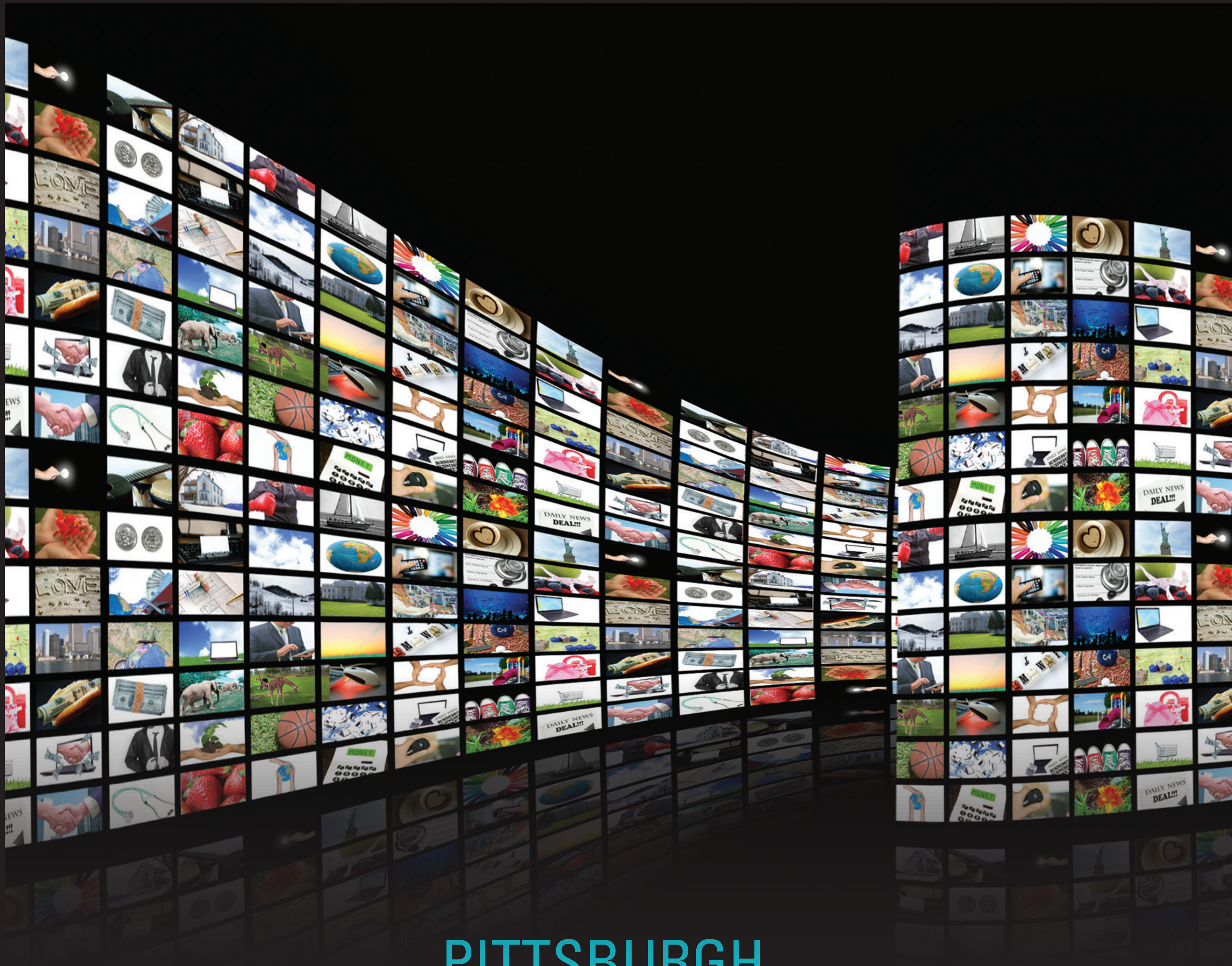
PITTSBURGH SUPERCOMPUTING CENTER
300 S. CRAIG STREET
PITTSBURGH, PENNSYLVANIA 15213

SPRING 2015



PITTSBURGH SUPERCOMPUTING CENTER

PEOPLE. SCIENCE. COLLABORATION.

PEOPLE. SCIENCE. COLLABORATION. | *SPRING 2015*

PSC.EDU

PITTSBURGH SUPERCOMPUTING CENTER provides university, government and industrial researchers with access to several of the most powerful systems for high-performance computing, communications and data storage and handling available to scientists and engineers nationwide for unclassified research. PSC advances the state of the art in high-performance computing, communications and data analytics and offers a flexible environment for solving the largest and most challenging problems in computational science. As a leading partner in XSEDE, the Extreme Science and Engineering Discovery Environment, the National Science Foundation's cyberinfrastructure program, PSC works with other XSEDE participants to harness the full range of information technologies to enable discovery in U.S. science and engineering.

FROM THE DIRECTORS

Ralph Roskies (left) and Michael Levine, PSC co-scientific directors

Welcome to the Spring 2015 edition of *People. Science. Collaboration.*

Many of you will have already heard the news: PSC will be fielding a new, powerful and novel computing and data processing system in support of the national science and engineering research effort (p. 14). With a planned production date in January 2016, the $9.65-million, NSF-funded *Bridges* will provide badly needed capabilities in the many existing and emerging fields of data-driven science. Bridges will provide resources to researchers, some practiced in the use of very high-performance computing, some using it for the first time, who are tackling some of the most profound computational challenges of the 21st century: extracting useful knowledge from the vast array of data now being generated by instruments, research initiatives and, through the Internet, individuals.

Besides fielding large resources, PSC also sustains a vigorous research program. Our Public Health Applications Group and its collaborators continue to blaze new trails in the development of lifesaving tools. Following a resounding success in transforming the vaccine delivery system of the West African Republic of Benin, they have begun intensive work with the Ministry of Health in the Republic of Mozambique, East Africa, to improve that nation's vaccine delivery and anti-malaria programs (p. 18). Our Advanced Networking Group continues to create new tools for improving electronic transfer of Big Data, with the Web10G tool for assessing problems with data connections taking its first step toward being included in major computer operating systems (p. 22). And our ongoing Data Exacell (DXC) Project, incorporating both our Blacklight and Sherlock supercomputers, is working with many, often new, users to develop novel techniques for storing, moving and analyzing vast datasets (p. 22).

We continue to support researchers in Pennsylvania and nationwide to do more intensive, detailed investigations that enable unexpected scientific insights. Work on the D.E. Shaw Research Anton supercomputer hosted at PSC continues to generate surprising results that illuminate just how dynamic and complex biomolecules can be. A study on the voltage-gated sodium channel protein has revealed the crucial differences by which local anesthetics and anti-epilepsy drugs interact with that protein, producing vastly different medical effects (p. 10).

In addition to serving as part of the DXC, Blacklight continues to enable unique research by virtue of its very large memory and large processor count. Blacklight has made possible a study of ancient climate patterns that may help explain how humanity survived a devastating population crash between 100,000 and 300,000 years ago (p. 4). Blacklight is also helping researchers to improve donor organ/recipient matching to save more lives (p. 8). It is helping to illuminate the process by which weeds became food sources for humanity, giving hints of how to improve disease resistance in wheat plants (p. 20). And it has proved invaluable to researchers trying to help users find video clips on the Web relevant to search terms without the labor-intensive step of humans first identifying what each video portrays (p. 16).

Our relationship with the National Science Foundation—and particularly its national network of supercomputing sites, XSEDE—continues to play a part in these advances, particularly through services like the Extended Collaborative Support Service for XSEDE supercomputing users, in which our staff continue to play a leading role.

Funding from NSF, from the National Institutes of Health, and non-governmental funding agencies such as the Bill and Melinda Gates Foundation, the Global Fund, and UNICEF all continue to fuel this success. In addition, support from the Commonwealth of Pennsylvania and funders such as the Buhl Foundation (p. 23) help us continue to carry out outreach to educate students from K-12 through graduate school and to local businesses seeking to better manage and use data.

We hope you find the following accounts of our efforts over the past 6 months enjoyable and edifying. We would like any feedback you have on this publication. Please see the inside back cover for instructions on participating in our online survey.

