RISE to the challenges of Intelligent systems

A prediction for future research

Joseph E. Gonzalez

Asst. Professor, UC Berkeley

jegonzal@cs.berkeley.edu

Machine Learning

Timescale: minutes to days

*Heavily studied ... primary focus of the **ML research***

amplab

APACHE Spark

MLbase

KeystoneML

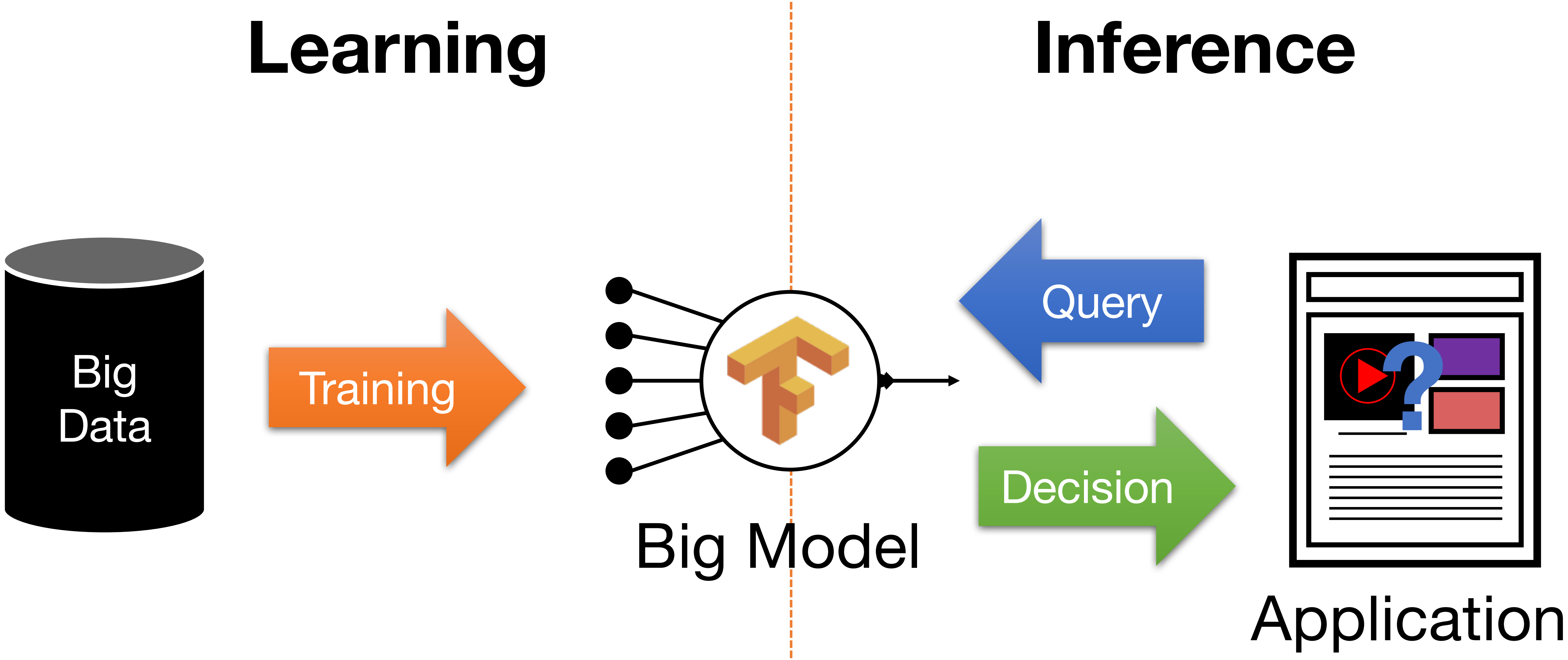
GraphX

Splash

CoCoA

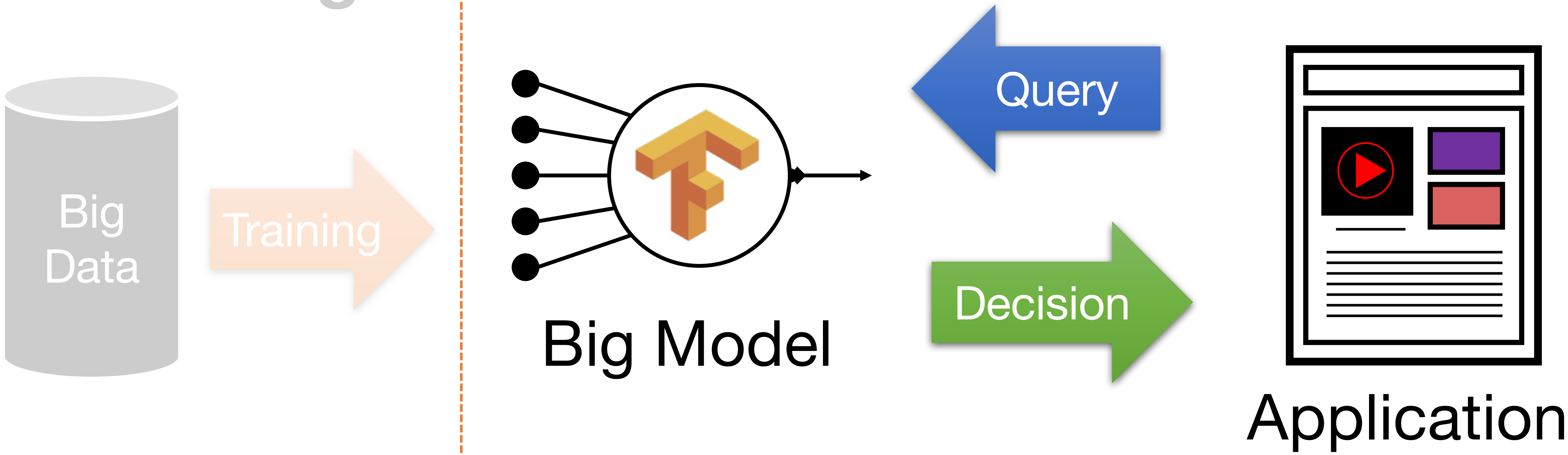
Please make a Logo!

