INVITED PAPER *Special Section on Log Data Usage Technology and Office Information Systems*

Data-Driven Decision-Making in Cyber-Physical Integrated Society

Noboru SONEHARA[†a)], ***Fellow***, **Takahisa SUZUKI**[†], ***Nonmember***, **Akihisa KODATE**[†], ***Senior Member***, **Toshihiko WAKAHARA**[†], **Yoshinori SAKAI**[†], ***Fellows***, **Yu ICHIFUJI**[††], ***Member***, **Hideo FUJII**[†††], ***and*** **Hideki YOSHII**[††††], ***Nonmembers***

SUMMARY The Cyber-Physical Integrated Society (CPIS) is being formed with the fusion of cyber-space and the real-world. In this paper, we will discuss Data-Driven Decision-Making (DDDM) support systems to solve social problems in the CPIS. First, we introduce a Web of Resources (WoR) that uses Web booking log data for destination data management. Next, we introduce an Internet of Persons (IoP) system to visualize individual and group flows of people by analyzing collected Wi-Fi usage log data. Specifically, we present examples of how WoR and IoP visualize flows of groups of people that can be shared across different industries, including telecommunications carriers and railway operators, and policy decision support for local, short-term events. Finally, the importance of data-driven training of human resources to support DDDM in the future CPIS is discussed.

key words: *Cyber-Physical Integrated Society, Data-Driven Decision Making, Internet of Persons, web of resources, lifelog, destination data management*

1. Introduction

Various kinds of information devices and sensors are being connected to networks by advanced information and communication technology (ICT) and information is being digitized for distribution, enabling much greater access to information by anyone from anywhere at any time. The result is that a society in which cyberspace and the real world are connected or fused is being formed. We refer to that society as the Cyber-Physical Integrated Society (CPIS) [1].

In such a CPIS, the actions and states of people and things in the real world are projected onto cyberspace via lifelog from smart phones and other such devices and from the Internet of Things (IoT). In cyberspace, changes in people and things in the physical world are reflected as changes in information. Such changes would be processed by data analysis and service synthesis based on scientific evidence with the results rapidly provided as feedback and with appropriate timing to individuals and society. In that way, the CPIS system can be controlled.

This kind of repeated information circulation between the real world and cyberspace can solve a variety of social problems and improve the lives of people. It will create new values, enable a transformation of industrial structure to an intellectual information industry and knowledge service industry, and create new opportunities for employment.

To achieve such goals, highly-public collection and analysis of a vast amount of social data, the scientific data-driven synthesis of high-quality knowledge services, and social and technical mechanisms for appropriate feedback to the CPIS are required.

As written above, this paper describes a Data-Driven Decision-Making (DDDM) support system for problem solving in the CPIS. The rest of the paper is organized as follows. In Sect. 2, the importance of DDDM in the CPIS is described. Section 3 focuses on destination data management systems as an example where DDDM support system can be applied by introducing a Web of Resources (WoR) and an Internet of Persons (IoP) system to visualize individual and group flows of people by analyzing collected Wi-Fi system usage log data. Specifically, we present and discuss examples of visualizing flow lines of inbound tourists, data sharing across different industries, including telecommunications carriers and railway operators, and policy decision support for local, short-term events that attract huge numbers of people. Section 4 concludes this paper by describing the importance of data-driven training of human resources in support of the future CPIS.

2. Problem Solving in the Cyber-Physical Integrated Society [2]

CPIS holds promise for solving a variety of social problems regarding the natural environment, energy, food supply, the prevention and amelioration of disasters, medical and nursing care, etc. To fulfill that promise, it is essential to strengthen the ability to extract knowledge and insight from huge and complicated digital data sets. Thus, our objective is to achieve a sustainable society by transforming data quantity into quality and creating an intellectual information industry and a knowledge service industry. This section is organized as follows. In 2.1, our concept and mechanism of DDDM support system is described. Then in 2.2, the potential for and challenges of CPIS in using and linking public surveys and lifelog data are discussed. Following that, 2.3 proposes a method compatible with both the use and the protection of lifelog data.

Manuscript received November 1, 2018.
Manuscript publicized July 4, 2019.
[†]The authors are with the Tsuda University, Tokyo, 151–0051 Japan.
[††]The author is with the Nagasaki University, Nagasaki-shi, 852–8521 Japan.
[†††]The author is with the Kintetsu Railway Co. Ltd., Nara-shi, 630–8025 Japan.
[††††]The author is with the Softbank Corp., Tokyo, 105–0021 Japan.
a) E-mail: sonehara@tsuda.ac.jp
DOI: 10.1587/transinf.2018OFI0002

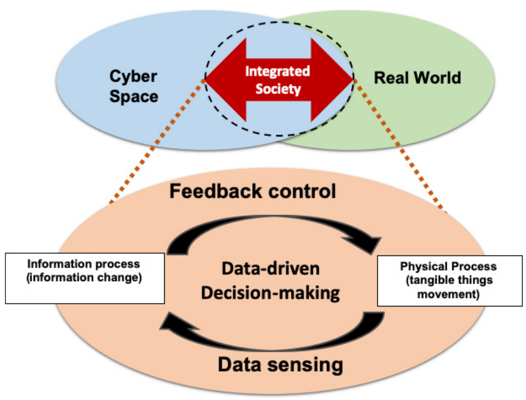


Fig. 1 Data-driven decision-making mechanism in the Cyber-Physical Integrated Society

2.1 Data-Driven Decision-Making [2]

In the past, decision-making has relied on subjective decisions and judgments made on the basis of partial or incomplete data. Although a wide variety of digital big data is currently available via networks, the ability of people and society to analyze such data is limited. Hence the quality of decision-making has not improved due to the decrease in analytical ability relative to the quantitative increase in data.

A solution to this problem is evidence-based decision-making. This requires the collection and analysis of big data acquired by information devices and various sensors to support decision-making and timely formulation of policy. DDDM should be used to support systems to make rational decisions and provide appropriate feedback to society. DDDM could raise the quality of decisions made by ordinary consumers and users as well as improving policymaking and managerial decisions.

