# Table of Contents

- Executive Summary .................................................. 3
- The Data Challenge for Healthcare .............................. 4
- On the Nature of Raw Data .......................................... 6
- Data Science Framework and Process ......................... 8
- Use Case 1: Public Health Influenza Data Sharing .......... 12
- Use Case 2: Analysis of Congestive Heart Failure Readmission Rates ................................................. 14
- Conclusion ............................................................... 16
Executive Summary

The cost of healthcare in the United States is growing at an unsustainable rate. In 2010, U.S. healthcare spending was $2.58 trillion, or 17% of U.S. gross domestic product (GDP) — a figure that on its own would be the fifth-largest economy in the world. By 2020, the percentage of GDP spent on healthcare is estimated to rise to nearly 20%.

Interoperable electronic health records (EHRs) for patient care hold tremendous potential to reduce the growth in costs. EHRs can help healthcare organizations improve chronic disease management, increase operating efficiencies, transform their finances, and improve patient outcomes.

However, EHR implementations are in various stages of maturity across the country, and their benefits have not been fully realized. One of the primary challenges healthcare decision-makers face is how to make meaningful use of the data collected, available, and accessible in EHRs.

By optimizing the use of data accessible in EHRs, we can uncover hidden relationships and identify patterns and trends in this diverse and complex information to improve chronic disease management, increase operating efficiencies, and transform healthcare organizations’ finances.

In this white paper, Leidos (formerly SAIC-maxIT-Vitalize) focuses on emerging capabilities in data science that can identify useful, actionable — and perhaps unanticipated — recommendations to improve healthcare delivery from large volumes of diverse and complex data. We present use case studies to examine some of these capabilities in action.
The Data Challenge for Healthcare

The cost of healthcare in the United States is growing at an unsustainable rate. It is already the most costly in the world, accounting for 17% of U.S. gross domestic product in 2010, with estimates predicting that percentage will grow to nearly 20% by 2020.\(^1\)

Considered on its own, the $2.58 trillion spent annually would be the fifth-largest economy in the world.

Interoperable electronic health records (EHRs) — comprising structured lab records with clinical test results as well as unstructured records of clinicians’ notes — are only beginning to play their expected role in reducing this growth rate. As more hospitals reach Stage 6 and Stage 7 of the HIMSS EMR Adoption Model, substantial benefits are possible,\(^2\) including healthcare delivery improvements in several core areas: optimizing resource productivity, including clinical and administrative labor and equipment scheduling; reducing the number of adverse drug events; and managing the revenue cycle.\(^3\)

EHRs hold the potential to improve chronic disease management, increase operating efficiencies, and transform healthcare organizations’ finances through a combination of enhanced data management, knowledge development, knowledge use, and practice change.

However, these benefits require that healthcare leaders be able to take action and make meaningful use\(^4\) of the data collected, available, and accessible in EHRs.

---

From Raw Data to Decision-Making

The variety and complexity of EHR data creates two fundamental challenges: collecting, storing, and providing access to EHR data, as well as exploring, inspecting, and consuming those data to uncover hidden relationships, identify trends and patterns, and improve patient outcomes.

The first challenge is far from solved conclusively, especially as it needs to address legal and regulatory requirements that themselves may change. However, much progress has been made in developing and deploying robust EHR systems in real-world healthcare practice across the United States and throughout the world.

This white paper focuses on the second challenge — transforming the raw data accessible in EHRs into improved understanding of patients’ health, healthcare practices, and opportunities to enhance both. As Marchand and Peppard note, simply deploying an IT data system does not inherently indicate “how to use the information it generates to make better decisions or gain deeper — and perhaps unanticipated — insights into key aspects of the business.”

It is the exploration of these data, and subsequent explanation of the relevant exploratory findings, that is essential to effective decision-making and subsequent implementation of an action plan.

We explore emerging capabilities in data science that can extract useful and actionable information from large volumes of diverse and complex data. We also use case studies to examine some of these capabilities in action.

What Is Data Science?

