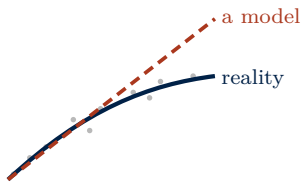
AI and Machine Learning
The Big Picture

What it is, where it works, and when to trust it



Day 1 · Executive Briefing · 40 minutes

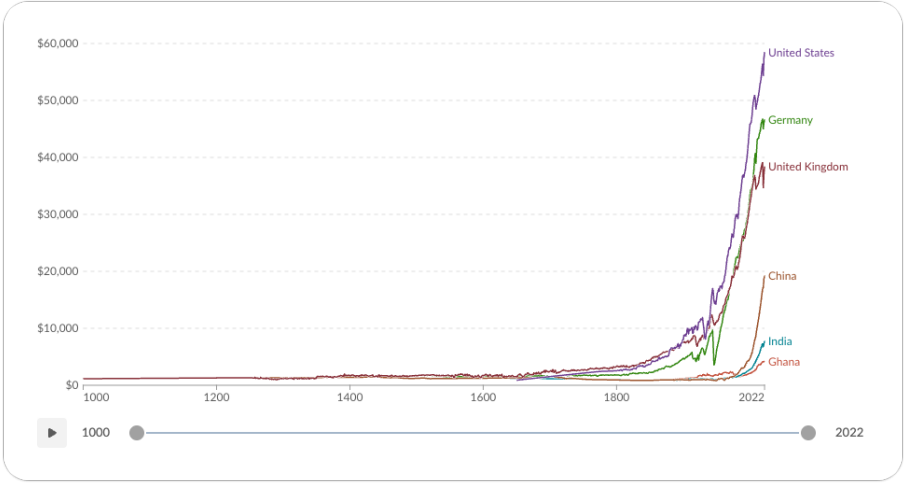
Fatih Kansoy · Worcester College, University of Oxford

Outline

1. Course orientation and motivation
2. From examples to predictions
3. Defining the prediction task
4. Model fitting and evaluation
5. Auditing predictive performance
6. Scope of predictive evidence

Part A · Why this matters

What happened and what is happening?



Source: Our World in Data

You used AI a dozen times before breakfast

Face unlock
recognises your face and unlocks the phone

Maps ETA
predicts when you will arrive

Spam filter
keeps junk out of your inbox

What to watch
suggests the next title for you

Fraud alert
flags an unusual payment

Autocomplete
guesses the next word

It is already running your day, quietly.

AI is already mainstream

72%

of surveyed firms use AI in at least one function

up from 55% a year earlier

~800M

weekly ChatGPT users

reportedly ~100M within two months of launch

$252bn

estimated corporate AI investment

in a single year

Source: McKinsey, State of AI 2024 (surveyed firms); OpenAI (Oct 2025); ~100M is a UBS/Similarweb estimate, not an OpenAI figure; Stanford AI Index 2025.

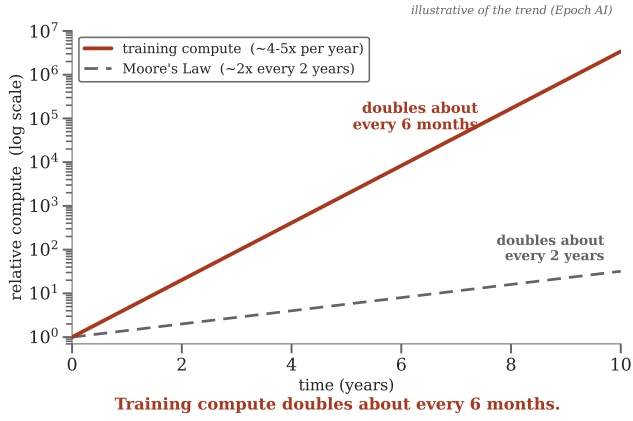
This is not a pilot. It is the operating system of modern business.

Why now: three forces arrived at once

Data
almost everything is now logged

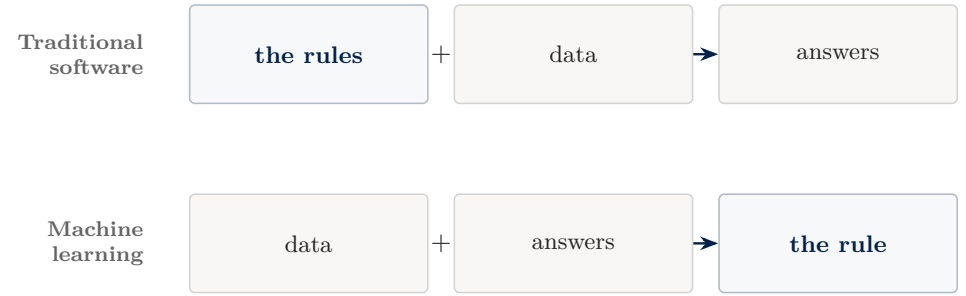
Compute
training power doubles about every 6 months

Algorithms
deep learning that scales with data



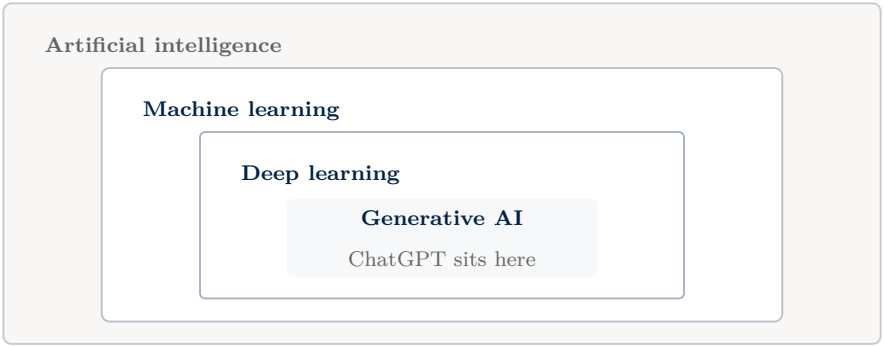
Source: Epoch AI (∼4–5x per year), a measured trend, not a physical law; contrast Moore's Law (∼2x every 2 years).

