José Ángel de Bustos

Senior Specialist Solution Architect,

Red Hat

Red Hat | intel

Red Hat OpenShift AI

Realizing value from AI/ML

Increasing velocity and consistency through MLOps

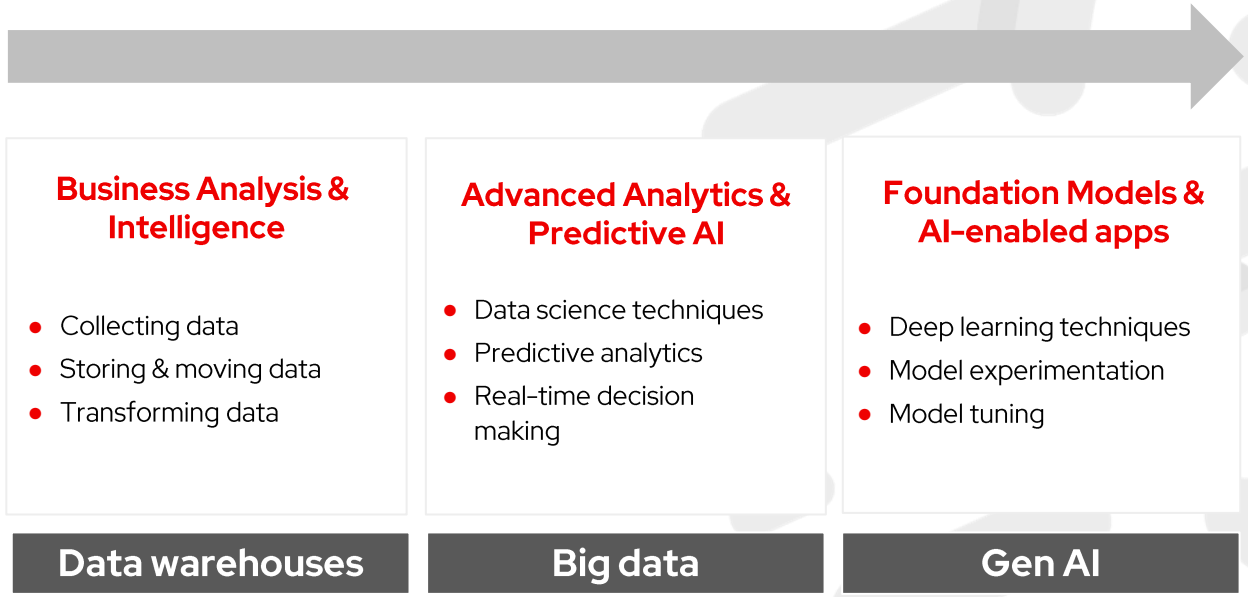
José Ángel de Bustos Pérez
Senior Specialist Solution Architect

Red Hat | intel

AI has undergone significant evolution

The evolution of AI: from Business Intelligence to Generative AI

- Predictive AI runs businesses today
- Foundation models provide a shortcut for realizing the value of AI



