The Pulse Is Not the People: A Code-Level Audit of Bluetooth Footfall Sensing and a Privacy-First Architecture for Urban Decisions

Author: Noel Sabia (ORCID 0009-0008-3064-7573)
Affiliation: —
Date: 14 November 2026
Companion artifact: `paper/main.md` (4–5 page hackathon report) · https://pulse.noelsabia.dev · code to be released after the hackathon

Abstract

Low-cost Bluetooth footfall counters are increasingly used to justify staffing, advertising, and urban-planning decisions. Their central claim is that they measure pedestrian density while preserving privacy. We audit that claim on the eight-sensor PAX counter network deployed in Heilbronn, Germany. A code-level review of the public 2026 firmware (commit e8be8bf) shows a **~10-second** passive BLE scan: the node counts distinct locally administered addresses, truncates the last two MAC bytes into a 16-bit ID, multiplies by a fixed factor (default 0.7), and sleeps for 900 seconds. The effective duty cycle is about 1.1% in that code. The deployed 2025 binary may differ; the data confirm the 0.7 scaling and per-wake export, not the scan duration or the 16-bit truncation. In the **12 June model window** (83,156 values), all values are exact multiples of 0.7; reconstructed counts have median 12 and maximum 418. Coverage in the **18 June coverage panel** (82,170 readings) is sparse: only 28.2% of 15-minute slots contain all eight sensors, 5.18% of all 96,232 density rows are zero, and, in the **12 June model window**, cross-sensor correlation falls from 0.58 to 0.38 after removing hour-of-week structure. Legally, MAC addresses are very likely personal data in this context; truncation does not anonymise — it produces a weak, transient identifier that may be pseudonymous in context; and passive sniffing under § 25 TDDDG is contested. We then propose a privacy-first architecture: thermal/radar/camera/probe/manual sensing with calibration, a network-aware hierarchical spatial model, a quantile forecast with conformal calibration, and a decision layer evaluated by regret rather than MAE alone. We show how the same decision-grade contract transfers directly to staffing, as an analytical extension to advertising, conditionally to waste routing, as a sibling to sensor placement and LoRaWAN planning, and not at all to sentiment or learning-guide tasks. The contribution is not a better people counter; it is an honest measurement model and a staged path to one.

1. Introduction

A store manager on Kaiserstraße knows next week's weather forecast and the staff roster. She does not know how many people will walk past next Saturday. The available data are counts from eight passive Bluetooth sensors; the decision is a shift plan. A dashboard that shows history does not close that gap. The manager needs a forecast, an explicit staffing rule, and a way to test a scenario before committing.

The Schwarz Digits
Pulse of the City

challenge asks teams to turn "privacy-preserving" pedestrian counts from Heilbronn into predictive tooling for advertising, store locations, and staffing [1]. The wording is doing a lot of work. "Pedestrian counts" implies people; "privacy-preserving" implies that the sensing layer has been designed around a privacy guarantee. Neither claim is self-evident. The counters are LoRaWAN-connected ESP32 devices that passively sniff Bluetooth Low Energy (BLE) advertisements and transmit aggregated device counts [2]. The challenge brief calls them privacy-preserving because they count devices rather than identify people [1]. That is a meaningful architectural choice, but it is not a measurement guarantee, and it is not a legal anonymisation guarantee.

This paper started as an attempt to write the technical report for a hackathon system, PulseCast, that turns the counts into a weekly staffing plan. The system works: on a 28-day holdout, a global weather-aware gradient-boosted model reaches MAE 3.32, WAPE 27.5%, and R^2 0.777 against a seasonal-naive baseline of 5.40, 44.7%, and 0.462. The plan is generated in 19 ms cold on v1.1 (contract: under two seconds) on a three-service deployment on STACKIT SKE, and the deployed system handled 120/120 concurrent SSE clients with zero connection errors and a p95 event latency of 331 ms. But the more we looked at the sensing layer, the clearer it became that the model was predicting a **different quantity** from the one the narrative claimed. The code and the data agree: each value is a short-window count of randomized BLE addresses scaled by a fixed factor. The "15-minute" cadence is the sleep cycle, not the integration window.

That finding changes the paper. A conventional technical report would present the model metrics and the deployment, note that the data are "estimated devices, not people," and stop. That is not enough. If a decision support system is built on a measurement that conflates flow and dwell, cannot distinguish a phone from a watch or a car, and samples 10 seconds out of every 15 minutes, then the honest contribution is a **measurement audit** plus a **reference architecture** for the next version. The model is not the hard part. The data contract is.

We make five contributions:

1. **A code-level audit** of the public 2026 Code for Heilbronn `lorax` firmware (commit e8be8bf), with the measurement chain: passive BLE scan, locally-administered-address filter, 16-bit truncation, per-sensor factor 0.7, 900-second sleep, RSSI threshold −80 dBm. The deployed 2025 binary may differ; the data confirm the 0.7 scaling and per-wake export, not the scan duration or truncation.
2. **A forensic data analysis** of 192,464 rows, confirming the 0.7 scaling in the 12 June model window (83,156 values, 100% multiples of 0.7), identifying the `density-01` calibration break, and quantifying coverage, zero-inflation, correlation, and missingness with the population stated for each statistic.
3. **A privacy and legal analysis**: why MAC addresses are likely personal data, why truncation does not anonymise and produces a weak, transient identifier that may be pseudonymous in context, why § 25 TDDDG / Art. 5(3) ePrivacy is the decisive unresolved question, and what a DPIA-compliant deployment would require.
4. **A privacy-first reference architecture**: a tiered sensing stack (thermal, radar, edge cameras, probe data, manual counts), a calibration program, a central-DP aggregation layer, a network-aware hierarchical spatial model, a quantile forecast with conformal calibration, and a decision layer evaluated by regret.
5. **A transferability matrix** across the six hackathon challenges, with explicit boundary conditions: direct for staffing, analytical for advertising, conditional for waste routing, sibling for sensor placement and LoRaWAN planning, and not applicable to cable-car sentiment or the learning guide.

We are deliberately conservative. We do not claim that the current network measures people, that the 0.7 factor is a validated penetration correction, that eight sensors support a city-wide heatmap, or that the staffing plan has proven ROI. We claim that the current system is a useful **relative activity index** for a local cluster, that its measurement semantics can be reconstructed from code and data, and that a privacy-first, calibrated version is feasible with a staged program.

The paper is organized as follows. Section 2 audits the measurement. Section 3 analyzes the privacy and legal problem. Section 4 proposes a privacy-first sensing stack. Section 5 develops the sparse-to-heatmap-to-decision pipeline. Section 6 tests how far the architecture generalizes across the six challenges. Section 7 sets out an evaluation agenda. Section 8 discusses limitations and what to keep from the hackathon build. Section 9 positions the work in related literature. Section 10 concludes.

2. The measurement problem

2.1 What the challenge claims

The challenge brief describes the data as “numbers of people measured in a given area,” collected by “privacy-preserving sensors” that “estimate crowd density by passively sniffing Bluetooth signals (MAC address advertisements) from nearby smartphones and smartwatches” and transmit “aggregated device counts” over LoRaWAN [1]. It also warns that PAX counters provide estimates because not everyone has a smartphone with Bluetooth turned on [1]. The brief is more careful than the headline: it distinguishes counts from people and asks teams to normalize. But the downstream narrative—store traffic, advertising exposure, staffing demand—treats the index as a footfall measure. That is the gap this section closes.

2.2 The code-level audit

The system is maintained by Code for Heilbronn. The public repository `cfhn/lorax` (mirror name `lora-frequenzmessung`) contains the `sensor-pax` firmware, a LoRaWAN/TTN aggregator, a Go backend, TimescaleDB, and a Grafana dashboard [3]. The firmware is derived from the open-source ESP32-Paxcounter / `libpax` lineage [4]. The audit below is of the public 2026 commit **e8be8bf**; the deployed 2025 binary may differ, and the data confirm the 0.7 scaling and per-wake export, not the 10-second scan or the 16-bit truncation.

