

# The Human Bottleneck in Data Analytics: Opportunities for Cognitive Systems in Automating Scientific Discovery

Yolanda Gil



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

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

Keynote at the  
Third Annual Conference on Advances in Cognitive Systems,  
May 28-31, 2015, Atlanta GA



# Theme of this Talk: **Knowledge-Driven Science Infrastructure**

---

Data-intensive computing is producing major advances

Scientists are still responsible for major aspects of the science process themselves, becoming unmanageable

▶ Human bottleneck

Great opportunities for cognitive systems



# Outline

---

1. The human bottleneck in data analytics
2. Related work on AI and cognitive aspects of scientific discovery
3. Semantic workflows to capture data analytics processes
4. Meta-reasoning to automate discovery
5. Discovery Informatics

# Data-Intensive Computing in Science



Sign In | RSS F

## WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

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

By Chris Anderson 06.23.08

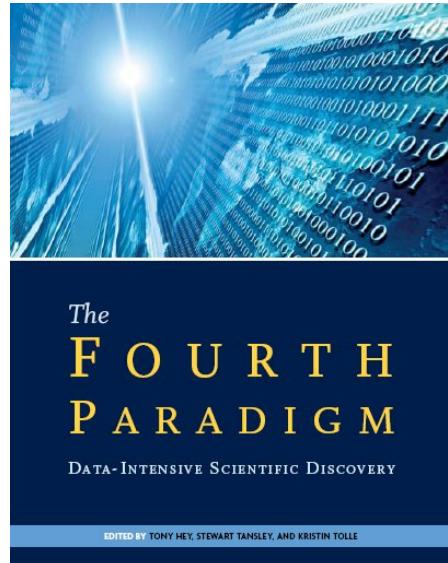


COMMUNICATIONS  
ACM

: wrong, but some are



stitute



Yolanda G.



## BIBLIOMETRICS AND CITATION ANALYSIS

From the Science Citation Index to Cybermetrics



gil@isi.edu

# Scientific Data Analysis

## Complex processes involving a variety of algorithms/software

**plink...**  
Whole genome association analysis toolset

Introduction | Basics | Download | References | Meta-analysis | Result annotation | Clamp

1. Introduction  
2. Basic information  
• Citing PLINK  
• Reporting problems  
• What's new?  
• PDF documentation

3. Download and general notes  
• Stable download  
• Development code  
• General notes  
• MS-DOS Notes  
• Windows Notes  
• Compilation  
• Using the command line  
• Viewing outputs history  
• Version history

4. Command reference table  
• List of commands  
• List of options  
• Underlined

5. Basic usage guide  
• Running  
• PED files  
• MAP files  
• Transposing  
• Long-to-short  
• Alkenes  
• Covariates  
• Clusters  
• Statistics

6. Data management  
• Recode  
• Reorder  
• Write SNPs  
• Update  
• Merge

**PennCNV**

Home  
Download  
Installation  
Tutorial  
• Quick Examples  
Input Files

**Burrows-Wheeler Aligner**

Introduction

Burrows-Wheeler Aligner (BWA) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. It implements two algorithms, bwa-short and dBWT-SW. The former works for query sequences shorter than 200bp and the latter for longer sequences up to around 100kbp. Both algorithms do gapped alignment. They are usually more accurate and faster on queries with low error rates. Please see the [BWA manual page](#) for more information.

[Mailing List](#)

[number variation detection in whole-genome SNP genotyping platforms](#) *Nucleic Acids Research* 36:e126, 2008

• Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, *et al.* [number variation detection in whole-genome SNP genotyping platforms](#) *Nucleic Acids Research* 36:e126, 2008

• Wang K, Chen Z, Tadesse MG, Glessner J, Grant SFA, Hakonarson H, *et al.* [number variation detection in whole-genome SNP genotyping platforms](#) *Nucleic Acids Research* 36:e138, 2008

### PennCNV: copy number variation detection

Welcome! PennCNV is a free software tool for Copy Number Variation (CNV) detection on Illumina and Affymetrix arrays. With appropriate input files, PennCNV implements a hidden Markov model (HMM) based segmentation-based algorithm in that it considers family information to generate false positive CNV regions, through a validation-based approach.



## Software

### Structure 2.3.X

The program *structure* is a free software to investigate population structure. It uses multivariate analysis to populations, assigning individuals to populations, identifying individuals and mixed individuals, and estimating migration rates between populations.

**GenePattern**

Getting Started  
Home  
Features  
Analyses Overview  
Analyses Modules  
Download  
Documentation/Tutorials  
Resources  
About  
Contact  
Contact  
Google Custom Search  
Search

Citing GenePattern  
Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nature Genetics* 38 no. 5 (2006): pp500-501.

GenePattern is a powerful genomic analysis platform that provides gene expression analysis, proteomics, SNP analysis and more. Its web-based interface provides easy access to these tools and allows users to build their own analysis pipelines that enable reproducible *in silico* research.

Software distribution: Allegro

Allegro is a complete linkage analysis package. Its features include parametric and non-parametric LOD score calculations and various features intended for genetic mapping. The software is available for download from the Broad Institute website.

### Genetic Analysis Software

PubMed Entrez BLAST OMIM Taxonomy Structure

Downloads

FASTLINK package  
FASTLINK executables for Windows  
FASTLINK executables for Mac OS X

**Overview**

Genetic linkage analysis is a statistical technique used to map genes and find the approximate location of disease genes. There was a standard software package for genetic linkage called LINKAGE. FASTLINK is a significantly modified and improved version of the main programs of LINKAGE that runs much faster.

# Problems (I): Efficiency and Quality

## ■ High cost

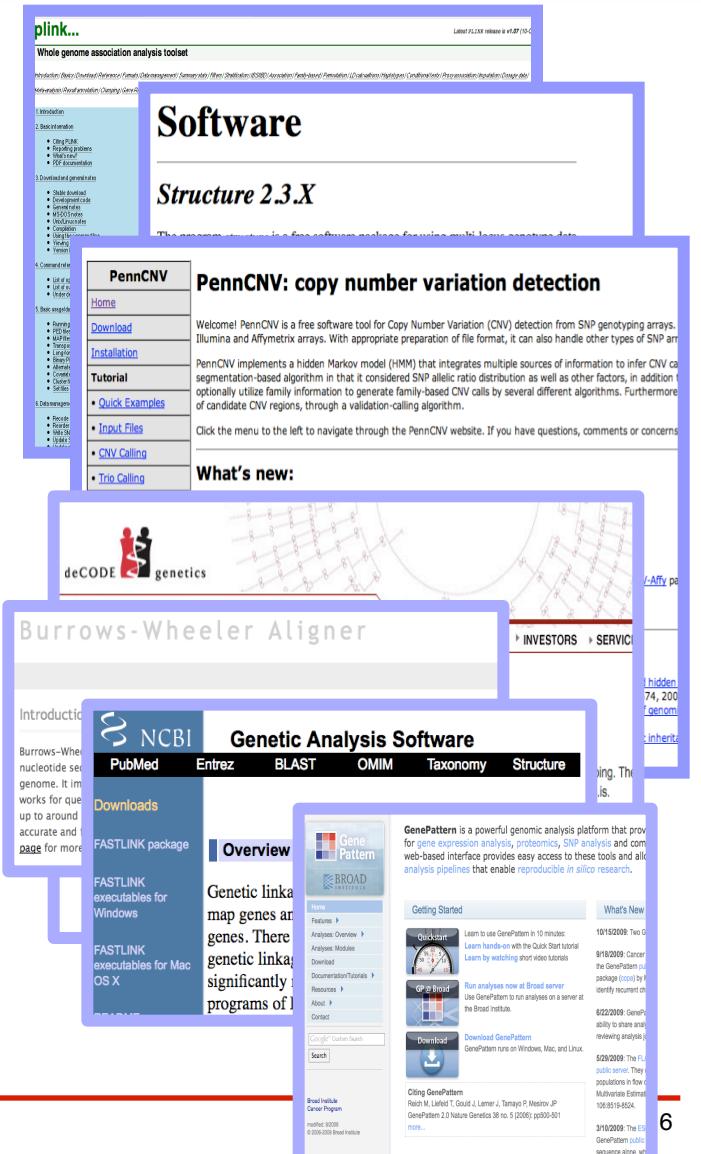
- “Scientists and engineers spend more than **60% of their time just preparing the data** for model input or data-model comparison” (NASA A40)

## ■ Quality concerns

- “We write QC code without thinking about the best way to do the WC. Such approaches perpetuate mediocrity. If someone did it right once, it would benefit many people.” (EC WF CQ)

## ■ Inefficiency

- “I often see that I’m repeating the work that 100 other people have been doing to obtain and process the data.” (EC WF CQ)



# Problems (II): Reproducibility

## Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

### Reporting Checklist For Life Sciences Articles

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more please read [Reporting Life Sciences Research](#).

**nature**

## A Biostatistic Paper Alleges Potential Harm To Patients In Two Duke Clinical Studies

By Paul Goldberg

## Human lives

exception.

A paper published in a journal that may be harmed by relying on biomarkers.

The paper

Friday, December 2, 2011 As of 12:00 AM New York 43° | 34°

THE WALL STREET JOURNAL | HEALTH

COMPUTER SCIENCE

## Accessible Reproducible

The New York Times

S  
rect. M  
tain a p  
edgeab  
to ope

Science

## Reliability

### Nobel Laureate Retracts Two Papers

R. KENNETH CHANG  
Retracted Scientific Studies: A Growing List

RETRACTED

CYFORUM

putation in research grows,  
needed to expand recording,

of I

Multipli

## Scientific integrity

新语丝

New Threads

新语丝

新语丝

新语丝

新语丝

新语丝

新语丝

USC Information Sciences Institute

Yolanda Gil

gil@isi.edu

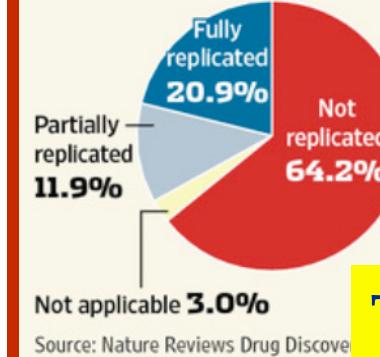
HEALTH INDUSTRY | DECEMBER 2, 2011

## Scientists' Elusive Goal: Reproducing Study Results

ing how it had halted nearly two-thirds of the experiments failed to match claims made

## No Cure

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.



## Financial



## Trust



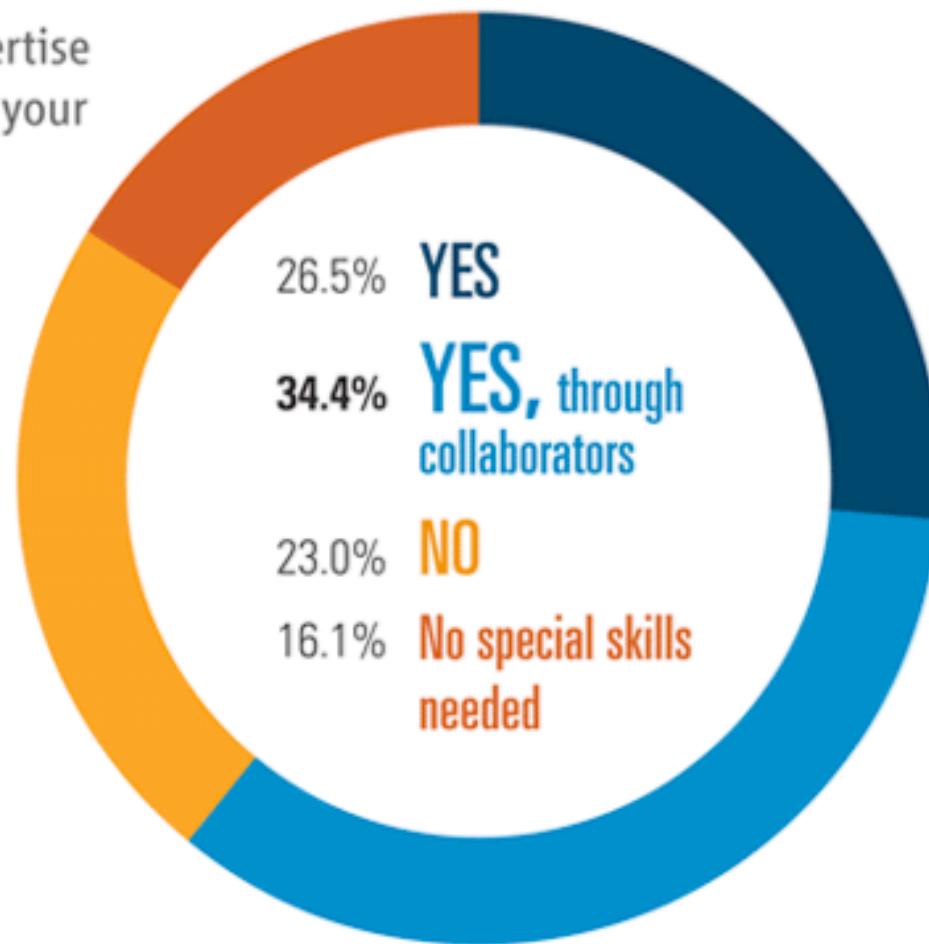
# Problems (III): Lack of Access to Data Analytics Expertise

---

*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.** ”

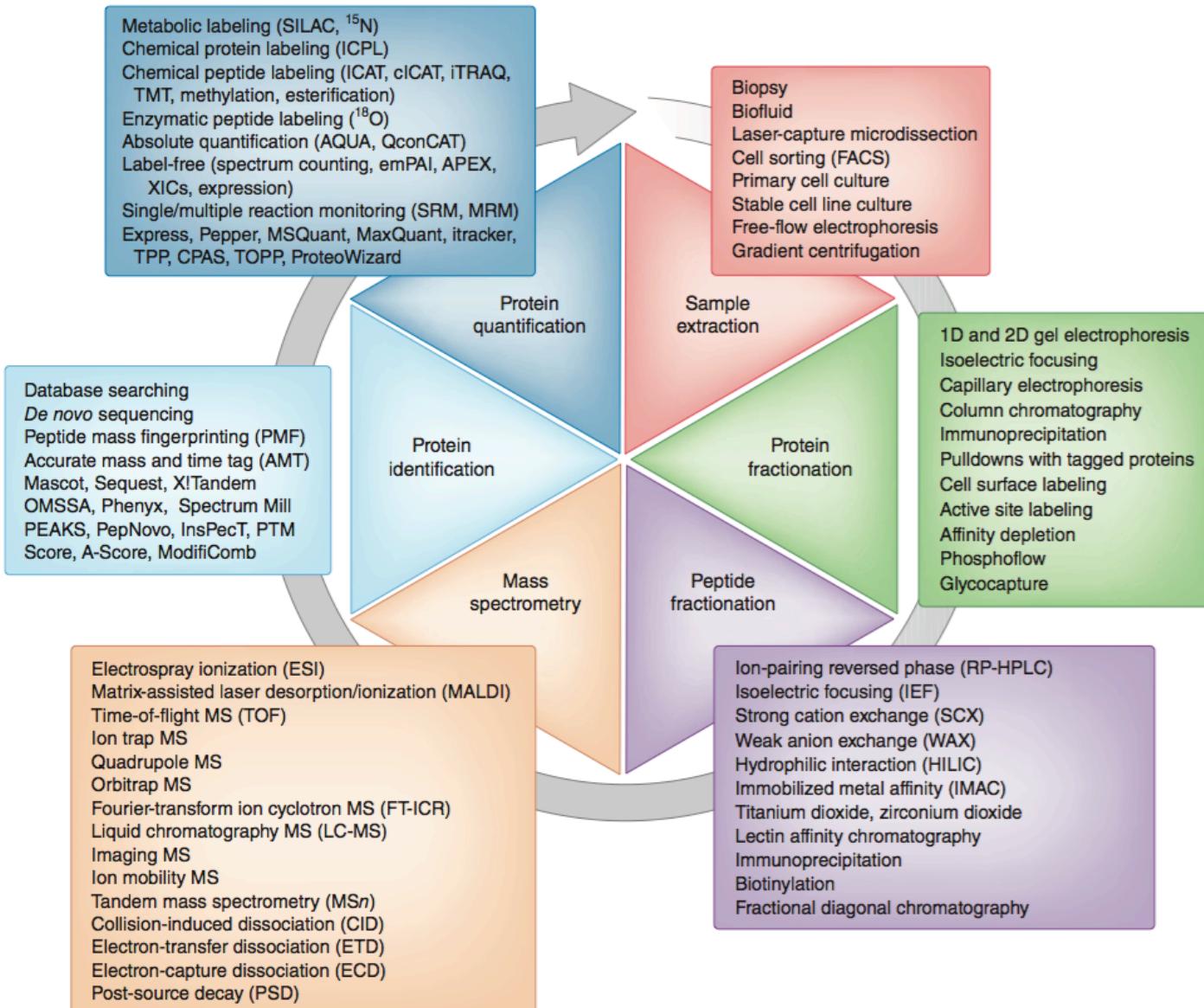


# Fragmentation of Expertise: An Example from Proteomics

Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. *Nat Biotechnol* 28, 695–709 (2010)

COMPUTATIONAL

EXPERIMENTAL



# 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
  - **Cognitive systems can be a game changer**

# Outline

---

1. The human bottleneck in data analytics
2. Related work on AI and cognitive aspects of scientific discovery
3. Semantic workflows to capture data analytics processes
4. Meta-reasoning to automate discovery
5. Discovery Informatics

