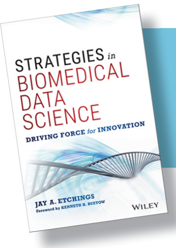
STRATEGIES *in* BIOMEDICAL DATA SCIENCE

DRIVING FORCE *for* INNOVATION

JAY A. ETCHINGS

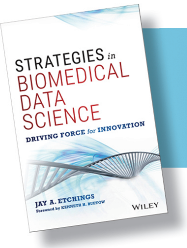
Foreword by KENNETH H. BUETOW

WILEY



STRATEGIES in
BIOMEDICAL
DATA
SCIENCE
DRIVING FORCE for INNOVATION
JAY A. ETCHINGS
Foreword by KENNETH H. BUETOW
WILEY

Contents



CHAPTER 1

Healthcare, History, and Heartbreak

Over the past decade, we have unlocked many of the mysteries about DNA and RNA. This knowledge isn't just sitting in books on the shelf nor is it confined to the workbenches of laboratories. We have used these research findings to pinpoint the causes of many diseases. Moreover, scientists have translated this genetic knowledge into several treatments and therapies prompting a bridge between the laboratory bench and the patient's bedside.

—Barack Obama on the Genomics and
Personalized Medicine Act (S. 976), March 23, 2007

While we are surely poised to continue to make tremendous medical advances—notably in personalized medicine, pharmacogenomics, and precision medicine—we are also facing substantial challenges. The challenges facing healthcare today are many, and if we do not adequately address them we risk missing opportunities, pushing the cost of care up, and slowing the pace of biomedical innovation. In briefly surveying the state of healthcare, it is not my intention to offer a political diagnosis or solution. Rather, it is my intention to use our current technical knowledge to point the way to practical solutions. For example, a long-theorized solution to health records management would be a single cloud-based system where healthcare information sharing exists universally. But if I were to present this as the best technical solution, it would not be my intention to also advocate for a shift to a single-payer healthcare system. As much as possible this book and the discussions in this chapter aim to avoid politics.

After decades of technological lag, biomedicine has started to embrace new technologies with increasing rapidity. Next-generation sequencing, mobile technologies, wearable sensors, three-dimensional medical imaging, and advances in analytic software now make it possible to capture vast amounts of information. Yet we still struggle with the collection, management, security, and thoughtful interpretation of all this information. At the same time, healthcare is changing quickly as the field grapples with new technologies and is transformed by mergers and new partnerships. As a complex adaptive system, healthcare is more than the sum of its parts, and it is always difficult to predict the future. But we do know that as the post–Affordable Care Act healthcare landscape takes shape, the industry is shifting toward digitally enabled, consumer-focused care models. Given these trends, technology will be granted many opportunities to improve patient care.

At the outset of this book it is worth surveying some of the top issues in healthcare. For many of you, these will be quite familiar. Whether you're an expert or not, you should feel free to skip ahead if you like. But it is my sincere hope that the background material will be of real value in bridging the gap between healthcare and biomedicine, on the one hand, and information technology (IT) and data management, on the other. Just as doctors in an age of increasing specialization can benefit from attending to the whole patient, it is very valuable for IT staff to have a more holistic and systemic understanding of healthcare.

TOP ISSUES IN HEALTHCARE

There are many, many sources that comment on the state of healthcare and biomedicine more broadly. Although I worked as a contractor for two of the country's largest Medicare/Medicaid contract holders, I am not a policy expert. But I have come to appreciate the importance of taking in the bigger picture. My admittedly incomplete survey of top healthcare issues is drawn from PwC's *Top Health Industry Issues of 2016* and PwC's *Top Health Industry Issues of 2015* [1]. These two brief reports offer compelling syntheses and analyses of current trends. In rereading these reports and reflecting on my own experiences in the field, I was struck by the number of top issues that are substantially or in part data or IT issues. Many of the top healthcare issues are centrally concerned with the storage, security, sharing, and analysis of data. In other words, IT and data management will be called on to make major contributions to advancing the dynamic healthcare field. Next I explore nine key issues impacting healthcare.

Mergers and Partnerships

As the health sector continues to change in response to the Affordable Care Act (2010), we are seeing many mergers and partnerships. "The ACA's emphasis on value and outcomes has sent ripples through the $3.2 trillion health sector, spreading and shifting risk in its wake. At the same time, capital is inexpensive, thanks to sustained low interest rates. Industry's response? Go big" [2]. Mergers between large insurance providers are consolidating the insurance market. In 2015, the second largest U.S. insurer, Anthem, made a $48.4 billion offer for health and life insurance provider Cigna. Mergers have also been common in the pharmaceutical field, including Pfizer's whopping $160 billion deal for specialty pharmaceutical star Allergan. While these deals are still awaiting regulatory approval, 2016 and 2017 will likely see more mergers and acquisitions. Many new partnerships are also being formed between pharmaceutical, life sciences, software, pharmacy, healthcare providers, and engineering companies, among others.

Mergers, acquisitions, and partnerships are driven by a number of larger market forces. Sometimes predicted lower IT or data costs drive consolidation. More often it is simply that IT and data will need to be able to respond nimbly to these changes. One of the largest challenges is postacquisition data management.