One cycle is:

```
wake from deep sleep
  → BLE passive scan (window 80 ms, interval 80 ms)
  → count distinct locally administered addresses in ~10 s
  → truncate the last two address bytes to a 16-bit ID
  → deduplicate within that scan only
  → multiply the integer count by factor (default 0.7)
  → send one LoRaWAN payload: sensor_id, 0, value, 1, battery
  → deep sleep 900 s
```

The key code facts are:

- `configuration.blecounter = 1`, `configuration.wificounter = 0` — BLE only [3].
- `configuration.blescantime = 30`, but the vendored `start_BLE_scan()` does not use `blescantime`; the main loop stops on the first 10-second report callback. The effective scan is about **10 seconds**, not 30 [3, 4].
- `libpax_counter_init(log, &count, 10, 1)` — 10-second report timer, cumulative count within the wake [3].
- `mac_add()` accepts only addresses with the **locally administered bit** set; it ignores universal/static addresses [3, 4].
- `id = (uint16_t *)(paddr + 4)` — the “hash” is a **16-bit truncation** of the last two bytes, not a cryptographic hash [3].
- `paxCount * cfg.factor` with `factor` default **0.7**; `sleep_time_sec` default **900 s** [3].
- RSSI threshold **−80 dBm**; BLE advertising queue size **30**; queue drops are logged, not counted in the payload [3].
- The backend stores the parsed value directly; the CSV is a raw export, not a 15-minute mean [3].

The practical consequence is the formula:

```
published_value ≈ 0.7 × U
U = distinct 16-bit IDs from locally administered BLE addresses seen in a ~10 s scan
sampling interval = ~900 s
duty cycle ≈ 10 / 910 ≈ 1.1%
```

This is not a 15-minute count, a 15-minute mean, or a count of people. It is a short-window count of a particular address regime, scaled by a provisioning constant.

2.3 A measurement model

Let `P` be the number of people present in the sensor’s catchment during the scan window, `d` the probability that a person carries a detectable advertising device, `m` the mean number of distinct advertising devices per carrier, and `q` the probability that at least one device advertises within the scan. The expected published value is:

```
E[y] = f · d · m · q · P
```

with `f = 0.7`. The factor can only be a correct people-per-device conversion if `d · m · q ≈ 1.43`. That is possible if `d` is high and `m ≈ 1.4`, but it is not measured. If `d = 0.5` and `m = 1.4`, `y/P ≈ 0.49` — a 2× undercount. If `d = 0.9` and `m = 1.2`, `y/P ≈ 0.76` — a 24% undercount. The scale is arbitrary without ground truth, and a single global factor cannot track seasonal, demographic, or device-mix changes.

The sampling design adds a second bias. A person who passes in 30 seconds has roughly a 4.4% chance of overlapping a 10-second scan every 900 seconds. A person who lingers for an hour can be counted in up to four consecutive bins. The index therefore conflates **flow** (people passing a point) with **dwell** (people staying in the area). It is not a screenline crossing count and not an occupancy count.

2.4 Data forensics

The delivered CSV contains 192,464 rows: 96,232 density readings and 96,232 battery readings. There are 12 device IDs; after merging the two `density-01` IDs and excluding three test sensors, eight usable locations remain in a ~580 m × 280 m cluster: `density-01`, `hn-goetzentuem-01`, `hn-hagenbuchersee-01`, `hn-hagenbuchersee-02`, `hn-marahaus-01`, `hn-neckarufer-01`, `hn-soleo-01`, and `hn-stadtgalerie-01`. The **12 June model window** (12 June–7 November 2026) contains 83,156 density values; the **18 June coverage panel** (18 June–7 November 2026) contains 82,170 density readings and is used for coverage statistics.

Quantization. Every one of those 83,156 values in the 12 June model window is an exact multiple of 0.7. The reconstructed integer `U` has median 12, p95 71, and maximum 418; 2.70% of the 83,156 values reconstruct to `U = 0`. A continuous physical series would not land on a 0.7 grid in 83,156 consecutive observations; the only plausible explanation is `value = round(0.7 × integer, 4)`. This is direct, data-internal confirmation of the 0.7 scaling and per-wake export; it does not independently confirm the scan duration or the 16-bit truncation [5].

The calibration break. The legacy `density-01` device (`863f75b0`, 16 October 2025–12 March 2026) is 100% integer-valued on a 1.0 grid: 0–28, mean 3.008, median 2.0, zero share 29.8%. The replacement (`f8c19730`, 16 April–7 November 2026) is 100% on the 0.7 grid: 0–63, mean 9.741, median 7.7, zero share 5.9%. The gap between the last legacy reading and the first new reading is **35.2065 days**, with no overlap, so a conversion factor cannot be estimated empirically. The level shift is ×3.24 in the mean and ×3.85 in the median, which is not explained by 0.7 alone; season, installation, and sensitivity are confounded. The safe conclusion is that the two series are not on the same scale and must not be pooled [5].

Cadence. The nominal 15-minute grid is not a 15:00 wall-clock grid. Current devices report every 15.27–15.37 minutes (median); the legacy device every 15.18 minutes with ~18 s jitter; the new `density-01` every 16.14 minutes with ~2.4 s jitter, producing ~89 readings/day. Flooring to `HH:00/15/30/45` is an approximation; two devices matched to the same slot can be up to ~15 minutes apart [5].

Coverage. In the **18 June coverage panel (82,170 readings)**, only **28.2%** of 15-minute slots contain all eight sensors; the modal count is seven (4,117 slots). Aggregating to 30, 60, or 120 minutes raises complete eight-sensor coverage to **63.4%, 69.7%, and 71.0%**, respectively. Median monthly coverage is 86.1%, but August 2026 falls to 54.2% on average and 23.3% at the worst location. Six of eight locations show a simultaneous multi-week degradation; `density-01` and `hagenbuchersee-01` do not. Coverage is strongly confounded with battery across devices (r = 0.84, n = 8), but the cause cannot be isolated [5].

Distribution and correlation. Across all **96,232 density rows**, 5.18% are zero, 10.36% are `<1`, and 39.57% are `<5`; in the **12 June model window** the shares are 2.70%, 8.32%, and 35.86%. There are no negative or NaN values and no duplicate timestamps; 113 intervals are shorter than 60 seconds across the core devices, 7 of them in the eight-sensor panel, and they look like retries. Raw cross-location Pearson correlation in the 12 June model window averages 0.58 (range 0.13–0.90), but drops to 0.38 after removing location-specific hour-of-week means and 0.48 at the daily level. Correlation is not evidence of shared unique visitors [5].

What the aggregates cannot tell us. The CSV contains no MAC addresses, address type, RSSI, unique IDs, or ground truth. MAC-randomization rate, penetration rate, unique-device count, dwell time, repeat visits, detection radius, and false-positive/false-negative rates cannot be measured from these aggregates. The 0.7 grid is a proxy for firmware processing, not a direct measurement of randomization or penetration. Any claim about those quantities must come from a calibration study.

Figures. Figure 1 shows the eight sensor locations. Figure 2 shows the hourly/weekday pulse. Figure 3 shows the per-device value grids and the `density-01` scale break.

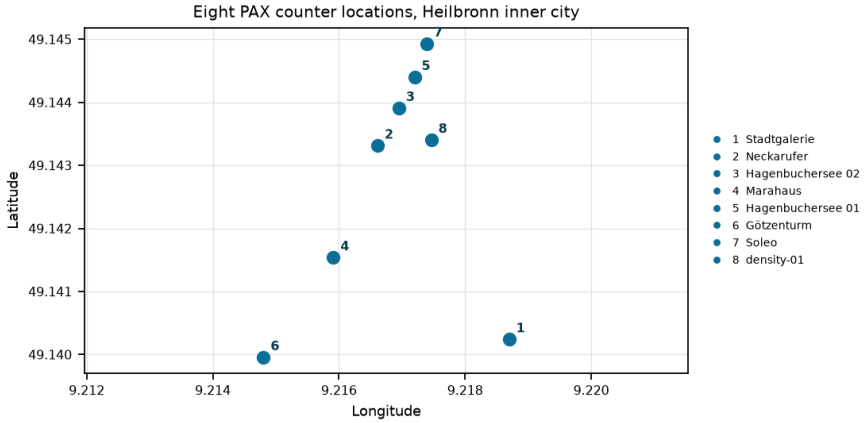


Figure 1. The eight sensor locations in the ~580 m × 280 m inner-city cluster. Measured points only; no city-wide surface.