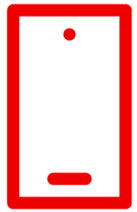
Every business has a use for AI/ML

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

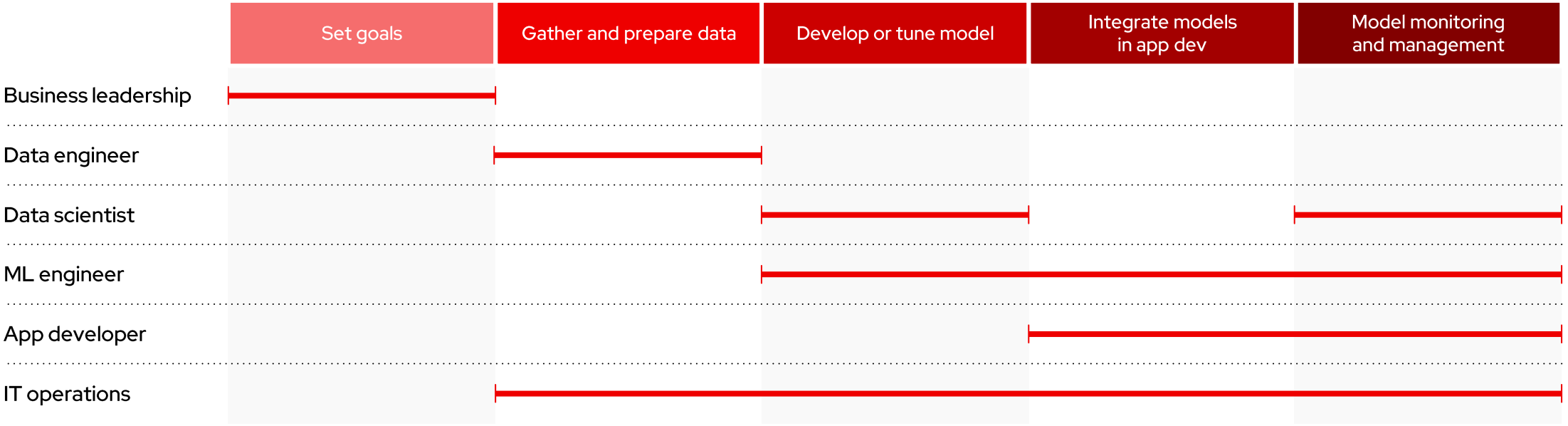
- Automated claims processing and handling
- Usage-based insurance services

Automotive

- Autonomous driving
- Predictive maintenance
- Improved supply chains

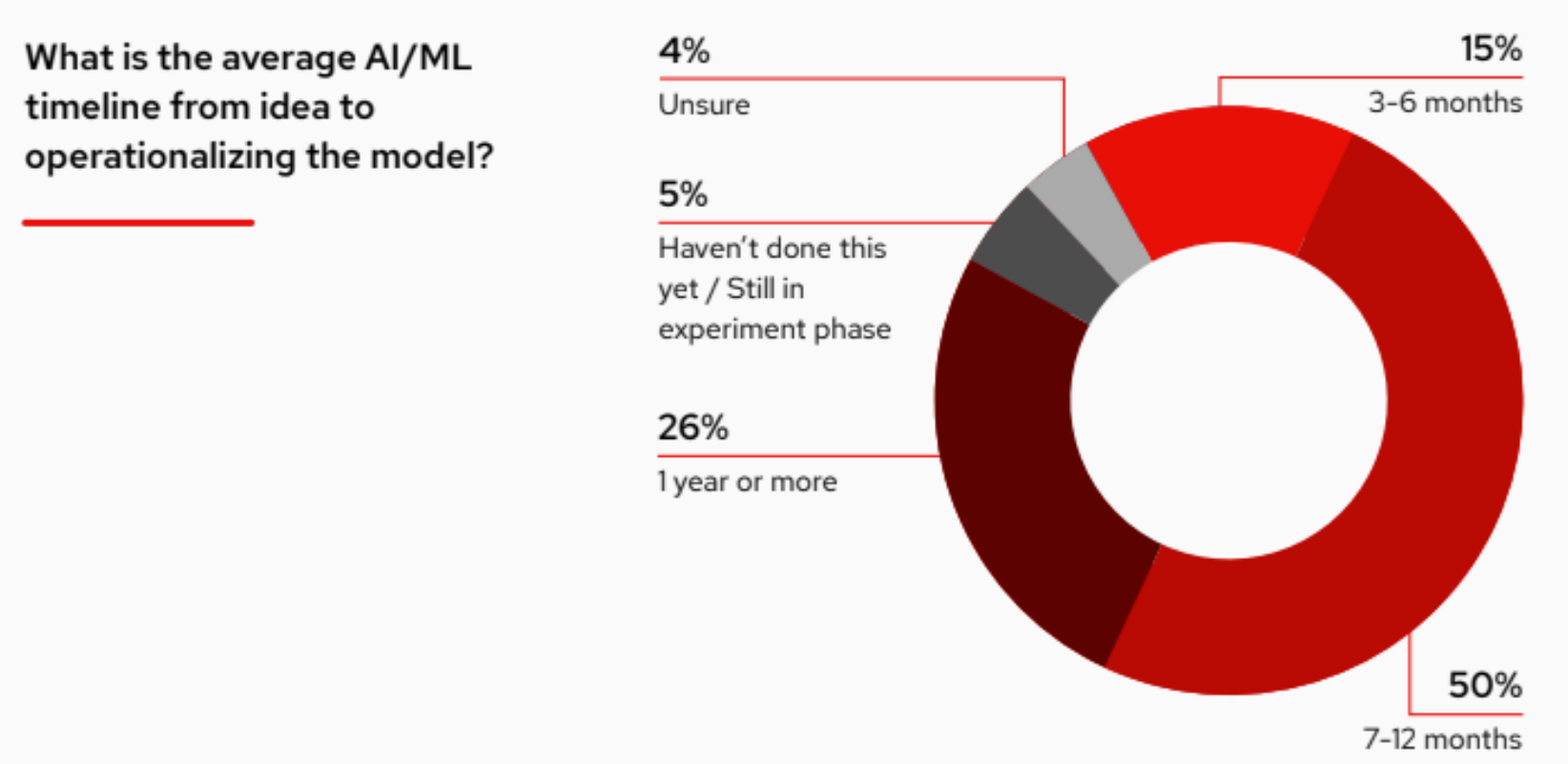
Operationalizing AI/ML requires collaboration

