Data Jackets Evolving and Connecting via History of IMDJ

Yukio Ohsawa Teruaki Hayashi

[*1] Dept. Systems Innovation, School of Engineering, The University of Tokyo

Here we show the structural changes in the connections among (1) the variables in data jackets (2) words in the expected outcomes of data use, and (3) supposed data use (analysis/simulation) scenarios represented in the 1603 data jackets stored from 2014 through 2018 (business years). These respectively mean (1) the features of data to be provided or created, not necessarily on the allowance to share the data contents, (2) the social expectation about how and for what the data can be used if shared, and (3) the main process of data use. The changes in the structure of the co-occurrence are visualized by use of the traditional sequence of graphs with KeyGraph, where we found that the connections are emerging for the recent two years, that means the ideas of data users are evolving with communication in IMDJ.

1. Introduction

Since we proposed the Innovators Marketplace on Data Jackets (IMDJ) in 2013 [Ohsawa 13], a human-driven platform for creating strategic use scenarios of data and for evaluating the use value of data, participants came to range across various domains of businesses and sciences. The starting workshops in this platform were organized in graduate and undergraduate schools in Japanese universities and exported to other countries Taiwan, India, etc. Then, participants in business sections came to overwhelm academia, with still continuing the activities in scientific projects.

The key technologies in IMDJ here are, simply put, (1) data jackets, (2) human interface for aiding participants in their thoughts and communications with connecting various data without seeing or touching the contents of data [Hayashi 18], and (3) the designed/created data and analysis tools for using data. Among these three, (1) and (2) are the inputs to the IMDJ as a system, whereas (3) is an output. The outcomes of each workshop in IMDJ are mostly the proposed strategic scenarios for data use/reuse (often including data analysis and synthesis) and the results of acting on those scenarios. However, item (3) is also positioned as an outcome or a side product of IMDJ. The relation of these inputs and outputs can be expressed by Eq.(1) through Eq.(3), representing a set of data jackets in (1) by DJs, the use scenario in (2) (called a solution in [Ohsawa 13, Hayashi 18, etc]) by Sol, and tools in (3) by Tool. Also, the requirements spoken in IMDJ from the viewpoints of data users are represented here by Req. And, Act means humans' actions and communications in data processing.

$$\text{DJs, Sol} \rightarrow \text{Req} \tag{1}$$

$$\text{Tool, Act} \rightarrow \text{Sol} \tag{2}$$

Each element above can be expressed by predicates, where DJs takes variables in the data as attributes, Sol takes a part of DJs into its action (Act) and as the inputs to the tools (Tools) it employs in Eq.(2). That is, Eq.(1) and Eq.(2) can be put into

$$\text{provide}(V_1),\ \text{use}(V_1,\ \text{req}) \rightarrow \text{realize}(\text{req}). \tag{3}$$

$$\text{process}(V_1,\ V_2),\ \text{compute}(\cup_{v \in V2}\, v),$$
$$\text{relate}(\cup_{v \in V3}\, v,\ \text{req} \mid V_3 \subset V_2) \rightarrow \text{use}(V_1,\ \text{req}). \tag{4}$$

Here, V_1 in provide(V_1) represents the initially given set of variables in DJs, i.e., the set of data jackets, and the predicate provide means the content of the data supposed to be provided. use(V_1, req) means to realize Sol with the variables in V_1 to satisfy the requirement req in Eq.(3), a situation represented by Req in Eq.(1). The use scenario use(V_1, req) is composed by more granular components, as Sol is in Eq.(2), that is to first preprocess the variables (choose the useful subset of V_1 and further process the chosen set by such procedure as noise filtering). Then, the relations of the variables related relatively directly to the requirements (e.g., the instruction signal) and other variables are computed with automated computation which may be machine learning. Yet the learned relation is not always useful for satisfying the requirement: for example, products sold well in the previous year should be modified to fit the interests of present years' customers. This, as well as tools with AI (Tool in Eq.(2)), actions and communications of humans play the most essential role in the real businesses. Thus, process and relate are represented by Act in Eq.(2). As a result, in IMDJ, all elements composing Eq.(3) and Eq.(4) should be communicated. In summary, the following are communicated in IMDJ.

