

Chapter 2

Framework

Abstract A framework of the applied data-centric social sciences is based on data-centric science. A methodology of data-centric science is very common and applicable to all the types of sciences. In this chapter, we will see a methodology used in applied data-centric sciences commonly.

2.1 Pipelines of Data-Centric Science

Generally, the data-centric investigation or data-driven study is constructed from the following steps:

- problem definition
- project design
- an explanatory data analysis
- data acquisition
- data collection
- data analysis
- interpretation
- decision-making

These steps construct a cycle to improve data quality, interpretation adequateness and effectiveness of decision-making. In order to understand the data-generating mechanism, we also should visit actual spots where the data are generated and confirm correspondence between the data and objects or concepts which they express. In general, the problem and project are unknown firstly. We may not clearly understand the problem which we need to solve and the project where we should work. To understand them, it is useful to come in touch with data of the problem or of the field where the project will be built. This type of activity is called *explanatory data analysis* [34].

In both the inductive and deductive approaches, in general, we face the so-called chicken-and-egg problem. This is a kind of causality dilemma. The problem definition

and the project building sometimes face the causality dilemma. This means that we need to build a project to understand details of the problem related to the project. The abductive approach may solve this causality dilemma. The abductive approach is defined as an approach to start to think the problem on a hypothesis. A plenty of data often becomes a starting point of the hypothesis. The explanatory data analysis provides us with knowledge on phenomena and characteristics of the problem.

2.2 Purpose, Goal and Proposal

We need a purpose or purposes for our research activity in order to justify our own activity. When we start our research or project, we have to ask ourselves whether we can improve our society, add new information to existing studies, solve a societal problem or propose a new policy to decision-makers. Our activity may simply create knowledge on a societal issue. Then, we need to consider how to contribute to our society by increasing knowledge on the societal issue.

The goal is different from the purpose. The goal of our research activity or our project should be concrete with some quantitative measures. For example, in the case of a business, the goal should be defined as a measurable improvement such as the number of consumers, the duration time for production, and so on. In the case of the academic research, the goal should find new things and/or propose new concepts or methods with some quantitative manners. How many or how much do we improve the activity or clarify the phenomena? To do so, we firstly need to grasp our current situation and determine an area where we can make a difference.

2.3 Project Design

In order to design our project or research activity, we must ask ourselves the following questions again and again during the project:

- What is our question?
- How do we ask and solve the question?
- What data do we need to answer the question?

We often start our project without any concrete goals. However, such a launch seems to create some problems during our own research activity. For example, imagine that we do not know what we should achieve and how to examine the issue. How do we feel about this situation?

Actually, we need to find a goal for our activity. Asking ourselves several questions may lead us to a concrete goal. To find an adequate question, a bird's eye view of the problem or phenomenon may help us. For example, we can ask ourselves as follows:

1. What is our research field?
2. Do we find any gaps between existing studies and general questions?
3. What is our focus?

4. What kinds of questions can we ask in our focus?
5. What can we expect to contribute based on our resources and skills?
6. Can we find any relationship between our standing point and the questions?

These questions may also help us to find a concrete goal for our project:

- What types of data do we need?
- How many or much data do we need?
- What types of data do we need to reach our goal?

Through these questions, we find a way to investigate our object to reach our goal.

How do you feel from these questions? You may not find any concrete answers to these questions. The main reasons why you cannot find any answers are because:

- a lack of information in the fields
- a lack of knowledge on the problem
- a lack of skills to solve the problem
- a lack of resources of the research

Then, you need to have an experience to treat the (even small) data on the problem or the field at least to find a concrete answer. You can start your explanatory data analysis from acquiring a small amount of data related to the field which you want to contribute to. And then, you will be able to find better answers to the problem.

Furthermore, during our research activity, it is important for us to often check whether our activity is adequate? To do so, it is useful to record logs of our own activities. In fact, documents or memos of our own activities help us to confirm our research activity. The research diary may be useful for this purpose. We can write our activity in research notes with dates. The software and procedures for computations should be also recorded. We need to check our activity during our research project repeatedly.

