

High-Resolution, Functional Mapping of Voxel, Vector, and Meta Datasets within the **Human BioMolecular Atlas Program (HuBMAP)**



Katy Börner & Paul Macklin

Intelligent Systems Engineering
School of Informatics, Computing & Engineering
Indiana University
Bloomington, IN

Rebuilding a Kidney Scientific Meeting
Bolger Center, Washington, D.C.

December 14, 2018

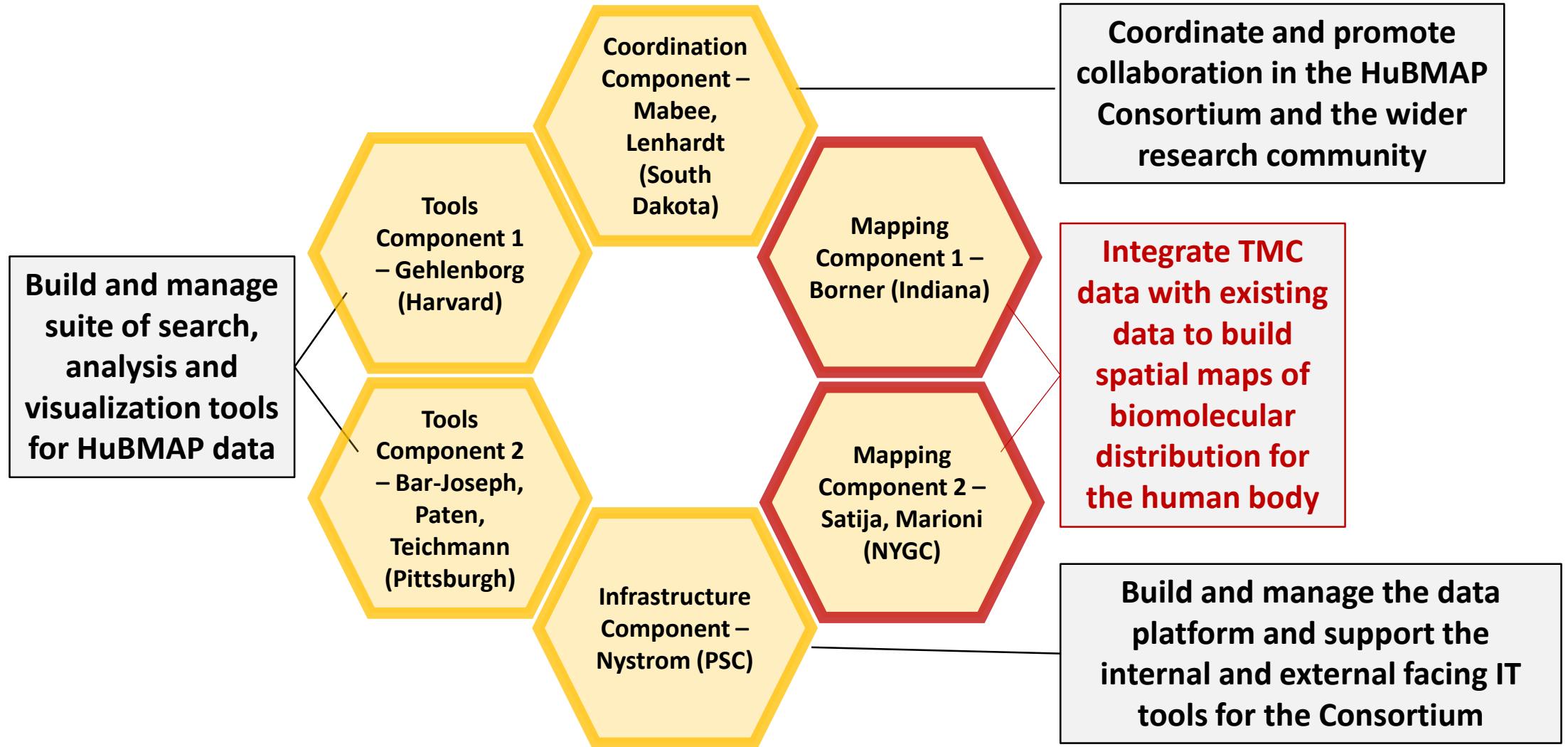


HuBMAP

The Human BioMolecular Atlas Program

<https://commonfund.nih.gov/HuBMAP>

HuBMAP: HIVE



HuBMAP: HIVE Mapping Components (MC)

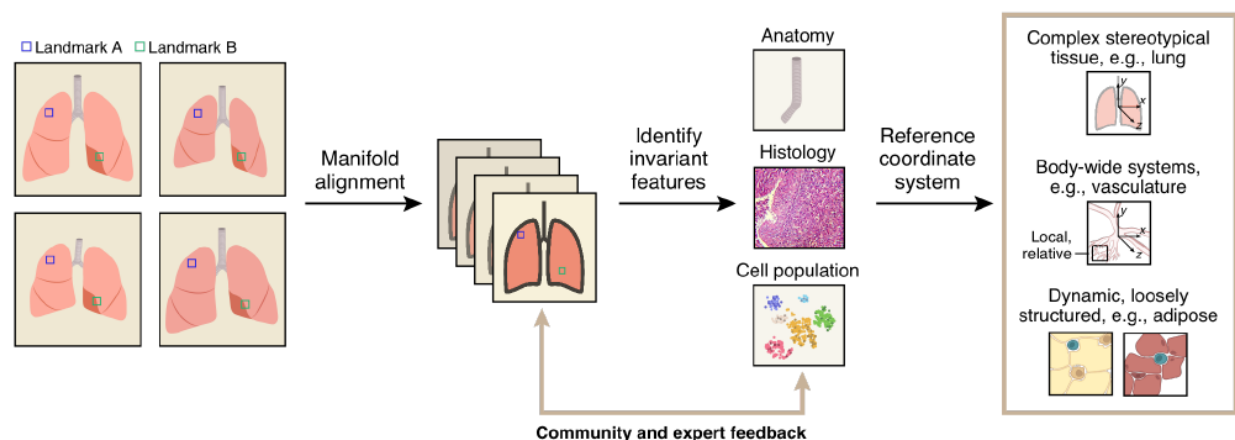
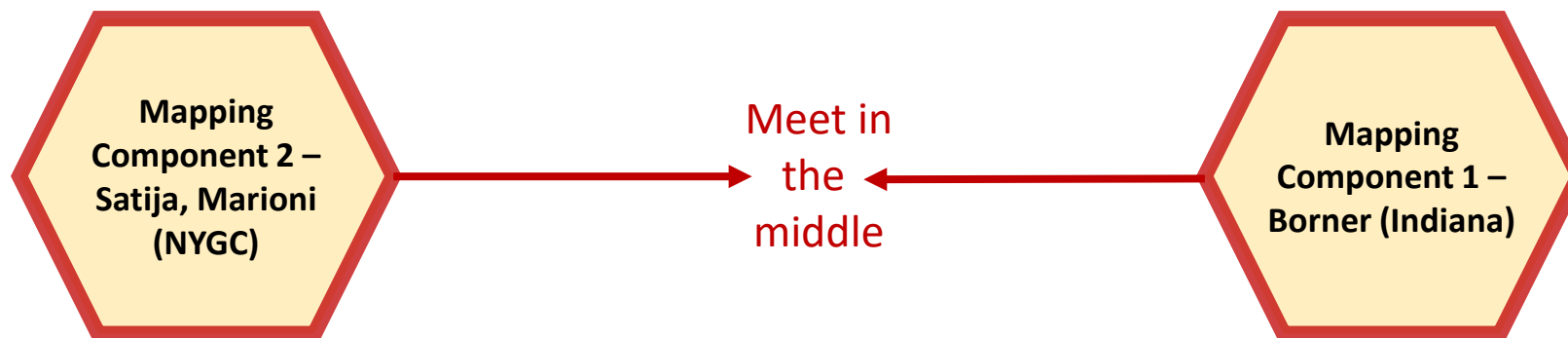


Figure 1: Overview of MC strategy to iteratively generate a reference coordinate system. We will work with TMC to annotate initial datasets with key features and ontologies, which will serve as 'landmarks' to align images across individuals. We will modify and adapt our strategy to diverse tissues, retaining a probabilistic framework that represents uncertainty due to measurement, and inter-individual variation.

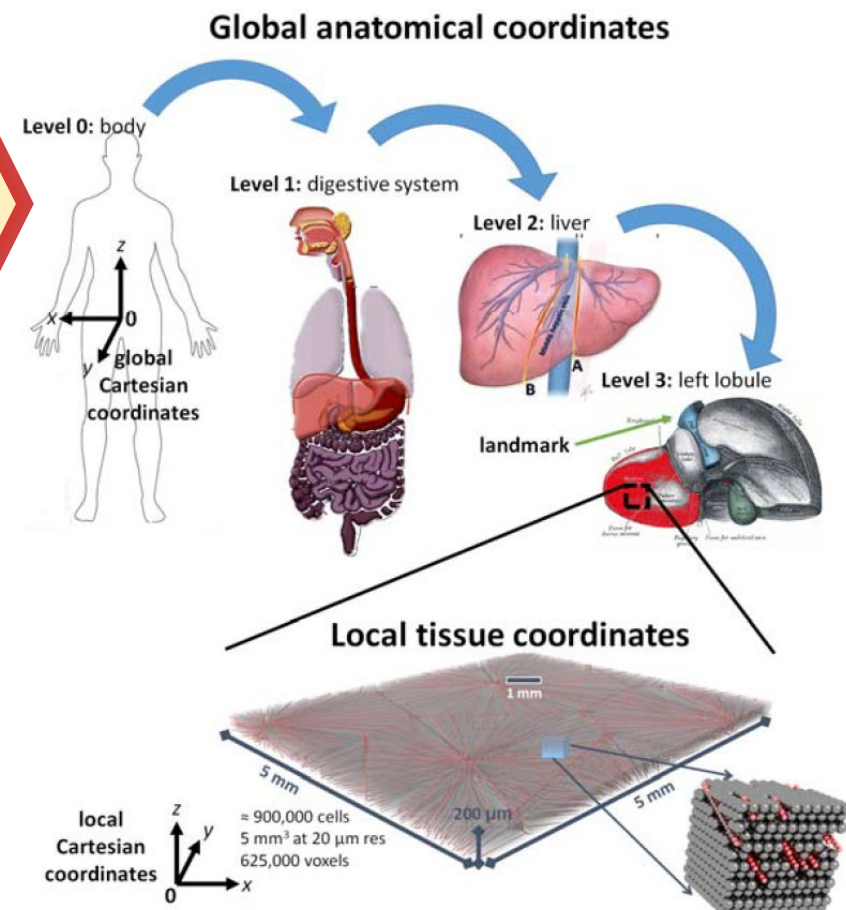





Fig. 5. CCF concept, navigating through the global anatomical coordinate system to insert a synthetic tissue sample (from PhysiCell⁴) into the left liver lobule with a local coordinate system.

HuBMAP: HIVE Mapping Component (MC)-Indiana

Griffin Weber  
Consultant
Harvard Medical School



Katy Börner     
HIVE MC-Indiana PI
Project Management
Indiana University

Paul Macklin   
HIVE MC-Indiana Co-PI
Data Standards
Indiana University

Lisel Record    
HIVE MC-Indiana PM
User Needs Analysis
Indiana University






Charlotte Smith 
Business Manager
Indiana University

Randy Heiland  
Ontologies
Indiana University

Sam Friedman  
Ontologies
Opto-Knowledge Systems, Inc.

Bruce Herr II 
User Interface Design
Indiana University

Ellen Quardokus  
CCF Literature Review
Indiana University

-  Project Management/Financial
-  Stakeholder Analysis, User Needs Analysis, User Studies
-  Common Coordinate Framework (CCF), Literature Review
-  User Interface Design
-  CCF Working Subgroup and CCF Workshop



Katy Börner
HIVE MC-Indiana PI
 Victor H. Yngve
 Distinguished Professor
 of Engineering and IS
 Indiana University (IU)
 Founding Director of
 CNS @katycns

Cyberinfrastructure for Network Science Center (CNS)

For 20+ years, CNS has developed tools, cyberinfrastructures and educational materials for diverse scientific communities.

Network Ties

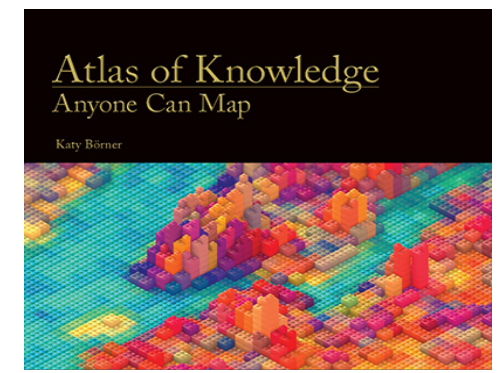
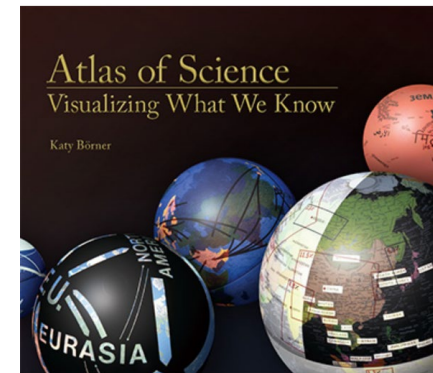
- Board of Trustees, Institute for Pure and Applied Mathematics (IPAM) since 2018.
- Humboldt Fellow, TU Dresden, Germany since 2017.
- Visiting Professor, Department of CS and Applied Cognitive Science, University of Duisburg-Essen, Germany since 2015.
- Visiting Professor, Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam, The Netherlands since 2012.

Research Focus

- Data Visualization Literacy
- Data Mining, Modeling, and Visualization
- Science of Science Studies
- Human Computer Interaction, Virtual Reality Interfaces
- Cognitive Science, AI
- MOOC Learning Analytics
- Cyberinfrastructure Design

Graphic Variable Types Versus Graphic Symbol Types

		Point	Line	Geometric Symbols
Spatial	x	quantitative		
	y	quantitative		
	z	quantitative		
Form	Size	quantitative	NA (Not Applicable)	
	Shape	qualitative	NA	
	Rotation	quantitative	NA	
	Curvature	quantitative	NA	
	Angle	quantitative	NA	
Retinal	Closure	quantitative	NA	
	Value	quantitative		
	Hue	qualitative		
Color	Saturation	quantitative		



Börner, Katy. 2015. *Atlas of Knowledge: Anyone Can Map*. Cambridge, MA: The MIT Press.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy, Andreas Bueckle, and Michael Ginda. Data Visualization Literacy: Definitions, Conceptual Frameworks, Exercises, and Assessments. Accepted.

NSF grant 1839167: TRIPODS+X: RES: Collaborative Research: Multi-Level Graph Representation for Exploring Big Data.



