The Transformative Potential of AI for Science:

Will AI Write the Scientific Papers of the Future?

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

https://tinyurl.com/yolandagil-aaai2020

Source: "*Will AI Write the Scientific Papers of the Future?*", Yolanda Gil, 2020 Presidential Address, Association for the Advancement of Artificial Intelligence (AAAI), February 2020.

A Personal Perspective on AI

A Personal Perspective on AI

- The AI community has always been
  - Visionary
  - Broad
  - Inclusive
  - Interdisciplinary
  - Determined
- And dare I say
  - Successful

AAAI

Association for the Advancement of Artificial Intelligence

A Personal View of Watershed Moments in AI: (I) 1980s

SOAR: AN ARCHITECTURE FOR GENERAL INTELLIGENCE

John E. Laird, Allen Newell and Paul S. Rosenbloom

University of Michigan
Carnegie-Mellon University
Stanford University

Learning representations by back-propagating errors

David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams

Integrating Planning and Learning: The PRODIGY Architecture

Manuela Veloso (veloso@cs.cmu.edu)
Jaime Carbonell (carbonell@cs.cmu.edu)
Alicia Pérez (aperez@cs.cmu.edu)
Daniel Borrajo [2] (dborrajo@fi.upm.es)
Eugene Fink (eugene@cs.cmu.edu)
Jim Blythe (jblythe@cs.cmu.edu)

Scripts Plans Goals and Understanding

Roger C. Schank and Robert P. Abelson

Theo: A Framework for Self-Improving Systems

Tom M. Mitchell, John Allen, Prasad Chalasani, John Cheng, Oren Etzioni, Marc Ringuette, Jeffrey C. Schlimmer

Elephants Don't Play Chess

Rodney A. Brooks
MIT Artificial Intelligence Laboratory, Cambridge, MA 02139, USA

Robotics and Autonomous Systems 6 (1990) 3-15

A Personal View of Watershed Moments in AI: 1990s

1990: Sphynx shows speaker independent large vocabulary continuous speech recognition – used it to write my PhD thesis

1992: Soar flies helicopter teams in simulations and is indistinguishable to commanders from human-controlled aircraft

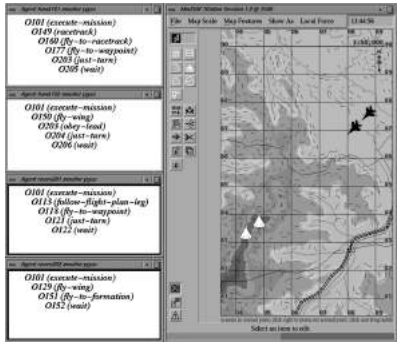
1992: TD-Gammon autonomously learns to play backgammon at human player levels

1995: SKICAT identifies five new quasars in the Second Palomar Sky Survey

1995: Navlab is the first trained car to drive autonomously on highways to cross the United States

1997: Deep Blue defeats human chess world champion and gets grandmaster-level rating

1999: RAX flew a spacecraft autonomously, demonstrating planning, monitoring, and fault repair



https://doi.org/10.1609/aimag.v16i1.1121

https://www.cs.cmu.edu/afs/cs/project/alv/www/navlab_video_page_files/Navlab_Kanade_version97.mpg

https://flic.kr/p/9Dpww

A Personal View of Watershed Moments in AI: 2000s

2000: The Gene Ontology is shown to describe over 15,000 gene products for drosophila, mouse, and yeast

2000: Kismet demonstrates and recognizes emotions

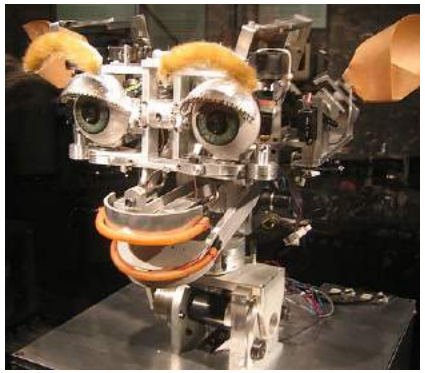
2003: USC ISI's statistical machine translation prototype beats hand-crafted commercial systems

2004: RDF semantic specifications become W3C recommendations for the GGG (Giant Global Graph)

2007: Stanley wins $1M for first autonomous high-speed off-road driving

2007: First robot soccer team against human players in exhibition game at RoboCup

2009: Pragmatic Chaos ensemble learning wins $1M competition to predict user film ratings



https://commons.wikimedia.org/w/index.php?curid=374949

https://commons.wikimedia.org/w/index.php?curid=20798358

A Personal View of Watershed Moments in AI: 2010s

2010: Siri voice-activated personal assistant is released as a smart phone app

2011: CMDragons team of 10 soccer robots coordinate plays for routine passing, interception, and goal scoring

2011: Watson takes first place in Jeopardy Q/A game defeating two human champions

2012: AlexNet scores improved 10 percentage points in the ImageNet visual recognition challenge

2013: Cognitive Tutor shows 8 percentile points average improvement in algebra in 25,000 students study

2016: Knowledge Graph used as semantic backbone in one third of 100B monthly searches

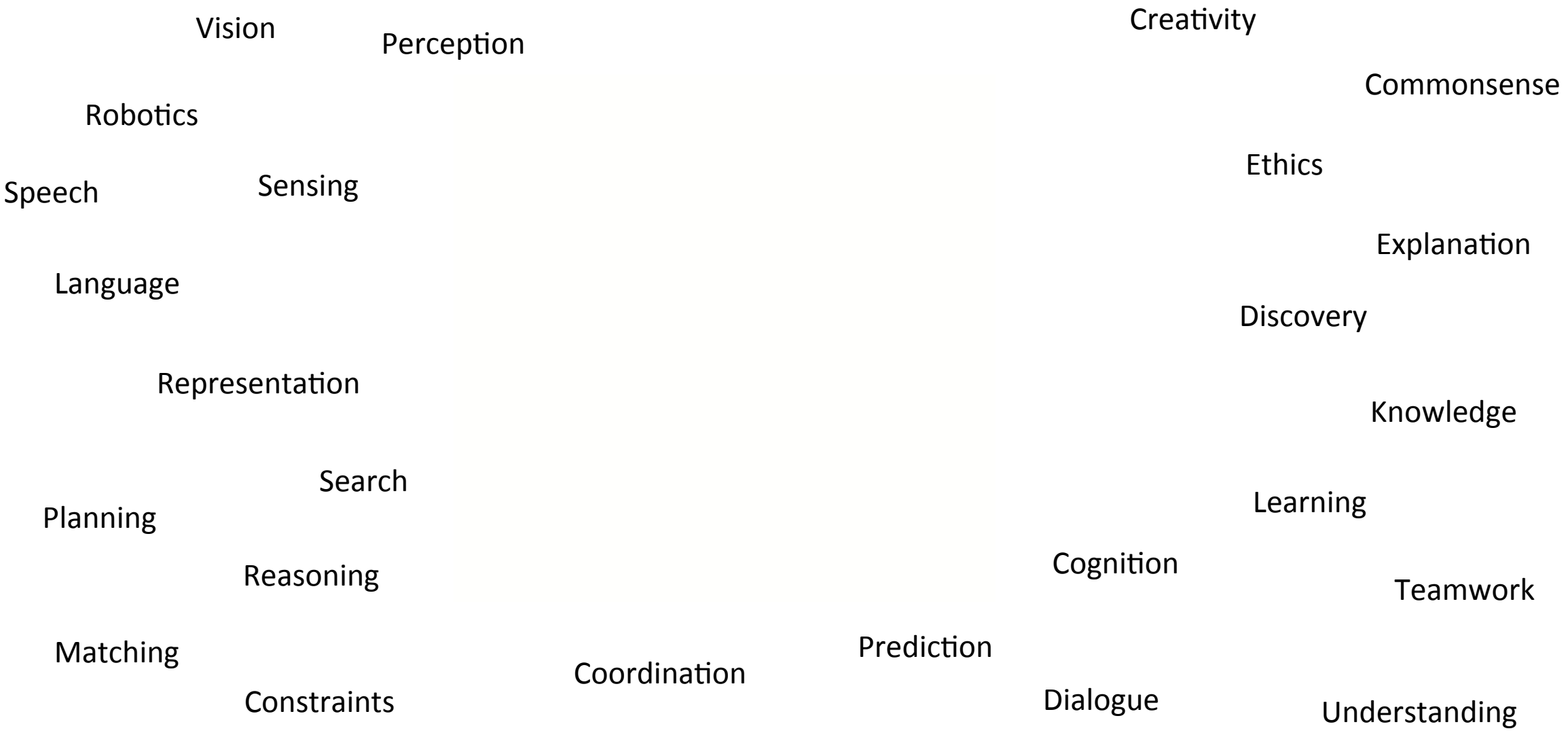
2019: Wikidata records 8 billion triples and over 880 billion edits, surpassing Wikipedia as most edited Wikimedia site

https://en.wikipedia.org/w/index.php?curid=54391045

https://www.youtube.com/watch?v=4QtBSDSC2pk

WIKIDATA

Diversity and Breadth of Advances in AI

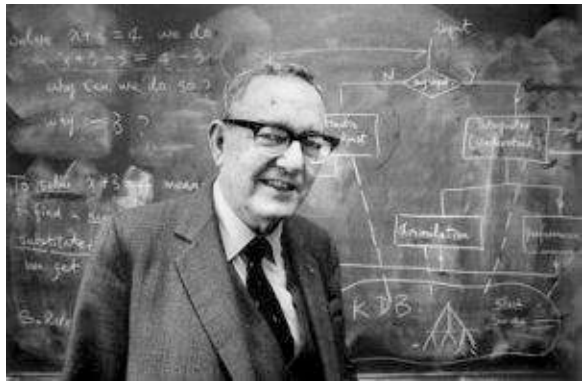


https://commons.wikimedia.org/w/index.php?curid=70268939

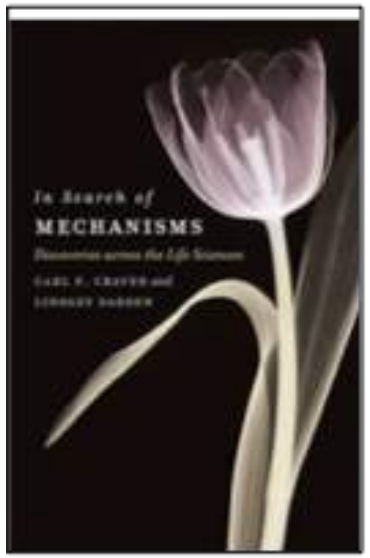
AI to Address Major Future Challenges

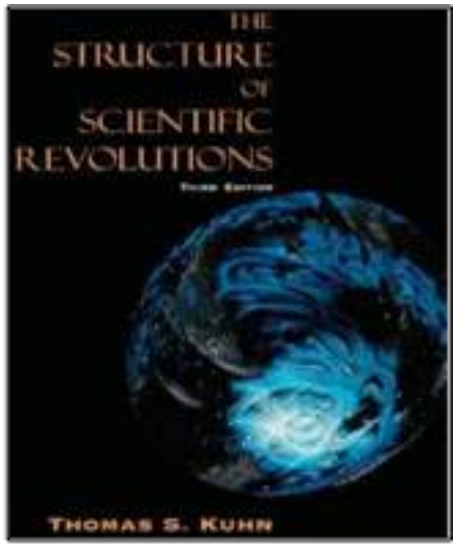
The Imperative for AI in Science

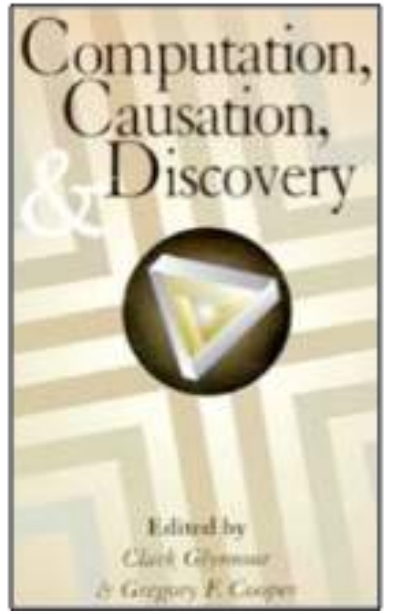
AI and Scientific Discovery



- [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]







Human Limitations Curb Scientific Progress [Gil DSJ'17]

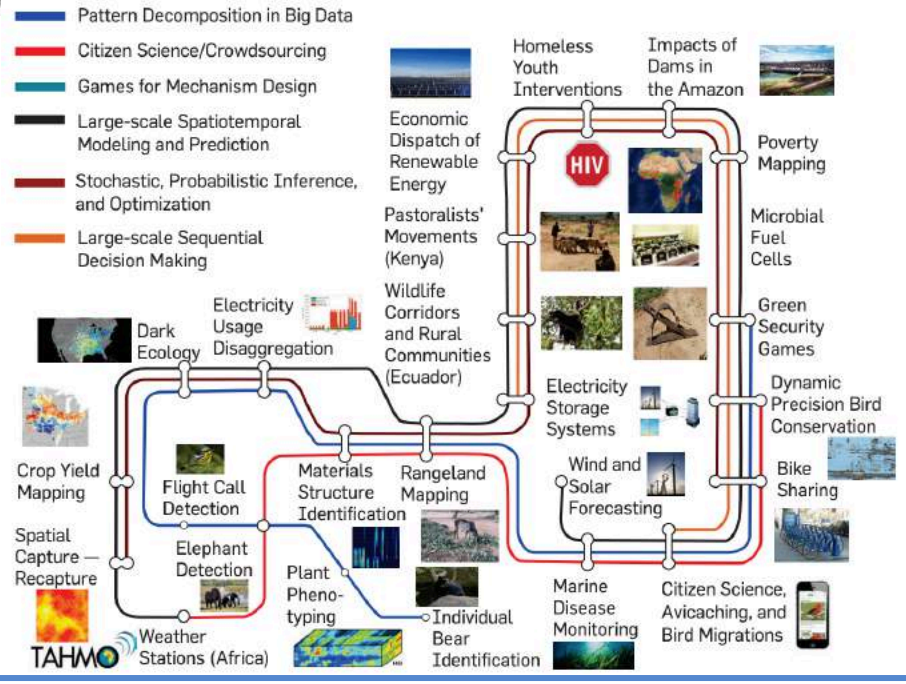
- Not systematic
  - e.g., [Peters et al PLOS 2014]
- Errors
  - e.g., [Herndon et al CJE 2013]
- Biases
  - e.g., [Anderson et al ACS 2014]
- Poor reporting
  - e.g., [Garijo et al PLOS 2013]

COMMUNICATIONS OF THE ACM

https://doi.org/10.1145/3339399

Computational Sustainability: Computing for a Better World and a Sustainable Future

By Carla Gomes, Thomas Dietterich, Christopher Barrett, Jon Conrad, Bistra Dilkina, Stefano Ermon, Fei Fang, Andrew Farnsworth, Alan Fern, Xiaoli Fern, Daniel Fink, Douglas Fisher, Alexander Flecker, Daniel Freund, Angela Fuller, John Gregoire, John Hopcroft, Steve Kelling, Zico Kolter, Warren Powell, Nicole Sintov, John Selker, Bart Selman, Daniel Sheldon, David Shmoys, Milind Tambe, Weng-Keen Wong, Christopher Wood, Xiaojian Wu, Yexiang Xue, Amulya Yadav, Abdul-Aziz Yakubu, Mary Lou Zeeman

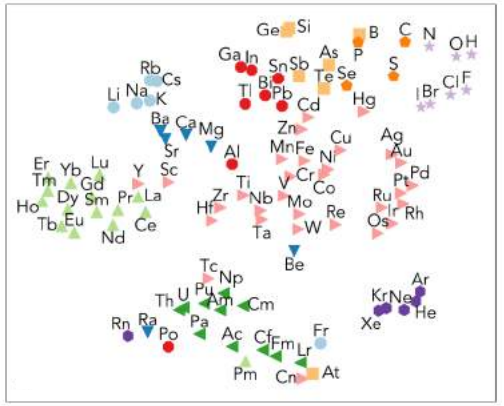


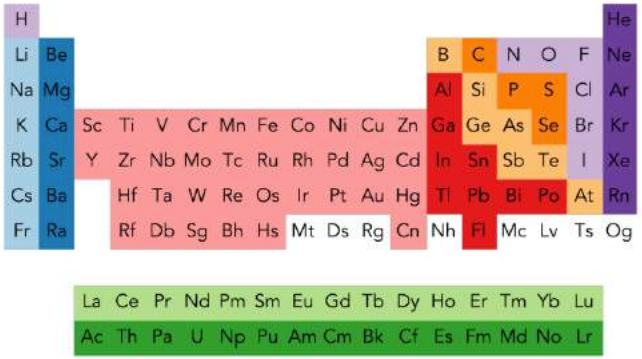
nature

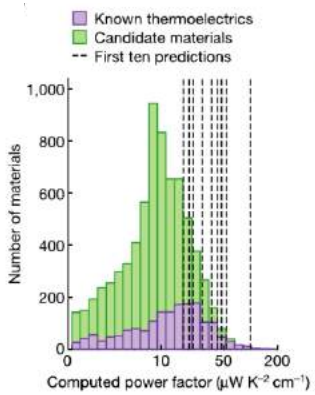
https://doi.org/10.1038/s41586-019-1335-8

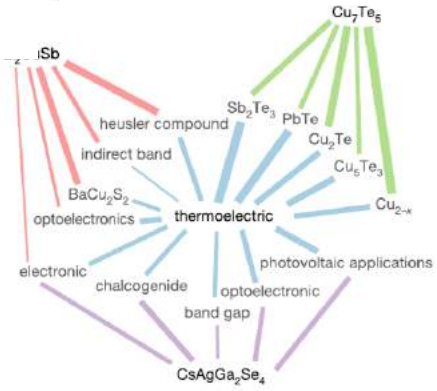
Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan[1,3*], John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2*] & Anubhav Jain[1*]









Knowledge-based Biomedical Data Science 2019

Tiffany J. Callahan[1], Harrison Pielke-Lombardo[1], Ignacio J. Tripodi[1,2], Lawrence E. Hunter[1]*

https://arxiv.org/pdf/1910.06710

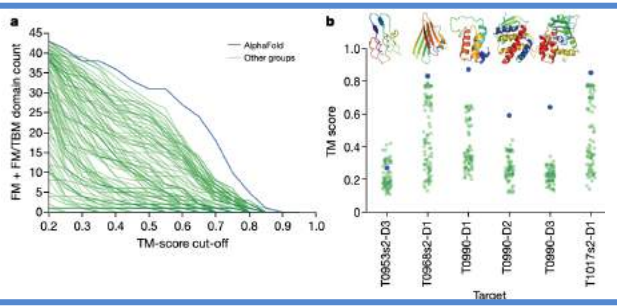
https://doi.org/10.1038/s41586-019-1923-7

nature

Improved protein structure prediction using potentials from deep learning

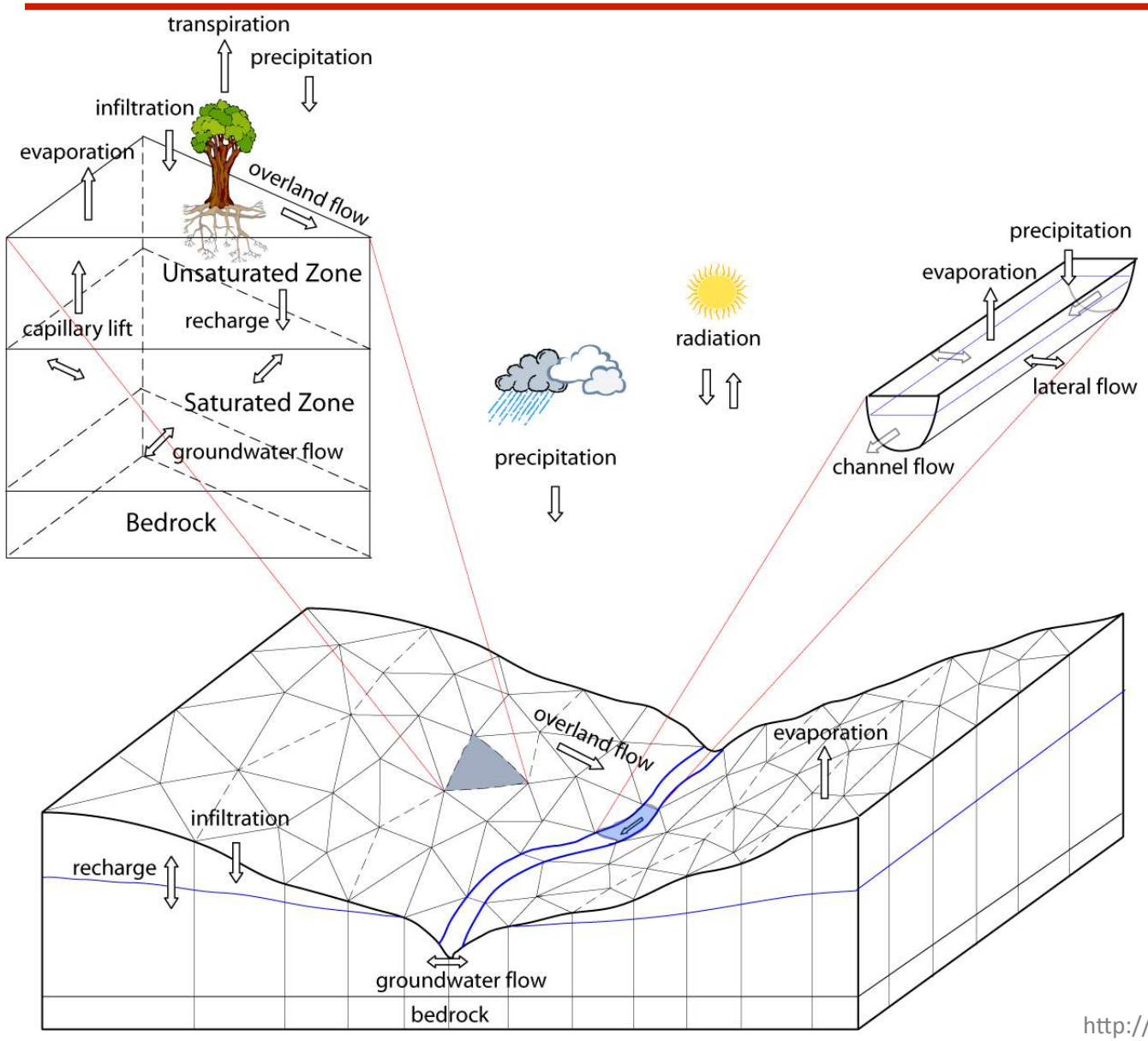
https://doi.org/10.1038/s41586-019-1923-7
Received: 2 April 2019
Accepted: 10 December 2019

Andrew W. Senior[1,4*], Richard Evans[1,4], John Jumper[1,4], James Kirkpatrick[1,4], Laurent Sifre[1,4], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1] & Demis Hassabis[1]



Capturing Scientific Knowledge

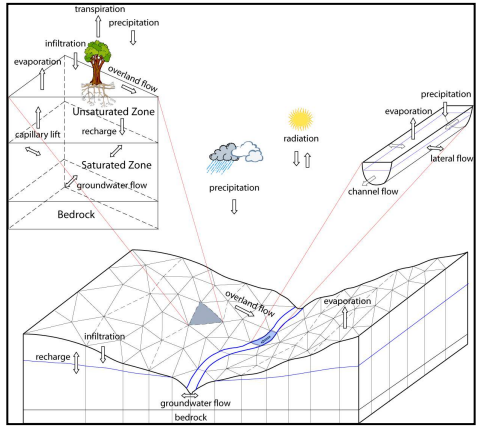
Scientific Knowledge



Process	Governing equation/model	Original governing equations	Semi-discrete form ODEs
Channel Routing	St. Venant Equation	$\frac{\partial h}{\partial t}+\frac{\partial(uh)}{\partial x}=q$	$\left(\frac{dc_i}{dt}=P_c-\sum Q_{gc}+\sum Q_{sc}+Q_{in}-Q_{out}-E_c\right)_i$
Overland Flow	St. Venant Equation	$\frac{\partial h}{\partial t}+\frac{\partial(uh)}{\partial x}+\frac{\partial(vh)}{\partial y}=q$	$\left(\frac{\partial h}{\partial t}=P_s-I-E_s-Q_{ss}+\sum_{j=1}^{3}Q_s^{\,j}\right)_i$
Unsaturated Flow	Richard Equation	$C(\psi)\frac{\partial\psi}{\partial t}=\nabla\cdot(K(\psi)\nabla(\psi+Z))$	$\left(\frac{dv_i}{dt}=I-q^e-ET_t\right)_i$
Groundwater Flow	Richard Equation	$C(\psi)\frac{\partial\psi}{\partial t}=\nabla\cdot(K(\psi)\nabla(\psi+Z))$	$\left(\frac{dv_i}{dt}=q^s+\sum_{j=1}^{3}Q_g^{\,j}-Q_t+Q_{gw}\right)_i$
Interception	Bucket Model	$\frac{dS_I}{dt}=P-E_I-P_n$	$\left(\frac{dS_I}{dt}=P-E_I-P_n\right)_i$
Snow melt	ISNOBAL	$\frac{dS_{snow}}{dt}=P-E_{snow}-\Delta w$	$\left(\frac{dS_{snow}}{dt}=P-E_{snow}-\Delta w\right)_i$
Evapotran-spiration	Penman-Monteith Method	$ET_0=\frac{\Delta(R_n-G)+\rho_a C_p\frac{(e_s-e_a)}{r_a}}{\Delta+\gamma(1+\frac{r_s}{r_a})}$	$\left(ET_0=\frac{\Delta(R_n-G)+\rho_a C_p\frac{(e_s-e_a)}{r_a}}{\Delta+\gamma(1+\frac{r_s}{r_a})}\right)_i$

