GIES MSBA WORKSHOP SERIES · WORKSHOP 5

ML Foundations II: Unsupervised Learning

Teach the research agent to find structure nobody labeled: group similar stocks, compress many columns into a few, and flag what doesn't fit.

Chapter 5 of the interactive textbook · dl-msba-workshop.azurewebsites.net/intro.html

BEFORE WE START

Scan to open today's chapter

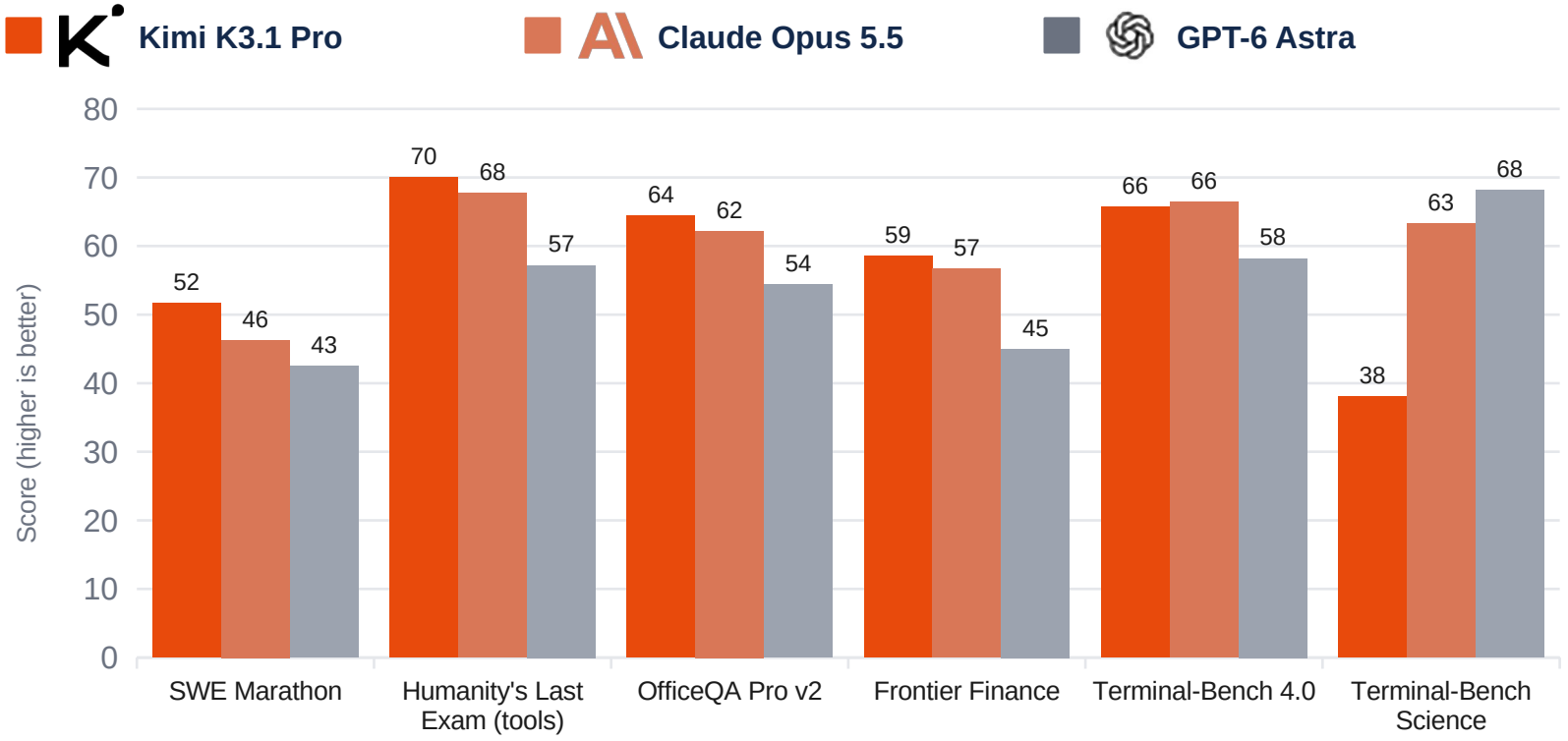


Chapter 5 · ML Foundations II

1. Scan the code with your phone camera, or open the textbook on your laptop.
2. Sign in with your @illinois.edu account (campus copy).
3. Keep it open: the notebook exercises at the end run there.

`dl-msba-workshop.azurewebsites.net/intro.html`

Kimi K3.1 Pro beats Opus 5.5 on 9 of 14 of its maker's benchmarks, but trails badly on science



WHY IT MATTERS FOR US

A Chinese lab is level with the frontier on agent and finance tasks.

More model choice for the research agent. Test before trusting: these are the vendor's own numbers.

Source: Moonshot AI benchmark chart for Kimi K3.1 Pro (vendor-run, all models at max thinking effort); Kimi blog, Oct 24, 2026 (Beijing time). Not yet independently verified.

BEFORE WE START · THURSDAY, OCT 22, 2026

00

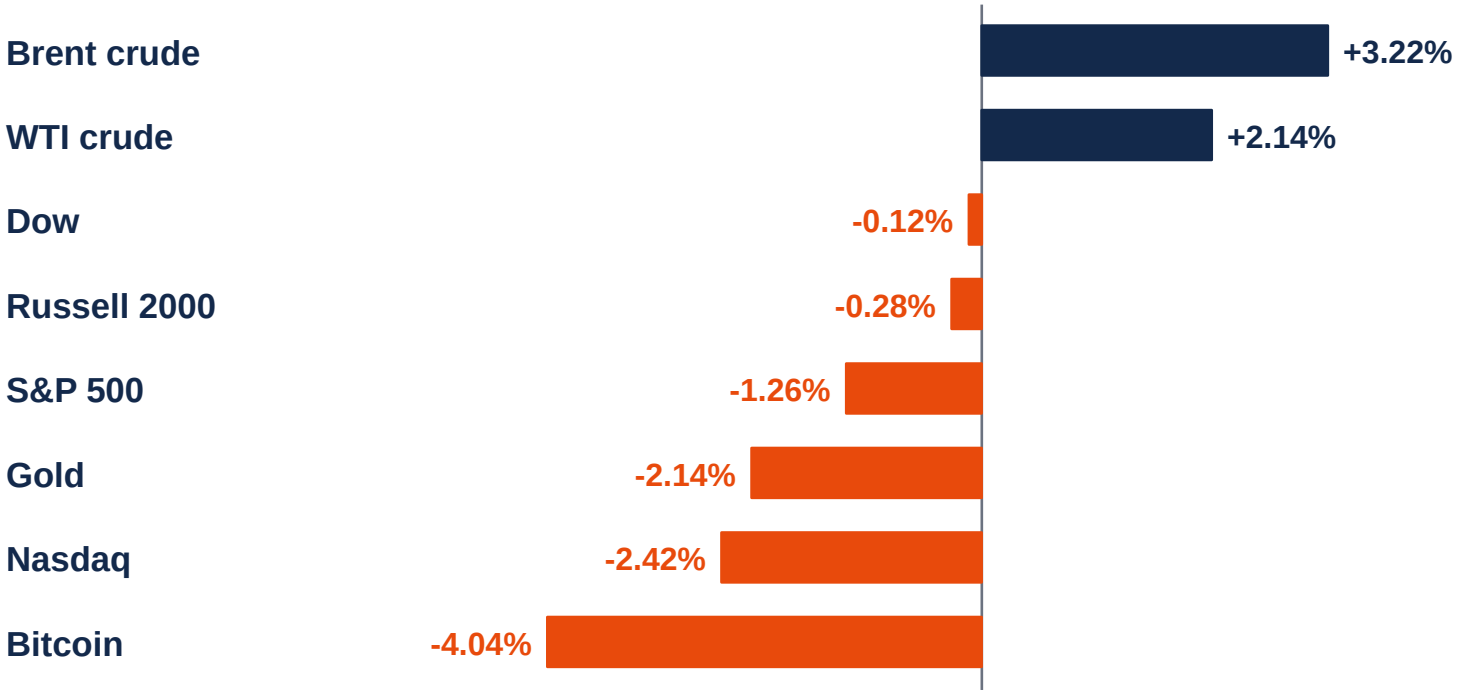
Market update

What moved today, and how today's tools would read it.

MARKET UPDATE · TODAY

Tech sold off, oil jumped and investors ran to safety: Nasdaq −2.4%, Brent +3.2%

% CHANGE TODAY



VIX (FEAR GAUGE)

19.11 +14.6% today

10-YR TREASURY YIELD

4.97% below 5% for the first time in a month

S&P 500 YEAR TO DATE

+13.3% through Oct 21 close

Intraday ≈1:30 pm ET, Yahoo Finance ticker board; 10-yr level via ^TNX; YTD from slickcharts (Oct 21 close). Numbers move; re-check before class.

Four stories behind today's moves

Nasdaq −2.4%

AI-spending worries hit tech

Meta −7.2% after it set up a new share class to fund AI spend; Lam Research −9% despite beating.

Brent +3.2%

Oil up on Middle East risk

More Houthi attacks in Saudi Arabia; Brent above $111, a five-month high.

VIX 19.1

Flight to safety

10-yr yield fell below 5%; VIX jumped; gold and bitcoin fell too.

Dow −0.1%

Rotation to defensives

Lockheed +6% to +9% after raising guidance; P&G, staples and health care up.

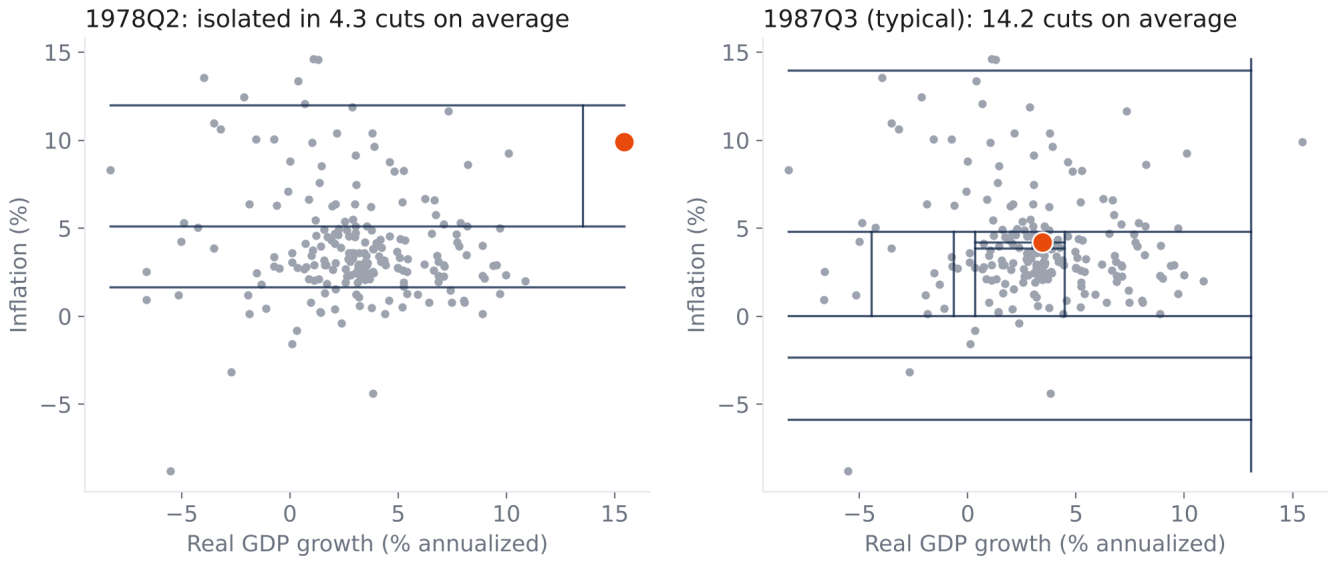
Next week: Fed decision Oct 27–28 (futures price an 87% chance of no change) · core PCE inflation Oct 29.

Sources: Reuters via Yahoo Finance (9:41 am ET), Yahoo ticker board (~1:30 pm ET), Benzinga premarket, TheStreet Pro, Investrade (Oct 22).

Each of today's questions maps to an unsupervised tool you'll learn in this session

Is this oil move normal, or a regime break?	→ **Anomaly detection**	Flags abnormal days with no definition of 'crisis'	Next slide
When has the market looked like this before?	→ **Nearest neighbors**	Finds the most similar past months by distance	After that
Which stocks are selling off together?	→ **Clustering**	Groups names by how they move, not by sector label	Section 2
Is it one big driver or many?	→ **PCA**	Shows if one factor (here: AI spend) explains most moves	Section 6

Isolation Forest finds the odd ones out by how few random cuts it takes to fence them off

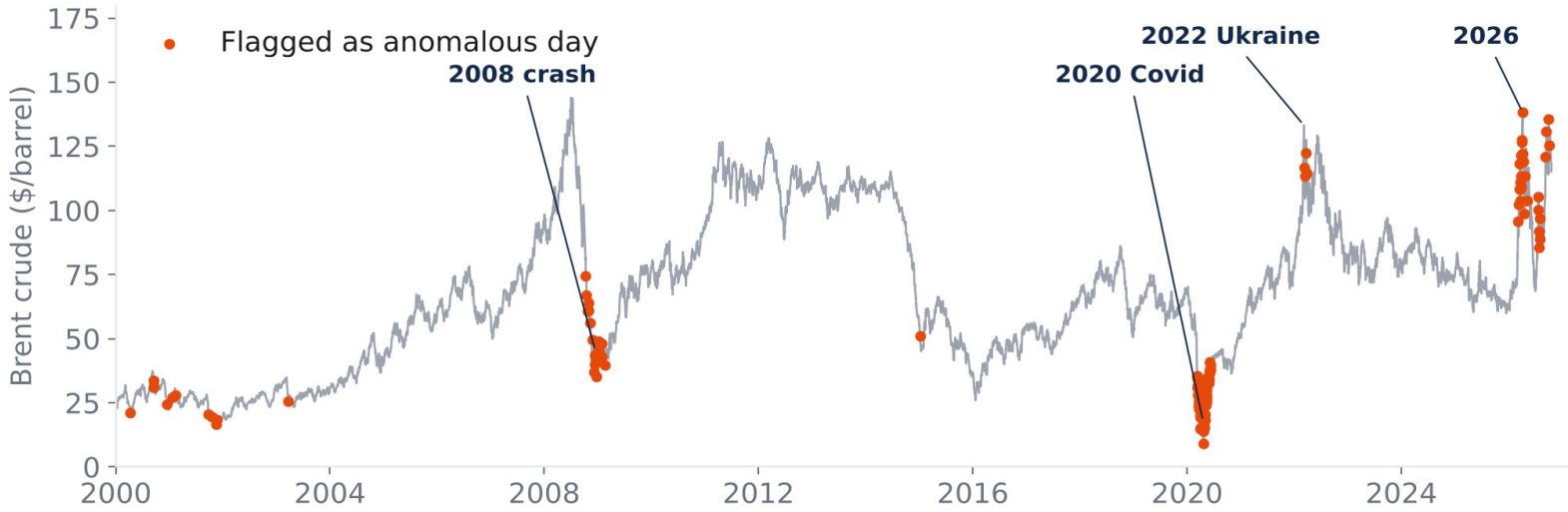


1. **Cut at random**
   Pick a column and a random threshold; keep the side your point is on.
2. **Count the cuts**
   Repeat until the point is alone. Odd points need few cuts.
3. **Average many trees**
   Hundreds of random runs give each point an oddness score.
4. **Flag the top few**
   You choose the share to flag (here 2% of days).

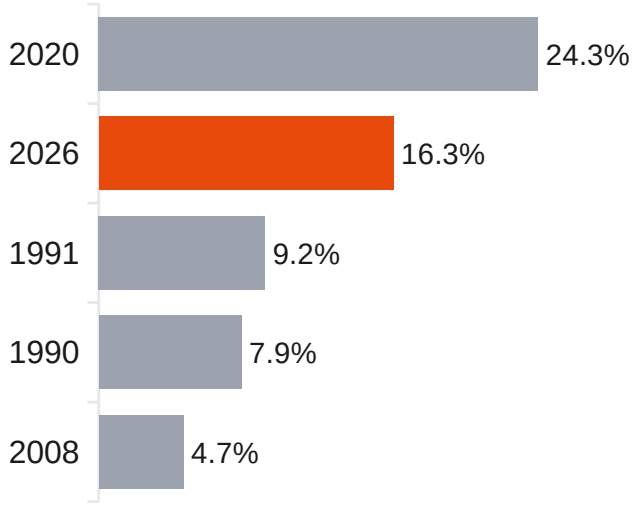
EVERYDAY EXAMPLE **Guess Who: “Wears a hat?” “Has a beard?” The one unusual face is found in two questions; a typical face takes many.**

Picture: real US quarters (GDP growth vs inflation). 1978Q2 isolated in 4.3 cuts on average; 1987Q3 needs 14.2.

