ML Development Workflow

Yanda Tao – 2024/10/21

ML workflow

Problem scoping

Deployment
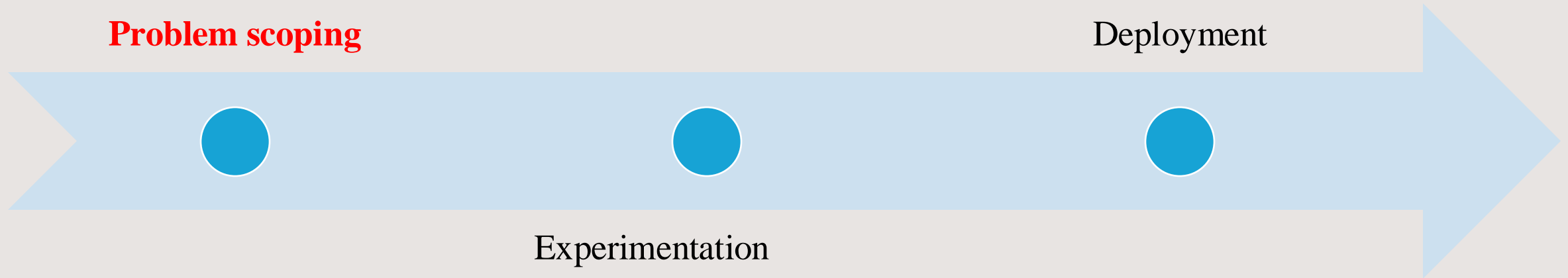
Experimentation

- *Choosing architecture (data, model)*
- *Training*
- *Evaluation*

Problem scoping

To define your problem, you may need to consider the following points:

- **Define business problem**

  - What are you trying to solve?

  - What is the business value?

- **Define ML problem**:

  - Which type of ML problems and which model?

  - What is your data?

  - What is your goal?

Business Problem (1):

To identify *what are you trying to solve*. You may need to consider:

Which industry are you working on?

What pain points are they facing?

Any specific real-world examples of your problem?

Which stakeholders do you plan to service?

Business Problem (2):

To define the *business value*. You may need to consider:

How can the ML solution reduce costs?

Whether your ML solution can create new revenue streams?

Whether your solution can help in mitigating risks?

Whether the solution will provide a competitive edge?

How can your ML solution provide a long-term sustainable value?

ML Problem (1)

The first step to define your *ML problems*:

Define type of ML problems: regression, classification

Define Model: linear regression, logistic regression, neural network, etc.

ML Problem (2)

What is your data?

Define data sources (where, how)

Define data processing requirements

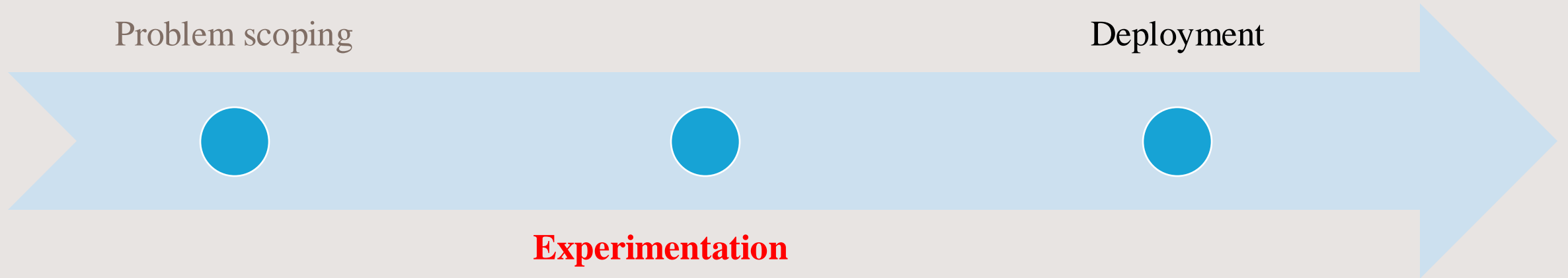Define the features you will use

ML Problem (3)