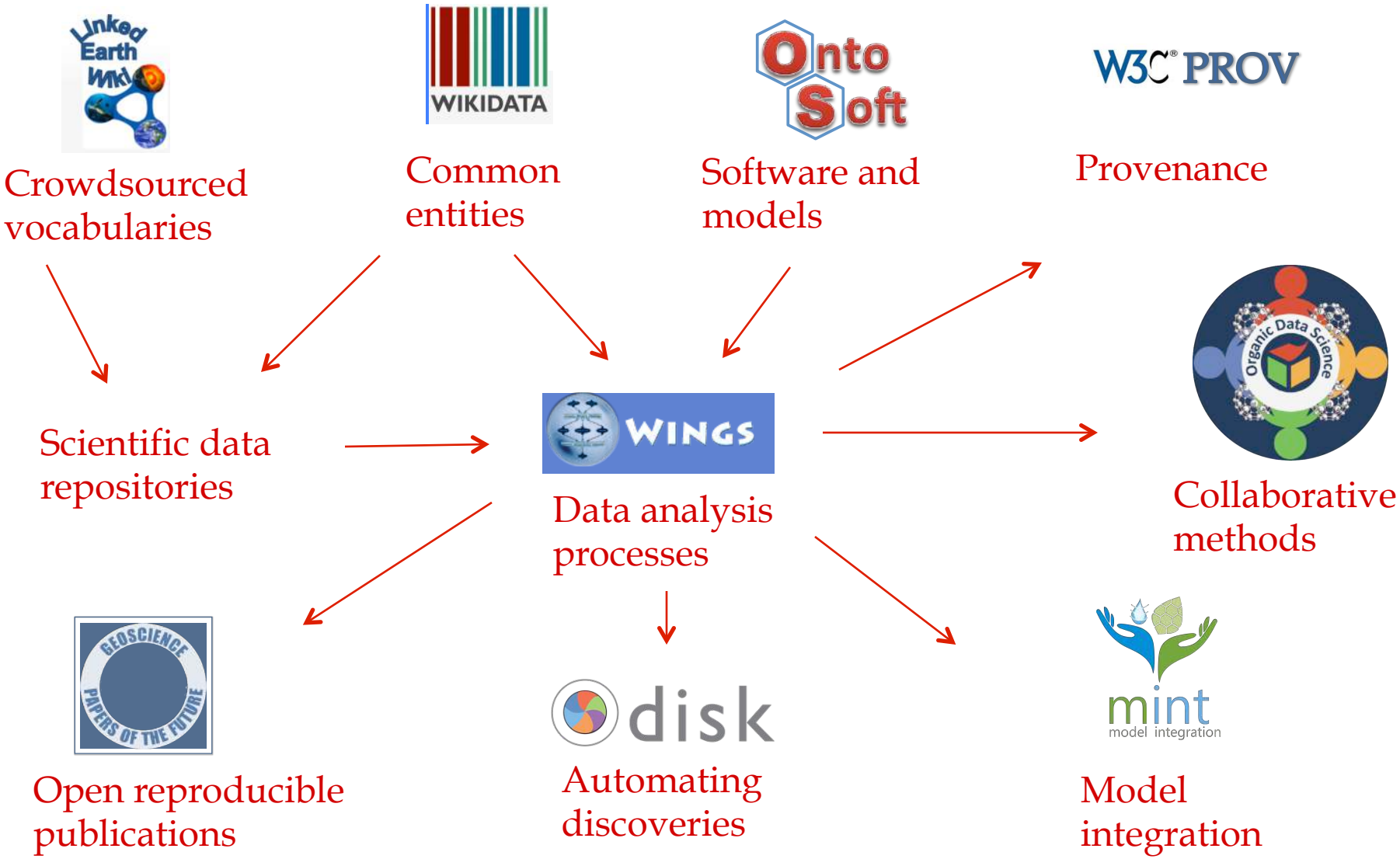
http://www.pihm.psu.edu/pihm_home.html

Supporting Compositionality of Scientific Knowledge

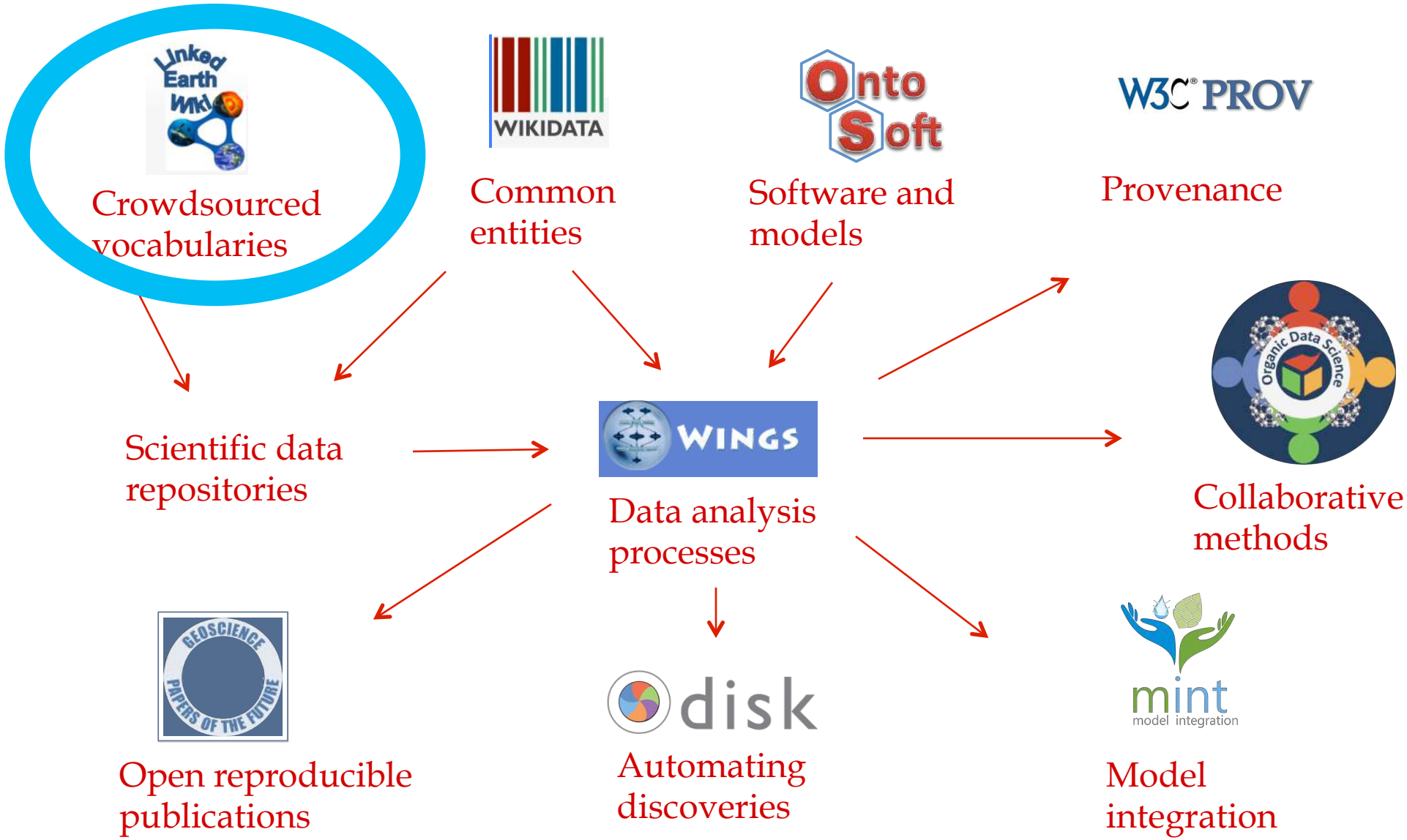


- Data formats
- Physical variables
- Constraints for use
- Adjustable parameters
- Interventions

Capturing Scientific Knowledge



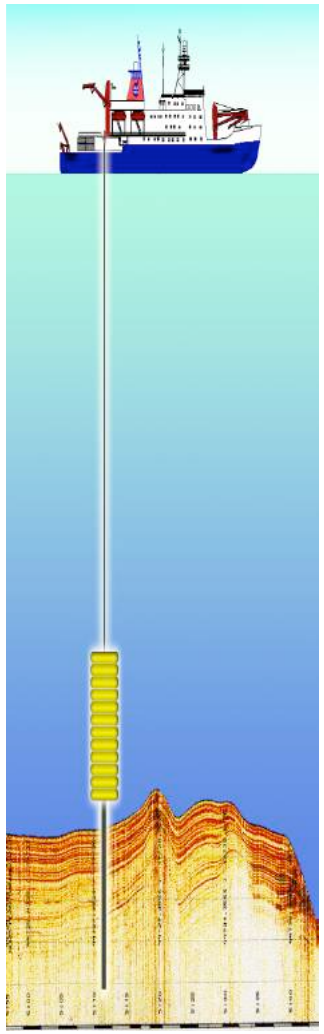
Capturing Scientific Knowledge



Low-Cost Creation of Scientific Vocabulary Standards

[Gil et al ISWC 2017; Khider et al PP 2019; Emile-Geay et al PAGES 2018]

Work with Deborah Khider, Julien Emile-Geay, Daniel Garijo (USC); Nick McKay (NAU)



https://commons.wikimedia.org/wiki/File:Gravity-corer_hg.png

https://commons.wikimedia.org/wiki/File:An_ice_core_segment.jpg

AGU100 ADVANCING EARTH AND SPACE SCIENCE

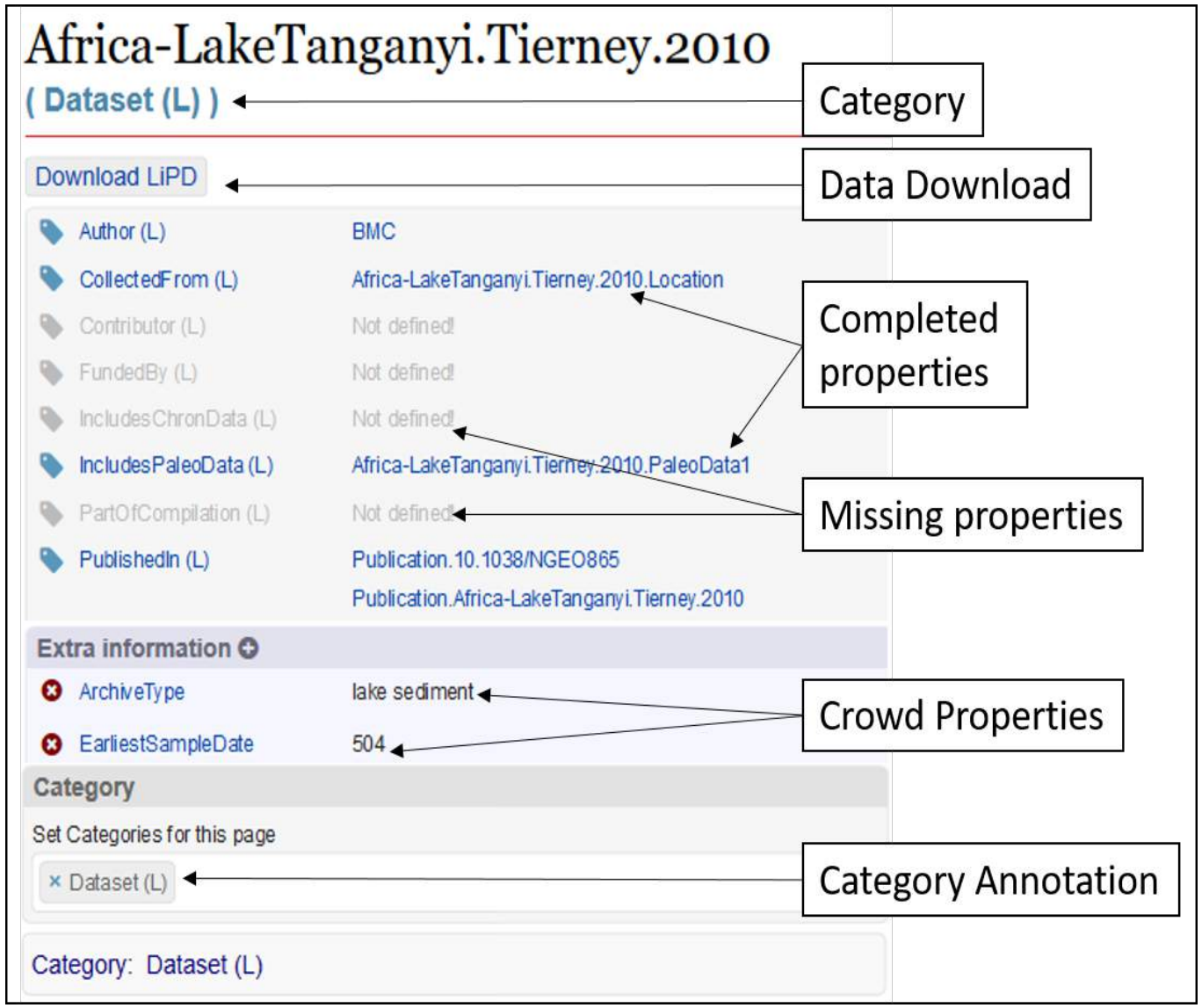
Problem: Diversity of requirements for metadata

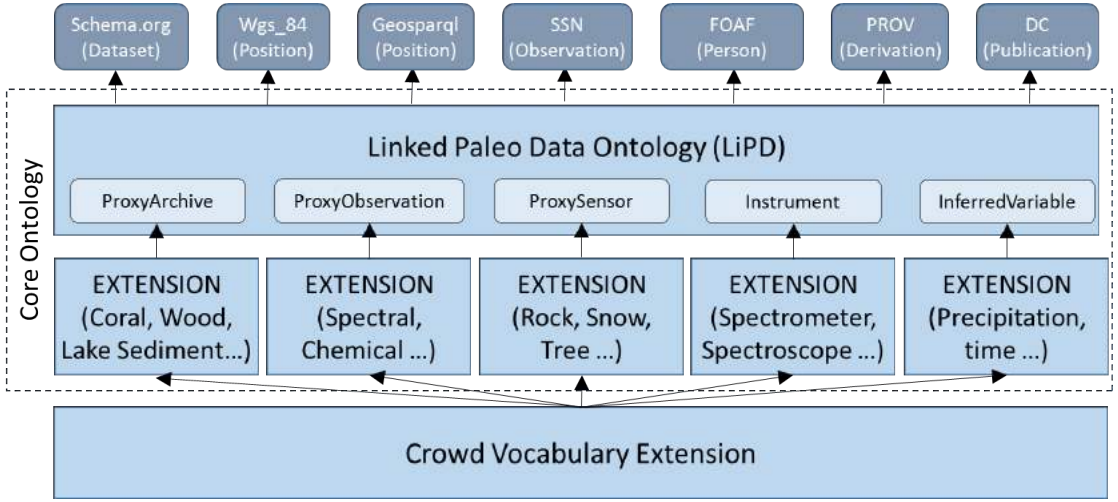
Approach: Semantic technologies used for controlled crowdsourcing facilitate creation of community standards to describe highly heterogeneous scientific data

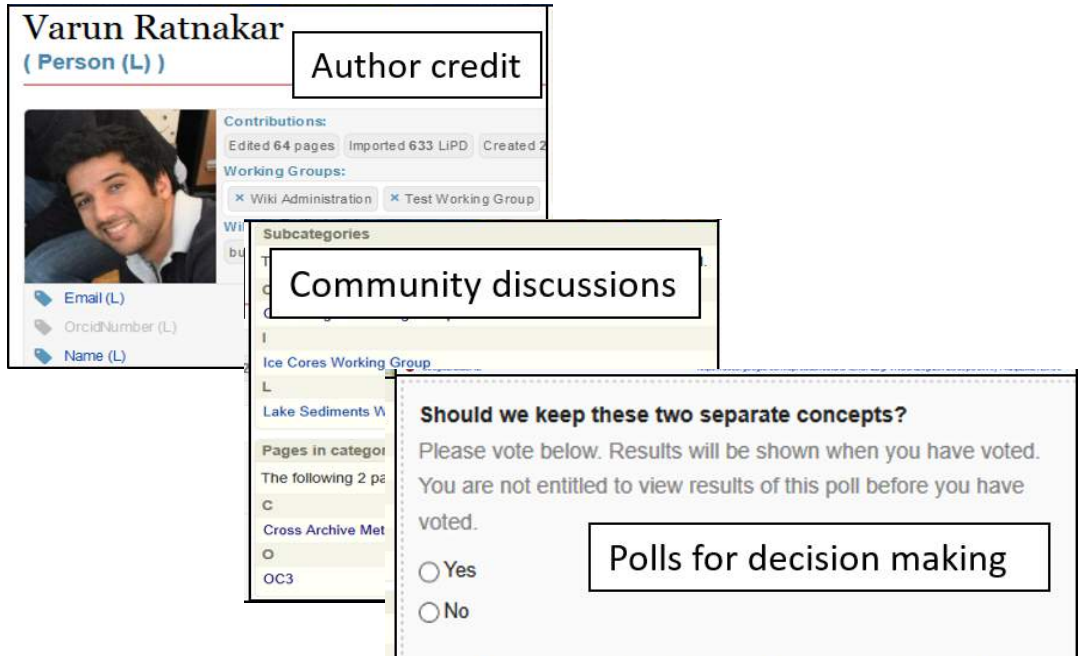
- Organic growth: As scientists annotate their datasets, they propose new metadata properties
- Crowdsourcing: Scientists proposed properties for reuse, vote on priorities
- Editorial oversight: Editors decide what properties will be in future versions

Results: A new standard for paleoclimate (PaCTS 1.0) with one (!!) single initial face-to-face meeting

Controlled Crowdsourcing: Continuous Ontology Growth



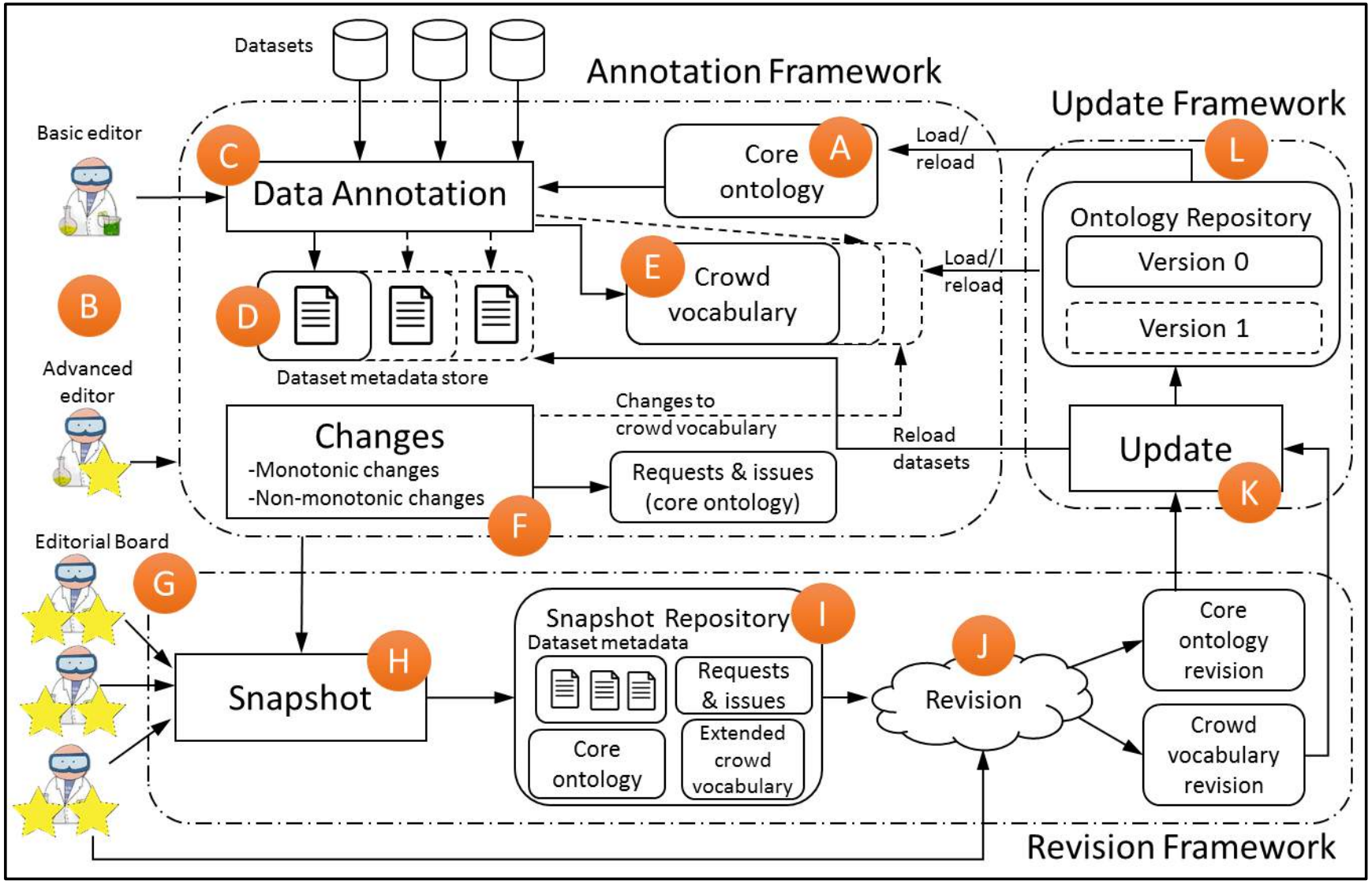




Living with a Live Ontology

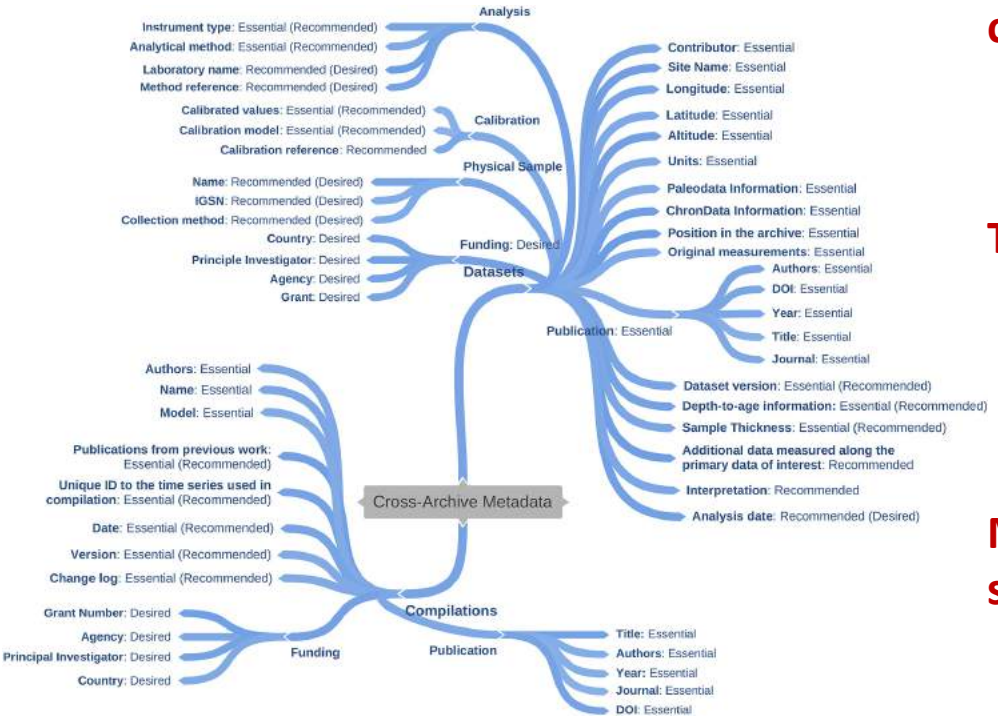
Challenge: Continuous revisions of ontologies + annotated data

- Maintain core + crowd ontology
- Editors decide when to update
  - Automated upgrades when monotonic changes
  - Semi-automated upgrades when non-monotonic changes



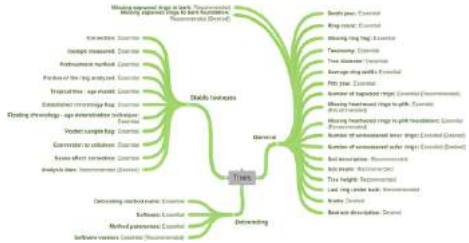
The Paleoclimate Community reporTing Standard (PaCTS)

