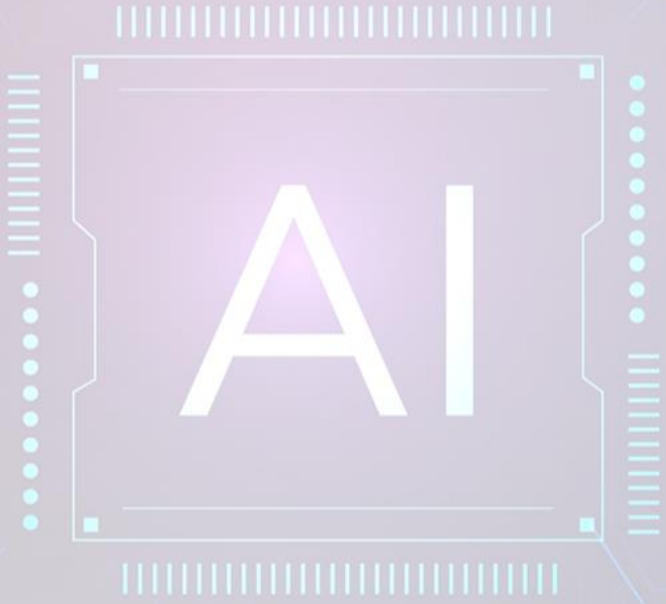
ML in Production

Part 1: ML Lifecycle

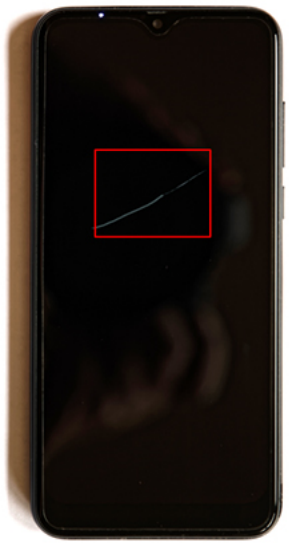
Ramin Toosi

ML Project Lifecycle

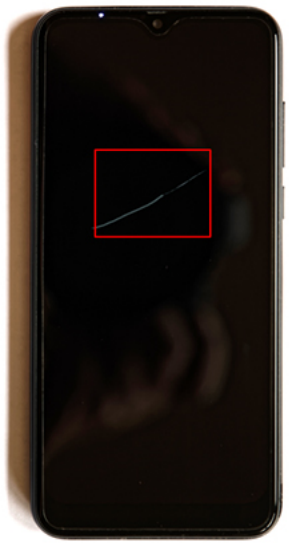


Deployment Example

Deployment Example

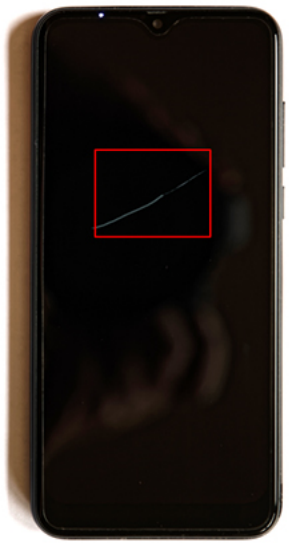


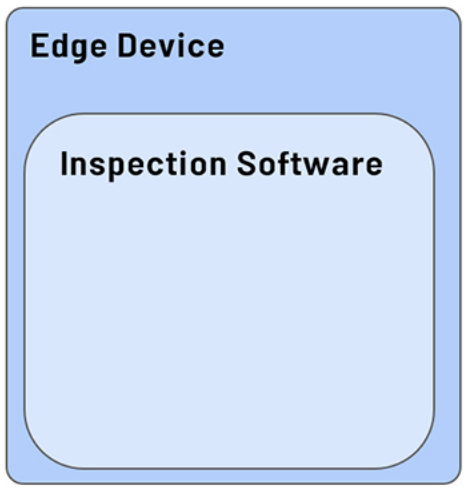
Deployment Example



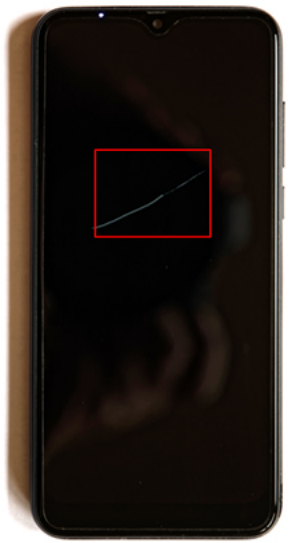
Edge Device

Deployment Example

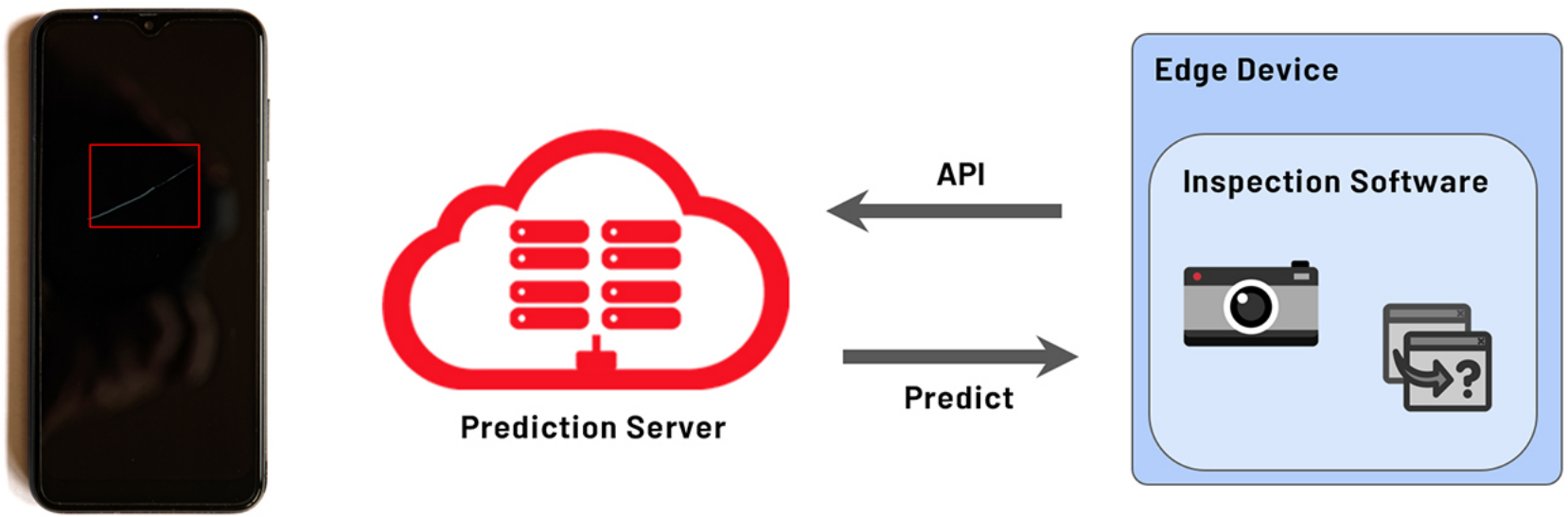




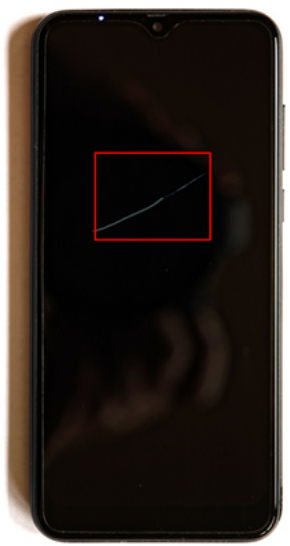
Deployment Example



Deployment Example



Deployment Example



Prediction Server

X → Y

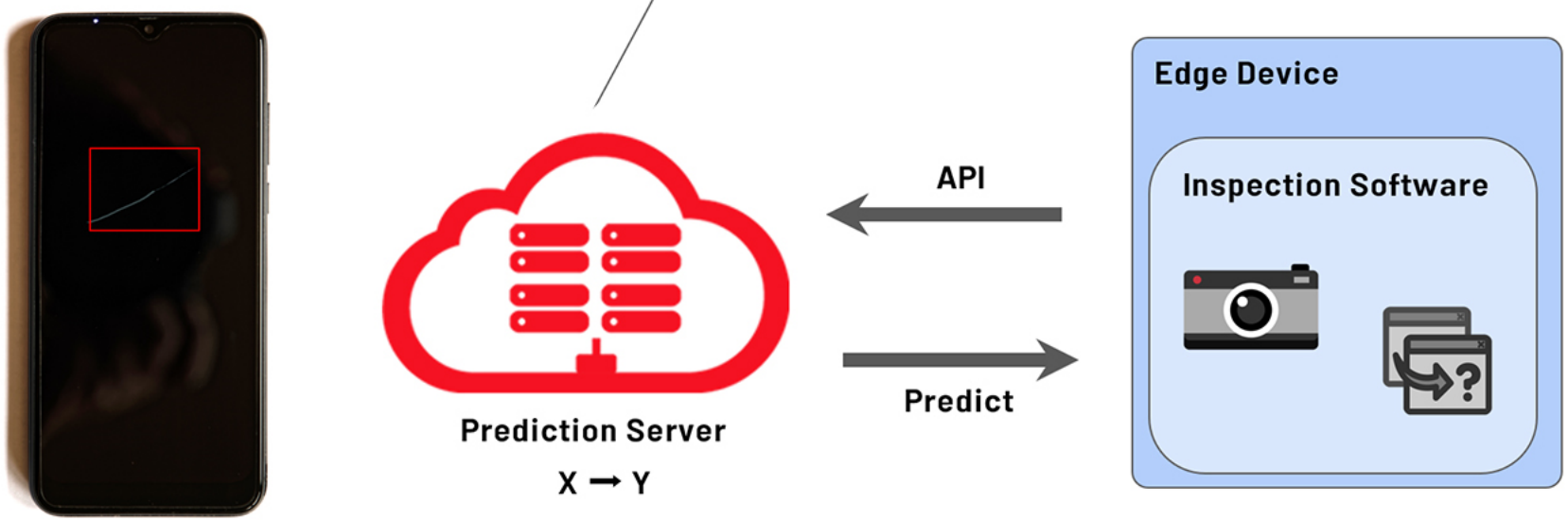
API

Predict

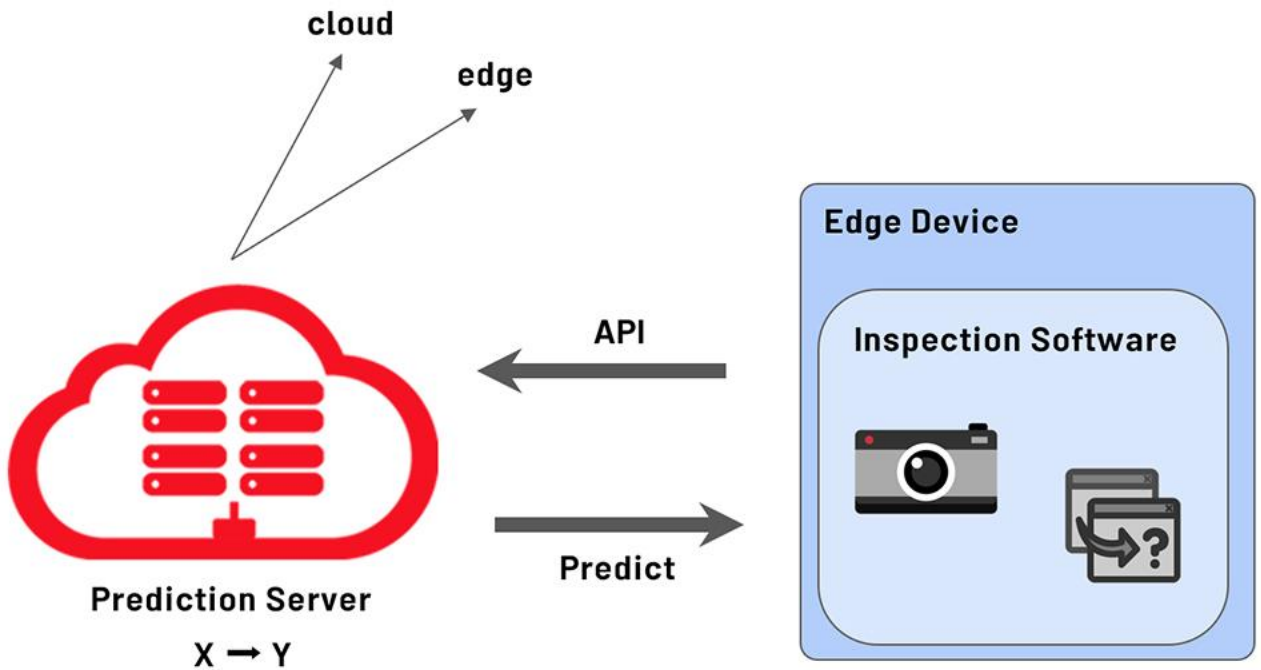
Edge Device

Inspection Software

Deployment Example



Deployment Example



Visual inspection example

Visual inspection example

Visual inspection example

ML in production

ML Model Code

ML in production

ML Project Code

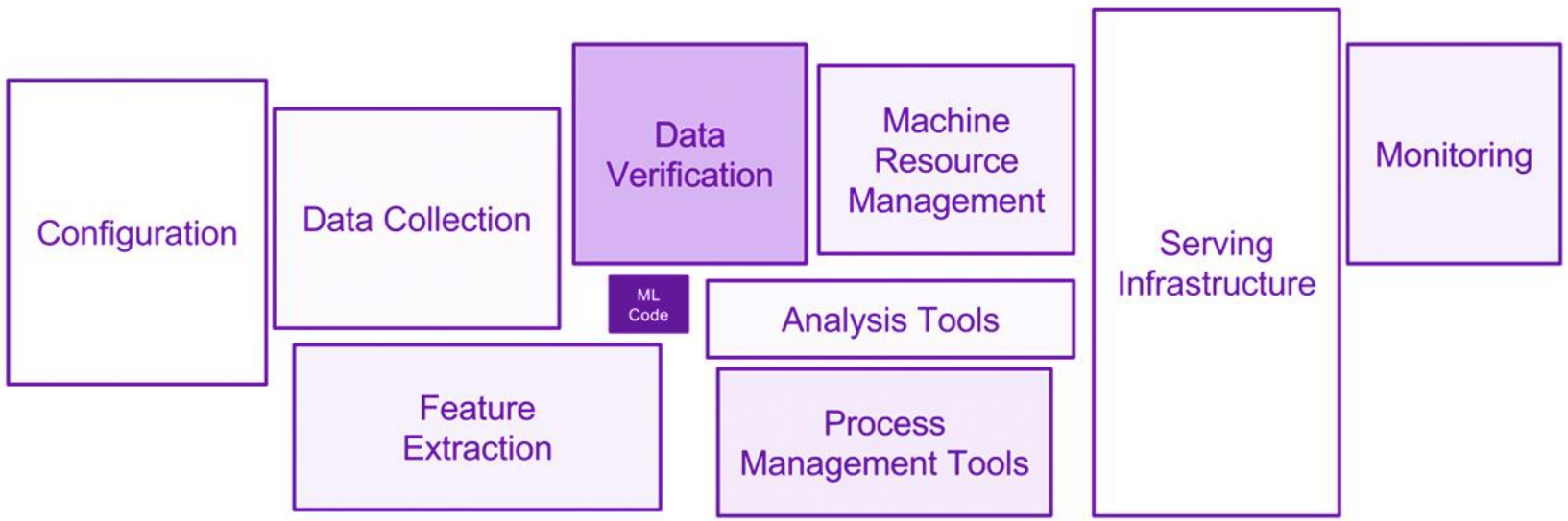
ML Model Code

ML in production

ML Project Code