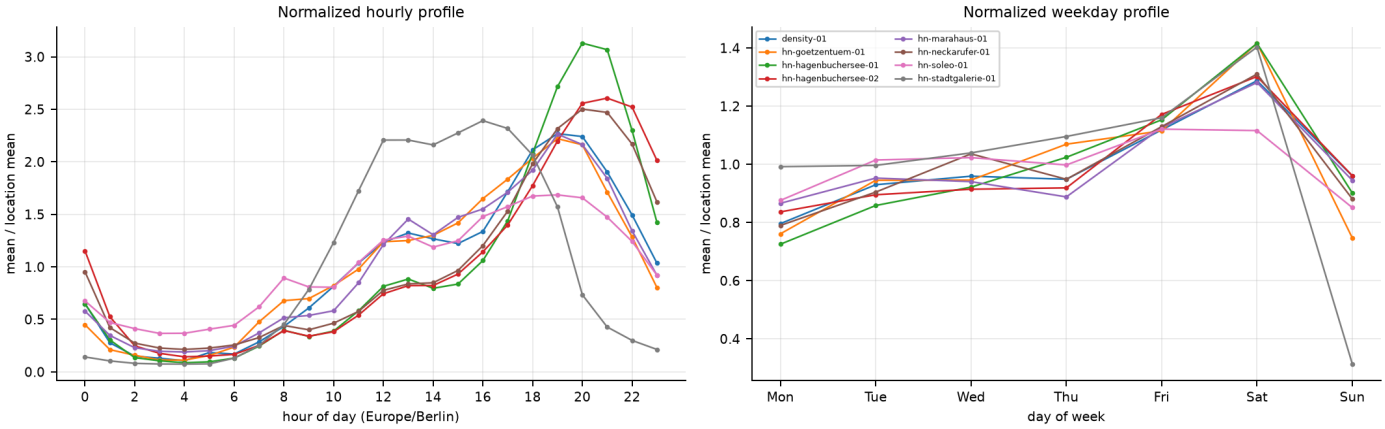


Figure 2. Mean index by hour and weekday across the eight sensors. The peak is 18:00–21:00 local; Saturday is the busiest day on average.

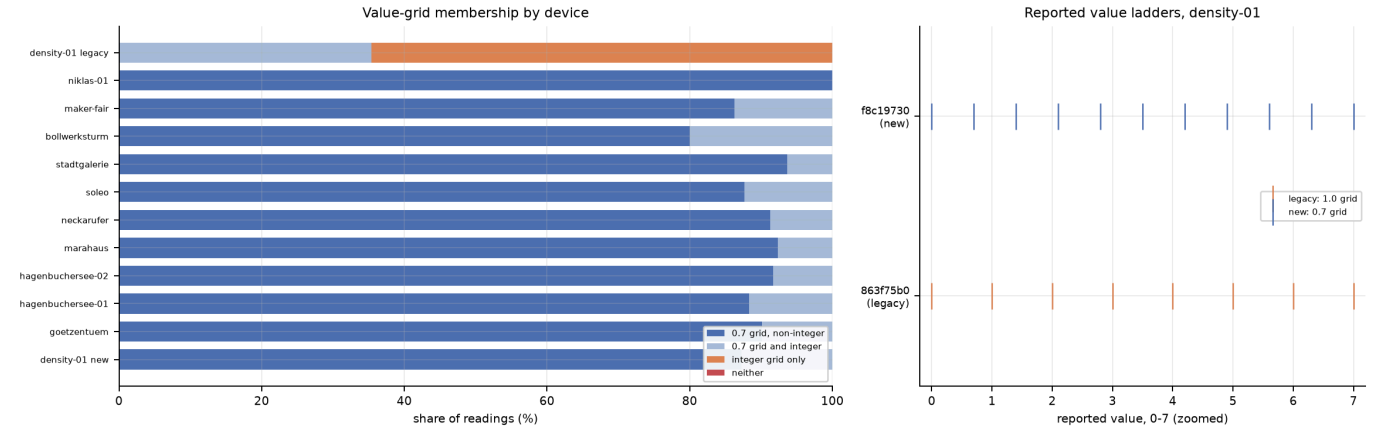


Figure 3. Left: per-device share of readings on the 0.7 grid, the integer grid, both, or neither. Right: the reported value ladders for the two `density-01` devices, showing the 1.0 versus 0.7 spacing.

2.5 Reliability taxonomy

The audit and the data suggest a taxonomy of failure modes, ordered by how much they matter for a decision:

Failure	Mechanism	Consequence
Short snapshot	~10 s scan every ~15 min	Misses peaks between scans; target is a sparse sample, not integrated flow
Flow/dwell conflation	Same person counted in consecutive bins; short passer-by often missed	A busy “passing” street and a “lingering” square can produce the same index
Randomized-only filter	Only locally administered addresses counted	Static/public-address devices invisible; non-phone random advertisers counted
16-bit truncation	Last two bytes used as ID	Collisions grow with density; undercount at busy times
No cross-scan dedup	Bitmap reset every wake	No dwell, repeat-visit, or cross-sensor dedup
Device ≠ person	Phones, watches, earbuds, cars, beacons	Unknown multiplier; one person can be several devices
Penetration unknown	Bluetooth on/off, discoverability, demographics	Absolute counts not identifiable without calibration
RSSI gate	−80 dBm threshold, body shielding, orientation	Detection zone is not a stable catchment
Queue drops	30-slot advertising queue	Undercount under density; not exposed in data
Gaps	LoRaWAN joins, power, maintenance	4–6 day gaps; missing not at random
Calibration factor	Default 0.7, undocumented	Single scalar cannot absorb penetration × devices/person × scan probability
Calibration break	Hardware replacement; 1.0 → 0.7 grid	Pooling the series fits the instrument, not the city

The first four are intrinsic to the design; the rest are deployment and calibration issues. A production system must address all of them, but the first four determine whether the metric can be interpreted at all.

2.6 What the existing model actually predicts

The PulseCast model is a global `HistGradientBoostingRegressor` trained on hourly means of the published index. On the 28-day holdout (n = 5,350), the production 7-day model reaches MAE 3.32, WAPE 27.5%, and R^2 0.777. The day-ahead weather model reaches 3.36, 27.8%, and 0.763. The seasonal-naive baseline is 5.40, 44.7%, and 0.462. Per-sensor MAE ranges from 2.26 (`density-01`) to 5.83 (`hn-stadtgalerie-01`). Weather is the main driver: removing it raises day-ahead MAE from 3.36 to 3.96; the holiday flag changes 3.96 to 3.84, a small effect.

These are credible relative results. They are also results **on the index**, not on people. The model is good at predicting a short-window count of randomized BLE addresses scaled by 0.7; that is a different achievement from predicting footfall. The MAE is in index units, and the index units are not people. A staffing rule that consumes the index inherits the scale ambiguity.

2.7 Why the audit matters

If the only use of the data were a relative “busier than usual” signal, the audit would be a caveat. But the PulseCast use case is staffing: the index is converted into required staff through a visitors-per-staff-hour assumption. That conversion is an absolute claim. If the index is 2× low because penetration is 0.5, the plan is 2× low. If the index conflates a passing street with a lingering square, the plan over-staffs the wrong hours. If the 0.7 factor is wrong, the plan is wrong by the same factor. The model’s MAE says nothing about these errors.

The right response is not to discard the data. It is to use them as a **relative activity index** for a local cluster, to calibrate them against ground truth before making absolute claims, and to state the measurement model explicitly. That is the bridge to the architecture in Sections 4–5.

3. The privacy problem

3.1 MAC addresses are likely personal data

The GDPR defines personal data as any information relating to an identified or identifiable natural person, including an online identifier, and asks whether identification is “reasonably likely” using all means available to the controller or another person (Art. 4(1), Recital 26) [13]. The CJEU’s *Breyer* judgment makes the test relative but broad: a dynamic IP address can be personal data where the controller has the legal means to obtain additional identifying information, and identification need not be in one person’s hands [14]. A MAC address is a device identifier; combined with sensor location, time, and deduplication, it can single out a device and its bearer, even if the controller does not know a name. The EDPB’s 2026 draft anonymisation guidelines use a hardware identifier such as a MAC address as an example of data that can identify a device and, in turn, an individual [17].

The practical conclusion is that a public-space footfall system should treat MAC addresses as personal data unless a documented anonymisation assessment proves otherwise. The burden is on the controller. “We hash, therefore it is anonymous” is not sufficient.

3.2 Hashing and truncation do not anonymise

Hashing or truncating a MAC address reduces direct identifiability, but it does not automatically anonymise. Truncation does not anonymise; it produces a weak, transient identifier that may be pseudonymous in context. An unsalted hash of a 48-bit MAC is enumerable; a short or truncated hash has a small space and frequent collisions; a keyed hash remains reversible by whoever holds the key. The EDPB’s pseudonymisation guidelines state that pseudonymised data that can be attributed to a person by additional information remain personal data, even if the pseudonymised data and the additional information are not in the same hands [16]. The French Conseil d’État’s *JCDecaux* decision is the clearest warning: truncating the MAC, adding a salt, and applying keyed hashing still left the controller able to identify and link records, and the purpose—measuring routes and repeat passages—was itself incompatible with anonymisation [18].

The Heilbronn firmware’s 16-bit truncation is weaker than a cryptographic hash. It is a dedup key with a 65,536-value space. That is a privacy-by-design choice in the sense that raw MACs are not persisted, but it is not an anonymisation guarantee. The operator’s blog says addresses are “hashed and stored for at most one hour” [2]; the public 2026 firmware resets the bitmap at each wake and has no one-hour store [3]. The two accounts disagree. The paper should report both and not treat either as a legal conclusion.

