Sep 28, 2021

New $15 million NSF grant launches Ohio State Imageomics Institute

TDAI faculty lead new field of study based on images of living organisms

The Ohio State University has been awarded a $15 million grant from the National Science Foundation to lead the creation of a new, interdisciplinary institute and establish a new field of study that has the potential to transform biomedical, agricultural and basic biological sciences.

The new entity, which will be called the Imageomics Institute, is an inaugural **institute for data-intensive discovery in science and engineering < https://www.nsf.gov/pubs/2021/nsf21519/nsf21519.htm>** created by the NSF as part of its **Harnessing the Data Revolution < https://www.nsf.gov/cise/harnessingdata/>** initiative. As such, the Imageomics Institute will be part of forming a national collaborative research network dedicated to computation-enabled discovery.

Led by faculty from Ohio State's **Translational Data Analytics Institute < https://tdai.osu.edu/>** , the Imageomics Institute will create a new field of study in which scientists use images of living organisms as the basis for understanding biological processes of life on Earth. This new approach, called imageomics, will utilize machine learning methodologies to extract from images biological traits such as the behavior or physical appearance of an individual, or even the distinguishing skeletal structure of a species.

Imageomics will leverage and extend the paradigm of knowledge-guided machine learning by structuring the algorithmic underpinnings around biological knowledge and continually guiding and refining them using new

Tanya Berger-Wolf

biological knowledge as it is generated.

Similar to genomics before it, which applied computation to the study of the human genome, imageomics will leverage computer science to help scientists extract meaning from an otherwise unwieldy amount of natural image data. Images to be studied include digital collections from museums, labs and other institutions, as well as photos taken by scientists in the field, camera traps, drones and even members of the public who have uploaded their images to platforms such as eBird, iNaturalist and Wildbook.

The ability to use image data for scientific discovery will get scientists closer to answering compelling questions such as: Are a baby zebra's stripes similar to its mother's, and what might the genetic mechanism be for inheriting the pattern? How do the skulls of bat species vary with environmental conditions, and what evolutionary adaptation drives that change? What subtle morphological features distinguish closely related species of fish from each other and why?

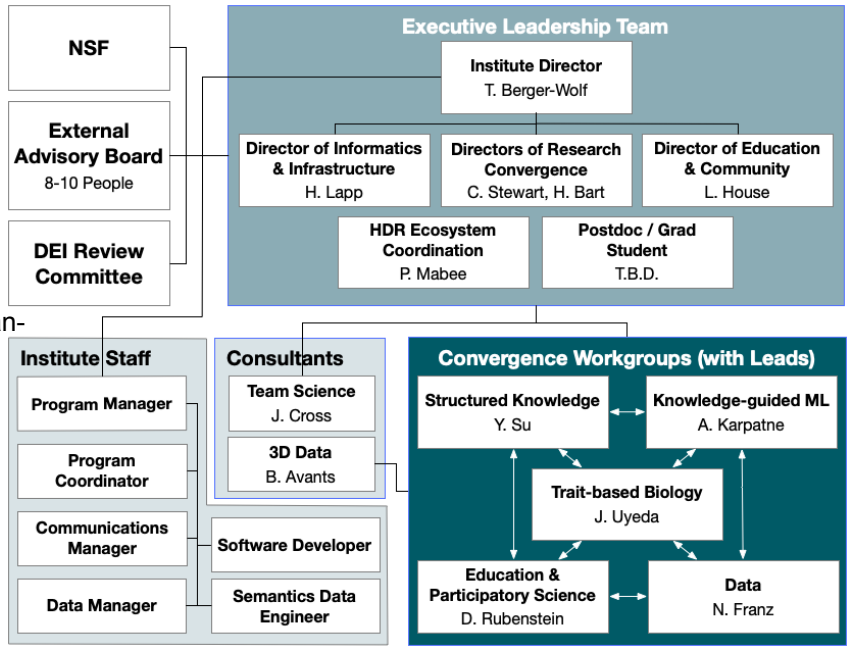
"So many pictures of organisms exist with scientific value that we've been unable to analyze at the right scale," said Tanya Berger-Wolf, faculty director of the Translational Data Analytics Institute and principal investigator for the Imageomics Institute. "This will change how we are able to see and understand the natural world. Computers help humans 'see' images differently and expose what we may otherwise miss."

In addition to Berger-Wolf, a professor of computer science and engineering, electrical and computer engineering, and ecology, evolution and organismal biology at Ohio State, the project team includes biologist and computer scientist co-investigators from Tulane University, Virginia Tech, Duke University and Rensselaer Polytechnic Institute; senior personnel from Ohio State, Virginia Tech and six additional institutions; and collaborators from more than 30 universities and organizations around the world.