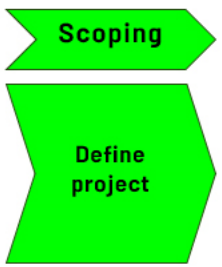
ML Model Code

Beyond the ML code

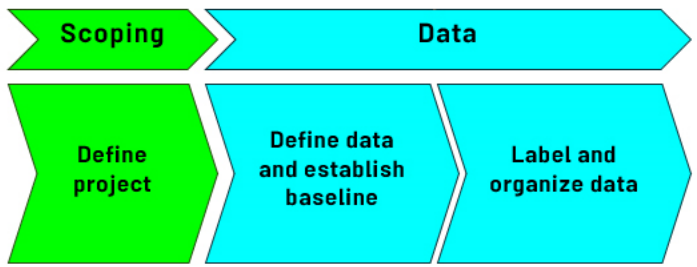


Ref: Hidden Technical Debt in Machine Learning Systems, NIPS 2015

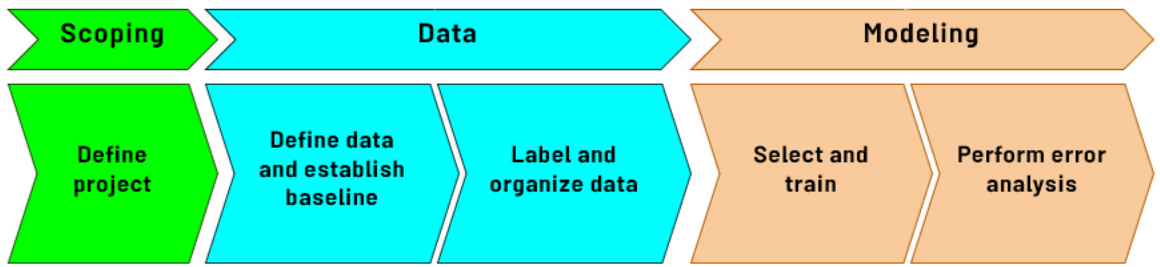
The ML project lifecycle



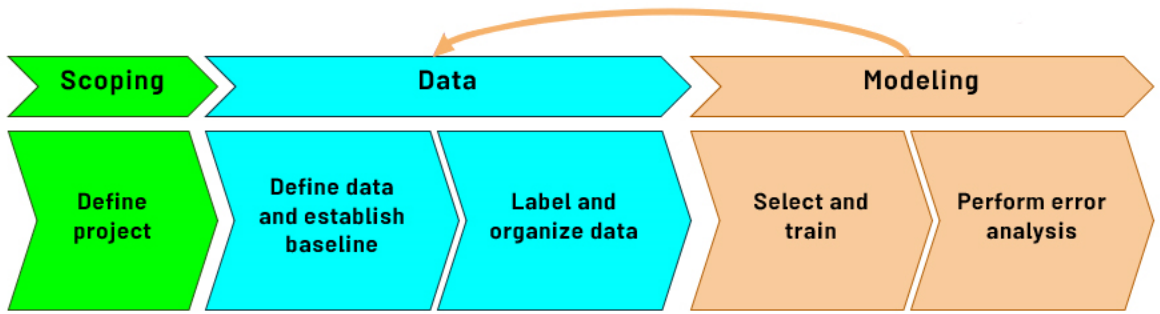
The ML project lifecycle



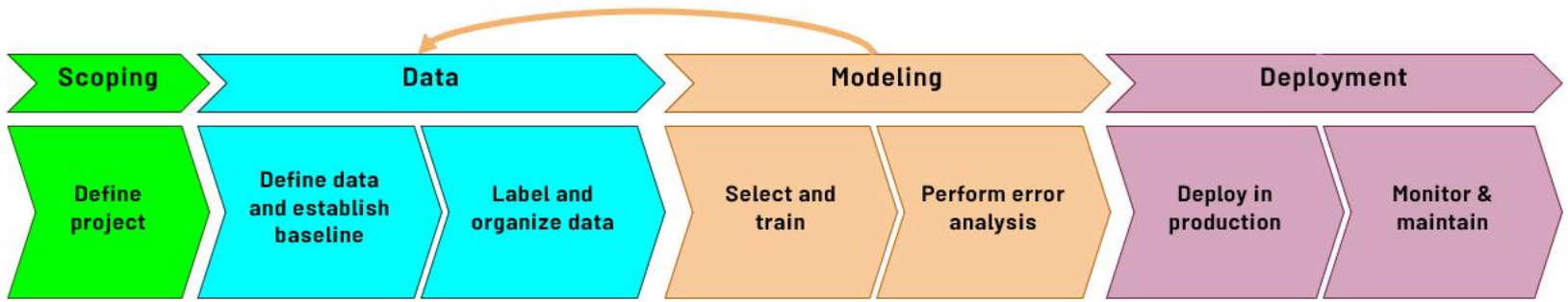
The ML project lifecycle



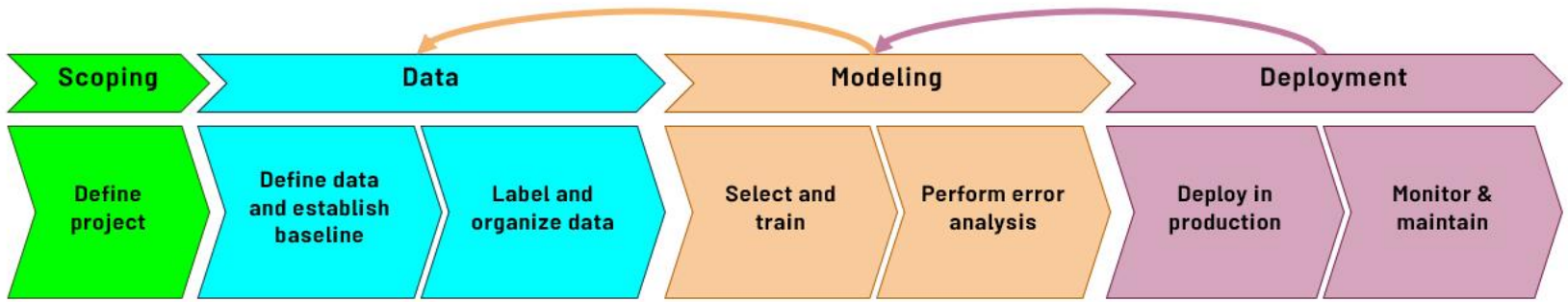
The ML project lifecycle



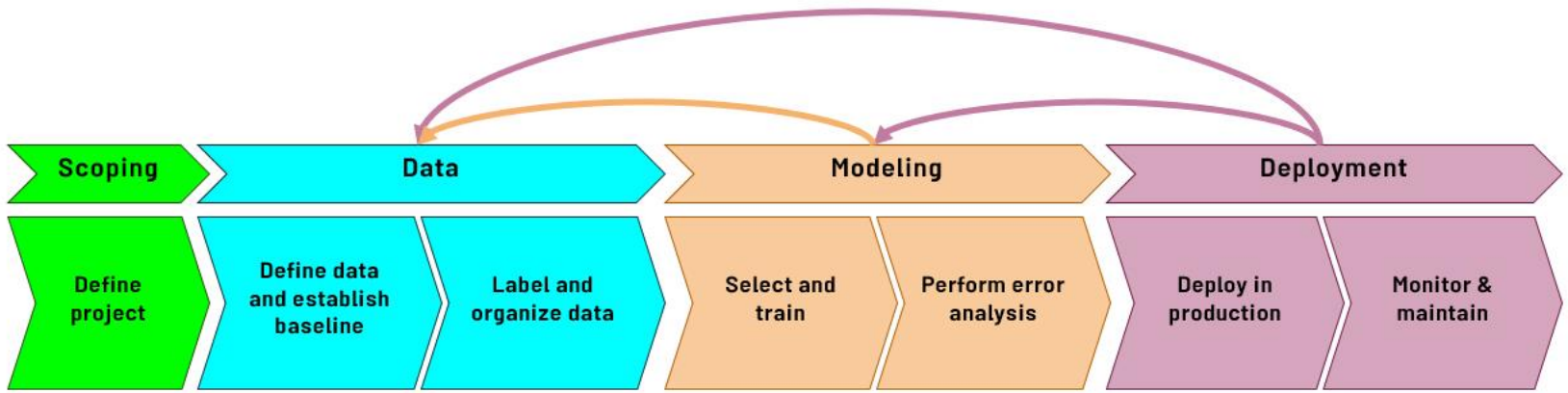
The ML project lifecycle



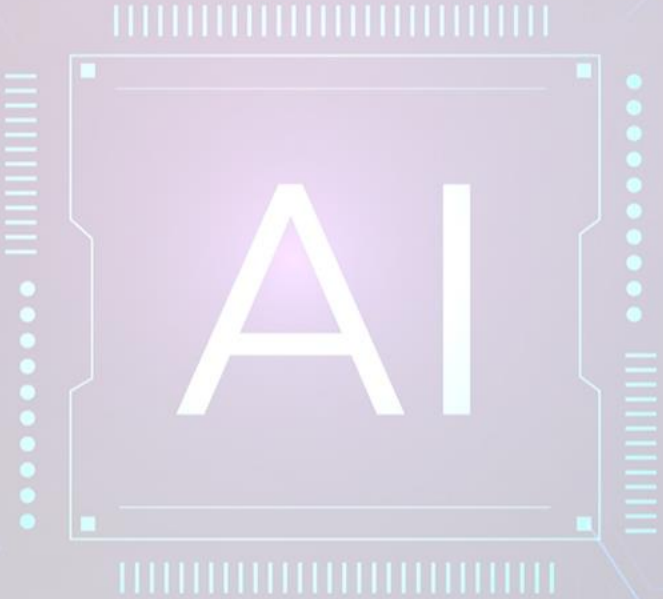
The ML project lifecycle



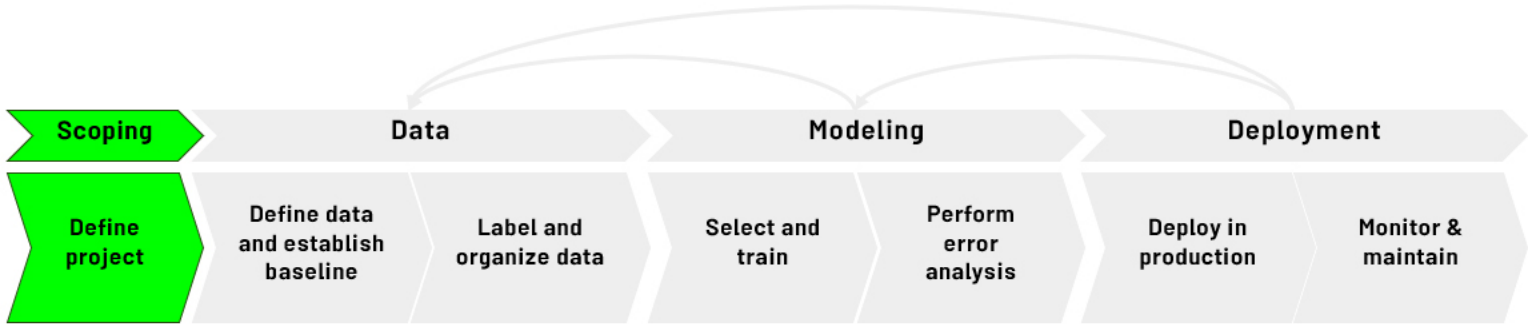
The ML project lifecycle



Case Study - Speech Recognition

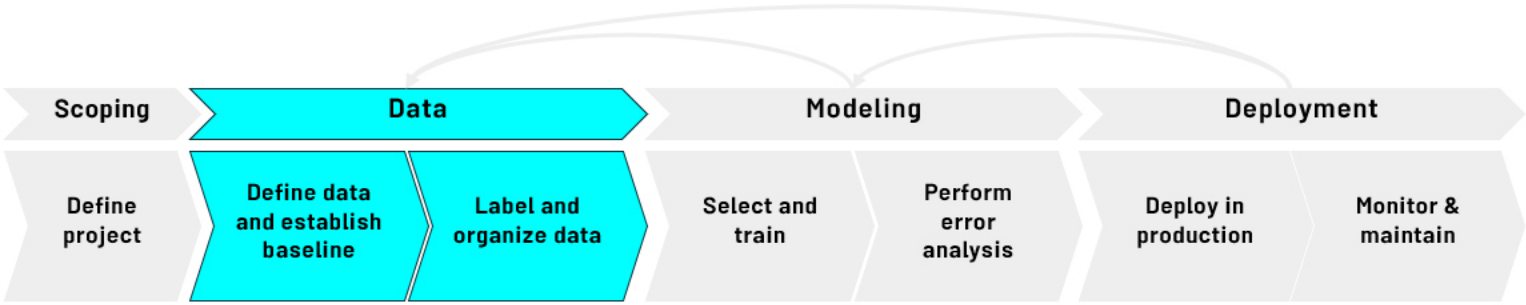


Speech recognition: Scoping stage



- 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

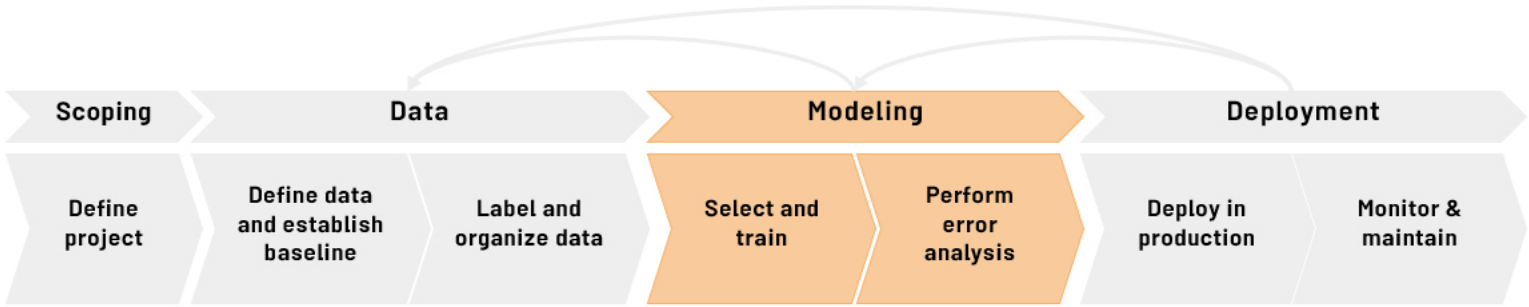
- Is the data labeled consistently?
- How much silence before/after each clip?
- How to perform volume normalization?