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



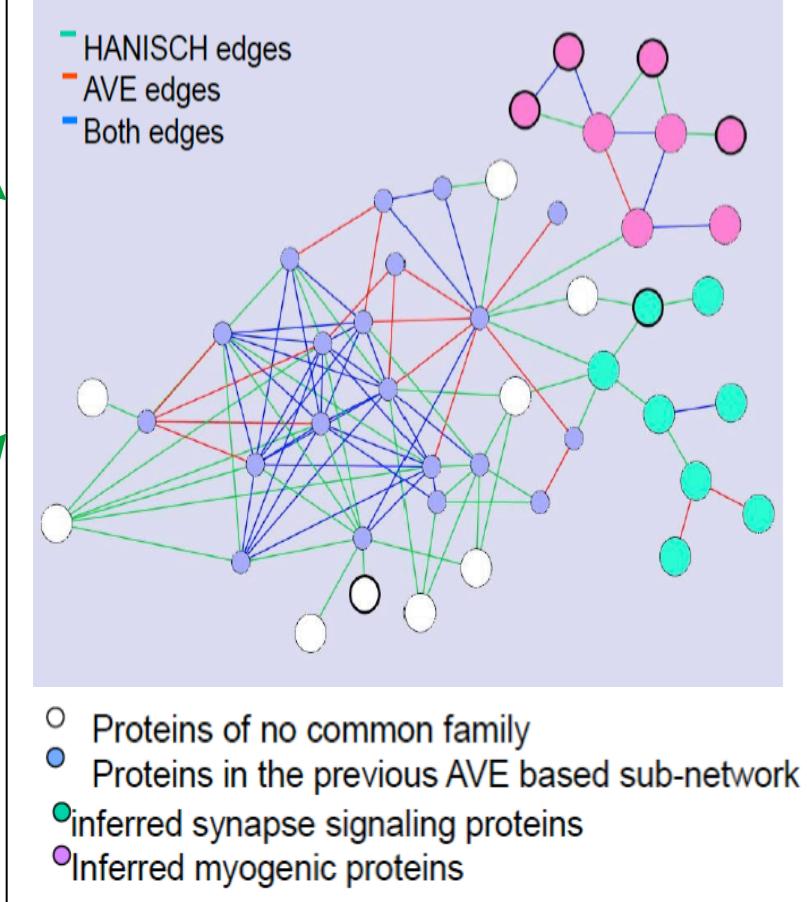
## Text extraction from publications

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

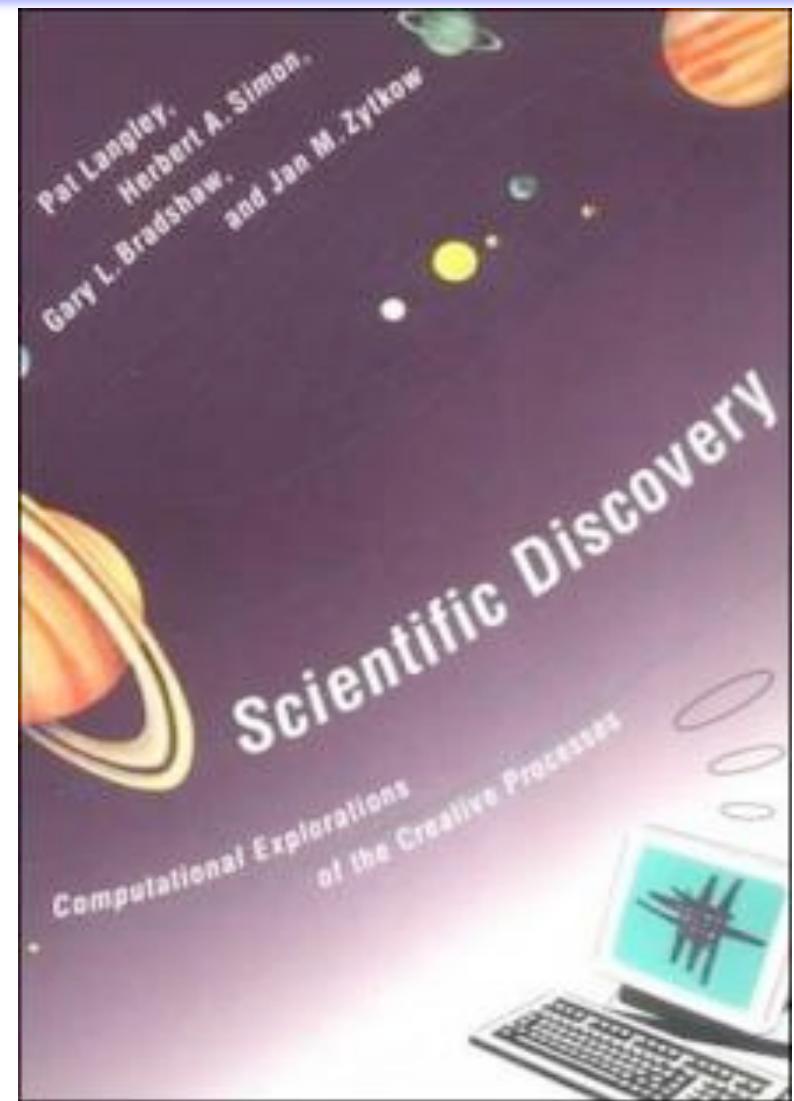
Ross D. King<sup>1,\*</sup>, Jem Rowland<sup>1</sup>, Stephen G. Oliver<sup>2</sup>, Michael Young<sup>3</sup>, Wayne Aubrey<sup>1</sup>, Emma Byrne<sup>1</sup>, Maria Liakata<sup>1</sup>, Magdalena Markham<sup>1</sup>, Pinar Pir<sup>2</sup>, Larisa N. Soldatova<sup>1</sup>, Andrew Sparkes<sup>1</sup>, Kenneth E. Whelan<sup>1</sup>, Amanda Clare<sup>1</sup>

Science

AAAS

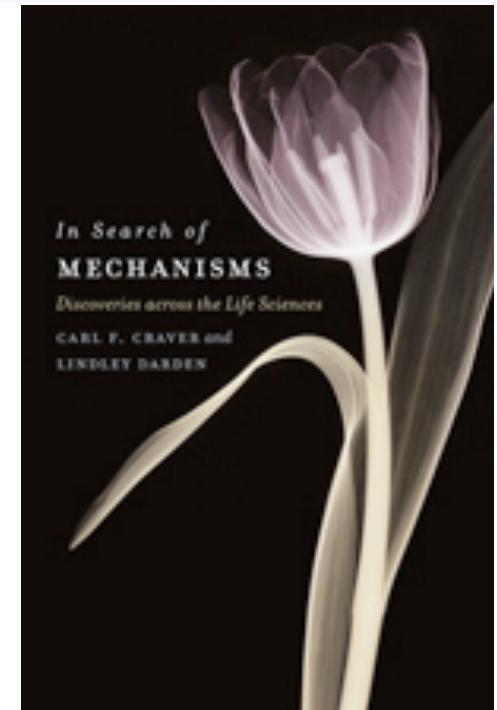
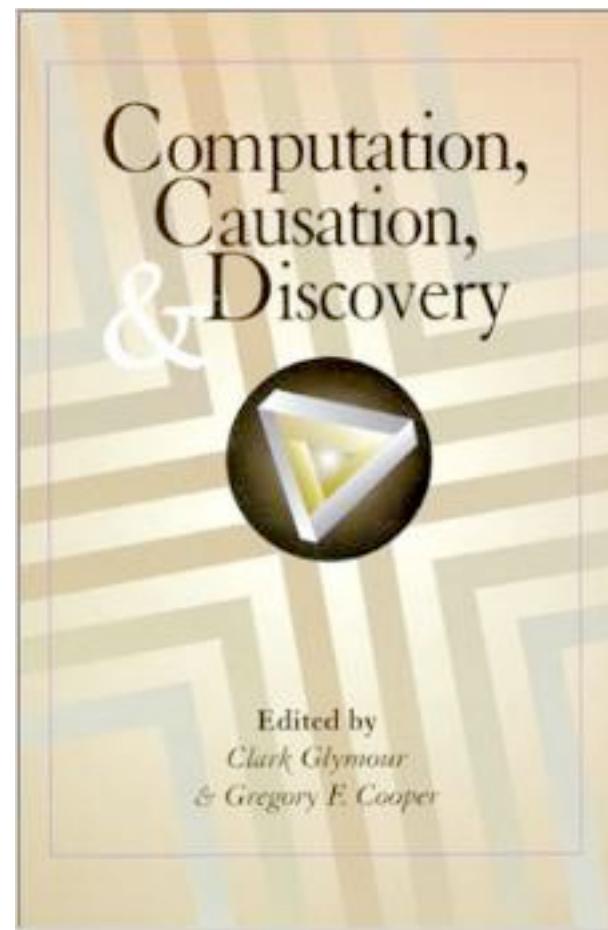
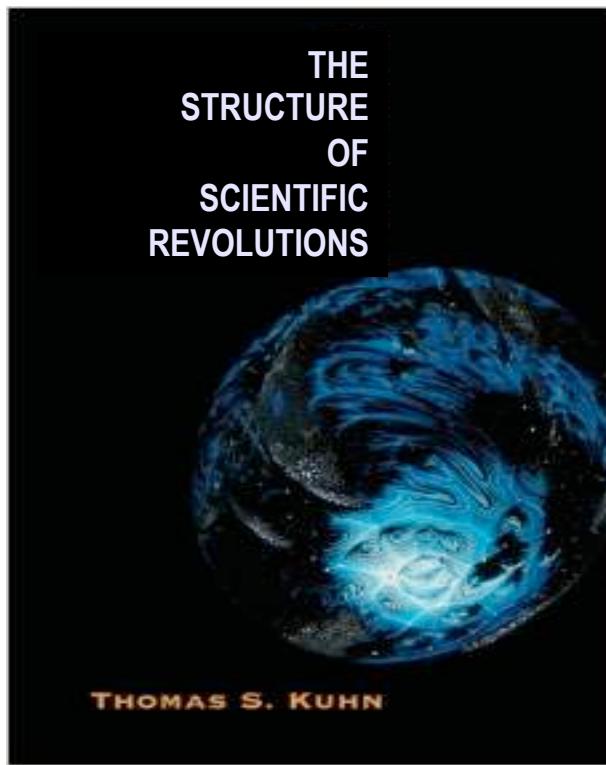
# Computational Scientific Discovery

- [Lenat 1976]
- [Lindsay et al 1980]
- [Langley 1981]
- [Falkenhainer 1985]
- [Kulkarni and Simon 1988]
- [Cheeseman et al 1989]
- [Zytkow et al 1990]
- [Simon 1996]
- [Valdes-Perez 1997]
- [Todorovski et al 2000]
- [Schmidt and Lipson 2009]



# Philosophy of Science

---

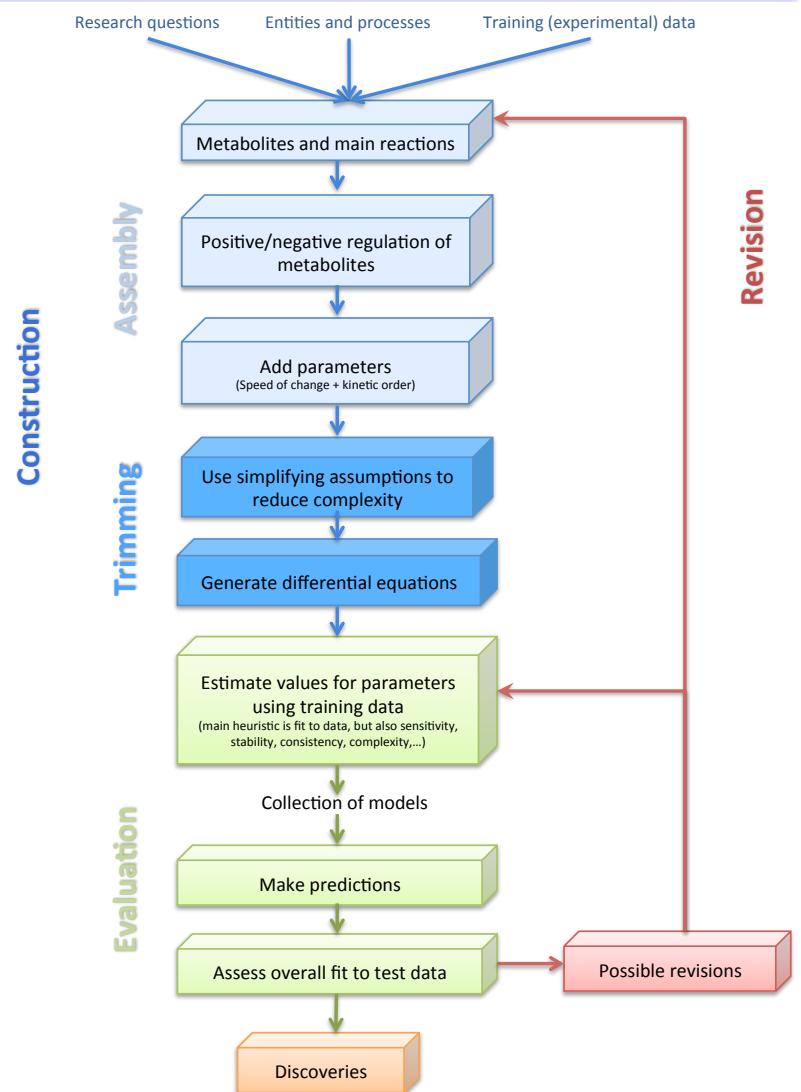


# Cognitive Science

A computational model of biological pathway construction [Chandrasekaran & Nersessian 2015]

1. Assembly
2. Trimming
3. Evaluation
4. Revision

*Adapted from [Chandrasekaran and Nersessian 2015], with thanks to Parag Mallick (Stanford), Dan Ruderman, and Shannon Mumenthaler of USC/PSOC.*



# Focus: Intelligent Science Assistants for Data Analysis

---

What is the state of the art?

What is a good problem to work on?

What is a good experiment to design?

What data should be collected?

What is the best way to analyze the data?

What are the implications of the experiments?

What are appropriate revisions of current models?

# Outline

---

1. The human bottleneck in data analytics
2. Related work on AI and cognitive aspects of scientific discovery
3. Semantic workflows to capture data analytics processes
4. Meta-reasoning to automate discovery
5. Discovery Informatics

# 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 AND OCEANIC ADMINISTRATION

COUNTDOWN TO LAUNCH

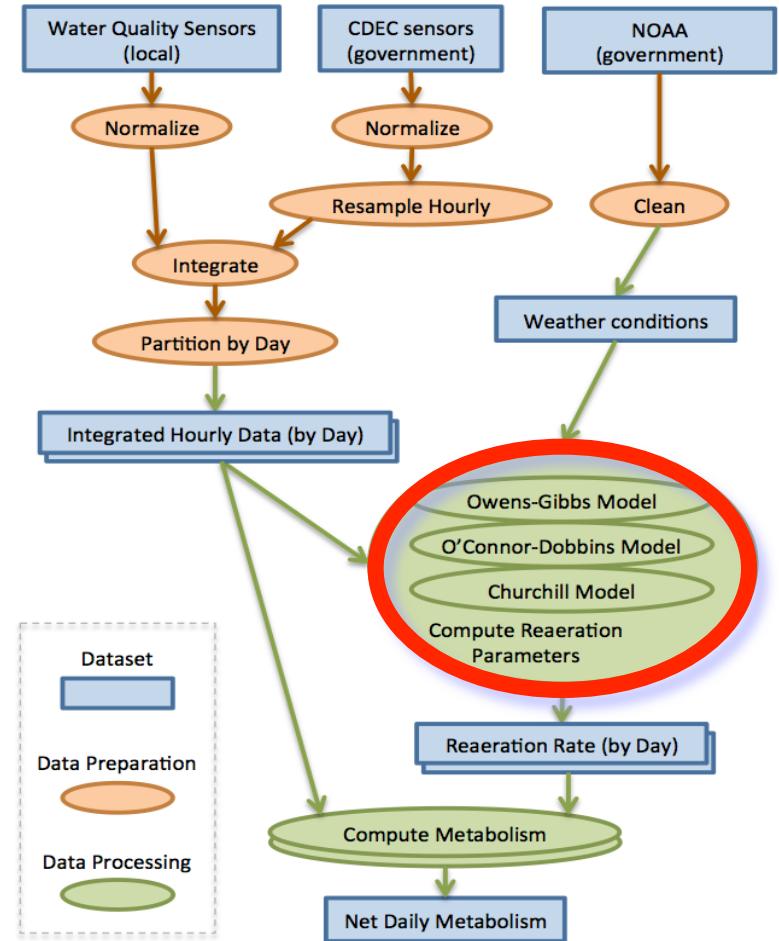
DEPARTMENT OF WATER RESOURCES California Data Exchange Center