Learning


Learning
Inference
Big
Data
Training
Query
Decision
Big Model
Application

Learning

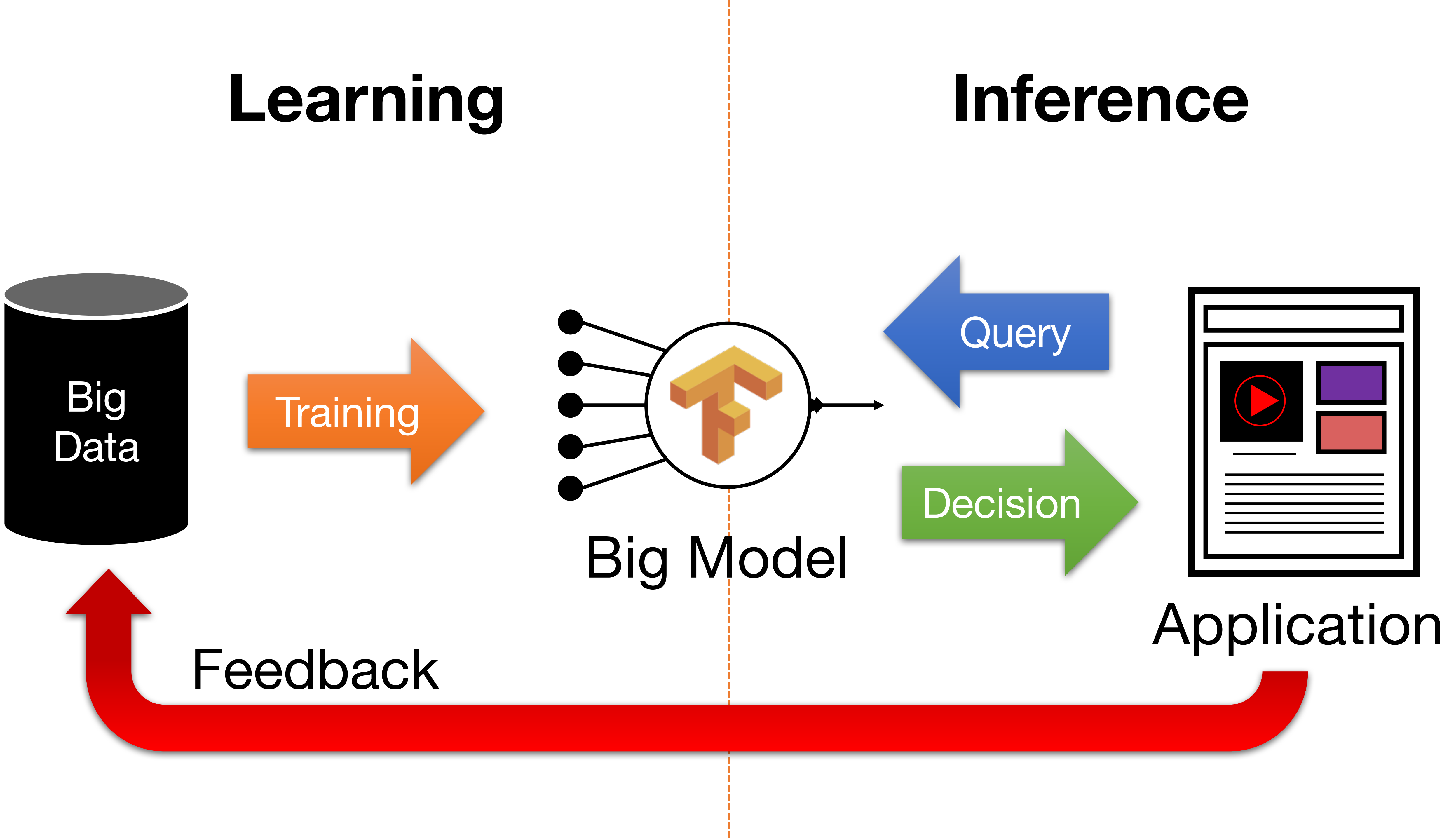
Inference



Often **overlooked**

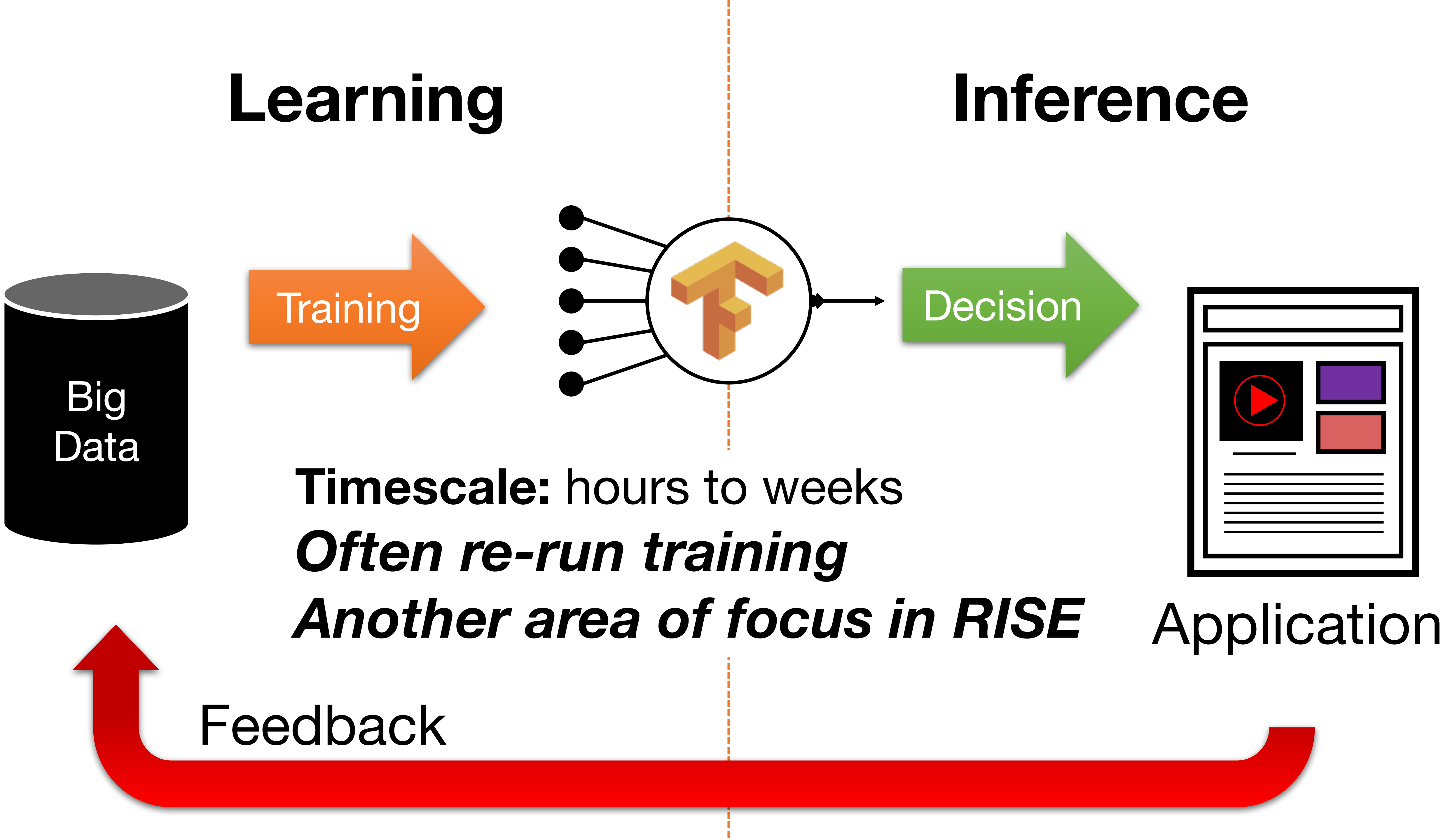
Timescale: ~10 milliseconds

A focus in the RISELab


Learning
Inference
Big
Data
Training
Big Model
Query
Decision
Application
Feedback