We stopped writing the rules



We stopped writing the rules and started learning them from examples.

One family of ideas, nested



Everything this week is machine learning on your own business data.

Course structure

Day	Analytical question	Main methods and outcome
1	What can these data credibly predict?	prediction contract, loss, baselines, data splitting, and evaluation
2	How do simple models estimate a number or probability?	linear regression, logistic regression, regularisation, and uncertainty
3	When does flexibility improve prediction?	decision trees, random forests, validation, and task-matched metrics
4	What structure can be learned without labels?	K-means, PCA, text representations, retrieval, and generative AI

Later methods cannot repair a target, timeline, or population that was defined incorrectly at the beginning.

Day 1 learning objectives

By the end of the session, students should be able to:

1. write a prediction contract specifying the unit, target, population, horizon, information cutoff, output, and intended use
2. match an evaluation to the task by choosing a loss, baseline, data split, and metric
3. diagnose timeline leakage, evaluation leakage, and misleading metric denominators
4. distinguish response prediction, operational decisions, and causal effects

The bank example develops these objectives; the parcel exercise and BIS application test whether students can apply them independently.

Part B · From examples to predictions

Supervised learning from labelled examples

A labelled example pairs:

- information available about a case, and
- the outcome subsequently observed for that case.

$$\underbrace{\text{historical features and outcomes}}_{\text{labelled examples}} \longrightarrow \underbrace{\text{estimate a rule}}_{\text{training}} \longrightarrow \underbrace{\text{predict an unseen case}}_{\text{application}}.$$

The objective is not to reproduce the historical records. It is to produce useful predictions for cases that were not used to estimate the rule.
Days 1–3 use this supervised-learning structure. Day 4 considers what can be learned when outcome labels are absent.

One observation: features and outcome

Consider one illustrative parcel-delivery record:

Record	Distance	Dispatch time	Rain	Delivered late?
P–1042	12.4 km	16:20	yes	yes

- **Unit:** one dispatched parcel.
- **Features $\mathbf{x}_i$:** distance, dispatch time, and rain; information available when the prediction is made.
- **Outcome y_i:** whether delivery was late, observed after delivery.
- **Identifier:** P–1042 names the record; it is not a predictive feature.

One labelled observation is therefore the pair $(\mathbf{x}_i, y_i)$.

Variables and observed values

Statistical notation separates quantities that vary across possible cases from the values recorded for one case.

Quantity	Population variable	Value in record i
Features	$\mathbf{X}$	$\mathbf{x}_i$
Outcome	Y	y_i

If a record contains p features,

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}).$$

- Capital letters describe variables across possible observations.
- Lowercase letters describe realised values in the data.
- x_{ik} is feature k recorded for observation i.

The historical sample

Collecting n labelled observations gives the historical sample

$$\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}.$$

Read this as:

> *"$\mathcal{D}_n$ contains n records; record i contains features $\mathbf{x}_i$ and observed outcome y_i."*

n	number of observations
p	number of features in each $\mathbf{x}_i$
i	index identifying a record

The table supplies observed examples. The relationship in the wider population remains unknown and must be estimated.

Numerical and binary outcomes

Days 1–3 use two principal forms of supervised prediction:

Outcome	Example	Model output
Numerical $Y \in \mathbb{R}$	delivery time, fare, or demand	predicted number $\widehat{y}_j$
Binary $Y \in \{0, 1\}$	late/on time, subscribe/not subscribe	predicted probability $\widehat{p}_j$

For a binary event,

$$Y = \begin{cases} 1, & \text{the event occurs,} \\ 0, & \text{the event does not occur.} \end{cases}$$

A binary class is not yet an operational action. It is obtained later by applying a threshold to an estimated probability.

The conditional prediction target

A central population quantity is the conditional mean:

$$m(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}].$$

Read this as:

> *"m of $\mathbf{x}$ is the expected, or average, value of Y among cases whose available features equal $\mathbf{x}$."*

When $Y \in \{0, 1\}$,

$$p(\mathbf{x}) \equiv \Pr(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}].$$

Thus $p(\mathbf{x}) = 0.30$ means that the event occurs for 30% of comparable cases in the population. It is an unknown population probability, not an observed value for one case.

Prediction and residual variation

Define the residual relative to the conditional mean:

$$\varepsilon = Y - m(\mathbf{X}).$$

It follows directly that

$$Y = m(\mathbf{X}) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid \mathbf{X}] = 0.$$

If comparable parcels have $m(\mathbf{x}) = 0.30$:

Observed outcome	y_i	Residual $y_i - 0.30$
Parcel is late	1	0.70
Parcel is on time	0	-0.30

Individual residuals need not be zero; their conditional average is zero by construction. Residual variation is relative to the information in $\mathbf{X}$, not necessarily permanently unknowable noise.