The mechanism of DDDM in CPIS is illustrated in Fig. 1. The system uses smart phones and other such portable ICT devices and the Web and Social Networking Services (SNS) as sensors. The data collected by these sensors is stored in a lifelog recording the social interactions, business activities, communication, etc. of people. The lifelog represents the sensed behavior of people in the real world and can be used to project that onto cyberspace. The lifelog data is analyzed in information space and high-quality knowledge services are synthesized. The knowledge services provide timely feedback to people, things and society. It is thus possible to execute the PDCA (Plan Do Check Act) cycle in real time and rapidly implement measures, which is something not possible with the conventional approach using social surveys and public statistics.

A DDDM support system accelerates the production, distribution, and consumption of intelligent information services and knowledge services. The general and widespread use of the DDDM support system will lead to the creation of a knowledge service industry for the CPIS.

2.2 Complementarity of Public Surveys and Lifelog Data [2]–[4]

Statistical surveys target random samples from strictly defined populations in order to be able to do unbiased estimation of the population's responses even with sample sizes. In contrast, lifelog data in cyberspace can be collected from a large number of samples in real time, although potentially having a problem with not being representative of the population. Thus, it is necessary to check the representativeness of collected lifelog data and correct for any biases by using unbiased public statistical data. If public statistical data and lifelog data are used to complement each other, it will be possible to get unbiased real time data from cyberspace. Up to now, public statistical data, such as population, tourism and number of foreign visitors data have been available annually or monthly, however newer data sources will enable these to be updated daily or even more frequently which will promote a knowledge service industry. In order to achieve this, it is necessary to simultaneously evaluate the quality of real time lifelog data and unbiased public statistical data. Recently, there has been research that estimates population and economic situation of residents using lifelog data and assess the accuracy of that estimation using public statistical data.

Moreover, public surveys and lifelog have different strengths and weaknesses. Subjective psychological data can be collected with public survey and objective data on actual actions can be collected through lifelog. Merging both types of data enables more precise analyses reflecting human actions and changes.

It has been pointed out that the reliability of representative survey data answered by self-report is decreasing. Even if samples were randomly selected, there is still the possibility that answers are biased due to intentional or unintentional factors. For example, there is the problem of recall since human memory is fallible, so that respondents may mis-remember actions they report on surveys. Because lifelog with automatically collected data may be less affected by such factors, the lifelog data would be expected to be more reliable than survey data.

2.3 Social Data Platforms Considering Daily Availability of Information Services [1]

Up to now, public surveys and online surveys have been used to collect data, but heightened awareness with respect to protection of personal information has been making it difficult to obtain cooperation in data collection in recent years.

On the other hand, various types of personal data, including lifelog, are accumulating in networks because of the growing popularity of high-performance mobile terminals and the rise of SNS. The lifelog includes records of personal activities such as travel and accommodation, interests and preferences, shopping, meals, and medical history. For that reason, a mechanism that both protects and allows effective use of lifelog data is required.

Focusing on the availability of frequently collected information services, we propose a method for protecting and using lifelog data. It is desirable that information services which are used in everyday life are available in emergencies as well, such as when disasters occur. For example, individual identifying information that may be needed during a disaster includes attributes for nearby people such as names, ages, addresses of persons who live in the affected area, health condition and whether they need assistance, language ability (e.g., whether they can understand Japanese). Although laws concerning the handling of personal data also apply during disasters, communications may be lost and administrative functions may cease in areas affected by a major earthquake, so it is not always possible to rapidly acquire and use peoples' personal data.

To address that problem, a system based on interoperable information services ensures seamless transition from tourist guide or shopping applications that use personal information covering location, interests, preferences, etc. in normal times to emergency evacuation guide applications and services that provide information on emergency medical care, access to relief goods, etc. that are used in the same manner when a disaster occurs.

In ordinary business and development, services are designed so as to establish an exchange of value for the benefits obtained from the disclosure of lifelog data in normal times. However, it is difficult to determine a uniform standard for disclosure conditions for all users. The reason is that, although people do not want their private information known by others, individuals value their private information differently for personal reasons. Thus, it is impossible to determine uniform criteria for deciding on disclosure that maintains a balance of protection and utility of personal information. Furthermore, it is desirable to have a simple user interface that enables the selecting of disclosure conditions even for persons that have low risk of personal data disclosure.

An example mechanism for lifelog data disclosure and utilization management is illustrated in Fig. 2. Such a system should be implemented to balance the protection of lifelog data in normal times while also ensuring the utility of lifelog data at times of disaster. Specifically, the prior consent of the person is obtained for the use of the lifelog data that is recorded in normal times only for response to a disaster. By doing so, it is not necessary to collect new lifelog data when a disaster occurs. The psychological barrier is low, because the consent is for disclosure of information that is limited to a time of disaster. The potential benefit of rescue, confirmation of safety, and automatic relief during a disaster provides sufficient incentive for ordinary citizens to agree to such use.

Also, to deal with the problem of being unable to determine a uniform set of disclosure conditions applicable to all people, each user would sets their disclosure conditions according to their own value judgments. The disclosure mechanism is a conditional opt-in for each type of personal infor-

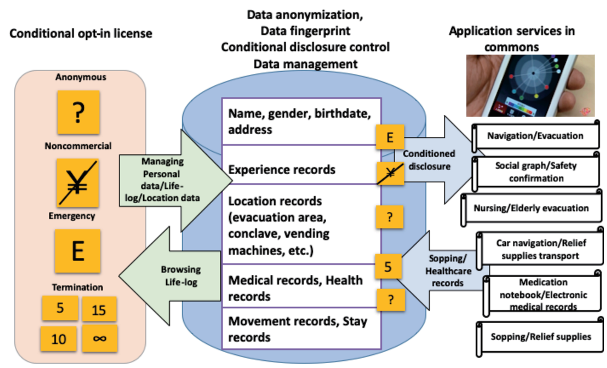


Fig. 2 Platform for protected use of lifelog by conditional opt-in

mation concerning the disclosure period after the disaster, limitation of the scope or purpose of disclosure, direct use for rescue or relief, etc., and indirect use for damage statistics, etc.

