Model Governance and Explainable AI
as tools for legal compliance and risk management

INNOQ

inlinity
trusting artificial intelligence

ISABEL BÄR
@isabel_baer

KILIAN KLUGE
@kilian_kluge

Let’s start with an example…

Amazon's sexist AI recruiting tool: how did it so wrong?

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

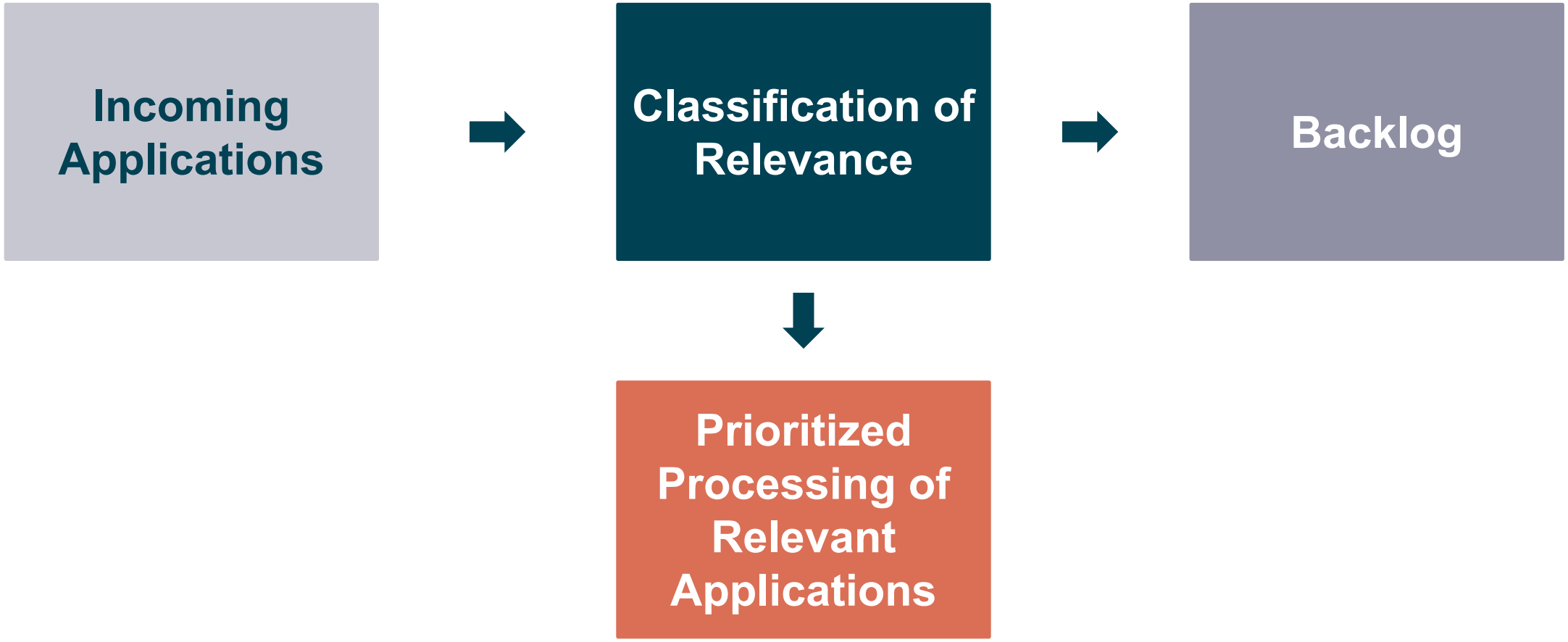
By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

