Red Hat OpenShift AI

Realizing value from AI/ML

Increasing velocity and consistency through MLOps

Ramón Gordillo
Principal Solution Architect,
Red Hat

AI is becoming a part of our everyday lives

Chat GPT

Stable Diffusion

watsonx Code Assistant

Llama 2

Meta AI

DALL·E 2

Gemini

GitHub Copilot

MISTRAL AI_

Every business has a use for AI/ML

Healthcare

- Increased clinical efficiency
- Faster/better diagnosis
- Improved outcomes

Financial services

- More personalized services
- Improved risk analysis
- Reduced fraud
- Better predictions

Telcos

- Better customer insights/experiences
- Optimized network performance & operations
- Improved threat detection

Insurance

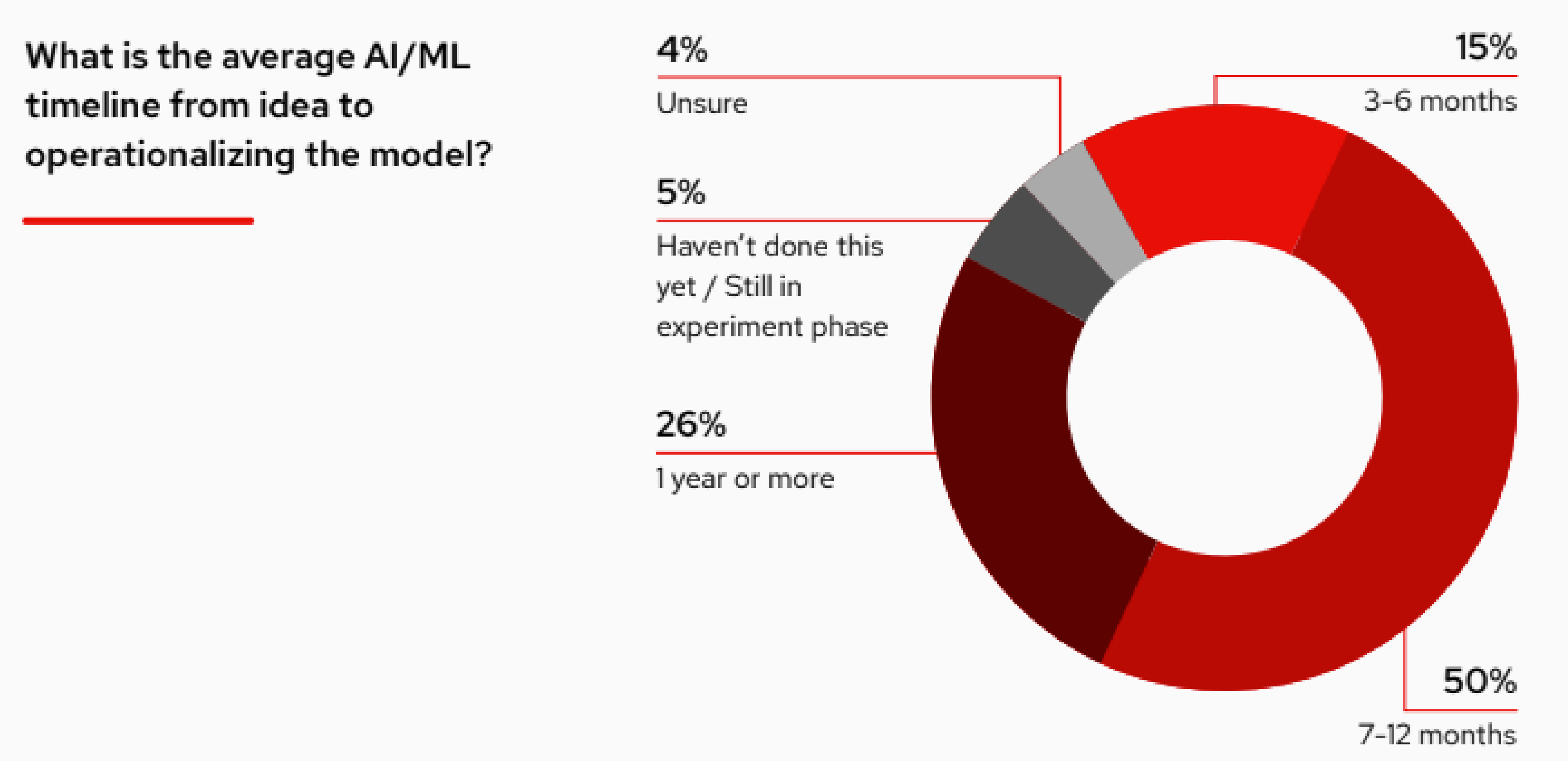
- Automated claims processing and handling
- Usage-based insurance services