Training estimates a prediction rule

$$\underbrace{\mathcal{D}_T = \{(\mathbf{x}_i, y_i) : i \in T\}}_{\text{training observations}} \xrightarrow{\mathcal{A}} \underbrace{\widehat{f}_T}_{\text{fitted prediction rule}} .$$

- T identifies the observations assigned to training.
- $\mathcal{A}$ is the learning algorithm: the procedure used to fit a rule.
- $\widehat{f}_T$ is the rule produced from training set T.
- The hat indicates estimation; the subscript records which training observations determined the fitted rule.

For numerical prediction, the aim is $\widehat{f}_T(\mathbf{x}) \approx m(\mathbf{x})$. For a probability model, the aim is $\widehat{f}_T(\mathbf{x}) \approx p(\mathbf{x})$.

Prediction for an unseen case

Let $j \notin T$ denote a case that was not used to fit the rule:

$$\underbrace{\mathbf{x}_j}_{\text{features of unseen case } j} \xrightarrow{\widehat{f}_T} \underbrace{\widehat{p}_j = \widehat{f}_T(\mathbf{x}_j)}_{\text{estimated event probability}} .$$

- $i \in T$ indexes observations whose features and outcomes were used for training.
- $j \notin T$ indexes a new or held-out observation.
- The fitted model receives $\mathbf{x}_j$, but it does not use y_j when producing the prediction.
- Each unseen case receives its own prediction.

The letters i and j have no inherent past or future meaning; this is the indexing convention used in the course.

Probability and observed outcome

Three related quantities must be kept separate:

Quantity	Notation	Meaning
Population probability	$p(\mathbf{x}_j)$	unknown event probability among comparable cases
Model estimate	$\widehat{p}_j$	probability estimated by the fitted rule for case j
Realised outcome	$y_j \in \{0, 1\}$	event or no event, observed after prediction

If $\widehat{p}_j = 0.30$, the model assigns case j an estimated 30% probability of the event. It does not predict the outcome $y_j = 0.30$; the later outcome is still either 0 or 1.
For a held-out case, the evaluator may possess y_j, but it remains excluded from model fitting and hidden until the prediction has been made.

Prediction and decision

Suppose a parcel has an estimated probability $\widehat{p}_j = 0.30$ of late delivery. Should it receive expedited handling?

$$a_j = d(\widehat{p}_j, \text{ benefits, costs, capacity, constraints}) .$$

Here a_j is the chosen action for parcel j.

1. The prediction model supplies an estimated probability.
2. The decision rule $d(\cdot)$ adds service objectives and operational constraints.
3. Two organisations can rationally choose different actions from the same probability because their costs, capacities, or obligations differ.

A prediction informs an action; it does not determine the action by itself.

Prediction and intervention

A second question is whether expedited handling would actually prevent a late delivery:

$$\underbrace{Y_j(1)}_{\text{outcome if expedited}} - \underbrace{Y_j(0)}_{\text{outcome under normal handling}}.$$

Question	Meaning	Evidence required
Prediction	which parcels are likely to be late?	performance on relevant unseen cases
Intervention	which late outcomes would be prevented by expediting?	randomisation or another credible causal comparison

Only one of $Y_j(1)$ and $Y_j(0)$ can be observed for the same parcel. Predictive association alone does not identify their difference.

From labelled data to prediction

The complete supervised-learning sequence is:

$$\underbrace{\mathcal{D}_T}_{\text{labelled training records}} \xrightarrow{\mathcal{A}} \underbrace{\widehat{f}_T}_{\text{fitted rule}} .$$

$$\underbrace{\mathbf{x}_j}_{\text{unseen features}} \xrightarrow{\widehat{f}_T} \underbrace{\widehat{p}_j}_{\text{prediction}} \xrightarrow{\text{later}} \underbrace{y_j}_{\text{observed outcome}} .$$

1. Define what one observation represents.
2. Record its features and observed outcome.
3. Use labelled training observations to estimate a rule.
4. Apply that rule to the features of an unseen case.
5. Compare the prediction with the outcome observed later.

The comparison in step 5 supplies evidence about predictive performance; it does not by itself establish the value or causal effect of an action.

The prediction rule does not define its own task

Data and a learning algorithm estimate $\widehat{f}_T$, but they do not determine what the prediction should mean.

Estimated from labelled data	Specified before fitting
fitted rule $\widehat{f}_T$	what one observation represents
fitted parameters and predictions	target event and prediction horizon
empirical performance on selected data	population and prediction time
	available features, loss, baseline, and evaluation design

Part C specifies these choices for the bank marketing campaign before any model or accuracy claim is accepted.

Part C · Defining the prediction task

A business objective is not a prediction task

"Improve the marketing campaign with AI" states an objective but leaves the statistical problem unresolved.

1. What does one observation represent, and which population should the result cover?
2. What outcome counts, and by when must it occur?
3. When is the prediction made, and which information exists then?
4. Is the required output a probability, a class, or a ranking?
5. Which errors matter, what baseline must be beaten, and which unseen cases will be used for evaluation?

These choices define the prediction task before an algorithm is selected.

Bank campaign: setting and observed outcome

A Portuguese bank contacted clients by telephone to offer a **term deposit**: savings held for a fixed period in return for interest. A campaign could involve more than one contact with the same client.

client and campaign information $\longrightarrow$ telephone contact(s) $\longrightarrow$ subscription outcome.