Isolation Forest flags 16.3% of 2026's oil trading days as abnormal, second only to 2020 in 40 years



SHARE OF DAYS FLAGGED, TOP 5 YEARS



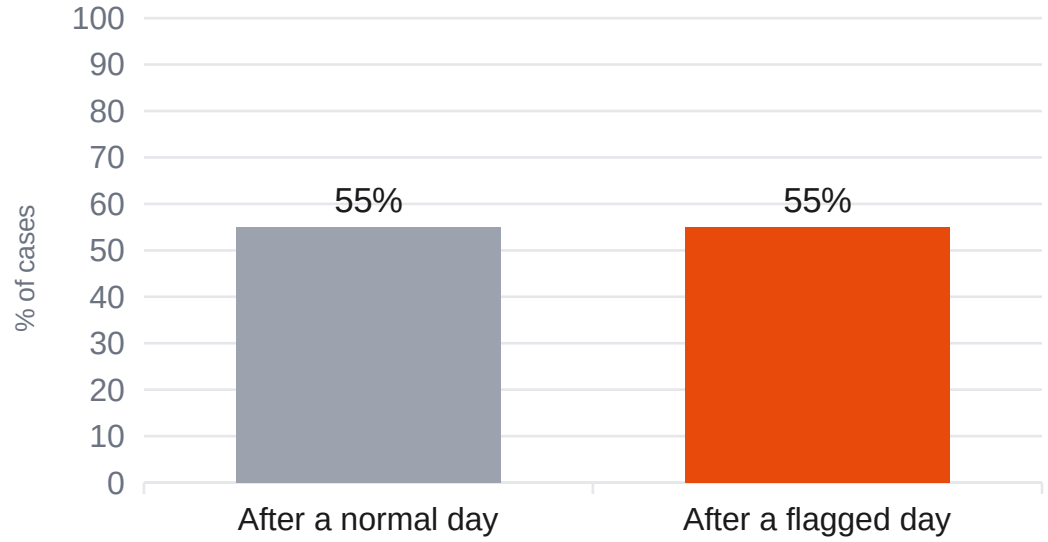
WHAT GOES IN: 3 NUMBERS PER DAY

1. **Today's change** — how far Brent moved that day
2. **20-day change** — the move over the past month
3. **20-day volatility** — how jumpy: now 79% vs 32% typical

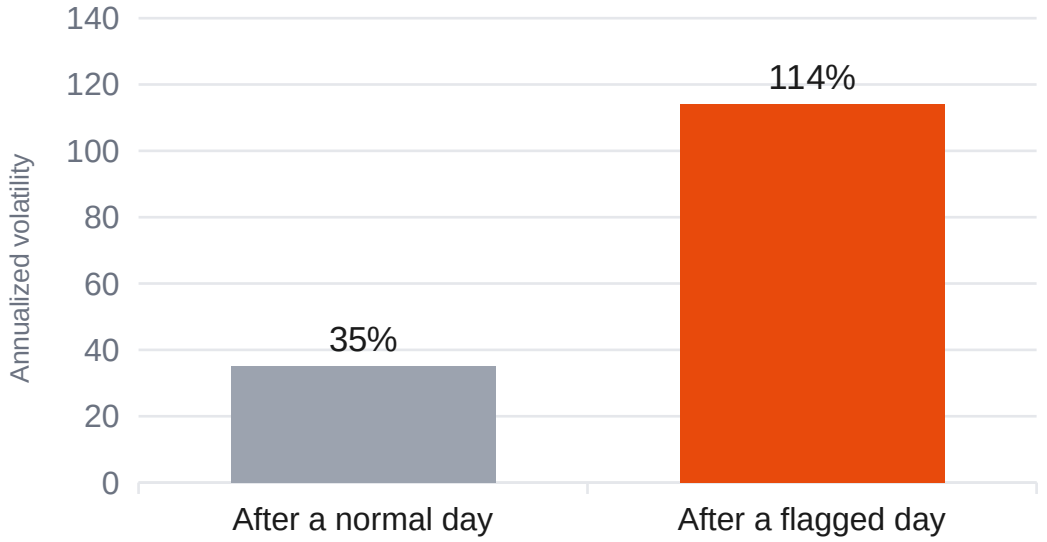
Computed: Isolation Forest (500 trees, 2% contamination) on daily Brent return, 20-day return and 20-day volatility, 1987–2026-10-20. Data: EIA via datasets/oil-prices (GitHub).

After a flagged oil day, the next move is far bigger, but no more likely to be up

Which way? Share of times oil was up 20 days later



How much? Volatility over the next 20 days



Use flags as a risk signal (size positions, hedge), not as a buy or sell signal.

Walk-forward: model refit each year on prior years only. Brent daily 1995–2026; 158 flagged vs 7,171 normal days. EIA via GitHub datasets.

Feed the anomaly score into last week's supervised model, next to valuation and momentum

INPUTS, EACH MONTH-END

Momentum	1-month and 12-month price change
Valuation	price vs its 5-year average
Brent–WTI spread	global vs US oil tightness
Volatility	how jumpy last month was
Anomaly score	from Isolation Forest (today's tool)

Regression

supervised, from last week (ridge: a line that resists overfitting)

PREDICTS NEXT MONTH'S

Brent return

which way, and how far

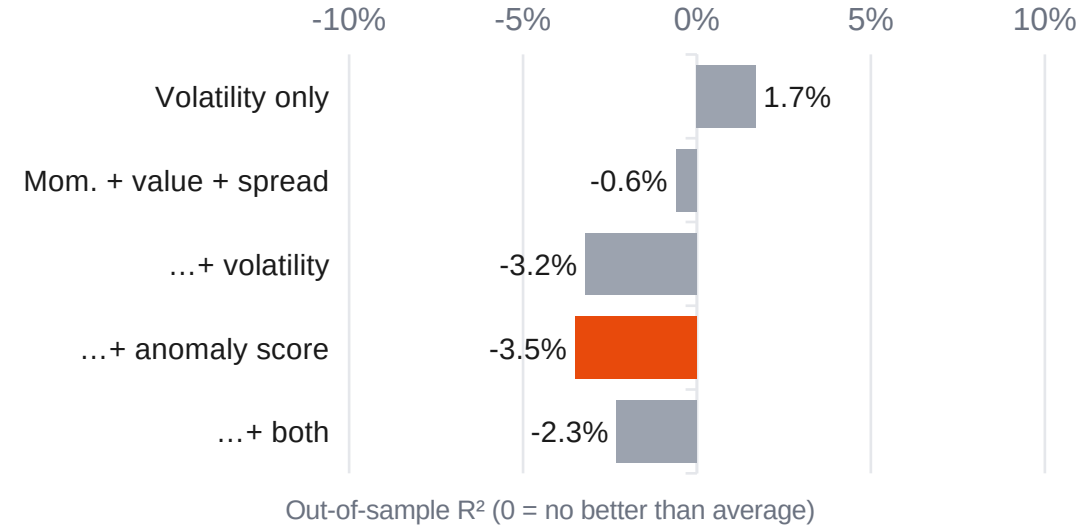
Brent volatility

how turbulent

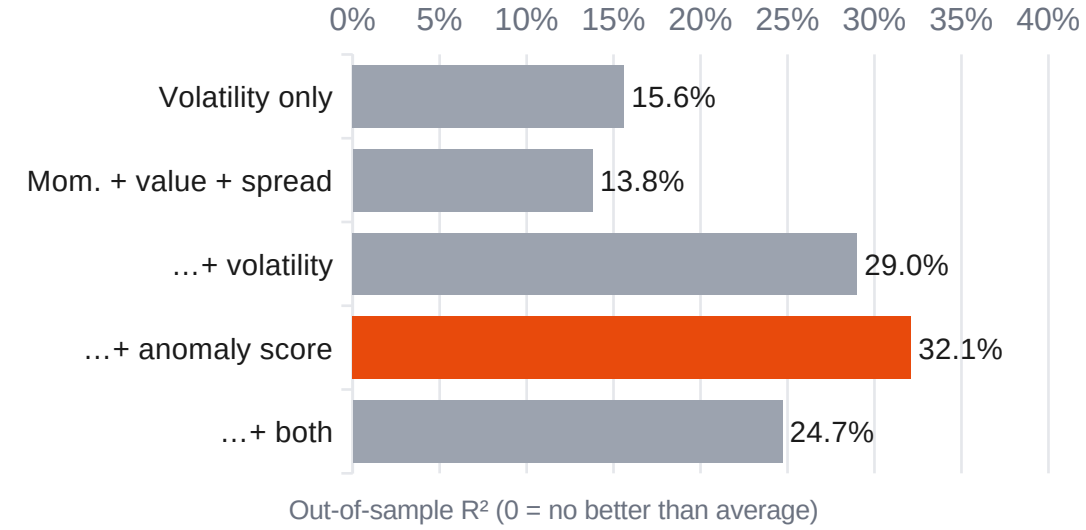
Walk-forward test: refit every month on past months only, scored on 261 months (2005–2026). Brent and WTI daily from EIA via GitHub datasets.

The anomaly score improves the risk forecast; nothing we tried predicts price direction

Next month's return (direction)



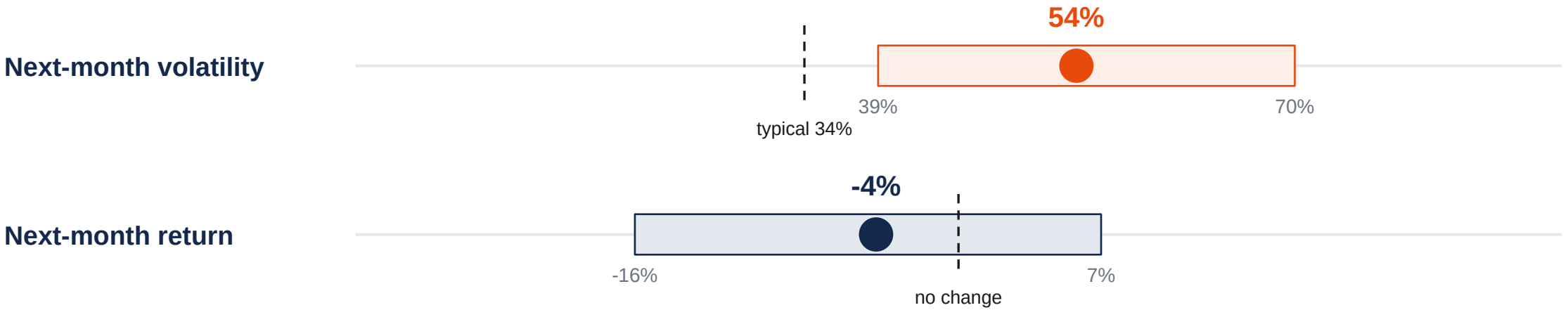
Next month's volatility (risk)



Best risk model swaps volatility for the anomaly score (32.1%). Using both double-counts: they're 0.77 correlated.

Walk-forward ridge regression, 261 monthly forecasts 2005–2026; R² vs the historical-average forecast.

Forecast for the next month: Brent volatility ≈54% vs 34% typical; direction is a coin flip



WHAT PUSHES THE RISK FORECAST ABOVE AVERAGE

+12.2 pts	+6.1 pts	-2.4 pts
Anomaly score (avg of month)	Anomaly days in month	Brent–WTI spread

Inputs as of 2026-10-20 (Brent $115.22). 80% bands from model residuals. Teaching example, not investment advice.

Nearest-neighbor search: the past months that look most like Sep 2026, and what stocks did next

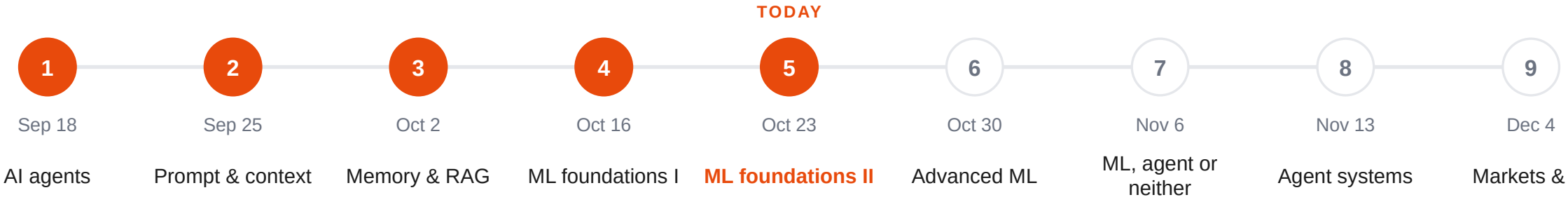
Month	S&P 500 3-mo	Brent 3-mo	Brent 1-mo	Gold 3-mo	VIX avg	S&P next 3 mo
Sep 2026 (now)	**+2.9%**	**+49.8%**	**+25.6%**	**+2.1%**	**15.8**	**?**
Sep 2004	-1.3%	+36.3%	+18.2%	+3.3%	14.1	**+7.0%**
Mar 2002	+0.8%	+27.0%	+20.1%	+6.3%	19.0	**-12.9%**
Apr 2015	+3.2%	+29.6%	+17.4%	-4.2%	13.5	**0.0%**
Aug 2005	+3.8%	+30.4%	+11.1%	+3.7%	13.0	**+1.1%**
Sep 1996	+0.9%	+22.0%	+14.1%	-0.5%	16.6	**+9.6%**

Same setup, very different outcomes (-12.9% to +9.6%): analogs frame the question; they don't answer it.

Computed: 5 nearest standardized months, 1990 to 2025, one per episode. Data: Shiller S&P 500, EIA Brent, gold, Cboe VIX (GitHub datasets). History, not a forecast.

WEEK BY WEEK

Each week the agent does more; today it finds patterns nobody labeled



Group, flag, apply to oil, and compress if time allows

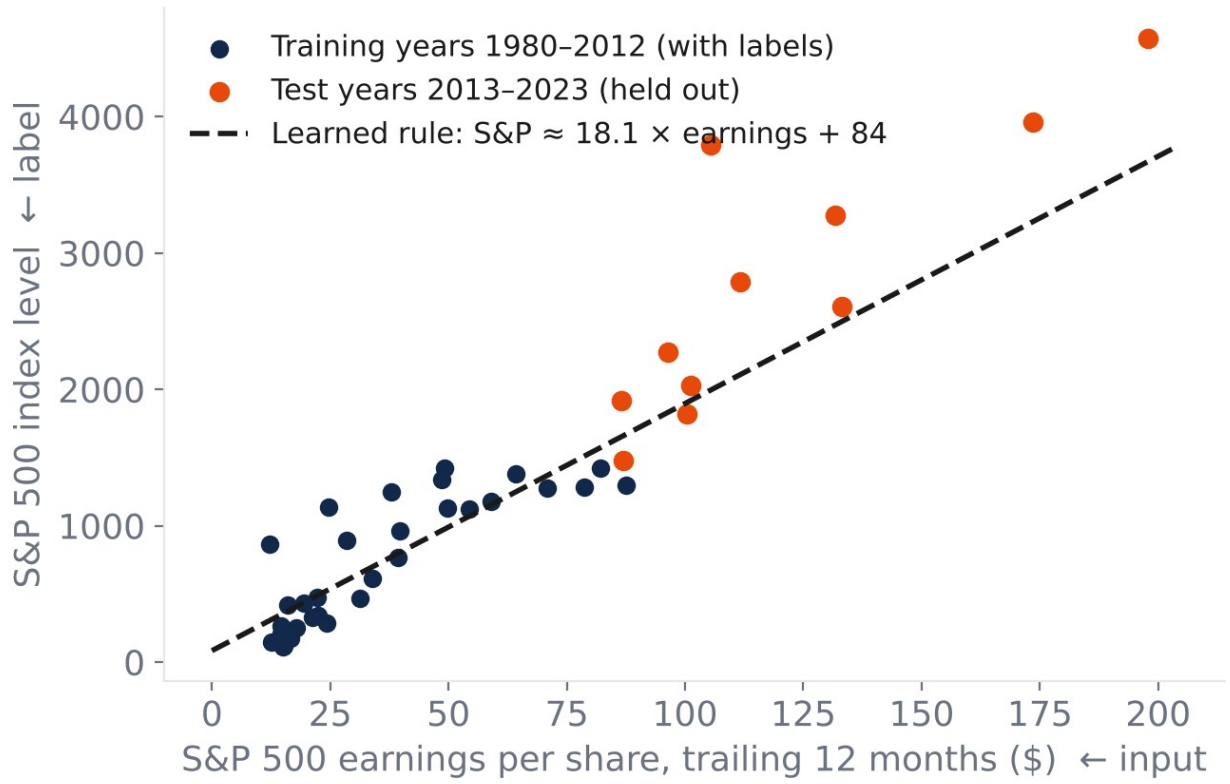
01	**Recap and background**	Last week, and its limits
02	**Clustering**	k-means and DBSCAN, by hand
03	**Anomaly detection**	Flag what doesn't fit: Isolation Forest
04	**At the firm**	Use cases and the research agent
05	**Case study: oil**	What moves oil, and what we can predict
06	**PCA (if time)**	Compress many columns: PCA, t-SNE, UMAP

01

Recap and background

Last week's supervised loop, where it runs out, and what changes without labels.

