WPSC 2026 · TRACK 13 — TECHNOLOGIES

UNIVERSITY OF TARTU
Department of Geography

Is it secret?
Is it safe?

Building-level synthetic population for Estonia

from open administrative data

Anto Aasa
Mobility Lab · University of Tartu · SyntPop v3



THE PROBLEM

A gap between what planners need and what's published

Planners need individuals

- Activity-based transport models
- Epidemic & contact networks
- Accessibility & exposure analysis

Statistics give aggregates

- Census microdata is locked
- Only marginal tables published
- Within-zone variation lost

Aggregates throw away the joint distribution — who lives with whom, and where.

POSITIONING

Why not machine learning?

ML (Bayesian nets, GANs)

— Need detailed microdata to train

— Need serious compute

— If data is locked — you can't start

Ours: rule-based microsimulation

— No seed microdata — fully open inputs

— Explicit, documented rules

— Transparent & reproducible

The trade-off: we depend on the conditional probabilities — but every number traces to a public source.

THE ESTONIAN CASE

Whole-country scale, open data only

1.36 M
synthetic persons

580 k
households

284 k
residential buildings

Why Estonia?

- Our Mobility Lab is building a national transport model — it needs agent-level inputs.
- No public census microdata; the register & 2021 census publish only marginal aggregates.
- If it works here with open data only, it travels to many countries in the same position.

DATA & PRIVACY ARCHITECTURE

The heart of “is it safe?”

OPEN — public APIs

- Age × sex, education, activity
- Occupation, population grid
- Building register

OPEN — spatial layers

- Building footprints
- Settlement boundaries
- Mother tongue by area

ANONYMIZED — secure env

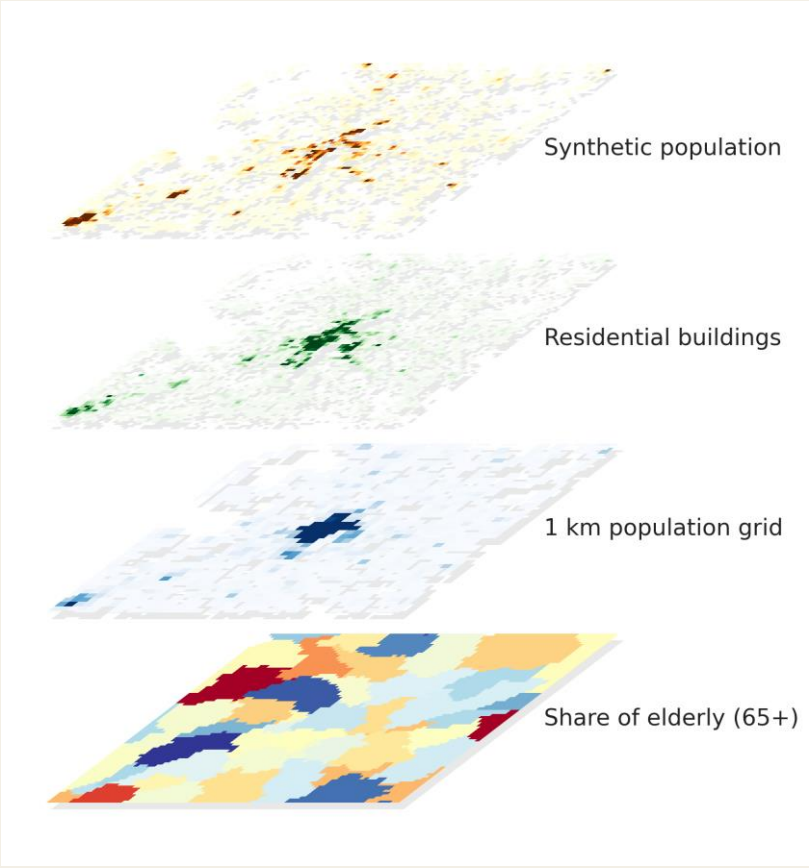
- Income brackets (k ≥ 3)
- Occupation, vehicle use
- Never exported as records