2.4 Data Acquisition

The data is recorded from some data-generating source. The data of society are currently available from web pages. Both personal and official web pages are a preliminary data source of our society. Electronic commerce systems are also sources of data for products and services. We can accumulate data on prices for goods and services from application programming interface (API) of some data providers. Data of financial markets, job opportunities, hotels, flights, traffic and so on are accumulated via Web API nowadays.

Web API is an application programming interface which can be used via the Internet. In Web API, there are several technologies to exchange commands and data between an API provider and users. Functions of natural language processing, geographical information systems (GIS), search engines and databases of e-commerce services are available as Web APIs. This list shows several examples:

- Yahoo! JAPAN text analytics WebAPI [37]
- Jalan vacant room information retrieval WebAPI [18]

- AB-ROAD travel retrieval WebAPI [2]
- Rakuten Web Service WebAPI [25]
- Google Translate WebAPI [14]

The secondary source of data is a sensor network. Several types of sensors have recently become available. Some of them can send data to a database server via the Internet directly. We require sensors that convert physical parameters to electrical signals. The sensor signals are converted into a form that can be converted to digital values. Analog-to-digital converters are included in the sensors. The sensors are connected with one another through wired or wireless network. This is sometimes referred to as Internet of things (IoT).

Machine-to-machine (M2M) solution is one of the implementations of IoT, which is provided from several vendors. Functional requirements of the M2M application are as follows:

- There are data that can be exchanged between a device and a server.
- There are device management capabilities provided by an M2M application.
- There are different components of which an M2M application is made.

Data management of the M2M application includes hierarchical structure of data elements. The data type can be associated with the data elements. Primitive data types such as string, integer, double, date, Boolean and byte array are supported. Users can define constraints for the data, identify the protocol to be used when exchanging a given data element and configure parameters to protocols. There are some commands that can be sent by a server to a device, and sets of events that can be sent by a device to a server.

2.5 Data Collection

Data collection is the process of gathering information. In the data collection, several types of sampling methods are known:

- simple random sampling
- systematic sampling
- snowball sampling
- comprehensive sampling

A simple random sampling means that we obtain a subset of individuals chosen from a larger set. Each piece of data is chosen randomly and has the same probability to be chosen at any stages.

A systematic sampling is to sample data according to some ordering scheme and then select elements at regular intervals through that ordered list, for example, selecting every 10th name from the telephone directory.

A snowball sampling is often used in sociology and statistics research. Snowball sampling is non-probability sampling where existing data recruits the potential data

which will be sampled in the future. Therefore, the sample group appears to grow like a rolling snowball. For example, suppose that we accumulate data of web pages from the World Wide Web. In this case, firstly, we choose a web page. Next, we select a page from a link included in the sampled web page. Repeating this procedure, we eventually collect data of web pages.

A comprehensive sampling means that we obtain all the data that we can cover. If we have sufficient computer resources and time, then we can conduct the comprehensive sampling.

The data are stored in computer systems as digital files such as CSV, TSV, XML and so on. These files can be inserted into database servers, which play an important role in data collection. A database management system (DBMS) is at the core of data collection. Some types of DBMS are recently available:

- relational database
- XML database
- object-oriented database
- document-oriented database

We need to handle several types of databases at the same time. In the data collection, we may need to determine the area of the data. A relational database management system (RDBMS) have a high affinity with CSV and TSV formats. XML formats can be transformed to CSV or TSV formats and can be handled by RDBMS. XML databases can be used to handle XML-formatted data directly.