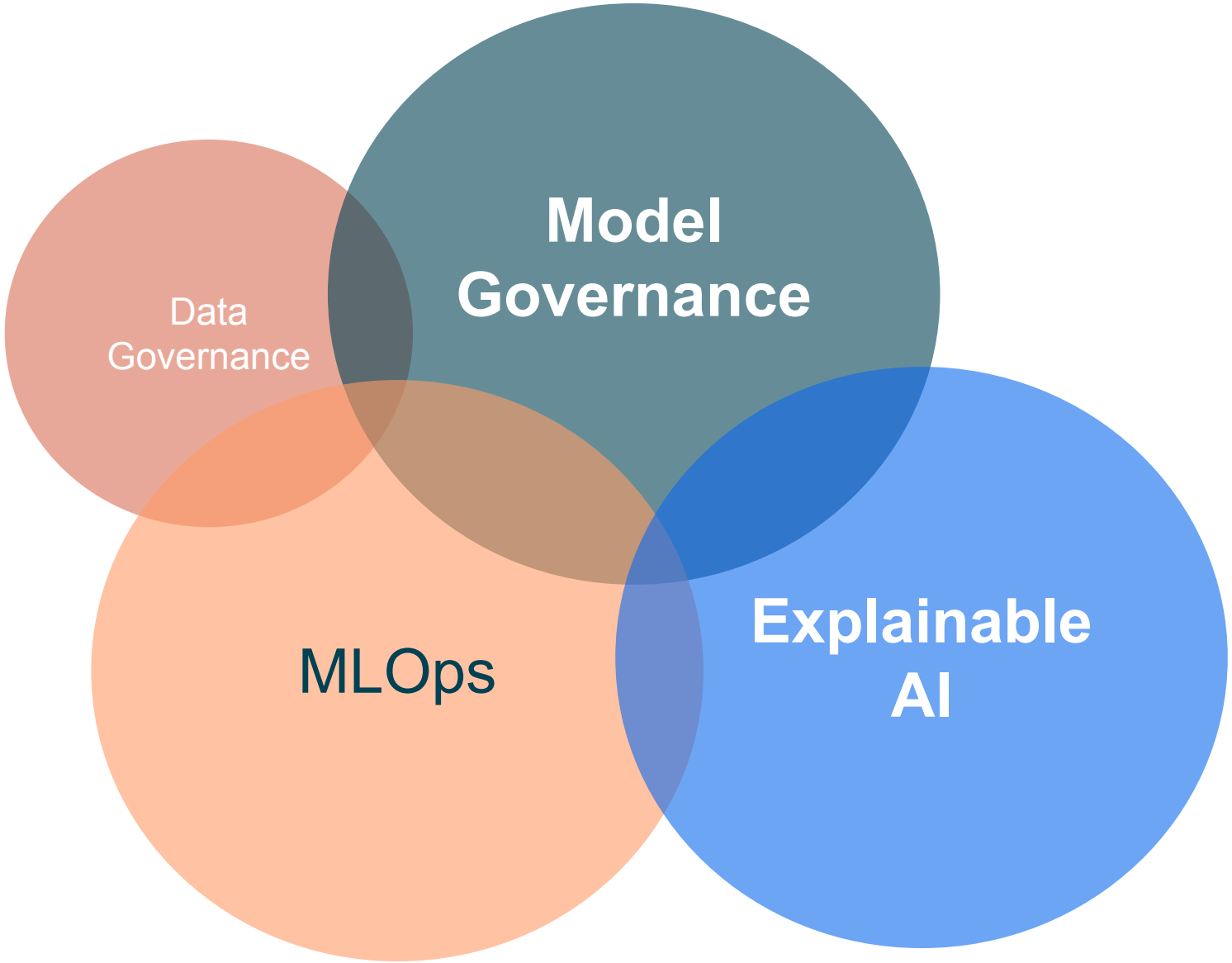
Amazon Shuts Down AI Hiring Tool for Being Sexist

Amazon created an artificial intelligence tool to help with recruiting, but it showed gender bias.

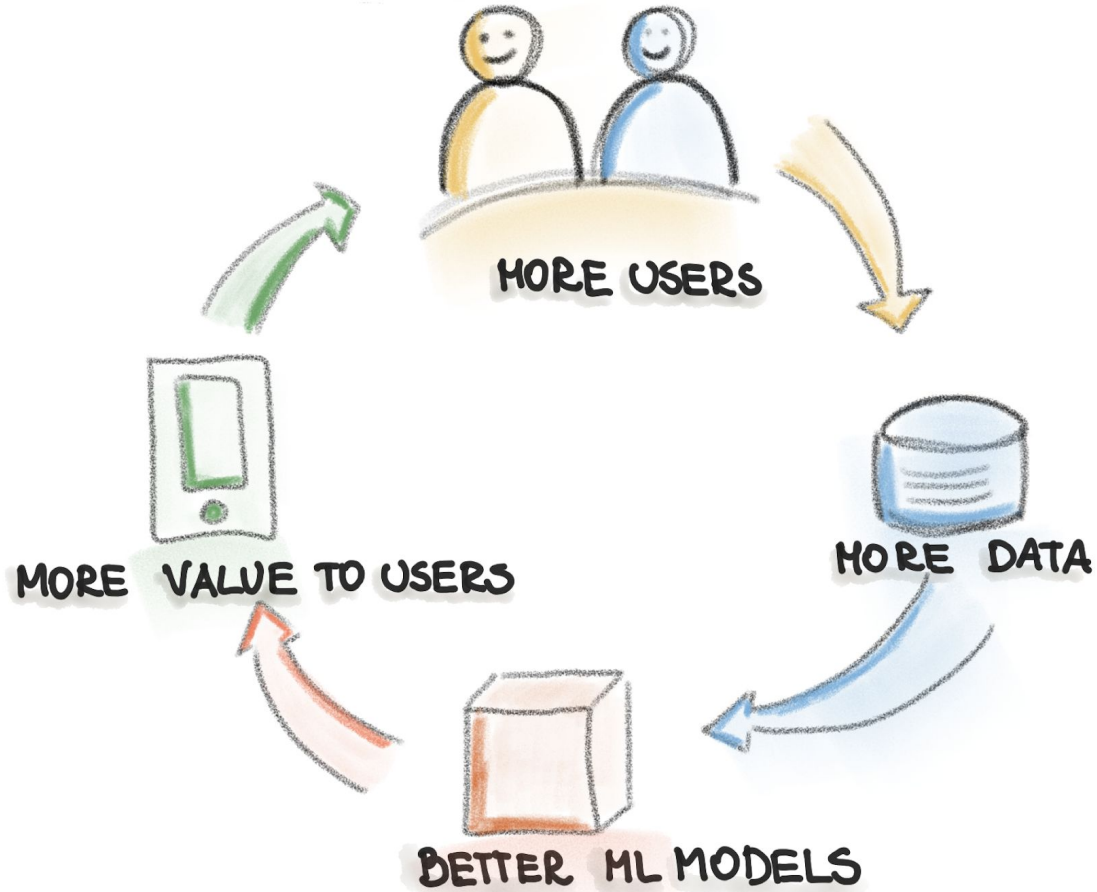
Automated Filtering of Job Applications



What does it take to successfully operate ML software in the long term?


