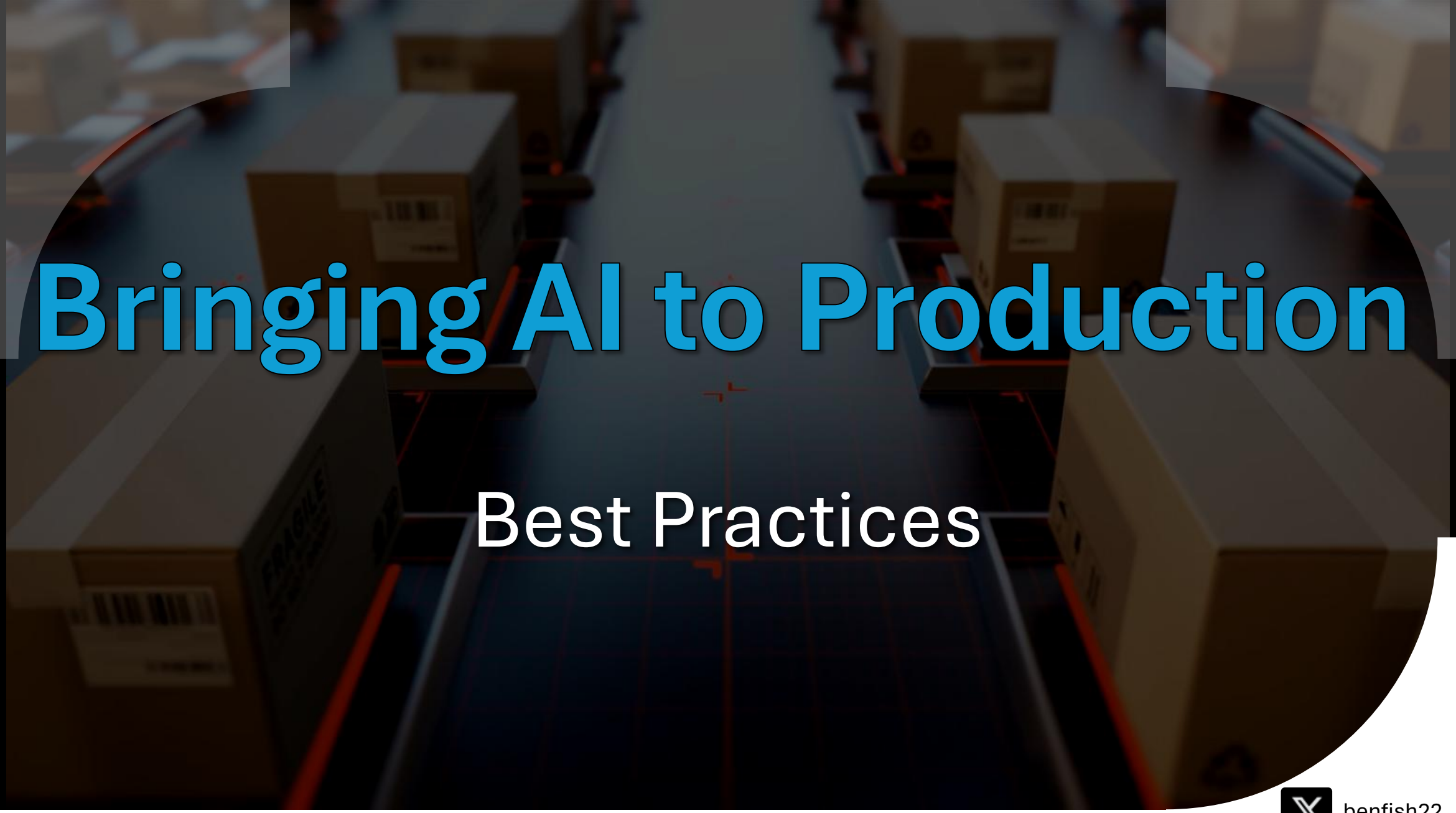


Bringing AI to Production

Best Practices

Ben Fishman

benfish22

benf22.github.io

Yearly investment in AI:

>100$ Billion

Thousands of companies

>20k papers

Yearly investment in AI:

>100$ Billion

Thousands of companies

>20k papers

And still...

80% - 90% of the AI projects **FAIL**

Why?

How to Improve?