Furthermore, we need to carefully consider the way to prevent data loss. The data loss badly affects results of data analysis and generate additional data acquisition and costs. Intentional and accidental deletion of files or data damages collections of data. Using a journaling file system and Redundant Arrays of Independent Disks (RAID) storage can protect against some types of software and hardware failure. Regular data backups are an effective method to recover the data from data-loss events. In fact, user errors or system failures cannot be prevented by regular backups, but we may quickly recover the system from such failures if we keep several versions of backups.

2.6 Data Validation

Data validation is one of the most important but the most time-consuming tasks [28]. Without clean data, data analysis and optimisation tools cannot work well. Data analysis and optimisation solutions always assume the presence of correct data. In many cases presence/inference of wrong data is even worse than absence of the data, and a harmful effect in decision-making will happen. Therefore, it is an important step for any researcher to verify and validate the accuracy and adequateness of the data. There are several types of validation methods:

- multiplexing data sources
- consistency check
 - item count validation test
 - range validation test
- finding outliers

Multiplexing data sources may help us to find data inconsistency. For example, suppose that we use macroeconomic statistics such as population or GDP. Then, we should collect the same data from two institutions. We can compare the same data elements obtained from the difference institutions at least. If quantities in the elements are different from each other, then we can understand that one of them is wrong or contains some error. This technique is, of course, applicable to other areas than macroeconomic data.

Consistency check is a common method for several types of data. In this case, a physical model of data-generating mechanism is useful. For example, causality, time and space can be used for this purpose. There are two types of data errors: systematics errors and random errors. Systematics errors can result from bugs of software to generate data or procedures. Thus, when they occur at all, they occur repeatedly. Systematic errors can produce three types of errors: (1) too many data elements, (2) too few data elements and (3) classification of data elements. The primary action of the data validation is to identify the occasions when systematic errors happen. (1) and (2) can be checked if we count the number of data elements. This is called *item count validation test*. (3) can be confirmed by checking the types of data and range of data. This is called *range validation test*. The range validation test is done by checking that all records are within specified ranges.

Random errors are generated as input errors or judgement errors. In general, random errors occur intermittently. This type of error can be detected as an outlier from other values. Both the range validation and item count validation tests can be used to detect random errors.

Range validation test is sometimes useful if the data are numeric or one of several options. If a data element is out of range, then we can determine that it is wrong data. When we use geographical information, we can use distance from a position as a norm of data. We may find incorrect data as some outliers from a relation of feature to the distance. When we use time series data, time order can be used to check the data consistency. If the time order is contrary or missing, then we may find incorrect data or missing data.

Outliers are defined as a data point that extremely differs from other data points. Ben-Gal [5] classifies outlier detection methods into univariate statistical methods and multivariate outlier detection. The earliest univariate methods for outlier detection use the assumption of an underlying known distribution of the data. An outlier can be detected by using mean of values included in a dataset and their standard deviation. If we assume that the values are sampled from a normal distribution, then the probability where the samples appear between the mean minus three times the standard deviation and the mean plus three times the standard deviation is 99.9%. Therefore, the values deviating from this range are detected as outliers. Barnett and Lewis [3]

showed statistical methods to identify outliers (Chauvenet’s criterion, Grubbs’s test for outliers [15], Peirce’s criterion [23], Dixon’s Q test [9], Thompson test). In the multivariate case, *Mahalanobis* distance [21] can be used. The *Mahalanobis* distance [21] for each multivariate data point \mathbf{x}_s ($s = 1, \dots, T$) is defined as

$$M_s = \left((\mathbf{x}_s - \bar{\mathbf{x}})^T \mathbf{V}_n^{-1} (\mathbf{x}_s - \bar{\mathbf{x}}) \right)^{1/2}, \quad (2.1)$$

where \mathbf{V} represents the sample covariance matrix defined as

$$\mathbf{V} = \frac{1}{T-1} \sum_{s=1}^T (\mathbf{x}_s - \bar{\mathbf{x}})(\mathbf{x}_s - \bar{\mathbf{x}})^T, \quad (2.2)$$

and $\bar{\mathbf{x}}$ the sample mean. A large value of M_s for the s -th data point indicates that it is an outlier.

