Part 1

Big Data, Clouds and Internet of Things

1

Big Data Science and Machine Intelligence

CHAPTER OUTLINE

1.1 Enabling Technologies for Big Data Computing

Over the past three decades, the state of high technology has gone through major changes in computing and communication platforms. In particular, we benefit greatly from the upgraded performance of the Internet and World Wide Web (WWW). We examine here the evolutional changes in platform architecture, deployed infrastructures, network connectivity and application variations. Instead of using desktop or personal computers to solve computational problems, the clouds appear as cost-efficient platforms to perform large-scale database search, storage and computing over the Internet.

This chapter introduces the basic concepts of data science and its enabling technologies. The ultimate goal is to blend together the sensor networks, RFID *(*radio frequency identification) tagging, GPS services, social networks, smart phones, tablets, clouds and Mashups, WiFi, Bluetooth, wireless Internet+, and 4G/5G core networks with the

Big-Data Analytics for Cloud, IoT and Cognitive Computing, First Edition. Kai Hwang and Min Chen.

Companion Website: http://www.wiley.com/go/hwangIOT

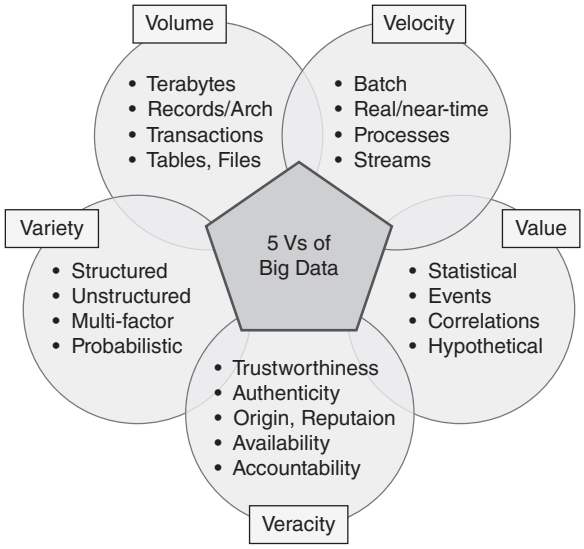


Figure 1.1 Big data characteristics: Five V's and corresponding challenges.

emerging Internet of Things (IoT) to build a productive big data industry in the years to come. In particular, we will examine the idea of technology fusion among the SMACT technologies.

1.1.1 Data Science and Related Disciplines

The concept of data science has a long history, but only recently became very popular due to the increasing use of clouds and IoT for building a smart world. As illustrated in Figure 1.1, today's big data possesses three important characteristics: data in large volume, demanding high velocity to process them, and many varieties of data types. These are often known as the five V's of big data, because some people add two more V's of big data: one is the veracity, which refers to the difficulty to trace data or predict data. The other is the data value, which can vary drastically if the data are handled differently.

By today's standards, one Terabyte or greater is considered a big data. IDC has predicted that 40 ZB of data will be processed by 2030, meaning each person may have 5.2 TB of data to be processed. The high volume demands large storage capacity and analytical capabilities to handle such massive volumes of data. The high variety implies that data comes in many different formats, which can be very difficult and expensive to manage accurately. The high velocity refers to the inability to process big data in real time to extract meaningful information or knowledge from it. The veracity implies that it is rather difficult to verify data. The value of big data varies with its application domains. All the five V's make it difficult to capture, manage and process big data using the existing hardware/software infrastructure. These 5 V's justify the call for smarter clouds and IoT support.

Forbes, Wikipedia and NIST have provided some historical reviews of this field. To illustrate its evolution to a big data era, we divide the timeline into four stages, as shown in Figure 1.2. In the 1970s, some considered data science equivalent to data logy, as noted by Peter Naur: "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

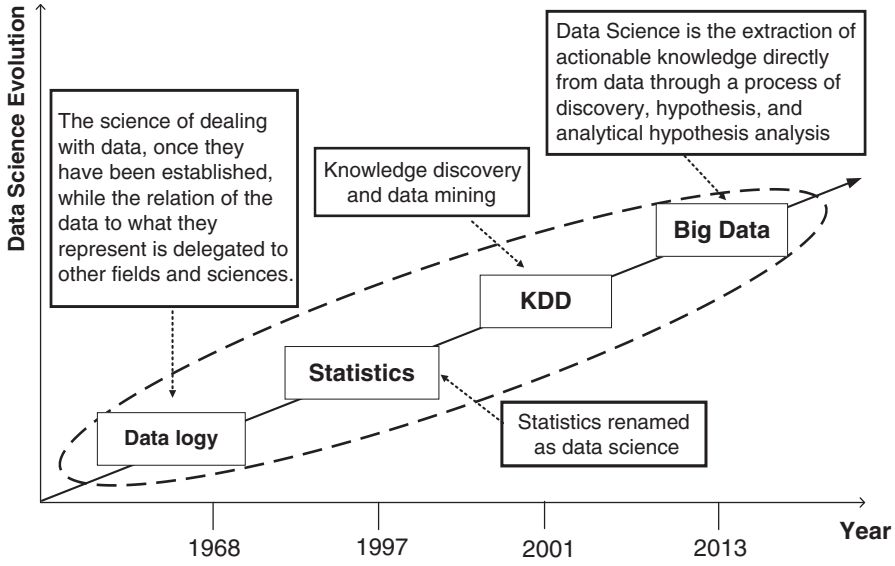


Figure 1.2 The evolution of data science up to the big data era.

At one time, data science was regarded as part of statistics in a wide range of applications. Since the 2000s, the scope of data science has become enlarged. It became a continuation of the field of data mining and predictive analytics, also known as the field of knowledge discovery and data mining (KDD).

In this context, programming is viewed as part of data science. Over the past two decades, data has increased on an escalating scale in various fields. The data science evolution enables the extraction of knowledge from massive volumes of data that are structured or unstructured. Unstructured data include emails, videos, photos, social media, and other user-generated contents. The management of big data requires scalability across large amounts of storage, computing and communication resources.

Formally, we define data science as the process of extraction of actionable knowledge directly from data through data discovery, hypothesis and analytical hypothesis. A data scientist is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, analytical skills and programming expertise to manage the end-to-end scientific process through each stage in the big data life cycle.

Today's data science requires aggregation and sorting through a great amount of information and writing algorithms to extract insights from such a large scale of data elements. Data science has a wide range of applications, especially in clinical trials, biological science, agriculture, medical care and social networks, etc [1]. We divide the value chain of big data into four phases: namely data generation, acquisition, storage and analysis. If we take data as a raw material, data generation and data acquisition are an exploitation process. Data storage and data analysis form a production process that adds values to the raw material.

In Figure 1.3, data science is considered as the intersection of three interdisciplinary areas: computer science or programming skills, mathematics and statistics, and

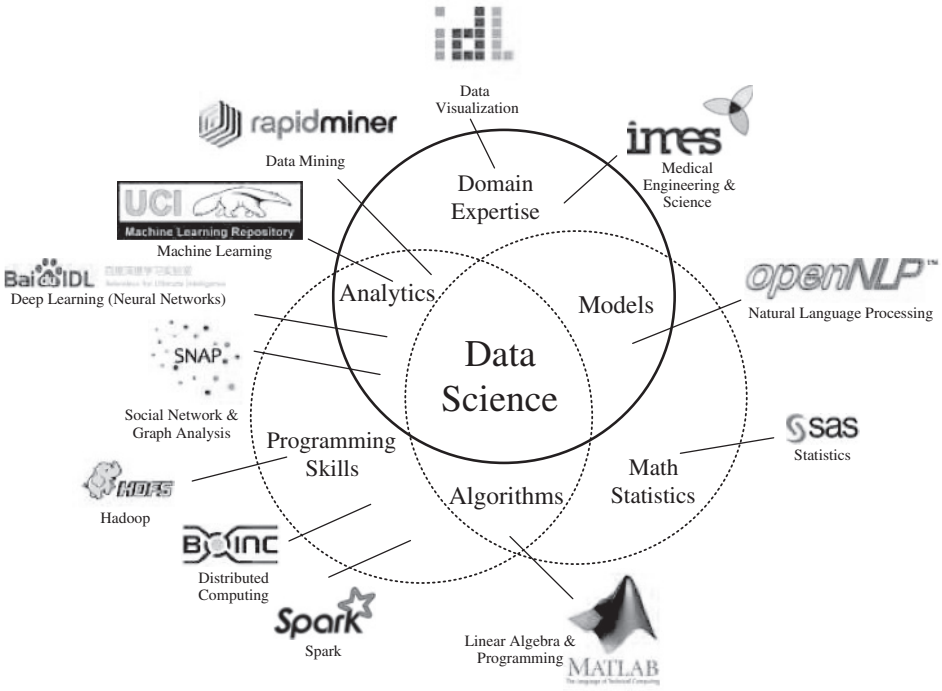


Figure 1.3 Functional components of data science supported by some software libraries on the cloud in 2016.

application domain expertise. Most data scientists started as domain experts who are mastered in mathematical modeling, data mining techniques and data analytics. Through the combination of domain knowledge and mathematical skills, specific models are developed while algorithms are designed. Data science runs across the entire data life cycle. It incorporates principles, techniques and methods from many disciplines and domains, including data mining and analytics, especially when machine learning and pattern recognition are applied.

Statistics, operations research, visualization and domain knowledge are also indispensable. Data science teams solve very complex data problems. As shown in Figure 1.3, when ever two areas overlap, they generate three important specialized fields of interest. The modeling field is formed by intersecting domain expertise with mathematical statistics. The knowledge to be discovered is often described by abstract mathematical language. Another field is data analytics, which has resulted from the intersection of domain expertise and programming skills. Domain experts apply special programming tools to discover knowledge by solving practical problem in their domain. Finally, the field of algorithms is the intersection of programming skills and mathematical statistics. Summarized below are some open challenges in big data research, development and applications:

- Structured versus unstructured data with effective indexing;
- Identification, de-identification and re-identification;

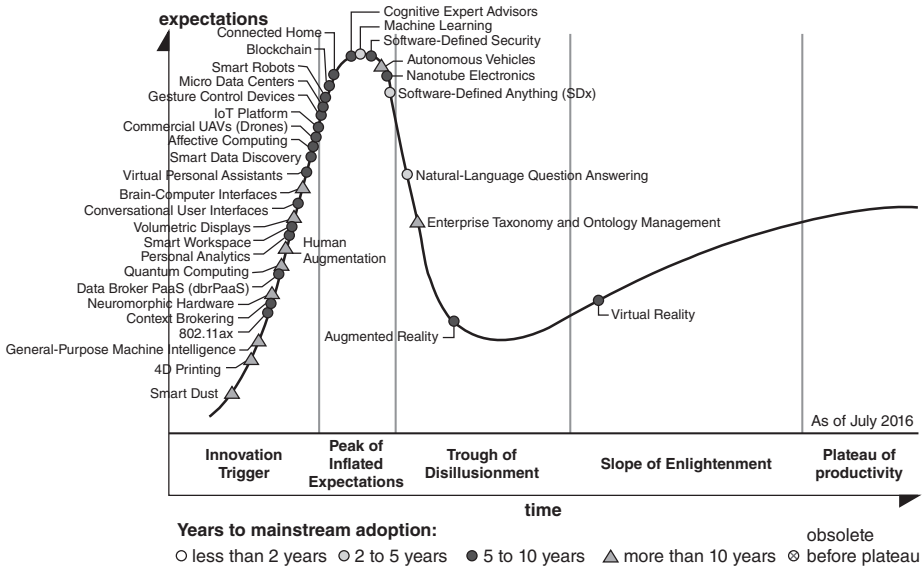


Figure 1.4 Hype cycle for emerging high technologies to reach maturity and industrial productivity within the next decade. (Source: Gartner Research, July 2016, reprinted with permission.) [19]

- Ontologies and semantics of big data;
- Data introspection and reduction techniques;
- Design, construction, operation and description;
- Data integration and software interoperability;
- Immutability and immortality;
- Data measurement methods;
- Data range, denominators, trending and estimation.

1.1.2 Emerging Technologies in the Next Decade

Garnter Research is an authoritative source of new technologies. They identify the hottest emerging new technologies in hype cycles every year. In Figure 1.4 we examine Gartner's Hype Cycle for new emerging technologies across many fields in 2016. The time taken for an emerging technology to become mature may take 2 to 10 years to reach its plateau of productivity. By 2016, the most expected technologies are identified at the peak of the hype cycle. The top 12 include cognitive expert advisors, machine learning, software defined security, connected home, autonomous vehicles, blockchain, nanotube electronics, smart robots, micro datacenters, gesture control devices, IoT platforms, and drones (commercial UAVs).

As identified by the dark solid circles, most technologies take 5 to 10 years to mature. The light solid circles, such as machine learning, software defined anything (SDx) and natural language answering, are those that may become mature in 2 to 5 years' time. Readers should check hype cycles released in previous years to find more hot technologies. The triangles identify those that may take more than 10 years of further development. They are 4-D printing, general-purpose machine intelligence, neuromorphic hardware, quantum computing and autonomous vehicles, etc. Self-driving cars were a

hot topic in 2016, but may need more time to be accepted, either technically or legally. The enterprise taxonomy and ontology management are entering the disillusion stage, but still they may take a long shot at becoming a reality.

Other hot technologies, like augmented reality and virtual reality, resulted in disillusionment, but they are heading towards industrial productivity now. At the early innovation trigger stage, we observe that Wifi 11.ac and context brokering are rising on the horizon, together with Data broker PaaS (dbrPaaS), personal analytics, smart workplace, conversational user interfaces, smart data discovery, affective computing, virtual personal assistant, digital security and people-literate technology. Many other technologies on the rising edge of the expectation curve include 3-D bio-printing, connected homes, biochips, software-defined security, etc. This hype cycle does include more mature technologies such as hybrid cloud computing, cryptocurrency exchange and enterprise 3-D printing identified in previous years.

