# Knowledge Structuring for Cross-disciplinary Data Exchange and Collaboration

**Teruaki Hayashi**

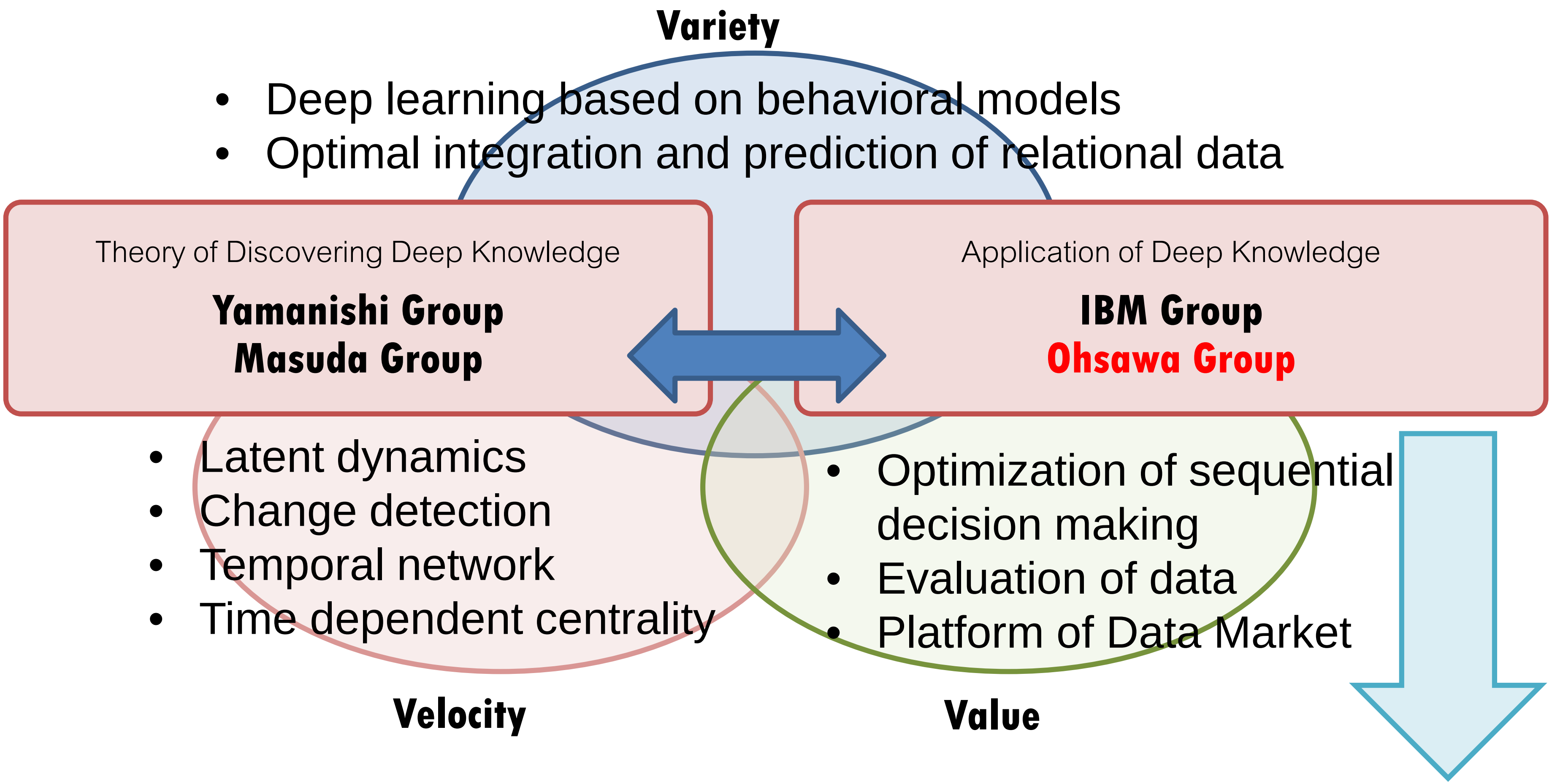
(Ohsawa Group, Yamanishi Team)

Department of Systems Innovation
School of Engineering
The University of Tokyo

**Yamanishi Team**

Discovering Deep Knowledge from Complex Data

and Its Value Creation




**Ohsawa Group**

Methods for Creating Data Market to Evaluate Utility Value of Data and Deep Knowledge

# Introduction

## Big Data

- Increasing capacity of storage
- Analysis method for heterogeneous data

## Personal Devices

- High granularity personal data
- Purchase logs, life logs, etc.

## Open Data

- The use of secondary data
- Massive amounts of data from the governments are available

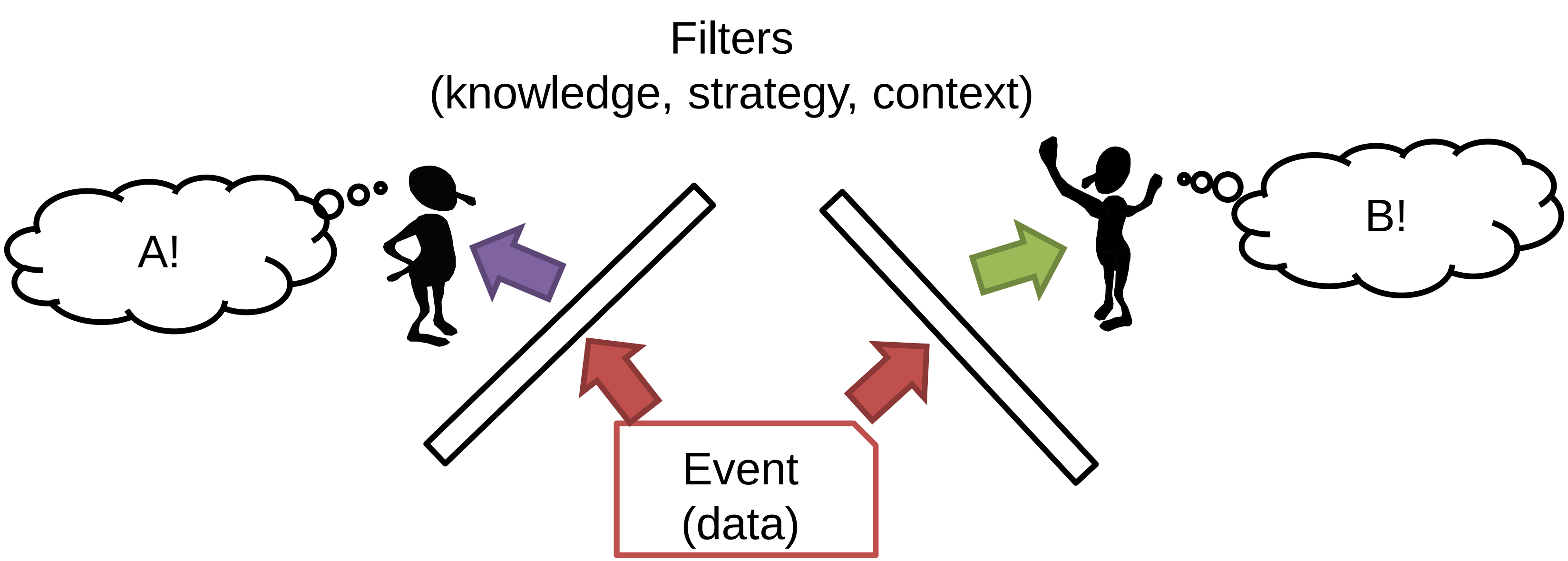
## Sensors

- Internet of Things
- High density data are available

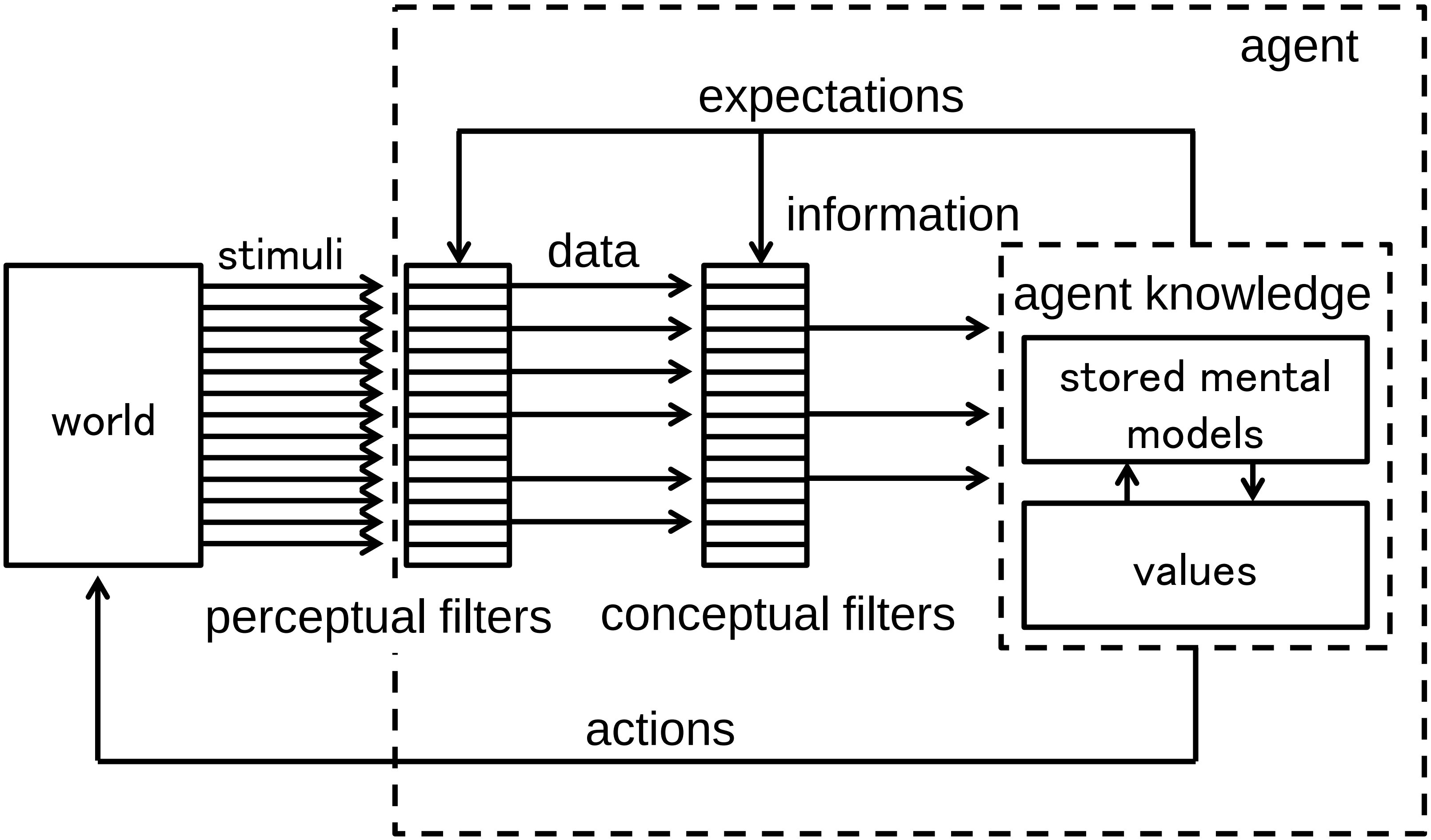
The potential benefits of reusing and analyzing massive amounts of data have been discussed by various stakeholders from diverse domains.

