



THE FOURTH PARADIGM IN PRACTICE

© 2012 Microsoft Corporation. All rights reserved

Except where otherwise noted, content in this publication is licensed under the Creative Commons Attribution 3.0 United States license, available at www.creativecommons.org/licenses/by/3.0/legalcode.



ISBN 978-0-9825442-1-1

Printed in the United States of America

The information, findings, views, and opinions contained in this publication are those of the authors and do not necessarily reflect the views of Microsoft Corporation or Microsoft Research. Microsoft Corporation does not guarantee the accuracy of any information provided herein.

Microsoft Research





CRAIG MUNDIE
*Chief Research and Strategy Officer
Microsoft Corporation*

Foreword

IN THE 20 YEARS since its founding, Microsoft Research has grown from a small group of researchers to more than 800 computer scientists and researchers at labs on four continents. Throughout this growth, the mission of Microsoft Research has remained consistent: to advance the state of the art in computer science and software engineering, and to take these technology advances to the public through our products.

As in many areas of research, the eventual applications and impact of computer-science research are challenging to predict and frequently different than originally envisioned. Over the years, we have seen many instances where Microsoft researchers have made advances that are at first incorporated into our products and technologies, but subsequently found to be uniquely impactful for an unrelated purpose in a very different field. Research at Microsoft helps drive forward the entire frontier of computing—and thus enables scientists to make significant progress on some of the major challenges facing society. That’s why we collaborate with leading academic researchers around the world on a broad range of issues related to health, the environment, education, and many other disciplines. By working on these non-computer-science projects we often discover new ways to hone and improve our own technologies—these are powerfully symbiotic relationships.

This collection of Science@Microsoft vignettes illustrates some of the progress researchers have

been making in a number of disciplines—from neurology and immunology to astronomy and climate change—and describes the technologies that have been deployed to gain these insights. We see our collaborative ventures and blue-sky research yielding returns both in the broader social arena and in improved products and services. It’s a classic case of “doing well by doing good.”

These stories also reveal interesting ways in which Microsoft and academic researchers are effectively applying computer science and technical computing research to fields far removed from computer science. With such multidisciplinary research collaborations, Microsoft Research is helping researchers reduce their time to new insights and accelerate the pace of scientific discovery. Microsoft’s unparalleled corporate support of basic research is fundamental to our long-term vision of transforming society through technology.

We have seen research on spam filters with machine learning lead to a new approach to develop an HIV vaccine. We have seen fundamental advances in computer vision and speech processing research lead to the Kinect hands-free controller. What will be the next breakthrough discovery, or cool new product? Who can say? That’s the beauty of doing basic research: you’re never certain where the results might lead, but you can be sure the journey will be worthwhile and full of surprises.

—CRAIG MUNDIE
Chief Research and Strategy Officer

Table of Contents



Genomics and Machine Learning

- 16 Using Model-Based Machine Learning to Understand Childhood Asthma
Understanding childhood asthma
- 18 Understanding the Genetic Causes of Human Disease
How researchers are untangling the complexity of human diseases
- 20 Algorithms that Can Handle the “Omics”
New developments in data analysis help address significant problems in biological research
- 22 Identifying Genetic Factors in Disease with Big Data
Analysis of large datasets helps researchers understand connections between heredity and disease

Earth and Environment

- 26 Amassing Global Data on Carbon and Climate
Enabling scientists to investigate the biological implications of persistent weather events
- 28 Simulating the Breathing of the Biosphere
Scientists address both the science and computational challenges
- 30 Data Deluge and Digital Watersheds
Using data to restore habitat for endangered fish
- 32 Improving Fuel Refining Technologies
How computer vision research can transform MRI
- 34 Transforming the Science of Behavioral Ecology
Exploring novel ways of investigating wildlife species

Health and Computer Vision

- 38 Understanding the Immune Response to HIV
Natural-killer cells play a direct role in fighting HIV
- 40 Medical Sensing via a Contact Lens
How this noninvasive technology monitors blood sugar levels
- 42 Improving Echocardiography
How scientists are improving the efficacy of echocardiography
- 44 Searching the Human Body
How researchers are untangling the complexity of human diseases

Computer Science and Technology

- 48 SenseCam
This device can aid those with debilitating memory impairment
- 50 Technology to Combat Counterfeit Products
How science is protecting consumers from purchasing fraudulent goods
- 52 Clinical Studies and Data Collection and Reuse
How software can help make clinical studies more effective
- 54 Microsoft TerraServer: an Imagery Scalability Story
Online, high-resolution maps

Table of contents continued next page ...

Table of Contents



Physics and Astronomy

- 58 WorldWide Telescope and Seamless Astronomy
How scientists are interacting with the universe
- 60 SkyServer: the Universe at Your Fingertips
How software can help make clinical studies more effective
- 62 Topological Quantum Computation
Faster, more powerful, more versatile computing

Biology and Life Sciences

- 66 Coevolution of Viruses and the Immune System
Understanding the human body's response to disease-causing pathogens
- 68 Software Verification Meets Biology
Biological models are bettering our understanding of biological phenomena
- 70 Programming Life
Using genetic devices to reprogram cell behavior and improve health, agriculture, and energy
- 72 How Many Species Are There?
Scientists are using expertise, tools, and technologies to find out

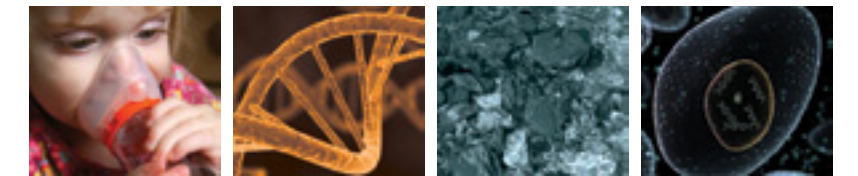
Citizen Science

- 76 GalaxyZoo
How volunteers are assisting professional astronomers to classify galaxies
- 78 Weatherathome.net
Anyone in the world can run a regional climate model
- 80 Rosetta@Home
Computer power—citizen participants help predict and design the three-dimensional structures

Scientific Tools

- 84 Chemistry Add-in for Word
Modify and share chemical information
- 86 NodeXL
NodeXL helps analyze and visualize its influence on individuals and societies
- 88 Scientific Workflow Workbench
Modify and share chemical information

Genomics and Machine Learning



Using Model-Based Machine Learning to Understand Childhood Asthma

ASTHMA AND ALLERGIES rank among the most common chronic disorders in children—and their incidence is on the rise. The question is why. Understanding the underlying causes of asthma might lead to prevention in susceptible children and better treatment for adults.

Although evidence from twin studies suggests a strong genetic component in asthma and allergies, few of these studies have identified the same genetic associations. Moreover, the role of the environment in asthma and allergies is evidenced by the rapid increase in the prevalence of these disorders over the last four to five decades, a time frame too short to be attributable to genetic factors alone. And indeed, various environmental exposures have been associated with the development of asthma and allergies. However, as with genetics, the data on the role of environmental factors are inconsistent, with the same environmental exposure showing increased risk, protection, or no effect, depending on the study.

The conflicting evidence on the effects of genetic variants and environmental exposures may be due in part to these factors having largely been studied separately. By contrast, this research seeks to model genetic and environmental factors jointly. The researchers view asthma as a complex disease that takes multiple forms, and therefore a central goal is to discover these underlying “phenotypes.” To do so, they are using a Bayesian model-based machine-learning approach, which, unlike conventional statistical analyses and black-

box machine-learning methods, easily allows the incorporation of rich, hierarchical structure derived from clinical background knowledge.

The project is a collaboration between Microsoft Research and the [University of Manchester](#). Microsoft Research provides the machine-learning expertise through its team of Chris Bishop, John Winn, Markus Svensén, and Nevena Lazic in the Machine Learning and Perception group, and has contributed [Infer.NET](#), a new framework that allows rapid construction of complex Bayesian models and performs efficient inference within those models. The University of Manchester brings world-class clinical expertise through Iain Buchan, Adnan Custovic, and Angela Simpson, along with a high-quality dataset collected by the [Manchester Asthma and Allergies Study](#).

Together, the researchers hope to build complex models that represent a broad range of important variables associated with asthma. And while the immediate goal is to study the development of asthma, a successful outcome can highlight the benefits of a model-based approach to the analysis of clinical data generally, which could have much broader applicability.

This work significantly contributed to the development of Infer.NET by providing real-world testing of its scalability and capabilities, and led to the first paper published using this technology. Improvements to Infer.NET resulting from this project have helped prepare the framework to enhance a range of Microsoft products.



CHRIS BISHOP
Distinguished Scientist
Machine Learning and
Perception Group
Microsoft Research Cambridge



ADNAN CUSTOVIC
Professor of Allergy
Head, Respiratory
Research Group
University of Manchester



“It is very exciting to take ideas from a technology context and apply them to such an important societal problem. At the same time, the challenges of modeling a disease as complex as asthma are pushing us to develop the machine-learning techniques in new directions.” —CHRIS BISHOP

“With Microsoft Research, we have created a multi-disciplinary collaborative partnership which will combine our world-leading expertise in birth cohorts and health informatics research with innovative computational statistical methods. We will work together to apply newly developed state-of-the-art data analysis techniques (of Bayesian inference and machine learning) to build complex models to describe different types of ‘asthma’ and investigate risk factors (both genetic and environment) for each asthma subtype. In doing so, we hope to understand the basic biological mechanisms that underlie the different asthmas, and to identify molecules which may be targets for future drug therapies.” —ADNAN CUSTOVIC

Understanding the Genetic Causes of Human Disease



JOHN WINN
*Senior Researcher
Machine Learning and
Perception Group
Microsoft Research Cambridge*



RICHARD DURBIN
*Joint Head of Human
Genetics
Wellcome Trust Sanger
Institute*

MANY COMMON DISEASES, including cardiovascular disease, cancer, and various psychiatric illnesses, arise from complex interactions between a person’s genetics and all the environmental influences he or she encounters over a lifetime. Untangling these factors and determining the underlying causes might lead to better prevention, diagnosis, and treatment of the diseases. Physicians might even develop individualized treatments that are based on a patient’s genetic make-up.

But analyzing multiple genetic and environmental factors is complicated, to say the least. Researchers have developed enormous genomics datasets but need statistical methods that can represent the complexity of human disease. In other words, they need both rich data about the behavior of human cells in various environmental contexts along with complex statistical models to analyze the resultant data.

Impossible? Hardly, thanks to the work of the [Machine Learning and Perception](#) group at Microsoft Research Cambridge and the [Wellcome Trust Sanger Institute](#). Senior Researcher John Winn and his colleagues have developed [Infer.NET](#), an advanced machine-learning framework for modeling and understanding very complex systems. Infer.NET allows the team to represent the complexity of human diseases in a way previously unachievable. The team at the Wellcome Trust Sanger Institute, led by Joint Head of Human Genetics Richard Durbin, brings world-class expertise in large-scale genomic sequencing and analysis of genomic data.

The project requires analyzing four types of data:

- **Genetic data**—All or key parts of the DNA sequence of an individual.
- **Functional genomic data**—Measurements, such as gene expression, that indicate the activity of individual genes in various body tissues.
- **Environmental data**—Information about an individual’s environmental exposures, such as smoking or sunbathing.
- **Disease data**—Physiological measurements and information about known diseases or symptoms that an individual has experienced.

This data is brought together in a single statistical model, so as to discover associations between the genome, cell function, environmental factors, and disease.

The researchers’ analysis is identifying correlations between genetics and the activity of genes in different tissues and the symptoms or characteristics of the individuals from whom the samples come. This is shedding new light on how variations in genetic makeup can make individuals susceptible to different diseases, giving researchers deeper understanding than ever before on the genetic causes of human disease.

The application of Infer.NET to bioinformatics led to the first parallel version of the framework, driving improvements to the design and implementation of this key technology. By pushing the scalability of Infer.NET, this project directly helped make this machine-learning framework ready to use in a number of Microsoft products.

“New, large-scale functional genomics datasets have the potential to tell us a huge amount about the underlying causes of many diseases. But data alone cannot achieve this—we need new statistical methods that are capable of representing the realistic complexity of human disease. I am very excited to play a part in developing such methods because they are the key to unlocking the secrets contained in these exciting new datasets and to making new inroads in understanding and treating disease.” —JOHN WINN

“The scale and complexity of genomic datasets is increasing exponentially with the recent revolution in DNA sequencing and related technologies. Working with Microsoft Research enables us to bring state-of-the-art ideas and methods in machine learning research to bear on these data sets, letting us shed light on key interactions involved in complex disease.” —RICHARD DURBIN

Algorithms that Can Handle the “Omics”



RICCARDO ZECCHINA
Professor
Theoretical Physics
Politecnico di Torino



JENNIFER CHAYES
Distinguished Scientist
Managing Director
Microsoft Research
New England



CHRISTIAN BORGS
Deputy Managing Director
Microsoft Research
New England

IMAGINE TWO NEANDERTHAL proto-scientists, standing before an enormous pile of rocks—big rocks, small rocks, smooth rocks, rocks with jagged edges.

“Well,” remarks one of the Stone Age researchers, “this is a lot of data here.”

“Yes,” replies his colleague, “if only we had some way to make sense of it all, I bet we could achieve a breakthrough in rock utilization.”

Today, biological researchers are confronting a similar dilemma: a wealth of data but an inadequate analytical toolkit. Increasingly, biologists are using genomic methods, such as expression profiling, next-generation sequencing, and RNAi screens, together with proteomic and metabolomic technologies, to discover the molecular basis of changes in living systems. Collectively, these methods are often referred to as “omics.”

Unfortunately, these advances in experimental technologies have, in many cases, outpaced the development of the bioinformatics tools that are needed to analyze the data. In other words, optimal analysis of large quantities of experimental data often require the solution of hard problems that are intractable by conventional computational tools. Fortunately, collaborative efforts between the [Massachusetts Institute of Technology](#) and [Microsoft Research](#) have resulted in new algorithms that often do quite well in analyzing these types of problems.

These methods, which extend ideas from statistical physics of disordered systems to problems in computer science, have provided novel distributive algorithmic schemes for solving large-scale optimization and inference tasks. Among the features of these new algorithms are computational efficiency, parallelizability, and flexibility to include heterogeneous prior knowledge and to integrate diverse data sources. The spectrum of applications ranges from constraint satisfaction and stochastic optimization problems over networks to graphical games and statistical inference problems.

The collaboration aims to apply these new algorithmic techniques to significant problems in biological research. Preliminary results have already led to the discovery of new functional genes, the prediction of protein contacts from sequence data, and the discovery of a new algorithm for message passing. These developments have the potential for broad application across computer science and to improve future Microsoft products.

The researchers’ current focus is on cancer genomics. In collaboration with the [Memorial Sloan-Kettering Cancer Center](#), they are working on the integration of different types of molecular data to reveal complex response pathways of relevance in cancer development. They hope that this work leads to not only advances in general algorithmic techniques for biological research but also to the development of drug targets for specific cancers.



“Using insights from the statistical physics of disordered systems, we have developed highly parallel algorithmic techniques for addressing a host of combinatorial optimization problems. The applications range from more conventional computer science problems like multicasting to problems of biological significance, including in particular an understanding of human gene regulatory networks—which, hopefully, will lead to the development of drug targets for various cancers.” —JENNIFER CHAYES

“Current biological research is generating a vast amount of data relevant to problems of human disease—from cancer to Alzheimer’s. But algorithmic techniques have not kept pace with the increase in experimental data. Our research with Riccardo Zecchina and his group in Turin is focused on closing this gap by developing new algorithmic techniques originating in statistical physics with domain-specific methods developed by computational biologists.” —CHRISTIAN BORGS

“We are filling an existing methodological gap between advanced algorithmic schemes that have their origin in statistical physics and the huge amount of molecular and evolutionary data made available by experimental technologies. We are developing computational tools which are intrinsically parallel and can be applied to large-scale inverse problems of biological interest and of clinical relevance in cancer genomics.” —RICCARDO ZECCHINA

Identifying Genetic Factors in Disease with Big Data



DAVID HECKERMAN
*Microsoft Distinguished Scientist
Senior Director
eScience Group
Microsoft Research Connections*



JENNIFER LISTGARTEN
*Researcher
eScience Group
Microsoft Research*



CHRISTOPH LIPPERT
*Student
Max Planck Institutes for
Developmental Biology and for
Intelligent Systems*

MEDICAL RESEARCHERS HAVE long known that many serious diseases—including heart disease, asthma, and many forms of cancer—are hereditary. Until fairly recently, however, there was no easy way to identify the particular genes that are associated with a given malady. Now, researchers can conduct genome-wide association studies—by sequencing the DNA of human subjects—enabling them to statistically correlate specific genes to particular diseases.

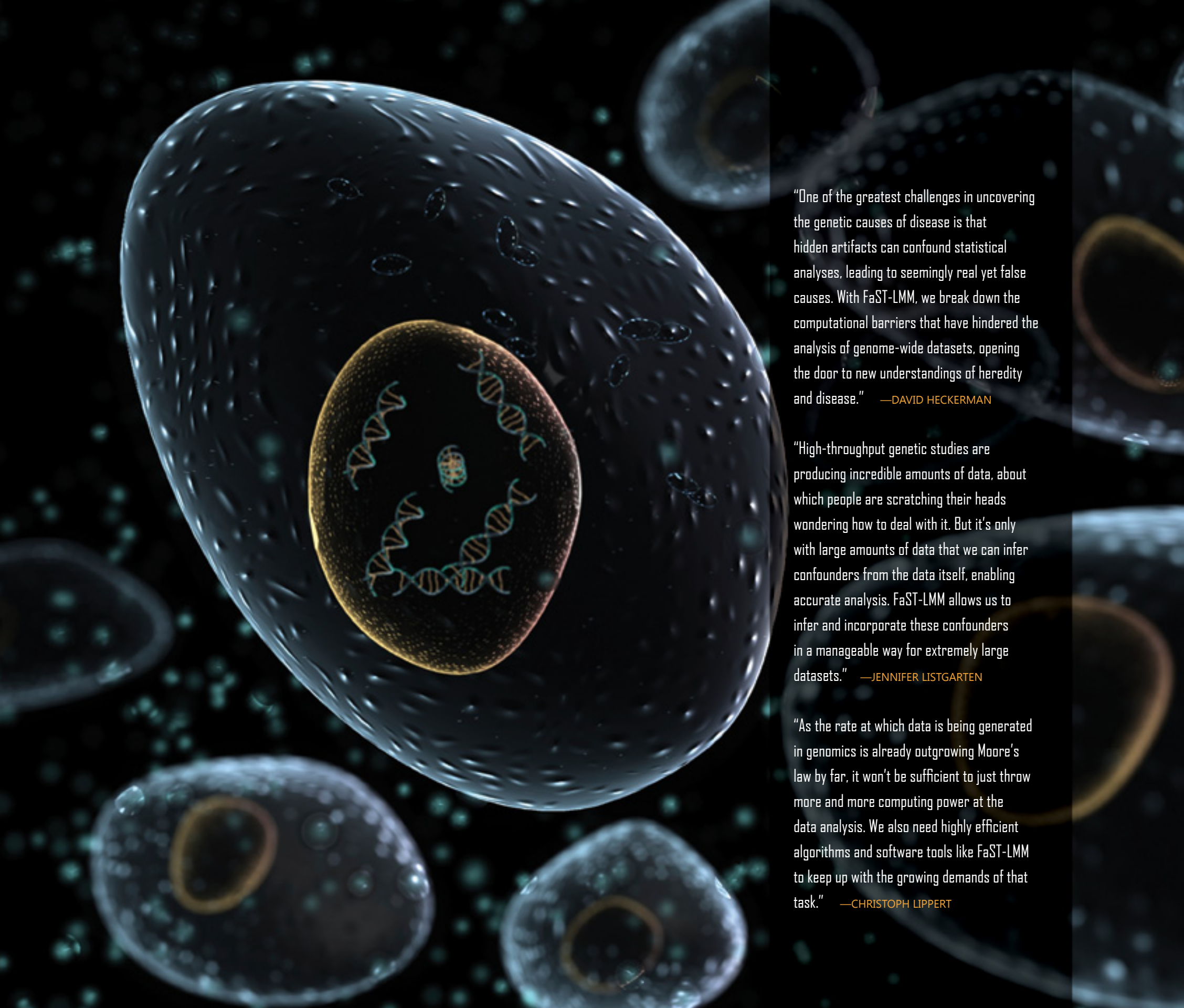
In order to study the genetics of a particular disease, researchers need a large sample of people who have the disorder, which means that some of these people are likely to be related to one another—even if it’s a distant relationship. This skews research results because certain positive associations between specific genes and the disease are false positives—the result of two people sharing a common ancestor. In other words, the research sample is not truly random, and researchers must statistically correct for the “confounding” that was caused by the shared ancestry of the subjects.

This is not an insurmountable statistical problem: there are so-called linear mixed models (LMMs) that can eliminate the confounding. However, an inordinately large amount of computer run time and memory are required to run LMMs to account for the relatedness among thousands of

research subjects. When working with the large datasets that offer the most promise for finding the connections between genetics and disease, the cost of the computer time and memory that these models require can quickly become prohibitive.

To address this problem, the Microsoft Research team developed Factored Spectrally Transformed Linear Mixed Model (FaST-LMM), an algorithm for genome-wide association studies that scale linearly in the number of individuals in both run time and memory use. FaST-LMM can analyze data for 120,000 individuals in just a few hours (whereas the current algorithms fail to run at all with data for just 20,000 individuals). The outcome: large datasets that are indispensable to genome-wide association studies are now computationally manageable from a memory and run-time perspective.

