Practical AI

Emanuele Fabbiani

Evaluation

1. Business Understanding

- Workshop
- Evaluation metrics
- Literature review

2. Data Gathering

- Source discovery
- Data preparation
- Quality assessment

3. Data Exploration

- Explorative analysis
- Data cleaning

4. Modelling

- Feature engineering
- Model design
- Model implementation

5. Evaluation

- Performance evaluation
- Time evaluation

6. Deployment

- Execution scheduling
- Performance tuning
- Integration

Why Evaluation?

The loss function is closely tied to the training algorithm. It is typically a mathematically well-behaved function, chosen to be **general** and facilitate the convergence of the training process. However, it is **rarely aligned** with the specific business needs of a given problem.

Therefore, model evaluation relies on **different metrics**, selected based on their relevance to the specific task.

In the following section, we will review some of the most common **metrics** for classification and regression, though many others exist.

When in doubt, a quick **Google search** can be helpful.

How do we evaluate the model?

We can use a linear regression model to **explore** the feature of a dataset. That's common in econometrics.

However, we are usually interested in making predictions.

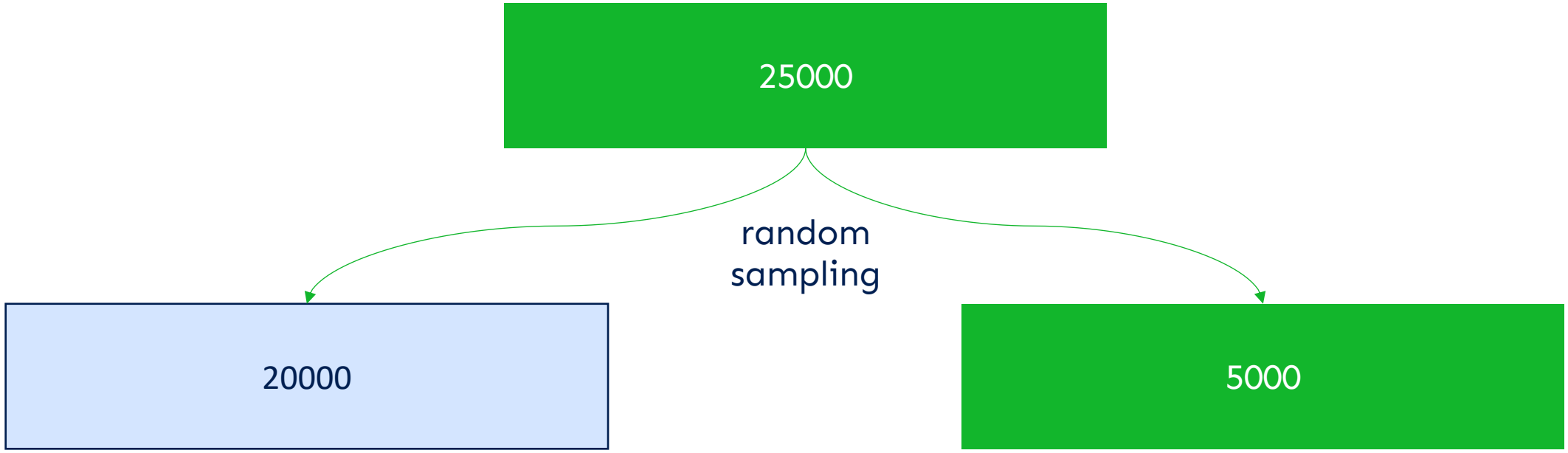
To make good predictions, the model must be able to **generalize**.

If we use all the data to train the model, we cannot evaluate how well it works on other data.

So, we need to **hold out** some data.

We define a **train** and a **test** set.

How do we evaluate the model?



Train set – used for parameter optimization

Test set – used for evaluation **only**

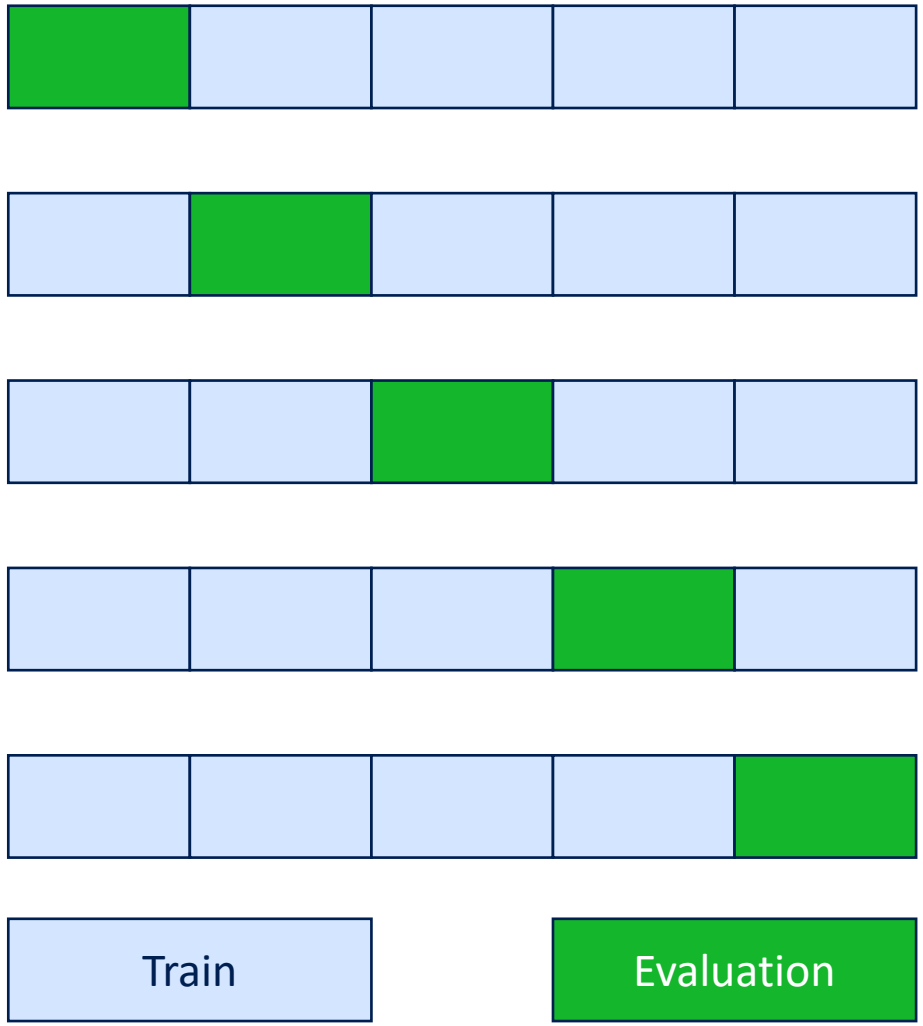
Cross Validation

K-fold cross-validation works by splitting the dataset into **k equally sized folds** (k=5 in the example).

