On General Principles for Hypothesis Interpretation

Hiroshi Ishikawa
Distinguished Professor
Tokyo Metropolitan University
ishikawa-hiroshi@tmu.ac.jp

TOKYO METROPOLITAN UNIVERSITY
東京都立大学

Short bio

- ***Hiroshi Ishikawa*** is a distinguished leading professor and an emeritus professor of Tokyo Metropolitan University (TMU). He is also the director of TMU Social Big Data Research Center.
- His research interests include database, data mining, and big data. He is also interested in applications of such technologies, including tourism, mobility, smart city engineering, medical, and science.
- He has published actively in international journals and conferences, such as ACM TODS, IEEE TKDE, VLDB, IEEE ICDE, and ACM SIGSPATIAL and MEDES. He has authored a dozen of books, which include books entitled *How to Make Hypotheses* (in Japanese, Kyoritsu Shuppan, 2021) and *Social Big Data Mining* (CRC, 2015).
- He received Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology of Japan in 2021. He also received DBSJ Award for Distinguished Achievement and Contribution from the Database Society of Japan in 2022. He is fellows of IPSJ and IEICE and members of ACM and IEEE.

Necessity to Interpret and Explain Hypothesis

- Why is it necessary to interpret the hypothesis in the first place?
- In a nutshell, this is to get the users (stakeholders) who are involved in the application of data analysis to accept the hypothesis. The parties involved in data analysis applications are as follows:
- *Analysts*
- *Field experts*
- *End users*
- Analysts and field experts need to determine if the entire analytical applications (i.e., big data application systems), in which the hypothesis plays a central role, are purely technically reliable.
- Service providers need to be accountable for such applications to be understood and used by service recipients. In other words, the beneficiary of the service has the right to explanation for the overall application, including the individual decisions of the service (Kaminski 2019).
- For that purpose, it is essential to interpret the hypothesis as a basis.
- For example, in scientific applications, analysts and experts are a team of data engineers and scientists. In social infrastructure applications, analysts are often data engineers, experts are decision makers, and end users are ordinary people who receive services.

Explanation in the Philosophy of Science

Explanation in the Philosophy of Science

- Here, we will first summarize the position of explanation in the philosophy of science. It includes the following types of explanation (Woodward and Lauren 2021).
- ***DN (Deductive Nomological) Model of Explanation***
- The *DN* model of explanation is composed by adding conclusions to the facts as input and the general rules. In this case, the facts and laws are collectively called an explanatory term (*explanans*), and the conclusion is called an explained term (*explanandum*). In the DN model, the explanandum is derived from the explanans by deductive reasoning.
- ***SR (Statistical Relevance) Model of Explanation***
- We define that one attribute is *statistically relevant* to another. That is, for a certain group p, attribute C is statistically relevant to attribute E only when the following condition is satisfied.
- $P(E \mid p.C) \neq P(E \mid p)$
- In the *SR* model of explanation, statistically relevant attributes are considered to be explanatory.

Explanation in the Philosophy of Science (continued)

- ***CM (Causal Mechanical) Model of Explanation***
- The CM model of explanation focuses on the causal process in which the cause produces the result. The causal process is characterized by the ability of the cause to transmit a mark. Furthermore, when two causal processes intersect each other and some change occurs, it is considered that there is a causal interaction between them. In this explanation model, the explanation is to trace the causal process and causal interaction leading to event E.
- ***Unificationist Model of Explanation***
- Phenomena that have been explained by applying different laws until then are now explained by one law. For example, Kepler's laws and Galileo's laws were integrated into Newton's laws. In other words, the fact that the number of inference patterns to be applied has decreased is considered as an explanation. This Unificationist model includes not only the law integration but also the phenomenon integration.
- ***Counterfactual Explanation***
- In order to identify cause-effect relationships, we assume possible worlds where no cause has occurred. By checking any change of the effect, it is judged whether the cause is really the cause. The closer such possible world is to the real world in which the cause occurred, the more valid this kind of explanation will be.

