Recent Developments in Data Science @ Berkeley

& DS-100

Principles and Techniques of Data Science

Joseph E. Gonzalez
Asst. Professor in EECS

jegonzal@berkeley.edu

Recent Developments in Data Science @ Berkeley

DS-100

Principles and Techniques of Data Science

ds100.org

Joseph E. Gonzalez
Asst. Professor in EECS

jegonzal@berkeley.edu

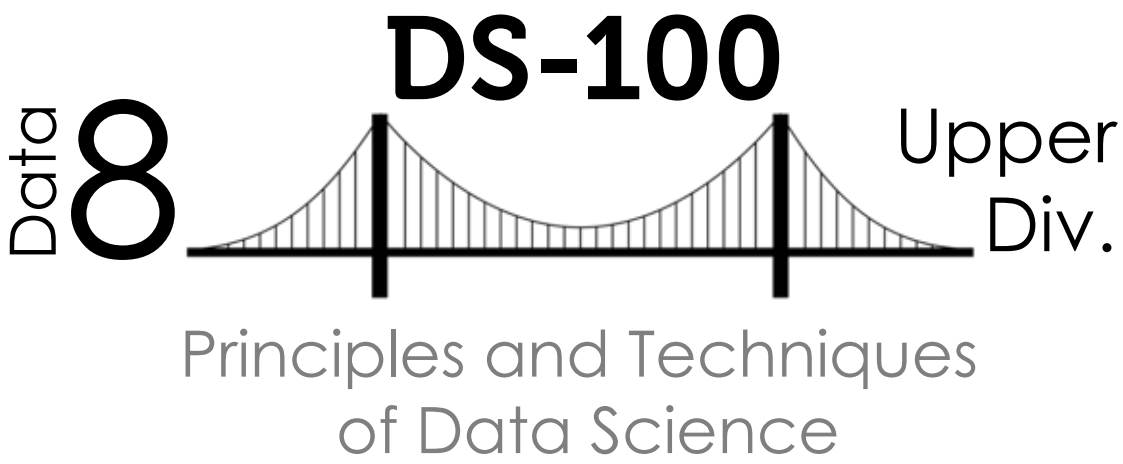
DS-100 Course Development

Started active course development in Spring 2016

- Reference: *CS 194-16 Introduction to Data Science*
  - **Initially:** focused more on tools
  - **Eventually:** fairly advanced many techniques and topics covered

Big Decisions

- Intermediate level
- Narrow scope
- Minimize pre-req. chain
- Strong stats. perspective



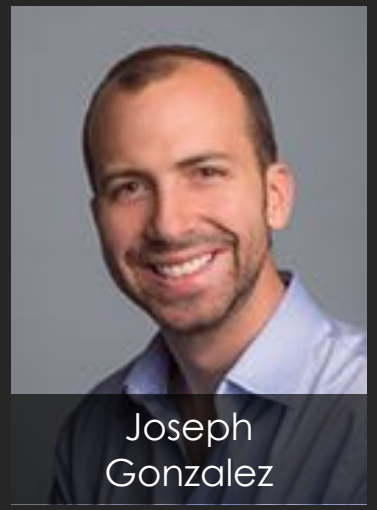
ds100.org

Our Goals

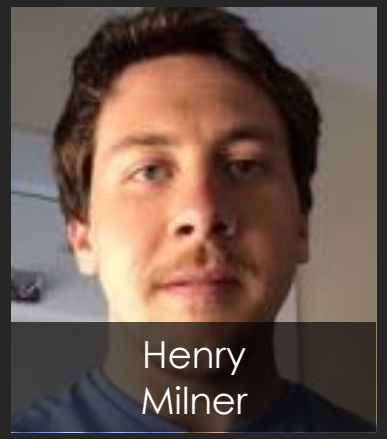
Prepare students for **advanced Berkeley courses** in data-management (CS186), machine learning (CS189), and statistics (Stat-154), by providing the necessary **foundation** and **context**

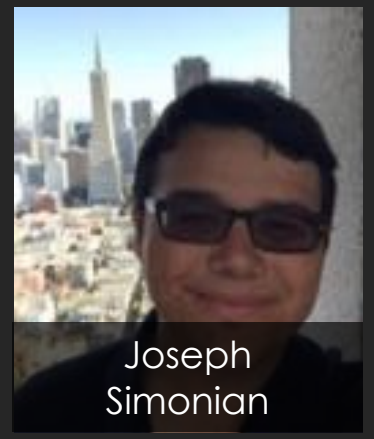
Enable students to start careers as data scientists by providing experience in working with ***real*** **data**, **tools**, and **techniques**

Empower students to apply **computational** and **inferential thinking** to tackle real-world problems



DS100 Created and Taught by Faculty and TAs With Diverse Background & Perspectives