The specific procedure for conditional opt-in involves the user registering their own lifelog via a website. At that time, the users make their own decisions concerning the handling of their lifelog and set the use conditions for access to their information. The lifelog is appropriately anonymized and stored in a database according to the use conditions as set by the user. The stored information is provided to service providers according to the use conditions of the information provider. The user can add, change, or delete the information items that are collected. In order to make it possible for anyone to easily determine the lifelog disclosure conditions, the data is managed with combinations of four licenses: 1) the information can be used only when data anonymization techniques have been applied, 2) the information can be used only for non-commercial purposes, 3) the information can be used only in emergencies, 4) the information can be stored only for a retention period specified in years and no longer.

In that way, lifelog data can be used in a data-driven policy decision system for rapid development of inappropriate relief and rescue plan.

3. Destination Data Management Systems

3.1 Tourism Information and Its Economic Impact

In Japan, the population is rapidly decreasing and many localities are considered to be declining due to population shifts to large urban areas. According to a report published in 2014 by the Japan Policy Council, the population of young females is expected to decline by 50% or more in a considerable number of municipalities by 2040. This has created concern that the number of municipalities will decrease in many regions of Japan. One measure for invigorating local economies and decreasing the risk of such decline is tourism. In 2018, 31.19 million people visited Japan, more than ever before, with most coming from Korea and China. Tourist spending has already reached 4.7 trillion JPY (assuming an average spend of 140,000 JPY).

Achieving forty million annual visits and a market size of 1 trillion JPY by 2020 are a policy objective of the Japanese government. Over 10 million domestic and foreign visitors are expected for the 2020 Tokyo Olympics and Paralympic games. Altogether, the equivalent of 40% of the Japanese population is expected to enjoy viewing and sightseeing for the event.

3.2 Destination Lifelog Data Management

For conventional tourism, data on the types of tourists, the types of services they wish to use, and the transportation functions that they use has been collected and analyzed. Most of such research involves independent collection and analysis of data on products and services that are specific to individual fields, including transportation, accommodation, entertainment, sightseeing, souvenir sales, and eating and drinking. In marketing, for example, data on purchase history and amount of payment collected would be a basis for further product recommendations and linking them to purchases. However, data such as transportation usage, accommodation, and purchases are not included in lifelog data.

Current local tourism promotion involves encouraging customers to stay and travel within the region, increasing spending in the process. In this scenario, only having data collected on transportation and accommodation or only on tourist movement and spending is insufficient. It is also essential to collect combined social data, including advertisement and information search, means of travel such as transportation and parking lots, sightseeing and event attendance, food and drink, lodging and shopping, etc. If a highly comprehensive data infrastructure could be implemented, it would be possible to support tourism policymaking and decisions such encouraging visitors to explore and spend more and also to increase visitor satisfaction. That would enable creation of an appealing destination social space in which visitors can enjoy smooth and comfortable transportation and accommodation services, commercial facilities that can be enjoyed while moving around and taking tours, and experiences such as participation in festivals, tourism, and sports events.

For more efficient operation and management of local tourism resources, it is essential that various stakeholders cooperate to continuously share and analyze various types of data, including transportation, accommodations, dining, sightseeing, entertainment, and shopping, and that tourism policies are planned and implemented based on the collection and analysis of data.

The destination lifelog data management system for executing the PDCA cycle for local tourism policy-making is illustrated in Fig. 3. The system comprises the WoR (Web of Resources), which is a data infrastructure for local tourism resources that uses Web reservation data, and the IoP (Internet of Persons), which is person trip data acquired using Wi-Fi system log. This system can be used to design new participatory, experience-based tourism services that

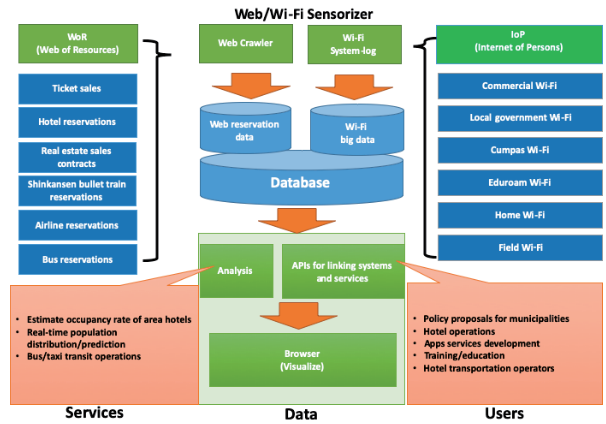


Fig. 3 Destination lifelog data management by Web and Wi-Fi device sensors

go beyond movement and purchasing. Furthermore, systematic collection and analysis of highly comprehensive individual lifelog of users and consumers enables cooperation of individual services and rational design of appealing local tourism services. Rational, data-driven support for sightseeing and touring activities for both service providers and users is also possible. The following sections describe specific R&D and social experiments using the WoR local tourism resources data infrastructure and the IoP personal trip data in the planning of local tourism policy for regional development.

3.3 Web of Tourism-Resources (WoR) Using Web Booking Log [5], [6]

Data-driven tourism policy planning requires data such as the number of tourists, occupancy rate of lodging facilities, and use rate of various means of transportation. Such data includes public statistics from local and national governments and the survey data of private enterprises. However, the International Visitor Survey of the Japan Tourism Agency, which provides public statistics on lodging and travel, is published only quarterly, so it is not suitable for use in rapid policy-making.

Most local governments do not have the data for rational decision-making concerning event timing and venue or the means to evaluate economic effects by holding events. For that reason, local governments and tourism operators need means for policymaking, such as knowing when and how many tourists visit and how to attract more tourists, and assessing the economic impact. Also, the cost of local governments individually conducting data surveys is huge, and if there are no standards for survey items and data, the individual results are very inefficient because they cannot be reused or compared.