Subjects and types of Explanation

Subjects of Explanation

- In order to interpret the hypothesis, it is necessary to explain the hypothesis. Basically, the structure of explanation consists of the subject of explanation and the action (method) of explanation. The correspondences between subjects and actions of the explanation are not always fixed. That is, the action of explanation may be dependent on the subject of explanation or may be independent. Here, the explanation will be described by focusing on the subject of the explanation.
- ***Subject of Explanation***
- The subjects of the explanation consist of the following basic components.
- How to generate Data (HD)
- How to generate Hypothesis (HH)
- What Features of hypothesis (WF)
- What Reasons for hypothesis (WR)

Subjects of Explanation (continued)

- ***How to generate Data (HD)***
- HD is to prepare the data necessary to make a hypothesis (model). Therefore, the procedure (i.e., algorithm) for that is the subject of explanation.
- Why are we trying to explain here with an algorithm rather than a program in a specific programming language? This is because algorithms are more abstract and easier to understand than programs. Furthermore, here, the process flow with the highest level of abstraction is the subject of explanation.
- Basically, data manipulation consists of the following.
- Operation for data search and data transformation
- Condition for data search
- In many cases, the above can be described at once by SQL language (Celko 2014) and SQL-like languages provided by various frameworks for development such as BigQuery (BigQuery 2022) .
- *The following subjects will be described later in associated methods for explanation.*
- *How to generate Hypothesis (HH)*
- *What Features of hypothesis (WF)*
- *What Reasons for hypothesis (WR)*

Types of Explanation

- Using these, the explanation is classified into the following two categories: Macroscopic and microscopic explanations.
- ***Macroscopic explanation***
- The macroscopic explanation is a general explanation and includes the following subjects.
- HD, HH, WF (, WR)
- Here, HD, HH, and WF are indispensable for the macroscopic explanation, and WR is targeted as needed.
- ***Microscopic explanation***
- The microscopic explanation is a detailed explanation and an individual explanation as compared with the macroscopic explanation. The microscopic explanation includes the following subjects in addition to the individual data C as the input that will produce the result R.
- (HH,) WF, WR
- Here WF and WR are indispensable for the microscopic explanation, and HH is targeted as needed. However, for WF and WR, only the parts directly related to the derivation of the result are targeted, not the whole of them, as much as possible.

Model-Dependent Methods for Explanation

Model-Dependent Methods for Explanation

- The action of explanation, that is, the method related to the presentation of explanation is summarized.
- The basic means for presenting an explanation is description by text. Furthermore, if multiple items are the subjects of explanation, they are presented so as to supplement the text using basic visualization methods such as tables, graphs, and figures.
- How to generate Data (HD)
- ***SQL***: SQL commands are presented as they are. SQL can be applied to different models, so it can be viewed as a model-agnostic method.

Model-Dependent Methods for Explanation (continued)

- How to generate Hypothesis (HH)
- We describe a method of presenting explanations that depends on the method of hypothesis generation.
- ***Inference***: The description of the whole laws and the inference rule (type) is presented with an awareness of abstraction.
- ***Problem-solving***: Regarding the description of the entire problem-solving procedure (algorithm), an outline is presented by pseudo-code with an awareness of abstraction. In particular, operations and conditions are focused on. A box model and a dataflow diagram (graph) are presented as auxiliary description of dependency relationships between variables.
- ***SQL***: The SQL commands are presented. In particular, operations and conditions are focused on.
- ***Regression and sparse regression***: In hypothesis generation that can be reduced to an optimization problem, an outline of the objective function and optimization algorithm are presented by pseudo code.
- ***NMF (Nonnegative Matrix Factorization)***: An outline of the objective function and optimization algorithm is presented by pseudo codes.
- ***Clustering***: An outline of the distance function, objective function, and optimization algorithm is presented by pseudo code.
- ***Decision tree***: An outline of the algorithm and index used for node division is presented by pseudo code.
- ***Random forest***: An outline of the algorithm and index used for node division is presented by pseudo code.
- ***Association rules***: An outline of the algorithm and the parameters such as minimum support and confidence is presented by pseudo codes.
- ***Neural Network***: An outline of the algorithm, loss function, and output function is presented by pseudo code.

