GBDT4CTRVis: Visual Analytics of Gradient Boosting Decision Tree for Advertisement Click-Through Rate Prediction

Wenwen Gao, Shangsong Liu, Yi Zhou, Fengjie Wang, Feng Zhou, Min Zhu*

Sichuan University

Speaker：Yi Zhou

时间：2023年7月23日 10:30-12:00
地址：融汇丽笙酒店一楼天空厅

ChinaVis 2023

CONTENTS

BACKGROUND

What is advertisement click-through rate (CTR) prediction?

Predicting whether Users will click on Ads displayed on certain digital Media

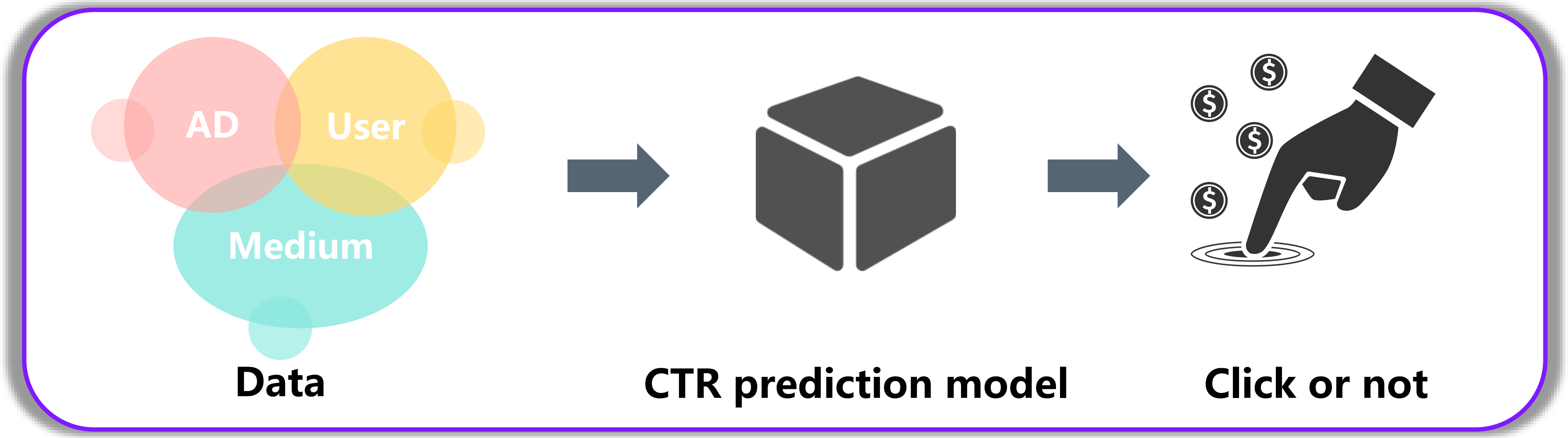
Exposing ads more accurately to target users in order to reduce advertising costs and increase ads revenue.

BACKGROUND

How to predict advertisement CTR?

Advertising analysts build CTR prediction models

Gradient Boosting Decision Tree (GBDT) is a widely used



Fail to **understand the impact of different features** and **analyze the decision-making and the iterative evolution process**

Difficulty in model tuning

BACKGROUND

Existing studies have shown that **interactive visualization** can provide **interpretability** to models and help overcome challenges in model development

- Existing studies do not analyze the **correlations between features** and do not support the distinction of different **categories of features** (ads, media, and users)
- Existing studies are difficult to explore the **iterative evolution process** of a large number of decision trees
- Existing studies do not support the global analysis from the **instance, feature, and model levels**.