FaST-LMM will enable researchers to analyze hundreds of thousands of individuals to find relationships between their DNA and their traits, identifying not only which diseases a given patient may get, but also which drugs will work best for that patient. FaST-LMM takes us one step closer to the day when physicians can provide their patients with personalized assessments of their risk of developing certain diseases and devise prevention and treatment protocols that are attuned to their unique hereditary makeup.



“One of the greatest challenges in uncovering the genetic causes of disease is that hidden artifacts can confound statistical analyses, leading to seemingly real yet false causes. With FaST-LMM, we break down the computational barriers that have hindered the analysis of genome-wide datasets, opening the door to new understandings of heredity and disease.” —DAVID HECKERMAN

“High-throughput genetic studies are producing incredible amounts of data, about which people are scratching their heads wondering how to deal with it. But it’s only with large amounts of data that we can infer confounders from the data itself, enabling accurate analysis. FaST-LMM allows us to infer and incorporate these confounders in a manageable way for extremely large datasets.” —JENNIFER LISTGARTEN

“As the rate at which data is being generated in genomics is already outgrowing Moore’s law by far, it won’t be sufficient to just throw more and more computing power at the data analysis. We also need highly efficient algorithms and software tools like FaST-LMM to keep up with the growing demands of that task.” —CHRISTOPH LIPPERT

Earth and Environment



Amassing Global Data on Carbon and Climate

AS CONCERNS OVER climate change mount, hard data on the interactions between the atmosphere and biosphere have become vital. Fortunately, over the past two decades, scientists have deployed sensor collections at several hundred locations worldwide, gathering continuous, long-term data on carbon balance across different climate zones and vegetative land covers. These records enable scientists to investigate the implications of fires and other disturbances and to study the impact of land-management approaches such as fertilization, grazing, and irrigation. They also let scientists examine the biological implications of persistent weather events, such as drought, or episodic events, such as major storms.

Sitting atop these sensor collections is [FLUXNET](#), a “network of networks.” Through FLUXNET, independent regional networks and individual field scientists come together to pursue synthesis science—crossing disciplines, data sources, and national boundaries. The FLUXNET dataset consists of more than 960 site-years of sensor data from more than 253 sites, as well as more than 100 related non-sensor field measurements. Contributions are ongoing and the data-processing algorithms are continually improved through experience.

Managing FLUXNET’s living dataset presents several computing challenges, which are shared by Dario Papale of the University of Tuscia in

Italy and Markus Reichstein (Max Planck Institute, Germany), working with Deb Agarwal (Lawrence Berkeley Laboratory, United States), Dennis Baldocchi (University of California, United States), Marty Humphrey (University of Virginia, United States), and Catharine van Ingen (Microsoft Research, United States). The team has constructed an advanced data server that is based on [Microsoft SQL Server](#) technologies and a collaboration portal that is based on [Microsoft SharePoint](#). Sensor data can be rapidly browsed over the Internet from the scientist’s desktop, and the portal enables communication among scientists via a private social networking site.

The ability to share data at this scale and diversity enables new insights, can reduce the uncertainty of existing model predictions, and has been the basis for landmark papers that challenge conventional wisdom. One such paper suggests that the availability of water may be more important than temperature for carbon fixation by plants, a conclusion that poses questions about existing predictions of ecosystem changes—such as tropical forest decline—in response to temperature change.

The project team designed a database for storing environmental data and explored how SQL databases, SQL Analysis Services data cubes, and Microsoft SharePoint can manipulate such data. Five other science projects in the field of climate change study have already used the database design.



DEB AGARWAL
Senior Scientist
Departmental Head, Advanced Computing for Science
Lawrence Berkeley National Laboratory



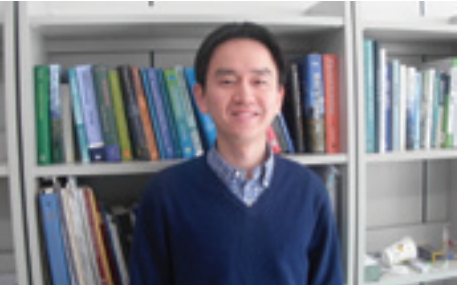
DENNIS BALDOCCHI
Professor
Biometeorology
University of California, Berkeley



“Enabling the more than 120 global groups studying the carbon cycle through [fluxdata.org](#) is helping us to better understand what cyber infrastructure is needed to further the science.” —**DEB AGARWAL**

“Our goal is to produce a data system that enables us to produce computations of ecosystem photosynthesis and evaporation ‘everywhere and all of the time.’ The computational demands of such a product are far beyond the capacity of the PCs in my lab. Working with Microsoft and the cloud computing system has enabled us to take ideas that were expressed on a classroom chalk board into reality.” —**DENNIS BALDOCCHI**

Simulating the Breathing of the Biosphere



YOUNGRYEL RYU
Assistant Professor
Environmental Ecology
Seoul National University



MARTY HUMPHREY
Associate Professor
Department of Computer
Science
University of Virginia

THE BIOSPHERE AND atmosphere interact. Climatic variables, including temperature, humidity, wind, and precipitation, affect individual plants and plant communities, while the type and diversity of plant cover influence the atmosphere. Understanding the breathing of the biosphere—the fluxes of carbon dioxide, water vapor, and trace gases between plants and the atmosphere—is a challenging task, involving numerous coupled, non-linear, biophysical processes.

At the regional to global scale, the computational demands require at least daily merging of several remote-sensing datasets. Scaling such a computation means scaling the science as well; for example, the science algorithm must encompass rainforests as well as croplands.

The [Breathing Earth System Simulator](#) (BESS) computation on the [MODISAzure](#) cloud addresses both the science and computational challenges. BESS scientists Youngryel Ryu and Dennis Baldocchi developed the science computation, which synthesizes data from satellite imagery, global climate models, and ground-based sensors. MODISAzure computer scientists, Marty Humphrey, Jie Li, You-Wei Cheah, and Catharine van Ingen addressed the computational challenges by building a four-stage data processing pipeline on [Windows Azure](#).

Cloud computing addresses three barriers to such a computation:

- Access to sufficient resources—The BESS computation used more than 500,000 CPU

hours, ingesting 14 terabytes of data from NASA and generating 1.5 terabytes of results. MODISAzure scales from 5 to 240 deployed virtual machines in the cloud.

- Tedium—The global-scale science computation breaks into more than 70,000 tasks and consumes 500,000 input files. MODISAzure marshals the right input files for each task and handles synchronization and error recovery.
- Complexity—The bulk of the input files were the result of preprocessing approximately 800,000 initial NASA satellite files into “sinusoidal” files that contain the daytime aggregate value at fixed, equal-sized pixels.

Computations like BESS enable researchers to examine complex climate changes globally and locally. Global-scale fluctuations can be simulated with climate models, extending their understanding of complex climate change phenomena.

The results can also be combined with local observations to address specific questions, such as the implications of rice farming in the Sacramento Delta region.

This was one of the first applications to use Windows Azure as a large computational platform. It provided the Windows Azure team with feedback and input that could be useful for future refinements to the cloud platform. It also proved how scientists can broaden their work from one PC to large-scale computations.

“The Microsoft cloud service, Azure, enabled me to study land-atmosphere interactions from 1 kilometer to a global scale. I learned how collaboration across disciplines can advance the science.” —YOUNGRYEL RYU

“We have been very excited to participate in the MODISAzure project, seeing first-hand what capabilities Windows Azure might be able to provide to eScience activities now and in the future. We really think this can enable a new generation of data-first explorations—we have been amazed at the results that Windows Azure has enabled the BESS scientists to achieve!”

—MARTY HUMPHREY

Data Deluge and Digital Watersheds

COMMODITY SENSORS AND Internet connectivity have created a veritable data deluge. Yet it remains a challenge to find, access, clean, and reuse data. That is particularly true when data from different sources is needed for synthesis science—bringing many diverse observations together to create a larger, holistic view.

Scientists at the [Berkeley Water Center](#), led by James Hunt, worked with the [National Marine Fisheries Service](#) on a synthesis challenge in the Russian River area. The watershed is the breeding ground for several species of fish, but wine-grape farming, urbanization, gravel mining, and other factors have affected the river. As a result, the fish have become endangered, and habitat restoration is critical.

To enable such studies, Microsoft Research’s Catharine van Ingen and [Lawrence Berkeley National Laboratory](#) researcher Deb Agarwal built digital watershed. Constructed on a [Microsoft SQL Server Database](#) and SQL Server Analysis Services data cube, the digital watershed enables simple interactive browsing of the assembled diverse

sensor and field observations. Data updates are automatically harvested from available government websites and are ingested from smaller spreadsheets or other sources. Some of the data are historic, dating back more than 100 years. Others are real-time measurements with only transient availability over the Internet.

Among the questions examined with the digital watershed was the impact of “frost dips.” During the season when wine-grape buds are setting, local farmers use sprinklers to avoid frost damage. Pumping from the river to supply the sprinklers causes transient dips in the river water level. Such dips can strand small hatchling fish, making them susceptible to predators or oxygen deprivation.

This research led to a fundamental change in how hydrologists get timely answers to their questions by using advanced computing methods. It established a model for how scientists can use technologies to help them solve problems with data processing so that they can remain focused on their sciences.



CATHARINE VAN INGEN
Partner Architect
eScience Group
Microsoft Research



JAMES HUNT
Professor
Civil and Environmental
Engineering
University of California, Berkeley



“The Turkish proverb, ‘Small keys unlock big doors,’ characterizes much environmental data. Remote sensing, ground-based sensors, direct measurements in the field, and historic events such as fires and floods all play a part.”

—CATHARINE VAN INGEN

“Our collaborative research with Microsoft and LBNL [Lawrence Berkeley National Laboratory] was essential in utilizing the extensive and widely distributed data to help environmental managers resolve fishery habitat questions in altered watersheds. Within environmental science, data on the state of water, air, land, and ecological resources are essential in understanding current conditions and providing a base for assessing future conditions. Modeling at multiple levels of complexity is a secondary component that is heavily dependent on data synthesis efforts at different spatial and temporal scales.”

—JAMES HUNT

Improving Fuel Refining Technologies

CLIMATE CHANGE IS a hot topic these days, with great attention focused on the need to reduce atmospheric carbon dioxide (CO₂) concentrations to slow the process of global warming. Such reduction will rely, in part, on developing better quality—that is, lower emission—automotive fuels, and among the most promising approaches to improving fuel quality is the use of gas-to-liquid (GTL) refining technologies.

But making GTL competitive with conventional oil refining technology will require improved design in trickle-bed reactors (TBRs). TBRs are essentially a

column filled with catalyst particles, through which gas and liquid flow downward. Improved design and operation of TBRs could reduce CO₂ emissions by 400 million metric tons a year, or about 2 percent of global CO₂ emissions.

TBRs are currently designed by using imprecise empirical correlations, which leads to significantly oversized reactors and higher operating and capital costs, thereby putting TBR technology at a competitive disadvantage. The best way to improve TBR design is to enhance understanding and modeling of the reactors through new phenomenological models and better closure laws for computational fluid dynamics simulations. Both of these require measurements of the velocity of the liquid and gas inside the reactor, as this will determine the overall performance of the reactor.

The only technique that can provide these measurements is magnetic resonance imaging (MRI), but the signal-to-noise ratio of conventional gas phase imaging is too low to achieve the required resolution. Collaboration between the [Magnetic Resonance Research Centre](#) at the University of Cambridge and Microsoft Research has produced a compressed sensing algorithm that enables gas phase velocity mapping at a resolution an order of magnitude greater than in the past. These measurements are of sufficiently high resolution to enable characterization of the interfacial velocity between the liquid and gas, which is critical to improved understanding of the reactor’s behavior and prospective design improvements.

In addition to its potential for helping to reduce atmospheric CO₂ emissions, the compressed sensing algorithm could lead to greater use of MRI as a means for testing hypotheses about the process under study. A particular target is to reduce data acquisition times by an order of

magnitude; this opens up new opportunities for studying chemical engineering processes, as well as enabling the implementation of magnetic resonance measurements with low magnetic field hardware, which would enable the use of on-site MRI as a process analytics tool.

Most recently, the researchers have been looking at radical techniques that avoid, altogether, the need to produce an image in order to perform a particular analysis. Producing an intermediate image is very costly in terms of the amount of data required and may be unnecessary when what is needed is a simple decision or an estimate of the value of a few parameters. An example is estimating the density and shape distribution of bubbles in a reactor. They have shown that this can be done directly, without any intermediate image, resulting in much shorter acquisition times and tolerance to noise. Ultimately, this could allow the use of more compact MRI machines, avoiding the need for strong magnetic fields and thus avoiding the need for supercooling. This would simplify the machinery in terms of both weight and cost, and allow certain kinds of measurements that were previously impossible. The resulting machines could have an impact in chemical engineering and, potentially, medical diagnosis.

In an increasingly data-driven world, where big data and real-time analytics become more prevalent with technologies like StreamInsight, Windows Azure, and Hadoop, probabilistic techniques, such as the compressed sensing algorithm developed during this research, can produce step changes to improve how researchers understand and use data competitively. The algorithm exemplifies the extensive expertise that Microsoft Research applies to developing a wide range of applications, which is vital to Microsoft’s development of more sophisticated analytical and data management tools.



LYNN GLADDEN
*Pro-Vice-Chancellor
for Research
Shell Professor
of Chemical Engineering
University of Cambridge*



ANDREW BLAKE
*Microsoft Distinguished Scientist
Managing Director
Microsoft Research Cambridge*

“The opportunity to develop technologies based on probabilistic inference that may change the nature of magnetic resonance imaging machines has been exciting. The results that we’ve been able to achieve together show how multidisciplinary research can create step changes in science.” —ANDREW BLAKE

“The project with Microsoft Research has enabled us to make significant advances in magnetic resonance imaging, which we have used in medical MRI as well as our main activity in chemical engineering. It has been possible to reduce imaging acquisition times by more than an order of magnitude and, in so doing, enabled us to study phenomena in multi-phase hydrodynamics that have not been studied previously using magnetic resonance or, indeed, any other experimental technique.” —LYNN GLADDEN

Transforming the Science of Behavioral Ecology



ROBIN FREEMAN
Researcher
Computational Ecology
Microsoft Research Cambridge



TIM GUILFORD
Professor of Animal
Behaviour
Department of Zoology
University of Oxford

UNDERSTANDING HOW THE behavior of species changes over time is vital to knowing how to protect and preserve key species. This is critical for those species that are vulnerable to changes in the environment, such as the response of global ecosystems to climate change and human activity. In particular, understanding the movement and spatial dynamics of individual species, between individuals and their environment, and spatial locations and patterns that are important for species survival, is vital. However, established techniques for studying the movement and behavior of individual animals are typically limited, inefficient, time consuming, and expensive.

The two most commonly used approaches to observing species are far from perfect. The first approach is a scientist spending years observing migratory behavior, for example, which tells only when animals depart and when they return (if at all), but nothing about what they do when, where, and why they migrate. The second approach includes tracking technologies, which are expensive, inflexible, inappropriate (in size, range, functionality, or weight, for example), and labor intensive, limiting both the applicability and scale of ecological and behavioral studies. In addition, even when data are collected, few if any computational and software tools exist to analyze the data easily and accurately or to enable the development and testing of predictive models that use the data. As a consequence, scientists understand remarkably little about the behavior of most key species, and less still about how their behavior is or will change as the environment changes.

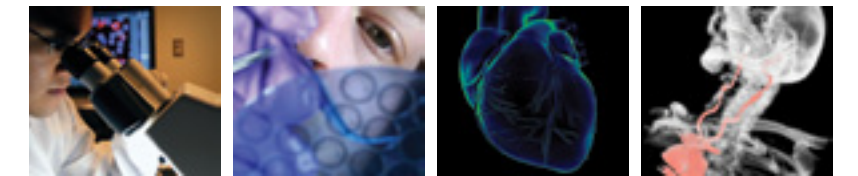
Robin Freeman, a zoologist in the [Computational Ecology and Environmental Sciences](#) group at Microsoft Research, has developed an open, reconfigurable, flexible, wirelessly-enabled, and low-cost tracking technology and set of software tools that address almost all of these problems. These technologies enable Freeman, his collaborators Tim Guilford, Ben Dean, and Holly Kirk at the [University of Oxford](#), and now, other researchers worldwide, to undertake previously impossible scientific studies collecting novel types of data and employing new kinds of analyses. These studies are initially focused on the migratory and foraging behavior of pelagic seabirds. The new platform is an open design with open software, so researchers can choose to modify the existing designs as their projects require. Microsoft Research has provided a number of solutions for most common tracking problems.

Furthermore, the group’s software tools for data analysis, modeling, and visualization enable Robin and his colleagues to analyze these new data to form a fundamentally better understanding of the movement and behavior of individuals, groups, and populations of important species, including the Manx shearwater, Hutton shearwater, black petrel, puffins, and guillemots. This project exemplifies how targeted research can help scientists understand social behavior and spatial dynamics at different levels. The rapid growth of personal devices that include sensors, such as smartphones, creates potential applications for the techniques used here to investigate how people interact in society.

“Our collaboration with Microsoft Research is very exciting and has allowed us to explore ideas that would have been impossible in a normal university setting. In particular, the development of this new device is opening a whole area of possibilities that we would not normally have been able to consider. Our work on the spatial ecology of highly pelagic seabirds is becoming increasingly relevant for the creation of protected areas and habitats. Our truly interdisciplinary team is allowing us to explore novels ways of investigating these species. Conducting such interdisciplinary research can be complex and challenging but is also extremely rewarding and, I believe, has enormous future potential.” —TIM GUILFORD

“Understanding the movement and behavior of wild animals is fundamental to preserving some of our most important, endangered, and engaging species. Leading this project has been immensely exciting, and the involvement of our collaborators at Oxford has been very rewarding and stimulating. These projects are often complex, and it’s amazing to be working with such a wonderful group of researchers.” —ROBIN FREEMAN

Health and Computer Vision



Understanding the Immune Response to HIV



MARCUS ALTFELD
Director, Program for Innate Immunity
Partners AIDS Research Center (PARC)



DAVID HECKERMAN
Microsoft Distinguished Scientist
Senior Director
eScience Group
Microsoft Research
Connections



CARL KADIE
Principal Research Software Design Engineer
eScience Group
Microsoft Research

ACCORDING TO THE [UNAIDS Report on the Global AIDS Epidemic 2010](#) by the Joint United Nations Programme on HIV/AIDS, HIV killed 260,000 children in 2009 with a disproportionate number of deaths in sub-Saharan Africa. Even in the United States, while no longer a death sentence, HIV requires expensive, life-long treatment. In 2011, investigators from the [Ragon Institute](#) of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University; [Imperial College London](#); the [National Cancer Research Institute](#); and [Microsoft Research](#) showed, for the first time, that the immune system’s natural killer (NK) cells play a direct role in fighting HIV. This knowledge opens a new path of research into ways to beat the virus.

Scientists have long known that NK cells play an important role in the control of viral infections, mounting short-lived but highly toxic assaults on infected cells. It’s logical to expect that NK cells would play a role in the control of HIV infections, and, in fact, various in-vitro and epidemiological studies suggest that NK cells do just that. However, it remained unknown whether NK cells directly

mediate anti-HIV immune pressure inside the human body. The first tell-tale signs that NK cells were affecting HIV were found by using a sophisticated software tool that was developed at Microsoft Research. The tool used almost a CPU-year of computation to sift through millions of possible clues as to how our immune system interacts with this deadly virus. Subsequent clinical and laboratory work that was performed by instructional collaborators resulted in evidence that the virus mutates in response to NK cell activity—by inference, confirming that NK cells play a direct role in fighting HIV. This knowledge opens a new path of research into ways to beat the virus, helping physicians in their long-running battle with HIV and AIDS.

Multiple Microsoft .NET and Windows HPC Server-based Microsoft technologies are being used in this effort to facilitate efficient software development and computing. In addition, the knowledge gained from working with scientists on the complex computational scenarios in this project has helped Microsoft make improvements to Windows HPC Server.



“This study suggests for the first time that NK cells can impose immune pressure on HIV, something that had previously been described only for T cells and antibodies, adding an additional cell to the repertoire of those with anti-HIV activity. The results of this study raise a number of interesting new questions. We need to better understand the molecular mechanisms that allow NK cells to recognize HIV-infected cells and learn how to manipulate these cells in humans for therapy or prevention.” —MARCUS ALTFELD

“Using the full breadth of Microsoft tools, including .NET and Windows HPC Server, allowed us to quickly code our algorithms. The big one-CPU-year run then finished in just over a day.” —CARL KADIE

“Our study provides hope that a greater appreciation of the NK-cell-mediated immune responses to HIV can lead to therapies that interrupt the virus’s evasive processes, thereby giving physicians another weapon in their long-running battle with HIV and AIDS.” —DAVID HECKERMAN

Medical Sensing via a Contact Lens

HAVING SUCCEEDED IN making computers faster, smaller, and cheaper, technologists are now focused on making computing more accessible—more effortlessly integrated into the user’s life—enabling more people to do more interesting things. Over the last few years, Microsoft has been working on creating [natural user interfaces](#) (NUIs) that make interacting with computers seamless, so that people can focus on completing their everyday tasks, building better relationships, and living better lives, even—or especially—when they’re on the go.

[The Functional Contact Lens](#) project is one such NUI endeavor. A collaboration between Babak Parviz and his Bio-Nanotechnology group at the [University of Washington](#) and Desney Tan and his [Computational User Experiences group](#) at Microsoft Research, the project aims to build a contact lens that provides the wearer with a fully configurable display of digital information.

In the initial phases of this project, the team designed and built prototype contact lenses that included an embedded LED display, a wireless data communication link, and a power harvesting unit. While difficult problems remain—for example, adequately focusing the light and dealing with the jitter of the contact lenses—this proof-of-concept is highly encouraging.