Michael Levine   Ralph Roskies

CONTENTS

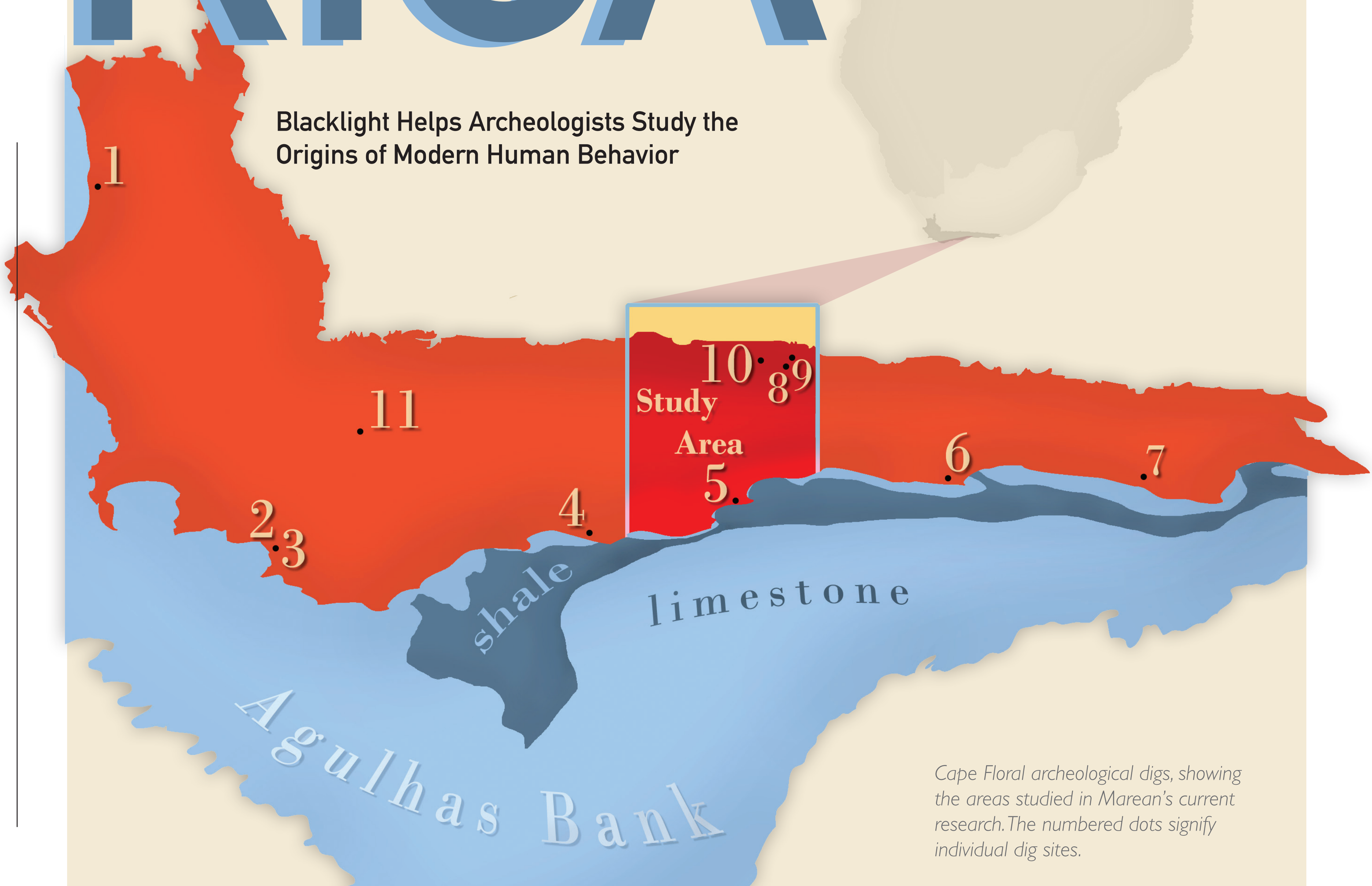
LIFEBOAT AFRICA

Blacklight Helps Archeologists Study the Origins of Modern Human Behavior

Picture an endangered species. Climate change has forced it into a small refuge. The population has crashed to thousands or fewer. Survival is by no means ensured.

This grim scenario may have happened. To us.

More amazing, not only did the human species survive its great population crash, we emerged from it with unparalleled, complex behaviors. This change allowed us to adapt our surroundings to suit us rather than the other way around. It made us "modern," able to live in virtually every environment on the Earth.



Cape Floral archeological digs, showing the areas studied in Marean's current research. The numbered dots signify individual dig sites.

"Humans cooperate with non-kin at spectacular levels of complexity," says Curtis Marean of Arizona State University. "So what we want to know is what are the contexts of evolution for those special features of humans? When did they arise, and why did they arise?"

To answer these questions, Marean and colleagues at Arizona State and in South Africa, Australia, Israel and France, have been studying the archeology of the Cape Floral Region of South Africa. They believe this region may be where the human species emerged from near-extinction to global domination.

EVIDENCE IN OUR GENES, IN THE GROUND

Judging from the interrelatedness of every human being, everyone on Earth today may descend from 15,000 or even fewer survivors of a great population crash. Based on genetic mutation rates, this crash happened an estimated 100,000 to 300,000 years ago.

During the last glacial maximum, Africa experienced a long dry phase that could explain why the human population crashed. At that time, the Cape Floral Region of South Africa is one of the few on the continent that shows any evidence of human habitation. Also at this time, people began displaying a number of behaviors that separate "modern" from "early" humans. They began to carry out complex thermal treatments of stone to make better tools. They started to engrave geometric patterns in ochre and bone.

"It seems very, very likely that the modern human lineage evolved … during a glacial stage when Africa was mostly dry and uninhabitable," Marean says. "We not only have a big brain, it's wired in a way that allows us to think in complex analogies, plan for the future, understand mathematics. We know that proclivity is imbedded in our genes. If we're trying to understand the process of that event, which leads to all modern humans, we need to understand the environment at that time."

BLACKLIGHT: INHERENT FLEXIBILITY

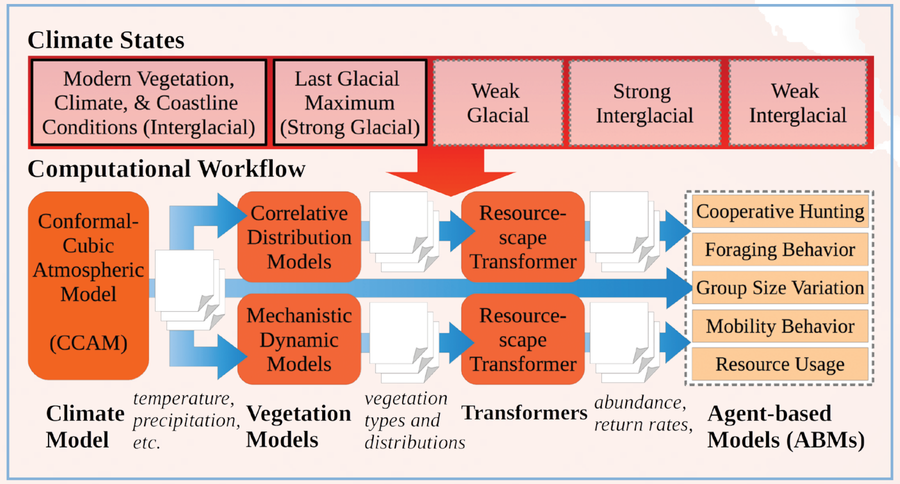
One archeological finding in particular caught the researchers' eyes. The Cape Floral Region is rich in shellfish, with a long, rich coastline, as well as tubers—plants with fleshy roots. Humans need a source of both protein and carbohydrate to survive, and this combination of foods would certainly do the trick. But were they there—and available enough to humans on foot—to sustain the human remnant at the time?

"Our project began as a straight archeological dig," says Marean. "We were interested in looking for evidence of modern human behavior, when it occurred, what types of behaviors are indicated, and so forth. And we were very successful at that."

"Then I realized that we needed much better climate and environmental contextual data to understand the archeological record we were excavating."

After a few years of working on commodity computers, the Arizona State researchers decided to "up the ante" by taking their models to PSC's Blacklight. An SGI Altix® UV 1000 shared-memory system, Blacklight offered a unique combination of large memory (up to 16 terabytes of memory available for a single job) and parallel processing (4,096 cores total) that served the requirements of two simulations that Marean and collaborators needed to run. The first was a "paleoclimate" model to see how currents and wind movements during the glacial maximum may have affected the local climate in the Cape Floral Region. The other will be an "agent based" model that simulates plants, animals and humans, making each grow, move and interact in a realistic way.