ChinaVis 2023

CONTENTS

SYSTEM OVERVIEW

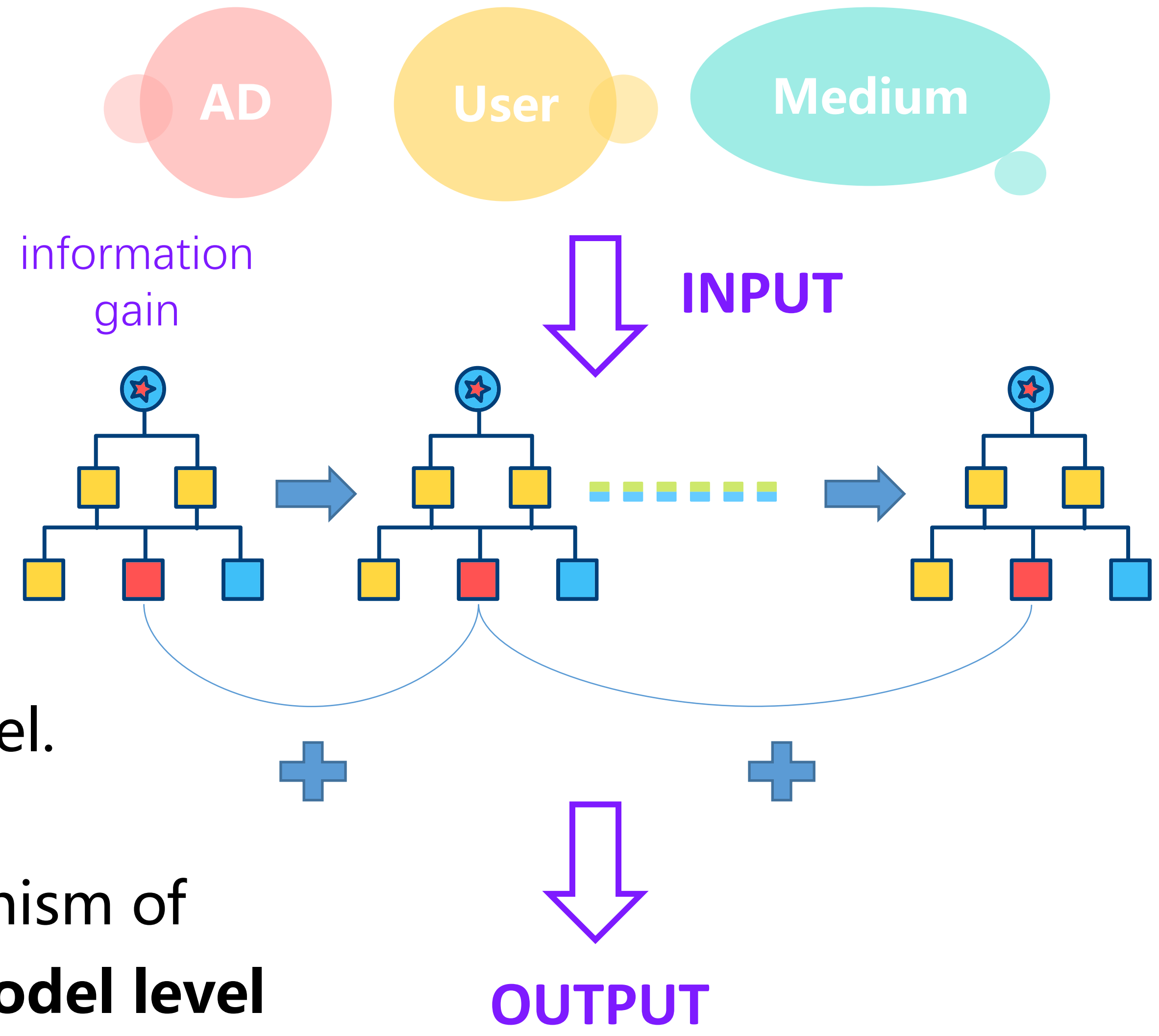
What do we need to do? How do we do it?

Requirements Analysis

R1: Explore the model's prediction results at the instance level.

R2: Analyzing model decision-making basis at the feature level.

R3: Understanding the model's decision-making mechanism at the model level.



GBDT4CTRVis helps advertising analysts understand the working mechanism of the GBDT-based CTR prediction model through **the instance, feature and model level** and facilitate the tuning process.

SYSTEM OVERVIEW

Pipeline of **GBDT4CTRVis**.

