ETH zürich

MIE lab

# 1st ACM SIGSPATIAL International Workshop on Reproducibility in Tracking Data Analysis and Mobility Research

Mobility Information Engineering Lab

13. November 2023

# Introduction

# Organizers

MIE lab

**Mobility Information Engineering Lab**, ETH Zürich

Nina Wiedemann

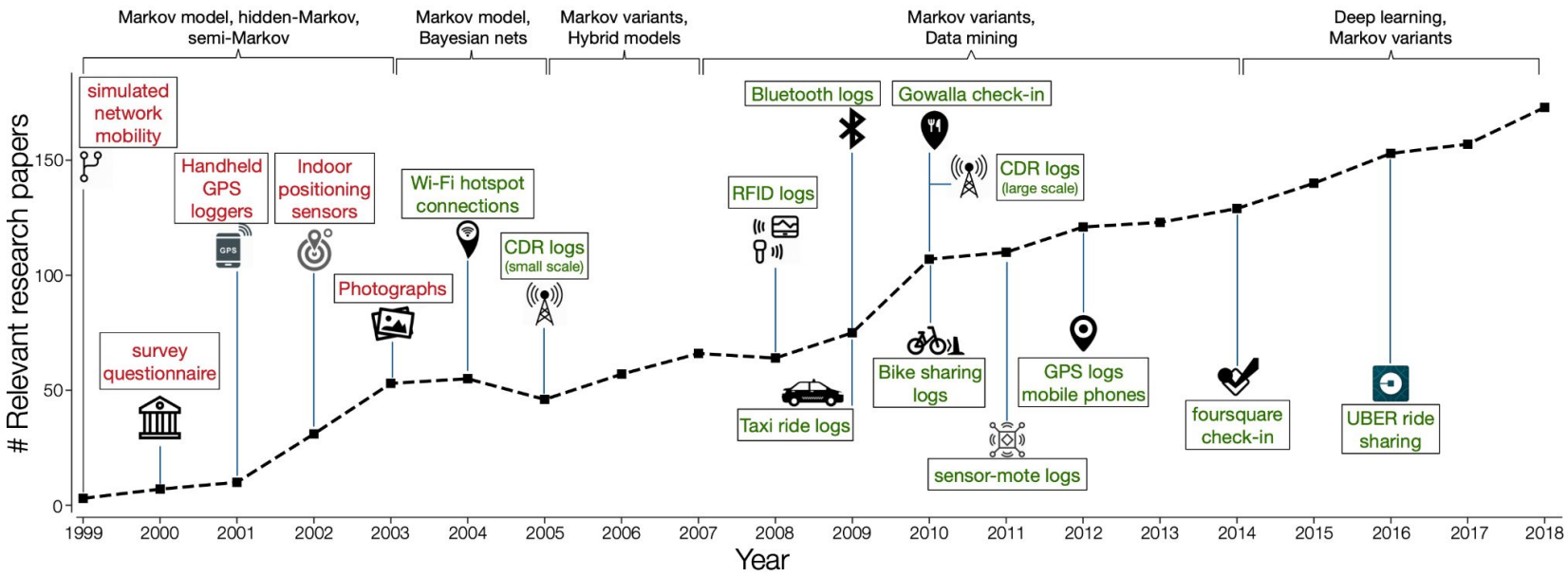