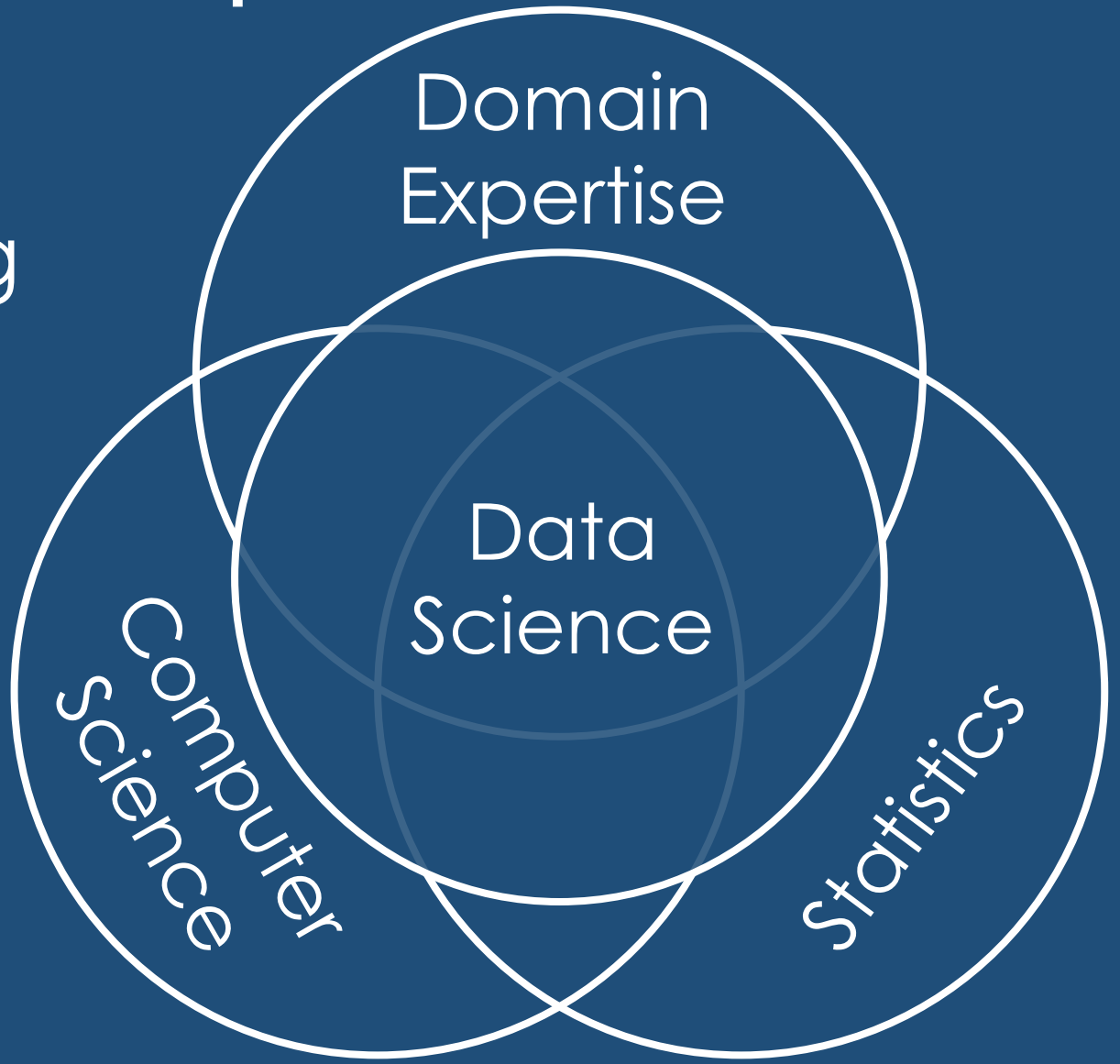


Data Science Requires Many Skills

Can't cover everything in DS100.

Instead we cover

- Key Concepts
  - … some details
- Connections
- How to learn …



Big Concepts in Data Science

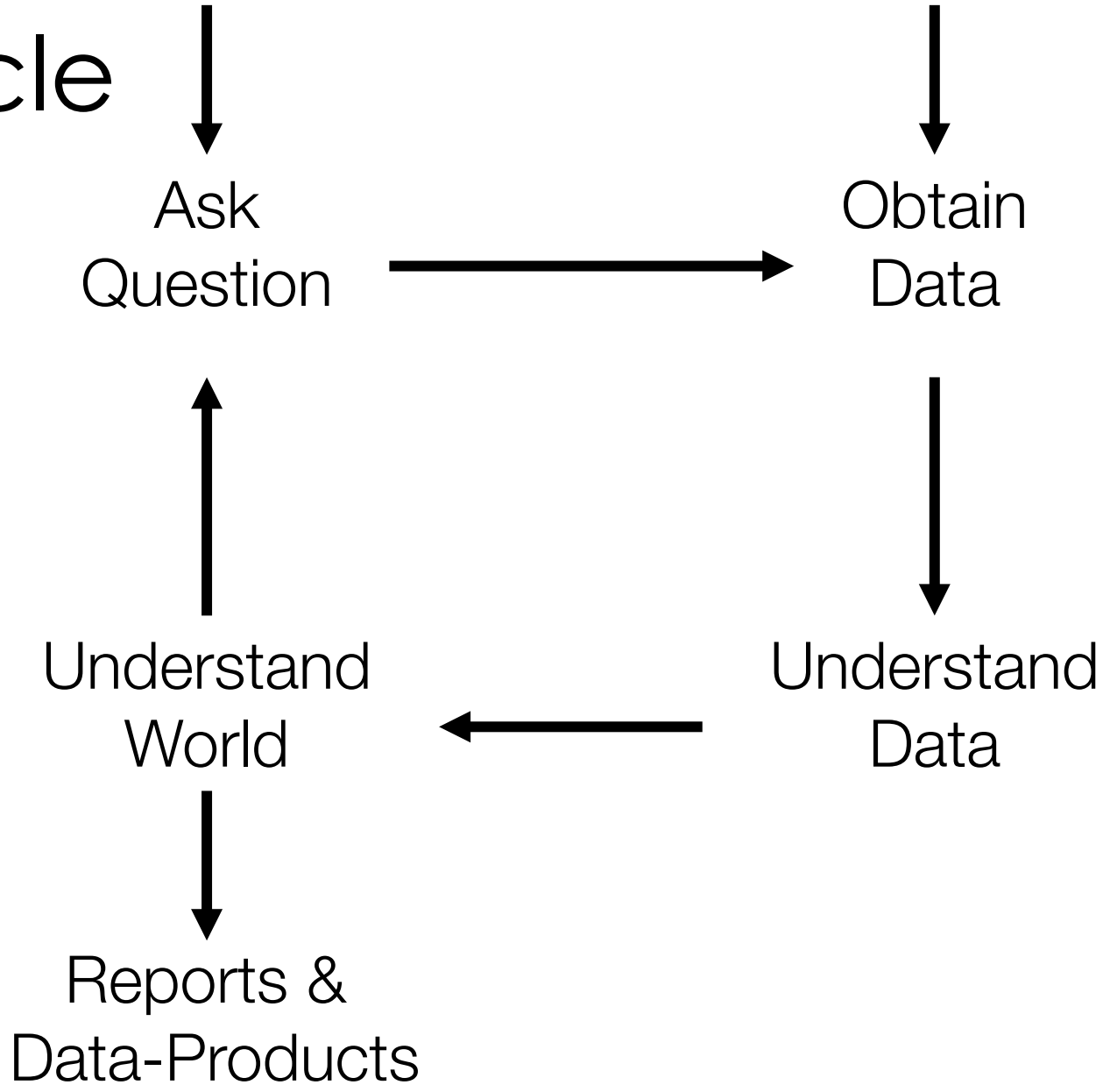
- Data preparation and representation
- Efficient and scalable data processing
- Question formulation and experimental design
- Exploratory data analysis and visualization
- Modeling fitting and inference
- Machine learning techniques and overfitting
- Validation and hypothesis testing