All cores

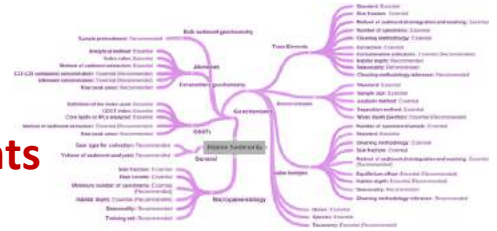


Ice cores

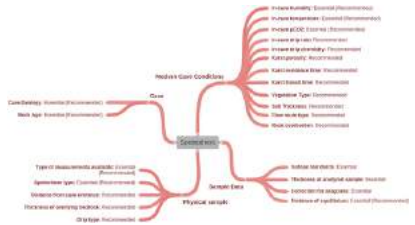
Trees



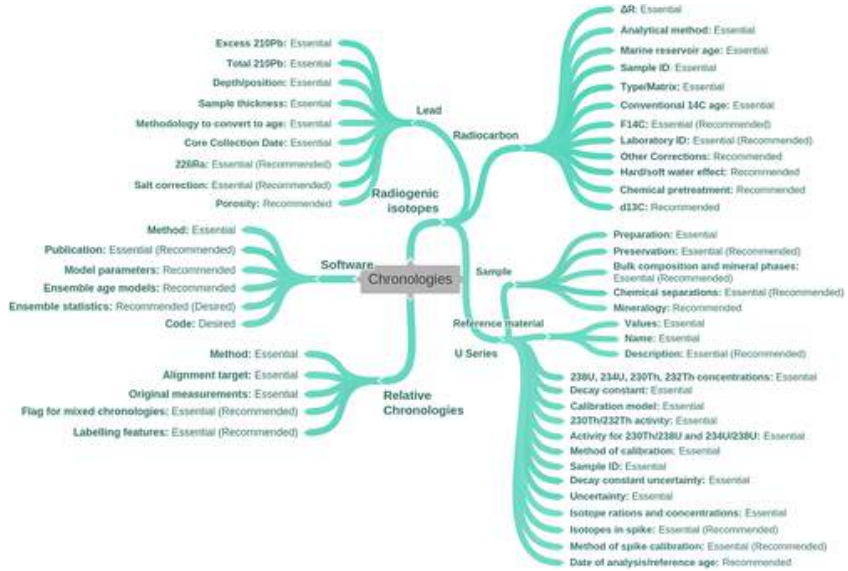
Marine sediments



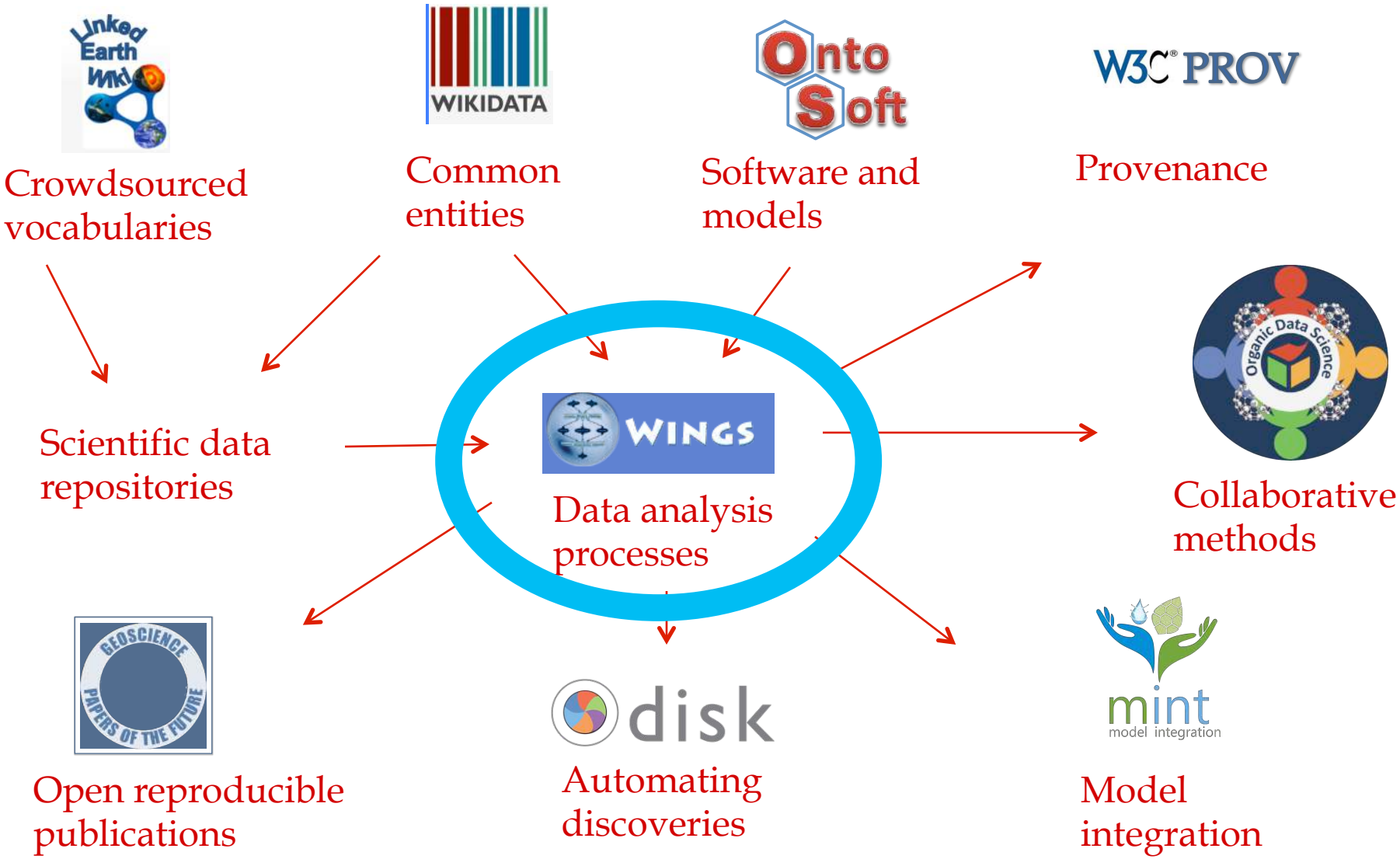
Speleothems



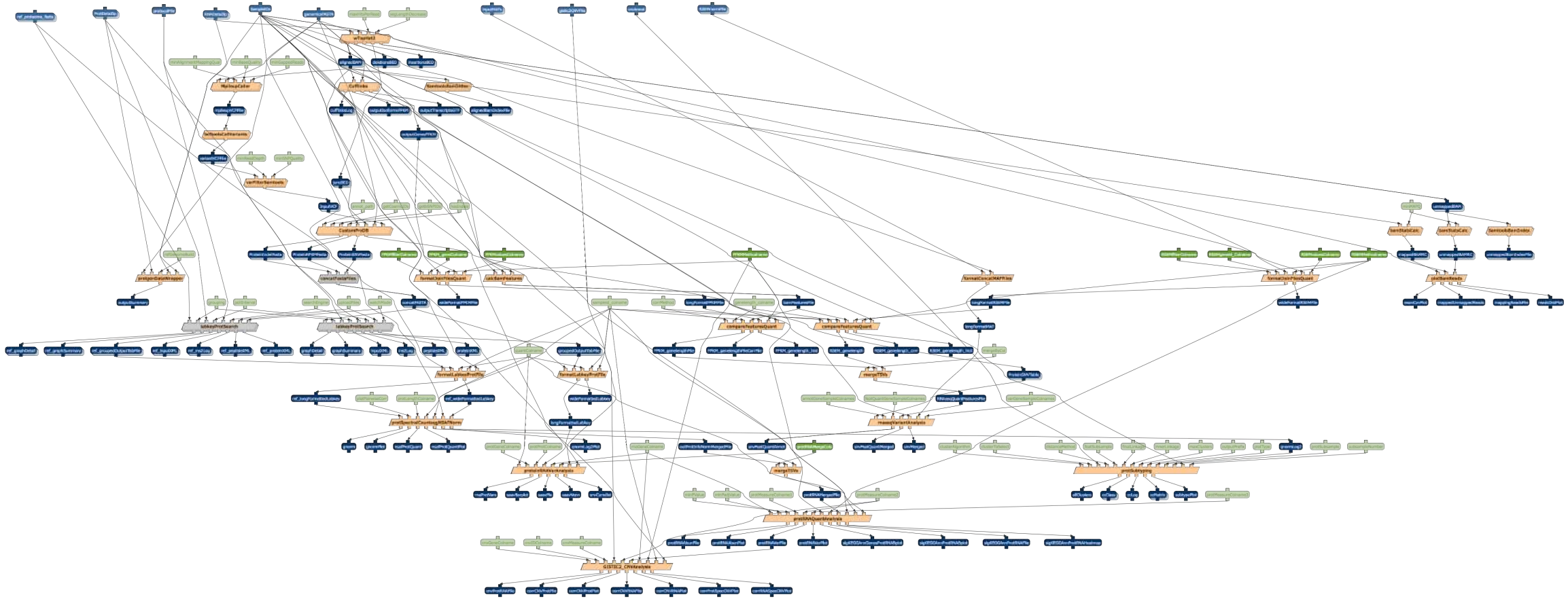
Chronologies



Capturing Scientific Knowledge



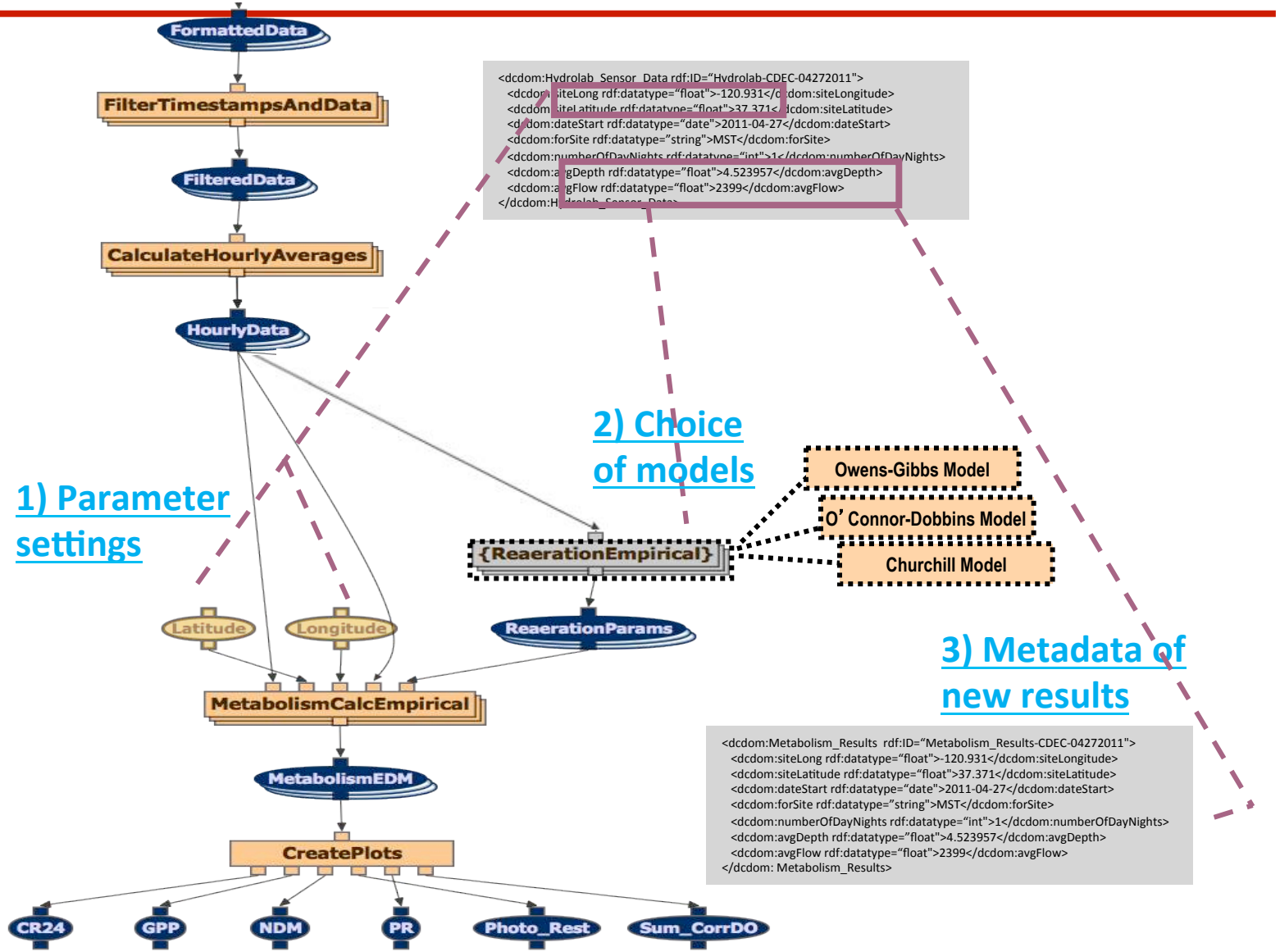
Workflows



Semantic Workflows

[Gil et al JETAI 2011; Gil et al IEEE IS 2011; Kim et al JETAI 2010; Gil IEMS 2014]

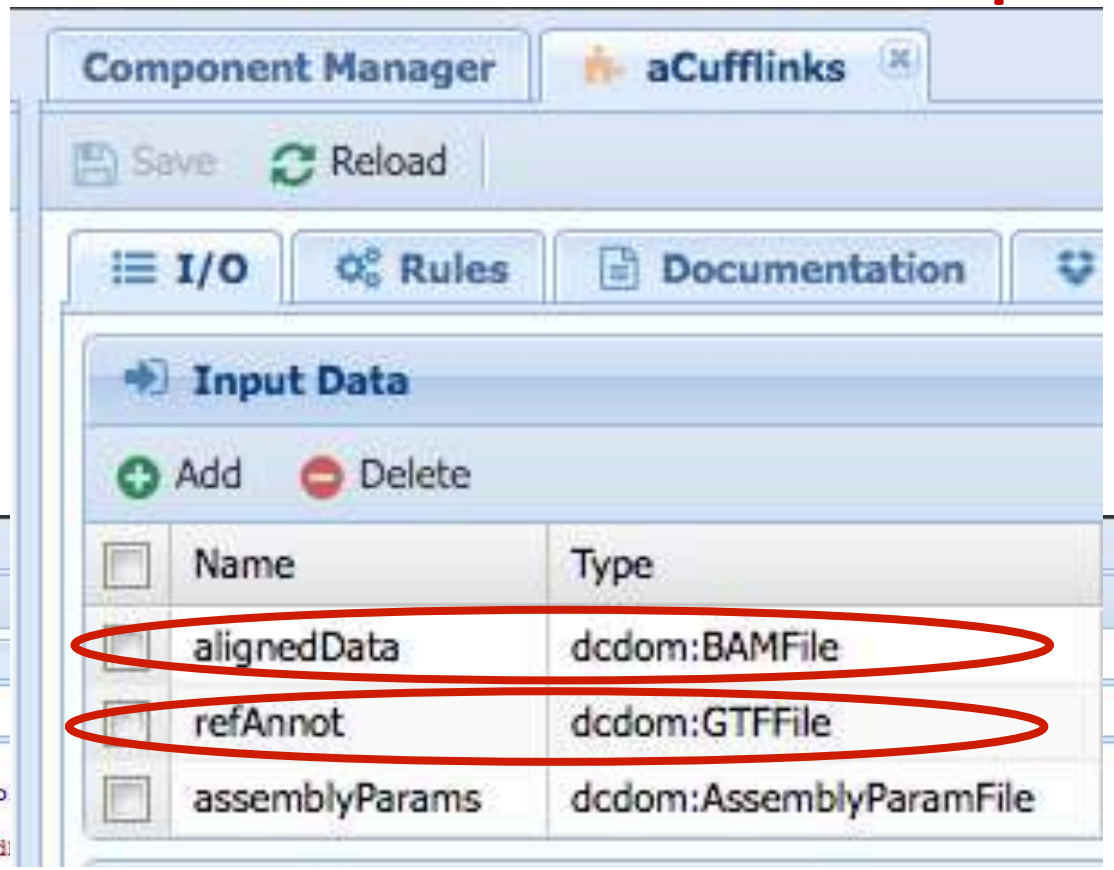
- Workflow constituents (step, input data, results, parameters) are assigned an identifier that can be referenced in constraints
- Input datasets have metadata that can be referenced in the constraints
- Constraints are used to customize a workflow to a given dataset:
  - Set parameters
  - Generate new metadata
  - Choose steps
  - Validation



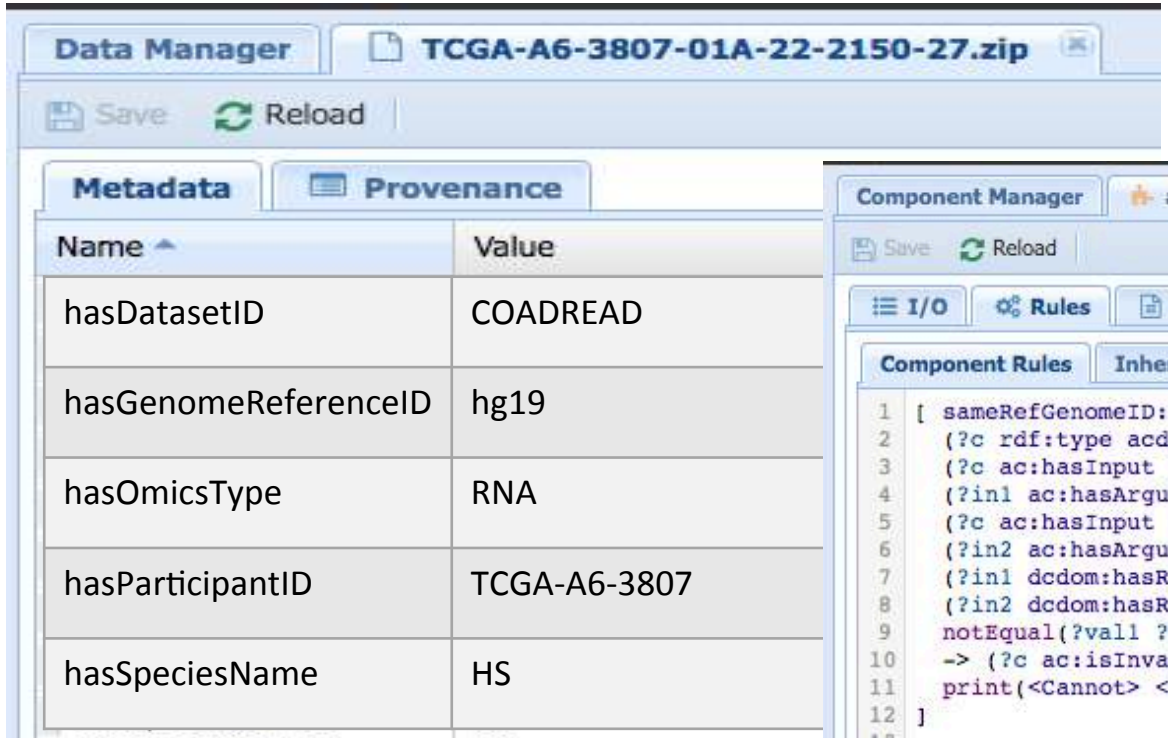
Example: Semantic Constraints on Workflows

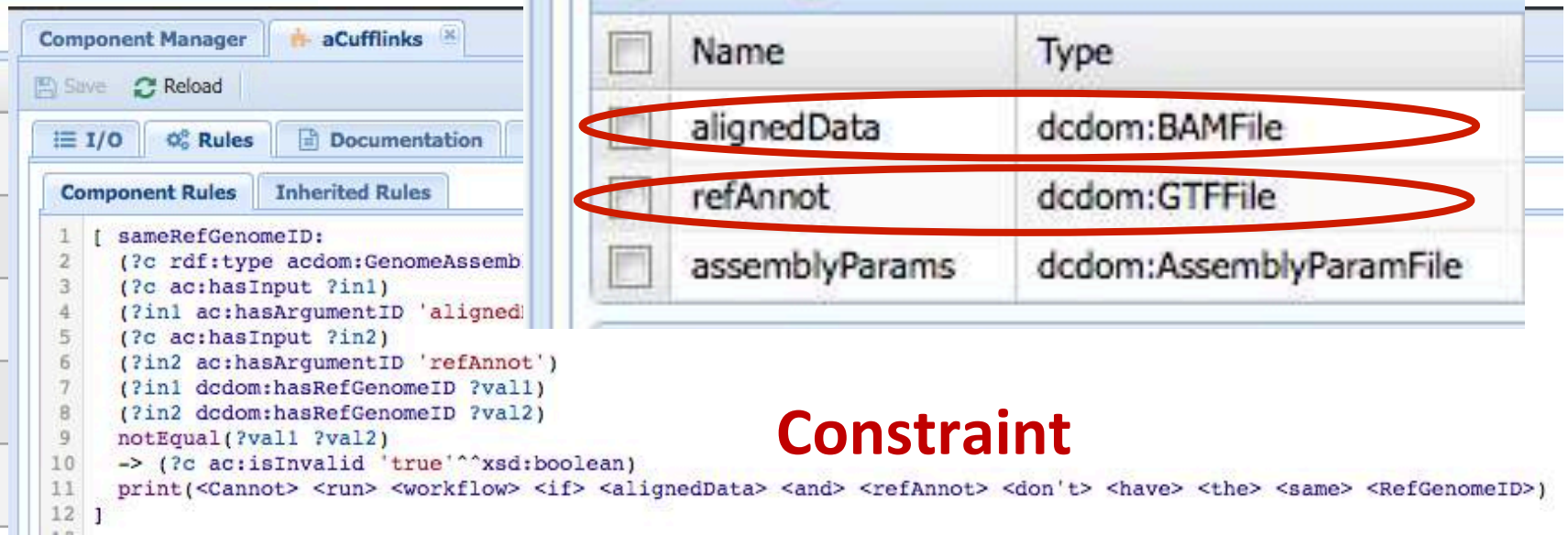
- The alignment step (TopHat) and the assembly step (Cufflinks) must be done with the same reference genome
  - Species, build, and version

Step



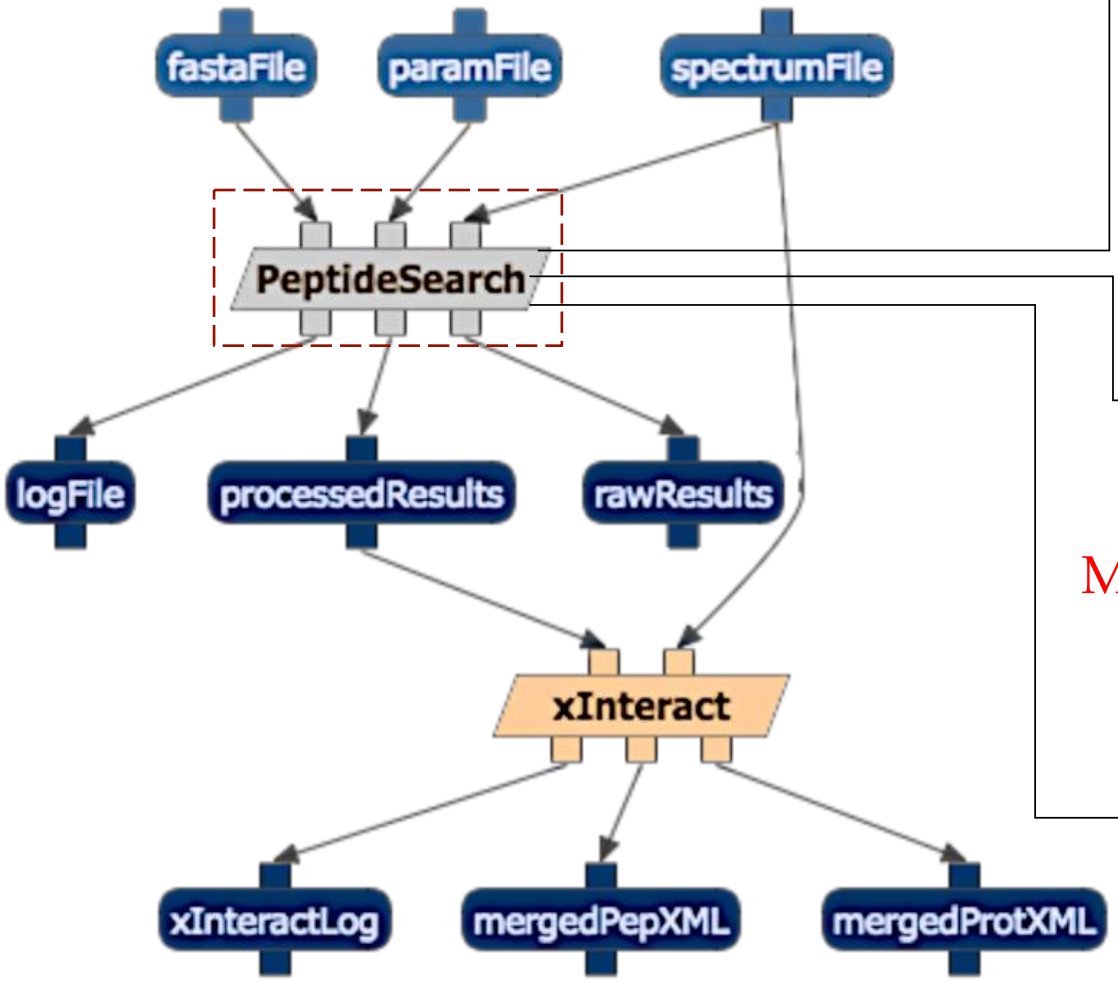
Data



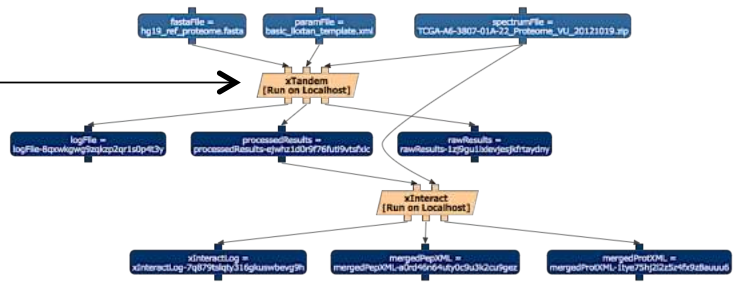


Constraint

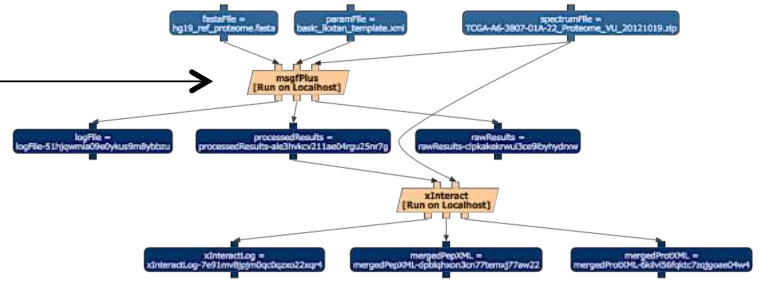
Abstractions in Semantic Workflows



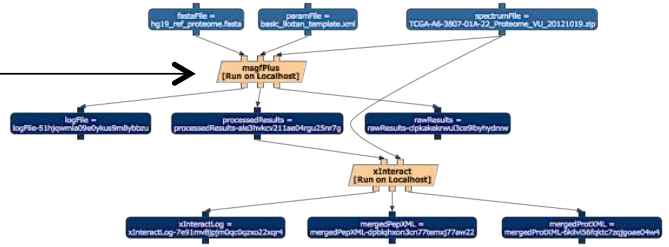
X!Tandem



MSGF+

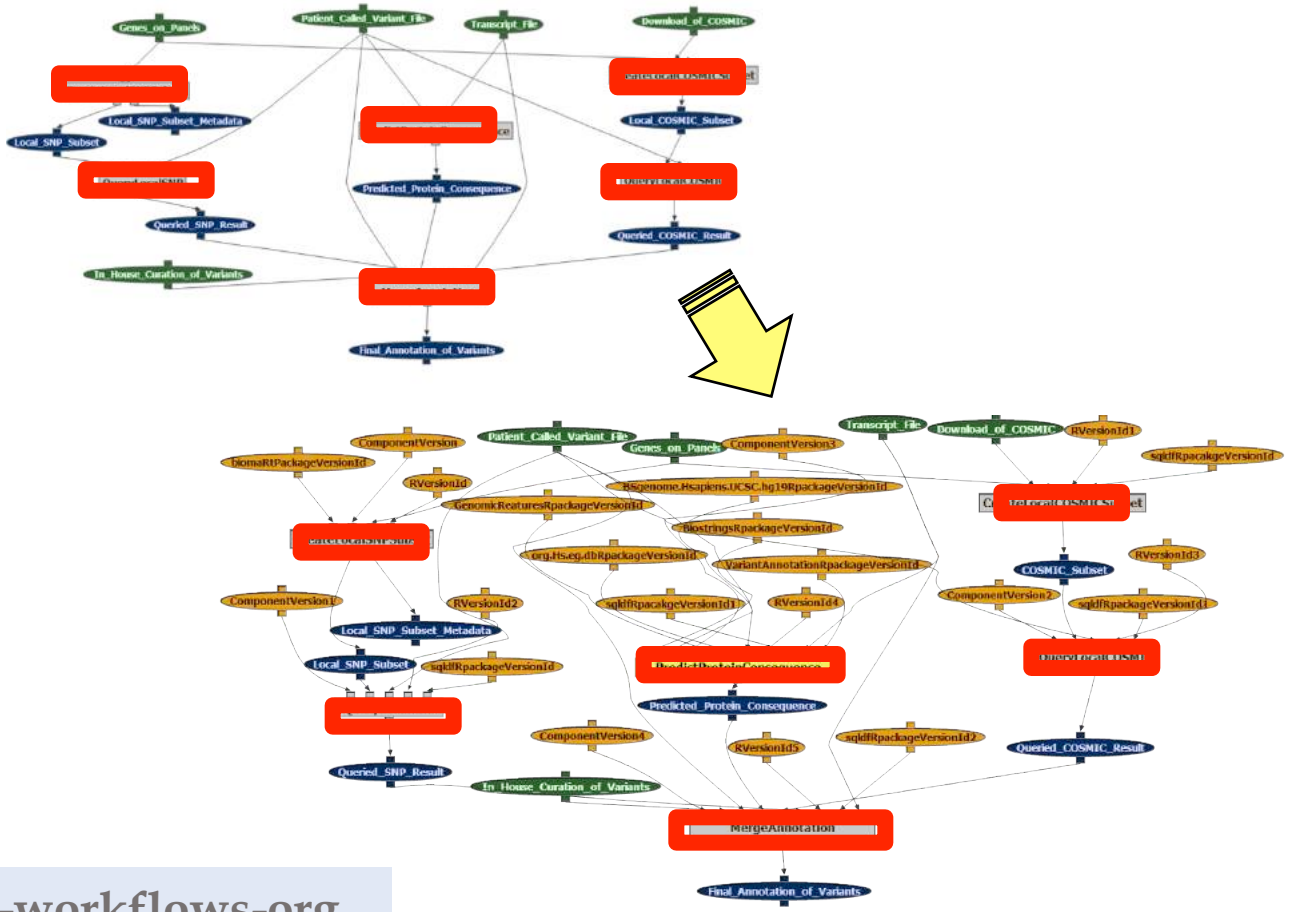


Myrimatch



WINGS: Workflow Composition

Workflow reasoning algorithms use constraint-based planning to generate executable workflows from high-level workflows



http://wings-workflows-org

unified well-formed req.

Seed workflow from request

seeded workflows

Find input data requirements

binding-ready workflows

Data source selection

bound workflows

Parameter selection

configured workflows

Workflow instantiation

workflow instances

Workflow grounding

ground workflows

Workflow ranking

top-k workflows

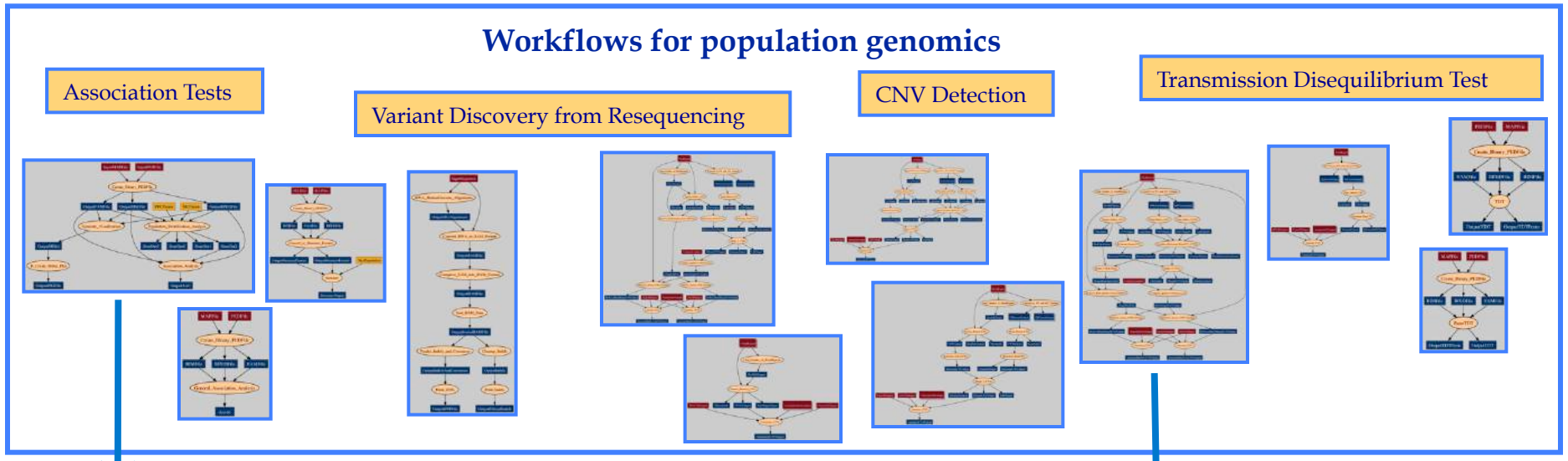
Workflow mapping

executable workflows

Reproducing Work in Population Genomics

[Gil et al 2012]

Work with Christopher Mason (Cornell University)

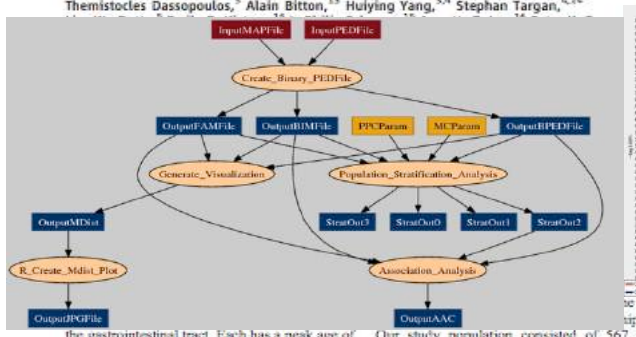


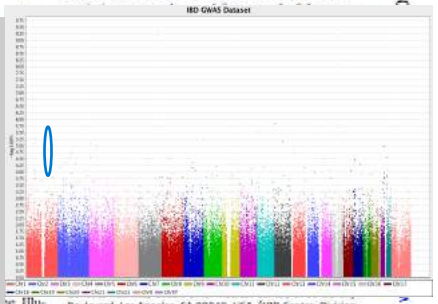
A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr, Kent D. Taylor, Steven R. Brant, John D. Rioux, Mark S. Silverberg, Mark J. Daly, A. Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths, Themistocles Dassopoulos, Alain Bitton, Huiying Yang, Stephan Targan,

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-

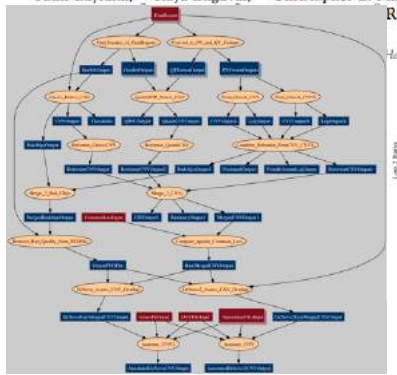


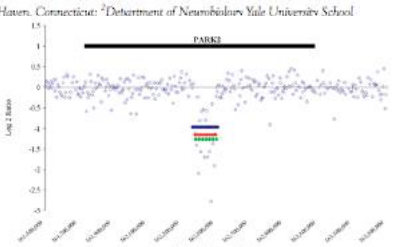


Boulevard, Los Angeles, CA 90048, USA. [4]IBD Center, Division of Gastroenterology, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, CA 90048, USA. [5]Harvey M.