Model
Governance
Data
Governance
MLOps
Explainable
AI

THE FLYWHEEL OF ML SYSTEMS



@visenger

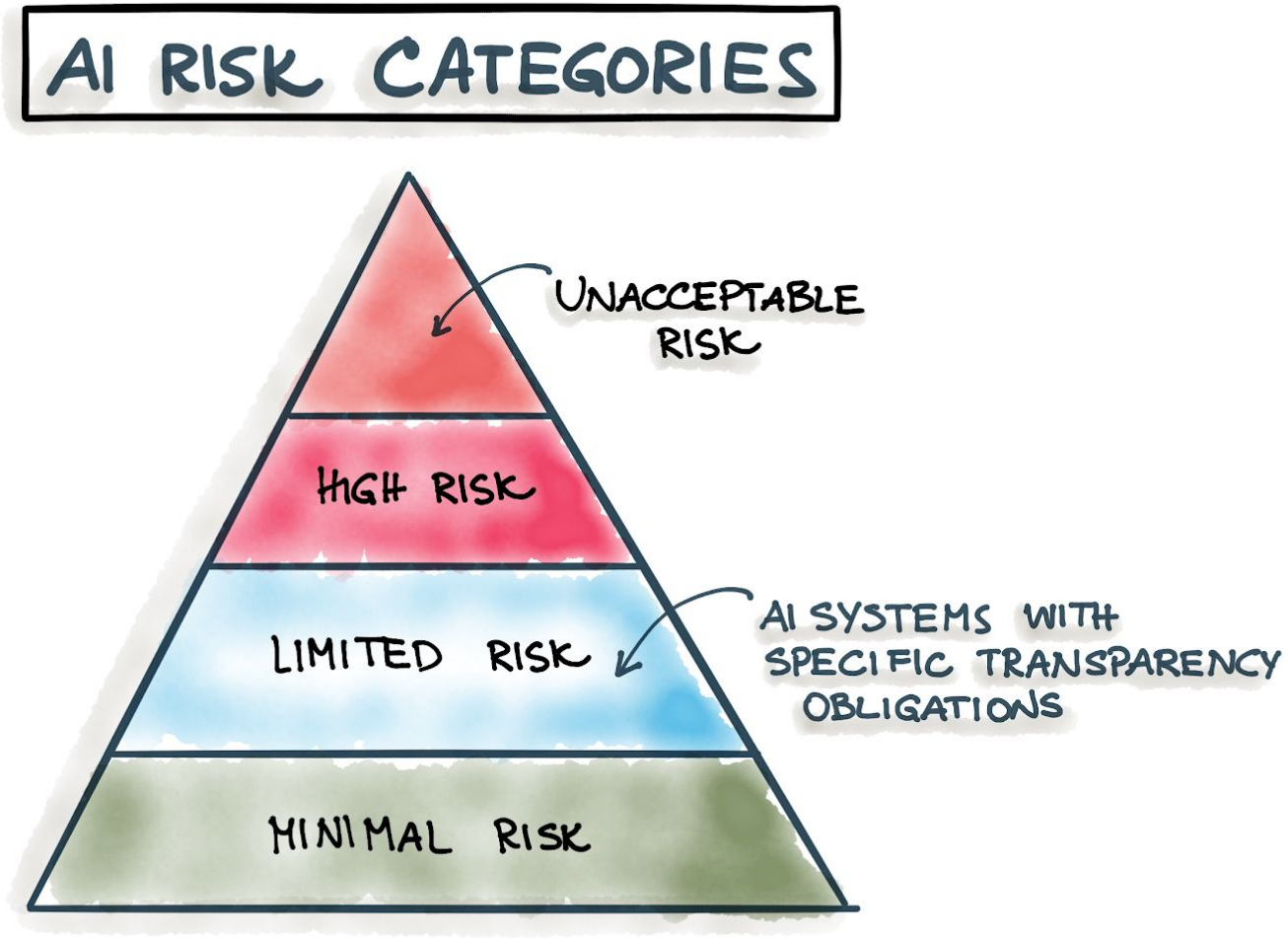
What is Model Governance?

Processes used by organizations to …

- implement policies
- track model activities and results
- manage risks

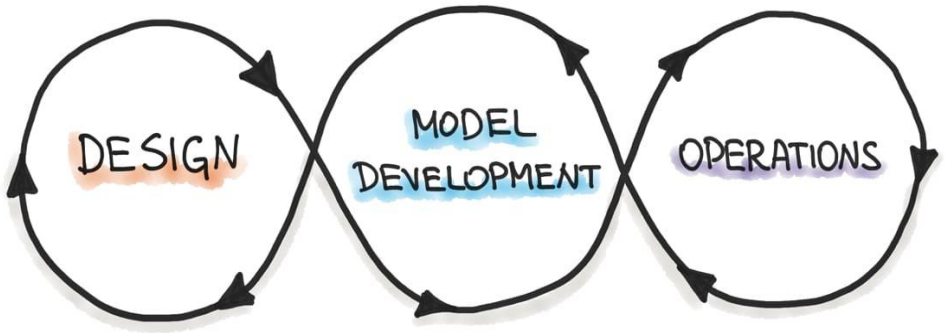
Existing Regulations

+




Business Risk
Reputation
Legal Concerns
Economic
Damage

Risks and Legal Requirements – an abstract problem?