Many providers in the healthcare space have grown through organic means and have survived on shoestring budgets. When compliance moved to the forefront, many chief information officers were granted grace periods to meet compliance and conducted internal audits, patching together existing components to meet the objectives. This expenditure had the systemic impact of preventing the distribution of funds toward infrastructure improvements. The maintenance of many legacy systems resulted, leaving organizations with out-of-date, proprietary, inflexible systems that were simply not designed to interoperate on the larger scale. Now when that smaller provider, which potentially maintains a large collection of Medicare/Medicaid accounts, is acquired by a larger entity, the most significant challenge is the integration of those legacy systems without impacting operational activities. The challenge of migrating years of patient data records into a system from an out-of-date platform encumbered by complex and tangled spaghetti code and created by a resource long since departed is substantial. The need to do so while maintaining business continuity drives many a large entity to maintain the down-level system for years following the acquisition.

Cybersecurity and Data Security

As more and more patient data is stored and shared, security is an increasing concern. Patient data typically contains individualized information. If that data is stolen, the risks of identity theft are substantial, and there exists a thriving black market for stolen health records. Data security breaches are relatively common. “During the summer of 2014, more than 5 million patients had their personal data compromised” [1]. These breaches are often costly for companies. Medical devices themselves can also be hacked. For example, in 2015 the government warned that “an infusion pump . . . could be modified to deliver a fatal dose of medication” [2].

The needs for elastic scalability, rapid provisioning, resource orchestration, high availability, and storage efficiency have contributed to the explosion in cloud providers and niche service offerings. However, this explosion has also opened holes in known security elements that were once sealed. Cloud security challenges can range from the innocuous VM sprawl, where virtual machines are orphaned in an on/off state and fall outside of the domain security policy for things as basic as patching and maintenance [3]. On the other end of the spectrum there would be virtualization hacking, where an adversary gains access to a host (a larger component [server] that houses multiple guests). Hypervisors or virtual machine monitors (VMMs) have been hardened over the years; however, they are only as fortified as their caretakers determine. One key determinant is the organizational structure or culture. A company that owned 100 bare-metal servers in a medium-size data center may have had 10 employees assigned to manage the environment and provide operational support. With the advent of virtualization, workforce reductions have taken place and the distribution ratio of humans to servers has changed. Between 1991 and 2006, a ratio of 1:100 was typical for a large company that provided operational support like a web hosting company [4]. These numbers do not include application specialists and development staff. In today's cloud and highly virtualized environments you could see 1:1,000 ratio of humans (admins) to guest (virtual machines/computers). Efficient providers like Rackspace.com and GoDaddy.com may have 10 to 20 times that ratio [5, 6].

A key component that supports that exponential ratio is robust resource orchestration, which supplies the common ecosystem bits such as backups, network routing, addressing, domain name space management, and availability. Years ago these elements had unique humans as designees owning the responsibility.

Now we can understand how an environment could grow organically, leading to VM sprawl that opens up security gaps. What can be done with orphaned guests long since forgotten by their caretakers?

Adversaries compromise vulnerable virtual machines and enlist armies of botnets or zombie computers assigned to unified tasks [7]. The best-known tasks are distributed denial-of-service (DDoS) attacks aimed at larger public targets, like universities or public businesses. Such large-scale attacks were described in the 11th Annual Worldwide Infrastructure Security Report [8]. Let us also remember that small-scale orphans, like zombies, can still make efficient spam servers, darknet servers, hubs for the distribution of pirated software, and so on. These examples of nefarious computing are familiar to administrators and have just moved to the cloud, where watchful eyes lack the granularity once associated with higher human-to-machine ratios. The relative anonymity behind these expansive and sometimes liberal usage models provides spammers, malicious code authors, and hacktivists opportunities to conduct their activities with relative impunity [9]. Private/public cloud Platform as a Service (PaaS) installations are typically the low-hanging fruit for these breaches, although recent evidence shows hackers targeting some larger Infrastructure as a Service (IaaS) vendors [10]. Hacking of PaaS or IaaS can be referred to more accurately as virtualization hijacking rather than hacking, as the virtual machines are hijacked and enlisted to perform some nefarious task.

Securing Multitenant Hosts

Cloud computing has a key characteristic, the virtualization layer. However, all virtualized systems and cloud systems have underlying components building up the infrastructure (e.g., network, storage, central processing unit, graphics processing unit, etc.) that were not dedicated or optimized specifically for virtualization until recently. The delivery of strong isolation capability in a multitenant environment is typically the first identified gap in security audits. What does it matter if security exists at the guest level if the underlying host can be exploited with a known UNIX kernel exploit? Addressing this security vulnerability can be accomplished through means that are beyond the scope of this text. The suggested reference materials align the host with Defense Information Systems Agency Security Technical Implementation Guide (DISA-STIG) guidelines [11].

Insider Threats

The threat of a malicious insider is a well-known constant to most organizations. This threat is controlled through authentication, authorization, access, and audit mechanisms. Internal or trusted users pose the most significant threat in the form of data leakage. The larger percentage of these potential incidents are managed through policy and audit mechanisms. Users will tend to navigate systems they have access to, whereas the average casual attacker will look for unsecured data and/or servers where a crime of opportunity may present itself. Strategies such as microsegmentation of domains can protect not only against the casual opportunist but also effectively minimize the attack footprint during a breach from an external adversary.

Data Integrity

Data integrity through policy-driven storage management should be central to any cloud, public or private. Accidental or intentional deletion or alteration of records without a backup of the original content can pose a serious security issue in the cloud. Mature organizations have determined classification of data types, which drive access-specific policies that dictate storage parameters and provide audit trails as well. The management policy must securely guarantee that unauthorized or unauthenticated entities are prevented from accessing private data. Exposure to data compromise increases exponentially in the cloud due to the number of transactions and interoperation between private and public cloud edges. Chapter 3 on data management discusses these principles at greater depth.

