AI Research Labs

The big picture · 3 maps · 22 steps

The big *picture*

Why you are learning what you are learning: how every lesson in the course helps build a machine learning model that works.

1 Building a model that predicts: the supervised learning workflow

2 Finding groups without labels: the clustering workflow

◆ Why any of it works



Free course with quizzes, progress tracking and Scoopy, an AI tutor:
airesearchlabs.si/learn/ml

Machine Learning Foundations · Edition November 2026 · © 2026 AI Research Labs. Free for personal study.

How it all fits *together*

Running example: Telco. A telecoms company with 10,000 customers wants to forecast each customer's **next bill** (a number), flag customers likely to **leave** in the next 30 days (a category), and **group** customers by how they use the service (no labels at all). One dataset, three jobs: follow it through the map.

MAP 1 · SUPERVISED LEARNING: BUILDING A MODEL THAT PREDICTS: THE SUPERVISED LEARNING WORKFLOW

1. **Frame the problem**
   Lessons 1.1, 1.2, 1.3 · feeds Get, split and prepare the data
2. **Get, split and prepare the data**
   Lessons 2.6, 2.8 · feeds Evaluate honestly, Control complexity, Turn predictions into decisions, Measure similarity
3. **A number or a category?**
   Lessons 4.1 · feeds Choose the model, Choose the model

REGRESSION LANE Predict a number

- R1 **Choose the model**
  Lessons 3.1, 3.2
- R2 **Fit it**
  Lessons 3.3, 3.4 · feeds Control complexity
- R3 **Read and check it**
  Lessons 3.5, 3.6, 3.8, 3.9 · feeds Control complexity, Choose the model
- R4 **Score it**
  Lessons 3.7 · feeds Evaluate honestly

CLASSIFICATION LANE Predict a category

- C1 **Choose the model**
  Lessons 4.2, 4.3, 6.4, 6.5 · feeds Turn predictions into decisions
- C2 **Fit it**
  Lessons 4.4, 4.5, 6.6 · feeds Control complexity
- C3 **Score it**
  Lessons 4.6, 4.7, 4.10, 4.11 · feeds Turn predictions into decisions, Evaluate honestly

4. **Evaluate honestly**
   Lessons 2.3, 2.4, 2.7 · feeds Control complexity
5. **Control complexity**
   Lessons 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 6.7 · feeds Turn predictions into decisions
6. **Turn predictions into decisions**
   Lessons 4.8, 4.9
7. **Deploy, monitor, repeat**
   Revisits 1.3, 2.8 · feeds Frame the problem

↺ Back to step 1 (frame the problem): modelling is a loop, not a line.

MAP 2 · UNSUPERVISED LEARNING: FINDING GROUPS WITHOUT LABELS: THE CLUSTERING WORKFLOW

1. **Measure similarity**
   Lessons 6.2
2. **Assign and update**
   Lessons 6.1
3. **Choose K and judge the result**
   Lessons 6.3 · feeds Control complexity, Measure similarity

UNDERNEATH BOTH MAPS: WHY ANY OF IT WORKS

- **No free lunch**
  Lessons 1.4 · feeds Choose the model, Choose the model, Evaluate honestly
- **Empirical vs population risk**
  Lessons 1.5 · feeds Evaluate honestly
- **Loss functions**
  Lessons 2.1 · feeds Fit it, Fit it
- **Empirical risk minimisation**
  Lessons 2.2 · feeds Fit it, Fit it, Control complexity
- **Bias–variance trade-off**
  Lessons 2.5 · feeds Evaluate honestly, Control complexity

In the order you *study*

The map follows the machine learning workflow, the order you build a model in, so its lessons come from all over the course. This page lists the lessons in course order, with the workflow step each one serves. Study in this order; use the map to see where each lesson fits.

SECTION	LESSONS	WORKFLOW STEP
01 **Machine Learning Foundations**	1.1 What is Machine Learning? 1.2 Types of Machine Learning 1.3 The ML Modelling Process	1 Frame the problem
	1.4 Fundamental Limits of Machine Learning	◆ No free lunch
	1.5 Empirical Risk and Population Risk	◆ Empirical vs population risk
02 **Statistical Learning and Model Evaluation**	2.1 Loss Functions and Risk	◆ Loss functions
	2.2 Empirical Risk Minimisation (ERM)	◆ Empirical risk minimisation
	2.3 Underfitting and Overfitting 2.4 Generalisation and the Generalisation Gap	4 Evaluate honestly
	2.5 The Bias–Variance Trade-off	◆ Bias–variance trade-off
	2.6 Training, Validation and Test Sets	2 Get, split and prepare the data
	2.7 Cross-validation	4 Evaluate honestly
	2.8 Data Leakage and Good ML Practices	2 Get, split and prepare the data
03 **Linear Regression**	3.1 Regression Fundamentals 3.2 Simple and Multiple Linear Regression	R1 Choose the model
	3.3 Ordinary Least Squares (OLS) 3.4 Calculating Regression Coefficients by Hand	R2 Fit it
	3.5 Interpreting Coefficients 3.6 Residuals and Regression Assumptions	R3 Read and check it
	3.7 R^2, Adjusted R^2, RMSE and Other Metrics	R4 Score it
	3.8 Standard Errors, Confidence Intervals and p-values 3.9 Multicollinearity and VIF	R3 Read and check it

