IMEN891K / IDSC501 / AXGS-501 · DATA ENGINEERING · FALL 2026

Midterm Review

Sangwoo Mo · POSTECH
Industrial & Management Engineering

Outline: Midterm Review

01 **Proposal Review**

02 **Midterm Review**

03 **By Course Code**

01 · Proposal Review

01 **Proposal Review**

02 Midterm Review

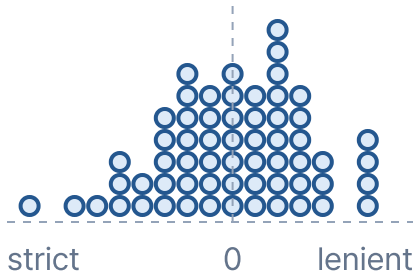
03 By Course Code

How Scores Were Processed

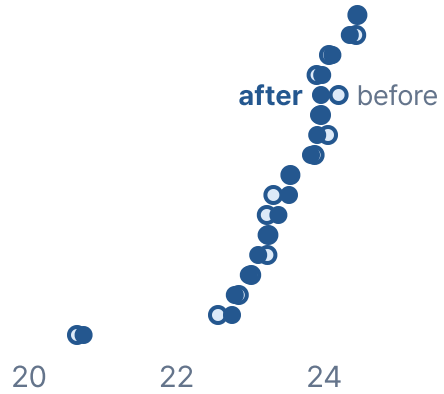
1. Adjust for each rater

per criterion: score = team + rater + error

Rater strictness

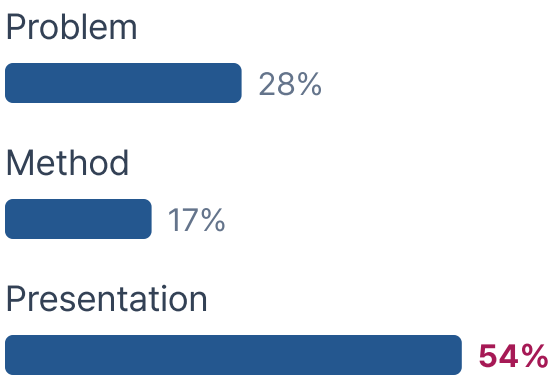


Subtract the effect



61 raters, points on the 30-point total

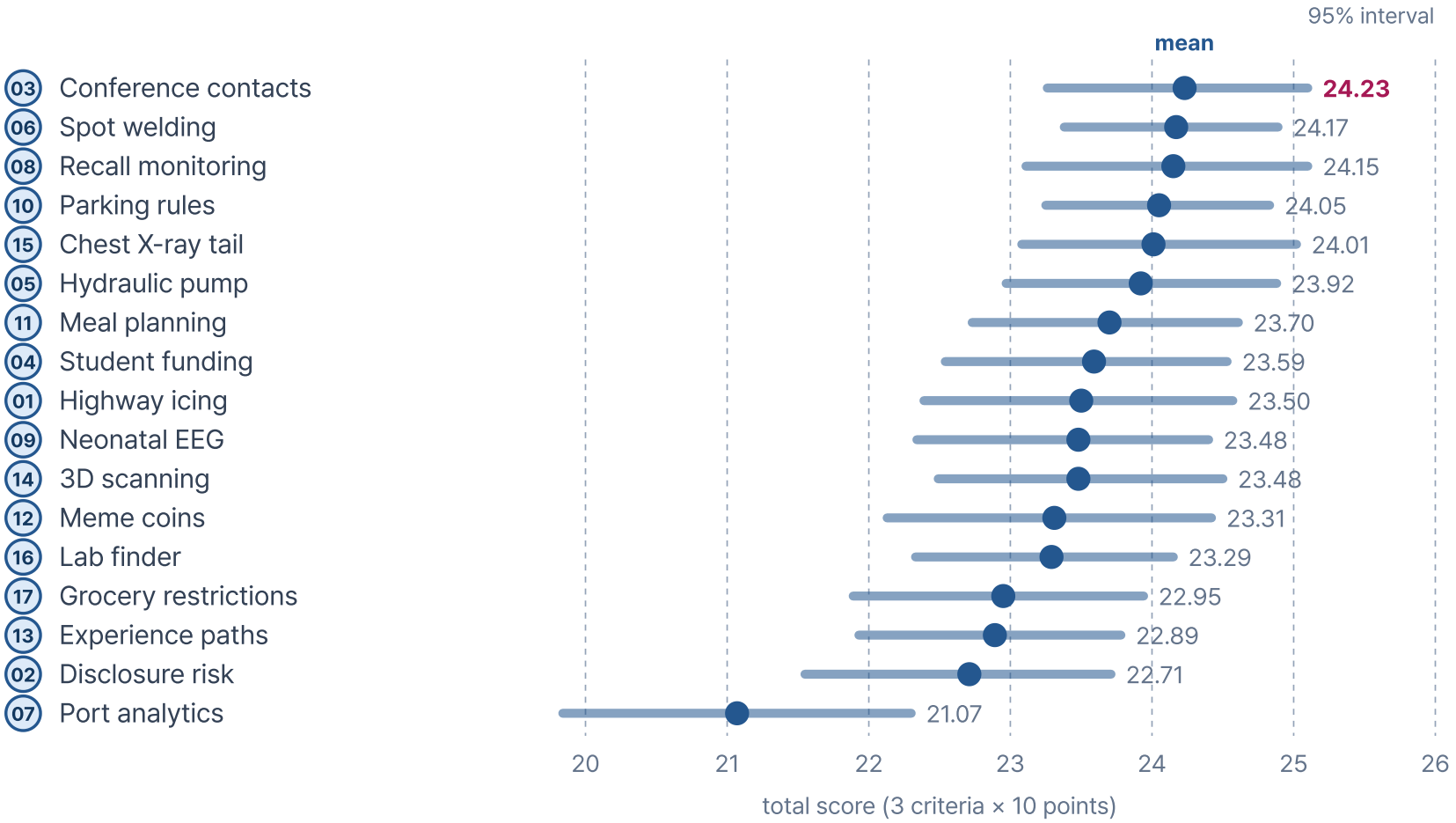
2. Balance the criteria



before: share of total variance
same mean and spread, then sum

Total / 30 = Problem / 10 + Method / 10 + Presentation / 10

Peer Scores

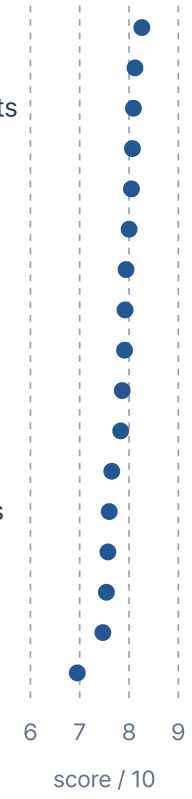

95% interval
mean
03 Conference contacts
06 Spot welding
08 Recall monitoring
10 Parking rules
15 Chest X-ray tail
05 Hydraulic pump
11 Meal planning
04 Student funding
01 Highway icing
09 Neonatal EEG
14 3D scanning
12 Meme coins
16 Lab finder
17 Grocery restrictions
13 Experience paths
02 Disclosure risk
07 Port analytics
24.23
24.17
24.15
24.05
24.01
23.92
23.70
23.59
23.50
23.48
23.48
23.31
23.29
22.95
22.89
22.71
21.07
20
21
22
23
24
25
26
total score (3 criteria × 10 points)

Peer Scores by Criterion

Final criterion scores add up to the total

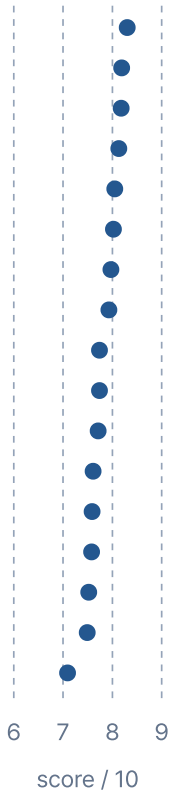
Problem

1	08	Recall monitoring	8.26
2	11	Meal planning	8.12
3	03	Conference contacts	8.09
4	10	Parking rules	8.07
5	01	Highway icing	8.05
6	15	Chest X-ray tail	8.00
7	06	Spot welding	7.94
8	05	Hydraulic pump	7.92
9	09	Neonatal EEG	7.91
10	04	Student funding	7.86
11	14	3D scanning	7.83
12	16	Lab finder	7.65
13	17	Grocery restrictions	7.60
14	13	Experience paths	7.57
15	12	Meme coins	7.54
16	02	Disclosure risk	7.47
17	07	Port analytics	6.95