MERCED RIVER NEAR STEVINSON

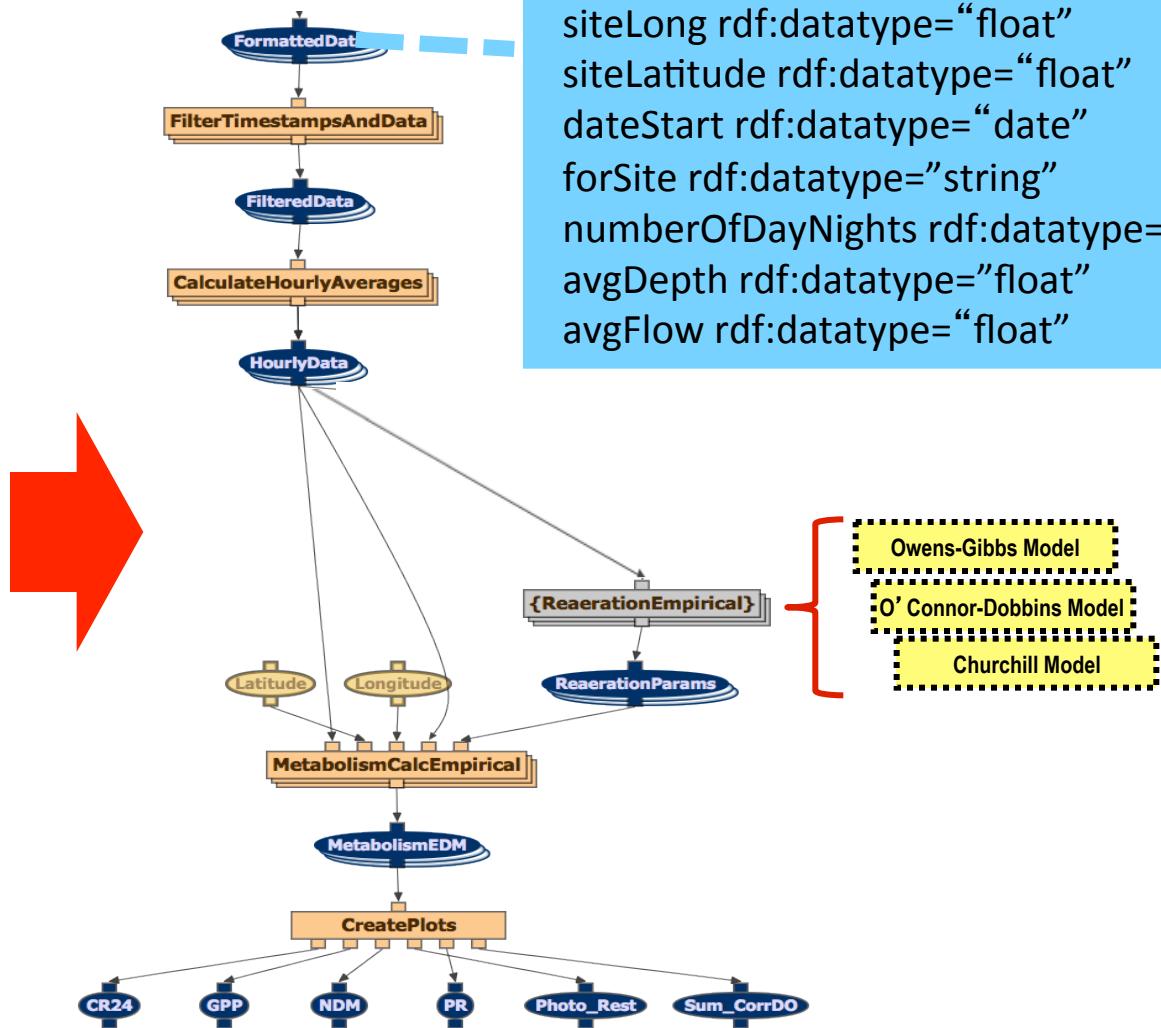
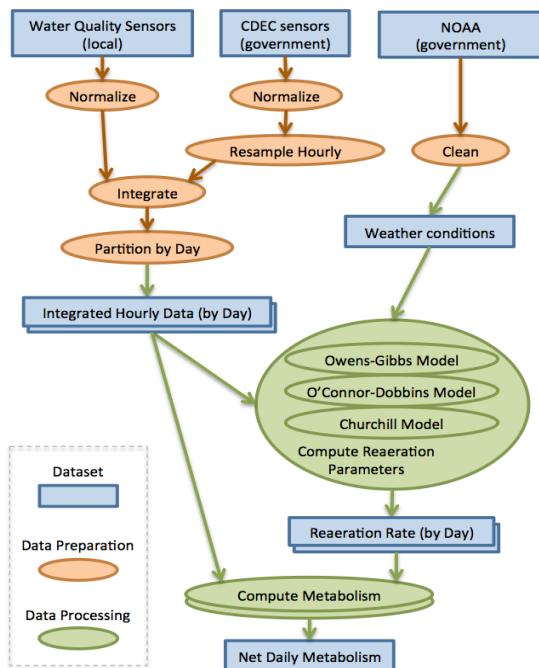
Station ID	MST	Elevation	62' ft
River Basin	MERCED R	County	MERCED
Hydrologic Area	SAN JOAQUIN RIVER	Nearby City	STEVINSON
Latitude	37.371000°N	Longitude	120.931000°W
Operator	CA Dept of Water Resources	Data Collection	

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

```

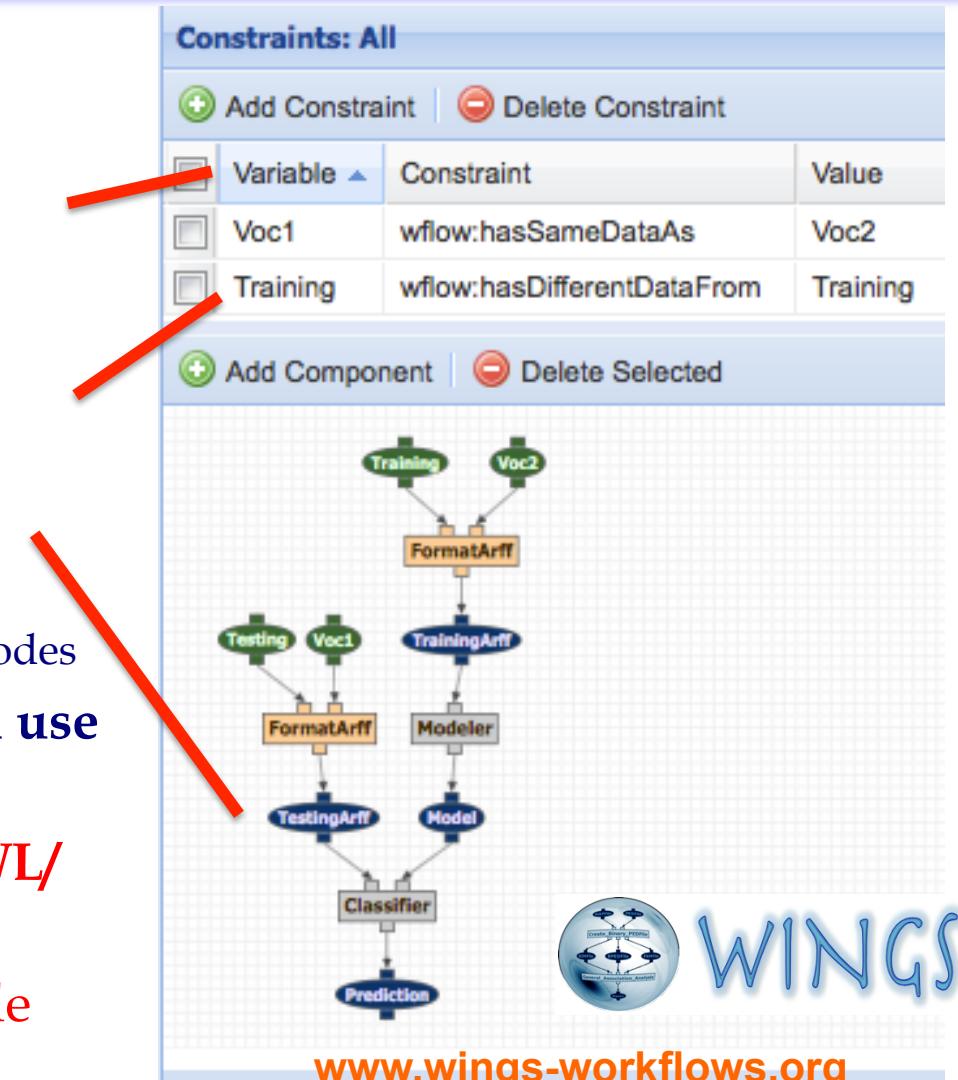
isa Hydrolab_Sensor_Data
siteLong rdf:datatype="float"
siteLatitude rdf:datatype="float"
dateStart rdf:datatype="date"
forSite rdf:datatype="string"
numberOfDayNights rdf:datatype="int"
avgDepth rdf:datatype="float"
avgFlow rdf:datatype="float"
  
```

# 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



# Semantic Components in WINGS

[Gil iEMSSs 2014]

**Wings Po**

Home Analysis

**Components: Default**

- + Add Component - Delete
- + CalculateHourlyAvera
- + ConvertToStandardFormat
- + CreateParametersF
- + CreatePlots
- + FilterTimestampsAndData
- + HourlyAverage
  - Aquaflow\_Hourly\_Average\_Inputs
- + MergeMetabolismResults
- + Metabolism
  - MetabolismDay
- + MetabolismCalcEmpirical
- + MetabolismCalculation
- + ReaerationEmpirical
  - ReaerationCM
  - ReaerationODM
  - ReaerationOGM
- + ReaerationPhysics
- + ReaerationEDM

**Classes of models/ components**

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 Data constraints**

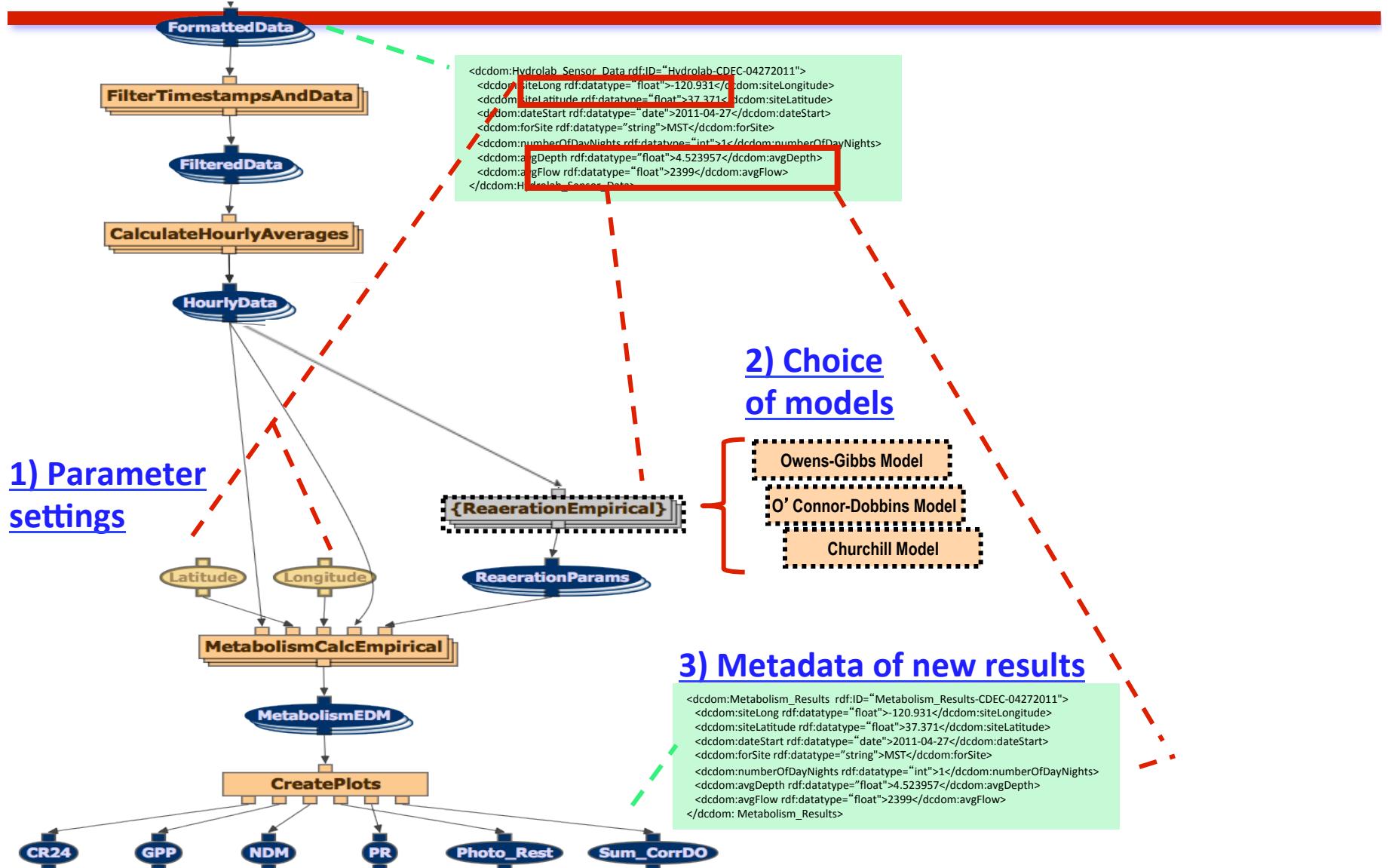
I/O Rules Inherited Rules Documentation		
Input Data		
Name	Type	Prefix
InputParameters	dcdom:Hourly_Averaged_Input	-i1
Input Parameters		
Name	Type	Prefix
velocity	xsd:float	-p2
slope	xsd:float	-p4
depth	xsd:float	-p1
flow	xsd:float	-n3
Output Data		
Name	Type	Prefix
K2Result		

**Use constraints**

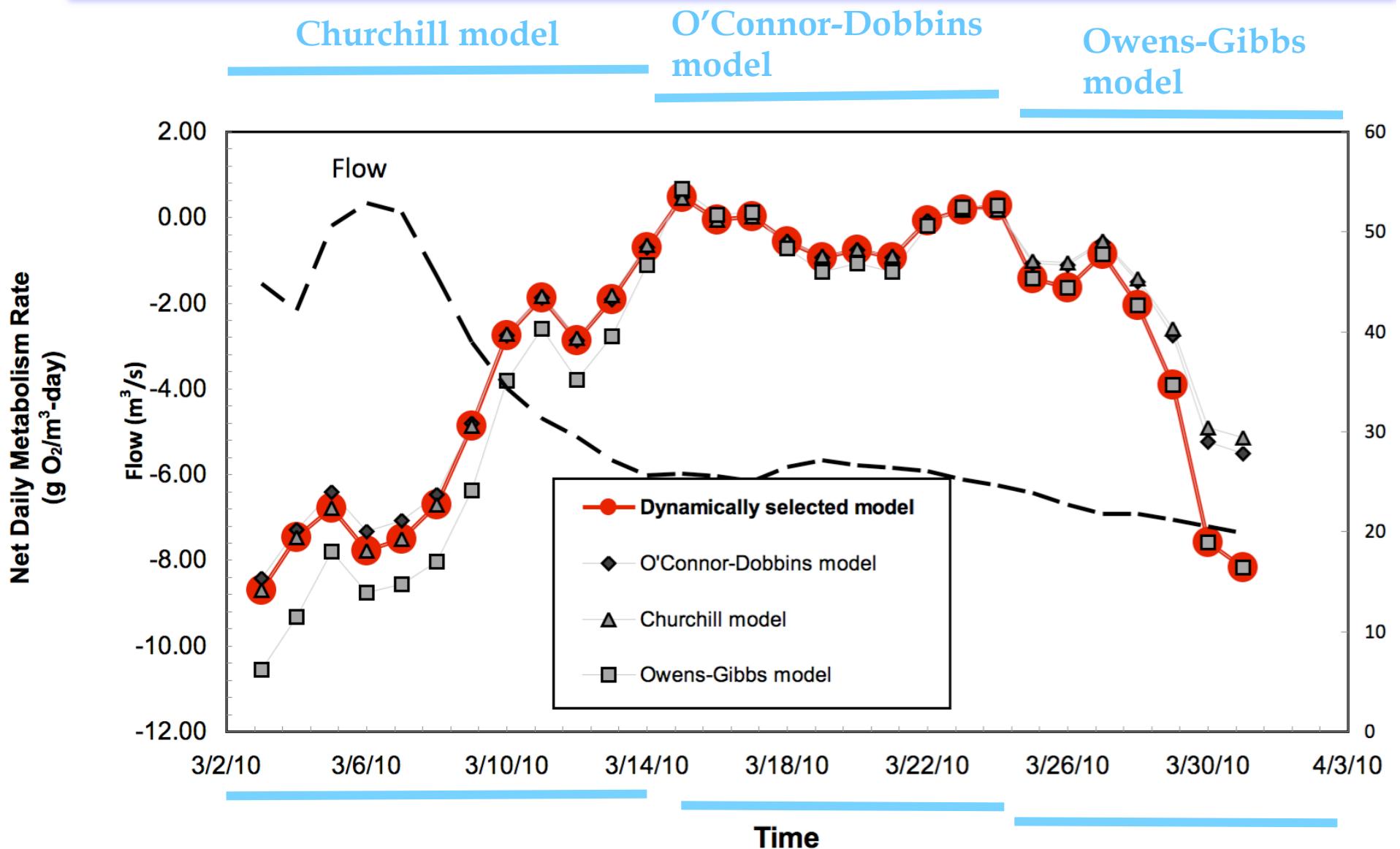
```
;; Depth must be over .6m
[ CMInvalidity1:
  (?c rdf:type pcdom:ReaerationCMClass)
  (?c pc:hasInput ?idv)
  (?idv pc:hasArgumentID
  'InputParameters')
  (?idv dcdom:depth ?depth)
  le(?depth '0.61')
  -> (?c pc:isValid 'true'))]
```