We therefore propose WoR as a standard infrastructure. The system collects booking log data from reservation websites for lodging and transportation, facilities, and events. WoR, which visualizes the movement of people in social situations, monitors reservation service websites to estimate

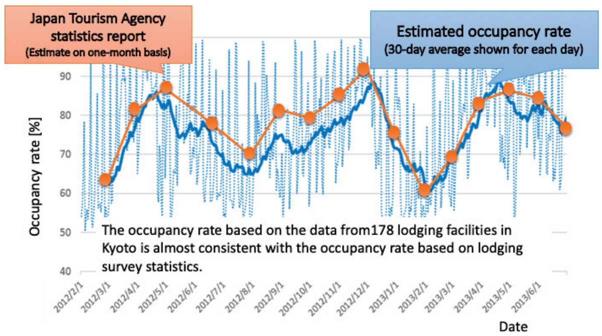


Fig. 4 Accommodation occupancy rate using WoR

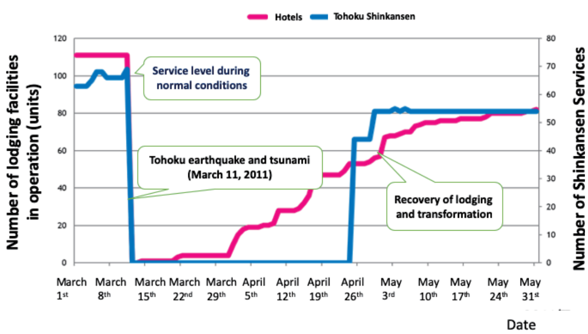


Fig. 5 Visualization of short-term recovery situation after an earthquake disaster using WoR

facility use rates in advance. The future utilization rates of facilities can be estimated from "reservation possible (vacancies, seats available, etc.)" and "reservation not possible (no vacancies, no seats available, etc.)" information for facilities.

The daily occupancy rates estimated from Web reservation data for 178 lodging facilities in Kyoto that have ten or more employees (monthly surveys conducted manually) are compared with public statistics in Fig. 4. Good accuracy was obtained in the estimates, with an error rate of 7% or less. It is therefore possible to use such data to support the planning of events to be held during less active periods.

WoR can be used to visualize the restoration of lodging and transportation systems after a disaster [7]–[9]. The operating status of the Tohoku Shinkansen bullet train and lodging facilities in the city of Sendai after the major earthquake disaster that occurred on March 11, 2011 is shown in Fig. 5. When the earthquake occurred, the bullet train services and lodging services became unavailable. After the earthquake, it became possible to visualize the short-term recovery situation of hotel and the bullet train services as infrastructure services, including electricity, water and sewage, and communication, were gradually restored. By accumulating Web booking log data over a long period of time in this way, it is possible to visualize the recovery situation for lodging facilities and the bullet train in times of disaster. It is also possible for the national government to use data for planning disaster recovery activities and providing transportation and accommodation for volunteers and recovery workers.

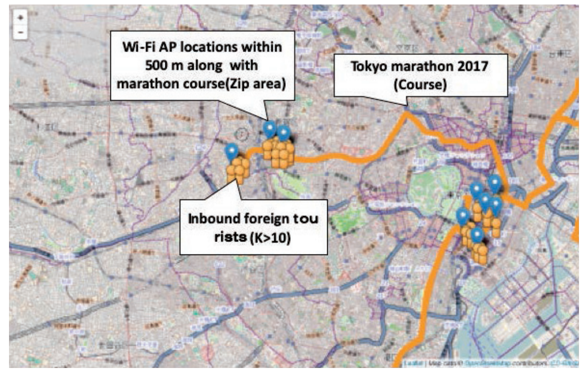


Fig. 6 Visualization of group flow for international tourists at the 2017 Tokyo Marathon: Departure area (from 10:00 to 16:00)

3.4 Internet of Persons (IoP) Using Wi-Fi System Log Data [10]–[12]

We introduce a system that visualizes the system log data of Wi-Fi access points to understand collective flow of people. Data sensing in a situation where all things are connected to the Internet and data is stored in cyberspace is referred to as the Internet of Things (IoT). In contrast to that, location sensing through mobile devices or network infrastructure stored in cyberspace can be referred to as the Internet of Persons (IoP).

Wi-Fi access point system log data is collected by telecommunications companies in the provision of free Wi-Fi access to international visitors to Japan. Consent to use that log data is obtained when the user applies for the service. Only anonymized data for which consent has been obtained is used for research purposes. The log data includes nationality, sex, and age.

A visualization of the movement of tourist groups (group flow) at the 2017 Tokyo Marathon based on data acquired using Wi-Fi device sensors is presented in Fig. 6. The figure is a map that visualizes the movement of foreign tourists on the marathon course (within 500 m) on February 26, 2017, in the vicinity of the starting point around the Shinjuku Metropolitan Government building. In consideration of privacy, each device number detected at a Wi-Fi access point is replaced by a random number and at least ten device numbers are processed collectively. The movement of device numbers is detected by postal area codes in units of one hour.

A visualization of group flow of tourists in the Kanto region in February 2017 is shown in Fig. 7. The map in the figure captures the dynamics of international tourist movement for Narita International Airport, Tokyo International Airport (Haneda), Yokohama, Tokyo, Funabashi, etc.

A map that presents a combined visualization of tourist movement and the hotel reservation situation and tourist spots in the central area of Nagasaki city for the year from July 2015 to June 2016 is shown in Fig. 8. Such data can support the discovery of new tourist spots, guest pick-up

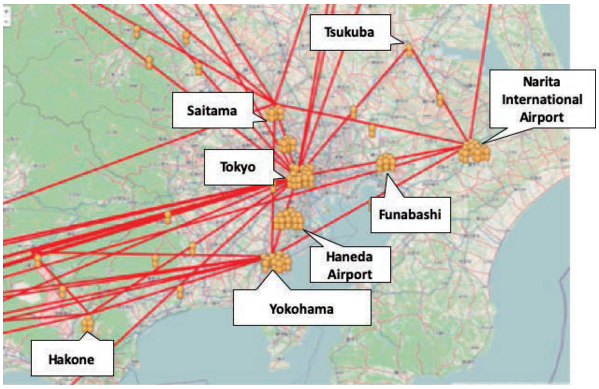


Fig. 7 Visualization of group flow for international tourists at the 2017 Tokyo Marathon: Kanto region (one month)