3.3 ePrivacy and § 25 TDDDG

The ePrivacy Directive’s Art. 5(3) requires prior consent for storing information in, or gaining access to information already stored in, terminal equipment, except for transmission or a service explicitly requested by the user [13]. Germany implements this in § 25 TDDDG, which applies regardless of whether the data are personal [20]. The decisive question is whether passive reception of a broadcast MAC is “access” to stored information. The better textual reading is that it is not: the sensor does not store anything in the phone, does not instruct the phone to disclose stored information, and does not retrieve data from the phone’s memory. German legislative materials use “Abruf” and “Auslesen” (retrieval/readout), which imply active access [39]. But the EDPB’s functional approach—MAC addresses are information stored in the terminal equipment, and the receiving entity need not be the one instructing the transmission—creates material risk the other way [15]. The DSK’s digital-services guidance lists reading MAC addresses among the cases that must be assessed under § 25, without resolving passive reception [38]. As of November 2026, we are not aware of a published EU or German decision on passive BLE/Wi-Fi sniffing under Art. 5(3)/§ 25.

If § 25 applies, prior consent is required, and a GDPR Art. 6 basis is not a substitute. The viable routes are a non-MAC sensor or a legal regime that validly disapplies § 25, which we do not see for general footfall measurement. Consent at city scale is practically impossible: walking past a sensor, leaving Wi-Fi on, or seeing a sign is not consent. That is why the architecture in Section 4 prefers thermal/radar/camera/probe sensing.

3.4 DPIA and lawful basis

If the data are personal data — as we argue they should be treated — a DPIA is very likely required. Continuous sensing of a publicly accessible inner-city area is the paradigm of "systematic monitoring of a publicly accessible area on a large scale" under Art. 35(3)(c) [13]. The Baden-Württemberg private-sector "Muss-Liste" explicitly lists offline tracking of customer movements in department stores/shopping centres and the non-intended use of device sensors or radio signals to determine location or movement over a substantial period (Nos. 2 and 10) [21]. As of November 2026, we are not aware of a public-sector positive list in Baden-Württemberg, so for the city the statutory criterion applies directly [40].

A public body acting in the performance of its public tasks cannot rely on Art. 6(1)(f); it needs Art. 6(1)(e) read with Art. 6(3) and a legal basis in Union or Member State law, or consent. A private association or company can use Art. 6(1)(f), but the balancing test is strict for public-space tracking that is invisible to most passers-by or conducted without meaningful notice, and German and Dutch DPAs have repeatedly found that the data subject's interests prevail [19, 44]. The DPIA must assess necessity and alternatives: manual counts, light barriers, thermal/radar, on-device video counting, and sample-based surveys. If a less intrusive sensor can meet the accuracy requirement, the MAC-based design is hard to justify.

3.5 Enforcement is real but fact-specific

German enforcement has been administrative rather than fine-based for this use case. The ULD Schleswig-Holstein investigated two city-marketing operators using passive probe requests and concluded there was no legal basis; both stopped [22]. The Bavarian authority inspected a fashion chain's Wi-Fi tracking test after a complaint; the company stopped [41]. The Hamburg authority's 2014/2015 report treats WLAN/Bluetooth indoor tracking as highly intrusive and personal-data processing requiring a legal basis [42]. As of November 2026, we are not aware of a published German fine specifically for passive Wi-Fi/BLE footfall tracking; that is not evidence of legality.

The Dutch Enschede case is instructive. The Dutch DPA fined the municipality €600,000 for Wi-Fi tracking in the city centre (pseudonymised MAC + location, used to count unique visitors) [43]. The fine was annulled by the Rechtbank Overijssel on 2 February 2024 because the DPA had not proved that the municipality processed personal data; the Raad van State upheld that outcome on 29 July 2026, largely on a procedural ground [19, 36, 37]. The case is not a safe harbour. It shows that the legal status is fact-specific and that the regulator must prove identifiability. The French

JCDecaux

case went the other way and is the stronger warning against relying on hashing/salting [18].

3.6 The operator's claim, examined

The Code for Heilbronn blog says the sensor "remembers the device addresses," hashes them, stores them for at most one hour, and transmits only the number of counted addresses [2]. The public 2026 firmware does not literally implement a one-hour store: it truncates the last two bytes into a 16-bit bitmap and resets the bitmap at each wake [3]. Both designs are materially better than storing raw MACs. Neither is a formal anonymisation guarantee:

- the raw MAC is observable at the sensor before truncation;
- the 16-bit ID has a small space and frequent collisions;
- the one-hour cache (if present) creates a one-hour linkability window per sensor;
- a keyed or unkeyed hash can be reversed or enumerated by an adversary with the key or the 48-bit space;
- the published per-sensor 15-minute counts can reveal a single device in low-traffic cells.

The honest formulation is: **the architecture is data-minimising at the edge, but the sensing stage still processes personal data, and the published aggregate may become personal data in small cells.** That is a privacy-by-design claim, not an anonymity claim.

3.7 What a defensible privacy claim requires

A defensible claim needs a threat model, a privacy unit, a formal or architectural guarantee, a retention policy, a DPIA, and a legal basis. The PET stack in Section 4.4 makes this concrete. The minimum for a pilot is:

- passive only; no probe requests, association requests, or OS APIs;
- no raw MAC retention; the sensor necessarily observes the address in RAM; the operator says a hashed cache may be kept for up to one hour, while the public 2026 code keeps only a 16-bit bitmap for the scan;
- per-sensor, per-window keyed hashing, not a 16-bit truncation;
- no cross-sensor or cross-day linkage;
- aggregate to zones/time bins with a k-threshold and suppression;
- central differential privacy at a trusted aggregator, with a device-day privacy unit and contribution bounds;
- a published budget ledger and explicit guarantee language;
- on-site signage, a real opt-out, and Art. 28 contracts or an Art. 26 arrangement.

If the project cannot document why § 25 does not apply, it should either obtain consent (impractical at city scale) or switch to a sensor that does not extract MAC addresses. That is the strongest privacy argument, and it is also the most reliable measurement argument: the alternatives do not have the same address-regime bias.

4. What privacy-first should mean

4.1 Sensing alternatives

The good news is that the MAC-based design is not the only option. The alternatives have different accuracy, cost, privacy, and deployment profiles. Table 1 summarizes the evidence.

Modality	Reported performance	Cost (2026 planning)	Privacy risk	Best use
BLE MAC sniffing (baseline)	5–12% of devices discoverable; 97%/92% count accuracy only when penetration is assumed 100% [10]; MAC randomization defeated for ~96% of Android phones [7]	€100–500	Very high	None for public footfall; at most a short-lived calibrated index under a DPIA
Thermal/IR array	89–95% doorway accuracy; degrades >30 °C, >7.5 m, groups [30]	€150–400/node	Low	Default at store entrances and narrow passages
mmWave radar	96% presence, 69% count outdoor; −2.5% aggregate bias [28]; 3-month F1 84%, flow r 60–85% [29]	€300–700	Low	Wide, weather-exposed crossings; presence/flow
Edge AI camera	13.8% day / 29.8% night MAPE outdoors [26]	€500–2,000; commercial €1,100–7,400	High (DPIA)	2–4 reference sites; pedestrian/cyclist classification
LiDAR	>97% disaggregate count; 1.6–4.3% 15-min error [27]	€1,000–8,000	Low–moderate	1–2 high-accuracy reference sites
Aggregated probe data	City-wide movement; coverage/bias unknown	Licence	High/commercial	City context, O/D, trends — not street-level truth
Manual counts	~11% clicker error; 2-hour observer fatigue	€25–40/hour	Low	Quarterly calibration audits
Sensor fusion	Hierarchical model; no single accuracy figure	€20k–60k development	Depends on inputs	The core of a defensible stack

Table 1. Sensing alternatives. Costs are planning estimates, not quotations; accuracy figures are from the cited studies and are context-dependent.

The table supports a clear ranking for Heilbronn. Thermal arrays are the best privacy/cost/accuracy trade-off at narrow gateways. Edge cameras are the accuracy benchmark but carry the heaviest DPIA and public-acceptance burden. mmWave and LiDAR are gap-fillers: good at presence/flow, weaker at exact counts, and best used at a few wide or high-value sites. Aggregated probe data are useful for city-wide context but are a non-probability sample; they cannot replace local ground truth. Manual counts are the only credible ground truth and should be a recurring program, not a one-off.

