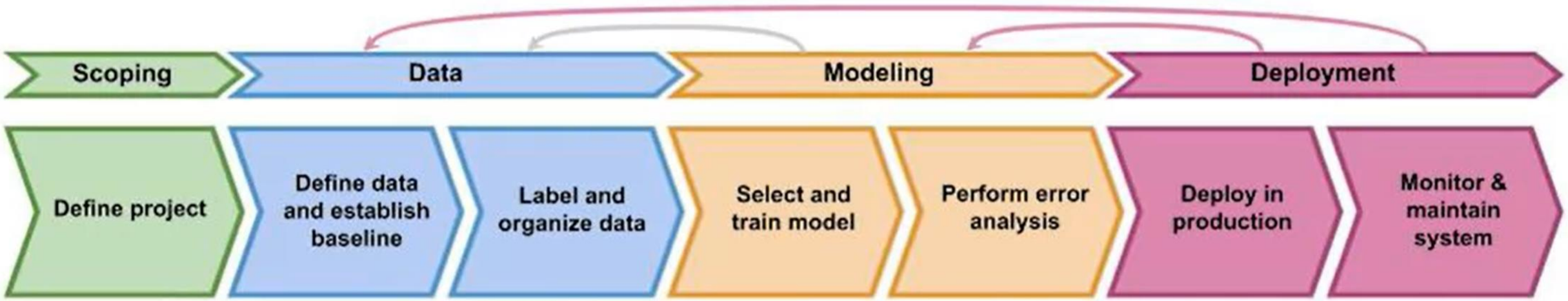
Machine Learning Engineering for Production (MLOps) Specialization


Scoping
Data
Modeling
Deployment
Define project
Define data and establish baseline
Label and organize data
Select and train model
Perform error analysis
Deploy in production
Monitor & maintain system

The MLOps process

- Designing the ML solution
  - Establishing functional and non-functional requirements
- ML Experimentation and Development
  - Delivering a stable ML model to be run in production
- ML Operations
  - Testing, versioning, continuous delivery, and monitoring

