



## 2. HEALTH AND WELLBEING

---





## *Introduction*

**SIMON MERCER** | Microsoft Research

**P**ART 2 OF THIS BOOK EXPLORES the remarkable progress and challenges we are seeing in the most intimate and personal of our sciences, the one with the most immediate impact on all of us across the planet: the science of health and medicine.

The first article sets the scene. Gillam et al. describe the progress of medical science over human history and make a strong case for a convergence of technologies that will change the face of healthcare within our lifetime. The remaining articles shed light on the convergent strands that make up this larger picture, by focusing on particular medical science challenges and the technologies being developed to overcome them.

Any assertion that the coming healthcare revolution will be universal is credible only if we can demonstrate how it can cross the economic and social divides of the modern world. Robertson et al. show that a combination of globally pervasive cell phone technology and the computational technique of Bayesian networks can enable collection of computerized healthcare records in regions where medical care is sparse and can also provide automated, accurate diagnoses.

An understanding of the human brain is one of the grand challenges of medicine, and Lichtman et al. describe their approach to the generation of the vast datasets needed to understand this most

complex of structures. Even imaging the human brain at the subcellular level, with its estimated 160 trillion synaptic connections, is a challenge that will test the bounds of data storage, and that is merely the first step in deducing function from form.

An approach to the next stage of understanding how we think is presented by Horvitz and Kristan, who describe techniques for recording sequences of neuronal activity and correlating them with behavior in the simplest of organisms. This work will lead to a new generation of software tools, bringing techniques of machine learning/artificial intelligence to generate new insights into medical data.

While the sets of data that make up a personal medical record are orders of magnitude smaller than those describing the architecture of the brain, current trends toward universal electronic healthcare records mean that a large proportion of the global population will soon have records of their health available in a digital form. This will constitute in aggregate a dataset of a size and complexity rivaling those of neuroscience. Here we find parallel challenges and opportunities. Buchan, Winn, and Bishop apply novel machine learning techniques to this vast body of healthcare data to automate the selection of therapies that have the most desirable outcome. Technologies such as these will be needed if we are to realize the world of the “Healthcare Singularity,” in which the collective experience of human healthcare is used to inform clinical best practice at the speed of computation.

While the coming era of computerized health records promises more accessible and more detailed medical data, the usability of this information will require the adoption of standard forms of encoding so that inferences can be made across datasets. Cardelli and Priami look toward a future in which medical data can be overlaid onto executable models that encode the underlying logic of biological systems—to not only depict the behavior of an organism but also predict its future condition or reaction to a stimulus. In the case of neuroscience, such models may help us understand how we think; in the case of medical records, they may help us understand the mechanisms of disease and treatment. Although the computational modeling of biological phenomena is in its infancy, it provides perhaps the most intriguing insights into the emerging complementary and synergistic relationship between computational and living systems.



# *The Healthcare Singularity and the Age of Semantic Medicine*

**I**N 1499, WHEN PORTUGUESE EXPLORER VASCO DA GAMA returned home after completing the first-ever sea voyage from Europe to India, he had less than half of his original crew with him—scurvy had claimed the lives of 100 of the 160 men. Throughout the Age of Discovery,<sup>1</sup> scurvy was the leading cause of death among sailors. Ship captains typically planned for the death of as many as half of their crew during long voyages. A dietary cause for scurvy was suspected, but no one had proved it. More than a century later, on a voyage from England to India in 1601, Captain James Lancaster placed the crew of one of his four ships on a regimen of three teaspoons of lemon juice a day. By the halfway point of the trip, almost 40% of the men (110 of 278) on three of the ships had died, while on the lemon-supplied ship, every man survived [1]. The British navy responded to this discovery by repeating the experiment—*146 years later*.

In 1747, a British navy physician named James Lind treated sailors suffering from scurvy using six randomized approaches and demonstrated that citrus reversed the symptoms. The British navy responded, 48 years later, by enacting new dietary guidelines requiring citrus, which virtually eradicated scurvy from the British fleet overnight. The British Board of Trade adopted similar dietary

<sup>1</sup> 15th to 17th centuries.

**MICHAEL GILLAM**  
**CRAIG FEIED**  
**JONATHAN HANDLER**  
**ELIZA MOODY**  
Microsoft

**BEN SHNEIDERMAN**  
**CATHERINE PLAISANT**  
University of Maryland

**MARK SMITH**  
MedStar Health Institutes  
for Innovation

**JOHN DICKASON**  
Private practice

practices for the merchant fleet in 1865, *an additional 70 years later*. The total time from Lancaster's definitive demonstration of how to prevent scurvy to adoption across the British Empire was 264 years [2].

The translation of medical discovery to practice has thankfully improved substantially. But a 2003 report from the Institute of Medicine found that the lag between significant discovery and adoption into routine patient care still averages 17 years [3, 4]. This delayed translation of knowledge to clinical care has negative effects on both the cost and the quality of patient care. A nationwide review of 439 quality indicators found that only half of adults receive the care recommended by U.S. national standards [5].

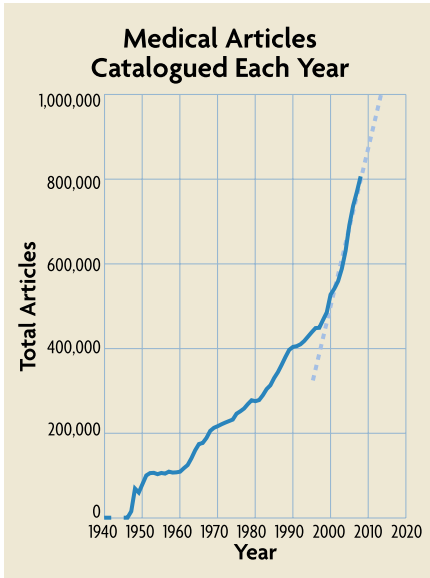
#### THE IMPACT OF THE INFORMATION EXPLOSION IN MEDICINE

Despite the adoption rate of medical knowledge significantly improving, we face a new challenge due to the exponential increase in the rate of medical knowledge discovery. More than 18 million articles are currently catalogued in the biomedical literature, including over 800,000 added in 2008. The accession rate has doubled every 20 years, and the number of articles per year is expected to surpass 1 million in 2012, as shown in Figure 1.

Translating all of this emerging medical knowledge into practice is a staggering challenge. Five hundred years ago, Leonardo da Vinci could be a painter, engineer, musician, and scientist. One hundred years ago, it is said that a physician might have reasonably expected to know everything in the field of medicine.<sup>2</sup> Today, a typical primary care doctor must stay abreast of approximately 10,000 diseases and syndromes, 3,000 medications, and 1,100 laboratory tests [6]. Research librarians estimate that a physician in just one specialty, epidemiology, needs 21 hours of study per day just to stay current [7]. Faced with this flood of medical information, clinicians routinely fall behind, despite specialization and sub-specialization [8].

The sense of information overload in medicine has been present for surprisingly many years. An 1865 speech by Dr. Henry Noyes to the American Ophthalmologic Society is revealing. He said that “medical men strive manfully to keep up their knowledge of how the world of medicine moves on; but too often they are the first to accuse themselves of being unable to meet the duties of their daily calling...” He went on to say, “The preparatory work in the study of medicine is so great, if adequately done, that but few can spare time for its thorough performance...” [9]

<sup>2</sup> [www.medinfo.cam.ac.uk/miu/papers/Hanka/THIM/default.htm](http://www.medinfo.cam.ac.uk/miu/papers/Hanka/THIM/default.htm)



**FIGURE 1.** *The number of biomedical articles catalogued each year is increasing precipitously and is expected to surpass 1 million in 2012.*

**COULD KNOWLEDGE ADOPTION IN HEALTH-CARE BECOME NEARLY INSTANTANEOUS?**

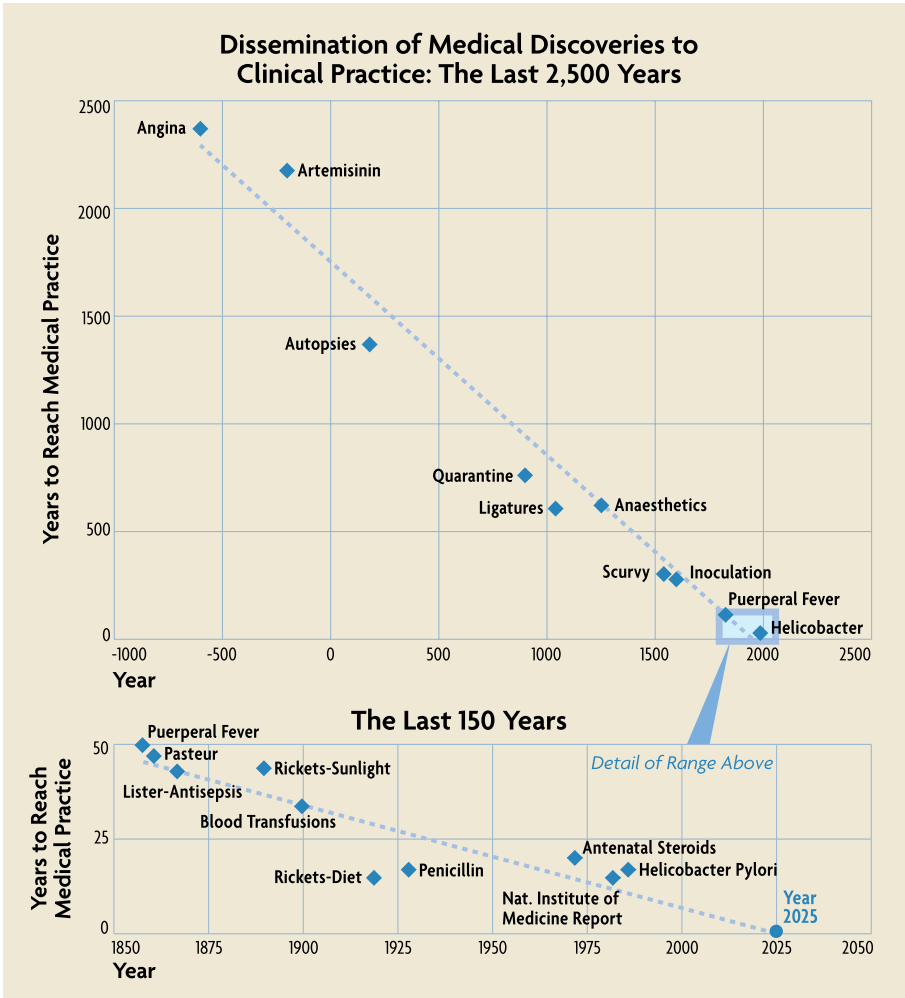
The speed at which definitive medical discoveries have broadly reached medical practice over the last two millennia has progressively increased, as shown in Figure 2 on the next page.

Focusing on the last 150 years, in which the effects of industrialization and the information explosion have been most acute, the trajectory flattens slightly but remains largely linear, as the figure shows. (An asymptotic fit yields an  $r^2$  of 0.73, whereas the linear fit is 0.83.)

Given that even the speed of light is finite, this trend will inevitably be asymptotic to the horizontal axis. Yet, if the linearity can be sufficiently maintained for a while, the next 20 years could emerge as a special time

for healthcare as the translation from medical knowledge discovery to widespread medical practice becomes nearly instantaneous.

The proximity of this trajectory to the axis occurs around the year 2025. In response to the dramatic computational progress observed with Moore’s Law and the growth in parallel and distributed computing architectures, Ray Kurzweil, in *The Singularity Is Near*, predicts that 2045 will be the year of the Singularity, when computers meet or exceed human computational ability and when their ability to recursively improve themselves can lead to an “intelligence explosion” that ultimately affects all aspects of human culture and technology [10]. Mathematics defines a “singularity” as a point at which an object changes its nature so as to attain properties that are no longer the expected norms for that class of object. Today, the dissemination path for medical information is complex and multi-faceted, involving commercials, lectures, brochures, colleagues, and journals. In a world with nearly instantaneous knowledge translation, dissemination paths would become almost entirely digital and direct.



**FIGURE 2.** While it took 2,300 years after the first report of angina for the condition to be commonly taught in medical curricula, modern discoveries are being disseminated at an increasingly rapid pace. Focusing on the last 150 years, the trend still appears to be linear, approaching the axis around 2025.

While the ideas around a technological singularity remain controversial,<sup>3</sup> the authors refer to this threshold moment, when medical knowledge becomes “liquid” and its flow from research to practice (“bench to bedside”) becomes frictionless and immediate, as the “Healthcare Singularity.”

