

Flock: Forecasting How Busy a Venue Will Be, and Counting How Many People Are There

Jayden Bansal

Flock Social LLC, Pennsylvania, USA

social@flockcorp.com

October 2026. A snapshot before the early-November retrain; to be updated with its held-out results.

Abstract

When friends make plans to meet up at a restaurant or other venue, they need an indication of the expected level of crowds in the venue. Google Maps offers a popular-times chart indicating the typical level of crowds at a given hour for a particular location, but this does not reflect the level of crowds on a particular night. This paper addresses two questions. Can AI tell how crowded a venue will be before you go? And can a credit-card-sized computer model be designed to count the number of people inside a venue using a thermal camera without identifying individuals? For the first question, my app Flock uses recent live data to correct a typical weekly crowd level pattern at each venue. Crowd levels are then assigned to one of five possible crowd levels: Quiet, Not Busy, Steady, Busy, or Packed. This prediction was tested on 119,939 instances of data that had not previously been used for this prediction. When the prediction was made one hour in advance, it correctly predicted the crowd level for the next live data provided by the data provider with a 54.9% success rate compared to a 30.2% success rate for the popular-times chart. On a 0-100 scale, this prediction also outperformed the popular-times chart, reducing the average error by 31.8% compared to the popular-times chart. This improvement largely comes from the most recent live data for each venue. On 13 October, this model switched to a simpler model that is based on the most recent live data. This also places a limit on the predictive capability of this model because for predictions made two to 24 hours in advance, where the most recent live data is farther behind, the simpler model only outperforms the popular-times chart by about three percentage points. I also tested this on real crowd data. Using 5.3 million publicly available hourly data records from 7 countries that were mostly individual counts of people, I found that using the most recent count to correct the weekly crowd level pattern resulted in the lowest one-hour error for 14 of 15 datasets. For the second question, the live data provided by the data provider are estimates and not actual crowd counts. For this question, I designed a small neural network called Owl. A neural network is a model that learns to recognize something by being shown many labeled examples instead of following rules that were written by a human. Owl learned to detect people in images from a low-resolution thermal camera. Owl runs on a Raspberry Pi, which is the size of a credit card. Owl was 94.3% correct on 300 previously unseen images of an office environment, showing either one or two people. The hand-written rule previously used to detect people (identifying warm spots) was only 40.7% correct. When tested on images from 12 other publicly available datasets, Owl was 40.4% correct compared to the 33.0% correct from the hand-written rule. However, Owl still does not outperform the hand-written rule on images containing three or more people in these combined datasets.

Keywords: crowd prediction; popular times; nowcasting; gradient boosting; thermal imaging; people counting; on-device inference; privacy by design

1 Introduction

1.1 Background and motivation

A group deciding where to go tonight wants one fact about each option: whether it will be busy when they arrive. The common answer in consumer maps is the popular-times chart: the venue's typical level for each hour of each weekday, averaged over past weeks, sometimes with a coarse live indicator beside it. I call this weekly pattern, lightly smoothed (Section 4.1), the venue's curve. Möhring et al. studied Google's chart as a behavioural signal [11]. The curve is a strong first guess, but by construction it cannot see the specific evening: the rain, the game two blocks away, the holiday, the week the place became popular.

I built Flock, an app for planning nights out with friends. Its crowd card shows one integer from 0 to 100, the served number, and one of five crowd levels: *Quiet*, *Not Busy*, *Steady*, *Busy* and *Packed*. The number is relative: 100 is the busiest that venue gets, and it does not count people. Whether that number is right decides whether a plan works. This project measures how much better than popular times that number can be, and what it takes.

1.2 Research question

Can AI tell how crowded a venue will be before you go? And can a credit-card-sized computer model be designed to count the number of people inside a venue using a thermal camera without identifying individuals? Sections 2 to 8 answer the first question, made testable as: for the venues and hours a working app actually serves, can a system that watches each venue's recent live readings beat the venue's own curve, and by how much? Sections 9 to 18 answer the second: can a neural network small enough for one core of a Raspberry Pi count the people in a room from a thermal camera better than the heat rule it replaces, with only counts leaving the device?

1.3 Hypotheses

H1 to H3 test the first question; Owl's benchmark and sealed tests test the second (Section 14).

- **H1.** Correcting the venue's curve with its own recent live readings reduces error against held-out live readings, on both the 0 to 100 scale and the five-level scale.
- **H2.** A model that predicts how an hour departs from the curve, using context such as weather, events and calendars but no recent history, beats the curve on cities it never trained on.

- **H3.** The value of a live reading decays within hours, so the weight it deserves falls with its age.

1.4 Contributions

For the first question, the main finding is that a venue’s freshest live reading is worth more than any correction to its curve: adding it lifts within-10 accuracy, the share of predictions within 10 points, from 31.2% to 55.2% (Section 6.6). Where the venue was read an hour earlier, the served number lands on the exact level 77.0% of the time (52.9% where the reading had changed since, against 29.1% for popular times; Sections 6.1 and 6.3). A venue’s departure from its curve correlates 0.74 from one hour to the next and 0.26 across two (Section 3.2), so these are one-hour results; from two to 24 hours ahead the corrected curve keeps a smaller, steady lead (Section 6.4). The contributions are:

- A serving path, in production since 26 September 2026, that corrects the venue’s curve with an offset (its recent median gap from the curve) and a nowcast (its latest reading, weighted by age), every weight fitted strictly before the readings that score it. On 119,939 live readings its weights never saw, it names the exact level one hour ahead 54.9% of the time, against 30.8% for the gradient-boosted model it replaced and 30.2% for popular times, and cuts the popular-times mean absolute error (MAE) by 31.8% on the provider’s relative 0 to 100 readings (Section 6.1). Persistence baselines (repeating the latest reading) locate the gain there, and a simplified rule built on that finding has served production since 13 October.
- Three properties of the curves on this provider’s readings: the curve is poorly calibrated, observed busyness rising about half a point per point of curve; departures stay correlated for one to two hours; and within-10 accuracy partly rewards landing on multiples of 5, as the readings do (Section 3).
- An external replication on 5,321,507 hourly records of counts and occupancy estimates in 315 dataset-site series in seven countries (1,787,058 scored hours), its analysis plan fixed before any data were downloaded: on counts the curve is close to calibrated and departures last longer, and adding the latest departure to the next hour’s curve still helps (Section 6.8).
- A gradient-boosted decision-tree model of the departure from the curve (the delta model) that beats predicting the mean, R^2 above zero, in three cities it never trained on (Section 6.2), with a complete account of its inputs (Section 5).
- Checks that outlast any one model: a gate that ships a model only if it beats what the app serves, tests that training and serving compute the same inputs, a dataset that records where each reading came from, and a retrain test whose split rule and thresholds were committed before its data existed, with a comparator chosen later (Sections 4 and 7.2).

1.5 Outline of the paper

Sections 2 to 8 answer the first question with the forecast: related work, preliminary findings, methods, what goes into the model, results, and discussion with the November retrain. Sections 9 to 18 answer the second with Owl, the people counter that runs on Flux, Flock’s venue sensor: it counts the people a thermal camera can see and sends a few numbers, never an image. Section 19 records how

both systems were built, Section 20 what limits each answer, and Section 21 the measurements that would strengthen them: the November retrain, and joining the two studies so that the forecast is checked against real counts. Section 22 answers both questions. Appendix A lists the choices made after seeing data.

2 Related Work

Sections 2 to 8 answer the first question.

Point-of-interest visit forecasting. The closest published task forecasts hourly visits to named points of interest (POIs). BysGNN [7] forecasts hourly visits to 400 POIs in each of five US cities from SafeGraph data, whose hourly visits also underlie mobility network models of epidemic spread [3]. It reads the previous 24 hours of visits at every venue in a city and forecasts the next six hours, a live feed a consumer app rarely holds. Against a Historical Average baseline (the same weekday over the previous month) that its authors liken to the popular-times chart, its MAE is 40% to 54% lower across the five cities at its 400-venue scale (Houston 8.860 to 4.095 visits); against last week’s value at the same hour the gain is 14% to 19% (both computed from its tables). Its more often quoted 4.34% to 6.71% margin is over the strongest competing graph model. Its mean absolute percentage error of 60% to 80% shows what precision the task admits, though that measure breaks down near zero [8]. Timokhin et al. [16] estimate hourly venue popularity from venue attributes and compare popular-times data with WiFi counts.

Occupancy prediction and granularity. Bollenbach et al. [2] report R^2 up to 0.87 for whole-beach visitor counts one step ahead, while section-level occupancy forecasts one to three days ahead have R^2 near zero and often below it (−1.26 to 0.42 across the best models). Following Moore et al. [12], they read R^2 (the share of variation a model captures) above 0.3 as weak, 0.5 as moderate and 0.7 as strong. My task sits at the granularity a crowd card answers: one venue and one hour.

Beating the trivial baseline is not guaranteed. In a beach-occupancy study [1], a feature-based random forest and a seasonal time-series model (SARIMA) both had higher MAE than naive persistence, though both had lower root mean squared error, which weights large misses more. In weekly foot-traffic forecasting, a simple weighted average beat an LSTM neural network [9]. These results are why every candidate in this project is scored against the served number and against persistence of the latest reading, the standard naive benchmark [8], on the same rows (Sections 4.5 and 6.1).

3 Preliminary Findings

While designing the correction, I measured the curve against verified live readings (Section 4.2). Each finding below fixed one design decision.

3.1 The curve is a poorly calibrated forecast

I joined every verified live reading to its venue’s curve value for the same weekday and hour, giving 9,903 paired rows. A reading’s departure from the curve flips sign as the curve rises (Figure 1). In hours whose curve value is at most 25% of the venue’s daily peak, observed busyness sits 15.60 points *above* the curve (1,162 rows). In hours above 75% of the peak it sits 13.56 points *below* (5,746 rows). The middle quarters deviate by +1.68 and −4.99.

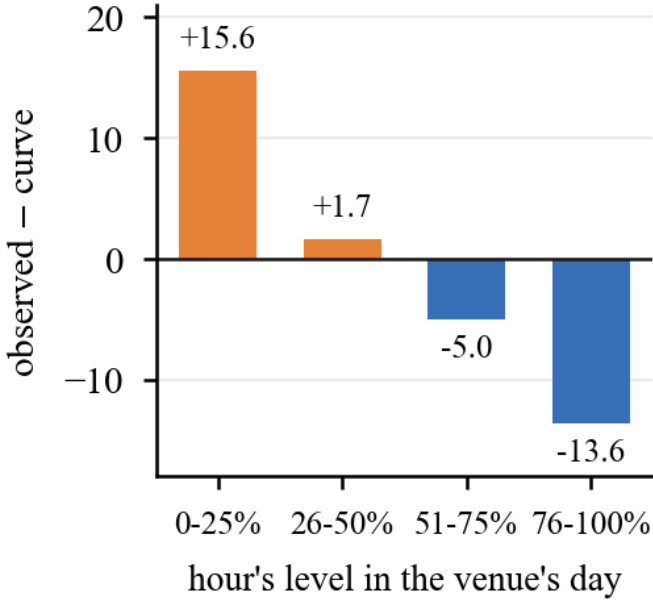


Figure 1. Observed minus curve, by the hour's level in the venue's own day (9,839 of the 9,903 paired rows).

Regressing observed busyness on the curve gives $observed = 18.82 + 0.466 \times curve$, with correlation $r = 0.372$ (Figure 2): each point of curve is worth about half a point of observed busyness, and the curve explains about 14% of the variance in live readings. The readings themselves are more spread out than the curve (standard deviation 33.2 against 26.4), and the slope is low because the correlation is low. The best forecast from such a curve shrinks it toward the mean, the regression effect in prediction [5, 6]. The 30-day test window (9 September to 8 October) repeats the pattern: $observed = 18.62 + 0.497 \times curve$, $r = 0.360$, with quiet hours 11.5 points above the curve and busy hours 13.9 below. The same effect explains why venues' offsets averaged about -11 (1,048 venues, 2026-09-06), which I had first read as a plain bias; on these 9,903 rows the mean departure is -6.3.

Design decision: one global offset cannot move quiet hours up and busy hours down, so the correction works per venue and uses the latest reading; a slope that bends the curve itself is weighed in Section 6.6.

3.2 A venue's recent state persists, then fades

In the first week of live collection (1 to 8 September) the provider's live readings were sticky: 80.9% of consecutive-hour readings of a venue were identical, and a venue's departure from its curve correlated 0.815 across one hour, 0.487 across two and 0.203 across three (Figure 3). Used as the prediction for the next hour, a reading one hour old landed within one crowd level 92.0% of the time and within 10 points 85.9% (3,165 readings). Nothing in a venue's curve carries this information. **Design decision:** the latest reading should enter the served number directly, at a weight that decays on the same timescale.

The stickiness did not last. Identical consecutive readings fell to 69.3% from 9 to 25 September, 44.8% from 26 September to 2 October and 47.5% from 3 to 8 October, and over the 30-day test window the departure correlates 0.739, 0.255 and 0.158 at one, two and three hours. Whether the early stickiness came from how often the provider refreshed its value or from when the collector read it cannot be told from these data.

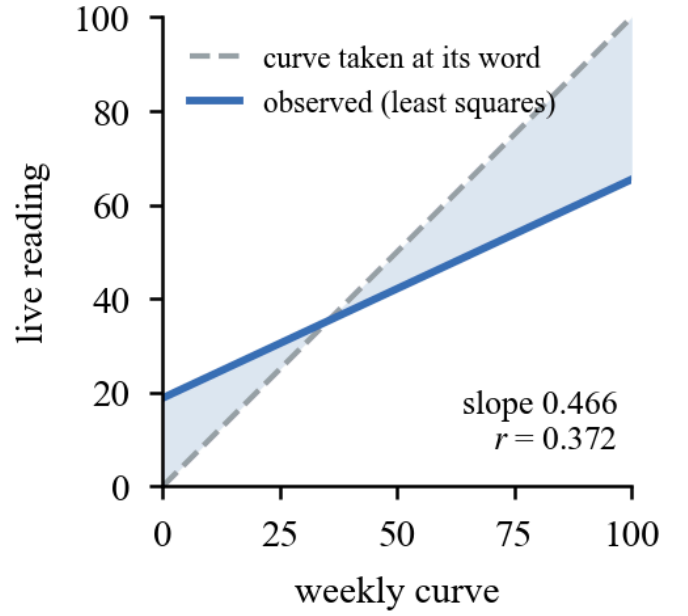


Figure 2. The least-squares fit of live readings on the curve (9,903 rows), against the identity line.

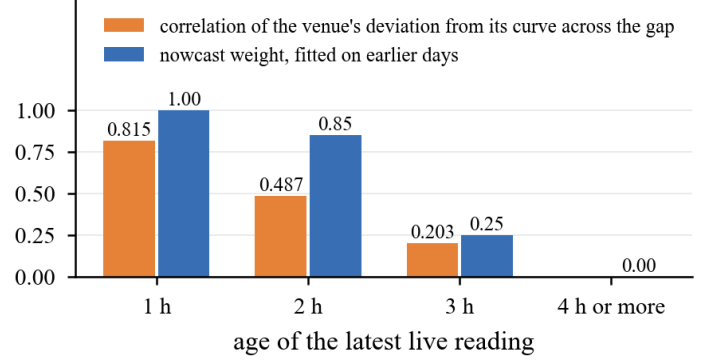


Figure 3. How a venue's departure from its curve persisted over 1 to 8 September, against the nowcast weights fitted on 1 to 5 September (Section 4.4). The fitted weights fall with age, more slowly than the correlation at two hours. Over the 30-day test the correlation at two hours fell to 0.26.

3.3 The within-10 metric rewards the grid

The provider's readings lie on a grid, or lattice, of multiples of 5, and the stored weekly value (popular times) is the provider's typical week as given, one per venue, weekday and hour, so it sits on the same grid. A forward split shows what that does to the metric. I fitted each venue's offset on its earlier readings and scored its later ones (894 venues, 3,930 later readings) at offset weights λ from 0 (no offset) to 1 (Figure 4).

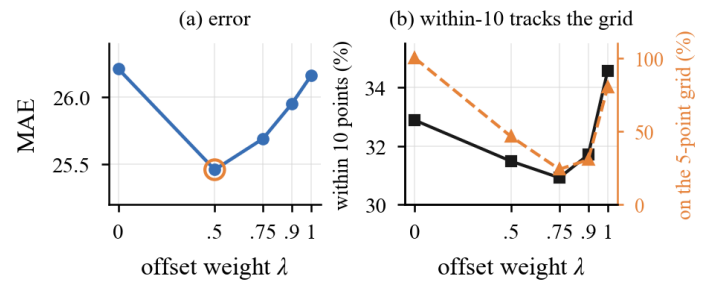


Figure 4. The forward offset test. (a) MAE is best at $\lambda = 0.5$. (b) Within-10 peaks at the two weights whose predictions land most often on the 5-point grid, not where the error is lowest.

MAE is best at $\lambda = 0.5$ (25.46, against 26.21 with no offset). Within-10 is highest at the two weights whose predictions most often land on the grid: $\lambda = 1$ (34.55%, 79.9% on-grid) and $\lambda = 0$ (32.88%, 100% on-grid). The median error is 20 at every weight. On grid-valued data, within-10 partly measures grid coincidence. **Design decision:** where a gridded and a non-gridded predictor are compared on within-10, the comparison reports the on-grid rate or projects both to the grid (Section 6.5). Production’s within-10 margins in Table 5, 24 points over popular times and over curve plus offset, are far larger than the within-10 variation across the offset weights tested here.

4 Methods

This section defines the task and the tests by which Section 6 judges H1 to H3.

4.1 Definitions and problem statement

Busyness and crowd levels. For a venue v and an hour h , busyness $y \in [0, 100]$ is a relative score normalised per venue, so that 100 is the busiest that venue gets; it is not a count of people. The app publishes the prediction $\hat{y}$, rounded, as the served number, and its crowd level, or *band*: *Quiet* (0 to 20), *Not Busy* (21 to 39), *Steady* (40 to 69), *Busy* (70 to 84) and *Packed* (85 to 100).

The weekly curve. Let $b(v, d, h)$ be the venue’s typical value for weekday d and hour h : the venue’s stored weekly value on its local clock, smoothed with its two neighbouring hours at weights 0.6, 0.2 and 0.2. The served number starts from this curve. The popular-times baseline is the stored weekly value itself, before smoothing, as a popular-times chart shows it. The smoothed curve is a whole number but not always a multiple of 5.

Problem. Given a venue v , a target hour h , the curve b , and the live readings of v taken strictly before h , produce $\hat{y}(v, h) \in \{0, \dots, 100\}$ and its crowd level, to minimise error against the live reading at h . The target hour’s own reading is never an input.

4.2 Data collection

Table 1. Data sources. The busyness provider is a commercial venue-busyness service; its scores are relative, 0 to 100 within each venue.

Source	Contribution
Weekly curves	The busyness provider’s typical week for each venue: 168 hourly values on a 21-point lattice of multiples of 5. These rows define the curve.
Live readings	The busyness provider’s reading for one venue at one hour on a real date, collected with a provenance label (live or provider forecast) since 2026-09-01, hourly from 3 September. About 1,380 Pennsylvania venues are scheduled every hour and Miami venues at two fixed hours a day. The 30-day window holds about 7,600 model-served live readings a day (realtime, with a curve above zero; Section 4.3) from 26 September to 2 October and about 4,600 a day from 3 to 8 October, after a drop on the provider’s side.
Weather	Temperature, humidity, wind, condition code and rain flag at the hour, from OpenWeather.
Events	Nearest ticketed event’s distance, type and attendance, and counts within 2 km, from Ticketmaster listings. For the November retrain, 627 professional and college games from TheSportsDB, refreshed daily.
Venue attributes	Category, rating, review count, price level and up to three place types, from a places listing. The venue’s location is used to look up weather, daylight and neighbours and is never a feature.
Calendars	Holiday, school break, holiday eve and special nights across 29 country and city scopes.

Source	Contribution
Owner readings	A venue’s own live 0 to 100 report, at training weight 0.10.
In-app reports	Users’ crowd reports, verified and aggregated per venue over 28 days (the feedback family of Table 3); none existed when v2.6.0 was trained.

Procedure. Table 1 lists every source. An hourly collector reads every tracked venue’s live value and stores one row per venue per hour, stamped with the venue’s local clock and a provenance label. Since 2026-09-25 the full sweep finishes inside the hour (8 min 42 s for a 5 a.m. sweep that called 358 venues and skipped the rest as closed, about 31 minutes for a 1,321-call evening sweep); before that it stopped at its 50-minute budget. Weekly curves and context tables are refreshed on their own schedules. Earlier curve snapshots were not kept, so the 30-day test (Sections 4.5, 6.1, 6.3 and 6.4) scores every configuration with the curves stored at the 8 October export, while its live inputs are limited to readings that existed when each forecast was made; the replay is therefore not a fully point-in-time reconstruction.

Size. When the delta model was trained, the corpus held 3,912,357 rows: 3,454,955 weekly and 457,402 realtime, across 34,784 venues in 34 cities, collected between 2026-03-10 and 2026-05-18. Realtime rows were scarce: 457,402 observations over 18,442 venues is 25 per venue across 168 weekly slots (26.3 on 2026-09-06: 466,179 rows over 17,722 venues after the 2026-09-05 duplicate-venue merge).

Integrity. Unique keys enforce one weekly row per (venue, weekday, hour) and one realtime row per (venue, date, hour). Before they existed, I found and repaired three faults: in 3,536,360 rows the hour column held a list position instead of the clock hour (position 18 held the venue’s midnight), there were 491,100 duplicate rows, and 63% of rows carried an impossible month-season pair. Every forecast accuracy figure in this paper comes from after that repair. A later repair cleared the weather from 3,246,456 weekly rows, which had carried the weather of their collection moment; the shipped model had trained on them before it. The 457,402 realtime rows collected before provenance was recorded cannot be separated into live readings and provider forecasts, and stay marked *unknown*. “Live” in this paper means the provenance-verified readings collected since 2026-09-01 (hourly from 3 September).

4.3 Variables and controls

Table 2 lists the variables and the controls that hold every comparison fixed.

Table 2. Variables and controls.

Type	In this project
Independent variable	The serving configuration or model that produces the number (Sections 4.4, 4.5 and 5.1): weekly curve, curve + offset, curve + offset + nowcast, two persistence baselines, rule engine, delta model v2.6.0, incumbent v2.5.0 (the previous production model).
Dependent variables	Within one level, exact level, band MAE (the mean number of levels between predicted and observed), MAE, R^2 , within 10 and within 15 points, bias.
Controlled: the rows	Every configuration is scored on the identical held-out readings, rebuilt through the production serving code.
Controlled: time	The offset and nowcast weights are fitted on 2026-09-01 to 09-05. The main evaluation covers 2026-09-09 to 10-08; the first test covered 09-06 to 09-08.

Type	In this project
Controlled: geography	Miami, Tokyo and Barcelona never enter the shipped delta model’s training, and each learner-comparison fold leaves out the city it scores. Temporal tests use a venue’s earlier readings, never the scored reading or a later one.
Controlled: population	Scored rows must be <i>model-served</i> (realtime, curve above zero), a test made by one function shared by trainer and evaluator.

Type	In this project
Reference	A constant “Not Busy” (a score of 30, the band’s midpoint), which shows how much a do-nothing answer earns on the level metrics.

4.4 System design

Figure 5 shows the system. Three data sources feed two pipelines. The online pipeline serves every request. The offline pipeline trains a delta model, which can replace the online path only by passing a two-part ship gate.