Data Science Lifecycle

High-level description of the data science workflow

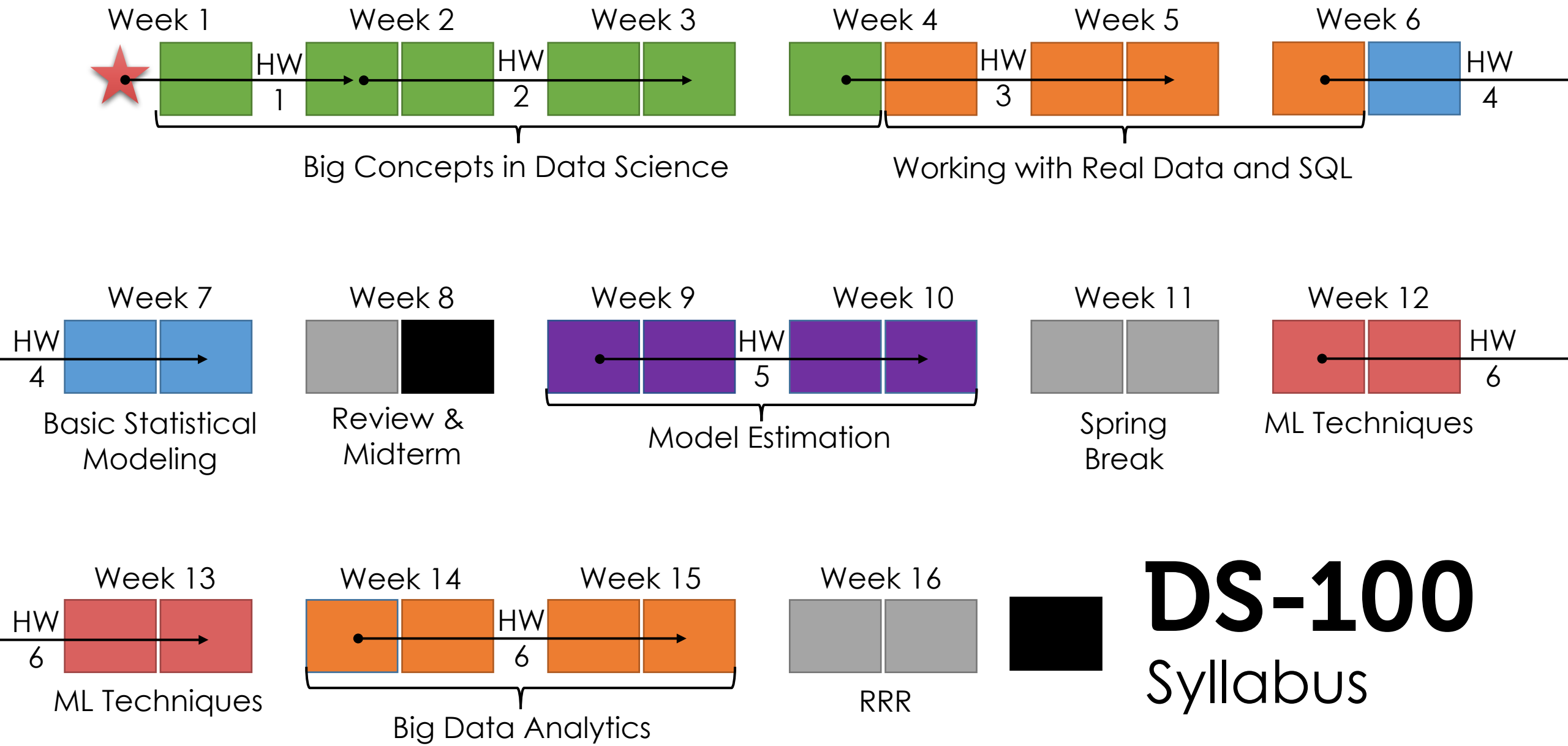
- Frame questions & design experiments
- Obtain and clean data
- Summarize and visualize data
- Inference and prediction

continuous process …



Using Real Tools

- Focus on Python programming language
- We will use various different technologies
  - Jupyter notebooks, pandas, numpy, matplotlib, SQL Server, github, Wrangler, plotly, tableau, Spark?, …
- *We **won't** teach students everything …*
  - Students will learn to **read documentation**
  - Students will learn to **teach themselves**
- ***BETA WARNING:*** things will break …
  - Students will learn how **to debug**
  - Students will learn how **to get help** (Piazza)



