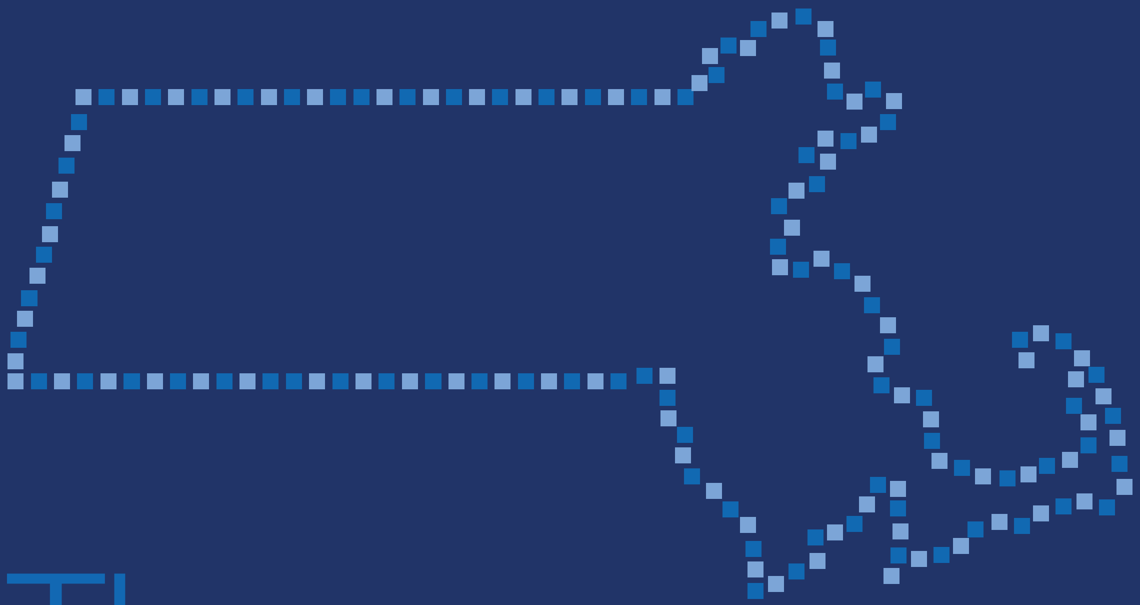
2014



The Massachusetts Big Data Report

A Foundation For Global Leadership

MASSACHUSETTS TECHNOLOGY COLLABORATIVE

massbigdata

MACP
MASSACHUSETTS COMPETITIVE PARTNERSHIP

MassTech: Who We Are

The Massachusetts Technology Collaborative, or MassTech, is an innovative public economic development agency which works to support a vibrant, growing economy across Massachusetts. Through our three major divisions - the **Innovation Institute**, **Massachusetts eHealth Institute** and the **Massachusetts Broadband Institute** - MassTech is fostering innovation and helping shape a vibrant economy.

We develop meaningful collaborations across industry, academia and government which serve as powerful catalysts, helping turn good ideas into economic opportunity. We accomplish this in three key ways, by:

FOSTERING the growth of dynamic, innovative businesses and industry clusters in the Commonwealth, by accelerating the creation and expansion of firms in technology-growth sectors;

ACCELERATING the use and adoption of technology, by ensuring connectivity statewide and by promoting competitiveness; and

HARNESSING the value of ***effective insight*** by supporting and funding impactful research initiatives.

MassTech: Our Mission

Our mission is to strengthen the innovation economy in Massachusetts, for the purpose of generating more high-paying jobs, higher productivity, greater economic growth and improved social welfare.

The Innovation Institute at MassTech

The Innovation Institute at MassTech was created in 2003 to improve conditions for growth in the innovation economy by:

- Enhancing industry competitiveness;
- Promoting conditions which enable growth; and
- Providing data and analysis to stakeholders in the Massachusetts innovation economy that promotes understanding and informs policy development.

The Innovation Institute convenes with and invests in academic, research, business, government and civic organizations which share the vision of enhancing the Commonwealth's innovation economy.

Using an innovative, stakeholder-led process, we have been implementing a "cluster development" approach to economic development. Projects, initiatives and strategic investments in key industry clusters throughout all regions of the Commonwealth are creating conditions for continued economic growth.

The Institute manages programs which focus on Advanced Manufacturing in the state, driving support for emerging sectors such as Big Data and Robotics and spurring programs which keep talented workers in the Commonwealth, whether through the Intern Partnership program or on entrepreneurship mentoring. Our mission is to strengthen the innovation economy in Massachusetts, for the purpose of generating more high-paying jobs, higher productivity, greater economic growth and improved social welfare.

MASSACHUSETTS TECHNOLOGY COLLABORATIVE

MACP
MASSACHUSETTS COMPETITIVE PARTNERSHIP

Dear Friends,

It is our pleasure to present the 2014 Mass Big Data Report: A Foundation for Global Leadership. Assembled and released with support from the Innovation Institute at the Massachusetts Technology Collaborative and the Massachusetts Competitive Partnership, this report represents a foundational analysis of the regional Mass Big Data ecosystem and its position as a global leader in the expanding fields of big data, open data, and analytics. As a special project of the Governor's Mass Big Data Initiative, this report seeks to provide an initial baseline understanding of the landscape of the Mass Big Data ecosystem and its challenges, opportunities, and strong potential for growth.

Through this work, we are pleased to report that the Mass Big Data ecosystem represents an extraordinarily fertile region for growth in data-driven enterprise and offers a unique combination of advantages on which to build the future of our data-rich world. With strengths across the spectrum of big data industry sectors and in key supporting areas such as talent development, research, and innovation, our region is producing the people, businesses, and products that fuel the explosive growth in this expanding field.

Through the Mass Big Data Initiative, the Commonwealth works in partnership with industry, academia, and the region's vibrant data-centric community to identify and address the exciting and transformative growth opportunities emerging from our expanding Mass Big Data ecosystem. Led by the Innovation Institute at the Massachusetts Technology Collaborative, the Initiative seeks to enhance growth conditions for big data in Massachusetts and to drive increased social benefits at the intersection of talent, advanced analysis, innovative technology, and public engagement with regional open data.

We invite you to read the report, share it widely and consider how you can participate in and benefit from the expansion of the Mass Big Data ecosystem in the Commonwealth.

Sincerely,

Pamela Goldberg
CEO
MassTech

Dan O'Connell
President & CEO
Massachusetts Competitive
Partnership

Pat Larkin
Director
Innovation Institute
at MassTech

Table of Contents

Executive Summary 4

I. Introduction 10

II. The Massachusetts Big Data Ecosystem 12

III. Competitive Position of Massachusetts 29

IV. Growth Prospects 40

V. Recommendations for Action 43

Appendix A. What Is Big Data? 47

Appendix B. List of Interviews 54

Appendix C. Profile of Survey Respondents 55

Appendix D. List of Keywords and Counts 56

Appendix E. Talent Pipeline & Higher Education Data 58

Appendix F. Referenced Tables 66

Acknowledgments 86

Executive Summary

On May 30, 2012, Governor Deval Patrick launched the Massachusetts Big Data Initiative, to leverage and expand the Commonwealth's position as a global leader in the rapidly growing big data sector. The Initiative, led by the Innovation Institute at the Massachusetts Technology Collaborative, has launched several pilot efforts to enhance and grow the region's vibrant and expanding Mass Big Data ecosystem, including strategic and collaborative partnership efforts with academia, industry and public sector organizations.

The purpose of the **2014 Mass Big Data Report** is to provide an assessment of the relative strengths and weaknesses of the Commonwealth in big data. The Mass Big Data Report is intended to highlight prospects for growth in areas such as talent and workforce, ecosystem, and public data access; and to identify opportunities to promote and expand the Mass Big Data sector, while enhancing the Commonwealth's position as a global leader. The 2014 Mass Big Data Report is intended to serve as a baseline assessment of the Massachusetts Big Data ecosystem and related economic factors. Subsequent updates to the report will track changes, trends, and metrics based on this foundational data.

Conducted by Nexus Associates and staff from the Innovation Institute at the Massachusetts Technology Collaborative, the study is based on a broad spectrum of sources, including interviews with 16 key industry stakeholders; the results of the first annual Mass Big Data Survey of over 60 Massachusetts big data companies; an analysis of publicly available federal, state, and university data; input from social media sources, including LinkedIn; and an extensive literature review.

Principal Findings

Close to 500 Companies Participate in the Massachusetts Big Data Ecosystem

The companies that make up the Mass Big Data ecosystem range from small start-ups with a handful of employees to large, well-established firms such as EMC, IBM, Akamai and Oracle. Mass Big Data companies are leaders in a wide variety of markets, including big data-enabled applications, data analysis tools, data management software, storage and other hardware, cloud services, and other supporting services. Many companies target a broad range of industries ("industry verticals"), including healthcare, life sciences, financial services, manufacturing, transportation, energy and utilities, telecommunications, e-commerce and retail trade, entertainment and media, social media, and marketing and advertising. There has been considerable acquisition activity among Mass Big Data companies in recent years as larger organizations seek to gain access to new technology or market share.

Research Centers Across the Commonwealth Differentiate the Mass Big Data Ecosystem

Massachusetts has a significant base of organizations with an interest in using big data to improve operations, to develop products, solutions, and services, and to inform decisions. Ten leading university and hospital-affiliated research centers across the Commonwealth provide an important foundation for advances in big data.

These centers are developing new technology platforms and analytical techniques, as well as using big data to address important research questions in healthcare, life sciences, communications, cyber security, transportation, energy, and other fields.

Nearly $20 Million in Federal Grants Awarded for Big Data Initiatives in Massachusetts
From 2006 to 2013, Massachusetts organizations received close to $20 million from the National Science Foundation, the National Institutes of Health and other federal agencies in support of research and educational activities related to big data. The federal Big Data Initiative has committed $200 million in funding nationwide for 2012-2017.

Investment Funding in Mass Big Data Companies Topped $2.5 Billion
More than 240 angel investor groups, venture capital firms, private equity firms, and strategic investors have invested more than $2.5 billion in at least 123 Massachusetts-based big data related companies since 2000. In the three largest investments, Hubspot, Jumptap, and Attivio have received $130.5 million, $101.5 million and 90.1 million, respectively.

Massachusetts Colleges and Universities Graduate Close to 5,600 Students Annually from 14 Data Science-related Programs
The Mass Big Data talent pipeline is robust and prepared to address the skills necessary in building the Mass Big Data ecosystem. Massachusetts offers a wide range of formal and informal educational opportunities for those interested in developing the skills identified as central to careers in big data. Massachusetts' colleges and universities graduate close to 5,600 students annually from 14 undergraduate and graduate data science-related programs, offering degrees in computer science and engineering, mathematics, statistics, physics, computational biology and other relevant fields. Hackathons, workshops, meet-ups and other industry-sponsored training are held on a regular basis on campuses across the Commonwealth. While most firms report these programs are generally well-aligned to the required skills, companies looking to fill positions report difficulty in recruiting sufficient numbers of qualified software engineers, data architects/engineers, and data scientists.

Massachusetts' Big Data Talent Density Among Highest in US
Massachusetts is a clear leader in per capita graduates from data science related programs as compared to other leading states, with a higher concentration of graduates in certain key degree programs, including biomathematics, bioinformatics, and computational biology.

Strength in Innovation: Data Integration Tools, Data Analysis Software, Data Management
Over two-thirds of the 485 companies researched develop big data applications for vertical industry markets, such as healthcare, life sciences and financial services. Nearly a third of the big data companies researched are in the data analysis software business.

5,250 Big Data Patents Granted in Massachusetts
Analysis of patent data provides insight into the technological strengths of organizations in Massachusetts. A total of 5,250 patents were granted to inventors in Massachusetts between 2008 and 2012 in 23 technology classes that relate to the processing and use of data.

Executive Summary

Overall Prospects for Growth:

Global Market for Big Data to Top $48 Billion

The overall global big data market is expected to top $48 billion by 2017, up from $11.6 billion in 2012. While hardware and services are expected to continue to account for the lion share of revenue, the fastest growth is likely to be in big data-enabled applications.

Big Data Applications in Healthcare, Life Sciences and Financial Services Most Promising

The vast majority of respondents view applications in healthcare, life sciences and financial services as "very promising" or "extremely promising" in terms of their prospects for substantial growth in Massachusetts. Study respondents ranked "data integration tools" as the highest growth area within big data technology and market advancements, followed by data management and data analysis software.

Significant Demand for Mass Big Data Jobs Predicted Over Next 12 Months

Over 50 local big data related firms in the 2013 Mass Big Data Survey reported that they are seeking to fill almost 400 big data-related jobs in Massachusetts over the next 12 months. Considering this figure is drawn from just under 10% of the firms in the Mass Big Data regional ecosystem with possible job openings, the extrapolated figure for the region as a whole could be as high as 3,000-4,000 jobs, before any projection adjustments for additional sector growth.

Massachusetts Big Data

INDUSTRY

485 companies

across the Mass Big Data Ecosystem

$2.5 billion invested

in 123 Mass Big Data-related companies since 2000

80 NEW

Big Data-related companies launched since 2010

INNOVATION

$20million in federal grants

awarded for big data research in Massachusetts since 2006

Since 2007 28 regional big data meet-up groups held

368 meet-ups

Mass. inventors granted

5,250 patents

in 23 Big Data technology classifications since 2008

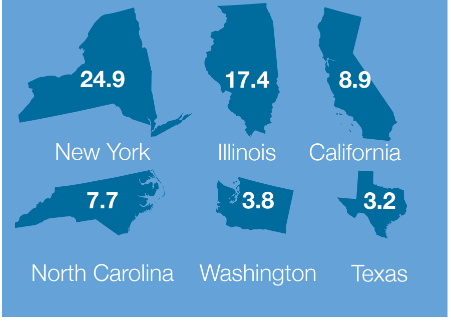
TALENT

5,600 students

from Mass. colleges and universities graduate from **14 data science-related programs** annually

79.5

Massachusetts has the **highest per capita** Big Data- related graduation concentration among leading states (per 100,000 population)



Source: 2013 Mass Big Data Study

Six Key Mass Big Data Priorities Identified for Action

The Mass Big Data Study identifies six key areas around which the regional data science community is moving to organize and engage in order to realize the full potential of the Mass Big Data ecosystem. Based on industry input to the 2013 Mass Big Data Survey, these identified elements leverage the unique attributes of the Mass Big Data ecosystem and offer a roadmap to driving increased economic opportunity and public benefit.

1. Strengthening Opportunities for Data Science Education and Training

Over two-thirds of survey respondents identified a need for new and refreshed data science programs in Massachusetts, citing growing demand for a big data workforce with training in computer science and mathematics/statistics, as well as familiarity with specific industry verticals. Other respondents suggested that, in parallel with creating new degree programs, courses in computer science and mathematics/statistics should also be integrated into a broader range of other existing degree programs to support multiple paths to the mix of interdisciplinary skills sought by employers. Suggestions supported an emphasis on developing bachelors' degree programs at public universities and matching professional certification courses with industry needs to extend the training for workers already in the labor force.

2. Increasing Regional Talent Retention and Industry Recruiting Success

Respondents highlighted that in an industry driven by talent resources, securing talent is a top industry priority, especially in the current context of high global demand for skilled data professionals. With an existing world-class talent pipeline in the region, industry growth can be enhanced by improving access and engagement between recent and rising graduates and local firms. Collaborative projects, hackathons and internships were cited as critical to engaging and expanding a local community of practitioners.

3. Expanded Access to Public Data

Respondents identified significant value in the availability of state and local public data sets in formats readily accessible by researchers, application developers, and others to create practical applications targeted at specific issues related to the delivery of public services and the quality of life in the Commonwealth. Strong initial efforts in health records, transportation, and education data should be expanded, regularized, and supported with improved public access to the data. Additional efforts to engage the developer community around the use of this data, through meet-ups, hackathons, and other events, were cited as critical to strengthening the Mass Big Data ecosystem.

4. Increased Awareness

Respondents felt that the Commonwealth should strengthen promotional efforts to raise regional, national, and international awareness of the strengths, assets, and ongoing leadership of top performers in the Mass Big Data ecosystem. Successful efforts would support increases in the attraction and retention of individual talent as well as companies. A broad-based campaign would increase buzz about Mass Big Data through websites, social media, and other press to highlight the innovative uses of big data around the region, the important role played by data scientists in industry verticals, and the success of big data related entrepreneurs in Massachusetts.

5. Mass Big Data Ecosystem Expansion

According to study respondents, additional Mass Big Data initiatives should accelerate regional innovation and company growth by supporting novel collaboration to stimulate partnerships and opportunities to enhance the unique innovation environment in the Commonwealth. Increasing cross-sector collaboration among university researchers, enterprise system suppliers, and software developers will improve use of existing strong regional expertise and assets. Strengthening ties among the major computer science research centers and celebrating student-led innovation and competition should increase opportunities for collaborative development of new technologies and products. Supporting the formation of new partnerships between big data firms and top Massachusetts industry verticals enables companies to open new big data markets and exploit opportunities in particular industry verticals.

6. Federal Grants

Study participants consistently recommended that Massachusetts big data companies and academic departments should actively seek out and apply for federal grants where appropriate and collaborations should be explored as early in the process as possible. Greater awareness of federal grants allows researchers to more effectively put together competitive proposals.

Big Data Defined

“Big Data” describes a range of data, data types, and tools to address the rapidly increasing amount of data that organizations around the globe are handling.[1] The amount of data collected, stored and processed by this diverse spectrum of organizations has grown exponentially. This has been driven, in part, by an explosion in the amount of data sourced from web-based transactions, social media and sensors. IDC projects that the digital universe will reach 40 zettabytes (ZB) by 2020, an amount that exceeds previous forecasts by 5 ZBs, resulting in a 50-fold growth from the beginning of 2010.[2]

There are a variety of ways for organizations to use big data to create value. Data can be used to develop a better understanding of customers and to tailor products and services for narrowly defined segments. Organizations can use data to monitor performance of key functions, identifying factors contributing to observed variances and highlighting needed remedial actions or new ways to optimize systems. Some use data to predict behavior or forecast events, and as a result, take appropriate action. Data can assist in helping to meet regulatory compliance or legal discovery requirements. Finally, organizations can use data as the building blocks for new products and services found across all industries.

[1] http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf

[2] IDC, The 2011 Digital Universe study, “Extracting Value from Chaos”, sponsored by EMC Corp.

Introduction

Governor Deval Patrick announces the Mass Big Data Initiative at MIT.

Background

In May 2012, Governor Deval Patrick officially launched the **Massachusetts Big Data Initiative to strengthen Massachusetts' position as a world leader** in the rapidly expanding global big data sector.[3] The Initiative called for a range of efforts to support this goal, including the establishment of an organizing committee — the Massachusetts Big Data Consortium — to help identify, design and inform efforts in the Commonwealth to accomplish this objective. It also called for the creation of a research and development matching grant program that could support big data investments, the establishment of a new support for big data focused internships, and the sponsorship of hack/reduce — a not-for-profit organization that creates opportunities for big data related innovation and training. The **Massachusetts Technology Collaborative** was charged with responsibility for implementing these and other steps to advance big data in the Commonwealth.

Additional details on specific initiative elements are available in Section 5.

[3] http://www.mass.gov/governor/pressoffice/pressreleases/2012/2012530-governor-announces-big-data-initiative.html

Purpose of the Study

The purpose of this study is to help inform and support further action by providing a baseline understanding of the Mass Big Data landscape, an objective assessment of the relative strengths and weaknesses of the region, and suggestions around additional steps that may promote the growth of the Mass Big Data ecosystem and strength Commonwealth's position as a global leader.

Methodology

The study is based on a series of interviews with key stakeholders,[4] a survey of companies (the first annual Mass Big Data Survey), an analysis of publicly available data, and an extensive literature review. The survey was sent via email to 403 companies. Of these, 19 were not delivered successfully due to an invalid email address. The survey remained open for approximately two weeks; 57 companies completed the survey, representing an effective response rate of 14 percent.[5,6]

Organization of Report

The report is organized into five sections:

- **Section One:** **Introduction**
- **Section Two:** **The Massachusetts Big Data Ecosystem** – Analysis of the ecosystem in the Massachusetts, assessment of existing big data companies, research and educational institutions, risk capital financing, and various government programs.
- **Section Three:** **Competitive Position of Massachusetts** – Analysis of the competitive position of Massachusetts compared to other regions of the country.
- **Section Four:** **Growth Prospects** – Discusses the potential for growth of the overall global market as well as specific segments and verticals in Massachusetts.
- **Section Five:** **Recommendations for Action** – Offers recommendations and appendices that provide a definition of big data, an explanation of uses and value, and a framework for assessing the structure of the market in terms of key business segments.

[4] See Appendix A.

[5] Item non-response varied by question.

[6] A classification of respondents by business segment and targeted verticals is presented in Appendix C.

The Massachusetts Big Data Ecosystem

> "Thanks to the proliferation of highly interactive websites, social networks, online financial transactions, and sensor-equipped devices, we are awash in data. With the right tools, we can begin to make sense of the data and use it to solve any number of pressing societal problems – but our existing tools are outdated and rooted in computer systems and technologies developed in the 1970s."[7]
>
> **SAM MADDEN, Associate Professor, Department of Electrical Engineering and Computer Science, MIT and Faculty Director of the "BIGDATA@CSAIL" Initiative**

Some 485 companies comprise the Mass Big Data cluster.

Companies offer myriad products and services based on a diverse set of technologies.

There is no standard definition of big data nor a standard for classifying big data businesses for statistical purposes such as the North American Industry Classification System (NAICS) used by government statistical agencies. As such, the process for identifying big data companies used in this report relies on a comprehensive keyword search of Crunchbase and LinkedIn profiles based on terms associated with different components of the technology platform, analytical techniques, and major uses.[8,9] The search revealed 485 companies in Massachusetts.

With the list in hand, an effort was made to identify market segments and verticals targeted by each company. Many companies targeted multiple areas. As shown in *Table 1*, while the cluster is diverse, two-thirds of 485 companies are involved in developing applications. Many of these companies, as well as those involved in other market segments, are focused on marketing and advertising, healthcare, and life sciences.

While businesses are active in a wide range of activities, the Commonwealth has a comparative advantage in key areas, including data integration, data management (specifically data warehouses and NoSQL/NewSQL databases), machine learning/predictive analysis, network analysis, semantic analysis, social media analysis, data visualization, and bioinformatics/genomics.

[7] http://www.csail.mit.edu/node/1750

[8] A list of the keywords used in the search is included in Appendix D.

[9] This was supplemented by a previous list of firms compiled by MTLC.

Mass Big Data Industry

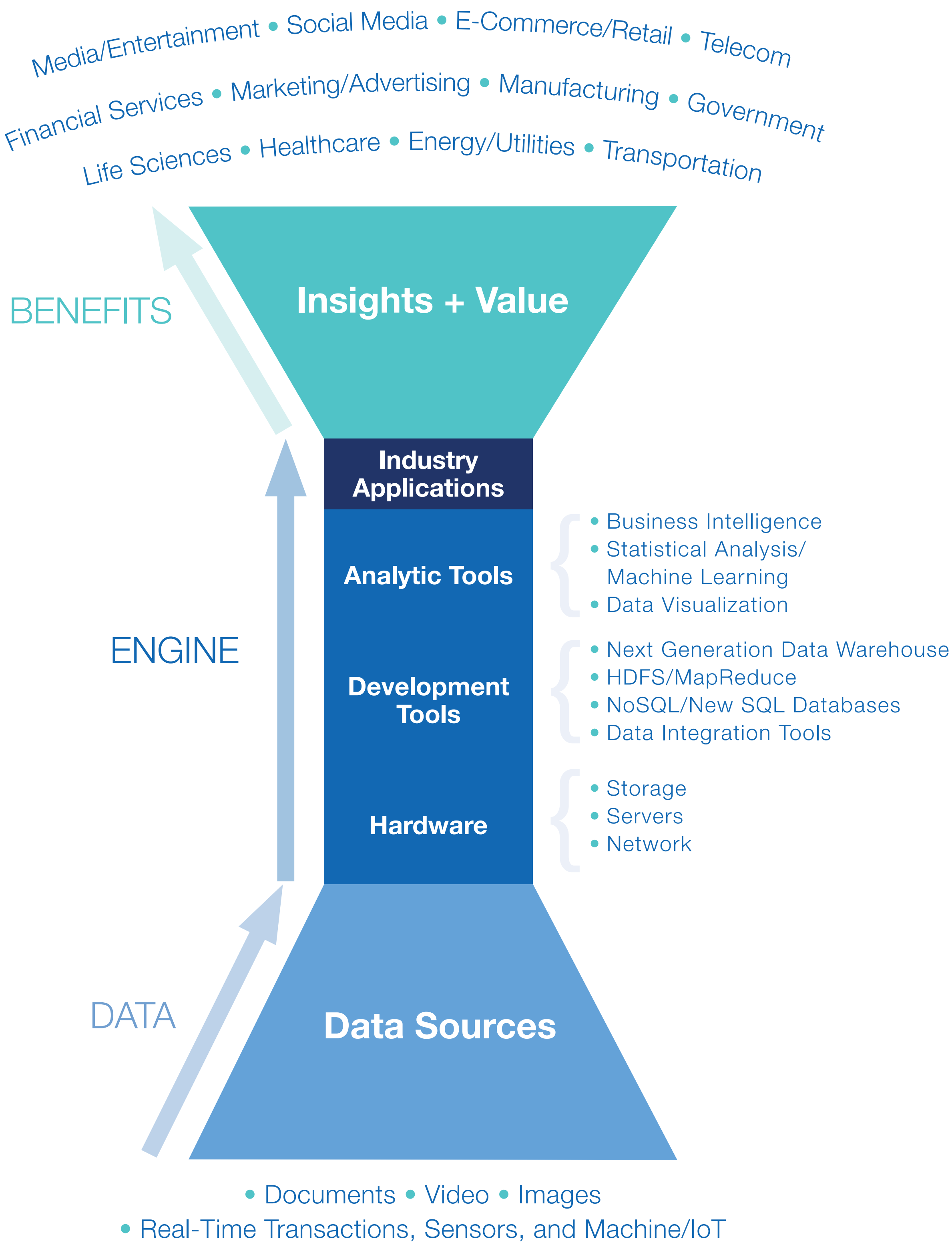

Media/Entertainment • Social Media • E-Commerce/Retail • Telecom
Financial Services • Marketing/Advertising • Manufacturing • Government
Life Sciences • Healthcare • Energy/Utilities • Transportation
BENEFITS
Insights + Value
Industry Applications
ENGINE
Analytic Tools
• Business Intelligence
• Statistical Analysis/ Machine Learning
• Data Visualization
Development Tools
• Next Generation Data Warehouse
• HDFS/MapReduce
• NoSQL/New SQL Databases
• Data Integration Tools
Hardware
• Storage
• Servers
• Network
DATA
Data Sources
• Documents • Video • Images
• Real-Time Transactions, Sensors, and Machine/IoT

TABLE 1) **Breakdown by Segment and Targeted Verticals**
(note: many firms target more than one segment or vertical)

Segment	N	Percent of 485
Data analysis software	160	33%
Data management software (incl. RDBMS, Hadoop, NoSQL, and NewSQL-based products).	101	21%
Consulting	78	16%
Business intelligence software	67	14%
Data integration software (software to ingest, extract, & transform data from multiple sources)	54	11%
Data visualization software	17	4%
Hardware, including computers, servers, storage and networking equipment	10	2%
Systems integration	10	2%
Applications geared to specific verticals such as e-commerce, financial services, healthcare, etc.	324	67%
Targeted Vertical	**N**	**Percent of 485**
Marketing and advertising	61	13%
Healthcare	60	12%
Life sciences	34	7%
Financial services	23	5%
Energy	11	2%
Social media	10	2%
Entertainment	6	1%
Homeland Security/Defense	5	1%
Education	4	1%
Telecommunications	4	1%
Transportation	3	1%
E-Commerce	2	0%
Manufacturing	1	0%

Companies range from small start-ups with a handful of employees to large, well established firms. Some firms such as EMC, Mathworks, and Akamai are long-established companies with deep roots in Massachusetts. These companies employ thousands of people in the Commonwealth. At the same time, there is also a great deal of vibrancy in this cluster with new firms being created at a rapid pace. Nearly 80 new big data companies were established in the past three years in Massachusetts *(See Appendix F, Table 31, "Age of Companies").*

In addition to a strong flow of homegrown companies, recent years have seen companies with headquarters outside of Massachusetts establishing new operations in the Commonwealth and/or acquiring local firms. These companies include major industry players such as IBM, Oracle, Google, and Yahoo. IBM alone currently employs upwards of 5,000 people in Massachusetts, primarily at the Littleton location of the IBM Mass Lab, the largest IBM Software Development Lab in North America.