"Um, today's weather?"

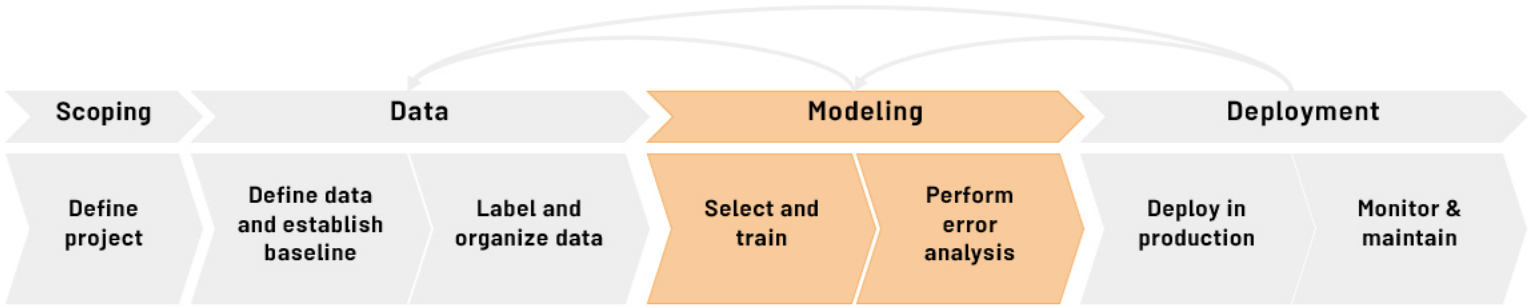
"Um… today's weather?"

"Today's weather?"

Speech recognition: modeling stage

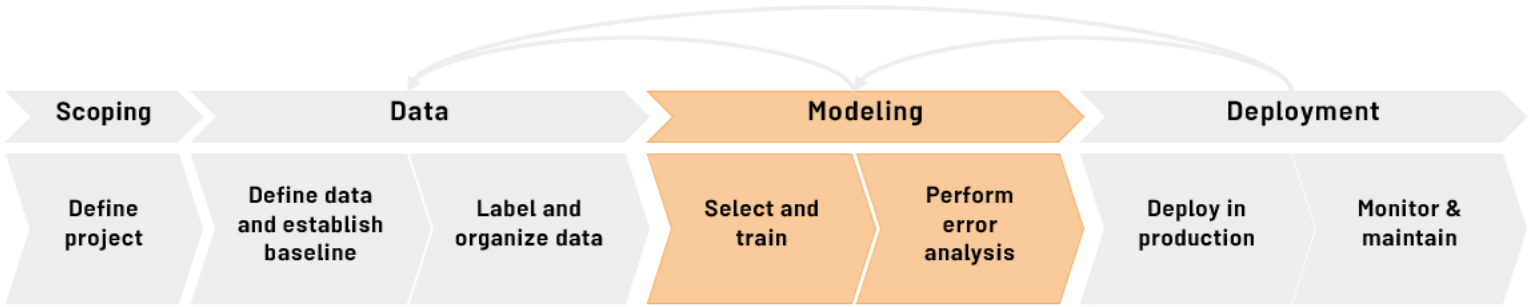


Speech recognition: modeling stage



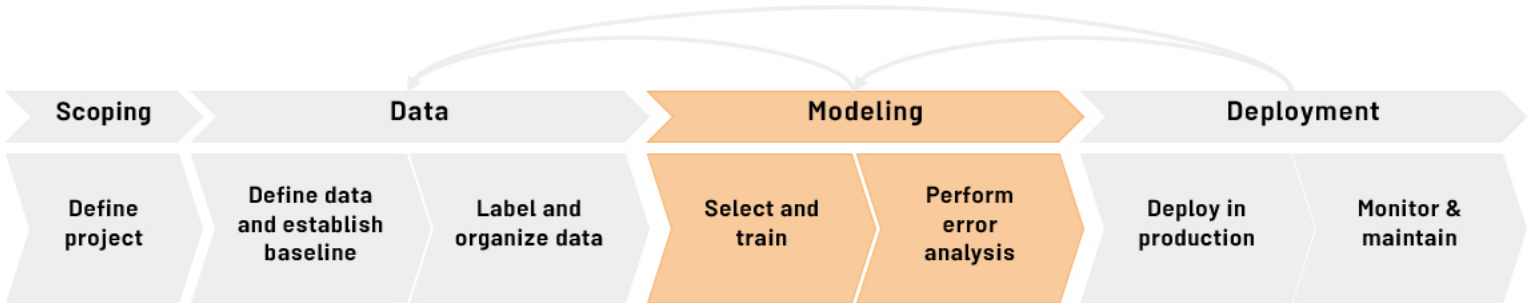
Code (algorithm/model)

Speech recognition: modeling stage



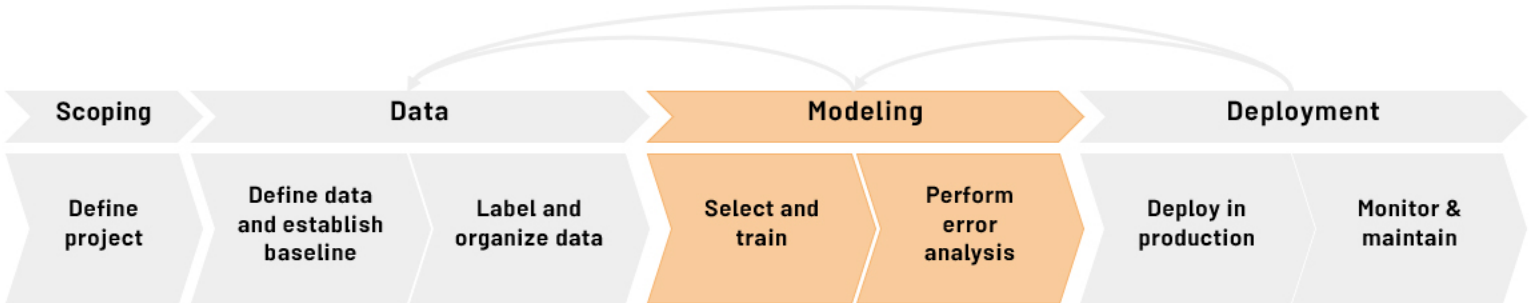
- Code (algorithm/model)
- Hyperparameters

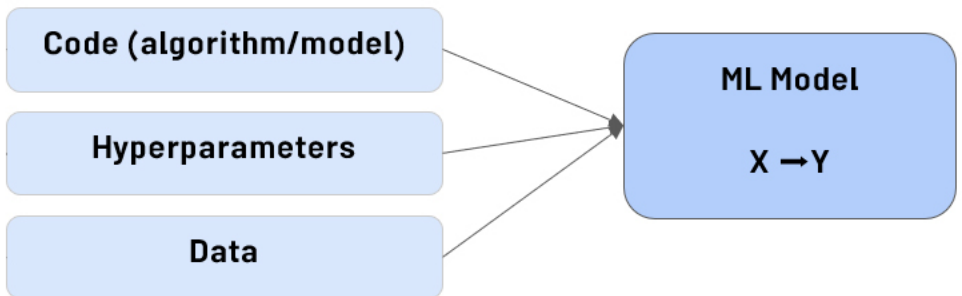
Speech recognition: modeling stage



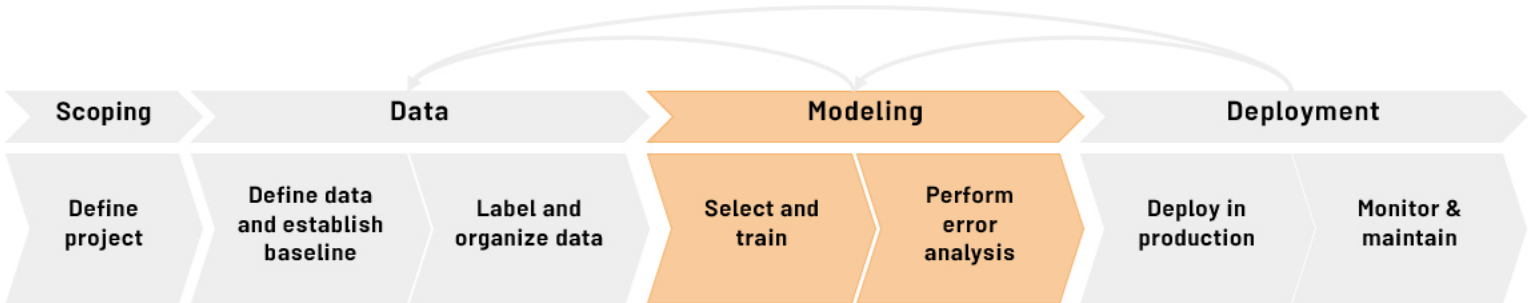
- Code (algorithm/model)
- Hyperparameters
- Data