Recently, the team took advantage of the fact that the lens comes into contact with bodily fluid—

tears, to be precise—to explore applicability of the lens for continuous medical sensing. Tests show that blood-glucose levels can be measured via special sensors embedded into the lens. This could be a boon to patients with type 1 diabetes, allowing them to monitor their blood-glucose levels without having to jab their fingertips several times a day. Moreover, the lens would monitor glucose levels continuously, a major improvement over the snapshot readings from periodic finger-stick blood draws.

Initially, the lens would record information on blood-glucose levels for review by patients and their physicians. Ideally, however, the lens will be perfected to automatically display alerts of abnormal glucose levels directly in the wearer’s view. Such alerts would prompt the patient to inject insulin or eat a high-glucose snack—and would fulfill the NUI goal of providing seamless computing interaction that improves quality of life.

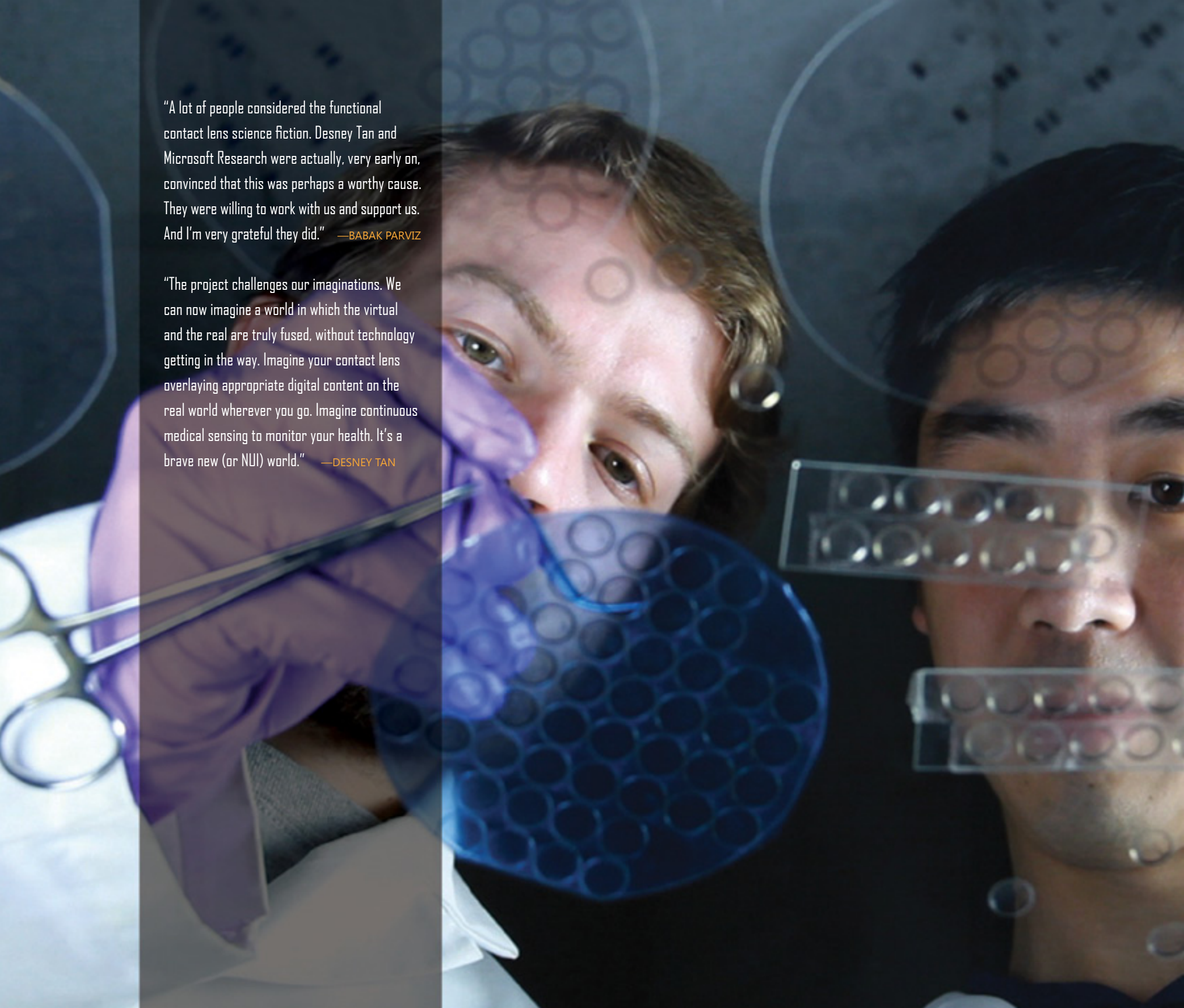
While the project team continues to explore uses of a fundamentally new type of natural user interface based on the functional contact lens, they have developed a number of novel methods for building devices such as sensors, radios, and antennas and integrating them onto a thin, flexible substrate. These methods are of significant interest to researchers in the fields of augmented reality, continuous health, and activity monitoring, and provide inspiration to Microsoft engineers who are developing future NUI technologies.



BABAK PARVIZ
*McMorrow Innovation
Associate Professor
Electrical Engineering
Department
University of Washington*



DESNEY TAN
*Senior Researcher
Computational User Experiences
Microsoft Research*



“A lot of people considered the functional contact lens science fiction. Desney Tan and Microsoft Research were actually, very early on, convinced that this was perhaps a worthy cause. They were willing to work with us and support us. And I’m very grateful they did.” —BABAK PARVIZ

“The project challenges our imaginations. We can now imagine a world in which the virtual and the real are truly fused, without technology getting in the way. Imagine your contact lens overlaying appropriate digital content on the real world wherever you go. Imagine continuous medical sensing to monitor your health. It’s a brave new (or NUI) world.” —DESNEY TAN

Improving Echocardiography

ECHOCARDIOGRAPHY HAS LONG been an important tool for diagnosing heart defects and diseases. However, its usefulness has been hampered by difficulties in quickly and accurately identifying the myocardium (heart muscle) in echocardiograms. Unfortunately, echocardiography images are of relatively low quality, with only about 40 percent of the clinical data considered of sufficient quality for automated analysis. This problem is particularly pronounced in 3-D echocardiography, where the large number of tissues that have similar appearance to the myocardium, including adjacent muscles and bright vessel walls, complicate the discrimination even more. Finally, the sheer amount of the data contained in a 3-D echocardiography study should be processed in a matter of a few seconds—ideally in real time—to be optimally useful in clinical practice. Most, if not all, current methods fail to meet this timeliness criterion.

The [University of Oxford](#) collaborated with [Microsoft Research Cambridge](#) to investigate automated methods for segmenting 3-D echocardiography to assist cardiologists in assessing heart performance. The initial plan was to combine the graph-cuts framework developed at Microsoft Research with 3-D echocardiographic fusion, a technique for improving image quality being developed by Oxford. This plan quickly led to the idea of applying random forests to

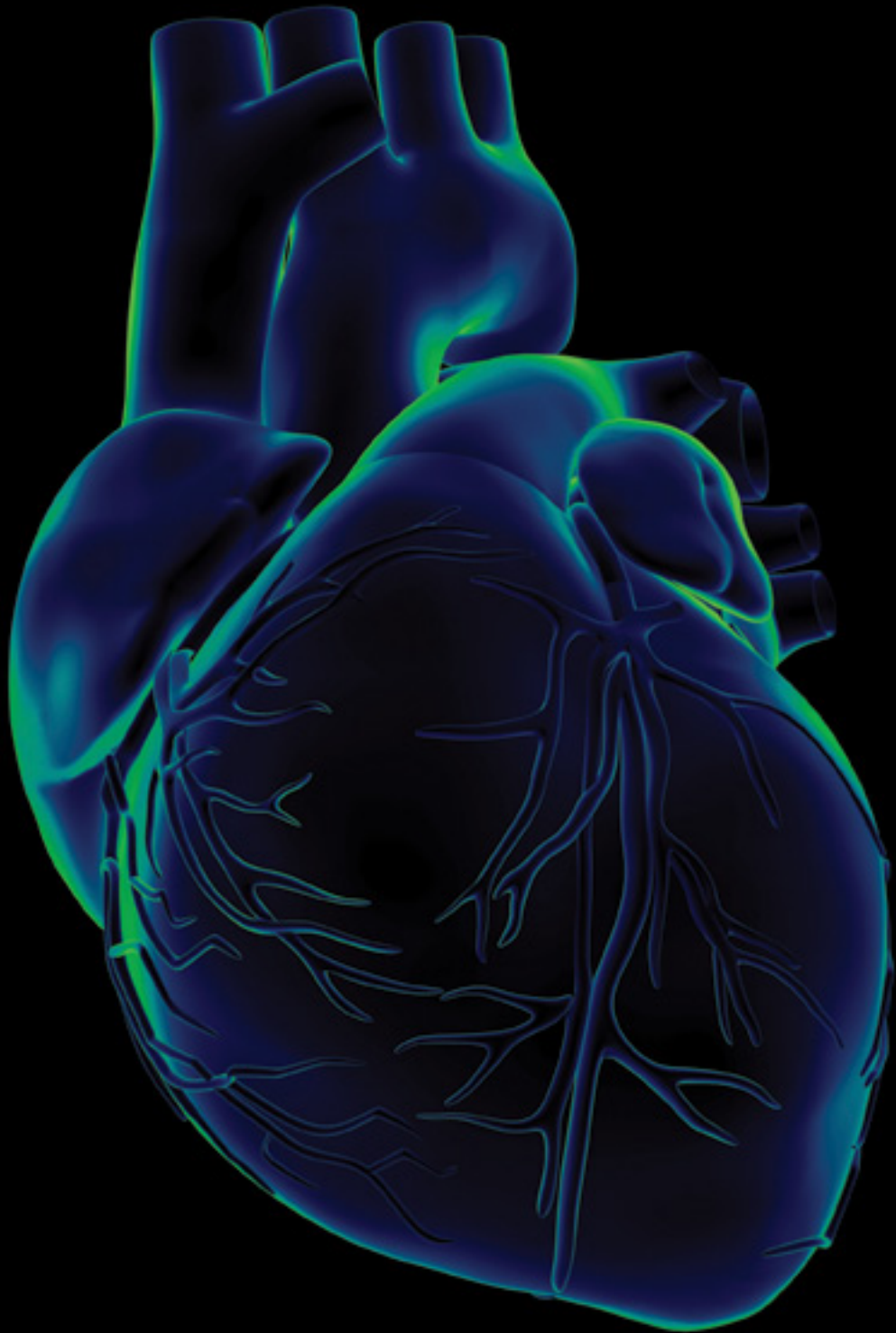
ultrasound image segmentation. (Random forests are discriminative classifiers developed recently in the machine-learning community.) The idea was to investigate whether the power of machine learning and training can lead to good segmentation results on medium-to-low quality data without the need for image fusion.

Recently, researchers have focused on using the temporal features, rather than just the intensity information, in a 3-D random forests framework to improve segmentation accuracy. They are also looking to establish efficient ways to train random forests for 3-D analysis. The project has demonstrated that the use of random forests allows scientists to obtain accurate delineations for the entire 3-D cardiac volume in a matter of seconds on a central processing unit (CPU), or even in real-time on a graphics processing unit (GPU). This class of techniques was also used by Andrew Blake and the team at Microsoft Research to produce the Kinect skeletal tracking technology. It is an example of how fundamental machine-learning research has widespread applications, ranging from medical imaging to consumer devices.

The improvement in the efficacy of echocardiography—a non-invasive diagnosis technique with no known risks or side effects—has major applications in the diagnosis of congenital heart disease and other cardiac conditions.

“The Oxford-Microsoft collaboration has provided a very effective way to combine the strengths of Microsoft in machine learning applied in computer vision with my group’s expertise in cardiovascular image analysis to provide new insight into how to analyze the quite challenging images of varying quality we meet in 3-D echocardiography. We published one of the first early papers in this area, and this has generated interest from a number of other groups around the world, as well as seeded a number of further projects in my laboratory looking at both other application domains, as well as methodological extensions suitable for 3-D and 4-D biomedical image analysis. This has been a great experience of translating research ideas from one research domain to another.” —ALISON NOBLE

“Alison Noble and her team have collaborated with us to apply ideas in machine learning to make a difference in medical imaging. It’s an example of how computer science underpins many of the advanced technologies that are becoming commonplace in hospitals around the world.” —ANDREW BLAKE



ALISON NOBLE
Technikos Professor of
Biomedical Engineering
Department of Engineering
Science
Oxford University



ANDREW BLAKE
Microsoft Distinguished Scientist
Managing Director
Microsoft Research Cambridge

Searching the Human Body



ANTONIO CRIMINISI
Senior Researcher
Computer Vision Group
Microsoft Research



RAJ JENA
Consultant Clinical Oncologist
Cambridge University Hospitals
NHS Foundation Trust

IN RECENT YEARS, great advances in medical technology have improved the accuracy by which medical professionals can diagnose and treat a myriad of ailments. With the ability to literally see inside a patient’s body—by using technologies such as magnetic resonance (MR), computed tomography (CT), and positron emission tomography (PET)—doctors now have a plethora of image information to help them address patients’ conditions. However, while the technology for image acquisition has improved enormously, deciphering the information buried in the pixels is a time-consuming process that is subject to the individual clinician’s experience and skill level.

Antonio Criminisi and his team at Microsoft Research Cambridge are addressing new challenges and opportunities in clinical routines. Through a project entitled InnerEye, Criminisi is combining medical expertise and machine-learning theory to design a system that makes computer-aided diagnoses from medical imagery. By working directly with clinicians, such as those from Addenbrooke’s Hospital in Cambridge, the team has been able to design the system for practical use from the ground up.

Given a CT or MR scan, a clinician’s key challenge is the identification of different body organs. Doctors typically view slices through the scan data, displayed in grey levels, and try to manually inspect different organs. By applying state-of-the-art machine-learning techniques, it is now possible

for the computer to automatically identify dozens of key body organs, as well as their size and location. In effect, the system extracts semantic information about the presence and position of organs from the image pixels. This information can be stored in conventional text files and is searchable. Thus, a doctor can quickly retrieve information about, for example, all patients with an enlarged spleen, or patients with kidney stones.

Criminisi’s team is also developing an efficient algorithm for the automatic detection and delineation of brain tumors and the identification of their constituent regions, such as the actively proliferating cells or the necrotic region. This task enables clinicians to better diagnose the tumor type and determine the best course of treatment.

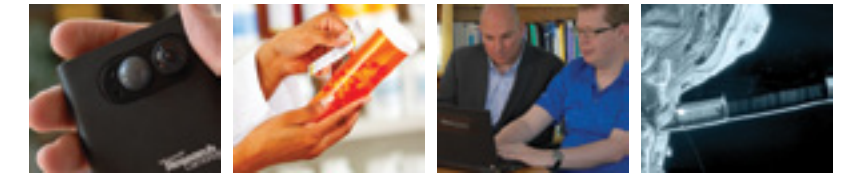
One of the key algorithms used in this work is called decision forests. This algorithm can “learn” discriminative features about human organs or tumors from vast amounts of example data. Interestingly, this technique is also the engine of the Microsoft Kinect skeletal tracking capability, exemplifying how fundamental machine-learning research has uses in both medicine and entertainment. Many of the technologies in the InnerEye project have now been integrated into the Microsoft Amalga health care platform, enabling doctors to more effectively work with patients; by automating some of the most difficult tasks, doctors have more time to focus on patient treatment.

“I work with a team that looks after patients with the most aggressive forms of brain cancer. We rely heavily on image information from CT and MRI scans to build up a picture of the tumor, where it is located, how it is growing, and how it is affecting the surrounding healthy brain tissue. We know that different treatments such as surgery, radiotherapy, and chemotherapy can target different areas of the tumor, which is why we tend to use these treatments in combination. Tracking these different areas of tumor through the course of a patient’s cancer treatment in order to determine the effectiveness of treatment is a non-trivial exercise. Whilst automated techniques have been developed to track the anatomy of the healthy brain, none of these techniques have been able to identify the different areas that make up a brain tumor. The exciting prospect of the InnerEye project is that machine-learning techniques might be able to accelerate, if not fully automate, the process of tumor tracking, allowing us as clinicians to make meaningful and objective assessments of treatments in the clinic for the immediate benefit of our patients.” —RAJ JENA

“Working on the InnerEye project is a fantastic opportunity to do interesting research with a direct impact on people’s lives. It is refreshing to see how modern machine-learning technology is now being employed by doctors to deliver better health care, more promptly. At the same time, medical needs are helping advance the state of the art in machine learning and computer vision research.” —ANTONIO CRIMINISI



Computer Science and Technology



SenseCam



STEVE HODGES
Principal Hardware Engineer
Sensors and Devices Group
Microsoft Research Cambridge



CHRIS MOULIN
Senior Lecturer
Cognitive Neuropsychology
University of Leeds

IMAGINE LIFE WITHOUT daily memories. Not recalling what you did yesterday or the day before. Many victims of Alzheimer’s disease and brain infections don’t have to imagine this memory-deficient existence—they live it. Anything that can help these sufferers retain memories is a godsend, which is why memory researchers greeted Microsoft Research’s [SenseCam](#) with such enthusiasm.

A small digital camera that is typically worn suspended from a neck strap, SenseCam automatically takes approximately two or three digital pictures every minute, in response to built-in sensors and a user-programmable timer. The camera easily stores a day’s worth of images on its internal memory card; these images can then be downloaded to a PC for viewing.

What has amazed memory researchers is the impact of the pictures. They don’t just jog the patient’s memory—studies indicate that they actually help in memory formation and retention. As one neuropsychologist working with the device observed, “Not only does it allow people to recall memories while they are looking at the images—which in itself is wonderful—but after an initial period of consolidation, it appears to lead to long-term retention of memories over many months, without the need to view the images repeatedly.” Moreover, with improved memory recall, patients exhibit greater confidence and higher self-esteem.

Initially developed in 1999, SenseCam was originally conceived as a general-purpose tool

that would aid in capturing visual data. Microsoft Researchers, however, quickly realized it had potential in helping patients with memory impairment and began collaboration with [Addenbrooke’s Hospital](#) in Cambridge, England, a facility renowned for its work with memory-loss patients.

The initial results at Addenbrooke’s led to a host of clinical studies funded by Microsoft Research, including research on SenseCam’s value in treating patients with epileptic amnesia, Alzheimer’s, and severe brain injuries, as well exploring the effect of SenseCam images on memory in healthy people.

This research project was so successful that it led to Microsoft licensing the technology to Vicon, which now distributes it as a product called [Revue](#). As well as being used in the study of memory impairment, the device also has many other possible applications, which include market research, documentary filmmaking, art projects, and classroom instruction. There is now a burgeoning international SenseCam research community comprising academics, clinicians, and practitioners from a variety of fields. More than 100 workshop, conference, and journal papers report on how SenseCam is being used, and there is an annual SenseCam Symposium where the community gathers to exchange ideas and results. It is a compelling example of how a research project has led to a commercially available product that is used in a wide variety of applications that benefit society.

“SenseCam has been a really inspiring project for me to be involved in. It’s amazing to see the technology get picked up by so many different researchers and clinicians all around the world. With so many different applications, it’s gone far beyond what we ever imagined. And in some cases it’s transforming people’s lives, which is incredibly rewarding.” —STEVE HODGES

“If SenseCam only gave you memories for things in the pictures, there would be no surprises, no magic, and no benefit to our research. What continues to amaze me, in my own personal use and in the SenseCam films we’ve made for others in the course of our research, is that SenseCam recovers memories for events, feelings, and details not captured in the lens and displayed on the screen. It’s not so much an *aide-mémoire* as a powerful cue for unlocking memories that you may have thought were lost.”

“For the memory researcher, it is going to be an invaluable tool for rehabilitation and measurement. The most important memory to someone is for their daily lives and events. It has been difficult to measure, observe, and rehabilitate these, because the psychologist can’t follow the person around. SenseCam can go anywhere with the person and thus provides the researcher or clinician with a set of materials fit for rehabilitation and research.” —CHRIS MOULIN



Technology to Combat Counterfeit Products



MANOS M. TENTZERIS
Professor
School of Electrical and
Computer Engineering
Georgia Institute of Technology



DARKO KIROVSKI
Senior Researcher
Machine Learning and Applied
Statistics
Microsoft Research

PRODUCT COUNTERFEITING IS a growing problem for legitimate manufacturers and their customers. A flood of sham products bedevil the software, computer hardware, pharmaceutical, entertainment, and fashion industries—everything from fake designer jeans to phony prescription drugs. The losses to the legitimate producers are substantial, as are the risks to consumers who unknowingly purchase fraudulent goods.

Sometimes buyers know they’re not getting the real deal—a Rolex for \$100 just can’t be real. But all too often, buyers pay full market price for a forgery, thinking they’re getting the genuine item. The latter instances are what we call true counterfeits, and these represent the gravest threats to legitimate manufacturers and the buying public.

A group of scientists, Darko Kirovski and Gerald DeJean at Microsoft and Manos Tentzeris, Vasileios Lakafosis, Anya Traille, Hoseon Lee, and Edward Gebara at the [Georgia Institute of Technology](#) (Georgia Tech), proposed the development of a technologically sophisticated certificate of authenticity (COA), an anti-counterfeiting device whose “signature” is extremely hard to copy but easy and convenient to authenticate. The proposed COA is a digitally signed physical object of fixed dimensions that has a random unique structure. Key among its requirements is that the COA be inexpensive to make and authenticate, but prohibitively expensive to replicate. Using radio-frequency electromagnetic “fingerprints” of dielectric and conductive resonators in the near-

field is the technological basis of the proposed COA.

