Data Science, AI/ML, Open Source, and MLOps

Data Science Primer for Non-Data Scientists

Jason Dudash

Emerging Technology, North America Office of Technology

@dudashtweets

https://www.linkedin.com/in/jasondudash/

Red Hat

What we'll discuss today

- **Data Science**
  - What, Why
- **Technology**
  - AI & Machine Learning
- **People and Teams**
- **Accelerating Solutions**
  - Open Source
  - MLOps

Data Science

DEFINITION: *data science [dey-tuh-sahy-uhns, dat uh]*

"Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data."

Source: https://www.ibm.com/cloud/learn/data-science-introduction

WHY IS DATA SCIENCE IMPORTANT?

Think of the amount of data industry generates – data is a strategic asset

Government

Smart City
Sensor-based asset monitoring

Manufacturing

Quality assurance

Retail

Digital in-store experience

Health–life science

Patient diagnosis/treatment

Energy

Monitoring and control

Automotive

Autonomous driving
Predictive maintenance

Financial Services

Fraud detection
Risk analysis

Telecommunications

Threat detection

Insurance

Automated claims processing

WHAT CAN DATA SCIENCE DO FOR ME?

Analysis

Detect patterns and trends and find relationships in data

Make Predictions

Make highly accurate guesses as to the likely outcome of a question based on history

Perform Classification

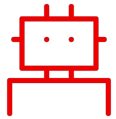
Given some input, group it into a known class of information (e.g. spam/not spam)

Form and validate hypothesis

Run experiments using ML techniques and algorithms, chart data, etc.

Gain insights from data

Building graphs, creating models, predicting, classifying, and analyzing data to understand it



Automate decision making

Integrate models with new and legacy software systems

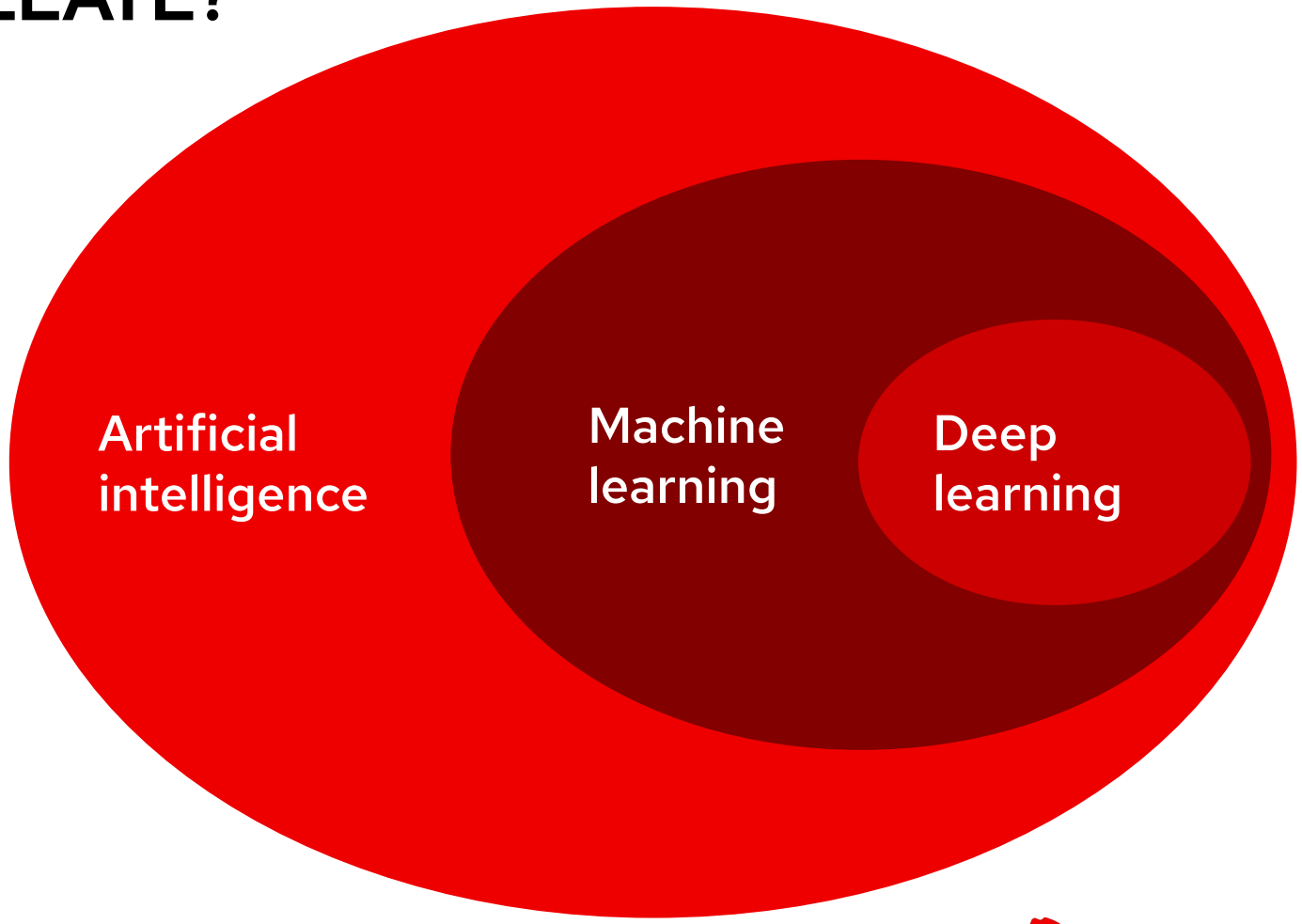
Create software services that can be integrated with traditional software to build intelligent systems