#### THE PROMISES OF A POST-HEALTHCARE SINGULARITY WORLD

Rofecoxib (Vioxx) was approved as safe and effective by the U.S. Food and Drug Administration (FDA) on May 20, 1999. On September 30, 2004, Merck withdrew it from the market because of concerns about the drug’s potential cardiovascular side effects. The FDA estimates that in the 5 years that the drug was on the market, rofecoxib contributed to more than 27,000 heart attacks or sudden cardiac deaths and as many as 140,000 cases of heart disease [11]. Rofecoxib was one of the most widely used medications ever withdrawn; over 80 million people had taken the drug, which was generating US\$2.5 billion a year in sales.<sup>4</sup>

Today, it is reasonable to expect that after an FDA announcement of a drug’s withdrawal from the market, patients will be informed and clinicians will immediately prescribe alternatives. But current channels of dissemination delay that response. In a post-Healthcare Singularity world, that expectation will be met. To enable instantaneous translation, journal articles will consist of not only words, but also bits. Text will commingle with code, and articles will be considered complete only if they include algorithms.

With this knowledge automation, every new medication will flow through a cascade of post-market studies that are independently created and studied by leading academics across the oceans (effectively “crowdsourcing” quality assurance). Suspicious observations will be flagged in real time, and when certainty is reached, unsafe medications will disappear from clinical prescription systems in a rippling wave across enterprises and clinics. The biomedical information explosion will at last be contained and harnessed.

Other scenarios of knowledge dissemination will be frictionless as well: medical residents can abandon the handbooks they have traditionally carried that list drugs of choice for diseases, opting instead for clinical systems that personalize health-care and geographically regionalize treatments based on drug sensitivities that are drawn in real time from the local hospital microbiology lab and correlated with the patient’s genomic profile.

<sup>3</sup> [http://en.wikipedia.org/wiki/Technological\\_singularity](http://en.wikipedia.org/wiki/Technological_singularity)

<sup>4</sup> <http://en.wikipedia.org/wiki/Rofecoxib>

Knowledge discovery will also be enhanced. Practitioners will have access to high-performance, highly accurate databases of patient records to promote preventive medical care, discover successful treatment patterns [12, 13], and reduce medical errors. Clinicians will be able to generate cause-effect hypotheses, run virtual clinical trials to deliver personalized treatment plans, and simulate interventions that can prevent pandemics.

Looking farther ahead, the instantaneous flow of knowledge from research centers to the front lines of clinical care will speed the treatment and prevention of newly emerging diseases. The moment that research labs have identified the epitopes to target for a new disease outbreak, protein/DNA/RNA/lipid synthesizers placed in every big hospital around the world will receive instructions, remotely transmitted from a central authority, directing the on-site synthesis of vaccines or even directed antibody therapies for rapid administration to patients.

#### **PROGRESS TOWARD THE HEALTHCARE SINGULARITY**

Companies such as Microsoft and Google are building new technologies to enable data and knowledge liquidity. Microsoft HealthVault and Google Health are Internet based, secure, and private “consumer data clouds” into which clinical patient data can be pushed from devices and other information systems. Importantly, once the data are in these “patient clouds,” they are owned by the patient. Patients themselves determine what data can be redistributed and to whom the data may be released.

A February 2009 study by KLAS reviewed a new class of emerging data aggregation solutions for healthcare. These enterprise data aggregation solutions (“enterprise data clouds”) unify data from hundreds or thousands of disparate systems (such as MEDSEEK, Carefx, dbMotion, Medicity, and Microsoft Amalga).<sup>5</sup> These platforms are beginning to serve as conduits for data to fill patient data clouds. A recent example is a link between New York-Presbyterian’s hospital-based Amalga aggregation system and its patients’ HealthVault service.<sup>6</sup> Through these links, data can flow almost instantaneously from hospitals to patients.

The emergence of consumer data clouds creates new paths by which new medical knowledge can reach patients directly. On April 21, 2009, Mayo Clinic announced the launch of the Mayo Clinic Health Advisory, a privacy- and security-enhanced

<sup>5</sup> [www.klasresearch.com/Klas/Site/News/PressReleases/2009/Aggregation.aspx](http://www.klasresearch.com/Klas/Site/News/PressReleases/2009/Aggregation.aspx)

<sup>6</sup> <http://chilmarkresearch.com/2009/04/06/healthvault-ny-presbyterian-closing-the-loop-on-care>

online application that offers individualized health guidance and recommendations built with the clinical expertise of Mayo Clinic and using secure and private patient health data from Microsoft HealthVault.<sup>7</sup> Importantly, new medical knowledge and recommendations can be computationally instantiated into the advisory and applied virtually instantaneously to patients worldwide.

New technology is bridging research labs and clinical practice. On April 28, 2009, Microsoft announced the release of Amalga Life Sciences, an extension to the data-aggregation class of products for use by scientists and researchers. Through this release, Microsoft is offering scalable “data aggregation and liquidity” solutions that link three audiences: patients, providers, and researchers. Companies such as Microsoft are building the “pipeline” to allow data and knowledge to flow through a *semantically interoperable* network of patients, providers, and researchers. These types of connectivity efforts hold the promise of effectively instantaneous dissemination of medical knowledge throughout the healthcare system. The Healthcare Singularity could be the gateway event to a new Age of Semantic Medicine.

Instantaneous knowledge translation in medicine is not only immensely important, highly desirable, valuable, and achievable in our lifetimes, but perhaps even inevitable.

#### REFERENCES

- [1] F. Mosteller, “Innovation and evaluation,” *Science*, vol. 211, pp. 881–886, 1981, doi: 10.1126/science.6781066.
- [2] J. Lind, *A Treatise of the Scurvy* (1753). Edinburgh: University Press, reprinted 1953.
- [3] E. A. Balas, “Information Systems Can Prevent Errors and Improve Quality,” *J. Am. Med. Inform. Assoc.*, vol. 8, no. 4, pp. 398–399, 2001, PMID: 11418547.
- [4] A. C. Greiner and Elisa Knebel, Eds., *Health Professions Education: A Bridge to Quality*. Washington, D.C.: National Academies Press, 2003.
- [5] E. A. McGlynn, S. M. Asch, J. Adams, J. Keeseey, J. Hicks, A. DeCristofaro, et al., “The quality of health care delivered to adults in the United States,” *N. Engl. J. Med.*, vol. 348, pp. 2635–2645, 2003, PMID: 12826639.
- [6] T. H. Davenport and J. Glaser, “Just-in-time delivery comes to knowledge management,” *Harv. Bus. Rev.*, vol. 80, no. 7, pp. 107–111, 126, July 2002, doi: 10.1225/R0207H.
- [7] B. S. Alper, J. A. Hand, S. G. Elliott, S. Kinkade, M. J. Hauan, D. K. Onion, and B. M. Sklar, “How much effort is needed to keep up with the literature relevant for primary care?” *J. Med. Libr. Assoc.*, vol. 92, no. 4, pp. 429–437, Oct. 2004.
- [8] C. Lenfant, “Clinical Research to Clinical Practice — Lost in Translation?” *N. Engl. J. Med.*, vol. 349, pp. 868–874, 2003, PMID: 12944573.
- [9] H. D. Noyes, *Specialties in Medicine*, June 1865.

<sup>7</sup> [www.microsoft.com/presspass/press/2009/apr09/04-21MSMayoConsumerSolutionPR.msp](http://www.microsoft.com/presspass/press/2009/apr09/04-21MSMayoConsumerSolutionPR.msp)

- 
- [10] R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Group, 2005, p. 136.
- [11] D. J. Graham, D. Campen, R. Hui, M. Spence, C. Cheetham, G. Levy, S. Shoor, and W. A. Ray, "Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclooxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study," *Lancet*, vol. 365, no. 9458, pp. 475–481, Feb. 5–11, 2005.
- [12] C. Plaisant, S. Lam, B. Shneiderman, M. S. Smith, D. Roseman, G. Marchand, M. Gillam, C. Feied, J. Handler, and H. Rappaport, "Searching Electronic Health Records for temporal patterns in patient histories: A case study with Microsoft Amalga," *Proc. Am. Med. Inform. Assoc.*, Washington, D.C., Nov. 2008.
- [13] T. Wang, C. Plaisant, A. Quinn, R. Stanchak, B. Shneiderman, and S. Murphy, "Aligning temporal data by sentinel events: Discovering patterns in electronic health records," *Proc. ACM CHI2008 Human Factors in Computing Systems Conference*, ACM, New York, Apr. 2008, pp. 457–466, doi: 10.1145/1357054.1357129.



# Healthcare Delivery in Developing Countries: Challenges and Potential Solutions

JOEL ROBERTSON  
DEL DEHART  
Robertson Research  
Institute

KRISTIN TOLLE  
DAVID HECKERMAN  
Microsoft Research

**B**RINGING INTELLIGENT HEALTHCARE INFORMATICS to bear on the dual problems of reducing healthcare costs and improving quality and outcomes is a challenge even in countries with a reasonably developed technology infrastructure. Much of medical knowledge and information remains in paper form, and even where it is digitized, it often resides in disparate datasets and repositories and in diverse formats. Data sharing is uncommon and frequently hampered by the lack of foolproof de-identification for patient privacy. All of these issues impede opportunities for data mining and analysis that would enable better predictive and preventive medicine.

Developing countries face these same issues, along with the compounding effects of economic and geopolitical constraints, transportation and geographic barriers, a much more limited clinical workforce, and infrastructural challenges to delivery. Simple, high-impact deliverable interventions such as universal childhood immunization and maternal childcare are hampered by poor monitoring and reporting systems. A recent *Lancet* article by Christopher Murray's group concluded that "immunization coverage has improved more gradually and not to the level suggested by countries' official reports of WHO and UNICEF estimates. There is an urgent need for independent and contestable monitoring of health indicators in an era of global initiatives that are target-



*The NxOpinion health platform being used by Indian health extension workers.*

oriented and disburse funds based on performance.” [1]

Additionally, the most recent report on the United Nations Millennium Development Goals notes that “pneumonia kills more children than any other disease, yet in developing countries, the proportion of children under five with suspected pneumonia who are taken to appropriate health-care providers remains low.” [2] Providing reliable data gathering and diagnostic decision support at the point of need by the best-trained individual available for care is the goal of public health efforts, but tools

to accomplish this have been expensive, unsupportable, and inaccessible.

Below, we elaborate on the challenges facing healthcare delivery in developing countries and describe computer- and cell phone–based technology we have created to help address these challenges. At the core of this technology is the NxOpinion Knowledge Manager<sup>1</sup> (NxKM), which has been under development at the Robertson Research Institute since 2002. This health platform includes a medical knowledge base assembled from the expertise of a large team of experts in the U.S. and developing countries, a diagnostic engine based on Bayesian networks, and cell phones for end-user interaction.

#### **SCALE UP, SCALE OUT, AND SCALE IN**

One of the biggest barriers to deployment of a decision support or electronic health record system is the ability to scale. The term “scale up” refers to a system’s ability to support a large user base—typically hundreds of thousands or millions. Most systems are evaluated within a narrower scope of users. “Scale out” refers to a system’s ability to work in multiple countries and regions as well as the ability to work across disease types. Many systems work only for one particular disease and are not easily regionalized—for example, for local languages, regulations, and processes. “Scale in” refers to the ability of a system to capture and benchmark against a single

<sup>1</sup> [www.nxopinion.com/product/knowledgemng](http://www.nxopinion.com/product/knowledgemng)

individual. Most systems assume a generic patient and fail to capture unique characteristics that can be effective in individualized treatment.

With respect to scaling up, NxKM has been tested in India, Congo, Dominican Republic, Ghana, and Iraq. It has also been tested in an under-served inner-city community in the United States. In consultation with experts in database scaling, the architecture has been designed to combine multiple individual databases with a central de-identified database, thus allowing, in principle, unlimited scaling options.

As for scaling out to work across many disease types and scaling in to provide accurate individual diagnoses, the amount of knowledge required is huge. For example, INTERNIST-1, an expert system for diagnosis in internal medicine, contains approximately 250,000 relationships among roughly 600 diseases and 4,000 findings [3]. Building on the earlier work of one of us (Heckerman), who developed efficient methods for assessing and representing expert medical knowledge via a Bayesian network [4], we have brought together medical literature, textbook information, and expert panel recommendations to construct a growing knowledge base for NxKM, currently including over 1,000 diseases and over 6,000 discrete findings. The system also scales in by allowing very fine-grained data capture. Each finding within an individual health record or diagnostic case can be tracked and monitored. This level of granularity allows for tremendous flexibility in determining factors relating to outcome and diagnostic accuracy.