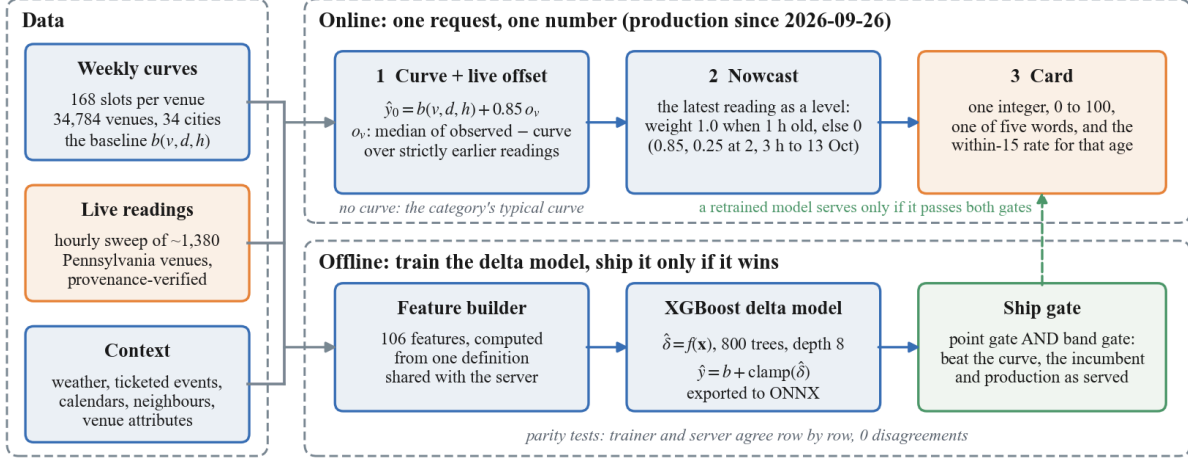


Figure 5. The system. One data column feeds an online serving path (in production since 2026-09-26) and an offline delta model, which serves only if it passes both halves of the ship gate.

Online correction. For each venue, let the offset o_v be the median of (*observed* - *curve*) over the venue’s newest 20 live readings taken strictly before the hour being scored and within 28 days, clamped to ± 50 and used only with at least two readings (otherwise 0). Let r_v be its most recent reading from an earlier hour, and a the age of that reading in hours. The served number is

$$\hat{y}_0(v, h) = b(v, d, h) + 0.85 \cdot o_v \quad (1)$$

$$\hat{y}(v, h) = (1 - w_a) \cdot \hat{y}_0(v, h) + w_a \cdot r_v \quad (2)$$

each step rounded and clipped to $[0, 100]$ as in production, with nowcast weights $w_1 = 1.0$, $w_2 = 0.85$, $w_3 = 0.25$ for readings one, two and three hours old, and $w_a = 0$ for $a \geq 4$. The offset weight and all four nowcast weights were chosen to minimise MAE on 2026-09-01 to 09-05 and held fixed through the 30-day test (served from 26 September, replayed before; Section 4.5). On 13 October the two- and three-hour weights were set to zero (Section 6.3). With $w_1 = 1.0$, a reading one hour old is served unchanged: for fresh readings the fit chose the “repeat the last reading” baseline of Section 3.2, and the offset takes over as readings age. The offset weight of 0.85 is the value this fit chose on those five days; the forward test of Section 3.3, on a different split, preferred 0.5. The card also publishes a confidence, an observed within-15 rate (Sections 5.2 and 6.3). Section 5 lists every step of the serving path and every input it reads.

Offline delta model. The model predicts how an hour departs from the curve, $\hat{\delta} = f(\mathbf{x})$, from 106 input features in eleven families, listed in full in Table 3. The venue’s own curve, the anchor to which the departure is added, is refused as an input. Inputs that would reveal the answer being predicted (label leakage) are excluded, following Kapoor and Narayanan’s taxonomy [10]; a weather fault found after training is recorded in Section 4.2. The learner is XGBoost [4], gradient-boosted decision trees: each tree splits the rows on one

input at a time and outputs a small correction, and 800 trees up to 8 splits deep add up to the prediction (learning rate 0.01, subsample 0.9, column subsample 0.8, minimum child weight 7, L1 0.5, L2 1.0). The score is rebuilt by a fixed *score chain*, clipped to $[0, 100]$ and rounded:

$$\hat{y} = \text{round}(\text{clip}(b + 0.5 \cdot o_v + \text{clamp}(\hat{\delta}, \pm 50), 0, 100)). \quad (3)$$

On the geographic holdout $o_v = 0$, because the offset uses verified live readings and these spring rows predate them. The v2.6.0 ship gate scored the delta with a ± 30 clamp and without rounding (Table 6); Section 6.2 also gives the model under Equation (3) exactly.

Ship gate. A candidate model must pass a point gate and a band gate.

- **Point gate:** on the held-out cities, beat the curve by 5 MAE or 0.10 R^2 , not raise MAE above the incumbent’s, and match the incumbent’s own within-10 on the same rows.
- **Band gate:** on at least 1,000 held-out live readings over 5 or more dates, it must:
 - beat production *as served* on within one level (the date-block 95% lower bound above -1 point);
 - not be worse on within-10 (the date-block 95% lower bound above -1 point);
 - beat the weekly curve on within one level (lower bound above 0) with band MAE no worse;
 - not fall below the rule engine on within one level;
 - hold every city with 300 or more readings within 2 points of production on within one level;
 - raise MAE by at most 1 point.

4.5 Evaluation protocol and statistics

- **Time holdout.** The served number is scored on every model-served live reading from 2026-09-09, the day after its first test, to 2026-10-08, the day of the export: 119,939 readings over 30 dates at

1,754 venues (Philadelphia 63,225, Lehigh Valley 49,503, Miami 7,211), each rebuilt through the production serving code. Two persistence baselines are scored beside it: rule (a), the newest earlier reading if at most 12 hours old, else popular times; and rule (b), the reading from the hour before if there is one, else curve plus offset. No reading in the window informed the offset and nowcast weights (Section 4.4). The first, three-day test, 4,183 readings from 2026-09-06 to 09-08 (Lehigh Valley 3,615, Miami 568), is reported beside it.

- **Geographic holdout.** The delta model is scored on the *gate slice*: the 67,249 realtime rows with a usable curve in the three held-out cities, a whole-city holdout [15] that is also the project’s validation set (Section 6.2).
- **Cross-validation.** Leave-one-city-out over the 30 training cities, the block design recommended for spatially structured data [15]. The corpus’s 34th city, Beijing, holds no rows in v2.6.0’s training frame.
- **Reproduction.** Every geographic-holdout number is recomputed from the saved model and holdout by scripts that pin seeds and thread counts, and those scripts also produce the intervals and the learner comparison (Tables 8 and 12).
- **Statistics.** Live-reading intervals come from date-block bootstraps, which recompute each result on 2,000 random redraws of whole observation dates [15]; geographic-holdout intervals use the two-way cluster bootstrap of Table 8. Tables 4 and 5 are computed from the per-reading rows of the 30-day window, so each of their intervals resamples 30 date blocks. Readings of one venue on one night are correlated, and the date blocks carry that: on the exact-level margin over popular times they make the interval about 4 times wider than if readings were independent, about 16 times the variance. A two-way bootstrap that resamples venues and dates together widens the intervals by a further 10% to 17%; where it changes a conclusion, Table 5 says so. In the results

tables, within one level is reported with exact level and band MAE beside it, because a constant scores nearly as high on it (Section 6.1).

4.6 Reproducibility safeguards

Following common reproducibility practice [14], training pins the device, thread count and seed, and a test confirms that two runs on the same data produce bit-identical predictions. The model is exported to ONNX, a portable format for trained models, and served in-process through ONNX Runtime [13]. Each feature family is computed by the trainer and the server from one definition, and tests require them to agree row by row with zero disagreements (the full list is in Section 19.6). An offline replay used for every live score is proven by a test to publish exactly what the serving function publishes.

5 What Goes into the Model

5.1 Two ways the number is made

Two pieces of arithmetic can produce the number on a crowd card, and they read different inputs. Since 26 September 2026 production has served the *online correction* of Equations (1) and (2): the venue’s own weekly curve, its trailing offset and its latest earlier live reading. In that mode the model stays loaded but is not run, so weather, events, calendars and neighbours do not reach the number. *Model mode*, the code’s default when no setting overrides it, runs the delta model v2.6.0-starling, which takes exactly 106 inputs, listed by family in Table 3. If the model files fail to load, a *rule engine* answers instead (Section 5.2, step 1).

The shipped model was trained on 1,934,988 rows from 30 cities, with weekly rows at weight 0.05 and spring realtime rows (provenance unknown) at 1.0, and with coordinates, venue identifiers, the venue’s own curve and every provenance column excluded.

Table 3. Every input of the delta model v2.6.0-starling (106 features), by family, with its source at serving time, refresh rate and use in the shipped trees. *Splits*: the family’s share of the 167,758 decision splits in the 800 trees; *roots*: trees that split on the family first; both counted from the shipped ONNX file. The last three rows are not inputs of v2.6.0: the venue’s curve is the anchor, and the families marked † are built for the November retrain.

Family	Features	Source at serving	Refresh	Use in the shipped trees
Time (19)	hour, weekday, month; sine and cosine of each; weekend; Friday or Saturday after 18:00; lunch, dinner, late night, morning; four season flags	The venue-local clock of the hour being scored	Every request	12.4% of splits; 349 roots, 274 of them on month. Split gain: month 20.1%, its sine and cosine 4.8%
Daylight (3)	daylight hours, hours after sunset, after-sunset flag	Solar estimate from the venue’s latitude and the month	Every request	11.2% of splits; 22 roots. Daylight hours alone 6.6%
Holidays (6)	holiday, school break, holiday eve, special night, special-night boost and suppress weights	Ten US federal holidays a year, US school-break windows, and 31 national holiday calendars with 1,351 dated special nights over 29 scopes (2025 to 2028)	Regenerated yearly	1.1% of splits; 55 roots, all on the suppress weight
Weather (18)	temperature, humidity, wind, rain flag; nine condition groups; temperature anomaly against a latitude-band by month climate table; four weather interactions	Current conditions keyed to about 1 km; a missing reading becomes the climate norm	Cached 30 min per location	19.5% of splits; 140 roots. Humidity, anomaly, temperature and wind each 3.8% to 4.6%
Events (16)	any event within 2 km; nearest event’s attendance and its log, and distance; its type as five flags (arts, family, music, sports, other); count, total attendance and its log within 2 km; large event; event × weekend, dinner, bar	Ticketed events under way at the scored hour within 2 km; a failed lookup enters as no event	Cached 1 h	1.3% of splits; 5 roots
Category curves (2)	typical value of the venue’s category at this weekday and hour; the same, refined by price and rating tier	Tables fitted on the training corpus, shipped in the model’s metadata	Fixed until retrain	9.5% of splits; 17 roots
Neighbours (2)	log count of venues with a curve within about 830 m; their mean curve value at this weekday and hour, the venue itself excluded	Stored weekly curves of nearby venues	Cached 24 h	13.8% of splits; 33 roots. Second and third most split-on features (7.1% and 6.7%)

Family	Features	Source at serving	Refresh	Use in the shipped trees
Venue (5)	category (13 classes), price level, rating, review count and its log	The venue's places listing; missing values take the corpus median, or 0 reviews	Every request	20.0% of splits; 95 roots. Log review count is the most split-on feature (9.3%)
Place types (30)	one flag for each of the 30 most common place types, from the venue's first three types	Same listing	Every request	9.4% of splits; 84 roots
Feedback (4)	mean crowd report, log report count, has-report flag, mean error of past predictions	Per-venue aggregate of verified in-app reports from the last 28 days	Cached 1 h	Never split on: no verified report existed at training time
Row type (1)	realtime or typical-week row	Always 1 at serving	Constant	1.9% of splits, 0 roots; split gain 6.1%; dropped for the retrain
Own curve (0)	refused as an input	The busyness provider's typical week, smoothed 0.6 / 0.2 / 0.2	Rebuilt after each hourly sweep	The model predicts the departure from this curve, and the served number adds that departure back
Live readings [†] (5)	last earlier reading's departure from the curve and its age; trailing offset and its count; the curve one hour earlier	The venue's own live readings strictly before the scored hour	Hourly	47% of a candidate's split gain in the September rehearsal
Sports games [†] (6)	a tracked team plays today; number of games; evening game; home game; distance to the home arena; arena within 10 km	Public sports schedules, for venues within 60 km of a tracked arena	Daily	0.02% of split gain in the rehearsal

Reading the last column. Split gain, a feature's share of the reduction in training error, is recorded for the month features (month, its sine and cosine) and the row-type flag; for the others Table 3 counts splits. A split count measures how often the trees consult a feature, and it favours continuous features with many cut points over binary flags, so it ranks families roughly. Two facts in it are exact. Ten features are never split on and cannot affect the output: the four season flags, the holiday flag, the four feedback aggregates and the family-event flag. And month carries 274 of the 800 first splits, a reflection of a corpus collected in March, April and May, which the retrain answers by dropping the month family.

Training and serving, side by side. Four differences between how v2.6.0 was trained and how model mode would serve it are known and recorded: the row-type flag is always 1 at serving; in September and October, month takes a value the spring corpus never held, and the climate table covers March to May, so the temperature anomaly is served as 0; the neighbour arithmetic became the training definition on 26 September, after the model shipped; and the shipped model's weekly rows each carried one weather reading from the moment the week was collected. None of the four touches the number production serves today, which reads no model.

5.2 The serving path

For a venue v at hour h , production computes the number in this order.

1. *Fallback.* If the model files fail to load, the rule engine answers from the hour, weekday, rating, review count, price level, place types and weather. With no positive curve value for the venue at that hour, the rule engine answered the same way until 14 October; since then the venue gets the typical curve of its category, refined by price and rating (Section 6.5).
2. *The weekly curve* $b(v, d, h)$, rounded and clipped to $[0, 100]$.
3. *The trailing offset* o_v of Section 4.4, at weight 0.85, each reading measured against the curve at its own hour.
4. *The nowcast.* The newest live reading from an earlier hour, if at most 12 hours old, pulls the number toward itself with weight 1.0 at an age of one hour and 0 after that (before 13 October, the weights of Section 4.4).
5. *The level.* *Quiet* up to 20, *Not Busy* to 39, *Steady* to 69, *Busy* to 84, *Packed* from 85.

6. *The confidence.* The card publishes the within-15 rate measured for the arithmetic that produced the number, on the 30-day window of Section 6.1: since 13 October, 83.0% when an hour-old reading entered and otherwise 42.7%, the overall within-15 rate of curve and offset alone over all 119,939 readings, not yet calibrated on readings without an hour-old reading. From 8 to 13 October it also published 40.7% and 38.6% for two- and three-hour readings, and until 8 October the three-day figures: 41.7% for curve and offset, and 87.4%, 62.8% and 43.5% for readings one, two and three hours old. The 30-day window lowered the last three, the two-hour figure most.

In model mode, step 2 becomes the curve plus the model's departure clamped to ± 50 . A *display map* (straight lines through 40 points pairing equally ranked v2.6.0 scores and live readings) then reshapes the scores without changing their order, and a *one-point push* moves rounded scores outside 25 to 65 one point toward the nearer end of the scale. The offset enters at weight 0.5, and the nowcast uses weights fitted on the model's own output. Live readings come from the hourly sweep of Section 4.2, at seven minutes past each hour; curves and offsets are rebuilt after each sweep.

At start-up the server refuses any model whose metadata does not record a passing ship gate, whose feature list names a feature serving cannot compute, whose ONNX input or output shape disagrees with its metadata, or whose published accuracy record is malformed, and it then serves the rule engine.

5.3 What the model never sees

No input identifies or locates a person. The only coordinates involved are the venue's, and even those are excluded as features: latitude reaches the model only through daylight, the climate band, the special-night rule and the neighbour box. A signed-in user's account number reaches the predictor only as the key of a per-user rate limit and is never a model input, and the training export has no user column. In-app feedback enters as four per-venue aggregates, which the shipped model never splits on. No crowd service reads data from Flux, the sensor of Sections 9 to 18.

6 Results

6.1 H1: the served number on held-out live readings

Table 4. The served number and its comparators on 119,939 live readings over 30 dates (2026-09-09 to 10-08), none of which informed the offset and nowcast weights (fitted on 09-01 to 09-05). Rows (a) and (b) are persistence baselines; (b) was found on this window, so its row is not out of sample. Production and the baselines carry 95% date-block intervals. Bold marks the best value among the configurations scored on the 30-day window. Bias is the mean of predicted minus observed. The last row is production on its first, three-day test, for reference.

Configuration	Within one level ↑	Exact level ↑	Within 10 ↑	Band MAE ↓	MAE ↓	Bias
Production to 13 Oct: curve + offset + nowcast	78.4% (77.9–78.9)	54.9% (53.9–55.7)	55.2% (53.9–56.2)	0.772	18.00 (17.73–18.28)	–1.24
(b) Hour-old reading, else curve + offset (production since 13 Oct)	79.0% (78.3–79.6)	54.7% (53.6–55.5)	54.9% (53.4–56.0)	0.756	17.68 (17.38–18.03)	–1.16
(a) Reading up to 12 h old, else popular times	74.3% (73.4–75.0)	53.2% (52.0–54.1)	55.1% (53.6–56.2)	0.866	20.05 (19.67–20.50)	+0.48
Curve + offset, no nowcast	70.5%	34.4%	31.2%	1.058	24.55	–1.45
Popular times (stored weekly value)	64.8%	30.2%	30.9%	1.155	26.40	+4.89
Rule engine	60.6%	22.1%	19.6%	1.287	29.63	+3.95
Before: v2.6.0 as served (map on)	58.3%	30.8%	26.9%	1.337	30.94	+13.94
<i>Constant “Not Busy” (reference)</i>	75.2%	14.8%	24.5%	1.270	28.37	–11.82
<i>Production on the first test (4,183 readings, 09-06 to 09-08)</i>	78.8%	56.0%	53.8%	0.735	17.06	–0.98

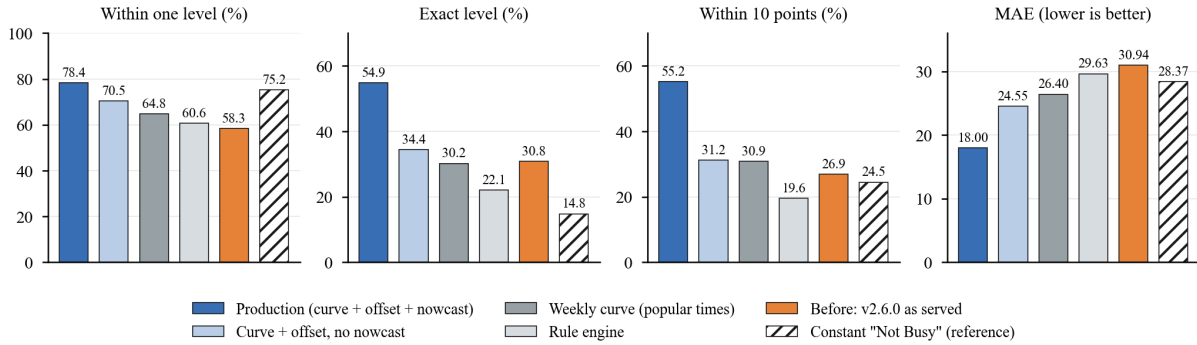


Figure 6. The served number against its comparators on 119,939 live readings over 30 dates (9 September to 8 October 2026).

Table 4 and Figure 6 give the result. Against the configuration it replaced, MAE fell 42% and the bias from +13.9 to –1.2; against popular times, one hour ahead and on the provider’s 0 to 100 readings, MAE is 31.8% lower and the exact level 24.7 points higher. The constant row sets the floor: real venue-hours are bimodal, with peaks at both ends of the scale (36.6% *Quiet* and 17.0% *Packed*), and “Not Busy” sits within one level of three of the five levels, so a constant scores 75.2% within one level while naming the exact level only 14.8% of the time (band MAE 1.270 against 0.772). It tied production on within one level in the three-day test; over the month production clears it by 3.1 points.

Across the month. The lead over popular times holds in every period and on every weekday. Production names the exact level 52.5% of the time from 9 to 25 September (39,000 readings, before the online correction went live), 55.7% from 26 September to 2 October (53,287, live) and 56.7% from 3 to 8 October (27,652, after the provider began returning fewer live readings; Section 7.2), against 29.2% to 32.2% for popular times. By weekday it runs from 54.4% to 56.7%, Friday and Saturday included, against 28.0% to 32.5%. Part of the rise reflects more venues having an hour-old reading as the sweeps sped up (35.3%, 55.5% and 52.6% of readings); holding the share of hour-old readings at the month’s average, the three periods’ exact-level rates are 59.0%, 53.0% and 54.8%. Miami is read at two fixed hours a day, eight hours apart, so its served number is curve plus offset alone, and the nowcast’s gain is a Pennsylvania result (112,728 readings).

Table 5. Production’s margin over each comparator on the same 119,939 readings, in points, with 95% date-block bootstrap intervals over the 30 dates. A negative MAE margin favours production. † The interval includes zero

under the two-way venue × date bootstrap (exact –0.05 to 0.58; within 10 –0.05 to 0.69).

Production minus	Within one level	Exact level	Within 10	MAE
(b) Hour-old reading, else curve + offset	–0.58 (–0.79 to –0.30)	+0.24 (0.03 to 0.51)†	+0.30 (0.02 to 0.65)†	+0.32 (0.15 to 0.44)
(a) Reading up to 12 h old, else popular times	+4.1 (3.6 to 4.7)	+1.7 (1.3 to 2.2)	+0.1 (–0.1 to 0.4)	–2.05 (–2.31 to –1.85)
Popular times (stored weekly value)	+13.6 (12.3 to 14.7)	+24.7 (23.1 to 25.8)	+24.3 (22.3 to 25.7)	–8.40 (–9.02 to –7.70)
Curve + offset, no nowcast	+7.9 (6.9 to 8.7)	+20.5 (18.9 to 21.7)	+24.1 (22.0 to 25.5)	–6.54 (–6.96 to –6.05)
Rule engine	+17.8 (16.7 to 18.6)	+32.8 (31.3 to 33.9)	+35.6 (34.0 to 36.8)	–11.63 (–12.08 to –11.02)
Before: v2.6.0 as served	+20.0 (18.7 to 21.5)	+24.1 (22.6 to 25.4)	+28.3 (26.5 to 29.6)	–12.93 (–13.72 to –12.14)
<i>Constant “Not Busy”</i>	+3.1 (1.1 to 5.1)	+40.1 (39.0 to 41.0)	+30.7 (28.9 to 32.0)	–10.37 (–11.33 to –9.45)

Every interval in Table 5 against popular times, curve plus offset, the rule engine, the replaced configuration and the constant excludes zero, under the two-way bootstrap as well, including the within-one-

level margins the three-day test could not settle. **H1 is supported.**

Where the accuracy comes from. Where the venue was read in the hour before (readings 41 to 77 minutes apart, 10th to 90th percentile), production serves that reading at full weight. Those rows, 48.3% of the window, carry 92% of the exact-level gain over popular times and 96% of the MAE gain; on 24.7% of all readings the value was unchanged from the hour before and the served number reproduces it exactly. The persistence baselines agree. Rule (a), the latest reading or else popular times, reaches 53.2% exact. Rule (b), the hour-old reading or else curve plus offset, is level with production on the exact level and slightly ahead within one level (0.6 points) and on MAE (0.32); the difference is the two- and three-hour nowcast weights (Section 6.3), and production has served rule (b) since 13 October. Under rule (b), where no hour-old reading exists (52% of readings), the number is curve plus offset. Over all 119,939 readings, the smoothing and the offset together lift within one level by 5.7 points over popular times and cut MAE by 1.85.