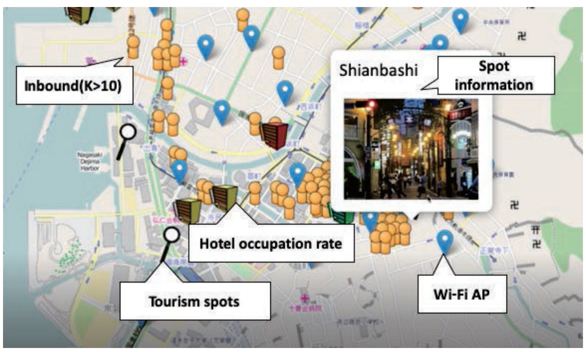


Fig. 8 Inbound collective tourists flow in Nagasaki-city (2017)

and hospitality services, assignment of service personnel by country of guest, efficient placement of signboards, and digital signage advertising that is presented at the appropriate timing.

Combining information on the movement of people obtained from Wi-Fi system log with the results of analyzing Web data, such as hotel reservations, in real time makes it possible provide tourist information in the appropriate language, introduce products that match the sex and age of visitors, and optimally place personnel for serving meals and product sales based on the movements of persons and groups, and their nationality, sex and age data. Another advantage is that supply and demand can be matched for consumption activities in real time, which is not been possible in the past.

3.5 Value–Adding through Data Sharing with Privacy Preservation [13], [14]

To increase the added value of data, it is important to share data among different businesses and to analyze integrated data. However, because of criticism and anxiety related to lack of protection of personal information, protection of privacy, and consumer awareness, there is been almost no attempt by businesses to increase the added value of the data they possess by matching and supplementing†. We therefore describe here a method for adding attribute information to the data of railway operators by analyzing and matching the data possessed by railway operators and telecommunications operators in an integrated manner, a method for increasing the accuracy of analysis by matching data possessed by different business operators, and a service for increasing the value and use of data.

In an experiment, Wi-Fi equipment was installed at ticket gates in the Kintetsu Nara station and operated from November 2017 to March 2018. The gate counting data acquired by the railway operator from automatic ticket machines was matched with the Wi-Fi system log data of the telecommunications operator to obtain mutually complementary data.

The Wi-Fi location information is data based on the Wi-Fi probe and association system log obtained from terminals of users of the “Public Wireless LAN Service Wi-Fi Spot” and “Free Wi-Fi PASSPORT” services offered by the telecom provider Softbank who consented to the acquisition of personal information, excluding users who later revoked permission to use.

We conducted experiments on the effectiveness of Wi-Fi access point detection of smart phones and other terminals. The data from the ticket counting machines was considered to represent the actual number of ticket users passing through that point, and that number was compared with the number detected by Wi-Fi. The relation between the ticket count and Wi-Fi detection and their correlation is shown in Figs. 9 and 10. We can see that there is a positive correlation ($R2 = 0.7179$). An empirical formula for estimating the actual number of persons from the detection count is given by the formula (1) below.

$$\text{Actual number} = \frac{\text{missing rate} \times \text{Wi-Fi detection count}}{\text{carrier's share} \times \text{Wi-Fi ON rate}} \tag{1}$$

Use of the estimation formula (1) to estimate the number of ticket users from the Wi-Fi detection count is shown

†For example, when it became known that the East Japan Railway Company was providing “Suuica use data from which names, telephone numbers, and shopping information were excluded and which was anonymized by converting dates of birth that included year, month and date to dates of birth that included only year and month and converting the Suica ID number to a different irreversible number” to Hitachi, Ltd., many users voiced criticism and anxiety that “there was lack of consideration for protection of personal information, protection of privacy, and consumer awareness”. To address that concern, an expert council was set-up to examine provision of Suica data to parties outside the company and the discussion results were summarized and published in October 2015. In the summary, it was concluded that “statistically processed analysis results shall be provided only for purposes of public good that are considered to obtain the understanding of users” and that “by accumulating the results of using big data to meet to the expectations of society with full attention given to consideration of users, it will be possible to improve convenience to users and also contribute to the social and economic development of the nation.”

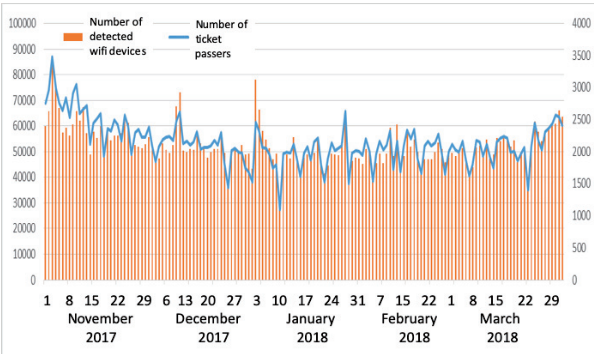


Fig. 9 Comparison between number of detected Wi-Fi devices and ticket passers (Nara-station, November 2017–March 2018)

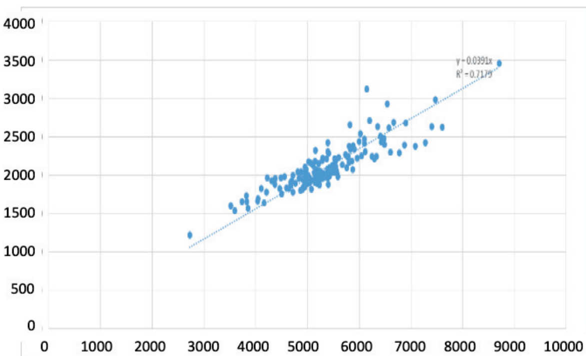


Fig. 10 Correlation between number of detected Wi-Fi devices and ticket passers (Nara-station, November 2017–March 2018)