# Why Data Exchange?

- Decision makers in the society recognize the different world, even though they see the same world (Metcalfe, 1998)
  - background knowledge and available opportunities
- The two problem solvers may construct different facts even if they observe the same event (data) (Hayashi et al., 2006)
  - the different perspectives, contexts and background knowledge

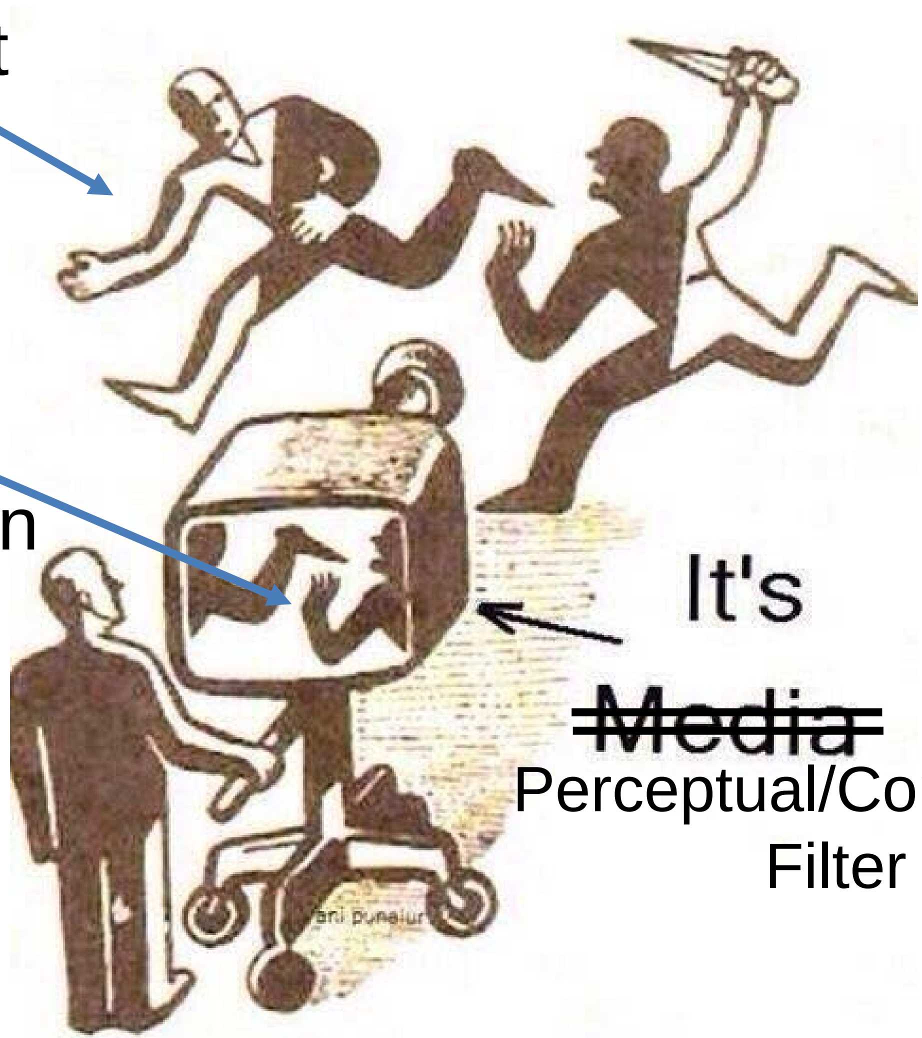


# Why Data Exchange?




The agent-in-the-world（Boisot & Canals, 2004）

# Why Data Exchange?




Data exchange are important to recognize the world correctly, and to encourage the cross-disciplinary data driven innovation.

# Our Approach

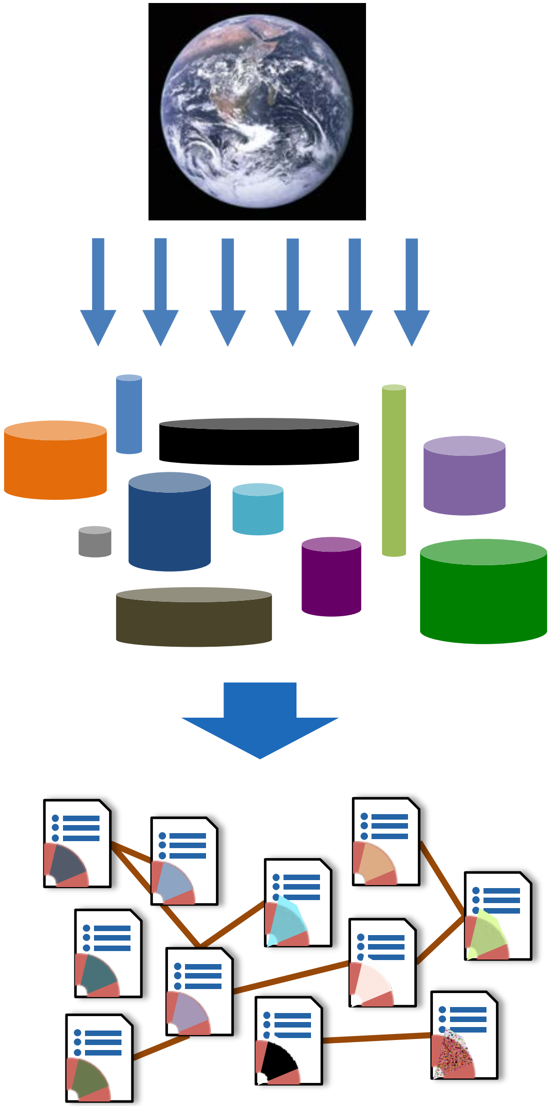
To encourage cross-disciplinary data exchange and collaboration…

It is important to understand the events in the world and the relationships of obtained data correctly.

Analyzing the structural features of the population of data from different domains, rather than analyzing individual data

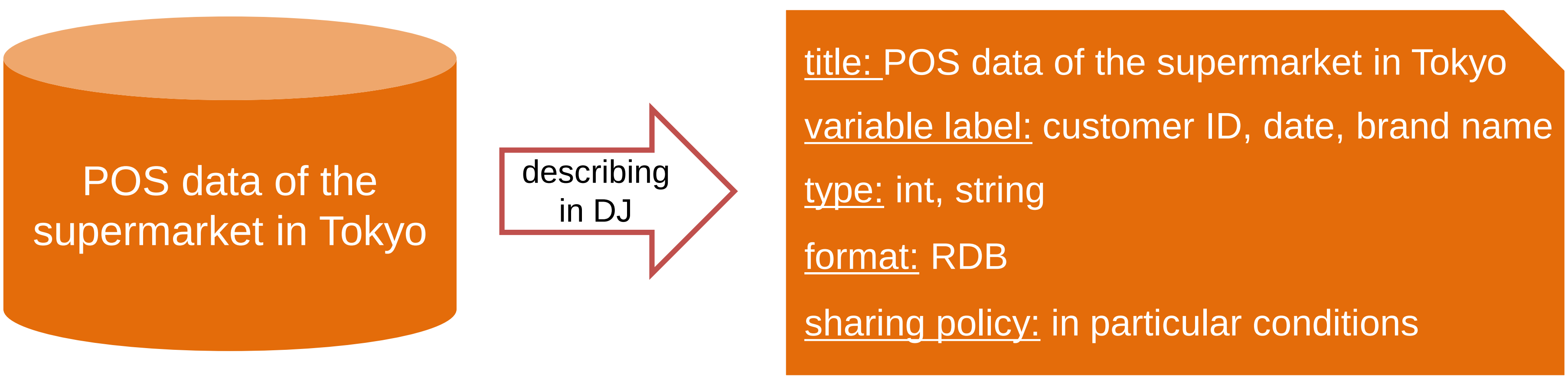
A data model to discuss different data on the same field is necessary to quantitatively evaluate the trends and features of the population of data.

It is effective to use metadata (data of data) obtained from different domains as the analysis subject.



# Data Jacket (DJ)

- **a structured summary of data described in natural language**
- DJ has been developed as a technique for sharing information about data and for considering the potential value of datasets, with the data itself hidden.
- Even if data itself is not open, by publishing DJ, data could be recognizable and understandable not only for humans, but also for machines.
- Published DJs enable data owners, data users and data analysts to understand the contents of each dataset, and start to communicate about data utilizations among stakeholders.




Y. Ohsawa et al., "Data Jackets for Externalizing Use Value of Hidden Datasets," 18th International Conference on Knowledge Based and Intelligent Information and Engineering System (KES2014), pp.946-953, Procedia Computer Science, Vol.35, pp.946-953, 2014.

# Examples of DJs

## Earthquake and related disaster information

**ID**

262

**OUTLINE**

This is the summary of earthquakes and related disasters. Will be updated for several times after the first release as new knowledge is obtained through investigation.

**VARIABLE LABEL**

Summary of the earthquake | Date and time of occurrence | Depth | Magnitude | Abolition time

Latitude (20:25 '14'' to 45.33 '19'') | Longitude (122:55 '59'' to 153:59 '25'') | Number of injured

Number of demolished houses | Number of partial damaged houses | Installation time | Epicenter

Local seismic intensity around the country | Number of deaths | House damage | Other damages

Presence or absence of the tsunami | Number of missing | Number of houses with the floor flooded

Number of inundation above floor level | Emergency Response Headquarters installation conditions

**SHARING POLICY**

With anyone

**COLLECTING COST**

Available at the website of Fire and Disaster Management Agency of Japan.

**FORMAT**

PDF

**TYPE**

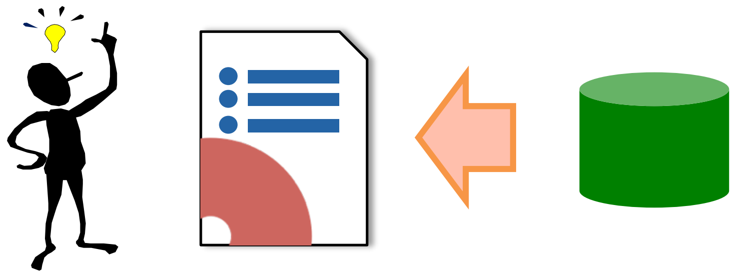
TEXT, TABLE

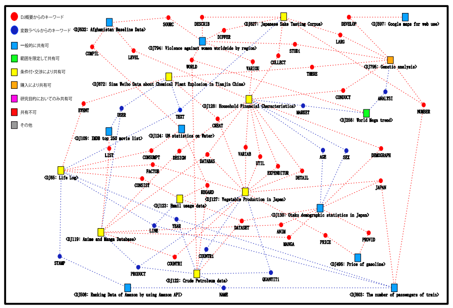
# What We can do with DJs

We can…

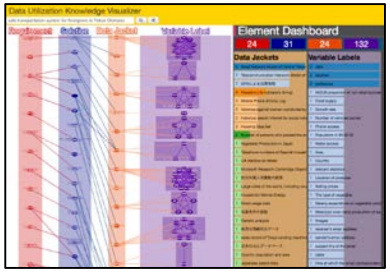
- understand who owns the data we are interested in.
- handle datasets in standardized format by describing each dataset in metadata.