Technology

Red Hat

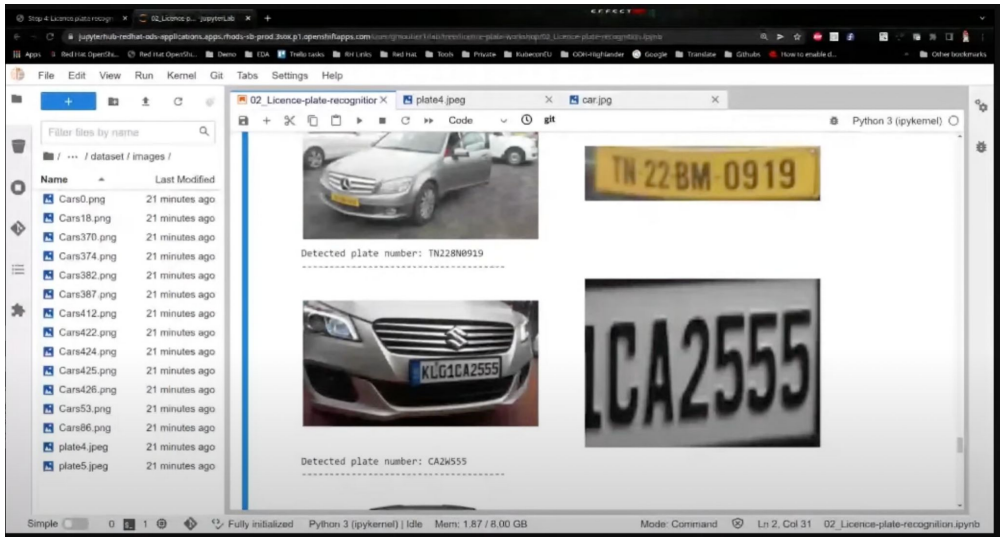
HOW DO THESE THINGS RELATE?

- Artificial Intelligence (AI)
- **Machine Learning (ML)**
- Deep Learning



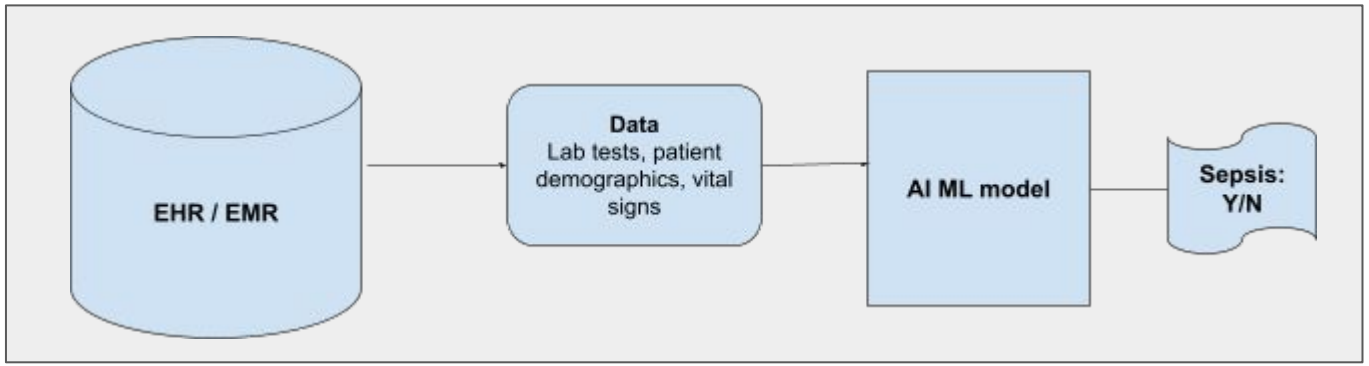
Red Hat

MACHINE LEARNING IN ACTION



License Plate Recognition

Chat Bots



Detecting Patient Sepsis at the Hospital

Image credit: Watson Assistant, HCA Sepsis Detection, License Plates Demo

MACHINE LEARNING TERMINOLOGY

- Data & Data Prep
- Training
- Models
- Serving/Inferencing
- Monitoring

Red Hat

People and Teams

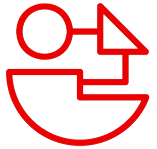
A DATA-CAPABLE PROJECT TEAM

Notice any roles that seem new to you?

Data Engineers

Data Scientists

Developers



Architects

IT Ops Engineers

WHAT A DATA-CAPABLE TEAM DOES

Scoping / Objective

Gather and Prep Data	Models and Training	Deployments & Integration	Ops & Monitoring
Data Engineering	**Data Science**	**Continuous Integration & Deployment**	**Monitor / alerts**
Data Ingestion	Data Splitting	Data Preprocessing	Consumption & optimization metrics
Data Cleansing	Feature Engineering	App Dev / Heuristics	Satisficing (Gating) metric
Data Analysis	Model Development	Inferencing Pipeline	Logging & Visualization
Data Transformation	Model Training	Deployment Targets	Explainability, Interpolation
Data Validation	Training Optimization	Deployment Patterns	Drift, Decay, Skew, Shift
	Model Validation		Improvements

PERSPECTIVE

around the industry

*"The story of enterprise Machine Learning - It took me **3 weeks to develop** the model. It's been **>11 months, and it's still not deployed**."*

~ IBM VP Dinesh Nirmal

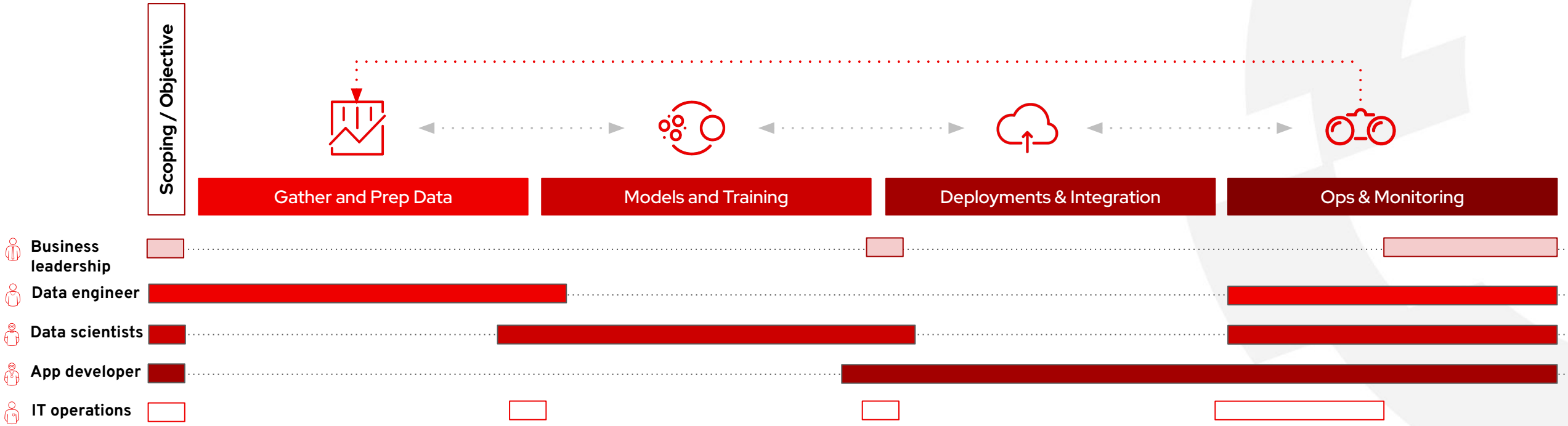
*"Only a small fraction of real-world ML systems is composed of the ML code... The required **surrounding infrastructure is vast and complex**."*