With regard to scaling out across a region, a challenge common to developing countries is the exceptionally diverse and region-specific nature of medical conditions. For example, a disease that is common in one country or region might be rare in another. Whereas rule-based expert systems must be completely reengineered in each region, the modular nature of the NxKM knowledge base, which is based on probabilistic similarity networks [4], allows for rapid customization to each region. The current incarnation of NxKM uses region-specific prevalence from expert estimates. It can also update prevalence in each region as it is used in the field. NxKM also incorporates a modular system that facilitates customization to terms, treatments, and language specific to each region. When region-specific information is unknown or unavailable, a default module is used until such data can be collected or identified.

#### **DIAGNOSTIC ACCURACY AND EFFICIENCY**

Studies indicate that even highly trained physicians overestimate their diagnostic accuracy. The Institute of Medicine recently estimated that 44,000 to 98,000

preventable deaths occur each year due to medical error, many due to misdiagnosis [5]. In developing countries, the combined challenges of misdiagnoses and missing data not only reduce the quality of medical care for individuals but lead to missed outbreak recognition and flawed population health assessment and planning.

Again, building on the diagnostic methodology from probabilistic similarity networks [4], NxKM employs a Bayesian reasoning engine that yields accurate diagnoses. An important component of the system that leads to improved accuracy is the ability to ask the user additional questions that are likely to narrow the range of possible diagnoses. NxKM has the ability to ask the user for additional findings based on value-of-information computations (such as a cost function) [4]. Also important for clinical use is the ability to identify the confidence in the diagnosis (i.e., the probability of the most likely diagnosis). This determination is especially useful for less-expert users of the system, which is important for improving and supervising the care delivered by health extension workers (HEWs) in developing regions where deep medical knowledge is rare.

#### **GETTING HEALTHCARE TO WHERE IT IS NEEDED: THE LAST MILE**

Another key challenge is getting diagnostics to where they are most needed. Because of their prevalence in developing countries, cell phones are a natural choice for a delivery vehicle. Indeed, it is believed that, in many such areas, access to cell phones is better than access to clean water. For example, according to the market database Wireless Intelligence,<sup>2</sup> 80 percent of the world's population was within range of a cellular network in 2008. And figures from the International Telecommunication Union<sup>3</sup> show that by the end of 2006, 68 percent of the world's mobile subscriptions were in developing countries. More recent data from the International Telecommunications Union shows that between 2002 and 2007, cellular subscription was the most rapid growth area for telecommunication in the world, and that the per capita increase was greatest in the developing world.<sup>4</sup>

Consequently, we have developed a system wherein cell phones are used to access a centrally placed NxKM knowledge base and diagnostic engine implemented on a PC. We are now testing the use of this system with HEWs in rural India. In addition to providing recommendations for medical care to the HEWs, the phone/

<sup>2</sup> [www.wirelessintelligence.com](http://www.wirelessintelligence.com)

<sup>3</sup> [www.itu.int](http://www.itu.int)

<sup>4</sup> [www.itu.int/ITU-D/ict/papers/2009/7.1%20teltscher\\_IDI%20India%202009.pdf](http://www.itu.int/ITU-D/ict/papers/2009/7.1%20teltscher_IDI%20India%202009.pdf)

central-PC solution can be used to create portable personal health records. One of our partner organizations, School Health Annual Report Programme (SHARP), will use it to screen more than 10 million Indian schoolchildren in 2009, creating a unique virtual personal health record for each child.

Another advantage of this approach is that the data collected by this system can be used to improve the NxKM knowledge base. For example, as mentioned above, information about region-specific disease prevalence is important for accurate medical diagnosis. Especially important is time-critical information about the outbreak of a disease in a particular location. As the clinical application is used, validated disease cases, including those corresponding to a new outbreak, are immediately available to NxKM. In addition, individual diagnoses can be monitored centrally. If the uploaded findings of an individual patient are found to yield a low-confidence diagnosis, the patient can be identified for follow-up.

#### **THE USER INTERFACE**

A challenge with cellular technology is the highly constrained user interface and the difficulty of entering data using a relatively small screen and keypad. Our system simplifies the process in a number of ways. First, findings that are common for a single location (e.g., facts about a given village) are prepopulated into the system. Also, as mentioned above, the system is capable of generating questions—specifically, simple multiple-choice questions—after only basic information such as the chief complaint has been entered. In addition, questions can be tailored to the organization, location, or skill level of the HEW user.

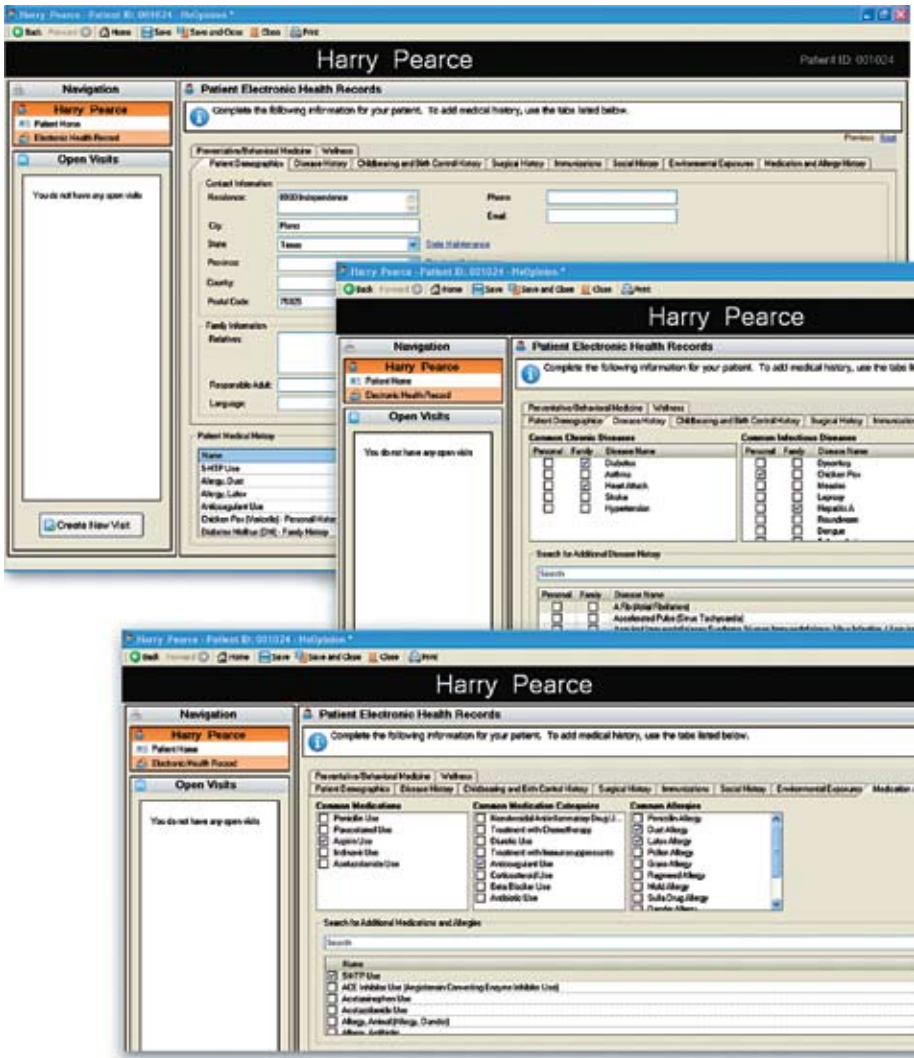
It is also important that the user interface be independent of the specific device hardware because users often switch between phones of different designs. Our interface application sits on top of a middle-layer platform that we have implemented for multiple devices.

In addition to simple input, the interface allows easy access to important bits of information. For example, it provides a daily summary of patients needing care, including their diagnosis, village location, and previous caregivers.

#### **DATA-SHARING SOLUTIONS**

Even beyond traditional legacy data silos (such as EPIC and CERNER) [5], barriers to sharing critical public health data still exist—including concerns about privacy and sovereignty. Data availability can also be limited regionally (e.g., in India and South Africa), by organizations (e.g., the World Health Organization,

308



NxOpinion's innovative approach, which shows data when you want it, how you want it, and where you want it, using artificial intelligence.

World Vision, or pharmaceutical companies), or by providers (e.g., insurance companies and medical provider groups). Significant public health value resides in each of these datasets, and efforts should be made to overcome the barriers to gathering data into shared, de-identified global databases. Such public datasets, while useful on their own, also add significant value to proprietary datasets, providing valuable generic context to proprietary information.

NxKM imports, manages, and exports data via *publish sets*. These processes allow various interest groups (governments, public health organizations, primary care providers, small hospitals, laboratory and specialty services, and insurance providers) to share the same interactive de-identified (privacy-preserving) global database while maintaining control of proprietary and protected data.

#### LOOKING FORWARD

Several challenges remain. While better educated HEWs are able to use these data collection and diagnostic decision support tools readily, other HEWs, such as Accredited Social Health Activists (ASHAs) and other front-line village workers, are often illiterate or speak only a local dialect. We are exploring two potential solutions—one that uses voice recognition technology and another that allows a user to answer multiple-choice questions via the cell phone's numeric keypad. Voice recognition technology provides added flexibility in input, but—at least so far—it requires the voice recognizer to be trained by each user.

Another challenge is unique and reproducible patient identification—verification that the subject receiving treatment is actually the correct patient—when there is no standard identification system for most under-served populations. Voice recognition combined with face recognition and newer methods of biometrics, along with a corroborating GPS location, can help ensure that the patient who needs the care is the one actually receiving treatment.

Another barrier is data integrity. For example, most rural individuals will report diagnoses that have not been substantiated by qualified medical personnel and could be erroneous. We have attempted to mitigate this issue by using an inference engine that allows for down-weighting of unsubstantiated evidence.

Deploying systems that work anywhere in the world can lead to the creation of a massive amount of patient information. Storing, reconciling, and then accessing that information in the field, all while maintaining appropriate privacy and security, are exceptionally challenging when patient numbers are in the millions (instead of tens of thousands, as with most current electronic health record

systems). Further, feeding verified data on this scale back into the system to improve its predictive capability while maintaining the ability to analyze and retrieve specific segments (data mine) remains difficult.

A final, and perhaps the greatest, obstacle is that of cooperation. If organizations, governments, and companies are willing to share a de-identified global database while protecting and owning their own database, medical science and healthcare can benefit tremendously. A unified database that allows integration across many monitoring and evaluation systems and databases should help in quickly and efficiently identifying drug resistance or outbreaks of disease and in monitoring the effectiveness of treatments and healthcare interventions. The global database should support data queries that guard against the identification of individuals and yet provide sufficient information for statistical analyses and validation. Such technology is beginning to emerge (e.g., [6]), but the daunting challenge of finding a system of rewards that encourages such cooperation remains.

#### **SUMMARY**

We have developed and are beginning to deploy a system for the acquisition, analysis, and transmission of medical knowledge and data in developing countries. The system includes a centralized component based on PC technology that houses medical knowledge and data and has real-time diagnostic capabilities, complemented by a cell phone-based interface for medical workers in the field. We believe that such a system will lead to improved medical care in developing countries through improved diagnoses, the collection of more accurate and timely data across more individuals, and the improved dissemination of accurate and timely medical knowledge and information.

When we stop and think about how a world of connected personal health records can be used to improve medicine, we can see that the potential impact is staggering. By knowing virtually every individual who exists, the diseases affecting that person, and where he or she is located; by improving data integrity; and by collecting the data in a central location, we can revolutionize medicine and perhaps even eradicate more diseases. This global system can monitor the effects of various humanitarian efforts and thereby justify and tailor efforts, medications, and resources to specific areas. It is our hope that a system that can offer high-quality diagnoses as well as collect and rapidly disseminate valid data will save millions of lives. Alerts and responses can become virtually instantaneous and can thus lead to the identification of drug resistance, outbreaks, and effective treatments in a fraction of the

time it takes now. The potential for empowering caregivers in developing countries though a global diagnostic and database system is enormous.

#### REFERENCES

- [1] S. S. Lim, D. B. Stein, A. Charrow, and C. J. L. Murray, "Tracking progress towards universal childhood immunisation and the impact of global initiatives: a systematic analysis of three-dose diphtheria, tetanus, and pertussis immunisation coverage," *Lancet*, vol. 372, pp. 2031–2046, 2008, doi: 10.1016/S0140-6736(08)61869-3.
- [2] *The Millennium Development Goals Report*. United Nations, 2008.
- [3] R. A. Miller, M. A. McNeil, S. M. Challinor, F. E. Masarie, Jr., and J. D. Myers, "The Internist-1/Quick Medical Reference Project—Status Report," *West. J. Med.* vol. 145, pp. 816–822, 1986.
- [4] D. Heckerman. *Probabilistic Similarity Networks*. Cambridge, MA: MIT Press, 1991.
- [5] L. Kohn, J. Corrigan, and M. Donaldson, Eds. *To Err Is Human: Building a Safer Health System*. Washington, D.C.: National Academies Press, 2000.
- [6] C. Dwork and K. Nissim, "Privacy-Preserving Datamining on Vertically Partitioned Databases," *Proc. CRYPTO*, 2004, doi: 10.1.1.86.8559.