**Understanding Data**





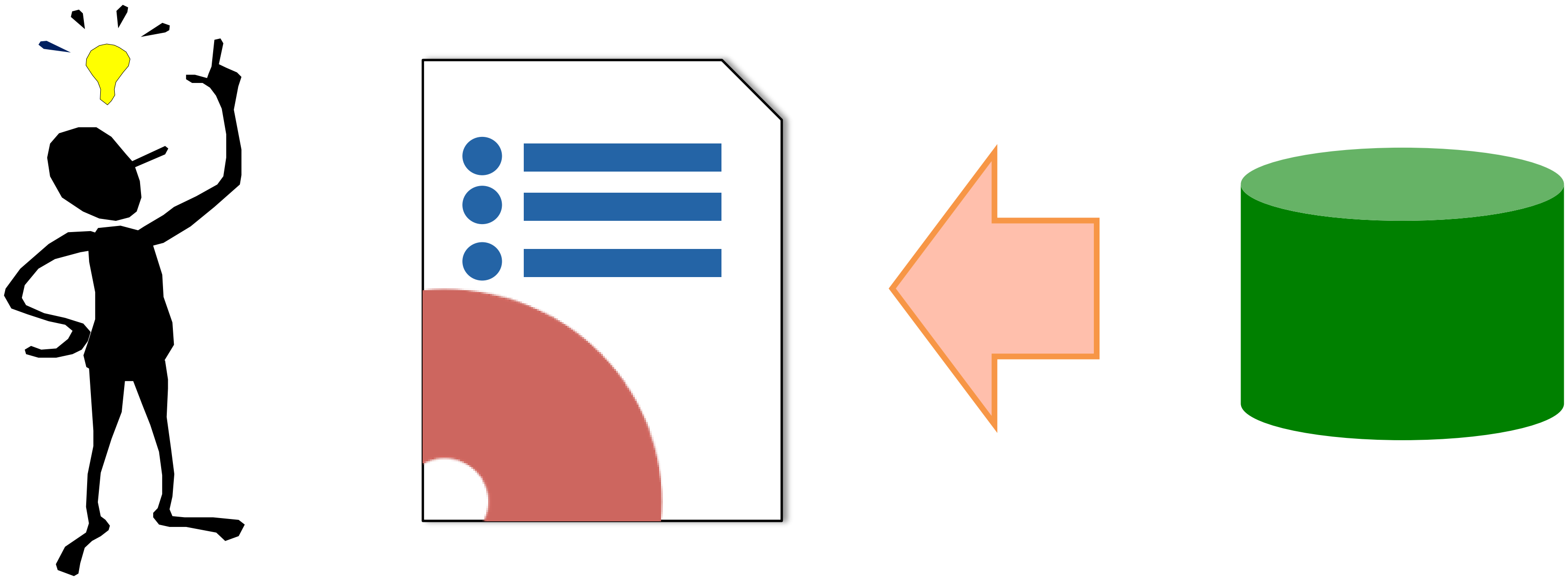
**Support Systems**





**Discussion for Data Utilization**

# Understanding Data

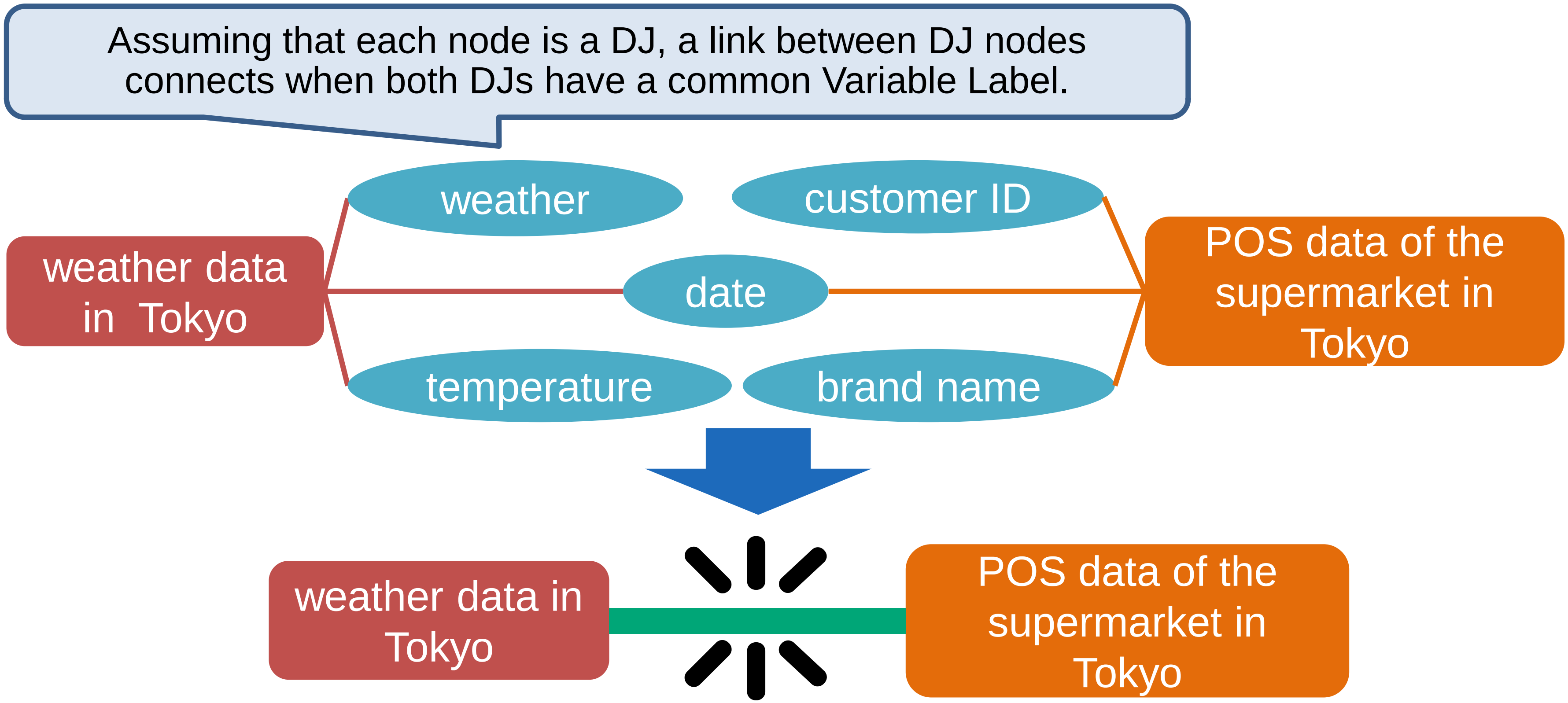


# Visualization of Data

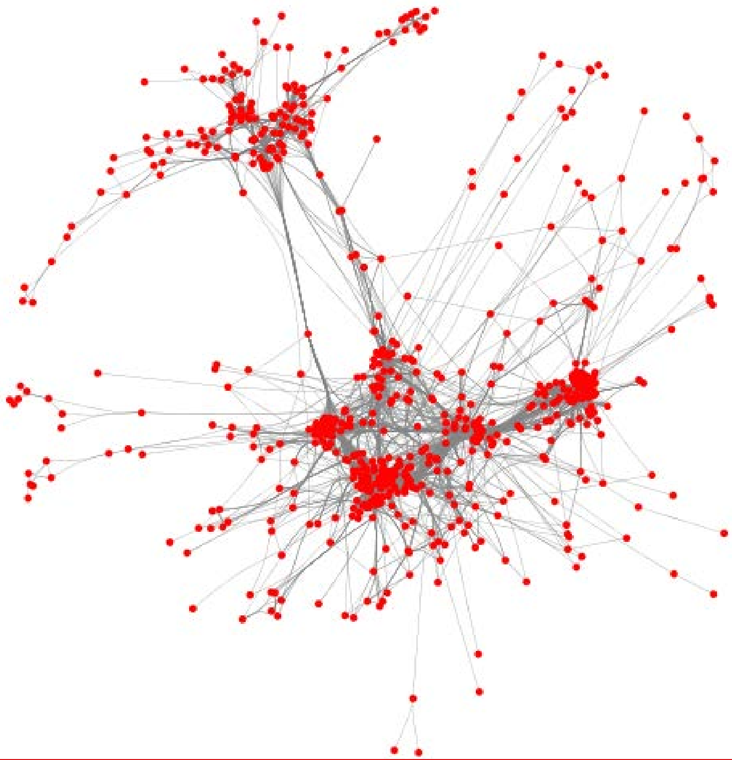
**Linkage of Data:**

**achieved by the combinations of variables in data**

datasets having common Variable Labels are highly likely to be combined



# Visualization of Data



| features | value |
|---|---|
| The number of links | 11077 |
| The number of nodes | 652 |
| Average degree | 33.98 |
| Density | 0.0522 |
| Average cluster coefficient | 0.703 |
| assortativity | 0.561 |
| diameter | 11 |
| Average shortest path | 3.442 |

**High average cluster coefficient and low density**

- close each other locally and sparse globally
- Similar data tends to connect strongly and consists the locally dense network

**Shorter average shortest path and high assortativity**

- The structure is similar to the network of human relations

# Visualization of Data

## Degree centrality

The indicator that how a node has linkages with other nodes

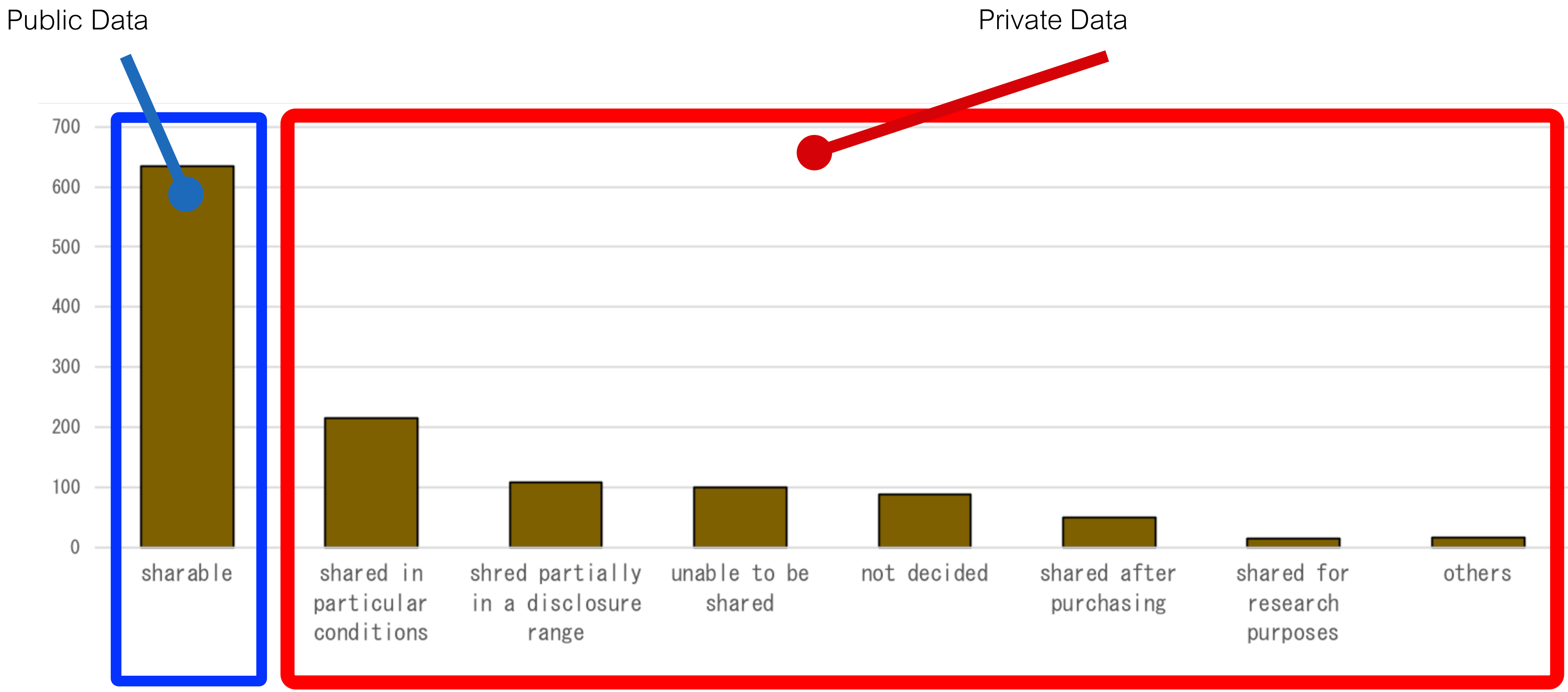
| Data name | Value |
|---|---|
| Facebook data | 0.192 |
| The locational information of public toilets | 0.187 |
| Twitter data | 0.186 |
| The data of earthquakes | 0.181 |
| The observation data of ozone layers | 0.180 |

## Betweeness centrality

The indicator that how a node bridges the nodes of other groups

| Data name | Value |
|---|---|
| The traffic data of highways | 0.221 |
| Social Networking Service data | 0.071 |
| Sales data of food consumption by areas | 0.064 |
| Happiness around the World | 0.054 |
| The book records data | 0.035 |

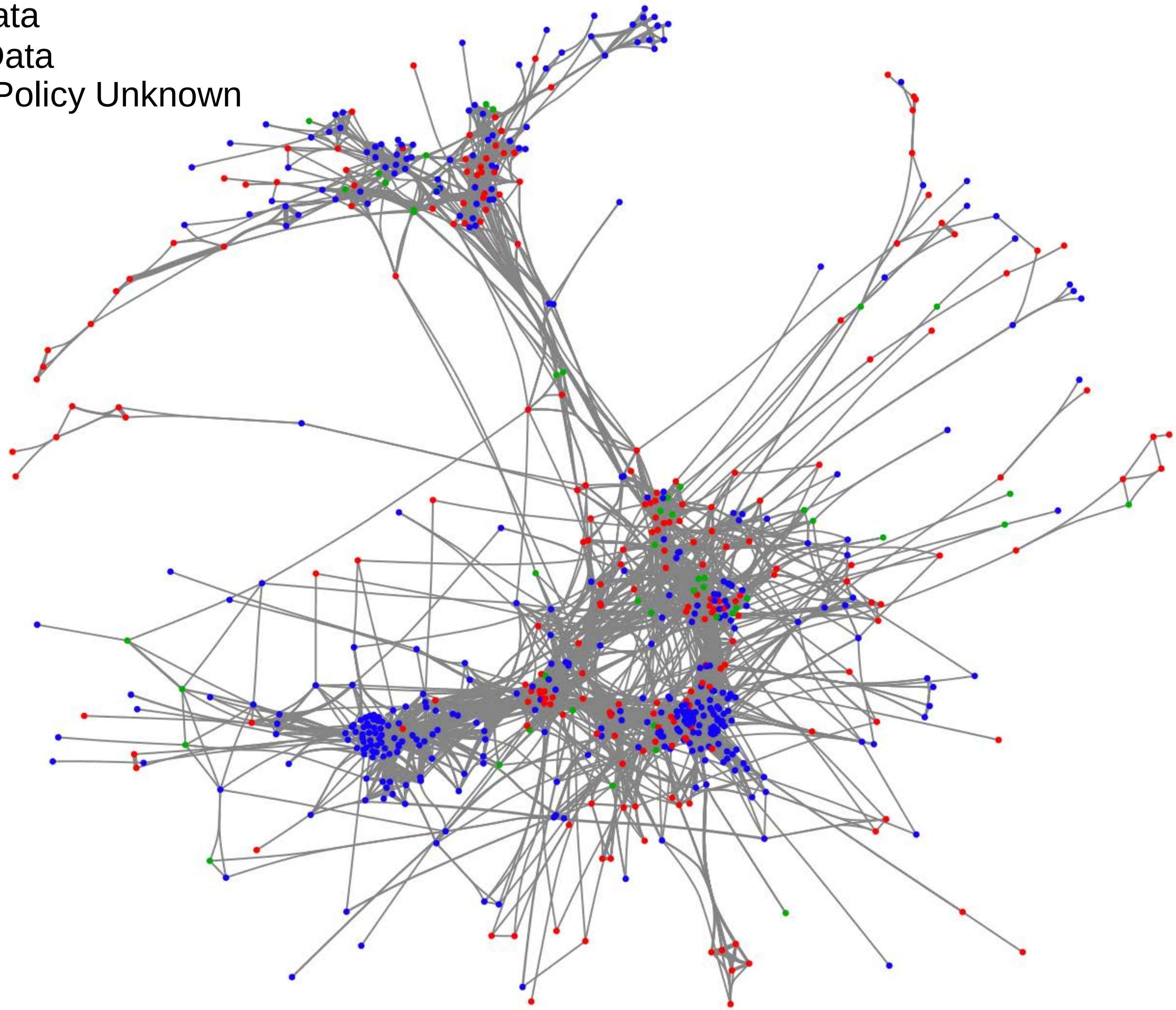
# Network Analysis with Sharing Policy




- Data is the economic goods in the Data Market
- The sharing policy is one of the important attributes of the data
- We analyzed the network of data considering the sharing policy