The term “data science” has been used in business for nearly a decade, encompassing the principally statistical analysis of historical, structured data. More recently, business intelligence has expanded to include data collected from social websites, such as determining purchase trends. However, it is only in the last few years that the term has been extended to define a broader field of study.

Leidos defines “data science” as the discipline of deriving relevant, actionable knowledge from raw data. These data are often collected for other purposes, such as facilitating business operations or complying with laws and regulations. This challenge has become more pervasive and problematic with the advances of technologies that collect, store, and manage very large, complex datasets.

Data science is distinct from the general concept of “big data” in that it recognizes the increasingly tight coupling among three skill areas: (1) strong domain expertise and knowledge of the available data resources; (2) analytical skills and techniques, often automated and based on mathematics, statistics, and machine learning; and (3) computer science and programming skills. All three capabilities are required to work with large datasets that are not amenable to traditional processing by database management systems.

Data science speaks to the process of extracting relevant insight from data through the use of technologies and tools, not simply the technologies and tools themselves.

6. The late Jim Gray coined a new term for the development of scientific knowledge, “the Fourth Paradigm” of scientific research, the seminal paper leading to http://oreilly.com/data/radarreports/what-is-data-science.csp
On the Nature of Raw Data

Opportunities to derive value from data depend to some extent on the nature of the raw data themselves. A dataset and the individual records within it can be characterized in several ways, including volume, velocity, and variety — the “3Vs” model commonly used to describe big data.

Type

The data in a record or dataset may be text strings or numeric values of various kinds, such as real or imaginary numbers, expressed separately or in arrays or matrices. Data arrays may be indexed or time-sequenced to form ordered data such as image, video, and audio records.

Structure

Any record or data element in a record may be unrelated to others (unstructured data), related in a fixed and formally defined way such as a hierarchy or relational linkage (structured data), or related in ways specific to the individual record or dataset (semi-structured data).

Volume

Volume refers both to the number of records in a dataset and the size of an individual record.

Roughly speaking, small datasets are those that can be managed, accessed, and analyzed with traditional file-based and commodity applications. Medium datasets require well-understood technologies, such as relational database management systems. Large datasets need specialized data management tools for analysis, storage and maintenance that must coordinate across processors and storage platforms.

Note that the actual sizes of “small”, “medium”, and “large” datasets continue to change as data storage and retrieval technology advances.8

Velocity

Velocity refers to how often new data arrive for collection, storage, and analysis, as well as how much data arrive at any one instant. High-velocity applications such as retail purchase transactions may require specialized tools to adequately handle incoming floods of data.

Opportunities to derive value from raw data depend on the nature of the data.

Variety

Variety describes both the number of datasets and the number of data types employed in analytics. Aggregating, aligning, linking, and correlating information across highly diverse datasets and data types is an ongoing challenge in data science for all applications, including emerging technologies to address the alignment of patient data across multiple EHRs.

---

8. For example, decoding the human genome originally took 10 years to process; now it can be achieved in one week, which depends on new technologies with improved volume storage and retrieval velocity performance.
Completeness, Richness, and Balance

These factors indicate how fully the dataset represents a desired population, activity, or other variable of interest. Completeness is the fraction of a given population represented within the dataset. Richness refers to the amount of data available for each entity (e.g., an individual or organization). Balance is a measure of how prevalent a given phenomenon is (or is estimated to be) in the dataset; for example, in a healthcare analytics context, a dataset including all in-hospital patients would have a different readmission rate than a dataset of patients with chronic diseases, since comorbidities are presumably more prevalent in the latter dataset.

Labeling

Labeling indicates to what extent objects, events, and other phenomena of interest are explicitly marked, e.g., an X-ray image with notes attached from a radiologist indicating whether a tumor was malignant or benign.

Healthcare Data Features and Implications

The data collected in healthcare operations exhibits tremendous variety and complexity, but fairly low velocity compared to extremely high-velocity applications such as credit-card processing. Similarly, although adoption of EHRs generates an unprecedented volume of data in healthcare delivery, the volume can be accommodated by technologies developed for much larger datasets managed in transactional systems, such as airline passenger data repositories.