The researchers' final supercomputer simulations will incorporate climate, natural resources and human behaviors.


Climate States
Modern Vegetation, Climate, & Coastline Conditions (Interglacial)
Last Glacial Maximum (Strong Glacial)
Weak Glacial
Strong Interglacial
Weak Interglacial
Computational Workflow
Conformal-Cubic Atmospheric Model (CCAM)
Correlative Distribution Models
Mechanistic Dynamic Models
Resource-scape Transformer
Resource-scape Transformer
Cooperative Hunting
Foraging Behavior
Group Size Variation
Mobility Behavior
Resource Usage
Climate Model
temperature, precipitation, etc.
Vegetation Models
vegetation types and distributions
Transformers
abundance, return rates,
Agent-based Models (ABMs)

PSC staff helped Curtis Marean's work with the support of several programs within the NSF's XSEDE network of supercomputing centers: Extended Collaborative Support Service; Novel and Innovative Projects Program; Campus Champions Program.

Thanks to Blacklight and support from PSC staff, the paleoclimate model was the first to simulate the Cape region during that period at a level of detail sufficient to shed light on whether the region was warm and wet enough to sustain sufficient shellfish and tuber populations—a major accomplishment in the field. Even better, the initial results suggest that the climate would indeed have supported the food sources humans needed, at a time when virtually no place else in Africa did.

Next the scientists will run the agent-based model, and explore the role of this environment in the emergence of modern behaviors. The project promises enough predictive power for the models and the archeological evidence to be tested against each other, another first.

"We certainly have debates over exactly how small the modern human population was at this point," Marean says. "And other people have argued that the progenitor population was in North Africa, or the Maghreb area ... It's going to be a while before we can say one way or another; but I think right now the Cape Floral Region hypothesis is a strong one. And like all good hypotheses, it's generating an enormous amount of good science."

FutureMatch: Enabling Better Organ Exchange Programs

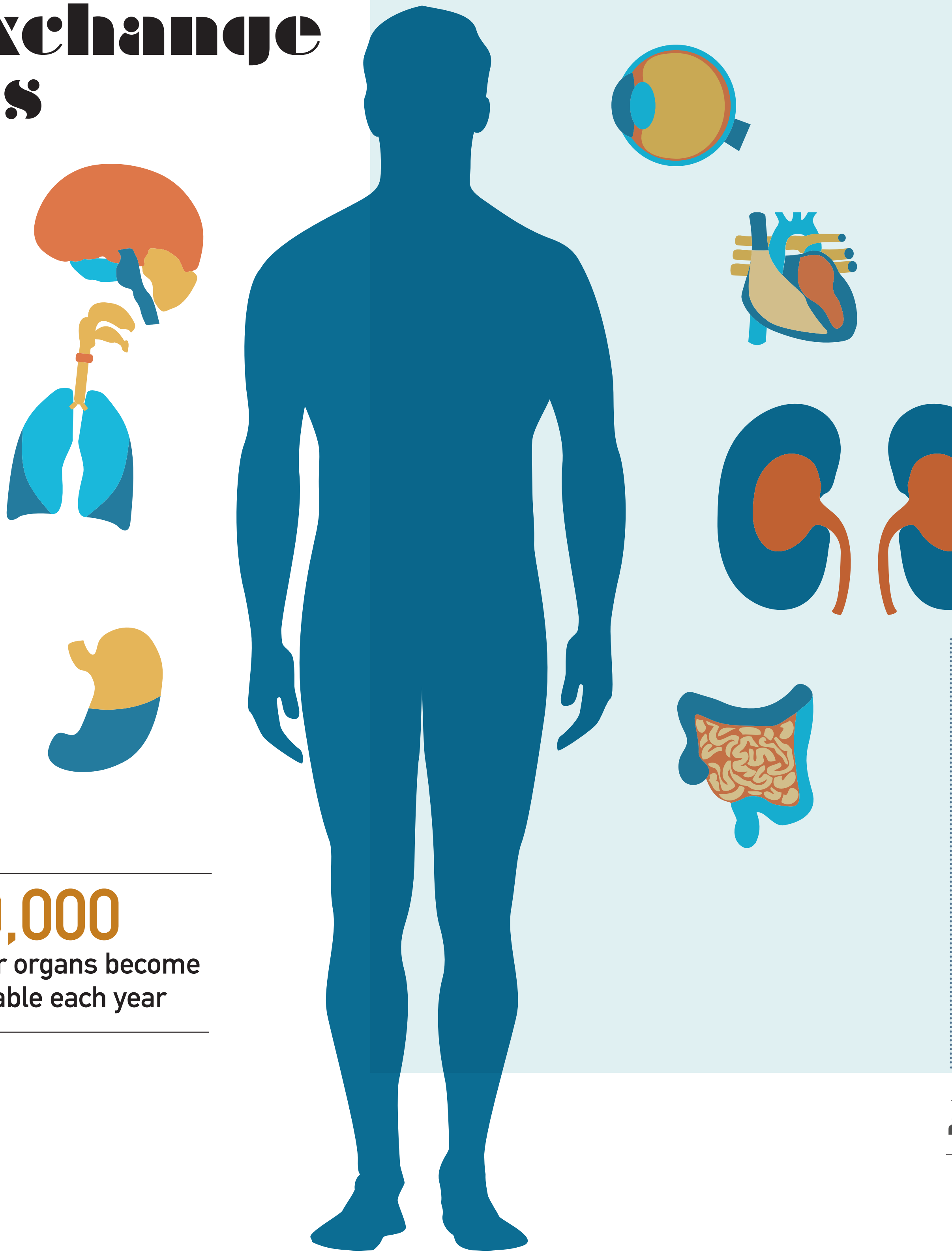
WHY IT'S IMPORTANT

Over 123,000 adults and children currently await organ transplants. Unfortunately, only about 30,000 donor organs become available each year. Every day, roughly 20 people die waiting for a "match." Donor cycles and chains, in which recipients and live donors who aren't compatible trade with other recipient/donor pairs, can alleviate donor-organ shortages. But as the number of donor/recipient pairs increases, it's difficult to calculate the best combination of trades. Tuomas Sandholm and his PhD students at Carnegie Mellon University have been working with the United Network for Organ Sharing (UNOS), the nonprofit organization that manages the national donor organ supply in the U.S. Their software automatically creates UNOS's kidney-paired donation transplant plans, optimizing organ matches at 142 transplant centers—about 60 percent of such facilities in the U.S.—twice a week.

123,000
adults and children currently await organ transplants

HOW BLACKLIGHT HELPED

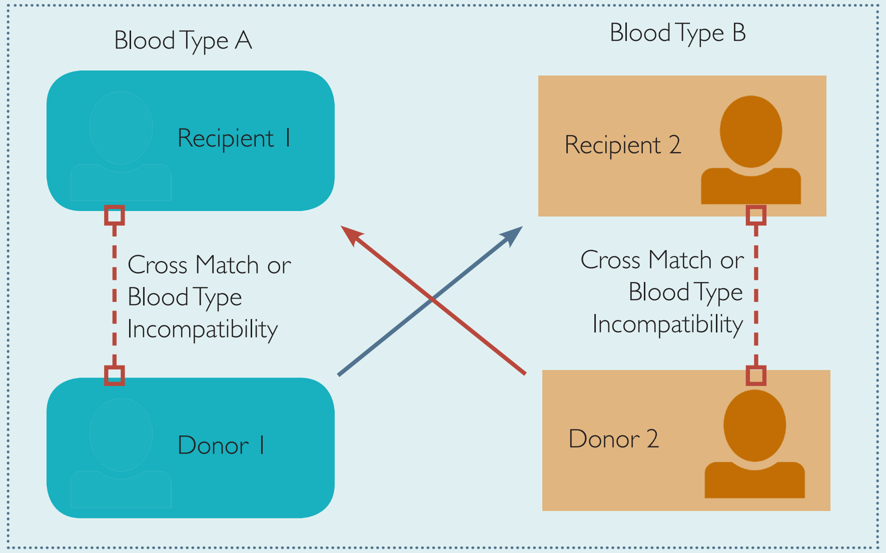
Calculating optimum donor exchanges requires both holding massive data in the computer's memory and many parallel computations. PSC's Blacklight supercomputer provided both of these capabilities, allowing the researchers to expand the number of donor/recipient pairs and broaden the criteria for matches so more hard-to-match patients could find donors. Their simulations on Blacklight suggested that both these goals could be achieved equitably, as defined by stakeholders. They also showed that a combined kidney and liver exchange could find more matches than separate exchanges. UNOS has already enacted many of their improvements.



