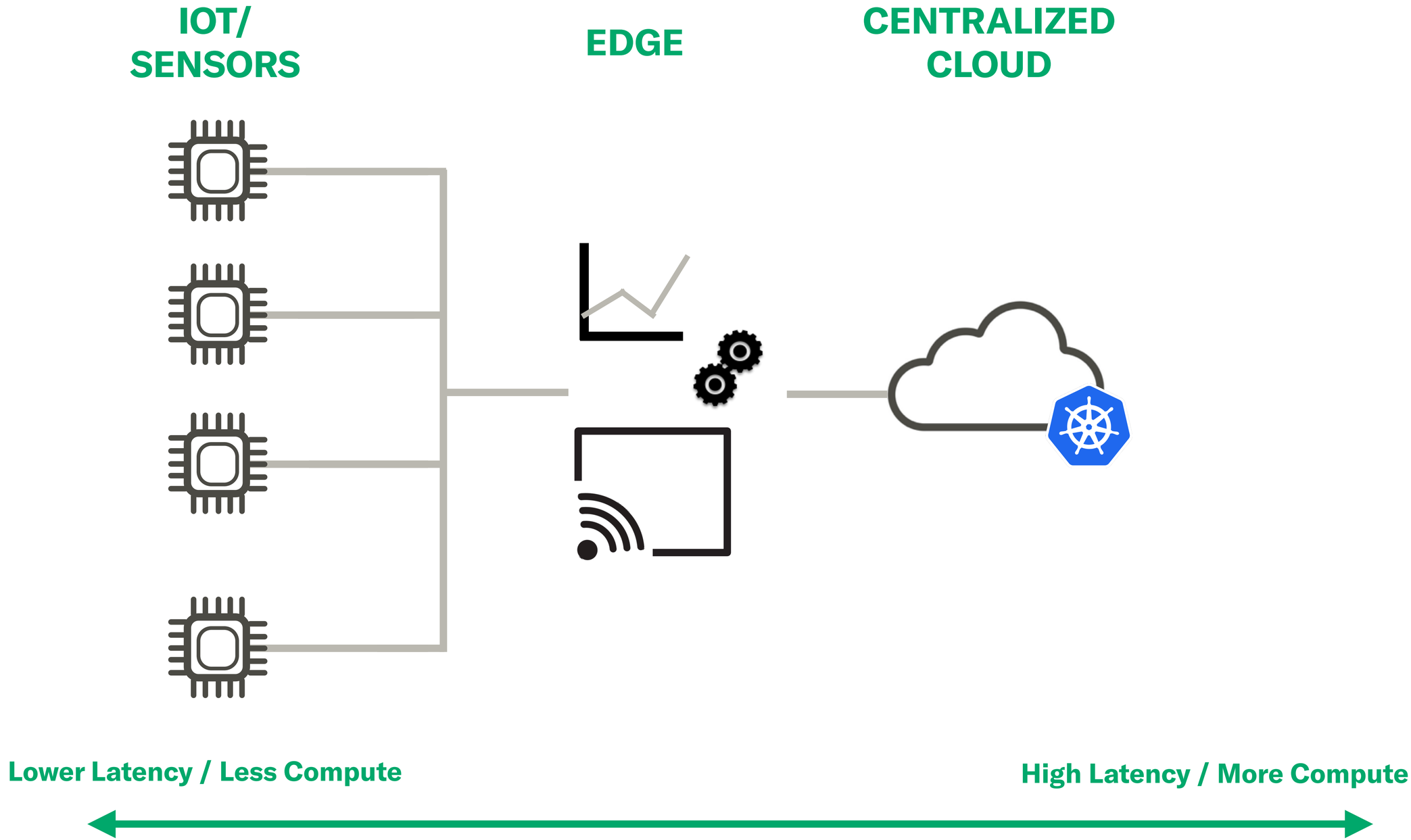
Cloud to Edge:

Architecting for the Next Generation

WESLEY REISZ

Cloud Native Architect /
Engineering Leader
Chairperson QCon SF &
Co-host of The InfoQ Podcast


IOT/
SENSORS
EDGE
CENTRALIZED
CLOUD
Lower Latency / Less Compute
High Latency / More Compute

WHY

Cost: Intelligent routing and optimization of compute

Performance Cost Optimization: Reduce latency / Improve performance

Legislation: Regulations & Compliance (GDPR)

New Use Cases: Such as the ability to influence the energy market through the aggregation of consumer devices.

Examples: Inferencing at the Edge, Reduce backhaul, Federated Learning, Enhanced CDN, Better Routing