Because healthcare data are so diverse and information-rich, the innovative, yet relevant, use of analytics must be a key aspect of decision-makers’ approach to understanding and intelligently using their data. Simple and common data queries and matching of explicitly labeled features (typical first-order approaches) are prohibitively time and resource-consuming, and quite often miss a large percentage of the useful intelligence that lies in healthcare data, waiting to be revealed and acted upon.
Data Science Framework and Process

Leidos has developed, implemented, and refined a framework for the practice of data science through extensive experience applying data science techniques to real-world problems in healthcare, national security, intelligence, commerce, and other fields. Success in these distinct domains relies on a framework of practice that is distinct from the general concept of “big data” in that it recognizes the tight coupling among three skill areas:

- Strong domain expertise and knowledge of the available data resources
- Analytical skills and techniques, based on mathematics, statistics, and machine learning
- Computer science and programming skills.

This framework comprises four fundamental stages, each of which is underpinned with a firm understanding of the business or clinical problem and healthcare delivery environment, though they may not proceed entirely linearly. As depicted in figure 1, the data science framework of Leidos is founded in resolving a business problem identified by decision-makers and results in actionable insight tied to this problem.

It is not uncommon to proceed through each stage to establish a basic data science workflow, then revisit earlier stages to accommodate shifts in the problem definition or environment resources, new data sources, additional analyses, and explore new information.

Data Collection

Data collection encompasses the activities from the moment raw data are gathered from their originating sources to the time of their placement (possibly in a transformed state) in repositories appropriate for storage, retrieval, and analysis. It also includes the planning and systems required for ongoing management and security of the collected data.

Data Curation

Data curation, where data are prepared for meaningful analysis, is perhaps the most important step in determining whether a data science project has the potential to produce measurable business value.

This stage includes, but is much broader than, the traditional extraction, transformation, and load (ETL) methods of both commodity and custom-developed big data management systems. Specifically, the first phase of data curation involves standardizing the raw data to uniform specifications. This typically involves data cleansing, in which potential errors (e.g., data corrupted in collection, data values outside meaningful or permissible ranges) are repaired.

The data science process should result in actionable insight for business problems.
and/or removed from the dataset, as well as reformatting and conversion of data values to standard units and formats. It may also include transformations such as conversion of audio files to text transcripts.

The second phase of data curation prepares standardized data for analysis. Data of different types and from different sources may be merged, aligned and linked, or aggregated into new records specific to an activity, process, or organization. Large datasets may be sampled to create smaller datasets that are more amenable to automated analysis and/or to correct statistical imbalances in the full dataset. Sampling may be inappropriate for some analysis activities, such as rare event or anomaly detection.

The dataset may be further enriched by the addition of new “features” generated from collected data. These features may be as simple as summary statistics (e.g., mean and standard deviation) computed from a numerical dataset, or as complex as a map generated from cellphone data of a criminal suspect’s locations leading up to the commission of a crime, or medication data of pharmacy locations that visualizes this one factor of patient engagement. Enrichment may also generate new representations of data, such as a network graph of clinical protocols and medications administered among patients created from clinician order entry records. Alternate data representations enable privacy protection, especially when an analytics project is not allowed the same access to data as the process that originated the data.

Computing these features is a key part of data curation. It relies on data scientists’ ability to explore and understand the data from both mathematical and business-problem-specific perspectives. They will use tools including SQL queries, off-the-shelf statistical packages, and multidimensional visualization applications to explore the data thoroughly, looking at both individual records and summary statistics and visualizations.

At the end of this stage, “counts” and “lists” of sets of prearranged record attributes can be gathered for direct communication with a data analyst or processed by more sophisticated analysis methods intended to generate findings that detect, identify, and explain previously unknown patterns.

Finally, data curation is where privacy protection is implemented, especially when an analytics project is not allowed the same access to data as the process that originated the data.

**Data Analysis**

This stage includes identification, selection, and execution of algorithmic and computational models and methods. It typically involves extension and novel integration of existing analytical approaches rather than the creation of new analytical methods.