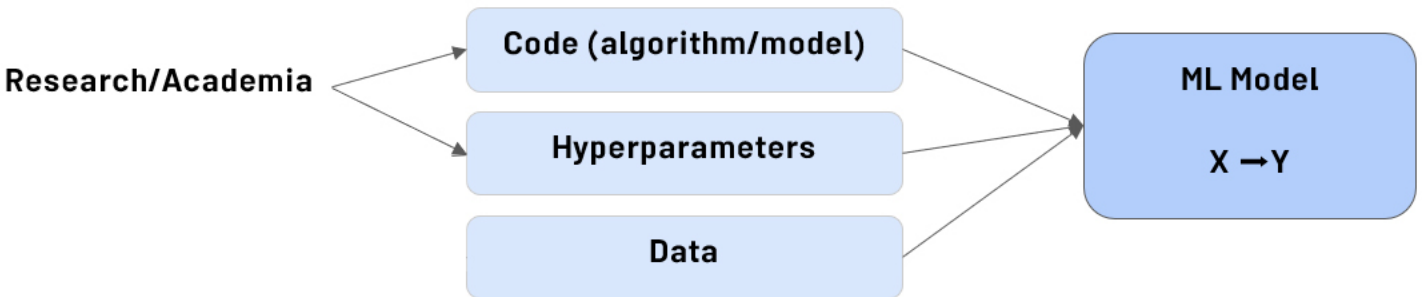
Speech recognition: modeling stage



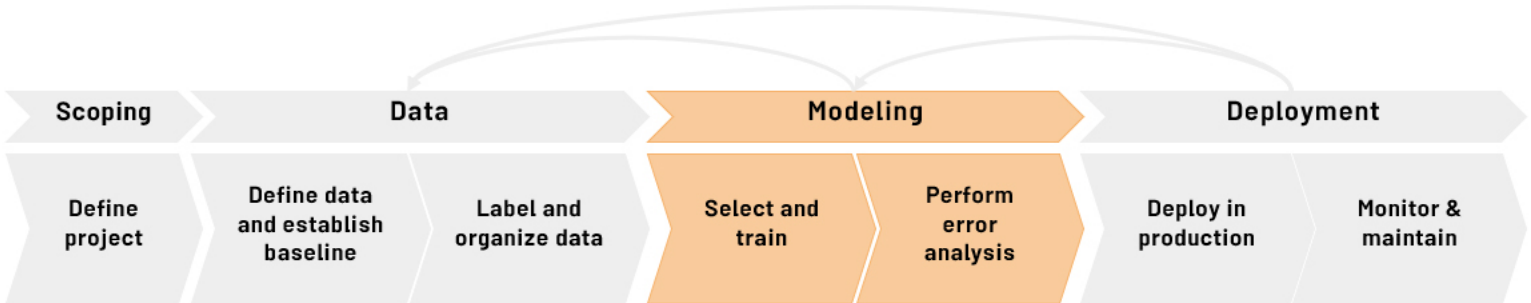


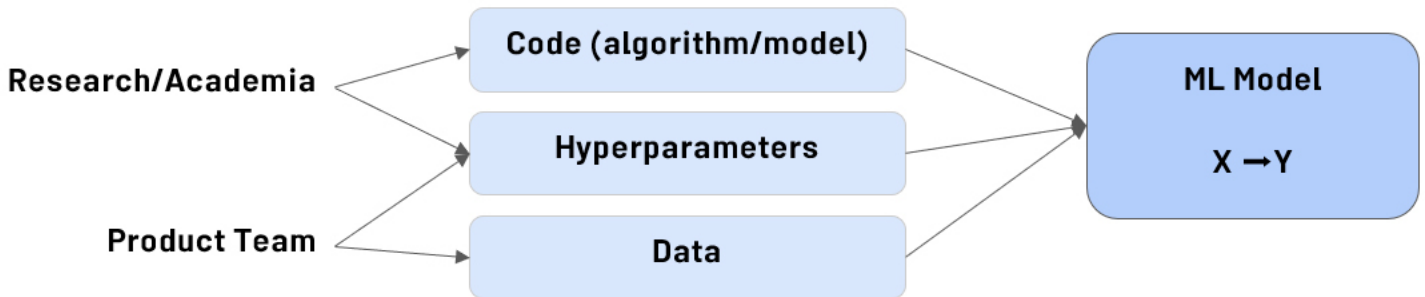
Speech recognition: modeling stage



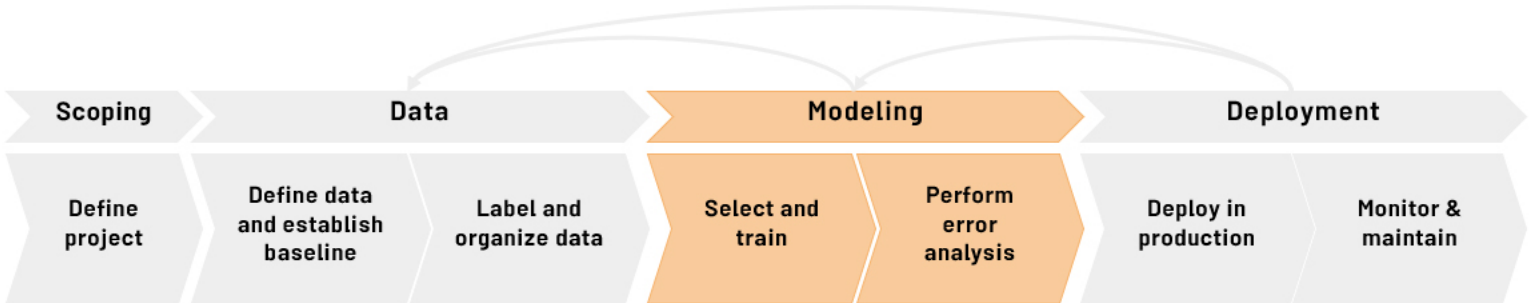


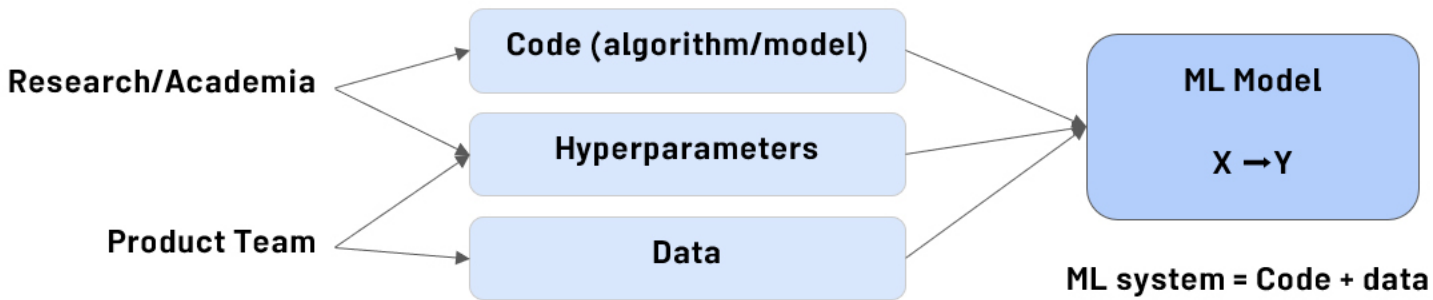
Speech recognition: modeling stage



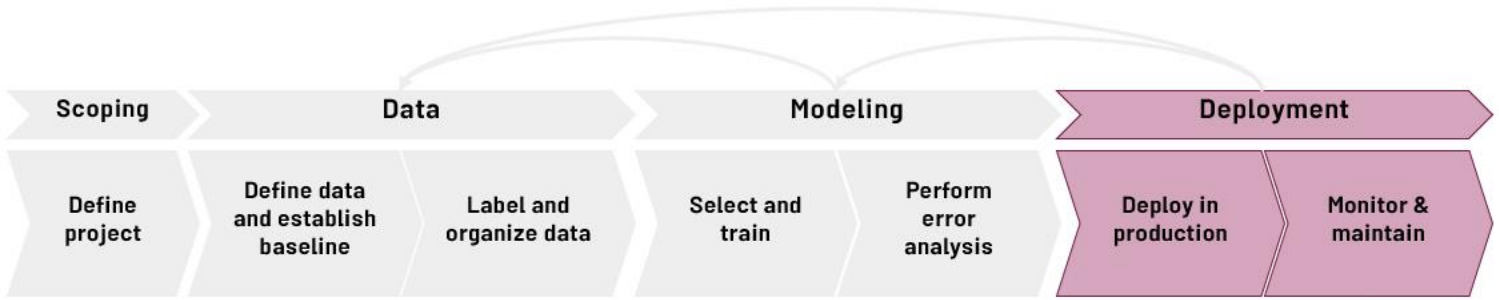


Speech recognition: modeling stage

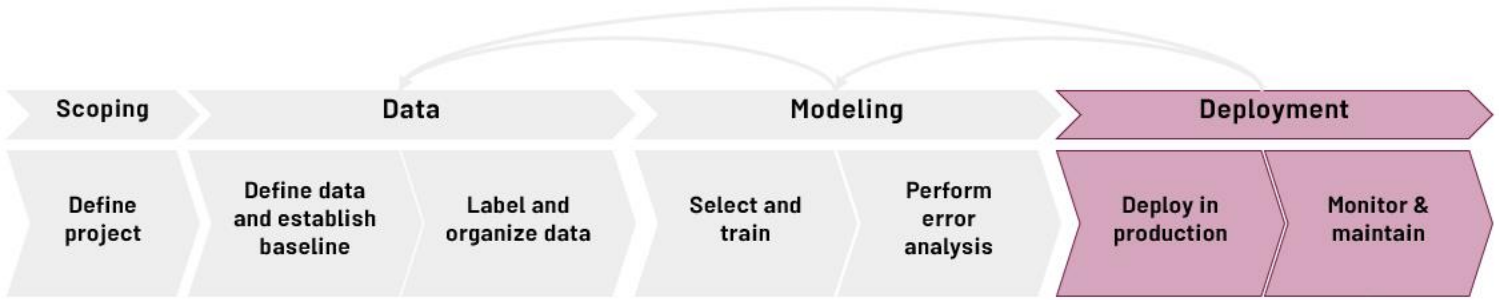


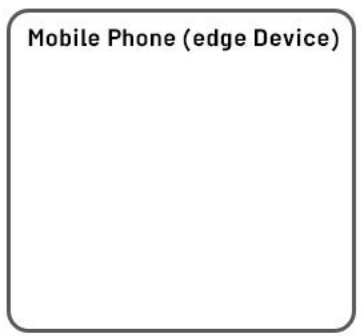


Speech recognition: modeling stage

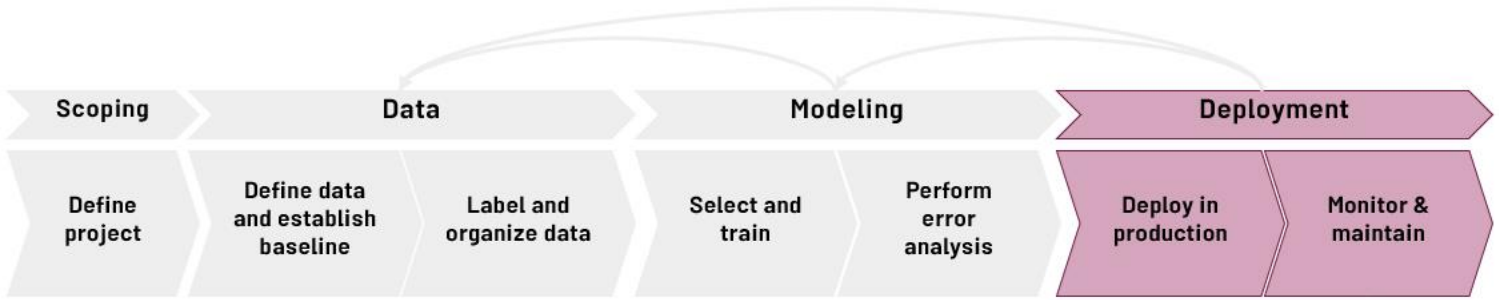


Speech recognition: modeling stage

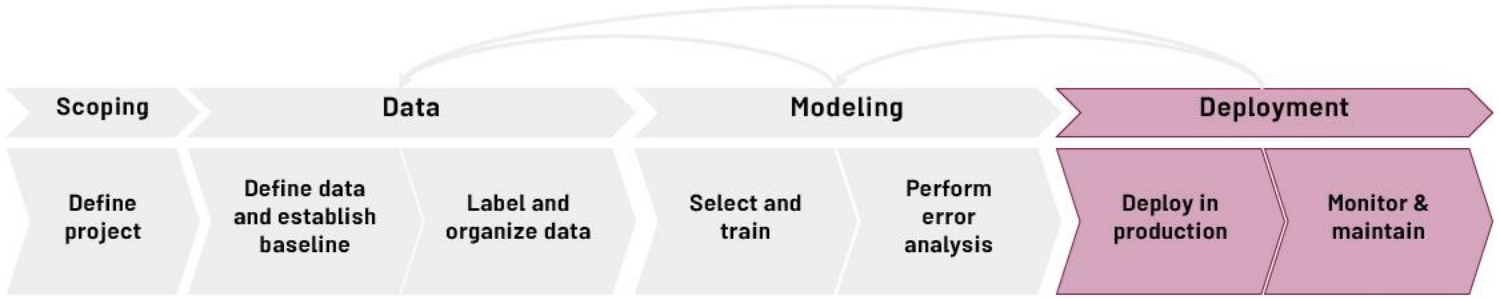




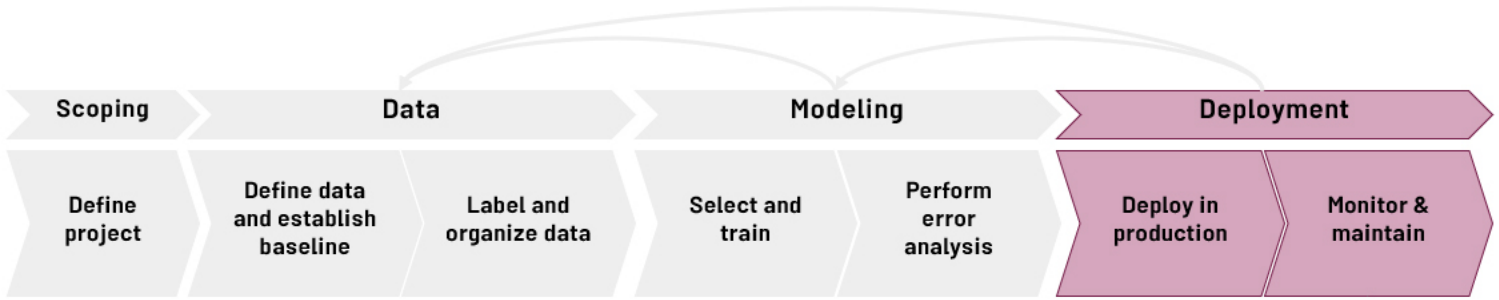
Speech recognition: modeling stage



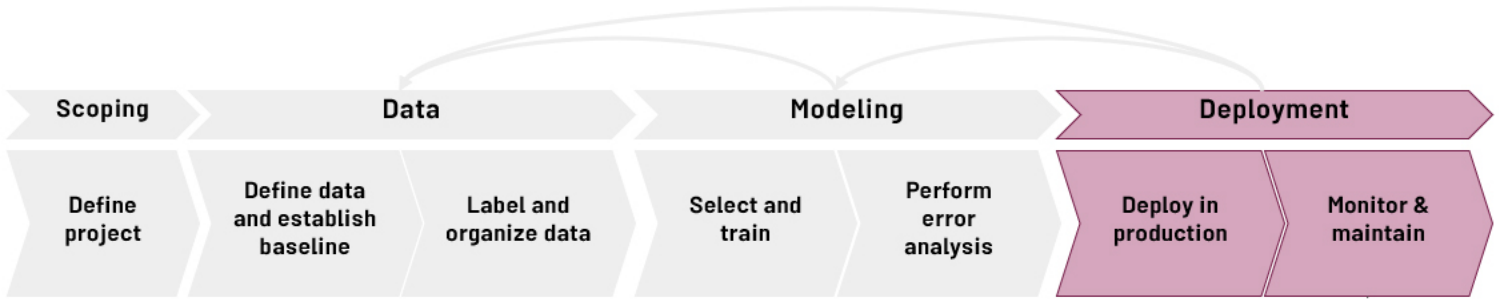
Speech recognition: modeling stage



