[Slide 1] Good afternoon, everyone! We are CS520 Group 23, presenting on 'Computing Rule-Based Explanations by Leveraging Counterfactuals.'

[Slide 2] This paper relates to 'Data Provenance' and introduces an advanced explanation system for complex machine learning models. Our presentation follows the structure of the paper.

[Slide 3] Before starting, let us talk about the concept 'Data Provenance’. Through the class, we learned that Data Provenance is a metadata that describes the data origin and creation process. However, here we explain Data Provenance as more straightforward through a slightly modified example from the example included in the paper we reviewed.

[Slide 4] Let us assume that a customer named Jesse applies for a loan and the bank uses an automated screening system. However, unfortunately, Jesse is deemed ineligible.

[Slide 5] Here, imagine you are a bank clerk. You can either reply that all of Jesse's features were rejected because they belong to the rejected group, or you can advise that he was rejected because he lacked certain features and that those features need to be updated.

[Slide 6] However, in any case, the bank's explanation system must be reasonable and fair, regardless of its complexity. Note that the bank's customers, including Jesse, are only interested in being evaluated for reasonable and fair reasons and are not interested in system detail. Therefore, we need to at least clarify where the basis and source of the data are.

[Slide 7] Let us recall once again the definition of data provenance mentioned earlier. Connecting it with the previous example, we can conclude that data provenance requires a reasonable algorithm for a system such as loan screening that encompasses provenance and application logic. With this goal in mind, the authors of the articles we reviewed introduced algorithms for more efficient and persuasive explanation systems.

[Slide 8] Now, let us talk about some terminology involved. We will briefly explain the key terms 'rule-based explanation', 'counterfactual explanation' and 'duality’.

[Slide 9] Before we begin the explanation, let us recall the example we looked at in the last step. We have seen that a bank can have two types of explanation system.

[Slide 10] In the first case, the system extracts the most common features (i.e., minimum number of) of features of customers who have already been rejected for a loan and checks whether the loan applicant meets the conditions. Although it does not confirm how customers are approved for a loan, the reasons for being rejected are clear and logical. This system judges the value of new data based on a set of the most common rules. This system can be said to be a rule-based explanation system.

[Slide 11] According to the rule-based explanation system, when the three properties of Relevance, Global Consistency, and Interpretability are satisfied for a specific instance, it becomes a system that can explain that instance. Relevance means that the instance must be relevant to the rule, which means that the rule must already contain the same value as the instance. Global consistency means that all instances corresponding to the rule must be consistent. For example, in a loan review system, this means that all instances included in the rule must have been rejected for loans. Finally, Interpretability states that rules should be the most common and simple and, therefore, have a small cardinality. In other words, the redundant rule should be excluded as much as possible.

One more concept that should be mentioned is the consistency of the data. In general, because it is realistically difficult to check all features, the scope of consistency checking in the paper is limited to the database, and this property is called data consistency.

[Slide 12] On the other hand, the counterfactual explanation system is an explanation system of a more purposeful nature. Like the second case in the previous example, through counterexamples, i.e., cases where the screening was passed, you can specify the features that a specific customer needs to change to pass the screening next time. The main purpose of the counterfactual explanation system is to identify counterfactual instances by checking whether the judgment result changes when the value of the instance included in the existing rule is changed. Therefore, a customer may be judged as unable to pass the loan screening unless he or she has an exception value for that feature, that is, a counterfactual instance.

[Slide 13] To be the counterfactual instance, it is necessary to satisfy two properties. First, the instance must be feasible and plausible with respect to the original instance value. Feasibility imposes constraints on the new values, while plausibility imposes constraints on how the new values in the counterfactual instance differ from the target instance. The authors propose to express these two features as PLAF predicates, which are composed of a conjunction of predicates for each feature of an instance. Second, using a function that calculates the difference from the old value to the new value of each feature with respect to the target instance, PLAF constraints are scored and ranked by their distance from the target instance.

[Slide 14] As seen earlier, the two explanation systems have their own pros and cons. The authors offer a new approach that integrates the two systems, but a rationale is needed to make it possible. In the paper, the author introduced the "Duality and Duality Theorem" and proved the validity of the algorithm to be presented in the future by proving that the two systems have "Duality" that complements each other. So what does ‘Duality’ mean? In a mathematical sense, the term 'dual' is close to meaning a symmetrical relationship. In this regard, the author devoted part of the space to explaining the concepts of 'duality' and 'duality Theorem' to prove that rule-based explanations and counterfactual explanations are, in fact, symmetrical and combinable systems.

[Slide 15] In the paper, the duality of the two systems is proved by proving the related lemmas and the Theorem. However, we can check duality in a more intuitive way.

[Slide 16] Through the data consistency mentioned by the author and the properties that a counterfactual instance must have, we were able to confirm that the corresponding instances consist only of common features. This means that despite the differences between some features, the two corresponding instances are symmetric, and each feature forms a one-to-one relationship. Since we know that a one-to-one relationship has an inverse function, we can easily verify that two instances, or two systems, are a dual relationship. Assume that you are matching different data schemas. Schemas that have the same characteristics in a one-to-one relationship are virtually identical and can be combined in any direction to produce equivalent results. In conclusion, the rule-based explanation system and counterfactual explanation system are dual relationships, so they can be used in combination.