The proposed technology, referred to as RF-DNA, would satisfy the key COA requirements. Each instance of the COA would cost less than one cent, and the COA reader is projected to cost around US\$100 in mass production. Because of the COA’s complex topology and interdependent fingerprint components, it would be extremely difficult and costly to reproduce illegally. Moreover, it would be resistant to wear and tear, as the fingerprint readout is contactless.

Gray-market piracy (high-quality counterfeits and illegal distribution of copies of software) costs Microsoft approximately US\$1–5 billion in losses annually. Piracy of other goods constitutes approximately 5 percent of all world trade annually, resulting in a serious threat to public safety, equity, job markets, as well as tax losses around the world.

RF-DNA offers a potential means of protecting producers and consumers from counterfeit goods—easily and at a low cost that is comparable with the costs of current, ineffective antipiracy features. Use of the COA could be extended to protect currency, checks, money orders, credit cards, licenses, passports, and a myriad of other legal documents. The anti-counterfeiting technology with cryptographically strong security could save Microsoft and worldwide resellers billions of dollars of revenue annually. Of course, it won’t help you if you buy a \$10 pair of Armani sunglasses—sometimes you just have to remember *caveat emptor*.

“The developed RF-DNA system features unprecedented anti-counterfeiting capabilities, effectively equivalent to ‘security keys’ of thousands of bits, by combining inkjet-printing ideas, ultra-low-cost ‘green’ materials, and near-field RFID technologies. It is a wonderful example of a very successful synergy between GT [Georgia Tech] and Microsoft Research, and it could set the foundation for near-ubiquitous anti-counterfeiting solutions in future Internet of Things or High-Speed Computing/Telecommunication/WSN Systems.”

—MANOS M. TENTZERIS

“RF-DNA comes to market with claims that are unparalleled in the world of anti-counterfeiting. What makes out a real challenge is to explore deployment scenarios as the business and operational landscape is excitingly complicated.”

—DARKO KIROVSKI

Clinical Studies and Data Collection and Reuse

CLINICAL STUDIES DRIVE medical research, but they're complex and costly affairs. In order to be scientifically useful, they must record large amounts of data on the health and treatment regimens of a carefully selected group of test subjects. The IT infrastructure to collect this valuable dataset is often created from scratch for each study, a time-consuming and expensive process that leads to non-standard systems that compile data in non-standard ways.

Could standardized software increase the effectiveness and reduce the cost of clinical studies, and thereby accelerate medical research? Together with colleagues from the CancerGrid project in the United Kingdom, researchers at Microsoft decided to find out.

Building on the work of [CancerGrid](#), which set out to improve information management for large-scale phase III clinical trials, researchers employed a semantics-driven approach and developed models, standards, and software for collecting medical research data. They broadened the scope of the work to include early-phase (phase I and II) studies, in which many more observations—including detailed molecular and imaging data—are made, but in a smaller number of subjects. Despite the importance of early-phase, experimental medicine in the development of new therapies, data from these studies is rarely reused

or combined. The researchers aimed to show how appropriate information systems support can be rapidly provided, facilitating study management and data reuse, at little or no cost to the researchers.

Their project focused on early-phase experimental studies in cancer, but the technology has proven to be widely applicable and is now deployed in institutes around the world. For example, it is being used to support a clinical study on the effectiveness of different ways of administering pneumonia vaccines to children in Nepal. Aside from allowing the trial to be run more cost-effectively, their system is collecting the data in a standardized way that will allow it to be easily reused in future studies.

The storage of clinical trials data required the creation of a robust and highly flexible IT infrastructure that would permit rapid reconfiguration to support different studies while maintaining the consistent data definitions needed to ensure that the data can be reused. [Microsoft SharePoint](#) and [Microsoft InfoPath](#) provided the needed robustness and configurability, demonstrating the value that these standard business technologies can bring to scientific research. The academic researchers shared their experiences with Microsoft product development teams, contributing to the design process for the next generation of [Microsoft Office](#) tools.



JIM DAVIES
Professor
Software Engineering
University of Oxford



SIMON MERCER
Director
Health and Wellbeing
Microsoft Research Connections



“The CancerGrid project is just one example of the value standard Microsoft tools can bring to basic scientific research.” —SIMON MERCER

“Working with Microsoft Research Connections has had a huge impact on our productivity. Our collaboration has made us more effective and more connected researchers. We’ve been able to develop the tools and technologies that our medical colleagues need.” —JIM DAVIES

Microsoft TerraServer: an Imagery Scalability Story

TODAY, ONLINE MAPPING websites enable PC and smartphone users to view traditional maps, aerial or satellite imagery, or “street view” images of their neighborhood, place of work, or vacation destination practically anywhere in the world. These applications did not exist until the late 1990s. Researchers led by database legend [James Gray](#) at Microsoft Research’s Bay Area Research Center in San Francisco, California, pioneered the early work of building a massively large image database of aerial and satellite imagery that standard web browsers can access without the need for special plug-ins or other applications.



GEORGE LEE
*Product and Services Lead
Orthoimagery
US Geological Survey*



TOM BARCLAY
*Development Manager
Bing Search
Microsoft*

The original motivation for the project was to test the scalability of a new version of the [Microsoft SQL Server](#) relational database management system (RDBMS). Working with the SQL Server team, the research project was to build a single database instance that was big (1 terabyte or larger); public (accessible on the Internet); interesting; accessible via standard web browsers (no plug-ins required); real (had a commercial purpose); fast; and easy to use, build, and deploy. Finding an interesting, real, and large dataset that wasn’t already widely available was a challenge. In researching potential datasets, the research team met with researchers at the [University of California, Santa Barbara](#) (UCSB), who were building an online digital corpus of geospatial information. Collaborating with UCSB, the team developed the [Microsoft TerraServer](#) database and web. At the time, circa 1996, the geographic information system (GIS) community was able to store and display street and world maps through web browsers. But the industry had not been able to build and deploy high-resolution imagery such as that found today on sites like Bing Maps or Google Maps.

The website then known as Microsoft TerraServer and now known as [Microsoft Research Maps](#) initially stored 2.3 TB of [U.S. Geological Survey](#) (USGS) grayscale “digital orthophoto quadrangle” (DOQ) imagery and 1 TB of declassified Russian military satellite data that were provided by Sovinformsputnik’s US partner, Aerial Images, Inc. The Microsoft TerraServer researcher’s novel approach was to take the very large images—varying from 25 MB to 200 MB each—and tile them into very small, 200 x 200 pixel JPEG compressed “tiles” that ranged from 6 KB to 36 KB, depending

on the image content. The pixels were selected from the source imagery such that a “seamless mosaic” of large expanses of Earth would appear to the user as a single, large image. The tiles were formatted to HTML tables of 3 x 2, 4 x 3, or 5 x 4 (in other words, the tiles were cut into 6, 12, or 20 pieces). Clickable arrows and zoom-in/out images placed around the table of images enabled users to pan and zoom around each logical collection of seamless tiles. In the conterminous United States, there were 10 logical images enabling a user to continuously pan north to south from Albuquerque, New Mexico, to Montana, or east to west from San Francisco, California, to Reno, Nevada, without changing logical seamless images.

The Microsoft TerraServer tiling and mosaic scheme proved to be a breakthrough in the GIS industry. The Microsoft TerraServer tiling approach is used by all major high-resolution imagery sites, including Google Earth, Google Maps, MapQuest, and Yahoo Maps. [Microsoft TerraServer / Microsoft Research Maps](#) are available to end users and researchers. Today, the site exclusively stores USGS original grayscale DOQ imagery, scanned USGS topographic maps (digital raster graphic, or DRG), and color imagery of major US cities photographed by the National GeoSpatial Agency (NGA) after 9/11. Microsoft continues to provide free and unencumbered access to this data via an HTML user interface, a SOAP/XML web service, and an OpenGIS Consortia compliant web map server. The site enables consumers as well as commercial and academic users to have both interactive and programmatic access to a total of 4.3 terabytes of imagery.

TerraServer was the first mapping service on the Internet with programmatic interfaces. At its release, it was the largest data collection accessible via web services. The TerraServer project generated a significant amount of feedback to the Microsoft SQL team on how to scale databases to large datasets. SQL Server was the foundation for the Virtual Earth technology.



“The TerraServer was the world’s largest public repository of high-resolution aerial, satellite, and topographic data. It was designed to be accessed by thousands of simultaneous users using Internet protocols via standard web browsers.” —TOM BARCLAY

“Orthoimagery is one of the most cost-effective methods for producing current maps, and it has been my dream that orthophotomaps be readily available to everyone. Today, orthoimagery is used in GIS systems at all levels of government and throughout the geospatial community. More importantly, it is readily available to the general public. Microsoft TerraServer and the technology behind it were the beginning steps toward making that dream a reality.” —GEORGE LEE

Physics and Astronomy



WorldWide Telescope and Seamless Astronomy

THE [SEAMLESS ASTRONOMY](#) project, led by Alyssa Goodman at [Harvard University](#), aims to apply new technologies, as they are imagined or developed, to the research environment that astronomy researchers find themselves in every day. Microsoft Research [WorldWide Telescope](#) (WWT) provides a fantastic set of linkages between data archives, data visualization, and online resources, including the refereed literature. By collaborating closely with other leaders of the NASA/NSF-sponsored [Virtual Astronomical Observatory](#) (VAO) program and researchers at Microsoft, participants in the Seamless Astronomy effort are working toward a “seamless” research environment in which researchers use their favorite tools to retrieve data, which is then used to conduct literature searches. Researchers then use other tools in

that environment to analyze, model, and simulate with the data in a freeform effort. The goal is for researchers to be able to focus on their research without being concerned about the various programs and services they use to help them with their work.

WorldWide Telescope is a free, interactive, virtual learning environment that combines terabytes of high-resolution images from ground- and space-based telescopes, astronomical data, and guided tours, enabling users to experience a seamless exploration of the universe. “Anybody who’s looked up at the sky has wondered about the nature of what they are seeing,” says Curtis Wong, a principal researcher at Microsoft Research who worked to bring Turing Award winner [Jim Gray’s](#) WWT vision to reality. “I think it’s a fundamental thing for people to be curious about the universe.”

WWT provides a unique opportunity to see and hear guided tours about the universe from astronomers and educators within the context of the sky. Students are able to pause a tour to get more in-depth information about any object from multiple sources on the web, examine other multispectral imagery of that object, or jump to a different tour that is related to that object. WWT makes this unprecedented level of data and imagery easy to access so that users of all ages can explore the universe.

As Goodman describes, the WWT technology, though still evolving, is just one example of new ways to interact with the large datasets that are accumulated by projects that are affiliated with the research labs, such as the Dataverse, Seamless Astronomy, and High Dimensional Data Visualization and Interaction. While in the past, data has typically been displayed in charts and graphs, in recent years, technology has enabled the collection of such large volumes of data that traditional display methods are inadequate. By displaying the data in three dimensions (with time as a fourth) or by interacting with it graphically, WWT enables researchers to not only better visualize and explain what they’ve found but also understand it better themselves, Goodman observes. “WWT is also one of the most extensive and powerful astronomical data environments that are accessible to the public for education,” she notes.

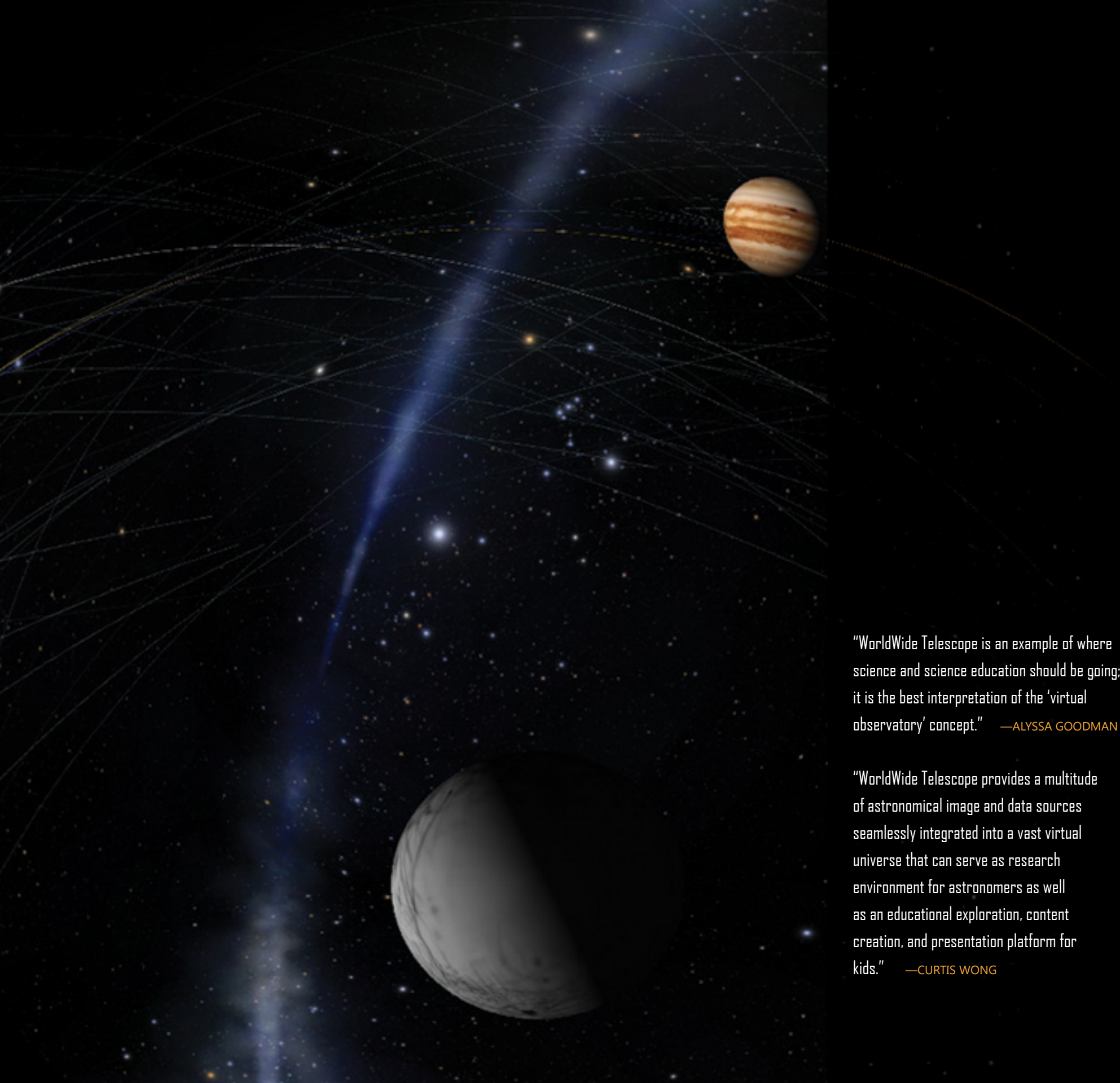
Through this research effort, Microsoft has gained valuable insight into geo-spatial visualization technologies. Researchers envision the data deluge extending beyond the sciences and into the business world. With information feeds and large volumes of streaming data, visualization techniques like the WWT will be valuable not only for research but also for the broader community of knowledge workers and data analysts.



ALYSSA GOODMAN
Professor
Astronomy
Harvard University



CURTIS WONG
Principal Researcher
eScience Group
Microsoft Research Connections



“WorldWide Telescope is an example of where science and science education should be going; it is the best interpretation of the ‘virtual observatory’ concept.” —ALYSSA GOODMAN

“WorldWide Telescope provides a multitude of astronomical image and data sources seamlessly integrated into a vast virtual universe that can serve as research environment for astronomers as well as an educational exploration, content creation, and presentation platform for kids.” —CURTIS WONG

SkyServer: the Universe at Your Fingertips

[SKYSERVER](#) IS THE multi-terabyte astronomy archive, containing the data of the [Sloan Digital Sky Survey](#) (SDSS) project, a collaborative effort among public and private organizations to create the most complete map of the northern sky. It presents the SDSS data as a web-accessible database, along with visual tools to analyze the data. The result is an SQL database with approximately 3 billion rows describing approximately 400 million celestial objects and 1 million spectra. It gives full graphical user interface (GUI) and SQL access to the SDSS data, allowing everyone the chance to use one of the world’s best telescopes. SkyServer includes 200 hours of online instruction to teach astronomy and computational science by using this data.

The ability to pose questions in a few hours and get answers in a few minutes changes the way scientists view the data; they can experiment interactively. When queries take three days and hundreds of lines of code, scientists ask fewer questions and so get far fewer answers. This and similar experiences prove that interactive access to scientific data and data mining tools can

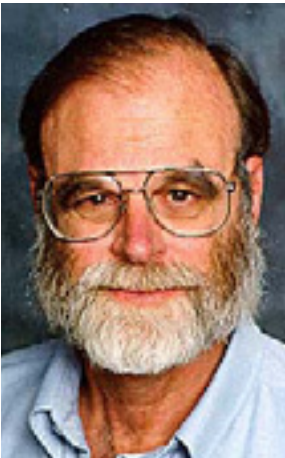
dramatically improve productivity.

The SkyServer is also an educational tool. Several interactive astronomy projects, from elementary to graduate level, have been developed in three languages (English, Japanese, and German). The SkyServer design has been cloned by several other observatories: Royal Observatory, Edinburgh, Cornell Arecibo Pulsar Search, Caltech Quest, Space Telescope Science Institute for Hubble and Galex datasets, and the National Optical Astronomical Observatories at Tucson. Also, SkyServer’s framework has been used as the template for radiation oncology and environmental sensing data.

SkyServer was one of the earliest projects that used web services to access and serve large datasets and image files. With many terabytes of data and imagery, the application explored the scaling capability of [Microsoft SQL Server](#) database technology. The project also inspired a new generation of astro-informaticians to take advantage of the power of relational database technology.



ALEXANDER S. SZALAY
Professor
Astrophysics and Computer
Science
Johns Hopkins University



JIM GRAY
Distinguished Engineer
eScience Group
Microsoft Research

“SkyServer is a good example of information at your fingertips for astronomers—letting scientists work in new, more productive ways. An astronomer can ask complex questions and see results almost instantly. It allows them to explore the data from desktops anywhere in the world.” —JIM GRAY

“With so much information in so many databases around the world, we cannot move all the data to where the analysis is being done; rather, we need to bring the analysis to the data by dividing the computation up among the archives.” —ALEXANDER S. SZALAY

Topological Quantum Computation



MICHAEL FREEDMAN
Director
Microsoft Station Q
Microsoft



CHARLES M. MARCUS
Professor
Physics
Harvard University

WITH THE GROWING demand for faster, more powerful, more versatile computing, and limitations on how small circuits can be built within the current paradigm, scientists inevitably must turn to a new era of intelligent computational devices. Researchers are reaching the limit of how fast and small they can make conventional microprocessors and urgently need to look to possible alternatives.

Quantum computers are one exciting avenue to explore. Research in quantum computing has offered many important new physical insights, as well as the potential of exponentially increasing the computational power that can be harnessed to solve important problems in energy, medicine, computer science, physics, mathematics, and material science. Quantum computing promises, quite literally, a quantum leap in the ability to execute complex quantum simulations across a wide range of applications.

Microsoft is working with universities around the world to develop the first quantum computer—a topological quantum computer. Since 2005, Microsoft’s collaboration with universities has driven a resurgence in condensed-matter physics, in the area of topological phases and materials. The unique basis of this approach to quantum computation is to use topological materials that by

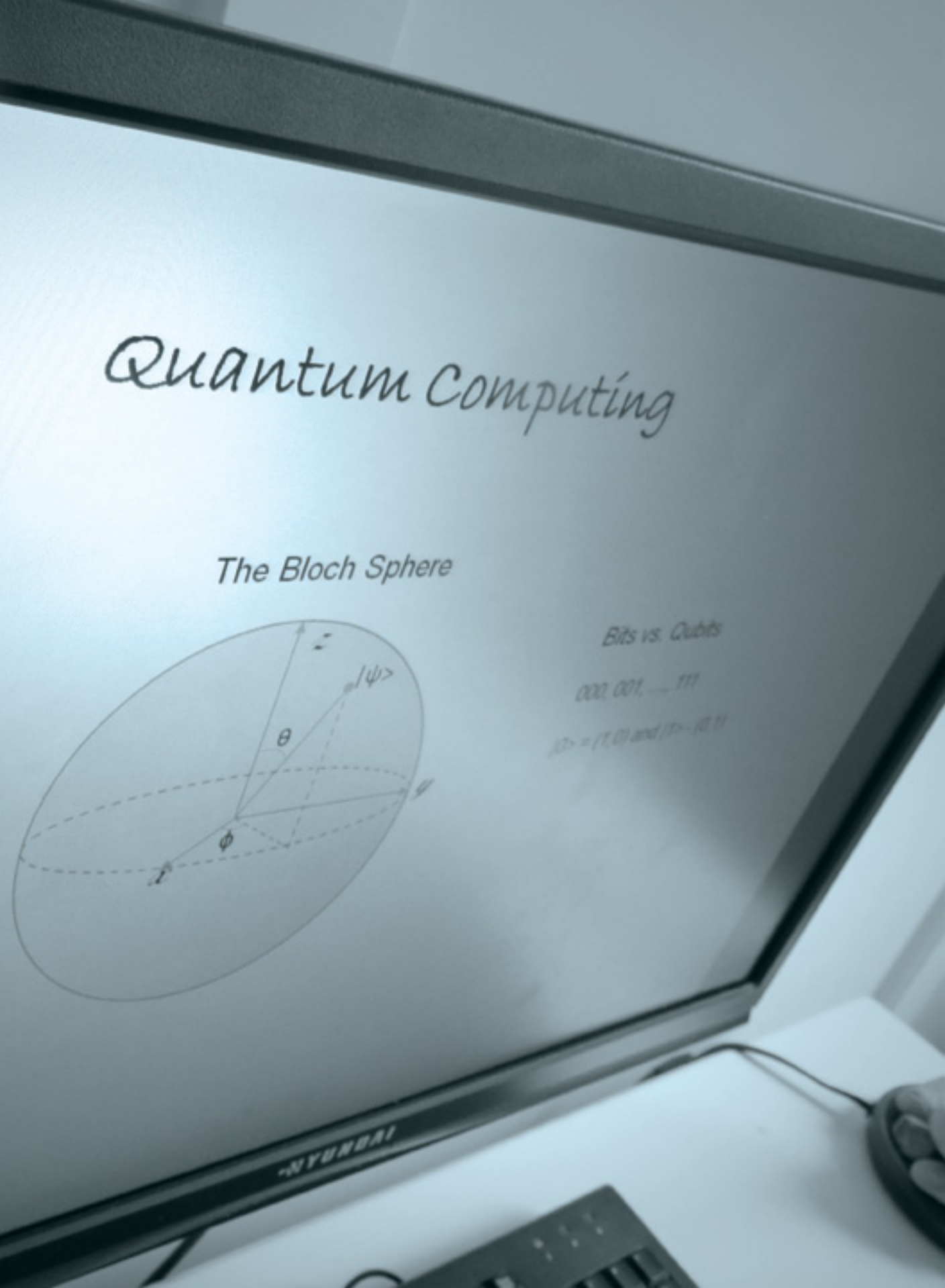
their nature limit errors. These are exotic, low-temperature systems that possess degrees of freedom that are immune to the action of local operators. By their topological nature, individual qubits and quantum gates are protected from errors. Examples of topological materials include fractional quantum Hall effect (FQHE) systems, Ising superconductor-topological insulator heterostructures (ISH), and 1D nanowires.