The gain holds where the value moved. On the 59,681 readings whose value differed from the venue’s previous reading, production names the exact level 41.0% of the time against 30.0% for popular times; on the 28,325 of those with a reading from the hour before, 52.9% against 29.1%.

6.2 H2: the delta model on cities it never trained on

Table 6. Point accuracy on the 67,249-row gate slice, scored as the v2.6.0 ship gate scored it: the predicted delta clamped at ± 30 and not rounded, with $\alpha_v = 0$ because the held-out rows predate verified live readings (display map off).

Predictor	MAE ↓	R ² ↑	Within 10 ↑
Weekly curve alone	31.48	−0.075	19.2%
Incumbent v2.5.0	30.77	−0.043	19.3%
Delta model v2.6.0	29.42	+0.040	20.7%

The model improves on the curve by 2.06 MAE (6.5%) and 0.115 R², and on the incumbent by 1.35 MAE. R² crosses zero: this is the first version to predict these cities better than the mean of the labels, a best-of-several result on the project’s validation set. The gain holds in every held-out city (Figure 7): 2.26 MAE in Miami, 2.21 in Barcelona and 1.72 in Tokyo, where within-10 is 23.6%, 20.9% and 16.9%. Every number in this section reproduces from the saved model and holdout (the incumbent from its tagged commit).

These three cities have been the project’s holdout since March: every version from v2.2 on was gated on them, and the score chain’s clamp and one-point push were fitted on the first 30% of their dates, so they are its validation set. The November retrain adds a time holdout (Section 7.2).

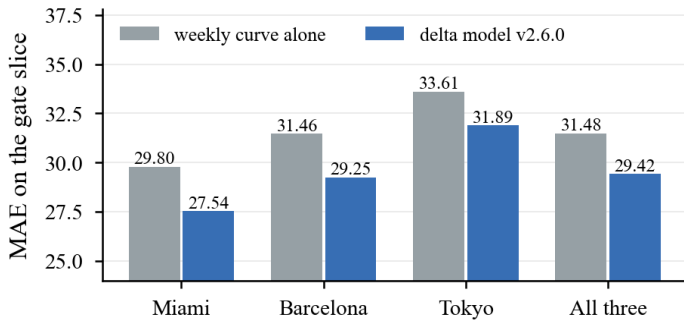


Figure 7. Delta model against the weekly curve on each held-out city of the gate slice.

On the same rows, measured on the five-level scale (Table 7), the model wins on every column: +2.6 points of exact level, +4.3 within one level, and 0.086 fewer levels of error on average.

Table 7. Level accuracy on the gate slice, with the delta clamped at ± 50 and not rounded, and a score of exactly 20 counted as *Not Busy* (display map off). Production counts 20 as *Quiet*; on that ladder the gains are +2.5 exact, +4.4 within one level and 0.089 fewer levels.

Predictor	Exact ↑	Within one ↑	Band MAE ↓	MAE ↓
Weekly curve	22.9%	57.7%	1.315	31.48
Delta model	25.5%	62.0%	1.230	29.35

Scored exactly as Equation (3) writes it, with the ± 50 clamp and rounding, the model reaches MAE 29.35, R² +0.042 and within-10 21.7%. Within-10 is a point higher than in Table 6, which clamped at ± 30 and did not round.

Table 8. The delta model’s margin over the weekly curve on the gate slice, with 95% intervals from a paired two-way cluster bootstrap (2,000 resamples, stratified by city) that resamples venues and city $\times$ month $\times$ weekday blocks. MAE and within 10 are scored as in Table 6, the level metrics as in Table 7.

Held-out rows	MAE	Within 10	Exact level	Within one level
All three cities	−2.06 (−2.75 to −1.44)	+1.51 (0.44 to 2.65)	+2.61 (1.55 to 3.79)	+4.31 (2.93 to 5.80)
Miami	−2.26 (−3.50 to −1.13)	+1.87 (−0.01 to 3.93)	+2.62 (0.65 to 4.78)	+4.83 (2.39 to 7.70)
Barcelona	−2.21 (−3.22 to −1.27)	+2.54 (0.38 to 4.87)	+4.27 (2.19 to 6.55)	+3.68 (1.79 to 5.76)
Tokyo	−1.72 (−2.69 to −0.83)	+0.50 (−0.66 to 1.85)	+1.71 (0.46 to 3.23)	+3.98 (2.01 to 6.17)

The holdout carries no venue or date column, so both clusters had to be recovered. Venues come from their static attributes: the daylight feature is an exact function of latitude and month, which identifies 2,644 venues, none holding more than one row per weekly slot. City $\times$ month $\times$ weekday is the finest date proxy the rows carry. The dependence is temporal: date blocks widen the MAE interval 7.4 times over treating rows as independent, where venues alone widen it 1.4 times. R² gains 0.115 (0.079 to 0.154), band MAE falls 0.086 (0.058 to 0.116), and against the incumbent v2.5.0 MAE falls 1.35 (1.00 to 1.72). Every margin excludes zero overall and in every city except within-10 in Miami and in Tokyo taken alone, so the claim for H2 rests on MAE and the level metrics.

Out of fold, where every training row is scored by a model that never saw its city, the 369,076 served realtime rows give MAE 27.54, R² 0.127, within-10 22.8% and within-15 33.3%. Refit with the same recipe and scored 15 days forward within the 30 training cities (103,254 realtime rows), the model scores MAE 29.02 against 30.23 for the curve (within-10 20.7% against 20.4%), and R² 0.078 against 0.016. The gain is smallest in the dinner (17:00 to 20:00) and evening (21:00 to 23:00) hours, 0.96 and 0.87 MAE.

Placed against prior work, the model’s 6.5% gain uses no recent history, where BysGNN’s larger reductions read the previous 24 hours at every venue [7]. The model’s R² of +0.040 is positive at a granularity where the nearest peer-reviewed occupancy study’s section-level forecasts a day or more ahead have R² near zero and often below it [2]. **H2 is supported on the held-out cities.** The model learned from spring rows of unknown provenance, and on verified live readings the online correction of Section 6.1 gives the stronger number, which is why production serves it (Sections 6.1 and 19.5).

6.3 H3: the value of a reading by its age

The served number's accuracy is set largely by how recently the venue was read (Figure 8). Over the 30-day window, 48.3% of readings (57,900 of 119,939) had a reading from the hour before, and on those the served number lands within 15 points 83.0% of the time, within one level 88.6%, names the exact level 77.0% of the time and has an MAE of 9.99. With a reading two hours old, within-15 falls to 40.7% (14,778 readings), and with one three hours old to 38.6% (7,794), level with curve and offset on the same rows (39.7% and 38.7%); with no reading in the last 12 hours it is 44.4%. On the three-day test the two-hour figure was 62.8%, and the card published it as its confidence until 8 October (Section 5.2, step 6); over the month nearly all of a reading's value sat in its first hour.

The month also refined the weights. On the 14,778 readings whose latest reading was two hours old, curve plus offset alone beat the nowcast at weight 0.85 by 2.73 MAE (25.83 against 28.56, interval 1.41 to 3.63) and 4.5 points within one level (2.4 to 6.1), and the nowcast was no better than popular times on MAE (27.34); its weight had been fitted on 1 to 5 September, when consecutive readings were identical four times in five. At three hours the 0.25 weight made no measurable difference. **H3 is supported:** the decay in accuracy follows the decay in correlation measured in Section 3.2, and is steeper. On 13 October production's two- and three-hour weights were set to zero, so it now serves rule (b), and the retrain gate compares against that stronger rule (Section 7.2).

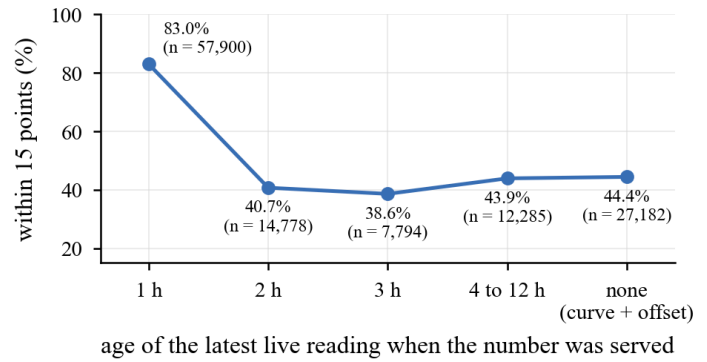


Figure 8. Accuracy by the age of the latest reading, over the 30-day window (119,939 readings). Since 13 October the card's confidence is the one-hour rate when an hour-old reading entered, and otherwise 42.7% (Section 5.2, step 6).

6.4 How far ahead

The headline results are one hour ahead. To measure longer horizons, each of the 119,939 readings was forecast again as if from h hours earlier, using only the readings that existed then (Table 9). Two or more hours ahead the latest reading is too old to use, and production serves the venue's curve corrected by its offset. That number stays ahead of popular times at every horizon to a day: by 3.0 to 3.4 points of exact level, 4.7 to 4.9 points within one level and 1.2 to 1.4 points of MAE, with date-block intervals that exclude zero at every horizon. Within 10 is the exception (0.6 to 0.8 points lower), the grid effect of Section 3.3. The large gains are one-hour gains; the smaller, steady lead is what a group checking the card hours in advance receives.

Table 9. Accuracy by forecast horizon on the same 119,939 readings, with the production rule served since 13 October: the hour-old reading at full weight when the forecast is made one hour ahead, and curve plus offset otherwise, using only readings from hourly slots at least h hours before the scored hour. Persistence is the newest of those readings, at any age; at 24 hours it is often the same hour the day before.

Hours ahead	Served exact ↑	Served within one ↑	Served MAE ↓	Popular times exact	Popular times MAE	Persistence exact
1	54.7%	79.0%	17.68	30.2%	26.40	53.2%
2	33.6%	69.7%	25.02	30.2%	26.40	32.3%
3	33.5%	69.6%	25.09	30.2%	26.40	30.3%
6	33.5%	69.6%	25.13	30.2%	26.40	28.4%
12	33.4%	69.5%	25.15	30.2%	26.40	28.5%
24	33.2%	69.5%	25.18	30.2%	26.40	33.1%

6.5 What coverage is worth

Coverage means that a venue has its own curve. A venue with no curve fell to the rule engine until 14 October. Since then it gets the typical curve of its category, refined by price and rating, which beat the rule engine on every scored slice: on 75,715 Philadelphia readings at 557 venues with each venue's own curve withheld, MAE 26.78 against 29.84 (−3.06, date-block interval −3.42 to −2.73) and the exact

level 28.7% against 21.1% (+7.6, 6.9 to 8.6), with the same direction in the Lehigh Valley and Miami. The comparison below predates that switch.

A delta model cannot score a venue that has no curve, so I compared the model path with the rule engine that served such venues. Both were given identical inputs on 5,025 verified live readings in Pennsylvania at venues that hold their own curve, with the events lookup disabled for both (Table 10, Figure 9).

Table 10. The two serving paths on the same 5,025 live readings (display map on, as served then).

Path	MAE ↓	Within 10 ↑	Within 15 ↑	Median error ↓	On-grid
Model path	28.22	31.7%	41.5%	21	41.5%
Rule engine	30.04	20.3%	29.2%	27	19.8%
Stored curve alone	26.61	30.0%	41.5%	20	100%

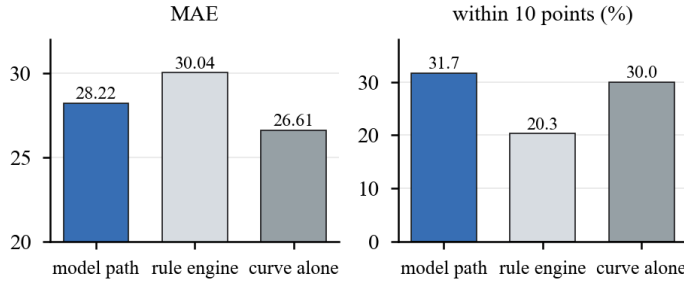


Figure 9. The model path and the rule engine on the same 5,025 live readings, with the curve alone for reference.

Through the model path, coverage is worth 1.82 MAE, 11.4 points of within-10 and six points of median error. Events were disabled in both paths, so their value is left unmeasured. Projected to the grid (Section 3.3), coverage is still worth 11.2 points of within-10, so the result is not a grid effect. The stored curve beats the model path on MAE by 1.61 while losing on within-10: on these bimodal labels the two metrics reward different predictions (Section 7.1).

6.6 Ablation experiments

Each test below changes one part of the system and holds the rest fixed.

Nowcast and offset. Table 4 separates the two corrections:

- *Smoothing and the offset* together take the popular-times row from 64.8% to 70.5% within one level, from 30.2% to 34.4% exact and from 26.40 to 24.55 MAE, and move the bias from +4.89 to -1.45. Within-10 ends where it began (30.9% to 31.2%): smoothing moves predictions off the 5-point grid and lowers it to 26.4%, and the offset lifts it back; this is the grid effect of Section 3.3.
- *The nowcast* then does most of the work: within-10 rises from 31.2% to 55.2%, and the exact level one hour ahead from 34.4% to 54.9%.

Shrinking the curve directly. The direct remedy for the poor calibration of Section 3.1 is to shrink the curve toward the mean. I measured it on the 9,903 paired rows (Table 11, Figure 10).

Table 11. Corrections of the curve’s calibration on 9,903 paired rows. The two per-venue rows are fitted and scored on the same readings, so they are in-sample estimates, not held-out results.

Predictor	MAE ↓	Within 10 ↑
Curve alone	26.86	31.6%
Curve + one global mean offset	25.88	28.9%
Global slope correction	25.49	21.8%
Per-venue offset, half weight (in-sample)	23.18	35.9%
Slope + per-venue offset (in-sample)	22.55	27.2%

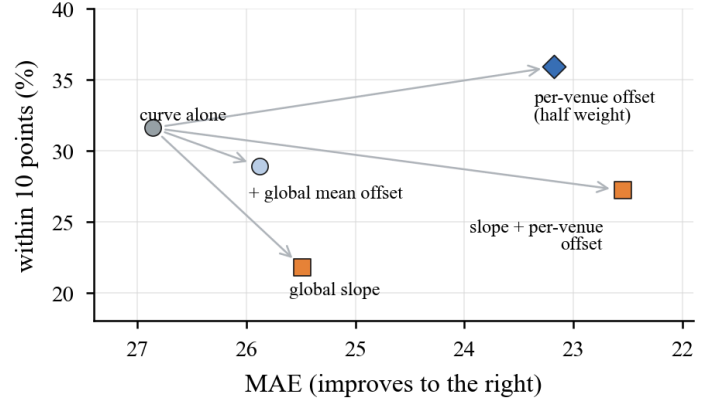


Figure 10. The corrections of Table 11. Corrections that narrow the spread (squares) improve MAE but cost within-10. The per-venue offset, the form of Equation (1), here at half weight, improves both.

Every correction that narrows the published spread improves MAE but costs within-10, so the system uses the per-venue offset. A leave-one-out test over 688 venues (11,804 held-out readings) repeats the pattern out of sample: a per-venue slope improves MAE by a further 0.99 over the per-venue offset but costs 3.19 points of within-10, while the offset itself improves MAE by 1.56 over the curve alone (25.97 to 24.41) at a cost of 0.1 point of within-10. On 8,895 September live readings, the half-weight offset moves MAE from 25.90 to 24.53 and the exact level from 32.1% to 38.0%, on five of five days and in both the Lehigh Valley and Miami.

The display map. From 2026-08-28 to 09-26 the 40-point display map of Section 5.2 sat between rounding and the choice of crowd level. On the 46,101 later held-out-city rows it was not fitted on, it raised within-10 by 8.4 points and MAE by 3.15. September’s live readings reversed that verdict: the map cost 10.5 points of within one level (57.4% with it against 67.9% without, on 7,920 readings), and on 26 September it left the served path, because curve plus offset bypasses it (model mode would still apply it).

Volume against depth. A paired test scored two models on the same 17,322 later rows. One was trained with the scored venues’ earlier history, 304,000 more rows in total; the other had never seen those venues. The model with the extra history scored 27.63 MAE against 27.14, worse by 0.49, in a single paired run. A two-head variant, learning the weekly curve and the departure separately, shows where the limit lies:

- trained on the unknown-provenance pool, it beats the shipped model by 0.88 MAE (95% interval 0.51 to 1.27);
- held to the 4,223 verified labels available at the time, it loses by 1.59 MAE (interval 0.86 to 2.45).

The limit is the number of verified labels per venue, which the live readings of Section 6.1 raise.

The learner. To test the choice of XGBoost, I held the features, the target and the score chain fixed and changed only the learning algorithm (the learner), rotating across the three held-out cities: each learner trained on two and was scored on the third (Table 12). Each fold gives a learner 135,000 to 165,000 rows from two cities, far fewer than the shipped model’s 1.93 million from 30, so the shipped model appears for reference. Neither boosted model was tuned, seeds and thread counts were pinned, and two runs gave identical predictions.

Table 12. Four learners and the weekly curve, with the same 106 features, target and score chain (Equation (3)), each learner trained on two held-out cities and scored on the third, pooled over the 67,249 gate rows. Bold marks the best learner in each column.

Learner	MAE ↓	R ² ↑	Within 10 ↑	Exact ↑	Within one ↑
XGBoost, the settings of Section 4.4	29.97	0.000	21.4%	25.1%	60.7%
LightGBM	29.80	+0.010	21.7%	25.3%	61.1%
Ridge regression	30.18	+0.003	20.3%	24.1%	59.9%
Category-by-hour mean delta	31.61	−0.095	19.7%	23.3%	57.7%
Weekly curve alone	31.48	−0.075	19.2%	22.9%	57.7%
<i>Shipped v2.6.0 (30 other cities)</i>	29.35	+0.042	21.7%	25.4%	62.0%

Against XGBoost, LightGBM is better by 0.18 MAE (0.07 to 0.29) and level with it on within-10; ridge regression is within 0.21 MAE (−0.02 to 0.49) but loses 1.09 points of within-10 (0.54 to 1.72) and 1.01 of the exact level; a category-by-hour average is no better than the curve. Among the learners tested, the two gradient-boosted ensembles did best. The shipped model beats every learner here by at least 0.45 MAE, but with 30 cities against two and far more rows, the comparison cannot separate the effect of cities from that of rows.

6.7 The retrain, rehearsed

This section rehearses the November retrain (Section 7.2), which will add the venue’s recent readings to the model’s inputs. Scored on the first test’s 4,183 readings (6 to 8 September), a model trained with five live features on 3,579 earlier live readings reached 79.7% within one level, +21.2 points over v2.6.0 as served (95% interval 15.8 to 24.2). The features were the last reading’s departure and its age, the trailing offset and its count, and the curve one hour earlier; they carry 47% of its split gain and are worth +10.0 points of within-10 over the same candidate without them. It trailed production by 12.2 points of within-10 (interval −16.3 to −2.7) and by 1.9 MAE. Treating the 4,183 readings as independent samples instead gives intervals of 19.3 to 23.1 and −14.3 to −10.1; the date-block intervals imply variances about 5 and 10 times larger, the first-test clustering estimates that Section 7.2 uses for its power calculation. The band gate refused it on the MAE guard and the three-date sample, and the within-10 criterion of Section 4.4 was added so that a loss like this one is refused on its own.

6.8 The findings on public real counts

Every figure above is measured against the provider’s estimates. To see how counted crowds behave, I repeated the analyses on 15 datasets of real counts under an analysis plan whose pass rules were fixed and hashed (fingerprinted, so any later change would show) before any count was downloaded; two amendments and horizons. The datasets are hourly pedestrian counters in Brisbane, Melbourne (2019 and 2025 to 2026), Dublin, Cork, Dún Laoghaire, Leeds, Auckland, Basel and Zurich; trail counters in Cambridge, Massachusetts and Seattle; hourly entries at 61 New York subway stations; and people

inside a university gym and in six reading areas of the KIT Library in Karlsruhe (seat estimates the library derives from Wi-Fi and calibrates on head-counts), used with the library’s permission. Together: 5,321,507 hourly records (training spans and scoring windows) in 315 dataset-site series (Melbourne’s two periods counted separately; physical sites are not deduplicated) in seven countries, of which 1,787,058 active target hours are scored in the main analysis. Each site was rescaled to 0 to 100 by its training-period 99th percentile (clipped at 100), curves were built from the eight weeks before each scored day, the main analysis scores active hours (eight-week curve at least 10), offsets and carried departures are clamped to ±50, and intervals are week-block bootstraps (Table 13, Figure 11).

Table 13. The provider’s live readings beside public real counts, one hour ahead. Real counts are pooled over 15 datasets with equal weight: 13 count people passing a sensor and two measure people inside (a gym’s counts and a library’s Wi-Fi seat estimates), shown on their own as well. Each source’s MAE is on its own 0 to 100 scale, so predictors compare within a column. The provider’s newest reading is the latest at any age; on counts every hour has one. Curve + latest departure was scored on the counts only.

Measure	Provider’s live readings	Real counts, 15 datasets	People inside: gym and library
Weekly curve as a forecast: slope, r	0.50, 0.36	0.92, 0.79	0.76, 0.62
Quiet hours, busy hours: observed minus curve	+11.5, −13.9	+0.3, −1.4	−0.5, −4.0
Departure correlation at 1, 2, 3 hours	0.74, 0.26, 0.16	0.74, 0.61, 0.54	0.83, 0.78, 0.73
MAE: weekly curve	26.40 (popular times)	9.58	14.36
MAE: newest reading as the forecast	21.94	10.36	8.23
MAE: curve + trailing offset	24.55	7.75	7.85
MAE: curve + latest departure	—	6.41	6.29

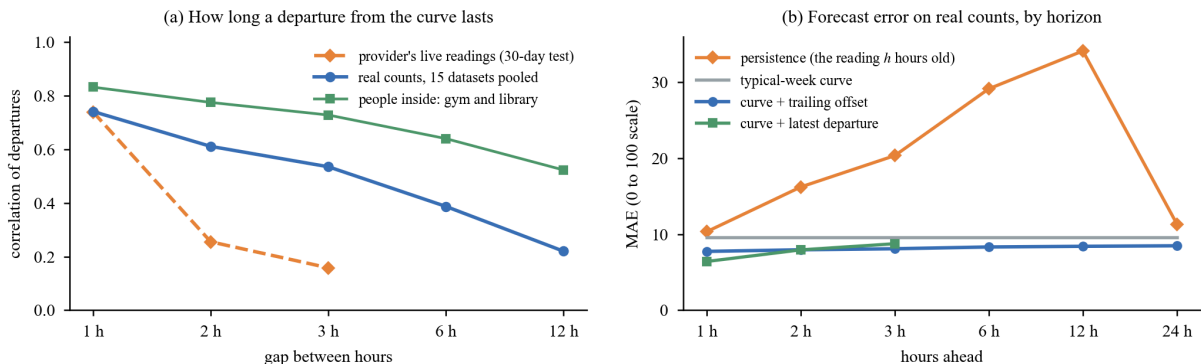


Figure 11. The provider’s live readings beside 15 real-count datasets. (a) How long a venue’s departure from its typical-week curve persists. (b) Forecast error on the real counts by how far ahead the forecast is made (pooled, active hours).

Part of the result carries over. A site's latest departure from its curve is its most valuable one-hour input, and the curve alone is never the best forecast: adding the latest departure to the next hour's curve gives the lowest one-hour error in 14 of the 15 datasets, 33% below the curve alone (MAE 6.41 against 9.58; registered without a pass rule), and the trailing offset improves on the curve at every horizon to 24 hours.