~ D. Sculley (Google), et al.
Hidden Technical Debt in Machine Learning Systems

*"A top complaint of data science, application development and delivery(AD&D) teams, and, increasingly, line-of-business leaders is the **challenge in deploying, monitoring, and governing machine learning models** in production. **Manual handoffs**, frantic monitoring, and loose governance prevent organizations from deploying more AI use cases."*

~ Forrester Report, 2020

CHALLENGE – BREAK DOWN SILOS

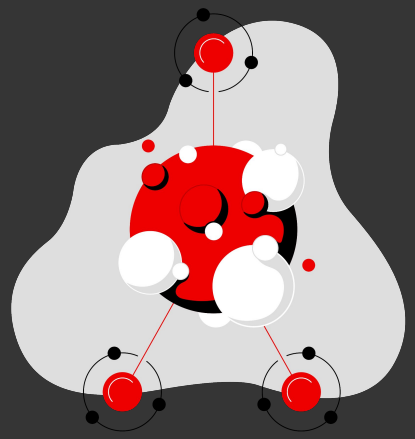


Red Hat

Accelerating Solutions

OPEN SOURCE COMMUNITIES

ARE THE DRIVING FORCE BEHIND TECHNOLOGY INNOVATION TODAY



Red Hat

CONSIDER OPEN SOURCE IN 2016

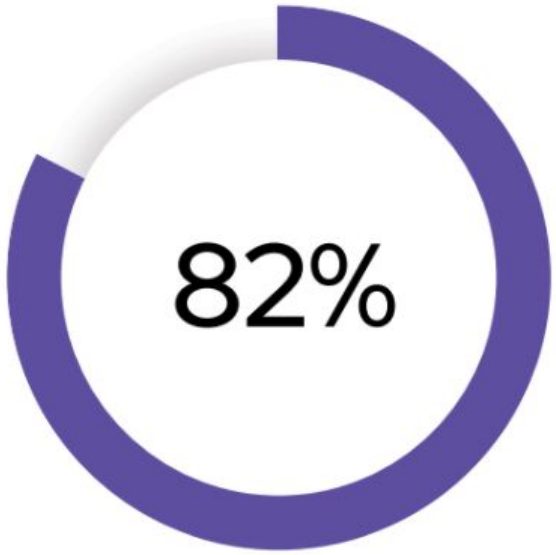
Innovate faster and break out of the proprietary loop

> An organization that does not fully consider open source options alongside the proprietary offerings they have traditionally procured is missing out on sound technologies, access to vibrant communities, and the opportunity to tap innovative new ways of working.
>
> Today, **failure to fully consider open source options is unwise**. Within a few short years, it will be unforgivably negligent.

Forrester
April 2016

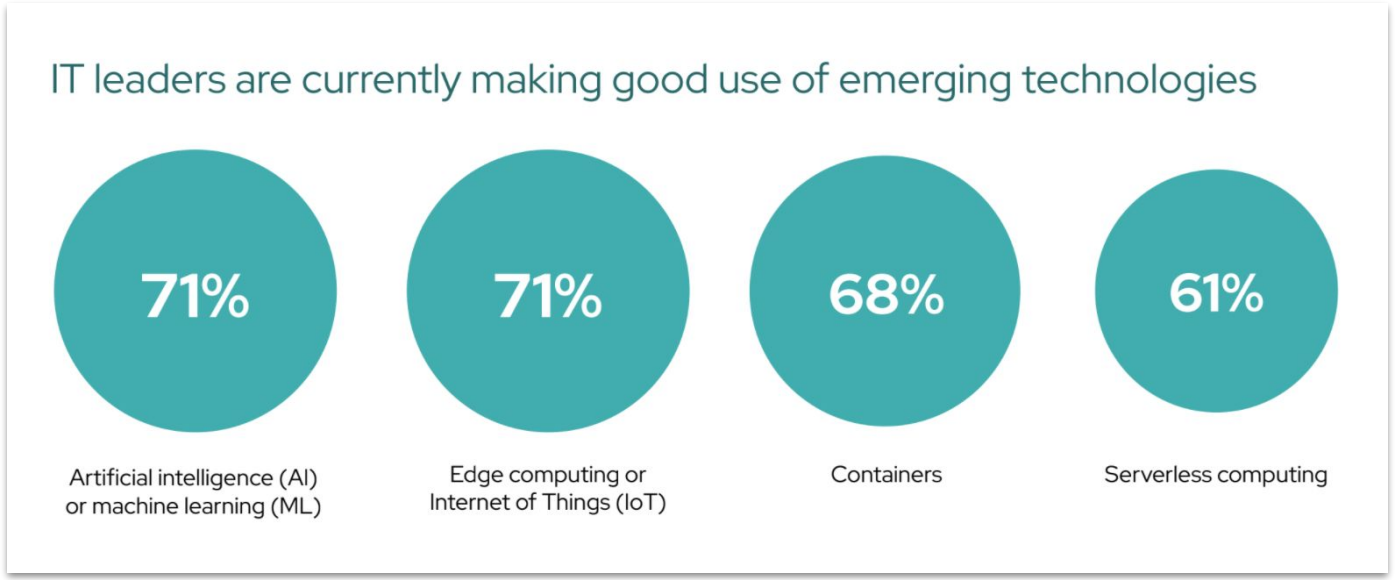
Source: Paul Miller and Lauren E. Nelson. Forrester, Open Source Powers Enterprise Digital Transformation, April 2016. ID: 133302.

NOW IN 2022



of IT leaders are more likely to select a vendor who contributes to the open source community.