Ben Fishman

Education

- Ph.D. Candidate – Computer Science
- M.Sc. Electrical Engineering
- B.Sc. Bio Medical Engineering

Industry

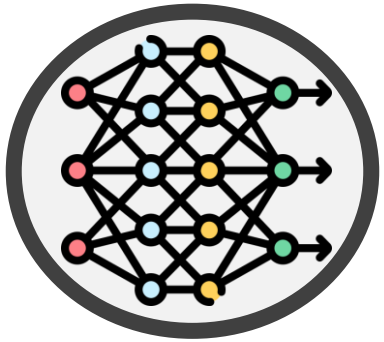
- Director of AI & Algo @ Microsoft
- Senior Data Scientist @ Microsoft
- Team Lead @ Mobileye

Background

- Computer Vision
- Audio
- Speech
- ML/ DL/ GenAI

The Main Reasons for Failure

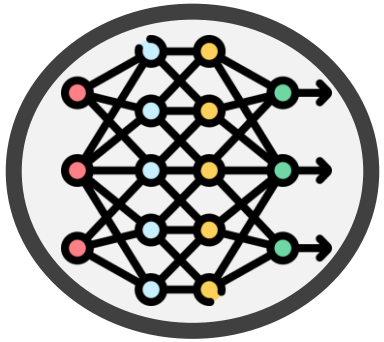
The Main Reasons for Failure



AI doesn't fit the
Core Problem

The Main Reasons for Failure

Wrong **Trajectory**

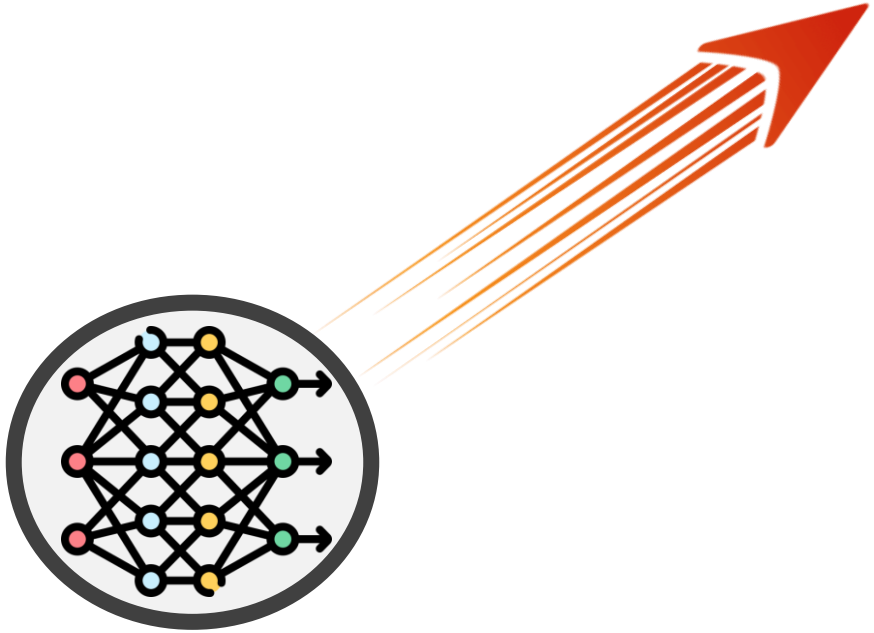


AI doesn't fit the **Core Problem**

The Main Reasons for Failure

Inefficient Development Process

Wrong **Trajectory**

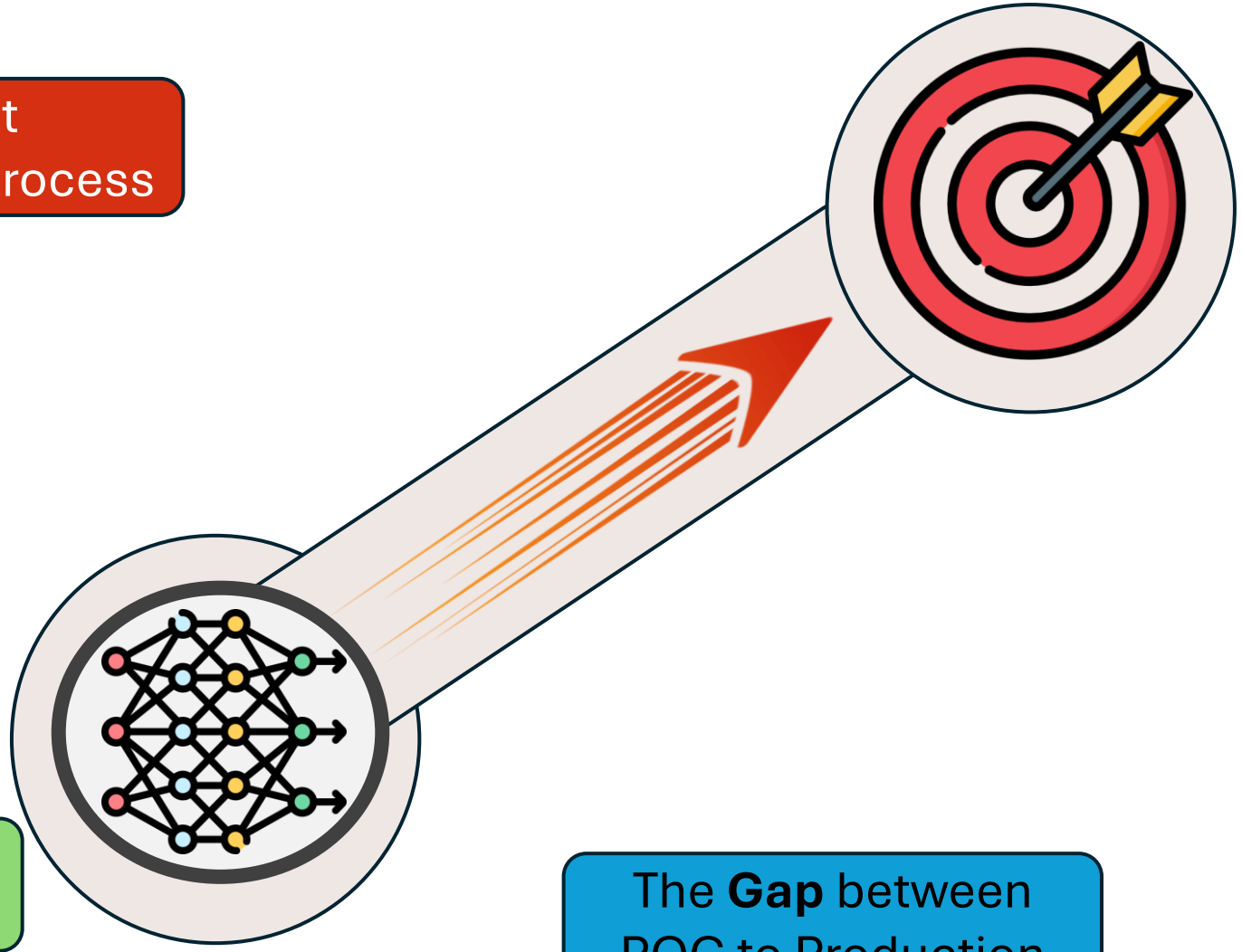


AI doesn't fit the **Core Problem**

