TrustyAI Introduction

Kubeflow Community Meeting

TrustyAI Team

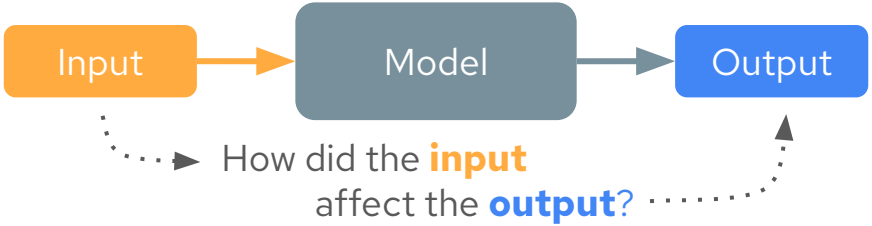
Why TrustyAI?

- Ethical AI
- Trustworthy AI
- Regulatory compliance
- Reliability
- Harmful content measures

What is Explainability?

- Explainability or XAI is the process of producing **human interpretable** explanations of complex model behaviour

What is Explainability?



- Explainability or XAI is the process of producing **human interpretable** explanations of complex model behaviour
- This is typically done by describing how **input features** affect the **model's outputs**

What is Fairness?

- AI fairness refers to the design, development, and deployment of AI systems in a way that ensures they operate equitably and do not include biases or discrimination against any individual or group