Some of the more mature technologies such as cloud computing, social networks, near-field communication (NFC), 3-D scanners, consumer telematics and speech recognition, that have appeared in hype cycles released from 2010 to 2015, do not appear in Figure 1.4. The depth of disillusionment may not be bad, because as interest wanes after extensive experiments, useful lessons are learned to deliver products more successfully. Those long-shot technologies marked by triangles in the hype cycle cannot be ignored either. Most industrial developers are near-sighted or very conservative in the sense that they only adopt mature technologies that can generate a profitable product quickly. Traditionally, the long-shot or high-risk technologies such as quantum computing, smart dust, bio-acoustic sensing, volumetric displays, brain–human interface and neurocomputers are only heavily pursued in academia.

It has been well accepted that technology will continue to become more human-centric, to the point where it will introduce transparency between people, businesses and things. This relationship will surface more as the evolution of technology becomes more adaptive, contextual and fluid within the workplace, at home, and interacting with the business world. As hinted above, we see the emergence of 4-D printing, brain-like computing, human augmentation, volumetric displays, affective computing, connected homes, nanotube electronics, augmented reality, virtual reality and gesture control devices. Some of these will be covered in subsequent chapters.

There are predictable trends in technology that drive computing applications. Designers and programmers want to predict the technological capabilities of future systems. Jim Gray's "Rules of Thumb in Data Engineering" paper is an excellent example of how technology affects applications and vice versa. Moore's Law indicates that the processor speed doubles every 18 months. This was indeed true for the past 30 years. However, it is hard to say that Moore's Law will hold for much longer in the future. Gilder's Law indicates that the network bandwidth doubled yearly in the past. The tremendous price/performance ratio of commodity hardware was driven by the smart phone, tablets and notebook markets. This has also enriched commodity technologies in large-scale computing.

It is interesting to see the high expectation of IoT in recent years. The cloud computing in mashup or other applications demands computing economics, web-scale data collection, system reliability and scalable performance. For example, distributed transaction processing is often practised in the banking and finance industries. Transactions represent 90% of the existing market for reliable banking systems. Users must deal with

multiple database servers in distributed transactions. How to maintain the consistency of replicated transaction records is crucial in real-time banking services. Other complications include shortage of software support, network saturation and security threats in these business applications.

A number of more mature technologies that may take 2 to 5 years to reach the plateau are highlighted by light gray dots in Figure 1.4. These include biochip, advanced analytics, speech-to-speech translation, machine learning, hybrid cloud computing, cryptocurrency exchange, autonomous field vehicles, gesture control and enterprise 3-D printing. Some of the mature technologies that are pursued heavily by industry now are not shown in the 2016 hype cycle as emerging technologies. These may include cloud computing, social networks, near-field communication (NFC), 3-D scanners, consumer telematics and speech recognition that appeared in the hype cycles in last several years.

It is interesting to see the high expectation of IoT in recent years. The cloud computing in mashup or hybrid clouds has already been adopted in the mainstream. As time goes by, most technologies will advance to better stages of expectation. As mentioned above, the depth of disillusionment may not be too bad, as interest wanes after extensive experiments, and useful lessons are learned to deliver products successfully. It should noted that those long-shot technologies marked by triangles in the hype cycle may take more than 10 years to become an industrial reality. These include the rising areas of quantum computing, smart dust, bio-acoustic sensing, volumetric displays, human augmentation, brain–human interface and neuro-business popular in the academia and research communities.

The general computing trend is to leverage more and more on shared web resources over the Internet. As illustrated in Figure 1.5, we see the evolution from two tracks of system development: HPC versus HTC systems. On the HPC side, supercomputers

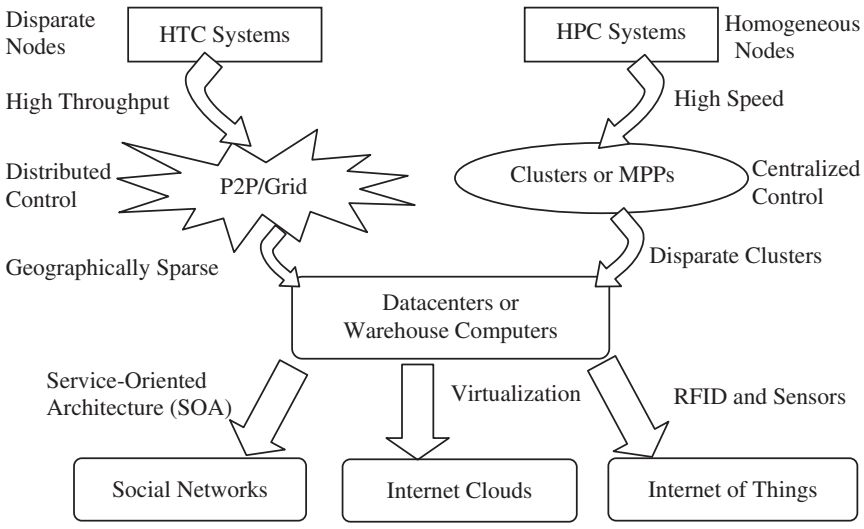


Figure 1.5 Evolutional trend towards parallel, distributed and cloud computing using clusters, MPPs, P2P networks, computing grids, Internet clouds, web services and the Internet of things. (HPC: high-performance computing; HTC: high-throughput computing; P2P: peer-to-peer; MPP: massively parallel processors; RFID: Radio Frequency Identification [2].)

(massively parallel processors, MPP) are gradually replaced by clusters of cooperative computers out of a desire to share computing resources. The cluster is often a collection of homogeneous compute nodes that are physically connected in close range to each other.

On the HTC side, Peer-to-Peer (P2P) networks are formed for distributed file sharing and content delivery applications. Both P2P, cloud computing and web service platforms place more emphasis on HTC rather than HPC applications. For many years, HPC systems emphasized raw speed performance. Therefore, we are facing a strategic change from the HPC to the HTC paradigm. This HTC paradigm pays more attention to high-flux multi-computing, where Internet searches and web services are requested by millions or more users simultaneously. The performance goal is thus shifted to measure the high throughput or the number of tasks completed per unit of time.

In the big data era, we are facing a data deluge problem. Data comes from IoT sensors, lab experiments, simulations, society archives and the web in all scales and formats. Preservation, movement and access of massive datasets require generic tools supporting high performance scalable file systems, databases, algorithms, workflow and visualization. With science becoming data centric, a new paradigm of scientific discovery is based on data intensive computing. We need to foster tools for data capture, data creation and data analysis. The cloud and IoT technologies are driven by the surge of interest in the deluge of data.

The Internet and WWW are used by billions of people every day. As a result, large datacenters or clouds must be designed to provide not only big storage but also distributed computing power to satisfy the requests of a large number of users simultaneously. The emergence of public or hybrid clouds demands the upgrade of many datacenters using larger server clusters, distributed file systems and high-bandwidth networks. With massive smart phones and tablets requesting services, the cloud engines, distributed storage and mobile networks must interact with the Internet to deliver mashup services in web-scale mobile computing over the social and media networks closely.

Both P2P, cloud computing and web service platforms emphasize high-throughput over a large number of user tasks, rather than high performance as often targeted in using supercomputers. This high-throughput paradigm pays more attention to the high flux of user tasks concurrently or simultaneously. The main application of the high-flux cloud system lies in Internet searches and web services. The performance goal is thus shifted to measure the high throughput or the number of tasks completed per unit of time. This not only demands improvement in the high speed of batch processing, but also addresses the acute problem of cost, energy saving, security and reliability in the clouds.

The advances in virtualization make it possible to use Internet clouds in massive user services. In fact, the differences among clusters, P2P systems and clouds may become blurred. Some view the clouds as computing clusters with modest changes in virtualization. Others anticipate the effective processing of huge datasets generated by web services, social networks and IoT. In this sense, many users consider cloud platforms a form of utility computing or service computing.

1.1.2.1. Convergence of Technologies

Cloud computing is enabled by the convergence of the four technologies illustrated in Figure 1.6. Hardware virtualization and multicore chips make it possible to have

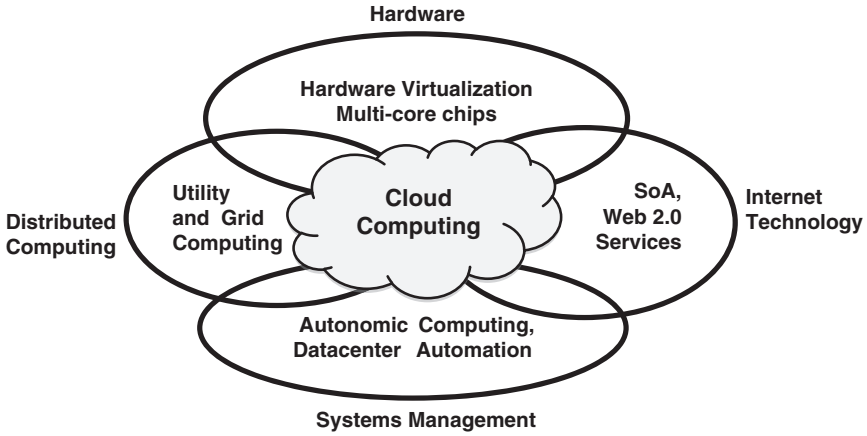


Figure 1.6 Technological convergence enabling cloud computing over the Internet. (Courtesy of Buyya, Broberg and Goscinski, reprinted with permission [3])

dynamic configurations in clouds. Utility and grid computing technologies lay the necessary foundation of computing clouds. Recent advances in service oriented architecture (SOA), Web 2.0 and mashups of platforms are pushing the cloud to another forward step. Autonomic computing and automated datacenter operations have enabled cloud computing.

Cloud computing explores the muti-core and parallel computing technologies. To realize the vision on data-intensive systems, we need to converge from four areas: namely hardware, Internet technology, distributed computing and system management, as illustrated in Figure 1.6. Today's Internet technology places the emphasis on SOA and Web 2.0 services. Utility and grid computing lay the distributed computing foundation needed for cloud computing. Finally, we cannot ignore the widespread use of datacenters with virtualization techniques applied to automate the resources provisioning process in clouds.

1.1.2.2. Utility Computing

Computing paradigms are attributed to different characteristics. First, they are all ubiquitous to our daily lives. Reliability and scalability are two major design objectives. Second, they are aimed at autonomic operations that can be self-organized to support dynamic discovery. Finally, these paradigms can be mixed with QoS (quality of service) and SLA (service-level agreement), etc. These paradigms and their attributes realize the computer utility vision.

Utility computing is based on a business model, by which customers receive computing resources from cloud or IoT service providers. This demands some technological challenges, including almost all aspects of computer science and engineering. For example, users may demand new network-efficient processors, scalable memory and storage schemes, distributed OS, middleware for machine virtualization, new programming models, effective resource management and application program development. These hardware and software advances are necessary to facilitate mobile cloud computing in various IoT application domains.

Table 1.1 Differences of three cloud service models from the on-premise computing in resources control under user, vendor and shared responsibilities.

Resources Types	On-Premise Computing	IaaS Model	PaaS Model	SaaS Model
App Software	User	User	Shared	*Vendor*
Virtual Machines	User	Shared	Shared	*Vendor*
Servers	User	*Vendor*	*Vendor*	*Vendor*
Storage	User	*Vendor*	*Vendor*	*Vendor*
Networking	Shared	*Vendor*	*Vendor*	*Vendor*

1.1.2.3. Cloud Computing versus On-Premise Computing

Additional computing applications are primarily executed on local hosts on premises. They appear as desktops, deskside, notebooks or tablets, etc. On-premise computing differs from cloud computing mainly in resources control and infrastructure management. In Table 1.1, we compare three cloud service models with the on-premise computing paradigm. We consider hardware and software resources in five types: storage, servers, virtual machines, networking and application software, as listed in the left-hand column of Table 1.1. In the case of on-premise computing at local hosts, all resources must be acquired by the users except networking, which is shared between users and the provider. This implies a heavy burden and operating expenses on the part of the users.

In the case of using an IaaS cloud like AWS EC2, the user only needs to worry about application software deployment. The virtual machines are jointly deployed by user and provider. The vendors are responsible for providing the remaining hardware and networks. In using the PaaS clouds, like Google AppEngine, both application codes and virtual machines are jointly deployed by user and vendor and the remaining resources are provided by the vendors. Finally, when the SaaS model is using the Saleforce cloud, everything is provided by the vendor, even including the app software. In conclusion, we see that cloud computing reduces users' infrastructure management burdens from two resources to none, as we move from IaaS to PaaS and SaaS services. This clearly shows the advantages for users in separating the application from resources investment and management.

1.1.2.4. Towards a Big Data Industry

As shown in Table 1.2, we had a database industry in the 1960 to 1990s. At that time most data blocks were measured as MB, GB and TB. Datacenters became widely in use

Table 1.2 Evolution of the big data industry in three development stages.

Stage	Databases	Data Centers	Big Data Industry
Time Frame	1960–1990	1980–2010	2010 and Beyond
Data Sizes	MB – GB -TB	TB – PB - EB	EB – ZB - YB
Market Size and Growth Rate	Database market, Data/Knowledge Engineering	$22.6 B market by IDC 2012, (21.5% growth)	$34 B in IT spending (2013), 4.4 M new big data jobs (2015), Gartner predicts it to exceed 100 B by 2020

from 1980 to 2010, with datasets easily ranging from TB to PB or even EB. After 2010, we saw the gradual formation of a new industry called big data. To process big data in the future, we expect EB to ZB or YB. The market size of the big data industry reached 34 billion in 2013. Exceeding 100 billion in big data applications is within reach by 2020.