Yolanda

# WINGS Specializes Workflow Based on Characteristics of Daily Data



# WINGS Dynamically Selects Appropriate Model Based on Daily Sensor Readings

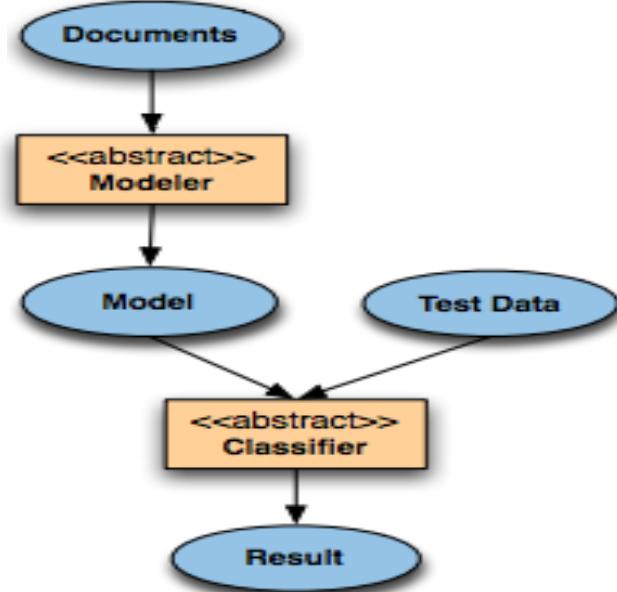


# Workflows Capture Data Analytics Expertise

[Hauder et al e-Science 2011]

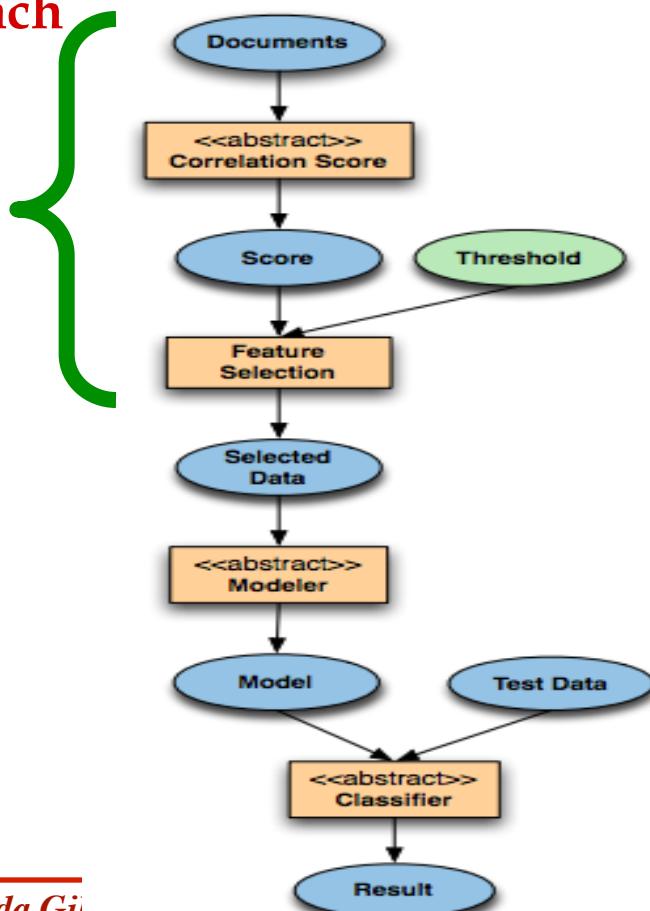
*Workflows for text analytics, joint work with Yan Liu (USC) and Mattheus Hauder (TUM)*

## Naïve Approach



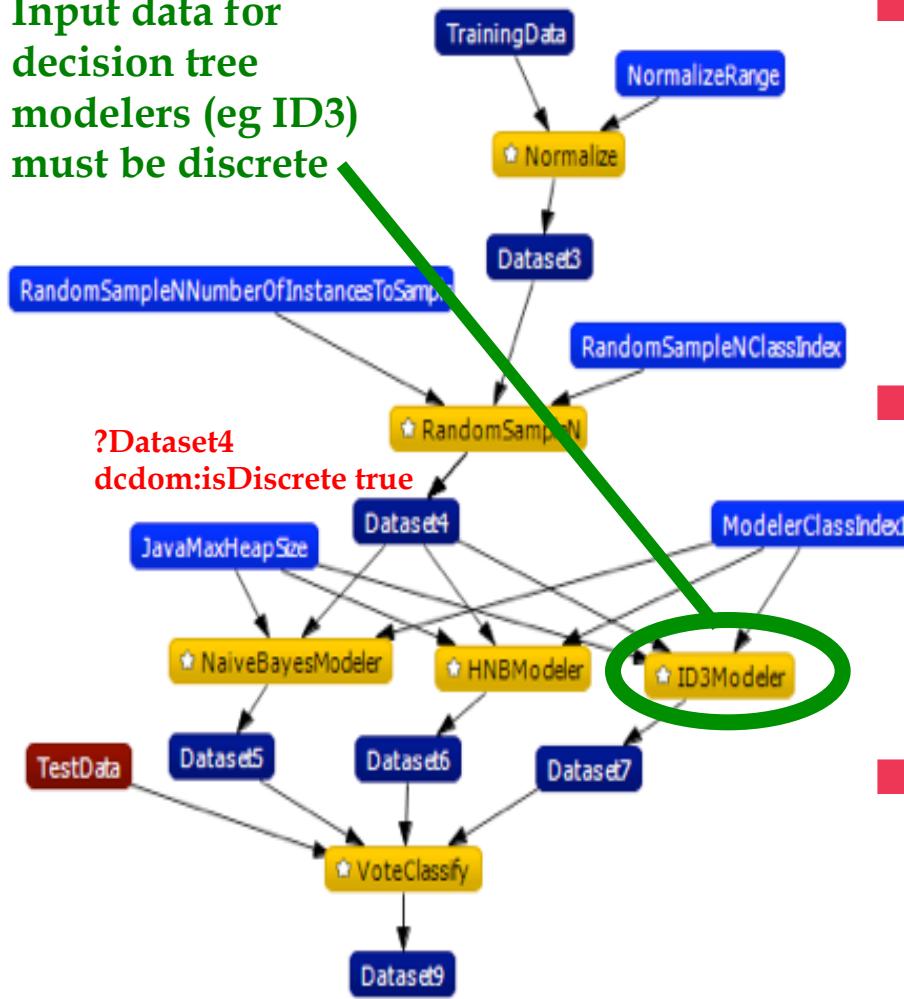
## Expert Approach

### Feature selection



# 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

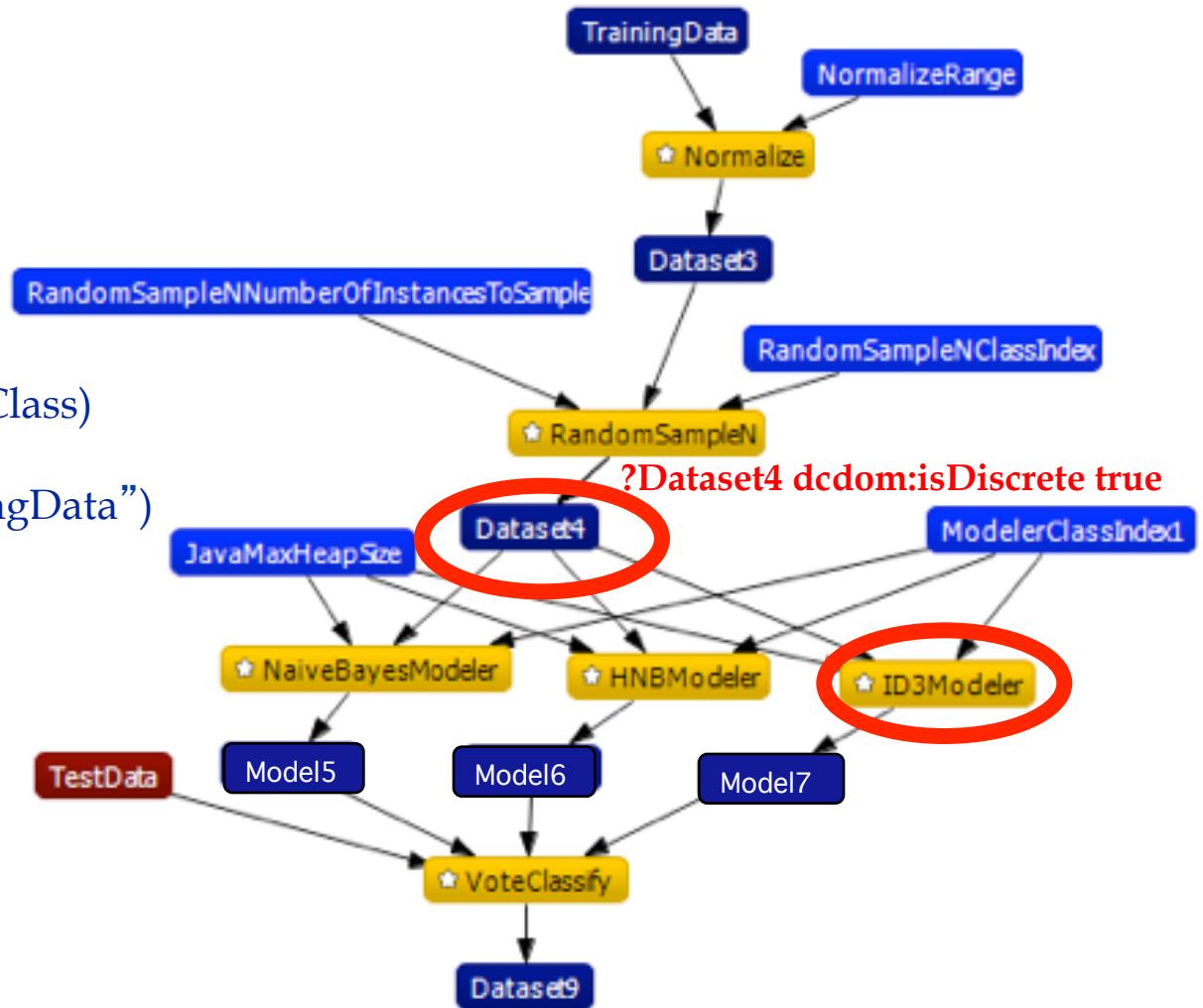
# Example (Step 1 of 5)

Rule in Component Catalog:

[modelerSpecialCase2:

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

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

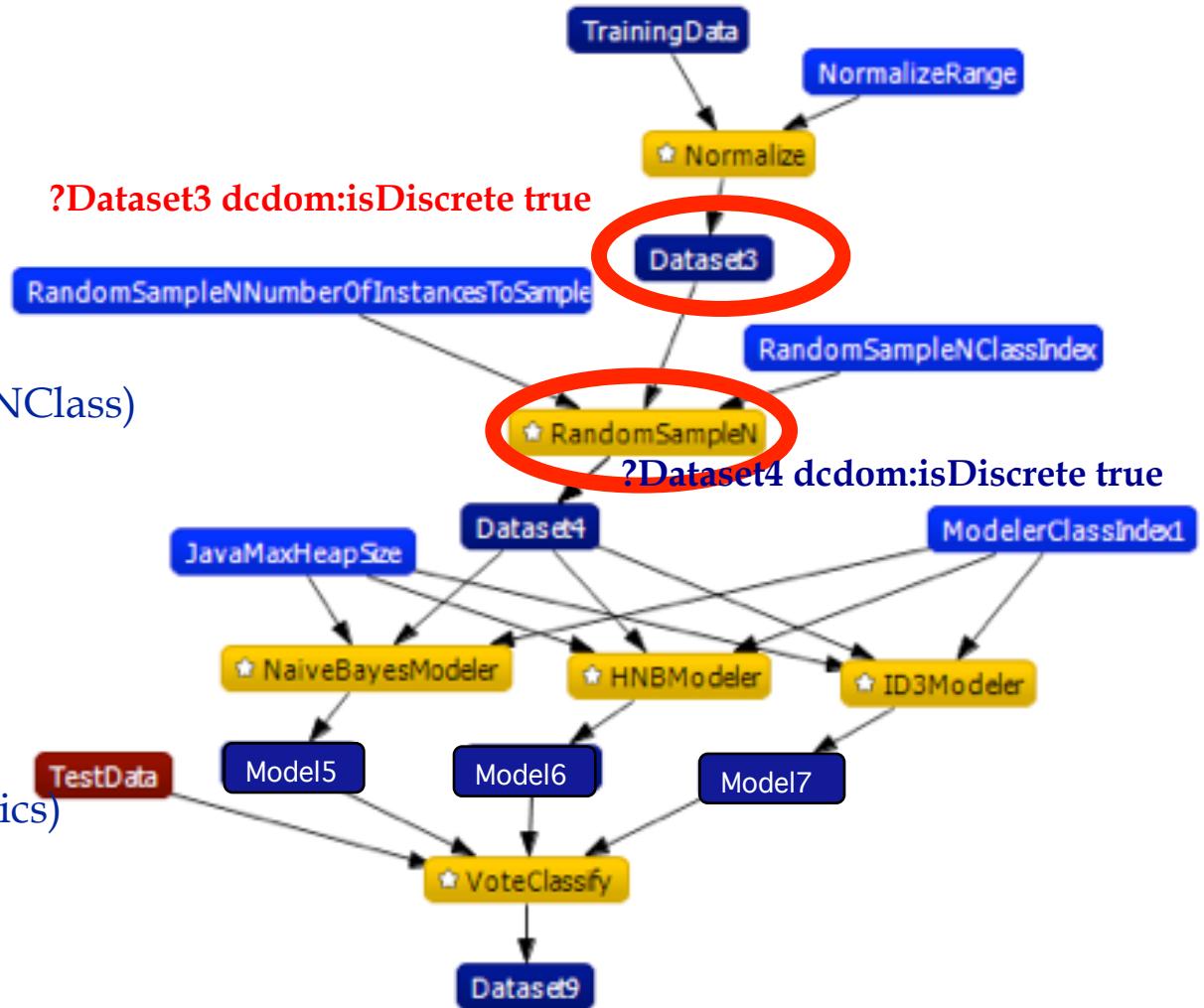


## Example (Step 2 of 5)

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)]
```



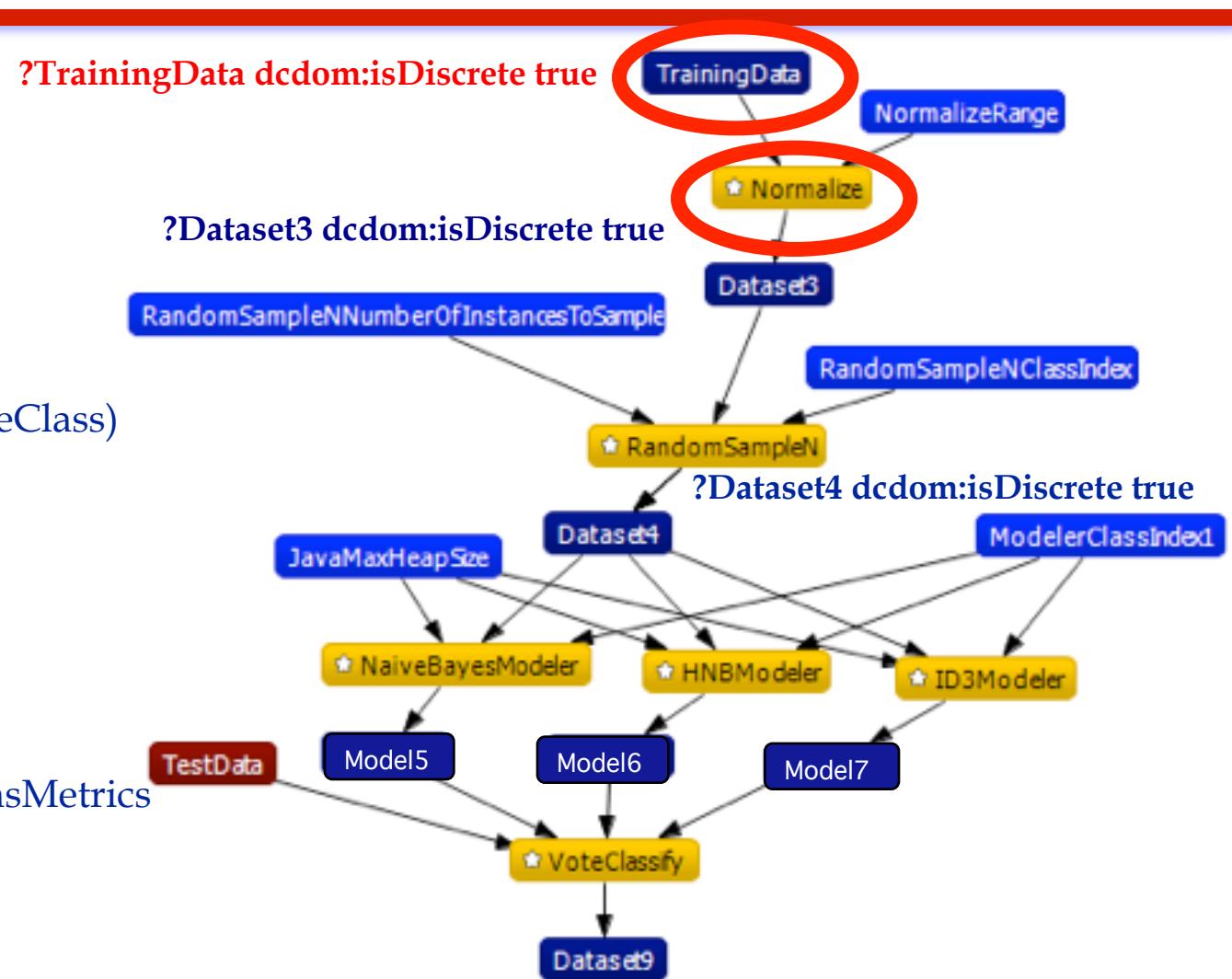
# Example (Step 3 of 5)

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)

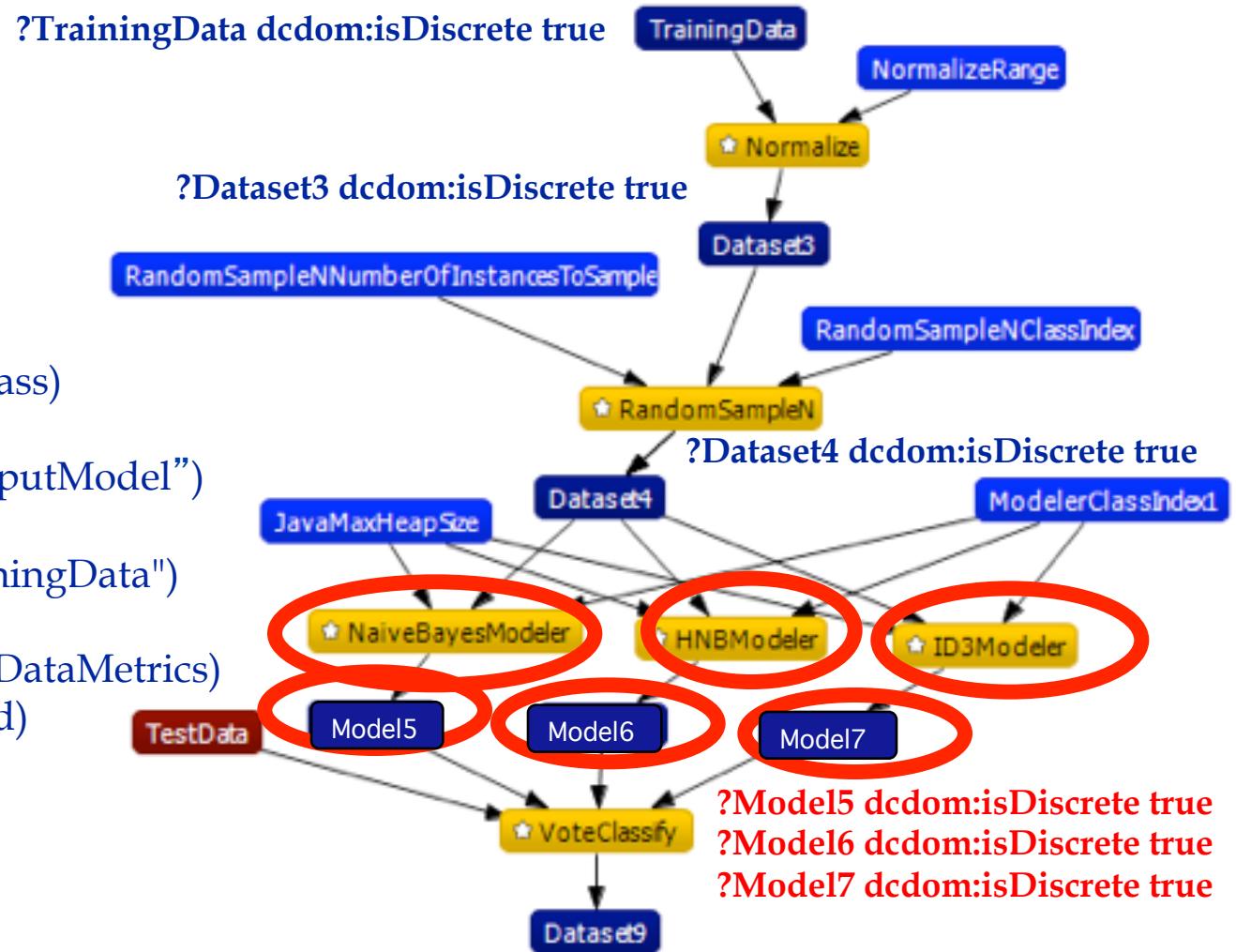
(?idv pc:hasArgumentID "trainingData")

(?idv ?p ?val)

(?p rdfs:subPropertyOf dc:hasDataMetrics)

notEqual(?p dcdm:isSampled)

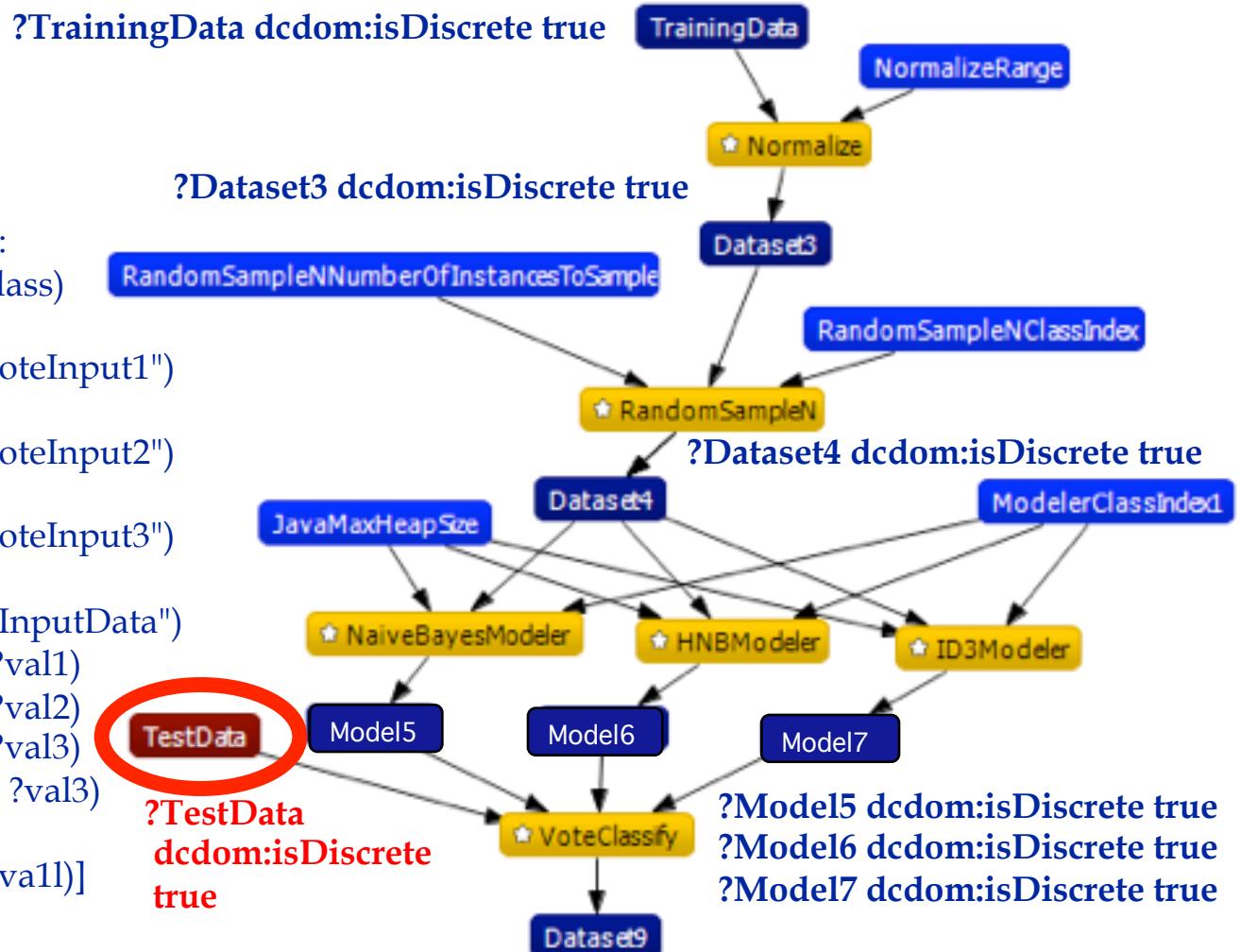
-> (?odv ?p ?val)]



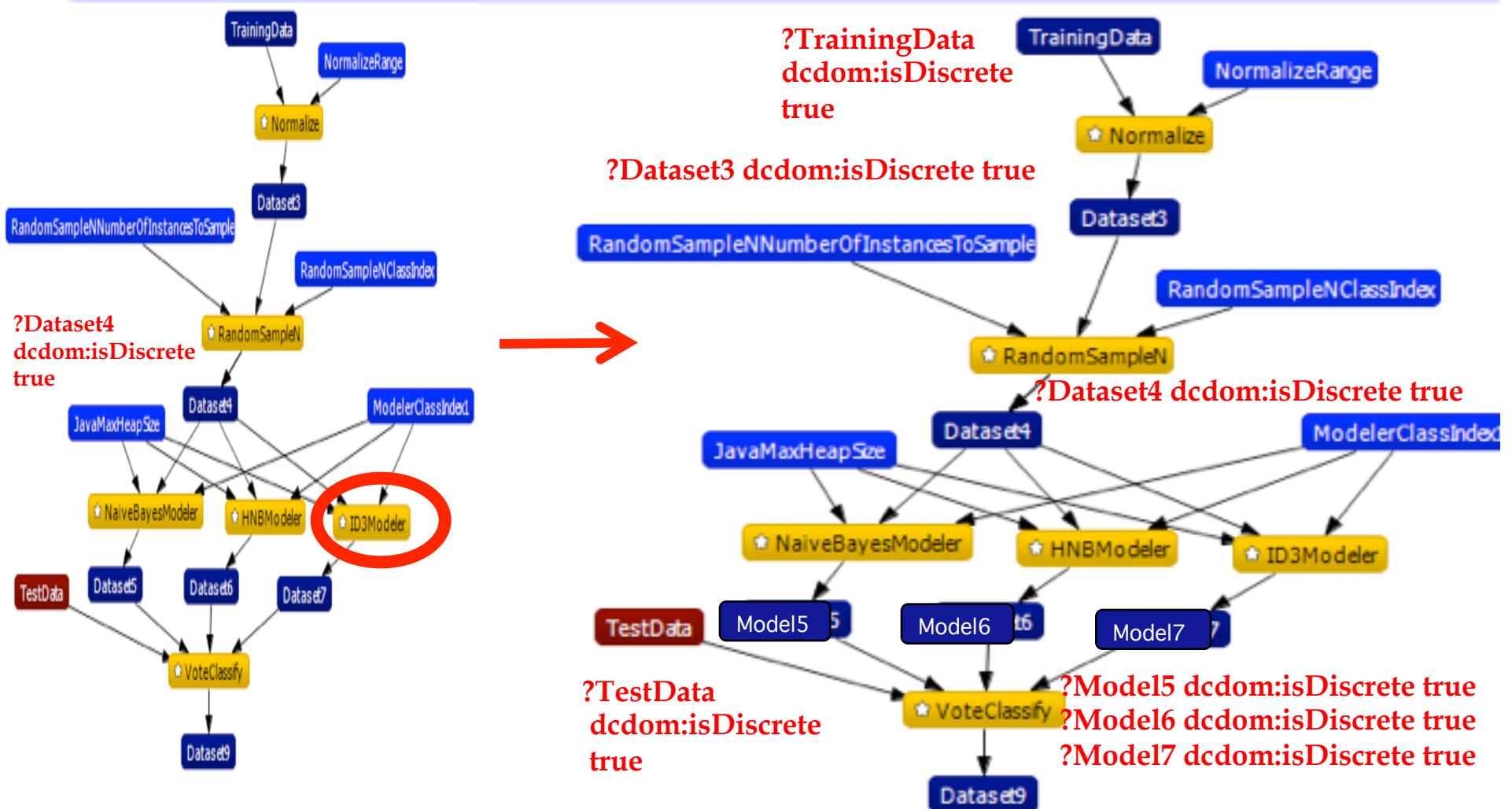
# Example (Step 5 of 5)

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)]
```

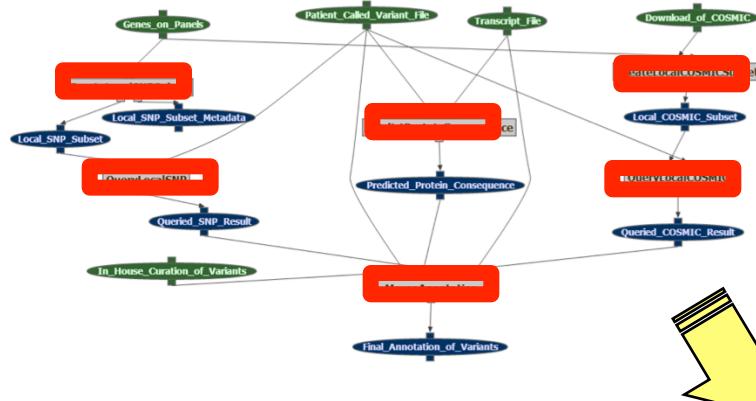


# WINGS Workflow Reasoners: Result

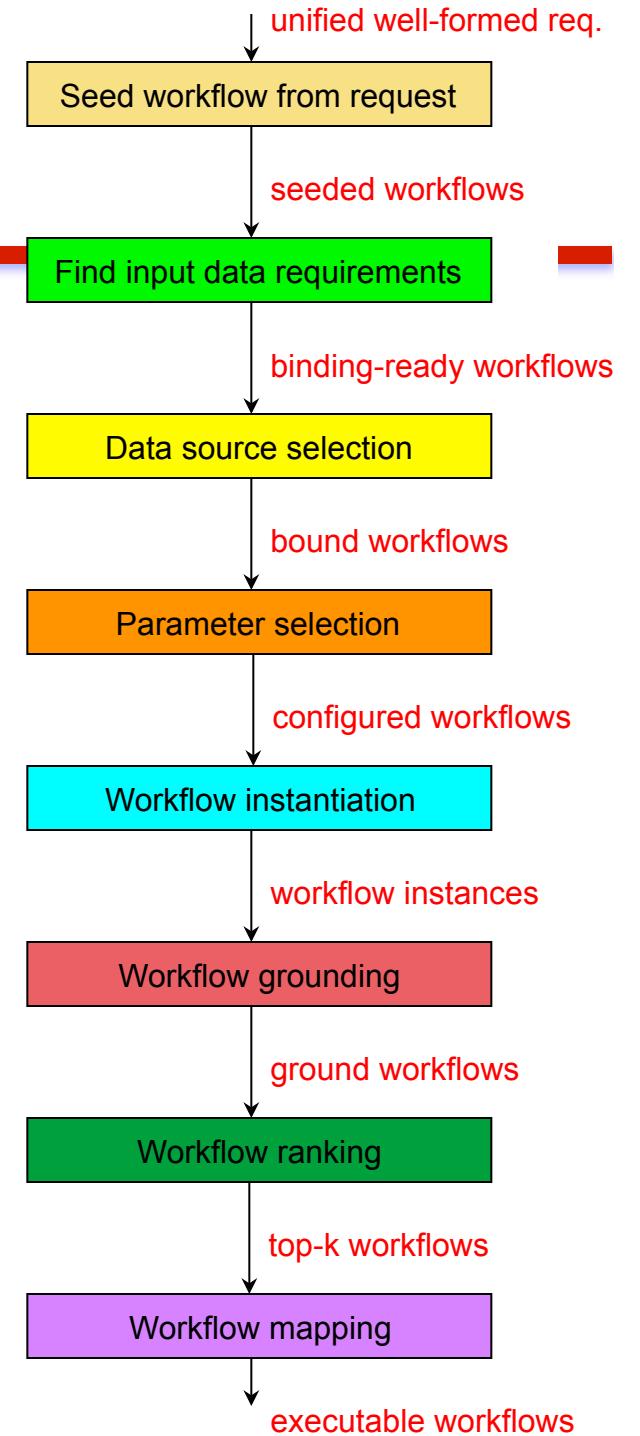
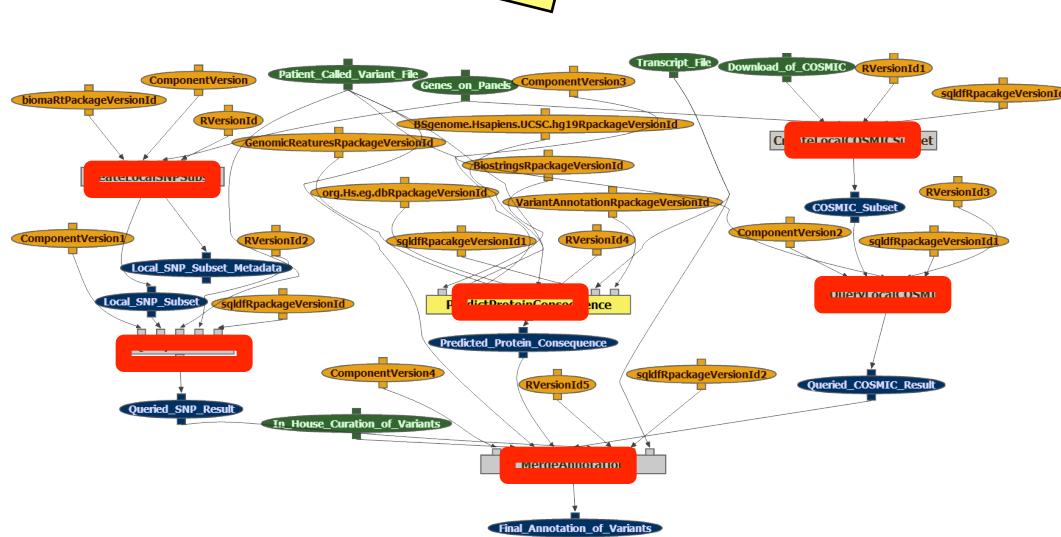