The provider-specific findings do not carry over as stated. The way they fail points to the provider's readings and curves, though the comparison also differs in what is counted, how curves are built and the labels' lattice. Real counts follow their curve closely (slope 0.92, r 0.79), so the raw hour-old count loses to the curve in 9 of the 15 datasets; under the registered pass rules, which require 95% intervals to clear thresholds fixed in advance, the first finding (the hour-old reading beating the curve on both error and exact level) replicated in 4 datasets, and the calibration and decay findings in none. Real departures last longer (0.61 two hours on, 0.54 three hours on, against the provider's 0.26 and 0.16): the provider's value swings around its curve more widely and settles back faster, while repeating itself hour to hour. On its readings the hour-old value used as a level, else curve plus offset (rule (b)), has the lowest one-hour MAE among those scored; on real counts it is the latest departure carried onto the curve.

The two datasets that count the people inside a place, rather than people passing a counter, side with the provider on the first finding. At the gym and the library the hour-old count beats the curve by a wide margin (exact level 66.0% against 48.1%, MAE 8.23 against 14.36), the curve fits less well (slope 0.76, r 0.62), departures persist longest (0.73 at three hours), and carrying the latest departure onto the curve is better still (MAE 6.29). For a real-count source such as Flux, the evidence favours that predictor.

Every other count I could use. Nine more checks ran after the registered analysis, outside its pass rules. (1) Six KIT libraries that close at night, which miss the registered rule only because Seatfinder logs nothing while closed, scored at their open hours (12,248 hours): the hour-old count beats the curve (MAE 8.13 against 15.32, exact 68.3% against 47.0%), departures persist (0.93 at one hour, 0.80 at three), and the latest departure on the curve is best again (MAE 5.32, exact 78.6%). (2) Five rooms at the National University of Singapore (ROBOD): on occupant counts (1,046 hours, up to 38 people) the hour-old count scores MAE 12.18 (10.66 to 13.78) against 21.63 (18.96 to 24.53) for a two-week curve, and on connected Wi-Fi devices (2,056 hours) 5.53 (4.92 to 6.15) against 11.77. (3) The 153 usable indoor Wi-Fi access points in the 24 buildings of the Universitat Politècnica de Catalunya's Campus Nord, Barcelona (71,926 hours, April to June 2023): the hour-old count beats the curve at 152 of 153, MAE 8.87 (8.66 to 9.08) against 16.60 (16.31 to 16.94). (4) To test whether the provider's readings follow real people at all, I set its typical week for 695 New York venues beside hourly entries at the nearest subway station (34 stations with at least five venues within 400 m, March to May 2026). The rhythms match closely (median r 0.74; 0.81 citywide), and at 28 of 34 stations the venues match their own station better than other stations (mean gain in r 0.033, interval 0.017 to 0.050), so the curve carries each block's own rhythm. Station entries are not the crowd inside a venue, so this checks rhythm rather than level. (5) Two Zurich series that the city's open-data team suggested (CC0), run through the registered engine and site rule: the footbridge counter at Cassiopeiasteg, the only one of five passing the site rule (4,526 scored hours), and the eight counting lines of the Hardbrücke tram stop (50,588 hours). On both, the latest departure on the curve is best (MAE 7.35 against the curve's 11.77, and 5.38 against 7.05), the curve is close to calibrated (slope 0.86 and 0.97) and departures last (0.54 and 0.46 at three hours). The hour-old count beats the curve on the footbridge (9.74) and loses at the tram stop (12.36), whose

commuter peaks the curve already predicts. (6) Seven more open series through the registered engine, with the KIT rule for venues that close. Where occupancy was measured, the latest departure on the curve was best on all four: 17 Aarhus libraries rebuilt from door counts (one-hour MAE 9.25 against the curve's 15.26), a Korean gym and pool from each visit's entry and exit (8.56 against 11.41), a university library's late-night head counts (5.08 against 12.91) and Wi-Fi devices in 12 Bologna libraries and museums (7.89 against 12.48). On entries it was best at six Orléans venues (14.02 against 16.58), level at a Japanese library (6.70 against 6.86) and worse at 12 Montpellier pools (17.32 against 15.65), whose departures barely persist (0.27 after an hour); there the curve with its offset was best. (7) The University of St. Gallen Library's two reading rooms, on the same service and used with its permission, at open hours (4,308 hours): as at KIT, the hour-old count beats the curve (MAE 7.79 against 22.29) and the latest departure is best (5.05, exact 79.5%). (8) York's two city-centre footfall cameras (City of York Council), scored to June 2025, before the camera gaps the Council flags (10,561 hours): the latest departure is best again (MAE 6.38 against the curve's 9.64), and the hour-old count alone does not beat the curve (10.23). (9) Toronto's island ferry (City of Toronto), as tickets scanned at the city terminal (2,060 hours): the latest departure is best (MAE 7.23 against the curve's 20.50) and the hour-old count beats the curve (7.94).

7 Discussion

7.1 Interpretation

The latest reading is the lever. Where an hour-old reading exists, using it at full weight is the strongest rule this project found (Section 6.1); where none exists, about half of venue-hours, production serves the corrected curve. On the provider's readings the latest value is itself the best one-hour forecast; on real counts, adding the latest departure to the next hour's curve is best (Section 6.8).

The delta model's gain holds in every held-out city. Without any observation of the venue's recent state, the spring data supports a 6.5% MAE reduction and an R^2 above zero (better than predicting the mean of the labels) at single-venue, single-hour granularity. On the gate slice the model's output correlates with the raw curve at 0.9638: the curve supplies the rhythm and most of the level, and the model a correction.

Readings per venue are the binding constraint. The spring corpus holds about 25 realtime rows per venue across 168 weekly slots. On 2026-09-06, 9,339 rows across 1,228 venues carried verified live provenance, a median of six each. In the one paired run of Section 6.6, 304,000 more rows of the scored venues' earlier history did not help. A reading's value lasts about an hour, so how often a venue is read matters more than how long its history is. Context does not stand in for it: in an offline test on 9 to 28 September, hourly rain, activity at Philadelphia's public bike-share docks near the venue and a local event calendar explained none of what rule (b) misses.

One number cannot sit near both ends of the scale. With 36.6% of readings *Quiet* and 17.0% *Packed*, MAE and within-10 can give opposite instructions, as the curve corrections of Section 6.6 show; a live reading improved both at once (MAE 24.55 to 18.00, within-10 31.2% to 55.2%).

Who it helps. For a group choosing where to go, the card is now right more often than not one hour ahead, measured against the provider's relative readings. Flock receives only venue-level scores from the busyness provider, built from aggregated, anonymised signals; no individual's data reaches Flock.

7.2 The November retrain

This paper is a snapshot taken before the retrain. The retrain will learn the live features rehearsed in Section 6.7 rather than blend a reading in after the fact, give live readings 80% of the total training weight, drop the month family and the row-type flag, and add the sports-schedule features. It is gated against production as served at gate time (rule (b) of Section 6.1 since 13 October) and ships only if it wins. The run moved from about 19 to 20 October to early November, after Halloween, in the hope of a held-out window that falls entirely after 8 October, the end of the month reported here. I will update this paper with the result: a model trained on a month of verified live readings, scored on a 14-day window, under the split rule and thresholds committed in September and the comparator chosen on 13 October.

The sample. From about 23:00 on 3 October the provider answered the same calls with fewer live readings (about 4,600 a day to 8 October, 61% of the week before). The drop did not lower accuracy on the readings that were returned (Section 6.1); readings never returned cannot be scored. From about 16:00 on 12 October the provider returned none for about 52 hours, then a trickle from one venue from 20:52 on 14 October. The registered rule anchors the window on the last date with live readings and needs at least 1,000 readings over 5 dates; it will be applied to the readings that exist on the day of the run. If it lands on 29 September to 12 October (window A), ten of those dates fall inside the month reported here, so production’s accuracy on them is already known. The retrain trains only on earlier dates, and no training setting was chosen on these ten days. The 30-day test of Section 6.1 is the last measurement of that method while production served it.

What was fixed in advance. Hypotheses H1 to H3 were stated after the analyses that test them (Appendix A). The retrain is the first test here whose hypotheses, split rule and decision thresholds were fixed in code before its data existed: the band gate’s thresholds and the 14-day time holdout were committed on 25 and 26 September 2026. The comparator was not: on 13 October, after the month’s analysis on dates that overlap a possible held-out window, production was simplified to rule (b), a stronger comparator. A window that overlaps those dates is held out from the retrain’s fitting, but the comparison is untouched by the 13 October choice only if every held-out date falls after 8 October.

- **H4.** The retrain scores higher than production as served on within one level over the held-out 14 days, and its date-block 95% lower bound is above -1.0 point.
- **H5.** It is not worse on within 10, by the same bound.
- **H6.** It beats the weekly curve on within one level with a lower bound above zero, and its band MAE is no worse than the curve’s.
- **Guards.** At least 1,000 readings over at least 5 dates; the window is held out from the retrain’s fitting (the registered wording, “production never saw the window”, does not hold for window A, ten of whose dates were used on 13 October to choose rule (b)); no city with 300 or more readings falls more than 2 points behind production on within one level; within one level is not below the rule engine’s; MAE rises by at most 1 point.
- **Analysis.** 2,000 date-block bootstrap resamples. The constant-answer reference and the configuration tested in this paper are recorded beside every verdict, as references only; neither changes the decision.
- **Decision.** The retrain ships only if all of these hold. Otherwise production keeps serving, and the result is reported either way.

Table 14. The smallest difference the held-out window can detect at 80% power (two-sided, 5%), in points, by the number of held-out readings. Each cell gives the value with readings treated as independent, then with the variance inflated 5 and 10 times, the clustering the first-test rehearsal of Section 6.7 measured, used here as an illustration; the 30-day window measured about 16 times on the exact-level margin over popular times (Section 4.5). The assumed rates are the one-hour rates of the configuration served until 13 October, over the 30-day window. The last column gives the readings needed to detect a 1-point difference with independent readings.

Metric (assumed rate)	15,000 held out	29,000 held out	60,000 held out	Readings for 1 point
Within one level (78.4%)	1.33 / 2.98 / 4.21	0.96 / 2.14 / 3.03	0.67 / 1.49 / 2.10	26,600
Exact level (54.9%)	1.61 / 3.60 / 5.09	1.16 / 2.59 / 3.66	0.80 / 1.80 / 2.55	38,900
Within 10 (55.2%)	1.61 / 3.60 / 5.09	1.16 / 2.59 / 3.66	0.80 / 1.80 / 2.54	38,800

With about 60,000 held-out readings, a 95% interval on a within-10 difference is about ± 0.6 points for independent readings and ± 1.3 to ± 1.8 points with 5 to 10 times inflation. With 5 to 10 times inflation, a retrain that exactly ties production would fail that bound and would need a within-10 win of about 0.3 to 0.8 points. The registered gate scores paired date-block differences, whose interval may be narrower, so the actual verdict can differ. The bound is strict on purpose: the rehearsal in Section 6.7 is the case it exists to stop.

8 Conclusions on the forecast

The first question asked whether AI can tell how crowded a venue will be before you go. One hour ahead, yes, and what does it is small: from 26 September to 13 October Flock served each venue’s weekly curve corrected by its recent offset and a nowcast of its latest reading, with weights fitted on earlier days. Replayed on 119,939 live readings from 9 September to 8 October that its weights never saw, it named the exact crowd level one hour ahead 54.9% of the time, within one level 78.4% and within 10 points 55.2%, against 30.8%, 58.3% and 26.9% for the gradient-boosted model it replaced, with an MAE 31.8% below popular times, all on the provider’s relative 0 to 100 estimates, not head counts. The three-day result it shipped on held on 29 times as many readings, on every weekday and in every period. The gain comes from fresh readings rather than a larger model: with an hour-old reading 77.0% of served crowd levels are exact, and production now serves that reading at full weight. Two to 24 hours ahead, when no fresh reading exists, the corrected curve still names the exact level about 3 points more often than popular times. On 15 public datasets of counts and occupancy estimates, crowds follow their curves far more closely than the provider’s readings do, and adding the latest departure to the next hour’s curve gave the lowest one-hour error in 14 of them. The gradient-boosted delta model, which production does not serve today, still beats the curve on three held-out cities without any recent-history input (Section 6.2).

Three lessons should carry over to other systems built on typical-week curves:

- score every forecast against the latest observation as well as the curve [8]: the curve was poorly calibrated on the provider’s readings (slope 0.50) and close to calibrated on real counts (0.92);
- where labels come in steps (here multiples of 5), read within- k rates beside the share of predictions landing on a step (Section 3.3);
- prefer fresher readings of each venue to a longer history of it: on the provider’s readings a reading’s value was concentrated in its first hour, though on real counts departures lasted longer.

9 Counting people on a thermal camera

Flock’s forecast predicts a relative score from 0 to 100 that counts no one. Owl answers the second question: can a credit-card-sized computer model be designed to count the number of people inside a venue using a thermal camera without identifying individuals? It is a neural network that runs on Flux, a small venue sensor built around a Raspberry Pi 5 and a FLIR Lepton 3.5 thermal camera, which records heat instead of light. No Flux is installed at a venue yet, and Flock’s crowd number does not read sensor data; the paper’s last sections set out how Flux’s counts could check the forecast. At 160×120 thermal pixels Owl is built to count people, never to identify them, and Flux reduces each frame to a count on the device and throws the frame away (Figure 12).

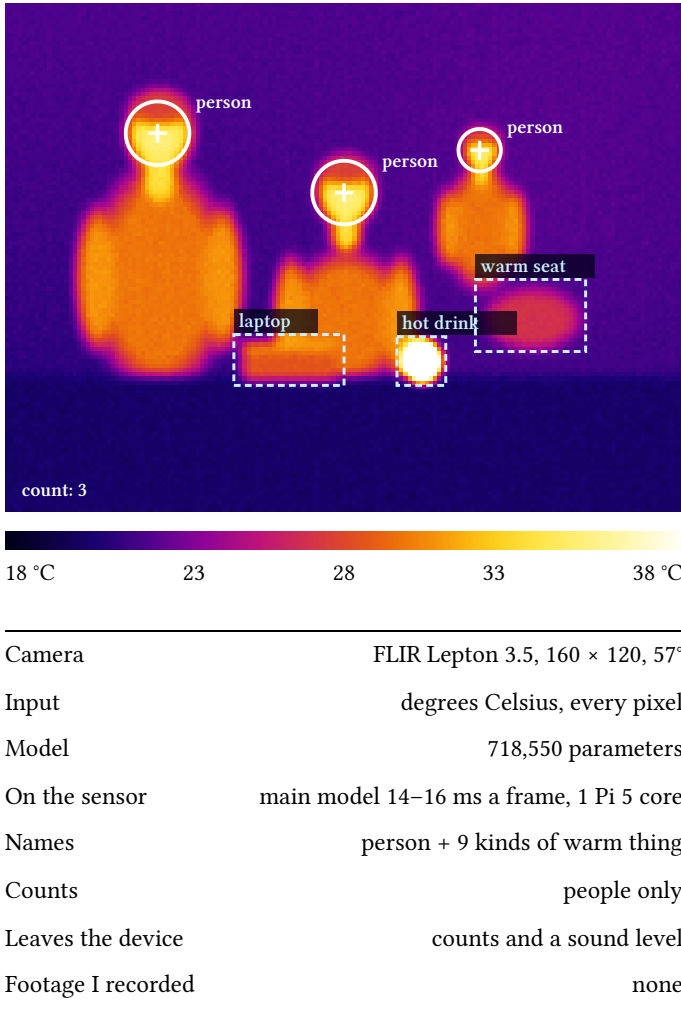


Figure 12. What Owl reads and what it reports. An illustrative 160×120 frame, drawn for this paper the way Owl’s training generator draws a scene, shown at the camera’s true resolution in the iron palette the Lepton uses. Owl puts a point on each person’s head and a named box on everything else warm. Only the points are counted: this frame counts as three.

The first counter on Flux, a heat-threshold rule, counted regions warmer than the room and large enough to be a person. It failed predictably: a hand at the lens counted as a person, spread fingers as two, a coffee mug or a laptop as somebody, and a packed room merged into one warm mass. Telling these apart is a question of shape, which a learned model does well.

Sections 9 to 18 describe Owl, the model that replaced the rule, and what building it showed. Four constraints shaped the work from the first day; the first two are the question’s own terms.

1. **It runs on the sensor.** One core of a Raspberry Pi 5, with no GPU and no round trip to a server.
2. **It names kinds of things and nothing more.** No identity, no face features, no tracking from one frame to the next. Only numbers leave the device: people counts and a sound level.
3. **I do not record people to train it.** Owl learns from frames I generate and from existing datasets.
4. **Every real frame is licensed.** Each source allows commercial use under its licence, or its authors gave written permission.

The contributions are a generator of thermal scenes matched to the Lepton 3.5 (Section 12.2); a 718,550-parameter network that marks each person with a point and names nine other kinds of warm object, checked for physical plausibility on the sensor and backed in busy frames by a second network, the crowd model (Sections 12.4 to 12.6); an evaluation against the rule on identical frames, on six real benchmark sets and on sealed tests kept out of every design choice (Section 13); and 15 findings from 19 full training runs plus five short trials (Section 15).

10 Background on thermal people counting

Thermal people counting is usually studied with low-resolution array sensors (thermopiles) on the ceiling, such as the 32×24 MLX90640 in PUT Thermo Presence [25], whose counting models have been benchmarked on resource-constrained hardware [34] and against deep detectors on low-resolution infrared sets [39], and with higher-resolution cameras for segmentation and pose [24, 26]. Owl borrows from three lines of work.

Objects as points. CenterNet [17] predicts a heat map, a grid of confidence scores (not temperatures) with a peak at each object’s centre, and finds objects at the peaks. Counting then reduces to counting peaks, which is exactly the quantity a venue needs. Its loss, the penalty training minimises, is CornerNet’s [32] penalty-reduced variant of focal loss [18], which handles the imbalance between a handful of peak pixels and thousands of background pixels.

Counting by density. Density-based counting [19, 20, 33] predicts a map whose values add up to the count. It degrades gracefully when objects overlap, where peak finding merges two touching heads into one.

Learning from generated data. A model trained on generated frames transfers to real ones only as far as the generator covers reality. Domain randomization [21] varies everything the model should ignore so that the real world looks like one more variation. Section 15 shows where that held for Owl and where it did not.

11 The sensor and the baseline

Flux reads the Lepton 3.5 through a PureThermal 3 board over USB. The camera is radiometric: each of its 160×120 pixels reports a temperature in degrees Celsius, across a 57° horizontal field of view [29]. Through that lens a head three metres away is about 8 pixels wide and a hand 30 cm from the lens about 45 pixels wide.

The rule divides the frame into cells of 4×4 pixels and estimates the room as the frame’s 20th percentile. It learns fixed warm spots such as radiators per cell over time, keeps cells warm enough to be skin, joins touching cells (diagonals included) into groups and counts groups of at least 6 cells, about one head at three metres. Its settings were tuned on a bench with a real person and against constructed scenes, and it works for well-separated people at moderate range. It has no notion of shape, which is the gap Owl fills. Throughout this paper the rule is scored on exactly the same frames as Owl.

12 Owl: method

Figure 13 shows the pipeline, from the training data to the counter on the sensor.

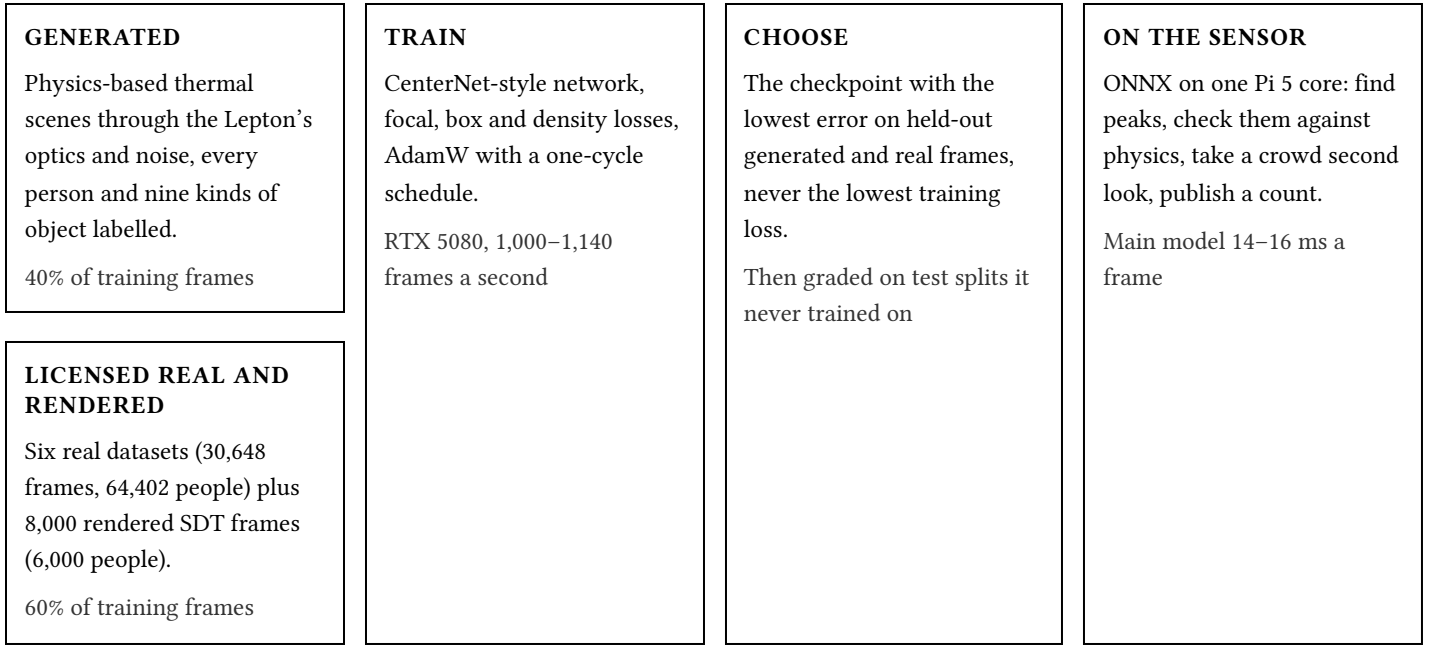


Figure 13. The pipeline. Data on the left, the deployed counter on the right. Checkpoints are selected on validation splits, and versions and settings on the real benchmark sets (Section 13). Accuracy is measured on those sets, fifteen rounds of sealed tests and generated stress scenes, and speed on the sensor itself.

12.1 Formulation

Each frame is labelled with one point per visible person (its head point), at the centre of the head, or at the centre of the visible body when the head is out of view. Every other warm object gets a box and one of nine kinds: hand, pet, hot drink, food, laptop, screen, heater, lamp and warm seat. Counting people is finding peaks in the person heat map. The other nine maps let the sensor’s display show why something was not counted, and they give the network somewhere to put the things it must not count.

12.2 A generator matched to the camera

Owl’s first version learned only from generated frames, and the generator stayed in every training batch after that (40% of frames for the shipped model). It draws a room at some temperature, people at every distance and pose, and the objects that fool a threshold, at twice the camera’s resolution, then averages down, blurs for the lens, and adds the camera’s noise. Table 15 lists the physical model, which replaced hand-picked temperatures after the first round of real-footage tests (Finding 4).

Table 15. The physical model behind the generated frames. Values were set from published thermography of skin and clothing.

Surface	Model
Skin	33.4 °C forehead in a 22 °C room, rising 0.14 °C for each degree the room is warmer
Clothing	A garment-specific share of the way from the room to the skin; a winter coat reads barely above the room
Hair	3–10 °C above the room

Table 16. Real training data. People are counted in the training split; for IISER they are labelled faces. Licences and permissions are listed in the acknowledgements.

