AI and Machine Learning

Day 4: From hidden structure to decisions you must answer for

Day 4 of 4

Fatih Kansoy
Oxford Summer School

Outline

Yesterday's five questions, audited in one pass

Keep your written answers in front of you.

1. A leaf predicts the training average of its region: Day 2's mean, restricted to one rectangle. A tiny leaf is memorisation in miniature.
2. For B trees of variance σ^2 with average correlation ρ, the average has variance $\rho\sigma^2 + (1-\rho)\sigma^2/B$. More trees shrink only the second term; the lever is decorrelating the trees, not adding more of them.
3. One validation set is one noisy draw. Five folds average that noise into a steadier judge.
4. PR beats ROC when positives are rare, as with the bank file's 11.3% subscribers.
5. The two audit questions: could every feature be known at prediction time, and does the split imitate deployment.

If any of these surprised you, reread the Day 3 slide behind it before Friday.

Thirty-two famous companies raise today's opening question

I selected 32 recognisable US-listed companies across seven sectors: Apple and Microsoft in technology, JPMorgan in financials, Exxon in energy, Netflix in communications, and 27 more. For each one I have nine years of daily returns.

Here is the question. On a normal trading day, how many dominant, mutually uncorrelated directions describe the movement of these 32 prices?

about 32 about 7 close to 1

> **Write your number down now and keep the paper. It is priced at the close of the day, with two computed verdicts.**

What you can do by the end of today

1. Group cases without labels, and judge the grouping without flattering it.
2. Find the few directions that carry most of the variation, and read their loadings.
3. Explain how embeddings connect this week's tools to modern AI.
4. Audit a deployed system for fairness, monitoring, and governance.

The first two answer today's opening question. The fourth closes a promise Day 1 made and has not yet kept.

The day runs from structure-finding to responsibility

question → data → loss → model → evaluation → decision → monitoring

- **Part I: grouping without labels.** K-means on the 32 companies, and how to judge a grouping.
- **Part II: directions that carry the variation.** PCA on the stocks, then on the US yield curve.
- **Part III: the same tools, one level up.** Embeddings, retrieval, and where modern AI sits.
- **Part IV: responsibility.** Fairness, monitoring, governance, and the close of the week.

Where today sits on Day 1's nesting map

On Day 1's map, Days 1 to 3 lived in one of four learning settings: supervised learning, data with answers attached.

- Parts I and II today are **unsupervised learning**: structure with no designated target.
- Part III touches **self-supervised learning**: the model predicts held-back parts of its own input, so the data supply the target for free. That trick powers generative AI.
- Reinforcement learning stays beyond this course.

By the end of today, two additional learning settings have been introduced. Reinforcement learning remains outside the course.

Where we are

Part I: grouping without labels

What can data tell you when nobody marks the right answer?

The 32 companies arrive with no target column. There is nothing to predict and no outcome to be wrong about. This part builds the one tool we need, K-means, and the discipline for judging what it returns.

When the label disappears, the evidence standard changes with it

All week, one column told us the truth: the fare, the subscription. Today no column does. Take thirty seconds: which pieces of the week's discipline die with that column?

Three go at once, because every judge this week compared a prediction with an answer:

- the loss on outcomes
- the baseline to beat
- the test audit of "correct"

Three checks replace them, each one askable of any grouping:

- **Separation:** the groups sit far apart, compared with the spread inside them.
- **Stability:** the same groups reappear under restarts and resampling.
- **Recognisability:** a person who knows the field would recognise the groups.

Treat every grouping as a proposal to be checked against these three.

The stock file: nine years of daily returns for 32 companies

- The data are daily log-returns: each one the day's percentage price change, measured on a log scale.
- The file runs from 5 January 2016 to 30 December 2024: 2,262 trading days per company.
- A second small file records each company's sector, seven sectors in all.

> Say what your data are. I pulled these returns from Yahoo Finance with the `yfinance` package: a convenience pull from a retail source, not a research-grade file. Its known flaw is survivorship bias: only companies that survived to today are in the file. These data teach structure on names you recognise; they support no research inference.

Standardised return histories make co-movement a distance

K-means needs one vector per case. Here a case is a company and its features are the same 2,262 trading dates. For company j, standardise its returns over time:

$$z_{tj} = \frac{r_{tj} - \bar{r}_j}{s_j}, \qquad \|\mathbf{z}_j - \mathbf{z}_\ell\|^2 = 2(T-1)(1-\rho_{j\ell}).$$

$\mathbf{z}_j$: company j's standardised return history · $\rho_{j\ell}$: the sample correlation between companies j and ℓ · $T = 2{,}262$

- high positive correlation gives a short distance
- zero correlation gives the middle distance $2(T-1)$
- negative correlation pushes the companies farther apart

So Euclidean K-means on these histories is a direct co-movement analysis. The seven course-authored sector labels are kept outside the fit and used only as a later comparison.

K-means minimises total squared distance to k centres

Choose a number of groups, k. K-means searches for k centre points and an assignment of every case to one centre, minimising:

$$\min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2.$$

C_j: the cases assigned to centre j · $\boldsymbol{\mu}_j$: that centre · $\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$: squared gaps added up feature by feature, Day 2's squared error again

- The double bars are notation, not a new idea; the two sums say: do that for every case in every cluster, and total it.
- You already own the central fact. On Day 2 you proved the mean minimises total squared error, so for any fixed group the best centre is that group's mean: the "means" in K-means.

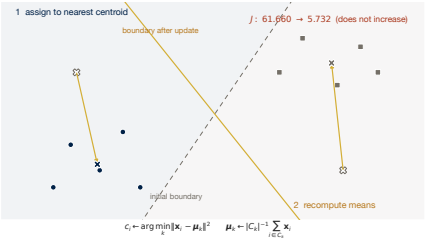
Appendix A writes out the algebra.

