

Patterns can mislead.

Exploratory data analysis, confounding and Simpson's paradox

SAME DATA, DIFFERENT VIEW

OVERALL



WITHIN GROUPS



Our path from Week 4

Preparation enables analysis; it does not answer the question for us.

WEEK 03

Clean the data

Correct verified errors; keep uncertainty visible.



WEEK 04

Transform safely

Scale and encode variables; prevent future-data leakage.



WEEK 05

Investigate

Compare groups, inspect patterns, and test assumptions.

A clean table and correct formula do not automatically give a meaningful comparison.

A surprising comparison

We will explain this result with the very same synthetic course records.

FULL-TIME REGISTRATIONS

74.5%



3,059 passes / 4,108 registrations

PART-TIME REGISTRATIONS

77.0%



429 passes / 557 registrations

But full-time can lead inside course groups. How?

By the end, you can answer four questions

Each question builds on the one before it.

01 What is the right unit?

Distinguish students from registrations.

02 What does the distribution show?

Identify centre, spread and shape.

03 What changes inside groups?

Compare conditional and pooled rates.

04 What can we conclude?

Separate association from causal claims.

EDA is a cycle of questions and checks

A useful workflow begins with the question, not with a plot gallery.



First decision: what does one row represent?

The correct denominator depends on the table's unit.

students.csv

1 student

1,200 rows

Student-level age and mode

enrolments.csv

1 course registration

4,665 rows

One student may appear again

activity.csv

1 student-week record

16,165 rows

Repeated measurements over time

Today's main pass-rate unit: ONE REGISTRATION

Two students can produce four registrations

One fictional student takes three courses; another takes one.

Student	Course	Pass?
A	C1	Yes
A	C2	Yes
A	C3	Yes
B	C1	No

75%

3 passes / 4 registrations

50%

1 student passed ≥ 1 course / 2 students

A percentage must name its denominator

Formula: pass rate = passes ÷ eligible registrations × 100%.

$$\text{Pass rate} = \frac{\text{number of passes}}{\text{number of eligible registrations}} \times 100\%$$

BAD REPORT

"429 people passed."

DEFENSIBLE REPORT

"429 of 557 registrations passed."

Start with a small example: counts are not rates

Both groups take the same test. Which one performs better?

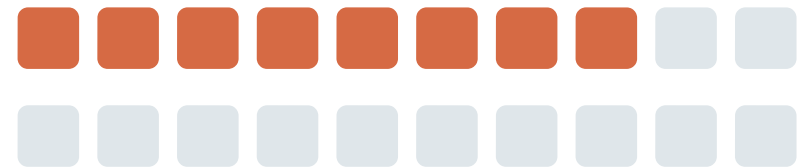
GROUP A



6 passes / 10 students

60%

GROUP B



8 passes / 20 students

40%

8 passes > 6 passes, but 40% < 60%. Always compare like denominators.

Question 1 • Which pass rate is higher?

Calculate both rates before choosing an answer.

Study mode	Passes	Registrations
Full-time	3,059	4,108
Part-time	429	557

A Full-time B Part-time C Same rate

Answer 1 • Part-time is higher overall

The comparison is correct for the pooled registration data.

FULL-TIME

74.5%



3,059 passes / 4,108 registrations

PART-TIME

77.0%



429 passes / 557 registrations

Difference $\approx$ 2.6 percentage points, not 2.6% relative change.

A numerical difference is a starting point

Before interpreting the gap, list alternatives that could explain it.

- **Different courses** Registration mix may differ by study mode.
- **Different starting levels** Prior preparation can influence passing.
- **Incomplete outcomes** Unknown outcomes can change denominators.

EDA asks “what else could produce this pattern?” before “what caused it?”

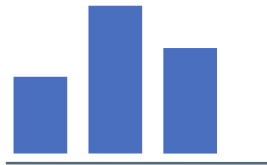
Pick a plot that matches your question

Categories and numbers need different visual summaries.

ONE CATEGORY

Bar chart

How many students per programme?



ONE NUMBER

Histogram

How are exam scores distributed?



TWO NUMBERS

Scatter plot

Do preparation and grade move together?



TWO CATEGORIES

Group percentages

Does pass rate change by mode?



How eight scores become a histogram

We count values in score ranges. Each bar is the count in one bin.

