

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Data Transformation: Normalization & Discretization

Lecture 7 · Wednesday 30 September 2026

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–5 min) Recap and today's roadmap
2. (5–17 min) Foundations: what transformation is; mean, σ and distance refreshed; why scale breaks distance
3. (17–35 min) Min-max, z-score, decimal scaling, robust z-score — every symbol explained, each worked by hand
4. (35–42 min) Data leakage, and a real experiment: k-NN on all 500 Zanzibar farms
5. (42–58 min) Discretization: bins → entropy → information gain → the best split point
6. (58–62 min) Scaling: Welford's one-pass algorithm for streaming data
7. (62–72 min) In-class exercise in pairs
8. (72–75 min) Summary, questions, references

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Apply	min-max, z-score, decimal scaling and the robust (MAD) z-score by hand — and say which suits which data
Explain	with numbers, why unscaled attributes distort Euclidean distance and k-NN
Avoid	data leakage by computing normalization statistics on the training set only
Compute	entropy and information gain, and use them to choose a discretization boundary
Describe	how Welford's algorithm normalizes a data stream in one pass with constant memory

Recap: Where Week 2 Left Us

IIT Madras Zanzibar

- Lecture 5: Euclidean / Manhattan distance treat every attribute's **numeric size** as if it were meaningful
- Lecture 6: binning grouped nearby values to **smooth** noise — today the same bins become **categories**

Lecture 5–6: the problem

Raw attributes, raw units

rainfall ≈ 680 – 2,020 mm

yield ≈ 0.1 – 11.5 tonnes

→ rainfall dominates any distance

Lecture 7: the fix

Transform before you mine

Normalization: numbers → a common scale

Discretization: numbers → categories

→ every attribute gets a fair vote

Data Transformation: Normalization & Discretization

IIT Madras Zanzibar

→ **Foundations: What, Why, and the Tools We Need**

Normalization Methods, Step by Step

Data Leakage & the Evidence

Discretization, Entropy & Information Gain

Scaling: Normalizing a Stream

In-Class Exercise

What Is Data Transformation?

IIT Madras Zanzibar

Definition. Data transformation converts attribute values into a form that is more suitable for mining — same records, same attributes, but the values are re-expressed.

Cleaning
Week 2

→

Integration
Week 2

→

Reduction
Tomorrow

→

Transformation
TODAY

Smoothing
remove noise
e.g. binning yields (Lecture 6)

Attribute construction
build a new attribute
e.g. revenue ÷ farm size = revenue per ha

Aggregation
summarize
e.g. daily sales → monthly totals

Normalization ← today
rescale to a common range
e.g. rainfall 680–2,020 mm → 0 to 1

Discretization ← today
numbers → labelled intervals
e.g. rainfall → Low / High

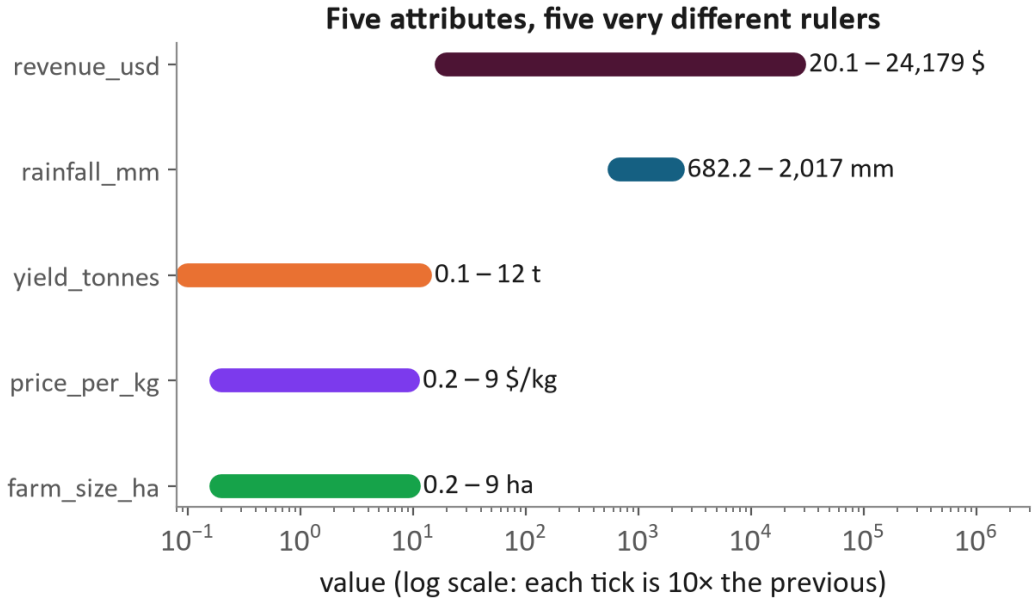
Concept hierarchy
values → higher-level concepts
e.g. village → district → region

Meet the Data: 500 Zanzibar Farms

IIT Madras Zanzibar

farm	crop	size (ha)	price ($/kg)	yield (t)	rainfall (mm)	revenue ($)
1	Cassava	0.25	0.26	—*	1,780.8	1,841
2	Rice	4.18	0.69	—*	1,765.6	2,526
…	…	…	…	…	…	…
500	…	…	…	…	…	…

**24 missing yields were filled with their crop's median (Week 2 cleaning), so all 500 farms are used.*



The problem in one sentence: the attributes are measured with different rulers — hectares, dollars, tonnes, millimetres — so their numbers differ in size by more than a thousand times.

Normalization puts every attribute on the same ruler before we compare or combine them.

Normalization in Plain Words

IIT Madras Zanzibar

Everyday example: who did better?

Student	Raw mark	Out of	Normalized
Amina	45	50	45/50 = **0.90**
Baraka	70	100	70/100 = 0.70

Raw numbers say Baraka (70 > 45).

On a common scale, **Amina did better** (0.90 > 0.70).

It does NOT change the order of values or the shape of the distribution — only the ruler.

Definition

Normalization rescales a numeric attribute, so its values fall in a small, common range — e.g., 0 to 1, or centered on 0 with a spread of 1.

Purpose

1. **Fairness**
   no attribute dominates because of its unit
2. **Comparability**
   different attributes on one ruler
3. **Correct algorithms**
   distances, gradients and PCA behave properly

Today: min-max · z-score · decimal scaling · robust z-score

Why Data Mining Cares About Normalization

IIT Madras Zanzibar

Needs normalization (combines attributes)

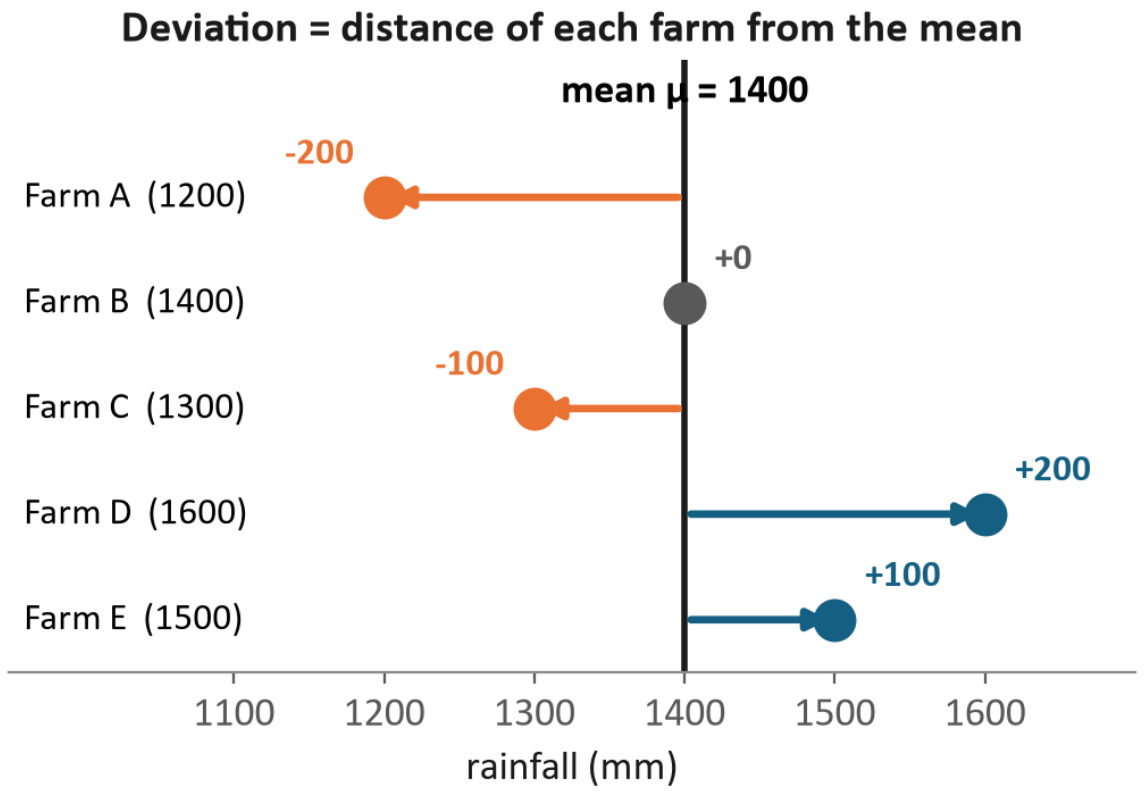
- ✓ **k-NN**
  Week 6: nearest records by distance
- ✓ **k-means clustering**
  Week 8: nearest cluster centre
- ✓ **PCA**
  Tomorrow: directions of largest variance
- ✓ **Neural nets / gradient descent**
  similar scales → faster, stable training

Usually does not need it (one attribute at a time)

- – **Decision trees**
  Week 7: one attribute vs a threshold
- – **Naive Bayes**
  per-attribute probabilities
- – **Association rules**
  Week 5: items, not magnitudes

Rule for your projects: if the algorithm combines attributes into a distance or a variance, normalize first.