SECTION	LESSONS	WORKFLOW STEP
04 **Logistic Regression and Classification**	4.1 Regression versus Classification	? A number or a category?
	4.2 Odds, Log-odds and the Sigmoid Function 4.3 The Logistic Regression Model	C1 Choose the model
	4.4 Maximum Likelihood Estimation 4.5 Cross-entropy / Log-loss	C2 Fit it
	4.6 The Confusion Matrix 4.7 Accuracy, Precision, Recall and Specificity	C3 Score it
	4.8 Sensitivity and Specificity Trade-offs 4.9 Threshold Selection and Classification Costs	6 Turn predictions into decisions
	4.10 ROC Curve and AUC 4.11 Class Imbalance	C3 Score it
05 **Regularisation and Model Selection**	5.1 Why Regularisation is Needed 5.2 Model Complexity and Feature Selection 5.3 Best Subset and Stepwise Selection 5.4 Ridge Regression (L2) 5.5 LASSO Regression (L1) 5.6 Understanding λ (Lambda) 5.7 Ridge versus LASSO 5.8 Mathematical Objectives and Worked Calculations 5.9 Selecting λ Using Cross-validation	5 Control complexity
06 **Clustering and Support Vector Machines**	6.1 K-means Clustering	2 Assign and update
	6.2 Distance Calculations and Cluster Assignment	1 Measure similarity
	6.3 Updating Centroids and Convergence	3 Choose K and judge the result
	6.4 SVM Fundamentals 6.5 Margins and Support Vectors	C1 Choose the model
	6.6 Hard-margin and Soft-margin SVM	C2 Fit it
	6.7 The C Parameter and Kernels	5 Control complexity

1 Frame the problem

What exactly are we predicting, and what decision will it drive?

KEY FORMULA

$$\hat{y} = \hat{f}(x)$$

Learn a rule $\hat{f}$ from examples (x_i, y_i) instead of writing it by hand.

FACT OR FIGURE

The type of the target decides the rest of the map: a number takes the regression lane, a category the classification lane, and no target at all the clustering workflow.

SKIP THIS STEP AND…

You optimise the wrong thing, such as predicting churn next week when the retention team needs a month's notice to act.

TELCO EXAMPLE

Telco's targets: next month's bill in £, and "leaves within 30 days" as yes or no. The decision: who the retention team phones.

Learn it

Taught in Section 1 (Machine Learning Foundations) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **1.1 What is Machine Learning?**
  Machine learning is learning a prediction rule from data instead of writing it by hand. What that means, how it differs from traditional programming, and why the decision matters more than the prediction.
- **1.2 Types of Machine Learning**
  Regression, classification, clustering and forecasting; supervised vs unsupervised learning; parametric vs non-parametric models. How to tell which kind of problem you are looking at.
- **1.3 The ML Modelling Process**
  The practical workflow from problem definition to deployment, why it is a loop and not a pipeline, and the difference between an algorithm and its complexity.

How it connects

- → **Get, split and prepare the data** (Step 2): The target decides what data to collect and how to split it
- ← **Deploy, monitor, repeat** (Step 7): It's a loop: what you learn changes how you frame the next model

Next: Get, split and prepare the data Back to the map

Online: airesearchlabs.si/learn/ml/map/frame

2 Get, split and prepare the data

How do I keep part of the data honest for the final check?

KEY FORMULA

$$z = \frac{x - \bar{x}_{\text{train}}}{s_{\text{train}}}$$

Scaling statistics come from the training rows only, then are reused on validation and test.

FACT OR FIGURE

The test set is used **once**. Every extra look turns it into a validation set and makes the reported score optimistic.

SKIP THIS STEP AND…

Standardise or select features before splitting and the held-out rows leak into training: cross-validation can report 0.93 for a model that does much worse on genuinely new customers.

TELCO EXAMPLE

6,000 customers to train, 2,000 to validate, 2,000 locked away for the final test. Split by customer, so one person never sits on both sides, and by date if you predict the future.

Learn it

Taught in Section 2 (Statistical Learning and Model Evaluation) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **2.6 Training, Validation and Test Sets**
  What each data split is for, typical split sizes, and why repeatedly checking the test set and picking the best model quietly ruins your performance estimate.
- **2.8 Data Leakage and Good ML Practices**
  The six common ways information leaks into model evaluation: test-set reuse, preprocessing before splitting, feature selection before CV, related records across splits, time leakage and target leakage. Each with a scenario, why it is wrong and the fix.