Model-Dependent Methods for Explanation (continued)

- What Features of hypothesis (WF)
- ***Inference and problem solving***: The description (pseudo-code) itself regarding the overall picture (that is, purpose) of the inference rules or problem-solving procedures is presented. In particular, the operations and conditions for achieving that purpose are listed. Alternatively, the entire variable dependency is presented.
- ***SQL***: All SQL commands are presented.
- ***Regression and sparse regression*** : In hypothesis generation that can be reduced to an optimization problem, especially in the case of linear regression, the fit of the model is presented by the coefficient of determination, and the importance of variables is presented by the standardized partial regression coefficients. Also, as a result of dimensionality reduction by regularization, selected variables are presented along with the importance. Furthermore, the causal relationship is presented by the path diagram.
- ***NMF***: The analysis results of the feature vectors are presented.
- ***Clustering***: The clustering results and silhouette coefficients as the overall evaluation of clustering are presented. All clusters and the data points belonging to them are presented in a low-dimensional scatter plot. In the case of hierarchical clustering, a dendrogram is presented.
- ***Decision tree***: The whole pictures of the decision tree and rules along the data structures are presented.
- ***Random forest***: The importance of variables calculated by integrating the degree of improvement of the division index (e.g., GiNi, entropy) is presented.
- ***Association rules***: Supports and confidences associated with all rules are presented.
- ***Neural Network*** : The structure of the model is presented in a schematic diagram. The loss function and the output function are presented.

Model-Dependent Methods for Explanation (continued)

- What Reason for hypothesis (WR)
- In addition to the presentation of the individual data *C* corresponding to the result *R*, the following methods for explanation is presented. Regarding the execution of the model, the creation method and features are presented as instantiated as possible.
- ***Inference and problem solving***: For inference and problem solving, the rules and procedures (i.e., operations and conditions) applied individually to derive R are presented using the tracer that interacts with the model execution system.
- ***SQL***: The executed SQL commands and the generated intermediate result are presented by the SQL database system.
- ***Regression and sparse regression***: Regression equations are presented in the same graph (chart) as individual data.
- ***NMF***: The analysis results of individual feature vectors are presented.
- ***Clustering***: The characteristics of the cluster to which the individual data belongs are presented. Such clusters and the data belonging to them are presented in a low-dimensional scatter plot.
- ***Decision tree***: The paths (subtrees) and rules (subset) applied to individual data are presented.
- ***Random forest***: The importance of variables is presented.
- ***Association rules***: The supports and confidences of individually applied association rules are presented.
- ***Neural Network*** : the part of the data that contributes to the individual result is presented by the methods using the value of the gradient such as Grad-CAM. Grad-CAM will be described in detail in a case study described below.

Model-Agnostic Methods for Explanation

Model-Agnostic Methods for Explanation

- ***LIME***
- LIME (Local Interpretable Model-Agnostic explanations) (Ribeiro et al. 2016) evaluates a function in the neighborhood z of the input in order to estimate the behavior of the model (that is, function) *f* at the input x. To explain the behavior of *f*, we approximate *f* with a simpler model (that is, surrogate function) *g*. At the same time, LIME minimizes the complexity of model *g* and improves the interpretability of the model. The input of the surrogate function can be interpreted by humans so that the factors of the result can be explained.
- First, the user gives concretely the related functions as follows in order to determine the model *g*.
- Function that defines the neighborhood: $\pi_x(Z') = \exp(-(x'-z')^2/\sigma^2)$
- Function that defines loss: $L()$
- Function that defines the complexity of the model: $\Omega()$
- Here, the values x' and z' at *g* correspond to the values x and z at *f*, respectively.
- Furthermore, the model *g* is specifically defined as follows.
- $g(z') = \Phi \cdot z'$ (linear regression)
- Here z' is a *coalition* vector. That is, if the value is 1, the corresponding feature exists, and if it is 0, the feature does not exist.
- We add Ω as a regularization term to find *g* that minimizes the following objective function.
- $L(f, g, \pi_x) + \Omega(g)$
- The following is a concrete description of the objective function.
- $\Sigma_{z'}[f(z) - g(z')]^2 \pi_x(z') + \infty \mathbf{1}(\|\Phi\|_0 > K)$
- Here **1** is an indicator function. The argument indicates the condition that the number of non- zeros is greater than *K*.