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

Fatih Bayrakli, Kaya Bilguvar, Christopher E. Mason, Michael L. DiLuna, Yasar Bayri, Richard P. Lifton, Matthew W. State,

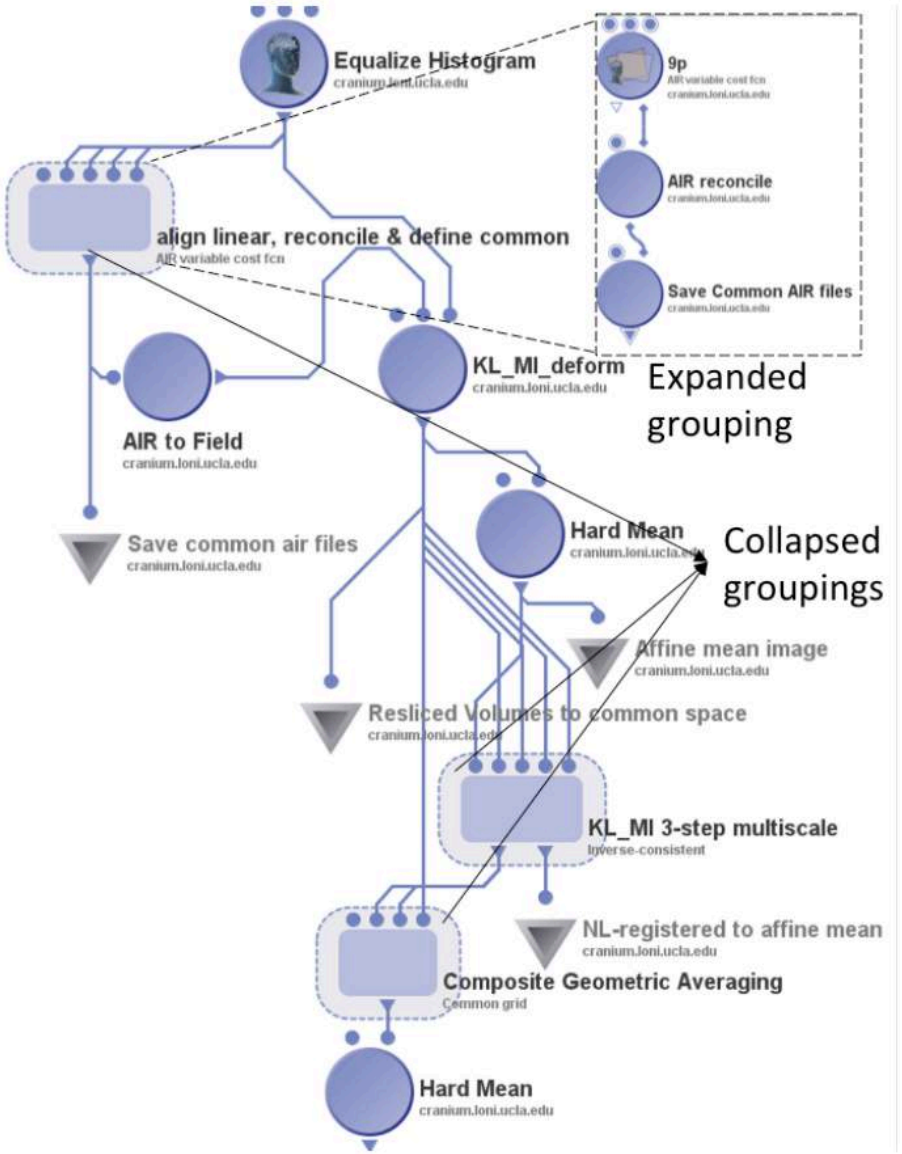


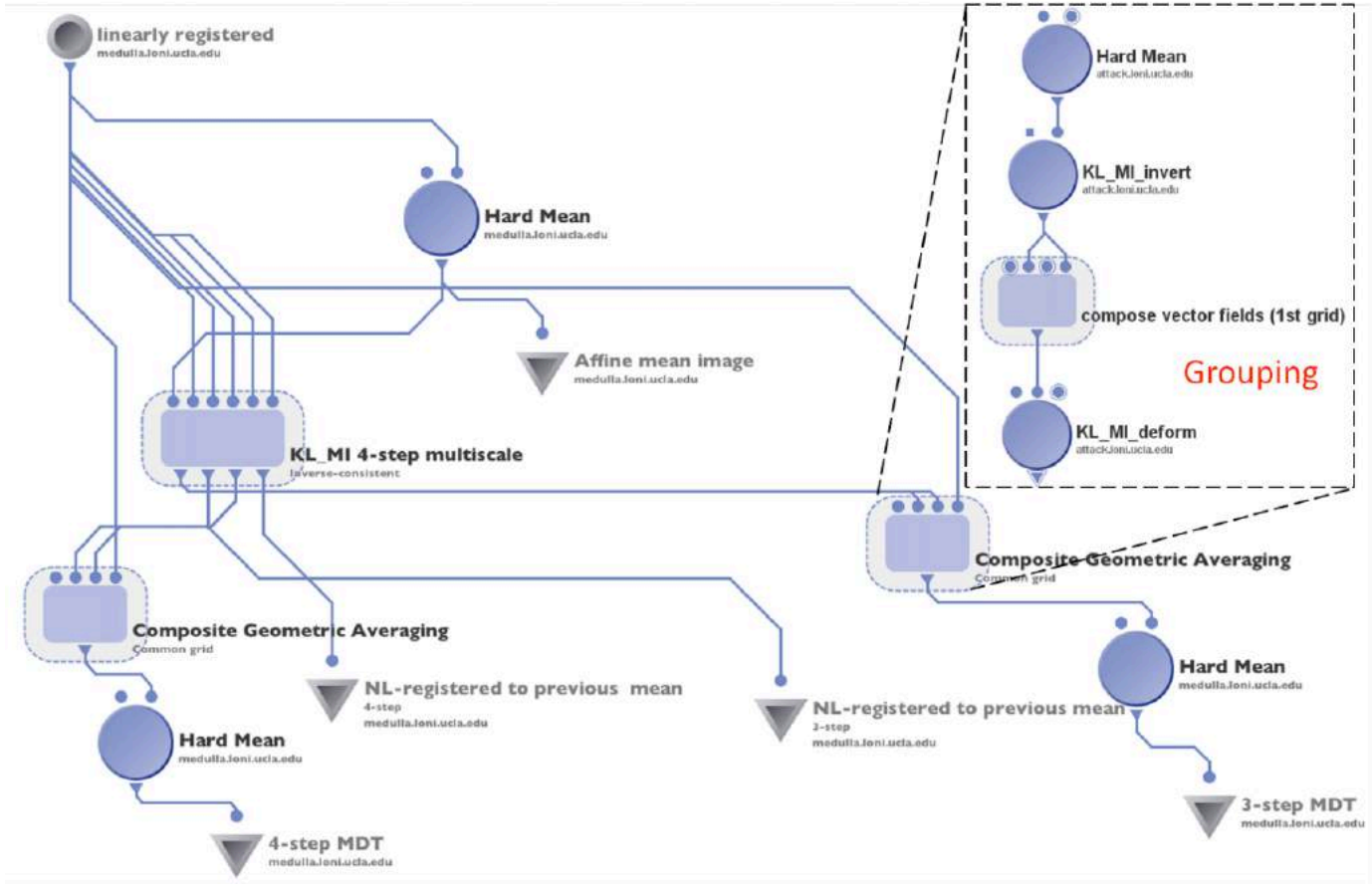


Learning Reusable Workflow Fragments

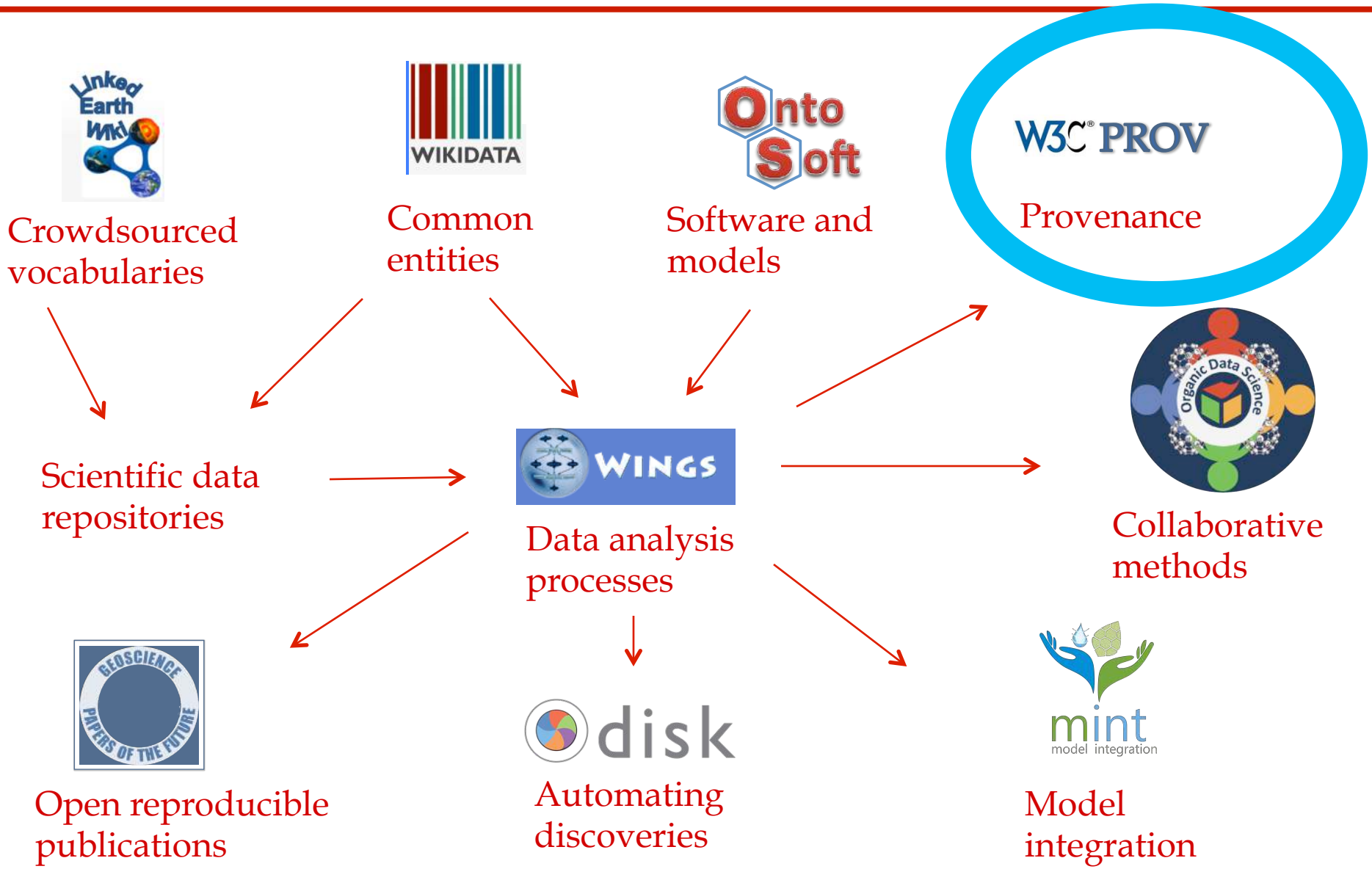
[Garijo et al FCGS 2017; Garijo et al eScience 2014]

Work with Daniel Garijo (USC) and Oscar Corcho (UPM)



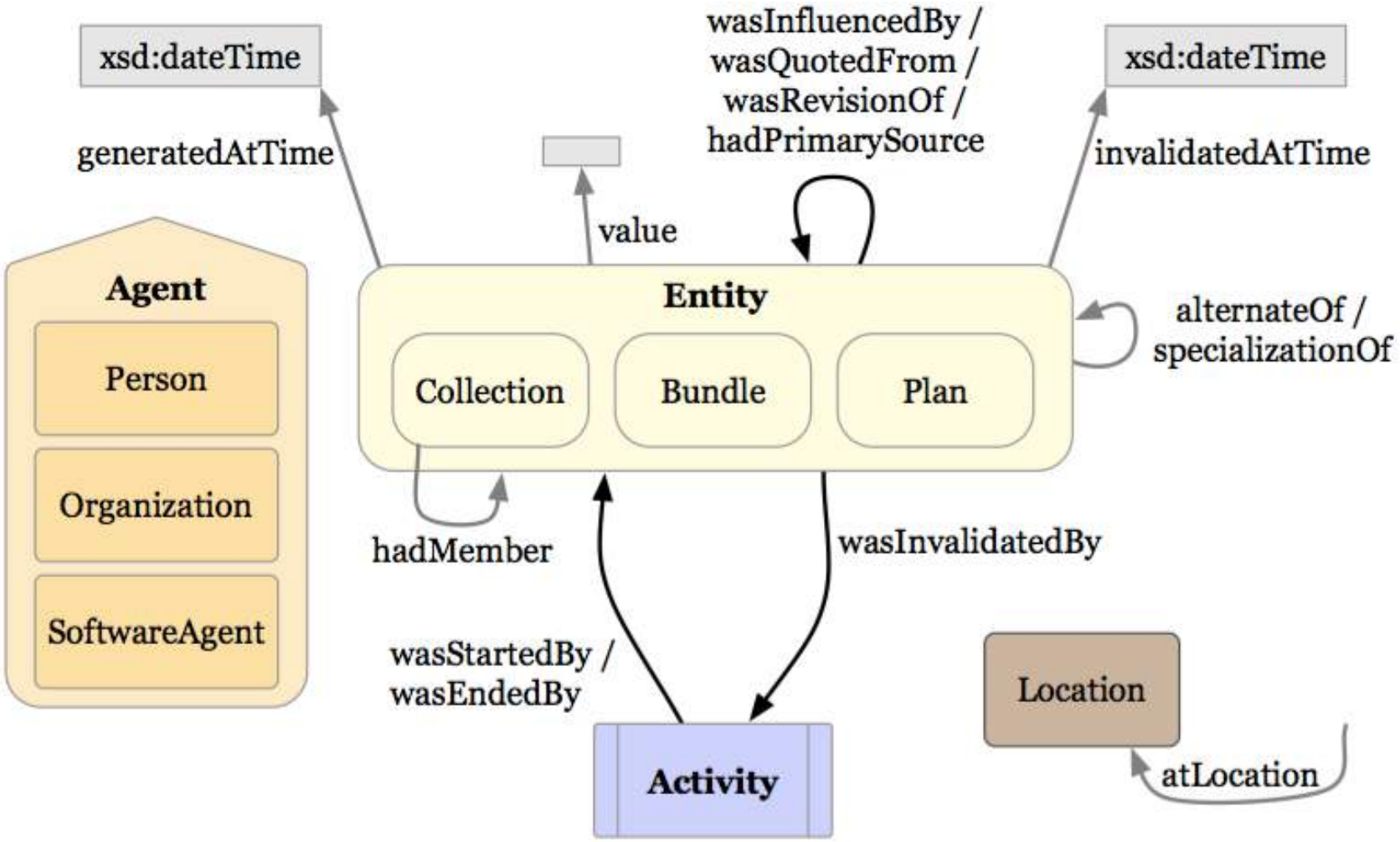


Capturing Scientific Knowledge



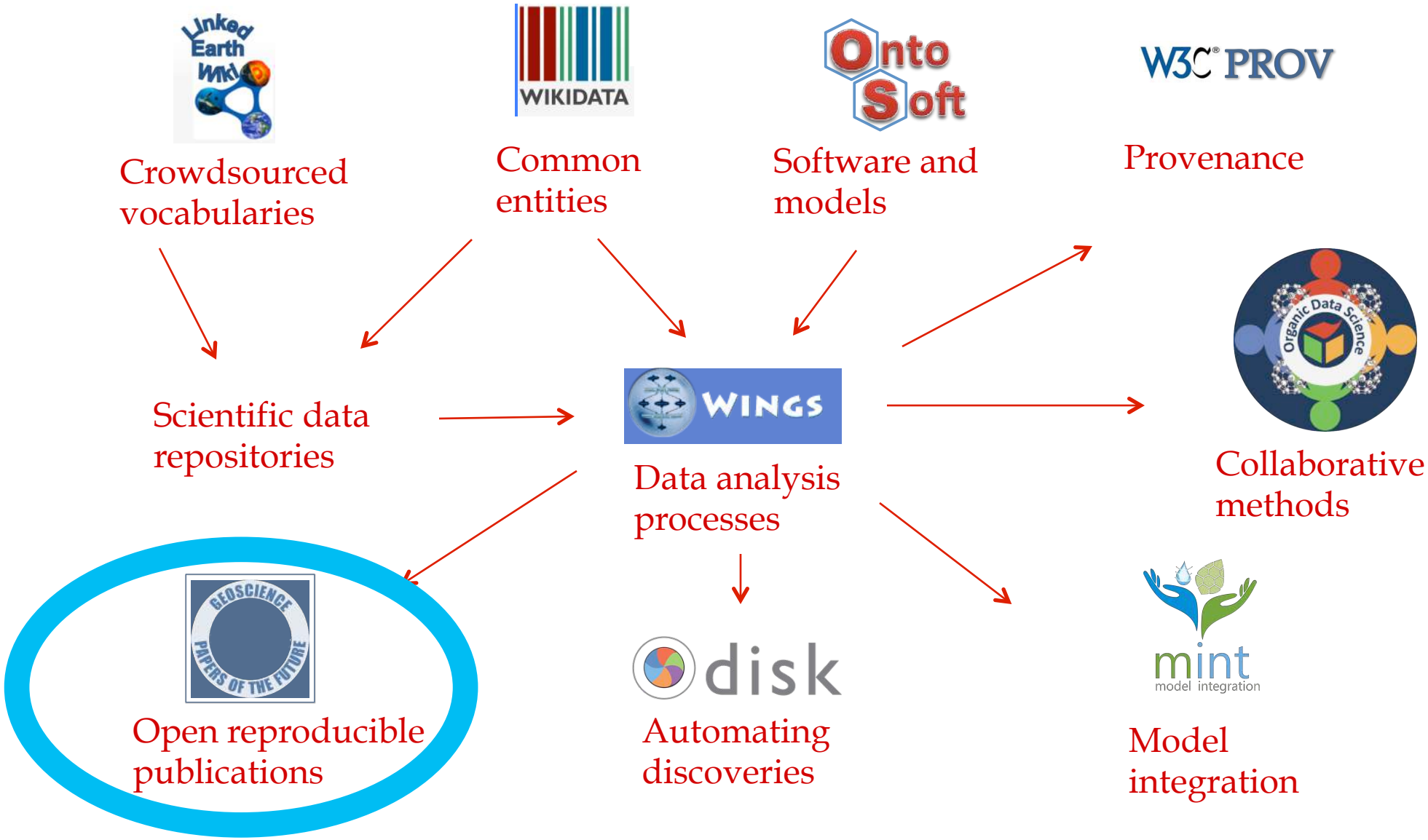
The W3C PROV Provenance Standard

[Gil and Miles 2013; Groth and Moreau 2013; Moreau et al 2014]



https://www.w3.org/2011/prov/

Capturing Scientific Knowledge



www.scientificpaperofthefuture.org

[Gil et al ESS 2016; Essawy et al EMS 2017; Goodman et al PLOS CB 2014]

Scientific Paper of the Future

Modern Paper

Text:
Narrative of the method, some data is in tables, figures/plots, and the software used is mentioned

Data:
Include data as supplementary materials and pointers to data repositories

Open Science

Sharing:
Deposit data and software (and provenance/workflow) in publicly shared repositories

Open licenses:
Open source licenses for data and software (and provenance/workflow)

Metadata:
Structured descriptions of the characteristics of data and software (and provenance/workflow)

Reproducible Publication

Software:
For data preparation, data analysis, and visualization

Provenance and methods:
Workflow/scripts specifying dataflow, codes, configuration files, parameter settings, and runtime dependencies

Digital Scholarship

Persistent identifiers:
For data, software, and authors (and provenance/workflow)

Citations:
Citations for data and software (and provenance/workflow)

AGU American Geophysical Union

Earth and Space Science
AN OPEN ACCESS AGU JOURNAL

Special Section: Geoscience Papers of the Future

NATURE REVIEWS | NEUROSCIENCE

Scanning the horizon: towards transparent and reproducible neuroimaging research

Russell A. Poldrack[1], Chris I. Baker[2], Joke Durnez[1,3], Krzysztof J. Gorgolewski[1], Paul M. Matthews[4], Marcus R. Munafò[5,6], Thomas E. Nichols[7], Jean-Baptiste Poline[8], Edward Vul[9] and Tal Yarkoni[10]

Towards the neuroimaging paper of the future

In this Analysis article, we have outlined a number of problems with current practice and made suggestions for improvements. Here, we outline what we would like to see in the neuroimaging paper of the future, inspired by related work in the geosciences[71].

SEG SOCIETY OF EXPLORATION GEOPHYSICISTS

GEOPHYSICS Call for Papers

Reproducible Research:
Geophysics Papers of the Future

AI magazine

On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications

Odd Erik Gundersen, Yolanda Gil, David W. Aha

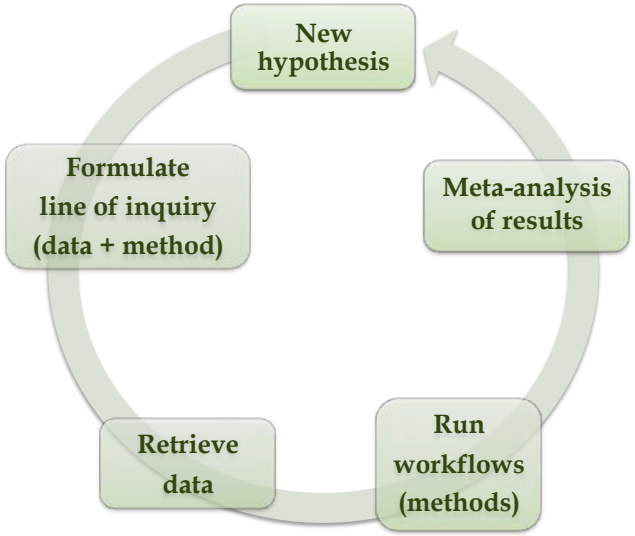
AI for Systematic Scientific Data Analysis

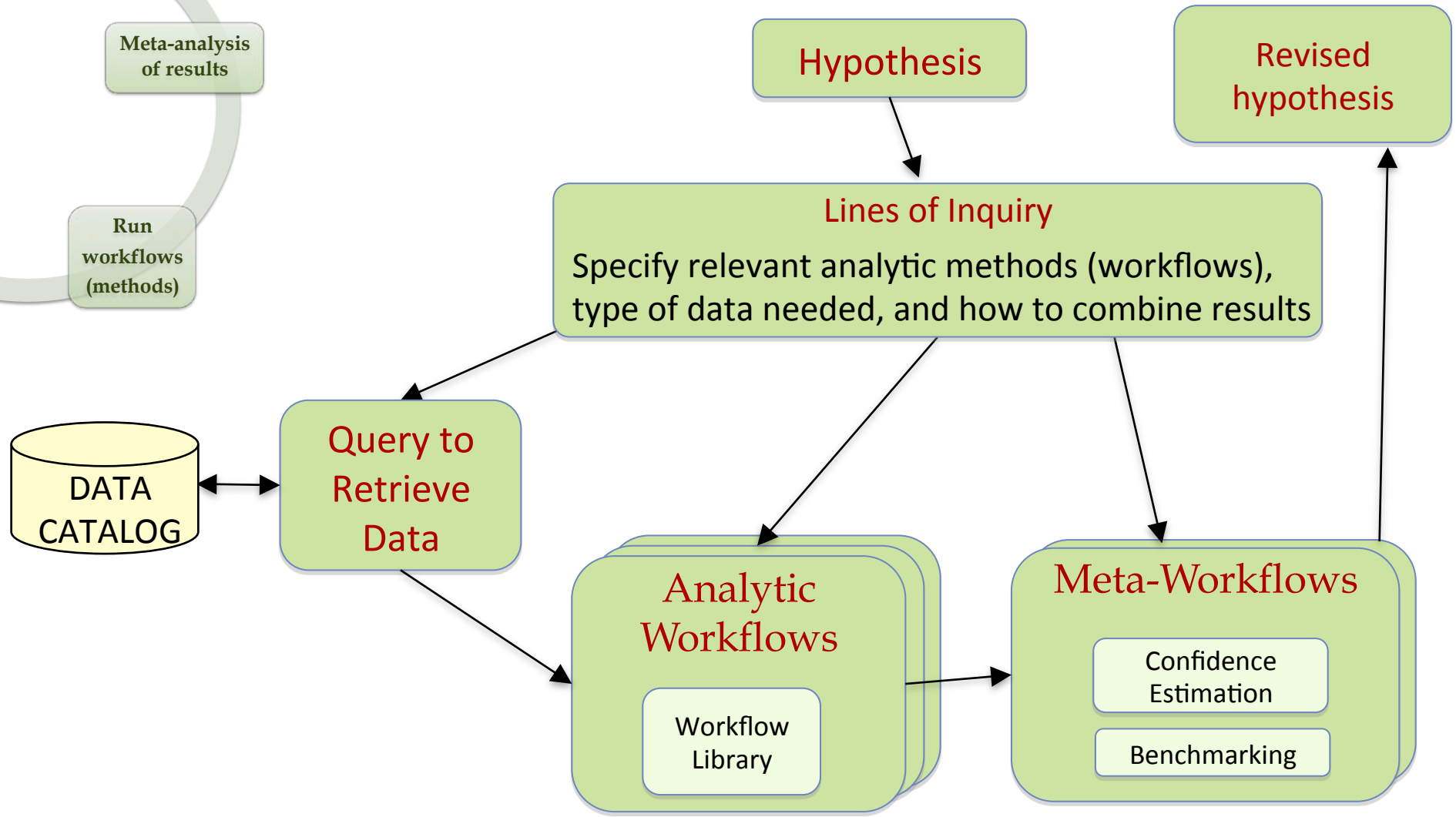
Automated Discovery: Hypotheses and Lines of Inquiry

[Gil et al ACS 2016; Gil et al AAAI 2017; Garijo et al 2017; Srivastava et al PSB 2019; Mallick et al 2020]

disk

With Parag Malick, Ravali Adusumilli, Hunter Boyce (Stanford); Arunima Srivastava (OSU); Daniel Garijo, Varun Ratnakar, Rajiv Mayani (USC/ISI); Thomas Yu (Sage Bionetworks)