Provide: to share/get the data to be used

Variables: the variables in V_1 to be given for use/reuse

Process: how the variables and their values are collected or selected, not only for learning efficiently but for fitting req. V_2 may include variables not included in V_1.

Compute: choice of the tools for computation e.q., tools with AI.

Relate: the action to activate the causality between V_3 and req. If V_3 does not fully explain the ways to satisfy req, additional actions should be discussed.

2. The Contents of Data Jackets

Each data jacket (DJ hereafter) is provided not only as the abstract of an existing dataset but also as the preparatory information for the communication exchanging or creating the above information. Therefore, a DJ may include these pieces of information by human-written text. It is certainly allowed that one dataset has multiple DJs by different authors because a DJ is desired to reflect the subjective thoughts or the vision of a participant in IMDJ. Thus, in the entry sheet of each DJ, the following pieces of information and comments are requested.

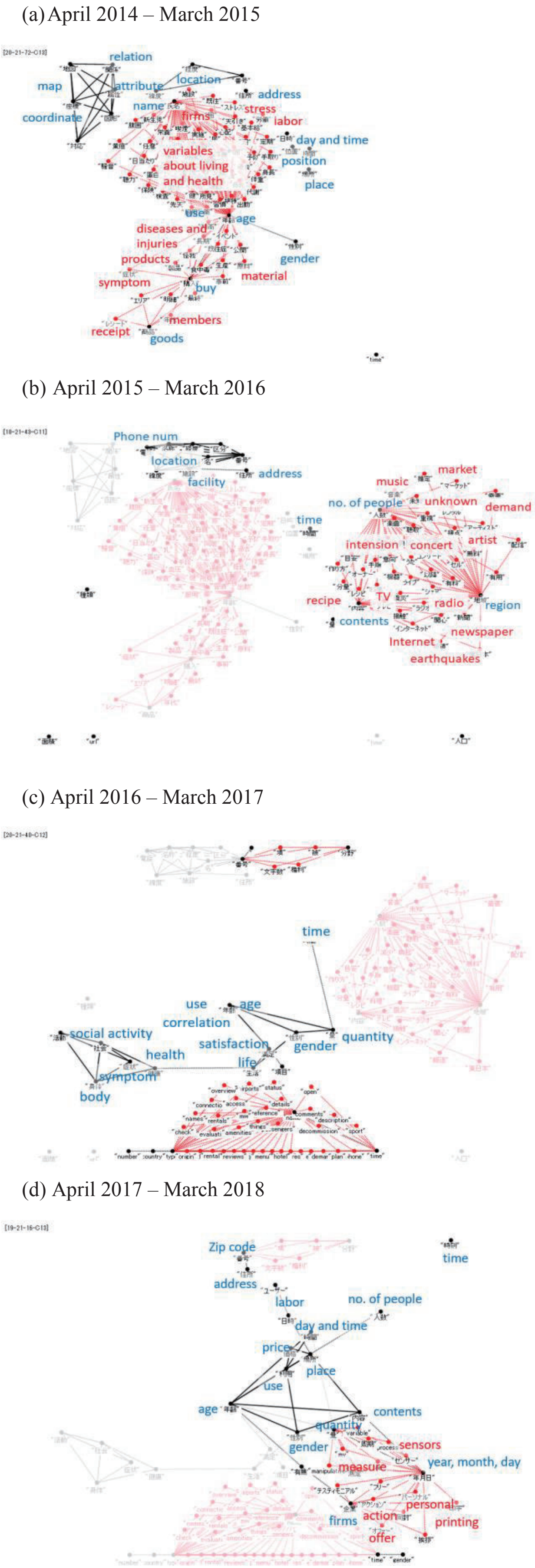
Contact: Ohsawa Lab, 7-3-1 Hongo, Bukyo-ku 113-8656 Tokyo, info@pands.sys.t.u-tokyo.ac.jp