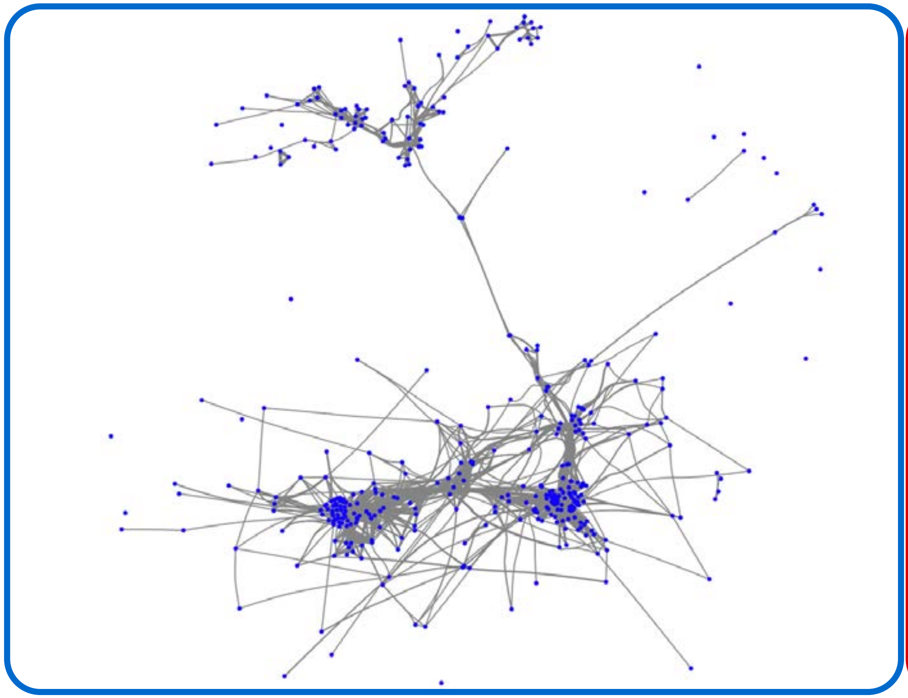
# Network Analysis with Sharing Policy

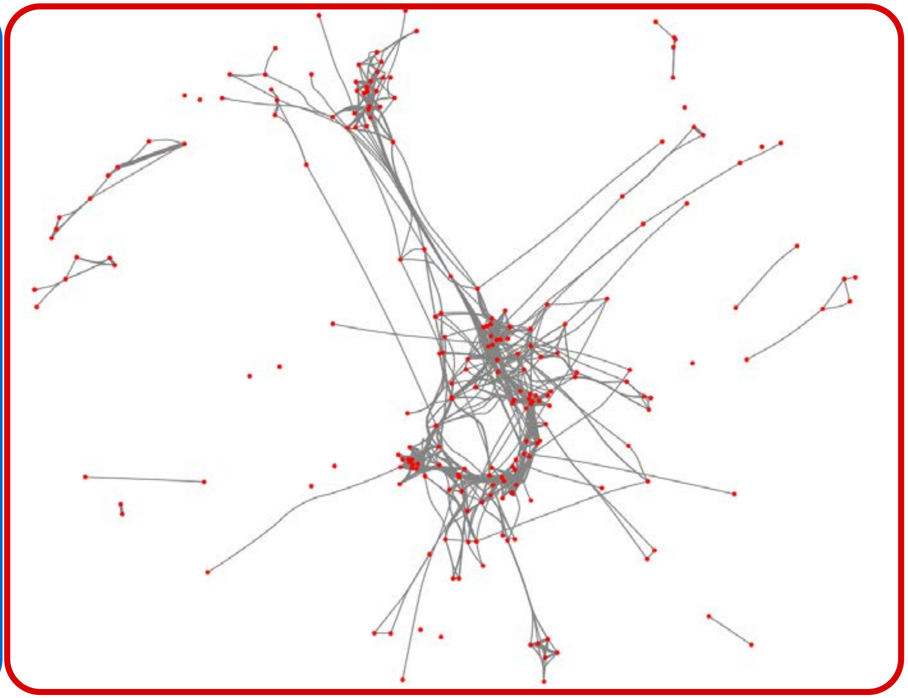
: Public Data
: Private Data
: Sharing Policy Unknown



# Network Analysis with Sharing Policy

| Features | Public Data | Private Data |
|---|---|---|
| The number of links | 5931 | 1068 |
| The number of nodes | 388 | 216 |
| Average degree | 30.57 | 9.89 |
| Density | 0.079 | 0.046 |
| Average cluster coefficient | 0.719 | 0.611 |
| assortativity | 0.556 | 0.456 |

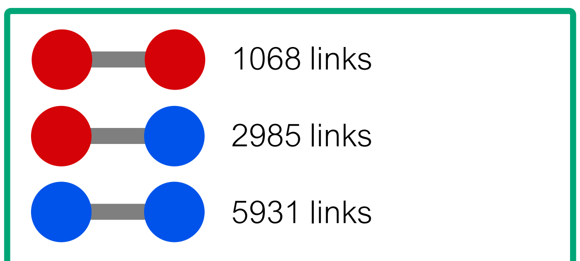


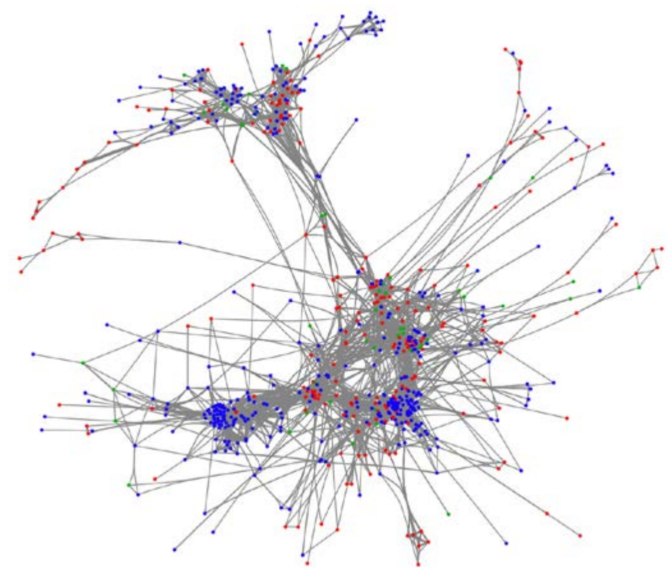


# Network Analysis with Sharing Policy

| Degree Centrality | Betweenness Centrality |
|---|---|
| Facebook data | The traffic data of highways |
| The locational information of public toilets | Social Networking Service data |
| Twitter data | Sales data of food consumption by areas |
| The data of earthquakes | Happiness around the World |
| The observation data of ozone layers | The book records data |

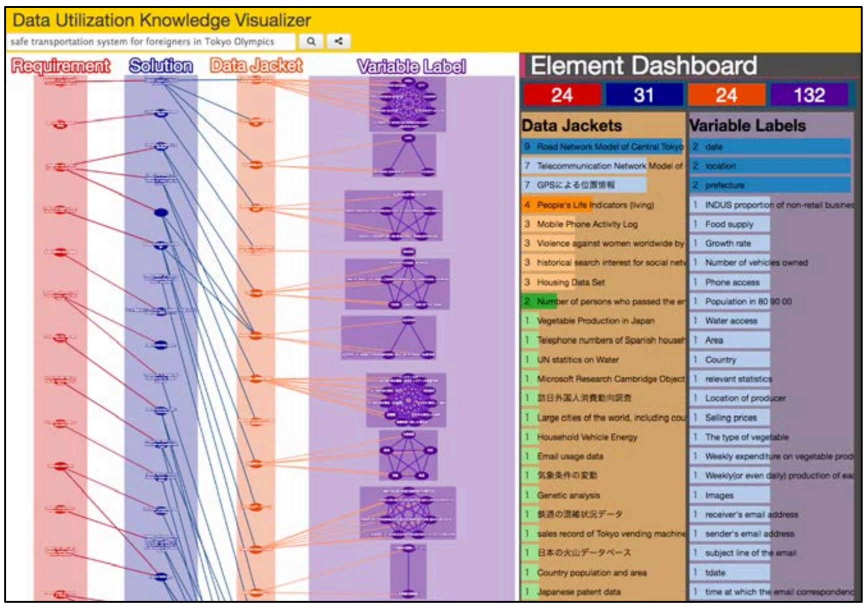
**Private Data exists more between Public Data than between Private Data.**

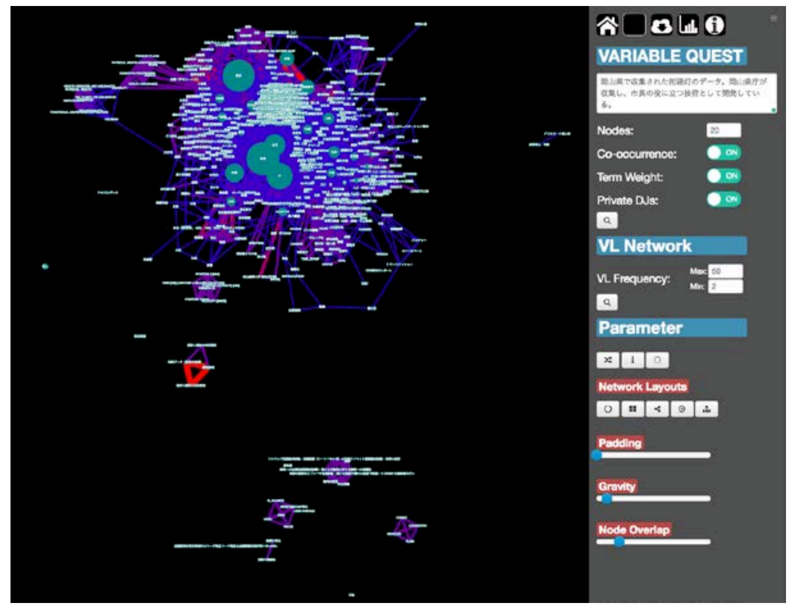


This result suggest that Private Data may play a role of combining data of different areas.

