Within the white-hot topic known as big data, the Internet of Things, or IoT, stands out as an example of how the 21st-century data deluge can transform business, healthcare, science, and society. But what exactly is the IoT and what needs to be done to fully realize its potential?

The term IoT was coined to describe data created by Internet-connected, mobile or stationary, embedded or stand-alone, objects and devices. It covers a wide range of applications: home appliances that automatically adjust lighting or temperature; portable medical devices that relay personal medical information to health care providers; sensors that report environmental data; machines and engines that stream performance and diagnostic data; and automobiles that recalibrate driving routes based on traffic and weather conditions. The growth of the IoT is staggering. The Internet now connects more objects and devices than people and today’s 25 billion Internet-connected devices are expected to double to 50 billion by 2020, equivalent to six devices per person worldwide (Evans, 2011). A 2015 McKinsey report projects up to $11.1 trillion in 2025 economic yield from IoT applications (Manyika, et al. 2015).

To harness the power of the IoT, many questions and challenges must first be addressed. Among them are:

- **Security**: What security challenges exist and how do we address them?
- **Provenance/Ownership**: Who will own shared data and publicly available data?
- **Privacy**: How will organizations protect employee privacy while using the IoT to track their performance and productivity?
- **Location**: Should data processing and analysis take place at the source of the data (i.e., the edge) or at centralized data centers? How can organizations compute and analyze data that is at the edge and in motion?
- **Standards**: Are data standards needed and possible, and if so, how do we develop them and promote widespread adoption?
- **Workforce development**: What training and skills will employees need to work with IoT data?

The NCDS brought together a small group of thought leaders to discuss these and other questions during a two-day workshop held July 29 and 30, 2015 at Cisco headquarters in San Jose, CA. Out of this workshop, the key IoT-related data challenges that science, industry, government, and the general public face today were enumerated. This white paper summarizes these challenges:

- Balancing ethical concerns such as privacy with the need to leverage data for competitive advantage.
- Protecting data from unwanted exposure, and ensuring data provenance.
- Navigating the maze of data standards, metadata schema, and ontologies to capitalize on their benefits.
- Handling big data, including storage and needed computational power.
- Developing training for individuals who must collect, share, manage, analyze, and preserve data.
The Industrial Internet presents unprecedented challenges for the employer-employee relationship (Ciocchetti 2010). Never before have employers been able to collect data on so many employee activities. UPS, for example, equips its delivery trucks with dozens of sensors that transmit data on everything from how quickly a driver delivers a package to whether or not a driver is regularly using a seatbelt (Goldstein 2014). The NFL fits all player uniforms with RFID chips to track performance on the field (Moynihan 2015).

In the age of interconnectivity, the extent of employees’ abilities to act freely within the workplace is no longer clear. A balance must be negotiated between the right of an employee to privacy and the right of an employer to use employee data to gain a competitive advantage. Although employees and employers have different concerns and objectives as they work through this issue, both can benefit from workplace monitoring. For example, a high-performing employee may want to show his/her performance data to a prospective employee or a new manager. Likewise, an employer may garner employee loyalty if the company clearly demonstrates the benefits of workplace monitoring (e.g., through employee raises and promotions) and when the data collection process is transparent. Transparency is crucial, as the IoT is in the “hype stage” of technology adoption, also known as the “privacy panic cycle” (Figure 1; Castro 2015). Thus, employees must be aware of and agree to the data that is collected about them and, when appropriate, maintain ownership of their data.

The IoT has also blurred the lines between private citizens and employees. Work places, hours, and devices are no longer fixed. Employees often work from home, after hours, and using personal devices. Personal email accounts and cell phones are frequently used for business communications, yet few businesses regulate such behavior.

New policies must be developed in parallel with new technologies to ensure transparency and maintain employee autonomy and privacy, while allowing employers to use workplace and employee data to improve the bottom line. Policies must also address when, where, and how an employer may collect data on its employees. Remediation plans must be part of these efforts in order to ensure a rapid response to data breaches or secondary use of employee data, especially sensitive data and proprietary information.