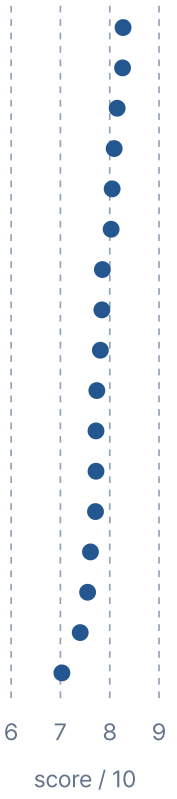
Method

1	03	Conference contacts	8.30
2	05	Hydraulic pump	8.19
3	08	Recall monitoring	8.18
4	10	Parking rules	8.13
5	01	Highway icing	8.05
6	09	Neonatal EEG	8.02
7	06	Spot welding	7.97
8	14	3D scanning	7.93
9	12	Meme coins	7.74
9	15	Chest X-ray tail	7.74
11	13	Experience paths	7.71
12	17	Grocery restrictions	7.61
13	16	Lab finder	7.59
14	04	Student funding	7.58
15	02	Disclosure risk	7.52
16	11	Meal planning	7.49
17	07	Port analytics	7.09



Presentation

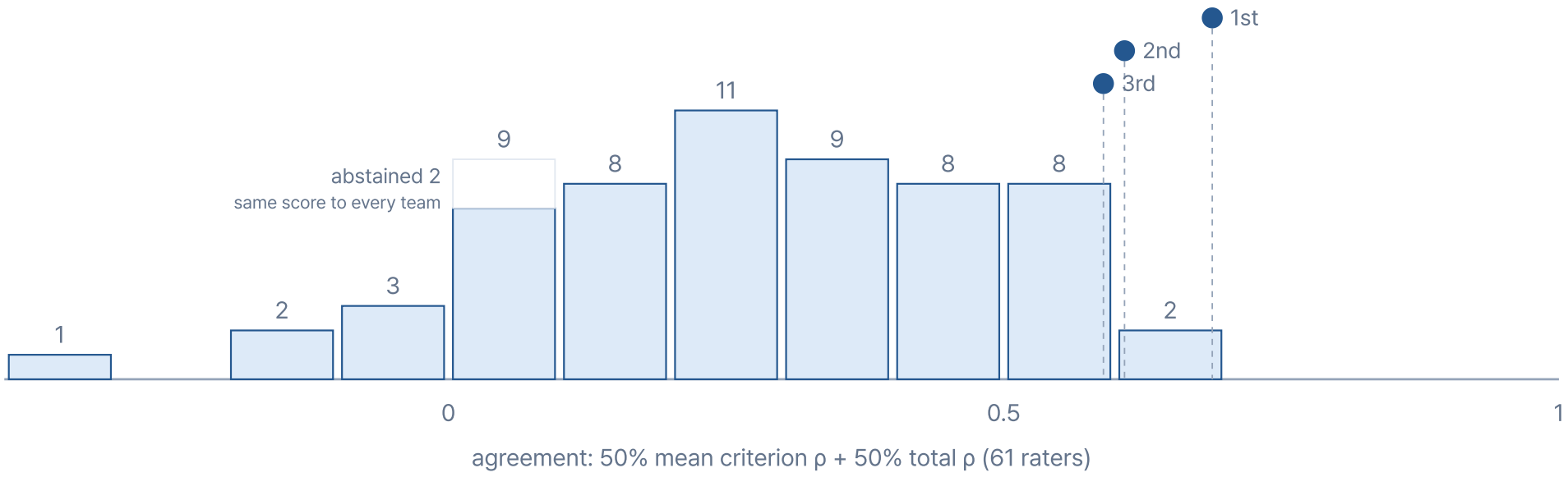
1	15	Chest X-ray tail	8.27
2	06	Spot welding	8.26
3	04	Student funding	8.15
4	11	Meal planning	8.09
5	16	Lab finder	8.05
6	12	Meme coins	8.03
7	10	Parking rules	7.85
8	03	Conference contacts	7.84
9	05	Hydraulic pump	7.81
10	17	Grocery restrictions	7.74
11	02	Disclosure risk	7.72
11	14	3D scanning	7.72
13	08	Recall monitoring	7.71
14	13	Experience paths	7.61
15	09	Neonatal EEG	7.55
16	01	Highway icing	7.40
17	07	Port analytics	7.03



Who Rated Well

The top three raters get bonus points

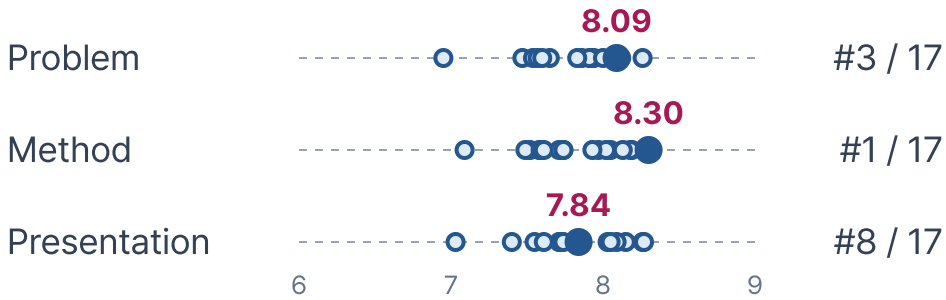
	Problem	Method	Presentation	Total	Combined
1st 이신영	0.574	0.499	0.578	0.825	**0.688**
2nd 박희령	0.747	0.361	0.537	0.669	**0.609**
3rd 박수영	0.440	0.363	0.675	0.687	**0.590**



agreement: 50% mean criterion ρ + 50% total ρ (61 raters)

Team 03: Conference contacts





Total: 8.09 + 8.30 + 7.84 = 24.23

AI feedback (slides only)

WHAT WORKED

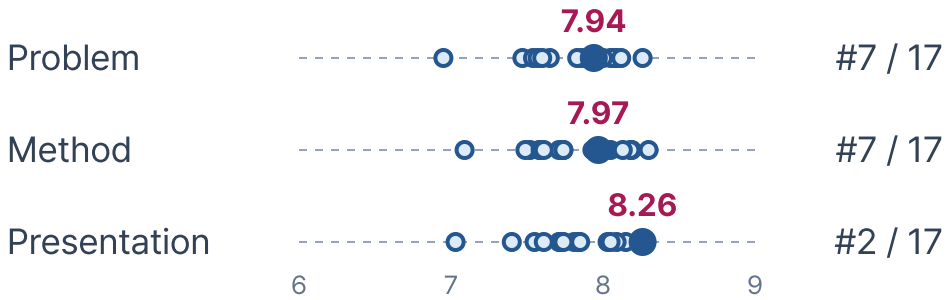
· PostgreSQL design: 16 entities with keys, provenance and sync metadata

WHAT WAS MISSING

· Ingestion examples only; no loaded counts or executed query results

Team 06: Spot welding





Total: 7.94 + 7.97 + 8.26 = 24.17

AI feedback (slides only)

WHAT WORKED

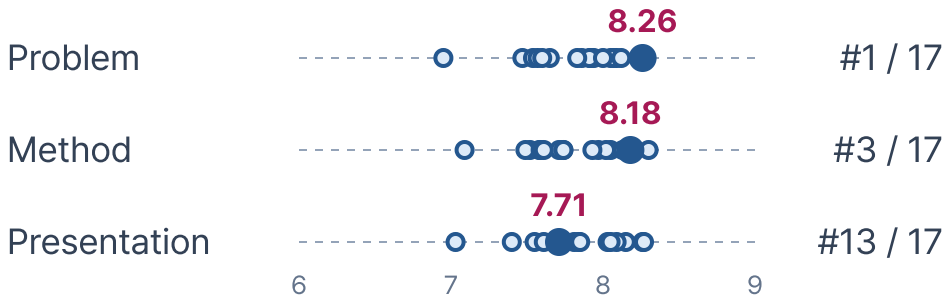
· Seven data issues diagnosed with numbers; 11,939 rows cut to 2,184 clean rows

WHAT WAS MISSING

· Modeling columns defined, but no database engine or store chosen

Team 08: Recall monitoring





Total: 8.26 + 8.18 + 7.71 = 24.15

AI feedback (slides only)

