AI Explainability WG

April 2023

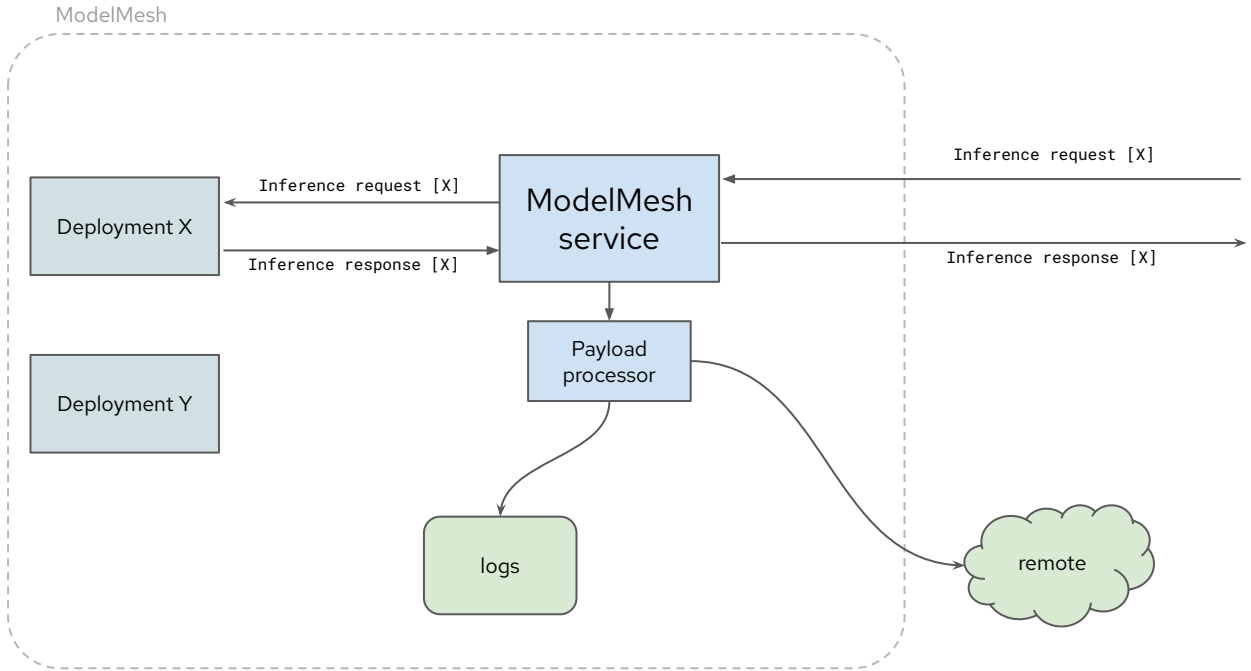
2023 Recap

- TrustyAI now a part of Open Data Hub 1.5.0[1]
- TrustyAI service 0.1.0
  - Integrated with ModelMesh[2]
  - Support for group fairness metrics
    - Currently SPD and DIR
  - PVC storage support
- Python TrustyAI 0.2.10
  - Workbench images available for Open Data Hub 1.5.0[3]
    - Metrics and explainers (LIME, SHAP, Counterfactuals) from Jupyter

[1]: http://opendatahub.io/news/2023-04-11/odh-release-1.5.0-blog.html
[2]: https://github.com/kserve/modelmesh/pull/84
[3]: https://github.com/opendatahub-io/notebooks