# Support Systems

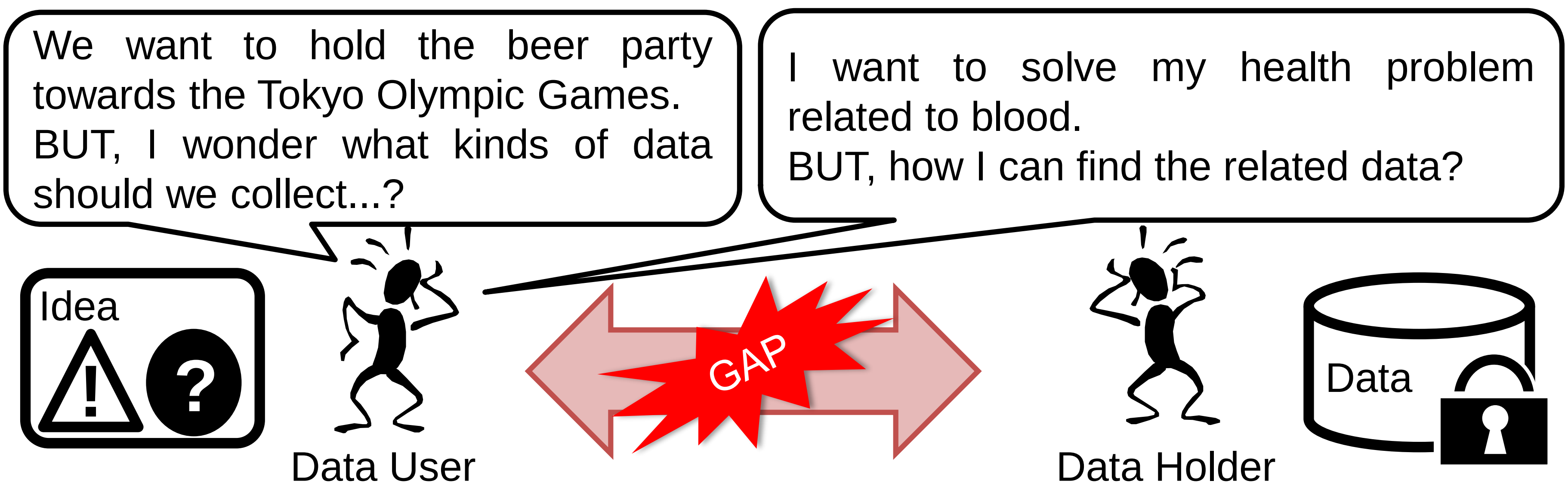


# Necessity of Support Systems

It is difficult for users to accurately obtain data corresponding to their own interests.

- the combination of databases may occur a serious violation of privacy (Acquisti & Gross, 2009; Xu et al., 2014)
- the size of the datasets is meaningless, and understanding the values of small data is more important (Boyd & Crawford, 2012)
- It is difficult to learn the kinds of data that are related to our interests as well as the means to obtain and utilize them (Hayashi & Ohsawa, 2016).

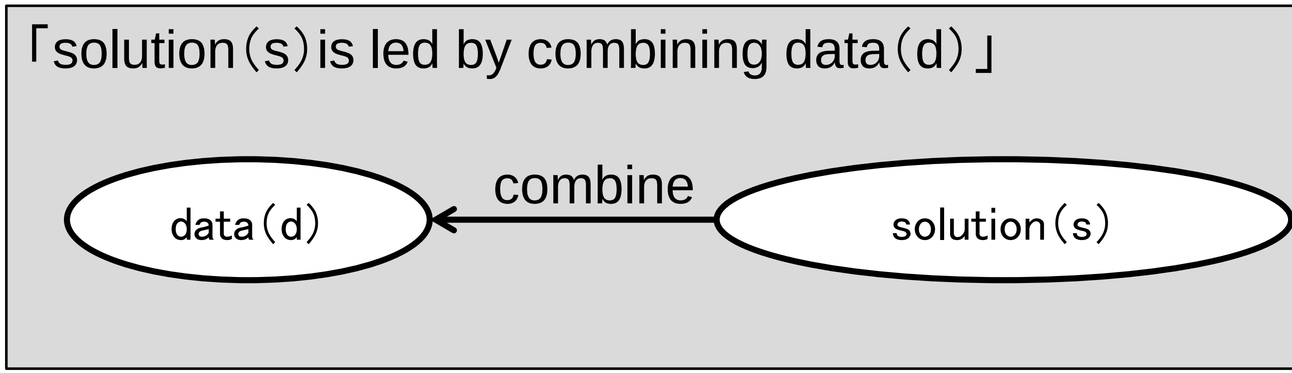


# Retrieval of Data

Structuring Knowledge of Data Utilization created in the data utilization workshops (Innovators Marketplace on Data Jackets)

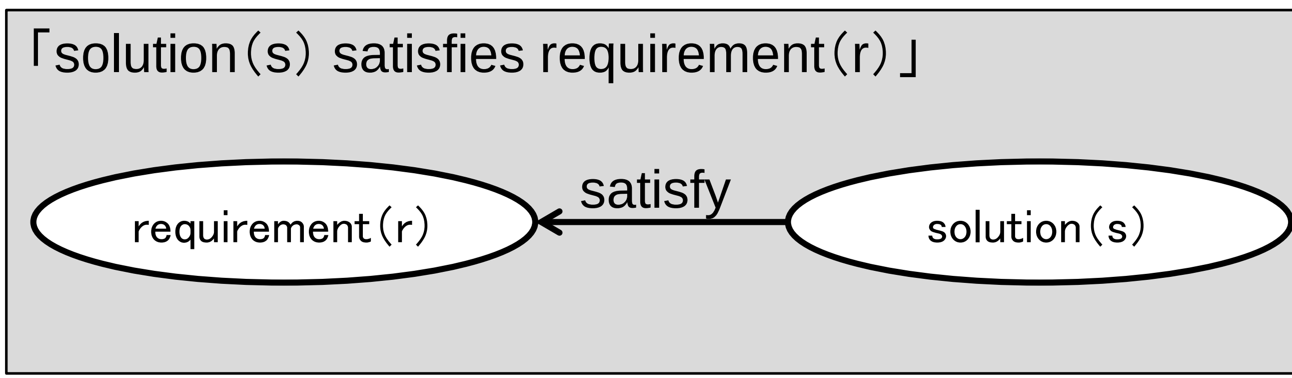
data utilization knowledge 1

***combine***$(s, d)$



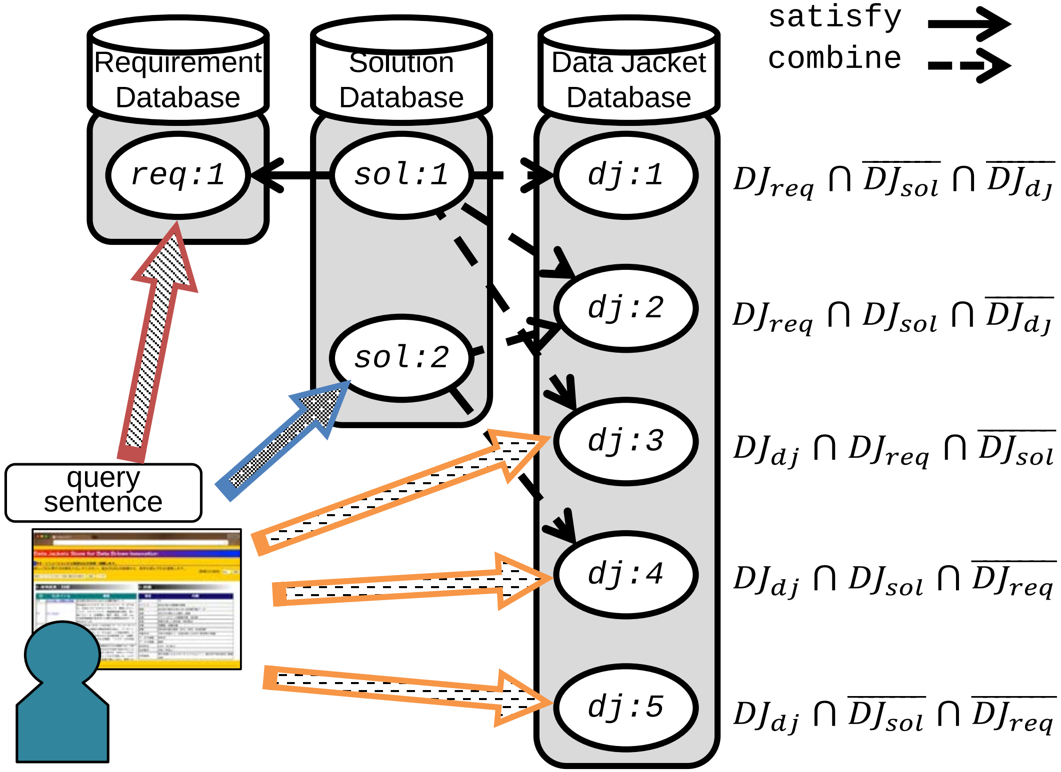

data utilization knowledge 2

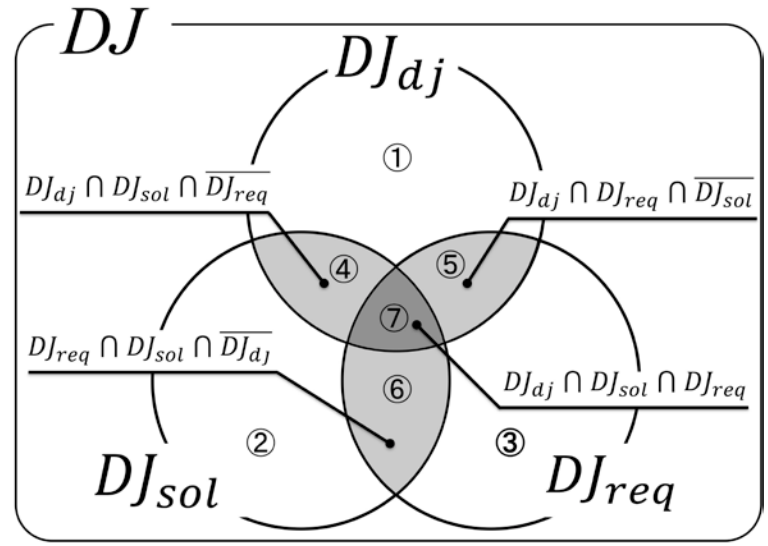
***satisfy***$(s, r)$




Reusing data utilization knowledge may be useful for Data Users to retrieve information about data related to their interests.

# Retrieval of Data




query sentence: $D_i=\{word_1, word_2, \cdots, word_j\}$ $(i, j \in N)$

return from DJ database: $DJ_{dj(Di)}=\cup_{j \in N} DJ_{dj(wordj)}$

return from Sol database: $DJ_{sol(Di)}=\cup_{j \in N} DJ_{sol(wordj)}$

return from Req database: $DJ_{req(Di)}=\cup_{j \in N} DJ_{req(wordj)}$

Set of DJs considering the numbers of retrieving.

T. Hayashi, Y. Ohsawa, "Data Jacket Store: Structuring Knowledge of Data Utilization and Retrieval System," Transactions of the Japanese Society for Artificial Intelligence, 31 (5),2016.