Henry Martin

Ye Hong

Martin Raubal

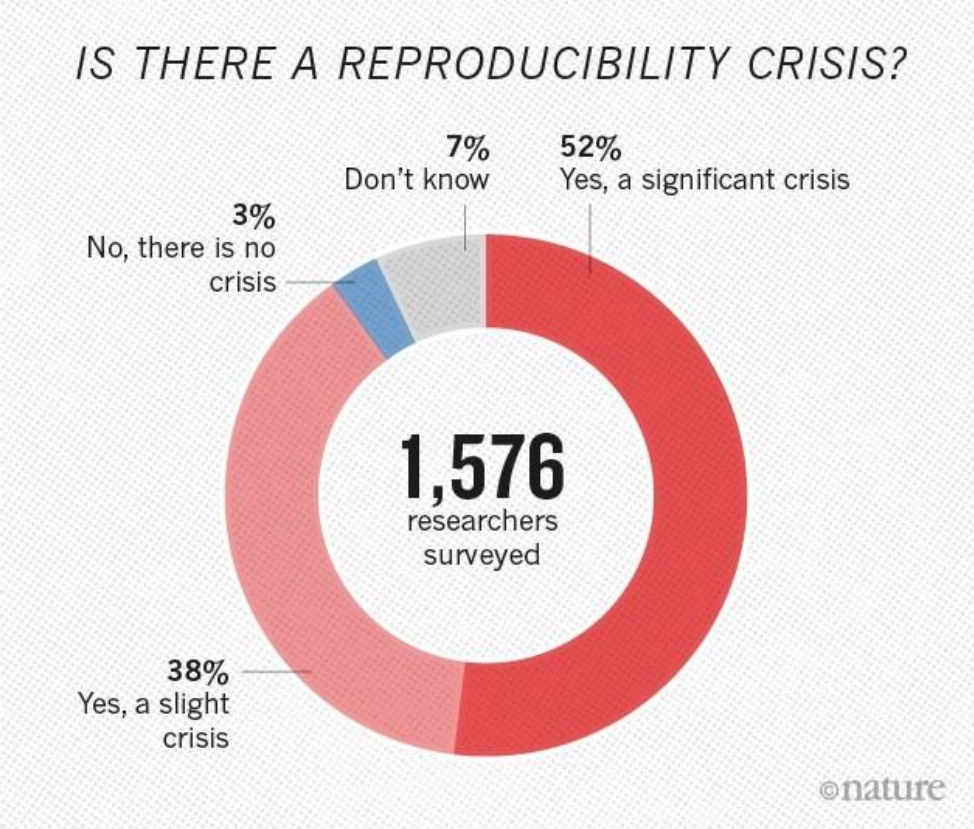
**Who are you?**

# Increasing number of papers in human mobility research




Kulkarni, Vaibhav, and Benoît Garbinato. "20 Years of Mobility Modeling & Prediction: Trends, Shortcomings & Perspectives." *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2019.