[Slide 17] Following the discussion, we introduce three algorithms, GeneticRule, GeneticCF, and GreedyCF. GeneticRule is the base, and the others extend it with counterfactual explanations.

[Slide 18] Due to time constraints, we recommend that you refer to our report for a detailed explanation, and only a brief explanation of each algorithm will be provided here.

First, GeneticRule is the base algorithm of the other two algorithms and is responsible for finding rule combinations from a given data set. Repeat crossover, mutation, and sorting by fitness score until you find K candidates that are consistent on both the D and s samples of the data set from the more general INST space.

The other two algorithms are basically extension algorithms to genetic rules and show an advanced form that includes new rules in existing rules by introducing a counterfactual explanation system through the CFRules function and the consistentCF function. The difference lies in the cardinality of the candidate rules. GreedyRuleCF differs from GeneticRuleCF in that it tries to achieve minimum cardinality by including only the optimal fit among candidates.

Here, a selectFittest function is used to classify the candidates, which is used in each algorithm to classify the candidates according to the degree of consistency and leave the optimal combination. However, in the case of the GreedyCF algorithm, it does not use this method because it uses a method of sorting according to cardinality.

[Slide 19] In the paper, the authors mentioned the details and results of the experiments they conducted to verify the effectiveness of the algorithm they developed. As in the case of algorithms, it is difficult to describe the specific details of individual experiments here, so only the results of the experiments will be described.

[Slide 20] The four data sets shown in the table above were used to evaluate the algorithm. As you go from left to right, you can see that the data set becomes more complex and atypical. Note that the counterfactual explanation model applied to the newly introduced algorithm was considered a black box and was reflected in the form of borrowing the GeCo model after evaluating 13 existing models. Evaluation of the model can be found at the GitHub address above. The CFRules and consistentCF included in the algorithm were not implemented by themselves but were applied using GeCo. Furthermore, the authors strengthened the effectiveness of the verification by including a comparison with existing systems Anchor and MinSetCover in the evaluation.

[Slide 21] The first evaluation carried out was of quality in terms of consistency and interpretability. In other words, the evaluation concerns the effectiveness of the algorithm. The authors evaluated three algorithms and two benchmark algorithms on the synthetic classifier version, which refers to the rule itself, and then performed the same evaluation on the real classifier version. When tested against synthetic classifiers, the new algorithms all outperformed the benchmarks, and when tested against real classifiers, the algorithms utilized by the counterfactual explanation system mainly outperformed other algorithms. Regarding the test results of the real classifier, the authors analyzed that the introduction of the counterfactual explanation system influenced improved consistency and effective redundancy control.

[Slide 22] The second evaluation performed was in terms of run-time comparison, that is, efficiency. When tested against synthetic classifiers, all new algorithms outperformed the benchmarks, but when tested against real classifiers, algorithms using counterfactual explanation systems showed a significant increase in run-time. The authors analyzed that the strong verification function provided by the counterfactual explanation system was the main cause of significantly increased run time with the complexity of the data.

[Slide 23] Finally, the authors included an attempt to analyze the test results evaluated previously in more detail using microbenchmarks. The result shows that as data sets become more complex, the performance of the underlying counterfactual explanation system deteriorates significantly. The author points out that this happens because as data sets become more complex, the system places a stronger emphasis on finding consistent candidates.

[Slide 24] In addition to previous evaluations, the authors acknowledged the limitations of the results and mentioned future improvement plans. In addition to the parts described in this paper, we would like to add some opinions we have discovered regarding the feasibility and effectiveness of the algorithm.

[Slide 25] The authors presented the limitations of the algorithm they developed, and the improvements needed in the future, organized into the five above. However, I think these five elements can be largely summarized in the 'generality' or 'versatility' aspect. Constraints represented by data consistency are useful for comparison with existing systems but limit the verification of whether the algorithm created is an effective explanation system. Furthermore, the data are not binary and are mutable. Above all, when data consistency is not assumed, the premise for the existing correlation proven through the duality Theorem becomes unstable. Furthermore, the need for a more developed counterfactual explanation model is that the algorithm introduced relies heavily on the black-box counterfactual explanation model, so it will be difficult to evaluate its value until the counterfactual explanation model, the corresponding algorithm, is established. In summary, the evaluation of the paper's algorithm can only be properly performed if the same performance and effect can be expected even when using a more realistic dataset, a more effective counterfactual explanation model must be introduced, and the existing premise of duality must be maintained.

[Slide 26] So far, in this paper, we have looked at the algorithm introduced by rule-based explanations by leveraging counterfactuals as a new explanation model for complex machine learning suitable for high risk. Our conclusion is that, although the algorithm has many limitations, it is meaningful to see that a more consistent and efficient approach to complex data automation can exist. Despite the ever-accelerating black-boxization of machine learning models, if a valid system can be maintained and developed while maintaining clear data provenance, the development of a better automated learning system will not be far away.

[Slide 27] Thank you for your attention. We hope you found this overview useful.