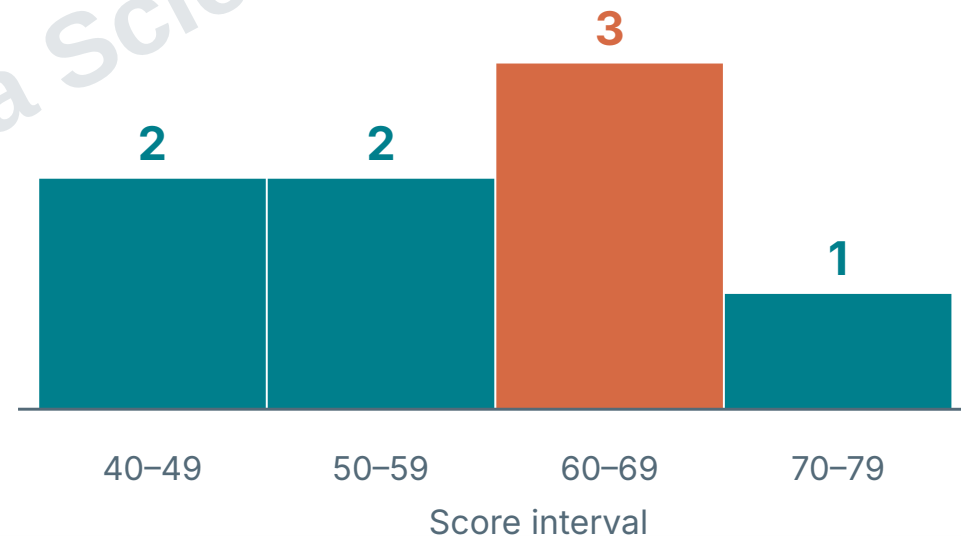
EIGHT GRADE VALUES

42 · 47 · 53 · 58

61 · 64 · 65 · 78

One number = one exam score

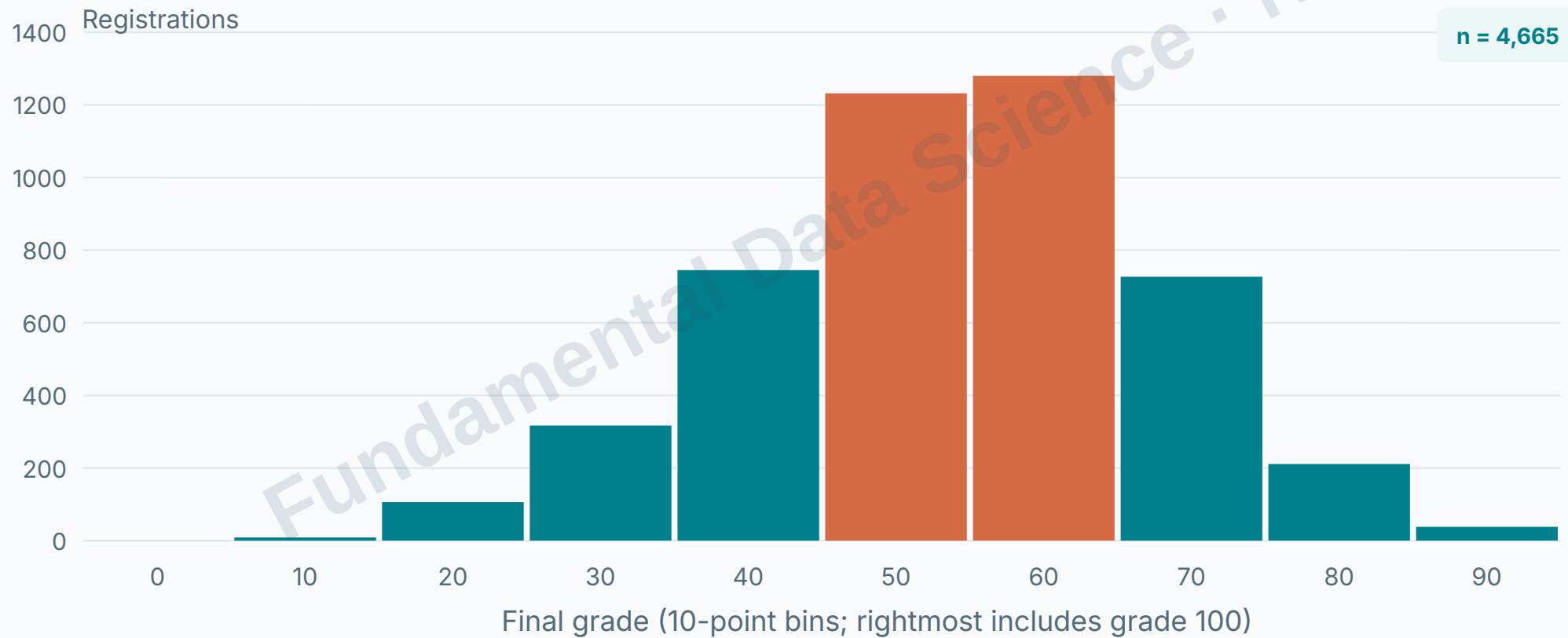
COUNT IN EACH BIN



Check: $2 + 2 + 3 + 1 = 8$ scores. Bars touch because the bins are consecutive.

A histogram shows where grades occur

The horizontal axis is the grade; the vertical axis counts registrations.



Read the histogram in four steps

A plot should produce a defensible written observation.

01

AXES

Horizontal: grade. Vertical: count.

02

DENOMINATOR

All 4,665 course registrations.

03

SHAPE

Most values cluster near 50–70.

04

WHAT WE CANNOT SAY

The graph does not explain why grades differ.

50–69.999... bins: $1,232 + 1,280 = 2,512$ registrations (53.8%).

The mean can be pulled by one large value

Constructed illustration: five weekly login counts.



10

Mean: $50 \div 5$

3

Median: middle value

Both are correct; they answer different “typical value” questions.

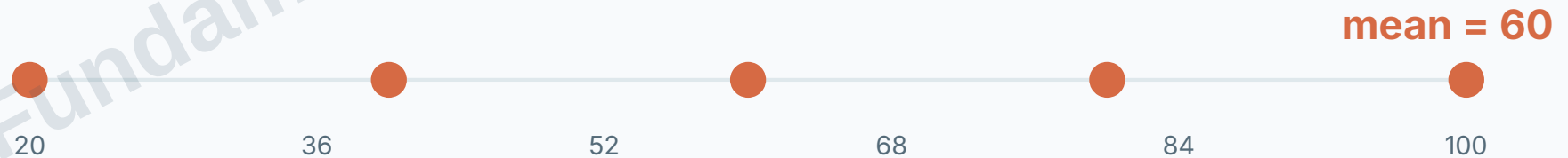
Equal averages can conceal different spread

Each row is a constructed group of five exam scores.