WHAT WORKED

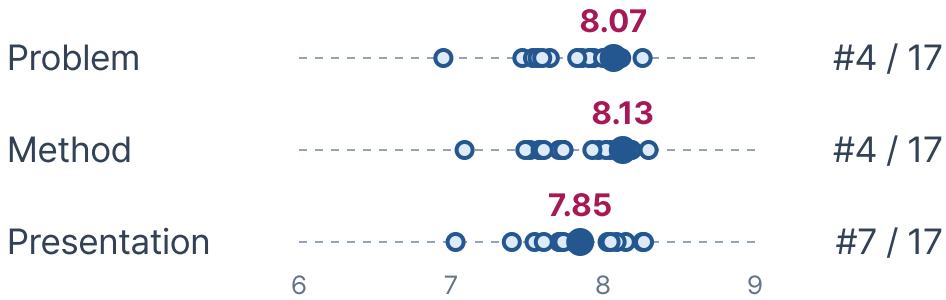
· 44,518 recall records from three APIs loaded into SQLite with four tables

WHAT WAS MISSING

· Seven problem-framing pages before the built pipeline and results

Team 10: Parking rules





Total: 8.07 + 8.13 + 7.85 = 24.05

AI feedback (slides only)

WHAT WORKED

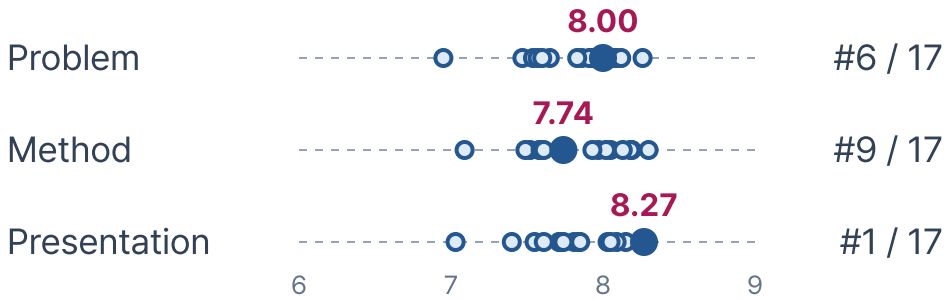
· Six public datasets joined to one row per building, 86,987 buildings, all counted

WHAT WAS MISSING

· Joined rows shown, but no database engine or primary keys

Team 15: Chest X-ray tail





Total: 8.00 + 7.74 + 8.27 = 24.01

AI feedback (slides only)

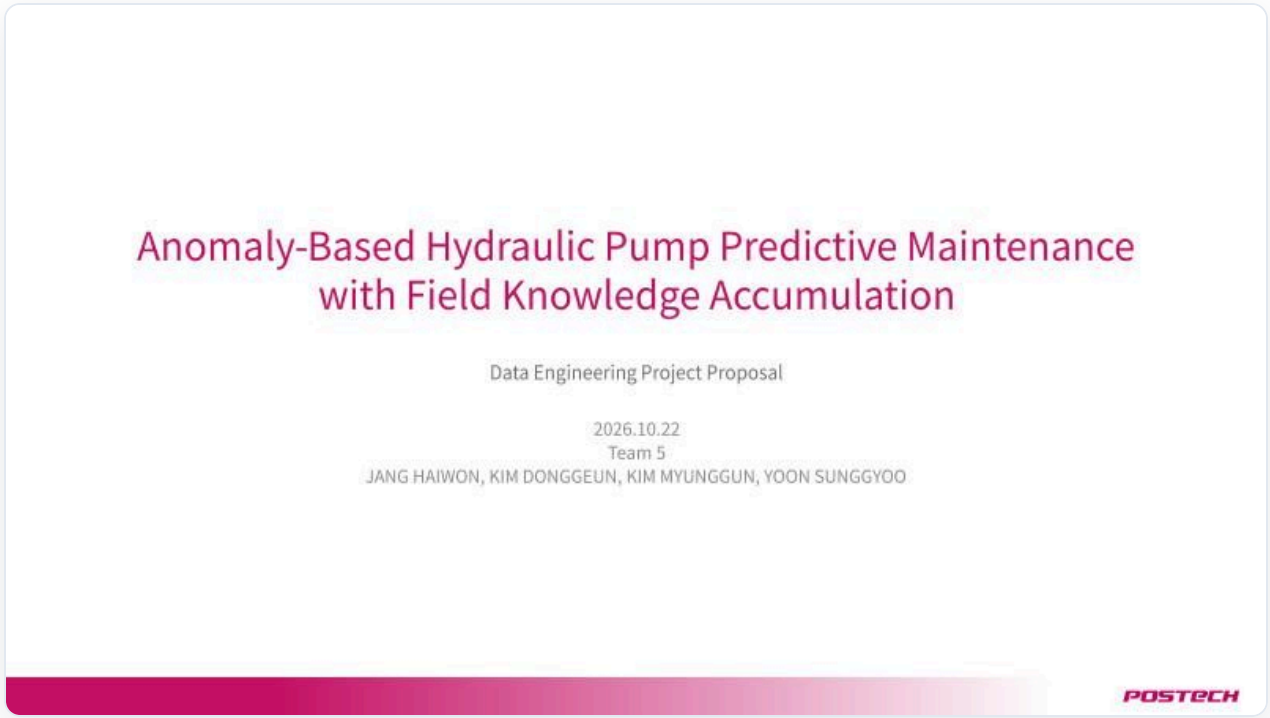
WHAT WORKED

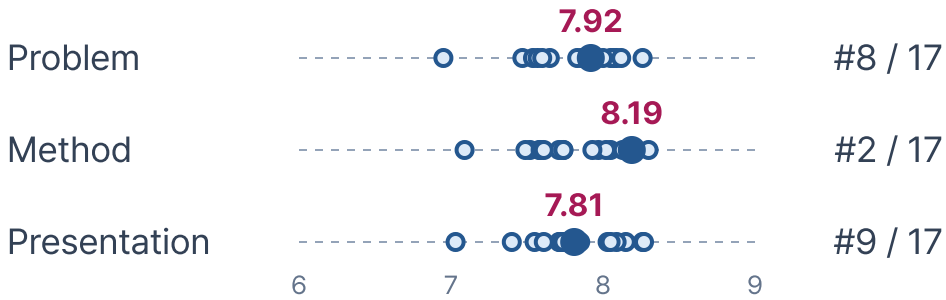
· Dataset profiled with numbers: 68,058 train images, 7 head classes hold 93.4%

WHAT WAS MISSING

· Synthetic generation method unnamed; no check that generated images are valid

Team 05: Hydraulic pump





Total: 7.92 + 8.19 + 7.81 = 23.92

AI feedback (slides only)

WHAT WORKED

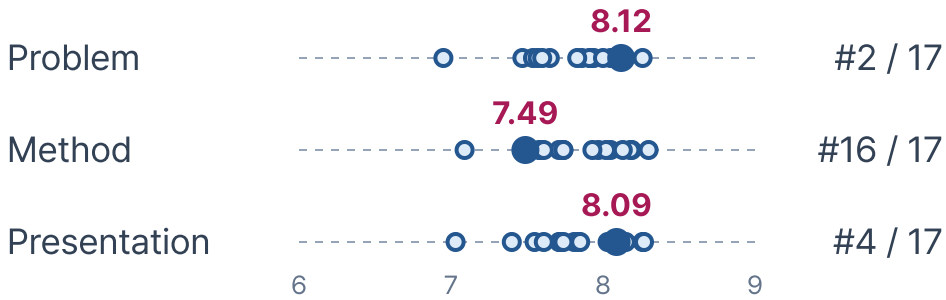
· Two-store design with candidates, selection criteria and four save checks

WHAT WAS MISSING

· No loaded sensor counts or results yet; demo still in preparation

Team 11: Meal planning





Total: 8.12 + 7.49 + 8.09 = 23.70

AI feedback (slides only)