During the data analysis, we may often find some outliers. In this case, we should check whether the outliers are consistent with the mechanism to generate the data or not.

Data quality problems are recognised as important tasks in data engineering. Detecting and removing errors and inconsistencies from data improve the quality of data. These tasks are called “data cleaning”. There is a big range of data cleaning commercial tools available in the market. Some of those are more generic in operation and others are solving a specific problem in a particular domain. Rahm and Do [24] also propose five phases of data cleaning approaches to construct an automated data-cleaning system for data warehousing:

- Data analysis
- Definition of transformation work flow and mapping rules
- Verification
- Transformation
- Backflow of cleaned data

The data analysis is needed in order to detect which kinds of errors and inconsistencies are to be removed. The definition of schema-related data transformations and mapping rules for data elements should be considered. The correctness and effectiveness of a transformation work flow and the transformation definitions should be tested and evaluated. The transformation steps are executed. After errors are removed, the cleaned data should also replace the dirty data in the original sources. To data quality management (DQM) to the data loaded in the system, we need define DQM rules that perform a variety of repair, clean up, and standardisation functions on incoming identity data values. These functions are implemented in recent Big Data

Analytics solutions such as IBM Netezza,¹ SAP Data Quality Management software² and Talend Enterprise Data Quality solution.³

2.7 Explanatory Data Analysis

One of the most important steps in the investigation is an explanatory analysis [34]. This is a kind of feasible study. The explanatory data analysis consists of the following steps:

1. visualise data and compute fundamental statistics
2. construct a model from ideas obtained from statistical analysis
3. estimate model parameters
4. validate or check an adequacy of the model with parameter estimates
5. interpret data using the estimated model
6. repeat 1–5 until we are satisfied with the interpretation

Concretely, this procedure can be drawn as

1. Make scatter plots between variables, draw time series, networks and spatial plots, and make a histogram from observations. From the plots, we can find some patterns and detect outliers of data. If we need a new axis of data, we define it or additionally start to collect data from environment. It is useful to compute descriptive statistics (mean, variance, quartile, skewness, and kurtosis). Changing granularity of data or spatio-temporal scales we need to compute these fundamental properties of data.
2. Applying methods of multivariate analysis (regression analysis, principal component analysis, spatial regression, and factor analysis) and time series analysis (autoregressive analysis) we need to determine relationship among variables and their temporal transitions (transition probabilities). These processes provide us with ideas on data generating mechanisms and stochastic models as an approximation of actual mechanism.
3. Repeating step 1 and step 3, we increase kinds of data and accumulate the number of observations as well as ideas of models for variables.
4. Realising the model, we attempt to estimate model parameters from observations. If we use a regression model, we will check goodness-of-fit of data for the model in terms of an explained variable and explanatory variables. A degree of freedom of the model is determined by using some criteria such as information criteria. Data bias (sampling bias, processing bias, and so on) should be taken into account in this step. During this step, we sometimes recognise a fault of data acquisition or data collection.

¹ IBM Netezza: <http://www-01.ibm.com/software/data/netezza/>.

² SAP Data Quality Management software: <http://www.sap.com/pc/tech/enterprise-information-management/software/data-quality/index.html>.

³ Talend Enterprise Data Quality solution: <http://www.talend.com/resource/data-quality.html>.

5. Repeatedly step 1 to step 4, we eventually accumulate our knowledge on phenomena and data. If we cannot reach an adequate interpretation, then we go back to previous steps.

2.8 Data Analysis

How do we analyse the data which we have in our problem or project? The applied data-centric social sciences are cyber-enabled and require the use of inductive strategies to define problems and solution. One of our final goals in the inductive approach is to find a model to explain the data-generating mechanism. If we have a good model to explain it, we have an ability to predict or to infer the phenomenon which we treat from a subset of the data.

