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

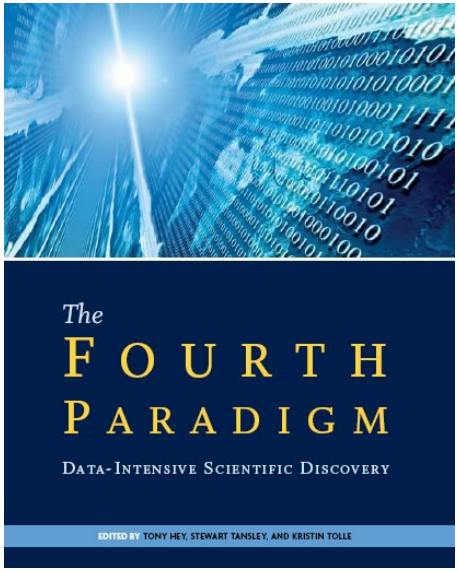
- Human bottleneck

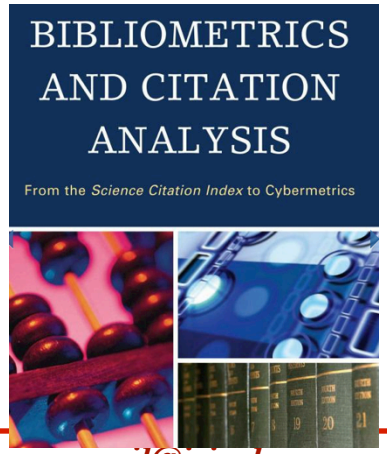
Great opportunities for cognitive systems

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

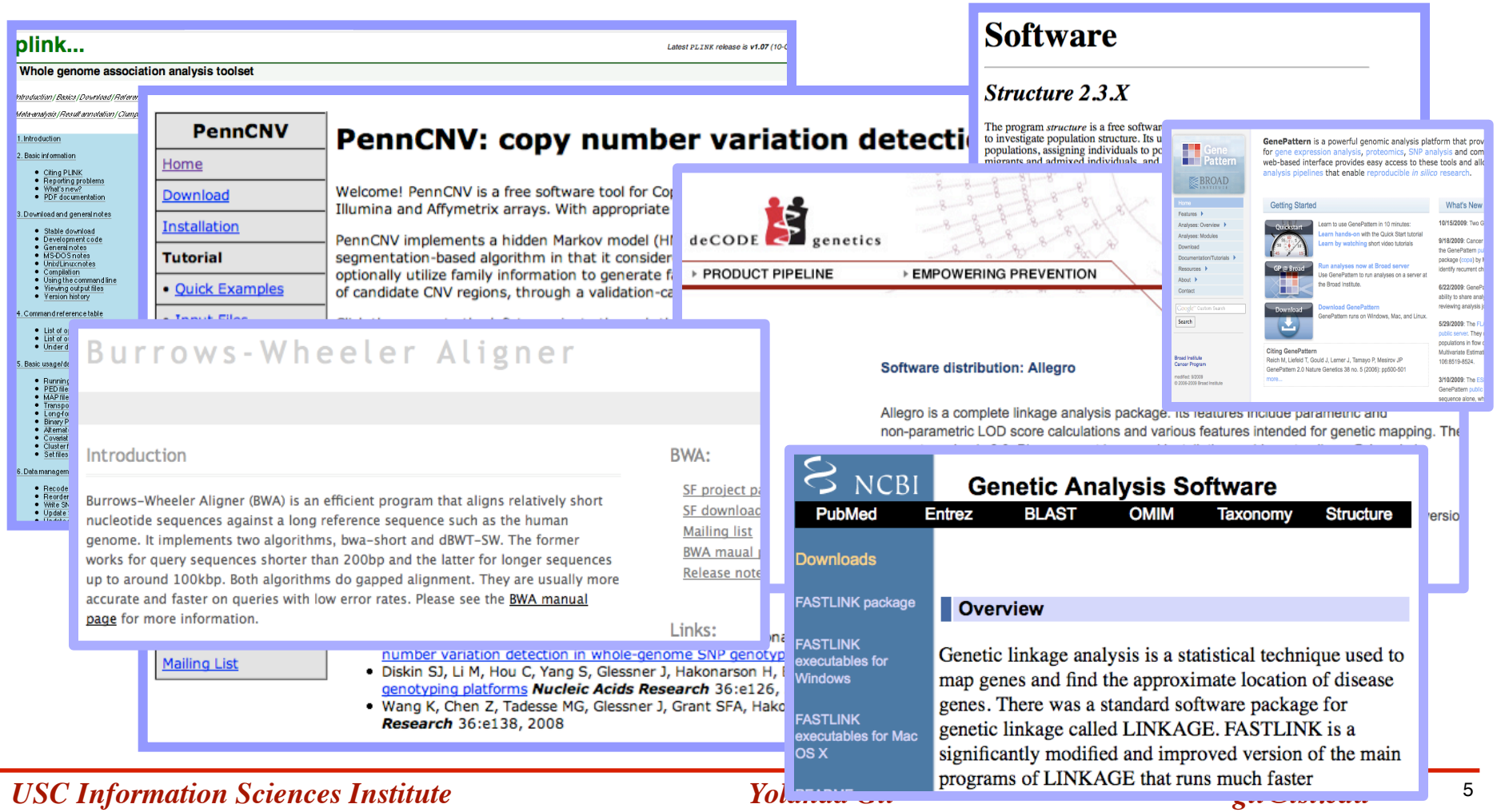
Data-Intensive Computing in Science





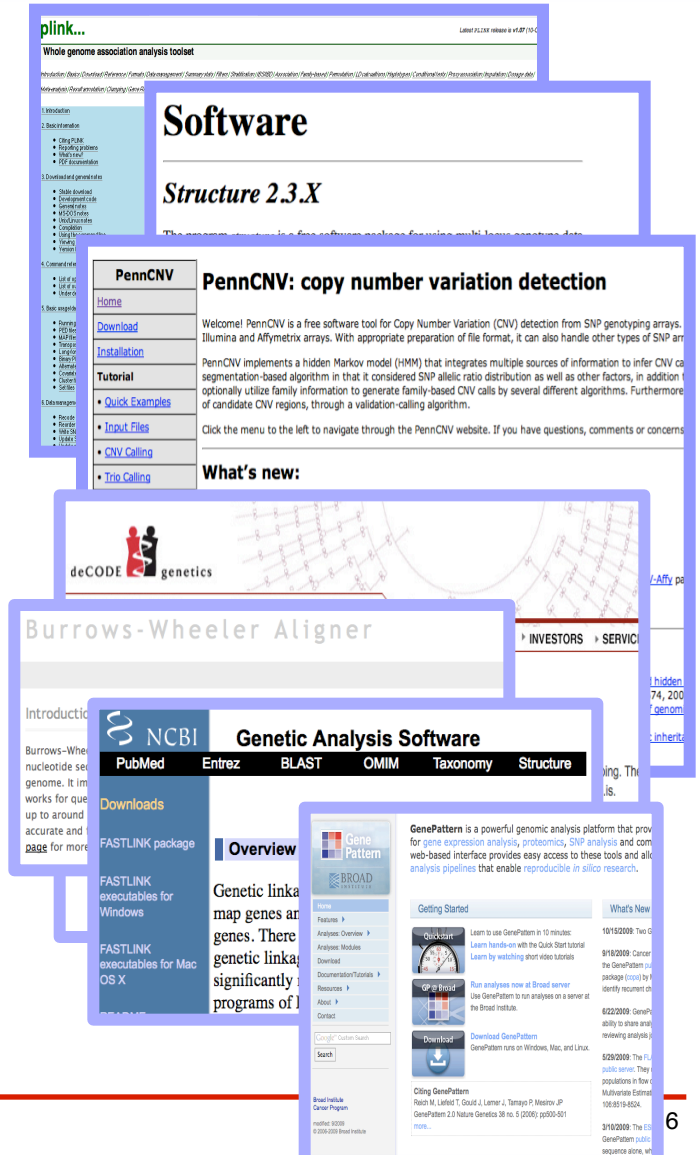
Scientific Data Analysis

- Complex processes involving a variety of algorithms/software



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)

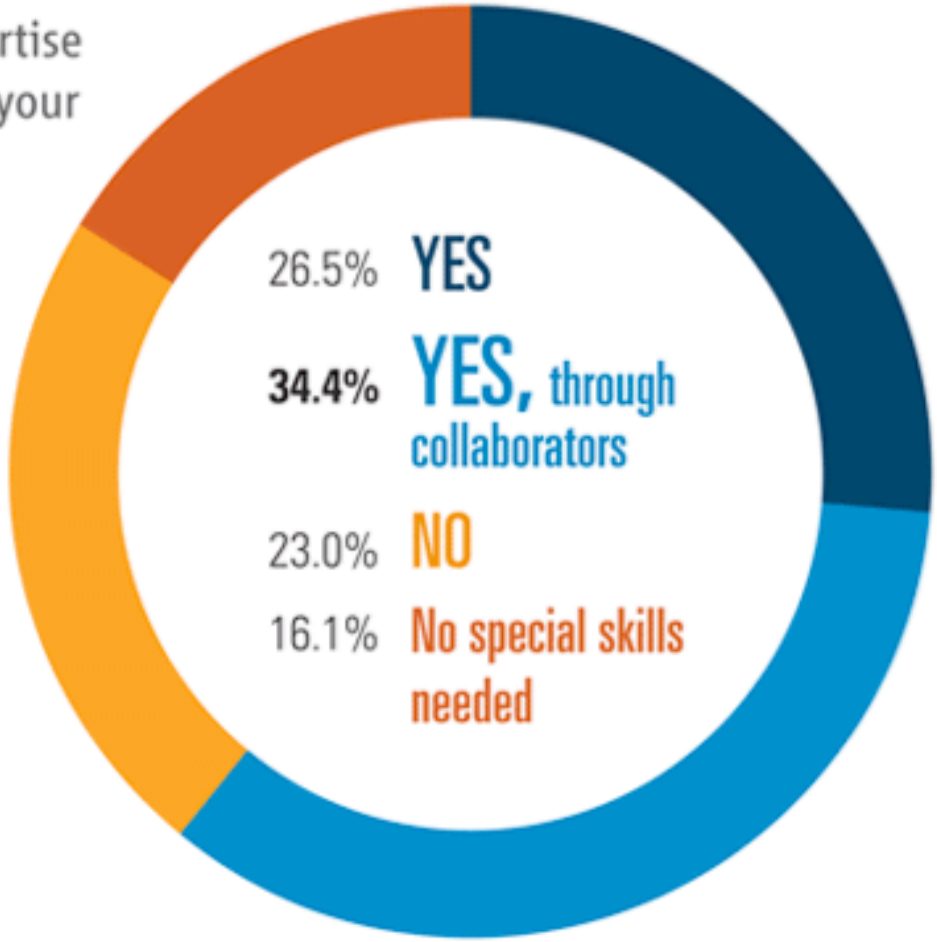


Problems (II): Reproducibility

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?

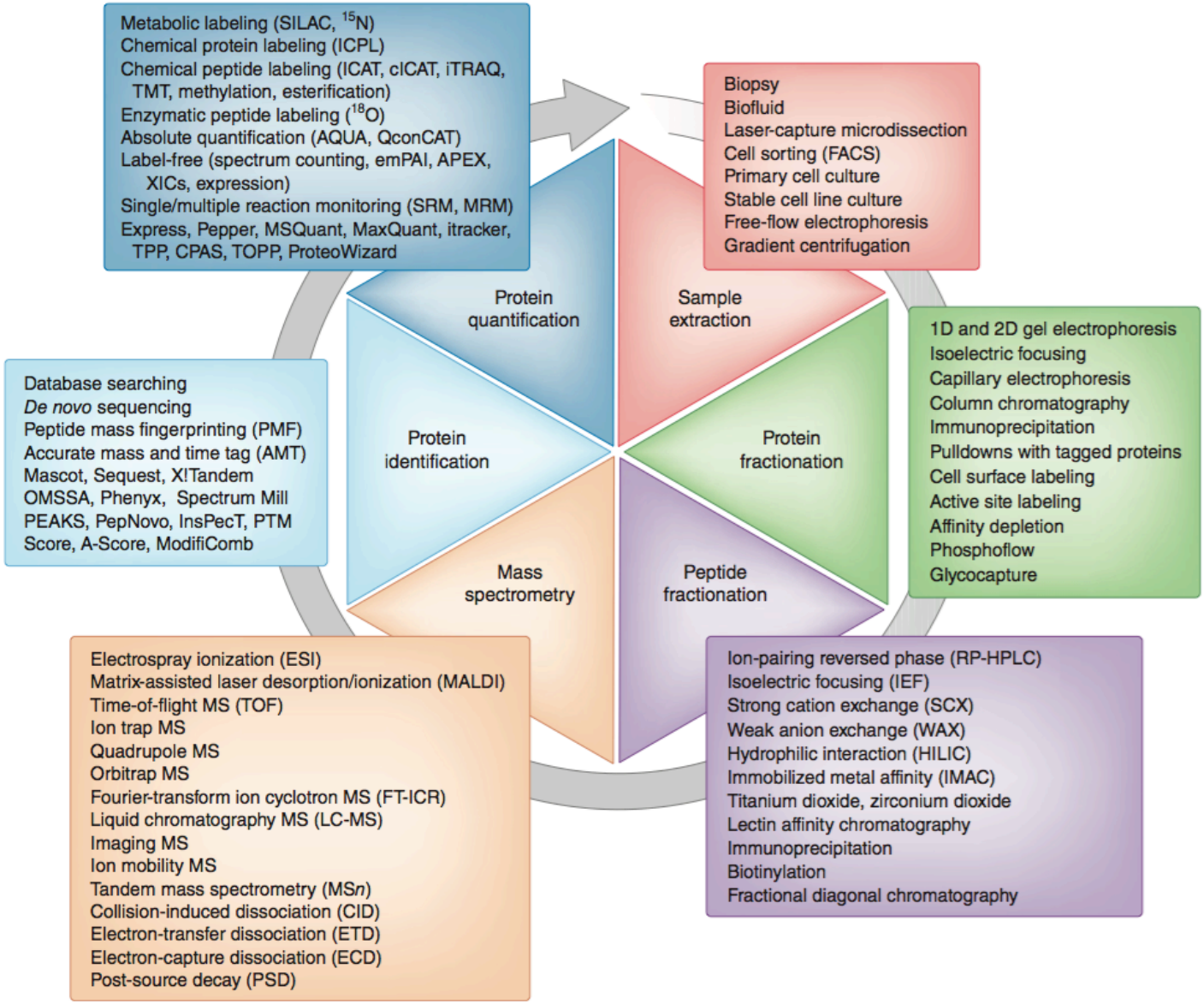


"The next few years [particularly in medicine] **the volume of data we need to analyze will expand exponentially.**"

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

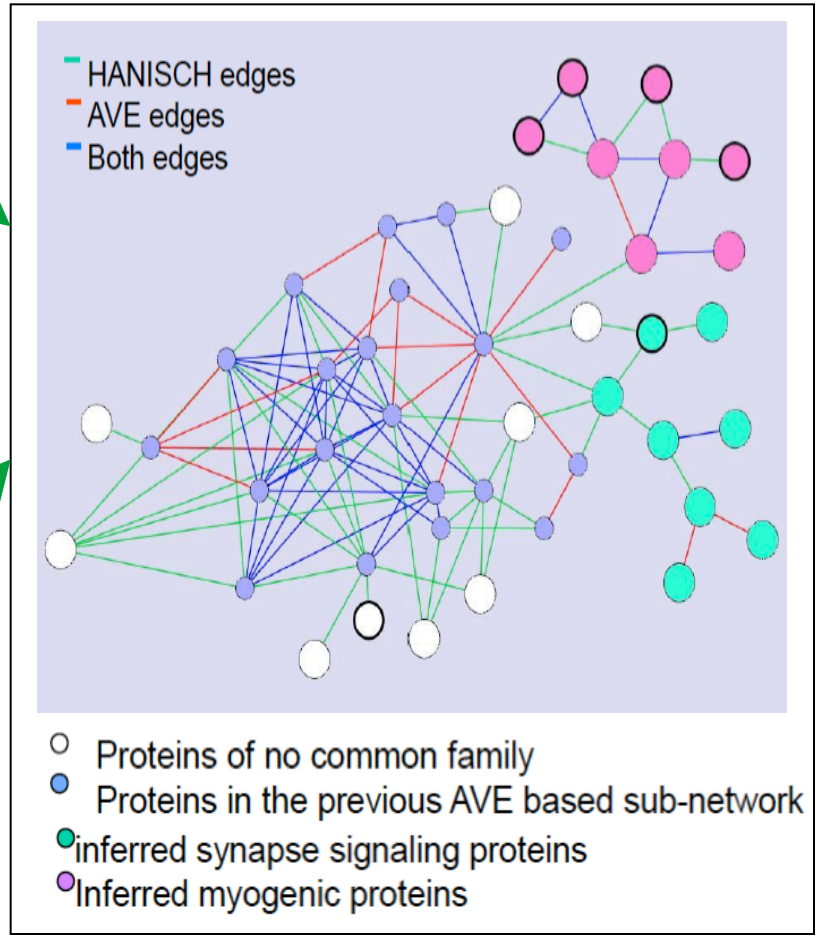
Text Extraction in Hanalyzer (L. Hunter, U. Colorado)

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 α2 (GlyR α2) [23]. The ability of Nova-1 to activate exon selection in neurons is antagonized by a second RNA-binding protein, brPTB (brain-enriched polypyrimidine tract-binding protein), which interacts with Nova-1 and inhibits its function [24]. DAZAP1 could function in a similar manner by binding to DAZL and inhibiting its function. Comparing the phenotypes of Dazl1 and Dazap1 single and double knock-out mice may provide some clues to the significance of their interaction. Dazl1 knock-out mice have already been generated and studied [6]. The spermatogenic defect in the male becomes apparent only after day 7 post partum when the germ cells are committing to meiosis (H. Cooke, personal communication). The genomic structure of Dazap1, delineated here, should facilitate the generating of Dazap1 null mutation.