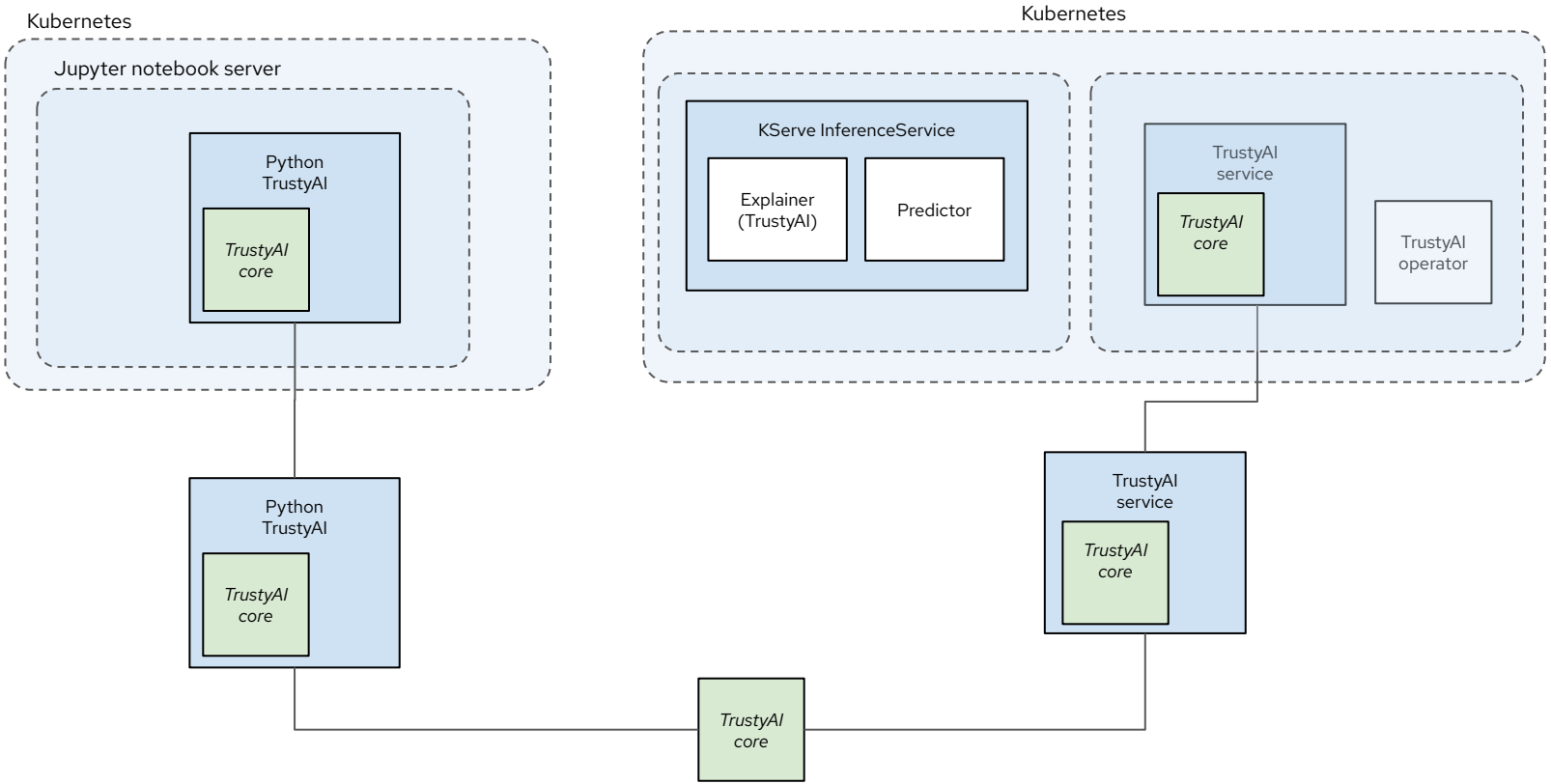
Architecture

What is TrustyAI?

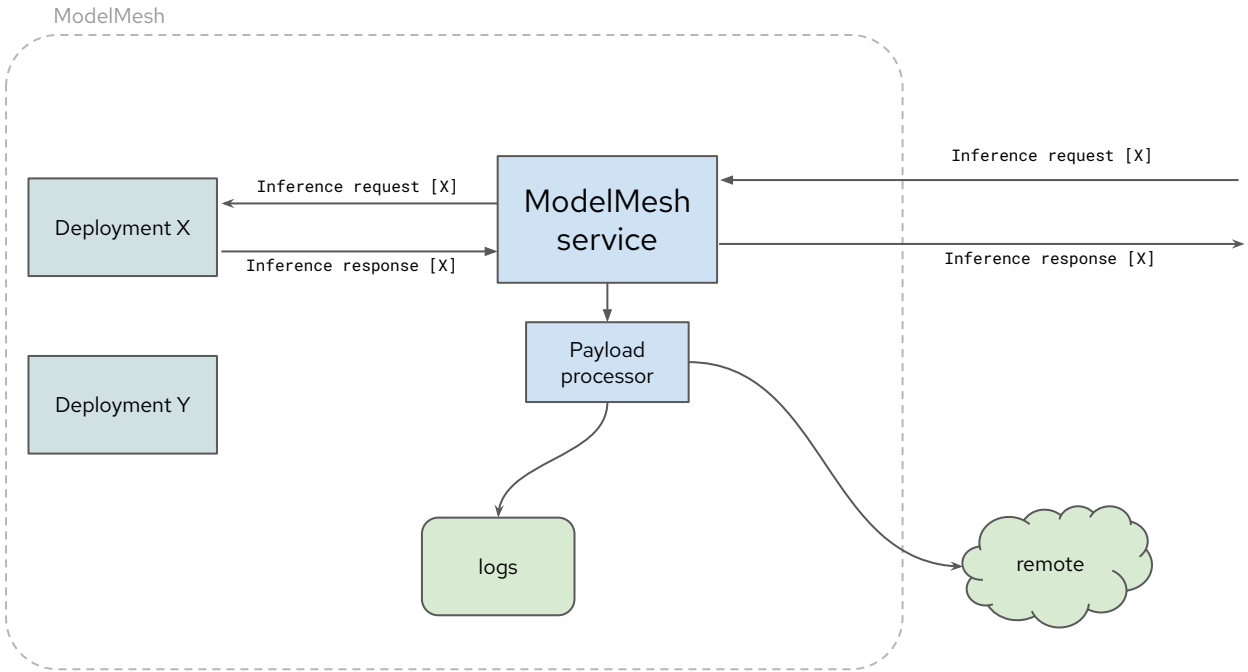
TrustyAI consists of several components, including:

- A core library[1] containing a suite of explainability algorithms and metrics to benchmark explanation quality and model fairness
- A Python module to interface with the core library, combining the ease-of-use of Python with the speed of the compiled Java library
- A REST service for fairness metrics and explainability algorithms including KServe and ModelMesh integration.
- A Kubernetes operator for TrustyAI service
- A KServe explainer that provides explanations for predictions made by using the built-in KServe explainer support