Encryption

Wide-area network (WAN) traffic is often the target for malicious access. In this section we will not address potential WAN attack vectors. In the previous section we touched on DDoS. But why settle for simple denial of service when you can instead steal a victim's traffic, take a few milliseconds to inspect or modify it, then pass it along to the intended recipient? As evidence, about 1,500 individual Internet Protocol (IP) blocks were hijacked in 2013 [12]. These events last from minutes to days, by attackers working from various countries. Current levels of encryption are often vulnerable to attack from persons with access to supercomputers. Current-generation Suite B encryption (AES 128–256 bit) modalities are gaining popularity and soon will become the standard, as will encryption of L2 extended tunnels (virtual local area networks or VLANs), or what now is being termed Encryption as a Service through public cloud providers. And automated certificate management for identity validation for intercloud connectivity will become standard.

As with many difficult issues, cybersecurity and data security is a question of balance. In particular: What is the right balance between privacy, on the one hand, and convenience and innovation, on the other? Likely this balance will continue to be dynamic. Strong security measures could be a competitive advantage.

In the past, many collaborative efforts have been met with insurmountable security parameters set into place years earlier prior to when many technologies emerged. A simple example would be the inability to share network ports on a device between development groups and production groups years after the VLANs technology was prevalent and accepted by many respected security and compliance professionals. A change to this policy came about in 2013. Centers for Medicare & Medicaid Services regulatory guidelines and DISA-STIG models reflect the update, but bear in mind that VLAN segmentation has been in wide use since 1985 as defined in IEEE 802.1Q. The DISA-STIGs comprise a library of documents that explain very specifically how computing devices should be configured to maximize security. Currently, there are over 400 STIGs, each describing how a specific application, operating system, network device, or smartphone should be configured. The DISA-STIGs represent the best practices in security but also are in the late majority or oftentimes laggards group when it comes to acceptance of new technologies. The consumer demand for innovative healthcare aims to find a balance between required security elements and the delivery of patient-focused healthcare [3].

Homomorphic Encryption

Homomorphic encryption allows for instant-read operations and/or calculations on encrypted information without the need for decryption prior to the read operation. The potential for a reliable model for homomorphic encryption exponentially increases security in cloud computing by facilitating encryption at rest with dual-factor authentication modes for intercloud data transfer or something as simple as computation supplied in the public cloud working on data in the private cloud [13, 14].

The Proliferation of Devices and Apps

Rapid growth in the use of smartphones and medical devices offers great opportunities to innovate how medical care is delivered. The shift toward handheld or "do-it-yourself" medicine is being driven both by a push to lower costs and by customers' desire for convenience. Improved cellular networks mean that consumers spend most of their time in areas with access to high-speed networks. This enables consumers to use apps and connected devices and, increasingly, to share this data in near real time with their medical providers.

From just 2013 to 2015 the percentage of consumers with at least one health or fitness app on their mobile device grew from 16% to 32%. According to one source, "86% of clinicians believe that mobile apps will become important to physicians for patient health management in the next 5 years" [1].

All these devices and apps can seamlessly collect data, but this means more and more data to analyze and archive. This also brings with it challenges around data integration and management.

From the perspective of the patient, there is not too much excitement in standing in line overnight at Best Buy or the Apple Store to get the latest and greatest glucose monitor or heart rate monitor; however, the mobile devices we all know and love may soon have these Internet of Things (IoT)–type features integrated into their hardware and software platforms, allowing for relationship analysis, precision medicine, and whole-life healthcare, where healthcare and condition management becomes fashionable.

A glucose meter that not only measures blood sugar but also tracks it over time will not only help the patient but the industry as well. Imagine if the integration extended to smart devices that tracked diet and exercise, building them into the "whole health" metric and providing exercise, dosing schedule, and dietary guidance to users. This is where the IoT becomes a game changer. Precision medicine can only be as precise as the data it measures or has access to. The ability to collect data from an array of sources will give new freedom to patients on managed care routines that previously could not exist.

Conversely, the challenge in this grand opportunity relates to the tenets of security in the cloud that we discuss throughout the book. There is a very real potential for an adversary to steal data from unsuspecting patients to use in nefariously. This very same deficiency is prevalent in public networked systems now where a power grid or water supply could be taken offline or suffer some type of service interruption. Suite B encryption components and/or emerging encryption techniques for securing network transmissions or data at rest or in flight will need to stay ahead of the would-be attacker. Homomorphic encryption and dynamic certificate services in theory offer next-generation protection.

New Sources of Data Both Public and Private

"High hopes surrounding big data investments in healthcare have been dampened by the challenge of converting large and diverse datasets into practical insights. In 2016, the health industry will begin to use these data in new ways, thanks to high-tech, so-called 'non-relational' databases" [1].

The term "nonrelational database" should be considered synonymous with NoSQL, which for clarification is not no SQL. In fact, NoSQL is short for "not only SQL," meaning *supporting* SQL and more. NoSQL represents a database that does not incorporate the table/key model that relational database management systems (RDBMSs) are built on.

Nonrelational databases are targeted for workloads that need data manipulation techniques and processes designed to provide solutions to data-intensive/big data challenges within the enterprise. Again, the nonrelational model is best known as NoSQL, but there are multiple flavors, all with their own strengths and weaknesses, and all should be considered prior to settling on a database model for your enterprise.