Semantic integration of biomedical databases

Generation of interesting new hypotheses



Robot Scientist [King et al 2009]

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

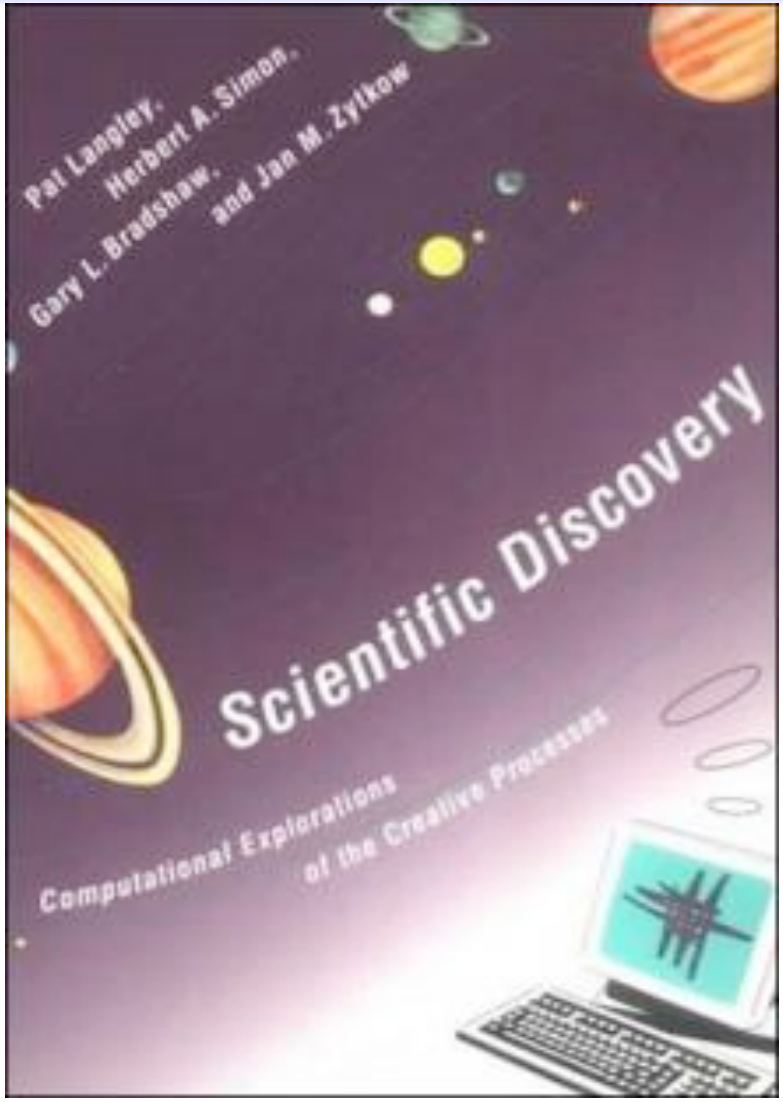
REPORT

The Automation of Science

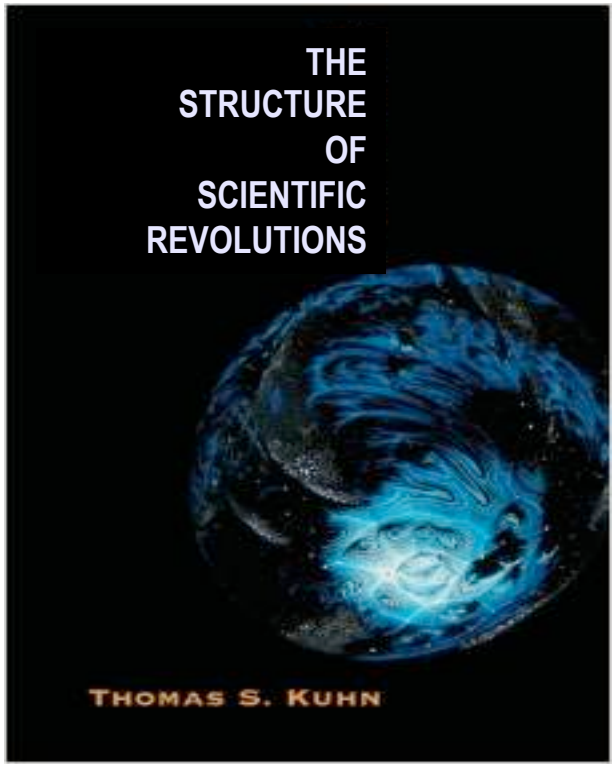
Ross D. King[1,*], Jem Rowland[1], Stephen G. Oliver[2], Michael Young[3], Wayne Aubrey[1], Emma Byrne[1], Maria Liakata[1], Magdalena Markham[1], Pinar Pir[2], Larisa N. Soldatova[1], Andrew Sparkes[1], Kenneth E. Whelan[1], Amanda Clare[1]

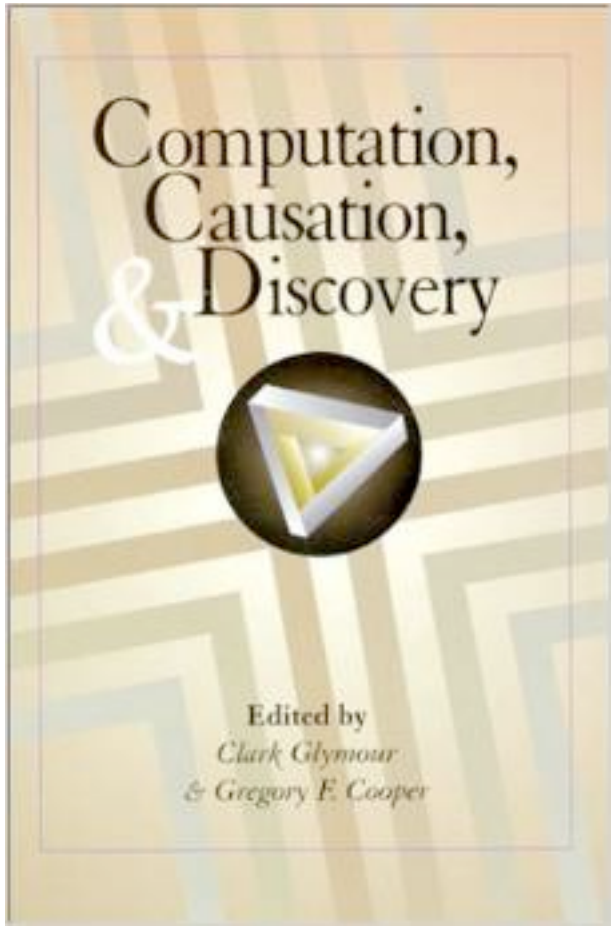
Computational Scientific Discovery

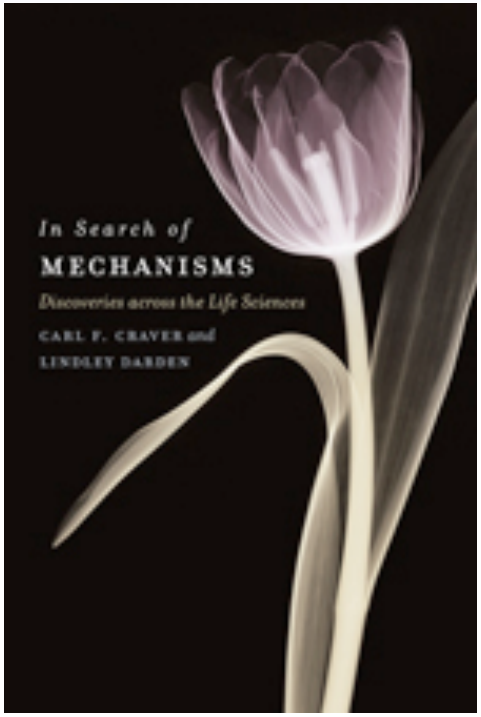
- [Lenat 1976]
- [Lindsay et al 1980]
- [Langley 1981]
- [Falkenhainer 1985]
- [Kulkarni and Simon 1988]
- [Cheeseman et al 1989]
- [Zytkow et al 1990]
- [Simon 1996]
- [Valdes-Perez 1997]
- [Todorovski et al 2000]
- [Schmidt and Lipson 2009]



Philosophy of Science



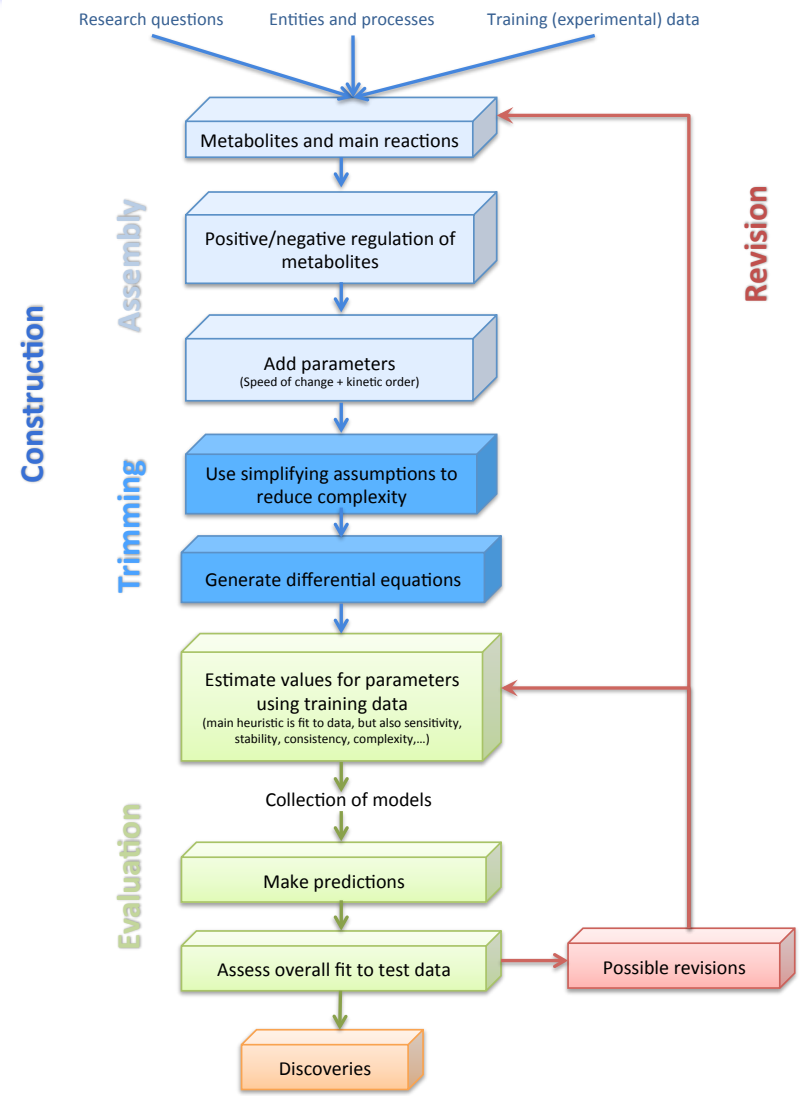




Cognitive Science

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?

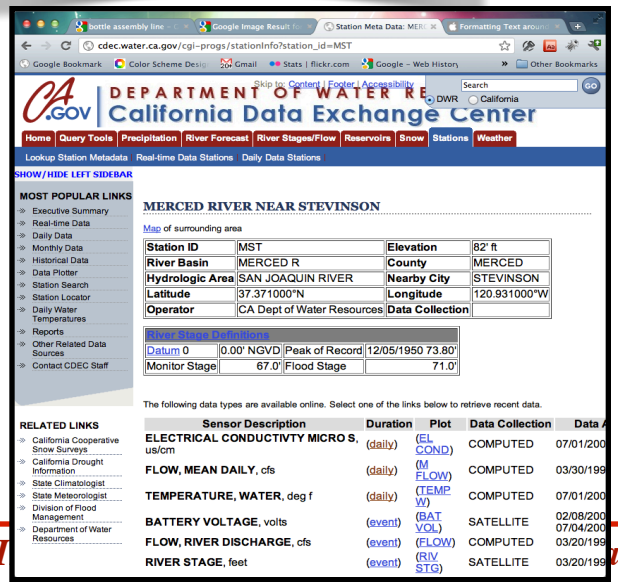
Outline

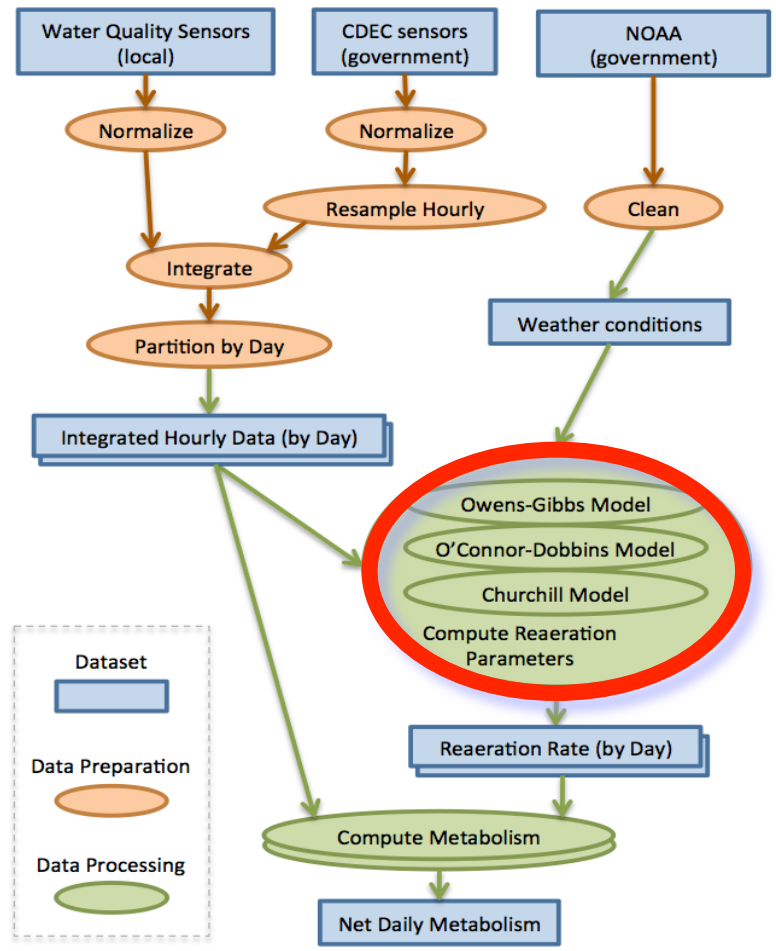
1. The human bottleneck in data analytics
2. Related work on AI and cognitive aspects of scientific discovery
3. Semantic workflows to capture data analytics processes
4. Meta-reasoning to automate discovery
5. Discovery Informatics

Timely Analysis of Environmental Data
[Gil et al ISWC 2011]

