red 'TRAINING' stamp over a person icon. Below it is a sidebar titled 'All Tasks' with an 'Expertise filter' dropdown and a search bar. The main content area lists several tasks under 'Train Hilary on contributing as owner':

- Train Felix on using organic data science wiki
- Train Hilary on using organic data science wiki
- Train Hilary on contributing as participant
 - Train Hilary on exploring tasks
 - Train Hilary on using the task explorer
 - Train Hilary on using subtask explorer
 - Train Hilary on using timeline explorer
 - Train Hilary on participating on tasks
 - Train Hilary on using person pages
 - Train Hilary on understanding basic task state
- Train Hilary on contributing as owner
- Train Gopal on using organic data science wiki
- Train Xuan on using organic data science wiki
- Train Matheus on using organic data science wiki

The screenshot shows a task documentation page for 'Train Hilary on contributing as owner'. The page has tabs for 'Page', 'Discussion', 'Read', 'View source', and 'View history'. The main content area includes:

- A green circular icon with the number '100' next to the title.
- The title: **Train Hilary on contributing as owner**.
- A list of sub-tasks:
 - Train Hilary on creating tasks
 - Train Hilary on using task alert
 - Train Hilary on organizing tasks
 - Train Hilary on understanding extended task states
- A properties box with the following details:
 - Type^M: medium
 - Progress^M: 100%
 - Start date^M: 25th Sep 2014
 - Target date^M: 4th Oct 2014
 - Owner^M: Hilary Dugan
 - Participants: Not defined!
 - Expertise: owning
- A legend at the bottom: 'Legend: M Mandatory | States: ■ Not defined, ■ Valid, ■ Inconsistent with parent'



Click on the documentation in the Properties box below to go over documentation, then do the practice below.

Practice: Complete the subtasks!

What Features Are Used to Manage Tasks?

The figure displays three screenshots of a web-based task management system, likely from the Organic Data Science platform. The heatmaps indicate user interaction with various features across the interface.

Screenshot 1: Project Overview

This screenshot shows the main project page for "G16 Workshop". It includes a summary table with the following data:

Type	Value
Progress	high
Start date	18th Jul 2014
Target date	24th Oct 2014
Owner	Paul Hanson
Participants	Chris Duffy
Expertise	Not defined!

The page also lists several tasks under "All Tasks" and "My Tasks".

Screenshot 2: Task Details

This screenshot shows a detailed view of a task titled "Develop Framework for Organic Data Science". The task details include:

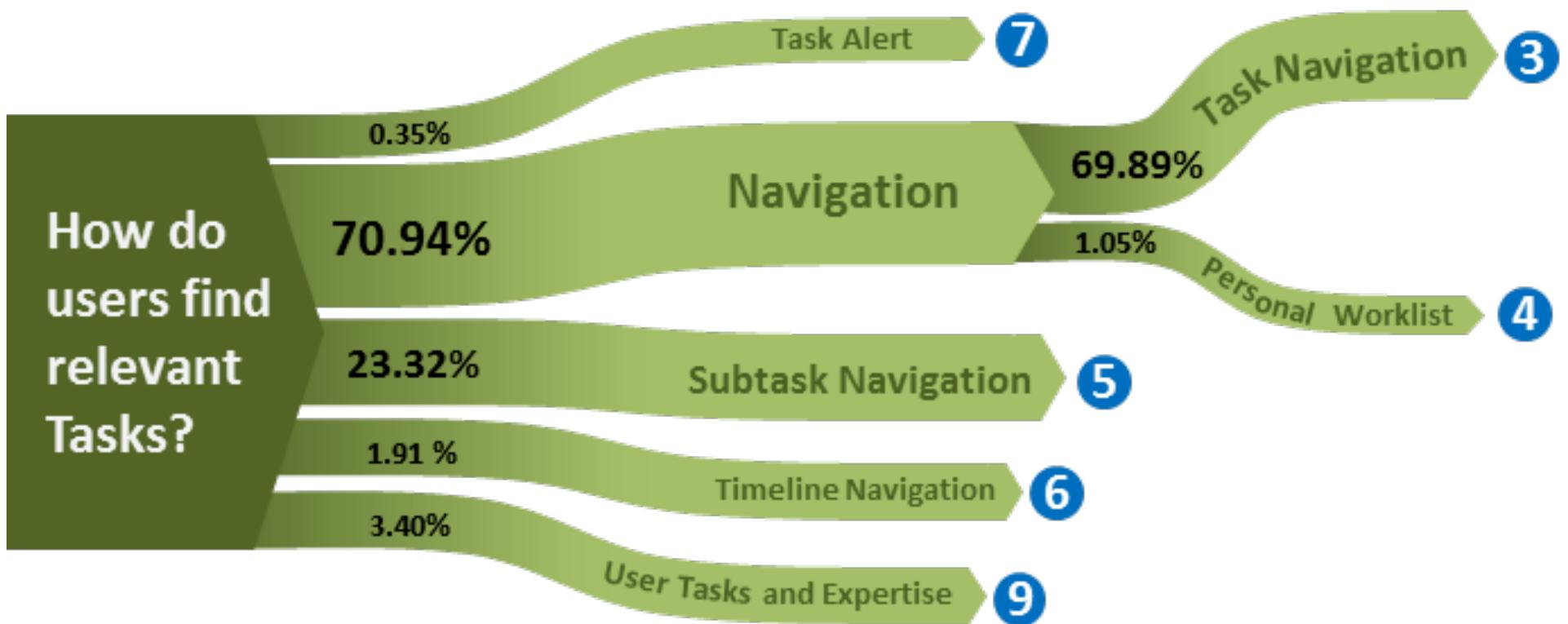
- Type: high
- Progress: 32%
- Start date: 1st Oct 2013
- Target date: 20th Jan 2014
- Owner: Paul Hanson
- Participants: Chris Duffy
- Expertise: Not defined!

The page also includes a description of the project goal and a sidebar with "Highest Contributors".

Screenshot 3: Main Page

This screenshot shows the main page of the Organic Data Science platform. It features a sidebar with links like "Main page", "Recent changes", and "Tools". The main content area includes sections for "What is Organic Data Science?", "Our Science Goal: The Age of Water and Carbon", and "Our Science Goal: The Age of Water and Carbon".

How Do Users Find Relevant Tasks?

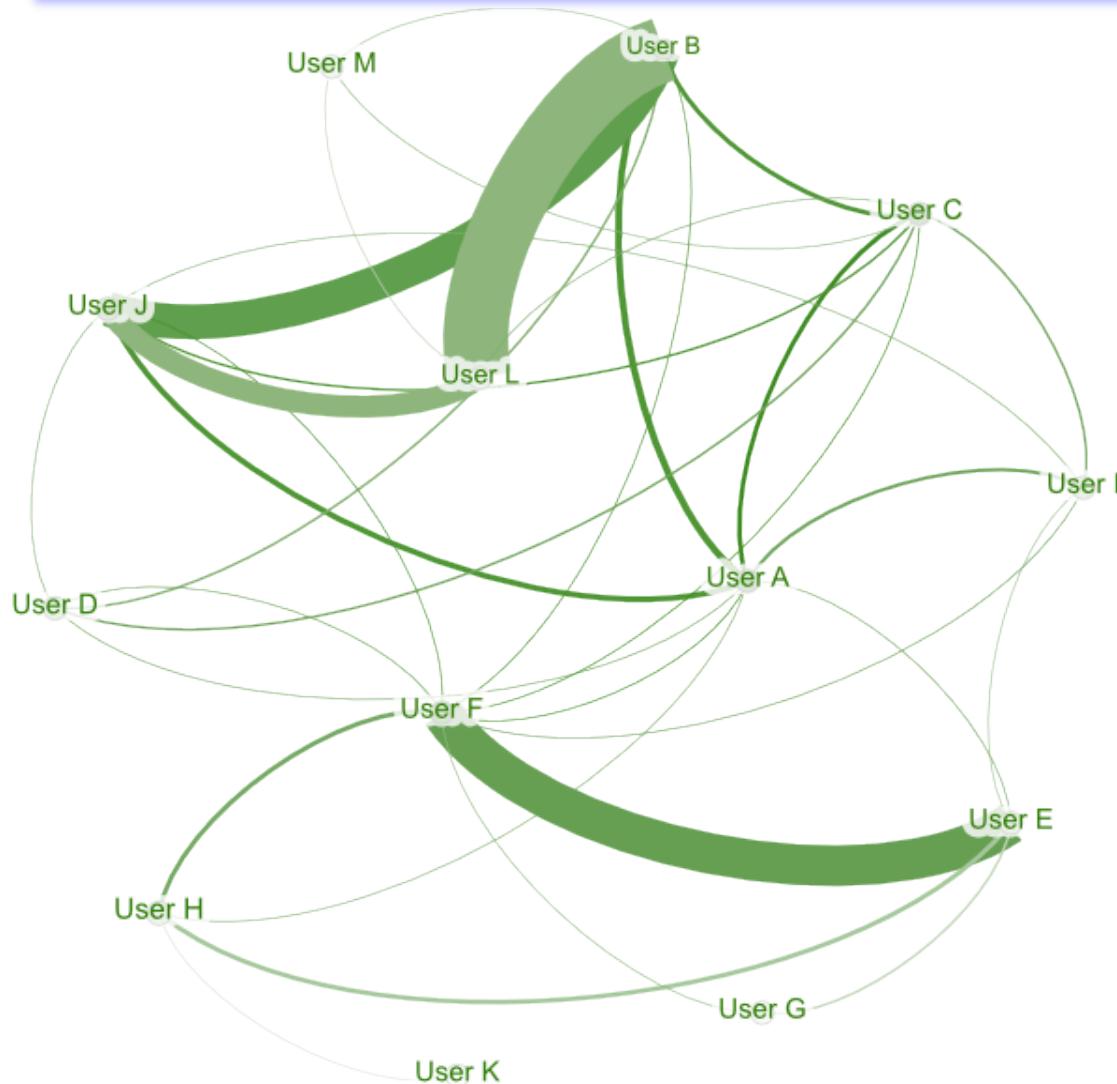


Are Users Collaborating?



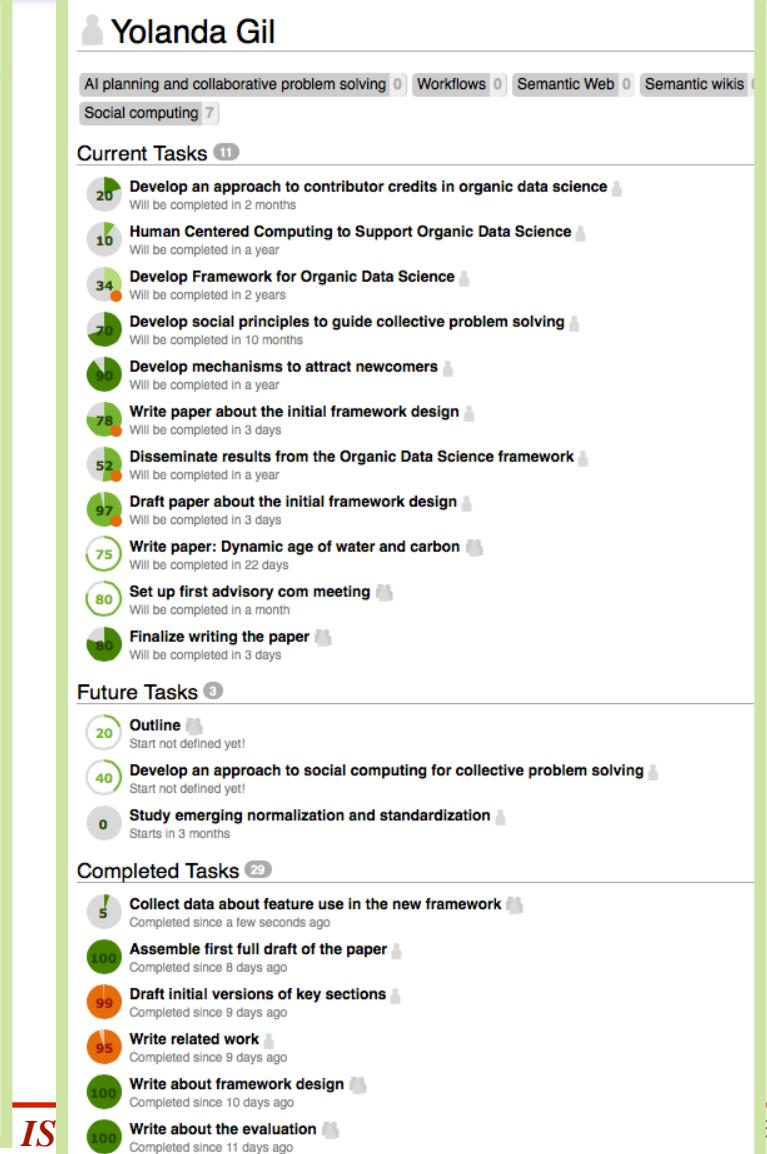
- (A)** 52% of tasks are viewed by more than one person
- (B)** 72% of tasks have more than one person signed up
- (C)** 19% of tasks have more than one person editing metadata
- (D)** 11% of tasks have more than one person editing content

What Does the Social Network of Collaborators Look Like?

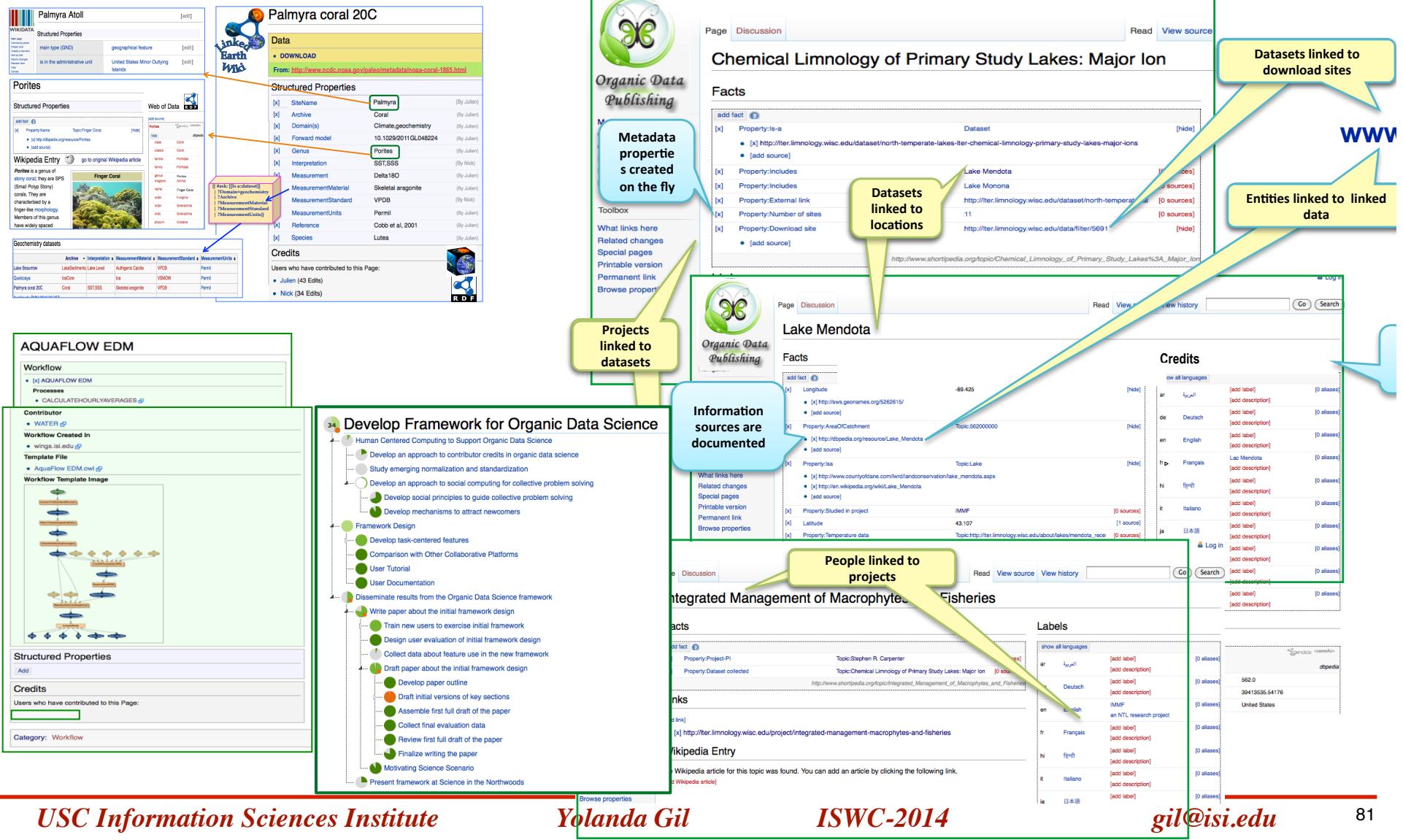


- Network of users (nodes) linked by shared tasks
- Links across all users
- Two distinct subgroups
 1. Water
 2. Software

A Semantic Challenge: Email-less Coordination for Projects



A Semantic Challenge: Open Science Processes



Semantic Challenges in Getting Work Done

■ To dos

- Managing personal to dos
- Managing coordinated to dos

■ Knowledge rich tasks in science

- Automatic paper generator
- A Web of semantic workflows/processes

■ Open science

- Email-less coordination of projects
- Open science processes

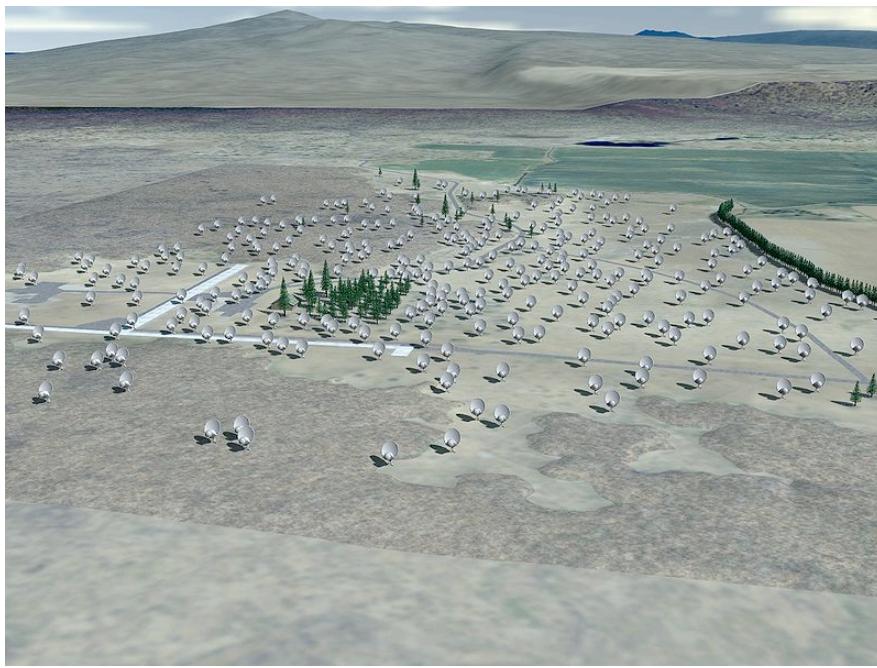
<http://www.isi.edu/~gil>

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

<http://www.organicdatascience.org>

<http://discoveryinformaticsinitiative.org>

“We need bigger glasses and more hands in the water” – J. Tarter, SETI Institute



Discovery Informatics: Knowledge-Rich Science Infrastructure

SCIENCE sciemag.org

10 OCTOBER 2014 • VOL 346 ISSUE 6206

ARTIFICIAL INTELLIGENCE

Amplify scientific discovery with artificial intelligence

Many human activities are a bottleneck in progress

By Yolanda Gil,¹ Mark Greaves,²
James Hendler,^{3*} Haym Hirsh⁴

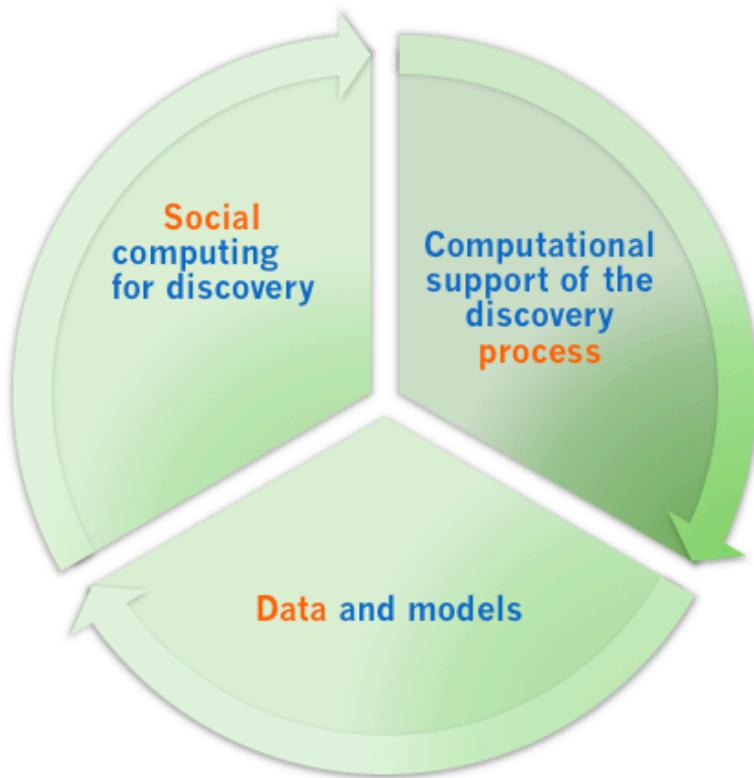
Technological innovations are penetrating all areas of science, making predominantly human activities a principal bottleneck in scientific progress while also making scientific advancement more subject to error and harder to reproduce. This is an area where a new generation of artificial intelligence (AI) systems can radically transform the prac-

increased the numbers of interested participants; Moore's law and steady exponential increases in computing power; and exponential increases in, and broad availability of, relevant data in volumes never previously seen. Those scientific efforts that have leveraged AI advances have largely harnessed sophisticated machine-learning techniques to create correlative predictions from large sets of "big data." Such work aligns well with the current needs of peta- and exascale science. However, AI has far broader capacity to ac-

information-finding beyond current search limitations.

We can project a not-so-distant future where "intelligent science assistant" programs identify and summarize relevant research described across the worldwide multilingual spectrum of blogs, preprint archives, and discussion forums; find or generate new hypotheses that might confirm or conflict with ongoing work; and even rerun old analyses when a new computational method becomes available. Aided by such a system, the scientist will focus on more of the creative aspects of research, with a larger fraction of the routine work left to the artificially intelligent assistant.

"AI-based systems that can represent hypotheses ... can reduce the error-prone human bottleneck in ... discovery."



PSB Workshop (January 2015)

KDD Workshop (August 2014):
<http://ailab.ist.psu.edu/idkdd14/>



AAAI Workshop (July 2014):
<http://discoveryinformaticsinitiative/diw2014>



AAAI Fall Symposium (Nov 2013):
<http://discoveryinformaticsinitiative/dis2013>

AAAI Fall Symposium (Nov 2012):
<http://discoveryinformaticsinitiative/dis2012>

Microsoft eScience Summit (Aug 2012)
Workshop on Web Observatories
for Discovery Informatics



PSB Workshop (Jan 2013):
on Computational Challenges of
Mass Phenotyping



NSF Workshop (Feb 2012):
<http://discoveryinformaticsinitiative/diw2012>

A View from Biomedical Research: The NIH Big Data To Knowledge (BD2K) Initiative

PEBOURNE

Professional Developments Worth Sharing

HOME

ABOUT

21
DEC
2013

Taking on the Role of Associate Director for Data Science at the NIH – My Original Vision Statement

by pebourne

On March 3, 2014 I will begin the job of Associate Director for Data Science (ADDS) at the National Institute of Health (NIH). I will report directly to NIH Director, Dr. Francis Collins. When I originally applied for the position in April 2013 I was asked to prepare a short vision statement. That statement

“Discovery informatics is in its infancy. Search engines are grappling with the need for deep search, but it is doubtful they will fulfill the needs of the biomedical research community when it comes to finding and analyzing the appropriate datasets. Let me cast the vision in a use case. As a research group winds down for the day algorithms take over, deciphering from the days on-line raw data, lab notes, grant drafts etc. underlying themes that are being explored by the laboratory (the lab’s digital assets). Those themes are the seeds of deep search to discover what is relevant to the lab that has appeared since a search was last conducted in published papers, public data sets, blogs, open reviews etc. Next morning the results of the deep search are presented to each member as a personalized view for further post processing. We have a long way to go here, but programs that incite groups of computer, domain and social scientists to work on these needs will move us forward.”

A View from Geosciences: The NSF EarthCube Initiative

Outcomes

Transform practices within the geosciences community spanning over the next decade

Provide unprecedented new capabilities to researchers and educators

Vastly improve the productivity of community

Accelerate research on the Earth system

Provide a knowledge management framework for the geosciences



EarthCube

GROUPS



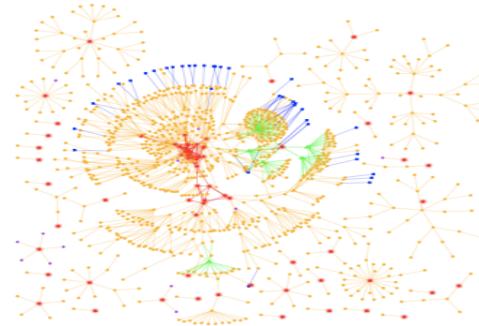
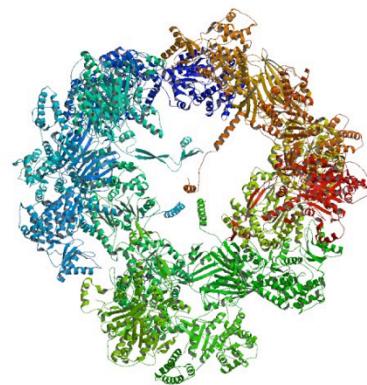
Data Workflows Semantics Governance

<http://www.earthcube.org/>

What Might the Future Look Like?

YOU: What are you working on?

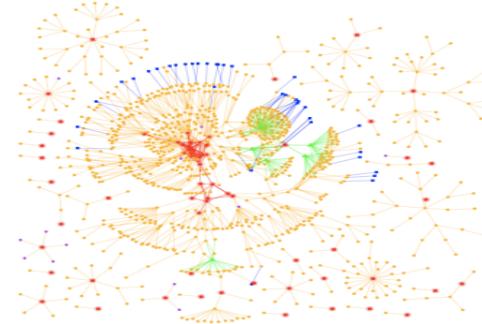
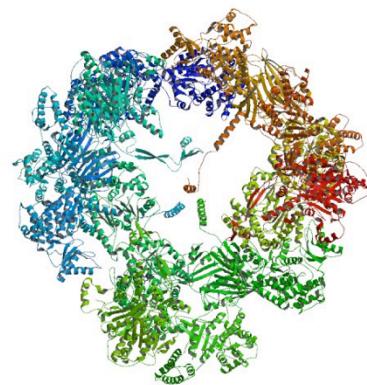
OTHER PERSON: I am really busy, working on...



In the Future

YOU: What are you working on?

OTHER PERSON: I am really busy, working on...



YOU: Yes, but aren't you glad that we can get our work done faster?

Thank you!



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

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

<http://www.organicdatascience.org>

<http://discoveryinformaticsinitiative.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 (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)
- *Biomedical workflows:* 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 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!*