## *Discovering the Wiring Diagram of the Brain*

JEFF W. LICHTMAN  
R. CLAY REID  
HANSPETER PFISTER  
Harvard University

MICHAEL F. COHEN  
Microsoft Research

**T**HE BRAIN, THE SEAT OF OUR COGNITIVE ABILITIES, is perhaps the most complex puzzle in all of biology. Every second in the human brain, billions of cortical nerve cells transmit billions of messages and perform extraordinarily complex computations. How the brain works—how its function follows from its structure—remains a mystery.

The brain's vast numbers of nerve cells are interconnected at synapses in circuits of unimaginable complexity. It is largely assumed that the specificity of these interconnections underlies our ability to perceive and classify objects, our behaviors both learned (such as playing the piano) and intrinsic (such as walking), and our memories—not to mention controlling lower-level functions such as maintaining posture and even breathing. At the highest level, our emotions, our sense of self, our very consciousness are entirely the result of activities in the nervous system.

At a macro level, human brains have been mapped into regions that can be roughly associated with specific types of activities. However, even this building-block approach is fraught with complexity because often many parts of the brain participate in completing a task. This complexity arises especially because most behaviors begin with sensory input and are followed by analysis, decision making, and finally a motor output or action.

At the microscopic level, the brain comprises billions of neu-

rons, each connected to other neurons by up to several thousand synaptic connections. Although the existence of these synaptic circuits has been appreciated for over a century, we have no detailed circuit diagrams of the brains of humans or any other mammals. Indeed, neural circuit mapping has been attempted only once, and that was two decades ago on a small worm with only 300 nerve cells. The central stumbling block is the enormous technical difficulty associated with such mapping. Recent technological breakthroughs in imaging, computer science, and molecular biology, however, allow a reconsideration of this problem. But even if we had a wiring diagram, we would need to know what messages the neurons in the circuit are passing—not unlike listening to the signals on a computer chip. This represents the second impediment to understanding: traditional physiological methods let us listen to only a tiny fraction of the nerves in the circuit.

To get a sense of the scale of the problem, consider the cerebral cortex of the human brain, which contains more than 160 trillion synaptic connections. These connections originate from billions of neurons. Each neuron receives synaptic connections from hundreds or even thousands of different neurons, and each sends information via synapses to a similar number of target neurons. This enormous fan-in and fan-out can occur because each neuron is geometrically complicated, possessing many receptive processes (dendrites) and one highly branched outflow process (an axon) that can extend over relatively long distances.

One might hope to be able to reverse engineer the circuits in the brain. In other words, if we could only tease apart the individual neurons and see which one is connected to which and with what strength, we might at least begin to have the tools to decode the functioning of a particular circuit. The staggering numbers and complex cellular shapes are not the only daunting aspects of the problem. The circuits that connect nerve cells are nanoscopic in scale. The density of synapses in the cerebral cortex is approximately 300 million per cubic millimeter.

Functional magnetic resonance imaging (fMRI) has provided glimpses into the macroscopic 3-D workings of the brain. However, the finest resolution of fMRI is approximately 1 cubic millimeter per voxel—the same cubic millimeter that can contain 300 million synapses. Thus there is a huge amount of circuitry in even the most finely resolved functional images of the human brain. Moreover, the size of these synapses falls below the diffraction-limited resolution of traditional optical imaging technologies.

Circuit mapping could potentially be amenable to analysis based on color coding of neuronal processes [1] and/or the use of techniques that break through the

diffraction limit [2]. Presently, the gold standard for analyzing synaptic connections is to use electron microscopy (EM), whose nanometer (nm) resolution is more than sufficient to ascertain the finest details of neural connections. But to map circuits, one must overcome a technical hurdle: EM typically images very thin sections (tens of nanometers in thickness), so reconstructing a volume requires a “serial reconstruction” whereby the image information from contiguous slices of the same volume is recomposed into a volumetric dataset. There are several ways to generate such volumetric data (see, for example, [3-5]), but all of these have the potential to generate astonishingly large digital image data libraries, as described next.

#### **SOME NUMBERS**

If one were to reconstruct by EM all the synaptic circuitry in 1 cubic mm of brain (roughly what might fit on the head of a pin), one would need a set of serial images spanning a millimeter in depth. Unambiguously resolving all the axonal and dendritic branches would require sectioning at probably no more than 30 nm. Thus the 1 mm depth would require 33,000 images. Each image should have at least 10 nm lateral resolution to discern all the vesicles (the source of the neurotransmitters) and synapse types. A square-millimeter image at 5 nm resolution is an image that has  $\sim 4 \times 10^{10}$  pixels, or 10 to 20 gigapixels. So the image data in 1 cubic mm will be in the range of 1 petabyte ( $2^{50} \sim 1,000,000,000,000,000$  bytes). The human brain contains nearly 1 million cubic mm of neural tissue.

#### **SOME SUCCESSES TO DATE**

Given this daunting task, one is tempted to give up and find a simpler problem. However, new technologies and techniques provide glimmers of hope. We are pursuing these with the ultimate goal of creating a “connectome”—a complete circuit diagram of the brain. This goal will require intensive and large-scale collaborations among biologists, engineers, and computer scientists.

Three years ago, the Reid and Lichtman labs began working on ways to automate and accelerate large-scale serial-section EM. Focusing specifically on large cortical volumes at high resolution, the Reid group has concentrated on very high throughput as well as highly automated processes. So far, their work has been published only in abstract form [3], but they are confident about soon having the first 10 terabytes of volumetric data on fine-scale brain anatomy. Physiological experiments can now show the function of virtually every neuron in a 300  $\mu\text{m}$  cube. The new EM data has the resolution to show virtually every axon, dendrite, and

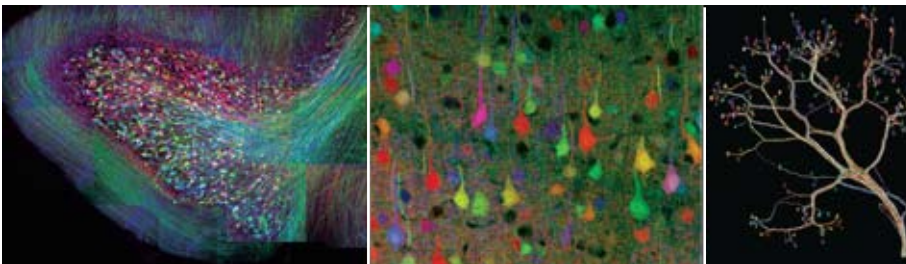
synapse—the physical connections that underlie neuronal function.

The problem of separating and tracking the individual neurons within the volume remains. However, some successes have already been achieved using exotic means. Lichtman's lab found a way to express various combinations of red, green, and blue fluorescent proteins in genetically engineered mice. These random combinations presently provide about 90 colors or combinations of colors [1]. With this approach, it is possible to track individual neurons as they branch to their eventual synaptic connections to other neurons or to the end-organs in muscle. The multi-color labeled nerves (dubbed “rainbow”), shown in Figure 1, are reminiscent of the rainbow cables in computers and serve the same purpose: to disambiguate wires traveling over long distances.

Because these colored labels are present in the living mouse, it is possible to track synaptic wiring changes by observing the same sites multiple times over minutes, days, or even months.

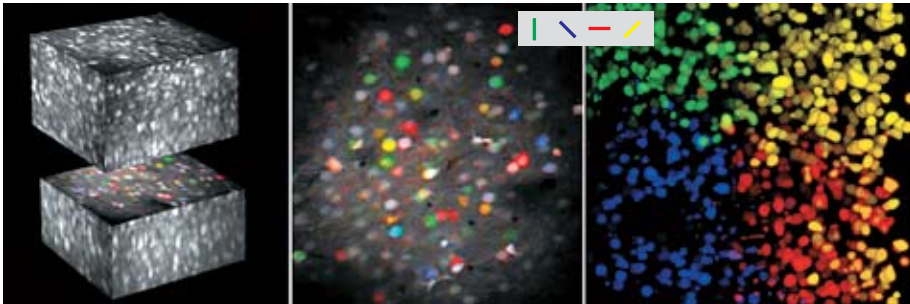
Reid's lab has been able to stain neurons of rat and cat visual cortices such that they “light up” when activated. By stimulating the cat with lines of different orientations, they have literally been able to see which neurons are firing, depending on the specific visual stimulus. By comparing the organization of the rat's visual cortex to that of the cat, they have found that while a rat's neurons appear to be randomly organized based on the orientation of the visual stimulus, a cat's neurons exhibit remarkable structure. (See Figure 2.)

Achieving the finest resolution using EM requires imaging very thin slices of neural tissue. One method begins with a block of tissue; after each imaging pass, a



**FIGURE 1.**

*Brainbow images showing individual neurons fluorescing in different colors. By tracking the neurons through stacks of slices, we can follow each neuron's complex branching structure to create the treelike structures in the image on the right.*



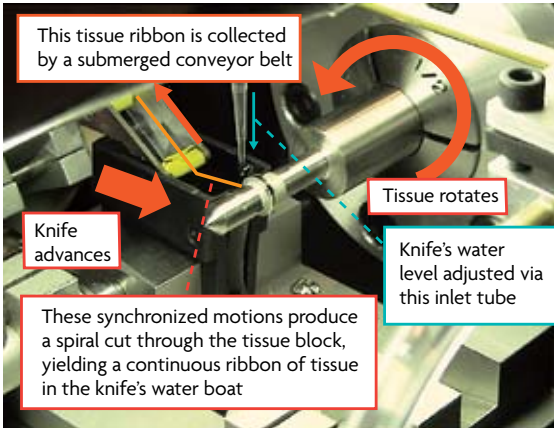
**FIGURE 2.**

*Neurons in a visual cortex stained in vivo with a calcium-sensitive dye. Left: A 3-D reconstruction of thousands of neurons in a rat visual cortex, obtained from a stack of images (300  $\mu\text{m}$  on a side). The neurons are color coded according to the orientation of the visual stimulus that most excited them. Center: A 2-D image of the plane of section from the left panel. Neurons that responded to different stimulus orientations (different colors) are arranged seemingly randomly in the cortex. Inset: Color coding of stimulus orientations. Right: By comparison, the cat visual cortex is extremely ordered. Neurons that responded preferentially to different stimulus orientations are segregated with extraordinary precision. This image represents a complete 3-D functional map of over 1,000 neurons in a 300x300x200  $\mu\text{m}$  volume in the visual cortex [6, 7].*

thin slice is removed (and destroyed) from the block, and then the process is repeated. Researchers in the Lichtman group at Harvard have developed a new device—a sort of high-tech lathe that they are calling an Automatic Tape-Collecting Lathe Ultramicrotome (ATLUM)—that can allow efficient nanoscale imaging over large tissue volumes. (See Figure 3 on the next page.)

The ATLUM [3] automatically sections an embedded block of brain tissue into thousands of ultrathin sections and collects these on a long carbon-coated tape for later staining and imaging in a scanning electron microscope (SEM). Because the process is fully automated, volumes as large as tens of cubic millimeters—large enough to span entire multi-region neuronal circuits—can be quickly and reliably reduced to a tape of ultrathin sections. SEM images of these ATLUM-collected sections can attain lateral resolutions of 5 nm or better—sufficient to image individual synaptic vesicles and to identify and trace all circuit connectivity.

The thin slices are images of one small region at a time. Once a series of individual images is obtained, these images must be stitched together into very large images



**FIGURE 3.**  
*The Automatic Tape-Collecting Lathe Ultramicrotome (ATLUM), which can allow efficient nanoscale imaging over large tissue volumes.*

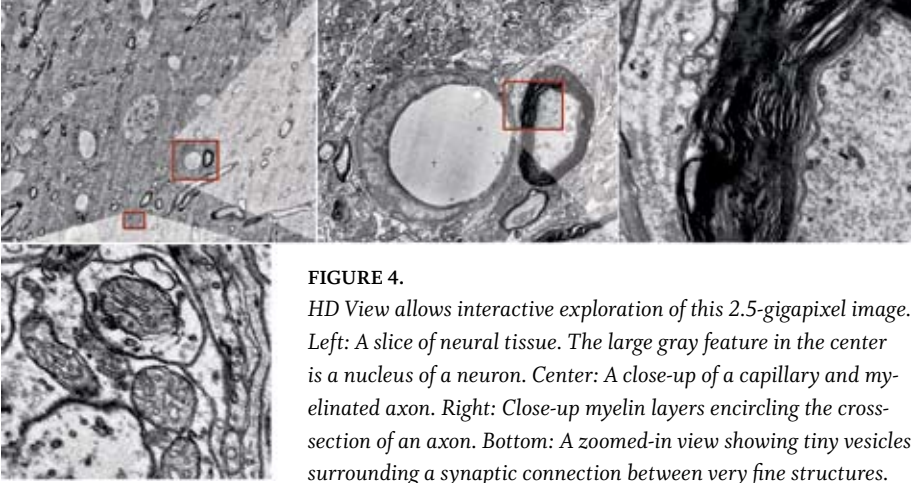
and possibly stacked into volumes. At Microsoft Research, work has proceeded to stitch together and then interactively view images containing billions of pixels.<sup>1</sup> Once these gigapixel-size images are organized into a hierarchical pyramid, the HD View application can stream requested imagery over the Web for viewing.<sup>2</sup> This allows exploration of both large-scale and very fine-scale features. Figure 4 shows a walkthrough of the result.