4.2 A recommended Heilbronn stack

A defensible pilot would use:

- **8–12 thermal nodes** at store entrances, the passages around Stadtgalerie, Marahaus and Götzenturm, and transit shelter entrances;
- **2–4 edge cameras** at wide plazas or complex intersections as the accuracy benchmark, with on-device inference and no frame retention;
- **1–2 mmWave/LiDAR nodes** at wide, weather-exposed crossings (Hagenbuchersee, Neckarufer);
- **aggregated probe data** for city-wide context and event flows, not street-level truth;
- **twice-yearly manual/video ground truth** at each site, stratified by time and weather;
- **transit validators** as a covariate where available.

This stack does not need MAC addresses. It also avoids the address-regime bias: thermal and radar count bodies, cameras count people, and manual counts calibrate the rest.

4.3 Calibration

Uncalibrated sensors are indexes, not measurements. The calibration program should estimate a site-specific detection probability and expansion factor, with uncertainty, and propagate that uncertainty into the decision. Options:

- **Ratio/expansion estimator:** `P_hat = k · y`, where `k = 1/r` and `r` is the detection rate measured against manual counts. This is standard sampling theory and is used in Bluetooth transport studies [10, 11].
- **Capture–recapture:** use two independent sensors or a sensor plus manual counts to estimate the unobserved population. This requires careful independence assumptions and is fragile for pedestrians.
- **Regression calibration:** regress ground truth on sensor counts plus covariates (time, weather, site), with a time-based holdout. This is the most flexible but can hide bias if the calibration sample is not representative.
- **Hierarchical Bayesian calibration:** model sensor-specific detection probabilities with partial pooling across sites. This is the right default when there are many sites and few counts per site; it produces 80% credible intervals and borrows strength.

The pilot should publish the calibration dates, the reference method, the estimated factor, and the interval. A single global 0.7 is not a calibration.

4.4 The privacy stack

The PET analysis supports a layered design:

- **No raw MAC retention.** If a MAC-based sensor is used at all, the sensor necessarily observes the address in RAM to deduplicate. The operator says a hashed cache may be kept for up to one hour; the public 2026 code keeps only a 16-bit bitmap for the scan. Hash at the edge with a per-sensor, per-window key, rotate the key, and destroy the old key. Do not use a plain hash, a static salt, or a 16-bit ID as a persistent identifier.
- **Edge aggregation.** Transmit counts only; no identifiers in the uplink.
- **No cross-sensor/cross-day linkage.** No trajectories, dwell, or repeat-visit metrics.
- **Central DP at a trusted city aggregator.** Use a device-day privacy unit and contribution bounds (e.g. at most four sensor-bin pairs per device per day). Publish fixed daily/weekly coarse releases with a budget ledger.
- **Do not use local DP for sparse 15-min bins.** Local DP error scales as $1/(\varepsilon\sqrt{n})$ versus central $1/(\varepsilon n)$, so it destroys utility at small-city scale. At $\varepsilon = 1$, a 14-device cell carries ±3 noise; a live 15-min public heatmap is not viable with a meaningful annual budget [23].
- **k-anonymity as a secondary check.** Four spatio-temporal points identify ~95% of people in a 1.5M-person mobility dataset [25]. In our judgement, generalizing to k = 20 would erase the local footfall signal. Use k as a suppression rule, not the primary guarantee.
- **No federated learning.** The pedestrian's phone does not run our code, so FL is not applicable to passive BLE sniffing.
- **No non-DP synthetic data as an anonymity claim.** Non-DP synthetic data has no formal guarantee; use DP synthesis plus membership-inference/Anonymeter evaluation for public release, not for the live dashboard.

The guarantee language matters. Say "privacy-enhanced," "aggregated with a k-threshold," or "central DP with $\varepsilon = \ldots$"; do not say "anonymous" unless a documented anonymisation assessment supports it. Publish what cannot be guaranteed: fine-grained public 15-min heatmaps, individual trajectories, repeat-visit metrics.

4.5 Pilot protocol

A 12-week pilot at six sites would be enough to estimate detection rates and calibrate the stack. The go/no-go criteria should be explicit:

- **MAPE ≤ 20%** against manual/video ground truth;
- **|bias| ≤ 10%** at the site level;

- **80% interval coverage ≥ 75%**;
- **no raw video or point-cloud retention**;
- **DPIA signed off**; legal basis documented;
- **cost within the pilot budget**.

A planning estimate is **€81k–161k** for year 1 including a **€20k–40k** backend allowance (**€61k–121k** excluding backend), and **€30k–50k** recurring, excluding staff time. These are estimates, not quotations; the point is that the privacy-first stack is not free, but it is comparable to the legal and reputational cost of a contested MAC deployment.

5. From sparse observations to decisions

5.1 Four estimands

Most overclaiming in urban sensing comes from letting one model silently answer four different questions:

1. the expected count at each sensor;
2. the latent spatial intensity field (heatmap);
3. the forecast of either;
4. the decision (staffing, routing, operations).

The heatmap is a state estimate; the staffing plan is a decision under uncertainty. They need different validation. The paper should name the estimand before choosing a model.

5.2 Spatial support

Eight sensors over a 580 m × 280 m bounding box give an equivalent uniform spacing of about **142 m** if they fill the area, or about **72 m** along a single street if they lie along one axis. A 100–150 m cell is defensible; a 25–50 m raster is a visualization, not an inference. The heatmap should be on the **walkable network**, not an isotropic Gaussian over building footprints. Every map should show the distance to the nearest sensor and a prediction-standard-deviation mask.

5.3 Spatial model

Footfall is a non-negative, overdispersed, zero-inflated count. A hierarchical negative-binomial model with a low-rank spatial field (Matérn GP/SPDE or a small set of spatial basis functions) is the statistically correct default [31, 32]. Ordinary kriging is a baseline; IDW is a visualization baseline. With eight sensors, a free CAR/BYM2 field on a fine grid is not identifiable—the prior would do almost all the work. Use a parametric Matérn covariance with a hierarchical prior, or fix the smoothness and estimate only range and nugget. Pool over time, but allow weekday/weekend and day/night non-stationarity. Report leave-one-sensor-out error and prediction intervals.

5.4 Forecast

The forecast core should be a global (panel) gradient-boosted model with quantile outputs and conformal calibration [33, 34]. Features: calendar, weather forecast, holidays, events (only if announced before issue time), POI/transit, lags, and sensor identity as a pooled effect. This is the strongest small-data baseline and is likely to beat an LSTM/Transformer/GNN on 8 series × 5 months. Deep spatiotemporal models and time-series foundation models should be benchmarks and ablations, not the default. The largest leakage risk is weather: training on observed weather and serving with forecast weather is optimistic. Evaluate with archived forecast weather, or a calibrated forecast-error simulation, and report both the observed-weather oracle and the deployable number.

5.5 Decision layer

The decision layer should be explicit about the objective and constraints. For staffing, the current heuristic is a reasonable first step:

```
required staff = ceil(index × k / visitors-per-staff-hour × safety factor)
  subject to minimum staff, shift length, opening hours, cost
```

`k` converts index units to visitors. It is **unvalidated**. Setting `k = 1` would treat one index unit as one visitor, which the measurement audit says is not justified; the plan is only as calibrated as `k`.

A production system could use a stochastic optimization or a robust optimization, but only when the utility and constraints are measured. The important move is to separate the predictor from the decision rule, propagate uncertainty, and evaluate **decision regret**: what would the plan have cost if the true footfall had been known? Without POS/queue/labor data, the paper can only compare to a flat schedule under stated assumptions. That is a heuristic comparison, not ROI.

5.6 Evaluation harness

A credible evaluation harness includes:

- **time-based holdouts** and rolling-origin evaluation;
- **leave-one-sensor-out** to test spatial generalization;
- **multiple seeds** for stochastic models;
- **Diebold–Mariano** tests for forecast differences;
- **calibration** (interval coverage, CRPS, quantile loss);
- **cost curves** and decision regret;
- **data provenance** (source, version, transformations, missingness);
- **model cards** and a metrics JSON.

This is standard MLOps hygiene [35], but it is often skipped in hackathon systems. The existing PulseCast metrics are a start: they include a seasonal-naive baseline, a time-based holdout, and a weather ablation. They do not include decision regret, spatial cross-validation, or calibration of the uncertainty band.

5.7 What the existing results show

The PulseCast results are a useful anchor. The production 7-day model reaches 3.32/27.5/0.777; the day-ahead model 3.36/27.8/0.763; the seasonal naive 5.40/44.7/0.462. Per-sensor MAE ranges 2.26–5.83. The weather ablation is large; the holiday effect is small. The mission brief's 3.36/27.8/0.779 was the quick baseline; the generated metrics are canonical: production 7-day 3.32/27.5/0.777, day-ahead 3.36/27.8/0.763. These results show that a global model can learn the daily/weekly shape and the weather sensitivity of the index. They do not show that the index is people, that the model generalizes across cities, or that the staffing plan improves outcomes. The honest reading is a **relative decision-support signal**.