The assign-and-update loop finds the centres

The search alternates two moves, each one easy:

1. **Assign:** give each point to its nearest centre.
2. **Update:** move each centre to the mean of its assigned points.

Run one round in your head. Assign can only lower the total: each point stays put or moves somewhere closer. Update lowers it again: the mean is the best centre for a fixed group. A total that only falls must settle.



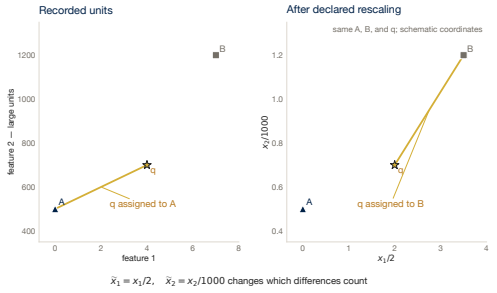
Schematic; the objective values are figure-internal, not fitted data.

One caution: the loop can settle into a good-but-not-best arrangement. Software restarts from random positions and keeps the best run. Our locked calculation uses 100 starts per k and seed 42.

Distance adds across features; scaling is a choice

The squared distance adds up squared differences, feature by feature. A feature measured in large units contributes large differences, and it quietly dominates the grouping.

The standard remedy is to standardise: shift each feature to mean zero and spread one, so every column counts alike. That is a modelling choice, not a law. In a customer file with income in euros and age in years, the scaling decision largely decides the clusters you get.



Schematic coordinates; the same points under two declared scalings.

In the stock analysis the columns are dates, but the scientific object is each company's history. Standardising within company removes volatility scale: K-means then responds to co-movement, not to the fact that one stock simply swings more.

The silhouette scores how well each case sits in its cluster

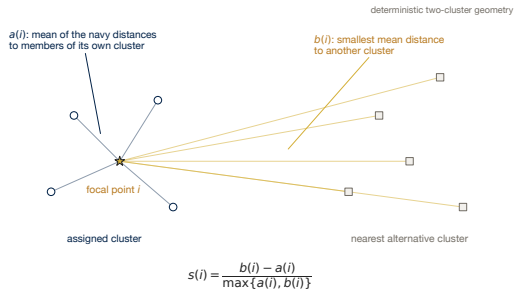
For case i:

$$s(i) = \frac{b_i - a_i}{\max(a_i, b_i)}.$$

a_i: mean distance to its own cluster · b_i: smallest mean distance to another cluster

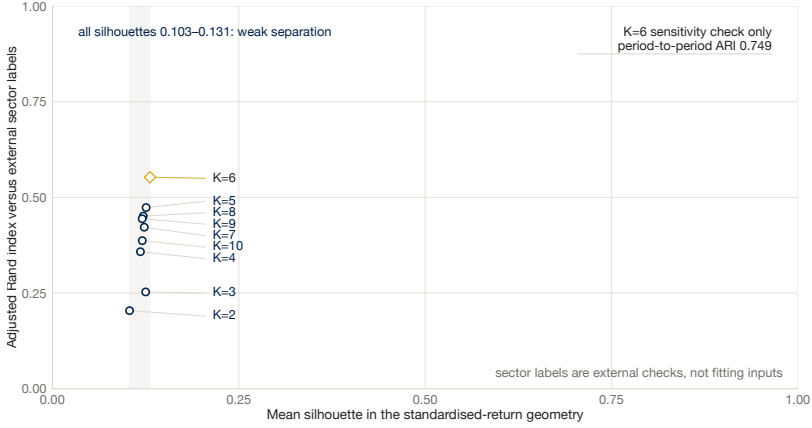
- near 1: firmly inside
- near 0: on a border
- below 0: nearer a rival group

We compare $k = 2, \ldots, 10$. No sector label chooses k.



Deterministic two-cluster geometry; not the stock data.

The evidence does not reveal one clean number of groups



Yahoo Finance convenience sample, 2016–2024; course-authored sector labels.

Three checks answer three different questions

For the working $k = 6$ partition:

- **Separation:** silhouette 0.131. The geometry is weak; every candidate from $k = 2$ to 10 lies between 0.103 and 0.131.
- **Recognisability:** sector ARI 0.553. The fitted groups align moderately with the seven external sector labels, despite having a different number of groups.
- **Sensitivity:** early-period versus late-period ARI 0.749 at fixed $k = 6$. This is one chronological check, not proof of population stability.

The frozen seed makes the calculation reproducible; it does not make $k = 6$ a fact. Nearby candidates and alternative initialisations remain plausible because the separation is weak.

> **The honest conclusion is useful structure with unresolved boundaries.**

Name clusters by behaviour, never by character

- The cluster labels 1 to 6 carry no meaning: swap two labels and nothing about the analysis changes.
- Meaning enters only when you profile each cluster: its size, its members, its average behaviour.

Then name what you observed:

- “Moves with energy prices” is a name the data can support.
- “The risky group” is not: it is an invented character, attached to every member.

The distinction matters beyond style. A character name attached to a group of people has a way of becoming an action taken against them.

Part IV shows exactly that happening.

Part I checkpoint: three things you can now do

You can now:

- run the assign-and-update loop in your head, and say why software uses many starting positions
- distinguish weak separation (0.131), moderate external agreement (0.553), and one chronological sensitivity check (0.749)
- name a cluster by its observed behaviour and refuse the character name

The groups are proposals, not discovered natural kinds. Part II turns the same matrix and asks a different question: which directions carry its variation?

Where we are

1. Grouping without labels
2. **The few directions that carry the variation**
3. The same tools, one level up
4. What a deployed system owes the people it scores
5. Close
6. Appendix

Part II: the few directions that carry the variation

Can a handful of directions describe 32 correlated series?

Part I grouped the companies. This part compresses their movements: it looks for the few underlying directions along which all 32 series move together.