Successful data analysis solutions do not take a “kitchen sink” approach of applying analysis methods indiscriminately and hoping for good outcomes. Instead, they are driven by the organization’s need for insightful and actionable intelligence and focus on producing relevant and accurate evidence informing such action.
As a general category, business intelligence analysis may use any of the many types of analysis to answer the question “What is going on in my business today?” It commonly focuses on obtaining answers to prescribed questions (e.g., “How is my administrative budget tracking against plan?”) and/or measuring known performance criteria (e.g., readmission rates), rather than searching for novel or unknown meaning in the data. The paragraphs below describe several common analysis methodologies.

- **Information Retrieval**
  Information retrieval uses direct manual or automated queries of a dataset to answer the question “What in this dataset is relevant to the query?” It typically relies on matching the query to explicitly represented features of the data (e.g., a patient’s name) without inference or prediction based on hidden features (e.g., an undiagnosed disease that could be inferred from the patient’s chart).

- **Classification**
  This method identifies to which of a prescribed set of classes a newly observed data point belongs. The classes may be explicitly defined by users, or may be learned by a classification algorithm based on other features of the data.

- **Clustering**
  Clustering is similar to classification, but labels data points based on their similarity (in one or more features) with each other. Data points are then grouped into likely classes based on metrics that may be deterministic (e.g., hierarchical and k-means), statistical (e.g., multivariate distributions and likelihood maximization), or network graph-based (e.g., centrality and between-ness).

- **Pattern Recognition and Matching**
  Pattern matching seeks to identify trends, sequences, and patterns in datasets. The patterns may be explicitly defined in advance, learned directly from data, and/or manually refined after initial learning from data.

- **Anomaly Detection**
  Anomaly detection is the reverse of pattern matching — looking for outliers from a known pattern, rather than the pattern itself. Anomalies may be structural, in which certain entities’ behavior is significantly different from that of peer entities (e.g., finding patient zero in a disease outbreak or fraudulent insurance claim filings), or temporal, when entities’ behavior changes significantly over time (e.g., the methods used by health payer companies to flag potentially fraudulent claims).

- **Predictive Modeling**
  Predictive modeling not only attempts to forecast future events, it also seeks to identify ways to optimize future outcomes (e.g., comparing a patient outcome forecast with and without a given medical intervention). The underlying characteristics of the original data determine the types of meaningful predictions and optimizations that are possible to generate.

- **Multirelational Network Analysis**
  Multirelational networks explore the connections and relationships among data points and entities represented in the data. Network analysis typically focuses on social networks, which analyze links among groups of people (e.g., provider referrals, patient satisfaction), and semantic networks, in which connections occur between types of knowledge (e.g., relationships between medical literature and the concepts and conclusions contained in clinicians’ notes).
Explanation via Interpretation and Communication of Analytic Findings

The end game of a decision support solution based on data science practices is to provide decision-makers with the intelligence they need to take action. This can happen only if leaders have timely access to clear, accurate, and relevant data analytic conclusions in a form that aligns with their missions, needs, and expectations. Communicating analytic conclusions may employ any of the following strategies. An example would be application of forensic analysis, for example longitudinal analysis of a drug’s efficacy in clinical trials, to understand and explain how the drug might affect readmission rates, with the goal of informing a hospital’s choice to incorporate it into their standard of care.

- Data Summarization
  Summarization distills key analytic findings relevant to a project’s objective and aligned with the consumer’s role and experience into a simpler and more concise form than a full report — essentially an executive summary. Analytic summaries should focus on the key aspects of analytic findings while enabling examination of the underlying evidence and analysis logic.

- Synthesis of Analytic Findings
  Synthesis unites seemingly separate findings, whether complementary or conflicting, to explain their combined effect on the decision-making process. Given the diversity of data types and levels of accuracy, timeliness, and relevance, synthesis can be the most difficult to accomplish effectively. However, it is crucial to leaders’ understanding, confidence, and ability to select appropriate actions, including a decision that the data are insufficient to warrant a change to current operations.