The model is trained on **k-1 folds** and tested on the remaining fold, repeating the process k times, with each fold serving as the test set once.

The final performance is calculated by averaging the metrics across all k iterations.

This approach ensures that the model is evaluated **on different subsets of data**, making it particularly useful when data is limited.



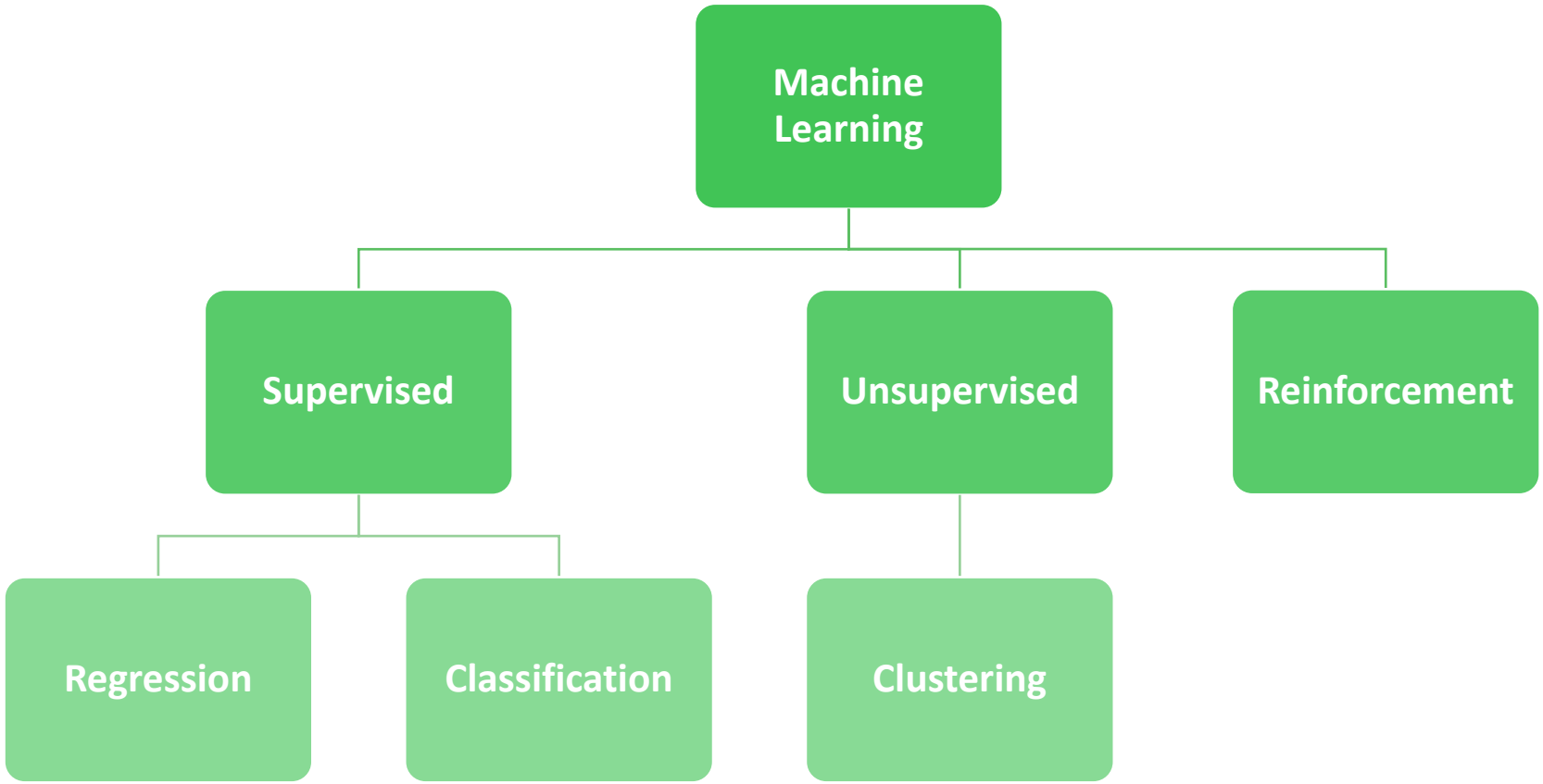
Test Set or Cross-Validation?

Apart from evaluating the final model, cross-validation is often used to select the model's hyperparameters. In this case, it is applied to the training set to preserve the test set for the final evaluation.

Creating a test set is **simpler** and faster but provides a **less reliable** estimate of the model's performance on unseen samples.

Cross-validation is **slower** and becomes even more time-consuming as k increases, but it offers a **more accurate** estimation of the model's performance.

Machine Learning tasks



Regression Metrics

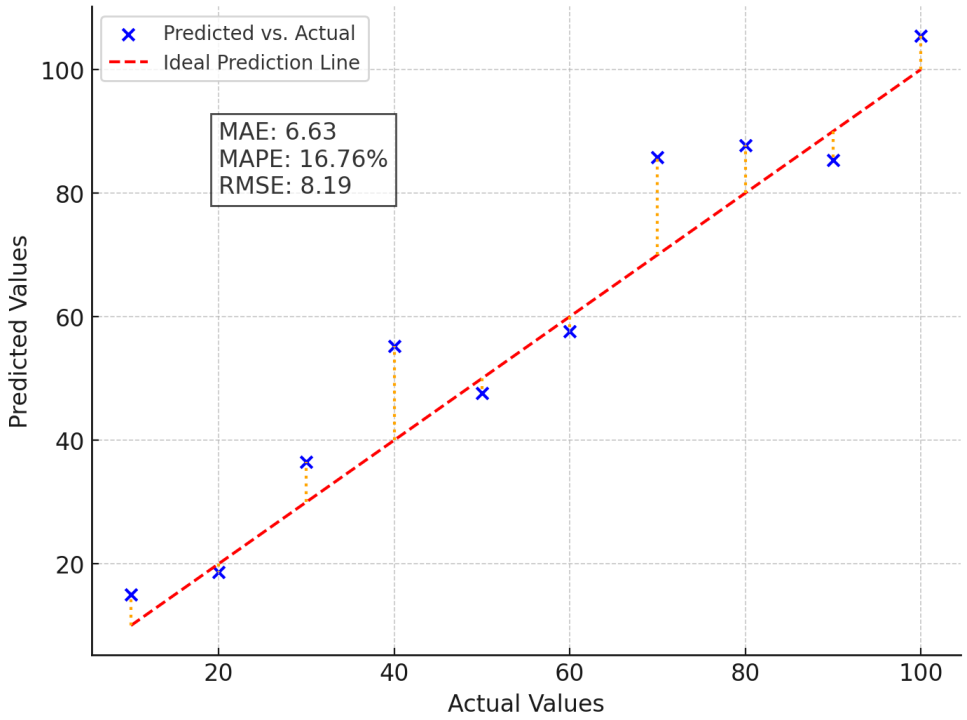
Let $i = 1 \ldots n$ be samples in a dataset, y_i be the target variable and $\hat{y}_i$ the prediction for sample i.

