COMP20008 Elements of Data Processing

INFORMATION TECHNOLOGY · UNIVERSITY OF MELBOURNE · SECOND-YEAR UNDERGRADUATE

EXAM REVISION SHEET
S2 2026 · **SIDE 1 OF 2**
Definitions · rules · formulas

Full guide + ask Sia about COMP20008

SIDE 1/2 EXAM MAP → method rules → formulas → comparisons → standard critiques | **S2 2026 · 50% FINAL EXAMINATION** | Revision priority: choose → calculate → interpret → critique

01 · Exam map COURSE.JSON · BRIEF

Continuous / project 50% | Examination 50%

```
Continuous / project assessment — 50%
· End-of-semester examination — 50%.
```

Permitted materials: not confirmed for this offering in the course materials we hold — check the exam instructions on the LMS.

REWARDED MOVE	WRITE IN THIS ORDER
Calculate	formula → substitution → answer → meaning
Compare	axis → A → B → consequence
Critique	failure cause → symptom → mitigation

02 · Pipeline + types C01

INPUT types → CLEAN scale → MODEL learn → OUT show → USE ethics

Source-derived pipeline: locate the stem first; the legal method belongs to that stage.

TEST	CLASS	EXAMPLES
Arithmetic meaningful?	Numerical	temperature, income; discrete counts / continuous range
Label or identifier?	Non-numerical	postcode, phone number
Ordered category?	Could-be-either	S/M/L; code 1/2/3 if justified

MEASUREMENT LADDER

```
nominal: labels < ordinal: order <
interval: equal gaps, no true 0 <
ratio: equal gaps + true 0
```

Do not classify by appearance. "3010" is non-numerical when it is a postcode. Separate this from dataset structure: table = structured; keys/tags = semi-structured (JSON/XML/HTML); free text/images = unstructured.

PRIVACY DECISION RULE

SELF person acts | LOCAL alter record | GLOBAL alter answer

Source-derived privacy levels: protection may act at the person, record or query-answer level.

RELEASE RISK	RULE
re-identification by quasi-identifiers	k-anonymity: every pattern matches ≥k records
one sensitive value throughout a k-group	homogeneity attack → require ≥l distinct values
query answer reveals one person	global DP: add noise scaled to sensitivity

03 · Clean before model C02

AUDIT: accurate · complete · consistent

timely · believable · interpretable

↓

DROP? | IMPUTE? | SCALE?

Source-derived quality gate: diagnose the dimension, then choose a repair and name its cost.

IF YOU SEE…	DO / SAY
rare, apparently random missingness	drop a few rows; loss is small
many/predictable missing values	impute: mean/median/mode, fill, regression or k-NN; may bias and dampen variance
mixed dates/names	standardise representation; de-duplicate
extreme value	verify; cap/remove or use robust method; never delete silently

SCALING — BITES BEFORE K-NN, K-MEANS, PCA

```
min-max: x' = (x - min)/(max - min) ∈
[0,1]
z-score: z = (x - x̄)/s → mean 0, SD 1
```

Min-max is bounded but one extreme stretches the range and compresses the rest. **z-score** uses standard-deviation units and is less squashed by an extreme. Scaling prevents units with a large numeric range from dominating distance/variance.

SAMPLING

Random: every record equally likely. **Stratified**: sample inside groups to preserve proportions. **Systematic**: every kth record. Always ask whether the scheme is representative; saving cost is not worth selection bias.

GROUND AN LLM ANSWER

query → retrieve chunks → augment → answer

permission gate surrounds retrieval

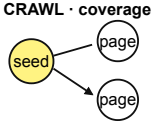
Source-derived RAG: retrieve permitted evidence, add it to the prompt, then generate.

Why: an LLM predicts plausible tokens rather than checking facts; retrieval grounds output and reduces hallucination. **Content partitioning:** chunk for relevance and enforce permissions so one user's data cannot enter another answer. **AUTOMAT:** Act-as, User, Targeted-action, Output, Mode, Atypical cases, Topic.