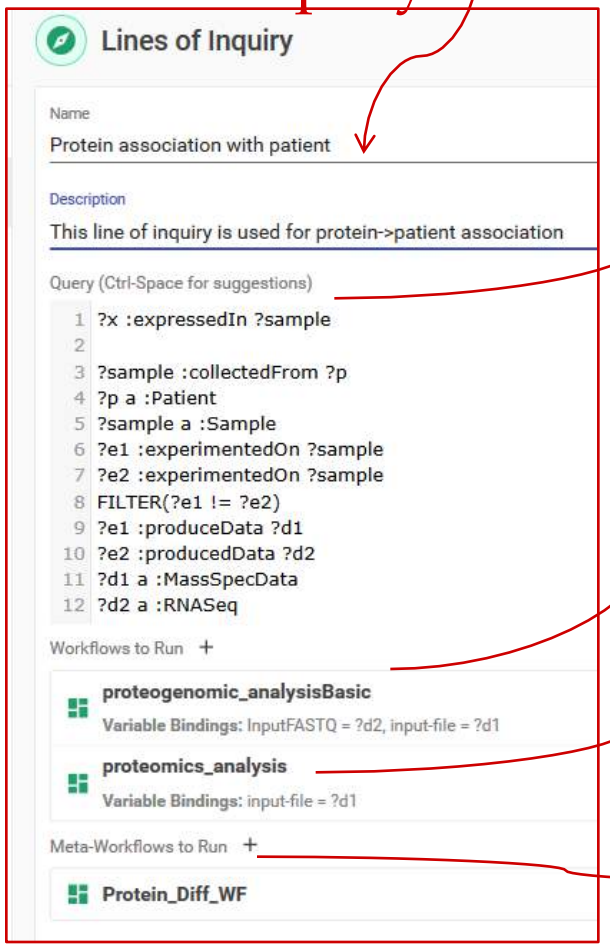


Automated Discovery in DISK

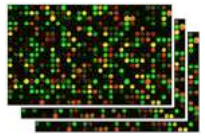
hypothesis

Protein PRKCDBP is expressed in samples of patient P36

lines of inquiry



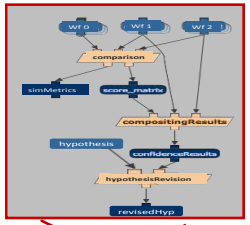
data



revision

PRKCDBP mutation is expressed in P36

workflows



meta-workflows

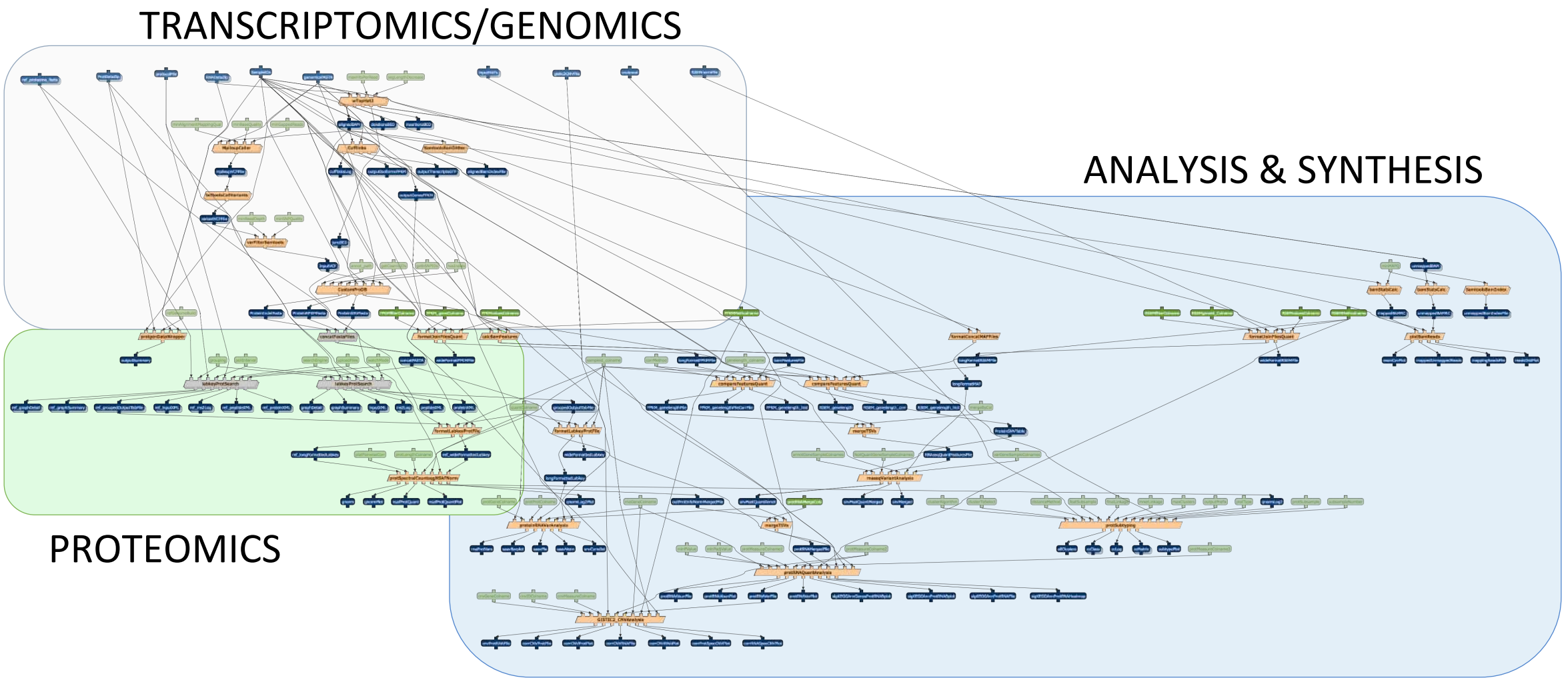
RDF

W3C® PROV

OPMW Workflow repository

Integrating Knowledge through Workflows

disk



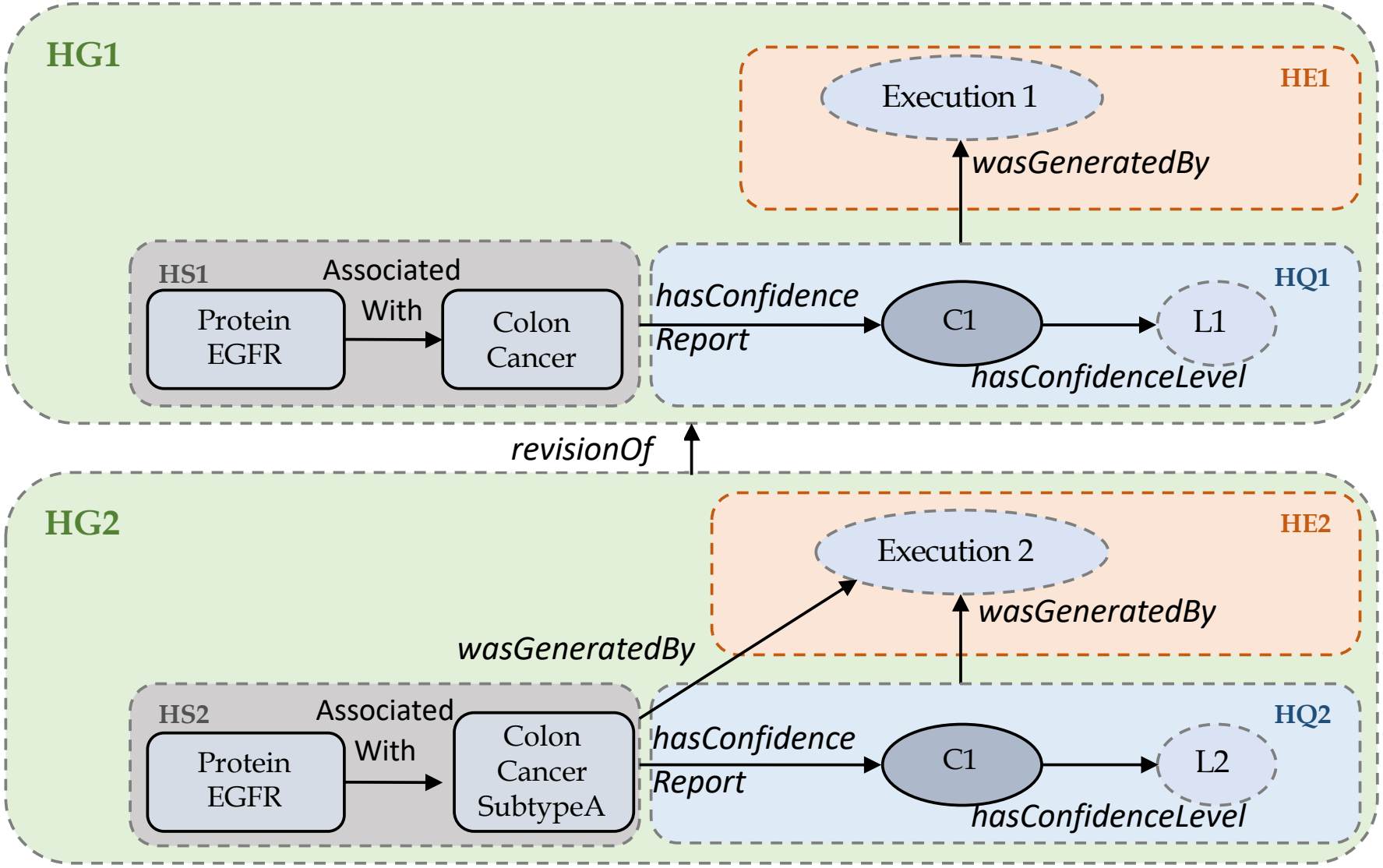
Representing Hypotheses

Statement

Qualifier

Evidence

Evolution



DISK Hypothesis Ontology:

http://disk-project.org/ontology/disk

Reproducing a Seminal Cancer Omics Paper

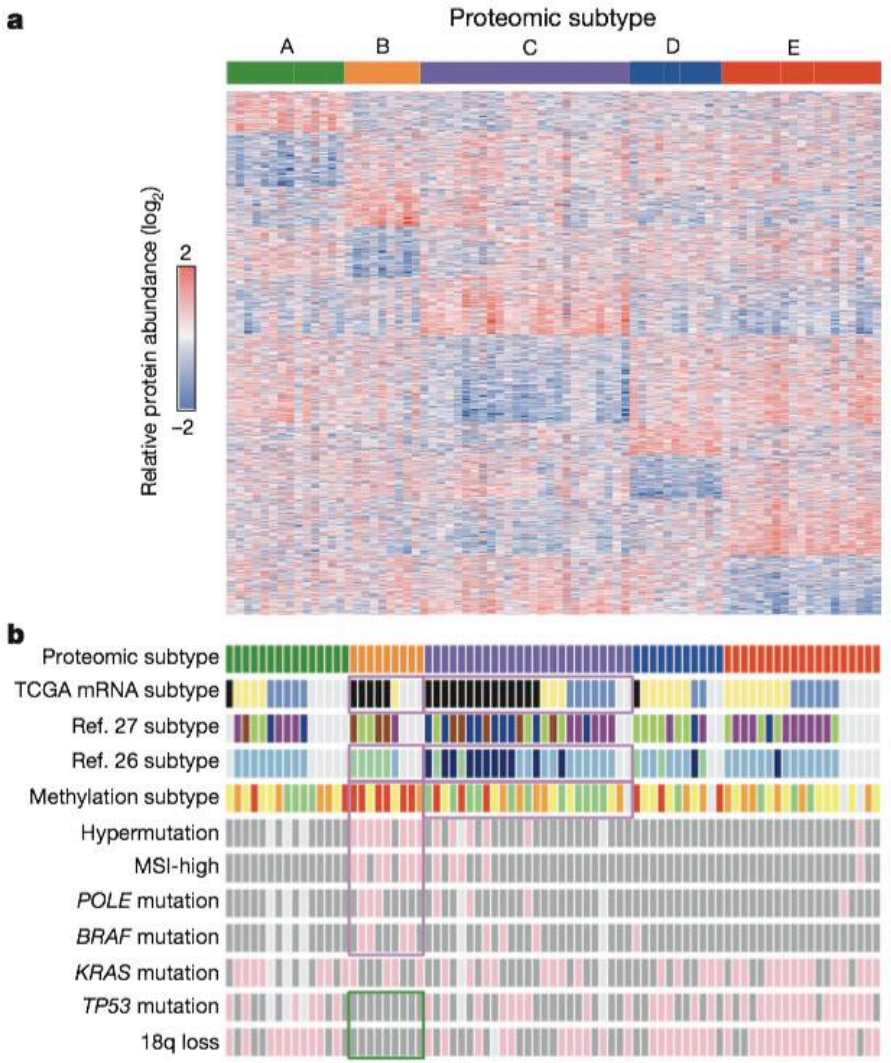
nature International weekly journal of science

Proteogenomic characterization of human colon and rectal cancer

Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C. Chambers, Lisa J. Zimmerman, Kent F. Shaddox, Sangtae Kim, Sherri R. Davies, Sean Wang, Pei Wang, Christopher R. Kinsinger, Robert C. Rivers, Henry Rodriguez, R. Reid Townsend, Matthew J. C. Ellis, Steven A. Carr, David L. Tabb, Robert J. Coffey, Robbert J. C. Slebos, Daniel C. Liebler & the NCI CPTAC

Affiliations | Contributions | Corresponding author

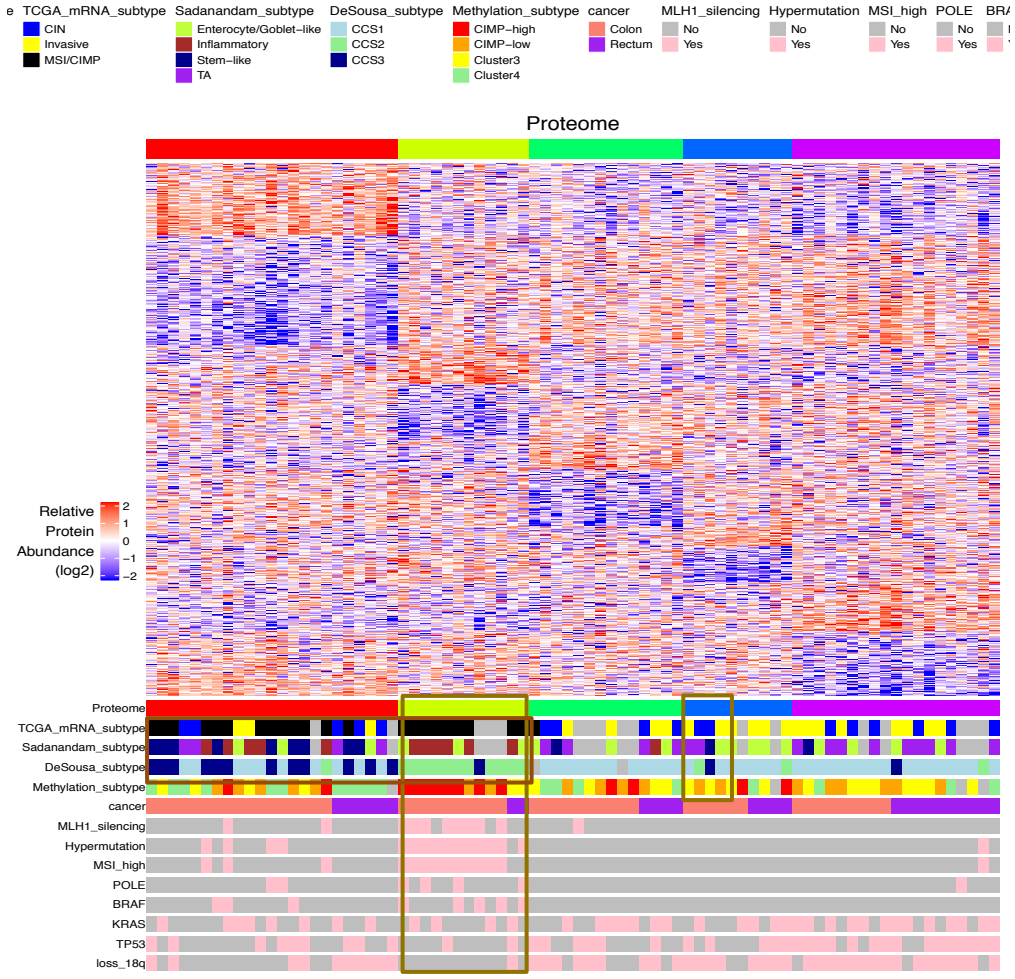
Nature **513**, 382–387 (18 September 2014) | doi:10.1038/nature13438
Received 19 September 2013 | Accepted 02 May 2014 | Published online 20 July 2014



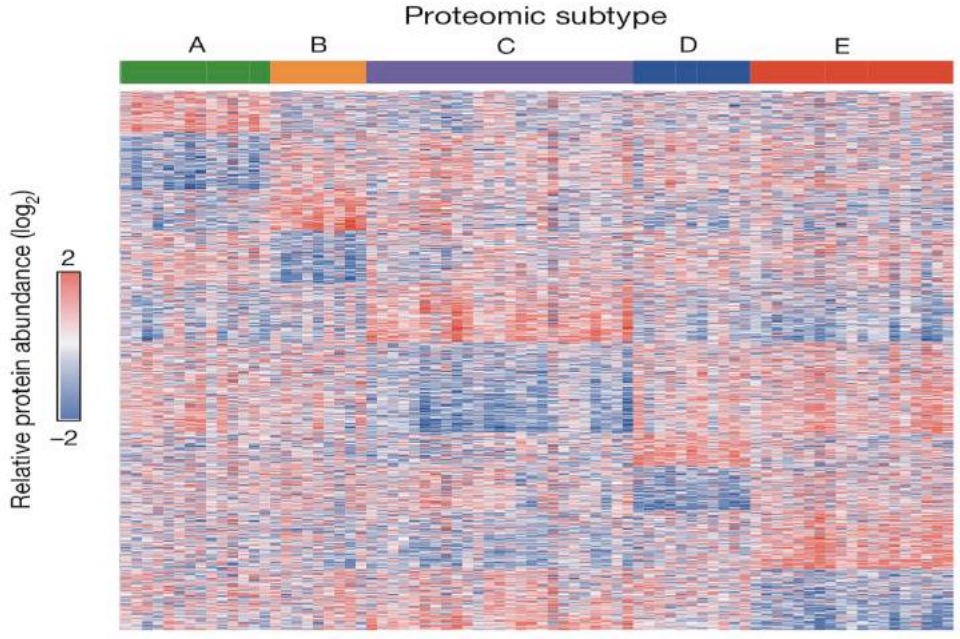
http://www.nature.com/nature/journal/v513/n7518/full/nature13438.html

Patient Proteomic Subtyping

Original
[Zhang et al 2014]

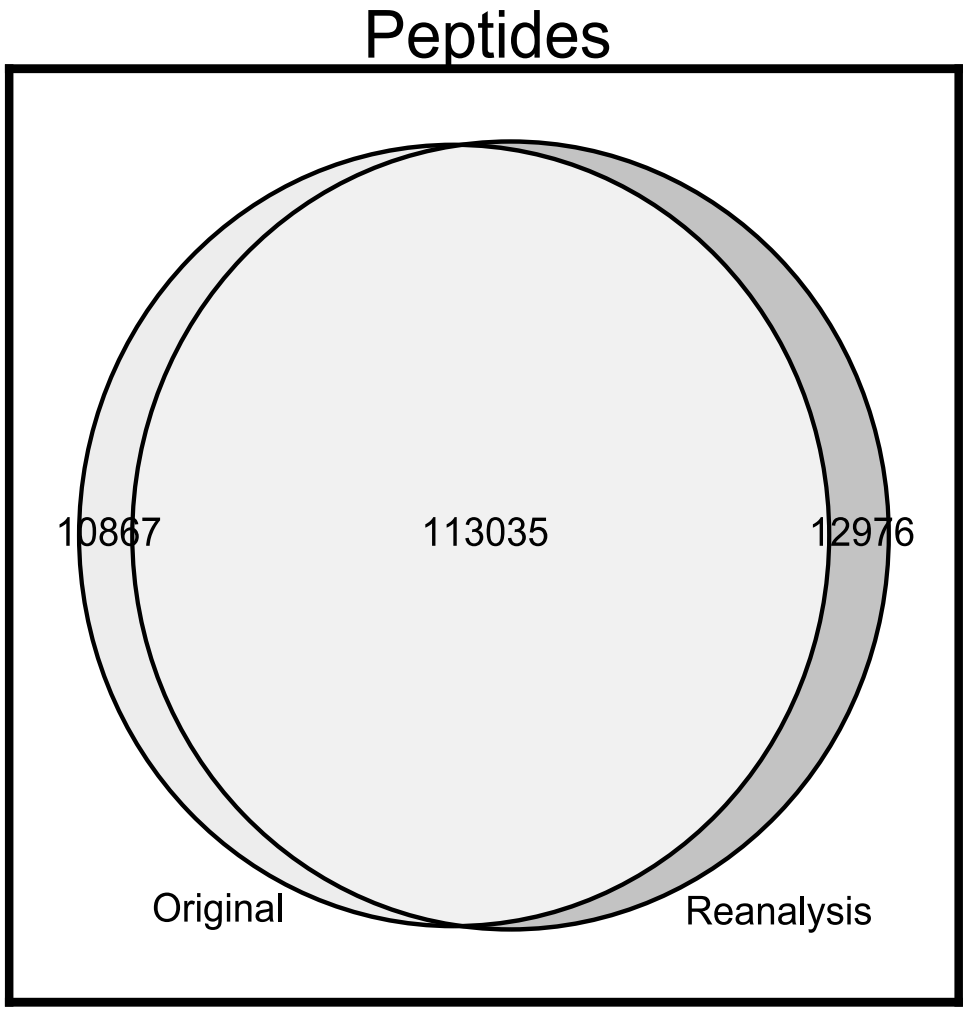


Reanalysis
[Mallick et al 2020]

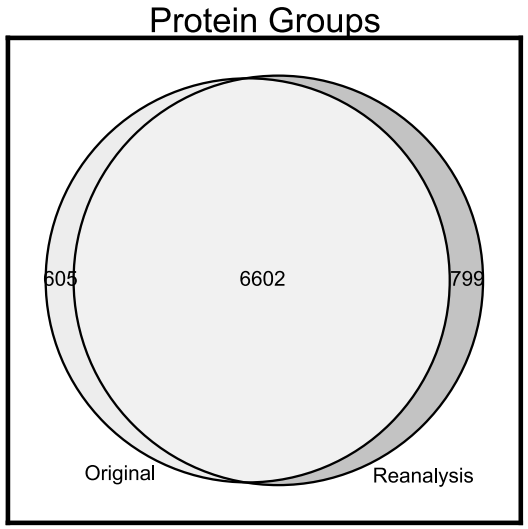


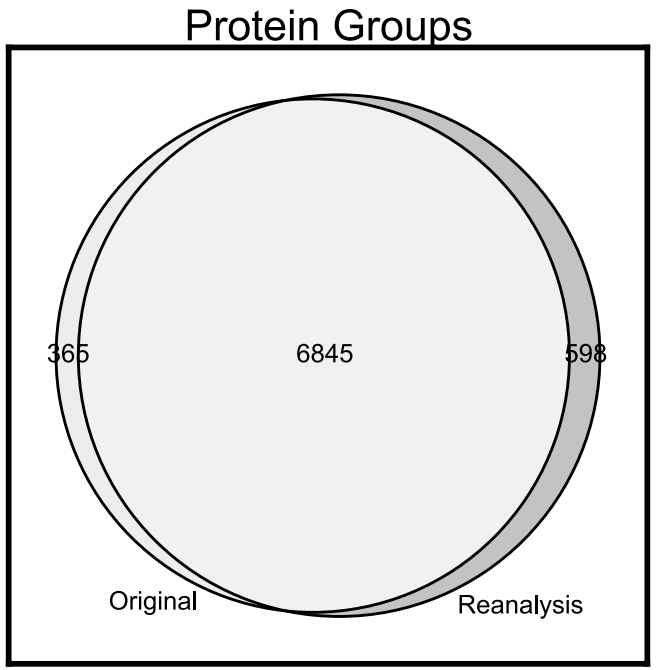
Large Differences in Protein Identification

10% Difference

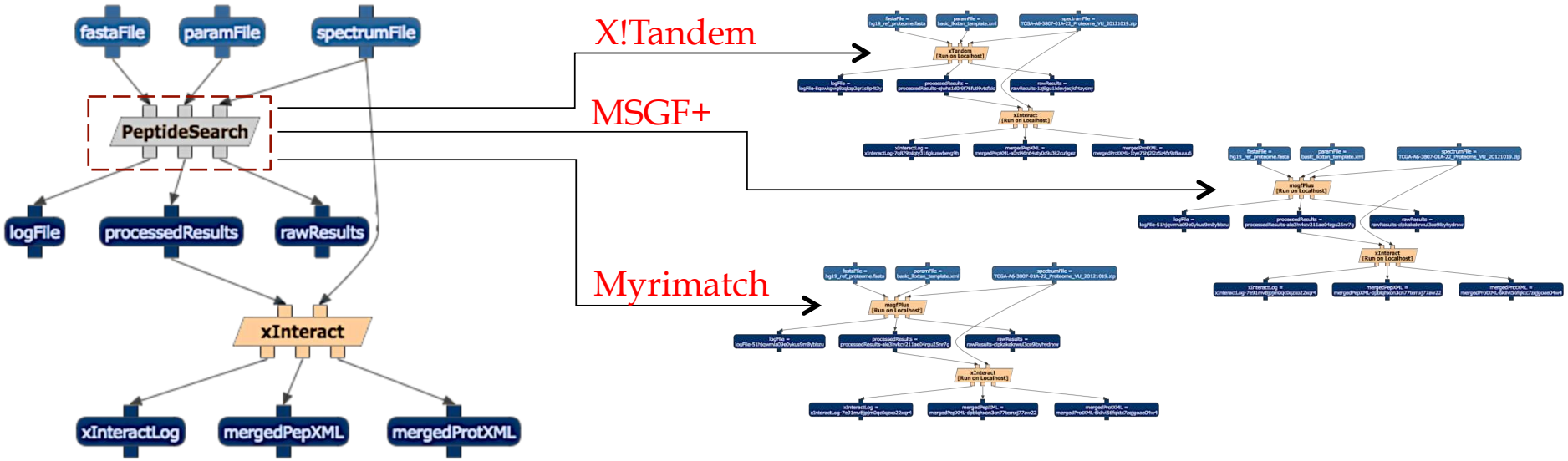


10% to 5% Difference





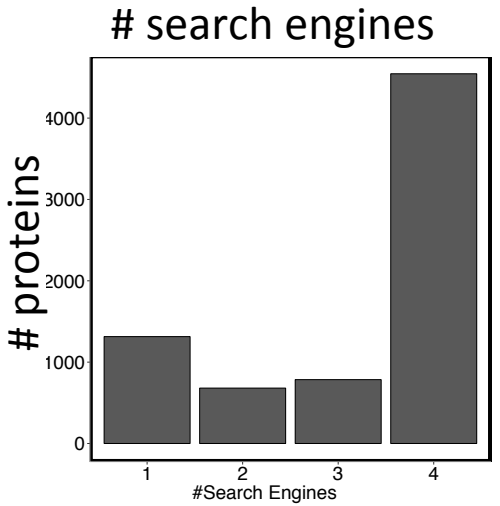
Multi-Omics: Sensitivity to Methods/Software

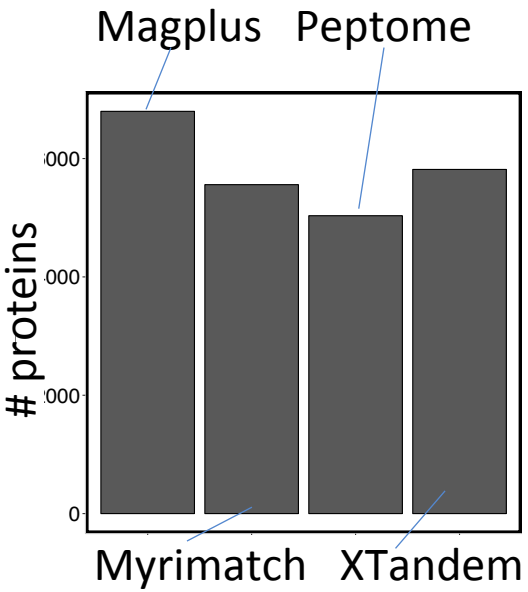


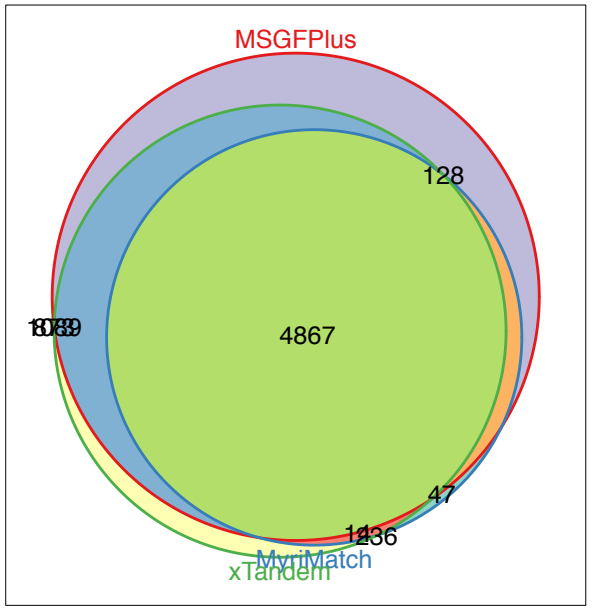
35% of protein identifications are not robust to changing just one analysis step

Detection of important proteins in colorectal cancer:

Protein	MSGF+	Myrimatch	X!Tandem
APC	Found	Not found	Found
PIK3CA	Found	Not found	Not found
VTI1A	Found	Not found	Found
PMS1	Found	Not found	Not found
NTHL1	Found	Found	Not found