Deployment example

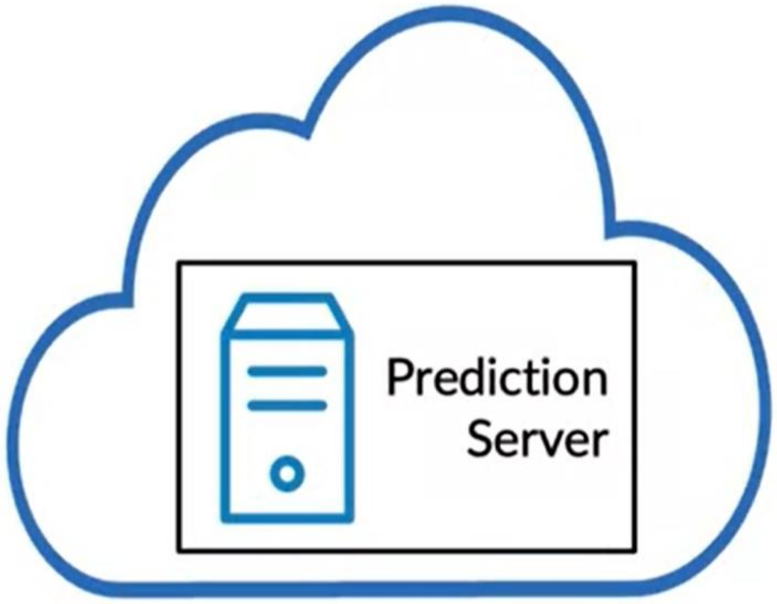
Photo from camera

Deployment example

Photo from camera

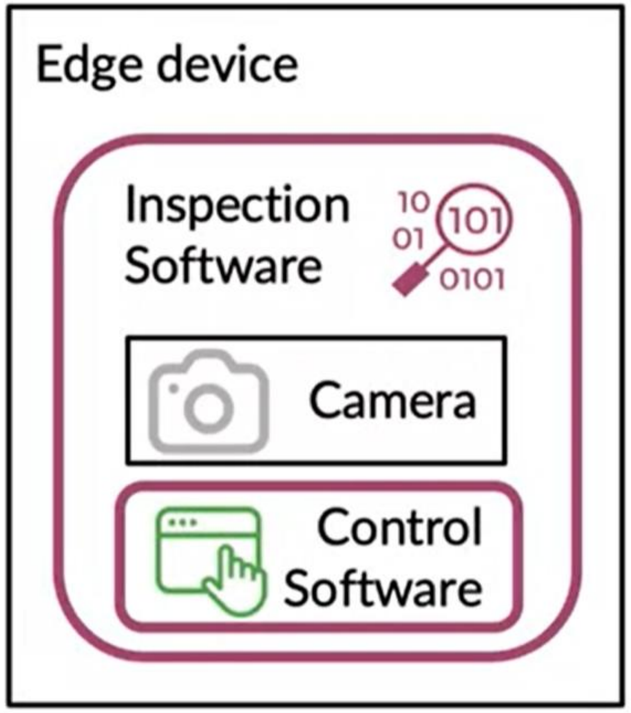
Deployment example

Photo from camera



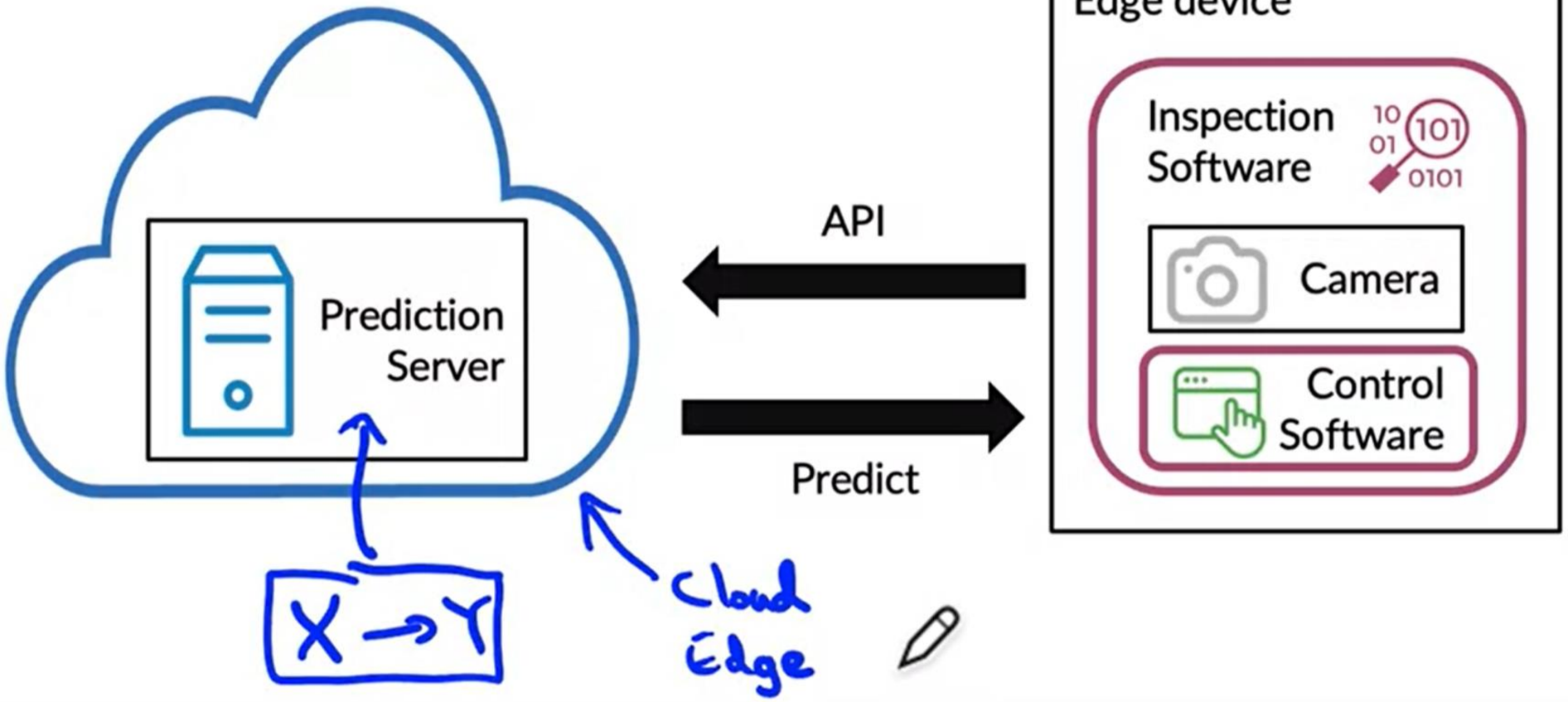
API

Predict



Deployment example

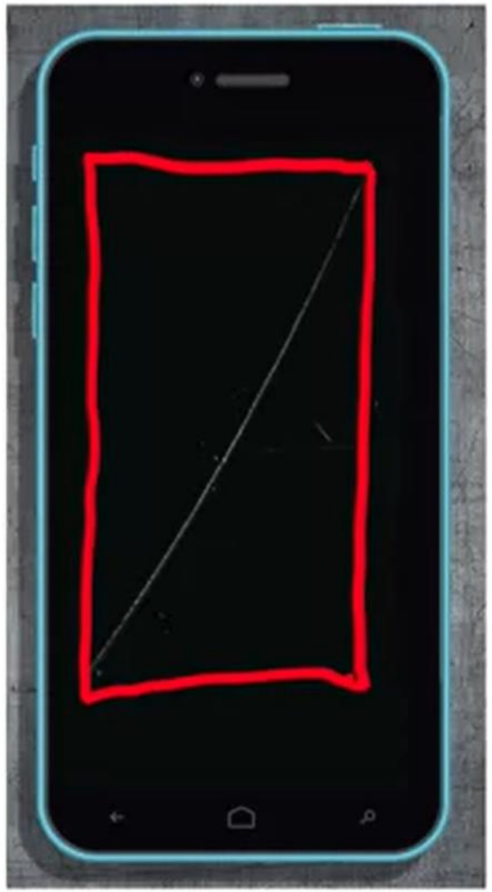
Photo from camera



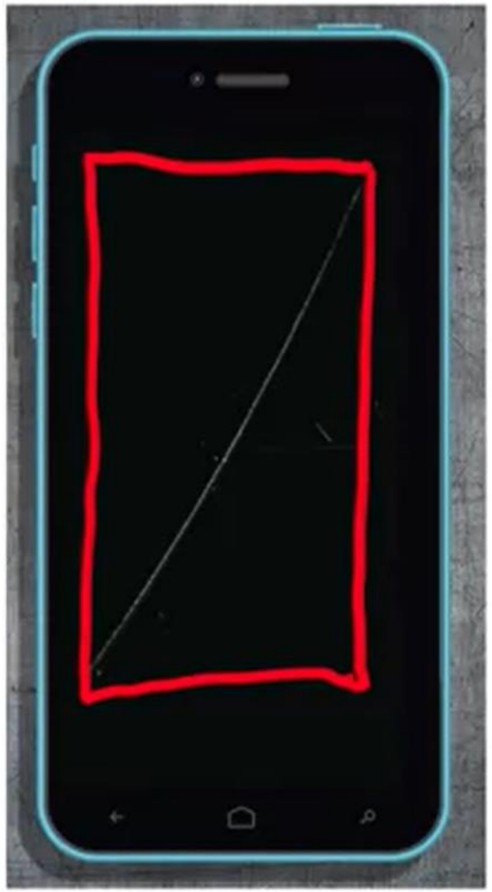
Visual inspection example

Visual inspection example

ok

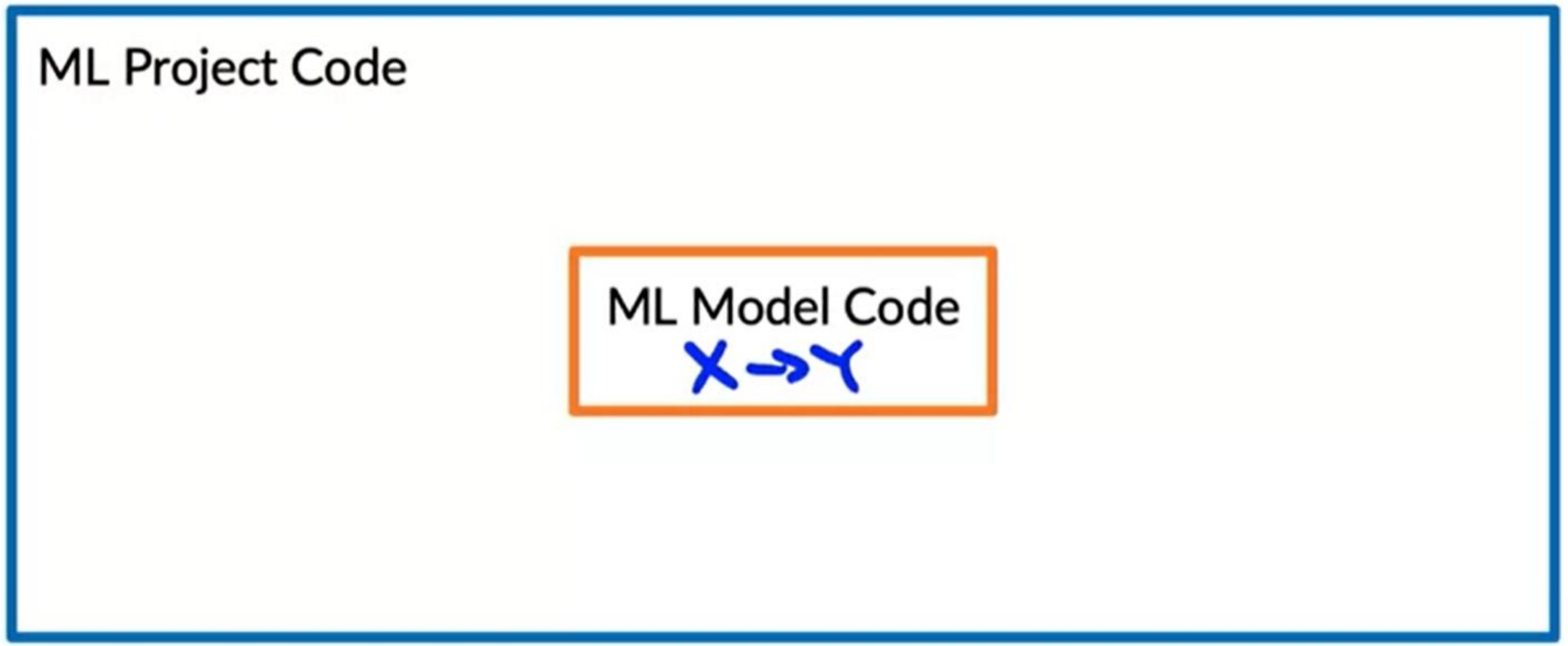


Visual inspection example

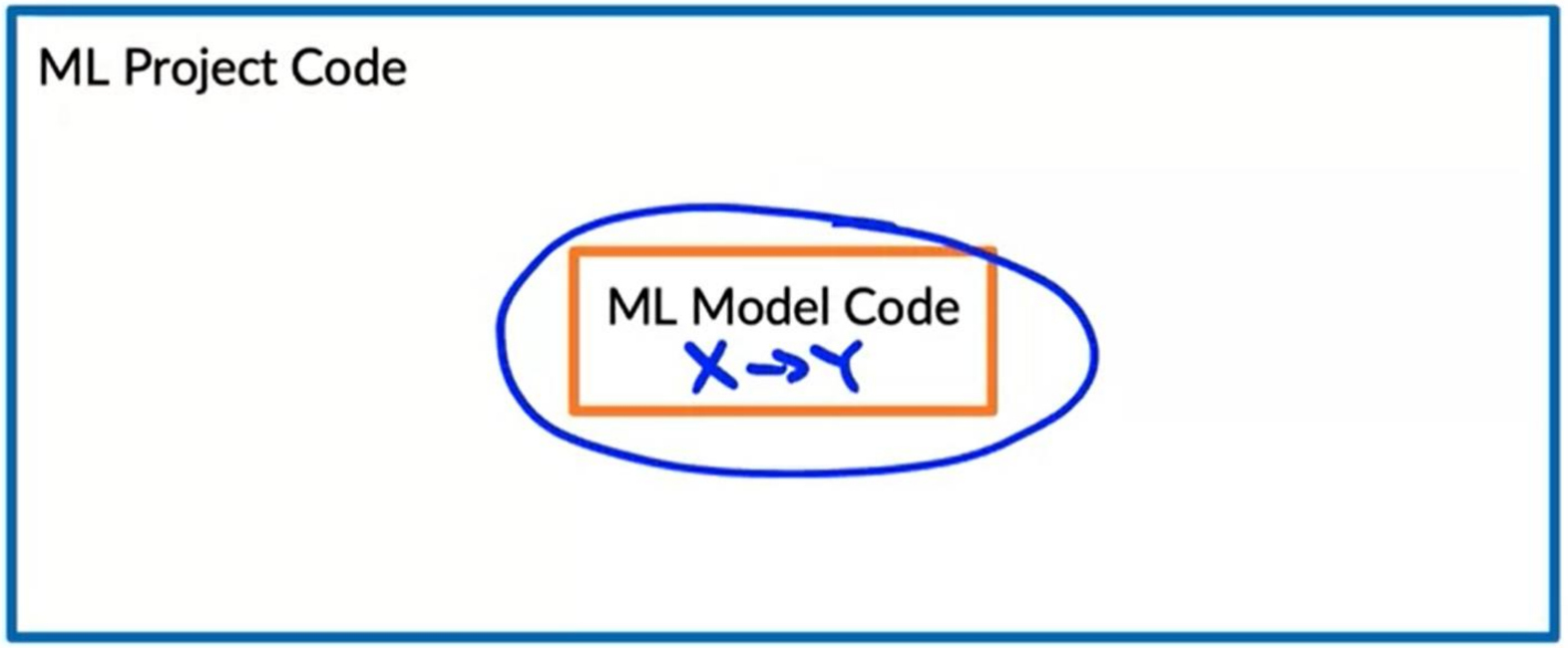


ML in production

ML in production

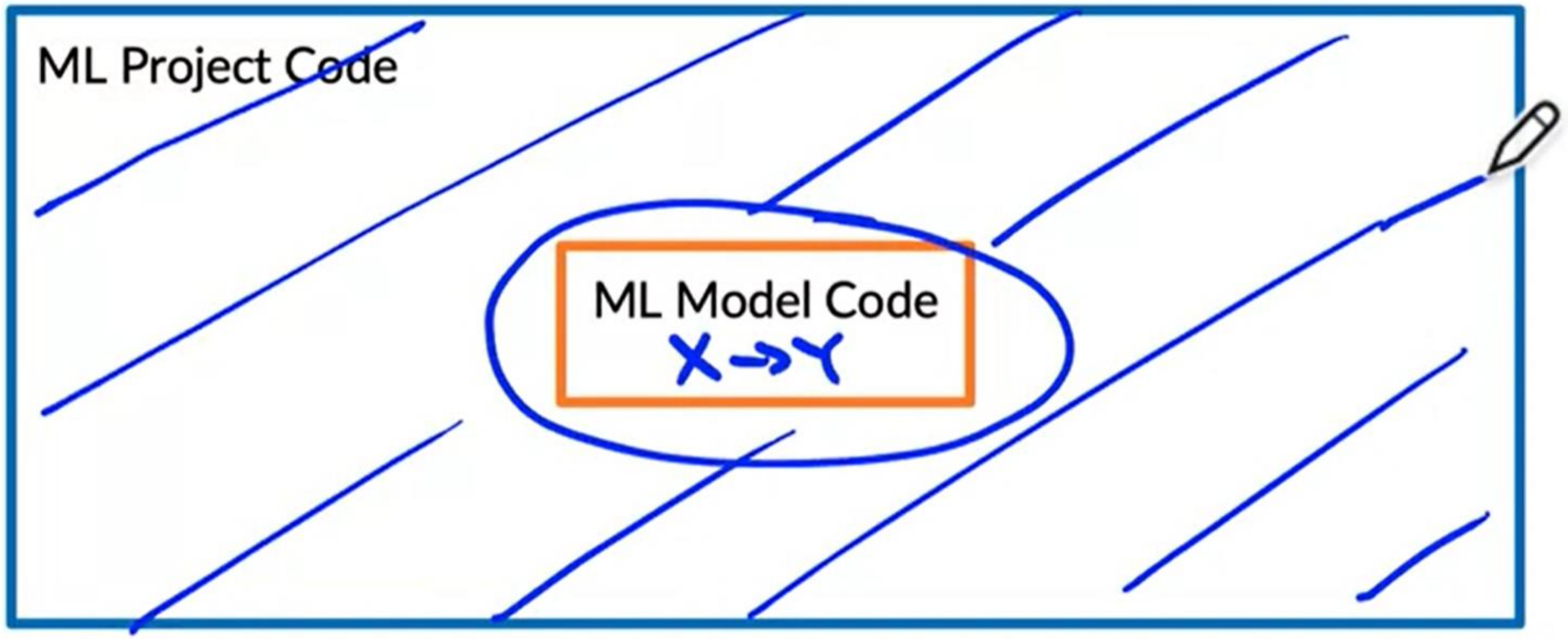


ML in production



"POC to Production Gap"

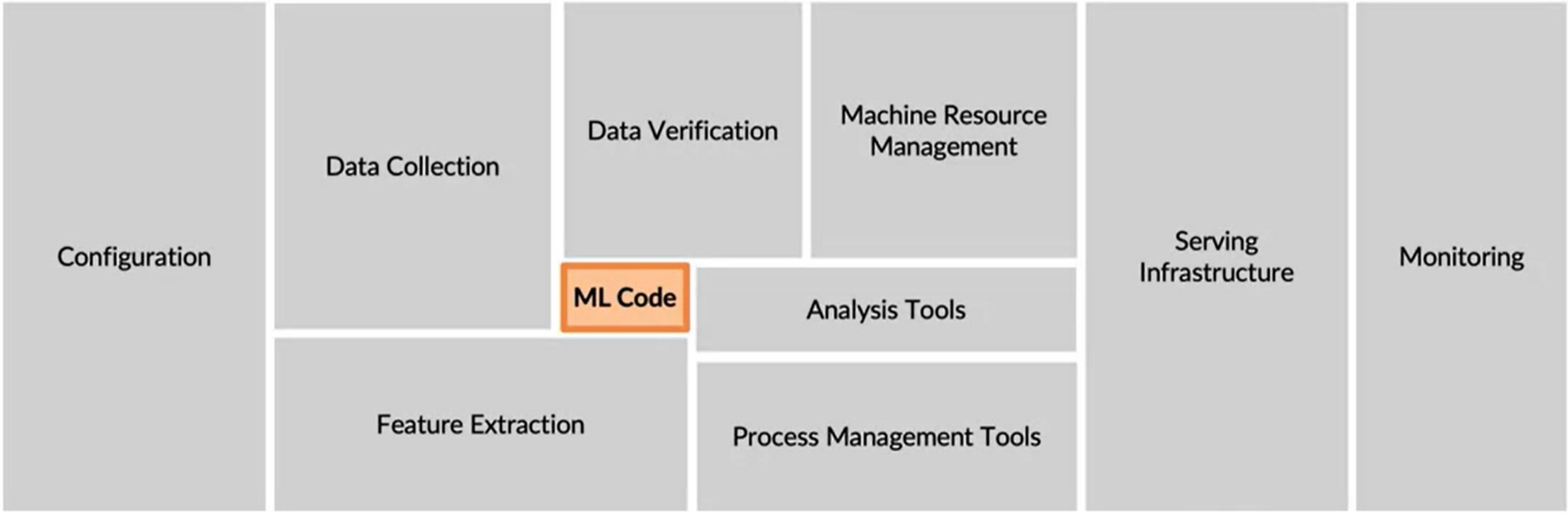
ML in production



5-10%

"POC to Production Gap"