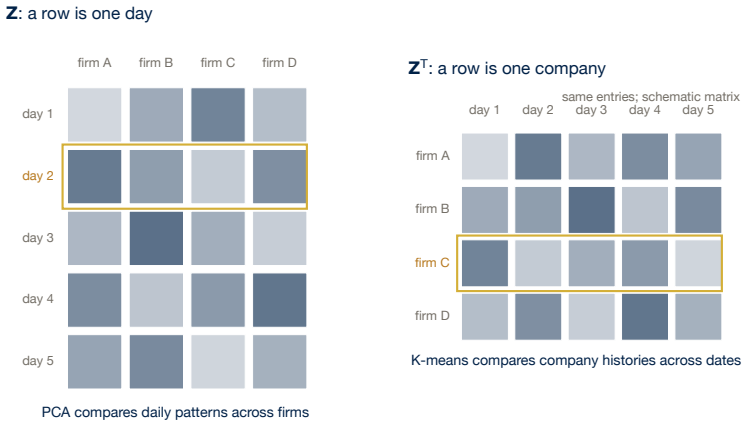
Correlated features repeat each other, so fewer directions may be enough

Our 32 return series correlate at 0.389 on average. Correlated series repeat part of each other's story, so far fewer uncorrelated directions may describe much of the variation.

Principal component analysis, PCA, makes that idea exact: it finds the directions that carry the most variation, in order, and tells you how much each one carries.

> **One caution, said once and kept all day: variance is spread, not importance. A direction can carry a large share of the variation and still be irrelevant to your task.**

The same matrix, transposed, asks a different question



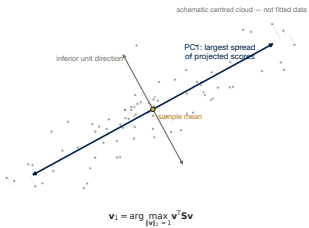
Schematic; the fitted stock matrix is $2{,}262 \times 32$.

For PCA, each day is now a point

In Part I each company was a point. Now each day is one point: 2,262 points in 32 dimensions, one axis per company. To see the cloud, start with two companies:

- plot one point per trading day: Apple's return across, Microsoft's up
- the two move together, so the points form a tilted oval, longest where both rise at once
- that long axis is the first principal component of a two-company world

Give every company its own axis: along which single direction does the cloud stretch the most?



Schematic centred cloud; not fitted data.

The first principal component: largest variance

Written as mathematics, the search for the long axis reads:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \widehat{\mathrm{Var}}(\mathbf{X}\mathbf{w}).$$

$\mathbf{w}$: a recipe of 32 weights, one per company, length fixed at one · $\mathbf{X}\mathbf{w}$: the 32 return columns blended into one weighted series · keep the recipe whose blended series has the largest spread

- The answer has a known form: $\mathbf{w}_1$ is the top eigenvector of the covariance matrix.
- An eigenvector, in plain words, is a direction the matrix stretches without turning; in the two-company picture, the long axis of the oval.

Appendix D sketches the derivation; software computes it in a numerically stable way.

Loadings describe the variables; scores describe the cases

The direction $\mathbf{w}_1$ has one entry per company. Those entries are the **loadings**. Projecting one day's returns onto the direction gives that day's **score**.

object	one number per	it tells you
loading	company (a variable)	how much this variable takes part in the direction
score	day (a case)	where this case sits along the direction

For stocks, both have a physical reading. Treat the direction as a basket of the 32 companies:

- each loading is a company's weight in the basket
- each day's score is the basket's combined move that day

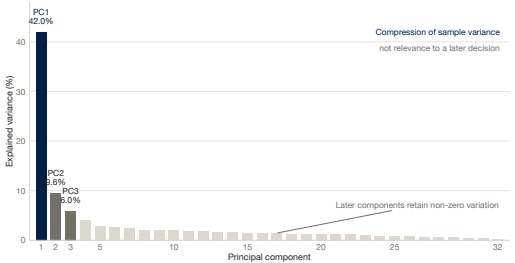
Keep the rule: loadings belong to columns, scores belong to rows.

One direction carries 42.0% of everything these stocks do

Each component's **explained variance ratio** is the share of total variation its scores carry. After standardising each series:

- PC1 carries 42.0%
- PC2 carries 9.6%, PC3 6.0%

Nearly half of the daily movement of 32 different companies is one shared movement. Said as compression: one number per day, that day's PC1 score, carries 42.0% of nine years of all 32 series.



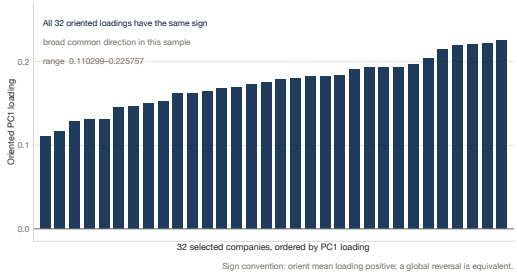
Computed from the frozen stock convenience sample, 2016–2024.

Your written number has met its first evidence; keep the paper. Before the loadings tell us: name a force that could move Apple, Exxon and JPMorgan on the same day.

Stock PC1 is a broad market-like direction in this sample

PC1's loadings: all 32 are positive, ranging from 0.110 to 0.226.

- the mathematics fixes a direction only up to its sign, so the shared sign is convention; the claim is the pattern
- positive weights across all 32 names make the component a broad, index-like basket in this sample
- a loading measures participation in this sample component; it is not CAPM beta, which requires a specified market return



Frozen teaching sample, 2016–2024.

The yield curve is a second dataset, and it poses a decision

- A Treasury yield is the interest rate the US government pays to borrow, and the maturity says for how long.
- Ten maturities, from 3 months to 30 years, read together form the curve. I take the ten yields daily, 2010 to 2026, from FRED.
- We work with each day's *changes*: 4,130 rows.