Model-Agnostic Methods for Explanation (continued)

- ***Kernel SHAP***
- Kernel SHAP (SHapley Additive exPlanations) (Lundberg and Lee 2017) creates an explanation model with a linear model as LIME does.
- Kernel SHAP minimizes the following objective function:
- $L\ () = \Sigma_{z'}\ [f\ (z) - g\ (z')]\ ^2 \pi_x\ (z')$
- Here z' is a coalition vector.
- The *Shapley* value is based on cooperative game theory. As shown below, each feature is regarded as a player, the average marginal contribution of each feature is set as the Shapley value, and the model g is defined by the linear sum of them.
- $g\ (z') = \Phi_0 + \Sigma_{i\ =\ 1,\ M}\ \Phi_i z'_i$
- Here, M is the number of types of features.
- Specifically, the objective function can be expressed using this as follows.
- $L\ () = [f\ (z) - C\Phi]\ ^2 \Pi$
- Here, $N = 2^M$, C is a $(N, M + 1)$ matrix, and is called a coalition matrix. Π is a (N, N) diagonal matrix and has the following diagonal components.
- $\Pi_{i,\ i} = (M - 1)/((M \text{ choose } |c_i|)\ |c_i|\ (M - |c_i|))$
- Please refer to the paper by Lundberg and Lee (Lundberg and Lee 2017) for these derivations.
- However, in the case of Kernel SHAP, there is no corresponding regularization of LIME.

Model-Agnostic Methods for Explanation (continued)

- ***Counterfactual Explanation***
- We try to explain the importance of features by observing how the result of a hypotheses (model) changes or the accuracy calculated based on the results changes when reducing or invalidating certain features.
- In other words, in the presence of event *C* and event *E*, if *C* is changed by some intervention *I* and *E* always changes, it can be explained that the result of *E* is caused by *C*. This method is called Counterfactual Explanation (Woodward 2004) in the context of the philosophy of science.
- However, the Counterfactual Explanation in the context of machine learning has a slightly different meaning. That is, there are many cases where the main purpose is to estimate the minimum features that substantially affect the results and accuracy (White and Garcez 2019).
- The following objective function is minimized for a counterfactual example *c* (Wachter et al. 2018).
- $\text{yloss}\,(f\,(c), y) + \lambda \text{dist}\,(c, x)$
- Here, the first term is the loss between $f\,(c)$ and the desired value y, and the second term is the distance between c and the test instance x.

Reference Architecture for Explanation

Reference Architecture for Explanation

- Reference Architecture for Explanation
- The framework that should be for explanation is summarized below (Ishikawa et al. 2020). The reference architecture for explanation generally consists of the following modules and interacts with the model generation and execution system (see Fig. 1).
- ***Explanation generation***: Generate an overall explanation by combining the subjects of explanation extracted through the explanation extraction.
- ***Visualization***: Visualize the overall explanation.
- ***Explanation extraction***: Extract the necessary subjects of explanation from the model generation (source), model execution (source), tracer (log data), and database (data).
- ***Trace***: Record the parts related to execution using the model generation, model execution, and database.
- Furthermore, the above module uses the following database.
- Model generation (source)
- Model execution (source)
- Trace (log data)
- Database (base)