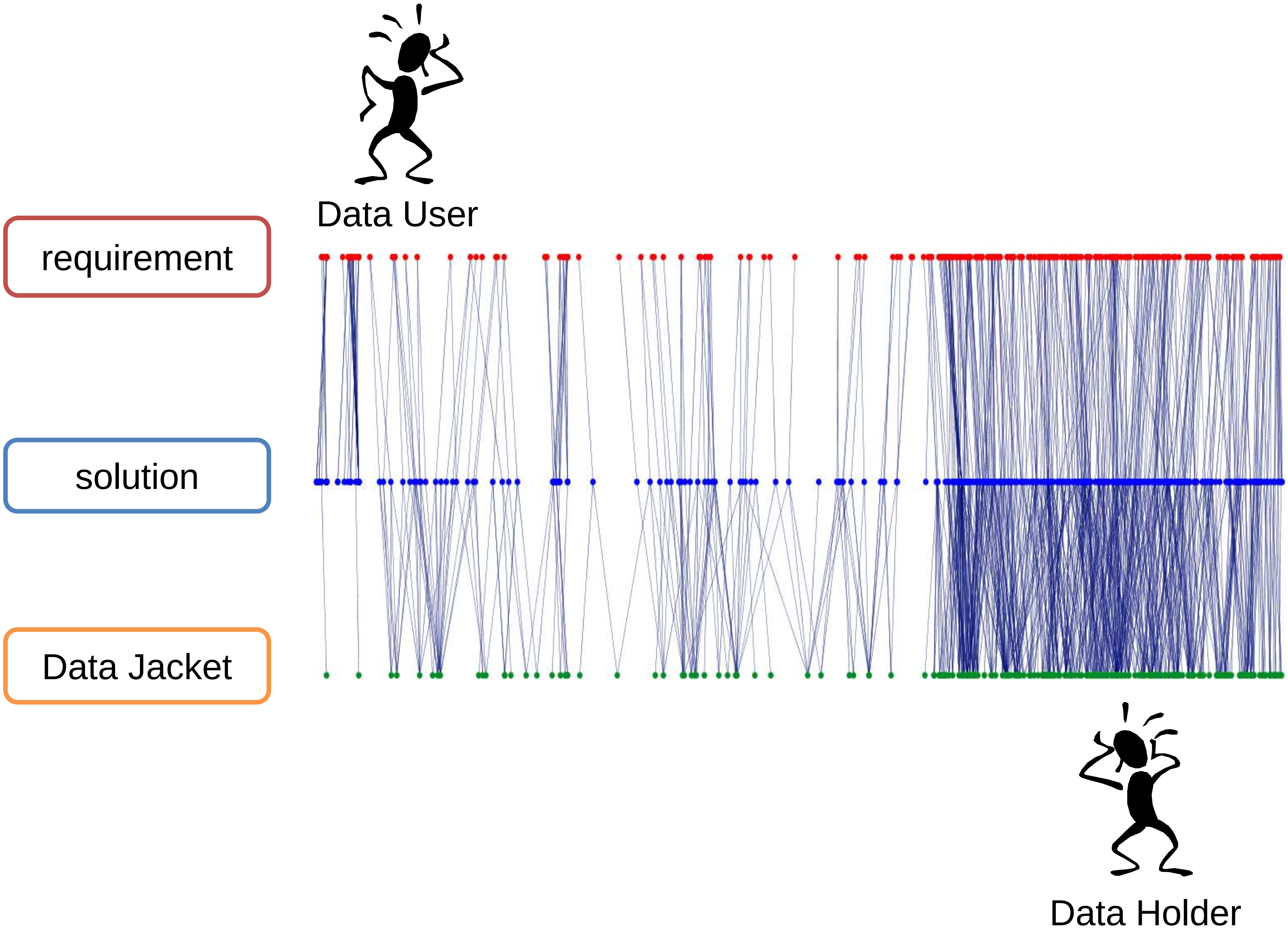
# Implementation

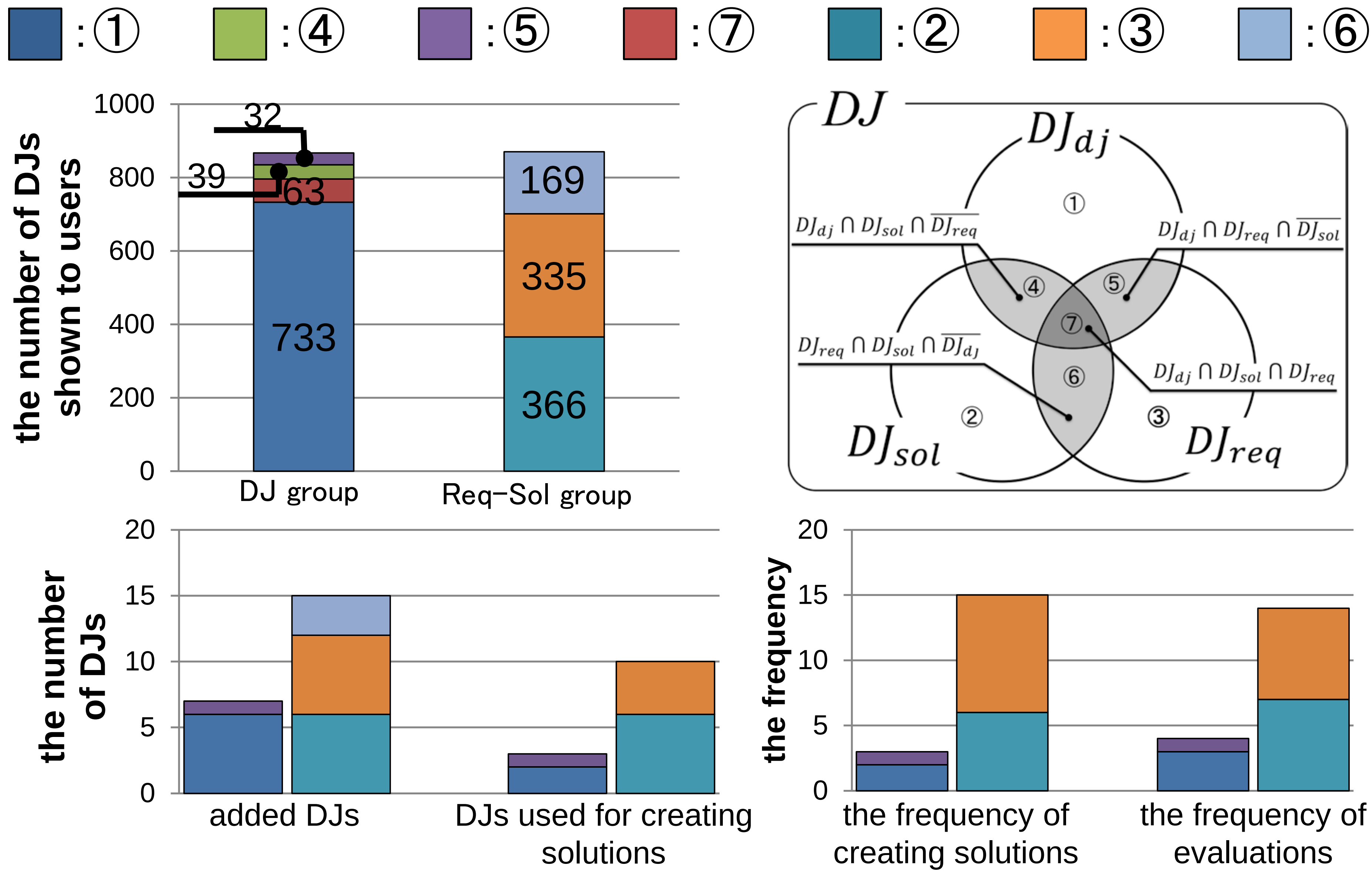
Knowledge graph of data utilization is represented based on a undirected graph ($G_W$)

$$G_w = (V_W, E_W)$$

| | |
|---|---|
| $V_W = \bigcup_{w \in \{req, sol, dj\}} V_w$ <br> $(V_i \cap V_j = \emptyset (i \neq j), V_i \neq \emptyset)$ | a set of nodes |
| $E_W = \left\{\{req_i, sol_j\}, \{sol_k, dj_l\} \middle\vert req, sol, dj \in W\right\}$ | a set of edges |

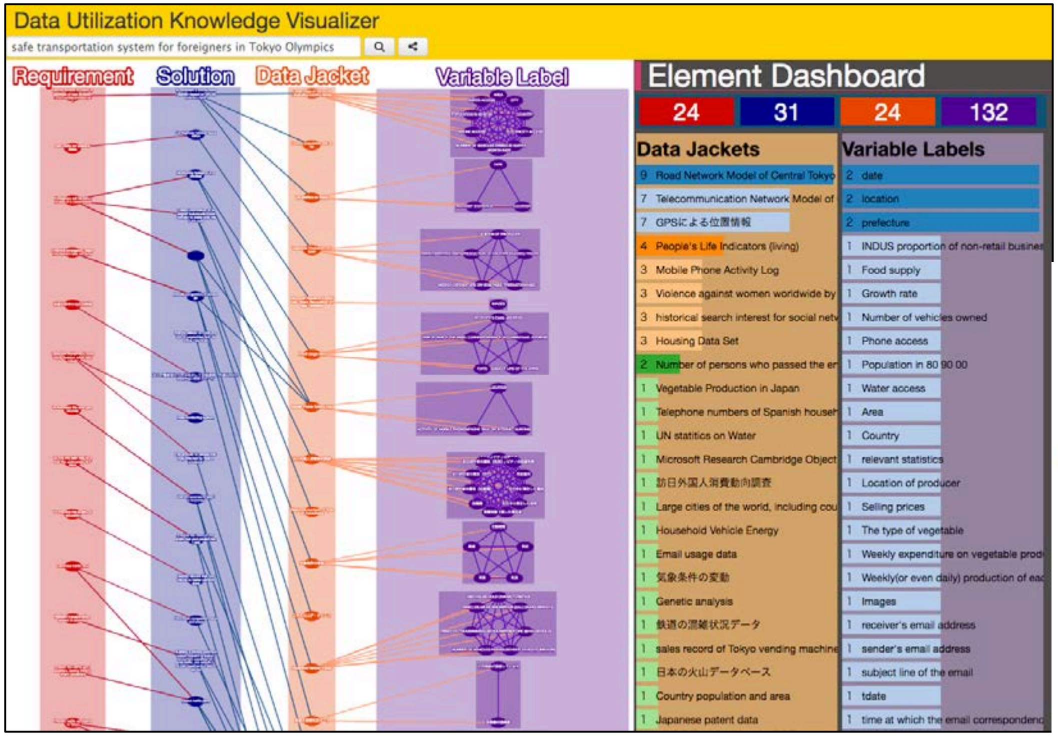
# Knowledge Graph of Data Utilization



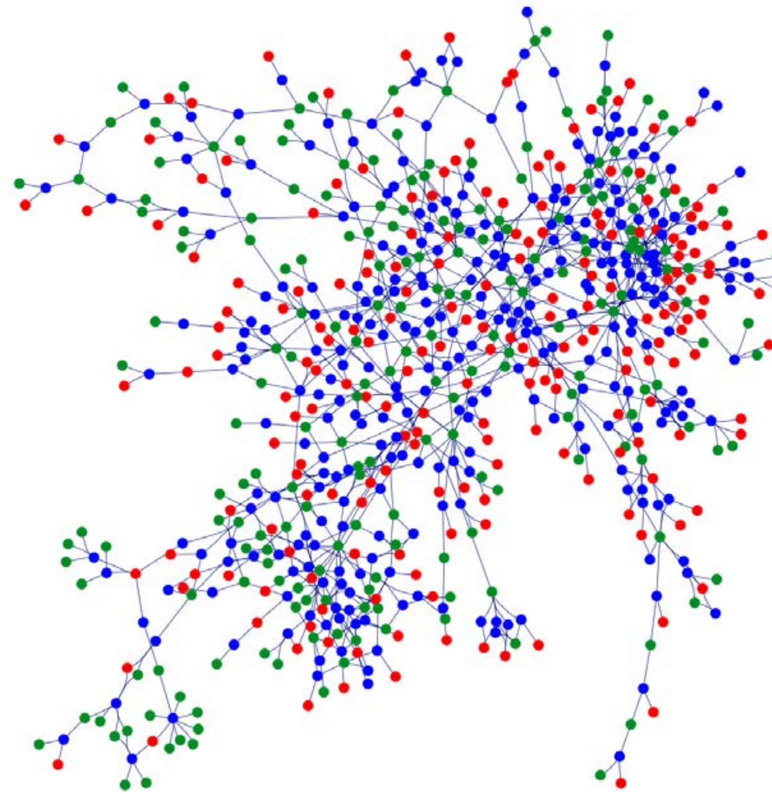



- The result shows that the structured knowledge of data utilization may support the significant discovery of data related to users' interests, even if users do not have sufficient knowledge of the data.

# Improving the Transparency of the Retrieval Process




Data/Knowledge Retrieval System (Data Jacket Store) by showing the retrieval process

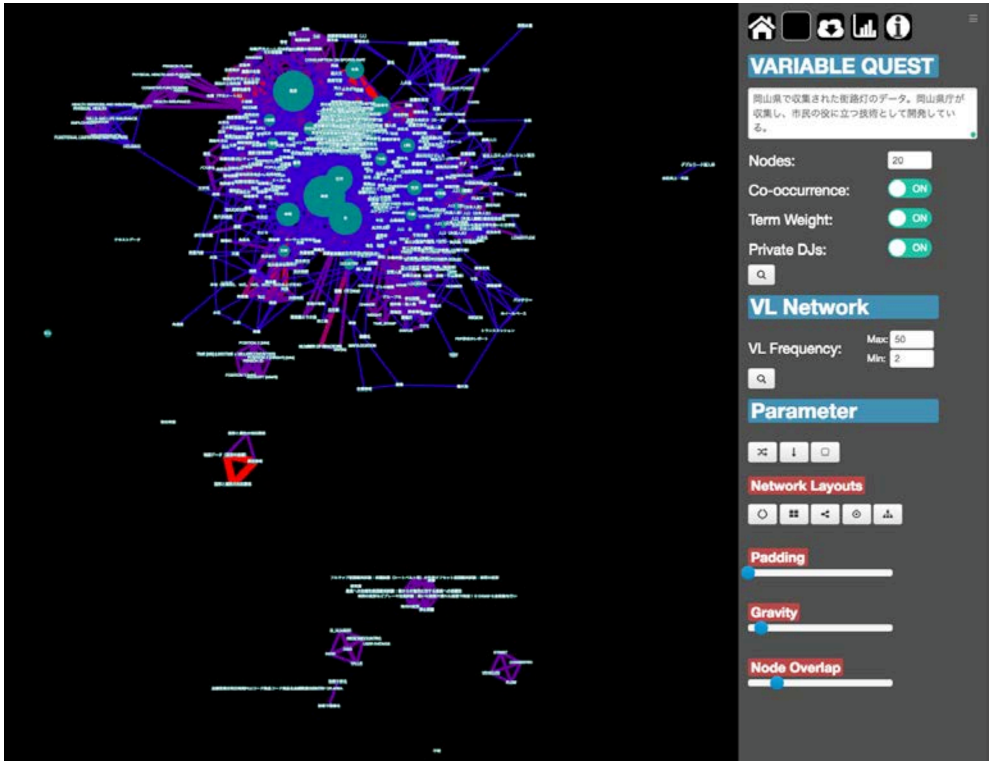


Structural analysis for evaluating knowledge elements in the data driven innovation

T. Hayashi, Y. Ohsawa, "Retrieval System for Data Utilization Knowledge Integrating Stakeholders' Interests," AAAI Spring symposium 2018 Beyond Machine Intelligence: Understanding Cognitive Bias and Humanity for Well-being AI, 2018.