(APAC = 77%, EMEA = 82%, LATAM = 83%, U.S. = 82%)



APAC		EMEA		LATAM		U.S.	
AI or ML	73%	AI or ML	70%	AI or ML	65%	AI or ML	75%
Edge or IoT	68%	Edge or IoT	69%	Edge or IoT	71%	Edge or IoT	73%
Containers	66%	Containers	63%	Containers	69%	Containers	73%
Serverless	58%	Serverless	64%	Serverless	58%	Serverless	61%

https://www.redhat.com/en/enterprise-open-source-report/2022

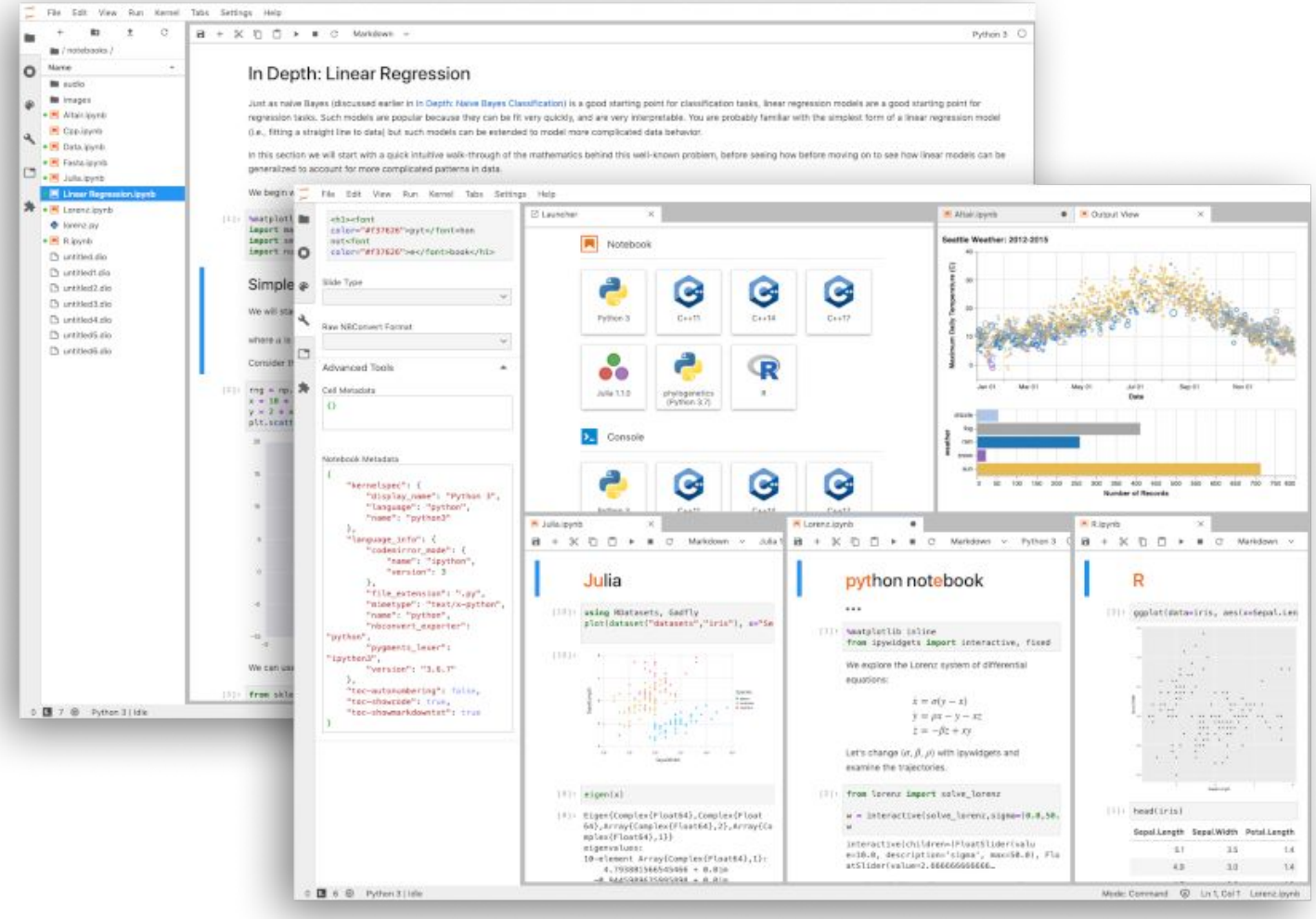
Red Hat

MOST DATA SCIENCE IS OPEN SOURCE

These are some of the most popular open source

(Scikit-Learn, TensorFlow, PyTorch, OpenCV, Kubeflow, etc.)

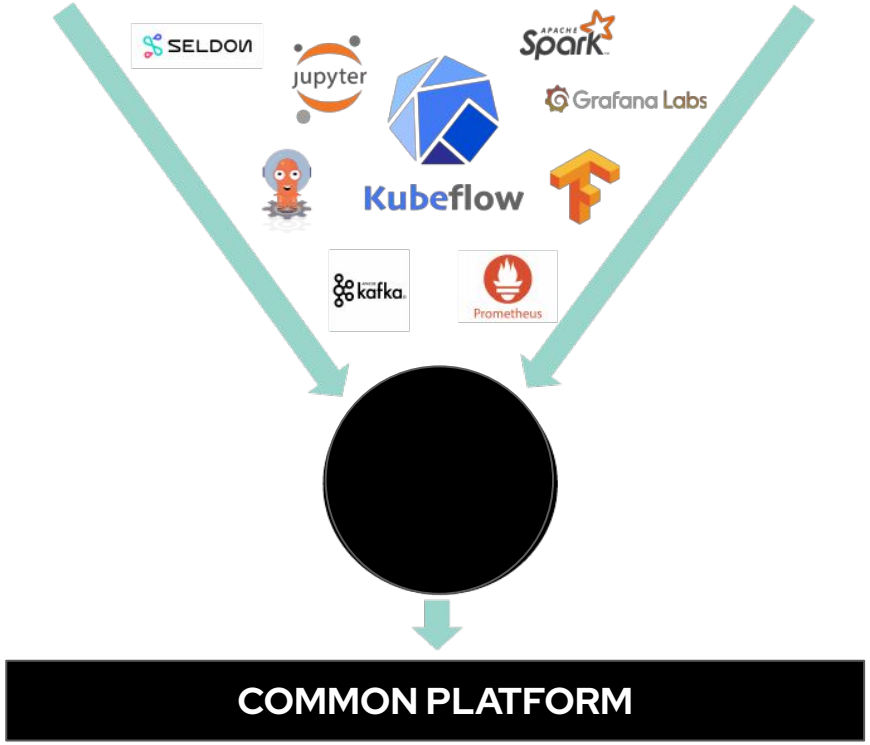
SINGLE PLAYER DATA SCIENCE



LAPTOP WITH SINGLE USER NOTEBOOK AND LOCAL TOOLS

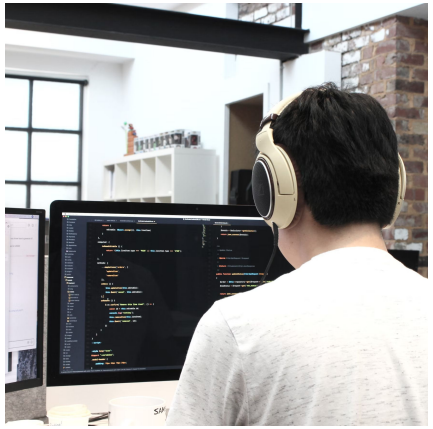
Image credit: Unsplash Glenn Carstens-Peters