WHY

Cost: Intelligent routing and optimization of compute

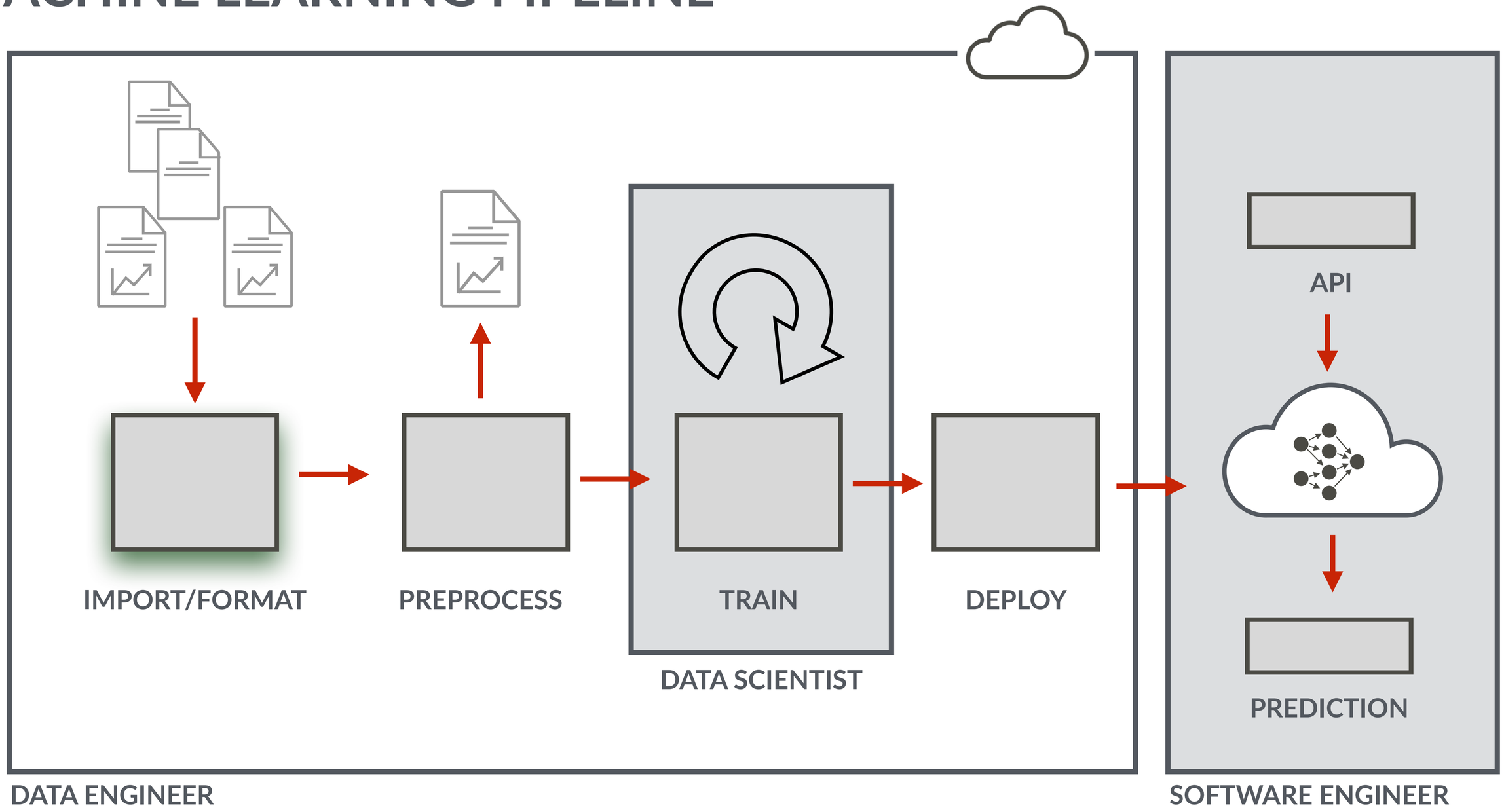
Performance Cost Optimization: Reduce latency / Improve performance

Legislation: Regulations & Compliance (GDPR)

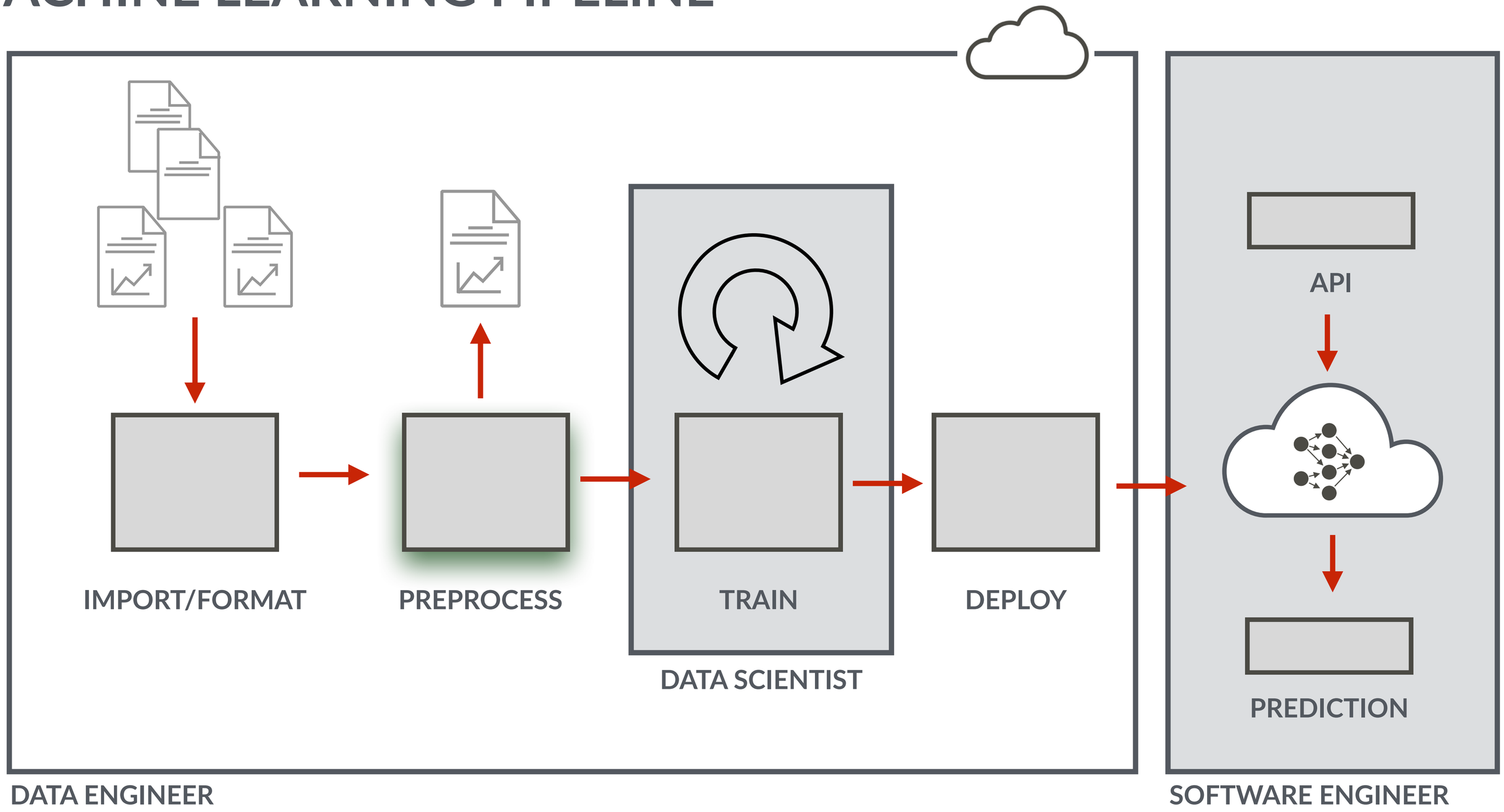
New Use Cases: Such as the ability to influence the energy market through the aggregation of consumer devices.