Learning

Inference



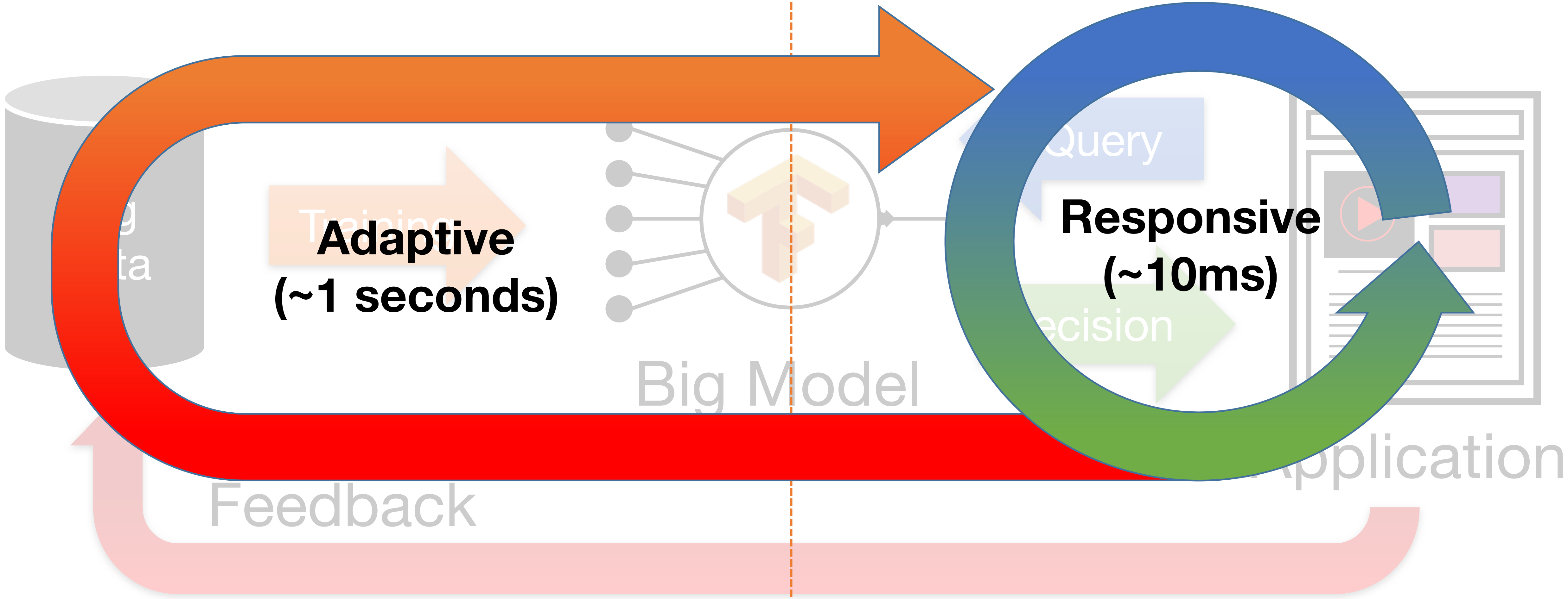
Timescale: hours to weeks

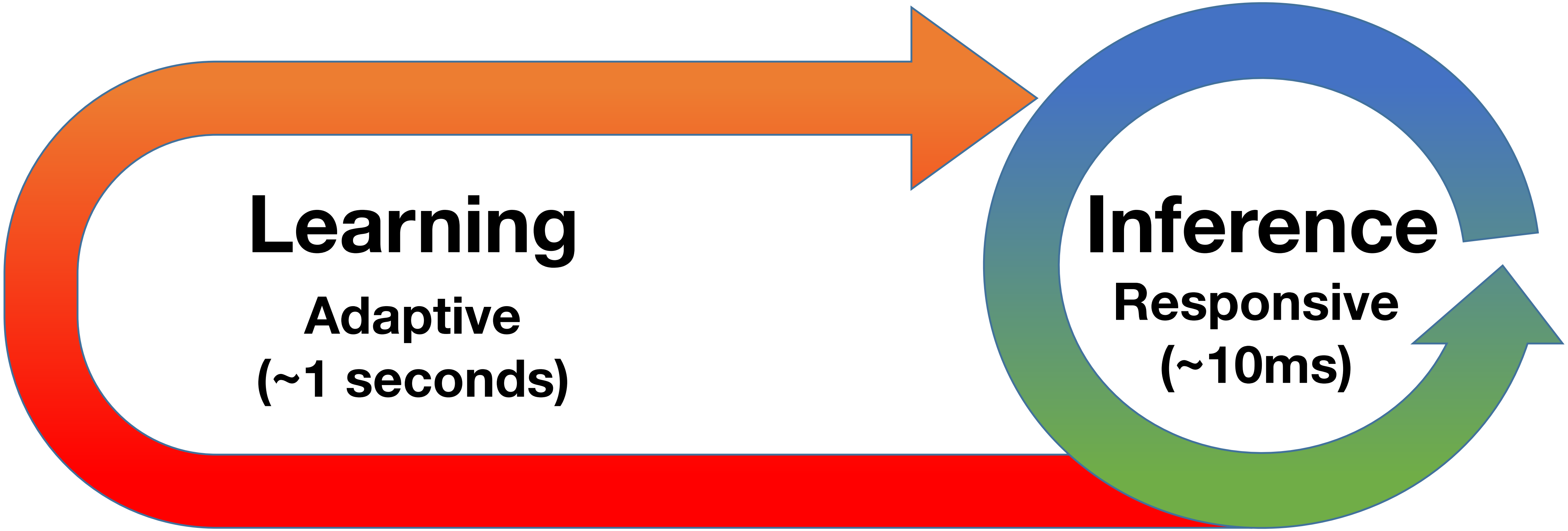
Often re-run training

Another area of focus in RISE

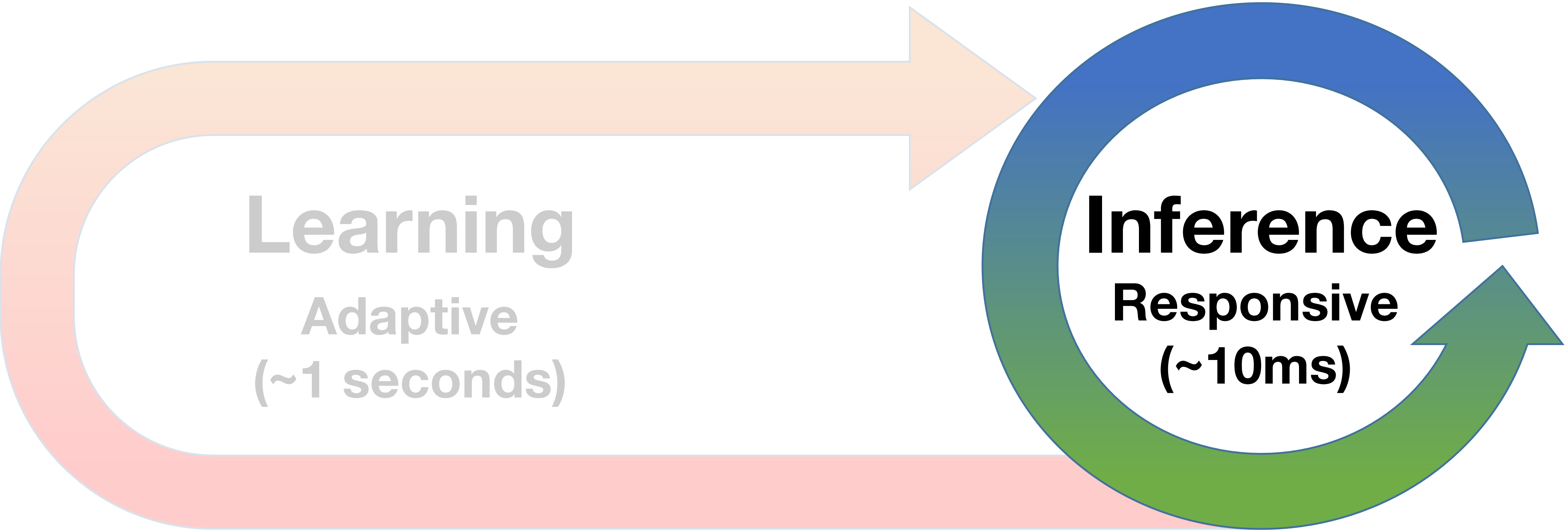
Learning

Inference





The focus of **Learning Systems** in **RISE**

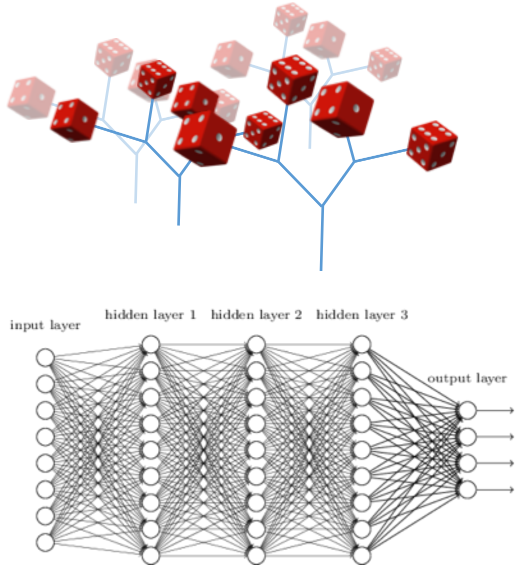


The focus of Learning Systems in RISE

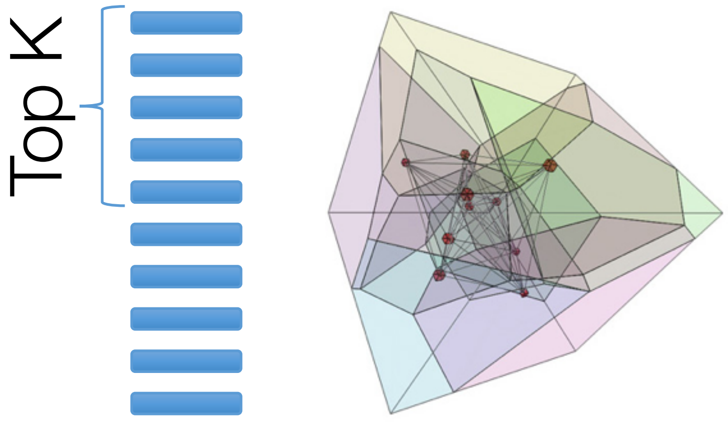
why is **Inference** challenging?

Need to render **low latency** (< 10ms) predictions for **complex**

Models



Queries



Features

SELECT * FROM
users JOIN items,
click_logs, pages
WHERE …

under **heavy load** with system **failures**.

Research in scalable **Inference**

Reducing Latency

- **Approximate caching** to address high-dim continuous features
- **Anytime predictions** study the tradeoff between accuracy and time during inference
- **Model compression** to reduce inference costs (memory and CPU)