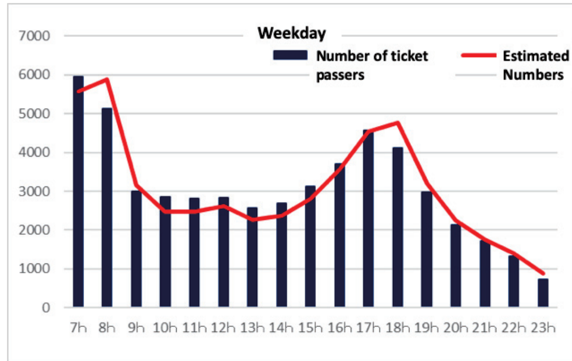


Fig. 11 Estimated real device number on weekdays by the experimental model (Nara-station, November 2017–March 2018)

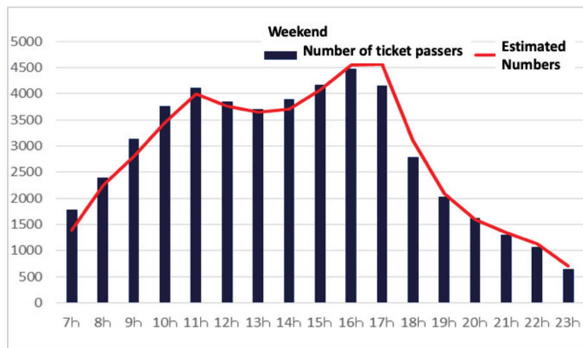


Fig. 12 Estimated real device number on weekend by the experimental model (Nara-station, November 2017–March 2018)

in Figs. 11 and 12. The parameters used in the experiment are the telecom carrier's share rate of 22.3% and the Wi-Fi on rate of 70%. With those values, we obtained an average of 1.94 (maximum value of 2.34 and minimum value of 1.54) with the daily detection missing rate. These results show that the number of ticket users can be estimated from the number of Wi-Fi detections.

The matching of data between different business operators in this way makes it possible to increase performance for estimation precision and to visualize attribute data. This method can be applied to data analysis for leisure facilities such as theme parks and commercial facilities such as department stores, and also holds promise for use in disaster prevention and amelioration, town planning, invigoration of tourist areas and shopping districts, and improvement of services for users.

3.6 Analysis of IoP at Shibuya Halloween 2018 and Policy Planning

Here, we describe an example of support for data-driven event policy planning. On October 31, 2018, an autonomous event was held by people celebrating Halloween in Shibuya. Many people gathered and formed crowds on the day before Halloween as well as on Halloween day itself. Some of the participants created disturbances on the street and other trouble with some even being arrested. For that reason, the Shibuya government suggested charging for the event or designating Yoyogi Park as the venue. Noting that many of the participants were not from Shibuya, there was also a concern that the measures to deal with the celebration disturbances would be paid for by Shibuya taxpayers.

There is no data collected on people coming to Shibuya to attend such local, short-term, large-scale events. Of use would be data on when they arrive, where they come from, what kind of persons they are and how many persons come, to provide scientific evidence for policymaking. We therefore used one of the IoP based commercial services, called Venue Vitalics by Cinarra Systems Japan [15] to estimate the device count during local event for policy planning in the near future.

In the experiment, we analyzed count and attribute data for people acquired from 632 access points installed within a radius of 300 m around Shibuya 109 (Dogenzaka, 2–chome, 29–1 in Shibuya-ku, Tokyo) on the day of the Halloween event to obtain the data for policymaking.

The number of Wi-Fi detections for the day of the Halloween event, October 31, 2018, and the following day are shown in Fig. 13. People began to gather around 17:00 on the day of the event and the number of people peaked at 20:00 when the arrival of 9,000 persons was detected. The scale of the event was estimated to be at least 90,000

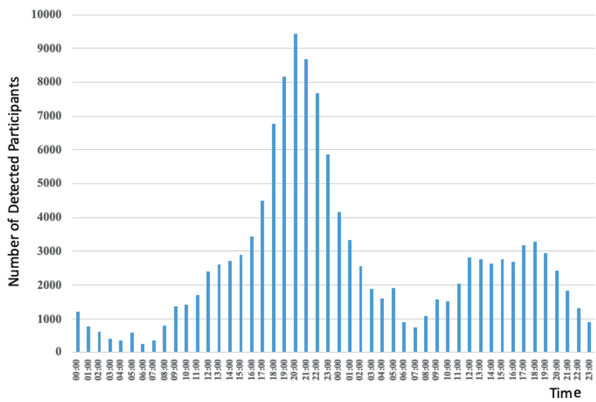


Fig. 13 Number of participants in Shibuya Halloween 2018 (Oct. 31.–Nov. 1.)

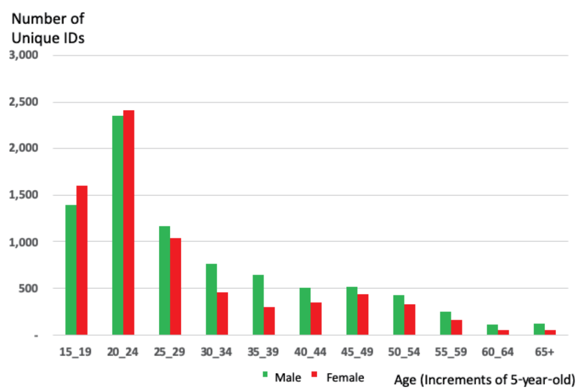


Fig. 14 Attributes (gender, generation) of participants in Shibuya Halloween 2018 (October 31. 18:00–24:00)

persons based on the number of Wi-Fi detections (missing area factor of 1.54). 5,000 participants remained in Shibuya, even after train service had stopped, and we can see a 2000-person peak at the time when the railway services started. Participation in the Shibuya Halloween event by residency of participant is shown in Fig. 14, where we can see that for October 31 from 18:00 to 24:00, the largest number of visitors were from Setagaya-ku, Shinjuku-ku, and Shibuya-ku, in that order. The distributions by sex and age are shown in Fig. 15, where we see that the age group from 20 to 24 accounted for the most participants and that males and females in the age groups from 15 to 19 and from 25 to 29 were also in the top positions.

Thus, we can see that IoP could be effective for providing data to support the setting of charges, designing the cost of security and cleanup, etc., the optimum placement of personnel, and estimating the economic impact when Shibuya-ku holds the Halloween event. Furthermore, matching the data with sales data, etc. could enable rational policy decisions.

