Module 3: Datasets, Algorithms, and Models (75)

- What is a dataset: structure and data quality
- Concepts: features, labels, training/test sets
- Intro to popular algorithms: linear regression, decision trees, k-means
- 📌 Demo: Google Colab – Basic example using linear regression
- 📌 Activity: Guided questions + small group discussion

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Why Data Matters in ML

- Algorithms are only as good as the data they use
- Data quality drives performance
- Foundation for training accurate models

What Is a Dataset?

- Structured collection of data
- Rows = instances or samples
- Columns = features or attributes
- May include a target variable (label)

Dataset Structure and Quality

- Features: measurable properties
- Labels: known outcomes (for supervised learning)
- Importance of clean, complete, and relevant data

Training and Test Sets

- **Training set**: used to train the model
- **Test set**: used to evaluate performance
- Often split: 70/30 or 80/20

Common Dataset Pitfalls

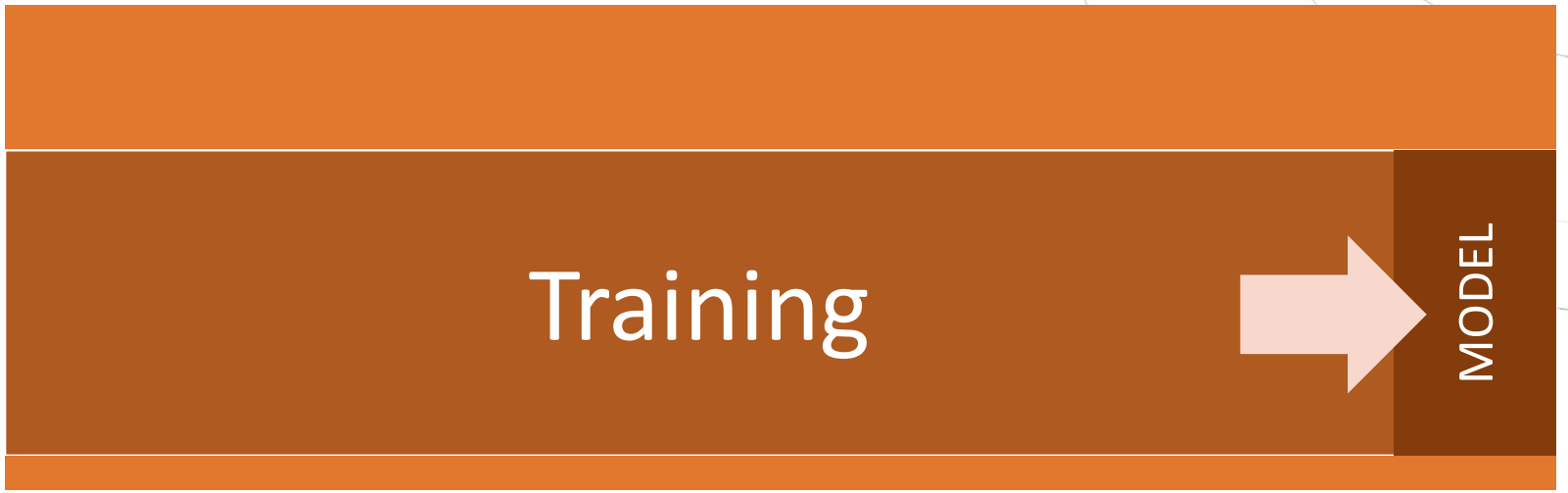
- Missing values
- Noisy or inconsistent data
- Data leakage
- Imbalanced classes