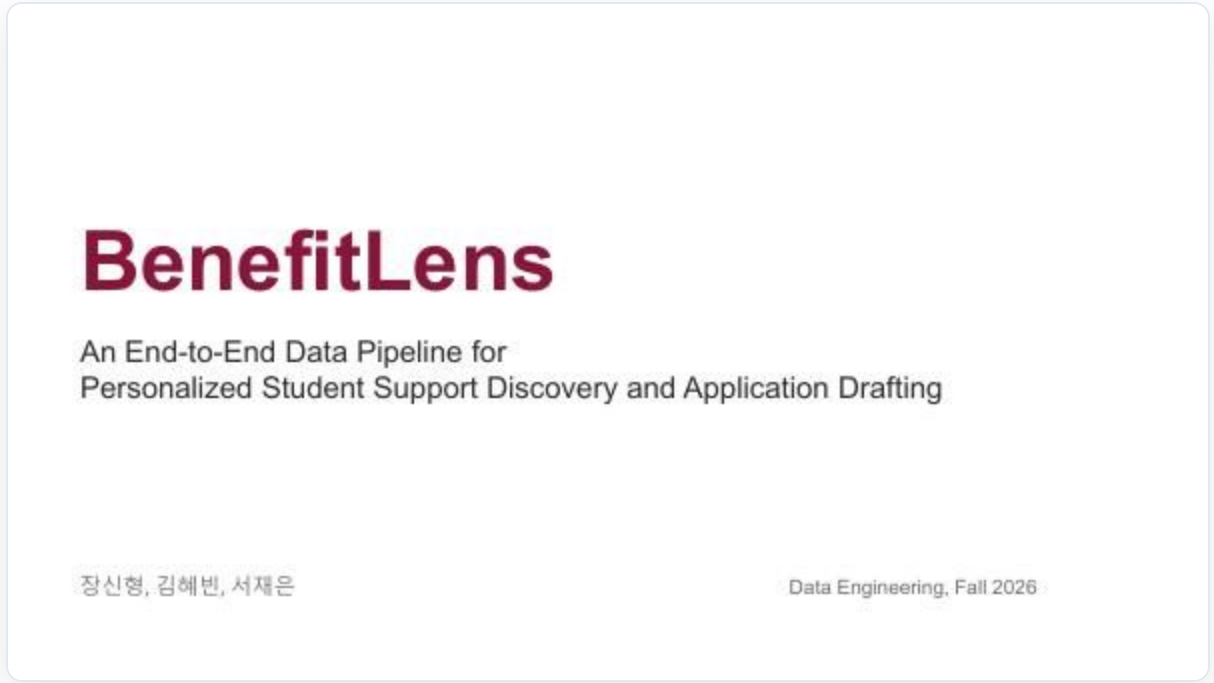
WHAT WORKED

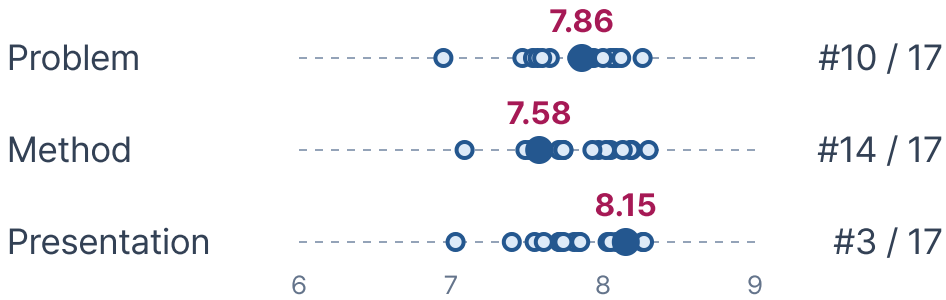
· Four stores designed: recipe, ingredient, rules in SQL; user profile as document

WHAT WAS MISSING

· Sample records only; collection volume and coverage not reported

Team 04: Student funding





Total: 7.86 + 7.58 + 8.15 = 23.59

AI feedback (slides only)

WHAT WORKED

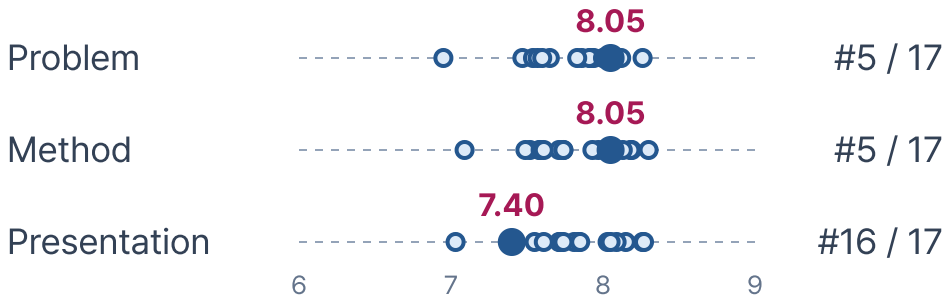
· One real notice already parsed into the JSONB schema with stored values

WHAT WAS MISSING

· Data so far is one program, three cases, five documents; no crawl count

Team 01: Highway icing





Total: 8.05 + 8.05 + 7.40 = 23.50

AI feedback (slides only)

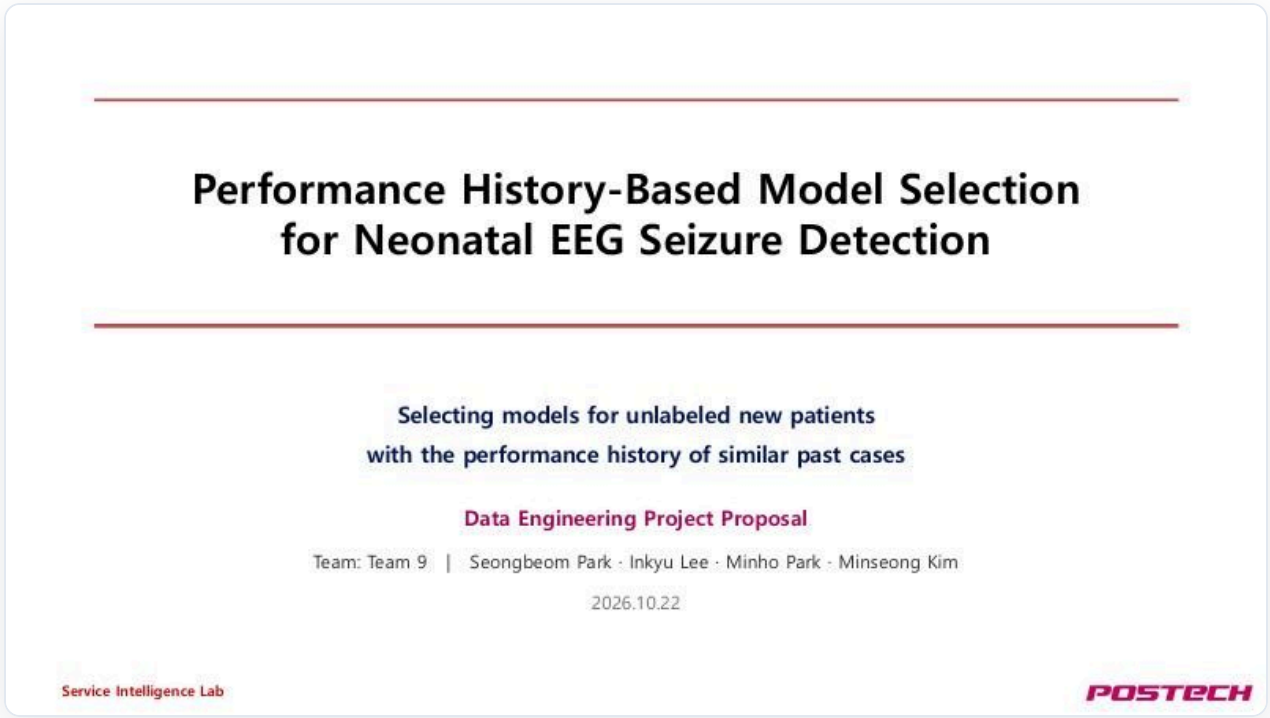
WHAT WORKED

· Seven sources with cadence, three held-out sensors, baseline and leakage-safe CV

WHAT WAS MISSING

· Wide scope: terrain model, risk prediction, live map and LLM agents

Team 09: Neonatal EEG



Performance History-Based Model Selection
for Neonatal EEG Seizure Detection

Selecting models for unlabeled new patients
with the performance history of similar past cases

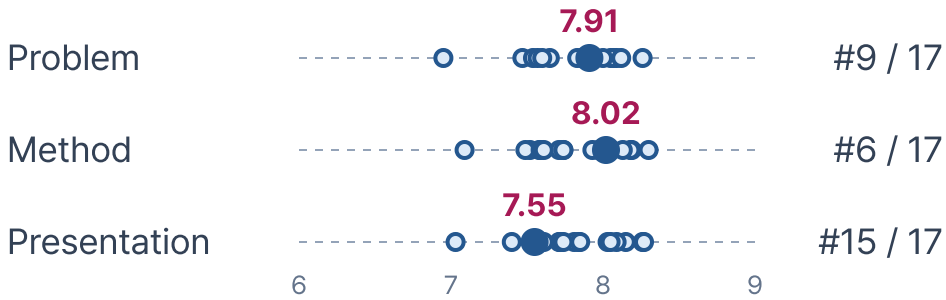
Data Engineering Project Proposal

Team: Team 9 | Seongbeom Park · Inkyu Lee · Minho Park · Minseong Kim

2026.10.22

Service Intelligence Lab

POSTECH



Total: 7.91 + 8.02 + 7.55 = 23.48

AI feedback (slides only)