Every member of your team plays a critical role in a complex process



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.

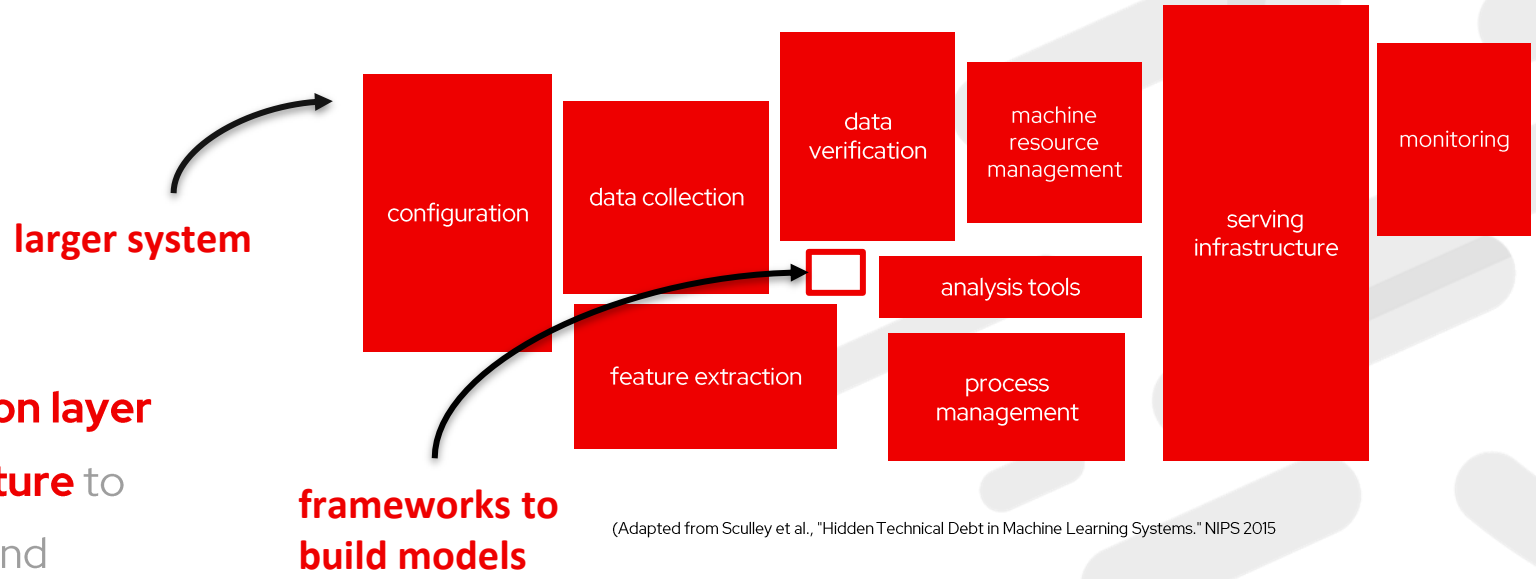


Source: Gartner Peer Insights, Open Source AI for Enterprise survey, 2023

Complexities of operationalizing models

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

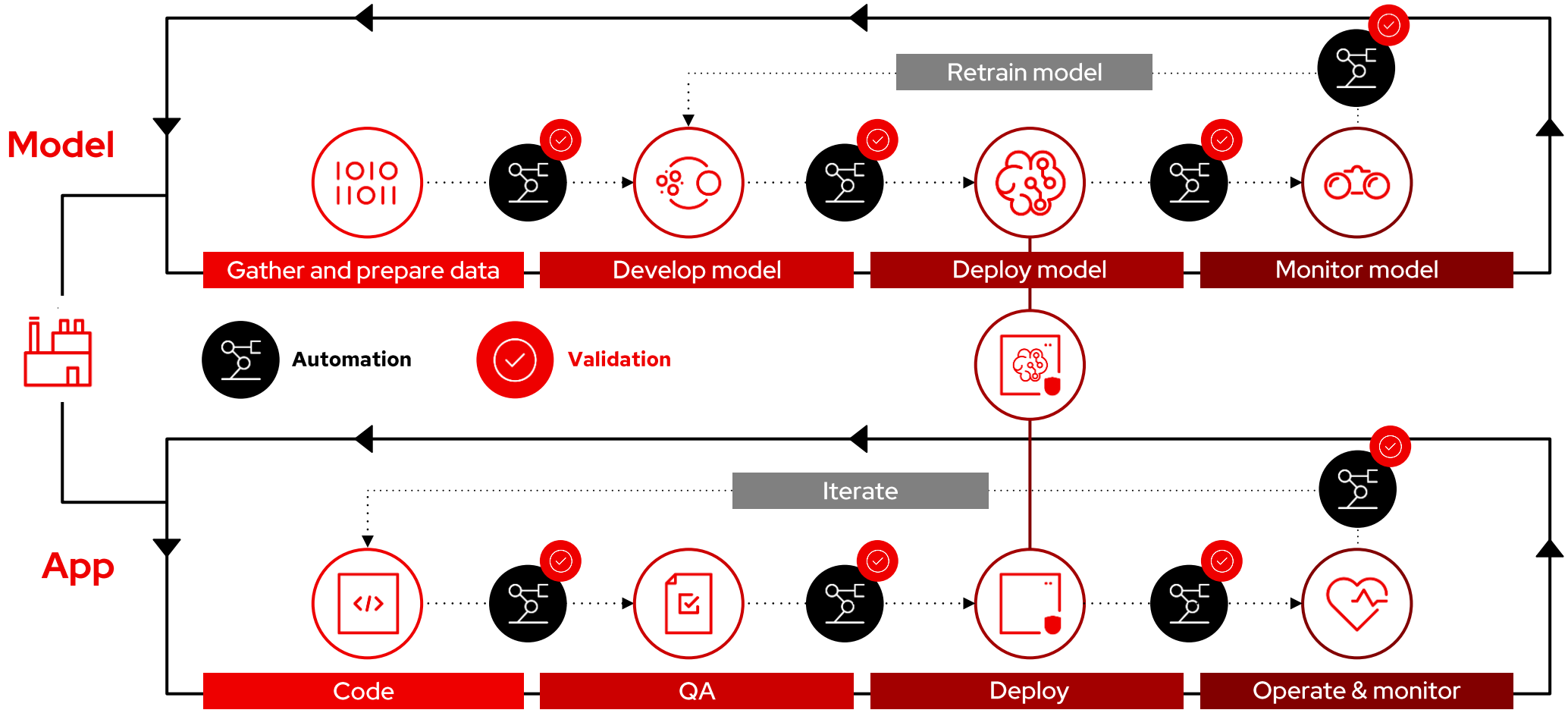
"**a common abstraction layer across any infrastructure** to give both developers and operations teams commonality in how applications are packaged, deployed, and managed."



(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

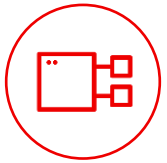
Training jobs require variable compute resource requirements with access to accelerators. Serving requires the ability to scale on demand based on inference requests

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.