Source	Camera and view	Stored as	Train frames	People	Use
--------	-----------------	-----------	--------------	--------	-----

Surface	Model
Pets	Coat near room temperature, eyes and ears 35–38 °C
A seat just left	4–8 °C above the room
Rooms	Typical rooms, plus 20% of frames drawn hot at 29–34 °C
Camera	Lens blur, noise, fixed-pattern column stripes, dead pixels, vignetting, several degrees of absolute error, gain error, a crooked mount, the frame after a flat-field correction

Scenes include standing, sitting and lying people, children, hats, hoods, glasses, masks, bald heads, bare torsos, winter coats, people hidden behind tables and bars, one person partly behind another, and crowds. They also include things that must not be counted: hands at the lens, reflections of people in glass and shiny floors, warm footprints, sun on the floor, candles, cold drinks and bulbs hotter than the camera’s range.

12.3 Licensed real data

Six real thermal datasets and one set of rendered frames entered training over the project (Table 16). Two are from the Lepton 3.5 itself. Sources that store temperatures (PUT, SDT) are used in degrees Celsius. Sources stored as 8-bit images carry no temperatures, so each frame is mapped onto a plausible room-to-skin range redrawn at each use (low end 16–28 °C, span 8–16 °C, brightness exponent 0.8–1.25), which stops the network from treating any single mapping as truth. IISER’s frames are colour images in the camera’s iron palette; I convert them back to one channel by luminance, which rises monotonically along that palette from black through purple, red, orange and yellow to white.

Source	Camera and view	Stored as	Train frames	People	Use
PUT Thermo Presence [25]	MLX90640 32 × 24, ceiling	°C	8,472	24,106	MIT licence
OpenThermalPose2 [26]	Side view, indoor and outdoor	8-bit	8,846	15,849	MIT licence
AAU VAP Trimodal [24]	Wall-mounted, meeting rooms	8-bit	3,964	5,466	CC BY 4.0
SenSys 2021 set [27]	32 × 24 rooms with hot laptops	8-bit	1,239	1,577	CC BY 4.0
SDT, TU Wien [22]	Lepton 3.5 , elevated indoor, one person or none	°C	4,800	3,600	Authors' permission
SDT rendered frames [22]	Rendered to the Lepton's look	°C	8,000	6,000	Authors' permission
Thermal Vision, IISER Bhopal [23]	Lepton 3.5 , classroom of 12 to 15 and lab of up to 6 (up to 10 and 5 labelled faces)	Iron palette	3,327	13,804	Authors' licence and permission

Real datasets label people and nothing else, so for real frames the loss on the other nine maps is switched off. AAU, SenSys and SDT mark bodies rather than heads; for them the head point is estimated from the box and its target is spread over the pixels the head could occupy ($\sigma = 5$ px), a fix that Finding 7 explains. Test splits are the datasets' own, or, where a dataset has none, contiguous ranges in time split 60/15/25 for training, validation and test: SDT by frame range, and IISER by time within each room with ten frames dropped at every boundary, because neighbouring frames are near copies.

12.4 Network

The network sees two channels. The first is absolute, $(t - 30) / 8$, so a 60 °C mug can never look like skin. The second is relative to the frame's median, $(t - \text{median}) / 4$, so the several degrees an uncalibrated Lepton reads high or low do not move a person out of range. An encoder halves the resolution four times, to strides 2, 4, 8 and 16, building 24, 48, 72 and 96 channels (maps of learned patterns), using residual blocks [30] with dilation at the deepest stage [31, 33] to widen the view: a hand filling half the frame must be seen as one hand and not five warm blobs. A top-down path fuses the deep features back to stride 4, a 40 × 30 grid. Three output branches, or heads, read that grid: ten heat maps (bias initialised to -4.6, so training starts from "nothing here"), a box head that predicts, on a log scale, the distance from each object's centre to each edge of its box, and for the crowd model a density head, constrained to positive values and initialised near empty. The main model has 718,550 parameters and exports to ONNX as a 2.9 MB file.

12.5 Training

The loss is CenterNet's penalty-reduced focal loss on the heat maps [17, 18], with $\alpha = 2$ and $\beta = 4$ as in CornerNet [32], masked by which maps a frame actually labels; an L1 loss on the log of those edge distances, at object peaks; and, for the crowd model, squared error per density cell plus a tenth of the error in the total. I train with AdamW (weight decay 10^{-4}), a one-cycle learning-rate schedule with 5% warm-up, gradient clipping at 5, and batches of 64. Fresh runs use a peak rate of 2×10^{-3} ; later versions start from the previous version, keeping every layer whose shape matches, at 2×10^{-4} . Eight CPU workers generate frames on the fly for an RTX 5080. The project ran 19 full training runs plus five short trials, 581,700 steps in all, which is 37.2 million training frames.

12.6 Decoding on the sensor

The sensor runs the ONNX file with ONNX Runtime [13] in full precision on one core. A person is a cell scoring above 0.4, on a confidence scale from 0 to 1, and at least as high as all eight neighbours. Rarer kinds must clear a higher bar before they are named, from 1.25 times the person threshold for hands and hot drinks to 1.75 times for pets, because they are learned from fewer examples. A person peak is dropped when a screen, laptop, heater, lamp or

warm seat is more confident in the same 3 × 3 neighbourhood. Two physical checks run on a 9 × 9 patch around each person: its warmest pixel must be at least 1 °C above the room, taken as the frame's 10th percentile (Finding 3), and its median must not exceed 45 °C, which no person reaches. When five or more people are found, the crowd model reads the same frame, and the published count is the highest of the main model's peaks, the crowd model's peaks and the rounded density total.

13 Owl: evaluation protocol

Checkpoints, benchmarks and sealed tests. Each run keeps the checkpoint (a saved copy of the network) with the lowest combined error on held-out generated and real validation frames, never the lowest training loss. Every version was then graded on the six test splits, which served as the development benchmarks for choosing versions, data shares, decoding settings and the crowd hand-over point. No test frame was ever trained on: the training cache holds training and validation frames only, with no exact duplicates of a test frame. The sealed tests, held out from all of these choices and each scored once apart from one recount after a frame-numbering error, are reported with the results.

Six real benchmark sets. Six test splits that no model trained on: the Lepton 3.5, the camera Flux uses, in indoor rooms (SDT, 2,000 frames) and in a crowded classroom and lab (IISER, 1,375 frames, 4,882 labelled people), a ceiling camera (2,347 frames, 8,043 people), rooms with hot laptops (265 frames), wall-mounted meeting rooms (813 frames, 1,061 people) and a side view (2,151 frames, 3,872 people). A frame is *exact* when the count equals the truth and *within one* when it is off by at most one person. The rule is scored on the same frames, as deployed and, separately, with its settings tuned on each set's validation split. Intervals are 95% block bootstraps, which recompute the score on 2,000 random redraws of whole runs of consecutive frames within a scene or sequence, because neighbouring frames are near copies: 50 SDT frames within a pose block, 25 IISER frames within a room, 100 PUT frames within a sequence, 20 SenSys frame numbers within a recording regime, 25 AAU annotated frames within a scene, and OpenThermalPose2 files grouped by subject and session prefix, with bare-numeric file names in runs of 50 (40, 56, 25, 14, 34 and 49 blocks). They describe sampling variation within these datasets, not variation across unseen venues.

Two properties of the sets matter. IISER's labels mark visible faces, so a person facing away is absent from its truth (a recount is in Section 14.3). And some splits share scenes: AAU's three meeting rooms and their people appear in training and test, IISER's test frames share their two rooms with training, and one PUT test sequence continues a training sequence. Removing that sequence and every test frame nearly identical to a training frame raises Owl's PUT score from 92.9% to 96.0% exact and moves no other set by more than 0.1 points, so near-copies do not inflate the results; AAU was not re-split, so its figure comes from rooms and people Owl trained on.

Stress test. 41 hard situations, each generated on purpose 150 times and scored alone, counted exactly as the sensor counts, crowd model included. Each case’s random seed is fixed, so a rerun draws the same frames. The cases come from the same generator as 40% of the training data and guided which cases training emphasised, so the stress test checks coverage of known hard cases; the real sets carry the headline numbers. Its intervals are Wilson intervals for a proportion, which treat the frames as independent.

Blind-spot report. Every real benchmark set split by situation (by pose for SDT and by number of people elsewhere), and every stress case. Anything under 90% exact or 97% within one is flagged.

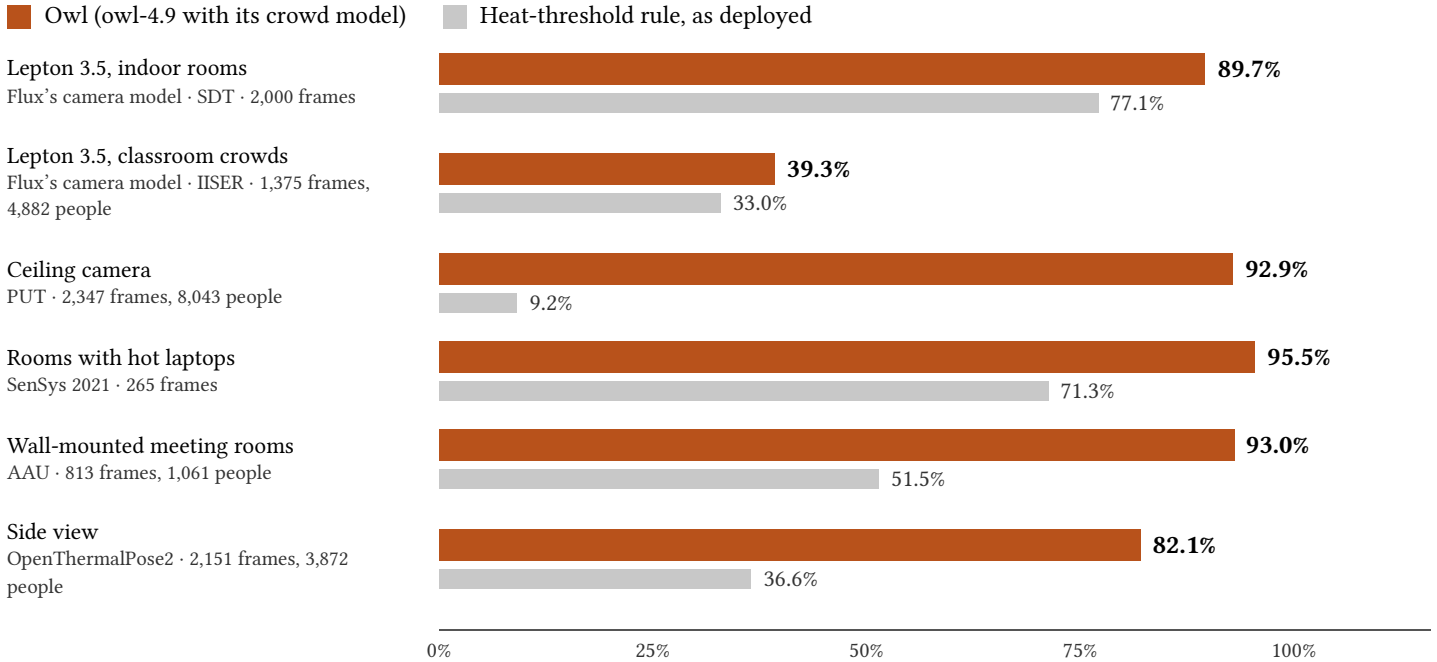


Figure 14. Frames counted exactly right on six real benchmark sets that no model trained on, owl-4.9 with its crowd model against the heat-threshold rule as deployed. The rule is scored on identical frames; Table 17 adds intervals and the rule tuned to each set.

Table 17. owl-4.9 with its crowd model against the heat-threshold rule on identical frames: the six real benchmark sets, the fifteen sealed tests and the stress test. Percent of frames, with 95% intervals for Owl’s exact rate: block bootstrap on the benchmarks; frame-level Wilson intervals on the sealed tests and the stress mean, which ignore dependence within a recording (the margins in the text resample blocks). The last column is the rule with its cell size, warmth margin and minimum group size tuned on each benchmark’s validation split; no sealed set was tuned.

Benchmark set	Owl exact	Owl within one	Rule exact, as deployed	Rule exact, tuned per set
Lepton 3.5, indoor rooms (SDT)	89.7 (85.4–93.4)	100	77.1	93.7
Lepton 3.5, classroom crowds (IISER)	39.3 (31.2–47.7)	73.9	33.0	40.2
Ceiling camera (PUT)	92.9 (86.9–97.2)	99.7	9.2	29.4
Rooms with hot laptops (SenSys 2021)	95.5 (91.7–98.2)	100	71.3	82.6
Wall-mounted meeting rooms (AAU)	93.0 (86.4–98.0)	99.1	51.5	75.5
Side view (OpenThermalPose2)	82.1 (74.4–89.2)	96.3	36.6	60.6

14 Owl: results

Figure 14 and Table 17 compare the shipped pair, owl-4.9 and its crowd model, with the rule on all six real benchmark sets. On five of them Owl counts 82–96% of frames exactly and is within one person on 96–100%, where the rule as deployed manages 9–77% exact. On the crowded classroom set, the hardest of the six development benchmarks, it is exact on 39% and within one on 74%, against the deployed rule’s 33% exact; on this set the two are statistically close on exact counts (a margin of 6.3 points, interval –5.5 to 17.6), and Owl’s lead grows with crowd size.

Benchmark set	Owl exact	Owl within one	Rule exact, as deployed	Rule exact, tuned per set
Sealed test: office, 1–2 people (OpenThermalAction)	94.3 (91.1–96.4)	100	40.7	–
Sealed test 2: another 32 × 24 sensor (SKKU)	33.7 (24.7–44.0)	86.5	79.8	–
Sealed test 2: outdoor handheld camera, people and dogs	41.0 (33.8–48.6)	86.1	16.9	–
Sealed test 3: 32 × 32 ceiling thermopile, 1–5 people (Klir)	9.2 (5.1–16.1)	32.1	18.3	–
Sealed test 4: lab room, 3 views (BU-TIV Lab)	38.0 (33.3–42.9)	81.8	7.3	–
Sealed test 4: atrium, 2 views (BU-TIV Atrium)	26.2 (23.1–29.7)	65.0	35.7	–
Sealed test 4: sports hall, 0–7 players (AAU soccer)	5.1 (3.3–7.7)	25.2	10.4	–
Sealed test 5: 12 public sets, pooled (Roboflow Universe)	40.4 (38.7–42.2)	78.4	33.0	–

Benchmark set	Owl exact	Owl within one	Rule exact, as deployed	Rule exact, tuned per set
Sealed test 6: 3 public sets, pooled (no account needed)	69.7 (67.0–72.3)	90.4	52.2	–
Sealed test 7: industrial hall, groups of 1–7 (UPPET)	8.8 (5.7–13.4)	29.8	20.5	–
Sealed test 8: a café and a restaurant, pooled	5.1 (1.4–16.9)	20.5	5.1	–
Sealed test 9: factory floor, hot robot arms (WLRI-HRC)	22.7 (19.0–26.9)	59.0	3.0	–
Sealed test 10: harbour front, outdoors (LTD)	16.7 (12.3–22.4)	44.0	3.8	–
Sealed test 11: room of 1–5 walkers, head height (LITIV)	3.1 (0.9–10.7)	28.1	18.8	–
Sealed test 12: 32 × 24 ceiling sensor, 0–5 people (MIDetection)	19.0 (15.7–22.8)	55.3	39.7	–
Sealed test 13: living rooms and offices, warm furniture (Thermal-IM)	64.0 (58.0–69.6)	90.8	54.4	–
Sealed test 14: 32 × 24 doorway sensor (TIDOS)	10.2 (7.9–13.1)	31.0	34.1	–
Sealed test 15: one person, 3 views (Almeria)	88.6 (86.0–90.7)	99.7	44.9	–
41-case stress test, mean	94.7 (94.1–95.2)	99.3	–	–

A tuned rule. Owl, one model for all six sets, leads the rule tuned on each set’s validation split by 13 to 64 points on four of them. On the two Lepton sets the tuned rule is level with or ahead of Owl: 93.7% against 89.7% on SDT, whose frames hold one person or none and where most of Owl’s errors are missed people (149 of 206), and 40.2% against 39.3% exact (84.6% against 73.9% within one) on IISER. Adding the sensor’s learned background mask made the tuned rule worse on every video set, because people who sit still fade into it.

A sealed test. To measure Owl on footage that guided nothing, I set aside 300 frames from the office test split of OpenThermalAction, a public MIT-licensed thermal set that no Owl version, setting or threshold had seen, spread across 269 clips of one or two people, at least 30 frames apart within a clip. Owl’s and the rule’s counts were computed and stored first. The frames were then counted visually by two separate passes of the same vision-language model (an AI that reads images), each blind to both outputs and to the other, with a third pass wherever they disagreed; the passes agreed on 93.9% of the 734 frames counted this way (these 300 and the 434 classroom frames recounted below), and on the sealed set the counts match the dataset’s own design (single-person and two-person clips) on 298 of 300 frames, though no human counter has checked them yet. Scored once, Owl counted 94.3% of the sealed frames exactly (95% interval 91.1 to 96.4) and was within one person on every frame. The deployed rule counted 40.7% exactly and 88.7% within one. Owl’s margin is 53.7 points (clip-block bootstrap 47.3 to 60.0).

Beyond its training cameras. A second sealed test asked how far Owl travels, with counts again stored before scoring and truth from each set’s own person boxes (duplicates merged). On 89 frames from a different 32 × 24 sensor (SKKU MLX90640, MIT licence), Owl counted 33.7% exactly against the rule’s 79.8%: it took close, blurred bodies for screens. On 166 outdoor photos from a handheld Seek

camera (Thermal Dogs and People, CC BY 4.0), Owl led 41.0% to 16.9%, though on the 70 frames showing only a dog it still counted a person in 59%. A further set from Voxel51, then the only one with many frames of three or more people, was scored and withdrawn because its licence could not be confirmed; Owl trailed the rule on it. A third sealed test used the one licence-clean public set then found with indoor groups of three or more, 109 unseen frames from a 32 × 32 ceiling thermopile (Klir et al., GPL-3.0): the rule did better, 18.3% exact against 9.2%, and 24.7% against 3.9% on the 77 frames of three to five people, as Owl undercounts the faint, blurred bodies of a sensor it never trained on.

Crowded rooms and a sports hall. A fourth sealed test added the two public sets of larger groups then obtainable. BU-TIV [35], used for evaluation with its authors’ permission, films an indoor lab and an atrium with 16-bit cameras at 512 × 512. On 395 Lab frames from three views, most holding three to five people, Owl counted 38.0% exactly (33.3–42.9) and 81.8% within one; the rule managed 7.3% and 29.1%. On the 338 Lab frames of three or more people Owl was exact on 34.6% and the rule on 3.6%, a margin of 31.1 points (block bootstrap 25.1 to 37.8). In the Atrium, 686 frames of mostly lone, distant people, the rule led 35.7% to 26.2%: one warm body far from the lens is what a threshold rarely misses and Owl, trained on closer views, sometimes drops, though only Owl got empty frames right (32.0% against 0%). Across 1,081 BU-TIV frames Owl led by 5.2 points (0.4 to 10.1), and by 16.3 on frames of three or more. On 393 crops of soccer filmed from high in a sports hall [36], players 25 to 50 pixels wide in 640-pixel-wide crops (6 to 12 in Owl’s input), both failed (5.1% exact for Owl, 10.4% for the rule).

Twelve more public sets. A fifth sealed test used the thermal person sets on Roboflow Universe whose licence permits this use and whose labels say which box is a person: 12 sets under CC BY 4.0 or a public-domain dedication, 2,976 unseen frames after one frame per source image, near-duplicate removal and the unseen check. Six other downloads were set aside before any prediction: two re-uploads of non-commercial datasets (LLVIP and FLIR ADAS), one traced to the source of the withdrawn Voxel51 set, and three whose classes do not say which label is a person. Pooled, Owl counted 40.4% exactly (38.7–42.2) against the rule’s 33.0%, a margin of 7.4 points (block bootstrap 3.2 to 11.6), and was within one person on 78.4% against 64.9%. In an exploratory split by sensor, Owl led by 10.0 points (5.0 to 14.5) on the nine sets from camera-resolution sensors and trailed by 6.5 (–14.7 to 1.4) on the three from 32 × 24 arrays; an occupancy set from a 32 × 32 array, reported apart because it comes from the group whose Lepton set Owl trained on, went the same way (6.6% against 70.0%). On the 222 pooled frames of three or more people the rule led, 20.3% to 9.9%.

Four more, with no account needed. A sixth sealed test used sets anyone can download under CC BY 4.0. On 530 frames from a 720 × 576 camera, almost all of one person, Owl counted 90.2% exactly against the rule’s 45.8%. On 519 frames from libraries, cafés, gyms and outdoor scenes at 256 × 192 (TPD, whose labels can miss a bystander), Owl’s exact agreement was 4.8 points below the rule’s, 58.2% against 63.0% (block interval –12.8 to 2.9), and on its 23 frames of three or more labelled people the rule led, 34.8% to 4.3%. On 83 frames from a 32 × 32 thermopile the rule led again, 25.3% to 10.8%. Pooled over these 1,132 frames, Owl led by 17.5 points (92 blocks, seed 20261014: 8.5 to 26.1); on the 84 pooled frames of three or more people the rule led, 17.9% to 4.8%. On 79 outdoor FLC-Rural frames, reported apart, Owl counted 22.8% exactly against 11.4%. The unseen check removed one set before any score was read: 474 of its 486 candidate frames matched frames Owl was trained or validated on (32 × 24 correlation 0.95 or more), because it is SenSys 2021, one of Owl’s training sources, under another title.

A hall, a café, a restaurant, a factory floor and a harbour.

Four more sealed tests ran on 15 October. UPPEP’s TPE split [37], evaluated with Fraunhofer IOSB’s permission and never to be trained on, films staged groups of one to seven people from high above an industrial hall, 10 to 40 pixels tall in Owl’s input; on 215 frames from 18 sequences (six sparsely labelled ones set aside before any prediction) both counters undercounted, Owl exact on 8.8% and the rule on 20.5% (margin -11.6 , block interval -25.7 to 0.5). The eighth test took the two public videos closest to a venue: in a café of 5 to 19 people neither counter was ever exact (Owl averaged 4.0 people where 8.9 were present), and in a restaurant of one to six each was exact on 2 of 10 distinct frames, Owl within one person on 8 and the rule on 6. On two held-out sessions of WLRI-HRC [38], a factory floor where one or two people work beside two hot robot arms, Owl led by 19.7 points (14.8 to 23.8), 22.7% exact against 3.0% on 432 frames; both overcounted (Owl on 74.1% of frames with one person or none, the rule on 99.5%), Owl’s false counts falling mostly on ceiling lights and a warm machine, rarely on the arms. On 209 held-out frames of the LTD harbour front [40], outdoors, Owl led by 12.9 points (7.8 to 18.1), 16.7% exact against 3.8%, on empty and sparse frames where the rule counted warm pavement and cars; on the 90 frames of three or more people, typically 9 pixels tall in Owl’s input, the rule led, 8.9% to 1.1%.

A room of walkers. An eleventh sealed test, also on 15 October, used the LITIV 2014 set [41] with its authors’ permission: one to five people walking, standing and passing each other in a 5 m room, filmed at head height and rectified to 480×360 , with the authors’ count of each annotated frame as truth and whole videos held out. On 64 frames from the two held-out videos, 205 people in all, Owl trailed by 15.6 points (block interval -28.9 to -3.3), 3.1% exact to the rule’s 18.8%. The rule counted every lone walker. Owl gave 76 person peaks for the 205 people, 75 of them on a person, and put 33 of its 68 screen detections on people. With three or more people both failed (rule 4.3%, Owl 2.2%).