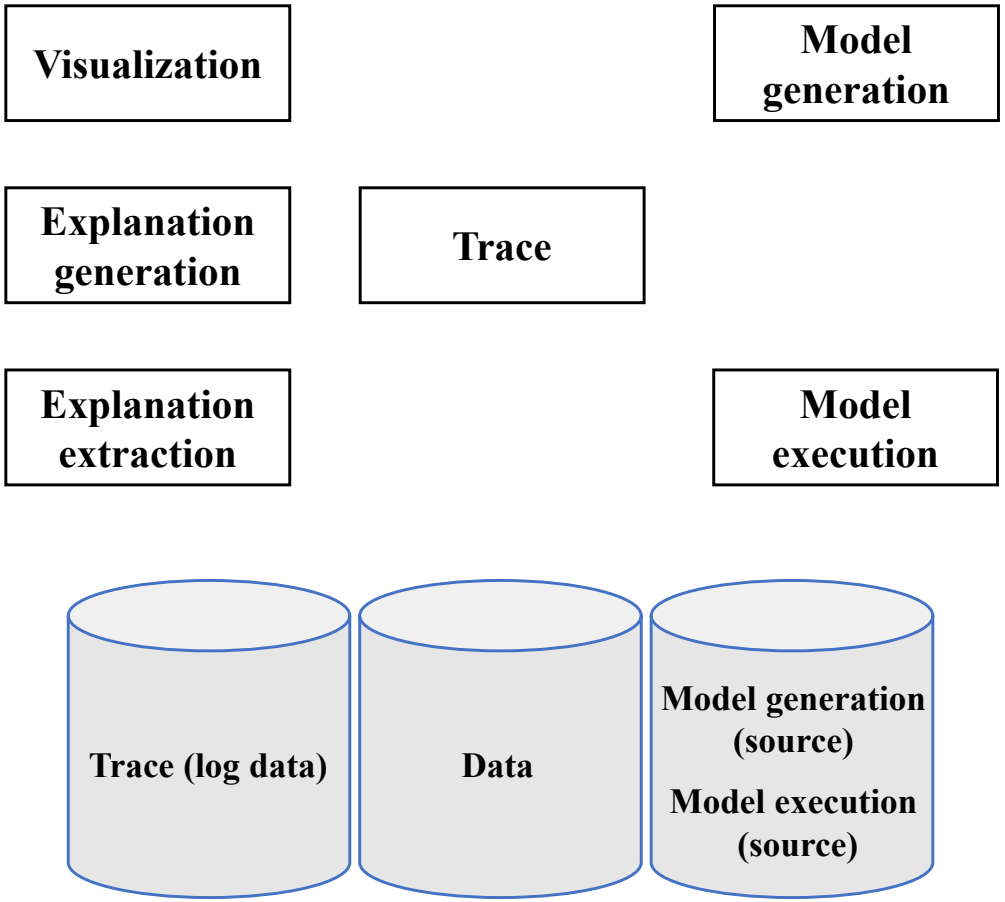


Figure 1. Reference architecture for explanation.

Case studies

Case 1: Discovery of Candidate Installation Site of Free Wi-Fi Access Point

- Case 1 exemplifies both a macroscopic explanation that presents the whole procedure for data and hypothesis generation and a microscopic explanation that presents the result. Let us take this case as an example of hypothesis generation using the difference between hypotheses.

Case 1: Discovery of Candidate Installation Site of Free Wi-Fi Access Point (continued)