Monitoring and capturing data about workplace behavior also impacts those behaviors and changes the work environment. Numerous organizations, including the Federal Trade Commission (FTC 2015), have expressed concerns related to privacy in the era of the IoT. However, it is unclear how behavior will change as a result of constant surveillance and measurement. New psychosocial approaches must be developed to measure the impact of data surveillance and monitoring on humans not only while at work, but also in public places and at home.
Data security is a challenge that makes CTOs loose sleep. The recent security breaches of major retailers, financial institutions, and healthcare systems (Figure 2), including Target, JP Morgan Chase, and Anthem Health, affected millions and caused not only personal anxiety but, in many cases, the release of sensitive information, including credit card numbers, bank account numbers, and personal health information. Moreover, recent research indicates that home devices, such as thermostats and webcams, can be compromised by novice hackers (Picchi 2015). Even more alarming is the recent demonstration of real-time hacking of a Jeep Cherokee—with the driver in it (Greenberg 2015).

The IoT, however, presents new challenges in data security (Kuukka 2014). For example, leaking proprietary data could seriously compromise the financial security of a business. Such leaks remain a real possibility, partly because of the historic separation between Information Technology (IT) and Operational Technology (OT) in most businesses (Hameed 2014). While the security needs of IT and OT differ, with IT focused on business operations (e.g., human resources, sales, etc.) and OT focused on mission-critical physical operations (e.g., engineering production workflows), the two branches must converge so that IT’s security policies, procedures, and incidents/breaches can inform those of OT and vice versa.

Device authentication poses another critical challenge related to data security. As businesses tap into data from external sources, such as sensors, and as businesses share data on their operations (e.g., customer base) or technologies, new approaches and methods to authenticate IoT and Industrial Internet devices are essential to ensure the integrity and security of all data. This is particularly important for distributed networks of objects and devices and for complex industrial operations that depend
on authentication as new devices are added to the system. In these complex industrial situations, IoT devices may come and go and a changing variety of devices must communicate with each other to keep operations running smoothly.

Data provenance, or a record of origin and ownership of data, must be interwoven into all security plans in order to ensure the validity of data, provide proper attribution, and prevent unauthorized access. As new devices and objects are created and added to existing networks, a balance must be reached between security needs and industry needs. Businesses will desire both access to data and the ability to maintain proprietary rights to data and trade secrets. A balance also must be reached between the rights of an individual or entity and the public good. For example, most people would agree that data about safety in the workplace or safety of a piece of equipment should be shared with the general populace. But few would likely support the release of individual or organizational data on performance, or secondary data derived from analytics. As more devices cross-communicate over networks and as data gets reused over its lifecycle, the question of who owns what data and how to maintain that ownership becomes critical. Dynamic policies and practices must be developed to maintain data provenance while promoting the secure sharing, reuse, and modification of data over the data lifecycle.
Numerous organizations, such as the International Organization for Standardization (ISO), the PCI Security Standards Council, the IEEE Standards Organization, and the Internet of Things Global Standards Initiative, have developed standards to guide electronic data and information sharing and use. However, these efforts remain fragmented and are sometimes industry- and application-specific (e.g., Aijaz & Aghvami, 2015; Keoh, et al. 2014).

Semantic Web techniques, such as structured vocabularies, may help supplement the efforts of standards organizations. Structured vocabularies, such as ontologies and metadata, provide meaning and context to data and can be used as standards in and of themselves. Ontologies enable computers to reason and abstract knowledge by structuring the vocabulary in terms of subject-predicate-object syntax and by applying first-order predicate logic to form inferences. Metadata is a form of rich annotation of data objects, enabling discovery, preservation, and provenance capture.

One universal ontology or metadata scheme while ideal, will not meet the needs of all industries. Instead, industry-specific ontologies (or metadata schemes) must be developed and mapped to one another. Standardized vocabularies should be developed with different user roles and needs in mind. For example, a buyer may want pricing information while an engineer may want specifications. Additionally, ontologies or other standardized vocabularies should support a hierarchy of terminology (similar to the way taxonomies do) so that applications can perform efficiently based on the level of complexity needed.

- On an industry-specific basis, ontologies should include terminology for the following levels of abstraction:
  - The lifecycle of objects (i.e., what digital information is required at each stage of the object’s lifecycle of design, build, operate, and maintain);
  - The personality of devices (i.e., how they function and relate to other entities); and
  - The most appropriate configuration management scheme (i.e., the most cost-effective engineering design).