For the stocks I standardised each series before PCA.

> Decide before I tell you: should we standardise the ten yield series too? Consider what the ten share that the 32 stocks did not.

The reasoning is about votes, and each recipe fits its data

- Standardising gives every series an equal vote on the shared directions; skip it, and the series with bigger swings vote louder.
- Stock volatilities differ widely, and one volatile company should not own the definition of co-movement: so there, every series got one spread.
- Daily yield changes already share one unit, percentage points, so a maturity that moves more should count for more.

The yield recipe is therefore covariance PCA on first differences, with no per-series scaling.

Both recipes are correct. Each fits its data. Blurring them, one recipe for everything, is a real and common error.

Three factors carry 95.7% of the yield curve's variation

component	share of total variation
PC1	80.2%
PC2	11.6%
PC3	3.9%
first three together	**95.7%**

One question before the three directions get their names. The stocks' first component carried 42.0%; this one carries 80.2%. What is different about ten maturities of one government's debt, compared with 32 companies from seven sectors?

- the maturities co-move far more tightly: their daily changes correlate at 0.62 on average, against the stocks' 0.389
- the more the columns repeat each other, the more one direction can carry

The loading patterns are level, slope, and curvature

- **PC1:** all ten loadings share one sign; the whole curve rises or falls together: the **level**.
- **PC2:** positive short, negative long, the sign changing between 5 and 7 years; the curve steepens or flattens: the **slope**.
- **PC3:** positive at both ends, negative in the middle, a U shape: the **curvature**.

Read each component's loadings as a shape across the ten maturities. As with the stocks, the mathematics fixes each direction only up to its sign; the patterns are the claims.

Litterman and Scheinkman (1991) described exactly these three factors in bond returns, and trading desks had names for them before any of this software existed.

An unsupervised method has rediscovered a result practitioners named by hand.

On 9 March 2020, one factor rebuilt most of a crisis day

Keeping only the first components compresses the data; how much of one real day survives? Take 9 March 2020, a large yield move during the COVID crash. That day is ten numbers, one change per maturity.

rebuilt from	share of that day's variation captured
the level factor alone	0.843
all three factors	0.980

- be careful with the labels: these two numbers are shares for that single day; the 80.2% and 95.7% are averages over all 4,130 days
- even 0.980 leaves 2% of a very large move unexplained, and on a crisis day that remainder is measured in money

> **A high share can still hide exactly the piece a risk system most needed to see.**

Part II checkpoint: three things you can now do

You can now:

- state what the first principal component is, and keep loadings and scores apart
- read 42.0% and 80.2% as explained-variance shares, and read loading patterns as level, slope, and curvature
- explain why K-means compares rows of $\mathbf{Z}^\top$ while PCA compares rows of $\mathbf{Z}$

One boundary matters: PCA found a broad common direction in this sample. It did not establish a causal factor, and its loadings are not CAPM betas. Both parts began by choosing a vector representation. Part III asks what changes when the representation itself is learned.

Where we are

1. Grouping without labels
2. The few directions that carry the variation
3. **The same tools, one level up**
4. What a deployed system owes the people it scores
5. Close
6. Appendix

Part III: the same tools, one level up

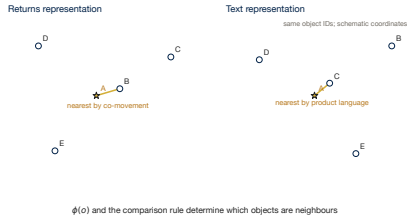
How do this week's tools reappear inside modern AI?

This part is short and carries no new numbers. Its point is a continuity claim: the systems you came to this course curious about are built on the machinery you now own.

Representation determines which objects become neighbours

Part I used a hand-designed representation: each company became a standardised return history, and the distance had an exact correlation interpretation.

- change the representation and the distances change
- the nearest neighbour may therefore change even when the distance calculation is flawless



Schematic coordinates; the same objects under two representations.

An embedding is a learned representation

Modern AI can learn the coordinates rather than receive them from the analyst. A word, image, document, or customer becomes a vector whose nearby points should be similar for the intended task. That learned vector is an **embedding**. The training signal determines what "nearby" means:

- a word's vector is learned from the words around it
- a customer's from the trail of what they buy
- an image's from predicting held-back parts of the picture

An embedding is not automatically meaningful. Its neighbours must be evaluated against the use, the population, and the errors that matter.

Vector arithmetic carries meaning: one famous example

The canonical example is word2vec (Mikolov et al., 2013), which learned a vector for each word from the words around it. Now complete this, in that vector space:

$$\text{king} - \text{man} + \text{woman} \approx ?$$

Every room answers at once: “queen”.

- that is the famous result, and it is real in this instance: meaning has become geometry, and arithmetic on the vectors respects it
- one caution: the example you just solved flatters the method. Later work found analogy arithmetic less reliable than the folklore suggests; treat it as a picture of the idea.

Once objects are vectors, this week's tools return with new names

Every document, image, and customer is now a vector, and you own exactly three tools for vectors. Make the mapping yourselves before the table does. A company holds a million support messages as embeddings; match each job to one of your tools:

- find the messages that mean the same as a new one
- discover the recurring themes
- draw the whole collection as one readable picture

this week's tool	one level up, it becomes
distance in feature space	semantic search: find the documents that mean the same
K-means	exploratory grouping: cluster text vectors, then inspect whether coherent themes appear
PCA	the map: compress embeddings to a picture you can read

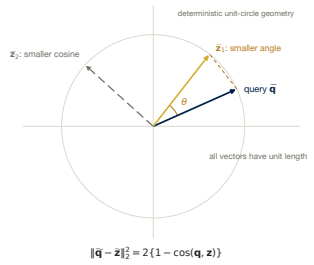
You had all three.