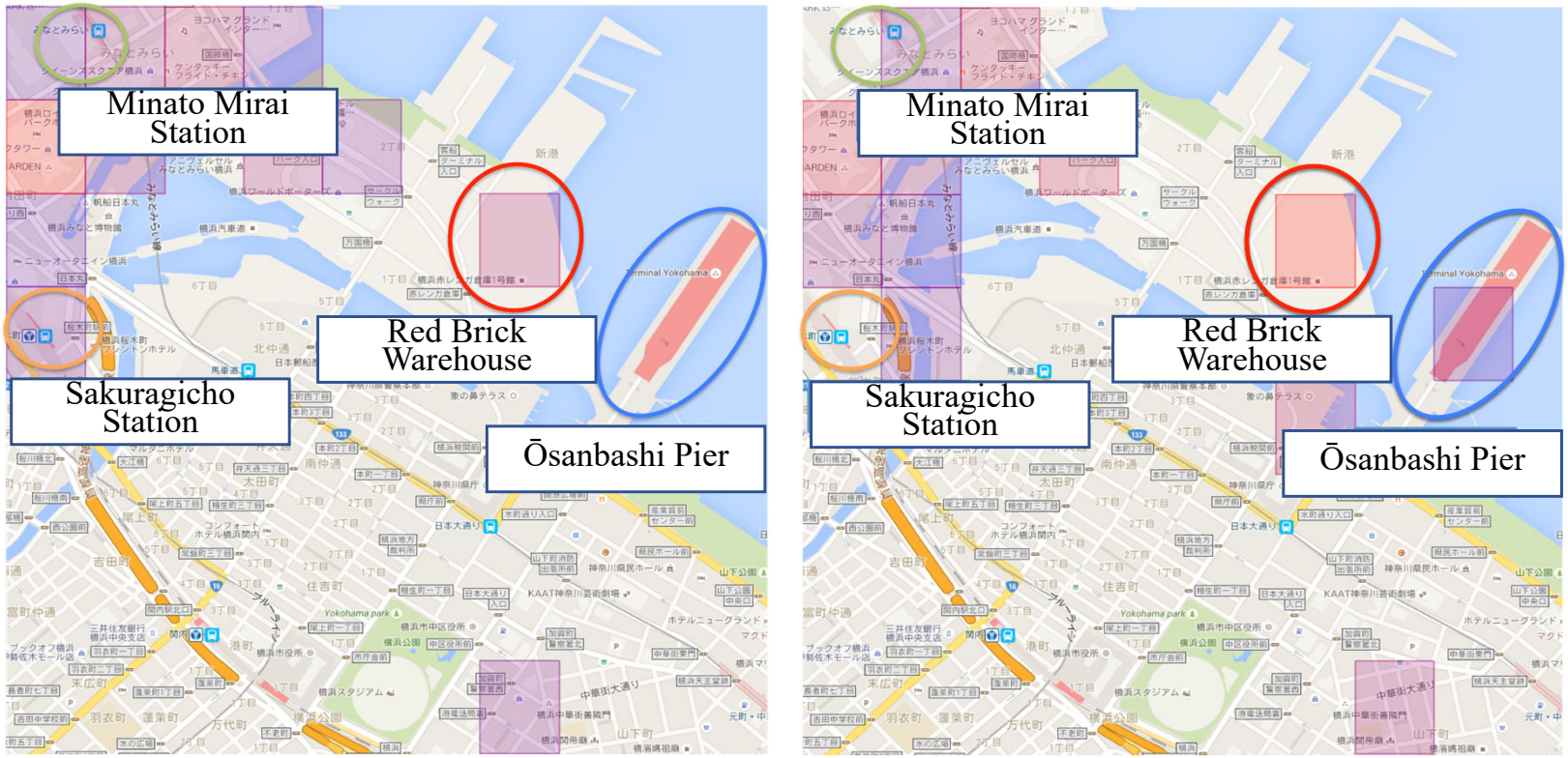
- (Specific purpose of the case in social science) This application is related to the social sciences. Travel agents are an example of parties involved in the EBPM (evidence-based policy making). They are confronted with the gap between the social needs that "a lot of foreign travelers want to use the Internet" and the desirable state that "an infrastructure with free Wi-Fi access spots for foreigners is available." Practically, it is important for them to identify areas with such gaps. Therefore, it is necessary to explain to EBPM decision-makers how to draw conclusions (i.e., detected area with the gaps).
- (Method used in this case) Only data posted by foreign visitors who are judged using a media-specific method (Mitomi et al. 2016) were prepared by selecting social data using SQL. For different results from different data sources, the final result was calculated using SQL as a procedure based on the inter-hypothesis difference method. Furthermore, we visualized the intermediate results that represent the state in the middle.