GROUP A



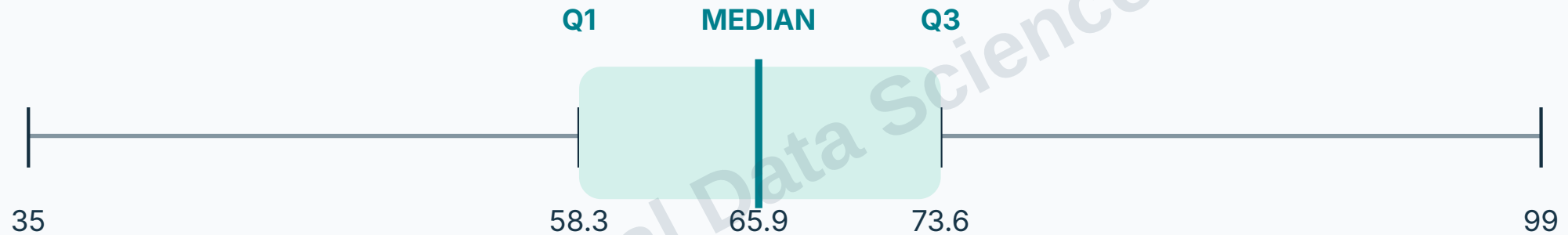
GROUP B



Group B is far more variable even though both groups have mean 60.

Quartiles describe the middle half

Entry qualification scores for 1,200 synthetic students.



15.3

Interquartile range: $73.6 - 58.3$

64

Full range: $99 - 35$

Unknown outcomes create denominator choices

Constructed cohort: 100 registrations, of which 20 have no recorded result.

Recorded outcome	Registrations
Pass	60
Fail	20
Unknown	20

75%

60 / 80 known outcomes

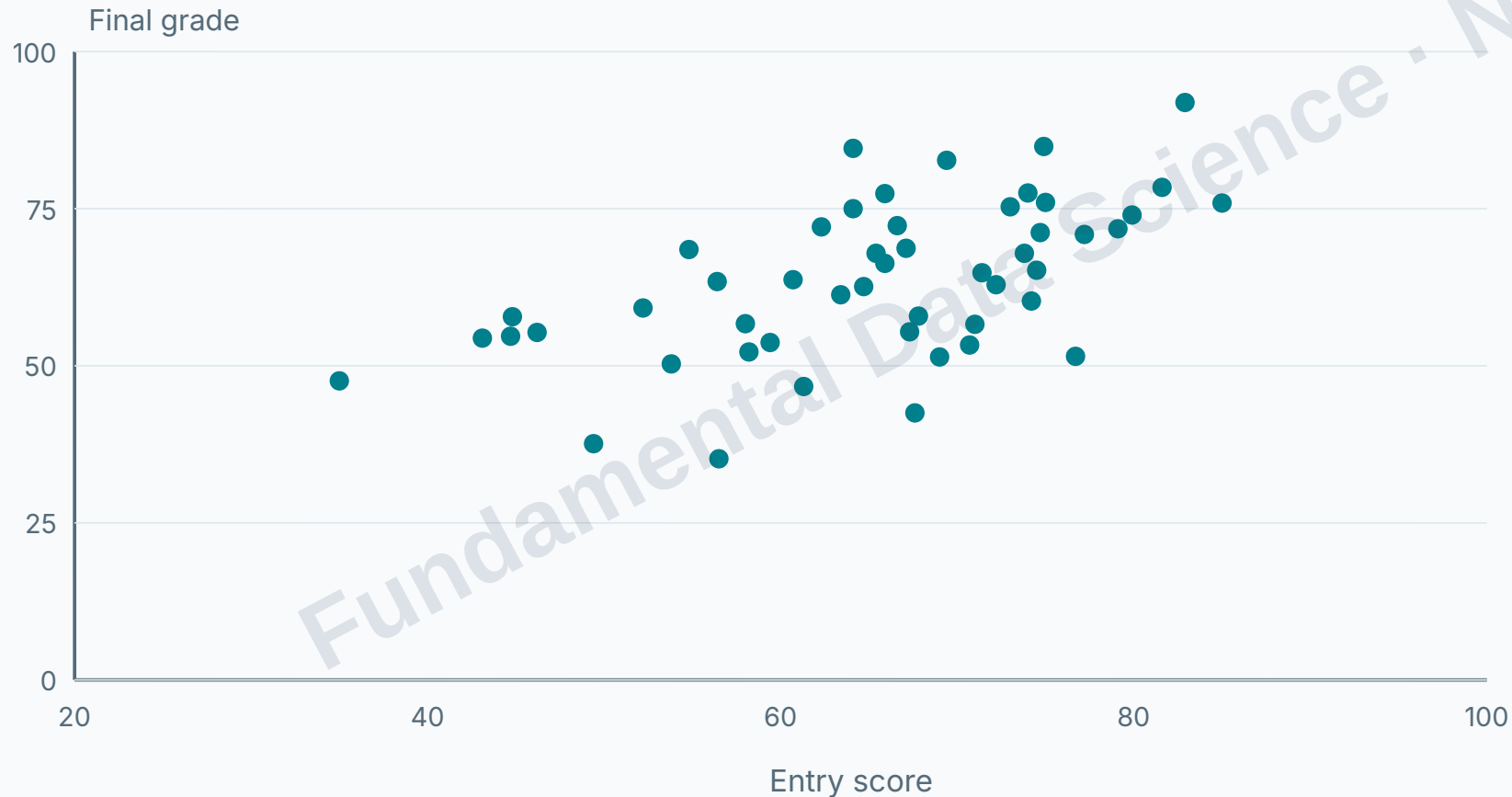
ALL 100 REGISTRATIONS

60% to 80%

The 20 unknown outcomes do not become failures by default.

