Design Explanation Microservices and Provenance: A Case Study of Explaining Cloud AI Service

Zerui Wang, Jun Huang, Anh Hoang Nguyen, Ding Li, Yan Liu

Concordia University

INTRODUCTION

The Need for Explainable AI (XAI) in Cloud AI Services:

- The current state of Cloud AI services is broad usage but lacks transparency and explainability.
- The Cloud AI services only provide general performance metrics but remain opaque on how the prediction is produced.

The Challenge of XAI for Cloud AI Services:

- Need of explanation results without unfolding the network structure of the learning model.
- XAI operations should be assessable at the same stage as learning performance evaluation.

XAI-as-a-Service:

- Designed using a microservice architecture to integrate AI models and XAI methods.
- Collect provenance data from XAI operations to enable traceability.

Case Studies:

- Results demonstrate the ability to generate reliable explanations for cloud AI services.
- Evaluation comprises XAI results evaluation and system-level evaluation.

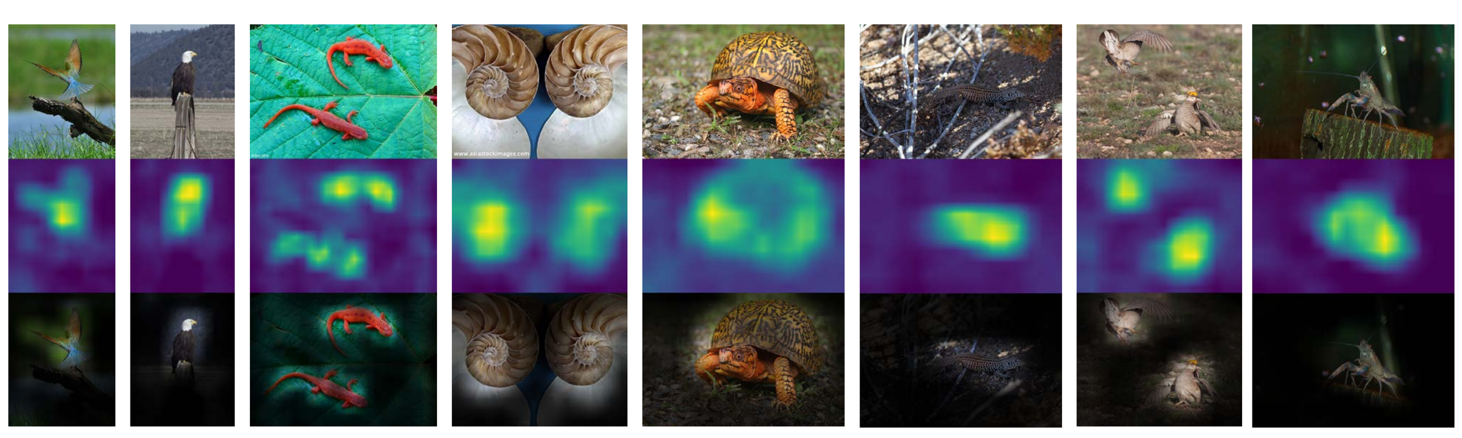

Figure 5. Imagenet data samples explanations

RESEARCH QUESTIONS

- *RQ1: What are the **key components for an XAI service** capable of handling diverse AI tasks, data types, and models?*
- *RQ2: **Which XAI methods** can **be integrated into the XAI service** to ensure practical and comprehensive explanations across various AI models and task domains?*
- *RQ3: What's an efficient **strategy for integrating XAI with cloud AI services**, managing **custom data**, and **executing diverse tasks** for reliable explanations?*
- *RQ4: How can XAI service guarantee **reproducible explanation** results, boosting transparency and reliability in AI decision-making?*

METHODOLOGY

XAI Service is Built on a Four-layered Microservice Architecture:

- **User Interface:** Allows users to view, access data, set up, and execute tasks.
- **Coordination Center:** Receives user requests, manages microservices, handles data representation, prepares data provenance, and evaluates the system performance.
- **Microservice Layer:** Encapsulates AI models, XAI methods, data provenance, and evaluations.
- **Data Persistence Layer:** Manages and stores datasets, operation data, XAI results, and evaluations.