The requirements surrounding ML infrastructure



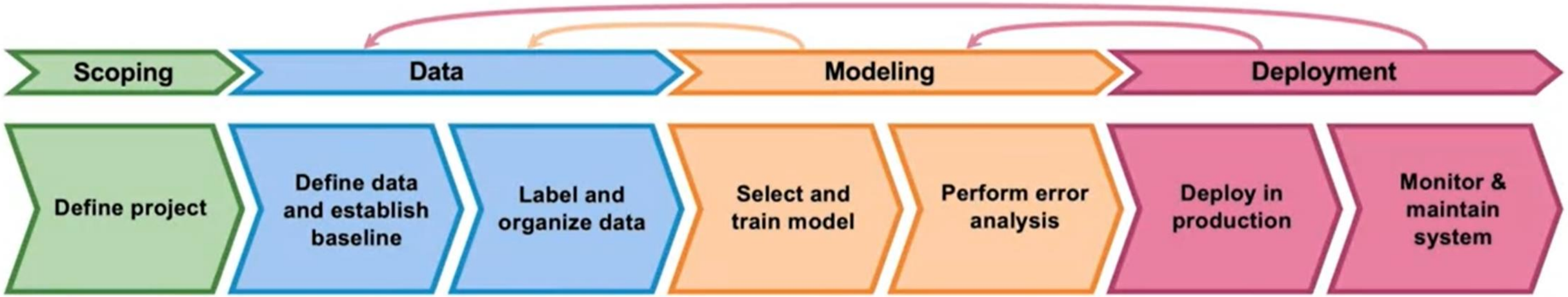
[D. Sculley et. al. NIPS 2015: Hidden Technical Debt in Machine Learning Systems]

DeepLearning.AI

The Machine Learning Project Lifecycle

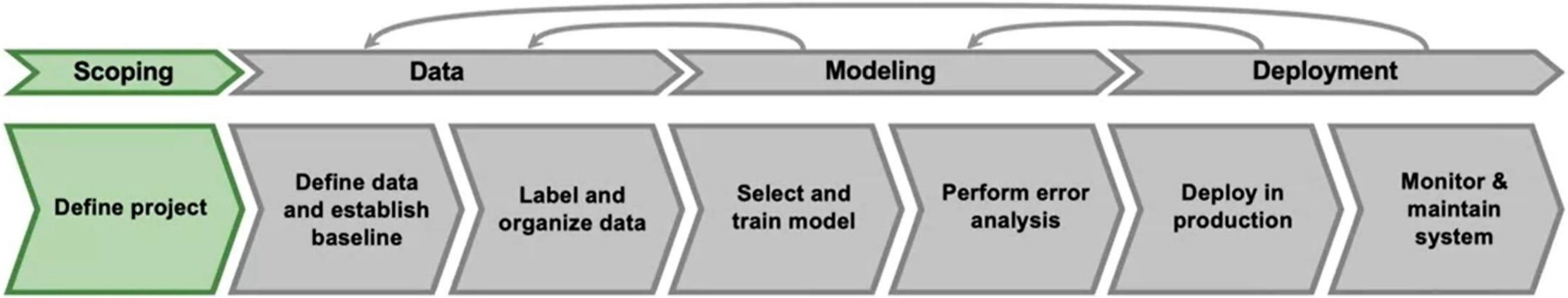
Steps of an ML project

The ML project lifecycle

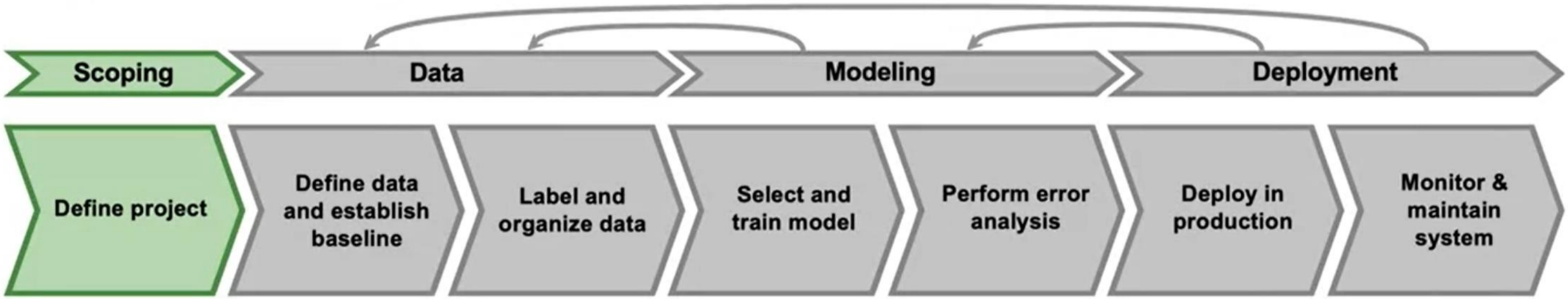


$X \rightarrow Y$

Speech recognition: Scoping stage

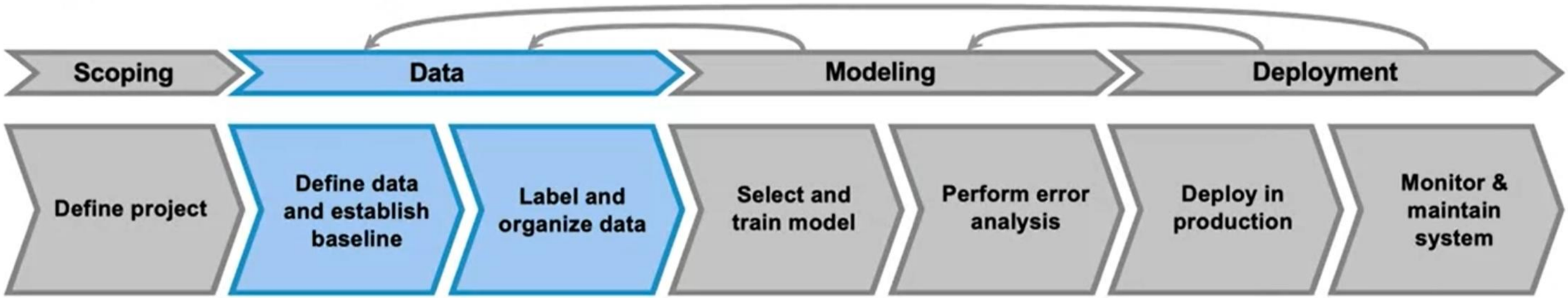


Speech recognition: Scoping stage



- Decide to work on speech recognition for voice search.
- Decide on key metrics:
  - Accuracy, latency, throughput
- Estimate resources and timeline

Speech recognition: Data stage



Define data

- Is the data labeled consistently?
- How much silence before/after each clip?
- How to perform volume normalization?

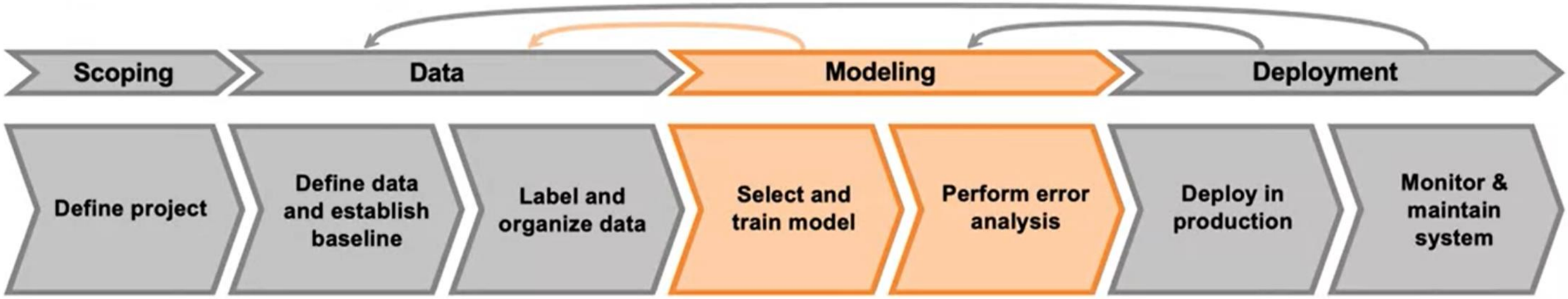
"Um, today's weather"

"Um... today's weather"

"Today's weather"

100ms 300ms 500ms

Speech recognition: Modeling stage



Research/Academia

Product Team

Code (algorithm/model)