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

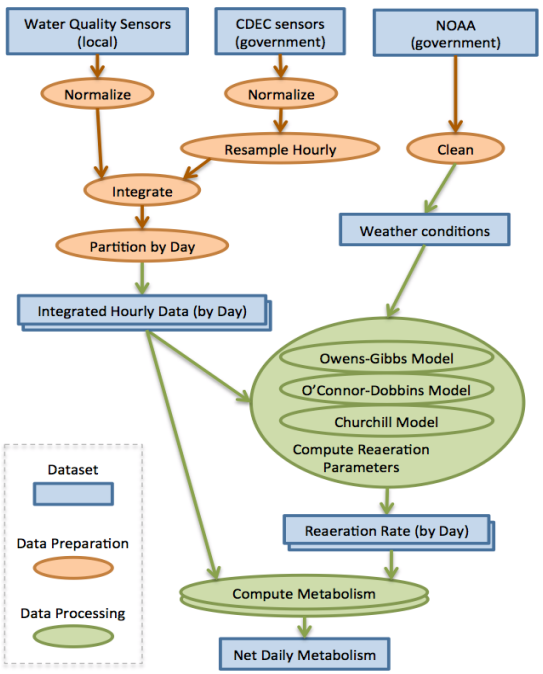
California's Central Valley:
- **Farming, pesticides, waste**
- **Water releases**
- **Restoration efforts**

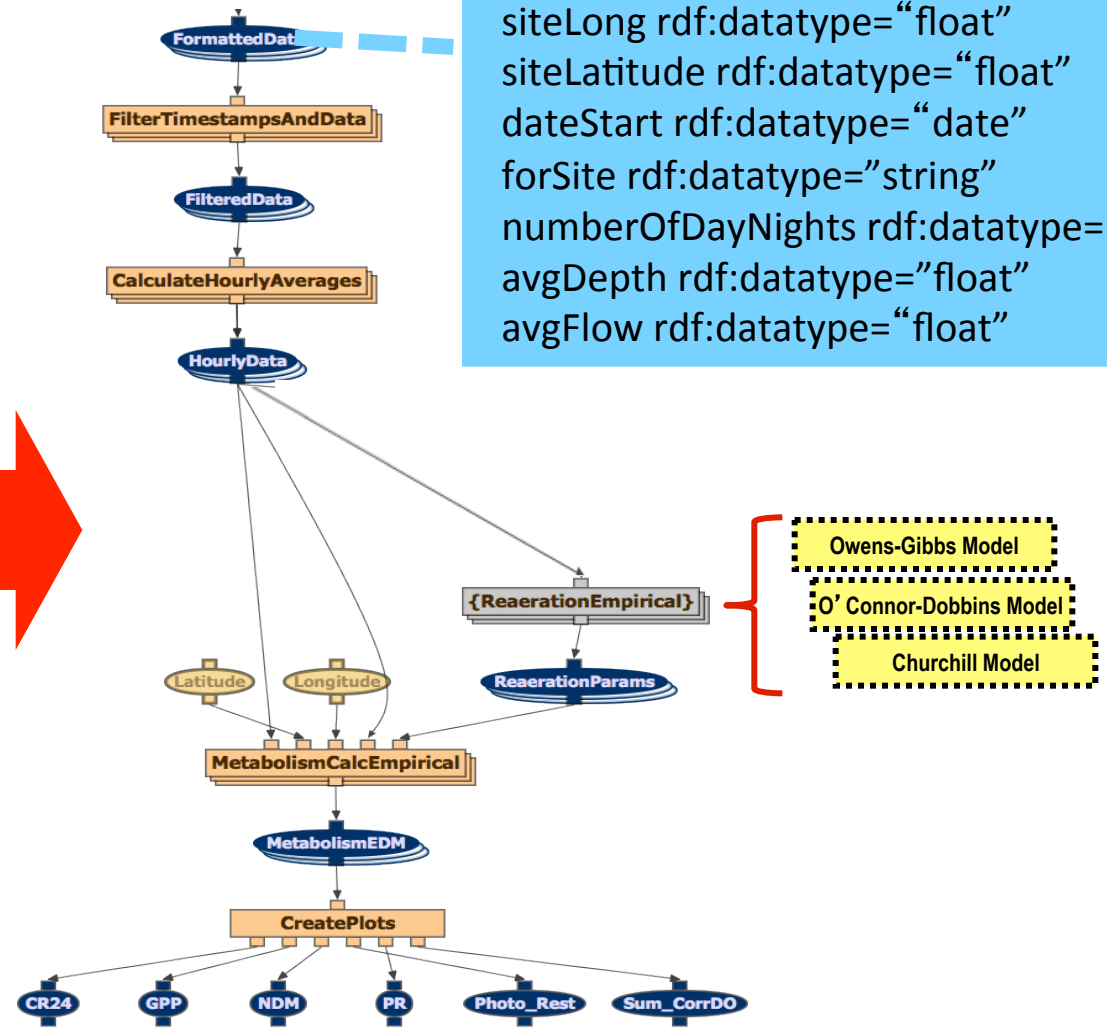




A Semantic Workflow

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"

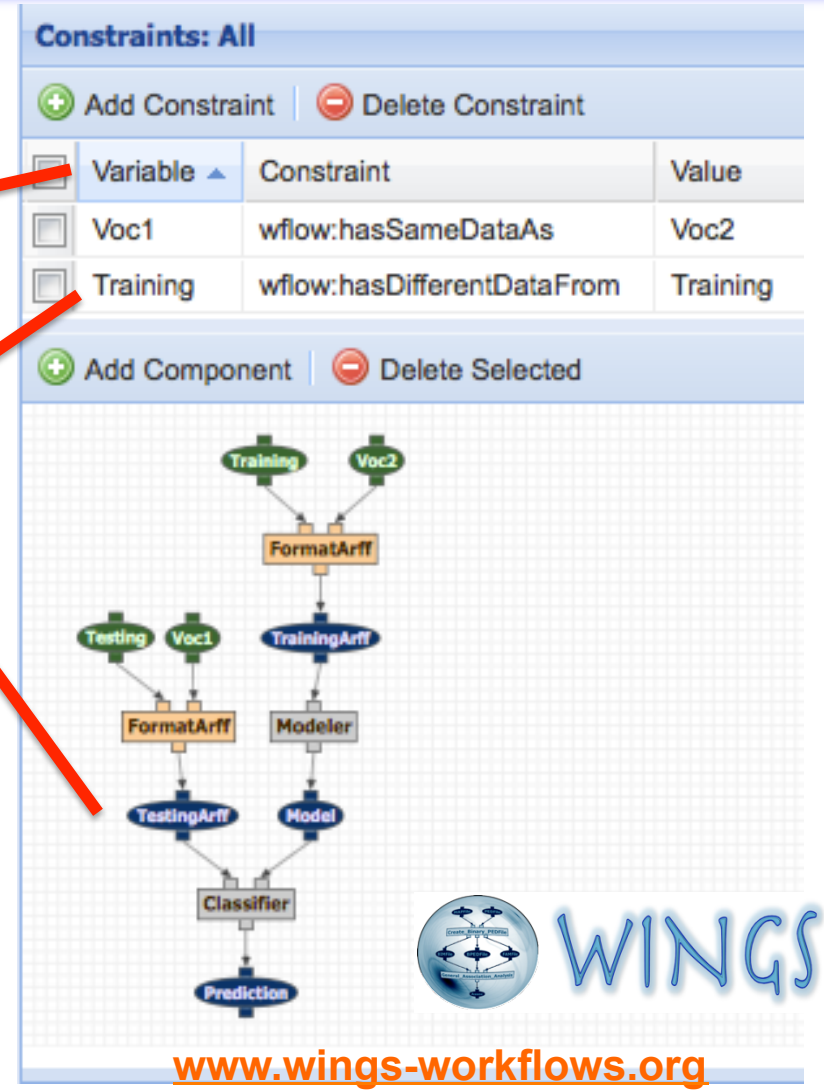




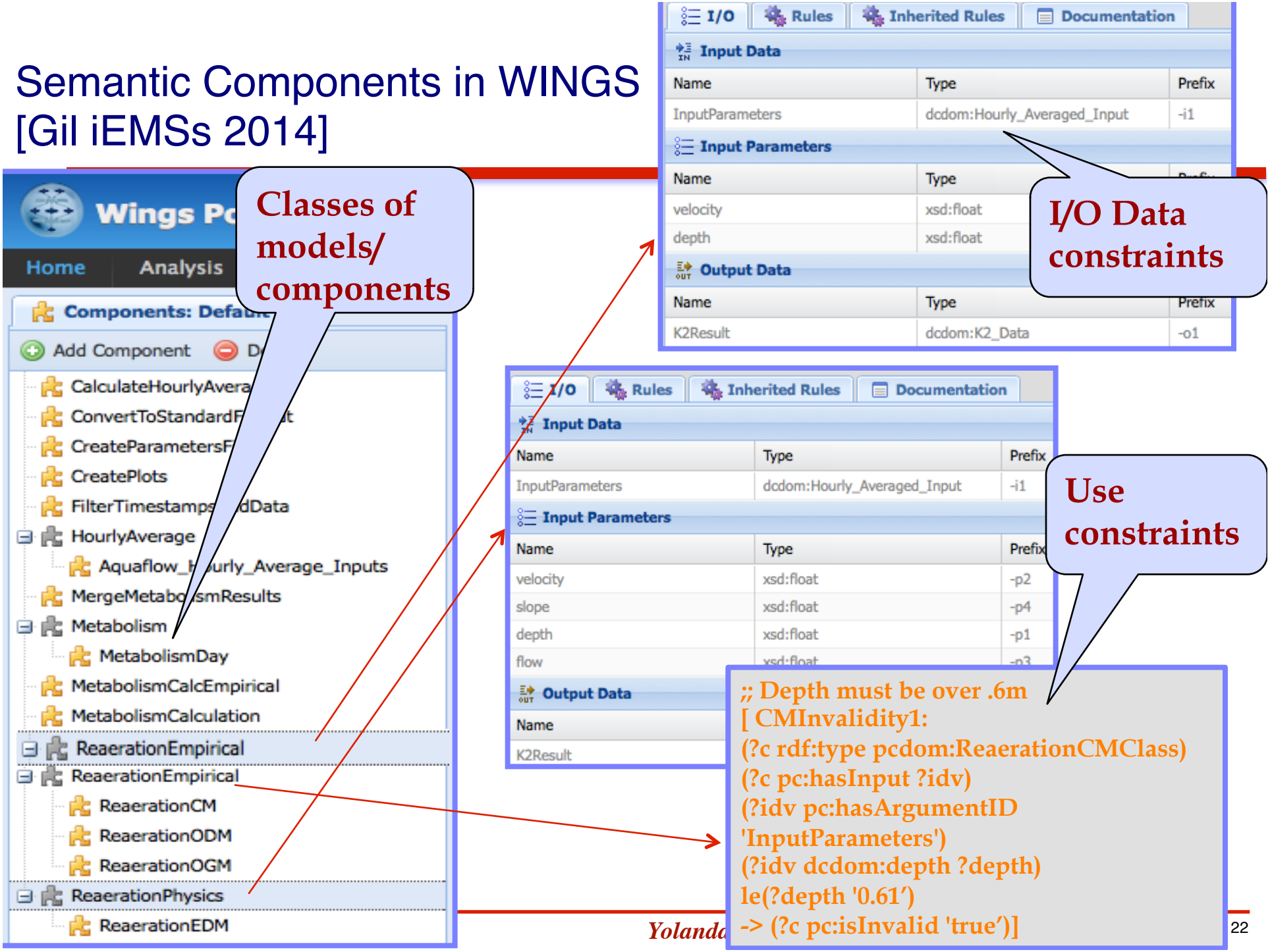
Semantic Workflows in Wings

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

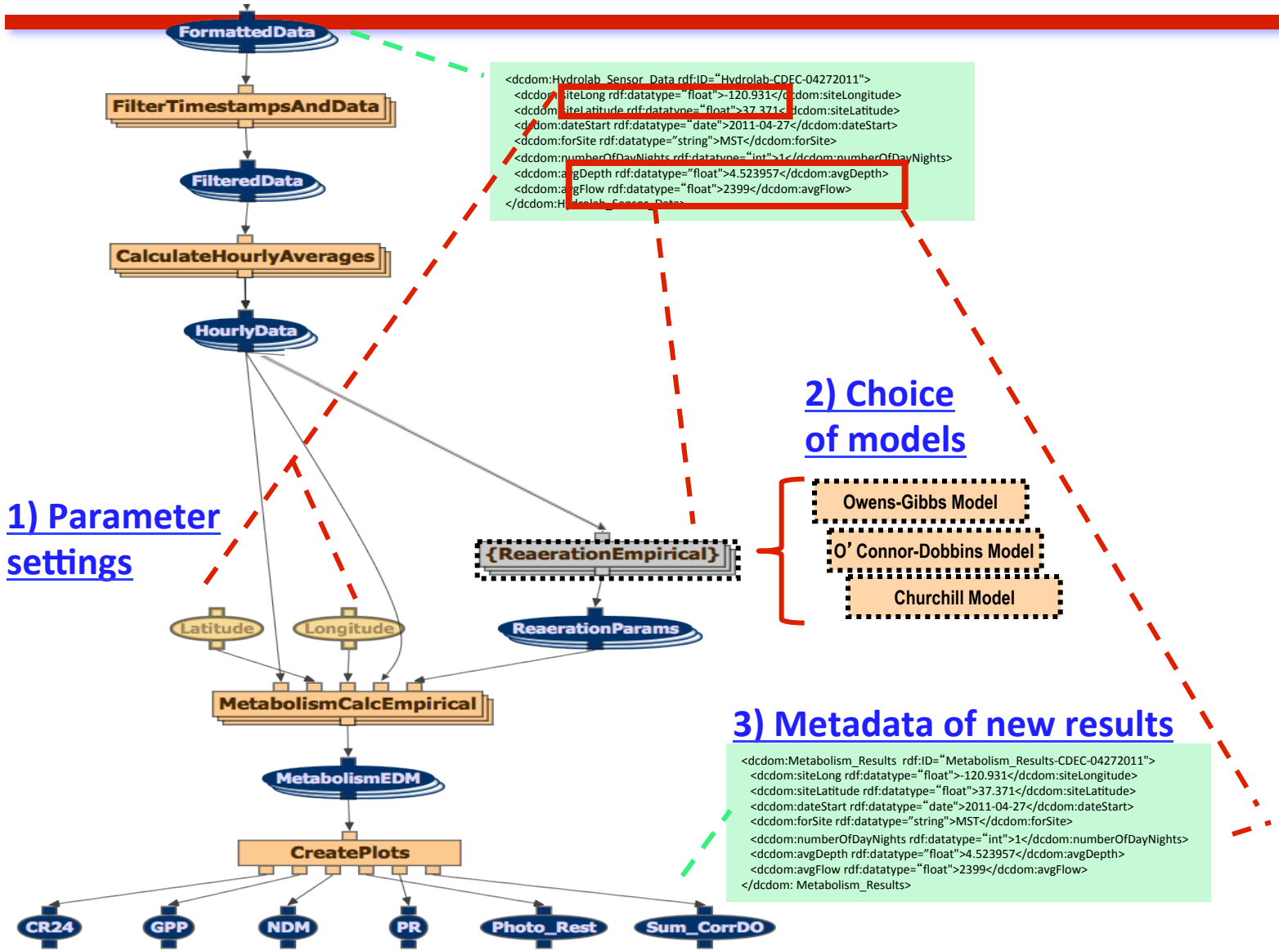
- **Workflows are augmented with semantic constraints**
  - Each workflow constituent has a **variable** associated with it
    - Workflow components, arguments, datasets
  - **Constraints** are used to restrict workflow variables
  - Can define **abstract classes of components**
    - Concrete components model exec. codes
- **Workflow reasoners propagate and use semantic constraints**
- **Uses semantic web standards: OWL/ RDF, SPARQL**
- Compilation of workflows to scalable execution infrastructure