Case 1: Discovery of Candidate Installation Site of Free Wi-Fi Access Point (continued)

- Explanation of Integrated Hypothesis
- The SQL-based procedure below uses a generalized difference method for generating a hypothesis using various data from Twitter and Flickr as a whole. As a result, we could discover tourist spots that were attractive for foreign tourists but had no accessible free Wi-Fi for them at least at the time of the experiment.
- (HD, HH, WF)
- **insert into ForeignerT select * from T: TweetDB where ForeignVisitorT (*)**
- From the TweetDB table, we create a database of foreign visitors using the function (ForeignVisitorT) as a filter condition, that determines the tweet poster as a foreign visitor based on the length of stay in Japan. We store the intermediate results in the ForeignerT table.
- (HD, HH, WF)
- **insert into ForeignerF select * from F: FlickrDB where ForeignVisitorF (*)**
- From the FlickrDB table, we create a database of foreign visitors using the function (ForeignVisitorF) as a filter condition, that determines foreign visitors based on the place of residence of the photo poster. We store the intermediate results in the ForeignerF table.

Case 1: Discovery of Candidate Installation Site of Free Wi-Fi Access Point (continued)

- (HD, HH, WF)
- **insert into GridT (Index) select ForeignerT.Index from ForeignerT group by ForeignerT.Index having count (*)> = ThT**
- We "group" foreign visitor's tweets ForeignerT based on the geotag (corresponding to "by" Index) attached to them. Furthermore, we obtain the index of the grid with the number of tweet posts above a certain number (ThT). In other words, this query retrieves free Wi-Fi access spots for foreign visitors. We store the intermediate result in GridT. Furthermore, the GridT is visualized by the 2D map (see Fig. 2 (a)).
- (HD, HH, WF)
- **insert into GridF (Index) select ForeignerF.Index from ForeignerF group by ForeignerF.Index count (*)> = ThF**
- We "group" foreign visitor's photos ForeignerF based on the geotags (corresponding to "by" Index) attached to them. Furthermore, we obtain the index of the grid with the number of photo posts above a certain number (ThF). In other words, this query retrieves spots visited by many foreign tourists. We store the intermediate results in GridF. Furthermore, the GridF is visualized by 2D map (see Fig. 2 (b)).



(a) (b)

Figure 2. (a) Visualization of Free WiFi spots based on posted tweets. (b) Visualization of tourist spots based on posted flickr photos.

Case 1: Discovery of Candidate Installation Site of Free Wi-Fi Access Point (continued)

- (HH, WF)
- **select * from GridF minus select * from GridT**
- This query obtains a set difference (set minus) between the spots visited by many foreign tourists represented by GridF and the spots providing accessible free Wi-Fi represented by GridT. Thus, we detect spots that are popular for foreign visitors but do not have accessible free Wi-Fi for them.
- (WR)
- For each of GridF and GridT, we present data for the results of interest (for example, Index of Osanbashi Pier). This tells us that GridF contains the Index, but GridT does not.
- At the same time, the following SQL command used to generate individual results is presented.
- **select * from GridF minus select * from GridT**