1.1.3 Interactive SMACT Technologies

Almost all applications demand computing economics, web-scale data collection, system reliability and scalable performance, such as in the banking and finance industries described above. In recent years, five cutting-edge information technologies: namely Social, Mobile, Analytics, Cloud and IoT, have become more demanding, known as the SMACT technologies. Table 1.3 summarizes the underlying theories, hardware, software and networking advances, and representative service providers of these five technologies. We will study these advances in subsequent chapters.

1.1.3.1. The Internet of Things

The traditional Internet connects machines to machines or web pages to web pages. The IoT refers to the networked interconnection of everyday objects, tools, devices or computers [4]. The things (objects) of our daily life can be large or small. The idea is to tag every object using radio-frequency identification (RFID) or related sensor or electronic technologies like GPS (global positioning system). With the introduction of IPv6 protocol, there are 2^{128} IP addresses available to distinguish all objects on the Earth, including all mobile, embedded devices, computers and even some biological objects. It is estimated that an average person is surrounded by 1000 to 5000 objects on a daily basis.

The IoT needs to be designed to track 100 trillion static or moving objects simultaneously. For this reason, the IoT demands unique addressability of all objects on the Earth. The objects are coded or labeled and IP-identifiable. They are instrumented and interconnected by various types of wired or wireless networks. In some cases they can interact with each other intelligently over the network. The term Internet of Things (IoT) is a physical concept. The size of the IoT can be large or small, covering local regions or a wide range of physical spaces. An IoT is not a just virtual network or logical network or peer-to-peer (P2P) network in cyber space. In other words, the IoTs are built in the physical world, even though they are logically addressable in cyberspace.

Communication among objects can be done in a variety of ways: For example, H2H refers to human-to-human, H2T for human-to-things, T2T for things-to-things, etc. The importance is to connect any things at any time and any place at low cost. By anything connections, we refer to between PCs, H2H (not using PCs but using mobile devices), H2T (using generic equipment) and T2T. By any-place connections, we refer to all the PCs, indoors, outdoors and on the move. By any-time, we imply connections at any time period: day time, night time, outdoor and indoors, and on the move, etc. The dynamic connections will grow exponentially into a new universal network of networks, called IoT. The IoT is strongly tied to specific application domains. Different application domains are embraced by different community circles or groups in our society. We simply call them the IoT domains or IoT networks accordingly.

Table 1.3 SMACT technologies characterized by basic theories, typical hardware, software tooling, networking and service providers needed.

SMACT Technology	Theoretical Foundations	Hardware Advances	Software Tools and Libraries	Networking Enablers	Representative Service Providers
Mobile Systems	Telecommunication, Radio Access Theory, Mobile Computing	Smart Devices, Wireless, Mobility Infrastructures	Android, iOS, Uber, WeChat, NFC, iCloud, Google Player	4G LTE, WiFi, Bluetooth, Radio Access Networks	AT&T Wireless, T-Mobile, Verizon, Apple, Samsung, Huawei
Social Networks	Social Science, Graph Theory, Statistics, Social Computing	Datacenters, Search Engines and WWW Infrastructure	Browsers, APIs, Web 2.0. YouTube, Whatsapp, WeChat, Massager	Broadband Internet, Software- Defined Networks	Facebook, Twitter, QQ, Linkedin, Baidu, Amazon, Taobao
Big Data Analytics	Data Mining, Machine Learning, Artificial Intelligence	Datacenters, Clouds, Search Engines, Big Data Lakes, Data Storage	Spark, Hama, DatTorrent, MLlib, Impala, GraphX, KFS, Hive, Hbase,	Co-Location Clouds, Mashups, P2P Networks, etc.	AMPLab, Apache, Cloudera, FICO, Databricks, eBay, Oracle
Cloud Computing	Virtualization Parallel and Distributed Computing	Server clusters, Clouds, Virtual Machines, Interconnection networks.	OpenStack, GFS, HDFS, MapReduce, Hadoop, Spark, Storm, Cassandra	Virtual Networks, OpenFlow Networks, Software-defined Networks	AWS, GAE, IBM, Salesforce, GoGrid Apache, Azure Rachspace, DropBox
Internet of Things (IoT)	Sensing Theory, Cyber Physics, Navigation, Pervasive Computing	Sensors, RFID, GPS, Robotics, Satellites, Zigbee, Gyroscope	TyneOS, WAP, WTCP, IPv6, Mobile IP, Android, iOS, WPKI, UPnP, JVM	Wireless LAN, PAN, MANET, WLAN Mesh, VANet, Bluetooth	IoT Council, IBM, Healthcare, SmartGrid, Social Media, Smart Earth, Google, Samsung

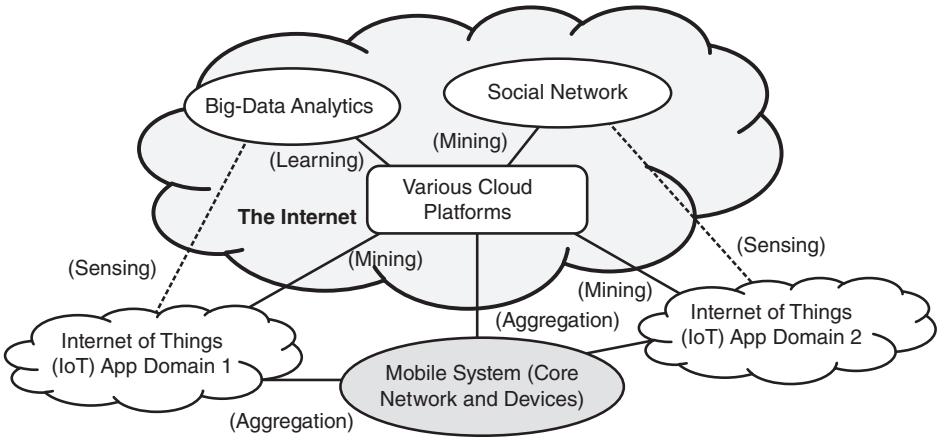


Figure 1.7 Interactions among social networks, mobile systems, big data analytics and cloud platforms over various Internet of Things (IoT) domains.

1.1.3.2. Interactions among SMACT Subsystems

Figure 1.7 illustrates the interactions among the five SMACT technologies. Multiple cloud platforms work closely with many mobile networks to provide the service core interactively. The IoT networks connect any objects including sensors, computers, humans and any IP-identifiable objects on the Earth, The IoT networks appear in different forms in different application domains. The social networks, such as Facebook and Twitter, and big data analytics systems are built within the Internet. All social, analytics and IoT networks are connected to the clouds via the Internet and mobile networks, including some edge networks like WiFi, Ethernet or even some GPS and Bluetooth data.

We need to reveal the interactions among these data-producing, transmission or processing subsystems in the mobile Internet system. In Figure 1.7, we label the edges between subsystems by the actions taking place between them. We briefly introduce below these interactive actions for five purposes: i) data signal sensing is tied to the interactions among IoT and social networks with the cloud platforms; ii) data mining involves the use of cloud power for effective use of captured data; iii) aggregation of data takes place between the mobile system; iv) IoT domains; and vi)the processing clouds. Machine learning forms the basis for big data analytics.

1.1.3.3. Interactions among Technologies

Large amounts of sensor data or digital signals are generated by mobile systems, social networks and various IoT domains. Sensing of RFID, sensor network and GPS generated data is needed to capture the data timely and selectively, if unstructured data were to be disrupted by noises or air loss. IoT sensing demands high quality of data, and filtering is often used to enhance the data quality. Chapter 3 is dedicated to various sensing operation in the IoT system:

- **Data Mining:** Data mining involves the discovery, collection, aggregation, transformation, matching and processing of large datasets. Data mining is a fundamental

operation incurred with the big data information system. The ultimate purpose is knowledge discovery from the data. Both numerical, textual, pattern, image and video data can be mined. Chapter 2 will cover the essence of big data mining in particular.

- **Data Aggregation and Integration:** This refers to data preprocessing to improve data quality. Important operations include data cleaning, removing redundancy, checking relevance, data reduction, transformation and discretization, etc.
- **Machine Learning and Big Data Analytics:** This is the foundation to use cloud's computing power to analyze large datasets scientifically or statistically. Special computer programs are written to automatically learn to recognize complex patterns and make intelligent decisions based on the data. Chapters 4, 5 and 8 will cover machine learning and big data analytics.

1.1.3.4. Technology Fusion to Meet the Future Demand

The IoT extends the Internet of computers to any object. The joint use of clouds, IoT, mobile devices and social networks is crucial to capture big data from all sources. This integrated system is envisioned by IBM researchers as a "smart earth" [22], which enables fast, efficient and intelligent interactions among humans, machines and any objects surrounding us. A smart earth must have intelligent cities, clean water, efficient power, convenient transportation, safe food supplies, responsible banks, fast telecommunications, green IT, better schools, health care and abundant resources to share. This sounds like a dream, which is yet to become a reality in the years to come.

In general, mature technology is supposed to be adopted quickly. The combined use of two or more technologies may demand additional efforts to integrate them for the common purpose. Thus integration may demand some transformational changes. In order to enable innovative new applications, core technology transformation presents a challenge. Disruptive technology is even more difficult to be integrated due to higher risk. They may demand more research and experimentation or prototyping efforts. This takes us on to consider technology fusion by blending different technologies together to complement each other.

All five SMACT technologies are deployed within the mobile Internet (also known as wireless Internet). The IoT networks may appear in many different forms at different application domains. For example, we may build in IoT domains for national defense, healthcare, green energy, social media and smart cities, etc. Social networks and big data analysis subsystems are built in the Internet with fast database search and mobile access facilities. High storage and processing power are provided by domain-specific cloud services on dedicated platforms. We still have a long way to go before we see widespread use of domain-specific cloud platforms for big data or IoT applications in the mobile Internet environment.

1.2 Social-Media, Mobile Networks and Cloud Computing

This section gives an overview of social networks, mobile devices and radio-access networks of all sorts for short-range and wide-range communications and data movement. Social and mobile cloud computing will be assessed. More detailed treatment of these topics can be found in Chapters 4, 7, 8 and 9.

Table 1.4 Summary of popular social networks and web services provided.

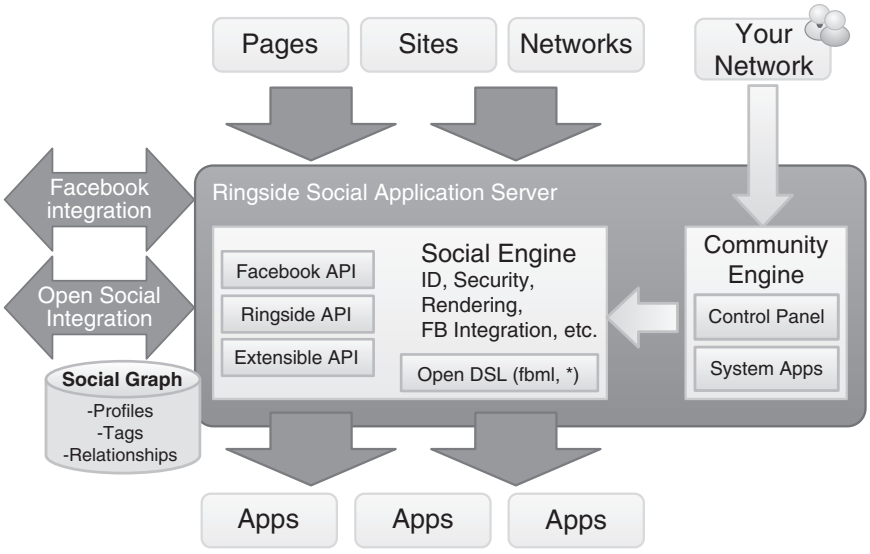
Social Network, Year and Website	Registered Active Users	Major Services Provided
Facebook, *2004* *http://www.facebook.com*	*1.65 billion users, 2016*	*Content sharing, profiling, advertising, events, social comparison, communication, play social games, etc.*
Tencent QQ *in China, 1999* *http://www.qq.com*	*853 million users, 2016*	*An instant messaging service, on-line games, music, ebQQ, shopping, microblogging, movies, WeChat, QQ Player, etc.*
Linkedin, *2002,* *http://www.linkedin.com*	*364 million users, 2015*	*Professional services, on-line recruiting, job listings, group services, skills, publishing, influences, advertising, etc.*
Twitter, *2006* *http://www.twitter.com*	*320 million users, 2016*	*Microblogging, news, alerts, short messages, rankings, demographics, revenue sources, photo-sharing, etc.*

1.2.1 Social Networks and Web Service Sites

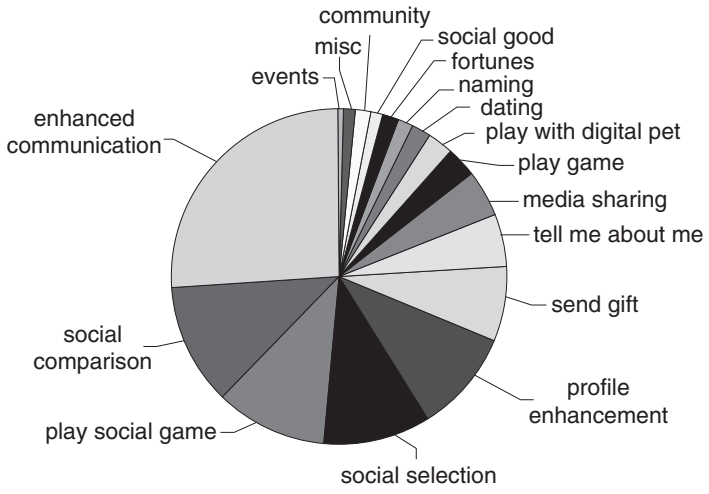
Most social networks provide human services such as friendship connections, personal profiling, professional services, entertainment, etc. In general, the user must register to become a member to access the website. Users can create a user profile, add other users as "friends", exchange messages, post status updates and photos, share videos and receive notifications when others update their profiles. In addition, users may join common-interest user groups, organized by workplace, school or college, or other characteristics, and categorize their friends into lists such as "People From Work" or "Close Friends", etc. In Table 1.4, we compare several popular social networks and introduce their services briefly.