Hyperparameters

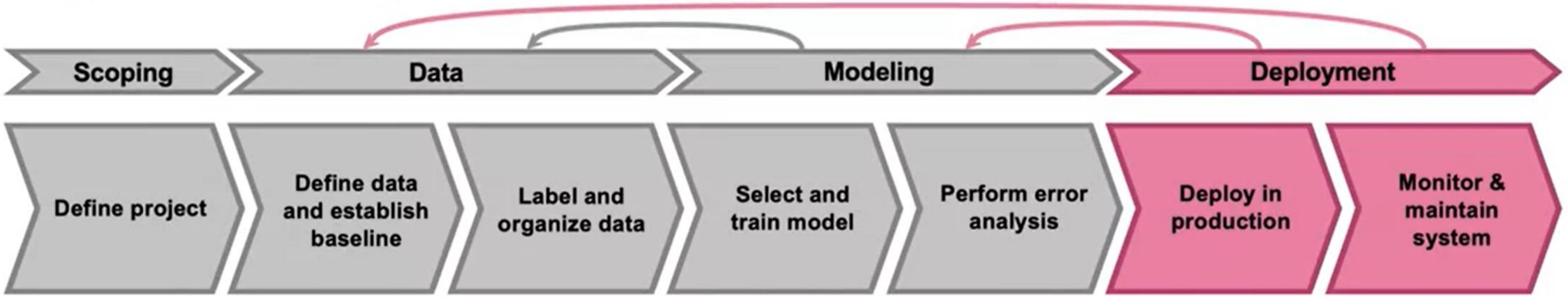
Data

ML Model

X → Y

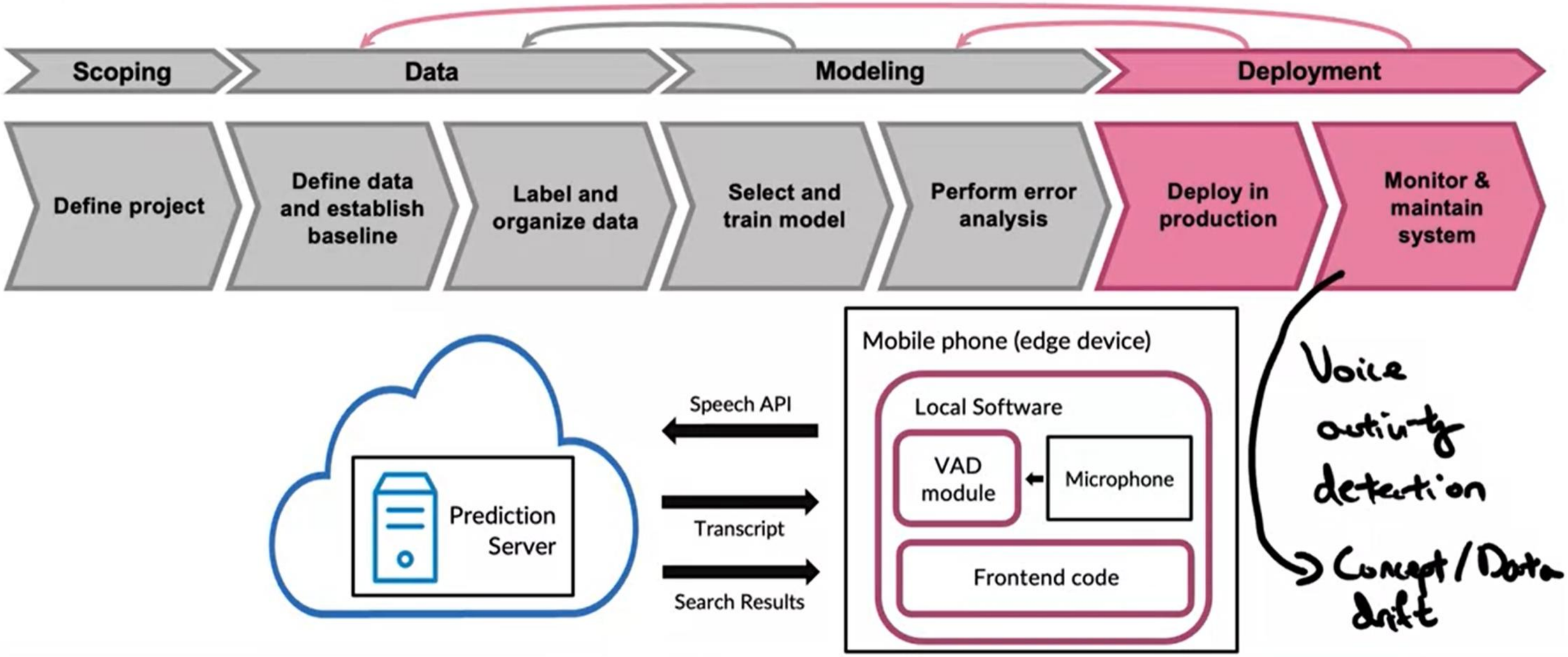
ML System = Code + Data

Speech recognition: Deployment stage

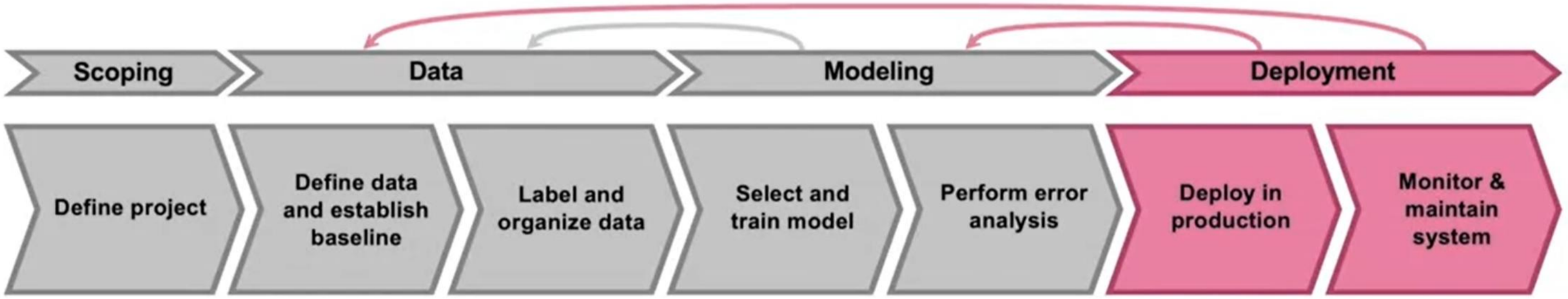


Mobile phone (edge device)

Speech recognition: Deployment stage

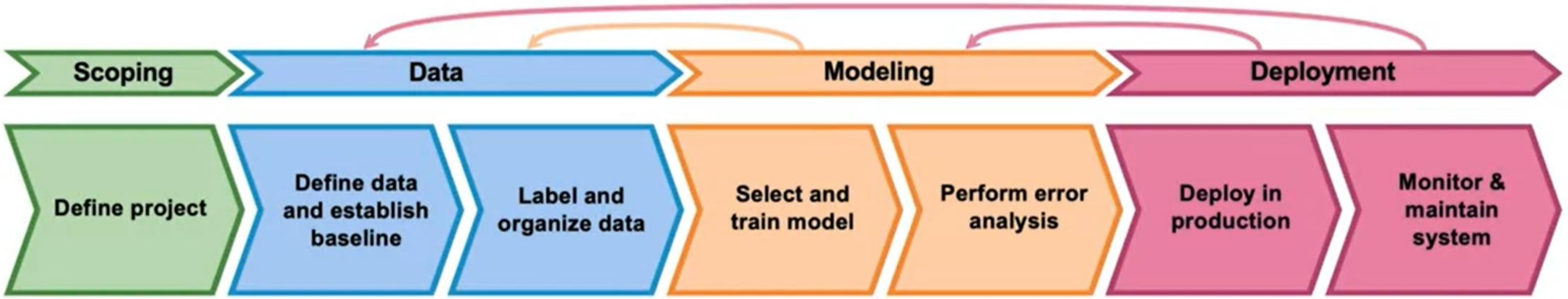


Course outline



1. Deployment

Course outline



1. Deployment
2. Modeling
3. Data

Optional: Scoping

MLOps (Machine Learning Operations) is an emerging discipline, and comprises a set of tools and principles to support progress through the ML project lifecycle.

Concept drift and Data drift

Speech recognition example

Training set: $x \rightarrow y$

- Purchased data, historical user data with transcripts

Test set:

- Data from a few months ago

Gradual change

How has the data changed?

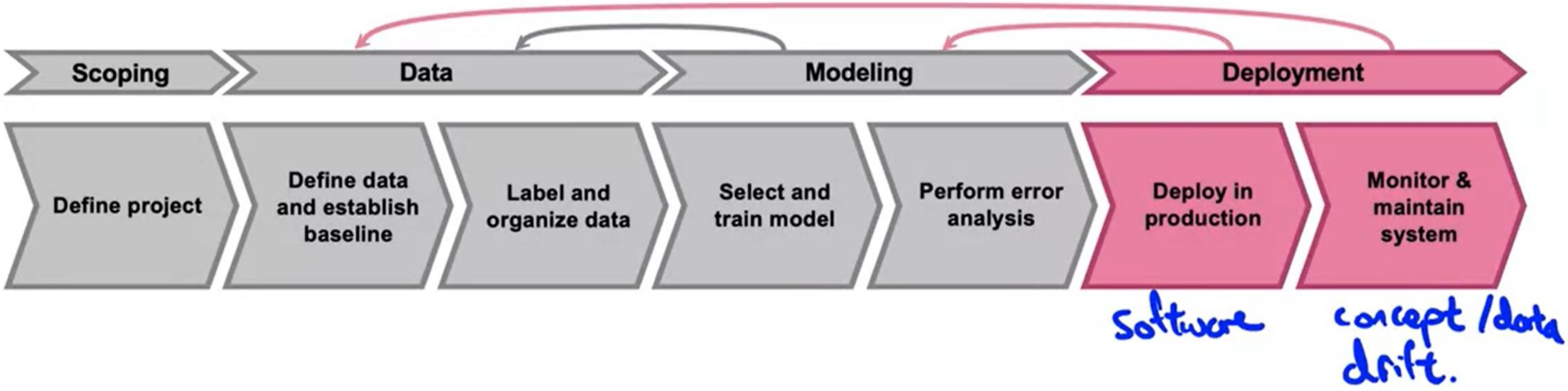
Software engineering issues

Checklist of questions

- Realtime or Batch
- Cloud vs. Edge/Browser
- Compute resources (CPU/GPU/memory)
- Latency, throughput (QPS)
- Logging

500ms, 1000 QPS

First deployment vs. maintenance



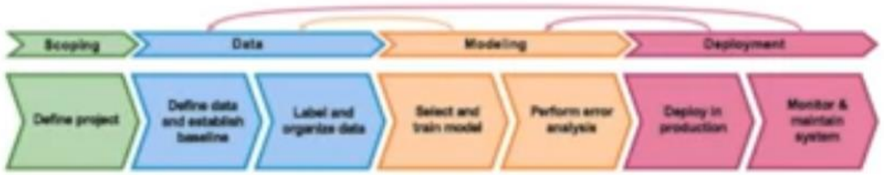
Common deployment cases

1. New product/capability
2. Automate/assist with manual task
3. Replace previous ML system

Key ideas:

- Gradual ramp up with monitoring
- Rollback

Visual inspection example



Visual inspection example

shadow mode

ML system shadows the human and runs in parallel.

Visual inspection example

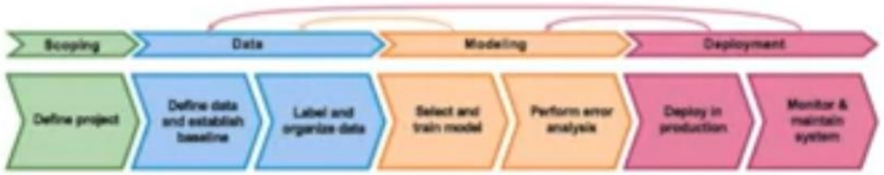
shadow mode

ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

Visual inspection example

shadow mode



Human

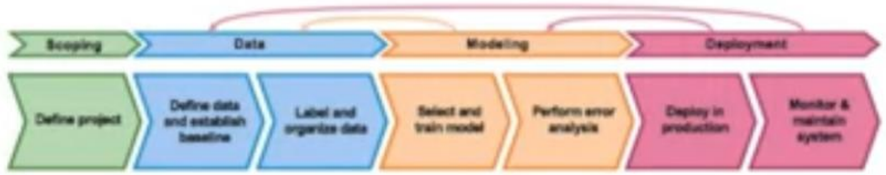
✓

ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

Visual inspection example

shadow mode



Human

✓

ML

✓

ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

Visual inspection example

shadow mode

Human ✓

ML ✓

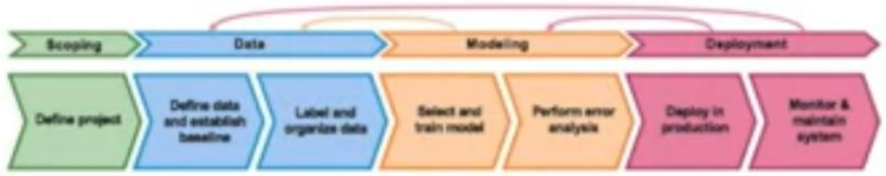
Human ✗

ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

Visual inspection example

shadow mode



Human ✓

ML ✓

Human ✗

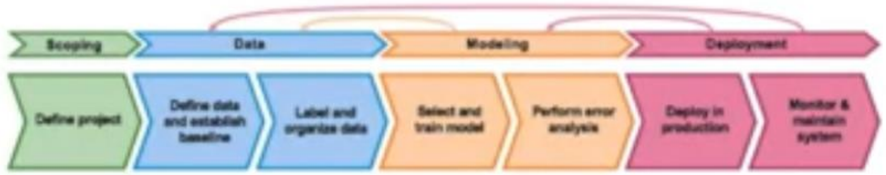
ML ✗

ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

Visual inspection example

shadow mode



Human ✓

ML ✓

Human ✗

ML ✗

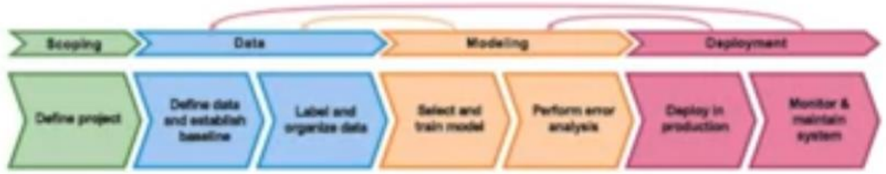
Human ✗

ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

Visual inspection example

shadow mode



Human ✓

ML ✓

Human ✗

ML ✗

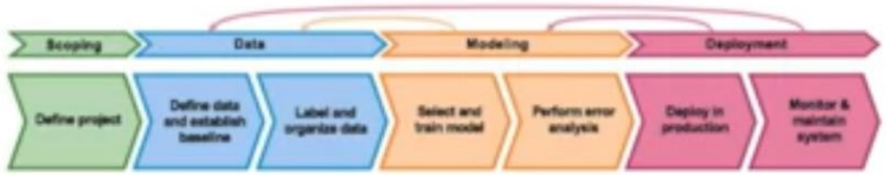
Human ✗

ML ✓

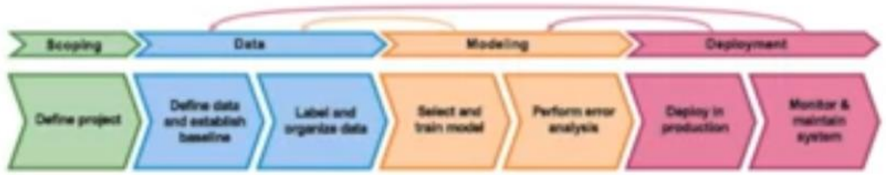
ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