Supervised learning: show the model inputs with their answers, and it learns the rule

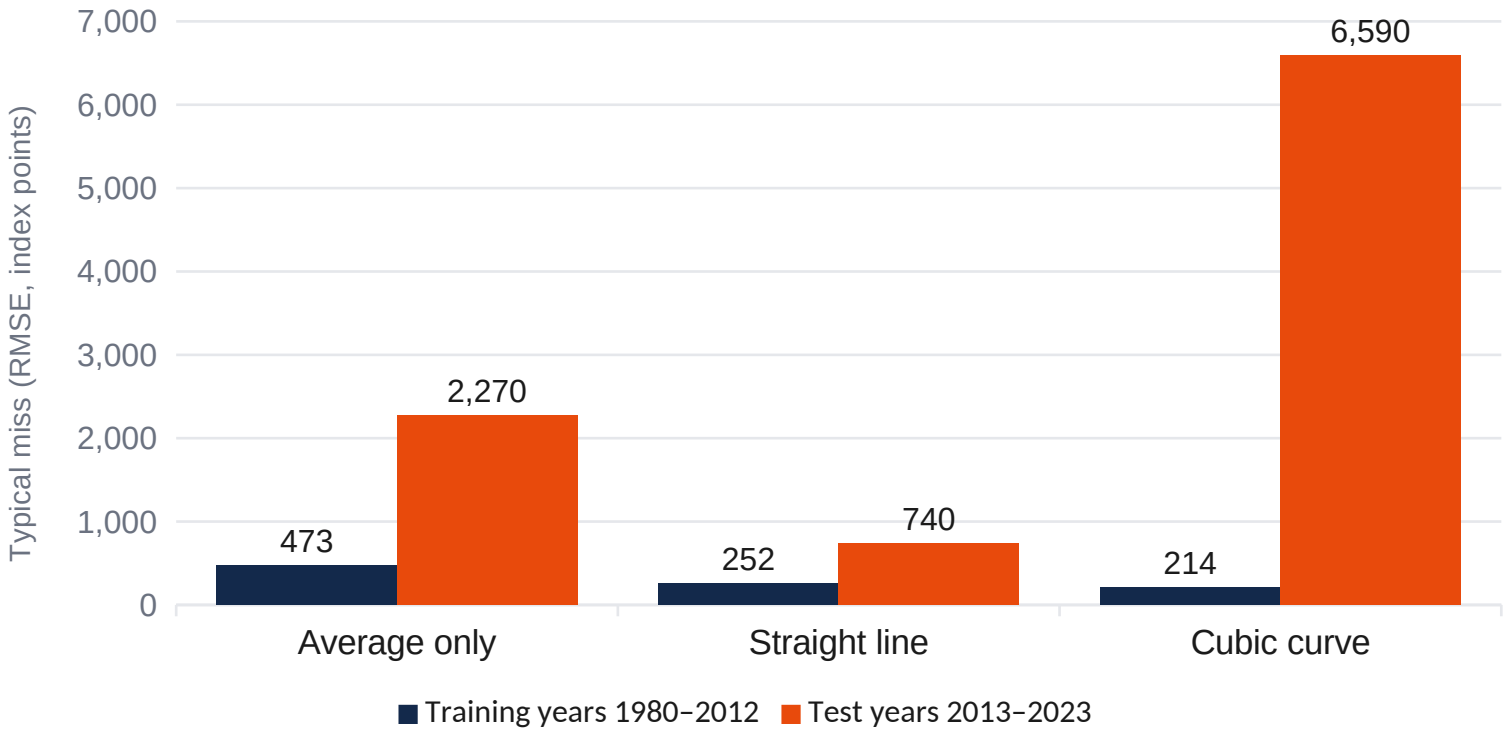


1. **Labeled examples**
   44 Januaries: earnings in, index level as the answer.
2. **Guess a rule**
   A straight line: index ≈ slope × earnings + intercept.
3. **Measure the miss**
   Loss: how far the guesses are from the answers.
4. **Adjust and repeat**
   Gradient descent nudges the line until the miss stops shrinking.

The key ingredient: a label column.

Data: Shiller S&P 500 dataset (github.com/datasets/s-and-p-500), January of each year 1980–2023.

Last week's lesson: the model that fit the past best forecast the future worst



WHAT IT SHOWED

Cubic: lowest training miss (214), worst test miss (6,590).

The straight line generalized best: 740 on years it never saw.

Grading on held-out labels is how supervised learning stays honest.

Recomputed for today from the Shiller data: models fitted on 1980–2012 Januaries, scored on 2013–2023.

WHERE SUPERVISED LEARNING RUNS OUT

Supervised learning only works when someone has already written down the answer

01

LABELS COST

Every row needs an answer

Nobody has tagged which of 500 stocks are true peers.

02

LABELS ARRIVE LATE

12 months

NBER declared the Dec 2007 recession start on Dec 1, 2008.

03

LABELS FIX THE QUESTION

Only finds what you named

It can't discover a market regime nobody has defined yet.

04

RARE EVENTS

Few examples to learn from

Fraud, defaults, crashes: too rare to label well.

Today: unsupervised learning starts from the inputs alone, no answer column needed.

NBER date: Business Cycle Dating Committee announcement, Dec 1, 2008 (reported by MPR News / AP).

Same table, one column missing: that column is the whole difference

SUPERVISED · LAST WEEK

Petal len.	Petal wid.	Sepal len.	Sepal wid.	Species (label)
1.4	0.2	5.1	3.5	setosa
3.5	1.0	5.0	2.0	versicolor
5.7	2.3	6.9	3.2	virginica
1.6	0.2	5.0	3.0	setosa
4.4	1.2	5.5	2.6	versicolor

Task: learn measurements → species, then grade against the answer.

UNSUPERVISED · TODAY

Petal len.	Petal wid.	Sepal len.	Sepal wid.	Species
1.4	0.2	5.1	3.5	?
3.5	1.0	5.0	2.0	?
5.7	2.3	6.9	3.2	?
1.6	0.2	5.0	3.0	?
4.4	1.2	5.5	2.6	?

Task: find which flowers belong together. No answer to grade against.

Rows: five real flowers from Fisher's Iris data (UCI / scikit-learn), measurements in cm.

DEFINITION

Unsupervised learning

Algorithms that find patterns in data without being told the right answer.

EVERYDAY EXAMPLE

Sorting a box of unlabeled family photos into piles by who and where.

Variables: faces in the photo, location, date taken

WHY IT MATTERS

Most real data is unlabeled. Labeling is slow and costly; this works on data as it is.

Unsupervised learning does three jobs: group, compress, flag

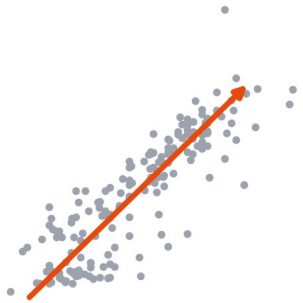


Group

CLUSTERING

Which rows belong together?

e.g. Stocks that trade alike → peer groups



Compress

DIMENSIONALITY REDUCTION

Can many columns become a few?

e.g. 50 risk measures → 3 factors



Flag

ANOMALY DETECTION

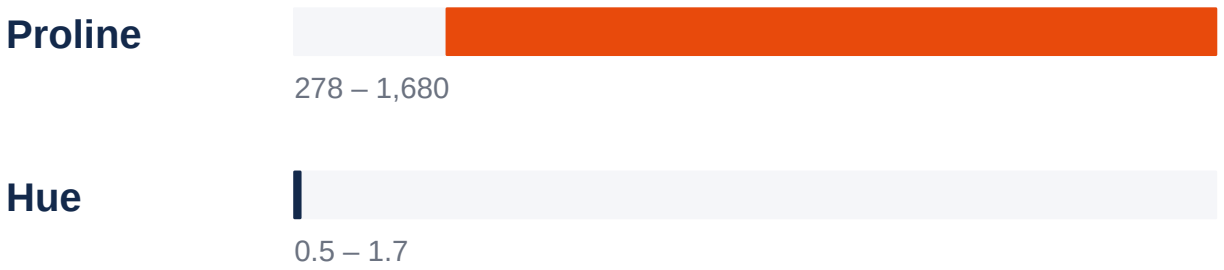
Which rows don't fit?

e.g. A quarter or a stock unlike the rest

BEFORE ANY ALGORITHM

Distance is everything today, so put columns on the same scale first

TWO WINE MEASUREMENTS, DRAWN TO THE SAME SCALE

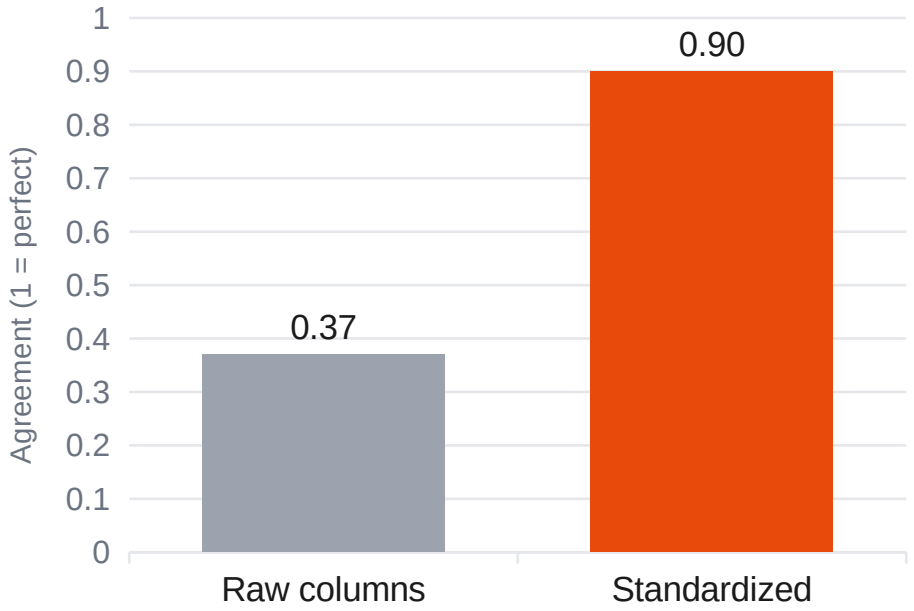


FIX: STANDARDIZE EACH COLUMN

`z = (value − column mean) ÷ column std. dev.`

Now every column gets a fair vote.

K-MEANS MATCH WITH THE 3 TRUE GRAPE VARIETIES



Computed: UCI Wine (178 wines, 13 columns), k-means k = 3; scikit-learn.

02

Clustering

Group similar rows: k-means, hierarchical, DBSCAN.

DEFINITION

k-means clustering

Split data into k groups so each point sits near its group's average point.

EVERYDAY EXAMPLE

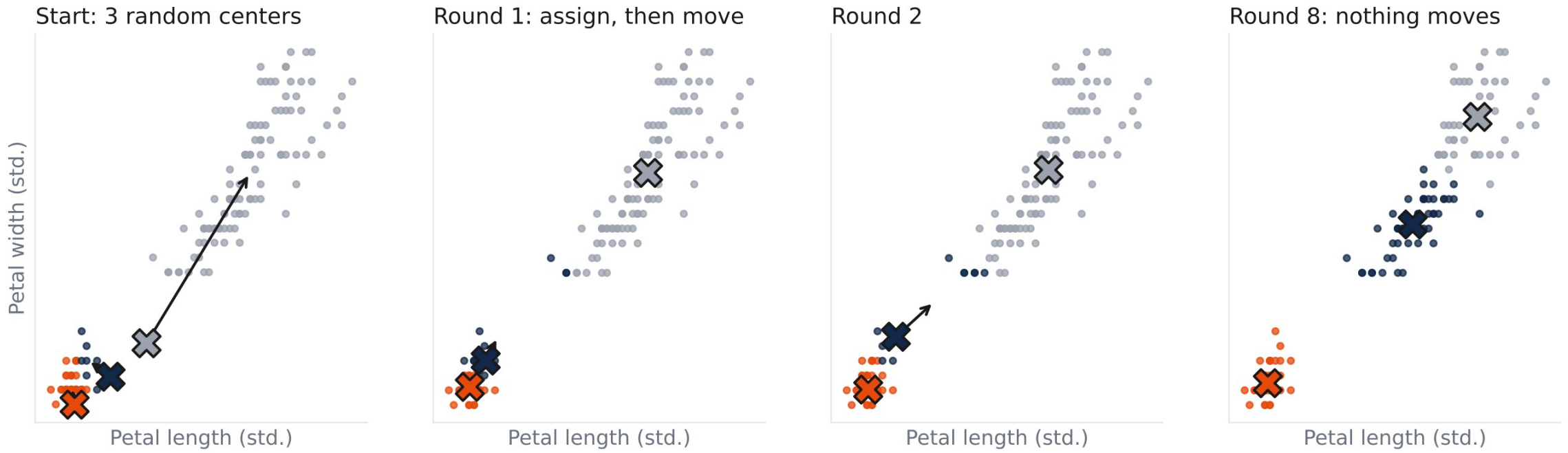
A TA forms 3 study groups so students in each group perform alike.

Variables: hours studied, exam score; k = 3 groups

WHY IT MATTERS

Fast and simple. The default first try for customer, product or stock segments.

k-means repeats two moves, assign and update, until the centers stop moving

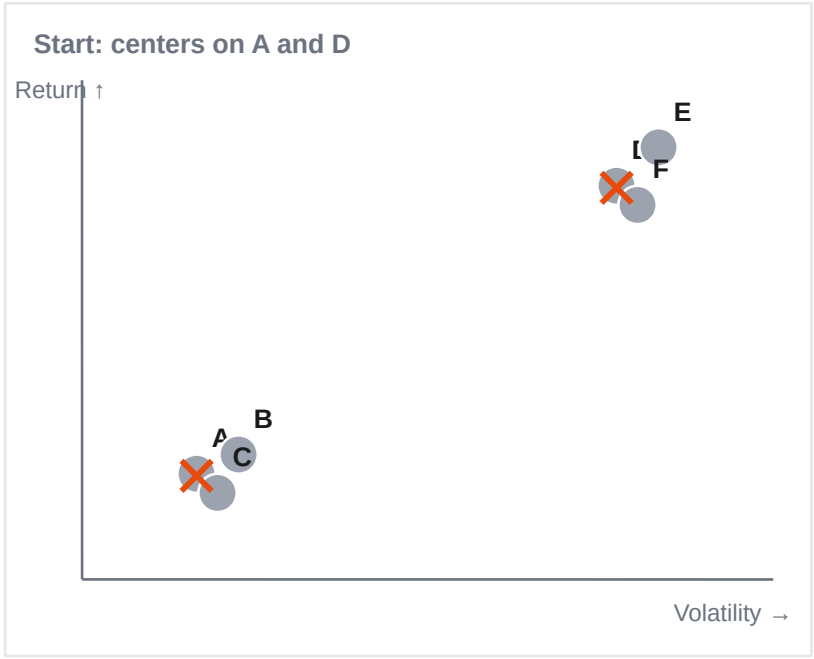


Assign each flower to its nearest ×, move each × to the middle of its flowers, repeat. Done in 8 rounds.

Computed: k-means on standardized Iris petal measurements (150 flowers); deliberately poor start.

k-means by hand: six stocks, two groups, start with two centers

Stock	Volatility	Return
A	10	5
B	12	6
C	11	4
D	30	20
E	32	22
F	31	19



Step 1 · Pick k and start

k = 2. Start centers on A (10, 5) and D (30, 20).

Step 2 · Measure distance

Straight-line distance:

$$\sqrt{((x_1-c_1)^2 + (x_2-c_2)^2)}$$

Illustrative numbers chosen to make the arithmetic easy; real use standardizes each column first.