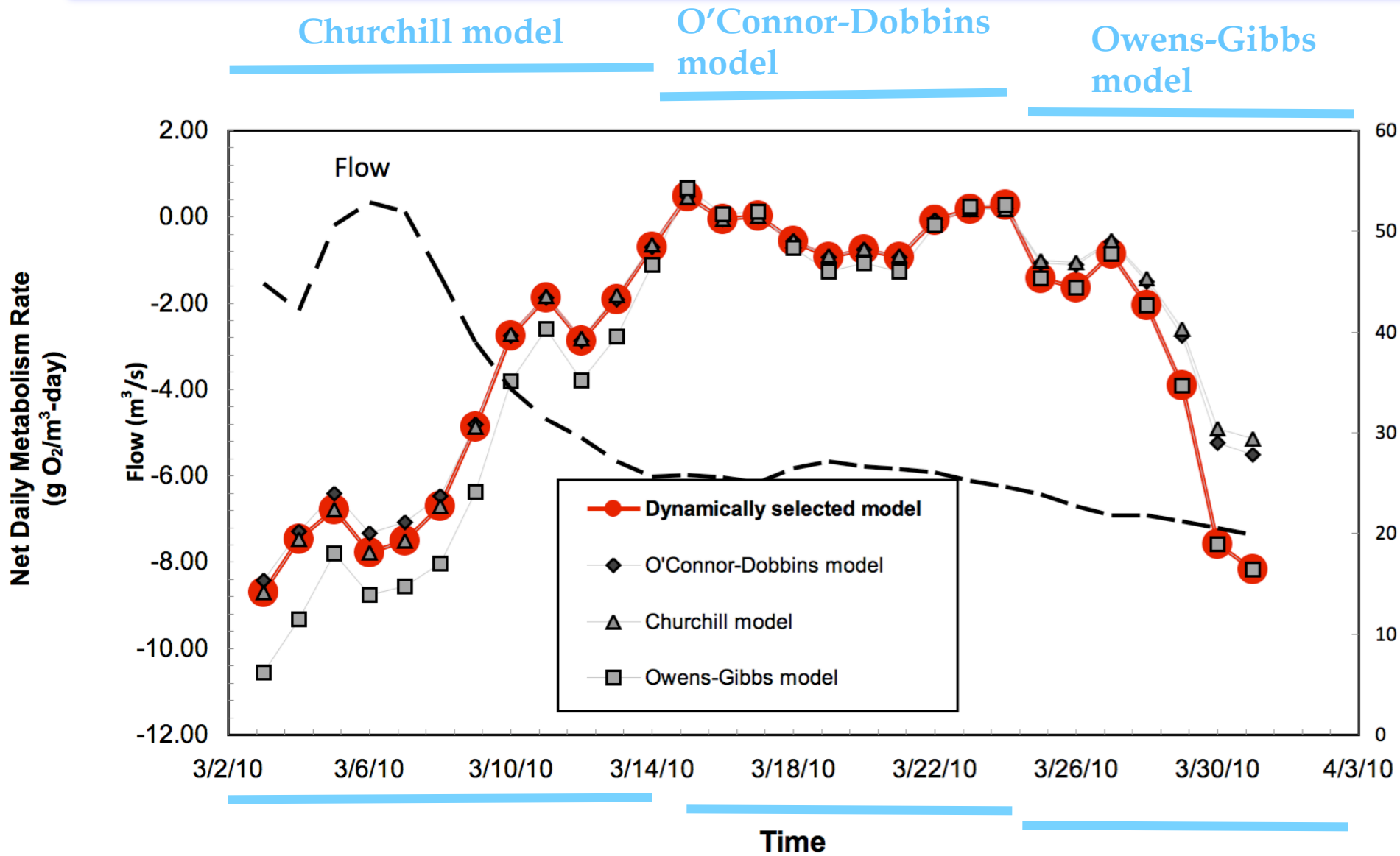
Semantic Components in WINGS [Gil iEMSs 2014]



WINGS Specializes Workflow Based on Characteristics of Daily Data



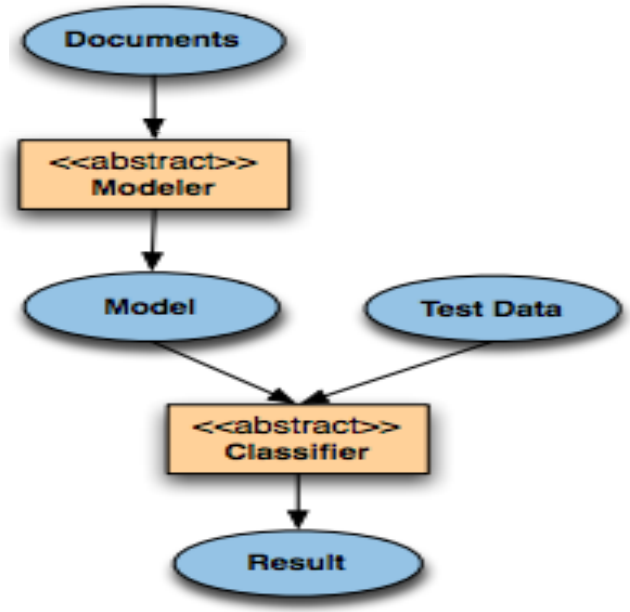
WINGS Dynamically Selects Appropriate Model Based on Daily Sensor Readings



Workflows Capture Data Analytics Expertise
[Hauder et al e-Science 2011]

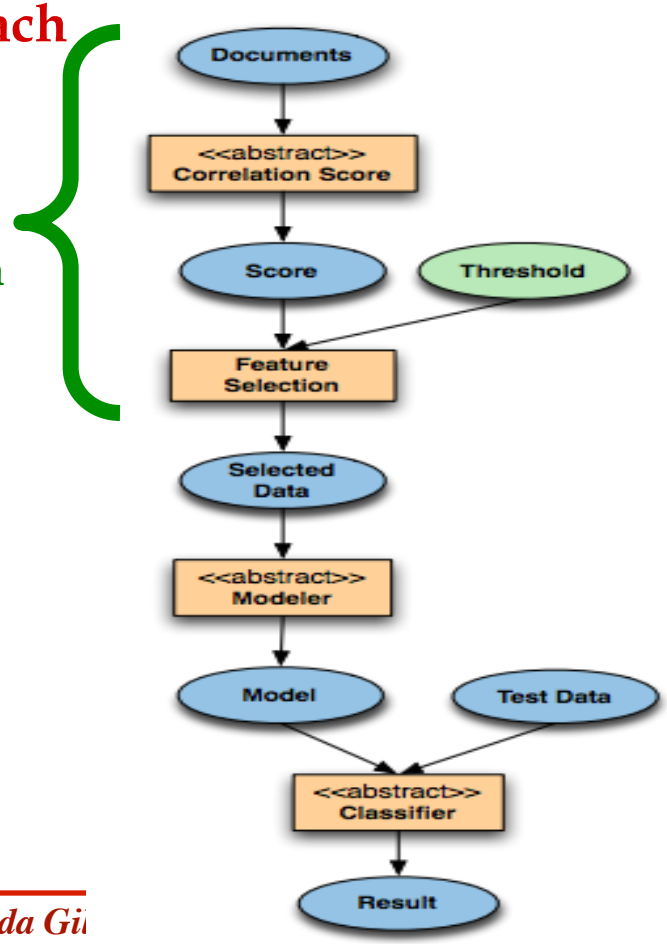
Workflows for text analytics, joint work with Yan Liu (USC) and Mattheus Hauder (TUM)

Naïve Approach

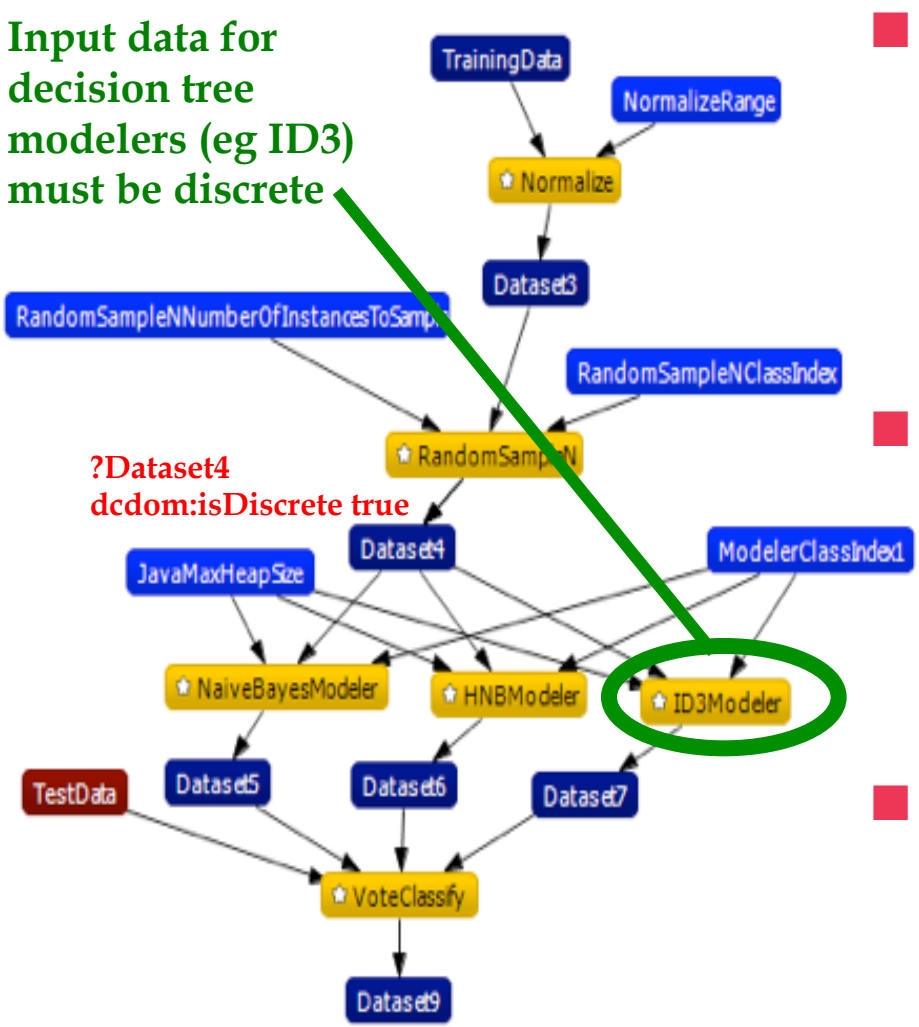


Expert Approach

Feature selection



WINGS Workflow Reasoners

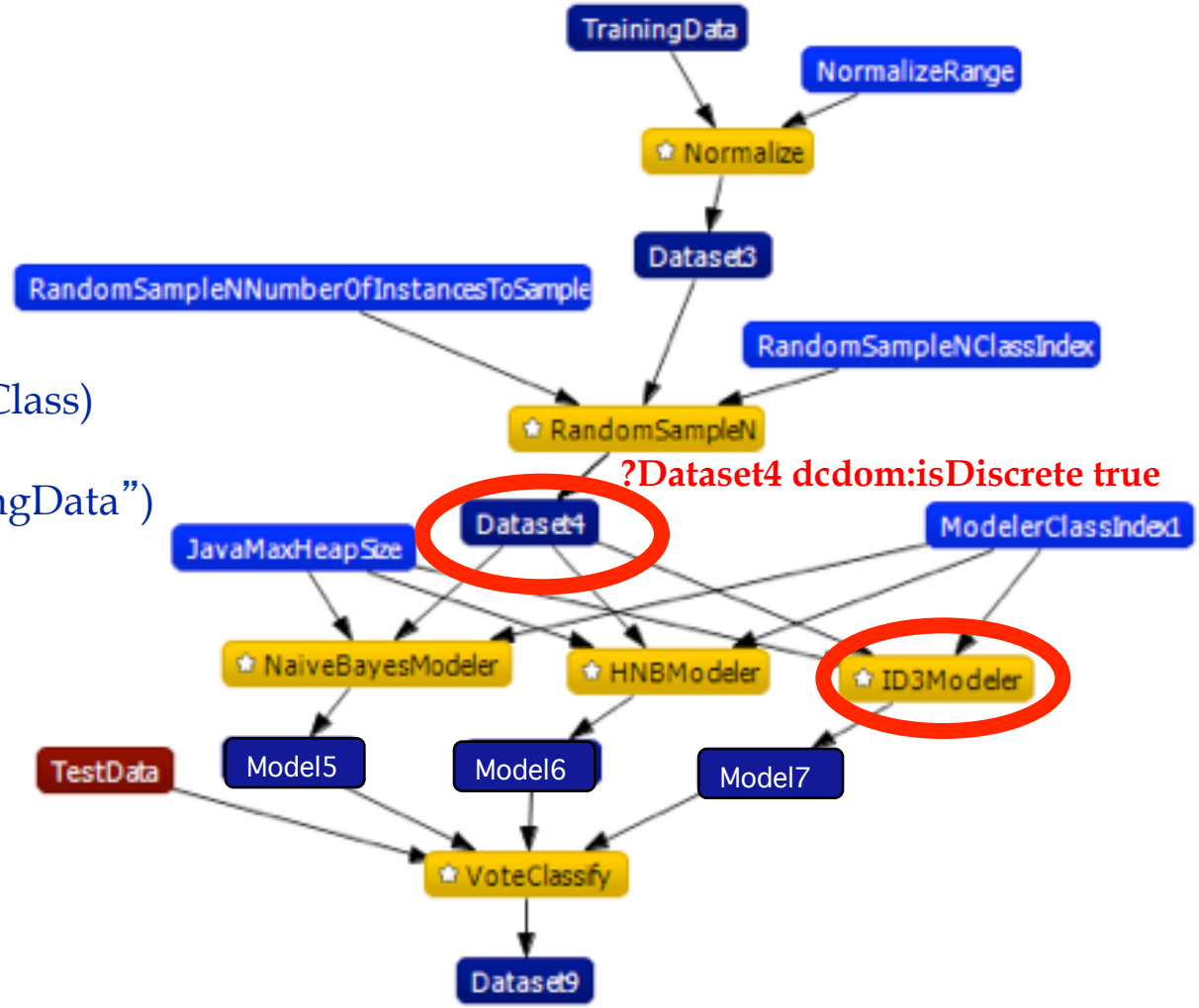


- **Key idea:** Skeletal planning, where constraints for each component are propagated through a fixed workflow structure (the skeleton)
- **Phase 1: Goal Regression**
  - Starting from final products, traverse workflow backwards
  - For each node, query for constraints on inputs
- **Phase 2: Forward 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")

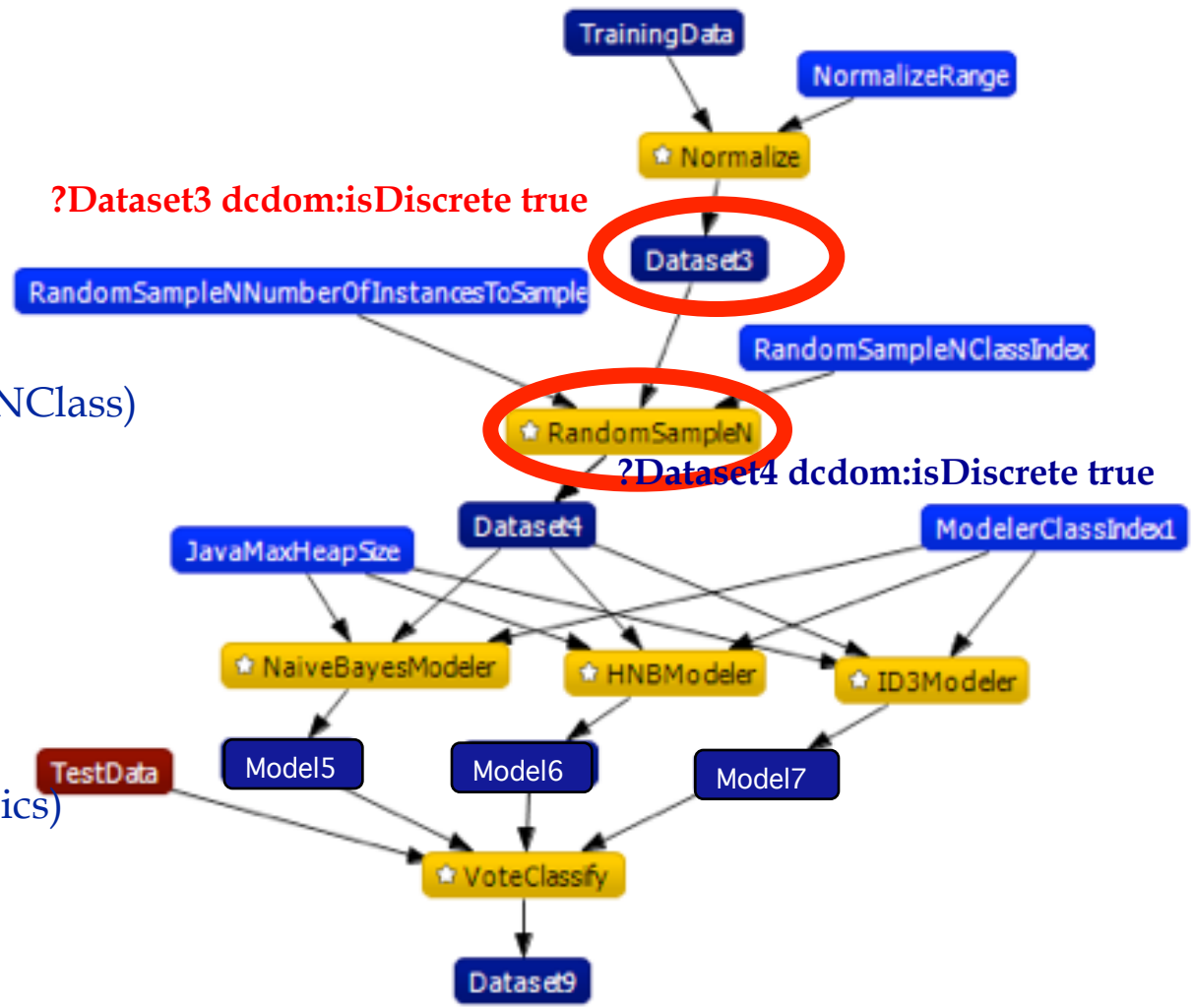
-> (?idv dcdom:isDiscrete "true"^^xsd:boolean)]