# WINGS Automatic Workflow Generation Algorithm [Gil et al JETAI 2011]

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



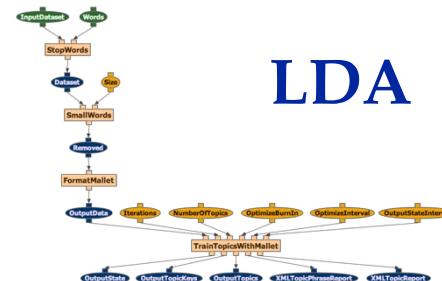
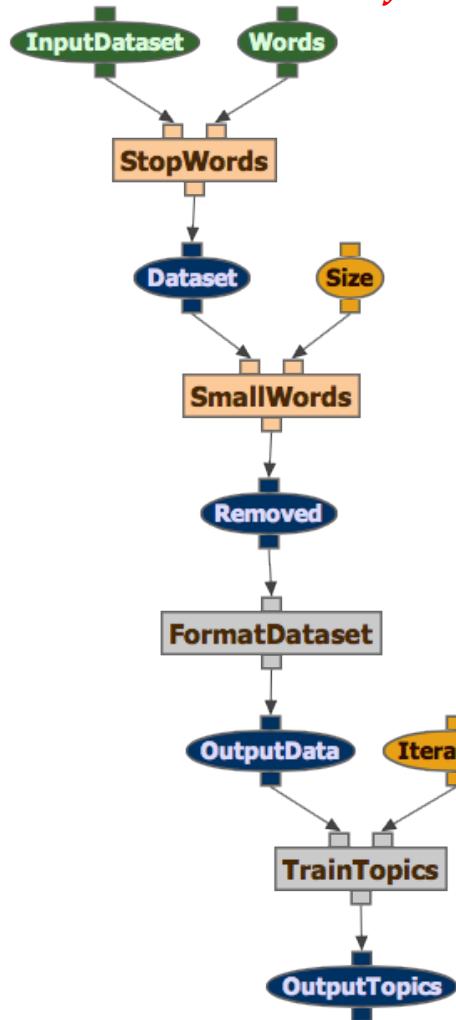
**“Pay-as-you-go” semantics**



# Benefits of Semantic Workflows:

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

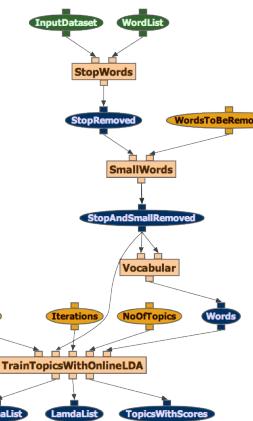
*Workflows developed with Y. Liu (USC) and C. Mattmann (JPL)*



LDA



Online LDA

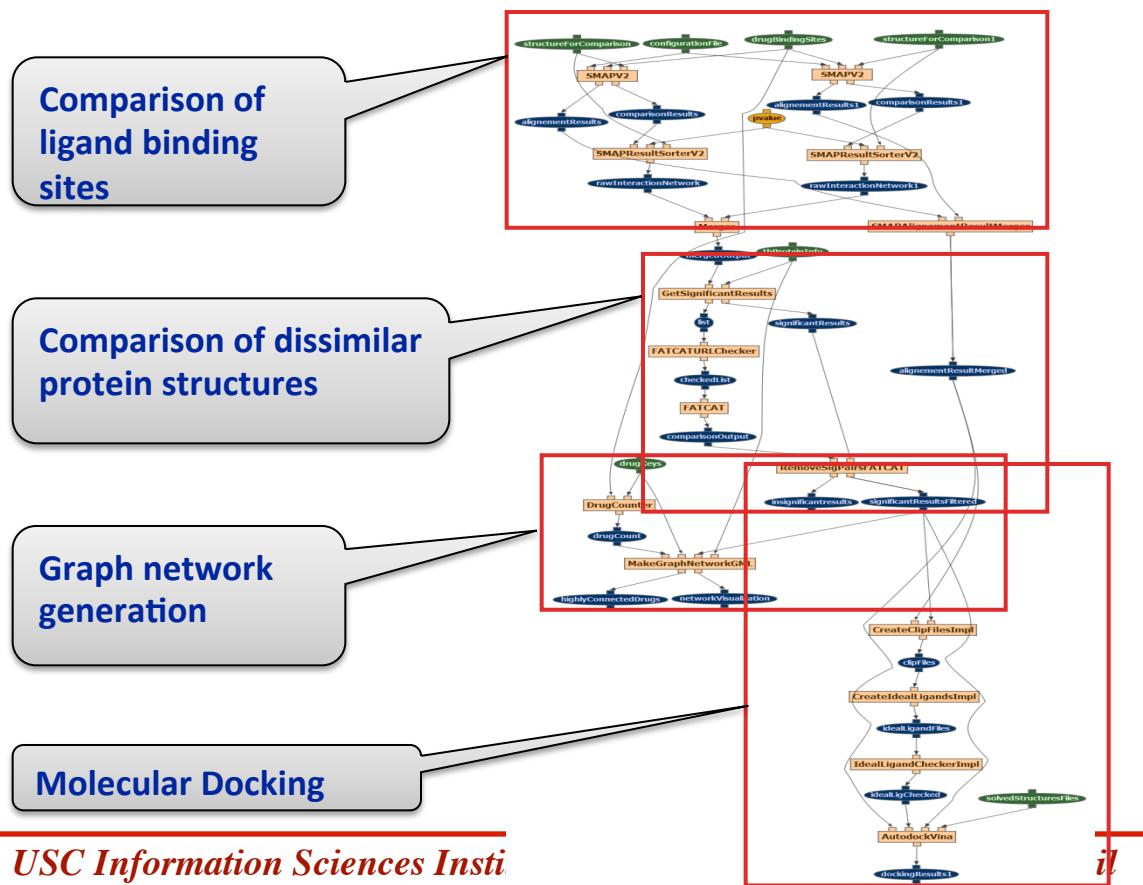


Parallel  
LDA

# 3) Capturing Expertise with Workflows: “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:

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

Novice Author	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

Comparison of dissimilar protein structures:

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

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

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

Docking

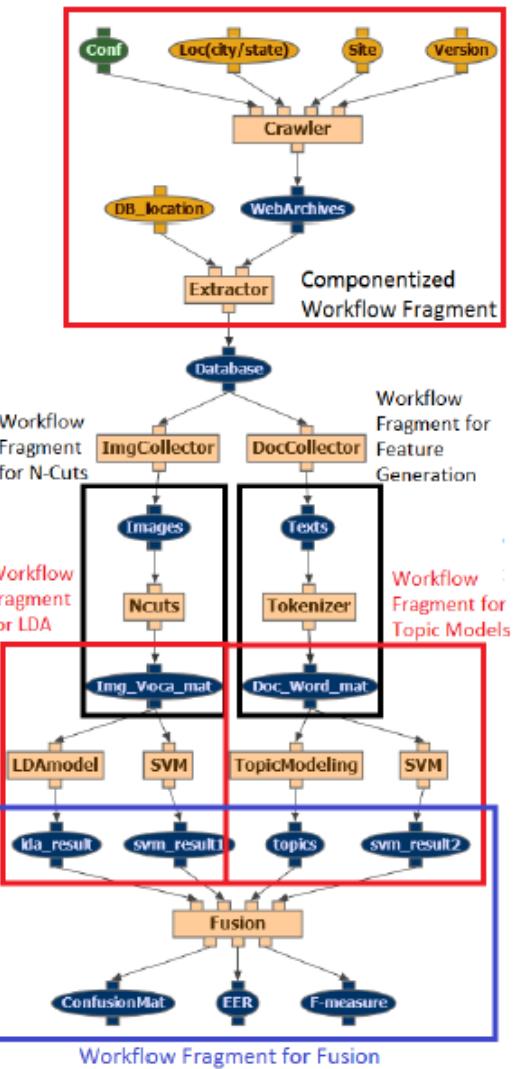
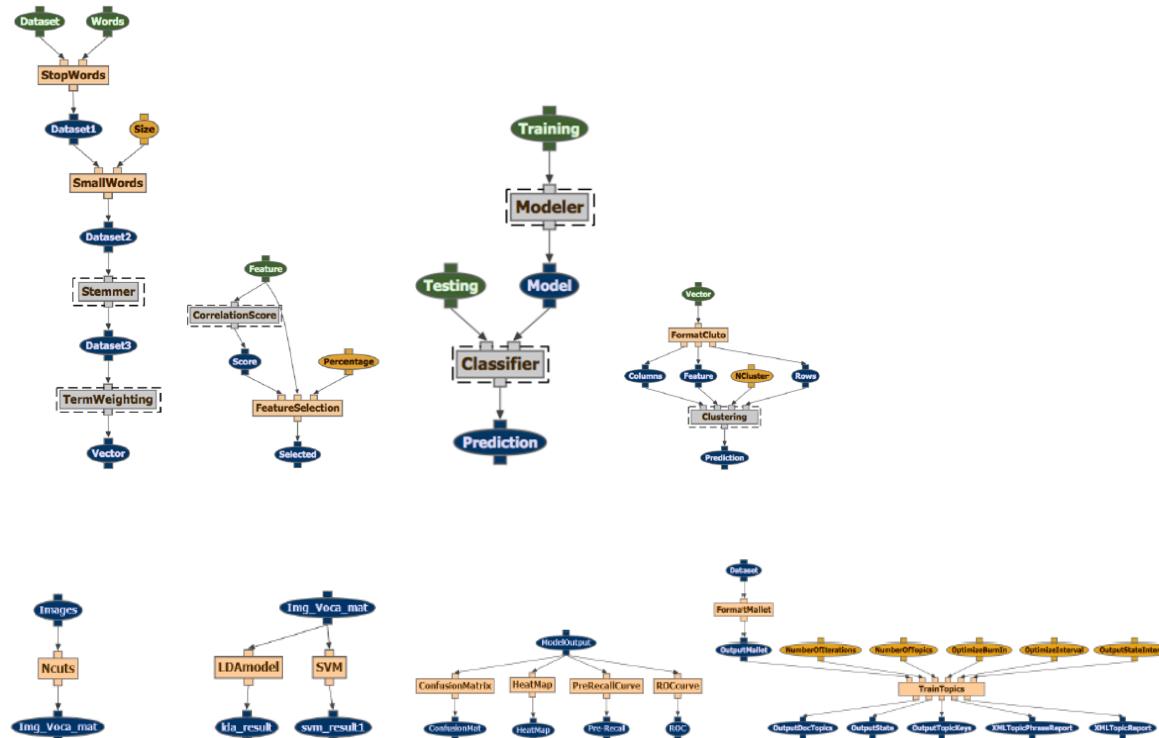
Minimal	CreateClip Files	Createideal Ligands	IdealLigand Checker	Autodock Vina
Novice	CreateClip Files	Createideal Ligands	IdealLigand Checker	Autodock Vina

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

# Benefits of Semantic Workflows:

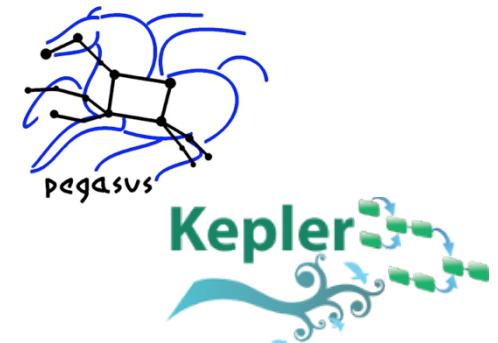
## 3) Efficiency Through Reuse [Sethi et al MM'13]

*Work with Ricky Sethi and Hyujoon Jo of USC*



# Related Work: Workflow Systems

- 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]



# Related Work: 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]

## 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.

# Outline

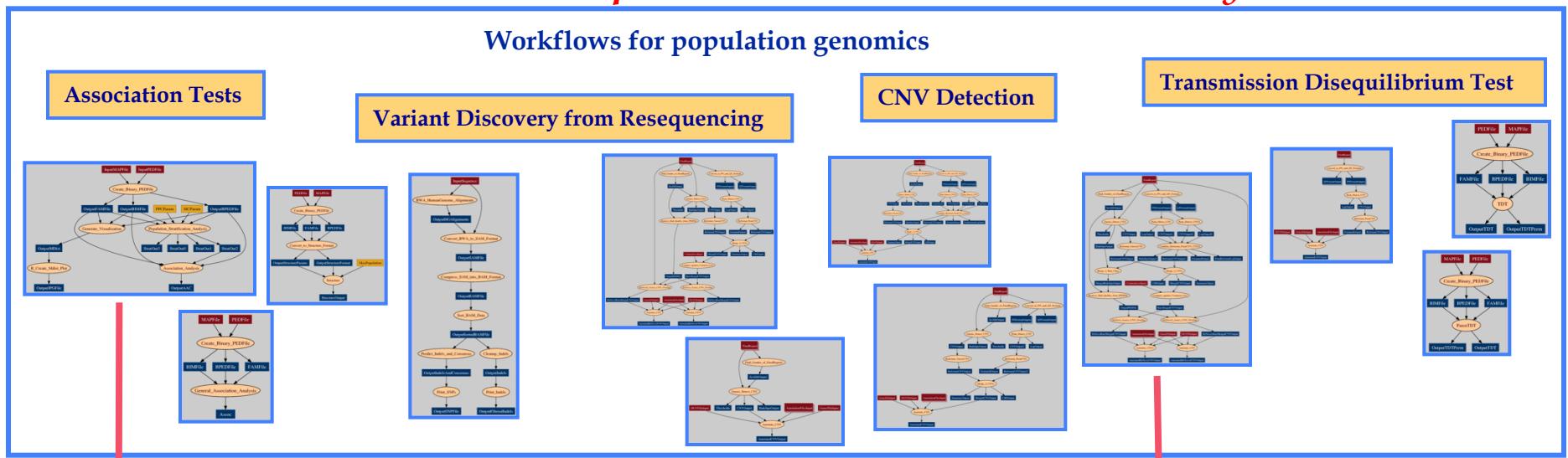
---

1. The human bottleneck in data analytics
2. Related work on AI and cognitive aspects of scientific discovery
3. Semantic workflows to capture data analytics processes
4. Meta-reasoning to automate discovery
5. Discovery Informatics

# A Workflow Library for Population Genomics

[Gil et al 2012]

*Work with Christopher Mason (Cornell University)*

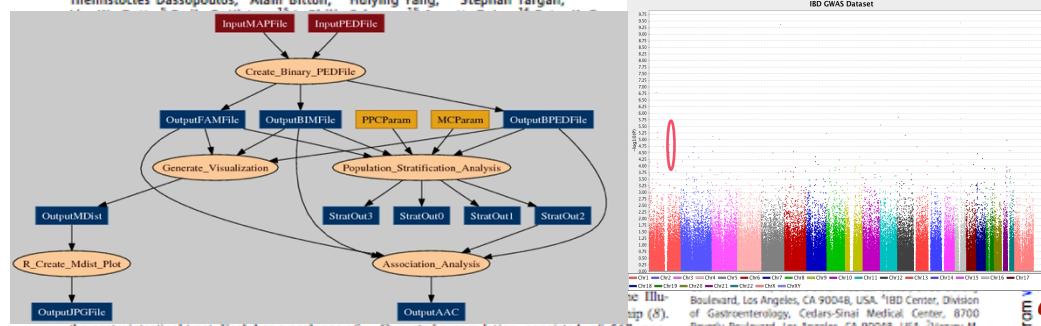


## A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr,<sup>3,2</sup> Kent D. Taylor,<sup>3,4</sup> Steven R. Brant,<sup>5,6</sup> John D. Rioux,<sup>7,8</sup> Mark S. Silverberg,<sup>9</sup> Mark J. Daly,<sup>3,10</sup> A. Hillary Steinhart,<sup>7</sup> Clara Abraham,<sup>11</sup> Miguel Regueiro,<sup>10</sup> Anne Griffiths,<sup>12</sup> Themistocles Dassopoulos,<sup>5</sup> Alain Bitton,<sup>13</sup> Hulying Yang,<sup>3,4</sup> Stephan Targan,<sup>14</sup>

candidate. 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).