For a recorded client–campaign observation i,

$$y_i = \begin{cases} 1, & \text{the client subscribed to the term deposit,} \\ 0, & \text{the client did not subscribe.} \end{cases}$$

The outcome is known only after the relevant campaign interaction.

The bank campaign data

The full file contains 41,188 client–campaign observations, 20 input columns, and the outcome y. The table shows a subset of the columns.

age	job	marital	education	contact	month	campaign	poutcome	duration	y
56	housemaid	married	basic.4y	telephone	may	1	nonexistent	261	no
57	services	married	high.school	telephone	may	1	nonexistent	149	no
37	services	married	high.school	telephone	may	1	nonexistent	226	no

- `campaign` counts contacts made for this client during the current campaign, including the last contact.
- `duration` records the length of the last contact in seconds.
- y records whether the client subscribed.

A row summarises one client–campaign observation; it is not simply one isolated telephone call.

From historical campaigns to a future prediction

The course uses the following task: immediately before the recorded final campaign contact, estimate whether the client will subscribe by the end of that campaign interaction.

$$\underbrace{\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}}_{\substack{\text{historical campaign records} \\ \text{outcomes observed}}} \xrightarrow{\text{fit}} \underbrace{\widehat{f}}_{\text{fitted response rule}} .$$

$$\underbrace{\mathbf{x}_j}_{\substack{\text{future or held-out case} \\ y_j \text{ not yet observed}}} \xrightarrow{\widehat{f}} \underbrace{\widehat{p}_j = \widehat{f}(\mathbf{x}_j)}_{\text{estimated subscription probability}} .$$

- i indexes historical records used to learn the relationship.
- j indexes a case whose outcome is unknown at prediction time.
- The model produces a separate $\widehat{p}_j$ for every eligible future case.

Interpreting a reported accuracy

Suppose the fitted system is reported to achieve **90% validation accuracy**. The claim cannot be interpreted without answering:

1. Which task, population, and prediction time were evaluated?
2. What accuracy is available from a no-features baseline?
3. How many subscribers were missed, and how many non-subscribers were incorrectly classified?
4. Were the evaluation records excluded from fitting and representative of the intended use?

The rest of Day 1 supplies the definitions and evidence required to answer these questions.

Sample and target population

Two contract fields answer *who*:

- **Unit**: one recorded client–campaign observation.
- **Sample**: contacts selected and recorded by the bank from May 2008 to November 2010.
- **Target population**: comparable future campaign observations to which the bank intends to apply the model.

Because the file contains only clients the bank chose to contact, predictive performance on these records does not automatically extend to all bank customers or to a differently selected campaign.

Target event, horizon, and prediction time

Three timing choices make the target testable:

- **Event**: subscription to the term deposit.
- **Horizon**: by the end of the recorded campaign interaction.
- **Prediction time**: immediately before the recorded final campaign contact begins.
- **Information cutoff**: no information created during or after that contact may be used.

This is the analytical contract chosen for the course. The data file alone does not determine when a prediction must be made.

Output and evaluation criterion

Output

- number
- probability
- class
- ranking

For a binary target, this course predicts a probability and chooses the action separately.

Evaluation

1. a loss for each miss
2. a no-features baseline
3. cases the model never saw

Part D develops all three.

The reported 90% is therefore incomplete: its baseline, error composition, and evaluation sample must still be established.

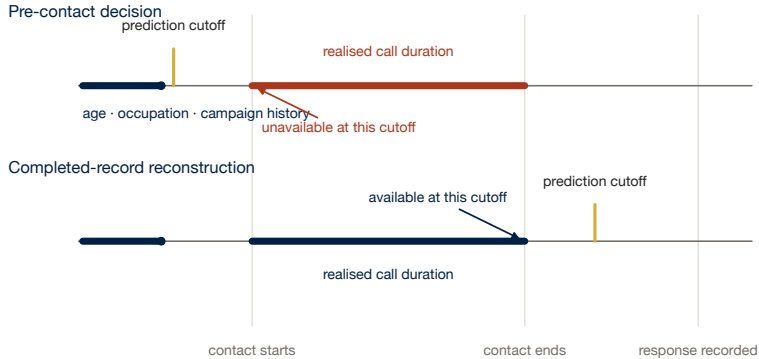
Feature availability at prediction time

The data table defined `duration` as the length of the last campaign contact. Apply the information cutoff:

1. The response probability is produced immediately before that contact begins.
2. Its duration is known only after the contact has ended.
3. The course's pre-contact model must therefore exclude `duration`.

Feature validity is determined by a simple test: would this value exist at the stated prediction time?

Feature validity depends on the task



Bank task timeline against the pre-final-contact information cutoff.

The prediction contract

1. **Unit**: what one row describes
2. **Population**: the cases the result must cover
3. **Target and horizon**: what outcome, by when
4. **Information cutoff**: the last usable moment of information
5. **Features**: what is available at prediction time
6. **Output**: number, probability, class, or ranking
7. **Evaluation**: loss, baseline, and unseen cases

Change one field and the statistical problem can change, even when the data table looks the same.

Bank campaign prediction contract

Field	This task
1. Unit	one recorded client–campaign observation
2. Population	comparable future campaign observations under the bank's contact process
3. Target and horizon	whether subscription is recorded by the end of the campaign interaction
4. Information cutoff	immediately before the recorded final contact begins
5. Features	information available at that moment; final-contact duration is excluded
6. Output	an estimated subscription probability for each eligible case
7. Evaluation	loss and classification metrics on held-out records, against a training-sample baseline

The contract determines how the reported 90% accuracy must be interpreted.