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)

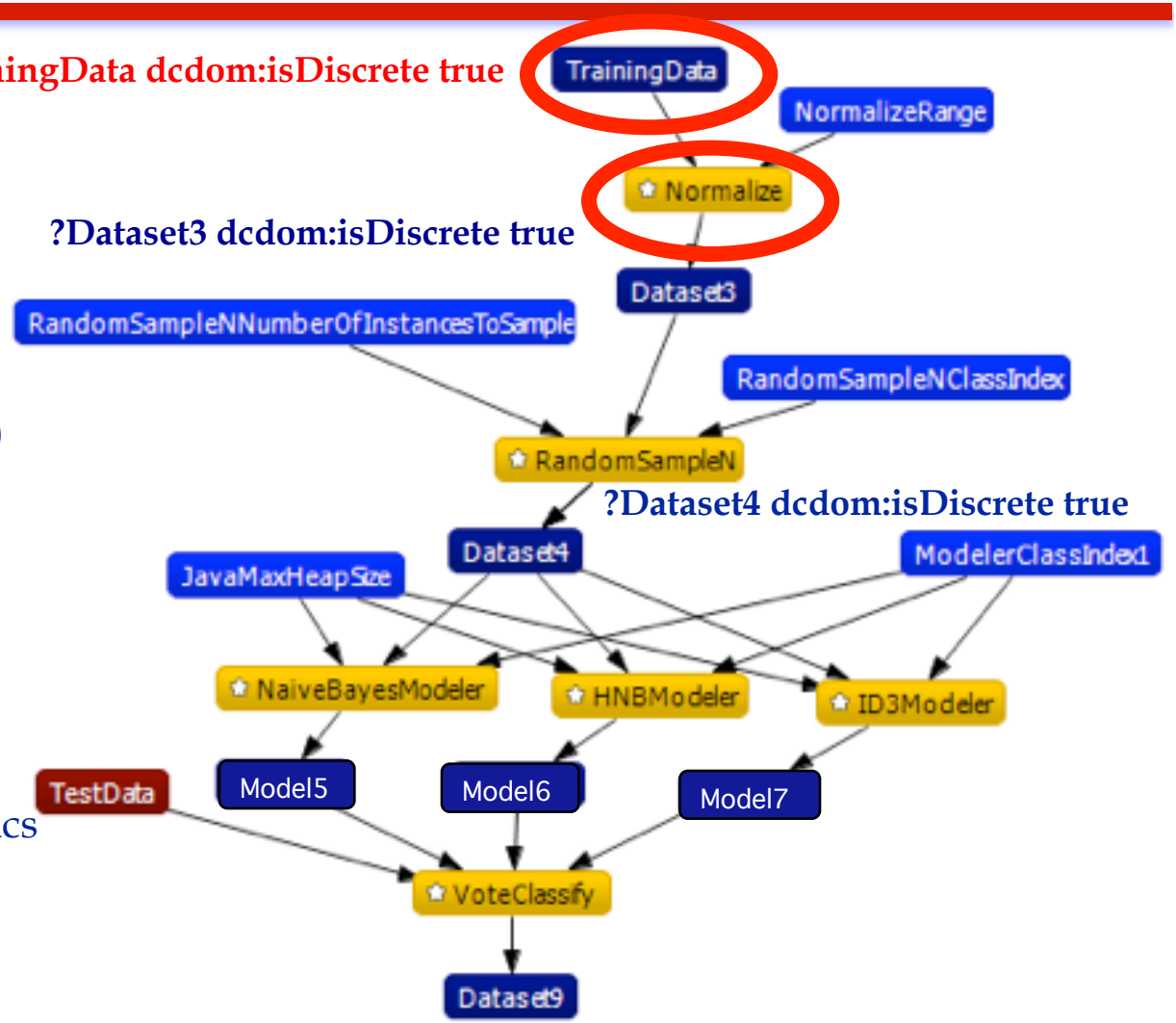
-> (?idv ?p ?val)]



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

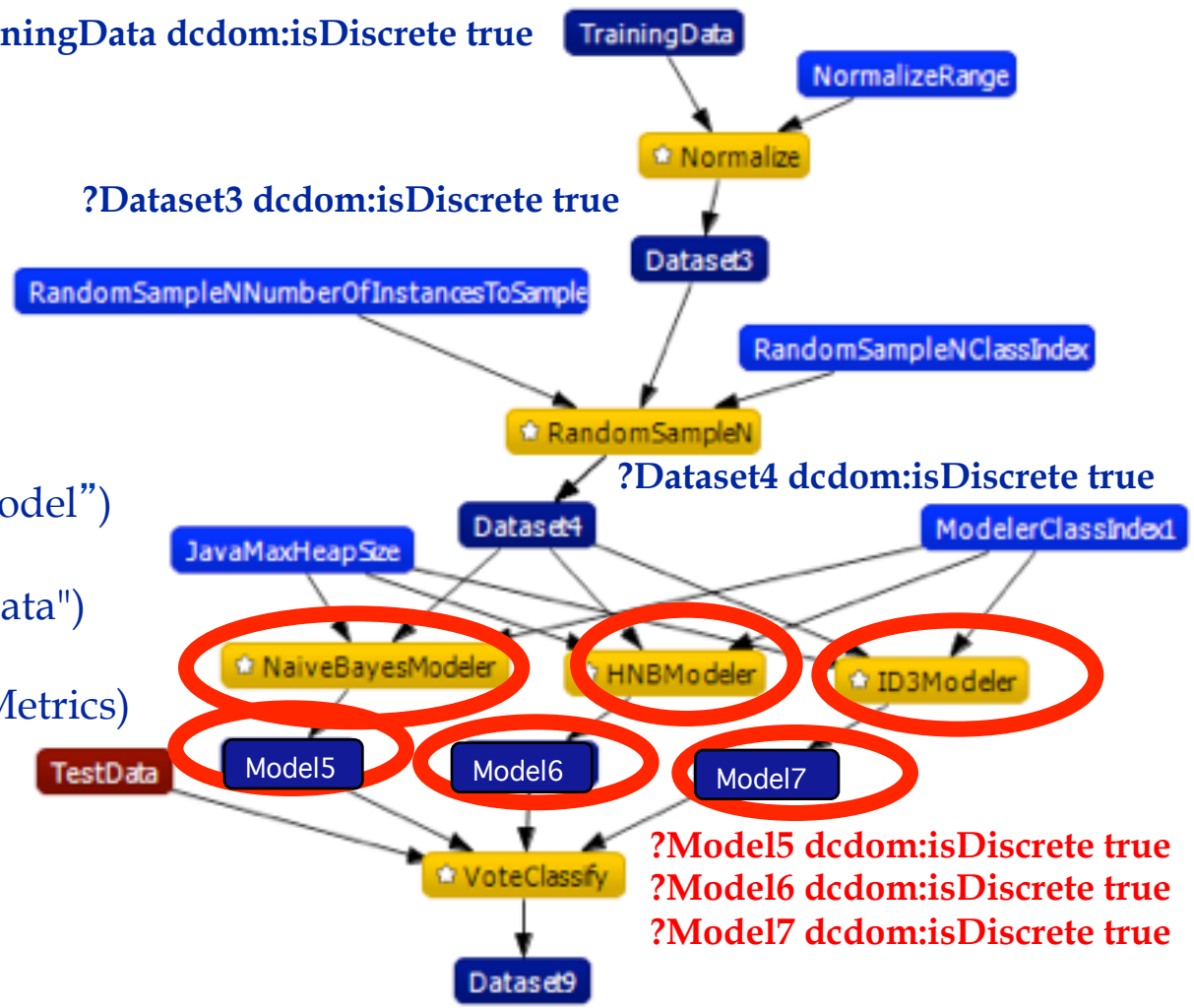
-> (?idv ?p ?val)]



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 dcdom:isSampled)

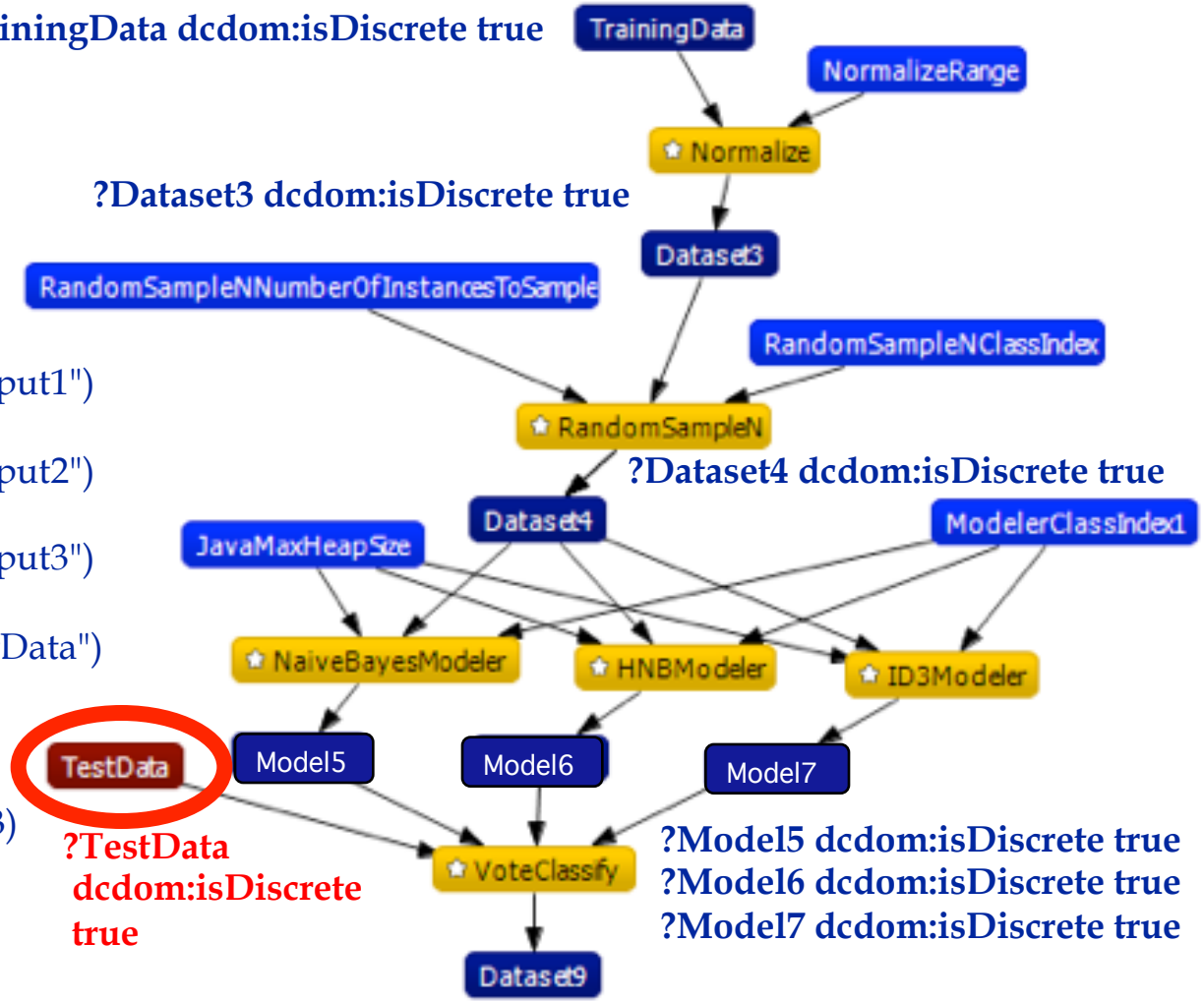
-> (?odv ?p ?val)]



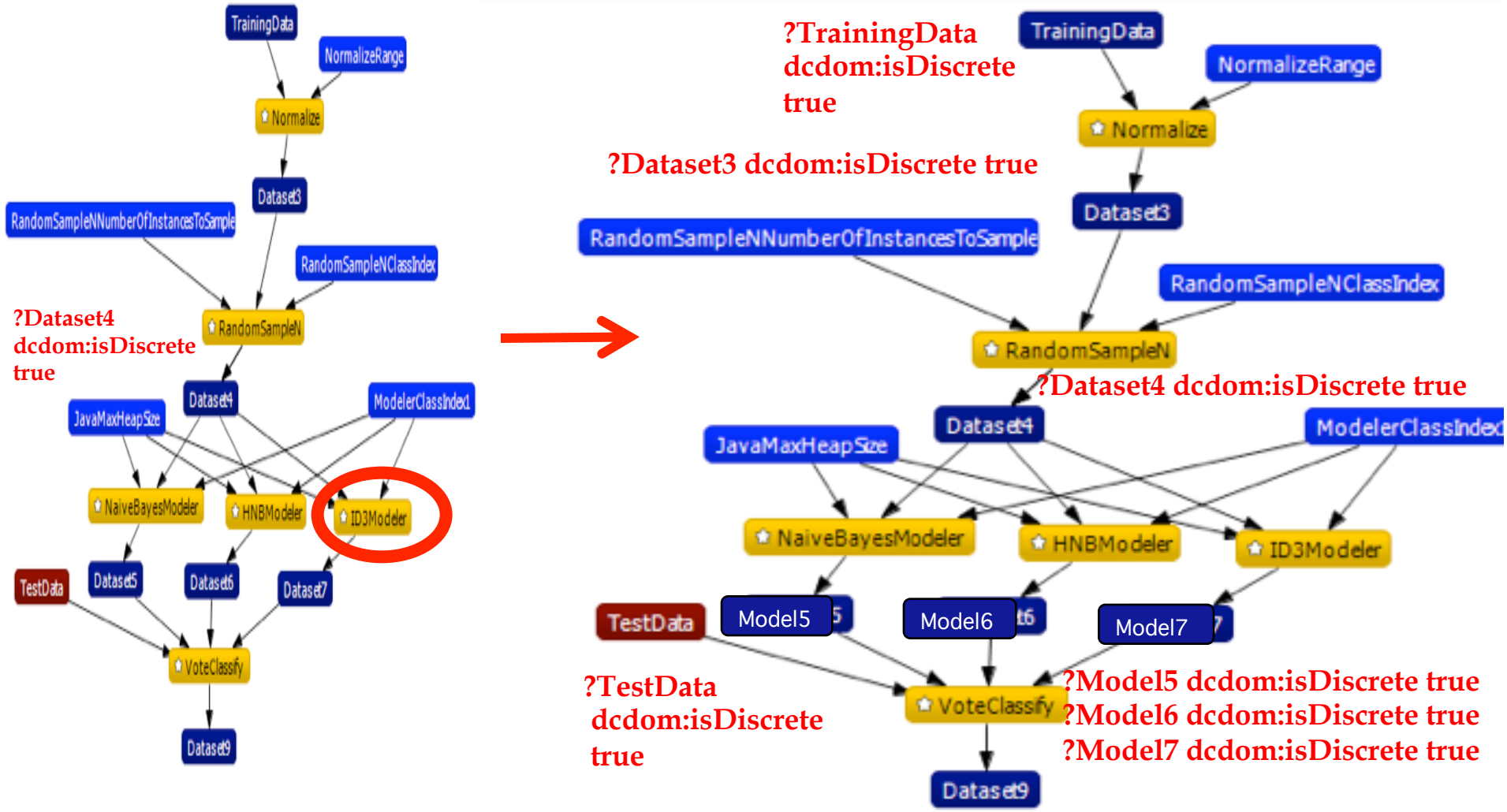
Example (Step 5 of 5)