Challenges to standards adoption include:

- Cost, both human labor and monetary, for developing standards and structuring data according to standards;
- Incentivizing contributions to standards or vocabulary development (McDonald 2014);
- Keeping up with new terminology for emerging products; and
- Detection of invalid data that may be introduced into the system.

Potential solutions to these challenges could include automating (e.g., data mining and machine learning) tasks such as seeding and mapping ontologies. Gamification techniques may encourage contributions; and crowdsourcing may be a cost-effective way to verify accuracy.
With increasing amounts of data captured from remote locations using sensors, video cameras, and other recording devices, historical paradigms of moving raw data to centralized sites for storage, computation, and analysis may be outdated. Indeed, storage and computation can, at least theoretically, take place at the data source, also known as the “edge.” Nonetheless, the computing power and storage capabilities required for optimization, prediction, machine learning, and other advanced analytics typically reside in central locations. In addition, industries increasingly demand real-time analysis to guide minute-to-minute decision-making, activities which often require moving data from a remote source to a centralized server for analysis.

The analytic challenges and required computational power for analytics at the edge are as yet unknown and must be understood in order to compare edge to center-based analytics. The quest to answer these questions must be guided by the increasing need for real-time analytics, and by weighing the monetary costs of data storage and the security risks of edge analytics. In addition, data scientists must develop hybrid algorithms for distributed analytics across the hierarchy from the edge to the center.

Many limitations exist for long-term data storage and curation at the edge, including smaller available “footprints” (Scheier 2010) and a lack of centralized control (Patki 2013). Furthermore, when analytics are moved to the edge, statistical power typically is reduced, which can compromise the validity of conclusions. A better understanding of how to maximize data storage and analytics at the edge will enable efficient edge analytics, including the recovery of edge data for downstream analytics.

Figure 3. Data requirements for Industrial Internet applications. (Source: IDC, A 3x3 opportunity matrix for big data and analytics, Publication #247766, April 2014, as reprinted in IDC/GE, Next generation, Industrial Internet apps: all about the data, September 16, 2015)
Data clean-up—the process of removing errors and inconsistencies caused by typographical mistakes, missing entries, invalid entries, etc—poses another challenge for edge analytics and storage. While data clean-up is critical, the process relies more on humans than on automation and typically takes place using a centralized server. These manual processes are impractical for the IoT, which often involves real-time analysis on data in motion. New data clean-up approaches that are automated and adapted for industry-specific data sources will be needed for analytics at the edge.

One key to comparing the edge to the center in storage and analytics is realizing where IoT data comes from and how it is likely to be used. IoT data often involves real-time, ever-changing data flows (data in motion), with many devices at different locations capturing many types of data (e.g., structured, semi-structured, unstructured) (Figure 3). Data flow is multifaceted and involves device-to-device, device-to-proxy, device-to-private cloud, and device-to-public cloud movement. The needs of the user—for volume, speed, cost efficiency, and more—as well as available networking, computing and storage architectures also factor into the edge vs. center comparison.

Finally, new approaches are needed for encrypting, tracking, and auditing data at the edge and in motion to maintain security for data stored, computed on, and moved between the edge and the center. While some systems enable tracking and auditing of data flow and access, such as Apache Kafka, today’s technologies are difficult to implement on a large-scale or with distributed data systems, and thus are impractical for the IoT. Furthermore, few existing approaches track the copying and deletion of data, whether authorized or not. New technical approaches must be developed to enable reliable, long-term audit trails to track data usage and compliance with accepted policies over time.
In the coming decades, the competitiveness of U.S. businesses will depend on their ability to use the IoT to speed global innovation in new technologies, products, and services. And that will require a talented workforce that includes expert data scientists, and data-savvy workers and managers.