A scatter plot shows two numerical variables

Points are 50 selected records from the synthetic DS100 student table.



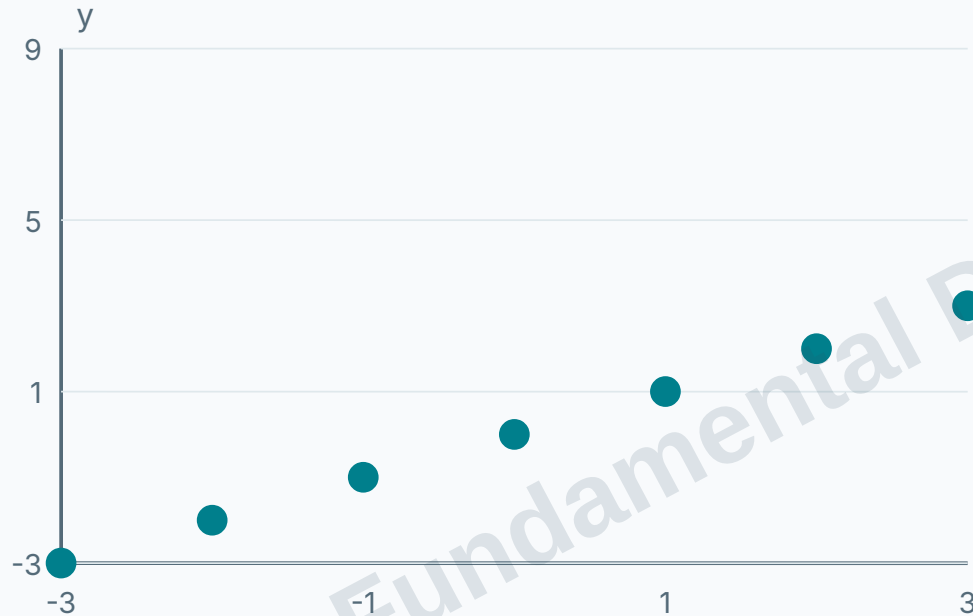
**Positive
association**

Both variables
still vary widely.

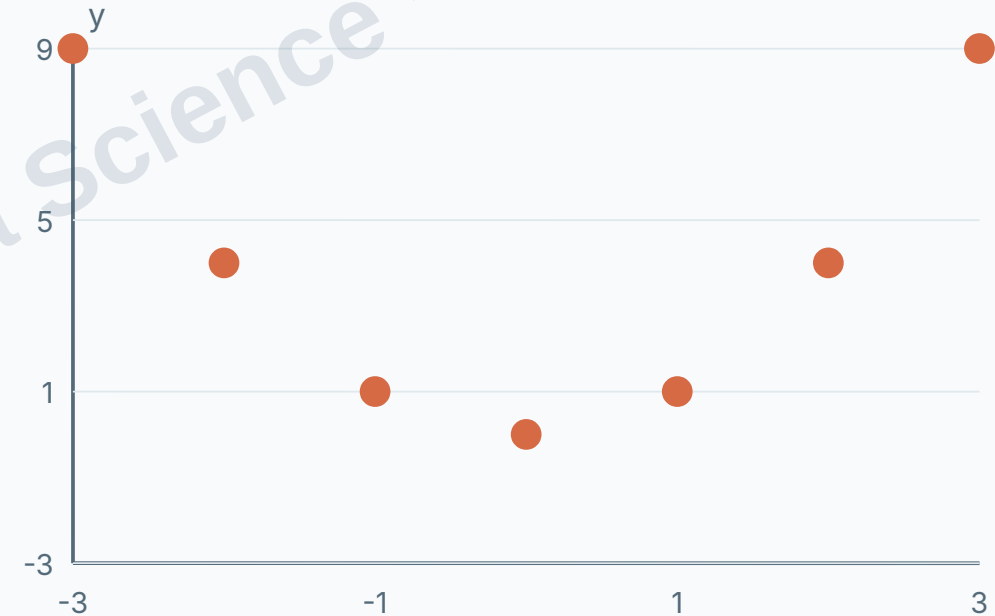
Illustrated points = 50 selected students; full 1,200-student correlation $\approx +0.48$.

A correlation coefficient can miss a curve

Constructed points: the second set has a strong U-shape but linear $r = 0$.



Linear pattern $r = +1$



Curved pattern $r = 0$

Describe a plot without overclaiming

A three-sentence template students can use for almost any EDA result.

1

OBSERVATION

Among 4,665 registrations, part-time has the higher pooled pass rate.

2

COMPARISON

The gap is about 2.6 percentage points using registration denominators.

3

LIMITATION

The result may change after comparing similar courses.

Before pooling groups, ask what differs

Different course choices could make overall rates hard to interpret.

Study mode

Who is being compared?

Course difficulty

Which subjects are included?

Grade threshold

What counts as passing?

Missing outcomes

Whose result is observed?

Next: the mathematical reason two correct group rates can reverse when combined.

10-minute break

When we return: a correct average will produce the opposite conclusion.

BEFORE YOU RETURN

Predict why the group mixture could matter.