To define your *goal*. You may need to:

Ensure the goals of the ML project align with the problem you defined

Some direct metrics such as accuracy, precision, recall, AUC, or ROC.
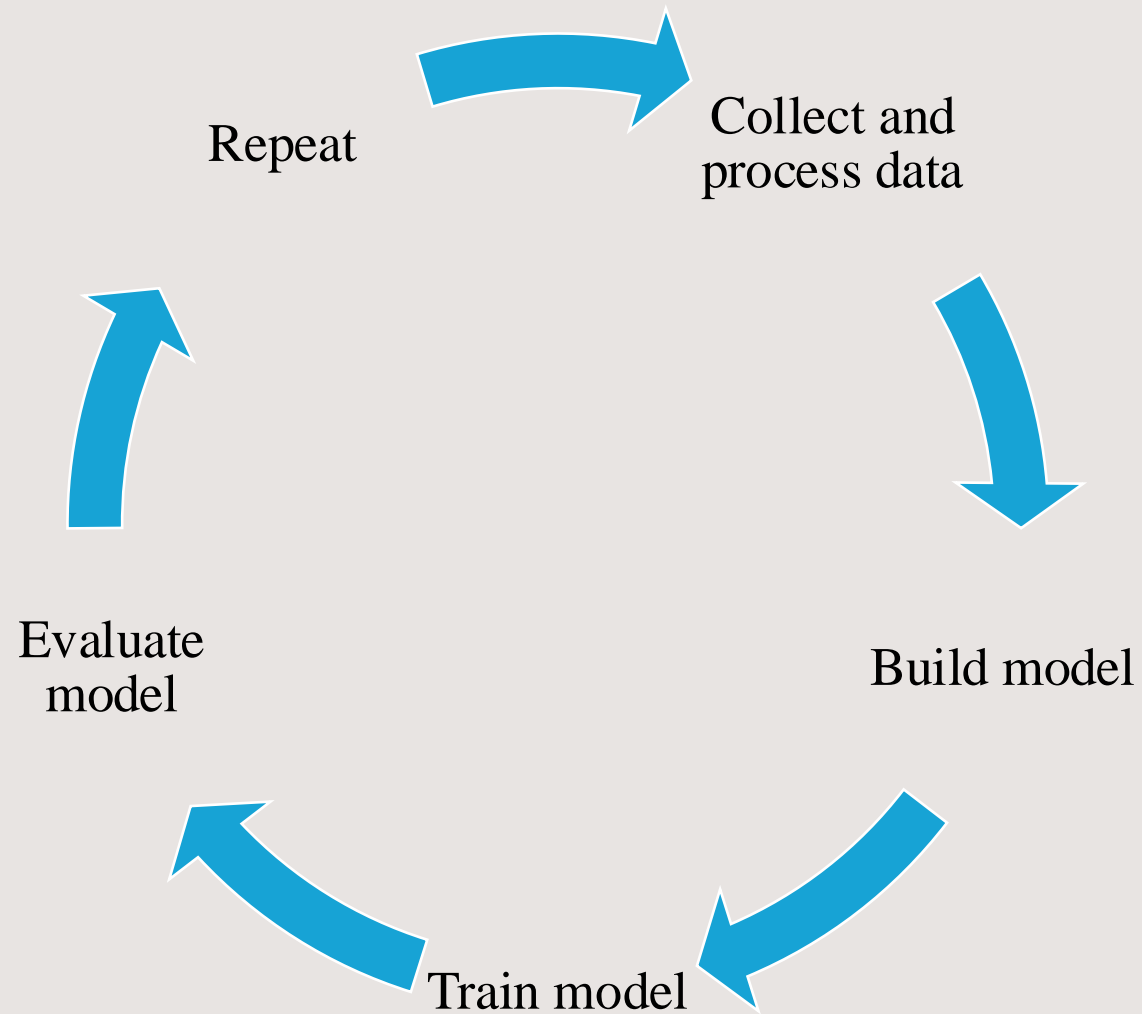
Some indirect metrics such as speed, scalability, and complexity.
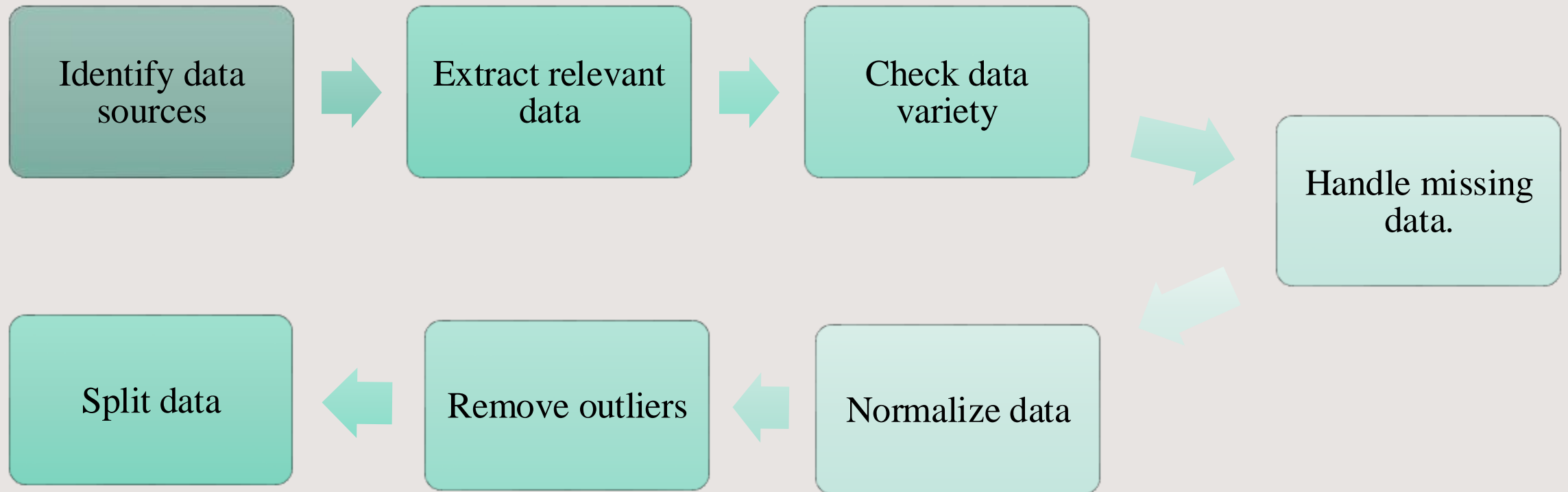
ML workflow