Among the ongoing external collaborators, Charles Marcus and the Marcus Lab at [Harvard University](#) have made great strides in fabricating and studying 1D quantum nanowires and FQHE systems. They aim to manipulate non-Abelian anyons—strange quantum quasi-particles that appear in two-dimensional systems—for use in quantum information processing, and ultimately for building a full-scale quantum computer. Microsoft’s collaboration with the [Marcus Lab](#) has revealed beautiful, intricate physics and a path toward building a topological quantum computer.

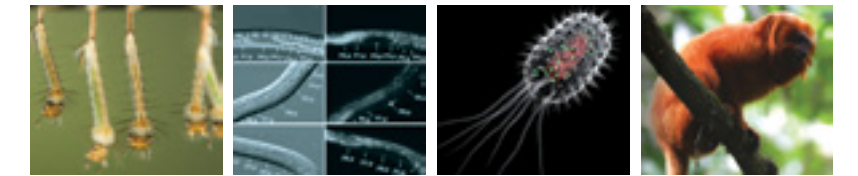
This project demonstrates the strong commitment at Microsoft Research to explore new computing paradigms that have great potential to influence the future of computing as well as the future of Microsoft.

“The mathematicians and theoretical physicists at Microsoft Station Q have created a vision of computation that completely reconsiders how information is processed and stored, intrinsically rooted in the laws of quantum mechanics, which appears poised to bring revolutionary change to technology and physical science. The challenges created for the experimentalists involved in this project are thrilling: first, discover a new class of particle, expected to exist in nature but not yet found; learn to create these particles on demand, move them, store them, wrap them around each other, and read out their states; then build a computer out of them. Perhaps this kind of vision is to be expected. After all, these are topologists—the notion “off on the horizon” means nothing to these people; the whole problem sits before them, all challenges equidistant. What compels the experimentalist to try is something different. Remarkably, each piece of the puzzle seems within reach, hard for sure, but not impossible. So, one step at a time, we proceed, occasionally getting the theoreticians to meet us half way between here and the horizon.” —CHARLES M. MARCUS

“We believe that topological materials, arguably among the most enigmatic objects in nature, appear to be essential to quantum information processing. Our challenge on the topological road is to physically realize the mathematically known topological phases of matter, and then to develop a scalable quantum device. If we can build it, we will be able to provide new answers to new questions and revolutionize our understanding of computation. Topological quantum computation is hard to resist.” —MICHAEL FREEDMAN



Biology and Life Sciences



Coevolution of Viruses and the Immune System



SIMON MALLAL
Director
Institute for Immunology and
Infectious Diseases
Murdoch University



TOMER HERTZ
Staff Scientist
Vaccine and Infectious
Disease Division
Fred Hutchinson Cancer
Research Center



NEBOJSA JOJIC
Principal Researcher
eScience Group
Microsoft Research

INSIDE THE HUMAN body, a battle rages between the immune system and disease-causing pathogens. Striving for an advantage, pathogens constantly evolve to evade detection by the immune system.

In 2010, Microsoft researchers—together with colleagues from [Murdoch University](#) in Western Australia, the [University of Western Australia](#), and [Fundación Ciencia para la Vida](#) in Chile—explored this evolutionary struggle. Their study focused on human leukocyte antigen (HLA) molecules, which sample cellular proteins and present them on the cellular surface for examination by the immune system.

When viruses infect a cell, they bring their own genetic material into the cell and use cellular resources to propagate. As a result, HLA molecules present viral proteins on the infected cell’s surface, spurring an immune attack on the “odd” cells. However, viruses often mutate to evade detection, altering the protein segments that HLA molecules are most likely to present.

On the other side, the distribution of the thousands of HLA variants present in human populations can change over many generations. This sets up an evolutionary game: viruses on one side, the immune system on the other. To analyze this contest, the researchers quantified HLA-binding preferences according to targeting efficiency, a novel measure that captures the correlation between HLA-binding affinities and the genetic conservation in the targeted regions. In theory, HLA molecules should draw attention to

protein segments that are shared across related viral species, as such regions should be functionally important and thus immutable.

Analysis of targeting efficiencies indicated that HLA molecules do indeed prefer to target such conserved regions. The magnitude of this preference varies in a way that shows evidence of target splitting, where two different HLA loci focus on different viral families. This phenomenon is consistent with theoretical biology predictions for predator-prey models and indicates that targeting efficiency as a measure of the HLA-virus links will be useful in analyzing viral evolution. Furthermore, in many cases the host’s total targeting efficiency scores for various viruses correlate with clinical outcomes, offering a potentially useful system of measures for analyzing infection outcomes in individual patients or entire human populations under different conditions, such as post-vaccination or following a previous viral infection.

This work was possible only by combining machine-learning techniques with large numbers of viral sequences. It illustrates that a computational approach can be just as important to biology as “wet lab” work for both formulating and testing new hypotheses. Several new fields of inquiry have stemmed from this work, including research on “correlation sifting,” a method for feature selection that improves upon standard LASSO approaches to a variety of tasks beyond biology, and which may be used to improve future Microsoft products that currently utilize LASSO or similar feature-selection algorithms.

“The findings are fascinating and of fundamental importance. We showed that the immune system focuses its attack on the segments of viruses that are most critical for their function. In response, different viruses have had to find different ways to survive our ability to detect and eliminate foreign forms of life within us. Some viruses have even been able to redirect our immune system attack in a way that enhances their survival.” —SIMON MALLAL

“This is one of the first large-scale analyses that use HLA binding prediction methods for studying host-pathogen co-evolution. The ability to scan a large number of viral genomes and to do this in parallel for a large number of HLA alleles provided compelling evidence that the immune system has evolved to focus surveillance on conserved regions of the pathogens with which it is co-evolving. We found that different HLA alleles have evolved to preferentially target specific viral families, showing additional evidence for the ongoing process of the adaptation of our immune system to the pathogens it encounters.” —TOMER HERTZ

“In this study, we sought to identify possible commonalities in HLA (human leukocyte antigen) binding preferences that would reveal patterns of optimization of this component of the immune system in response to the variation in pathogens. What we found was fascinating. For example, for the first time, we saw that sequence variation patterns in very different viral families show mutual dependencies that can only be detected when these are observed through the prism of our immune responses.” —NEBOJSA JOJIC



Software Verification Meets Biology



JASMIN FISHER
Researcher
Programming Principles and
Tools Group
Microsoft Research



BYRON COOK
Principal Researcher
Programming Principles and
Tools Group
Microsoft Research



NIR PITERMAN
Lecturer
Department of Computer
Science
University of Leicester

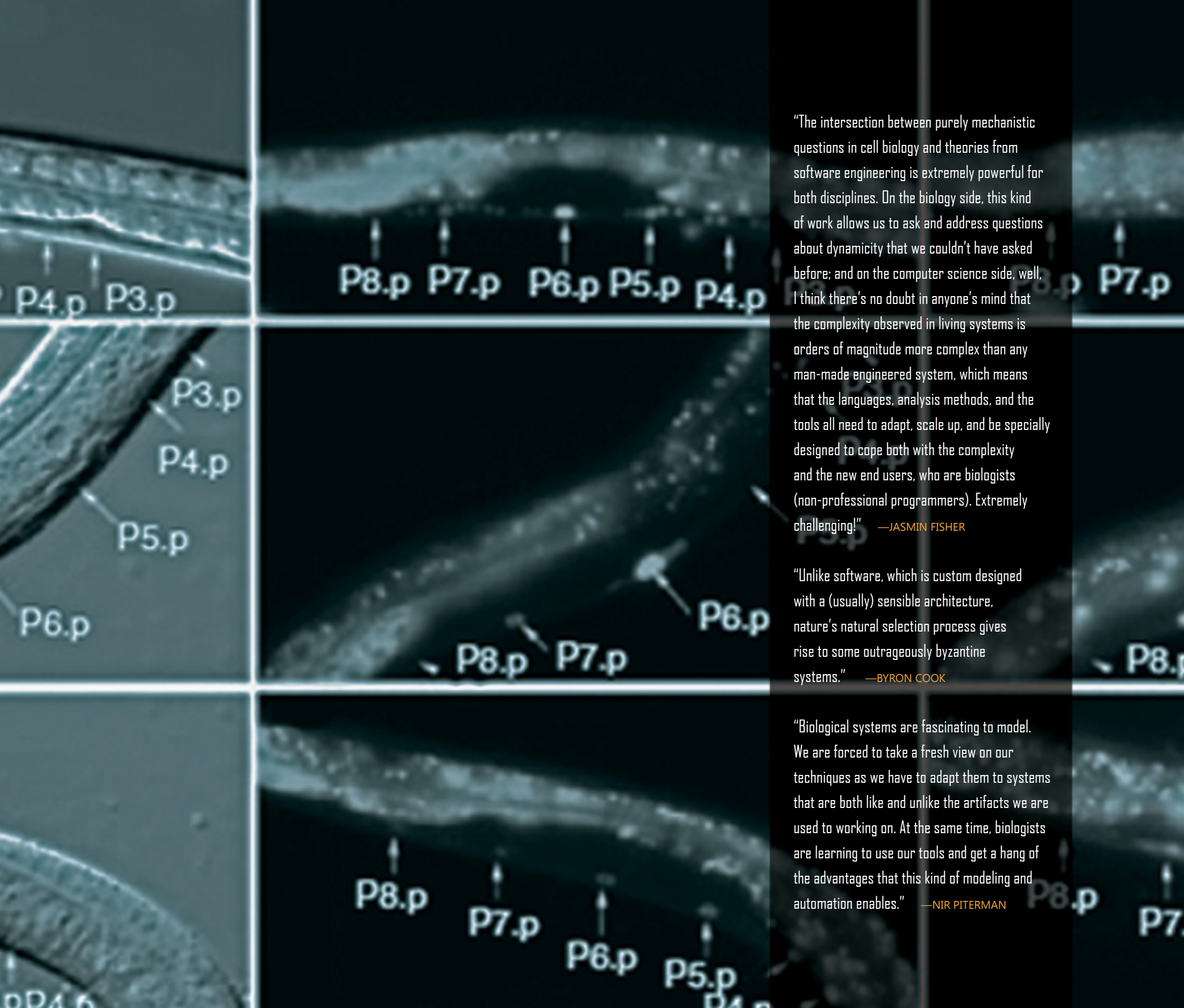
THE TRADITIONAL SCIENTIFIC method—every schoolchild learns it: question, hypothesis, experiment, data analysis, and conclusions. Critical to this process is the construction of models whose predictions can be compared to the real universe, and vice versa. In biology, for example, scientists deal with models of the mechanisms by which embryonic cells differentiate to become one type of cell in the course of cell division. They want to know whether the absence of a particular protein or gene in the cell determines which type of cell the embryonic cell becomes. The problem is that models of biological systems are so complex that it is difficult to extract useful predications from them.

Is there a way to answer queries about intractably large models of biological systems automatically? This is the problem a group of researchers set out to solve at Microsoft Research. They have developed an approach called executable biology, which takes advantage of Microsoft’s strength in automatic software proving and formal methods. Executable biology takes analytical tools developed to study reasoning about computer systems and uses them to model and analyze biological systems. Their solution adapts recently discovered automatic-program verification techniques that were originally developed for the analysis of high-performance concurrent software. In particular, they have applied techniques for proving termination of these systems. Many biological models can be viewed as concurrent systems with far more concurrently executing threads than would be found in an average software system.

Nonetheless, with some adaptation, the techniques have shown great scalability.

In collaboration with Nir Piterman at the [University of Leicester](#), an expert with unique experience in using formal methods in biological modeling, the researchers are developing an efficient procedure for proving stabilization of biological systems modeled as qualitative networks or genetic regulatory networks. For scalability, their procedure uses modular proof techniques, where state-space exploration is applied only locally to small pieces of the system, rather than the system as a whole. It exploits the observation that, in practice, the form of modular proofs can be restricted to a very limited set. For completeness, their technique falls back on a non-compositional counterexample search.

By using this procedure, they have analyzed a number of challenging examples, including a 3-D model of mammalian epidermis and a model of metabolic networks operating in type 2 diabetes. In cases where previous stabilization-proving techniques were known to succeed, this technique attained results far faster, and it obtained new results in cases where tools had previously failed. Biologists can now ask much deeper questions about their models and expect to have tools that automatically answer them. By applying Microsoft Research’s software analysis expertise and tools to a biological scenario, this project is helping to create new business opportunities and partnerships in areas such as biotechnology and health care.



“The intersection between purely mechanistic questions in cell biology and theories from software engineering is extremely powerful for both disciplines. On the biology side, this kind of work allows us to ask and address questions about dynamicity that we couldn’t have asked before; and on the computer science side, well, I think there’s no doubt in anyone’s mind that the complexity observed in living systems is orders of magnitude more complex than any man-made engineered system, which means that the languages, analysis methods, and the tools all need to adapt, scale up, and be specially designed to cope both with the complexity and the new end users, who are biologists (non-professional programmers). Extremely challenging!” —JASMIN FISHER

“Unlike software, which is custom designed with a (usually) sensible architecture, nature’s natural selection process gives rise to some outrageously byzantine systems.” —BYRON COOK

“Biological systems are fascinating to model. We are forced to take a fresh view on our techniques as we have to adapt them to systems that are both like and unlike the artifacts we are used to working on. At the same time, biologists are learning to use our tools and get a hang of the advantages that this kind of modeling and automation enables.” —NIR PITERMAN

Programming Life

EACH ONE OF the 10^{17} cells that make up an individual, each cell that makes up a plant, each stem cell, and even a “simple” bacterium or a white blood cell, is a remarkable biological “machine.” At the heart of these biological machines is a molecular “program” governing sophisticated biological computation, information processing, and decision making—from energy production and consumption, to effective response to attack and malfunction, to what actions need to be taken in response to changes in the environment. It’s a program that determines, for example, plant growth and agricultural yield, which in turn affects global carbon cycling. When this program malfunctions, it can lead to autoimmune disease, cancer, and viral infections.

What if it were possible to program cells? Scientists could address some of the greatest challenges facing humanity. For instance, it might be possible to program human cells or even the entire immune system to prevent or tackle disease in new ways that would transform medicine. Plant cells could be programmed to improve crop yields, to solve a problem for which there is currently no viable solution: how to feed a global population of 9 billion people. It might even be possible to design and program artificial cells to

cheaply generate sustainable, global sources of energy, for example, by artificial photosynthesis. This isn’t science fiction, but, until very recently, nor has it been “science fact.” While the current state of the art has shown how bacteria can be modified to attack specific types of cancer cells, how yeast can be modified to make the world’s most effective antimalarial drug, and how bacteria can be engineered to convert sunlight into electricity, there is a major barrier to making progress: designing cellular behavior currently involves complex, laborious, and highly error-prone methods, often based on trial and error.

This work has not only pushed the boundaries of biological computing, it has also helped Microsoft refine [Microsoft Visual Studio](#) products. In the process of building a visual tool for scientists, the research team used Microsoft Automatic Graph Layout (MSAGL), which provides an essential part of the tool’s user interface, and is part of Visual Studio. With scenarios that pushed the limits of MSAGL’s capabilities, the researchers helped the Visual Studio team expand its capabilities—and also provided direct input into making it more robust. This teamwork between Microsoft researchers and product development groups simultaneously helps advance science and improvements to Visual Studio technology.



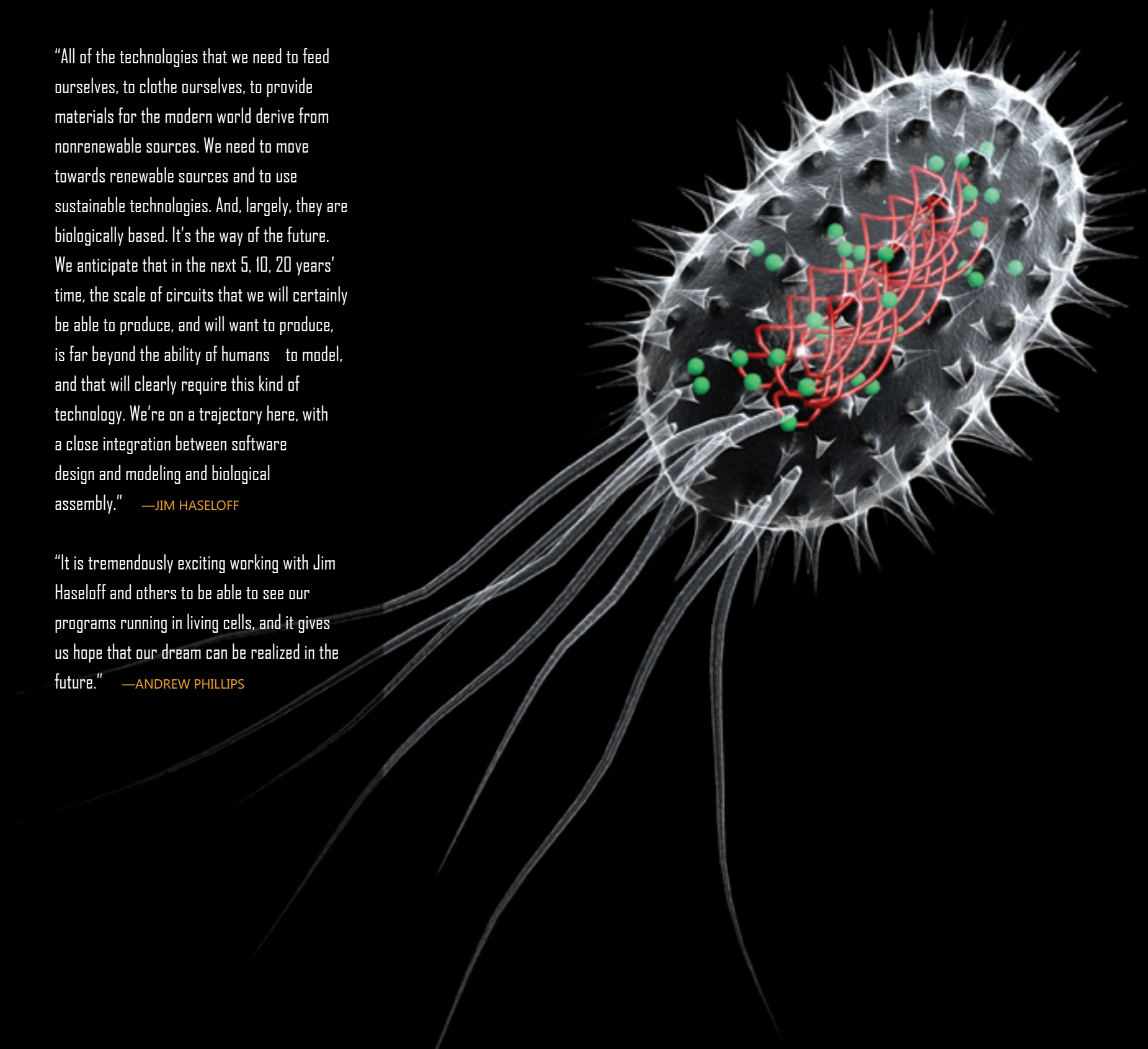
JIM HASELOFF
Plant Biologist
Department of Plant Sciences
University of Cambridge

ANDREW PHILLIPS
*Head of Biological
Computation Group*
Microsoft Research Cambridge



“All of the technologies that we need to feed ourselves, to clothe ourselves, to provide materials for the modern world derive from nonrenewable sources. We need to move towards renewable sources and to use sustainable technologies. And, largely, they are biologically based. It’s the way of the future. We anticipate that in the next 5, 10, 20 years’ time, the scale of circuits that we will certainly be able to produce, and will want to produce, is far beyond the ability of humans to model, and that will clearly require this kind of technology. We’re on a trajectory here, with a close integration between software design and modeling and biological assembly.” —JIM HASELOFF

“It is tremendously exciting working with Jim Haseloff and others to be able to see our programs running in living cells, and it gives us hope that our dream can be realized in the future.” —ANDREW PHILLIPS



How Many Species Are There?