Step 3: each stock joins the center it's closest to

WORKED: STOCK B (12, 6)

to center A (10, 5) $\sqrt{((12-10)^2 + (6-5)^2)} = \sqrt{5} \approx 2.24$

to center D (30, 20) $\sqrt{((12-30)^2 + (6-20)^2)} = \sqrt{520} \approx 22.80$

2.24 < 22.80 → B joins A's group

Result after one pass

Group 1: A, B, C (low risk, low return)
Group 2: D, E, F (high risk, high return)

Stock	Dist. to A	Dist. to D	Joins
A	**0.00**	25.00	**Group 1**
B	**2.24**	22.80	**Group 1**
C	**1.41**	24.84	**Group 1**
D	25.00	**0.00**	**Group 2**
E	27.80	**2.83**	**Group 2**
F	25.24	**1.41**	**Group 2**

Steps 4–5: move each center to its group's average, then repeat until nothing changes

STEP 4 · NEW CENTER = AVERAGE OF THE GROUP

Group 1 (A, B, C)

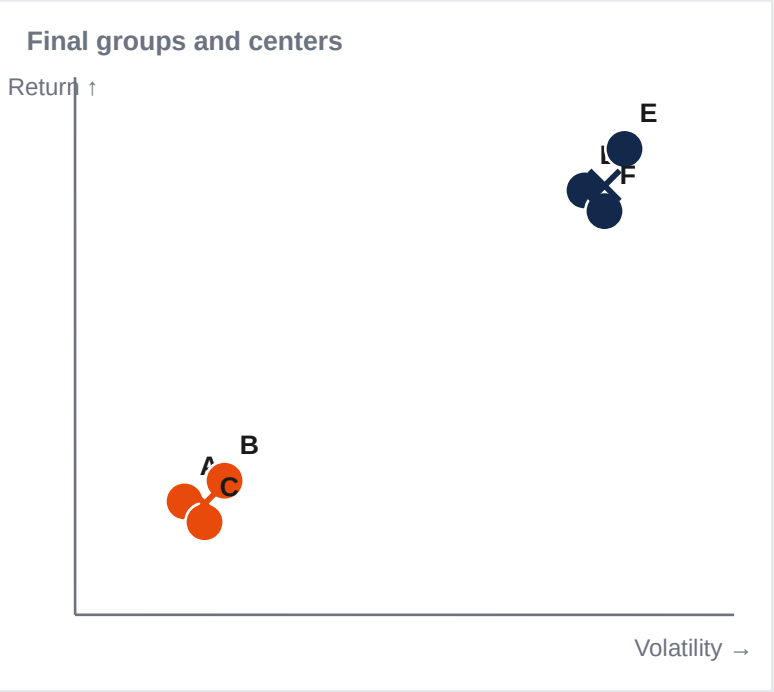
`vol (10+12+11)/3 = 11` `return (5+6+4)/3 = 5` → **(11, 5)**

Group 2 (D, E, F)

`vol (30+32+31)/3 = 31` `return (20+22+19)/3 = 20.3` → **(31, 20.3)**

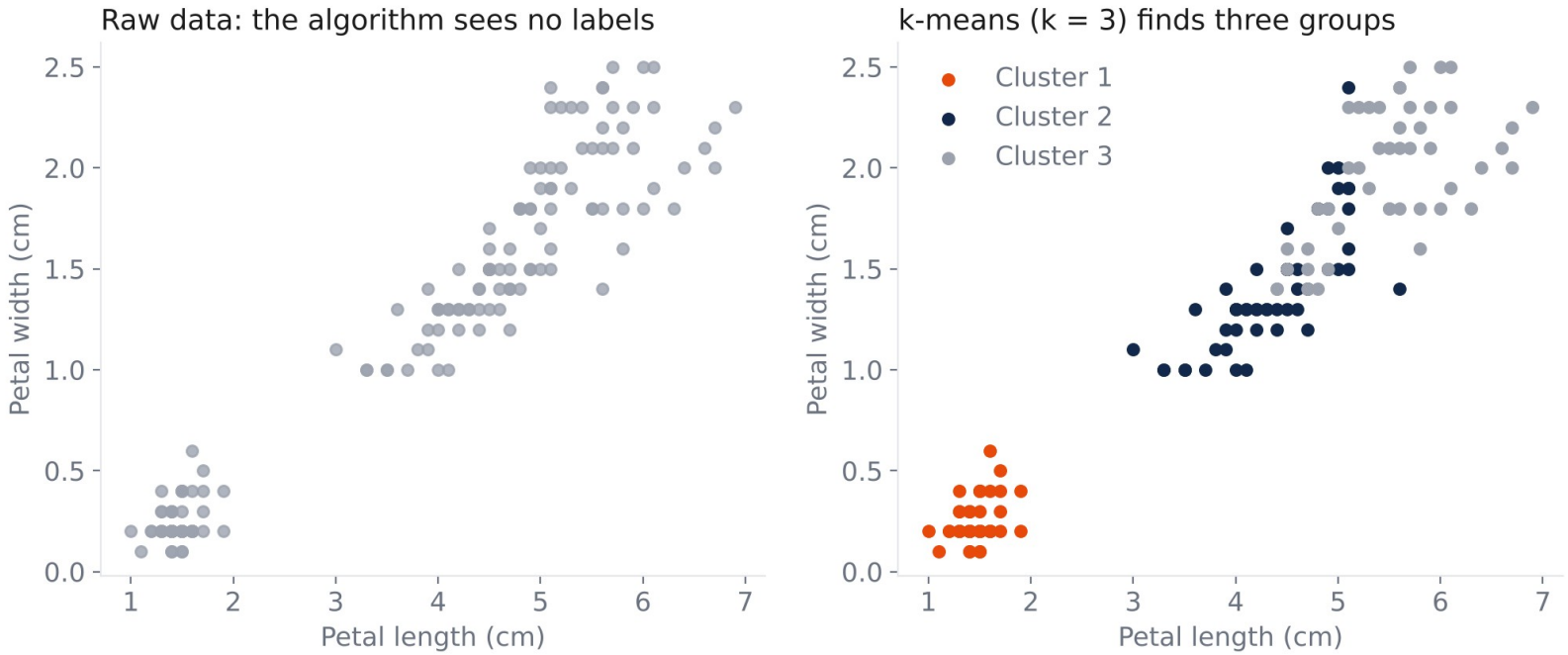
STEP 5 · REPEAT

Re-measure with the new centers: every stock is still closest to its own group's center, so nothing switches. Done.



Total squared distance to centers fell from 17 to 10.7: each round can only lower it, which is why k-means always stops.

Given only petal measurements, k-means rediscovers the three iris species



Left: what the algorithm sees. Right: the three groups it found.

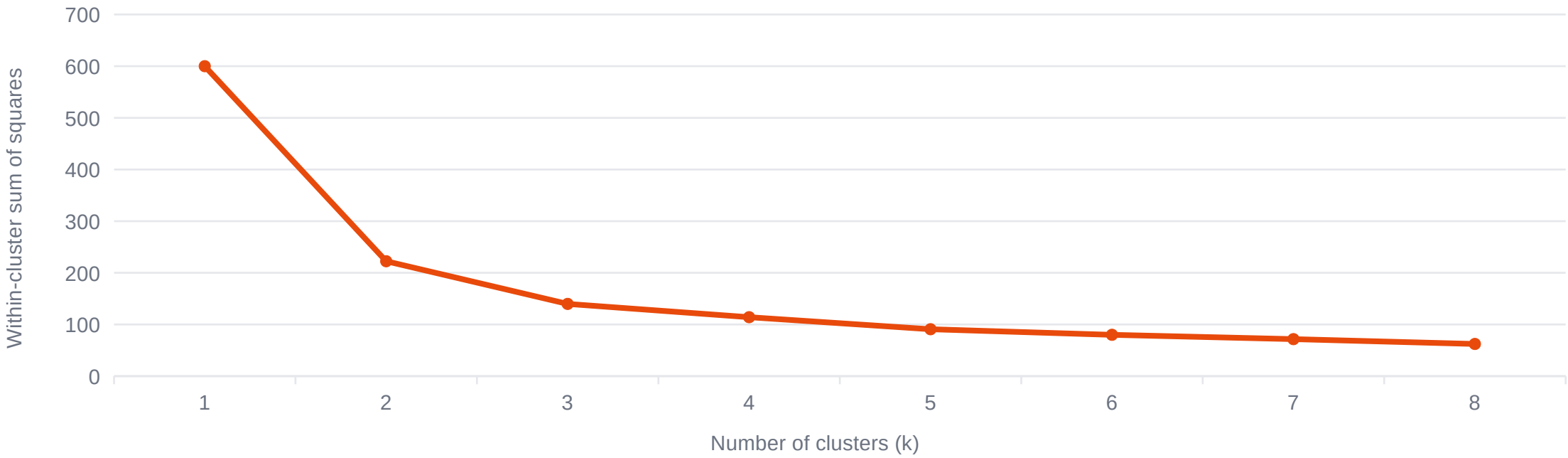
Data: Fisher's Iris (1936), 150 flowers, 4 measurements; UCI / scikit-learn. Plot shows 2 of 4 features.

83%

of flowers land in a cluster dominated by their own species, with no labels used.

Computed: k-means (k=3) on standardized Iris features; Adjusted Rand Index 0.62.

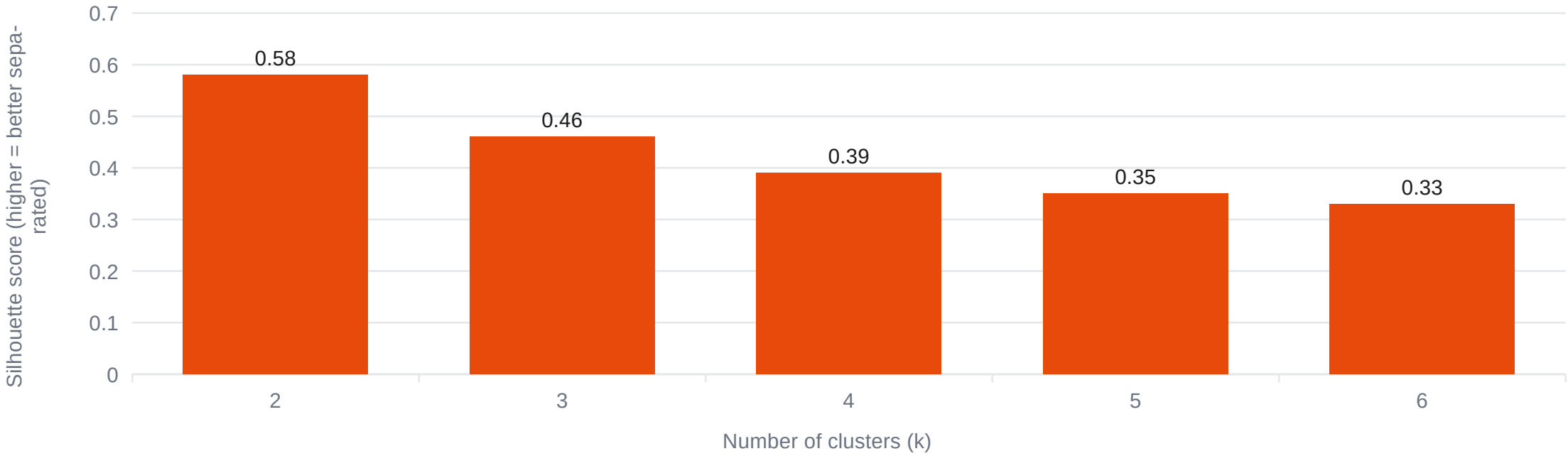
Adding clusters always lowers error, but gains flatten after k = 3



Look for the bend: past it, extra clusters buy little.

Computed: k-means on standardized Iris features (150 flowers), k = 1 to 8.

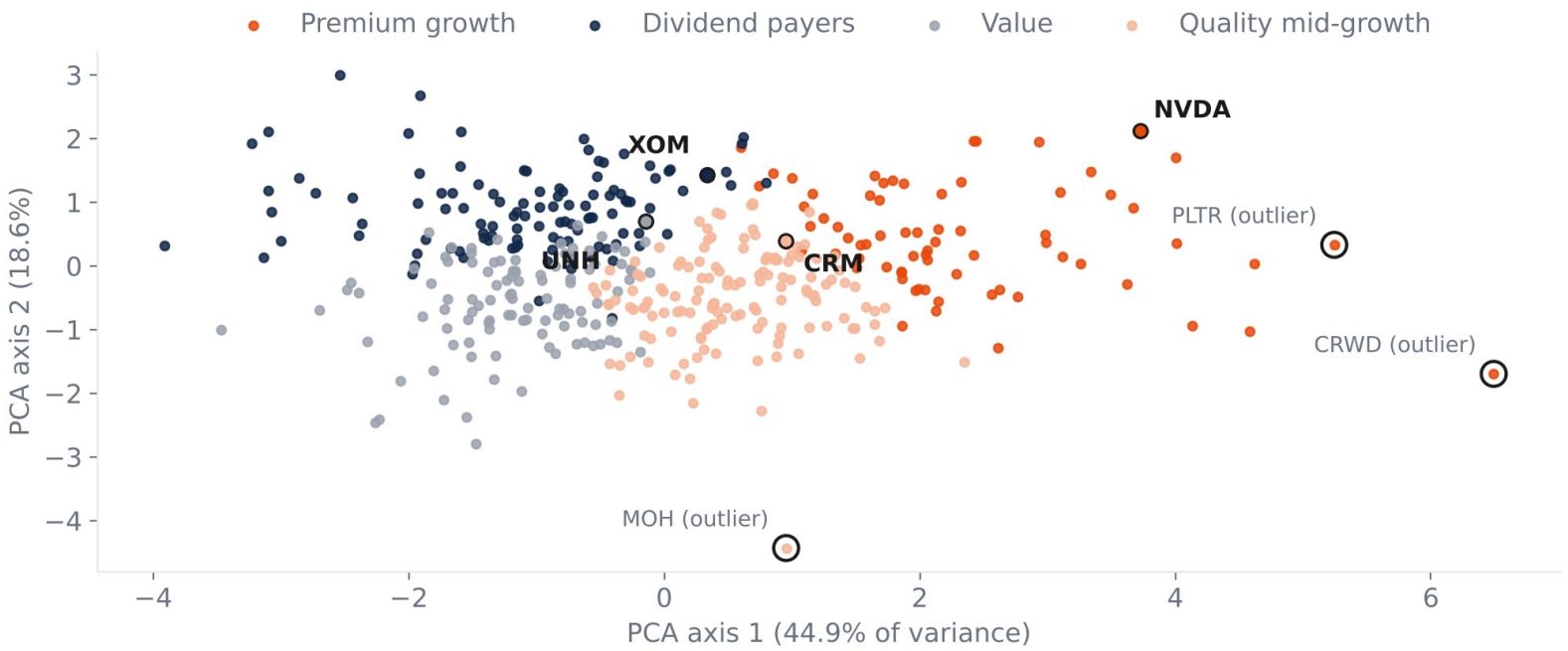
Silhouette prefers k = 2: metrics can disagree with the 'true' answer



Two species overlap, so the cleanest split is 2 groups, not 3.

Computed: k-means on standardized Iris features; silhouette from scikit-learn.

k-means on five valuation ratios splits 426 S&P 500 companies into four peer groups



Each dot is a company, laid out by PCA. Circled: the three most unusual names.

Data: S&P 500 constituent financials (Yahoo Finance snapshot via datasets/s-and-p-500-companies-financials, GitHub); not current prices.

Read each group from its typical member: four styles appear without sector labels

Premium growth

70 companies

Median P/E	38.3
Dividend yield	0.6%
Price / book	14.5
Price / sales	8.6

LARGEST

NVDA · AAPL · GOOGL · MSFT

Dividend payers

110 companies

Median P/E	19.2
Dividend yield	3.7%
Price / book	2
Price / sales	2.9

LARGEST

XOM · CVX · BAC · MS

Value

106 companies

Median P/E	16.2
Dividend yield	1.3%
Price / book	2.3
Price / sales	1.2

LARGEST

UNH · DIS · CB · MPC

Quality mid-growth

140 companies

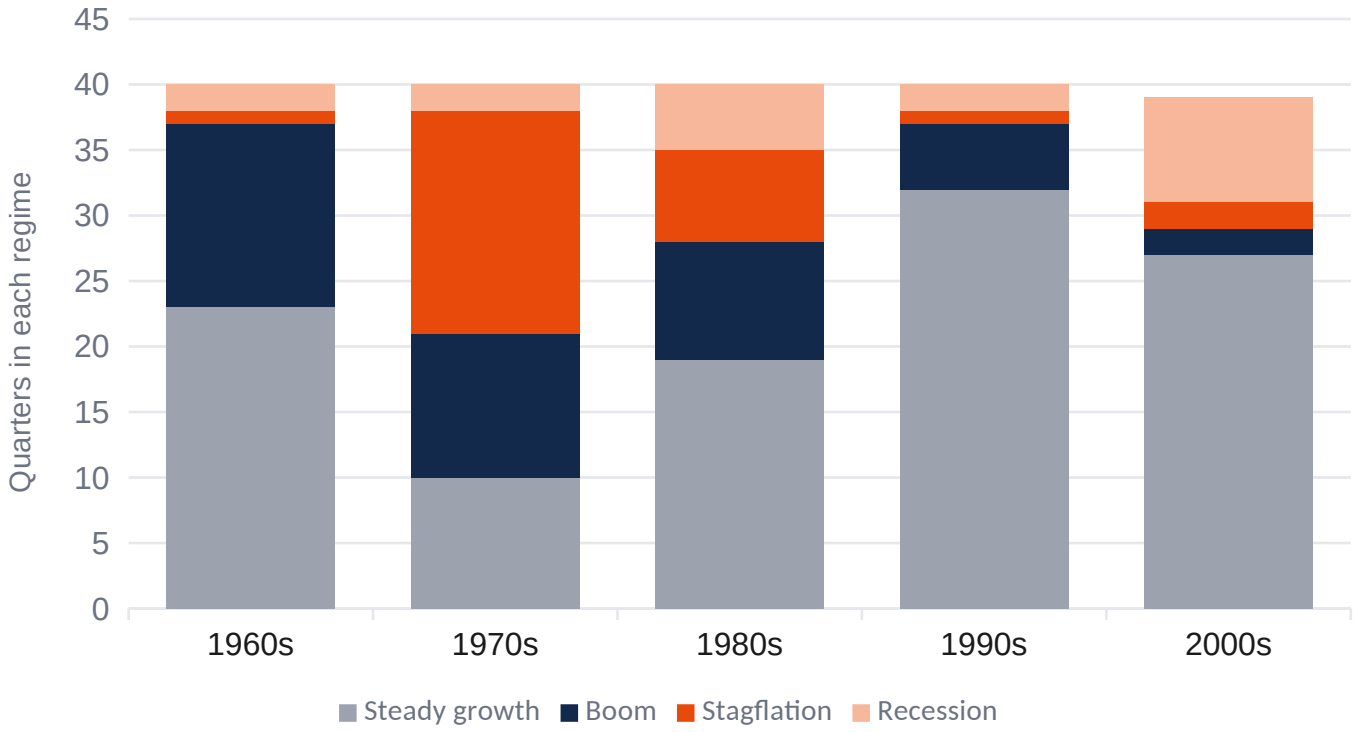
Median P/E	30
Dividend yield	1%
Price / book	5.2
Price / sales	4.1

LARGEST

RTX · AXP · CRM · QCOM

Medians per group. All 426 companies median: P/E 23.7, yield 1.4%. Source as previous slide.