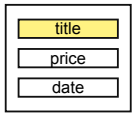
TRAP *"Impute" is incomplete: name the method and the bias ↔ information-loss trade-off.*

04 · Capture → clean text C03–04

CRAWL · coverage



SCRAPE · fields



Source-derived contrast: crawling follows links; scraping extracts named DOM fields.

AXIS	CRAWLING	SCRAPING
Goal	coverage/indexing	structured data
Move	frontier; breadth-first links	targeted DOM parsing
Output	URLs/raw pages	field table
Control	robots.txt + visited set	find_all() + get_text()

SCRAPING ASSEMBLY LINE

```
seed URL → fetch HTML → parse DOM →
find_all(target nodes) → get_text(strip
tags) → standardise → clean matrix
```

Flawed crawler: no visited set means duplicates re-enter the frontier, cycles repeat, unique coverage stalls and termination fails. Deduplicate before enqueue/fetch; stop at the target count of unique URLs.

sentence split → tokenise normalise → lemma → filter / n-grams

Source-derived text pipeline: order matters; clean tokens feed BoW/TF-IDF.

LEMMATISE	STEM
dictionary + grammar; valid base: better→good	suffix rule; may be non-word: studies→studi

Splitting needs exceptions for abbreviations and decimals. Lowercase/strip punctuation to merge forms. Lemmatisation shrinks sparsity; stop-word removal drops ubiquitous noise; n-grams restore short local order but enlarge vocabulary.

FAILURE SYMPTOM	DIAGNOSIS → REPAIR
"Dr. Lee" becomes two sentences	naive full-stop split → add abbreviation/number exceptions
"The", "the", "the," become three features	missing normalisation → lowercase + strip punctuation
chase/chased/chasing stay separate	surface forms → lemmatise to one dictionary form
crawler revisits navigation forever	no visited set → check before enqueue and fetch

ANSWER SHAPE *Name the stage, show its input→output, then state the downstream benefit.*

05 · Vectorise + visualise C05

fixed shared vocabulary

	the	fox	hen	chase
D1	2	1	0	1
D2	1	0	2	2

Source-derived BoW: rows are comparable only after vocabulary alignment and zero padding.

BoW counts terms over a fixed vocabulary. It is simple, interpretable, fixed-length and language-agnostic; it also discards order/semantics, is sparse/high-dimensional and lets common terms dominate.

DISTANCE + WEIGHTING

```
d(x,y) = √(Σᵢ(xᵢ-yᵢ)²)
TF-IDF(t,d) = TF(t,d) × IDF(t)
IDF(t) = log(N/df(t))
```

USE	RULE	FAILURE
raw TF	frequency within document	ubiquitous words swamp topics
TF-IDF	down-weight high df; up-weight distinctive terms	still sparse; still loses order
n-grams	retain short phrases	larger vocabulary

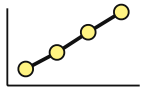
CHART DECISION

categories	bar	time/order	line
2 numeric	scatter	distribution/groups	boxplot
matrix	heatmap	—	—

raw: spread visible



bins: tidy trend



Source-derived binning trap: averages smooth noise, hide variance and can exaggerate a trend.

REPRESENTATION DIAGNOSIS

NEED	REPRESENTATION	REASON
interpretable counts	BoW	fixed coordinates; easy distance
topic distinction	TF-IDF	distinctive terms outweigh ubiquitous ones
short phrase/order	n-grams	retain local sequence, but expand vocabulary
word context	windowed co-occurrence	count neighbours around each target occurrence

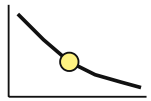
For a critique, separate **representation loss** (order/semantics discarded) from **weighting failure** (common terms dominate) and match the fix to the failure.

06 · Cluster + reduce C06–08

k-means loop



SSE elbow



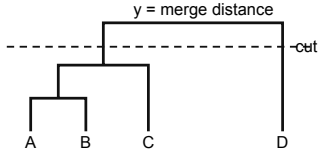
Source-derived k-means and elbow: fit improves with k, so do not minimise SSE blindly.

K-MEANS — WHERE IT BITES

```
repeat: nearest-centroid assignment →
centroid = mean of members →
convergence
SSE = Σclusters Σpoints ‖x - μcluster‖²
```

Elbow: plot SSE against k; pick the bend where marginal reduction collapses. Advantage: simple/visual. Limitation: bend may be ambiguous/subjective; silhouette or Calinski–Harabasz are alternatives. **Outlier:** squared distance inflates SSE and drags a centroid or consumes a cluster. Remove/cap after verification, scale, or use k-medoids/density methods.

METHOD	CLUSTER DISTANCE / CHOICE	SIGNATURE
k-means	nearest mean; set k first	flat; outlier-sensitive
single link	minimum pair distance	chaining
complete link	maximum pair distance	compact
average link	mean pair distance	compromise
VAT	reordered distance matrix	dark diagonal blocks suggest k



Source-derived dendrogram: cut horizontally; count crossed branches for the cluster count.

PCA

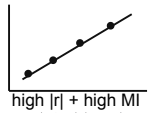
Project onto orthogonal directions of maximum variance: PC1 most, PC2 next. Variance ≠ target relevance. A low-variance direction may separate classes; scale first because PCA is variance-sensitive.

COMPARE CLUSTERINGS

Treat each clustering as a categorical labelling and compute **NMI**: near 1 means the labellings agree; near 0 means they disagree. It compares partitions even when cluster names differ. Do not confuse this use with VAT, which diagnoses cluster tendency before fitting.

07 · Dependence → prediction C08–12

linear



high |r| + high MI

y = x²



r≈0; MI high

Source-derived dependency test: Pearson sees a line; MI sees any dependence.

UNCERTAINTY FORMULAS

```
H(X) = -Σᵢ pᵢ log₂pᵢ
MI(X,Y) = H(Y) - H(Y|X)
NMI = MI/min(H(X),H(Y))
```

Pearson r ∈ [−1,1]: signed linear association; misses curved/non-monotone structure. **MI**: 0 iff independent; detects non-linear/categorical dependence, but has no sign and continuous data needs discretisation. Entropy is 0 when deterministic; binary entropy is 1 bit at 50/50.

STEM	CHOOSE / WARNING
predict class from labelled rows	classification: k-NN/tree
predict continuous value	regression: $\hat{y}=b_0+b_1x$; least squares
small local neighbourhood	k-NN; scale; k small noisy, k large smooth
interpretable recursive rules	tree; depth controls overfit
near-unique category	information gain biased to high cardinality → pure leaves, memorisation

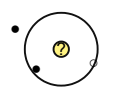
EVALUATION — DERIVE FROM TP/FP/FN/TN

```
Accuracy=(TP+TN)/N
Precision=TP/(TP+FP)
Recall=TP/(TP+FN)
F1=2PR/(P+R)
```

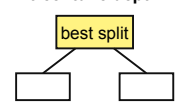
Metric by error cost: false negative costly (fraud/cancer) → recall; false positive costly → precision. Accuracy can hide failure on imbalanced data. k-fold CV rotates the held-out fold and averages scores; every row is tested once.

Leakage: target or future-only data appears among features; "perfect" training MSE can be invalid because the value is unavailable at prediction. Audit availability before fitting.

k-NN: scale, tune k



tree: tune depth



Source-derived supervised choices: neighbourhood size controls locality; tree depth controls complexity.

Feature selection: filter methods rank features (correlation/MI), wrapper methods search subsets by model performance, and embedded methods select during fitting. **Verification:** public key → recompute statement hash → compare with signed hash; a blockchain edit changes a hash and breaks later links.