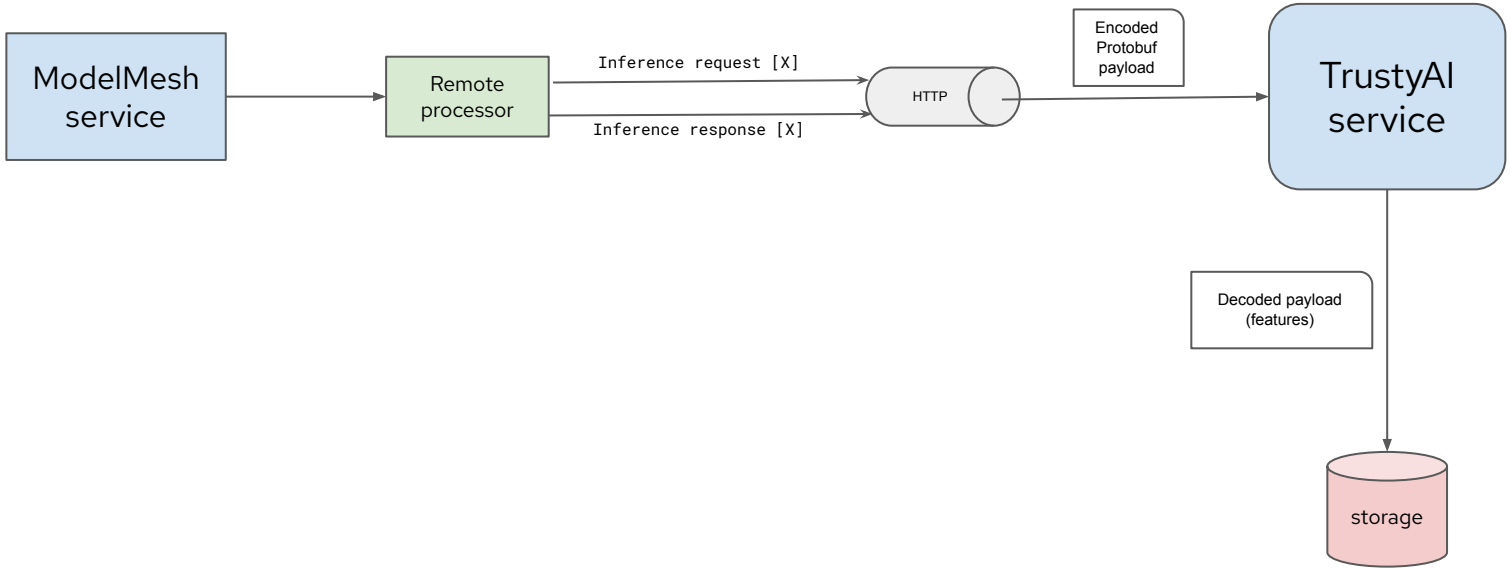
TrustyAI service

ModelMesh payload processor

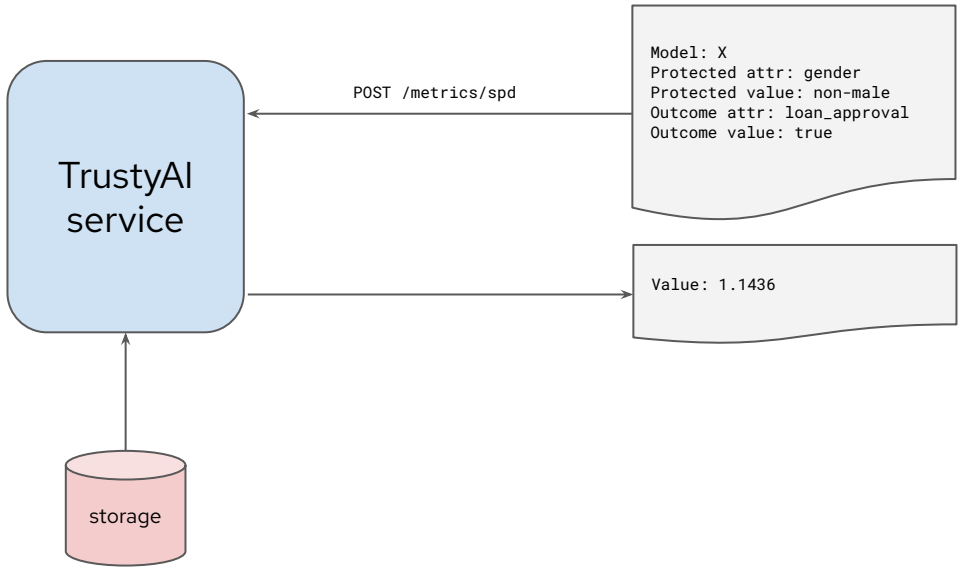


https://github.com/kserve/modelmesh/pull/84

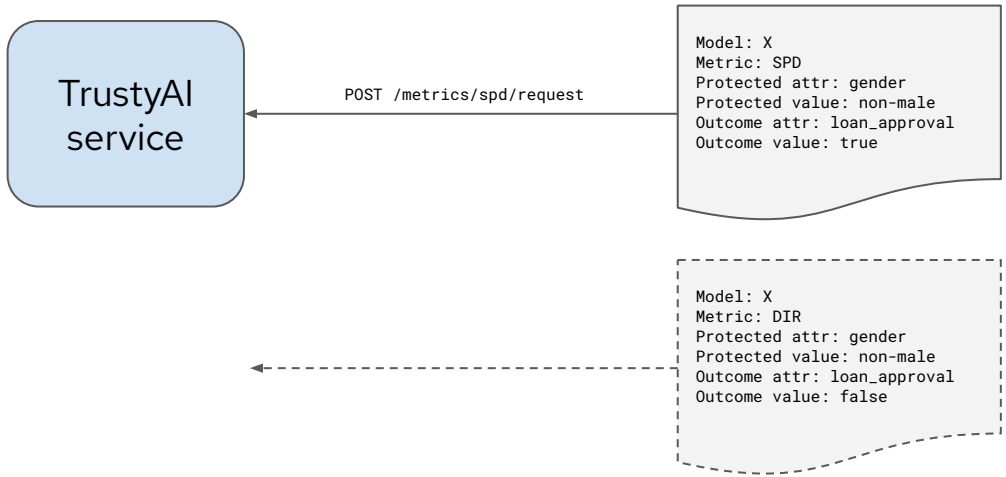
Payload storage



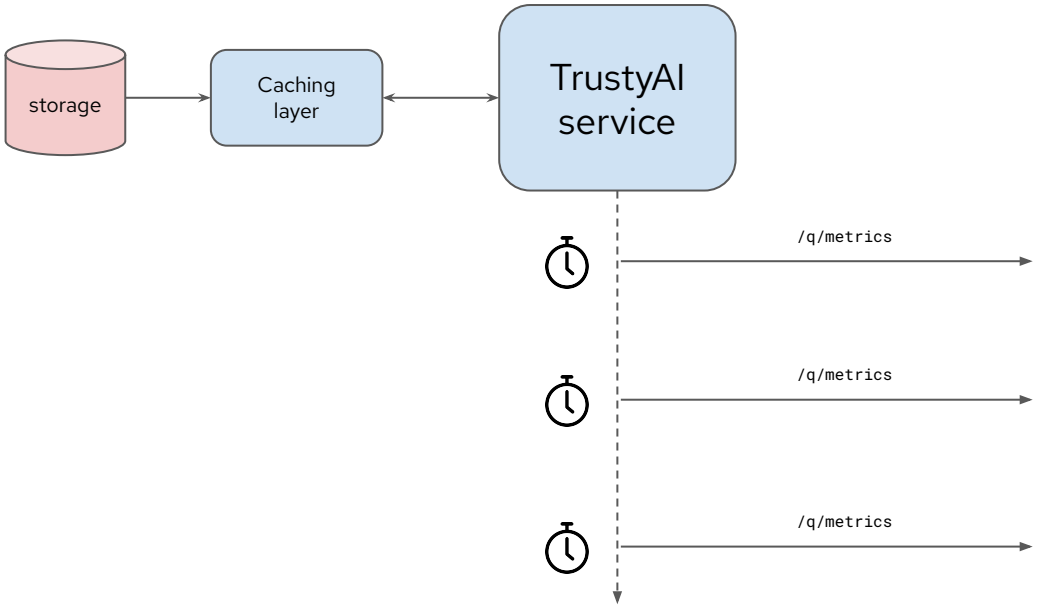
“Ad-hoc” requests



Scheduled metrics



Scheduled metrics



```
trustyai_spd{
        favorable_value="1",
        instance="trustyai:8080",
        job="trustyai-service",
        model="example",
        outcome="income",
        privileged="1",
        protected="gender",
        request="e4bf1430-cc33-48a0-97ce",
        unprivileged="0"
}
```

Roadmap 2023

TrustyAI 2023 roadmap

July 2023 TrustyAI 0.2.0

- *Explainers*
  - Support for explainers LIME, SHAP, CF at service level
- *Metrics*
  - Flexible scheduling/batching
  - Improve service metadata endpoints
    - Include available categories

September 2023 TrustyAI 0.3.0

- *Explainers*
  - Support for external explainability libraries
- *Metrics*
  - Additional metrics
  - Metrics statistical tests

December 2023 TrustyAI 0.4.0

- *Storage*
  - Wider storage support (database backends)
- *Explainers*
  - NLP explainability support
- *Metrics*
  - Support for user-defined historical windows