The exact skills required of the IoT workforce are still unclear. As recently as 2011, a McKinsey Global Institute Report argued for talent in data analytics and put forth a projection that by 2018, the United States will have a shortfall of 140,000 to 190,000 “deep” analysts and an additional shortfall of 1.5 million general analysts and managers with analytical expertise. That same report predicted that investments in analytical talent and training would produce a greater than 60% increase in operating margins across industry sectors. However, McKinsey’s follow-up reports have found that analytical skills alone are insufficient to harness the power of big data, with industry leaders reporting minimal return on investments in analytics, largely because analytical findings are too often not actionable (Callinan, et al. 2014; Court 2015). The failure of analytics to guide decision-making is driving a need for a new type of talent, something McKinsey refers to as the “translator” (Ariker, et al. 2014). As envisioned, a translator bridges two complementary fields, for example computer science and energy or statistics and automotives (Figure 4). Data scientists are theoretically well positioned to serve as translators, but the requisite interdisciplinary skill set has yet to be defined.

**Figure 4:** Examples of interdisciplinary training to meet today’s needs in data science for the Internet of Things. (Image: RENCI, University of North Carolina at Chapel Hill.)
Moreover, traditional training programs remain specialized for fields such as information technology, computer science, and statistics. While these programs have value, they will not meet emerging needs (Mintz 2014). The development of new training programs will require an understanding of the interdisciplinary skill set for a data scientist and a clearly defined training path that details the steps required to obtain those skills. Training must begin well before college and will need to continue throughout a worker’s career. Also, academia must move beyond outdated department-based training and education to embrace new models for education, training, and professional development, including modular courses and professional certificate programs. The academic structure itself must be modified to encourage innovation and interdisciplinary training, including on-the-job training and a tenure system that emboldens faculty to seek cross-departmental training and career opportunities. Furthermore, new technologies can provide new kinds of educational experiences, many of which hold promise but have yet to be fully evaluated. These include simulation-based learning, gamification, and MOOCs (massive open online courses).

Perhaps most important is the need for collaboration between academia and industry. If the U.S. is to lead the world in innovations gleaned from data, it must lead in producing a pipeline of talent for the IoT. Fundamental to creating this pipeline will be academic-industry collaborations that lead to educational programs and new data science models that tackle the real-world problems of industry. Industry needs and priorities must guide new academic interdisciplinary training programs, including industry practicums, internships, and apprenticeships. Such partnerships must serve as the cornerstone of all efforts to create and evaluate new interdisciplinary training programs in data science.
SUMMARY

The data science challenges of the IoT have exposed critical needs for new policies, workforce training, and research and development. U.S. industries will fall behind in the global economy if major investments are not made by academia, government, and industry to meet these needs. Financial investments alone will be ineffective in the absence of true partnerships among the businesses who want to make use of IoT data and the colleges, universities, and community colleges that train data scientists.

This white paper outlines some of the key issues that need to be addressed and the conversations that must be initiated among stakeholders related to the IoT. It is designed to stimulate further discussions that lead to concrete plans for action. Recognizing what issues need to be addressed, and gathering people together to discuss those issues and brainstorm on possible solutions, are the first steps when confronting any new challenge or opportunity, and the IoT qualifies as both.

Ultimately, we see our workshop, this paper, and the continuing efforts of the National Consortium for Data Science as contributions to a multi-faceted effort to help American industry take full advantage of big data and the Internet of Things. By fully understanding the challenges and opportunities of this new digital frontier, we can plan strategically for a future in which the U.S. maintains and enhances its leadership role in the global economy.

To learn more about our efforts, please contact us at info@data2discovery.org.

ABOUT THE NCDS

The National Consortium for Data Science (NCDS, http://datascienceconsortium.org) is a collaboration of leaders in academia, industry, and government formed to address the data challenges and opportunities of the 21st century. The NCDS was founded as a mechanism to help the U.S. take advantage of the ever-increasing flow of digital data in ways that result in new jobs and industries, advances in healthcare, transformative discoveries in science, and competitive advantages for U.S. industry. The NCDS works to: engage broad communities of data experts; coordinate data science research priorities that span disciplines and industries; facilitate the development of education and training programs; and apply expertise to data challenges in science, business, and government.

Current NCDS members are: Cisco Systems, INC, Deloitte Consulting, Drexel University, Duke University, EMC, GE Research, IBM, MCNC, North Carolina State University, RENCI, RTI International, University of North Carolina at Chapel Hill, University of North Carolina at Charlotte, University of North Carolina at Greensboro, and University of North Carolina General Administration.
REFERENCES