Case 2: Central Peak Crater

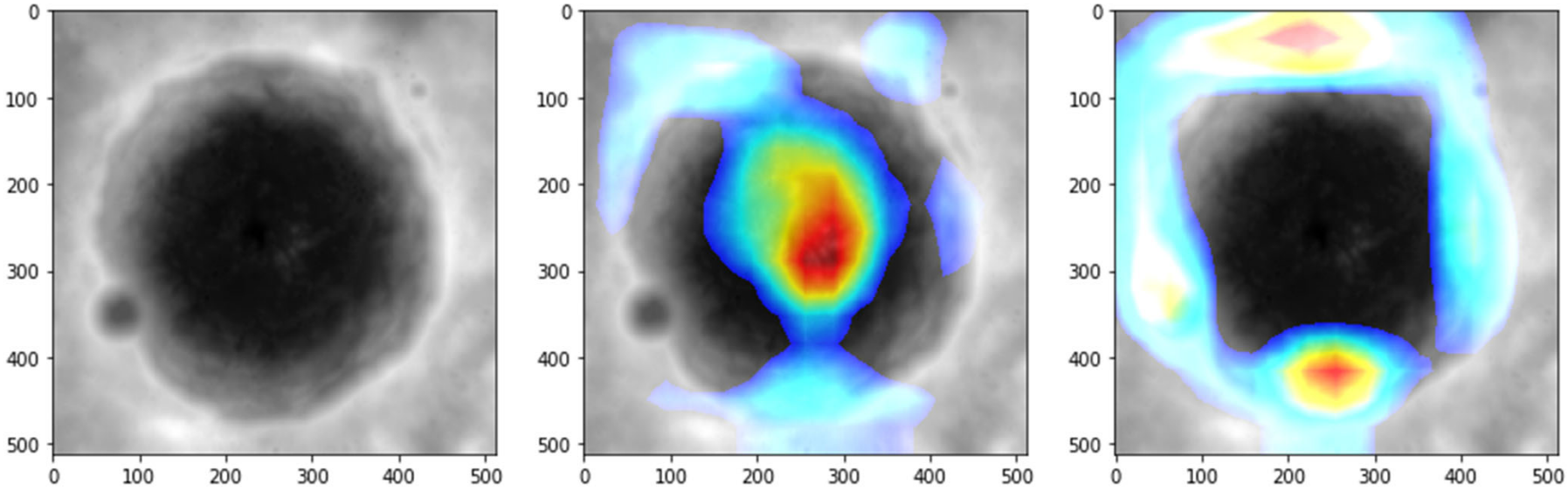
- Case 2 describes a microscopic explanation of the basis of individual judgment of model classification. Here, the example of the central peak crater is used to describe an example of explanation of individual judgment grounds.

Case 2: Central Peak Crater (continued)

- (Specific purpose of the case in natural science) This case is also related to lunar and planetary science. In order to understand the internal structure and activity of the moon, it is considered to use the materials inside the moon as a clue. The central peak in the crater formed by impacts of meteorites is attracting attention as a place where materials below the lunar surface are exposed on the lunar surface. However, not all craters with a central peak on the moon (hereinafter referred to as the central peak crater) have been identified. Therefore, it is scientifically meaningful to create a catalog of central peak craters. However, since identification of central peak craters has been manually done by experts, it has taken a lot of efforts and time. So automatic identification has become the focus of scientists. Therefore, it is also necessary to explain to the relevant scientists the reason why the craters included in the candidates discovered by machine learning are judged to be the central peak craters.
- (Method used in use case) RPSD method (Yamamoto et al. 2017) was used to detect craters only to prepare training data for CNN. Next, the trained CNN was applied to find central peak craters, including unknown and known ones. Furthermore, Grad-CAM (Selvaraju et al. 2020) was used to examine the evidence for determining central peak craters.

Case 2: Central Peak Crater (continued)

- Explanation of Results
- (WR)
- To confirm the reason for the classification result, we visualize the contribution area (that is, individual evidence) of the input image that affects the model's decision, that is, output label. For that purpose, we use Grad-CAM (Selvaraju et al. 2020), which is one of the methods for visualizing the contribution area of each label in the input image. The left part (see Fig. 3) is the input image, the central part is the contribution area of the "Central Peak Crater" label, and the right part is the contribution area of the "Normal Crater" label. The contribution area of the "Central Peak Crater" has an area of high intensity inside the crater, and the emphasized area covers the central peak. On the other hand, there is no emphasized area for the central peak in the contribution area of the "Normal Crater." Therefore, the area occupied by the central peak is considered to contribute to the classification of the "Central Peak Crater" label.



(a) Original image

(b) Contribution area for “central-peak crater”

(c) Contribution area for “normal crater”

Figure 3. Contribution areas for “central-peak crater” and “normal crater” by using Grad-CAM.

Summary

- Necessity to Interpret and Explain Hypothesis
- Explanation in the Philosophy of Science
- Subjects and types of Explanation
- Model-Dependent Methods for Explanation
- Model-Agnostic Methods for Explanation
- Reference Architecture for Explanation
- Case studies