Clustering 200 US quarters on growth and inflation finds four regimes; the 1970s stand out



Stagflation

Growth 1.1% · inflation 10% · 28 quarters

Boom

Growth 7.9% · inflation 3.6% · 42 quarters

Steady growth

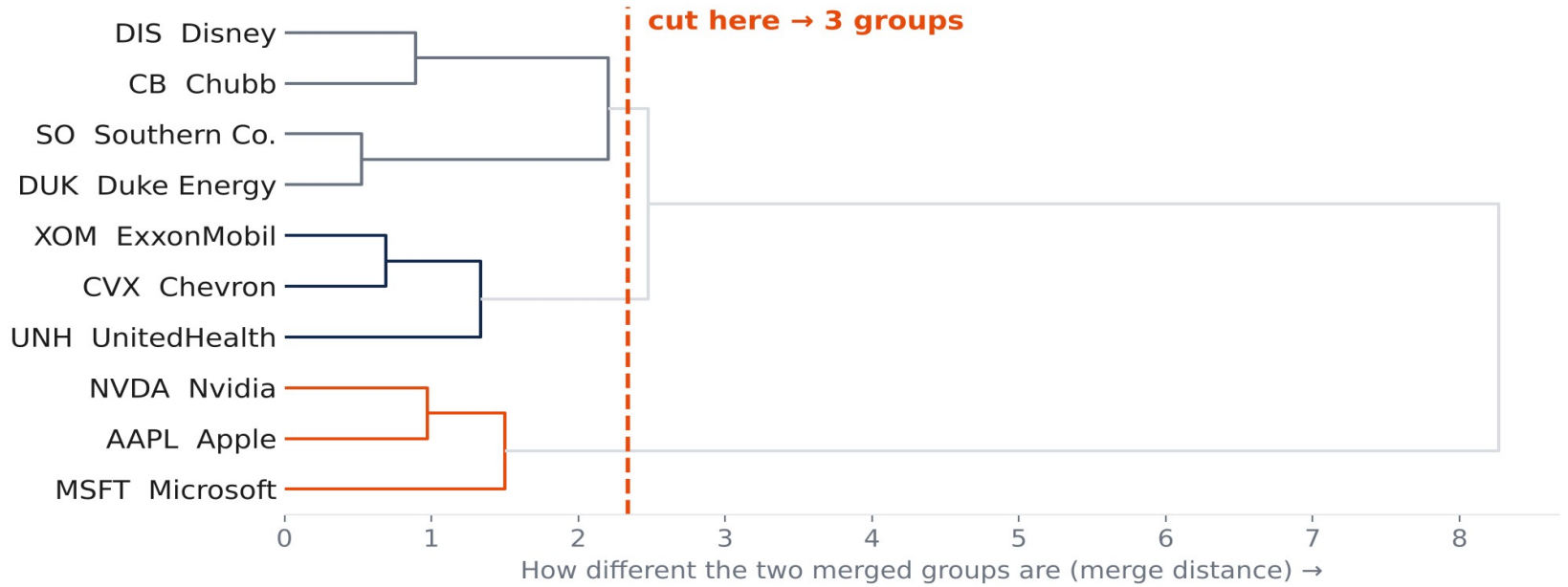
Growth 2.9% · inflation 3% · 112 quarters

Recession

Growth -3.1% · inflation 1.6% · 20 quarters

Computed: k-means (k = 4) on standardized real GDP growth and inflation, US 1959Q2–2009Q3 (statsmodels macrodata). 1959 omitted from the chart.

Hierarchical clustering merges the two most similar companies, again and again, into a tree



1 Start
Every company is its own group.

2 Merge
Join the two closest: Duke + Southern first.

3 Repeat
Keep merging. Longer lines = less alike.

4 Cut
Slice the tree where you want: here, 3 groups.

Tech giants · oil majors + UnitedHealth · utilities, Chubb and Disney. No sector labels used.

Computed: Ward linkage on 5 standardized valuation ratios; S&P 500 Yahoo Finance snapshot (GitHub datasets).

DEFINITION

DBSCAN

Groups points that sit in dense crowds; points in sparse areas are labeled noise.

EVERYDAY EXAMPLE

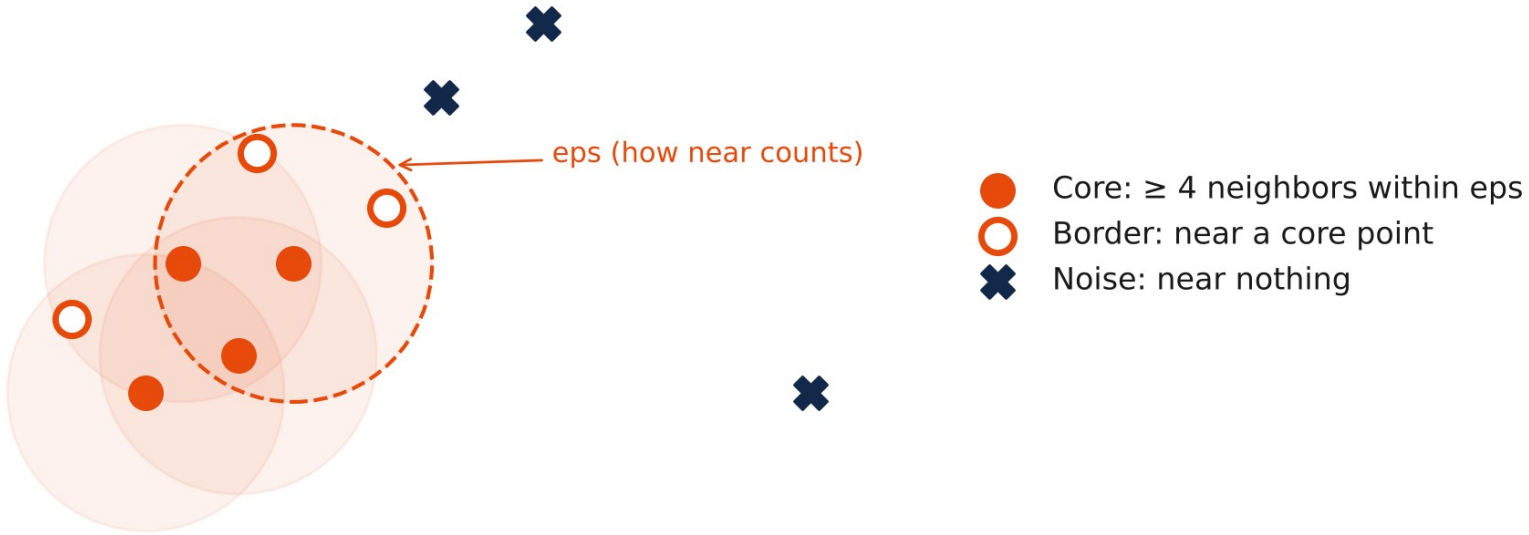
From a drone, people standing close together at a festival form groups; loners are left out.

Variables: position; how close counts as 'near' (eps); minimum crowd size

WHY IT MATTERS

Finds odd-shaped groups, needs no k, and flags outliers for free.

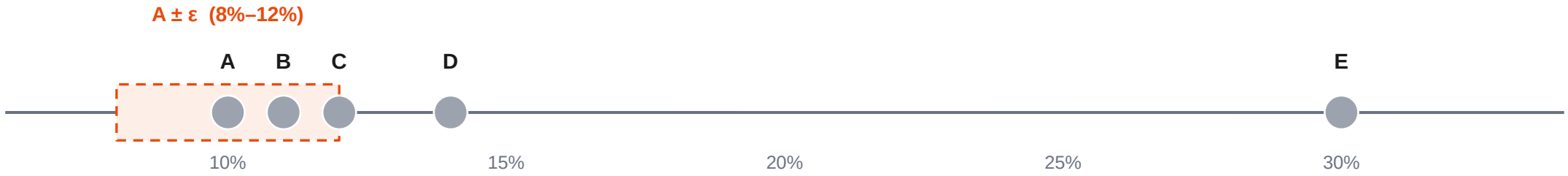
DBSCAN labels every point core, border or noise using one radius and one count



Clusters grow outward from core points; anything a core point can't reach is noise.

Illustrative diagram: 10 hand-placed points, eps radius shown, min. 4 neighbors (including itself).

DBSCAN by hand: five stocks, one radius, one count



ε = 2 points

Two stocks are neighbors if their volatility differs by 2 or less

MinPts = 3

A stock is 'core' if 3 stocks (itself included) sit within ε

Example: A = 10%

Window 8%–12% holds A, B, C = 3 → A is core

Illustrative volatilities: A 10%, B 11%, C 12%, D 14%, E 30%. One feature keeps distance simple: |difference|.

Draw a window around each stock, count who's inside, then label it core, border or noise

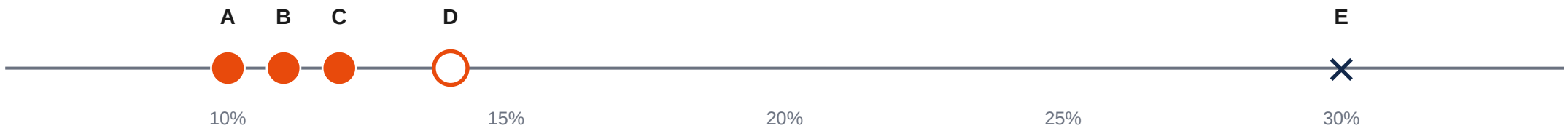
Stock	Vol.	Window (±ε)	Inside	Count	Label
A	10%	8–12%	A, B, C	3	**Core**
B	11%	9–13%	A, B, C	3	**Core**
C	12%	10–14%	A, B, C, D	4	**Core**
D	14%	12–16%	C, D	2	**Border (next to core C)**
E	30%	28–32%	E	1	**Noise**

RESULT

Cluster 1: A, B, C, D

Noise: E

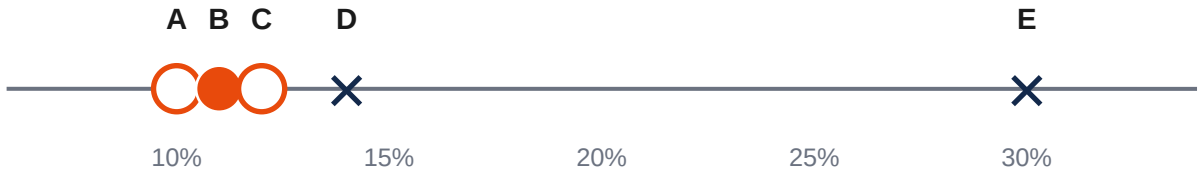
No k chosen. E is left out, not forced into a group.



● *core* ○ *border* ✕ *noise. ε = 2, MinPts = 3. Illustrative numbers.*

Change ε and the answer changes: too small is all noise, too big is one blob

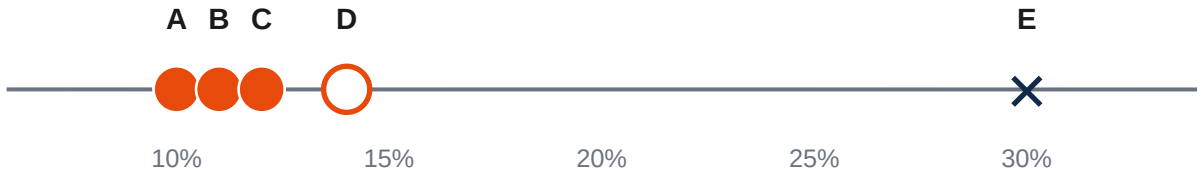
ε = 1



Cluster A, B, C · noise D, E

Tight: D drops out

ε = 2



Cluster A–D · noise E

Sensible: E is the outlier

ε = 20



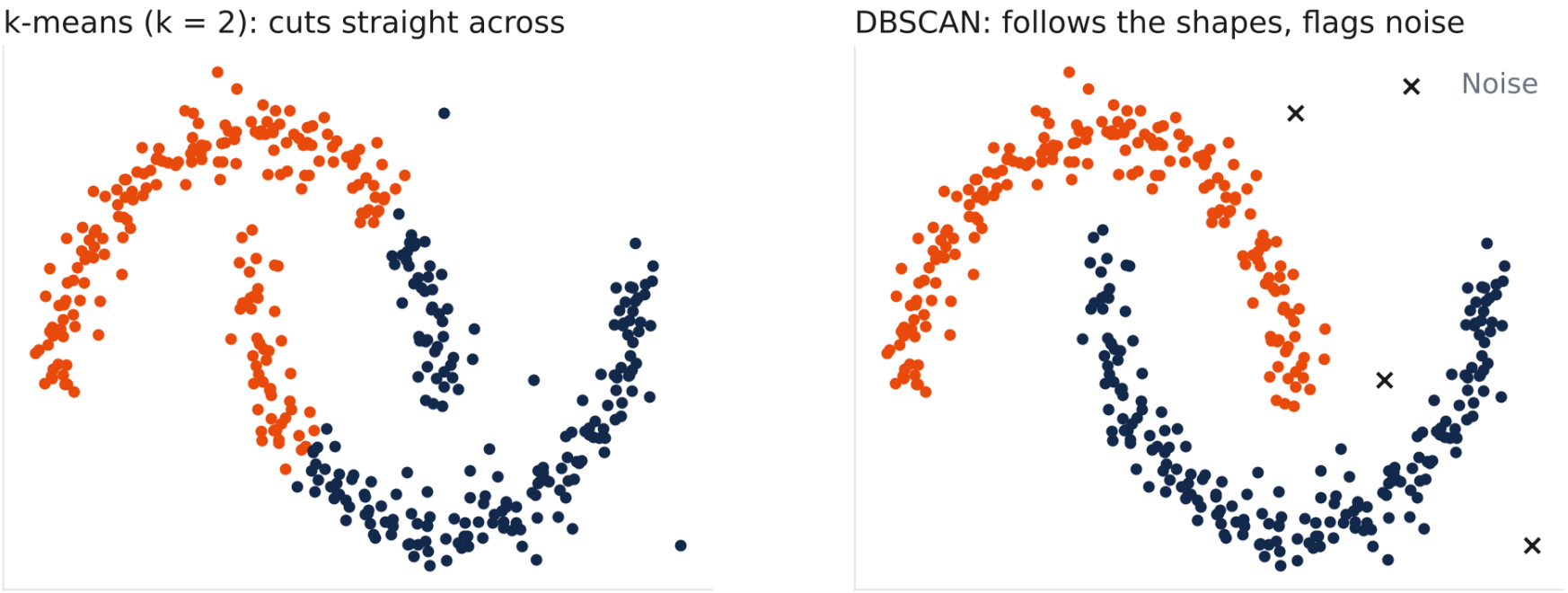
One cluster A–E · no noise

Too loose: E swallowed

How to pick ε: sort each point's distance to its MinPts-th neighbor and look for the bend.

MinPts = 3 throughout. ● core ○ border ✕ noise.

k-means forces round groups; DBSCAN follows the real shape and flags noise



Same 412 points. DBSCAN separates both arcs and flags 3 stray points.

Synthetic benchmark: scikit-learn make_moons (400 points, seed 0) + 12 uniform stray points. Illustrative shape, not measured data.

CLUSTERING CHEAT SHEET

Pick the clustering method by the shape of your problem

Method	You choose	Best when	Watch out
k-means	k	Large data, roundish groups	Outliers pull centers
Hierarchical	Where to cut	Want nested groups	Slow on big data
DBSCAN	eps, min size	Odd shapes, noise	Uneven density

CHECKPOINT · TAKE 30 SECONDS

The PM wants stocks grouped into peer sets, but a few stocks fit no group. Which method?

> **CONTEXT** Return correlations for a few hundred stocks; a handful of names trade unlike anything else.

A k-means with a large k, so every stock gets a group

B DBSCAN: groups dense clusters and labels the loners as noise

C PCA: compress the returns to two components

CHECKPOINT · ANSWER

The PM wants stocks grouped into peer sets, but a few stocks fit no group. Which method?

CONTEXT Return correlations for a few hundred stocks; a handful of names trade unlike anything else.

A k-means with a large k, so every stock gets a group

✓ **B. DBSCAN: groups dense clusters and labels the loners as noise**

C PCA: compress the returns to two components

WHY

DBSCAN needs no k, follows irregular group shapes, and leaves stocks that belong nowhere as noise instead of forcing them in.

03

Anomaly detection

Flag what doesn't fit the pattern.

DEFINITION

Anomaly detection

Learn what normal looks like, then flag the rows that look least like it.

