

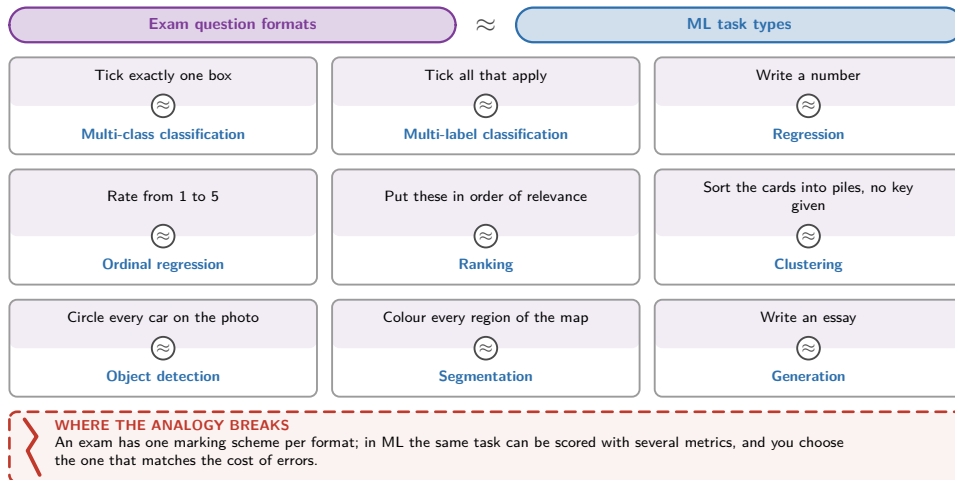


Task Types Atlas

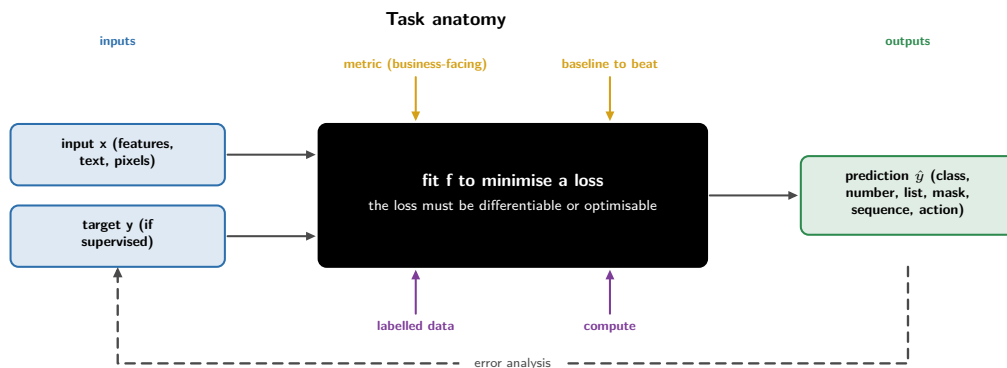
Model atlas | Every ML task type as one row: input → output, training loss, typical metric, baseline to beat, where it is taught; an identification key and the pairs people confuse

#ModelAtlas #TaskTypes #ProblemFraming #Reference

What a task type is



Task types are question formats on an exam paper: the format fixes how answers are marked before anyone answers



The anatomy every row of this sheet fills in

Loss is not metric

The **loss** is what training minimises (smooth, e.g. cross-entropy); the **metric** is what you report and decide on (often non-differentiable, e.g. F1, NDCG@10). Losses live in losses-and-training, metrics in ML 07 and metrics-and-kpis.