The Structure of Science

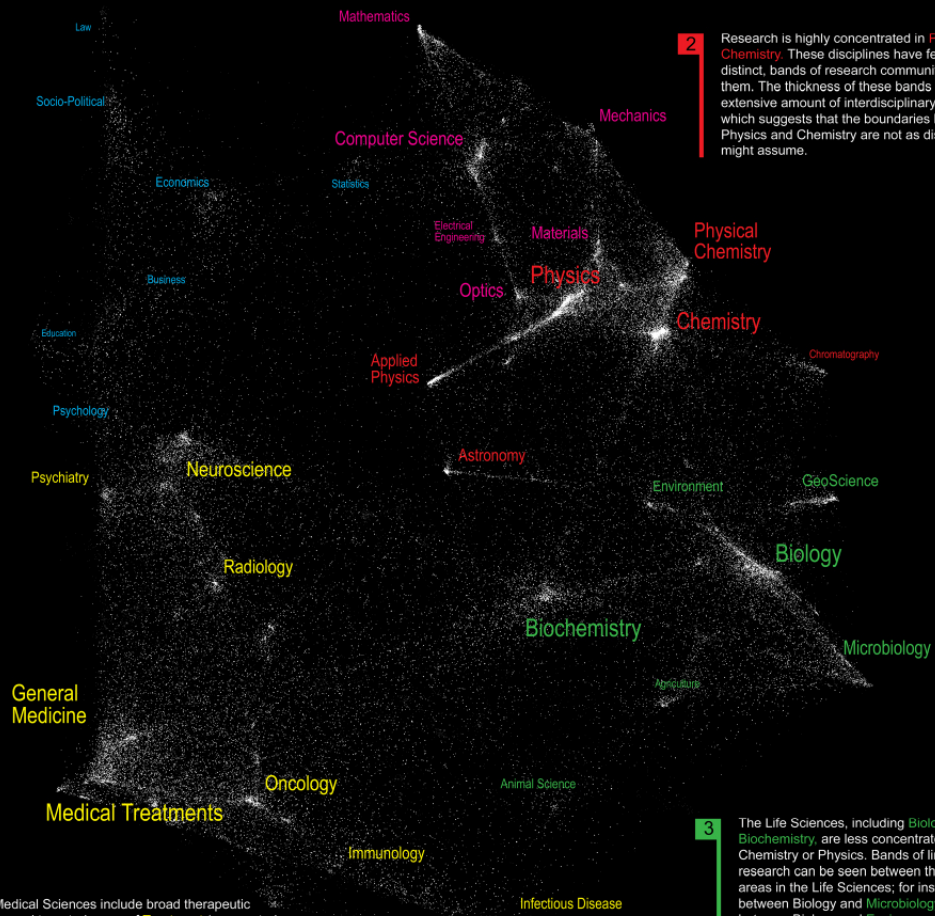
5 The Social Sciences are the smallest and most diffuse of all the sciences. **Psychology** serves as the link between Medical Sciences (Psychiatry) and the Social Sciences. **Statistics** serves as the link with Computer Science and Mathematics.

1 **Mathematics** is our starting point, the purest of all sciences. It lies at the outer edge of the map. **Computer Science**, **Electrical Engineering**, and **Optics** are applied sciences that draw upon knowledge in Mathematics and Physics. These three disciplines provide a good example of a linear progression from one pure science (Mathematics) to another (Physics) through multiple disciplines. Although applied, these disciplines are highly concentrated with distinct bands of research communities that link them. Bands indicate interdisciplinary research.

2 Research is highly concentrated in **Physics** and **Chemistry**. These disciplines have few, but very distinct, bands of research communities that link them. The thickness of these bands indicates an extensive amount of interdisciplinary research, which suggests that the boundaries between Physics and Chemistry are not as distinct as one might assume.

3 The Life Sciences, including **Biology** and **Biochemistry**, are less concentrated than Chemistry or Physics. Bands of linking research can be seen between the larger areas in the Life Sciences; for instance between Biology and Microbiology, and between Biology and Environmental Science. Biochemistry is very interesting in that it is a large discipline that has visible links to disciplines in many areas of the map, including Biology, Chemistry, Neuroscience, and General Medicine. It is perhaps the most interdisciplinary of the sciences.

4 The Medical Sciences include broad therapeutic studies and targeted areas of **Treatment** (e.g. central nervous system, cardiology, gastroenterology, etc.) Unlike Physics and Chemistry, the medical disciplines are more spread out, suggesting a more multi-disciplinary approach to research. The transition into Life Sciences (via Animal Science and Biochemistry) is gradual.



We are all familiar with traditional maps that show the relationships between countries, provinces, states, and cities. Similar relationships exist between the various disciplines and research topics in science. This allows us to map the structure of science.

One of the first maps of science was developed at the Institute for Scientific Information over 30 years ago. It identified 41 areas of science from the citation patterns in 17,000 scientific papers. That early map was intriguing, but it didn't cover enough of science to accurately define its structure.

Things are different today. We have enormous computing power and advanced visualization software that make mapping of the structure of science possible. This galaxy-like map of science (left) was generated at Sandia National Laboratories using an advanced graph layout routine (VxOrd) from the citation patterns in 800,000 scientific papers published in 2002. Each dot in the galaxy represents one of the 96,000 research communities active in science in 2002. A research community is a group of papers (9 on average) that are written on the same research topic in a given year. Over time, communities can be born, continue, split, merge, or die.

The map of science can be used as a tool for science strategy. This is the terrain in which organizations and institutions locate their scientific capabilities. Additional information about the scientific and economic impact of each research community allows policy makers to decide which areas to explore, exploit, abandon, or ignore.

We also envision the map as an educational tool. For children, the theoretical relationship between areas of science can be replaced with a concrete map showing how math, physics, chemistry, biology and social studies interact. For advanced students, areas of interest can be located and neighboring areas can be explored.

Nanotechnology

Most research communities in nanotechnology are concentrated in **Physics**, **Chemistry**, and **Materials Science**. However, many disciplines in the Life and Medical Sciences also have nanotechnology applications.

Proteomics

Research communities in proteomics are centered in **Biochemistry**. In addition, there is a heavy focus in the tools section of chemistry, such as **Chromatography**. The balance of the proteomics communities are widely dispersed among the Life and Medical Sciences.

Pharmacogenomics

Pharmacogenomics is a relatively new field with most of its activity in **Medicine**. It also has many communities in **Biochemistry** and two communities in the Social Sciences.

Science related Wikipedian ACTIVITY

This visualization explores the activity of science, math, and technology (SMT) related articles in the English-language Wikipedia (<http://en.wikipedia.org>). The central image shows 659,388 articles (circles). Overlaid is a 37 x 37 grid of relevant half-inch sized images.

Blue, green, and yellow circles represent the 3,599 math, 6,474 science, and 3,164 technology related articles respectively. The larger the size of a circle the higher the likelihood it is that type of article. The four corners show activity patterns of SMT articles.

Article Edit Activity
Articles are size coded based on how frequently they have been edited from Feb. 6, 2001 to April 6, 2007. More consideration is given to current and major edits. Larger circles have been edited more frequently than smaller circles.

2007 Major Edits
Articles are size coded based on how many major edits they received from January 1st, 2007 to April 6th, 2007. Larger circles have received more edits than smaller circles. The highest number of major edits was 2,627.

For the central image, each article is size coded based on the likelihood that it is math, science, or technology related.

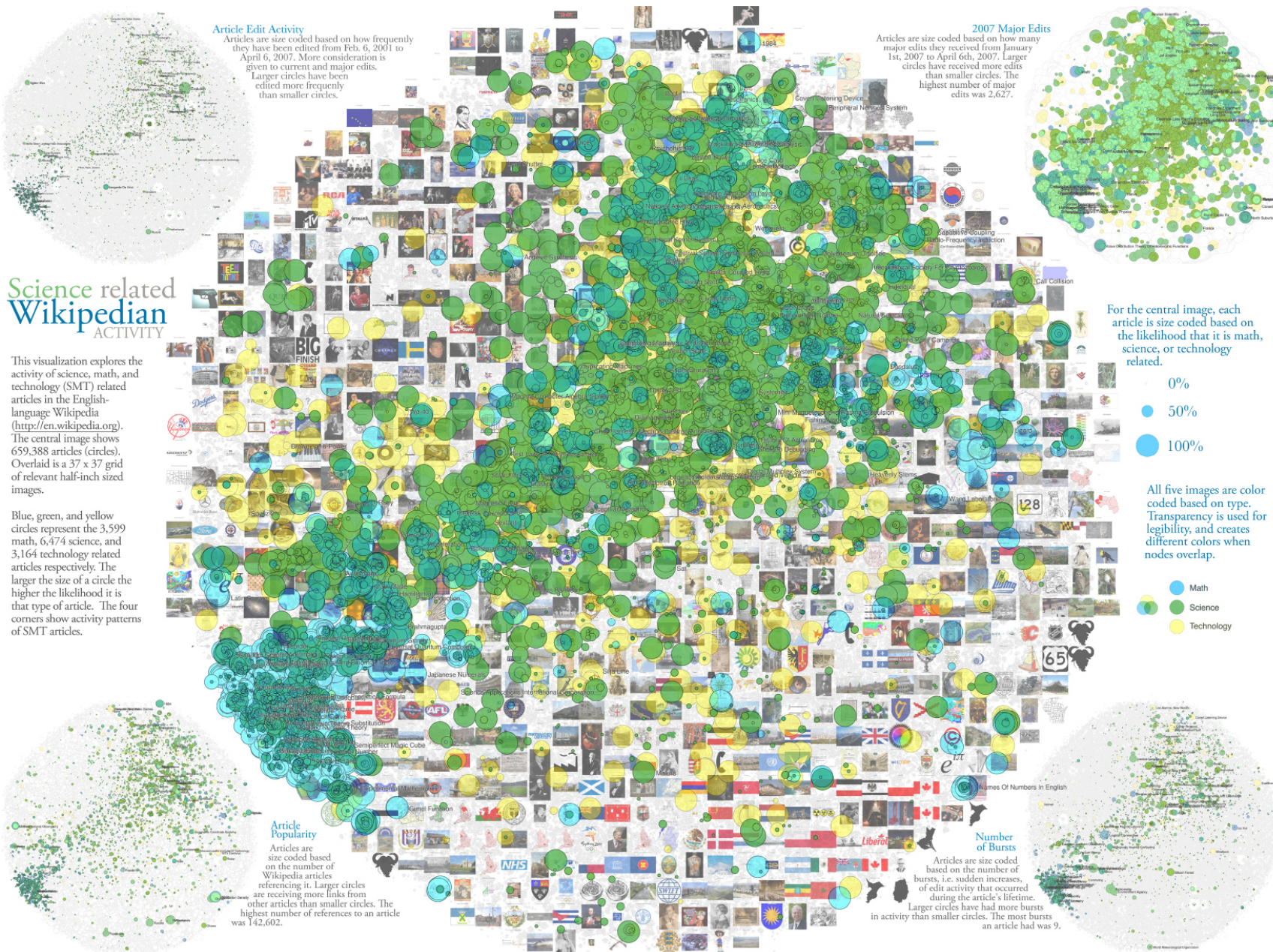
- 0%
- 50%
- 100%

All five images are color coded based on type. Transparency is used for legibility, and creates different colors when nodes overlap.

- Math
- Science
- Technology

Article Popularity
Articles are size coded based on the number of Wikipedia articles referencing it. Larger circles are receiving more links from other articles than smaller circles. The highest number of references to an article was 142,602.

Number of Bursts
Articles are size coded based on the number of bursts, i.e. sudden increases, of edit activity that occurred during the article's lifetime. Larger circles have had more bursts in activity than smaller circles. The most bursts an article had was 9.



Diseasome

The Human Disease Network

Explore online at <http://diseasome.eu>

Statistics

of Nodes: 516
 # of Edges: 1188
 Density: 0,0089
 Average Degree: 9,20
 Diameter: 15
 Average Shortest Path: 6,5

Top 5 Diseases

1. Deafness
2. Leukemia
3. Colon Cancer
4. Retinitis Pigmentosa
5. Diabetes Mellitus

Top 5 Genes

1. TP53
2. PAX6
3. FGFR2
4. RTN4
5. MSH2

Description

The map presents a network of 516 diseases linked by 1188 known disorder-gene associations, indicating the common genetic origin of many diseases.

HOW TO USE THE MAP

The map offers a rapid visual reference of the genetic links between disorders and a valuable global perspective for physicians, genetic researchers, and biomedical researchers alike. This new approach may lead to relatively efficient targeting of their affected genes, improve the understanding of the causes of disease, and the functions of particular genes.

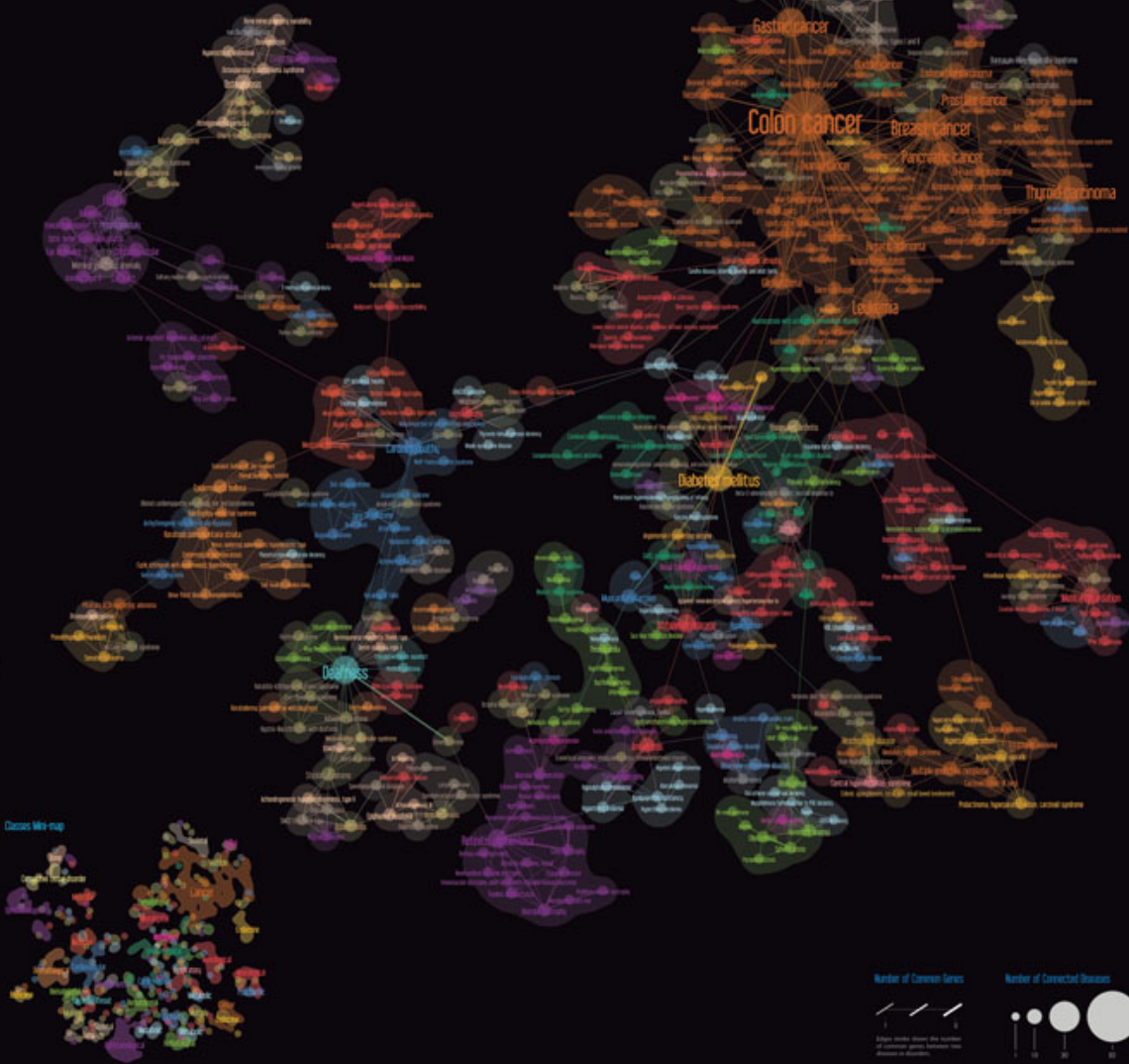
NETWORK VISUALIZATION TECHNIQUES APPLIED

The map was drawn using the force-directed layout algorithm ForceAtlas2 in Gephi. Node color corresponds to the disorder class to which the disease belongs, and the size is proportional to its node degree, the overall number of links. Link's width is proportional to the number of genes that are involved in both diseases and colored with the average color between source and target nodes. Isolated diseases are not shown and only the great component has been kept. The Clusters filter map labels most remarkable disorder clusters and shows largest visual clusters.