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

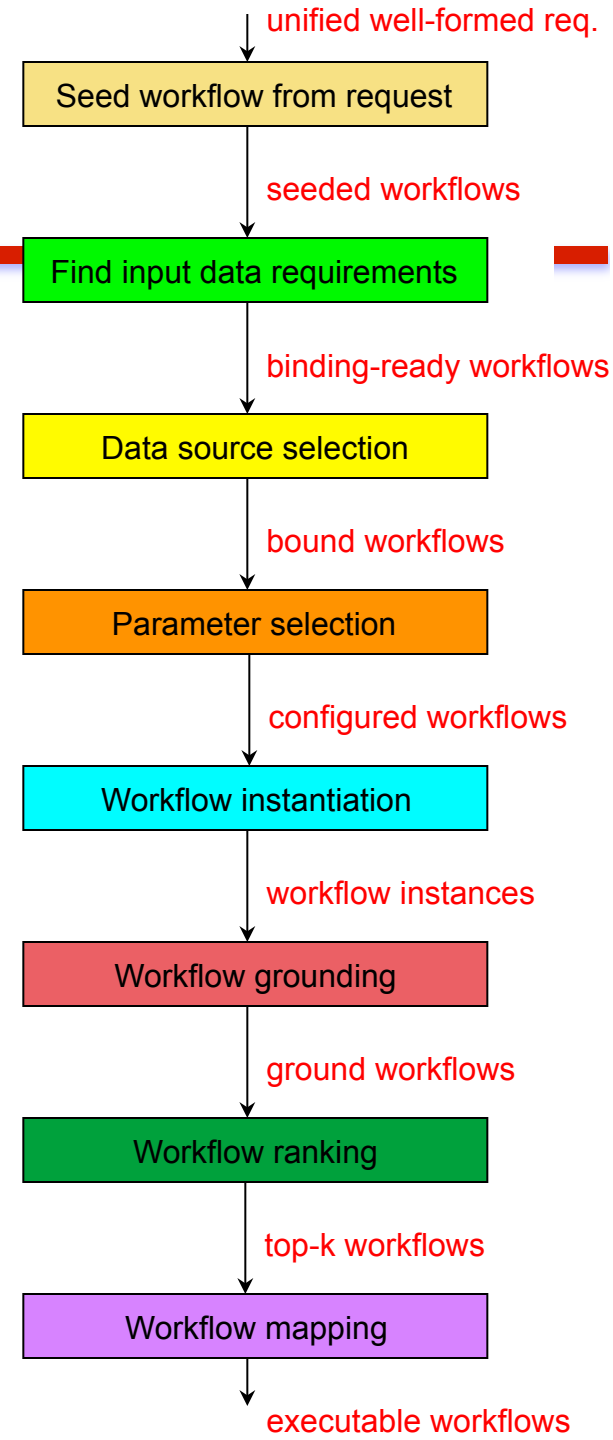


WINGS Automatic Workflow Generation Algorithm [Gil et al JETAI 2011]

Work with P. Gonzalez (UCM) and Jihie Kim (ISI)

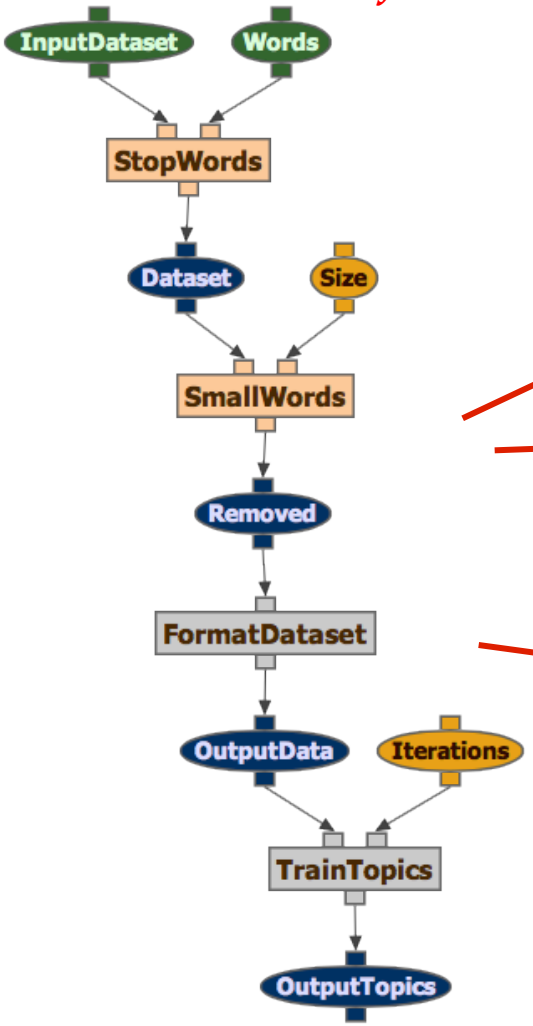
Workflows with S. McWeeney & C. Zhang (OHSU)

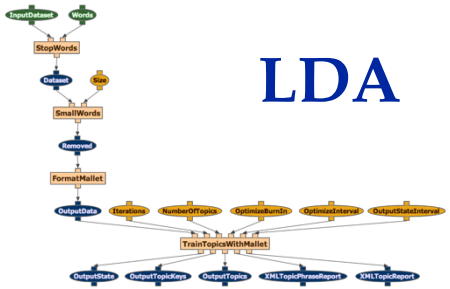
"Pay-as-you-go" semantics



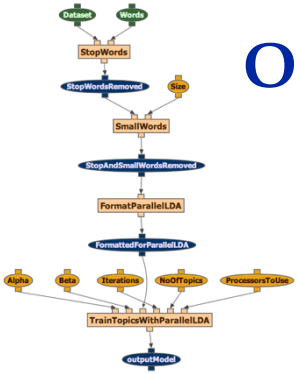
Benefits of Semantic Workflows:
1) Automatic Workflow Elaboration [Gil et al WORKS'13]

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

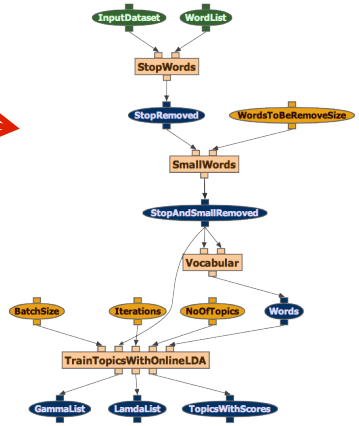




LDA



Online LDA

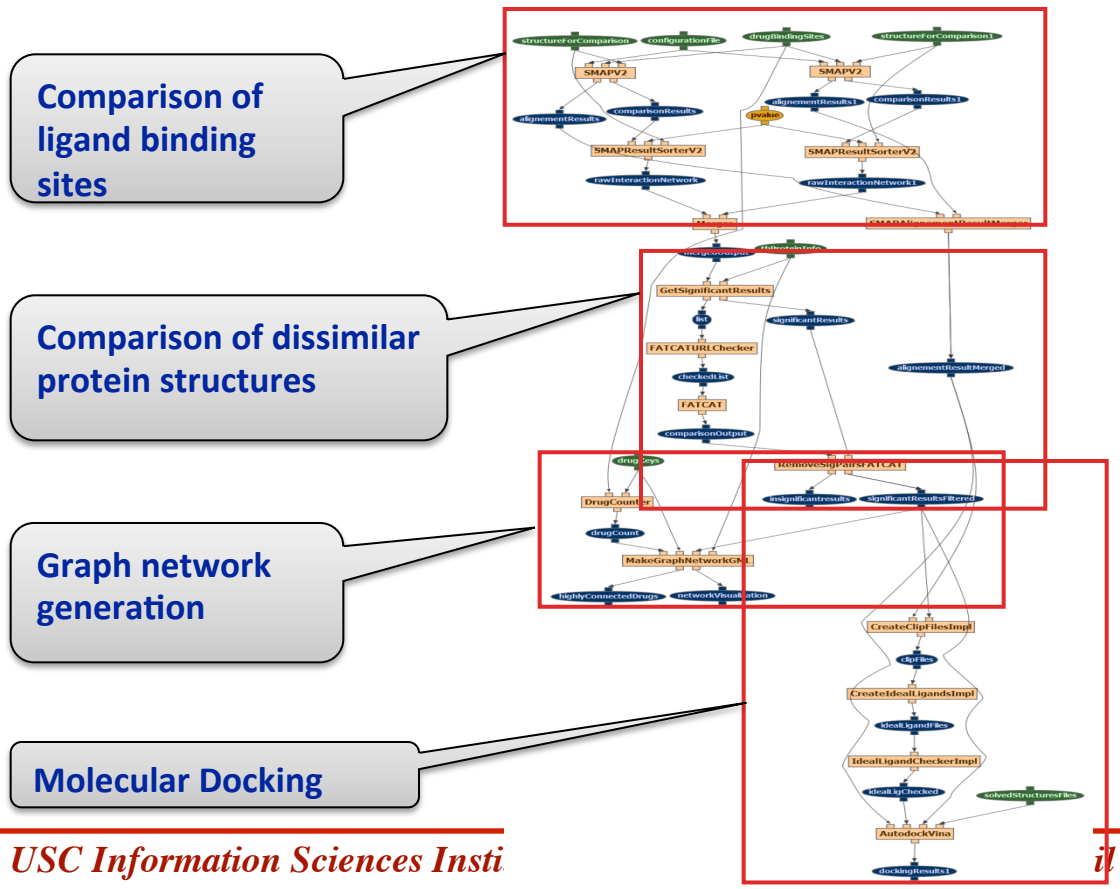


Parallel LDA

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:

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

Comparison of dissimilar protein structures:

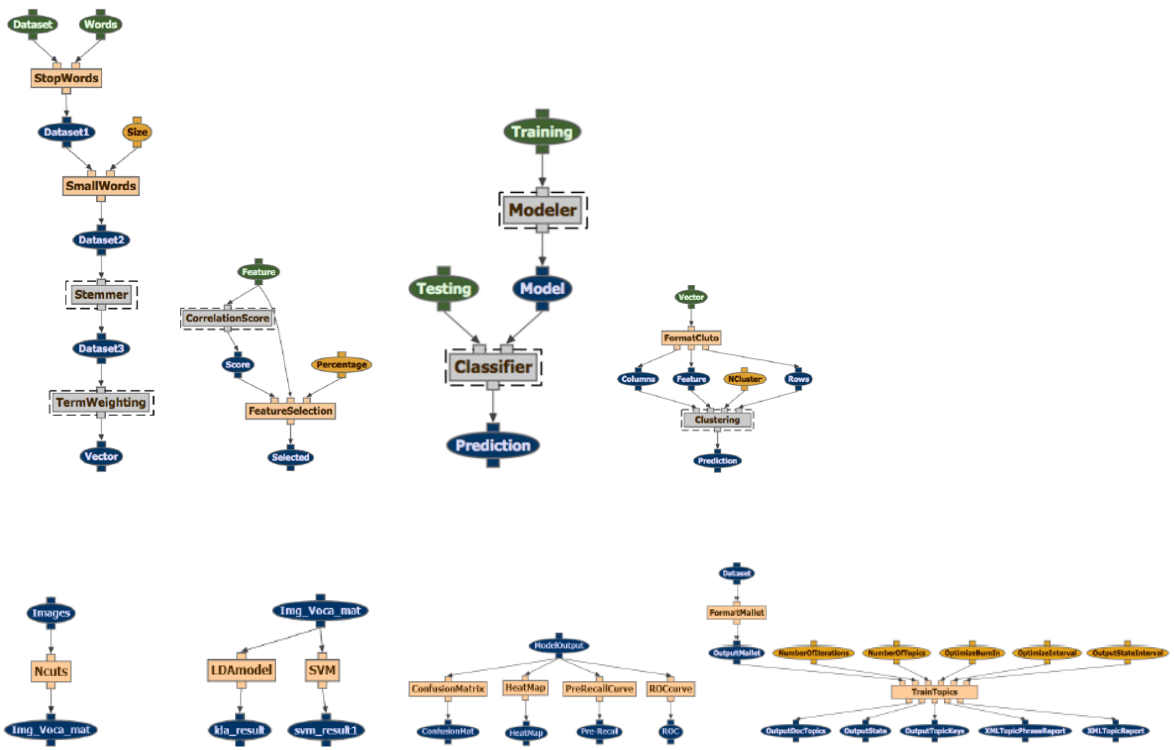
GetSignificant Results	FATCAT URLChecker	FATCAT	Remove Significant Pairs	Minimal
GetSignificant Results	FATCAT URLChecker	FATCAT	Remove Significant Pairs	Novice
GetSignificant Results	FATCAT URLChecker	FATCAT	Remove Significant Pairs	Author

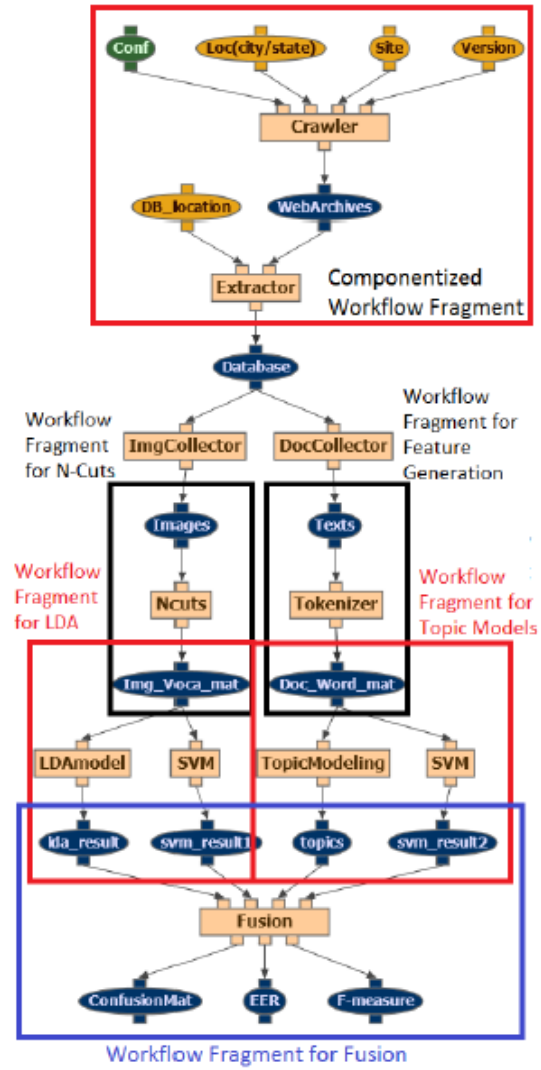
Docking

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

Benefits of Semantic Workflows: 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]

http://www.isi.edu/~gil/publications.php

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)

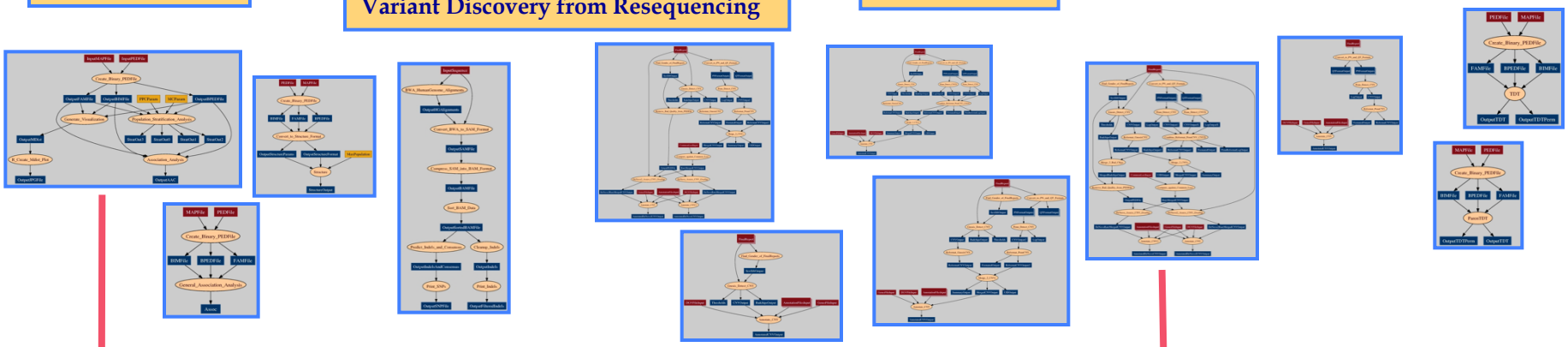
Workflows for population genomics

Association Tests

Variant Discovery from Resequencing

CNV Detection

Transmission Disequilibrium Test

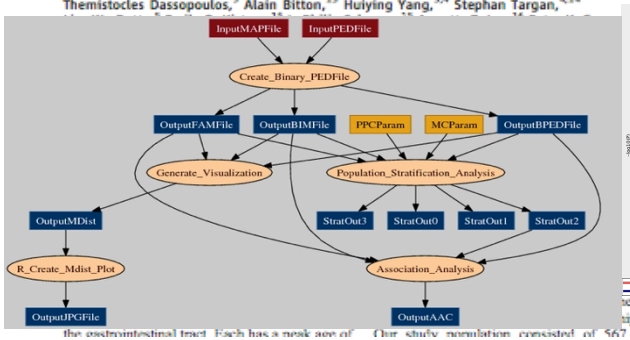


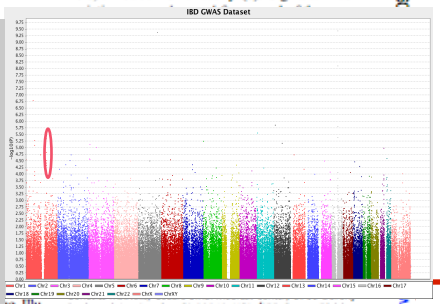
A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

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,[8,10] A. Hillary Steinhart,[9] Clara Abraham,[11] Miguel Regueiro,[1] Anne Griffiths,[12] Themistocles Dassopoulos,[5] Alain Bitton,[13] Huiying Yang,[3,4] Stephan Targan,[4,14]

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-



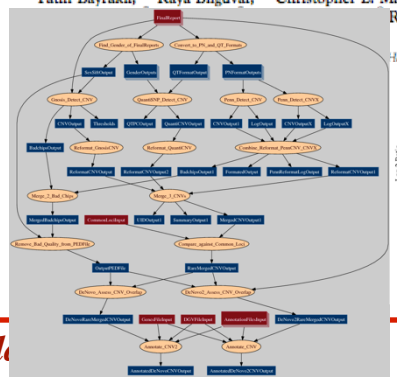


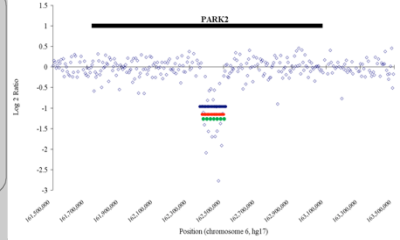
Boulevard, Los Angeles, CA 90048, USA. [4]IBD Center, Division of Gastroenterology, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, CA 90048, USA. [5]Harvey M.