Once the images are captured and stitched, multiple slices of a sample must be stacked to assemble them into a coherent volume. Perhaps the most difficult task at that point is extracting the individual strands of neurons. Work is under way at Harvard to provide interactive tools to aid in outlining individual “processes” and then tracking them slice to slice to pull out each dendritic and axonal fiber [8, 9]. (See Figure 5.) Synaptic interfaces are perhaps even harder to find automatically; however, advances in both user interfaces and computer vision give hope that the whole process can be made tractable.

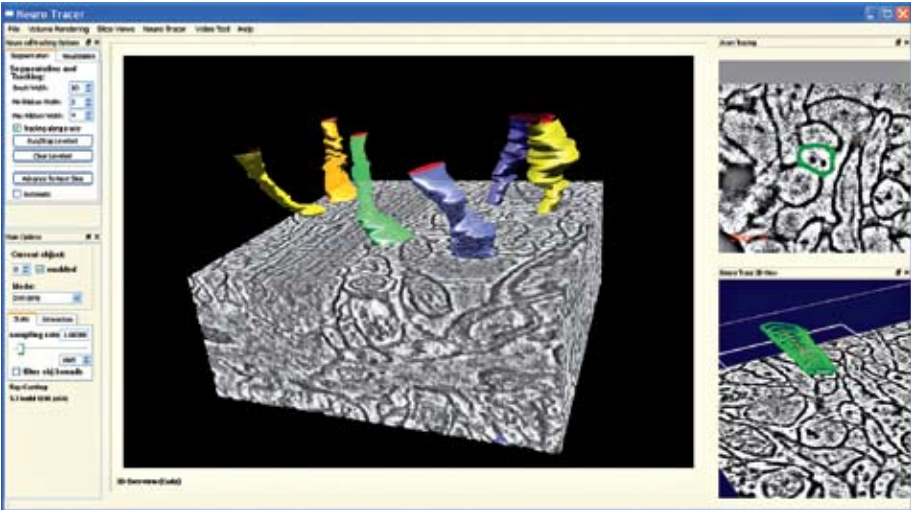
Decoding the complete connectome of the human brain is one of the great challenges of the 21st century. Advances at both the biological level and technical level are certain to lead to new successes and discoveries, and they will hopefully help answer fundamental questions about how our brain performs the miracle of thought.

<sup>1</sup> <http://research.microsoft.com/en-us/um/redmond/groups/ivm/ICE>

<sup>2</sup> <http://research.microsoft.com/en-us/um/redmond/groups/ivm/HDView>



**FIGURE 4.** HD View allows interactive exploration of this 2.5-gigapixel image. Left: A slice of neural tissue. The large gray feature in the center is a nucleus of a neuron. Center: A close-up of a capillary and myelinated axon. Right: Close-up myelin layers encircling the cross-section of an axon. Bottom: A zoomed-in view showing tiny vesicles surrounding a synaptic connection between very fine structures.



**FIGURE 5.** NeuroTrace allows neuroscientists to interactively explore and segment neural processes in high-resolution EM data.

## REFERENCES

- [1] J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, and J. W. Lichtman, "Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system," *Nature*, vol. 450, pp. 56–62, 2007, doi: 10.1038/nature06293.
- [2] S. Hell, "Microscopy and its focal switch," *Nature Methods*, vol. 6, pp. 24–32, 2009, doi: 10.1038/NMeth.1291.
- [3] D. Bock, W. C. Lee, A. Kerlin, M. L. Andermann, E. Soucy, S. Yurgenson, and R. C. Reid, "High-throughput serial section electron microscopy in mouse primary visual cortex following in vivo two-photon calcium imaging," *Soc. Neurosci. Abstr.*, vol. 769, no. 12, 2008.
- [4] W. Denk and H. Horstmann, "Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure," *PLoS Biol.*, vol. 2, p. e329, 2004, doi: 10.1017/S1431927606066268.
- [5] K. J. Hayworth, N. Kasthuri, R. Schalek, and J. W. Lichtman, "Automating the Collection of Ultrathin Serial Sections for Large Volume TEM Reconstructions," *Microsc. Microanal.*, vol. 12, pp. 86–87, 2006.
- [6] K. Ohki, S. Chung, Y. H. Ch'ng, P. Kara, and R. C. Reid, "Functional imaging with cellular resolution reveals precise microarchitecture in visual cortex," *Nature*, vol. 433, pp. 597–603, 2005, doi:10.1038/nature03274.
- [7] K. Ohki, S. Chung, P. Kara, M. Hübener, T. Bonhoeffer, and R. C. Reid, "Highly ordered arrangement of single neurons in orientation pinwheels," *Nature*, vol. 442, pp. 925–928, 2006, doi:10.1038/nature05019.
- [8] W. Jeong, J. Beyer, M. Hadwiger, A. Vazquez, H. Pfister, and R. Whitaker, "Scalable and Interactive Segmentation and Visualization of Neural Processes in EM Datasets," *IEEE Trans. Visual. Comput. Graphics*, Oct. 2009.
- [9] A. Vazquez, E. Miller, and H. Pfister, "Multiphase Geometric Couplings for the Segmentation of Neural Processes," *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, June 2009.



# *Toward a Computational Microscope for Neurobiology*

**ERIC HORVITZ**  
Microsoft Research

**WILLIAM KRISTAN**  
University of California,  
San Diego

**A**LTHOUGH GREAT STRIDES HAVE BEEN MADE in neurobiology, we do not yet understand how the symphony of communication among neurons leads to rich, competent behaviors in animals. How do local interactions among neurons coalesce into the behavioral dynamics of nervous systems, giving animals their impressive abilities to sense, learn, decide, and act in the world? Many details remain cloaked in mystery. We are excited about the promise of gaining new insights by applying computational methods, in particular machine learning and inference procedures, to generate explanatory models from data about the activities of populations of neurons.

## **NEW TOOLS FOR NEUROBIOLOGISTS**

For most of the history of electrophysiology, neurobiologists have monitored the membrane properties of neurons of vertebrates and invertebrates by using glass micropipettes filled with a conducting solution. Mastering techniques that would impress the most expert of watchmakers, neuroscientists have fabricated glass electrodes with tips that are often less than a micron in diameter, and they have employed special machinery to punch the tips into the cell bodies of single neurons—with the hope that the neurons will function as they normally do within larger assemblies. Such an approach has provided data about the membrane voltages and action

potentials of a single cell or just a handful of cells.

However, the relationship between neurobiologists and data about nervous systems is changing. New recording machinery is making data available on the activity of large populations of neurons. Such data makes computational procedures increasingly critical as experimental tools for unlocking new understanding about the connections, architecture, and overall machinery of nervous systems.

New opportunities for experimentation and modeling on a wider scale have become available with the advent of fast optical imaging methods. With this approach, dyes and photomultipliers are used to track calcium levels and membrane potentials of neurons, with high spatial and temporal resolution. These high-fidelity optical recordings allow neurobiologists to examine the simultaneous activity of populations of tens to thousands of neurons. In a relatively short time, data available about the activity of neurons has grown from a trickle of information gleaned via sampling of small numbers of neurons to large-scale observations of neuronal activity.

Spatiotemporal datasets on the behaviors of populations of neurons pose tantalizing inferential challenges and opportunities. The next wave of insights about the neurophysiological basis for cognition will likely come via the application of new kinds of computational lenses that direct an information-theoretic “optics” onto streams of spatiotemporal population data.

We foresee that neurobiologists studying populations of neurons will one day rely on tools that serve as *computational microscopes*—systems that harness machine learning, reasoning, and visualization to help neuroscientists formulate and test hypotheses from data. Inferences derived from the spatiotemporal data streaming from a preparation might even be overlaid on top of traditional optical views during experiments, augmenting those views with annotations that can help with the direction of the investigation.

Intensive computational analyses will serve as the basis for modeling and visualization of the intrinsically high-dimensional population data, where multiple neuronal units interact and contribute to the activity of other neurons and assemblies, and where interactions are potentially context sensitive—circuits and flows might exist dynamically, transiently, and even simultaneously on the same neuronal substrate.

#### COMPUTATION AND COMPLEXITY

We see numerous opportunities ahead for harnessing fast-paced computations to assist neurobiologists with the science of making inferences from neuron popula-

tion data. Statistical analyses have already been harnessed in studies of populations of neurons. For example, statistical methods have been used to identify and characterize neuronal activity as trajectories in large dynamical state spaces [1]. We are excited about employing richer machine learning and reasoning to induce explanatory models from case libraries of neuron population data. Computational procedures for induction can assist scientists with teasing insights from raw data on neuronal activity by searching over large sets of alternatives and weighing the plausibility of different explanatory models. The computational methods can be tasked with working at multiple levels of detail, extending upward from circuit-centric exploration of local connectivity and functionality of neurons to potentially valuable higher-level abstractions of neuronal populations—abstractions that may provide us with simplifying representations of the workings of nervous systems.

Beyond generating explanations from observations, inferential models can be harnessed to compute the *expected value of information*, helping neuroscientists to identify the best next test to perform or information to gather, in light of current goals and uncertainties. Computing the value of information can help to direct interventional studies, such as guidance on stimulating specific units, clamping the voltage of particular cells, or performing selective modification of cellular activity via agonist and antagonist pharmacological agents.

We believe that there is promise in both automated and interactive systems, including systems that are used in real-time settings as bench tools. Computational tools might one day even provide real-time guidance for probes and interventions via visualizations and recommendations that are dynamically generated during imaging studies.

Moving beyond the study of specific animal systems, computational tools for analyzing neuron population data will likely be valuable in studies of the construction of nervous systems during embryogenesis, as well as in comparing nervous systems of different species of animals. Such studies can reveal the changes in circuitry and function during development and via the pressures of evolutionary adaptation.

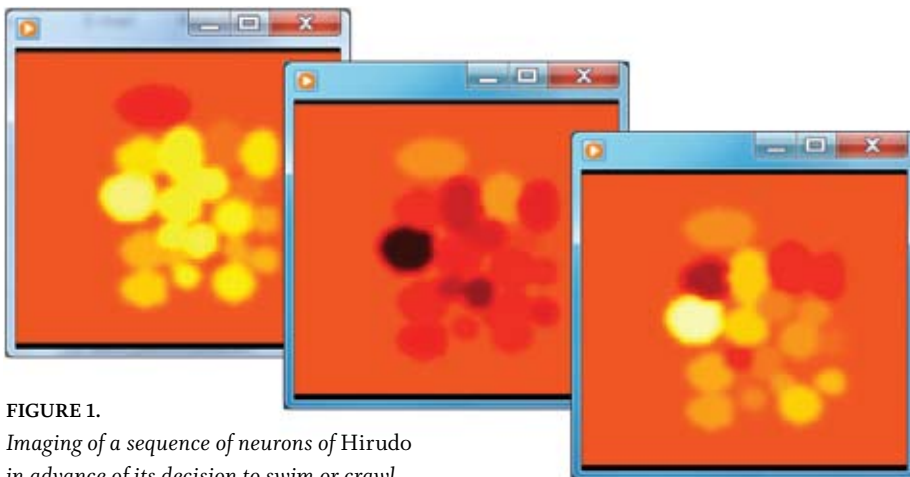
#### **SPECTRUM OF SOPHISTICATION**

Neurobiologists study nervous systems of invertebrates and vertebrates across a spectrum of complexity. Human brains are composed of about 100 billion neurons that interact with one another via an estimated 100 trillion synapses. In contrast, the brain of the nematode, *Caenorhabditis elegans* (*C. elegans*), has just 302 neurons. Such invertebrate nervous systems offer us an opportunity to learn about the prin-

ciples of neuronal systems, which can be generalized to more complex systems, including our own. For example, *C. elegans* has been a model system for research on the structure of neuronal circuits; great progress has been achieved in mapping the precise connections among its neurons.