The deployed system provides a second kind of evidence. The public app is at https://pulse.noelsabia.dev and the API at https://api-pulse.noelsabia.dev. The v1.1 images are 16.9 MB (api), 249 MB (web), and 952 MB (ml). Plan generation is 19 ms cold; the contract is under two seconds. In the v1.1 120-client Cloudflare k6 run, 120/120 SSE connections opened and closed with zero connection errors; p95 event latency was 331 ms, p95 scenario latency

646 ms, presenter POST latency 49.60 ms, and the run lasted 4m57.9s. These numbers show that the delivery path works at the demo scale. They do not validate the measurement or the decision quality.

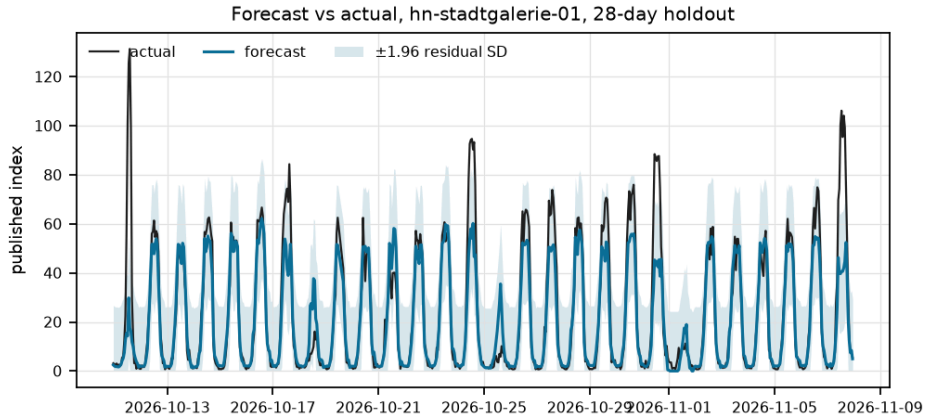


Figure 4. Forecast versus actual for one sensor over the 28-day holdout. The model captures the daily/weekly shape but misses spikes; the band is a planning aid, not a calibrated 95% interval unless coverage is reported.

6. How far it generalizes

6.1 The six challenges

The OKF_Vault bundle contains six challenges: Pulse of the City (footfall), WasteRouting.AI (glass fill levels), Measure Where It Matters (heat-sensor placement), LoRaWAN network planning, Riding the Sentiment (cable-car debate), and the Learning Guide (family offers). They are not six versions of the same problem. They differ in what is observed, whether a spatial field exists, whether the decision is forward-looking, and whether the decision changes the observation process.

6.2 The shared contract

The shared abstraction is a **decision-grade sensing contract**, not a universal model:

```
observation:   y_t      = H_t x_t + e_t
field:         x_hat_t  = I(y_<=t, H_t, c_t, G)
forecast:      x_hat_t+h = F(x_hat_<=t, c_t+h; theta)
decision:      a*       = argmax_a E[U(a, x_t+h)]  subject to constraints
evaluation:    L(y, x_hat) + R(a*, x_oracle) + privacy leakage
```

`x_t` is a latent state over spatial units; `H_t` is the observation operator; `c_t` is context available at decision time; `G` is geometry or graph structure; `I` is spatial inference; `F` is temporal forecasting; `U` is the decision utility. The key distinction is **state estimation** versus **design**: in staffing, the pipeline estimates a state and acts on it; in sensor placement, the decision changes what is observed next. Calling both “forecasting” hides the inversion.

6.3 Transfer matrix

Challenge	Class	What transfers	What must change	Evidence
Pulse / staffing	Direct	Observation handling, local field, weather-aware forecast, staffing heuristic, deployment	Store-specific assumptions; calibration; outcome data	MAE 3.32 vs 5.40 naive
Pulse / advertising	Analytical, not tested	Same field and forecast; allocation objective replaces staffing	Ad inventory, cost, reach, dwell, conversion; city-wide sites for a real campaign	No result; analytical extension
Waste routing	Conditional	Temporal forecast, clustering, configurable VRP, holiday/season context	Coordinates or zone map, capacity/type, fleet/depot/shift, overflow ground truth	No result; dataset documented
Sensor placement	Sibling	Observation model, spatial inference, error-vs-cost evaluation	Decision changes `H_t`; active learning/optimal design; no temporal forecast	Protocol only
LoRaWAN planning	Sibling	Link/coverage prediction, candidate-site ranking, address-weighted coverage	Packet-loss estimate, buildings/terrain, longer window, GPS calibration	No result; dataset documented
Cable-car sentiment	Non-application	At most an event/attention signal	No physical field, no forecast target, no allocation decision; different NLP stack	Descriptive task
Learning guide	Non-application	At most a location/context consumer	No sensing, no field, no forecast; matching and child-privacy regime	Product task

Table 2. Transferability matrix. The stack generalizes as a contract across a subset; two challenges are siblings and two are out of scope.

6.4 Case studies

Staffing (empirical). The only evidence-backed forecast; the decision layer is unevaluated. The path is: PAX index → weather-aware global forecast → staffing heuristic → plan. The next step is calibration and decision-level evaluation.

Advertising (analytical extension). The same field and forecast can rank the eight locations and time windows by expected exposure. This is a decision-layer substitution, not a new sensing study. It cannot support city-wide ad planning or effectiveness claims without more sites and ad-economics data.

Waste routing (conditional). The temporal-forecast and decision-rule pattern transfers, but the spatial object becomes a container graph, not a metric field. The dataset has no coordinates, capacity, fleet, depot, or shift data; a route-saving claim would be driven by assumed geometry. The honest output is a method and a data-acquisition plan, not a result.

Sensor placement and LoRaWAN (siblings). The spatial-inference layer transfers, but the decision is a design decision: where to place sensors or gateways. The evaluation is an error-versus-cost curve, not a temporal forecast. LoRaWAN has real data and explicit candidate sites, but the missing packet-loss baseline is a serious limitation. Both are separate “spatial design” contributions, not applications of the forecast core.

6.5 Falsifiers

The generalization claim is only credible if it can fail. The paper should state the tests:

1. Waste forecast fails a naive baseline after accounting for irregular sampling and `overfull` encoding.
2. LoRaWAN link model cannot beat a distance-only baseline on held-out gateways.
3. Advertising allocation is unstable under plausible weather scenarios.
4. Store-location ranking with eight points is no better than random at city scale.

5. Decision regret is large despite low MAE.
6. Aggregate counts can be re-identified, or the privacy claim fails the threat model.

A paper that reports these tests, even if some fail, is stronger than one that asserts six-way generalization.

7. Evaluation agenda

7.1 Measurement

- Ground-truth counts at each site, stratified by time, weekday, and weather; manual counts are the reference, with inter-rater reliability.
- Penetration and device-multiplicity study for any MAC-based sensor, with a documented sampling design.
- Duty-cycle and detection-radius study; report the probability that a person with a given dwell time is detected.
- False-positive/false-negative audit using co-located thermal/camera/manual counts.

7.2 Forecast

- Rolling-origin evaluation, not a single 28-day holdout.
- Leave-one-sensor-out and leave-one-site-out.
- Calibration: interval coverage, CRPS, quantile loss.
- Decision regret: compare the plan to an oracle plan and a flat schedule under explicit assumptions.
- Weather ablation with **archived forecast weather**, not only observed weather.

7.3 Privacy

- DPIA before deployment; DPO sign-off; Art. 36 consultation if residual high risk.
- Threat model: singling out, linkability, inference, reconstruction, sensor compromise.
- Formal guarantee: central DP with a stated ε/δ, privacy unit, contribution bound, and budget ledger; k-threshold as a secondary check.
- Re-identification test on the published aggregates; small-cell suppression audit.
- Legal basis and § 25 memo; signage and opt-out usability test.

7.4 System

Measured on v1.1: 120/120 SSE connections opened and closed, zero connection errors, p95 event latency 331 ms, p95 scenario latency 646 ms, presenter POST 49.60 ms, over 4m57.9s; plan generation 19 ms cold (contract < 2 s); images 16.9 MB (api), 249 MB (web), 952 MB (ml). Remaining checks:

- cold deployment time not measured;
- failure drills: FastAPI down, Open-Meteo down, tunnel down, pod restart;
- observability: health/ready, data/model versions, staleness labels;
- a repeat of the 120-client run after any image or cluster change.