Privacy guarantee

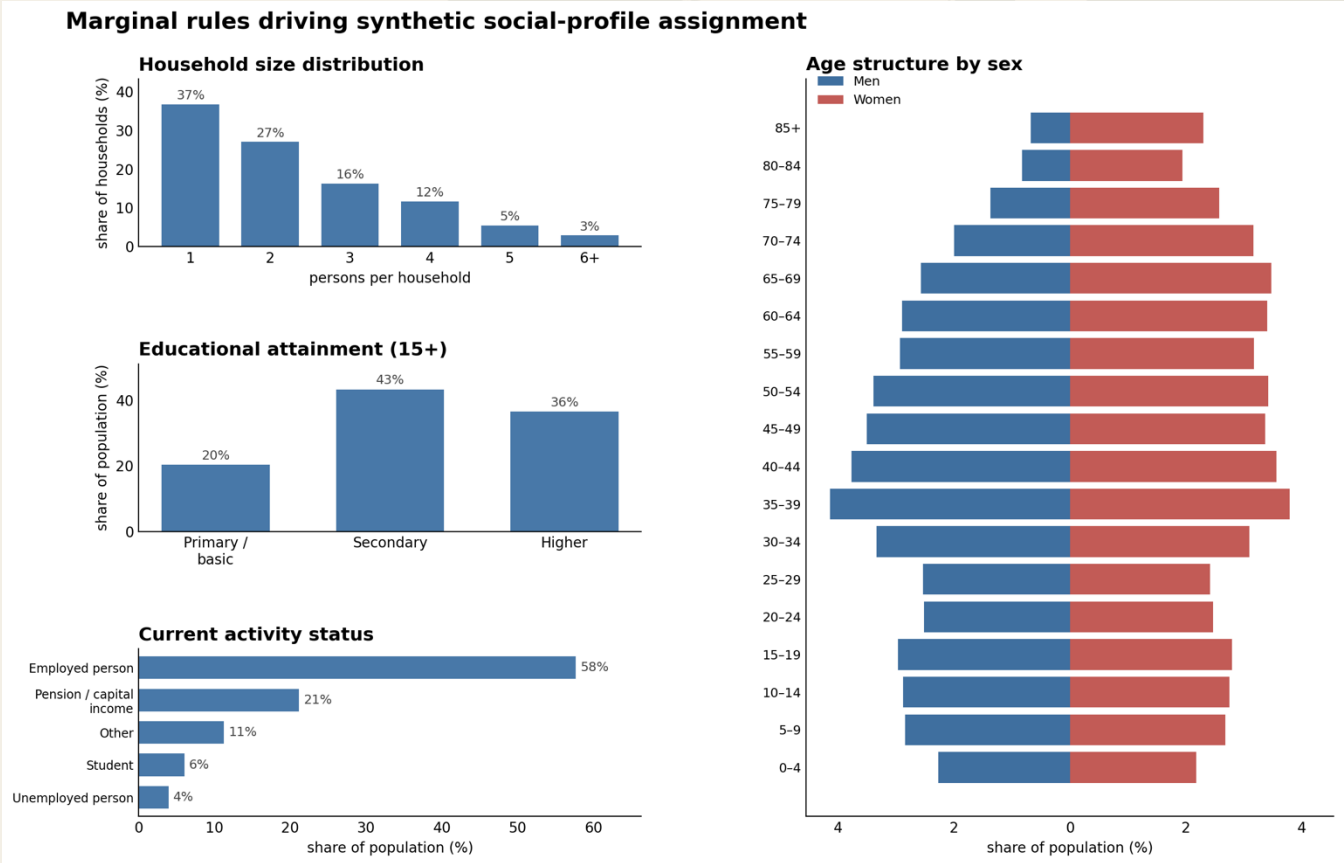
Microdata never leaves the secure environment — only locality-level distributions (k ≥ 3) are exported. Individual records are never released.