Training, Validation, and Test

- Split process into: training, validation, and test

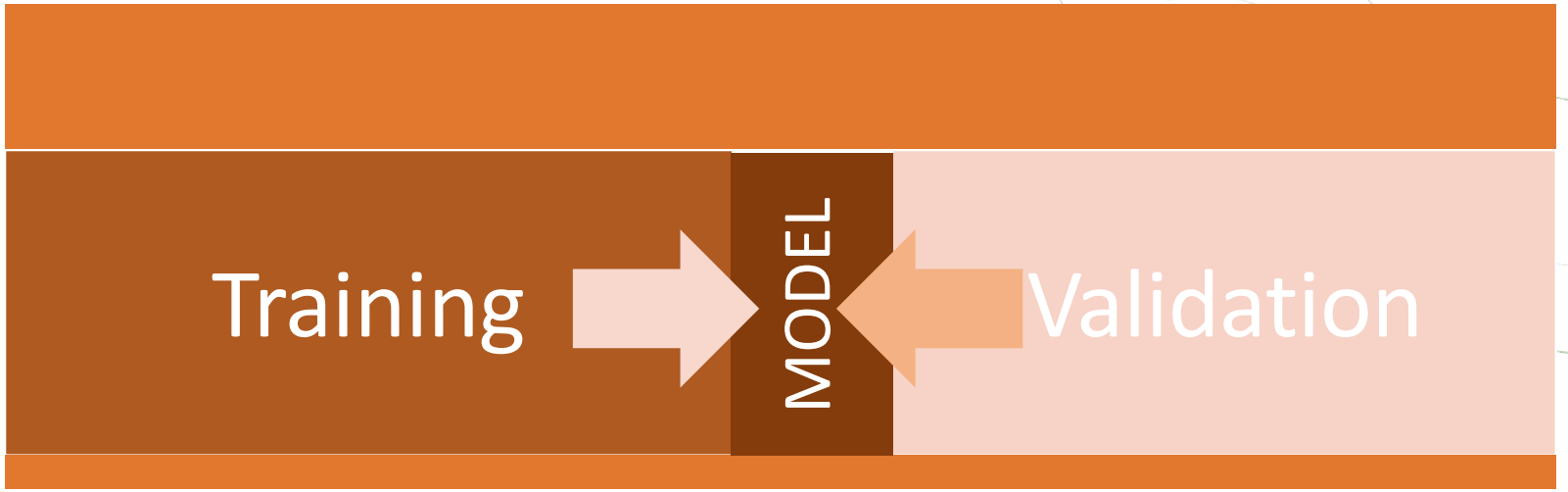
Training

- Learn the model over the training set



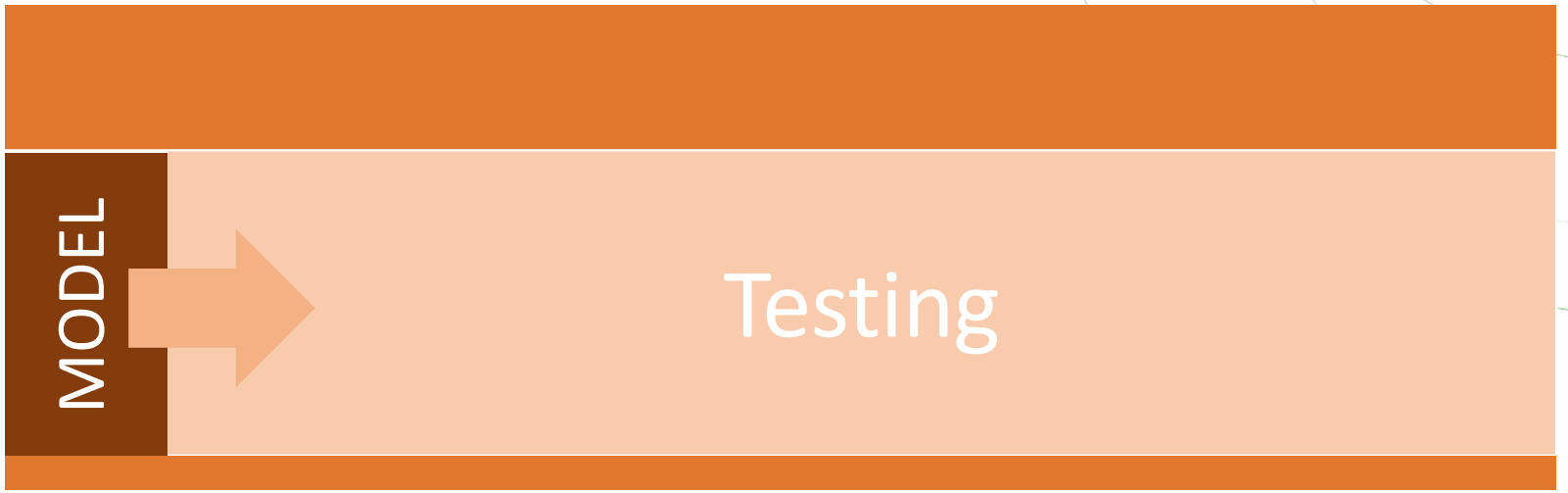
Validation