MULTIPLAYER DATA SCIENCE



- **Consistency** across team members
- Scalable **end-to-end** tools and components
- Provide ETL tools used by **Data Engineers**
- Provide development tools for **Data Scientists**
- Provide middleware services needed by **Developers**
- **Standardized** data science environments
- Provide tools to **orchestrate** ML and app delivery
- AI/ML **pipelines** and long processing tasks
- Provide **monitoring** tools for models and services

MULTIPLAYER DATA SCIENCE



MONITORING, GOVERNANCE, SECURITY, ORCHESTRATION

MULTI-USER NOTEBOOKS

DATA SERVICES

APP SERVICES

HYBRID PLATFORM (ON-PREM OR CLOUD)

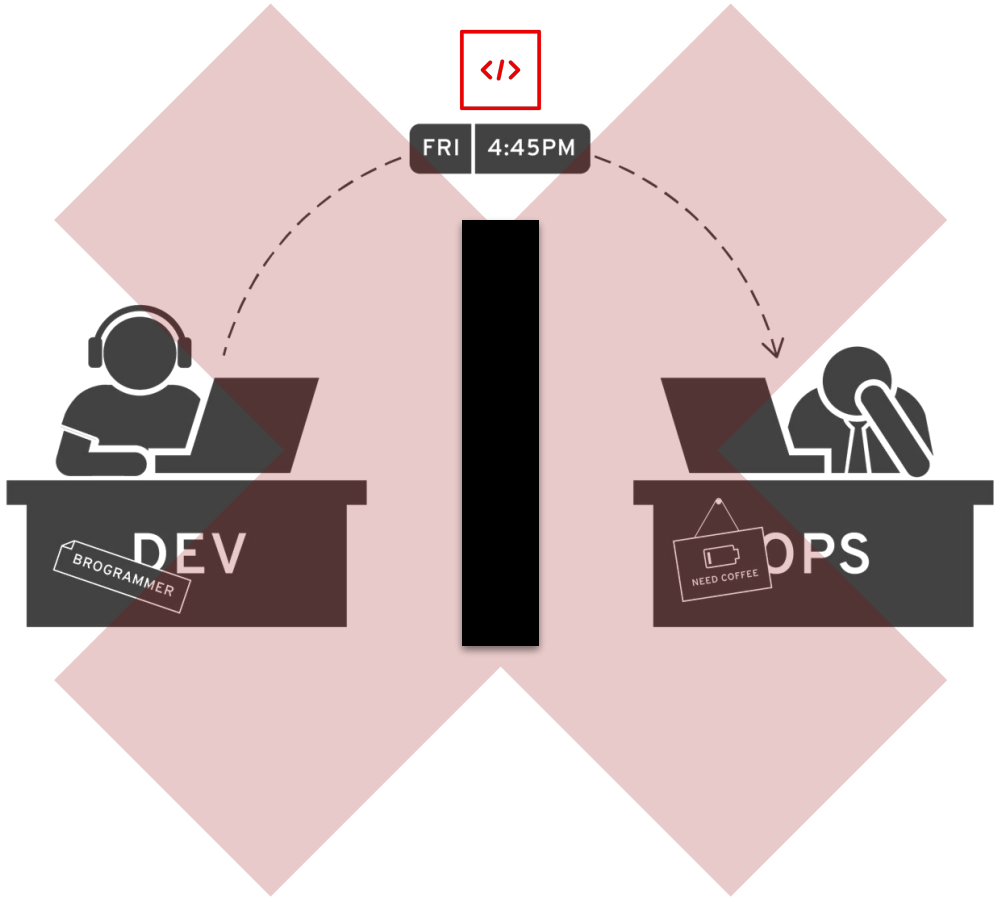
SCALABLE COMPUTE

NETWORKED STORAGE

GPUs

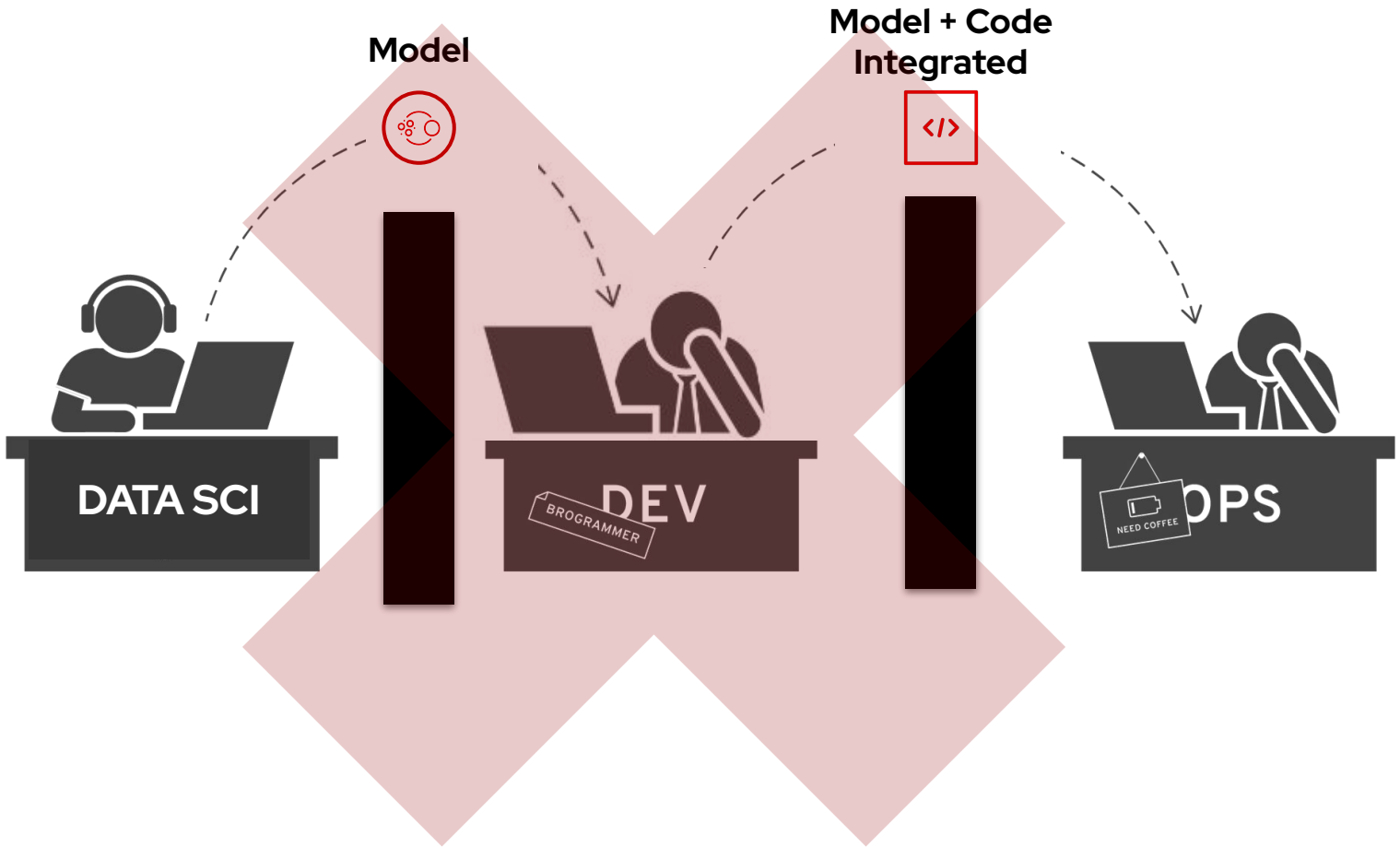
Image credit: Unsplash Tim van der Kuip, True Agency