The Disorder Class Interactions graph below shows the interaction level between disorder classes, representing the number of shared genes, up to 80.

NOTES
 The network Disorder Network
 Bastin M, Heymann S, Valleron AJ, Collet B, Tardif R, Barthelemy A (2017)
 PLoS Med 14(10): e1002262

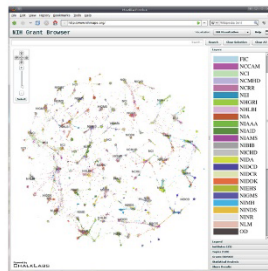
Disorder Class Interactions



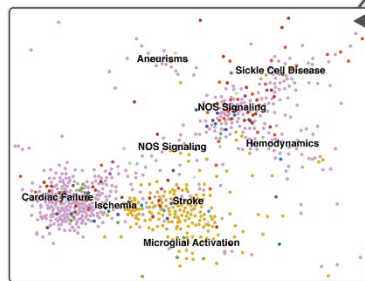
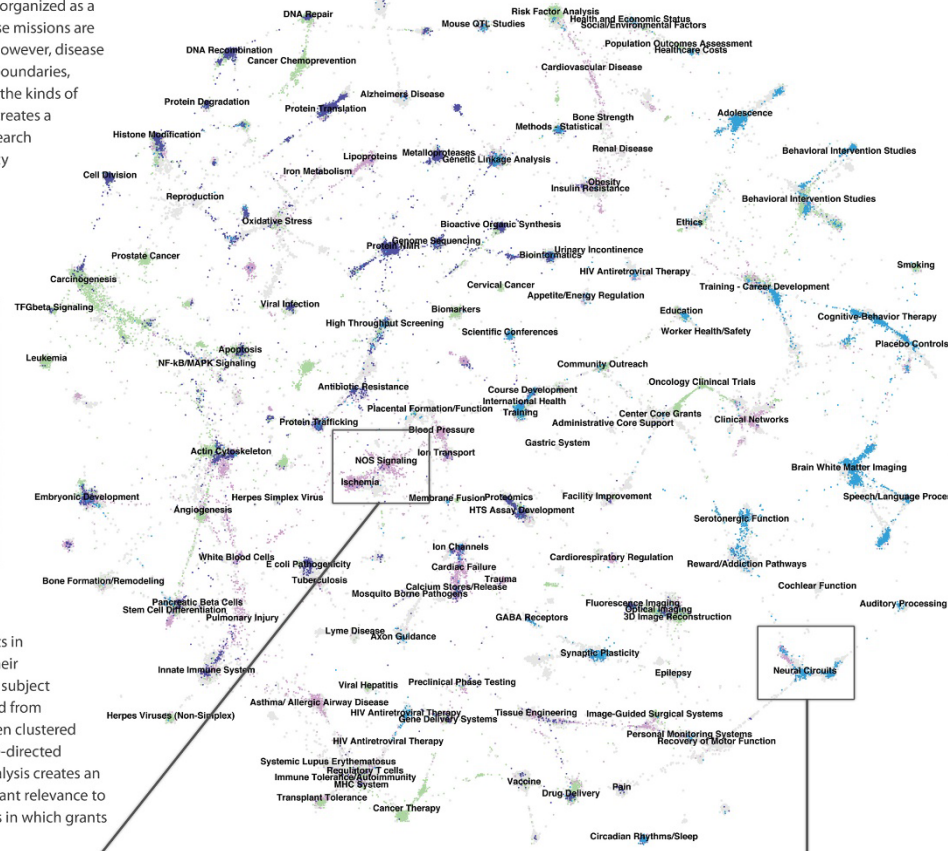
A Topic Map of NIH Grants 2007

Bruce W. Herr II (Chalklabs & IU), Gully Burns (ISI), David Newman (UCI), Edmund Talley (NIH)

The National Institutes of Health (NIH) is organized as a multitude of Institutes and Centers whose missions are primarily focused on distinct diseases. However, disease etiologies and therapies flout scientific boundaries, and thus there is tremendous overlap in the kinds of research funded by each Institute. This creates a daunting landscape for decisions on research directions, funding allocations, and policy formulations. Shown here is devised an interactive topic map for navigating this landscape, online at www.nihmaps.org. Institute abbreviations can be found at www.nih.gov/icd.

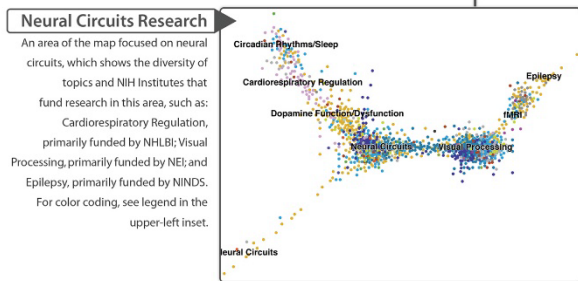


Topic modeling, a statistical technique that automatically learns semantic categories, was applied to assess projects in terms used by researchers to describe their work, without the biases of keywords or subject headings. Grant similarities were derived from their topic mixtures, and grants were then clustered on a two-dimensional map using a force-directed simulated annealing algorithm. This analysis creates an interactive environment for assessing grant relevance to research categories and to NIH Institutes in which grants are localized.



Cardiac Diseases Research

An area of the map focused on cardiovascular function and dysfunction. Cardiac Failure (primarily funded by NHLBI) is typically clustered next to Stroke (NINDS), since these are the two major medical emergencies associated with ischemia, which results from a restricted blood supply. Also localized in this area are grants focused on Nitric Oxide (NOS) Signaling, a major biochemical pathway for vasodilation, and grants on Hemodynamics, Sickle Cell Disease, and Aneurysms.

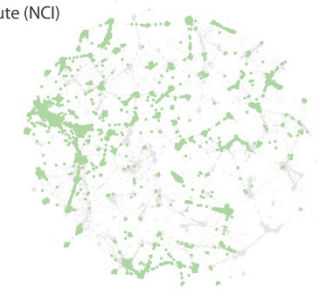


Neural Circuits Research

An area of the map focused on neural circuits, which shows the diversity of topics and NIH Institutes that fund research in this area, such as Cardiorespiratory Regulation, primarily funded by NHLBI; Visual Processing, primarily funded by NEI; and Epilepsy, primarily funded by NINDS. For color coding, see legend in the upper-left inset.

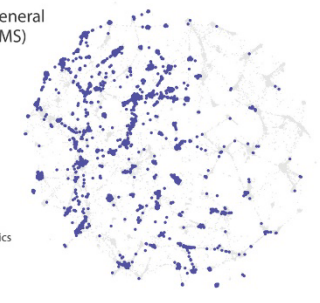
National Cancer Institute (NCI)

- TOP 10 TOPICS
- 1 Oncology Clinical Trials
 - 2 Cancer Treatment
 - 3 Cancer Therapy
 - 4 Carcinogenesis
 - 5 Risk Factor Analysis
 - 6 Cancer Chemotherapy
 - 7 Metastasis
 - 8 Leukemia
 - 9 Prediction/Prognosis
 - 10 Cancer Chemoprevention



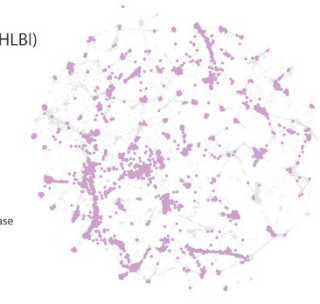
National Institute of General Medical Sciences (NIGMS)

- TOP 10 TOPICS
- 1 Bioactive Organic Synthesis
 - 2 X-ray Crystallography
 - 3 Protein NMR
 - 4 Computational Models
 - 5 Yeast Biology
 - 6 Metalloproteases
 - 7 Enzymatic Mechanisms
 - 8 Protein Complexes
 - 9 Invertebrate/Zebrafish Genetics
 - 10 Cell Division



National Heart, Lung, and Blood Institute (NHLBI)

- TOP 10 TOPICS
- 1 Cardiac Failure
 - 2 Pulmonary Injury
 - 3 Genetic Linkage Analysis
 - 4 Cardiovascular Disease
 - 5 Atherosclerosis
 - 6 Hemostasis
 - 7 Blood Pressure
 - 8 Asthma/ Allergic Airway Disease
 - 9 Gene Association
 - 10 Lipoproteins



National Institute of Mental Health (NIMH)

- TOP 10 TOPICS
- 1 Mood Disorders
 - 2 Schizophrenia
 - 3 Behavioral Intervention Studies
 - 4 Mental Health
 - 5 Depression
 - 6 Cognitive-Behavior Therapy
 - 7 AIDS Prevention
 - 8 Genetic Linkage Analysis
 - 9 Adolescence
 - 10 Childhood





Lisel Record
HIVE MC-Indiana PM
 Associate Director of CNS, IU.
 15 years of curatorial
 experience at nonprofit and
 public institutions.

Project Manager Focus

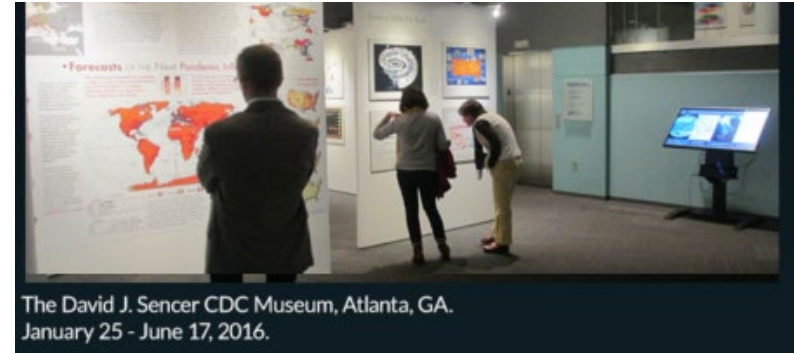
Managing projects, proposal
 development, and strategic
 planning efforts as well as directing
 outreach activities and working
 with Center collaborators.

Expertise with

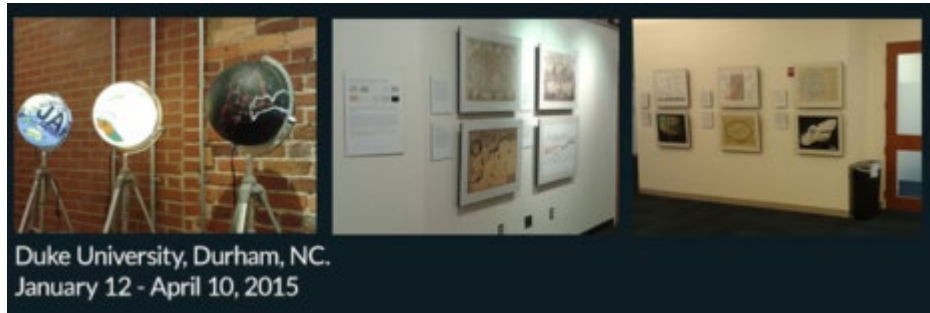
Science Communication, Outreach.
 Co-curator (with Börner) of the *Places
 & Spaces: Mapping Science* exhibit,
<http://scimaps.org> since 2013.



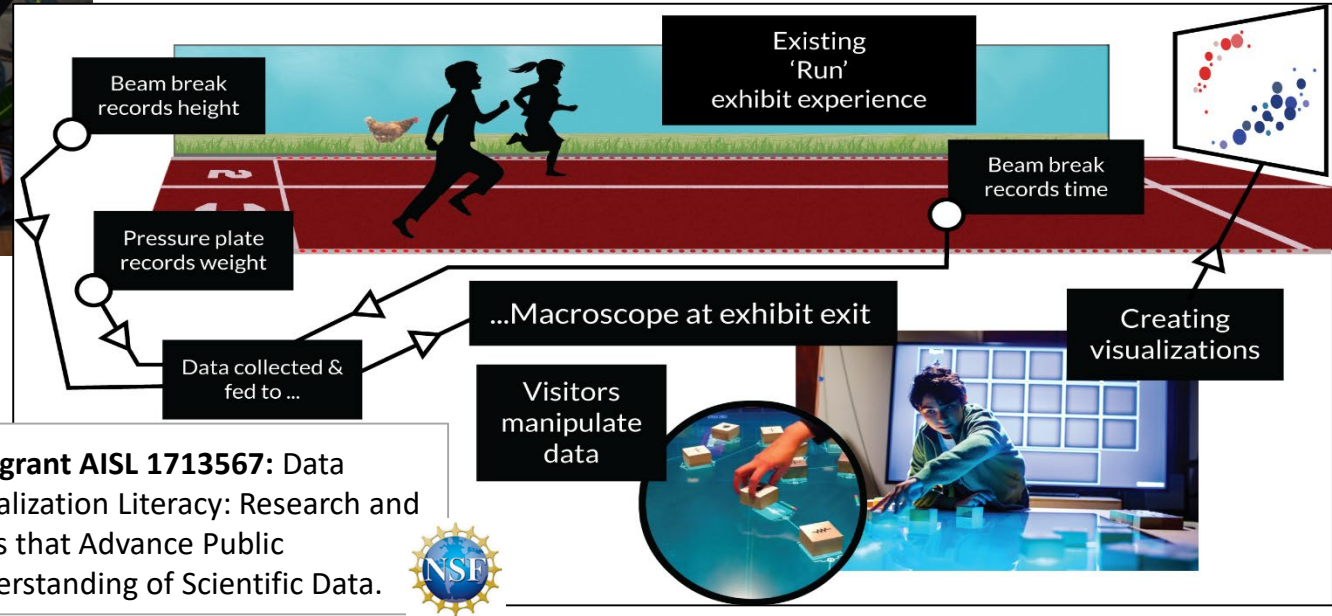
Exhibit Numbers
 14 Years
 100 Maps of S&T
 16 Macroscopes
 240 Authors
 350+ Venues
 4M Website visits



The David J. Sencer CDC Museum, Atlanta, GA.
 January 25 - June 17, 2016.