See **Figure 1** for a visual representation of the XAI service API architecture.

CASE STUDY RESULTS

TABLE II
CAM-BASED PREDICTION CHANGES DISTRIBUTION STATISTIC

Statistics	*Grad-CAM*	*Grad-CAM++*	*Eigen-CAM*	*Layer-CAM*	*XGrad-CAM*
	CAM-based XAI methods using ResNet				
$Mean$	27.0461	25.9198	63.2657	**25.5333**	27.0476
STD [a]	26.5806	25.9853	34.7627	**25.7503**	26.5802
P_{25} [b]	4.8230	**4.6346**	29.8233	4.6865	4.8394
P_{50} [b]	17.3631	15.9730	76.7189	**15.8819**	17.3631
P_{75} [b]	43.5777	40.7204	95.1838	**39.3082**	43.5777
	CAM-based XAI methods using DenseNet				
$Mean$	**25.5995**	26.4856	69.3031	26.9097	36.1336
STD [a]	**25.1781**	25.5827	31.9986	25.4721	29.4764
P_{25} [b]	**5.1250**	5.6601	43.4602	6.5918	8.5504
P_{50} [b]	**16.4289**	18.1160	85.6432	18.5324	29.4022
P_{75} [b]	**38.8719**	41.9311	95.5513	42.1751	61.8364

[a]: Standard Deviation, [b]: Percentile

The case study utilizes the **ImageNet dataset** to perform **XAI methods** on **Cloud AI services.** Results show varying **prediction changes** of different XAI methods, offering insights into the explanation results.

This table illustrates **prediction changes** statistics for five distinct **CAM-based methods.** The smaller the prediction changes after masking, the higher the XAI result accuracy. Thus, this table serves as a valuable guide for assessing each method.

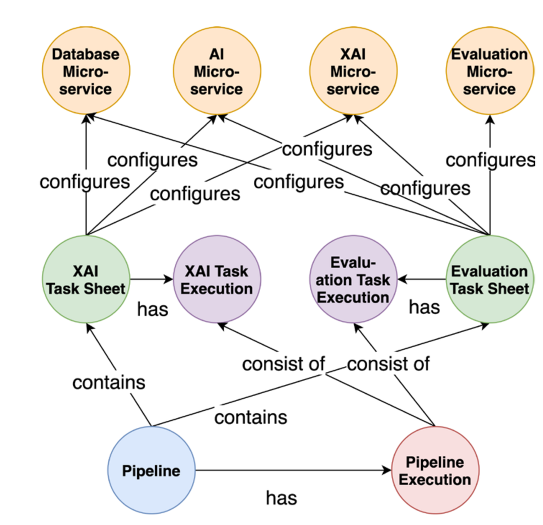

Figure 2. The nodes and edges definition for XAI provenance data.
This robust, structured provenance data representation enhances reproducibility, allowing for precise replication of settings and procedures.

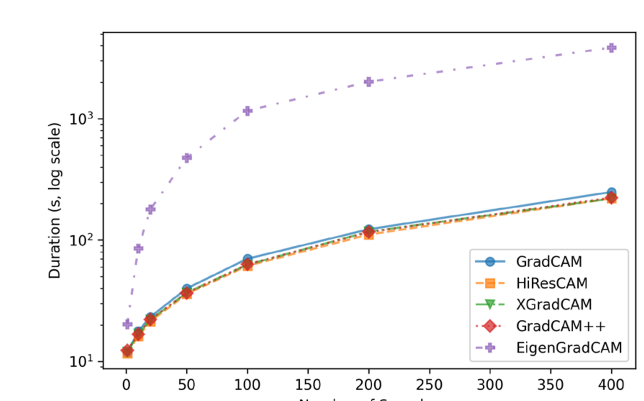

Figure 3. Response time of XAI microservices for varying sample sizes
The chart visualizes the relationship between the response time of various XAI microservices and the size of input samples. Each curve represents a distinct XAI method, demonstrating its performance scalability with increasing data volume.