Then,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}$$



Regression Metrics

Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are **absolute error metrics**, meaning they share the same unit of measure as the target. RMSE penalizes larger errors more heavily since each error is squared, whereas MAE treats all errors equally.

MAPE is a **relative error metric** and may be easier for non-technical stakeholders to interpret. However, MAPE **cannot be used** when the target is negative, zero, or close to zero, as the target appears in the denominator.

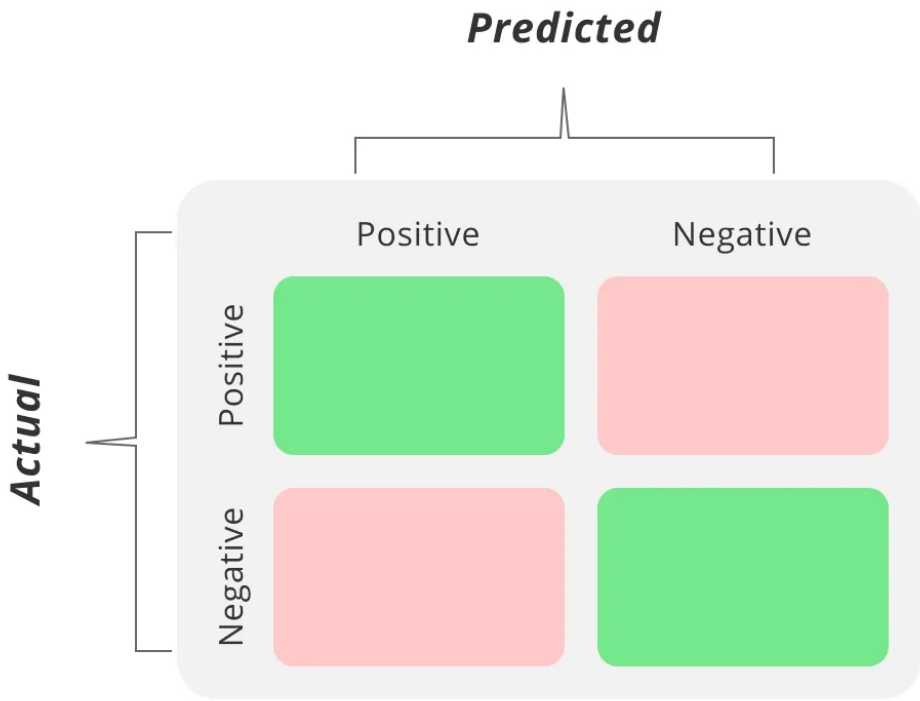
Additionally, MAPE implies that errors are less significant when the target is low, whereas they are more substantial when the target is high. This assumption may not hold in all use cases. Over time, alternative metrics such as Mean Absolute Scaled Error (MASE) and Symmetric Mean Absolute Percentage Error (SMAPE), commonly used in forecasting, have been proposed to address these issues.

Classification Metrics

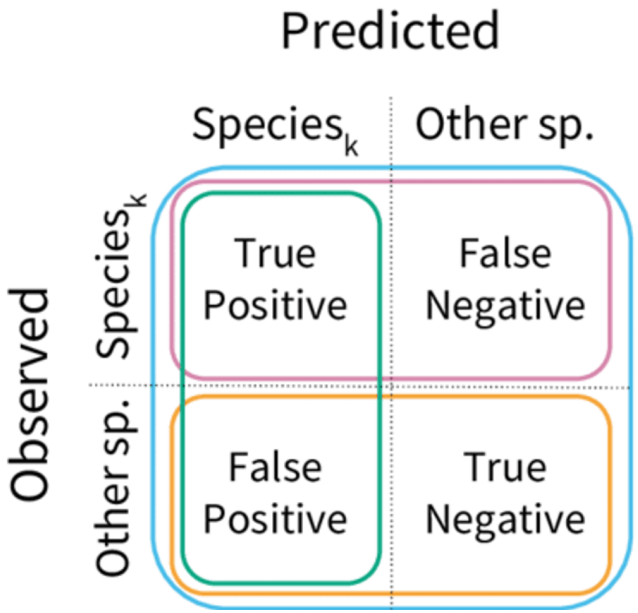
Classification has various metrics, all derived from a single visualization: the **confusion matrix**.

The confusion matrix displays **predicted labels** in the columns and **actual labels** in the rows. Each element represents the number (or fraction) of samples corresponding to a specific (row, column) combination.

For example, the element (1,1) represents the number (or fraction) of samples that are positive and correctly predicted as such.



Classification Metrics

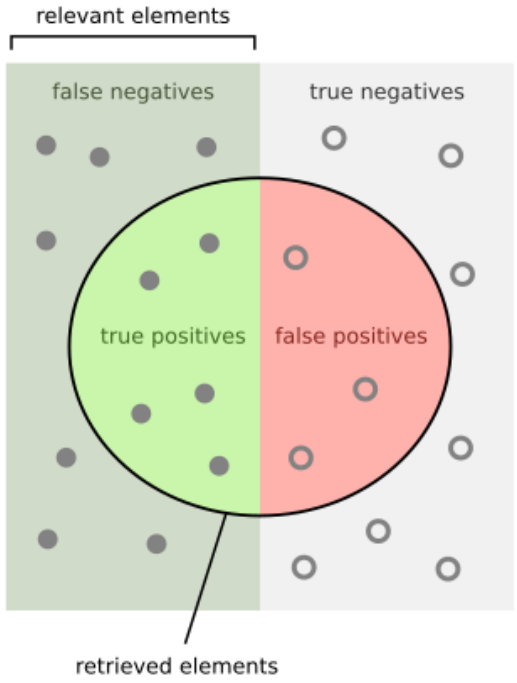


Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$

Specificity $= \frac{TN}{TN + FP}$

Precision $= \frac{TP}{TP + FP}$

Recall $= \frac{TP}{TP + FN}$



How many retrieved items are relevant?