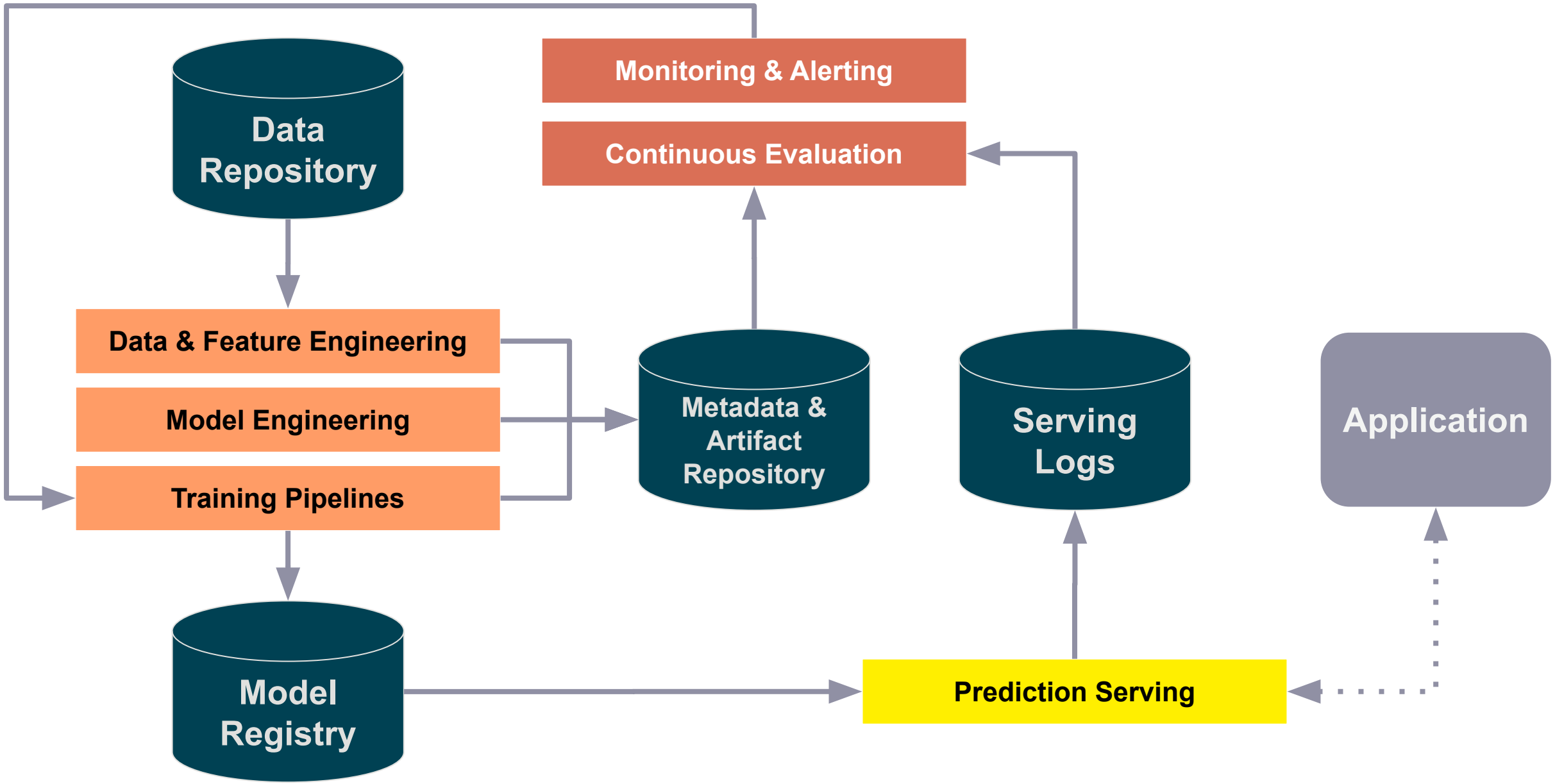


+

Model Governance

Model Governance is a problem that can be solved technically

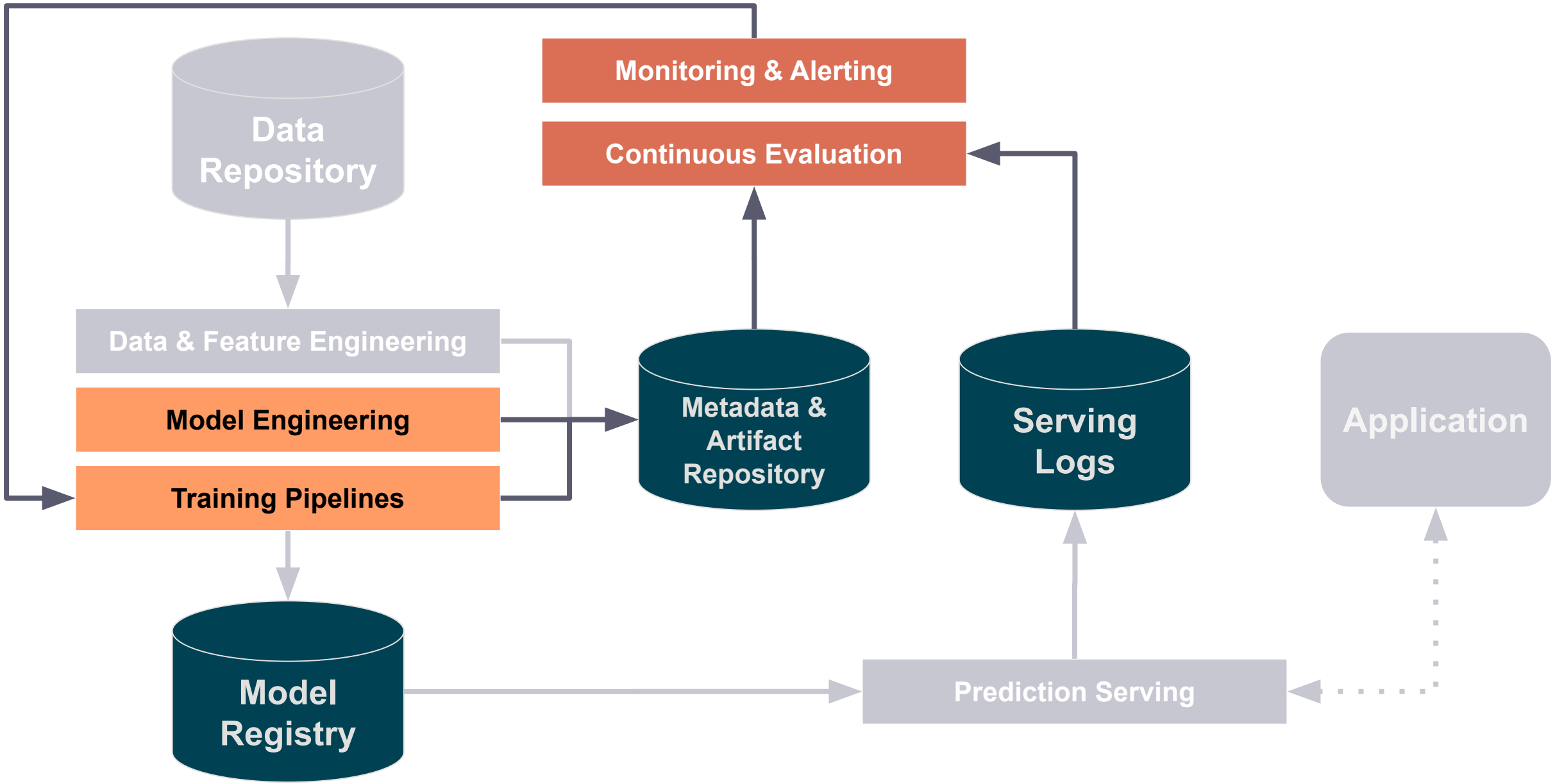
So how can MLOps infrastructure help to manage risks and comply with legal requirements?


Monitoring & Alerting
Data Repository
Continuous Evaluation
Data & Feature Engineering
Model Engineering
Training Pipelines
Metadata & Artifact Repository
Serving Logs
Application
Model Registry
Prediction Serving