Comparison: fare prediction after trip completion

Field	This task
1. Unit	one completed New York yellow-taxi trip
2. Population	completed city trips like those in the file
3. Target and horizon	the fare recorded when the trip ended, in dollars
4. Information cutoff	after completion: an after-the-fact classroom prediction
5. Features	recorded trip information, including distance and duration
6. Output	a number, in dollars
7. Evaluation	dollar error on held-out trips, against the training-mean baseline

The course uses 50,000 filtered trips from January 2024. Unlike the bank task, trip duration is available at the stated prediction time.

Concept check: the bank prediction contract

Without looking back, state:

1. the unit and target population
2. the event, horizon, and information cutoff
3. the allowed features, output, loss, baseline, and held-out evidence

Then test `duration`: does it exist at the cutoff? Day 2's models operate inside these fields; changing a field changes the statistical problem.

Part D · Model fitting and evaluation

Fitted rules depend on the training sample

Apply the same learning procedure $\mathcal{A}$ to two possible training samples:

$$T \xrightarrow{\mathcal{A}} \widehat{f}_T, \qquad T' \xrightarrow{\mathcal{A}} \widehat{f}_{T'}.$$

- The subscript records which observations determined the fitted rule.
- Different finite samples can produce different fitted parameters and predictions.
- The fitting criterion must therefore be stated explicitly, and performance must be assessed beyond the records used for fitting.

Part D begins by defining the penalty that selects $\widehat{f}_T$.

Squared loss selects the smaller total penalty

Illustrative data: fares of \$10, \$14, and \$30. Candidate A always predicts \$18; candidate B always predicts \$14.

	Candidate A: \$18		**Candidate B: \$14**	
Fare	miss	squared	miss	squared
\$10	−8	64	−4	16
\$14	−4	16	0	0
\$30	+12	144	+16	256
Total		**224**		272

Candidate B hits one fare exactly, but squared loss keeps A because $224 < 272$.

Change the loss, and the winning prediction can change

Same three trips, same two candidates, but now the penalty is the absolute miss.

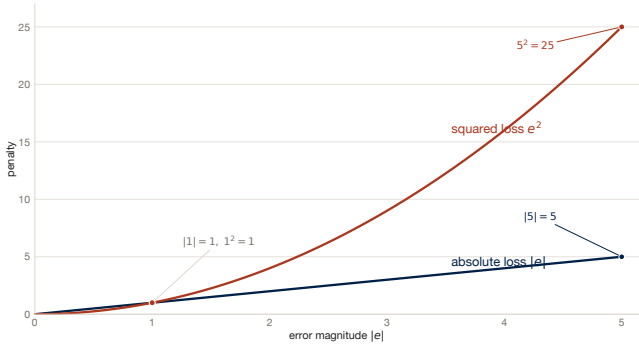
$$\text{Candidate A: } 8 + 4 + 12 = 24, \qquad \text{Candidate B: } 4 + 0 + 16 = 20.$$

- Under absolute loss, B wins because $20 < 24$.
- Candidate A, \$18, is the **mean**; candidate B, \$14, is the **median**.
- Squared loss points to the mean; absolute loss points to the median.

"Best prediction" has no meaning until the loss is named. Day 2 proves the squared-loss result.

Squaring makes large errors dominate

- A $12 miss costs 12 under absolute loss.
- The same miss costs 144 under squared loss.
- Large misses therefore pull much harder on the squared-loss winner.



Absolute and squared penalties as error grows; illustrative.

Fitting means minimising average training loss

The constant-fare calculation has this general form:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{|T|} \sum_{i \in T} L(y_i, f(\mathbf{x}_i)).$$

$\mathcal{F}$: allowed rules · L: one-case penalty · T: training cases · $|T|$: their count · $\arg\min$: the rule attaining the smallest average

- Our menu $\mathcal{F}$ contained two constant rules: \$18 and \$14.
- Other models change the menu and search method; the comparison by average training loss remains.

The loss encodes which mistakes matter

The penalty rule is called the **loss**, written $L(y, f(\mathbf{x}))$. Choose it to match the task:

Target	Loss	What it penalises
a number	squared loss	large misses especially strongly
a number	absolute loss	every miss in direct proportion to its size
a probability	binary log loss	confident probabilities contradicted by the outcome
a class	0–1 loss	every wrong class equally

The bank uses probability-level loss; a spam filter may use class error. Ranking measures arrive on Day 3.

Training loss is observable; population risk is the goal

> Picture a student who memorises last year's answer sheet. On last year's paper: 100%. On this year's paper: failure. Model fitting only ever looks at last year's paper.

Now the same idea in symbols:

$$\widehat{R}_T(f) = \frac{1}{|T|} \sum_{i \in T} L(y_i, f(\mathbf{x}_i)), \qquad R_P(f) = \mathbb{E}_P\left[L(Y, f(\mathbf{X}))\right].$$

T: the training cases · P: the target population · $\mathbb{E}_P$: the average over new cases drawn from it

- $\widehat{R}_T(f)$ is the loss we can compute and minimise.
- $R_P(f)$ is performance on new population cases—the quantity we actually care about.

Held-out data estimate the second quantity without reusing the cases that fitted the rule.