Title of the data (e.g., "POS data in a supermarket")
Outline (e.g., "the sequence of goods bought by customers who come to buy consumption items")
Collecting Cost (e.g., "depends on the system.")
Sharing Policy (e.g., "can be shared after negotiation, if the condition fits the business of the supermarket.")
Types (e.g., "a table of text, numbers, and symbols.")
Formats (e.g., "XML")
Variables/Attributes (e.g., "product class, product category, product name, customer ID, and date and time of purchase.")
Analysis /Simulation (e.g., "apply sequence analyzer such as RNN or visualizer such as Tangled String, for predicting and/or explaining the future strategies in marketing. Predicting may not be trustworthy because the market in the past and the future tend to differ, but the explanation is absolutely required.")
Outcome (understanding the reasons for customers' behaviors)
Anticipation (expected strategies to satisfy existing and forthcoming customers, to be obtained from the data)
What kind of data/tools do you wish to have? (additional data or AI tools that should be combined with the current dataset)

The information corresponding to **Provide** in Section 1 is included in **Collecting Cost** and **Sharing Policy** here. **Variables** in Section 1 are in **Variables/Attributes** here. **Process** and **Compute** in Section 1 are provided in **Analysis /Simulation** here that includes to select variables and use tools with AI, because it is not easy to separate this preprocess and the main step of computation because ordinary users tend to trust experts or automated tools as a package for setting hyperparameters in the preprocess and learning (computation). It is noteworthy here that IMDJ is a platform to connect requirements of sheer users of data and AI tools, and the details of settings in the post-process are dealt with in the later step of action planning. Finally, **Relate** in Section 1 is expected to be filled in **Outcome** and **Anticipation** above, although the entries to the blanks for these two items are optional. In summary, from the viewpoint to detect the social interaction of the data use/reuse process, the three items **Collecting Cost, Sharing Policy, Variables/Attributes** represent the pre-process, whereas **Outcome** and **Anticipation** represent the post-process. **Analysis/simulation** shows a part of the main process.

3. The visualized changes in DJ for 5 years

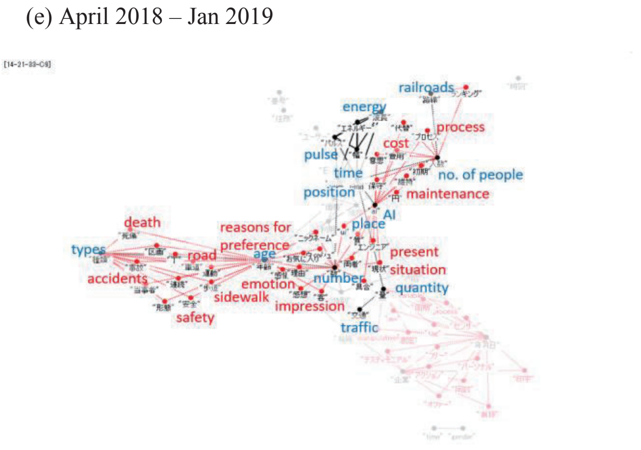
Kamishibai (K)-KeyGraph shows the structural variation of the co-occurrence graphs of items in the partial data for each period, in a given sequence [Ohsawa 10]. The black nodes with blue letters show frequent words, connected by red ones represented by less frequent words. The sequence of visualized graphs is shown for explaining the structural changes, but the distances such as obtained from the coordinates in distribution representation vectors are not reflected here. This visualization rule is common in Figure 1, 2, and 3. The same item appears at the same position in the 2D coordinate, so the changing is intuitively grasped at a glance. In Figure 1, 2, and 3, respectively, the changing in the connections among the variables in **Variables/Attributes (Collecting Cost, Sharing Policy** are cut because of the difference in the granularity of information**)**, among words in **Outcome** plus **Anticipation,** and among words in **Analysis /Simulation process**, are visualized.