DEVOPS HAS SOLVED A LOT OF IT CHALLENGES



Red Hat

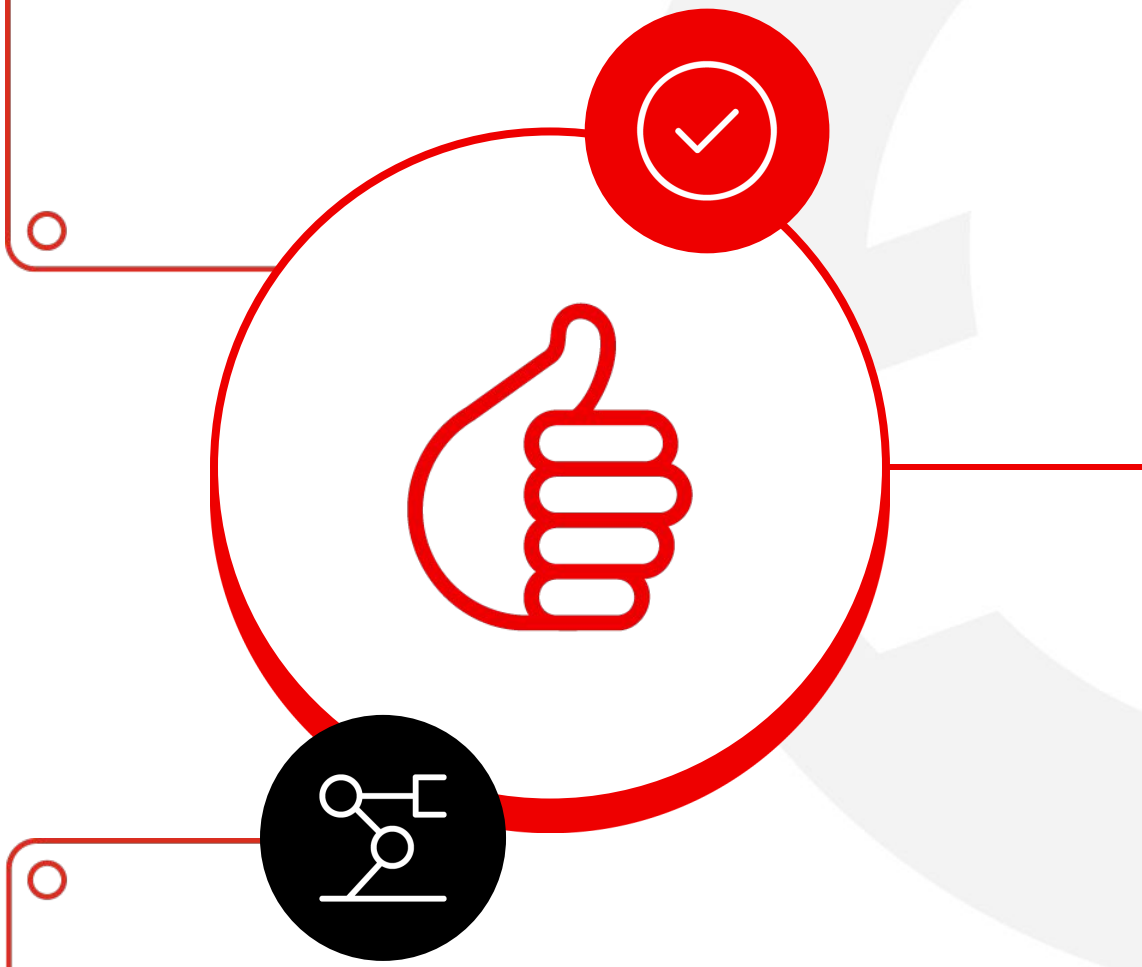
AVOID MAKING PAST MISTAKES AGAIN



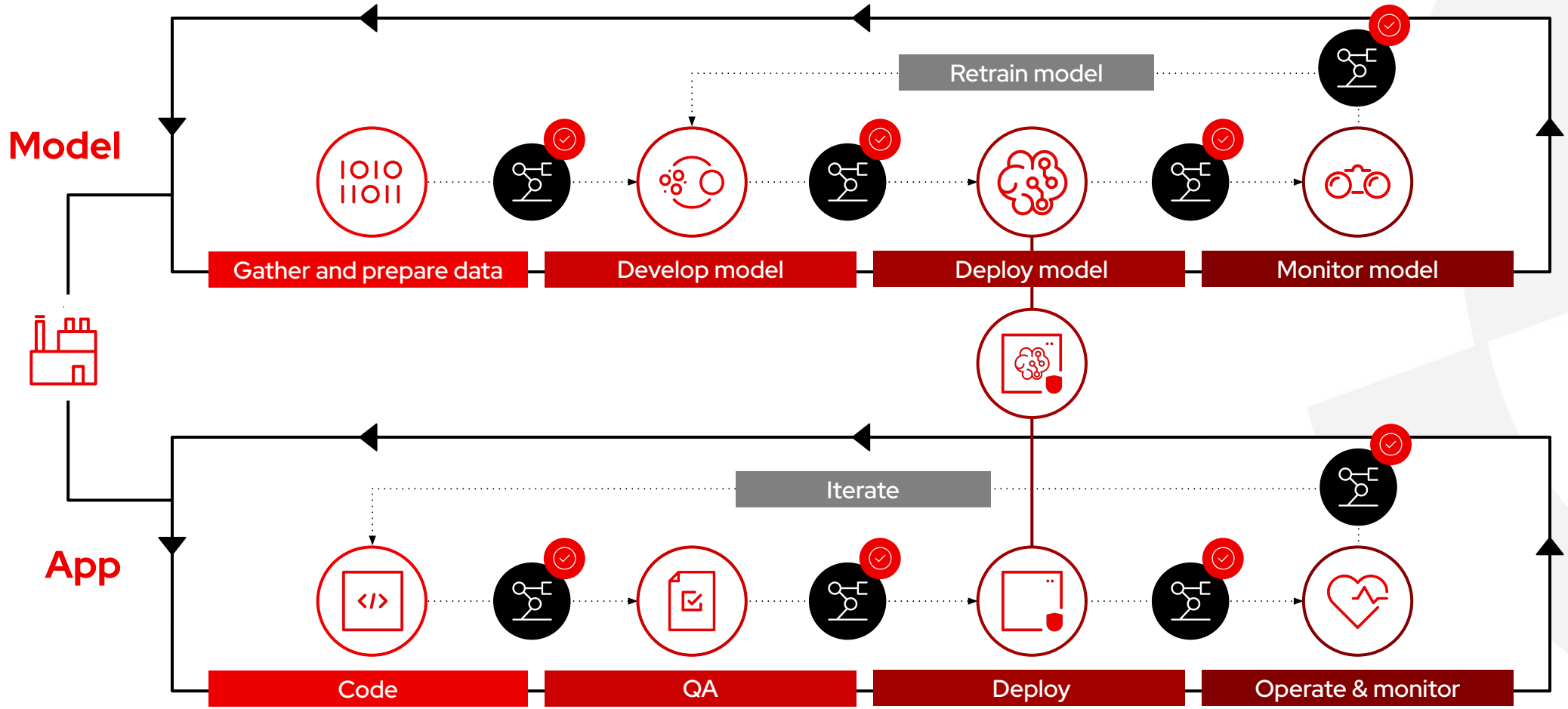
Red Hat

MLOps

The **practices, culture,** and **tools** that aim to **reliably** and **efficiently deploy** and **maintain AI/ML models** in production.



Red Hat



PIPELINES / CONTINUOUS INTEGRATION / EVERYTHING AS CODE / AUTOMATE EVERYTHING / RAPID FEEDBACK