Automotive

- Autonomous driving
- Predictive maintenance
- Improved supply chains

Operationalizing AI is still a challenging process

Half of respondents (50%) say their average AI/ML timeline from idea to operationalizing the model is 7-12 months.



Source: Gartner Peer Insights, Open Source AI for Enterprise survey, 2023

The reality of enterprise IT environments

Infrastructure

50%

of apps being run by enterprises are on hybrid cloud.[1]

41%

claim that IT Agility is one of main reasons to run apps on hybrid cloud[1]

Bare metal
Virtualization

Private cloud
Edge
Public cloud

Applications

AI/ML
Analytics
Serverless

Cloud-native services and microservices

Java™
.Net
ISV

Processes and people

Developer tools

Methods and practices

Culture and policy

Source: Red Hat detail. "2024 Global Tech Trends," Feb. 2024.

The reality of enterprise IT environments

Infrastructure

50%

of apps being run by enterprises are on hybrid cloud.[1]

41%

claim that IT Agility is one of main reasons to run apps on hybrid cloud[1]

Bare metal
Virtualization

Private cloud
Edge
Public cloud

Applications

AI/ML
Analytics
Serverless

Cloud-native services and microservices

Java™
.Net
ISV

Processes and people

Developer tools

Methods and practices

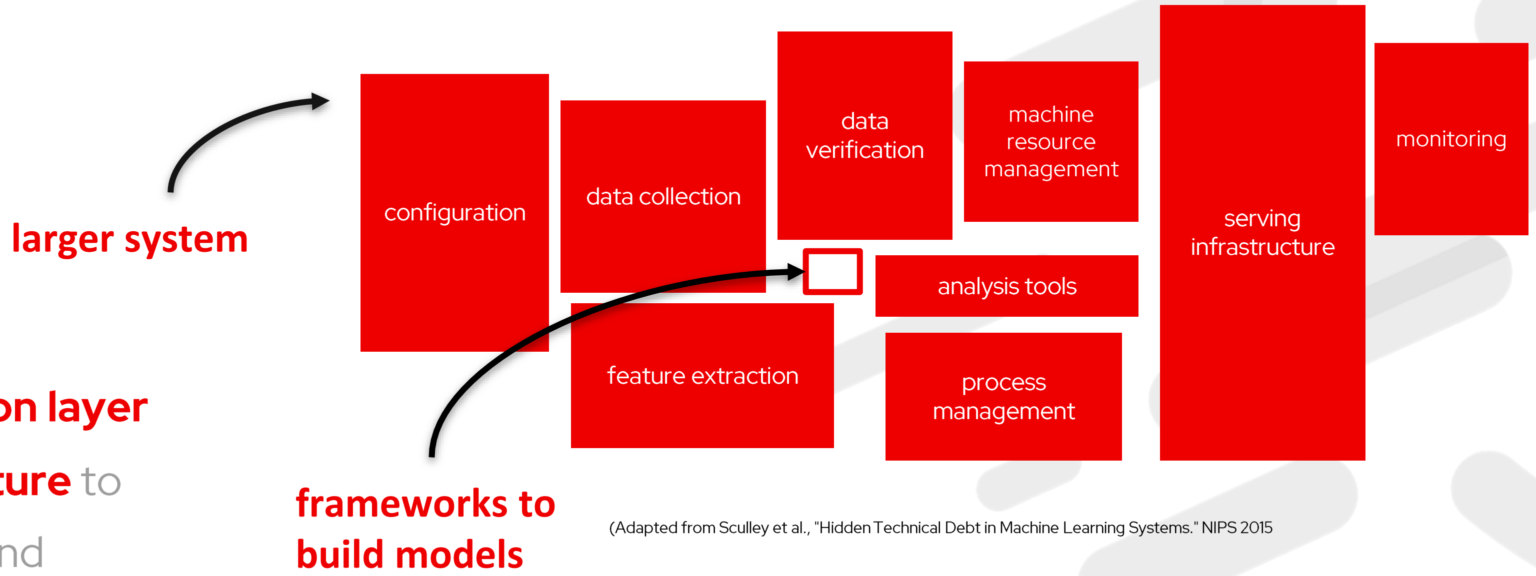
Culture and policy

Source: Red Hat detail. "2024 Global Tech Trends," Feb. 2024.