(a) April 2014 – March 2015

(b) April 2015 – March 2016

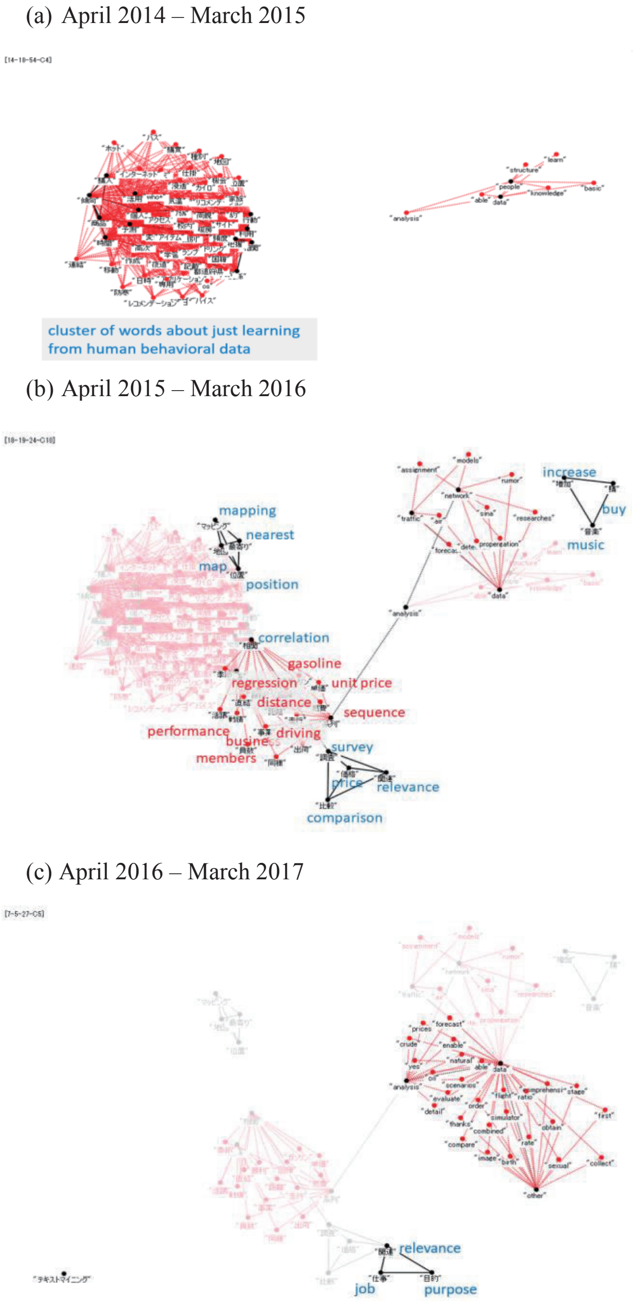
(c) April 2016 – March 2017

(d) April 2017 – March 2018



(e) April 2018 – Jan 2019

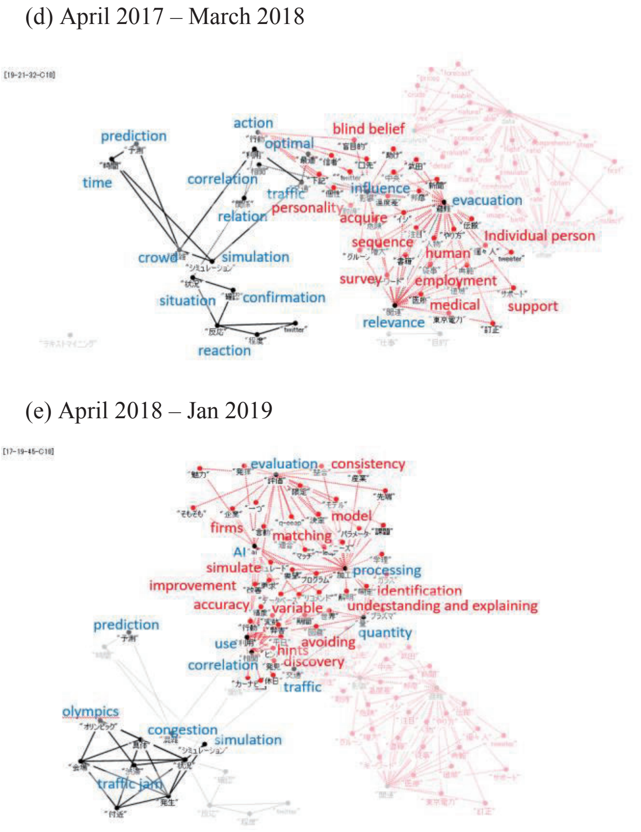
Figure 1. The changes in the connection of variables in DJs.



(a) April 2014 – March 2015

(b) April 2015 – March 2016

(c) April 2016 – March 2017



(d) April 2017 – March 2018

(e) April 2018 – Jan 2019

Figure 2. The changes in the connection of words in the expected outcomes of data use, described in DJs.