The workhorse measure of likeness is cosine similarity

Cosine similarity asks how nearly two vectors point the same way, ignoring their lengths:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}.$$

the dot in the numerator multiplies matching entries and adds them · near 1: the same direction · near 0: approximately orthogonal in this representation



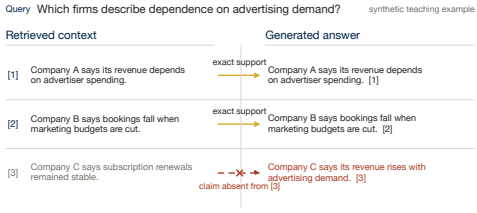
Deterministic unit-circle geometry.

Ignoring length is the design: after a suitable representation and normalisation, a short note and a long report can still point in a similar direction. Whether that direction captures topic, style, or something else depends on how the vectors were constructed.

Auditing retrieval-augmented generation, link by link

A common RAG design answers in two steps:

- **Retrieve:** rank passages with lexical, dense, or hybrid search.
- **Generate:** write an answer conditioned on what was retrieved.



Synthetic teaching example; not model output.

Now audit it: a fluent, confident, wrong answer arrives, Day 1's lesson in new clothes. Which link do you check first? Both failures happen, so evaluate the links separately:

- a wrong document fetched gives a fluent answer built on the wrong evidence
- a right document may still be used unfaithfully

Day 1's map is complete, and Part III closes with it

Place today on the map:

- Parts I and II filled in unsupervised learning
- Part III sits with self-supervised learning and generative AI: embeddings are learned by predicting held-out context, a target the data supply for free
- supervised learning, Days 1 to 3, surrounds it all as the evaluation discipline

You can now:

- say what an embedding is and what cosine similarity measures
- name the week's three tools in their level-up forms: search, topics, maps
- audit a retrieve-then-generate pipeline link by link, in evidence-chain terms

One sentence carries us into the last part. Every audit this week asked whether a system was right. Part IV asks what remains when it is wrong: who bears the errors, and who answers for them.

Where we are

1. Grouping without labels
2. The few directions that carry the variation
3. The same tools, one level up
4. **What a deployed system owes the people it scores**
5. Close
6. Appendix

Part IV: what a deployed system owes the people it scores

Who bears the errors, and who answers for them?

This is the question Day 1 left open on purpose, and the week ends by answering it. The tools are the ones you already hold: proxies, error rates, thresholds, monitoring. We point them at deployed systems.

On Day 1 I promised you this conversation, and it starts with Amazon

Day 1, final hour: the Amazon CV tool. Reconstruct it before I do: what the tool penalised, and what it had been trained on.

> The record, again. Amazon built an experimental tool to screen CVs, trained on roughly ten years of CVs submitted to Amazon, most of which came from men. The tool penalised CVs containing the word "women's" and downgraded graduates of two all-women's colleges. Amazon said it was never used by its recruiters to evaluate candidates. The project was experimental, and scrapped (Dastin, Reuters, 2018).

Day 1's verdict stopped, on purpose, at: the model reproduced the pattern in its data. Whether that pattern should drive any decision belongs to Day 4. Today is Day 4.

> **What discipline stops a faithfully learned pattern from becoming an unfair decision?**

Each technical choice carried a responsibility question

Nothing in this part is new machinery. Each choice you made this week has a second reading:

day	the technical choice	the responsibility question attached
1	which data, which population	who is in the file, and who is missing from it
2	which loss, which threshold	which mistakes you priced, and against whom
3	which split, which audit	whether the evidence imitates real deployment
4	who is grouped and scored	which groups bear the errors

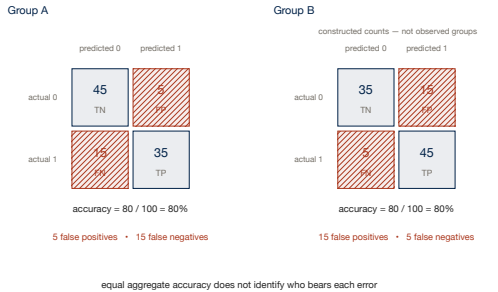
Responsible AI reads the same choices for their effect on people.

Error rates by group reveal who bears the mistakes

Compute Day 1's confusion matrix separately for each group g the decision touches:

$$\text{recall}_g = \frac{TP_g}{TP_g + FN_g},$$

$$\text{FPR}_g = \frac{FP_g}{FP_g + TN_g}.$$



Constructed counts; not observed groups.

FPR_g: the share of group g's innocent cases that the model wrongly flags

Read the second in plain words, because it carries the day's weight:

- a model can look fine overall while one group's innocent members receive most of the false accusations
- with a nationality in place of g, that sentence is the case that closes today

No rule can deliver every fairness criterion at once

Fairness can be demanded in several forms:

- equal recall across groups
- equal false-positive rates
- probabilities that mean the same thing in every group

When the true rate of the outcome differs across groups, no rule can deliver all of these at once. The reason is arithmetic, not ideology: Appendix F lets you check it in a ten-second invented example, next to the formal definitions.

Choosing which criterion to enforce is therefore a decision about which harm matters most.

A health algorithm under-served Black patients; race was never an input

Obermeyer, Powers, Vogeli and Mullainathan (*Science*, 2019) studied a widely used US health-risk algorithm. It under-served Black patients. Race was not a feature.

Find the leak before I show it. It is not in the features, so a feature audit comes back clean.

Take Day 1's contract, all seven fields. Which field is the door?

The harm entered through field three: the target

The algorithm predicted healthcare **cost** as a proxy for healthcare **need**.

health need → care received → [cost, the proxy target] → score → support

- the break is in the second arrow: at the same level of sickness, less money was spent on the care of Black patients
- so a model trained on cost read equally sick patients as healthier: an equal predicted score did not mean an equal illness burden
- correcting the target raised the share of Black patients flagged for extra help from 17.7% to 46.5%
- the authors estimated such tools are applied to roughly 200 million Americans a year

The lesson generalises: the target is itself a choice, and a proxy target can carry a bias no feature audit will find. Part II left this warning in miniature.