How it connects

- → **Evaluate honestly** (Step 4): Cross-validation reuses the training split
- → **Control complexity** (Step 5): λ is chosen on validation data, with scaling fitted inside each training fold
- → **Turn predictions into decisions** (Step 6): The threshold is chosen on validation data, never on the test set
- → **Measure similarity** (Step 1): The same scaling rules apply to clustering
- ← **Frame the problem** (Step 1): The target decides what data to collect and how to split it

Previous: Frame the problem **Next:** A number or a category? Back to the map

Online: airesearchlabs.si/learn/ml/map/data

? A number or a category?

Is the target a quantity, or one of a few classes?

KEY FORMULA

$$P(y = 1 \mid x) \in [0, 1]$$

A class needs a probability; a straight line through 0/1 labels leaves that range.

FACT OR FIGURE

With more than two unordered classes, coding them 1, 2, 3 invents an order and spacing that don't exist.

SKIP THIS STEP AND…

Fit a straight line to 0/1 churn labels and some customers get a "probability" of 1.3 or −0.2.

TELCO EXAMPLE

The bill forecast takes the regression lane; the churn flag takes the classification lane. Same customers, same columns.

Learn it

Taught in Section 4 (Logistic Regression and Classification) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **4.1 Regression versus Classification**
  When the outcome is a category, we model the probability of each class. Why ordinary linear regression is a poor fit for yes/no outcomes, and why logistic "regression" is a classifier.

How it connects

- → **Choose the model** (Regression lane · step R1): Numbers go to the regression lane
- → **Choose the model** (Classification lane · step C1): Categories go to the classification lane

Previous: Get, split and prepare the data **Next:** Choose the model or Choose the model Back to the map

Online: airesearchlabs.si/learn/ml/map/fork

R1 Choose the model

What shape of relationship am I assuming?

KEY FORMULA

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

FACT OR FIGURE

"Linear" means linear in the coefficients: adding x^2 or $x_1 x_2$ as predictors is still linear regression.

SKIP THIS STEP AND…

Assume a straight line when the effect bends and the model underfits: the error stays large however much data you add.

TELCO EXAMPLE

bill $= \beta_0 + \beta_1 \cdot$GB used $+\beta_2 \cdot$minutes $+\beta_3 \cdot$contract months $+\varepsilon$.

Learn it

Taught in Section 3 (Linear Regression) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **3.1 Regression Fundamentals**
  Regression predicts a number from inputs. The model, the noise term, what "fitting" means, and how predictions are made, using house prices as a running example.
- **3.2 Simple and Multiple Linear Regression**
  One predictor or many, the model is a weighted sum of inputs. Why a model with X² or interaction terms is still "linear" regression: it only has to be linear in the coefficients.

How it connects

- ← **A number or a category?** (The fork): Numbers go to the regression lane
- ← **Read and check it** (Regression lane · step R3): A pattern in the residuals means the model form is wrong: transform a variable or add a term
- ← **No free lunch** (Foundation): No regression model is best everywhere, so try several

Previous: A number or a category? **Next:** Fit it Back to the map

Online: airesearchlabs.si/learn/ml/map/r-choose

R2 Fit it

Which coefficients fit the data best?

KEY FORMULA

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

the coefficients that minimise $\mathrm{RSS} = \sum_i (y_i - \hat{y}_i)^2$.

FACT OR FIGURE

With one predictor, $\hat{\beta}_1 = S_{xy}/S_{xx}$: three sums and a division, which you can do by hand in a few minutes.

SKIP THIS STEP AND…

If a predictor barely varies (S_{xx} small), its fitted slope is mostly noise.

TELCO EXAMPLE

OLS might find each extra GB adds about £2.10 to the bill, holding minutes and contract fixed.

Learn it

Taught in Section 3 (Linear Regression) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- 3.3 **Ordinary Least Squares (OLS)**
  OLS picks the line that minimises the residual sum of squares. Residuals and RSS, the design matrix and its dimensions, the normal equations, the closed-form solution, and when it breaks.
- 3.4 **Calculating Regression Coefficients by Hand**
  The slope and intercept formulas for simple regression, a full worked example with a table of deviations, and why a predictor with little variation gives an unreliable slope.

How it connects

- → **Control complexity** (Step 5): Ridge and LASSO are this same objective plus a penalty
- ← **Loss functions** (Foundation): Squared loss gives OLS
- ← **Empirical risk minimisation** (Foundation): OLS is ERM with squared loss

Previous: Choose the model **Next:** Read and check it Back to the map

Online: airesearchlabs.si/learn/ml/map/r-fit

R3 Read and check it

What do the coefficients mean, and can I trust them?

KEY FORMULA

$$t = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}, \qquad \mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

FACT OR FIGURE

A VIF of 10 means that coefficient's variance is 10 times what it would be if its predictor were uncorrelated with the others.