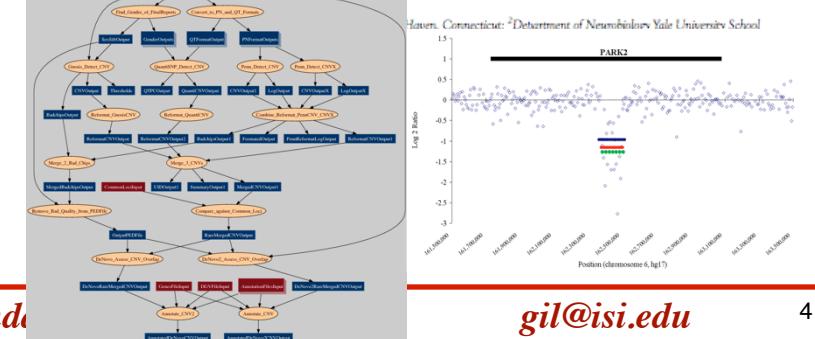
We next tested for association of *IL23R* markers in an independent ileal CD case-control cohort, consisting of 401 patients and 433 controls, all of Jewish ancestry (8). Significant as-



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

Fatih Bayrakli,<sup>1,2</sup> Kaya Bilgavar,<sup>1,3</sup> Christopher E. Mason,<sup>3,5</sup> Michael L. DiLuna,<sup>1,3</sup> Yasar Bayri,<sup>1,3</sup> Richard P. Lifton,<sup>5,6</sup> Matthew W. State,<sup>3-5</sup>

1,2Yale University School of Medicine, New Haven, Connecticut; <sup>3</sup>Department of Neurobiology, Yale University School

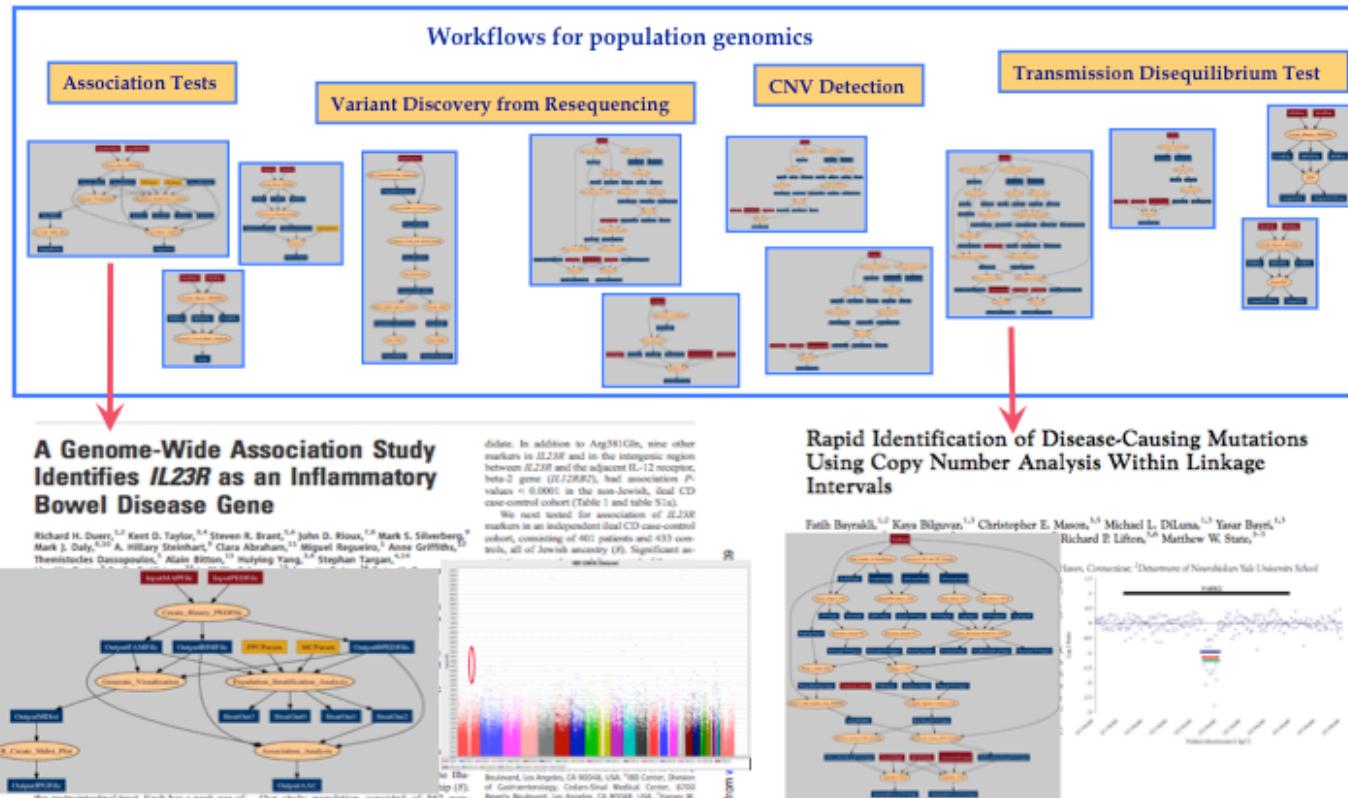


gil@isi.edu

# A Grand Challenge: Automatic Analysis of Entire Data Repositories

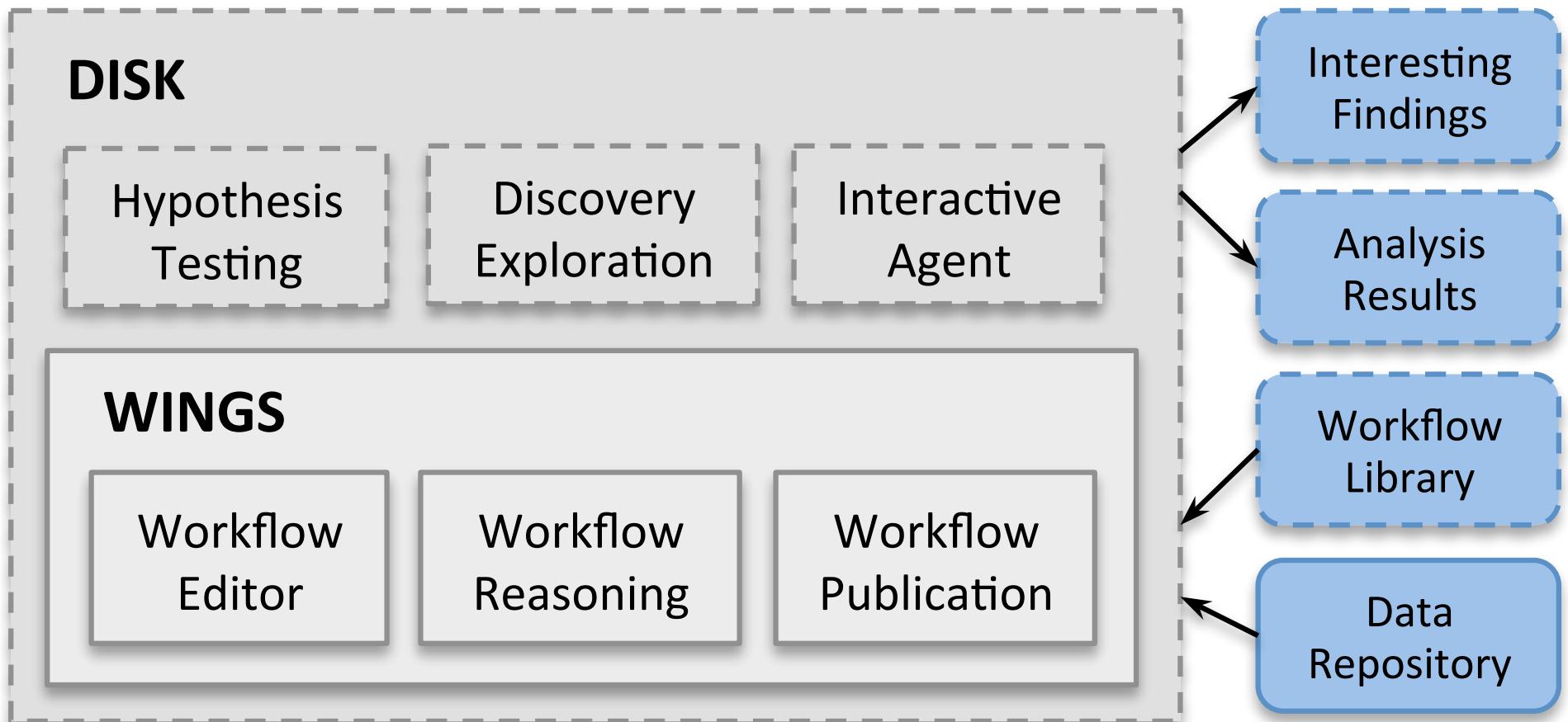
■ Capture knowledge about analytic methods

- Run workflows in existing data repositories
- Report new findings



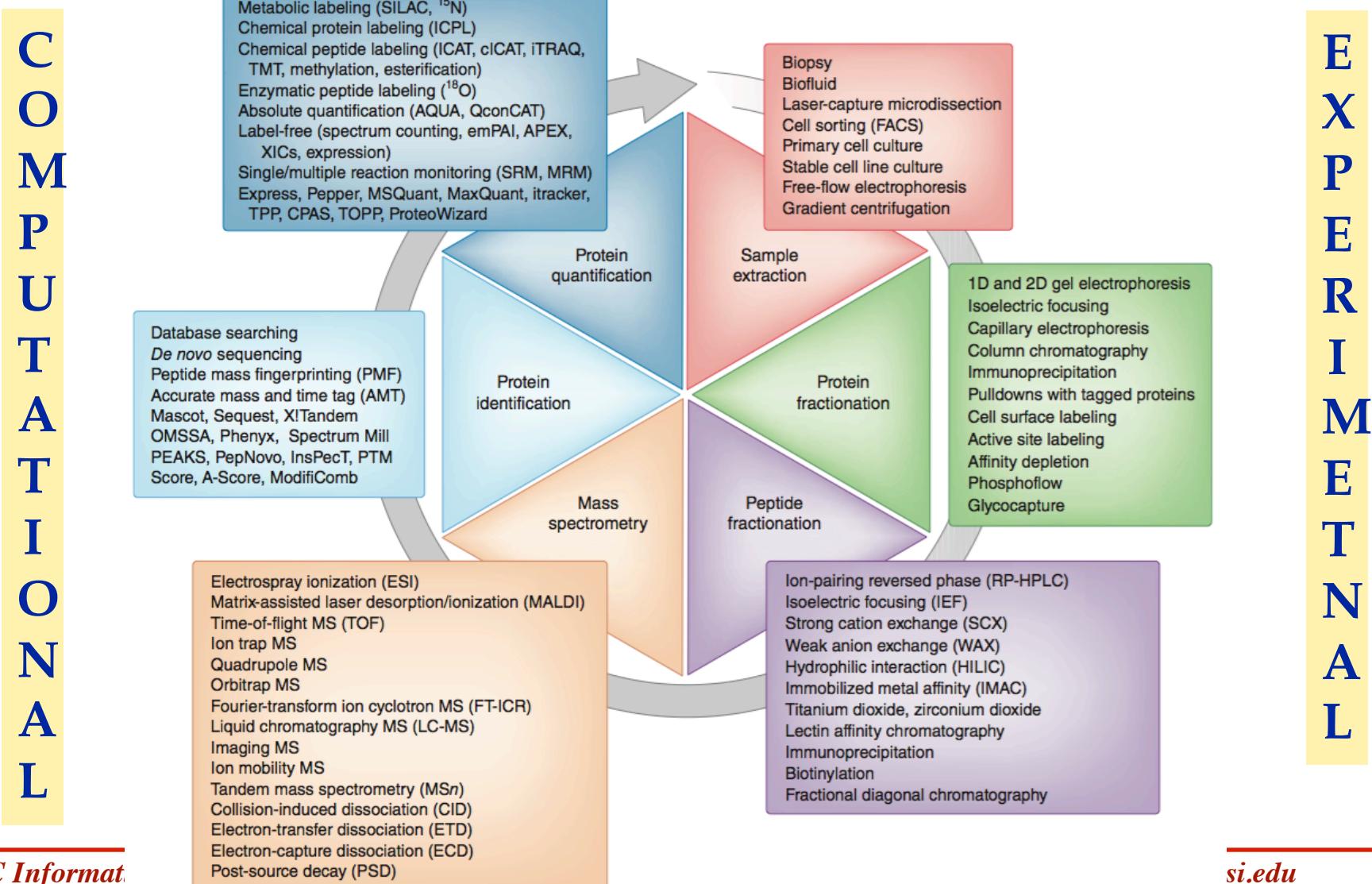
# Meta-Workflows for Identifying Interesting Findings of Analysis Workflows

*Work with Parag Mallick (Stanford University)*

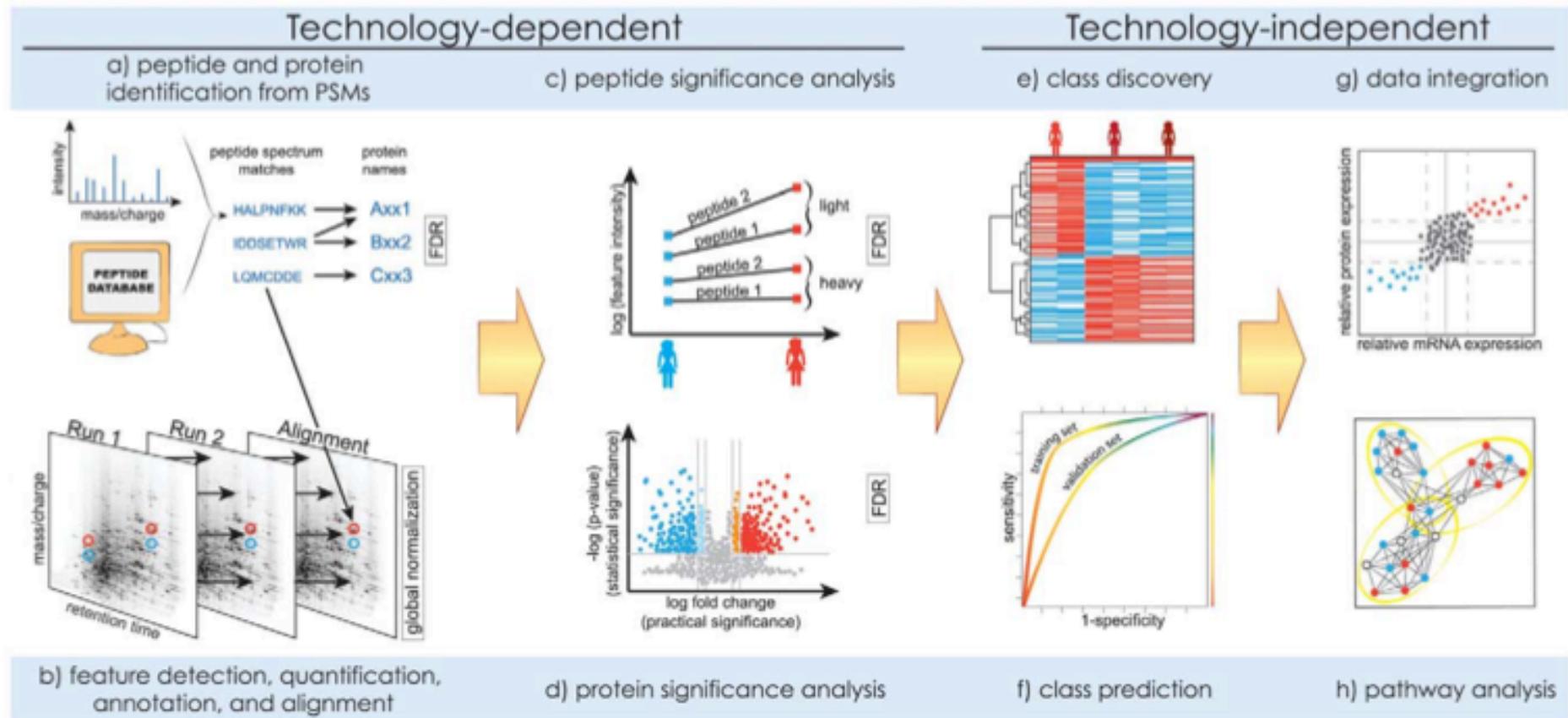


# A Wide Range of Computational Workflow Options: Automated Process Would Be Systematic for Entire Data Repositories

Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. *Nat Biotechnol* 28, 695–709 (2010)



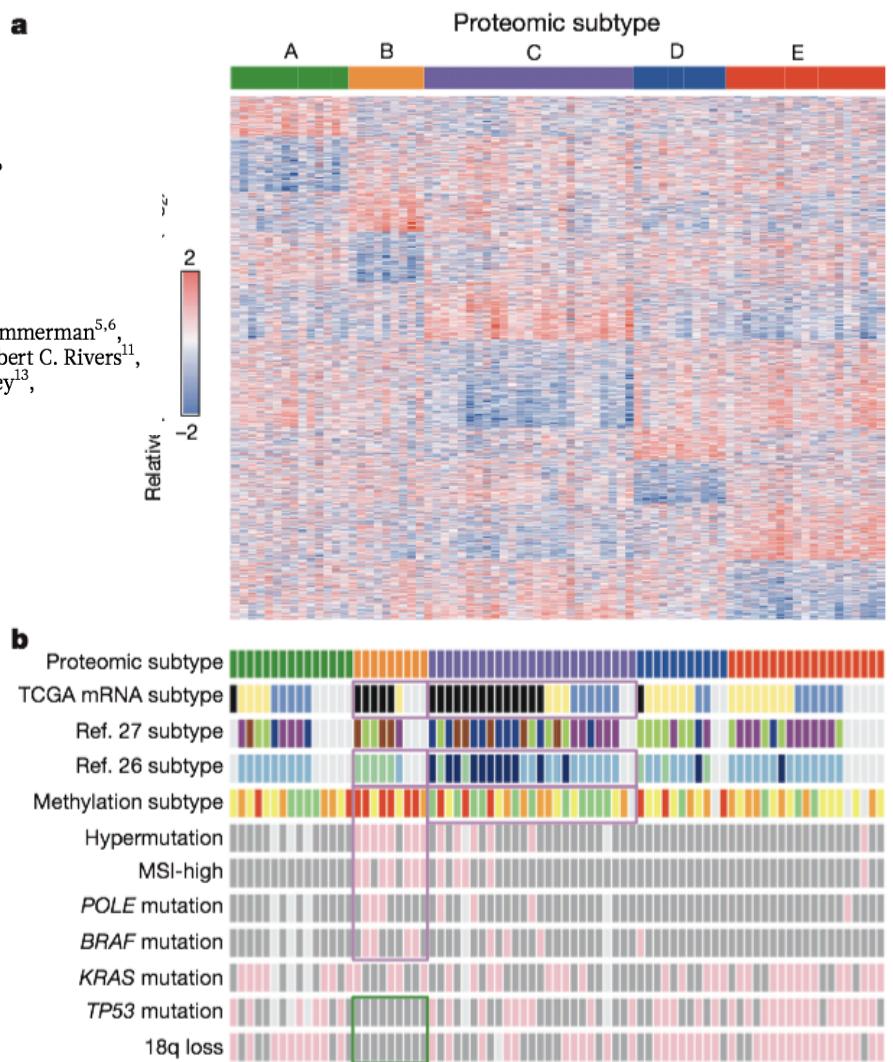
# Upstream Processing Affects Downstream Results: Automated Process Would Avoid Errors