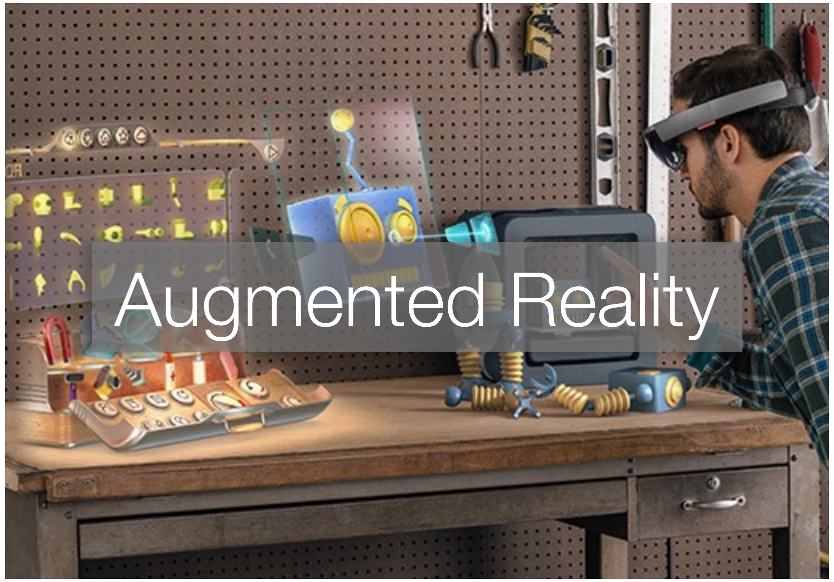
Improving Throughput

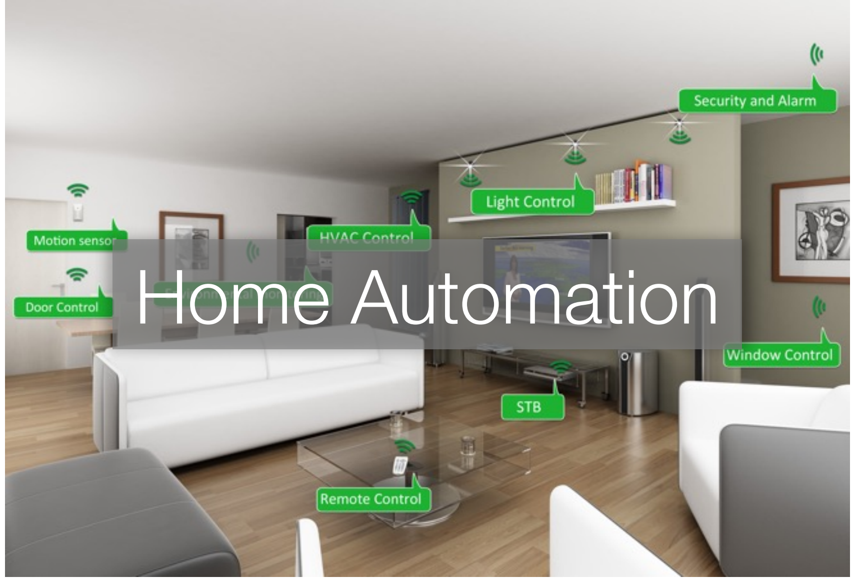
- **Batching technique** to leverage parallel hardware
- **Model cascades** to separate simple and complex queries

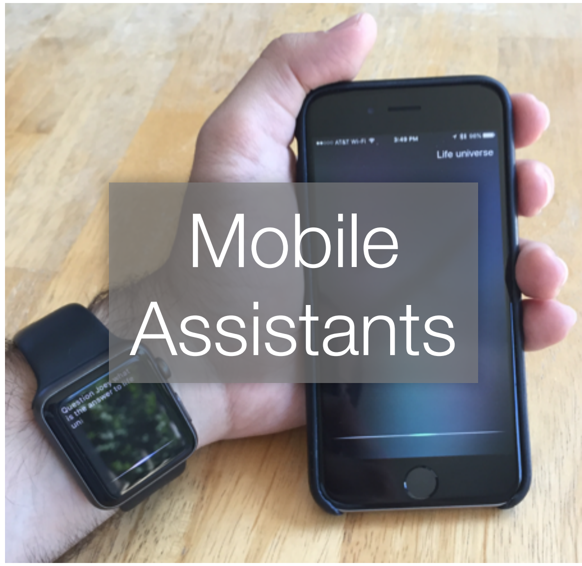
System Failures

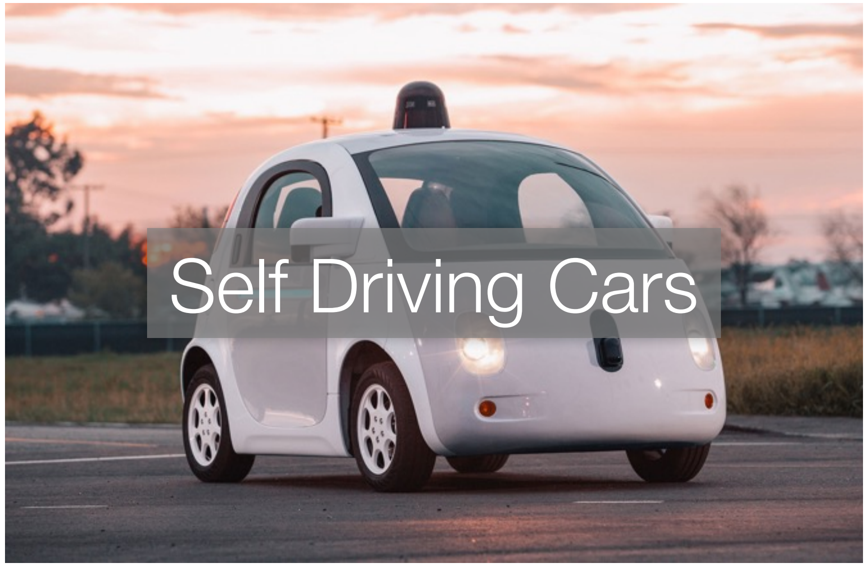
- **Graceful degradation** as models and resources fail
- **Abstractions** to communicate loss of performance to end-user app.

Inference is central to many new apps.









Inference is moving beyond the cloud

Opportunities

- ➢ Reduce latency and improve privacy
- ➢ Address network partitions

Research Challenges

- ➢ Minimize **power consumption**
- ➢ **Limited hardware** & long life-cycles
- ➢ **Protect models** from attack
- ➢ Develop new **hybrid models** to leverage cloud and devices