30,000
donor organs become available each year

> *"You have a decision variable for each possible donor/recipient chain, and each possible cycle, a huge number of decision variables ... So if you have more memory, as we did with Blacklight, you can just bring in more variables. This enabled us to run a massive number of simulations within our new learning algorithms that learn to match better in a dynamic kidney exchange setting."*
>
> *—Tuomas Sandholm, Carnegie Mellon University*

Diagram of an exchange between otherwise incompatible pairs.


Blood Type A
Blood Type B
Recipient 1
Recipient 2
Cross Match or Blood Type Incompatibility
Cross Match or Blood Type Incompatibility
Donor 1
Donor 2

PSC staff helped Tuomas Sandholm's work with the support of programs within the NSF's XSEDE network of supercomputing centers: Extended Collaborative Support Service; Novel and Innovative Projects Program.

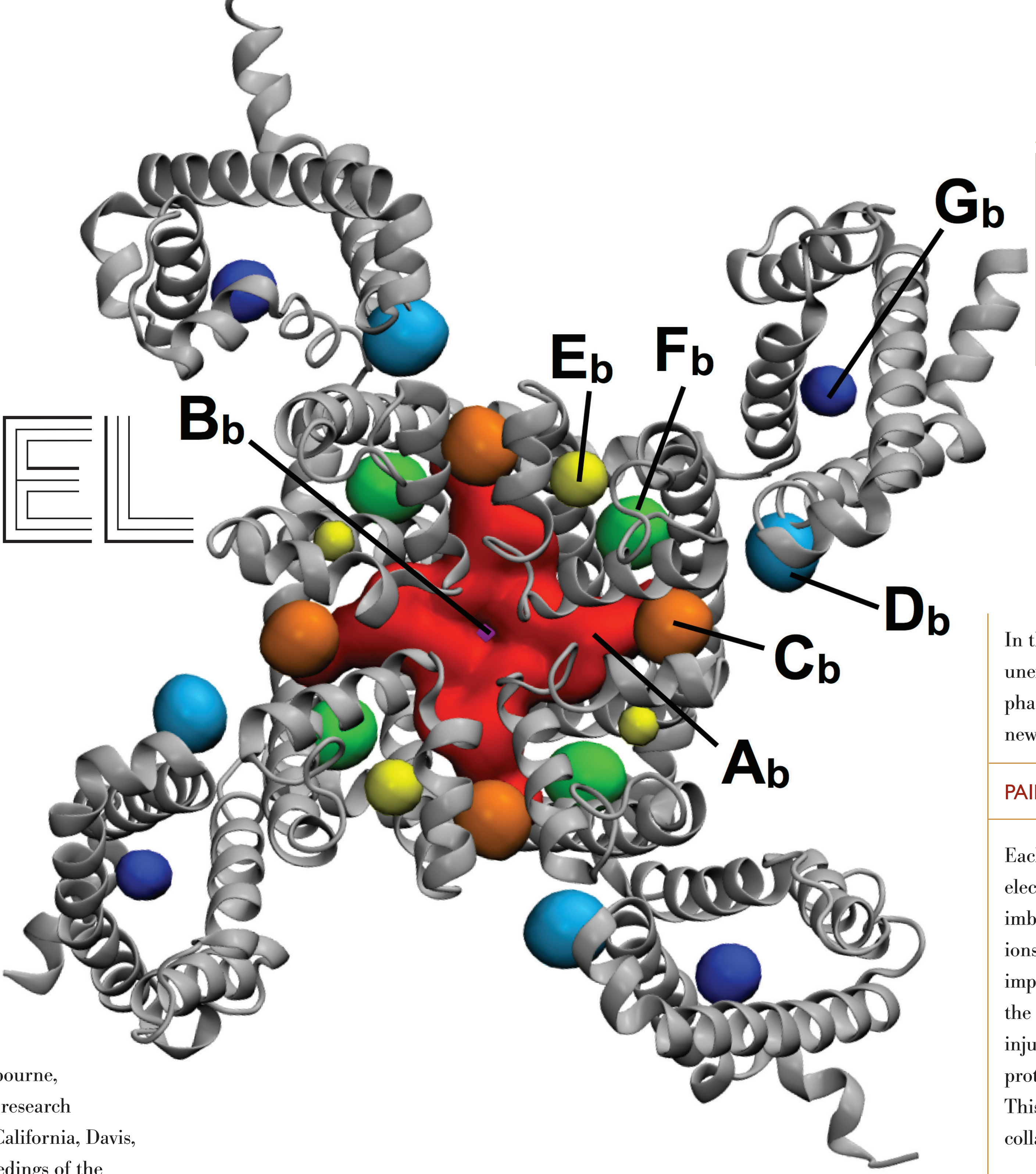
JANUS CHANNEL

Anton Simulations Reveal How Pain, Epilepsy Drugs Work through Same Target Protein

If you tried to imagine two neurological conditions as different from each other as possible, epilepsy and non-headache pain would be a good choice. Epilepsy develops when nerve cells in the brain start to fire in a too-regular pattern, overwhelming normal brain activity. Pain is a matter of sensation, when damage to tissues or other causes activate pain-sensitive nerves in the skin or other peripheral parts of the body, far from the brain.

Astonishingly, researchers recently discovered that drugs for these two very different maladies work by affecting the same nerve cell protein.

The mystery of how such vastly different conditions could share a similar therapeutic lynchpin has recently been solved, thanks to molecular simulations using the Anton supercomputer at PSC. Cèline Boiteux of RMIT University in Melbourne, Australia, and Toby Allen's research group at the University of California, Davis, have reported in the Proceedings of the National Academy of Sciences, USA, how they used the Anton supercomputer at PSC to identify two very different mechanisms by which the painkiller benzocaine and the anti-epileptic drug phenytoin affect the voltage-gated sodium channel protein.



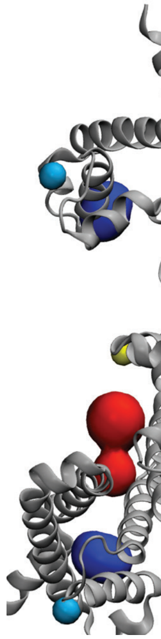
Topical painkiller benzocaine (colored blobs) sticks to the sodium channel protein (gray, looking down through the channel through the cell membrane) in a series of spots (Gb through Ab), finally settling into a number of positions in the pore of the channel, blocking any sodium passage (Ab).

In the process, the researchers have uncovered unexpected new molecular targets that pharmaceutical researchers may use to develop new pain-and-seizure control medications.

PAIN, EPILEPSY: A COMMON THREAD?

Each nerve cell is like a little battery, storing electro-chemical energy by maintaining an imbalance of charged sodium and potassium ions outside and inside the cell. A nerve cell impulse happens when a trigger event—in the case of a pain-sensing neuron, a physical injury for example—causes "gated channel" proteins in the cell membrane to open up. This lets the ions flow across the membrane, collapsing the ion imbalance.

Gated channels both cause and react to this collapse of the ion imbalance. Unaffected channels nearby those that have activated open in response. This causes their neighbors to open as well, the cascade spreading the signal.



Pain and epilepsy are very different in many ways but share a common feature. Both require neural over-activity to happen—and so we could potentially stop both by targeting the sodium channel that underlies that activity.

"The idea is that if you stop the gated sodium channels from working, you stop the propagation of the pain message," says Boiteux, first author of the study. "For an anti-epileptic agent it's a bit more subtle...We want to slow down the affected neurons. And one way to do that is to stop some of the ion channels from working."

THE VALUE OF TIME: ENTER ANTON

The researchers believed that Anton, a supercomputer developed by D. E. Shaw Research and hosted at PSC, could provide a window into that difference. Anton is hardwired for long-timescale molecular dynamics simulations and so is ideal for this task.

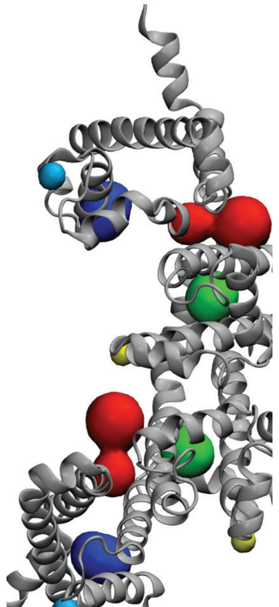
"We realized that one of the big issues with other computing systems was that on the typical time scale that we could access—in the range of 100 nanoseconds—we had to make a lot of assumptions about where the drug was going to bind to the channel," says Allen, the principal investigator of the study. "We knew by experiment that some binding sites play a key role in drug binding," but starting with the drugs in or near the known critical sites might prevent the researchers from discovering unknown critical sites. With an Anton allocation, by comparison, the researchers could simulate large biomolecules over time frames 10 to 100 times longer.