In the data analysis, we can have several types of tools: segmentation, change-point-detection, parameter estimation, classification, correlation, quantification and so on. These methods and tools are addressed in Chap. 3 in detail.

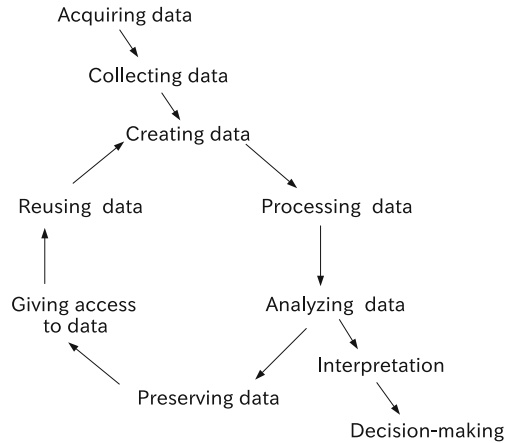
2.9 Data Life-Cycle

Data is often used for a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding is over. Following projects may need data to analyse them or add to the data. The secondary analysis of the data will need the data to analyse them for purposes other than those the primary project intended. As shown in Fig. 2.1, the data creation, recycle and reuse is an ongoing process. The data life-cycle is constructed from the following elements:

- creating data
- processing data
- analysing data
- preserving data
- giving access to data
- reusing data

The data is created from both new data collection which is acquired from the research activity and old data collection which was created at the past activity. Processing data and analysing data are done in the current research project, as we have seen above. The data created in the activity should be preserved as an archive. Giving access to data means that the authors of the data transfer ownership to others and waive the authors' right, which places the work into the public domain. In the European Union, there is the database right. In some countries, there may be no protection for collections of data. When we reuse the data in our own publications, we should indicate the licence under which we are reusing the data in order to make readers to recognise the data reused.

Fig. 2.1 A schematic illustration of data life-cycle



2.10 Social Implementation

2.10.1 Examples

Recently, data-oriented services have launched in several branches of commercial sectors. Facebook [12] and Twitter [35] are currently first examples of social media. ResearchGate [26] and Google Scholar [13] are academic examples. In the case of tourism management systems, Expedia [11], Ebookers [10] and Tripadvisor [32] are good examples. SurveyMonkey [29] enables us to design, collect and analyse our own surveys.

After the Earth Summit, which was held in Rio de Janeiro in 1992, the finiteness of our environment and the importance of monitoring our society was recognised. Several international institutions have issued sustainability indicators in order to guide and facilitate decision-making.

Consequently, social statistics databases from public sectors are available. United Nations Statistical Databases (UNSD)⁴ provide sustainable development indicators as well as macro statistics. United Nations Commission on Sustainable Development (CSD) indicators (CSDIs) for Sustainable Development are measured in 14 themes:

- poverty
- natural hazards
- economic development
- governance
- atmosphere
- global economic partnership
- health
- land

⁴ <http://unstats.un.org/unsd/databases.htm>.

- consumption and production patterns
- education
- oceans, seas and coasts
- biodiversity
- demographics
- freshwater

The European Union also selects eleven headline indicators as Sustainable Development Indicators (SDIs)⁵:

- socioeconomic development
- sustainable consumption and production
- social inclusion
- demographic changes
- public health
- climate change and energy
- sustainable transport
- natural resources
- global partnership
- good governance

The World DataBank of the World Bank is a free and open comprehensive data service on socioeconomic-technological systems, which provides several perspectives as macroeconomic indicators related to human activities [31]. Much of the data from the statistical systems of 188 member countries of the International Bank for Reconstruction and Development (IBRD).

Helbing and Baliatti proposed the 85 online repositories for the socio-economic sciences [17]. They classified these databases into 18 categories such as:

1. Internet and historical snapshots
2. information retrieval engines
3. text mining on the Web
4. social data sharing
5. conflict data
6. data in economics and finance
7. scientific collaboration data
8. social sciences
9. urban data
10. traffic data
11. open maps
12. logistic data
13. health data
14. climate and environmental data

⁵ <http://epp.eurostat.ec.europa.eu/portal/page/portal/sdi/indicators>.