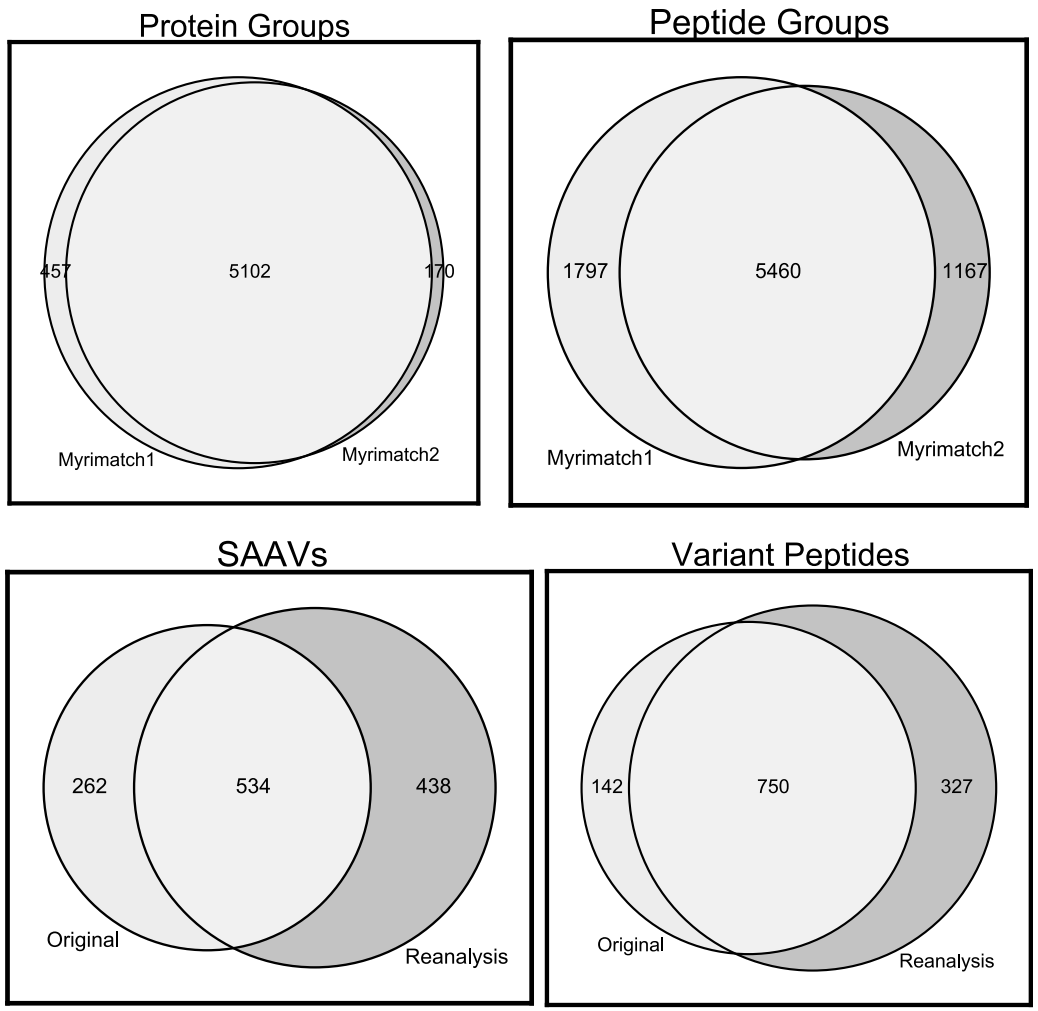




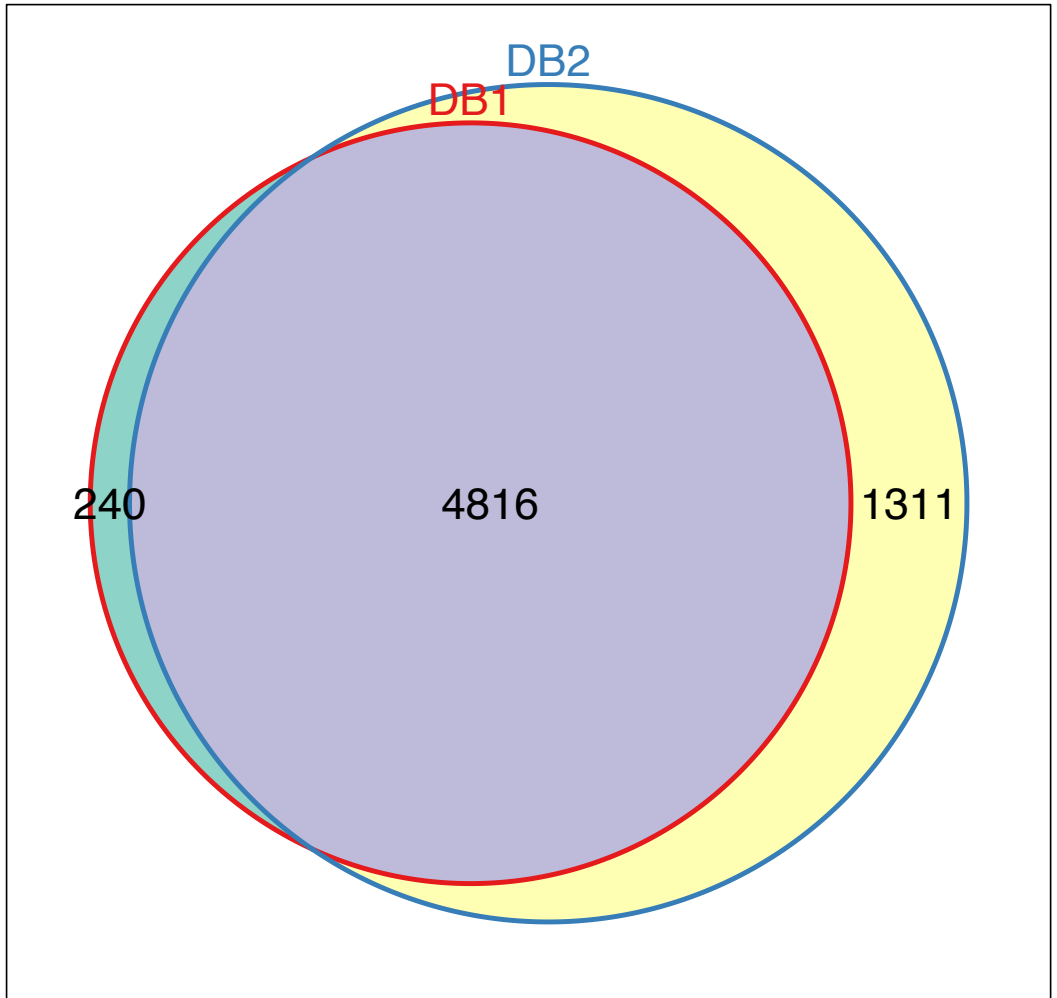


Multi-Omics: Method Sensitivity

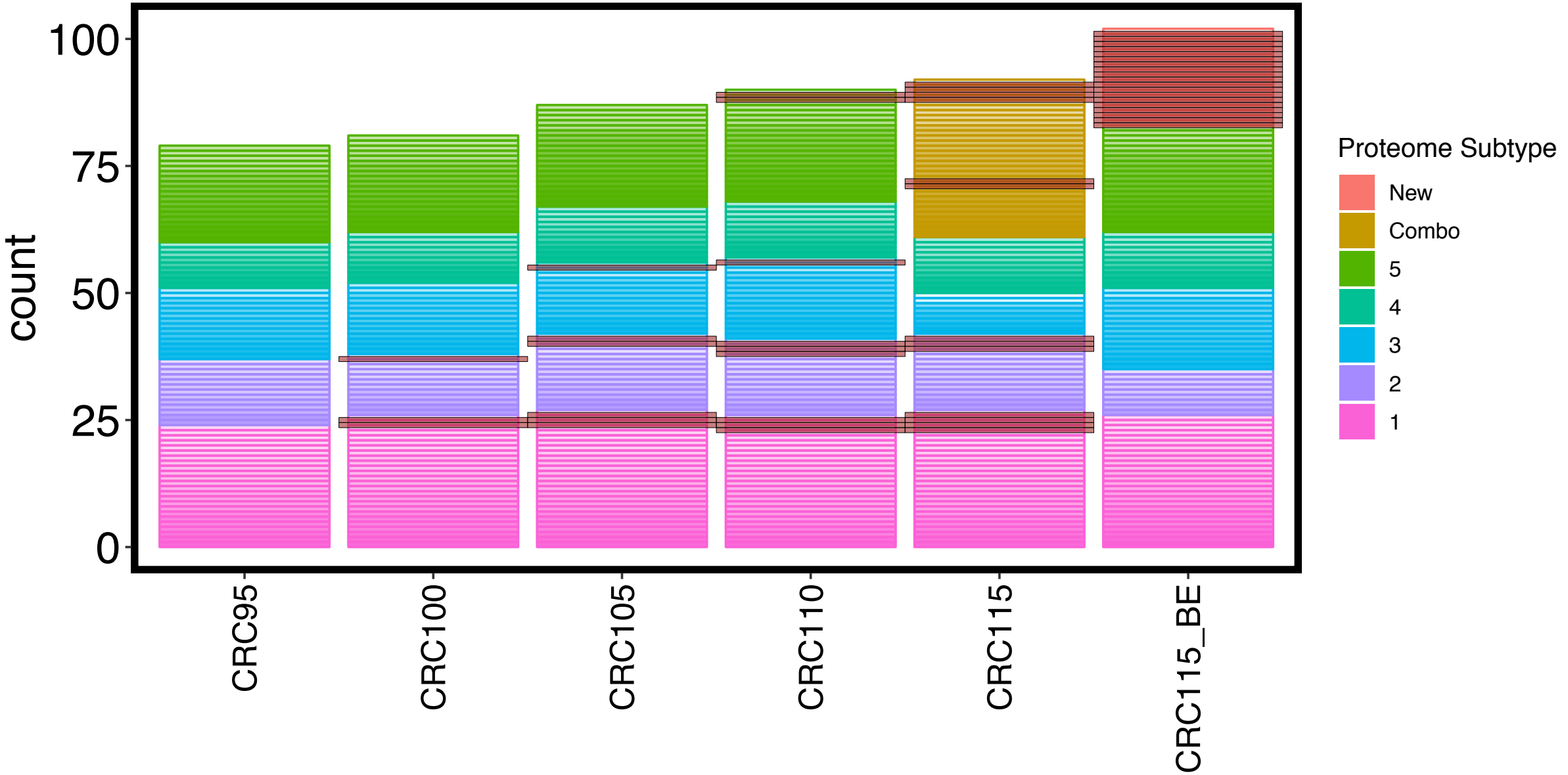
Different Parameters



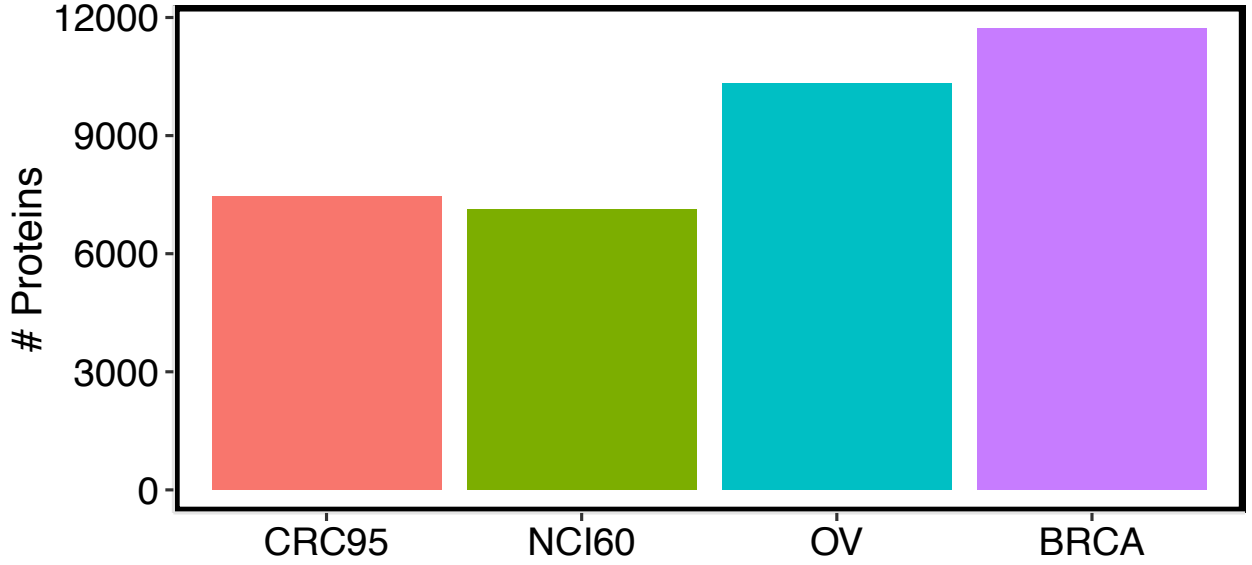
NCBI vs UCSC Reference DB

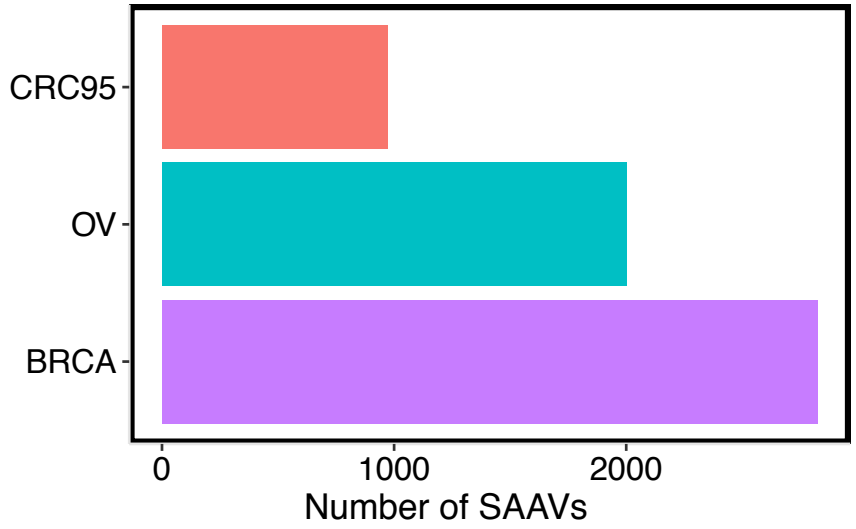


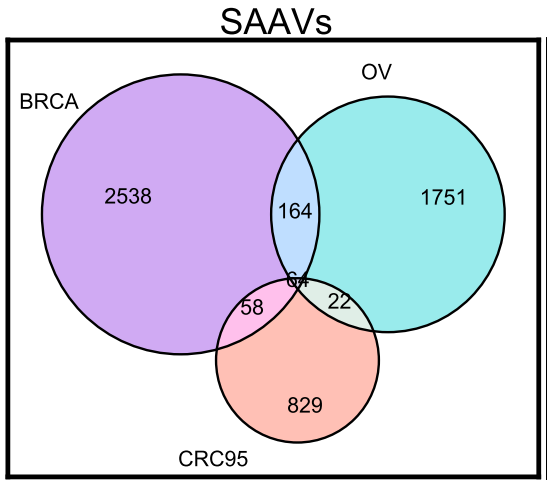
Revising Hypotheses with Additional Data

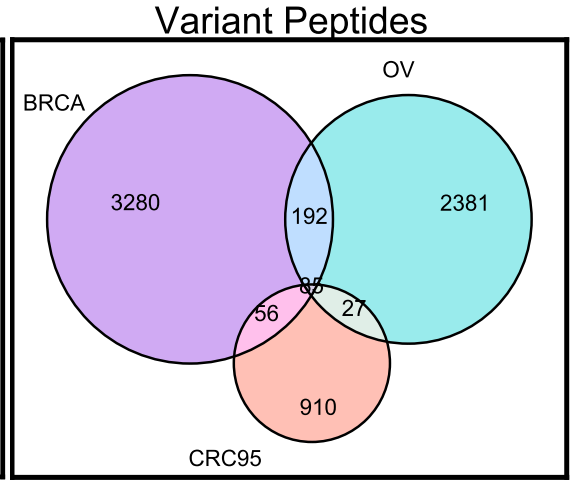


Analyzing and Comparing Different Datasets









Benchmarking for DREAM Challenges

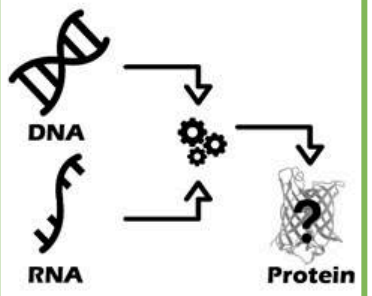
DREAM CHALLENGES

NCI-CPTAC DREAM Proteogenomics Challenge

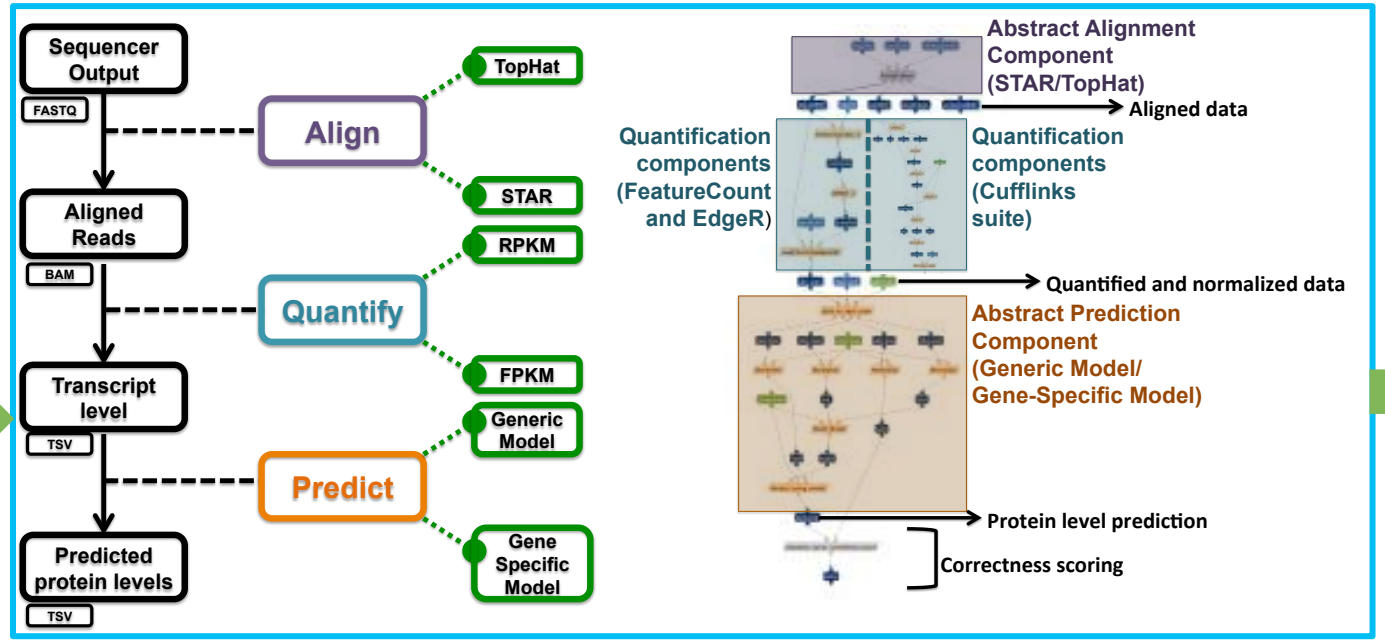
(Closed)

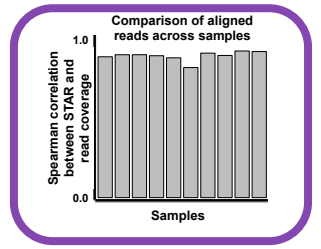
This challenge will use public and novel proteogenomic data generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) to benchmark an understanding of the interfaces between different layers of information in a population of cancer cells.

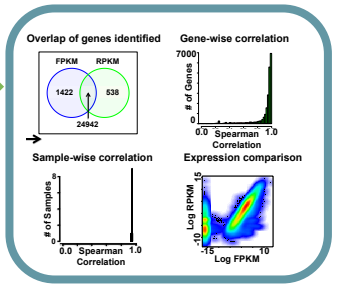
Predict protein abundances from mRNA and genetic data

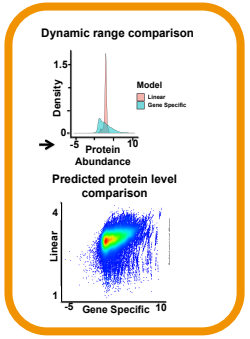


CPTAC DATA SETS









Alignment	Quantification	Predictive Model	Correctness Score	Time Taken
STAR	FPKM	Linear	0.2161	~29 hrs
STAR	RPKM	Linear	0.2155	~20 hrs
STAR	FPKM	Gene-Specific	0.9064	~29 hrs
STAR	RPKM	Gene-Specific	**0.9124**	~20 hrs
TopHat	RPKM	Linear	0.2053	~103 hrs
TopHat	RPKM	Gene-Specific	0.9080	~103 hrs

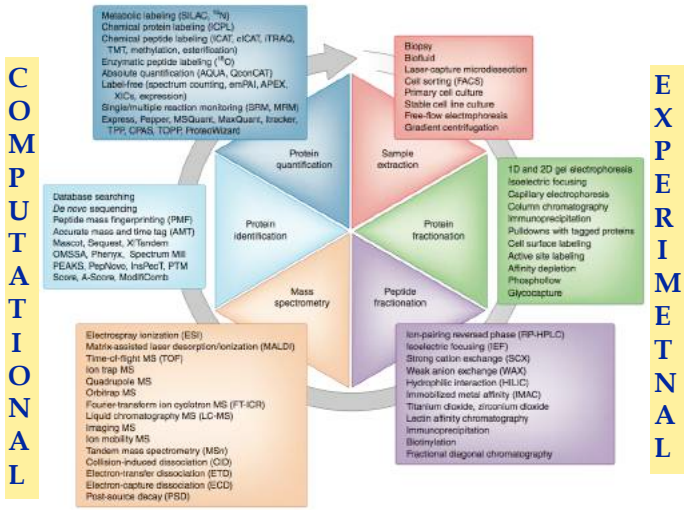
AI for Automating Discovery

Projection of Data Volumes in 2025

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

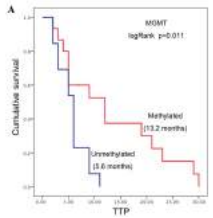
doi:10.1371/journal.pbio.1002195.t001

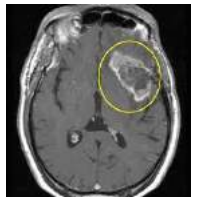
Siloed Expertise

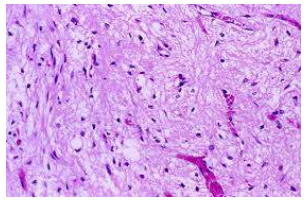


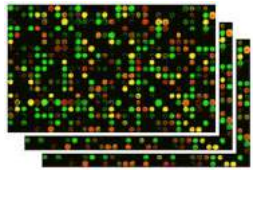
doi: 10.1038/nbt.1658

Analytic Complexity for Each Type of Data







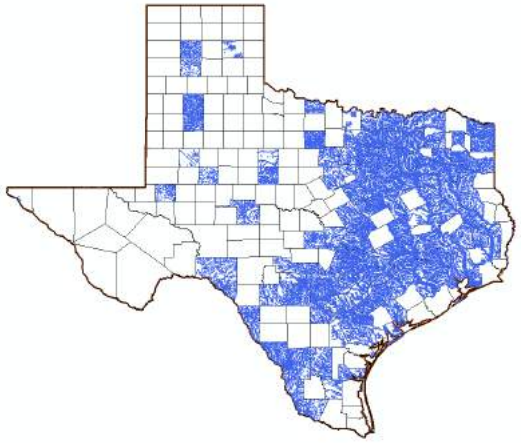


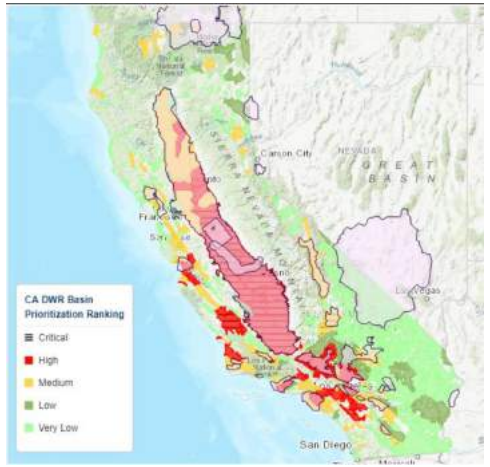
AI for Interdisciplinary Science Frontiers

MINT: Model INTegration

[Gil et al IEMS 2018; Garijo et al eScience 2019]

Collaboration with Daniel Garijo, Deborah Khider, Craig Knoblock, Ewa Deelman, Rafael Ferreira (USC/ISI), Vipin Kumar (UM), Scott Peckham (CU), Chris Duffy & Armen Kemanian (PSU), Kelly Cobourn (VT), Suzanne Pierce (UT)





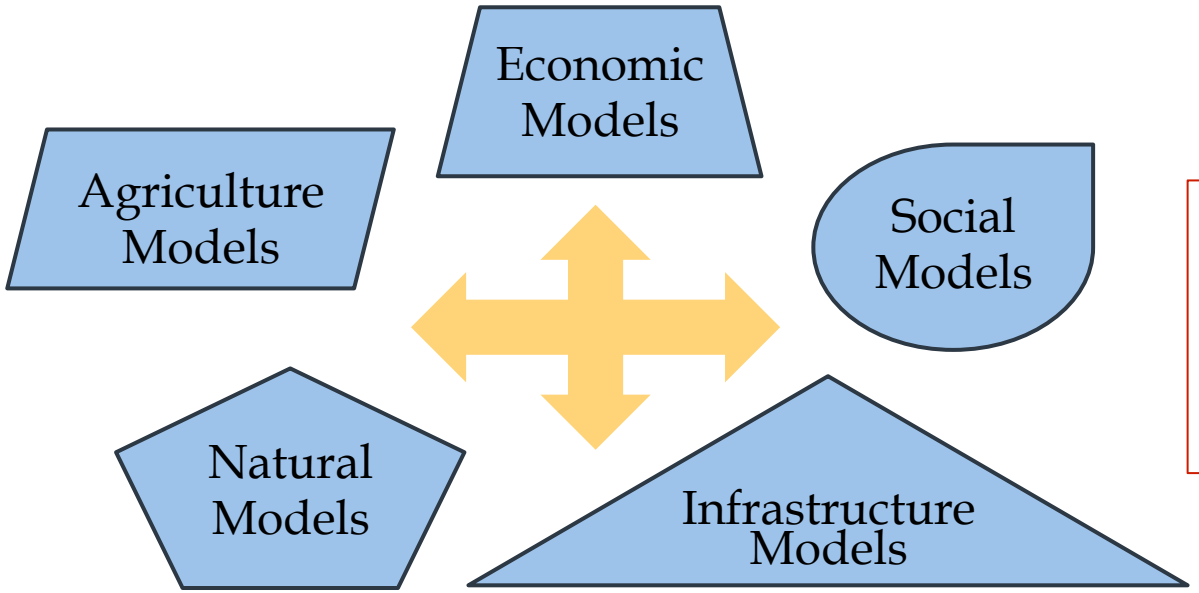
Food-Energy-Water Nexus

https://www.publicdomainpictures.net/en/view-image.php?image=256538&picture=agriculture-farming

https://commons.wikimedia.org/wiki/File:FEMA_-_43937_-_Flood_Damaged_Tennessee.jpg

https://www.pexels.com/photo/bird-birds-blue-sky-drought-1178291/

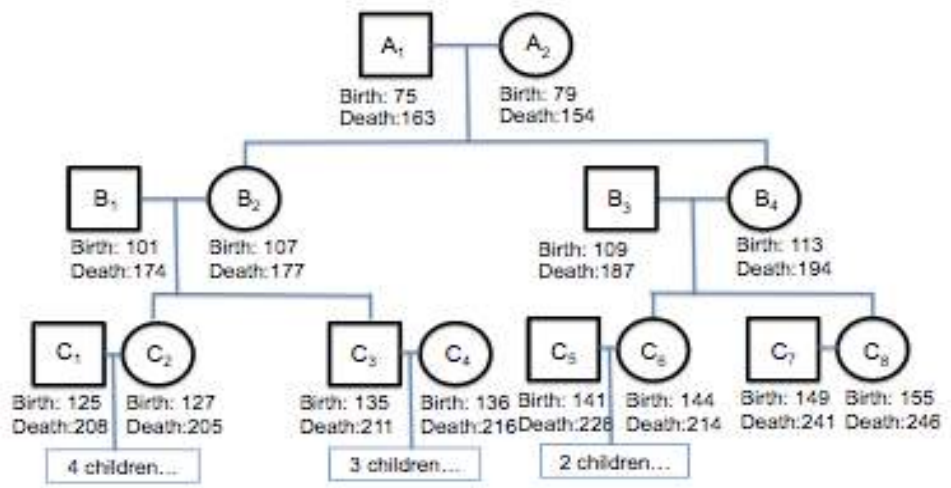
Integrated Modeling

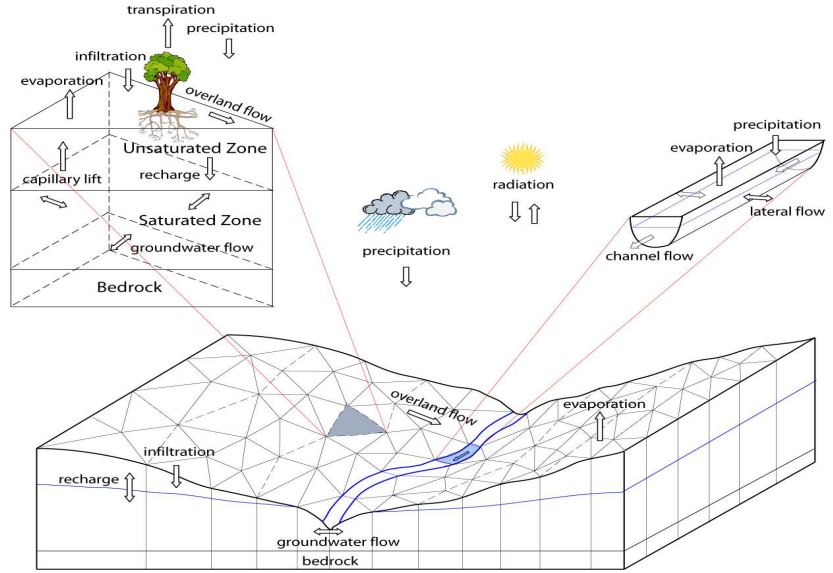


- Takes months/years
- Currently a craft
- Increased demand

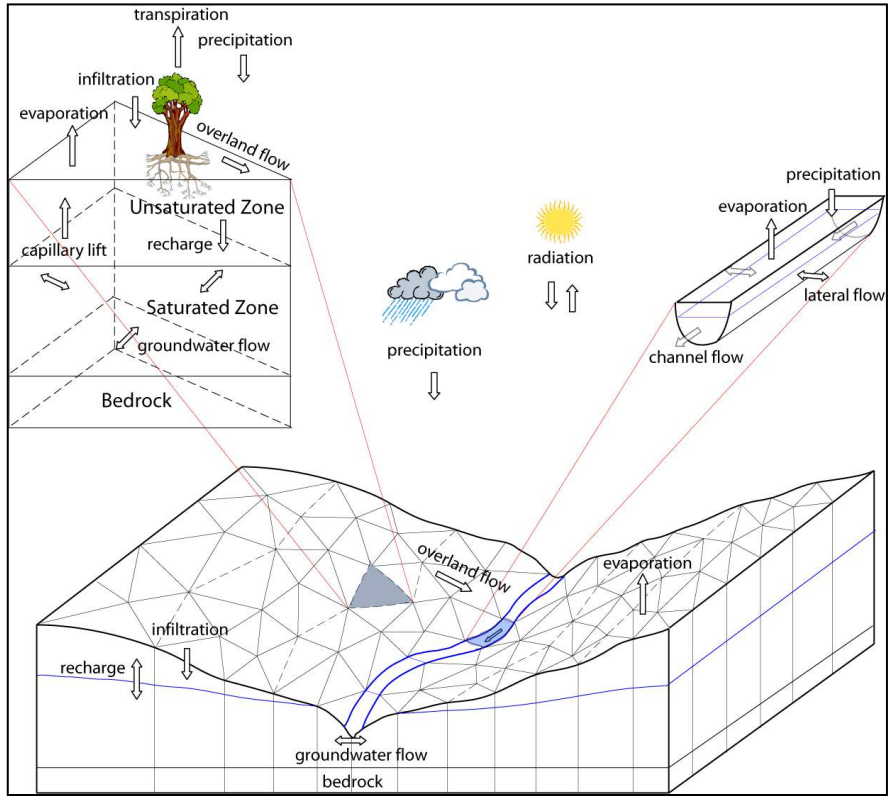
Mediation at many levels

- Modeling scope
- Model set up
- Variable mapping
- Data ingestion
- Spatial gridding