Now let's consider whether a nonrelational database might be a good fit for your use case. If you have designed an application that dynamically creates, harvests, or stores rapidly changing data that spans multiple data types from standard structured, semistructured, and unstructured at large volume and or velocity, you have a strong use case. If your application is intolerant of the 12- to 18-month waterfall development cycle and requires research-style agility in the form of agile sprints, iterating quickly and pushing code weekly, daily, or even multiple times a day, then you may desire this database model. In past well-known database management system (DBMS) models, something as simple as adding a schema extension to an existing database could require a herculean effort where paid service engagements were the only way to ensure success. Perhaps your application was once siloed, serving a finite audience, but is now delivered as a service requiring high availability (always on), global access, and it must have in-built flexibility to be accessible from a host of devices that did not exist a few years ago. This is another situation in which a nonrelational database might work well. And last and possibly most significant, research institutions are adopting scale-out architectures deployed on commodity X86 hardware and largely are vendor agnostic in their open source movement toward distributed cloud computing models.

The mainframe style of large monolithic servers, storage area networks, and locked-in vendor-driven infrastructures are simply not producing competitive results any longer.

These new ways of using data depend on consumers and organizations sharing data. One area where data access promises to improve is clinical trial data. In Europe, clinical trial data for approved drugs is publicly available. "As of October 2014, 520 organizations—including physician groups, patient advocates, government regulatory bodies and one large pharmaceutical company, GlaxoSmithKline—had signed the AllTrials petition, which calls for 'all trials registered, all results reported'" [1].

DATA MANAGEMENT

Another key area is research data management, which has become the number one topic discussed between vice presidents of research and CIOs at the university level. In its simplest terms, research data management amounts to the organization of data throughout its life cycle in your organization. The research life cycle could begin with a data download, generation of de novo artifacts through to the research process, publication and dissemination, and ending with the archiving of valuable results for inclusion in future research initiatives.

It is important to understand that research data management is a key component of the research process and should be as seamless, nonintrusive, and efficient as possible to meet the expectations and requirements of the researcher, the institution/university, funding organization, and legislation. It is also critical that a human within the organization owns the process. One of the hottest new job roles in almost any organization is the director of data management or director of research data management in the case of the university. A quick look at ZipRecruiter shows 1,129 director of data management jobs in Wilmington, Delaware, alone [15]. On a personal note, Arizona State University recently appointed a director of research data management after a 15-month search and recruitment process.

For researchers, research data management should be seamless and nonintrusive to the actual science objectives of that research. The policy of maintaining data artifacts throughout the life span of the funding opportunity has been extended indefinitely, mainly due to the advent of nonrelational database storage. The results from research today that fell short of its objective may be highly useful in discoveries of public value in future years. NoSQL implementations, such as MongoDB, that can store massive amounts of diverse data while keeping the data "warm" and readily available to search queries and perhaps machine learning mechanisms of the future. Be prepared to hear the term "deep learning" as machine learning crawlers continually index data in these massive online collections. The topic of research data management is discussed in depth in chapters 3 and 4 and should be considered central to sustainability in research institutions.

A 2011 white paper further explains the importance and changing conventions of research date management:

> The scientific process is enhanced by managing and sharing research data. Good research data management practice allows reliable verification of results and permits new and innovative research built on existing information. This is important if the full value of public investment in research is to be realized. These principles have been recognized by key stakeholders: most Research Councils now have policies in place which encourage or mandate the creation of a research data management plan and the deposit of research data in a recognized data center where such exist. Many leading journals require underlying datasets also to be published or made accessible as part of the essential evidence base of a scholarly article. [16]

The Precision Medicine Initiative

In 2015 President Barack Obama launched the Precision Medicine Initiative. Its mission is "To enable a new era of medicine through research, technology, and policies that empower patients, researchers, and providers to work together toward development of individualized care" [17].

The White House's website explains further:

> The future of precision medicine will enable health care providers to tailor treatment and prevention strategies to people's unique characteristics, including their genome sequence, microbiome composition, health history, lifestyle, and diet. To get there, we need to incorporate many different types of data, from metabolomics (the chemicals in the body at a certain point in time), the microbiome (the collection of microorganisms in or on the body), and data about the patient collected by health care providers and the patients themselves. Success will require that health data is portable, that it can be easily shared between providers, researchers, and most importantly, patients and research participants. [17]

In March 2015, Dr. Francis Collins, director of the National Institutes of Health, tasked a Working Group of his Advisory Committee to the Director to develop a plan for creating and managing a large research cohort. On September 17, 2015, Dr. Collins accepted the framework outlined in the Working Group report and began building the infrastructure so that participants can begin enrolling in the cohort in 2016 [18].

Research institutions have long understood that the more we learn, the less we know or the more we realize we do not know. The idea had been that once the code of the genome was decoded, we would have the insight to eradicate many disease conditions once and for all. However, this great discovery in modern science served to illuminate how much we really did not know about how gene signaling and biomarkers interoperate. There is a huge demand now for greater insights into the biological, environmental, and behavioral factors that influence disease conditions. A staggering number of diseases lack any reliable and reproducible treatments. Precision medicine offers the best approach for disease prevention and treatment. The science of evaluating individual variability in genes, hereditary factors, lifestyle, and environment factors for each unique patient promises to unlock targeted treatments that will be breakthroughs in the coming years. Evaluation of epigenetics for changes in organisms caused by modification of gene expression rather than alteration of the genetic code itself is just one of the emerging areas directly related to precision medicine. Over the past 10 years we have witnessed significant advances in precision medicine; however, there is quite a bit of work to be done at the federal, public institution, and private industry level to realize precision medicine's full value. In this author's opinion, we are still at a point where we know only a minuscule amount of what precision medicine holds for the future of medicine.