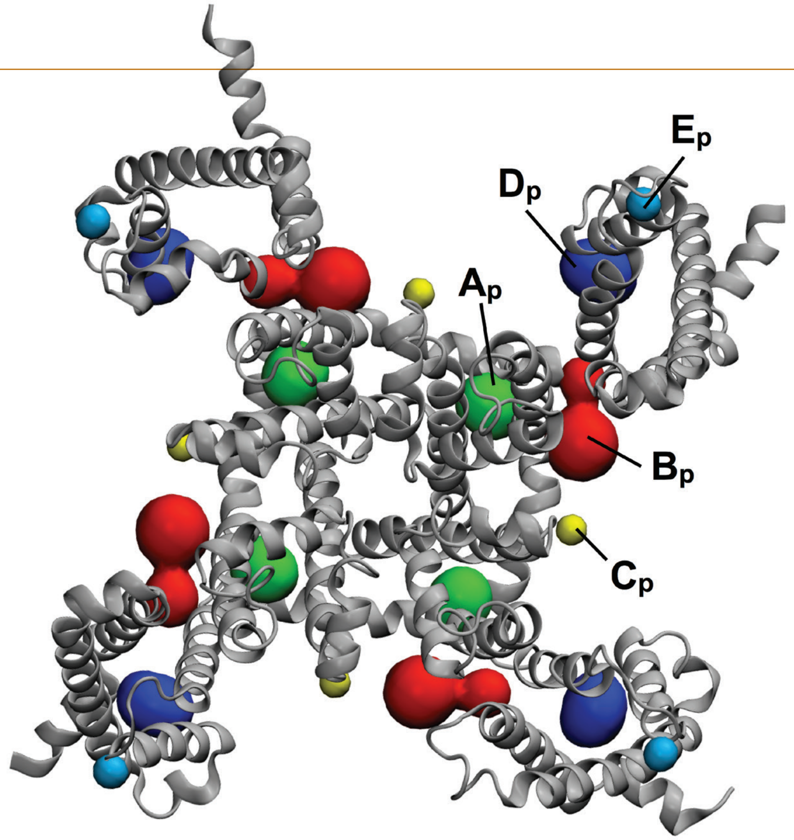
"We decided to use the power of Anton to not make any assumption at all of where the drug was going to bind," Boiteux adds. "We just put the drug into the solution and waited to see what would happen without any other input on our part."

The group concentrated on the voltage-gated sodium channel from bacteria, because the full three-dimensional structure of this protein was already known, unlike the human version. They could test human drugs against this protein because previous research had revealed it to react to the drugs in the same way human sodium channels do.

SURPRISING, REASSURING RESULTS

Simulating benzocaine and phenytoin with the membrane-bound gated channel provided both reassuringly predictable results and some big surprises. Benzocaine traveled to its known binding spot in the channel, blocking the ability of sodium ions to get through. But it used a highway of intermediate attachment points to get there that researchers had been entirely unaware of.





Anti-epilepsy drug phenytoin (colored blobs) sticks to the sodium channel in a more random series of spots (Ep through Ap), eventually settling into a place outside the channel (Ap) that closes the central pore indirectly.

"We could see benzocaine going straight to the site we knew was really important, which was really reassuring," suggesting that the simulation was recreating the real channel accurately, Boiteux says. "What we didn't expect was that all these other sites kind of aligned along the protein, leading to this major site and allowing the drug to enter the protein and block it." These intermediate sites could be potential drug-development targets.

Phenytoin had a very different way of approaching the channel. Unlike benzocaine, simulated phenytoin was unable to enter the channel quickly. Instead, it took a more random route, eventually sticking to a point on the exterior of the protein, shutting the channel indirectly and less completely than benzocaine—helping to explain the difference in the drugs' effects.

"What was interesting about phenytoin was that it was able to close the channel even without getting inside," Allen says, "and the amino acids it uses to close the gate are conserved between the bacterial and mammalian channels. This suggests the bacterial channel will be a good initial model for human anti-seizure drug discovery."

bridging THE GAP

One major focus of Bridges will be to allow neuroscientists to create fine maps of brain function, such as this network of 500 neural regions reconstructed from simulated magnetic resonance imaging. Some sequencing involves matching billions of short DNA sequences. (image above)

Ramsey, Joseph D., Ruben Sanchez-Romero, and Clark Glymour. (2014) Non-Gaussian methods and high-pass filters in the estimation of effective connections. Neuroimage 84:986-1006.

Bridges Will Bring High-Performance Computing and Data Analysis Capabilities to New Fields

Today, investigators in important, data-intensive fields such as cancer genomics, the digital humanities and machine learning increasingly find that they need the data-handling and analytic power of high-performance computers operating in a system designed specifically for manipulating and storing very large amounts of data.

With its new *Bridges* environment, PSC will provide these researchers with a new level of computational and data-handling capability that meets with their needs. Supported by a $9.65-million National Science Foundation (NSF) award, *Bridges* will be delivered by Hewlett-Packard (HP) based on an architecture designed by PSC. Construction is scheduled to begin in October 2015, with a target production date of January 2016.

"We designed *Bridges* to benefit new communities and to bring the power of high-performance computing to Big Data. The community's response to learning about *Bridges* has been overwhelmingly positive, confirming *Bridges*' transformational potential for research."
—*Nick Nystrom, Bridges Principal Investigator and Project Leader, PSC*

"*Bridges* represents a new approach to supercomputing that helps keep PSC and Carnegie Mellon University at the forefront of high-performance computing. It will help researchers tackle and master the new emphasis on data that is now driving many fields."
—*Subra Suresh, President, Carnegie Mellon University*

"The ease of use planned for *Bridges* promises to be a game-changer. Among many other applications, we look forward to its helping biomedical scientists here at Pitt and at other universities unravel and understand the vast volume of genomic data currently being generated."
—*Patrick D. Gallagher, Chancellor, University of Pittsburgh*

BRIDGING TO NEW RESEARCH COMMUNITIES

To support data analytics and to help research communities that have not traditionally used the tools of conventional high-performance computing (HPC), *Bridges* will offer features for flexibility and ease of use normally associated with lower-powered systems:

- extensive interactivity to let users work as they do on their own personal computing devices, complementing thought and providing immediate feedback, rather than requiring users to submit jobs and await their completion as in the usual supercomputing environment.
- web-browser-based gateways that will launch jobs and manage workflows transparently, behind-the-scenes, on users' behalf. This will let them harness *Bridges*' capabilities without having to master the more arcane tools of HPC.
- users' software and tools such as the Hadoop ecosystem, Python, R, MATLAB® and Java. Persistent databases and web servers will enable modern, flexible, and easily extensible software architectures.
- software portability and reproducibility, greater user control over software and environments and cloud interoperability through virtualization.

Together, these new features will let users easily transition software and data from their local computers to *Bridges*, and in doing so, open new, larger, and wider ranging lines of research.

BRIDGING TO THE DATA-INTENSIVE ERA

One of the greatest challenges to researchers in the 21st century is to extract knowledge from data so large and complex that its effective analysis and management requires substantial innovation. Using software developed at PSC and hardware developed by HP, Intel and NVIDIA, *Bridges* will provide flexible performance capabilities, combining large-memory nodes that can be applied to analyze huge amounts of data at unprecedented speeds with hundreds of smaller nodes for analyses that can be partitioned. These components are supported by a high-performance data management system and shared file system linked by an interconnect architected specifically for this purpose.

BRIDGING TO UNIVERSITY CAMPUSES

At peak times users often swamp a university's computational resources. A pilot project with fellow Pennsylvania institution Temple University will connect that campus with *Bridges* to give their researchers additional computing capacity at times of unusually high use, as well as to let *Bridges* offload suitable work to the Temple cluster.