Examples: ***Inferencing at the Edge***, Reduce backhaul, Federated Learning, Enhanced CDN, Better Routing

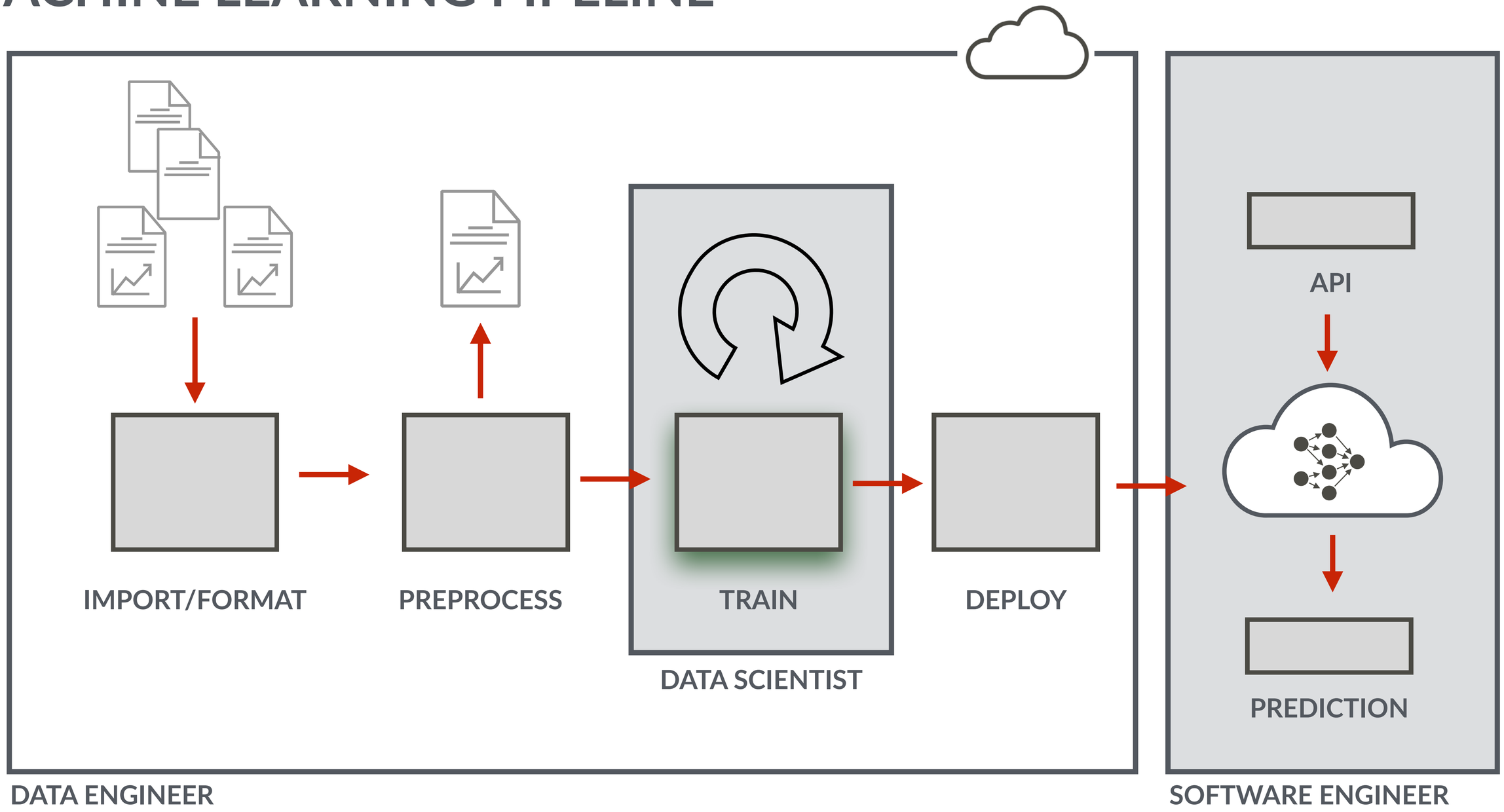
MACHINE LEARNING PIPELINE



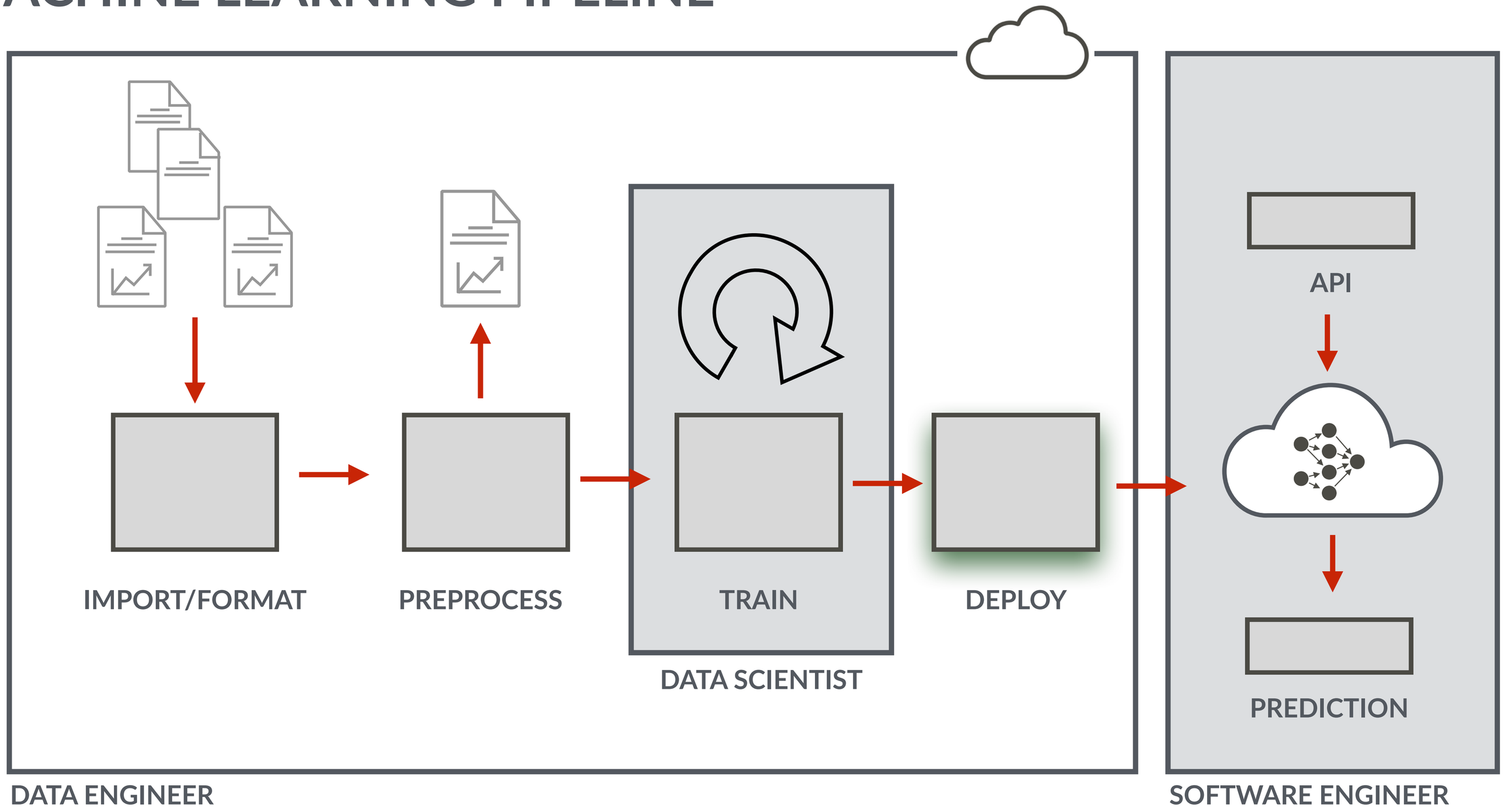
MACHINE LEARNING PIPELINE



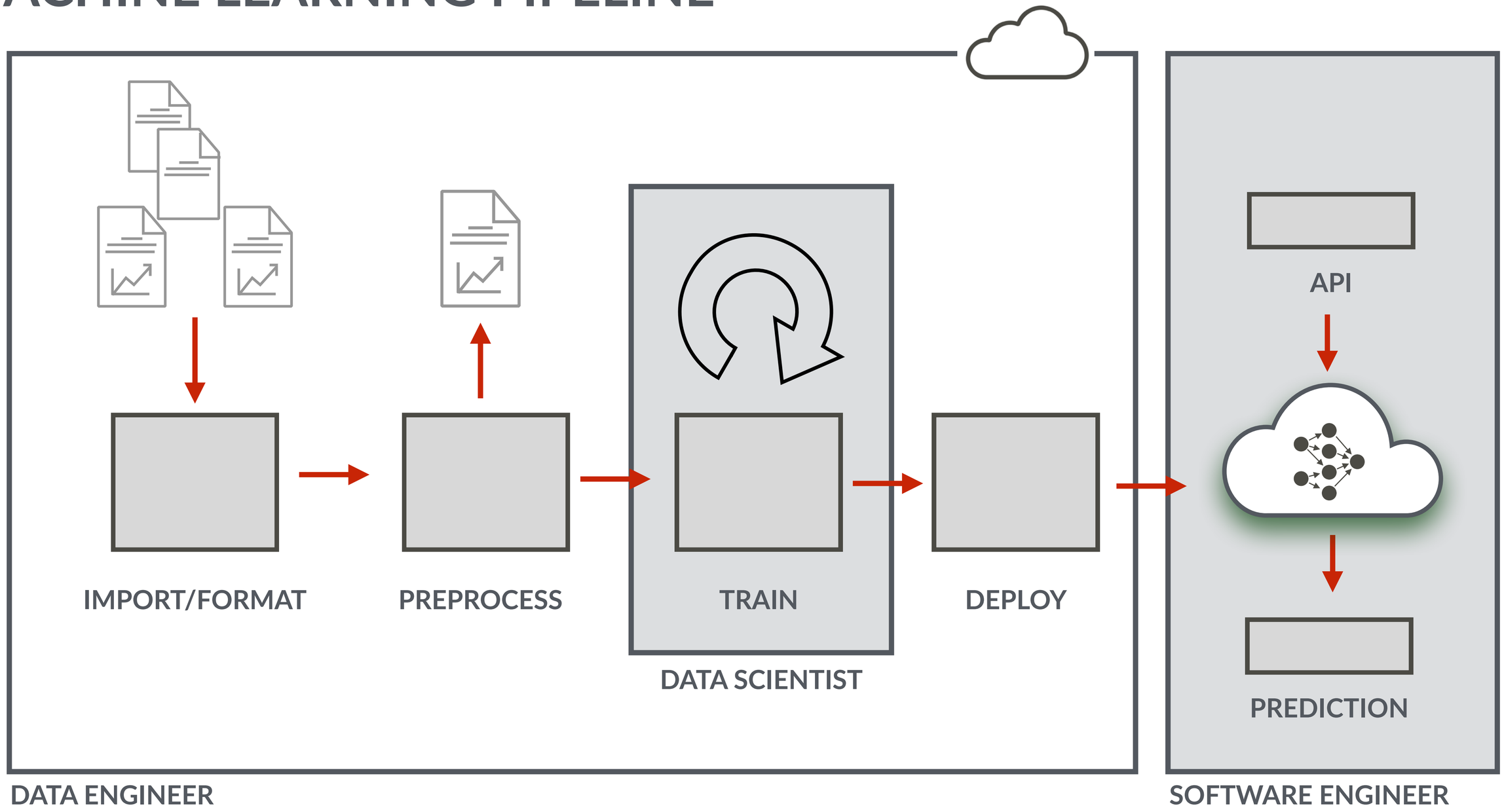
MACHINE LEARNING PIPELINE

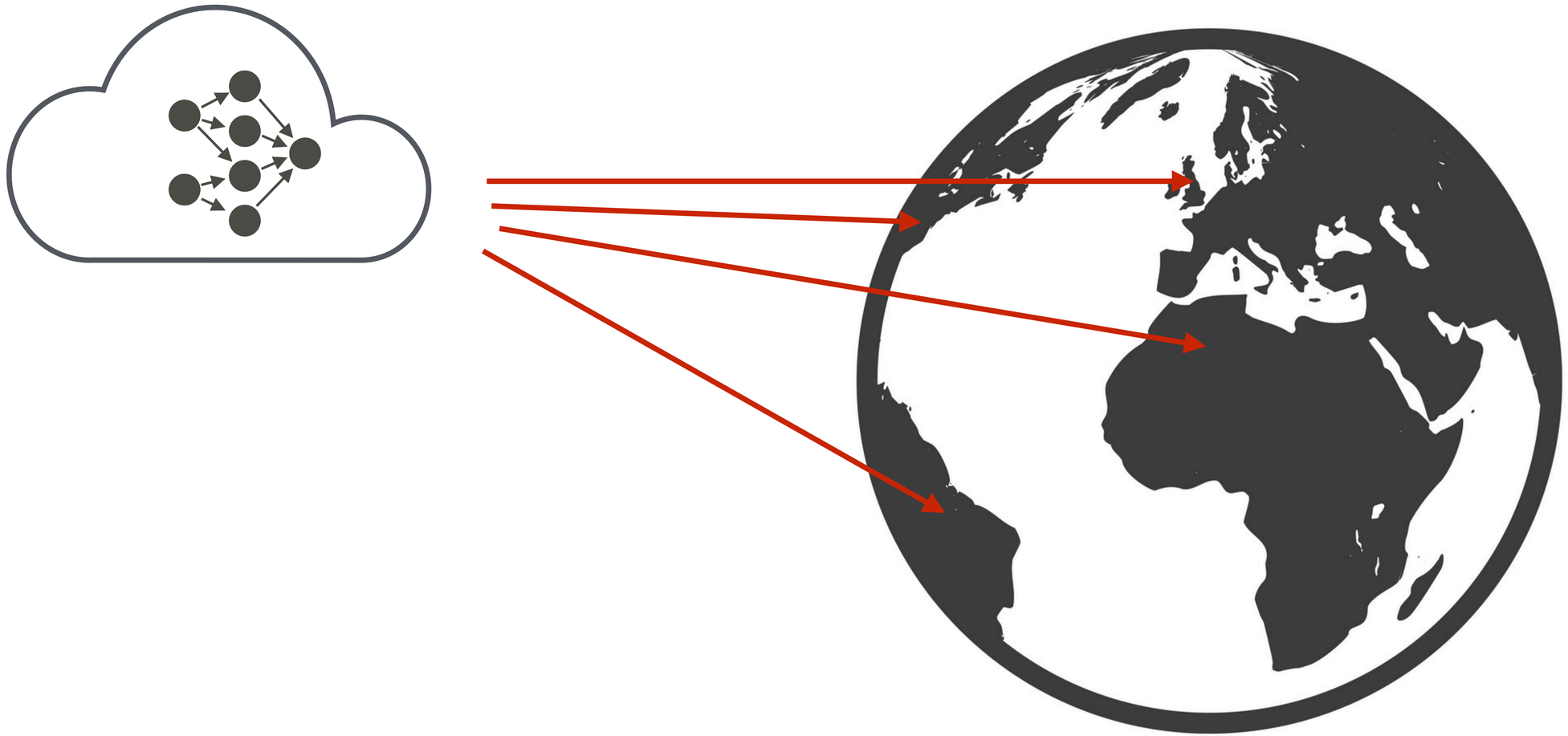


MACHINE LEARNING PIPELINE



MACHINE LEARNING PIPELINE






S
T
A
R
Situation
Task
Action
Result
Introduce the situation to the employer and set the context