Automation

Validation

Red Hat

Takeaway Resources

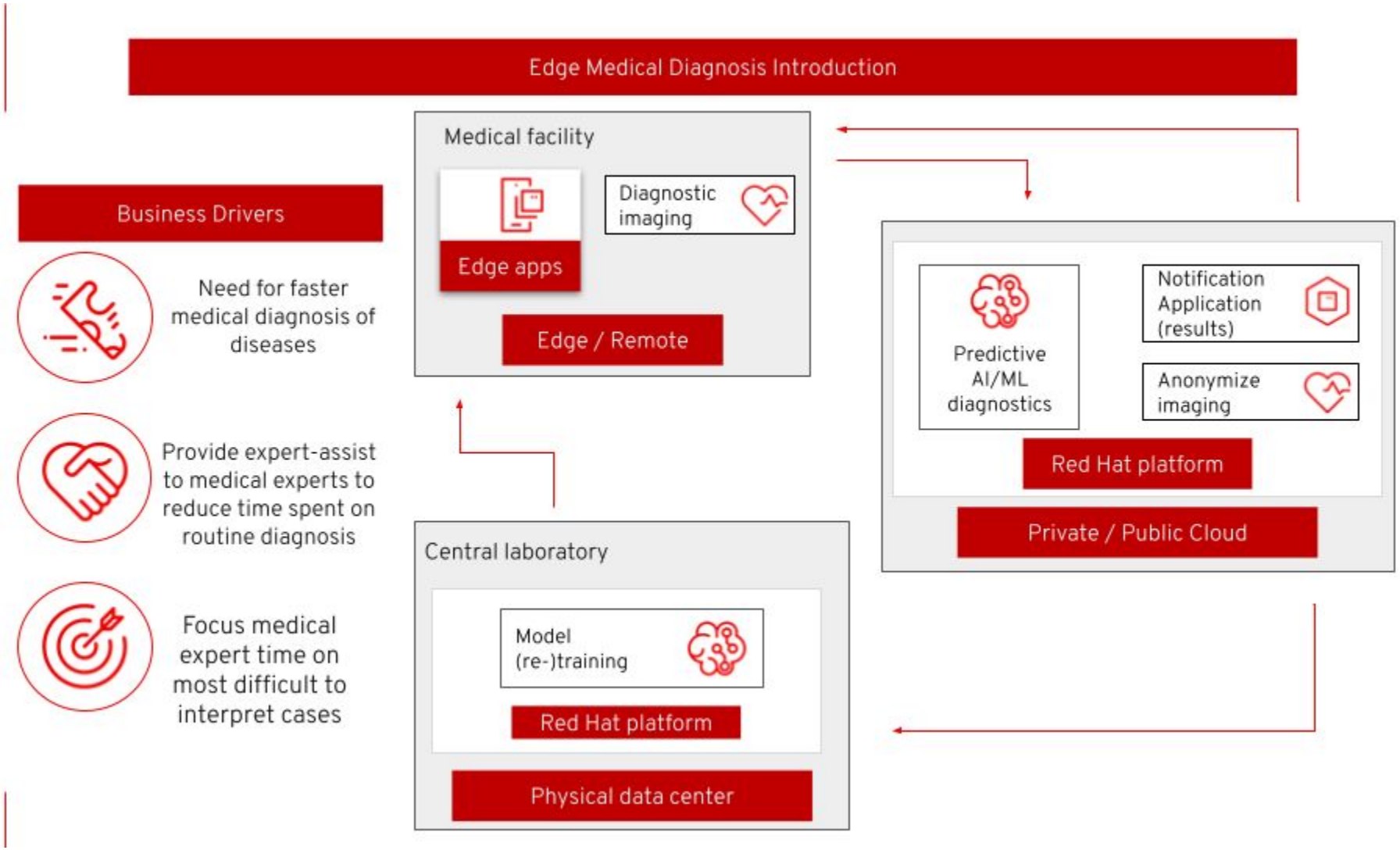
- **What is Machine Learning?**
  - https://www.redhat.com/en/blog/what-machine-learning
  - https://www.youtube.com/watch?v=O0XtsKFs5xI
- **Red Hat OpenShift Data Science**
  - https://www.redhat.com/en/technologies/cloud-computing/openshift/openshift-data-science
- **Open Data Hub**
  - https://opendatahub.io/