TEACHING THE MACHINE TO SEE

BLACKLIGHT "TRAINS" VIDEO SEARCH SYSTEM FOR COMPETITION VICTORY

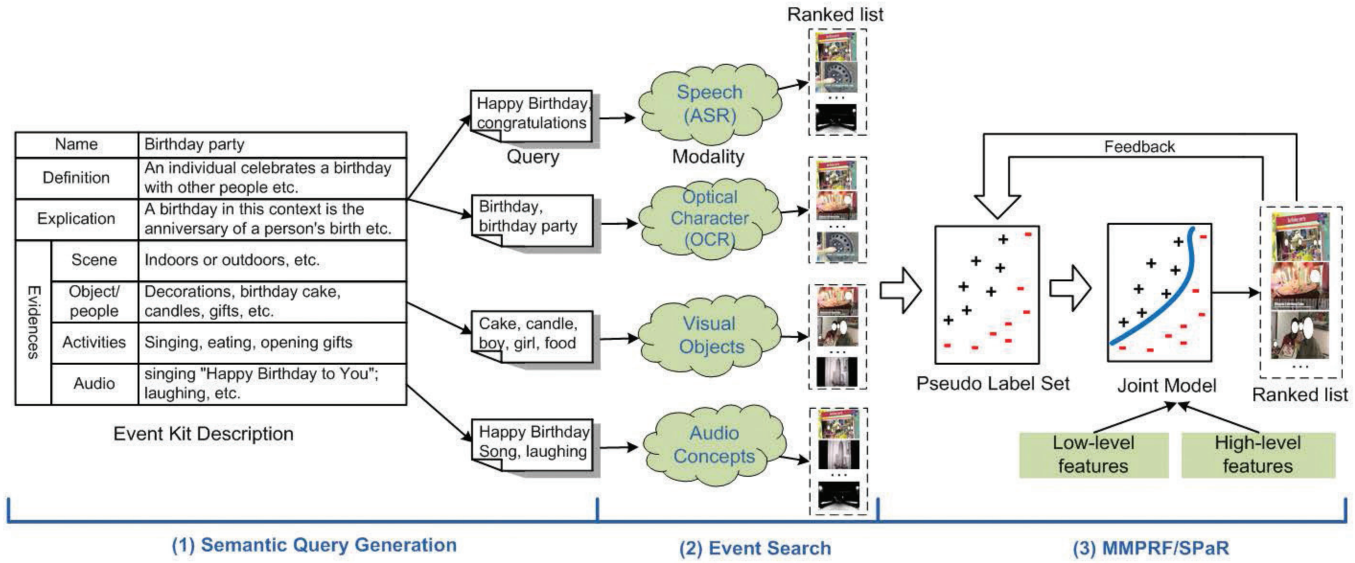
WHY IT'S IMPORTANT

A good example that we live in the era of Big Data is that, as we've moved from super-8-film home movies to ever-present smartphones, we've all begun to generate so much visual imagery that we seldom look at a given video more than once. Worse, when we do want to find a video clip, it's lost among thousands of others. Machine intelligence researchers Shoou-I Yu and Lu Jiang, working with colleagues on Carnegie Mellon University's Alexander Hauptmann's Informedia project and at PSC have developed E-Lamp, a system of "event detectors" designed to search for events in videos without human intervention. Such a detector could help us all keep better tabs of our video-electronic lives.

HOW THE DATA EXACELL AND BLACKLIGHT HELPED

> "When we tried to train the system on our own computer cluster, we were overloading our file system. Blacklight gave us eight times the speed, and we were not breaking the file system."
> —*Shoou-I Yu, Carnegie Mellon University*

The task of finding a video of a birthday party, for example, is fairly easy for a person. But it's extremely hard for a computer: All the cues a machine might use to spot a video, including color, shape, sounds and even captions can be misleading. Using PSC's Data Exacell system to manage a vast volume of data and its Blacklight supercomputer's large number of processors and very large memory, the researchers were able to employ a huge number of potential clues. The team was also able to develop a larger number of "concept detectors"—elements in the E-Lamp system for searching for specific things, such as birthday parties. The team increased the input of "training" videos into E-Lamp from 0.2 million video shots in its 2013 version to more than 2.7 million shots in the current version. They also increased the number of concept detectors from 346 to over 3,000. At the National Institute of Standards and Technology's 2014 TREC Video Retrieval Evaluation workshop, E-Lamp outperformed all other competitor systems in searching for videos based on either queries given to the researchers ahead of time or queries sprung on them at the competition.


Name
Birthday party
Definition
An individual celebrates a birthday with other people etc.
Explication
A birthday in this context is the anniversary of a person's birth etc.
Evidences
Scene
Indoors or outdoors, etc.
Object/ people
Decorations, birthday cake, candles, gifts, etc.
Activities
Singing, eating, opening gifts
Audio
singing "Happy Birthday to You"; laughing, etc.
Event Kit Description
Happy Birthday congratulations
Query
Birthday, birthday party
Cake, candle, boy, girl, food
Happy Birthday Song, laughing
Speech (ASR)
Modality
Optical Character (OCR)
Visual Objects
Audio Concepts
Ranked list
Feedback
Pseudo Label Set
Joint Model
Ranked list
Low-level features
High-level features
(1) Semantic Query Generation
(2) Event Search
(3) MMPRF/SPaR

E-Lamp consists of a series of tools that start with a definition of a kind of event (left), and then scans videos for sounds or images that are associated with those definitions (center). The machine can't know—it can only make statistical guesses. So the final step is to rank the possible "hits," with users providing feedback that help the system learn (right).

> "The system builds a model for detecting a concept, then tests that model. Then it builds an improved model and tests that. It asks itself, 'If I have features that look like this, will they help me do the best job in discriminating videos with dogs from videos without dogs?'"
> —*Alexander Hauptmann, Carnegie Mellon University*

> "If a video used to train the system is misleading—for example, a 'vacation video' that shows people changing a tire—it can be a disaster for accurately identifying a vacation. Blacklight allowed us to use more sample videos, and when you have more 'correct' samples and a smart algorithm, the training is more robust to the misleading samples."
> —*Lu Jiang, Carnegie Mellon University*

PSC staff helped Alexander Hauptmann's work with the support of programs within the NSF's XSEDE network of supercomputing centers: Extended Collaborative Support Service; Novel and Innovative Projects Program.

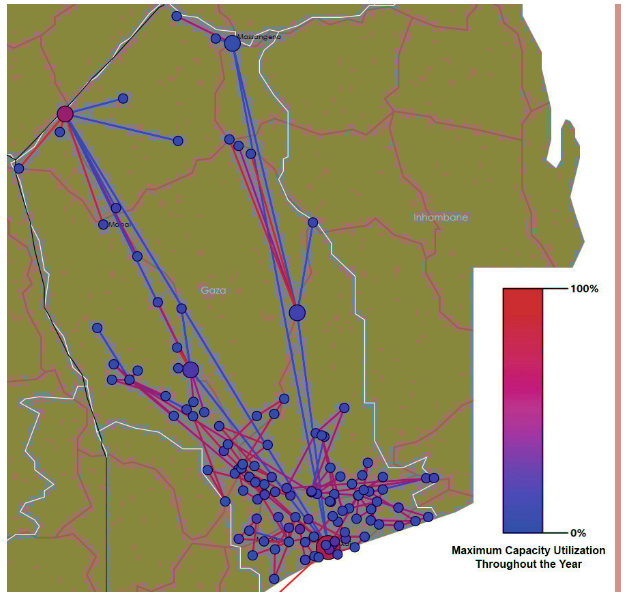
PUTTING TOOLS IN THE RIGHT HANDS

PSC Workshops Transform Public Health in East Africa

GETTING VACCINES WHERE THEY ARE NEEDED

In a series of workshops sponsored by the Mozambique Ministry of Health, PSC's Public Health Applications Group and their colleagues at the Johns Hopkins School of Public Health have trained ministry officials and public health stakeholders in that East African country to use two lifesaving computational tools.

Beginning last spring, the HERMES Logistics Modeling Team of PSC and Johns Hopkins collaborated with VillageReach to train stakeholders in using the HERMES supply chain simulation software, modeling vaccine delivery scenarios that participants proposed. Participants included Ministry of Health Expanded Program on Immunization (EPI) managers, Provincial Directorates of Health EPI managers from three provinces, UNICEF and World Health Organization vaccine supply chain specialists, local supply chain officers and representatives from the University of Eduardo Mondlane and VillageReach. As reported in the Fall 2014 *People. Science. Collaboration.*, a third-party analysis had concluded that the HERMES tool saved children's lives while reducing costs in a pilot vaccine program in the Republic of Benin in West Africa, leading to national changes in that country's vaccination policies.