Prediction tasks: labels and numbers

Task	Input → output	Training loss	Typical metric	Baseline	Taught in
Binary classification	$x \rightarrow P(y = 1)$	binary cross-entropy	ROC-AUC, PR-AUC, F1	majority class; logistic regression	ML 03, 07
Multi-class	$x \rightarrow \text{one of } K \text{ (softmax)}$	categorical cross-entropy	accuracy, macro-F1, top-k	most frequent class	ML 03, 07
Multi-label	$x \rightarrow K \text{ independent yes/no}$	sum of K binary CE	micro/macro-F1, Hamming loss	predict each label's prior	ML 07
Regression	$x \rightarrow \text{real number}$	MSE (MAE, Huber if outliers)	RMSE, MAE, R^2	training mean (median for MAE)	ML 02
Quantile / probabilistic	$x \rightarrow \text{quantiles or a distribution}$	pinball loss, negative log-likelihood	pinball, CRPS, interval coverage	empirical quantiles	ML 13, 15
Ordinal regression	$x \rightarrow \text{ordered class}$	cumulative-link (ordinal logistic)	MAE on ranks, weighted kappa	median class	ML 03
Count regression	$x \rightarrow 0, 1, 2, \dots$	Poisson deviance	Poisson deviance, MAE	mean count	ML 03
Survival (time to event)	$x \rightarrow \text{survival curve, censored labels}$	Cox partial likelihood	C-index, integrated Brier	Kaplan-Meier curve	ML 15 (recall)
Ranking (learning to rank)	query + candidates → ordered list	pairwise (RankNet), listwise (LambdaRank)	NDCG@k, MRR, MAP	BM25 or popularity order	LP (ranking)
Recommendation	user, context → ranked items	BPR, sampled softmax, click BCE	recall@k, NDCG@k, online CTR	most popular items	ML 14
Retrieval	query → top-k of millions	contrastive (InfoNCE)	recall@k, MRR	BM25 keyword search	LLM 02, 11

Supervised prediction. Taught in: ML = machine-learning, LP = learning-paradigms folder (sheet number after the folder)

Loss	Formula	Used for
Binary cross-entropy	$-\frac{1}{n} \sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$	binary, multi-label (per label)
Categorical cross-entropy	$-\frac{1}{n} \sum_i \log p_{i, y_i}$	multi-class, next token
Mean squared error	$\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2$	regression
Pinball (quantile)	$\frac{1}{n} \sum_i \max(\tau e_i, (\tau - 1)e_i)$	quantile regression, forecasts

where n = number of examples (count); y_i = true label (0/1) or true value of example i (class or target unit); p_i = predicted probability of the positive class (0-1); p_{i, y_i} = predicted probability of the true class of example i (0-1); $\hat{y}_i$ = predicted value (target unit); τ = target quantile, e.g. 0.9 (0-1); e_i = error, true value minus predicted value (target unit)

The four losses that cover most rows (recall only; derivations in losses-and-training)

🔍 Unsupervised, sequential and perception tasks					
Task	Input → output	Training loss	Typical metric	Baseline	Taught in
Clustering	x → cluster id	k-means inertia, GMM likelihood	silhouette; ARI if labels exist	one cluster; random split	ML 10
Dimensionality reduction	x (d dims) → z ($k \ll d$)	reconstruction error	explained variance, trustworthiness	raw features	ML 11
Anomaly detection	x → anomaly score	none, or one-class objective	PR-AUC, precision@k on labelled cases	z-score or IQR rule	ML 12
Density estimation	x → p(x)	negative log-likelihood	held-out log-likelihood, bits/dim	single Gaussian fit	ML 15, DL 11
Forecasting	past series + covariates → next h values	MSE, MAE or pinball	MASE, WAPE, CRPS	naive (last value), seasonal naive	ML 13
Sequence labelling	tokens → one tag per token	per-token CE, CRF likelihood	entity-level F1	most frequent tag per word	P&G 09, SPAI
Structured prediction	x → tree, graph, sequence with constraints	structured hinge, CRF likelihood	exact match, task-specific F1	greedy local classifier	LP (structured)

No labels, or labels over time: the baseline matters even more because metrics are harder to read

Task	Input → output	Training loss	Typical metric	Baseline	Taught in
Image classification	image → class	cross-entropy	top-1 / top-5 accuracy	linear probe on a pretrained backbone	P&G 01
Object detection	image → boxes + classes + scores	box L1 / GloU + focal or CE	mAP@[.5:.95]	pretrained COCO detector	P&G 01
Semantic segmentation	image → class per pixel	pixel CE, Dice	mIoU	pretrained segmenter, zero-shot SAM masks	P&G 01
Instance / panoptic segmentation	image → one mask per object	detection + mask BCE	mask AP, panoptic quality	Mask R-CNN-style model	P&G 01