Predictive gain means beating the matching baseline

A **baseline** is the best no-features guess under the chosen loss:

Judged by	The no-features baseline
squared loss	the training mean
absolute loss	the training median
binary log loss	the training response-rate baseline
accuracy	the training majority class

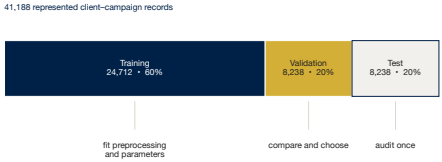
- For event probabilities, the training yes-share is the **response-rate baseline**, also called prevalence.
- For accuracy, predict the majority class for every case.
- Claim predictive gain only after beating the matching baseline on unseen cases.

Day 2 proves why squared loss selects the training mean.

Training, validation, and test data have different jobs

A fixed 60/20/20 split, stratified: each part keeps the same 11.3% yes-rate, so no part is accidentally easier than the others.

Split	Records	Yes	Its job
Training	24,712	2,784	fit the model
Validation	8,238	928	compare, choose
Test	8,238	928	audit once



Bank file: 41,188 records, stratified 60/20/20.

> **Keep the test set locked until the final audit; repeated checking turns it into another validation set.**

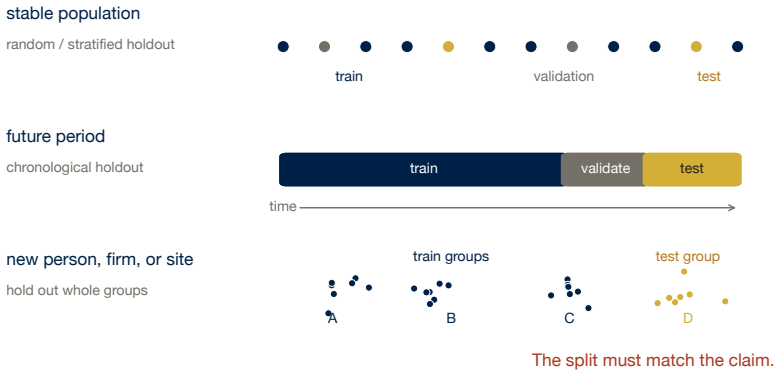
Overfitting appears as a widening train–held-out gap

A flexible menu can memorise training noise and accidents:

- training loss continues to fall
- held-out loss stops improving, or rises
- the gap between them becomes the warning signal

One further caution: our random split answers one question, new records from a stable population. Appendix C lists the splits for future periods, new clients, and new places.

The holdout design must match the deployment claim



Three holdout designs: stable population, future period, new entities.

Checkpoint: can you explain why the winner won?

Answer three concrete questions:

1. Why did \$18 beat \$14 under squared loss, but lose under absolute loss?
2. Which no-features baseline matches accuracy?
3. Which split fits, which split chooses, and which split is used once?

If you can answer all three, you have the equipment needed to judge the bank's 90% claim.

Part E · Auditing predictive performance

A better score can answer the wrong question

Follow what happens when **duration** enters the pre-final-contact model:

1. Long calls correlate with subscriptions, so validation accuracy rises.
2. Duration is recorded only after the final call ends.
3. At prediction time, the deployed model cannot know it.

> **Timeline leakage**: a feature that will not exist at the moment the prediction must be made.

The arithmetic is real, but it evaluates a model that cannot be used for the stated task.

Split first; fit preprocessing on training data only

> **Evaluation leakage**: information from the validation or test data leaks into how you prepare or choose the model. The evaluation then no longer imitates real deployment.

1. **Wrong order**: standardise all 41,188 records, then split. Test information has already influenced training rows.
2. **Correct order**: split first; estimate preprocessing quantities on training data; apply the frozen transformation elsewhere.

One timing test catches most feature leakage

For the pre-final-contact model, decide for each column before you read the answers: would the value exist before the recorded final contact begins?

Column	Available?	Role
age	yes	known before the final contact
month of the call	yes	scheduled in advance
previous campaign outcome	yes	recorded history
duration of the final call	**no**	timeline leakage
the subscription outcome	**no**	the target itself

Ask the same question of every candidate feature before fitting: would this value exist at the prediction moment?

Monitoring catches relationships that decay

Google Flu Trends estimated US flu activity from flu-related search queries:

- It fitted search behaviour from one period.
- In February 2013, its estimate exceeded the CDC measure by more than twofold.
- From August 2011 to September 2013, it overestimated flu prevalence in 100 of 108 weeks (Lazer et al., *Science*, 2014).
- Google stopped publishing the estimates in August 2015.

Evaluation is not a one-time event: behaviour changes, so monitored performance can decay.

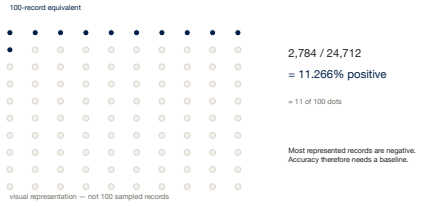
Predicting “no” for everyone already scores 88.7%

Start with the no-features guess:

$$\text{training yes rate} = \frac{2{,}784}{24{,}712} = 11.27\%$$

$$\text{all-no accuracy} = \frac{7{,}310}{8{,}238} = 88.74\%$$

Because “no” is the training majority class, the baseline predicts “no” for every validation record.



Training split: 2,784 of 24,712 records subscribed.