Problem scoping

Deployment

Experimentation

- *Choosing architecture (data, model)*
- *Training*
- *Evaluation*

Experimentation
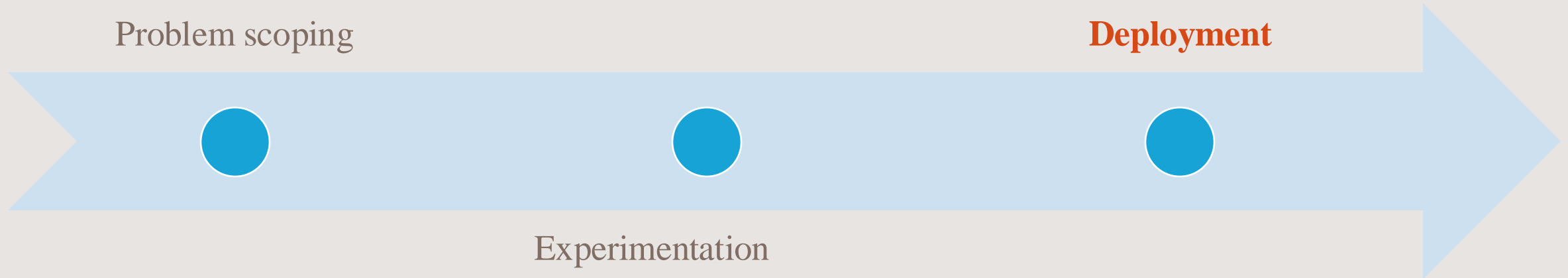
Collect and process data