SKIP THIS STEP AND…

Skip the residual plots and collinearity checks and a predictor can show the "wrong" sign or a huge p-value while the predictions still look fine.

TELCO EXAMPLE

Minutes and calls are 0.95 correlated, so their separate coefficients swing wildly between samples even though the bill forecast barely changes.

Learn it

Taught in Section 3 (Linear Regression) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **3.5 Interpreting Coefficients**
  What a regression coefficient means, the "holding other predictors fixed" clause, units and standardisation, why coefficients can change sign, and why association is not causation.
- **3.6 Residuals and Regression Assumptions**
  The assumptions behind OLS inference, and how residual plots reveal curvature, non-constant variance, autocorrelation and outliers. How to describe a residual plot in an exam.
- **3.8 Standard Errors, Confidence Intervals and p-values**
  How uncertain is each coefficient? Standard errors, t-statistics, p-values and confidence intervals, with a full worked example, a real regression output to read, and what a p-value is not.
- **3.9 Multicollinearity and VIF**
  When predictors carry nearly the same information, coefficients become unstable and standard errors explode, even though predictions can stay fine. The variance inflation factor, how to compute it, and what to do.

How it connects

- → **Control complexity** (Step 5): Ridge steadies the coefficients that multicollinearity makes unstable
- → **Choose the model** (Regression lane · step R1): A pattern in the residuals means the model form is wrong: transform a variable or add a term

Previous: Fit it **Next:** Score it Back to the map

Online: airesearchlabs.si/learn/ml/map/r-read

R4 Score it

How far off are the predictions, in units people understand?

KEY FORMULA

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \qquad \text{RMSE} = \sqrt{\tfrac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$$

FACT OR FIGURE

In-sample R^2 never falls when you add a predictor, even pure noise. Out-of-sample R^2 can be negative: worse than predicting the average.

SKIP THIS STEP AND…

Report training R^2 and every extra feature looks like progress.

TELCO EXAMPLE

A validation RMSE of £6 means a typical forecast is about £6 off the real bill.

Learn it

Taught in Section 3 (Linear Regression) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **3.7 R^2, Adjusted R^2, RMSE and Other Metrics**
  MAE, MSE, RMSE, MAPE and average error; R^2 as improvement over predicting the mean; out-of-sample R^2 that can go negative; and adjusted R^2, which penalises useless predictors. With worked calculations.

How it connects

- → **Evaluate honestly** (Step 4): Only the out-of-sample score counts

Previous: Read and check it **Next:** Evaluate honestly Back to the map

Online: airesearchlabs.si/learn/ml/map/r-score

C1 Choose the model

How do I turn inputs into a class, or a probability of one?

KEY FORMULA

$$p = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \beta^\top x$$

An SVM instead finds the boundary with the widest margin, $2/\|w\|$.

FACT OR FIGURE

Logistic regression is linear regression on the log-odds: each unit of x_j multiplies the odds by e^{β_j}.

SKIP THIS STEP AND…

Use a model with no probability output and you can't set a cost-aware threshold later.

TELCO EXAMPLE

$z = -2 + 0.8 \cdot$support calls $- 0.05 \cdot$contract months. A customer with $z = 1$ has $p = \sigma(1) \approx 0.73$ of leaving.

Learn it

Taught in Section 4 (Logistic Regression and Classification) and Section 6 (Clustering and Support Vector Machines) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **4.2 Odds, Log-odds and the Sigmoid Function**
  Probability, odds and log-odds are three views of the same thing. How to convert between them by hand, and how the sigmoid function squashes any score into a probability.
- **4.3 The Logistic Regression Model**
  The model as a linear score passed through a sigmoid, or as linear regression on the log-odds. Interpreting coefficients as odds ratios, computing predictions by hand, and the linear decision boundary.
- **6.4 SVM Fundamentals**
  A support vector machine is a linear classifier that chooses the separating line with the widest safety margin. The idea, the ±1 label convention, the decision rule, and the link to regularisation.
- **6.5 Margins and Support Vectors**
  The geometry of the margin: distance from a point to the boundary, the margin width 2/‖w‖, what support vectors are, and a complete SVM solved by hand.

How it connects

- → **Turn predictions into decisions** (Step 6): The probability is what a threshold acts on
- ← **A number or a category?** (The fork): Categories go to the classification lane
- ← **No free lunch** (Foundation): No classifier is best everywhere, so try several

Previous: A number or a category? **Next:** Fit it Back to the map

Online: airesearchlabs.si/learn/ml/map/c-choose

C2 Fit it

Which coefficients make the observed labels most probable?

KEY FORMULA

$$\mathrm{CE} = -\frac{1}{n}\sum_{i}\left[y_i \ln \hat{p}_i + (1-y_i)\ln(1-\hat{p}_i)\right]$$

The SVM's hinge loss is $\max(0,\ 1 - y\,f(x))$.

FACT OR FIGURE

Predicting $\hat{p} = 0.01$ for a customer who does leave costs $-\ln 0.01 \approx 4.6$ in cross-entropy; predicting 0.5 costs only 0.69.