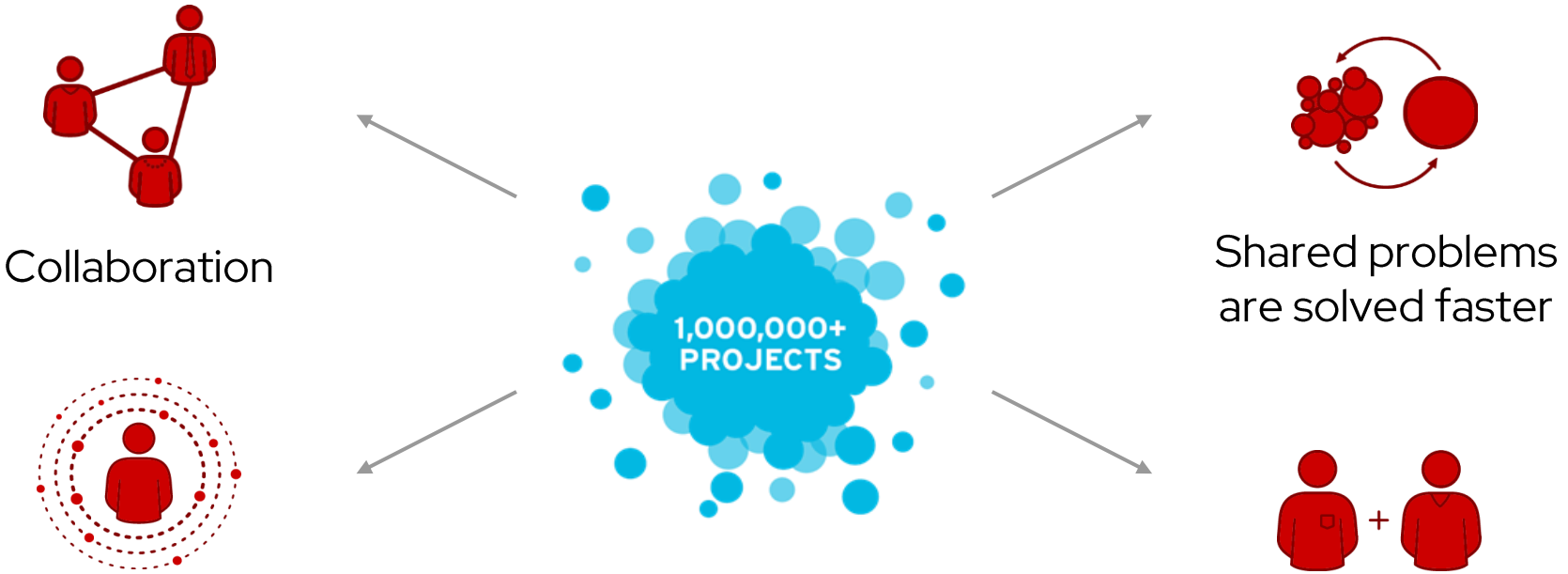
Agility

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



Collaboration

Shared problems are solved faster

Transparency/Security (both access and the ability to act)

Working together creates standardization

Red Hat's AI portfolio

Trust | **Choice** | **Consistency**

Models

RHEL AI

Base Model | Alignment Tuning | Methodology & Tools | Platform Optimization & Acceleration

AI platform

OpenShift AI

Development | Serving | Monitoring & Lifecycle | Resource Management

AI enabled portfolio

Lightspeed portfolio

Usability & Adoption | Guidance | Virtual Assistant | Code Generation

AI enabled apps

App & Developer services

Model Evaluation and Testing | App Connectivity | Secure Supply Chain

Open Hybrid Cloud Platforms

Red Hat Enterprise Linux | Red Hat OpenShift | Red Hat Ansible Platform

Acceleration | Performance | Scale | Automation | Observability | Security

Partner Ecosystem

Hardware | Software | Accelerators | Models | Delivery

Red Hat Enterprise Linux AI

Foundation Model Platform

Seamlessly develop, test and run best of breed, open source Granite generative AI models to power your enterprise applications.

The model is the new platform.

Open Granite models

Highly performant, fully open source, collaboratively developed Granite language and code models from the community, fully supported & indemnified by Red Hat and IBM.