Many neurobiologists choose to study simpler nervous systems even if they are motivated by questions about the neurobiological nature of human intelligence. Nervous systems are derived from a family tree of refinements and modifications, so it is likely that key aspects of neuronal information processing have been conserved across brains of a range of complexities. While new abstractions, layers, and interactions may have evolved in more complex nervous systems, brains of different complexities likely rely on a similar neuronal fabric—and there is much that we do not know about that fabric.

In work with our colleagues Ashish Kapoor, Erick Chastain, Johnson Apacible, Daniel Wagenaar, and Paxon Frady, we have been pursuing the use of machine learning, reasoning, and visualization to understand the machinery underlying decision making in *Hirudo*, the European medicinal leech. We have been applying computational analyses to make inferences from optical data about the activity of populations of neurons within the segmental ganglia of *Hirudo*. The ganglia are composed of about 400 neurons, and optical imaging reveals the activity of approximately 200 neurons at a time—all the neurons on one side of the ganglion. Several frames of the optical imaging of *Hirudo* are displayed in Figure 1. The brightness



**FIGURE 1.**  
*Imaging of a sequence of neurons of Hirudo in advance of its decision to swim or crawl.*

of each of the imaged neurons represents the level of depolarization of the cells, which underlies the production of action potentials.

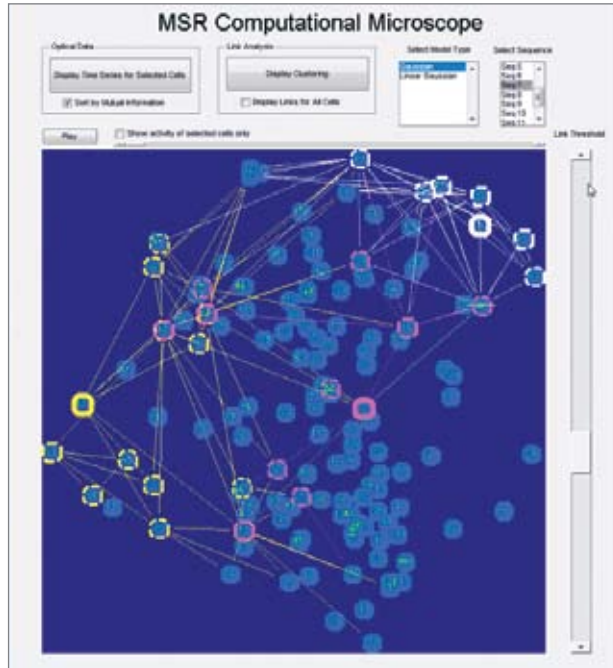
We are developing analyses and assembling tools in pursuit of our vision of developing computational microscopes for understanding the activity of neuronal populations and their relationship to behavior. In one approach, we generate graphical probabilistic temporal models that can predict the forthcoming behavior of *Hirudo* from a short window of analysis of population data. The models are generated by searching over large spaces of feasible models in which neurons, and abstractions of neurons, serve as random variables and in which temporal and atemporal dependencies are inferred among the variables. The methods can reveal modules of neurons that appear to operate together and that can appear dynamically over the course of activity leading up to decisions by the animal. In complementary work, we are considering the role of neuronal states in defining trajectories through state spaces of a dynamical system.

#### EMERGENCE OF A COMPUTATIONAL MICROSCOPE

We have started to build interactive viewers and tools that allow scientists to manipulate inferential assumptions and parameters and to inspect implications visually. For example, sliders allow for smooth changes in thresholds for admitting connections among neurons and for probing strengths of relationships and membership in modules. We would love to see a world in which such tools are shared broadly among neuroscientists and are extended with learning, inference, and visualization components developed by the neuroscience community.

Figure 2 on the next page shows a screenshot of a prototype tool we call the MSR Computational Microscope, which was developed by Ashish Kapoor, Erick Chastain, and Eric Horvitz at Microsoft Research as part of a broader collaboration with William Kristan at the University of California, San Diego, and Daniel Wagenaar at California Institute of Technology. The tool allows users to visualize neuronal activity over a period of time and then explore inferences about relationships among neurons in an interactive manner. Users can select from a variety of inferential methods and specify modeling assumptions. They can also mark particular neurons and neuronal subsets as focal points of analyses. The view in Figure 2 shows an analysis of the activity of neurons in the segmental ganglia of *Hirudo*. Inferred informational relationships among cells are displayed via highlighting of neurons and through the generation of arcs among neurons. Such inferences can help to guide exploration and confirmation of physical connections among neurons.

**FIGURE 2.**  
Possible connections and clusters inferred from population data during imaging of Hirudo.



**FIGURE 3.**  
Inferred informational relationships among neurons in a segmental ganglion of Hirudo. Measures of similarity of the dynamics of neuronal activity over time are displayed via arcs and clusters.

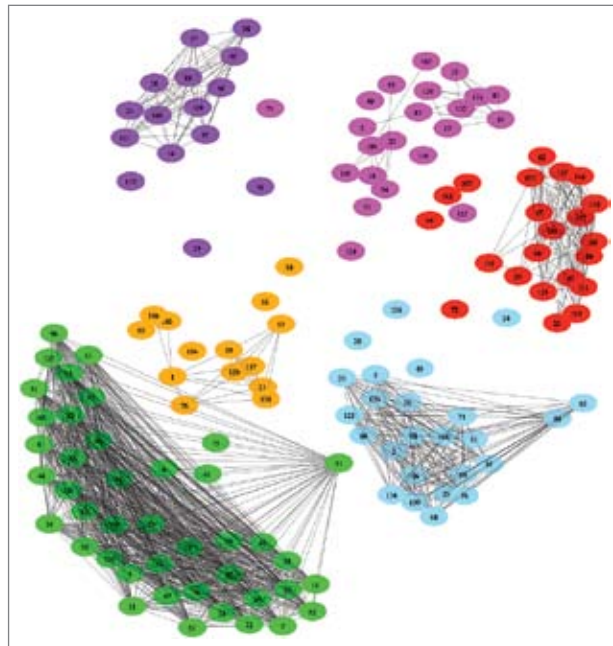


Figure 3 shows another informational analysis that spatially clusters cells that behave in a similar manner in the ganglia of *Hirudo* over a set of trials. The analysis provides an early vision of how information-theoretic analyses might one day help neurobiologists to discover and probe interactions within and between neuronal subsystems.

We are only at the start of this promising research direction, but we expect to see a blossoming of analyses, tools, and a broader sub-discipline that focuses on the neuroinformatics of populations of neurons. We believe that computational methods will lead us to effective representations and languages for understanding neuronal systems and that they will become essential tools for neurobiologists to gain insight into the myriad mysteries of sensing, learning, and decision making by nervous systems.

#### REFERENCES

- [1] K. L. Briggman, H. D. I. Abarbanel, and W. B. Kristan, Jr., “Optical imaging of neuronal populations during decision-making,” *Science*, vol. 307, pp. 896–901, 2005, doi: 10.1126/science.110.





# *A Unified Modeling Approach to Data-Intensive Healthcare*

IAIN BUCHAN  
University of Manchester

JOHN WINN  
CHRIS BISHOP  
Microsoft Research

**T**HE QUANTITY OF AVAILABLE HEALTHCARE DATA is rising rapidly, far exceeding the capacity to deliver personal or public health benefits from analyzing this data [1]. Three key elements of the rise are electronic health records (EHRs), biotechnologies, and scientific outputs. We discuss these in turn below, leading to our proposal for a unified modeling approach that can take full advantage of a data-intensive environment.

## **ELECTRONIC HEALTH RECORDS**

Healthcare organizations around the world, in both low- and high-resource settings, are deploying EHRs. At the community level, EHRs can be used to manage healthcare services, monitor the public's health, and support research. Furthermore, the social benefits of EHRs may be greater from such population-level uses than from individual care uses.

The use of standard terms and ontologies in EHRs is increasing the structure of healthcare data, but clinical coding behavior introduces new potential biases. For example, the introduction of incentives for primary care professionals to tackle particular conditions may lead to fluctuations in the amount of coding of new cases of those conditions [2]. On the other hand, the falling cost of devices for remote monitoring and near-patient testing is leading to more capture of objective measures in EHRs, which can provide

less biased signals but may create the illusion of an increase in disease prevalence simply due to more data becoming available.

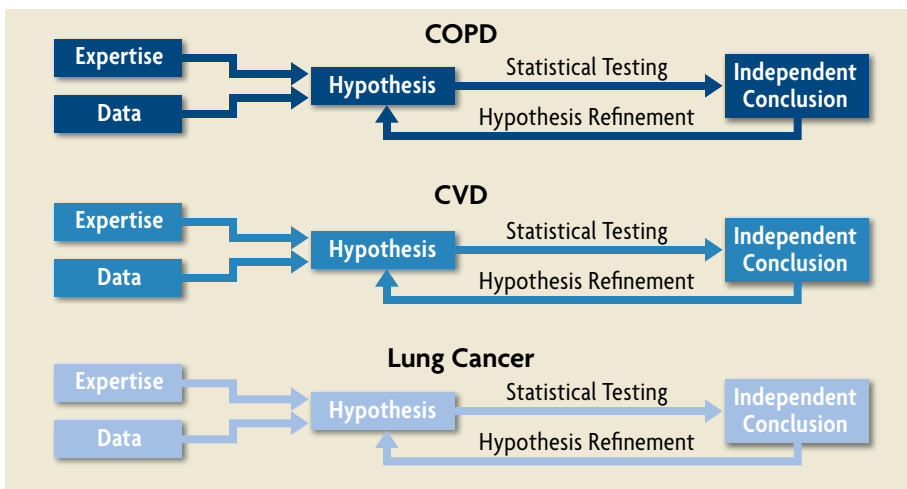
Some patients are beginning to access and supplement their own records or edit a parallel health record online [3]. The stewardship of future health records may indeed be more with individuals (patients/citizens/consumers) and communities (families/local populations etc.) than with healthcare organizations. In summary, the use of EHRs is producing more data-intensive healthcare environments in which substantially more data are captured and transferred digitally. Computational thinking and models of healthcare to apply to this wealth of data, however, have scarcely been developed.

### **BIOTECHNOLOGIES**

Biotechnologies have fueled a boom in molecular medical research. Some techniques, such as genome-wide analysis, produce large volumes of data without the sampling bias that a purposive selection of study factors might produce. Such datasets are thus more wide ranging and unselected than conventional experimental measurements. Important biases can still arise from artifacts in the biotechnical processing of samples and data, but these are likely to decrease as the technologies improve. A greater concern is the systematic error that lies outside the data landscape—for example, in a metabolomic analysis that is confounded by not considering the time of day or the elapsed time from the most recent meal to when the sample was taken. The integration of different scales of data, from molecular-level to population-level variables, and different levels of directness of measurement of factors is a grand challenge for data-intensive health science. When realistically complex multi-scale models are available, the next challenge will be to make them accessible to clinicians and patients, who together can evaluate the competing risks of different options for personalizing treatment.

### **SCIENTIFIC OUTPUTS**

The outputs of health science have been growing exponentially [4]. In 2009, a new paper is indexed in PubMed, the health science bibliographic system, on average every 2 minutes. The literature-review approach to managing health knowledge is therefore potentially overloaded. Furthermore, the translation of new knowledge into practice innovation is slow and inconsistent [5]. This adversely affects not only clinicians and patients who are making care decisions but also researchers who are reasoning about patterns and mechanisms. There is a need to combine the mining



**FIGURE 1.** Conventional approaches based on statistical hypothesis testing artificially decompose the healthcare domain into numerous sub-problems. They thereby miss a significant opportunity for statistical “borrowing of strength.” Chronic obstructive pulmonary disease (COPD), cardiovascular disease (CVD), and lung cancer can be considered together as a “big three” [6].

of evidence bases with computational models for exploring the burgeoning data from healthcare and research.

Hypothesis-driven research and reductionist approaches to causality have served health science well in identifying the major independent determinants of health and the outcomes of individual healthcare interventions. (See Figure 1.) But they do not reflect the complexity of health. For example, clinical trials exclude as many as 80 percent of the situations in which a drug might be prescribed—for example, when a patient has multiple diseases and takes multiple medications [7]. Consider a newly licensed drug released for general prescription. Clinician X might prescribe the drug while clinician Y does not, which could give rise to natural experiments. In a fully developed data-intensive healthcare system in which the data from those experiments are captured in EHRs, clinical researchers could explore the outcomes of patients on the new drug compared with natural controls, and they could potentially adjust for confounding and modifying factors. However, such adjustments might be extremely complex and beyond the capability of conventional models.