The model’s 90.2% accuracy is about 1.4 percentage points above this baseline, not 90 points above zero.

The confusion matrix exposes what accuracy hides

For the valid pre-final-contact model on the validation set, call a record positive when $\widehat{p}_i \geq 0.50$:

	Predicted positive	Predicted negative	Total
Actually positive	$TP = 216$	$FN = 712$	928
Actually negative	$FP = 99$	$TN = 7{,}211$	7,310
Total	315	7,923	8,238

- Subscribers: 216 correctly flagged; 712 missed.
- Non-subscribers: 99 false alarms; 7,211 correctly rejected.

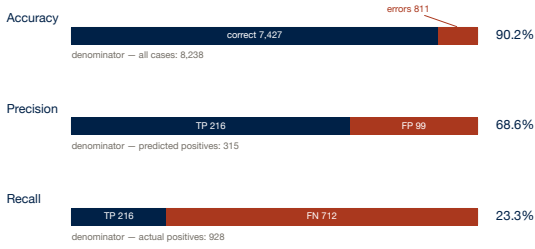
Accuracy adds the diagonal cells and hides which kind of error produced the total.

One model can have 90.2% accuracy and 23.3% recall

$$\text{Accuracy} = \frac{TP + TN}{8{,}238} = 90.2\%$$

$$\text{Precision} = \frac{TP}{TP + FP} = 68.6\%$$

$$\text{Recall} = \frac{TP}{TP + FN} = 23.3\%$$



Accuracy: all records · Precision: flagged records · Recall: actual subscribers

Validation set at $\tau = 0.50$: the three metrics divide by different denominators. Definitions in Appendix D.

At $\tau = 0.50$, the model finds fewer than one quarter of actual subscribers despite its high accuracy.

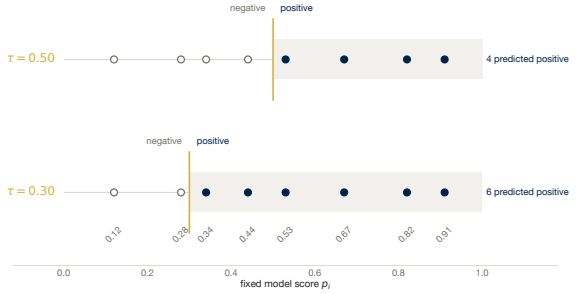
The threshold converts probabilities into class decisions

A probability becomes a class only after you choose a threshold τ:

$$\widehat{y}_i = \mathbf{1}[\widehat{p}_i \geq \tau].$$

$\mathbf{1}[\cdot]$: 1 if the condition holds, else 0 · τ: the cut-off you choose

- $\tau = 0.50$ is a software default, not a law.
- Moving τ changes every cell of the confusion matrix.



The same fixed scores under two thresholds; illustrative.

Tomorrow we derive τ from the costs of the two kinds of mistake. Day 3 studies whole curves of thresholds at once.

Held-out accuracy answers prediction—not causation

Job	What it asks for
Prediction	the likely outcome for a case like this one
Explanation	the relationships that organise the fitted prediction
Intervention	the change in outcome an action would cause

A feature can predict strongly while ruining the other two jobs:

- **Proxy**: postcode predicts repayment because it stands in for income.
- **Consequence**: a symptom predicts a disease because it follows the disease.
- **Selection**: retention calls predict saved customers because staff called the clients they already judged saveable.

A predictive association alone cannot reveal whether a feature is a proxy, a consequence, or part of a selection process.

A model can reproduce historical bias accurately

Amazon's experimental CV-screening tool illustrates the limit:

- It trained on roughly ten years of submitted CVs, most from men.
- It penalised the word "women's" and downgraded graduates of two all-women's colleges.
- Amazon said recruiters never used it to evaluate candidates; the project was scrapped (Dastin, Reuters, 2018).

The model reproduced a pattern in its data. Whether that pattern should guide a decision is a separate Day 4 question.

Who will subscribe is not who will be persuaded

Prediction question

Among records like these, who tends to subscribe?

Intervention question

For whom would making a call change the outcome?

The bank file cannot **identify** the second quantity because:

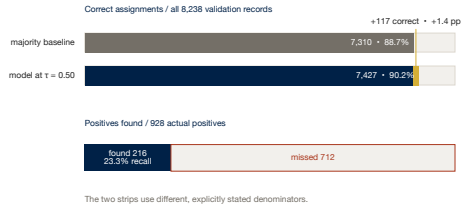
- it records the bank's chosen calling strategy
- it contains no comparable group left uncalled
- alternative explanations therefore remain

A high response score describes association, not the causal effect of calling. Day 2 introduces the uplift question.

The 90% result is real—but narrow

The claim survives as evidence for the contract we wrote, and for nothing wider:

- unrounded: 90.16% against a baseline of 88.74%, a gain of about 1.4 points
- at $\tau = 0.50$, recall is 23.3%: most subscribers are missed
- a validation result, not the untouched test audit
- tied to every choice: population, cutoff, features, split, threshold



The validation evidence behind the 90% claim, read against its baseline and its recall.

Checkpoint: diagnose the score before trusting it

Name the failure in each case:

1. A feature exists only after the prediction moment.
2. A preprocessing step used the test-set mean.
3. A response probability is presented as the effect of calling.

Then state the baseline and recall that qualify the 90% accuracy.

The answers are timeline leakage, evaluation leakage, and an unsupported intervention claim.