Tool 1 Refreshed: Mean (μ) and Standard Deviation (σ)



Five farms' rainfall: 1200, 1400, 1300, 1600, 1500 mm

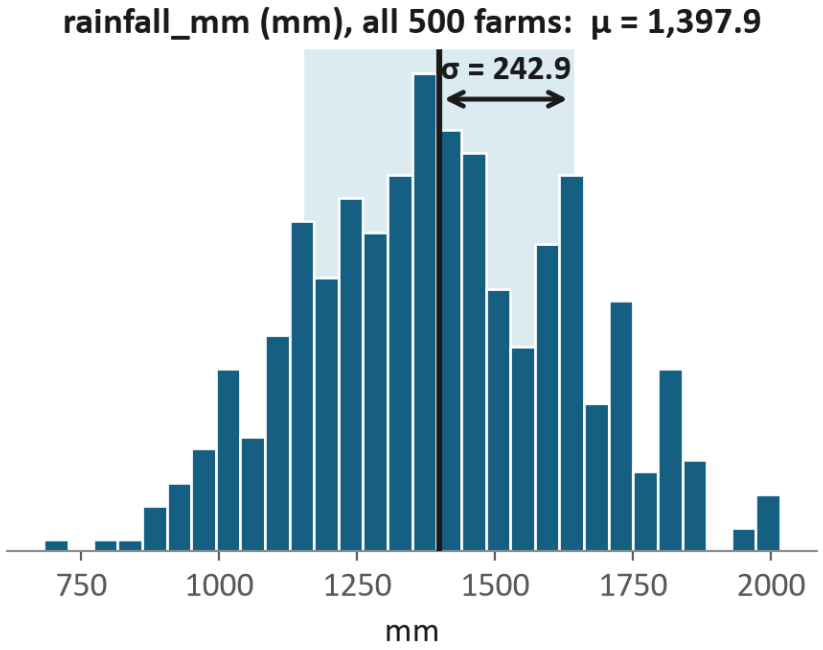
1. **Mean μ = Σv / n**
   7000 / 5 = 1400 mm (the balance point)
2. **Deviations v − μ**
   −200, 0, −100, +200, +100 (they sum to 0)
3. **Square and average**
   (40,000+0+10,000+40,000+10,000) / 5 = 20,000
4. **Square root → σ**
   √20,000 = 141.4 mm (a typical deviation)

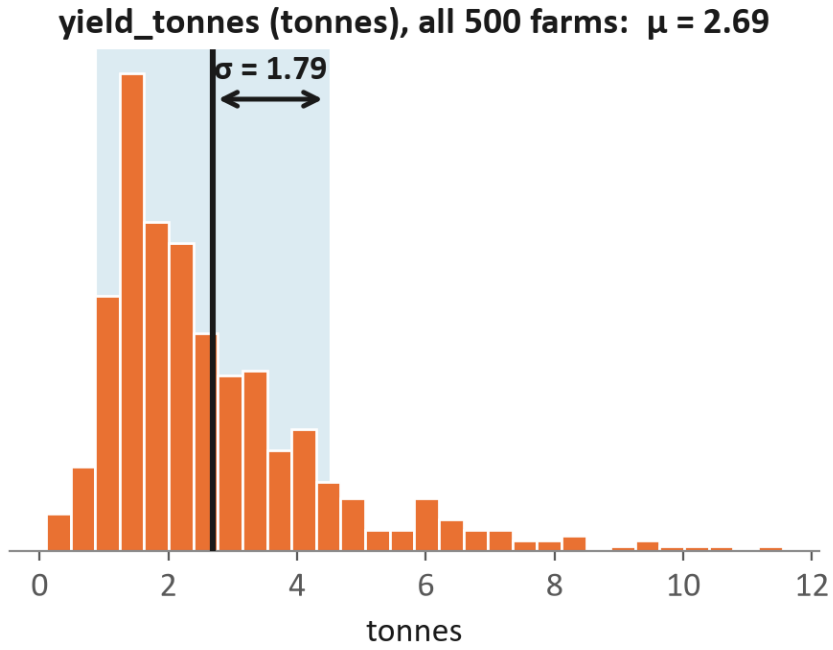
$$\sigma = \sqrt{(\Sigma \, (v - \mu)^2 / n)}$$

μ = the average; σ = the typical distance from the average

In words: σ tells you how spread out an attribute is, measured in that attribute's own units.

Where Our Dataset's μ and σ Come From





Rainfall
μ = 1,397.9 mm
σ = 242.9 mm

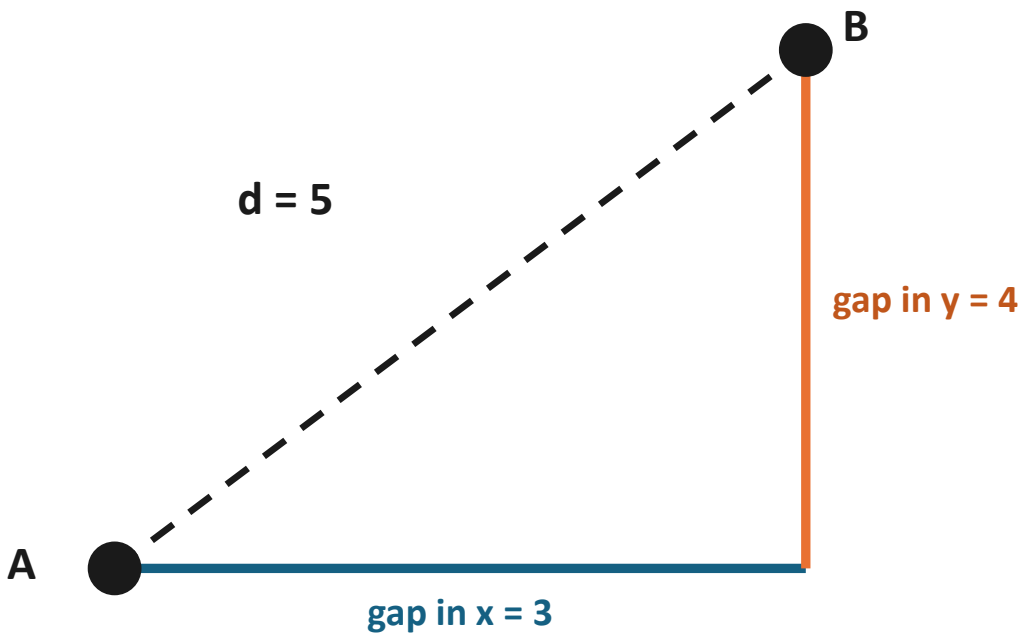
Yield
μ = 2.69 t
σ = 1.79 t

Same recipe, 500 values: add, divide by 500 → μ; deviations, square, average, square-root → σ. Shaded band = μ ± σ (about 2 in 3 farms).

Tool 2 Refreshed: Euclidean Distance

IIT Madras Zanzibar

Distance = Pythagoras on a graph



$$d(A, B) = \sqrt{((x_A - x_B)^2 + (y_A - y_B)^2)}$$

add one squared gap per attribute, then square-root

1 **Gaps**
x: 3, y: 4

2 **Square and add**
$3^2 + 4^2 = 9 + 16 = 25$

3 **Square root**
$d = \sqrt{25} = 5$

Watch this: each attribute adds its gap SQUARED — so a gap of 60 adds 3,600, while a gap of 2.5 adds only 6.25.

Worked Example: A Distance That Lies

IIT Madras Zanzibar

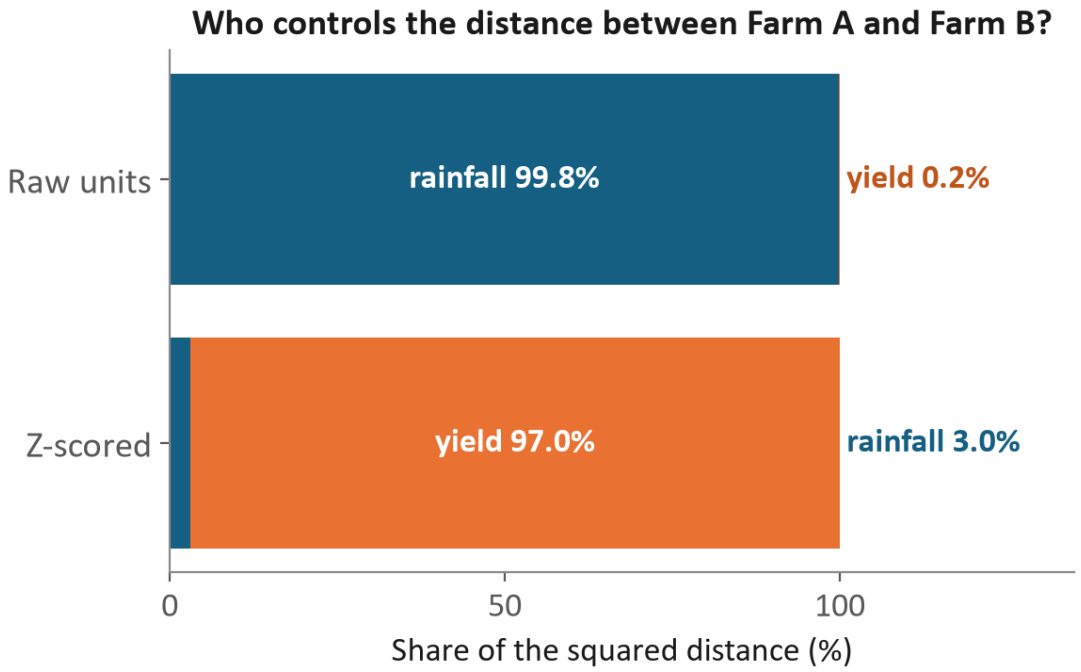
Two Zanzibar farms, raw units

	rainfall (mm)	yield (t)
Farm A	1200	1.5
Farm B	1260	4.0
gap	60	2.5
gap²	**3600**	6.25

1 **Raw distance**
√(3600 + 6.25) ≈ 60.05 → almost pure rainfall

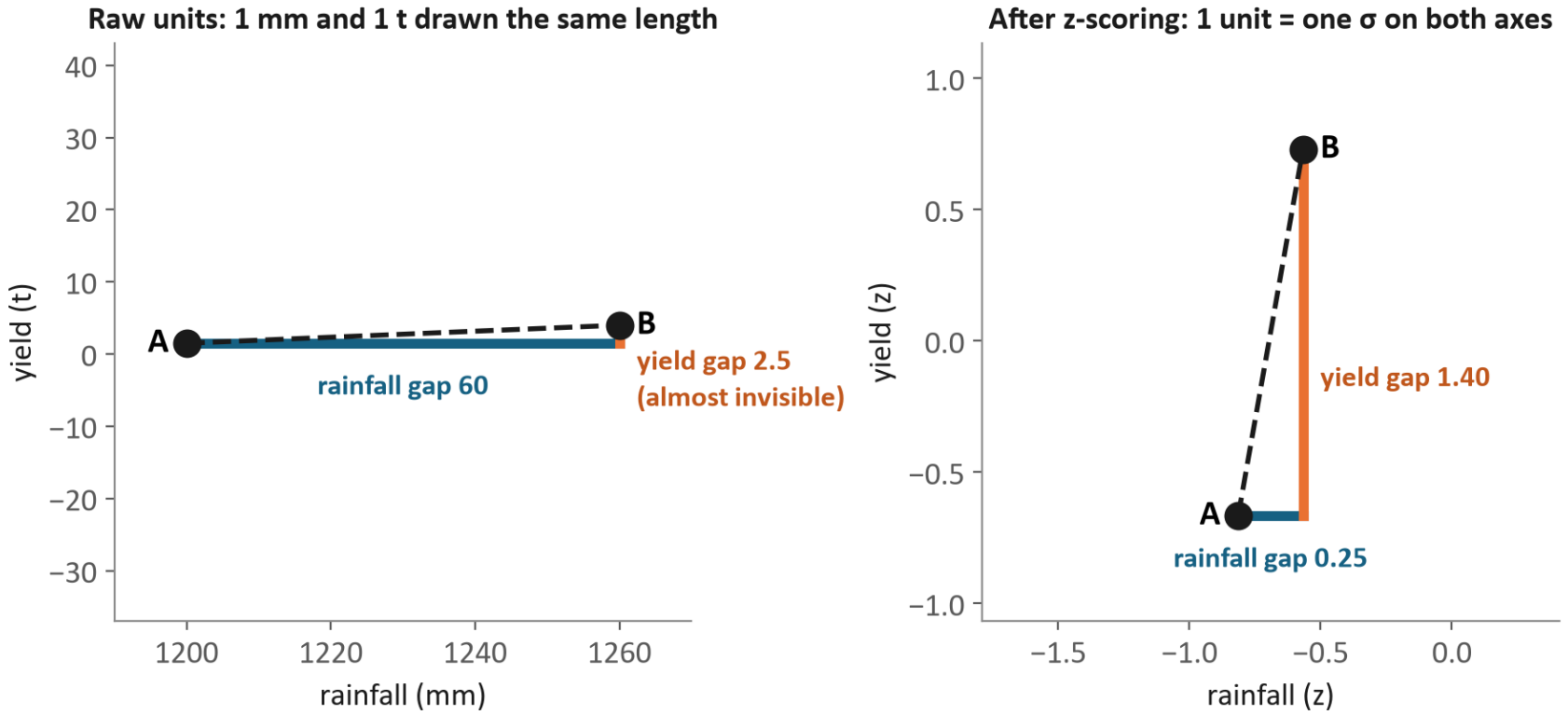
2 **Divide each gap by its σ (242.9 mm; 1.79 t)**
$(60/242.9)^2 = 0.061$ $(2.5/1.79)^2 = 1.950$