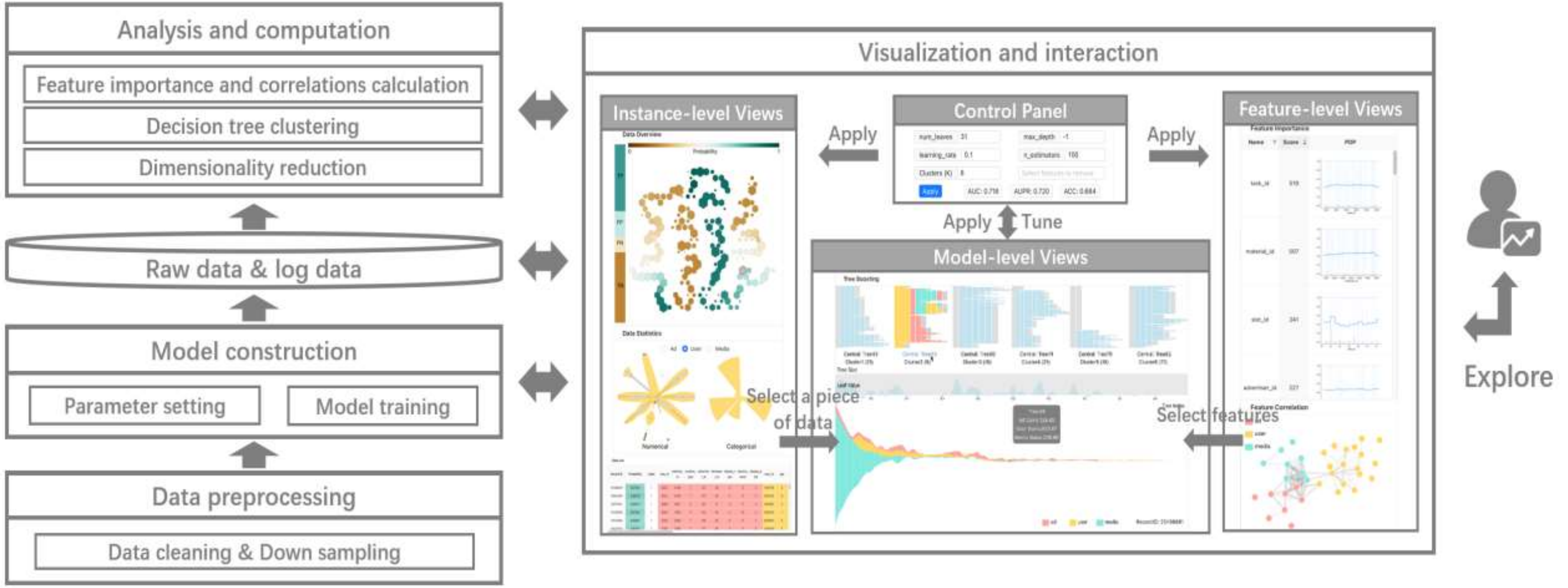
Consists of four main modules:

① Data preprocessing

② Model construction

③ Analysis and computation

④ Visualization and interaction



SYSTEM OVERVIEW

① Data preprocessing

- The advertising CTR prediction dataset is publicly available from the Huawei 2020 DIGIX algorithm competition
- Each record has 36 fields, one is the label for advertising click behavior (0 or 1), and the remaining 35 fields can be divided into three categories of features: **ad, medium, and user**
- We first perform data cleaning and **downsampling** of the dataset

① Model construction

- Implement the GBDT model by **LightGBM**
- There are four main hyperparameters: 1. Maximum number of leaf nodes of the decision tree (**num_leaves),** 2. Maximum depth of the decision tree (**max_depth**), 3.Number of decision tree (**n_estimators**), 4.Learning rate (**learning_rate**)

