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 (III): Lack of Access to Data Analytics Expertise

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

“...the volume of data we need to analyze will expand exponentially.”

- 26.5% YES
- 34.4% YES, through collaborators
- 23.0% NO
- 16.1% No special skills needed

*Science, Dec 2011*
Fragmentation of Expertise: An Example from Proteomics

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
Semantic integration of biomedical databases

The significance of the interaction between DAZAP1 and DAZL 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 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 DNA binding proteins is exemplified by the neuron-specific nuclear RNA binding protein Nova-1. Nova-1 regulates the alternative splicing of the pre-mRNA encoding neuronal inhibitory glycine receptor (GHR). The ability of Nova-1 to activate exon selection in neurons is antagonized by a second DNA binding protein, RB1, trans-enriched polyproline tract binding protein, which interacts with Nova-1 and inhibits its function. DAZAP1 could function in a similar manner by binding to DAZL and inhibiting its function. Comparing the phenotypes of Dazl and Dazap1 single and double knock-out mice may provide some clues to the significance of their interaction. Dazl knock-out mice have already been generated and studied. 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 genetic structure of Dazap1 delineated here, should facilitate the generating of Dazap1 null mutation.

Generation of interesting new hypotheses

Proteins of no common family
Proteins in the previous AVE based sub-network
Inferred synapse signaling proteins
Inferred myogenic proteins
Robot Scientist [King et al 2009]
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

THE STRUCTURE OF SCIENTIFIC REVOLUTIONS

Computation, Causation, & Discovery

In Search of MECHANISMS
Discoveries across the Life Sciences
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

[Diagram showing data analysis flow]

[Map of California’s Central Valley]

[Image of water quality sensors]

[Image of database interface]

[Image of weather conditions and models]

[Image of data processing steps]

[Image of Owens-Gibbs Model, O’Connor-Dobbins Model, Churchill Model]

[Image of reaeration rate calculation]

[Image of net daily metabolism calculation]
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

www.wings-workflows.org
Semantic Components in WINGS [Gil iEMSs 2014]

Classes of models/components

I/O Data constraints

Use constraints

; Depth must be over .6m [ CMInvalidity1:
(?c rdf:type pcdom:ReaerationCMClass)
(?c pc:hasInput ?idv)
(?idv pc:hasArgumentID 'InputParameters')
(?idv dcdom:depth ?depth)
le(?depth '0.61')
-> (?c pc:isInvalid 'true')]
WINGS Specializes Workflow Based on Characteristics of Daily Data

1) Parameter settings

2) Choice of models

Owens-Gibbs Model
Connor-Dobbins Model
Churchill Model

3) Metadata of new results

WINGS Specializes Workflow Based on Characteristics of Daily Data
WINGS Dynamically Selects Appropriate Model Based on Daily Sensor Readings

Churchill model

O’Connor-Dobbins model

Owens-Gibbs model
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
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