Keypoints / pose	image → joint coordinates	heatmap MSE	OKS-based AP, PCK	pretrained pose model	P&G 01, 08
Speech recognition (ASR)	audio → text	CTC or seq2seq CE	word error rate (WER)	pretrained Whisper-class model	P&G 05

Perception. P&G = perception-and-generation folder

Step	Statement	Go to / result
1a	The output is a category	→ go to 2
1b	The output is not a category	→ go to 4
2a	Exactly one category per item	→ go to 3
2b	Any number of categories per item	Multi-label classification
3a	Categories have an order (1-5 stars)	Ordinal regression
3b	No order between categories	Binary / multi-class classification
4a	The output is a number	→ go to 5
4b	The output is structured	→ go to 6
5a	The number is for future time steps	Forecasting
5b	The number describes the present item	Regression
6a	The output is an ordered list of items	Ranking / retrieval / recommendation
6b	Something else	→ go to 7
7a	The output is regions of an image or signal	Detection / segmentation
7b	Something else	→ go to 8
8a	The output is new content (text, image, audio)	Generation
8b	Something else	→ go to 9
9a	The output is an action that changes the world	Control / RL
9b	No labels: find structure or outliers	Clustering / anomaly detection

Identification key: answer each pair of statements to name the task. Read 1a or 1b, then go to the step given

⚡ Generation, language and decision tasks					
Task	Input → output	Training loss	Typical metric	Baseline	Taught in
Text generation	prompt → text	next-token cross-entropy	perplexity, human or LLM-judge preference	template or retrieval answer	LLM 03
Image generation	text or noise → image	denoising MSE (diffusion), adversarial	FID, CLIP score, human preference	retrieval of a stock image	DL 12-13, P&G 03
Machine translation	source text → target text	token CE (teacher forcing)	BLEU, chrF, COMET	existing MT system	LLM, P&G 09
Summarisation	long text → short text	token CE	ROUGE, faithfulness rate	lead-3 (first three sentences)	LLM
Question answering	question (+ context) → span or answer	span start/end CE or token CE	exact match, token F1	best-overlap sentence	LLM 11

Text-to-speech	text → waveform	spectrogram L1 / codec-token CE	MOS, WER of re-transcription	concatenative TTS	P&G 05
Control (RL)	state → action	policy-gradient or TD loss	average return, success rate	random or rule policy; PID	RL 01-05
Contextual bandit	context → one arm	regret-minimising update	cumulative regret, CTR	uniform A/B split	RL 06
Planning / opti- misation	goal + model → plan	none (search) or learned value	plan cost, success rate	greedy heuristic	algorithms, SPAI

Most of these are conditional generation; LLMs now solve many of them with one model, but the metric and baseline still differ per task

Scalar	one real number; linear head, MSE e.g. house price
Probability vector	K values summing to 1; softmax head, CE e.g. digit 0-9
Bit vector	K independent probabilities; sigmoid head, BCE e.g. photo tags
Ordered list	scores sorted per query; ranking loss e.g. search results
Mask or box set	per-pixel classes or variable-size set of boxes e.g. tumour outline
Sequence	tokens produced one by one; CE per token e.g. translation
Action	a choice that changes the next input; return-based loss e.g. robot torque

Output shapes: the type of $\hat{y}$ decides the head of the model and the loss

Multi-class softmax, one answer	vs	Multi-label K sigmoids, any subset
Tell: Can two labels be true at once?		
Regression random CV allowed	vs	Forecasting time-ordered splits only
Tell: Is the target in the future of the features?		
Detection boxes	vs	Segmentation pixel masks
Tell: Is a rectangle precise enough?		
Semantic segmentation class per pixel	vs	Instance segmentation mask per object
Tell: Must two touching cars be counted as two?		
Imbalanced classification labelled rare class	vs	Anomaly detection few or no labels
Tell: Do you have enough labelled positives to train on?		
Retrieval recall stage, millions → 100	vs	Ranking precision stage, 100 → 10
Tell: Is it about not missing, or about ordering the top?		
Extractive QA / summary copies spans	vs	Abstractive writes new text
Tell: Can every output word be found in the source?		