Duke University, Durham, NC.
 January 12 - April 10, 2015



**NSF grant AISL 1713567: Data
 Visualization Literacy: Research and
 Tools that Advance Public
 Understanding of Scientific Data.**





Bruce Herr II
HIVE MC-Indiana
 Lead System Architect
 at CNS, IU. 15y of
 software development
 (9y in industry).

Development Focus

- Data Federation
- Data Mining and Visualization
- Human Computer Interfaces
- Cyberinfrastructure Design

Open Source Software

CNS develops data-driven, user-focused software, visualizations, and web applications.

CNS's **CIShell** (cishell.org) is a data analysis and visualization framework, based on Java **OSGi industry-standard**.

It runs analysis and visualization algorithms within a single wrapper that enables transparent format conversion and **workflow queueing**.

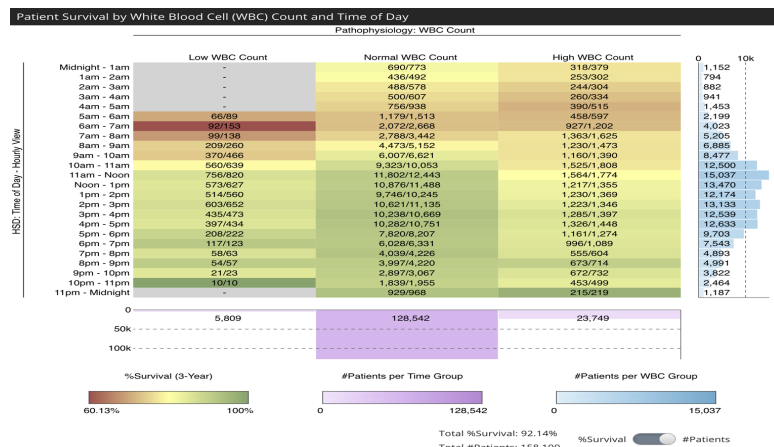
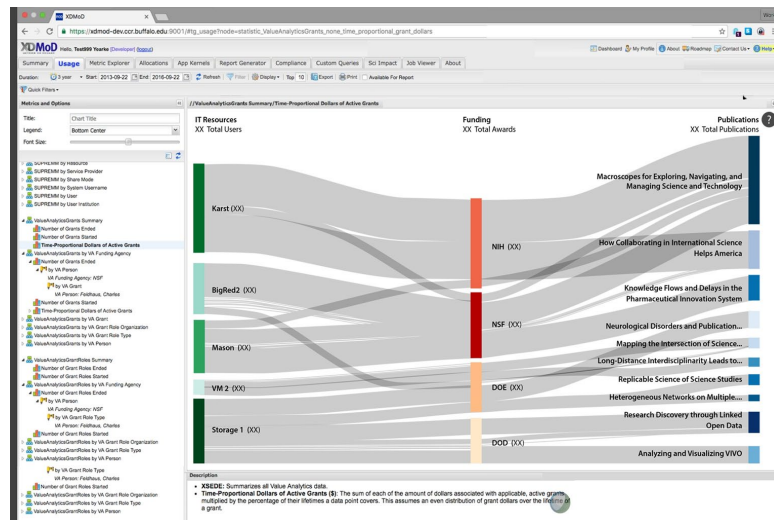
CIShell tools include the Science of Science (Sci2), Network Workbench (NWB), and Epidemiology (Epic) Tools that support temporal, geospatial, topical, and network analysis and vis. (180 algs, 225,000 downloads)

CNS's **Web Visualization Framework (WVF)** is a highly configurable packaging of several industry-standard web libraries (Angular, D3, head.js, Bootstrap, and others). Open source under commercially-compatible 3-clause BSD or MIT license.

Network Ties

Collaborations with developer teams at

- SMM Science Museum, <https://github.com/cns-iu/xmacroscope>
- nanoHUB at Purdue, <https://nanohub.org>
- XSEDE Metrics on Demand (XDMoD) <https://xdmod.ccr.buffalo.edu>
- i2b2: Informatics for Integrating Biology & the Bedside, <https://www.i2b2.org>
- VIVO Researcher Networking, <https://duraspace.org/vivo>



https://demo.cns.iu.edu/client/hsd/static/heatmap_hour.html



VIVO Ontology Classes

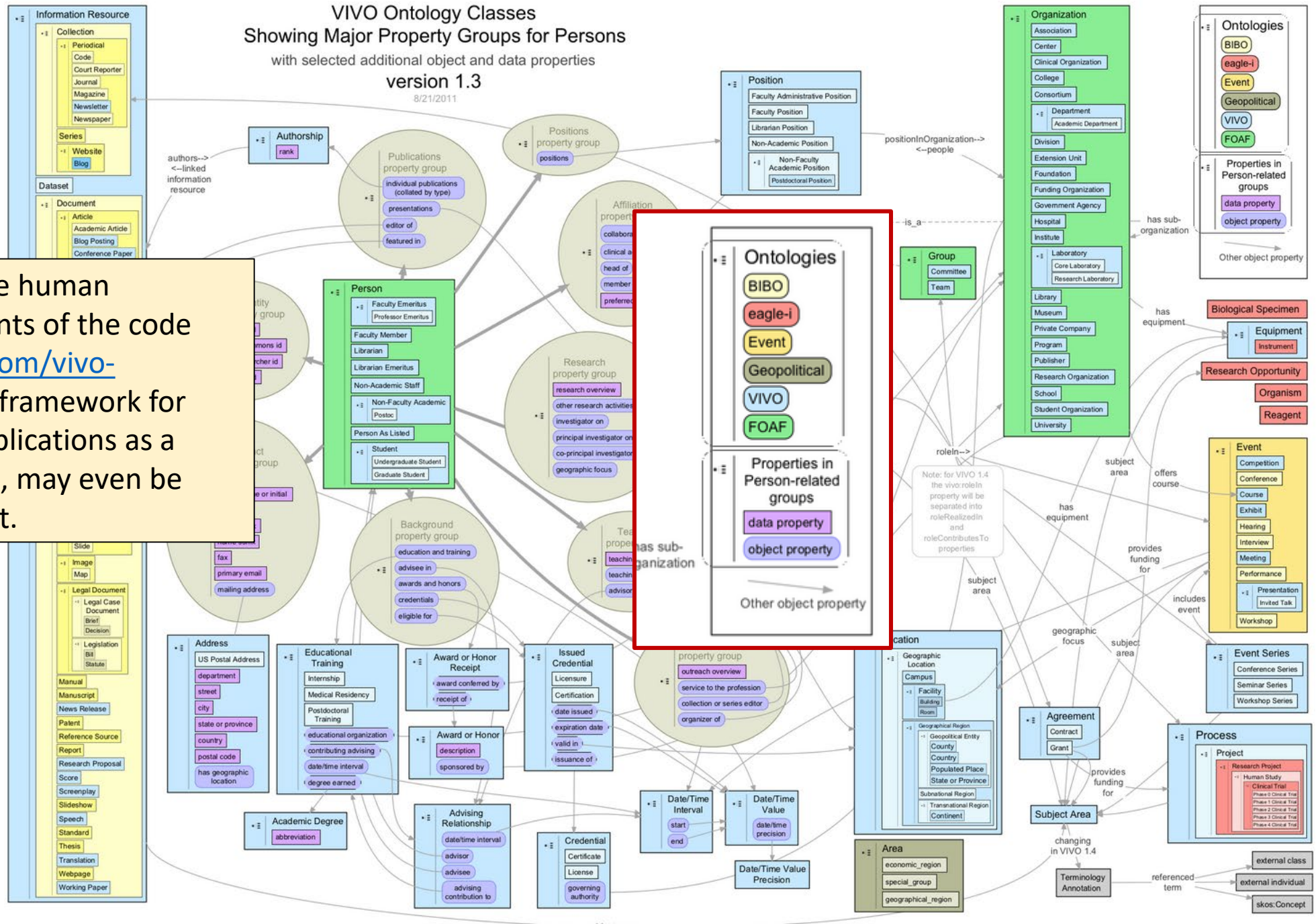
Showing Major Property Groups for Persons

with selected additional object and data properties

version 1.3

8/21/2011

VIVO can be applied to the human biomolecular atlas. Elements of the code like Vitro (<https://github.com/vivo-project/Vitro>), a full stack framework for building semantic web applications as a reference implementation, may even be repurposed for this project.



Börner, Conlon, Corson-Rikert, and Ding. 2012. *VIVO: A Semantic Approach to Scholarly Networking and Discovery*. Williston, VT: Morgan & Claypool Publishers.



Griffin Weber, MD, PhD

Associate Professor
Biomedical Informatics
Harvard Medical School

<http://weber.hms.harvard.edu>

"Profiles" social networking platform for scientists

Keywords

Last Name: [Input field]

Institution: [Dropdown menu]

Find People

Menu

- About Profiles
- Edit My Profile
- Manage Proxies
- Logout

Halamka, J is my...

- Collaborator
- Advisor (Current)
- Advisor (Past)
- Advisee (Current)
- Advisee (Past)

My Network

- John, Halamka
- Isaac, Kohane
- Korneth, Mandi
- Shawn, Murphy

History

- John Halamka
- George Church
- Mark Zeidel
- Any Goldberg

John David Halamka, M.D.

Academic Title: Associate Professor of Medicine
Administrative Title: Chief Information Officer
Department: Medicine- Beth Israel-Deaconess
Institution: Beth Israel Deaconess Medical Center
Address: 1135 Tremont St, Roxbury Crossing, MA 02120
Telephone: 617754-8002
Fax: 617754-8015
Email: jhalamka@caregroup.harvard.edu

Concepts

- Medical Records Systems
- Computerized Medical Record Linkage
- Patient Identification Systems
- Computer Security
- Regional Medical Programs

Co-Authors

- Kohane, Isaac
- Mandi, Kenneth
- Rind, David
- Safra, Charles
- Stair, Thomas

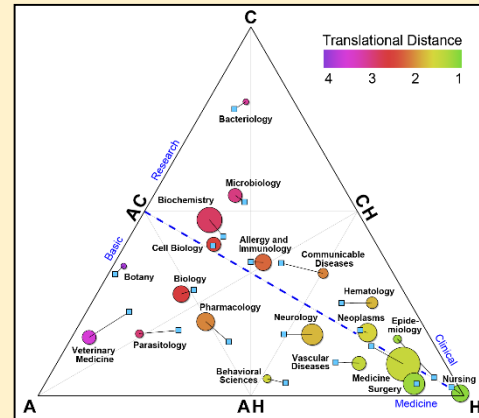
Similar People

- Middleton, Glockford
- Safra, Charles
- Kohane, Isaac
- Mandi, Kenneth
- Berglund, Bryan

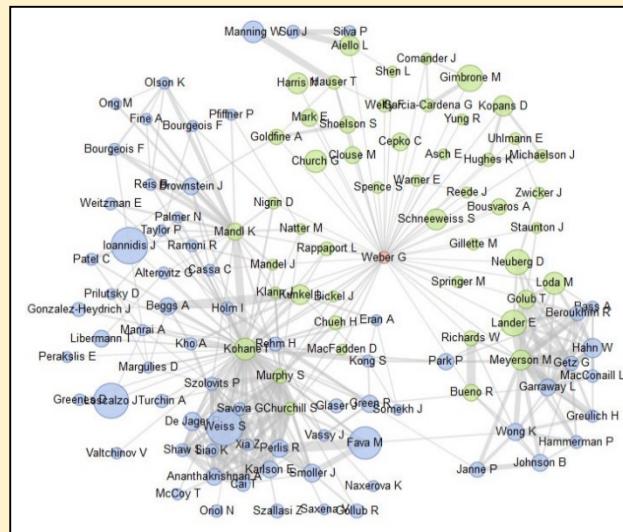
Publications

- Levine M, Adida B, Mandi K, Kohane J, Halamka J. What are the benefits and risks of fitting patients with radiofrequency identification devices. PLoS Med. 2007 Nov 27;4(11):e322.
- Halamka JD, Mandi KD, Tang PC. Early experiences with personal health records. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):1-7.
- Halamka J, Fournier GA. MA-SHARE MedInfo-ED medication history pilot. Am J Med Qual. 2008 Sep-Oct;21(5):296-8.
- Halamka J, Jusis A, Stubbsfield A, Westhus J. The security implications of VeriChip cloning. J Am Med Inform Assoc. 2006 Nov-Dec;13(6):601-7.
- Halamka JD. Patients should have to opt out of national electronic care records. AGAINST. BMJ. 2006 Jul 1;333(7557):41-2.
- Halamka J. On the road to RHIOs. Health Manag Technol. 2006 Jun;27(6):8, 10.
- Halamka JD. Health information technology: shall we wait for the evidence? Ann Intern Med. 2006 May 16;144(10):775-6.
- Halamka J, Aranow M, Ascenzo C, Bates DW, Berry K, Debor O, Fefferman J, Glaser J, Heindold J, Stanley J, Stone DL, Sullivan TE, Tripathi M, Wilkinson B. E-Prescribing collaboration in Massachusetts: early experiences from regional prescribing projects. J Am Med Inform Assoc. 2006 May-Jun;13(3):239-44.
- DeBor O, Diamond C, Grodecki D, Halamka J, Overhage JM, Shirky C. A tale of three cities—where RHIOs meet the NHIN. J Healthc Inf Manag. 2006;20(3):63-70.
- Halamka J. Early experiences with E-prescribing. J Healthc Inf Manag. 2006;20(2):12-4.

Triangle of Biomedicine



Social Network Analysis



"i2b2" query & analysis tool for clinical databases

The screenshot shows the 'Clinical Query 2' interface in a browser window. It includes a 'Query Tool' section with a 'Query Name' field (Hyper-BETA-Male@02:28:25), a 'Temporal Constraint' dropdown, and a grid for defining query groups (Group 1, Group 2, Group 3) with various filters like 'Gender', 'Race', 'Diagnoses', and 'Medications'. There are also buttons for 'Run Query', 'Clear', 'Print Query', and 'New Group'. The 'Query Status' section at the bottom shows 'Finished Query: Hyper-BETA-Male@02:28:25' with a compute time of 10.8 seconds and a patient count of 390003.