There has been considerable acquisition activity in recent years as large companies seek to gain access to compelling technology or market share.
Massachusetts has witnessed a significant number of acquisitions of big data companies. From 2010 to 2013, there were 24 of these acquisitions, including the high-profile 2011 acquisitions of Vertica Systems Inc. by Hewlett-Packard and Endeca by Oracle. Acquisitions from 2013 included Crashlytics, Spindle and Blue Fin Labs by Twitter; Trusteer by IBM, Humedica Inc. by United Health, and Jumptap by Millenial Media *(See Appendix F, Table 32, "Recent Acquisitions of Companies in Massachusetts"*). Many point to this activity as recognition of the value created in the Mass Big Data ecosystem, as firms seek to access the innovative technologies and top level talent associated with the acquired firms. Other incentives that drive individual acquisitions may come from a number of important business factors, including: market-segment dominance by large incumbents which increases incentives for smaller companies to merge with large ones in order to compete successfully; the preference of many customers for comprehensive, integrated solutions which single small companies may have difficulty in providing; the challenges to small companies of achieving their own public offering; and the financial incentive that many investors and managers often have to sell a company rather than grow it to scale.

Companies headquartered in Massachusetts have also acquired firms in order to gain new technology, to broaden product offerings, and to tap new markets. EMC, for example, has made a series of acquisitions in recent years, some of which have targeted big data related firms, including acquisitions of XtremeIO, Syncplicity, and Silver Tail Systems *(see Appendix F, Table 33, "Recent Acquisition of Companies by EMC")*. In 2010, IBM acquired Netezza Corporation, headquartered in Marlborough and a global leader in data analytics and data appliances. Integrated into IBM's big data business, Netezza continues to help expand IBM's business analytics initiatives and help global clients gain faster insights into their business information.

The Commonwealth's innovation ecosystem provides a strong competitive environment for Mass Big Data Companies.

The decision-making of firms is strongly influenced by the region's competitive environment, which is itself a product of a variety of factors, including: the local landscape of firm strategy and rivalry, demand conditions, availability of factor inputs (labor, technology and capital), and the strength of supporting institutions and industries.[10] Respondents to the 2013 Mass Big Data Survey generally report that the Commonwealth offers a favorable environment for big data companies. As shown in *Table 2*, over 75% of respondents agree that there is strong regional demand for products and services and 90% agree that there is a healthy rivalry among companies in the region. Most respondents agree that Massachusetts boasts a large, well-qualified labor

[10] This is based on the framework developed by Michael Porter in the *Competitive Advantage of Nations*.

supply; strong collaboration between companies and universities (but less so with healthcare institutions); and a venture capital community that is interested in making investments in this sector. However, most also believe the cost of doing business in Massachusetts is relatively high and that government policies are not yet sufficiently supportive of their types of business. (Note: the survey was administered during a period when a new potential tax on computer software service was debated but ultimately rescinded.)

TABLE 2) **Competitive Environment**

Statement	Of those offering and opinion...				
	Don't Know	Strongly Disagree	Disagree	Agree	Strongly Agree
There is strong demand for our types of products/services from customers in Massachusetts	8.3%	3.1%	21.2%	42.4%	33.4%
There is a healthy rivalry among competitors in Massachusetts	16.7%	0.0%	10.0%	73.3%	16.7%
There are strong collaborations among companies in Massachusetts	5.6%	3.0%	53.0%	35.3%	8.8%
There are strong collaborations between companies and universities in Massachusetts	13.9%	3.2%	41.9%	38.7%	16.1%
There are strong collaborations between companies and healthcare institutions in Massachusetts	41.7%	4.8%	57.1%	33.3%	4.8%
There is an extensive supplier network in Massachusetts	30.6%	4.0%	40.1%	48.0%	8.1%
Venture capitalists in Massachusetts are interested in making investments in this sector	13.9%	9.6%	12.9%	35.5%	41.9%
There is a large pool of people in Massachusetts with the skills that we need	5.6%	3.0%	35.3%	53.0%	8.8%
The cost of doing business in Massachusetts is relatively low	8.6%	25.1%	56.2%	18.7%	0.0%
Government policies in Massachusetts are supportive of our type of business	30.6%	16.0%	44.1%	36.0%	4.0%

Source: 2013 Mass Big Data Survey

A Closer Look at Key Elements of the Mass Big Data Competitive Environment

Massachusetts has a significant base of companies and organizations with an interest in big data. Massachusetts is home to world-class hospital systems *(See Appendix F, Table 34, "Largest Hospital Systems in Massachusetts by Size and Revenue")*, Fortune 500 companies *(See Appendix F, Table 35, "Fortune 500 Companies Headquartered in Massachusetts"*), and hundreds of other firms,[11] which have an interest in big data. As an example, Partners Healthcare owns and operates seven hospitals, four medical groups and five community health centers; it also has significant contractual and financial relationships with a number of other healthcare organizations. More than 500,000 lab exams are performed each day throughout its network. Adding to this, are all the images that are produced, genomic data generated, and medical records that need to be integrated and amended. This represents a massive volume of clinical data. The organization's Research Patient Data Register (RPDR) alone includes some 2 billion data elements for 6.3 million patients. The RPDR and other data are used to identify patients for clinical trials, conduct trials in silico, and assess comparative effectiveness.[12] The RPDR framework provides the foundation for a collaborative effort among hospitals to share data. Five institutions in the Boston area — Beth Israel Deaconess Medical Center, Children's Hospital Boston, Brigham and Women's, Massachusetts General Hospital and the Dana Farber Cancer Center — are part of the Shared Health Research Informatics Network (SHRINE). The open-source software is also being made available to other hospital groups across the country under government and privately funded initiatives. A planned Innovation Center, to be launched in 2014 as a partnership between Baystate Health and the Massachusetts Life Sciences Center, is organizing to provide access for researchers and innovators to healthcare data which will accelerate the development of healthcare informatics that can improve care and lower costs.

Some $2.5 billion has been pumped into Massachusetts-based big data related companies since 2000.

At least 123 companies in Massachusetts involved in various big data business segments have raised risk capital since 2000. Led by Hubspot and Jumptap — both in marketing applications that involve big data — the total amount raised by companies has approached $2.5 billion *(Table 3)*.

[11] In 2009 (latest year for which data are available), there were roughly 3,000 firms in Massachusetts with 500 or more employees.

[12] Comparative effectiveness studies also require claims data. In this regard, Partners Healthcare has found it difficult to use the APCD. Interview with Shawn Murphy, Medical Director, Research Computing, Partners HealthCare.

TABLE 3) **Top 25 Massachusetts-based Recipients of Risk Capital**

Company		VC Investment ($ million)
1	Hubspot	130.5
2	Jumptap	101.5
3	Attivio	90.1
4	GlassHouse Technologies	86.8
5	Protein Simple	85.3
6	Visible Measures	82.3
7	ExaGrid Systems	72.1
8	Celeno	68.2
9	Carbonite, Inc.	66.0
10	Endeca (Oracle)	65.0
11	Humedica, Inc.	63.0
12	Netezza (acquired by IBM PureData System 2009)	61.5
13	EnterpriseDB	56.6
14	CambridgeSoft	54.3
15	Akorri	48.7
16	Dataxu	45.8
17	Navic Networks (acquired by Microsoft)	42.0
18	Dataupia	40.0
19	Sepaton Inc.	37.5
20	VideoIQ	35.0
21	Compete	33.0
22	StreamBase Systems, Inc. (TIBCO)	32.0
23	Vertica Systems, Inc. (acquired by HP)	30.5
24	PatientsLikeMe	29.0
25	Kalido	28.6
	Sub-total	**1,485.3**
	Other (98 companies)	**994.5**
	Total	**2,479.8**

Source: Crunchbase, August 2013

Funding has been provided by 243 different venture capital firms, private equity firms, angel investor groups, and strategic investors.

Research centers in Massachusetts provide an important foundation for advances in big data technologies and uses.

A number of research centers have been established in Massachusetts to develop new technology platforms, to advance analytical techniques, and to use data to address important research questions in health, communications, cybersecurity, transportation, energy, and other fields. These include Boston University's Rafik B. Hariri Institute for Computing and Computational Science & Engineering, the Broad Institute, the Dana Farber Cancer Institute's Center for Cancer Computational Biology, Harvard University's Institute for Quantitative Social Science, MIT's BigData@CSAIL (Computer Science and Artificial Intelligence Lab) Initiative, UMass Amherst's Institute for Computational Biology, Biostatistics & Bioinformatics, and WPI's Center for Research in Exploratory Data and Information Analysis *(Appendix F, Table 36,"Selected Research Centers in Massachusetts")*.

The Commonwealth is also home to research centers that generate significant amounts of data and provide opportunities and test beds for the development of collaborative synergies and practical big data applications in partnership with the region's big data industry and academic players. The NSF-sponsored Ocean Observatories Initiative, led by and headquartered at the Woods Hole Oceanographic Institute, deploys, monitors and analyzes data from sensors and autonomous underwater vehicles in areas across the Atlantic.

In addition to major research centers, various groups have been established within institutions to collaborate on various research projects. For example, the Data Research Group at Northeastern University consists of seven faculty members with expertise and research projects in machine learning, spatial indexing, the semantic web, and data base management. Similarly, researchers in Harvard School of Public Health, Engineering and Social Systems division are using big data to develop a better understanding of "the complex behavior of human societies". Research projects across the region deal with a wide spectrum of topics, including transportation, education, social networks, and crime. This work is supported by funding from a range of sources including the federal government, foundations, private companies, and other sources.

Between 2007 and June 2013, the federal government provided at least a $19.4 million to institutions in Massachusetts for research and educational activities related to big data *(Appendix F, Table 37, "Federal Funding of Big Data Projects by Institution, 2007-2013")*.[13] Sources included a broad range of programs sponsored by the National Science Foundation, National Institutes of Health, US Geologic Survey, and the Department of Energy.[14] Activities deal with the development of new database structures, search queries, and machine learning algorithms as well as techniques for data mining and analysis of complex events. A number of projects focus on using new technologies to improve education and training.

State institutions have also committed funds explicitly to exploring big data. For example, in 2012, the University of Massachusetts awarded nearly $750,000 through the President's Science and Technology Initiatives Fund to support six research projects in areas deemed important to the Massachusetts economy. Three projects revolved around "big data" analytics; i) The Big Data Informatics Initiative ($136,250 to detect financial fraud using large scale data sets; ii) The Institute for Computational Biology, Biostatistics & Bioinformatics ($97,500) to use "big" life science data to improve and individualize clinical practices; and iii) mHealth-based Behavioral Sensing and Interventions ($185,000) to develop wearable sensor software.

[13] This excludes funds that went to companies under DOD and DARPA-funded programs.

[14] Abstracts of federally funded research projects are available on the MassTech website.

Grants totaling more than $1.1 million were provided in 2012 and 2013.

In 2012 and 2013, the founding institutions of the **Massachusetts Green High-Performance Computing Center** ("MGHPCC") sponsored a seed-fund program to catalyze sustained cross-institution research collaborations. Grants totaling more than $1.1 million were provided over two years. Projects span three key facets of research computing: the use of computers as a tool for scientific discovery, development of application software that enable new types of research, and computer science research that points the way toward next generation "exascale" computer systems. Many of these projects deal with big data. Examples include: i) designing cloud and big data platforms for scientific and hpc applications; ii) genome-scale characterization of chromosonal aberrations using parallelizable compression algorithms; iii) automated segmentation of vessel network structures in large image stack sets; iv) development of future generation "exascale" software platforms; and v) the use of high-performance computing to automate medical imaging analysis, and vi) development of a next-generation, on-demand service for managing and processing massive amounts of genome information.

Additional details on the MGHPCC are available on page 28.

Massachusetts offers a wide range of formal and informal educational opportunities for people interested in developing skills required for careers in big data.

Formal Education: Colleges and universities offer a wide range of degree and certificate programs big data. Companies that work with big data seek employees with a range of key competencies that draw strongly on skills gained from computer science, mathematics/statistics, physics and other data-intensive disciplines. In 2012, colleges and universities in Massachusetts graduated nearly 5,600 students with relevant credentials *(Table 4)*.[15]

TABLE 4) Number of Big Data Related Degrees Granted in Massachusetts, 2012

CIP	Title	Number of Schools	Number of Graduates				
			BA/BS	MA/MS	PhD	Certificate (a)	Total
11	Computer and Information Sciences	47	1,441	934	93	11	**2,479**
14.09	Computer Engineering	7	137	161	8	0	**306**
27	Mathematics and Statistics	45	1,014	133	65	17	**1,229**
40.08	Physics	27	336	106	152	0	**594**
26.0203, 26.0206	Biophysics and Molecular Biophysics	4	3	1	18	0	**22**
40.0202, 40.0403, 40.0603	Astrophysics, Atmospheric Physics and Dynamics, and Geophysics and Seismology	7	24	3	2	0	**29**
26.11	Biomathematics, Bioinformatics, and Computational Biology	7	6	57	41	1	**105**
51.2706	Medical Informatics	2	0	25	0	0	**25**
45.0603	Econometrics and Quantitative Economics	1	37	0	0	0	**37**
14.37	Operations Research	2	0	17	10	0	**27**
52.12	Management Information Systems and Services	11	64	76	0	0	**140**
52.13	Management Sciences and Quantitative Methods	7	219	280	0	31	**530**
30.06	Systems Science and Theory	2	1	31	0	0	**32**
30.08	Mathematics and Computer Science	2	28	0	0	0	**28**
	Total		**3,310**	**1,824**	**389**	**60**	**5,583**

Notes: (a) Post-baccalaureate certificate
Source: National Center for Education Statistics

[15] A breakdown of graduates by institution is included in Appendix D.

Some of these "big data relevant" programs in Massachusetts are in fact multidisciplinary courses of study that combine computer science and mathematics/statistics with application to specific domains within or related to big data itself. While the particular term "data science" is present in relatively few program titles, many of the programmatic requirements for these degrees include data science-driven courses. The following is a selection of "data science driven" programs from across the Commonwealth (*For the complete list and detailed descriptions of the programs, see Appendix E, "Talent Pipeline & Higher Education Data")*.

- **Bentley University, Graduate Certificate in Business Analytics**
- **Boston University, Master of Science in Systems Engineering**
- **Harvard University, Master of Science in Computational Science and Engineering**
- **Massachusetts Institute of Technology, Master of Science in Operations Research**
- **Northeastern University, Master of Science in Health Informatics**
- **Worcester Polytechnic Institute (WPI), Master of Science in Data Scienc**e

As shown in *Table 5*, most survey respondents felt that students graduating from colleges and universities in Massachusetts are equipped with skills needed by employers, but the number of graduates is insufficient to fill the Mass Big Data cluster's open positions. The few respondents that reported current programs as being inadequate point to the need for more training in Hadoop and a broader variety of databases as well as machine learning.

TABLE 5) **Perceptions of Colleges and Universities in Massachusetts**

Perceptions	Of those offering and opinion...				
	Don't Know	Strongly Disagree	Disagree	Agree	Strongly Agree
Colleges and universities in Massachusetts are producing graduates with needed skills (a)	13.16%	3.0%	18.2%	57.6%	21.2%
Colleges and universities in Massachusetts are producing graduates with needed skills, but the number of graduates is insufficient	12.82%	0.0%	35.3%	35.3%	29.4%
Colleges and universities in Massachusetts are producing graduates with needed skills, but they are leaving the state after graduation	36.84%	0.0%	16.7%	66.7%	16.7%

Note: (a) For consistency, the statement in the original survey was reworded in the affirmative.

A significant percentage of graduates currently leave Massachusetts upon graduation, but within these numbers, there are important differences among schools and degree levels. *Table 6* presents placement data for the class of 2012 computer science graduates from the University of Massachusetts Amherst and MIT. Graduates with a bachelor's degree from UMass are more likely to enter the labor market directly upon graduation compared to those from MIT (93 percent versus 46 percent). Moreover, UMass graduates who enter the labor market and are successful in securing a job are more likely to take a job in Massachusetts than MIT graduates (88 percent versus 48 percent). While differences remain, half or more of graduates with advanced degrees from both schools take jobs outside of the Commonwealth.

TABLE 6) **Placement of 2012 Computer Science Graduates**

School	Degree	Percent Further Education	Percent in Labor Market	Percent Employed	Of Those Employed, Percent Employed in MA	Percent Other
UMass Amherst	Bachelors	7%	93%	84%	88%	9%
	Masters	20%	80%	80%	50%	0%
	PhD	0%	100%	100%	29%	0%
MIT	Bachelors	54%	46%	44%	48%	2%
	Masters	15%	85%	85%	38%	0%
	PhD	0%	100%	100%	49%	0%

Sources: 2012 MIT Graduating Student Survey. Response rates: PhD – 76%; Master's – 69%; and Bachelor's – 78%.
2012 UMASS Amherst Graduating Senior Survey (OIR). Response rate: 62%. UMass CS website. Response rates: Masters – 24%; PhD – 100%.

Outside of collegiate-based educational programs, there are numerous other channels for individuals to obtain requisite skills and expertise. A number of companies provide training that is specifically related to their products.

Nearly 1,000 people have completed EMC's five-day course.

For example, **EMC** has broadened the types of training that it offers to include professional development. In this regard, the company offers two data science courses, including a one-day course aimed at business managers and a five-day course aimed at data scientists as shown in Table 7. The latter covers various data management technologies, including Hadoop, as well as advanced analytical techniques using R and other packages. Participants put what they have learned into practice in a final lab session. EMC offers the training online or at one of EMC's education centers. Alternatively, arrangements can be made to conduct training at a customer's premises. Participants who opt to take and pass an exam receive EMC ProvenTM Professional Data Scientist Associate (EMCDSA) certification. To date, nearly 1,000 people have completed the five-day course. In addition, nearly 900 college and universities have access to course materials through EMC's Academic Alliance program. EMC has worked with a number of local schools, including Babson College, to develop data science programs.[16]

TABLE 7) **EMC Data Science Training Courses**

Course	Duration	Topics
Data Science and Big Data Analytics	Five days	• End-to-end data analytics lifecycle • Using R to execute basic analytics methods • Advanced analytics and statistical modeling for Big Data – Theory and Methods • Advanced analytics and statistical modeling for Big Data – Technology and Tools • Lab: Putting it all together
Data Science and Big Data Analytics for Business Transformation	One day	• Deriving Business Value from Big Data • Leading Analytic Projects • Developing Data Science Teams • Driving Innovation via Analytic Projects

Source: https://education.emc.com

Established in late 2012, hack/reduce is an effort to train and retain data scientists in the Boston area.[17]

The organization has received financial support from technology vendors, venture capital firms, and the state government.[18] hack/reduce organizes hackathons, workshops and other events. In 2013, it organized 12 hackathons, 22 workshops and seminars, and 38 big-data focused meetups. The hackathons, workshops, and meetups are centered on particular technologies such as ElasticSearch and VoltDB or on specific issues such as forecasting the success of product launches and providing nutritional information on combat rations to soldiers. Over 1,000 participants are estimated to have taken part in these training and collaboration activities in 2013.

Similarly, H@cking Medicine was established in 2011 with the support of the Martin Trust Center for MIT Entrepreneurship "to teach entrepreneurs and clinicians the skills necessary to launch disruptive healthcare businesses."[19] In addition to conferences and seminars on various topics, it organizes hackathons, which focus on healthcare-related issues. The first hackathon was undertaken in concert with Athena Healthcare in May 2013; the second was in September 2013 at Brigham and Women's Hospital.

Governor Patrick at the hack/reduce launch.

Eric Haynes/Governor's Office

[17] Nexus Associates, interviews for 2014 Mass Big Data Report.
[18] A progress report issued in April 2013 states that the organization had received $25,000 from the Commonwealth and $375,000 from partners.
[19] http://hackingmedicine.mit.edu/mission

Finally, more than 25 meet-up groups have been established in recent years in the Boston/Cambridge area that relate to big data.[20] Some function as user groups organized around specific software such as Python, R, Hadoop, and NoSQL. Others are oriented to analytical techniques or application in particular domains. Three examples are presented in *Table 8*.

TABLE 8) Examples of Meet-up Groups

Meet-up Group	Est.	Description	Members	Events
Boston Predictive Analytics	2010	The goal of this meet-up is present informative lectures, hands-on tutorials, networking events, etc, towards helping the local community further it's understanding and proficiency regarding Predictive Analytics. The group has three main focal points: business applications, advanced mathematics, and computer science; with topics covering Recommender Systems, Machine Learning, Google Analytics, Data Visualization, Social Media / Text Analytics, and related topics.	2,559	32
Boston Hadoop User Group	2009	Goal of most meetings will be build data models that attendees can use themselves; make data mining and data analytics accessible to everyone; and increase awareness of open source data mining tools.	1,644	33
Data Science Group	2012	This group will concentrate on understanding the tools and skill-sets needed to become an effective Data Scientist. They explore all topics related to the data lifecycle including acquiring new data sets, parsing new data sets, filtering and organizing data, mining data patterns, advanced algorithms, visually representing data, telling stories with data and softer skills such as negotiations and selling your ideas based upon data.	1,096	10

Source: Meetup.com downloaded August 20, 2013

Boston Python Workshop.

[20] See Appendix F, Table 51

The Commonwealth has taken steps to improve access to data, strengthen computing resources, and extend broadband coverage.

1. Massachusetts Open Data Initiative

In November 2009, the Commonwealth launched the “Open Data Initiative”, an important effort to increase government transparency and to improve public access to open data sets. The resulting “Open Data Catalog” website [21] provides links to available datasets in a range of categories: economy, education, energy, environmental, financial, geography, health, housing, licensing, municipalities, population, public safety, technology, and transportation.[22]

2. Electronic medical records and claims data

An enormous amount of information is generated and used in the course of caring for the citizens of the Commonwealth. Physicians, surgeons, nurses, pharmacists, psychologists and other healthcare professionals as well as insurance payers routinely collect, store, review, analyze and transmit health care-related information. The meaningful use of electronic heath records [23] is key to ensuring that healthcare focuses on the needs of the patient, is delivered in a coordinated manner, and yields positive health outcomes at the lowest possible cost.

- **MassHIway**
  The Commonwealth was one of the first states in the country to pass legislation that requires all health care providers to adopt electronic health record systems and connect these systems to a health information exchange. The health information exchange in Massachusetts — known as the Massachusetts HIway — is a secure network for sharing electronic health records and other health-related information among hospitals, doctors’ offices, pharmacies, skilled nursing facilities, laboratories, and other healthcare-related organizations.

- **All-Payer Claims Database (APCD)**
  Massachusetts is one of only nine states that have a functional All-Payer Claims Database (APCD) in operation [24] — a central repository of healthcare claims and payment data. Maintained by the Center for Health Information and Analysis (CHIA), the APCD includes medical, pharmacy, and dental claims

[21] https://wiki.state.ma.us/confluence/display/data/Data+Catalog

[22] https://wiki.state.ma.us/confluence/display/data/Open+Data+Meeting

[23] An electronic health record (EHR) contains information on an individual patient, including demographic data, medical history, diagnoses, medications, allergies, immunization status, vital signs, lab results, radiology images, clinical notes, and insurance and billing information. An EHR is generated and maintained within a particular institution such as a hospital, long-term care facility, clinic, or physician office. EHR systems can include computerized provider order entry (CPOE), electronic prescribing, clinical decision support, patient reminders, and calculation of clinical quality and efficiency measures. If linked through a health information exchange, information contained in an EHR can be shared across different institutions.