DS-100

Syllabus

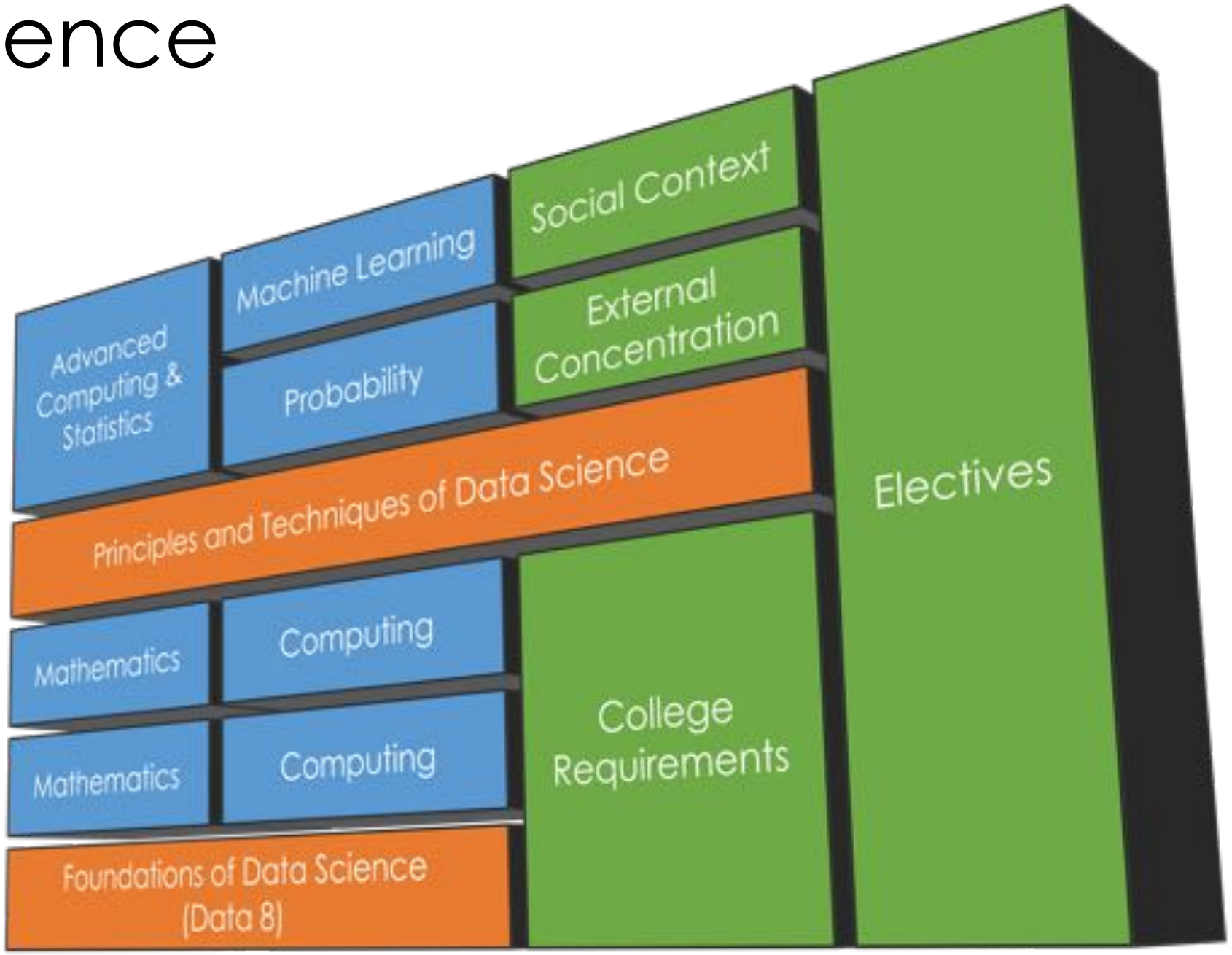
http://www.ds100.org/sp17/syllabus

The Data Science Major …

We are working on it!

Goals

- Interdisciplinary
- Personalized
- Technically deep
- Contextualized
- Pragmatic



& DS-100

Principles and Techniques of Data Science

Recent Developments in Data Science @ Berkeley

Joseph E. Gonzalez
Asst. Professor in EECS

jegonzal@berkeley.edu

ds100.org

Research …

rise.cs.berkeley.edu

Berkeley's lab tradition

- 5-6 Projects addressing big problem
- Bringing faculty from different areas (in CS)

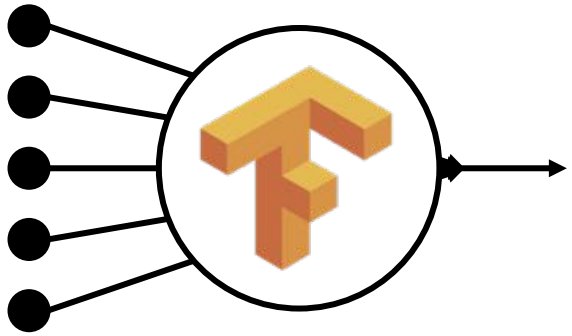
Berkeley's lab tradition

- Working for 5-6 years on a new major problem
- Bringing faculty from different areas