Record Linkage Across Datasets

TYPES OF DATA	STRUCTURED DATA		UNSTRUCTURED DATA	
	1	2	1	2
Medication	OTC medication Medication filled	Medication prescribed Dose Route NDC RxNorm	Medication instructions Allergies Out-of-pocket expenses	Medication taken Diaries Herbal remedies Alternative therapies
Demographics		HL7		
Encounters	Employee sick days	Visit type and time		Chief complaint
Diagnoses	Death records	SNOMED ICD-9 CPT ICD-9		Differential diagnosis
Procedures		LOINC Pathology, histology ECG Radiology		
Diagnoses (ordered)		Lab values, vital signs SNPs, arrays		REPORTS TRACKING IMAGES
Diagnoses (results)				DIGITAL CLINICAL NOTES PHYSICAL EXAMINATIONS
Genetics		23andMe.com		PAPER CLINICAL NOTES
Social history		Police records		TWEETS FACEBOOK POSTINGS
Family history		Ancestry.com		
Symptoms		Indirect from OTC purchases Fitness club memberships, grocery store purchases		
Lifestyle		Credit CARD PURCHASES		
Socioeconomic		Census records, Zillow, LinkedIn		
Social network		Facebook friends, Twitter hashtags		
Environment		Climate, weather, public health databases, HealthMap.org, GIS maps, EPA, phone GPS		

Probabilistic linkage to validate existing data or fill in missing data

Examples of biomedical data:

- Pharmacy data
- Health care center (electronic health record) data
- Claims data
- Registry or clinical trial data
- Data outside of health care system

Ability to link data to an individual:

- Easier to link to individuals
- Harder to link to individuals
- Only aggregate data exists

Data quantity: More, Less

Probabilistic linkage to obtain new types of data



Paul Macklin, Ph.D.
 Associate Professor
 Intelligent Systems Eng.
 Indiana University
<http://MathCancer.org>
[@MathCancer](https://twitter.com/MathCancer)

Ties to software community

- Boolean signal networks (Inst. Curie)
- SBML ODE signal networks (IU)
- High-throughput model exploration (Argonne National Lab)
- Cloud-hosted simulations in nanotherapy (IU, Purdue)
- 3-D rendering (ParaView)
- Broader open source comp bio community via MultiCellDS standards work
 - Chaste
 - Tissue Simulation Toolkit
 - CompuCell3D
 - Morpheus
 - ...

Multicellular systems biology

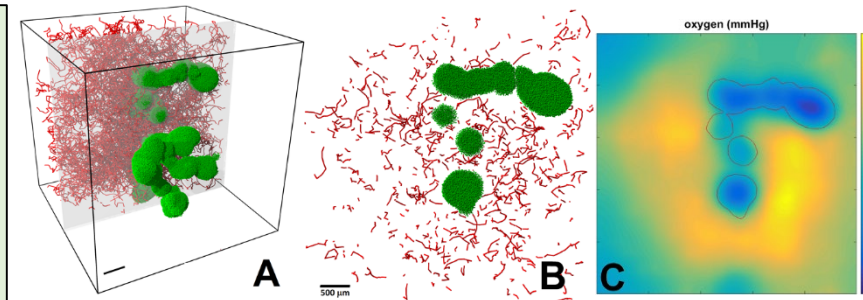
- Diffusive transport of growth substrates, signaling factors, ...
- ECM models
- Off-lattice cell models.
- Multicellular data standards

Open source software

- BioFVM for transport
- PhysiCell for cell modeling
- MultiCellDS for standards
- OpenSource.MathCancer.org

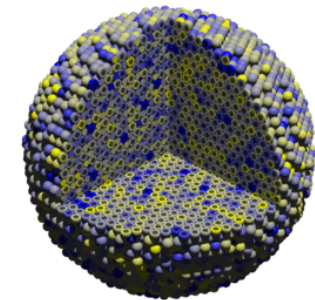
Close experimental ties

- Breast cancer hypoxia in mouse models: Daniele Gilkes (JHU)
- Breast cancer invasion in organoid models: Andy Ewald (JHU)
- Colon cancer metabolism in organoids: Shannon Mumenthaler (USC)
- Synthetic biology: David Kehoe (IU)
- High-end experimental, microscopy communities via NCI CSBC/PSON



Using BioFVM to solve oxygen release in a highly-vascularized 3-D tumor tissue

Current time: 0 days, 0 hours, and 0.00 minutes
 18317 agents



Using BioFVM+PhysiCell to simulate immunosurveillance of 3-D tumors

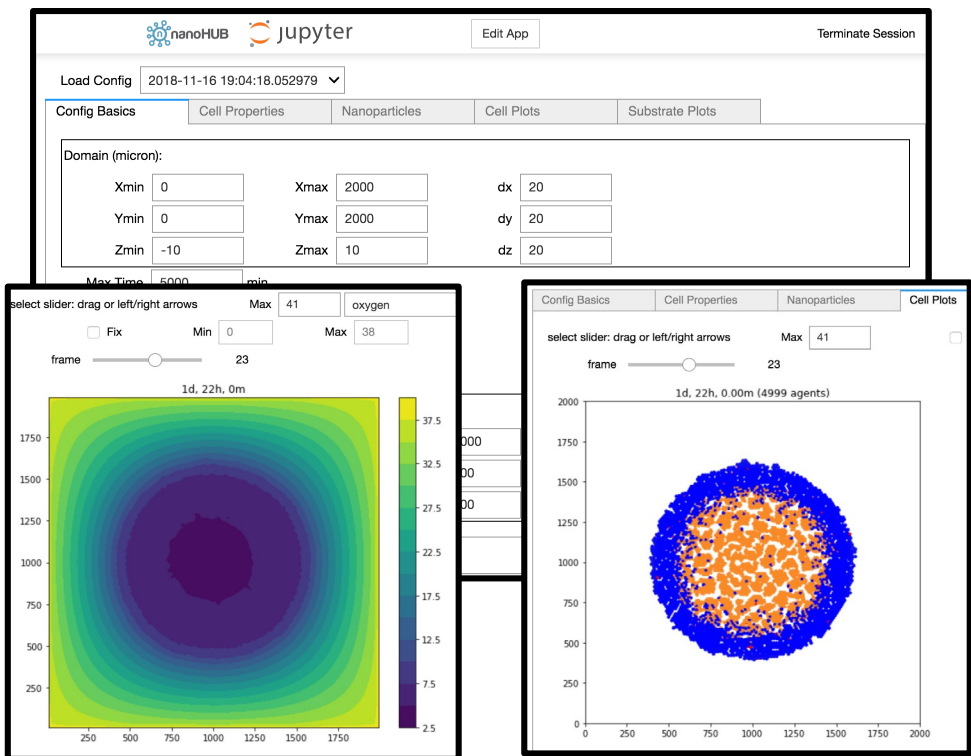
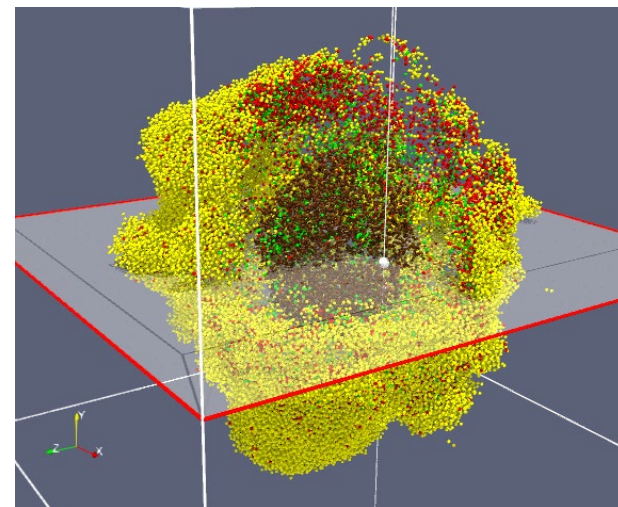
[\[Watch on YouTube \(4K\)\]](#)



Randy Heiland
 Research Associate,
 Macklin Lab
 Indiana University
<http://rheiland.github.io>

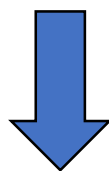
Building on open source
 software and data standards:
 Python, Jupyter, VTK, ITK,
 Slicer, ParaView, PhysiCell,
 MultiCellIDS, ISA-Tab, ...

Software engineering, scientific
 data analysis/vis, cybersecurity.



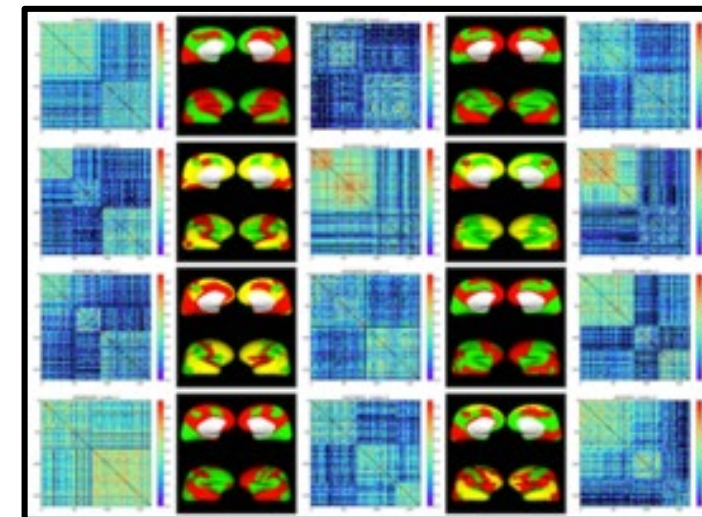
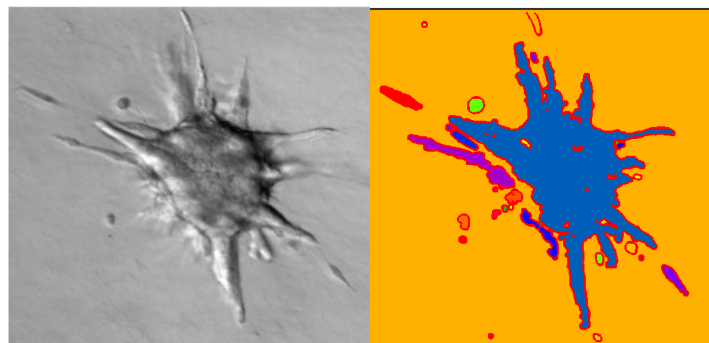
Jupyter notebook GUIs

MultiCellIDS (.xml)



Python
 conversion
 script

ISA-Tab





Samuel Friedman, Ph.D.
 Staff Scientist
 Opto-Knowledge
 Systems, Inc. (OKSI)
 Torrance, CA
[@CompCancer](#)

Multicellular biology

- Lead Developer for MultiCellular Data Standard (MultiCellDS)
- Integration of pre-existing standards when possible and practical

Expertise with Geographic Information System (GIS)

- Performing data fusion with need to translate between multiple modalities
- Uncertainty and variability quantification with data fusion
- Creation of geographic maps based on imaging processing data pipelines
- Algorithm development for ideal surveying of geographic domains

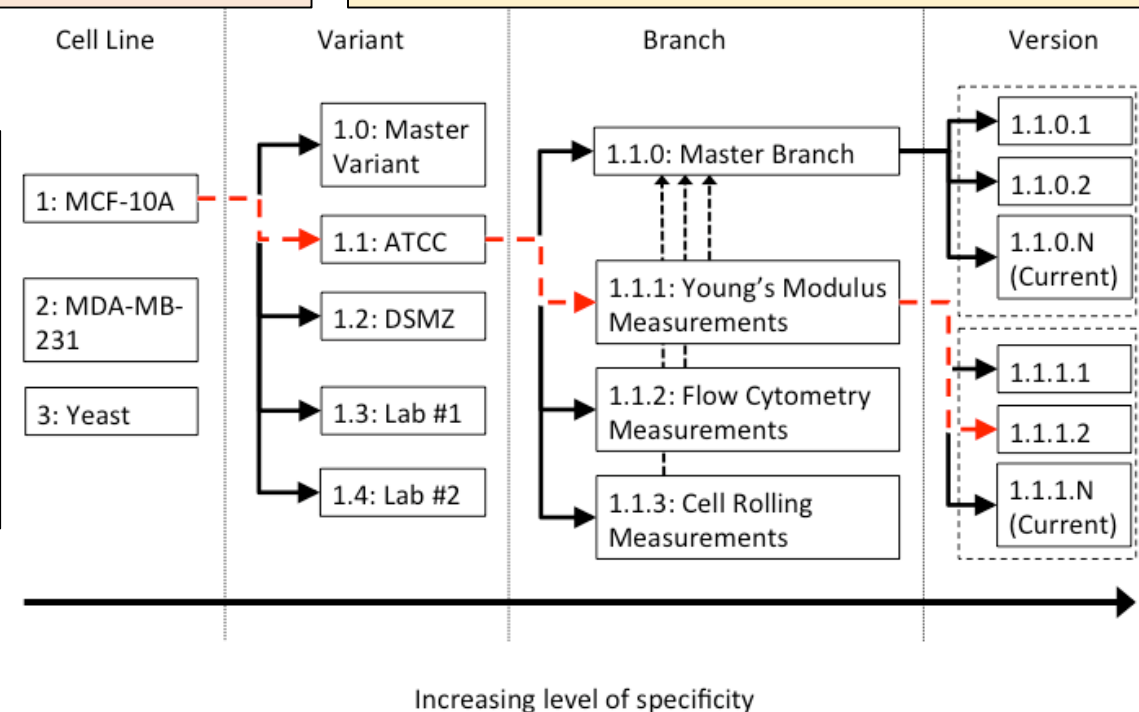
Ties to COMBINE Community

- Active member of the Computational Models In Biology Network (COMBINE)
 - SBML, CellML, NeuroML...
 - Multicellular modeling
 - Annotation of computational models with ontologies (Washington; Seattle Children's Hospital)
- Preliminary discussions of integration of Electronic Health Records (EHRs) with Dave Nickerson (Auckland)
 - Understanding context of the data is critical to understanding

Open source software

- Developing APIs for MultiCellDS for multiple languages
- Previously developer for HTCondor

Hierarchical Organization of Digital Cell Lines in MultiCellDS





Ellen M. Quardokus

Research Associate, 28 yrs of research & laboratory and project management (25 yrs @ IU)