Part II · A 200-student support programme

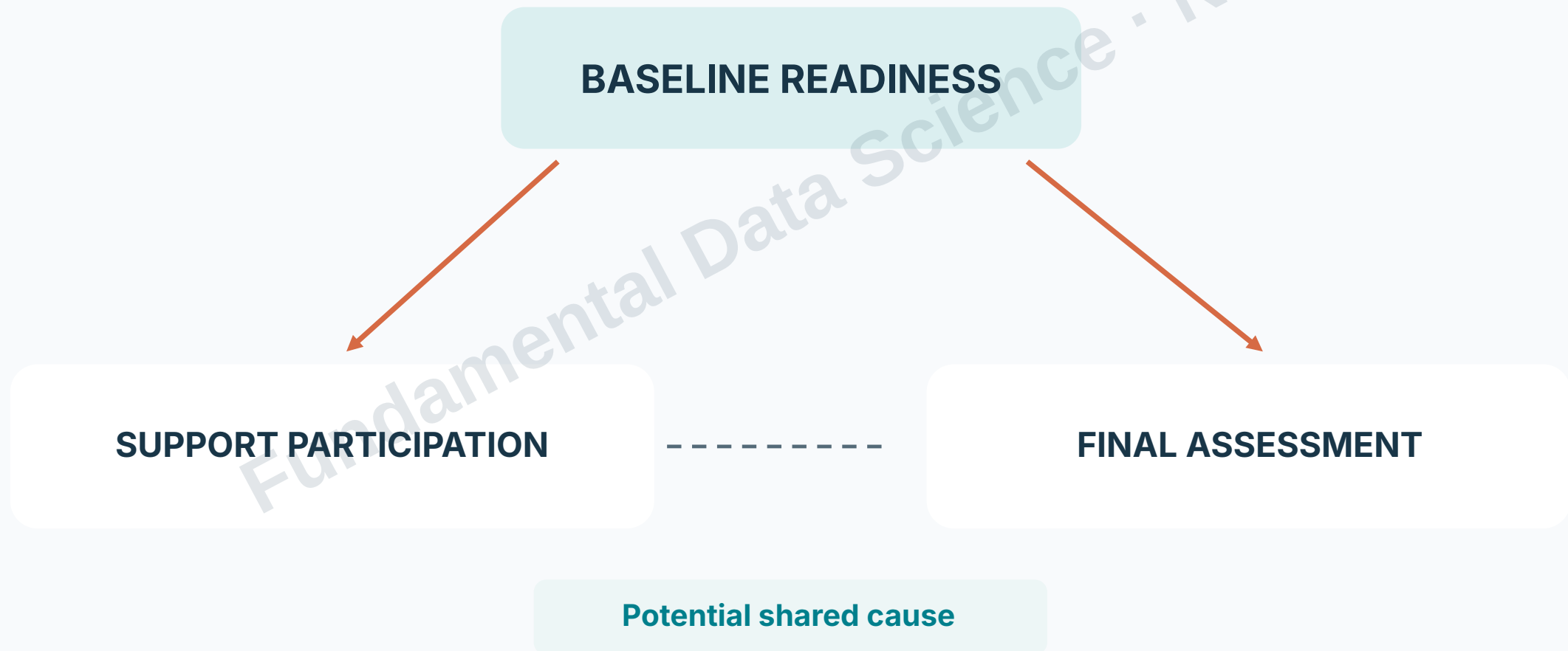
Small invented numbers let us see Simpson's paradox by hand.



This is an observational comparison: participation is NOT randomly assigned.

Readiness may influence both variables

This is a plausible explanation, not a proven causal diagram.



Question 2 • Calculate the rates in each group

Every cell is “passed / total students”. Work within each readiness level first.

Readiness	Support	No support
Lower	48 / 80	10 / 20
Higher	18 / 20	64 / 80
All students	? / 100	? / 100

Within lower readiness, who leads? Within higher readiness, who leads?

Answer 2a · Support leads within both levels

At matched recorded readiness, the observed gaps are +10 percentage points.

LOWER READINESS

Support



60%

No support



50%

48/80 vs 10/20

HIGHER READINESS

Support



90%

No support



80%

18/20 vs 64/80

Answer 2b · Pooling reverses the result

Now combine the numerators and denominators before dividing.