COMP20008 Elements of Data Processing

INFORMATION TECHNOLOGY · UNIVERSITY OF MELBOURNE · SECOND-YEAR UNDERGRADUATE

EXAM REVISION SHEET
S2 2026 · **SIDE 2 OF 2**
Self-test · answers · traps

Full guide + ask Sia about COMP20008

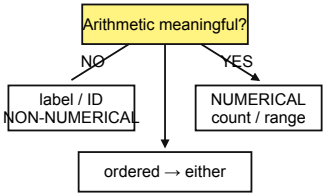
SIDE 2/2 WORKED MINI-EXAMPLES → judgement cards → traps → exam-morning recall

SELF-TEST: COVER ANSWERS FIRST

Show the path; finish every number with meaning

08 · Drill: classify + clean

WE C01–02



Source-derived Q1 decision tree: answer from meaning, not typography.

VALUE	ANSWER PATH
postcode "3010"	arithmetic meaningless → non-numerical identifier
temperature °C	numerical, continuous, interval: equal gaps; no true zero
T-shirt S/M/L	ordered category → could-be-either; ordinal 1/2/3 if justified
phone number	identifier; leading zeros matter → non-numerical

Answer: non-numerical; numerical; could-be-either; non-numerical. Each label needs its reason.

QUALITY MINI-ANSWER

Inconsistent representation: "13/02/19", "13th Feb 2019", "2019-02-13" in one date column → standardise before parse/sort. **Missing:** blank DOB or "20 years ago" breaks age calculation → drop if rare/random; otherwise justify imputation.

IF YOU SEE X	→ DO Y
distance model + mixed units	scale first
bounded range, no extremes	min-max
extremes / SD units desired	z-score
few random missing rows	drop, state information loss
many model-able gaps	impute, state bias/variance cost

EXAM-MORNING PLAN

1. Confirm the LMS instructions and timing; recite the formulas once, then stop cramming.
2. Recall six anchors: type test; scaling; vector alignment; linkage rule; metric denominators; privacy distinction.
3. For every stem: identify stage/output, choose method, show working, state meaning, add one limitation.
4. If stuck, write the decision rule and a concrete consequence; a method name alone is not a critique.

SAY ALOUD *Accuracy, completeness, consistency, timeliness, believability, interpretability.*

09 · Drill: BoW distance

WE C04–05

Text: "The fox saw the hen and chased it. The hen turned and chased the fox." Window = two words either side; exclude target.

	the	saw	chased	it	turned	and
fox	2	1	1	0	0	0
hen	1	0	2	1	1	1

six component differences, each ±1 → √6

Source-derived worked vector: align the union vocabulary before subtracting.

1. Contexts: fox → the=2, saw=1, chased=1. Hen → chased=2, it=1, the=1, turned=1, and=1.
2. Union vocabulary: {the, saw, chased, it, turned, and}. Pad absent terms with 0.
3. fox=(2,1,1,0,0,0); hen=(1,0,2,1,1,1).

SUBSTITUTE ALL SIX DIMENSIONS

$d = \sqrt{(2-1)^2+(1-0)^2+(1-2)^2+(0-1)^2+(0-1)^2+(0-1)^2}$
$= \sqrt{1+1+1+1+1+1} = \sqrt{6}$

After lemmatisation: chased→chase, turned→turn, saw→see. Hen context becomes chase=2, it=1, the=1, turn=1, and=1. Counts collapse onto valid base forms, shrinking and de-sparsifying the vocabulary.

TF-IDF JUDGEMENT

Cooking posts cluster poorly under raw TF because "the", "recipe", "cup", "minutes" dominate every vector. IDF down-weights terms appearing in many documents and up-weights "sourdough", "tempering", "marinade": distinctive words distinguish topics.

TF count × IDF log(N/df) = topic weight

Source-derived weighting: a term matters when frequent here but uncommon across documents.

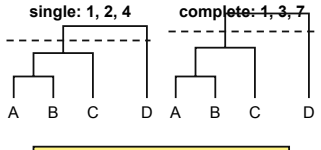
CLASSIC TRAPS — ANTIDOTES

- **BoW:** "dog bites man" = "man bites dog"; add n-grams if local order matters.
- **Elbow:** SSE falls with every larger k; choose the bend, never the raw minimum.
- **Single link:** one near bridge can chain remote groups.
- **PCA:** high variance may be irrelevant to the target.
- **Tree:** high-cardinality purity may be memorisation.
- **RAG:** grounding reduces hallucination, but chunking and permissions still govern quality and leakage.

10 · Drill: linkage + clusters

WE C06–07

Points: A=1, B=3, C=4, D=8. Pair distances: AB=2, AC=3, AD=7, BC=1, BD=5, CD=4.



cut below final merge → {A,B,C} | {D}

Source-derived worked dendrogram: same two-cluster cut, different merge heights.

1. Smallest distance BC=1 → merge {B,C} for both.
2. **Single:** {BC}–A=min(2,3)=2; then {ABC}–D=min(7,5,4)=4. Heights **1,2,4**.
3. **Complete:** {BC}–A=max(2,3)=3; {BC}–D=max(5,4)=5, so merge A; then {ABC}–D=max(7,5,4)=7. Heights **1,3,7**.
4. Cut just below the last merge → {A,B,C} and {D}.

IF YOU SEE X	→ DO Y
need k before fitting?	k-means: use VAT/elbow; elbow may be subjective
nested relationships / choose k later	hierarchical + dendrogram cut
straggly chaining	single linkage caused it
compact-diameter preference	complete linkage
far outlier + squared distance	centroid dragged; robustify or change method

PCA / DEPENDENCE CARD

If PC1 has 95% variance, it is useful for compression/visualisation, *not proof of relevance*. If $x=-3,\ldots,3$ and $y=x^2$, symmetry cancels the linear trend: $r\approx 0$. Yet y is determined by x, so MI/NMI is high. Pearson limitation = linear only; MI limitation = no sign and continuous data needs binning.

MINI SELF-TEST	ANSWER PATH
SSE curve has no clear bend	elbow is ambiguous; report subjectivity and consider silhouette/Calinski–Harabasz
VAT has three dark diagonal blocks	three groups of mutually close points; try k=3
two partitions use different label names	compare assignments with NMI; names themselves do not matter
PC1 explains most variance	compression cue only; test relevance to the target before discarding PCs

Full-credit clustering answer: define the distance rule, carry each merge height, state the cut, then name the geometry it favours.

11 · Drill: metrics + leakage

WE C09–10

Fraud classifier, N=1000: TP=30, FP=20, FN=10, TN=940.

PREDICTED \ ACTUAL	positive	negative
positive	TP 30	FP 20
negative	FN 10	TN 940

precision row · recall column

Source-derived matrix: orient the cells before touching a formula.

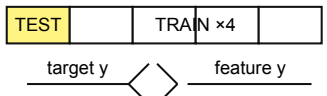
CARRY THE NUMBERS TO THE END

```
Accuracy=(30+940)/1000=0.97
Precision=30/(30+20)=0.60
Recall=30/(30+10)=0.75
F1=2(0.60·0.75)/(0.60+0.75)=0.90/1.35≈0.667
```

Judgement: prioritise recall because a missed fraud is the costly false negative. Accuracy 0.97 is misleading: legitimate transactions dominate, so a majority-only model can look strong while catching little fraud.

IF COSTLIEST ERROR IS…	OPTIMISE	WHY
false negative	recall	catch actual positives
false positive	precision	trust positive predictions
both	F1	harmonic balance

5-fold CV: rotate held-out fold



target y — feature y

copy target → MSE 0 = leakage

Source-derived evaluation: honest held-out scoring; a target-feature shortcut is invalid.

Tree trap: ~80 nearly unique colour values create pure leaves because information gain favours high cardinality; the tree memorises. Coding colours 1–80 fabricates order and thresholds. **k-NN trap:** new users have tiny Jaccard overlap $|A\cap B|/|A\cup B|$; use content/hybrid/popularity priors.

