1

WHAT IS DATA INDUSTRY?

The next generation of information technology (IT) is an emerging and promising industry. But, what's truly the "next generation of IT"? Is it the next generation mobile networks (NGMN), Internet of Things (IoT), high-performance computing (HPC), or is it something else entirely? Opinions vary widely.

From the academic perspective, the debates, or arguments, over specific and sophisticated technical concepts are merely hype. How so? Let's take a quick look at the essence of information technology reform (IT reform) – digitization. Technically, it is a process that stores "information" that is generated in the real world from the human mind in digital form as "data" into cyberspace. No matter what types of new technologies emerge, the data will stay the same. As the British scholar Viktor Mayer-Schonberger once said [1], it's time to focus on the "I" in the IT reform. "I," as information, can only be obtained by analyzing data. The challenge we expect to face is the burst of a "data tsunami," or "data explosion," so data reform is already underway. The world of "being digital," as advocated some time ago by Nicholas Negroponte [2], has been gradually transformed to "being in cyberspace."[1]

With the "big data wave" touching nearly all human activities, not only are academic circles resolved to change the way of exploring the world as the "fourth paradigm"[2] but industrial community is looking forward to enjoying profits from

[1]Cyberspace, invented by the Canadian author William Gibson in his science fiction of *Neuromancer* (1984).

[2]The fourth paradigm was put forwarded by Jim Gray. http://research.microsoft.com/en-us/um/people/gray.

The Data Industry: The Business and Economics of Information and Big Data, First Edition. Chunlei Tang.

"inexhaustible" data innovations. Admittedly, given the fact that the emerging data industry will form a strategic industry in the near future, this is not difficult to predict. So the initiative is ours to seize, and to encourage the enterprising individual who wants to seek means of creative destruction in a business startup or wants to revamp a traditional industry to secure its survival. We ask the reader to follow us, if only for a cursory glimpse into the emerging big data industry, which handily demonstrates the properties property of the four categories in Fisher–Clark's classification, which is to say: the resource property of primary industry, the manufacturing property of secondary industry, the service property of tertiary industry, and the "increasing profits of other industries" property of quaternary industry.

At present, industrial transformation and the emerging business of data industry are big challenges for most IT giants. Both the business magnate Warren Buffett and financial wizard George Soros are bullish that such transformations will happen. For example,[3] after IBM switched its business model to "big data," Buffett and Soros increased their holdings in IBM (2012) by 5.5 and 11%, respectively.

1.1 DATA

Scientists who are attempting to disclose the mysteries of humankind are usually interested in intelligence. For instance, Sir Francis Galton,[4] the founder of differential psychology, tried to evaluate human intelligence by measuring a subject's physical performance and sense perception. In 1971, another psychologist, Raymond Cattell, was acclaimed for establishing Crystallized Intelligence and Fluid Intelligence theories that differentiate general intelligence [3]. Crystallized Intelligence describes to "the ability to use skills, knowledge, and experience"[5] acquired by education and previous experiences, and this improves as a person ages. Fluid Intelligence is the biological capacity "to think logically and solve problems in novel situations, independently of acquired knowledge."[5]

The primary objective of twentieth-century IT reform was to endow the computing machine with "intelligence," "brainpower," and, in effect, "wisdom." This all started back in 1946 when John von Neumann, in supervising the manufacturing of the ENIAC (electronic numerical integrator and computer), observed several important differences between the functioning of the computer and the human mind (such as processing speed and parallelism) [4]. Like the human mind, the machine used a "storing device" to save data and a "binary system" to organize data. By this analogy, the complexities of machine's "memory" and "comprehension" could be worked out.

What, then, is data? Data is often regarded as the potential source of factual information or scientific knowledge, and data is physically stored in bytes (a unit of measurement). Data is a "discrete and objective" factual description related to an event,

[3]IBM's centenary: The test of time. The Economist. June 11, 2011. http://www.economist.com/node/18805483.

[4]https://en.wikipedia.org/wiki/Francis_Galton.

[5]http://en.wikipedia.org/wiki/Fluid_and_crystallized_intelligence.

and can consist of atomic data, data item, data object, and a data set, which is collected data [5]. Metadata, simply put, is data that describes data. Data that processes data, such as a program or software, is known as a data tool. A data set refers to a collection of data objects, a data object is defined in an assembly of data items, a data item can be seen as a quantity of atomic data, and an atomic data represents the lowest level of detail in all computer systems. A data item is used to describe the characteristics of data objects (naming and defining the data type) without an independent meaning. A data object can have other names [6] (record, point, vector, pattern, case, sample, observation, entity, etc.) based on a number of attributes (e.g., variable, feature, field, or dimension) by capturing what phenomena in nature.

1.1.1 Data Resources

Reaping the benefits of Moore's law, mass storage is generally credited for the drop in cost per megabyte from US$6,000 in 1955 to less than 1 cent in 2010, and the vast change in storage capacity makes big data storage feasible.

Moreover, today, data is being generated at a sharply growing speed. Even data that was handwritten several decades ago is collected and stored by new tools. To easily measure data size, the academic community has added terms that describe these new measurement units for storage: kilobyte (KB), megabyte (MB), gigabyte (GB), terabyte (TB), petabyte (PB), exabyte (EB), zettabyte (ZB), yottabyte (YB), nonabyte (NB), doggabyte (DB), and coydonbyte (CB).