4. Discussions

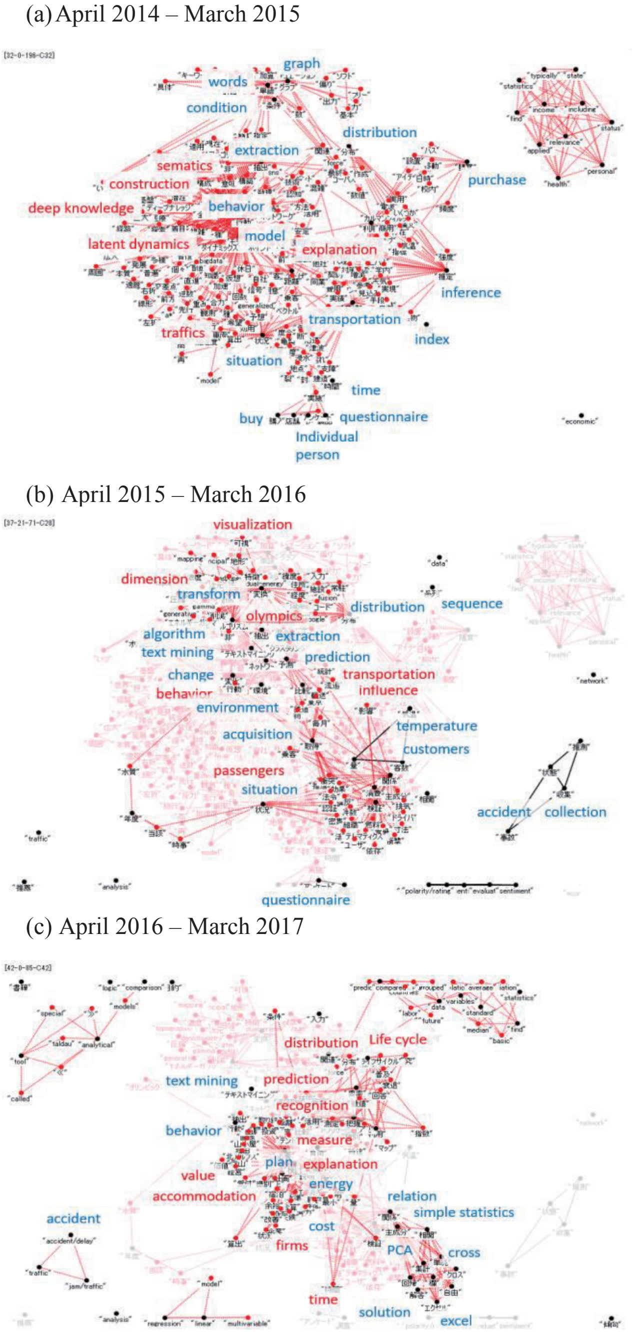
The numbers of nodes and edges in graphs are set equally for all periods (years), to make the comparison for different periods. As in Fig. 1 (a), the connections among the variables have been sparse in 2014, in that the human (personal) data and geographic data have been separated into clusters. This tendency lasted till 2015, but the human data and physical data including data from sensors have got coupled into a cluster in 2017. Then, a larger cluster in 2018 appeared that integrates the human data and physical data into AI-driven innovations toward businesses with emotional designs and marketing. The cluster that emerged at the center of Fig.1 (c) for 2016 may be interpretable as a transient path from the localized clusters in (a) to the grown-up one in (e). This integration of clusters is also found in Fig.2. In Fig.2(d) for the year 2017, the simulation and data analysis are coupled into a cluster on services for health care, simulation-based urban traffic management [Yoshimura 17], disaster protection etc., linked to show social demands for AI-driven innovations in Fig.2 (e).

Such a connection of local clusters is not so obvious in Fig.3. However, the last cluster in (e) connect visions toward data-driven innovation and the full sequence is coherently addressed to realize “explanation” as well as “prediction” from sequential data (highlighting spaciotemporal information), putting these keywords at the center of the graphs. The last Fig. 3(e) involves “theory” used for two meanings: (1) mathematical models for investigating latent dynamics [Yamanishi 15], and (2) the respect to domain knowledge highlighted in model-based sciences [Magnani 17], corresponding to the return to “model” in Fig.3 (a) (the assumption of low-dimension causality in using sparse data [Rish 14] is itself a model that may not work in explaining behaviors in the complex interaction with unexpected events).