The threshold you derived on Day 2 is an instrument of policy

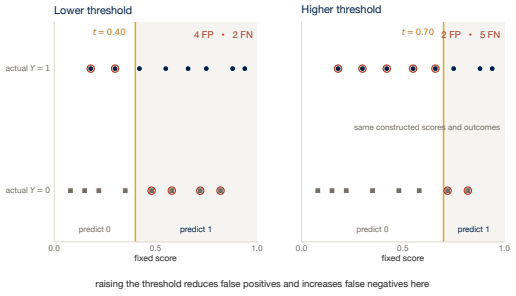
Recall the dial you built yourselves:

$$\tau = \frac{C_{\text{FP}}}{C_{\text{FP}} + C_{\text{FN}}},$$

the false-positive cost as a share of the two error costs combined

With the hypothetical costs of 4 and 80 it produced 0.048, and the model flagged many more clients.

Read it again with today's eyes:



Constructed scores and outcomes; the same file under two thresholds.

- whoever sets those two costs decides how many false positives exist and, with the group rates, who receives them
- that threshold is policy, and someone identifiable should own it and answer for it

Explanation has three jobs, each with its own customer

"The model should be explainable" means three distinct things. Keep them apart, because a system can deliver one and fail the others.

- **Debugging:** the builder asks why the model behaves as it does, to find leaks and errors. Customer: the team.
- **Case explanation:** one scored person asks, why me, and what would change the outcome? Customer: the person.
- **Accountability:** the institution must show its decision process to an auditor, a court, a regulator. Customer: the public.

The case we close with failed the last two.

Monitoring is the chain's last link, and Google Flu Trends shows why

Recall Day 1's parable in one line: Google Flu Trends looked excellent on the data it was built from, then overestimated flu in 100 of 108 weeks. What failed has a name.

- **Drift**, in plain words: the world moves away from the data the model was fitted on.
- The only defence is monitoring: re-check calibration and error rates on fresh outcomes, on a schedule, with a named person responsible.

$$\text{question} \to \text{data} \to \text{loss} \to \text{model} \to \text{evaluation} \to \text{decision} \to \boxed{\text{monitoring}}$$

Day 1 promised this link. It is now built.

A deployed model can write its own flattering data

Drift can also be self-inflicted, and the bank model you audited this week is one retraining away from it. Watch the loop, one step at a time:

1. the model selects who is acted on: the bank calls only the clients it scores high
2. only the selected are observed: a subscription can be recorded only for a client somebody called
3. tomorrow's data tilt towards the model's own choices: the next campaign's file fills with the model's favourites
4. retrained on that file, the model grows more confident in the same choices

- this is a **feedback loop**: the model's own decisions write tomorrow's data, and retraining on the same pipeline makes it worse, not better
- the fix is oversight of the loop itself; one classic instrument is to leave a small random sample of decisions untouched by the model

Hold the pattern. The case we close with ran this loop for years.

In the Netherlands, a risk score wrongly accused about 26,000 parents

Between 2005 and 2019, the Dutch tax authority used a self-learning risk-scoring **classifier** to flag possible childcare-benefit fraud. This was the supervised kind you met on Day 2, not a clustering method.

The tax authority wrongly accused around 26,000 parents and ordered them to repay large sums; estimates of those affected run to around 35,000 recipients. Families fell into debt; some lost homes and jobs. More than a thousand children, with some estimates running to 3,500, were taken into care. I give you that figure hedged because the public record contests it. When a number is contested, say so.

In January 2021 the entire Rutte cabinet resigned over the affair.

Before the next slide, take one minute. Write down every failure you can name in what I just described, using this week's vocabulary.

Every failure in that system has a name you learned this week

Check your list against the table. The mechanism of the discrimination: the tax authority used nationality as a proxy for risk.

what happened	the week's name for it
nationality entered the score as a proxy	Day 1's proxy trap: postcode standing in for income
the false positives fell on one group	today's group-wise error rates, never computed
no explanation, no effective appeal	the accountability job of explanation, failed
enforcement outcomes fed the training data	today's feedback loop
a risk score was treated as a finding of fraud	a score became a verdict: Day 2's line, crossed

Nothing here required new mathematics to catch. It required someone to ask this week's questions before acting.

Institutions now write these duties down: three names to know

I name these so you recognise the vocabulary; teaching the law is another course.

- **The EU AI Act** (Regulation 2024/1689): a risk-tier approach. The higher the risk of the use, the heavier the obligations on documentation, oversight, and monitoring.
- **GDPR Article 22:** rights for people subject to decisions made solely by automated means with significant effects.
- **The model card:** the documentation habit. One short document stating a model's purpose, data, performance overall and by group, known limits, and monitoring plan. Appendix G lists the fields.

The direction of all three is the same: the questions of this part are becoming written duties with owners.

Part IV checkpoint: three things you can now do

You can now:

- attach a responsibility question to each of the week's technical choices, in one table
- compute error rates by group, and explain why the fairness criteria cannot all hold at once
- audit a deployment story for proxies, group rates, appeal, feedback, and monitoring, and name what the Dutch system lacked

That is the toolkit Day 1's Amazon slide asked for. What remains is the close of the week.

Where we are

1. Grouping without labels
2. The few directions that carry the variation
3. The same tools, one level up
4. What a deployed system owes the people it scores
5. Close
6. Appendix

Your opening guess: close to one was nearest, but incomplete

At the start of today you wrote down how many dominant uncorrelated directions describe the 32 prices.

One strong shared direction is real. A single direction carries 42.0% of the stocks' variation, and its twin carries 80.2% of the yield curve's.

