MONASH University

WEEK 6 TUTORIAL ACTIVITY — BEAT YOUR BASELINE

Building AI That Works · LO6 · ~90 minutes · four lanes, one standard of proof

> **Goal.** Build the thin slice of your MVP's core feature, and **prove it beats the dumb thing that already works**. Everyone in the room makes the same five moves. Only the lane differs.

Teams of 3–4 (cross-campus via Google Meet). Deck: `BEX3202_week_06_prediction_pt2_activity_slides.pptx`. You need: a Google account with Gemini, your team's GitHub repo with an empty `DIARY.md`, a shared doc, and **the ten items you brought from the pre-class**. A notebook (Colab) is needed for Lane 1 only.

> **The scorer.** Move 4 is counting, not judging, so it is written for you: open `lane_tools/BEX3202_week6_scorer.ipynb` in Google Colab, run the setup cell, scroll to your lane, paste your numbers in. It prints your build next to your baseline and hands you the honest sentence for field 5. Building the thing (Move 3) is still yours — the scorer only counts what you decided.

FIRST: DECLARE YOUR LANE

Your project makes one of four kinds of claim. Pick the lane of **the one MVP feature you are building** — not of your whole product.

Lane	Your system…	Typical projects
1 · Predict a number	outputs a quantity	demand, stock, staffing, queue times
2 · Flag or classify	sorts things into buckets	fraud, verification, triage, churn, risk
3 · Extract from documents	reads a document and returns fields	forms, invoices, receipts, records
4 · Retrieve and generate	answers from *your* corpus	enquiry handling, policy Q&A, support

If your product is genuinely two lanes — many are — pick the one you are building **today**. The other lane is next week's problem.

THE FIVE MOVES (~75 MIN)

Everyone makes these moves at the same time. The whole class answers the same question at each step; only your answer is lane-specific.

MONASH University

#	Move	Time	What you do	✅ Done when
1	**Name your baseline**	7 min	Write down the dumb thing that already works without AI	It's on the board, in one line
2	**Build your eval set**	20 min	~20 items with the *right answer* recorded by hand	The link is on the board — **you cannot start Move 3 until it is**
3	**Build the thin slice**	30 min	Direct AI to build the one feature. Spec → prompt → test → iterate	It runs on your eval set without you editing it each time
4	**Run both, get the number**	20 min	Run the baseline *and* your build over the same eval set	You have two numbers that can be compared
5	**The honest sentence**	10 min	Write the sentence. Price one use	Lane Card fields 5 and 6 are filled

Move 2 is the one people skip. It is unglamorous and building is fun. It is also the only reason anyone should believe your result, so the room enforces it: no team starts building until its eval set is on the board.

TEAM DELIVERABLE — THE LANE CARD

One card per team, in the shared doc. Same six fields whatever your lane.

1. Our lane and our one feature: _____

2. Our baseline (the dumb thing that already works): _____

3. Our eval set: N = _____ · came from _____ · labelled by _____

4. The number: baseline _____ | our build _____ | metric _____

5. The honest sentence:

> *"On ____ [items], our ____ scored ____ against the baseline's ____. That **is / is not** enough to ____, because ____."*

6. What one use costs us: _____ · and at 100 uses/day: _____

Field 5 is the point of the whole week. Field 6 is the question almost nobody answered in Assessment 1 — exactly one proposal in fifty-seven gave a real unit cost.

TWO RULES, ENFORCED

1. **Beat the benchmark.** If your build cannot beat the dumb thing, the dumb thing is your product and you have saved yourself a term of work. That is a *result*, not a failure — report it.
2. **Never test on what you built on.** Whatever you tuned against, tested against or looked at while prompting is contaminated. Your eval set must be items the build has never seen.

LANE ANNEXES

Find your lane. Everything else in this document applies to everyone.

Lane 1 — Predict a number

Baseline	Seasonal-naive (same period last year), last value, or the group mean. Whichever is most defensible for your series.
Eval set	The **last 12 periods**, held out in time order. Not a random sample.
The number	MAE, or MASE if you want it scale-free. Lower beats baseline.
The trap	**Shuffling time.** If your split is random, your model has seen the future and your result is fiction.
Free route	Colab + pandas. A lag/trend model is enough today.

Fallback data: `BEX3202_week_06_prediction_pt2_sample_data_cafe_sales.csv` *Optional depth (post-class):* `01_weeks/week_05_prediction_pt1/forecasting_lab/` — R and Python, nine scripts each. **Not today.** Today's build is a simple lag/trend model; the lab is for after class.