Example

(Time Permitting)

Example

EDGE MEDICAL DIAGNOSIS

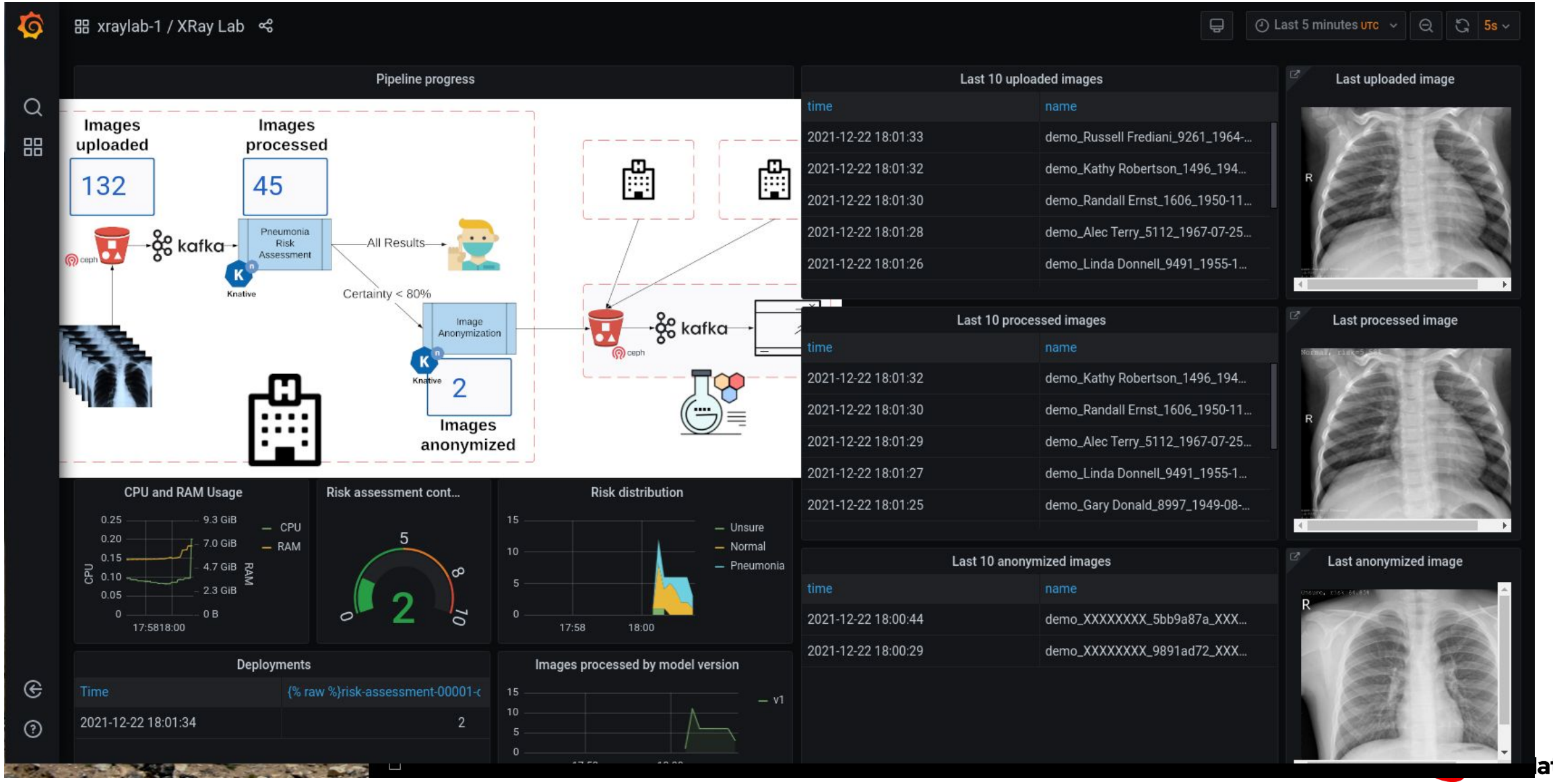


Source: Validated Patterns

Red Hat

Example

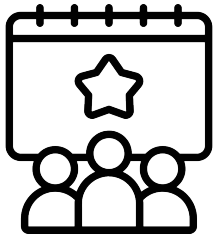
EDGE MEDICAL DIAGNOSIS



Source: Validated Patterns

DevNation
FEDERAL 2022

October 5, 2022 | 9:00am - 4:30pm
Convene | 1201 Wilson Blvd 30th floor, Arlington, VA 22209



REGISTER HERE:
red.ht/DevNationFederal2022

WE HOPE TO SEE YOU THERE!

Join us for:

4 Main Stage Sessions

3 Breakout Technical Sessions

2 Hands-On Labs

- Ethical AI
- Building Modern Machine Learning Pipelines with AWS
- Data Driven Decisions Panel
- Introduction to Quantum
- **LAB:** License plate recognition
- **LAB:** Credit card fraud detection with AI/ML

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat