DATA1002 Informatics: Data and Computation

SCHOOL OF COMPUTER SCIENCE · THE UNIVERSITY OF SYDNEY · UNDERGRADUATE

Full guide + ask Sia about DATA1002

SIDE 1/2 EXAM MAP · METHODS · FORMULAS · DECISION RULES · COMPARISONS — **CLOSED-BOOK MEMORY SHEET** — S2 2026 · recall → transfer → check

0 · Exam map

UNIT OUTLINE

EXAM 50 | W 10 | 2 | R 20 | C 10 | P 8

Written Examination 50% · Weekly Coding Tasks 10% · Early-feedback MCQ 2% · Practice Coding Test 0% · Project Stage 1 Report 20% · Python Coding Test 10% · Project Stage 2 Poster 8%.

PERMITTED-MATERIALS TRUTH · DATA1002 SEMESTER 2 2026 UNIT OUTLINE, ASSESSMENT TABLE

"Written exam — Exam CLOSED book exam with MC and short essay questions"

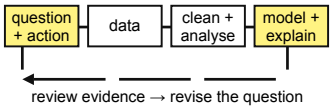
Priority: definitions → relationship → quantities → mechanism → boundary. The final is the 50% row: two-hour MCQ and short essay, with a 40% minimum.

1 · Frame the decision

C1

Data science lifecycle = iterative question framing, acquisition, cleaning, analysis, modelling, communication and review. It is not a linear software project: iteration, evidence and decision use come before modelling.

Decision question = bounded question naming actor · outcome · evidence · action. A *project failure mode* prevents valid, usable or maintainable results.



Source figure C1, redrawn: an answer loops back when evidence or use fails.

ASK FIRST	FAILURE SIGNAL
Who acts, on what outcome?	Technically correct output has no usable decision.
What evidence is needed?	Wrong provenance, stakeholder or implementation path.
What would change the claim?	No boundary or counter-case can be stated.

PREDICTION ERROR

$e_i = y_i - \hat{y}_i$ · observed minus predicted; sign shows over/under-prediction only after the target and scale are named.

EXAM CHECK → *A fluent model is still a failed project if the question, stakeholders, provenance or implementation path are wrong.*

1b · Reproducible synthesis

C12 · COURSE DATA

Reproducible analysis means data choices, code, outputs and interpretation can be rerun and inspected. A *poster narrative* orders question → data → method → result → evaluation → limitation.

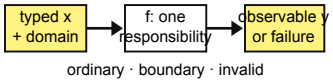
2 · Values, types, contracts

C2

Python value = a concrete object: number, string, Boolean or collection. Its *data type* fixes representation, valid operations and interpretation. Trace what reads, changes and depends on it.

FUNCTION CONTRACT

y = f(x) · specified typed input x → observable output y under documented type and boundary conditions.



Source figure C2, redrawn: contract first, then three independent tests.

TEST	WHAT IT EXPOSES
Ordinary	Expected transformation on a representative valid value.
Boundary	Off-by-one, empty input, limit and transition behaviour.
Invalid	Type/domain failure: signal or recover; never silently corrupt.

Automated test = reproducible comparison of behaviour/output with an expected result for specified input. State the expected result *before* execution. Change one condition at a time so a mismatch localises the defect.

2b · Trace, do not guess

C2 · C3

STEP	RECORD
Before	value, type, collection state, control condition
Operation	branch taken, iteration item, call arguments
After	new state, return/output, invariant still true?

Minimal example: expose one rule; explain why it works; name what breaks it; state failure behaviour. One success is not evidence for other types, missing values or boundaries.

WRITTEN-ANSWER STEM → *"For input x of type …, the condition … selects …; the observable result is …; the boundary case changes this because …."*

2c · Contract failure grid

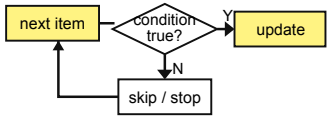
C2

FAILURE	OBSERVABLE RESPONSE
wrong type	reject or convert only under a documented rule
missing value	represent explicitly; do not let absence masquerade as data
empty collection	define return, identity value or deliberate failure
boundary value	test inclusion/exclusion at the exact transition
runtime error	handle without silently corrupting the result

3 · Branch, iterate, parse

C3

Conditional branch selects statements using a Boolean condition. *Iteration* repeats over a sequence or while a condition holds. *Text parsing* extracts structured information from character data using explicit rules.



Source figure C3, redrawn: every parser assumption becomes a branch or loop test.

LOOP ACCUMULATION

$s_k = s_{k-1} + x_k$ · define the initial value and invariant: after k items, what exactly does s_k mean?

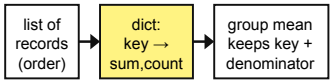
PARSER ASSUMPTION	BOUNDARY TEST
delimiter	missing, repeated or unexpected separator
case	upper/lower variants; normalise explicitly
encoding	character outside assumed representation
field presence	empty or missing field; do not shift columns silently

Invariant test: choose an initial value that makes the claim true at k=0; show one update preserves it. A plausible final total cannot repair a wrong starting state.

4 · Choose the structure

C4

List = ordered mutable collection accessed by position. *Dictionary* = mapping from unique keys to values for direct lookup/update. *Aggregation* = combine observations into grouped counts, totals, means or rates.



Source figure C4, redrawn: collection → keyed state → defensible summary.

GROUPED MEAN

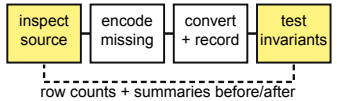
$\bar{x}_g = (1/n_g)\ \Sigma i \in g\ x_i$ · maintain both sum and count for each key; divide inside that group.

NEED	USE	TRAP
sequence/position	list	looking up a group by repeated scanning
keyed lookup/update	dictionary	losing the grouping key
fair comparison	rate/mean	ranking totals when size/exposure differs

5 · Cleaning as an audit

C5

Data provenance records where data came from and how they were collected, transformed and used. A *missing-value convention* distinguishes absent, unknown, inapplicable or withheld values. A *data-quality rule* tests validity, consistency, completeness, uniqueness or timeliness.



Source figure C5, redrawn: every cleaning decision leaves an inspectable trace.

MISSINGNESS RATE

rmiss = nmissing / nrows · numerator and denominator refer to the same in-scope field and records.

STAGE	REQUIRED EVIDENCE
Inspect	types, ranges, source, metadata, missing encodings
Transform	explicit rule; invalid conversions exposed
Account	row counts and summaries before/after each step
Validate	plausibility, type, uniqueness and completeness checks
Decide	exclude, impute or represent missingness; document why

Boundary: the percentage quantifies recorded absence, not its cause. Clean-looking data need not be representative, correctly labelled or fit for the decision.

5b · Quality decisions

C5

IF YOU SEE…	DO…
sentinel used as value	replace only when documentation establishes the convention
conversion failure	surface it; inspect affected records; never silently drop
duplicate key	test expected uniqueness before any join
missing target/feature	state the rule and downstream effect on the in-scope set

SHORT ESSAY → *name the assumption, its row-level effect, the check that detects failure, and how the decision would change.*

10 · Annotation before modelling

C10

Natural language processing represents, analyses or generates human-language data. *Data annotation* assigns labels or structured information. *Inter-annotator agreement* measures consistency among people using the same scheme.

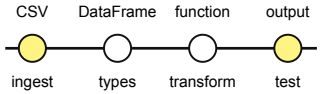
OBSERVED AGREEMENT

Po = nagreements / nitems

6 · Reusable data work

C6

Python function = named reusable block with inputs, operations and optional return. *Module* = reusable file/library namespace. *DataFrame* = labelled two-dimensional table whose columns may have different types.



Source figure C6, redrawn: explicit inputs and outputs make each step testable.

ROW-COUNT INVARIANT

nout = nin - nremoved + nadded

Record counts before/after every filter and join. A filter should lose only rows failing its predicate. A join may duplicate keys: unexplained change → inspect key uniqueness and join type before continuing.

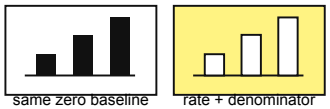
BOUNDARY	QUESTION
function	one responsibility? typed input? returned result?
module	does reuse reduce duplication without hiding state?
CSV	delimiter, header, encoding, missing convention?
DataFrame	labels, column types, key uniqueness, in-scope rows?

Limit: modularity improves testability but does not validate the semantics of a column, transformation or external source.

7 · Visual comparison

C7

Visual encoding maps data to position, length, colour, shape or another channel. A *comparison baseline* is the reference value/group. *Chart integrity* means scale, selection and decoration do not misrepresent data or uncertainty.



Source figure C7, redrawn: comparable scale and denominator precede visual polish.

NORMALISED RATE

rg = xg / ng · numerator and denominator must describe the same population and period.

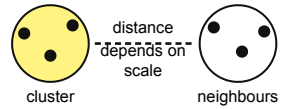
Choose the chart after naming the comparison. Show denominator/rate unit and evidence boundary. Salience can amplify a biased comparison; polish never repairs poor data.

COMPARISON NEED	INTEGRITY CHECK
group magnitude	same scale and in-scope population?

8 · Similarity → recommendation

C8

Clustering groups by defined similarity/distance. A *distance metric* quantifies feature-space dissimilarity. A *recommendation* ranks suggestions from user/item/similarity information and an objective.



Source figure C8, redrawn: representation and scale determine distance.

EUCLIDEAN DISTANCE

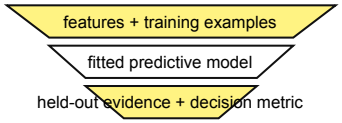
$d(x,z) = \sqrt{[\Sigma_{j=1}^{p}\ (x_j - z_j)^2]}$

Path: define p features → justify/standardise scales → difference, square, sum, square-root → rank neighbours → test usefulness for the decision. A cluster is not a natural label; it is an output of representation, scaling, metric and algorithm choices.

9 · Predict and evaluate

C9

Predictive model = fitted rule mapping features to an unknown outcome for new observations. *Evaluation metric* quantifies one aspect of performance. *Generalisation* is performance on relevant new data beyond training examples.



Source figure C9, redrawn: held-out evidence, not training fit, supports generalisation.

CLASSIFICATION ACCURACY

Accuracy = (TP + TN) / (TP + TN + FP + FN)

DECISION	REQUIRED CHECK
split + metric	fit on training examples; evaluate held-out evidence with a decision-aligned metric
threshold + claim	inspect FP/FN, class balance, subgroup error and relevance to new data

Training accuracy and one aggregate score can hide overfitting, subgroup error and an unusable threshold.

11 · AI reliability and fairness

C11

Generative model produces from learned patterns/context. *Algorithmic fairness* examines group errors, benefits and burdens. *Human oversight* checks evidence and consequences before action.

SELECTION-RATE GAP

$\Delta r = rA - rB$ · diagnostic signal, not a complete fairness judgement or remedy.

DATA1002 Informatics: Data and Computation

SCHOOL OF COMPUTER SCIENCE · THE UNIVERSITY OF SYDNEY · UNDERGRADUATE

Full guide + ask Sia about DATA1002

SIDE 2/2 WORKED PATHS · JUDGEMENT CARDS · CLASSIC TRAPS · EXAM-MORNING RETRIEVAL

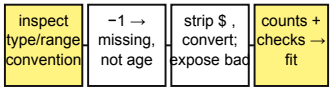
CLOSED-BOOK MEMORY SHEET

S2 2026 · recall → transfer → check

12 · Worked 1: cleaning bug

COURSE WORKED EXAMPLE · 6 MARKS

Given: a CSV stores missing ages as `-1` and prices as strings such as `'$1,240'`. Outline a testable pipeline before model fitting.



Worked path from the guide: known encodings become explicit missingness and numeric values.

1. **Inspect** types, ranges and missing conventions. [1]
2. Replace documented age sentinel **−1** with missing. [1]
3. Remove currency symbols/separators; convert price; expose invalid conversion. [1]
4. Record row counts and summaries before/after each transform. [1]
5. Test plausible ages and numeric nonmissing prices. [1]
6. Fit only after documenting how missing ages are excluded, imputed or represented. [1]

ANSWER TO THE END

```
-1 is no longer evidence of a young age; '$1,240' becomes numeric 1240 only after documented character removal; failures stay visible; model input and exclusions are auditable.
```

TRAP → *A model cannot distinguish a real value from a coding convention unless the pipeline makes the distinction.*

12b · Audit questions

C5 · C6

ASK	EVIDENCE IN ANSWER
What changed?	rule + affected rows + before/after count
What could fail?	invalid conversion, wrong sentinel, duplicate key
How detected?	range/type/uniqueness/completeness test
Why defensible?	provenance and documented downstream choice

Record the first transition where actual state diverges from the expected state.

12c · Formula sprint

C1–C12

RELATIONSHIP	MEANING / CHECK
$e_i = y_i - \hat{y}_i$	signed residual; target unit and scale
$y = f(x)$	typed input/output contract
$s_k = s_{k-1} + x_k$	initial value + loop invariant
$\bar{x}_g = (1/n_g)\Sigma x_i$	same-key sum and denominator
rmiss=nmissing/nrows	recorded absence, not cause

13 · Worked 2: accumulator

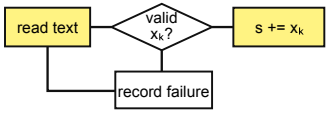
C3

Task: verify a text-processing loop that aggregates accepted values.

RELATIONSHIP

$s_k = s_{k-1} + x_k$

LINE OF WORKING	WHAT EARNS THE CHECK
Define	s_k = sum of accepted x values after k processed items.
Base	At k=0, s_0 must make the invariant true before any item.
Branch	Parser rule decides whether current text yields a valid x_k.
Update	Accepted: add x_k once. Rejected: preserve s and signal/record why.
Tests	ordinary, boundary and invalid strings; one condition changes at a time.



Source C3 drill: branch controls whether one update preserves the invariant.

Conclusion: correctness follows from a true base case plus an update that preserves the stated meaning of s_k; a plausible final total alone proves neither.

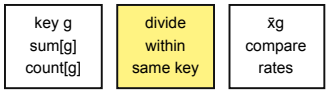
14 · Worked 3: grouped mean

C4

RELATIONSHIP

$\bar{x}_g = (1/n_g)\,\Sigma_{i \in g}\, x_i$

1. Choose a dictionary because the **group is a key**.
2. For every record in group g, update both `sum[g]` and `count[g]`.
3. Compute $\bar{x}_g$ only from the matching sum and n_g.
4. Check by reconstructing `sum[g] = x̄g × ng`; then interpret "per observation in group g".



Source figure C4, redrawn: never lose the key or denominator.

TRAP → *Ranking raw totals rewards group size; it answers a different question from a group mean or rate.*

14b · Row-count drill

C6

RELATIONSHIP

`nout = nin - nremoved + nadded`

OPERATION	EXPECTED / INSPECT
filter	only predicate failures removed; check missing/boundary
join	join type + key multiplicity; check duplicates/unmatched
conversion	count stable unless policy removes exposed failures

15 · Worked 4: distance

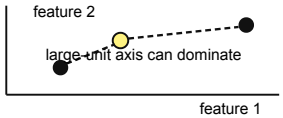
C8

Task: explain a cluster or recommendation from two represented observations.

RELATIONSHIP

$d(x,z) = \sqrt{[\Sigma_{j=1}^{p}\ (x_j - z_j)^2]}$

1. Name every feature j and its unit/meaning.
2. For each coordinate, compute $x_j - z_j$; square it so signs do not cancel.
3. Sum across the p features; take the square root.
4. Repeat only after scaling is justified; then rank neighbours or explain assignment.
5. Evaluate usefulness against the decision—not resemblance alone.



Source C8 drill: the nearest point can change when feature scales change.

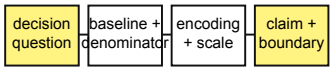
ANSWER BOUNDARY

```
"These are nearest under this representation, scaling and Euclidean metric; the result is not a discovered natural category."
```

16 · Visual judgement cards

C7

IF THE TASK ASKS…	DO
compare unequal group sizes	compute rg=xg/ng; show denominator or rate unit
show change against reference	name the baseline before choosing an encoding
judge trustworthiness	inspect scale, selection, uncertainty and decoration
communicate a conclusion	state what the data support and the evidence boundary



Source figure C7, redrawn: the chart is the last step, not the first.

TRAP → *A polished chart can make a weak or biased comparison more salient; it cannot make the comparison valid.*

16b · Rate self-test

C7

RELATIONSHIP

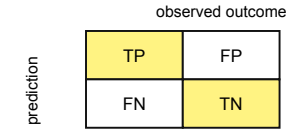
$r_g = x_g / n_g$

1. Define x_g as the counted event/quantity for group g.
2. Define n_g as the matching in-scope population or exposure for the *same period*.
3. Compute within each group before comparison.
4. Show the rate unit or denominator with the visual.

17 · Worked 5: model evaluation

C9

Task: evaluate a classifier without mistaking training fit for useful performance.



New drill figure: inspect all four cells before accepting the aggregate score.

RELATIONSHIP

`Accuracy = (TP + TN) / (TP + TN + FP + FN)`

1. Keep training examples for fitting; use held-out evidence for evaluation.
2. Populate TP, TN, FP and FN; denominator = all held-out cases.
3. Compute accuracy; compare with class balance and the decision baseline.
4. Judge unequal consequences of FP versus FN and inspect subgroup error.
5. Conclude only about relevant new data represented by the evaluation set.

ANSWER BOUNDARY

```
Accuracy counts correct predictions but can conceal class imbalance, unequal error costs, subgroup failure and an unusable threshold.
```

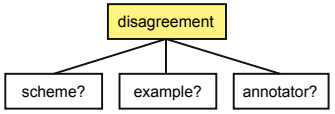
18 · Annotation diagnosis

C10

RELATIONSHIP

`Po = nagreements / nitems`

OBSERVED PATTERN	DIAGNOSIS TO TEST
low agreement	ambiguous text, unclear category boundaries, weak examples, or inconsistent annotators
high raw agreement	one label may dominate; show label-frequency context
model disagrees with label	separate model error from reference-label uncertainty
changed domain	annotation scheme and examples may not transfer

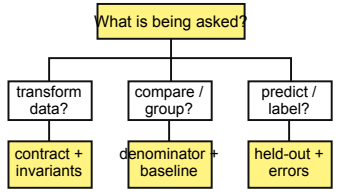


Source figure C10, redrawn: disagreement is evidence to diagnose, not merely blame.

TRAP → *Observed agreement is easy to compute but does not adjust for agreement expected from label frequencies.*

19 · Method chooser

C1–C11



New decision tree: identify the job, then select the check that can falsify it.

IF YOU SEE X	→ DO Y
broad request	actor/outcome/evidence/action → lifecycle
small computation	type + y=f(x) contract + three tests
repeated text	branch + loop invariant + parser boundaries
group records	dictionary key + sum/count + within-key mean
messy table	provenance → explicit rule → counts → invariant
unequal groups	normalised rate; same population and period
similarity	features + scale + metric → qualified neighbours
prediction	held-out set + metric + confusion cells + transfer
label disagreement	Po + label frequencies + scheme diagnosis
AI-assisted decision	verify source + uncertainty + subgroup effect + reviewer

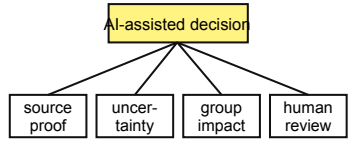
20 · Classic traps

ALL CHAPTERS

WRONG MOVE	REPAIR
model before decision question	state actor/outcome/evidence/action first
one successful input	ordinary + boundary + invalid
total as fair comparison	retain group count/exposure denominator
clean means representative	separate quality rules from sampling/labels/fitness
join count change ignored	inspect uniqueness and join type
chart chosen before comparison	baseline + denominator before encoding
cluster treated as truth	qualify feature, scaling, metric, algorithm
training score as generalisation	use relevant held-out evidence
fluent output as evidence	verify provenance; require accountable review

21 · AI decision audit

C11



Source figure C11, redrawn: four checks before action on generated output.

SELECTION-RATE DIAGNOSTIC

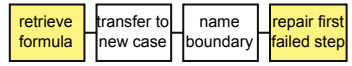
r_A = selectedA / eligibleA · r_B = selectedB / eligibleB · $\Delta r = r_A - r_B$

1. Give each numerator and eligible-population denominator.
2. Subtract rates using an explicit group order so the sign is interpretable.
3. Inspect uncertainty and whether comparison groups are relevant.
4. Examine error, benefit and burden distributions—not only aggregate performance.
5. Assign accountable review before action; document evidence and limitations.

Conclusion: the gap is a diagnostic signal. It cannot alone establish validity, explain mechanism or choose an ethical remedy. Fluent output is not verified evidence.

23 · Exam-morning plan

RETRIEVAL CONTROL



Source practice figure, redrawn: retrieval is finished only after transfer and repair.

1. **Blank page:** write the eleven relationships; define denominators.
2. **Say aloud:** lifecycle, provenance, contract, invariant, baseline, generalisation, oversight.
3. **Changed case:** run the chooser; show mechanism and boundary.
4. **Error log:** correct the first unsupported inference.
5. **Check:** precise definitions, signs/orders, units, denominators and evidence-bounded claim.

ANSWER SKELETON

```
Define → relationship → quantities → transformation → independent check → task interpretation → boundary.
```

LAST RECALL → *A strong DATA1002 answer makes the decision visible, the transformation testable, and the limitation operational.*