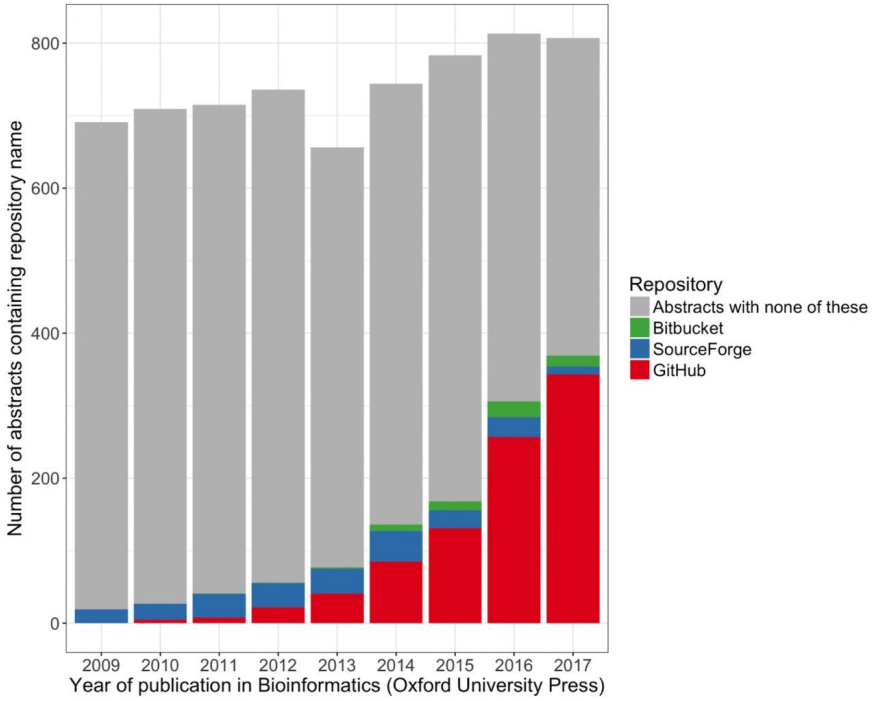
# The reproducibility crisis



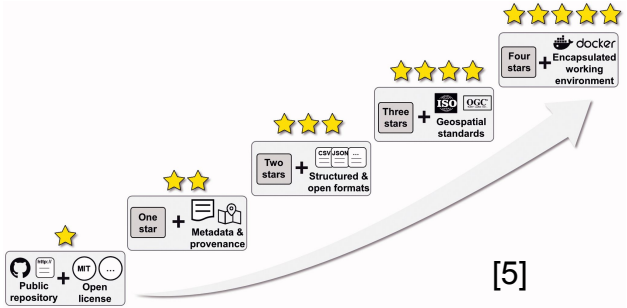

41586_2016_BF533452a_Fig1_HTML.jpg

# Efforts towards reproducibility

- Open-source code on GitHub
- Public datasets and benchmarks
- Open-source libraries



# Efforts towards reproducibility in spatial data science




**Publications on reproducibility**

[1] Goodchild, M. F. (2022). Reproducibility and replicability in spatial science. In *Handbook of Spatial Analysis in the Social Sciences* (pp. 518-527). Edward Elgar Publishing.

[2] Pebesma, Edzer. "Reproducing Spatial Data Science Publications." *Spatial Data Science Symposium* (2023).

[3] Daniel Nüst & Edzer Pebesma (2021) Practical Reproducibility in Geography and Geosciences, *Annals of the American Association of Geographers*, 111:5, 1300-1310

[4] Brunsdon, Chris, and Alexis Comber. "Opening practice: supporting reproducibility and critical spatial data science." *Journal of Geographical Systems* 23.4 (2021): 477-496.

[5] John P. Wilson, Kevin Butler, Song Gao, Yingjie Hu, Wenwen Li & Dawn J. Wright (2021) A Five-Star Guide for Achieving Replicability and Reproducibility When Working with GIS Software and Algorithms, *Annals of the American Association of Geographers*, 111:5, 1311-1317

# Agenda

| | |
|---|---|
| 08:00 | Introduction |
| 08:10 - 08:40 | Keynote by Edzer Pebesma on “Reproducibility in Tracking Data Analysis” |
| 08:40 - 08:55 | Talk by Ville Mäkinen on Geo Privacy platform |
| 09:00 - 10:30 | Trackintel introduction and tutorial |
| 10:30 - 11:00 | Coffee Break |
| 11:00 - 11:45 | Hands-on session on tracking data preprocessing |
| 11:45 - 12:00 | Discussion |

ETH zürich

MIE lab

# Reproducibility in tracking data analysis

# Installation instructions

Clone the trackintel-tutorial repository into a local folder

- `cd your-desired-folder`
- `git clone https://github.com/mie-lab/trackintel-tutorial.git`
- `cd trackintel-tutorial`

Install the reprotrack environment from the provided environment.yml file (with conda)

- `conda env create -f sigspatial23/environment.yml`
- `conda activate reprotrack`