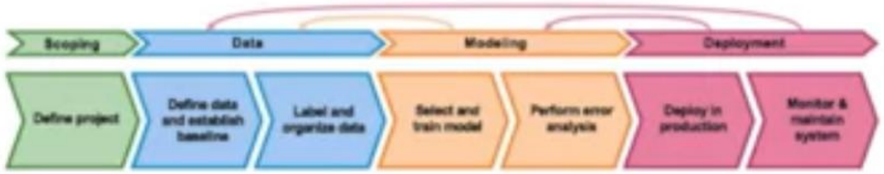
Canary deployment



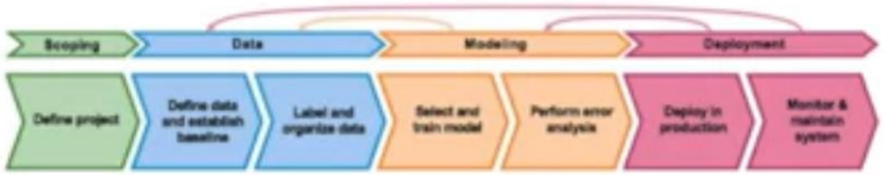
Canary deployment



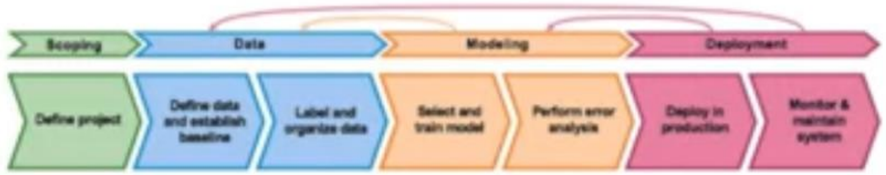
Canary deployment



Canary deployment

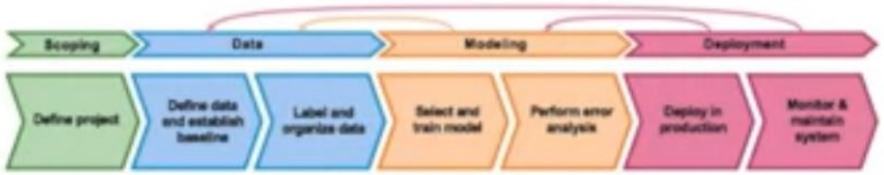


Canary deployment



- Roll out to small fraction (say 5%) of traffic initially.
- Monitor system and ramp up traffic gradually.

Canary deployment



- Roll out to small fraction (say 5%) of traffic initially.
- Monitor system and ramp up traffic gradually.

Blue green deployment

Phone images

Blue green deployment

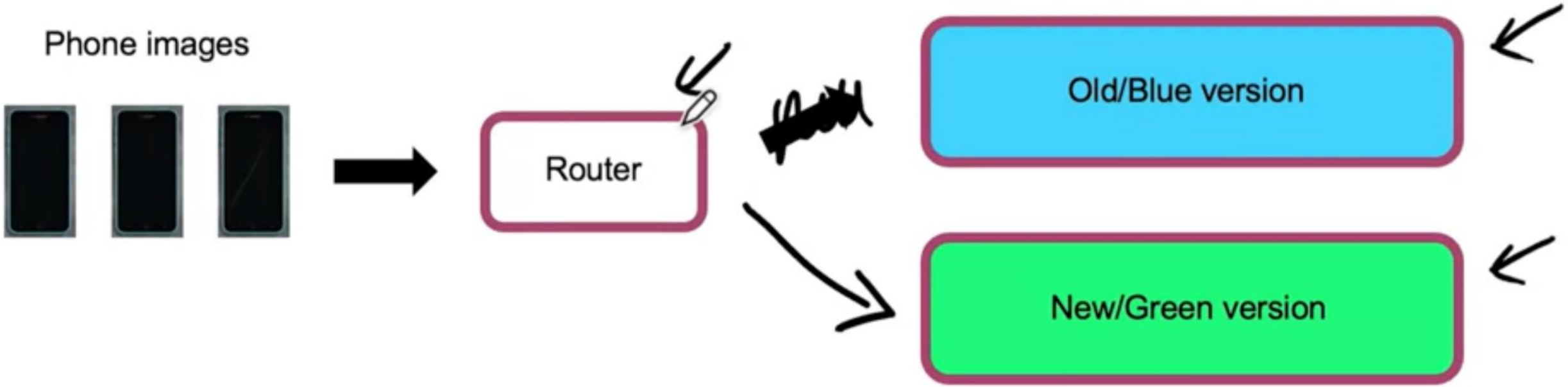
Phone images

Router

Old/Blue version

New/Green version

Blue green deployment



Easy way to enable rollback

Degrees of automation

Human

human

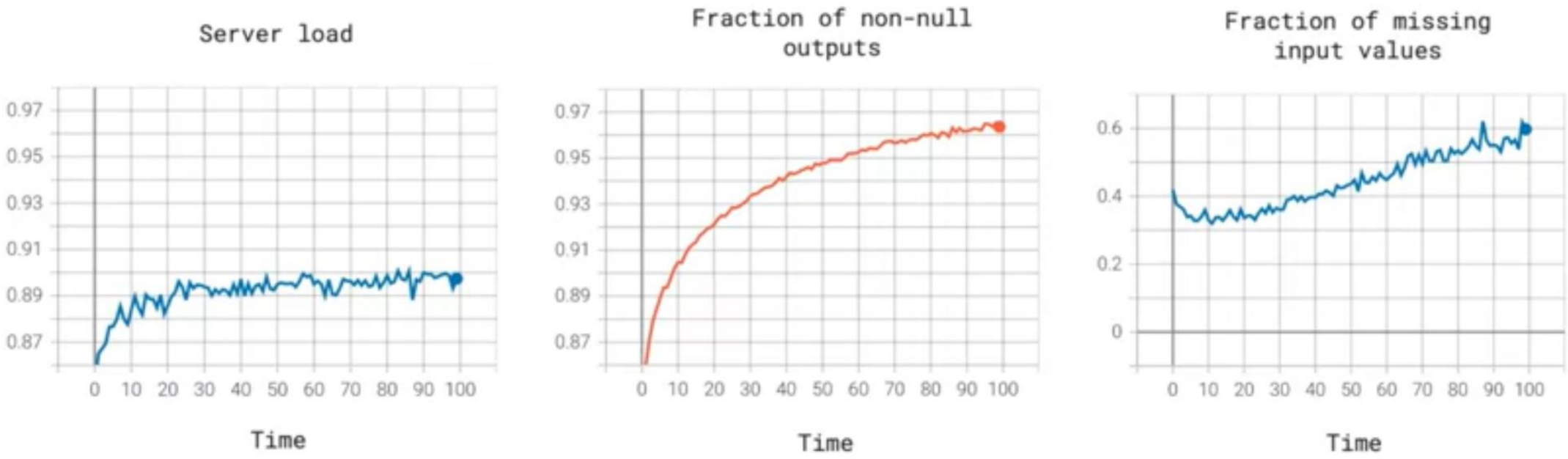
AI

Human only → Shadow mode → AI assistance → Partial automation → Full automation

human in the loop

You can choose to stop before getting to full automation.

Monitoring dashboard

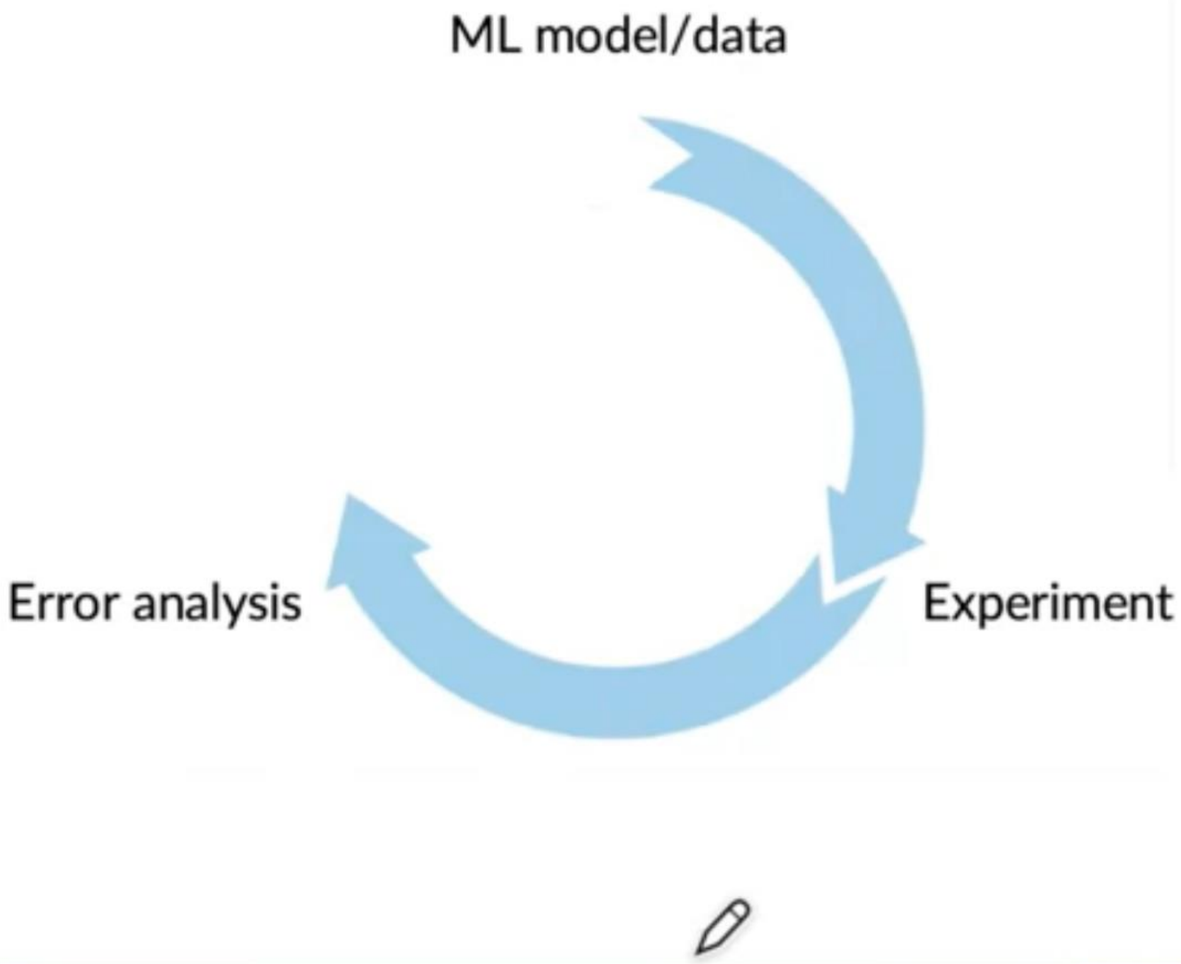


- Brainstorm the things that could go wrong.
- Brainstorm a few statistics/metrics that will detect the problem.