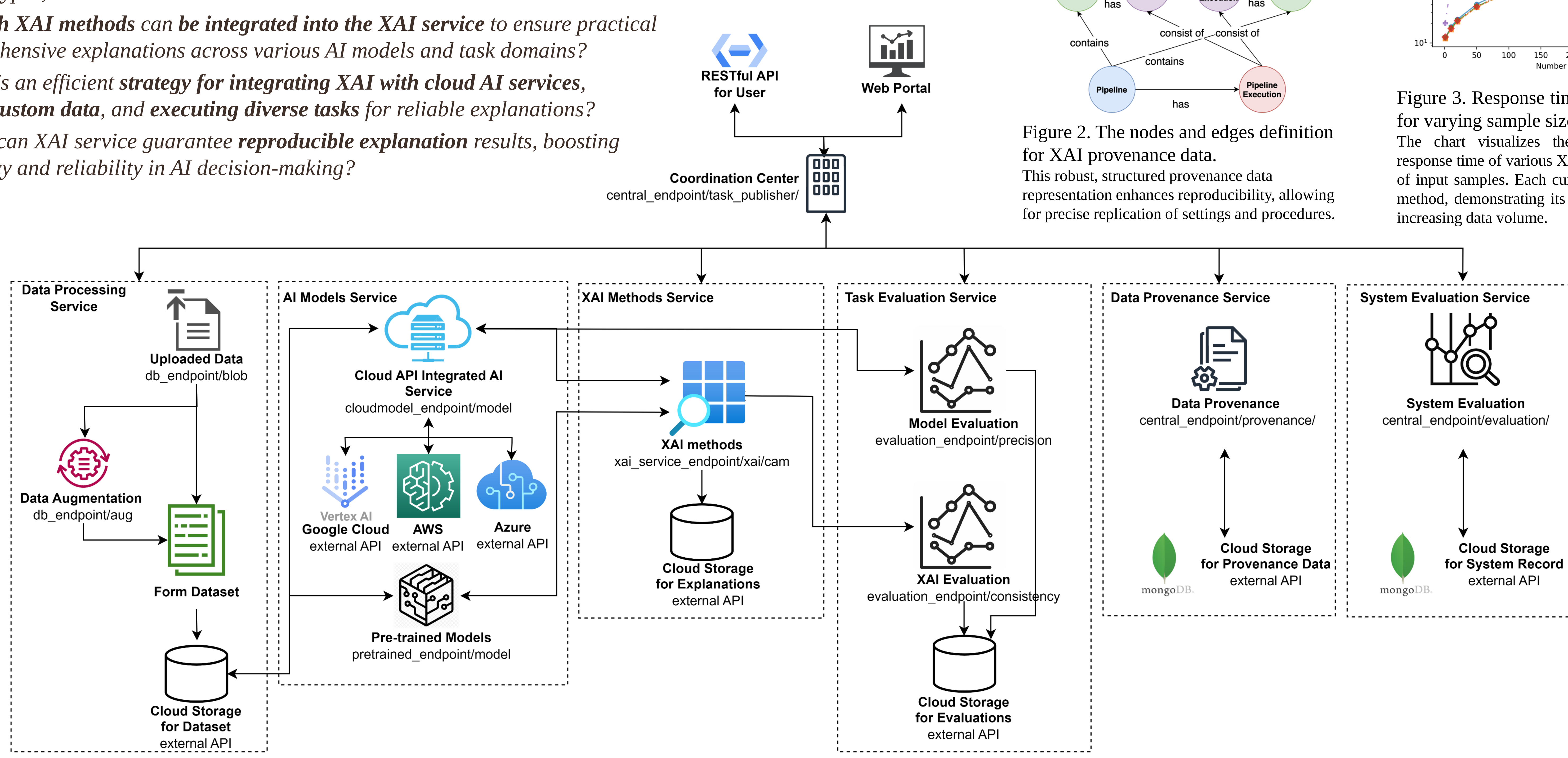

Figure 1. XAI Service API Architecture Diagram

RELATED WORKS

TABLE I
COMPARISON OF XAI FRAMEWORKS

Framework	Publisher	Supported Data Types	Supported XAI Methods	Results Presentation	Results Evaluation	Deployment	Compatibility with cloud AI services
Dalex [1]	Warsaw University	Tabular	1, 2, 3, 4, 5	Plot/Array	No	Standalone	Lacks explicit cloud support
Explainability 360 [2]	IBM	Tabular/Image /Text	1, 2, 6	Plot/Dashboard	No	Docker	Lacks explicit cloud support
InterpretML [3]	Microsoft	Tabular/Text	1, 2, 3, 10	Plot/Dashboard	No	Standalone	Lacks explicit cloud support
Captum [4]	Meta	Images/Text	1, 2, 6, 7, 8	Attribution Plot	Robustness Metrics	Standalone	Lacks explicit cloud support
OmniXAI [5]	Salesforce	Tabular/Image /Text/Timeseries	1, 2, 3, 4, 5, 6, 7, 8, 9	Plot/ Dashboard	No	BentoML	Lacks explicit cloud support
Vertex XAI	Google	Tabular/Image /Text	1, 2, 8	Attribution Plot	No	Vertex AI	Vertex AI
XAI service	**This work**	**Tabular/Image /Text**	**1, 2, 3, 4, 5, 7**	**Plot/Dashboard/ Result data**	**Consistency Metrics**	**Docker**	**Compatible API**

Supported XAI Methods: 1. LIME (Local Interpretable Model-agnostic Explanations) [6], 2. SHAP (SHapley Additive exPlanations) [7], 3. PDP (Partial Dependence Plots) [8], 4. ICE (Individual Conditional Expectation) [9], 5. ALE (Accumulated Local Effects) [10], 6. LRP (Layer-wise Relevance Propagation) [11], 7. CAM (Class Activation Mapping) [12], 8.Integrated Gradients [13], 9. Counterfactual Explanations [14], 10. Decision Rules [15]

REFERENCE

[1] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, and P. Biecek, “dalex: Responsible machine learning with interactive explainability and fairness in python,” The Journal of Machine Learning Research, vol. 22, no. 1, pp. 9759–9765, 2021.