Part F · Scope of predictive evidence

Trust the evidence chain, not the headline number

The bank case gives a reusable structure:

$$\text{trustworthy claim} = \text{prediction contract} + \text{baseline} + \text{held-out evidence} + \text{decision scope}.$$

Contract: what is predicted · Baseline: what no-feature prediction already achieves · Evidence: performance on unused cases · Scope: what the result can justify

- Here, 90.16% is about 1.4 points above the 88.74% baseline.
- Recall is 23.3%, and the result is validation evidence, not the untouched test audit.
- The score predicts response; it does not identify the effect of calling.

The method travels: define, compare, hold out, and limit the claim.

Five review questions for tomorrow

Write short answers tonight; we audit them in tomorrow's first minutes.

1. What baseline did the 90% model have to beat, and what was its value?
2. A new column makes your validation score jump. What do you check first?
3. Why may the test set be used only once?
4. Accuracy 90.2% and recall 23.3% describe the same model. Explain how both can be true.
5. Why does a high response probability alone not show that calling causes subscriptions?

Every answer is on today's slides. None needs a computer.

Day 1 built the judge; Day 2 begins the model list

A number, a probability, and a ranking are outputs, not model names. Today established how any model will be judged: contract, loss, baseline, split, and scope. The application adds an auditable rule system, not a fitted machine-learning family.

Tomorrow we fit and interpret two named models:

- linear regression learns taxi fares from 50,000 completed trips
- its slope becomes dollars per mile
- logistic regression turns an additive index into bank response probabilities through a sigmoid
- decision costs determine when a probability is enough to act on

Baselines, losses, metrics, and thresholds help judge or use a model; they are not additional models.

Reference · Tools behind the argument

Appendix A: symbols used across the week

Symbol	Meaning
n	number of records
p	number of encoded features
$\mathbf{x}_i \in \mathbb{R}^p$	feature vector of record i
y_i	observed target of record i
$\mathbf{X} \in \mathbb{R}^{n \times p}$	the encoded data table
$\mathcal{D}_n$	the n pairs $(\mathbf{x}_i, y_i)$

Symbol	Meaning
$\mathcal{F}$	the menu of candidate rules
$L\big(y, f(\mathbf{x})\big)$	penalty for one prediction
$\widehat{R}_T(f)$	average penalty on training set T
$R_P(f)$	expected penalty in population P
τ	classification threshold

- Raw categorical, text, image, or audio features need an encoding before entering a numerical model.
- If $S = 1$ marks entry into the sample, the fitted relationship $\mathbb{P}(Y \mid \mathbf{X}, S = 1)$ describes records like those collected, and nothing wider without further argument.

Appendix B: four learning settings

- **Supervised learning** fits a mapping from inputs to observed targets. Days 1 to 3 live here.
- **Unsupervised learning** searches for structure without a designated target, such as clusters or low-dimensional summaries.
- **Self-supervised learning** constructs training targets from unlabelled data, for example predicting a missing or next word, then trains with supervised-style optimisation.
- **Reinforcement learning** learns a policy from rewards produced by actions and subsequent states.

One system can combine several settings: a generative assistant may use self-supervised pretraining, supervised fine-tuning, preference information, retrieval, and policy constraints.

Appendix C: choose a holdout that matches use

Deployment question	Holdout principle	What goes wrong if you ignore it
new records from a stable population	random or stratified split	subsets drift to different event rates
a future period	chronological split	the model gets to peek at the future
an entirely new person or firm	hold out whole entities	the model looks good only because it already knows these people
a later record for a known entity	chronological or rolling split	training sees the client's future history
a new location or institution	geographic or site holdout	local patterns are assumed to travel without evidence

The bank file is date ordered with no reliable client identifier, so our stratified split demonstrates model-selection discipline; pick the row that matches your claim.

Appendix D: each metric changes the denominator

Measure	Denominator	What it reports
Accuracy	all cases	the share of correct class assignments
Precision	predicted positives	the share of flagged cases where the event is present
Recall	actual positives	the share of real events that are flagged
Specificity	actual negatives	the share of real non-events that are rejected

Each row reports a share, and the denominator names the group it is a share of. All four depend on the chosen threshold. Probability-level measures such as log loss do not.

Reference: a seven-question prediction audit

1. What is one record, and which population must the result cover?
2. What outcome, by when, and from which information cutoff?
3. Would every feature be available at the moment of prediction?
4. Which loss judges the prediction, and what does the matching no-features baseline score?
5. Were training, validation, and test kept apart, with the test set used once?
6. Did any preprocessing step see the held-out data?
7. Is the claim predictive, explanatory, or interventional, and does the evidence match?

References

- Moro, S., Cortez, P., and Rita, P. (2014). "A Data-Driven Approach to Predict the Success of Bank Telemarketing." *Decision Support Systems*, 62, 22–31. Data: UCI Machine Learning Repository.
- New York City Taxi and Limousine Commission. *Yellow Taxi Trip Records*, January 2024.
- Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). "The Parable of Google Flu: Traps in Big Data Analysis." *Science*, 343(6176), 1203–1205.
- Butler, D. (2013). "When Google Got Flu Wrong." *Nature*, 494, 155–156.
- Dastin, J. (2018). "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." Reuters, 10 October 2018.
- James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning*. Springer.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Springer.
- Shmueli, G. (2010). "To Explain or to Predict?" *Statistical Science*, 25(3), 289–310.
- Hernán, M. A., and Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.