“It’s exciting to be part of this team of scientists venturing into a new space between their respective disciplines, toward something completely different,” Berger-Wolf said. “It is everything that is great about interdisciplinary collaboration.”

Management and Coordination Plan

The *Imageomics* Institute team combines deep complementary expertise from different fields of biology (biodiversity, ecology, evolution, development, genomics), computer science (ML, computer vision, knowledge computing, software engineering), information science (data and metadata standards), and community engagement, spanning eleven institutions. To coordinate and manage expertise, programs, and activities across institutions and team members, the Institute will be overseen by a cross-disciplinary and cross-institutional Executive Leadership Team, which works with Institute staff, consultants, and Convergence Workgroups.



Executive Leadership Team (ELT)

The ELT ensures that priorities, programs, and activities promote the Institute's mission and goals. It also answers to NSF, and, in conjunction with annual retreats, meets with the External Advisory Board (comprised of 8-10 experts in the fields represented by the Institute) to review progress, accomplishments, and strategic plans. The ELT is comprised as follows.

Tanya Berger-Wolf (OSU), PI, will serve as Institute Director, and will provide the overall leadership and coordination, setting the strategic priorities. Computational Ecologist, OSU TDAI Director, Wild Me Director.

Henry Bart, Jr. (Tulane), Co-PI, and **Charles Stewart** (RPI), Co-PI, will serve as Co-Directors for Research Convergence, responsible for connecting, coordinating, and prioritizing data and research convergence efforts across the Institute. Bart is a fish systematist and biodiversity information specialist with extensive research experience using fish specimen media (2D, 3D, radiograghs, CT scans, videos) and metadata. Stewart is a computer scientist with expertise in CV applied to ecology and environmental conservation as well as start-up and large-scale project leadership. **Hilmar Lapp** (Duke), Co-PI, will serve as Director of Informatics & Infrastructure, overseeing software research and development for the Institute's information and data infrastructure, dissemination, and data standards. Lapp has extensive experience with structured knowledge technologies in biology, is a veteran in collaborative open-source software development, and has led a variety of community cyberinfrastructure building programs. **Leanna House** (VT), Sr. Personnel, will serve as Director of Education & Community, coordinating education, outreach, and community efforts across all participating institutions. She brings expertise in Bayesian statistical modeling with an emphasis in visualization, uncertainty analyses, human-computer interaction, and education. **Paula Mabee** (NEON), Sr. Personnel, will ensure the Institute coordinates effectively with the HDR ecosystem at all levels. She is an evolutionary and developmental biologist, Chief Scientist and Observatory Director for NEON, and has extensive experience coordinating activities and people across distributed research projects and networks.

Convergence Workgroups (CWGs)

CWGs will form one of the Institute's key mechanisms to bring about cross-institutional and cross-disciplinary integration. Each CWG is led by Key Personnel, who will recruit participants from the Institute's research groups, leadership, consultants, staff, and from the respective research community. CWGs coordinate regularly with each other, and inform as well as coordinate with the ELT on priorities, needs, and outcomes. The CWGs and their leads are as follows. *Knowledge-guided ML* (KGML) will be led by **Anuj Karpatne** (VT), Co-PI. Karpatne is a computer scientist, member of the Sanghani Center for AI and Data Analytics, and brings expertise in the area of data mining, ML, and KGML, a field he is pioneering. *Structured Knowledge* will be led by **Yu Su** (OSU), Sr. Personnel. Su is Assistant Professor in Computer Science and TDAI

affiliate. He brings expertise in NLP, ML, and knowledgebases. *Trait-based Biology* will be led by **Josef Uyeda** (VT), Sr. Personnel. Uyeda is Assistant Professor in Biological Sciences, and brings expertise in modeling trait evolution on phylogenies, systematics, and their integration with ontologies. *Data* will be led by **Nico Franz** (ASU), Sr. Personnel. Franz is Director of the Biodiversity Knowledge Integration Center, and brings expertise in systematics and biodiversity big data. *Education and Participatory Science* will be led by **Daniel Rubenstein** (Princeton), Sr. Personnel, Professor in Zoology, Ecology and Evolutionary Biology. Rubenstein brings expertise in behavioral ecology and conservation.

Other Senior Personnel and Consultants

Murat Maga (UW-SCRI), Sr. Personnel: Associate Professor, Pediatrics. Dr. Maga brings expertise in evolutionary developmental biology, craniofacial biology, 3D imaging, computational anatomy and statistical shape analyses. He will participate in the Data CWG, and in research and outreach activities that relate to 3D datasets, particularly of model organisms. His lab will oversee deployment of the portal for community annotation of 3D datasets using the Institute's cyberinfrastructure.