SUPPORT

$48 + 18 = 66$ passes

$80 + 20 = 100$ students

66%

NO SUPPORT

$10 + 64 = 74$ passes

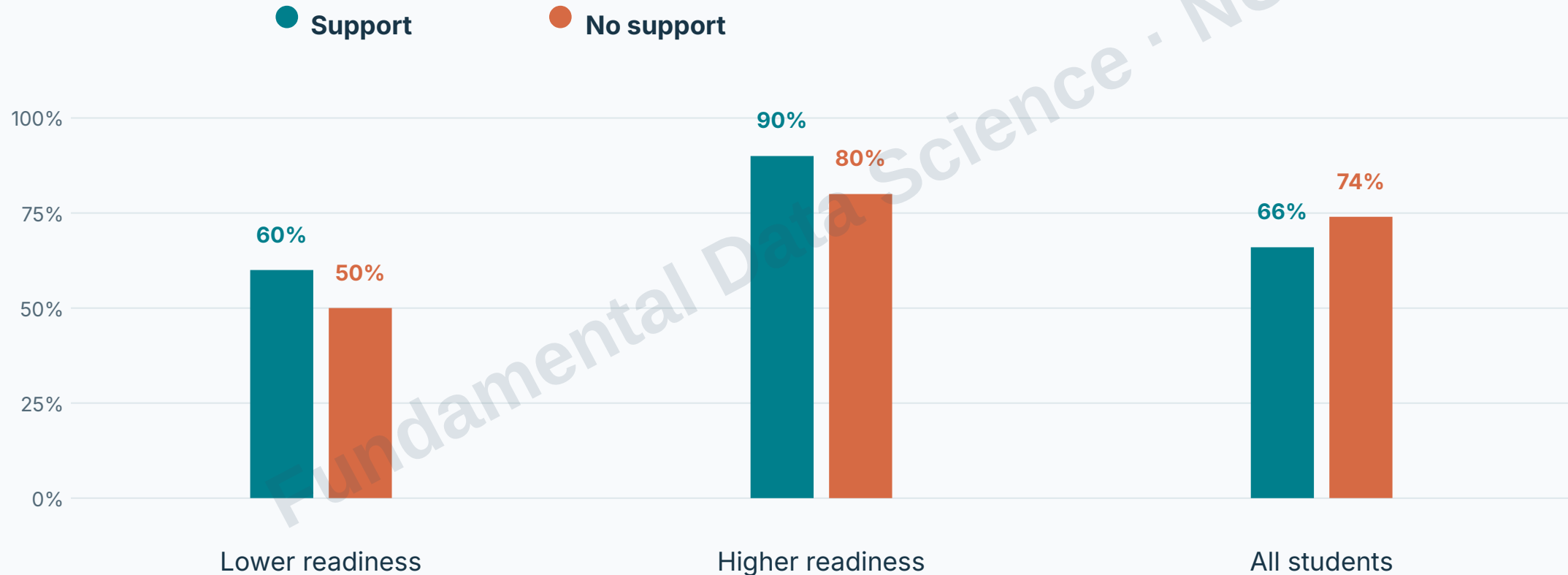
$20 + 80 = 100$ students

74%

Within both groups: support leads. Overall: no support leads.

Three comparisons, two different conclusions

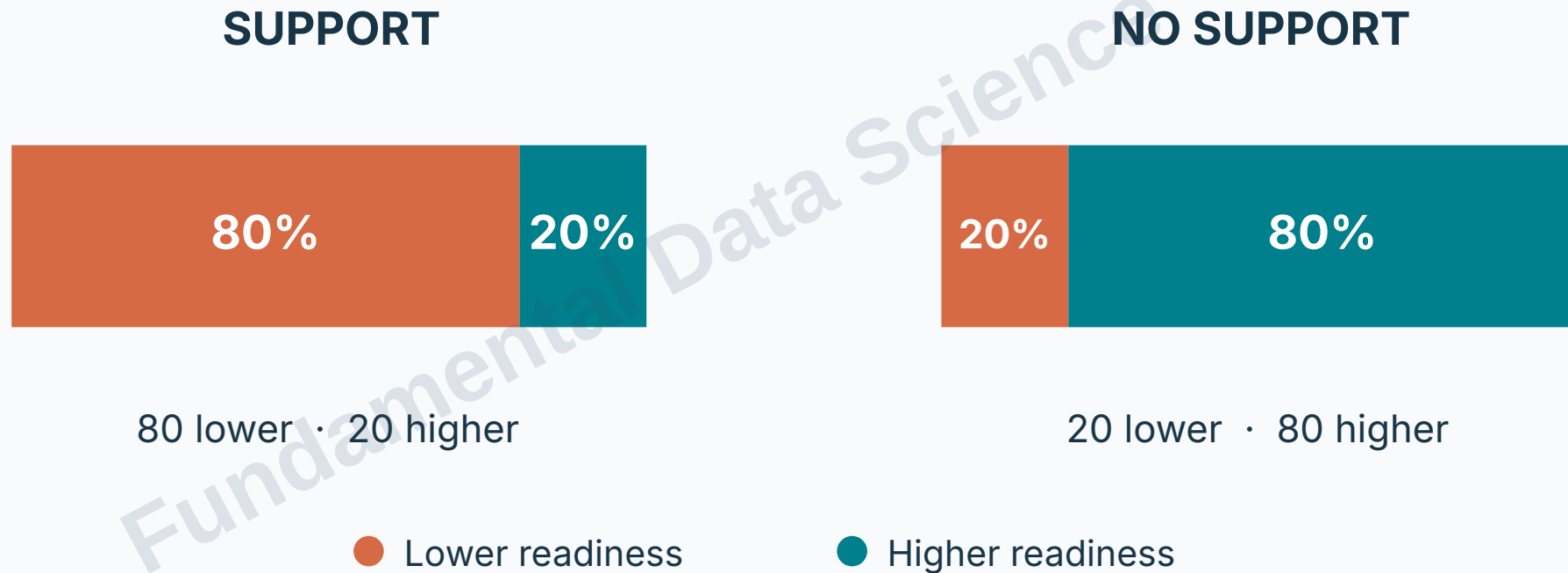
The subgroup bars and pooled bars use different mixtures.



Same 200 students. Different group weights.

The groups have opposite readiness mixes

Share of lower versus higher readiness among each participation group.



Different composition changes the weight given to each stratum's pass rate.

The overall rate is a weighted average

Every subgroup contributes its size $\times$ its observed rate.

$$\text{Overall pass rate} = w_1 \times p_1 + w_2 \times p_2$$

- w_1, w_2 Group shares: n_1/n and n_2/n
- p_1, p_2 Rates inside each group
- $w_1 + w_2$ Must equal 1

Why? passes = $n_1p_1 + n_2p_2$; divide by $n_1 + n_2$.

Why not simply average 60% and 90%?

The two readiness groups do not contain the same number of students.