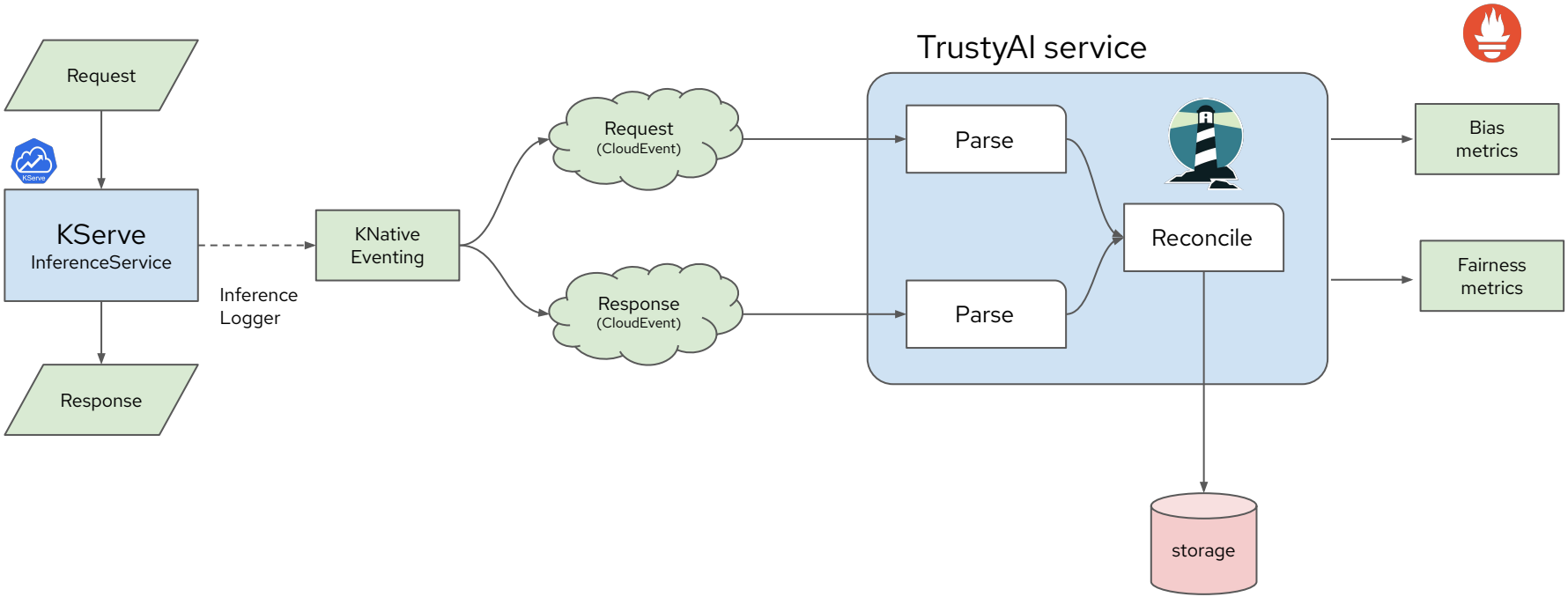
[1] - Core library implemented mainly in Java


Kubernetes
Jupyter notebook server
Python
TrustyAI
TrustyAI
core
Kubernetes
KServe InferenceService
Explainer
(TrustyAI)
Predictor
TrustyAI
service
TrustyAI
core
TrustyAI
operator
Python
TrustyAI
TrustyAI
core
TrustyAI
service
TrustyAI
core
TrustyAI
core

TrustyAI


ModelMesh
Deployment X
Deployment Y
Inference request [X]
Inference response [X]
ModelMesh service
Inference request [X]
Inference response [X]
Payload processor
logs
remote

TrustyAI


Request
KServe
InferenceService
Response
Inference
Logger
KNative
Eventing
Request
(CloudEvent)
Response
(CloudEvent)
TrustyAI service
Parse
Parse
Reconcile
storage
Bias
metrics
Fairness
metrics

TrustyAI

Related Projects

- trustyai-ood, a library for model certainty enablement with out-of-distribution (OOD) detection
- trustyai-detoxify, algorithms and tools for detecting and fixing hate speech, abuse and PII disclosure (HAP) in text generated by LLMs

TrustyAI vs Competitors

TrustyAI vs Competitors

	TrustyAI	Fiddler.ai	CogScale Certifai	Seldon Alibi	IBM AIF/AIX360
Accessibility	Free + open source	$50k annually Closed source	$8.8-52k annually Closed source	BSL 1.1	Free + open source
Development	Python, Java (potentially R)	Python	Python + YAML + CLI	Python	Python
Algorithm Implementations	Custom, efficient implementations	Closed source	Closed source	Imported from other libraries	Implemented + Imported
Releases in 2024	9	7	0	2	2
Activity (commits in last month)	26	-	-	0	1
Supported Model Types	Tabular, Time-series, LLM*, CV*	Tabular, CV, LLM	Tabular	Tabular	Tabular, Timeseries
Integrations	ODH, KServe, ModelMesh	AWS	Paid subscriptions get: AWS, AZURE, RHOS	Removed from KServe	-