EVERYDAY EXAMPLE

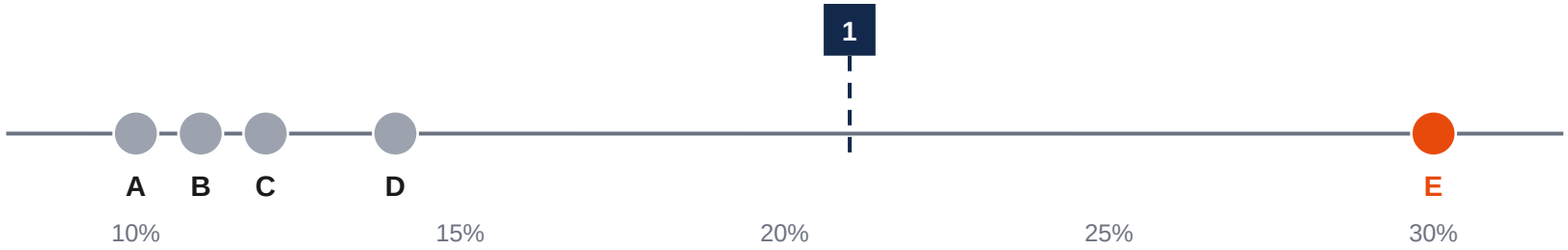
Your bank texts you about a $900 charge in another country at 3 a.m.

Variables: amount, location, time of day

WHY IT MATTERS

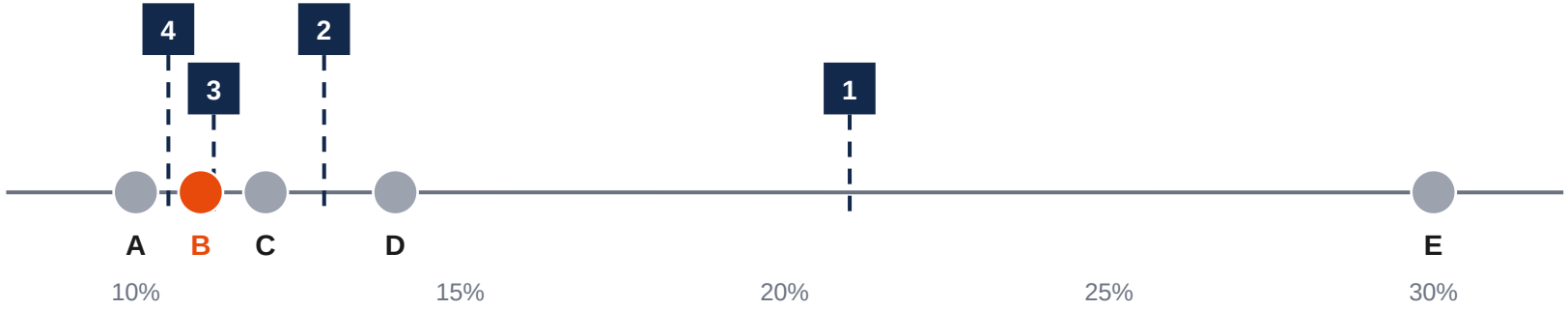
Fraud and errors are rare and unlabeled, so you can't train a classifier on them.

Cut at random until a stock stands alone: E takes 1 cut, B takes 4



Isolating E (30%) — 1 cut

Cut 1 at 21%: E is the only stock above it, so it is alone.



Isolating B (11%) — 4 cuts

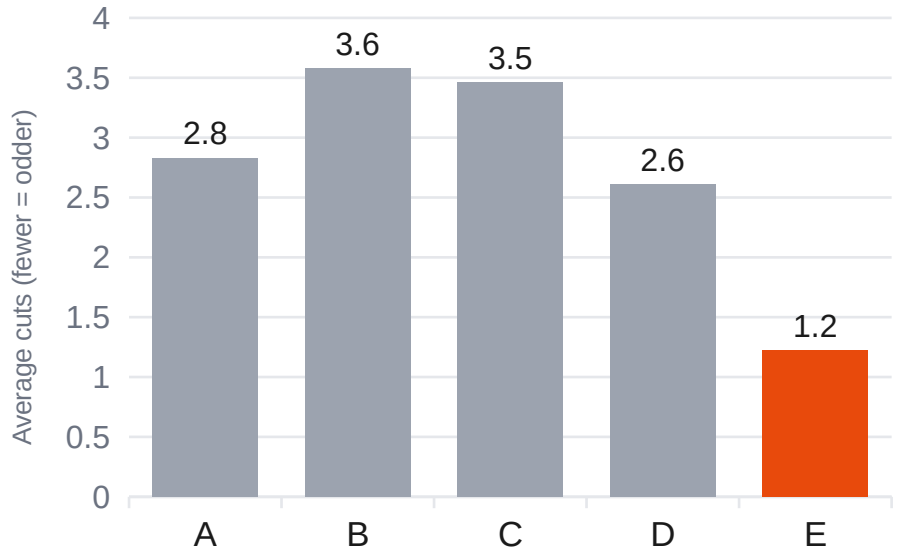
Cut 1 at 21% leaves A B C D; cut 2 at 12.9% leaves A B C; cut 3 at 11.2% leaves A B; cut 4 at 10.5% leaves B alone.

Rule: pick a random point between the lowest and highest stock left; keep the side your stock is on.

Same illustrative volatilities as DBSCAN: A 10%, B 11%, C 12%, D 14%, E 30%. Cut points are one actual random draw.

Repeat with many random trees: the fewest cuts on average marks the anomaly

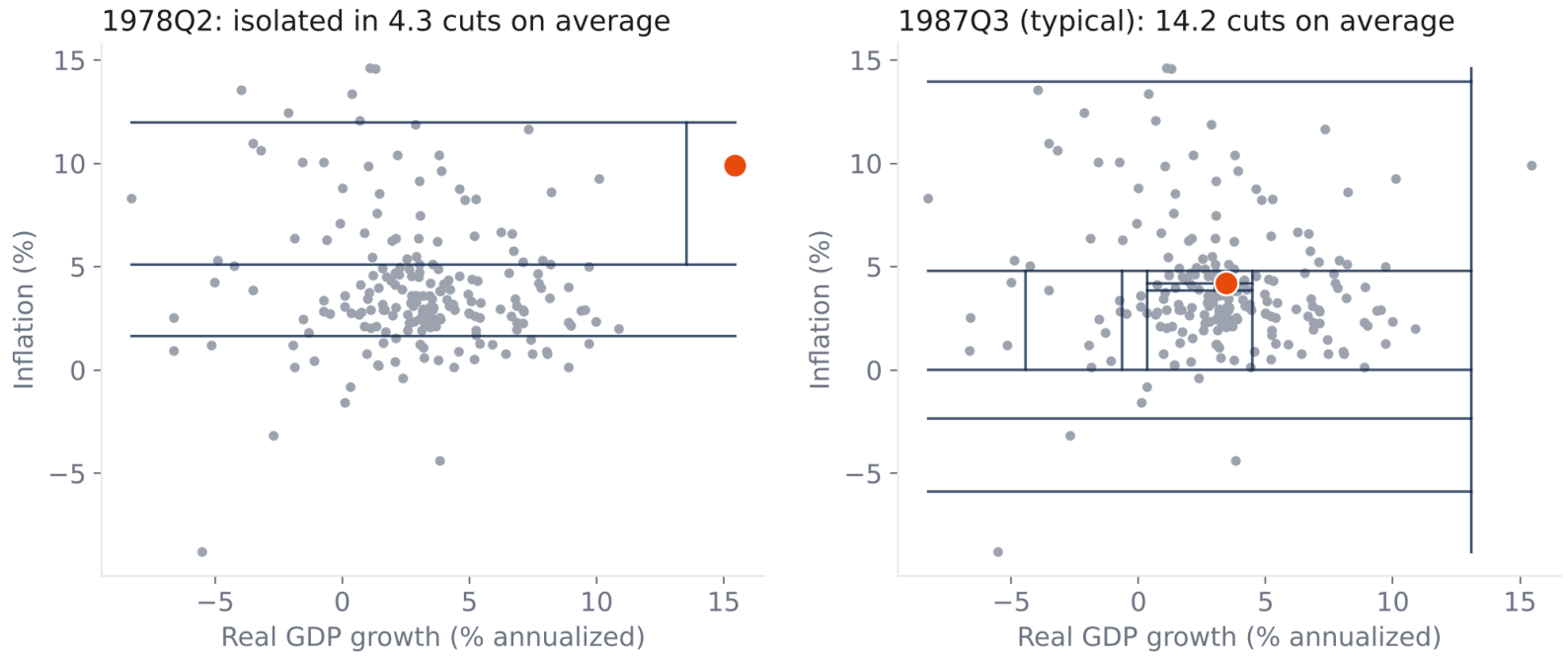
Stock	Vol.	Tree 1	Tree 2	Tree 3	Avg, 10,000 trees
A	10%	4	2	2	**2.8**
B	11%	4	2	4	**3.6**
C	12%	3	3	4	**3.5**
D	14%	2	3	3	**2.6**
E	30%	1	2	1	**1.2**



Flag the top 20% (1 of 5): E. D comes next: a little apart, but not flagged.

Trees 1–3 are actual random draws; averages from 10,000 simulated trees. 'Top 20%' is the contamination setting you choose.

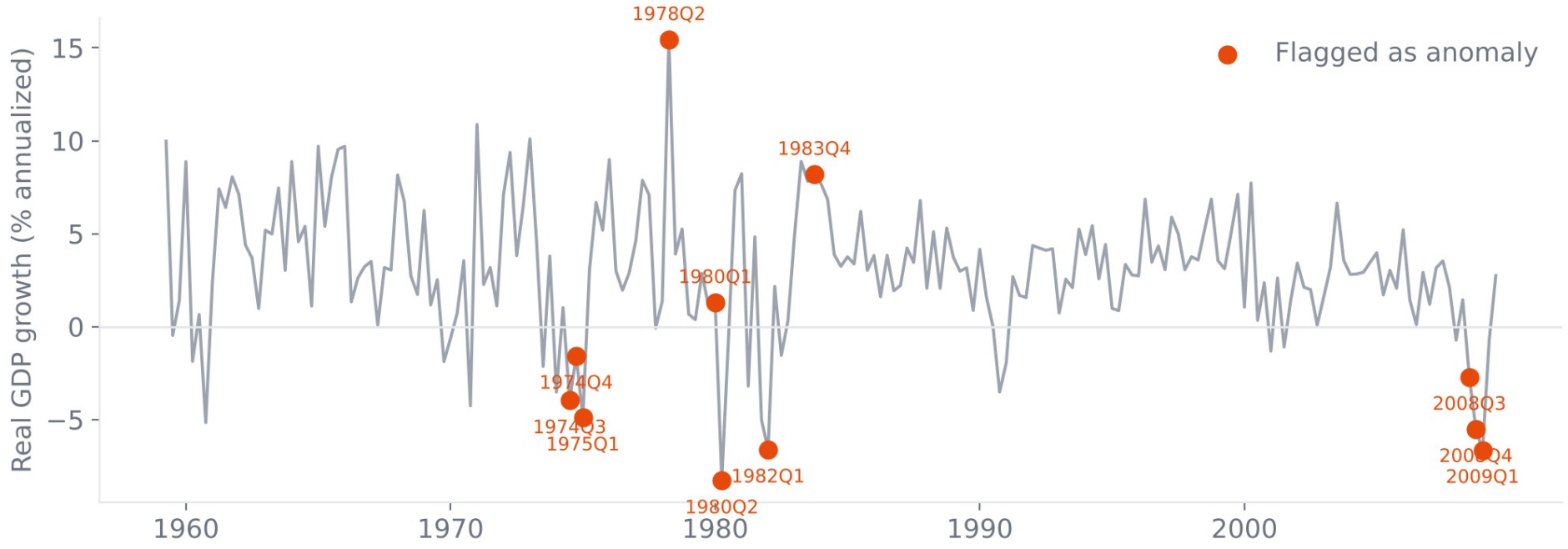
Random cuts isolate odd points fastest: 1978Q2 in ~4 cuts, a typical quarter in ~14



Each line is one random cut. Few cuts to fence a point off = anomaly.

Computed: random cuts on US GDP growth and inflation, 1959–2009 (statsmodels); 2,000 runs.

With no recession labels, Isolation Forest flags the 1970s shocks, 1980-82 and 2008



Flags cluster around oil shocks, the Volcker recessions and the financial crisis.

Data: US quarterly macro 1959-2009 (statsmodels macrodata). Features: real GDP growth, inflation, change in unemployment.

On the same 426 companies, Isolation Forest surfaces the names an analyst should check first

Company	What stands out	Its value	S&P median
CrowdStrike (CRWD)	P/E	3,989	23.7
Molina Healthcare (MOH)	P/E	2,387.8	23.7
Palantir Technologies (PLTR)	Price / sales	78.7	3.1
Charter Communications (CHTR)	P/E	2.8	23.7
Vici Properties (VICI)	Dividend yield	7.9%	1.4%

Isolation Forest (500 trees) on the five standardized ratios; top 5 scores. Yahoo Finance snapshot, not current.

CHECKPOINT · TAKE 30 SECONDS

Isolation Forest flagged 1978Q2, a boom quarter. Is the model wrong?

> **CONTEXT** US real GDP grew 15.4% annualized in 1978Q2; most flagged quarters are recessions.

A Yes: anomaly detection should only flag crashes

B No: unusual in either direction gets flagged; an analyst decides which flags matter

C Yes: the contamination setting must be too high

CHECKPOINT · ANSWER

Isolation Forest flagged 1978Q2, a boom quarter. Is the model wrong?

CONTEXT US real GDP grew 15.4% annualized in 1978Q2; most flagged quarters are recessions.

A Yes: anomaly detection should only flag crashes

✓ **B. No: unusual in either direction gets flagged; an analyst decides which flags matter**

C Yes: the contamination setting must be too high

WHY

The model only knows 'unlike the rest'. A record boom is as unusual as a crash. Interpreting flags is the human's job.

04

At the firm

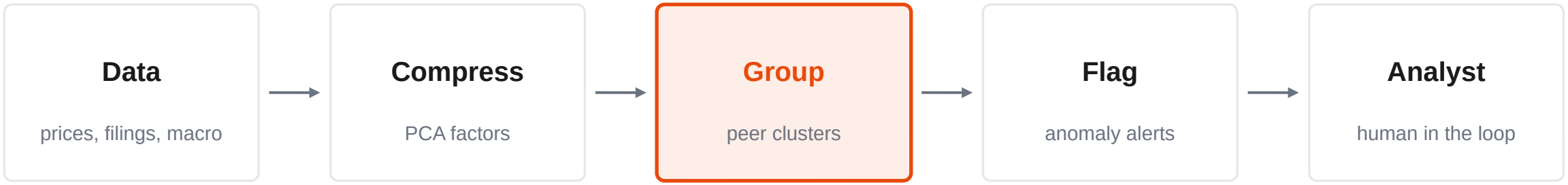
How Champaign Capital Research puts these to work.

USE CASES

Each method maps to a real research task at the firm

Task	Method	Input	Output
Find a stock's true peers	k-means / hierarchical	Return correlations, ratios	Peer groups
Spot market regimes	k-means on macro	Growth, inflation, rates	Regime labels
Build risk factors	PCA	Daily stock returns	Factor exposures
Catch unusual moves	Isolation Forest	Price, volume, filings	Ranked alerts
Segment clients	k-means / DBSCAN	Holdings, activity	Client segments

In our agent, unsupervised models narrow the field; the analyst decides



The agent proposes groups and flags; a human approves before any decision.

IN CODE

Each method is about three lines in scikit-learn

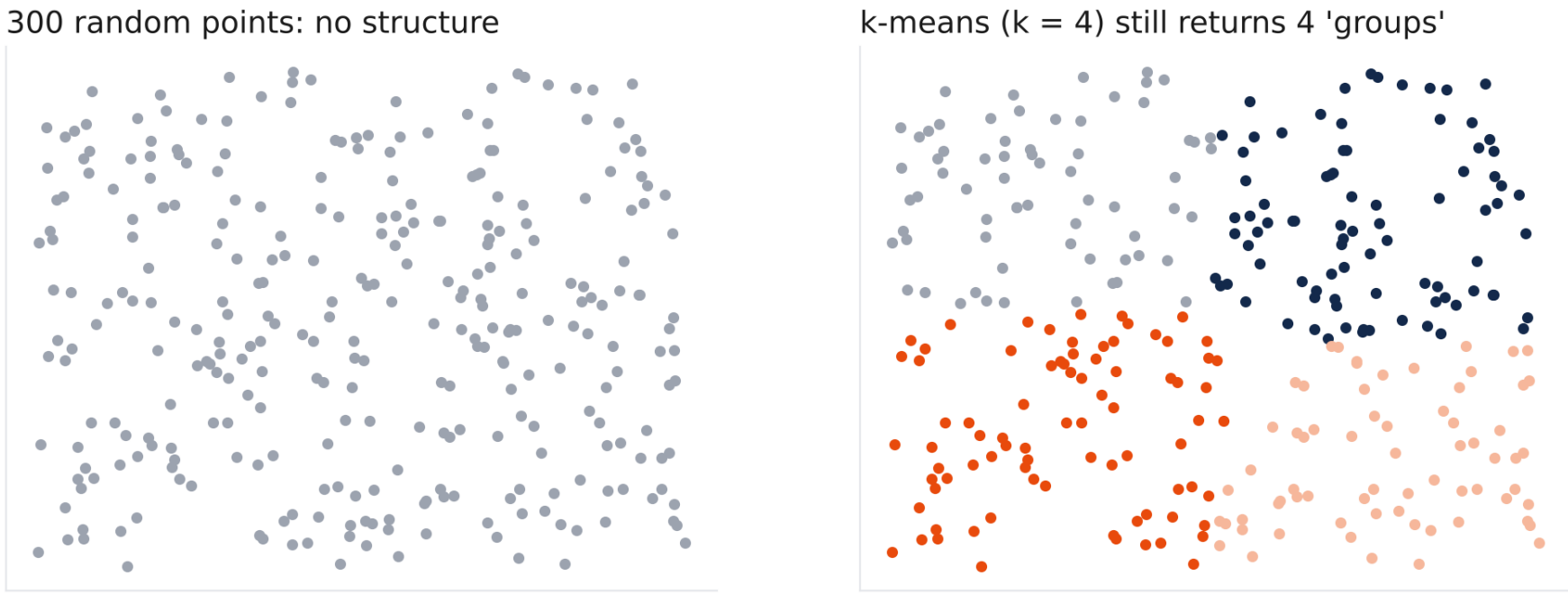
```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans, DBSCAN
from sklearn.decomposition import PCA
from sklearn.ensemble import IsolationForest

X = StandardScaler().fit_transform(df)
groups = KMeans(n_clusters=3).fit_predict(X)
factors = PCA(n_components=2).fit_transform(X)
flags = IsolationForest(contamination=0.05).fit_predict(X)
```



The hard part isn't the code; it's choosing features and judging results.

k-means always returns k groups, even when the data has no groups at all



Random noise, yet a silhouette of 0.39: similar to some 'real' results earlier.