15. energy
16. reality mining
17. other open data initiatives

These databases define situations of our world from several dimensions such as economy, environment, technology and societies. Some of them have been updated and expanded currently. The current data availability of databases obviously enhance our research and business environment. A future data availability will expand our research to capture our world from data-centric point of view more than the current.

2.10.2 Privacy and Public Utility

2.10.2.1 Data Protection Act

The Data Protection Act 1998 is a United Kingdom Act of Parliament which defines UK law for processing data on identifiable living people [8]. It provides us with the ability to control the area and purpose where our personal information is used by organisations, businesses or the government with the contract at the time when we provide our personal information. Everyone who is responsible for using data has to follow strict rules called *data protection principles*. They must make sure the information is:

- used fairly and lawfully
- used for limited, specifically stated purposes
- used in a way that is adequate, relevant and not excessive
- accurate
- kept for no longer than is absolutely necessary
- handled according to people's data protection rights
- kept safe and secure
- not transferred outside of the company without adequate protection

Fundamentally, usage of personal information other than primary purposes is not permitted. Specifically, ethnic background, political opinions, religious beliefs, health, sexual health and criminal records are sensitive information to be treated with stronger legal protection.

However, the implementation of data protection principles strongly depends on countries. For instance, U.S.-based service providers mostly implement their own privacy policy as self-regulations. The mindset behind this could be summarised as “agree, or stay out”. Users who want to use services must accept terms and conditions before they provide their privacy data. European data protection laws are more based on laws and regulations. European privacy protection research has identified three major protection goals of privacy, equivalent to the well-known protection goals of common security such as confidentiality, availability and integrity. In addition to

them, transparency, intervenability and unlinkability are to be considered. Confidentiality is an opposite concept to availability. Transparency is opposite to unlinkability. Integrity is opposite to intervenability. These axes show trade-off relationships in privacy protection.

Demand of secondary usage of data is observed. There is a trade-off relationship between privacy and transparency. Thus, the task of implementing transparency services is a crucial part of all electronic commerce under the regulation.

Anonymisation is one of the key technologies. The data related to privacy is often used for secondary purposes after anonymisation. Several types of anonymising techniques are available. For example, deletion of personal identification numbers and random shuffling are often used. In this case, some methods can keep statistical properties of the data. This technique is called *data anonymisation*. In a data anonymisation process, a real-world application of a privacy-preserving technology, which is called the synthetic data generation, is needed [20]. A plenty of data processing techniques regarding privacy-preserving are recently proposed [1, 22, 33, 36]. The privacy-preserving computations consist of several computations executed in partitioned databases while keeping privacy [1]. The privacy-preserving record linkage techniques [36] allow the linking of databases between organisations while at the same time preserving the privacy of these data. Tsubaki considers a way to evaluate the value of informative data for prediction under partial disclosure [33].

In the study of synthetic data generation [20], there are three types of privacy definition. l -diversity, (d, γ) -privacy and differential privacy. l -diversity can protect against adversaries with background knowledge, but it does not always guarantee privacy when there is a semantic relationship between distinct sensitive values. (d, γ) -privacy is a probabilistic privacy definition in which an adversary believes in some prior probability appearing in the data. Differential privacy is a privacy definition that the anonymisation algorithm should not give additional information about the remaining individual to the adversary who knows complete information about all individuals in the data except one. Jensen also proposes a decentralised solution for supporting an anonymised collection of transparency-relevant information based on the service-oriented principles [19].