- Optimize the predictive capability of the model using the validation set



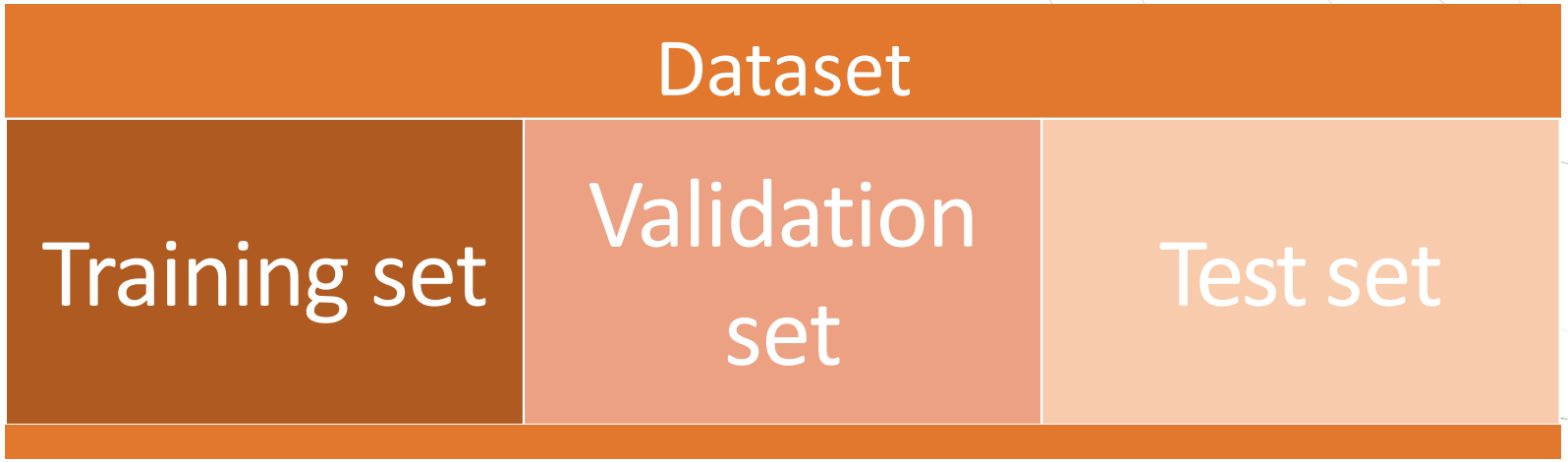
Testing

- See how well the model works on the test set (unseen before)
- The error gives an unbiased estimate of the predictive power of a model