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]

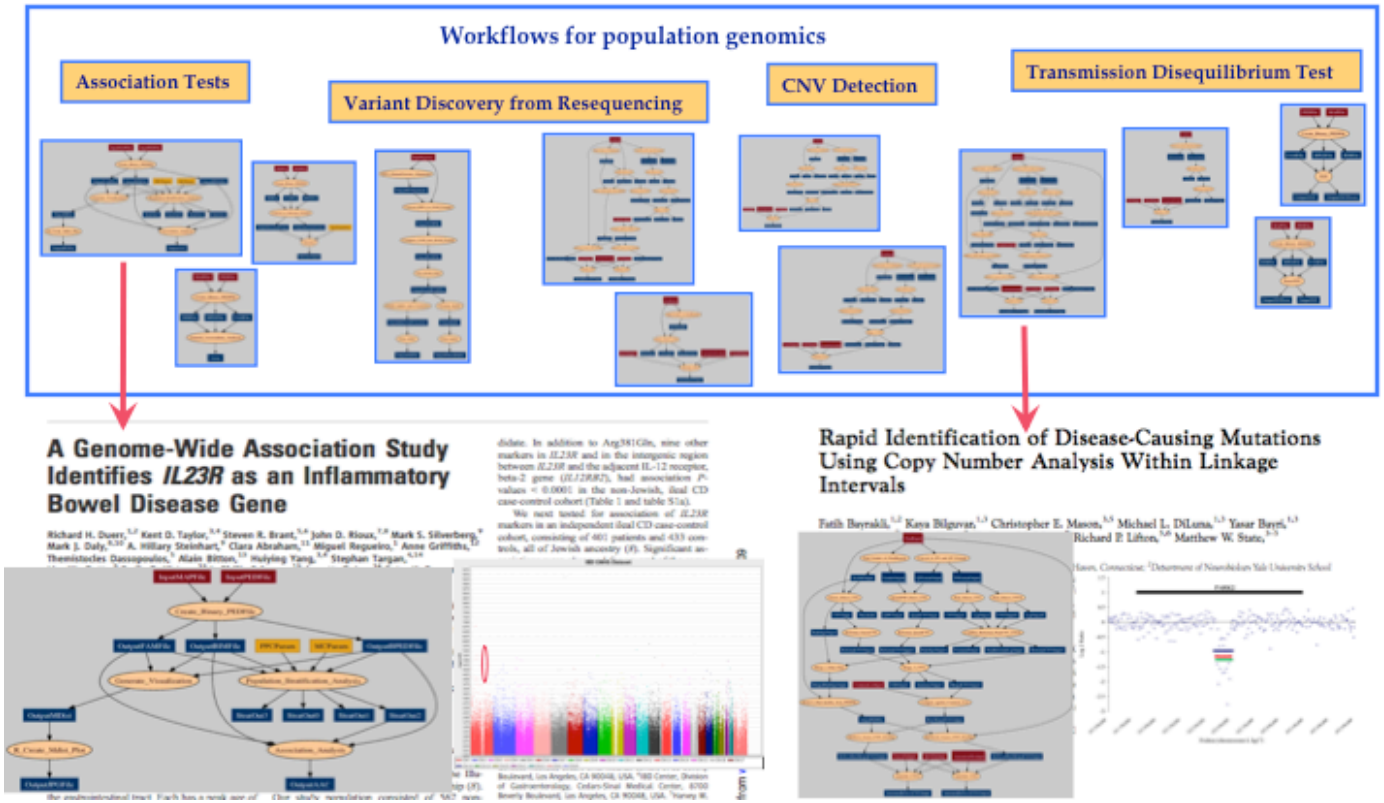
Haven, Connecticut: [2]Department of Neurobiology Yale University School





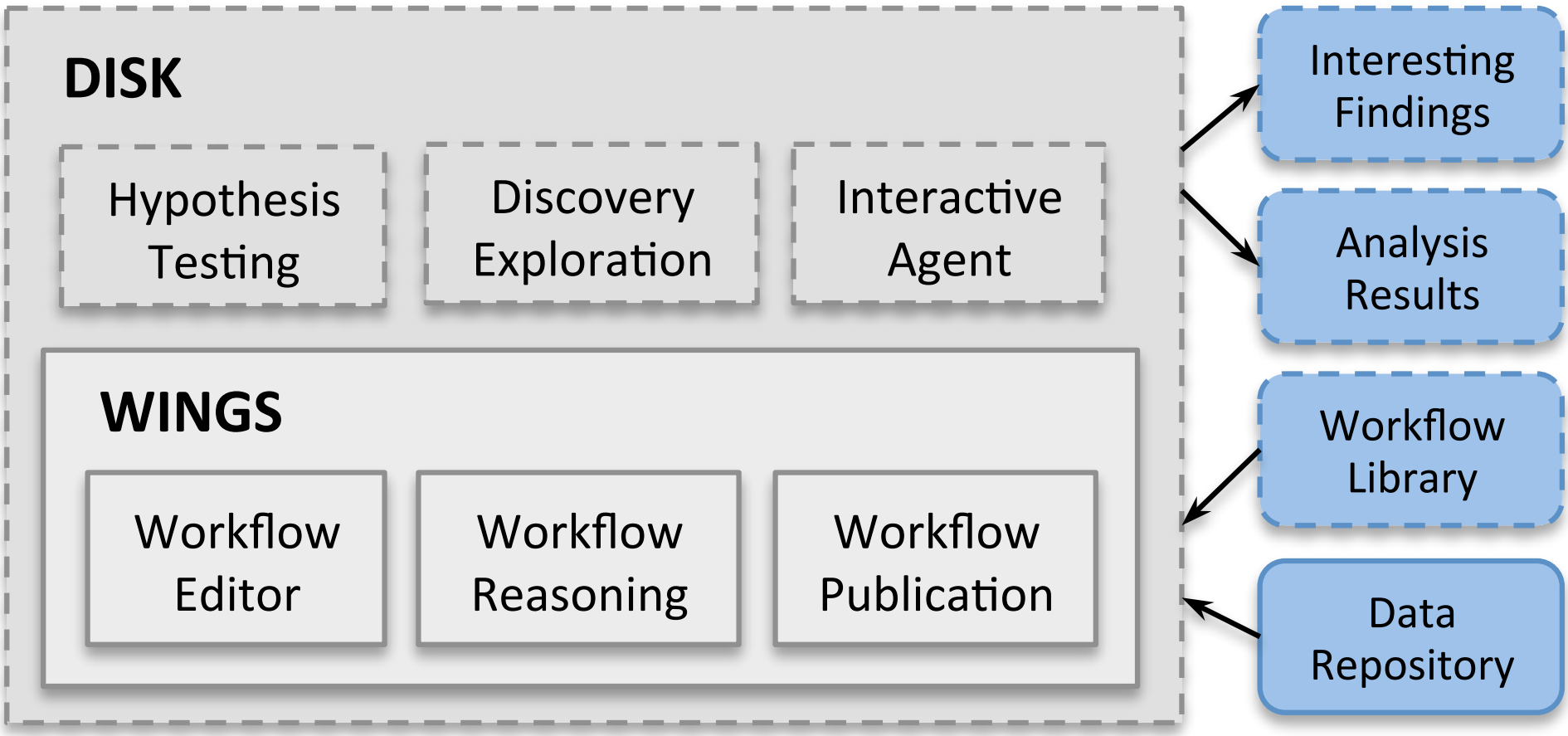
A Grand Challenge: Automatic Analysis of Entire Data Repositories

- Capture knowledge about analytic methods
  - Run workflows in existing data repositories
  - Report new findings



Meta-Workflows for Identifying Interesting Findings of Analysis Workflows

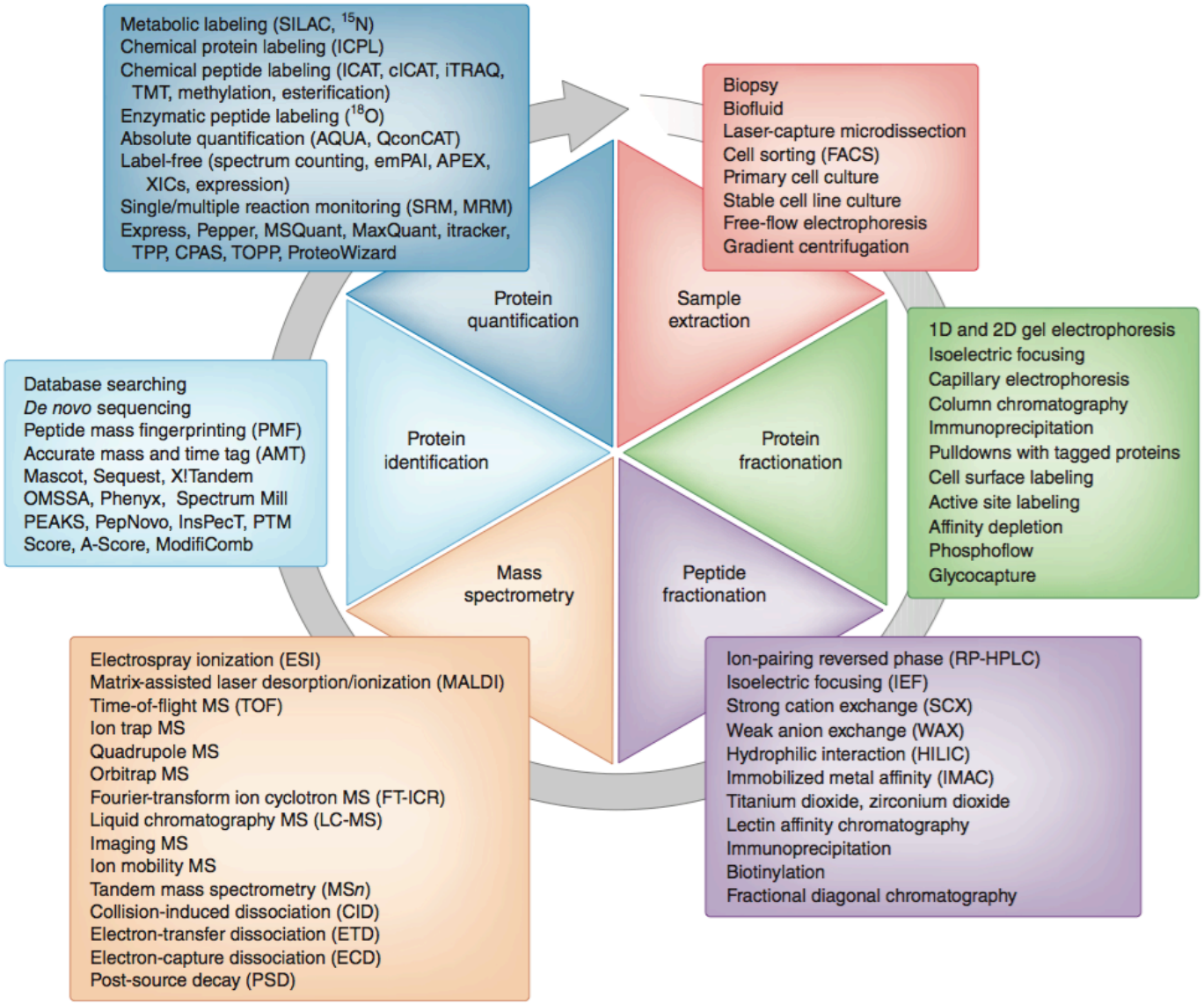
Work with Parag Mallick (Stanford University)



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

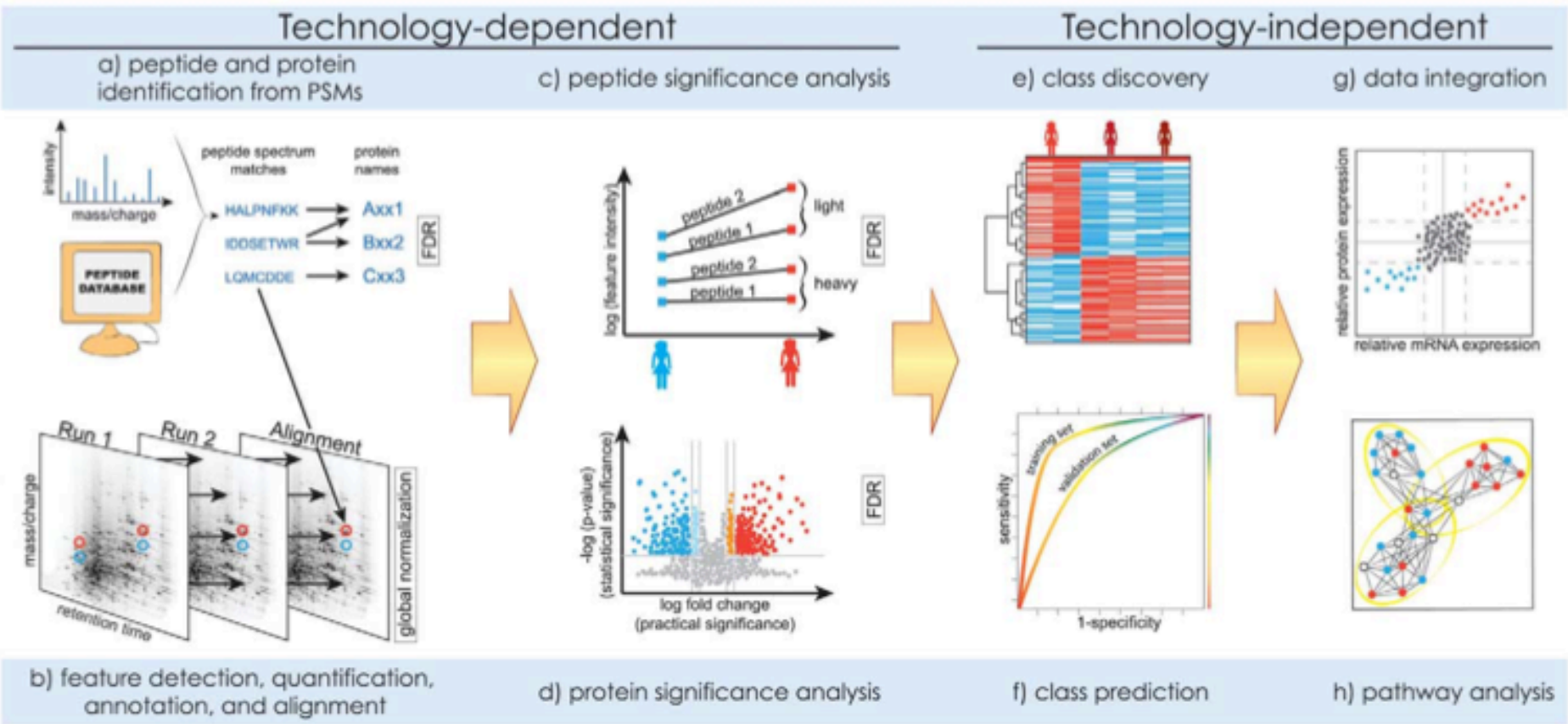
Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. *Nat Biotechnol* 28, 695–709 (2010)