3 **Normalized distance**
√(0.061 + 1.950) ≈ 1.42 → driven by yield



Lesson: an attribute's influence on distance should come from its information, not from its unit of measurement.

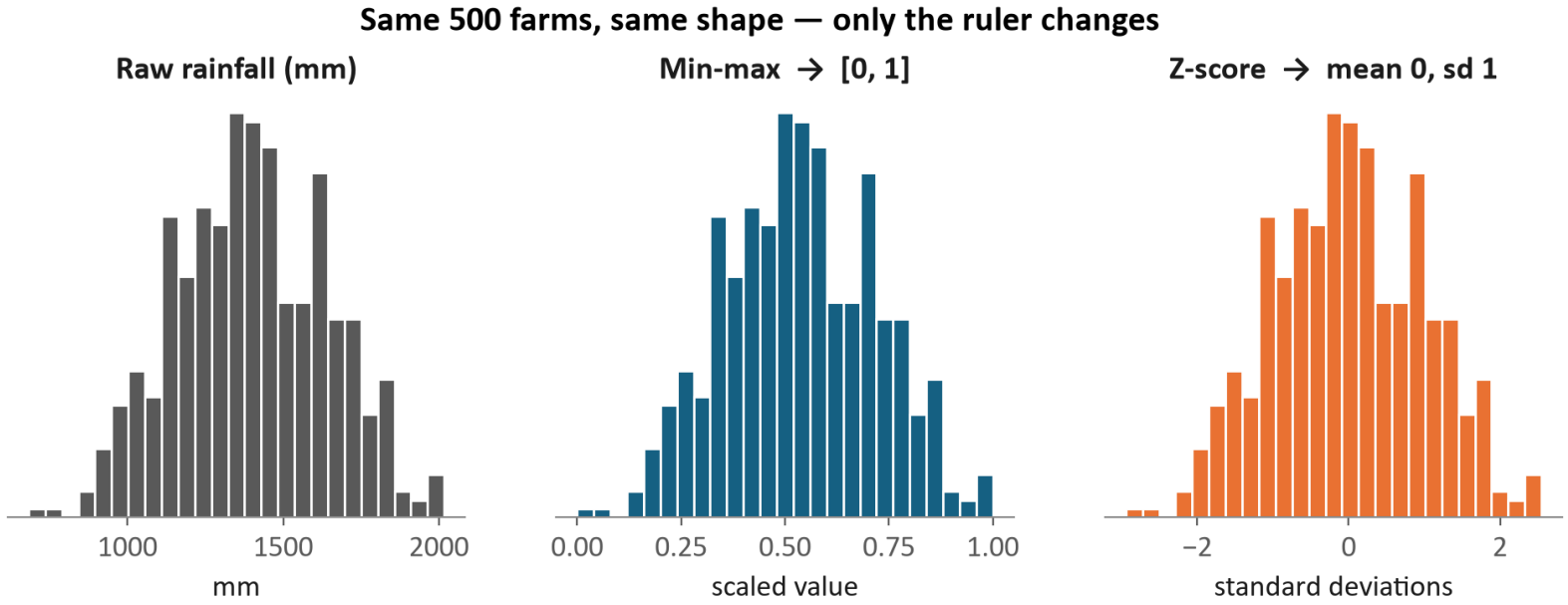
See It: The Same Two Farms, Drawn to Scale



What changed? not the farms — only the ruler on each axis. After normalization, 1 unit on either axis means 'one typical spread' of that attribute.

One Recipe Behind Every Normalization

$$v' = (v - \text{reference}) \,/\, \text{spread}$$



1 **Min-max**
reference = min, spread = max − min

2 **Z-score**
reference = mean μ, spread = σ

3 **Decimal scaling**
reference = 0, spread = 10^j

Data Transformation: Normalization & Discretization

IIT Madras Zanzibar

Foundations: What, Why, and the Tools We Need

→ **Normalization Methods, Step by Step**

Data Leakage & the Evidence

Discretization, Entropy & Information Gain

Scaling: Normalizing a Stream

In-Class Exercise

Min-Max: Reading the Formula

IIT Madras Zanzibar

$$v' = (v - \min) / (\max - \min) \times (new_max - new_min) + new_min$$

with the usual target 0 to 1 (new_min = 0, new_max = 1) this becomes simply v′ = (v − min) / (max − min)

Symbol	What it means	In our example
v	the original value for one record	Farm A rainfall = 1200 mm
v′	the new, normalized value (the answer)	?
min	smallest value of the attribute in the data	1000 mm
max	largest value of the attribute in the data	1600 mm
new_min, new_max	the range we want to land in	0 and 1

In plain words

'What fraction of the way from the smallest to the largest value is v?'

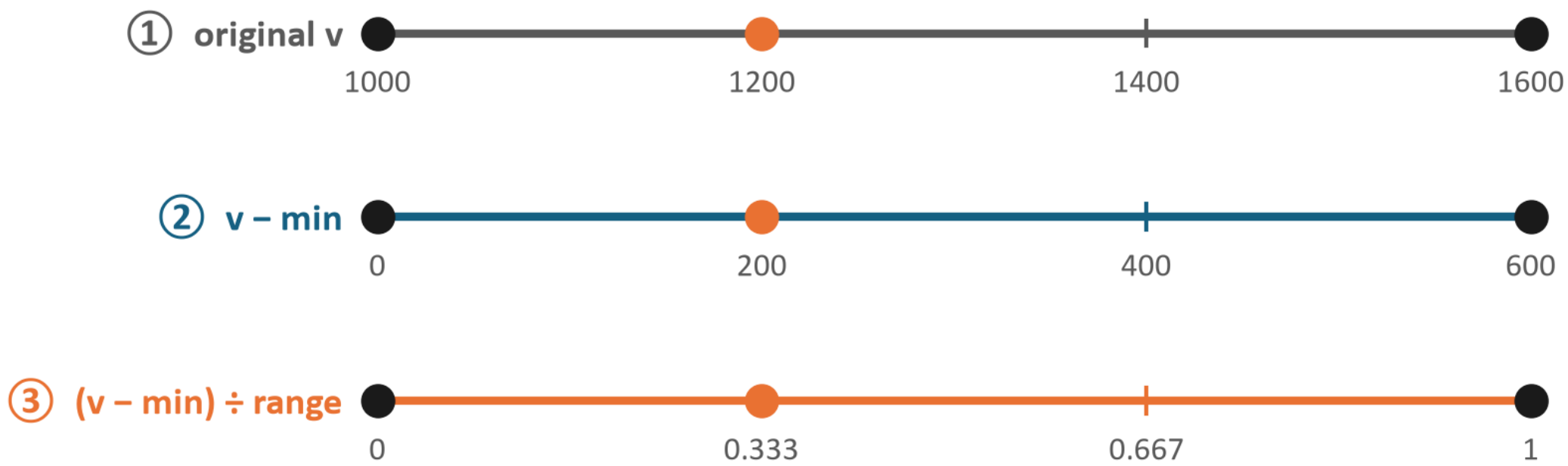
Like a fuel gauge: empty = 0, full = 1, half a tank = 0.5.

Remember: min and max describe the whole column; v is one cell in it.

Min-Max: Where (v – min) and ÷ (max – min) Come From

IIT Madras Zanzibar

Step ② slides everything left so the smallest value sits at 0; step ③ squeezes so the largest sits at 1