SKIP THIS STEP AND…

Judge a probability model on accuracy alone and you can't tell a confident wrong answer from a hesitant one.

TELCO EXAMPLE

Maximum likelihood picks the β that make Telco's past leavers look likely to leave and its stayers likely to stay.

Learn it

Taught in Section 4 (Logistic Regression and Classification) and Section 6 (Clustering and Support Vector Machines) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **4.4 Maximum Likelihood Estimation**
  How logistic regression is fitted: choose coefficients that make the observed labels as probable as possible. The likelihood, the log-likelihood, a worked example, and why there is no closed form.
- **4.5 Cross-entropy / Log-loss**
  Cross-entropy is the negative average log-likelihood. Why it punishes confident wrong predictions so hard, how to compute it, and how it differs from accuracy as a measure.
- **6.6 Hard-margin and Soft-margin SVM**
  The hard-margin optimisation problem, why it fails on overlapping data, slack variables and hinge loss, the soft-margin objective, and how hinge loss compares with logistic and 0–1 loss.

How it connects

- → **Control complexity** (Step 5): Hinge loss plus the C parameter is a regularised objective
- ← **Loss functions** (Foundation): Cross-entropy and hinge loss
- ← **Empirical risk minimisation** (Foundation): Logistic regression is ERM with cross-entropy

Previous: Choose the model **Next:** Score it Back to the map

Online: airesearchlabs.si/learn/ml/map/c-fit

C3 Score it

How good is it at catching the class that matters?

KEY FORMULA

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}$$

FACT OR FIGURE

If 2% of customers leave, a model that always says "stays" is 98% accurate and catches nobody.

SKIP THIS STEP AND…

Optimise accuracy on imbalanced data and you get exactly that useless model.

TELCO EXAMPLE

An AUC of 0.85: pick a random leaver and a random stayer, and 85% of the time the model rates the leaver as riskier.

Learn it

Taught in Section 4 (Logistic Regression and Classification) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **4.6 The Confusion Matrix**
  Once you pick a threshold, every prediction is a true positive, false positive, true negative or false negative. How to build the table, read it, and which errors are "misses" and which are "false alarms".
- **4.7 Accuracy, Precision, Recall and Specificity**
  Every metric you can compute from a confusion matrix, what question each one answers, a full worked example, and why accuracy can be dangerously misleading.
- **4.10 ROC Curve and AUC**
  The ROC curve shows every sensitivity–specificity trade-off a model can offer. AUC summarises it as the probability a random positive outranks a random negative. How to compute both by hand.
- **4.11 Class Imbalance**
  When the interesting class is rare, a model can score high accuracy while catching nothing. Baselines, the right metrics, threshold moves, oversampling and stratified sampling, and the oversampling traps.

How it connects

- → **Turn predictions into decisions** (Step 6): Precision and recall move as the threshold moves
- → **Evaluate honestly** (Step 4): Score on held-out data, not training data

Previous: Fit it **Next:** Evaluate honestly Back to the map

Online: airesearchlabs.si/learn/ml/map/c-score

4 Evaluate honestly

Will it work on customers it has never seen?

KEY FORMULA

$$\mathrm{CV}_K = \frac{1}{K}\sum_{k=1}^{K}\mathrm{Err}_k$$

FACT OR FIGURE

In 5-fold cross-validation every row is used for validation exactly once and for training four times.

SKIP THIS STEP AND…

Pick the model with the lowest training error and you pick the one that memorised the noise.

TELCO EXAMPLE

Training error 2%, validation error 15%: a 13-point generalisation gap that says the churn model overfits.

Learn it

Taught in Section 2 (Statistical Learning and Model Evaluation) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **2.3 Underfitting and Overfitting**
  Too simple and a model misses the pattern; too flexible and it memorises noise. How to recognise each from training and validation errors, and what to do about it.
- **2.4 Generalisation and the Generalisation Gap**
  Generalisation is performing well on new data. The generalisation gap is how much worse a model does there than on its training data. What makes the gap grow or shrink, and what to do when test scores disappoint.
- **2.7 Cross-validation**
  K-fold cross-validation step by step, a worked comparison of two models, why each fold is an independent experiment, and when ordinary random folds are the wrong tool.

How it connects

- → **Control complexity** (Step 5): Cross-validation is how λ and C are chosen
- ← **Get, split and prepare the data** (Step 2): Cross-validation reuses the training split
- ← **Score it** (Regression lane · step R4): Only the out-of-sample score counts
- ← **Score it** (Classification lane · step C3): Score on held-out data, not training data
- ← **No free lunch** (Foundation): Why candidates are compared on held-out data
- ← **Empirical vs population risk** (Foundation): Held-out data estimates population risk
- ← **Bias–variance trade-off** (Foundation): Explains the generalisation gap

Previous: Score it or Score it **Next:** Control complexity Back to the map

Online: airesearchlabs.si/learn/ml/map/evaluate

5 Control complexity

How simple can the model be and still capture the pattern?