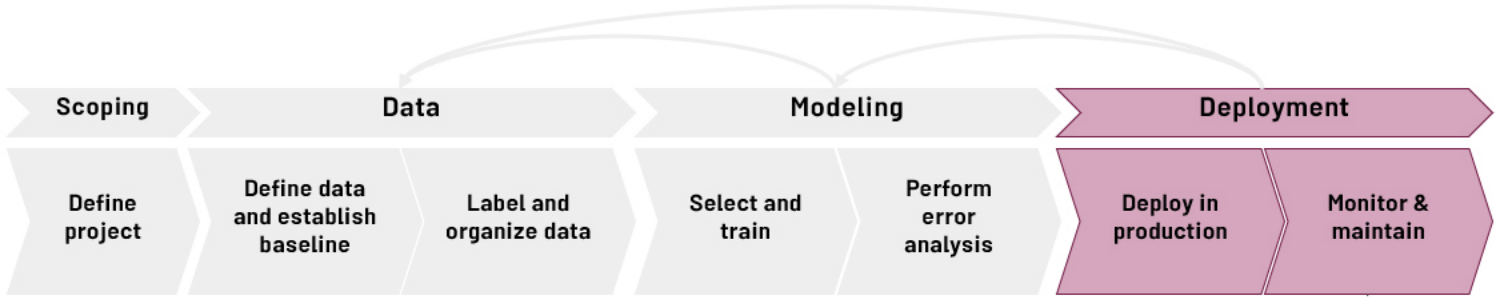
Speech recognition: modeling stage

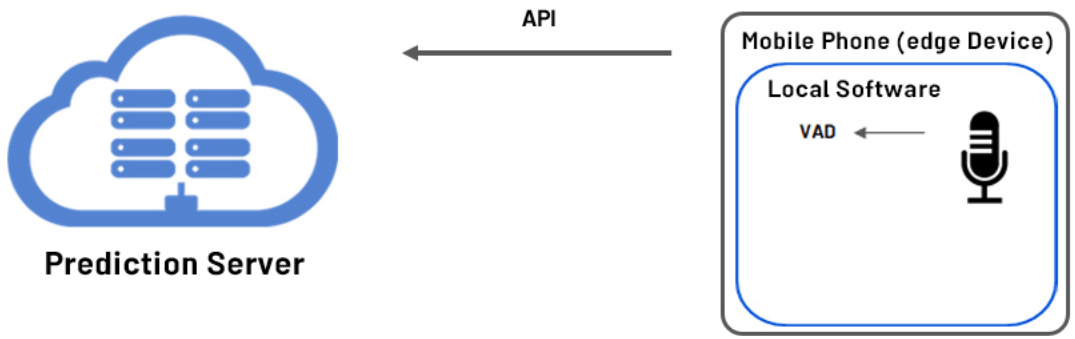


Speech recognition: modeling stage

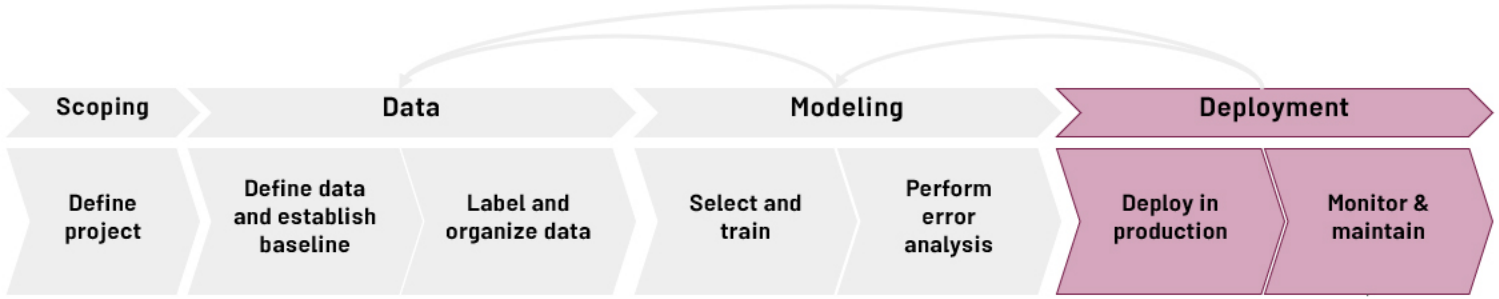


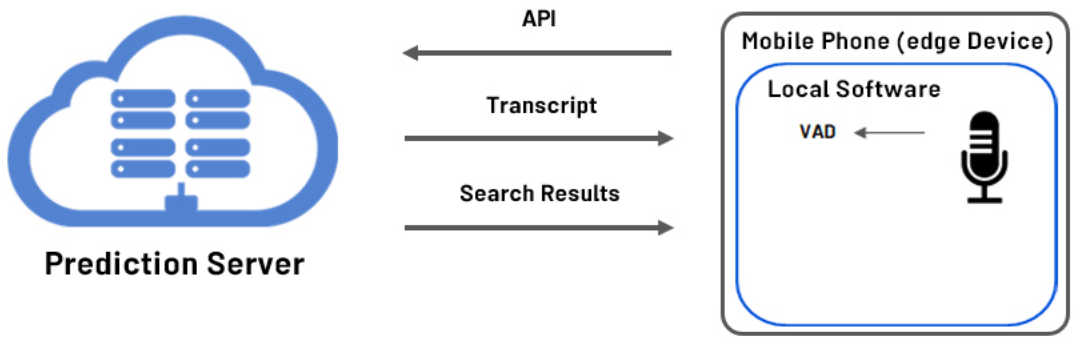
Speech recognition: modeling stage



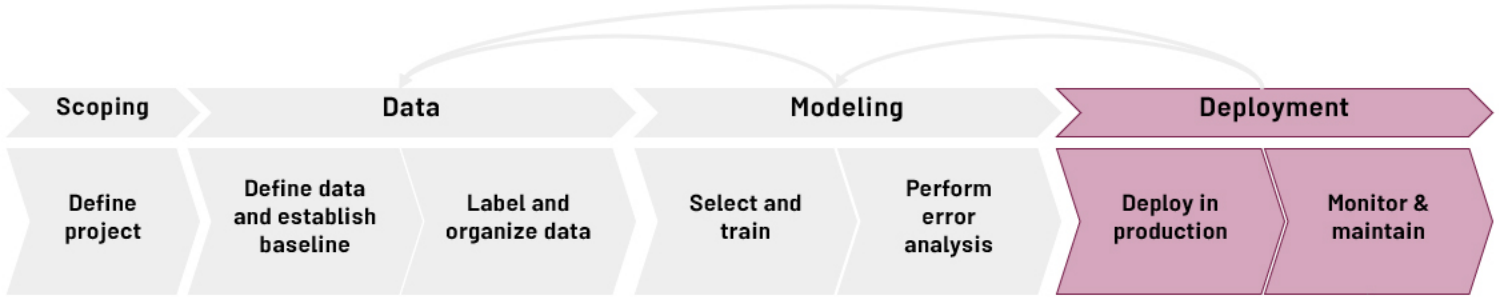


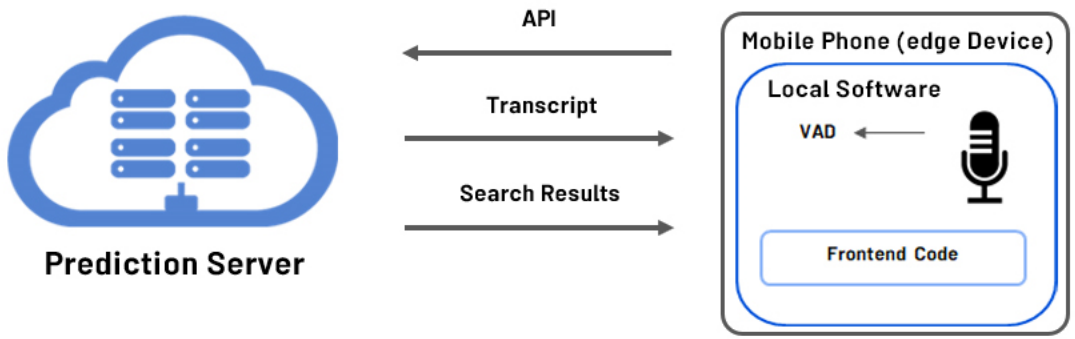
Speech recognition: modeling stage



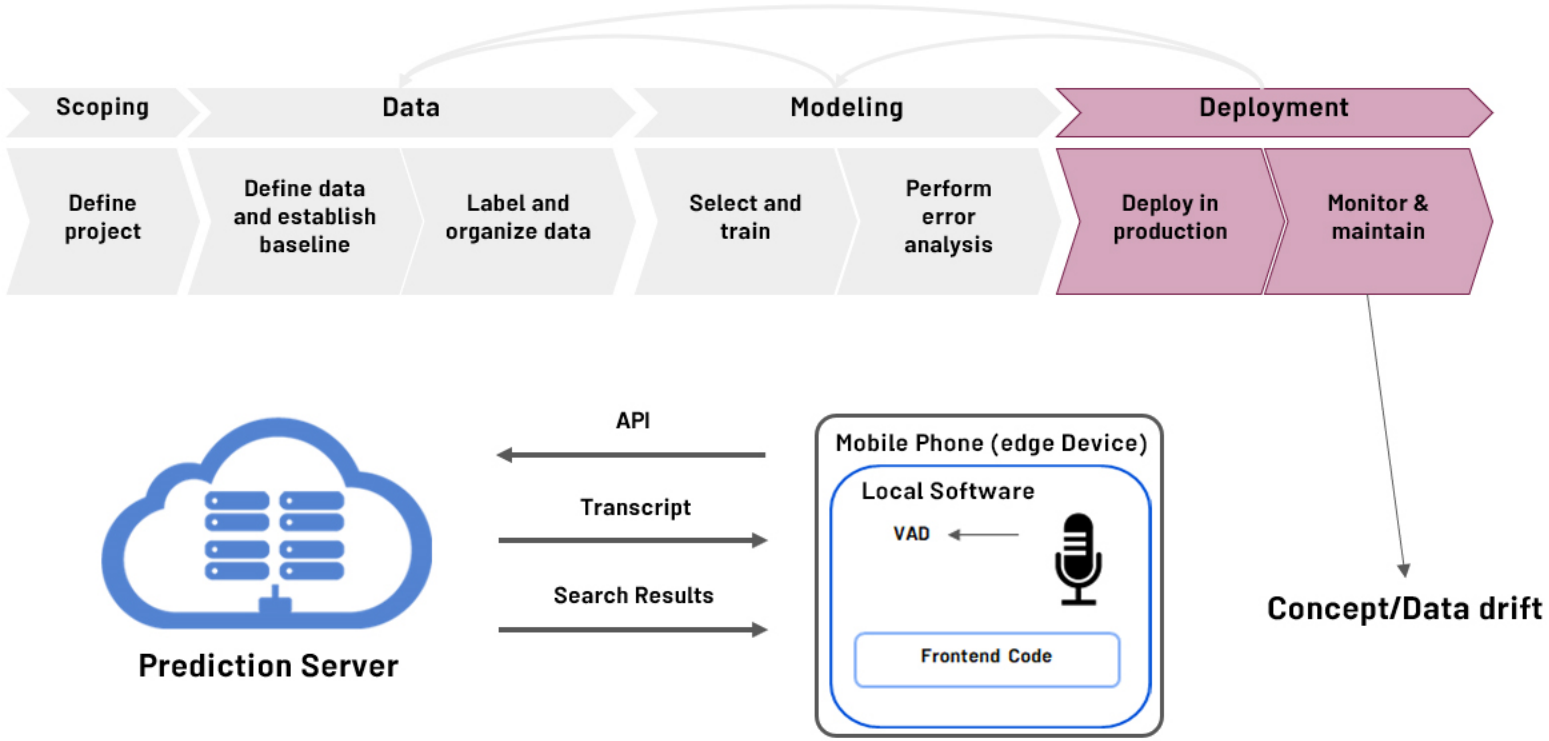


Speech recognition: modeling stage

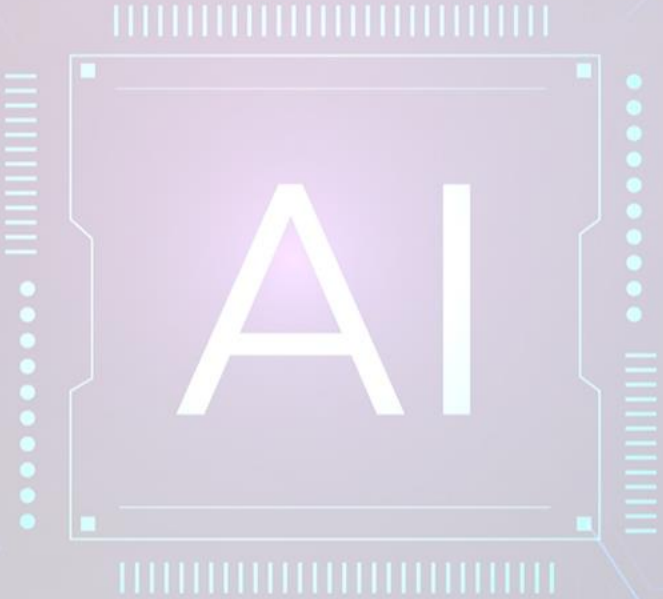




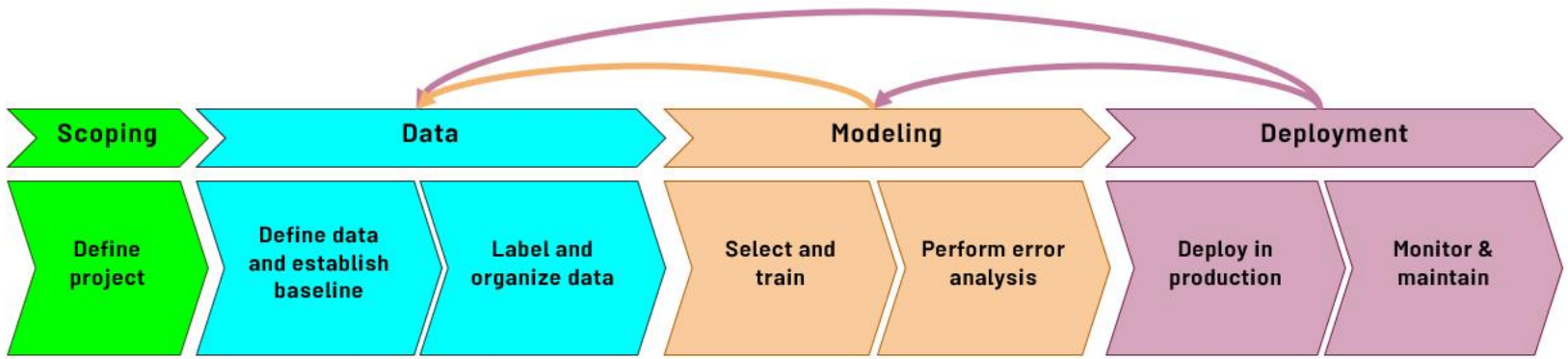
Speech recognition: modeling stage



Outline



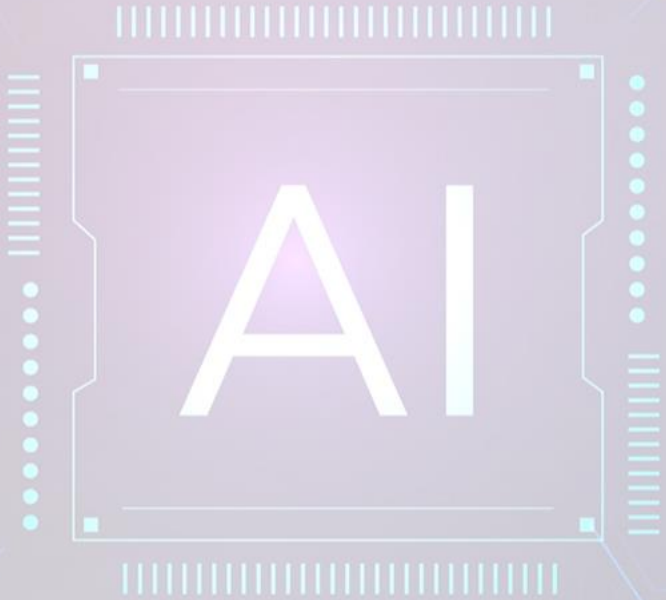
Outline



- 1- Deployment
- 2- Modeling
- 3- Data
- 4- 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

Deployment - and what makes it hard?