REGRESSION + SELECTION CHECK

PROMPT	FULL-CREDIT MOVE
interpret b_1	$\Delta\hat{y}$ per +1 unit x; least squares minimises squared residuals
report R^2	share of y-variation explained; high R^2 is not proof of causality
choose features	filter ranks correlation/MI; wrapper searches subsets; embedded selects during fitting
evaluate honestly	hold out unseen rows or rotate k folds; keep test sealed

12 · Drill: privacy + trust

WE C11–12

Scenario: table is 2-anonymous on {age-band, postcode}; a two-record group {50–59, 30**} has Diabetes for both rows.

k=2: WHO hidden	l=2: WHAT hidden
50–59 · 30** Diabetes	50–59 · 30** Diabetes
50–59 · 30** Diabetes	50–59 · 30** other condition
same sensitive value	≥2 sensitive values
HOMOGENEITY	**less utility**

Source-derived privacy fix: k-anonymity hides identity; l-diversity protects the sensitive value.

1. 2-anonymity holds: every QI pattern matches at least one other row.
2. Yet locating the target in the group reveals Diabetes with certainty: **homogeneity attack**.
3. Enforce l-diversity: each QI group has ≥l distinct sensitive values, via coarser grouping/merging/suppression.
4. Trade-off: more generalisation improves privacy but reduces utility.

query truth **R=12** → mean-0 noise → release 14 or 11

more noise → privacy ↑ · accuracy ↓

Source-derived global differential privacy: query, perturb, release; sensitivity sets noise scale.

TERM	FAST DISTINCTION
copyright	expression; automatic
patent	novel invention/process; registered, time-limited
trademark	brand identifier
self / local / global	individual protects / alter record / alter query answer

Digitally signed credential: obtain issuer public key → recompute exact statement hash → verify signature resolves to the same hash. Match proves issuer authenticity + content integrity. **Chain:** each header stores the previous-header hash; changing earlier data changes its hash and breaks later links.

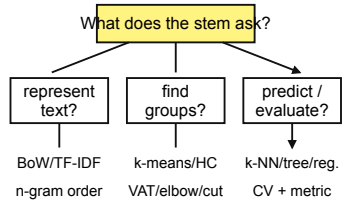
TRAP *Mean-0 noise does not mean zero noise; one release varies while repeated expectation returns the truth.*

BIAS DIAGNOSIS

SIGNAL	DIAGNOSIS
under-represented group	representation bias → worse performance
flawed feature/label proxy	measurement bias → misses target
sensitivity	one-record maximum change → noise scale

13 · Method chooser

ALL CHAPTERS



New decision figure: method selection starts from the question's output, not a favourite algorithm.

IF YOU SEE X	→ DO Y / SAY WHY
digits used as an identifier	non-numerical; arithmetic meaningless
web links/frontier/coverage	crawler + robots.txt + visited set
named fields in one DOM	scrape: find_all → get_text
common words blur topics	TF-IDF
distance/variance method	scale first
unknown k / tendency	VAT then elbow; state ambiguity
curved dependence	MI/NMI, not Pearson alone
imbalanced positives	precision/recall/F1, not accuracy alone
too-perfect training score	audit leakage; evaluate held out
shared sensitive value inside k-group	homogeneity → l-diversity
model answer needs sources	RAG retrieve → augment → generate; enforce permission gate

CLASS → FORMULA → MEANING → CRITIQUE

name the limitation + consequence + fix

New exam-morning loop: the interpretation and limitation are part of the answer.

FINAL CHECK *Units and denominators; train/validation/test roles; target unavailable at prediction; no unsupported causal language.*

EVALUATE CUES

SIGNAL	CHECK / FIX
binned trend looks strong	inspect raw scatter: binning exaggerates a trend and hides spread
fluent but false	hallucination predicts likely tokens without grounding; use permitted RAG chunks
credential trust	public key → recompute hash → verify signature; match = authenticity + integrity
near-zero Pearson	inspect scatter or use MI before declaring independence