A ceiling sensor in offices and homes. A twelfth sealed test, on 18 October, used MlDetection [43], by permission: a 32×24 MLX90640 on the ceilings of offices, homes and labs, pointing down or at 45° , with the authors’ person boxes as truth and their six test clips, from rooms outside their training set, held out whole. On 479 frames (1,041 people), Owl trailed by 20.7 points (block interval -27.6 to -13.8), 19.0% exact to the rule’s 39.7%, and by 17.9 on frames of three or more (21.6% against 39.4%). It counted 4.0 people a frame where 2.2 were present, and 8.5 in empty rooms: the sealed tests map each image’s 5th to 99.5th percentile onto 20 to 34°C , and that stretch turns a still room’s small differences into warm spots. Given the sensor’s own temperatures, as on Flux, Owl led by 12.1 points (6.1 to 18.2), 32.6% to 20.5%, in a second comparison also fixed before any prediction. The sensor maker’s software (Melexis), whose output ships with the set, was exact on 49.8%.

Heat left on furniture. A thirteenth sealed test, on 20 October, used Thermal-IM [44] (BSD-3-Clause): one person in living rooms and offices, filmed a second after touching or sitting on furniture that keeps the heat. One frame from each held-out clip (261) was counted visually in two blind passes of a vision-language model, which agreed on every frame and have not yet been checked by people. Owl counted 64.0% exactly against the rule’s 54.4% (block interval of the difference -13.7 to 34.3). On frames with one person it saw a second on 25.8% (the rule 44.5%), mostly on chairs and cushions still warm from that person, and both counted someone in all five empty rooms whose chair was still warm: a weakness the generated stress test, where warm footprints score 98%, does not show. Thermal-IM’s training clips did not train Owl. Its only person labels are body poses

in the colour camera’s image; one mapping onto the thermal image put the head on the head in about 21 of 40 checked frames, short of the 36 set in advance, and no frame could be confirmed empty.

People passing under a doorway sensor. A fourteenth sealed test, on 20 October, used TIDOS [45], by permission: a 32×24 MLX90640, not Flux’s 160×120 Lepton, looking down from above two classroom doors. Its labels count the room, not the view, so its 519 frames were counted visually in two blind passes of a vision-language model (agreeing on 516; not yet checked by people). Owl trailed by 23.9 points (block interval -29.5 to -18.7), 10.2% exact to the rule’s 34.1%, finding people in 96% of empty frames, where the stretch made noise into warm spots. On the sensor’s own temperatures it got 95% of empty frames right but split bodies that fill the view, exact on 62.2% to the rule’s 67.2%.

One person at the Lepton’s size. A fifteenth sealed test, on 21 October, used the University of Almeria’s multi-view set [46], by permission: one person posing in a furnished room, filmed by three wall devices whose 160×120 thermal frames hold no temperatures. Its 699 test frames, each showing one person, were counted visually in two blind passes of a vision-language model (agreeing on all; not yet checked by people). Owl led by 43.6 points (block interval 38.6 to 48.9), 88.6% exact to the rule’s 44.9%; of its 80 errors, 55 were a second person at one of two fixed spots where nobody stood.

14.1 From owl-1 to owl-4.9

Owl went through 16 named versions in four days; Figure 15 follows those shipped on the five original real sets (Findings 1, 2 and 9).

■ Owl, exact ---- the rule

Missing points were not measured for that version.

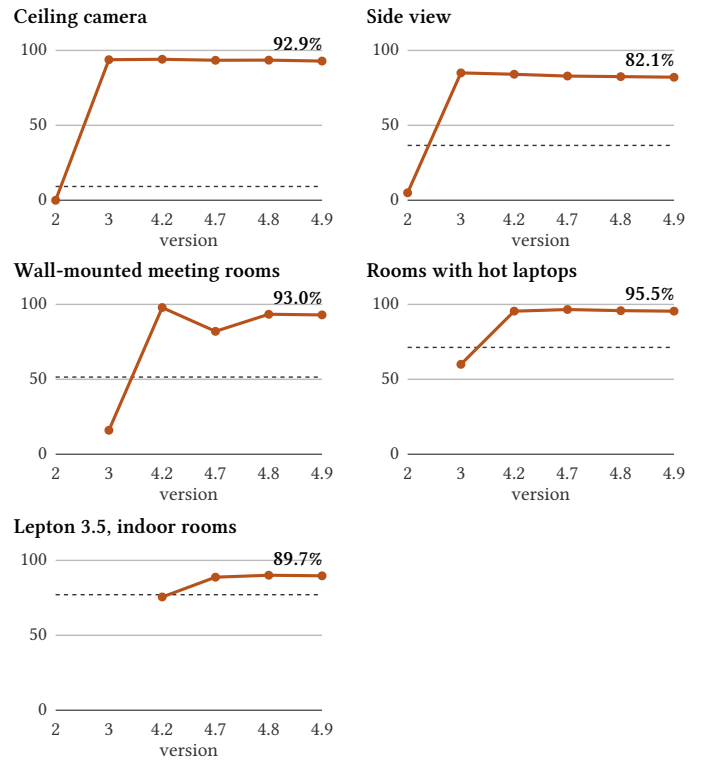


Figure 15. Exact counts on each real benchmark set’s test split by version. owl-4.2 was the first shipped with meeting rooms, laptop rooms and measured physics; owl-4.7 and owl-4.8 added frames from the Lepton 3.5; owl-4.9 added real classroom crowds. owl-3’s side-view figure includes the background fix of Finding 3 (74% before it).

14.2 Like-for-like comparison of the last two versions

When owl-4.9 was trained I re-graded owl-4.8 and owl-4.9 on the full test splits with identical code (Table 18). On the five earlier sets no exact-count rate moved by more than 0.6 points and no within-one rate by more than 0.8.

Table 18. owl-4.8 and owl-4.9, each with its crowd model, graded together on full test splits. Exact / within one, in percent. The owl-4.8 stress mean comes from a run before each case’s seed was fixed; owl-4.9’s is the seeded rerun.

Test set	Rule exact	owl-4.8	owl-4.9
Lepton 3.5, indoor rooms	77.1	90.1 / 100	89.7 / 100
Lepton 3.5, classroom crowds	33.0	11.5 / 20.6	39.3 / 73.9
Ceiling camera	9.2	93.5 / 99.7	92.9 / 99.7
Rooms with hot laptops	71.3	95.8 / 100	95.5 / 100
Wall-mounted meeting rooms	51.5	93.4 / 99.9	93.0 / 99.1
Side view	36.6	82.5 / 96.6	82.1 / 96.3
41-case stress test, mean exact	–	94.5	94.7

14.3 Crowds on the Lepton 3.5

The IISER classroom set is the only real crowd I have from the Lepton 3.5: rows of 12 to 15 students, up to ten with a visible face, each row partly hidden behind the one in front. owl-4.8 had never seen anything like it and overcounted badly, averaging 8.1 people against 3.6 labelled faces. Figure 16 separates the two halves of the fix. Training the main model on the classroom frames lifted exact counts from 11.5% to 28.6%. Retraining the crowd model as well took them to 39.3%, and on frames with six to ten labelled people moved within-one from 3.4% to 96.1%.

Owl’s strength is the crowd: on frames with six to ten labelled people it is within one 96.1% of the time, against the deployed rule’s 4.5%. Sparse classroom frames are the opposite case. Against the face

labels, the rule is exact on 37.9% and 72.7% of frames labelled with one or two people, against Owl’s 14.6% and 22.4%, and on the 88 frames labelled empty Owl reports 2.6 people on average. Every classroom frame labelled empty or with one face falls in IISER’s test split (training holds 17 two-face classroom frames and none with fewer), so Owl never trained on an empty or one-person classroom. The 334 test frames labelled with two or fewer faces, and 100 others, were recounted visually, every person with any part in view, in the blind two-pass procedure of the sealed test: the face labels undercount (frames labelled empty held 1.1 people on average, and none was empty), which narrows Owl’s error on sparse frames. On frames holding one or two people the rule remains more accurate (85% exact, against Owl’s 21% to 22%), in part because Owl puts points on warm seat backs; on frames holding three or more, Owl leads (43.5% exact, against the rule’s 21.1%). In an exploratory check that mixes the two label definitions, with the 434 recounted frames substituted for their face labels, the IISER test reads 41.5% exact and 77.7% within one for Owl, against 40.4% and 69.2% for the rule.

When the crowd model steps in is a single setting, chosen on the classroom frames and the stress test: handing over earlier was worse on both (Table 19), so the sensor keeps five.

Table 19. When to ask the crowd model, owl-4.9 pair. The stress means and blind-spot counts come from runs before each case’s seed was fixed, which can differ by up to 10 points between draws; the seeded rerun of the shipped setting gives a stress mean of 94.7%.

Crowd model asked from	Classroom exact	Within one	Stress mean	Blind spots
3 people	36.9	72.3	94.3	12
4 people	37.5	73.1	94.6	10
5 people (shipped)	39.3	73.9	94.9	11

■ owl-4.8 with its crowd model ■ owl-4.9, old crowd model ■ owl-4.9 with its crowd model

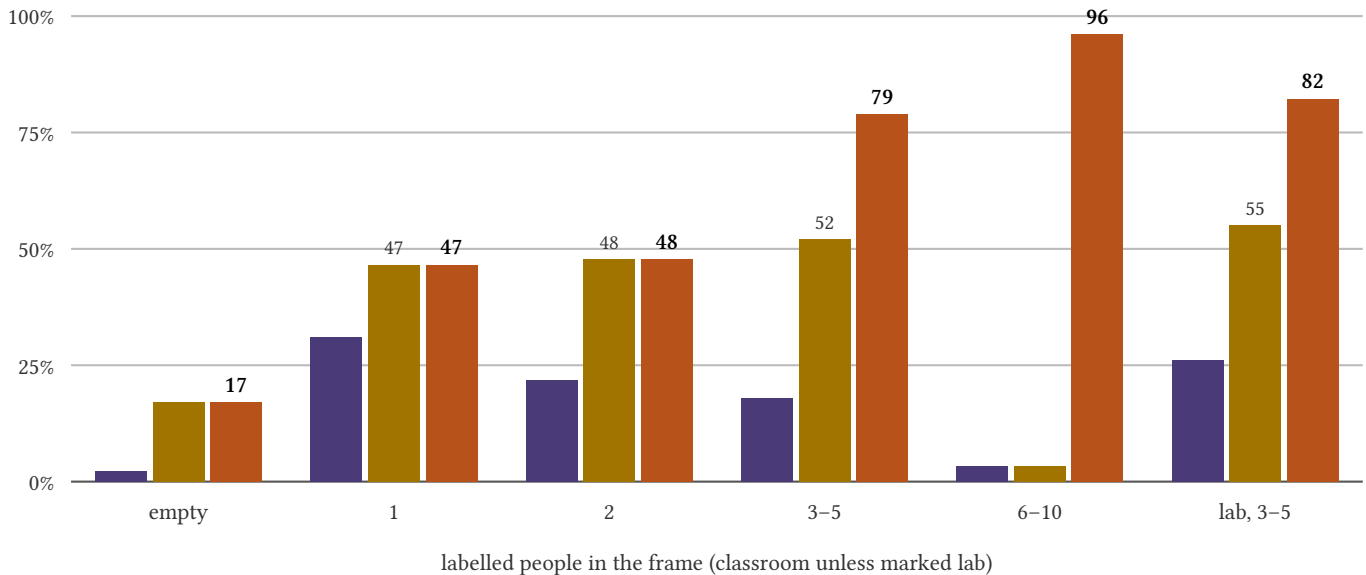


Figure 16. Frames within one person of the truth on IISER’s held-out classroom and lab frames, by how many people were labelled in the frame. The crowd model’s contribution is the difference between the gold and orange bars; it decides the six-to-ten bucket almost on its own.

14.4 Hard cases

Figure 17 shows all 41 stress cases for the shipped model. In generated scenes, every warm object that fools the rule by design is

counted correctly on 98 to 100% of stress frames, where the deployed rule manages 14% to 85%: a hand at the lens, a hot mug, a monitor, a radiator, a pet, a reflection, warm footprints, a candle; real heat left on furniture is harder (sealed test 13). Four cases fall under 90% exact

(Figure 17): people lying flat, a bare torso at the lens, one person behind another (99.3% within one) and a crowd of ten (31.3%, 95% interval 24.5 to 39.1; 75.3% within one).

14.5 Naming everything else

On 3,000 held-out generated frames, owl-4.9 finds the nine other kinds as Table 20 shows. These names are shown on the sensor’s screen to explain why something was not counted.

Table 20. Object naming on held-out generated frames. Found: share of real objects with a peak. Right when named: share of named peaks that were that kind.

Kind	Objects	Found	Right when named
person	10,461	74.9%	89.7%
hand	682	69.5%	84.3%
pet	617	67.1%	92.4%
hot drink	836	84.2%	83.0%
food	441	81.2%	91.6%
laptop	768	68.2%	79.8%
screen	491	60.5%	80.7%
heater	794	78.7%	85.5%
lamp	316	77.5%	79.0%
warm seat	586	58.7%	83.1%

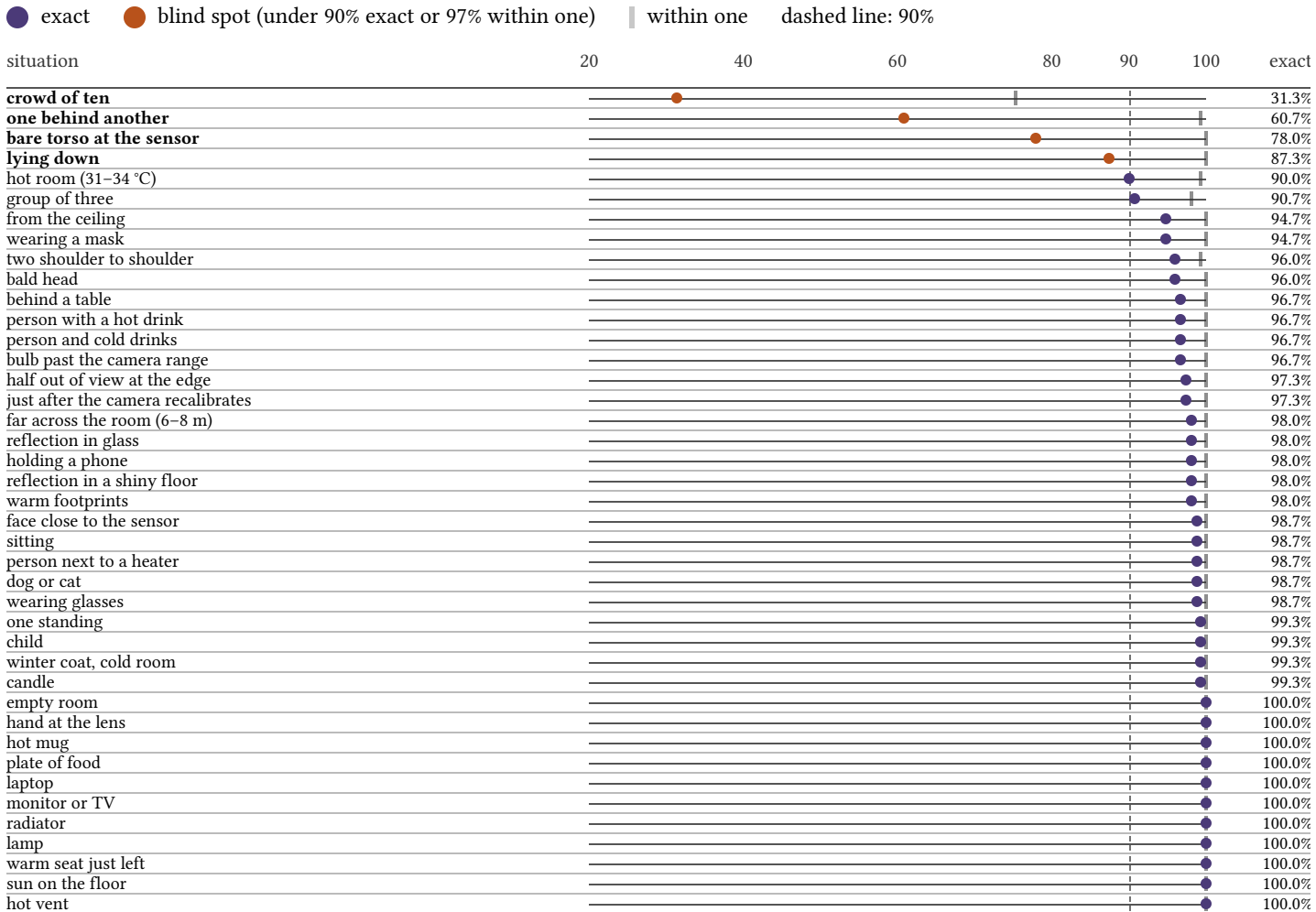


Figure 17. The 41-case stress test for owl-4.9 with its crowd model, 150 generated frames per case from a seed fixed per case, sorted from hardest to easiest.

The same 3,000 frames are deliberately harder than real rooms, with crowds of up to ten drawn often. On them owl-4.9 counts 59.3% exactly and 84.3% within one, against the rule’s 20.2% and 49.9%; empty rooms are right 92.3% of the time and single people 80.1%. The generator grew harder with each version, so these synthetic figures are not comparable across versions.

14.6 Speed and size

On one core of the sensor’s Pi 5, the main model of owl-4.8 and owl-4.9 takes 14–16 ms per frame to run and decode, measured by the sensor’s own self test (ten reads) on its own camera. A frame with five or more people also runs the crowd model, a network of the same size (722,655 weights in its exported file against 717,422 for the main model, whose 718,550 training parameters shrink slightly at export), which roughly doubles that frame’s time; that path has not yet been timed on the Pi.

15 Owl: findings

Every change to Owl was kept or thrown out on the real benchmark sets. These are the results that shaped it, in the order I found them, including the alternatives set aside. Stress-test figures here come from runs before each case’s seed was fixed and can differ by up to 10 points between draws.

F1 *Generated frames alone fail on real footage*

owl-2 counted 80% of generated frames exactly and found 73–94% of the warm objects, but counted 0% of real ceiling frames and 5% of side-view frames exactly: it had only seen clean drawn shapes and read a real room’s texture as people.

F2 *A little real data fixes most of it*

A 500-step test with real frames mixed in went from 0% to 81% exact on the real ceiling set. The full run, owl-3, reached 94% on the ceiling set and 74% on the side view. From then on, generated frames also gained real-looking texture and low-resolution blocking.

F3 *Judge warmth against the cool end of the frame*

The physical check compared each person with the room, taken as the frame’s median; when a person fills the view the median is the person, and the check threw out the person just found. Taking the room as the 10th percentile lifted the real side view from 74% to 85% exact and 86% to 98% within one, hot rooms from 47% to 68%, and a bare torso at the sensor from 16% to 30%, with the ceiling set and every false-alarm case unchanged.

F4 *Physics beats tricks*

Two training changes that seemed sensible, skewing the absolute-temperature channel and drawing real group frames three times as often, changed nothing on the ceiling and side-view sets and cost a few points on unseen rooms. Replacing hand-picked temperatures with measured skin and clothing physics lifted the unseen meeting rooms from 24% to 33% exact and from 58% to 85% within one.

F5 *A bigger network did not help*

Widening the network from 24 to 32 base channels tied the smaller one (94.2% and 73.9% against 93.8% and 74.2% on the ceiling and side view). Accuracy was limited by data, so later versions added data rather than width.

F6 *More of a hard case can cost more than it gains*

owl-4.4 drew more people lying down and more bare bodies at the sensor, the two single-person cases still under 90%. The real meeting rooms fell from 97.9% to 60.9% exact with no measurable gain on the targeted cases, and the change was reverted.

F7 *A sharp target at a guessed point teaches the model to miss*

A finer 80×60 output grid (owl-5) keeps heads a few pixels apart as two people. But on the datasets whose head points are estimated from a body box, a sharp target at the guess taught it to miss people: the meeting rooms fell from 97.9% to 67.2%. Spreading those targets over the pixels the head could occupy fixed the cause, and became standard for every later version.

F8 *Density counts crowds that peaks merge*

Used as a second look in crowds, the finer-grid model lifted a crowd of ten from 41% to 63% within one. Back on the 40×30 grid, a density head, whose total does not fall when heads touch, lifted it to 82% and put one person behind another at 100% within one, with the four real benchmark sets then in use unchanged. It became the crowd model (owl-4.3).

F9 *Data shares trade off, and the right data ends the trade*

Adding real Lepton frames (SDT) improved results on that camera, but every way of making room for them hurt the meeting rooms: cutting the meeting rooms’ share to 15% cost 25 points, and owl-4.6 gained 13 points on the Lepton while losing 14 on the meeting rooms. owl-4.7 chose the camera Flux uses (88.8% against owl-4.2’s 75.6%). Adding SDT’s rendered Lepton-style frames in owl-4.8 improved both at once (Figure 18).

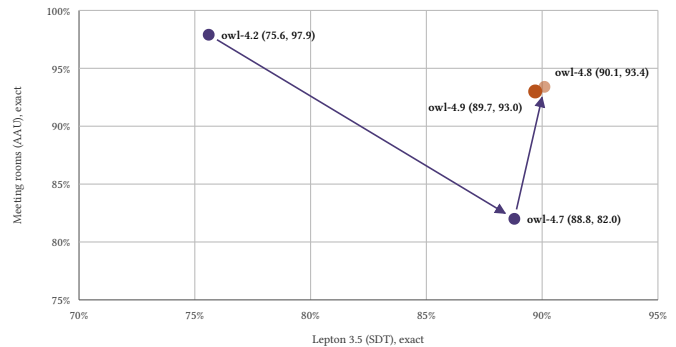


Figure 18. The trade between the Lepton 3.5 frames (SDT, horizontal) and the wall-mounted meeting rooms (AAU, vertical). owl-4.7 bought the camera with the meeting rooms; owl-4.8, trained with rendered Lepton frames as well, recovered the meeting rooms without giving the camera back. owl-4.9 differs from owl-4.8 by under one point on both sets.

F10 *Validation scores do not compare across versions*

When training drew group frames three times as often, the validation draw reused that weighting by accident, and owl-3.2 looked worse than owl-3 on what was really a harder mix. Validation follows each run’s own mix of sources, so a validation score picks a checkpoint within its run and is never compared across versions.

F11 *Lower precision cost more than it saved*

An 8-bit export ran 2.1 times faster at a quarter of the size, but cut owl-3’s exact count on the real ceiling set from 94% to 86%, and to 84% when calibrated on real frames. At 14–16 ms per frame in full precision, speed was never the constraint, so the sensor keeps full precision.

F12 *A newer base can make a worse crowd model*

Retraining the crowd model starting from owl-4.8 instead of its own earlier weights dropped a crowd of ten to 18.7% exact. The older crowd model stayed until real crowd data arrived.

F13 *A real room shows what no test set does*

On the sensor in a real room, owl-3 named a pet where there was none, sometimes counted a monitor, and missed a bare torso near the camera. Three changes answered them: higher naming bars for rarer kinds, dropping a person where a screen-like object is more confident, and bare skin and bodies filling the view in the generated frames. On the stress test a bare torso at the sensor went from 16% under the first owl-3 decoder (30% after Finding 3) to 78% under owl-4.9 with the seeds fixed.

F14 *Real crowds on the Lepton 3.5 were the missing data*

After owl-4.8, training on the same data had stopped helping. Only after IISER Bhopal licensed its Lepton 3.5 classroom crowds to me did the crowd numbers move: 11.5% to 39.3% exact, an average error of 4.66 people to 1.03, and rooms of six to ten from 3.4% to 96.1% within one, while the other benchmark sets held steady (Section 14.2).