Indiana University

ellenmq@indiana.edu

HuBMAP MC-Indiana CCF literature review

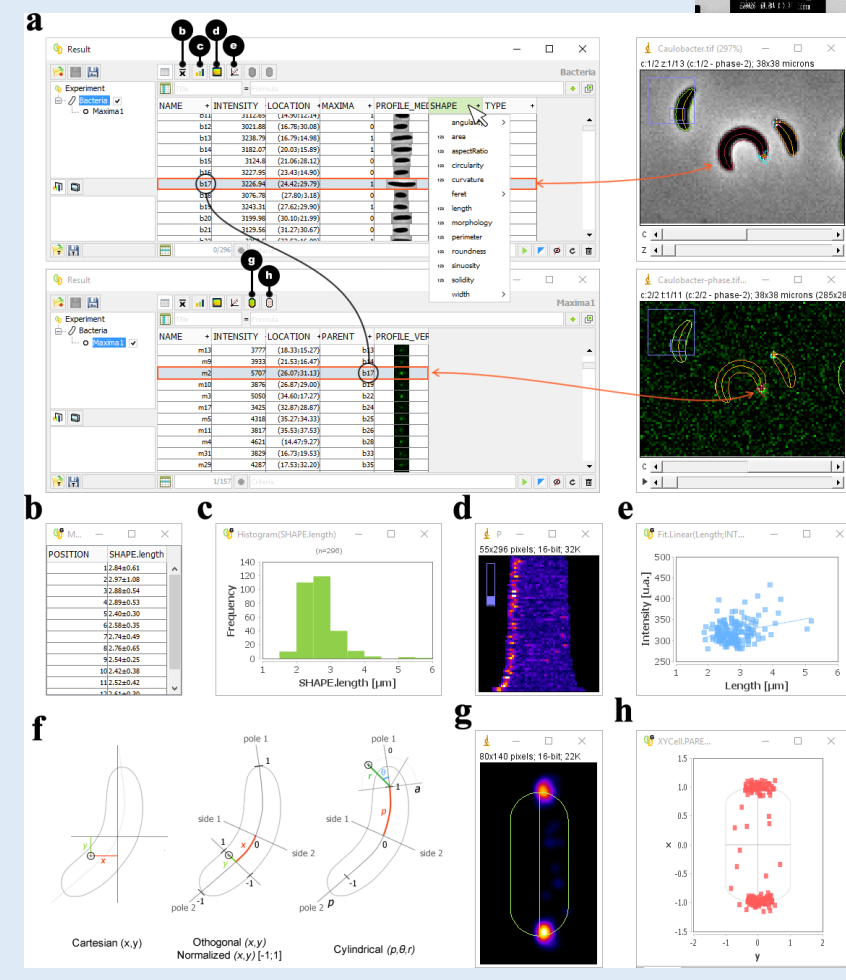
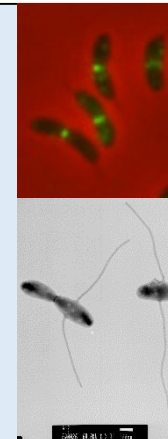
Common Coordinate Frameworks of Human BioMolecular Data: Desirable Properties, Methods, Data Structures and Exemplary Interfaces

1. Introduction: Challenges and Opportunities
2. Atlas Building Considerations Across Scales: Whole Body vs Organs, Tissues, Cells, Subcellular
3. Common Coordinate Framework (CCF) Construction for Different Scales and Regions of The Body
4. Ontology Building and Editing from Existing Ontologies
5. User Interfaces and Visualization of Ontologies & Data
6. Conclusions and Future Work
7. References

Part of IU team that submitted Seed Network proposal to the Chan-Zuckerberg/Helmsley Charitable Trust-The Human Gut Cell Atlas (GCA)

Extensive Microscopy Expertise

- Fluorescence, Immunofluorescence
- Electron (SEM/TEM)
- Sample preparation
- Image Acquisition
- Image analysis & Troubleshooting
- Microscope & software training
- Video Tutorials for MicrobeJ analysis ImageJ plug-in



- Research Project Management
- Database design & management
- Scientific software beta testing
- Software Interface collaborations
- Computer software training

Extensive Expertise in Techniques & Troubleshooting:

- Molecular Biology
 - PCR, Illumina Next Generation Sequencing (NGS) library construction, genomic & Sanger sequence analysis, DNA/RNA isolation & microarray analysis
- Protein Biochemistry
 - Protein purification
 - Enzymatic assays
 - Custom Antibody Production & Testing
 - Western blotting
 - Radioactive metabolic labeling
- Microbial Genetics
- Fluorescence-activated cell sorting (FACS)

Teaching Experience:
Human Anatomy & Physiology
Human Genetics
Comparative Anatomy

MC-Indiana

User Needs Analysis



HuBMAP

The Human BioMolecular Atlas Program

Data Visualization Literacy: Definitions, Conceptual Frameworks, Exercises, and Assessments

Katy Börner¹, Andreas Bueckle¹, Michael Ginda¹

¹Indiana University

69
70
71
72
73
74
75
76
77

1
2
3
4
5
6
7
8
9

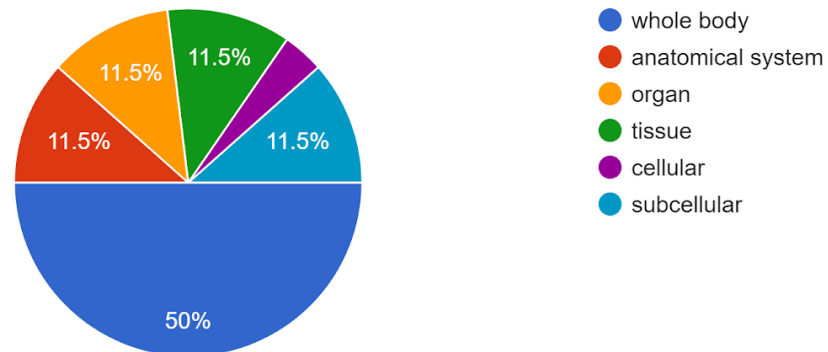
Insight Needs	Data Scales	Analyses	Visualizations	Graphic Symbols	Graphic Variables	Interactions
<ul style="list-style-type: none"> • categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships 	<ul style="list-style-type: none"> • nominal • ordinal • interval • ratio 	<ul style="list-style-type: none"> • statistical • temporal • geospatial • topical • relational 	<ul style="list-style-type: none"> • table • chart • graph • map • tree • network 	<ul style="list-style-type: none"> • geometric symbols <ul style="list-style-type: none"> point line area surface volume • linguistic symbols <ul style="list-style-type: none"> text numerals punctuation marks • pictorial symbols <ul style="list-style-type: none"> images icons statistical glyphs 	<ul style="list-style-type: none"> • spatial <ul style="list-style-type: none"> position • retinal <ul style="list-style-type: none"> form color optics motion 	<ul style="list-style-type: none"> • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion

User Needs Analysis: Gathering User Stories

- Data was collected during the HIVE kickoff meeting held October 11-12, 2018.
- 26 user stories were provided from NIH and HIVE award project participants
- Data collection took place via a Google Form.
- Please note that many stakeholders are not represented here. Additional data collection continues.

At what level would you like to enter the map:

26 responses



As a computational biologist, I would like to use HubMAP to integrate data across platforms, modalities, and organs so that I can

*genes,
ell
common*

As a cell biologist, I would like to use HubMAP to analyze data so that I can develop research in individual labs

Identify Stakeholders / User Needs Analysis

Current Study:

1. Demographics
2. Key Functionality
3. Suggest Other Experts
 - Snowball Sample
5. User Studies
 - Invitation to participate in user studies for prototype interfaces to the HUBMAP Atlas.

Planned Study

Review Existing Atlases

Ask experts to comment, e.g., on

- Overall Functionality: <http://www.emouseatlas.org/emap/home.html>, <http://www.emouseatlas.org/emap/imageBrowse/>
- Visual Interface: <https://www.openanatomy.org/atlas-pages/>
- Ontology-Tissue Browser: <https://liveratlas.org>
- Access to Data: <https://www.igp.uu.se/research/hpa/workflow>

Collaboration Opportunity:

Please share your input via <https://bit.ly/2ROGGpg>.

Collaboration Opportunity:

User Needs Analyses

Please let us know if you are interested to help specify the look-and-feel plus functionality of qualitatively new interfaces to tissue data.

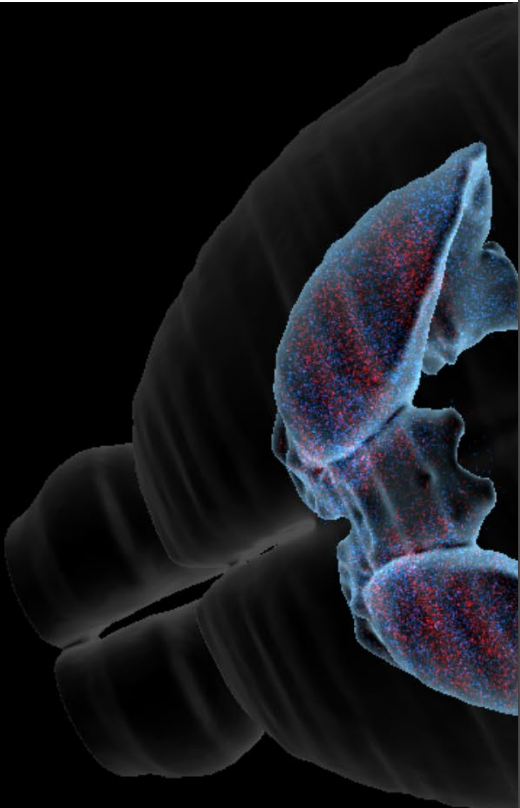
User Studies

Please let us know if you are interested to explore and provide feedback on initial user interface prototypes.

EPFL Blue Brain Cell Atlas [about](#) [contact](#)

Search or select a region

- Basic cell groups and regions
 - Cerebrum
 - Cerebral cortex
 - Cortical plate
 - Isocortex
 - Olfactory areas
 - Hippocampal formation
 - Hippocampal region
 - Retrohippocampal region
 - Cortical subplate
 - Cerebral nuclei
 - Striatum
 - Pallidum
 - Brain stem
 - Interbrain
 - Thalamus
 - Hypothalamus
 - Periventricular zone
 - Periventricular region
 - Hypothalamic medial zone
 - Hypothalamic lateral zone
 - Midbrain
 - Hindbrain
 - Cerebellum
 - Cerebellar cortex
 - Cerebellar nuclei
 - fiber tracts



<https://bbp.epfl.ch/nexus/cell-atlas>


THE HUMAN PROTEIN ATLAS [Search](#) [Fields »](#)



















[MENU](#) [HELP](#) [NEWS](#)

THE HUMAN PROTEOME : THE TISSUE ATLAS








THE TISSUE AND ORGAN PROTEOMES¹ Explore the proteomes of specific tissues and organs

The expression for all protein-coding genes in all major tissues and organs in the human body can be explored in this interactive database, including numerous catalogues of proteins expressed in a tissue-restricted manner.



Brain		Heart	
Adrenal gland		Skeletal muscle	
Parathyroid gland		Gastrointestinal tract	
Thyroid gland		Salivary gland	
Lung		Esophagus	
Bone marrow and lymphatic tissues		Stomach	
Bone marrow		Duodenum	
Lymph node		Small intestine	
Spleen		Colon	
Appendix		Pancreas	
Liver		Kidney	
Gallbladder		Breast	
Testis		Cervix	
Epididymis		Endometrium	
Seminal vesicle		Ovary	
Prostate		Placenta	
Adipose tissue		Skin	

Extended tissue profiling¹

Extended brain samples		Extended skin samples	
Mouse brain		Full section adrenal gland	
Eye		Thymus	
Lactating breast			

<https://www.proteinatlas.org/humanproteome/tissue>

MC-Indiana

Common Coordinate
Framework (CCF)



HuBMAP

The Human BioMolecular Atlas Program

Concept CCF: Anatomic coordinates

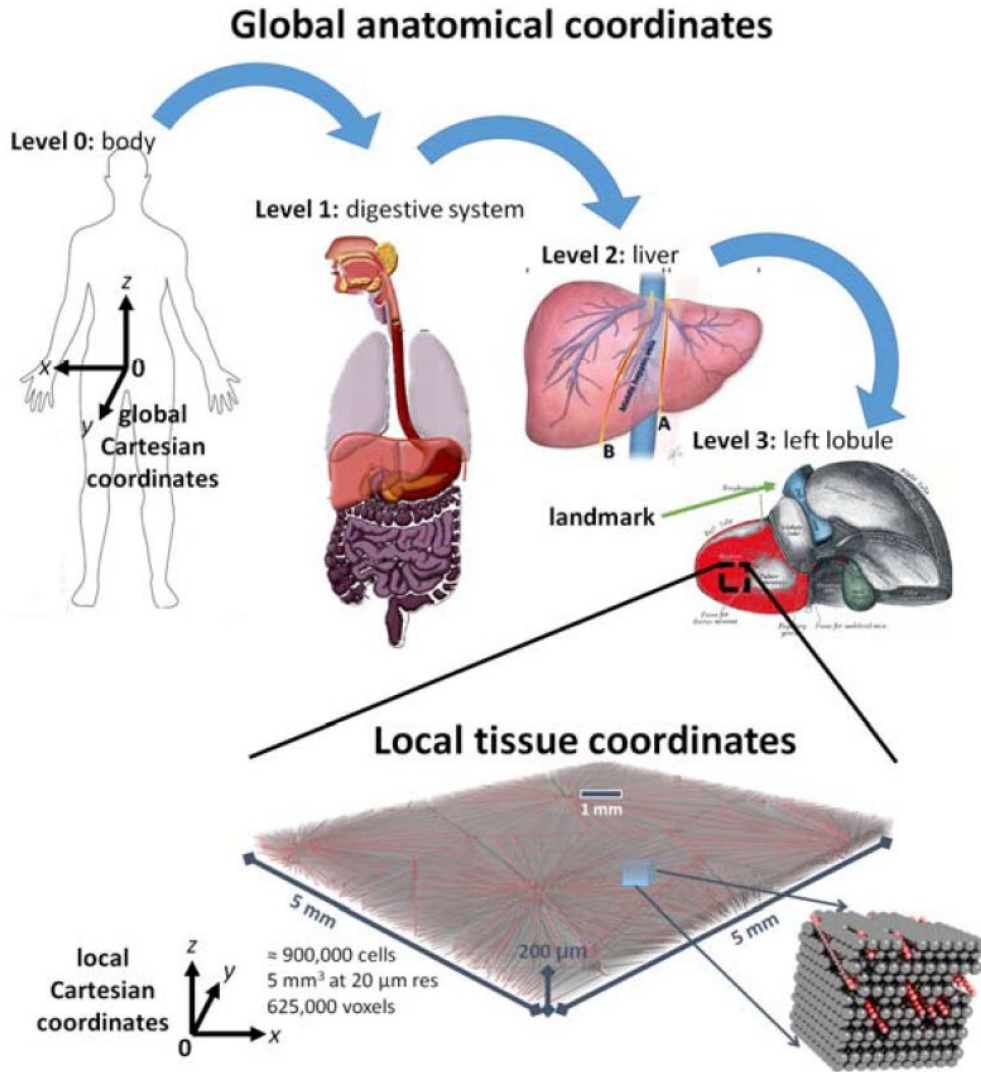


Fig. 5. CCF concept, navigating through the global anatomical coordinate system to insert a synthetic tissue sample (from PhysiCell⁴) into the left liver lobe with a local coordinate system.

Motivation:

Global coordinates will initially difficult.