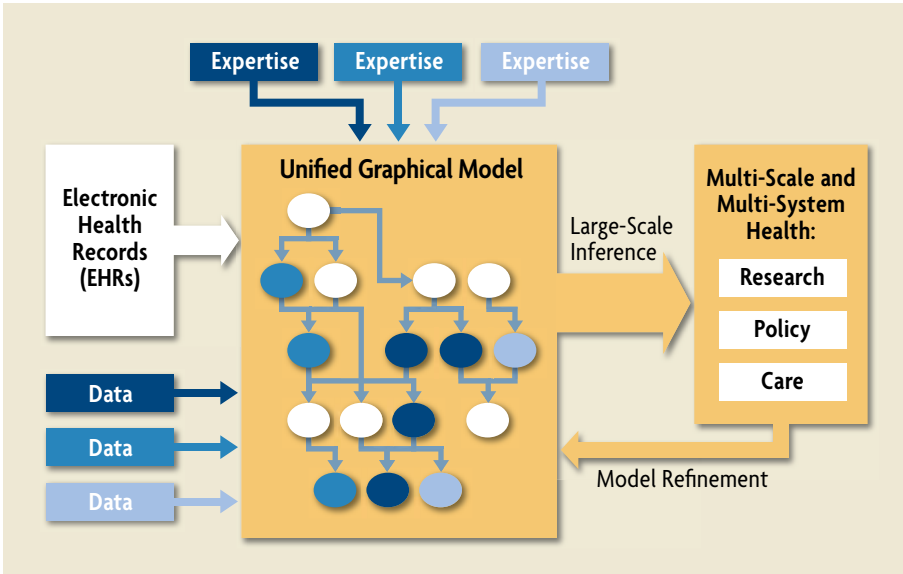


FIGURE 2.

We propose a unified approach to healthcare modeling that exploits the growing statistical resources of electronic health records in addition to the data collected for specific studies.

### A UNIFIED APPROACH

We propose a unified modeling approach that can take full advantage of a data-intensive environment without losing the realistic complexity of health. (See Figure 2.) Our approach relies on developments within the machine learning field over the past 10 years, which provide powerful new tools that are well suited to this challenge. Knowledge of outcomes, interventions, and confounding or modifying factors can all be captured and represented through the framework of probabilistic graphical models in which the relevant variables, including observed data, are expressed as a graph [8]. Inferences on this graph can then be performed automatically using a variety of algorithms based on local message passing, such as [9]. Compared with classical approaches to machine learning, this new framework offers a deeper integration of domain knowledge, taken directly from experts or from the literature, with statistical learning. Furthermore, these automatic inference algorithms can scale to datasets of hundreds of millions of records, and new tools such

as Infer.NET allow rapid development of solutions within this framework [10]. We illustrate the application of this approach with two scenarios.

In scenario 1, an epidemiologist is investigating the genetic and environmental factors that predispose some children to develop asthma. He runs a cohort study of 1,000 children who have been followed for 10 years, with detailed environmental and physiological measures as well as data on over half a million of the 3 million genetic factors that might vary between individuals. The conventional epidemiology approach might test predefined hypotheses using selected groups of genetic and other factors. A genome-wide scanning approach might also be taken to look for associations between individual genetic factors and simple definitions of health status (e.g., current wheeze vs. no current wheeze at age 5 years). Both of these approaches use relatively simple statistical models. An alternative machine learning approach might start with the epidemiologist constructing a graphical model of the problem space, consulting literature and colleagues to build a graph around the organizing principle—say, “peripheral airways obstruction.” This model better reflects the realistic complexity of asthma with a variety of classes of wheeze and other signs and symptoms, and it relates them to known mechanisms. Unsupervised clustering methods are then used to explore how genetic, environmental, and other study factors influence the clustering into different groups of allergic sensitization with respect to skin and blood test results and reports of wheezing. The epidemiologist can relate these patterns to biological pathways, thereby shaping hypotheses to be explored further.

In scenario 2, a clinical team is auditing the care outcomes for patients with chronic angina. Subtly different treatment plans of care are common, such as different levels of investigation and treatment in primary care before referral to specialist care. A typical clinical audit approach might debate the treatment plan, consult literature, examine simple summary statistics, generate some hypotheses, and perhaps test the hypotheses using simple regression models. An alternative machine learning approach might construct a graphical model of the assumed treatment plan, via debate and reference to the literature, and compare this with discovered network topologies in datasets reflecting patient outcomes. Plausible networks might then be used to simulate the potential effects of changes to clinical practice by running scenarios that change edge weights in the underlying graphs. Thus the families of associations in locally relevant data can be combined with evidence from the literature in a scenario-planning activity that involves clinical reasoning and machine learning.

#### THE FOURTH PARADIGM: HEALTH AVATARS

Unified models clearly have the potential to influence personal health choices, clinical practice, and public health. So is this a paradigm for the future?

The first paradigm of healthcare information might be considered to be the case history plus expert physician, formalized by Hippocrates more than 2,000 years ago and still an important part of clinical practice. In the second paradigm, a medical record is shared among a set of complementary clinicians, each focusing their specialized knowledge on the patient's condition in turn. The third paradigm is evidence-based healthcare that links a network of health professionals with knowledge and patient records in a timely manner. This third paradigm is still in the process of being realized, particularly in regard to capturing the complexities of clinical practice in a digital record and making some aspects of healthcare computable.

We anticipate a fourth paradigm of healthcare information, mirroring that of other disciplines, whereby an individual's health data are aggregated from multiple sources and attached to a unified model of that person's health. The sources can range from body area network sensors to clinical expert oversight and interpretation, with the individual playing a much greater part than at present in building and acting on his or her health information. Incorporating all of this data, the unified model will take on the role of a "health avatar"—the electronic representation of an individual's health as directly measured or inferred by statistical models or clinicians. Clinicians interacting with a patient's avatar can achieve a more integrated view of different specialist treatment plans than they do with care records alone.

The avatar is not only a statistical tool to support diagnosis and treatment, but it is also a communication tool that links the patient and the patient's elected network of clinicians and other trusted caregivers—for what-if treatment discussions, for example. While initially acting as a fairly simple multi-system model, the health avatar could grow in depth and complexity to narrow the gap between avatar and reality. Such an avatar would not involve a molecular-level simulation of a human being (which we view as implausible) but would instead involve a unified statistical model that captures current clinical understanding as it applies to an individual patient.

This paradigm can be extended to communities, where multiple individual avatars interact with a community avatar to provide a unified model of the community's health. Such a community avatar could provide relevant and timely information for use in protecting and improving the health of those in the community. Scarce community resources could be matched more accurately to lifetime healthcare needs,

particularly in prevention and early intervention, to reduce the severity and/or duration of illness and to better serve the community as a whole. Clinical, consumer, and public health services could interact more effectively, providing both social benefit and new opportunities for healthcare innovation and enterprise.

## CONCLUSION

Data alone cannot lead to data-intensive healthcare. A substantial overhaul of methodology is required to address the real complexity of health, ultimately leading to dramatically improved global public healthcare standards. We believe that machine learning, coupled with a general increase in computational thinking about health, can be instrumental. There is arguably a societal duty to develop computational frameworks for seeking signals in collections of health data if the potential benefit to humanity greatly outweighs the risk. We believe it does.

## REFERENCES

- [1] J. Powell and I. Buchan, "Electronic health records should support clinical research," *J. Med. Internet Res.*, vol. 7, no. 1, p. e4, Mar. 14, 2005, doi: 10.2196/jmir.7.1.e4.
- [2] S. de Lusignan, N. Hague, J. van Vlymen, and P. Kumarapeli, "Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research," *Prim. Care Inform.*, vol. 14, no. 1, pp. 59–66, 2006.
- [3] L. Bos and B. Blobel, Eds., *Medical and Care Compunetics 4*, vol. 127 in Studies in Health Technology and Informatics series. Amsterdam: IOS Press, pp. 311–315, 2007.
- [4] B. G. Druss and S. C. Marcus, "Growth and decentralization of the medical literature: implications for evidence-based medicine," *J. Med. Libr. Assoc.*, vol. 93, no. 4, pp. 499–501, Oct. 2005, PMID: PMC1250328.
- [5] A. Mina, R. Ramlogan, G. Tampubolon, and J. Metcalfe, "Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge," *Res. Policy*, vol. 36, no. 5, pp. 789–806, 2007, doi: 10.1016/j.respol.2006.12.007.
- [6] M. Gerhardsson de Verdier, "The Big Three Concept - A Way to Tackle the Health Care Crisis?" *Proc. Am. Thorac. Soc.*, vol. 5, pp. 800–805, 2008.
- [7] M. Fortin, J. Dionne, G. Pinho, J. Gignac, J. Almirall, and L. Lapointe, "Randomized controlled trials: do they have external validity for patients with multiple comorbidities?" *Ann. Fam. Med.*, vol. 4, no. 2, pp. 104–108, Mar.–Apr. 2006, doi: 10.1370/afm.516.
- [8] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [9] J. Winn and C. Bishop, "Variational Message Passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, 2005.
- [10] T. Minka, J. Winn, J. Guiver, and A. Kannan, Infer.NET, Microsoft Research Cambridge, <http://research.microsoft.com/infernet>.





# *Visualization in Process Algebra Models of Biological Systems*

**LUCA CARDELLI**  
Microsoft Research

**CORRADO PRIAMI**  
Microsoft Research -  
University of Trento  
Centre for Computational  
and Systems Biology and  
University of Trento

IN A RECENT PAPER, NOBEL LAUREATE PAUL NURSE calls for a better understanding of living organisms through “both the development of the appropriate languages to describe information processing in biological systems and the generation of more effective methods to translate biochemical descriptions into the functioning of the logic circuits that underpin biological phenomena.” [1]

The language that Nurse wishes to see is a formal language that can be automatically translated into machine executable code and that enables simulation and analysis techniques for proving properties of biological systems. Although there are many approaches to the formal modeling of living systems, only a few provide executable descriptions that highlight the mechanistic steps that make a system move from one state to another [2]. Almost all the techniques related to mathematical modeling abstract from these individual steps to produce global behavior, usually averaged over time.

Computer science provides the key elements to describe mechanistic steps: algorithms and programming languages [3]. Following the metaphor of molecules as processes introduced in [4], process calculi have been identified as a promising tool to model biological systems that are inherently complex, concurrent, and driven by the interactions of their subsystems.

Causality is a key difference between language-based modeling approaches and other techniques. In fact, causality in concurrent languages is strictly related to the notion of concurrency or independence of events, which makes causality substantially different from temporal ordering. An activity A causes an activity B if A is a necessary condition for B to happen and A influences the activity of B—i.e., there is a flow of information from A to B. The second part of the condition defining causality makes clear the distinction between precedence (related only to temporal ordering) and causality (a subset of the temporal ordering in which the flow of information is also considered) [5]. As a consequence, the list of the reactions performed by a system does not provide causal information but only temporal information. It is therefore mandatory to devise new modeling and analysis tools to address causality.

Causality is a key issue in the analysis of complex interacting systems because it helps in dissecting independent components and simplifying models while also allowing us to clearly identify cross-talks between different signaling cascades. Once the experimentalist observes an interesting event in a simulation, it is possible to compact the previous history of the system, exposing only the preceding events that caused the interesting one. This can give precise hints about the causes of a disease, the interaction of a drug with a living system (identifying its efficacy and its side effects), and the regulatory mechanisms of oscillating behaviors.

Causality is a relationship between events, and as such it is most naturally studied within discrete models, which are in turn described via algorithmic modeling languages. Although many modeling languages have been defined in computer science to model concurrent systems, many challenges remain to building algorithmic models for the system-level understanding of biological processes. These challenges include the relationship between low-level local interactions and emergent high-level global behavior; the incomplete knowledge of the systems under investigation; the multi-level and multi-scale representations in time, space, and size; and the causal relations between interactions and the context awareness of the inner components. Therefore, the modeling formalisms that are candidates to propel algorithmic systems biology should be complementary to and interoperable with mathematical modeling. They should address parallelism and complexity, be algorithmic and quantitative, express causality, and be interaction driven, composable, scalable, and modular.

#### LANGUAGE VISUALIZATION

A fundamental issue in the adoption of formal languages in biology is their

usability. A modeling language must be understandable by biologists so they can relate it to their own informal models and to experiments.