KEY FORMULA

$$\underbrace{\text{RSS} + \lambda \sum_j \beta_j^2}_{\text{ridge}} \qquad \underbrace{\text{RSS} + \lambda \sum_j |\beta_j|}_{\text{LASSO}}$$

FACT OR FIGURE

$\lambda = 0$ gives ordinary least squares; as $\lambda \to \infty$ every coefficient goes to 0. In an SVM, C works the other way round: a small C means strong regularisation.

SKIP THIS STEP AND…

With hundreds of candidate features and only a few thousand customers, an unpenalised model fits the training set beautifully and next month badly.

TELCO EXAMPLE

LASSO keeps 12 of Telco's 40 features and sets the rest exactly to zero. λ is chosen by 5-fold cross-validation on standardised training data.

Learn it

Taught in Section 5 (Regularisation and Model Selection) and Section 6 (Clustering and Support Vector Machines) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **5.1 Why Regularisation is Needed**
  Regularisation trades a little training fit for a simpler, more robust model. The two competing objectives, the benefits of parsimony, and the kinds of datasets where regularisation matters most.
- **5.2 Model Complexity and Feature Selection**
  Complexity has many dials: number of features, coefficient size, depth, K. Why adding features always helps in-sample, why it eventually hurts out-of-sample, and how feature selection fits in.
- **5.3 Best Subset and Stepwise Selection**
  Best subset selection tries every combination of predictors; forward and backward stepwise approximate it greedily. The algorithms step by step, how many models each fits, and a worked forward-stepwise example.
- **5.4 Ridge Regression (L2)**
  Ridge adds a penalty on the sum of squared coefficients. Its closed-form solution, why it shrinks, why it always has a solution, how it handles correlated predictors, and two worked examples.
- **5.5 LASSO Regression (L1)**
  LASSO penalises the sum of absolute coefficients, which shrinks some exactly to zero: built-in variable selection. Soft thresholding, a worked comparison with ridge, and when to prefer each.
- **5.6 Understanding λ (Lambda)**
  What happens at λ = 0, small, intermediate, large and infinite λ. Why regularisation never beats OLS on training error, and how bias, variance and coefficients move along the way.
- **5.7 Ridge versus LASSO**
  Why LASSO produces exact zeros and ridge does not, explained with the constraint geometry of a disc versus a diamond. A full side-by-side comparison with subset selection, plus an interactive shrinkage explorer.
- **5.8 Mathematical Objectives and Worked Calculations**
  Norms, penalised objectives and how to compute ridge, LASSO and subset "errors" by hand, including choosing between two candidate fits, and penalised logistic regression.

- **5.9 Selecting λ Using Cross-validation**
  λ is a hyperparameter: choose it by cross-validation on standardised training data, then test once. The full procedure, the one-standard-error rule, a case study, and the mistakes to avoid.
- **6.7 The C Parameter and Kernels**
  C trades margin width against training violations, exactly like an inverse regularisation strength. Kernels let an SVM draw curved boundaries by working in a richer feature space. Linear, polynomial and RBF kernels, and tuning C and γ.

How it connects

- → **Turn predictions into decisions** (Step 6): The tuned model's scores are what the threshold is set on
- ← **Get, split and prepare the data** (Step 2): λ is chosen on validation data, with scaling fitted inside each training fold
- ← **Fit it** (Regression lane · step R2): Ridge and LASSO are this same objective plus a penalty
- ← **Read and check it** (Regression lane · step R3): Ridge steadies the coefficients that multicollinearity makes unstable
- ← **Fit it** (Classification lane · step C2): Hinge loss plus the C parameter is a regularised objective
- ← **Evaluate honestly** (Step 4): Cross-validation is how λ and C are chosen
- ← **Choose K and judge the result** (Step 3): K is a complexity dial too: more clusters always fit the data more closely
- ← **Empirical risk minimisation** (Foundation): Regularisation is ERM plus a penalty
- ← **Bias–variance trade-off** (Foundation): Why a penalty can lower test error

Previous: Evaluate honestly **Next:** Turn predictions into decisions Back to the map

Online: airesearchlabs.si/learn/ml/map/complexity

6 Turn predictions into decisions

Where do I draw the line, given what each mistake costs?

KEY FORMULA

$$t^* = \frac{C_{FP}}{C_{FP} + C_{FN}}$$

for calibrated probabilities.

FACT OR FIGURE

If a wasted retention call costs £5 and a lost customer £95, the cost-minimising threshold is $5/100 = 0.05$, not 0.5.

SKIP THIS STEP AND…

Keep the default 0.5 and you let most at-risk customers walk away to save a few phone calls.

TELCO EXAMPLE

Phone every customer with $\hat{p} > 0.05$. Choose that threshold on validation data, then report the test result once.