[24] The other eight states are ME, NH, VT MD, TN, MN, KS, and UT.
http://www.mass.gov/chia/docs/p/apcd/apcd-overview-updated-2013-04-11.pdf

data for all payers covering residents of Massachusetts. While the primary purpose of the APCD is to improve the efficiency routing claims data to state agencies and facilitating planning and administration, the APCD is also intended to be used for research that supports lower costs and better care.[25] Qualified researchers may seek CHIA approval to access data, in compliance with state and federal privacy laws and regulations. Privacy rules developed by CHIA limit the scope of data requests, including requiring that only the minimum data necessary for the study be released and specifying steps to ensure the physical security of data files. CHIA accepts applications from state agencies, system providers & payers, and researchers. Through August 20, 2013, eight applications had been approved; all from either public agencies or universities *(See Appendix F, Table 52, "Applications for the Use of APDC")*.

3. Massachusetts Green High Performance Computing Center (MGHPCC)

Opened in November, 2012, the MGHPCC is a groundbreaking collaboration among Boston University, Harvard University, Massachusetts Institute of Technology, Northeastern University, the University of Massachusetts, and the Commonwealth. Located in Holyoke, Massachusetts, and running on green hydroelectric power, the state-of-the-art data center helps to provide cutting-edge research computing resources for the five participating institutions and broader research community. The computing resources housed in the center include one of only five "Atlas Tier 2 Centers" in the United States, established to enable the analysis of data from the Large Hadron Collider (LHC) at CERN. This represents a significant computing resource for MIT's Joint Program on the Science and Policy of Global Change, along with many others. In February 2013, the Massachusetts Life Sciences Center[26] awarded $4.54 million to MGHPCC to expand its capacity for life sciences-related research and data analysis. The new computing system – Commonwealth Computational Cloud for Data Driven Biology – will be dedicated to "enhancing life sciences research through large-scale computation and big data analytics."[27] The state provided a $25 million grant to help defray the initial construction cost of the facility in addition to $14.5 million in New Market Tax Credit through MassDevelopment.

4. Mass Broadband 123

The Massachusetts Technology Collaborative received $45.4 million in federal grants under the American Recovery and Reinvestment Act (ARRA) and $44.3 million in state funds to build and operate a 1,200 mile fiber optic network designed to provide high-speed internet access to over 1.200 public institutions in 120 communities in the western and north central parts of the state. The Massachusetts Broadband Institute (MBI), an operating division of the Massachusetts Technology Collaborative, completed construction, testing, and handover of this 'middle mile' network in February 2014. The system is designed to be capable of transferring data at speeds of roughly two gigabytes per second and provides the backbone to expand high-speed internet access to regions that lack broadband connectivity.[28]

[25] Only de-identified data is made available with the following exceptions: state agencies, subject to state and federal laws and regulations protecting patient privacy; providers and payers for "carrying out treatment and coordinating care among providers;" and consumers accessing data on services they personally received.

[26] The Massachusetts Life Sciences Center (MLSC) is a quasi-public agency of the Commonwealth of Massachusetts tasked with implementing the Massachusetts Life Sciences Act, a 10-year, $1-billion initiative that was signed into law in June of 2008. The MLSC's mission is to create jobs in the life sciences and support vital scientific research that will improve the human condition. This work includes making financial investments in public and private institutions that are advancing life sciences research, development and commercialization as well as building ties among sectors of the Massachusetts life sciences community.

[27] http://www.mghpcc.org/blog/mghpcc-recipient-of-major-mlsc-gran

[28] The FCC considers broadband to be capable of download speeds of at least 4 megabits per second (Mbps). To understand how fast broadband is, a 4 Mbps connection could download a 3-minute song in about 8 seconds or a 90-minute standard-definition movie in just under 30 minutes. (http://broadband.masstech.org/what-we-do/what-broadband)

Chiaki Hayashi

Massachusetts Exhibits Relative Strengths Across Multiple Dimensions

It is useful to think about the relative strengths of Massachusetts along four dimensions: **business, technology, talent, and capital.**

Business Strengths

The strength of a region depends on the success of companies in the market place. In this regard, *Table 9* provides a list of companies sorted by annual revenue (2012) generated through the sale of products and services related to big data. Three of the top 67 companies are headquartered in Massachusetts: EMC, Attiva, and Basho.[29]

[29] Attivio sells enterprise-class software that supports SQL and simple search-style queries to retrieve information via reports, dashboards and custom interfaces. Basho Technologies offers an NoSQL database (Riak) and cloud storage software.

TABLE 9) Worldwide Big Data Revenue by Vendor, 2012

Vendor		HQ	Revenue (US$ millions)			Share of Big Data Revenue		
			Total	Big Data	Big Data as % of Total	Hardware	Software	Services
1	IBM	NY	$103,930	$1,352	1%	22%	33%	44%
2	HP	CA	$119,895	$664	1%	34%	29%	38%
3	Teradata	OH	$2,665	$435	16%	31%	28%	41%
4	Dell	TX	$59,878	$425	1%	83%	0%	17%
5	Oracle	CA	$39,463	$415	1%	25%	34%	41%
6	SAP	Germany	$21,707	$368	2%	0%	67%	33%
7	**EMC**	**MA**	**$23,570**	**$336**	**1%**	**24%**	**36%**	**39%**
8	Cisco Systems	CA	$47,983	$214	<1%	80%	0%	20%
9	Microsoft	WA	$71,474	$196	<1%	0%	67%	33%
10	Accenture	Ireland	$29,770	$194	1%	0%	0%	100%
11	Fusion-io	UT	$439	$190	43%	71%	0%	29%
12	PwC	NY	$31,500	$189	1%	0%	0%	100%
13	SAS Institute	NC	$2,954	$187	6%	0%	59%	41%
14	Splunk	CA	$186	$186	100%	0%	71%	29%
15	Deloitte	NY	$31,300	$173	1%	0%	0%	100%
16	Amazon	WA	$56,825	$170	<1%	0%	0%	100%
17	NetApp	CA	$6,454	$138	2%	77%	0%	23%
18	Hitachi	Japan	$112,318	$130	<1%	0%	0%	100%
19	Opera Solutions	NY	$118	$118	100%	0%	0%	100%
20	Mu Sigma	IL	$114	$114	100%	0%	0%	100%
21	TCS	India	$$10,170	$82	1%	0%	0%	100%
22	Palantir Technologies	CA	$78	$78	100%	0%	63%	38%
23	Intel	CA	$53,341	$76	<1%	83%	0%	17%
24	MarkLogic *	CA	$78	$69	88%	0%	63%	38%
25	Booz Allen Hamilton	VA	$5,802	$68	1%	0%	0%	100%
26	Cloudera *	CA	$61	$61	100%	0%	47%	53%
27	Actian	CA	$46	$46	100%	0%	63%	38%
28	SGI	CA	$769	$43	6%	83%	0%	17%
29	Capgemini	France	$14,020	$42	<1%	0%	0%	100%
30	1010data	NY	$37	$37	100%	0%	0%	100%
31	10gen *	NY	$36	$36	100%	0%	42%	58%
32	Alteryx	CA	$36	$36	100%	0%	55%	45%
33	Google	CA	$50,175	$36	<1%	0%	0%	100%
34	Guavus	CA	$35	$35	100%	0%	67%	33%
35	VMware	CA	$3,676	$32	1%	0%	71%	29%
36	ParAccel	CA	$24	$24	100%	0%	44%	56%
37	TIBCO Software	CA	$1,024	$24	2%	0%	53%	47%
38	MapR *	CA	$23	$23	100%	0%	51%	49%
39	**Attivio**	**MA**	**$26**	**$21**	**80%**	**0%**	**62%**	**38%**
40	Fractal Analytics	CA	$20	$20	100%	0%	0%	100%
41	Pervasive Software	TX	$51	$19	37%	0%	59%	41%
42	Hortonworks *	CA	$18	$18	100%	0%	0%	100%
43	Informatica	CA	$812	$17	2%	0%	78%	22%
44	QlikTech	PA	$321	$16	5%	0%	74%	26%
45	DataStax *	CA	$15	$15	100%	0%	59%	41%
46	**Basho ***	**MA**	**$14**	**$14**	**100%**	**0%**	**63%**	**38%**
47	Microstrategy	VA	$595	$13	2%	0%	59%	41%
48	Tableau Software	WA	$130	$13	10%	0%	59%	41%
49	Couchbase *	CA	$12	$12	$100%	0%	64%	36%
50	Kognitio	UK	$12	$12	100%	0%	47%	53%

Note: * Vendors with primary focus on Hadoop and NoSQL.

Source: http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017

Industry Self-Identification by Region — Location Quotient Analysis

A location quotient (LQ) is a widely used indicator to measure the degree of concentration of industries, occupational skills or other assets of a region relative to the nation (or other reference area) in order to reveal the particular strengths of the region.[30] LinkedIn is one of, if not the most wide used social media career resources that provides firms and prospective employees with the opportunity to describe themselves, their business, and their training. The service reports that it has roughly 93 million individual users in the United States as of March, 2014, which represents approximately 29% of estimated U.S. population. In *Table 10*, the cells highlighted in yellow denote the metropolitan areas with companies that were at least 20 percent more likely to list the respective keywords in their LinkedIn profiles than in the nation as whole.

While Massachusetts has strengths across a wide range of areas, the results suggest that the Commonwealth has a particularly high degree of specialization in a number of areas, including data integration, data management (specifically NoSQL/NewSQL and data warehouses), machine learning/predictive analysis, network analytics, semantic analysis, social media analysis, data visualization, and bioinformatics/genomics. Associated keywords for these areas are at least 30 percent more likely to be found in profiles of companies in Massachusetts than in the US as a whole. Conversely, the table suggests that the relative strengths of the greater Washington, DC area in cybersecurity, NYC and Chicago in automated trading, NYC and Los Angeles in ad targeting /serving, and the San Francisco Bay Area in Hadoop and related technologies.

TABLE 10) **Location Quotient – Business Segment**

Business Segment	BOS	CHI	DCA	DFW	NYC	PHL	RDH	LAX	SAN	SFO	SEA
Big Data	1.9	0.7	1.1	0.7	1.1	0.7	1.0	0.7	0.8	2.0	1.6
Data Integration	1.4	1.1	1.2	1.0	0.8	1.2	1.2	1.1	1.0	0.8	0.4
Data Management	1.3	0.9	0.7	1.0	0.8	1.0	0.8	1.2	0.6	1.4	1.0
NoSQL/SQL	1.6	0.9	0.5	0.4	0.9	1.1	0.8	0.9	0.5	1.6	0.9
Hadoop	1.0	0.6	0.6	0.9	0.8	0.6	1.0	1.4	0.9	1.9	1.5
Data Warehouse	1.2	1.2	0.8	1.6	0.8	1.0	0.8	1.2	0.4	1.1	0.8
Data analysis and visualization	0.8	1.0	1.0	1.1	1.0	1.1	0.9	0.9	1.0	0.9	1.0
BI and business analytics	0.6	1.1	1.0	1.5	0.9	1.3	1.1	1.0	0.9	0.7	1.2
Data mining and analysis	0.7	0.9	1.0	0.8	1.0	1.0	0.8	0.8	1.0	0.8	0.9
Data science, machine learning / predictive analytics	1.5	1.2	0.7	0.7	1.3	0.8	0.6	1.0	1.6	1.5	0.8
Semantic analysis	1.2	0.5	1.3	0.8	1.4	0.9	1.3	1.2	0.5	1.4	1.0
Geo-spatial analysis	0.8	0.3	1.2	0.5	0.7	0.5	1.3	0.4	0.7	0.6	0.9
Image analysis	1.0	0.6	0.6	0.5	0.8	0.7	0.8	1.3	1.4	1.0	0.4
Data visualization	1.5	0.9	1.3	0.3	1.0	1.0	0.6	0.8	1.5	1.0	1.1
Selected applications	0.9	1.0	1.2	0.7	1.1	0.7	1.4	1.3	1.2	0.9	0.8
Sentiment and social media analysis	1.2	0.9	0.8	0.8	1.3	0.4	1.6	1.2	0.7	1.5	0.8
Bioinformatics and genomics	2.3	0.1	1.4	0.7	0.8	0.3	2.5	0.9	2.6	1.3	1.2
Web analytics	0.6	1.3	0.5	0.6	0.8	0.8	1.7	1.5	1.1	0.6	1.1
Network analytics	1.3	0.9	0.9	0.9	0.8	0.2	0.6	0.7	0.7	0.8	0.4
Fraud, threat and risk detection	0.7	0.8	1.3	1.0	0.9	1.2	1.0	1.1	1.4	0.6	0.6
Cybersecurity	0.4	0.3	5.0	0.2	0.3	0.2	0.4	0.9	2.5	0.4	0.4
Automated trading	-	5.9	0.2	0.5	3.4	1.0	1.2	0.8	-	0.4	0.9
Ad targeting / serving	0.8	1.1	0.2	0.5	3.1	-	1.2	3.5	0.3	1.6	0.2

Source: 2013 analysis of LinkedIn data (Nexus Associates)

[30] Specifically, the location quotient is a ratio of the relative concentration of a particular type of asset in a region to the nation (or other reference area). It is defined as (xi/x) / (Xi/X). An LQ greater than one indicates that the concentration of the asset in the region is greater than in the nation (or other reference area). As such, this may signify an area of comparative advantage.

Technological Strengths

An analysis of patent data provides some insight into the technological strengths of organizations in Massachusetts (*See Appendix F, Table 53, "Patents Issued to Massachusetts' Inventors, 2008 to 2012"*). While bearing in mind that data around patent registration will lag roughly 2 or more years behind the actual work that was patented, this provides a rough approximation of activity when compared across regions. A total of 5,250 patents were granted to inventors in Massachusetts between 2008 and 2012 in 23 technology classes that relate to the processing and use of data.[31,32] With roughly five percent of patents granted for these classes in the US, Massachusetts ranks fifth among all states with respect to the absolute number of patents in these classes and fourth on a per capita basis.[33,34]

Three classes — 705, 707 and 709 — account for 38 percent of patents granted to inventors in Massachusetts within this field of technology. The shares of the latter two in Massachusetts are roughly equivalent to that of the United States, suggesting no particular comparative advantage. Moreover, although more than 500 patents in Class 705 were granted to inventors in Massachusetts between 2008 and 2012, this class of technology is relatively underrepresented in the Commonwealth. Class 705 is a collection of 20-plus data processing techniques that relate to marketing and advertising, electronic shopping, insurance, stock/bond trading, healthcare management, reservation systems, and other "business method" applications.

As shown in the last column, there are four classes (highlighted in yellow) with concentrations that are at least 20 percent greater in Massachusetts than in the US as a whole — 703, 711, 717 and 718 — suggesting that the region has a comparative advantage in these technologies. For example, the share of patents accounted for by Class 703 in Massachusetts is twice as large as the share in the US as whole. Class 703 deals with processes or apparatus for sketching or outlining of layout of a physical object or part; representing a physical process or system by mathematical expression; modeling a physical system that includes devices for performing arithmetic and some limited logic operation upon an electrical signal, which is a continuously varying representation of physical quantity; modeling to reproduce an electronic, electrical, or nonelectrical device or system to predict its performance or to obtain a desired performance; and processes or apparatus that allows the data processing system to interpret and execute programs written for another kind of data processing system.

[31] Patents are organized based on the US Patent Classification System (USPC). Under the USPC, a "class" generally delineates one technology from another; "subclasses" delineate processes, structural features, and functional features of the subject matter encompassed within the scope of the class.

[32] Patent origin is determined by the residence of the first-named inventor listed on the patent grant.

[33] Absolute number of patents granted: California - 33,243; Washington - 11,691; Texas - 9,924; New York - 6,356; Massachusetts - 5,250; Total US - 105,576. Patents per 100,000 population: Washington 169.5; Vermont - 88.5; California - 87.4; Massachusetts – 79.0; Oregon - 58.2; Total US - 33.2.

[34] There are no classes for which Massachusetts ranks first on either dimension.

Massachusetts has the highest per capita Big Data-related graduation concentration

among leading states (per 100,000 population)

Talent
Massachusetts colleges and universities offer courses geared toward the needs of organizations in the Mass Big Data ecosystem, with graduates produced and accessible from a concentrated region.

The strength and expansion of the Mass Big Data ecosystem is dependent on the availability and the accessibility of qualified talent. The regional talent pipeline in turn depends, to a great extent, on the number of graduates from relevant programs and the size and dynamics of the local labor market. In this area, Massachusetts is a clear leader in per capita graduates from data science related programs as compared to other leading states *(Table 11)*.

TABLE 11) **Graduates Per 100,000 Population**

CIP	Title	US	MA	CA	IL	NC	NY	TX	WA
Total	**All Degrees**	**46.2**	**79.5**	**36.7**	**71.1**	**38.8**	**60**	**32.6**	**36.4**
11	Computer and Information Sciences	23.15	32.79	17.74	33.72	19.09	29.4	14.91	15.83
14.09	Computer Engineering	2.71	4.6	3.9	1.52	2.25	2.61	1.93	1.59
27	Mathematics and Statistics	9.57	18.49	8.69	12.39	10.47	16.37	6.98	9.51
40.08	Physics	3.12	8.94	2.95	3.27	2.96	4.23	1.97	3.45
26.0203, 26.0206	Biophysics and Molecular Biophysics	0.09	0.33	0.13	0.24	0.04	0.2	0.07	0
40.0202, 40.0403, 40.0603	Astrophysics, Atmospheric Physics and Dynamics, and Geophysics and Seismology	0.15	0.44	0.28	0.12	0	0.06	0.33	0.22
26.11	Biomathematics, Bioinformatics, and Computational Biology	0.42	1.58	0.35	0.21	0.73	0.71	0.2	0.22
51.2706	Medical Informatics	0.13	0.38	0.09	0.58	0	0.04	0.1	0.1
45.0603	Econometrics and Quantitative Economics	0.14	0.56	0.34	0	0.17	0.1	0.17	0
14.37	Operations Research	0.37	0.41	0.57	0.19	0.18	3.04	0.21	0
52.12	Management Information Systems and Services	3.93	2.11	0.47	5.14	2.29	2.39	4.16	4.38
52.13	Management Sciences and Quantitative Methods	2.13	7.97	1.09	12.79	0.41	0.54	1.47	0.55
30.06	Systems Science and Theory	0.15	0.48	0.01	0.54	0.15	0.32	0	0.46
30.08	Mathematics and Computer Science	0.08	0.42	0.12	0.4	0.01	0.03	0.02	0.06
30.16	Accounting and Computer Science	0	0	0	0	0	0	0.02	0
30.3	Computational Science	0.01	0	0.01	0	0	0	0.03	0

Absolute numbers of graduates are available in *Appendix E, Table 30*, *"Talent Pipeline & Higher Education Data"*.
Source: National Center for Education Statistics (NCES) and US Census Bureau.

While NCES data provide insight on the number and pattern of college graduates, LinkedIn can be used to provide an indication of the pool of people in particular jobs by region. An analysis of LinkedIn data focused on three positions: data scientists, data engineers/architects, and software engineers. A 2012 Harvard Business Review article titled, “Data Scientist: The Sexiest Job of the 21st Century,” called attention to the demand for employees with an emerging set of particular competencies in order to capitalize on the promise of big data. The authors refer to the emergence of “a new key player in organizations: the data scientist,” which is defined as “a high-ranking professional with the training and curiosity to make discoveries in the world of big data.”[35,36]

An analysis of data available on LinkedIn reveals 2,149 members throughout the United States who have the title of “Data Scientist.”[37] As shown in *Table 12*, 129 people carry this title in the Greater Boston Area, representing six percent of the current pool of data scientists in the country and placing the region third behind San Francisco and New York. The other two jobs shown — data engineers/architects and software engineers — have different patterns; however, in neither case is Massachusetts ranked first.

TABLE 12) **Number of People in LinkedIn with Selected Job Titles**

City	Data Scientist	Data Engineer / Architect	Software Engineer	Total
SFO	706	484	45,624	46,814
BOS	**129**	**286**	**17,542**	**17,957**
SEA	276	669	14,073	15,018
NYC	114	443	10,855	11,412
LAX	113	212	8,933	9,258
DCA	62	176	8,374	8,612
CHI	63	305	6,927	7,295
AUS	19	232	6,644	6,895
SAN	25	46	6,086	6,157
RDH	67	97	4,934	5,098
DFW	69	77	4,509	4,655
PHI	59	209	4,123	4,391
PIT	13	32	2,413	2,458
Subtotal – 13 cities	**1,715**	**3,268**	**141,037**	**146,020**
Total – US	**2,149**	**6,642**	**238,968**	**247,759**

[35] Thomas H. Davenport and D.J. Patil, Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* October 2012.

[36] In 2011, the McKinsey Global Institute (MGI) published a report on big data. While touting the significant promise of big data, the report notes significant challenges to realize this potential, particularly with respect to the shortage of people with requisite expertise. MGI models the market for “deep analytical talent” in the United States between 2008 and 2018, which is equated to expertise in statistics and/or machine learning. MGI estimates that 156,000 people with deep analytical talent were employed in 2008 based, in part, on BLS occupational data for actuaries, mathematicians (including technicians and scientists), operational research analysts, statisticians, industrial engineers, epidemiologists, and economists. MGI expects an additional 161,000 graduates with requisite training to enter the labor market by 2018 drawn from the fields of computer and information sciences, mathematics and statistics, engineering, physical sciences and science technology, biological and biomedical sciences, social sciences, and business. Given an expected loss of 32,000 people through attrition, the total supply of people with requisite talent in 2018 is put at 285,000. MGI expects demand to increase to a total of 425,000 to 475,000 positions by 2018, leaving a shortage of 140,000 to 190,000 people with “deep analytical talent.” MGI also projects a shortage of 1.5 million “data-savvy managers and analysts who have the skills to be effective consumers of big data insights—i.e., capable of posing the right questions for analysis, interpreting and challenging the results, and making appropriate decisions.

[37] Downloaded on June 6, 2013.

Table 13 shows a per capita view that takes the size of the cities into account. Boston ranks third, second, and third in terms of data scientists, data engineers/architects, and software engineers, respectively.

Table 14 takes another cut at the data using location quotients, which demonstrate the relative concentration of the jobs in Massachusetts and other metropolitan areas compared to the US. The concentration of these positions in Massachusetts is three to five times greater than the US; however, other areas, particularly the San Francisco Bay Area have much higher concentrations of data scientists, data engineers/architects and software engineers.

TABLE 13) **Number of Jobs with Selected Titles per 100,000 Population**

City	Data Scientist	Data Engineer / Architect	Software Engineer	Total
SFO	1.89	1.46	35.71	39.05
SEA	0.45	0.23	6.95	7.63
BOS	**0.26**	**0.41**	**6.94**	**7.61**
AUS	0.16	0.22	5.56	5.94
RDH	0.06	0.12	3.62	3.80
SAN	0.13	0.13	2.68	2.93
DCA	0.24	0.14	1.96	2.34
CHI	0.15	0.12	1.18	1.44
PIT	0.04	0.04	1.19	1.27
NYC	0.10	0.12	1.02	1.23
LAX	0.08	0.10	0.93	1.11
DFW	0.04	0.09	0.70	0.84
PHI	0.02	0.03	0.66	0.71
Subtotal - 13 cities	**0.22**	**0.20**	**3.72**	**4.14**
Total - US	**0.07**	**0.07**	**1.27**	**1.40**

Source: LinkedIn. Downloaded August 20, 2013

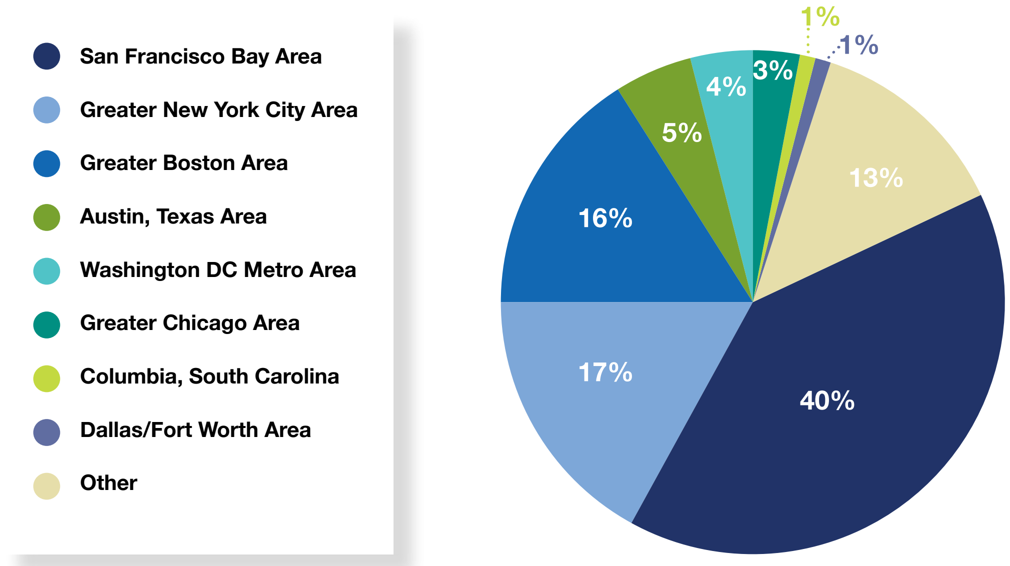
TABLE 14) **Location Quotient**

City	Data Scientist	Data Engineer / Architect	Software Engineer	Total
SFO	23.15	5.13	13.45	13.31
BOS	**4.06**	**2.91**	**4.97**	**4.90**
AUS	5.34	2.50	3.53	3.52
RDH	5.89	2.13	3.46	3.45
SEA	4.65	2.82	3.30	3.30
DCA	2.84	3.57	2.43	2.47
SAN	1.15	0.68	2.52	2.46
PIT	0.80	0.64	1.34	1.32
DFW	0.41	1.64	1.30	1.30
CHI	0.97	1.51	0.96	0.97
NYC	2.03	1.59	0.93	0.96
PHI	1.43	1.64	0.90	0.92
LAX	0.69	0.64	0.84	0.84
Subtotal - 13 cities	**3.03**	**1.87**	**2.24**	**2.24**
Total - US	**1.00**	**1.00**	**1.00**	**1.00**

Source: LinkedIn. Downloaded August 20, 2013

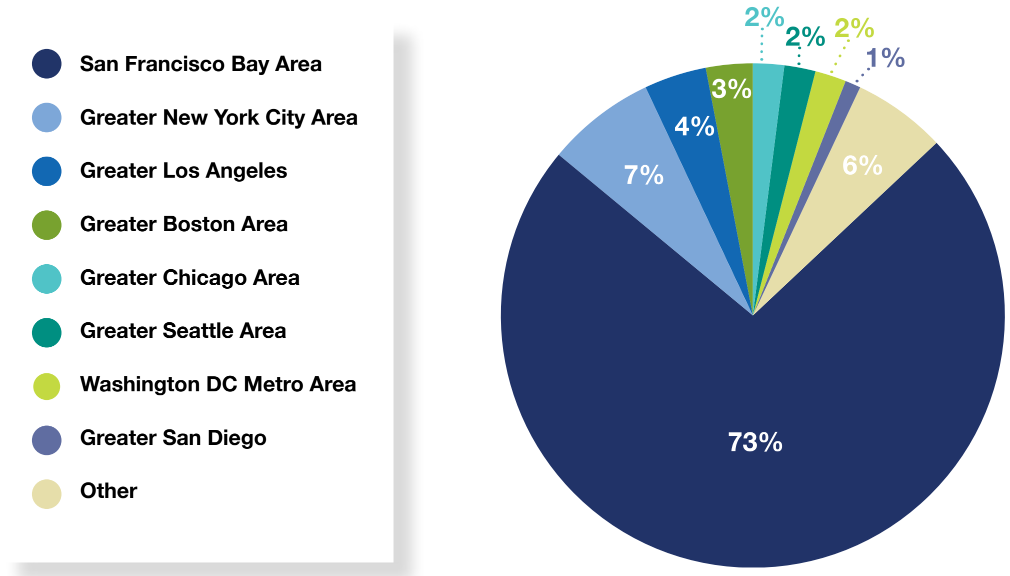
LinkedIn can be used to shed light on the universities from which people graduated and where they are currently employed, based on their own self-identification as a "data scientist" or similar. Of graduates from MIT and Harvard who are employed in jobs with the title of "Data Scientists", roughly 84 percent are currently working outside the state, including 45 percent who have taken jobs in the San Francisco Bay Area and 17 percent who are in Greater NYC *(Figure 1)*. Conversely, a much smaller fraction of graduates from Stanford and UC Berkeley leave the San Francisco Bay Area *(Figure 2)*. Almost three-quarters of data scientists listed on LinkedIn as having graduated from the two schools are currently employed in the surrounding regions. Only three percent have ventured east to Greater Boston.