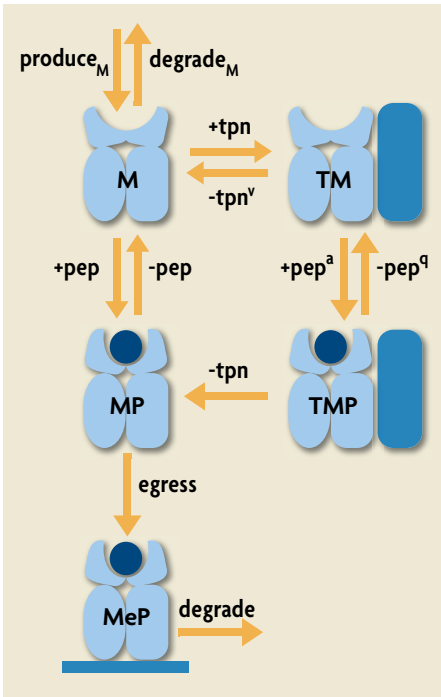
One attempt by biologists to connect formal languages and informal descriptions of systems involved the use of a constrained natural language organized in the form of tables that collect all the information related to the structure and dynamic of a system. This narrative representation is informative and structured enough to be compiled into formal description that is amenable to simulation and analysis [6, 7]. Although the narrative modeling style is not yet visual, it is certainly more readable and corresponds better to the intuition of biologists than a formal (programming) language.

The best way to make a language understandable to scientists while also helping to manage complexity is to visualize the language. This is harder than visualizing data or visualizing the results of simulations because a language implicitly describes the full kinetics of a system, including the dynamic relationships between events. Therefore, language visualization must be dynamic, and possibly reactive [8], which means that a scientist should be able to detect and insert events in a running simulation by direct intervention. This requires a one-to-one correspondence between the internal execution of a formal language and its visualization so that the kinetics of the language can be fully reflected in the kinetics of the visualization and vice versa.

This ability to fully match the kinetics of a general (Turing-complete) modeling language to visual representations has been demonstrated, for example, for pi-calculus [9], but many practical challenges remain to adapting such general methods to specific visualization requirements. (See Figure 1 on the next page.) One such requirement, for example, is the visualization and tracking of molecular complexes; to this end, the BlenX language [10] and its support tools permit explicit representation of complexes of biological elements and examination of their evolution in time [11]. (See Figure 2 on page 103.) The graphical representation of complexes is also useful in studying morphogenesis processes to unravel the mechanistic steps of pattern formation. (See Figure 3 on page 104.)

## ANALYSIS

Model construction is one step in the scientific cycle, and appropriate modeling languages (along with their execution and visualization capabilities) are important, particularly for modeling complex systems. Ultimately, however, one will want to analyze the model using a large number of techniques. Some of these techniques may be centered on the underlying mathematical framework, such as the analysis of



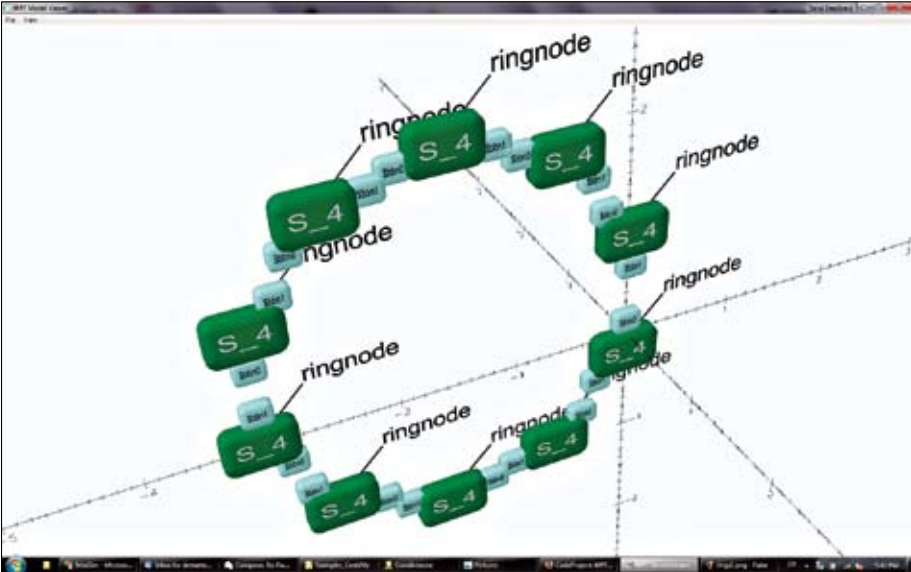
**FIGURE 1.** This diagram can be placed in 1:1 correspondence with formal stochastic pi-calculus models [9, 12, 13] so that one can edit either the diagrams or the models. The nodes represent molecular states (the node icons are just for illustration), and the labeled arcs represent interactions with other molecules in the environment. The models use a biochemical variant of pi-calculus with rate weight as superscripts and with +/- for binding and unbinding.

tactic relationships between genes, genomes, and proteins. An entirely new avenue of research is the investigation of the semantic equivalences of biological entities populating complex networks of interactions. This approach could lead to new visions of systems and reinforce the need for computer science to enhance systems biology.

Biology is a data-intensive science. Biological systems are huge collections of in-

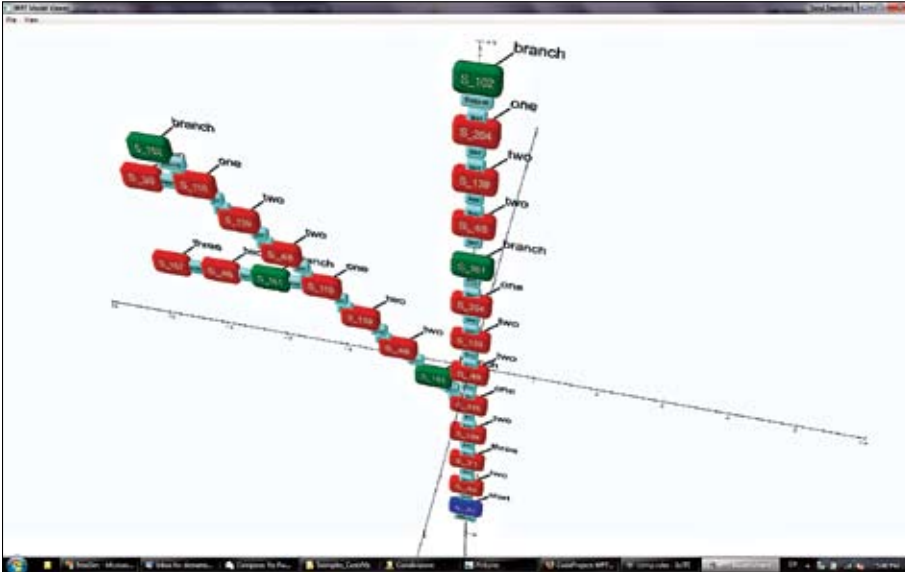
differential equations, Markov chains, or Petri nets generated from the model. Other techniques may be centered on the model description (the language in which the model is written). For example, we may want to know whether two different model descriptions actually represent the same behavior, by some measure of behavior equivalence. This kind of model correspondence can arise, for example, from apparently different biological systems that work by the same fundamental principles. A similar question is whether we can simplify (abstract) a model description and still preserve its behavior, again by some measure of behavior equivalence that may mask some unimportant detail.

Behavioral equivalences are in fact a primary tool in computer science for verifying computing systems. For instance, we can use equivalences to ensure that an implementation is in agreement with a specification, abstracting as much as possible from syntactic descriptions and instead focusing on the semantics (dynamic) of specifications and implementations. So far, biology has focused on syntactic



**FIGURE 2.**  
*The green S boxes in the diagram represent entities populating the biological system under consideration. The light blue rectangles attached to the green boxes represent the active interfaces/ domains available for complexation and decomplexation. The diagram shows how the simulation of the BlenX specification formed a ring complex and provides the position and the connections between boxes for inspection.*

teracting components. The last decade of research has contributed to identifying and classifying those components, especially at the molecular level (gene, metabolites, proteins). To make sense of the large amount of data available, we need to implicitly represent them in compact and executable models so that executions can recover the available data as needed. This approach would merge syntax and semantics in unifying representations and would create the need for different ways of storing, retrieving, and comparing data. A model repository that represents the dynamics of biological processes in a compact and mechanistic manner would therefore be extremely valuable and could heighten the understanding of biological data and the basic biological principles governing life. This would facilitate predictions and the optimal design of further experiments to move from data collection to knowledge production.



**FIGURE 3.**

The green, red, and blue *S* boxes in the diagram represent different species populating the biological system under consideration. The light blue rectangles attached to the boxes represent the active interfaces/domains available for complexation and decomplexation. The diagram elucidates how patterns are formed in morphogenesis processes simulated by *BlenX* specifications.

### ANALYSIS VISUALIZATION

Executable models need visualization to make their execution interactive (to dynamically focus on specific features) and reactive (to influence their execution on the fly). Execution is one form of analysis; other analysis methods will need visualization as well. For complex systems, the normal method of “batch” analysis, consisting of running a complex analysis on the model and then mining the output for clues, needs to be replaced with a more interactive, explorative approach.

Model abstraction is an important tool for managing complexity, and we can envision performing this activity interactively—for example, by lumping components together or by hiding components. The notion of lumping will then need an appropriate visualization and an appropriate way of relating the behavior of the original components to the behavior of the lumped components. This doesn’t mean visualizing the modeling language, but rather visualizing an abstraction function between

models. We therefore suggest visualizing the execution of programs/models in such a way that the output is linked to the source code/model specification and the graphical abstraction performed by the end user is transformed into a formal program/model transformation. The supporting tool would then check which properties the transformation is preserving or not preserving and warn the user accordingly.

All the above reinforces the need for a formal and executable language to model biology as the core feature of an *in silico* laboratory for biologists that could be the next-generation high-throughput tool for biology.

#### ACKNOWLEDGMENTS

The authors thank Andrew Phillips and Lorenzo Dematté for preparing the figures.

#### REFERENCES

- [1] P. Nurse, “Life, Logic and Information,” *Nature*, vol. 454, pp. 424–426, 2008, doi: 10.1038/454424a.
- [2] J. Fisher and T. Henzinger, “Executable Cell Biology,” *Nature Biotechnology*, vol. 25, pp. 1239–1249, 2007, doi: 10.1038/nbt1356.
- [3] C. Priami, “Algorithmic Systems Biology: An opportunity for computer science,” *Commun. ACM*, June 2009, doi: 10.1145/1506409.1506427.
- [4] A. Regev and E. Shapiro, “Cells as computation,” *Nature*, vol. 419, p. 343, 2002, doi: 10.1038/419343a.
- [5] P. Degano and C. Priami, “Non-interleaving semantics of mobile processes,” *Theor. Comp. Sci.* vol. 216, no. 1–2, pp. 237–270, 1999.
- [6] M. L. Guerriero, J. Heath, and C. Priami, “An automated translation from a narrative language for biological modelling into process algebra,” *Proc. of CMSB 2007*, LNBI 4695, 2007, pp. 136–151, doi: 10.1007/978-3-540-75140-3\_10.
- [7] M. L. Guerriero, A. Dudka, N. Underhill-Day, J. Heath, and C. Priami, “Narrative-based computational modelling of the Gp130/JAK/STAT signalling pathway,” *BMC Syst. Biol.*, vol. 3, no. 1, p. 40, 2009, doi: 10.1186/1752-0509-3-40.
- [8] S. Efroni, D. Harel, and I. R. Cohen, “Reactive Animation: Realistic Modeling of Complex Dynamic Systems,” *Computer*, vol. 38, no. 1, pp. 38–47, Jan. 2005, doi: 10.1109/MC.2005.31.
- [9] A. Phillips, L. Cardelli, and G. Castagna, “A Graphical Representation for Biological Processes in the Stochastic Pi-calculus,” *Trans. Comput. Syst. Biol.*, VII - LNCS 4230, 2006, pp. 123–152, doi: 10.1007/11905455\_7.
- [10] L. Dematté, C. Priami, and A. Romanel, “The BlenX Language: a tutorial,” *Formal Meth. Comput. Syst. Biol.*, LNCS 5016, 2008, pp. 313–365, doi: 10.1145/1506409.1506427.
- [11] L. Dematté, C. Priami, and A. Romanel, “The Beta Workbench: a computational tool to study the dynamics of biological systems,” *Brief Bioinform.*, vol. 9, no. 5, pp. 437–449, 2008, doi: 10.1093/bib/bbn023.
- [12] C. Priami, “Stochastic pi-calculus,” *Comp. J.*, vol. 38, no. 6, pp. 578–589, 1995, doi: 10.1093/comjnl/38.7.578.
- [13] A. Phillips and L. Cardelli, “Efficient, Correct Simulation of Biological Processes in Stochastic Pi-calculus,” *Proc. Comput. Meth. Syst. Biol.*, Edinburgh, 2007, pp. 184–199, doi: 10.1007/978-3-540-75140-3\_13.