Robust **Inference** is critical

Self "*Parking*" Cars

Self "*Driving*" Cars

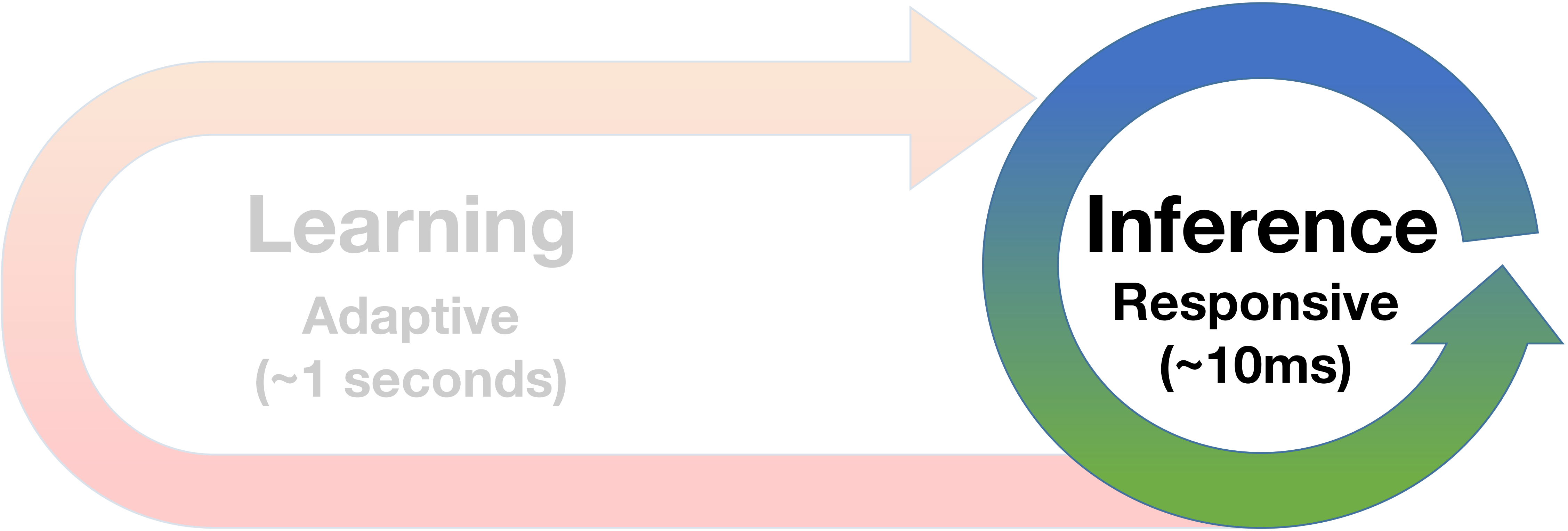
Chat AIs

Research in Robust **Inference**

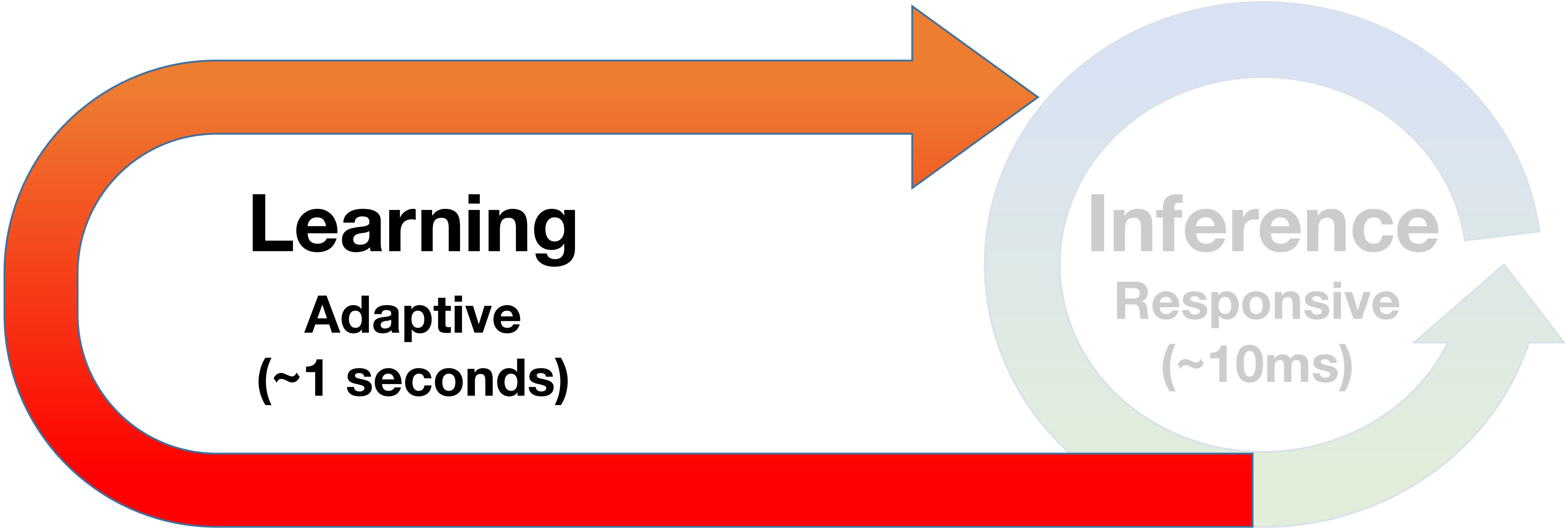
How do we

- identify inputs that are **outside the domain** of the model
  - nighttime images for a daytime model
- recognize **poorly performing models** without feedback
  - e.g., feature and label distribution deviates from training data
- **Calibrate** and **communicate uncertainty** in predictions
  - e.g., ensembles & CIs → increased **overhead** …

at scale with rapidly changing models and data?



The focus of **Learning Systems** in **RISE**



Closing the Loop

The focus of Learning Systems in RISE

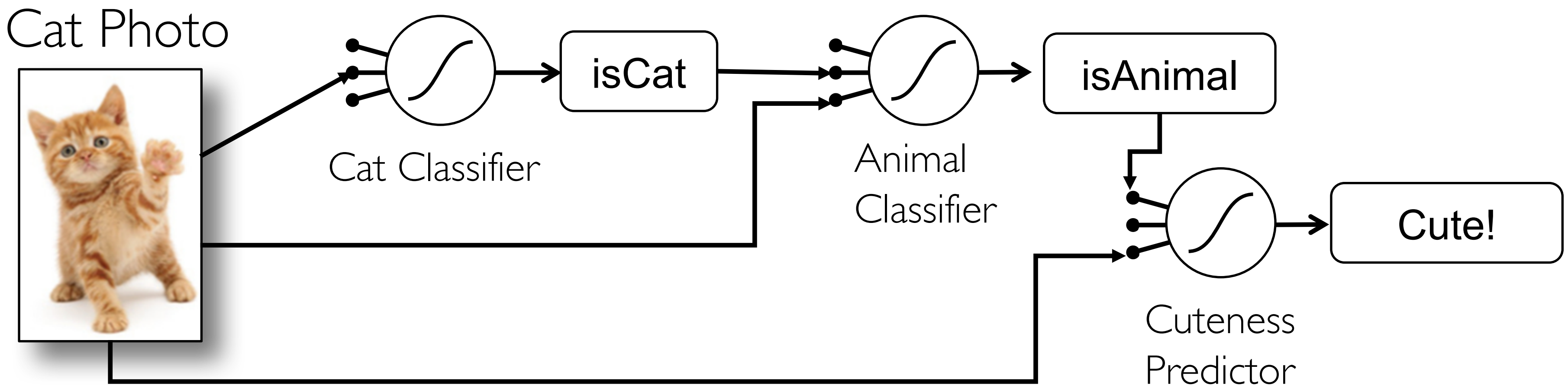
Why is **Closing the Loop** challenging?