7.5 Proposed experiments on the existing CSV

These are real analyses that do not require new sensors:

1. **Quantization audit** — confirm 100% multiples of 0.7; reconstruct `U`; compare `density-01` regimes. (Done.)
2. **Coverage/zero-inflation** — per-sensor missingness, zero runs, correlation with battery. (Done.)
3. **Cross-sensor correlation** — raw, detrended, daily; effective number of independent sensors. (Done.)
4. **Subsampling** — train on 1-in-k rows (30/60/120 min) and measure MAE/WAPE degradation. This quantifies the value of temporal sampling.
5. **Leave-one-sensor-out** — train on seven, test on the held-out sensor. This tests shared structure.
6. **Factor sensitivity** — rescale counts by 0.5/0.7/1.0/1.4 and compare staffing plans. This shows why calibration matters for decisions.
7. **Decision-level evaluation** — compare the plan to a flat schedule on coverage and cost under stated assumptions.

Subsampling and factor sensitivity are the most informative for the paper's thesis: they show that the model's accuracy is not the same as decision quality, and that a single calibration constant can change the plan.

8. Discussion and limitations

The main limitation is the measurement, not the model. The sensor samples about 10 seconds out of every 15 minutes, counts only randomized BLE addresses, truncates the address to 16 bits, multiplies by a fixed 0.7 factor, and cannot distinguish people from devices. The published value is a useful relative index, but it is not a census and not a 15-minute count. Device penetration is unknown. A single global factor cannot absorb device ownership, multiple devices per person, Bluetooth settings, and the short scan window.

Coverage is also a limitation. Eight sensors in a 580 m × 280 m cluster cannot support a city-wide heatmap or a store-location recommender. In the 18 June coverage panel, only 28.2% of 15-minute slots have all eight sensors, and gaps are not random. The 28-day holdout is short, the weather is observed rather than forecast, and there is no event log, POS data, or labor-cost data. The deployment has one API replica, no SSO, no HA, and no live sensor ingestion. The 15-minute and 1-hour horizons are replay/nowcast, not a live feed.

The privacy claim also needs care. The application stores no MAC addresses, and only aggregate counts leave the sensor. But the sensing stage still processes link-layer identifiers at the edge; the 16-bit ID is a pseudonymous dedup key, not a cryptographic hash; and passive MAC sensing in public space has a contested legal basis in Germany. The "privacy-preserving" label is an architectural claim, not a formal guarantee. A production deployment would need a DPIA, a documented legal basis, data minimization, retention limits, and, where possible, a sensor that does not extract MAC addresses at all.

The paper is also an opinionated design proposal, not a controlled trial. The alternative architecture has not been deployed at Heilbronn scale; the cost figures are planning estimates; the legal analysis is not legal advice. The tone should be constructive: Code for Heilbronn is a volunteer civic-tech group that built a low-cost, privacy-conscious prototype under real constraints. The critique is of the claim and the measurement design, not the volunteers' intent.

What to keep from the hackathon build. The decision-first framing, the global model, the time-based holdout, the seasonal-naive baseline, the transparent staffing heuristic, the cached fallback, the build-time batch, and the honest limitations are all worth keeping. The plan screen is the product; the map and charts support it.

What to replace. The sensing layer (thermal/radar/camera/probe/manual), the calibration (site-specific, with uncertainty), the privacy stack (no raw MAC retention, central DP, k-threshold), the spatial model (network-aware hierarchical field), the forecast (quantile GBM with conformal calibration and forecast weather), and the platform (auth, observability, SLOs, multi-tenancy) in that order. Do not scale the model before fixing the measurement.

9. Related work

Passive counting. Bluetooth and Wi-Fi sensing has a long history in pedestrian counting. Classic Bluetooth inquiry detected 2–7% of pedestrians in early studies [11]; static Bluetooth sensors produced usable trends but required site-specific correction factors [12]; Wi-Fi/Bluetooth sensors can estimate densities and wait times when penetration is known [10]. MAC randomization reduces linkability but is not a census, and fingerprinting can defeat it [7, 8, 9]. Our audit is consistent with this literature and makes the sampling design explicit.

Forecasting. Gradient-boosted trees remain a strong baseline for tabular, covariate-rich, small-data problems [33]. Weather is a known driver of pedestrian activity, but its magnitude is context-dependent; we treat it as predictive, not causal. The heatmap/forecast literature offers a menu of spatial and temporal models; at eight sensors, the statistically principled choice is a hierarchical spatial field plus a global quantile forecast, not a deep model [31, 32, 34].

Privacy and law. The legal analysis draws on the GDPR, *Breyer*, the EDPB pseudonymisation and anonymisation guidelines, *JCDecaux*, the Dutch Enschede case, and German DPAs [13–22]. The technical privacy analysis draws on differential privacy, k-anonymity, membership inference, and mobility uniqueness [23–25]. The conclusion is that "privacy-preserving" is an architectural claim that must be specified, not a label.

Urban sensing and decision support. The lesson from commercial mobility products is to make the parameters explicit: consent, aggregation, DP, retention, and audit. The sensing alternatives literature provides the accuracy/cost/privacy trade-offs [26–30]. The MLOps literature provides the evaluation hygiene [35]. The contribution here is to connect the measurement audit to the decision layer and to the legal/privacy constraints in one place.

10. Conclusion

The PulseCast case is a useful cautionary tale. A small, weather-aware model can turn a noisy inner-city index into a weekly staffing plan that beats a seasonal-naive baseline and runs in under two seconds on a real deployment. But the index is not people. Each value is a short BLE scan scaled by 0.7, sampled every 15 minutes, with unknown penetration, incomplete coverage, and a hardware calibration break. The "privacy-preserving" label is an architectural claim, not a formal guarantee; the 16-bit truncation is not a hash; and passive MAC sensing in public space has a contested legal basis.

The constructive path is clear:

1. **Calibrate** the index against manual/video ground truth, with site-specific factors and uncertainty.
2. **Change the sensing layer** to thermal/radar/camera/probe/manual, with a DPIA and no raw MAC retention.
3. **Model the field** as a network-aware hierarchical count process, at a 100–150 m support, with uncertainty.
4. **Forecast** with a global quantile GBM, forecast weather, and conformal calibration.
5. **Evaluate the decision**, not just MAE: decision regret, cost curves, and a pilot with go/no-go criteria.
6. **Publish the limits**: what the data do not measure, what the privacy guarantee is, and what would falsify the claim.

The honest conclusion is narrower and more useful than a city-scale claim: the decision loop works, the relative forecast is credible, and the next step is calibration and a privacy-first sensing layer. We would rather ship a plan a manager can use and state what it does not measure than promise a city-wide heatmap the data cannot support.

Data, code, and AI disclosure

The raw challenge CSV is not redistributed. Derived aggregates, the trained model artifacts, the metrics JSON, and the figure-generation scripts can be shared subject to the organizers' permission. The code for the audit and the PulseCast system will be released after the hackathon; the live app is at https://pulse.noelsabia.dev and the API at https://api-pulse.noelsabia.dev. The research notes that support the audit are in `paper/research/` and can be deposited alongside the paper or linked from the repository.

AI tools were used for drafting and editing. The authors checked each number and citation; the legal analysis is a research draft, not legal advice, and some sources are guidance, DPA positions, or inferences rather than binding law. AI tools are not authors and take no responsibility for the content. A deployment should obtain a written opinion from German counsel, especially on the § 25 TDDDG question.

References

[1] Schwarz Digits / Code for Heilbronn. *The Pulse of the City*, challenge brief, 2026. Local: `OKF_Vault/Schwarz Digits/FCH Challenge The _Pulse of the City_.pdf`.

[2] Code for Heilbronn. "Passantenzahlen messen mit LORA," 18 September 2025. https://codeforheilbronn.de/blog/2025/09/passantenzahlen-messen-mit-lora/

[3] Code for Heilbronn. `lorax` / `sensor-pax` firmware and backend, commit `e8be8bf5c11e1b76c88c3f176382c6df99eea989` (develop, 18 June 2026). https://codeberg.org/cfhn/lorax/commit/e8be8bf5c11e1b76c88c3f176382c6df99eea989

[4] `ESP32-Paxcounter` v3.8.0 and `libpax`. https://github.com/cyberman54/ESP32-Paxcounter; https://github.com/dbinfrago/libpax

[5] PulseCast data-quality analysis. `paper/research/07-data-quality.md`, 2026.

[6] Bluetooth SIG. "Enhancing device privacy and energy efficiency with Bluetooth Randomized RPA Updates," 14 May 2025. https://www.bluetooth.com/blog/enhancing-device-privacy-and-energy-efficiency-with-bluetooth-randomized-rpa-updates/

[7] J. Martin, T. Mayberry, C. Donahue, L. Foppe, L. Brown, C. Riggins, E. C. Rye, and D. Brown. "A study of MAC address randomization in mobile devices and when it fails." *Proceedings on Privacy Enhancing Technologies*

2017(4):365–383. DOI: 10.1515/popets-2017-0054.