WRONG: GIVE BOTH GROUPS EQUAL WEIGHT

$$(60\% + 90\%) / 2$$

75%

Assumes a 50:50 group mix

CORRECT: USE THE ACTUAL 80:20 MIX

$$0.80 \times 60\% + 0.20 \times 90\%$$

66%



80 lower

20 higher

Weights = subgroup size ÷ group total. Use the observed 80/100 and 20/100.

The two weighted calculations side by side

Different weights explain the 66% versus 74% reversal.

SUPPORT: 80 LOWER, 20 HIGHER

$$0.80 \times 60\% + 0.20 \times 90\%$$

$$48\% + 18\%$$

$$= 66\%$$

NO SUPPORT: 20 LOWER, 80 HIGHER

$$0.20 \times 50\% + 0.80 \times 80\%$$

$$10\% + 64\%$$

$$= 74\%$$

At the same mix, the ranking changes again

Illustrative direct standardization: use 50% lower, 50% higher for both groups.

SUPPORT • COMMON 50:50 MIX

75%

$$0.50 \times 60\% + 0.50 \times 90\%$$

NO SUPPORT • COMMON 50:50 MIX

65%

$$0.50 \times 50\% + 0.50 \times 80\%$$

These are hypothetical standardized rates, not observed new data.

Simpson's paradox, in one sentence

A pooled association can run in the opposite direction to the association within groups.

WITHIN LOWER: 60% > 50% WITHIN HIGHER: 90% > 80%

POOLED: 66% < 74%

No contradiction: the two pooled groups contain different readiness mixes.

Now return to our actual course example

All figures below come from the 4,665 synthetic course registrations.

POOLED PASS RATE

Full-time

74.5%

Part-time

77.0%

Next question:

What if we compare similar types of courses first?

Which courses are we comparing?

We grouped twelve courses into two broad categories for this EDA check.

THREE QUANTITATIVE SERVICE COURSES

MA110 · CH110 · ST110

Lower pass-rate course group

NINE OTHER COURSES

Other nine course codes

Higher pass-rate course group

The question: does full-time versus part-time look the same within each course group?

Inside both broad course groups, full-time leads

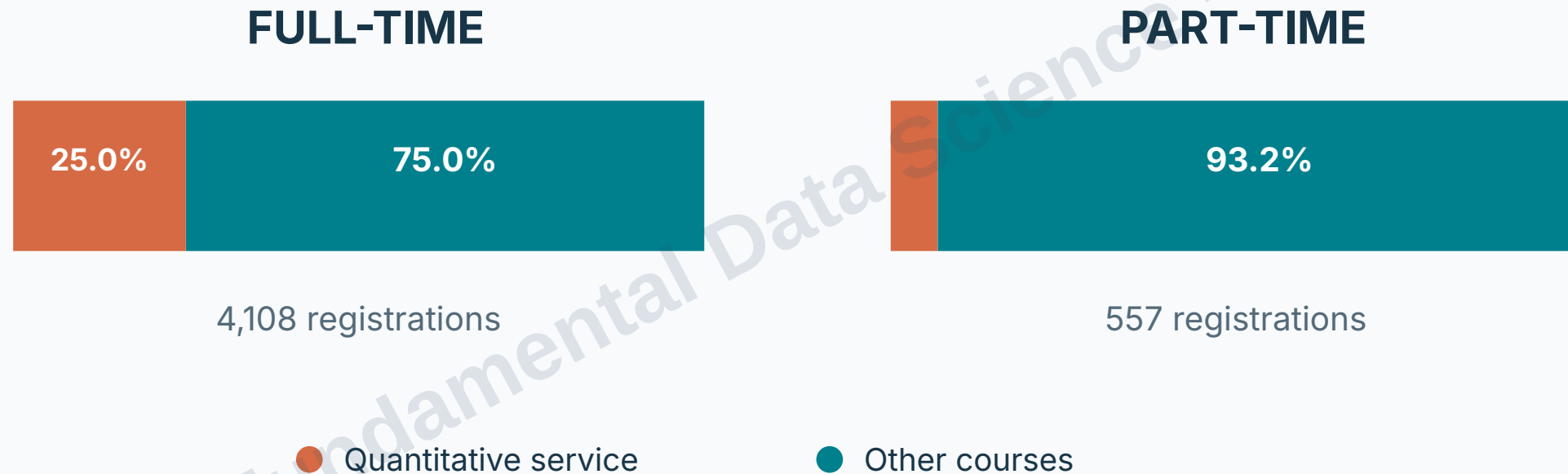
Every cell is passes / registrations in the synthetic course dataset.

Course group	Full-time	Part-time
Quantitative service	305 / 1,027 = 29.7%	7 / 38 = 18.4%
Other courses	2,754 / 3,081 = 89.4%	422 / 519 = 81.3%
ALL COURSES	3,059 / 4,108 = 74.5%	429 / 557 = 77.0%

Group-specific order: full-time > part-time. Overall order: part-time > full-time.

Part-time students register for a different mix

Course composition supplies unequal weights in the overall comparison.



The higher-pass-rate group of courses receives much more weight for part-time.

Reconstruct the course totals with the weights

The pooled percentages are weighted, not simple, averages.

FULL-TIME · 25% / 75%

$$0.25 \times 29.7\%$$

$$+ 0.75 \times 89.4\%$$

$$\approx 74.5\%$$

PART-TIME · 6.8% / 93.2%

$$0.0682 \times 18.4\%$$

$$+ 0.9318 \times 81.3\%$$

$$\approx 77.0\%$$

Calculated from unrounded fractions; intermediate displayed rates are rounded.

Broad course groups can still hide differences

A course with few registrations gives an unstable percentage.

ST110 • SMALL CELL

1 / 9

One extra pass would change
11.1% → 22.2%.