Build model

Train model

Evaluate model

Repeat

Experimentation

Collect and process data

```
Identify data sources  →  Extract relevant data  →  Check data variety  →  Handle missing data.

Split data  ←  Remove outliers  ←  Normalize data  ←
```

ML workflow
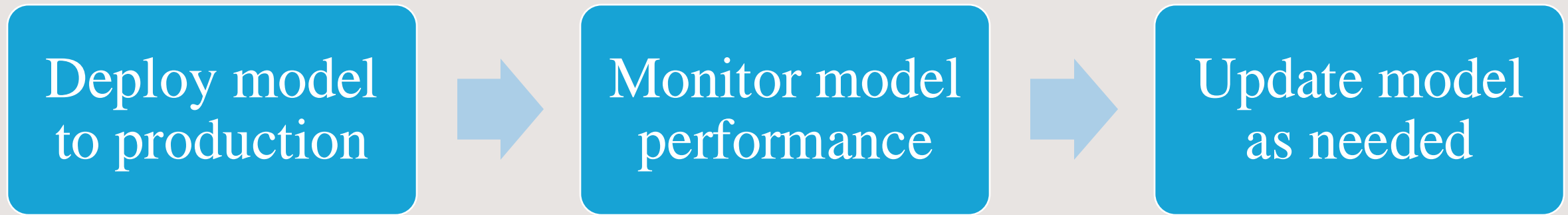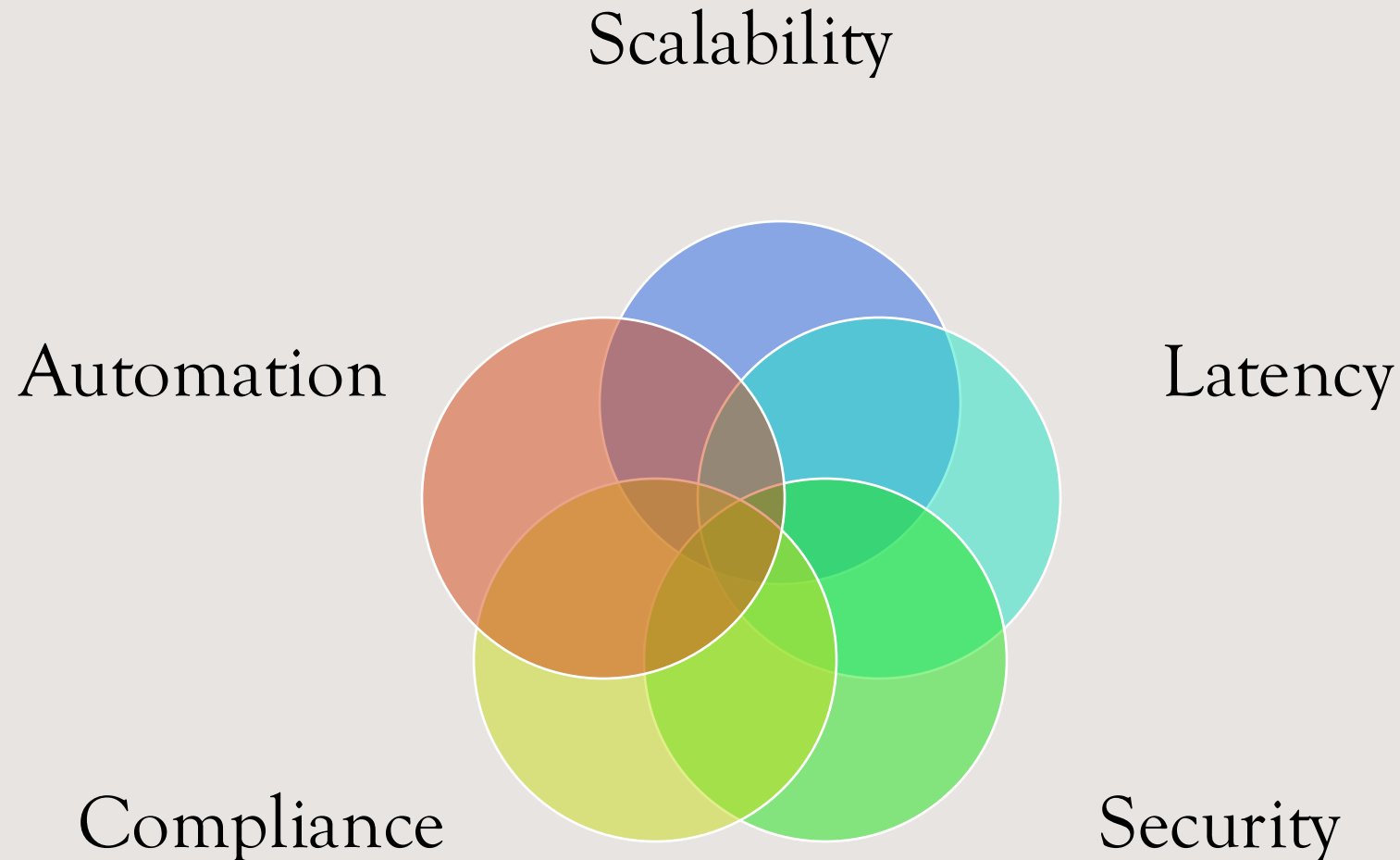
Problem scoping