COMPUTATIONAL



EXPERIMENTAL

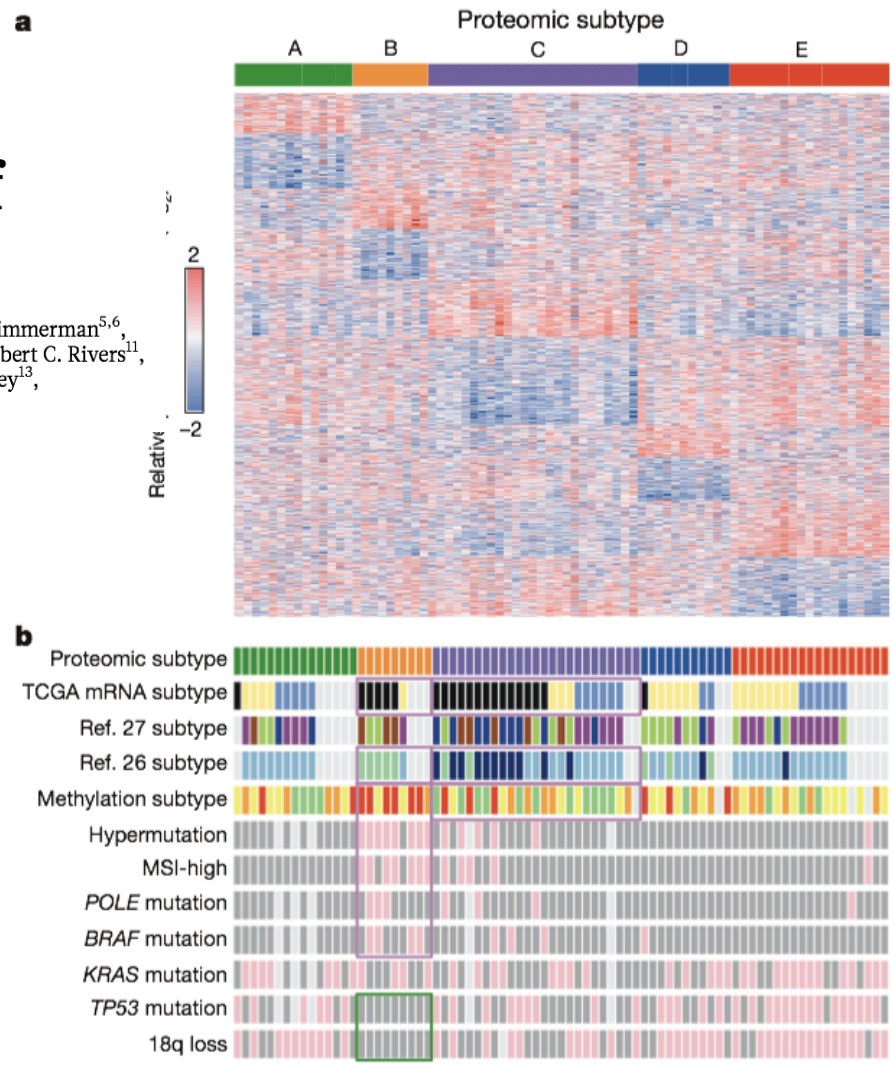
Upstream Processing Affects Downstream Results: Automated Process Would Avoid Errors



Compartmentalized Expertise: Automated Process Would Cover Multiple Expertise Areas

Proteogenomic characterization of human colon and rectal cancer

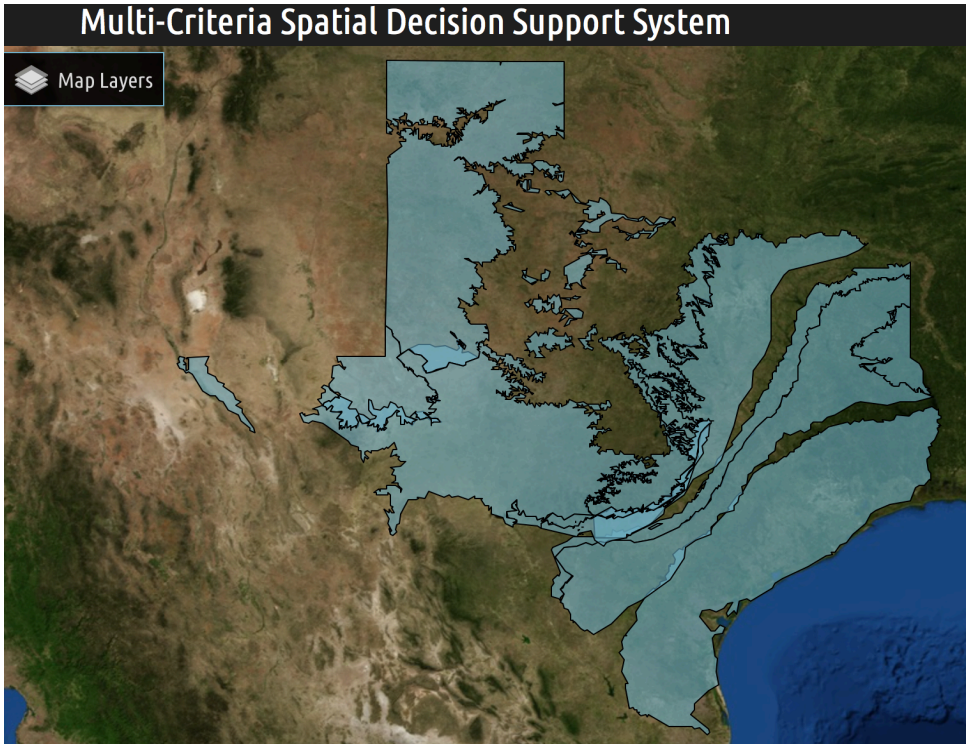
Bing Zhang[1,2], Jing Wang[1], Xiaojing Wang[1], Jing Zhu[1], Qi Liu[1], Zhiao Shi[3,4], Matthew C. Chambers[1], Lisa J. Zimmerman[5,6], Kent F. Shaddox[6], Sangtae Kim[7], Sherri R. Davies[8], Sean Wang[9], Pei Wang[10], Christopher R. Kinsinger[11], Robert C. Rivers[11], Henry Rodriguez[11], R. Reid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

Work with Suzanne Pierce (University of Texas Austin)

Organic Data Science

All Tasks | My Tasks 38

computer scien | search

Develop Framework for Organic Data Sci
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

John Talk Preferences Watchlist Contributions Log out

Page | Discussion

Search

Your Overdue Tasks
Write about the evaluation
a day ago

Write paper about the initial framework design

Draft paper about the initial framework design

Timeline | SubTasks

TODAY

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%

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, Inconsitent with parent

Write paper about the initial framework design
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

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.

Properties

Add

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

SMW Semantic MediaWiki

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

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 sciencemag.org Science AAAS 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 Hirsh[4]**

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 practice. Such work aligns well with the current needs of peta- and exascale science. However, AI has far broader capacity to accelerate 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.

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."

"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

Science AAAS

PERSPECTIVE

COMPUTER SCIENCE

Automating Science

David Waltz[1], Bruce G. Buchanan[2]

+ 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 Geoscieces: 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

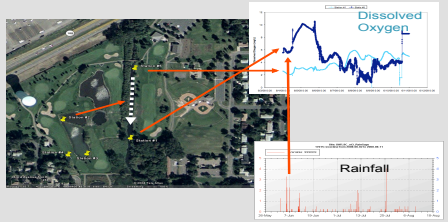
http://www.earthcube.org/

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."

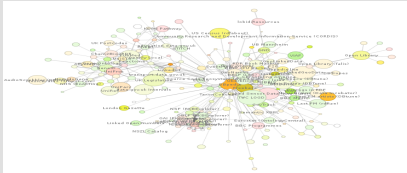
Geospatial Reasoning



Geospatial Pattern Matching: Discovering Flow Anomalies

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

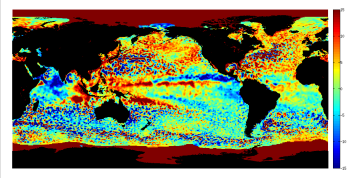
Information Integration



Semantic Metadata: Entity Linking Across Data Sources

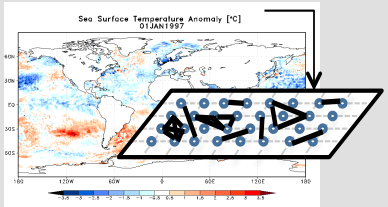
- Name-based and structure-based mapping of entities
- Semi-automatic integration of diverse data sources

Machine Learning



Pattern Mining: Monitoring Ocean Eddies

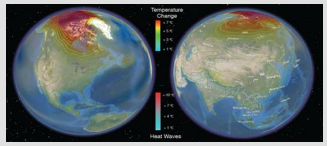
- Spatio-temporal pattern mining of satellite data using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks



Network Analysis: Climate Teleconnections

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

http://climatechange.cs.umn.edu/



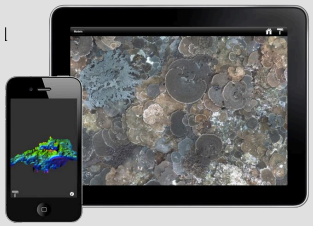
Extremes and Uncertainty: Heat waves, heavy rainfall

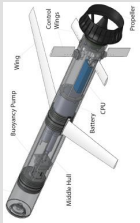
- Extreme value theory in space-time and dependence of extremes on covariates
- Spatiotemporal trends in extremes and physics-guided uncertainty quantification

Change Detection: Monitoring Ecosystem Disturbances

- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Created a comprehensive catalogue of global changes in vegetation, e.g. fires, deforestation, and insect damage

Robotics

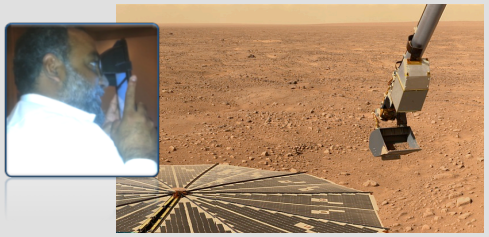




Offline Models from AUV data: Models of Coastal Zones

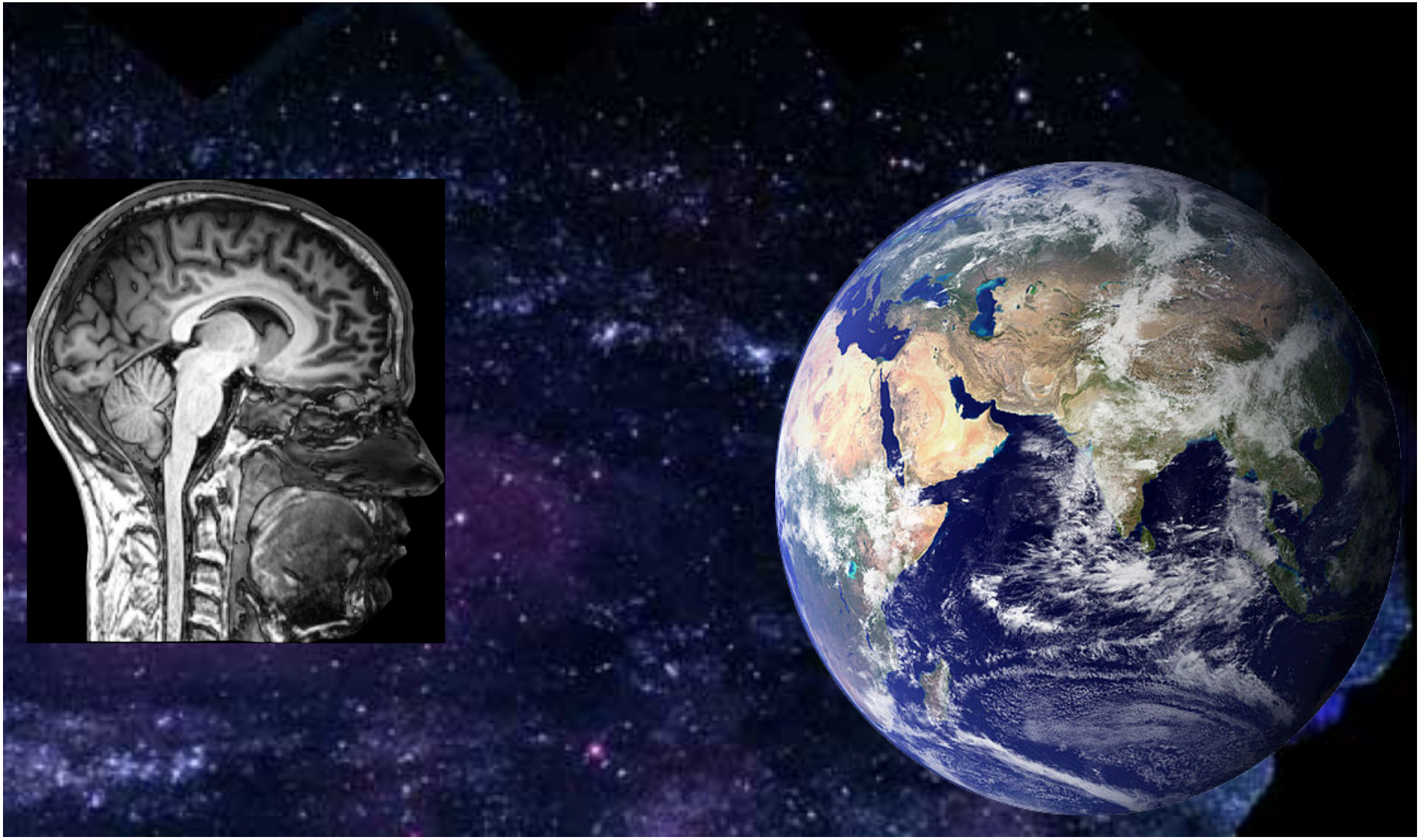
- Georeferenced mapping and 3D reconstruction
- Long-term autonomy for AUV gliders includes in-situ mass-spectrometry

Augmented Reality

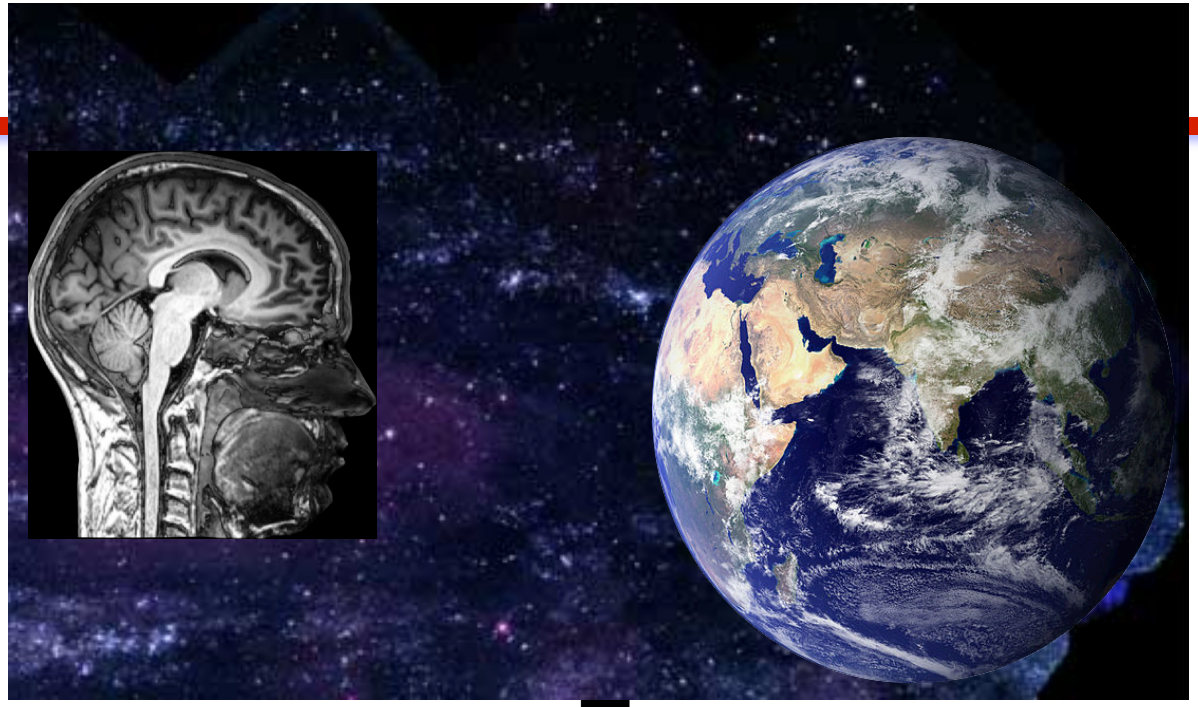


Tablet-based Augmented Reality: Exploring Remote Locations

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



http://commons.wikimedia.org/wiki/File:MRI_brain_sagittal_section
http://commons.wikimedia.org/wiki/File:Earth_Eastern_Hemisphere
http://www.nasa.gov/mission_pages/swift/bursts/uv_andromeda.htm



http://commons.wikimedia.org/wiki/File:Mano_cursor.s

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!*