Complexities of operationalizing models

"a consistent application platform for the management of existing, modernized, and cloud-native applications that runs on any cloud."

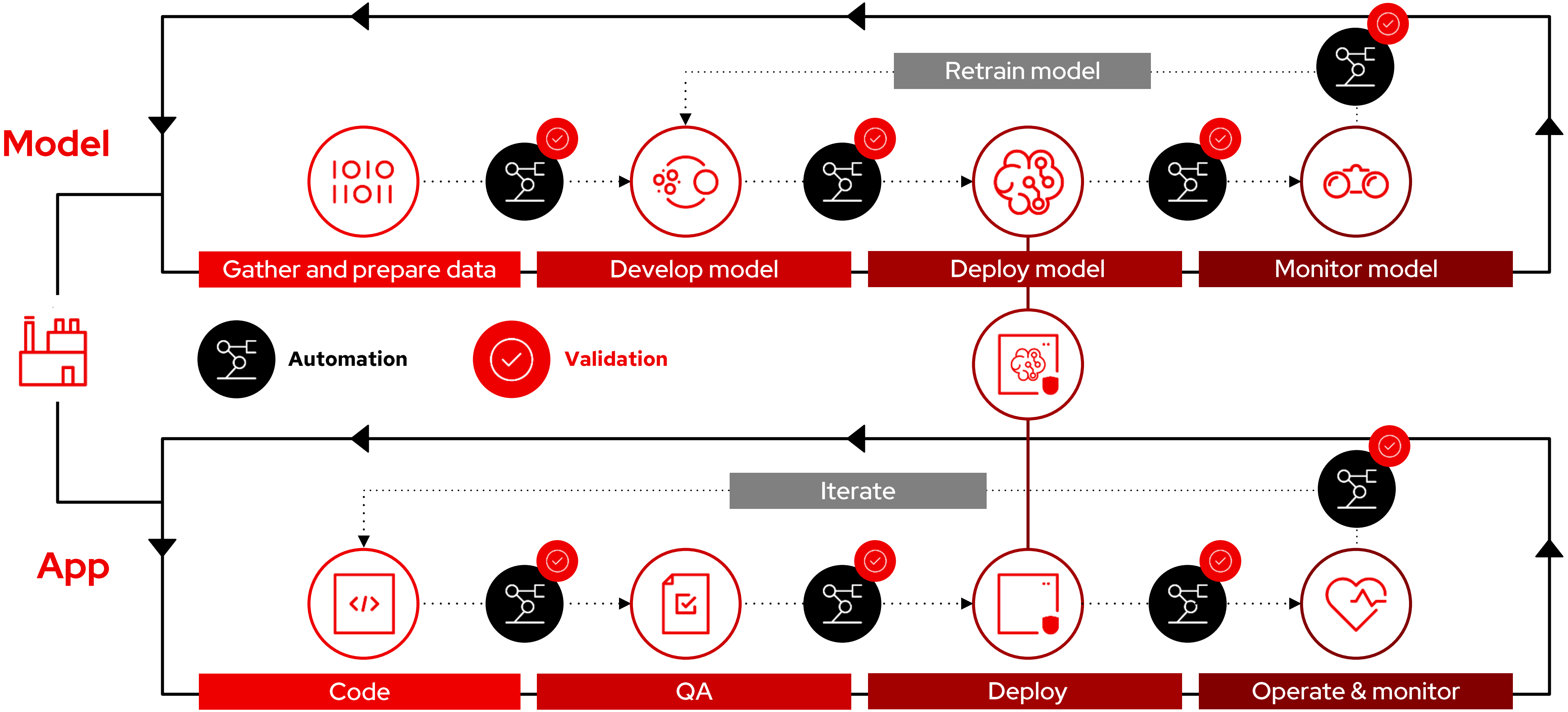
"a common abstraction layer across any infrastructure to give both developers and operations teams commonality in how applications are packaged, deployed, and managed."