Yasin Bakiş (Tulane), Sr. Personnel: Sr. Manager of Biodiversity Informatics and Data Engineering. He brings expertise in biodiversity informatics. He will participate in the Data CWG, and in research with fish specimen images (2D, 3D, radiograghs, CT scans, videos) and image metadata.

James Balhoff (RENCI), Sr. Personnel: Sr. Research Scientist. Balhoff brings expertise in the integration of bio-ontologies and corresponding software services. He will lead extending the Phenoscape KB for accessing and integrating ontologies and other structured knowledge, and will be participating in the Structured Knowledge CWG.

Bryan Carstens (OSU), Sr. Personnel: Professor, Ecology, Evolution and Organismal Biology. He brings expertise in molecular ecology, systematics, and ML analyses of biodiversity data. He will participate in the Trait-based Biology CWG, and the Carstens lab will participate in extraction and analysis of trait data at the species level, particularly the application of these data to species delimitation.

Wei-Lun Chao (OSU), Sr. Personnel: Chao is Assistant Professor in Computer Science and OSU TDAI. He brings expertise in CV and ML. He will participate in the Knowledge-guided ML CWG, and in developing KGML, especially data- and sample efficient ML, transfer learning and novelty detection.

Wasila Dahdul (UCI), Sr. Personnel: Data Curation Librarian. Dahdul brings expertise in data curation and ontologies. She will participate in the Trait-based Biology CWG, supply as well as update the needed source ontologies and phylogenies to guide ML, and assess new *Imageomics* traits.

Jennifer Cross, Consultant, is Director of the Institute for Research in the Social Sciences (IRISS) at CSU. She will lead the Team Science and Team Integration activities for the Institute (see below).

Brian Avants (Invicro Solutions), Consultant, is a computer scientist with specialization in biomedical image analysis, and chief software architect of Advance Normalization Tools (ANTs), a high-performing open-source image analysis ecosystem that will be used for processing 3D datasets. He will provide his expertise to relevant workgroups and research teams regarding computational optimizations for large datasets.

Managing Team Science and Collaboration

To facilitate and train the Institute's Key Personnel in effective Team Science practices, we will engage the Team Science Consulting services of IRISS (Consultant Dr. Cross). IRISS provides a full range of activities to support planning, ideation, convergence, facilitation, training, growth, and evaluation of team science [136]. Funding for this consultation service is included in the budget.

Cross-organizational collaboration will be facilitated via the following activities. Convergence Workgroup teams will meet monthly to brainstorm and discuss results and papers, providing opportunities for postdocs and graduate students to lead meetings, and practice presenting and discussing research results. An Institute Seminar Series will be jointly run with partners and collaborators and will include lectures open to the scientific public. Institute All-Hands Retreats will be organized annually (OSU yr 1, then sites TBD) to set priorities, plan programs, and review progress versus milestones. The retreats will also serve to promote cohesiveness, networking, collaboration, and professional development among Institute members. Funding for the retreats, including organization and travel, is in the budget.

Members of the team, including the ELT, have long-standing experience with distributed collaborative projects that conduct almost all of their work and communication virtually, using technologies such as Zoom, Slack, Github, Google Docs, Overleaf, etc. This proposal itself is a convergence achieved entirely online.

Timeline, Milestones, and Metrics of Success

The impact of novel synthetic and convergent research is complex and often requires a long-term horizon [113,225,242]. With this in mind, our shorter term metrics of success include community growth and diversity of backgrounds; diversity of participation in *Imageomics* events; new collaborations formed, publications with new co-authorships, and new proposals resulting from *Imageomics*; the use of *Imageomics* as a concept; reuse of Institute tools and methods; and new discoveries made for structured knowledge and ML.