Concept drift and data drift

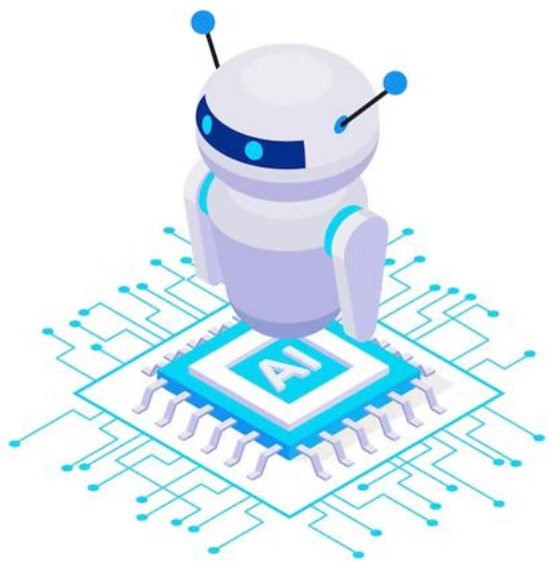
- Speech recognition example 
  - Training set:
    - Purchased data, historical user data with transcripts
  - Test set:
    - Data from a few month ago
- How has the data changed?

Gradual Change

Sudden change

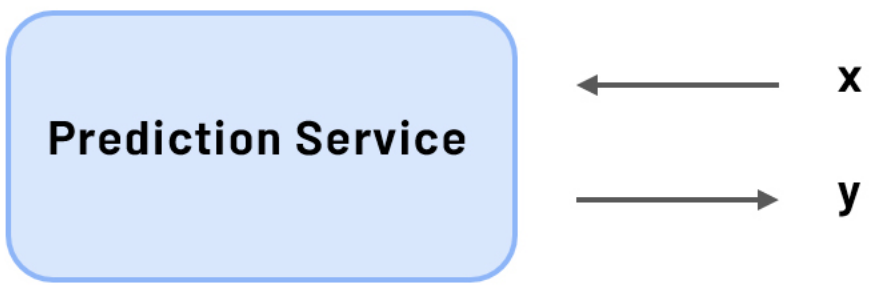
Concept drift and data drift

- **Data drift**
  - Distribution change of X
- **Concept drift**
  - Change in mapping
  - X→Y

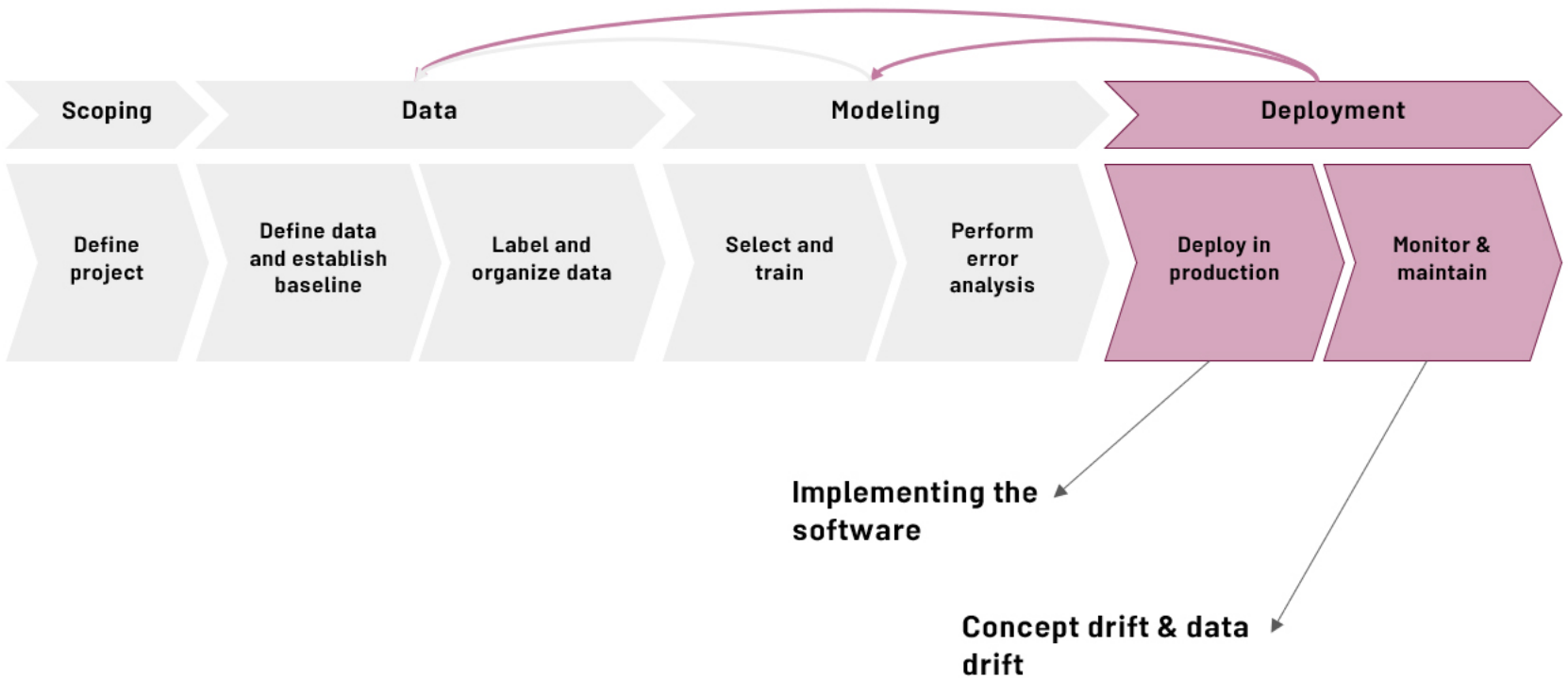


Software engineering issues

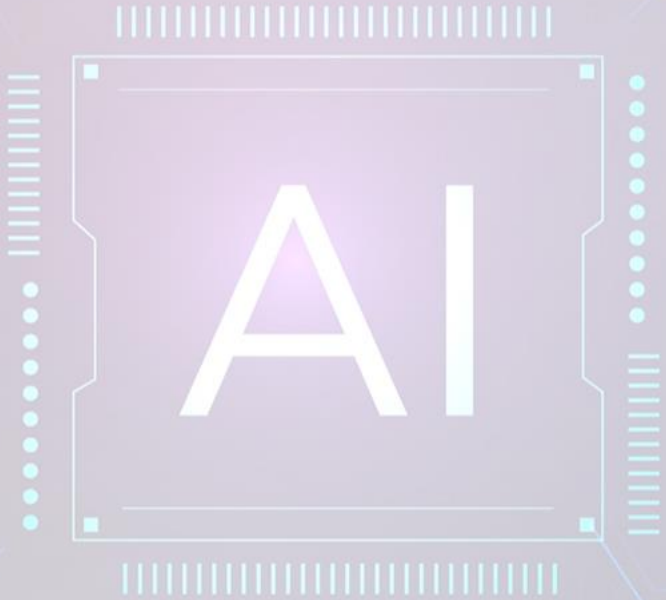
- Checklist of questions
  - Realtime or Batch
  - Cloud vs. Edge/Browser
  - Compute resources (CPU/GPU/memory)
  - Latency, throughput (QPS)
  - Logging
  - Security and privacy



First deployment vs. maintenance

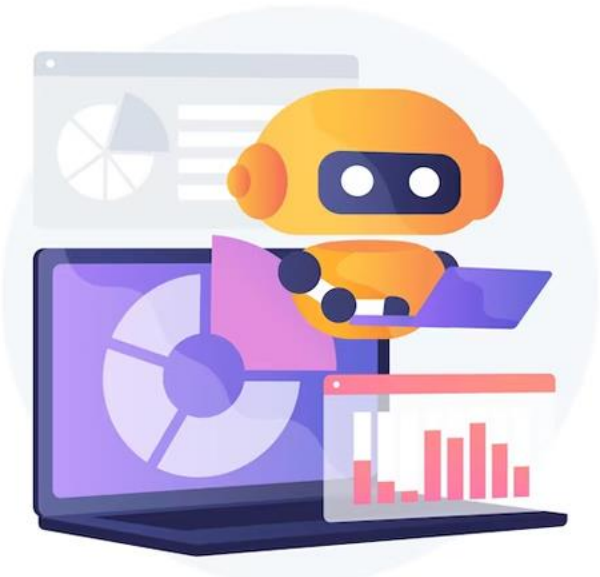


Common Deployment patterns

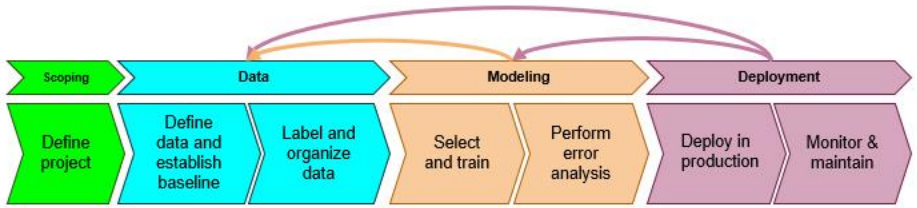


Common deployment cases

- 1- New product/capability
- 2- Automated/assist with manual task
- 3- Replace previous ML system
- Key ideas
  - Gradual ramp up with monitoring
  - Rollback



Visual inspection example



Human ✓

ML ✓

Human ✗

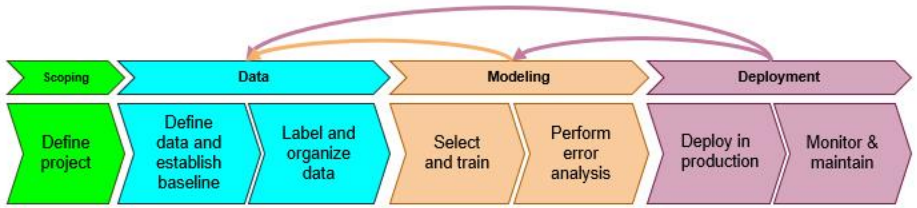
ML ✗

Human ✓

ML ✗

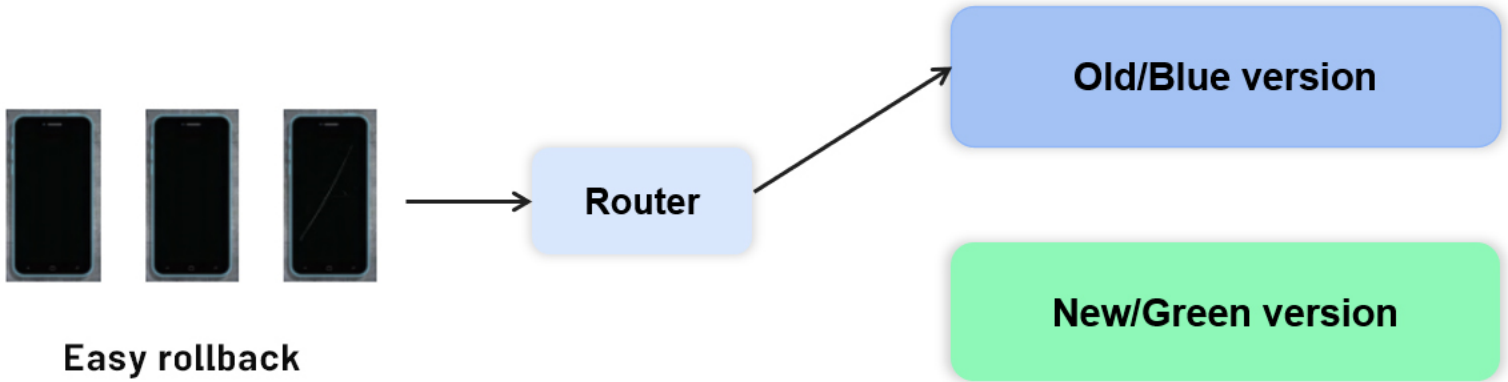
- ML system shadows the human and runs in parallel.
- ML system's output not used for any decisions during this phase.
- Sample outputs and verify predictions of ML system.