Precision =

How many relevant items are retrieved?

Recall =

Classification Metrics

F1 score and Matthews Correlation Coefficient (MCC) **consider all the elements of the confusion matrix**, addressing the limitations of accuracy on imbalanced datasets.

The F1 Score is the harmonic mean of precision and recall:

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

The MCC is a sort of "correlation" between the predicted and actual classification:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Why Isn’t Accuracy Enough?

Not all mistakes are created equal.

Consider the problem of determining whether a patient needs a specific exam to check for a certain disease. If the model incorrectly predicts that the patient does not need the exam when they do (a false negative), the consequence could be fatal. Conversely, if the model predicts that the patient needs the exam when they do not (a false positive), it leads to unnecessary costs for the healthcare system and wasted time for the patient. Since a false negative is **far more dangerous** than a false positive, prioritizing a **very high recall** is essential, even at the **expense of precision**.

Moreover, many datasets are **unbalanced**, meaning some classes are significantly rarer than others. In such cases, even poorly performing models can achieve high accuracy, making accuracy an unreliable metric for evaluation.

Business Metrics

An ML model is just a tool to achieve a business goal. The real success metrics aren't about the model itself but the **business problem it solves**.

Each problem requires its metrics, but there are two key principles to follow:

- Define business metrics **before** building the model (see the problem statement section).
- Ensure the metrics are as closely tied as possible to the **economic impact** of your ML system.

Discussion

What’s the best metric for a credit rating model?

~~Accuracy~~

~~F1 Score~~

~~Recall~~

~~Binary Cross-Entropy~~


amc

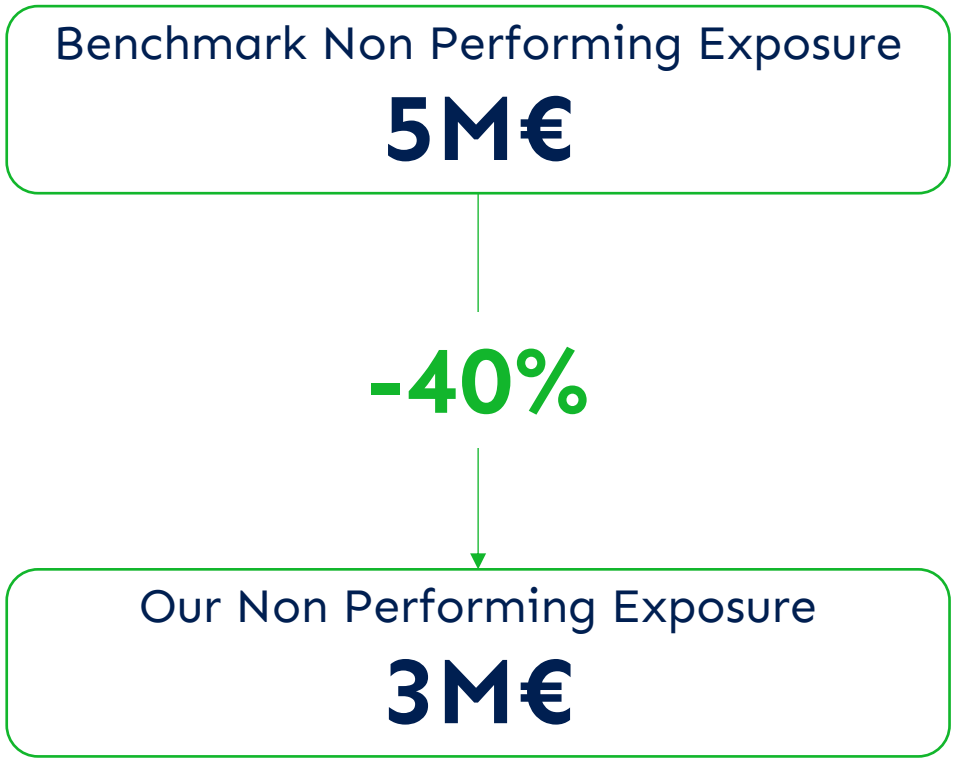
Money

Business stakeholders care about one thing.

Money.

How many loans can we issue, and how much money might we lose from defaults?

Lending 100M€



Deployment

1. Business Understanding

- Workshop
- Evaluation metrics
- Literature review

2. Data Gathering

- Source discovery
- Data preparation
- Quality assessment

3. Data Exploration

- Explorative analysis
- Data cleaning

4. Modelling

- Feature engineering
- Model design
- Model implementation

5. Evaluation

- Performance evaluation
- Time evaluation

6. Deployment

- Execution scheduling
- Performance tuning
- Integration

Omission Alert

This is a brief introduction to a broad and captivating topic.

While it is not required for this class, you are welcome to explore the basic ideas of MLOps: this website is a good starting point.