Learn it

Taught in Section 4 (Logistic Regression and Classification) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- 4.8 **Sensitivity and Specificity Trade-offs**
  Moving the threshold trades misses for false alarms. Why you can't raise sensitivity and specificity together with the same model, and how to reason about which to favour.
- 4.9 **Threshold Selection and Classification Costs**
  When a miss and a false alarm cost different amounts, choose the threshold that minimises expected cost. The theory, a worked table, and how to choose it without leaking the test set.

How it connects

- ← **Get, split and prepare the data** (Step 2): The threshold is chosen on validation data, never on the test set
- ← **Choose the model** (Classification lane · step C1): The probability is what a threshold acts on
- ← **Score it** (Classification lane · step C3): Precision and recall move as the threshold moves
- ← **Control complexity** (Step 5): The tuned model's scores are what the threshold is set on

Previous: Control complexity **Next:** Deploy, monitor, repeat Back to the map

Online: airesearchlabs.si/learn/ml/map/decide

7 Deploy, monitor, repeat

Does it still work next quarter?

KEY FORMULA

$$P_{\text{new}}(x, y) \neq P_{\text{train}}(x, y)$$

Every guarantee assumes new data looks like the training data.

FACT OR FIGURE

Test scores describe the past. When prices, products or customers change, the assumption behind every score in this course breaks quietly.

SKIP THIS STEP AND…

Deploy and forget: a churn model trained before a price rise keeps scoring confidently while its recall falls.

TELCO EXAMPLE

Telco re-checks the confusion matrix every month and retrains when recall drops. That sends it back to step 1: modelling is a loop.

Revisit

Taught in Section 1 (Machine Learning Foundations) and Section 2 (Statistical Learning and Model Evaluation) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **1.3 The ML Modelling Process**
  The practical workflow from problem definition to deployment, why it is a loop and not a pipeline, and the difference between an algorithm and its complexity.
- **2.8 Data Leakage and Good ML Practices**
  The six common ways information leaks into model evaluation: test-set reuse, preprocessing before splitting, feature selection before CV, related records across splits, time leakage and target leakage. Each with a scenario, why it is wrong and the fix.

How it connects

- → **Frame the problem** (Step 1): It's a loop: what you learn changes how you frame the next model

Previous: Turn predictions into decisions **Next:** Frame the problem Back to the map

Online: airesearchlabs.si/learn/ml/map/monitor

1 Measure similarity

What does it mean for two customers to be "close"?

KEY FORMULA

$$d(x, c) = \sqrt{\sum_j (x_j - c_j)^2}$$

FACT OR FIGURE

Unscaled, monthly minutes (0 to 3,000) swamps contract length (1 to 24 months) in the distance.

SKIP THIS STEP AND…

Cluster on raw units and you have really clustered on one variable.

TELCO EXAMPLE

Standardise GB used, minutes and tenure before measuring any distances.

Learn it

Taught in Section 6 (Clustering and Support Vector Machines) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **6.2 Distance Calculations and Cluster Assignment**
  Euclidean and Manhattan distance, how to assign points to centroids by hand, a fully worked 2-D example, and why unscaled features let one variable dominate the clusters.

How it connects

- ← **Get, split and prepare the data** (Step 2): The same scaling rules apply to clustering
- ← **Choose K and judge the result** (Step 3): Poor clusters often mean the features need rescaling or a different distance

Next: Assign and update Back to the map

Online: airesearchlabs.si/learn/ml/map/u-distance

2 Assign and update

How does K-means find the groups?

KEY FORMULA

$$J = \sum_i \|x_i - \mu_{c(i)}\|^2$$

Assign each point to its nearest centroid, move each centroid to its cluster's mean, repeat.

FACT OR FIGURE

Neither step can increase J, so the algorithm always stops.

SKIP THIS STEP AND…

Feed it category codes (red = 1, green = 2) and the "means" are meaningless.

TELCO EXAMPLE

With K = 3, Telco's customers settle into heavy streamers, big talkers and light users.

Learn it

Taught in Section 6 (Clustering and Support Vector Machines) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **6.1 K-means Clustering**
  K-means finds K groups in unlabelled data by alternating two simple steps. What clustering is for, the algorithm in four steps, what it minimises, and how it differs from K-nearest neighbours.

Previous: Measure similarity **Next:** Choose K and judge the result Back to the map

Online: airesearchlabs.si/learn/ml/map/u-assign

3 Choose K and judge the result

How many groups, and are they any use?

KEY FORMULA

Plot J against K and look for the elbow.

FACT OR FIGURE

J always falls as K grows (with $K = n$, $J = 0$), so you can't choose K by minimising J. K-means also converges only to a local optimum: run several random starts.

SKIP THIS STEP AND…

Pick K by lowest J and every customer becomes their own segment.

TELCO EXAMPLE

J drops sharply up to K = 3 and then flattens: three segments, each one the marketing team can name and act on.