WHAT WORKED

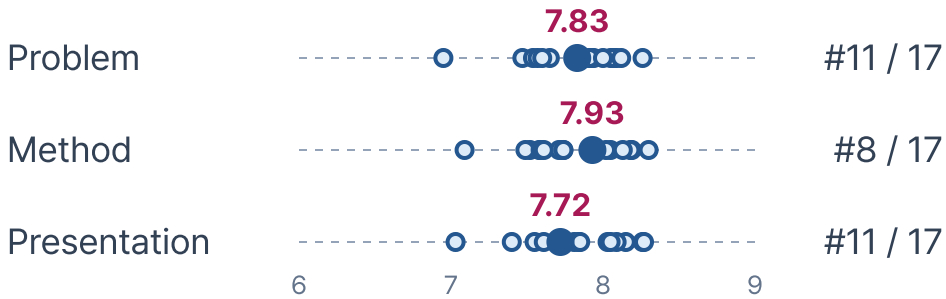
· Concrete public dataset: 79 neonates, 19 channels, 256 Hz, three annotators

WHAT WAS MISSING

· No model trained or selection performance measured yet

Team 14: 3D scanning





Total: 7.83 + 7.93 + 7.72 = 23.48

AI feedback (slides only)

WHAT WORKED

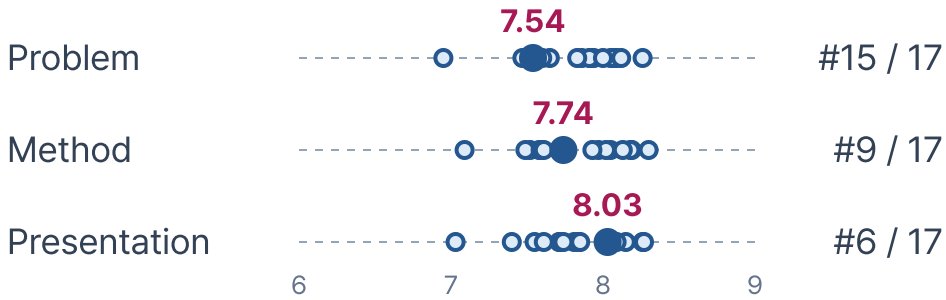
· Real EDA already run: coverage versus views for bottle and dragon

WHAT WAS MISSING

· Object dataset name and total size not specified

Team 12: Meme coins





Total: 7.54 + 7.74 + 8.03 = 23.31

AI feedback (slides only)

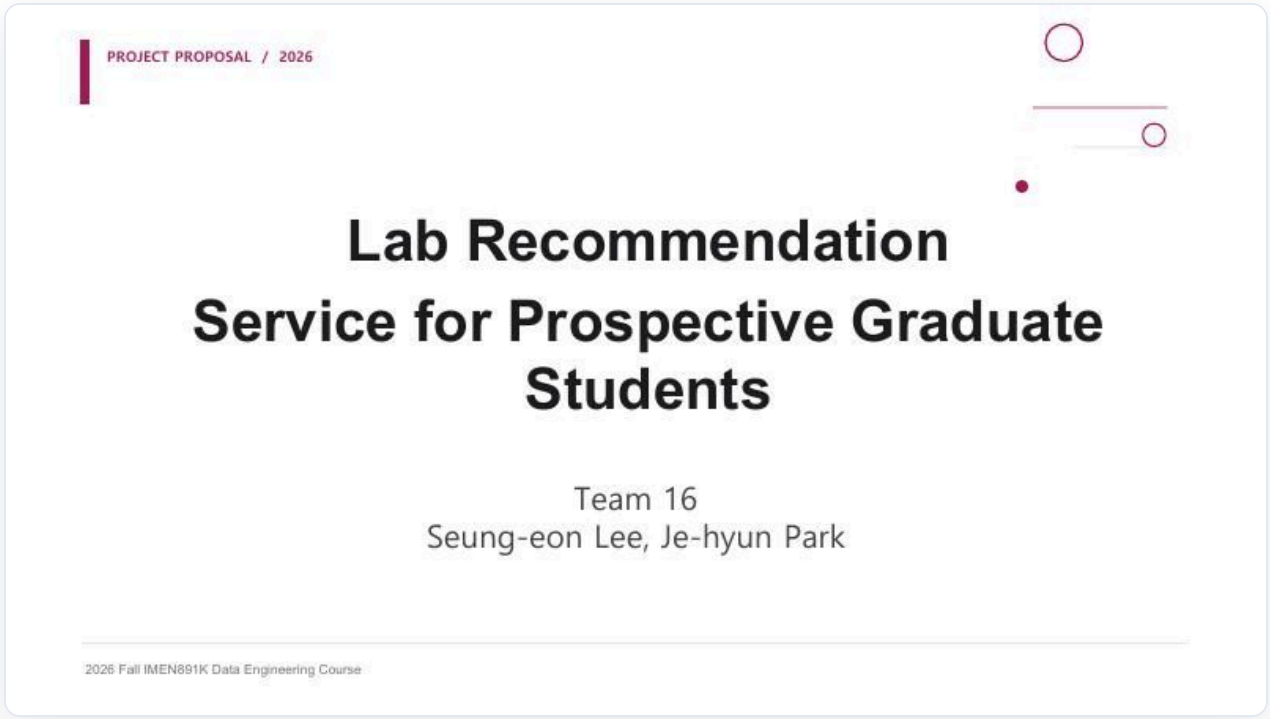
WHAT WORKED

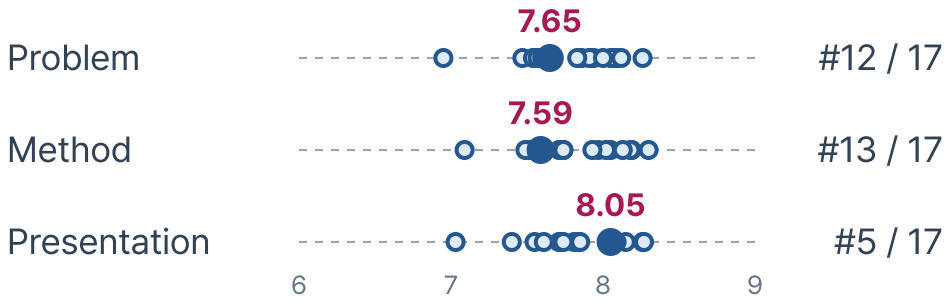
· 14,773 daily price rows for 10 coins already crawled from CoinMarketCap

WHAT WAS MISSING

· Social data not collected yet; Reddit access still pending

Team 16: Lab finder





Total: 7.65 + 7.59 + 8.05 = 23.29

AI feedback (slides only)

WHAT WORKED

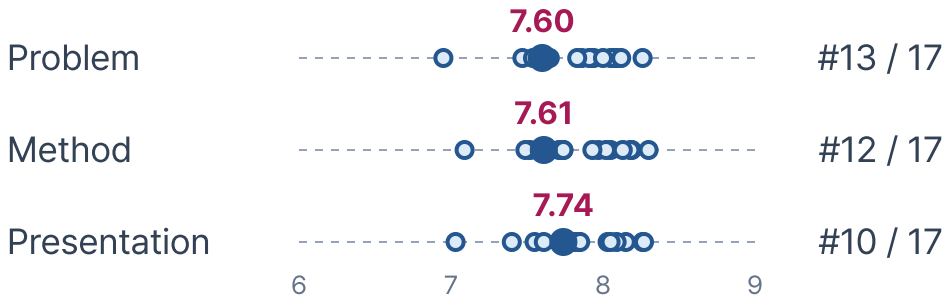
· Clear schema and raw-to-clean normalization rules with real DI page examples

WHAT WAS MISSING

· Seven sample rows shown; full crawl coverage and totals unreported

Team 17: Grocery restrictions





Total: 7.60 + 7.61 + 7.74 = 22.95

AI feedback (slides only)

WHAT WORKED

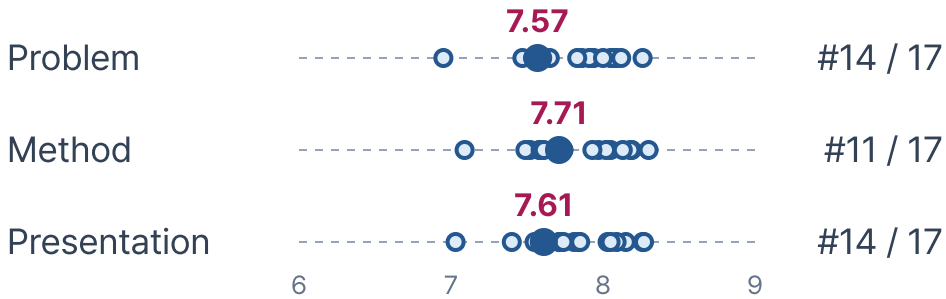
· 6.2 million products collected from seven official APIs and open datasets