ChinaVis 2023

CONTENTS

VISUALIZATION

A: Control Panel

B, C, D: Instance-level Views

G, H: Feature-level Views

E, F: Model-level Views

VISUALIZATION

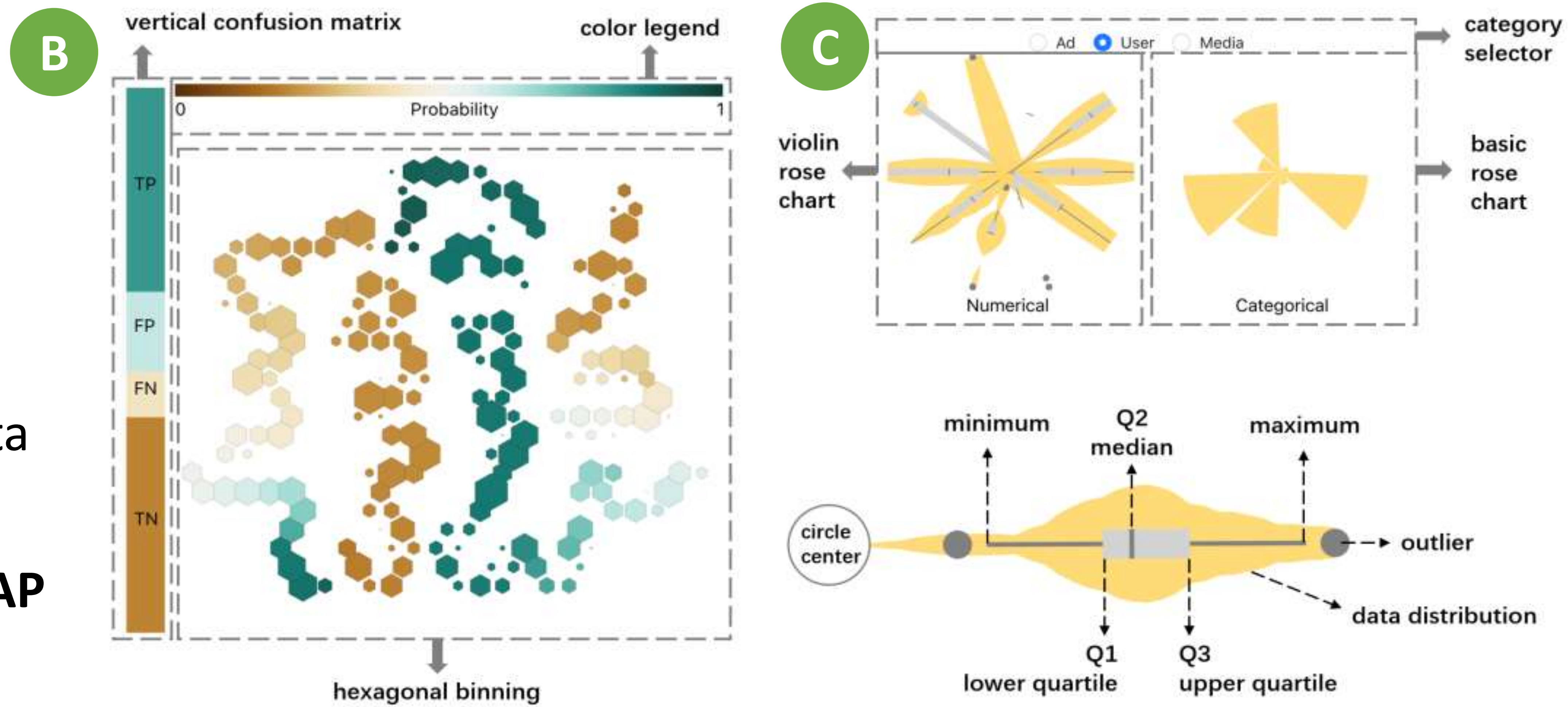
B, C, D: Instance-level Views

Hexagonal bining: **Data Overview**

- uses hexagons to aggregate similar data samples that fall within its boundaries after dimensionality reduction by **UMAP**

Rose charts: **Data Statistics**

Table: **Data Details**



D

Record ID	Probability	Label	city	city_rank	device_name	device_size	career	gender	net_type	residence	emui_version
4227959	0.93191	1	287	2	52	193	9	3	2	11	24
33457928	0.93162	1	287	2	29	141	9	3	2	11	20