Model Performance & Management

Ensure high model performance through

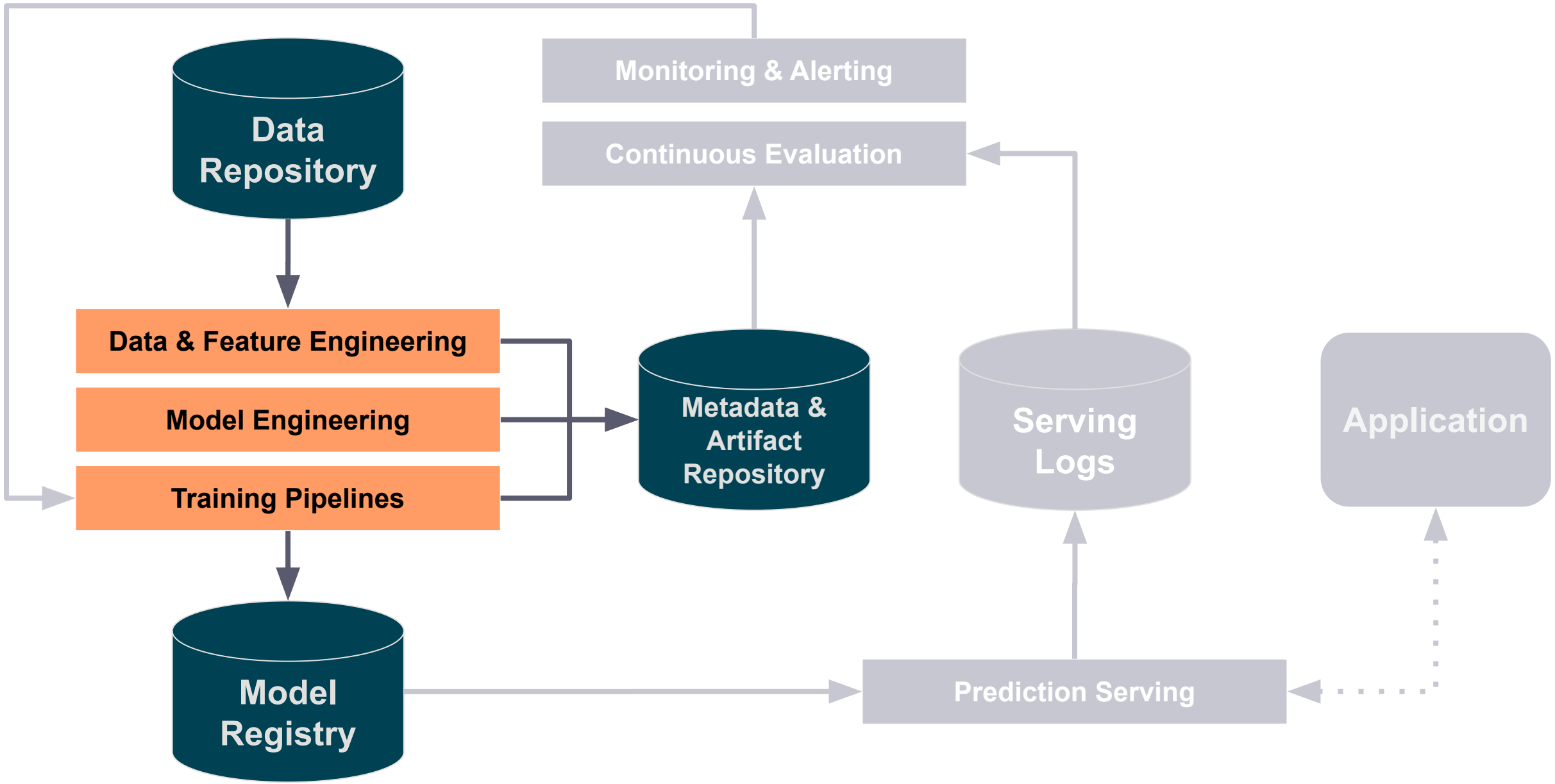
- collection of model metrics
- monitoring of model performance
- alert functions
- enabling quick re-training with fresh data
- enabling changes to the training procedures


Monitoring & Alerting
Data Repository
Continuous Evaluation
Data & Feature Engineering
Model Engineering
Training Pipelines
Metadata & Artifact Repository
Serving Logs
Application
Model Registry
Prediction Serving

Reproducibility & Documentation

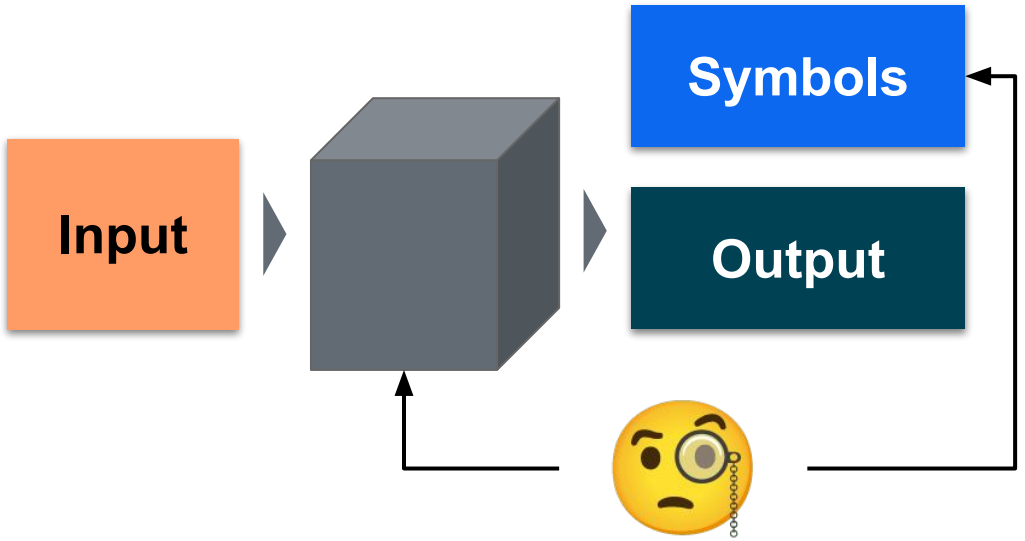
Ensure reproducibility through

- versioning of datasets
- versioning of data/feature engineering and training code
- versioning of models
- metadata and artifact tracking