Learn it

Taught in Section 6 (Clustering and Support Vector Machines) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **6.3 Updating Centroids and Convergence**
  The update step, why K-means always converges but only to a local optimum, the effect of a bad start, choosing K with the elbow method, and the main ways K-means goes wrong.

How it connects

- → **Control complexity** (Step 5): K is a complexity dial too: more clusters always fit the data more closely
- → **Measure similarity** (Step 1): Poor clusters often mean the features need rescaling or a different distance

Previous: Assign and update Back to the map

Online: airesearchlabs.si/learn/ml/map/u-judge

Underneath both maps · Foundation

No free lunch

THE IDEA

No algorithm is best on every problem.

WHY IT MATTERS

Learning works only because we assume something about the problem, so we compare several models instead of trusting one.

Learn it

Taught in Section 1 (Machine Learning Foundations) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- `1.4` **Fundamental Limits of Machine Learning**
  The No Free Lunch theorem, why no algorithm is best for every problem, and the conditions that make learning from data possible at all.

How it connects

- → **Choose the model** (Regression lane · step R1): No regression model is best everywhere, so try several
- → **Choose the model** (Classification lane · step C1): No classifier is best everywhere, so try several
- → **Evaluate honestly** (Step 4): Why candidates are compared on held-out data

Next: Empirical vs population risk Back to the map

Online: airesearchlabs.si/learn/ml/map/f-nfl

Underneath both maps · Foundation

Empirical vs population risk

THE IDEA

$\hat{R}(f) = \frac{1}{n}\sum_i L(y_i, f(x_i))$ estimates $R(f) = \mathbb{E}[L(Y, f(X))]$

WHY IT MATTERS

We can only compute the error on our sample, but we care about the error on everyone. The gap between them is the central problem.

Learn it

Taught in Section 1 (Machine Learning Foundations) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **1.5 Empirical Risk and Population Risk**
  Population risk is the error we truly care about; empirical risk is the error we can compute. The gap between them is the central problem of machine learning.

How it connects

- → **Evaluate honestly** (Step 4): Held-out data estimates population risk

Previous: No free lunch **Next:** Loss functions Back to the map

Online: airesearchlabs.si/learn/ml/map/f-risk

Underneath both maps · Foundation

Loss functions

THE IDEA

squared, absolute, 0–1, cross-entropy, hinge

WHY IT MATTERS

A loss scores one prediction. Choosing it is choosing what "good" means.

Learn it

Taught in Section 2 (Statistical Learning and Model Evaluation) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **2.1 Loss Functions and Risk**
  A loss function scores one prediction; risk averages it. Squared, absolute, percentage, 0–1, cross-entropy, cost-weighted and hinge loss: what each punishes, where each is used, and how to compute them by hand.

How it connects

- → **Fit it** (Regression lane · step R2): Squared loss gives OLS
- → **Fit it** (Classification lane · step C2): Cross-entropy and hinge loss

Previous: Empirical vs population risk **Next:** Empirical risk minimisation Back to the map

Online: airesearchlabs.si/learn/ml/map/f-loss

Underneath both maps · Foundation

Empirical risk minimisation

THE IDEA

$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$

WHY IT MATTERS

OLS and logistic regression are both ERM. Penalised ERM is regularisation.

Learn it

Taught in Section 2 (Statistical Learning and Model Evaluation) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **2.2 Empirical Risk Minimisation (ERM)**
  ERM chooses the model in a family with the lowest training loss. OLS and logistic regression are both ERM. Why minimising training loss is necessary but not sufficient.

How it connects

- → **Fit it** (Regression lane · step R2): OLS is ERM with squared loss
- → **Fit it** (Classification lane · step C2): Logistic regression is ERM with cross-entropy
- → **Control complexity** (Step 5): Regularisation is ERM plus a penalty

Previous: Loss functions **Next:** Bias–variance trade-off Back to the map

Online: airesearchlabs.si/learn/ml/map/f-erm

Underneath both maps · Foundation

Bias–variance trade-off

THE IDEA

$$\mathbb{E}[(y - \hat{f})^2] = \text{Bias}^2 + \text{Var} + \sigma^2$$

WHY IT MATTERS

Flexible models trade bias for variance. This is why complexity must be controlled at all.

Learn it

Taught in Section 2 (Statistical Learning and Model Evaluation) of the course. The map follows the machine learning workflow, the order you build a model in; the course teaches in the order that is easiest to learn, so the numbers can jump.

- **2.5 The Bias–Variance Trade-off**
  Expected prediction error splits into bias², variance and irreducible noise. Why flexible models trade bias for variance, illustrated with the thought experiment of many training sets.

How it connects

- → **Evaluate honestly** (Step 4): Explains the generalisation gap
- → **Control complexity** (Step 5): Why a penalty can lower test error

Previous: Empirical risk minimisation Back to the map

Online: airesearchlabs.si/learn/ml/map/f-bv