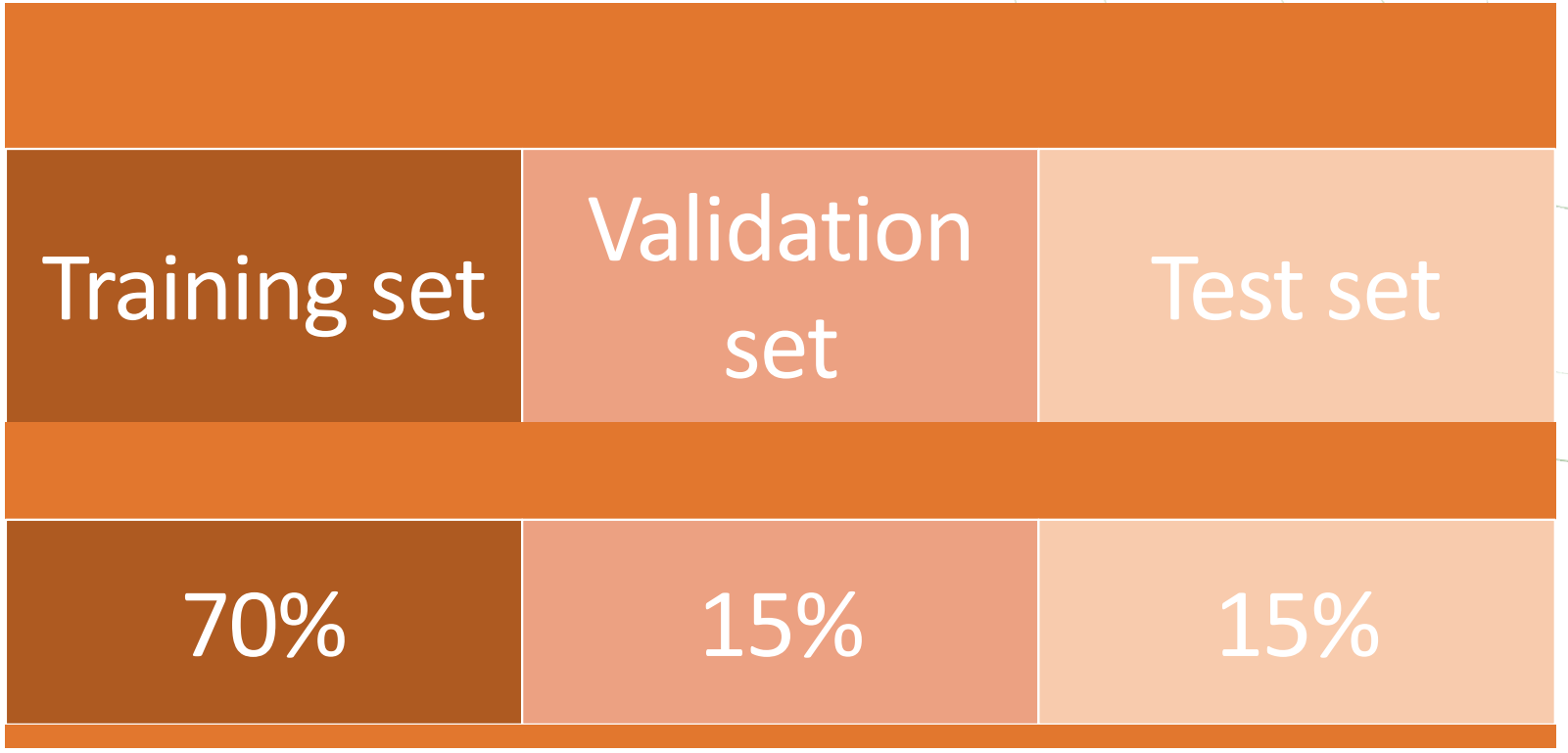
Training, Validation, and Test

- Split data into three sets: training, validation, and test



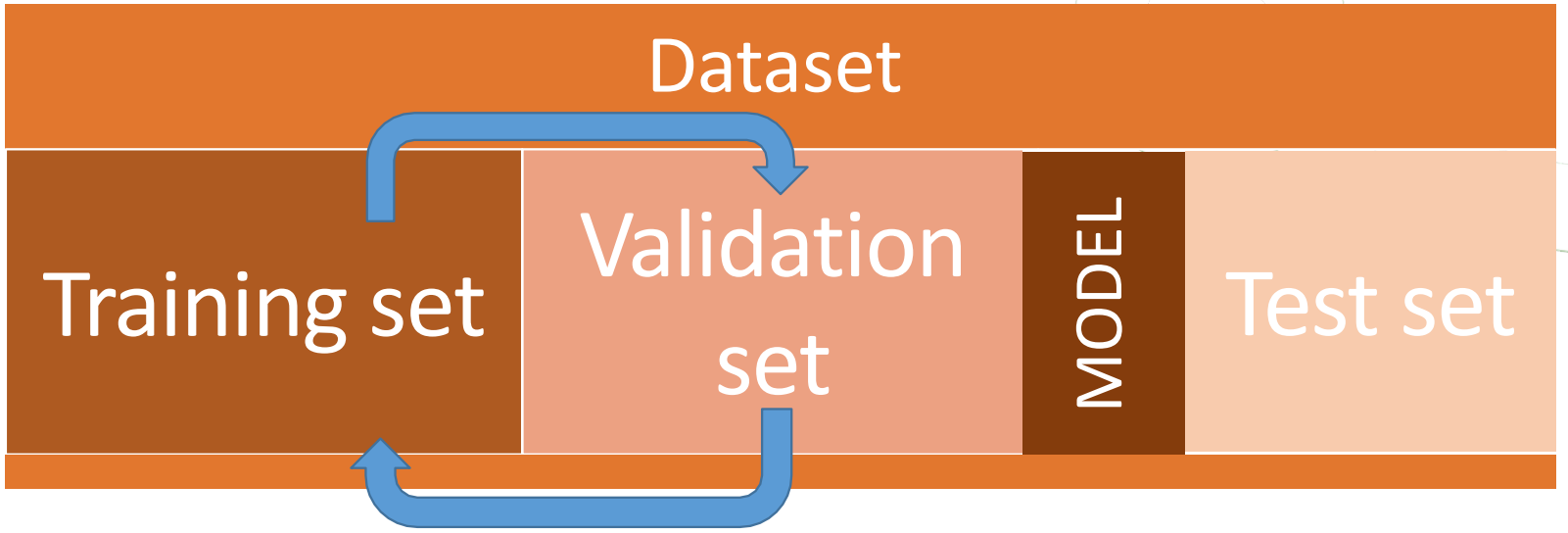
Training, Validation, and Test

- Split data into three sets: training, validation, and test



Cross-Validation

- Repeat the iteration on training + validation multiple times
- 10-fold cross-validation: pick 10 times random subsets as training and validation, and average the quality of the results



Intro to Algorithms: Linear Regression

- Predicts **continuous** values
- Models relationships between features and target
- Simple, interpretable model

Intro to Algorithms: Decision Trees

- Tree-like structure for decisions
- Easy to understand and visualize
- Handles classification and regression

Intro to Algorithms: K-Means

- Unsupervised clustering algorithm
- Groups data into k clusters
- Useful for segmentation tasks

Supervised Learning Models

Model	Description
Linear Regression	Predicts a continuous value using a linear relationship between input and output.
Logistic Regression	Used for binary classification (yes/no, 0/1).
Decision Trees	Tree-like models used for both classification and regression tasks.
Random Forest	Ensemble of decision trees for more robust predictions.
Support Vector Machines (SVM)	Finds the best boundary between classes. Effective in high-dimensional spaces.
K-Nearest Neighbors (K-NN)	Classifies a sample based on the majority class of its 'k' nearest neighbors.
Naive Bayes	Probabilistic model based on Bayes' theorem, good for text classification.
Gradient Boosting Machines (GBM)	Powerful ensemble model that builds trees sequentially (e.g., XGBoost, LightGBM).
Neural Networks	Highly flexible models that mimic the human brain, used in deep learning.

(Used when the data has labeled outputs)

Unsupervised Learning Models