Synthetic: 300 uniform random points (seed 3); k-means k = 4. Illustrative, not measured data.

PITFALLS

No labels means the easiest mistakes are invisible

- Forgetting to scale lets one column decide everything
- The algorithm always returns groups, even in pure noise
- Check stability: resample, rerun, see if groups hold

Final test: can you name each cluster in plain English?

05

Case study: what moves oil?

Everything from today, applied to oil in 2026: five experiments and what held up.

One question, five experiments: what moves oil, and can we predict it?

1

UNSUPERVISED

Flag

Isolation Forest marks abnormal oil days

→

2

SUPERVISED

Forecast

Regression predicts next month's return and risk

→

3

HONEST TEST

Search

2,047 variable combos, scored on unseen 2019–26

→

4

HOLDOUT

Stress-test

Train to 2025, then face 2026's war

→

5

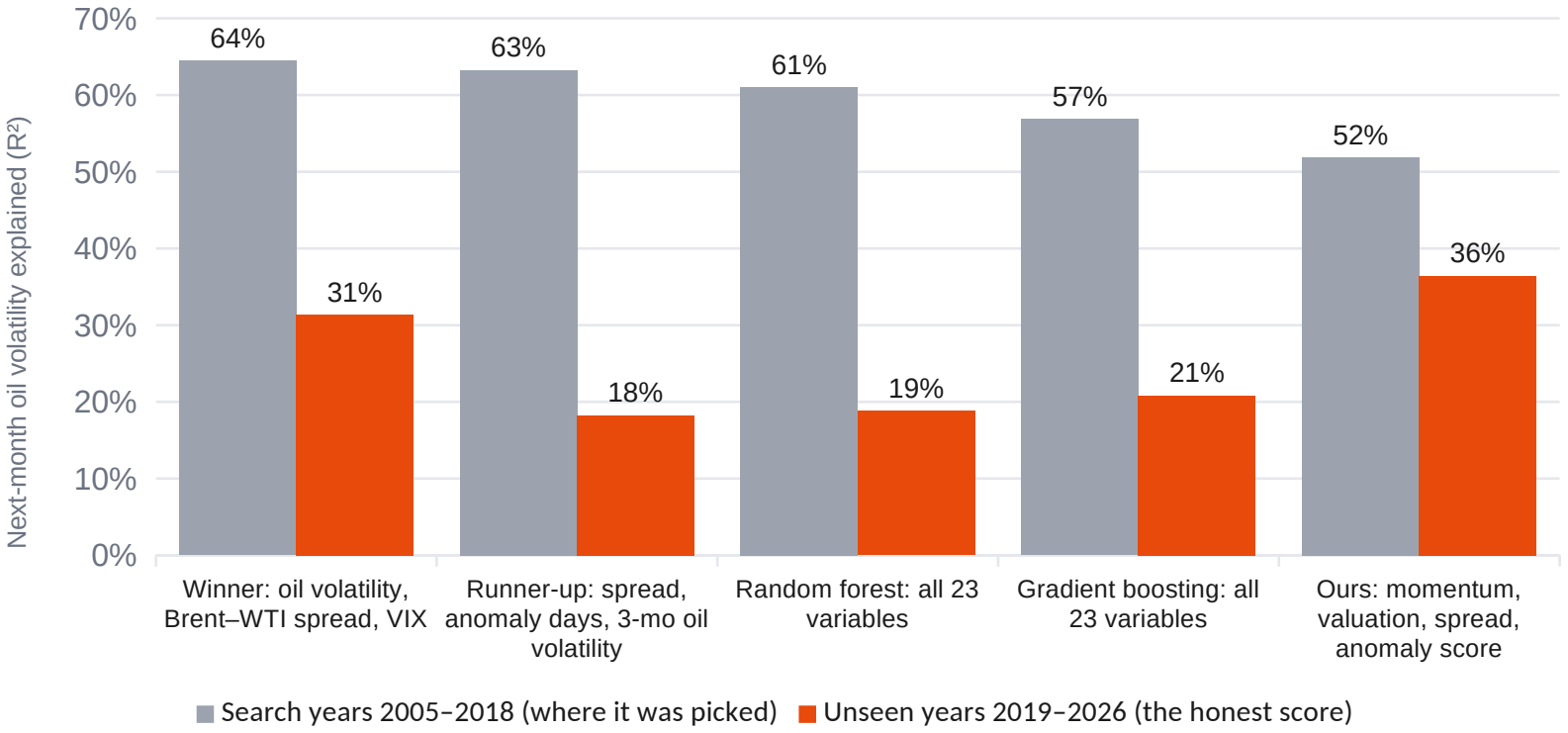
ATTRIBUTION

Explain

War events vs Trump posts vs markets

Data: Brent and WTI daily (EIA), VIX, gold, natural gas, S&P 500, FX and 10-yr yields (GitHub datasets); Trump posts (Twitter and Truth Social archives).

The best of 2,047 combos lost half its score on unseen years; simple models held up best



DIRECTION (UP OR DOWN)

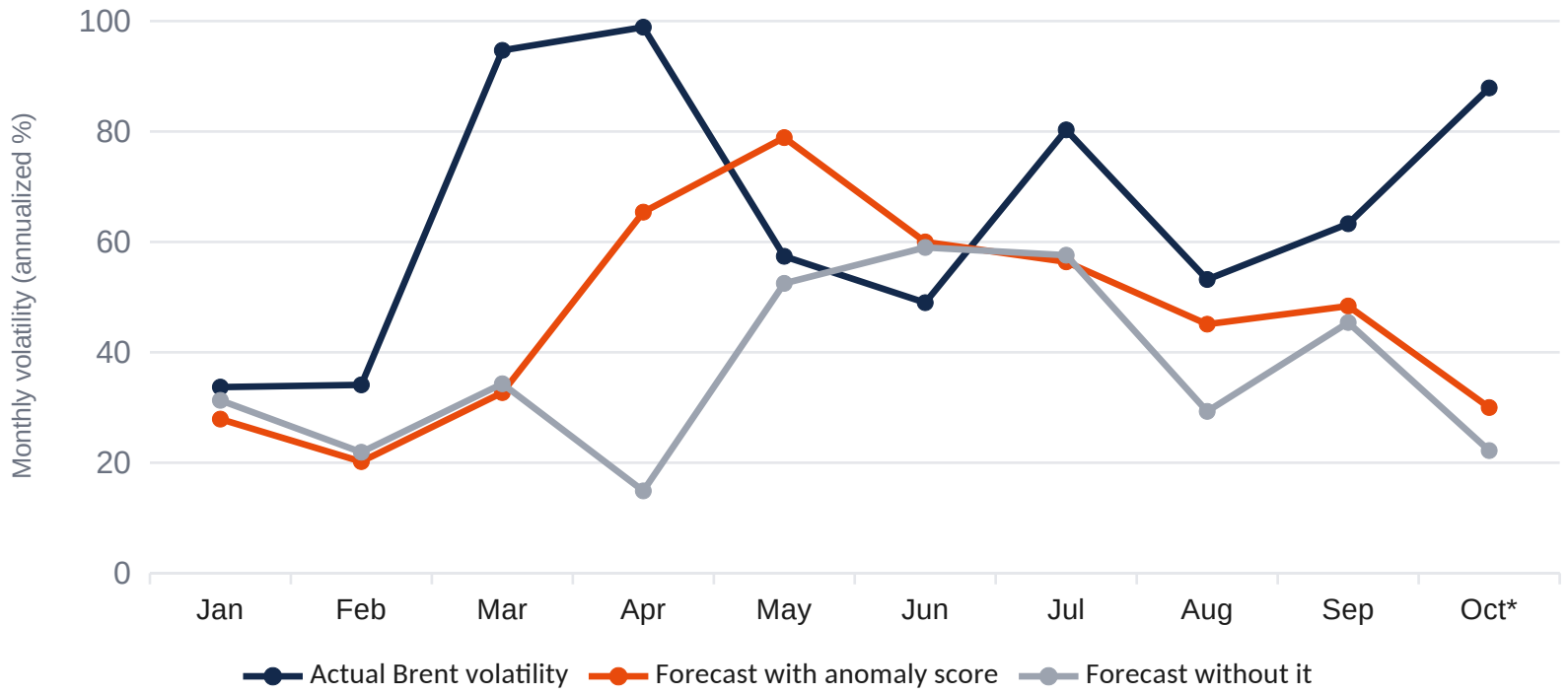
54.3%

right on unseen years, vs 56.5% for "always guess up"

Only 9.5% of 2,047 combos beat a simple average. Price direction isn't in these 23 variables.

Walk-forward ridge regression, refit yearly; combos chosen on 2005–2018 only. Pool: 23 price, macro and anomaly variables.

Trained only through 2025, the model with the anomaly score kept up with the war; without it, it fell behind



WITH ANOMALY SCORE

+25.7%

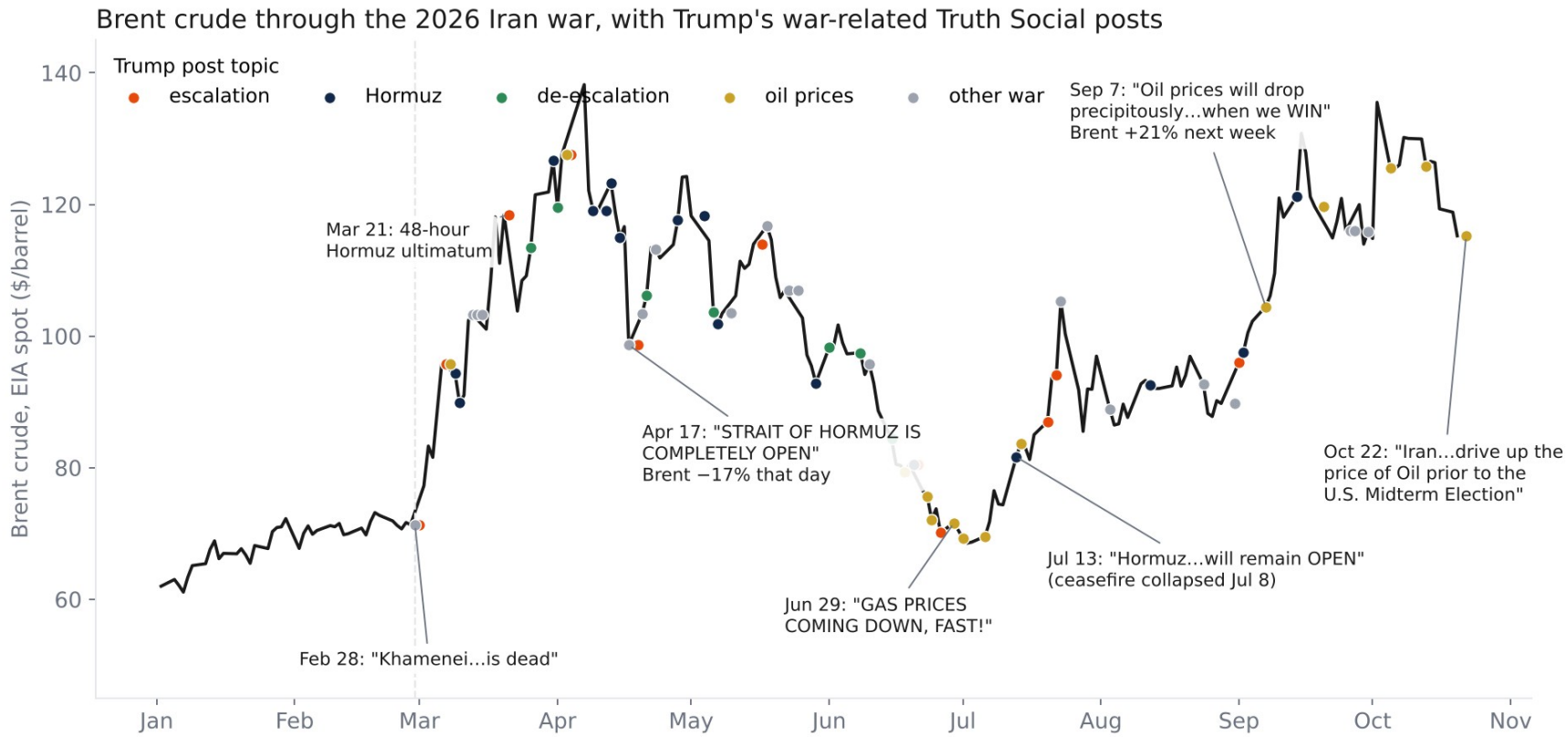
WITHOUT IT

-24.3%

R^2 on 2026 vs the historical average. Neither saw March coming.

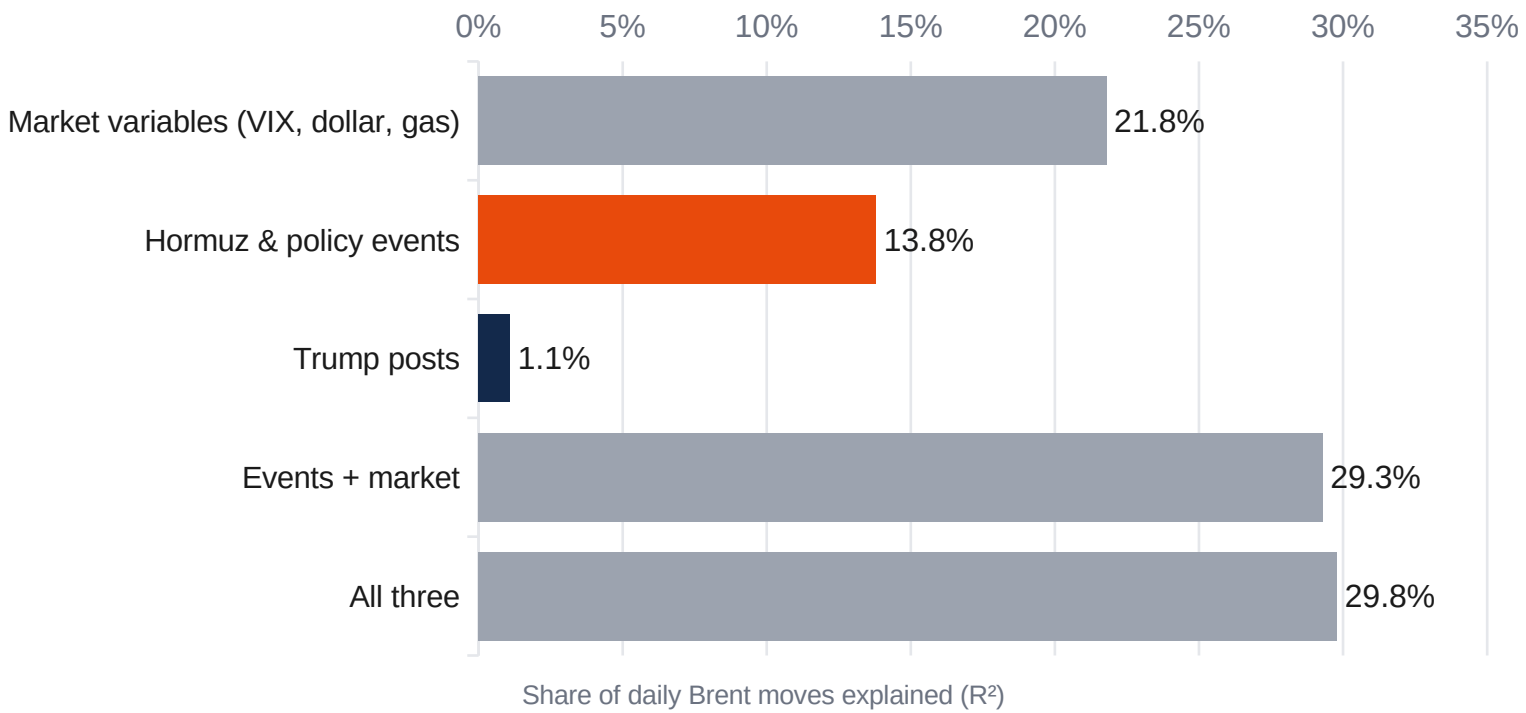
**Oct 1–20 only. With: anomaly score + anomaly days, VIX level + change, Brent 1-mo momentum, Brent–WTI spread. Without: momentum, valuation, spread, oil volatility.*

Oil followed the Strait of Hormuz; Trump's posts mostly narrated it



Sources: Brent spot from EIA via datasets/oil-prices (GitHub), through Oct 20, 2026. Posts: sample of 92 war-related Truth Social posts from trumpstruth.org (Feb 28–Oct 22, 2026); not a complete archive.