To put this in perspective, we have, thanks to a special report, "All too much: monstrous amounts of data,"[6] in *The Economist* (in February 2010), an ingenious descriptions of the magnitude of these storage units. For instance, "a kilobyte can hold about half of a page of text, while a megabyte holds about 500 pages of text."[7] And on a larger scale, the data in the American Library of Congress amounts to 15 TB. Thus, if 1 ZB of 5 MB songs stored in MP3 format were played nonstop at the rate of 1 MB per minute, it would take 1.9 billion years to finish the playlist.

A study by Martin Hilbert of the University of Southern California and Priscila López of the Open University of Catalonia at Santiago provides another interesting observation: "the total amount of global data is 295 EB" [7]. A follow-up to this finding was done by the data storage giant EMC, which sponsored an "Explore the Digital Universe" market survey by the well-known organization IDC (International Data Corporation). Some subsequent surveys, from 2007 to 2011, were themed "The Diverse and Exploding Digital Universe," "The Expanding Digital Universe: A Forecast of Worldwide Information," "As the Economy Contracts, The Digital Universe Expands," "A Digital Universe – Are You Ready?" and "Extracting Value from Chaos."

The 2009 report estimated the scale of data for the year and pointed out that despite the Great Recession, total data increased by 62% compared to 2008, approaching 0.8 ZB. This report forecasted total data in 2010 to grow to 1.2 ZB. The 2010 report forecasted that total data in 2020 would be 44 times that of 2009, amounting to 35

[6] http://www.economist.com/node/15557421.

[7] http://www.wisegeek.org/how-much-text-is-in-a-kilobyte-or-megabyte.htm.

ZB. Additionally the increase in the amount of data objects would exceed that amount in total data. The 2011 report brought us further to the unsettling point that we have reached a stage where we need to look for a new data tool to handle the big data that is sure to change our lifestyles completely.

As data organizations connected by logics and data areas assembled by huge volumes of data reach a "certain scale," those massive different data sets become "data resources" [5]. The reason why a data resource can be one of the vital modern strategic resources for humans – even possibly exceeding, in the twenty-first century, the combined resources of oil, coal, and mineral products – is that currently all human activities, and without exception including the exploration, exploitation, transportation, processing, and sale of petroleum, coal, and mineral products, will generate and rely on data.

Today, data resources are generated and stored for many different scientific disciplines, such as astronomy, geography, geochemistry, geology, oceanography, aerograph, biology, and medical science. Moreover various large-scale transnational collaborative experiments continuously provide big data that can be captured, stored, communicated, aggregated, and analyzed, such as CERN's LHC (Large Hadron Collider),[8] American Pan-STARRS (Panoramic Survey Telescope and Rapid Response System),[9] Australian radio telescope SKA (Square Kilometre Array),[10] and INSDC (International Nucleotide Sequence Database Collaboration).[11] Additionally INSDC's mission is to capture, preserve, and present globally comprehensive public domain biological data. As for economic areas, there are the data resources constructed by financial organizations and the economic data, social behavior data, personal identity data, and Internet data, namely the data generated by social networking computations, electronic commerce, online games, emails, and instant messaging tools.

1.1.2 The Data Asset

As defined in academe, a standard asset has four characteristics: (1) it should have unexpired value, (2) it should be a debit balance, (3) it should be an economic resource, and (4) it should have future economic benefits. The US Financial Accounting Standards Board expands on this definition: "[assets are] probable future economic benefits obtained or controlled by a particular entity as a result of past transactions or events."[12] Basically, by this definition, assets have two properties: (1) an economic property, in that an asset must be able to produce an economic benefit, and (2) a legal property, in that an asset must be controllable.

Our now common understanding is that the intellectual asset, as one of the three key components[13] of intellectual capital, is a "special asset." This is based on the

[8] http://public.web.cern.ch/public/en/LHC/LHC-en.html.
[9] http://pan-starrs.ifa.hawaii.edu/public.
[10] http://www.ska.gov.au.
[11] http://www.insdc.org.
[12] http://accounting-financial-tax.com/2009/08/definition-of-assets-fasb-concept-statement-6.
[13] In the book *Value-Driven Intellectual Capital*, Sullivan argues that intellectual capital consists of intellectual assets, intellectual property and human assets.

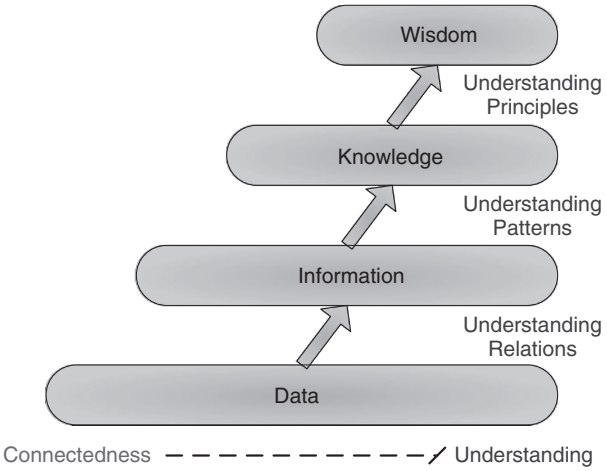

Figure 1.1 DIKW pyramid. Reproduced by permission of Gene Bellinger

concept of intellectual capital introduced in 1969 by John Galbraith, an institutional economist of the Keynesian school, and later expanded by deductive argument due to Annie Brooking [8], Thomas Stewart [9], and Patrick Sullivan [10]. In more recent years the concept of intellectual asset was further refined to a stepwise process by the British business theorist Max Boisot, who theorized on the "knowledge asset" (1999) [11]; by Chicago School of Economics George Stigler, who added an "information asset" (2003) [12]; and by DataFlux CEO Tony Fisher, who suggested a "data asset" specification process (2009) [13] that would closely follow the rules presented in the DIKW (data, information, knowledge, and wisdom) pyramid shown in Figure 1.1.