STUART PIMM
*Doris Duke Chair of
Conservation Ecology
Nicholas School of the
Environment
Duke University*



LUCAS JOPPA
*Scientist
Computational Ecology and
Environmental Science Group
Microsoft Research Cambridge*

HOW MANY SPECIES are there? While the question has an almost childlike simplicity, the answer has proven elusive. Biologists agree that the list of known species—numbering about 2 million—is woefully incomplete. Estimates of how many more species exist range from 5 to 50 million, a practically meaningless span.

Biologists have long sought to identify areas where effective conservation could save the most species. Biodiversity hotspots—places with extreme rates of habitat loss as well as unusually high numbers of endemic species—are priorities. But with so many species as yet unknown, one has to wonder: could and would their discovery change those priorities?

Moreover, those unknown species are likely to have small geographic ranges and to be rare within their habitats, and thus they would be prime candidates for extinction. So by figuring out how many species are “missing” from the record, Lucas Joppa, an ecologist in the [Computational Science Laboratory](#) at Microsoft Research—together with Stuart Pimm, a world-leading expert in conservation ecology at [Duke University](#), and David Roberts at [University of Kent](#)—are providing quantitative estimates for how many more species might be threatened and endangered.

The unique insight that the team brings is the acknowledgment of an inherent social dimension to the process of species description. People (taxonomists) describe species, so the number of species described must surely depend on the number of people actively describing them. By incorporating human effort into their statistical

model, the researchers can predict, with measured confidence, the numbers of species remaining to be discovered.

Pimm and Joppa defined and refined the species model from a quantitative viewpoint, providing deep insight into how one might potentially account for unknown species by using a novel proxy parameterization around the number of taxonomists in a given field. Then, by using new scientific software tools and technologies being developed by Microsoft’s Computational Science Laboratory, they are taking massive amounts of hugely dispersed data, bringing them together in a computationally powerful manner, and applying statistical models to make predictive assessments of the total number of species.

The new approach to conservation science is helping to push the boundaries of current spatial database technology. The project uses [Microsoft SQL Server](#) in novel ways that change how scientists think about analyzing spatial data, helping to bring new insights across a range of application scenarios. The researchers are working with the product team to push the limits of SQL Server’s capabilities, while their feedback is contributing to the development process for future versions, including SQL Azure.

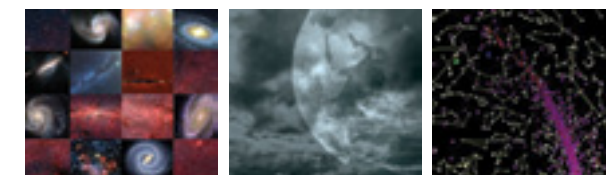
The future of human life depends upon the consequences of the massive extinction crisis currently facing humanity. The unknown-species estimates present quantitative evidence that can improve the effectiveness of environmental regulation and conservation—and help the planet.



“We know we have an incomplete catalogue of life. If we don’t know how many species there are, we may be underestimating or overestimating how fast species are becoming threatened or extinct, and if we don’t know where they live, then how can we prioritize places and policy for conservation? What if the places we are ignoring now turn out to be amongst those with the most unknown species? Trying to answer these great unknowns keeps me excited about the work I do, because it is not only hugely challenging intellectually, but the answers we are seeking are so important for addressing some of the most dire environmental problems of our generation.” —LUCAS JOPPA

“Understanding biodiversity—how much there is, where it is, how fast we are losing it (and what we can do to prevent that loss)—is a major environmental challenge for this new century. What we know already constitutes a vast amount of information, data stream in from around the globe daily, and yet we know we must make decisions on incomplete data, while modeling where the data gaps lie. Microsoft’s commitment to handling, visualizing, and modeling environmental data makes it a leading player in this effort.” —STUART PIMM

Citizen Science



GalaxyZoo

ONE OF THE most successful citizen-science efforts, [Galaxy Zoo](#) enlists individuals worldwide to assist in the classification of galaxies. It represents the world’s largest astronomical collaboration, bringing professional astronomers together with hundreds of thousands of volunteers. Independent assessments show that classifications provided by Galaxy Zoo volunteers are as accurate as those from professional astronomers. The results have informed research into the formation of elliptical galaxies and provided a sample of merging galaxies of unprecedented breadth and fidelity. Some classifications have highlighted the dominance in some environments of red spirals, in which

star formation has been rapidly and mysteriously extinguished. The influence of Galaxy Zoo stretches beyond the band of astronomers who seek to understand the evolution of the universe. The [Zooniverse](#), which grew out of Galaxy Zoo, now hosts 10 projects inviting volunteers to do everything from transcribing ancient papyri to searching for planets around other stars. Meanwhile, the rich datasets that citizen science of this sort provides also inspire new approaches to machine learning—something that will be essential to enable automated and human classifiers to cope with the next generation of surveys, which will produce terabytes of data every night.

“Galaxy Zoo’s initial success took us by surprise, stretching our server capacity and overwhelming the science team. Collaboration with Microsoft Research helped us understand how to make the most of the project’s potential.”

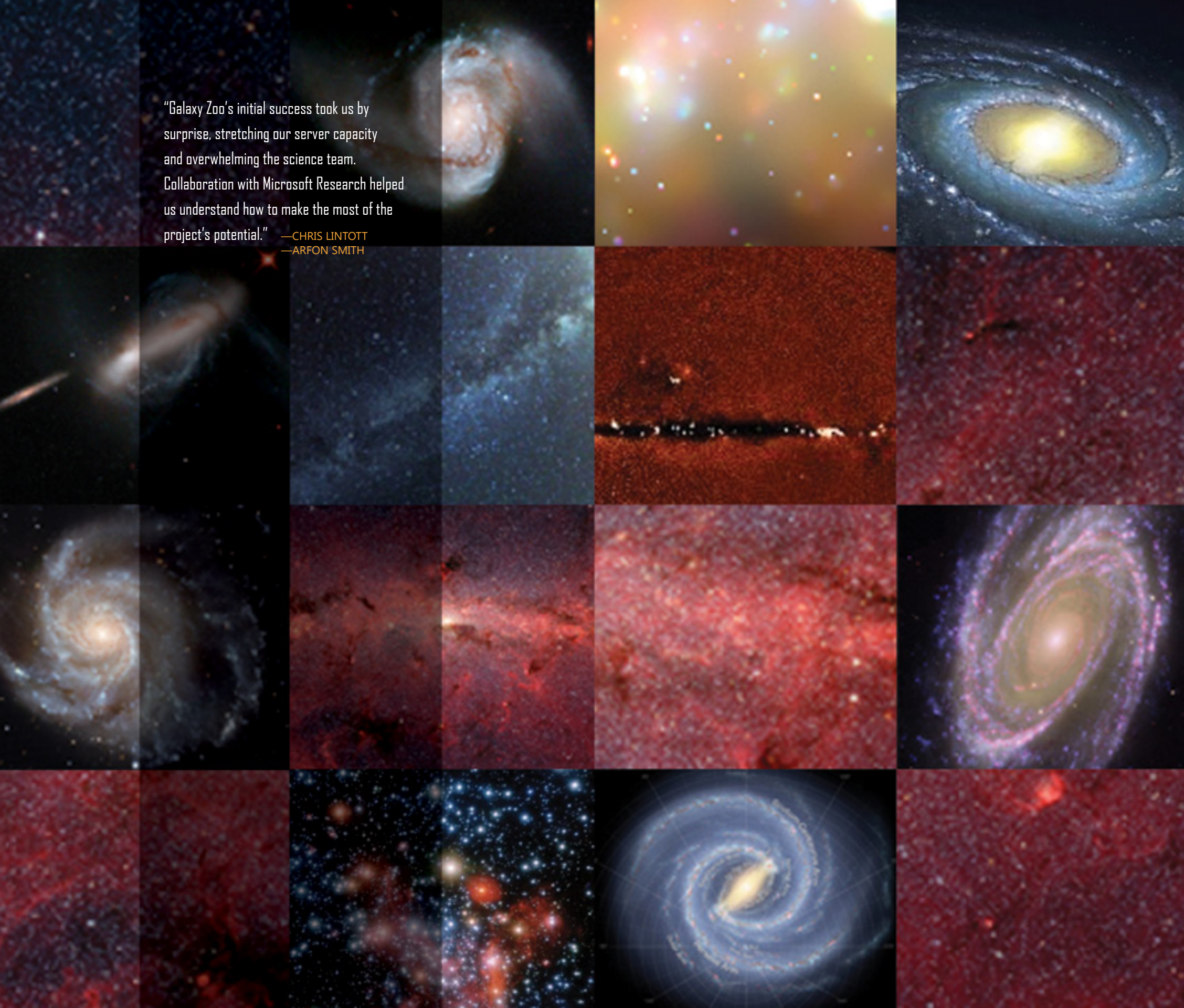
—CHRIS LINTOTT
—ARFON SMITH



CHRIS LINTOTT
Researcher
Department of Physics
University of Oxford



ARFON SMITH
Director
Citizen Science
Adler Planetarium



Weatherathome.net

EVIDENCE OF GLOBAL climate change is now unequivocal, but what does that mean for the weather where you live? Will damaging weather events increase, or might your weather become less extreme? To answer these questions, climateprediction.net has partnered with the [Met Office](#) (the United Kingdom’s national weather service) to create [weatherathome.net](#), an international project that is open to anyone with a computer and Internet access.

With support from Microsoft Research, weatherathome.net will enable anyone in the world to download and run a regional climate model on their home computer. The model is initially available for three target regions: Europe, the western United States, and southern Africa. Participants produce simulations that will enable scientists to estimate how often heat waves, floods,

and hurricanes will strike in the next few decades. The initiative will also indicate how much of the blame for these events can be attributed to greenhouse gas emissions caused by humans.

The model has been developed by the Met Office, and results from different regions are being used directly by scientists who specialize in the climates of those regions. The European region is being analyzed by the Met Office and by Oxford, Edinburgh, and Leeds universities; southern Africa by the University of Cape Town; and the western United States by Oregon State University. Results are also made available to scientists who are interested in climate impacts in the various regions.

The first results from the weatherathome.net experiment were recently published in the journal, *Geophysical Research Letters* (February 2012).

“The support that climateprediction.net has received from Microsoft Research for developing the weatherathome project has been invaluable and is very gratefully acknowledged. Microsoft’s support has been vital in enabling a whole new modeling capability that will much more accurately inform decision-making processes in the face of uncertain future changes to weather—and in particular, extreme weather patterns.” —SUZANNE ROSIER

“Many scientists and decision makers, seeking to understand the consequences of global climate change, have used climate modeling results to describe future climates. Previously, descriptions of future climates have had to choose between a fairly high number (100+) of simulations using global models with coarse spatial resolution (100+km) missing important features like mountain ranges, or finer spatial resolution from regional models but only one or two simulations, which is not enough to describe the probability of different future climates. For instance, one model might suggest an increase in winter precipitation of 10 percent and another suggests a decrease of 5 percent. Weatherathome.net offers an ideal way to achieve both the fine spatial resolution and the high number of simulations that scientists and decision makers desire. In addition to Microsoft Research, the western US part of the project is jointly supported by both science- and decision-oriented agencies: US Geological Survey, Bureau of Land Management, and the California Energy Commission.” —PHILIP W. MOTE



SUZANNE ROSIER
Research Scientists & Joint
Coordinator
climateprediction.net



PHILIP W. MOTE
Professor
College of Oceanic and
Atmospheric Sciences
Oregon State University

Rosetta@Home

ALTHOUGH MOST CITIZEN-SCIENCE experiments allow volunteers to observe the work that their computer is performing on behalf of the project, there has been no mechanism to visually present the overall results and the aggregate contributions of individual volunteers or teams. Thanks to a first-of-its-kind feedback system collaboratively developed by Microsoft Research and the [University of Washington](#) (UW), that’s no longer the case.

The system was created for the UW’s [Rosetta@Home](#) project, which uses the distributed computing power of citizen participants to help predict and design the three-dimensional structures of natural and synthetic proteins by using minimum energy calculations. In Rosetta@Home, an individual computational run generates a folded protein conformation, along with two key metrics: energy (Rosetta score of the computed protein structure, which is analogous to free energy) and RMSD (root mean square deviation from the experimentally determined structure). A typical experiment consists of generating a large number of conformations in an effort to find the

lowest energy structure, which, ideally, should also have the lowest RMSD. By using [Microsoft SQL Server](#), the researchers created an Internet-based feedback report that illustrates the contributions of individuals or teams, displayed in the context of overall results. A SQL Server relational database tracks individual results and processes reporting queries on demand, and its associated reporting services render plots from query results—graphics that can be fully integrated with the project’s public website. The feedback reports have been extremely popular with participants, who find it reinforcing to see the actual impact of their contribution. Moreover, the reporting system has also proven to be useful to investigators inside the UW lab, allowing them to monitor the progress of experiments and view results quickly.

Because user engagement is critical to the success of community computing projects like Rosetta@Home, there’s every reason to believe that this reporting solution can encourage participation in a variety of citizen-science projects.



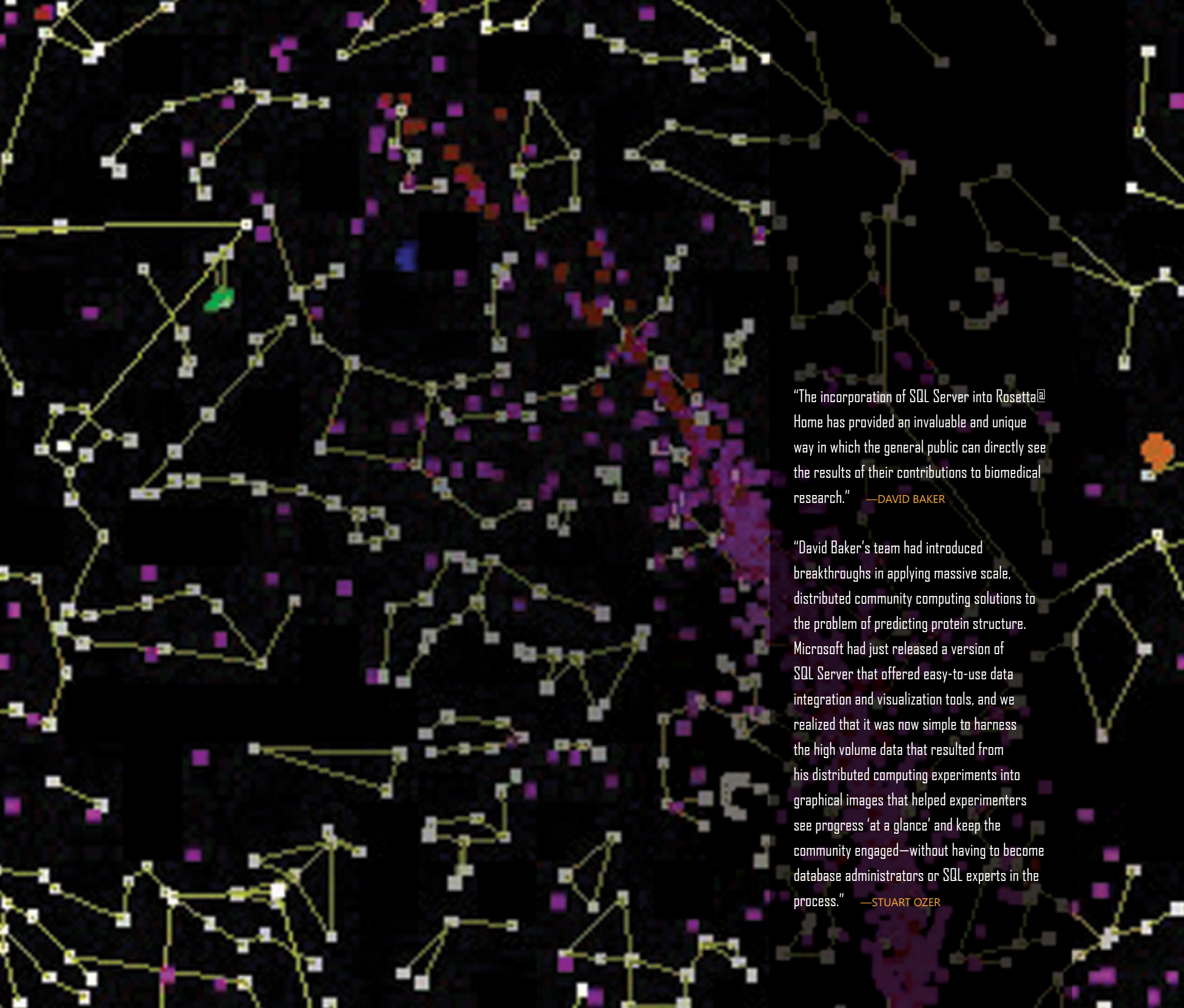
DAVID BAKER
Professor
Biochemistry
University of Washington



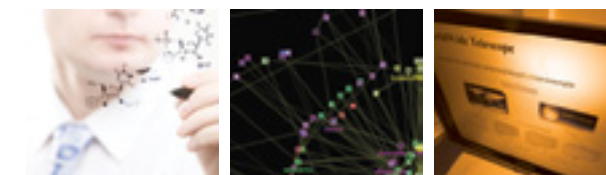
STUART OZER
Partner Group Program
Manager
Microsoft Server and Tools
Microsoft

“The incorporation of SQL Server into Rosetta@Home has provided an invaluable and unique way in which the general public can directly see the results of their contributions to biomedical research.” —DAVID BAKER

“David Baker’s team had introduced breakthroughs in applying massive scale, distributed community computing solutions to the problem of predicting protein structure. Microsoft had just released a version of SQL Server that offered easy-to-use data integration and visualization tools, and we realized that it was now simple to harness the high volume data that resulted from his distributed computing experiments into graphical images that helped experimenters see progress ‘at a glance’ and keep the community engaged—without having to become database administrators or SQL experts in the process.” —STUART OZER



Scientific Tools



Chemistry Add-in for Word

THIS POWERFUL TOOL simplifies the authoring of chemical information in [Microsoft Word](#). The [Chemistry Add-in for Word](#) project was developed in collaboration with Peter Murray-Rust and Joe Townsend from the University of Cambridge's [Unilever Centre for Molecular Science Informatics](#).

The add-in harnesses a chemistry-specific extensible markup language (XML)—Chemical Markup Language (CML)—that allows scientific information to be captured and expressed more easily at the authoring stage. The Chemistry Add-in for Word enables the scientific and academic research community to author chemical content, represent it in a variety of ways, and include the data behind those structures—right

in a Word document. The add-in makes chemistry documents open, readable, and easily accessible to humans and computers.

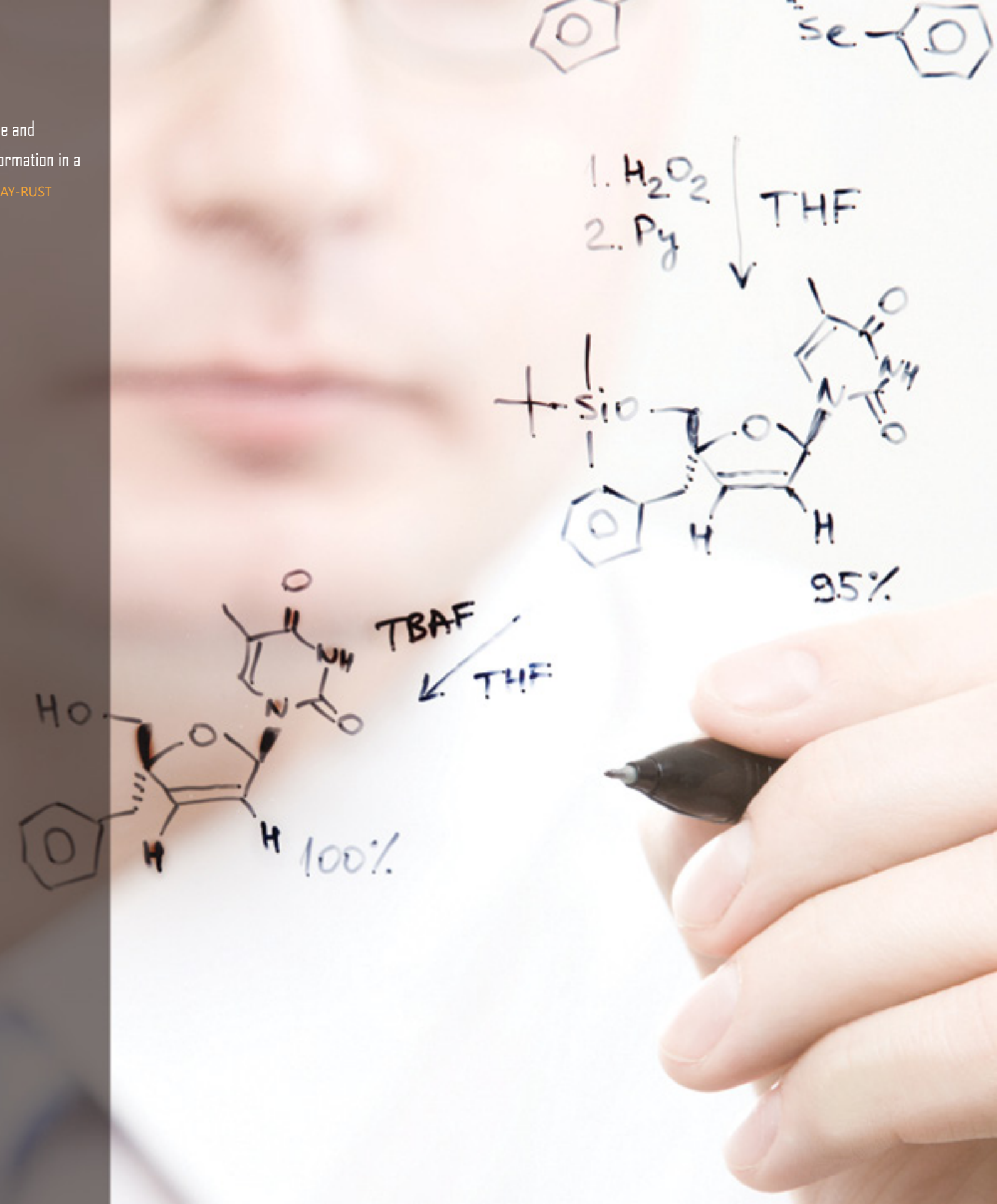
Scientists, researchers, and students can use the Chemistry Add-in for Word to insert chemical structures directly into Word documents; render print-ready, two-dimensional diagrams; contribute to simpler pre-publication processes and richer information discovery scenarios; preserve chemical information for archival purposes; and store and expose chemical information in a semantically rich manner.