The Main Reasons for Failure

Inefficient Development Process

Wrong **Trajectory**



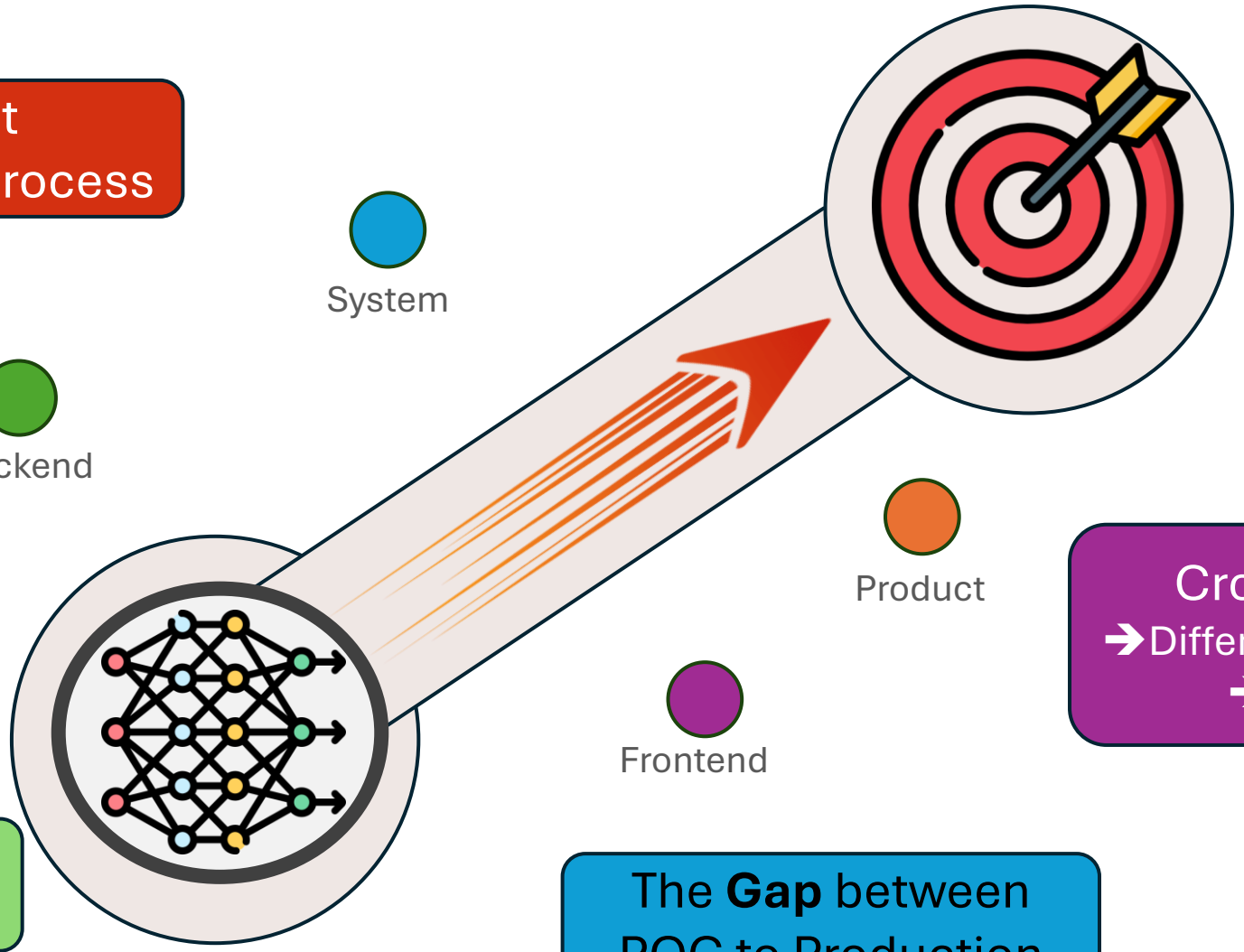
AI doesn't fit the **Core Problem**

The **Gap** between POC to Production

The Main Reasons for Failure

Inefficient Development Process

Wrong **Trajectory**



Cross Domain Teams
➔Different Development Process
➔ Lack of Trust in AI

AI doesn't fit the **Core Problem**

The **Gap** between POC to Production

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

Application Oriented Algo

Application Oriented Algo

Run E2E (sooner rather than later)

- Better **understanding the app**
- **Simple** (off-the shelf) → **Complex**

Application Oriented Algo

App: Detecting pedestrians crossing the road

Run E2E (sooner rather than later)