According to the ISO 27001:2005 standard, data assets are an important component of information assets, in that they contain source code, applications, development tools, operational software, database information, technical proposals and reports, job records, configuration files, topological graphs, system message lists, and statistical data.

We therefore want to treat *data asset* in the broadest sense of the term. That is to say, we want to redefine the data asset as data exceeding a certain scale that is owned or controlled by a specific agent, collected from the agent's past transactions involved in information processes, and capable of bringing future economic benefits to the agent.

According to Fisher's book *The Data Asset*, the administrative capacity of a data asset may decide competitive advantages of an individual enterprise, so as to mitigate risk, control cost, optimize revenue, and increase business capacity, as is shown in Figure 1.2. In other words, the data asset management perspective should closely follow the data throughout its life cycle, from discovery, design, delivery, support, to archive.

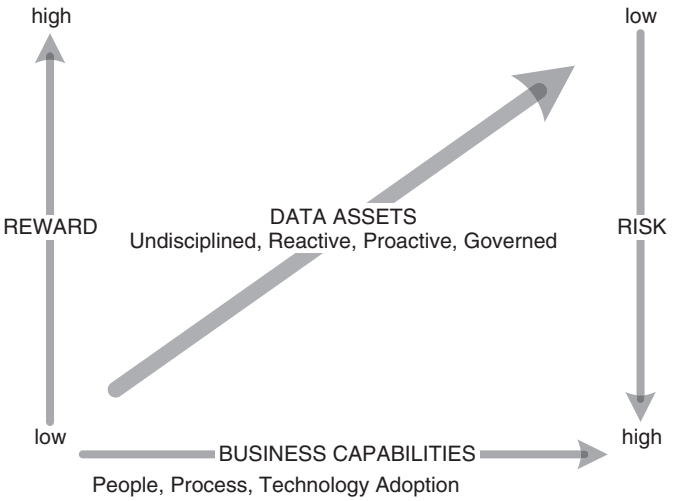

Figure 1.2 Advantages of managing data assets. Reproduced by permission of Wiley [13]

Our view[14] is that the primary value of data assets lies in the willingness of people to use data, and for some purpose as is reflected by human activities arising from data ownership or application of data. In a sense, data ownership, which defines and provides information about the rightful owner of data assets, depends on the "granularity of data items." Here is a brief clinical example of how to determine data ownership. Diagnostic records are associated with (1) patient's disease status, in terms of disease activity, disease progression, and prognosis, and (2) physician's medical experience with symptoms, diagnosis, and treatments. Strictly speaking, the patient and physician are both data owners of diagnostic records. However, we can minimize diagnostic records to patient's disease status, namely reduce its granularity such that only the patient takes data ownership of the diagnostic records.

1.2 INDUSTRY

The division of labor mentioned in one of Adam Smith's two classic works *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), is generally recognized as the foundation of industry [14], the industry cluster, and other industry schemes.

Industry is the inevitable outcome of the social division of labor. It was spawned by scientific and technological progress and by the market economy. Industry is in fact a generic term for a market composed of various businesses having interrelated benefits and related divisions of labor.

[14]This view is based on a discussion my peers and I had with Dr. Yike Guo, who is the founding director at Data Science Institute as well as Professor of Computer Science, Imperial College London.

1.2.1 Industry Classification

In economics, classification is usually the starting point and the foundation of research for industries. Industries can be classified in various ways:

- *By Economic Activity.* Primary industry refers to all the resource industries dealing with "the extraction of resources directly from the Earth," secondary industry to industries involved in "the processing products from primary industries," tertiary industry to all service industries, and the quaternary industry to industries that can significantly increase the industrial profits of other industries. The classification of tertiary industries is due to Fisher (1935) and the classification of quaternary industries is due to Clark (1940).
- *By Level of Industrial Activity.* There are three levels: use of similar products as differentiated by an "industrial organization," use of similar technologies or processes as differentiated by an "industrial linkage," and use of similar economic activities as differentiated by an "industrial structure."
- *By a System of Standards.* For international classification standards, we have the North American Industry Classification System (NAICS), International Standard Industrial Classification of All Economic Activities (ISIC), and so forth.

Of course, industries can be further identified by products, such as the chemical industry, petroleum industry, automotive industry, electronic industry, meatpacking industry, hospitality industry, food industry, fish industry, software industry, paper industry, entertainment industry, and semiconductor industry.

1.2.2 The Modern Industrial System

Computational optimization, modeling, and simulation as a paradigm not only produced IT reform of the information industry but also a fuzzy technology border, as new trends were added to the industry, such as software as a service, embedded software, and integrated networks. In this way, IT reform atomized the traditional industries and transformed their operation modes, thus prompting the birth of a new industrial system. The industries in this modern industrial system include, but are not limited to, the knowledge economy, high-technology industry, information industry, creative industries, cultural industries, and wisdom industry.