1 **Start**
values 1000 … 1600 mm

2 **Shift: subtract min**
smallest → 0; 1200 → 200

3 **Shrink: ÷ range (600)**
largest → 1; 1200 → 0.333

Min-Max: Worked Example

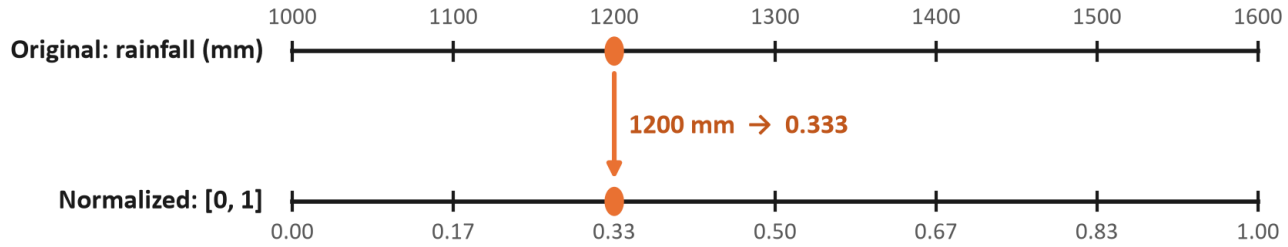
IIT Madras Zanzibar

Normalize Farm A's rainfall (1200 mm) to 0 … 1

1 Range
max − min = 1600 − 1000 = 600

2 Shift
v − min = 1200 − 1000 = 200

3 Scale
200 / 600 = 0.333

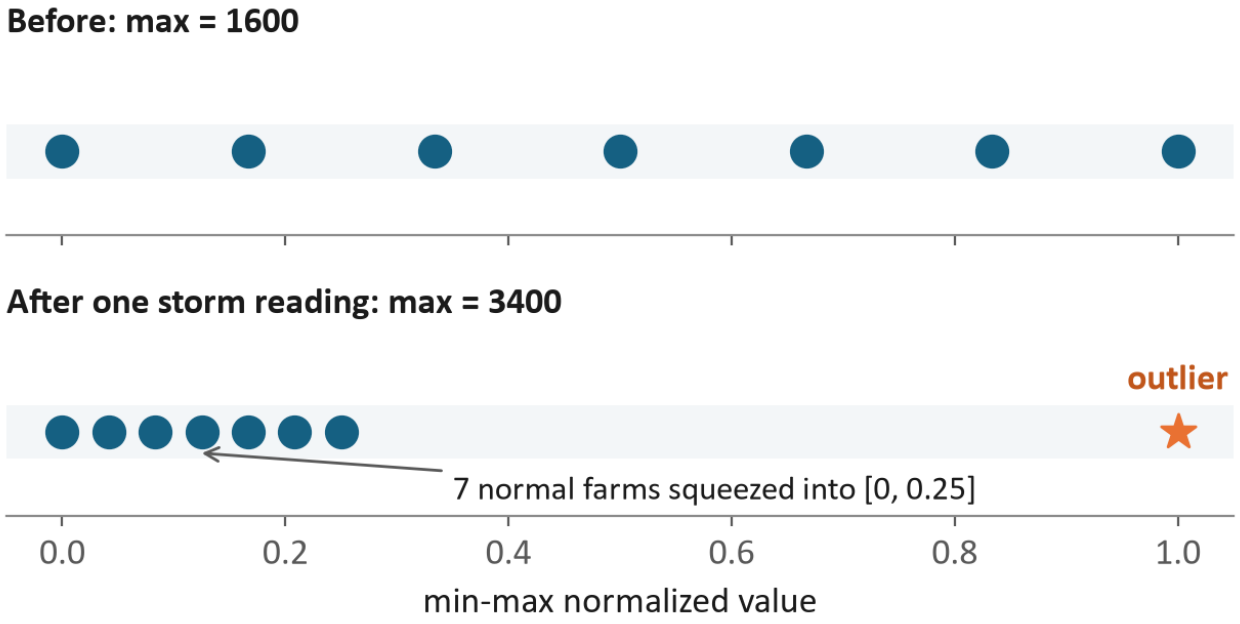


Check it
1000 → 0 1600 → 1
1300 (midpoint) → 0.5 ✓

Target −1 … 1 instead?
0.333 × (1 − (−1)) + (−1)
= 0.667 − 1 = **−0.333**

Real use: photo pixels 0–255 ÷ 255 → 0…1 for neural networks.

Min-Max Risk: One Outlier Rescales Everyone

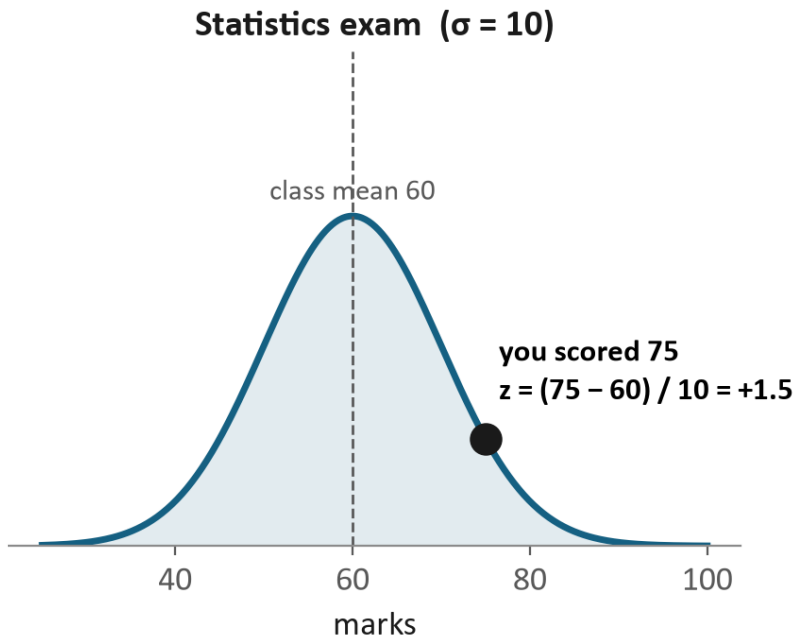


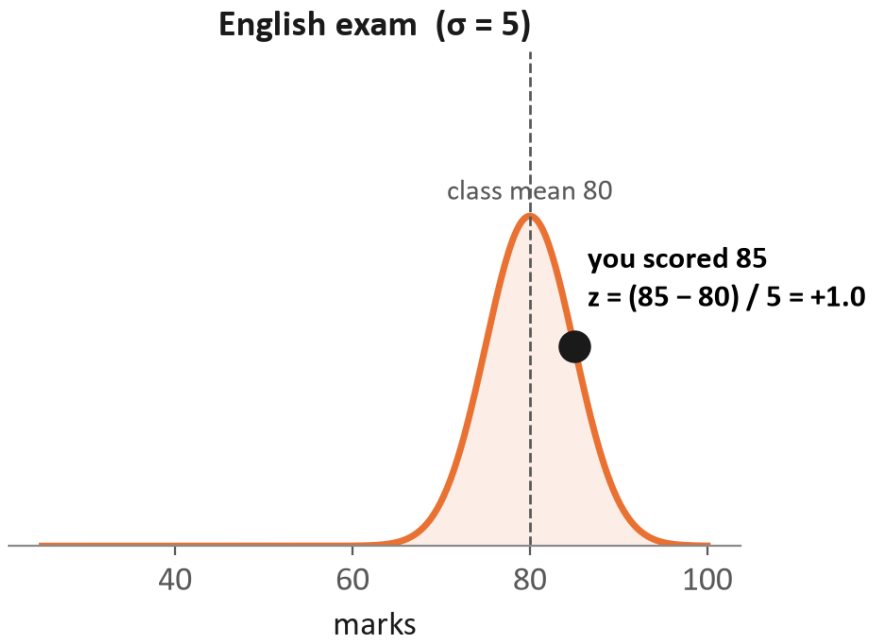
rainfall	max = 1600	max = 3400
1000	0.000	0.000
1200	0.333	**0.083**
1400	0.667	**0.167**
1600	1.000	**0.250**
3400	—	1.000

Farm 1200 did not change
yet its value fell from 0.333 to 0.083

Rule of thumb: use min-max only after outlier handling (Lecture 6), or when the range is fixed by definition (percentages, pixel values 0–255, sensor limits).

What Is a Z-Score? Start With Exams





Definition

A z-score says how many standard deviations a value lies above (+) or below (−) the average.

z	0	+1	−2	beyond ±3
meaning	average	1 σ above	2 σ below	unusual

Z-Score: Reading and Building the Formula

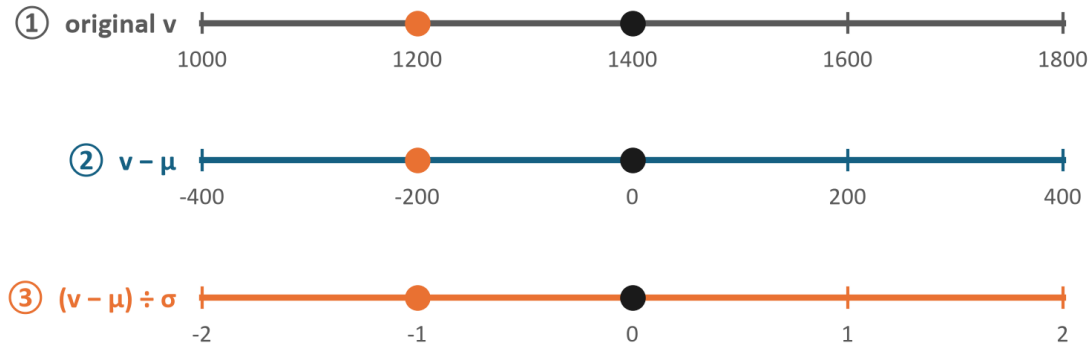
IIT Madras Zanzibar

$$z = (v - \mu) / \sigma$$

result: mean 0, standard deviation 1

Symbol	What it means	In our example
v	original value	1200 mm
μ	mean of the attribute	1400 mm
σ	standard deviation	200 mm
z	new value, in 'number of σ'	?

Step ② moves the average to 0; step ③ re-measures distances in units of σ (μ = 1400, σ = 200)



1 **Shift: v − μ**
1200 − 1400 = −200

2 **Re-measure: ÷ σ**
−200 / 200 = −1.0

Contrast: min-max anchors at the smallest value; z-score anchors at the average value.

Z-Score: Worked Example and Why It's the Default

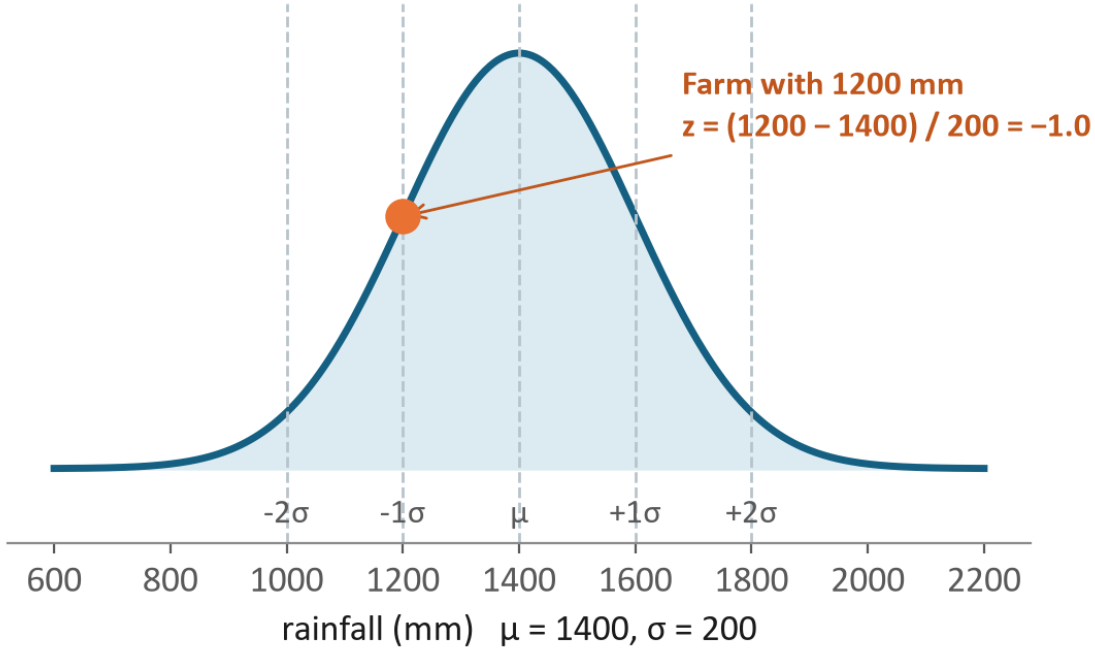
$$z = (v - \mu) / \sigma$$

μ = 1400 mm, σ = 200 mm

1 **Centre**
1200 − 1400 = −200 mm

2 **Scale**
−200 / 200 = −1.0

3 **Say it in words**
one standard deviation below the average farm



Why it's the default: a new extreme value shifts μ and σ only a little, so the scale is stable. Same idea as Pearson's r (Lecture 4) and the z-score outlier rule (Lecture 3). Not bounded to 0…1.