BIOSIMILARS, DRUG PRICING, AND PHARMACEUTICAL COMPOUNDING

Biosimilars, drug pricing, and pharmaceutical compounding are key drivers and benefits of precision medicine initiatives. They pave the road for pharmacogenomics. Biosimilars are often created using genetic technology and made from sugars, proteins, or engineered cells and/or tissues. It is worth clarifying that not all biosimilars are made from genetic technology. Two popular and well-known biologics are adalimumab (Humira) for rheumatoid arthritis and trastuzumab (Herceptin) for breast cancer. Due to the complex research required to engineer these treatments, they can be expensive; biologics can run $50,000 a year or more. It is no easy feat to create a biosimilar drug; it is far more challenging than creating the generic version of a brand-name drug, which typically just re-creates the same chemical recipe in a different preparation. Most common drugs are made from chemicals having a known chemical structure. Biosimilar and biologic drugs are far more complex.

Biosimilar drugs, while similar to the biologic drugs in target and purpose, have "allowable differences because they are made from living organisms," according to the FDA [19]. The FDA keeps a watchful eye on this emerging science, which again is another product of pharmacogenomics. Recently the FDA has gained the authority to approve biosimilar products under a provision of the Affordable Care Act.

PROMISING AREAS OF INNOVATION

The top issues in healthcare, discussed earlier, will be met at least in part by advances in data management, IT, and other technical innovations. While these innovations will likely occur in somewhat surprising ways and cut across numerous fields, I see six areas as particularly promising and worth a relatively high-level discussion.

The Internet of Things

Goldman Sachs, in a 2014 equity research report, listed the Internet of Things as a megatrend, explaining:

> The Internet of Things (IoT) is emerging as the third wave in the development of the Internet. The 1990s' fixed Internet wave connected 1 billion users while the 2000s' mobile wave connected another 2 billion. The IoT has the potential to connect 10X as many (28 billion) "things" to the Internet by 2020, ranging from bracelets to cars. [20]

Cisco Internet Business Solutions Group, as shown in Figure 1.1, places the number of interconnected devices even higher: at 50 billion by 2020.

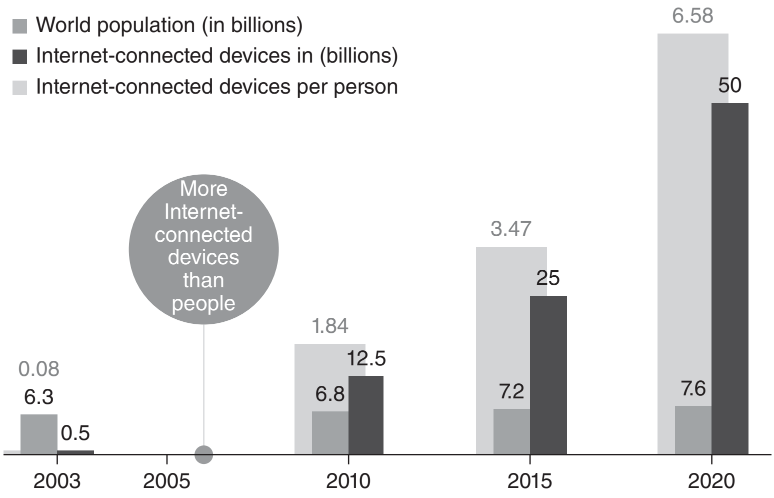


Figure 1.1 Growth in Internet-Connected Devices by 2020
Cisco IBSG, April 2011

Now is a terrific time for biomedicine to jump onboard IoT as sensor prices have dropped by 50% in the past 10 years and processing also has dropped as raw computational cycles have dropped 50 times in price per floating-point operation per second (FLOP). Wireless networks are ubiquitous, with the price of Internet bandwidth declining as much as 40 times over the past 10 years; smartphones are ever more common and increasingly incorporated into automobiles; and Internet Protocol version 6 (IPv6) is in wide deployment across large-scale provider networks. Software-defined networking (SDN) is prevalent on provider backbones, and of course big data analysis is on the minds of many software development firms readying for the volumes of data that will be produced by the IoT.

Figure 1.2 shows many of the companies currently developing IoT projects. There is no shortage of investor interest; some analysts estimate that as much as $5 billion in mergers, acquisitions, and venture capital will be raised in 2016. There are certainly infrastructure challenges to the collection and aggregation of data; however, the question remains: Can wearables (IoT) improve healthcare? As we continue to learn in the realm of big data analysis, the challenge lies in interpreting the data and creating applications that make the data actionable. We already know that wearables can collect wide varieties of data. For example, exercise and activity levels, sleep quality measurements, heart-rate