Supporting Compositionality of Scientific Knowledge



http://www.pihm.psu.edu/pihm_home.html

- Data formats
- Physical variables
- Constraints for use
- Adjustable parameters
- Interventions

Ontologies for Unambiguous Physical Variables

Work by Scott Peckham and Maria Stoica (CU)

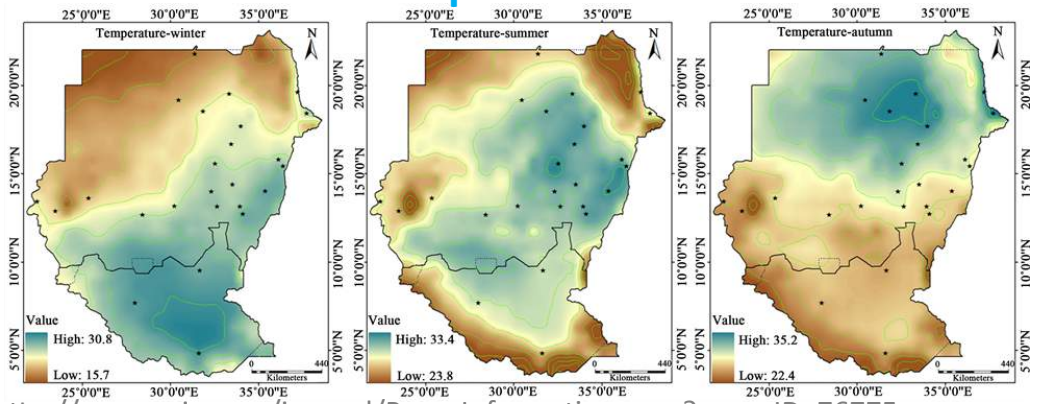
- Ontology of standard scientific names
  - Eg SSN: `watershed_outlet_water__volume_outflow_rate` is more precise than "streamflow" or "discharge"

cycles_soil Cycles soil description file. Soil files typically have a suffix of .soil, but any naming convention can be used as long as it matches the soil file na

Label	Long Name	Description	Standard Name	Units
RV	rock volume	Rock volume expressed as volume over total volume	soil_rock__volume_fraction	m3 m-3
DZ	soil layer thickness	Soil layer thickness	soil_layer__thickness	m
SLOPE	slope of the field	Average slope of field of interest	land_surface__slope	m m-1
LAYER	soil layer number	Soil layer number from 1 to an integer that is user defined (or database defined)		
SOM	soil organic matter	Soil organic matter per unit of non-rock soil expressed as percentage over total mass	soil_matter~organic__mass_fraction	kg kg-1 x-100
SILT	silt percentage	Silt mass per unit of soil mass (no rocks) and expressed as a percentage	soil~no-rock_silt__mass_fraction	%
BD	bulk density	Soil mass dry and wihtout rock divided by the sampled volume	soil~no-rock~dry__mass-per-volume_density	Mg m-3
CLAY	clay percentage	clay particle size fraction size fraction of each soil layer in %.	soil_clay_particle__volume_fraction	%

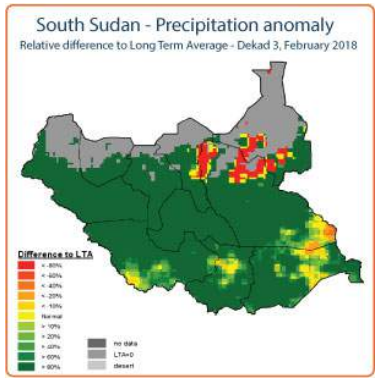
Spatial Datasets of Varying Quality

Temperature

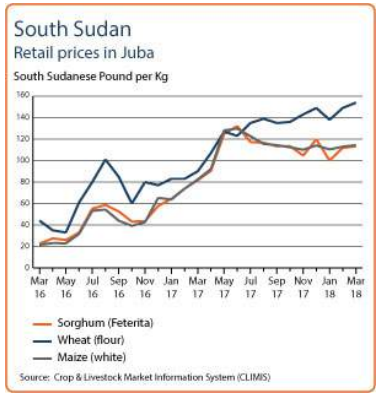


http://www.scirp.org/journal/PaperInformation.aspx?paperID=76775

Precipitation

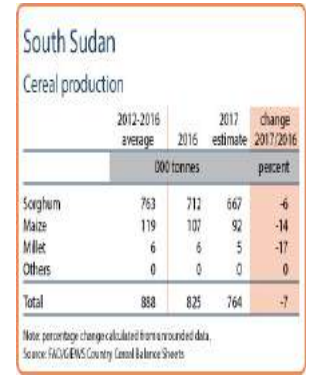


Crop prices

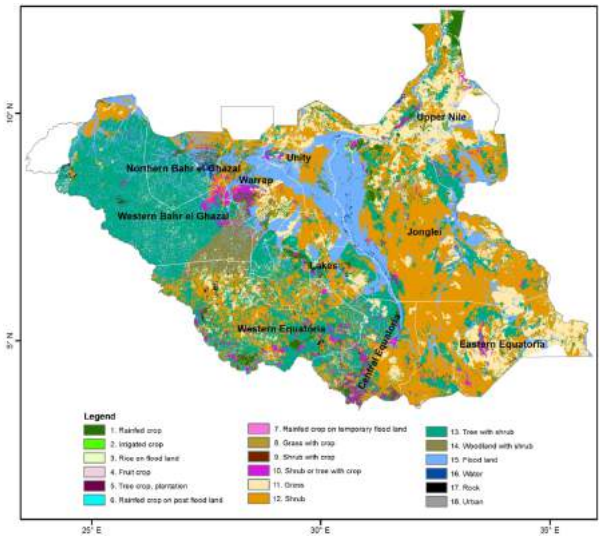


http://www.fao.org/giews/countrybrief/country.jsp?code=SSD

Production

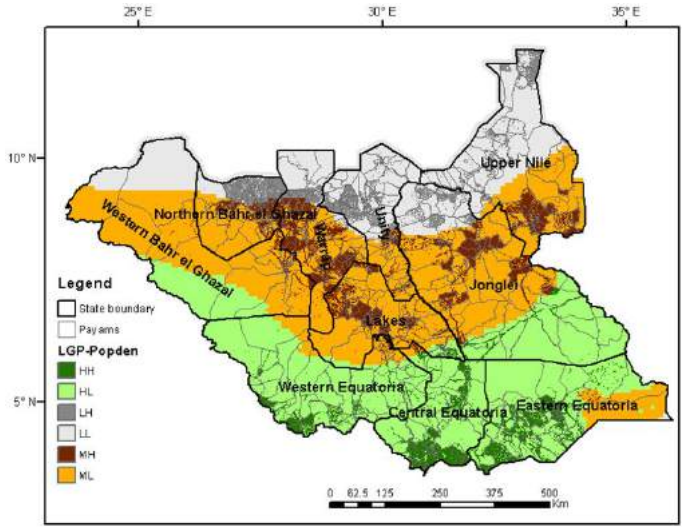


Land use



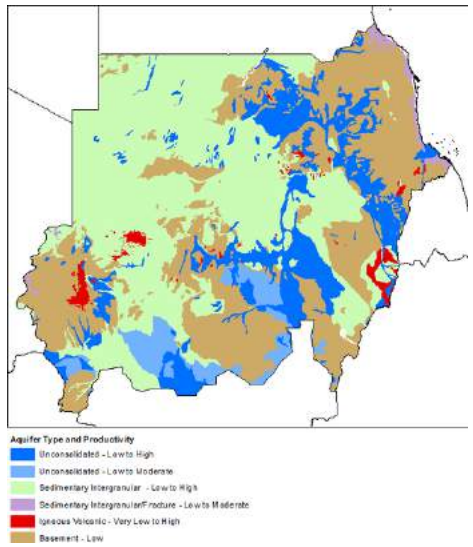
Diao, X. et al. *DOI: 10.5772/47938*

Agricultural potential



Diao et al DOI:10.5772/47938

Aquifer productivity and recharge



http://earthwise.bgs.ac.uk/index.php/Hydrogeology_of_Sudan

Automated Data Transformations

Work by Craig Knoblock, Yao-Yi Chiang Jay Pujara (USC)

Create a saturation file for each mesh cell

pg.gw file

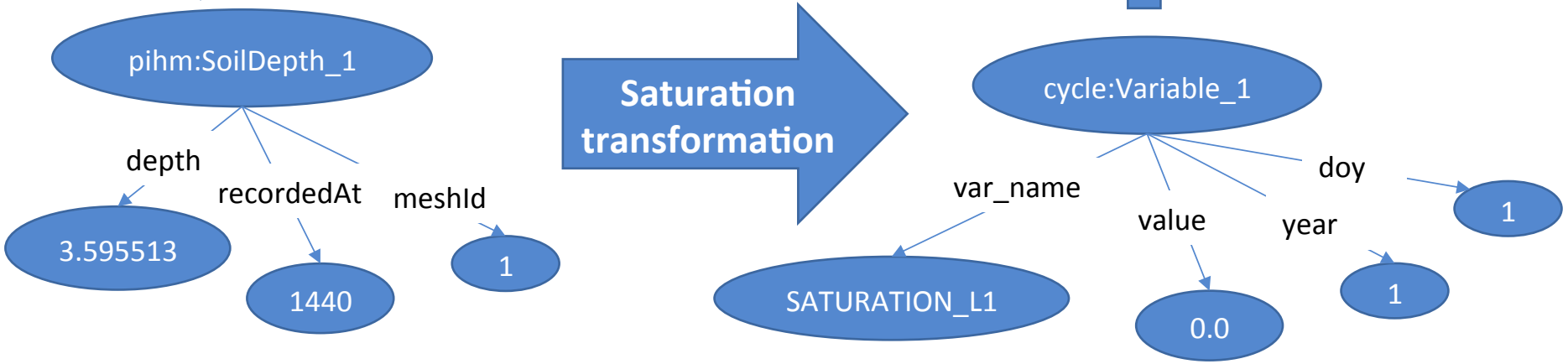
Time	Mesh 1	Mesh 2	...	Mesh n
1440	3.595513	6.534754	...	3.771523
2880	3.59[illegible]509	6.534728	...	3.771488
4320	3.59[illegible]505	6.534702	...	3.771453
...	...	...	...	...

Cycles.REINIT file

ROT	DOY	VARIABLE	VALUE
1	1	INFILTRATION	0.0
1	1	SATURATION_L1	0.0
1	1	SAT[illegible]TION_L2	0.012
...	...	...	...

Create a data graph for each cell

Write the data graph as a row



Hydrology model gives depth until ground water

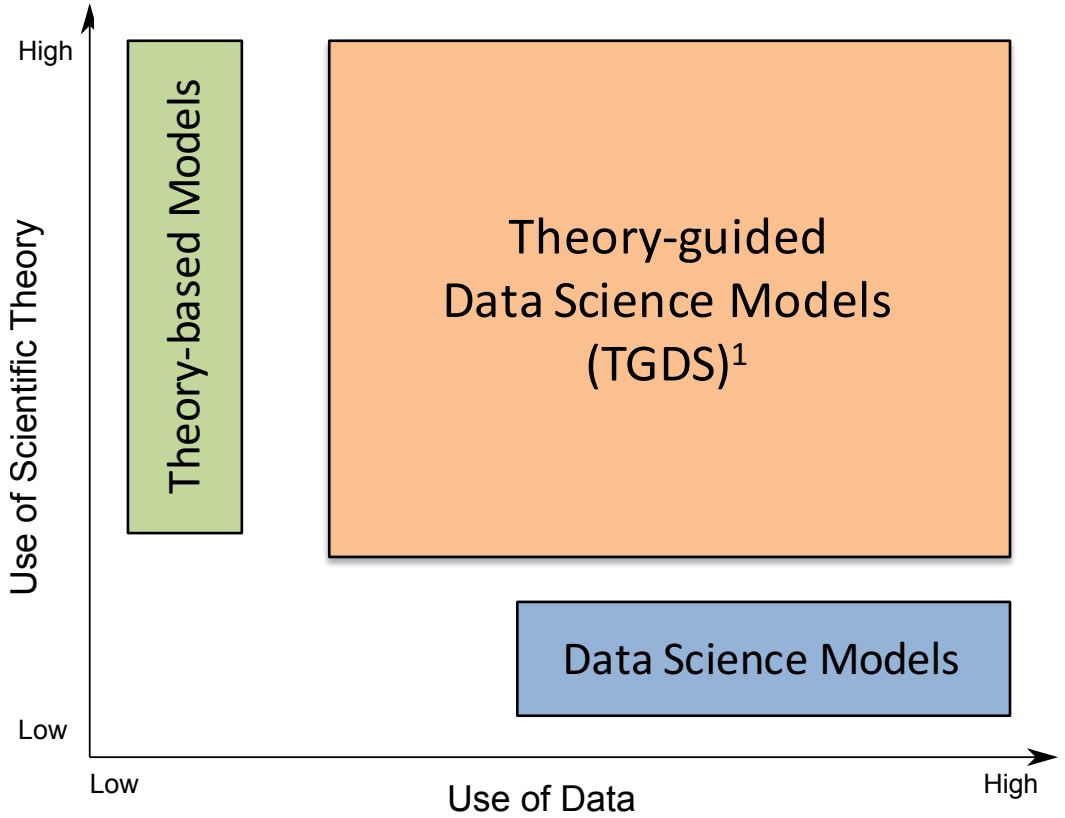
Agriculture model requires saturation

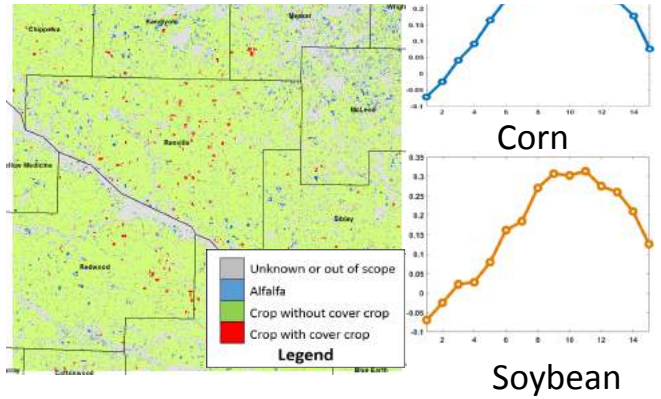
Creating Virtual Gauges When No Data Available

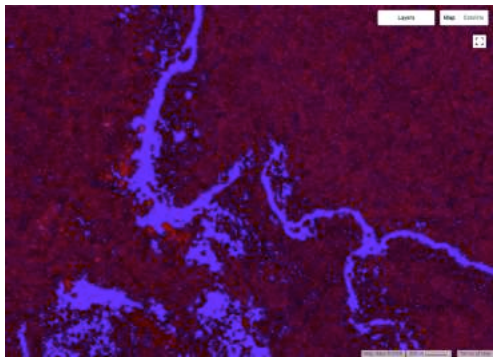
mint
model integration

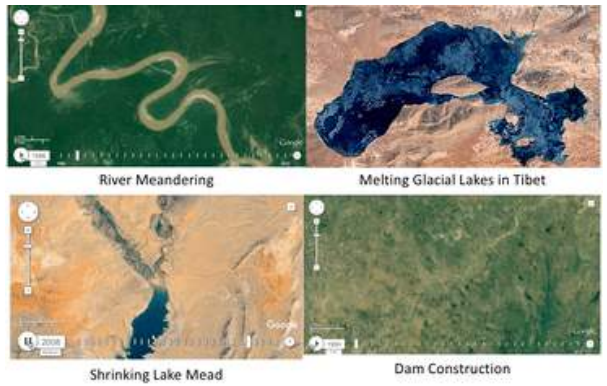
Work by Vipin Kumar and Ankuah Khandelwal (UM)

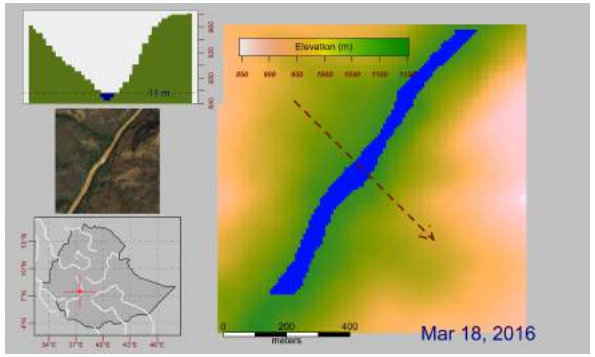
Theory-guided data science incorporates biophysical laws into machine learning from remote sensing data











Mapping Models to Interventions and Decisions

Work by Armen Kemanian and Yuning Shi (PSU)

Drivers and adjustable parameters

Weather

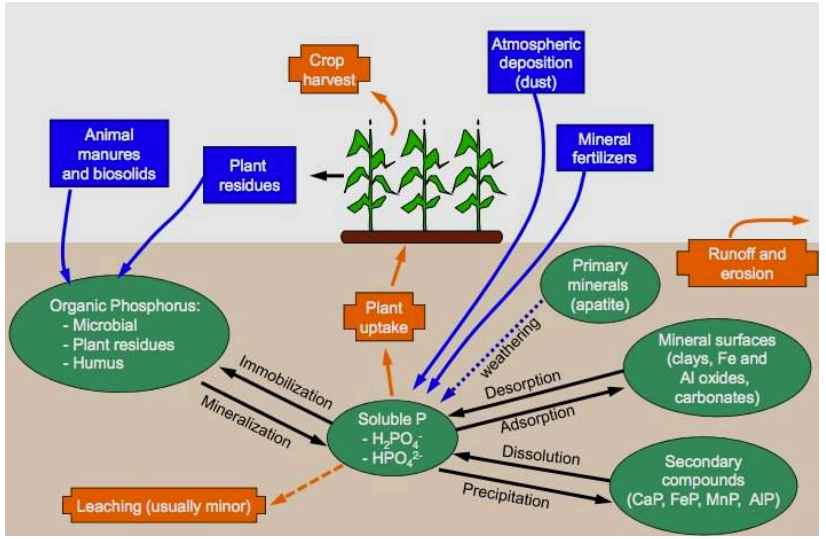
Crop types

Fertilizer

Planting dates

Weed factor

Land use



https://commons.wikimedia.org/wiki/File:Phosphorus_Cycle_copy.jpg

Model parameters

Nitrogen stress

Soils

Solar radiation

Response variables

Crop yield

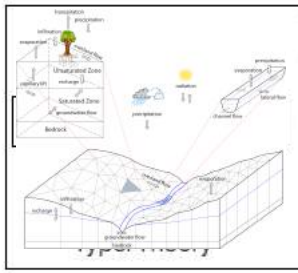
Rich Representations of Models

Work with Daniel Garijo, Deborah Khider, Varun Ratnakar, Maximiliano Osorio (USC)

Model Catalog

Search models

Search on: Full text



The Soil & Water Assessment Tool (SWAT)

The Soil & Water Assessment Tool (SWAT) is a small watershed to river basin-scale model used to simulate the quality and quantity of surface and ground water and predict the environmental impact of land use; land management practices and climate change (https://swat.tamu.edu/ 2019)

Keywords: Soil, watershed, surface water, ground water, en... More details

2 versions

Category: Agriculture
Type: Theory-

Cycles

Cycles simulates the productivity and the water-carbon and nitrogen balance of soil-crop systems subject to climate conditions and a large array of management constraints

Keywords: agriculture, cycles, crop growth, weather, soil, cr... More details

4 versions

Category: Economy
Type: Theory-

Economic aggregate crop supply response model (EACS)

The Aggregate crop supply response model (EACS) describes the aggregate crop supply response model for the country of South Sudan. This is a regional-scale aggregate model of agricultural supply for a specified set of crops (cassava; groundnuts; maize; sesame seed; and sorghum).

Keywords: economy, land use, crop production, fertilizer c... More details

Data formats

Input files

Name	Description	Value on setup	Format
cycles_weather	Cycles weather file	-	weather

Output files

Physical variables

Cycles configuration (version 0.9.4) exposing additional parameters such as weeds fraction and fertilizer rate
Authors: Rafael Silva

- **Time interval:** 1 day
- **Grid details:**
  - **Type:** PointBasedGrid
  - **Dimentions:** 0D
  - **Spatial resolution:** Point
- **Processes:** Respiration, Nitrogen mineralization and immobalization, Biomass growth, Humification, Transpiration, Nitrogen uptake, Precipitation, Solar radiation, Management, Nitrogen transport
- **Download:** cycles-0.9.4-alpha.zip?raw=true

Constraints

Name	Description	Value on setup
crop_name	Name of the crop to run the simulation for	Maize (default)
start_planting_day	Day of the year when the planting started The range is from 1 to 365	100 (default)
end_planting_day	Day of the year when the planting ended	149 (default)

Adjustable parameters

cycles_weather — Cycles weather file

Label	Long Name	Description	Standard Name
RHn	relative humidity minimum	relative humidity minimum	atmosphere_air_water-vapor__min_of_relative_s
Solar	solar radiation of the day	Shortwave incoming radiation on a horizontal surface at elevation of the location	land_surface-horizontal_radiation-incoming-sho
Wind	wind speed	Wind speed at 1.5 m	land_surface_wind__speed

MINT: From Models to Solutions

Work with Deborah Khider (USC); Suzanne Pierce and Lissa Pearson(UT); Chris Duffy and Lele Shu (PSU)

The Soil & Water Assessment Tool (SWAT)

The Soil & Water Assessment Tool (SWAT) is a small watershed to river basin-scale model used to simulate the quality and quantity of surface and ground water and predict the environmental impact of land use; land management practices and climate change (https://swat.tamu.edu/ 2019)

Keywords: Soil, watershed, surface water, ground water, en... More details

2 versions

Cycles

Cycles simulates the productivity and the water-carbon and nitrogen balance of soil-crop systems subject to climate conditions and a large array of management constraints

Category: Agriculture
Type: Theory-

Keywords: agriculture, cycles, crop growth, weather, soil, cr... More details

4 versions

Economic aggregate crop supply response model (EACS)

The Aggregate crop supply response model (EACS) describes the aggregate crop supply response model for the country of South Sudan. This is a regional-scale aggregate model of agricultural supply for a specified set of crops (cassava; groundnuts; maize; sesame seed; and sorghum).

Category: Economy
Type: Theory-

Keywords: economy, land use, crop production, fertilizer c... More details

Data formats

Input files

Name	Description	Value on setup	Format
cycles_weather	Cycles weather file	-	weather

Physical variables

Output files 9 Parameters 8

Cycles configuration (version 0.9.4) exposing additional parameters such as weeds fraction and fertilizer rate
Authors: Rafael Silva

- **Time interval:** 1 day
- **Grid details:**
  - **Type:** PointBasedGrid
  - **Dimentions:** 0D
  - **Spatial resolution:** Point
- **Processes:** Respiration, Nitrogen mineralization and immobalization, Biomass growth, Humification, Transpiration, Nitrogen uptake, Precipitation, Solar radiation, Management, Nitrogen transport
- **Download:** cycles-0.9.4-alpha.zip?raw=true

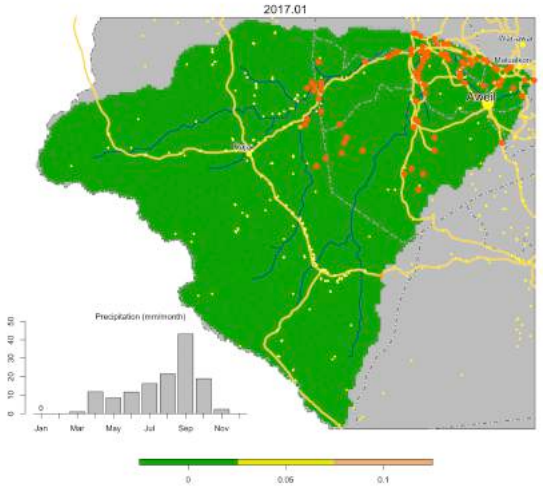
Constraints

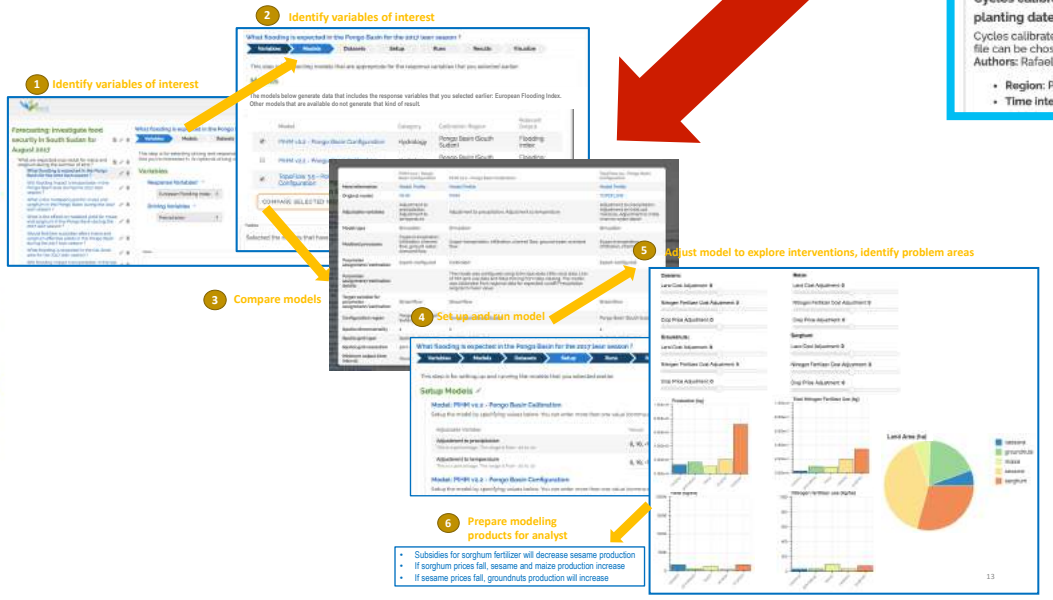
Name	Description	Value on setup
crop_name	Name of the crop to run the simulation for	Maize (default)
start_planting_day	Day of the year when the planting started The range is from 1 to 365	100 (default)
end_planting_day	Day of the year when the planting ended	149 (default)

Adjustable parameters

cycles_weather Cycles weather file

Label	Long Name	Description	Standard Name
RHn	relative humidity minimum	relative humidity minimum	atmosphere_air_water-vapor__min_of_relative_sa
Solar	solar radiation of the day	Shortwave incoming radiation on a horizontal surface at elevation of the location	land_surface-horizontal_radiation-incoming-shor
Wind	wind speed	Wind speed at 1.5 m	land_surface_wind__speed