Worked Example: Same Five Farms, Both Methods

IIT Madras Zanzibar

Rainfall of the five farms: $\mu = 1400$ mm, $\sigma = 141.4$ mm (computed earlier), min = 1200, max = 1600, range = 400

Farm	rainfall v	min-max $(v - 1200) / 400$	z-score $(v - 1400) / 141.4$
A	1200	0.00	−1.41
B	1400	0.50	0.00
C	1300	0.25	−0.71
D	1600	1.00	+1.41
E	1500	0.75	+0.71

1. **Order preserved**
   A < C < B < E < D in every column

2. **Different zero**
   min-max: the smallest farm; z: the average farm

3. **Different unit**
   min-max: the whole range; z: one σ

Exam tip: state the statistics you used (min/max or μ/σ), show subtraction and division as separate steps, and sanity-check: the smallest value → 0 (min-max) and the average value → 0 (z-score).

Decimal Scaling: Move the Decimal Point

IIT Madras Zanzibar

$$v' = v / 10^j$$

j = the smallest whole number that makes every |v'| < 1

Trick: count the digits of the largest value → j

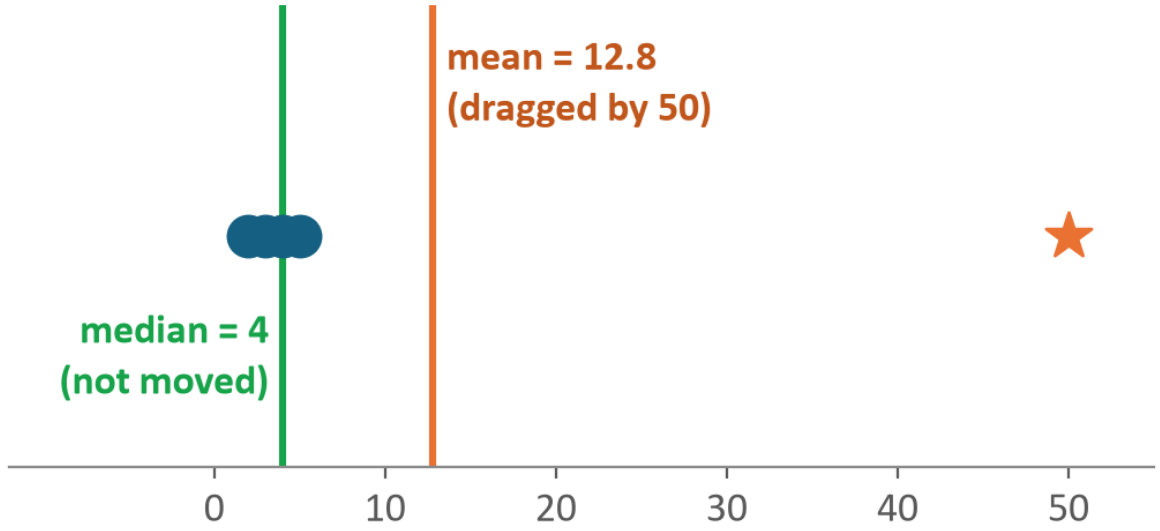
try j	24,179 / 10^j	< 1 ?
3	24.18	no
4	2.42	no
5	**0.242**	**yes ✓**

revenue v ($)	v / 10^5
20.06	0.000201
2,207.37	0.022074
12,271.00	0.122710
24,179.42	0.241794

- Keeps every digit — only the decimal point moves
- Reverse it: multiply by 10^j

In practice: rare in modern pipelines; common in legacy systems and exams. It does not centre the data and is still set by the single largest value.

Robust Z-Score (1 of 2): Why the Mean Can Mislead



1 **MAD by hand**

$|v - 4| = 2, 1, 0, 1, 46 \rightarrow$ sorted 0, 1, 1, 2, 46 $\rightarrow$ middle = 1

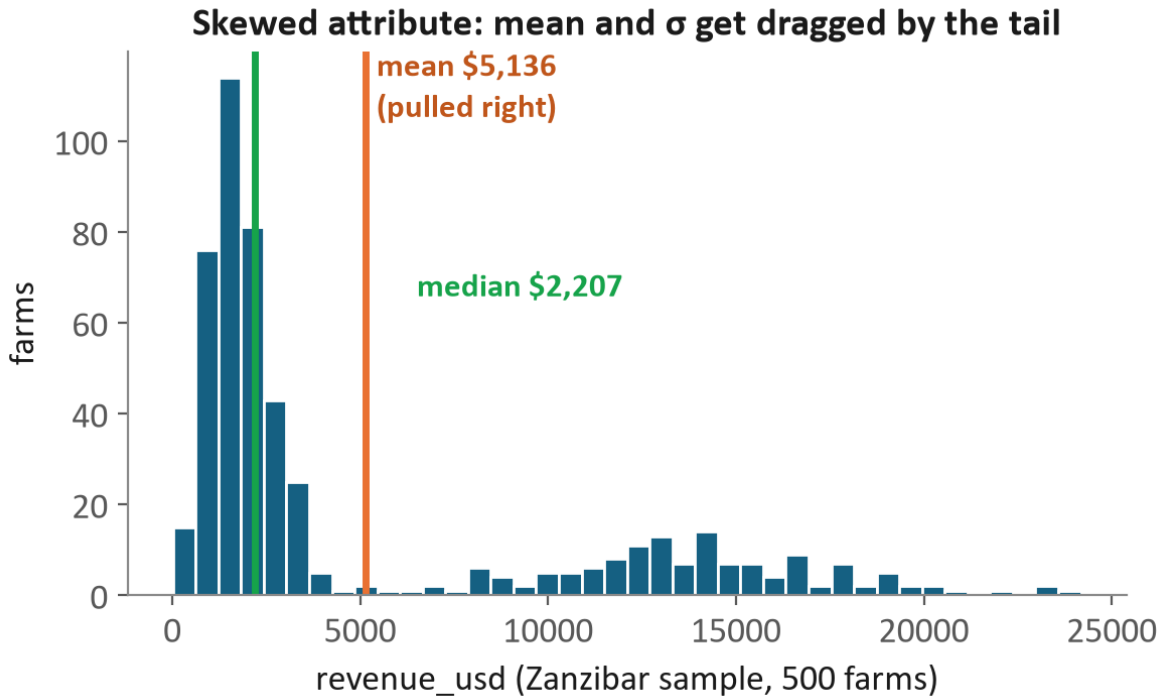
$$M = 0.6745 \cdot (v - \text{median}) / \text{MAD}$$

MAD = median of $|v - \text{median}|$; flag if $|M| > 3.5$

values 2, 3, 4, 5, 50	standard z	robust M
centre	mean 12.8	median 4
spread	σ 18.6	MAD 1
score of 50	2.0 (missed)	**31.0 (flagged)**

Masking: an outlier inflates σ, which shrinks its own z-score. Median and MAD are not dragged by it, so the robust score exposes it.

Robust Z-Score (2 of 2): On Our Revenue Data



Richest farm: revenue = $24,179

	standard z	robust M
centre	mean 5,136	median 2,207
spread	σ 5,745	MAD 984
score	3.31	**15.1**

1 **Standard z**
(24,179 − 5,136) / 5,745 = 3.31

2 **Robust M**
0.6745 × (24,179 − 2,207) / 984 = 15.1

Use it when: the histogram shows a long tail — revenue, income, prices, website visits.

Normalization in the Real World

IIT Madras Zanzibar

1 Child health (WHO)

height-for-age z-score; z < −2 flags stunting

z-score

2 Phone cameras & image AI

pixel 0–255 ÷ 255 → 0…1 for neural networks

min-max

3 Banks & credit scoring

income (TZS) and age (years) standardized for models

z-score

4 Music recommendation

tempo 60–200 bpm vs danceability 0–1 → similar songs

min-max / z

5 IoT & farm sensors

temperature, humidity, soil moisture combined

z-score

6 National examinations

raw marks → standardized scores across papers

z-score

Common thread: whenever different measurements are combined or compared, they are first put on one ruler.

Choosing a Normalization Method

IIT Madras Zanzibar

Method	Formula	Best when	Watch out for
Min-max	$(v - \min)/(\max - \min)$	you need a fixed range, e.g. 0 … 1	one new extreme rescales everything
Z-score	$(v - \mu)/\sigma$	roughly symmetric data; the usual default	μ and σ are pulled by outliers
Robust z (MAD)	$0.6745(v - \text{median})/\text{MAD}$	skewed data or known outliers (revenue)	MAD = 0 if over half the values are identical
Decimal scaling	$v / 10^j$	simplicity; legacy systems	no centring; set by the maximum

Default: z-score for distance-based methods (k-NN, k-means, PCA). Switch to the robust version when a histogram shows a long tail.

Data Transformation: Normalization & Discretization

IIT Madras Zanzibar

Foundations: What, Why, and the Tools We Need

Normalization Methods, Step by Step

→ **Data Leakage & the Evidence**

Discretization, Entropy & Information Gain

Scaling: Normalizing a Stream

In-Class Exercise

Train, Test, and What 'Leakage' Means

IIT Madras Zanzibar

1

Training set

the rows the model learns from — like revision material

2

Test set

rows kept aside to measure performance on unseen data — like the final exam

3

Leakage

any information from the test set reaching training — like seeing the exam paper early

How normalization leaks: computing μ, σ (or min, max) from ALL rows lets the test rows shape the preprocessing.

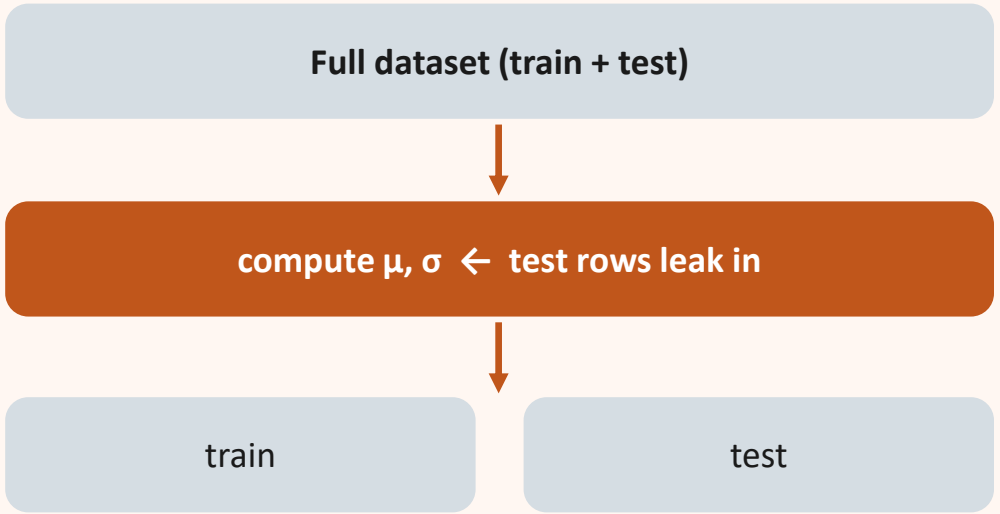
Result: accuracy looks better than it will be on genuinely new farms.