Facebook is by far the largest social networking service provider, with over 1.65 billion users. The Tencent QQ network is the second largest social network based in China. The QQ network has over 800 million users. It is really the Facebook in China, with extended services such as email accounts, entertainment and even some web business operations. Linkedin is a business-oriented social network providing professional services. It is highly used by large business enterprises in recruiting and search for talent. Twitter offers the largest short text message and blogging services today. Other sites are on-line shopping networks or tied to special interest groups.

Example 1.1 Facebook Platform Architecture and Social Services Provided
With 1.65 billion active users worldwide in 2016, Facebook keeps huge personal profiles, tags and relationships as social graphs. Most users are in the US, Brazil, India, Indonesia, etc. The social graphs are shared by various social groups on the site. This website has attracted over 3 millions active advertisers with $12.5 billion revenue reported in 2014. The Facebook platform is built with a collection of huge datacenters with a very large storage capacity, intelligent file systems and searching capabilities. The web must resolve the traffic jams and collisions among all its users. In Figure 1.8(a), the infrastructure of the Facebook platform is shown.



(a) Facebook infrastructure



(b) Facebook application distribution

Figure 1.8 The Facebook platform offering over 2.4 millions of user applications [6].

The platform is formed with a huge cluster of servers. The requests are shown as pages, sites and networks entering the Facebook server from the top. The social engine is the core of the application server. This social engine handles IS, security, rendering and Facebook integration operations. Large numbers of APIs are made available to benefit users to use more than 2.4 millions of applications. Facebook has acquired

Table 1.5 Service functionality of the Facebook platform.

Function	Short Description
Profile Pages	Profile picture, bio information, friends list, user's activity log, public messages
Graph Traversal	Access through users' friends list on profile pages, with access control
Communication	Send and receive messages among friends, instant messaging, and micro blogging
Shared Items	Photo album with built-in access control, embedded outside videos on profile page
Access Control	Access control levels: Only me, Only friends, Friends of friends, and Everyone
Special APIs	Games, calendars, mobile clients, etc.

Insragram, WhatsApp, Qculus VR and PrivateCore applications. The social engine executes all user applications. Open DSL is used to support application executions. The service functionalities of Facebook include six essential items, as summarized in Table 1.5.

Facebook provides blogging, chat, gifts, marketplace, voice/video calls, etc. Figure 1.8(b) shows the distribution of Facebook services. There is a community engine that provides networking services to users. Most Facebook applications are helping users to achieve their social goals, such as improved communication, learning about self, finding similar others, engaging in social play and exchanges. Therefore, Facebook appeals more in the private and personal domains. ■

1.2.2 Mobile Cellular Core Networks

A cellular network or mobile network is a wireless network distributed over land areas called cells, each served by at least one fixed-location transceiver, known as a cell site or base station. In a cellular network, each cell uses a different set of frequencies from neighboring cells, to avoid interference and provide guaranteed bandwidth within each cell. Mobile communications systems have revolutionized the way people communicate, joining together communications and mobility. Figure 1.9 shows the progress of mobile core networks for wide-range communications, having gone through five generations of development, while short-range wireless communication has also upgraded in data rate, QoS and applications during the same period.

Evolution of wireless access technologies has just entered the fourth generation (4G). Looking at the past, wireless access technologies have followed different evolutionary paths aimed at performance and efficiency in a high mobile environment. The first generation (1G) has fulfilled the basic mobile voice communication needs, while the second generation (2G) has introduced the capacity and coverage. The third generation (3G) is a quest for data at higher speeds to open the gates for truly "mobile broadband" experience. The fourth generation (4G) provides access to a wide range of telecommunication services, including advanced mobile services, supported by mobile and fixed networks, which are fully packet switched with high mobility and data rates.

As the mobile communications industry traveled a long way from 2G to 4G, now 5G aims to change the world by connecting anything to anything. Different from its

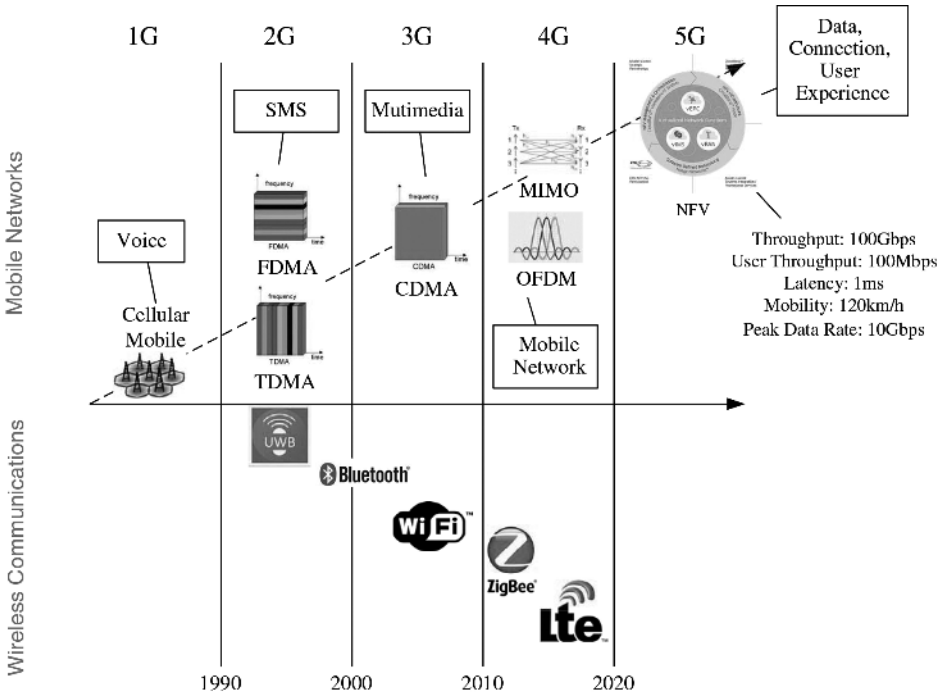


Figure 1.9 Mobile core networks for wide-range communications have gone through five generations, while short-range wireless networks upgraded in data rate, QoS and applications.

previous versions, the research of 5G is not only focusing on new spectrum bands, wireless transmissions, cellular networking, etc., for an increase in capacity. It will be an intelligent technology to interconnect the wireless world without barriers. To meet the requirements of the 5G to enable higher capacity, higher rate, more connectivity, higher reliability, lower latency, larger versatility and application-domain specific topologies, new concepts and design approaches are needed. Current standardization work for 4G may influence the introduction of promising radio features and network solutions for 5G systems.

New network architectures, extending beyond heterogeneous networks and exploiting new frequency spectrum (e.g. mmWave), are emerging from research laboratories around the world. In addition to the network side, advanced terminals and receivers are being developed to optimize network performances. Splitting the control and data planes (currently studied in 3GPP) is an interesting paradigm for 5G, together with massive multi-input multi-output (MIMO), advanced antenna systems, software-defined networking (SDN), Network Functions Virtualization (NFV), Internet of Things (IoT) and cloud computing.

1.2.3 Mobile Devices and Internet Edge Networks

Mobile devices appear as smart phones, tablet computers, wearable gear and industrial tools. Global users of mobile devices exceeded 3 billions in 2015. The 1G devices,

used in the 1980s, were mostly analog phones for voice communication only. The 2G mobile networks began in the early 1990s. Digital phones appeared accordingly for both voice and data communications. As shown in Figure 1.9, 2G cellular networks appear as GSM, TDMA, FDMA and CDMA, based on different division schemes to allow multiple callers to access the system simultaneously. The basic 2G network supports 9.6 Kbps data with circuit switching. The speed was improved to 115 Kbps with packet radio services. Up to 2015, 2G networks were still in use in many developing countries.

Since 2000, 2G mobile devices have been gradually replaced by 3G products. The 3G networks and phones are designed to have 2 Mbps speed to meet the demand of multimedia communications through the cellular system. The 4G LTE (Long Term Evolution) networks appeared in the 2000s. They were targeted to achieve a download speed of 100 Mbps, upload speed of 50 Mbps and a static speed of 1 Gbps. The 3G system is enabled by better radio technology with MIMO smart antennas and OFDM technology. The 3G systems have received widespread deployment now, but could be replaced gradually by 4G networks. We expect the mixed use of 3G and 4G networks for at least another decade. The 5G networks may appear beyond 2020 with a target speed of at least 100 Gbps.

1.2.3.1. Mobile Core Networks

The cellular radios access networks (RAN) are structured hierarchically. Mobile core networks form the backbone of today's telecommunication systems. The core networks have gone through four generations of deployment in the past three decades. The 1G mobile network was used for analog voice communication based on the circuit switching technology. The 2G mobile network started in the early 1990s to support the use of digital telephones in both voice and data telecommunications exploring packet switching circuits. Famous 2G systems are the GSM (Global System for Mobile Communications) developed in Europe and the CDMA (Code Division Multiple Access) system developed in the US. Both GSM and CDMA systems are deployed in various countries.

The 3G mobile network was developed for multimedia voice/data communications with global roaming services. The 4G system started in the early 2000s based on the LTE and MIMO radio technologies. The 5G mobile networks are still under heavy development, which may appear in 2020. The technology, peak data rate and driven applications of the five generations of cellular mobile networks are summarized in Table 1.6. Speedwise, the mobile systems improved from 1 Kbps to 10 Kbps, 2 Mbps and

Table 1.6 Milestone mobile core networks for cellular telecommunication.

Generation	1 G	2 G	3 G	4 G	5 G
Radio and Networks Technology	Analog phones, AMPS, TDMA	Digital phones GSM, CDMA	CDMA2000, WCDMA, and D-SCDMA	LTE, OFDM, MIMO, software-steered radio	LTE, Cloud-based RAN
Peak Mobile Data Rate	8 Kbps	9.6–344 Kbps	2 Mbps	100 Mbps	10 Gbps–1 Tbps)
Driving Applications	Voice Communication	Voice/Data Communication	Multimedia Communication	Wide band Communication	Ultra-speed Communication

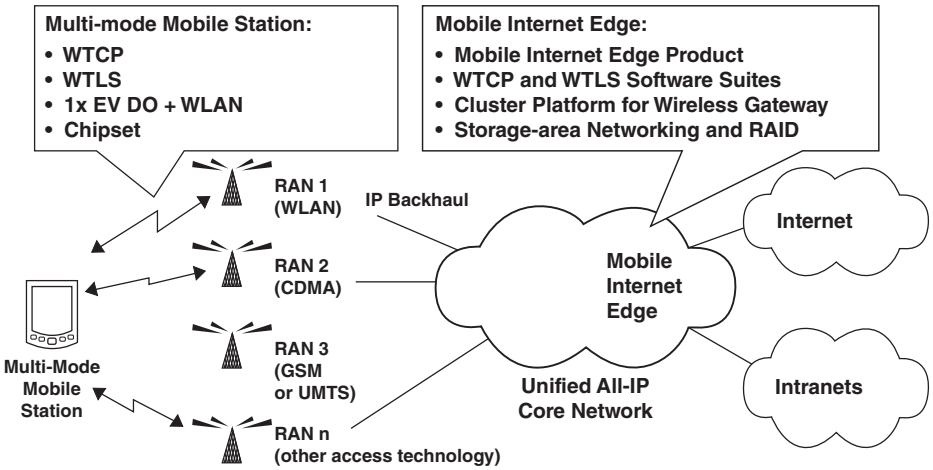


Figure 1.10 The interactions of various radio-access networks (RANs) with the unified all-IP based mobile core network, Intranets and the Internet.

100 Mbps in four generations. It is projected that the upcoming 5G system may achieve a 1000 increase in data rate to 100 Gbps or higher. The 5G system may be built with remote radio head (RRH) and virtual base stations installed in CRAN (Cloud-based Radio Access Networks).

1.2.3.2. Mobile Internet Edge Networks

Most of today's wireless and mobile networks are based on radio signal transmission and reception at various operating ranges. We call them radio access networks (RANs). Figure 1.10 illustrated how various RANs are used to access the mobile core networks, which are connected to the Internet backbone and many Intranets through mobile Internet edge networks. Such an Internet access infrastructure is also known as the wireless Internet or mobile Internet by the pervasive computing community. In what follows, we introduce several classes of RANs known as WiFi, Bluetooth, WiMax and Zigbee networks. Generally speaking, we consider several short-range wireless networks, such as wireless local-area network (WLAN), wireless home-area network (WHAN), personal-area network (PAN) and body-area network (BAN), etc. These wireless networks play a key role in mobile computing and IoT applications.

1.2.3.3. Bluetooth Devices and Networks

Bluetooth is a short-range radio technology named after a Danish King dating back to the 9th century. A Bluetooth device operates in 2.45 GHz industrial scientific medical band, as specified by IEEE 802.15.1 Standard. It transmits omni-directional (360°) signals with no limit on line of sight, meaning data or voice can penetrate solid non-metal object. It supports up to 8 devices (1 master and 7 slaves) in a PAN called *Piconet.* Bluetooth devices have low cost and low power requirements. The device offers a data rate of 1 Mbps in *ad hoc* networking with 10 cm to 10 meters in range. It supports voice or data communication between phones, computers and other wearable

devices. Essentially, Bluetooth wireless connections are replacing most wired cables between computers and their peripherals such as mouse, keyboard and printers, etc.