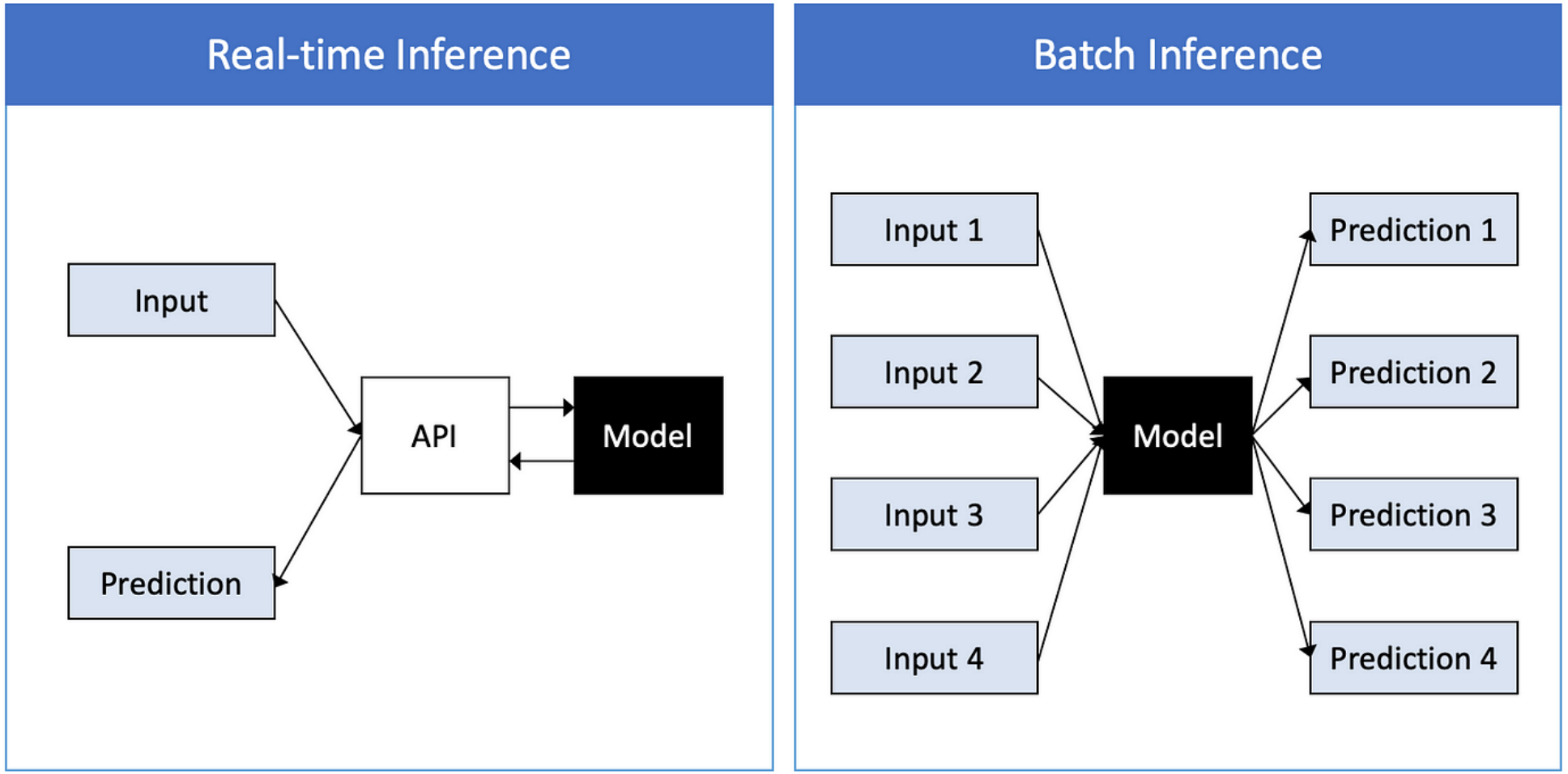
Deployment?

Once the model is ready, the next step is to integrate it into a **software application** and make it accessible to users or other systems within your company.

At this stage, several key factors need to be considered:

- **Retraining strategy**: How often should the model be retrained to maintain accuracy? What triggers a retraining—time intervals, data drift, or performance degradation?
- **Prediction mode**: Should predictions be generated in real-time or in batch?
- **Training time**: How long can the training process take without disrupting operations?
- **Prediction time**: What are the acceptable response times for generating predictions?
- **Specific requirements**: Are there any regulatory, security, or infrastructure constraints that need to be addressed?
- **Data flow**: how will the model receive its input data? Will it pull data from a database, receive API requests, or process streaming data in real time?

Batch vs Real-Time Predictions



Discussion

What's the best deployment mode for Netflix's recommender system?

Discussion

What's the best deployment mode for YouTube's policy violation detection model?

Change Management

Change Management?

Change management is a **structured approach** to guiding individuals, teams, and organizations from their current state to a desired future state while minimizing resistance and maximizing adoption.

When implementing a new ML/AI solution in an organization, key considerations include:

- **Resistance** from end users who may feel threatened.
- **Lack of trust** in the technology.
- **Risks** associated with potential **errors** in ML/AI solutions.
- Organizational **changes** resulting from the new technological implementation.

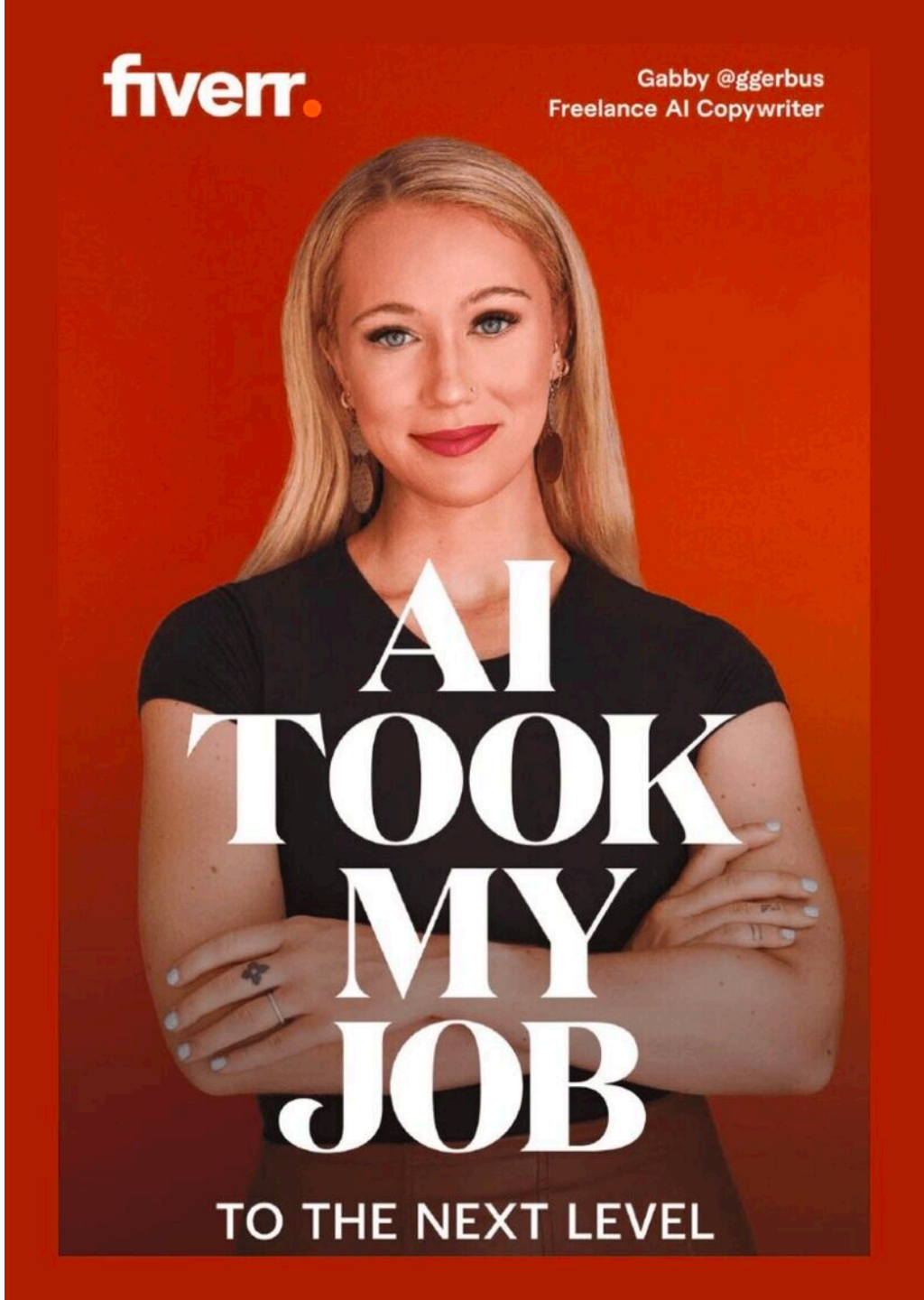
A change management plan is instrumental in mitigating issues after the introduction of the new technology.