VISUALIZATION

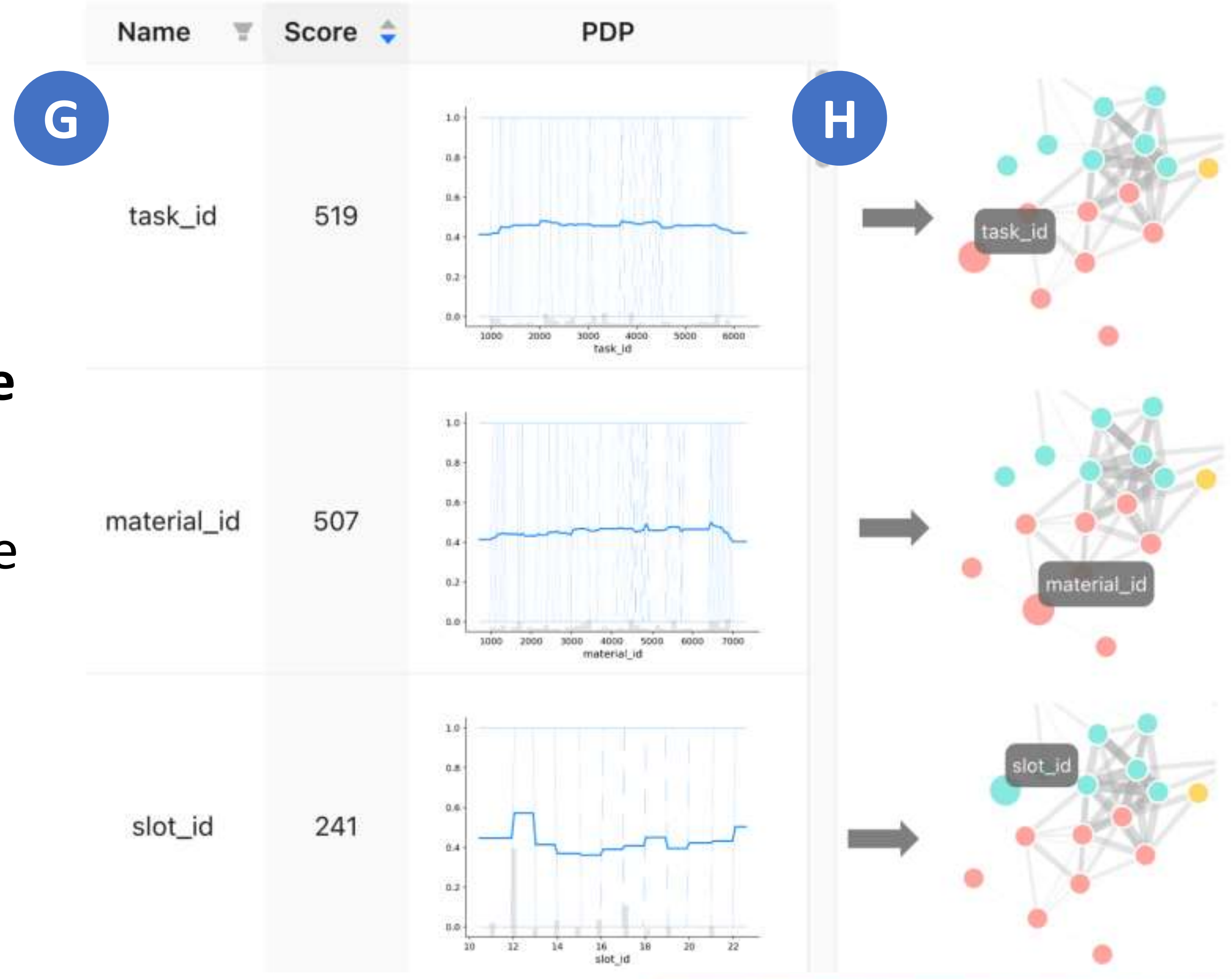
G, H: Feature-level Views

List combined with dual-axis plot: **Feature Importance**

- Feature importance represents the number of times the feature is selected as the splitting feature in all decision trees