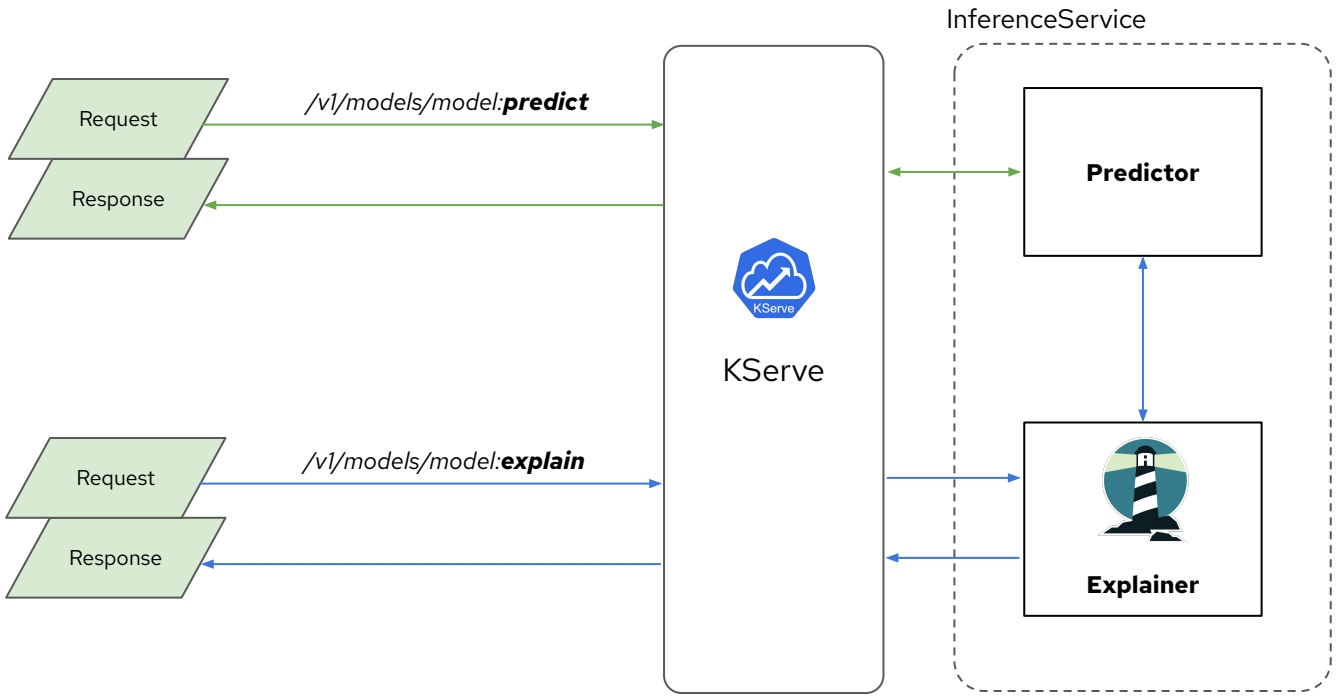
*ongoing development

Source:
TrustyAI Vs Competitors Market Analysis

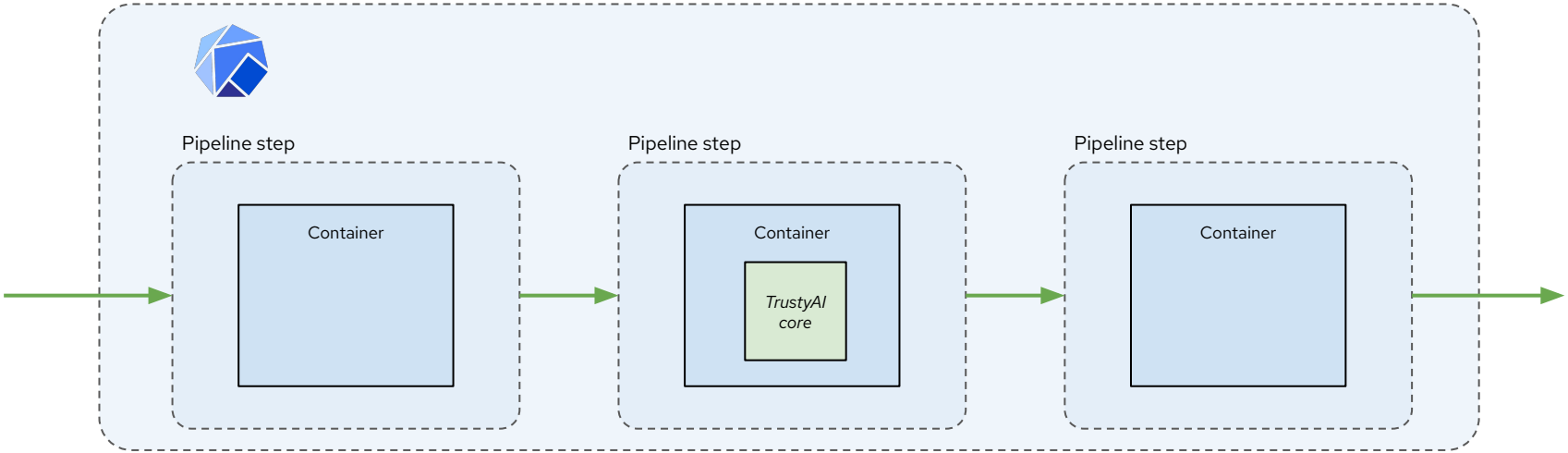
Integrations with Kubeflow

Integrations with Kubeflow

- Pre-built TrustyAI Kubeflow Notebooks image on Open Data Hub and Red Hat OpenShift AI
- KServe integration that provides explanations for predictions made by AI/ML models using the built-in KServe explainer support for LIME and SHAP
- Model Registry
- Feast
- *Pipelines*


InferenceService
Request
Response
/v1/models/model:**predict**
KServe
KServe
Predictor
Request
Response
/v1/models/model:**explain**
Explainer

Kubeflow pipeline



- Real-time bias/fairness calculations
- Data drift