The Chemistry Add-in for Word is available to scientists, researchers, students, and the general public to facilitate the sharing of scientific information.



PETER MURRAY-RUST
Researcher
Unilever Center for Molecular
Science Informatics
University of Cambridge

"Chemistry Add-in provides a simple and flexible way to include chemical information in a Word document." —**PETER MURRAY-RUST**



NodeXL

[NODEXL](#) IS A POWERFUL and easy-to-use interactive network visualization and analysis tool that uses [Microsoft Excel](#) for representing generic graph data, performing advanced network analysis, and visual exploration of networks. NodeXL supports multiple social network data providers that import graph data (nodes and edge lists) into Excel. The import features of NodeXL explore social media by pulling data from personal email indexes on the desktop, Twitter, Flickr, YouTube, Facebook, and web hyperlinks.

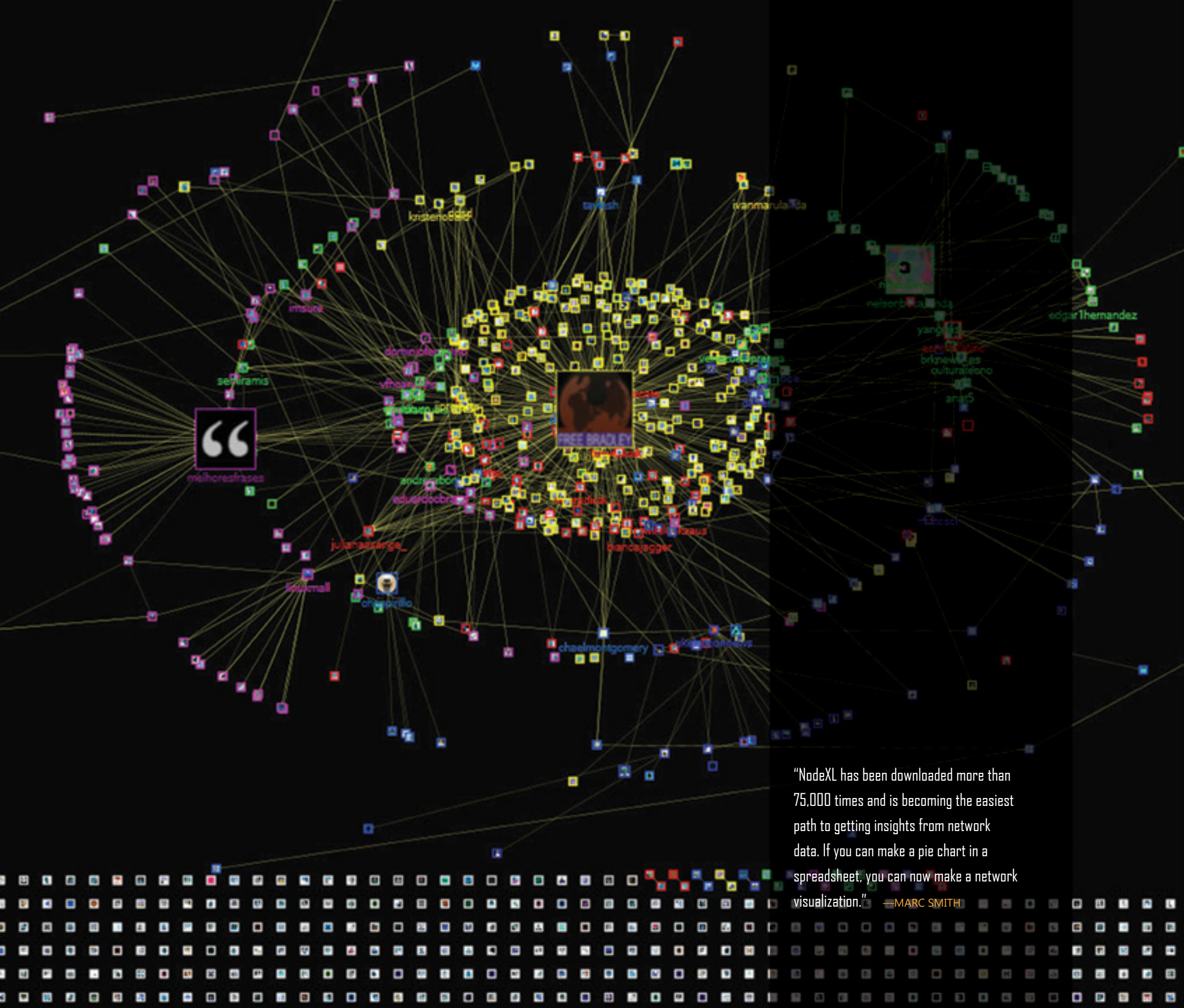
NodeXL allows non-programmers to generate

useful network statistics and metrics quickly and create visualizations of network graphs. Filtering and display attributes can be used to highlight important structures in the network.

NodeXL is a project from the Social Media Research Foundation with support from Microsoft Research Connections and Microsoft Research Cambridge, with additional contributions from researchers at the [University of Porto](#), [University of Maryland](#), [Connected Action Consulting](#), [Stanford University](#), [Oxford University](#), [Australian National University](#), and [Illinois Institute of Technology](#).



MARC SMITH
Chief Social Scientist
Connected Action
Consulting Group



“NodeXL has been downloaded more than 75,000 times and is becoming the easiest path to getting insights from network data. If you can make a pie chart in a spreadsheet, you can now make a network visualization.” —MARC SMITH

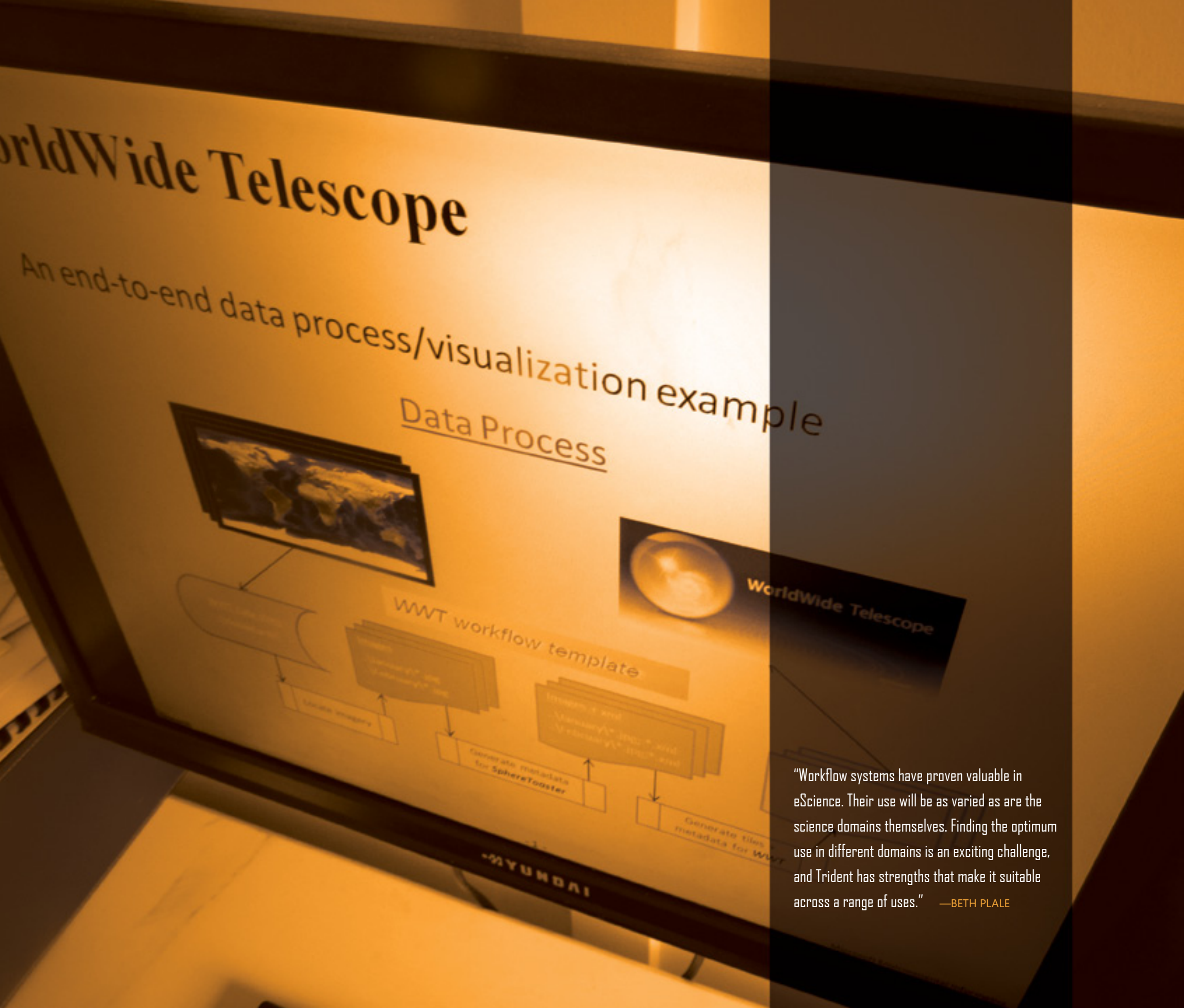
Scientific Workflow Workbench

PROJECT TRIDENT: [Scientific Workflow Workbench](#) is an open-source toolkit based on the [Windows Workflow Foundation](#) that addresses scientists’ need for a flexible, powerful way to analyze large, diverse datasets. It includes graphical tools for creating, running, managing, and sharing workflows.

As part of Vortex2, a project funded by the National Science Foundation to study tornados, Project Trident was used by Indiana University to invoke and manage hourly workflows. By using a 512-core HPC super computer at the university, the workflow automatically obtained current weather data and then generated a short-term forecast, ultimately producing 240 forecasts and more than 9,000 weather images for researchers to access by their smartphones in the field.



BETH PLALE
Director
Data to Insight Center
Indiana University



“Workflow systems have proven valuable in eScience. Their use will be as varied as are the science domains themselves. Finding the optimum use in different domains is an exciting challenge, and Trident has strengths that make it suitable across a range of uses.” —BETH PLALE

Biographies



CHRIS BISHOP is a distinguished scientist at [Microsoft Research Cambridge](#), where he leads the Machine Learning and Perception group. He is also vice president of the [Royal Institution of Great Britain](#) and professor of computer science at the [University of Edinburgh](#). He is a fellow of the Royal Academy of Engineering, a fellow of the Royal Society of Edinburgh, and a fellow of [Darwin College Cambridge](#). His research interests include machine learning and its applications.



ADNAN CUSTOVIC is a professor of allergy and a head of the Respiratory Research Group at the [School of Translational Medicine](#) at the [University of Manchester](#). He is a world-leading authority on the primary prevention of allergic diseases, and he focuses on the relationship between allergic disease and the environment and primary prevention of asthma and allergic diseases. He is currently secretary of the British Society for Allergy and Clinical Immunology and president of the European Academy of Allergy and Clinical Immunology.



JOHN WINN is a senior researcher in the [Machine Learning and Perception](#) group at [Microsoft Research Cambridge](#) and co-creator of the Infer.NET inference framework. His main research interests are in machine learning, machine vision, and computational biology.



RICHARD DURBIN is joint head of human genetics at the [Wellcome Trust Sanger Institute](#) and leader of the Genome Informatics group. He leads a team of researchers that has worked on evolutionary probabilistic methods and is interested in population genome sequencing. He is involved in projects ranging from SGRP yeast sequence variation and the TreeFarm database of animal gene families to new methods for handling genetic variation data and the ACEDB genome database.



JENNIFER CHAYES is distinguished scientist and managing director of [Microsoft Research New England](#) in Cambridge, Massachusetts, which she co-founded in July 2008. Before this, she was research area manager for mathematics, theoretical computer science, and cryptography at [Microsoft Research Redmond](#). She joined [Microsoft Research](#) in 1997, when she co-founded the [Theory Group](#). Her research areas include phase transitions in discrete mathematics and computer science, structural and dynamical properties of self-engineered networks, and algorithmic game theory.



CHRISTIAN BORGES is deputy managing director of the [Microsoft Research New England](#) lab in Cambridge, Massachusetts. He is also an affiliate professor of mathematics at the University of Washington. Since joining Microsoft in 1997, he has become one of the world leaders in the study in phase transitions in combinatorial optimization, and more generally, in the application of methods from statistical physics and probability theory to address problems of interest to computer science and technology. He is one of the top researchers in the modeling and analysis of self-organized networks, such as the Internet, the World Wide Web, and social networks.



RICCARDO ZECCHINA is professor of theoretical physics at the [Politecnico di Torino](#) in Italy. His interests are in topics at the interface between statistical physics and computer science. His current research activity is focused on combinatorial and stochastic optimization, probabilistic and message-passing algorithms, and interdisciplinary applications of statistical physics (in computational biology, graphical games, and statistical inference).



DAVID HECKERMAN is a Microsoft distinguished scientist and senior director of the [eScience](#) group at Microsoft Research. His research interest is focused on learning from data. The models and methods he uses are inspired by work in the fields of statistics and data analysis, machine learning, probability theory, decision theory, decision analysis, and artificial intelligence. His recent work has concentrated on using graphical models for data analysis and visualization in biology and medicine with a special focus on the design of HIV vaccines.



JENNIFER LISTGARTEN is a researcher in the [eScience](#) group at Microsoft Research. Her work focuses on the development and application of statistical and machine-learning methods for the analysis of high-throughput, biologically-based data.



CHRISTOPH LIPPERT is a graduate student at the [Max Planck Institutes](#) for Developmental Biology and for Intelligent Systems in Tübingen, Germany. In summer 2012, he will join the [eScience](#) group at Microsoft Research. His research focuses on the development of probabilistic models in genomics. He has contributed analysis tools that assist biologists with their research in genetic mapping of human diseases and the analysis of genetic and phenotypic variation in plants and other model organisms.



DEB AGARWAL is a senior scientist, Advanced Computing for Science departmental head, and the Data Intensive Systems group lead at the [Lawrence Berkeley National Laboratory](#). She is also working with the [Berkeley Water Center](#) at the University of California, Berkeley, where she has been leading the team developing advanced cyber infrastructure for geosciences since 2005. Her projects involve research, development, and deployment of computing technologies to support collaborative scientific research. Her current research focus is on the development of data server infrastructure to significantly enhance data browsing and analysis capabilities.



DENNIS BALDOCCHI is a professor of biometeorology at the [University of California, Berkeley](#). His research focuses on the physical, biological, and chemical processes that control the exchange of trace gases between ecosystems and the atmosphere. His current projects include a long-term study on carbon dioxide, water vapor, and energy exchange of an oak savanna and annual grassland in the foothills of California, and measurements of methane exchange from rice, a drained peatland, and a restored wetland in the Sacramento-San Joaquin Delta.



MARTY HUMPHREY is an associate professor in the [Department of Computer Science](#) at the University of Virginia. He has been on the faculty of the University of Virginia since 1988. His research interests include the use of cloud computing for cancer research and for environmental research, such as using Windows Azure to model and manage large watershed systems.



YOUNGRYEL RYU is assistant professor of environmental ecology at [Seoul National University](#). He was a main contributor to the project that simulated the breathing of the biosphere by using BESS computations on the MODIS Azure cloud while he was an intern working with Catharine van Ingen at Microsoft Research in 2010. His current research topics include land-atmosphere interactions and the application of computing resources for data-intensive science.



CATHARINE VAN INGEN is partner architect in the Microsoft Research [eScience](#) group. Her research explores how commercial software and tools can be used to enable synthesis science in environmental research science. A key challenge in such studies is addressing not only very large datasets from satellites and ground sensors, but also the small, irregular, ancillary, and categorical data that are necessary for scientific understanding.



JAMES HUNT is professor of [civil and environmental engineering](#) at University of California, Berkeley. A fundamental challenge in all instances of his research topics is how to deal with vast and widely distributed data.



LYNN GLADDEN is pro-vice-chancellor for research for the [University of Cambridge](#). She is the Shell Professor of Chemical Engineering, and head of the Department of [Chemical Engineering and Biotechnology](#), where she leads the activities in the Magnetic Resonance Research Centre. She is a fellow of both the Royal Society and Royal Academy of Engineering, and was appointed Commander of the Order of the British Empire (CBE) in 2009. She is a member of the council of the Engineering and Physical Sciences Research Council.



ANDREW BLAKE is a Microsoft distinguished scientist and managing director of [Microsoft Research Cambridge](#). He is a fellow of the [Royal Academy of Engineering](#), the Royal Society, and the [IEEE](#). He received the Royal Academy of Engineering Silver Medal in 2006, the IET Mountbatten Medal in 2007, and the Royal Academy of Engineering MacRobert Award in 2011 with his colleagues for their machine-learning contribution to Microsoft Kinect human motion capture. His research interests include probabilistic principles of computer vision software, with applications to motion capture, user interface, image editing, remote collaboration, and medical imaging.



TIM GUILFORD leads the [Oxford Navigation Group](#) in animal behavior at the Department of Zoology, Oxford University. His research explores the mechanisms and processes of animal navigation and movement, principally in avian systems. He has pioneered a number of techniques for understanding avian navigation, leading to new insights and understanding. Recently, he has developed a variety of methods for investigating highly pelagic seabirds, including endangered and vulnerable species.



ROBIN FREEMAN is a researcher in [computational ecology](#) at Microsoft Research Cambridge and a research fellow in CoMPLEX (the Centre for Mathematics and Physics in the Life Sciences and Experimental Biology) at [University College London](#). He is also a research associate with the Animal Behaviour group at [Oxford University](#). His research focuses on topics at the interface between behavior, ecology, and computation. He is particularly interested in the application of computational techniques for analyzing animal behavior, from systems to autonomously record the behavior of animals in the wild to the development and application of machine-learning techniques to analyze these data.



MARCUS ALTfeld is the director of the Program for Innate Immunity at the Partners AIDS Research Center (PARC) and the [Division of Infectious Diseases](#) at the Massachusetts General Hospital in Boston, and he is an associate professor at [Harvard Medical School](#).



CARL KADIE is principal research software design engineer in the [eScience](#) group at Microsoft Research. He is also an affiliate of the Machine Learning and Applied Statistics (MLAS) and Adaptive Systems and Interaction (ASI) groups at Microsoft Research. His research interests include creating practical machine learning algorithms for recommendation, spam detection, and, most recently, biology.



BABAK PARVIZ is the McMorro Innovation Associate Professor in the Electrical Engineering Department and the associate director of the Micro-Scale Life Sciences Center at the [University of Washington](#). His research at the interface between biology and electrical engineering has led to a rewarding collaboration with Microsoft Research to develop functional contact lenses by using novel methods in computer science and engineering.



DESNEY TAN is a senior researcher at Microsoft Research, where he manages the [Computational User Experiences](#) group in Redmond, Washington, as well as the Human-Computer Interaction group in Beijing, China. He was honored as one of MIT Technology Review's 2007 list of 35 innovators under 35 for his work on brain-computer interfaces, and he was named one of SciFi Channel's Young Visionaries at TED 2009, as well as one of Forbes' Revolutionaries—radical thinkers and their world-changing ideas—for his work on [whole body computing](#).



ALISON NOBLE is the Technikos Professor of Biomedical Engineering in the [Oxford University Department of Engineering Science](#) and a fellow of St Hilda's College, Oxford. She is a director of the [Biomedical Image Analysis \(BioMedIA\) Laboratory](#), a multi-disciplinary research group working in the area of biomedical imaging and image analysis. She is a senior member of the [IEEE](#), a fellow of the [IET](#), and a fellow of the [Royal Academy of Engineering](#) (FREng).



ANTONIO CRIMINISI is a senior researcher in the [Computer Vision](#) group at Microsoft Research Cambridge. His current research interests are in the area of visual metrology and 3-D reconstruction from single and multiple images with application to computer vision, visual arts, and art history.



RAJ JENA is a consultant clinical oncologist for [Cambridge University Hospitals NHS Foundation Trust](#), visiting fellow in the [Faculty of Engineering and Physical Science at the University of Surrey](#), and a visiting scientist at the [European Centre for Nuclear Research](#) (CERN). His research interests are in the application of advanced imaging techniques and radiotherapy treatment to improve outcomes for patients with central nervous system tumors.



STEVE HODGES leads the [Sensors and Devices](#) group at [Microsoft Research Cambridge](#), which applies its skills across both hardware and software layers to deliver compelling new user experiences. The ultimate goal of the group is to better understand how advances in technology will impact traditional computing and the ways in which people use and interact with computing devices. His personal research focuses around sensing, ubiquitous computing, and new technologies for display and interaction.