Model	Description
K-Means Clustering	Groups data into ‘k’ clusters based on feature similarity.
Hierarchical Clustering	Builds a tree of clusters (dendrogram) for hierarchical grouping.
DBSCAN	Density-based clustering for discovering clusters of varying shapes.
Principal Component Analysis (PCA)	Dimensionality reduction technique that keeps variance.
t-SNE / UMAP	Non-linear dimensionality reduction for visualization.
Autoencoders	Neural network-based model for data compression and reconstruction.

(Used when the data has no labels)

Reinforcement Learning Models

Model	Description
K-Means Clustering	Groups data into ‘k’ clusters based on feature similarity.
Hierarchical Clustering	Builds a tree of clusters (dendrogram) for hierarchical grouping.
DBSCAN	Density-based clustering for discovering clusters of varying shapes.
Principal Component Analysis (PCA)	Dimensionality reduction technique that keeps variance.
t-SNE / UMAP	Non-linear dimensionality reduction for visualization.
Autoencoders	Neural network-based model for data compression and reconstruction.

(Learning via interaction and feedback in an environment)

Other Important Models / Techniques

Model	Description
Ensemble Methods	Combines multiple models (e.g., Bagging, Boosting, Stacking).
Time Series Models	e.g., ARIMA, LSTM (for forecasting).
Transformer Models	State-of-the-art models in NLP (e.g., BERT, GPT).
GANs (Generative Adversarial Networks)	Generates new data similar to training data (used in image generation).

Choosing the Right Algorithm

- Depends on task: classification, regression, clustering
- Consider data size, type, quality
- Trade-off: accuracy vs interpretability

Summary and Q&A

- Datasets = fuel for ML
- Features, labels, train/test split
- Intro to core algorithms: **regression**, **trees**, **clustering**

Bias

Over the Town
Marc Chagall (1918)