Lane 2 — Flag or classify

Baseline	Majority class ("always say legitimate"), or **one hand-written rule** you can state in a sentence.
Eval set	20 cases **you labelled yourself**, including **5 you find genuinely ambiguous**. The hard ones are where the value is.
The number	Of 20, how many right — **and which kind of wrong**. Count false alarms and misses separately; they are not the same mistake.
The trap	19 of your 20 cases are "fine", the model says "fine" every time, and scores 95%. You met this in Week 5 as the accuracy lie.
Free route	Gemini with your rule written into the prompt, or a simple rule in a spreadsheet.

Then the Week 6 question, which Week 5 did not ask: **where do you set the threshold, and what happens to the cases you are unsure about?** Most working systems route the uncertain middle to a human review queue rather than guessing. Say what yours does.

Fallback data: `lane2_refund_claims.csv`

Lane 3 — Extract from documents

Baseline	A regex or template — or honestly, **a stopwatch on someone typing it in by hand**. That is what you are replacing.
Eval set	10 real documents where **you typed out the correct field values first**. That hand-typing *is* your ground truth; there is no shortcut.
The number	Fields correct ÷ fields total. Plus seconds per document, against the manual time.
The trap	Testing on the same document you tuned the prompt against. It will look perfect and mean nothing.
Free route	Paste the image or PDF into Gemini with a **named field schema** ("return JSON with these exact keys: …") and ask for structured output. No OCR library, no install.

Fallback data: `lane3_documents/` (6 documents + `answer_key.csv`)

Lane 4 — Retrieve and generate

Baseline	Keyword search over the same corpus — or "just link them to the FAQ page".
Eval set	20 real questions, each with the **correct answer** *and* **the passage it comes from**.
The number	Of 20, how many were answered correctly **and** cited the right source. Both, not either.
The trap	Two of them. **Grading by vibes** — it sounds right, so you tick it. And **answering from model memory** rather than your corpus, which works until the question is about something only your documents know.
Free route	Upload your folder to Gemini, and require a citation in every answer.

Fallback corpus: `lane4_corpus/` (café policies + `questions.csv` with answers and source passages)

DEBRIEF (~10 MIN, WHOLE CLASS)

One team per lane reads out field 5. Four different problems, four different metrics — and one identical shape. That shape is the week: **name the baseline, build the eval set, get the number, say the honest sentence.**

It is also what Assessment 3 rewards. Criterion 3 asks whether you tested honestly and fed the findings back. Today is the first entry in that record.

Before you leave: commit diary entry 1. Not tonight — now.

MONASH University

FACILITATOR NOTES (DO NOT DISTRIBUTE)

Run the clock in common, the content in lanes. Every team answers the same question at the same minute. Lead A calls the moves to the whole room in lane-agnostic language and never enters a breakout. Lead B owns Meet, parity, and lane routing.

Lead A — the metronome. At each move boundary, read two or three answers off the shared board aloud: one strong, one weak, one from the other campus. Fix the weak one live. The 7-minute baseline call is the highest-value block in the two hours — protect it and enforce the stop.

Lead B — lane sweep, not team sweep. Visit all Lane 3 rooms in sequence, then all Lane 4, and so on. One lane's context at a time beats context-switching per room. Maintain the lane map. **Gate Move 3 via the board** — check for the eval set link rather than asking teams whether they've done it.

Lane 2's fallback dataset is rigged for the debrief — use it. On `lane2_refund_claims.csv` (60 claims, 13 suspicious):

Approach	Accuracy	Suspicious caught
"Always say genuine"	**78%**	**0 of 13**
Rule: new account **or** over $300	58%	9 of 13 (wrongly flags 21)

The do-nothing baseline scores twenty points higher and catches nothing. Put both rows on screen and ask which one they would deploy. The answer is obviously the *worse-scoring* one — and that is the entire Week 6 move from "what's the accuracy?" to "which error costs more, and where do we set the line?"

The five traps to catch live:

- Any random split in Lane 1 (leakage — the most common error in the room).
- A Lane 2 team celebrating 95% on 19 negatives.
- A Lane 3 team testing on their tuning document.
- A Lane 4 team grading by vibes, with no source passage recorded.
- Any team that declares victory without ever running the baseline.

If a lane has one team, Lead A takes it personally rather than leaving it uncoached. **If a lane has one team per campus**, merge them into a cross-campus lane table.

Pre-flight (before class): confirm document upload works on the real Monash Gemini endpoint **from both a Malaysian and an Australian account** — Lane 3 depends on it entirely. If it fails or throttles, drop Lane 3 to 3 documents each and say so at the top of the hour.

If a team has no repo, that is 20 minutes gone. Check the pre-class board post the day before and chase the teams who haven't posted.

Contingency: this sheet works on paper. If Meet or Gemini fails, teams do Moves 1, 2 and 5 by hand — name the baseline, build the eval set, write the sentence. That still captures most of the LO6 value; only Move 3 needs the tooling.

Do not release Assessment 1 feedback the morning of this seminar — the first eight minutes will be about marks instead of builds.