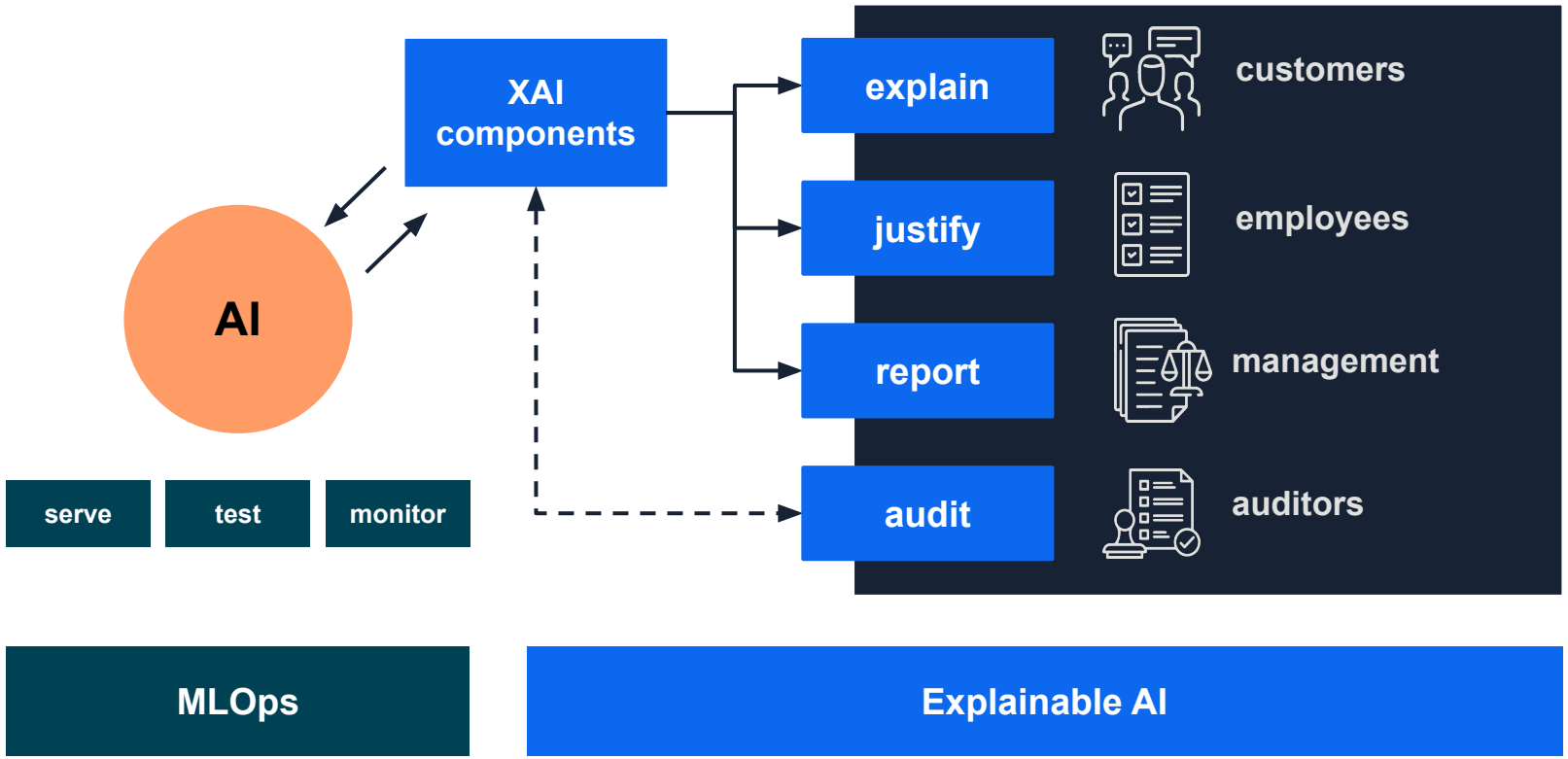

Monitoring & Alerting
Data Repository
Continuous Evaluation
Data & Feature Engineering
Model Engineering
Training Pipelines
Metadata & Artifact Repository
Serving Logs
Application
Model Registry
Prediction Serving

Interpretability

Understand your models' behavior on a technical level.



Explainability



NIST: “Four Principles of Explainable AI”

Explanation

Delivers or contains accompanying evidence or reason(s) for outputs and/or processes

Meaningful

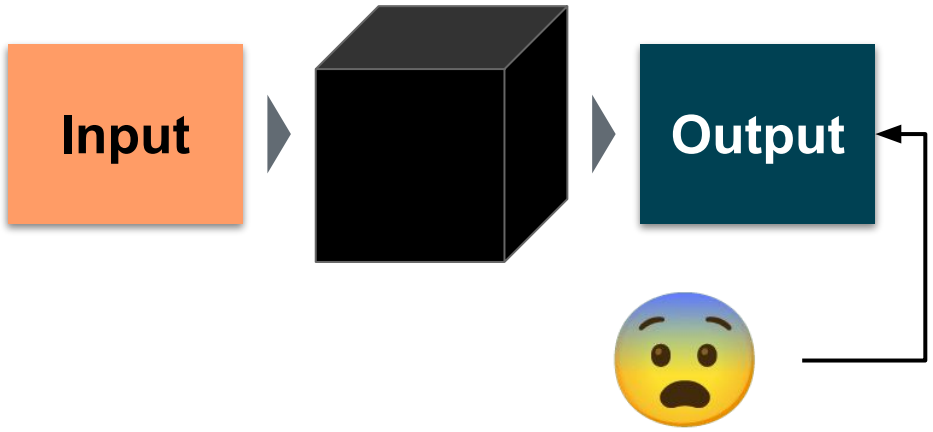
System provides explanations that are understandable to the intended consumer(s)