3.7 Applications in Other Countries and Across Borders

The example applications shown in this paper are conducted within Japan and as such follow the Japanese Personal Information Protection Act. Personal data and privacy protection regulations or laws of each country are enacted based on the OECD guidelines. In practice, regulations on protection of personal information and its operation vary from country to country. Therefore, when conducting international surveys and research, it is necessary to collect, manage and operate personal data complying with the regulations of each country. For example, the General Data Protection Regulation (GDPR) [16], which is the legal regulation of the EU, protects a considerably wide range of information as "personal data" rather than the concept of "personal information" under the Japanese Personal Information Protection Act. Therefore, for Japanese researchers who do comparative surveys in Europe and Japan, they also need to follow GDPR. Furthermore, this legislation stipulates the guarantee of rights to automatic processing and strict countermeasures against infringement such as guaranteeing the right to refuse automatic collection and processing of personal data including profiling of data subjects (Article 22 [16]). Besides many issues are pointed out in connection with utilization of social big data [17], [18]. Data related to personal information and privacy, various information such as behavior history, purchase history, communication log, access records and "likes" can be recorded easily and used, must be handled very carefully. In addition, confirmation is necessary such as whether the consent of the person is taken or whether the purpose of use is clearly indicated. In addition, software robots which automatically collect data, such as crawlers, and systems interacting with Artificial Intelligence which analyze collected data can also be problematic if they are used without notifying the users properly [19].

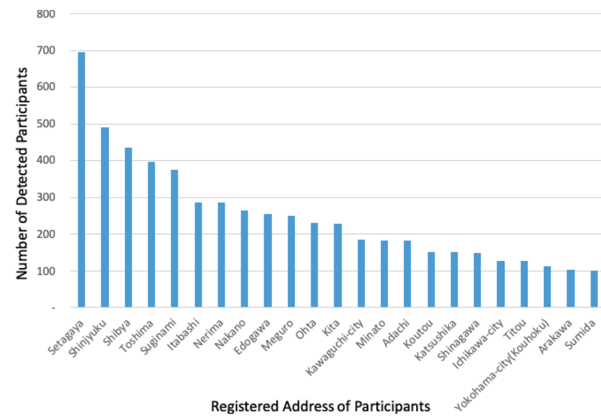


Fig. 15 Attributes (registered address) of participants in Shibuya Halloween 2018 (October 31. 18:00–24:00)

4. Conclusion

We have introduced and described Web and Wi-Fi device sensors that uses Web reservation data and Wi-Fi system log data as lifelog sensors for people. We also described its use in analyzing facility utilization, tourist visits to Japan, and the flow lines and attributes of event participation. We described how the data-driven policy decision system consisting of WoR and IoP can support rational decision-making and managerial decisions.

It is estimated that the 2020 Tokyo Olympic and Paralympic games will attract over 10 million visitors from Japan and abroad, and the policy target is for 40 million foreign visitors in 2020 [20].

A data-driven information and social systems for planning and executing policies for tourism, travel, and disaster prevention and amelioration is an important issue for the future. To respond to a rapid increase in the number of visitors to Japan in a short time, the ability to use of public data such as population and number of foreign visitors in real time and real-time visualization of group flow with full consideration given to privacy are important. To achieve that, construction of a social data infrastructure that enables Wi-Fi data cooperation by local governments, private enterprise, and universities and integrated operation and management of such data is an urgent task.

Further future work is the training of personnel that can use data in the collection, processing, analysis, and management of highly-public social big data for rational, data-driven policymaking and decision-making.

For that purpose, data science education is very important. We believe that data science is a new science that synthesizes existing fields of study into an objective-oriented science. That requires an educational system for human resource development and data science teaching materials.

The elements of data science are considered to be the subjects listed below.

1. Statistics, especially statistical analysis and hypothesis testing
2. Network science for collecting data, particularly Internet science, including Web/SNS
3. Data and information science, particularly the legal aspects of using data from the viewpoint of protecting personal information, and information theory, including data anonymization techniques.
4. The use of data to solve social problems, particularly methodologies for determining what kinds of data are effective for social problems.

An essential feature of the data science lecture is that it is accompanied by exercises. Concerning the above item 4, in particular, techniques for obtaining the data required for solving problems based on field examples are required. Conversely, the ability to consider what kinds of problems can be solved with data that has already been obtained will be important. Also important for data science education are teaching materials that dive deeply into each field for persons who have practical experience and introductory materials for teaching data science to beginners at universities, who are learning under the traditional literary system.

References

[1] N. Sonehara, "Identity Data Commons for Human-centric Cyber-Physical Integrated Society," IEICE ISS Journal, vol.17, no.4 (69), pp.21–22, 2013.

[2] N. Sonehara, "Big Data Driven ICT Innovation in Cyber-Physical Integrated Society -A Proposal on Identity Data Commons for Harmonizing Protection with Utilization," IEICE ISS-LOIS, 2014-01-LOIS-ICM, 2014.

[3] T. Kobayashi, Y. Ogawa, T. Suzuki, and H. Yamamoto, "News audience fragmentation in the Japanese Twittersphere," Asian Journal of Communication, vol.29, no.3, pp.274–290, 2018. DOI:10.1080/01292986.2018.1458326

[4] T. Suzuki, T. Kobayashi, and J. Boase, "Political Conversations as Civic Engagement: Examining Patterns from Mobile Communication log in Japan," Mobile Media, Political Participation, and Civic Activism in Asia, pp.57–79, 2016.

[5] Y. Ichifuji and N. Sonehara, "Visualization of Recovery Situation in Disaster Area by Using Web Reservation Data," Proceeding of 2014 IEEE DOI 10.1109/ARES.2014.67, pp.440–446, 2014.

[6] Y. Ichifuji and N. Sonehara, "Web data driven tourism policy supporting system," ESTRELA, no.266, pp.10–15, 2016.

[7] Y. Ichifuji and N. Sonehara, "Visualization for Social System Resilience by Web Reservation Data," The IEICE Trans. Inf. & Syst., vol.95(5)-D, pp.1100–1109, 2012.