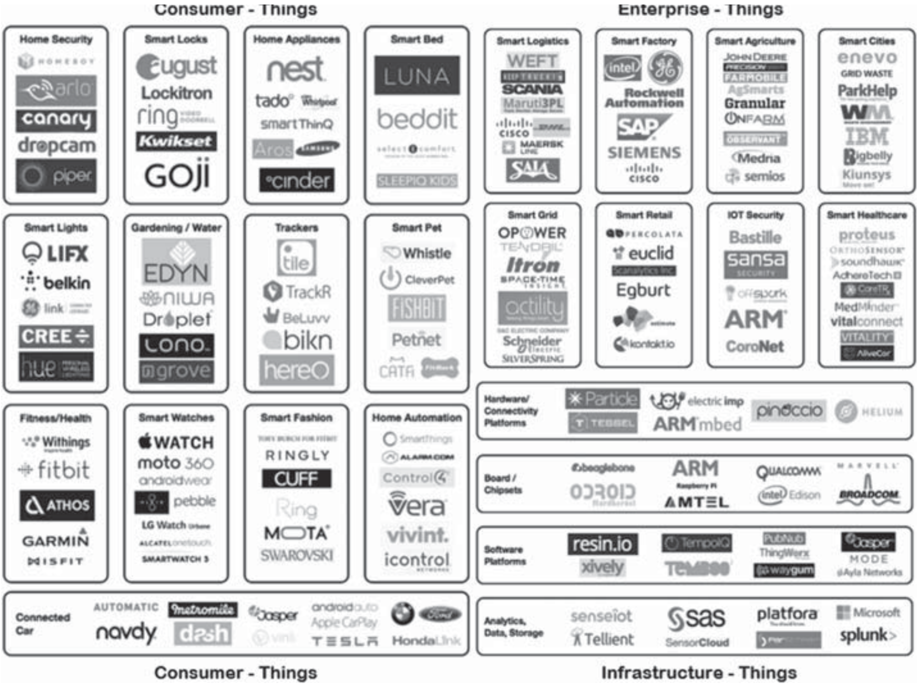


Figure 1.2 Brands in the IoT Realm

values, and blood sugar readings can benefit patients with risk factors for congestive heart failure, diabetes, and arrhythmias. In the next year we expect these devices to collect even more biometric data, including: detection of galvanic skin response through noninvasive methods, which can be critical in monitoring and understanding the stress response; blood glucose levels, which have the potential to reduce the cost and burden of managing diabetes; and tracking of pulse transit time to monitor blood pressure in real time.

Continued innovation in this realm aims to make healthcare devices ubiquitous. There are only two main potential roadblocks to widespread adoption. The first pertains to how device data will be treated by FDA regulatory guidelines for delivery of clinical care. Many conditions require close monitoring from FDA-listed devices, and this should remain standard practice. However, even when treating these conditions, and certainly when monitoring less critical patients, wearable devices can provide contextual, continuous data that helps connect the dots between regulated medical device readings and provider encounters.

The FDA's role continues to grow and evolve as it races to keep pace with emerging technologies. The FDA works to protect consumers while attempting to not inhibit discoveries of public value. This challenge extends to the IoT and medical devices. The FDA, in collaboration with the National Health Information Sharing Analysis Center, the Department of Health and Human Services, and the Department of Homeland Security, hosted a public workshop titled "Moving Forward: Collaborative Approaches to Medical Device Cybersecurity" in 2016. [21]. Such collaborations will likely become more common in the coming years.

The second challenge area is security. Insecure software/firmware, inability to provide physical security, potential identity theft, other unforeseen privacy concerns, insecure network access methods, weak encryption or lack thereof, and insufficient authentication/authorization all come to mind as both challenges and potential opportunities as the IoT continues to grow. The IoT faces all the same security challenges as, and perhaps more than, other types of networking and data sharing.

Data Visualization and Imaging

One path to understanding data is by seeing it. Data visualizations can be key to unlocking the path from data to information to knowledge. The ability to collect and explore complex data leads to an inevitable desire to see the data for interpretative analysis through the next generation of tools. According to Suhale Kapoor, the cofounder and executive vice president of Absolutdata Analytics: "Visuals will come to rule: The power of pictures over words is not a new phenomenon—the human brain has been hardwired to favour charts and graphs over reading a pile of staid spreadsheets. This fact has hit data engineers who are readily welcoming visualization softwares that enable them to see analytical conclusions in a pictorial format" [22]. Tools that present information in complex data as visual representations have matured and continue to grow in adoption.

Visualization leverages knowledge from data, driving more adaptive and dynamic visualization tools. The charts and graphs of the past are still compelling in their own right, although they are static and lack the real-time feel of the adaptive nature of "live data." Data visualization tools will continue to move beyond graphs and dynamically open new windows into the potential of simulations for every science domain imaginable. Dynamic dashboards, three-dimensional simulations, and automated diagnostic systems populated by incoming data sources reflecting up-to-the-minute fluctuations reveal hidden insights that would otherwise go unnoticed. These are perhaps some of the most exciting opportunities for biomedicine on the horizon.

Visualization has undergone a recent refresh as the advent of big data required a translational tool to become "human readable." Looking at human genomic data, which for the most part is a collection of base pairs, tells even the most astute researcher little. However, with the right visualization tool, that same data becomes more than coordinates; it becomes a map, or perhaps the Google Earth of life sciences. It should be noted that although we can sequence a human genome for $1,000, this does not include any analysis. The most significant costs are in the postprocess and analytics phases.