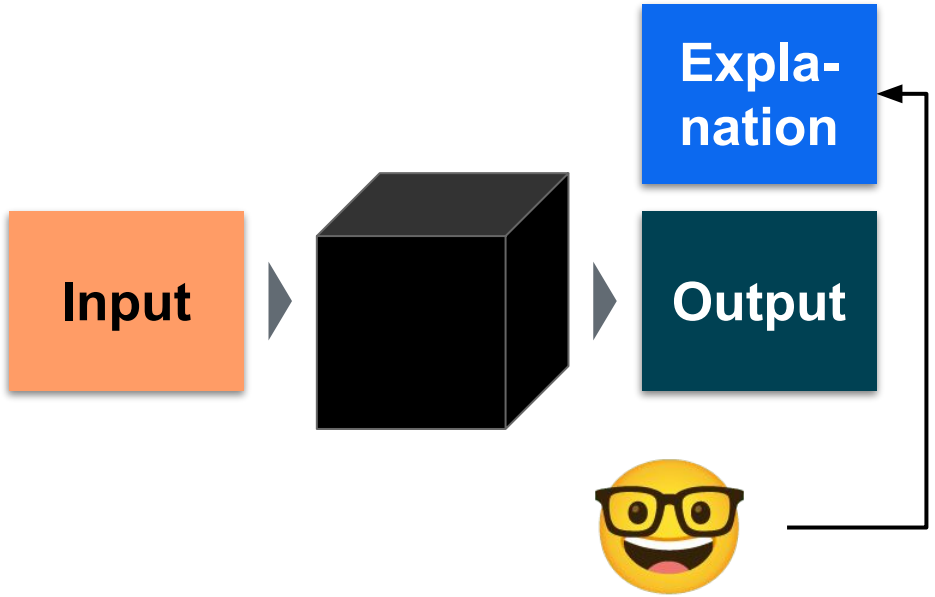
Explanation Accuracy

Explanation correctly reflects the reason for generating the output and/or accurately reflects the system’s process

Knowledge Limits

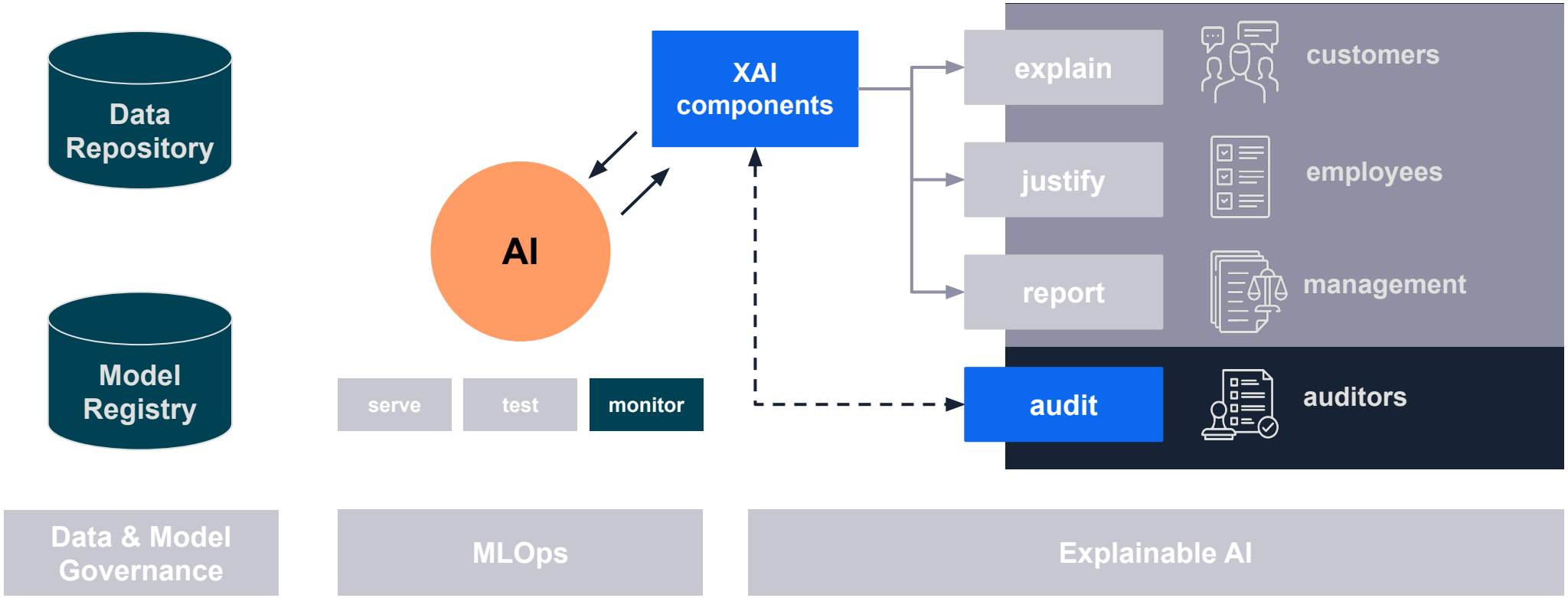
System only operates under conditions for which it was designed and when it reaches the sufficient confidence in its ouput





- **Why was this candidate selected?** → **Anchors:** Which features does the recommendation rely upon?
- **Did "the AI" consider all relevant facts?** → **Feature Importance:** Which features contributed positively/negatively?
- **How certain is "the AI"?** → **Counterfactuals:** What would need to change to reverse the recommendation?

Auditability



Where do we go from here?

Get in touch!

ISABEL BÄR
@isabel_baer

isabel.baer@innoq.com

KILIAN KLUGE
@kilian_kluge

kilian.kluge@inlinity.ai

INNOQ

inlinity
trusting artificial intelligence