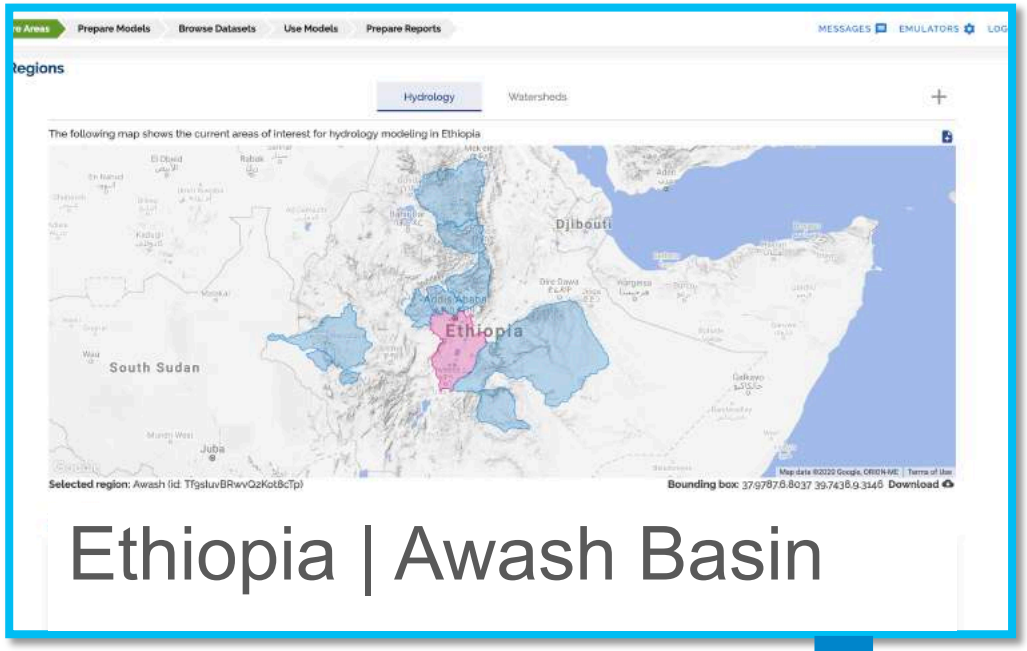




Model Portability: Data Scarce Regions

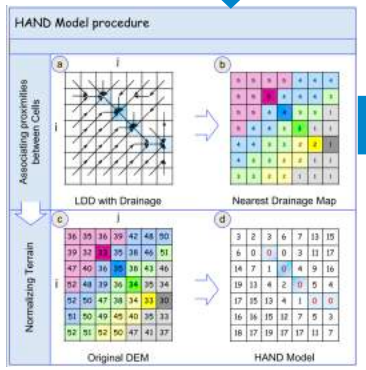
Height Above Natural Drainage Model (HAND)

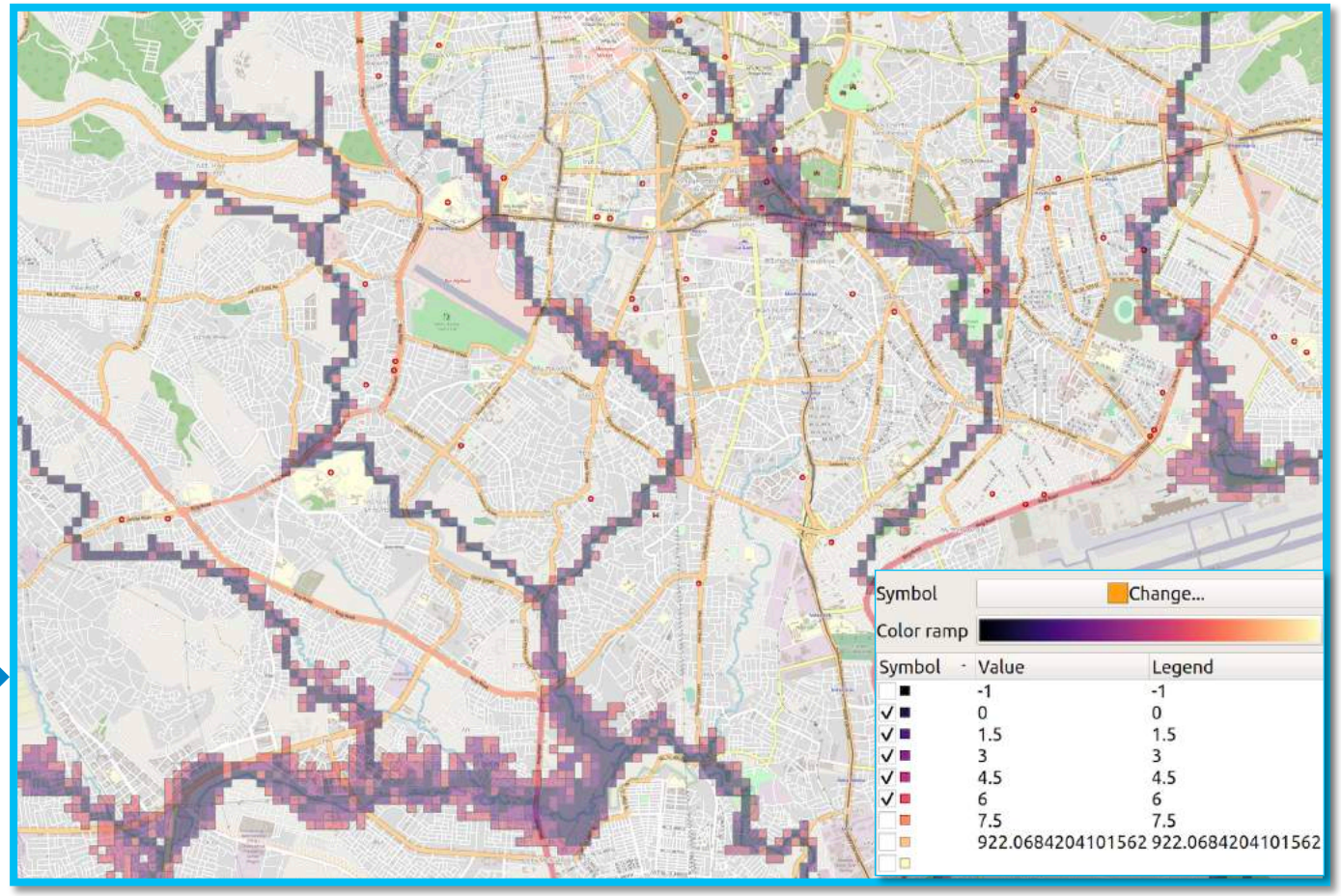
Work with Suzanne Pierce, Daniel Hardesty Lewis, David Arctur Paola Passalacgua (UT), David Tarboton (Utah State); Misty Porter, Mary Hill (KU)



Results provide first pass vulnerability for flood risk

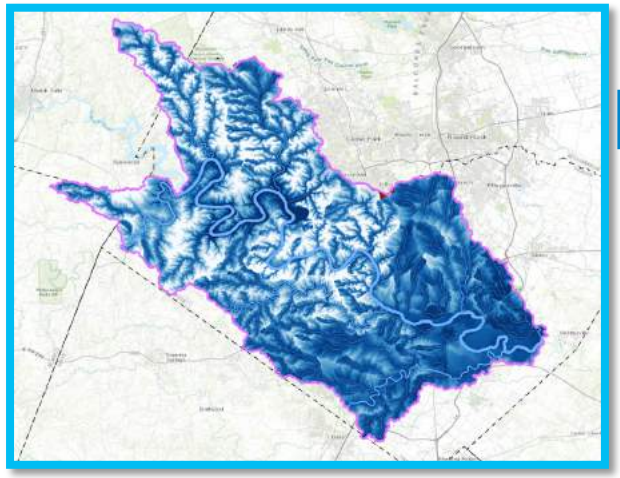
Pixels with colors on right of scale (brighter red) indicate higher change of inundation or flooding

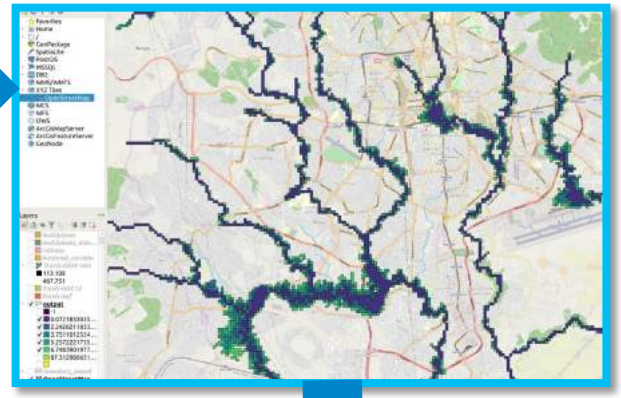




Model Portability: Data Rich Regions

Work with Suzanne Pierce, Daniel Hardesty Lewis, David Arctur and Paola Passalacgua (UT)

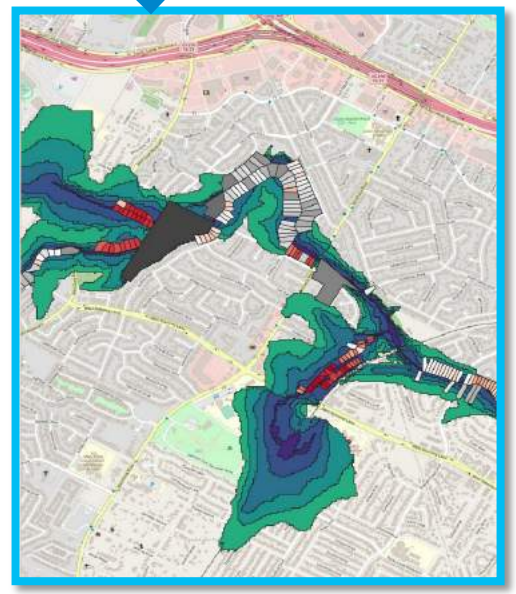




Combining

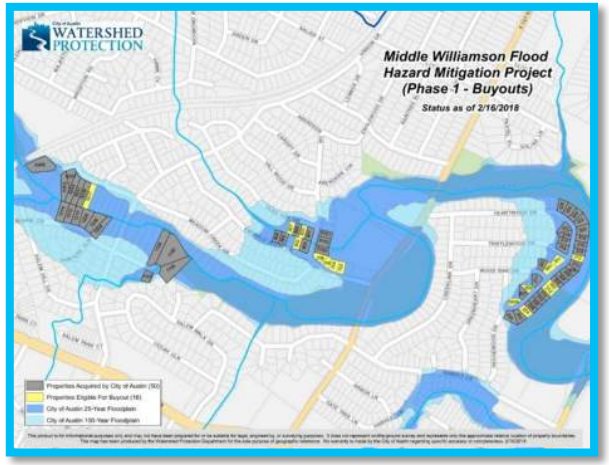
- Terrain (10 m)
- Population
- Urban Land Use

Travis County



Comparison: Williamson County

Manually Constructed Buyout Map

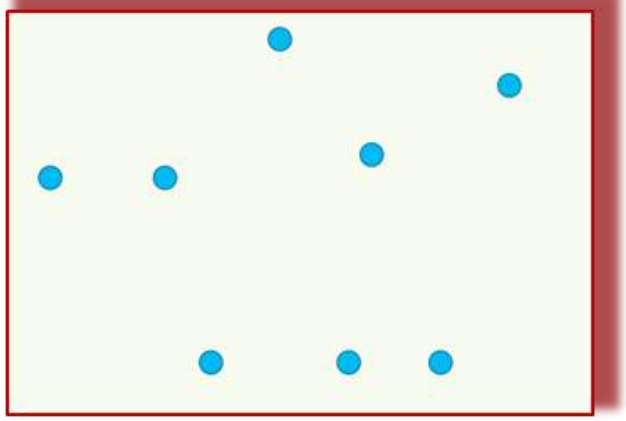


MINT Generated Vulnerability Map

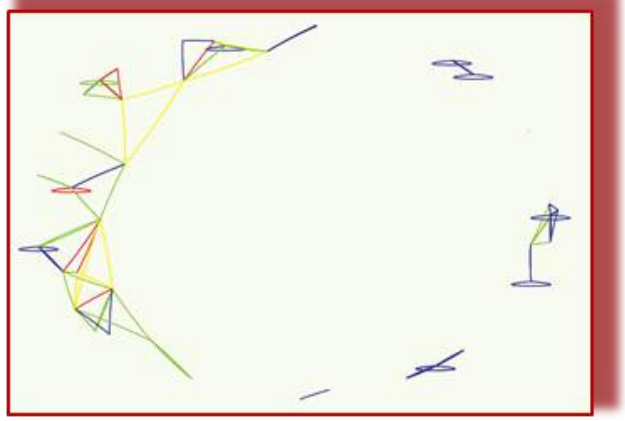
A Perspective on the Future

Tackling Complex Scientific Phenomena

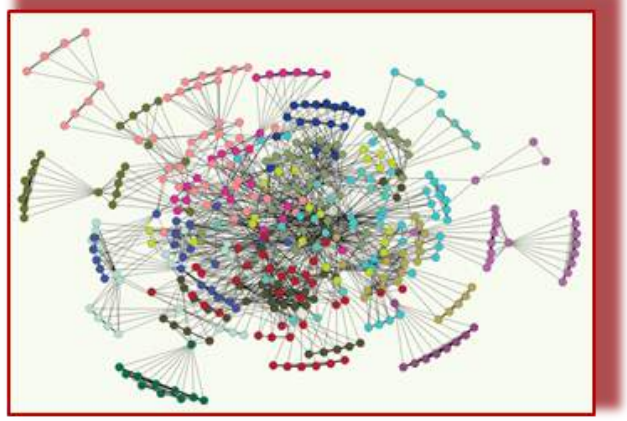
Single authorship



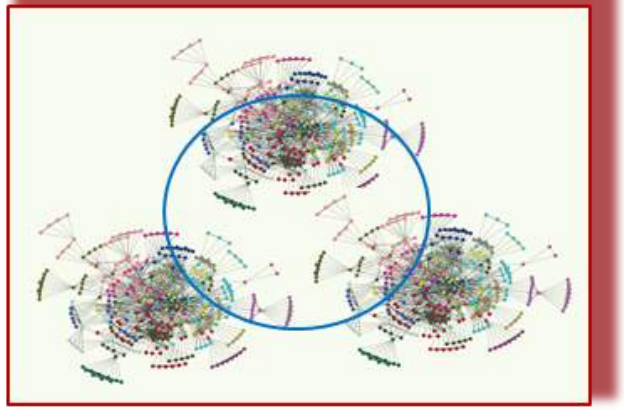
Co-authorship



Large number of co-authors



Community as author



Evolution of the scientific enterprise from [Barabasi, 2005] extended with the ATLAS Detector Project at the Large Hadron Collider [The ATLAS Collaboration, 2012].

Consider the Atlas Collaboration

Atlas collaboration

- Approximately 4,000 authors
- Worked in subgroups that coordinated with one another
- Collaboration lasted many years

Today, large scientific collaborations take significant time and effort and therefore are not very frequent

How can we change this?

The Importance of Process

Freestyle Chess Champion
Anson Williams

"The winner was revealed to be not a grandmaster with a state-of-the-art PC but a pair of amateur American chess players using three computers at the same time. Their skill at manipulating and "coaching" their computers to look very deeply into positions effectively counteracted the superior chess understanding of their grandmaster opponents and the greater computational power of other participants.

Weak human + machine + **better process** was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process."

– *Garry Kasparov, 2010*

http://www.thechessdrum.net/blog/2007/12/21/anson-williams-king-of-freestyle-chess/

Future AI Systems as Partners for Discovery

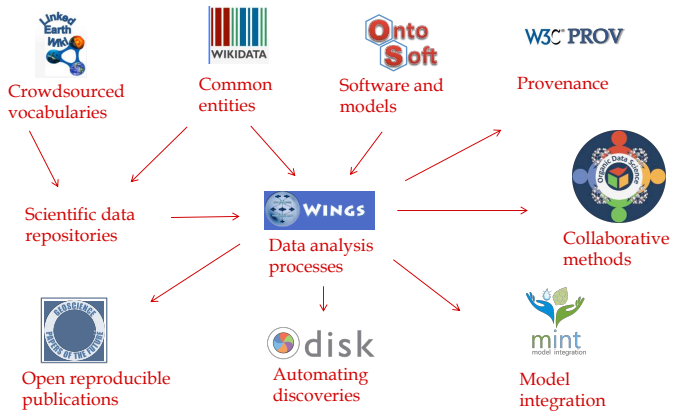
[Gil DSJ 2017]

Thoughtful AI: Principles for Partnership

Rationality	Behavior is governed by explicit knowledge structures
Context	Seek to understand the purpose and scope of tasks
Initiative	Proactively learn new knowledge relevant to their task
Networking	Access external sources of knowledge and capabilities
Articulation	Respond with persuasive justifications and arguments
Systems	Facilitate integration & collaboration with humans/systems
Ethics	Behavior that conveys scope and uncertainty

These are important research challenges for AI

Will AI Write the Scientific Papers of the Future?



Thoughtful AI: Principles for Partnership	
Rationality	Behavior is governed by explicit knowledge structures
Context	Seek to understand the purpose and scope of tasks
Initiative	Proactively learn new knowledge relevant to their task
Networking	Access external sources of knowledge and capabilities
Articulation	Respond with persuasive justifications and arguments
Systems	Facilitate integration & collaboration with humans/systems
Ethics	Behavior that conveys scope and uncertainty

YOUR IDEAS HERE!

Scientific Paper of the Future

Modern Paper

Text: Narrative of the method, some data is in tables, figures/plots, and the software used is mentioned

Data: Include data as supplementary materials and pointers to data repositories

Open Science

Sharing: Deposit data and software (and provenance/workflow) in publicly shared repositories

Open licenses: Open source licenses for data and software (and provenance/workflow)

Metadata: Structured descriptions of the characteristics of data and software (and provenance/workflow)

Reproducible Publication

Software: For data preparation, data analysis, and visualization

Provenance and methods: Workflow/scripts specifying dataflow, codes, configuration files, parameter settings, and runtime dependencies

Digital Scholarship

Persistent identifiers: For data, software, and authors (and provenance/workflow)

Citations: Citations for data and software (and provenance/workflow)

Formulating Grand Challenges

COMMUNICATIONS OF THE ACM

January 2019

REVIEW ARTICLES

Intelligent Systems for Geosciences: An Essential Research Agenda

By Yolanda Gil, Suzanne A. Pierce, Hassan Babaie, Arindam Banerjee, Kirk Borne, Gary Bust, Michelle Cheatham, Imme Ebert-phoff, Carla Gomes, Mary Hill, John Horel, Leslie Hsu, Jim Kinter, Craig Knoblock, David Krum, Vipin Kumar, Pierre Lermusiaux, Yan Liu, Chris North, Victor Pankratius, Shanan Peters, Beth Plale, Allen Pope, Sai Ravela, Juan Restrepo, Aaron Ridley, Hanan Samet, Shashi Shekhar

Many aspects of geosciences pose novel problems for intelligent systems research. Geoscience data is challenging because it tends to be uncertain, intermittent, sparse, multiresolution, and multi-scale. Geosciences processes and objects often have amorphous spatiotemporal boundaries. The lack of ground truth makes model evaluation, testing, and comparison difficult. Overcoming these challenges requires breakthroughs that would significantly transform intelligent systems, while greatly benefitting the geosciences in turn. Although there have been significant and

AI magazine

Spring 2016

Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery

Hiroaki Kitano

AI reproducing articles → AI as research assistant → AI as co-author

The Next Two Decades

2025: AI can generate automatically new complex scientific analyses using open data

2025: AI detects when it is missing knowledge and can seek and read new scientific papers on target topics

2030: AI can generate and test sophisticated hypotheses about complex physical phenomena

2030: AI can reproduce the results in 80% of the articles in a scientific journal

2035: AI can describe a scientific experiment and discuss sophisticated aspects of it

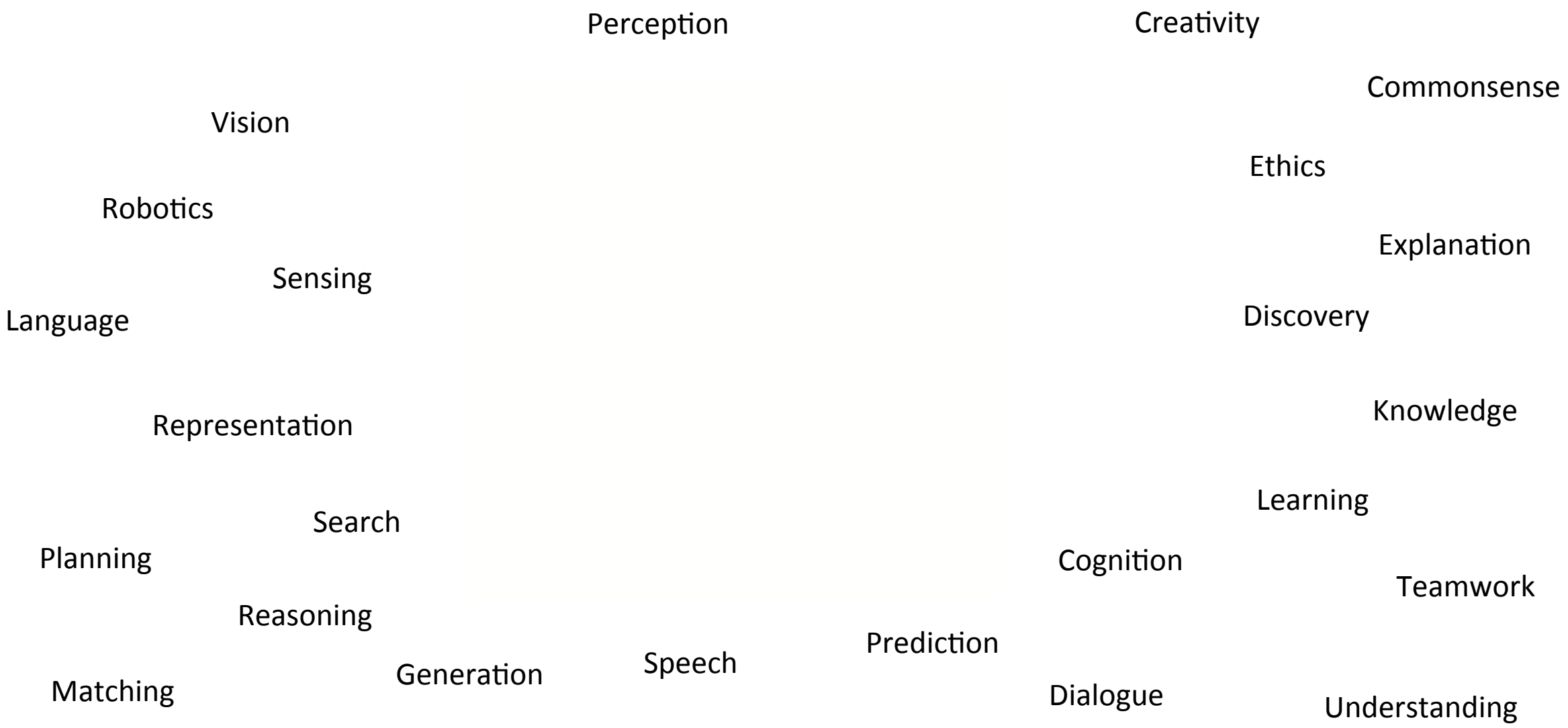
2035: AI can compare scientific experiments and papers and contrast their merits

2040: AI can teach advanced theories in some scientific domain effectively to students

2040: AI can formulate research questions and generate novel contributions in some scientific domain

AI to Address Major Future Challenges

Diversity and Breadth of Advances in AI



https://commons.wikimedia.org/w/index.php?curid=70268939

Will AI Write the Scientific Papers of the Future?

- The AI community has always been
  - Visionary
  - Broad
  - Inclusive
  - Interdisciplinary
  - Determined
- And dare I say
  - Successful

So… the answer may be yes?

https://www.flickr.com/photos/losalamosnatlab/22043600248
http://www.thechessdrum.net/blog/2007/12/21/anson-williams-king-of-freestyle-chess/

Will AI Write the Scientific Papers of the Future?

- AI researchers have been:
  - Visionary
  - Broad
  - Inclusive
  - Interdisciplinary
  - Determined
- And dare I say
  - Successful

So… the answer may be yes?

- Humans:
  - Not systematic
  - Errors
  - Biases
  - Poor reporting

So… the answer is definitely yes?

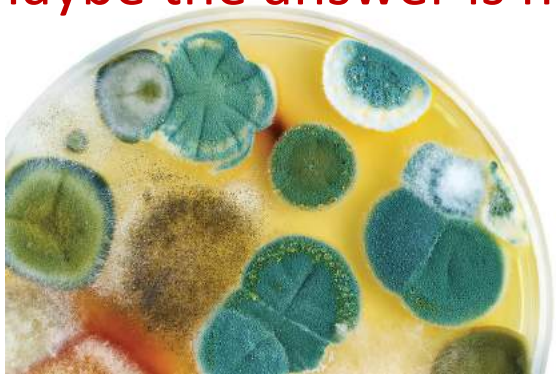
https://www.flickr.com/photos/losalamosnatlab/22043600248
http://www.thechessdrum.net/blog/2007/12/21/anson-williams-king-of-freestyle-chess/

Will AI Write the Scientific Papers of the Future?

- The AI community has always been
  - Visionary
  - Broad
  - Inclusive
  - Interdisciplinary
  - Determined
- And dare I say
  - Successful

So… the answer may be yes?

Maybe the answer is no:



Pennicillin discovery resulted from human error…

- Not systematic
- Errors
- Biases
- Poor reporting

So… the answer is definitely yes?

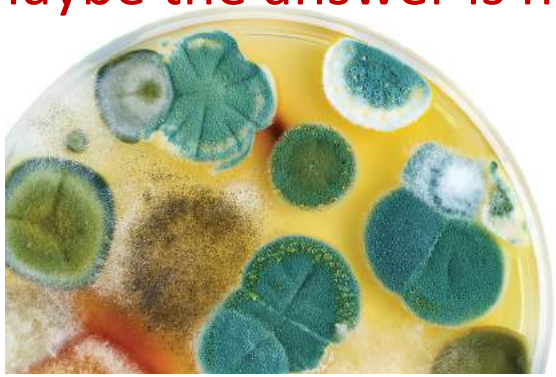
https://www.flickr.com/photos/losalamosnatlab/22043600248
http://www.thechessdrum.net/blog/2007/12/21/anson-williams-king-of-freestyle-chess/

Will AI Write the Scientific Papers of the Future?

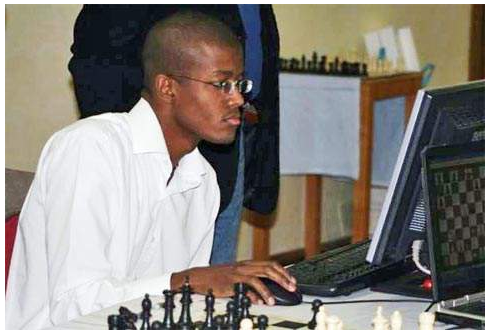
- The AI community has always been
  - Visionary
  - Broad
  - Inclusive
  - Interdisciplinary
  - Determined
- And dare I say
  - Successful

So… the answer may be yes?

Maybe the answer is no:



Pennicillin discovery resulted from human error…



…and humans make unique contributions…

- Not systematic
- Errors
- Biases
- Poor reporting

So… the answer is definitely yes?

https://www.flickr.com/photos/losalamosnatlab/22043600248
http://www.thechessdrum.net/blog/2007/12/21/anson-williams-king-of-freestyle-chess/

Thank you!











- Varun Ratnakar, Daniel Garijo, Deborah Khider, Maximiliano Osorio, Hernan Vargas (USC)
- *Workflows*: Jihie Kim, Ewa Deelman, Karan Vahi; Rafael Ferreira, Rajiv Mayani, Hyunjoon Jo, Yan Liu, Dave Kale (USC); Ralph Bergmann (U Trier); William Cheung (HKBU); Oscar Corcho (UPM); Pedro Gonzalez, Gonzalo Castro (UCM); Paul Groth (UA); Ricky Sethi (FSU); Carole Goble (UM); Chris Mattmann, Paul Ramirez, Dan Crichton, Rishi Verma (JPL); Natalia Villanueva (UTEP)
- *Linked Earth and Organic Data Science*: Julien Emile-Geay, Deborah Khider (USC); Nick McKay (NAU); Felix Michel and Matheus Hauder (TUM); Chris Duffy (PSU); Paul Hanson, Hilary Dugan, Craig Snortheim (U Wisconsin); Jordan Read (USGS); Neda Jahanshad (USC)
- *Biomedical workflows*: Phil Bourne, Sarah Kinnings (UCSD); Chris Mason (Cornell); Joel Saltz, Tahsin Kurk (Emory U.); Jill Mesirov, Michael Reich (Broad); Shannon McWeeney, Christina Zhang (OHSU); Parag Mallick, Ravali Adusumilli, Hunter Boyce (Stanford U.)
- *Geosciences workflows*: Paul Hanson (U Wisconsin), Tom Harmon & Sandra Villamizar (U Merced), Tom Jordan & Phil Maechlin (USC), Kim Olsen (SDSU); Suzanne Pierce (UT); Chris Duffy & Armen Kemanian (PSU); Scott Peckham & Maria Stoica (CU)
- *And many others!*