WHAT WAS MISSING

· Dense slides and wide scope: meal plans, chat, OCR on top of catalog

Team 13: Experience paths





Total: 7.57 + 7.71 + 7.61 = 22.89

AI feedback (slides only)

WHAT WORKED

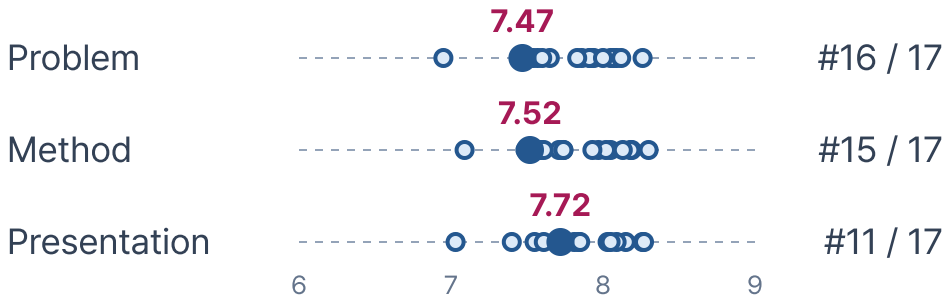
· Clear worked example: filter-delay case shows what the system returns

WHAT WAS MISSING

· No papers fetched or experience records counted yet

Team 02: Disclosure risk





Total: 7.47 + 7.52 + 7.72 = 22.71

AI feedback (slides only)

WHAT WORKED

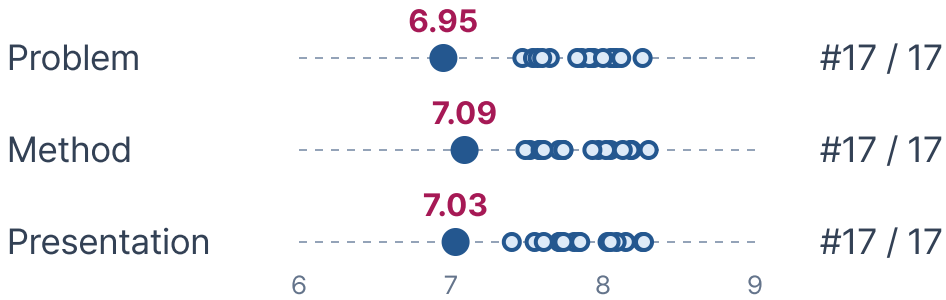
· Real OpenDART filings counted: 1,512 rights issues in one year

WHAT WAS MISSING

· Stores named, but no table definitions or keys shown

Team 07: Port analytics





Total: 6.95 + 7.09 + 7.03 = 21.07

AI feedback (slides only)

WHAT WORKED

· Clear pain point: BPA portal serves fixed tables needing domain knowledge

WHAT WAS MISSING

· Too abstract: agent workflow only, no port data loaded or schema

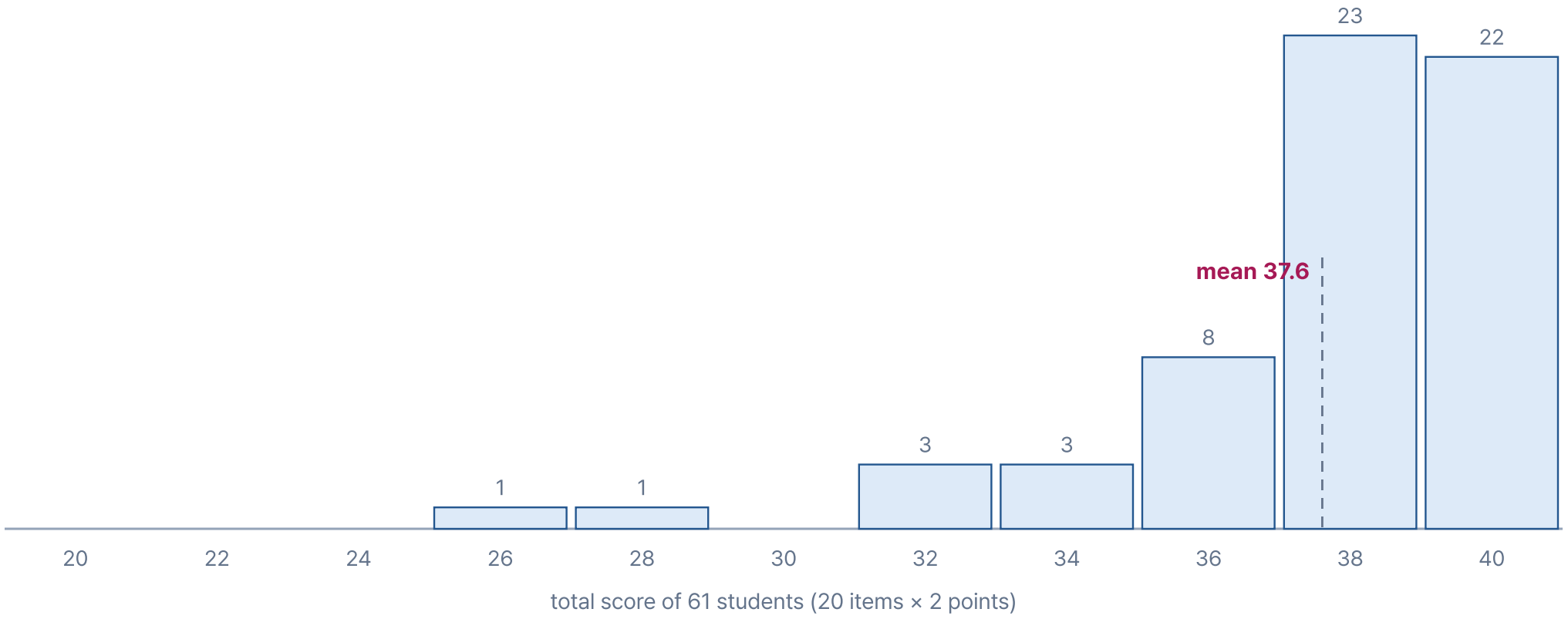
02 · Midterm Review

01 Proposal Review

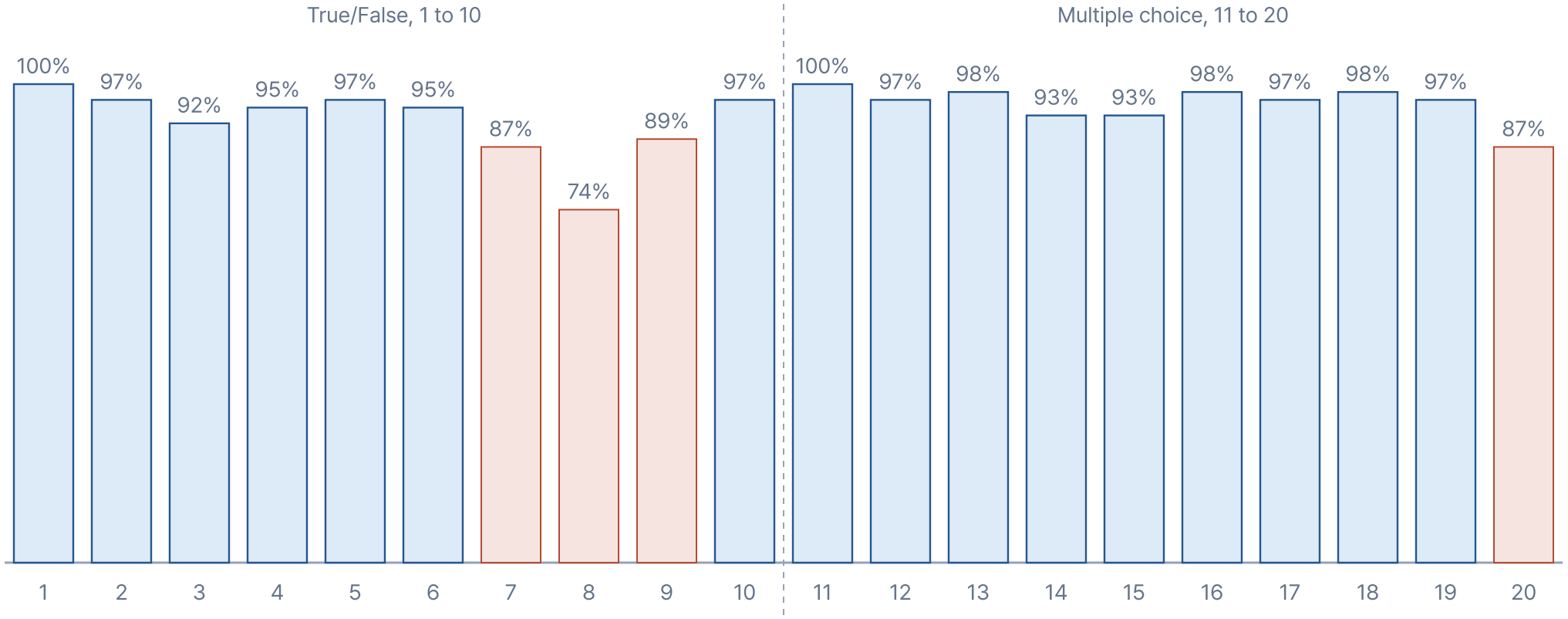
02 **Midterm Review**

03 By Course Code

Score Distribution



Item Scores


True/False, 1 to 10
Multiple choice, 11 to 20
100%
97%
92%
95%
97%
95%
87%
74%
89%
97%
100%
97%
98%
93%
93%
98%
97%
98%
97%
87%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Q1 Sensor as Encoder

A temperature sensor acts as an encoder: it converts a physical signal into a form that can be recorded digitally.