One question tells each pair apart

PROBLEM Daily sales 10, 12, 11, 13, 12. A model forecasts 13 and 13 for the next two days; actuals are 14 and 13. Is the model better than naive?

GIVEN

- $\text{MASE} = \text{MAE of the model on the test days} / \text{MAE of the naive one-step forecast on the training days}$
- naive one-step forecast: tomorrow = today

1 Naive in-sample errors

$|12-10|, |11-12|, |13-11|, |12-13| = 2, 1, 2, 1$

Why: each day predicted by the day before

2 Naive MAE

$(2 + 1 + 2 + 1) / 4 = 1.5$

Why: the scale of a free forecast

3 Model MAE on test

$(|14-13| + |13-13|) / 2 = 0.5$

Why: mean absolute error of the model

4 MASE

$0.5 / 1.5 = 0.33$

Why: below 1 beats naive

ANSWER $\text{MASE} \approx 0.33$: the model's errors are a third of the naive forecast's

CHECK ✓ A MASE of 1 would mean no better than copying yesterday (ML 13)

Baselines are numbers too: score the naive forecast before the model

Interview trap

"Which metric for a 1 % fraud model?" Accuracy is wrong (99 % for always saying no). Answer PR-AUC or recall at a fixed precision, and name the baseline (prevalence = PR-AUC of a random model, here 0.01).

✓ Myths and recap

MYTH

✗ Accuracy works for any classification task

✗ Forecasting is regression with a date column

✗ LLMs make task types obsolete

✗ Generated content cannot be evaluated

FACT

✓ Only for balanced classes and equal error costs
Use PR-AUC, F1 or cost-weighted metrics when one class is rare.
Source: ML 07, 09

✓ Splits must respect time and the baseline is the naive forecast
Shuffled K-fold trains on the future and inflates scores.
Source: ML 13

✓ One model, many tasks, but each task still has its own metric and baseline
A chatbot answering questions is scored on exact match or faithfulness, a classifier on F1.
Source: LLM 15

✓ It is evaluated by reference metrics, learned metrics and human or LLM judges
BLEU/ROUGE, FID, COMET and pairwise preference rates are all standard.
Source: metrics-and-kpis

Framing mistakes that cost whole projects

Output type → task
category, number, list, mask, sequence, action

Task → loss
CE, MSE, pinball, ranking, CTC

Task + costs → metric
chosen before training

Always a baseline
majority, mean, naive, BM25, popularity

Time in the target?
forecasting rules apply

Two labels at once?
multi-label, not multi-class

Pin these next to your problem statement

Q1 Output head and loss for multi-label classification?

A K independent sigmoid outputs, sum of K binary cross-entropies; one threshold per label.

Q2 Baseline and metric for a demand forecaster?

A Naive or seasonal-naive forecast; MASE (below 1 beats naive), WAPE or CRPS for probabilistic forecasts.

Q3 Typical metric for object detection?

A mAP averaged over IoU thresholds 0.5 to 0.95 (COCO-style).

Q4 Why is ordinal regression not multi-class?

A Its classes are ordered: predicting 1 star for a 5-star review must cost more than predicting 4; multi-class CE ignores that order.

Q5 Which loss gives a 90 % prediction interval's upper bound?

A Pinball (quantile) loss with $\tau = 0.9$.

Q6 Retrieval vs ranking in a search system?

A Retrieval cheaply narrows millions of items to about 100 (recall matters); ranking orders those 100 with a heavier model (NDCG@10 matters).

Q7 Metric for speech recognition?

A Word error rate: (substitutions + deletions + insertions) / words in the reference; lower is better.

Q8 Baseline for summarisation of news?

A Lead-3: the first three sentences of the article, a strong baseline on news datasets.

Q9 When is a problem anomaly detection rather than classification?

A When labelled anomalies are too few or keep changing; you model normal data and score distance from it.