1.2.3.4. WiFi Networks

WiFi access point or WiFi networks are specified in the IEEE 802.11 Standard. So far, they have appeared as a Series of 11 a, b, g, n, ac and ay networks. The access point broadcasts its signal in a radius of less than 300 ft. The closer it is to the access point, the faster will be the data rate experienced. The maximum speed is only possible within 50–175 ft. The peak data rates of WiFi networks have improved from less than 11 Mbps in 11b to 54 Mbps in 11g and 300 Mbps in 11n networks. The 11n and 11ac network applies OFDG modulation technology with the use of multiple input and multiple output (MIMO) radio and antenna to achieve its high speed. WiFi enables the fastest WLAN in a mesh of access points or wireless routers. They offer almost free access to the Internet within 300 ft at many locations today.

1.2.4 Mobile Cloud Computing Infrastructure

Mobile devices are rapidly becoming the major service participants nowadays. There is a shift of user preferences from traditional cell phones and laptops to smart phones and tablets. Advances in the portability and capability of mobile devices, together with widespread 3G/4G LTE networks and Wi-Fi accesses, have brought rich mobile application experiences to the end users. Mobile cloud computing is a model for elastic augmentation of mobile device capabilities via ubiquitous wireless access to cloud storage and computing resources. This is further enhanced by context-aware dynamic adaption to the changes in the operating environment. Figure 1.11 shows a typical mobile environment for offloading large jobs to remote clouds from mobile device holders.

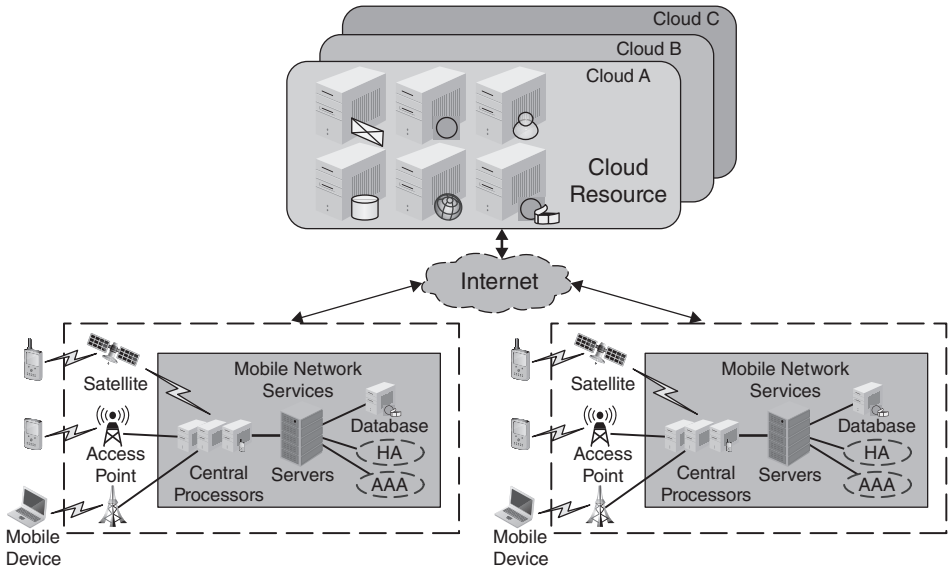


Figure 1.11 The architecture of a mobile cloud computing environment.

With the support of mobile cloud computing (MCC), a mobile user basically has a new cloud option to execute its application. The user attempts to offload the computation through WiFi, cellular network or satellite to the distant clouds. The terminal devices at the user end have limited resources, i.e. hardware, energy, bandwidth, etc. The cellphone itself is infeasible to finish some compute-intensive tasks. Instead, the data related to the computation task is offloaded to the remote cloud. Special cloudlets were introduced to serve as wireless gateways between mobile users and the Internet. These cloudlets can be used to offload computations or web services to remote clouds safely. Details of cloudlets for mobile clouds will be described in Chapter 2.

1.3 Big Data Acquisition and Analytics Evolution

Big data analytics is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in higher business intelligence or scientific discovery, such as more effective marketing, increased revenue, etc. The primary goal of big data analytics is to help companies make better business decisions by enabling data scientists and other users to analyze huge volumes of transaction data that may be left untapped by conventional business intelligence (BI) programs.

1.3.1 Big Data Value Chain Extracted from Massive Data

Data science, data mining, data analytics and knowledge discovery are closely related terms. In many cases, they are used interchangeably. These big data components form a big data value chain built up of statistics, machine learning, biology and kernel methods. Statistics cover both linear and logistic regression. Decision trees are typical machine learning tools. Biology refers to artificial neural networks, genetic algorithms and swarm intelligence. Finally, the kernel method includes the use of support vector machines. These underlying theories and models will be studied in Chapters 4, 5 and 6. Their applications will be covered in Chapters 7, 8 and 9.

Compared with traditional datasets, big data generally includes masses of unstructured data that need more real-time analysis. In addition, big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and incurs new challenges, for example on how to effectively organize and manage such data. At present, big data has attracted considerable interest from industry, academia and government agencies. Recently, the rapid growth of big data mainly comes from people's daily life, especially related to the Internet, Web and cloud services.

Example 1.2 Expected Growth of Big Data from 2010 to 2020 and Economic Gains
The size of a big dataset varies with time. Table 1.7 shows some representative data sizes from 2010 to 2020. These numbers simply give the reader the impression of a steady increase with time. By the 2015 standard, a dataset of 1 TB is labeled as big data. The massive volumes bring in values in economic gains as revealed in the bottom rows. Just location sensitive applications can deliver $800 billion revenue in 10 years. Big data

Table 1.7 The Growth of big data from 2010 to 2020 and expected economic value in two typical big data applications.

Big Data Sources, Observed Growth in Data Size and Year	Data Size
Global data generated in 2 days in 2011	1.82 ZB
Pictures uploaded to Facebook in 2014	759 million pieces
Storage capacity of American manufacturing industry in 2010	966 PB
Number of RFID tags scanned in 2011 to 2020	12 million to 209 billion
Data captured during a computer geek's 2.450 million hours	200 TB
Personal location data services could reach the level of $800 billion dollars in 10 years	
Savings from healthcare analysis and treatment could exceed $300 billion dollars in the US	

applied for healthcare may save $300 billions in medical expenses in the US. The big data industry is gradually shaping up over time. ■

At present, data has become an important production factor that could be comparable to material assets and human capital. As multimedia, social media and IoT are fast evolving, business enterprises will collect more information, leading to an exponential growth of data volume. Big data will have a huge and increasing potential in creating values for businesses and consumers. The most critical aspect of big data analytics is big data value. We divide the value chain of big data into four phases: data generation, data acquisition, data storage and data analysis. If we take data as a raw material, data generation and data acquisition are exploitation processes, as data storage must use clouds or datacenters. Data analysis is a production process that utilizes the raw material to create new value.

The rapid growth of cloud computing and IoT also triggers the sharp growth of data. Cloud computing provides safeguarding, access sites and channels for data assets. In the paradigm of IoT, sensors worldwide are collecting and transmitting data to be stored and processed in the cloud. Such data in both quantity and mutual relations will far surpass the capacities of the IT architectures and infrastructure of existing enterprises, and its real-time requirement will greatly stress the available computing capacity. The following example highlights some representative big data values driven by the massive data volume involved.

1.3.1.1. Big Data Generation

The major data types include Internet data, sensory data, etc. This is the first step of big data. Given Internet data as an example, huge amounts of data in terms of searching entries, Internet forum posts, chatting records and microblog messages, are generated. Those data are closely related to people's daily lives, and have similar features of high value and low density. Such Internet data may be valueless individually but, through the exploitation of accumulated big data, useful information such as habits and hobbies of users can be identified, and it is even possible to forecast users' behaviors and emotional moods.

Moreover, generated through longitudinal and/or distributed data sources, datasets are more large-scale, highly diverse and complex. Such data sources include sensors, videos, clickstreams, and/or all other available data sources. At present, main sources of

big data are the operation and trading information in business enterprises, logistics and sensing information in the IoT, human interaction information and position information in the Internet world, and data generated in scientific research, etc.

1.3.2 Data Quality Control, Representation and Database Models

In Table 1.8, we summarize interesting properties and attributes that affect the data quality. We introduce the methods, architectures and tools for big data analysis. Our studies by no means cover all progress made in this field. We identify the key concepts and some representative tools or database models used in this context. Big data sources come from business transactions, textual and multimedia contents, qualitative knowledge data, scientific discovery, social media and sensing data from IoT. The quality of the data is often poor due to the massive volume, data variety due to unpredictable data types, and data veracity for lack of traceability.

The quality control of big data involves a circular cycle of four stages: i) we must identify the important data quality attributes; ii) to access the data relies on the ability to measure or assess the data quality level; iii) then we must be able to analyze the data quality and their major causes; and finally iv) we need to improve the data quality by suggesting concrete actions to take. Unfortunately, none of these tasks are easy to implement. In Table 1.8, we identify the important attributes towards data quality control. Among these data quality control dimensions, the intrinsic attributes and representational and access control mechanisms are equally important.

Data can be represented in many different ways. Four major representation models are suggested for big data: i) the <key, value> pairs is often used to distribute data in MapReduce operations (to be presented in Chapter 5). The Dynamo Volldemort is a good example to use key-value pairs; ii) table lookup or relational database such as

Table 1.8 Attributes for data quality control, representation and database operations.

Category	Attributes	Basic Definitions and Questions To Ask
Intrinsic and Contextual	Accuracy and Trust	Data correctness and credibility: true, fake or accurate?
	Integrity and Reputation	Biased or impartial data? Reputation of data source?
	Relevance and Value	Data relevance to task at hand and value added or not?
	Volume and Completeness	Data volume tested and any value present?
Representation	Easy to Comprehend	Data clarity and easy to understand without ambiguity?
	Interpretability and Visualization	Data well represented in numbers, textual, graphs, image, video, profiles or metadata, etc.?
Accessibility and Security	Access Control	Data availability, access control protocols, easy to retrieve?
	Security Precautions	Restricted access or integrity control from alteration or deletion?

Google's BigTable and Cassandra software; iii) graphic tools like GraphX used in Spark for social graph analysis, and iv) special database systems such as MongoDB, SimpkleDB and CouchDB commonly used by the big data community.

1.3.3 Big Data Acquisition and Preprocessing

Loading is the most complex procedure among the three, which includes operations such as transformation, copy, clearing, standardization, screening and data organization. A virtual database can be built to query and aggregate data from different data sources, but such a database does not contain data. On the contrary, it includes information or metadata related to actual data and its positions. Such two "storage-reading" approaches do not satisfy the high performance requirements of data flows or search programs and applications. Compared with queries, data in such two approaches is more dynamic and must be processed during data transmission.

Generally, data integration methods are accompanied with flow processing engines and search engines:

1) **Data Selection:** Select a target dataset or subset of data samples on which the discovery is to be performed.
2) **Data Transformation:** Simplify the datasets by removing unwanted variables. Then analyze useful features that can be used to represent the data, depending on the goal or task.
3) **Data Mining:** Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
4) **Evaluation and knowledge representation:** Evaluate knowledge pattern, and utilize visualization techniques to present the knowledge vividly.

1.3.3.1. Big Data Acquisition

This includes data collection, data transmission and data pre-processing. As the second phase, data acquisition also includes data collection, data transmission and data pre-processing. During big data acquisition, once we collect the raw data, we utilize an efficient transmission mechanism to send it to a proper storage management system to support different analytical applications. The collected datasets may sometimes include much redundant or useless data, which unnecessarily increases storage space and affects the subsequent data analysis. Table 1.9 summarizes major data acquisition methods and preprocessing operations.

For example, high redundancy is very common among datasets collected by sensors for environmental monitoring. Data compression technology can be applied to reduce the redundancy. Therefore, data pre-processing operations are indispensable to ensure

Table 1.9 Some big data acquisition sources and major preprocessing operations.

Collection Sources	Logs, sensors, crawlers, packet capture and mobile devices, etc.
Preprocessing Steps	Integration, Cleaning and Redundancy elimination
Data Generators	Social Media, Enterprises, Internet of Things, Internet, Bio-Medical, Government, scientific discovery, environments, etc.

efficient data storage and exploitation. Data collection is to utilize special data collection techniques to acquire raw data from a specific data generation environment. Many common data collection and generation sources and data generators are introduced below.

1.3.3.2. Log Files

CAs are one widely used data collection method, and log files are record files automatically generated by the data source system, so as to record activities in designated file formats for subsequent analysis. Log files are typically used in nearly all digital devices. For example, web servers record in log files the number of clicks, click rates, visits, and other property records of web users. To capture activities of users at the websites, web servers mainly include the following three log file formats: public log file format (NCSA), expanded log format (W3C) and IIS log format (Microsoft).

All three types of log files are in the ASCII text format. Databases other than text files may sometimes be used to store log information to improve the query efficiency of the massive log store. There are also some other log files based on data collection, including stock indicators in financial applications and determination of operating states in network monitoring and traffic management.

1.3.3.3. Sensors

Sensors are common in daily life to measure physical quantities and transform physical quantities into readable digital signals for subsequent processing (and storage). Sensory data may be classified as sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature, etc. Sensed information is transferred to a data collection point through wired or wireless networks, for applications that may be easily deployed and managed, for example video surveillance system.

1.3.3.4. Methods for Acquiring Network Data

At present, network data acquisition is accomplished using a combination of web crawler, word segmentation system, task system and index system, etc. Web crawler is a program used by search engines for downloading and storing web pages [28]. Generally speaking, web crawler starts from the uniform resource locator (URL) of an initial web page to access other linked web pages, during which it stores and sequences all the retrieved URLs. Web crawler acquires a URL in the order of precedence through a URL queue and then downloads web pages, and identifies all URLs in the downloaded web pages, and extracts new URLs to be put in the queue.