Canary deployment

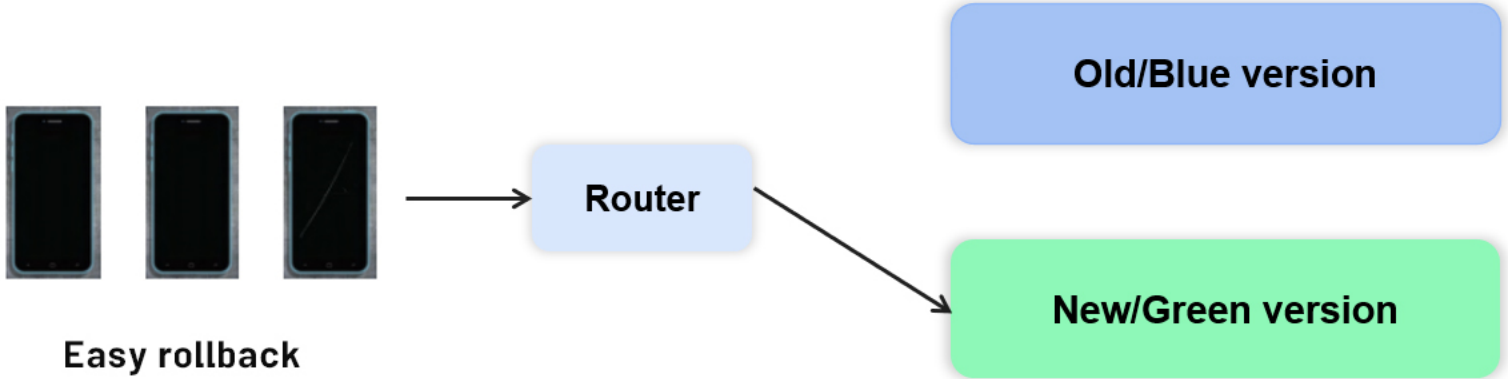


- Roll out to small fraction (say 5%) of traffic initially.
- Monitor system and ramp up traffic gradually.

Blue green deployment



Blue green deployment



Degrees of automation

Human only

Degrees of automation

Human only

Shadow Mode

Degrees of automation

Human only → Shadow Mode → AI assistant

Degrees of automation

Human only → Shadow Mode → AI assistant

Degrees of automation

Human only → Shadow Mode → AI assistant

Degrees of automation

Human only → Shadow Mode → AI assistant → Partial Automation

Degrees of automation

Human only → Shadow Mode → AI assistant → Partial Automation

Degrees of automation

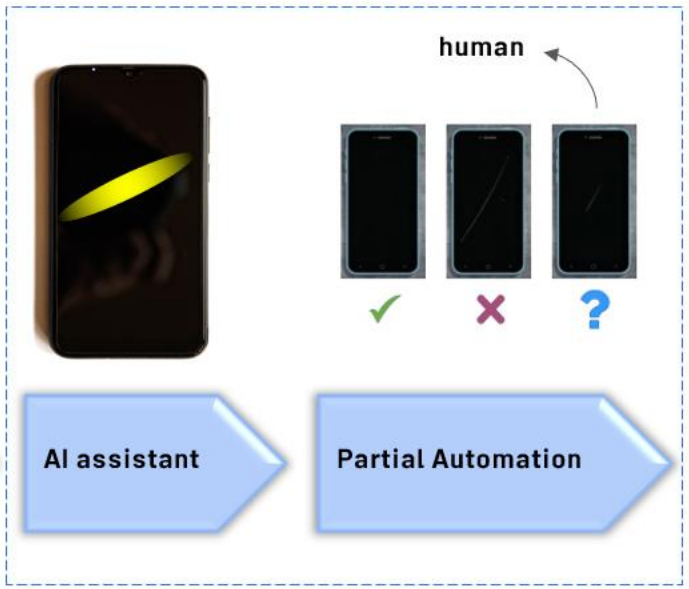
Human only → Shadow Mode → AI assistant → Partial Automation → Full Automation

Degrees of automation

Human only → Shadow Mode → AI assistant → Partial Automation → Full Automation

You can choose to stop before getting to full automation.

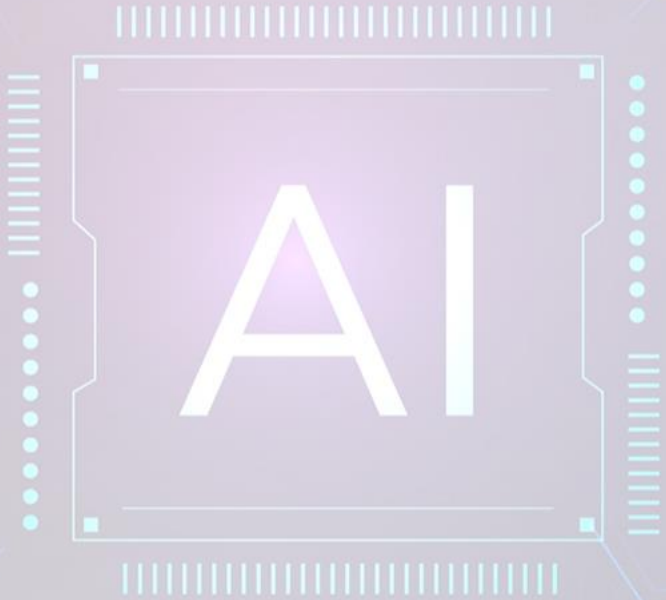
Degrees of automation



Human only → Shadow Mode → AI assistant → Partial Automation → Full Automation

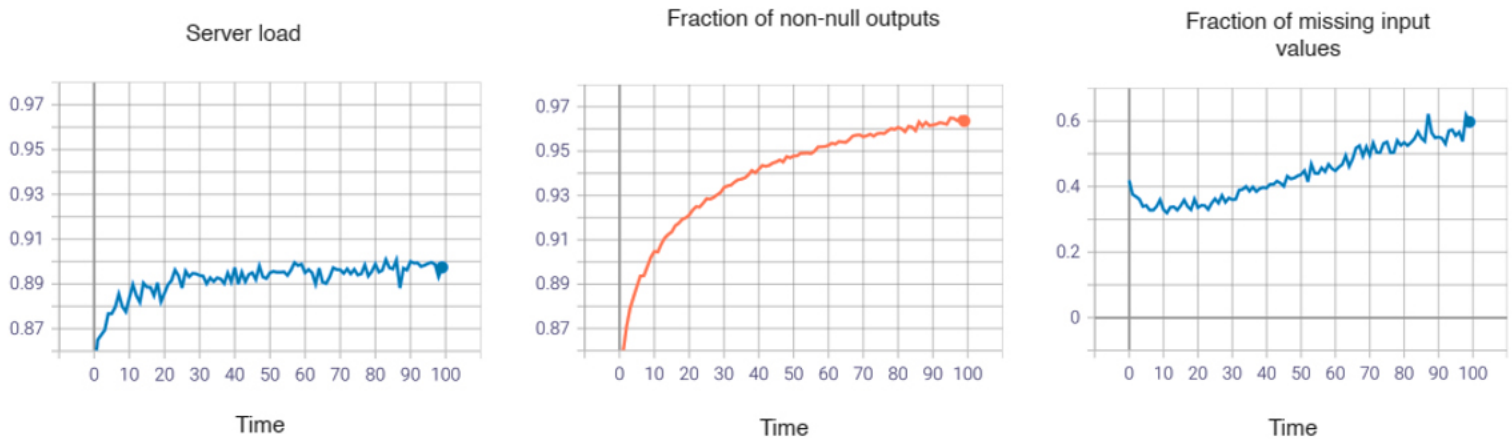
You can choose to stop before getting to full automation.

Monitoring



Monitoring dashboard

- Brainstorm a few statistics/metrics that will detect the problem.
- It is ok to use many metrics initially and gradually remove the ones you find not useful.
- Brainstorm the things that could go wrong.



Examples of metrics to track

Software Metrics	Memory, compute, latency, throughput, server load

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

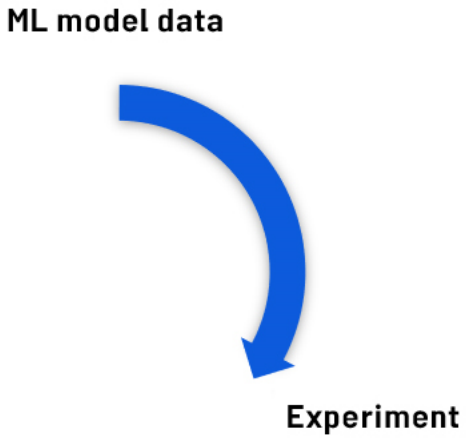
Examples of metrics to track

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

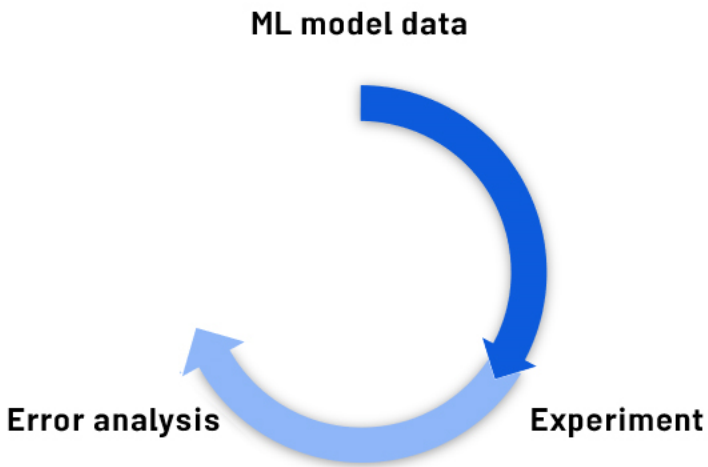
Just as ML modeling is iterative, so is deployment

ML model data

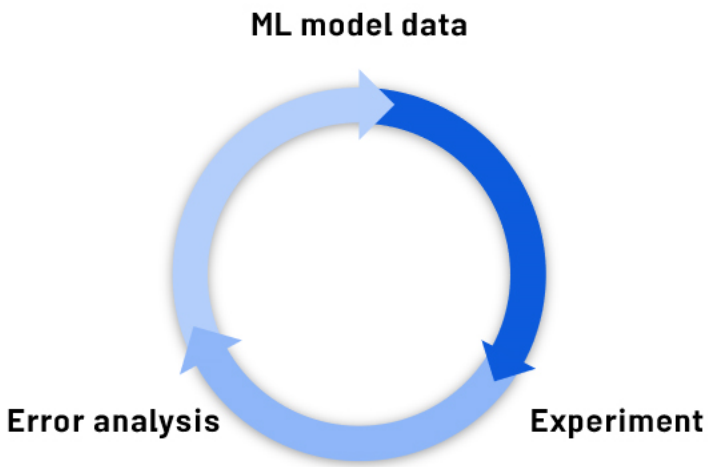
Just as ML modeling is iterative, so is deployment



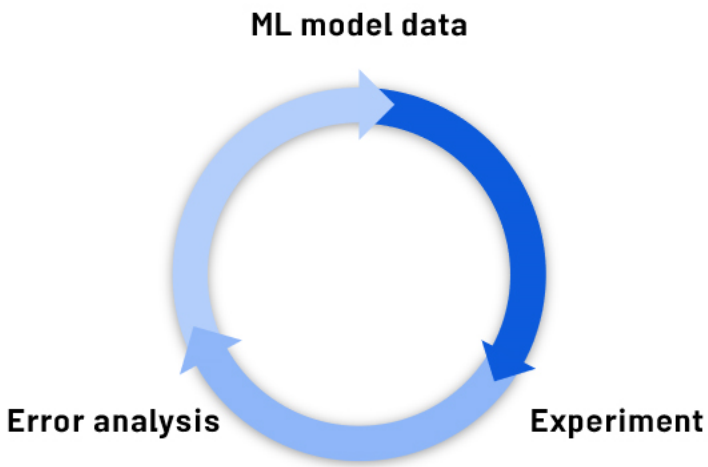
Just as ML modeling is iterative, so is deployment



Just as ML modeling is iterative, so is deployment

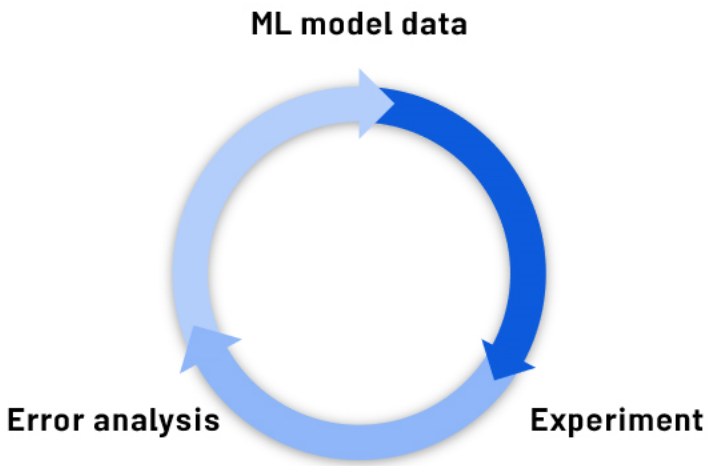


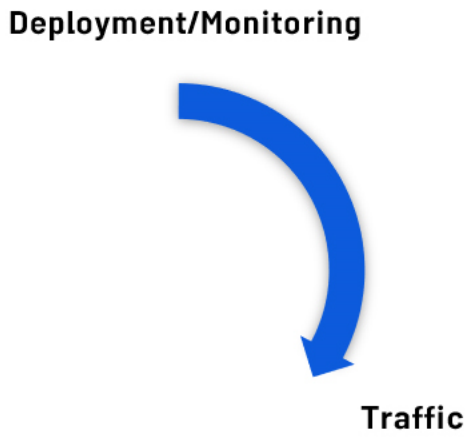
Just as ML modeling is iterative, so is deployment