InstructLab model alignment

Scalable, cost-effective solution for enhancing LLM capabilities efficiently for a wide range of applications, making knowledge & skills contributions accessible to a wide range of users

Optimized bootable model runtime instances

Granite models & InstructLab tooling packaged as a bootable RHEL image, including Pytorch/runtime libraries, hardware optimized inference for Nvidia, Intel and AMD that can run anywhere and provides onramp to OpenShift AI for scale and lifecycle & watsonx for agent integration and governance.

Enterprise support, lifecycle & indemnification

Trusted enterprise platform, 24x7 production support, extended model lifecycle and model IP indemnification

Red Hat | intel

Introducing: **InstructLab**

Open source community project for GenAI model development

Developers submit skills as pull requests

InstructLab

Requests are triaged by community reviewers

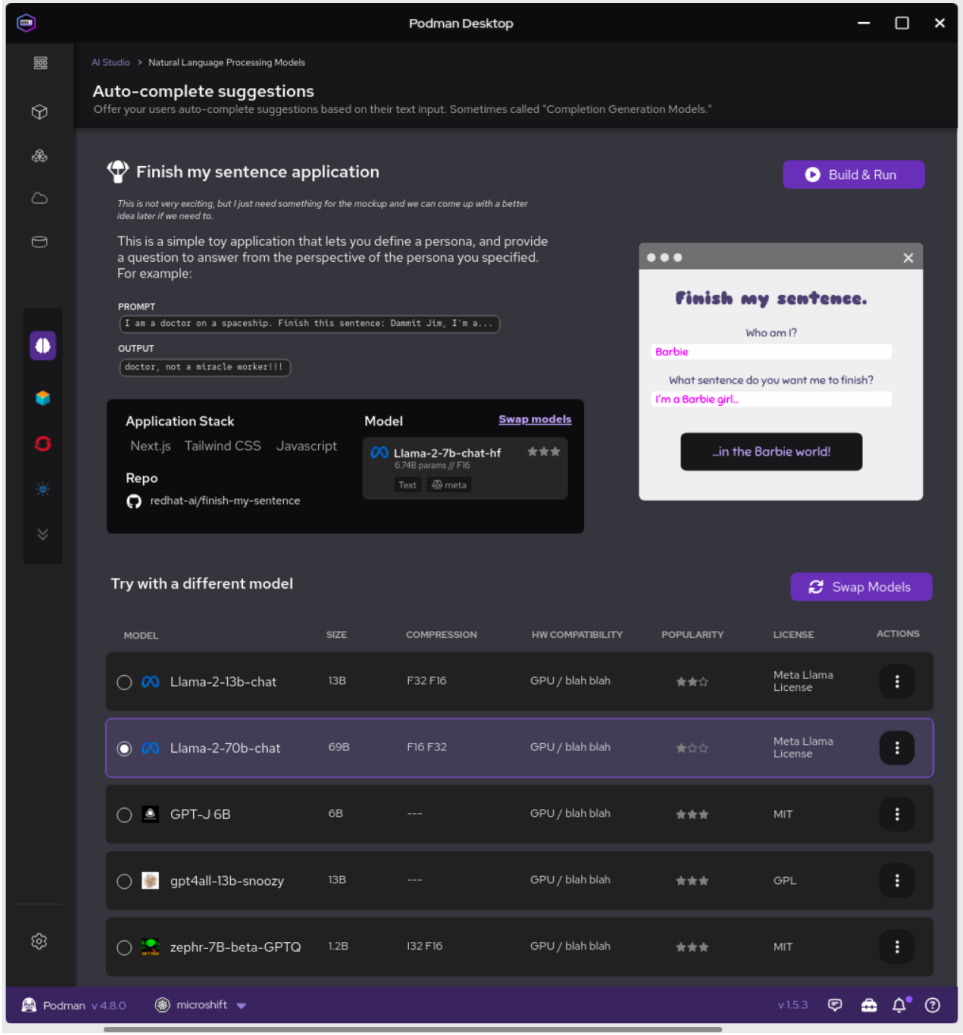
Triaged skills are used to to generate synthetic data and train the community model

Trained model is released periodically

Red Hat | intel.