Golden rule: split first; learn every statistic from the training set only; reuse it unchanged on the test set.

Data Leakage: Fit on Train, Apply to Test

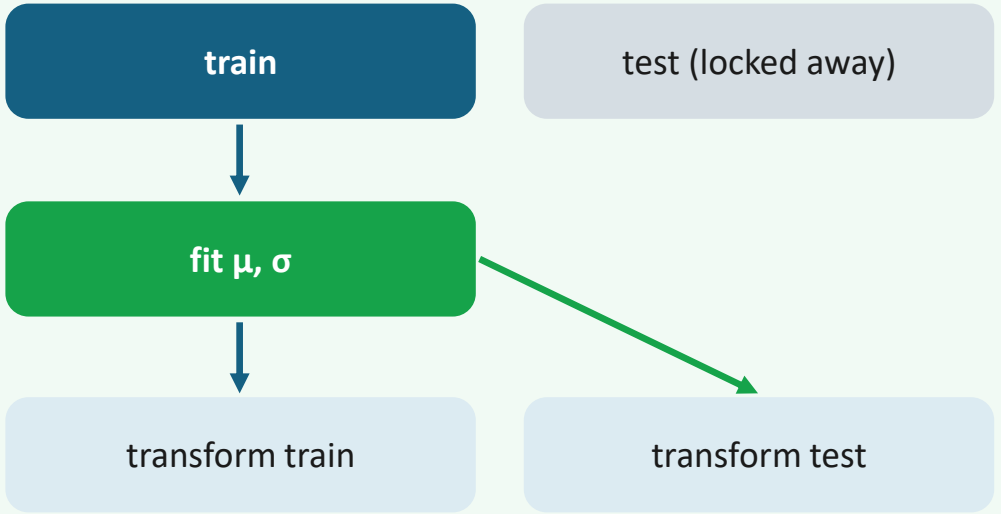
IIT Madras Zanzibar

✗ Wrong: statistics from ALL data



Reported accuracy is too optimistic

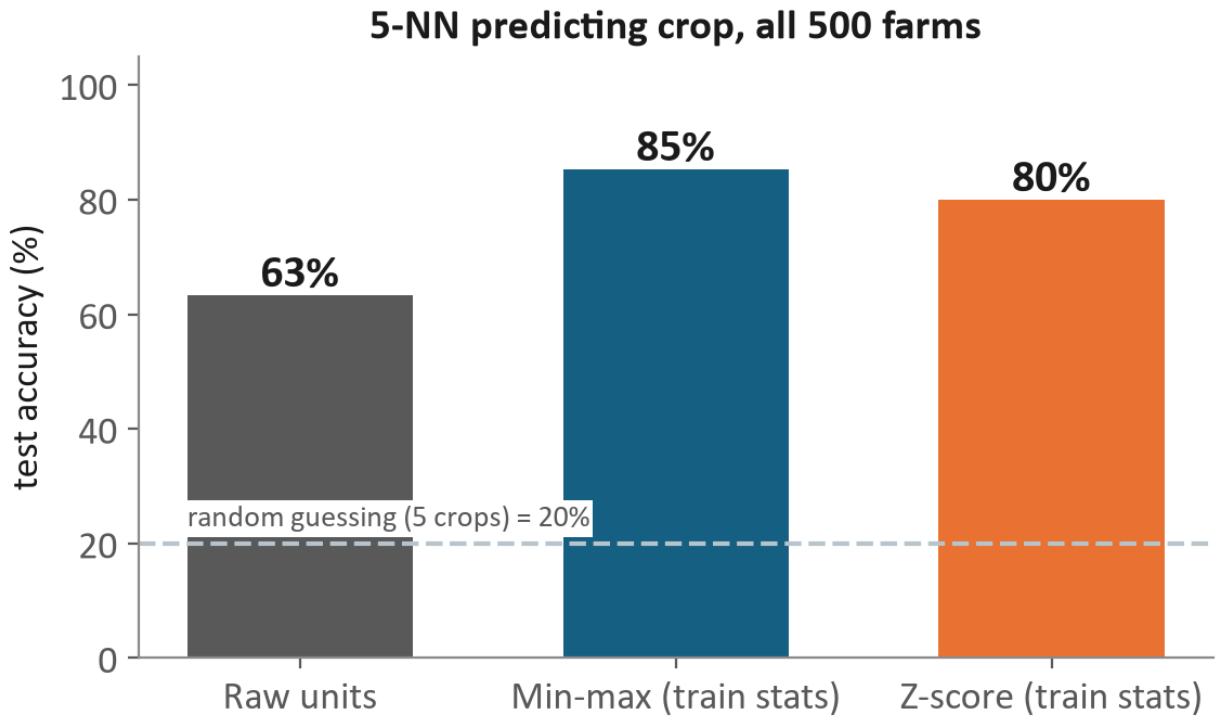
✓ Right: statistics from TRAIN only



Same μ, σ reused — test stays unseen

In scikit-learn: scaler.fit(X_train) once → scaler.transform(X_train) and scaler.transform(X_test). Never fit on X_test.

Evidence: Normalization and k-NN on All 500 Farms



Set-up (Week 1 dataset, all 500 farms)

- 5 numeric attributes → predict crop (5 classes)
- 5-NN: vote of the 5 nearest farms by distance
- 350 train / 150 test; scaler fitted on train only

1 **Raw units: 63% correct**
revenue (≈ 20 – 24,000) swamps price_per_kg (≈ 0.2 – 9)

2 **Scaled: 80 – 85% correct**
+17 to +22 points with the same algorithm

Implication: for any distance-based method (k-NN in Week 6, k-means in Week 8), normalization is a prerequisite, not decoration.

Data Transformation: Normalization & Discretization

IIT Madras Zanzibar

Foundations: What, Why, and the Tools We Need

Normalization Methods, Step by Step

Data Leakage & the Evidence

→ **Discretization, Entropy & Information Gain**

Scaling: Normalizing a Stream

In-Class Exercise

What Is Discretization?

IIT Madras Zanzibar

Definition. Discretization replaces a continuous numeric attribute by a small number of intervals, each with a label.

Everyday example: marks → grades

Mark (%)	0–49	50–59	60–69	70–79	80–100
Grade	F	D	C	B	A

On our data: rainfall → label

rainfall (mm)	< 1250	1250 – 1600	> 1600
label	Low	Medium	High

Unsupervised: ignore labels — equal-width, equal-frequency
Supervised: use labels — entropy / information gain

Purpose

1. **Algorithms need categories**
   e.g. Apriori (Week 5)
2. **Readable rules**
   'IF rainfall = High THEN …'
3. **Less noise, less data**
   many values → one label
4. **Concept hierarchies**
   OLAP roll-up (Week 4)

Also: top-down (split) vs bottom-up (merge, e.g. ChiMerge)

Discretization in the Real World

IIT Madras Zanzibar

1 Health: BMI (WHO)

< 18.5 Underweight · 18.5–24.9 Normal · 25–29.9 Overweight · ≥ 30 Obese

2 Income tax

income → brackets, each with its own rate

3 Marketing

age → 18–24, 25–34, 35–44 …

4 Weather services

rainfall amount → advisory / warning levels

5 Credit ratings

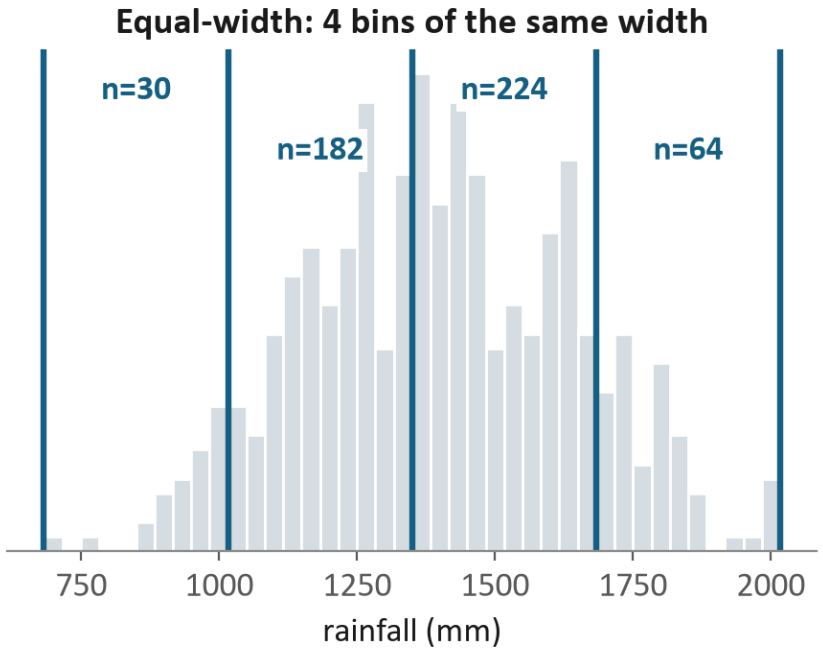
risk assessment → AAA, AA, A, BBB …

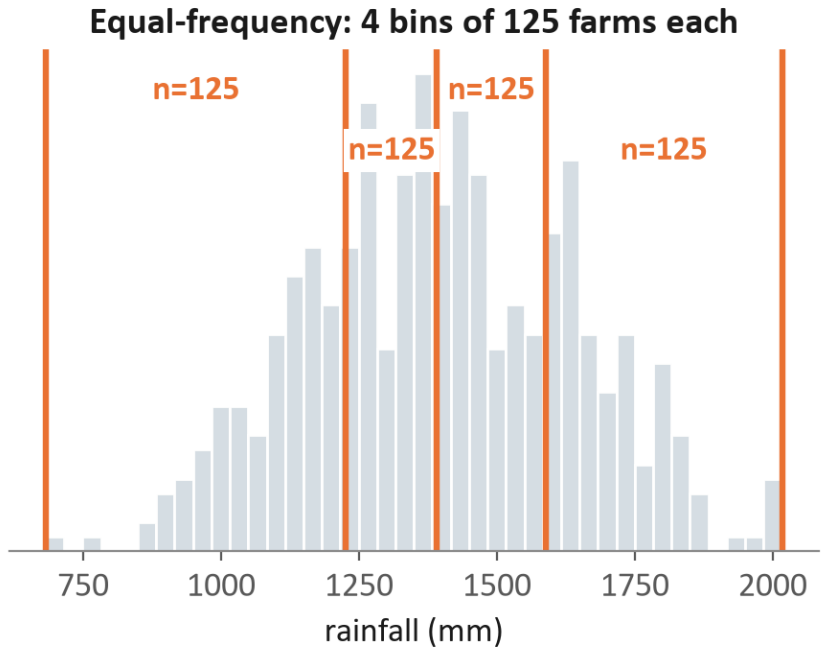
6 Medicine

blood pressure → normal, elevated, hypertension

Common thread: good boundaries separate outcomes that matter. Entropy finds such boundaries from data.