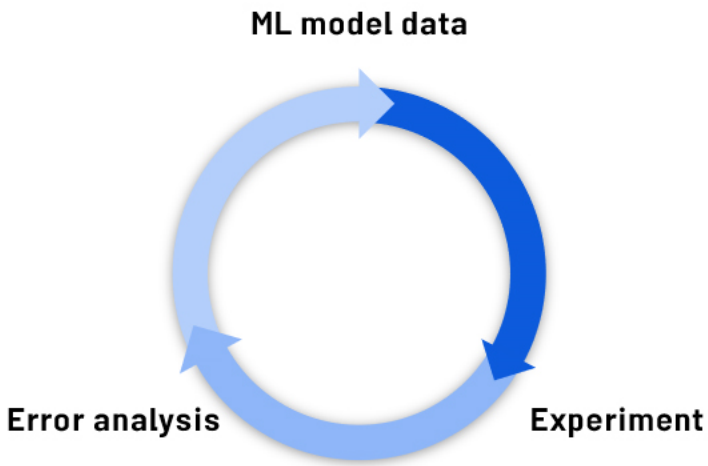
Deployment/Monitoring

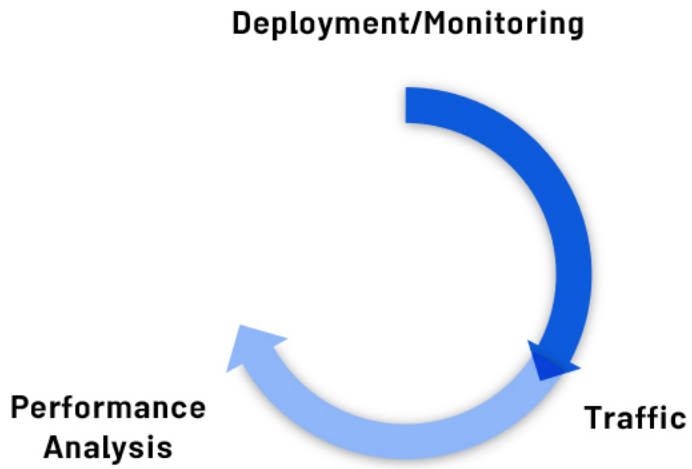
Just as ML modeling is iterative, so is deployment



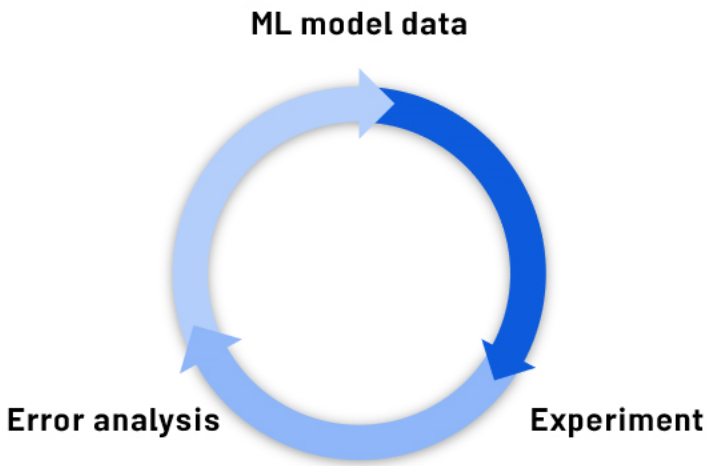


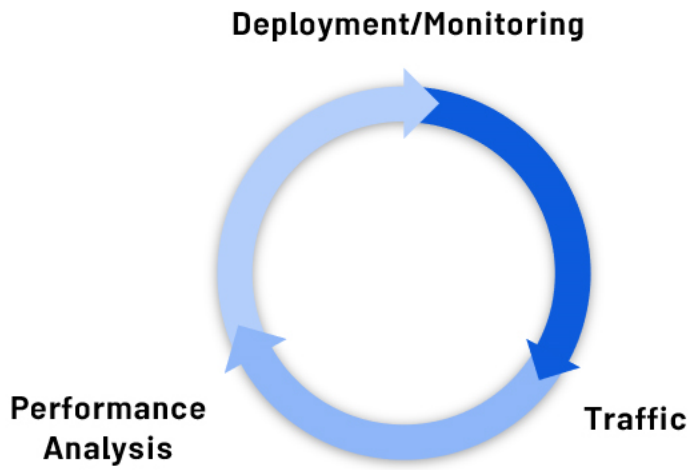
Just as ML modeling is iterative, so is deployment



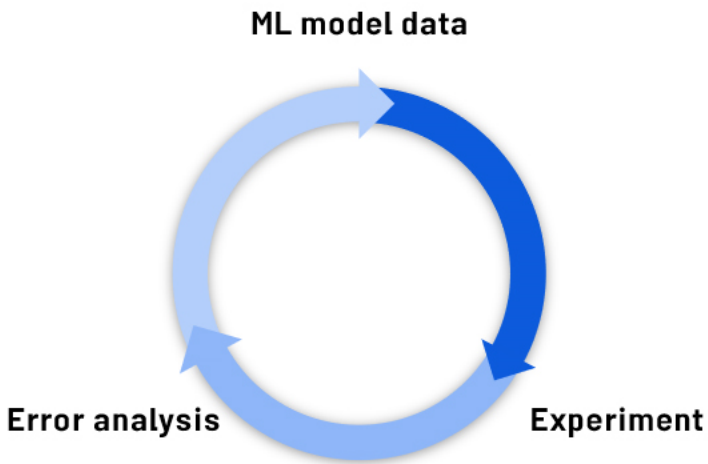


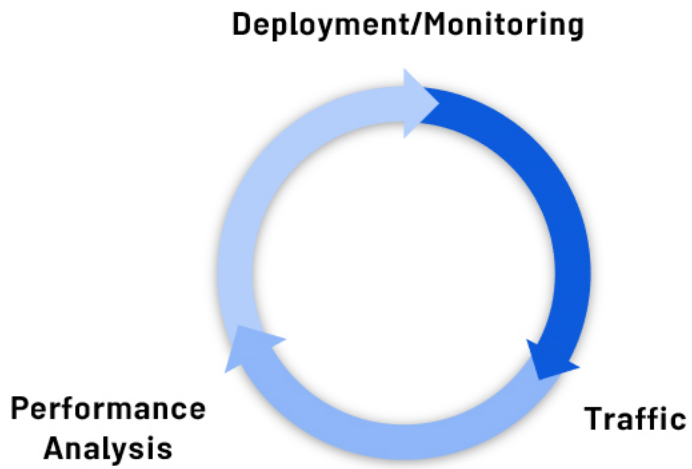
Just as ML modeling is iterative, so is deployment





Just as ML modeling is iterative, so is deployment

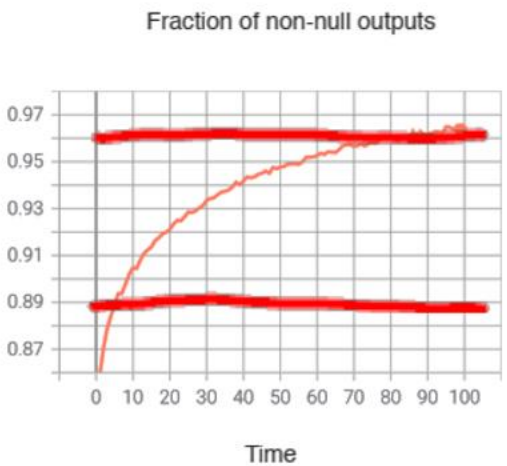


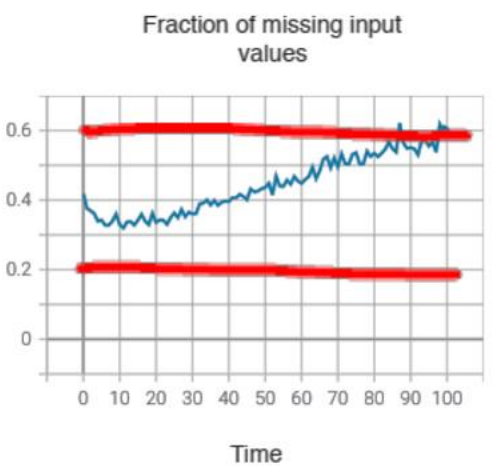


Iterative process to choose the right set of metrics to monitor.

Monitoring Dashboard

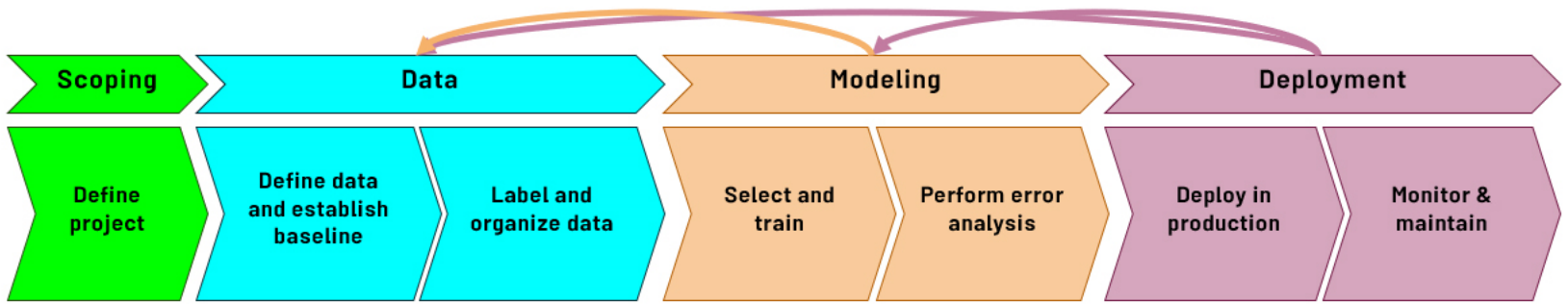
- Adapt metrics and thresholds over time
- Set thresholds for alarms



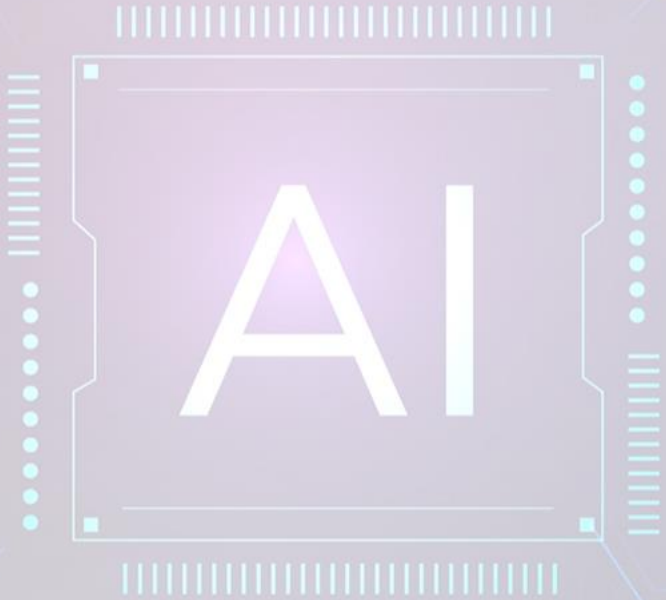


Model Maintenance

- Manual retraining
- Automatic retraining



Pipeline monitoring



Speech recognition example

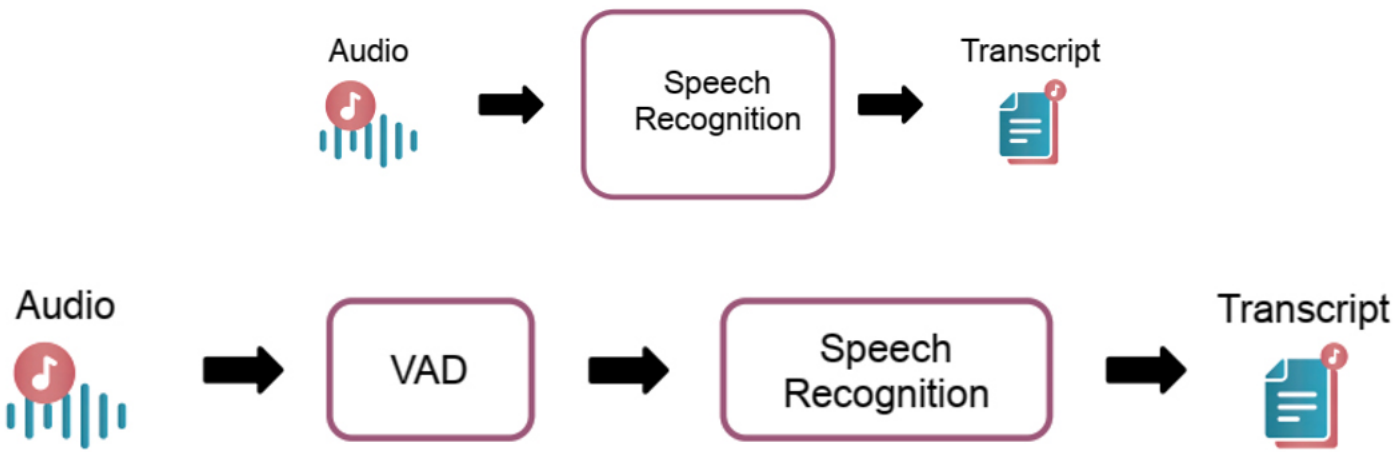
- Some cell phones might have VAD clip audio differently, leading to degraded performance

Speech recognition example

- Some cell phones might have VAD clip audio differently, leading to degraded performance

Speech recognition example

- Some cell phones might have VAD clip audio differently, leading to degraded performance

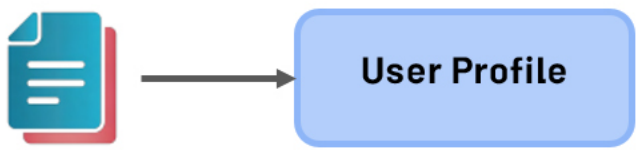


User profile example

User data

User profile example

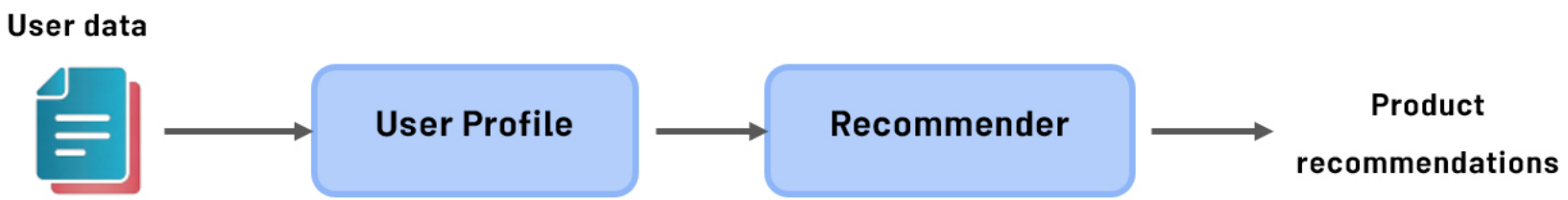
User data



User profile example

User data

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.