[2] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović et al., “One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques,” arXiv preprint arXiv:1909.03012, 2019.

[3] H. Nori, S. Jenkins, P. Koch, and R. Caruana, “Interpretml: A unified framework for machine learning interpretability,” arXiv preprint arXiv:1909.09223, 2019.

[4] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, “Captum: A unified and generic model interpretability library for pytorch,” 2020.

[5] W. Yang, H. Le, S. Savarese, and S. C. Hoi, “Omnixai: A library for explainable ai,” arXiv preprint arXiv:2206.01612, 2022.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “why should I trust you? Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in neural information processing systems, vol. 30, 2017.

[8] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” Annals of statistics, pp. 1189–1232, 2001.

[9] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” journal of Computational and Graphical Statistics, vol. 24, no. 1, pp. 44–65, 2015.

[10] D. W. Apley and J. Zhu, “Visualizing the effects of predictor variables in black box supervised learning models,” Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 82, no. 4, pp. 1059–1086, 2020.

[11] A. Binder, S. Bach, G. Montavon, K.-R. Müller, and W. Samek, “Layer-wise relevance propagation for deep neural network architectures,” in Information science and applications (ICISA) 2016. Springer, 2016, pp. 913–922.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[13] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in International conference on machine learning. PMLR, 2017, pp. 3319–3328.

[14] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” Harv. JL & Tech., vol. 31, p. 841, 2017.

[15] H. Yang, C. Rudin, and M. Seltzer, “Scalable bayesian rule lists,” in International conference on machine learning. PMLR, 2017, pp. 3921–3930.