In the data validation service, privacy issues are important. Soni et al. propose three types of data validation concepts [28]: producer-centric, customer-centric and reporting-centric. In the provider-centric approach, the actual processing of the data is performed on provider's side, which implies that the relevant data is transferred from consumer to provider. The customer-centric approach shows high privacy but low latency and low efficiency. In the consumer-centric approach, the processing is performed on the consumer side. This results in low privacy but high latency and high efficiency. In the reporting-centric approach, the processing of rules is performed as it is in the consumer-centric approach; however, the flagged data is transferred to the provider for reporting purposes. This shows medium privacy and medium latency but low efficiency.

2.10.3 Problems in Social Implementation

We need to carefully consider social implementation to data-centric social sciences. We eventually recognise several types of problems relating to social implementation of the data-centric approach:

- data lifetime
- data accuracy
- manipulation

The available time of individuals is finite. We often face the problem of data lifetime. Imagine that some data can be shared in some individuals in order to decide their behaviour. The data are created step by step and change gradually. The previous data may mislead the behaviour of the decision-maker. How do we distinguish old data from the latest data? I think that we should observe physical environment and interpret the data with a linkage with the actual environment.

The second problem is data accuracy. The data accuracy should be confirmed based on data from other sources or improved by using several validation procedures. If we found some differences between two databases, then we understand that we need to validate the data from these databases. These data errors may mislead both individual and social behaviour. The manipulation by data is sometimes observed. Some of them are used for the purpose of controlling social behaviour in public sectors or commercial sectors. The Libor (London Interbank Offered Rate) scandal was a series of fraudulent actions. This was that several world's banks obtained profits by manipulating the Libor interest rate illegally [4, 27].

2.10.4 Application of Data Analysis Techniques

Data analysis techniques can be used to detect fraud behaviour. These techniques were firstly employed by banks, telephone companies and insurance companies. The techniques for fraud detection are classified into two main categories including artificial intelligence and statistical techniques [6]. Some of the examples of statistical data analysis techniques are as follows:

- data preprocessing for detecting, validating, correcting error and filling up of incorrect and missing data
- computation of user profile
- matching algorithms for detecting incongruities in the behaviour of users or transactions, which are compared with earlier known profiles or models

To apply these techniques to actual situations, we need to access private data. However, public utility in commercial transactions is sometimes prioritised in comparison with privacy. Other examples are found in drug development. A database concerning medication delivery to the patients has recently been analysed for the purpose of drug

design. Patients' medical data are recorded as electronic health records (EHRs). There are studies on a privacy-protecting information system for controlled disclosure of EHR related to personal data to third parties [16]. The automated healthcare-data-mining system is studied as applications of web technology to healthcare for remote patients [30]. The data-mining service extracts information from data based on a correlation between lifestyle and health data.