Unsupervised First: Equal-Width vs Equal-Frequency





Equal-width

same interval size; counts can be very unequal

Equal-frequency

same number of farms in each bin

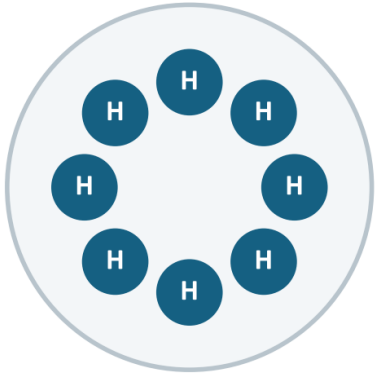
What both ignore

the class label — does the cut separate High from Low yield?

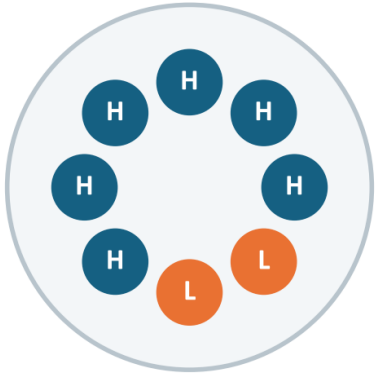
Entropy: How Mixed Is a Group?

IIT Madras Zanzibar

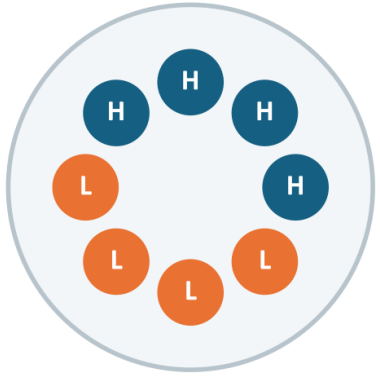
8 High, 0 Low
H = 0 (no surprise)



6 High, 2 Low
H = 0.811 (some surprise)



4 High, 4 Low
H = 1.0 (maximum surprise)



$$H = -\Sigma\, p_i \log_2 p_i$$

p_i = share of the group in class i

p	$\log_2 p$	why
1	0	$2^0 = 1$
0.5	−1	$2^{-1} = ½$
0.25	−2	$2^{-2} = ¼$
0.75	−0.415	calculator

Calculator tip: $\log_2 x = \ln x \div \ln 2$ e.g. $\ln 0.75 \div \ln 2 = -0.2877 \div 0.6931 = -0.415$

In words: entropy = how surprised you expect to be by the label. 0 = pure group; 1 = perfect 50/50 mix (two classes).

Entropy: Worked Example

$$H(S) = -\Sigma_i\, p_i \log_2 p_i$$

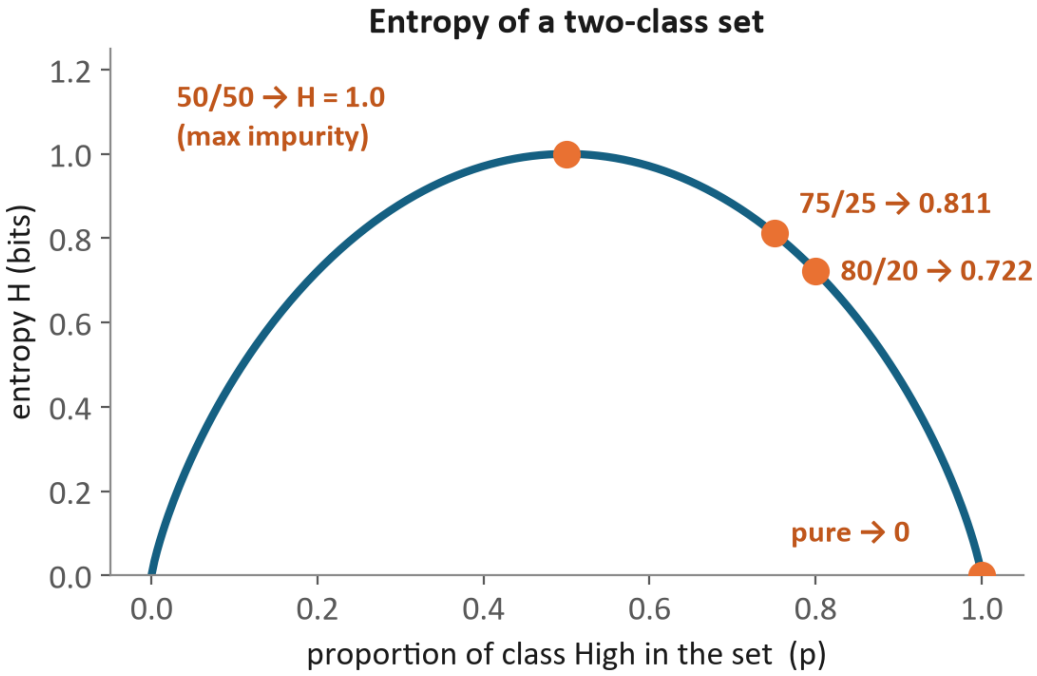
S = the group; p_i = share of class i; $0 \cdot \log_2 0$ counts as 0

A group with 3 Low and 1 High

1 **Proportions**
p(Low) = 3/4 = 0.75, p(High) = 1/4 = 0.25

2 **Logs (table on previous slide)**
$\log_2 0.75 = -0.415$, $\log_2 0.25 = -2$

3 **Multiply, add, flip the sign**
$-(0.75 \times -0.415 + 0.25 \times -2) = 0.311 + 0.500 = 0.811$



Anchor values: pure → 0; 50/50 → 1.0; 75/25 → 0.811; 80/20 → 0.722. Memorize these four.

Information Gain: How Much Does a Split Help?

IIT Madras Zanzibar

$$IG = H(\text{before}) - [\,(n_left / n) \cdot H(\text{left}) + (n_right / n) \cdot H(\text{right})\,]$$

entropy before the split minus the size-weighted entropy after it

Before
one mixed group:
H(before)

→

Split
cut at a boundary
into left and right

→

After
weighted average of
H(left) and H(right)

→

Gain
how much mixing
was removed

Symbol	What it means	In our example
n	number of records in the group	8 farms
n_left / n	share of records on the left side	4/8 or 3/8
IG	entropy removed by the split (bits)	bigger = better split

Worked Example 1: Information Gain at Rainfall = 1350

IIT Madras Zanzibar

rainfall (mm)	1000	1100	1200	1300	1400	1500	1600	1700
yield class	L	L	L	H	H	H	H	L

1 **Before: 4 L, 4 H**
H(before) = 1.000 (50/50)

2 **Left (< 1350): 3 L, 1 H**
H(left) = 0.811

3 **Right (≥ 1350): 3 H, 1 L**
H(right) = 0.811

4 **Weighted after**
(4/8)(0.811) + (4/8)(0.811) = 0.811

5 **Information gain**
IG = 1.000 − 0.811 = 0.189

Both sides still mixed → small gain

Read it: 0.189 bits of uncertainty removed — modest. Can a different boundary do better?

Worked Example 2: Information Gain at Rainfall = 1250

IIT Madras Zanzibar

rainfall (mm)	1000	1100	1200	1300	1400	1500	1600	1700
yield class	L	L	L	H	H	H	H	L

1 **Left (< 1250): 3 L, 0 H**
pure group → H(left) = 0

2 **Right (≥ 1250): 4 H, 1 L**
80/20 → H(right) = 0.722

3 **Weighted after**
(3/8)(0) + (5/8)(0.722) = 0.451

4 **Information gain**
IG = 1.000 − 0.451 = 0.549

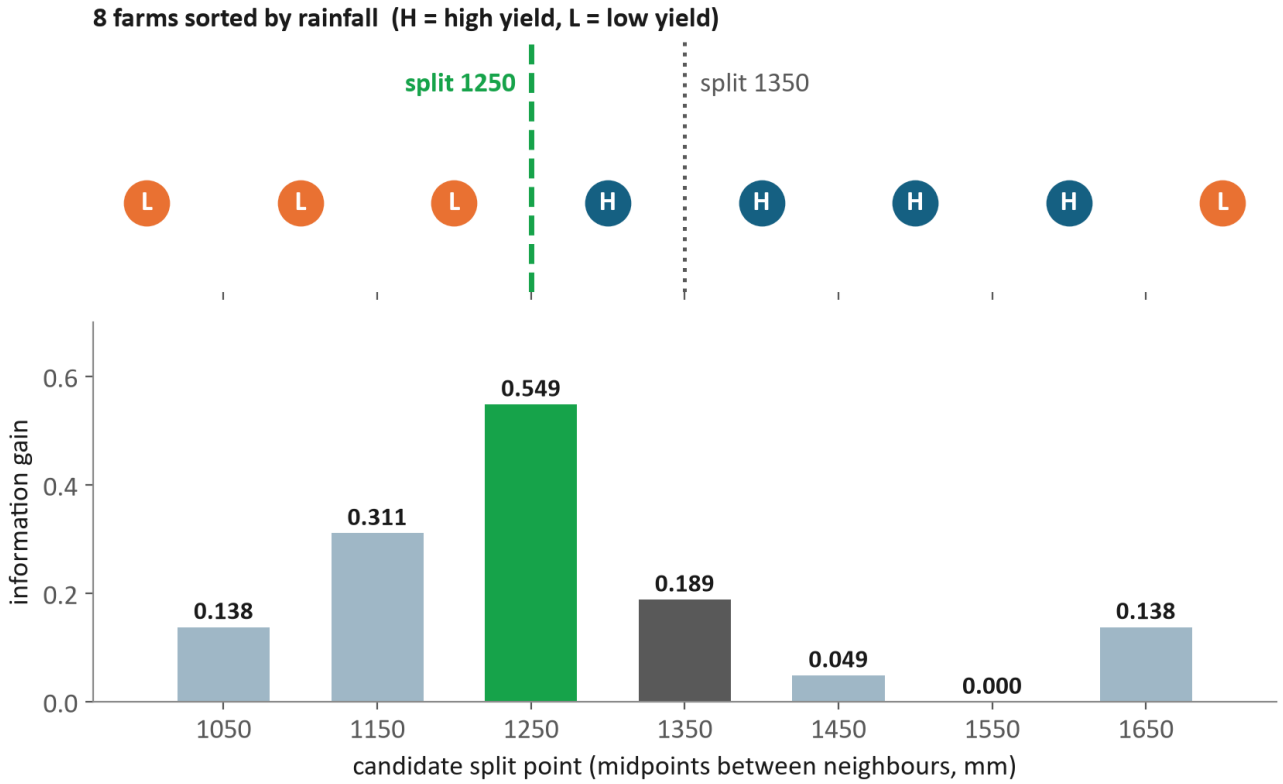
Compare

split 1350 → IG = 0.189

split 1250 → IG = **0.549** (≈ 2.9× better)

Why 1250 wins: it isolates a completely pure group. Purity on one side is worth more than mild improvement on both sides.