(Adapted from Sculley et al., "Hidden Technical Debt in Machine Learning Systems." NIPS 2015

Source: https://www.redhat.com/en/technologies/cloud-computing/openshift

Lifecycle for operationalizing models



Challenges

Training, Serving & Monitoring

Workload management

Training jobs require variable compute resource requirements with access to accelerators. Serving requires the ability to scale on demand based on inference requests

Orchestration

Consistency in repeatable and secure pipelines for data ingestion and processing through to model build and staging. Deployment across multiple platforms often leads to varying methodologies.

Platform and vendor complexity

Machine learning models typically optimized for specific hardware platforms which vary based on each model and use case. Adopting emerging technologies introduces risk.

Fleet management

Insights into model performance and quality are inconsistent and varied across the enterprise. Lack of model transparency increases risk within deployments.

Challenges

Model Lifecycle

Rollout coordination

Friction in handoffs between data science, application developer, and devops teams leads to high quality experiments never making it into production.

Software supply chain

Multiple orchestration platforms and bespoke build processes introduce risk into the software supply chain through lack of auditability, traceability, and transparency.

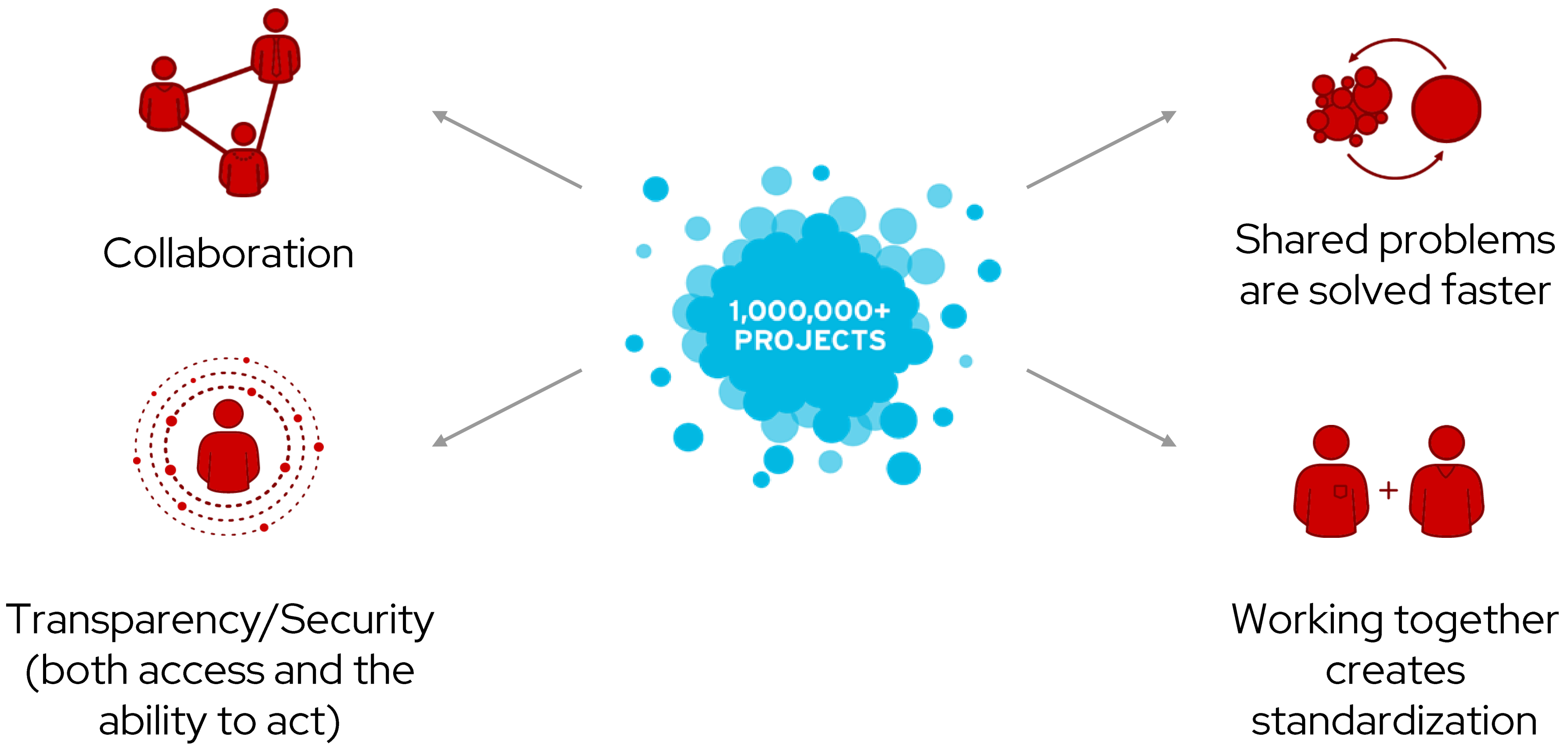
Agility

The ability to maximize value out of AI/ML is driven by more and more experiment iterations. Manual process and interventions reduce overall volume of runs.

Loss of confidence

Repeated failures in model rollout leads to lack of confidence in AI/ML which limits the overall potential of the business.