True ✓

False

100% right

Why. An instrument turns an event in the world into a record; a sensor encodes temperature as numbers

Q2 Correlation and Cause

Production volume and defect rate are strongly correlated. This proves that lowering production will lower the defect rate.

True

False ✓



97% right

Why. Correlation cannot say what happens under an intervention; a shared cause, such as rushed overtime, can drive both

Q3 Missing Values Before Filling

Some customer records have missing values. The team should fill them all before checking why they are missing.

True

False ✓



92% right

Why. Why a value is missing decides how to treat it; missingness itself can carry information

Q4 Schema Alignment vs Entity Matching

Two systems use different customer IDs. Giving their columns the same names is enough to match each customer.

- True
- False ✓



95% right

Why. Same column names align the schema; deciding which records are the same customer is entity matching, a separate step

Q5 Scaling Before Distance

A model measures customer similarity using age and income. Scaling helps stop income’s larger numbers from dominating the distance.

True ✓

False



97% right

Why. Distance adds squared differences; income in millions swamps age in tens. Scaling puts both on one footing

Q6 One-Hot and Order

A model must distinguish low, medium and high priority. One-hot encoding tells the model which level is higher.

True

False ✓



95% right

Why. One-hot gives each level its own column with no order between them; ordered levels need ordinal encoding

Q7 Selection vs Extraction

A team wants fewer inputs while keeping each variable's meaning. Feature selection does this by combining variables into new ones.

True

False ✓



87% right

Selection: keep a subset



Extraction: combine into new ones



Why. Selection keeps a subset of the original variables, so each keeps its meaning; combining them is extraction, such as PCA

Common slip: both reduce inputs, only extraction makes **new** ones

Q8 Keys Between Tables

A relational database can store each customer's details once. Orders can refer to that customer using a key.

True ✓

False



74% right

Customer

id	name
c1	Kim
c2	Lee

Order

order	cust_id
o1	c1
o2	c1
o3	c2

Why. Customers live once in their own table; each order carries the customer key, a foreign key

Common slip: copying customer details into **every** order

Q9 Flexible Documents

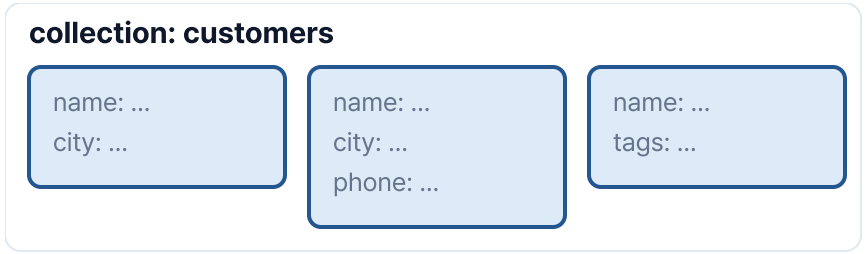
Some customer profiles have fields that others do not. A document store can keep these profiles in the same collection.

True ✓

False



89% right



Why. A collection does not force one schema; documents with extra fields sit side by side

Common slip: treating a collection like a table with **fixed** columns

Q10 Results as Inputs

A SQL query’s result can be a table in another query. A MongoDB aggregation stage can pass its documents to the next stage.

True ✓

False



97% right

Why. Both compose: a SQL result is a table another query reads; each stage hands documents to the next

Q11 API or Crawling

You need data from two sites: one has an API; the other only has web pages. Which methods fit the first and second sites?

A Web crawling; API calls	**B** API calls; web crawling ✓
C API calls; sensor measurement	**D** Synthetic data; web crawling

100% right

Why. Use the route the source offers: an API returns structured records; pages without an API need crawling

Q12 Typical Value Under Outliers

A few houses have prices far above the rest. Which measure of a typical price is less affected by these extreme values?

A Mean

B Range

C Standard deviation

D Median ✓



97% right

Why. The median depends on order, not size: a few huge prices move the mean, not the middle value

Q13 Same Statistics, Different Shapes

Two stores have the same means, variances and correlation for sales and ad spending. Why still compare their scatter plots?

A Matching statistics mean the data have the same shape

B Matching statistics mean the data have the same outliers

C Different patterns can produce the same statistics ✓

D A scatter plot cannot show anything beyond the correlation



98% right

Why. Anscombe's quartet: four datasets share mean, variance and correlation yet look completely different

Q14 Smoothing a Noisy Trend

Sensor readings fluctuate rapidly because of noise. You want a smoother trend on a dashboard. Which method fits?

A Moving average ✓

B One-hot encoding

C kNN imputation

D Z-score standardization



93% right

Why. A moving average replaces each point with the mean of its neighbors, damping short-term noise

Q15 Sampling Every Factory

You need a sample that includes every factory. Some factories have far more records than others. What should you do?

A Take the first records in the combined dataset

B Sample only from the largest factory

C Randomly sample at least one record from every factory ✓

D Remove unusually large or small values



93% right

Why. Group by factory and draw at least one record from each, so no factory is left out

Q16 SELECT-FROM-WHERE

This query builds a product list. What does the list contain?

```
SELECT name
FROM Product
WHERE category = 'Electronics';
```

A Names of Electronics products ✓

B All fields of Electronics products

C Names of all products

D The number of Electronics products



98% right

Why. SELECT picks the name column; WHERE keeps only Electronics rows

Q17 Basic JOIN

Every product has a matching company. This query builds a supplier report. What does it show?

```
SELECT p.name, c.country
FROM Product p JOIN Company c
  ON p.company_id = c.company_id;
```

A The number of products in each country

B Every possible product–company pair

C Each product and its category

D Each product and the country of its company ✓



97% right

Why. ON pairs each product with the company whose ID matches; company_id is unique, so one row per product

Q18 GROUP BY and Average

A dashboard runs this query. What does each result row describe?

```
SELECT category, AVG(price) AS average_price
FROM Product
GROUP BY category;
```

A Each product and its price

B One average price for all products

C Each category and its average product price ✓

D Each company and its average product price



98% right

Why. GROUP BY category makes one row per category; AVG(price) is computed inside each group

Q19 Access Pattern to NoSQL Type

A service must (i) find a session by its token and (ii) follow links between friends. Which types fit, in that order?

A Graph; key-value

B Key-value; graph ✓

C Document; wide-column

D Wide-column; document



97% right

Why. A token lookup is a get by key: key-value. Following friend links is traversal: graph

Q20 Aggregation Stages

Each sale records a city, product category and amount. You need total Electronics sales in each city. Which steps?