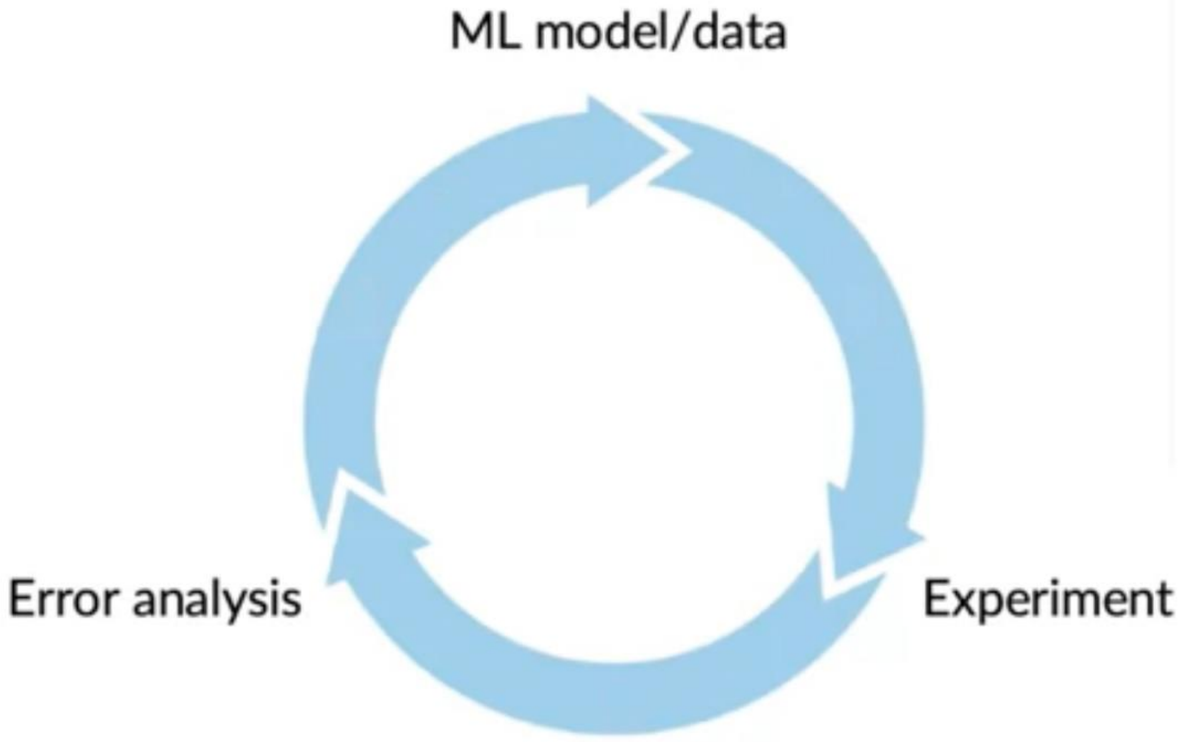
Examples of metrics to track

Software metrics:	Memory, compute, latency, throughput, server load
Input metrics: 	Avg input length Avg input volume Num missing values Avg image brightness
Output metrics: 	# times return " " (null) # times user redoes search # times user switches to typing CTR

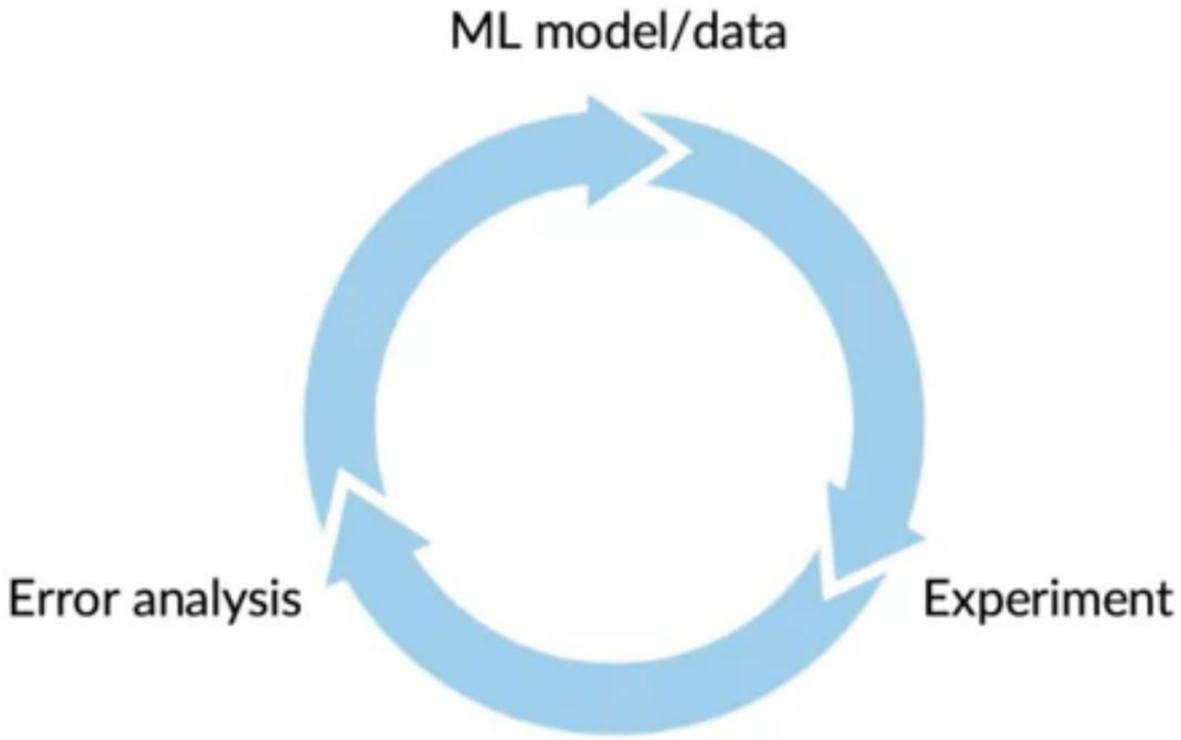
Just as ML modeling is iterative, so is deployment

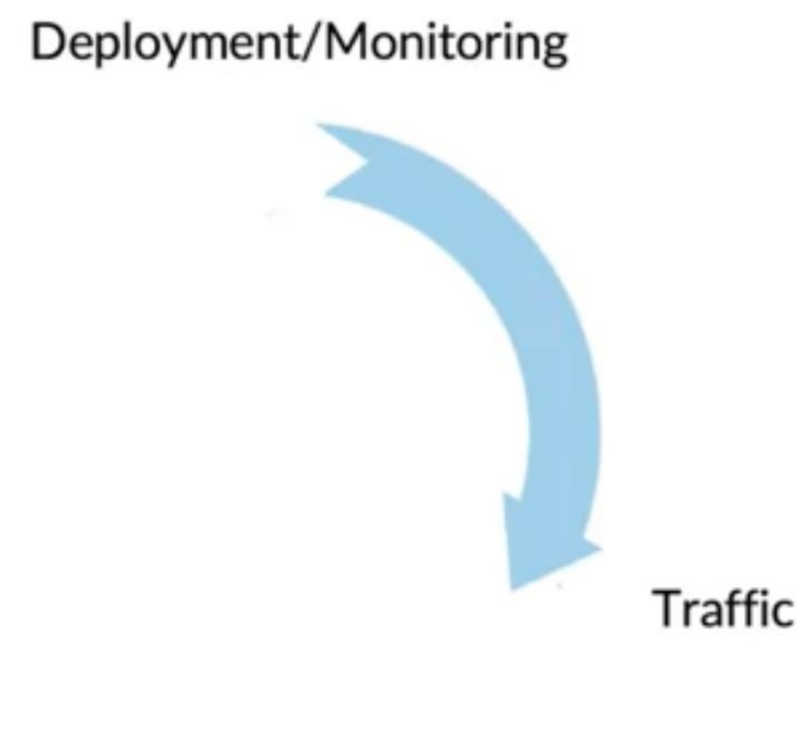


Just as ML modeling is iterative, so is deployment

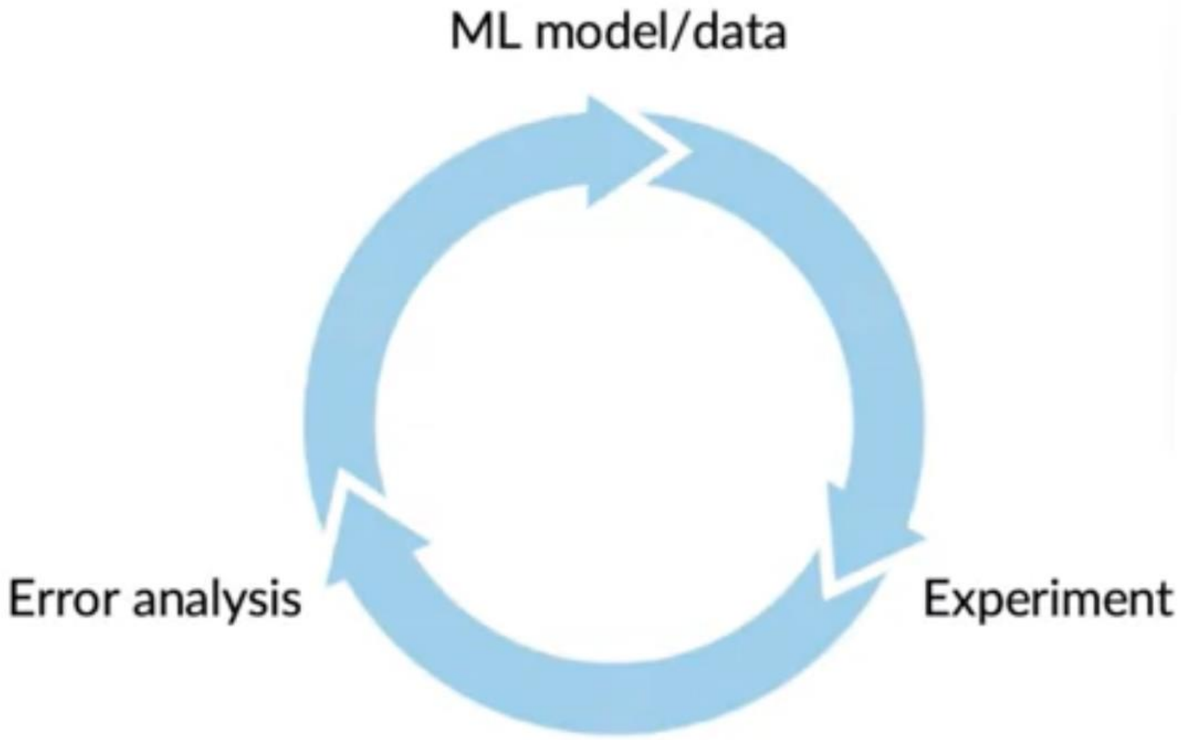


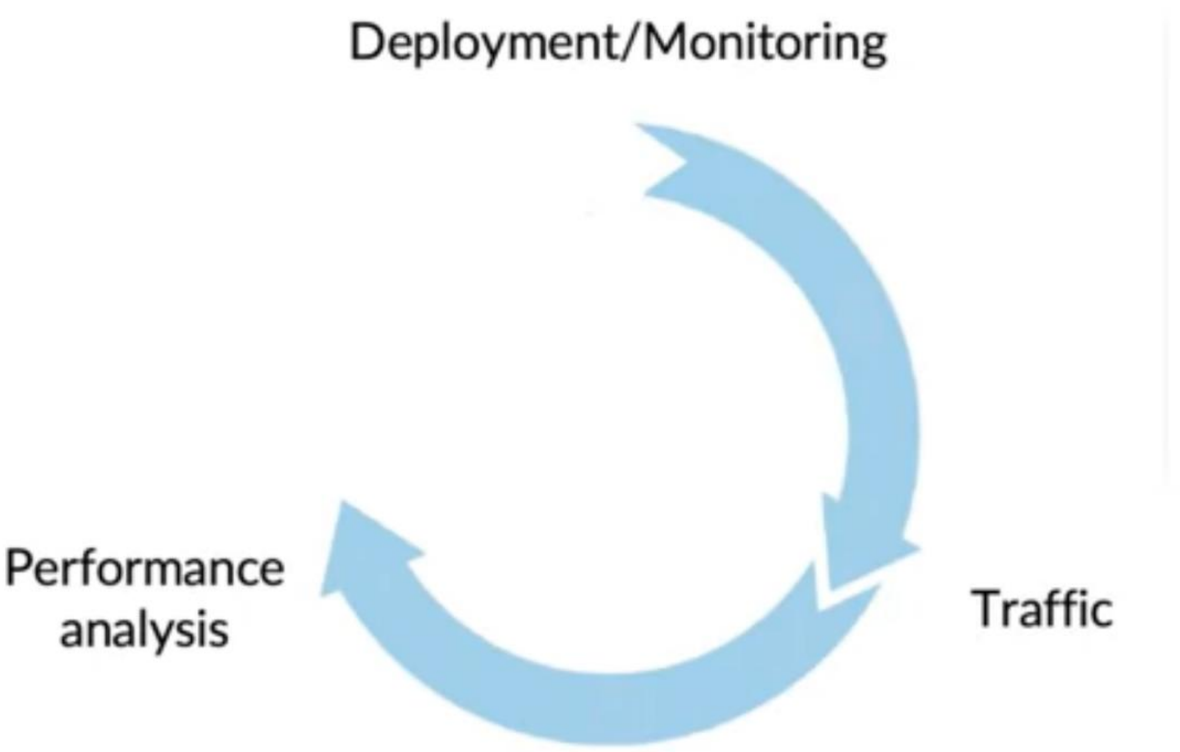
Just as ML modeling is iterative, so is deployment