WHAT GOES IN

Data layers and the marginal rules



Input data layers → synthetic population



Marginal rules driving attribute assignment

PIPELINE AT A GLANCE

A reproducible R pipeline in three stages

01 SYNTHESIS

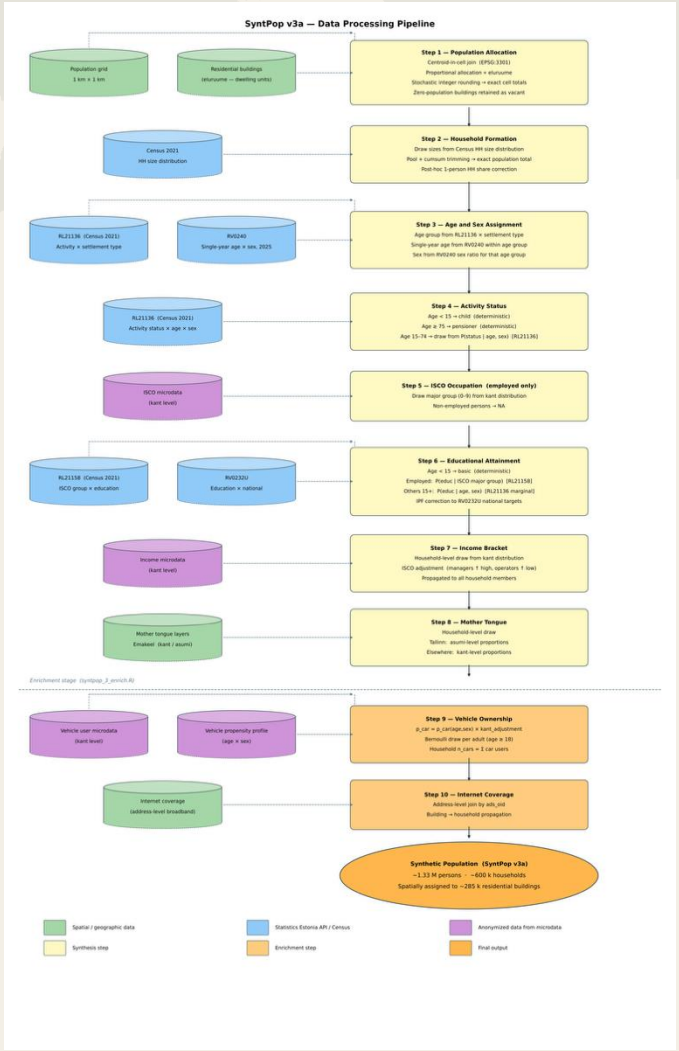
Allocate → households → age / sex → activity, occupation, education, income, mother tongue

02 ENRICHMENT

Vehicle ownership · internet access

03 COMMUTING

Work + school assignment · commute mode — new

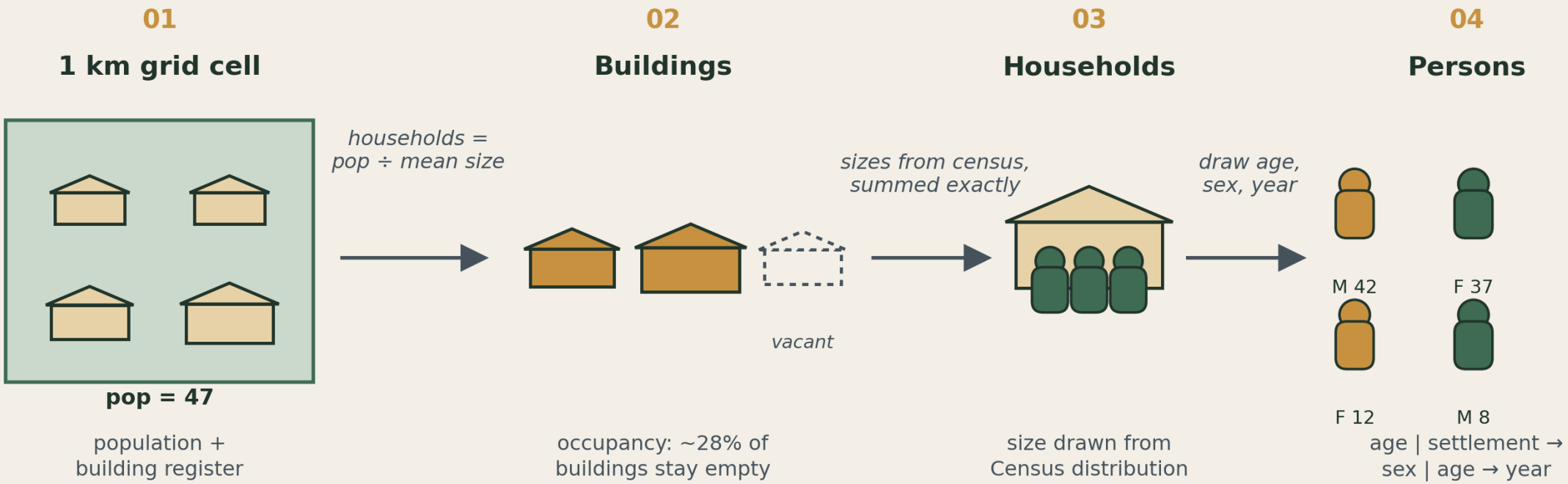


10 steps · open + anonymized inputs · checkpoint resume · RDS / CSV / GeoPackage.