References

1. Abbasi, S., Cimato, S., Damiani, E.: Toward secure clustered multi-party computation: a privacy-preserving clustering protocol. In: Mustofa, K., Neuhold, E., Tjoa, A., Weippl, E., You, I. (eds.) *Information and Communication Technology. Lecture Notes in Computer Science*, vol. 7804, pp. 447–452. Springer, Berlin (2013)
2. AB-ROAD travel retrieval WebAPI: <http://webservice.recruit.co.jp/ab-road/>
3. Barnett, V., Lewis, T.: *Outliers in Statistical Data*. Wiley, New York (1998)
4. BBC Timeline: Libor-fixing scandal (6 Feb 2013) <http://www.bbc.com/news/business-18671255>. Accessed 29 Mar 2014
5. Ben-Gal, I.: Outlier detection. In: Maimon, O., Rockach, L. (eds.) *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, pp. 131–147. Springer, New York (2005)
6. Chan, P.K., Wei, F., Prodomidis, A.L., Stolfo, S.J.: Distributed data mining in credit card fraud detection. *Intel. Syst. Appl. IEEE* **14**(6), 67–74 (1999)
7. Chauvenet, W.: *A Manual of Spherical and Practical Astronomy V.II*, 1st edn. Lippincott, Philadelphia (1863) (Reprint of 1891 5th edn: Dover, NY (1960))
8. Data Protection Act 1998: <http://www.legislation.gov.uk/ukpga/1998/29/contents>
9. Dean, R.B., Dixon, W.J.: Simplified statistics for small numbers of observations. *Anal. Chem.* **23**(4), 636–638 (1951)
10. Ebookers: <http://www.ebookers.com/>
11. Expedia: <http://www.expedia.co.jp/>
12. Facebook: <https://www.facebook.com/>
13. Google Scholar: <http://scholar.google.co.jp/>
14. Google Translate WebAPI: <https://developers.google.com/translate/>
15. Grubbs, F.E.: Procedures for detecting outlying observations in samples. *Technometrics* **11**(1), 1–21 (1969)
16. Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N., Müller, G.: Aspects of privacy for electronic health records. *Int. J. Med. Inform.* **80**(2), e26–e31 (2011)
17. Helbling, D., Ballezzi, S.: From social data mining to forecasting socio-economic crises. *Eur. Phys. J. Spec. Top.* **195**(1), 3–68 (2011)
18. Jalan vacant room information retrieval WebAPI: <http://www.jalan.net/jw/jwp0000/jww0001.do>
19. Jensen, M.: Towards privacy-friendly transparency services in inter-organizational business processes. In: 2013 IEEE 37th Annual Computer Software and Applications Conference Workshop, pp. 200–205 (2013)
20. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 277–286 (2008)
21. Mahalanobis, P.C.: On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* **2**(1), 49–55 (1936)
22. Mehta, S.R., Vinterbo, S.A., Little, S.J.: Ensuring privacy in the study of pathogen genetics. *Lancet Infect. Dis.* **14**, 70016 (2014)

23. Peirce, B.: Criterion for the rejection of doubtful observations. *Astron. J.* **2**(43), 161–163 (1852)
24. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. *IEEE Bull. Tech. Comm. Data Eng.* **23**(4), 3–13 (2000)
25. Rakuten Web Service WebAPI: <http://webservice.rakuten.co.jp/api/ichibaitemsearch/>
26. ResearchGate: <http://www.researchgate.net/>
27. Reuters: Libor scandal may cost banks \$14 billion in settlements: analysts (12 July 2012) <http://uk.reuters.com/article/2012/07/12/uk-libor-scandal-estimates-idUKBRE86B1EE20120712>. Accessed 29 Mar 2014
28. Soni, S., Mehta, S., Hans, S.: Towards providing data validation as a service. In: 2012 IEEE 9th International Conference on Services Computing, pp. 570–577 (2012)
29. SurveyMonkey: <https://www.surveymonkey.com/>
30. Takeuchi, H., Kodama, N., Hashiguchi, T., Hayashi, D.: Automated healthcare data mining based on a personal dynamic healthcare system. In: Engineering in Medicine and Biology Society, EMBS '06. 28th Annual International Conference of the IEEE, 30 Aug–3 Sept 2006, pp. 3604–3607 (2006)
31. The World DataBank of the World Bank: <http://data.worldbank.org>
32. Tripadvisor: <http://www.tripadvisor.com/>
33. Tsubaki, H.: Valuation of partly disclosed datasets for prediction. In: icdmw, 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 733–734 (2013)
34. Tukey, J. W.: Exploratory Data Analysis. Addison-Wesley, Reading (1977)
35. Twitter: <https://twitter.com/>
36. Vatsalan, D., Christen, P., Verykios, V.S.: A taxonomy of privacy-preserving record linkage techniques. *Inform. Syst.* **38**(6), 946–969 (2013)
37. Yahoo! JAPAN text analytics WebAPI: <http://developer.yahoo.co.jp/webapi/jp/>

Applied Data-Centric Social Sciences
Concepts, Data, Computation, and Theory

Sato, A.-H.

2014, XXIII, 281 p. 71 illus., 20 illus. in color., Hardcover

ISBN: 978-4-431-54973-4