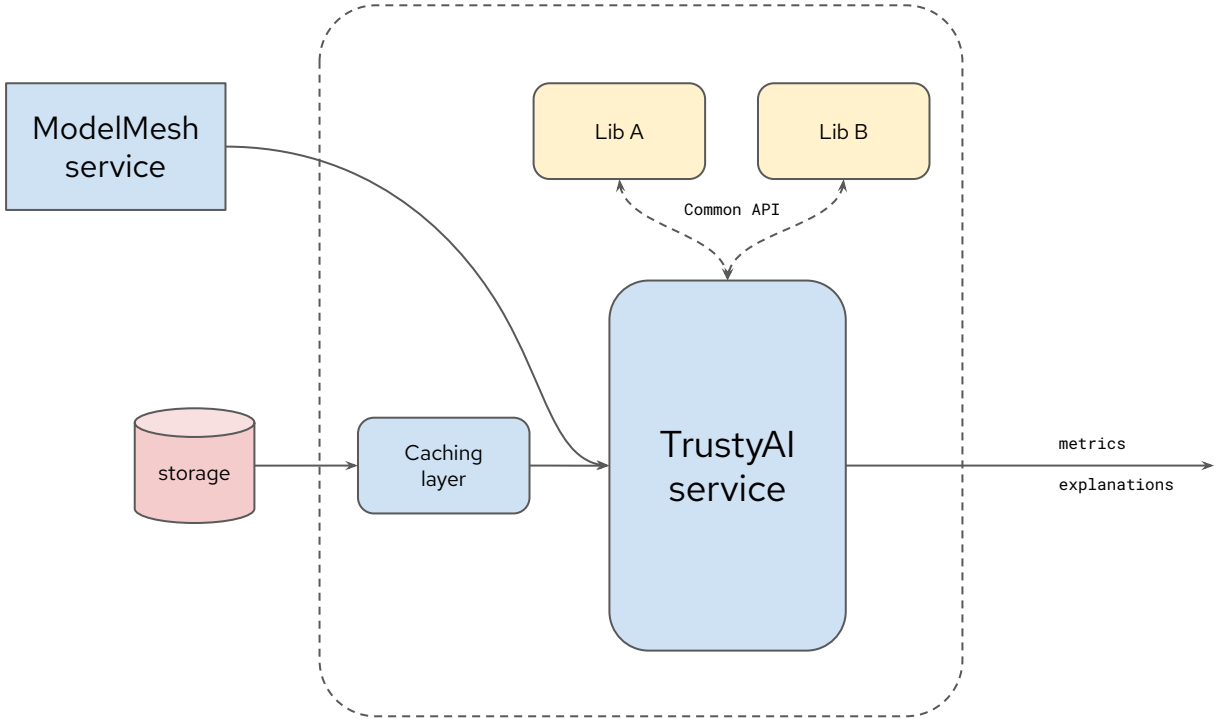
External library support

External library support

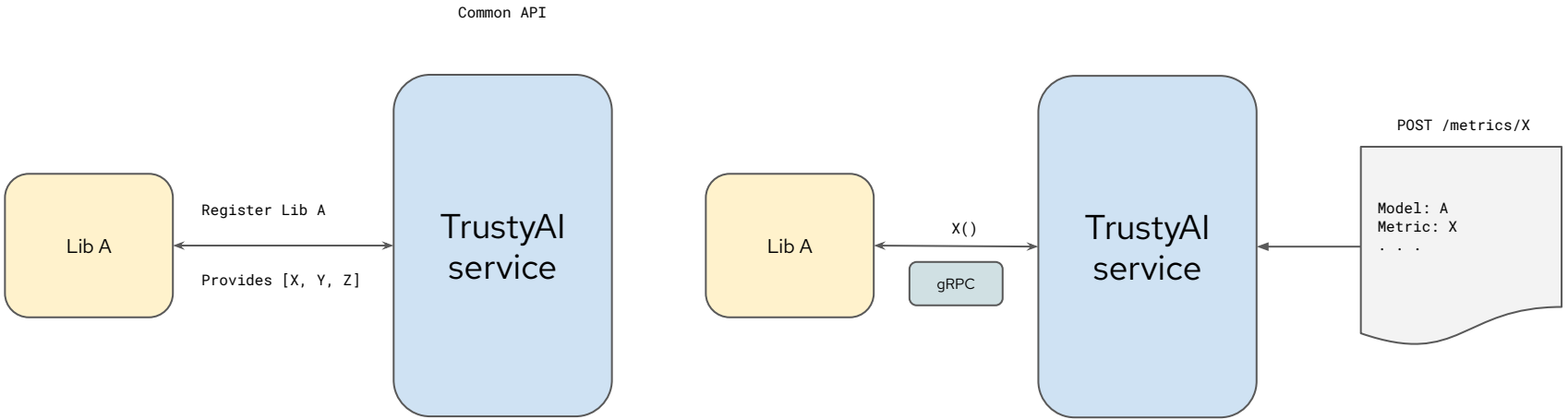
- Define architecture and API for TrustyAI extensibility
  - Both metrics and explainability
- KServe: Enable explainers to work with ModelMesh[1]

[1]: https://github.com/kserve/kserve/issues/2388

External library support



External library support



LLMs and explainability

LLMs and explainability

- “...more formal theories and principles to understand, characterize, and explain the abilities or behaviors of LLMs are still missing..”
  - A Survey of LLMs (https://arxiv.org/abs/2303.18223)
- Teaching LLMs to Self-Debug
  - https://arxiv.org/abs/2304.05128
- Complementary Explanations for Effective In-Context Learning
  - https://arxiv.org/abs/2211.13892
- WebBrain
  - https://arxiv.org/abs/2304.04358