- Narrative Creation
  Narratives place analytic findings in a larger context relevant to the target audience(s), who may have roles, experience, and perspectives different from data scientists, or even leaders of the organization performing data science efforts. An example would be an explanation of how decreasing readmission rates for a particular medical condition would likely affect (positively or negatively) not just the hospital where care is given but also the surrounding community.

- Forensic Analysis
  Forensic analysis examines historical data to develop analytic findings that may provide insights on future plans over multiple time scales. Longitudinal analysis of a drug’s efficacy in clinical trials, with the goal of informing a hospital’s choice to incorporate it into their standard of care, is an example of forensic analysis.

- Visualization
  Visualization creates image-based static or dynamic representations of data and analytic findings (e.g., line charts, histograms, maps, network diagrams) to complement direct examination of the data themselves. These visual representations may also have interactive elements that enable viewers to select data subsets, observe data patterns over time, and otherwise perform basic analysis. Visualization requires care in selecting data types, as well as the cultural tolerance and acceptance of visuals as actionable evidence, but can be a powerful method of communicating findings and recommendations.
USE CASE 1:  
Public Health Influenza Data Sharing

Homeland Security Presidential Directive 21° states the necessity of establishing a national biosurveillance system for human health that integrates state, regional, and community-level capabilities. To better support early detection and rapid response to major infectious disease outbreaks or pandemics, the Centers for Disease Control (CDC) wanted to expand the real-time exchange of clinical health information among health information exchanges (HIEs) and local, state, and federal public health organizations.

CDC awarded Leidos a contract to design and develop an integrated HIE “to connect public health with HIEs to improve public health’s real-time understanding of communities’ population health and healthcare facility status.”

Leidos led a team of bioinformaticists, systems engineers, HIE operators, and subject matter experts in state and local public health requirements to assemble the Northwest Public Health Information Exchange (NW-PHIE) to help achieve this goal.

A key challenge was establishing a real-time, multidirectional information flow across federal, state, and local jurisdictions. Until this project, biosurveillance platforms such as RODS, ESSENCE, and BioSense collected de-identified biosurveillance information, but did not share or visualize these data across platforms. No syndromic surveillance data feeds were being sent to public health from hospitals in central and eastern Washington State.

The starting points for NW-PHIE’s syndromic surveillance requirements were the American Health Informatics Community (AHIC) Biosurveillance Use Case and the MBDS. NW-PHIE augmented the AHIC requirements with additional requirements from local and state public health agencies.

The Leidos team developed a data format specification that described how the highly diverse clinical data would be transmitted to public health. We followed best practices and industry standards, such as those from HL7.

Leidos next mapped the data fields in the HL7 messages to the clinical care information provided by the various data sources to meet the data output specifications. Through these alignments, information for a single patient is split into many discrete biosurveillance HL7 messages triggered by events at a healthcare facility, such as registration, lab order, lab result, etc.

To make clinical data useful for population health purposes, several analyses were needed. First, individual messages are reassembled into a longitudinal, comprehensive view of a visit encounter. Next, the encounters are classified according to surveillance criteria, counted over a fixed time interval, and paired with an appropriate denominator, such as total visit volume or catchment population. From these data, absolute counts and rates of illness within the patient population are obtained.

11. Real-time Outbreak and Disease Surveillance Laboratories (RODS), https://www.rods.pitt.edu/site/
To assist public health departments in evaluating the clinical data, we developed tools and algorithms to help visualize data and support the development of appropriate public health responses. The resulting analysis is fed back to clinicians via health alerts, investigation findings, and case criteria updates.

Now, thanks to NW-PHIE, 16 hospitals (representing close to a million patient encounters a year) are sending a full set of syndromic surveillance data to public health. These feeds proved to be vitally important in helping manage the 2009 H1N1 outbreak and are available to help manage future disease outbreaks.
USE CASE 2:
Analysis of Congestive Heart Failure Readmission Rates

Patients with a diagnosis of congestive heart failure (CHF) pose a challenge to health delivery systems. According to the Centers for Medicaid & Medicare Services (CMS), the rate of readmission within 30 days of discharge for CHF patients ranges from 18.2% to 33.8%\(^{14}\), compared to the average readmission rate of approximately 24.8% for all patients in the United States.