Introducing: **Podman AI Lab**

Simple developer access to local containers and AI



- Run and code against local models quickly on your laptop (Mac, Windows, & Linux)
- Accelerate AI adoption, by easing concerns around data access, data privacy, & security
- Local developer workflow for model fine-tuning
- Path to production - Easy to package and deploy apps and models direct to OpenShift AI all they way down to bare metal
- Simple access to Red Hat developer subscriptions

Red Hat | intel.

Tech Preview

Introducing: **image mode for Red Hat Enterprise Linux**

Making operating systems as simple as containers

```
FROM quay.io/rhel9/rhel-9-
bootc:latest

RUN dnf -y install [software]
[dependencies] && dnf clean all

ADD [application]
ADD [configuration files]

RUN [config scripts]
```

Allows teams, infrastructure, and ecosystems to converge on a single container-native workflow to manage everything from their applications to the underlying operating system. build, test and distribute the operating system as if it was any other container.

- **All customers** benefit from the simplicity and portability across their traditional and hybrid cloud environments
- **DevOps teams** finally have a native OS with their favorite CI/CD & GitOps workflows solving last-mile problems
- **Security teams** will be delighted by the cryptographic validation, transparent provenance, and tool consolidation
- **Partners** will love how easy it is to build and distribute on a trusted base operating system

Red Hat | intel

Tech Preview

Image mode for RHEL

A container-native workflow for the life cycle of a system

```
FROM rhel9/rhel-bootc:latest

RUN dnf install -y [software]
[dependencies] && dnf clean all

ADD [application]
ADD [configuration files]

RUN [config scripts]
```

Build

A *bootc* base image & container file is all that's needed to describe a system, applications, and dependencies. Use your existing container tools or pipelines to quickly create and test images.

Deploy

Easily convert to a VM/cloud image or deploy on bare metal using RHEL's installer. The container image includes full hardware drivers, but not cloud agents by default.

Manage

Designed for modern GitOps & CI/CD driven environments. Systems will auto-update from the container registry by default. More advanced control and automation is available via custom rollouts (e.g. Ansible). Intelligence via Insights and on-prem content curation via Satellite are planned for the future.

Red Hat OpenShift AI

Model development

Interactive, collaborative UI for exploratory data science, and model training, tuning and serving

Model serving

Model serving routing for deploying models to production environments

Model monitoring

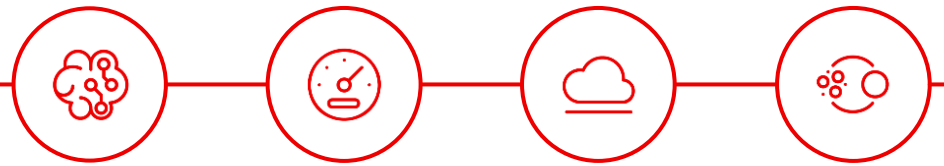
Centralized monitoring for tracking models performance and accuracy

Data & model pipelines

Visual editor for creating and automating data science pipelines

Distributed workloads

Seamless experience for efficient data processing, model training, tuning and serving



Announcing: **Red Hat Lightspeed across Red Hat platforms**

Intelligent, natural language GenAI processing capabilities designed to extend existing IT skills

OpenShift Lightspeed will be available in Technology Preview later in 2024

RHEL Lightspeed is currently in the planning stages and availability will be announced