[8] Y. Ichifuji and N. Sonehara, "Visualization of Recovery Situation in Disaster Area by Using Web Reservation Data," The 4th International Workshop on Resilience and IT-Risk in Social Infrastructures (RISI), Proc. of the 9th International Conference on Availability, Reliability and Security (ARES), pp.440–446, 2014.

[9] Y. Ichifuji and N. Sonehara, "Estimation of disaster recovery condition based on Web reservation data," 3rd UN WCDRR Public Forum 158, Workshop on Establishing Resilient Life-Space in the Cyber-Physical Integrated Society, 2015.

[10] N. Koide, Y. Ichifuji, H. Yoshii, and N. Sonehara, "Estimation of national tourism statistics based on Wi-Fi association log data," IEEE 2nd International Workshop on Big Data for Sustainable Development, 2016. DOI: 10.1109/BigData.2016.7840847

[11] A. Fukuzaki, N. Koide, and N. Sonehara, "TTPP method for validation of k-anonymity on Wi-Fi association log data," Proceedings of the 10th ACM Conference on Security and Privacy in Wireless and Mobile Networks, pp.275–276, 2017.

[12] N. Sonehara, N. Koide, and A. Fukuzaki, "Evidence based policy making and decision making," NII Press release, 2017. http://www.nii.ac.jp/userimg/press_20170314.pdf

[13] H. Fujii, Y. Mizoguchi, Y. Morie, H. Yoshii, and N. Sonehara, "On big data analysis integrating ticket gate counting data with Wi-Fi access log data," SIG IOT IPSJ, IOTS 2018, 2018.

[14] N. Sonehara, "SDC (Social Data Commons) to Form COI (Community of Interest) across Different Industries and Organizations," Organizational Science, vol.51, no.4, pp.15–23, 2018.

[15] Venue Vitalics™, https://cinarra.co.jp/solution/venue-vitalics/, Accessed on June 20, 2019.

[16] European Commission, "General Data Protection Regulation," 2016. http://data.europa.eu/eli/reg/2016/679/oj

[17] D.S. Terzi, R. Terzi, and S. Sagiroglu, "Asurvey on security and privacy issues in big data," 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST), 14–16 Dec. 2015. DOI: 10.1109/ICITST.2015.7412089

[18] N. Joshi and B. Kadhiwala, "Big data security and privacy issues A survey," 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), 2017. DOI: 10.1109/IPACT.2017.8245064

[19] Y. Tanaka, A. Kodate, and T. Bolt, "Data Sharing System Based on Legal Risk Assessment," Proceedings of the 5th Multidisciplinary International Social Networks Conference, Article no.17, 2018. DOI: 10.1145/3227696.3227715

[20] K. Imai, N. Sonehara, et. al., "Part2 the impact on society and economy," Research and Legislative Reference Bureau National Diet Library, Research Materials 2017-6, 2017.

Noboru Sonehara is a Professor of the College of Policy Studies and a director general of the Research Institute of Policy Studies at Tsuda University, and Honorary Professor of National Institute of Informatics. He is doing education and research on Social Big-Data infrastructure and Data-Driven Policy Sciences. He received the B.E. degree, the M.E. degree and the Ph.D. degree from Shinshu University in 1976, 1978 and 1994, respectively. He joined NTT, and has been engaged in R&D of facsimile communication and digital commerce systems. He is a member of IIEEJ and ITE, and a Fellow of IEICE.

Takahisa Suzuki is an Associate Lecturer of the College of Policy Studies at Tsuda University. He received M.E. from Tokyo Institute of Technology in 2008, PhD from the Graduate University for Advanced Studies, Japan in 2015. His research focuses on media effects and computational methods.

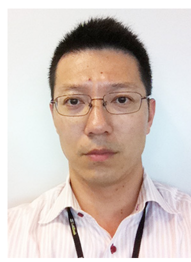
Akihisa Kodate is a Professor of the College of Policy Studies at Tsuda University. He received B.E., M.E. and D.E. degrees from Waseda University in 1992, 1994 and 2000, respectively. He was a researcher at University of Hannover, Germany and a visiting Associate Professor at Waseda University. He has been engaged in the research of information systems, such as content analysis and lifelogging services. He is a senior member of IEICE, a member of IEEE and IPSJ.

Toshihiko Wakahara is a Visiting Professor of the Research Institute for Policy Studies at Tsuda University. He received B.E., M.E. and D.E. degrees from Tokyo Institute of Technology, Tokyo, Japan in 1970, 1972 and 1998, respectively. He joined NTT Electrical Communication Laboratories in 1972 and was a Visiting Professor of Waseda University from 1999 to 2003 and Professor at Fukuoka Institute of Technology as a Professor from 2003 to 2018. He received the ISS Distinguished Service Award in 2015. He is a Fellow of IEICE, a Fellow of IIEEJ and a member of IPSJ.

Yoshinori Sakai is a Visiting Professor of the Research Institute for Policy Studies at Tsuda University and Emeritus Professor at Tokyo Institute of Technology. He received a B.S., M.S., and a Dr. of Eng. degree from the University of Tokyo in 1969, 1971, and 1974, respectively. He joined NTT in 1974, and since then has been researching PCM–FDM transmission methods, video teleconferencing methods, image encoding, and multimedia interfaces. He was Director of Tokyo Shibuya Study Center, the Open University of Japan and President of IEICE.

Yu Ichifuji is an Associate Professor of Center for Information and Communication Technology at Nagasaki University. He received B.E., M.E. and Doctor of Informatics from Tohoku University in 2003, 2005 and 2010, respectively. His research focuses on person trip pattern recognition of tourists using Wi-Fi system log.

Hideo Fujii is an Executive Chief Researcher, Kintetsu Railway Co. Ltd. His research focuses on effective utilization of corporate data and understanding of people flow.



Hideki Yoshii is a manager of Data Technology and Development Section, SoftBank Corp.. His research focuses on use and protection of mobile phone-based lifelog. He is a Super Creator for bigdata processing, analyzing and sharing, certificated by Information-technology Promotion Agency, Japan in 2006.