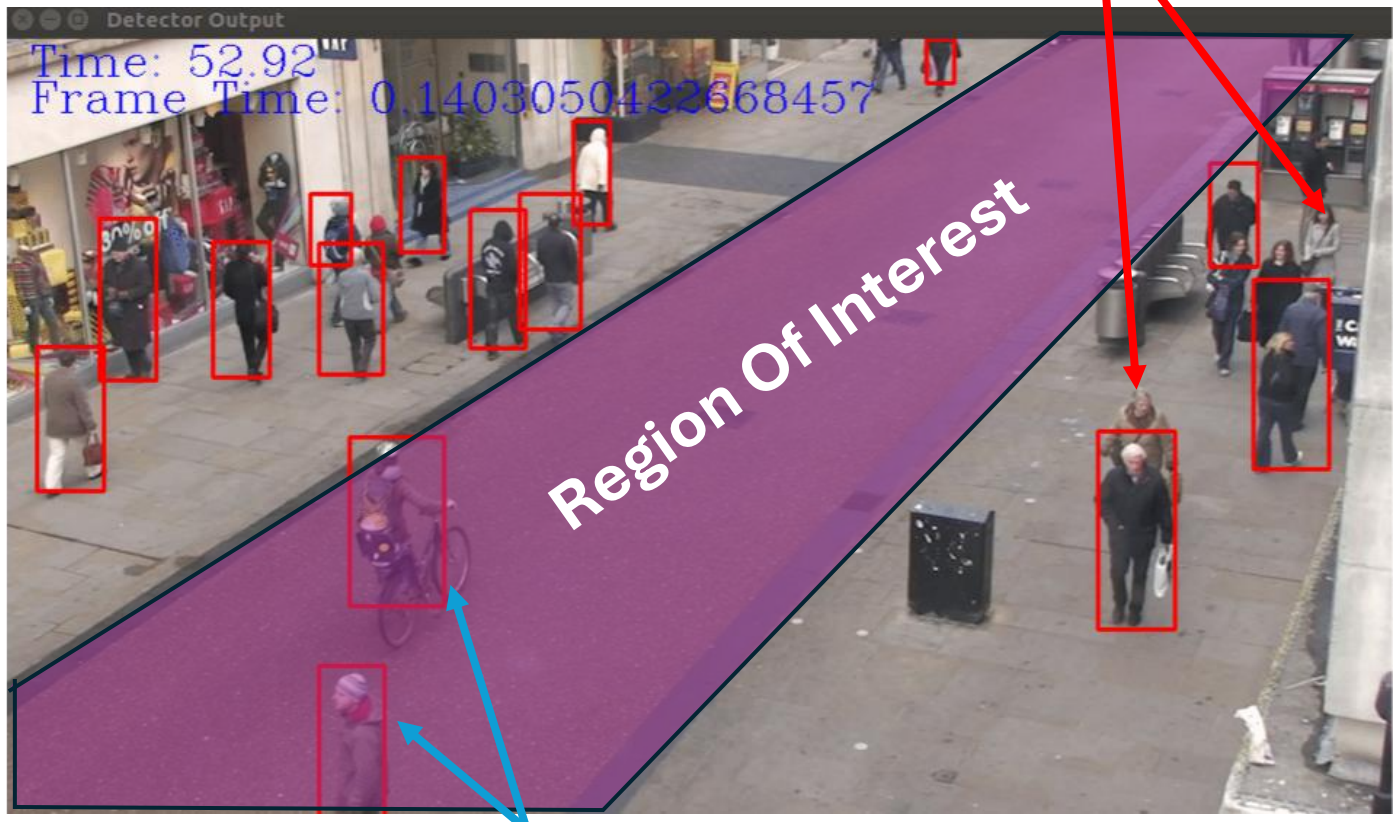
- Better **understanding the app**
- **Simple** (off-the shelf) → **Complex**

Be Relevant

- **Test** (only) what you need
- **Fix** (only) what you need

Application Oriented Algo