# Compartmentalized Expertise: Automated Process Would Cover Multiple Expertise Areas

## Proteogenomic characterization of human colon and rectal cancer

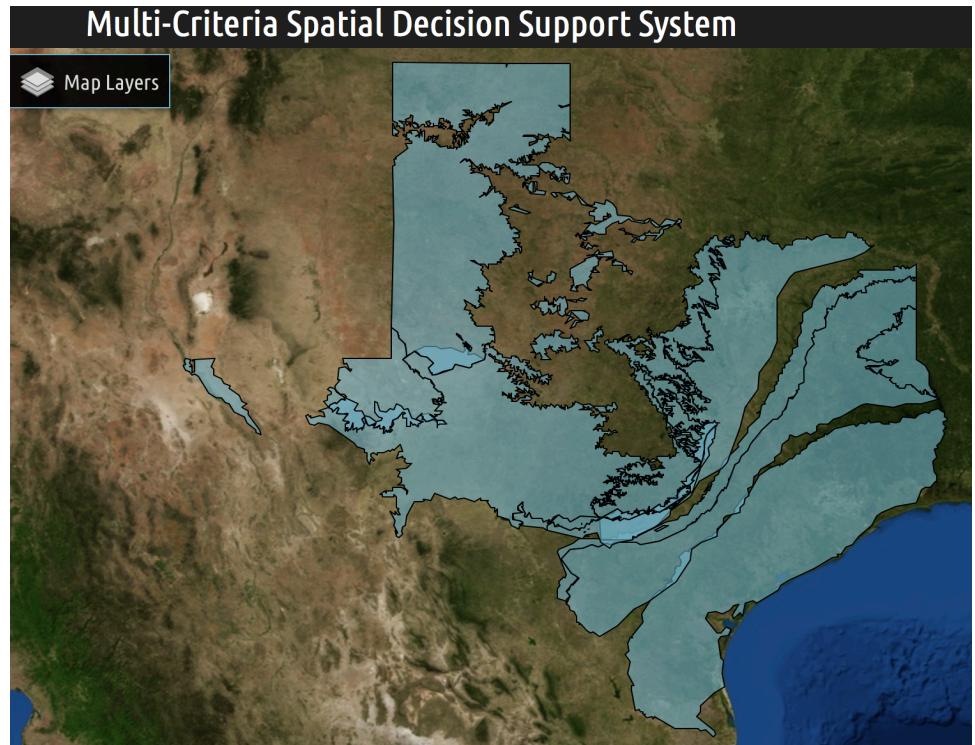
Bing Zhang<sup>1,2</sup>, Jing Wang<sup>1</sup>, Xiaojing Wang<sup>1</sup>, Jing Zhu<sup>1</sup>, Qi Liu<sup>1</sup>, Zhao Shi<sup>3,4</sup>, Matthew C. Chambers<sup>1</sup>, Lisa J. Zimmerman<sup>5,6</sup>, Kent F. Shaddox<sup>6</sup>, Sangtae Kim<sup>7</sup>, Sherri R. Davies<sup>8</sup>, Sean Wang<sup>9</sup>, Pei Wang<sup>10</sup>, Christopher R. Kinsinger<sup>11</sup>, Robert C. Rivers<sup>11</sup>, Henry Rodriguez<sup>11</sup>, R. Reid Townsend<sup>8</sup>, Matthew J. C. Ellis<sup>8</sup>, Steven A. Carr<sup>12</sup>, David L. Tabb<sup>1</sup>, Robert J. Coffey<sup>13</sup>, Robbert J. C. Slebos<sup>2,6</sup>, Daniel C. Liebler<sup>5,6</sup> & the NCI CPTAC\*



# Water Resource Modeling

*Work with Suzanne Pierce (University of Texas Austin)*

- Texas has over 33 diverse groundwater cases, can use with initial state conditions, parameter settings, and decision variables
- Different user groups (land use planning, environmental protection, and economic growth) have different analysis goals
- Automated process would customize the analysis



# Organic Data Science: Collaborative Workflow Development [Gil et al IUI 2015; ESWC 2015]

The screenshot illustrates the Organic Data Science platform for collaborative workflow development, featuring a central dashboard and various user interface components.

**Central Dashboard:** The top right shows a user profile for "John" with options to "Talk", "Preferences", "Watchlist", "Contributions", and "Log out". A search bar is also present. A callout box highlights "Your Overdue Tasks" with a red exclamation mark icon and the message "Write about the evaluation a day ago".

**Task Management:** The main area displays a task tree under "Draft paper about the initial framework design". The tree includes:

- Develop paper outline (100% complete)
- Draft initial versions of key sections (26% complete)
- Assemble first full draft of the paper
- Collect final evaluation data
- Finalize writing the paper (75% complete)

A timeline bar indicates progress from "TODAY" back to "0%".

**Task Properties:** A detailed view of the "Draft paper about the initial framework design" task is shown on the right, labeled "2a". It lists:

- Type<sup>M</sup>: medium
- Progress<sup>M</sup>: 21% (locked)
- Start date<sup>M</sup>: 22nd Aug 2014
- Target date<sup>M</sup>: 13th Oct 2014
- Owner: John Smith
- Participants: James Williams, Steven Johnson
- Expertise: computer science, collaboration

A legend at the bottom defines icons: M Mandatory, States: Not defined (grey), Valid (green), Inconsistent with parent (yellow).

**Workflow Actions:** On the left, a context menu for a task is open, labeled "2b", with options: Cut (highlighted with a cursor), Paste, Rename, Delete, and To Toplevel. A counter "38" indicates the number of tasks in the "My Tasks" list.

**Semantic MediaWiki Integration:** A logo for "Semantic MediaWiki" (SMW) is located in the bottom left corner.

**Text Overlay:** A descriptive text box at the bottom states: "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."

**Form Fields:** A "Properties" form at the bottom contains fields: "Submitted to" (IUI-2015) and "(by John)".

# Outline

---

1. The human bottleneck in data analytics
2. Related work on AI and cognitive aspects of scientific discovery
3. Semantic workflows to capture data analytics processes
4. Meta-reasoning to automate discovery
5. Discovery Informatics



# Discovery Informatics

Science Challenges for Intelligent Systems

<http://discoveryinformaticsinitiative.org>



## NSF Workshop on Discovery Informatics

February 2-3, 2012

Arlington, VA

## Final Workshop Report

August 31, 2012

USC Information Sciences Institute

Yolanda Gil

PSB Workshop on Discovery Informatics in Biological and Biomedical Sciences (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



# Discovery Informatics

SCIENCE [scinemag.org](http://scinemag.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,<sup>1</sup> Mark Greaves,<sup>2</sup>  
James Hendler,<sup>3\*</sup> Haym Hirsh<sup>4</sup>

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."***

# Discovery Informatics

RPR      Review of Policy Research

326

## The Promise and Potential of Big Data: A Case for Discovery Informatics

Vasant G. Honavar

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

PERSPECTIVE

COMPUTER SCIENCE

Automating Science

David Waltz<sup>1</sup>, Bruce G. Buchanan<sup>2</sup>

[+ Author Affiliations](#)

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





# 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*

**“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/>



# 2015 NSF Workshop on Intelligent Systems for Geosciences

<http://is-geo.org>

“Intelligent systems must incorporate existing scientific knowledge and the user’s context. This would enable novel forms of reasoning and learning about geosciences data.”

### Geospatial Reasoning

**Geospatial Pattern Matching: Discovering Flow Anomalies**

- Scalable geospatial temporal pattern matching
- Retrospective detection of when contaminants entered an ecosystem

### Machine Learning

**Pattern Mining: Monitoring Ocean Eddies**

- Spatio-temporal pattern mining of satellite data using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks

**Network Analysis: Climate Teleconnections**

- Scalable method for discovering related graph regions
- Discovery of novel climate teleconnections

### Robotics

**Offline Models from AUV data: Models of Coastal Zones**

- Georeferenced mapping and 3D reconstruction
- Long-term autonomy for AUV gliders includes in-situ mass-spectrometry

### Information Integration

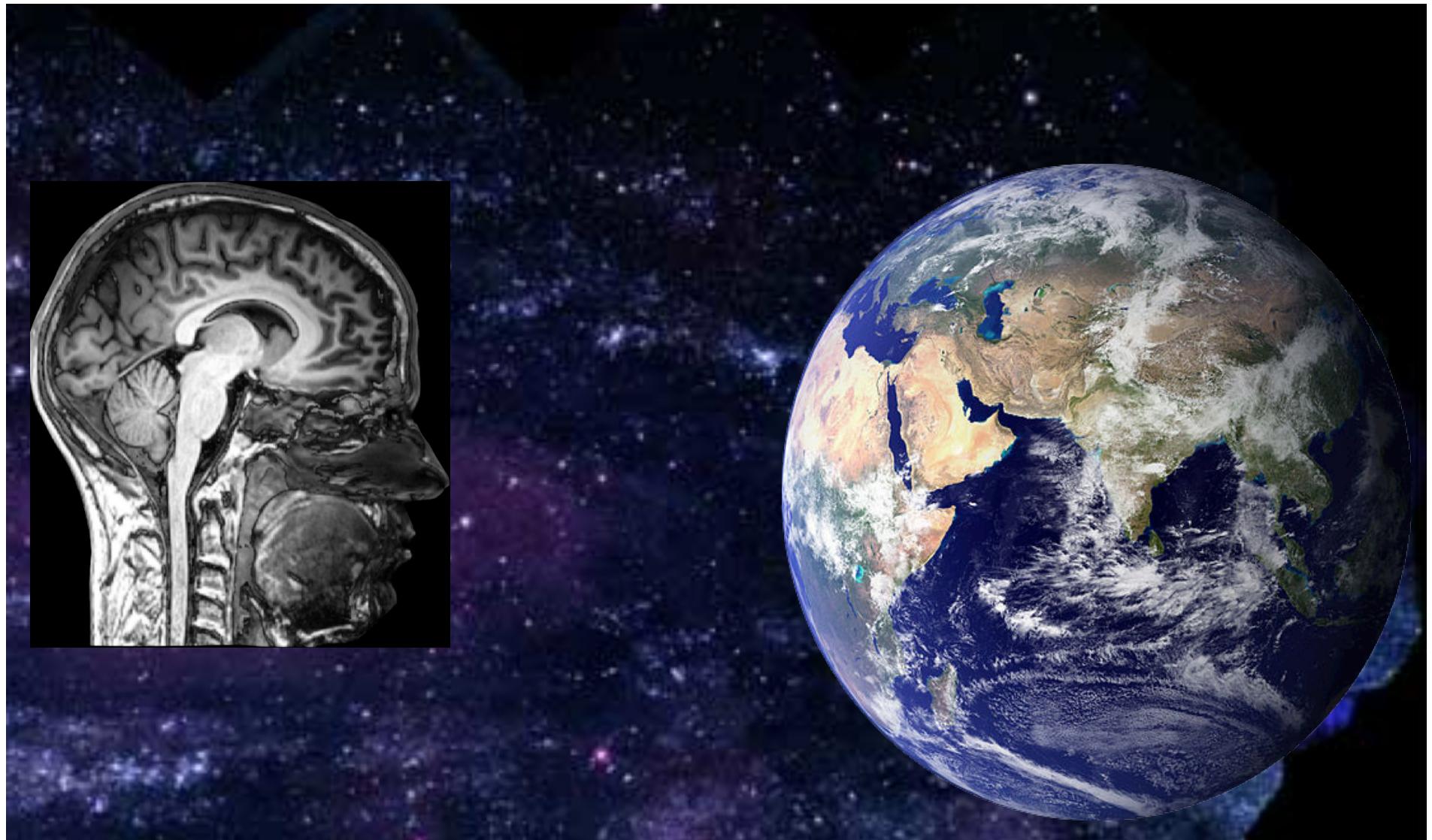
**Semantic Metadata: Entity Linking Across Data Sources**

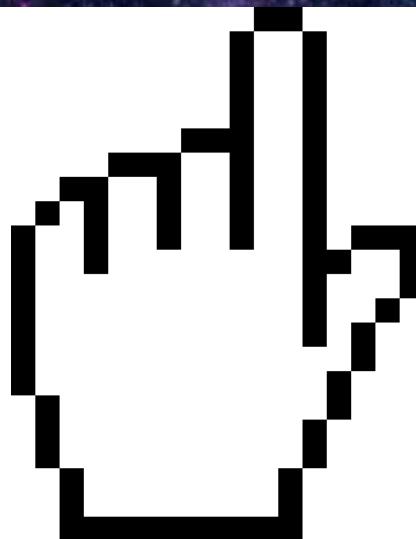
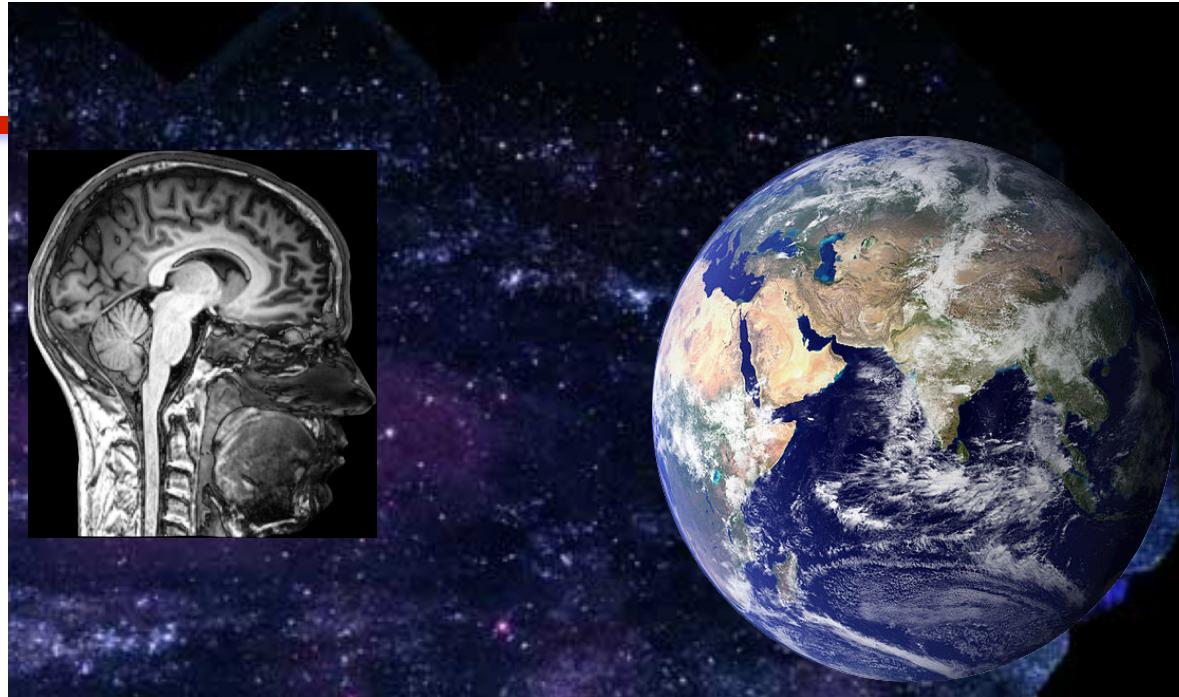
- Name-based and structure-based mapping of entities
- Semi-automatic integration of diverse data sources

### Augmented Reality

**Tablet-based Augmented Reality: Exploring Remote Locations**

- Low-cost tablet-based virtual reality displays
- Virtual presence in inaccessible or previously visited locations





# 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), Neda Jahanshad (USC)
- *Biomedical workflows:* Phil Bourne & Sarah Kinnings (UCSD), Parag Mallick (Stanford U.) Chris Mason (Cornell), Joel Saltz & Tahsin Kurk (Emory U.), Jill Mesirov & Michael Reich (Broad), Randall Wetzel (CHLA), Shannon McWeeney & Christina Zhang (OHSU)
- *Geosciences workflows:* Chris Duffy (PSU), Paul Hanson (U Wisconsin), Tom Harmon & Sandra Villamizar (U Merced), Tom Jordan & Phil Maechlin (USC), Kim Olsen (SDSU)
- *And many others!*