A Keep Electronics; group by category; sum amounts

B Keep Electronics; group by city; count the sales

C Group all sales by city; sum amounts; sort totals

D Keep Electronics; group by city; sum amounts ✓



87% right

$match category = Electronics

$group _id: city, total: $sum amount

one row per city; $sum works inside $group

Why. Filter first ($match), then group by city ($group), then sum the amounts ($sum)

Common slip: grouping by **category** gives one total, not one per city

03 · By Course Code

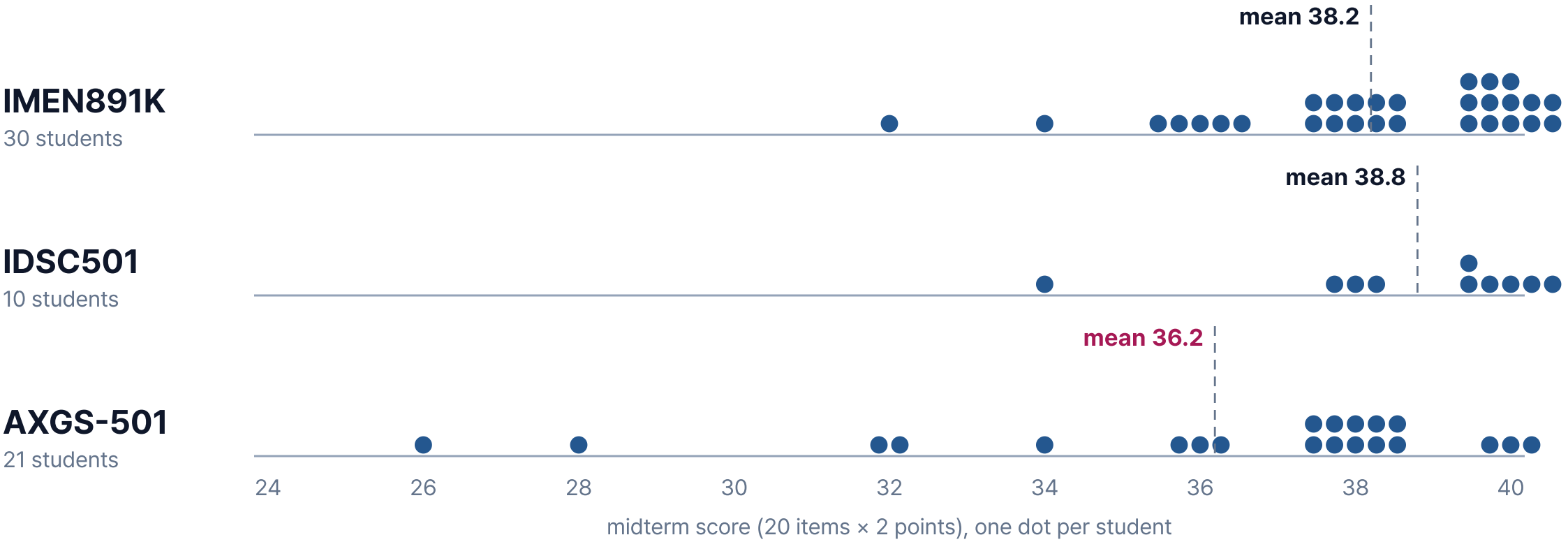
01 Proposal Review

02 Midterm Review

03 **By Course Code**

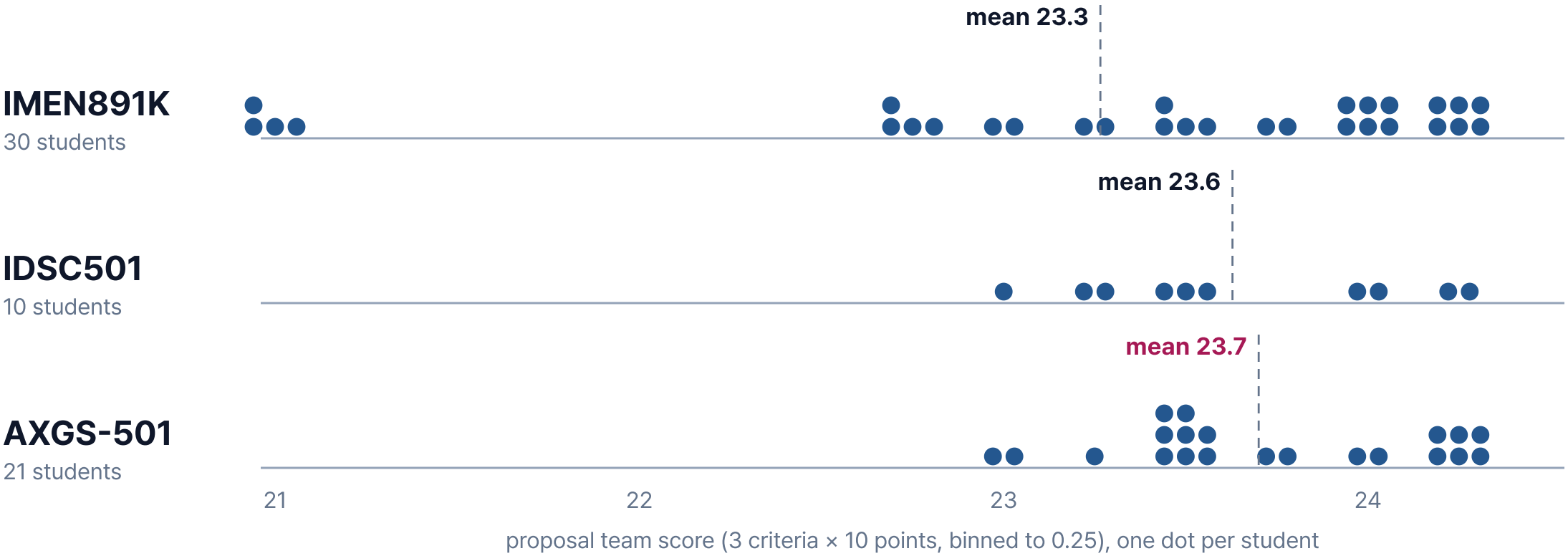
Midterm by Course Code

AXGS-501 averages 2 points lower, from a longer low tail



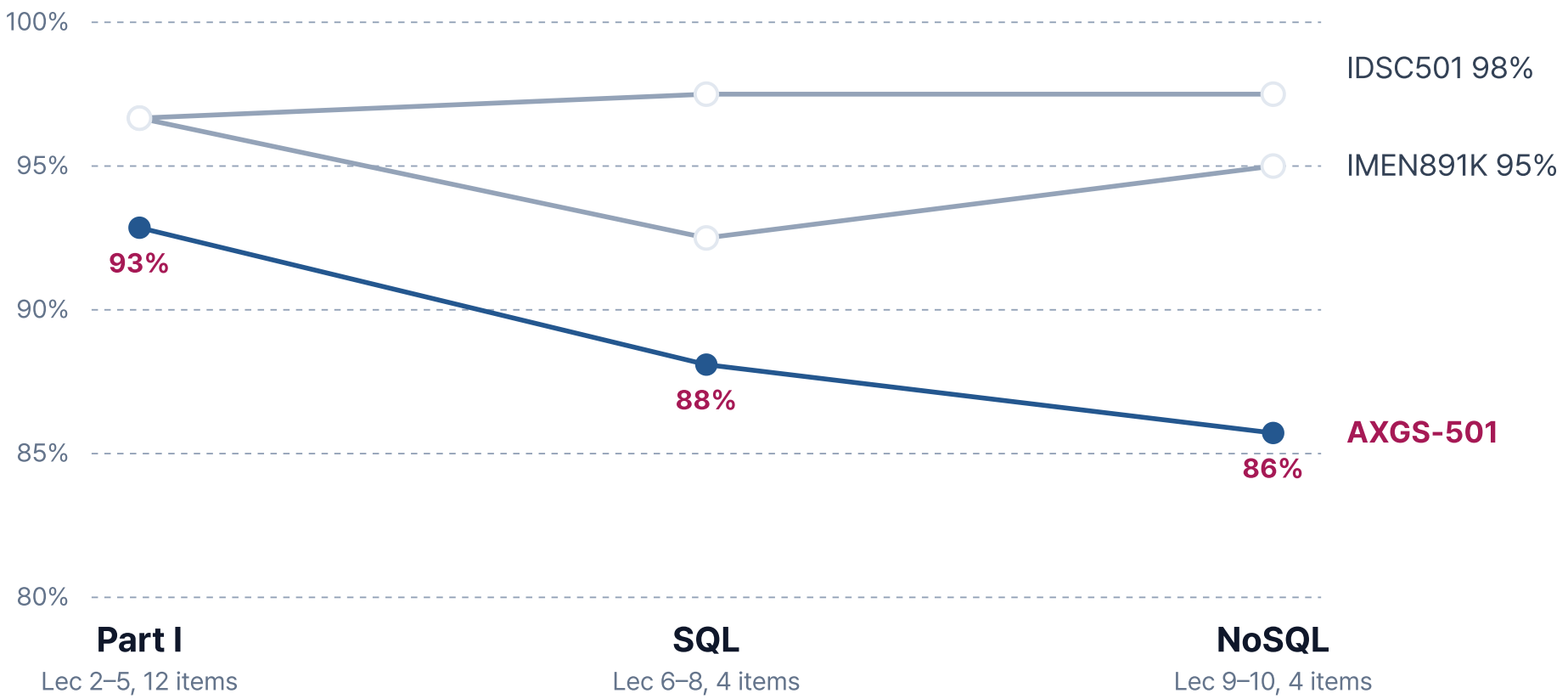
Proposal by Course Code

Team scores are shared, and the three codes sit within half a point



Correct Rate by Lecture Block

The AXGS-501 gap grows from Part I to SQL and NoSQL



Thank you for your hard work on the proposal and midterm!