App: Detecting pedestrians crossing the road



Run E2E (sooner rather than later)

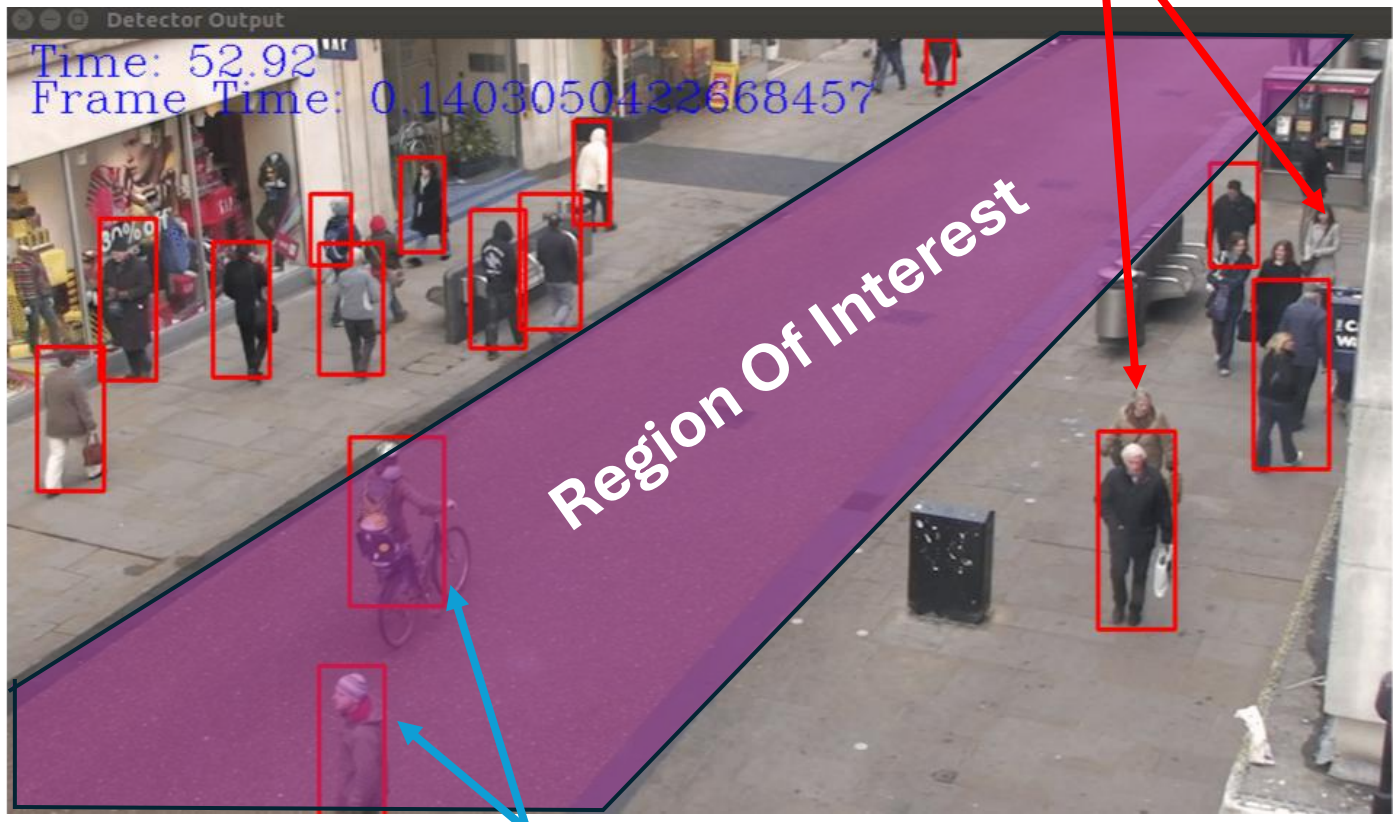
- Better **understanding the app**
- **Simple** (off-the shelf) → **Complex**