This process is repeated until the web crawler is stopped. Data acquisition through a web crawler is widely applied in applications based on web pages, such as search engines or web caching. Traditional web page extraction technologies feature multiple efficient solutions and considerable research has been done in this field. As more advanced web page applications are emerging, some extraction strategies are used to cope with rich Internet applications. The current network data acquisition technologies mainly include traditional Libpcap-based packet capture technology, zero-copy packet capture technology, as well as some specialized network monitoring software such as Wireshark, SmartSniff and WinNetCap.

1.3.3.5. Big Data Storage

Big data storage refers to the storage and management of large-scale datasets while achieving reliability and availability of data accessing. The explosive growth of data has more strict requirements on data storage and management. We consider the storage of big data as the third component of big data science. The storage infrastructure needs to provide information storage service with reliable storage space, and it must provide a powerful access interface for query and analysis of a large amount of data.

Considerable research on big data promotes the development of storage mechanisms for big data. Existing storage mechanisms of big data may be classified into three bottom-up levels: file systems, databases and programming models. File systems are the foundation of the applications at upper levels. Google's GFS is an expandable distributed file system to support large-scale, distributed, data-intensive applications. GFS uses cheap commodity servers to achieve fault tolerance and provides customers with high-performance services. GFS supports large-scale file applications with more frequent reading than writing. However, GFS also has some limitations, such as a single point of failure and poor performances for small files. Such limitations have been overcome by Colossus, the successor of GFS.

In addition, other companies and researchers also have their solutions to meet the different demands for storage of big data. For example, HDFS and Kosmosfs are derivatives of open source codes of GFS. Microsoft developed Cosmos to support its search and advertisement business. Facebook utilizes Haystack to store the large amount of small-sized photos. Taobao also developed TFS and FastDFS. In conclusion, distributed file systems have become relatively mature after years of development and business operation. Therefore, we will focus on the other two levels in the rest of this section.

1.3.3.6. Data Cleaning

Data cleaning cleanses and preprocesses data by deciding strategies to handle missing fields and alter the data as per the requirements. Data cleaning is a process to identify inaccurate, incomplete or unreasonable data, and then to modify or delete such data to improve data quality. Generally, data cleaning includes five complementary procedures: defining and determining error types, searching and identifying errors, correcting errors, documenting error examples and error types, and modifying data entry procedures to reduce future errors.

During cleaning, data formats, completeness, rationality and restriction should be inspected. Data cleaning is of vital importance to keep data consistency, which is widely applied in many fields, such as banking, insurance, retail industry, telecommunications and traffic control. In e-commerce, most data is electronically collected, which may have serious data quality problems. Classic data quality problems mainly come from software defects, customized errors or system mis-configuration. Some consider data cleaning in e-commerce by using crawlers and regularly re-copying customer and account information.

The problem of cleaning RFID data is examined next. RFID is widely used in many applications, for example inventory management and target tracking. However, the original RFID features low quality, which includes a lot of abnormal data limited by the physical design and affected by environmental noise. The probabilistic model was developed to cope with data loss in mobile environments. We could build a system to automatically correct errors of input data by defining global integrity constraints.

1.3.3.7. Data Integration

Data integration is the cornerstone of modern commercial informatics, which involves the combination of data from different sources and provides users with a uniform view of data. This is a mature research field for traditional database. Historically, two methods have been widely recognized: data warehouse and data federation. Data warehousing includes a process named ETL (Extract, Transform and Load). Extraction involves connecting source systems, selecting, collecting, analyzing and processing necessary data. Transformation is the execution of a series of rules to transform the extracted data into standard formats. Loading means importing extracted and transformed data into the target storage infrastructure.

1.3.4 Evolving Data Analytics over the Clouds

Big data analytics is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in higher business intelligence or scientific discovery, such as more effective marketing, increased revenue, etc. Big data sources must be protected in web server logs and Internet clickstream data, social media activity reports, mobile-phone call records and information captured by sensors or IoT devices. Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics and data mining.

In Figure 1.12, we specify the goals and requirements of today's cloud analytics evolved from the basic analysis used in handling small data in the past. In the past, we handled "small data" objects in terms of MB to GB, as shown on the left-hand side of Figure 1.12. On the *x*-axis, we evolve from small data to "big data", which ranges from TB to PB based

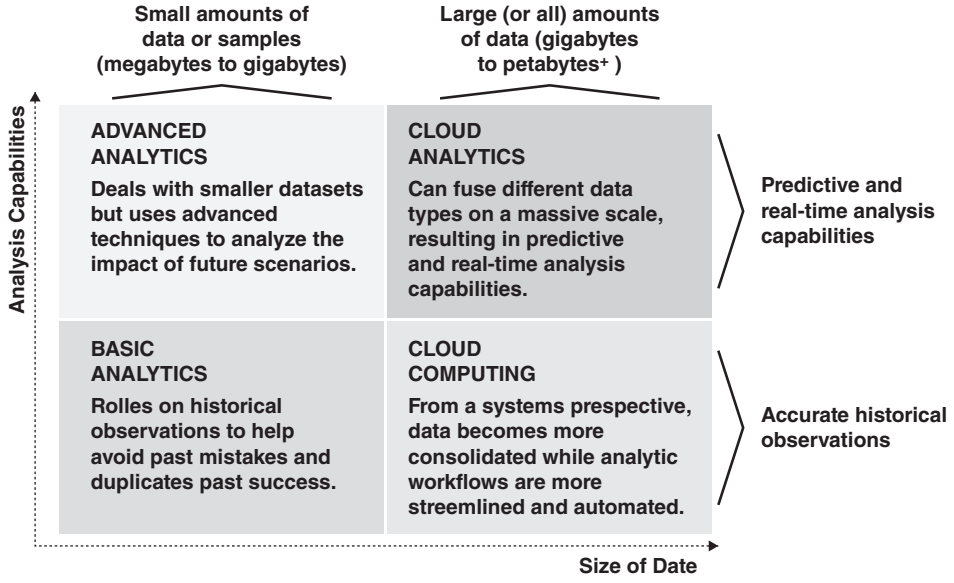


Figure 1.12 The evolution from basic analysis of small data (MB to GB) in the past to sophisticated cloud analytics over today's big datasets (TB~PB).

on the 2015 standard. On the *y*-axis, we show the analytics capability in two ascending levels: accurate historical observations versus predictive and real-time analysis capabilities. The performance space is divided into four subspaces:

1) The **basic analysis** of small data relies on historical observations to help avoid past mistakes and duplicate past successes.
2) The **advanced analytics** system on small data is improved from the basic capability to use advanced techniques to analyze the impact of future scenarios.
3) As we move to **cloud computing**, most existing clouds provide a better coordinated analytics workflow in a streamlined and automated fashion, but still lack predictive or real-time capabilities.
4) For an ideal **cloud analytics** system, we expect to handle scalable big data in streaming mode with real-time predictive capabilities.

Traditional data analysis means the use of proper statistical methods to analyze massive first-hand data and second-hand data, to concentrate, extract and refine useful data hidden in a batch of chaotic data, and to identify the inherent law of the subject matter, so as to develop functions of data to the greatest extent and maximize the value of the data. Data analysis plays a huge guidance role in development plans for a country, as well as understanding customer demands and predicting market trends by business enterprises. Big data analysis can be deemed as the analysis of a special kind of data. Therefore, many traditional data analysis methods may still be utilized for big data analysis. Several representative traditional data analysis methods are examined in the following, many of which are from statistics and computer science.

In general, we build a cloud for big data computing with a layered structure, as illustrated in Figure 1.13. At the bottom layer, we have the cloud infrastructure management

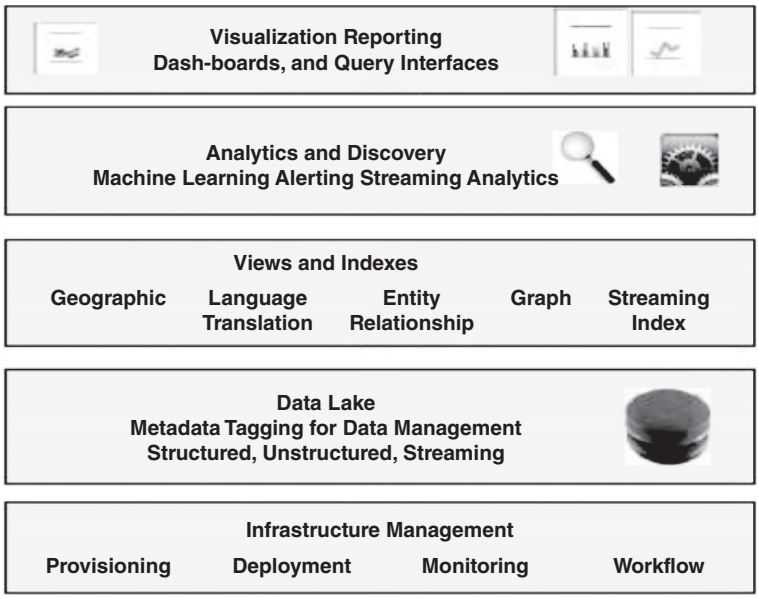


Figure 1.13 Layered development of cloud platform for big data processing and analytics applications.

control, which handles resources provisioning, deployment of agreed resources, monitoring the overall system performance and arranging the workflow in the cloud. All big data elements collected from all sources form the data lake. Data could be structured or unstructured or come-and-go in streaming mode. This lake stores not only raw data but also the metadata for data management.

At the middle layer, we need to provide views and indexes to visualize and access data smoothly. This may include geographic data, language translation mechanisms, entity relationship, graphs analysis and streaming index, etc. At the next higher level, we have the cloud processing engine which includes data mining, discovery and analytics mechanisms to perform machine learning, alerting, and data stream processing operations. At the top level, we have to report or display the analytics results. This includes visualization support for reporting with dashboards and query interfaces. The display may take the form of histograms, bar graphs, charts, video, etc.

1.4 Machine Intelligence and Big Data Applications

In this section, we try to link machine intelligence to big data applications. Machine intelligence is attributed to smart clouds applied IoT sensing, and data analytics capabilities. First, we reveal the relationship between data mining and machine learning. Then we give an overview of important big data applications. Finally, we present the key concept of cognitive computing and their applications.

1.4.1 Data Mining and Machine Learning

We classify data mining into three categories: association analysis, classification and cluster analysis. Machine learning techniques are divided into three categories: supervised learning, unsupervised learning and other learning methods including reinforcement learning, active learning, transfer learning and deep learning, etc.

1.4.1.1. Data Mining versus Machine Learning

Data mining and machine learning are closely related to each other. Data mining is the computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The overall goal of the data-mining process is to extract information from a dataset and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization and online updating.

Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions. These two terms are commonly confused, as they often employ the same methods and overlap significantly. Machine learning is closer to applications and end user. It focuses on prediction, based on known properties learned from the training data. As shown in Figure 1.14, we divide machine learning techniques into three categories: i) supervised learning such as regression model,

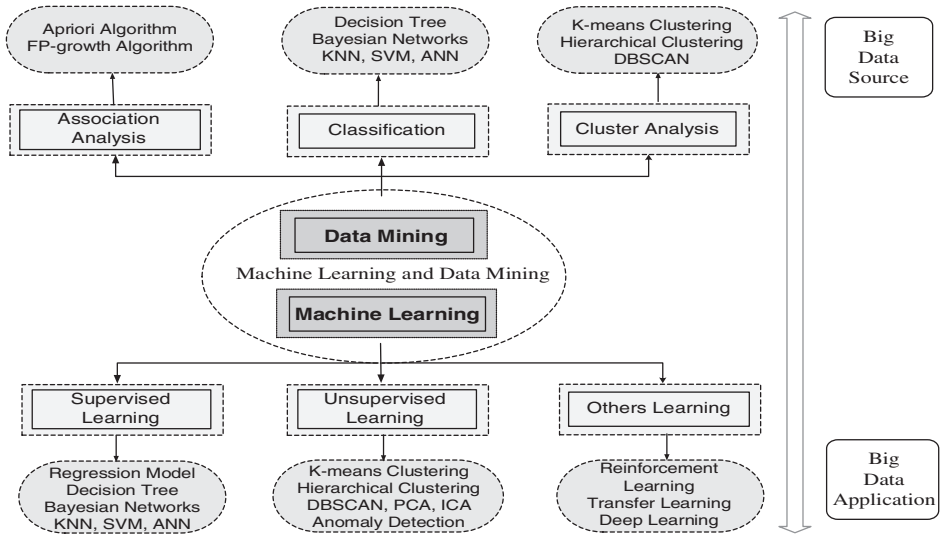


Figure 1.14 The relationship of data mining and machine learning.

decision tree, etc.; ii) unsupervised learning, which includes clustering, anomaly detection, etc.; and iii) other learning, such as reinforcement learning, transfer learning, active learning and deep learning, etc.

Data mining is closer to the data source. It focuses on the discovery of unknown properties of the data, which is also considered as the analysis step of knowledge discovery in databases. As shown in Figure 1.14, the typical data mining techniques are classified into three categories: i) association analysis includes Apriori algorithm and FP-growing algorithm; ii) classification algorithm includes decision tree, support vector machine (SVM), k-nearest-neighbor, Naïve Bayesian, Bayesian belief network and artificial-neural-network (ANN), etc.; and iii) clustering algorithm includes K-means and density-based spatial clustering of applications with noise.

The analysis of big data is confronted with many challenges but the current research is still in the beginning phase. Considerable research efforts are needed to improve the efficiency of data display, data storage and data analysis. The research community demands a more rigorous definition of big data. We demand structural models of big data, formal description of big data, and a theoretical system of data sciences, etc. An evaluation system of data quality and an evaluation standard of data computing efficiency should be developed.