# VARIABLE QUEST (VQ)

- VARIABLE QUEST (VQ) is the network visualization of VLs using the matrix-based inferring method of VLs by unifying co-occurrence graphs [Hayashi & Ohsawa, 2017].





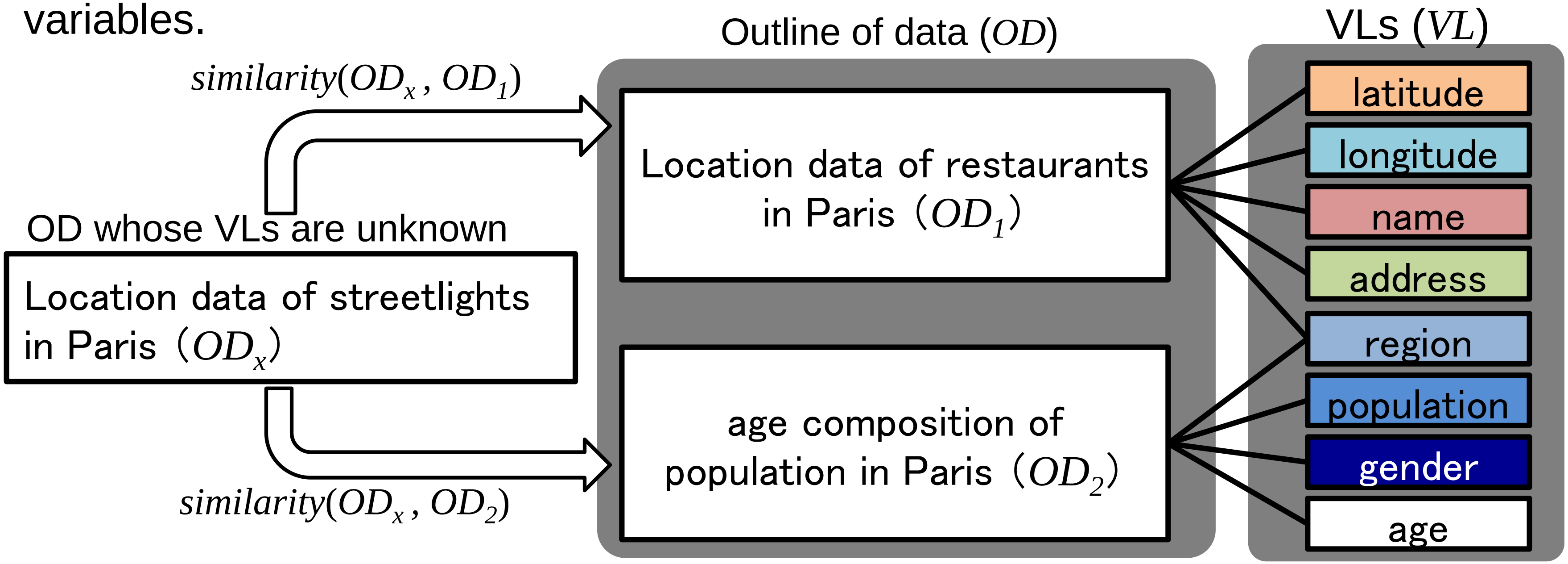
T. Hayashi, Y. Ohsawa, "Matrix-based Method for Inferring Variable Labels Using Outlines of Data in Data Jackets," The Pacific-Asia Conference on Knowledge Discovery and Data Mining 2017 (PAKDD2017), 2017.

T. Hayashi, Y. Ohsawa, "VARIABLE QUEST: Network Visualization of Variable Labels Unifying Co-occurrence Graphs," IEEE-ICDM Workshops 2017, pp.577-583, 2017.

# Inferring VLs

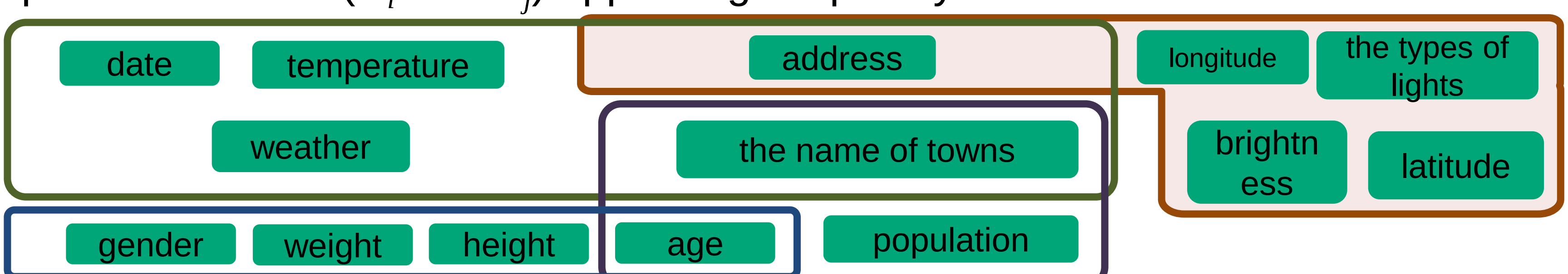
Model 1:

When a pair of datasets is similar each other, the datasets have the similar variables.

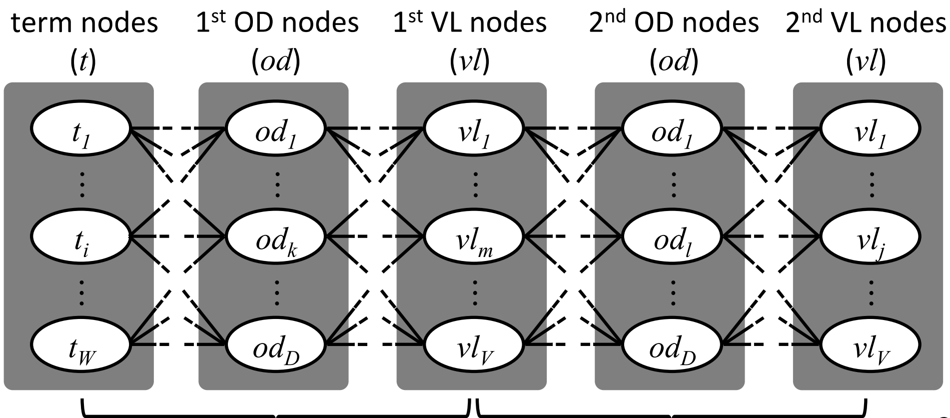



Model 2:

A pair of variables ($vl_i$ and $vl_j$) appearing frequently in the same datasets.



# Term-VL Matrix EC




a Term-VL matrix $E$ ($=MR^{\mathrm{T}}$)

$$
\begin{array}{c|c}
 & \begin{matrix} vl_1 & \cdots & vl_m & \cdots & vl_V \end{matrix} \\
\begin{matrix} t_1 \\ \vdots \\ t_i \\ \vdots \\ t_W \end{matrix} &
\begin{bmatrix} e_{11} & \cdots & e_{1m} & \cdots & e_{1V} \\ \vdots & & \vdots & & \vdots \\ e_{i1} & \cdots & e_{im} & \cdots & e_{iV} \\ \vdots & & \vdots & & \vdots \\ e_{W1} & \cdots & e_{Wm} & \cdots & e_{WV} \end{bmatrix}
\end{array}
$$

a VL co-occurrence matrix $C$ ($=RR^{\mathrm{T}}$)

$$
\begin{array}{c|c}
 & \begin{matrix} vl_1 & \cdots & vl_j & \cdots & vl_V \end{matrix} \\
\begin{matrix} vl_1 \\ \vdots \\ vl_m \\ \vdots \\ vl_V \end{matrix} &
\begin{bmatrix} c_{11} & \cdots & c_{1j} & \cdots & c_{1V} \\ \vdots & & \vdots & & \vdots \\ c_{m1} & \cdots & c_{mj} & \cdots & c_{mV} \\ \vdots & & \vdots & & \vdots \\ c_{V1} & \cdots & c_{Vj} & \cdots & c_{VV} \end{bmatrix}
\end{array}
\Rightarrow
$$

a Term-VL matrix $EC$ ($=MR^{\mathrm{T}}RR^{\mathrm{T}}$)

$$
\begin{array}{c|c}
 & \begin{matrix} vl_1 & \cdots & vl_j & \cdots & vl_V \end{matrix} \\
\begin{matrix} t_1 \\ \vdots \\ t_i \\ \vdots \\ t_W \end{matrix} &
\begin{bmatrix} g_{11} & \cdots & g_{1j} & \cdots & g_{1V} \\ \vdots & & \vdots & & \vdots \\ g_{i1} & \cdots & g_{ij} & \cdots & g_{iV} \\ \vdots & & \vdots & & \vdots \\ g_{W1} & \cdots & g_{Wj} & \cdots & g_{WV} \end{bmatrix}
\end{array}
$$

1. When $OD_x$ is given, a $W$-dimensional feature vector of $OD_x$ ($\boldsymbol{od_x}$) is obtained after the pre-processing.
2. By comparing the similarity of $\boldsymbol{od_x}$ and each $W$-dimensional feature vector of VL ($vl_j$) in the matrix $EC$, a scored set of VLs are obtained.

$$c_{ij} = \sum_{k=1}^{|D|} r_{ik} r_{kj}$$

$$g_{ij} = \sum_{m=1}^{|V|} \left( \sum_{k=1}^{|D|} v_{ik}\, r_{km} \right) \left( \sum_{l=1}^{|D|} r_{ml}\, r_{lj} \right)$$

The Term-VL matrix $EC$ is equivalent to the adjacency matrix of the 5-partite graph, which consists of 5-disjoint sets of nodes.

$g_{ij}$ represents the number of paths from the $i$th term ($t_i$) to the $j$th VL ($vl_j$) in the 2nd VL nodes, by way of the 1st OD nodes, the 1st VL nodes, and the 2nd OD nodes.

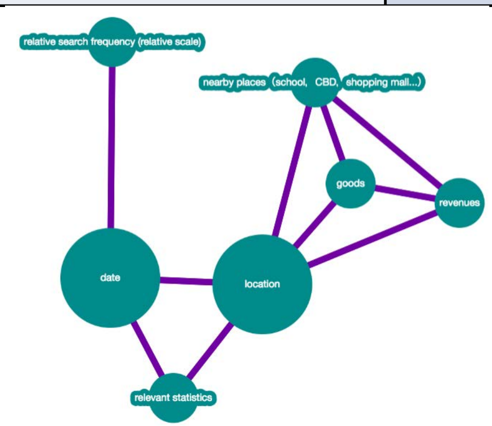
# Implementation

Co-occurrence graph of VLs is represented based on a weighted undirected graph ($G_S$)

