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

[Slide 2] This paper is related 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 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 Jesse is rejected because all of Jesse's features belong to the rejected group, or you can advise that she was rejected because she 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 details. 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 this with the previous example, we can conclude that the provenance of the data requires a reasonable algorithm for a system, such as loan screening, that encompasses the provenance and the application logic. With this goal in mind, the authors of the paper 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 of customers who have already been rejected for a loan and checks whether the loan applicant meets the conditions. Although it does not identify how customers can be 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. 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. 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] Counterfactual explanation systems, on the other hand, are more purposive explanatory systems. In the second case, the banking system was able to specify the counterexamples related to customer’s features that led to the rejection. As such, the main purpose of a counterfactual explanation system is to identify counterfactual cases by checking whether the judgment outcome changes when the values of the cases included in the existing rules change.

[Slide 13] To be a 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, the PLAF constraints are scored and ranked by their distance from the target instance.

[Slide 14] The authors propose a new approach that integrates the two systems mentioned so far but requires a theoretical foundation to make it possible. In the paper, the author introduced the "duality and duality theorem" and proved the validity of the proposed algorithm by proving that the two systems have "duality" that complements each other. So, what does ‘duality’ mean? In a mathematical sense, the term "double" comes close to meaning a symmetrical relationship. Therefore, the author devotes 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, symmetric 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] By the consistency of the data 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. 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] First, GeneticRule is the base algorithm of the other two algorithms and is responsible for finding rule combinations from a given data set. The other two algorithms are basically extension algorithms to genetic rules but utilizing counterfactual explanation system through the CFRules function and the consistentCF function. However, GreedyRuleCF differs from GeneticRuleCF in that it tries to achieve minimum cardinality by including only the optimal fit among candidates.

[Slide 19] In the paper, the authors mentioned the details and results of the experiments they carried out to verify the effectiveness of the algorithm they developed.

[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. As benchmarking systems, the existing Anchor and MinSetCover systems were introduced.

[Slide 21] The first evaluation performed 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 in the synthetic classifier version, which refers to the rule itself, and then performed the same evaluation in the real classifier version. When tested against synthetic classifiers, all new algorithms outperformed 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 running time comparison, that is, efficiency. When tested against synthetic classifiers, all new algorithms outperformed 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 the 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, these limitations are ultimately based on the lack of generality of the data set and the classifier. Due to this constraint, the algorithm may not be practical in practice and the validity of the algorithm may weaken. Furthermore, because the counterfactual explanation model relies heavily on the black box, the newly developed algorithm is ultimately incomplete.

[Slide 26] So far, in this paper, we have looked at the algorithm introduced by rule-based explanations that use 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 there can exist a more consistent and efficient approach to complex data-based decision automation.

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