Mozambique's vaccine delivery system, as modeled by HERMES

"The HERMES workshop in Mozambique was a major first step in building an in-country team of modelers, as we began by introducing the group to basic modeling concepts and ended with participants building and running their own models," says PSC's Leila Haidari, HERMES project coordinator. "This workshop also marked the first time in-country supply chain experts in any country used the HERMES user interface.

Participants left with new skills and interest in continuing modeling efforts, while we left with valuable feedback on our software and initial modeling results."

STOPPING THE SPREAD OF MALARIA

In another project in cooperation with the Mozambique Ministry of Health, the JANUS: Decision Support team of PSC and Johns Hopkins is assessing potential ways in which the country's mosquito control, case management and public health communication campaigns to combat malaria may be optimized. In May and November 2014, members of the JANUS team traveled to Mozambique to conduct workshops to present their results, garner feedback, and determine the most important questions of various stakeholders to guide the modeling efforts. Workshop participants included members of the Ministry of Health, The Global Fund, the U.S. President's Malaria Initiative, in-country personnel who implement Mozambique's malaria control program and others.

"Our JANUS workshops provided us with invaluable experiences and gave us a deeper and richer understanding of the needs, problems and nuances to implementing malaria control programs in Mozambique," says Johns Hopkins' Sarah Bartsch, project coordinator for JANUS. "Getting this feedback allows our team to better assist our in-country partners. It was great to see our work having a positive impact and knowing it is truly benefiting the Mozambican people."

Malaria is one of the largest public health concerns in Mozambique, accounting for 40 percent of all outpatient visits, 60 percent of admissions to pediatric wards, and almost 30 percent of all hospital deaths. The JANUS team was founded by Johns Hopkins Bloomberg School of Public Health and is funded in part by The Global Fund, an organization that gives $4 billion each year to fight diseases and support public health in more than 140 countries.

All Flesh Is Grass?

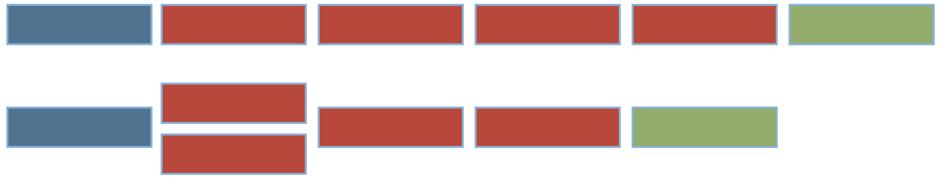
Blacklight Helps Researchers Untangle Genome of Wheat Progenitor

> "Assembling a genome is like putting together a jigsaw puzzle. With goatgrass, it was like putting together a jigsaw puzzle with lots of blue sky and clouds."
> —*Michael Schatz*
> *Cold Spring Harbor Laboratory*

WHY IT'S IMPORTANT

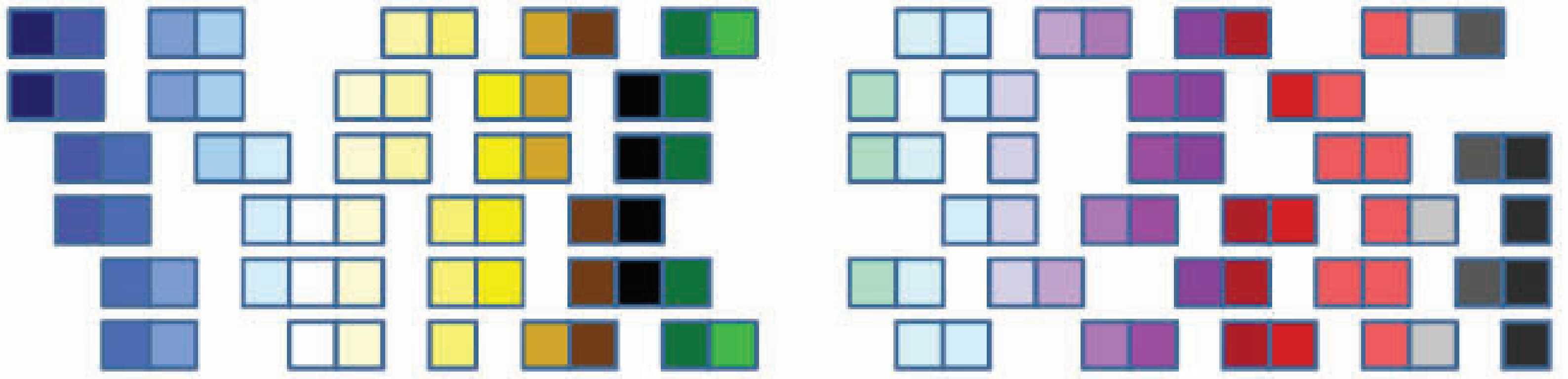
Wheat, a species of grass, provides more protein for human consumption—more flesh—than any other plant. Globally, we harvest 725 million metric tons of wheat every year. It's a fundamental human food source, found in many common items including breads, pastas and cereals. But wheat is a complicated plant, a hybrid of several progenitor species.

Bread wheat has six sets of chromosomes, compared with humans' two, and far more repetition in its genome—the collection of genes and other DNA that directs an organism's biology. Engineering improved disease resistance, drought tolerance or increased yields into wheat is difficult partly because the size and repetition of wheat DNA make it difficult to create an accurate, complete genome for the grain.



When a species' DNA sequence has multiple copies of nearly identical genes (top, red boxes), there's a danger that the computer will confuse two of them as the same gene (bottom, two red boxes stacked), and the final DNA sequence assembly will be missing one of those genes.

Whole genome sequencing involves matching billions of short DNA sequences.



HOW BLACKLIGHT HELPED

Michael Schatz and colleagues at Cold Spring Harbor Laboratory, New York, have studied wheat in part by determining the DNA sequence for *Ae. tauschii*, or goatgrass. This grass is one of the species our ancestors bred to create modern wheat. To determine the DNA sequence of a genome, researchers must first split it into short DNA fragments because current DNA sequencers cannot read more than a few hundred nucleotides at once from the many billions in the genome. Then they match overlaps in these short DNA sequences to put them together into their original order. But species like goatgrass and wheat with lots of sequence repetition can "fool" the matching process causing genes to escape detection.

Staff at PSC helped Schatz modify the gold-standard ALLPATHS-LG genome assembly software to handle big, repetitive genomes. Thanks to PSC's Blacklight supercomputer and its huge shared memory, the software could now assess many millions of potential ways of assembling small DNA fragments at once, without traveling back and forth to data storage. This greatly speeded the computation. Schatz's work on Blacklight detected at least 230 genes that had been missed in earlier attempts to assemble the goatgrass genome.

> "It's notoriously difficult to electronically store the de Bruijn graphs that represent possible overlaps of DNA sequences. Blacklight provided us with the very large memory necessary to do that."
> —*Michael Schatz*
> *Cold Spring Harbor Laboratory*

PSC staff helped Michael Schatz's work with the support of programs within the NSF's XSEDE network of supercomputing centers: Extended Collaborative Support Service; Novel and Innovative Projects Program.

DATA EXACELL MARKS FIRST YEAR WITH USER-DRIVEN ADVANCES

User input and experience have helped PSC's Data Exacell (DXC) complete a successful first year of operation. Thanks to the ongoing dialog with users, the DXC team has advanced the system's software and gained a new understanding of the advanced and novel hardware used to create it.

DXC's mission is to develop and test new hardware and software technologies as well as system architectures to support data-intensive research. These advances will be used at PSC and made available to the broader community for the benefit of the National Science Foundations's research-user community. DXC was funded by a major $7.6-million grant from the NSF's *Data Infrastructure Building Blocks* program in 2013. It was supplemented by a $1.2-million grant last year to add database capabilities required by the users.

User-motivated improvements have enabled the DXC to grow smoothly from 50 million to 250 million user files. In addition to inaugural users in fields that were relatively new to supercomputing, such as radio astronomy, machine learning and large-scale genomics, the system has served users from fields completely new to supercomputing. These include analyzing Twitter data for early epidemic warning, untangling coincidental correlations from cause-effect relationships in cancer and even how 19th-century authors handled character gender roles. This wide set of perspectives has helped PSC staff improve and test the DXC and, in the process, helped to guide the design of PSC's newest, user-friendly supercomputer, *Bridges* (See p. 14).