Knowledge Economy The "knowledge economy" is a term introduced by Austrian economist Fritz Machlup of Princeton University in his book *The Production and Distribution of Knowledge in the United States* (1962). It is a general category that has enabled the classification of education, research and development (R&D), and information service industries, but excluding "knowledge-intensive manufacturing," in "an economy directly based on the production, distribution, and use of knowledge and information," in accord with the 1997 definition by the OECD (Organization for Economic Co-operation and Development).

High-Technology Industry The high-technology industry is a derivative of the knowledge economy that uses "R&D intensity" and "percentage of R&D employees" as a standard of classification. The main fields are information, biology, new materials, aerospace, nuclear, and ocean, and characterized by (1) high demand for scientific research and intensity of R&D expenditure, (2) high level of innovativeness, (3) fast diffusion of technological innovations, (4) fast process of obsolescence of the prepared products and technologies, (5) high level of employment of scientific and technical personnel, (6) high capital expenditure and high rotation level of technical equipment, (7) high investment risk and fast process of the investment devaluation, (8) intense strategic domestic and international cooperation with other high-technology enterprises and scientific and research centers, (9) implication of technical knowledge in the form of numerous patents and licenses, (10) increasing competition in international trade.

Information Industry The "information industry" concept was developed in the 1970s and is also associated with the pioneering efforts of Machlup. In 1977 it was advanced by Marc Uri Porat [15] who estimated the predominant occupational sector in 1960 was involved in information work, and established Porat's measurements. The North American Industry Classification System (NAICS) sanctioned the information industry as an independent sector in 1997. According to the NAICS, the information industry includes three establishments engaged "(1) producing and distributing information and cultural products," "(2) providing the means to transmit or distribute these products as well as data or communications," and "(3) processing data."

Creative Industries Paul Romer, an endogenous growth theorist, suggested in 1986 that countless derived new products, new markets, and new opportunities for wealth creation [16] could lead to the creation of new industries. Although Australia put forward in 1994 the concept of a "creative nation," Britain was first to actually give us a manifestation of the "creative industries" when it established a new strategic industry with the support of national policy. According to the UK Creative Industries Mapping Document (DCMS) definition, creative industries as an industry whose "origin (is) in individual creativity, skill and talent and which has a potential for wealth and job creation through the generation and exploitation of intellectual property (1998)." This concept right away swept the globe. From London it spread to New York, Tokyo, Paris, Singapore, Beijing, Shanghai, and Hong Kong.

Cultural Industries The notion of a culture industry can be credited to the popularity of mass culture. The term "cultural industries" was coined by the critical theorists Max Horkheimer and Theodor Adorno. In the post-industrial age, overproduction of material similarly influenced culture, to the extent that the monopoly of traditional personal creations was broken. To criticize such "logic of domination in post-enlightenment modern society by monopoly capitalism or the nation state," Horkheimer and Adorno argued that "in attempting to realise enlightenment values of reason and order, the holistic power of the individual is undermined."[15] Walter Benjamin, an eclectic thinker also from the Frankfurt School, had the opposite view.

[15]http://en.wikipedia.org/wiki/Culture_industry.

He regarded culture as due to "technological advancements in art." The divergence of those views reflects the process of culture "from elites to the common people" or "from religious to secular," and it is such argumentations that accelerated culture industrialization to emerge as the "cultural industry." In the 1960s, the Council of Europe and UNESCO (United Nations Educational, Scientific and Cultural Organization) changed "industry" to the plural form "industries," to effect a type of industry economy in a broader sense. In 1993, the UNESCO revised the 1986 cultural statistics framework, and defined the cultural industries as "those industries which produce tangible or intangible artistic and creative outputs, and which have a potential for wealth creation and income generation through the exploitation of cultural assets and production of knowledge-based goods and services (both traditional and contemporary)." Additionally what cultural industries "have in common is that they all use creativity, cultural knowledge, and intellectual property to produce products and services with social and cultural meaning." The cultural industries therefore include cultural heritage, publishing and printing, literature, music, performance art, visual arts, new digital media, sociocultural activities, sports and games, environment, and nature.

Wisdom Industry Taking the lead in exalting "wisdom," in a commercial sense, IBM has been a vital player in the building of a "Smarter Planet" (2008). In the past IBM had advanced two other such commercial hypes: "e-Business" in 1996 and "e-Business on Demand" in 2002. These commercial concepts, as they were expanded both in connotation and denotation, allowed IBM to thus explore both market depth and width. With the intensive propaganda related to Cloud computing and the IoT, there are now hundreds of Chinese second-tier and third-tier cities that have discussed constructing a "Smart City." In the last couple of years IBM has won bids for huge projects in Shenyang, Nanjing, Shenzhen, among other places. To the best of our knowledge, however, the wisdom industry, which has only temporarily appeared in China, is based on machines and, we believe, will never have the ability to possess wisdom, knowledge, and even information, without the human input of data and thus data mining.

From these related descriptions of industries, we can see that cultural industries have a relatively broad interpretation. The United States treats cultural industries as copyright industries in the commercial and legal sense, whereas Japan has shifted to the expression "content industries" based on the transmission medium. In the inclination to emphasize "intellectual property" over "commoditization," the wisdom industry, knowledge economy, and information industry (disregarding the present order of appearance) are externally in compliance with the DIKW pyramid. The information industry may be further divided into two sectors. The first sector is the hardware manufacturing sector that includes equipment manufacturing, optical communication, mobile communication, integrated circuit, display device, and application electronics. The second is the information component of the services sector that includes the software industry, network information service (NIS), digital publishing, interactive entertainment, and telecommunications service.[16] The wisdom industry, which is essentially commercial hype despite being labeled "an upgraded version of creative

[16]https://en.wikipedia.org/wiki/Telecommunications_service.

industries," is no more than a use of "human beings" disguised as industrial carriers to "machines."

1.3 DATA INDUSTRY

From the foregoing description one could say that the information industry may be simply understood as digitization. Technically, IT is a process that stores "information" generated in the real world by human minds in digital form amassed as "data" in cyberspace, as is the process of producing data. In time the accumulated data can be sourced from multiple domains and distinct sectors.

The mining of "data resources" and extracting useful information already is seemingly "inexhaustible" as data innovations keep on emerging. Thus, to effectively endow all the data innovations with a business model – namely industrialization – would call for us to rename this strategic emerging industry, which is strong enough to influence the world economy, "data industry." The data industry is the reversal, derivation, and upgrading of the information industry.

1.3.1 Definitions

Connotation and denotation are two principal ways of describing objects, events, or relationships. Connotation relates to a wide variety of natural associations, whereas denotation consists in a precise description. Here, based on these two types of descriptions, we offer two definitions, in both a wide and a narrow sense, for the data industry.

In a wide sense, the data industry has evolved three technical processes: data preparation, data mining, and visualization. By these means, the data industry connotes rational development and utilization of data resources, effective management of data assets, breakthrough innovation of data technologies, and direct commoditization of data products. Accordingly, by definition then, the existing industrial sectors – such as publishing and printing, new digital media, electronic library and intelligence, digital content, specific domain data resources development, and data services in distinct sectors – should be included in the data industry. To these we should add the existing data innovations of web creations, data marketing, push services, price comparison, and disease prevention.

In a narrow sense, the data industry is usually divided into three major components: upstream, midstream, and downstream. In this regard, by definition, the data industry denotes data acquisition, data storage, data management, data processing, data mining, data analysis,[17] data presentation, data product pricing, valuation, and trading.

1.3.2 An Industry Structure Study

To understand profitability of a new industry, one must look at the distinctive structure that shapes the unfolding nature of competitive interactions. On the surface, the data industry is extremely complex. However, there are only four connotative factors associated with the data industry. These factors include: data resources, data assets,

[17]In this book, from the perspective of *Data Science*, I try to distinguish data mining and traditional data analysis tools or techniques, the latter refer to data analysis.

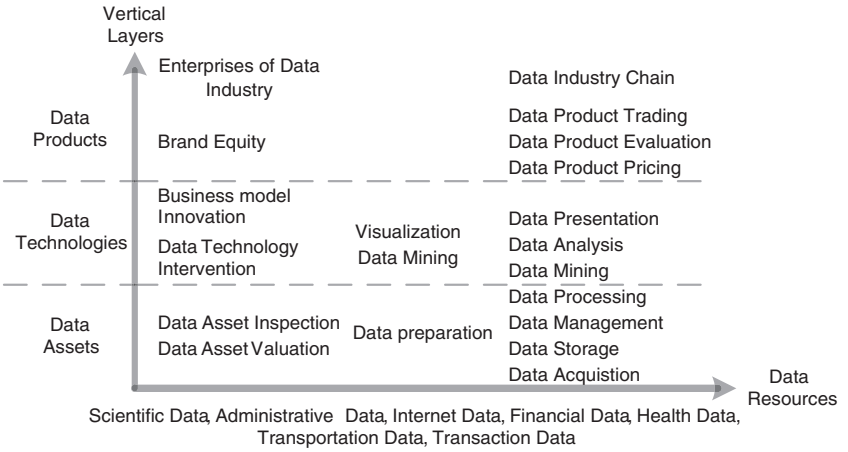

Figure 1.3 Structure of the data industry

data technologies, and data products. In a nutshell, from a vertical bottom-top view, the structure of the data industry (as shown in Figure 1.3) could be expressed by (1) data assets precipitation that forms the foundation of the data industry, (2) data technologies innovation as its core, and (3) data products circulation as its means. Theoretically, these three layers rely on data sources via mutually independent units that form underlying substructures, and then vertically form the entire data industry chain.

Technology Substructure The essence of the industry is to cope with conversion technologies. The corresponding term for the data industry is "data science," which is "a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD)."[18]

Peter Naur, a Danish pioneer in computer science and Turing award winner, once coined a new word – dataloy – in 1966 because he disliked the term computer science. Subsequently datalogy was adopted in Denmark and in Sweden as datalogi. However, Naur lost to William Cleveland of Purdue University in the influence of new word combinations or coinages, despite the fact that Naur was far more known than Cleveland. In 2001, Cleveland suggested a new word combination – data science – as an extension of statistics and, using that term, published two academic journals *Data Science Journal* and *The Journal of Data Science* (on the two disciplines of statistics) in 2002 and 2003, respectively. Cleveland's proposal has had an enormous impact over the years. Whenever or wherever people mentioned "data analysis" now, they first associate it with statistical models. Yet this curious episode did not stop data technologies from evolving.

Back to the technology substructure, the data industry has developed through the following three steps.

Step 1: *Data Preparation.* Similar to the geological survey and analysis during mineral exploration [5], data preparation determines data quality and

[18]https://en.wikipedia.org/wiki/Data_science.

selection of follow-up mining. These methods include (1) judgment of the availability of a data set (e.g., if a data source is isomeric and the data set is accessible); (2) analysis of the physical and logical structure of a data set; and (3) metadata acquisition and integration.

Step 2: *Data Mining.* As an efficient and scalable tool, data mining draws on ideas [6] from other disciplines. The ideas include (1) query optimization techniques, like indexing, labeling, and join algorithms, to enhance query processing from traditional database technologies; (2) sampling, estimation, and hypothesis testing from statistics; (3) search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning; and (4) high-performance or parallel computing, optimization, evolutionary computing, information theory, signal processing, and information retrieval from other areas. In general, data mining tasks are divided into two categories [6]: predictive tasks and descriptive tasks. Both of these data mining processes utilize massive volumes of data and exploratory rules to discover hidden patterns or trends in the data that cannot be found with traditional analytical tools or by human intuition.

Step 3: *Visualization.* The idea of visualization originated with images created by computer graphics. Exploration in the field of information visualization [17] became popular in the early 1990s, and was used to help understand abstract analytic results. Visualization has remained an effective way to illuminate cognitively demanding tasks. Cognitive applications increased in sync with the large heterogeneous data sets in fields such as retail, finance, management, and digital media. Data visualization [18], an emerging word combination containing both "scientific visualization" and "information visualization," has been gradually accepted. Its scope has been extended to include the interpretation of data through 3D graphics modeling, image rendering, and animation expression.

Resource Substructure Data resources have problems similar to those of traditional climate, land, and mineral resources. These include an uneven distribution of resource endowments, reverse configuration of production and use, and difficulties in development. That is to say, a single property or combination of properties of a data resource (e.g., diversity, high dimensionality, complexity, and uncertainty) can simultaneously reflect the position and degree of priority for a specific region within a given time frame so as to directly dictate regional market performance.

The resource substructure of the data industry consists of (1) a resource spatial structure (i.e., the spatial distribution of isomorphic data resources in different regions); (2) a resource type structure (i.e., the spatial distribution of non-isomorphic data resources in the same region); (3) a resource development structure (i.e., the spatial-temporal distribution of either to-be-developed data resources or having-been-developed data resources that were allowed for development); (4) a resource utilization structure (i.e., the spatial-temporal distribution of multilevel deep processing of having-been-developed data resource); and (5) a resource protection

structure (i.e., the spatial-temporal distribution of protected data resources according to a specific demand or a particular purpose).

Sector Substructure The sector substructure of the data industry is based on the relationships of various data products arising from the commonness and individuality in the processing of production, circulation, distribution, and consumption.

In regard to the information industry, sub-industries of the data industry may have two methods of division. First is whether data products are produced. This can be divided into (a) nonproductive sub-industry and (b) productive sub-industry. In this regard data acquisition, data storage, and data management belong to the nonproductive sub-industry, and in the productive sub-industry, data processing and data visualization directly produce data products while data pricing, valuation, and trading indirectly produce data products. Second is whether data products are available to a society. Data product availability can be divided into (c) an output projection sub-industry and (d) an inner circulation sub-industry, whereby the former provides data products directly to society and the latter provides data products within a sub-industry or to other sub-industries.

1.3.3 Industrial Behavior

Industrial behavior of the data industry is concentrated on four areas: data scientist (or quant [19]), data privacy, product pricing, and product rivalry.

Data Scientist Victor Fuchs, often called the "Dean of health economists", named the physician "the captain of the team" in his book *Who Shall Live? Health, Economics, and Social Choice* (1974). Data scientists could be similarly regarded the "captains" of the data industry.

In October 2010, *Harvard Business Review* announced[19] that the data scientist has been becoming "the sexiest job of the 21st century." Let's look at what it means to be called "sexiest." It is not only the attraction of this career path that is implied, it is more likely the art implied by "having rare qualities that are much in demand." The authors of this HBS article were Thomas Davenport and D. J. Patil, both men well known in academe and in industrial circles. Davenport is a famous academic author, and the former chief of the Accenture Institute for Strategic Change (now called Accenture Institute for High Performance Business, based in Cambridge, Massachusetts). Davenport was named one of the world's "Top 25 Consultants" by *Consulting* in 2003. Patil is copartner at Greylock Partners, and was named the first US Chief Data Scientist by the White House in February 2015. In the article they described the data scientist as a person having clear data insights through the use of scientific methods and mining tools. Data scientists need to test hunches, find patterns, and form theories. Data scientists not only need to have a professional background in "math, statistics, probability, or computer science" but must also have "a feel for business issues and empathy for customers." In particular, the top data scientists should be developers of new data mining algorithms or innovators of data products and/or processes.

[19]https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

According to an earlier report by the McKinsey Global Institute,[20] data scientists are in demand worldwide and their talents are especially highly sought after by many large corporations like Google, Facebook, StumbleUpon, and Paypal. Almost 80% of the related employees think that the yearly salary of this profession is expected to rise. The yearly salary for a vice president of operations may be as high as US$132,000. MGI estimated that "by 2018, in the United States, 4 million positions will require skills" gained from experience working with big data and "there is a potential shortfall of 1.5 million data-savvy managers and analysts."

Data Privacy Russian-American philosopher Ayn Rand wrote in his 1943 book *The Fountainhead* that "Civilization is the progress toward a society of privacy." As social activities increasingly "go digital," privacy becomes more of an issue related to posted data. Every January 28 is designated as Data Privacy Day (DPD) in the United States, Canada, and 47 European countries, to "raise awareness and promote privacy and data protection best practices."[21]

Private data includes medical and social insurance records, traffic tickets, credit history, and other financial information. There is a striking metaphor on the Internet: computers, laptops, and smart phones are the "windows" – that is to say, more and more people (not just identifying thieves and fraudsters) are trying to break them into your "private home," to access your private information. The simple logic behind this metaphor is that your private data, if available in sufficient quantity for analysis, can have huge commercial interest for some people.

Over the past several years, much attention has been paid to private data snooping, and to the storage of tremendous amounts of raw data in the name of national security. For instance, in 2011, Google received 12,271 requests to hand over its users' private data to US government agencies, and among them law enforcement agencies, according to company's annual Transparency Report. Telecom operators responded to "a portion of the 1.3 million"[22] law enforcement requests for text messages and phone location data were largely without issued warrants. However, a much greater and more immediate data privacy threat is coming from large number of companies, probably never even heard of, called "data brokers."[23] They are electronically collecting, analyzing, and packaging some of the most sensitive personal information and often electronically selling it without the owner's direct knowledge to other companies, advertisers, and even the government as a commodity. A larger data broker named Acxiom, for example, has boasted that it has, on average, "1,500 pieces of information on more than 200 million Americans [as of 2014]."[23]

No doubt, data privacy will be a central issue for many years to come. The right of transfer options for private electronic data should be returned to owners from the handful of companies that profiteer by utilizing other people's private information.

Product Pricing We use the search engine (a primary data product) to demonstrate how to price a product. It is noteworthy that a search engine is not really software and

[20] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
[21] http://en.wikipedia.org/wiki/Data_Privacy_Day.
[22] http://www.wired.com/2012/07/massive-phone-surveillance.
[23] http://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information.

is not really free. As early as 1998, Bill Gross, the founder of GoTo.com, Inc. (now called Overture), applied for a patent for search engine pricing.

Today's popular search engines operate using an open and free business model, meaning they do not make money from users but instead are paid by advertisers. There are two types of advertising in the search engine. One is the pay-per-click (PPC) model used by Google, whereby no payment is solicited from the advertiser if no user clicks on the ad. The other is the "ranking bid" model "innovated" by Baidu, whereby search results are ranked according to the payment made by advertisers. Google, in October 2010, adjusted its cost-per-click (CPC) pricing by adding a 49% premium to wrestler-type advertising sponsors[24] who want to take the optimum position in the results. Cost-per-click is similar to Baidu's left ranking that has existed for long time and contributes almost 80% of the revenue from advertisements.

Compared to the search engine, targeted advertising is a more advanced data product. Targeted advertising consists of community marketing, mobile marketing, effect marketing, interaction innovation, search engine optimization (SEO), and advertisement effect monitoring. Despite the fact that targeted advertising "pushes" goods information to the consumers, its vital function is to "pull," to exploit the vicissitudes and chaotic behavior of consumers. One way is by classifying users through the tracking and mining of Cookie files in users' browsers and then associating these classes by matching related products along with sponsor rankings. Another way is to monitor users' mouse movements by calculating residence time to try and determine the pros and cons of an interactive pop-up ad. Yet, there are far more than these two ways to target consumers, such as by listening to background noise (music, wind, breathing, etc.) produced by a user's laptop microphone. In sum, the purpose of targeted advertising is to nudge customer interest preferences to the operational level, with plenty of buying options to increase revenue, enhance the interactive experience, retain customer loyalty, and reduce the cost of user recall.

Product Rivalry

Oligarch Constraint: e-Books Here we only focus on the content of e-Books in digital form, without their carriers – computers, tablets, smart phones, and other electronic devices.

When Richard Blumenthal, the senior US Senator from Connecticut, served as Attorney General of Connecticut, he sent a letter of inquiry in August 2010 to Amazon regarding antitrust scrutiny on the pricing of e-Books. Blumenthal, undoubtedly, thought that the accord between sellers and publishers on e-Book pricing was bound to the increase chance of monopoly pricing, in "driving down prices in stock and pushing up prices in sales" adopted by Amazon to suppress smaller competitors. Today, there are over 3.5 million e-Books available in the Kindle Store of Amazon, and most of them are sold for less than US$10.

[24]Wrestler-type advertising sponsors refer to those who are willing to pay high to advertise their poor-quality products.

Weakening of Oligarch Restriction: eBooks Searched via Price Comparison In December 2010, Google entered this chaotic e-Book market, and claimed an "all about choice" strategy, which is to say, (1) any devices, including Android and iOS devices, browsers, special eBook readers (e.g., Amazon's Kindle, Barnes & Noble's Nook); (2) any book Google would ultimately provide, amounting to more than 130 million e-Books worldwide, with an initial 3 million volumes online including scanned edition of unique copies;[25] (3) any payment options, e-Books that could be bought from Google's Checkout using various payment options. Interestingly, Google added a function allowing price comparison of its e-Books to thousands of cooperative retailers. James McQuivey, a vice president and principal analyst at Forrester Research, commented that Google opened a gate to about 4,000 retailers who previously did not have the capability to invest in large-scale technology necessary to surmount powerful market competition.

1.3.4 Market Performance

Relative to this text three metric standards are used for measuring the market performance of the data industry: product differentiation, efficiency and productivity, and competition.

Product Differentiation In traditional industries, despite excluding a purely competitive market as well as an oligopoly market [20], product variety is considered to differ by some degree of innovation to allow for product differentiation (or simply differentiation). In other words, launching a new product is better than changing the packaging, advertising theme, or functional features of a product. In the data industry, however, we cannot say that product variety is more relevant to innovation compared to product differentiation. For example, for search engines, a plurality of search engines (e.g., keyword search tools, image search engines) belong to product differentiation, but they are based on different data technologies and depend on supply, demand, and consumers of various walks of life, in communities and organizations. Particularly, data product differentiation is controlled by the scale and diversity of data resources, such that even using the same algorithm in different domains will result in different data products.

Efficiency and Productivity Efficiency is "the extent to which time, effort, or cost is used well for the intended task or function,"[26] and this is usually classified into technical efficiency (and technological advance where the time factor is considered), cost efficiency, allocative efficiency, and scale efficiency. Productivity is an efficiency of production activities [21], and this can be expressed as a function, namely the ratio of output to inputs used in the production process. When the production process involves a single input and a single output, the production function can

[25]These so-called unique copies originated from the controversial Google Books Library Project launched in 2004 under the assistance of five partners: Harvard University Library, Stanford University Library, Oxford University Library, Michigan University Library, and New York Public Library.
[26]http://en.wikipedia.org/wiki/Efficiency.

be used to indicate productivity. When integrating multiple inputs or multiple outputs, total factor productivity is needed to show a change (increase or decrease) in productivity.

In 2001, after researching the efficiency and productivity of the information industry between 1995 and 1999, economist Dale Jorgenson of Harvard [22] pointed out that, over 50% of the entire technological advance in the US economy should be attributed to the technological advances in information hardware manufacturing. China and Japan also witnessed fast advancement in this sector. However, recent research shows[27] that from 2002 to 2006, the total factor productivity of China's software industry was 3.1% while the gain in technical efficiency was only 0.9%. This shows clearly that expensive hardware replacement and slow software innovations can no longer rapidly push economic growth. In addition, users have begun customizing data products according to their own demands, instead of buying standards-based servers, software, and solutions.

Competition Unlike other industries, competition in the data industry covers political, economic, military, and cultural areas – from the microscopic to the macroscopic and from virtual to real. Big data has already encroached on such fields, directly affecting our lives, as aerospace, aviation, energy, electric power, transportation, healthcare, and education. But the data industry faces competition both within a nation's borders and beyond its borders, which is to say, international competition. In the future it is probable that this international data competition will cause nations to compete for digital sovereignty in accord with the scale and activity of the data owned by a country and its capability of the interpreting and utilizing data. Cyberspace may prove to be another gaming arena for great powers, besides the usual border, coastal, and air defense tactics.

In the United States, in 2003, the White House published *The National Strategy to Secure Cyberspace*, a document that defines the security of cyberspace as a subset of Homeland Security. The US Air Force (USAF) answered that call in December 2005 when it enlarged the scope of its operational mission to fly and fight in air, space, and cyberspace. One year later, during a media conference, the USAF announced the establishment of an Air Force Cyberspace Command (USCYBERCOM). In March 2008, the USCYBERCOM released its strategic plan, and set new requirements for the traditional three missions of the USAF. These include (1) global vigilance: perception and transfer; (2) global reach: connection and transmission; and (3) global power: determent and crackdown. In 2009, President Obama personally took charge of a cyberspace R&D project where the core content is data resource acquisition, integration and processing, and utilization. In the same year, Obama issued a presidential national security order that set the cybersecurity policy as a national policy priority, and defined cyberspace crime as unauthorized entry and acquisition of data. In September 2010, the US military forces successfully destroyed the nuclear facilities of Iran through the virus "stwxnet" that was hidden in a flash driver, starting a war in

[27] Source: Li, He. A Study of Total Factor Productivity of Software Industry in China. Master's Thesis of Zhejiang Technology and Business University, 2008.

cyberspace. On March 29, 2012, the Obama administration released the *Big Data Research and Development Initiative*, which included the Department of Defense, Defense Advanced Research Projects Agency, National Science Foundation, National Institutes of Health, Department of Energy, and US Geological Survey. The six federal departments and agencies made commitments to invest over 200 million dollars altogether, "to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data." In another related development,[28] the Pentagon has approved a major expansion of the USCYBERCOM in January 2013 over several years, "increasing its size more than fivefold" – "Cyber command, made up of about 900 personnel, will expand to include 4,900 troops and civilians." Recently, the USCYBERCOM appears to be more urgent the need to reach "a goal of 6,000 person"[29] by the end of 2016.

So far as the United States is concerned, it has implemented an entire force operational roadmap of both internal and external cyberspace as well as data resource protection, utilization, and development. By now, Russia, Britain, Germany, India, Korea, and Japan are doing similar work.

It should be noted that unlike previous warfare, this so-called sixth-generation war[30] is showing much heavier dependence on the industry sector. For example, when the United States carried out its cyberspace maneuvers, the participants included multiple government departments and related private sector companies, in addition to the operational units. Future international data industry competitions will ultimately shape the competitive advantages of all countries in cyberspace.

[28]http://www.washingtonpost.com/world/national-security/pentagon-to-boost-cybersecurity-force/2013/01/27/d87d9dc2-5fec-11e2-b05a-605528f6b712_story.html.
[29]http://www.defenseone.com/threats/2015/02/us-cyber-command-has-just-half-staff-it-needs/104847.
[30]Russia refers the cyber warfare as the "sixth-generation war."