Describe the task you had to complete, including the expectations and challenges it would involve
Explain what you did, and how you did it
End with the results of your efforts, including accomplishments, rewards, and impact

SITUATION

- ***Java developer*** working for an enterprise. Production code is deployed Java/Spring
- Data Scientists work in ***Python*** & Jupyter Notebooks
- Cloud Native Deployments are ***K8s***.

TASK

- ***Find*** a way to work with the data scientists to deploy a machine learning models using enterprise approved tools (where possible).

- **Inferencing** at the edge

ACTION

- ***Create*** a virtual env to work with the data scientist in Juypter Notebooks

- ***Test / Iterate*** it with a Flask web service.

- ***Harden*** it and prep it to deploy it with Java Sprint

RESULT

- Stateless ***Dockerized*** container ready to deploy at the edge
- Roadmap on how we might deploy this to the edge

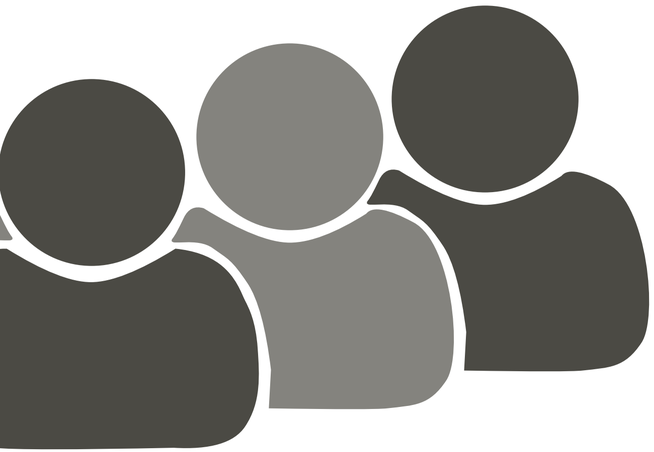
DEMO

- Jupiter Notebook
- Python Service (recreate it)
- Spring Boot Service (import model)