We know there are 3 billion diploid base pairs but 6 billion haploid sequences (because half come from your mother and half from your father). Discoveries mined from this extensive collection of data can be examined in postprocess pipelines in genome browsers with relative ease, comparatively speaking, when considering the enormity of the collection. Numerous methods have been developed to automate the analysis of genomic data. Nonetheless, the visual exploration of alterations in cancer genomes, epigenomes, and transcriptomes in multidimensional data sets and of the relationships among these alterations presents specific challenges. Schroeder, Gonzalez-Perez, and Lopez-Bigas's paper details the many offerings of targeted tool sets for genomics visualization [23].

Data Storage

The race is on to capture your data. The big cloud players, including Amazon Web Services, Dropbox, BOX, and Microsoft Azure, are in a heated competition to become the one-stop shop for institutions everywhere to move their data to. These providers are planning for the future and are not simply after flat file, unstructured, typical internal data assets. They are also planning for the storage of research data, web platform data, structured data from various

RDBM systems, and mobile device and sensor data gathered in emerging IoT efforts. About 85% of IoT data at this point is of no use to the companies that collect it. However, someday it may be, and that makes it worth retaining. The challenge now for institutions is determining which approach is sustainable over the long haul from compliance, financial, and availability perspectives. Is Hadoop, either in the cloud (Amazon Elastic MapReduce) or on premises, a viable solution? Should we still look toward an infrastructure of large-scale storage area networks that have evolved into data management platforms in many cases? Is a distributed server–based storage system (ephemeral) the best method to ensure robust scalability? In the coming chapters we take a deeper look at these options, share our experiences, and allow readers to determine which option or options will best meet their enterprise requirements.

Data Analytics

Data scientists, statisticians, and analysts are quickly moving to the forefront of many environments. Enterprises now expect that data scientists will be able to wrangle data, mine valuable insights, build complex models, identify difficult to discern patterns and relationships, and use data to predict future outcomes. Researchers and senior leaders recognize that uncovering insights that are locked away in the vast amounts of data is critical. Prescriptive analytics have replaced descriptive analysis or sentiment analysis, and statistical testing is now more commonplace in research sciences. Enabling the personalized medicine revolution will require a diverse collection of analytic tools including R, Apache Mahout, and Spark as well as custom tools to fuel novel genomic analysis and the integration of multidimensional molecular and clinical data.

The challenges of big data are well understood, while the benefits at this point are only imagined. Due to architected deployment options, our ability to gain insights from diverse forms of data previously considered difficult data types and formats is now much greater. With the democratization of big data, deep data analysis is nearly limitless in its scope.

Compute Capabilities

Data-intensive (big data) analytic frameworks and traditional high-performance computing (HPC) have evolved along diverse paths over the past 10 years. However, they are slowly converging again as institutions begin to find target workloads for each. Hadoop is certainly not the magic bullet, and the traditional "big iron" community cluster does not meet all aspects of many mature research pipelines. Typically, Hadoop (big data) is regarded to be about "data" and HPC clusters are about “computational compute.” The current confluence of big data, computation, and analytics is driving “data-intensive compute” with the goal of solving the grand challenges. Solving these grand challenges will require a fundamental redesign of enterprise infrastructures as well as legacy thought processes. Thoughtful choreography of a diverse collection of physical and logical capabilities that perform as an integrated whole will attempt to overcome the physical limitations of Moore’s Law. Delivery, open access, and support of these diverse environments will enable tomorrow’s brilliant researchers to reveal new horizons in deep machine learning, artificial intelligence, cryptography, and complexity sciences, continuing to push the boundaries of emergent architectures toward realized quantum computing.

Cloud

The cloud is ubiquitous, and widespread adoption at extreme volumes will continue as cloud offerings right-size their consumption models for greater efficiency and competitiveness. Data is the driver for much of the public cloud’s growth, and the cloud is becoming more useful as it expands from PaaS offerings to IaaS and raw compute (bare metal nodes versus virtual nodes). Analytic tools are also on the rise as Microsoft Analytics (which utilizes Hortonworks for its underlying Hadoop distribution), Amazon Redshift, and Google BigQuery gain ground and customer footprint.

Playing it safe in your cloud adoption is no longer considered a progressive strategy. The most prevalent strategy from 2016 to 2020 will be to determine which components to maintain in the on-premises portion of your hybrid cloud and which to farm out to public providers, and how to remain cloud vendor agnostic with a high availability, elastic, and dynamically mobile public presence. When one public provider fails to meet the desired service-level workloads, enterprises will adaptively relocate to another provider or to the on-premises capacity without service interruption. Cloud solutions and services will continue to innovate to support this model.

Emerging compliance models, legacy applications, and laggards will hold a portion of the IT roadmap landlocked on premise. But make no mistake: A paradigm shift is under way in how organizations understand and approach cloud adoption. As touched on earlier, not every aspect of research technology will be cloud ready in 2017 or perhaps even 2020; quite candidly, research institutions represent a relatively small customer base and therefore will be a bit behind public enterprise adoption. The current trend will continue as elements continue to move into the cloud. As we work through the remainder of the text, one specific aim is to inspire thought around hybrid-cloud architectures, highlighting achievable outcomes through integration of on-premises and public cloud-based resources.

CONCLUSION

Healthcare and biomedical research are incredibly dynamic sectors. Healthcare is still adapting to the mandates and policies of the Affordable Care Act, while biomedical research is expanding into precision medicine, the IoT, and advanced computational modeling (to name just a few innovative directions). There are many substantial challenges facing healthcare and biomedical research. Notable for this book is the fact that many of these challenges have substantial data or IT components. In other words, many will be solved, at least in part, by advances in data storage, analytics, network, and systems architecture. The chapters that follow explore these and other topics.

NOTES

1. Health Research Institute. December 2014. "Top Health Industry Issues of 2015." https://www.pwc.com/us/en/health-industries/top-health-industry-issues/assets/pwc-hri-top-healthcare-issues-2015.pdf.
2. Health Research Institute. December 2015. "Top Health Industry Issues of 2016." https://www.pwc.com/us/en/health-industries/top-health-industry-issues/assets/2016-us-hri-top-issues.pdf.
3. Steven Warren. December 9, 2008. "What Is Your Best Definition of VM Sprawl?" *TechRepublic*. http://www.techrepublic.com/blog/virtualization-coach/what-is-your-best-definition-of-vm-sprawl/.
4. Mark Verber. 2008. "How Many Administrators Are Enough?" Updated December 1. http://www.verber.com/mark/sysadm/how-many-admins.html.
5. Rackspace Hosting. "Rackspace: Managed Dedicated & Cloud Computing Services." https://www.rackspace.com/.
6. GoDaddy. "Domain Names | The World's Largest Domain Name Registrar—GoDaddy." https://www.godaddy.com/.
7. Imperva Incapsula. "Botnet DDoS Attacks." https://www.incapsula.com/ddos/ddos-attacks/botnet-ddos.html.
8. Arbor Networks. 2016. *Worldwide Infrastructure Security Report*. https://www.arbornetworks.com/images/documents/WISR2016_EN_Web.pdf.
9. SearchSecurity. "Definition: What Is Hacktivism?" http://searchsecurity.techtarget.com/definition/hacktivism.
10. Kevin Kell. November 11, 2013. "EC2 Security Revisited." Learning Tree Blog EC2 Security Revisited Comments. http://blog.learningtree.com/en/ec2-security-revisited/.
11. IASE. "Security Technical Implementation Guides (STIGs)." http://iase.disa.mil/stigs/Pages/index.aspx.
12. Dyn Research. November 19, 2013. "The New Threat: Targeted Internet Traffic Misdirection." http://research.dyn.com/2013/11/mitm-internet-hijacking/.
13. Craig Gentry. September 2009. *A Fully Homomorphic Encryption Scheme*. PhD Thesis. https://crypto.stanford.edu/craig/craig-thesis.pdf.
14. Andy Greenberg. November 3, 2014. "Hacker Lexicon: What Is Homomorphic Encryption?" Wired.com. https://www.wired.com/2014/11/hacker-lexicon-homomorphic-encryption/.
15. ZipRecruiter. "1,129 Director Data Management Jobs in Wilmington, Delaware." https://www.ziprecruiter.com/jobs/delaware/wilmington/director-data-management.
16. A. Whyte and J. Tedds. September 1, 2011. "Making the Case for Research Data Management." Digital Curation Centre. http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm.
17. White House. "White House Precision Medicine Initiative." https://www.whitehouse.gov/precision-medicine.

18. National Institutes of Health. "Precision Medicine Initiative Cohort Program." https://www.nih.gov/precision-medicine-initiative-cohort-program.

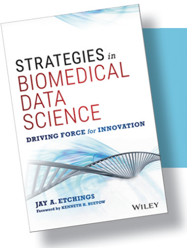
19. U.S. Food and Drug Administration. Updated August 2015. "Information for Consumers (Biosimilars)." http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsare-DevelopedandApproved/ApprovalApplications/TherapeuticBiologicApplications/Biosimilars/ucm241718.htm.

20. Goldman Sachs. September 3, 2014. "Internet of Things Primer." http://www.goldmansachs.com/our-thinking/outlook/internet-of-things/iot-report.pdf.

21. U.S. Food and Drug Administration. 2016. "Moving Forward: Collaborative Approaches to Medical Device Cybersecurity," Public workshop, January 20–21. http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm474752.htm.

22. Quoted in Bruce Robbins, "5 Predictions for 2016 on Data, Analytics and Machine Learning." January 3, 2016. Data Science Central. http://www.datasciencecentral.com/profiles/blogs/5-predictions-for-2016-on-data-analytics-and-machine-learning.

23. M. Schroeder, A. Gonzalez-Perez, and N. Lopez-Bigas. 2013. "Visualizing Multidimensional Cancer Genomics Data." *Genome Medicine* 5, no. 9, https://genomemedicine.biomedcentral.com/articles/10.1186/gm413.



About the Author

Jay Etchings is the director of Operations for Research Computing at Arizona State University (ASU). Research Computing is a new initiative led by the university's most senior leaders addressing fluid technical environments that support highly computational workloads, petascale data analysis, next-generation cybercapabilities, and emerging network innovation such as 100Gbps Internet2 and Science DMZ. It is a mission to architect, develop, deploy, and support innovative architectures addressing emerging challenges of fourth-paradigm science. Research Computing serves the ASU research community and also publishes on its original research initiatives and presents at conferences and events. Current projects include development and support of next-generation big data solutions for research medicine; the Hortonworks–Mayo–ASU genomics platform; development of biomedical informatics tool sets for use in highly parallel environments; and OpenStack Cloud development for life sciences utilizing Big Data ecosystem components (data-intensive research), including integration and advancement of the Internet2 Innovation platform, extension of ubiquitous research compute to all, and the proliferation of software-defined networking. Mr. Etchings also has appointments with the Open Daylight Foundation (Linux Foundation), the Open Networking Foundation, Internet2, and the Open Fog Consortium and has served at the National Science Foundation as a proposal panel reviewer.