CMS views these variations as an opportunity to decrease the rate of CHF readmission. Beginning in October 2012, CMS has penalized payments to health systems that exceed the average rates of readmission. Other diagnosis categories, such as hypertension, also trigger penalties from CMS when readmission rates exceed prescribed norms.

To address this business challenge, Leidos conducted a research and development effort to study whether our information-based health approach to clinical challenges could be used to create a model that could calculate an accurate CHF readmission probability based on a broad set of clinical and non-clinical factors.

A multidisciplinary team including experts in commercial and federal healthcare, health informaticists, physicians, scientists, statisticians, and mathematicians conducted initial research into causes of CHF readmission. They identified three types of factors that may affect readmission of CHF patients: patient-specific factors (clinical and demographic), hospital-specific factors (practices and continuity), and socioeconomic factors (support and access).

Our team used these SME findings to develop our pilot prediction model for CHF readmission rates, in the form of a Bayesian inference network (INET) model with multiple nodes and relationships (links). Each node depicted on a network graph represented a grouping of data that affects CHF readmission (directly or indirectly), such as whether a patient has extensive comorbid conditions or a social support system after discharge. The model, which includes multi-path relationships (chains of inference), then generates predictive values by statistically weighting and combining the complementary or competing chains of inference.

Note that this use case focuses on analysis; requirements for data collection and curation would need to be considered based on a specific implementation of this proof-of-concept experiment.

However, in preparation for a decision on technology investment into data collection and curation, the team also investigated whether electronic patient data were available to populate the model’s parameters directly. Figure 2 illustrates the availability of EHR data in our statistical analysis of more than 5,000 CHF patient encounters in a 10-hospital system between 2010 and 2011.

---

Having constructed the network and populated it with data, the team investigated a set of outcomes for three use cases identified by the team's clinical SMEs, using historical “hard” CHF data and subjective, but expert, judgments.

The INET model’s findings — including outcome likelihood (chance the patient would readmit) and “top N” indicators that, if addressed correctly, would reduce the readmission — were presented to a focus group consisting of physicians, executive leaders, informaticists, and statistical analysts from a 10-hospital university health system and a large, inner-city safety net hospital.

The focus group members reviewed the outcomes against their own CHF readmission experience and concluded that the INET model would be a useful decision-support tool that provides insight for hospital clinical and administrative staff members responsible for identifying CHF patients at risk prior to discharge.
Conclusion

Healthcare organizations are transitioning to EHRs en masse for the clear benefits they offer in improving collaboration and information sharing, reducing inefficiency and administrative effort, and ultimately improving patient care. Implementation progress varies widely, but many organizations are already realizing some level of benefit simply by using EHRs instead of paper records.

Data science is a deep, flexible, customizable next level of opportunity for organizations using EHRs. Healthcare decision-makers can choose the depth, breadth, and complexity of the data sharing, analytics, and explanation tools they deploy based on their missions, budgets, and patients’ needs.

As healthcare organizations grow, adapt, and serve changing populations, the data science mix of methods can be expanded or tailored (including scaled-out methods to meet “big data” needs) to keep delivering benefits. At every point, data science can be harnessed in different ways to derive useful, actionable information from the complexity of data in EHRs.
ABOUT LEIDOS

Leidos is a FORTUNE 500® science and technology solutions leader working to address some of the world’s toughest challenges in national security, health and engineering. The Company’s 20,000 employees support vital missions for our government and the commercial sector, develop innovative solutions to drive better outcomes and defend our Nation’s digital and physical infrastructure from ‘new world’ threats. Headquartered in Reston, Virginia, Leidos reported annual revenues of approximately $5.77 billion for its fiscal year ended January 31, 2014 after giving effect to the spin-off of the company’s technical services and information technology business. For more information, visit www.Leidos.com.

FOR MORE INFORMATION

Julie A. Rosen, Ph.D.
Chief Scientist, Leidos Health
Julie.A.Rosen@leidos.com

Visit us at www.leidos.com/health