5. Conclusions

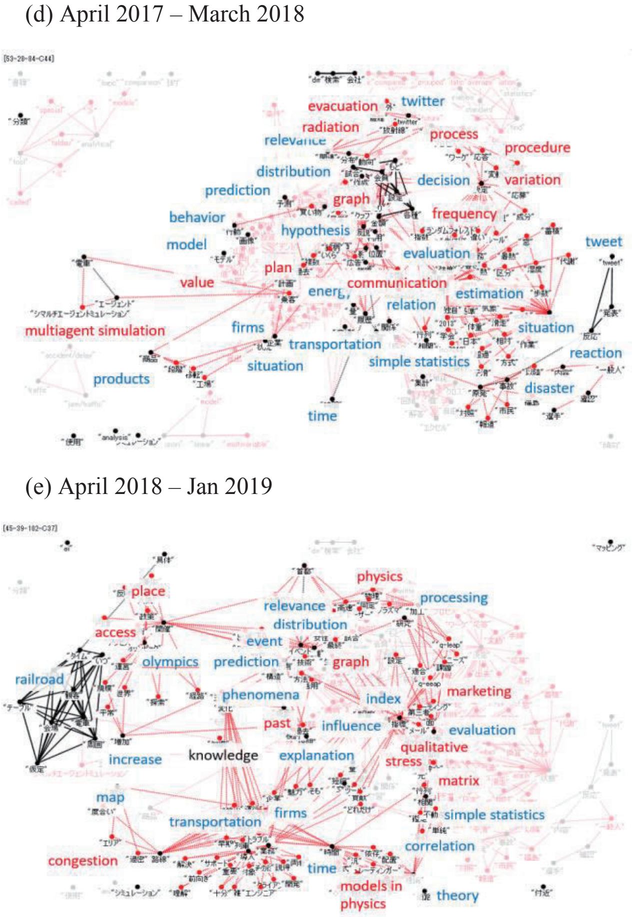
Above we find the human-centric data engineering process has been considered so far in IMDJ, reflecting the communication onto the entries of new DJs. DJs came to be included in the DTA standard of data catalog. We assume this reflects the potential expectation of data users, whose communication in the market of data triggers the evolution of representation of data.

Acknowledgments This work was supported by JST CREST JPMJCR1304, JSPS KAKENHI JP16H01836, JP16K12428, and industrial collaborators in the DJ workgroup.


(a) April 2014 – March 2015

(b) April 2015 – March 2016

(c) April 2016 – March 2017


(d) April 2017 – March 2018

(e) April 2018 – Jan 2019

Figure 3. The changes in the connection of words in the analysis/simulations described in DJs.

References

[Hayashi 18] Hayashi, T., Ohsawa, Y., "Inferring Variable Labels Using Outlines of Data in Data Jackets by Considering Similarity and Co-occurrence," *Int'l J. of Data Science and Analytics* 6(4), 351-361 (2018)

[Magnani 17] Magnani, L., Bertolotti, T., Springer Handbook of Model-Based Science (2017)

[Mudinas 18] Mudinas, A, Zhang, D., Levene, M., Market Trend Prediction using Sentiment Analysis: Lessons Learned and Paths Forward, in *Proc. Wisdom* 2018, London.

[Ohsawa 13] Ohsawa, Y., Hayashi, T., Kido, H., Liu, C., et al, Data Jackets for Synthesizing Values in the Market of Data *Procedia Computer Science* 22, 2013, pp. 709-716

[Ohsawa 10] Ohsawa, Y., Ito, T., and Kamata. IM. Kamishibai KeyGraph: Tool for Visualizing Structural Transitions for Detecting Transient Causes, *NMNC* 6 (2) 1-15 (2010)

[Rish 14]Rish, I., Grabarnik, G.A., *Sparse Modeling Theory, Algorithms, and Applications,* CRC, 2014.

[Yoshimura 17] Yoshimura, S., Fujii, H., "MATES: Multi-Agent based Traffic and Environmental Simulator -Core Technologies and Practical Applications-", 8th Int'l Conf. Computational Methods (ICCM 2017) (2017).

[Yamanishi 15] Hayashi, Y., Yamanishi, K., Sequential network change detection with its applications to ad impact relation analysis, *KDD* 29(1) 137–167（2015）