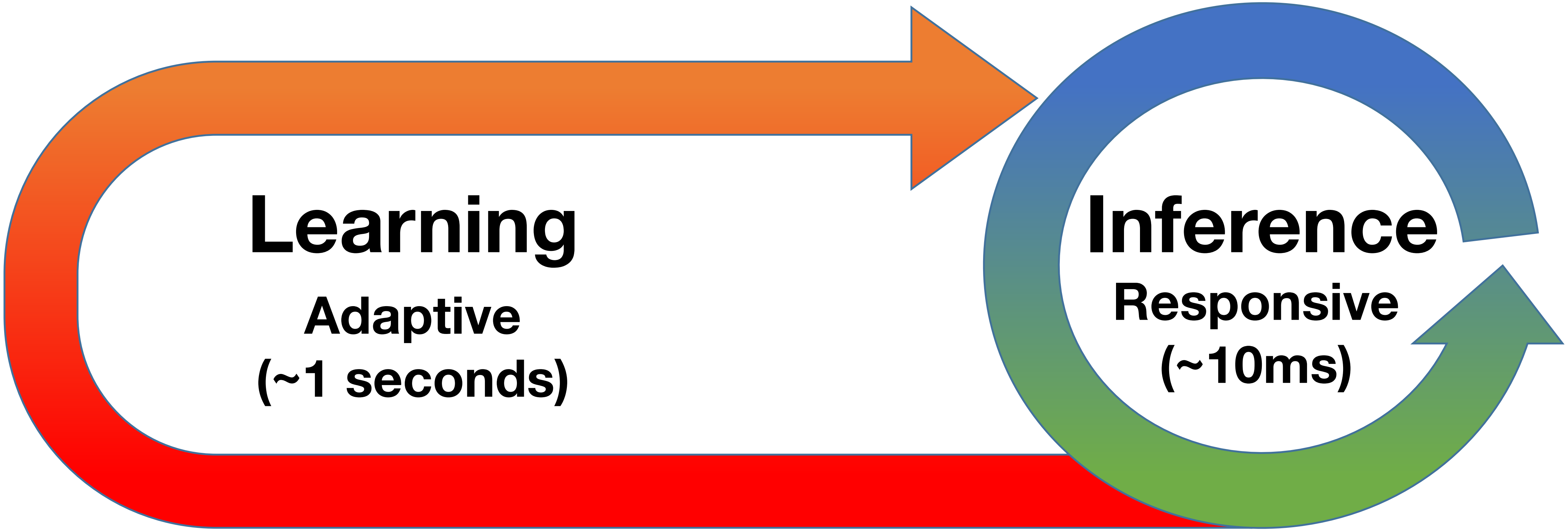
- Combines multiple **systems** with different design goals
  - **Latency** vs **Throughput**
- Exposes system to **feedback loops**
  - *If we only play the top songs how will we discover new hits?*
- Must address **concept drift** and **temporal variation**
  - How do we **forget the past** and **model time directly**
  - **Model complexity** should evolve with data
- Personalization and delayed reward → emphasis on **MTL** and **RL**
- Learning with complex **model dependencies**
- **Robust learning** against **adverserial data**

Feedback and Model Dependencies

How do we:

- automatically **identify feedback** and model **dependencies**?
- distributed learning with bandits: ***theoretical results*** → ***systems***
  - *tradeoff comm., conv., & computational overhead*
- collect sufficient training data for **counterfactual analysis**
- learn with complex **dependencies**:





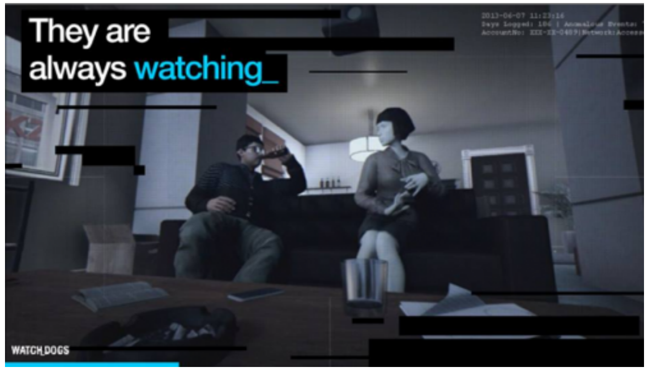
Security

The focus of **Learning Systems** in **RISE**

Security: Protecting Queries

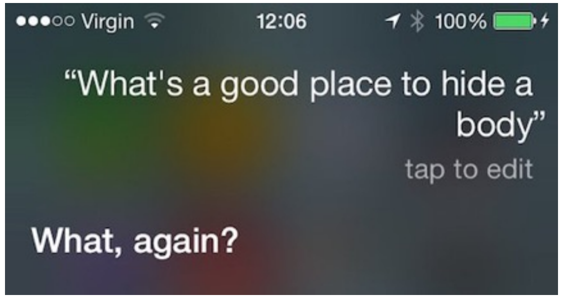
Intelligent systems asked to render predictions on **sensitive queries**.

AR/VR Systems

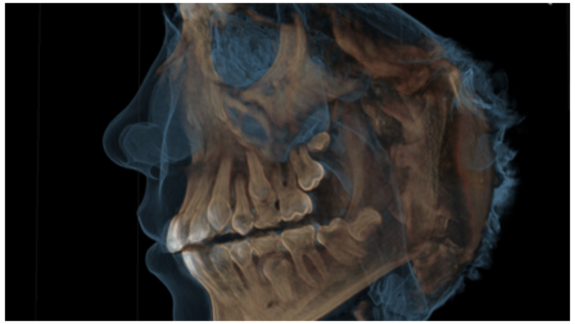


Home Monitoring

Voice Technologies



Medical Imaging



Protect the **query** and **prediction** while hosting models **in the cloud**.

