AI Explainability WG

May 2023

May 2023 updates

- TrustyAI service
  - Improvements to metrics endpoints
    - Calculation data size configurable per-metric
    - Metadata includes representative values
    - More robust REST payload validation
    - Per-metric configurable thresholds
- Python TrustyAI 0.2.12
  - Mostly fixes around the Tyrus visualisations[1]
- Community
  - Architectural Decision Records (ADRs) for proposals[2]

[1]: https://pypi.org/project/trustyai/0.2.12/
[2]: https://github.com/trustyai-explainability/community/tree/main/adr

Community

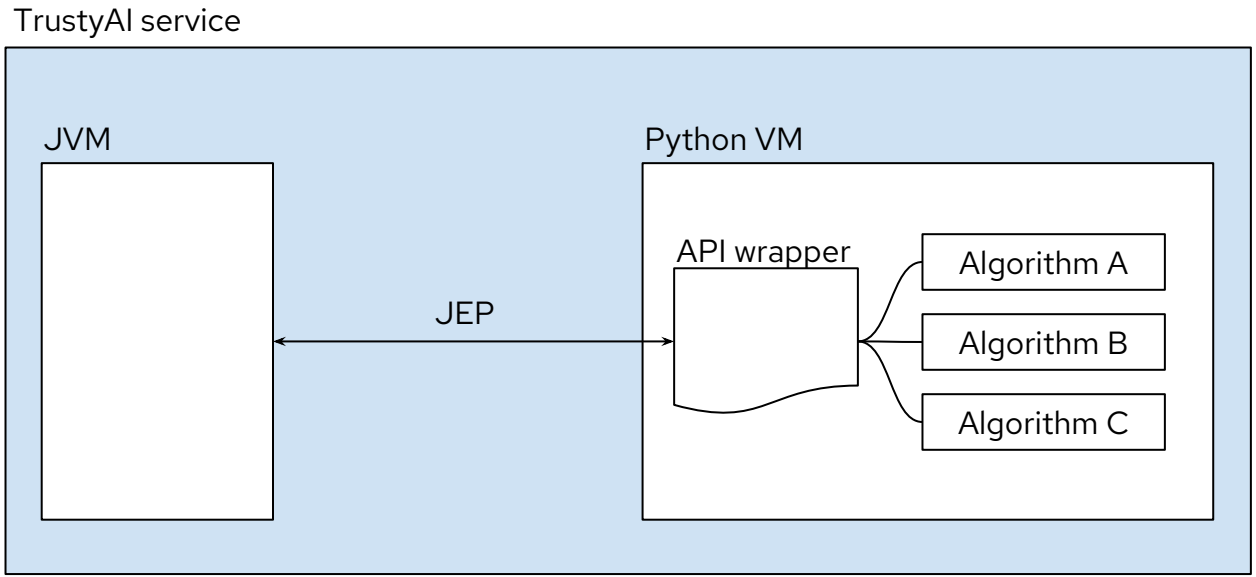
Community

- Community repository[1]
  - Community information
  - Contribution guidelines
  - Code of conduct
- Architectural Decision Records (ADRs)
  - Approved ADRs:
    - ADR-0001: TrustyAI external library integration[2]
    - ADR-0002: Metrics and XAI namespaces[3]

[1]: https://github.com/trustyai-explainability/community
[2]: https://github.com/trustyai-explainability/community/blob/main/adr/ADR-0001-trustyai-external-library-integration.md
[3]: https://github.com/trustyai-explainability/community/blob/main/adr/ADR-0002-metrics-and-xai-namespaces.md

TrustyAI external library integration

External library support



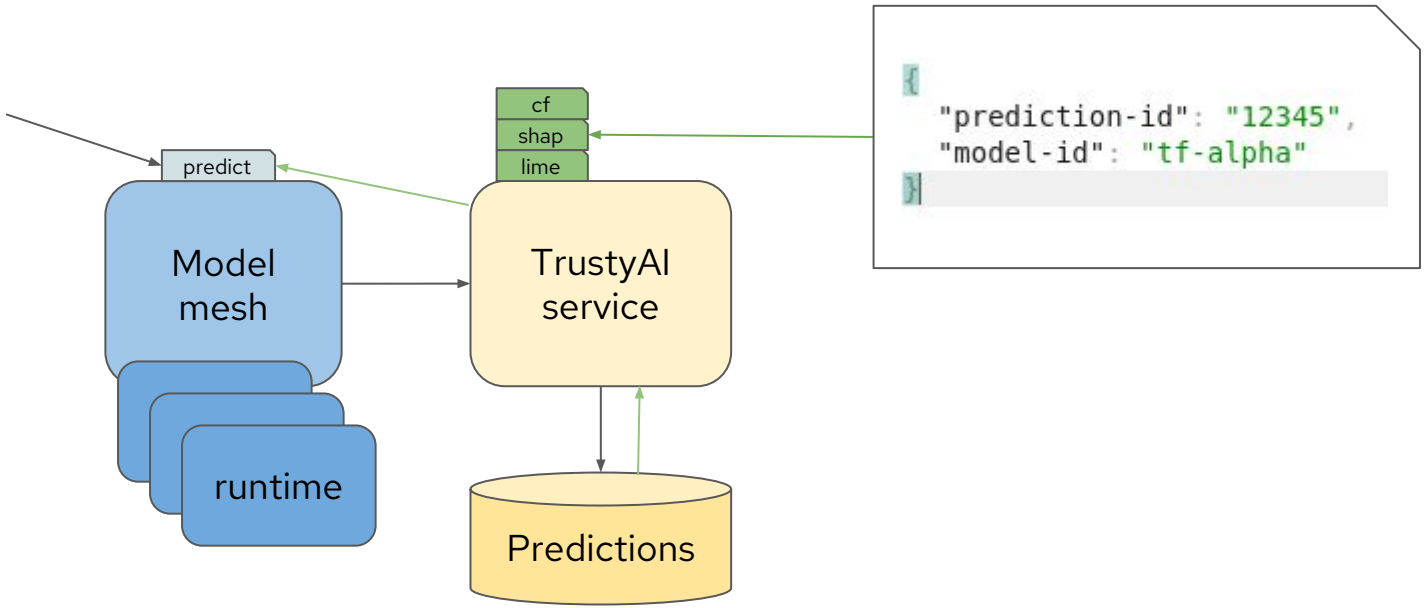
https://github.com/trustyai-explainability/community/blob/main/adr/ADR-0001-trustyai-external-library-integration.md#proposal

External library support

- Communication API
  - Ongoing development
  - Determine general requirements for metrics / explainers
    - Representative APIs
    - Data structures for marshalling
    - Performance benchmarks

TrustyAI Explainers Endpoints

Explainers Endpoints



https://github.com/trustyai-explainability/trustyai-explainability/pull/164

Roadmap

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