Many solutions of big data applications claim they can improve data processing and analysis capacities in all aspects, but there exists no unified evaluation standard and benchmark to balance the computing efficiency of big data with rigorous mathematical methods. The performance can only be evaluated by an implemented and deployed system, which could not horizontally compare advantages and disadvantages of various solutions. Before and after the use of big data, the efficiencies are also hard to compare. In addition, since data quality is an important basis of data preprocessing, simplification and screening, it is another urgent problem to effectively evaluate data quality.

The emergence of big data triggers the development of algorithm design, which has transformed from a computing-intensive approach into a data-intensive approach. Data transfer has been a main bottleneck of big data computing. Therefore, many new computing models tailored for big data have emerged and more such models are on the horizon. Machine intelligence is critical to solve challenging issues existing in big data applications. Machine intelligence is obtained through machine learning.

- **Supervised Machine Learning**: includes the following categories:
  a) **Regression Model:** Decision Tree, SVM;
  b) **Bayesian Classifier:** Hidden Markov Model;
  c) **Deep Learning:** to be explained in Chapter 8.
- **Unsupervised Machine Learning**: includes:
  a) **Dimension Reduction:** Principal component analysis (PCA);
  b) **Clustering:** finding a partition of the observed data in the absence of explicit labels indicating a desired partition. Chapter 7 will be devoted to these unsupervised models.
- **Other Machine Learning techniques**:
  a) **Reinforcement Learning: Markov decision processes (MDPs)** provide a mathematical framework for modeling decision making in situations where the outcomes are partly random and partly under the control of a decision maker.
  b) **Transfer Learning:** Through transfer learning, the time-consuming and labor-intensive processing costs can be reduced extensively. After a certain time of labeling and validation through transfer learning, the training sets are established. Among various key big data technologies, machine intelligence is the key component. The machine learning techniques for big data computing will be studied in detail in Chapters 3, 4 and 5.

1.4.2 Big Data Applications – An Overview

A large number of big data applications have been reported in the literature. We will treat big data and cloud applications in Part III of this book. Here in Table 1.10, we simply give a global overview of various big data applications. The US National Institute of Standards and Technology (NIST) has identified 52 application cases of big data application. These tasks are grouped into 9 categories. As a matter of fact, many data driven applications have emerged in the past two decades. For example, business intelligence has become a prevailing technology in business applications. Network search engines are based on a massive data-mining process. We briefly introduce these applications as follows:

1.4.2.1. Commercial Applications

The earliest business data was generally structured data, which was collected by companies from old systems and then stored in RDBMSs. Analytical technologies used in such systems prevailed in the 1990s and were intuitive and simple, for example reports, instrument panels, special queries, search-based business intelligence, online transaction processing, interactive visualization, score cards, predictive modeling and data mining. Since the beginning of the 21st century, networks and websites have provided a unique opportunity for organizations to have online display and directly interact with customers.

Table 1.10 Application categories of big data: from TBs to PBs (NIST 2013).

Category	Brief Description	Example Applications
Government	National Archives and Records, Federal/State Administration, Census Bureau, etc.	CIA, FBI, Police Forces, etc.
Business and Commercial	Finance in Cloud, Cloud Backup, Mendeley (Citations), Web Search, Digital Materials, etc.	Netflix, Cargoing shipping, On-line shopping, P2P
Defense and Military	Sensors, Image surveillance, Situation Assessment, Crisis Control, Battle management, etc.	Pentagon, Home Security Agency
Health Care and Life Science	Medical records, Graph and Probabilistic analysis, Pathology, Bioimaging, Genomics, Epidemiology, etc.	Body-Area Sensors, Genomics, Emotion control
Deep Learning, Social media	Self-driving car, geolocate images/cameras, Crowd Sourcing, Network Science, NIST benchmark datasets	Machine learning, Pattern Recognition, Perception, etc.
Scientific Discovery	Sky Surveys, Astronomy and Physics, polar science, Radar Scattering in Atmosphere, Metadata, Collaboration, etc.	Large Hadron Collider at CERN and Belle Accelerator in Japan
Earth Environment	Earthquake, Ocean, Earth Observation, Ice sheet Radar scattering, Climate simulation datasets, Atmospheric turbulence identification, Biogeochemistry.	AmeriFlux and FLUXNET gas sensors, IoT for smart earth
Energy Research	New Energy Resources, Wind power, Solar systems, green computing, etc.	SmartGrid Project

Abundant products and customer information, including click stream data logs and user behavior, etc., can be acquired from the websites. Product layout optimization, customer trade analysis, product suggestions and market structure analysis can be conducted by text analysis and website mining technologies. The quantity of mobile phones and tablet PC first surpassed that of laptops and PCs in 2011. Mobile phones and Internet of Things based on sensors are opening a new generation of innovation applications, and searching for larger capacity of supporting location sensing, people oriented and context operation.

1.4.2.2. Network Applications

The early Internet mainly provided email and webpage services. Text analysis, data mining and webpage analysis technologies have been applied to the mining of email content and building search engines. Nowadays, most applications are web-based, regardless of their application field and design goals. Network data accounts for a major percentage of the global data volume. Web has become the common platform for interconnected pages, full of various kinds of data, such as text, images, videos, pictures and interactive content, etc. Advanced technologies are in great demand in semi-structured or unstructured data.

For example, the image analysis technology may extract useful information from pictures, for example face recognition. Multimedia analysis technologies are applied to automated video surveillance systems for business, law enforcement and military applications. Online social media applications, such as Internet forums, online communities, blogs, social networking services and social multimedia websites, etc., provide users with great opportunities to create, upload and share content. Different user groups may search for daily news and publish their opinions with timely feedback.

1.4.2.3. Big Data in Scientific Applications

Scientific research in many fields is acquiring massive data with high-throughput sensors and instruments, such as astrophysics, oceanology, genomics and environmental research. The US National Science Foundation (NSF) has recently announced the BIGDATA Research Initiative to promote research efforts to extract knowledge and insights from large and complex collections of digital data. Some scientific research disciplines have developed massive data platforms and obtained useful outcomes.

For example, in biology, iPlant applies network infrastructure, physical computing resources, coordination environment, virtual machine resources, and inter-operative analysis software and data service to assist research, educators and students, in enriching all plant sciences. iPlant datasets have high varieties in form, including specification or reference data, experimental data, analog or model data, observation data and other derived data. Big data has been applied in the analysis of structured data, text data, website data, multimedia data, network data and mobile data.

1.4.2.4. Application of Big data in Enterprises

At present, big data mainly comes from and is mainly used in business enterprises, while BI and OLAP can be regarded as the predecessors of big data application. The application of big data in business enterprises can enhance their production efficiency and competitiveness in many aspects. In particular, in marketing, with correlation analysis of big data, business enterprises can accurately predict the behavior of consumers.

On sales planning, after comparison of massive data, business enterprises can optimize their commodity prices. On operation, such enterprises can improve their operation efficiency and operation satisfaction, optimize the input of the labor force, accurately forecast personnel allocation requirements, avoid excess production capacity and reduce labor costs. On supply chain, using big data, business enterprises may conduct inventory optimization, logistic optimization and supplier coordination, etc., to mitigate the gap between supply and demand, control budgets and improve services.

Example 1.3 Banking Use of Big Data in Financing and e-Commerce Applications
In the finance community, the application of big data has grown rapidly in recent years. For example, China Merchants Bank utilizes data analysis to recognize that such activities as "Multi-times score accumulation" and "score exchange in shops," are effective for attracting quality customers. By building a customer loss early warning model, the bank can sell high-yield financial products to the top 20% customers in loss ratio so as to retain them. As a result, the loss ratios of customers with Gold Cards and Sunflower Cards have been reduced by 15% and 7%, respectively.

By analyzing customers' transaction records, potential small and micro-corporate customers can be effectively identified. Utilizing remote banking, cloud referral platforms can help implement cross-selling, and considerable performance gains have been observed in recent years. Obviously, the most classic application is in e-commerce. Tens of thousands of transactions are conducted in Taobao and the corresponding transaction time, commodity prices and purchase quantities are recorded every day.

More importantly, such information matches age, gender, address and even hobbies and interests of buyers and sellers. Data Cube of Taobao is a big data application on the Taobao platform, through which merchants can be aware of the macroscopic industrial status of the Taobao platform, market conditions of their brands and consumers' behaviors, etc., and accordingly make production and inventory decisions. Meanwhile, more consumers can purchase their favorite commodities at more preferable prices.

The credit loan of Alibaba automatically analyzes and judges whether to provide loans to business enterprises through the acquired enterprise transaction data by virtue of big data technologies, while manual intervention does not occur in the entire process. It is disclosed that, so far, Alibaba has lent more than RMB 30 billion Yuan, with the rate of bad loans at only about 0.3\%, which is a great deal lower than those of other commercial banks. ■

1.4.2.5. Healthcare and Medical Applications

The healthcare industry is growing rapidly and medical data is a continuously and rapidly growing complex data, containing abundant and various information values. Big data has unlimited potential for effectively storing, processing, querying and analyzing medical data. The application of medical big data will profoundly influence human health. The IoT is revolutionizing the healthcare industry. Sensors collect patient data, then microcontrollers process, analyze and communicate the data over wireless Internet. Microprocessors enable rich graphical user interfaces. Healthcare clouds and gateways help analyze the data with statistical accuracy. A few simple examples are given below. More on health care IoT and big data applications will be studied in Chapters 4, 5, 8 and 9.

Example 1.4 Big data Applications in Healthcare Industry

Aetna Life Insurance Company selected 102 patients from a pool of 1000 patients to complete an experiment to help predict the recovery of patients with metabolic syndrome. In an independent experiment, it scanned 600,000 laboratory test results and 180,000 claims through a series of detection test results of metabolic syndrome of patients in three consecutive years. In addition, it summarized the final result into an extreme personalized treatment plan to assess the dangerous factors and main treatment plans of patients.

In this way, doctors may reduce morbidity by 50% in the next 10 years, by prescribing statins and helping patients to lose weight by as much as 5 lb, or suggesting patients should reduce the total triglyceride in their bodies if the sugar content in their bodies is over 20%. The Mount Sinai Medical Center in the US utilizes technologies of Ayasdi, a big data company, to analyze all genetic sequences of *Escherichia Coli*, including over 1 million DNA variants, to know why bacterial strains resist antibiotics. Ayasdi's technology uses Topological data analysis, a brand new mathematic research method, to understand data characteristics.

HealthVault of Microsoft offers an excellent application of medical big data launched in 2007. The goal is to manage individual health information in individual and family medical equipment. Presently, health information can be entered and uploaded with mobile smart devices and imported into individual medical records by a third-party agency. In addition, it can be integrated with a third-party application with the software development kit (SDK) and open interface. ■

1.4.2.6. Collective Intelligence

With the rapid development of wireless communication and sensor technologies, mobile phones and tablet computers have integrated more and more sensors, with increasingly stronger computing and sensing capacities. As a result, crowd sensing is taking to the center stage of mobile computing. In crowd sensing, a large number of general users utilize mobile devices as basic sensing units to conduct coordination with mobile networks for distribution of sensed tasks and collection and utilization of sensed data. The goal is to complete large-scale and complex social sensing tasks. In crowd sensing, participants who complete complex sensing tasks do not need to have professional skills.

Crowd sensing modes represented by Crowdsourcing have been successfully applied to geotagged photograph, positioning and navigation, urban road traffic sensing, market forecasting, opinion mining and other labor-intensive applications. Crowdsourcing, a new approach to problem solving, takes a large number of general users as the foundation and distributes tasks in a free and voluntary way. Crowdsourcing can be useful for labor-intensive applications, such as picture marking, language translation and speech recognition.

The main idea of Crowdsourcing is to distribute tasks to general users and to complete tasks that users could not individually complete or do not anticipate to complete. With no need for intentionally deploying sensing modules and employing professionals, Crowdsourcing can broaden the sensing scope of a sensing system to reach the city scale and even larger scales. Crowdsourcing was applied by many companies before the emergence of big data. For example, P&G, BMW and Audi improved their R&D and design capacities by virtue of Crowdsourcing.

In the big data era, Spatial Crowdsourcing is a hot topic. The operation framework of Spatial Crowdsourcing is shown as follows. A user may request the service and resources related to a specified location. Then the mobile users who are willing to participate in the task will move to the specified location to acquire related data (i.e. video, audio or pictures). Finally, the acquired data will be sent to the service requester. With the rapid growth of mobile devices and the increasingly complex functions provided by such devices, it is forecast that Spatial Crowdsourcing will be more prevalent than traditional Crowdsourcing, for example Amazon Turk and Crowdflower.

1.4.3 Cognitive Computing – An Introduction

The term cognitive computing is derived from cognitive science and artificial intelligence. For years, we have wanted to build a "computer" that can compute as well as learn by training, to achieve some human-like senses or intelligence. It has been called a "brain-inspired computer" or a "neural computer". Such a computer will be built with special hardware and/or software, which can mimic basic human brain functions such

as handling fuzzy information and perform affective, dynamic and instant responses. It can handle some ambiguity and uncertainty beyond traditional computers.

To this end, we want a cognitive machine that can model the human brain with the cognitive power to learn, memorize, reason and respond to external stimulus, autonomously and tirelessly. This field has been also called "neuroinformatics". Cognitive computing hardware and applications could be more affective and influential by design choices to make a new class of problems computable. Such a system offers a synthesis, not just of information sources but of influences, contexts and insights. IBM describes the systems that learn at scale, reason with purpose and interact with humans.

Cognitive computing hardware and applications could be more affective and influential by design choices to make a new class of problems computable. Such a system offers a synthesis not just of information sources but of influences, contexts and insights. In other words, cognitive computing systems make some well-defined "context" computable. IBM describes the systems that learn at scale, reason with purpose and interact with humans naturally.