Security: Protecting Models

Data is a core **asset** & models capture the **value** in data

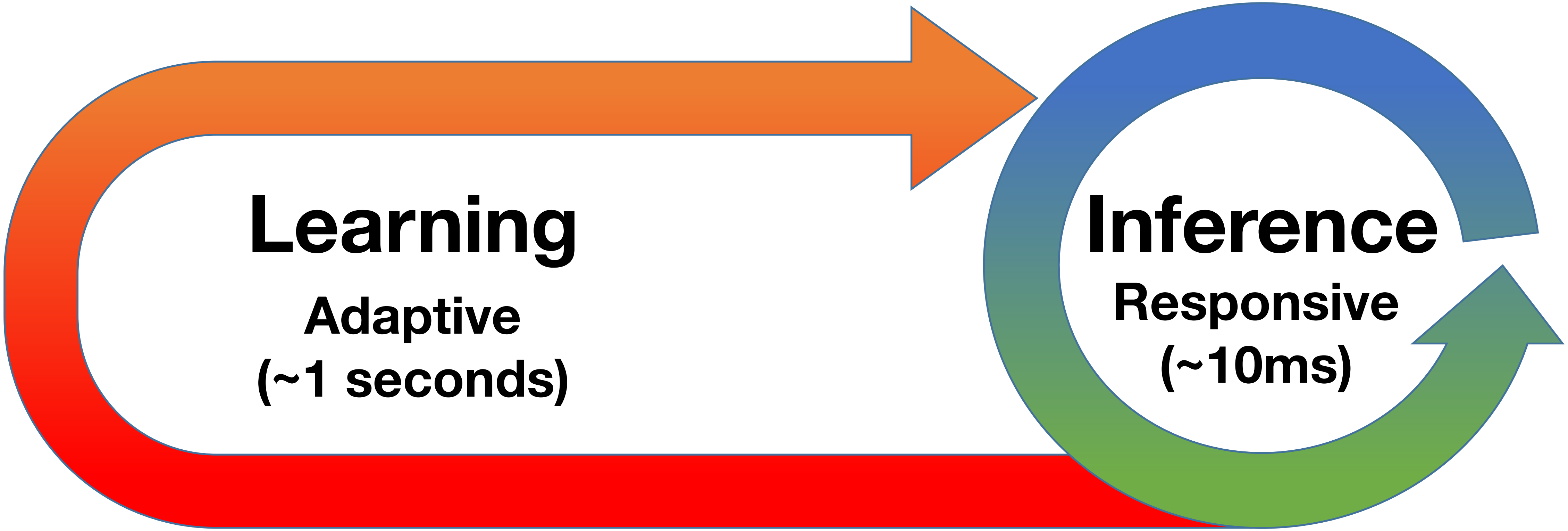
- **Expensive**: many engineering & compute hours to develop
- Models can **reveal private information** about the data

How do we **protect models** from being stolen?

- Prevent them from being copied from devices (**DRM**? **SGX**?)
- Defend against **active learning** attacks on decision boundaries

How do we identify when models have been stolen?

- **Watermarks** in decision boundaries?



The focus of **Learning Systems** in **RISE**

Motivating Example

KUNA

Home video security systems

Technology

- ➢ AC Powered Lamp
- ➢ Commodity ARM processor
- ➢ 720HD Video
- ➢ Microphone & Speaker
- ➢ Infrared Motion Sensors

Goals:

- ➢ Detect, identify, and record people
- ➢ Notify homeowner and open channel of comm.

Similar Technologies

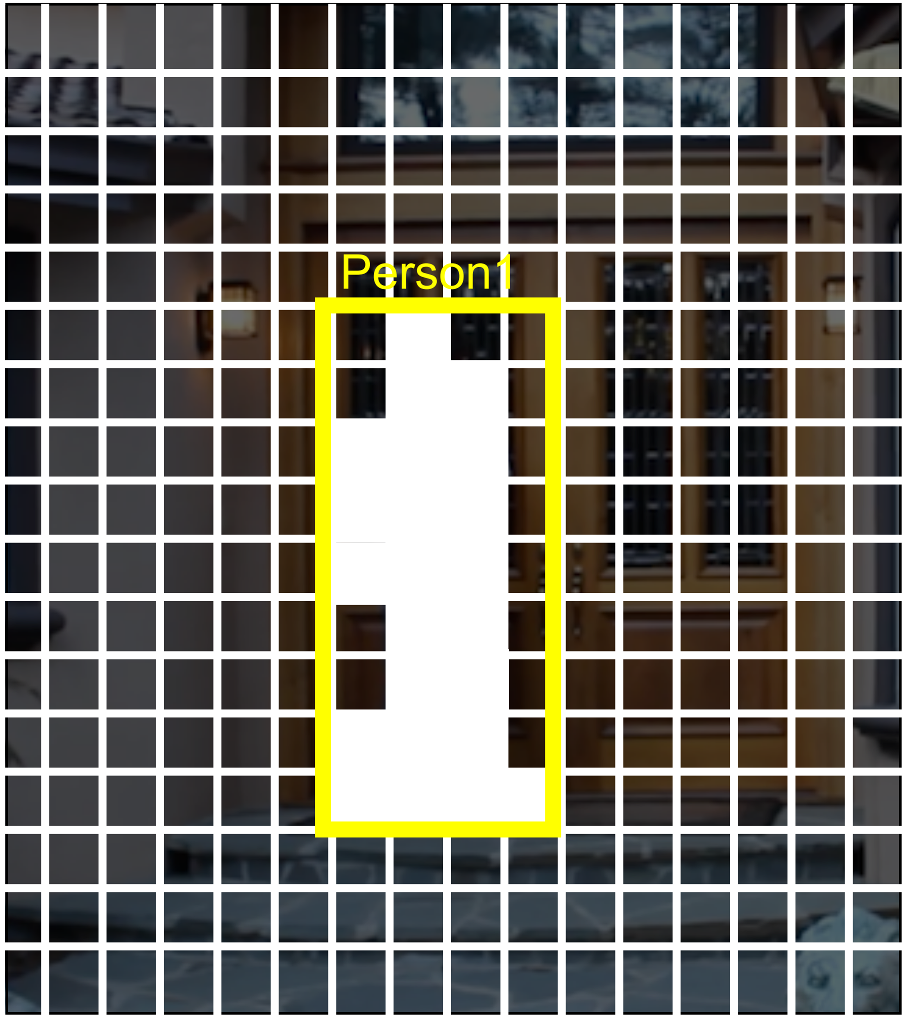
Battery Operated
Wireless Camera System

Powered
Indoor Wireless Camera System

Key challenges

- Need to recognize people and notify home-owner **in real-time**
  - Package delivery → user must connect to camera and talk to person
- Limited, **commodity processors** on devices
  - in some cases (Arlo) **limited power**
- Sending video to cloud is expensive: **$**, **power**, and **bandwidth**
- **Security:** Video stream may contain **sensitive information**
  - records when you leave …
  - a camera in every room …

How does KUNA work?



Fast onboard pixel-level filter identifies suspicious change

Key frames are sent to EC2 for further processing

More sophisticated processing to reduce false positives (**costly GPU time**)

KUNA future technology challenges

- Improved **on-device classification** to reduce false positives that are **processed by cloud**.
- **On-device learning** to identify user specific patterns
  - e.g., the shrub in front of my house moves with the wind
- More **efficient prediction rendering** in the cloud
  - Running full CV pipeline on all images is very costly

Future:

- Event characterization: *“Package delivery at 1:33 PM”*
- Automatic user interaction: *“Hi can I help you …”*

The focus of **Learning Systems** in **RISE**

amplab → **RISELab**

natural progression of research and
an exciting opportunity
to address new challenges

RISELab All Hands Options

Goals:

- Foster **greater collaboration** and improve **research quality**
- Give everyone a **broad perspective** of work in the RISELab
- All hands should be **enjoyable** and **rewarding** (beyond food)

Presentations on:

1. **topic area** overview covering more than one paper
2. **work in progress** with emphasis on getting feedback
3. **proposals** for new projects and finding collaborations
4. **debates** on hot topics in research (competing perspectives)