METHOD CORE

From grid to households: occupancy-based allocation

Household-first, not building-first — every step an explicit, inspectable rule.



Multi-scale: population grid → buildings (with vacancy) → households → individual attributes.

VALIDATION

Demographic realism — and a solved problem

- ✓ **Age groups (6) —** all within 0.6 pp
- ✓ **Education 15+ —** matches register target
- ✓ **Income brackets —** 18 / 19 / 26 / 19 / 18 — on target
- ✓ **Mother tongue —** national + Tallinn neighbourhood level
- ✓ **Mean household size —** on target at 2.35 (was 2.15)

Household size was my old "structural gap" — the limitation I used to apologize for is now a result.

Commuting: work + school + mode

Assign where people go

— Workplace for ~603 k employed

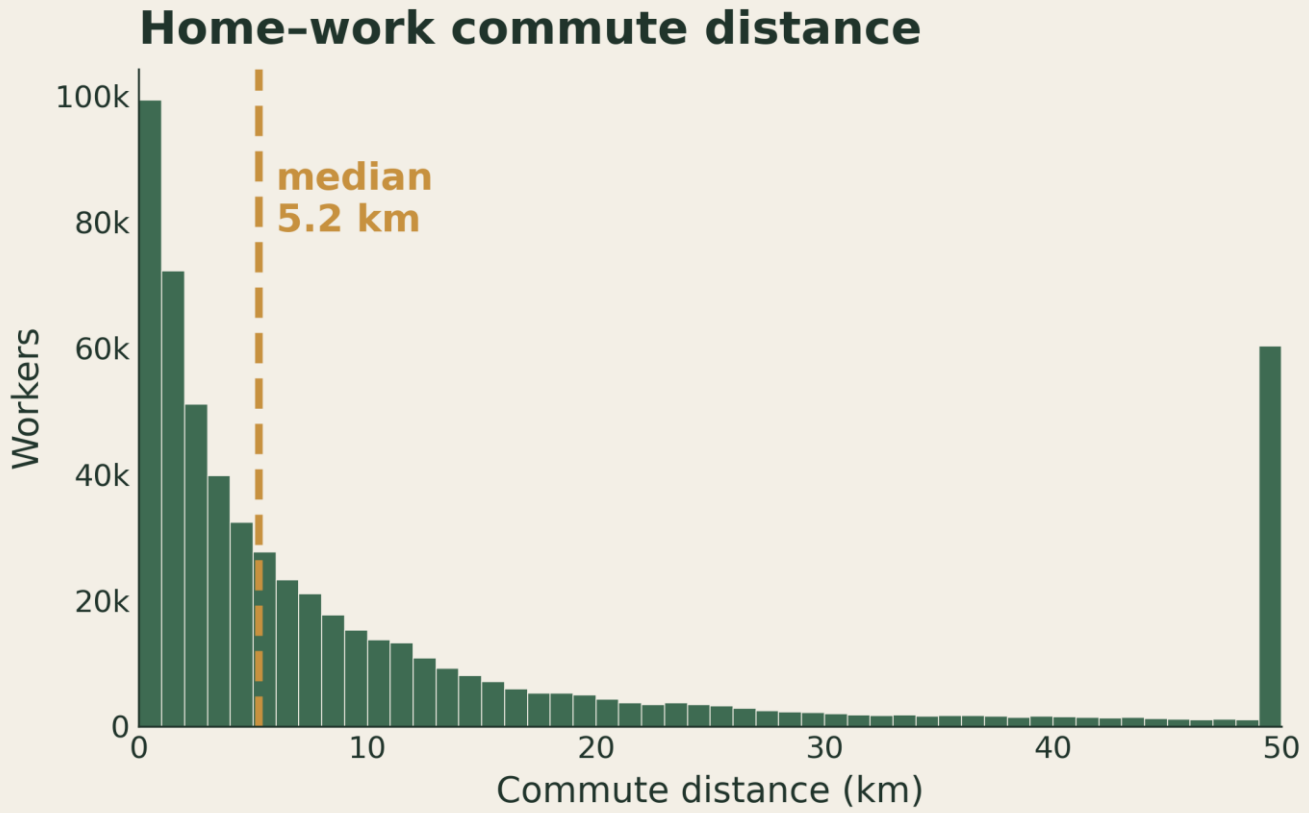
— School for ~193 k pupils & students

Blend three sources

— Mobile positioning — behavioural OD backbone

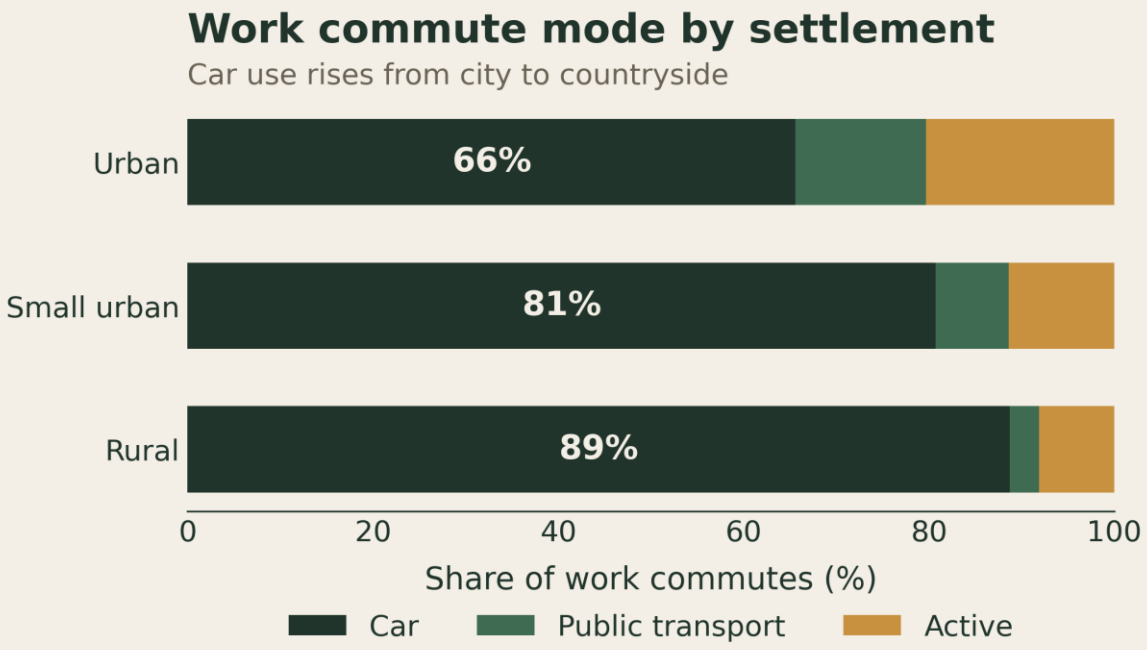
— Register matrix — fine spatial detail

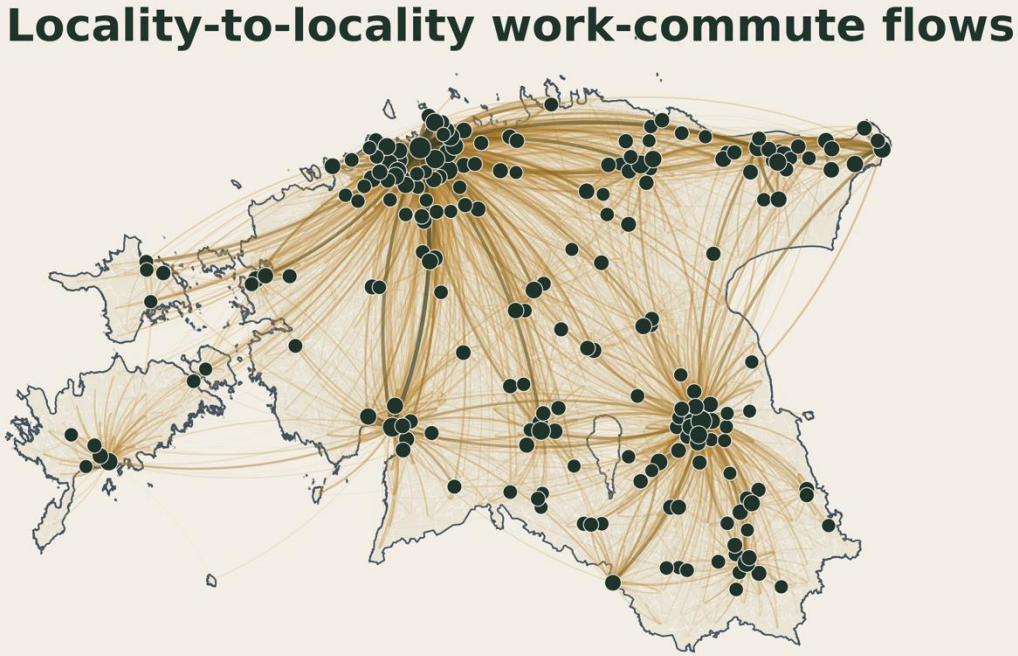