FIGURE 1) **Location of "Data Scientists" Who Graduated from MIT and Harvard**



Source: LinkedIn. Downloaded August 20, 2013.

FIGURE 2) **Location of "Data Scientists" Who Graduated from Stanford and UC Berkeley**



Source: LinkedIn. Downloaded August 20, 2013.

In all likelihood, California, particularly San Francisco, will continue to draw graduates from Massachusetts and other states given the sheer magnitude of demand for qualified employees. As shown in *Table 15*, the number of jobs posted for data scientists, data engineers/architects, and software engineers is far greater in San Francisco than in Boston. (New York also has more opening than Boston for the first two positions.)

TABLE 15) **Jobs Posted on LinkedIn**

City	Data Scientist	Data Engineer / Architect	Software Engineer	Total
SFO	84	65	1,591	1,740
BOS	**12**	**19**	**322**	**353**
SEA	16	8	247	271
NYC	19	23	202	244
LAX	11	13	121	145
DCA	14	8	115	137
CHI	14	11	112	137
AUS	3	4	102	109
SAN	4	4	85	93
RDH	1	2	62	65
DFW	3	6	47	56
PHI	1	2	40	43
PIT	1	1	28	30
Subtotal - 13 cities	**183**	**166**	**3,074**	**3,423**
Total - US	**209**	**211**	**3,985**	**4,405**

Source: LinkedIn. Downloaded August 20, 2013.

Investment

Companies headquartered in Massachusetts have been successful in attracting venture capital and other strategic investment. They do not rely solely on local VC firms. Demonstrating that capital is mobile, 75 percent of the firms that provided funding to Massachusetts companies are based outside of the Commonwealth as shown in *Table 16*. Put another way, 45 percent of the 123 venture-backed companies received financing from firms based in California and 22 percent from firms headquartered in NY.

TABLE 16) **Source of Investment in Massachusetts Companies**

Headquarter State		Location of Headquarters of Firms Investing in MA Companies		MA Companies Receiving Investment from Firms Headquartered in Location	
		Number	% of Total	Number	% of Total
1	CA	64	26%	55	45%
2	MA	61	25%	72	59%
3	NY	28	12%	27	22%
4	Israel	11	5%	9	7%
5	Great Britain	9	4%	9	7%
6	VA-DC	7	3%	12	10%
7	TX	6	2%	4	3%
8	WA	6	2%	5	4%
9	Canada	5	2%	5	4%
10	CT	5	2%	4	3%
	Subtotal	**202**	**83%**	**NA**	**NA**
	Other	**41**	**17%**	**32**	**26%**
	Total	**243**	**100%**	**NA**	**NA**

Note: Excludes acquisitions.
Source: MassTech research team based on data from crunchbase.com downloaded September 2013.

Massachusetts Share of Federal Funding for Big Data

The federal government is also an important source of funding for big data initiatives. In this regard, the Obama Administration announced a "National Big Data Research and Development Initiative" on March 29, 2012. The initiative has three primary goals: i) advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data; ii) harness these technologies to accelerate the pace of discovery in science and engineering, strengthen national security, and transform teaching and learning; and iii) expand the workforce needed to develop and use Big Data technologies. To launch the initiative, six Federal departments and agencies announced more than $200 million in new commitments *(Table 17)*.[38]

Over the past 18 months (January 2012 to June 2013), organizations in Massachusetts received awards totaling almost $7.6 million under programs associated with the federal government's big data initiative, placing the state fifth in terms of the absolute volume of funding received and fourth on a per capita basis *(See Appendix F, Table 38, "Federal Funding for Big Data")*. While most monies went to universities and not-for-profit organizations, some were channeled to private companies. For example, in 2013, Raytheon BBN secured roughly $4.3 million under the DOD Warfighters Interface Technologies Advanced Research Programs.[39]

Companies in Massachusetts view California and New York as principal competition

Almost 60 percent of survey respondents ranked California as the top location for businesses in the big data sector *(Table 18)*. California has not launched state-sponsored initiatives which specifically target big data.

[38] The National Institutes of Health announced in March 2012 that it would make the 1000 Genomes Project dataset (200 terabytes) freely available to researchers Amazon Web Services (AWS) cloud.

[39] Machine Sciences, Inc and Scientific Systems Company, Inc., received federal awards prior to 2012.

TABLE 17) Federal Big Data Initiative – Announced Commitments

Dept./Agency	Initiative/Project	Funding
Defense Advanced Research Projects Agency (DARPA)	XDATA program to develop computational techniques and software tools for analyzing large volumes of data	$100 million over four years
Defense Department	Autonomous systems and improved situational awareness for warfighters	$60 million
Department of Energy	Scalable Data Management, Analysis and Visualization Institute, led by the Lawrence Berkeley National Laboratory[41]	$25 million over five years
National Science Foundation	Establishment of AMPLab at the University of California, Berkeley (a)	$10 million over five years
	Support of undergraduate training in using graphical and visualization techniques for complex data	$2 million
	Support to a research group focusing on protein structures and biological pathway	1.4 million
	Grants for EarthCube – a system to allow geoscientists to access, analysis and share information	N/A
National Science Foundation / National Institutes for Health	Core Techniques and Technologies for Advancing Big Data Science & Engineering ("Big Data") to develop and evaluate new algorithms, statistical methods, technologies, and tools for improved data collection and management, data analytics and e-science collaboration environments.	N/A
US Geological Survey	Big Data for Earth Systems	N/A

Notes (a) AMP stands for "Algorithms, Machines, and People". AMPLab is a five-year collaborative effort at UC Berkeley, involving students, researchers and faculty from a wide array of computer science and data-intensive application domains. It aims to address issues raised by "the massive explosion in online data, the increasing diversity and time-sensitivity of queries, and the advent of crowdsourcing." https://amplab.cs.berkeley.edu/about.

TABLE 18) Percentage of Respondents Ranking Locations for Big Data (Ranked 1-8)

State	1	2	3	4	5	6	7	8	Average Ranking
California	58.8%	20.6%	8.8%	2.9%	0.0%	0.0%	5.9%	2.9%	**2.0**
Massachusetts	26.5%	35.3%	14.7%	11.8%	2.9%	2.9%	2.9%	2.9%	**2.6**
New York	5.9%	23.5%	20.6%	14.7%	2.9%	8.8%	8.8%	14.7%	**4.2**
Washington	0.0%	2.9%	23.5%	17.7%	29.4%	8.8%	5.9%	11.8%	**4.8**
Virginia	0.0%	5.9%	8.8%	17.7%	11.8%	26.5%	26.5%	2.9%	**5.4**
Texas	5.9%	2.9%	8.8%	11.8%	17.7%	17.7%	14.7%	20.6%	**5.5**
North Carolina	0.0%	5.9%	8.8%	11.8%	17.7%	20.6%	17.7%	17.7%	**5.6**
Illinois	2.9%	2.9%	5.9%	11.8%	17.7%	14.7%	17.7%	26.5%	**5.9**

Massachusetts businesses participating in the Mass Big Data Study rank the Commonwealth second and New York third

Massachusetts' competitive position centers around the unique combination of business, technology and talent strengths, with some of the top global big data firms making their headquarters in the Commonwealth, including EMC, Attiva, and Basho. The concentration and access to big data talent is higher in Massachusetts than other technology focused regions throughout the United States. Respondents to the Mass Big Data Study rank Massachusetts third after California and New York for building big data businesses. Federal funding for big data initiatives through an announced $200 million commitment will be a factor in regional advancement as well.

[41] The SDAV Institute brings together researchers from six national laboratories and seven universities.

The Global Big Data market is expected to top $48 billion by 2017, up from $11.6 billion in 2012

The total Big Data market reached roughly $11.6 billion in 2012 as shown in *Table 19*. Hardware and services account for the 80 percent of revenue earned in 2012. Wikibon expects the market to top $47 billion by 2017, with applications representing the fastest growing segment. The Commonwealth is home to companies that are active in all of these areas.[40]

TABLE 19) **Projected Big Data Revenue, 2012 - 2017**

Market Segment	2012		2017	
	US$ Billion	Percent	US$ Billion	Percent
Hardware	4.27	37.0%	15.41	31.8%
Computer	2.29	19.8%	7.53	15.5%
Storage	1.75	15.2%	6.95	14.3%
Networking	0.23	2.0%	0.93	1.9%
Services	5.04	43.6%	20.9	43.1%
Professional services	4.42	38.3%	17.59	36.3%
Cloud services (Xaas)	0.62	5.4%	3.31	6.8%
Software	1.25	10.8%	4.77	9.8%
SQL	0.88	7.6%	2.51	5.2%
Hadoop and other infrastructure software	0.24	2.1%	1.14	2.4%
NoSQL	0.13	1.1%	1.12	2.3%
Applications	0.99	8.6%	7.38	15.2%
Total	**11.55**	**100.0%**	**48.46**	**100.0%**

Source: http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017.

[40] http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017. Wikibon includes the following products and services under the umbrella of Big Data: Hadoop software and related hardware; NoSQL database software and related hardware; next-generation data warehouses/analytic database software and related hardware; Non-Hadoop Big Data platforms, software, and related hardware; in-memory – both DRAM and flash – databases as applied to Big Data workloads; Data integration and data quality platforms and tools as applied to Big Data deployments; advanced analytics and data science platforms and tools; application development platforms and tools as applied to Big Data use cases; business intelligence and data visualization platforms and tools as applied to Big Data use cases; analytic and transactional applications as applied to Big Data use cases; and Big Data support, training, and professional services.

Specific industry segments/verticals may offer greater promise for growth in Massachusetts

The survey asked respondents to rank business segments in terms of the relative promise for growth in Massachusetts and indicate which verticals offered the potential for substantial growth in the Commonwealth. Results are presented in *Tables 20 and 21*. Respondents generally view segments related to the technology platform the most promising, led by data integration software. In terms of verticals, respondents highlighted healthcare, life science and financial services.

TABLE 20) **Percentage of Respondents Ranking Growth Prospects for Big Data Business Segments (Ranked 1-9)**

Segment	1	2	3	4	5	6	7	8	9	Average Ranking
Data integration software	33.3%	13.3%	13.3%	20.0%	6.7%	6.7%	3.3%	3.3%	0.0%	3.0
Data analysis software	13.3%	20.0%	23.3%	10.0%	6.7%	20.0%	3.3%	0.0%	3.3%	3.7
Data management software	13.3%	23.3%	13.3%	13.3%	16.7%	6.7%	6.7%	6.7%	0.0%	3.8
Business intelligence software	13.3%	13.3%	23.3%	10.0%	16.7%	10.0%	6.7%	6.7%	0.0%	3.9
Data visualization software	3.3%	6.7%	3.3%	30.0%	20.0%	13.3%	20.0%	3.3%	0.0%	4.9
Applications geared to specific verticals	6.7%	16.7%	10.0%	3.3%	10.0%	16.7%	20.0%	10.0%	6.7%	5.1
Systems integration	3.3%	3.3%	3.3%	3.3%	10.0%	6.7%	30.0%	40.0%	0.0%	6.5
Consulting / training	6.7%	0.0%	6.7%	6.7%	3.3%	10.0%	10.0%	23.3%	33.3%	6.9
Hardware	6.7%	3.3%	3.3%	3.3%	10.0%	10.0%	0.0%	6.7%	56.7%	7.1

TABLE 21) **Prospect for Substantial Growth in Big Data Verticals in Massachusetts**

Industry	Of those offering and opinion...						V+E
	Don’t Know	Not Promising	Slightly Promising	Moderately Promising	Very Promsing	Extremely Promising	
Healthcare	16.13%	0.0%	0.0%	3.9%	42.3%	53.8%	96.1%
Life sciences	16.13%	0.0%	0.0%	7.7%	46.2%	46.2%	92.3%
Financial services	15.63%	0.0%	3.7%	18.5%	40.7%	37.0%	77.8%
E-Commerce	28.13%	0.0%	8.7%	39.1%	39.1%	13.1%	52.2%
Education	21.88%	4.0%	12.0%	32.0%	32.0%	20.0%	52.0%
Energy	37.50%	5.0%	15.0%	30.0%	25.0%	25.0%	50.0%
Social media	18.75%	11.5%	15.4%	26.9%	23.1%	23.1%	46.2%
Telecommunications	25.81%	4.4%	13.0%	43.5%	26.1%	13.0%	39.1%
Homeland Sec./Defense	18.75%	7.7%	7.7%	46.2%	11.5%	26.9%	38.5%
Manufacturing	32.26%	4.8%	23.8%	38.1%	23.8%	9.5%	33.3%
Transportation	43.75%	11.1%	16.7%	50.0%	22.2%	0.0%	22.2%
Entertainment	28.13%	21.7%	30.4%	30.4%	8.7%	8.7%	17.4%

Note: Several respondents added the “internet of things” as another area offering significant promise for growth.

Big data companies expect to add a significant number of jobs in Massachusetts over the next 12 months.

The companies that responded to this survey question are currently seeking to fill 294 positions in Massachusetts and expect to add a total of 387 jobs in the Commonwealth over the next 12 months.

The survey asked respondents to identify the three job positions that they are finding most difficult to fill. The most frequently mentioned positions are shown in *Table 22*.[41]

TABLE 22) **Difficult to Fill Positions**

Job Title	Among Top Three	Hardest to Fill
Software engineer	68%	39%
Data engineer/architect	32%	10%
Data scientist	26%	13%
Marketing and sales	26%	10%
Product manager	23%	6%

Other positions included engineers, data analyst, and consultant.

The required qualifications vary by position *(See Appendix F, Table 49, "Required Degrees by Job Types" and Table 50, "Required Knowledge of Specific Tools").* In general, companies are looking for candidates with a bachelor's degree; although a few companies indicated that the minimum requirement is a masters degree or PhD, particularly for data scientists and data engineers/architect positions.

A review of job postings for "Data Scientists" posted on LinkedIn sheds light on the qualifications sought by companies for these positions. In general, companies are looking for individuals who are well grounded in statistics and computer science. Required degrees include computer science, economics, statistics, mathematics, physics and/or other quantitative field. Companies want people with hands-on experience working with large datasets and a firm grasp of Hadoop and NoSQL as well as traditional relational databases. In this regard, they are looking for candidates who are proficient in programming languages (such as perl, ruby, python, C/C++, java), data extraction languages (SQL), and statistics packages (such as R, Matlab, SAS). Excellent communication skills and an ability to work in teams are a must. Depending on the specific role, domain expertise may also be required.

"Data Architects" or "Data Engineers" are responsible for designing reliable and scalable data platforms to collect and process very large amounts of data (structured and unstructured) as well as a standard interfaces to support data analysis. A review of job postings on LinkedIn suggests that companies are generally looking for people with substantial experience building large-scale, distributed data processing systems who have expertise in traditional RDBMS as well as big data architectures (Hadoop, Pig, Hive), NoSQL data stores, and analytical tools. An ability to work in a team is critical as is good communication skills. Typically, companies are looking for candidates with a bachelor's or advanced degree in computer science or computer engineering.

[41] Mathworks accounts for 200 of the open positions.

Recommendations for Action

Clearly, the Commonwealth has significant assets upon which to strengthen the foundation of the Mass Big Data ecosystem and expand its global leadership. The findings in this report suggest particular areas of opportunity and recommended steps to help unleash the transformative potential of the sector to enhance economic, public, and social benefits across the Commonwealth.

Companies believe that action is needed across a broad range of issues to ensure the growth of the Massachusetts Big Data Ecosystem

The majority of respondents to the 2013 Mass Big Data Survey who voiced an opinion indicated that each of the steps listed below were very, or extremely important *(Table 23)*. Respondents to the survey identified the top priorities to address as; i) securing increased federal funding in support of big data research and innovation, ii) coordinating efforts to expand the talent pool available to industry, and iii) widening the engagement between industry and colleges & universities. These top priorities establish a clear focus around partnering with Massachusetts colleges and universities to increase their engagement with and support for the wider Mass Big Data ecosystem.

TABLE 23) **Importance of Potential Actions**

Potential Actions	Of those offering and opinion...						V+E
	Don't Know	Not Promising	Slightly Promising	Moderately Promising	Very Promsing	Extremely Promising	
Take steps to increase Massachusetts' share of federal grants for big data initiatives	5.88%	3.1%	12.5%	12.5%	43.8%	28.1%	71.9%
Support efforts to increase the supply of workers with needed skills	6.06%	3.2%	6.5%	19.4%	38.7%	32.3%	71.0%
Foster greater collaboration between companies and universities	5.88%	3.1%	3.1%	25.0%	31.2%	37.5%	68.7%
Provide financial incentives to attract and retain companies	11.76%	13.3%	3.3%	16.7%	30.0%	36.7%	66.7%
Develop new data science curricula in colleges and universities in the state	8.57%	6.2%	9.4%	21.9%	25.0%	37.5%	62.5%
Foster greater collaboration between companies and healthcare institutions	14.71%	3.4%	13.8%	20.7%	34.5%	27.6%	62.1%
Pilot the use of new technology platforms or applications to demonstrate value	11.76%	13.3%	6.7%	20.0%	36.7%	23.3%	60.0%
Make government data readily available to companies	5.88%	9.4%	3.1%	28.1%	28.1%	31.2%	59.4%
Make healthcare data readily available to companies	17.65%	14.3%	3.6%	25.0%	28.6%	28.6%	57.1%
Invest in university-based R&D to address critical technology issues	11.76%	3.3%	16.7%	26.7%	23.3%	30.0%	53.3%
Build a capacity within state and city government to use big data to address social concerns	11.76%	13.3%	13.3%	20.0%	30.0%	23.3%	53.3%
Strengthen marketing efforts to promote Massachusetts as a dominant player in big data	5.71%	15.2%	9.1%	24.2%	15.2%	36.4%	51.5%
Provide access to an open cloud computing infrastructure geared to big data	5.88%	12.5%	12.5%	25.0%	31.2%	18.8%	50.0%

Source: 2013 Mass Big Data Survey

The Mass Big Data Ecosystem is faced with a range of compelling opportunities

Important opportunities for growth in the regional ecosystem can be met through collaboration among industry, academia, government, and the public in efforts that help to realize the full potential of Mass Big Data.

Strengthen Awareness and Promotional Efforts.

The Commonwealth already boasts a tremendous landscape of strong big data assets and has an opportunity to increase business activity and economic growth by further improving the interconnectedness of the Mass Big Data ecosystem. Efforts to strengthen promotional efforts around the Mass Big Data ecosystem, with a particular focus on the talent pipeline to attract and retain talent, will support new connections, collaborations, innovations, and increased and more efficient leveraging of the strong regional big data workforce. A broad-based campaign would create buzz about Mass Big Data through the use of websites, social media, and press coverage. This could highlight the innovative uses of big data, the important role played by data scientists and other key members of teams, and the success of entrepreneurs in Massachusetts.

Expand Opportunities for Data Science Education and Training.
More than two-thirds of survey respondents felt that there is a need for new data science programs in Massachusetts, citing growing demand for employees with training in both computer science and mathematics/statistics as well as knowledge of specific domains. Others respondents suggested that, in parallel to the creation of any new degree, courses in computer science and mathematics/statistics should be integrated into a broader range of degree programs and that an emphasis be placed on developing flexible degree or certificate programs for midcareer professionals that allow for part-time, evening, or online classes.

The Commonwealth can respond to this need by assisting colleges and universities in their efforts to develop bachelors' degree programs and also flexible, professional certification courses for those potential students who are already part of the labor force. This represents a strong investment as employees already employed in the Commonwealth are more likely to remain in-state after receiving training. Valuable programs would involve practical experience addressing real world problems using large datasets and leading edge data management and analytical tools. Capstone projects, hackathons and internships could be used to provide such training while at the same time engaging students with a local community of practitioners.

Provide Better Access to Public Data and Health Records.
State and local governments in the Commonwealth have made great strides to identify and release public data. Best practices from these efforts can inform follow-on steps and the process of opening data sets can be accelerated. Making public data readily available in a machine-readable format enables use by researchers and application developers to develop the next generation of analysis and tools with direct regional benefit. A crucial element of this effort is the establishment of a centralized open data repository. Such a resource would support and be fed by the existing open data policies and requirements of state institutions and can help management efforts to hold these organizations accountable for providing data in keeping with their requirements. Standard application programming interfaces (APIs) can be developed to help developers build apps using public data. This is particularly important for very large datasets and dataset that change frequently. Information on new datasets should be posted on website and distributed through social media such as Twitter and Facebook. Beyond building the system, robust steps can be taken to engage the developer community, including meet-ups, hackathons, and other events.

Strengthen Opportunities for Company Growth through Novel Collaboration in Industry Verticals with Academia.
Support the development of opportunities for industry and academia to exchange information and work together to leverage new capabilities to address challenges in specific industry verticals. Help to foster beta sites and testbeds for next generation technologies and innovative solutions grounded in the details and requirements of industry sectors.

Accelerate Regional Innovation and Public Benefit By Leveraging Open Data from Government and Other Sources.
Provide opportunities for university researchers, industry and civic-minded coders to collaborate in new partnerships around the development of next generation technologies, tools, and analytics. The Commonwealth

can create momentum through initiative efforts including launching award programs to spur the development of big data software applications focused on the Commonwealth. The Commonwealth holds the potential to encourage developers to use public data to create practical applications targeted at specific issues related to the delivery of public services and the quality of life in the Commonwealth. This begins with making machine-readable data readily available through standard APIs. However, effort is needed to engage the developer community. In addition to hackathons and other events, the Commonwealth could establish an annual competition for apps that use public data. Such a competition may be organized around specific issues such as encouraging public transportation, promoting recycling, improving public education, facilitating access to public series, and promoting wellness.

Establish a Center of Excellence to Address Public issues: Health, Energy, Education & Transportation.

The Commonwealth should establish a center of excellence, which focuses on using big data to address public issues. The center should bring together individuals from academia, government and industry with requisite expertise. The center should be involved in the full lifecycle of a project, including obtaining data, developing analytical datasets, conducting analysis, developing insights, and communicating results to different audiences. Working in concert with state and local government, the center should address a combination of immediate problems and long-term issues in the Commonwealth.

Take Steps to Secure a Greater Share of Federal Grants.

The federal government issues new solicitations for funding programs routinely. The Commonwealth should establish a capacity to monitor the development of solicitations pertaining to big data, offering researchers early intelligence on planned solicitations, including potential foci, eligibility, and evaluation criteria. Information should be forwarded to relevant researchers in universities, not-for-profit institutions and private companies. Advanced awareness of programs in the pipeline within agencies would allow researchers greater time to put together competitive proposals. In this regard, where appropriate, the potential for collaboration among institutions and companies should be explored as early in the process as possible. In some cases, cost sharing may be mandatory.[44] Massachusetts should consider establishing a fund to help meet cost-sharing requirements. State contributions should be contingent on securing industry support for the proposed project.

Through the work of its Mass Big Data Initiative, the Commonwealth is actively engaged in the work of supporting public and private sector collaboration among the region's leading industry, academic, and non-profit actors to expand opportunities for growth across the Mass Big Data Ecosystem.

[44] NSF only requires mandatory cost-sharing for five programs: Major Research Instrumentation Program, Robert Noyce Scholarship Program, Engineering Research Centers, Industry/University Cooperative Research Centers, and the Experimental Program to Stimulate Competitive Research. NSF prohibits voluntary committed cost sharing in all proposals, including those that require mandatory cost sharing. (Note: "committed" contributions are subject to audit.) Under NSF policy, organizations may, at their own discretion, contribute voluntary uncommitted cost sharing to NSF-sponsored projects. However, these resources are not auditable by NSF and are not allowed to be included in the proposed budget or budget justification.

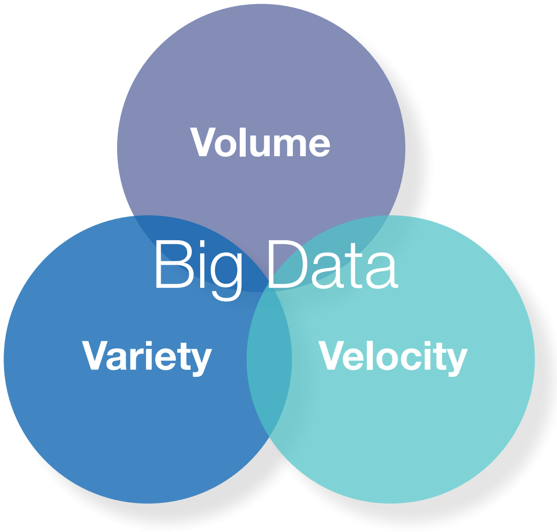
What is Big Data?

Big data refers to datasets that are large, complex and generated at high speed.
Pick up any newspaper, business magazine or scientific journal and you'll find a discussion of "Big Data." Organizations are using big data to create new products and generate insights into a wide range of phenomena. Applications are wide spread, including fraud detection, customer sentiment analysis, ad personalization, stock trading, drug discovery, health care delivery, energy efficiency, and management of computer and telecommunication networks.

While the precise etymology is unclear, the phrase "Big Data" appears to have been coined in the mid-1990s by researchers at Silicon Graphics International (SGI) to describe the rapidly increasing amount of data that organizations were handling.[43] Since then, the amount of data being collected, stored and processed has grown exponentially, driven, in part, by an explosion in web-based transactions, social media and sensor use.

IDC projects that the digital universe will reach 40 zettabytes (ZB) by 2020, an amount that exceeds previous forecasts by 5 ZBs, resulting in a 50-fold growth from the beginning of 2010.[44,45]

"Big Data" is neither a technology nor an industry; it is a term that applies to data that cannot be processed or analyzed using traditional techniques in a timely or cost-effective manner. Typically, Big Data is defined in terms of three characteristics of data streams:[46]



[43] http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf
[44] A petabyte is 10^15 or 1,000,000,000,000,000 bytes. A zettabyte is equal to 10^21 bytes.
[45] IDC/EMC Digital Universe Study (2012).
[46] Some commentators have added another two terms: veracity and value.

High Volume. Big data refers to massive datasets that are orders of magnitude larger than data managed in traditional databases. While the overall scale of data being collecting and stored is certainly impressive, the real issue is the amount of data handled by *individua*l organizations. A few statistics illustrate. Facebook has more than one billion active users with 150 billion friend connections.[47] Every bit of new content — news feeds, messages, events, photos and ads — is stored and tracked along with the massive amount of data contained in weblogs. More than 500 terabytes of new data are loaded into the company's databases every day with the largest Hadoop cluster capable of storing more than 100 petabytes. [48] The need to store and process massive amount of data is not limited to commercial concerns. For example, the Large Hadron Collider generates ~15 petabytes of data per year — equivalent to a CD stack roughly 20 km high.[49] Similarly, the planned Large Synoptic Survey Telescope will produce ~20 terabytes of data per night, resulting in 60 petabytes of raw data and a catalog database of 15 petabytes over ten years of operations. The total volume of data after processing will be on the order of several hundred petabytes.[50]

High Variety. The increase in volume has been accompanied by an increase in the complexity of data that organizations store and process. Up until recently, attention was focused on structured data, i.e., data that are neatly formatted based on a pre-defined formal schema (e.g., relational database). However, most data do not fit this description. A great deal of data is unstructured, including text, image, video, audio and sensor data. Semi-structured data, as the name implies, is a mix of structured and unstructured elements. This includes, for example, XML and other markup languages.

High Velocity. There are two aspects of the need for speed. The first centers on the ability to handle data as they arrive. While some data are generated periodically, others such as machine data are delivered in a constant stream. Taking the Large Hadron Collider as an example again, the 150 million sensors in the facility deliver data 40 million times per second. The second aspect relates to how fast data need to be processed. While processing historical data for business intelligence reporting or more in-depth analysis might need to be completed within minutes or hours, other tasks are more time sensitive. Certain types of transactions such as processing a trade or placing a targeted ad require the ability to process data in milliseconds.

The value of big data lies in their use.

There are five broad ways in which organizations can use big data to create value. First, organizations can use data to develop a better understanding of their customers and tailor product and services for narrowly defined segments. Second, organizations can use data to monitor performance of key functions, identifying factors contributing to observed variances and pointing to needed remedial actions or ways to optimize systems. Third, organizations can use data to predict behavior or forecast events, and as a result, take appropriate action. Fourth, organization can use data to meet regulatory compliance or legal discovery requirements. Finally, organization can use data as the building blocks for new products and services. These uses are found across virtually all industries as illustrated in *Table 24.*

[47] Annual Report, 2012.

[48] http://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day

[49] http://cds.cern.ch/record/1165534/files/CERN-Brochure-2009-003-Eng.pdf

[50] http://www.lsst.org/lsst/science/concept_data

TABLE 24) **Examples of Uses in Different Industries**

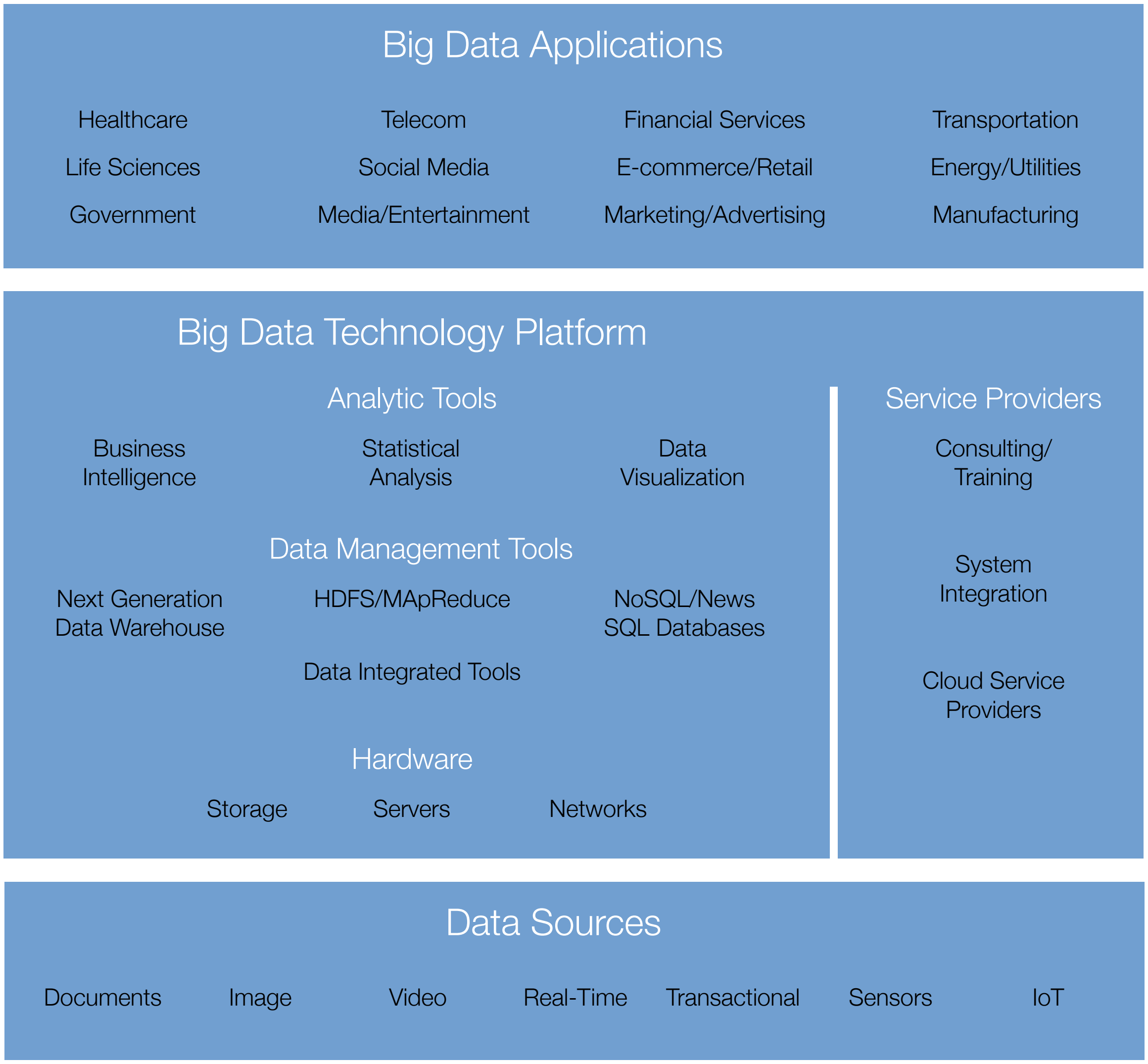
Financial Services	Marketing and Advertising	E-commerce / Retail Trade	Social Media
• Fraud detection and analysis • Credit risk scoring and analysis • Risk analysis and mitigation • Automated trading algorithms • Compliance and regulatory reporting • Legal discovery • Customer sentiment analysis • Targeting product offerings	• Customer sentiment analysis • Campaign analysis • Trading / pricing of ads • Personalized web content / emails • Ad targeting / serving	• Click stream analysis • Customer sentiment analysis • Analysis of buying behavior • Call center / log analysis • Point of sale transaction analysis • Development and application of pricing models • Personal web content delivery • Ad targeting / serving • Inventory management	• Ad targeting / serving • Customized content and promotion • Location based services
Media and Entertainment	**Telecommunication**	**Manufacturing**	**Transportation**
• Customer sentiment analysis • Content streaming • Search and recommendation optimization • Customized content and promotions • Ad targeting / serving	• Customer sentiment analysis • Analysis of buying behavior • Analysis of usage patterns • Call center / log analysis • Location-based services • Network analysis and optimization • Predictive maintenance	• Process control • Capacity utilization and forecasting • Supply chain analysis and management • Predictive maintenance • Analysis of warranty claims	• Location tracking • Capacity utilization and forecasting • Development and application of pricing models • Fuel consumption analysis • Predictive maintenance
Energy and Utilities	**Healthcare**	**Life Sciences**	**Government**
• Smart meter analytics • Compliance audits • Real-time demand forecast and pricing • Network analysis and optimization • Predictive maintenance	• Clinical trials in silico • Comparative effectiveness • Social media analysis to detect disease or treatment patterns • Capacity utilization and forecasting • Patient monitoring • Personalized medicine • Billing compliance	• Genomic sequencing • Drug discovery • Drug surveillance / monitoring	• Fraud detection and analysis • Threat analysis • Analysis of crime patterns • Weather forecasting • Cyber security

Source: Literature review

What is Big Data?

The market can be divided into three major segments.
The structure of the Big Data market is depicted in the figure below *(Figure 3)*. It consists of vendors of components of big data technology platforms, service big data applications providers, and developers of big data applications. The latter can be divided into two groups: companies that are developing commercial applications that are enabled by big data and companies that are building their own applications to use big data internally to run operations and inform decisions.

FIGURE 3) **Big Data Business Segments**



Big Data Applications

Healthcare
Life Sciences
Government

Telecom
Social Media
Media/Entertainment

Financial Services
E-commerce/Retail
Marketing/Advertising

Transportation
Energy/Utilities
Manufacturing

Big Data Technology Platform

Analytic Tools

Business Intelligence
Statistical Analysis
Data Visualization

Data Management Tools

Next Generation Data Warehouse
HDFS/MApReduce
NoSQL/News SQL Databases
Data Integrated Tools

Hardware

Storage
Servers
Networks

Service Providers

Consulting/ Training
System Integration
Cloud Service Providers

Data Sources

Documents
Image
Video
Real-Time
Transactional
Sensors
IoT

Source: Nexus Associates

A brief discussion of key business segment and enabling technologies follows:

- **Developers of Big Data Applications.** Numerous companies are developing new products enabled by the availability of data and development of new technology platforms. As discussed above, applications cut across a wide range of functions and industry verticals. Examples include customer sentiment analysis, analysis of buying behavior and churn, customized content delivery, ad targeting / serving, price optimization, location-based services, network analysis, and fraud detection and analysis.

- **Big Data Technology Platforms.** Big Data involves new technologies that enable the storage, processing and analysis of data. In general, a Big Data platform lets users store large amounts of structured and unstructured data in native format and process/analyze the data in parallel using server-class, commodity components. Customers are looking for integrated solutions.

- **Business Intelligence, Data Visualization and Analytic Software**. Analysis of big data draws on both business intelligence tools for reporting and advanced statistical techniques for data mining, machine learning, and predictive analysis. With respect to the later, it should be noted that SQL analytic functions can be used for various types of analyses, including full text search, funnel analysis, sentiment analysis, pattern matching and predictive modeling. That said, R – an open source statistical computing and graphing package – is frequently used to analyze large data sets. Revolution Analytics (CA) was founded in 2009 to support the R community and develop an enterprise version of the software. Systems packages offered by companies such as Oracle and Tibco also build on R. Major players such as IBM (SPSS), Mathworks, Oracle, SAS, SAP, and Tibco have all developed applications to enable users to draw on Hadoop and NoSQL databases. The same is true of vendors of data visualization tools such as Tableau Software. In this regard, Gartner expects that Hadoop will be embedded in roughly 65 percent of “prepackaged analytic applications with advanced analytics” by 2015.[51]

- **Data Management.** Big data is giving rise to the need for new data management techniques. Traditional relational database management systems (RDBMS) are hard pressed to keep up with the sheer amount and different types of data that need to be handled as well as the requirement for greater speed. In general, big data technology platform are built around the Apache Hadoop framework, noSQL/newSQL databases, and next generation data warehouses. These new approaches to data storage and processing are being used to complement rather than replace traditional methods. Moreover, the three technologies themselves are complementary. For example, next generation data warehouses can incorporate both Hadoop and NoSQL. Industry observers cite the need for users to adopt an appropriate set of technologies depending on the particular use, including databases that are optimized for specific (vertical) applications. Over time, the preferred deployment strategy is likely to center on a unified architecture.[52]

  Apache Hadoop. Managed by the Apache Software Foundation, Hadoop is an open source software framework for storing and processing large datasets across multiple computer clusters.[53] It was inspired by work originally done at Google in early 2000s. Hadoop is written in Java and is designed to run on commodity hardware, allowing scale-out at relatively low cost. The core components of Hadoop are the Hadoop Distributed File System (HDFS), which enables large

[51] Gartner. www.cio.com. January 24, 2013.

[52] Nexus Associates, interviews for 2014 Mass Big Data Report.

[53] The Hadoop Project Management Committee includes representatives of Cloudera, Facebook, Hortonworks, InMobi, Jive, LinkedIn, Microsoft, StumbleUpon, Twitter, WANdisco, and Yahoo! (http://hadoop.apache.org/who.html. Downloaded April 8, 2013.).

volumes of multi-structured data to be stored across multiple servers in a cluster with a high degree of redundancy, and Hadoop MapReduce, which enables parallel processing of data stored across servers in the Hadoop cluster.[54] Software releases are made available under the Apache 2.0 license.

A number of companies have developed commercial products based on Apache Hadoop, including three startups – Cloudera (CA), MapR Technologies (CA), and Hortonworks (CA). All three are contributing to the Apache Hadoop project, while also developing commercial enterprise distribution products based on different strategies. Cloudera was founded in 2009 by the initial developers of Hadoop. It is the current leader, focusing on creating a full Hadoop management suite and extending the framework to accommodate real-time analytics. MapR (CA) was founded in 2009 and opened for business in 2011. Given certain limitations of HDFS, MapR replaced it with its own proprietary file system. Hortonworks (CA) was spun-out of Yahoo in 2011. It is committed to a 100% open-source Hadoop distribution   Other vendors of Hadoop distributions include IBM (NY), Intel (CA) and Microsoft (WA). In addition, a number of companies such as Hadapt (MA) have taken the approach of integrating SQL and Hadoop into one platform. At this point, there is one common set of Hadoop APIs, enabling some degree of compatibility. All of these companies have entered into partnerships with vendors of complementary tools and systems.

NoSQL/NewSQL databases. NoSQL databases are highly scalable, non-relational databases (such as columnar, document, key-value, object and graph databases) designed to handle large volumes of data, particularly in applications requiring near real time processing. More than 150 NoSQL databases have been developed, including Accumulo, Aerospike. Cassandra, CouchDB, DynamoDB, HBase, MemcacheDB, MongoDB, Neo4J, Redis, and Riak. Most of these are open source. NewSQL is a type of RDBMS that seeks to provide the same scalable performance of NoSQL systems for online transaction processing while still maintaining ACID guarantees. Examples include Clustrix, NuoDB and Volt DB.

The last few years have seen the emergence of a relatively large number of firms aimed at offering NoSQL and NewSQL database software, including commercial versions of the open source software. These include companies such as 10gen (MongoDB), Aerospike (Aerospike), DataStax (Cassandra), Sqrrl (Accumulo), and Riak (Riak).

Next generation data warehouses. These warehouses are designed specifically to accommodate large volumes of data and provide near real time results in response to SQL queries. The fundamental characteristics of these warehouses center on using advanced data compression, columnar architectures, shared-nothing architectures, and massively parallel processing (MPP) capabilities deployed on commodity hardware. Some approaches make use of in-memory data processing.

[54] A number of related components have been developed to support or build on Hadoop, including Flume and Sqoop (enable users to collect data from multiple sources and integrate them into Hadoop), Hive (originally developed by Facebook, enables users to write SQL-like queries, which are then converted to MapReduce), Casandra and HBase (non-relational databases), HCatalog, Pig, Mahout, (data mining application implemented using MapReduce), Oozie, and Zookeeper.

This market has seen a fair amount of consolidation in recent years as leading vendors have acquired early stage companies with innovative technology. IBM (NY) acquired Neteeza (MA) in 2010; EMC (MA) purchased Greenplum (CA) in 2010; Hewlett Packard (CA) bought Vertica Systems (MA) in 2011; and Teradata (OH) acquired Aster Data System (CA) in 2011. These companies are positioning their products as complementary to Hadoop and NoSQL. For example, IBM has developed a platform – BigInsights – based on Apache Hadoop. This is packaged with various proprietary modules such as InfoSphere, Cognos BI tools, and SPSS analytical software. EMC Greenplum partnered with MapR to release a partly proprietary Hadoop distribution in May 2011. Teradata partnered with Hortonworks to integrate Hadoop into the Aster Discovery Platform. Oracle has also come out with data warehouse appliances that incorporates Cloudera's Hadoop distribution along with its own proprietary NoSQL database.

- **Data integration software.** Data integration involves retrieving and merging data from disparate data sources for specific uses. A number of tools have been developed for this purpose. Much of the activity has centered on tools for extracting, transforming, and loading data, typically into a data warehouse. Major vendors of enterprise ETL software, include IBM,[55] Informatica, Oracle, and SAS. Recent years have seen the emergence of a number of vendors such as Pentaho and Talend, which are offering products and services based on open source software. In addition to ETL, data integration also involves tools for data replication, data federation, data synchronization, and data cleaning. Driven by buyer demands, vendors are building capacities to provide comprehensive tool sets through internal development and/or acquisition of companies with complementary products.

- **Hardware.** The tremendous increase in the volume of data being generated is giving rise to a significant increase in demand for hardware to transfer, store and process data. EMC leads the data storage market, followed by NetApp, Hewlett Packard, and IBM.

Service Providers

- Cloud service providers. Big data platforms can be deployed in public clouds (as well as on-premise or in private clouds). Amazon Web Services (AWS) is arguably the most successful in this space. Amazon Elastic MapReduce is a web service that enables users to process vast amounts of data. It utilizes a hosted Hadoop framework running on Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3). Users have the option of using either Amazon's Hadoop distribution or MapR. In addition, with Amazon DynamoDB — a fully managed NoSQL database service — users can store data in solid-state stores. AWS also provides alternative NOSQL databases, including Cassandra and MongoDB. The company also offers a petabyte-scale data warehouse service — Amazon Redshift — enabling users to analyze their data using their own business intelligence tools. It is optimized for datasets ranging from a few hundred gigabytes to a petabyte or more and costs less than $1,000 per terabyte per year, a tenth the cost of most traditional data warehousing solutions.[56] Google launched a new service — BigQuery — and introduced MapR as a service via Google Compute Engine. Microsoft recently teamed with Hortonworks to offer the company's Hadoop distribution via Microsoft Azure. A raft of other companies is offering cloud-based DB as a service.

- **System integrators, consultants and support companies.** Numerous companies are offering services to businesses that are building big data capacities.

[55] In 2005, IBM acquired Ascential Software Corp (Westborough, MA).
[56] http://aws.amazon.com/redshift/

Appendix B

List of Interviews

The following leaders from industry, academia, and government participated in interviews in support of the 2014 Mass Big Data Report:

John Baker, *Founder,* DataKin & The Data Science Group

Justin Borgman, *Chief Executive Officer and Co-Founder,* Hadapt

Puneet Batra, *former Chief Data Scientist,* Kyruss

John Cardente, *Distinguished Engineer & Big Data Future Building Blocks Team Leader,* EMC

David Dietrich, *Advisory Technical Education Consultant,* Big Data & Data Science, EMC

Phil Francisco, *VP, Data Management Products & Strategy at IBM Information Management,* IBM

Leo Hermacinski, *CEO,* dSide Technologies

Ze Jiang, *CEO and Founder,* iQuartic

Bill Kiczuk, *VP & Chief Technology Officer,* Raytheon

Marilyn Kramer, *Deputy Executive Director,* Center for Health Information and Analysis

Dave Laverty, *Vice President Marketing,* Big Data & Analytics, IBM

Sam Madden, *Faculty Director,* BigData@CSAIL, MIT

Shawn Murphy, *Director of Research Information Systems,* Partners HealthCare

Steve Papa, *Founder and former CEO,* Endeca

Paul Sonderegger, *Big Data Strategist,* Oracle

Matthew Trunnell, *Chief Technology Officer,* Broad Institute

TABLE 25)

Segment	n	Percent
Hardware, including computers, servers, storage and networking equipment	5	8.9%
Data integration software, i.e., software to ingest, extract, transform and load data from multiple sources	29	51.8%
Data management software, including those built on RDBMS, Hadoop, NoSQL, and NewSQL.	25	44.6%
Business intelligence software	22	39.3%
Data visualization software	15	26.8%
Data analysis software	33	58.9%
Applications geared to specific verticals such as e-commerce, financial services, healthcare	24	42.9%
Systems integration	9	16.1%
Consulting / training	20	35.7%
Other	11	19.6%

TABLE 26)

Targeted Verticals	n	Percent
E-Commerce	11	22.5%
Education	10	20.4%
Energy	8	16.3%
Entertainment	8	16.3%
Financial services	19	38.8%
Healthcare	26	53.1%
Homeland Security/Defense	10	20.4%
Life sciences	17	34.7%
Manufacturing	9	18.4%
Social media	12	24.5%
Telecommunications	10	20.4%
Transportation	10	20.4%
Other	15	30.6%

List of Keywords and Counts

Segment	KW Classification	Keyword
A	1	Big data
B	2	Data Integration
C	3	Apache Accumulo
C	3	Cassandra
C	3	CouchDB
C	3	Google BigTable
C	3	Hbase
C	3	In-memory
C	3	MongoDB
C	3	NewSQL
C	3	NoSQL
C	4	Data warehouse
C	4	Hadoop
C	4	Hive
C	4	MapReduce
C	4	Massively parallel processing
D	5	Machine data
D	5	Machine-generated data
D	5	Machine-to-machine (M2M)
D	5	Unstructured data
E	6	Business analytics
E	6	Business intelligence
E	6	Data analytics (and analysis)
E	6	Enterprise analytics (and analysis)
E	6	Informatics
E	7	Advanced analytics (and analysis)
E	7	Data mining
E	7	Data science
E	7	Machine learning
E	7	Mahout
E	7	Predictive analytics (and analysis)
E	7	Real-time analytics (and analysis)
E	8	Natural language processing
E	8	Semantic analytics (and analysis)
E	8	Text mining
E	9	Sentiment analytics (and analysis)
E	9	Social media analytics (and analysis)
E	10	Geo-spatial analytics (and analysis)
E	10	Location analytics (and analysis)
E	11	Facial recognition
E	11	Image analytics (and analysis)
E	11	Video analytics (and analysis)
E	12	Bioinformatics
E	12	Data visualization

Segment	KW Classification	Keyword
E	13	Web analytics (and analysis)
E	14	Network analytics (and analysis)
E	15	Social CRM
E	15	Social network analytics (and analysis)
F	16	Fraud detection
F	16	Risk analysis
F	16	Threat detection
F	17	Cybersecurity
F	18	Automated trading
F	19	Ad serving
F	19	Ad targeting
F	20	Genome (and genomic) sequencing

TABLE 27) **Keyword Counts**

	US	BOS	CHI	DCA	DFW	NYC	PHL	RDH	LAX	SAN	SFO	SEA
A = Big Data	**1,496**	**105**	**51**	**144**	**33**	**41**	**215**	**29**	**17**	**23**	**417**	**77**
B = Data Integration	**673**	**36**	**37**	**67**	**20**	**28**	**72**	**23**	**9**	**14**	**78**	**10**
C = Data Management	**1,361**	**65**	**63**	**80**	**42**	**59**	**150**	**36**	**13**	**16**	**274**	**44**
2 = noSQL/SQL	453	27	20	21	6	16	54	14	4	5	99	13
3 = Hadoop	347	13	10	19	9	18	36	6	4	6	93	17
4 = Data Warehouse	561	25	33	40	27	25	60	16	5	5	82	14
D = Data	**239**	**21**	**15**	**17**	**10**	**4**	**30**	**4**	**4**	**7**	**43**	**10**
E = Data analysis and visualization	**11,344**	**346**	**589**	**926**	**371**	**403**	**1,500**	**353**	**117**	**232**	**1,347**	**377**
6 = BI and business analytics	5,758	139	329	478	255	215	714	207	70	99	566	219
7 = data mining and analysis	3,226	87	147	282	73	98	430	93	28	66	357	92
8 = data science, machine learning / predictive analytics	1,331	73	78	73	28	51	223	29	9	42	272	37
9 = semantic analysis	204	9	5	23	5	9	37	5	3	2	40	7
10 = geo-spatial analysis	70	2	1	7	1	1	6	1	1	1	6	2
11 = i mage analysis	322	12	10	16	5	16	32	6	3	9	43	4
12 = Data visualization	433	24	19	47	4	13	58	12	3	13	63	16
F = Selected applications	**2,837**	**95**	**141**	**280**	**60**	**137**	**392**	**53**	**44**	**71**	**338**	**75**
13 = sentiment and social media analysis	489	22	22	34	11	22	85	5	9	7	101	13
14 = bioinformatics and genomics	249	21	1	29	5	8	27	2	7	13	45	10
15 = web analytics	796	19	51	33	14	45	81	17	15	17	71	29
16 = network analytics	148	7	7	11	4	4	16	1	1	2	17	2
17 = fraud, threat and risk detection	726	19	28	79	22	30	84	25	8	20	56	15
18 = cybersecurity	215	3	3	91	1	7	8	1	1	11	12	3
19 = automated trading	71	-	21	1	1	2	32	2	1	-	4	2
20 = ad targeting / serving	143	4	8	2	2	19	59	-	2	1	32	1
Total	**17,950**	**668**	**896**	**1,514**	**536**	**672**	**2,359**	**498**	**204**	**363**	**2,497**	**593**

TABLE 28) Number of Graduates in Massachusetts and Competing States (2012)

CIP	Title	US	MA	CA	IL	NC	NY	TX	WA
11	Computer and Information Sciences	72,678	2,479	6,749	4,341	1,861	5,754	3,886	1,092
14.09	Computer Engineering	8,510	306	1,482	196	219	511	502	110
27	Mathematics and Statistics	30,053	1,229	3,304	1,595	1,021	3,204	1,819	656
40.08	Physics	9,801	594	1,123	421	289	828	513	238
26.0203, 26.0206	Biophysics and Molecular Biophysics	276	22	51	31	4	39	18	-
40.0202, 40.0403, 40.0603	Astrophysics, Atmospheric Physics and Dynamics, and Geophysics and Seismology	467	29	108	15	-	11	86	15
26.11	Biomathematics, Bioinformatics, and Computational Biology	1,329	105	132	27	71	138	53	15
51.2706	Medical Informatics	411	25	36	75	-	7	25	7
45.0603	Econometrics and Quantitative Economics	436	37	129	-	17	20	45	-
14.37	Operations Research	1,177	27	217	25	18	595	56	-
52.12	Management Information Systems and Services	12,333	140	177	662	223	467	1,085	302
52.13	Management Sciences and Quantitative Methods	6,679	530	415	1,647	40	106	382	38
30.06	Systems Science and Theory	470	32	4	70	15	62	-	32
30.08	Mathematics and Computer Science	242	28	44	51	1	5	4	4
30.16	Accounting and Computer Science	15	-	-	-	-	-	4	-
30.3	Computational Science	37	-	4	-	-	-	7	-
	Total	**144,914**	**5,583**	**13,975**	**9,156**	**3,779**	**11,747**	**8,485**	**2,509**

Source: National Center for Education Statistics (NCES)

TABLE 29) **Graduates from MA Colleges and Universities (2012)**

CIP	Title	Number of Schools	Number of Graduates				
			BA/BS	MA/MS	PhD	Certificate (a)	Total
11	**Computer and Information Sciences**	**47**	**1,441**	**934**	**93**	**11**	**2,479**
	Amherst College		13	0	0	0	13
	Assumption College		4	0	0	0	4
	Bard College at Simon's Rock		3	0	0	0	3
	Bentley University		43	57	0	0	100
	Boston College		72	0	0	0	72
	Boston University		44	285	4	0	333
	Brandeis University		26	111	1	7	145
	Bridgewater State University		19	11	0	0	30
	Clark University		6	13	0	0	19
	College of Our Lady of the Elms		8	0	0	0	8
	Curry College		15	0	0	0	15
	Eastern Nazarene College		1	0	0	0	1
	Endicott College		8	4	0	0	12
	State University		25	24	0	0	49
	Framingham State University		15	0	0	0	15
	Gordon College		9	0	0	0	9
	Hampshire College		12	0	0	0	12
	Harvard University		37	10	7	1	54
	ITT Technical Institute-Norwood		17	0	0	0	17
	Technical Institute-Wilmington		11	0	0	0	11
	Massachusetts College of Liberal Arts		8	0	0	0	8
	Massachusetts Institute of Technology		174	105	43	0	322
	Merrimack College		12	0	0	0	12
	Mount Holyoke College		5	0	0	0	5
	Northeastern University		125	162	7	0	294
	Regis College		1	0	0	0	1
	Salem State University		13	0	0	0	13
	Simmons College			0	2	0	2
	Smith College		12	0	0	0	12
	Springfield College		2	0	0	0	2
	Stonehill College		5	0	0	0	5
	Suffolk University		29	11	0	0	40
	The New England Institute of Art		48	0	0	0	48
	Tufts University		44	15	3	0	62
	University of Massachusetts-Amherst		77	21	14	0	112
	University of Massachusetts-Boston		53	25	2	0	80
	University of Massachusetts-Dartmouth		16	10	0	0	26
	University of Massachusetts-Lowell		182	42	3	3	230
	University of Pheonix-Boston Campus		7	0	0	0	7
	Wellesley Collegee		10	0	0	0	10

TABLE 29 CONTINUED) Graduates from MA Colleges and Universities (2012)

CIP	Title	Number of Schools	Number of Graduates				
			BA/BS	MA/MS	PhD	Certificate (a)	Total
11	**Computer and Information Sciences (cont.)**						
	Massachusetts College of Liberal Arts		8	0	0	0	8
	Massachusetts Institute of Technology		174	105	43	0	322
	Merrimack College		12	0	0	0	12
	Mount Holyoke College		5	0	0	0	5
	Northeastern University		125	162	7	0	294
	Regis College		1	0	0	0	1
	Salem State University		13	0	0	0	13
	Simmons College			0	2	0	2
	Smith College		12	0	0	0	12
	Springfield College		2	0	0	0	2
	Stonehill College		5	0	0	0	5
	Suffolk University		29	11	0	0	40
	The New England Institute of Art		48	0	0	0	48
	Tufts University		44	15	3	0	62
	University of Massachusetts-Amherst		77	21	14	0	112
	University of Massachusetts-Boston		53	25	2	0	80
	University of Massachusetts-Dartmouth		16	10	0	0	26
	University of Massachusetts-Lowell		182	42	3	3	230
	University of Pheonix-Boston Campus		7	0	0	0	7
	Wellesley Collegee		10	0	0	0	10
	Wentworth Institute of Technology		76	0	0	0	76
	Western New England University		16	0	0	0	16
	Westfield State University		24	0	0	0	24
	Wheaton College		11	0	0	0	11
	Williams College		11	0	0	0	11
	Worcester Polytechnic Institute		69	28	7	0	104
	Worcester State University		24	0	0	0	24
14.09	**Computer Engineering**	**7**	**137**	**161**	**8**	**0**	**306**
	Boston University		22	36	2	0	60
	Eastern Nazarene College		1	0	0	0	1
	Northeastern University		35	101	4	0	140
	Tufts University		9	0	0	0	9
	University of Massachusetts-Amherst		33	0	0	0	33
	University of Massachusetts-Dartmouth		16	10	2	0	28
	University of Massachusetts-Lowell		21	14	0	0	28
27	**Mathematics and Statistics**	**45**	**1,014**	**133**	**65**	**17**	**1,229**
	Amherst College		22	0	0	0	22
	Assumption College		16	0	0	0	16
	Bard College at Simon's Rock		2	0	0	0	2
	Bentley University		212	0	0	0	212

TABLE 29 CONTINUED) Graduates from MA Colleges and Universities (2012)

CIP	Title	Number of Schools	Number of Graduates				
			BA/BS	MA/MS	PhD	Certificate (a)	Total
14.09	**Mathematics and Statistics (cont.)**						
	Boston College		57	2	0	0	59
	Boston University		40	7	7	0	54
	Brandeis University		41	2	7	3	53
	Bridgewater State University		45	0	0	0	45
	Clark University		16	0	0	0	16
	College of Our Lady of the Elms		3	0	0	0	3
	College of the Holy Cross		45	0	0	0	45
	Eastern Nazarene College		5	0	0	0	5
	Emmanuel College		4	0	0	0	4
	Fitchburg State University		6	0	0	0	6
	Framingham State University		10	0	0	0	10
	Gordon College		4	0	0	0	4
	Hampshire College		2	0	0	0	2
	Harvard University		109	38	18	0	165
	Lasell College		1	0	0	0	1
	Lesley University		13	0	0	0	13
	Massachusetts College of Liberal Arts		5	0	0	0	5
	Massachusetts Institute of Technology		96	0	22	0	118
	Merrimack College		3	0	0	0	3
	Mount Holyoke College		25	0	0	0	25
	Nichols College		2	0	0	0	2
	Northeastern University		31	13	3	0	47
	Salem State University		6	10	0	0	16
	Simmons College		7	0	0	0	7
	Smith College		17	0	0	14	31
	Springfield College		2	0	0	0	2
	Stonehill College		11	0	0	0	11
	Suffolk University		3	0	0	0	3
	Tufts University		25	6	1	0	32
	University of Massachusetts-Amherst		102	17	6	0	125
	University of Massachusetts-Boston		11	0	0	0	11
	University of Massachusetts-Dartmouth		113	0	0	0	113
	University of Massachusetts-Lowell		27	13	0	0	40
	Wellesley College		20	0	0	0	20
	Western New England University		2	0	0	0	2
	Westfield State University		20	0	0	0	20
	Wheaton College		7	0	0	0	7
	Wheelock College		13	0	0	0	13
	Williams College		53	0	0	0	53
	Worcester Polytechnic Institute		42	25	1	0	68
	Worcester State University		9	0	0	0	9

TABLE 29 CONTINUED) **Graduates from MA Colleges and Universities (2012)**

CIP	Title	Number of Schools	Number of Graduates				
			BA/BS	MA/MS	PhD	Certificate (a)	Total
40.08	**Physics**	**27**	**336**	**106**	**152**	**0**	**594**
	Amherst College		6	0	0	0	6
	Boston College		16	2	7	0	25
	Boston University		17	22	17	0	56
	Brandeis University		14	3	4	0	21
	Bridgewater State University		6	0	0	0	6
	Clark University		6	1	2	0	9
	College of the Holy Cross		10	0	0	0	10
	Gordon College		4	0	0	0	4
	Hampshire College		3	0	0	0	3
	Harvard University		51	19	57	0	127
	Massachusetts College of Liberal Arts		4	0	0	0	4
	Massachusetts Institute of Technology		83	2	37	0	122
	Merrimack College		3	0	0	0	3
	Mount Holyoke College		8	0	0	0	8
	Northeastern University		13	13	5	0	31
	Simmons College		1	0	0	0	1
	Smith College		5	0	0	0	5
	Stonehill College		3	0	0	0	3
	Tufts University		5	5	3	0	13
	University of Massachusetts-Amherst		25	5	8	0	38
	University of Massachusetts-Boston		3	6	0	0	9
	University of Massachusetts-Dartmouth		7	7	0	0	14
	University of Massachusetts-Lowell		7	17	9	0	33
	Wellesley College		3	0	0	0	3
	Wheaton College		8	0	0	0	8
	Williams College		14	0	0	0	14
	Worcester Polytechnic Institute		11	4	3	0	18
26.0203, 26.0206	**Biophysics and Molecular Biophysics**	**4**	**3**	**1**	**18**	**0**	**22**
	Boston University		0	1	5	0	6
	Brandeis University		1	0	1	0	2
	Harvard University		0	0	12	0	12
	Northeastern University		2	0		0	2
40.0202, 40.0403, 40.0603	**Astrophysics, Atmospheric Physics and Dynamics, and Geophysics and Seismology**	**7**	**24**	**3**	**2**	**0**	**29**
	Boston College		1	0	0	0	1
	Boston University		2	0	0	0	2
	Harvard University		14	0	0	0	14
	Massachusetts Institute of Technology		0	3	2	0	5
	Smith College		1	0	0	0	1
	Wellesley College		2	0	0	0	2
	Williams College		4	0	0	0	4

TABLE 29 CONTINUED) Graduates from MA Colleges and Universities (2012)

CIP	Title	Number of Schools	Number of Graduates				
			BA/BS	MA/MS	PhD	Certificate (a)	Total
26.11	**Biomathematics, Bioinformatics, and Computational Biology**	**7**	**6**	**57**	**41**	**1**	**105**
	Boston University		0	32	20	0	52
	Brandeis University		0	3	0	1	4
	Emmanuel College		1	0	0	0	1
	Harvard University		0	11	17	0	28
	Massachusetts Institute of Technology		1	1	4	0	6
	Northeastern University		0	10	0	0	10
	Simmons College		4	0	0	0	4
51.2706	**Medical Informatics**	**2**	**0**	**25**	**0**	**0**	**25**
	Brandeis University		0	2	0	0	2
	Northeastern University		0	23	0	0	23
45.0603	**Econometrics and Quantitative Economics**	**1**	**37**	**0**	**0**	**0**	**37**
	Tufts University		37	0	0	0	37
14.37	**Operations Research**	**2**	**0**	**17**	**10**	**0**	**27**
	Massachusetts Institute of Technology		0	5	10	0	15
	Northeastern University		0	12	0	0	12
52.12	**Management Information Systems and Services**	**11**	**64**	**76**	**0**	**0**	**140**
	American International College		1	0	0	0	1
	Bay Path College		0	27	0	0	27
	Bentley University		0	3	0	0	3
	Boston University		0	29	0	0	29
	Fisher College		5	0	0	0	5
	Framingham State University		14	0	0	0	14
	Nichols College		3	0	0	0	3
	Salem State University		6	0	0	0	6
	University of Massachusetts-Dartmouth		27	0	0	0	27
	Western New England University		3	0	0	0	3
	Worcester Polytechnic Institute		5	17	0	0	22
52.13	**Management Sciences and Quantitative Methods**	**7**	**219**	**280**	**0**	**31**	**530**
	American International College		11	0	0	0	11
	Bentley University		0	21	0	31	52
	Boston University		0	46	0	0	46
	Bridgewater State University		190	22	0	0	212
	Lasell College		0	39	0	0	39
	Northeastern University		0	151	0	0	151
	Suffolk University		18	1	0	0	19
30.06	**Systems Science and Theory**	**2**	**1**	**31**	**0**	**0**	**32**
	Boston University		0	28	0	0	28
	Worcester Polytechnic Institute		1	3	0	0	4
30.08	**Mathematics and Computer Science**	**2**	**28**	**0**	**0**	**0**	**28**
	Massachusetts Institute of Technology		26	0	0	0	26
	Springfield College		2	0	0	0	2
Total			**3,310**	**1,824**	**389**	**60**	**5,583**

University Data Science Degree Program Descriptions:

Bentley University, Graduate Certificate in Business Analytics

The Certificate in Business Analytics is intended to provide students with a solid grounding in applied statistical methods with an emphasis on the use of appropriate software tools. Courses provide students the opportunity to see how these methods are currently used in different business areas. Internships are offered as an option (See course requirements, *Appendix F, Table 45, "Bentley Required Courses for Graduate Certificate in Business Analytics")*.

Boston University, Master of Science in Systems Engineering[57,58]

Systems Engineering is a cross-disciplinary program, offered by the College of Engineering in cooperation with faculty from the Graduate School of Arts & Sciences and the School of Management. The program integrates courses from Engineering, Computer Science, Mathematics, and Management *(See course requirements in Appendix F, Table 41, "Boston University Course Requirements Master of Science in Systems Engineering")*. The coursework requirements for the MS degree include three core courses, two courses in one of the concentration areas, and a thesis or graduate project. In addition, students are required to fulfill a practicum requirement of their program through either (a) internship or employment in industry, government, or non-profit organization or (b) satisfactory completion of a project-based graduate course approved by the division.

Harvard University, Master of Science in Computational Science and Engineering

The Harvard University School of Engineering and Applied Science (SEAS) established the Institute for Applied Computational Science (IACS) in September 2010. IACS is responsible for "launching a unique interdisciplinary education and research program in computational science and engineering (CSE)." A new one-year master's degree program in Computational Science and Engineering is starting in fall 2013 and a two-year master's program is slated to begin in 2014. The course of study was developed with input from industry and national labs as well as from Harvard faculty. As noted in the program description, graduates are expected to be able to do the following: i) produce a computational solution to a problem that is reproducible and can be comprehended by others in the same field; ii) communicate across disciplines and collaborate in a team; iii) model complex systems appropriately with consideration of efficiency, cost, and data availability; iv) use computation for advanced data analysis; v) create or enable a breakthrough in a domain in science; vi) take advantage of parallel and distributed computing and other emerging modes of computation, both in algorithms and in code implementation; vii) evaluate and compare multiple computational approaches to a scientific challenge and choose the most appropriate and efficient one; and viii) apply techniques and tools from software engineering to build robust, reliable, and maintainable software *(See course requirements, Appendix F, Table 46, "Harvard Required Courses for MS in Computational Science and Engineering")*.

Massachusetts Institute of Technology, Master of Science in Operations Research[59,60]

The master's degree (SM) program prepares graduates for a professional career that usually involves applications of operations research. In addition to course requirements, students must demonstrate computer literacy and proficiency in English. In addition, degree candidates are required to write and present a thesis based on independent, usually applied, research *(See course requirements, Appendix F, Table 40, MIT Course*

[57] The Systems Engineering program at BU offers an undergraduate minor as well as ME, MS and PhD degrees.
[58] Worcester Polytechnic Institute offers a BS and MS in System Dynamics.
[59] MIT also offers a PhD in Operation Research.
[60] Northeastern offers an MS degree in Operation Research

Requirements – MS Operations Research). Massachusetts Institute of Technology, PhD in Computational and Systems Biology. As states in the course description, "The emerging field of systems biology represents an integration of concepts and ideas from the biological sciences, engineering disciplines, and computer science… Recent advances in biology, including the Human Genome Project and massively parallel approaches to probing biological samples, have created new opportunities to understand biological problems from a systems perspective. Systems modeling and design are well established in engineering disciplines but are relatively new to biology… Spanning the School of Engineering and the School of Science, [doctorate] program integrates coursework and research opportunities in biology, engineering, mathematics, microsystems, and computer science with interdisciplinary courses in computational and systems biology ..." *(See course requirements, Appendix F, Table 43, "MIT Required Courses PhD in Computational and Systems ")*.

Northeastern University, Master of Science in Health Informatics[61,62]
The health informatics program is multidisciplinary, drawing on the College of Computer and Information Science and Bouvé College of Health Sciences. The aim is for students to learn how to use information technology and information management concepts and methods in healthcare delivery. Graduates may assume roles in a wide range of health-related organizations, including hospitals, physician groups, HMOs, software companies, pharmaceutical and biotech companies, clinical research organizations, and government agencies *(See course requirements Appendix F, Table 39 "Northeastern University Course Requirements Master of Science in Health Informatics")*.

Worcester Polytechnic Institute (WPI), Master of Science in Data Science
In November, 2013, WPI's announced it would offer an interdisciplinary Master's Degree in Data Science as part of a joint program though faculty members at its the Departments of Computer Science, Mathematical Sciences, and the School of Business. The program is planned to emphasize development of a cross-disciplinary technical and scientific background, with specific focus on areas including machine learning, statistical modelling, data warehousing, predictive modeling, and large-scale database architecture and management. The program is intended to prepare students to apply and advance state-of-the-art data analytic tools and methods (including data mining, big data algorithms, and data visualization) in order to develop transformative solutions to problems across a range of domains; to use the knowledge and skills they gain in analytics, computing, statistics, and business intelligence to understand and explain their results and their applicability and validity; and to serve as visionary leaders and project managers in data analytics. The program offers both a master of science degree or completion of a graduate certificate in data science. The degree can also be completed as part of a combined five-year BS/MS program. A graduate certificate can be earned by completing 18 credits of relevant graduate coursework; the credits can later be applied to complete the master of science degree.

It is also worth noting that Harvard University began offering a data science course in the statistics department in Spring 2013 – Statistics 221. Statistical Computing and Visualization. The course "emphasizes rigorous methods for the full cycle of a typical data-intensive problem solution, including defining the problem within a context, designing a method to solve it, evaluating its properties, implementing it, communicating the findings, and generalizing to a product or a statistical method." It covers current theory and philosophy of building models for data, computational methods, and tools such as d3js, parallel computing with MPI, and R. Students are required to complete an industry-sponsored project.

[61] In addition to the MS in Health Informatics, Northeastern offers a Ph.D. in Personal Health Informatics, Advanced Standing Science in Health Informatics, Graduate Certificate in Health Informatics Management and Exchange, Graduate Certificate in Health Informatics Privacy and Security, and a Graduate Certificate in Health Informatics Software Engineering.

[62] Brandies University also offers a MS in Health and Medical Informatics.

Referenced Tables

TABLE 31) Age of Companies

Company Age	Number of Companies	Percent of Companies
1-3 years	77	13.9%
4-10 years	220	39.9%
11-20 years	166	30.1%
21 or more years	69	12.5%
N/A	20	3.6%
Total	**552**	**100%**

Source: Crunchbase, LinkedIn and company websites

TABLE 32) Recent Acquisitions of Companies in Massachusetts

Company	Founded	Product	Acquired By	Date
Crashlytics	2011	Software for Big data crash analysis	Twitter	2013
Spindle	2010	Mobile application	Twitter	2013
BlueFin Labs	2008	Social analytics for TV	Twitter	2013
Trusteer	2006	Provider of endpoint cybercrime prevention solutions	IBM	2013
Jumptap	2005	Provider of mobile advertising solutions	Millennial Media	2013
StreamBase Systems, Inc.	2003	Server platform software	TIBCO Software	2013
OrderMotion	1994	Media tracking and analytics tool	Netsuite	2013
Humedica, Inc.	1979	Bioinformatics analysis	United Health	2013
Expressor Software	2007	Data integration software	QlikTech	2012
Digital Reef	2006	Software for eDiscovery services and digital information governance applications	TransPerfect	2012
Vela Systems	2005	Provides of a web-based platform for all field and management users	Autodesk	2012
Memento	2002	Enterprise fraud and compliance platform	FIS	2012
Dataspora	2008	Consultantancy focusing on predictive analysis	Via Science	2011
Vlingo	2006	Provider of voice to text technology and natural language processing software	Nuance Communications	2011
Akorri	2005	storage and virtualization management software	NetApp	2011
Tatto Media	2005	advertising targeting platform	Ozura World	2011
Vertica Systems Inc	2005	Real-time analytics database	Hewlett Packard	2011
SensorLogic	2002	Cloud M2M service	Gemalto	2011
Endeca	1999	Data management, web commerce and business intelligence software, enabling enterprises to analyze semi-structured and unstructured data	Oracle	2011
Oco, Inc	1999	Enterprise-business analytics software as a service (SaaS)	Deloitte	2011
Navisite	1998	Cloud enabled enterprise hosting and application management services	Time Warner Cable	2011
geoVue	1996	Provider of location-based decision support systems	Vesata Enterprises	2011
Art Technology Group	1991	eCommerce software and related on-demand commerce optimization applications.	Oracle	2011
CambridgeSoft	1986	Provider of software and services	PerkinElmer	2011
Blackwave	2006	Platform for the storage and delivery of video over IP	Juniper Networks (CA)	2010
Quattro Wireless	2006	Mobile advertising company	Apple	2010
Netezza	2000	High-performance data warehouse appliances and advanced analytics applications.	IBM (NY)	2010
Vaultus Mobile	2000	Mobile application platform	Antenna Software (NJ)	2010
Phase Forward, Inc.	1997	Data management software for clinical trials and drug safety	Oracle (CA)	2010
Unica	1992	Provider of cloud-based marketing software	IBM	2010
Card.ly		Internet provider of online mini business card	Workface	2010
Agencyport Software	2000	Provider of web technologies and robust business intelligence tools	Sword	2009
Ibrix	2000	File serving software for cluster, grid, and enterprise computing environments	Hewlett Packard (CA)	2009
Mazu Networks, Inc.	2000	Network behavior analysis software	Riverbed Technology (SP, BRAZIL)	2009
DuPont Photonics Tech.	2003	Computer hardware provider	Enablence Technologies	2008
Compete	2000	Consumer behavior data for channel optimization and media effectiveness	Taylor Nelson Sofres (UK)	2008
Navic Networks	2000	Television networks tools that use real-time data to optimize delivery and placement of targeted video ads	Microsoft (WA)	2008
SecureMedia,	1996	digital content security software	MediaXstream (NJ)	2008
AnchorPoint	1984	Telecommunications Management (TM) solutions provider	MTS	2008
Azima DLI	1966	machine condition monitoring and assessment software	Azima	2008
Applix	1983	MOLAP database server and related presentation tools	Cognos (CAN) (a)	2007

Notes: (a) Cognos was subsequently acquired by IBM in 2008.

Source: Crunchbase, company websites and news reports.

TABLE 33) Recent Acquisitions of Companies by EMC

Company	Location	Date Founded	Product / Services	Acquisition Date
Likewise Software	Bellevue, WA	2004	Authentication software	2012
Pivotal Labs	San Francisco, CA	1989	Full-service software development	2012
Silicium Security	Vaudreuil, CAN	1999	Enterprise security software to detect malware	2012
Silver Tail Systems	Menlo Park, CA	2008	Predictive analytics software to prevent fraud and abuse on websites	2012
Syncplicity	Menlo Park, CA	2008	File synchronization, back-up, and sharing software	2012
Watch4Net	Montreal, CAN	2000	Service assurance software for networks	2012
XtremIO	Cupertino, CA Herzelyia, ISR	2009	Solid state storage devices	2012
NetWitness	Herndon, VA	2006	Network security monitoring software	2011
Bus-Tech, Inc.	Bedford, MA	2007	Mainframe networking and storage products for data center connectivity applications	2010
Greenplum	San Mateo, CA	2003	Database software for BI and data warehousing applications	2010
Isilon Systems	Seattle, WA	2001	Enterprise data storage	2010

Source: Crunchbase, company websites and news reports.

TABLE 34) Largest Hospital Systems in Massachusetts by Size and Revenue

Name	Owns and Operates	Total Beds	Significant Contractual and Financial Relationships	Hospital Net Patient Service Revenue ($million)
Partners Healthcare	7 hospitals (a) 4 medical groups 5 community health centers	2,682	11 medical groups 3 community health centers	4,360
UMass Memorial HealthCare	5 hospitals 1 community health center	1,038	6 medical groups 2 community health centers	1,660
Steward Health Care System	10 hospitals	1,763	24 medical groups 7 community health centers	1,386
Beth Israel Deaconess Medical Center	3 hospitals 1 medical group 1 community health center	759	12 medical groups 6 community health centers	1,152
Children's Hospital Boston	1 hospital 1 community health center	384	1 medical group	1,016
Baystate Health	3 hospitals, 1 medical group 2 community health centers	831	22 medical groups 1 commnunity health center	922

Notes: (a) including Massachusetts General Hospital, Brigham and Women’s Hospital , and Dana-Farber Cancer Institute

Source: bluecrossmafoundation.org/delivery-system-map/report/1254

TABLE 35) Fortune 500 Companies Headquartered in Massachusetts

State Rank	Fortune Rank	Company	City	Industry	Revenues ($millions)
1	84	Liberty Mutual Insurance Group	Boston	Insurance: property and casualty	34,671
2	114	Staples	Framingham	Specialty retailer	25,022
3	117	Raytheon	Waltham	Aerospace and defense	24,857
4	121	Massachusetts Mutual Life Insurance	Springfield	Insurance: life and health	24,226
5	125	TJX	Framingham	Specialty retailer: apparel	23,192
6	139	EMC	Hopkinton	Computer peripherals	20,008
7	182	Global Partners	Waltham	Wholesalers: diversified	14,836
8	225	Thermo Fisher Scientific	Waltham	Scientific, Photographic, and Control Equip.	11,780
9	262	State Street Corp.	Boston	Commercial bank	10,207
10	335	Boston Scientific	Natick	Medical Products and Equipment	7,622
11	476	Biogen Idec	Weston	Pharmaceuticals	5,049

TABLE 36) Selected Research Centers in Massachusetts

Institution	Center	Founded	Description
Boston University	Center for Computational Science	1990	Focuses on efforts to coordinate and promote computationally based research, foster computational science education, and provide for the expansion of computational resources and support.
	Center for Reliable Information Systems and Cyber Security	2005	A DOD designated National Center of Academic Excellence in Information Assurance Education. Established to promote and coordinate research on reliable and secure computation and on information assurance education by providing increased opportunities for collaboration among researchers from cognate fields.
	Rafik B. Hariri Inst. for Computing and Computational Science & Engineering	2011	Aims to initiate, catalyze, and propel collaborative, interdisciplinary research and training initiatives for the betterment of society by promoting discovery and innovations through the use of computational and data-driven approaches, and by supporting advances in the science of computing inspired by challenges in arts and sciences, engineering, and management disciplines.
Broad Institute	n/a	2003	A large-scale scientific collaboration in genomics and chemical biology that grew out of major initiatives at Harvard and MIT. Collectively, these projects aim to assemble a complete picture of the molecular components of life, define the biological circuits that underlie cellular responses, uncover the molecular basis of major inherited diseases, unearth all the mutations that underlie different cancer types, discover the molecular basis of major infectious diseases, and transform the process of therapeutic discovery and development.
Dana Farber Cancer Institute	Center for Cancer Computational Biology	2009	Focuses on genomic and computational biology approaches that open new ways of understanding cancer by improving analysis and interpretation of genomic data through integration with information derived from other sources, including publicly available data. Also supports analysis and interpretation of genomic and other large-scale data to further basic, clinical, and translational research.
Harvard	Center for Research on Computation and Society	2002	Focuses on development of new ideas and technologies designed to address fundamental computational problems arising from societal issues, such as privacy, security, and crowdsourcing. The Harvard Center for Research in Computation and Society's integrative approach combines research on computer science and technology informed by societal events to reach their research goals.
	Center for Systems Biology	1999	Overall goal is to find general principles that help explain the structure, behavior, and evolution of cells and organisms.
	Institute for Quantitative Social Science:	2005	Focuses on quantitative research in the social sciences across many disciplines.
MIT	Computer Science and Artificial Intelligence Lab (a)	2003	Focuses on artificial intelligence, systems, and theory. Goal is to apply knowledge on human intelligence, extending functional capabilities of machines, human/machine interactions to engineer solutions with global impact. In 2012, MIT was selected by Intel to host a new research center focusing on big data. This was subsumed under Bigdata@CSAIL. It focuses on identifying and developing technologies to solve the next generation data challenges.
	Media Lab	1985	Focuses on efforts that combine seemingly disparate research areas to uncover ways to radically improve the way people live, learn, express themselves, work and play.
	Operations Research Center	1953	Aims to apply advanced analytical methods to help make better decisions. The center's research activities cover both methodological research (i.e. mathematical programming and combinatorial optimization, cluster analysis, network design) and application domains (i.e. flexible manufacturing systems, air traffic control, epidemiology).

TABLE 36 CONTINUED) **Selected Research Centers in Massachusetts**

Institution	Center	Founded	Description
Northeastern	Center for Complex Network Research	2004	How networks emerge, what they look like, and how they evolve; and how networks impact on understanding of complex systems.
	Center for Interdisciplinary Research on Complex Systems	1995	Aims at elucidating fundamental aspects of the structure and function of complex physical and biological systems across multiple levels of organization using a combination of quantitative state-of-the-art experimental and theoretical research tools. Ongoing research projects span biomolecular systems, physiological systems from neuroscience to cardiac nonlinear dynamics, nanosystems from nanomaterials design to nanotribology, and complex interfacial systems in materials science from microstructural pattern formation in alloys to crystal decohesion and crack propagation.
	Institute for Information Assurance	2005	An NSA/Department of Homeland Security Center of Academic Excellence in Information Assurance Research and Education. Examines cyber security from three perspectives: (1) Network security spanning multiple network communication layers, such as sensors and wireless devices; (2) Information integrity, including threats such as viruses and insider attacks; (3) Hardware and software system vulnerabilities in information infrastructures.
Partners Health-Care	Informatics for Integrating Biology and the Bedside	2004	One of seven NIH-funded National Center for Biomedical Computing. Focuses on develop software and methodologies to enable clinical researchers to accelerate translation of genomic and "traditional" clinical findings into novel diagnostics, prognostics, and therapeutics. It has developed open source software, which enables researchers to combine genomic and molecular research with data and observations from electronic health records. i2b2 has created a web-based query and data sharing network called SHRINE (a).
UMASS Amherst	Center for Intelligent Information Retrieval	1992	Focuses on developing technology that provides effective and efficient access to large networks of heterogeneous, multimedia information.
	Institute for Computational Biology, Biostatistics & Bioinformatics	2012	Aims to apply computational, biomedical and translational research to the life sciences through high-level analytic methods. Activities focus on catalyzing intellectual exchange and connections among participating departments, pursuing extramural funding to support the development of educational opportunities, and engaging life science and IT companies to identify shared interests for future collaborations.
	Life Sciences Center	Planned	The planned facility will house three research centers: Biosensors and Big Data Center, the Healthcare Informatics and Technology Innovation Center, and the Models to Medicine Center. The Biosensors and Big Data Center will focus on developing techniques to continuously analyze patient data in real time.
UMASS Lowell	Center for Advanced Computation and Telecommunications	1993	Compute-intensive modeling of physical and information systems. Members of the Center have undertaken research in the areas medical imaging, acoustics, fluid dynamics, heat transfer, control, probabilistic modeling, information processing and communication networks.
	Ctr. for Computer Machine/ Human Intelligence Networking & Dist. Systems	2001	Research, training and education to help advance research in the analytical, experimental and operational aspect of computer hardware and software, Data Engineering, and Information Technologies that have influence on Data and "Big Data" Knowledge Extraction, Engineering and Services, and Machine/Human Computational Intelligence.

TABLE 36 CONTINUED) **Selected Research Centers in Massachusetts**

Institution	Center	Founded	Description
WPI	Center for Research in Exploratory Data and Information Analysis	NA	Research in data exploration and knowledge discovery, and to the application of this research in scientific, industrial, and commercial domains. Verticals: bioinformatics; e-commerce; earth and space science; security; communication networks; healthcare. Specific areas: knowledge discovery in databases; data mining; information visualization; machine learning; pattern recognition; statistics; signal analysis.

Notes: (a) The Shared Health Research Information Network (SHRINE) helps researchers overcome one of the greatest problems in population-based research. Eligible investigators may use the SHRINE web-based query tool to determine the aggregate total number of patients at participating hospitals who meet a given set of inclusion and exclusion criteria. The SHRINE network currently covers six million patients and provides more than 10 billion medical facts from the five participating institutions: Beth Israel Deaconess Medical Center, Boston Children's Hospital, Brigham and Women's Hospital, Dana-Farber Cancer Institute, and Massachusetts General Hospital. Additional pilot efforts are underway in California and other states.

TABLE 37) Federal Funding of Big Data Projects by Instituion, 2007-2013

Institute	Funding	Project Titles
Massachusetts Institute of Technology	4,472,529	• An Array Oriented Data Management System for Massive Scale Scientific Data • Kreyol-based Cyberlearning for a New Perspective on the Teaching of STEM in local Languages • Physical Database Design for Next-Generation Databases • Quantum Optomechanics on Multiple Mass Scales • Scalable and Secure Database as a Service • Social Robots as Mechanisms for Language Instruction, Interaction, and Evaluation in Pre-School Children (b) • Technology to Support Mathematical Argumentation (c)
Concord Consortium	2,912,271	• Integrating Sensors and Simulations to Improve Learning • Technologies in Support of Student Experimentation
University of Massachusetts Amherst	1,922,506	• Connecting the Ephemeral and Archival Information Networks • High-Performance Complex Processing of Continuous Uncertain Data • Probing Astrophysics Frontiers With Gravitational Wave Bursts • Support for Young Researchers to attend the International Intelligent Tutoring Systems Conference 2012 • Topical Positioning System (TPS) for Informed Reading of Web Pages
Northeastern University	1,643,891	• A Scalable Search Tool for Interesting Patterns in Scientific Data • Collection Construction Methodologies for Learning-to-Rank • Exploring Data in Multiple Clustering Views • Using Archival Resources to Conduct Data-Intensive Internet Research • Center for Historical Information and Analysis
Worcester Polytechnic Institute	1,498,299	• Complex Event Analytics • Managing Discoveries in Visual Analytics • Query Mesh – A Novel Paradigm for Query Processing
Tufts University	1,422,177	• Bridging Student, Scientific, and Mathematical Models with Expressive Technologies • Interdisciplinary Machine Learning Research and Education
Harvard University	1,386,764	• A Prototype WorldWide Telescope Visualization Lab Designed in the Web-based Inquiry Science Environment • Center for Historical Information and Analysis • DataBridge – A Sociometric System for Long-Tail Science Data Collections • Representation, Modeling and Inference for Large Biological and Infor Networks • Center for Historical Information and Analysis
Boston University	1,358,981	• Algorithms for Tandem Repeat Variant Discovery Using Next Generation Sequencing Data • Center for Historical Information and Analysis • Entity Selection and Ranking for Data-Mining Applications • Linguistically Based ASL Sign Recognition as a Structured Multivariate Learning Problem
University of Massachusetts Lowell	1,356,376	• Querying Rich Uncertain Data in Real Time • Transforming Science Learning with an Interactive Web Environment for Data Sharing and Visualization (d)
Brandeis University	586,385	• A Development Environment for Query Optimizer Engineering • An efficient, versatile, scalable, and portable storage system for scientific data
Springfield Technical Community College	549,458	• Exploring the Virtual World of Contextualized English Language Acquisition
Woods Hole Oceanographic Institute	124,009	• Articulating Cyberinfrastructure Needs of the Ocean Ecosystem Dynamics Community • Interoperability Testbed-Assessing a Layered Architecture for Integration of Existing Capabilities
TERC Inc	121,147	• Technology to Support Mathematical Argumentation
OpenAirBoston.net	48,500	• Mobile Pathways for 21st Century Learning
Education Development Center	46,053	• EarthCube Education End-User Workshop
Total	**19,449,346**	

Notes: (a) in partnership with Machine Science, Inc.

TABLE 38) **Federal funding for Big Data by State**

State	2012	2013 (June)	Total 2012 to 2013 (June)	Funding per Capita
CA	13,748,887	2,804,971	16,553,858	0.44
NY	7,386,227	3,481,660	10,867,887	0.56
WI	7,200,967	1,677,413	8,878,380	1.55
IL	7,221,999	396,352	7,618,351	0.59
MA	**6,109,283**	**1,466,378**	**7,575,661**	**1.14**
PA	4,015,443	2,713,881	6,729,324	0.53
NJ	2,679,830	2,961,402	5,641,232	0.64
MI	3,316,194	1,049,216	4,365,410	0.44
IN	4,003,752	281,029	4,284,781	0.66
CO	3,781,218	184,256	3,965,474	0.76
WA	715,469	2,727,567	3,443,036	0.50
MD	2,717,196	400,000	3,117,196	0.53
FL	2,866,945	126,027	2,992,972	0.15
TX	1,282,816	662,760	1,945,576	0.07
AZ	1,929,359		1,929,359	0.65
RI	1,816,685		1,816,685	1.73
MN	1,759,794		1,759,794	0.33
NC	1,728,754		1,728,754	0.18
UT	1,418,895	99,947	1,518,842	0.53
IA	135,000	1,300,000	1,435,000	0.47
DC	1,103,892	138,874	1,242,766	1.97
OH	970,778		970,778	0.08
LA	770,103	150,000	920,103	0.20
VA	193,494	449,850	643,344	0.08
GA	281,711	324,024	605,735	0.06
ME	549,291		549,291	0.41
SC	444,302		444,302	0.09
AL	399,394		399,394	0.08
OR	345,000		345,000	0.09
MO	278,533		278,533	0.05
MS	205,459		205,459	0.07
ID	170,811		170,811	0.11
KS	70,000		70,000	0.02
MT	50,000		50,000	0.05
State unspecified	10,387,062	8,981,353	19,368,415	NA
Total	**92,054,543**	**32,376,960**	**124,431,503**	**0.40**

TABLE 39) Northeastern University Course Requirements Master of Science in Health Informatics

Required: • Introduction to Health Informatics and Health Information Systems • The American Health Care System • Health Informatics Capstone Project	
Health Informatics (two): • Health Systems Lab • The Business of Health Care Informatics • Creation and Application of Medical Knowledge	**Technical (two):** • Database Design, Access, Modeling, and Security • Strategic Topics in Programming for Health Professionals • Data Management in Health Care • Key Standards in Health Informatics Systems
Business Management (two): • Organizational Behavior, Work Flow Design, and Change Management • Management Issues in Healthcare Information Technology • Project Management	**Elective Courses (two):** • Design for Usability in Health Care • Emerging Technologies in Healthcare • Introduction to Genomics and Bioinformatics • Legal and Social Issues in Health Informatics • Public Health Surveillance and Informatics

TABLE 40) MIT Course Requirements – MS Operations Research

Required: • Introduction to Mathematical Programming or Optimization Methods • Applied Probability • Statistical Learning and Data Mining or other advanced statistics subject	
Four additional graduate level courses. The following list indicates subjects frequently taken as electives.	
Applied Operations Research: • Logistical and Transportation Planning Methods • Engineering Risk-Benefit Analysis	**Economics and Finance:** • Options and Futures Markets • Advanced Financial Economics I
Optimization Techniques: • Dynamic Programming and Stochastic Control • Nonlinear Programming • Advanced Algorithms • Network Optimization • Combinatorial Optimization • Systems Optimization: Models and Computation	**Probabilistic Modeling and/or Simulation:** • Demand Modeling • Introduction to Numerical Simulation • Discrete Stochastic Processes • Queues: Theory and Applications • System Dynamics
Transportation: • Airline Schedule Planning • Carrier Systems • Transportation Flow Systems • Logistics Systems • Logistics and Supply Chain Management	

TABLE 41) Boston University Course Requirements Master of Science in Systems Engineering

Core: • Dynamic Systems Theory or Dynamic Programming and Stochastic Control • Optimization Theory and Methods • Advanced Stochastic Modeling and Simulation or Stochastic Processes or Probability with Statistical Applications	
Concentrations	
Computational & Systems Biology: • Molecular Bioengineering I • DNA and Protein Sequence Analysis • Computational Biology: Genomes, Networks, Evolution • Nonlinear Dynamics in Biological Systems • Adv. Signals and Systems Analysis for Biomedical Engineering • Structural Bioinformatics • Protein and Genomic Systems Engineering • Computational Genomics I	**Control Systems:** • Dynamic Systems Theory • Process Modeling and Control • Precision Machine Design and Instrumentation • Robot Motion Planning • Optimal and Robust Control • Recursive Estimation and Optimal Filtering • Adaptive Control • Advanced Process Control • Dynamic Programming and Stochastic Control • Discrete Event and Hybrid Systems • Vision Robotics and Planning • Nonlinear Systems and Control
Energy & Environmental Systems: • Game Theory • Sustainable Power Systems • Electrochemistry of Fuel Cells and Batteries • Energy & Environmental Economics • Public Control of Business • Solar Energy Systems • Regional Energy Modeling • Clean Technology Business Models	**Network Systems:** • Computer Communication Networks • Networking the Physical World • Wireless Communications • Queuing Systems • Randomized Network Algorithms • Mobile Networking and Computing • Communication Networks Control
Operations Research: • Simulation • Dynamic Programming and Stochastic Control • Advanced Stochastic Modeling and Simulation • Advanced Optimization Theory and Methods • Queuing Systems • Combinatorial Optimization and Graph Algorithms • Discrete Event and Hybrid Systems • Advanced Scheduling Models and Methods	**Production & Service Systems:** • Production Systems Analysis • Sustainable Power Systems • Discrete Event and Hybrid Systems • Production System Design • Advanced Scheduling Models and Methods • Creating Value Through Operations and Technology • Supply Chain Management

TABLE 42) MIT Required Courses for BS in Mathematics with Computer Science

- Differential Equations
- Linear Algebra
- Design and Analysis of Algorithms,
- Introduction to Electrical Engineering and Computer Science
- Introduction to Algorithm
- Mathematics for Computer Science or Principles of Discrete Applied Mathematics or Principles of Discrete Applied Mathematics
- Automata, Computability, and Complexity or Theory of Computation
- Elements of Software Construction or Computer System Engineering

TABLE 43) **MIT Required Courses PhD in Computational and Systems Biology**

<table>
<tr><td colspan="2"><b>Required:</b>
• Topics in Computational and Systems Biology
• Modern Biology
• Computational Biology
• Research Group Rotations (four two-month rotations)</td></tr>
<tr><td colspan="2">Four additional “advanced” electives. The following list indicates subjects frequently taken as electives.</td></tr>
<tr><td><b>Biology:</b>
• Principles and Practice of Drug Development
• Cell Biology: Structure and Functions of the Nucleus
• Eukaryotic Cell Biology: Principles and Practice
• Immunology
• Genetic Neurobiology
• Developmental Neurobiology
• Regulation of Gene Expression
• Topics in Metabolic Biochemistry
• Topics in Protein Biochemistry
• Nucleic Acids, Structure, Function, Evolution and Their Interactions with Proteins
• Topics of Mammalian Development and Genetics
• The Protein Folding Problem
• Cancer Biology</td><td><b>Biological Engineering:</b>
• Biomolecular Kinetics and Cellular Dynamics
• Fields, Forces, and Flows in Biological Systems
• Analysis of Biological Networks
• Cell and Tissue Engineering
• Glycomics
• Biological Engineering II: Instrumentation and Measurement
• Physical Biology
• Tools for Assessing Biological Function</td></tr>
<tr><td rowspan="3"><b>Brain and Cognitive Science:</b>
• Cellular and Molecular Neurobiology: The Brain and Cognitive Sciences III
• Neural Basis of Learning and Memory
• The Visual System
• Cognitive Neuroscience
• Neurology, Neuropsychology, and Neurobiology of Aging
• Biochemistry and Pharmacology of Synaptic Transmission
• Cellular Neurophysiology
• Developmental Neurobiology
• Animal Behavior
• Introduction to Computational Neuroscience
• Neural Plasticity in Learning and Development
• Genetic Neurobiology
• Cognitive Artifacts and Architectures
• Sensation and Perception
• Statistical Learning Theory and Applications
• Introduction to Neural Networks</td><td><b>Electrical Engineering and Computer Science:</b>
• Dynamic Programming and Stochastic Control
• Dynamic Systems and Control
• Integrated Microelectronic Devices
• Design & Fabrication of Microelectromechanical Devices
• Advanced Algorithms
• Network Optimization
• Integer Programming and Combinatorial Optimization
• Computational Functional Genomics</td></tr>
<tr><td><b>Mechanical Engineering:</b>
• Introduction to Numerical Simulation
• Optical Engineering
• Design and Fabrication of Microelectromechanical Devices</td></tr>
<tr><td><b>Civil and Environmental Engineering:</b>
• Nonlinear Dynamics and Waves</td></tr>
<tr><td><b>Chemistry:</b>
• Enzymes: Structure and Function
• Bioorganic Chemistry
• Biophysical Chemistry
• Biophysical Chemistry and Molecular Design
• Practical Macromolecular Crystallography</td><td><b>Chemical Engineering:</b>
• Production Systems Analysis
• Sustainable Power Systems
• Discrete Event and Hybrid Systems
• Production System Design
• Advanced Scheduling Models and Methods
• Creating Value Through Operations and Technology
• Supply Chain Management</td></tr>
<tr><td><b>Mathematics:</b>
• Stochastic Processes
• Introduction to Numerical Methods
• Introduction to Modeling and Simulation
• Nonlinear Dynamics and Chaos
• Combinatorial Optimization</td><td><b>Physics:</b>
• Statistical Mechanics I
• Statistical Mechanics II
• Nonlinear Optics
• Systems Biology
• Statistical Physics in Biology
• Biological Physics</td></tr>
</table>

TABLE 44) Clark University Required Courses of MA in GIS for Development and Environment

Required: • Advanced Vector GIS • Advanced Raster GIS • Introduction to Remote Sensing • GISDE Professional Seminar • Master's Final Research Requirement	
Seven additional graduate level courses. The following list indicates a sampling of "Skill" and "Policy" Electives. Students may take courses offered by the other three graduate programs in International Development and Social Change, Community Development and Planning, or Environmental Science and Policy, or in other departments.	
Skill Electives: • Python Programming • Computer Programming for GIS • Web mapping and Open Source GIS • Environmental Applications of GIS • Introduction to Quantitative Methods • Intermediate Quantitative Methods • Advanced Remote Sensing • Landscape Ecology • Concepts and Applications in Spatial Analysis • GIS and Land Change Science • GIS and accuracy assessment	**Policy Electives:** • Decision Methods for Environmental Management and Policy • US Environmental Pollution Policy • Biogeochemical Cycles and Global Change • Environmental Toxicology • Climate change, Energy and Development • Community Development Decision Making and Negotiations • Economic Fundamentals for International Development • Humanitarian Assistances in Complex Emergencies/ Disasters • Sustainable Consumption and Production • Fundamentals of Environmental Science • Participatory Development Planning • Seminar in Human Dimensions of Global Change: Impacts and Societal Responses • The Climate System and Global Environment Change

TABLE 45) Bentley Required Courses for Graduate Certificate in Business Analytics

- Quantitative Analysis for Business and Finance
- Intermediate Statistical Modeling for Business
- Business and Economic Forecasting
- Time Series Analysis
- Data Mining
- Customer Data Analysis and Relationship Marketing
- Data Management and Systems Modeling
- Data Warehousing and Data Mining
- The Macroeconomics of Financial Markets
- Market Structure and Firm Strategy
- Marketing Research and Analysis
- Internship in Business Data Analysis (optional)

TABLE 46) Harvard University Required Courses for MS in Computational Science and Engineering

- Data Science
- Computational Design of Materials
- Advanced Scientific Computing: Numerical Methods
- Computing Foundations of Computational Science
- Computational Fluid Dynamics
- Interdisciplinary Seminar in Computational Science and Engineering
- Advanced Scientific Computing: Stochastic Optimization Methods
- Systems Design for Computational Science

TABLE 47) Required Courses for NYU MS in Data Science

• Introduction to Data Science. Introduces students to basic algorithms and software tools, teaches how to deal with data, representing data, and methodology. Provides hands-on experience using Torch, a software system being developed at NYU and other research centers that has a large data science library
• Statistical and Mathematical Methods. Introduces basic statistical and mathematical methods needed in the practice of data science. It covers basic methods in probability, statistics, linear algebra, and optimization.
• Machine Learning and Computational Statistic. Covers a wide variety of topics in machine learning, pattern recognition, statistical modeling, and neural computation. It covers the mathematical methods and theoretical aspects, but primarily focuses on algorithmic and practical issues.
• Big Data. Covers methods and tools for automatic knowledge extraction from very large datasets. Methods include on-line learning, feature hashing, class embedding, distributed databases, map-reduce framework, CUDA GPU programming, and applications.
• Inference and Representation. Covers graphical models, causal inference, and advanced topics in statistical machine learning.
• Capstone Project and Presentation in Data Science

Source: http://cds.nyu.edu/academics/ms-in-data-science/curriculum/required-courses/

TABLE 48) Required Courses for IDSE Certification of Professional Achievement in Data Sciences

• Algorithms for Data Science
• Probability & Statistics
• Probability & Statistics
• Exploratory Data Analysis and Visualization

Source: http://idse.columbia.edu/certification

Appendix F

Referenced Tables

TABLE 49) Required Degrees by Job Type

Degree	Software Engineer	Data Scientist	Data Engineer	Marketing and Sales
Electrical Engineering	✓			
Computer Engineering	✓		✓	
Computer Science	✓	✓	✓	✓
Mathematics / Statistics of other analytic-intensive field	✓	✓	✓	✓
Other				✓

Source: 2013 Mass Big Data Survey.

TABLE 50) Required Knowledge of Specific Tools

Tool	Requirement	Software Engineer	Data Scientist	Data Engineer	Marketing and Sales
RDBMS, OLAP, OTLP, SQL	A plus, but not essential	✓			✓
	Essential	✓		✓	✓
ETL, Flume, Sqoop	A plus, but not essential	✓		✓	
	Essential	✓		✓	
Hadoop, HDFS, MapReduce	A plus, but not essential	✓	✓	✓	✓
	Essential		✓	✓	✓
NoSQL	A plus, but not essential	✓	✓	✓	✓
	Essential	✓	✓	✓	✓
NewSQL	A plus, but not essential	✓	✓	✓	✓
	Essential		✓	✓	✓
SPSS, SAS, MATLAB, R	A plus, but not essential	✓	✓	✓	✓
	Essential	✓	✓		
Hbase, PIG, Hive	A plus, but not essential	✓	✓		
	Essential		✓	✓	
Java, Python	A plus, but not essential	✓	✓	✓	
	Essential	✓	✓	✓	✓
Tableau, Gephi, Flare, etc.	A plus, but not essential	✓	✓	✓	
	Essential				

Source: 2013 Mass Big Data Survey.

TABLE 51) Meet-up Groups

Name	Founded	Description	Members	Number of Meet-ups
The Boston Python User Group	2007	Boston Python is the world's largest local Python user group. Meet other local Python developers, learners, employers, and enthusiasts of all kinds.	3,828	110
Boston Predictive Analytics	2010	The goal of this meet-up is present informative lectures, hands-on tutorials, networking events, etc, towards helping the local community further it's understanding and proficiency regarding Predictive Analytics. Our group has three main focal points: business applications, advanced mathematics, and computer science; with topics covering Recommeder Systems, Machine Learning, Google Analytics, Data Visualization, Social Media / Text Analytics, and related topics.	2,559	32
Boston Hadoop User Group	2009	Goal of most meetings will be build data models that attendees can use themselves; make data mining and data analytics accessible to everyone; and increase awareness of open source data mining tools. Our Members: Any and everyone who is interested in doing data mining and analytics without the hassle of coding.	1,644	33
Greater Boston useR Group (R Programming Language)	2011	R is a free and open programming language for statistical computing, data analysis, and graphical visualization. The Greater Boston useR Group seeks to bring this community together to share ideas, discuss R related topics, and provide direction for new and experienced users.	1,212	19
Boston Data Visualization	2011	This meet-up will aim to bring … anyone interested in data visualization together, fostering a community, creating a space for learning and enabling new partnerships. The meet-up will host talks, hack days, workshops/ tutorials and personal work showcases.	1,165	11
New England Artificial Intelligence	2011	Our group is for those interested in AI, machine learning, forecasting, recommendation systems, and building smarter applications. We share experience and knowledge in the field, and help each other with ideas and projects.	1,140	15
Data Science Group	2012	This group will concentrate on understanding the tools and skill-sets needed to become an effective Data Scientist. We will explore all topics related to the data lifecycle including acquiring new data sets, parsing new data sets, filtering and organizing data, mining data patterns, advanced algorithms, visually representing data, telling stories with data and softer skills such as negotiations and selling your ideas based upon data.	1,096	10
Big Data Boston	2012	Big Data Boston is for people with a passion for analytics & insights that are derived from the extreme information generated today. This group is for the small start up and the big company, the individual and the group, anyone in Boston that wants to make it the capital for Big Data!	861	16
Boston Cloud Services – All things Cloud, SaaS, PaaS, XaaS	2009	A group dedicated to sharing, evangelizing and promoting the next big wave in technology, Cloud based services: software (SaaS), Platform (PaaS) as-a-service etc with a focus on the end users of cloud services; specifically people who have dealt or are dealing with the move from on-premise to the cloud and either have faced or are facing issues, Business & Technical like: security, management, integration etc. How product management or IT changes when you're building or using a Cloud / SaaS product.	745	15
Boston Data Mining	2013	The Boston Data Mining meet-up focuses on making data mining accessible to everyone. It aims to be inclusive to all regardless if you are a beginner or expert in statistical modeling, machine learning, data mining, or any of the analytics (business, predictive, etc). Our talks will focus on the practical side of data mining and analytics by showing attendees how to perform data modeling in a non-programming environment. To achieve that goal we employ RapidMiner; an open source data mining platform used to quickly and easily prototype data modeling processes.	575	Upcoming

TABLE 51 C0NTINUED) **Meet-up Groups**

Name	Founded	Description	Members	Number of Meet-ups
Big Data Innovation Boston	2013	This is a group to connect Big Data enthusiasts in Boston for networking, presentations, workshops, demonstations and much more. This group is ideal for: * Big Data Startup Companies * Entrepreneurs * Developers and Programmers * Investors * Big Data Recruiters * Big Data Service Providers	510	2
Boston Hacking Predictive Analytics App	2009	This is your chance to prototype your ideas before going to VCs or for seed funding. We all have dream applications that need machine learning and predictive analytics. We can hack those app. using available machine learning APIs (from Microsoft, Google, BigML, BPA EyeQ, and so on). No need to learn those complicated math or gymnastics of matrices. We are the Big-data web, data driven web, and Web^n!	473	26
Boston Algorithmic Trading	2012	Boston Algorithmic Trading is for anyone interested in creating and using algorithms in the financial markets. We arrange monthly talks from practicing quants, algorithmic traders, trading technology experts, and academics. Our focus is practical, rather than theoretical. We enjoy talking about how to automate the purchase and sale of securities using statistics, machine learning, data mining, and algorithms.	412	6
Open Analytics Boston	2012	A group devoted to the use and development of open source, big data, agile intelligence solutions, for the Boston Metro area. Join our group if interested in solving real business problems utilizing open source, big data analytical solutions.	399	2
Data Mining for Marketers	2012	Data mining! Predictive Modeling! Statistical Modeling! Predictive analysis! What does this all mean for marketers, and how can it work in your organization? This group aims to explore how data mining is leveraged by marketing departments and firms to predict behavior and optimize response and profitability.	292	7
Boston Storm Users	2012	Boston Storm Users is a group for developers using or hoping to learn about Twitter's Storm real-time data processing framework. We get together to discuss best practices, to exchange ideas and to learn how to apply Storm to various engineering challenges.	267	5
Graph Database Boston	2012	Developers interested in learning about and working with graph databases for social, spatial, hierarchical or other highly connected data sets. We host hands-on lab sessions, technology reviews, topical lectures, and plenty of social beer nights. Curious about graphs, want to brush up your non-RDBMS skills? Join us!	265	4
Big Data Meet-up Boston	2013	This is a group to connect Big Data enthusiasts in Boston for networking, presentations, workshops, demonstations and much more. This group is ideal for: * Big Data Startup Companies * Entrepreneurs * Developers and Programmers * Investors * Big Data Recruiters * Big Data Service Providers	249	2
The North American VoltDB Meet-up Group	2012	This is a group for application developers who aspire to make the impossible possible – wicked smart folks who thrive on the challenge of blazing new trails, turning things upside down, and building a whole new breed of applications that will change the world.	239	5
Big Data Developers – Boston	2013	This is an IBM sponsored Big Data meet-up group. Geared towards developers, data scientists and ALL Big Data enthusiasts, our meetups provide an opportunity to work hands on with the solutions and tools in our Big Data portfolio. Our Meetups typically include a 45-60 min (max) presentation that serves as an introduction and overview for a specific Big Data technology. It is followed by ~3 hours to collaborate with fellow developers and apply your Big Data skills. We provide a cloud environment that you can run through the browser of your laptop at NO cost to you. Our meetups are FREE. Meet-up topics include: – Hadoop-based analytics – Stream Computing – Text Analytics – Visualization and Discovery tools for Big Data – Big Data App Development – Deep dives into the technologies that makes big data processing possible – Anything and everything about Big Data.	214	Upcoming

TABLE 51 C0NTINUED) Meet-up Groups

Name	Founded	Description	Members	Number of Meet-ups
Elasticsearch Boston	2012	Elasticsearch is picking up steam in Bean Town. Let's share our experiences with it and strengthen the local community around it by introducing newcomers to the great features it has to offer.	214	6
Learning Analytics Boston	2012	This group will provide a place to meet and discuss potential use of existing technologies and practice in the service of improving education. I started this group because there is a rich community across disciplines that offer a wide variety of complex approaches to this topic that merit review in the context of the the education sector. I look forward to reviewing technologies, organizing some hands on labs, participating in dialogue, and pursuing innovative approaches with you.	204	5
Boston Analytics Professionals	2013	The Analytics Professionals group was created to provide a forum for those in Boston to share tips, tricks and knowledge about using an analytic platform such as ParAccel to drive business value. With the variety and volume of data available these days from a number of sources (RDBMS, web logs, sensor data, social media, etc.) we're interested in discussing how you begin to make sense of all this data and derive real business value from it.	155	1
The Boston Vertica User & Modern Bi Meet-up Group	2012	You should join this group if you use Vertica and/or would like to learn more, share information and best practices and meet really awesome people. This is also a place to discuss modern business intelligence (BI) and techniques of dealing with Really Big Data. There are many of us who are not satisfied with traditional monster slow and expensive data warehouses and BI systems and are switching to Vertica. Vertica is sponsoring and participating in our group. Compete inc will be partnering in this meetups groups effort.	124	2
Data Science for the Bottom Billion	2012	Are you interested in data science but also passionate about international development? Then this is the meet-up for you! Let's get together to explore how the rapidly growing areas of data science and big data can transform the world of international development and contribute to improving the lives of the billions of underprivileged and poor people in the world.	116	2
Oracle NoSQL & Big Data - Boston	2013	Join us to learn about Oracle Big Data/NoSQL and related technologies, including use cases, new features and exciting networking with other professionals.	114	4
Big Data Analytics, Discovery & Visualization	2013	All things Big-data, Data Visualization, Data Discovery & Analysis. We will have meetups where we could learn and share best practices around Big-Data Analytics, Discovery & Visualization and it's impact on businesses. This group is relevant for data scientists, business executives seeking to learn what big-data discovery could do to their businesses & professionals who are interested to learn what data discovery is all about.	99	Upcoming
Boston Smart Data Meet-up Group	2013	Smart Data" refers to the high-value data used to inform business decisions. Our interest is in comprehensive enterprise solutions, from ingestion of data through to business insight. This includes: large-scale physical systems, virtual/cloud systems, operational systems such as Hadoop, BI tools, and techniques to drive business decisions	99	2

TABLE 52) **Applications for the Use of APDC**

Applicant	Project Title	Posted	Status
Mass. Health Connector	Risk Adjustment per Affordable Care Act	Jul. 13, 2012	Approved
Center for Health Policy and Research, Univ. of Mass. Med. School	Patient Centered Medical Home Evaluation	Aug. 14, 2012	Approved
Mass. Dept. of Public Health, Preventive and Behavioral Medicine, Univ. of Mass. Med. School	Health Care Reform and Disparities in the Care and Outcomes of Trauma Patients	Sep. 14, 2012	Amended
Mass. Dept. of Public Health, Bureau of Community Health and Prevention	Evaluation of Mass in Motion and Community Transformation Grants	Sep. 14, 2012	Approved
Mass. Dept. of Public Health, Tobacco Cessation and Prevention Program	Utilization of tobacco treatment in Massachusetts to quit smoking	Sep. 14, 2012	Approved
Mass. Dept. of Public Health, Center for Birth Defects Research and Prevention	Surveillance of Congenital Heart Defects (CHDs)	Sep. 14, 2012	NA
Mass. Dept. of Public Health, Bureau of Substance Abuse Services	Substance Abuse Treatment Needs and Services Gap Analysis	Nov. 14, 2012	Approved
Nat'l Bureau of Economic Research, Univ. of Penn. Yale Univ.	The Effects of Fragmentation in Health Care	Dec. 14, 2012	Approved
Harvard School of Public Health	Will the Academic Innovations Collaborative Increase the Value of Primary Care and Improve Providers' and Trainees' Experiences?	Jan. 8, 2013	Approved
Mass. Dept. of Public Health, Bureau of Infectious Disease	STD, HIV, and Viral Hepatitis Testing, Treatment, and Screening Trends	Jan. 9, 2013	Amended
Center for Health Policy and Research, Univ. of Mass. Med. School	Massachusetts Patient Centered Medical Home Initiative Shared Savings Project	Feb. 13, 2013	Approved
Kyruus, Inc	Understanding Provider Expertise and Behavior	Feb. 13, 2013	Pending
Yale Univ. and the Nat'l Bureau of Economic Research/Univ. of Penn.	Maternal and Paternal Health and Children's Healthcare Access and Use	Feb. 13, 2013	Pending
Mass. Health Quality Partners	Practice Pattern Variation Analysis (PPVA) Program	Feb. 19, 2013	Amended
MassHealth, Exec. Office of Health and Human Services, and Univ. of Mass. Med. School	Child Health Care Quality Measurement - Core Measure Set Testing	Mar. 13, 2013	Amended
Harvard School of Public Health	Understanding High-Cost Patients in Massachusetts	Apr. 11, 2013	Amended
Dr. Arnold Epstein and Dr Amy Boutwell	Analysis of the Massachusetts All-Payer Claims Database to Describe the Epidemiology of Readmissions	Apr. 12, 2013	NA
Kyruus, Inc.	Promoting Transparent Clinical Expertise	Aug. 12, 2013	NA

Source: http://www.mass.gov/chia/researcher/hcf-data-resources/apcd/accessing-the-apcd.html

TABLE 53) Patents Issued to Massachusetts' Inventors, 2008 to 2012

Class	Title	Number of Patents in MA	Shares of US	Location Quotient
700	DP: Generic Control Systems or Specific Applications	159	5.0%	1.0
701	DP: Vehicles, Navigation, and Relative Location	57	1.4%	0.3
702	DP: Measuring, Calibrating, or Testing	234	5.1%	1.0
703	DP: Structural Design, Modeling, Simulation, and Emulation	231	10.5%	2.1
704	DP: Speech Signal Processing, Linguistics, Language Translation, and Audio Compression/Decompression	148	5.6%	1.1
705	DP: Financial, Business Practice, Management, or Cost/Price Determination	527	3.8%	0.8
706	DP: Artificial Intelligence	127	6.0%	1.2
707	DP: Database and File Management or Data Structures	703	5.4%	1.1
708	Arithmetic Processing and Calculating	43	5.3%	1.1
709	Multicomputer Data Transferring	806	5.6%	1.1
710	Input/Output	168	4.9%	1.0
711	Memory	412	7.2%	1.4
712	Processing Architectures and Instruction Processing	59	3.9%	0.8
713	Support (Electrical Computers and Digital Processing Systems)	238	3.8%	0.8
714	Error Detection/Correction and Fault Detection/Recovery	256	4.0%	0.8
715	DP: Presentation Processing of Document, Operator Interface Processing, and Screen Saver Display Processing	274	5.3%	1.1
716	Computer-Aided Design, and Analysis of Circuits and Semiconductor Masks	66	2.2%	0.5
717	DP: Software Development, Installation, and Management	286	7.8%	1.6
718	Virtual Machine Task or Process Management or Task Mgt./Control	95	6.3%	1.3
719	Interprogram Communication or Interprocess Communication	74	5.2%	1.1
720	Dynamic Optical Information Storage or Retrieval	1	1.0%	0.2
725	Interactive Video Distribution Systems	45	2.2%	0.4
726	Information Security	241	5.3%	1.1
Total		**5,250**	**5.0%**	**1.0**

Source: Nexus Associates based on USPTO and US Census.

Acknowledgments

Special Thanks to

Massachusetts Competitive Partnership

Daniel O'Connell, *President & CEO,* Massachusetts Competitive Partnership
William H. Swanson, *Chairman of the Board, Chairman,* Raytheon
Jack M. Connors Jr., *Co-founder & Former Chairman,* Hill Holliday Connors Family Office
Roger W. Crandall, *President & CEO,* Mass Mutual Financial Group
John F. Fish, *Founder, President & CEO,* Suffolk Construction
Gary Gottlieb, *President & CEO,* Partners HealthCare
Joseph L. Hooley, *President & CEO,* State Street Corporation
Abigail P. Johnson, *President, Fidelity Financial Services,* Fidelity Investments
Robert K. Kraft, *Founder, Chairman & CEO,* The Kraft Group
Jeffrey M. Leiden, *Chairman, President & CEO,* Vertex Pharmaceuticals
David H. Long, *Chairman, President & CEO,* Liberty Mutual
Thomas J. May, *President & CEO,* Northeast Utilities
Brian T. Moynihan, *President & CEO,* Bank of America
Robert L. Reynolds, *President & CEO,* Putnam Investments
Ronald L. Sargent, *Chairman & CEO,* Staples
Laura J. Sen, *President & CEO,* BJ's Wholesale Club
Joseph M. Tucci, *Chairman, President & CEO,* EMC

Interviews

John Baker, *Founder,* DataKin & The Data Science Group
Justin Borgman, *Chief Executive Officer & Co-Founder,* Hadapt
Puneet Batra, *former Chief Data Scientist,* Kyruus
John Cardente, *Distinguished Engineer & Big Data Future Building Blocks Team Leader,* EMC
David Dietrich, *Advisory Technical Education Consultant, Big Data & Data Science,* EMC
Phil Francisco, *VP, Data Management Products & Strategy at IBM Information Management,* IBM
Leo Hermacinski, *CEO,* dSide Technologies
Tom Hopcroft, *President & CEO,* Mass Technology Leadership Council
Ze Jiang, *CEO & Founder,* iQuartic
William F. Kiczuk, *VP & Chief Technology Officer,* Raytheon
Marilyn Kramer, *Deputy Executive Director,* Center for Health Information & Analysis
Dave Laverty, *Vice President Marketing, Big Data & Analytics,* IBM
Sam Madden, *Faculty Director, BigData@CSAIL,* MIT
Shawn Murphy, *Director of Research Information Systems,* Partners HealthCare
Steve Papa, *Founder & former CEO,* Endeca
Paul Sonderegger, *Big Data Strategist,* Oracle
Matthew Trunnell, *Chief Technology Officer,* Broad Institute

Massachusetts Big Data Consortium Organizing Committee

Mohamad Ali, *Chief Executive Officer,* Workforce Optimization Division, Aspect Software

Gregory Bialecki, *Secretary for Housing & Economic Development,* Commonwealth of Massachusetts

Pamela Goldberg, *Chief Executive Officer,* Massachusetts Technology Collaborative

John Goodhue, *Executive Director,* Massachusetts Green High Performance Computing Center

Tom Hopcroft, *President & CEO,* Mass Technology Leadership Council

Patrick Larkin, *Director, Innovation Institute at MassTech*

Chris Lynch, *Hack-Reduce & Atlas Ventures Former Chief Executive Officer,* Vertica/HP

Jeffrey Nick, *Senior Vice President & Chief Technology Officer,* EMC Corporation

Steve Papa, *Founder & Fomer CEO, Endeca*

Andrei Ruckenstein, *Vice President & Associate Provost for Research,* Boston University

Daniela Rus, *Professor,* Department of Electrical Engineering & Computer Science, Director, CSAIL, Massachusetts Institute of Technology

Jit Saxena, *Founder & Chairman,* Netezza

Bob Zurek, *Senior Vice President,* Products Epsilon

Acknowledgments

Massachusetts Technology Collaborative Board of Directors

Executive Committee

The Honorable Gregory P. Bialecki, *Board Chairperson,* Massachusetts Technology Collaborative; Secretary, Executive Office of Housing and Economic Development, Commonwealth of Massachusetts

Donald R. Dubendorf, Esq., *Board Vice-Chairperson,* Massachusetts Technology Collaborative; Attorney, Dubendorf Law

Emily Nagle Green, *President and Chief Executive Officer*, Smart Lunches

Alain Hanover, *Co-founder,* CommonAngels, and Active Mentor, MIT Venture Mentoring Service

Robert Johnson, *President,* Becker College

Pamela Reeve, *Vice Chair,* The Commonwealth Institute

Mitchell G. Tyson, *Principal,* Tyson Associates

Board Members

Martin Aikens M.E.A., *Consultant*

Robert L. Caret, PhD, *President,* University of Massachusetts

Julie Chen, PhD. *Vice Provost for Research,* Francis College of Engineering, University of Massachusetts Lowell

Leland Cheung*, City Councilor,* City of Cambridge

Justin Dangel, *Chief Executive Officer,* Consumer United

Joseph Dorant, *President,* Massachusetts Organization of State Engineers and Scientists (MOSES)

Stephen W. Director, PhD, *Provost,* Northeastern University

Richard M. Freeland, PhD, *Commissioner,* Massachusetts Department of Higher Education

Dana Mohler-Faria, PhD, *President,* Bridgewater State College

The Honorable Glen Shor, *Secretary, Executive Office for Administration and Finance,* Commonwealth of Massachusetts

Andre Ruckenstein, *Founding President,* Massachusetts Green High Performance Computing Center; Professor of Physics and Former Vice President for Research, Boston University

Benjamin I. Schwartz, *Novelist; Principal,* Emphasis Consulting Group

Frederick Sperounis, PhD, *Executive Vice Chancellor,* University of Massachusetts Lowell

Karl Weiss, PhD, *Professor Emeritus,* Northeastern University

Acknowledgments

Innovation Institute Governing Board

Chairperson, Donald R. Dubendorf, Esq, *Attorney-at-Law*, Dubendorf Law; Board Vice-Chairperson, Massachusetts Technology Collaborative Board of Directors

Ex Officio Members

The Honorable Gregory P. Bialecki *Secretary, Executive Office of Housing and Economic Development,* Commonwealth of Massachusetts; Board Chairperson, Massachusetts Technology Collaborative Board

Pamela W. Goldberg *Chief Executive Officer,* Massachusetts Technology Collaborative

Marty Jones *President and Chief Executive Officer,* MassDevelopment

Governing Board Members

Julie Chen, PhD*, Vice Provost for Research,* University of Massachusetts - Lowell

C. Jeffrey Cook, *Partner,* Cohen Kinne Valicenti & Cook LLP

Thomas G. Davis, *Executive Director,* The Greater New Bedford Industrial Foundation

Priscilla H. Douglas, PhD, *Principal,* P.H. Douglas & Associates

Patricia M. Flynn, PhD, *Trustee Professor of Economics & Management*, Bentley University

Amy K. Glasmeier, PhD, *Head, Department of Urban Studies & Planning,* Massachusetts Institute of Technology

Mary K. Grant, PhD, *President,* Massachusetts College of Liberal Arts

Michael A. Greeley, *General Partner,* Flybridge Capital Partners

Emily Nagle Green, *President & Chief Executive Officer,* Smart Lunches LLC

C. Jeffrey Grogan, *Former Partner,* Monitor Group LP

Richard K. Lester, PhD, *Department Head of Nuclear Science & Engineering & Co-Chair of Industrial Performance Center*, Massachusetts Institute of Technology

Teresa M. Lynch, *Former Senior Vice President & Director of Research,* Initiative for a Competitive Inner City

Daniel O'Connell, *President,* Massachusetts Competitive Partnership

Timothy Rowe, *Founder & Chief Executive Officer*, Cambridge Innovation Center

Pieter J. Schiller, *Partner Emeritus,* Advanced Technology Ventures

Stephen C. Smith, Executive Director, Southeastern Regional Planning & Economic Development District

Mitchell G. Tyson, *Principal,* Tyson Associates

Karl Weiss, PhD, *Professor Emeritus,* Northeastern University

Jack M. Wilson, PhD, *President Emeritus & University Distinguished Professor of Higher Education,* Emerging Technologies, and Innovation, University of Massachusetts

Phyllis R. Yale, *Partner*, Bain & Company

Patrick Larkin, *Director, Innovation Institute at MassTech; Deputy Director, Massachusetts Technology Collaborative*.

Acknowledgments

2014 Mass Big Data Report Staff

Content Creation:

Project Director: **Christopher Scranton**, *Senior Manager for Big Data & Technology Initiatives,* Innovation Institute at MassTech

Mary Rose Greenough, *Senior Advisor,* Innovation Institute at MassTech

Steve Hopley, *Senior Manager for Big Data Programs,* Innovation Institute at MassTech

Lauren Evans, *Research & Analysis Intern,* Innovation Institute at MassTech

Consultants:

Research & Analysis: **Nexus Associates**

Report Design: **Yellow Inc.**

THE MASSACHUSETTS BIG DATA REPORT | *A Foundation for Global Leadership* | *April 2014*