CHRIS MOULIN is a senior lecturer in cognitive neuropsychology at the [University of Leeds](#), United Kingdom. His current research interests focus on neuropsychological impairments of memory. In particular, he is interested in the interaction of executive function and long-term memory.




MANOS M. TENTZERIS is a professor with the [Georgia Tech School of Electrical and Computer Engineering](#) (GT-ECE). Currently, he is the head of GT-ECE Electromagnetics Technical Interest Group. Also, he is the head of the [ATHENA](#) (Agile Technologies for High-performance Electromagnetic Novel Applications) research group and has established academic programs about highly integrated/multilayer packaging for RF and wireless applications that use ceramic and organic flexible materials, plastic/paper-based RFIDs and sensors, inkjet-printed electronics and antennas, RF nanostructures, SOP-integrated (ultrawideband, multiband, conformal) antennas, “green” RF electronics, wearable RF, power scavenging, wireless power transfer and RF biomonitoring, and implantable devices.





DARKO KIROVSKI has been a researcher at Microsoft Research since 2000, first in the Crypto and Anti-Piracy group and now in the [Machine Learning and Applied Statistics](#) group. His research interests include: intellectual property protection and embedded system design, in particular, code compression for embedded processors, debugging using combined emulation-simulation, engineering change, symbolic debugging, design watermarking, and forensic engineering.


more biographies ...

Biographies

 **JIM DAVIES** is a professor of software engineering in the [Department of Computer Science](#) at the University of Oxford, and he directs a program of advanced, professional education in software engineering, teaching advanced techniques to people working full-time in the industry. He is leading the development of semantics-driven technology for medical research and electronic governance and a related program of work into the automatic generation of systems from reusable models of structure and functionality.


 **SIMON MERCER** is the director of the [Health and Wellbeing](#) group for Microsoft Research Connections. He has a background in zoology and worked as director of software engineering at Gene Codes Corporation before moving to Microsoft Research in 2005. In his current role at Microsoft Research, he manages a range of collaborations between Microsoft and academia in the area of health care research. His interests include bioinformatics, synthetic biology, translational medicine, and the management of scientific data.


 **TOM BARCLAY** is partner architect and development manager of Bing Imagery Technologies and the Bing Search team at Microsoft. He has been the lead researcher on [Microsoft TerraServer](#) since 1996. His interests include database design, scientific computing, and data mining.


 **GEORGE LEE** is the [US Geological Survey](#) (USGS) product and services lead for orthoimagery. His use of the orthoimagery method during his collaboration with Microsoft Research to develop Microsoft TerraServer exemplifies the successful use of real-world scenarios to challenge and advance computing technologies.


 **ALYSSA GOODMAN** is professor of astronomy at [Harvard University](#) and a research associate of the [Smithsonian Institution](#). With her research group at the [Harvard-Smithsonian Center for Astrophysics](#) in Cambridge, Massachusetts, she studies the dense gas between the stars. She has worked with Microsoft Research for several years on the WorldWide Telescope project. Her innovative use of WWT in research (such as with the [Seamless Astronomy](#) project) and science education (for example, the [WorldWide Telescope Ambassadors Program](#)) has significantly helped the Microsoft Research team expand the capabilities of WWT to assist researchers and educators with their work.


 **CURTIS WONG** is principal researcher with [Microsoft Research Connections](#), focusing on interaction, media, and visualization technologies. He has received more than 25 patents in areas such as interactive television, media browsing, visualization, search, gaming and learning. An amateur astronomer since childhood, he was the driving force in bringing the WorldWide Telescope project to reality. He gives credit to colleague and Turing Award winner Jim Gray, who encouraged him to build his dream of creating an environment for scientific research as well as public education in astronomy. He is particularly proud that millions of people from every continent on the earth are using WorldWide Telescope.


 **JIM GRAY** was a distinguished engineer in the Microsoft Research Bay Area [eScience](#) group. His long career with Microsoft, Digital Equipment, Tandem Computers, and IBM produced seminal work in relational database management systems, transaction processing systems, and the sciences. His vision and work on applying computing technologies to data-intensive sciences inspired the collection of insightful essays in [The Fourth Paradigm: Data-Intensive Scientific Discovery](#).


 **ALEXANDER S. SZALAY** is professor of astrophysics and computer science at the [Johns Hopkins University](#). He is a cosmologist whose work spans a broad area from astrophysics to statistics and computer science. He collaborated with Jim Gray to extend the ideas from the SkyServer to the Virtual Observatory and to other scientific disciplines. He currently works on building large scientific databases and data-intensive computing.


 **CHARLES M. MARCUS** is a professor of physics at [Harvard University](#) and former scientific director of the Center for Nanoscale Systems. His main research interest is experimental condensed matter physics, including solid-state quantum information processing, nanofabrication of electronic devices, and electron transport. Much of his ongoing research has focused on clean, ballistic semiconductor structures, such as chaotic quantum dots, with more recent work emphasizing novel fabrication approaches and systems, effects of electron spin, measurements of electron decoherence, and potential applications of nanostructures to quantum information and quantum computing.


 **MICHAEL FREEDMAN** is the director of [Microsoft Station Q](#) and a technical fellow at Microsoft. His main research interest is topological states of matter and the construction of mathematical models which illuminate these. He leads the Microsoft Station Q team, which is the Microsoft Research group that is working on topological quantum computing by combining research from math, physics, and computer science. He received the Fields Medal, the highest honor in mathematics, for solving the long-standing Poincaré conjecture in four dimensions. He has also received numerous other awards including the Oswald Veblen Prize in Geometry, a MacArthur Fellowship, and the National Medal of Science.


 **SIMON MALLAL** is director of the [Institute for Immunology and Infectious Diseases](#) at Murdoch University and a clinical immunologist and immunopathologist at [Royal Perth](#) Hospital. He has had a longstanding research interest in the major histocompatibility complex (MHC) and genetic influences on clinical outcomes in HIV and other diseases. More recently, he has focused on viral adaptation to HLA-restricted immune responses and the implications of this for HIV vaccine immunogen design.


 **TOMER HERTZ** is currently a staff scientist in the [Vaccine and Infectious Disease Division](#) at the [Fred Hutchinson Cancer Research Center](#) in Seattle, Washington, working on computational immunology. His general research interests are in machine learning and computational immunology. He was a main contributor to this research project during his postdoctoral training in the Microsoft Research eScience group, working with Nebojsa Jojic.


 **NEBOJSA JOJIC** is a principal researcher in the eScience group at Microsoft Research. He is interested in machine-learning approaches to building compact representations of natural signals that reveal patterns of scientific and practical importance. His interest in computational immunology was spurred by similarities that he found between types of data and inference tools used in biology and engineering.


 **JASMIN FISHER** is a researcher in the [Programming Principles and Tools](#) group at Microsoft Research. She is one of the founders of the field of executable biology and a leader in the area of formal methods in biology.


 **BYRON COOK** is a principal researcher in the [Programming Principles and Tools](#) group at Microsoft Research. He is a leader in automatic program verification. He pioneered the research on practical proofs of termination of programs making termination proofs a useful and practical tool.


 **NIR PITERMAN** is a lecturer in the [Department of Computer Science](#) at University of Leicester. He worked with the team to develop the analysis techniques, in particular suggesting an important enhancement that further accelerated the performance of the analysis.


 **ANDREW PHILLIPS** is head of the [Biological Computation Group](#) in the Computational Sciences Laboratory at Microsoft Research Cambridge. His research is in developing visual programming languages and tools for simulating and analyzing complex models of biological systems. He was recipient of a prestigious MIT Technology Review TR35 award in 2011.


 **JIM HASELOFF** is a plant biologist working at the [Department of Plant Sciences](#), University of Cambridge, where he leads a synthetic biology lab. His scientific interests are focused on the engineering of plant morphogenesis by using microscopy, molecular genetics, and computational and synthetic biology techniques.


 **LUCAS JOPPA** is a scientist in the [Computational Ecology and Environmental Science](#) group in the Computational Science Lab at Microsoft Research Cambridge. He leads the conservation research unit, which develops and accelerates better, predictive, actionable, and systemic conservation science, tools, and technologies in areas of societal importance His research covers all aspects of the conservation spectrum, from gathering the next-generation data necessary for better scientific understanding to developing new predictive methods and models.


 **STUART PIMM** is Doris Duke Chair of Conservation Ecology at the [Nicholas School of the Environment](#) at Duke University and one of the most cited scientists working in the field of conservation biology. He was the recipient of the Dr. A. H. Heineken Prize from the Royal Netherlands Academy of Arts and Sciences in 2006 and the Tyler Prize for Environmental Achievement in 2010.


 **ARFON SMITH** is the director of citizen science at the [Adler Planetarium](#) in Chicago and technical lead of the [Zooniverse](#). Before joining the Zooniverse, he earned his doctorate in astrochemistry from the [University of Nottingham](#) in 2006 and then worked as a senior developer in the production software group at the [Wellcome Trust Sanger Institute in Cambridge](#).


 **CHRIS LINTOTT** is a researcher in the [Department of Physics](#) at the University of Oxford, where he leads a team of scientists, educators, and developers in the development of citizen science projects. Passionately committed to scientific outreach, he is best known as co-presenter of the BBC's long-running [Sky at Night](#) series.


 **SUZANNE ROSIER** is a research scientist, joint coordinator of [climateprediction.net](#), and coordinating author of the associated climateeducation.net project. She helped launch climateprediction.net's first regional modeling initiative, "weatherathome," which is modeling limited areas of the world in sufficient detail to enable scientists to deduce future changes in extreme weather patterns. She is now helping to develop this experiment for the Australia/New Zealand region—due for launch in mid-2012—and will be analyzing the results as they become available.


 **PHILIP W. MOTE** is a professor in the [College of Oceanic and Atmospheric Sciences](#) at Oregon State University, director of the Oregon Climate Change Research Institute for the Oregon University system, and director of the [Oregon Climate Service](#). His current research interests include scenario development, regional climate modeling with a superensemble generated by volunteers' personal computers, and adaptation to climate change.

 **DAVID BAKER** is a professor at the [University of Washington](#), an investigator of the [Howard Hughes Medical Institute](#), and a member of the [National Academy of Sciences](#). He is a world leader in protein structure prediction and design. He leads the Baker Laboratory with a research focus on the prediction and design of protein structures and protein-protein interactions.

 **STUART OZER** currently leads a team in the [Microsoft Server and Tools](#) business, working with the largest-scale and most demanding data-intensive workloads on [SQL Server](#) and [Windows Azure](#). He is a veteran of Microsoft Research's [eScience](#) team, where he worked on academic collaboration efforts to apply modern database-centered technologies and workflows to diverse problems in biological sequence analysis, protein structure, and environmental sensor networks.

 **PETER MURRAY-RUST** is a researcher at the Unilever Center for Molecular Science Informatics in the Department of Chemistry at the [University of Cambridge](#), England. His research in molecular informatics brings tools from computer science to chemistry, biosciences, and Earth sciences, integrating humans and machines in managing information. He is one of the creators of the Chemical Markup Language (CML), an expanding XML representation of molecular science including molecules, spectra, reactions, computational chemistry, and solid state. He campaigns for open data in science, and is on the advisory board of the Open Knowledge Foundation and a co-author of the Panton Principles for Open Data.

 **MARC SMITH** is a sociologist specializing in the social organization of online communities and computer-mediated interaction. He founded and managed the [Community Technologies Group](#) at Microsoft Research in Redmond, Washington, and led the development of social media reporting and analysis tools for Telligent Systems. He is the chief social scientist of [Connected Action consulting group](#) and lives and works in Silicon Valley, California. He is a co-founder of the Social Media Research Foundation, which is dedicated to open tools, open data, and open scholarship related to social media.

 **BETH PLALE** is director of the [Data to Insight Center](#), which is associated with the [Indiana University Pervasive Technology Institute](#). She is an experimental computer scientist whose research interests include data management, data-driven computing, and the preservation of scientific and scholarly data sets. She is a professor in the [School of Informatics and Computing](#) at Indiana University.

Afterword

AFTERWORD FROM THE EDITORS

This book contains a collection of vignettes on a variety of scientific research projects that either use state-of-the-art Microsoft technologies and research, or have led to pioneering new Microsoft technology concepts. As described, the use of advanced computer science technologies—as well as the use of robust product software—has led to a wide range of new scientific insights. Along the way, the application of Microsoft technologies to emerging scientific problems has significantly enhanced research, development, and products.

In the mid-1990s, Jim Gray of Microsoft Research recognized that the next “big data” challenges for database technology would come from science and not from commerce. He also identified the technical challenges that such data-intensive phenomena would pose for scientists, as well as the key role that IT and computer science could play to enable future scientific discoveries. Gray called this new mode of data-intensive scientific discovery the “fourth paradigm,” and contrasted it with traditional methodologies of experiment, theory, and computer simulation. In designating this new mode, he promoted the need for computer scientists to work collaboratively as equals with scientists in other fields to develop the necessary tools and technologies for this new type of scientific exploration—but not to replace the first three paradigms (experimental, theoretical, and computational). The book, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, contains a collection of essays that expand on Gray’s vision for this new, fourth paradigm. Written by scientists from various fields, the essays explore the future impact of the explosion of scientific data to research areas such as earth,

environment, health, and wellbeing, as well as the scientific infrastructure that is required to support data-intensive science. Many of the projects in this book demonstrate the fourth paradigm in practice.

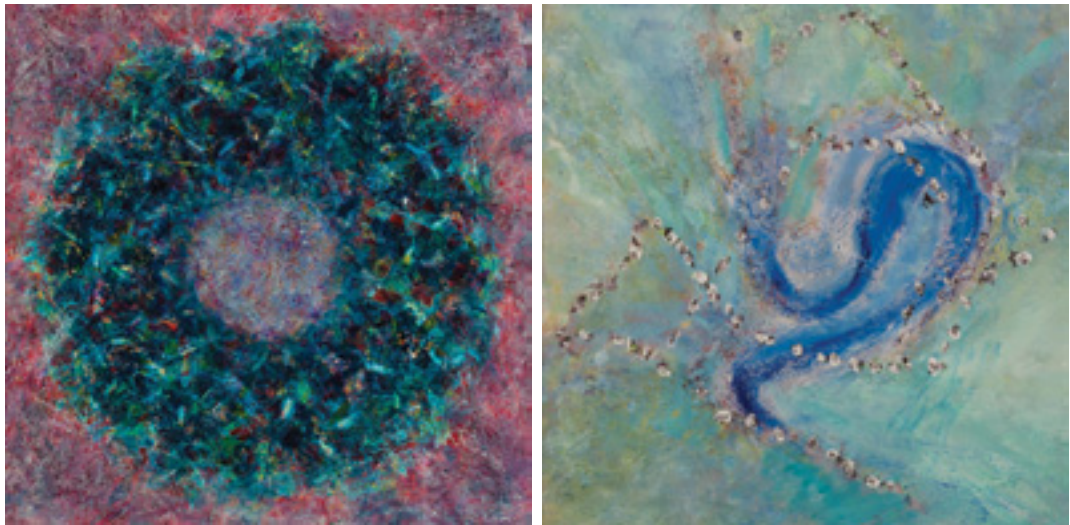
In 2005, Stephen Emmott of Microsoft Research assembled an international group of computer scientists and scientific researchers at a meeting in Venice to discuss the evolution and potential impact of computer science and computing on scientific research in the next fifteen years. The outcome was a manifesto: *Towards 2020 Science*, which described both the challenges and the opportunities arising from the increasing symbiosis of computing, computer science, and the sciences. In the short time since that document was published, many new research activities have emerged from a broad range of science disciplines spanning genomics and proteomics, Earth sciences and climatology, nano-materials, chemistry, and physics. The “2020 science” vision inspired some of the projects described in this book.

Also in 2005, Microsoft Chief Research and Strategy Officer Craig Mundie started the Technical Computing Initiative at Microsoft, a program that enabled computer scientists at Microsoft Research to work with scientists at universities. As described in this book, these collaborative science projects put the visions of “fourth paradigm” and “2020 science” into practice. The success of these projects constitutes a set of examples of how computer scientists can collaborate with scientists in other fields to advance computing technologies and sciences to solve major scientific challenges together. As Mundie says in his introduction, this is genuinely a case of “doing well by doing good!”

Senior editors: Stephen Emmott, David Heckerman, Tony Hey
Editors: Kenji Takeda, Yan Xu

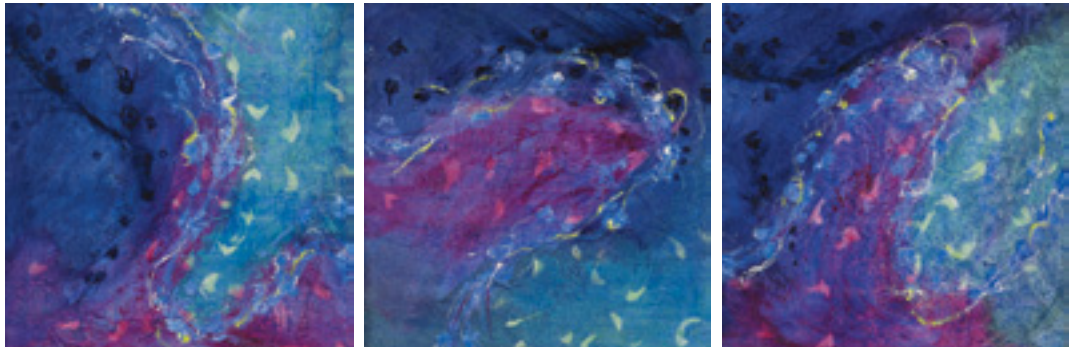
Artist Interpretations

Molecular science inspired these abstract works by Tessa Coe



THE SOUND OF SILENCE

SIMPLY IRRESISTIBLE



SWINGS AND ROUNDABOUTS

THE SOUND OF SILENCE

An amazing molecule called hsp60 assists wrongly folded proteins to achieve their correct shape. This large molecule is classed as one of the so-called molecular chaperones. Within its barrel-shaped center, hsp60 can isolate a protein from counterproductive chance interactions and even guide a misshapen protein to fold up into its correct final form. Unless proteins are exactly the “right” shape, the process within the body for which they were intended will not happen properly.

SIMPLY IRRESISTIBLE

I chose a small molecule called insulin to look at when I first became interested in proteins and the processes by which they fold up into their correct shape. Insulin comprises just two short chains of small molecules. One chain has a couple of short spirals in it; the other just loops, and that’s all. Within this small structure, as in the other globular proteins, powerful hydrophobic interactions help determine the molecule’s final folded shape.

SWINGS AND ROUNDABOUTS

Protein chains usually fold up spontaneously into complex structures. One of the main driving forces in the folding process is the need for all the hydrophobic molecules within the protein—those molecules that are not “happy” being surrounded by water—to be hidden away at the center of the protein. In contrast, charged and ionic hydrophilic regions of the molecule are edged to the surface of the protein, where they can be surrounded by water. These opposing tendencies begin the protein’s drive towards a final self-formed shape.

ABOUT THE ARTIST: Tessa Coe

Originally trained as a physicist at Imperial College in London, Tessa Coe worked for some years as a research scientist before an interlude in industry. Her research interests were biological, however, centered on trying to understand how and why molecules cross the outer surfaces of cells. Biology has flourished since those days, but she has maintained a keen interest in the amazing discoveries related to cells, genes, and proteins. It is these astonishing stories that she draws upon in her most recent work.

Coe has always loved to draw, but her serious interest in art began when she was a student. The physics department at Imperial College is across the street from the Royal College of Art and quite close to the Victoria and Albert Museum. Because Imperial was an all science and engineering college, Coe considers herself lucky to have had such ready access to these great art institutions. During her undergraduate years, she enjoyed wonderful lunchtime art history lectures, after which she could explore the galleries and museums of London and consolidate what she had learned.

At various stages in her life, Coe has attended courses in drawing, painting, and art history, most notably, at Winchester School of Art, the Tower Arts Centre in in Winchester, and at the Juno Studio in Braishfield, in Hampshire. She has been a full-time painter and exhibitor for the last 15 years.

Acknowledgements

The editors would like to thank all of the university researchers who kindly agreed to be interviewed for this collection. They were extremely patient with our requests and without their input, this book would not have been possible. We are grateful to the Microsoft researchers who participated in these projects and to Andrew Herbert, former director of the Microsoft Research Cambridge laboratory, for his invaluable support of computational science and the UK-based projects. We also would like to thank Dan Fay, Simon Mercer, and Jim Pinkelman of Microsoft Research Connections for their constructive input throughout the course of putting this book together. We would like to express our gratitude to Alyssa Felda of Microsoft Research Connections and her team of proofreaders and art designers; and to Kimberley Lane of Microsoft Research Connections for her support of putting together a Creative Commons license for this book. Finally, we thank Rick Rashid, chief research officer at Microsoft Research, for his support and guidance, and Craig Mundie, Microsoft chief research and strategy officer, for the inspiration behind Microsoft’s Technical Computing Initiative.

EDITORIAL BY:

David Heckerman, Kenji Takeda, Stephen Emmott, Tony Hey, Yan Xu

BOOK PHOTOGRAPHY

Chris Towey, Director of Photography for: Medical Sensing via a Contact Lens; Understanding the Immune Response to HIV; Clinical Studies and Data Collection and Reuse

Stock Photos: Dreamstime, IStockphoto, Getty Images

© 2012 Microsoft Corporation. All rights reserved.

Microsoft provides this material solely for informational and marketing purposes.
MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT.



www.microsoft.com/scienceatmicrosoft



Microsoft®