- N universal Cartesian, Barycentric, coordinate system to date
- Each sample is from a different individual. No expectation of aligned boundaries, scaling, etc.

However:

- Anatomic information allows localization of a sample
- Each sample has its own local coordinates

Concept:

- Use hierarchical anatomic information to localize the sample within the body (functionally standardized between people)
- Emplace local Cartesian coordinates
- Record extra metadata on distances to known landmarks
- Eventually define global coordinates via landmarks
- Leverage, combine, and extend elements in existing ontologies

Existing technologies:

- Physics, Cell, Metadata, Units, and other ontologies
- MultiCellIDS (next slide)
- ApiNATOMY and other anatomic annotations / ontologies

MultiCellDS goal:

Standardize (extracted) multicellular data:

- **Metadata**
- **Cell line parameters**
- **Single-time snapshots**
- **Time course data**

Multidisciplinary Team:

- Clinicians
- Biologists
- Data Scientists
- Mathematicians

International Team:

- USA, UK, Netherlands, Germany (for now)

Involvement of the open source community:

- Chaste, Morpheus, Tissue Simulation Toolkit, PhysiCell, BioFVM, CellSys*

MultiCellDS: a community-developed standard for curating microenvironment-dependent multicellular data

Authors:

Samuel H. Friedman¹, Alexander R. A. Anderson², David M. Bortz³, Alexander G. Fletcher⁴, Hermann B. Frieboes⁵, Arinadrezza Ghaffarizadeh¹, David Robert Grimes⁶, Andrea Hawkins-Daarud⁷, Stefan Hoehme⁸, Edwin F. Juarez^{1,9}, Carl Kesselman¹⁰, Roeland M.H. Merks^{11,12}, Shannon M. Mumenthaler¹, Paul K. Newton¹³, Kerri-Ann Norton¹⁴, Rishi Rawat¹, Russell C. Rockne¹⁵, Daniel Ruderman¹, Jacob Scott¹⁶, Suzanne S. Sindi¹⁷, Jessica L. Sparks¹⁸, Kristin Swanson⁷, David B. Agus¹, Paul Macklin^{19,*} (corresponding author)

¹ Lawrence J. Ellison Institute for Transformative Medicine, University of Southern California, Los Angeles, CA USA

² Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, FL USA

³ Applied Mathematics, University of Colorado, CO USA

⁴ School of Mathematics & Statistics and Bateson Centre, University of Sheffield, Sheffield, United Kingdom

⁵ Bioengineering, University of Louisville, Louisville, KY USA

⁶ Cancer Research UK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Oxford United Kingdom

⁷ Mathematical NeuroOncology, Mayo Clinic, Phoenix, AZ USA

⁸ Institute of Computer Science, University of Leipzig, Leipzig, Germany

⁹ Electrical Engineering, University of Southern California, Los Angeles, CA USA

¹⁰ Information Sciences Institute, University of Southern California, Marina del Rey, CA USA

¹¹ Life Sciences Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

¹² Mathematical Institute, Leiden University, Leiden, The Netherlands

¹³ Aerospace and Mechanical Engineering, University of Southern California, Los Angeles, CA USA

¹⁴ Systems Biology Laboratory, Johns Hopkins University, Baltimore, MD USA

¹⁵ Mathematical Oncology, City of Hope, Duarte, CA USA

¹⁶ Translational Hematology and Oncology Research, Cleveland Clinic, Cleveland, OH USA

¹⁷ Applied Mathematics, University of California, Merced, CA USA

¹⁸ Chemical, Paper, and Biomedical Engineering, Miami University, Oxford, OH USA

¹⁹ Intelligent Systems Engineering, Indiana University, Bloomington, IN USA

* corresponding author: email: macklinp@iu.edu, www: <http://MathCancer.org>

Abstract:

Exchanging and understanding scientific data and their context represents a significant barrier to advancing research, especially with respect to information siloing. Maintaining information provenance and providing data curation and quality control help overcome common concerns and barriers to the effective sharing of scientific data. To address these problems in and the unique challenges of multicellular systems, we assembled a panel composed of investigators from several disciplines to create the MultiCellular Data Standard (MultiCellDS) with a use-case driven development process. The standard includes (1) digital cell lines, which are analogous to traditional biological cell lines, to record metadata, cellular microenvironment, and cellular phenotype variables of a biological cell line, (2) digital snapshots to consistently record simulation, experimental, and clinical data for multicellular systems, and (3) collections that can logically group digital cell lines and snapshots. We have created a MultiCellular DataBase (MultiCellDB) to store digital snapshots and the 200+ digital cell lines we have generated. MultiCellDS, by having a fixed standard, enables discoverability, extensibility, maintainability, searchability, and sustainability of data, creating biological applicability and clinical utility that permits us to identify upcoming challenges to uplift biology and strategies and therapies for improving human health.

Metadata:

- Creator, curator, contact ...
- Versioning
- Provenance: data sources

Phenotype dataset:

- Microenv. context
- Hierarchical Phenotype
 - size, cycle, death, motility, secretions, ...
- Easily extended with molecular-level details

Digital Cell Lines:

For a single cell type:

- Metadata
- One or more phenotype datasets

Digital Snapshots:

- Metadata
- Spatial sampling of ME
- Spatial list of blood vessel segments
- Spatial list of all cells and their phenotypes
- Can use cell densities, too

MultiCellDS: a community-developed standard for curating microenvironment-dependent multicellular data

Authors:

Samuel H. Friedman¹, Alexander R. A. Anderson², David M. Bortz³, Alexander G. Fletcher⁴, Hermann B.

MultiCellDS goal:

Standardize (extracted) multicellular data:

- Metadata
- Cell line parameters
- Single-time s
- Time course

Multidisciplin

- Clinicians
- Biologists
- Data Scientist
- Mathematici

International

- USA, UK, N
- Germany (f

Involvement source comm

- Chaste, Mo
- Tissue Simu
- Toolkit, Phy
- BioFVM, CellSys*

Digital cell line

Metadata

Phenotype measurement set

Microenvironment descriptors

Phenotype measurements

Phenotype measurement set

Microenvironment descriptors

Phenotype measurements

Phenotype measurement set

Microenvironment descriptors

Phenotype measurements

Digital snapshot

Metadata

Microenvironment

Variables

Mesh information

Variable data

Basement membranes

Vascular network

Cellular Information

Digital cell lines

Mesh information

Cell populations

Metadata:

- Creator, curator, contact ...
- Versioning
- Provenance: data sources

Phenotype dataset:

- Microenv. context
- Hierarchical Phenotype
 - size, cycle, death, motility, secretions, ...
- Easily extended with molecular-level details

Digital Cell Lines:

For a single cell type:

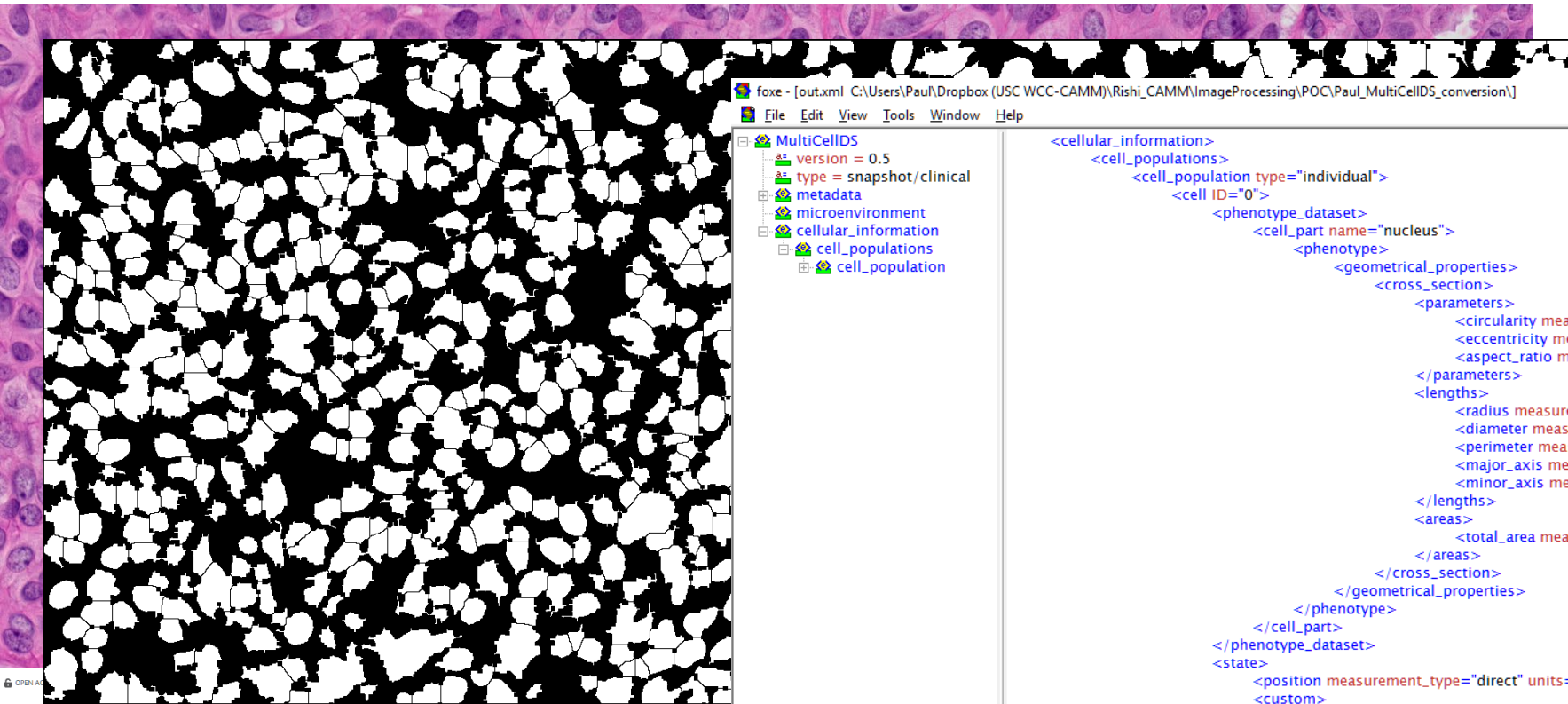
- Metadata
- One or more phenotype datasets

Digital Snapshots:

- Metadata
- Spatial sampling of ME
- Spatial list of blood vessel segments
- Spatial list of all cells and their phenotypes
- Can use cell densities, too

have generated. MultiCellDS, by having a fixe searchability, and sustainability of data, creati identify upcoming challenges to uplift biology

Early example: Breast cancer pathology



```
MultiCellDS
  version = 0.5
  type = snapshot/clinical
  metadata
  microenvironment
  cellular_information
  cell_populations
    cell_population
      <cellular_information>
        <cell_populations>
          <cell_population type="individual">
            <cell ID="0">
              <phenotype_dataset>
                <cell_part name="nucleus">
                  <phenotype>
                    <geometrical_properties>
                      <cross_section>
                        <parameters>
                          <circularity measurement_type="direct" units="dimensionless">7.3628793e-001</circularity>
                          <eccentricity measurement_type="direct" units="dimensionless">7.8810628e-001</eccentricity>
                          <aspect_ratio measurement_type="direct" units="dimensionless">6.1553918e-001</aspect_ratio>
                        </parameters>
                      </cross_section>
                    </phenotype>
                  </cell_part>
                </phenotype_dataset>
              </cell_population>
            </cell_populations>
          </cellular_information>
        </cell_population>
      </cell_populations>
    </cellular_information>
  </MultiCellDS>
```

Computational Pathology to Discriminate Benign from Malignant Intraductal Proliferations of the Breast

Fei Dong, Humayun Irshad, Eun-yeong Oh, Melinda F. Lervill, Elena F. Brachtel, Nicholas C. Jones, Nicholas W. Knoblauch, Laleh Montaser-Kouhsari, Nicole B. Johnson, Luigi K. F. Rao, Beverly Faulkner-Jones, David C. Wilbur, Stuart J. Schnitt, Andrew H. Beck

Published: December 9, 2014 • DOI: 10.1371/journal.pone.0114885

Saves	Citation
2,510 Views	2 Shares

Article	Authors	Metrics	Comments	Related Content
---------	---------	---------	----------	-----------------

Abstract
Introduction
Materials and Methods
Results
Discussion
Author Contributions
References