amplab


Big
Data


Training



Big Model


amplab


APACHE
Spark


ML base


KeystoneML

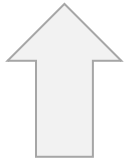

GraphX


Splash


CoCoA

RISE Lab

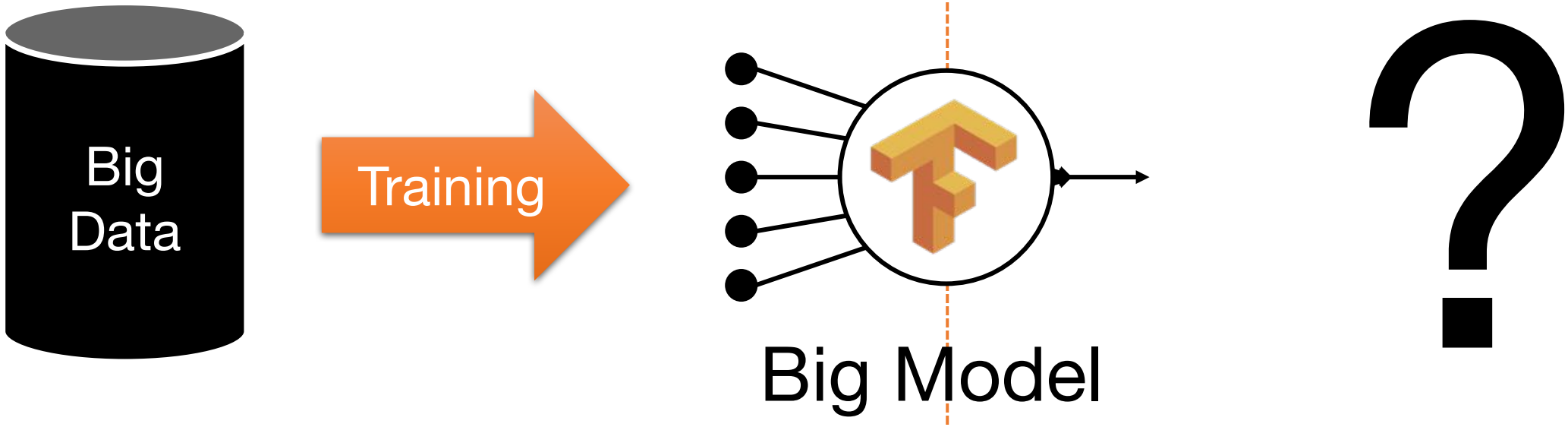
From live data to real-time decisions



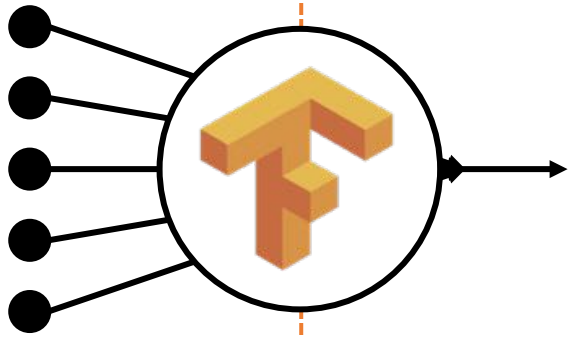
AMP Lab

From batch data to advanced analytics

Learning



Learning

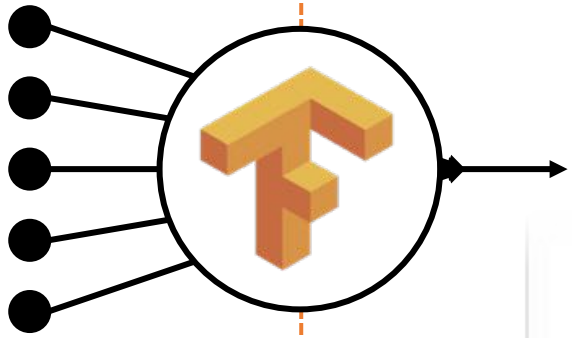


Big Model

Learning

Big Data

Training



Big Model

Conference Papers

Dashboards and Reports

Learning

Big Data

Training

Big Model

Conference Papers

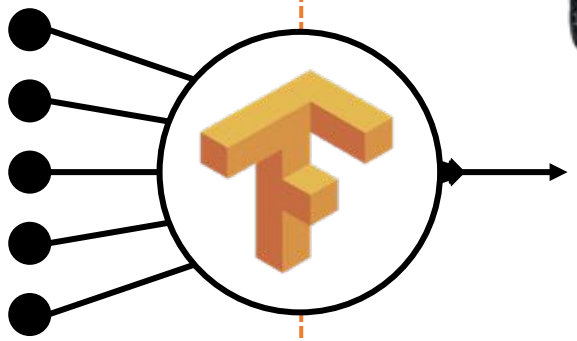
Dashboards and Reports

Drive Actions

Learning

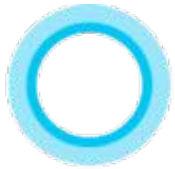
Big Data

Training



Big Model

Drive Actions

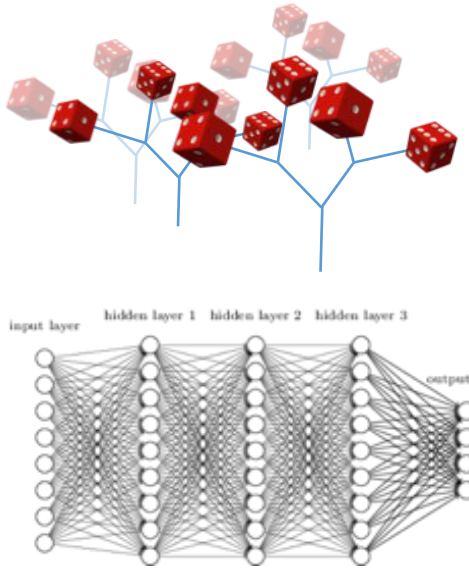


Hi, I'm Cortana.

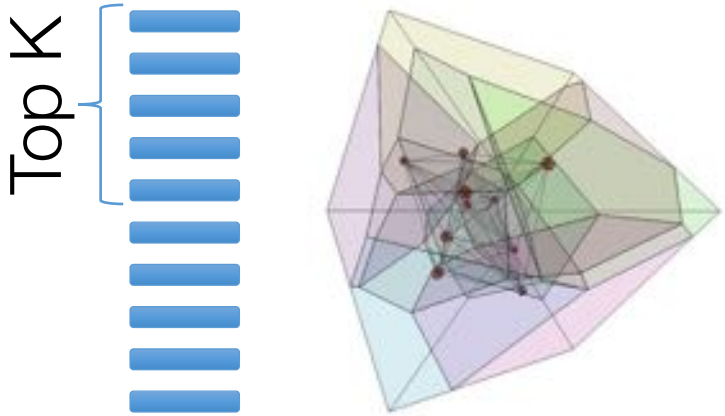
why is **Inference** challenging?

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

Models



Queries



Features

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

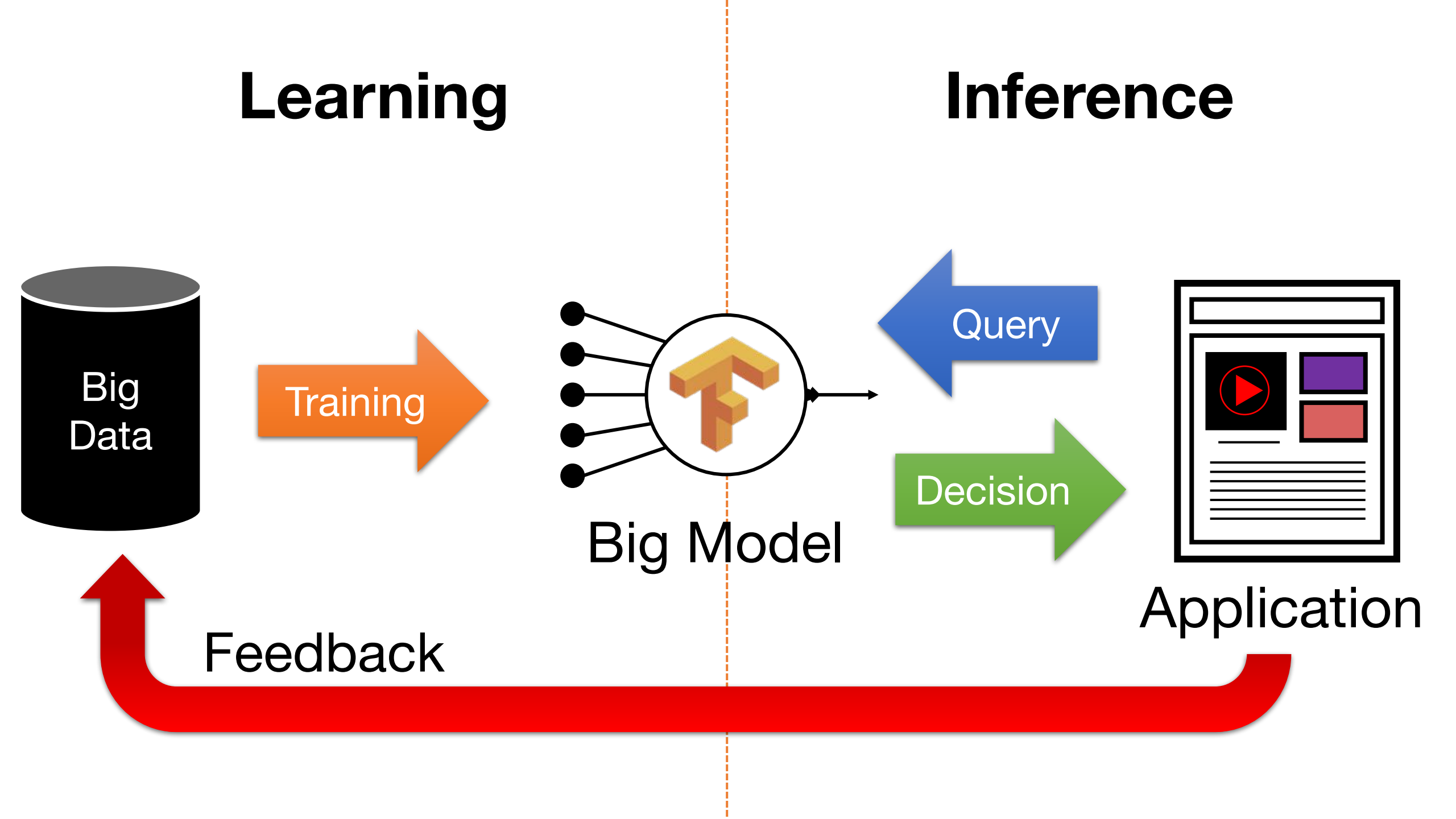
under **heavy load** with system **failures**.