Input data for decision tree modelers (e.g., ID3) must be discrete.
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 ?val1)]
```
WINGS Workflow Reasoners: Result

?Dataset4 dcdom:isDiscrete true

?Dataset3 dcdom:isDiscrete true

?Dataset4 dcdom:isDiscrete true

?TestData dcdom:isDiscrete true

?Model5 dcdom:isDiscrete true

?Model6 dcdom:isDiscrete true

?Model7 dcdom:isDiscrete true
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

- Seed workflow from request
- Find input data requirements
- Data source selection
- Parameter selection
- Workflow instantiation
- Workflow grounding
- Workflow ranking
- Workflow mapping

unified well-formed req.
seeded workflows
binding-ready workflows
bound workflows
configured workflows
workflow instances
ground workflows
top-k workflows
executable workflows
Benefits of Semantic Workflows:  
1) Automatic Workflow Elaboration [Gil et al WORKS’13]

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

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
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


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

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


COMPUTATIONAL

ExPERIMENTAL

Metabolic labeling (SI-LAC, $^{15}\text{N}$)
Chemical protein labeling (ICPL)
Chemical peptide labeling (ICAT, cICAT, iTRAQ, TMT, methylation, esterification)
Enzymatic peptide labeling ($^{13}$O)
Absolute quantification (AQUA, OqonCAT)
Label-free (spectrum counting, emPAI, APEX, XICs, expression)
Single/multiple reaction monitoring (SRM, MRM)
Express, Pepper, MSQuant, MaxQuant, Iteracker, TPP, CPAS, TOPP, ProteoWizard

Biopsy
Biofluid
Laser-capture microdissection
Cell sorting (FACS)
Primary cell culture
Stable cell line culture
Free-flow electrophoresis
Gradient centrifugation

1D and 2D gel electrophoresis
Isoelectric focusing
Capillary electrophoresis
Column chromatography
Immunoprecipitation
Pulldowns with tagged proteins
Cell surface labeling
Active site labeling
Affinity depletion
Phosphoflow
Glycocapture

Electrospray ionization (ESI)
Matrix-assisted laser desorption/ionization (MALDI)
Time-of-flight MS (TOF)
Ion trap MS
Quadrupole MS
Orbitrap MS
Fourier-transform ion cyclotron MS (FT-ICR)
Liquid chromatography MS (LC-MS)
Imaging MS
Ion mobility MS
Tandem mass spectrometry (MSn)
Collision-induced dissociation (CID)
Electron-transfer dissociation (ETD)
Electron-capture dissociation (ECD)
Post-source decay (PSD)

Database searching
De novo sequencing
Peptide mass fingerprinting (PMF)
Accurate mass and time tag (AMT)
Mascot, Sequest, X!Tandem
QMSSA, Phenyx, Spectrum Mill
PEAKS, PepNovo, In~S~pect, PTM
Score, A-Score, ModifiComb

Ion-pairing reversed phase (RP-HPLC)
Isoelectric focusing (IEF)
Strong cation exchange (SCX)
Weak anion exchange (WAX)
Hydrophilic interaction (HILIC)
Immobilized metal affinity (IMAC)
Titanium dioxide, zirconium dioxide
Lectin affinity chromatography
Immunoprecipitation
Biotinylation
Fractional diagonal chromatography
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 Zhang1,2, Jing Wang1, Xiaojing Wang1, Jing Zhu1, Qi Liu1, Zhiao Shi3,4, Matthew C. Chambers5, Lisa J. Zimmerman5,6, Kent F. Shaddox8, Sangtae Kim7, Sherri R. Davies8, Sean Wang1, Pei Wang10, Christopher R. Kinsinger11, Robert C. Rivera11, Henry Rodriguez11, R. Reid Townsend6, Matthew J. C. Ellis8, Steven A. Carr12, David L. Tabb1, Robert J. Coffey11, Robbert J. C. Slebos11,9, Daniel C. Liebler5,6 & 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)
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
Amplify scientific discovery with artificial intelligence

Many human activities are a bottleneck in progress

By Yolanda Gil, Mark Greaves, James Hendler, Haym Hirsh

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 of science. AI advances have largely harness sophisticated machine-learning techniques to create correlational predictions from large sets of data. Such work aligns well with the current needs of peta- and exascale science. However, AI has far broader capacity to achieve 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.”
The Promise and Potential of Big Data: A Case for Discovery Informatics

Vasant G. Honavar
"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

Data Workflows Semantics Governance

http://www.earthcube.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
- Retropective detection of when contaminants entered an ecosystem

**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

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

Machine Learning

**Extremes and Uncertainty:**
Heat waves, heavy rainfall
- Extreme value theory in space-time and dependence of extremes on covarables
- Spatio-temporal trends in extremes and physics-guided uncertainty quantification

**Change Detection:**
Monitoring Ecosystem Disturbances
- Robust unsupervised 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

Information Integration

**Semantic Metadata:**
Entity Linking Across Data Sources
- Name-based and structure-based mapping of entities
- Semi-automatic integration of diverse data sources

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
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 (UTrier), 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!**