EC102 • EXCEPTION

30 / 34

Part-time 88.2% vs
full-time 86.3%.

Part-time has the lower rate in 11 of 12 individual courses, not all 12.

Question 3 · Did support cause the difference?

We observed higher rates for support within each readiness group.

Can we conclude support CAUSED the higher conditional rates?

- A** Yes, because the association survives grouping.
- B** Not yet; non-random selection may still matter.
- C** No, support must be harmful because the pooled rate is lower.

Answer 3 • Association is not causation

B is the defensible answer for this observational example.

WHAT WE OBSERVED

Higher pass rates among support participants within both readiness groups.

WHAT WE HAVE NOT ESTABLISHED

That support caused the difference: motivation, time, or preparation may still differ.

Causal questions require a suitable design and clearly stated assumptions.

The timing of a variable matters

Group by meaningful pre-existing factors; be cautious with variables affected later.

BEFORE SUPPORT

Baseline readiness

A potential common cause that can be examined.

AFTER SUPPORT

Attendance at tutoring

May itself be changed by support: do not automatically adjust.

A more detailed split is not always a more valid causal comparison.

Exploration and causal evaluation ask different things

Use EDA to generate checks and state what remains unknown.

EXPLORATORY QUESTION

Where is the pattern?

Does it reverse within groups?

Could a third variable matter?

CAUSAL QUESTION

Would changing the treatment alter outcomes?

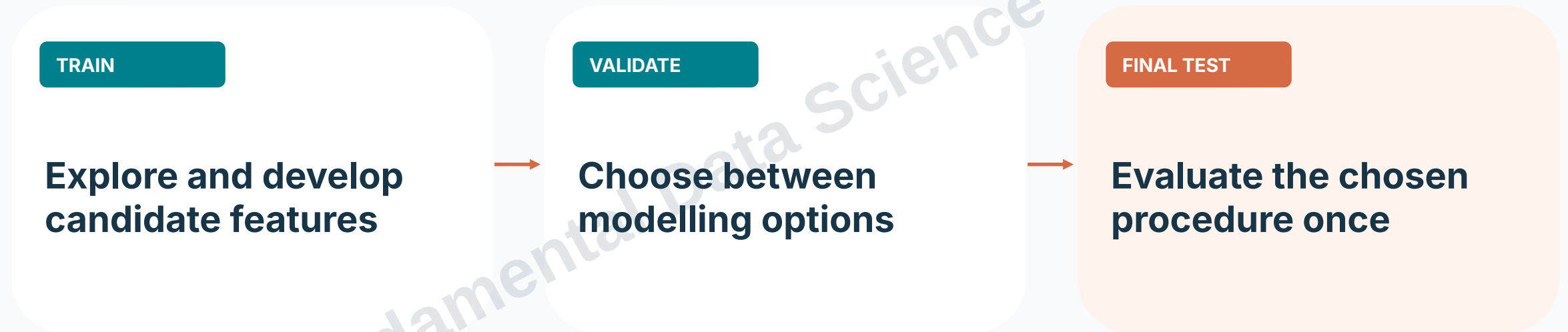
What design identifies the effect?

Which alternatives must be ruled out?

A rigorous EDA answer can finish with “we cannot conclude causation.”

Week 4 still matters: respect the test boundary

EDA for feature selection must not peek at the final test outcomes.



For new-student prediction, keep a student's registrations in the same split.

A defensible conclusion has three parts

Model wording using the actual synthetic course data.

OBSERVATION

Across 4,665 course registrations, pooled pass rates are 77.0% part-time and 74.5% full-time.

EXPLANATION OF THE REVERSAL Full-time has the higher observed rate inside both broad course groups, but the group mixes differ.

LIMITATION

This comparison does not identify a causal effect of study mode; course mix and small cells need care.

Lab 5 • Show the analysis, then defend it

Concept-focused implementation: explain each choice and its consequence.

- 1 QUESTION** State an explicit population and comparison.
- 2 DATA** Check join keys, registration rows and missing n.
- 3 EDA** Compare pooled and within-course pass rates.
- 4 INTERPRET** Write one observation, explanation and limitation.

Implementation may be short. The quality of reasoning is what you must demonstrate.

Five questions to ask before trusting a pattern

Use these prompts in labs, reports, and later model evaluation.

- 1 What does one row represent?
- 2 Which denominator defines the percentage?
- 3 What is the distribution and how was it measured?
- 4 Does the result change within meaningful groups?
- 5 **What can the data show — and what can they not prove?**

Reading · University teaching sources

Official open course notes and textbooks used to inform this lecture.

UC BERKELEY

Data 100: Data Cleaning and EDA; Visualization I/II

[Open source ↗](#)

HARVARD

Statistics 110: lectures 5–6 (Simpson's paradox)

[Open source ↗](#)

PENN STATE

STAT 100: Lessons 2, 5 and 6

[Open source ↗](#)

OPENINTRO

Introduction to Modern Statistics, 2e, chapters 4–6

[Open source ↗](#)

ISLP

James et al. (2023), applications in Python

[Open source ↗](#)

Appendix · Data provenance and calculations

Documentation to make updates and auditing easier.

Analysis unit	Course registration (not unique student)
Join	enrolments.csv → students.csv on student_id
Population	4,665 enrolments; 1,200 distinct students
Pass definition	Final grade ≥ 50 in the course dataset
Two broad groups	MA110, CH110, ST110 vs the other nine courses
Pooled rates	$3,059/4,108 = 74.46\%$; $429/557 = 77.02\%$