Be Relevant

- **Test** (only) what you need
- **Fix** (only) what you need

Application Oriented Algo

App: Detecting pedestrians crossing the road



Run E2E (sooner rather than later)

- Better **understanding the app**
- **Simple** (off-the shelf) → **Complex**

Be Relevant

- **Test** (only) what you need
- **Fix** (only) what you need

Align with other stake holders

- **Product**
- **System**

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

Invest in Data & Evaluation

Evaluation

Data

Invest in Data & Evaluation

Evaluation

Data

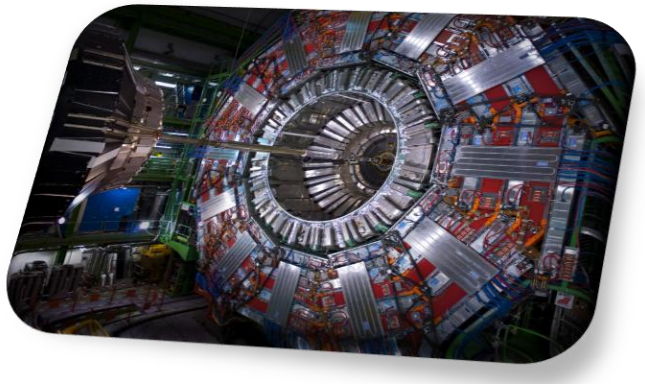
❖ **Companies tend to underestimate their importance**

Invest in Data & Evaluation

Evaluation

Data

- Companies tend to underestimate their importance
- The compass & accelerator of the development process



Invest in Data & Evaluation

Evaluation

Invest in Data & Evaluation

Evaluation

Evaluation Engine

Accurate

Reliable

Easy to use

Fast

Focuses you

Raises problems fast

Provides confidence

Time Investment!

Invest in Data & Evaluation

Evaluation

Evaluation Engine

Accurate

Reliable

Easy to use

Fast

→

Focuses you

Raises problems fast

Provides confidence

Time Investment!

Metrics

Raw Metrics
Not enough

→

Refined Metrics
Fits your application

Invest in Data & Evaluation

Evaluation

Evaluation Engine

Accurate

Reliable

Easy to use

Fast

Focuses you

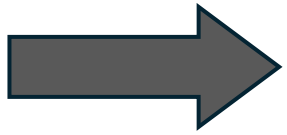
Raises problems fast

Provides confidence

Time Investment!