AI/ML innovation driven by open source



Red Hat Enterprise Linux

- Secure, composable, and compliant platform with a consistent administration experience across platforms
- Support for core AI/ML libraries, hardware, and accelerators

Red Hat OpenShift

- Simplified deployment, scaling, and management of AI/ML training and serving
- Cloud-hosted or self-managed options across data center, public cloud, and edge
- Extend DevOps to the entire ML lifecycle, and enable collaboration across teams

Red Hat Ansible Automation Platform

- Build, provision, and manage applications and infrastructure across platforms
- Automate application deployments, installations, upgrades, and day-to-day management repeatable and reliable

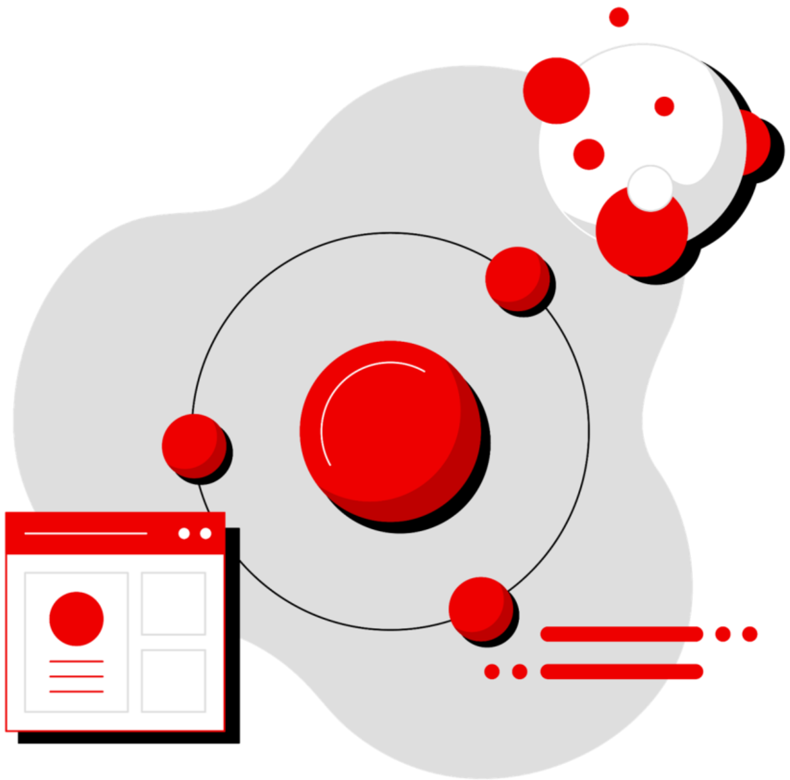
Red Hat OpenShift AI

Red Hat OpenShift AI provides an integrated platform for building, training, tuning, deploying and monitoring AI-enabled applications, predictive and foundation models securely and at scale across hybrid-cloud environments.

Built on top of Red Hat OpenShift delivers a consistent, streamlined and automated experience to help organizations rapidly innovate and deliver AI-enabled apps into production.

Use the power of enterprise-ready open source

Set yourself and your teams up for success with a solid foundation



The AI/ML ecosystem is complex

- Technologies are rapidly evolving
- Vendor landscape is constantly changing
- No single vendor can provide everything you need
- Organizations need a supported, secure enterprise version of open source tools and technologies for AI/ML
- Success with AI/ML starts with having a solid foundation to build upon