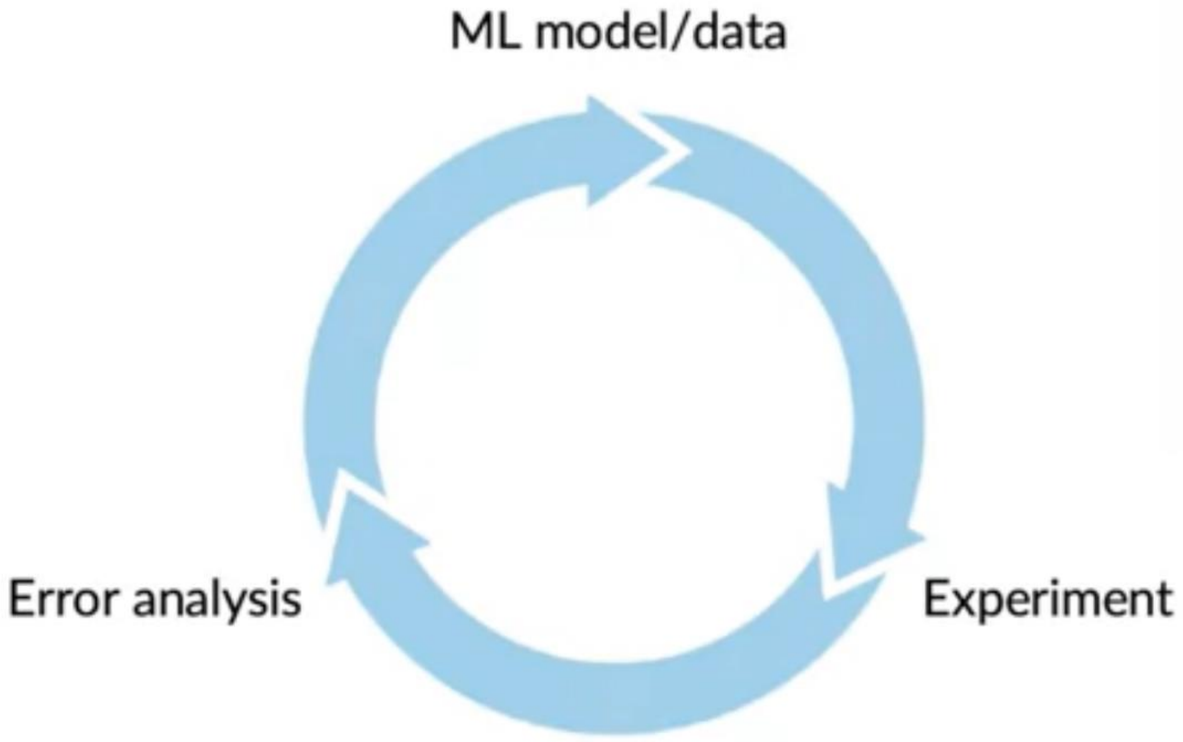


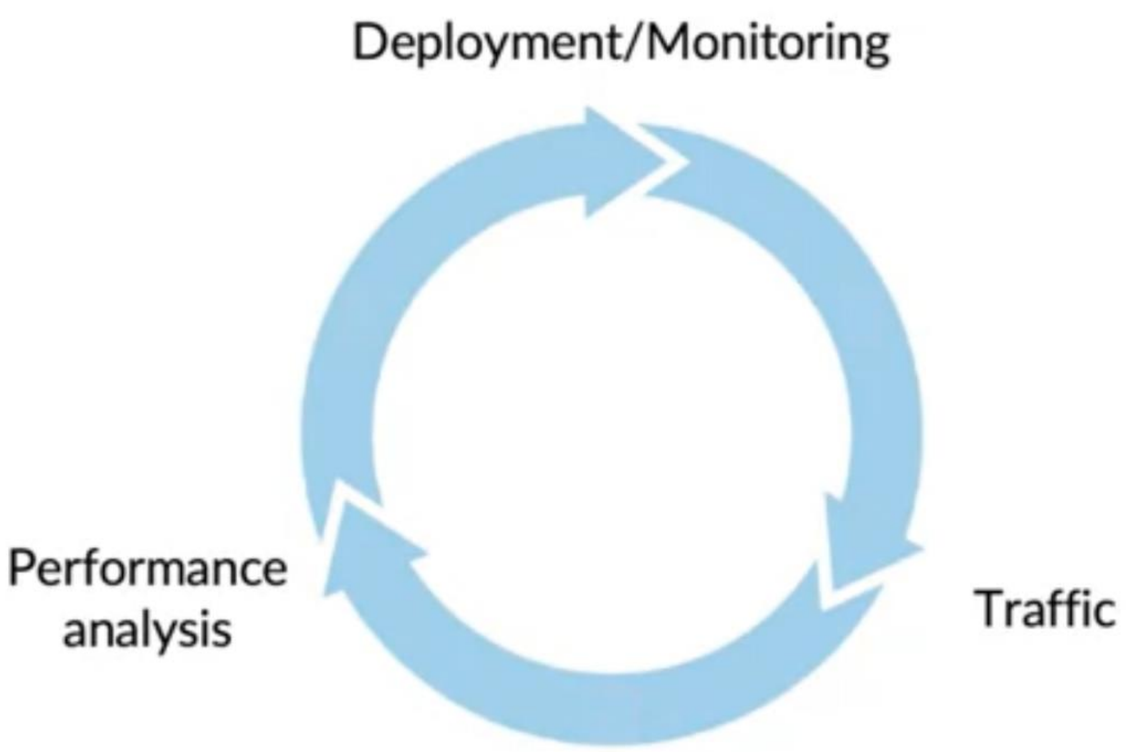
Just as ML modeling is iterative, so is deployment



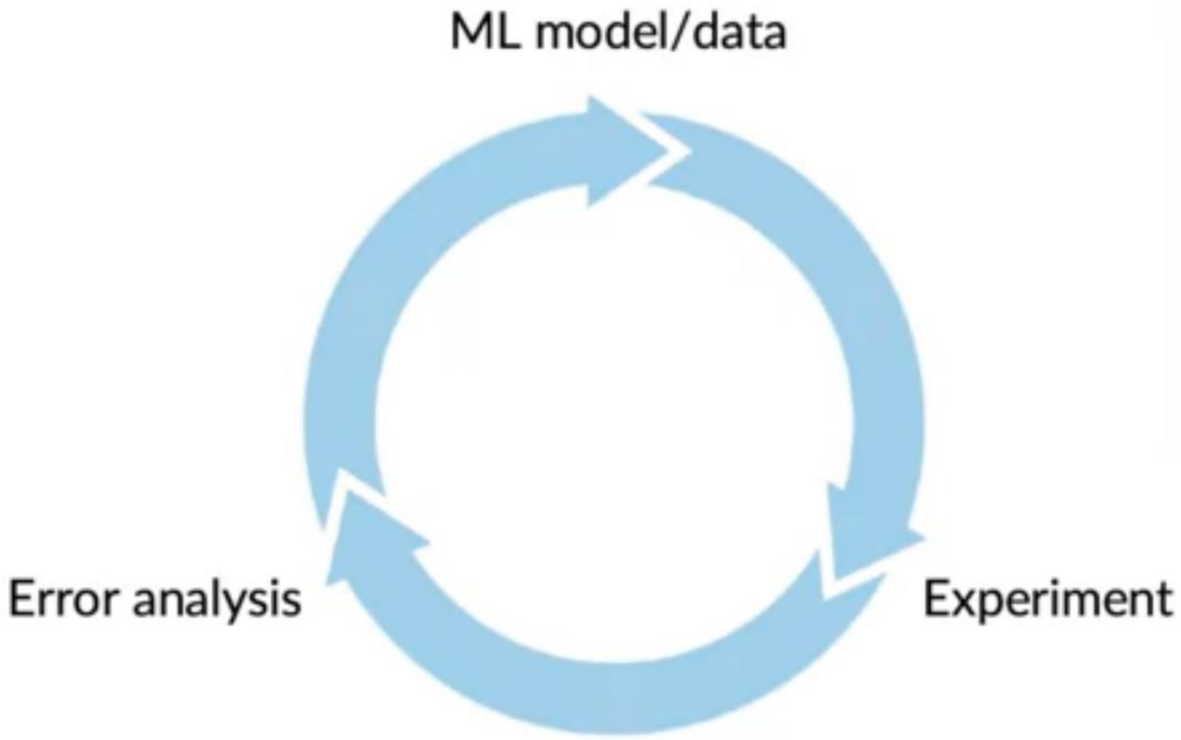


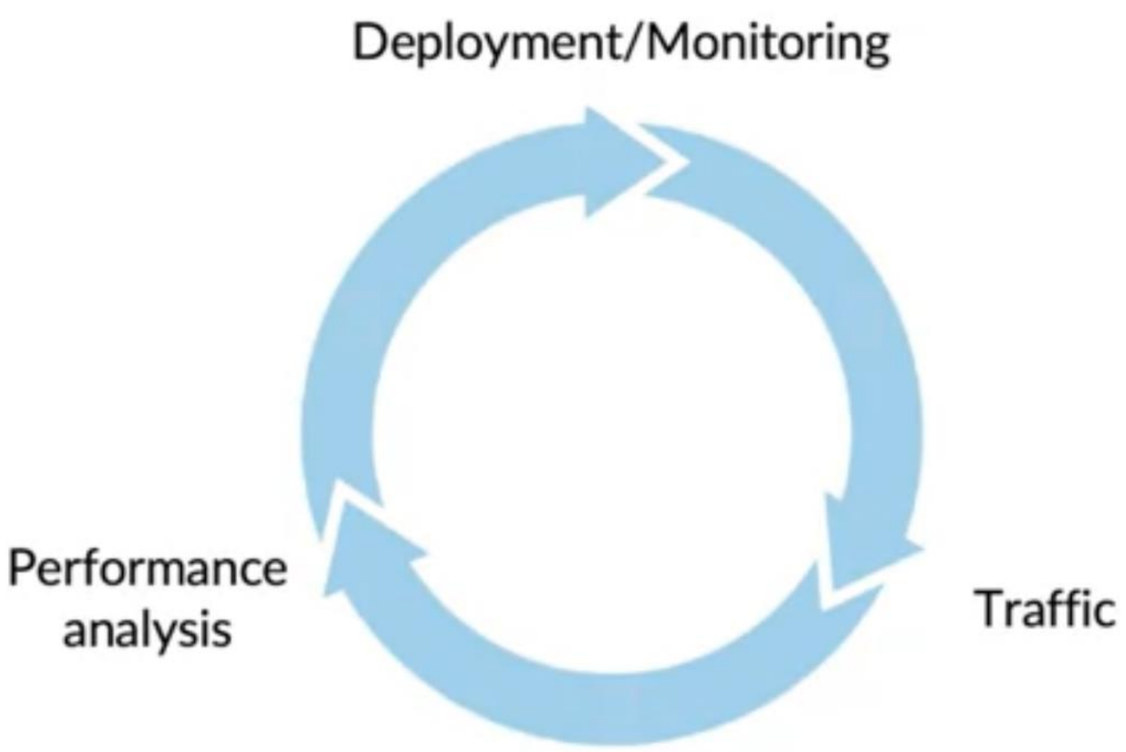
Just as ML modeling is iterative, so is deployment