$$G_S = (V_S, E_S, f, h)$$

$$f: V_s \to L_{V_s},\ h: E_s \to L_{E_s}$$

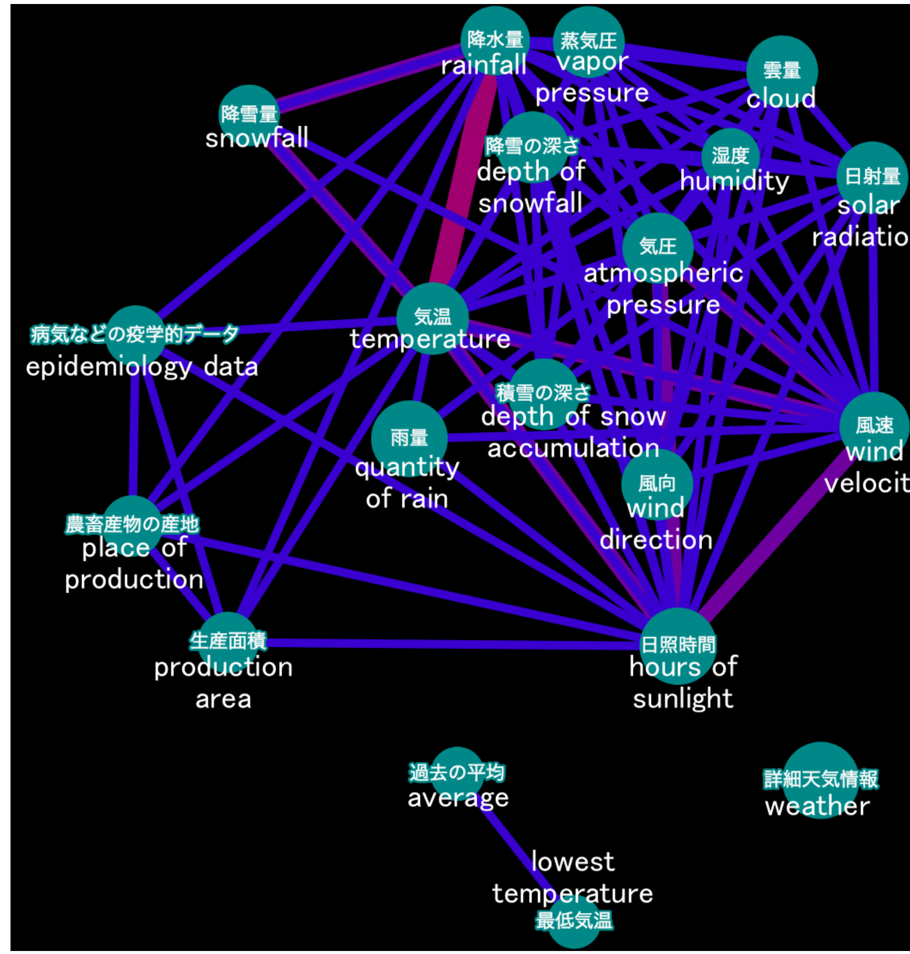
| | |
|---|---|
| $V_S = \{vl_i \in S\}$ | a set of nodes |
| $E_S = \left\{(vl_i, vl_j)_{dj_k} \middle| vl_i, vl_j \in S, vl_i \neq vl_j\right\}$ | a set of edges |
| $L_{V_S} = \{\text{frequency}(vl_i) \mid vl_i \in S\}$ | the frequency of VLs |
| $L_{E_S} = \left\{\text{link}(vl_i, vl_j) \middle| vl_i, vl_j \in S, vl_i \neq vl_j\right\}$ | the frequency of co-occurrences of a pair of VLs |
| $\text{link}(vl_i, vl_j) = \sum_{k=1}^{D} \lvert vl_i \rvert_{dj_k} \lvert vl_j \rvert_{dj_k}$ | |



# Example 1

$OD_x$: Japan weather data provided by Japan Meteorological Agency, which includes information about the temperature and weather of prefectures.

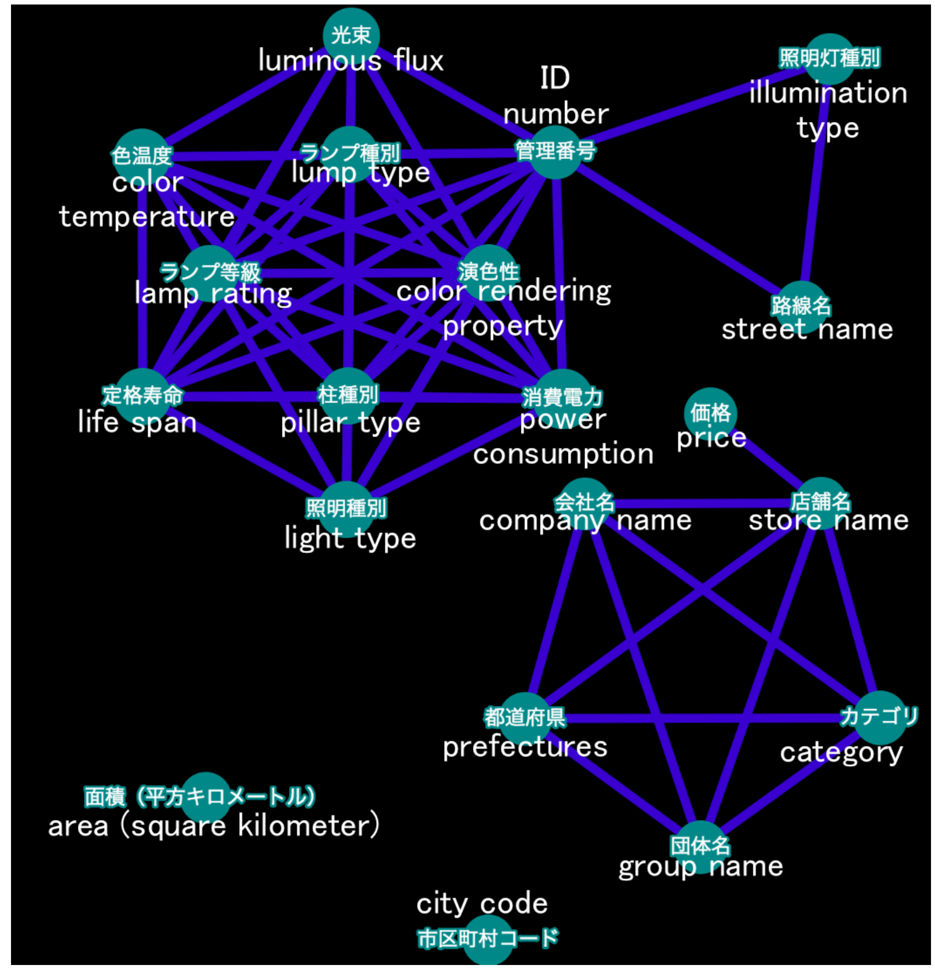
| Term-VL Matrix EC | |
|---|---|
| VL | Similarity |
| weather | 0.318 |
| hours of sunlight | 0.313 |
| rainfall | 0.307 |
| temperature | 0.293 |
| vapor pressure | 0.293 |
| solar radiation | 0.293 |
| depth of snowfall | 0.293 |
| wind velocity | 0.293 |
| weather | 0.318 |
| hours of sunlight | 0.313 |
| ⋮ | ⋮ |



# Example 2

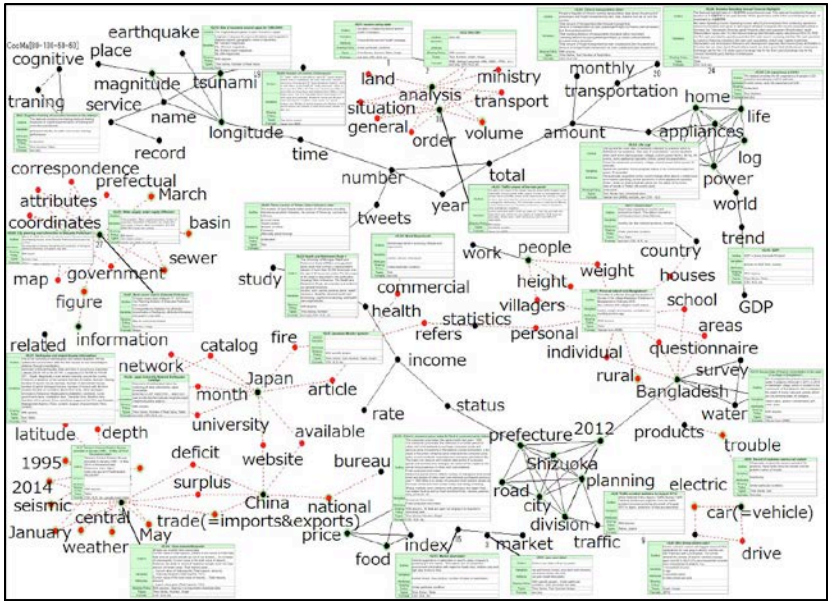
$OD_x$: The information about location and installation of streetlights in Paris.

| Term-VL Matrix EC | |
|---|---|
| VL | Similarity |
| light type | 0.557 |
| life span | 0.557 |
| color temperature | 0.557 |
| lump type | 0.557 |
| color rendering | 0.557 |
| pillar type | 0.557 |
| luminous flux | 0.557 |
| power consumption | 0.533 |
| ID number | 0.505 |
| illumination types | 0.465 |
| ⋮ | ⋮ |



# Discussion for Data Utilization

# Innovators Marketplace on Data Jackets (IMDJ)

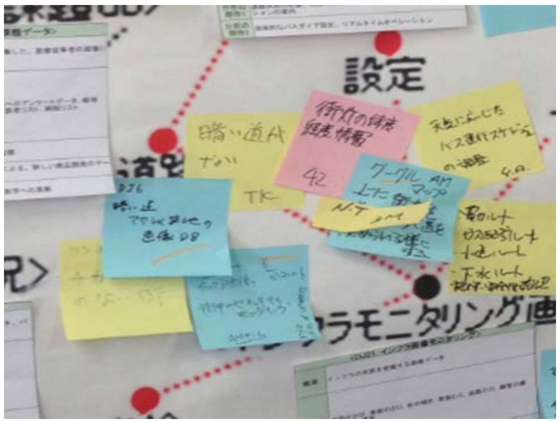


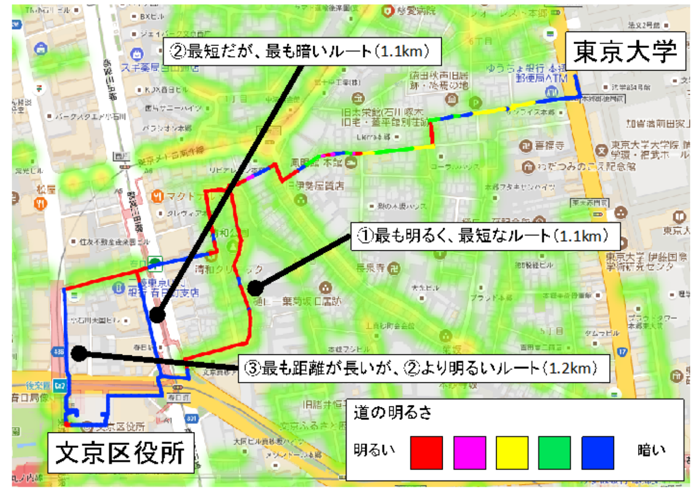

- IMDJ is a gamified workshop for discussing the data utilization.
- Data owners provide their datasets as DJs, data analysts create solutions for solving data users' problems which are stated as requirements, and evaluate them.


Y. Ohsawa, T. Hayashi, and H. Kido, "Restructuring Incomplete Models in Innovators Marketplace on Data Jackets," Springer Handbook of Model-Based Science, L. Magnani, T. Bertolotti (eds), Springer, pp.1015-1031, 2017.

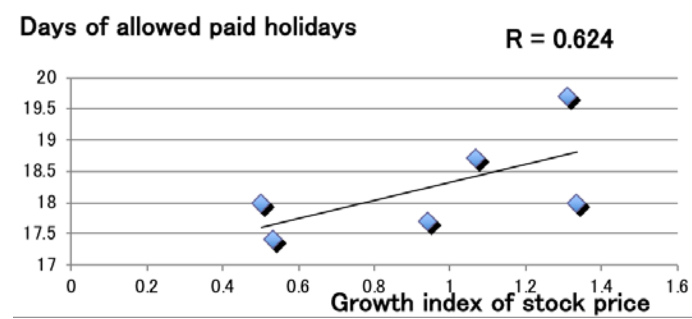
# Examples of Use Cases （1/2）

## Streetlight data & map data



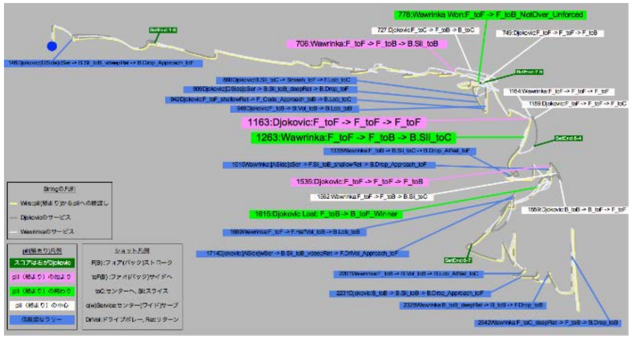




## Paid holidays & stock price data



# Examples of Use Cases（2/2）

Visualizing the sequence of tennis and strategies




Understanding the latent dangerous locations from the history of bike data


Support system for football players and trainers

# Summary

- The potential benefits of reusing and analyzing massive amounts of data have been discussed by various stakeholders from diverse domains.
- However, it is difficult to learn the kinds of data that are related to our interests.
- We introduce our latest technologies for activating cross-disciplinary data exchange and collaboration by structuring the knowledge of data utilization using Data Jacket (DJ).

Thank you for your listening!