Resistance from End Users

Resistance often stems from the fear of losing control, power, or being forced to change established habits.

To gain support, a good strategy is to demonstrate how the technology can **assist** users in their work rather than replace them. Typically, ML and AI handle the less creative, more **mechanical** aspects of the job.

Additionally, securing management support is essential for driving process changes that facilitate adoption.

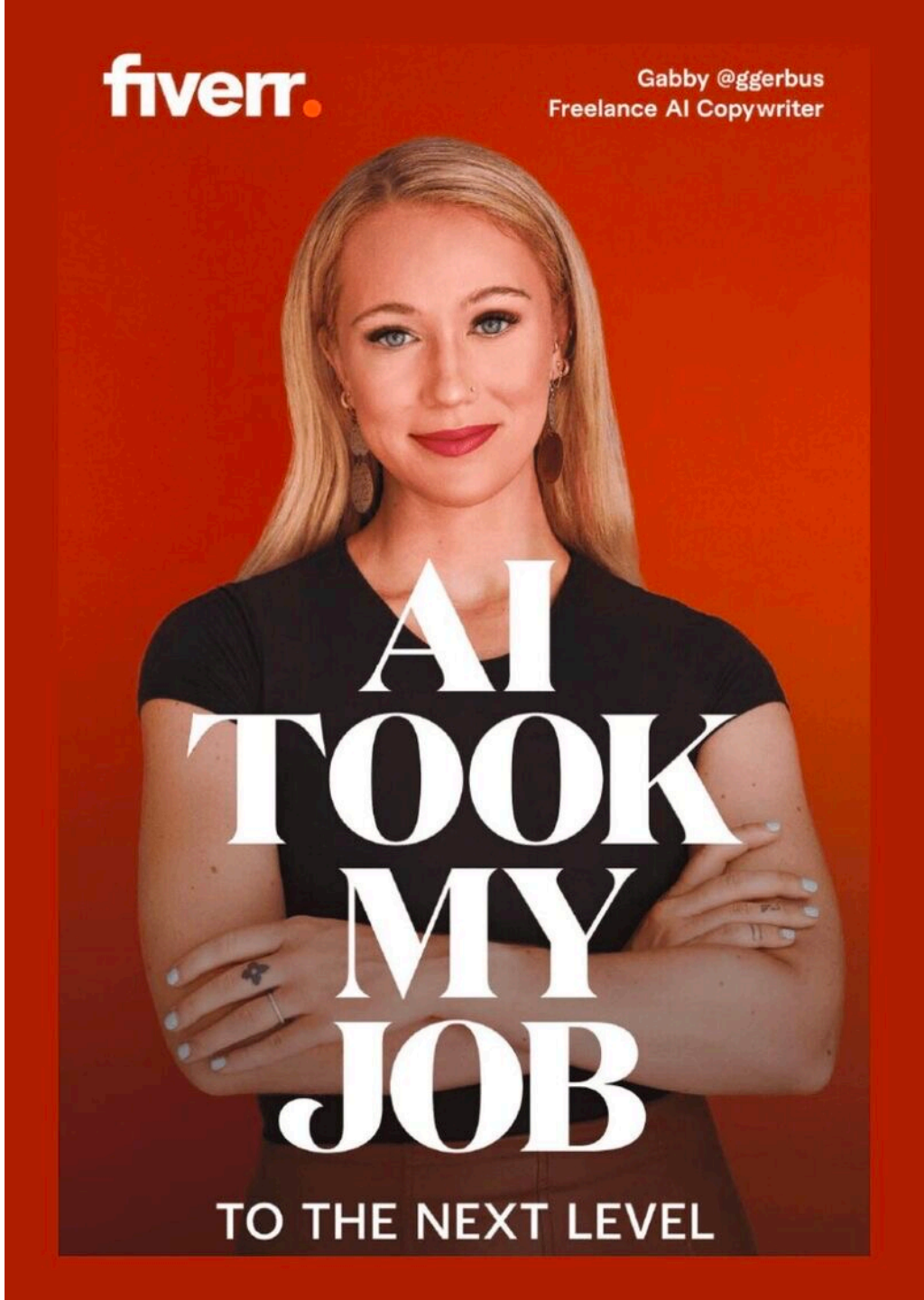


Lack of Trust

ML and AI systems are often perceived as "**black boxes**", leading to their outputs being double-checked or ignored.

To prevent this pattern, which would negate the benefits of an ML solution, strategies include:

- Establishing a period of **parallel operation** between human and ML outputs to demonstrate the model's performance.
- Providing **explanations** for the model's output, if available.
- Conducting **training sessions** to demonstrate the solution.
- Being transparent about the solution's **limitations**.