[8] J. K. Becker, D. Li, and D. Starobinski. "Tracking anonymized Bluetooth devices." *Proceedings on Privacy Enhancing Technologies*

2019(3):50–65. DOI: 10.2478/popets-2019-0036.

[9] M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, and F. Piessens. "Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms." *AsiaCCS 2016*

, 413–424. DOI: 10.1145/2897845.2897883.

[10] A. Kurkcu and K. Ozbay. "Estimating pedestrian densities, wait times, and flows with Wi-Fi and Bluetooth sensors." *Transportation Research Record*

2644(1):72–82, 2017. DOI: 10.3141/2644-09.

[11] T. Nicolai and H. Kenn. "About the relationship between people and discoverable Bluetooth devices in urban environments." *Mobility '07*

, 72–78. DOI: 10.1145/1378063.1378076.

[12] Y. Malinovskiy, N. Saunier, and Y. Wang. "Analysis of pedestrian travel with static Bluetooth sensors." *Transportation Research Record*

2299:137–149, 2012. DOI: 10.3141/2299-15.

[13] Regulation (EU) 2016/679 (GDPR), https://eur-lex.europa.eu/eli/reg/2016/679/oj; Directive 2002/58/EC (ePrivacy), Art. 5(3), https://eur-lex.europa.eu/eli/dir/2002/58/oj.

[14] CJEU. Case C-582/14, *Patrick Breyer v Bundesrepublik Deutschland*, 19 October 2016. https://curia.europa.eu/juris/liste.jsf?num=C-582/14

[15] EDPB. *Guidelines 2/2023 on Technical Scope of Art. 5(3) of ePrivacy Directive*, v2.0, 7 October 2024. https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202302_technical_scope_art_53_eprivacydirective_v2_en_0.pdf

[16] EDPB. *Guidelines 01/2025 on Pseudonymisation*, 16 January 2025. https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf

[17] EDPB. *Guidelines 02/2026 on Anonymisation*, v1.0, 7 July 2026 (public consultation). https://www.edpb.europa.eu/system/files_en?file=2026-09%2Fedpb_guidelines_202602_anonymisation_v1_en_0.pdf

[18] Conseil d'État. *JCDecaux France*, no. 393714, 8 February 2017. https://www.conseil-etat.fr/fr/arianeweb/CE/decision/2017-02-08/393714

[19] Autoriteit Persoonsgegevens. *Wifi and bluetooth*. https://www.autoriteitpersoonsgegevens.nl/en/themes/internet-and-smart-devices/wifi-and-bluetooth

[36] Rechtbank Overijssel.

Enschede / Autoriteit Persoonsgegevens

, ECLI:NL:RBOVE:2024:594, 2 February 2024.
[37] Afdeling bestuursrechtspraak van de Raad van State.
Autoriteit Persoonsgegevens / College van B&W Enschede

, ECLI:NL:RVS:2026:4403, 29 July 2026.
[20] Telekommunikation-Digitale-Dienste-Datenschutz-Gesetz (TDDDG), § 25. https://www.gesetze-im-internet.de/ttdsg/__25.html
[21] LfDI Baden-Württemberg. *Muss-Liste* nach Art. 35 Abs. 4 DS-GVO, Nos. 2 and 10. https://www.baden-wuerttemberg.datenschutz.de/wp-content/uploads/2018/05/Liste-von-Verarbeitungsvorg%C3%A4ngen-nach-Art.-35-Abs.-4-DS-GVO-LfDI-BW.pdf
[22] ULD Schleswig-Holstein. *Tätigkeitsbericht 2019*, §5.4.8. https://www.datenschutzzentrum.de/tb/tb37/uld-37-taetigkeitsbericht-2019.pdf
[23] C. Dwork and A. Roth. “The algorithmic foundations of differential privacy.”
Foundations and Trends in Theoretical Computer Science

9(3–4):211–407, 2014. DOI: 10.1561/0400000042.
[24] A. Pyrgelis, C. Troncoso, and E. De Cristofaro. “Knock knock, who’s there? Membership inference on aggregate location data.” *NDSS 2018*. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_05B-5_Pyrgelis_paper.pdf
[25] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. “Unique in the crowd: The privacy bounds of human mobility.”
Scientific Reports

3:1376, 2013. DOI: 10.1038/srep01376.
[26] B. Cyganek, P. Janyst, M. Knapik, Ł. Przebinda, M. Ptaszkiewicz, and T. Tajmajer. “Evaluation of an embedded computer vision system for outdoor media audience analysis.” *PP-RAI 2022*. https://wydawnictwo.umg.edu.pl/pp-rai2022/pdfs/07_pp-rai-2022-101.pdf
[27] A. Lesani, E. Nateghinia, and L. F. Miranda-Moreno. “Development and evaluation of a real-time pedestrian counting system for high-volume conditions based on 2D LiDAR.”
Transportation Research Part C

114:20–35, 2020. DOI: 10.1016/j.trc.2020.01.018.
[28] G. Vaidya and M. A. Zúñiga. “Exploiting mmWave and deep-learning models to estimate people count in urban scenarios.”
IoTDI 2024

, 73–84. DOI: 10.1109/IoTDI61053.2024.00011.
[29] W. Wang, G. Vaidya, A. K. Bhattacharjee, F. Fioranelli, and M. A. Zúñiga. “A long-term study of mmWave sensing in an outdoor urban scenario.”
DCOSS-IoT 2023

, 240–247. DOI: 10.1109/DCOSS-IoT58021.2023.00049.
[30] C. Perra, A. Kumar, M. Losito, P. Pirino, M. Moradpour, and G. Gatto. “Monitoring indoor people presence in buildings using low-cost infrared sensor array in doorways.”
Sensors

21(12):4062, 2021. DOI: 10.3390/s21124062.
[31] N. A. C. Cressie.
Statistics for Spatial Data

(rev. ed.). Wiley, 1993.
[32] F. Lindgren, H. Rue, and J. Lindström. “An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach.”
Journal of the Royal Statistical Society: Series B

73(4):423–498, 2011. DOI: 10.1111/j.1467-9868.2011.00777.x.
[33] J. H. Friedman. “Greedy function approximation: A gradient boosting machine.”
Annals of Statistics

29(5):1189–1232, 2001. DOI: 10.1214/aos/1013203451.
[34] V. Vovk, A. Gammerman, and G. Shafer.
Algorithmic Learning in a Random World

. Springer, 2005.
[35] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. “Hidden technical debt in machine learning systems.”
NeurIPS 2015

, 2503–2511.
[38] Datenschutzkonferenz (DSK). *Orientierungshilfe der Aufsichtsbehörden für Anbieter:innen von digitalen Diensten (OH Digitale Dienste)*, v1.2, November 2024, paras 22, 26. https://www.datenschutzkonferenz-online.de/media/oh/OH_Digitale_Dienste.pdf
[39] Deutscher Bundestag. *Entwurf eines Gesetzes zur Regelung des Datenschutzes und des Schutzes der Privatsphäre in der Telekommunikation und bei Telemedien (TTDSG)*, BT-Drs. 19/27441, p. 38. https://dserver.bundestag.de/btd/19/274/1927441.pdf
[40] LfDI Baden-Württemberg and Ministerium für Justiz und Migration Baden-Württemberg. *Prüfung der Notwendigkeit einer Datenschutzfolgenabschätzung – Schwellenwertanalyse*, 29 November 2024. https://jum.baden-wuerttemberg.de/fileadmin/redaktion/m-jum/intern/PDF/Migration/Erlasse_und_Anwendungshinweise/2024/29.11.2024_Anlage_01_Schwellwertanalyse.pdf
[41] Bayerisches Landesamt für Datenschutzaufsicht (BayLDA). *Tätigkeitsbericht 2015/16*, §§3.10, 22.10. https://www.lda.bayern.de/media/baylda_report_07.pdf
[42] Der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit (HmbBfDI). *25. Tätigkeitsbericht Datenschutz 2014/2015*, §V.9. https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Taetigkeitsberichte_Datenschutz/Taetigkeitsberichte_PDF/25.TB_2014-2015.pdf
[43] Autoriteit Persoonsgegevens. *Boete wifitracking Enschede*, 29 April 2021, updated 29 July 2026. https://www.autoriteitpersoonsgegevens.nl/documenten/boete-wifitracking-enschede
[44] Autoriteit Persoonsgegevens. *Legal bases for WiFi tracking and Bluetooth tracking*. https://www.autoriteitpersoonsgegevens.nl/en/themes/internet-and-smart-devices/wifi-and-bluetooth/legal-bases-for-wifi-tracking-and-bluetooth-tracking