Robust **Inference** is critical

Self "*Parking*" Cars

Self "*Driving*" Cars

Chat AIs

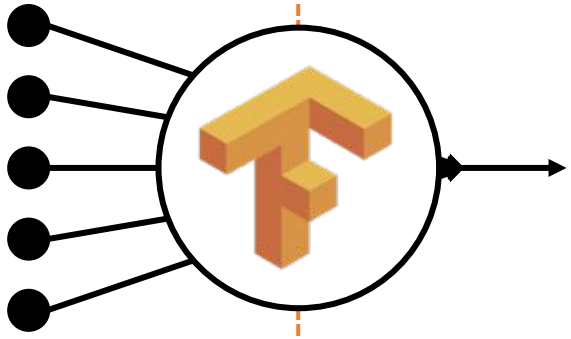

Learning
Inference
Big Data
Training
Query
Decision
Big Model
Application
Feedback

Learning

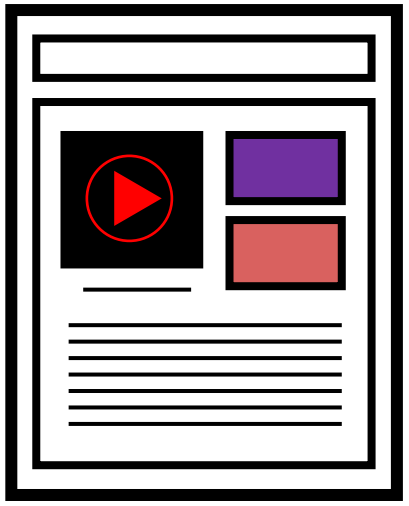
Inference

Big Data

Training



Decision



Application

Timescale: hours to weeks

Often re-run training

Another area of focus in RISE

Feedback

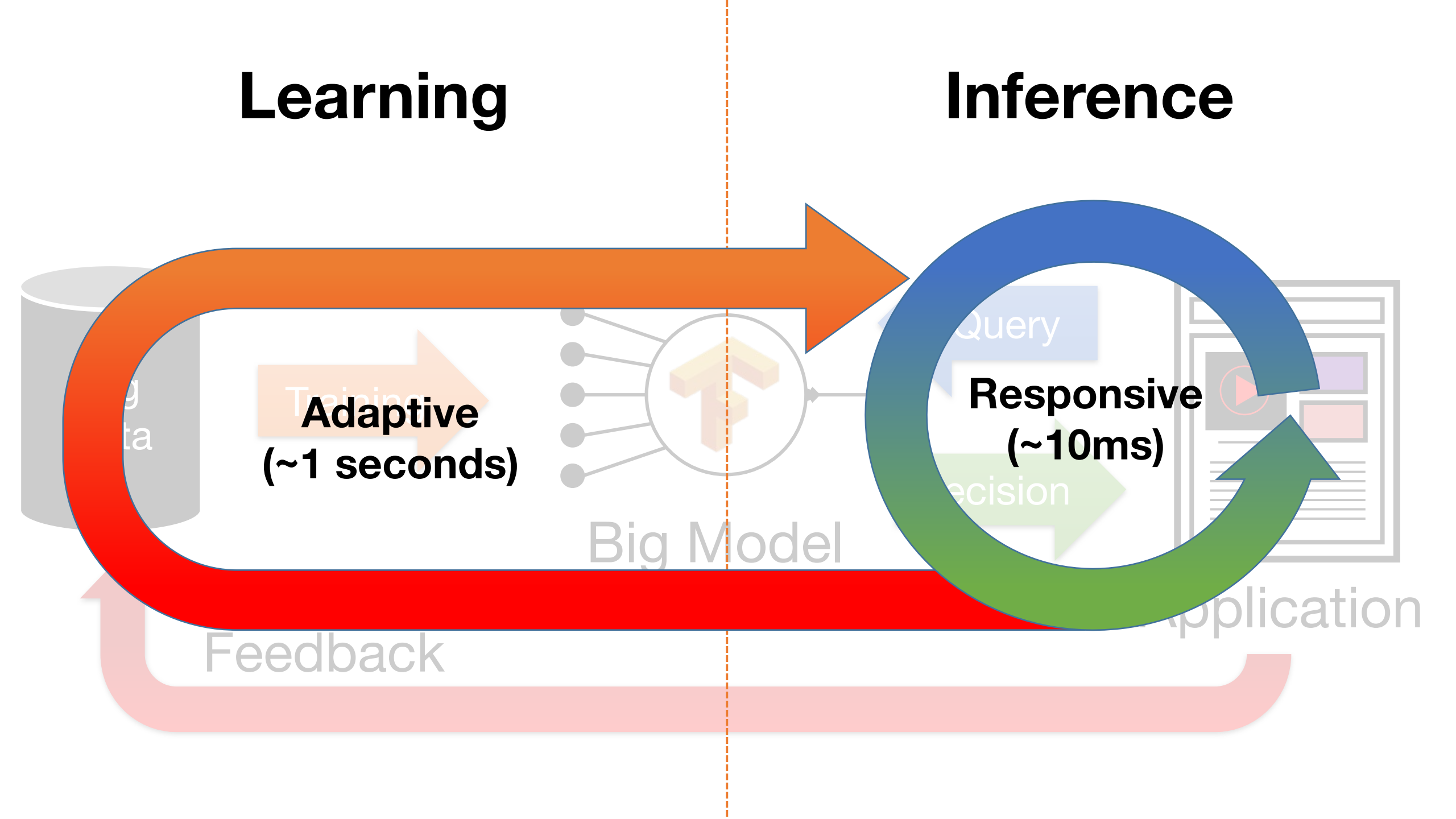
Why is **Closing the Loop** challenging?