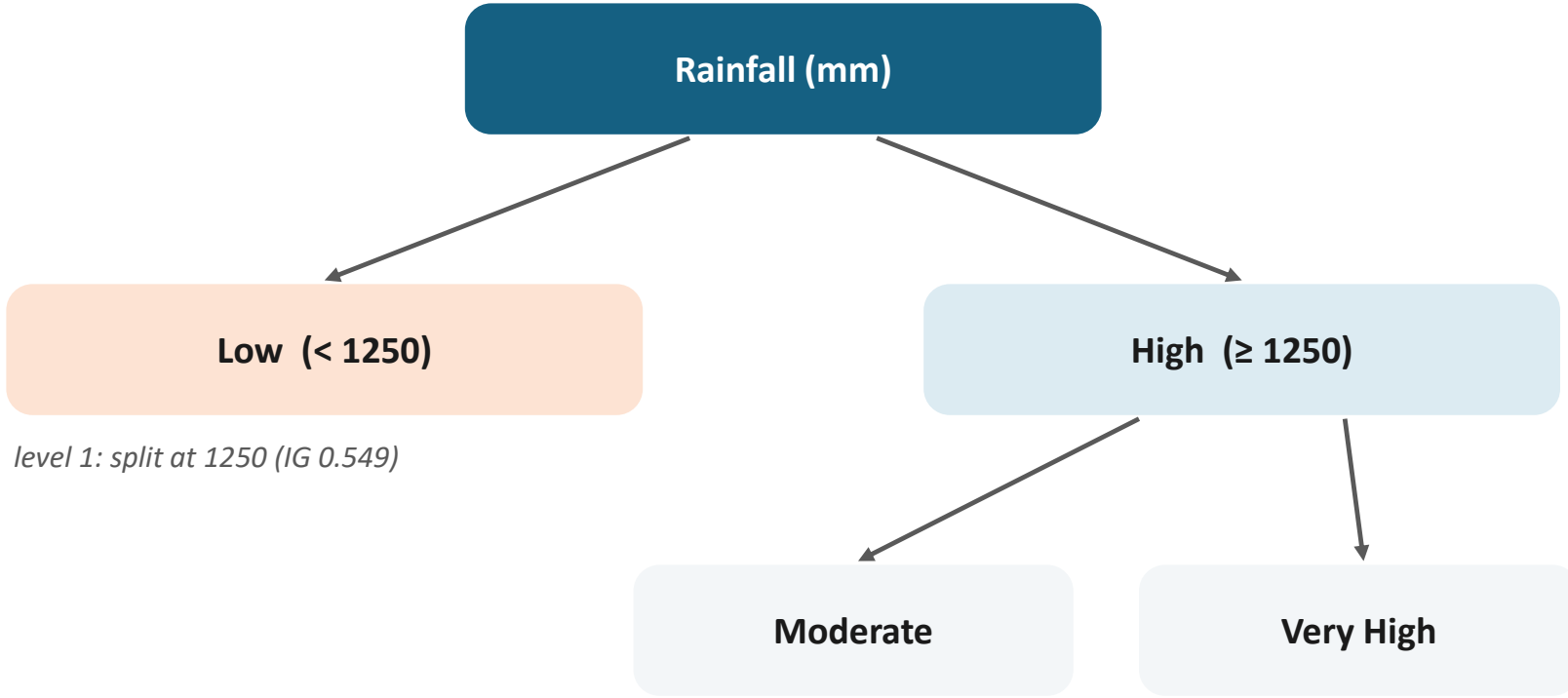
The Algorithm: Try Every Boundary, Keep the Best



1. **Sort** order the values once
2. **Candidates** midpoints between neighbours (n − 1)
3. **Score** information gain for each candidate
4. **Choose** maximum gain → 1250 mm (IG 0.549)

More intervals? repeat inside each side; stop when the gain is no longer worth it (Fayyad & Irani's stopping rule).

From Split Points to a Concept Hierarchy



level 1: split at 1250 (IG 0.549)

level 2: a second entropy split inside 'High'

Roll up ↑
fewer, coarser categories

Drill down ↓
finer categories

Week 4: the same moves on a data cube

Result: one numeric attribute can now be analysed at several levels of detail.

Data Transformation: Normalization & Discretization

IIT Madras Zanzibar

Foundations: What, Why, and the Tools We Need

Normalization Methods, Step by Step

Data Leakage & the Evidence

Discretization, Entropy & Information Gain

→ **Scaling: Normalizing a Stream**

In-Class Exercise

Welford's Algorithm: Normalize a Stream in One Pass

IIT Madras Zanzibar

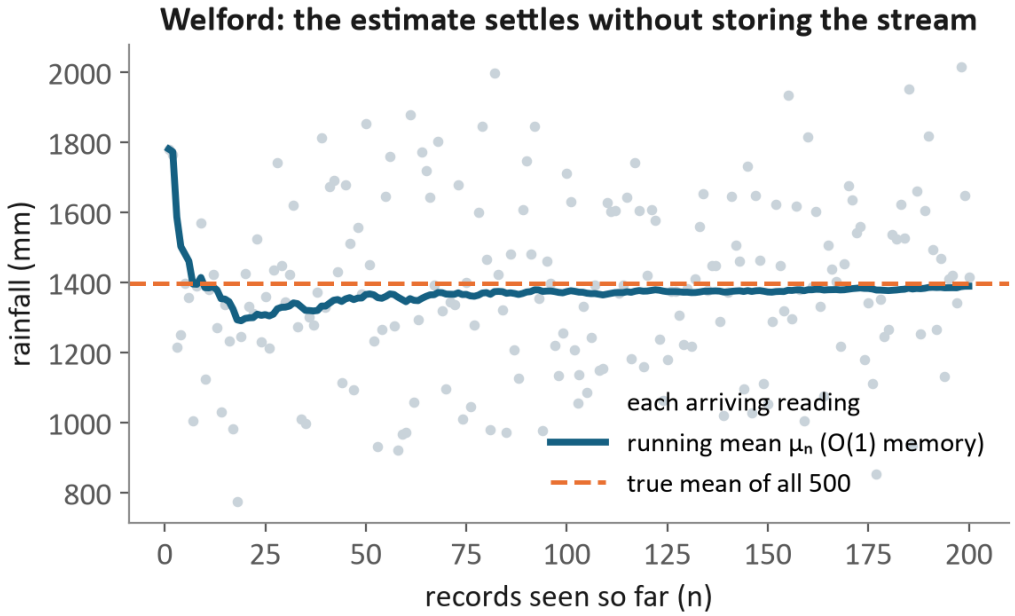
$$\delta = x_n - \mu_{n-1} \quad \mu_n = \mu_{n-1} + \delta / n \quad M_2 \mathrel{+}= \delta\,(x_n - \mu_n)$$

$\sigma^2 = M_2 / n$ at any moment · keeps only 3 numbers · one pass

Trace: readings arrive 1200, 1400, 1300

n	x_n	δ	μ_n	M_2
1	1200	—	1200	0
2	1400	200	1300	0 + 200·100 = 20,000
3	1300	0	1300	20,000 + 0 = 20,000

✓ **σ^2 = 20,000 / 3 = 6,667 → σ = 81.6**
same as the long way



Streaming z-score: $z = (x_n - \mu_n) / \sigma_n$ using the current estimates — no second pass.

Real use: weather stations, mobile-money fraud checks, IoT sensors — anywhere data never stops.

Data Transformation: Normalization & Discretization

IIT Madras Zanzibar

Foundations: What, Why, and the Tools We Need

Normalization Methods, Step by Step

Data Leakage & the Evidence

Discretization, Entropy & Information Gain

Scaling: Normalizing a Stream

→ **In-Class Exercise**

In-Class Exercise (10 minutes, pairs)

IIT Madras Zanzibar

IN-CLASS TASK

farm_size_ha in a sample: min = 0.2, max = 4.5, μ = 1.4, σ = 0.9

Labelled farms (yield class): 0.3 L · 0.5 L · 0.8 H · 1.0 H · 1.5 H · 2.0 H · 2.5 L · 3.0 L

1. Min-max normalize farm_size = 2.0 to the range 0 … 1.
2. Z-score normalize the same value and say in words what the answer means.
3. Compute the entropy of the 8 labelled farms, then the information gain of a split at farm_size = 1.25.
4. Your gain in Q3 should surprise you. Find a better split point and compute its gain.

Solution

IIT Madras Zanzibar

1 Min-max
(2.0 − 0.2) / (4.5 − 0.2) = 1.8 / 4.3 = 0.419

2 Z-score
(2.0 − 1.4) / 0.9 = 0.667 → ⅔ σ above average

3 Entropy before
4 H, 4 L → H = 1.000

4 Split at 1.25
left 0.3–1.0: 2 L, 2 H (H = 1); right 1.5–3.0: 2 H, 2 L (H = 1)

5 Gain
1 − (4/8·1 + 4/8·1) = **0.000** → useless split

6 Better: split at 0.65 (or 2.25)
left: 2 L pure; right: 4 H, 2 L → H = 0.918

IG(0.65) = 1 − (2/8 · 0 + 6/8 · 0.918) = 1 − 0.689 = 0.311

Lesson: a boundary in the middle of the range can carry zero information. Only the class labels tell you where a useful boundary is — that is why supervised discretization exists.

Summary & Exam Checkpoints

IIT Madras Zanzibar

Normalize before distance

Raw units let the biggest numbers dominate. On all 500 farms, k-NN rose from 63% to 80–85%.

Match method to the data

Min-max: fixed range. Z-score: the default. Robust z: long tails and outliers.

Fit on train only

Learn min/max or μ/σ from the training set; reuse them on test. Anything else is leakage.

Entropy picks boundaries

IG = H(before) − weighted H(after). Try every midpoint; keep the maximum. Returns in Week 7.

Common Questions

IIT Madras Zanzibar

Q **Do I normalize categorical attributes like crop or region?**

A No. Normalization is for numbers. Categories are encoded differently (e.g. one-hot) — Week 6.

Q **Does normalization change the correlation between two attributes?**

A No. Pearson's r already divides by σ, so linear rescaling leaves it unchanged.

Q **Do decision trees need normalization?**

A No — a tree compares one attribute with a threshold at a time, so scale never mixes. Distance methods do need it.

Q **Is entropy-based discretization the same as a decision tree?**

A It is the same calculation on one attribute. A tree repeats it across all attributes, level by level (Week 7).

How This Feeds Your Assessment — and What's Next

IIT Madras Zanzibar

Assessment

Quiz 1: normalize a value by hand; compute entropy and information gain for a split

Mini-project: justify your normalization choice and fit it on training data only (graded step)

Next: Lecture 8 (Thursday)

Data reduction: too many rows → sampling; too many columns → PCA

Tutorial 3

Normalization, discretization and PCA on the Zanzibar data — Mini-Project announced

Before Thursday: revise covariance (Lecture 4). PCA is built directly on it.

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed. Morgan Kaufmann — Section 2.5 (Data transformation).
- Fayyad, U. M. & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. Proc. IJCAI, 1022–1027.
- Iglewicz, B. & Hoaglin, D. C. (1993). How to Detect and Handle Outliers. ASQC Quality Press. (Modified z-score, 3.5 threshold.)
- Welford, B. P. (1962). Note on a method for calculating corrected sums of squares and products. Technometrics, 4(3), 419–420.
- Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379–423. (Entropy.)
- World Health Organization (2006). WHO Child Growth Standards. (Height-for-age z-scores.)