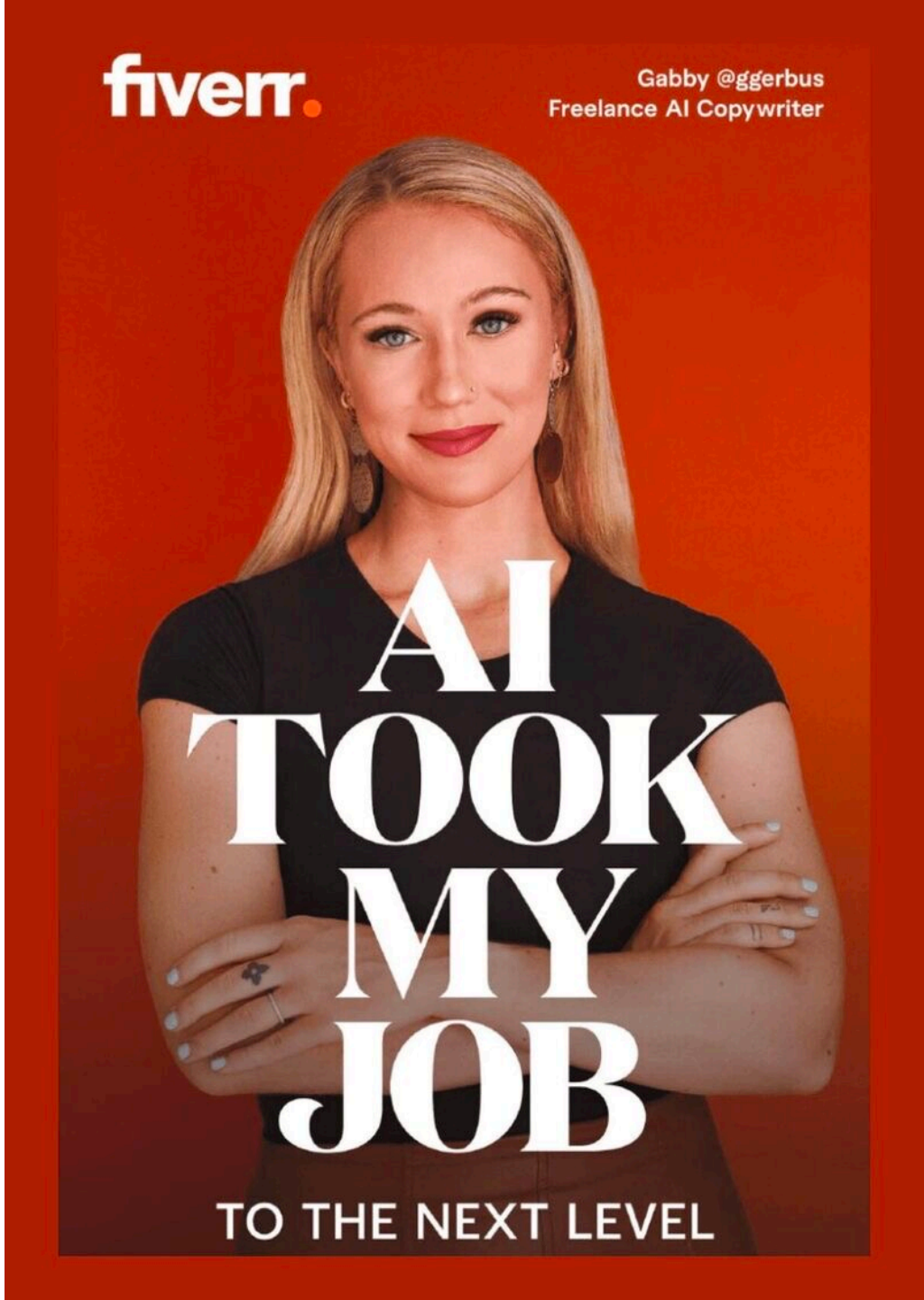


Risks

Useful ML and AI systems are employed to make **decisions**, but incorrect decisions can be costly. Every ML or AI system is prone to errors.

Before adopting such a system, a thorough analysis of output **uncertainty** must be conducted, followed by appropriate **risk management measures**, including financial considerations.

When possible, initiating adoption with human **supervision** helps mitigate these risks.