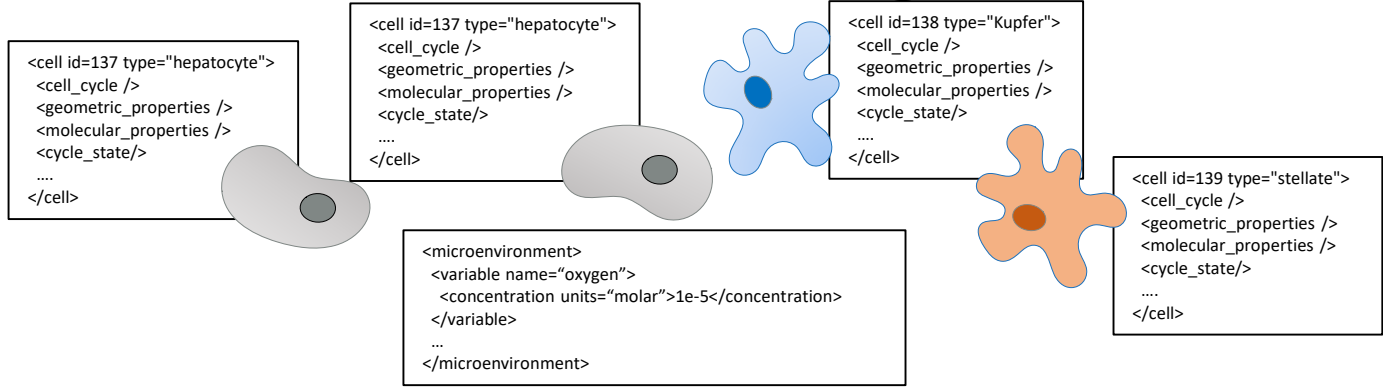
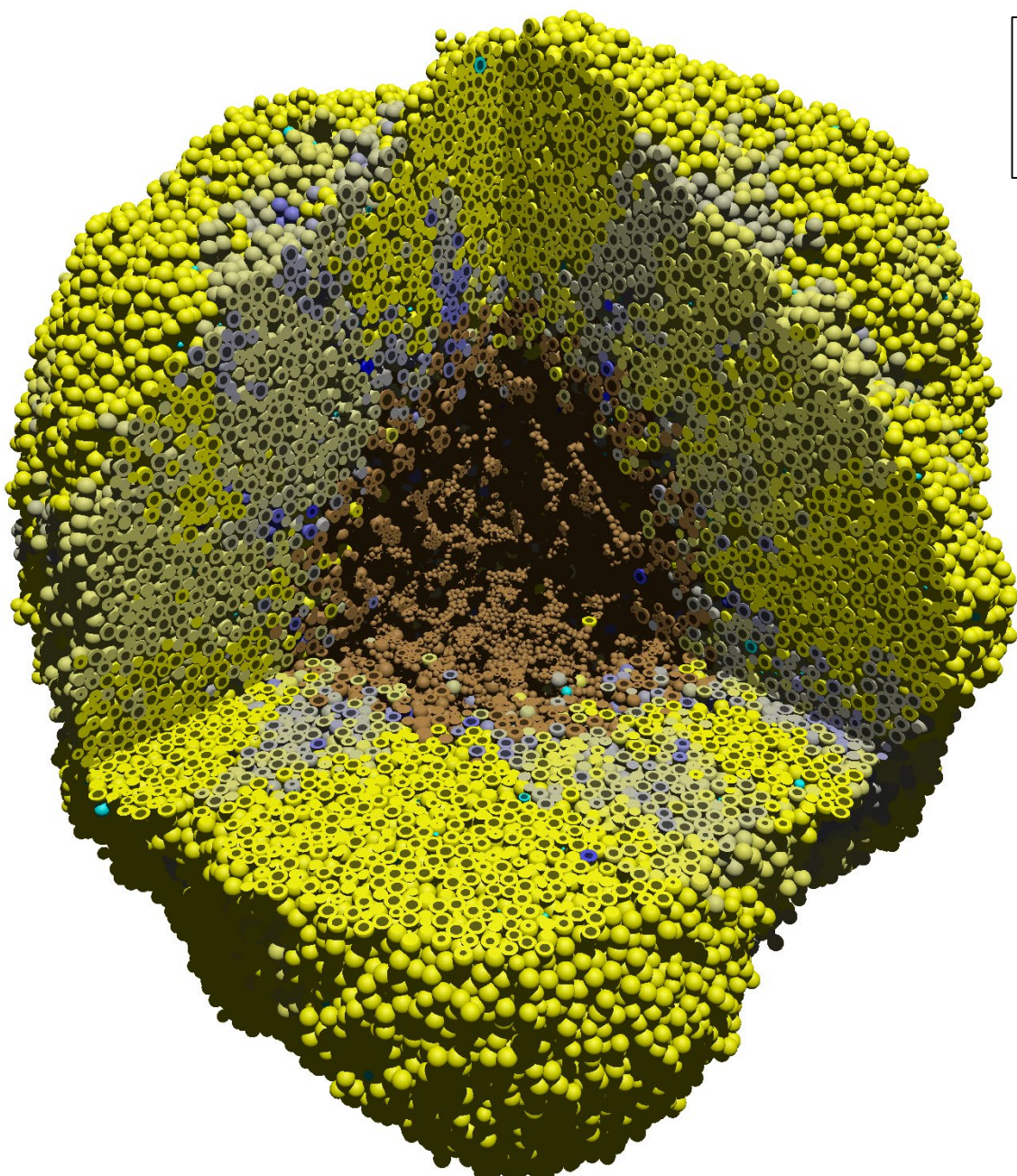
The categorization of intraductal proliferative lesions of the breast based on routine light microscopic examination of histopathologic sections is in many cases challenging, even for experienced pathologists. The development of computational tools to aid pathologists in the characterization of these lesions would have great diagnostic and clinical value. As a first step to address this issue, we evaluated the ability of computational image analysis to accurately classify DCIS and UDH and to stratify nuclear grade within DCIS. Using 116 breast biopsies diagnosed as DCIS or UDH from the Massachusetts General Hospital (MGH), we developed a computational method to extract 392 features corresponding to the mean and standard deviation in nuclear size and shape, intensity, and texture across 8 color channels. We used L1-regularized logistic regression to build classification models to discriminate DCIS from UDH. The top-performing model contained 22 active features and achieved an AUC of 0.95 in cross-validation on the MGH dataset. We applied this model to an external validation set of 51

Download PDF

Print Share

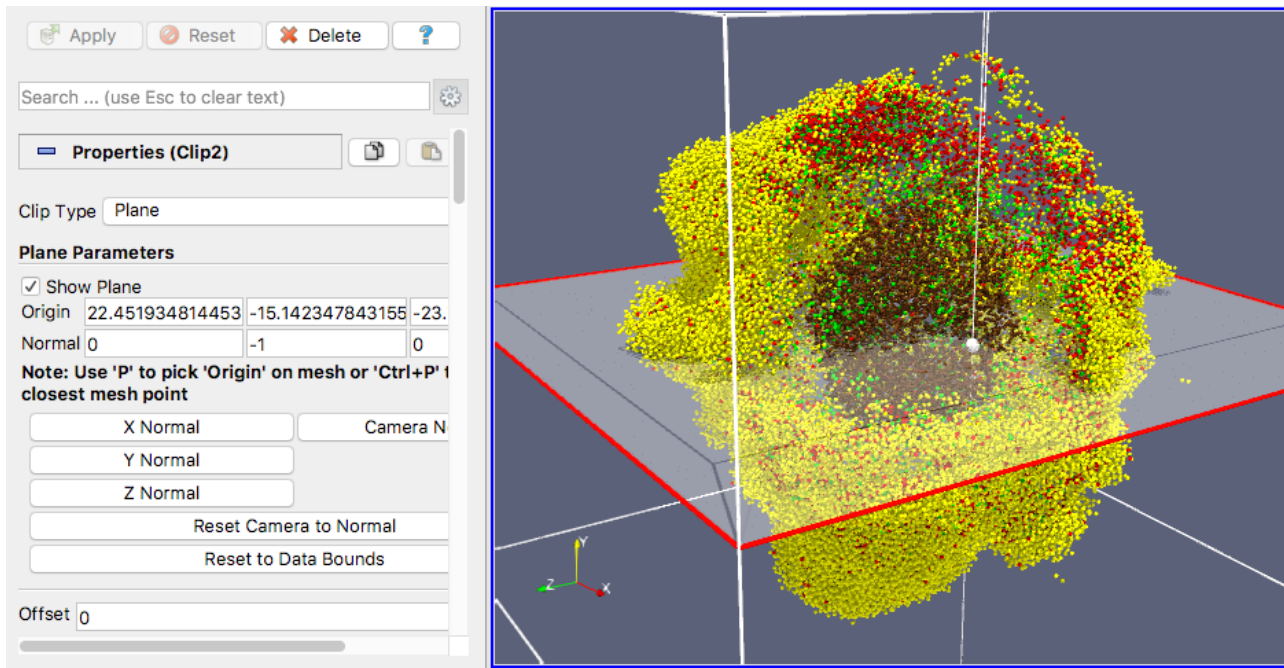
Reddit
Google+
StumbleUpon
Facebook
LinkedIn
CiteULike
Mendeley
PubChase
Twitter
Email

Digital snapshots allow semantic, multiscale annotation for image-based data

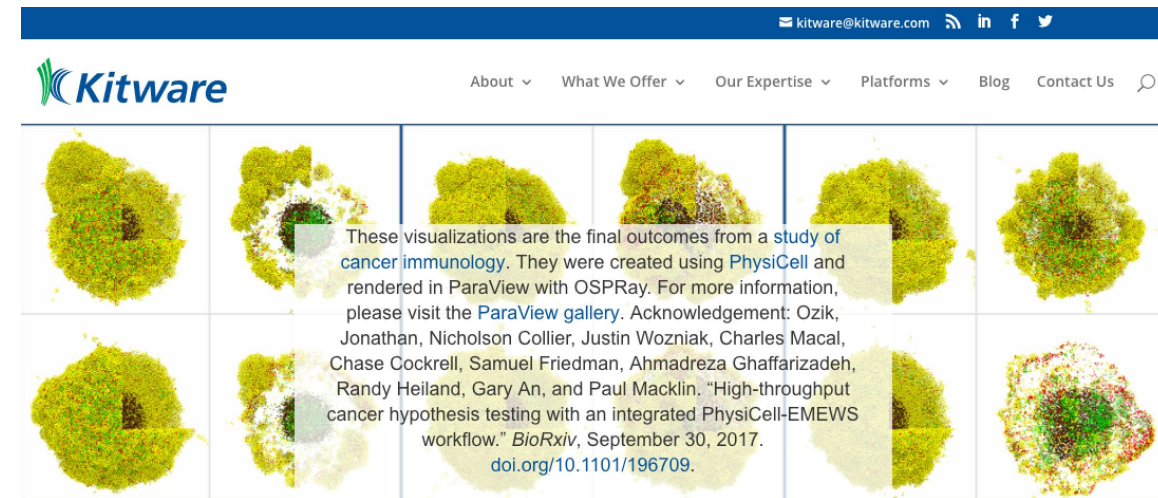
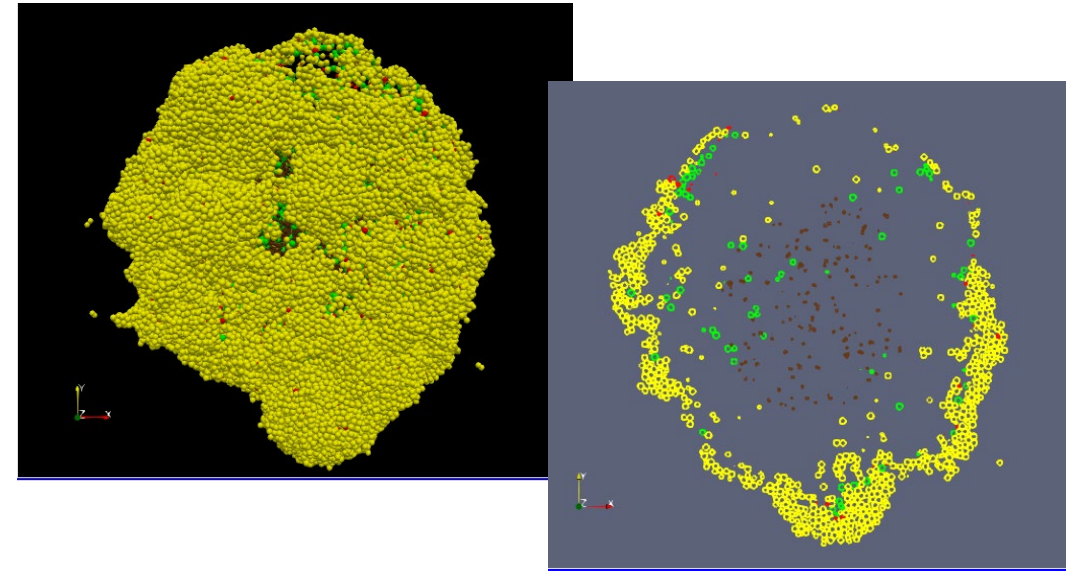


```
><cellular_information>  
>  ><cell_populations>  
>    ><cell_population type="individual">  
>      ><custom>  
>        ><simplified_data type="matlab" source="BioFVM">  
>          ><filename>output00000540_cells.mat</filename>  
>        </simplified_data>  
>        ><simplified_data type="matlab" source="PhysiCell">  
>          ><labels>  
>            ><label index="0" size="1">ID</label>  
>            ><label index="1" size="3">position</label>  
>            ><label index="4" size="1">total_volume</label>  
>            ><label index="5" size="1">cell_type</label>  
>            ><label index="6" size="1">cycle_model</label>  
>            ><label index="7" size="1">current_phase</label>  
>            ><label index="8" size="1">elapsed_time_in_phase</label>  
>            ><label index="9" size="1">nuclear_volume</label>  
>            ><label index="10" size="1">cytoplasmic_volume</label>  
>            ><label index="11" size="1">fluid_fraction</label>  
>            ><label index="12" size="1">calcified_fraction</label>  
>            ><label index="13" size="3">orientation</label>  
>            ><label index="16" size="1">polarity</label>  
>            ><label index="17" size="1">migration_speed</label>  
>            ><label index="18" size="3">motility_vector</label>  
>            ><label index="21" size="1">migration_bias</label>  
>            ><label index="22" size="3">motility_bias_direction</label>  
>            ><label index="25" size="1">persistence_time</label>  
>            ><label index="26" size="1">motility_reserved</label>  
>            ><label index="27" size="1">oncoprotein</label>  
>            ><label index="28" size="1">elastic_coefficient</label>  
>            ><label index="29" size="1">kill_rate</label>  
>            ><label index="30" size="1">attachment_lifetime</label>  
>            ><label index="31" size="1">attachment_rate</label>  
>          </labels>  
>        </simplified_data>  
>      </custom>  
>    </cell_population>  
>  </cell_populations>  
></cellular_information>
```

Build on existing, open source tools: ParaView

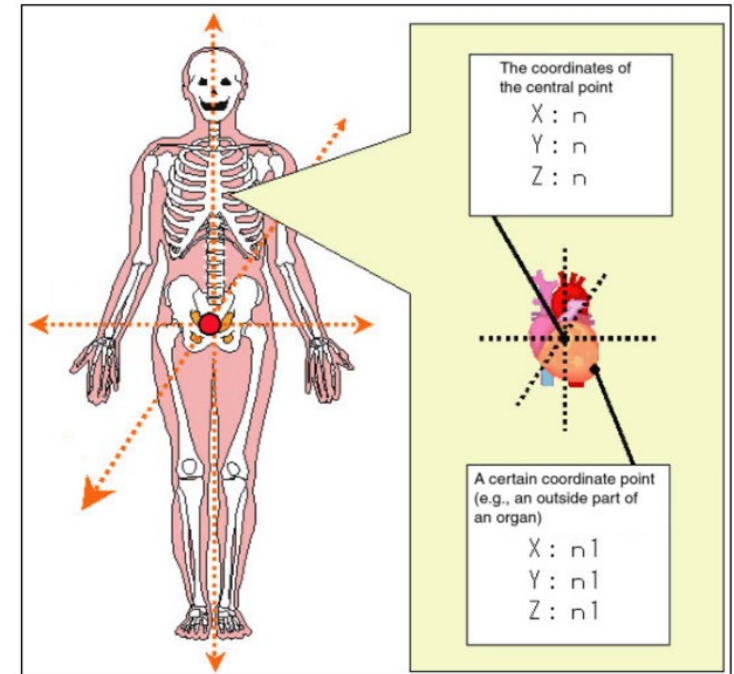


Interactive viewing, clipping, slicing, animation



Initial 9 Months

- Highly modular design,
 - Using existing APIs, and
 - Aligned with existing standards.
-
- Collaborate with the best teams.
 - Document well to speed up adoption.
 - Code will be released as open source under the commercially-compatible 3-clause BSD or MIT license.
-
- New visualizations will be taught in IVMOOC.cns.iu that students from 100 countries take each Spring.



Open Questions

- What things do existing ontologies and user interfaces well (e.g., organs, 3D rendering) and NOT do well (e.g., uncertainty, variability).
- Should the MC teams focus on whole body, only organs (e.g., those that TMCs focus on), or one organ?
- 2D or 3D? Most datasets are 2D but human body is 3D.
- Single CCF or multiple, e.g., male/female? Unique opportunity to compute, visualize, and understand variations across individuals.
- What datasets will become available when? How to avoid that first organ-specific data that becomes available impacting the development a general CCF?
- Many more ...