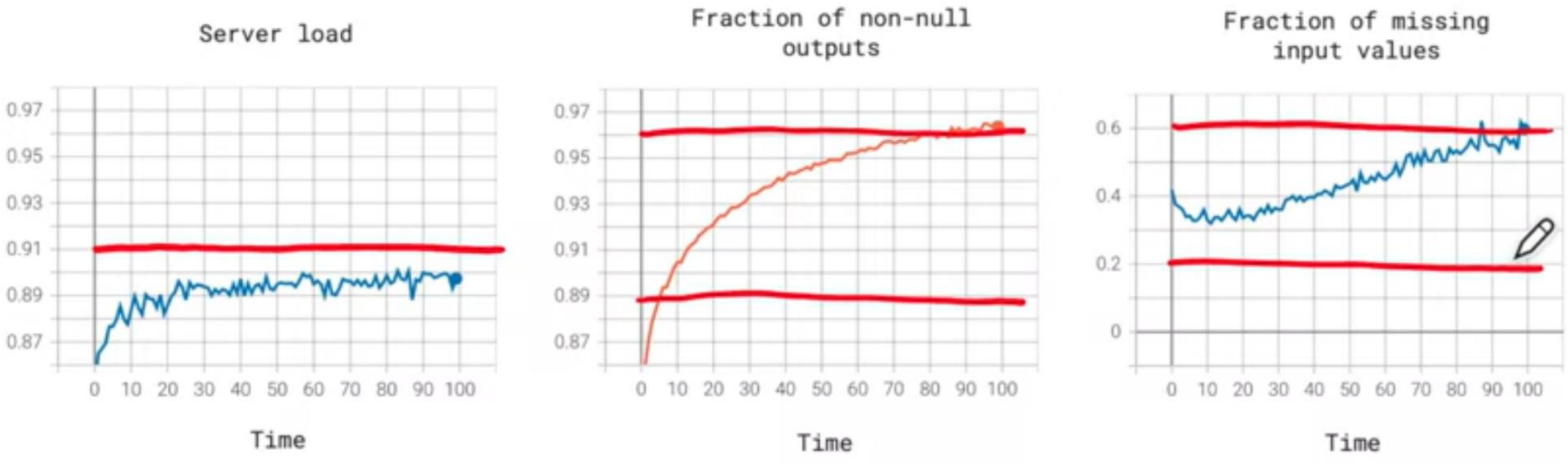
Just as ML modeling is iterative, so is deployment





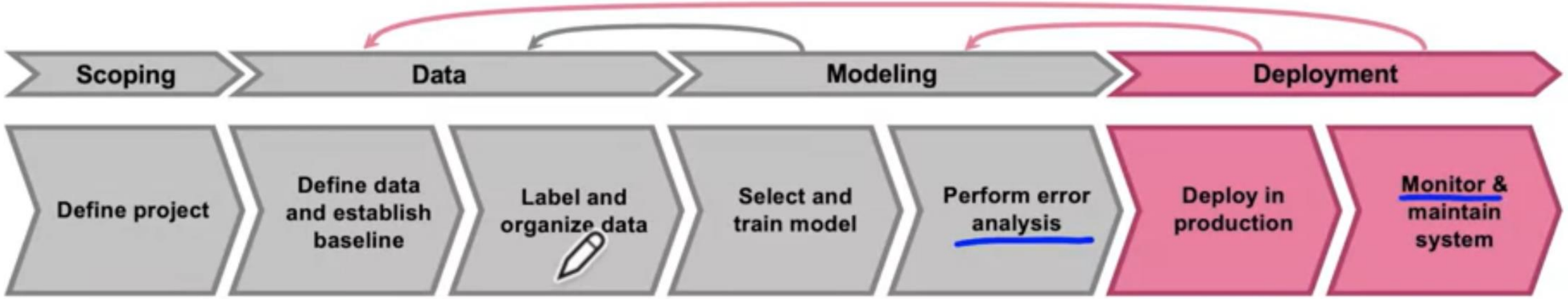
Iterative process to choose the right set of metrics to monitor.

Monitoring dashboard



- Set thresholds for alarms
- Adapt metrics and thresholds over time

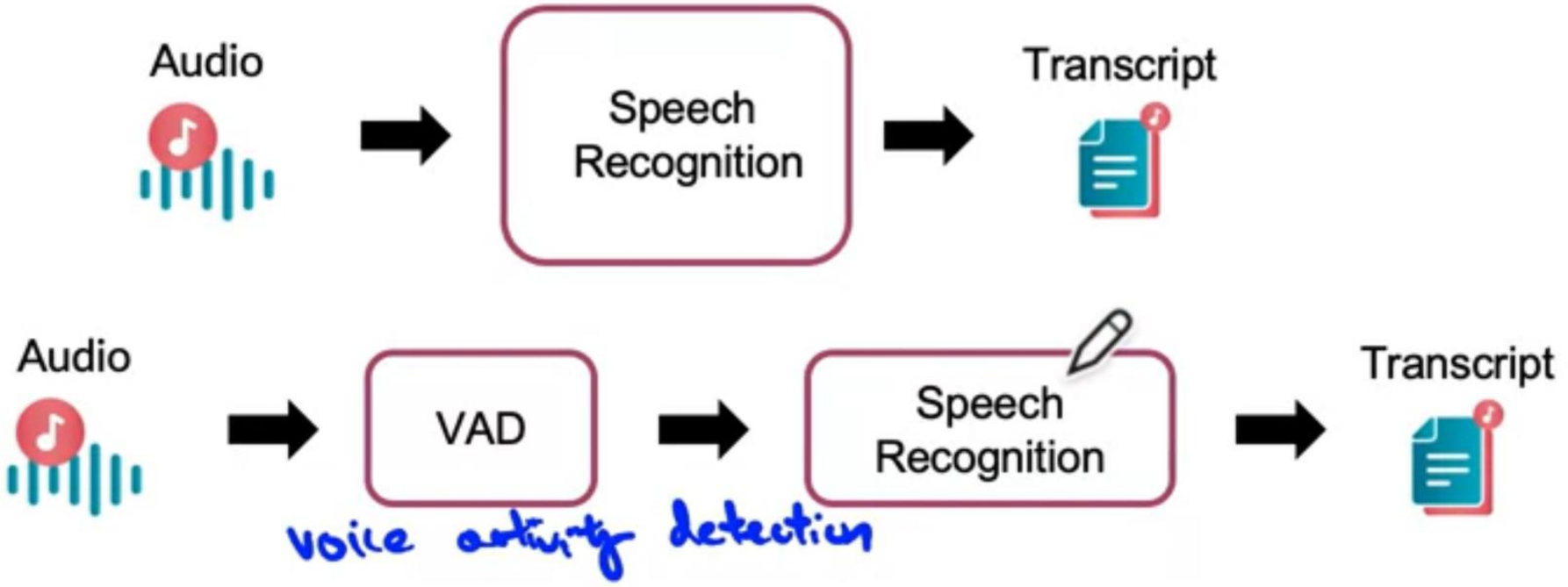
Model maintenance



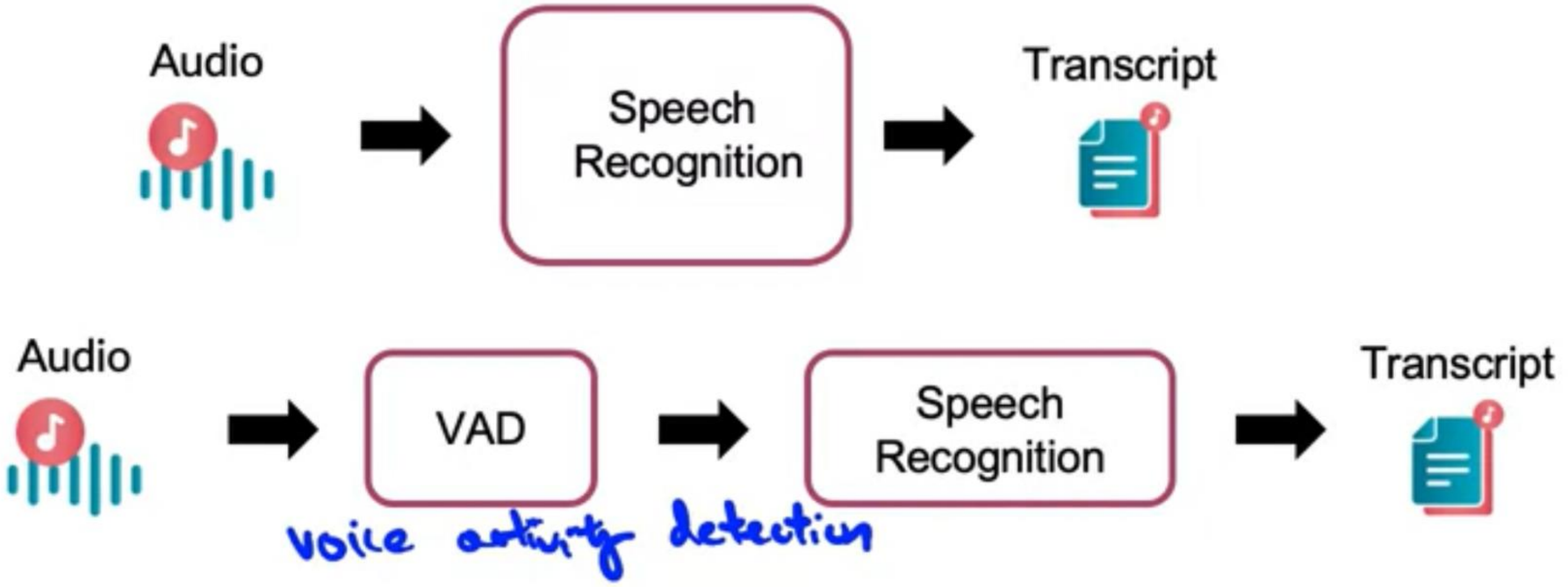
- Manual retraining
- Automatic retraining

Speech recognition example

Speech recognition example

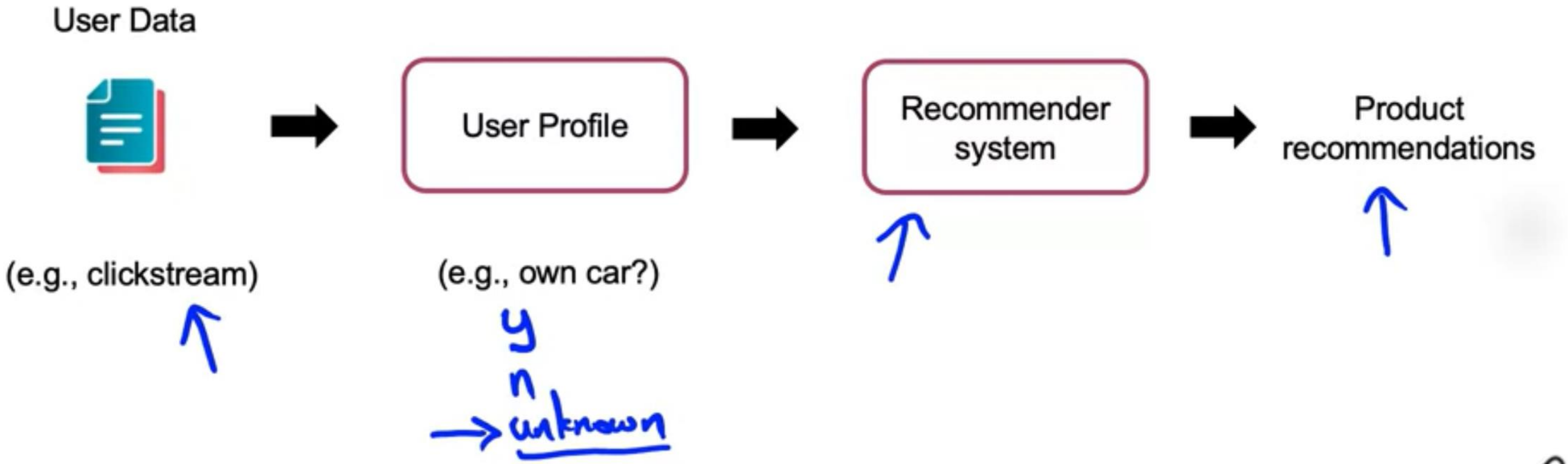


Speech recognition example



Some cellphones might have VAD clip audio differently, leading to degraded performance

User profile example



Metrics to monitor

Monitor

- Software metrics
- Input metrics
- Output metrics

How quickly do they change?

- User data generally has slower drift.
- Enterprise data (B2B applications) can shift fast.