Metrics

Raw Metrics
Not enough



Refined Metrics
Fits your application

Human in The Loop

Invest in Data & Evaluation

Data

Invest in Data & Evaluation

Data

- Data Collection
- Data Annotation
- Data Analysis
- Data Engineering
- Data Roles
- Datasets Creation

- Legal
- Privacy
- Responsible AI
- Curation
- Auto Labeling

Invest in Data & Evaluation

Data

- Data Collection
- Data Annotation
- Data Analysis
- Data Engineering
- Data Roles
- Datasets Creation

Data Roles

- Data Manager
- Data Analyst
- Data Engineers
- Data Collectors
- Data Annotators

- Legal
- Privacy
- Responsible AI
- Curation
- Auto Labeling

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

Make it a Factory

Make it a Factory

Problem: Many technical challenges ➔ Inefficient work + Focus distraction

Make it a Factory

Problem: Many technical challenges ➔ Inefficient work + Focus distraction
Solution: Solve them

Make it a Factory

Problem: Many technical challenges ➔ Inefficient work + Focus distraction
Solution: Solve them

It's not appealing and not deliverable – BUT it is crucial for success

Make it a Factory

Problem: Many technical challenges ➔ Inefficient work + Focus distraction
Solution: Solve them

It's not appealing and not deliverable – BUT it is crucial for success

Infra/ MLOps

Commonality

Teamwork

Make it a Factory

Problem: Many technical challenges ➔ Inefficient work + Focus distraction
Solution: Solve them

It's not appealing and not deliverable – BUT it is crucial for success

Infra/ MLOps

- Data Pipelines
- Parallelization
- Training Acceleration

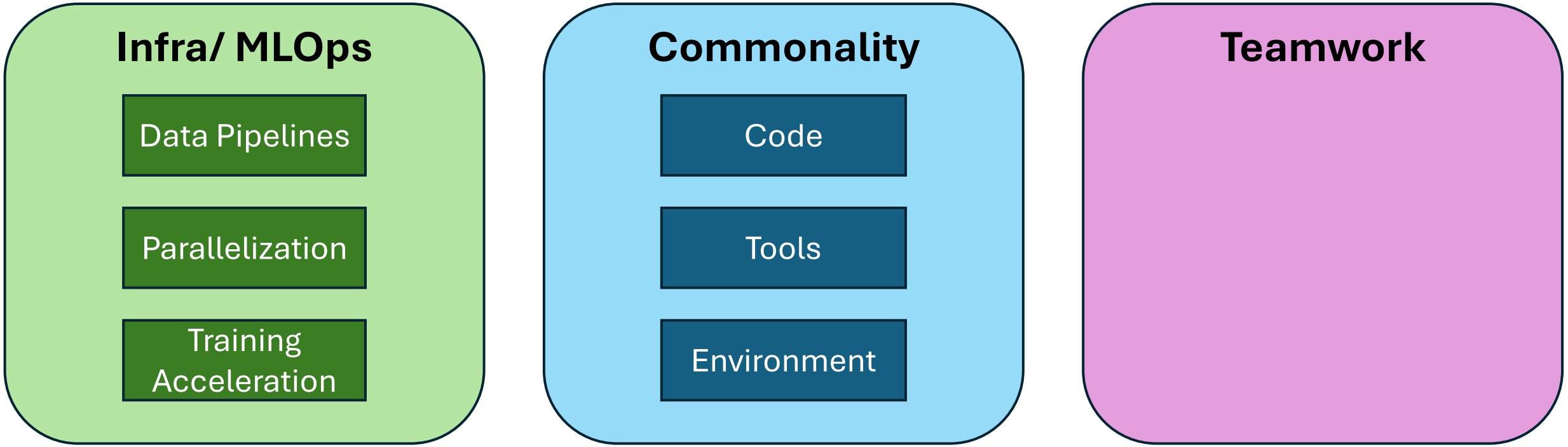
Commonality

Teamwork

Make it a Factory

Problem: Many technical challenges ➔ Inefficient work + Focus distraction
Solution: Solve them