WHAT ABOUT DEPLOYING

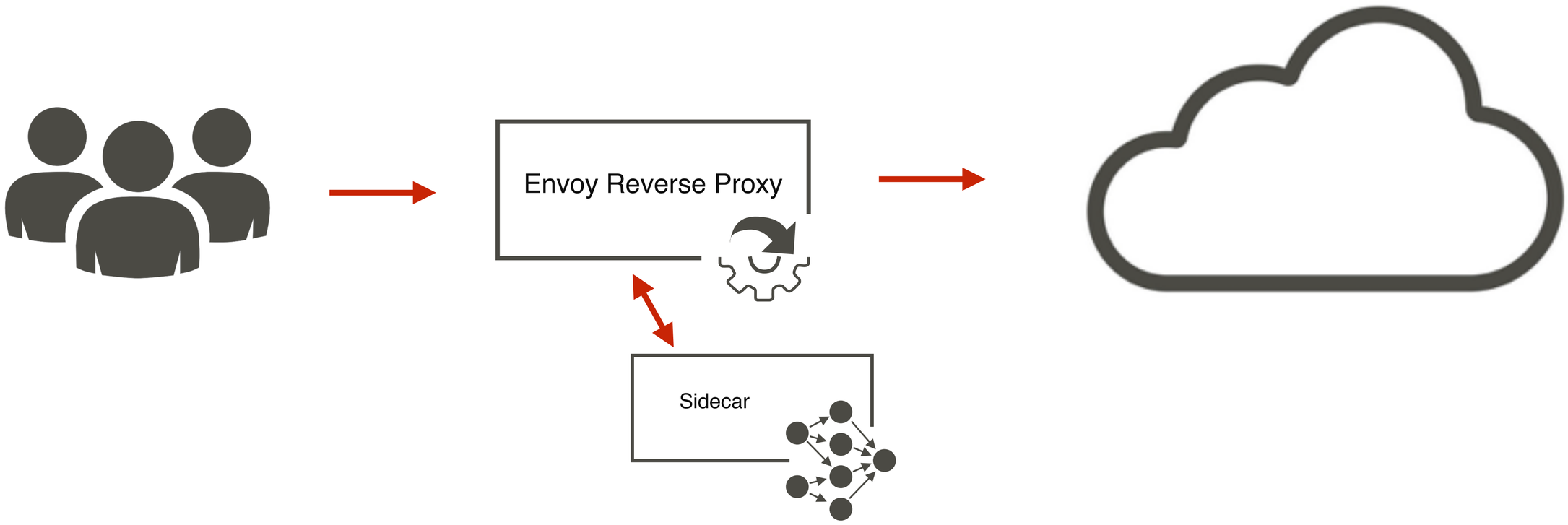
DEVICE EDGE

INFRASTRUCTURE EDGE

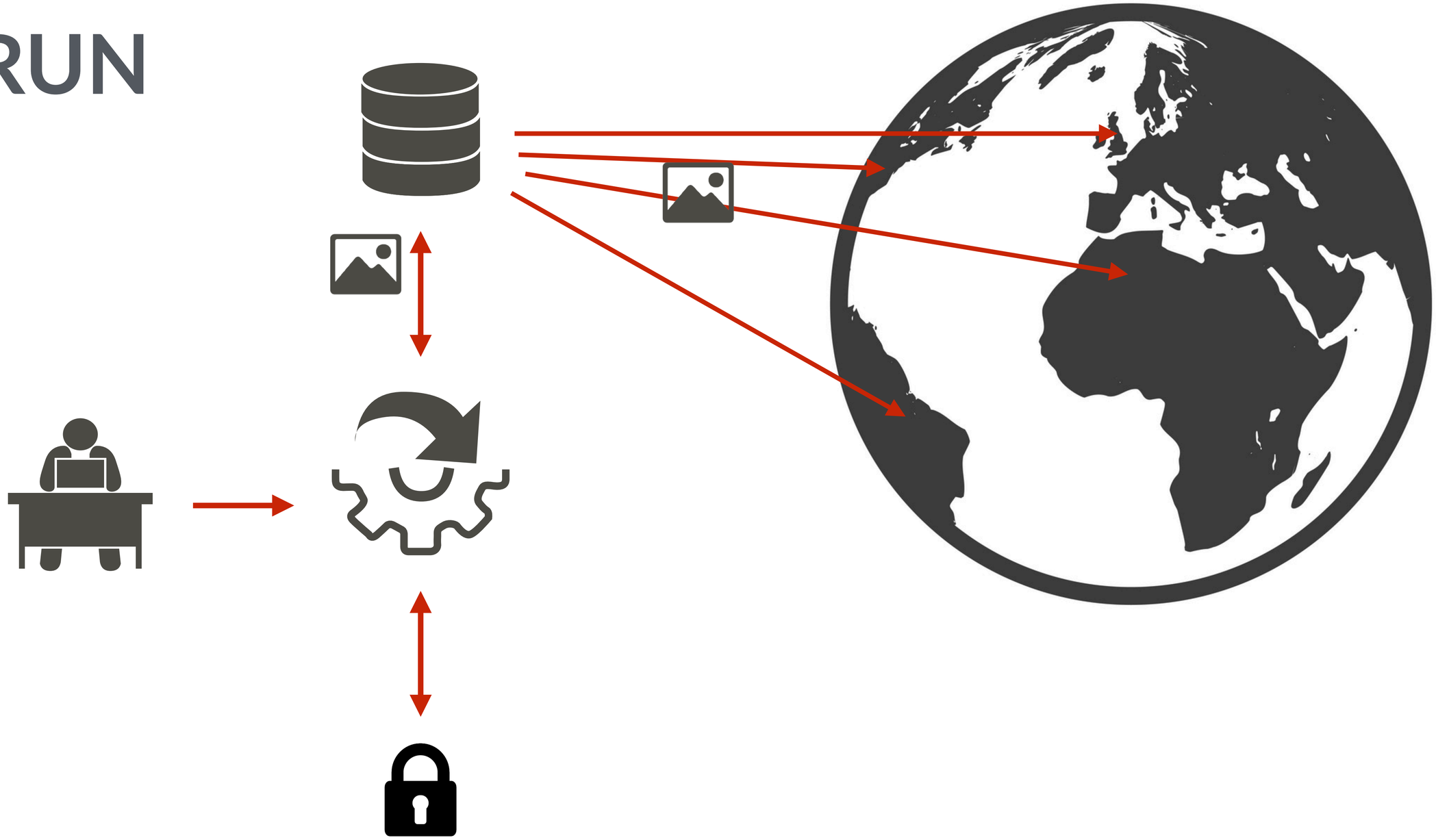


LAST MILE

BUILD

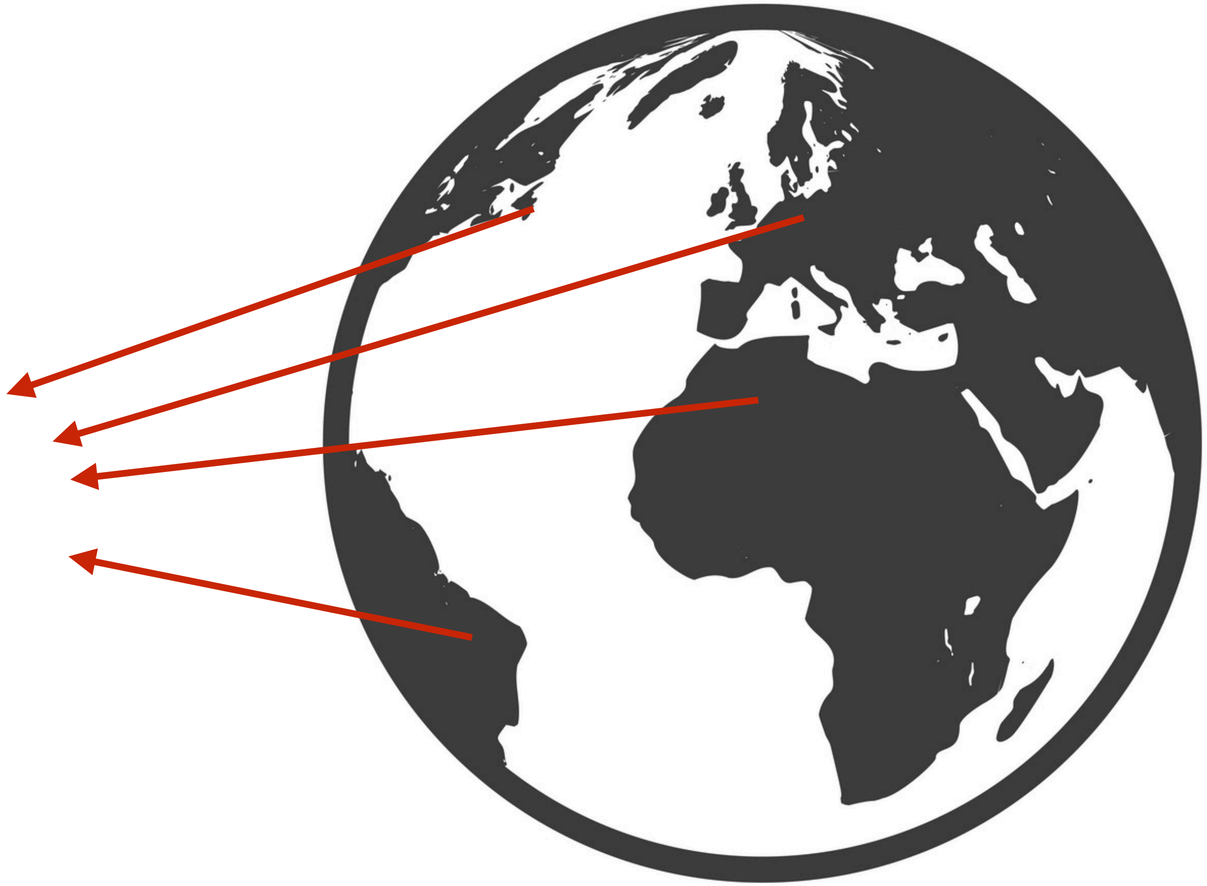


RUN



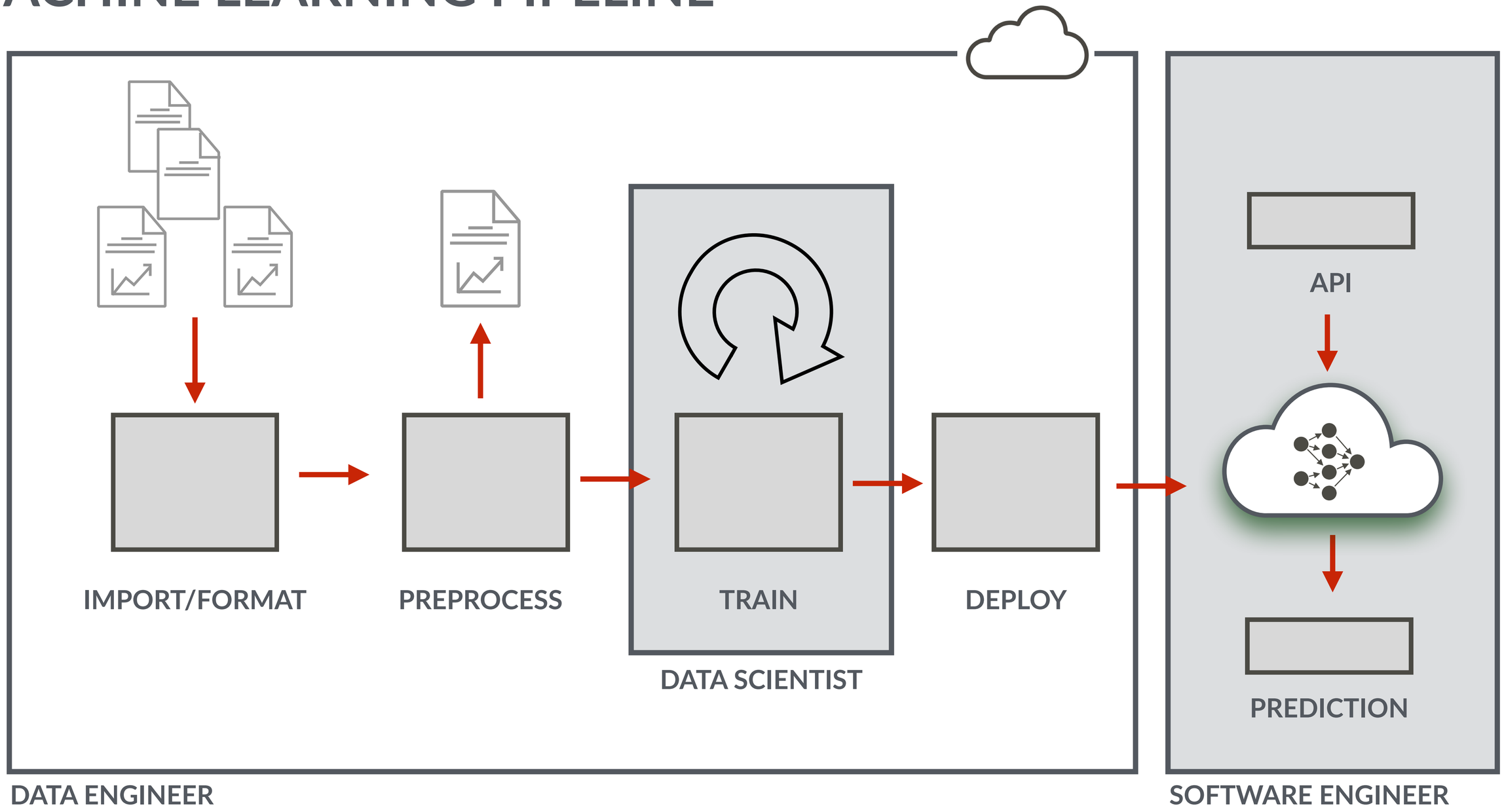
MANAGE

Cloud



Global POPs

MACHINE LEARNING PIPELINE



Use Cases

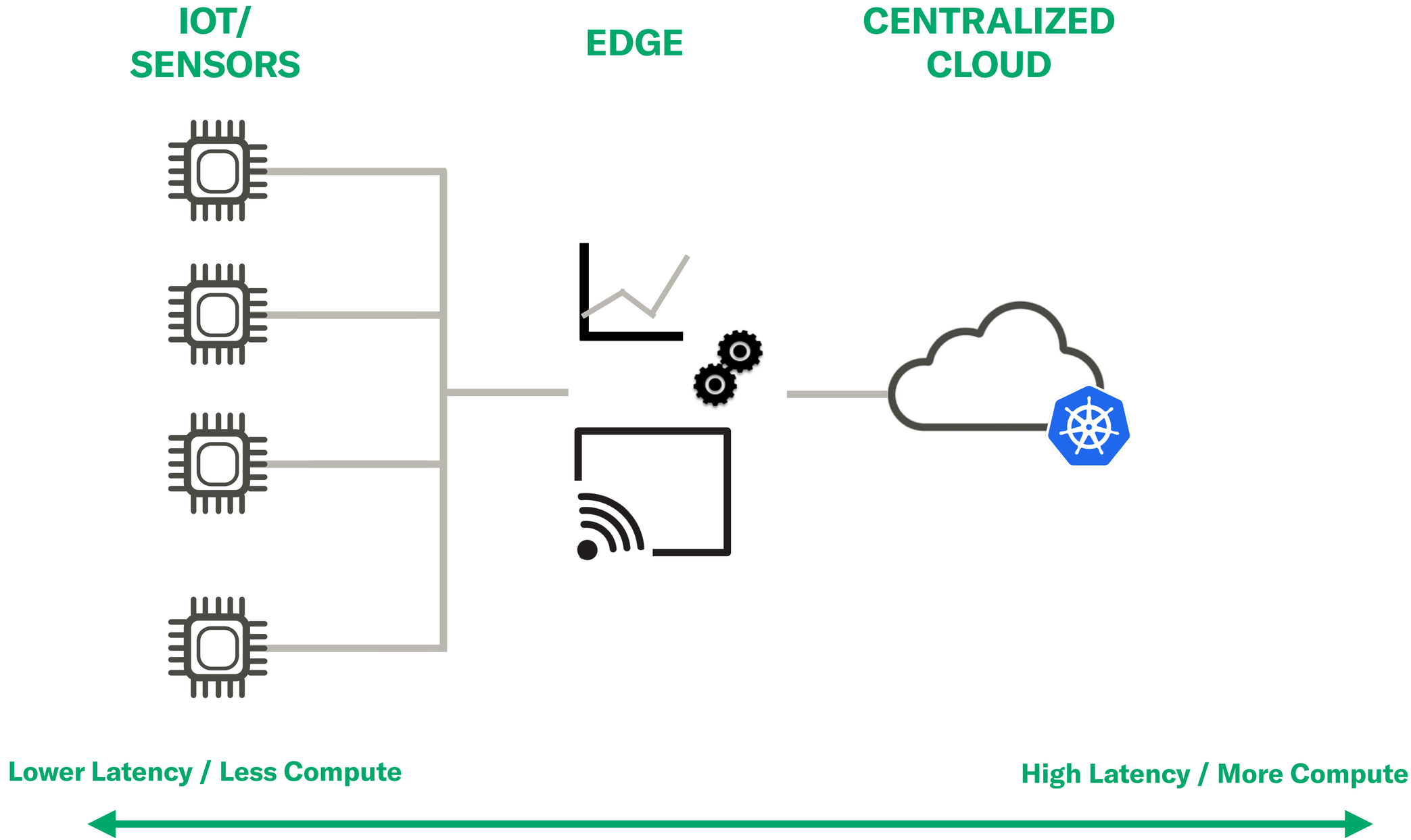
What Can I Do at the Edge

Cloudflare:
Service Worker API implementation for the Cloudflare platform. Brings a server less style approach to running JavaScript workloads on their Points of Presence.
Source: https://www.infoq.com/presentations/cloudflare-workers/

Facebook Live:
Stream is sent via RTMP (Real-Time Messaging Protocol) to a geographically local PoP. The connection is forwarded over an internal Facebook network to a Facebook data-centre. When you see a live stream in your feed and you click on it the player requests the manifest. If it isn't already on your local PoP the request goes to the data centre to get the manifest, and then fetches the media files in 1 sec clips. As they get sent back they are cached on the PoP if they aren't there already.
