



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

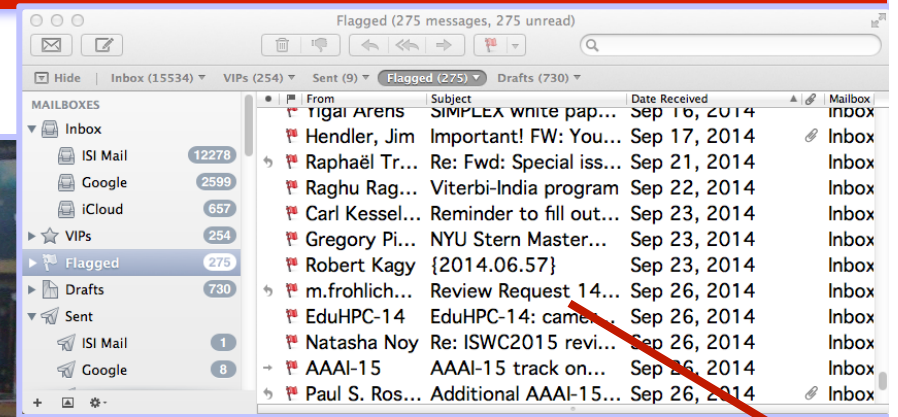
Outline

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 lists2 semantic challenges
2. Knowledge rich tasks in science
 - Semantic workflows2 semantic challenges
3. Collaborative tasks in science
 - Organic data science2 semantic challenges
4. Closing thoughts

To Dos



Email requests

Daily to-dos

To Dos

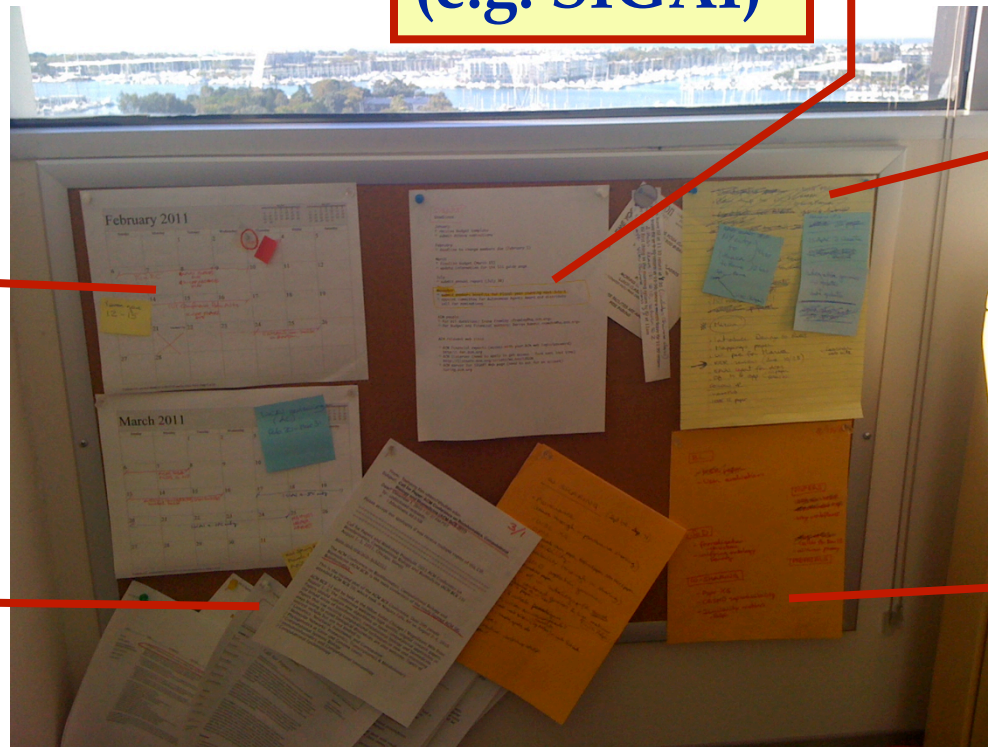
Monthly travel
and deadlines

By annual
timeline
(e.g. SIGAI)

This
week

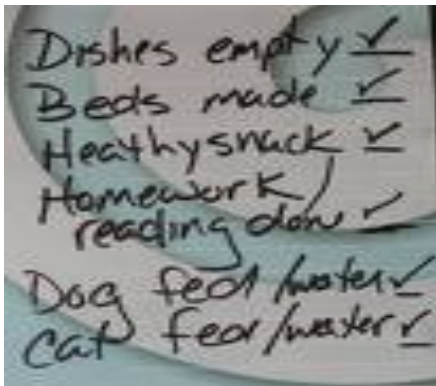
CFPs,
BAAs

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!

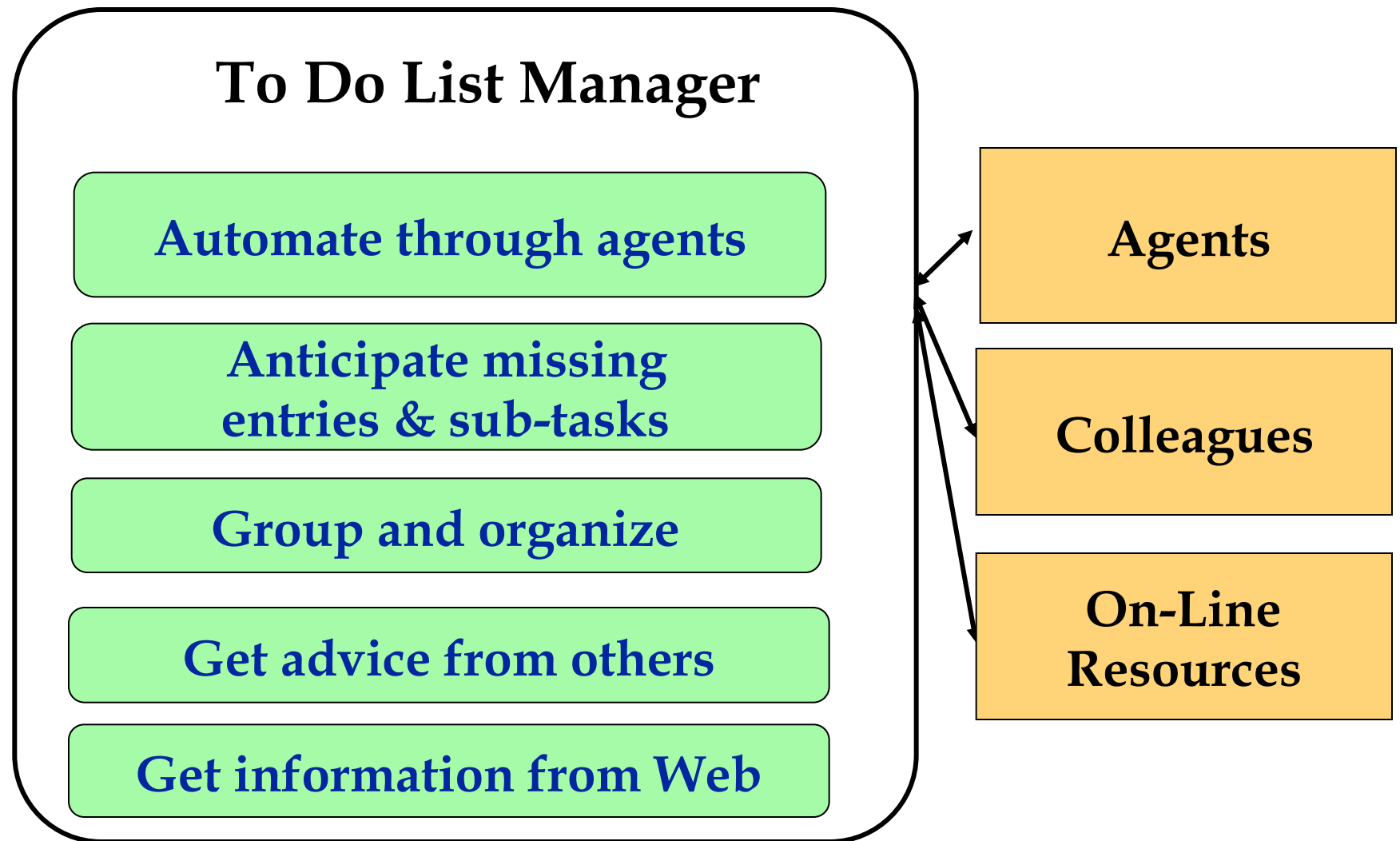
- ~1500 items collected from ~325 people
- Many are not amenable to automation
- Many could be automated fully or in part

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

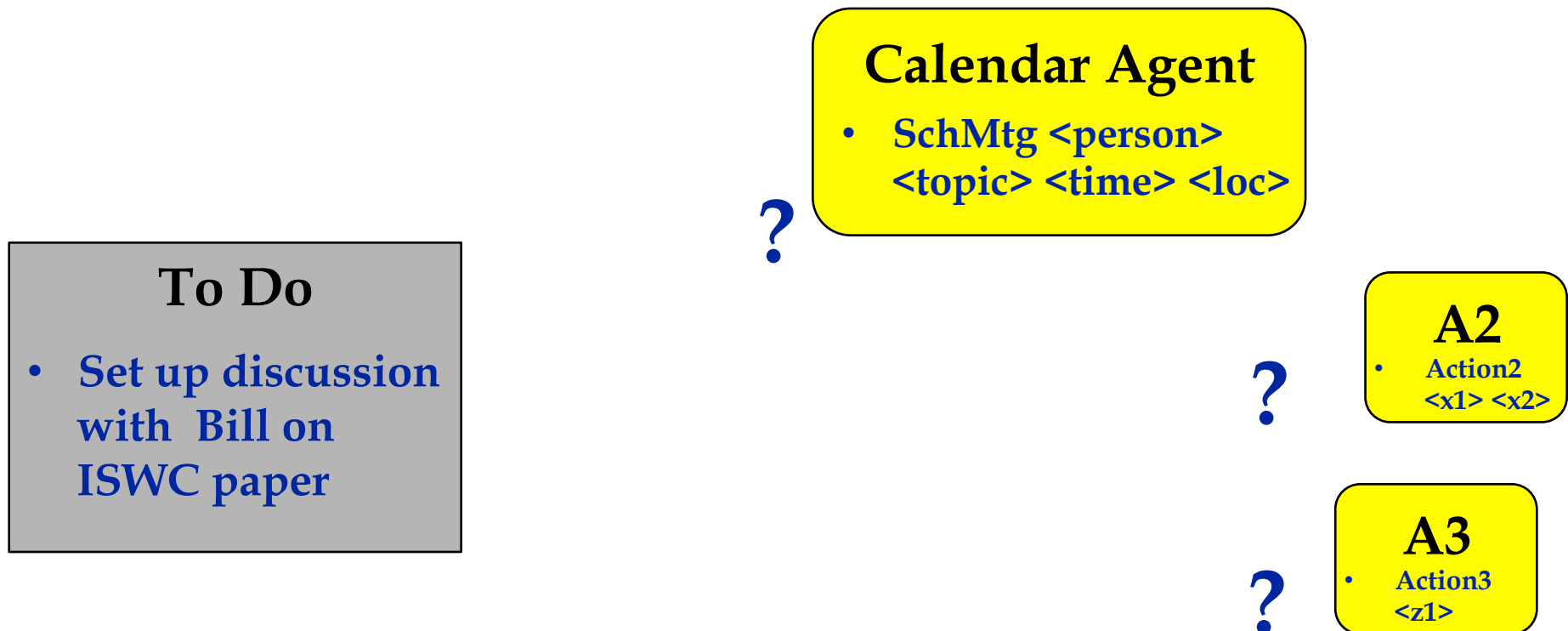
Opportunities



Agents:

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

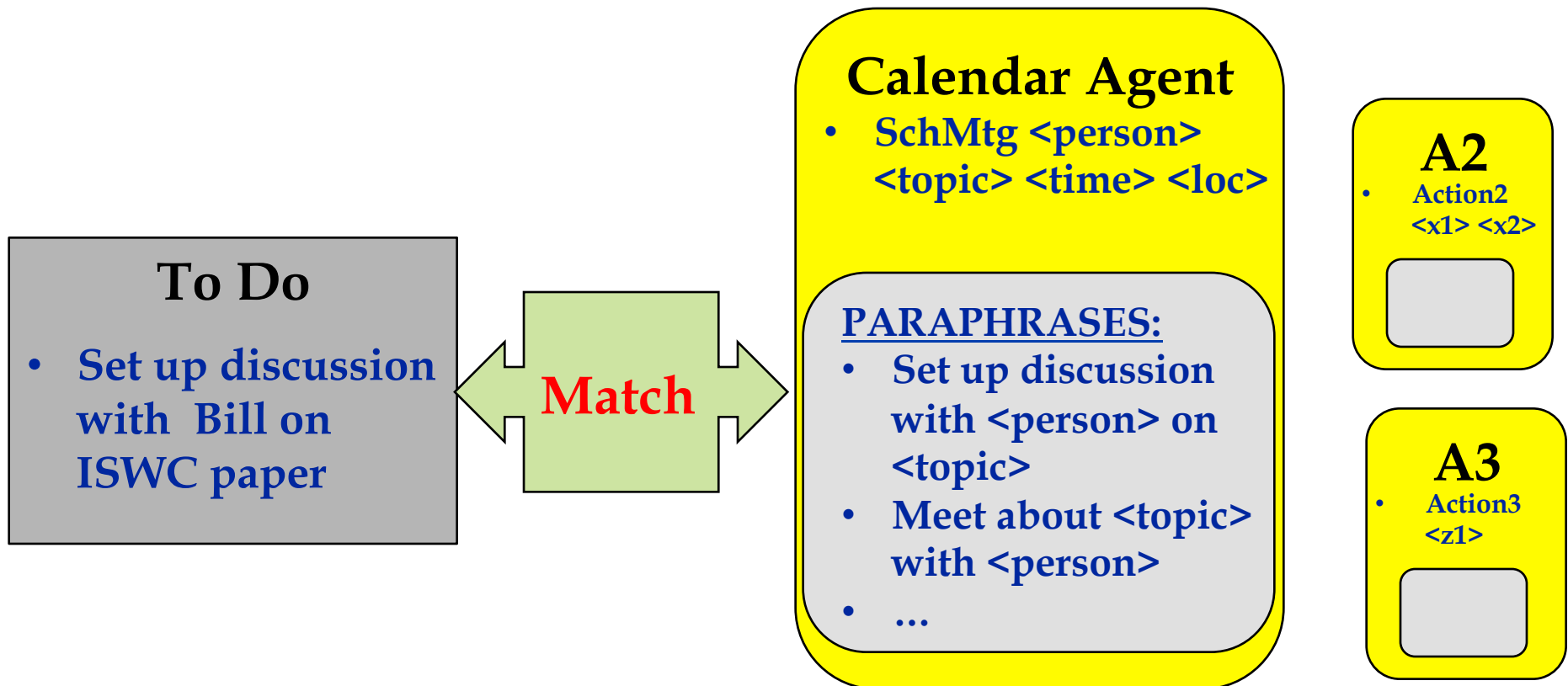
- Match agent capabilities to user's to dos



Agents:

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

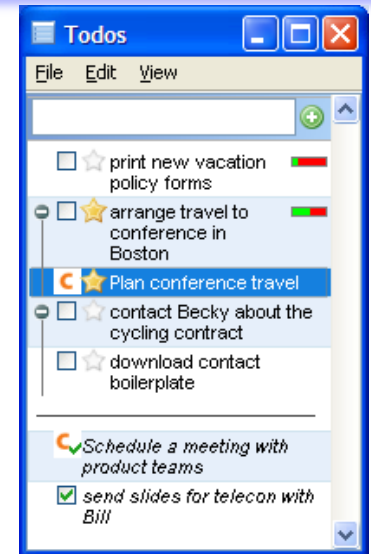
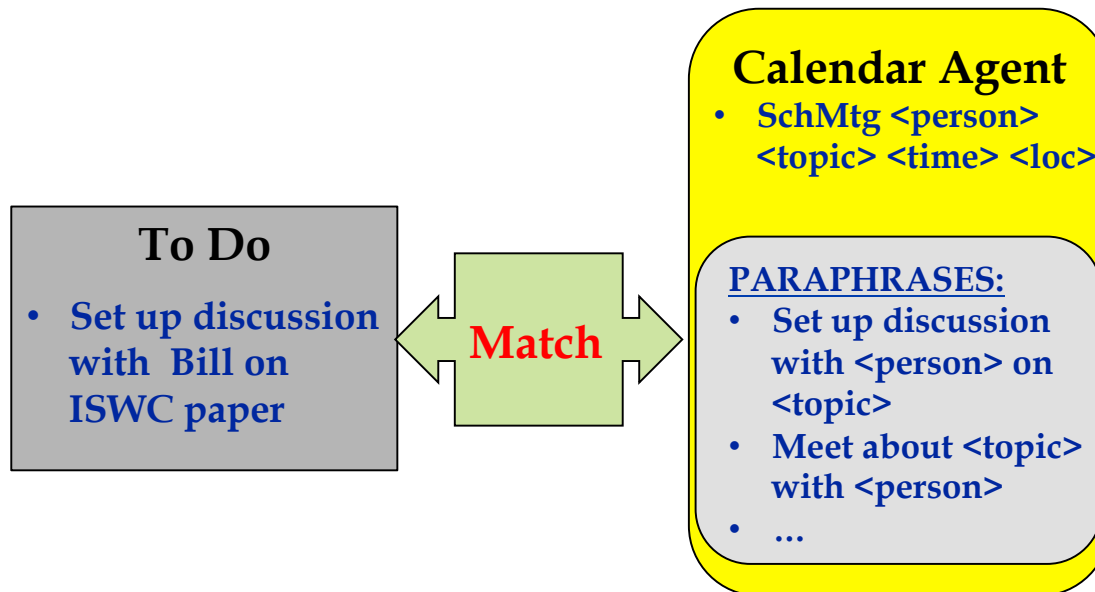
- Use paraphrase patterns of agent capabilities to match them to user's to dos



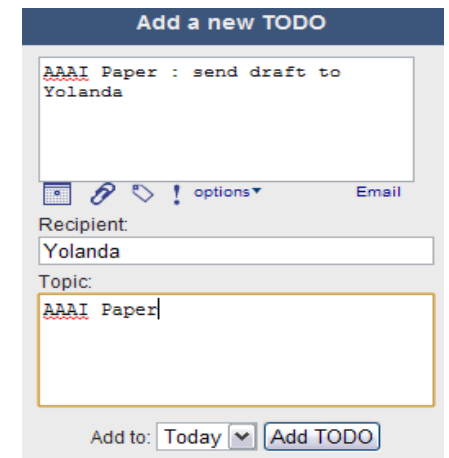
Agents:

Beamer for CALO and Radar [Gil & Ratnakar AAI 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

Knowledge



To Do List Manager

Automate through agents

**Anticipate missing
entries & sub-tasks**

Group and organize

Get advice from others

Get information from Web

Paraphrase Game [Chklovski 2005]



"this can help you"

Another Way To Say It:

this could be helpful



TRY



HINT



GIVE UP

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, AAI 2005]


You Have Taught Me: **Learner2**

A **copier** is also typically used to **duplicate** a **document**.

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**.

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

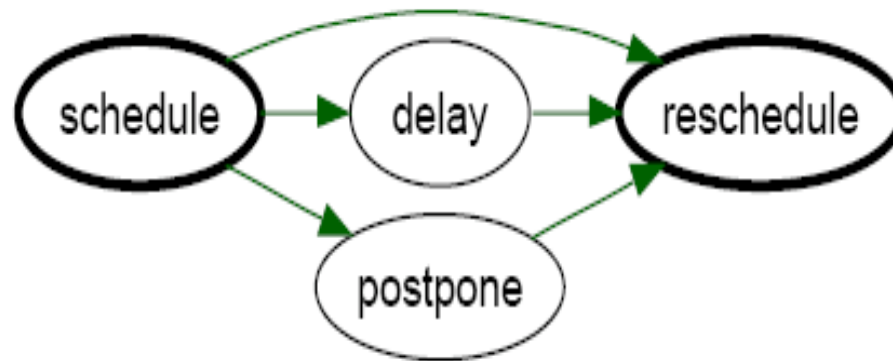
As an admin assistant, if helping with **setting up a videoconference**, if you need to deal with a **conference time**, an important activity may be: **agree upon** it

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

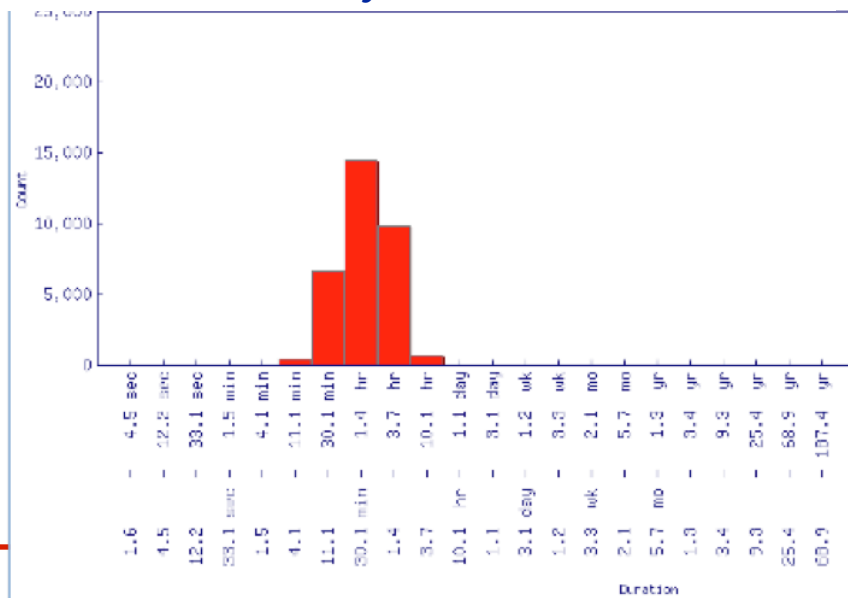
Possible problem: When **attending a meeting**, **not having an LCD projector** may cause a problem.

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

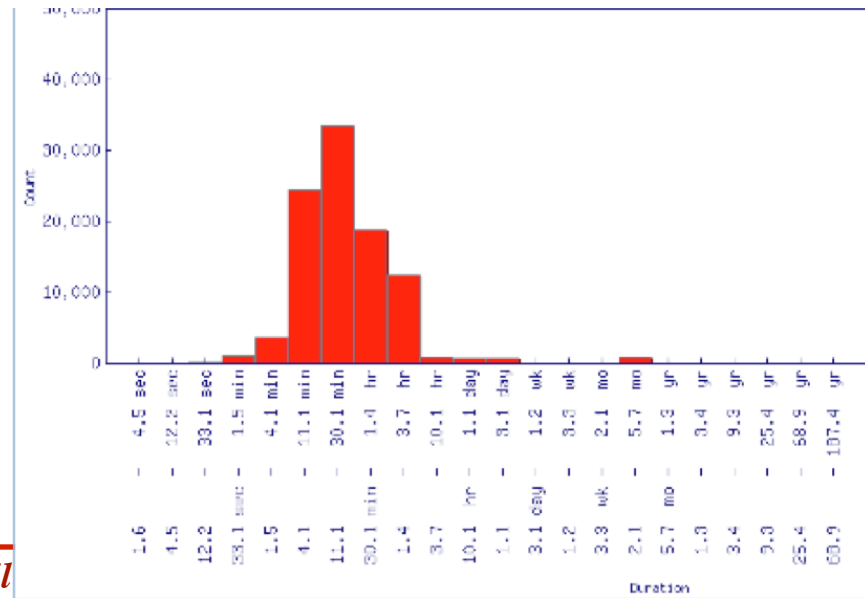
VerbOcean [Chkovski and Pantel, IJCNLP 2005]



“Lunch” - likely duration 1hr



“Presentation” - likely duration 10mins



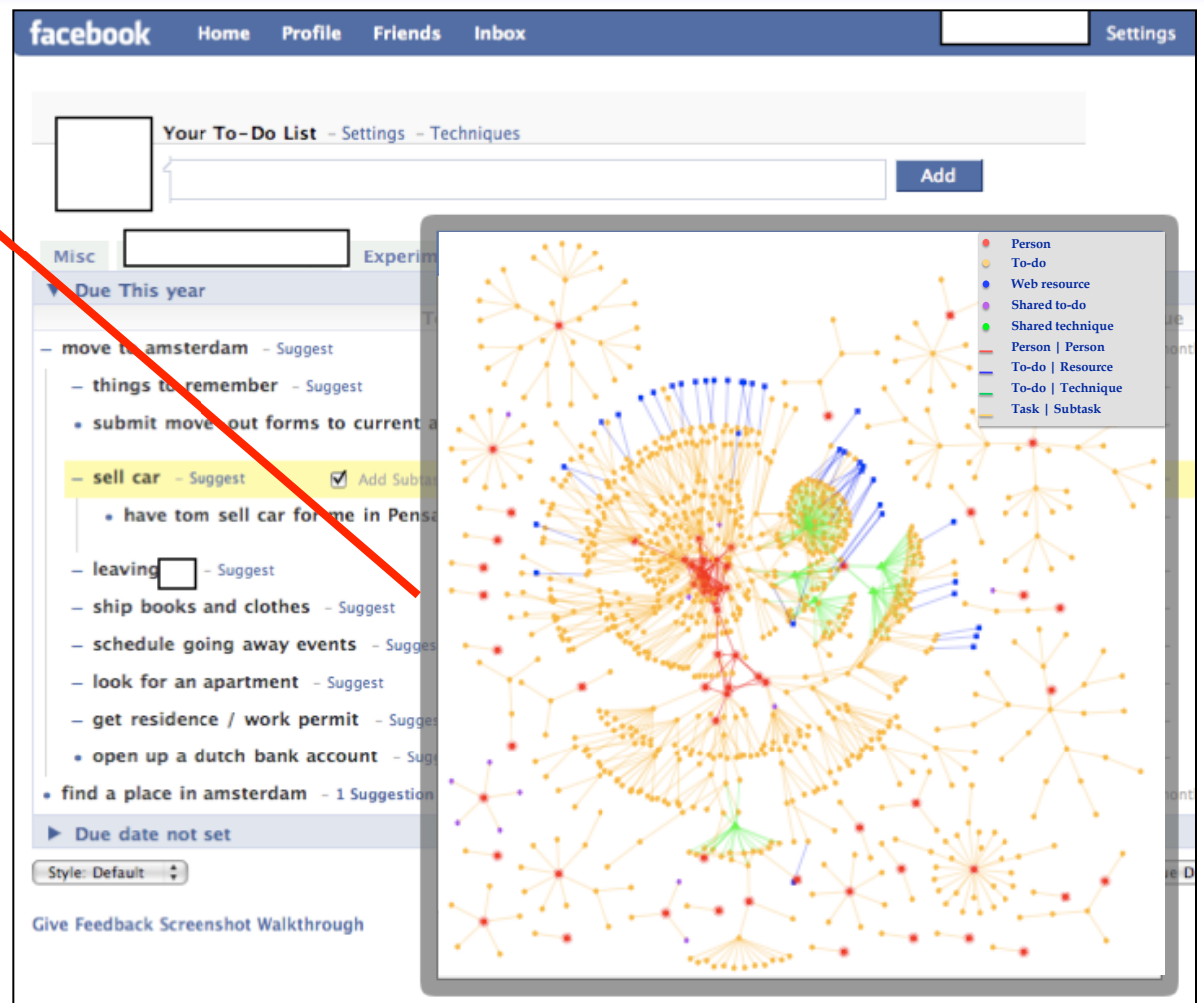
Managing To Dos through Colleagues: Social Task Networks [Groth et al 2010]

■ 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]



[About](#) [API](#) [Contact](#)

pack for a camping trip

Submit

[\[debug\]](#) Displaying 6 How-Tos that we think may match. [Tell us if they do!](#) Search powered by [Google](#).

pack for a camping trip

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 over do 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 Vrandečić: “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

[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

[7h remaining in 3 tasks >](#)

Client A web

[choose color scheme](#)

Client liked the pastels/gradients scheme.

Monday 1 Aug 2005, 7:29 pm

Client A web

[design draft](#)

Called joan. Confirmed that black/chrome color scheme is preferred over orange/fuschia.

Sunday 31 Jul 2005, 1:51 pm

Tasks

Return library books in

Set task view here

- Pick up the milk Today
- Finish TPS reports
- Take over the world

1 Supermarket

- [Pick up the milk](#)

| Task | Progress | Due Date |
|------------------------|----------|-----------|
| Buy Aruba tickets | 0/1 | 8/2/2005 |
| Get oil change | 0/0.5 | |
| Design draft | 4/8 | 7/29/2005 |
| Call Mike to get specs | 0/2 | 8/2/2005 |
| Pitch design refresh | 0/1 | |

Todos

- print new vacation policy forms
- arrange travel to conference in Boston
- Plan conference travel**
- contact Becky about the cycling contract
- download contact boilerplate
- Schedule a meeting with product teams
- send slides for telecon with Bill

Add a new TODO

AAAA Paper : send draft to Yolanda

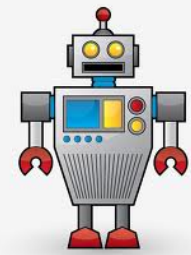
Recipient: Yolanda

Topic: AAAA Paper

Add to: Today

To Do List Manager

Agents/services, other people, advice web sites



Amazon search results for 'book'.

Weather for Marina del Rey, CA (90292).

Today in Marina del Rey: Sunny, 81°F.

amazon search

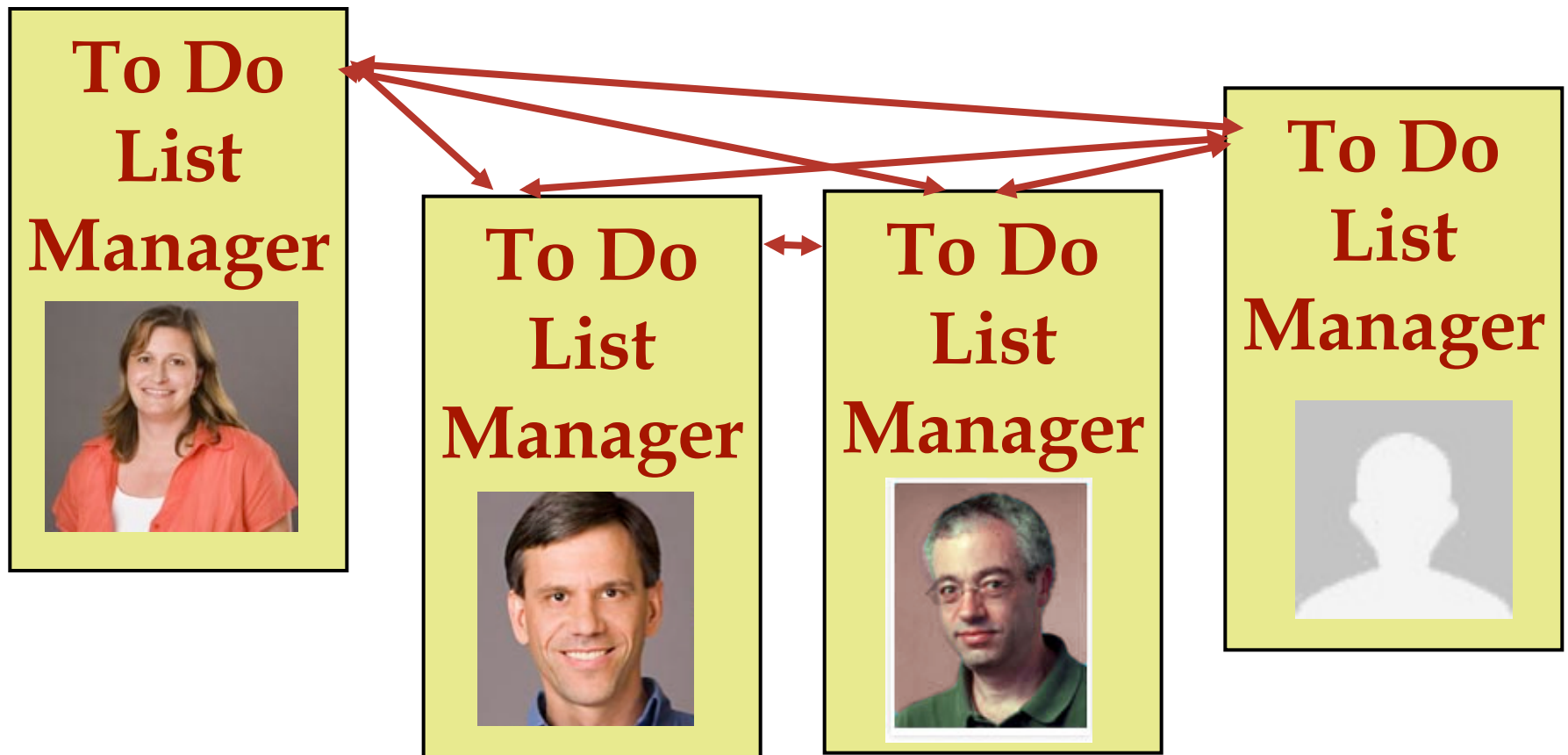
URL builder: Base: http://www.amazon.com/

Fetch Page: http://www.amazon.com/

Fetch Page: http://www.amazon.com/

Fetch Page: http://www.amazon.com/

A Semantic Challenge: Coordinating To Dos of Different People



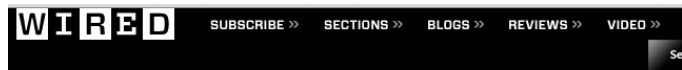
Semantic Challenges in Getting Work Done

- 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

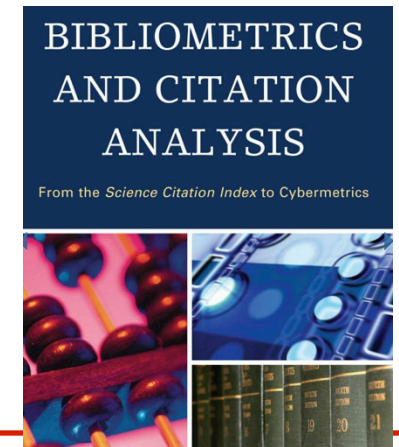
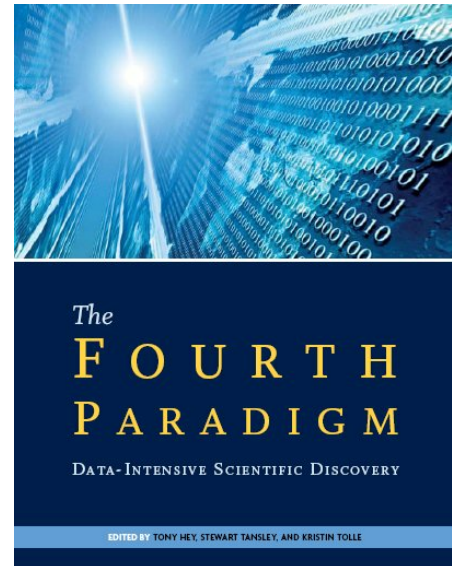
SCIENCE : DISCOVERIES

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

ISI 2011

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)



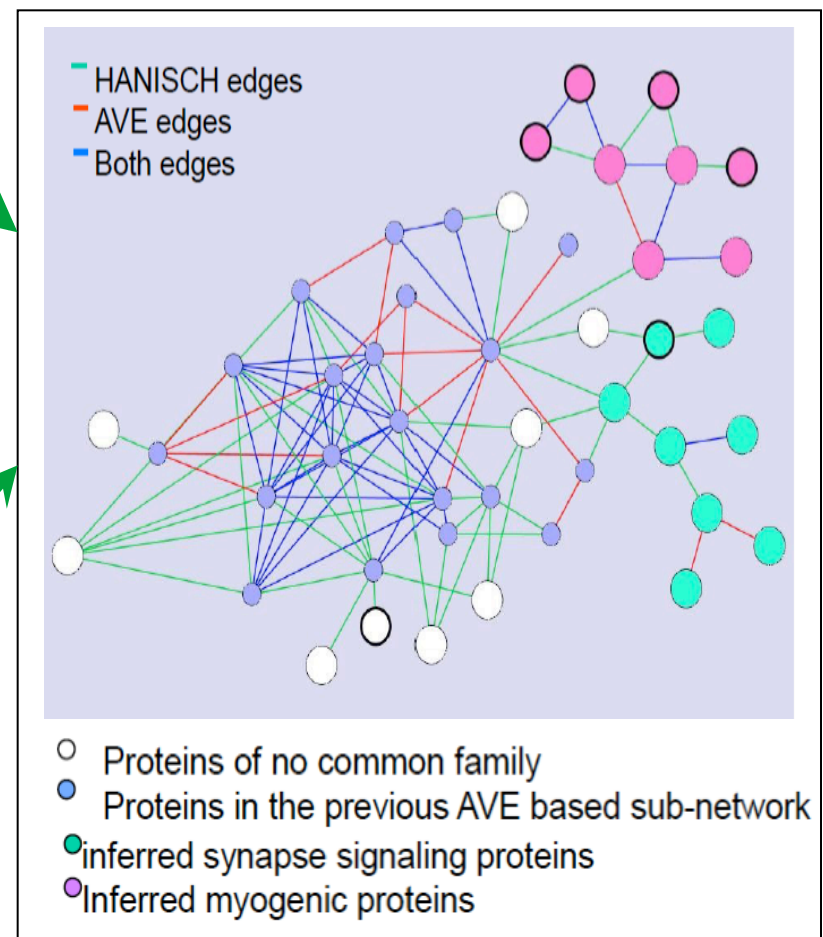
Text extraction
from publications

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^{1,2}, 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



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



CA.GOV DEPARTMENT OF WATER RESOURCES California Data Exchange Center

Home Query Tools Precipitation River Forecast River Stages/Flow Reservoirs Snow Stations Weather

Lookup Station Metadata Real-time Data Stations Daily Data Stations

SHOW/HIDE LEFT SIDEBAR

MOST POPULAR LINKS

- Executive Summary
- Real-time Data
- Daily Data
- Monthly Data
- Historical Data
- Data Plotter
- Station Search
- Station Locator
- Daily Water
- Temperatures
- Reports
- Other Related Data Sources
- Contact CDEC Staff

MERCED RIVER NEAR STEVINSON

Map of surrounding area

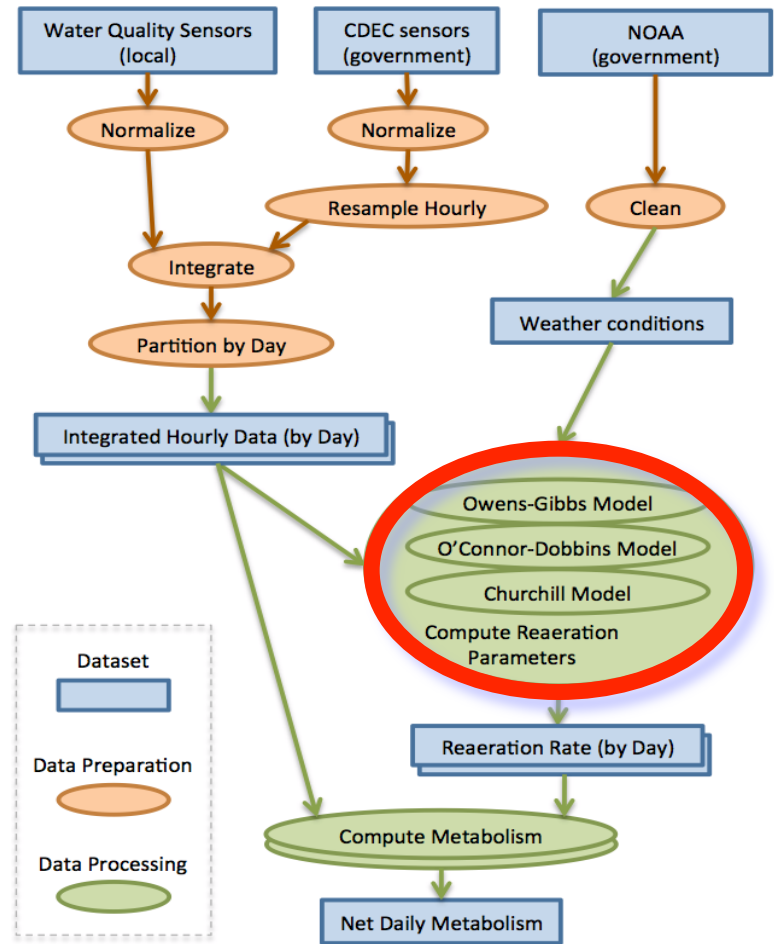
| | | | |
|-----------------|--|-------------|--------------|
| Station ID | MST | Elevation | 82' ft |
| River Basin | MERCED R | County | MERCED |
| Hydrologic Area | SAN JOAQUIN RIVER | Nearby City | STEVINSON |
| Latitude | 37.371000°N | Longitude | 120.931000°W |
| Operator | CA Dept of Water Resources Data Collection | | |

River Stage Definitions

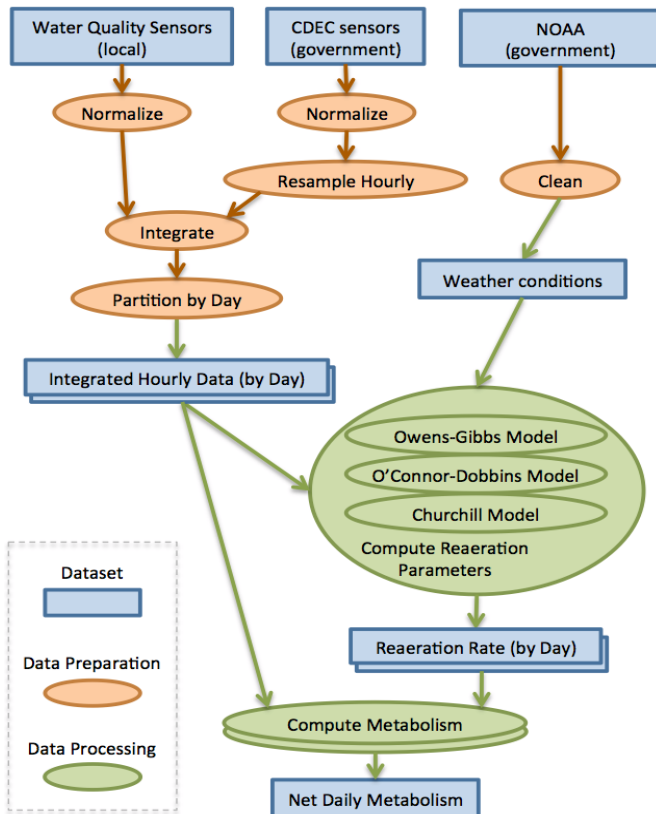
| | | | |
|---------------|------------|----------------|-------------------|
| Datum 0 | 0.00' NGVD | Peak of Record | 12/05/1950 73.80' |
| Monitor Stage | 67.0' | Flood Stage | 71.0' |

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 07/04/200 |
| FLOW, RIVER DISCHARGE, cfs | (event) | (FLOW) | COMPUTED | 03/20/199 |
| RIVER STAGE, feet | (event) | (RIV STG) | SATELLITE | 03/20/199 |

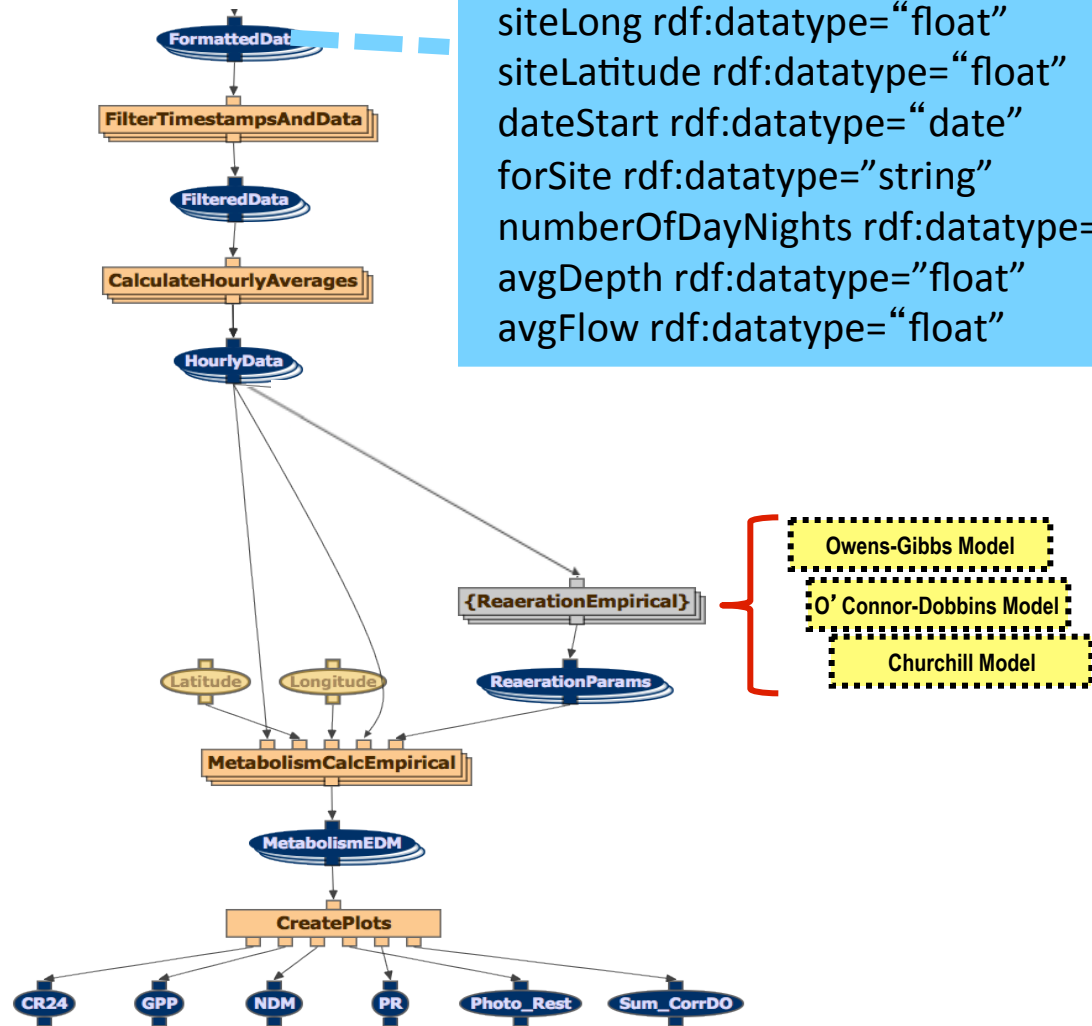


A Semantic Workflow



```

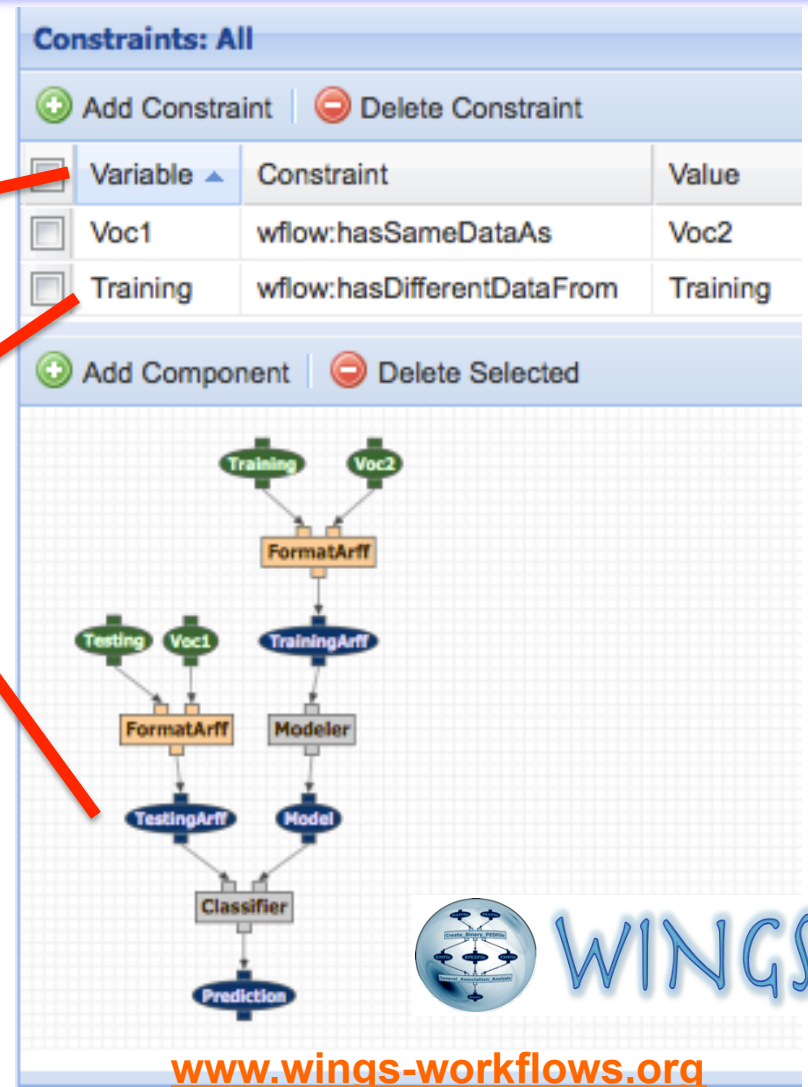
DailySensorData
  isa Hydrolab_Sensor_Data
  siteLong rdf:datatype="float"
  siteLatitude rdf:datatype="float"
  dateStart rdf:datatype="date"
  forSite rdf:datatype="string"
  numberOfDayNights rdf:datatype="int"
  avgDepth rdf:datatype="float"
  avgFlow rdf:datatype="float"
    
```



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 iEMSs 2014]

Classes of models/ components

Wings Po
Home Analysis

Components: Default

- Add Component
- CalculateHourlyAverage
- ConvertToStandardFormat
- CreateParametersFile
- CreatePlots
- FilterTimestampsAndData
- HourlyAverage
 - Aquaflow_Hourly_Average_Inputs
- MergeMetabolismResults
- Metabolism
 - MetabolismDay
 - MetabolismCalcEmpirical
 - MetabolismCalculation
- ReaerationEmpirical
 - ReaerationEmpirical
 - ReaerationCM
 - ReaerationODM
 - ReaerationOGM
 - ReaerationPhysics
 - ReaerationEDM

| I/O | | |
|------------------|-----------------------------|--------|
| 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 Data constraints

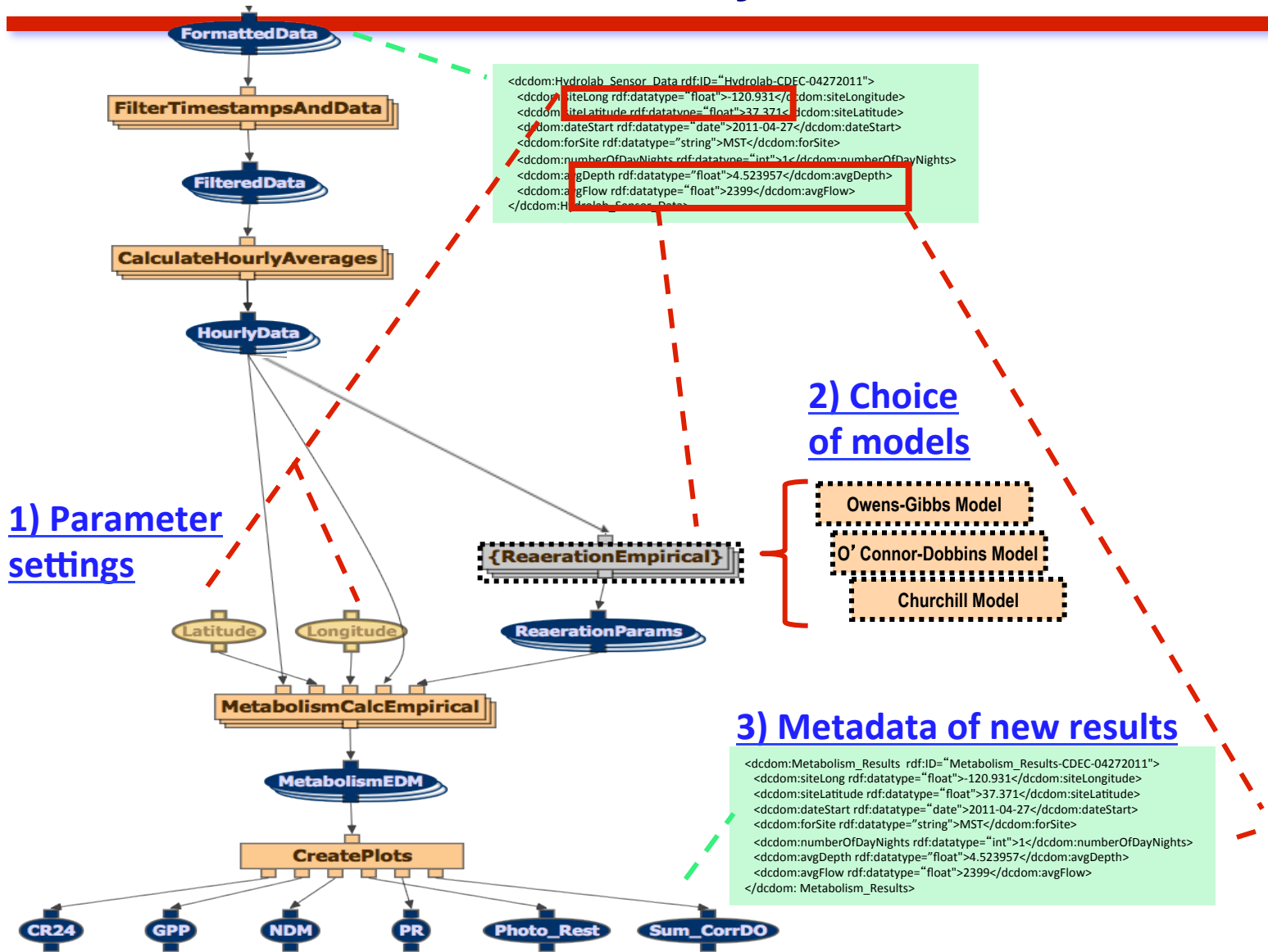
| I/O | | |
|------------------|-----------------------------|--------|
| Input Data | | |
| Name | Type | Prefix |
| InputParameters | dcdom:Hourly_Averaged_Input | -i1 |
| Input Parameters | | |
| Name | Type | Prefix |
| velocity | xsd:float | -p2 |
| slope | xsd:float | -p4 |
| depth | xsd:float | -p1 |
| flow | xsd:float | -p3 |
| Output Data | | |
| Name | Type | Prefix |
| K2Result | | |

Use constraints

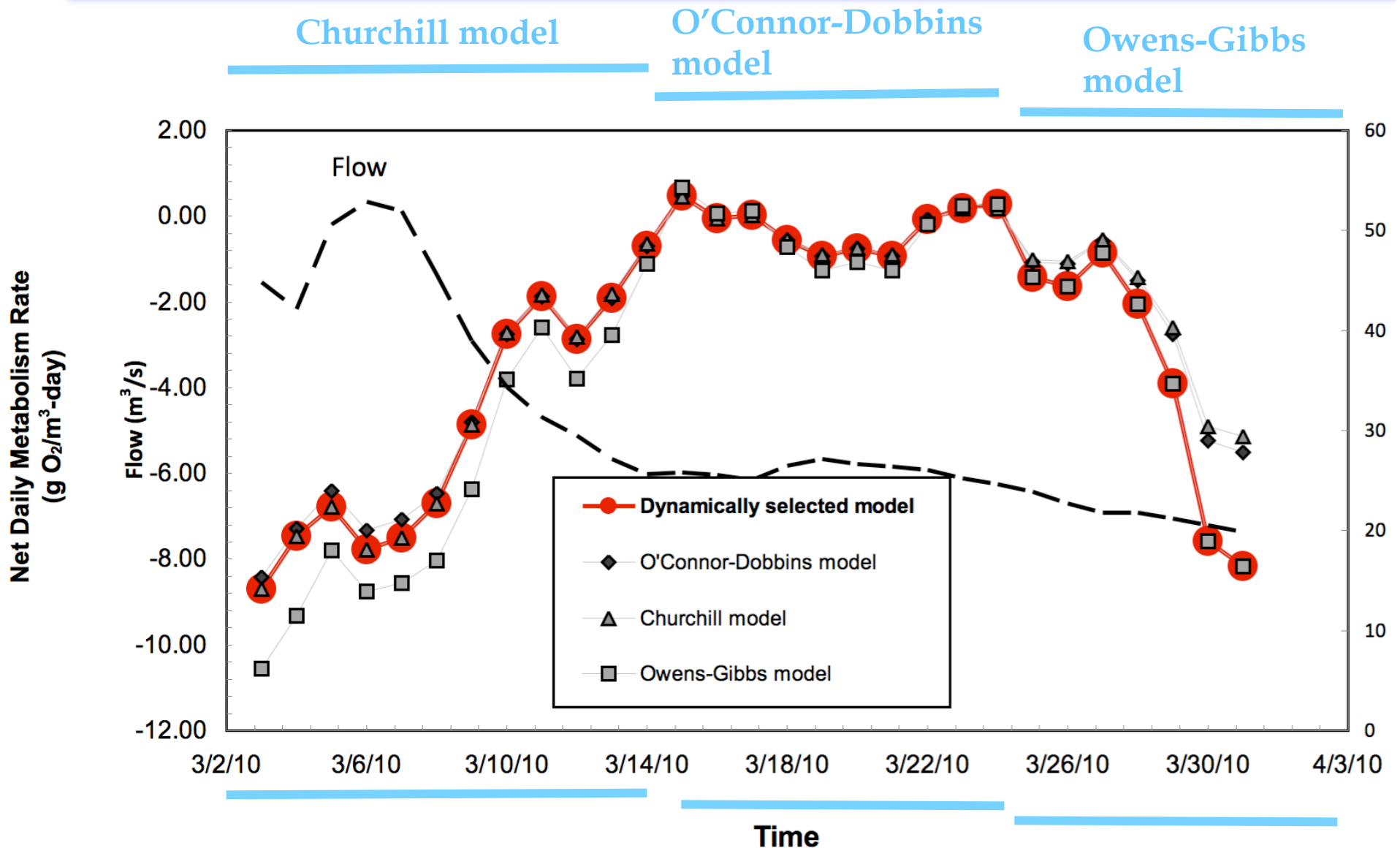
```

;; 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:isInvalid 'true')]
    
```

WINGS Specializes Workflow Based on Characteristics of Daily Data

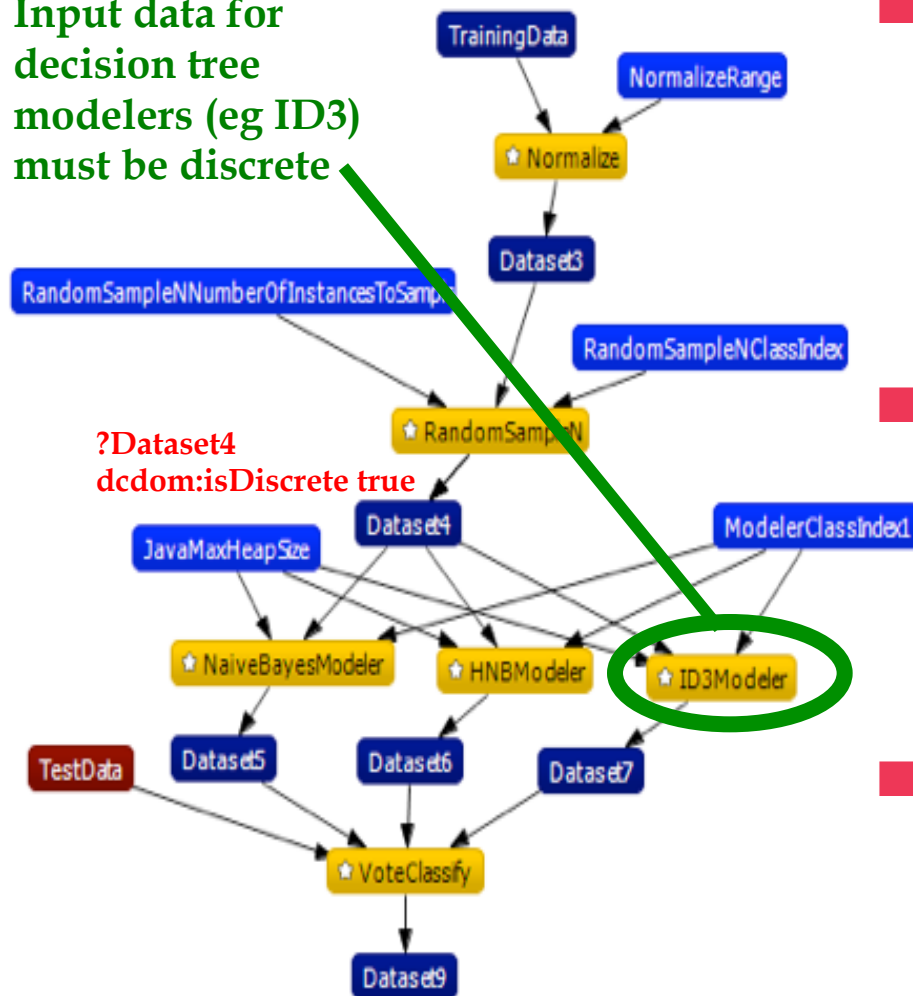


WINGS Dynamically Selects Appropriate Model Based on Daily Sensor Readings



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:

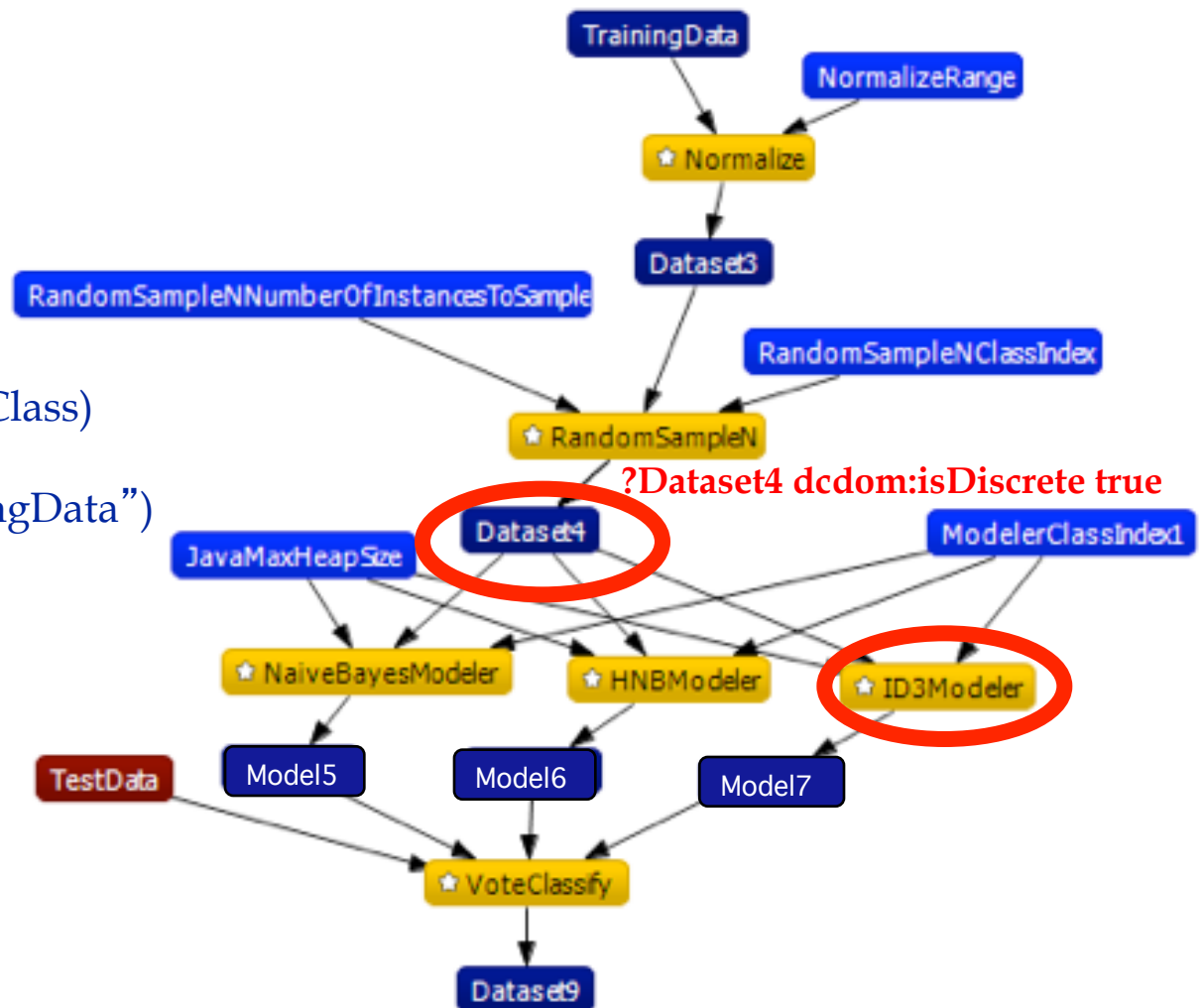
[**modelerSpecialCase2:**

(?c rdf:type pcdom:ID3ModelerClass)

(?c pc:hasInput ?idv)

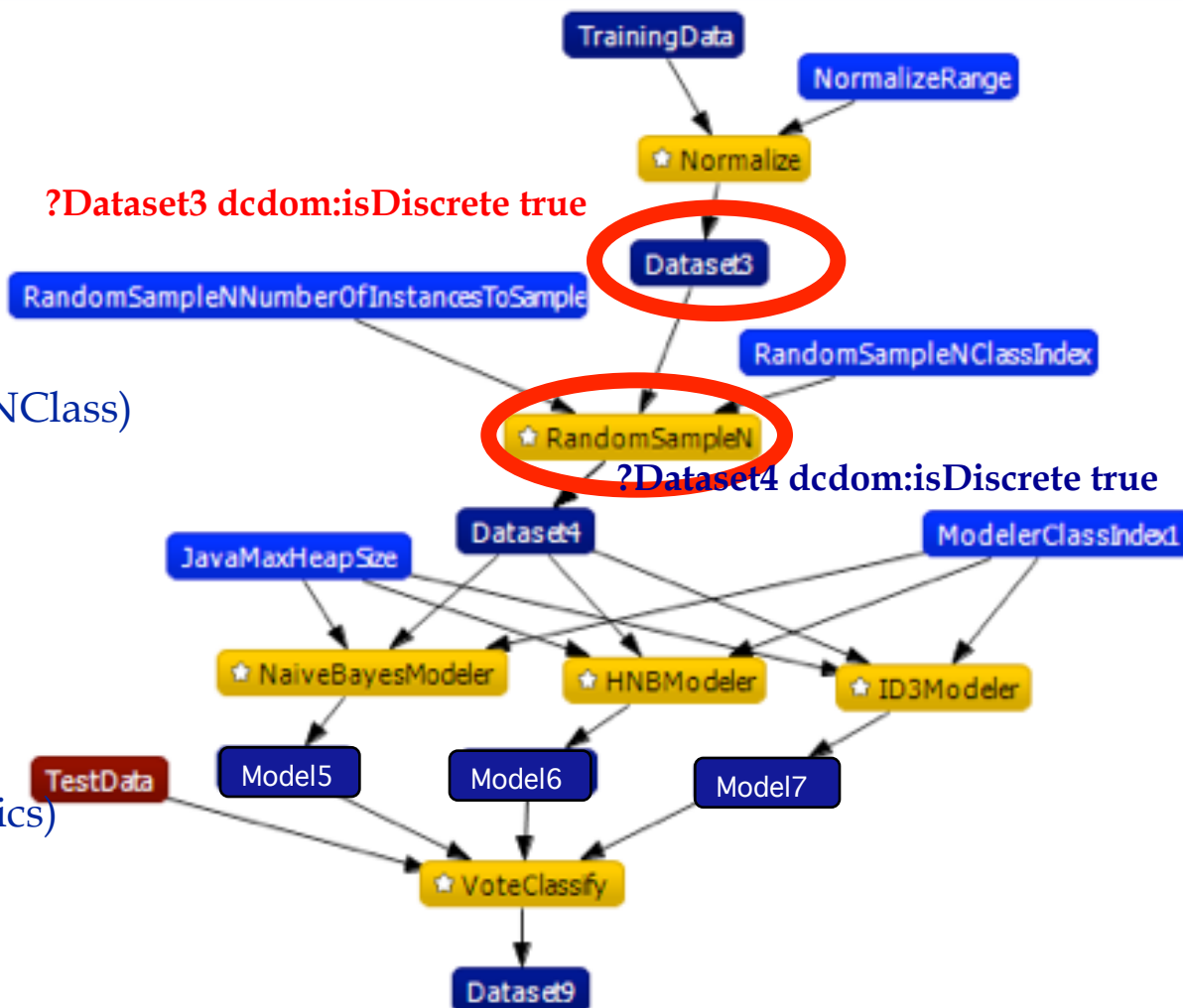
(?idv pc:hasArgumentID "trainingData")

-> (?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)

?TrainingData dcdom:isDiscrete true

?Dataset3 dcdom:isDiscrete true

?Dataset4 dcdom:isDiscrete true

Rule in Component Catalog:

[normalizerTransfer:

(?c rdf:type pcdom:NormalizeClass)

(?c pc:hasOutput ?odv)

(?odv pc:hasArgumentID

"normalizeOutputData")

(?c pc:hasInput ?idv)

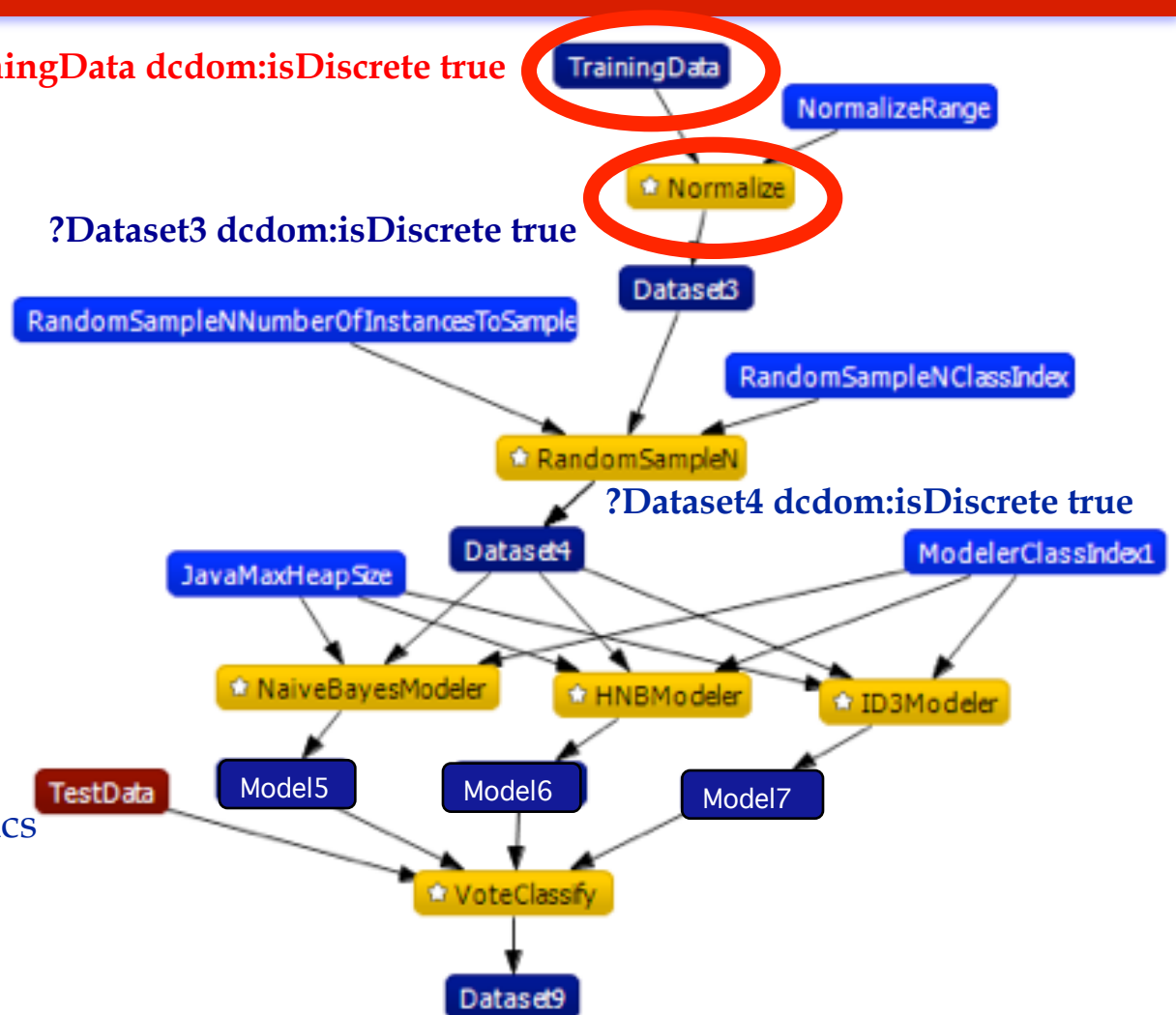
(?idv pc:hasArgumentID

"normalizeInputData")

(?odv ?p ?val)

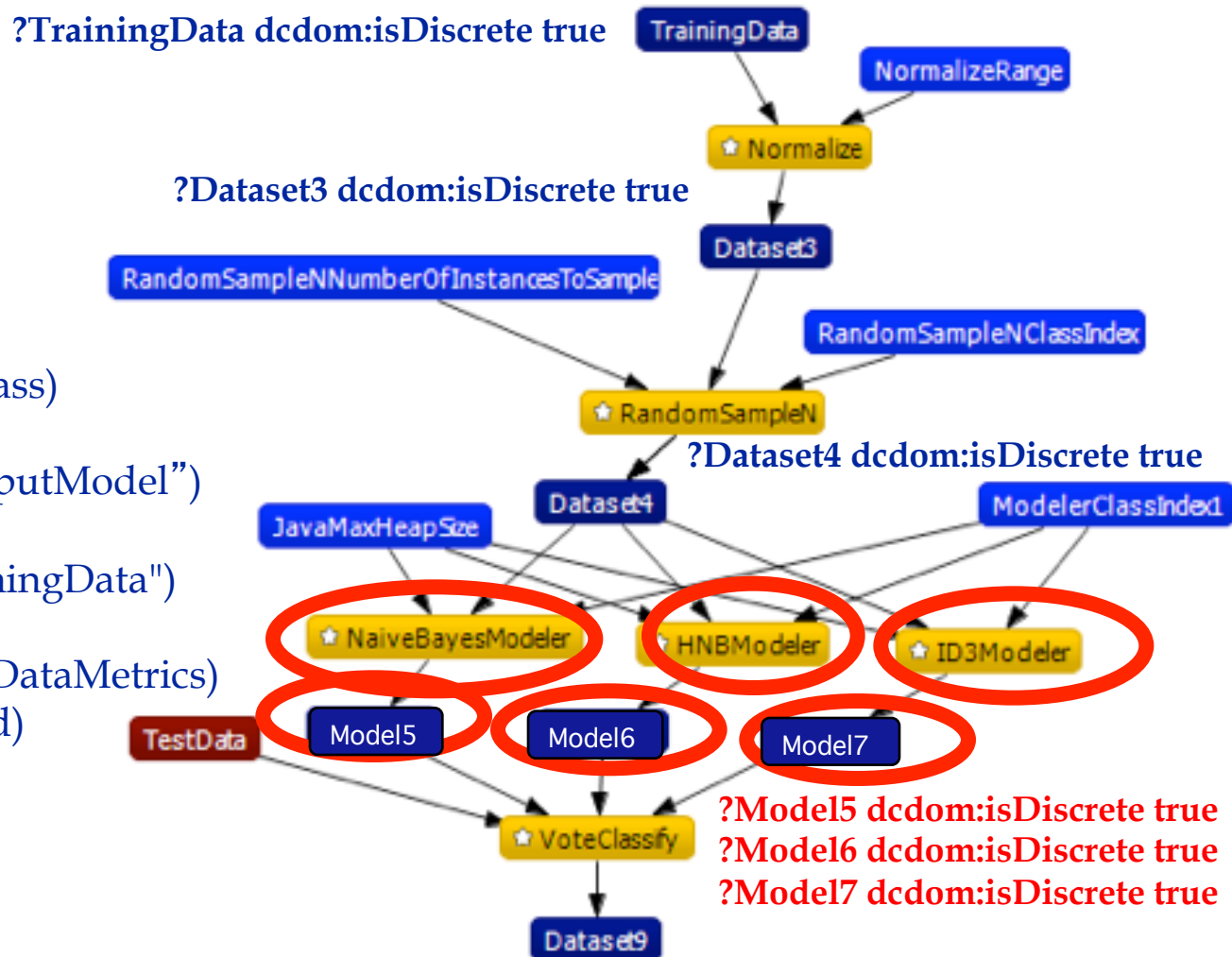
(?p rdfs:subPropertyOf dc:hasMetrics

-> (?idv ?p ?val)]



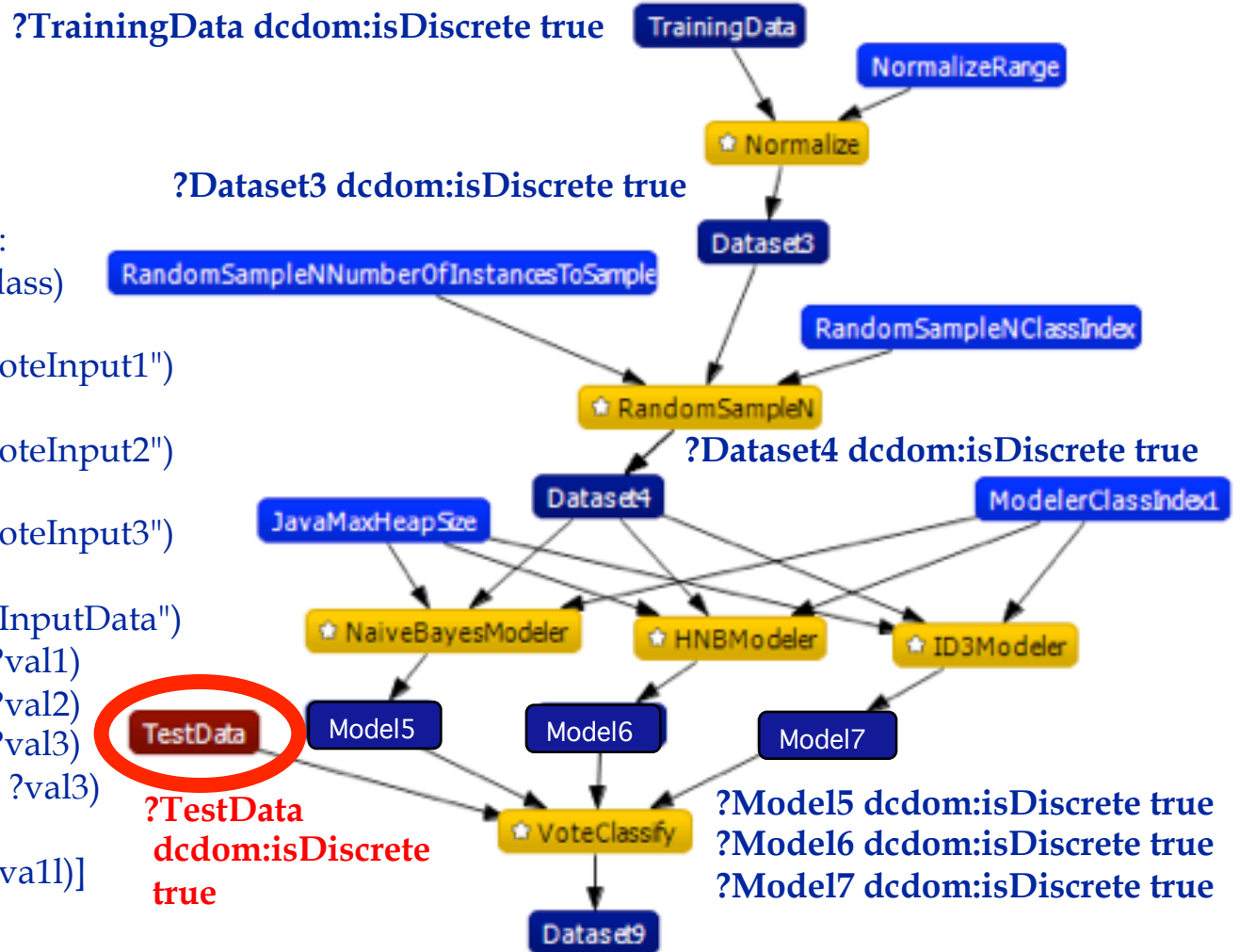
Example (Step 4 of 5)

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 dcdom:isSampled)
 -> (?odv ?p ?val)]

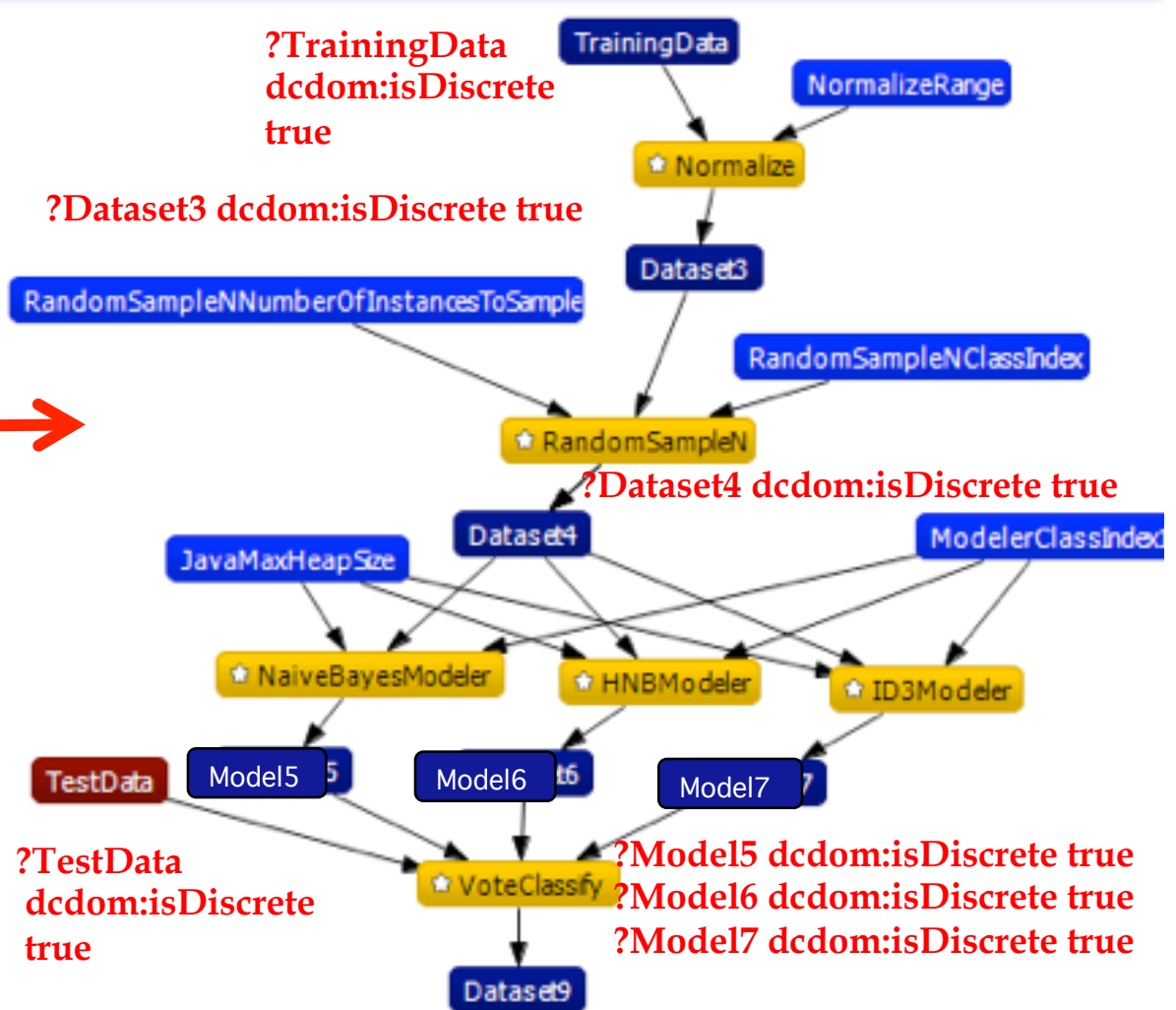
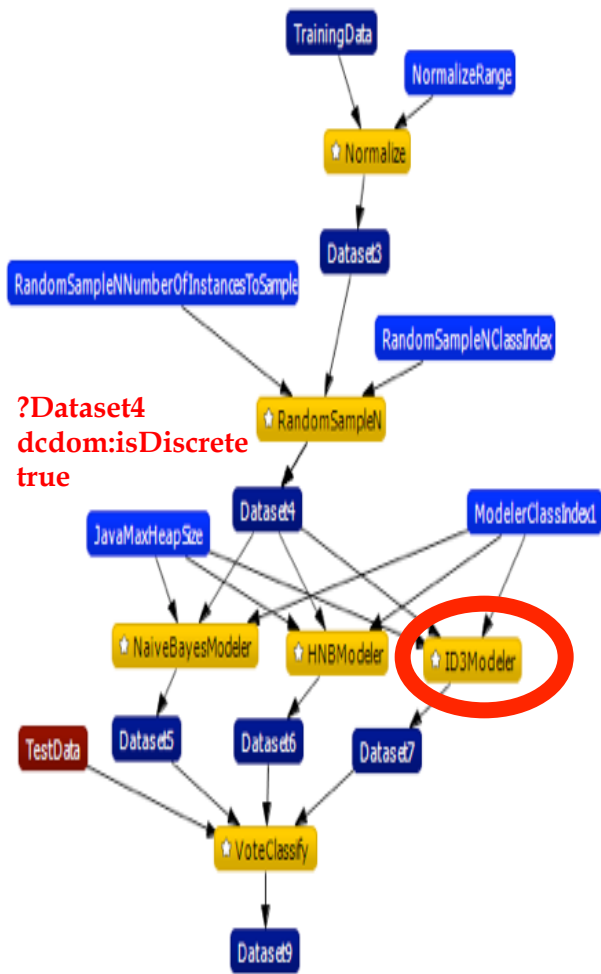