During the war, Hormuz events and market fear explained oil's daily moves; the posts added almost nothing



ON HORMUZ & POLICY EVENT DAYS

12 of 13

moved the expected way: up on closures and strikes, down on ceasefires and reopening

≈4.29 pts per event day. ~70% of daily moves stay unexplained.

Daily Brent, Mar 2–Sep 22, 2026 (139 days). OLS; events from a sourced war timeline; 92 sampled posts.

Four lessons for building the research agent

01 Risk is predictable; direction isn't

≈30% of next-month volatility explained out of sample; up/down no better than a coin.

02 Unsupervised signals pay off in breaks

The anomaly score adds little in calm years but kept the model on track in 2026.

03 Simple beats complex

Best-of-2,047 and 23-variable ML lost half their score on unseen years.

04 Watch events, not posts

Hormuz status explained moves; posts mostly narrated them, after the fact.

In the agent: flag the regime → forecast risk → alert on Hormuz-type events → a human decides.

Scan to get the code and data, then run it yourself

Oil case study · code + data + agent skill

1. Open in Colab: run all cells, about 1 minute, no install.
2. Or on your laptop: pip install, then python run.py.
3. Using a coding agent? Give it skill/oil-case-study/SKILL.md and ask it to test a new variable.

github.com/ashcastelinocs124/msba-ai-workshop/tree/main/case-studies/oil-2026

06

Dimensionality reduction (if time)

Compress many columns into a few: PCA, t-SNE, UMAP.

DEFINITION

Principal Component Analysis (PCA)

Replace many correlated columns with a few new ones that keep most of the variation.

EVERYDAY EXAMPLE

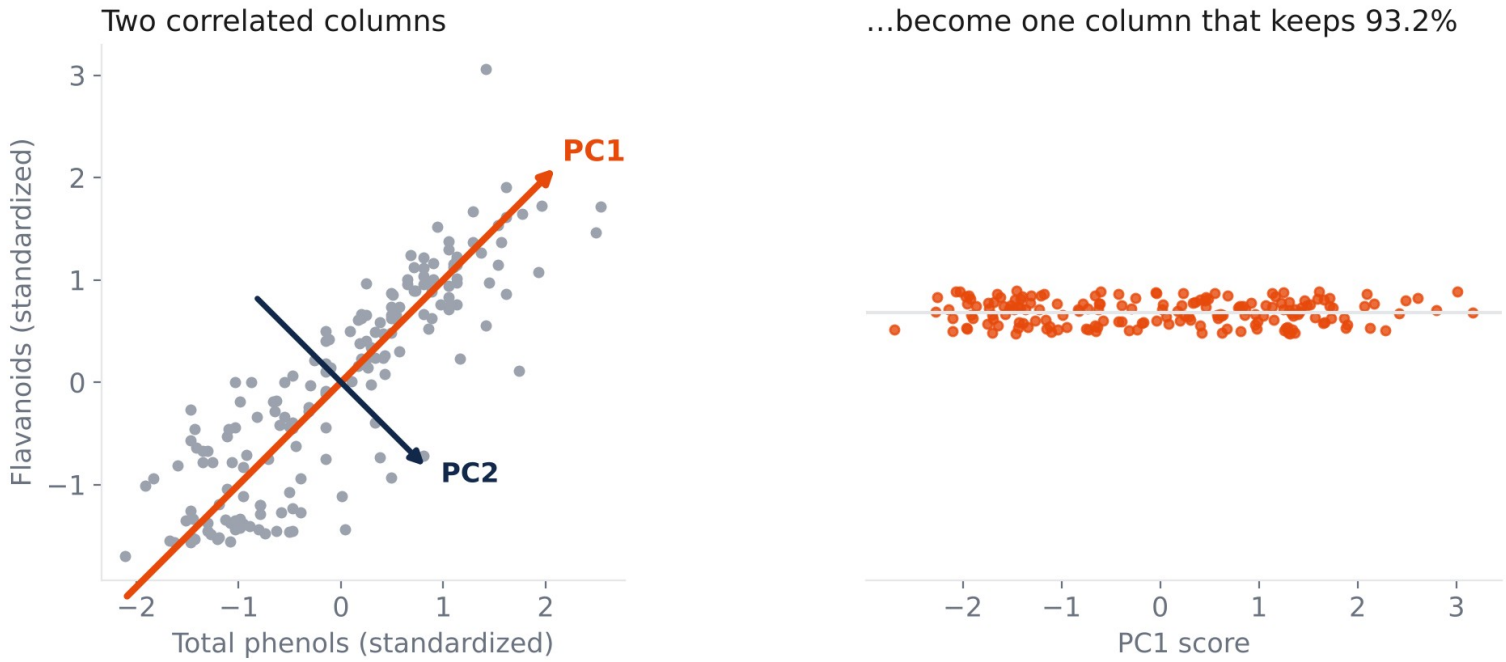
Summarizing a student's 10 course grades as one 'overall strength' score plus one 'math vs writing' score.

Variables: 10 grades in → 2 component scores out

WHY IT MATTERS

Fewer columns: faster models, less noise, and you can finally plot the data.

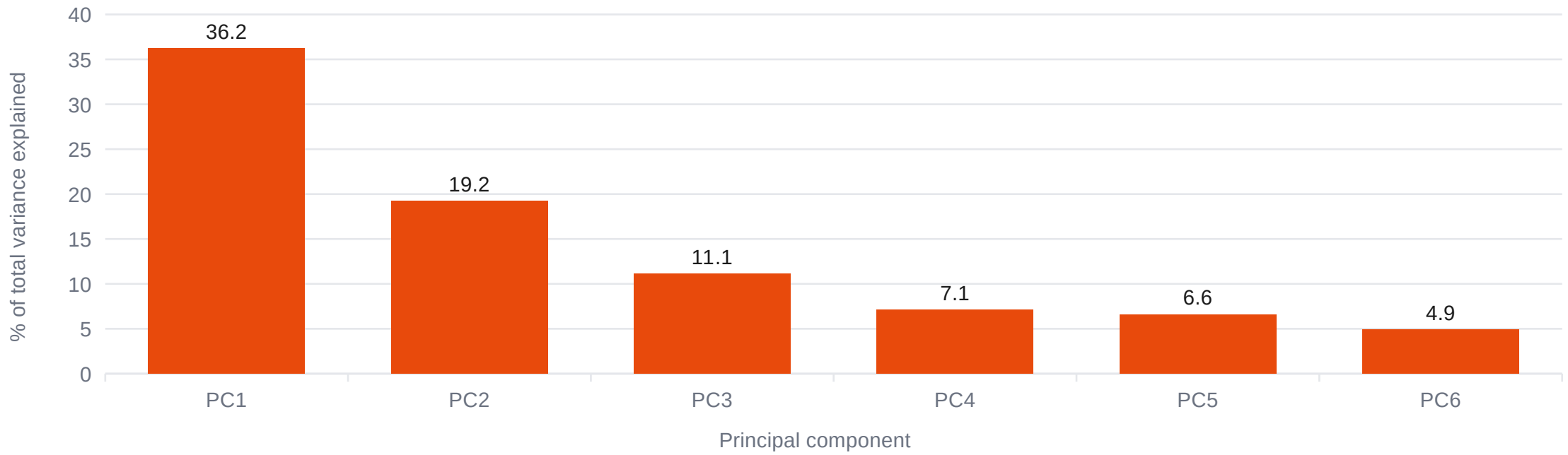
PCA finds the direction the data stretches most and measures every point along it



Left: two wine measurements that move together. Right: each wine's position along PC1.

Computed: UCI Wine, total phenols vs flavanoids (correlation 0.86), standardized; scikit-learn PCA.

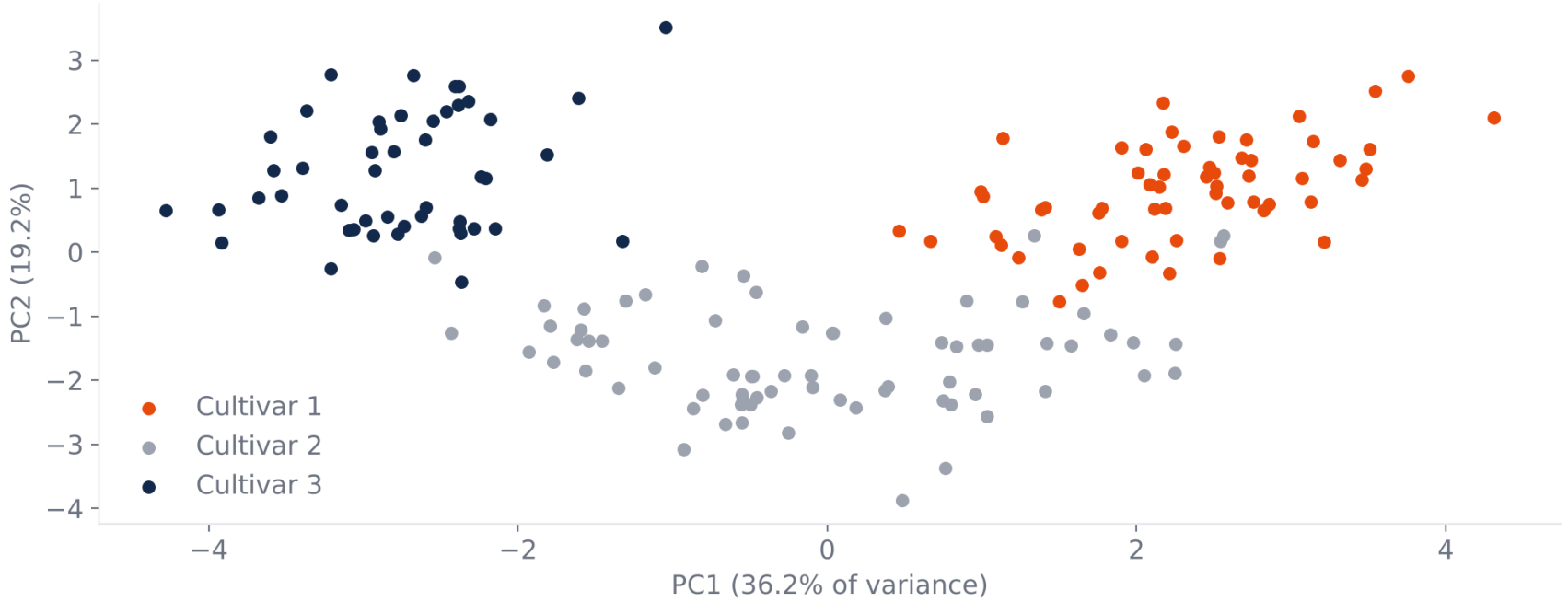
Two components keep 55% of the information in 13 chemical measurements



Each extra component adds less; most signal lives in the first few.

Computed: PCA on standardized UCI Wine data (178 wines, 13 chemical measurements); scikit-learn.

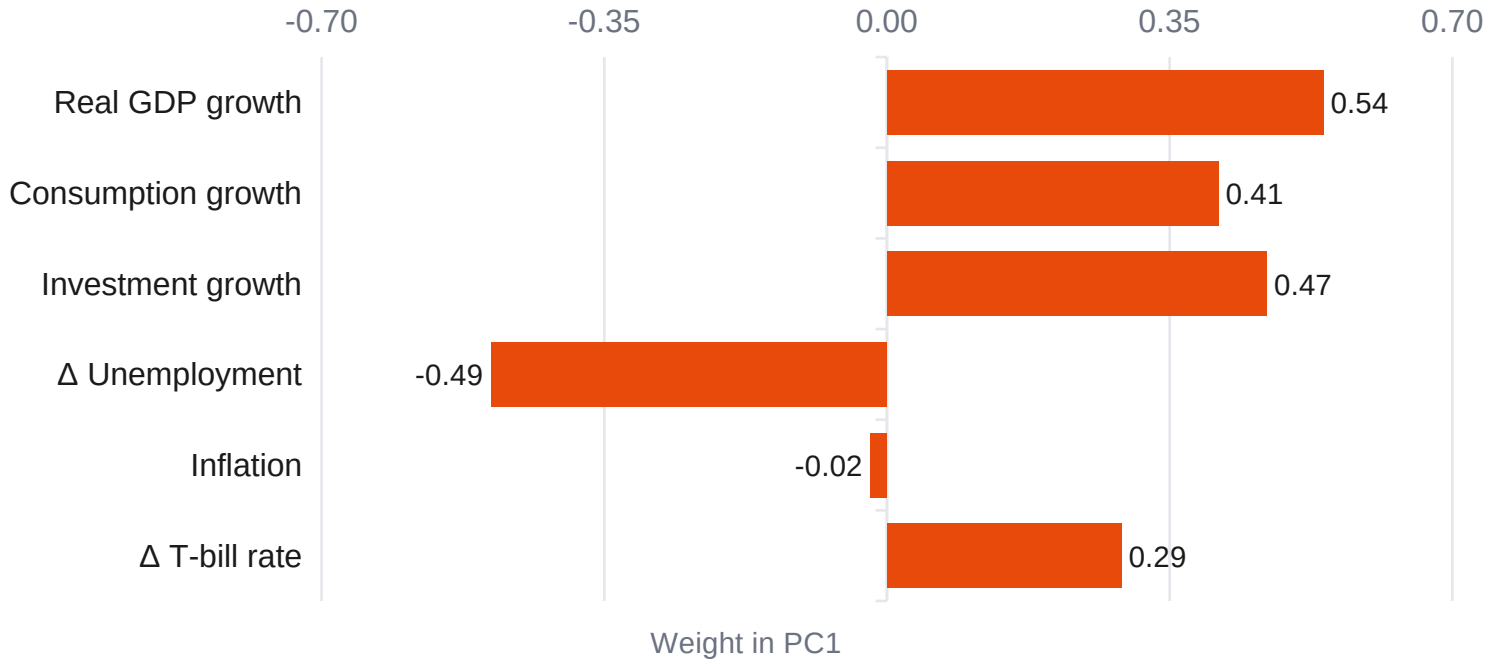
Plotted on just two components, the three grape varieties separate on their own



Color is added afterwards to check; PCA never saw the variety.

Computed: PCA on standardized UCI Wine data (178 wines); colors = known cultivar, not used by PCA.

One component summarizes 50.2% of six macro series: it is the business cycle



HOW TO READ IT

Growth, spending and investment pull one way; unemployment pulls the other.

High PC1 = expansion.

Low PC1 = contraction.

Inflation barely loads: it lives on PC2.

Computed: PCA on standardized quarterly changes, US 1959–2009 (statsmodels macrodata). Δ = change from prior quarter.

For pictures of complex data, t-SNE reveals groups that PCA blurs together

PCA: digits overlap



t-SNE: ten clear islands



1,797 handwritten digits, 64 pixels each, squeezed into 2D two ways.

Computed: scikit-learn Digits (UCI, 1,797 images, 8×8 px); PCA vs t-SNE (perplexity 30). Colors = true digit.

WHICH ONE?

Use PCA to feed models; use t-SNE or UMAP to look at data

PCA

Linear and fast

Components are interpretable

Reusable on new data

t-SNE / UMAP

Non-linear, great pictures

Distances between islands mean little

Mainly for exploration

Model inputs: PCA. Visual exploration: t-SNE or UMAP.

What to remember

01 Unsupervised learning finds structure without labels; judge it by usefulness

02 k-means: assign to nearest center, move centers, repeat. DBSCAN: window, count, label

03 Isolation Forest flags rare rows by how fast random cuts isolate them

04 Oil case: risk is forecastable, direction isn't; watch events, not posts

05 PCA compresses columns for models; t-SNE and UMAP are for pictures

Questions?

Next: Workshop 6, Advanced ML — deep learning and transformers (Oct 30).

GO DEEPER

scikit-learn user guide: Clustering · Decomposition · Novelty and outlier detection

James et al., An Introduction to Statistical Learning, ch. 12 (Unsupervised Learning)

Data: Iris, Wine, Digits (UCI via scikit-learn) · US macrodata (statsmodels)

Today's chapter

BEFORE YOU LEAVE

Scan to mark your attendance

Paste the series attendance QR code here

Attendance · Workshop 5

Fri, Oct 23, 2026 · 1:30–3:30 pm · 3007 BIF

1. Point your phone camera at the code.
2. Fill in the short attendance form.
3. Attend 3 sessions to earn credit for the workshop.