Source: https://www.infoq.com/podcasts/sachin-kulkarni-facebook-live/

Chick fil-A:
We think of our Edge Computing environment as a "micro private cloud". By this, we mean that we provide developers with a series of helpful services and a place to deploy their applications on our infrastructure.
Source: https://medium.com/@cfatechblog/edge-computing-at-chick-fil-a-7d67242675e2

Tesla:
A Virtual Power Plant (VPP) is a network of distributed energy-resources (often solar, wind, and batteries) that are aggregated to provide smarter and more flexible power generation, distribution, and availability. Tesla's VPP consists of vertically integrated hardware and software, including both cloud and edge computing.
Source: https://www.infoq.com/news/2020/03/tesla-vpp


IOT/
SENSORS
EDGE
CENTRALIZED
CLOUD
Lower Latency / Less Compute
High Latency / More Compute

KEY TAKEAWAYS

- ***Edge*** is running workloads between the user and the cloud. It can be divided into two parts (using the the ***last mile*** as a boundary). **Device edge** is closest to the user and ***Infrastructure edge*** closest the cloud.
- Edge works ***reduce network costs***, ***improve speed/latency***, and ***enables new use cases*** around *machine learning, performance, compliance, & cost optimization*.
- It’s time to ***rethink*** how we deploy workloads. No longer is it just a defacto Cloud deployment.

Cloud to Edge:

Architecting for the Next Generation

WESLEY REISZ

Cloud Native Architect /
Engineering Leader
Chairperson QCon SF &
Co-host of The InfoQ Podcast

Architect Team Presentation

Technical Depth + Breadth – Geek Speak

60 Minutes

Presentation Guidelines

- Candidate can use any combination of slides, whiteboard, demo
- First 45 minutes: interviewer(s) will not interrupt the session
- Last 15 minutes: interviewer(s) ask technical depth/breadth questions


S
T
A
R
Situation
Task
Action
Result
Introduce the situation to the employer and set the context
Describe the task you had to complete, including the expectations and challenges it would involve
Explain what you did, and how you did it
End with the results of your efforts, including accomplishments, rewards, and impact