1.4.3.1. System Features of Cognitive Computing

In a way, a cognitive system redefines the relationship between humans and their pervasive digital environment. They may play the role of assistant or coach for the user, and they may act virtually autonomously in many situations. The computing results of a cognitive system could be suggestive, prescriptive or instructive in nature. Listed below are some characteristics of cognitive computing systems:

- **Adaptive in learning:** They may learn as information changes, and as goals and requirements evolve. They may resolve ambiguity and tolerate unpredictability. They may be engineered to feed on dynamic data in real time, or near real time.
- **Interactive with users:** Users can define their needs as a trainer of the cognitive system. They may also interact with other processors, devices and cloud services, as well as with people.
- **Iterative and stateful**: They may redefine a problem by asking questions or finding additional source input if a problem statement is ambiguous or incomplete. They may "remember" previous interactions iteratively.
- **Contextual in information discovery**: They may understand, identify and extract contextual elements such as meaning, syntax, time, location, appropriate domain, regulations, user's profile, process, task and goal. They may draw on multiple sources of information, including both structured and unstructured digital information, as well as sensory inputs such as visual, gestural, auditory or sensor provided.

1.4.3.2. Differences with Current Computers

Cognitive systems differ from current computing applications in that they move beyond tabulating and calculating based on preconfigured rules and programs. Although they are capable of basic computing, they can also infer and even reason based on broad objectives. Cognitive computing systems can be extended to integrate or leverage existing information systems and add domain or task-specific interfaces and tools. Cognitive systems leverage today's IT resources and coexist with legacy systems into the future. The ultimate goal is to bring computing even closer to human thinking and become a fundamental partnership in human endeavors.

Table 1.11 Related fields to neuroinformatics and cognitive computing.

Subject Areas	Brief Description of The Field	Technology Support
Artificial Intelligence	Study of cognitive phenomena to implement the human intelligence in computers	Pattern recognition, robotics, computer vision, speech processing, etc.
Learning and Memory	Study of human learning and memory mechanisms and build them in future computers	Machine learning, database systems, memory enhancement, etc.
Languages and Linguistics	Study of how linguistic and language are learned and acquired, and how to understand novel sentences	Language and speech processing, machine translation, etc.
Perception and Action	Study of the ability to take in information via the senses such as vision and hearing, etc. Haptic, olfactory and gustatory stimuli fall into this domain	Image recognition and understanding, behavioral science, brain imaging, psychology and anthropology
Neuro-informatics	Neuroinformatics stands at the intersection of neuroscience and information science	Neurocomputers, artificial neural nets, deep learning, aging, disease control, etc.
Knowledge Engineering	The study of big data analysis, knowledge discovery, transformation and creativity process	Datamining, data analytics, knowledge discovery and system construction

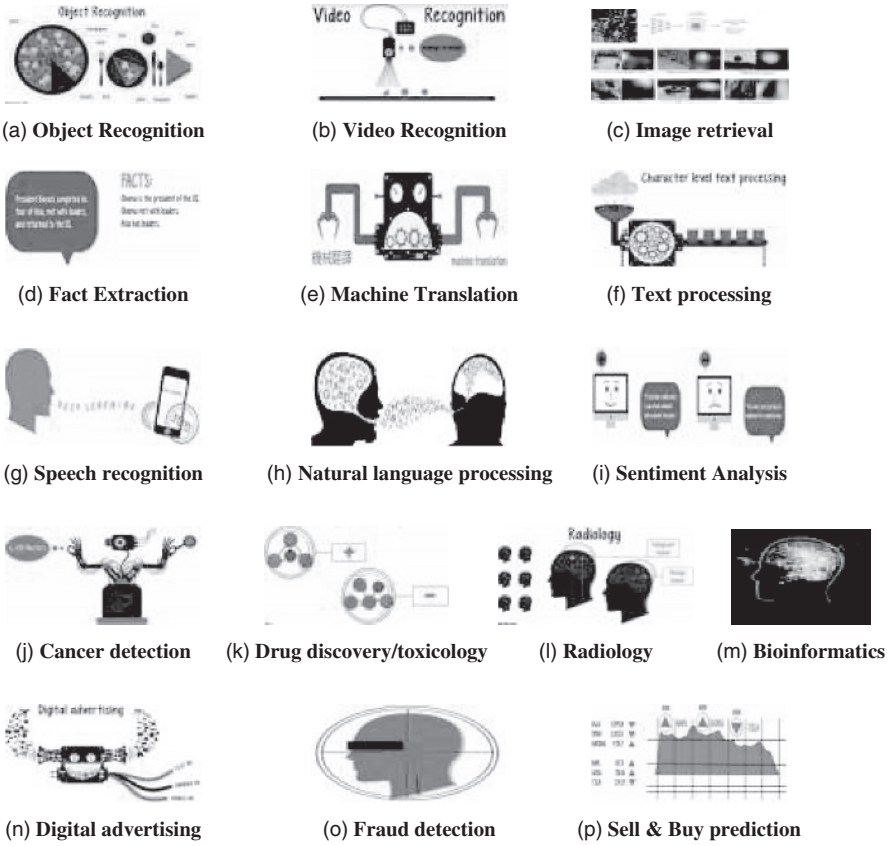
Cognitive science is interdisciplinary in nature. It covers the areas of psychology artificial intelligence, neuroscience and linguistics, etc. It spans many levels of analysis from low-level machine learning and decision mechanisms to high-level neural circuitry to build brain-modeled computers. Related fields to neuroinformatics and cognitive computing are summarized in Table 1.11. In Chapter 3 and subsequent chapters, we will further explore these cutting-edge technologies.

1.4.3.3. Applications of Cognitive Machine Learning

Cognitive computing platforms have emerged and become commercially available, and evidence of real-world applications is starting to surface. Organizations have adopted and used these cognitive computing platforms for the purpose of developing applications to address specific use cases, with each application utilizing some combination of available functionality. Examples of such real-world cases include: i) speech understanding; ii) sentiment analysis; iii) face recognition; iv) election insights; v) autonomous driving; and vi) deep learning applications. Many more examples are available in cognitive computing services. These demystify the possibilities into real-world applications. Figure 1.15 lists all important cognitive machine learning applications.

Among these big data applications:

a) object recognition;
b) video interpretation;
c) image retrieval;



(a) **Object Recognition** (b) **Video Recognition** (c) **Image retrieval**

(d) **Fact Extraction** (e) **Machine Translation** (f) **Text processing**

(g) **Speech recognition** (h) **Natural language processing** (i) **Sentiment Analysis**

(j) **Cancer detection** (k) **Drug discovery/toxicology** (l) **Radiology** (m) **Bioinformatics**

(n) **Digital advertising** (o) **Fraud detection** (p) **Sell & Buy prediction**

Figure 1.15 Machine and deep learning applications classified in 16 categories.

are related to machine vision applications. Text and document tasks include:

a) fact extraction;
b) machine translation; and
c) text comprehension.

On the audio and emotion detection side, we have:

a) speech recognition;
b) natural language processing, and
c) sentiment analysis tasks.

In medical or healthcare applications, we have:

a) cancer detection;
b) drug discovery;
c) toxicology and radiology; and
d) bioinformatics.

Additional information on cognitive machine learning applications can be found on the youtube website: www.youtube.com/playlist?list =PLjJh1vlSEYgvGod9wWiydumYl8hOXixNu

In business and financial applications, we have (n) digital advertising, (o) fraud detection and (p) sell and buy prediction in market analysis. Many of these cognitive tasks are awaiting automation. Some of the identified applications involve plenty of raw data in text of trillions of words in various languages, visual data in billions of images and videos, audio in 400 days of speech, user queries and marketing messages, plus knowledge and social media graphs in billions of labeled tuples.

1.5 Conclusions

This chapter introduces the basic definitions and key concepts of big data science and cognitive computing. The purpose is to prepare our readers for studying the in-depth treatment in subsequent chapters. Smart clouds are supported by IoT sensing and big data analytics. More coverage of smart clouds is given in Chapters 3, 4 and 9. We emphasize the interactions or fusion of the SMACT technologies for big data processing. Social-media networking and mobile access of the cloud services are introduced in Section 1.2. These topics will be further studied in Chapters 2, 3, 8 and 9. We introduce basics of data mining, machine learning, data analytics and cognitive computing in Sections 1.3 and 1.4. These big data topics are further studied in subsequent chapters.

Homework Problems

1.1 Briefly characterize the differences in the following computing paradigms and technologies: Clouds, datacenters, virtualization, supercomputers, Internet technologies, web services, utility computing and service computing:
- **a)** Cloud computing versus supercomputing applications?
- **b)** What are the similarity and differences between clouds and datacenters?
- **c)** The conventional Internet versus the Internet of things?
- **d)** What is utility computing versus service computing?
- **e)** Why is virtualization crucial to the use of today's clouds?

1.2 Hype cycle is updated every year. You have learned from Figure 1.4 about the progress up to July 2016. Check with the Wikipedia with the latest Gartner Report on Hype Cycle and discuss the new changes compared with the 2016 report.

1.3 This homework requires you to do some research. Write an updated assessment report of the SMACT technologies. Discuss the strength and weakness and pros/cons of each technology. You need to dig out a few relevant technical reports or white papers from relevant industries, especially from major industrial players such as Facebook, AT&T, Google, Amazon and IBM, etc. Reading some published papers at leading *ACM/IEEE Magazines or Conferences* would be useful to make insightful assessment with concrete evidence.

1.4 Compare conventional on-premise desktop computing with the three cloud service models: IaaS, PaaS and SaaS. The resources and user application software are divided in five categories: user applications, virtual machines, servers, storage and networking. Each resource category could be controlled by user, vendor, or shared between user and vendor. Indicate appropriate control labels as User, Vendor and Shared in these four computing models. Justify your labels with reasoning.

1.5 After studying the basic concept of mobile clouds in Section 1.2.4, try to answer the following questions with an updated survey of the major providers of mobile cloud services. Checking Wikipedia under mobile clouds or mobile cloud computing may be helpful in finding the answers. Additional information can be also found in *IEEE MobileCloud Conferences* or the Special issue on Mobile Cloud Computing in *IEEE Transactions on Cloud Computing or Service Computing.*

1.6 Briefly explain the problems (challenges) associated with four "V's" of big data characteristics. (1) Volume, (2) Velocity, (3) Variety, and (4). Veracity. Discuss the resources demand and associated processing requirement and limitations.

1.7 In data science, what are the intersection field of the application domain expertise and the field of mathematics or statistics background? Also explain the intersection field of programming skills and the required mathematics or statistics background.

1.8 Consider the following two cloud/IoT service applications, and find out more examples from the literature about smart cities in Example 1.5 and healthcare cloud services in Example 1.4. You need to report your findings on how machine learning and big data analytics can help out in their success stories.

1.9 Explain why big data can be more cost-effectively handled by clouds than by using supercomputers. Why big data scientists require domain expertise. Also explain the differences in supervised versus unsupervised machine learning techniques.

1.10 In Figure 1.3, we have identified a number of software tools provided by various companies or research centers. Consider the following three software packages: The MatLab library for computing algorithms, the UCI machine learning repository for data analytics, and the OpenNLP for natural language processing. Find out from their websites or the literature about their functionalities and usage requirements for big data computing.

References

1 B. Baesens, *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Wiley, 2015.
2 L. Barroso and U. Holzle, The datacenter as a computer: An introduction to the design of warehouse-scale machines. In: *Synthesis Lectures on Computer architecture*. M. Hill (ed.), Morgan Claypool, 2009.

3 R. Buyya, J. Broberg and A. Goscinski (eds), *Cloud Computing: Principles and Paradigms*. Wiley Press, US, February 2011.
4 H. Chaouchi, *The Internet of Things*. Wiley, 2010.
5 M. Chen, *Big Data Related Technologies*. Springer Computer Science Series, 2014.
6 S. Farnham, *The Facebook Association Ecosystem*. O'Reilly Radar Report, 2008.
7 J. Gobbi, R. Buyya, S. Marusic and M. Palaniswarni, Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer System*s, 29, 1645–1660, 2013.
8 J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Third Edition. Morgan Kaufmann, 2012.
9 U. Hansmann, et al., *Pervasive Computing: The Mobile World*, Second Edition. Springer, 2003.
10 T. Hey, S. Tansley and K. Tolle (eds), *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
11 M. Hilber and P. Lopez, The world's technological capacity to store, communicate and compute information. *Science*, 332(6025), 2011.
12 K. Hwang, G. Fox and J. Dongarra, *Distributed and Cloud Computing*. Morgan Kaufamann, 2011,
13 K. Hwang and D. Li, Trusted cloud computing with secure resources and data coloring. *IEEE Internet Computing*. October 2010.
14 R, Liu, *Introduction to Internet of Things*. Science Press, Beijing, 2011.
15 H. Karau, et al., *Learning Spark: Lightning Fast Data Analysis*. O'Reily, 2015.
16 M. Rosenblum and T. Garfinkel, Virtual machine monitors: current technology and future trends. *IEEE Computer*, May 2005, 39–47.
17 S. Ryza, et al., *Advanced Analytics with Spark*. O'Reily, 2015.
18 J. Smith and R. Nair, *Virtual Machines: Versatile Platforms for Systems and Processes*. Morgan Kaufmann, 2005.
19 Hype Cycle, http://www.gartner.com/newsroom/id/2819918
20 Mark Weiser, The computer for the 21st century. *Scientific American*, 1991.
21 Y. Li, and W. Wang, Can mobile cloudlets support mobile applications? *IEEE INFOCOM*, April 2014, 1060–1068.
22 J. Kelley, III, Computing, Cognition and The Future of Knowing, IBM Corp, October 2015. http://www.research.ibm.com/software/IBM Research/multipmedia/Computing_Cognition_WhitePaper.pdf