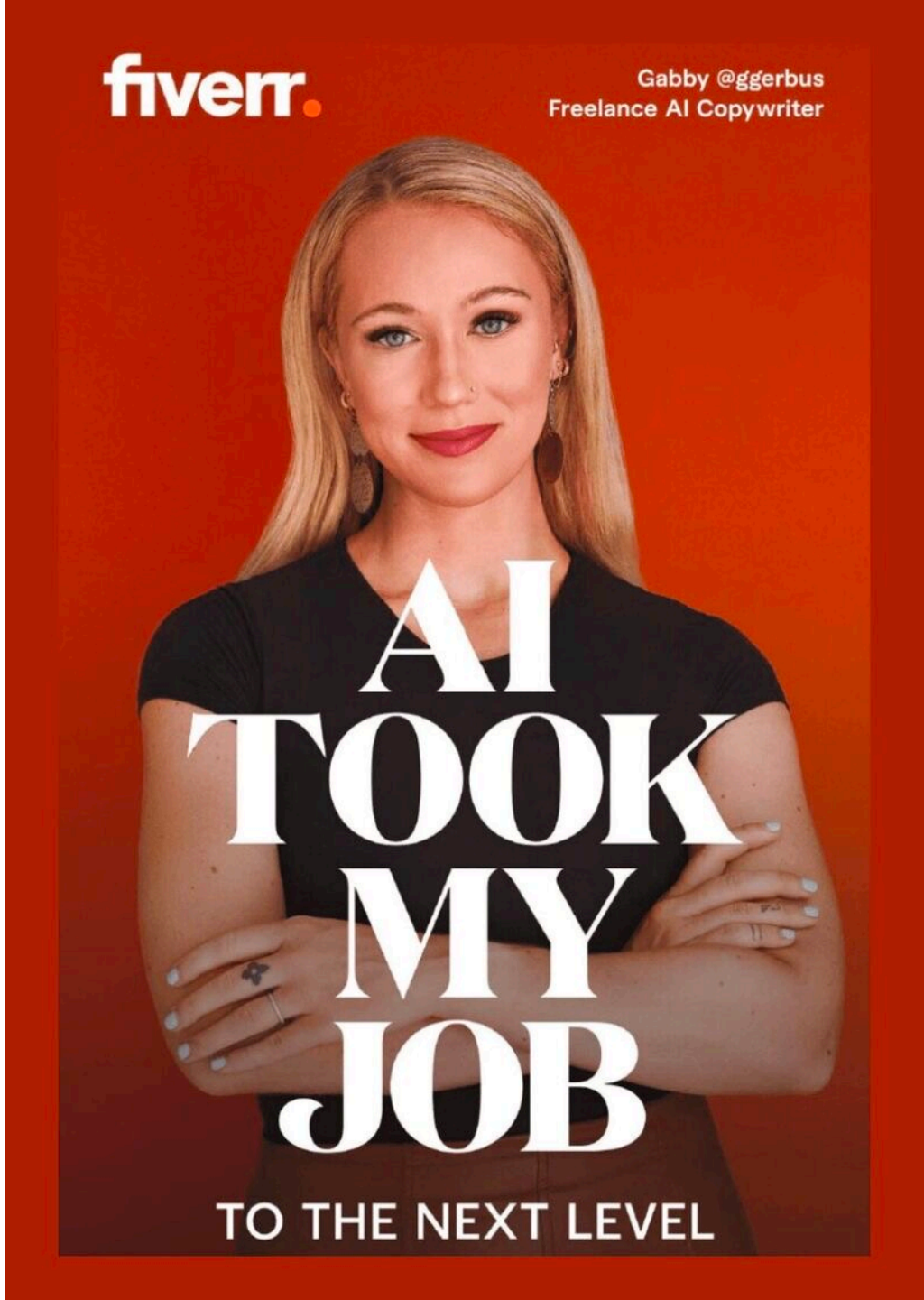
Organisational Changes

Due to the systematic introduction of AI and ML, roles may **evolve**, and departments may be created or eliminated.

Many organizations now have a central AI team, along with engineers embedded in various departments to bridge technology and business (**hub and spokes** model).

Furthermore, with AI being implemented at scale, some roles may **shift** from **operational** to more **supervisory** functions.

Upper management should acquire sufficient knowledge and vision to effectively support these changes.



What Can Go Wrong?

Zillow

Founded in 2006, Zillow was a leading real estate listings platform in the United States for years.

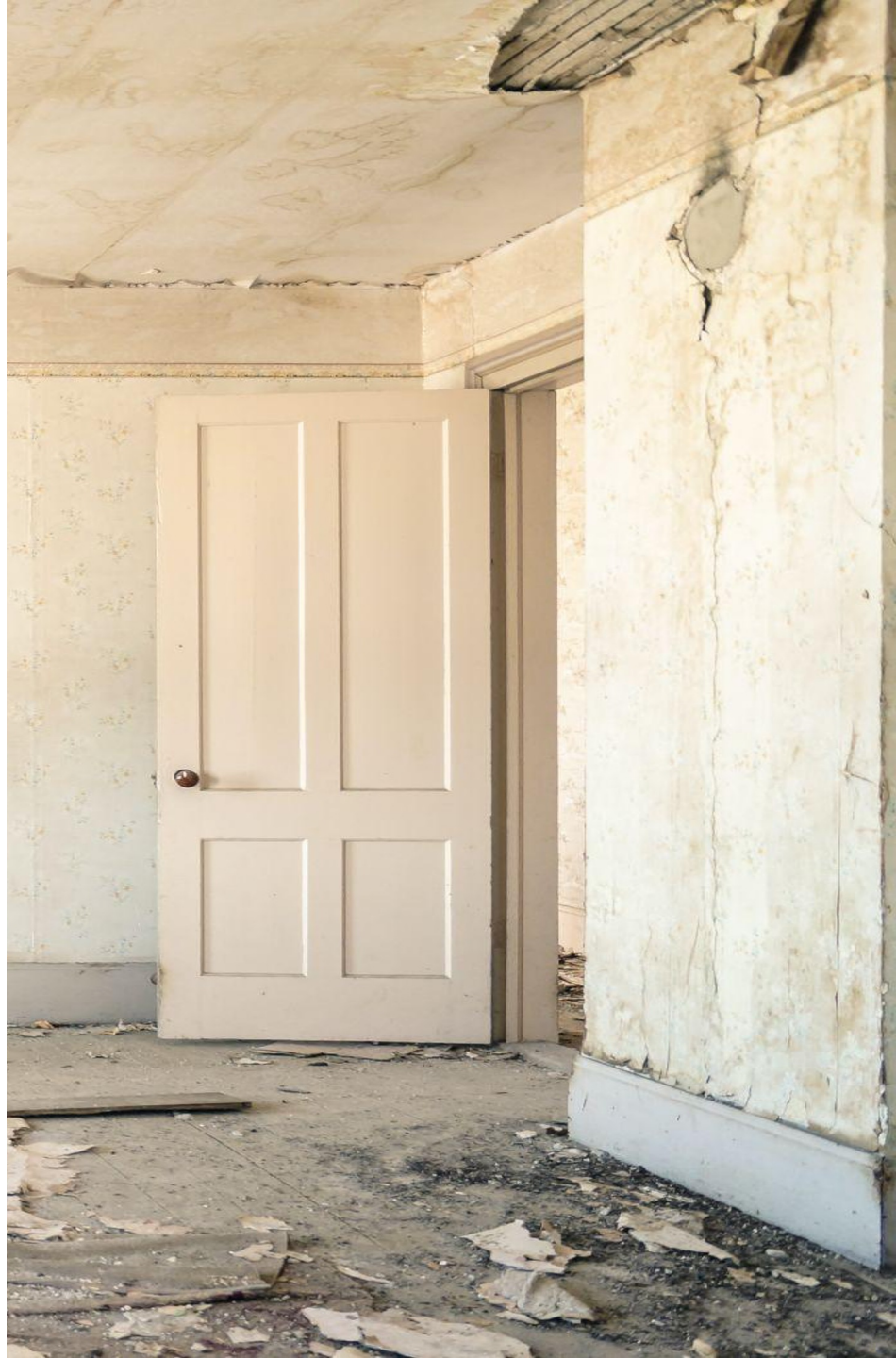
In 2018, the company entered the **house flipping business**, purchasing properties in areas with anticipated price increases and reselling them for a profit.

And Its Downfall

In November 2021, Zillow announced it was shutting down its house flipping business.

After accumulating over **$500 million** in losses, the company was forced to sell off most of its acquired properties and **lay off 25% of its workforce**.

On the day of its quarterly earnings announcement, Zillow's stock **plunged by more than 30%**.



What Happened?

Zillow's CEO put it this way:

"We couldn’t accurately predict changes in property values in either direction. Our mistakes were **far bigger** than we ever thought possible.”

Machine learning and AI are powerful tools, but they are **no magic wands**. To deliver reliable and valuable results, they need a business structure built to support it.

https://medium.com/@donlelef/zillow-a-machine-learning-failure-e6a1fb59316e


The End