Convergence	Milestone	Y 1	Y 2	Y 3	Y 4	Y 5
Team	Institute structure and staff set up					
	Effective team science & evaluation					
	DEI evaluation plan					
	External advisory board					
Data	Morphological traits, phenotype ontologies					
	Relatedness information					
	Skeletal comb. with RGB, morphology, relatedness					
	Convergence of 3D, skeletal, shape, text, online data					
	Fish and Zebra data					
	Butterfly data					
	Bird data (2D image + text from eBird), Bat data					
	Full data convergence					
Research	Proof-of-concept results built on BGNN					
	Alignment of Data and Structured Knowledge					
	Imageomics scene graph					
	Defined research projects					
	First open ML competition					
	ML predicts a new testable bio-hypothesis					
	Automated inference of existing traits (for subset)					
	Inference of new trait definitions					
	Enriched image feature ontologies (ProPNet)					
	End-to-end workflow convergence ("model" organisms)					
	ML enriches structured biological knowledge					
	Full workflow convergence and automation					
Infrastructure	Curation and validation infrastructure					
	Annotation tools for 2D					
	API prototype for KGML, expanding Phenoscape					
	Annotation tools for 3D and text					
	APIs for existing online platforms					
	Published ML models, datasets, tools and metadata					
	Imageomics tool deployment to bio-community					
	Prototype API for "Images in, traits out" (for subset)					
	Full API (certain animal taxa) and tool deployment					
Education	Strategic education plan with partners					
	Summer data camp					
	Princeton teacher training					
	HDR curricular materials developed and released					
	Field interdisciplinary science course					
Community	Imageomics.org website, social media, communication					
	HDR Ecosystem integration					
	Tutorials, demos, workshops (virtual, in person)					
	Citizen Science Events (GGR - 2024)					
	Imageomics Forum					
	Catalyzed spin-off projects					

Research, Education, Broadening Participation, and Knowledge Transfer

Integration of the efforts to research, educate, transfer knowledge, and broaden participation in STEM is *essential* for the Institute's success, and we designed its structure to facilitate and promote this, including multi-institutional DEI Review Committee. Human-machine partnerships researched within the context of *Imageomics* rely directly on the convergence of ideas and human input, which can only be accomplished with the participation of individuals with diverse expertise, race, gender, culture, age, etc. Participants are recruited from interdisciplinary collaborations, formal and informal education endeavours, and outreach and citizen science events, all of which are opportunities for knowledge transfer. In promoting the democratization of *Imageomics* science, we make STEM accessible to engage people with diverse backgrounds and interests, transfer *Imageomics* ideas, use research outcomes to enable conservation, and collect data for research, closing the cycle.

HDR Institute: Imageomics: A new frontier of biological information powered by knowledge-guided machine learning

The traits that characterize living organisms—in particular, their morphology, physiology, behavior and genetic make-up—enable them to cope with forces of the physical as well as the biological and social environments that impinge on them. Moreover, since function follows form, traits provide the raw material upon which natural selection operates, thus shaping evolutionary trajectories and the history of life. Interestingly, most living organisms, from microscopic microbes to charismatic megafauna, reveal themselves visually and are routinely captured in copious images taken by humans from all walks of life. The resulting massive amount of image data has the potential to further our understanding of how multifaceted traits of organisms shape the behavior of individuals, collectives, populations, and the ecological communities they live in, as well as the evolutionary trajectories of the species they comprise. Images are increasingly the currency for documenting the details of life on the planet, and yet traits of organisms, known or novel, cannot be readily extracted from them. Just like with genomic data two decades ago, our ability to collect data at the moment far outstrips our ability to extract biological insight from it. The Institute will establish a new field of IMAGEOMICS, in which biologists utilize machine learning algorithms (ML) to analyze vast stores of existing image data—especially publicly funded digital collections from national centers, field stations, museums and individual laboratories—to characterize patterns and gain novel insights on how function follows form in all areas of biology to expand our understanding of the rules of life on Earth and how it evolves.

This Institute will introduce structured knowledge from the biological sciences to guide and structure ML algorithms to enable biological trait discovery from images, establishing the field of **Imageomics**. With images captured and annotated by scientists and the public serving as the basis for the work, the Institute's convergent approach uses structured biological knowledge to provide scientifically validated inductive biases and rich supervision for ML, and ML will in turn enrich the body of biological knowledge. The resulting ML models and tools will help to make what was hidden visible, so that scientists from a wide range of biological communities can discover and infer the traits of organisms; assess shared similarities and differences between individuals, populations and species; and come to see the world in new ways. **Imageomics** will accelerate and transform the biomedical, agricultural and basic biological sciences as they seek to understand and control genes that relate to particular phenotypes and enable an overarching understanding of how the genome evolved in tandem with the organismal phenome. Because traits are the essential links between genes and the environment, using ML to help characterize them will lead to emergent understandings of how they function. Harnessing the insights that arise from these new visualizations will stimulate the use of new genetic technologies, such as CRISPER, and more nuanced ecological practices, such as modified land use schemes that emerge from better understanding the connections between individual decision-making within species and their impact on their population dynamics. With the emergence of new and better targeted practices that generate fewer unintended consequences, the new linkages resulting from a better understanding of traits and their consequences will bolster the nation's bioeconomy. In addition, by leveraging and expanding existing diverse, inclusive and intellectually wide-ranging collaborative networks, the Institute will also educate the next generation of scientists and engage the broader public in scientific inquiry and knowledge discovery so that **Imageomics** can transform and democratize science for public good.