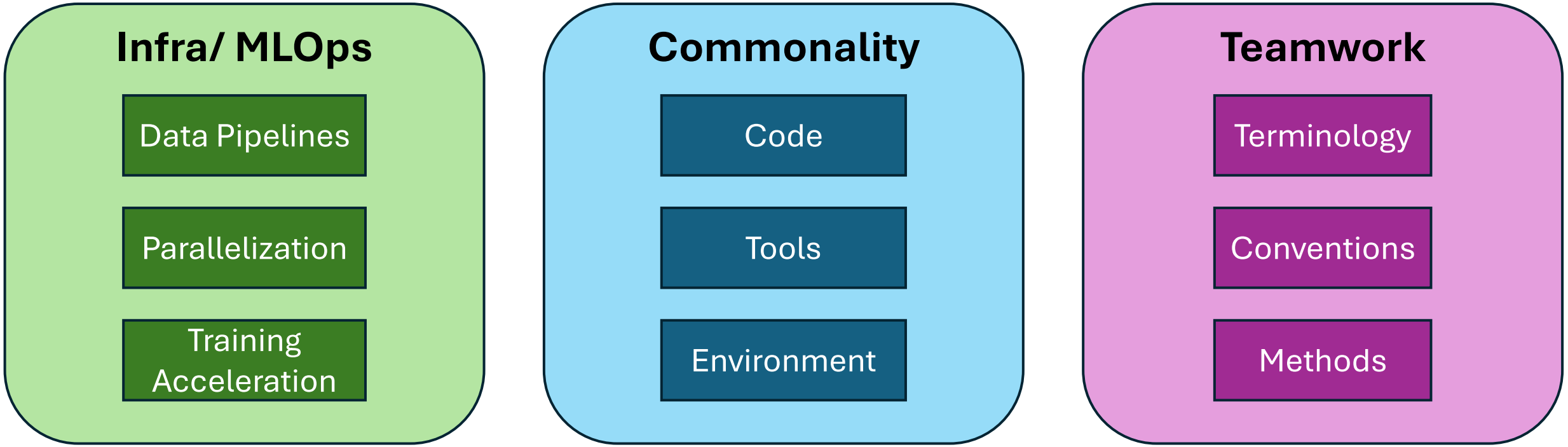
It's not appealing and not deliverable – BUT it is crucial for success



Make it a Factory

Problem: Many technical challenges ➔ Inefficient work + Focus distraction
Solution: Solve them

It's not appealing and not deliverable – BUT it is crucial for success



The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

Divide & Conquer

Divide **large** problems into **smaller** problems ➔

Easier to solve them

Divide & Conquer

Time

Modules

People

Divide **large** problems into **smaller** problems ➔

Easier to solve them

Divide & Conquer

Time

Modules

People

Divide & Conquer

Time

Modules

People

Split your timeline into smaller segments:

- To improve in small **incremental steps**
- To **name & control** your status
- To be able to **reproduce**
- To show **progress**

Divide & Conquer

Time

Modules

People

Split your timeline into smaller segments:

- To improve in small **incremental steps**
- To **name & control** your status
- To be able to **reproduce**
- To show **progress**

Version everything

Code Version

Algo Version

Data Version

Dataset Version

Metrics Version

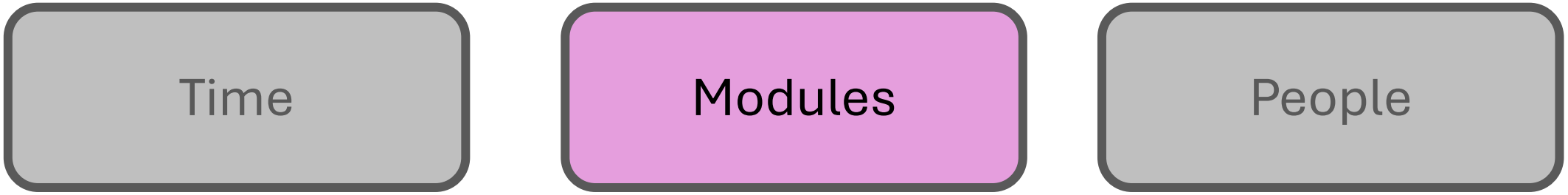
Divide & Conquer

Time

Modules

People

Divide & Conquer



Algo Modules

Modularity rather than E2E models

Dev Modules

Separate module for each functionality

(Data preparation, training …)

Divide & Conquer

Time

Modules

People

Algo Modules

Modularity rather than E2E models

Smaller tasks are simpler

Interpretability

Debuggability

Dev Modules

Separate module for each functionality

(Data preparation, training ...)

Faster Development

Better Control

Reproducability

Divide & Conquer

Time

Modules

People

Algo Modules

Modularity rather than E2E models

Smaller tasks are simpler

Interpretability

Debuggability

Dev Modules

Separate module for each functionality

(Data preparation, training ...)

Faster Development

Better Control

Reproducability

Modulate more than you already have

Divide & Conquer

Time

Modules

People

Divide & Conquer

Time

Modules

People

Split the work between team members

Separate roles in the team

The Principals

Application oriented algo

Invest in Data & Evaluation

Make it a factory

Divide & Conquer

Don’t be naïve

Be in touch

Bring Them Home