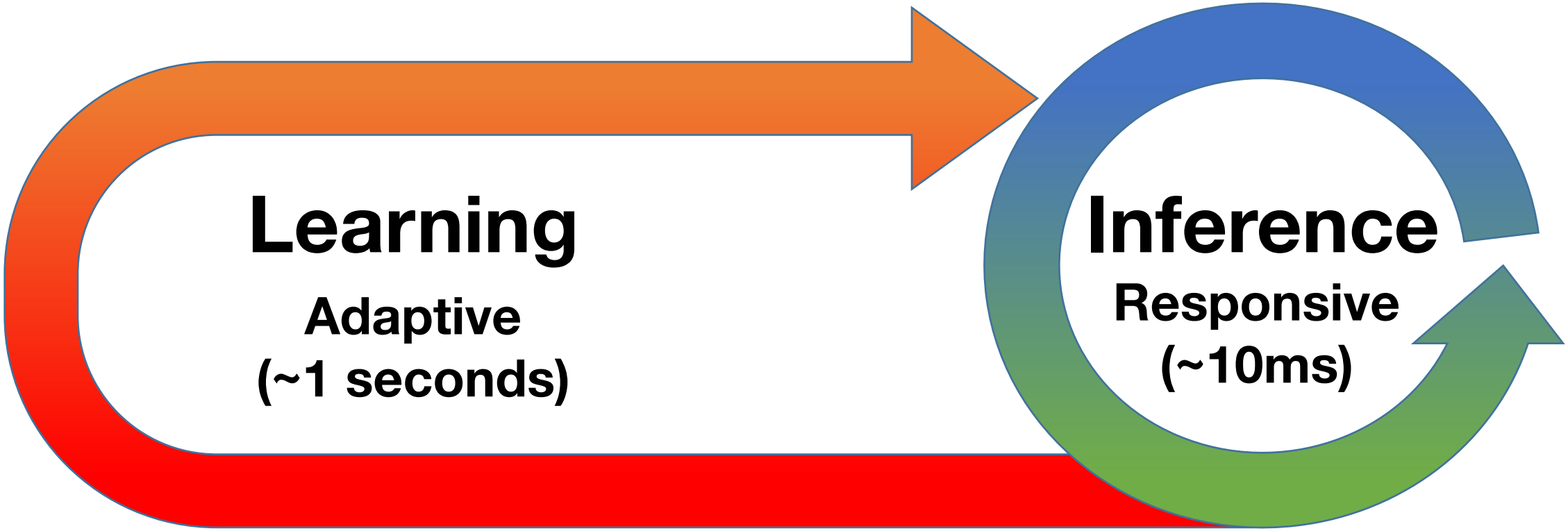
Implicit and **Delayed**
Feedback

Self Reinforcing
Feedback Loops

$$\frac{d}{dt}\left[\right.$$ 

World Changes
at varying rates


Learning
Inference
Adaptive
(~1 seconds)
Responsive
(~10ms)
Big Model
Feedback



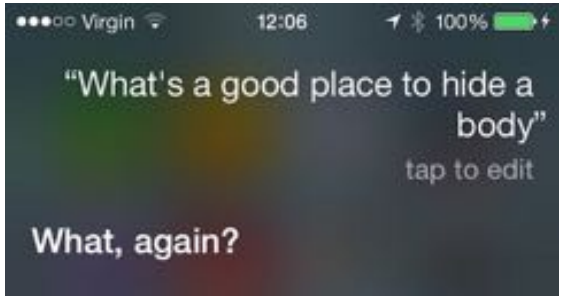
Secure

Intelligence in **Sensitive Contexts**

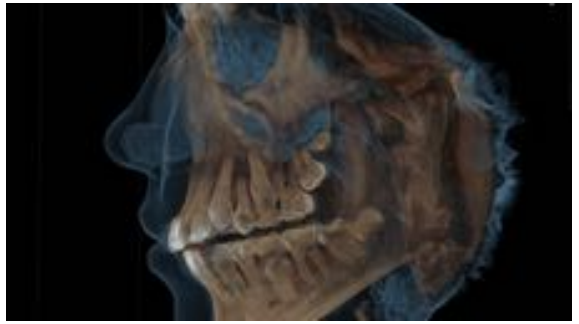
AR/VR Systems

Home Monitoring

Voice Technologies



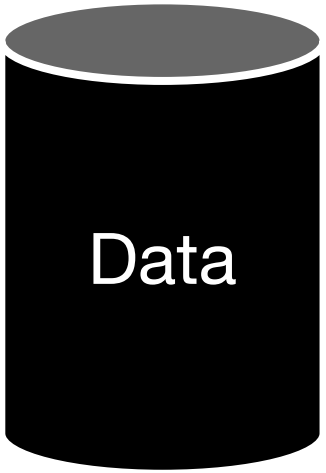
Medical Imaging



Protect the **data**, the **model**, and the **query**

Protect the **data**, the **model**, and the **query**

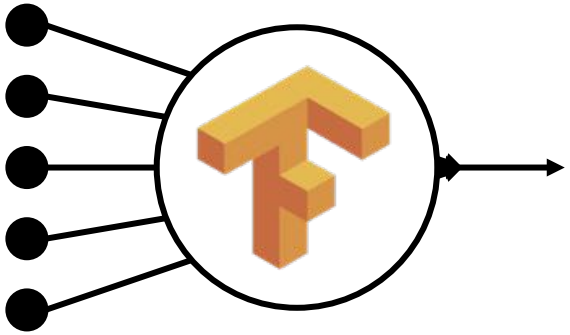
High-Value **Data** *is* ***Sensitive***