— Travel survey — mode & calibration



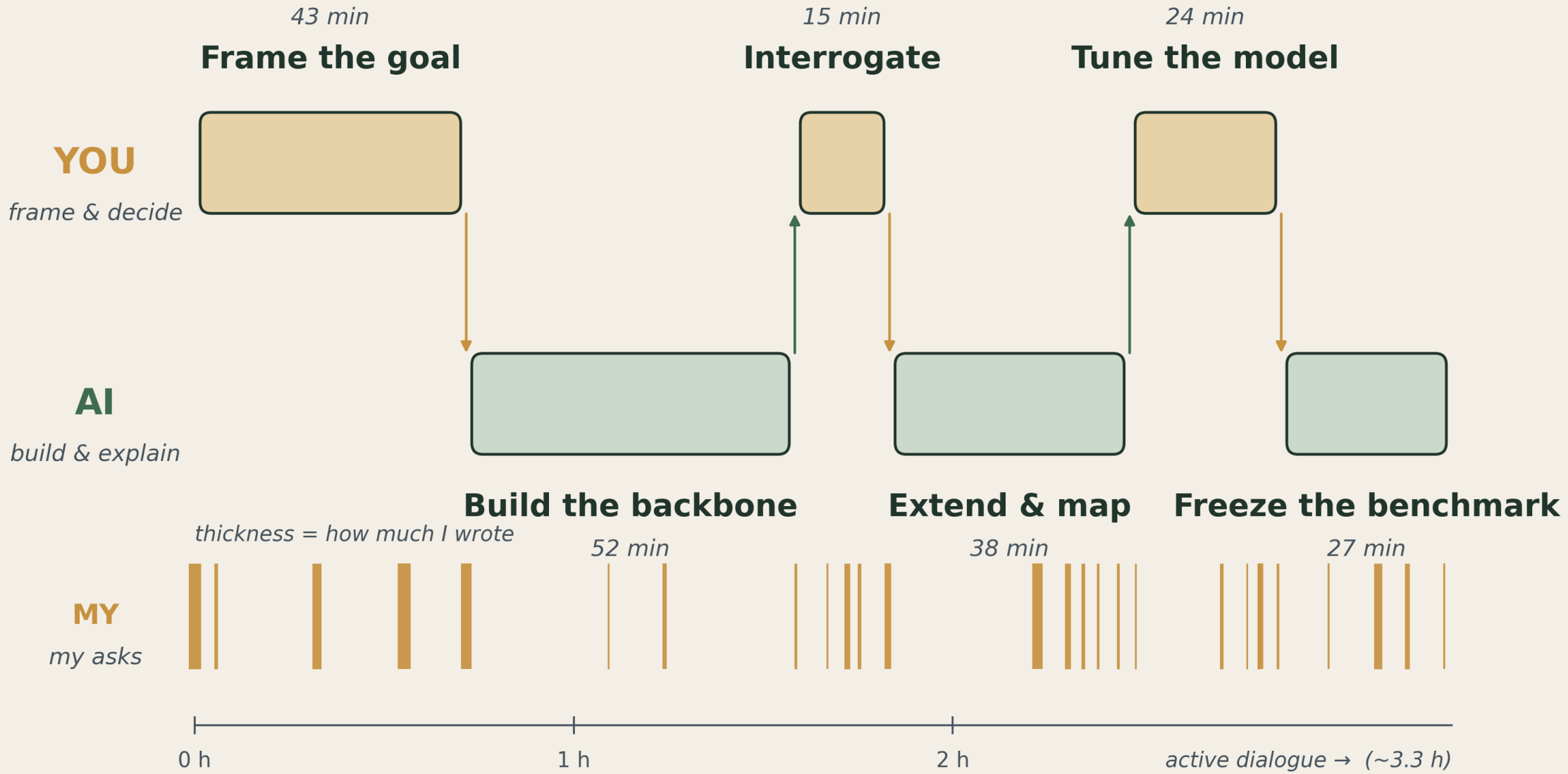
MOBILITY RESULTS

Realistic flows — frozen as a benchmark





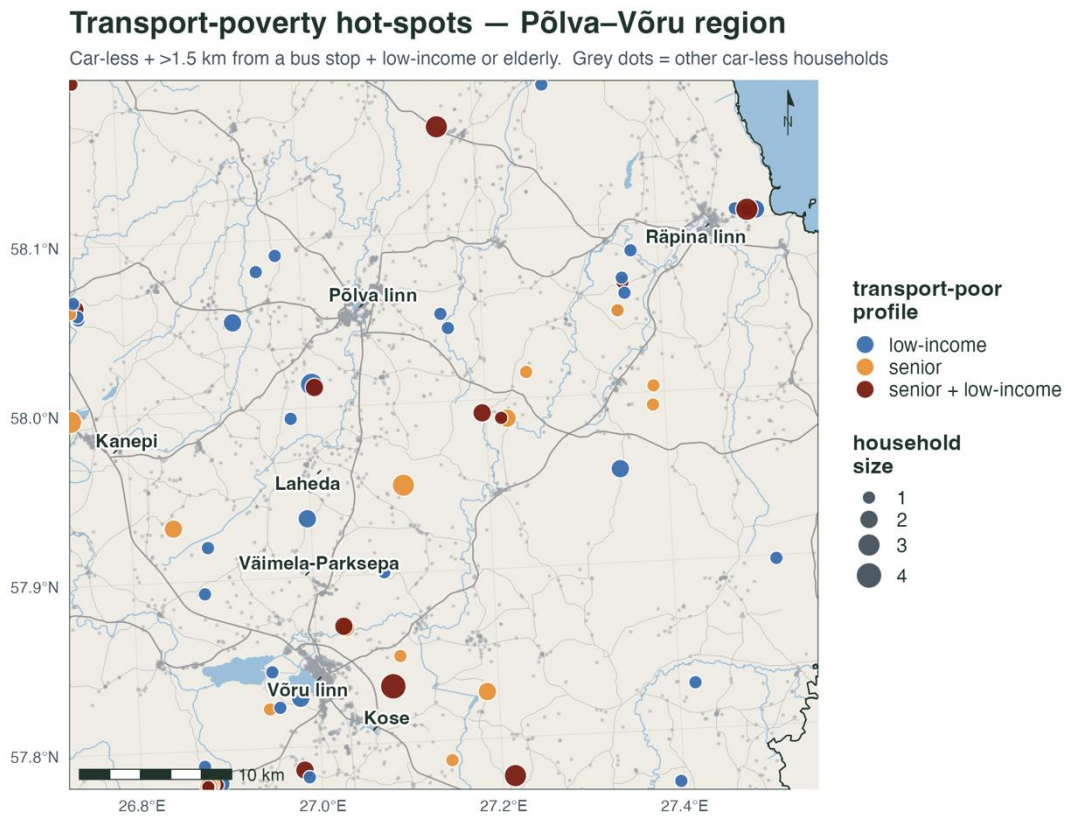
How the commuting layer was built — a dialogue



The model is the residue of a conversation.

PLANNING APPLICATION

Finding transport poverty



What only the joint micro-data shows

5,048 car-less households — the aggregate headline

108 + over 1.5 km from a bus stop

68 + low-income / elderly = transport-poor

→ the car-less headline overstates the target 74×

Car-less rate: town 27.7% ≈ country 25.9% — but transport poverty is 100% rural.

Aggregates give 5,000 car-less; the micro-data finds the 68 who are actually stranded — and where.

UNIVERSITY OF TARTU
Department of Geography

A building-level synthetic population for a whole country — now in motion, from open data alone.

Is it secret? Is it safe?

Both — through open data and k ≥ 3 anonymization.

Anto Aasa · Mobility Lab · University of Tartu · anto.aasa@ut.ee