Deployment

Experimentation

- *Choosing architecture (data, model)*
- *Training*
- *Evaluation*

Deployment (1)

Deploy model to production → Monitor model performance → Update model as needed

Deployment (2)

Scalability

Automation

Latency

Compliance

Security

Term project: Build Supervised ML Model

Your task is to build a supervised ML model to solve a real-world problem.

Follow the ML workflow (except deployment)

You can choose any dataset

You can use any ML library (scikit-learn is recommended)

You can use any ML models covered in the course

Demo

Background: In today's business environment, which heavily relies on credit card payments, credit card fraud has become an increasingly serious issue. Such fraudulent activities not only result in direct financial losses for consumers but also cause significant negative impacts on businesses' reputation and operations, including loss of consumer trust, reputational damage, and increased compliance costs.

Now let's try to define the problem in the given context together by answering the following question:

1. What's your business problem:
   - Refer to Page 4 and Page 5
2. What's your ML problem:
   - Refer to Page 6 and Page 8

Group Project - Proposal

In the proposal, you need to answer the following questions:

- **What's your business problem**: Clearly define the business problem you're trying to solve and its significance.

- **What's your ML problem**: Translate the business problem into an ML problem, justifying your modeling choice.

Format Requirement: 1 page PDF. An additional page is allowed for references or appendix **only**.

Font size 12, Times New Roman, single-spaced. All margins should be 1 inch (2.54 CM).

Group Project - Report

The report should contain the following contents:

- **Project Overview**

- **Model Development**

Format Requirement: Jupyter notebook on EdStem.
- Up to 3,000 words for Project Overview.
- No limit for Model Development (code).

Group Project - Presentation

The presentation is expected to contain:

- Brief overview of the business problem
- Description of data sources
- Presentation of ML techniques selected and rationale
- Results of your work
- Business insight gained from the project

Format Requirement: 10 minutes for each group, with 2 minutes Q&A.
- PDF copy of presentation slides
- All members are encouraged to participate

Group Project - Grading

Grading: This project will be evaluated based on the following three aspects:

- Business:
  - Clear definition of the business problem.
  - The novelty and potential value of the project.
  - Feasibility of implementing and deploying the solution
  - Supporting evidence and references
- Technical:
  - The novelty and appropriateness of the chosen techniques.
  - Consideration and comparison of alternative techniques or similar methods.
  - Effective utilization of data
  - Successful implementation of the proposed solution
- Submission Quality:
  - The quality and professionalism of the slides and team presentation.
  - The clarity of the report's structure and writing.
  - Adherence to formatting requirements (e.g., page and length limits).

TEAM MEETING