Now price the three options:

- about 32: too many for a useful low-dimensional account; the 0.389 average correlation shows substantial shared movement
- about 7: confuses external sector categories with uncorrelated PCA directions
- close to 1: nearest the dominant-direction answer, but 42.0% is not the whole system

> **One is the dominant direction, not a complete description.**

The grouping answer remains unresolved

The clustering evidence answers a different question:

- candidate silhouettes from 0.103 to 0.131 show weak separation
- at the working $k = 6$, sector ARI 0.553 shows moderate alignment with an external taxonomy
- period ARI 0.749 is one chronological sensitivity check, not a guarantee of stability

Six is a useful resolution for examining the sample, not a discovered natural count. Check your paper: the opening question asked for directions, whereas K-means supplied proposed groups.

> **Dimension and grouping are different questions, even on the same matrix.**

The week ends with an honest inventory

Depth	Models and methods
fit / evaluate	linear and logistic regression; decision tree; random forest
calculate / operate	baselines; rule-based text system; K-means; PCA; BM25; TF-IDF/LSA and cosine retrieval
interpret only	gradient boosting; learned embeddings; optional zero-shot classification; retrieve-then-generate systems

A rule-based classifier is not machine learning; PCA is not a predictive model; RAG is not one model. Accurate names keep the claim proportional to what you actually did.

You can now fit or operate a compact set of methods and audit a wider set without claiming four-day mastery.

The week's one thread runs through all four days

- **Day 1:** prediction is not causation. A model describes the file it was fitted on.
- **Day 2:** a coefficient is not a cause, and a probability is not a decision until costs enter.
- **Day 3:** a good score is not skill until the audit passes: valid features, a split that imitates deployment.
- **Day 4:** predicting risk is not the same as acting on people, and acting carries duties.

$$\text{question} \to \text{data} \to \text{loss} \to \text{model} \to \text{evaluation} \to \text{decision} \to \text{monitoring}$$

Any result you meet after this course can be walked along this chain, link by link.

> **One discipline underneath: say exactly what the evidence shows, and no more.**

Friday asks you to demonstrate the week's outcomes

The assessment takes place on Friday. It asks you to show the week's outcomes on concrete material:

- frame a task as a contract
- name the loss and the baseline
- audit a score for leakage and split design
- price a decision with costs
- judge a deployment for fairness and monitoring

Prepare by rehearsing. The five questions two slides ahead are the rehearsal set. For each day, be able to state its one thread sentence and reproduce its central numbers from the slides.

Good answers this week have all looked the same way: a claim, its evidence, and its scope, in that order.

Two writing habits carry you through Friday and beyond

The deployment conclusion, three sentences. I would use this model only for this decision, at this threshold, because of this evidence. I would not use it for that purpose, because of that risk. I would monitor these numbers, on this schedule.

The strong AI disclosure. Weak: "I used an AI tool." Strong: say what the tool did, what you checked and changed, and what responsibility you keep. The strong form is the one this course expects, on Friday and afterwards.

Both habits force the same three ingredients: claim, evidence, scope.

Five questions rehearse Friday, and every answer is on the slides

1. Why is the K-means centre a mean? Answer in one sentence, using what Day 2 proved.
2. Read a loading pattern: what exactly makes PC2 of the yield curve "slope"?
3. Your clustering scores a silhouette of 0.131, sector ARI 0.553, and period ARI 0.749. What different question does each answer?
4. Name the proxy mechanism in the Dutch case, and its Day 1 twin.
5. Write the three-sentence deployment conclusion for the bank model this week audited.

Write your answers tonight. On Friday, they are the shape of the answers that score well.

The week ends on a symmetry worth keeping

The yield curve hid one benign factor: the level, moving every maturity together, harmless and useful once seen.

A benefits system hid one harmful proxy: nationality, moving the false positives onto one group of families.

The mathematics of hidden structure was the same in both rooms. The difference was whether anyone checked what the hidden thing was doing before acting on people. That check, in all its forms, is what this week trained you to do.

Keep the discipline; the models will change

The methods you met this week will age. The discipline will not:

- a written contract before anyone fits anything
- a baseline before any praise
- a judge the model has never met
- a decision priced in costs, owned by a person
- a monitored life after deployment, with group rates on the dashboard

You arrived able to use these systems. You leave able to interrogate them. Thank you for four days of careful work; go and use it.

Where we are

1. Grouping without labels
2. The few directions that carry the variation
3. The same tools, one level up
4. What a deployed system owes the people it scores
5. Close
6. Appendix

Appendix A: the K-means objective, in full algebra

These pages are optional; nothing on Friday requires them.

For a fixed cluster C_j, the centre $\boldsymbol{\mu}$ minimising $\sum_{i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$ is found by differentiating and setting to zero: $\boldsymbol{\mu} = \frac{1}{|C_j|} \sum_{i \in C_j} \mathbf{x}_i$, the group mean, exactly Day 2's argument, coordinate by coordinate.

The total spread also decomposes:

$$\sum_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \underbrace{\sum_j \sum_{i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2}_{\text{within clusters}} + \underbrace{\sum_j n_j \|\boldsymbol{\mu}_j - \bar{\mathbf{x}}\|^2}_{\text{between clusters}} .$$

$\bar{\mathbf{x}}$: the overall mean · n_j: the size of cluster j

The left side is fixed by the data, so minimising the within-cluster term is the same as maximising the between-cluster term: tight groups, far apart.

Appendix B: choosing k without external reference labels

Two standard aids, best read together:

- **The elbow:** plot the total within-cluster squared distance against k. It always falls as k grows; look for the bend where further clusters stop paying.
- **The silhouette:** compute the average $s(i)$ for each candidate k and prefer higher values, while remembering the main deck's warning about chasing the score.

Neither replaces judgement about the use. A marketing team that can run three campaigns has a reason for $k = 3$ that no curve can supply.