F15 *Validation stopped improving early in the run*

owl-4.9's real validation score stopped rising about a quarter of the way through its run (Figure 19). The rest of the run brought no sustained gain, so the next experiment is new data: more real crowds and side views from this camera.

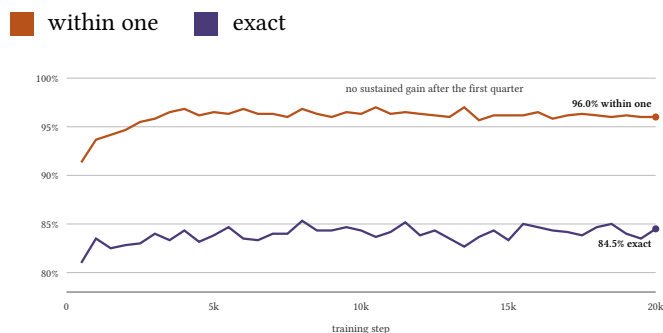


Figure 19. owl-4.9's real validation score every 500 steps over its 20,000-step run. Both curves are flat after about 5,000 steps. The validation mix includes the new classroom crowds, so these values are lower than earlier versions' and are not comparable with them.

16 Privacy by design

Flux is built to tell people how busy a place is. It must never become a camera that watches them.

- **Kinds, never who.** Owl's outputs are ten kinds of thing. It has no identity output, no face class and no face features. IISER labels faces, and I use each face box only as the point where a head is; Owl is trained to locate people, never to predict who they are.
- **No memory between frames.** Owl reads each thermal frame on its own and keeps no track of anyone. Flux's separate doorway sensor follows a crossing only for the moment it takes to tell an arrival from a departure.
- **Numbers are all that leave.** The network's points and boxes exist only in memory. Boxes are drawn only on units with a screen and are never sent. What leaves the sensor is a handful of numbers: people counts and a sound level, never an image.
- **Enforced by tests.** Tests in the sensor's suite fail the build if Owl gains an identity output or carries anything between frames, or if an image, a point or a box leaves the device.
- **Data used as licensed.** Every dataset is used as its licence or permission allows. None of their images are redistributed or published, and none are shipped with the model. IISER's licence also forbids using its data to identify anyone.

The model files are kept private, like Flock's crowd-forecasting model. The generator, network and training and grading code are public. Without a model on disk the sensor counts with the rule, and says so.

17 Future work on the counter

The clearest next step is data: further thermal crowd datasets have been requested, and the gap that matters most is real radiometric crowds from the Lepton 3.5 in venues, measured against hand counts. Flux is built to carry a doorway counter, an 8×8 time-of-flight sensor that tells arrivals from departures; it first ran on the bench in October 2026 and has not yet counted a real doorway. Flux's software already combines its running total with Owl's count of people in view and estimates how long people stay as occupancy divided by the arrival rate (Little's law [28]); tested in a real doorway, that would answer how many people are inside as well as how many are visible, without identifying anyone. Finally, each crowd decision currently rests on one frame; agreeing a count across consecutive frames, without following individuals, may steady crowded readings further.

18 Conclusions on the counter

The second question asked whether a credit-card-sized computer model can be designed to count the number of people inside a venue using a thermal camera without identifying individuals. Yes in simple rooms, not yet in crowds, and not yet inside a working venue. On one core of a Raspberry Pi, at 14 to 16 ms a frame, Owl counts people on a 160×120 thermal camera and sends only counts, with no identity output and no memory between frames (Section 16). It is 82 to 96% exact on five of six development benchmarks (rule 9 to 77%) and 94.3% on a sealed office set (rule 40.7%), and leads the rule when 15 further public sets are pooled in two sealed tests, on a factory floor beside hot robot arms and on an outdoor harbour front. It falls short on low-resolution arrays, people walking past a head-height camera and dense groups in a sports hall, an industrial hall or a café; the November retrain adds training data for all but the industrial hall, whose authors allow evaluation only.

19 How it was built

Both studies were built in seven months, from the first collector in March 2026 to the sensor's doorway counter in October. The forecast's spring-trained model, rebuilt after an August audit, gave way in September to a correction built on live readings; Owl replaced the heat rule in four days. Table 21 gives the timeline; dates are Eastern time.

19.1 The forecast: corpus and first models (March to May 2026)

The first collector was committed on 10 March 2026 and widened to 34 cities the same day. From the start it gathered two kinds of row from the busyness provider: weekly rows, one typical week per venue of 168 hourly values on the provider's 0 to 100 scale, and realtime rows, one reading per venue-hour on a real date. Event listings were attached the same week, and a model pipeline with an ONNX inference service followed on 14 March. In early May the collector reached the 13 of its 34 cities it had not yet tried: a probe of 100 venues in each found 996 of 1,300 (76.6%), and those entered the corpus. Beijing, where 37 of 100 were found, stayed thin and holds no rows in v2.6.0's training data, so 30 cities train the model and three are held out (Section 4.5). The last spring reading is dated 18 May 2026, and by then the corpus held 3,912,357 rows across 34,784 venues in 34 cities (Section 4.2).

The first models predicted busyness directly, with the venue's popular-times curve among their inputs. An 80-feature model of this kind scored far higher on validation than on held-out cities: it had learned to copy the curve, and coordinates acted as a lookup table for the five cities it was trained on. That result set three design decisions that still stand. Coordinates were dropped as features. Cross-validation holds out a whole city at a time. And the label became the *delta*, how far the hour departs from the venue's own curve, with the served score the curve plus a clamped correction. On 10 May, v2.2.1 read the curve from the server's table, the first fix making training and serving agree (train/serve parity). No accuracy figure from before August is used as evidence, because the corpus was later repaired (Section 4.2).

19.2 An end-to-end audit, and v2.6.0 (August 2026)

In August I audited the whole pipeline and found two silent failures. The model had never loaded in production, because the package that reads its ONNX file was missing from the server, so the rule engine had answered every request. The collector had written no rows since 19 May while reporting success, because its provenance check compared zero rows written with zero rows labelled. Both now fail loudly. On 12 August a series of versions was trained. v2.3 trained on the rows production serves, realtime rows with a usable curve, because weekly rows carry a delta label of zero, which had taught earlier models to shrink every correction. v2.3 to v2.4 failed their gates, and v2.5.0 shipped that afternoon. The next six days were spent on the data and the pipeline:

- *Train/serve parity* (14 August). The server was not building the inputs training built: live predictions had been computed at coordinates 0, 0, serving read all of a venue's place types where training read the first three, and the event-size estimator had drifted from the one that labelled the corpus. A parity break now stops the model from loading.
- *The clock* (15 August). The weekly collector had stored the provider's slot number as the hour; the provider's day runs from 06:00, so slot 18 was the venue's midnight. A database fix corrected 3,536,360 rows, and a row with no declared clock is now refused.
- *A second audit* (15 August). A read-only pre-retrain audit listed 21 findings, 8 of them blocking, among them 491,100 duplicate rows, an impossible month on 62.9% of rows and a gate that scored rows production never serves. All eight were fixed before the next run.
- *Provenance* (15 August). The 457,402 spring realtime rows cannot be split into live readings and provider forecasts; two attempts to recover the split failed against null controls. Those rows stay marked unknown, and since 4 September they are excluded from training by default.
- *A category leak* (16 August). Two category features had been fitted on every city, so a held-out city helped build its own scores. They are now refitted without it.

The first retrain in which every stage had been checked shipped as v2.6.0-starling on 18 August, with the result of Table 6. The gate's old within-10 floor of 29.2% had been derived before the clock fix, so it was replaced by a measurement, the incumbent's 19.3% on Table 6's rows, which v2.6.0 beat at 20.7%. Of two copies, trained on a CPU and a GPU and within 0.02 MAE of each other, the CPU copy shipped because its run reproduces bit for bit.

19.3 Alternatives tested, and the decisions they produced (August to September 2026)

After v2.6.0, every proposed change faced a gate before it could ship, and each result is recorded so the experiment is not repeated.

- *Calendar features* (19 August). Varying the month alone across the year moved v2.6.0's raw correction by 27.31 points on average, against 5.92 for the hour: with one spring of data, the calendar features had absorbed drift in the collection. v2.7, without month and season, failed its gate (MAE 29.89, within-10 19.9%), so v2.6.0 stayed; the November retrain replaces the month features with live readings (Section 7.2).
- *Calibration* (20 August). About 40 output corrections were fitted on the earliest 30% of dates and scored on 46,101 later rows. There a quantile display map gained 8.40 points of within-10 while raising MAE by 3.15. It went live on 28 August; September's live readings later showed it cost within-one accuracy, and it left the served number on 26 September (Section 6.6).
- *Hour ranking* (20 August). Within one venue-night the model ordered two hours correctly 62.7% of the time and the curve 63.1%, over 31,498 pairs, so the app's "go later" advice reads the curve and stays silent below a 10-point gap.
- *Event provenance* (20 August). 22 of 34 cities held no row with a nearby event because their events were never collected. A new provenance column now separates a measured absence from a missing lookup.
- *Game nights* (30 to 31 August). Games of seven teams became six features. Game nights are 7.1 points busier in the labels, but the spring training data held about 19 days of games, and the features did not move within-10. They stay switched off until the retrain.
- *The two-head model* (5 September). It had separate heads for weekly and live rows. Held to the verified live labels then available, it lost to the shipped model, and a model given 304,000 more rows of the same kind lost to one trained without them in a single paired run (Section 6.6). Both results pointed to the venue's own recent readings as the input to build.
- *Forty-two proposals* (6 September). Six rounds of research proposed 42 changes; an automated reviewer, told to refute each one, left seven standing. The one shipped was the per-venue trailing offset (Section 6.6).
- *The curve's range* (6 to 8 September). The poor calibration of Section 3.1 invited a slope correction; it lowered MAE but cost within-10, so the system keeps the per-venue offset (Section 6.6). The same week measured what coverage is worth (Section 6.5), which made adding venues with a curve the collector's first job.

19.4 Live collection (September 2026)

Live collection restarted on 1 September for about 1,400 venues in Philadelphia and the Lehigh Valley, and every row now carries a provenance label. Calls settled at one a second after the provider's rate limit throttled the first day's 600 a minute. The sweep went hourly from 20:08 on 3 September (00:08 UTC on the 4th) and now runs at seven minutes past each hour. An open-hours filter stopped calling venues that their own curve says are shut, cutting the mean sweep from 1,414 calls to 993. From 5 September, Miami, one of the three held-out cities, is read at two fixed hours a day, so the gate has live readings from a city training never sees. From 6 September the venues users open lead every sweep, and from 25 September calls run in parallel, up to eight at once, so the full sweep finishes inside the hour. The pool of verified live readings grew from 4,223 on 4 September to 9,339 over 1,228 venues on 6 September.

19.5 The switch to curve, offset and nowcast (25 September to 13 October 2026)

On 25 September a replay of the serving code (the band replay) over every live reading in the 8 September export measured, for the first time, what the card is scored on: whether the served number lands in

the same crowd level as the live reading. As served, v2.6.0 with the display map landed within one level 57.4% of the time on 7,920 readings, and a reading one hour old predicted the next within one level 92.0% of the time. Eleven causes of error were ranked by the points each could recover, and the two largest needed no retrain. Two serving switches followed: the curve plus 0.85 of the trailing offset, and the nowcast of Equation (2), with weights fitted on 1 to 5 September and scored on 6 to 8 September. The switches went through six rounds of adversarial review by automated tools, one of which caught that the offset could see the hour it was scoring; the fix restricted it to readings strictly before the hour. They went live in production at 18:55 on 26 September, on the three-day result in the last row of Table 4. On 25 and 26 September the retrain’s split rule and thresholds were committed (Section 7.2), and on 26 September the retrain was rehearsed (Section 6.7). On 8 October a fresh read-only export scored the switches on every live reading since 9 September (Section 6.1). On 13 October, when persistence baselines showed that a two-hour reading no longer helped, production switched to rule (b), the simpler rule the retrain gate now compares against (Sections 6.1 and 6.3).

19.6 The parity tests

The train/serve parity faults of Section 19.2 were the hardest to see, so the safeguards built since are tests that training and serving agree:

- the ONNX export verifies feature order, and serving refuses a model file listing a feature it cannot build;
- before any fold is fitted, a run with no city held out must reproduce the shipped category features bit for bit, and a control that rebuilds them the old, leaky way must be caught;
- the curve smoothing, the neighbour features (on 8,006 readings), the live features, the game features and 62 calendar, weather and event columns are each computed by trainer and server on shared test data, with zero disagreements allowed;
- the collector writes the event values serving computes, checked over 400 listings;
- the band replay must publish what the real prediction function publishes on every row;
- a leak test changes every reading at or after a row’s hour, including its own label, and requires none of the row’s inputs to move;
- the serving responses of v2.6.0 must match recorded answers under every switch combination.

19.7 Flux, the sensor (May to October 2026)

The sensor began as backend tables and an endpoint on 2 May 2026, tested end to end before any device existed. The demo unit’s design was fixed on 4 May, at about \$1,000 of parts: a Raspberry Pi 5, a FLIR Lepton 3.5 on a PureThermal 3 board, a 4G board, an analog microphone, an infrared break-beam and a battery. The Lepton was chosen over the 32 × 24 MLX90640 for its 160 × 120 temperature image (19,200 pixels against 768), and the sensor program moved onto it on 26 August. The first bench day on a Pi 5 was 6 September: the camera came up on the first try, and pooling the image into 4 × 4 cells stopped a bare head and a clothed torso from reading as separate people. On 21 September the enclosure was designed in OpenSCAD and the device split in two, a base with the Pi, battery and screen, and a wall head with the sensors looking down at a doorway. On 7 October an I2C decibel meter replaced the analog microphone; it measures A-weighted sound level on its own processor, so the Pi will read a number and never audio. On 8 October the base and the wall head were merged into one box, because the doorway sensor and the decibel meter need short I2C wires.

19.8 Owl, the training chain (26 to 29 September 2026)

Owl’s versions, runs and findings are in Sections 12.5, 14.1 and 15. Licence requests went out on 27 September; TU Wien granted SDT on 28 September, and IISER Bhopal licensed its Lepton 3.5 classroom crowds on 29 September. A dataset whose authors had not replied was not used, and another was left out for its copyleft labels; November trains it only into a second candidate, beside one built without any copyleft set. The shipped pair, owl-4.9 and its crowd model, draws 60% of its training frames from licensed datasets (SDT’s rendered Lepton frames among them) and 40% from the generator.

19.9 The doorway counter (September and October 2026)

The doorway counter went through three designs. The infrared break-beam of the May plan became a one-sided infrared proximity module on 21 September, with one part to mount and nothing to align. On 25 September it became a VL53L8CX time-of-flight sensor, which looks down at the doorway as an 8 × 8 grid of distances and tells arrivals from departures. Its driver was written on 26 September, before the part arrived, by porting the manufacturer’s driver onto the Python standard library. On 7 October the counter ran on the bench for the first time, reading 265 frames.

Table 21. Timeline of both studies, 2026. F: the forecast; C: the counter.

Date	Track	Milestone	Result
10 Mar	F	First collector; 34 cities	Weekly curves and readings begin
Mar to Apr	F	Direct models with the curve as an input	Copied the curve; delta label adopted
2 May	C	Sensor tables and endpoint	Tested before any device existed
4 May	C	Demo unit designed; Lepton 3.5 chosen	About \$1,000 of parts
10 May	F	v2.2.1 reads serving’s curve table	First parity fix
18 May	F	Last spring reading	Corpus of 3,912,357 rows
12 Aug	F	Audit; v2.3 to v2.4 fail; v2.5.0 ships	Model loading and collector checks rebuilt
14 to 16 Aug	F	Parity sweep, clock fix, 21-finding audit, leak closed	3,536,360 rows corrected
18 Aug	F	v2.6.0-starling ships	MAE 29.42 against 31.48, 67,249 rows
19 Aug	F	v2.7 without month and season	Fails gate, MAE 29.89
20 Aug	F	About 40 calibrations; quantile map	Map +8.40 within-10, +3.15 MAE (46,101 forward rows)
26 Aug	C	Sensor program moves to the Lepton 3.5	768 to 19,200 pixels
30 Aug	F	Game-night features	Within-10 unchanged; held for the retrain
1 Sep	F	Live collection restarts with provenance	One call a second
5 Sep	F	Two-head model	MAE 1.59 worse than v2.6.0
5 Sep	F	Miami holdout hours begin	Two runs a day
6 Sep	F	Demand-first sweeps; trailing offset ships at half weight	MAE 25.90 to 24.53
6 Sep	C	First bench day on a Pi 5	4 × 4 pooling
6 to 8 Sep	F	Curve range measured; slope refused	Slope 0.466; coverage worth 1.82 MAE

Date	Track	Milestone	Result
21 Sep	C	Enclosure; base and wall head	
25 Sep	F	Band replay; parallel sweeps; retrain rules committed	As served 57.4% within one
26 Sep	F	Curve + offset + nowcast in production, 18:55	Exact level 56.0% on the three-day test (the model it replaced: 32.2%)
26 Sep	F	Retrain rehearsed with live features	79.7% within one; refused by the gate
26 Sep	C	owl-1, generated frames only	
27 Sep	C	owl-2 fails on real frames; owl-3 and owl-4.2	Real ceiling 0% to 94%
28 Sep	C	Crowd model; owl-4.7 and owl-4.8; SDT permission	Lepton 3.5 set: 75.6% to 90.1% exact
29 Sep	C	owl-4.9 and its crowd model ship	Crowd error 4.66 to 1.03 people
3 Oct	F	Provider's live coverage drops	Live readings fall to about 61% of the week before
7 Oct	C	Doorway counter's first frames; decibel meter chosen	265 frames
8 Oct	F	Month-long test of the served number	54.9% exact one hour ahead, 119,939 readings
12 Oct	F	Provider returns no live readings from about 16:00	
13 Oct	F	Persistence baselines scored; production simplified	Hour-old reading at full weight, else curve + offset
13 Oct	C	Owl re-scored with intervals, a tuned rule and a seeded stress test; sealed test scored once; IISER frames recounted	Sealed test 94.3% exact on 300 unseen frames; rule 40.7%
14 Oct	F	Retrain moved to early November	Rules unchanged
14 Oct	F	Replication on public counts, plan hashed before download; KIT grants its counts	15 datasets, 5.3 million counts; curve + latest departure lowest in 14 of 15
14 Oct	C	Sealed tests 2 and 3	The rule leads on two low-resolution sensors
14–15 Oct	F, C	More than 40 data requests sent	
15 Oct	C	Sealed test 4: BU-TIV (with permission) and a sports hall	Lab room, 3+ people: Owl 34.6% exact, rule 3.6%
15 Oct	C	Owl joins the November retrain; new sets split by recording	Gate fixed before training
15 Oct	C	Sealed test 5: 12 public Roboflow sets	Pooled: Owl 40.4% exact, rule 33.0%
15 Oct	C	Sealed test 6: four sets that need no account	Three pooled: Owl 69.7% exact, rule 52.2%
15 Oct	C	Sealed tests 7 to 11: an industrial hall, a café and a restaurant, a factory floor, a harbour, a room of walkers	Crowds defeat both; Owl leads on the factory and harbour, the rule on the walkers

19.10 Collecting the data, and asking for it (September and October 2026)

The forecast's readings come from the provider's API through Flock's own hourly collector (Section 19.4), stored with their provenance and never redistributed. Everything else came from searching for public data and asking for what was not public. A survey of open occupancy data found no public set that pairs a busyness score with counts of people, so the replication of Section 6.8 and its extensions use counts alone: 31 sources and 6,497,751 stored hourly records, credited in the acknowledgements. The 15 core datasets were analysed under a plan hashed before their first download; the 16 extensions are exploratory, and of 247 imported campus Wi-Fi access points, 153 qualified for scoring.

Where a licence was missing or unclear, I asked. On 14 and 15 October I sent more than 40 requests: to city transport and football teams (Wellington, York, Toronto, Amsterdam), university groups (KIT Library, Munich, UCL's GeoDS, Unisport Bern, Kassel), pool, gym and venue operators (Zurich, SWM Munich, Fitnesspark, The Cliffs), occupancy companies (Occuspace, hystreet, TouringPlans, NickOccupancy) and the owners of 26 thermal datasets, among them BU-TIV, TIDOS, LLVIP, RGBT-CC and two Lepton sets, one filmed in Prague's National Museum. KIT Library granted its Seatfinder counts, through its API and with credit, and the University of St. Gallen Library its own. BU-TIV's first author granted evaluation of the Atrium and Lab sequences within a day (sealed test 4) and has been asked about training. UPPET's authors allowed evaluation, not training (sealed test 7), WLRI-HRC's authors confirmed its CC BY 4.0 licence for evaluation and training (sealed test 9), the LITIV lab at Polytechnique Montréal granted evaluation and training (sealed test 11), KU Leuven allowed academic use of MlDetection (evaluated in sealed test 12) and VIPER, Boston University research use of TIDOS (sealed test 14), and the University of Almería evaluation and training on its multi-view set (sealed test 15). Fitnesspark declined, so its counts, and a public archive of them, were left out. The rest are pending.

The same rules held for every source. A set with no licence and no permission is not used (IPHD). A portal that refuses scripted requests is asked by email, not worked around (York). A set whose uploader licensed material it did not own is withdrawn, even after scoring (the Voxel51 thermal set). No frame of a test set is published. Where a set's labels are faces (IISER), each face box is used only as the position of a head.

20 Limitations

The first four limit the answer to the first question; the last two limit the answer to the second.

- **Relative readings, not headcounts.** The main forecast study's labels are the provider's relative readings on a per-venue 0 to 100 scale, estimates from its own system. The forecast therefore predicts the provider's next live value, and a score of 70 means different crowds at different venues. No count inside a venue has checked these readings, though their weekly rhythm tracks the subway entries beside 695 New York venues (median r 0.74, Section 6.8). On 15 public real-count datasets the provider-specific findings do not hold as stated, so they describe the provider's estimates rather than crowds in general. But 13 of those datasets count people passing a sensor; the two registered occupancy datasets and the KIT, Singapore and Barcelona series added afterwards side with the provider: the hour-old reading beats the curve (Section 6.8).
- **The test window.** The served accuracy rests on 119,939 readings over 30 dates in Philadelphia, the Lehigh Valley and Miami, three

of the 34 cities in one season; the nowcast's gain comes from the two Pennsylvania regions alone. The online correction shipped on a three-day test, short of the five dates the band gate asks of a model; the 30-day test confirms it. Production's current rule was identified on this window (Section 6.1) and has not yet been scored on dates not used to choose it; the retrain's window supplies such dates only if they all fall after 8 October (Section 7.2). When the served number says *Packed* it is right far more often than the configuration it replaced (67.4% against 28.0%), but catches slightly fewer Packed hours (45.3% against 49.2%, on 20,445 readings).