- Medical Info.
- Home video
- Finance

Models capture **value** in data

- Core Asset
- Sensitive



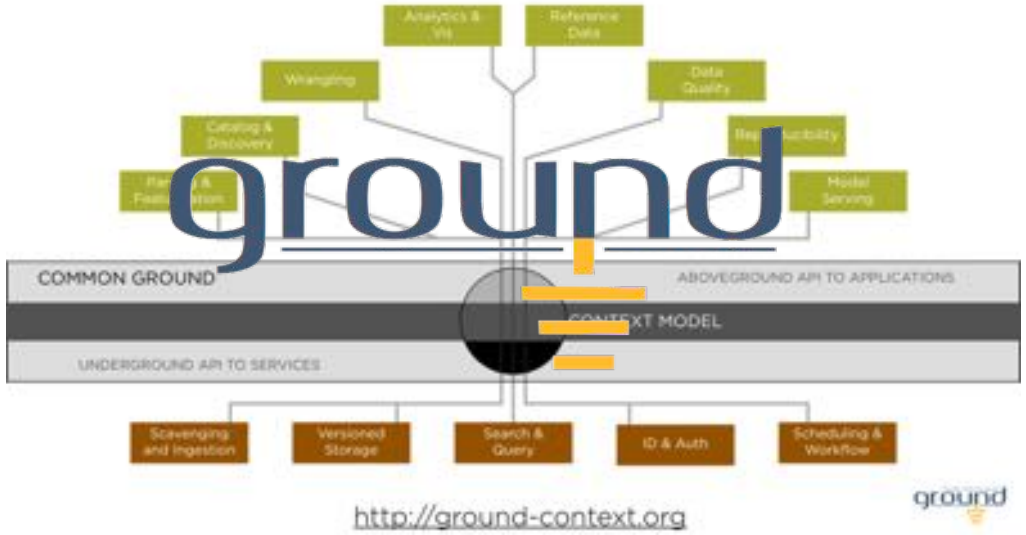
Queries can be as sensitive as the data

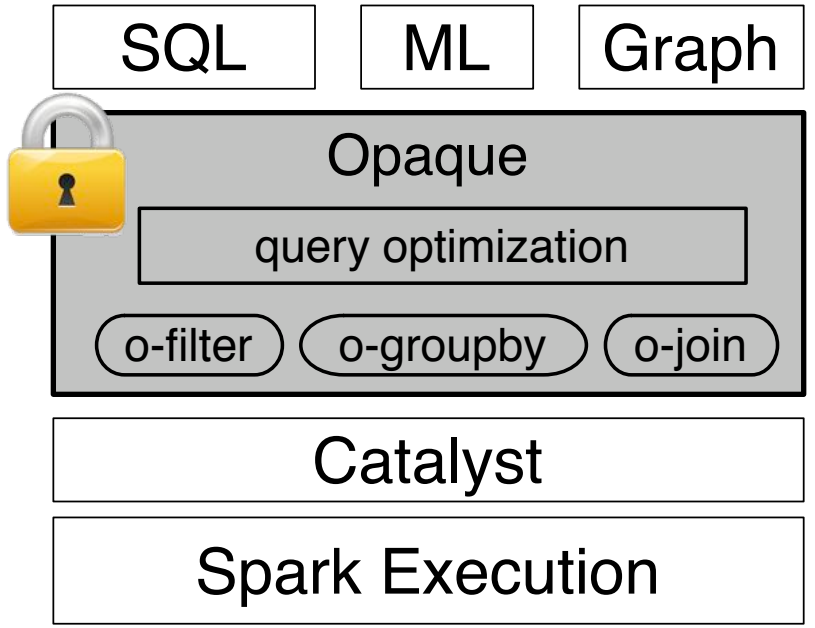
riselab

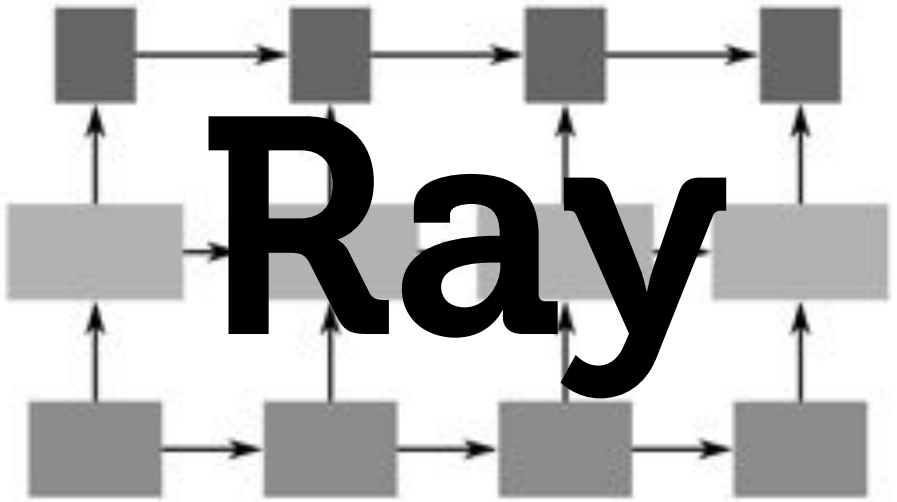
UC Berkeley

We are developing new technologies that will enables applications to make low-latency intelligent decision on live data with strong security guarantees.

A few early projects …







How can I get involved in Research for RISE (or anywhere on campus)

1. Learn about the ongoing projects:
   - https://rise.cs.berkeley.edu
   - https://github.com/ucbrise
2. Email faculty and **grad students** to see if they have openings or are looking for help
   - *Have an idea for the project!*
3. Try to attend seminars hosted by the lab
   - We will start posting these soon!

Joseph E. Gonzalez (jegonzal@berkeley.edu)