Example (Step 5 of 5)

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 ?va1l)]

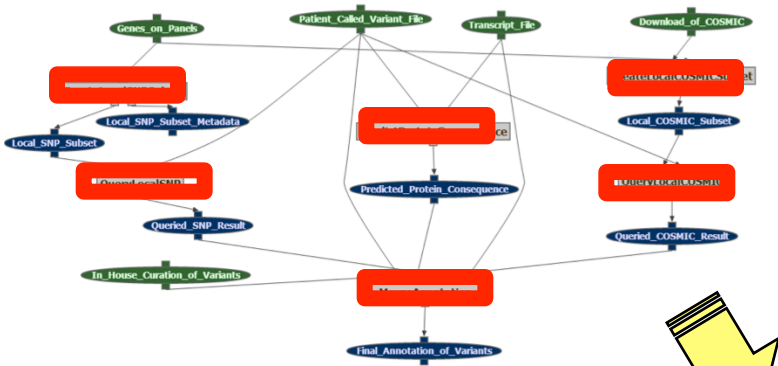


WINGS Workflow Reasoners: Result

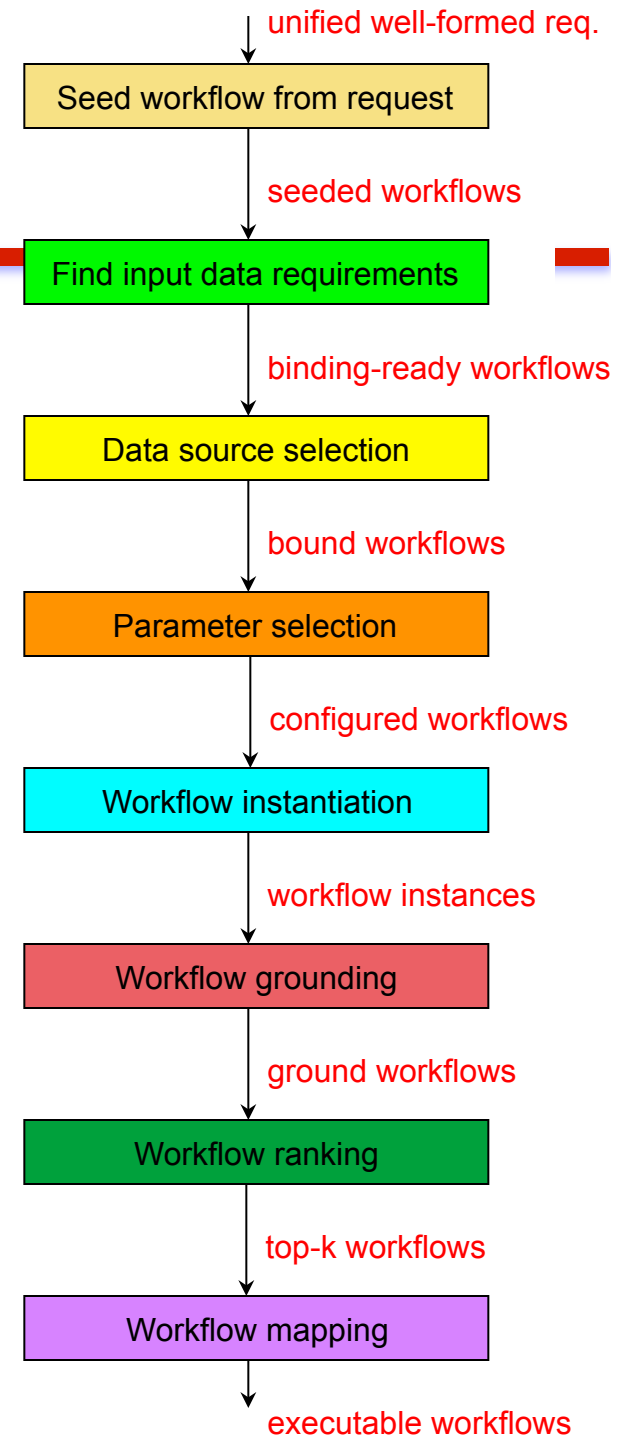
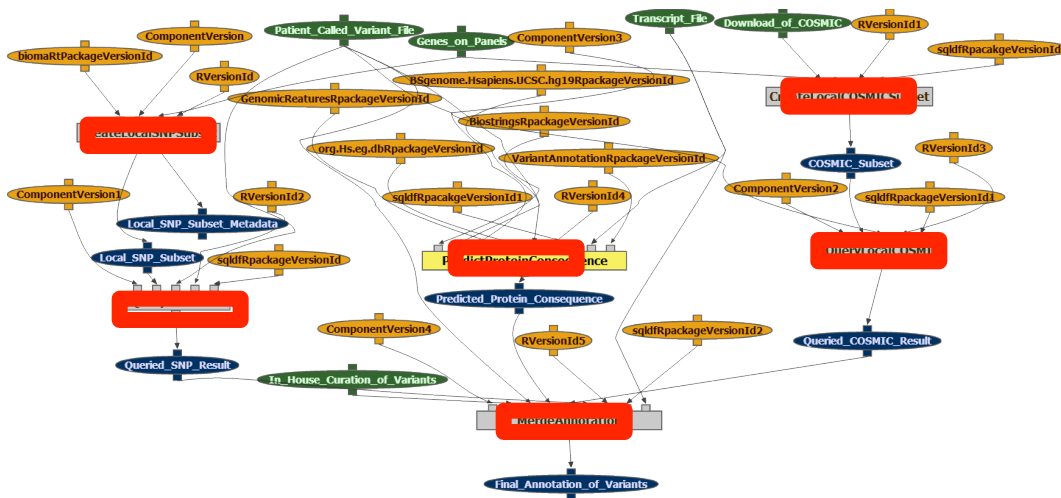


WINGS Automatic Workflow Generation Algorithm [Gil et al JETAI 2011]

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



“Pay-as-you-go” semantics



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 displays the TurboSoft Portal interface. The top navigation bar includes 'Home', 'Software', 'File Types', and 'Community'. The main content area is divided into two panes. The left pane shows a hierarchical tree of software components, with 'PIHM' selected. The right pane shows the 'Describe Software' interface for the 'PIHM' component, featuring tabs for 'I/O', 'Assumptions', 'Standard Names', 'Summary', and 'Con'. The 'I/O' tab is active, showing a table of inputs and outputs.

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

CSDMS Standard Names [Peckham iEMSs 2014]

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



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

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

CF CSDMS

Standard Names

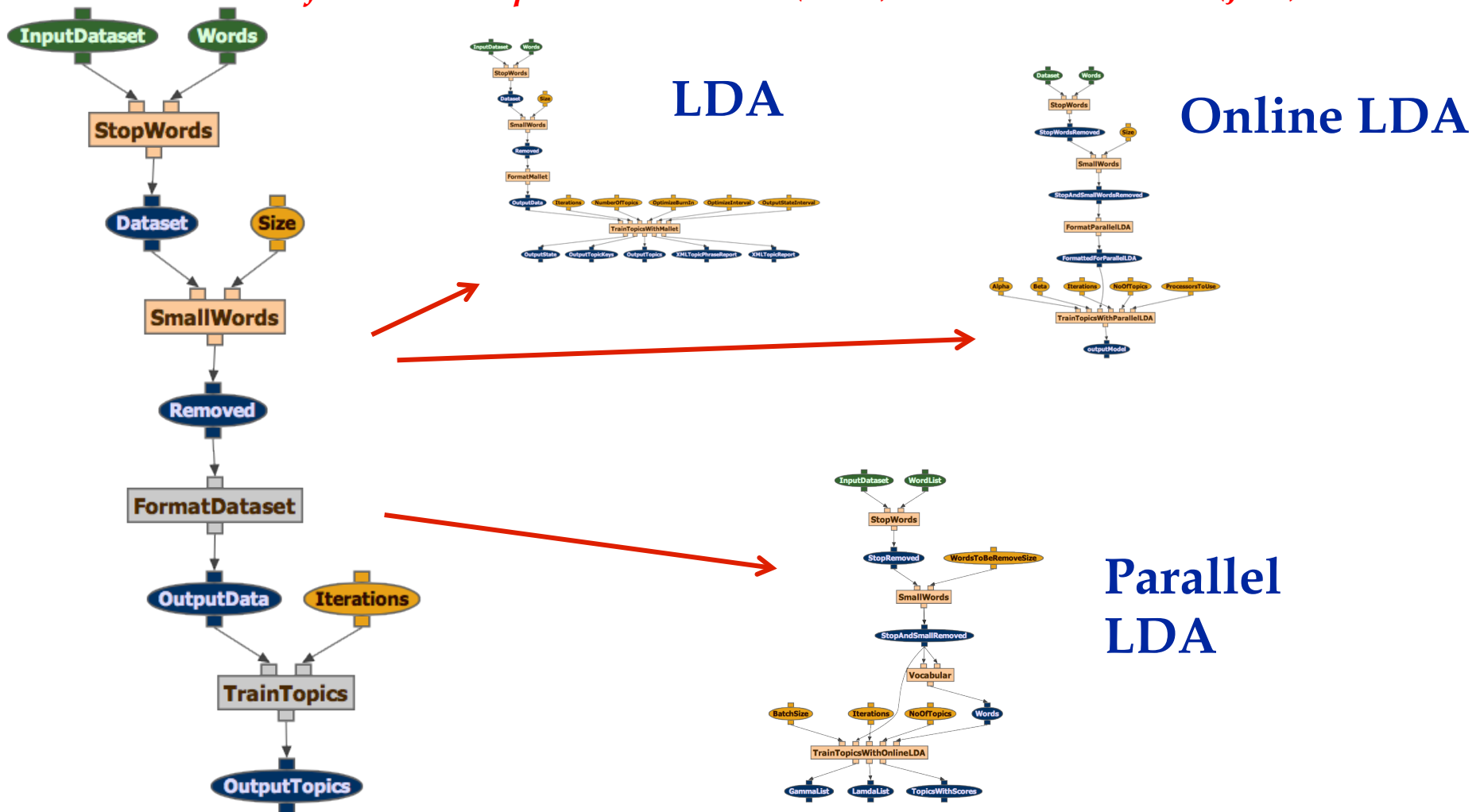
+ Add Standard Name - Delete

| <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 | <div style="border: 1px solid red; padding: 2px;">None</div> albedo aspect_angle backscattered_shortwave_irradiation_flu brutsaert_emissivity_canopy_factor brutsaert_emissivity_cloud_factor | |

Benefits of Semantic Workflows:

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

Workflows developed with Y. Liu (USC) and C. Mattmann (JPL)



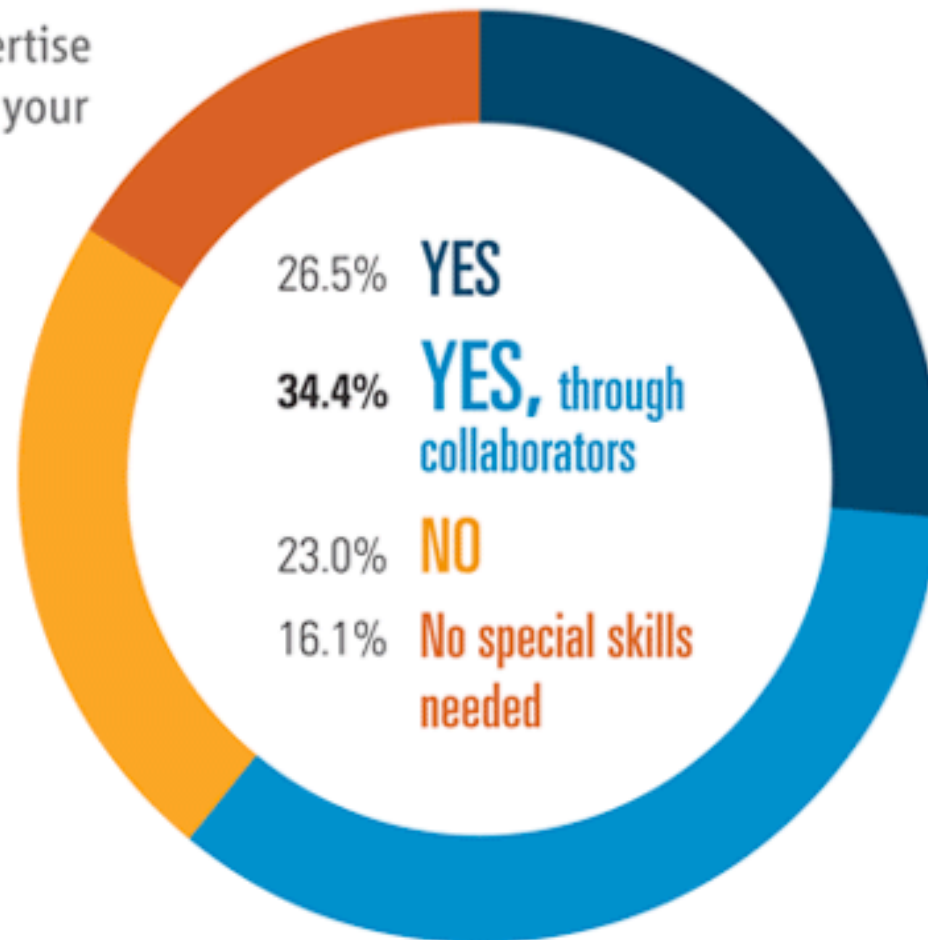
Benefits of Semantic Workflows:

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.”

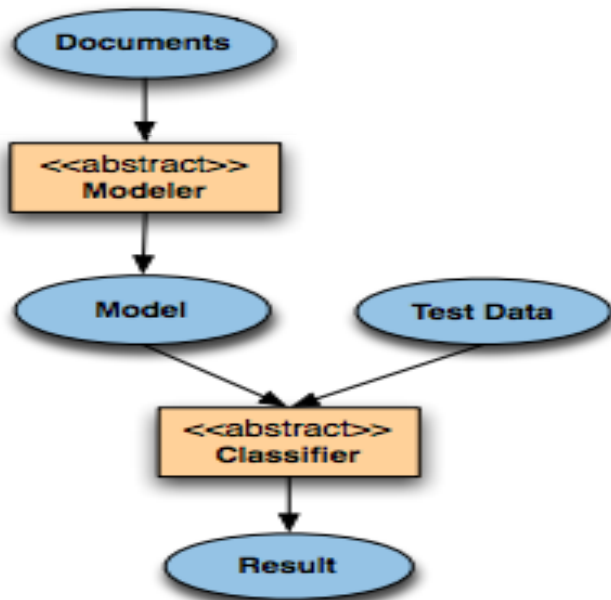


Capturing Expertise through Workflows

[Hauder et al e-Science 2011]

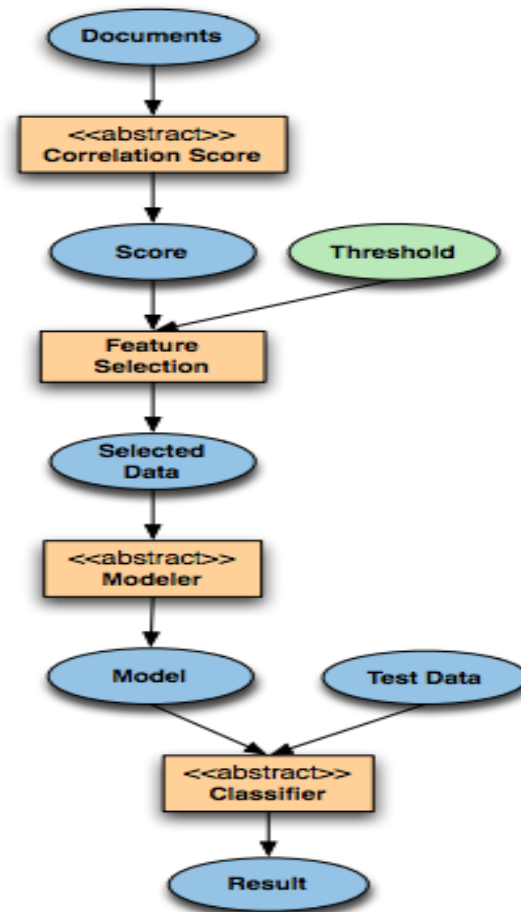
Workflows for text analytics, joint work with Yan Liu (USC) and Mattheus Hauder (TUM)

Naïve Approach



Expert Approach

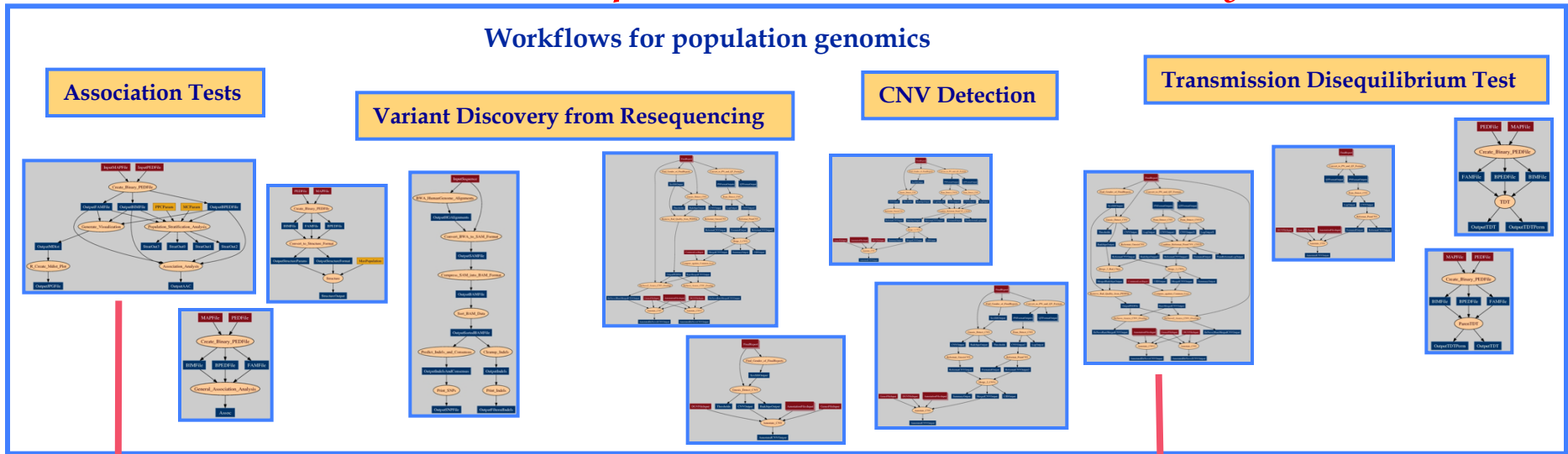
Feature selection



Capturing Expertise [Gil et al 2012]

Work with Christopher Mason (Cornell University)

Workflows for population genomics

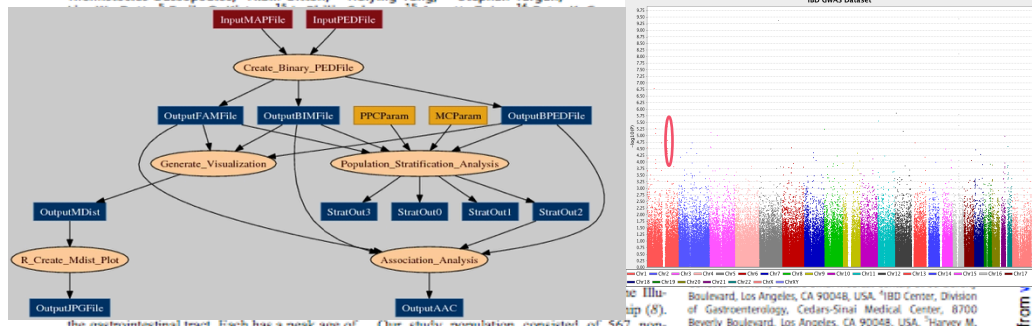


A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

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,^{2,10} A. Hillary Steinhart,⁴ Clara Abraham,¹¹ Miguel Regueiro,¹ Anne Griffiths,¹² Themistocles Dassopoulos,¹ Alain Bitton,¹³ Huiying Yang,^{3,4} Stephan Targan,^{3,14}

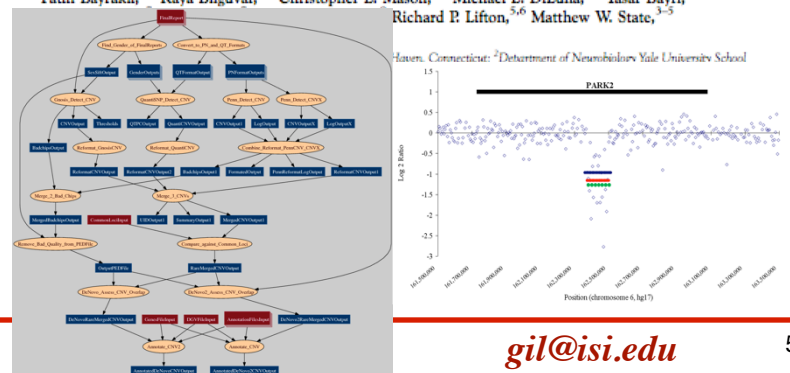
didate. In addition to Arg381Gln, nine other markers in *IL23R* and in the intergenic region between *IL23R* and the adjacent IL-12 receptor, beta-2 gene (*IL12RB2*), had association *P*-values < 0.0001 in the non-Jewish, ileal CD case-control cohort (Table 1 and table S1a).

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-



Rapid Identification of Disease-Causing Mutations Using Copy Number Analysis Within Linkage Intervals

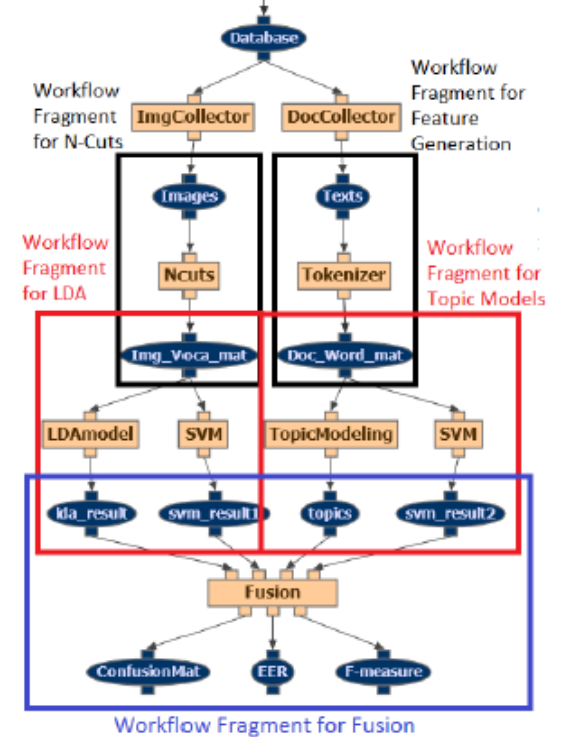
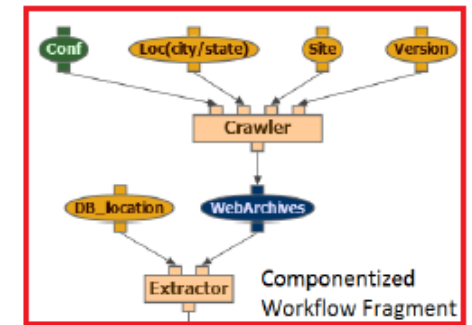
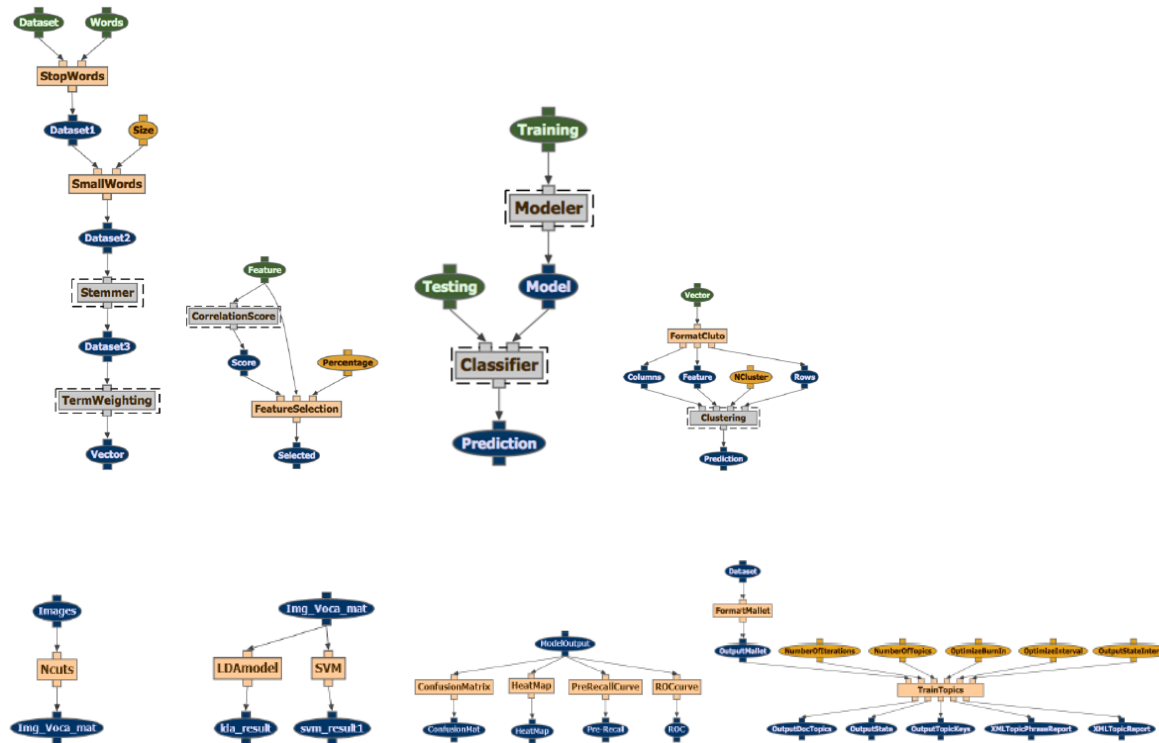
Fatih Bayrakli,^{1,2} Kaya Bilguvar,^{1,3} Christopher E. Mason,^{3,5} Michael L. DiLuna,^{1,3} Yasar Bayri,^{1,3} Richard P. Lifton,^{5,6} Matthew W. State,^{3,5}



Benefits of Semantic Workflows:

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

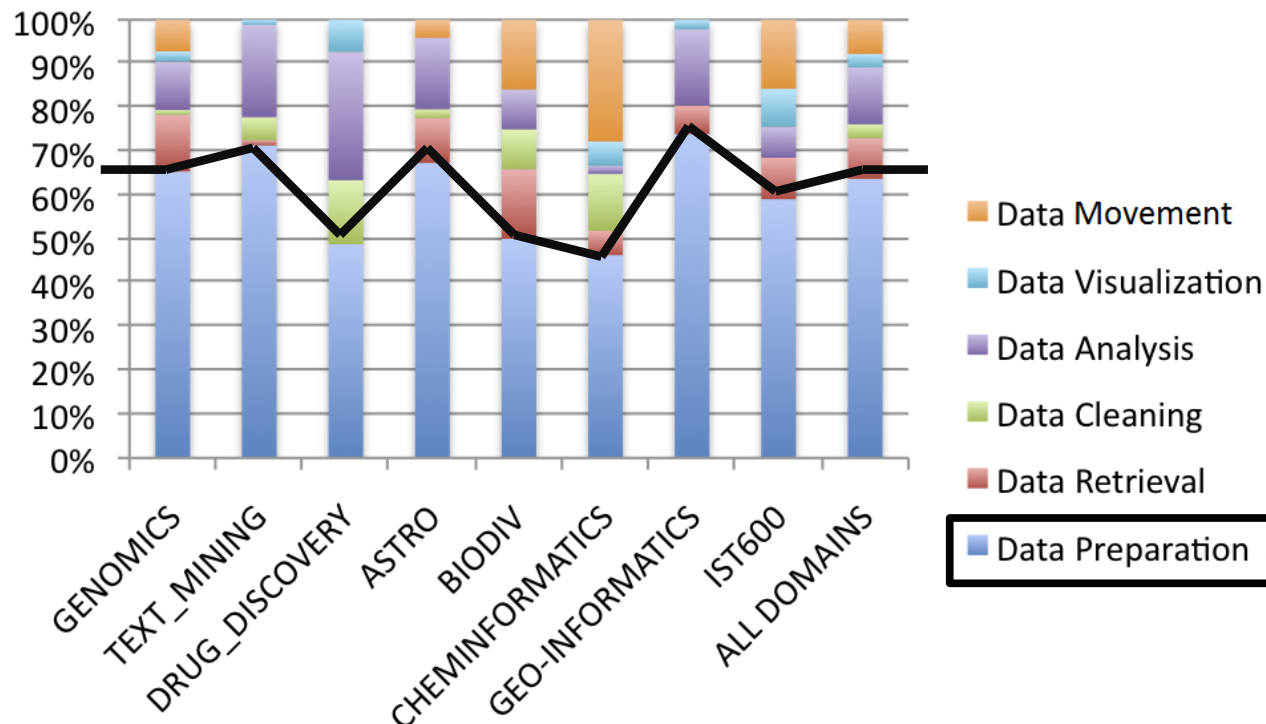
Work with Ricky Sethi and Hyujoon Jo of USC



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)

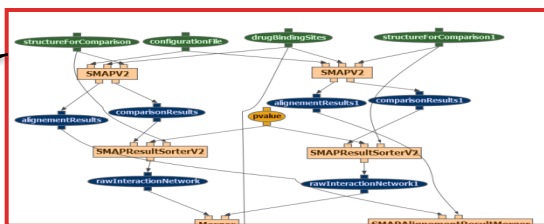


Measuring Time Savings with “Reproducibility Maps” [Garijo et al PLOS CB12]

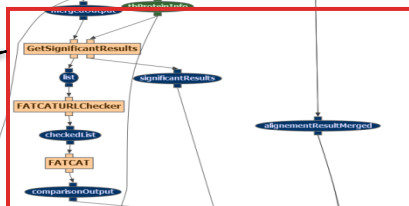
Work with D. Garijo of UPM and P. Bourne of UCSD

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

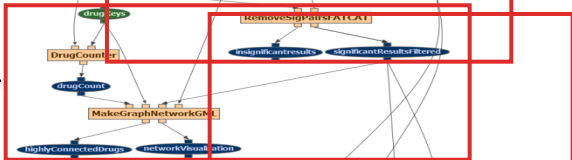
Comparison of ligand binding sites



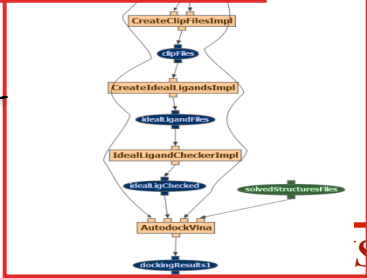
Comparison of dissimilar protein structures



Graph network generation



Molecular Docking



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 |

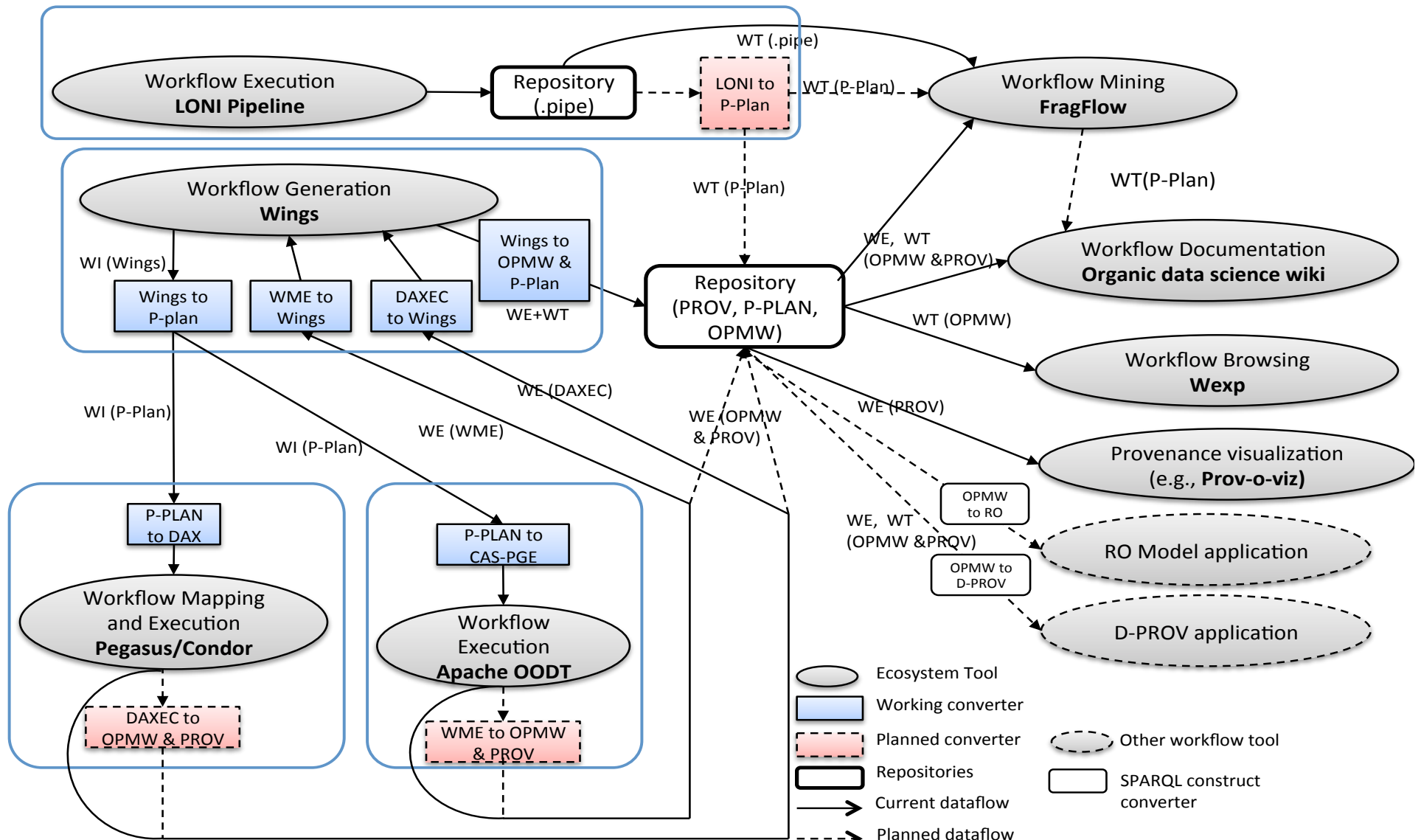
Docking

| | | | | |
|------------------|---------------------|---------------------|---------------|---------|
| CreateClip Files | CreateIdeal Ligands | IdealLigand Checker | Autodock Vina | Minimal |
| CreateClip Files | CreateIdeal Ligands | IdealLigand Checker | Autodock Vina | Novice |
| CreateClip Files | CreateIdeal Ligands | IdealLigand Checker | Autodock Vina | Author |

Benefits of Semantic Workflows:

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

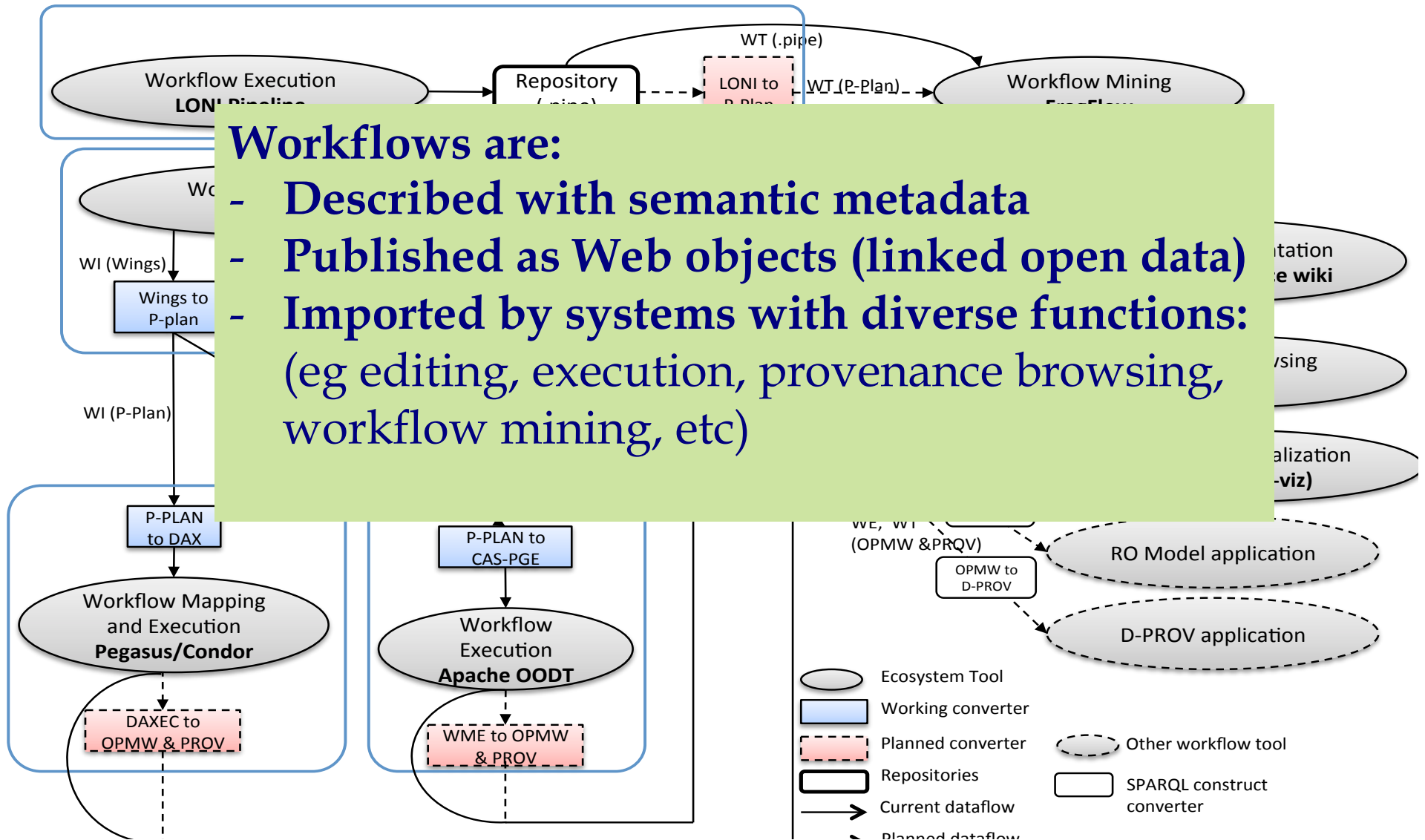
Work with D. Garijo and O. Corcho of UPM



Benefits of Semantic Workflows:

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

Work with D. Garijo and O. Corcho of UPM



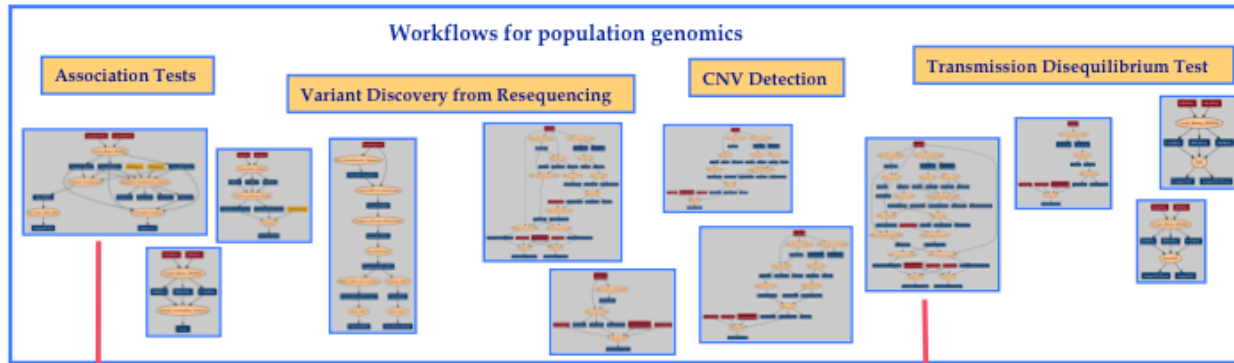
Some Readings

- 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

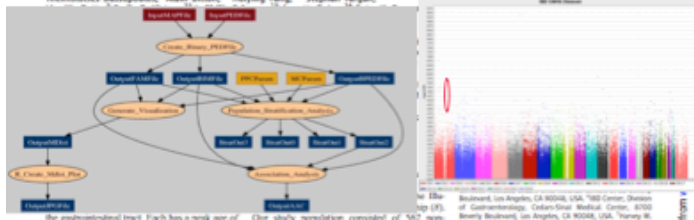


A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr,^{1,2} Kent D. Taylor,^{1,2} Steven K. Brant,^{1,2} John D. Rioux,^{1,2} Mark S. Silverberg,³ Mark J. Daly,^{1,2,4} A. Hillary Steinhart,⁵ Clara Abraham,^{1,2} Miguel Regueiro,⁶ Anne Griffiths,⁷ Theodoros D'Amico,⁸ Hideo Hirota,⁹ Huiying Tang,^{1,2} Stephan Taylor,^{1,2}

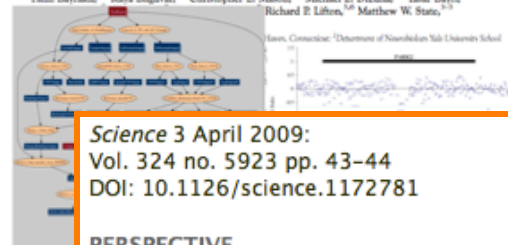
diabetes. In addition to Arg381Gln, nine other markers in *IL23R* and in the isogenic region between *IL23R* and the adjacent *IL-12* receptor, beta-2 gene (*IL23RA*), had association *P*-values < 0.0001 in the non-Jewish, IBD CD case-control cohort (Table 1 and table S1a).

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



Rapid Identification of Disease-Causing Mutations Using Copy Number Analysis Within Linkage Intervals

Fatih Beyrali,^{1,2} Kaveh Bilgiver,^{1,3} Christopher E. Mason,^{1,3} Michael L. DLane,^{1,3} Yasar Bayar,^{1,3} Richard E. Lifton,^{1,3} Matthew W. State,^{1,3}



Science 3 April 2009:
Vol. 324 no. 5923 pp. 43-44
DOI: 10.1126/science.1172781

PERSPECTIVE
COMPUTER SCIENCE
Automating Science

David Waltz¹, Bruce G. Buchanan²

± Author Affiliations

Computers with intelligence can design and run experiments, but learning from the results to generate subsequent experiments requires even more intelligence.



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

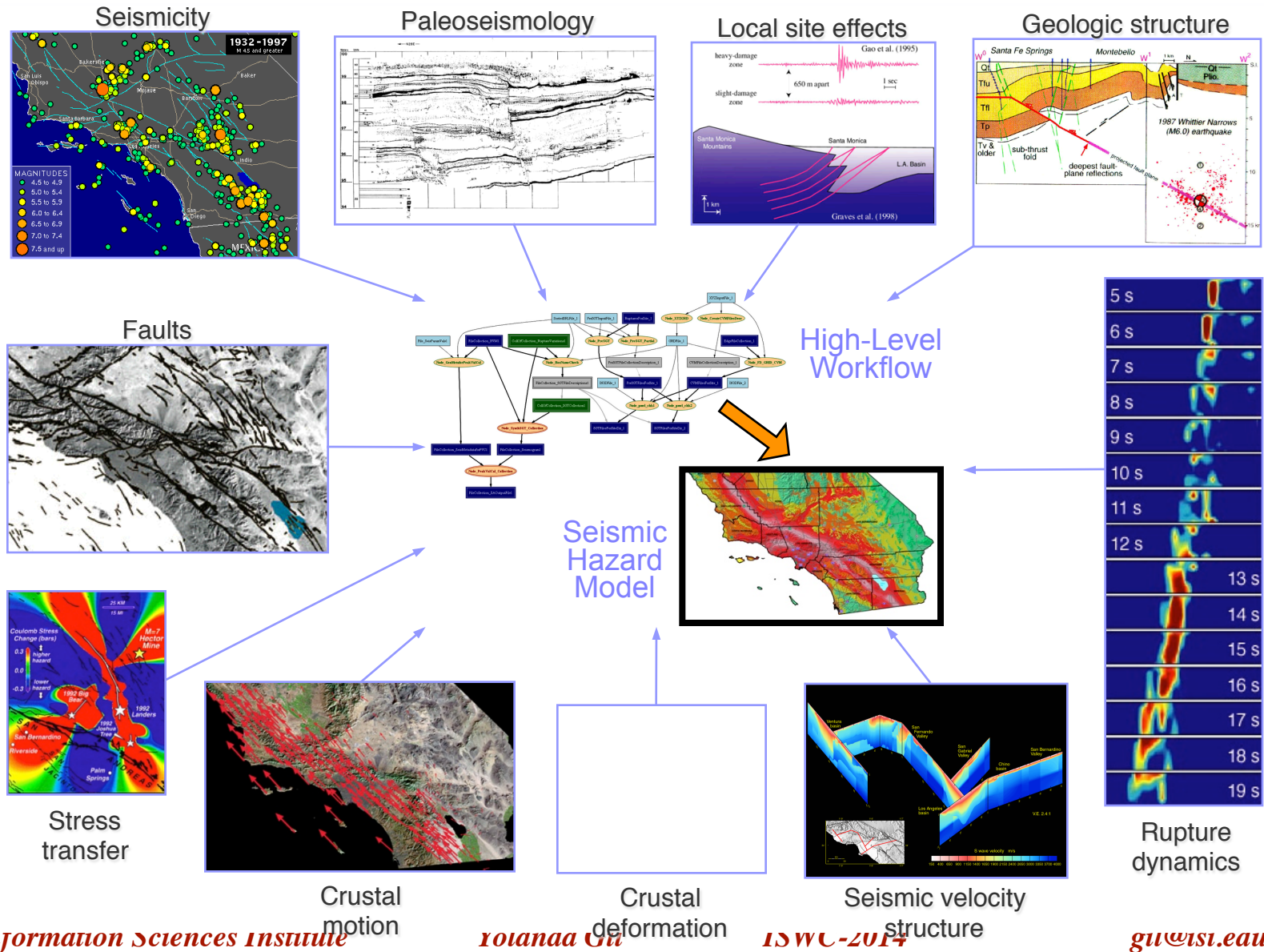
■ Knowledge rich tasks in science

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

■ Open science

Collaboration to Develop Workflows [Gil et al 2007]

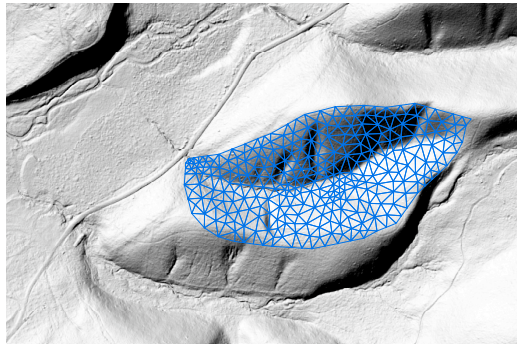
Slide from T. Jordan of USC and SCEC



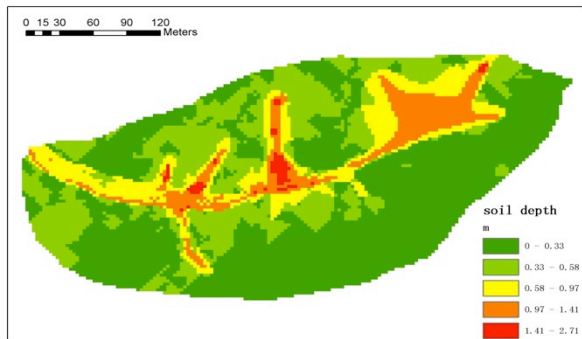
Understanding the “Age of Water”

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

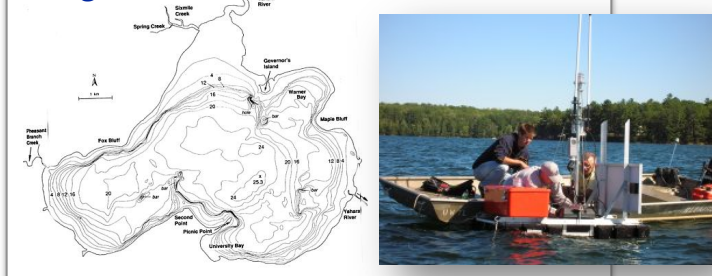
Lidar-derived numerical mesh



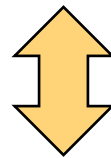
Mapping Bedrock GP Radar



High-resolution sensor network data

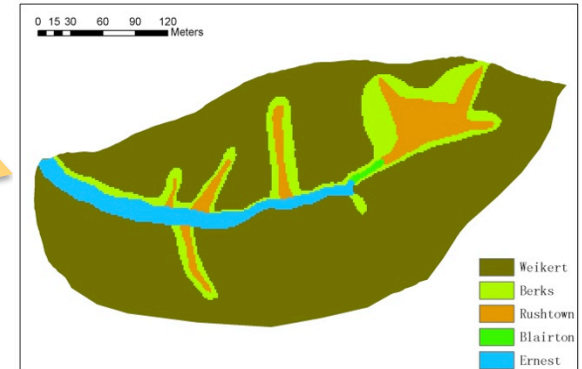


Linking Catchment
Model-Data Assets
Supported by NSF GEO-CZO

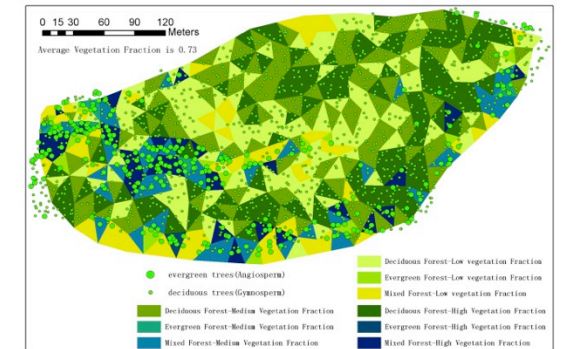


Linking Lake Model-Data
Assets Supported by
NSF BIO-GLEON, USGS CIDA

Research Soil Survey



High Resolution Vegetation Mapping



Models of lake hydrodynamics and water quality



Read JS, et al. 2014. Ecological Modelling. 291C: 142-150. doi:10.1016/j.ecolmodel.2014.07.029

A New Kind of Collaborative Platform

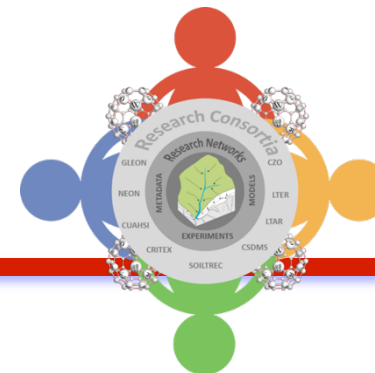
■ Taxonomy of Science Communities [Bos et al 2007]

| | |
|-------------------------------------|------------|
| 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

0 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 and phosphorus data

Obtain initial phytoplankton data

Gather stream flow data for North Temperate Lakes

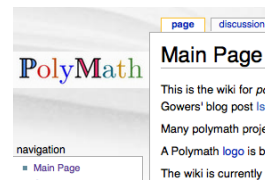
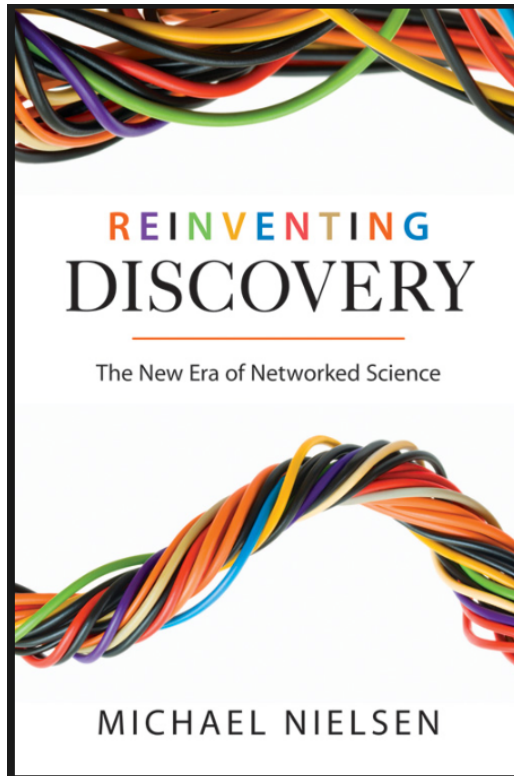
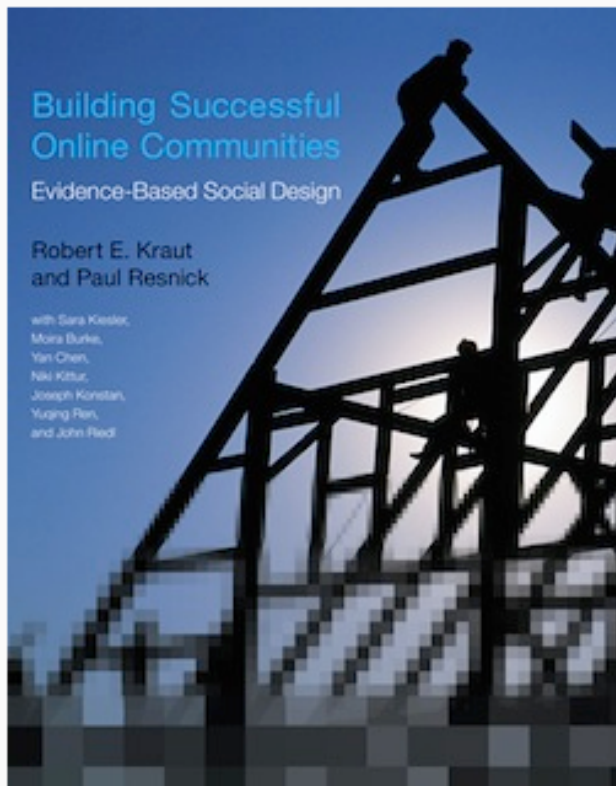
Collect daily volumetric stream flow data

Obtain water temperature data for North Temperate Lakes

Obtain or estimate nutrient and carbon concentration data for each major inflow

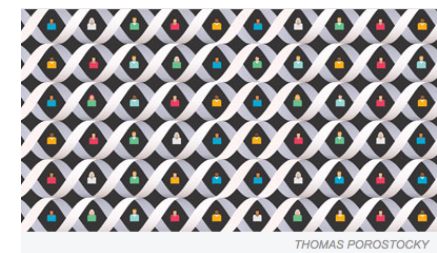
- Many tasks involved
- Necessary data resides in different repositories
- Different people understand different kinds of data
 - Where it is
 - How to use it
- If other data needed, unclear who has it

Social Principles for Online Communities



The polymath blog

January 26, 2014
Two polymath (of a sort) proposed projects
Filed under: discussion, polymath projects | 101 views | 0 0 0
Tags: Conjecture, polymath-projects, Riemann hypothesis
This post is meant to propose and discuss a polymath project and
I. A polymath proposal: Conjectures.
One of the interesting questions regarding the polymath endeavor
is "Can polymath be used to develop a theory/new area?"
My idea is to have a project devoted to developing a theory of "com-
plexity" in a finite set of points in a well-developed area of math-
ematics (not much, I suppose that for such a project the first discus-
sion should be more general.)
In general (but perhaps more so for an open-ended project), I'm
looking for interesting areas that have been explored, and that are
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 mathematical projects. [massively collaborative mathematics possible?](#)

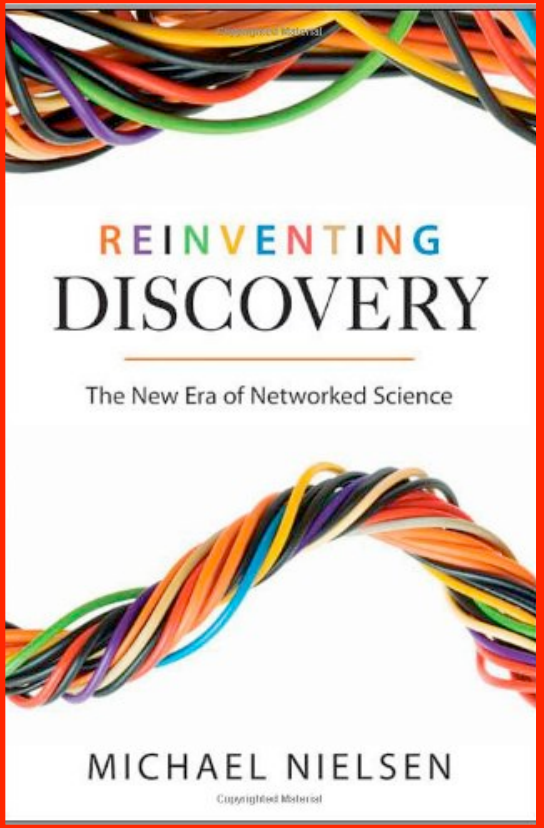
Many polymath projects will be proposed, planned, and run at [This Blog](#).

A Polymath logo is being trialled. If you have more suggestions, please add them to the talk page.

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

Existing polymath projects

- **Polymath1**: New proofs and bounds for the density Hales-Jewett theorem. I
- **Polymath2**: Must an "explicitly defined" Banach space contain c_0 or l_p ? Initial
- **Mini-polymath1**: Solving Problem 6 of the 2009 International Mathematical Olympiad.
- **Polymath3**: The polynomial Hirsch conjecture. Proposed July 17, 2009; launched
- **Polymath4**: A deterministic way to find primes. Proposed July 27, 2009; launched
- **Polymath5**: The Erdős discrepancy problem. Proposed Jan 10, 2010; launched
- **Mini-polymath2**: Solving Problem 5 the 2010 International Mathematical Olympiad. Proposed
- **Polymath6**: Improving the bounds for Roth's theorem. Proposed Feb 5, 2011.
- **Mini-polymath3**: Solving a problem from the 2011 International Mathematical Olympiad.
- **Mini-polymath4**: Solving a problem from the 2012 International Mathematical Olympiad.
- **Polymath7**: Establishing the Hot Spots conjecture for acute-angle triangles.
- **Polymath8**: Improving the bounds for small gaps between primes.




**REINVENTING
DISCOVERY**

The New Era of Networked Science

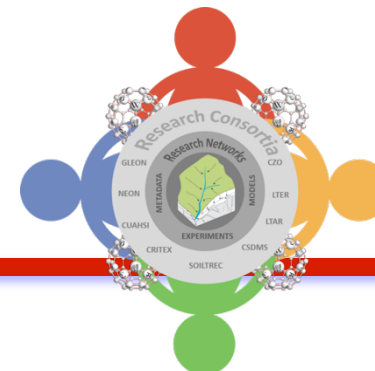
MICHAEL NIELSEN

Copyrighted Material



Tricki
a repository of mathematical know-how

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 displays a web-based task management system. At the top, a navigation bar includes a user profile for 'John' with options for 'Talk', 'Preferences', 'Watchlist', 'Contributions', and 'Log out'. A search bar is also present. The main content area is titled 'Your Overdue Tasks' and shows a task 'Write about the evaluation' that is 'a day ago' overdue. Below this, a task titled 'Draft paper about the initial framework design' is highlighted. This task is decomposed into several sub-tasks, each with a progress bar: 'Develop paper outline' (100%), 'Draft initial versions of key sections' (26%), 'Assemble first full draft of the paper' (0%), 'Collect final evaluation data' (0%), and 'Finalize writing the paper' (75%). A vertical red line indicates 'TODAY'. A detailed view of the 'Draft paper about the initial framework design' task is shown in a pop-up window, listing its sub-tasks: 'Develop paper outline', 'Draft initial versions of key sections', 'Assemble first full draft of the paper', 'Collect final evaluation data', 'Review first full draft of the paper', and 'Finalize writing the paper'. The task properties are listed below: Type (medium), Progress (21%), Start date (22nd Aug 2014), Target date (13th Oct 2014), Owner (John Smith), Participants (James Williams, Steven Johnson), and Expertise (computer science, collaboration). A legend explains the progress bar colors: Not defined (grey), Valid (green), and Inconsistent with parent (yellow). A text box explains the plan: '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.' The interface also shows a sidebar with a task tree, a search bar, and a context menu for the selected task.

1 Page Discussion

7 Your Overdue Tasks

5 Write about the evaluation
a day ago

6 Timeline SubTasks

5

10 Draft paper about the initial framework design

21

Develop paper outline 100%

Draft initial versions of key sections 26%

Assemble first full draft of the paper 0%

Collect final evaluation data 0%

Finalize writing the paper 75%

TODAY

2a

21 Draft paper about the

Develop paper outline

Draft initial versions of key sections

Assemble first full draft of the paper

Collect final evaluation data

Review first full draft of the paper

Finalize writing the paper

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

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

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.

2b

Properties

Add

[x] Submitted to UII-2015 (by John)

3 All Tasks My Tasks 36

4

computer sci search

Organic Data Science

Develop Framework for Organic Data Science

Framework Design

Disseminate results from the Organi

Write paper about the initial fra

Train new users to exercis

Design user evaluation of

Collect data about feature

Draft paper about the initia

Cut

Paste

Rename


Delete

To Toplevel

8











SMW
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

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 | 0.0000999999997473788 | (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_201:
- AF_EDM_Execution_2_March_201:

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

Organic Data Science

All Tasks

Expertise filter search

- 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 states
- 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

Page **Discussion** Read View source View history

Train Hilary on using organic data science wiki

100 Train Hilary on contributing as owner

- Train Hilary on creating tasks
- Train Hilary on using task alert
- Train Hilary on organizing tasks
- Train Hilary on understanding extended task states

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**

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?

Organic Data Science

All Tasks | My Tasks (29) | Filter | Search

- Develop Framework for Organic Data Science
- Select core lake and catchment models
- Develop mathematical model of age of water
- Implement the lake model for North Temperate
- Document the GLM lake model
- Couple the lake and catchment models
- Develop a computational model for water
- Engage broader community
- G16 Workshop**
- Set up first advisory committee meeting
- Document the PIHM catchment model
- Implement the catchment model for North

Main page | Recent changes | Random page

Tools: What links here | Related changes | Upload file | Special pages

Page: Discussion | Read | Edit | View history | Search

Engage broader community

G16 Workshop

Materials

| | |
|--------------------------|---------------|
| Type ^M | high |
| Progress ^M | 0% |
| Start date ^M | 18th Jul 2014 |
| Target date ^M | 24th Oct 2014 |
| Owner ^M | Paul Hanson |
| Participants | Chris Duffy |
| Expertise | Not defined! |

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

Workshop: Modeling the Age of Water (The first steps toward the age of carbon)

Sunday, October 26th 2014
Jovance – The meeting venue for GLEON 16

The purpose of this workshop is to give participants first hand experience with lake catchment hydrology modeling, as well as lake and reservoir hydrodynamic modeling, while at the same time learning the principles underlying the use of stable isotopes to age water and identify its flow paths. These are the initial steps toward using integrated catchment-lake models to identify source and flow paths of both water and carbon through lake and reservoir catchments. This work is conducted within the context of "organic team science", which is an

Organic Data Science

Page: Discussion | Read | Edit | View history | Search

Develop Framework for Organic Data Science

Organic Data Science

Develop mathematical model of age of water

Implement the lake model for North Temperate

Document the GLM lake model

Couple the lake and catchment models

Develop a computational model for water

Engage broader community

Document the PIHM catchment model

Implement the catchment model for North

Main page | Recent changes | Random page

Tools: What links here | Related changes | Upload file | Special pages

Framework Design

Develop framework results from the Organic Data Science framework

Type^M high

Progress^M 37%

Start date^M 1st Oct 2013

Target date^M 30th Oct 2013

Owner^M

Participants

Expertise

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

A Platform for Organic Data Science

Our goal is to allow scientists to formulate and resolve science processes through an open framework that facilitates ad-hoc participation and entice collaborators based on attractive science goals. We are pursuing a social computing approach that takes into account human aspects such as incentives and participation, while providing the fabric for representing and coordinating tasks involved in accomplishing science goals. Our approach will openly expose science tasks, facilitating inspection and engagement of new potential contributors. The collaboration will grow in an organic way, drawing in people and other contributions from existing data providers and cyberinfrastructure resources.