- Toxic content filtering/scoring/masking
- Global explainability

TrustyAI

Future Work

Future Work

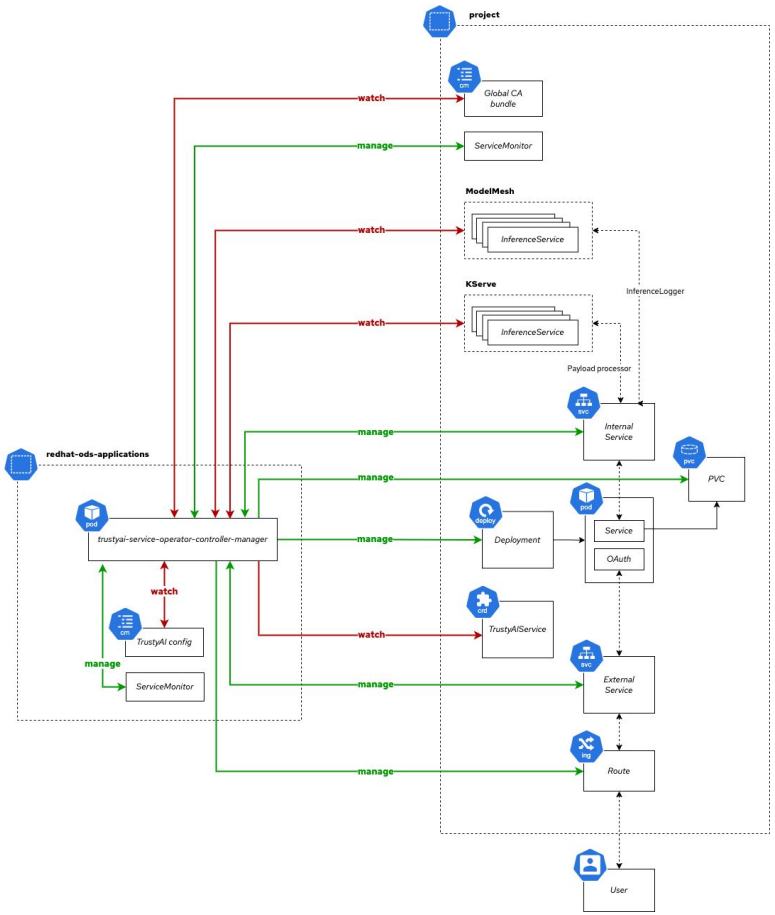
- Generalize our algorithms to support arbitrary numeric data, which will provide support for vision, audio, etc.
- Incorporating other open-source projects such as the lm-evaluation-harness to provide LLM evaluation capabilities

Resources

Resources

- TrustyAI Java library
- TrustyAI Python Bindings
  - github docs tutorial examples
- Community
  - slack roadmap discussion board
- Documentation

TrustyAI Kubernetes architecture



Architectural details available at https://github.com/opendatahub-io/architecture-decision-records/tree/main/architecture/components/explainability

TrustyAI