Node-link chart: **Feature Correlation**

- Spearman's correlation coefficient



VISUALIZATION

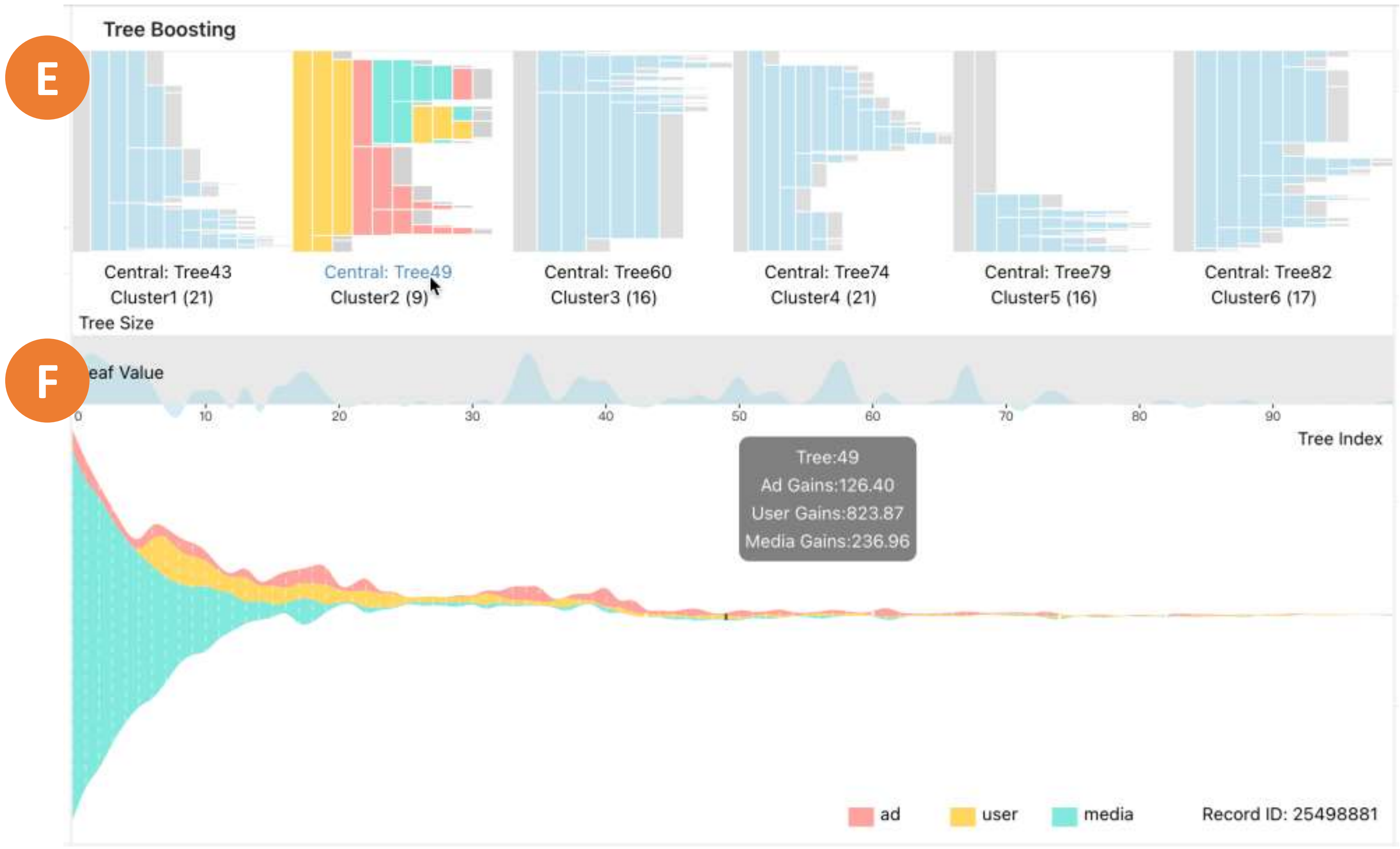
E, F: Model-level Views

Icicle: **the most representative K Decision Trees**

- Zhang-Shasha algorithm calculates the tree edit distance
- K-Medoids algorithm clusters the trees using a tree distance matrix