An important accomplishment of the DXC work has been to enhance the performance of SLASH2, the PSC-developed software that handles data for DXC. Improvements included optimizing SLASH2 for the GeneTorrent software toolset for using the Cancer Genomics Hub and creating a Big-Data-optimized alternative to rsync, a popular utility for transferring files. Another important component of DXC is to test and optimize the use of new, faster hardware for data storage, including solid-state disks and future SAS3-standard disk drives. PSC was the first customer to receive and test SAS3-capable hardware from Super Micro, Inc.

Future improvements to DXC may include innovations that allow multi-site users to employ their own authentication methods securely within the DXC system and methods to minimize the need for copying remote data to use it, reducing both data storage and networking demand.

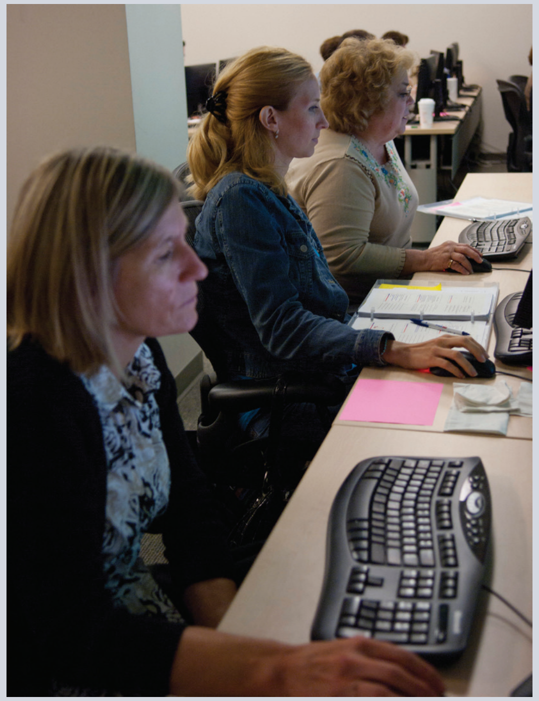
WEB10G CODE SUBMITTED FOR LINUX INCLUSION

Web10G, software developed by PSC and National Center for Supercomputing Applications staff to diagnose network problems, may soon be finding its way to a computer near you. The National Science Foundation-funded project has developed a means for extracting information from the common TCP/IP network protocol, for the first time giving network administrators detailed data necessary to diagnose and fix a host of networking problems.

"The code for Web10G has been submitted to the Linux kernel team for review and inclusion," says Chris Rapier, PSC network applications engineer. Once included in Linux—the operating system favored by many programmers and computer scientists—the software will be a step closer to inclusion in consumer operating systems such as MacOS and Windows. "This process might take some time, but I've been invited to speak at NetDev 0.1—the primary conference for Linux network developers. I'll be leading an open-ended discussion on the value of instrumentation like Web10G."

It may be a bit surprising to learn that even network engineers have never been able to look inside a TCP/IP connection to make fine-tuned assessments of what's going wrong with it. This is an unintended consequence of TCP/IP's original design. Web10G is a way to recover data on an individual connection so that network administrators and even individual users can tell why a network connection has failed or slowed.

Rapier presented Web10G and its applications for enhancing scientific workflow at several major symposia in 2014, including serving as an invited speaker at the Chinese American Networking Symposium at New York University and Internet2's I2 Tech Exchange in Indianapolis. At the latter, he also introduced Insight, a new interface that allows users to visualize their network connections, easily identify poorly performing connections and to submit Web10G data to administrators to aid in diagnostics.



PSC EDUCATION: BUHL INTERIM REPORT IS OCCASION TO TAKE STOCK

PSC's interim report to the Henry C. Frick Fund of the Buhl Foundation, submitted halfway through a three-year grant, provided an opportunity to take stock of the center's ongoing Innovative Approaches to STEM Education (IASE) program.

IASE is built on prior PSC programs in "computational reasoning," the use of modeling and simulation tools as well as pedagogical models. The program is using computational reasoning to help high school science and math teachers introduce their students to computational tools and to use those tools to teach course content. PSC developed the Computation and Science for Teachers (CAST) program based on tools developed by the Maryland Virtual High School and, with funding from the DSF Charitable Foundation, created a Professional Development program for teachers (available online at http://mvhs.shodor.org/).

The three-year 2012 Buhl grant continued CAST as IASE, combining its resources with those developed by PSC's Better Educators of Science for Tomorrow (BEST) program, which helps high school teachers incorporate computational tools into their biology curricula, and PSC's CMIST (Computational Modules in Science Teaching) program, which brings innovative science tutorials into secondary school classrooms, focusing on integrative computational biology, physiology and biophysics.

In the first phase of the program, PSC worked with teachers from the Pittsburgh School for the Creative and Performing Arts and the Pittsburgh Science and Technology Academy. A 2013 "Summer Institute" held at PSC brought teachers from these schools together to discuss and address STEM teaching challenges and introduced computational reasoning and the first modeling tool from CAST, interactive Microsoft Excel spreadsheets. Last year's Summer Institute opened IASE to more teachers from other Pittsburgh-area schools. Topics continued the exploration of both computational reasoning by introducing agent-based modeling in NetLogo and other simulation resources available over the Internet. The 2015 Summer Institute will focus on the last of the CAST modeling tools, the Vensim software, an industry-standard software package for creating computer simulations (http://vensim.com). Another focus will be the dual challenge faced by teachers: a standardized-test-oriented curriculum that makes incorporating new computational tools difficult, along with new written standards requiring such tools.

Future work will concentrate on developing pilot programs for introducing computational tools into curricula, comparing pilot schools with those without such programs.

PSC Resources for Data-Driven Science

Blacklight is an SGI Altix® UV1000 supercomputer designed for memory-limited scientific applications in fields as different as biology, chemistry, cosmology, machine learning and economics. Funded by the National Science Foundation (NSF), Blacklight carries this mission out with partitions with as much as 16 terabytes of coherent shared memory.

Sherlock is a YarcData Urika™ (Universal RDF Integration Knowledge Appliance) data appliance with PSC enhancements. It enables large-scale, high-performance graph analytical processing through massive multithreading (128 hardware threads per processor), a shared address space, sophisticated memory access optimizations and support for heterogeneous applications. Sherlock is funded by the NSF.

Anton is a special purpose supercomputer designed to dramatically increase the speed of molecular dynamics simulations, allowing biomedical researchers to understand the motions and interactions of proteins and other biologically important molecules over much longer time periods than previously possible. Designed and built by D. E. Shaw Research (DESRES), the Anton machine hosted at PSC was provided without cost by DESRES for non-commercial use by the national biomedical research community.

The **Data Supercell** (DSC) is a PSC-designed and built system for managing and archiving petabyte-scale data for researchers and industrial users. The DSC provides low-latency, high-capacity, high-reliability, high-bandwidth and low-cost data storage and retrieval.

The **Data Exacell** (DXC) is an NSF-funded pilot project to provide hardware and software building blocks to support data-intensive research projects. DXC is based on the unique, PSC-developed data storage architecture in the *Data Supercell*, combined with the very-large-memory capabilities of *Blacklight* and the graph-analytics capabilities of *Sherlock*, as well as specialized database capabilities. PSC experts are working with multiple research groups both to refine DXC architecture and to extend the NSF's support for new fields of science.

For more information on PSC's resources for data-driven science, go to WWW.PSC.EDU

Pittsburgh Supercomputing Center is a joint effort of Carnegie Mellon University and the University of Pittsburgh. It was established in 1986 and is supported by several federal agencies, the Commonwealth of Pennsylvania and private industry.

PSC gratefully acknowledges significant support from the following:

The National Science Foundation
The National Institutes of Health
The National Energy Technology Laboratory
The National Oceanographic and Atmospheric Administration
The U. S. Department of Defense
The U. S. Department of Energy

The Commonwealth of Pennsylvania
D. E. Shaw Research
Cray Inc.
Silicon Graphics, Inc.
The Buhl Foundation
Bill and Melinda Gates Foundation
The Global Fund
UNICEF

Design & production: Shandra Williams
Senior writer: Ken Chiacchia
Managing editor: Vivian Benton
Project coordination: Cheryl Begandy

Printing: HEETER Printing
Printed on Sappy McCoy paper, a premium sheet with 10
percent post-consumer waste fiber with vegetable-based inks

We would like your feedback on People. Science Collaboration by participating in a short survey, which you can find at https://www.surveymonkey.com/s/FRFPGRB. You can also scan the QR code using your mobile device.

Contribute to PSC's nonprofit, academic mission at psc.edu/donate.