Organic Data Science

Page: Discussion | Read | Edit | View history | Search

Main Page

Contents [hide]

- 1 What is Organic Data Science?
- 2 Our Science Goal: The Age of Water and Carbon
- 3 Technical and Social Aspects of Organic Data Science
- 3.1 Ongoing Technical Activities
- 4 Contributing to this project
- 5 Acknowledgments

Highest Contributors

- 936 Edits
- 643 Edits
- 1925 Edits
- 407 Edits
- 154 Edits
- 139 Edits
- 102 Edits
- 95 Edits
- 23 Edits
- 15 Edits

What is Organic Data Science?

We are investigating Organic Data Science, a new approach aimed to allow scientists to formulate and resolve science processes through an open framework that facilitates ad-hoc participation and entice collaborators based on attractive science goals. Organic Data Science allows scientists to formulate and resolve science processes through an open framework that facilitates ad-hoc participation and entice collaborators based on attractive science goals.

Accomplishing this requires three elements:

1. a scientific approach to tackle a problem of the age of water, will facilitate ad-hoc collaborations in science.
2. a technical approach that facilitates interdisciplinary collaborations, and
3. a social approach to engage the community.

Read more about Organic Data Science.

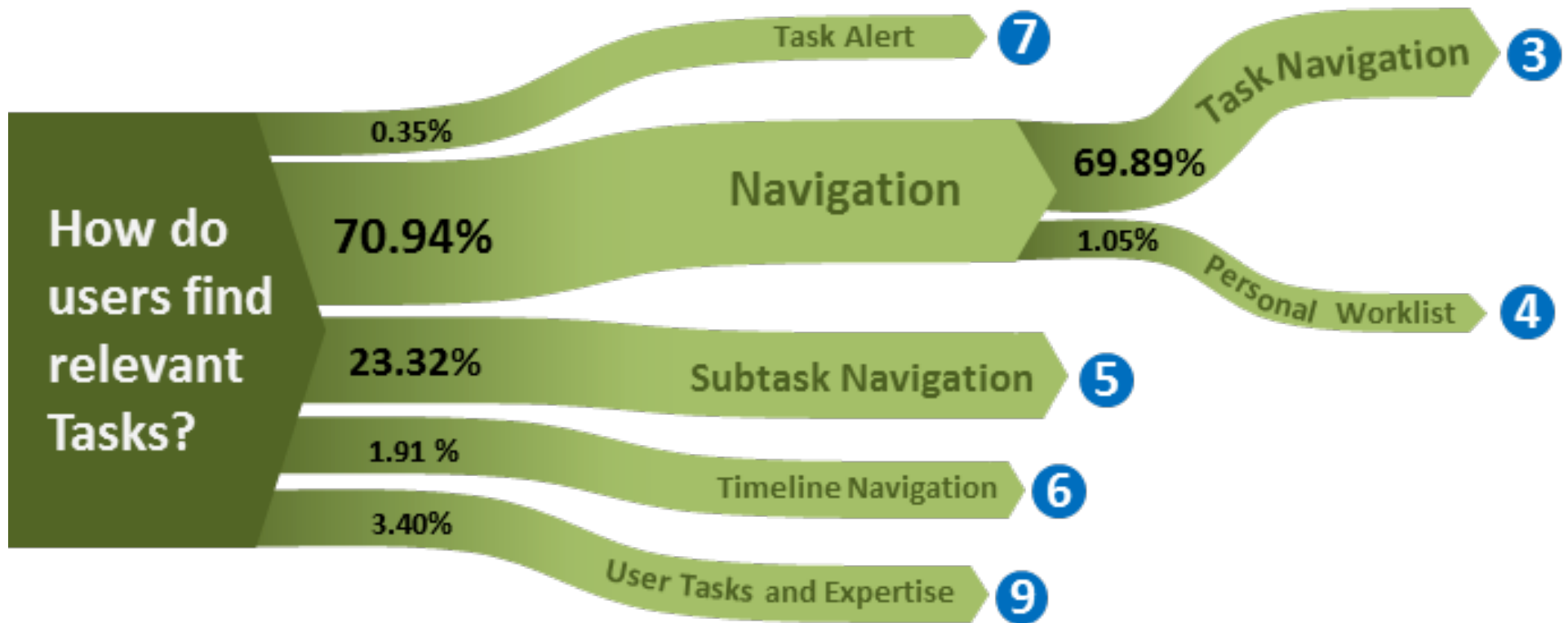
Our Science Goal: The Age of Water and Carbon

Organic Data Science

Main page | Recent changes | Random page

Tools: What links here | Related changes | Upload file | Special pages | Printable version

How Do Users Find Relevant Tasks?

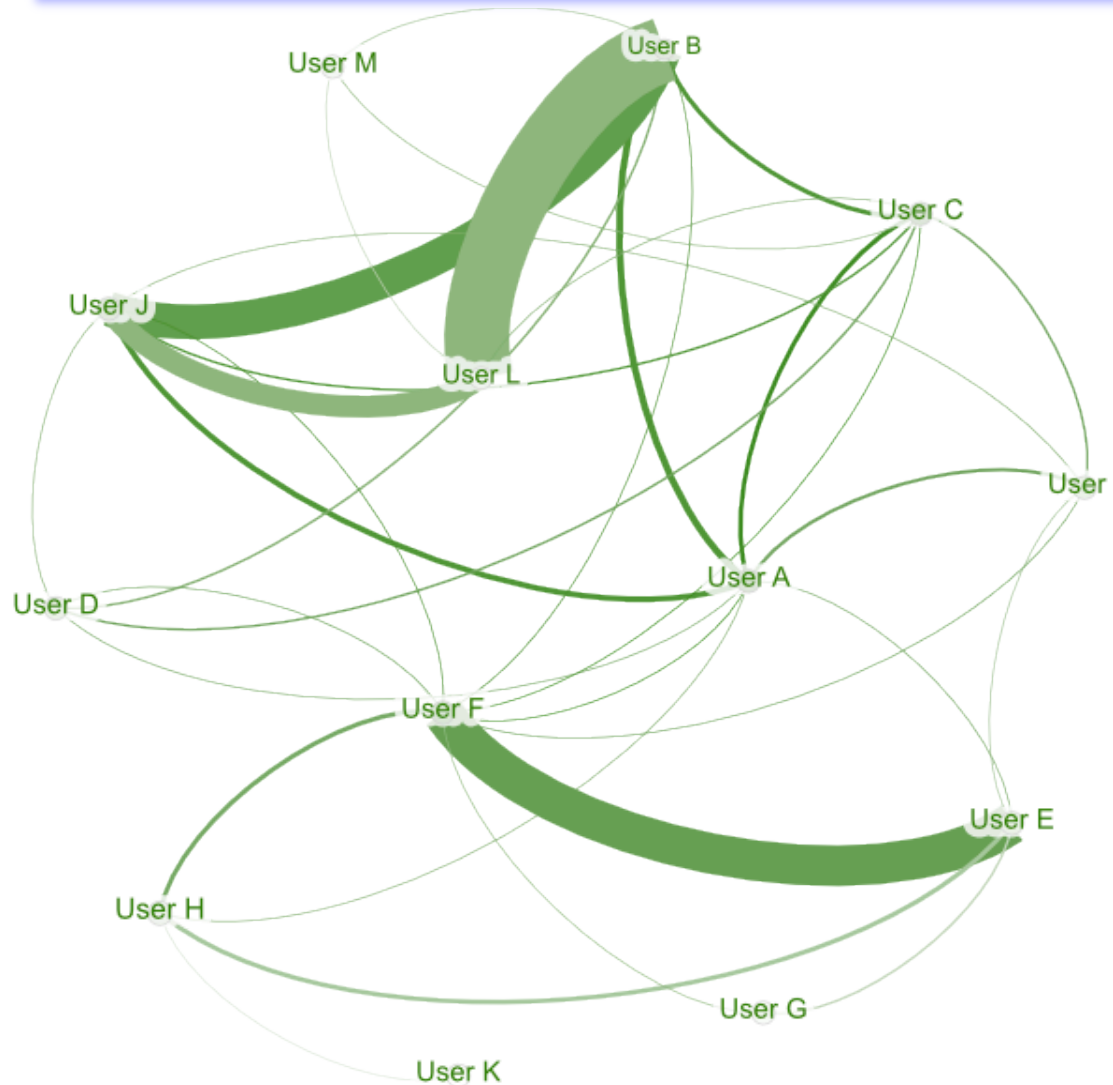


Are Users Collaborating?



- 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

34 Develop Framework for Organic Data Science

- Human Centered Computing to Support Organic Data Science
 - Develop an approach to contributor credits in organic data science
 - Study emerging normalization and standardization
 - Develop an approach to social computing for collective problem solving
 - Develop social principles to guide collective problem solving
 - Develop mechanisms to attract newcomers
- Framework Design
 - Develop task-centered features
 - Comparison with Other Collaborative Platforms
 - User Tutorial
 - User Documentation
- Disseminate results from the Organic Data Science framework
 - Write paper about the initial framework design
 - Train new users to exercise initial framework
 - Design user evaluation of initial framework design
 - Collect data about feature use in the new framework
 - Draft paper about the initial framework design
 - Develop paper outline
 - Draft initial versions of key sections
 - Assemble first full draft of the paper
 - Collect final evaluation data
 - Review first full draft of the paper
 - Finalize writing the paper
 - Motivating Science Scenario
 - Present framework at Science in the Northwoods

Yolanda Gil

AI planning and collaborative problem solving 0 Workflows 0 Semantic Web 0 Semantic wikis 0
Social computing 7

Current Tasks 11

- 20 Develop an approach to contributor credits in organic data science
Will be completed in 2 months
- 10 Human Centered Computing to Support Organic Data Science
Will be completed in a year
- 34 Develop Framework for Organic Data Science
Will be completed in 2 years
- 78 Develop social principles to guide collective problem solving
Will be completed in 10 months
- 78 Develop mechanisms to attract newcomers
Will be completed in a year
- 78 Write paper about the initial framework design
Will be completed in 3 days
- 52 Disseminate results from the Organic Data Science framework
Will be completed in a year
- 97 Draft paper about the initial framework design
Will be completed in 3 days
- 75 Write paper: Dynamic age of water and carbon
Will be completed in 22 days
- 80 Set up first advisory com meeting
Will be completed in a month
- 97 Finalize writing the paper
Will be completed in 3 days

Future Tasks 3

- 20 Outline
Start not defined yet!
- 40 Develop an approach to social computing for collective problem solving
Start not defined yet!
- 0 Study emerging normalization and standardization
Starts in 3 months

Completed Tasks 29

- 5 Collect data about feature use in the new framework
Completed since a few seconds ago
- 100 Assemble first full draft of the paper
Completed since 8 days ago
- 99 Draft initial versions of key sections
Completed since 9 days ago
- 95 Write related work
Completed since 9 days ago
- 100 Write about framework design
Completed since 10 days ago
- 100 Write about the evaluation
Completed since 11 days ago

A Semantic Challenge: Open Science Processes

Palmyra coral 20C

Data

- DOWNLOAD
- From: <http://www.nodc.noaa.gov/palco/metadata/nodc-coral-1985.html>

Structured Properties

- SiteName: Palmyra (By Julien)
- Archive: Coral (By Julien)
- Domain(s): Climate geochemistry (By Julien)
- Forward model: 10.1029/2011.GL048224 (By Julien)
- Genus: Porites (By Julien)
- Interpretation: SST,SST (By Nick)
- Measurement: Delta18O (By Julien)
- MeasurementMaterial: Skeletal aragonite (By Julien)
- MeasurementStandard: VPDB (By Nick)
- MeasurementUnits: Permil (By Julien)
- Reference: Cobb et al, 2001 (By Julien)
- Species: Lutea (By Julien)

Credits

Users who have contributed to this Page:

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

Chemical Limnology of Primary Study Lakes: Major Ion

Facts

- Property: is-a: Dataset
- Property: includes: Lake Mendota, Lake Monona
- Property: External link: <http://ter.limnology.wisc.edu/dataset/north-temperate-lakes-iter-chemical-limnology-primary-study-lakes-major-ions>
- Property: Number of sites: 11
- Property: Download site: <http://ter.limnology.wisc.edu/data/filter/5691>

Callouts:

- Metadata properties created on the fly
- Datasets linked to locations
- Entities linked to linked data
- Projects linked to datasets
- Information sources are documented
- WWW
- Datasets linked to download sites

AQUAFLOW EDM

Workflow

- AQUAFLOW EDM

Contributor

- WATER
- wings.isi.edu

Workflow Template Image

Develop Framework for Organic Data Science

- Human Centered Computing to Support Organic Data Science
- Develop an approach to contributor credits in organic data science
- Study emerging normalization and standardization
- Develop an approach to social computing for collective problem solving
- Develop social principles to guide collective problem solving
- Develop mechanisms to attract newcomers
- Framework Design
 - Develop task-centered features
 - Comparison with Other Collaborative Platforms
 - User Tutorial
 - User Documentation
- Disseminate results from the Organic Data Science framework
 - Write paper about the initial framework design
 - Train new users to exercise initial framework
 - Design user evaluation of initial framework design
 - Collect data about feature use in the new framework
 - Draft paper about the initial framework design
 - Develop paper outline
 - Draft initial versions of key sections
 - Assemble first full draft of the paper
 - Collect final evaluation data
 - Review first full draft of the paper
 - Finalize writing the paper
 - Motivating Science Scenario
 - Present framework at Science in the Northwoods

Lake Mendota

Facts

- Property: Longitude: -89.425
- Property: AreaOfCatchment: Topic:562000000
- Property: is-a: Topic:Lake
- Property: Studied in project: IMMF
- Property: Latitude: 43.107
- Property: Temperature data: Topic:http://ter.limnology.wisc.edu/about/lakes/mendota_receiv

Credits

Users who have contributed to this Page:

- ar العربية
- de Deutsch
- en English
- fr Français
- hi हिन्दी
- it Italiano
- ja 日本語
- Log in

Integrated Management of Macrophytes and Fisheries

Facts

- Property: Project-PI: Topic:Stephen R. Carpenter
- Property: Dataset collected: Topic:Chemical Limnology of Primary Study Lakes: Major Ion
- Property: is-a: Topic:an NTL research project
- Property: External link: <http://ter.limnology.wisc.edu/project/integrated-management-macrophytes-and-fisheries>

Labels

- Property: 562.0
- Property: 39413535.54176
- Property: United States

Callouts:

- People linked to projects

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

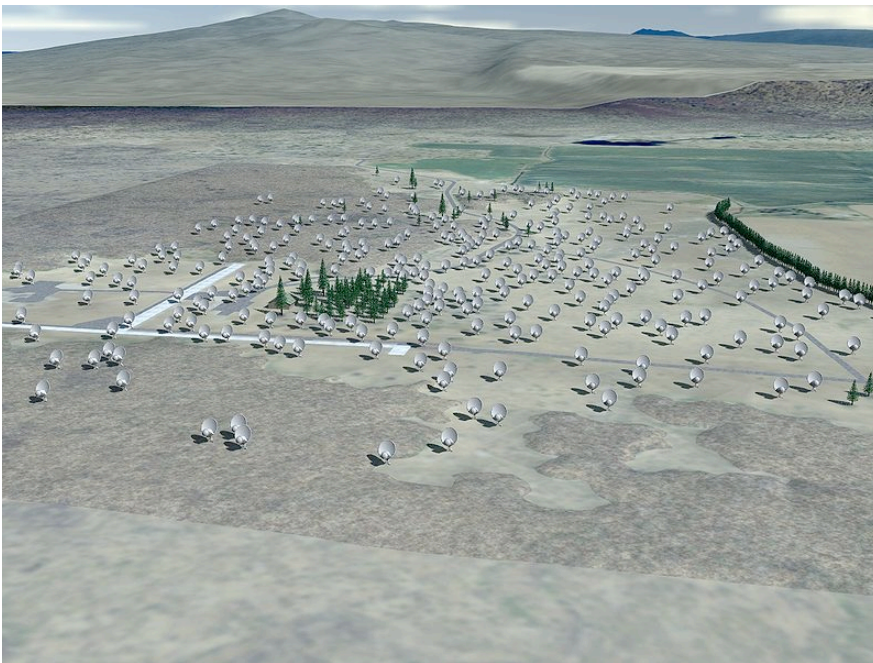
<http://www.isi.edu/~gil>

<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 sciencemag.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

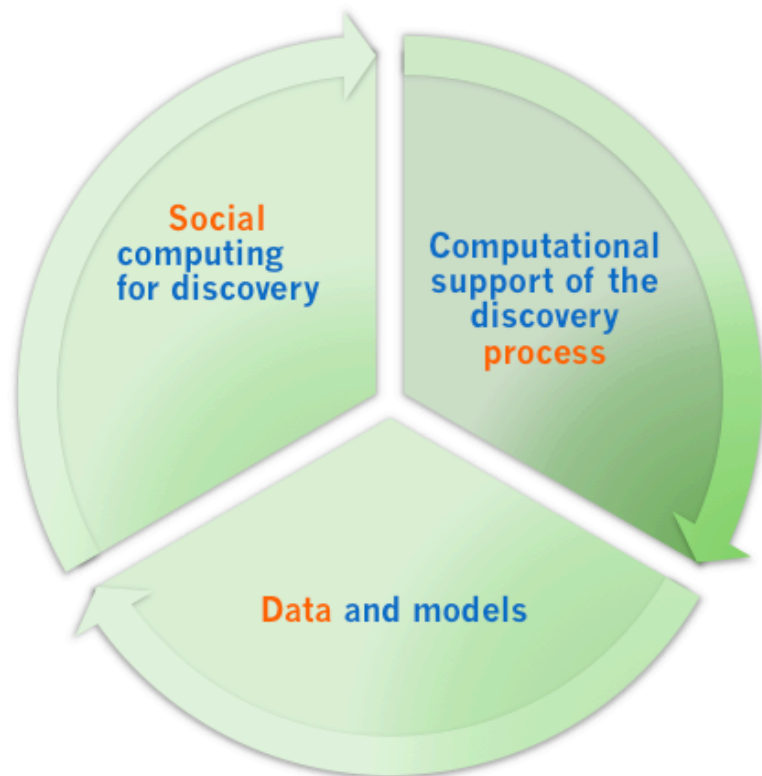
Technological innovations are penetrating all areas of science, making predominantly human activities a principal bottleneck in scientific progress while also making scientific advancement more subject to error and harder to reproduce. This is an area where a new generation of artificial intelligence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



GROUPS



Data

Workflows

Semantics

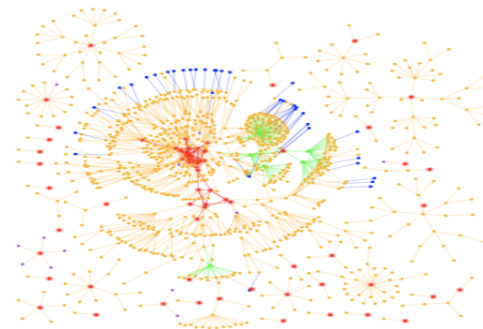
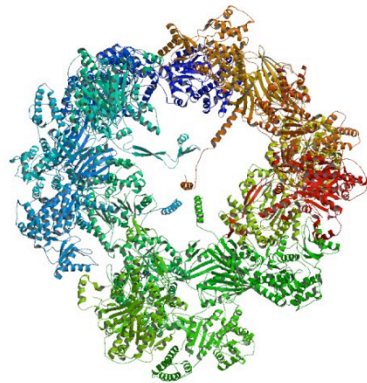
Governance

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

What Might the Future Look Like?

YOU: What are you working on?

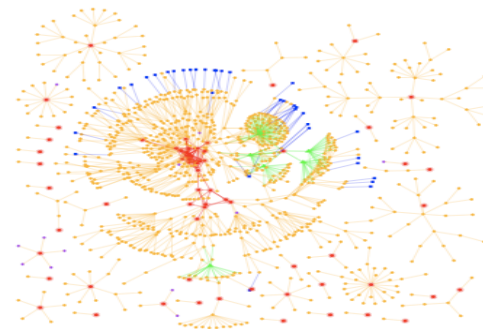
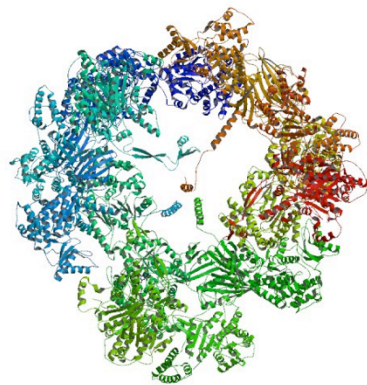
OTHER PERSON: I am really busy, working on...



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!*