- **The delta model's data and holdout.** Section 6.2 rests on spring rows of unknown provenance, and its three held-out cities are also the project's validation set. On verified live readings the spring-trained model loses to the weekly pattern: MAE 30.94 as served (display map on) against 26.40 for popular times over the 30-day test, and 27.52 against the curve's 26.42 for the model without the display map or offset, on 7,920 readings in the 8 September export. Ten of its 106 features are never used (Section 5.1), and latitude, dropped as a feature, re-enters through daylight hours. The learner comparison of Table 12 is untuned and confined to three cities.
- **A silent feed.** Since 12 October the provider has returned almost no live readings. Over 28 replayed outage start dates, losing the hour-old reading costs 20.7 points of exact level at once, then about 0.4 points a week for three weeks; an offset kept for a month would still beat the bare curve by 3.6 points, but each reading expires after 28 days, and production then falls back to the curve. Until a fix deployed on 15 October, the serving path also dropped offsets whose stored row was over seven days old.
- **Owl in the field.** Owl has not been measured in a working venue (Section 21), and its six real sets also served as development benchmarks (Section 13). The first sealed test (Section 14) is one or two people against a plain wall, from the lab whose side-view set Owl trained on, and its counts, like the IISER recounts, came from an AI model reading the frames, not from counting in the room. Owl transfers unevenly to cameras it was not trained on. The rule beat it on 32×24 and 32×32 arrays (sealed tests 2, 3, 5, 12 and 14; on test 12 Owl led when given the sensor's real temperatures), in BU-TIV's atrium of lone, distant people, in a room of walkers (test 11) and on pooled frames of three or more people (tests 5 and 6), and dense or distant groups defeated both counters (tests 4, 7, 8 and 10). Outdoors Owl counted a person in most frames showing only a dog. The hardest cases remain people hidden by people (a side view of three to five people is exact on 40% of frames, 85.5% within one), lying and sitting people on the Lepton (exact on 81% and 83% of SDT frames, which hold one person or none) and sparse classrooms, where Owl counts some warm seat backs.
- **No Flux is installed at a venue.** The one unit is a demonstration unit, Flock's crowd number reads no sensor data, and the doorway counter has run on the bench alone.

21 What comes next: the retrain, and joining the forecast to the counter

This section sets out what would strengthen each answer: the retrain for the first, a recording on Flux's own camera for the second, and counts from a working venue for both. None has been run yet.

21.1 The November retrain

Section 7.2 sets out the forecast's retrain. Owl retrains at the same time. Sealed tests 2 to 6 and 8 to 11 have been scored and their failures studied, so they are no longer sealed. Each is split by whole

recording into a part that trains Owl and a part that only evaluates it, with near copies across the split removed from training. UPPET (test 7) and TIDOS (test 14) are never trained on, and BU-TIV only if its authors grant training permission. The split was fixed on 15 October, before any training, and extended on 19 and 20 October with MIDetection's training and validation rooms and held-out test clips, and with OpenThermalAction's for the candidate allowed copyleft labels (184,576 frames, the earlier rows unchanged, each version hashed); the freeze date will be reported with the results. For the November retrain, the synthetic frames can add the heat a body leaves on a seat, cushion or table edge, shaped like the contact, starting 60 to 90% of the way from room to skin temperature and fading over minutes from the rim inward. It was off for owl-4.9 and for every result in this paper. Sealed test 13's errors motivated that change, so it cannot show an unbiased gain from it; the gain will be judged on validation data and the Flux recording. On validation parts carved from the new sets' training parts, owl-4.9 is exact on 31.5% of frames (2.5% with three or more people) against 83.5% on its own validation parts, rescored the same way: the gap the new data must close. The untouched tests, scored once after the freeze, are the Flux recording (Section 21.2), the labelled validation split of LTDv2 (a newer, larger release of the LTD harbour set [40]) and 215 thermal frames of LITIV 2012 [42], a set without per-person labels, whose model-assisted counts come from Section 14's blind two-pass procedure (95.3% agreement; not yet checked by people) and on which neither Owl nor the rule has run. Owl's gate was fixed on 15 October, before any training: on the pooled evaluation parts of tests 2 to 4 the new model must beat owl-4.9 on exact count, with a block-bootstrap interval above zero; it must gain at least 5 points on frames of three or more people; and it may lose no more than 2 points on any of the six benchmark sets or sealed test 1. I will add both results here, whichever way each gate decides.

21.2 A harder sealed test for Owl

The next sealed test will be recorded on Flux's own camera, with consent, in a real room of one to fifteen people counted as it is filmed, and kept unseen until the model it grades is final; sparse rooms, where Owl counts warm seat backs (Section 14.3), and dense groups are what it must answer first.

21.3 Counts as ground truth

The forecast predicts busyness normalised per venue, so that 100 is the busiest that venue gets (Section 4.1). A score of 70 could mean a dozen people in a small bar or two hundred in a large one. A Flux unit in a venue that Flock already reads would give that venue an absolute scale: its headcount, paired with the busyness provider's live score for the same hours, would show how many people a given score means at that venue, and whether the five crowd levels match what a group sees at the door. Its doorway counter, once proven on a real door, would add arrivals and departures, so the sensor could estimate how many are inside as well as in view (Section 17).

It would also give the served number a fresher reading: its strength is the latest reading, whose value sits in the first hour (Sections 3.2 and 6.3, Figure 8), and a sensor reporting every 30 seconds makes that reading seconds old, where readings now arrive at most hourly (Section 4.2). On real counts like Flux's, Section 6.8's replication favours adding the latest departure to the next hour's curve.

21.4 What would be measured first

- One unit, installed with the owner's agreement, at a venue the hourly collector already reads.
- Hand counts of the people in Owl's view on several busy nights, to score Owl in that room before its counts serve as

ground truth, and of arrivals and departures to score the doorway counter.

- Hour by hour over several weeks: the busyness provider's live score against Flux's count, the score per person, and how often the crowd level matches what the count shows.

22 Conclusion

Can AI tell how crowded a venue will be before you go? Yes for the next hour, and only slightly further ahead. What does it tell: each venue's weekly curve corrected by its own recent live readings, which named the provider's crowd level one hour ahead 54.9% of the time, against 30.2% for the popular-times chart and 30.8% for the larger gradient-boosted model (Section 8). Those are the provider's estimates, not head counts. Can a credit-card-sized computer model be designed to count the number of people inside a venue using a thermal camera without identifying individuals? Yes in simple rooms, not yet in crowds, and not yet inside a working venue: on one Raspberry Pi core, sending only counts, Owl was exact on 94.3% of a sealed office set against the heat rule's 40.7%, and dense groups still defeat it (Section 18). The November retrain and a first recording in a venue (Section 21) will test both answers again.

Acknowledgements

Owl would not count real rooms without the people who shared their data. The Computer Vision Lab at TU Wien (Christopher Pramerdorfer, Julian Strohmayr and Martin Kampel) granted permission to use the SDT dataset, real and rendered. Arijit Samal and Haroon R. Lone of IISER Bhopal licensed the Thermal Vision dataset and granted permission to use it. PUT Thermo Presence (PUT Vision Lab) and OpenThermalPose2 (ISSAI) are used under the MIT licence; the AAU VAP Trimodal dataset (Aalborg University) and the SenSys 2021 low-resolution set are used under Creative Commons Attribution 4.0. The second sealed test uses the SKKU MLX90640 set under the MIT licence and Thermal Dogs and People under Creative Commons Attribution 4.0.

Data sources. The public counts of Section 6.8 come from Brisbane City Council; the City of Melbourne Pedestrian Counting System; Dublin City Council and the National Transport Authority; Cork City Council; Dún Laoghaire-Rathdown County Council (via Smart Dublin); Leeds City Council (Data Mill North) and the City of York Council (York Open Data), both containing public sector information licensed under the Open Government Licence v3.0; the City of Toronto (contains information licensed under the Open Government Licence – Toronto); Heart of the City, Auckland; Kanton Basel-Stadt; Stadt Zürich (Open Data Zürich: the Bahnhofstrasse counts, measured by hystreet.com, and, under CC0, the Tiefbauamt's pedestrian counters and the Verkehrsbetriebe Zürich counts at Hardbrücke); the Cambridge Water Department; the Seattle Department of Transportation; the Metropolitan Transportation Authority (data.ny.gov); Crowdedness at the Campus Gym by Nick Rose (Kaggle, ODbL v1.0); and the KIT Library, Karlsruhe Institute of Technology, and the University of St. Gallen Library, whose Seatfinder occupancy data are used with their permission. The open series of the sixth check come from Aarhus Kommune (opendata.dk, CC BY 4.0); Boryeong City (data.go.kr 15134612; no use restrictions); the University of Alberta Library (Borealis, doi:10.5683/SP3/TCY5XY, CC0); Saga City (BODIK, CC BY 4.0); Orléans Métropole (data.orleans-metropole.fr, Licence Ouverte v2.0, data of 15 October 2026); Montpellier Méditerranée Métropole (data.montpellier3m.fr, ODbL); and the Comune di Bologna (BolognaWifi Affollamento, CC BY 4.0). The room data come from ROBOD (Tekler, Ono, Peng, Zhan, Lasternas and Chong, National University of Singapore; figshare, CC BY 4.0). The campus Wi-Fi data come from Universitat Politècnica de Catalunya, Space and Time User Distribution in a University Campus (Mendeley Data, CC BY 4.0). The ceiling-sensor frames of the third sealed test come from Klir et al., Multi-Person Localization based on a Thermopile Array Sensor with Machine Learning and a Generative Data Model (GitHub, GPL-3.0). The fourth sealed test uses BU-TIV (Wu, Fuller, Theriault and Betke, Boston University), evaluated with its authors' permission, and the Thermal soccer dataset (Gade and Moeslund, Aalborg University; Kaggle, CC BY-SA 4.0). The fifth sealed test uses twelve sets from Roboflow Universe, each under CC BY 4.0 or a public-domain dedication and credited to its publisher: Thermal Human Detection; yolo-thermal-pedestrian-indoor; CIIS Thermal; Outcomex Thermal Person Detection; drone-work Thermal People; football-3arjv

Thermal Human 4; mlx90640; thermalobjectdetection MLX90640 Camera Person Detection; Mondragon Unibertsitatea Thermal Camera Person Detection; pnusafety.net Thermal Human Detection; thermla-mlx90640; and toinonen Thermal Human; with IISER Bhopal's Occupancy Estimation (MIT) reported apart. The sixth uses the VILNIUS TECH thermal pose dataset (Vdovjak and Sledevic; Mendeley Data, CC BY 4.0), TPD (Dhillon, Chaturvedi, Sharma and Kumar, Bennett University; Zenodo, CC BY 4.0), Poudel's thermopile occupancy set (Zenodo, CC BY 4.0) and FLC-Rural (Dhillon, Sharma and Kumar; Zenodo, CC BY 4.0). The seventh to fifteenth use UPPE's TPE images (Cormier, Specker and Beyerer, Fraunhofer IOSB; CC BY-NC-SA 3.0, by permission), Thermal People Cafe (uon101) and Thermal Occupancy Counter (QuietBites Project) from Roboflow Universe, WLRI-HRC (Süme, Ponomarjova, Wendt and Rupitsch, University of Offenburg; Zenodo), the Long-term Thermal Drift dataset (Nikolov et al., Aalborg University; Kaggle), the LITIV 2014 thermal-visible dataset (Bilodeau, Torabi and colleagues, Polytechnique Montréal; by permission), MIDetection (Vandersteegen and colleagues, KU Leuven EAVISE with Melexis; by permission, with training on its training rooms allowed from 19 October), Thermal-IM (Tang, Ye, Ma and Zhao, Tsinghua University and MIT; BSD-3-Clause), TIDOS (Cokbas, Ishwar and Konrad, Boston University; by permission for research, never trained on, as commercial use of a model trained on it needs a Boston University licence) and the Multi-View Visible-Thermal set (Lupián and colleagues, Universidad de Almería; Mendeley Data, CC BY-NC 3.0, by permission); the untouched November test uses LITIV 2012 from Polytechnique (by permission). Except where noted, they are used under CC BY 4.0 or a public-domain dedication. Venue busyness readings come from a commercial venue-busyness provider, which has no part in and does not endorse this work. Weather data comes from OpenWeather (<https://openweathermap.org/>) under the Open Database License (ODbL). Sports schedules come from TheSportsDB and event listings from Ticketmaster. No dataset named here is released with this paper.

Every number behind Section 6.8 and Owl's tests is kept in Flock's database, in six research tables apart from the app's own: the 6,497,751 hourly public counts with each source's licence and the attribution it requires; 20,083 frame-level count records (9,242 frames across fifteen rounds of sealed tests, 483 frames of the separately reported IISER occupancy set and 434 IISER recounts, each with Owl's and the rule's count, and owl-4.9's count on 9,924 validation frames from 27 sets), beside the owner, licence and terms of each of the 53 thermal datasets; the 184,576-frame November split (Section 21.1); and 1,031 result rows, each with the file that computes it.

References

- [1] J. Bollenbach, S. Neubig, A. Hein, R. Keller, and H. Krcmar. 2022. Using machine learning to predict POI occupancy to reduce overcrowding. In *INFORMATIK 2022* (LNI P-326). Gesellschaft für Informatik, Bonn, 393–408. https://doi.org/10.18420/inf2022_34
- [2] J. Bollenbach, S. Neubig, A. Hein, R. Keller, and H. Krcmar. 2024. Enabling active visitor management: local, short-term occupancy prediction at a touristic point of interest. *Information Technology & Tourism* 26, 3 (2024), 521–552. <https://doi.org/10.1007/s40558-024-00291-2>
- [3] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589, 7840 (2021), 82–87. <https://doi.org/10.1038/s41586-020-2923-3>
- [4] T. Chen and C. Guestrin. 2016. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [5] J. B. Copas. 1983. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Methodological)* 45, 3 (1983), 311–335. <https://doi.org/10.1111/j.2517-6161.1983.tb01258.x>
- [6] F. Galton. 1886. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886), 246–263. <https://doi.org/10.2307/2841583>
- [7] A. Hajisafi, H. Lin, S. Shaham, H. Hu, M. D. Siampou, Y.-Y. Chiang, and C. Shahabi. 2023. Learning dynamic graphs from all contextual information for accurate point-of-interest visit forecasting. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23)*. ACM, 1–12. <https://doi.org/10.1145/3589132.3625567>
- [8] R. J. Hyndman and A. B. Koehler. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* 22, 4 (2006), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- [9] S. Islam, D. Gandhi, J. Elarde, T. Anderson, A. Roess, T. F. Leslie, H. Kavak, and A. Züfle. 2021. Spatiotemporal prediction of foot traffic. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising (LocalRec '21)*. ACM, 1–8. <https://doi.org/10.1145/3486183.3490997>
- [10] S. Kapoor and A. Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4, 9 (2023), 100804. <https://doi.org/10.1016/j.patter.2023.100804>

- [11] M. Möhring, B. Keller, R. Schmidt, and S. Dacko. 2021. Google Popular Times: towards a better understanding of tourist customer patronage behavior. *Tourism Review* 76, 3 (2021), 533–569. <https://doi.org/10.1108/TR-10-2018-0152>
- [12] D. S. Moore, W. I. Notz, and M. A. Fligner. 2013. *The Basic Practice of Statistics* (6th ed.). W. H. Freeman and Company, New York.
- [13] ONNX Runtime developers. ONNX Runtime. Software, accessed October 2026. <https://onnxruntime.ai/>. Server: onnxruntime-node 1.29.0; sensor: the Debian python3-onnxruntime package, version 1.21.
- [14] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and H. Larochelle. 2021. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research* 22, 164 (2021), 1–20.
- [15] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillaere-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 8 (2017), 913–929. <https://doi.org/10.1111/ecog.02881>
- [16] S. Timokhin, M. Sadrani, and C. Antoniou. 2020. Predicting venue popularity using crowd-sourced and passive sensor data. *Smart Cities* 3, 3 (2020), 818–841. <https://doi.org/10.3390/smartcities3030042>
- [17] X. Zhou, D. Wang, and P. Krähenbühl. 2019. Objects as points. arXiv:1904.07850. <https://doi.org/10.48550/arXiv.1904.07850>
- [18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV ’17)*. IEEE, 2999–3007. <https://doi.org/10.1109/ICCV.2017.324>
- [19] V. Lempitsky and A. Zisserman. 2010. Learning to count objects in images. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*. 1324–1332.
- [20] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’16)*. IEEE, 589–597. <https://doi.org/10.1109/CVPR.2016.70>
- [21] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 23–30. <https://doi.org/10.1109/IROS.2017.8202133>
- [22] C. Prämendorfer, J. Strohmayer, and M. Kampel. 2020. SDT: a synthetic multi-modal dataset for person detection and pose classification. In *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1611–1615. <https://doi.org/10.1109/ICIP40778.2020.9191284>
- [23] A. Samal and H. R. Lone. 2025. Thermal vision: pioneering non-invasive temperature tracking in congested spaces. *Smart Health* 36 (2025), 100576. <https://doi.org/10.1016/j.smhl.2025.100576>. Dataset: <https://osf.io/9z5a4/>
- [24] C. Palmero, A. Clapés, C. Bahnsen, A. Mögelmoose, T. B. Moeslund, and S. Escalera. 2016. Multi-modal RGB–depth–thermal human body segmentation. *International Journal of Computer Vision* 118, 2 (2016), 217–239. <https://doi.org/10.1007/s11263-016-0901-x>. Dataset: <https://vap.aau.dk/>
- [25] P. Aszkowski and M. Piechocki. 2022. Thermo Presence: the low-resolution thermal image dataset and occupancy detection using edge devices. In *Proceedings of the 3rd Polish Conference on Artificial Intelligence (PP-RAI 2022)*. Publishing House of Gdynia Maritime University, Gdynia, 49–52. Dataset: <https://github.com/PUTvision/thermo-presence>
- [26] A. Kuzdeuov, M. Zakaryanov, A. Tleuliyev, and H. A. Varol. 2025. OpenThermalPose2: extending the open-source annotated thermal human pose dataset with more data, subjects, and poses. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 7, 4 (2025), 902–913. <https://doi.org/10.1109/TBIOM.2025.3575499>. Dataset: <https://huggingface.co/datasets/issai/OpenThermalPose2>
- [27] S. Zhu, T. Voigt, D. F. Perez-Ramirez, and J. Eriksson. 2021. A low-resolution infrared thermal dataset and potential privacy-preserving applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys ’21)*. ACM, 552–555. <https://doi.org/10.1145/3485730.3493692>. Dataset: <https://doi.org/10.5281/zenodo.5574233>
- [28] J. D. C. Little. 1961. A proof for the queuing formula: $L = \lambda W$. *Operations Research* 9, 3 (1961), 383–387. <https://doi.org/10.1287/opre.9.3.383>
- [29] Teledyne FLIR. FLIR Lepton 3.5 radiometric thermal camera module. Product documentation, accessed October 2026.
- [30] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’16)*. IEEE, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [31] F. Yu and V. Koltun. 2016. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR 2016)*. arXiv:1511.07122.
- [32] H. Law and J. Deng. 2018. CornerNet: detecting objects as paired keypoints. In *Computer Vision – ECCV 2018 (Lecture Notes in Computer Science)*. Springer, Cham, 765–781. https://doi.org/10.1007/978-3-030-01264-9_45
- [33] Y. Li, X. Zhang, and D. Chen. 2018. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR ’18)*. IEEE, 1091–1100. <https://doi.org/10.1109/CVPR.2018.00120>
- [34] M. Piechocki, M. Kraft, T. Pajchrowski, P. Aszkowski, and D. Pieczyński. 2022. Efficient people counting in thermal images: the benchmark of resource-constrained hardware. *IEEE Access* 10 (2022), 124835–124847. <https://doi.org/10.1109/ACCESS.2022.3225233>
- [35] Z. Wu, N. Fuller, D. Theriault, and M. Betke. 2014. A thermal infrared video benchmark for visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW ’14)*. IEEE, 201–208. <https://doi.org/10.1109/CVPRW.2014.39>
- [36] R. Gade and T. B. Moeslund. 2018. Constrained multi-target tracking for team sports activities. *IPSJ Transactions on Computer Vision and Applications* 10 (2018), 2. <https://doi.org/10.1186/s41074-017-0038-z>
- [37] M. Cormier, A. Specker, and J. Beyerer. 2025. UPPET: unified pedestrian pose estimation in thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW ’25)*. IEEE, 4551–4560. <https://doi.org/10.1109/CVPRW67362.2025.00440>
- [38] S. Süme, K.-M. Ponomareva, T. M. Wendt, and S. J. Rupitsch. 2025. Enhancing human–robot collaboration with thermal images and deep neural networks: the unique thermal industrial dataset WLRI-HRC and evaluation of convolutional neural networks. *Journal of Sensors and Sensor Systems* 14, 1 (2025), 37–46. <https://doi.org/10.5194/jsss-14-37-2025>
- [39] D. Läufer, S. Braun, S. Süme, and U. Himmelsbach. 2024. Comparison of approaches for human detection with low-resolution infrared data sets using deep learning. In *2024 21st International Conference on Ubiquitous Robots (UR)*. IEEE, 596–602. <https://doi.org/10.1109/UR61395.2024.10597494>
- [40] I. Nikolov, M. P. Philipsen, J. Liu, J. V. Dueholm, A. S. Johansen, K. Nasrollahi, and T. B. Moeslund. 2021. Seasons in drift: a long term thermal imaging dataset for studying concept drift. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1 (NeurIPS Datasets and Benchmarks 2021).
- [41] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi. 2014. Thermal–visible registration of human silhouettes: a similarity measure performance evaluation. *Infrared Physics & Technology* 64 (2014), 79–86. <https://doi.org/10.1016/j.infrared.2014.02.005>
- [42] A. Torabi, G. Massé, and G.-A. Bilodeau. 2012. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding* 116, 2 (2012), 210–221. <https://doi.org/10.1016/j.cviu.2011.10.006>
- [43] M. Vandersteegen, W. Reusen, K. Van Beeck, and T. Goedemé. 2023. Person detection using an ultra low-resolution thermal imager on a low-cost MCU. In *Image and Vision Computing (IVCNZ 2022)*. Springer, 33–47. https://doi.org/10.1007/978-3-031-25825-1_3
- [44] Z. Tang, W. Ye, W.-C. Ma, and H. Zhao. 2023. What happened 3 seconds ago? Inferring the past with thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR ’23)*. IEEE. Thermal-IM dataset, BSD-3-Clause.
- [45] M. Cokbas, P. Ishwar, and J. Konrad. 2020. Low-resolution overhead thermal tripwire for occupancy estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW ’20)*. IEEE. <https://doi.org/10.1109/CVPRW50498.2020.00052>. TIDOS dataset, by permission.
- [46] M. Lupión, A. Polo-Rodríguez, J. Medina-Quero, J. F. Sanjuan, and P. M. Ortigosa. 2024. 3D human pose estimation from multi-view thermal vision sensors. *Information Fusion* 104 (2024), 102154. <https://doi.org/10.1016/j.inffus.2023.102154>. Dataset: <https://doi.org/10.17632/zsp6hgvcz.1>, by permission.

Appendix A: Choices made after seeing data

H1 to H3 followed Section 3’s findings and several Section 6 analyses; Section 7.2 is the first test with rules fixed in advance, though its comparator, production as served, was simplified on 13 October. Section 3.2’s measurements (1 to 8 September) overlap the three test days and informed the nowcast (weights fitted on 1 to 5 September). Section 6.1’s 30-day window was scored after the three-day result was known. The display map, offset weight and original nowcast weights were chosen on earlier rows; the serving clamp and one-point push were fitted on the first 30% of the held-out cities’ dates, inside the gate slice, and those cities gated every version from v2.2 on (Section 6.2). Of the “no event nearby” rows (56.1% of the corpus), 59.5% were defaults correlated with city, so Section 6.5 disables the events lookup for both arms. The persistence baselines, two-way bootstrap, age-mix standardisation, changed-value readings and test-window stickiness were computed on 13 October, after this paper’s first version; rule (b), identified on the 30-day window and served since, has no out-of-sample figure yet. The replication’s curve-plus-latest-departure predictor had no pass rule. Owl’s sealed tests: test 4’s soccer sample grew from 98 to 393 crops after a frame-numbering fix, and test 5’s bootstrap blocks were regrouped after first estimates (intervals only); every other plan change (tests 6 to 9, 11, 12 and LITIV 2012 after its count) preceded any prediction, and crowd-size comparisons were planned from test 7 on.