Area: **evolution of the tree size**

Streamgraph: **evolution of the information gain**



ChinaVis 2023

CONTENTS

BACKGROUND

SYSTEM OVERVIEW

VISUALIZATION

EVALUATION

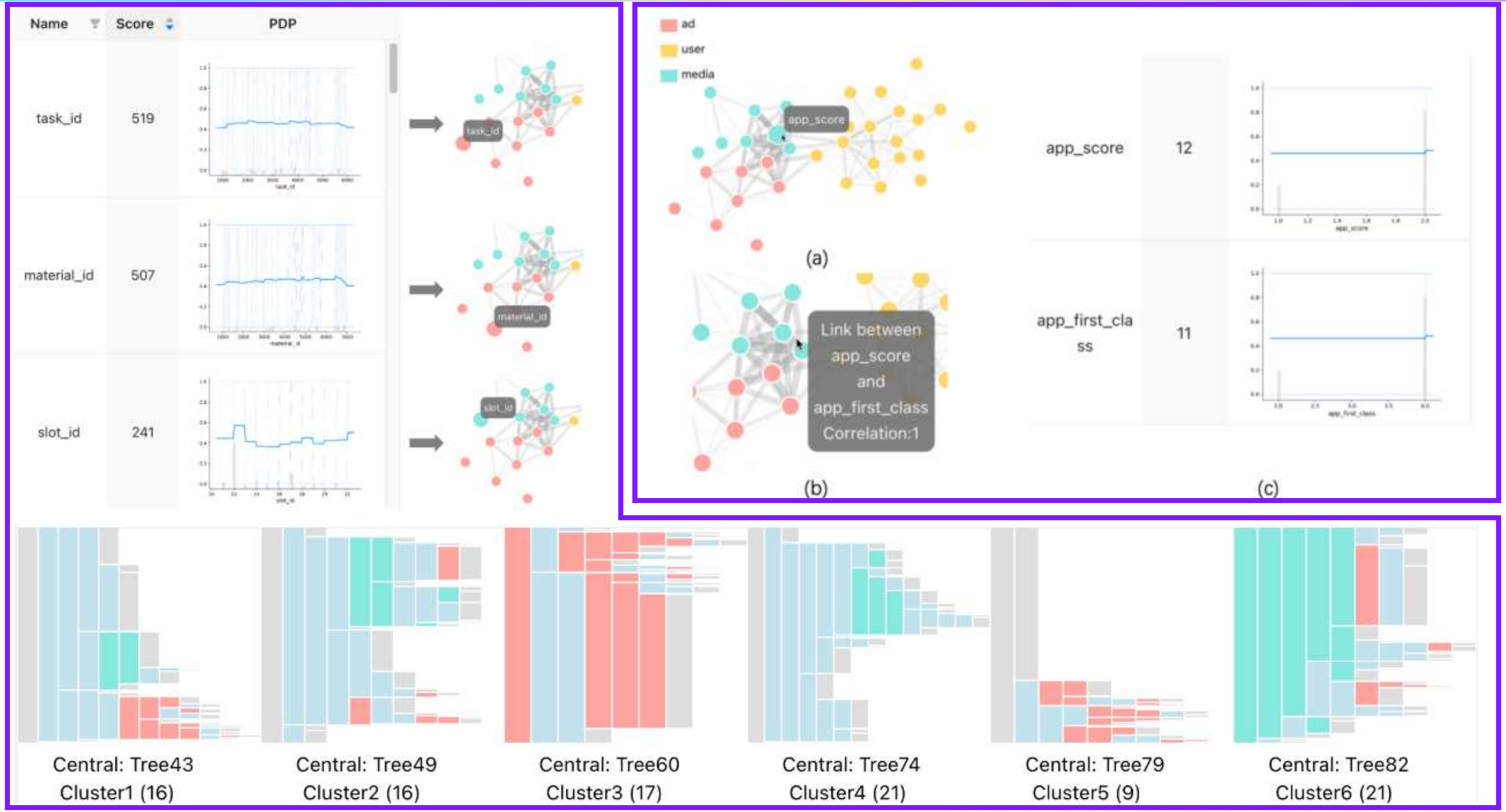
CONCLUSION

EVALUATION

Case Study

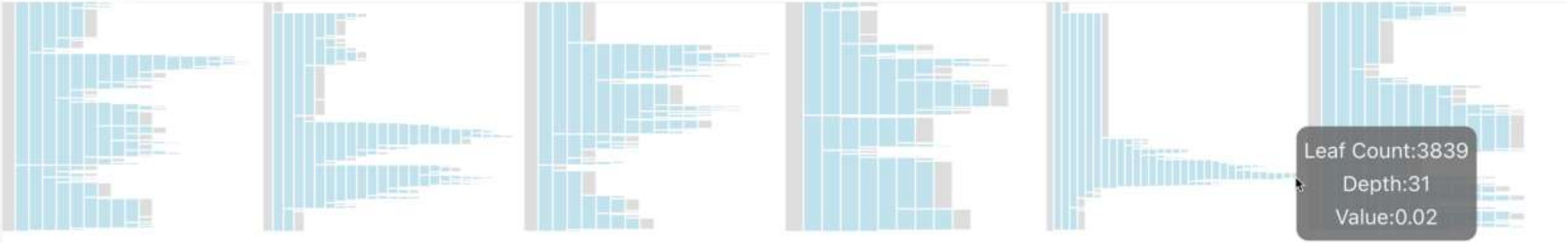
EVALUATION

Case Study - Analyzing and selecting features

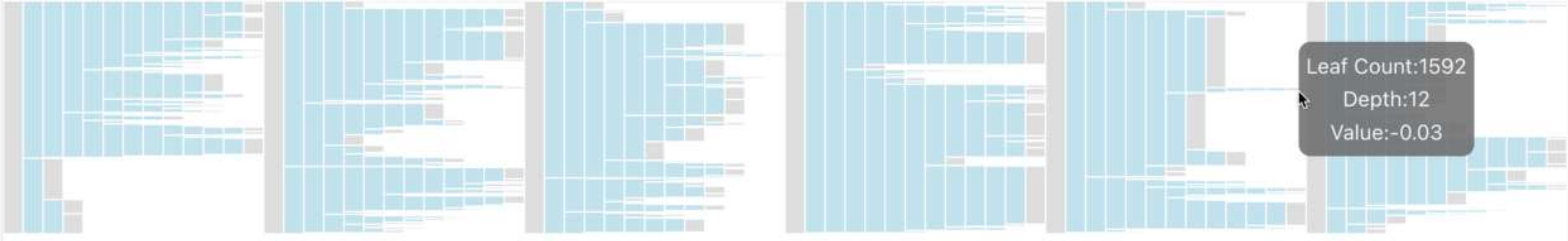


EVALUATION

Case Study - Analyzing and Tuning Model Structures



(a) num_leaves=100, max_depth=-1

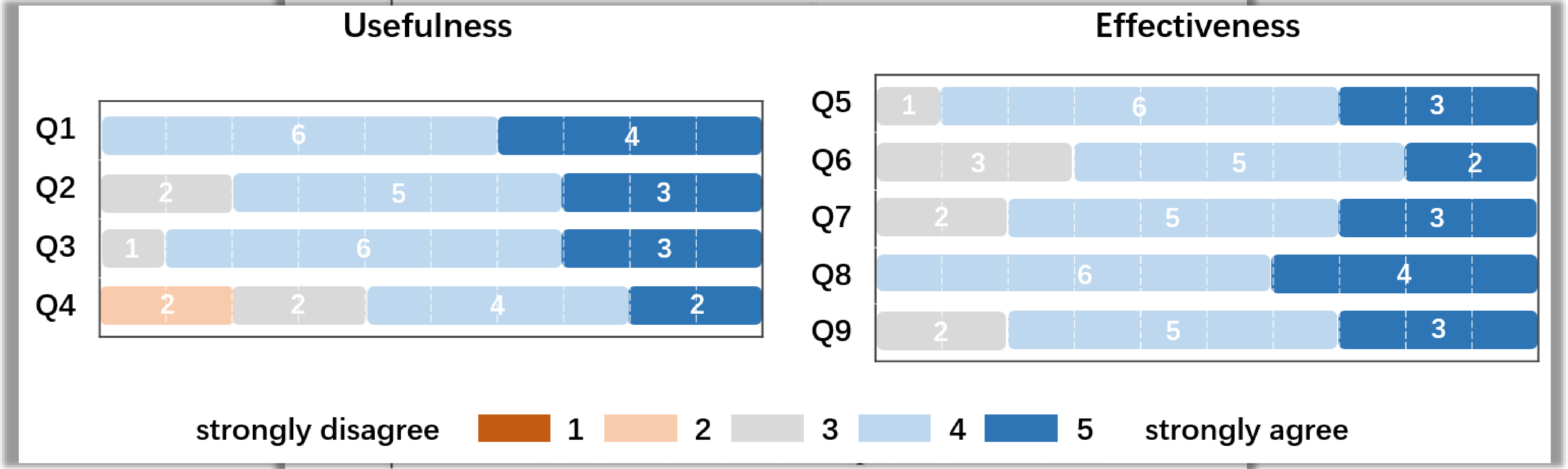


(b) num_leaves=100, max_depth=12

EVALUATION

Expert Evaluation

No.	Question
Q1	I think it's easy to learn the system
Q2	I think it's easy to understand the visual design of the



ChinaVis 2023

CONTENTS

CONCLUSION

GBDT4CTRVis helps advertising analysts **understand** the working mechanism of GBDT-based CTR prediction model from three levels: **instance, feature, and model**, and facilitate the process of **model tuning**

Limitation & Future Work

- Improving system **response time**: Using high-performance devices and optimizing the time complexity of algorithms.
- Enriching model **tuning strategies**: Applying more parameter optimization techniques such as grid search.
- Optimizing **visualization and interaction design**: Enriching the system's functionality.
- **Generalization**: Applied to other fields that use GBDT for binary prediction.

ChinaVis 2023

THANK YOU FOR LISTENING

Wenwen Gao, Shangsong Liu, Yi Zhou, Fengjie Wang, Feng Zhou, Min Zhu*

Sichuan University

中国•重庆 Chongqing•China