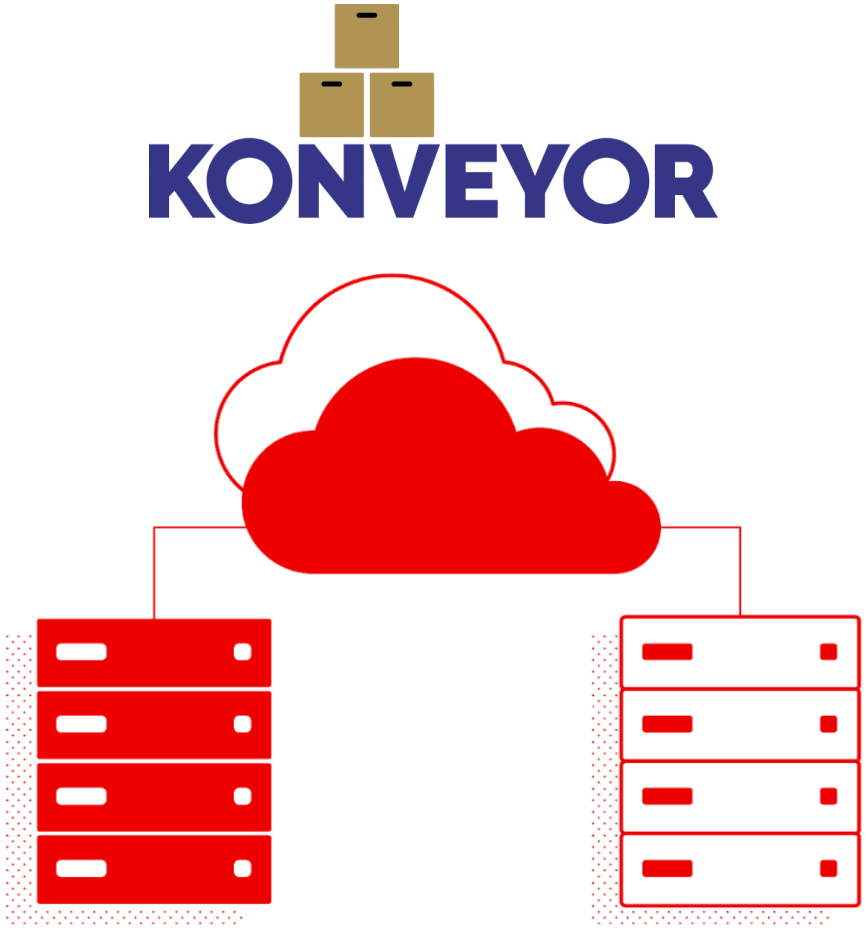
Improve the productivity and efficiency of ops and developers by integrating AI into cluster administration and the operating system

Simplify enterprise planning and administration, improve performance and enhance security

More easily navigate the complexities of enterprise IT in the hybrid cloud

Red Hat | intel

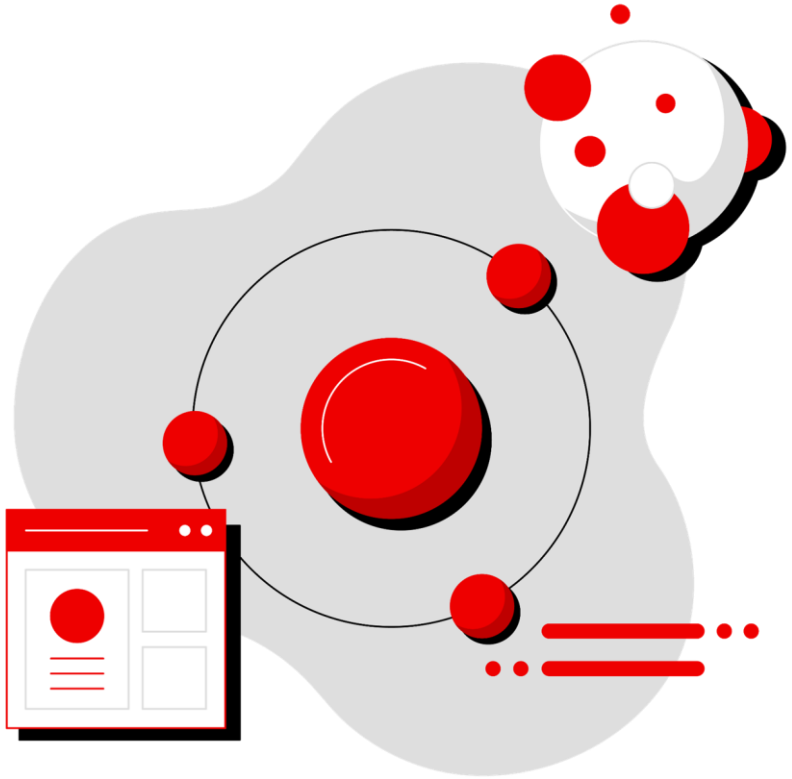
Announcing: **Konveyor GenAI for the Konveyor Community**



- GenAI applied to application modernization efforts
- Workflow-integrated LLMs
- Generated code directly within IDEs
- Successful migrations build strong recommendations
- Roadmap for Red Hat migration toolkit for applications

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