Appendix C: the adjusted Rand index, formally

Consider all pairs of cases. Two labellings agree on a pair when both put the pair together, or both put it apart. The Rand index is the share of agreeing pairs. Random labellings score high on it by accident, so the adjusted version subtracts the chance level:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]},$$

RI: the share of agreeing pairs · $\mathbb{E}[\mathrm{RI}]$: its expectation over random labellings with the same group sizes

ARI is 0 on average for random labelling and 1 for perfect agreement, which is the gloss the main deck used for 0.553.

Appendix D: the PCA derivation, sketched

With centred data and sample covariance matrix S, the first direction solves:

$$\max_{\mathbf{w}} \ \mathbf{w}^\top S\,\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{w} = 1.$$

Introduce a multiplier λ for the constraint and differentiate: this gives $S\mathbf{w} = \lambda\mathbf{w}$. The candidates are the eigenvectors of S. Each candidate achieves variance equal to its λ, so the maximiser is the eigenvector with the largest eigenvalue. Later components repeat the argument at right angles to the earlier ones.

In practice software obtains the same quantities through the singular value decomposition of the data matrix, which is numerically more stable than forming S.

Appendix E: reconstruction from the first r components

Collect the first r loading vectors as columns of V_r and the scores as $Z_r = XV_r$. The rank-r reconstruction of the centred data is:

$$\widehat{X}_r = Z_r V_r^\top,$$

V_r: the first r loading vectors, as columns · Z_r: the corresponding scores

and the global share of variance it retains is the sum of the first r explained-variance ratios: 80.2% for $r = 1$ and 95.7% for $r = 3$ on the yield changes.

A single day's reconstruction quality is a different number, computed on that day's own centred values: 0.843 and 0.980 for 9 March 2020.

Appendix F: four fairness criteria

For groups g and a flagging rule, define selection rate, recall (true-positive rate), false-positive rate, and calibration by group.

criterion	requirement across groups
demographic parity	equal selection rates
equal opportunity	equal recall
predictive parity	equal precision
calibration by group	fitted probabilities match observed rates in each group

The impossibility statement: when the true outcome rate differs across groups, a rule cannot satisfy calibration by group and equal error rates at the same time. The only exceptions are degenerate cases, such as a perfect predictor.

Appendix F, continued: the impossibility in numbers

Feel it in ten seconds:

- Group A: 100 people, 50 truly positive; the rule flags 50 and catches 40.
- Group B: 100 people, 10 truly positive; the rule flags 10 and catches 8.

Precision is 0.8 in both groups and recall is 0.8 in both groups: two criteria match exactly. Now the false-positive rates. Group A: 10 false positives among 50 negatives, 0.20. Group B: 2 among 90, 0.022.

The innocent members of the higher-rate group bear nine times the false accusations, and no threshold removes the gap: it comes from the different sizes of the two innocent pools.

Choosing the criterion is therefore part of the decision itself. For a fraud accusation, where a false positive is catastrophic for the person, equal false-positive rates is the natural priority.

Appendix G: the model card's fields

A model card is one short document, kept with the model, answering:

- **Purpose and scope:** the decision it supports, and the uses it must not serve
- **Data:** provenance, population, time range, known gaps
- **Performance:** the audited metrics, overall and by group, with the split design stated
- **Limits:** leakage risks, drift expectations, groups with thin data
- **Monitoring:** which numbers are watched, how often, and who owns the model

Filling one in for the bank model takes an hour. Every part of it uses only this week's vocabulary.

Appendix H: three clustering alternatives in one page

- **Hierarchical clustering** merges the closest groups step by step, producing a tree; you cut the tree at the granularity you need, and no k is fixed in advance.
- **DBSCAN** grows clusters from dense regions and labels sparse points as noise; it finds non-spherical shapes that K-means cannot, at the price of two density settings.
- **Mixture models** treat each cluster as a probability distribution and return membership probabilities rather than hard assignments.

All three answer the same question as K-means with different assumptions about what a group is: compact balls, nested merges, dense regions, or overlapping distributions.

The judging discipline, separation, stability, recognisability, applies unchanged.

Appendix I: the evidence chain, link by link

Revision material for Friday. Each row names where the week built one link of the boxed evidence chain.

link	where you built it
question, data	Day 1: the prediction contract, the leakage tests
loss	Days 1 and 2: squared error, log loss, and their baselines
model	Days 2 and 3: lines, sigmoids, trees, ensembles; today, structure without labels
evaluation	Days 1 to 3: splits, cross-validation, one test audit
decision	Day 2: the cost threshold, and the uplift boundary
monitoring	today: drift, feedback loops, group rates, and a named owner

References: data and methods

- Stock returns: daily log-returns for 32 US-listed companies, 5 January 2016 to 30 December 2024, retrieved from Yahoo Finance via the `yfinance` package; sector labels assigned per company. A convenience teaching extract, as stated in Part I.
- Yield curve: US Treasury constant-maturity yields (FRED series DGS3MO to DGS30), daily, 2010 to 2026. Federal Reserve Bank of St. Louis, fred.stlouisfed.org.
- Litterman, R., and Scheinkman, J. (1991). "Common Factors Affecting Bond Returns." *Journal of Fixed Income*, 1(1), 54–61.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." arXiv:1301.3781.
- James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning*. Springer, ch. 12.

References: cases and governance

- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*, 366(6464), 447–453.
- Parlementaire ondervragingscommissie Kinderopvangtoeslag (2020). *Ongekend onrecht* (Unprecedented Injustice). Dutch House of Representatives; Amnesty International (2021). *Xenophobic Machines*.
- Regulation (EU) 2024/1689 (the EU AI Act); Regulation (EU) 2016/679 (GDPR), Article 22.
- Dastin, J. (2018). "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." Reuters, 10 October 2018. (Reused from Day 1.)
- 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. (Reused from Day 1.)