Test the environment configuration

- `python -c "import trackintel; print(trackintel.__version__)"`

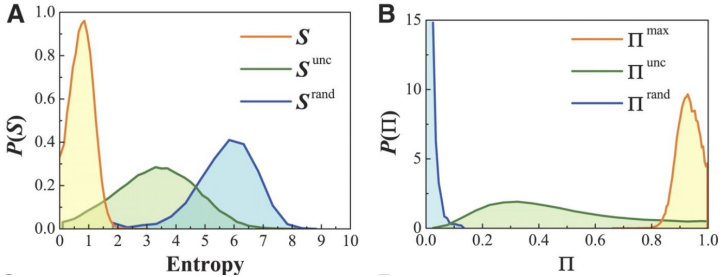
# Human mobility research

➢ **Modelling & understanding human mobility**

**Returners and explorers dichotomy in human mobility**

**The scales of human mobility**

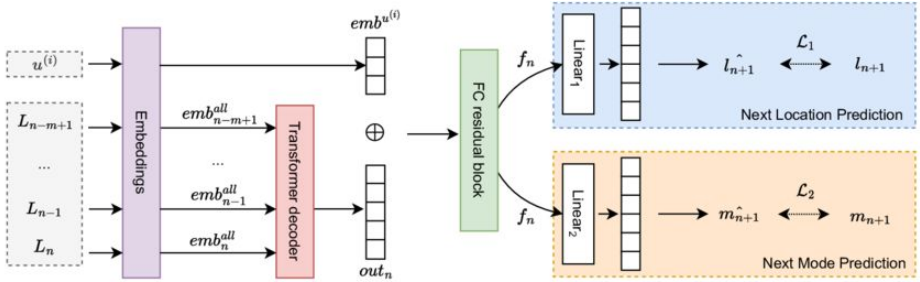
**Limits of Predictability in Human Mobility**



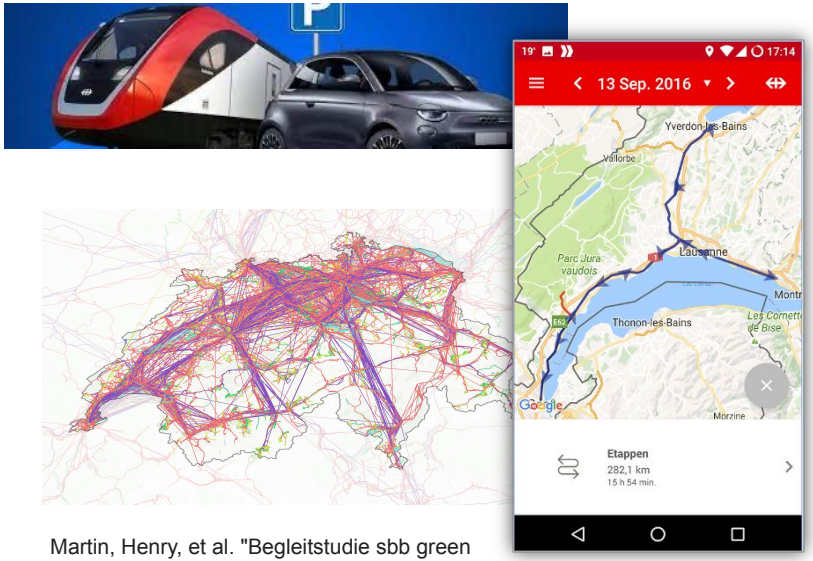
# Human mobility research

➢ **Predicting & analyzing human mobility for applications (e.g. for transportation)**

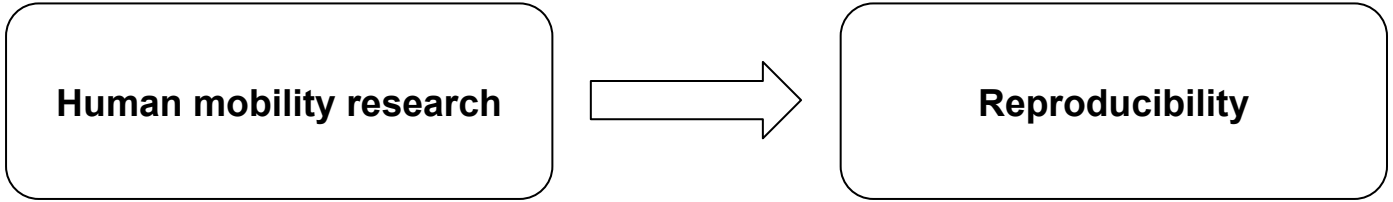
*Next place prediction*




Hong, Ye, Henry Martin, and Martin Raubal. "How do you go where? improving next location prediction by learning travel mode information using transformers." *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 2022.




Martin, Henry, et al. "Begleitstudie sbb green class-abschlussbericht." *Arbeitsberichte Verkehrs-und Raumplanung* 1439 (2019).

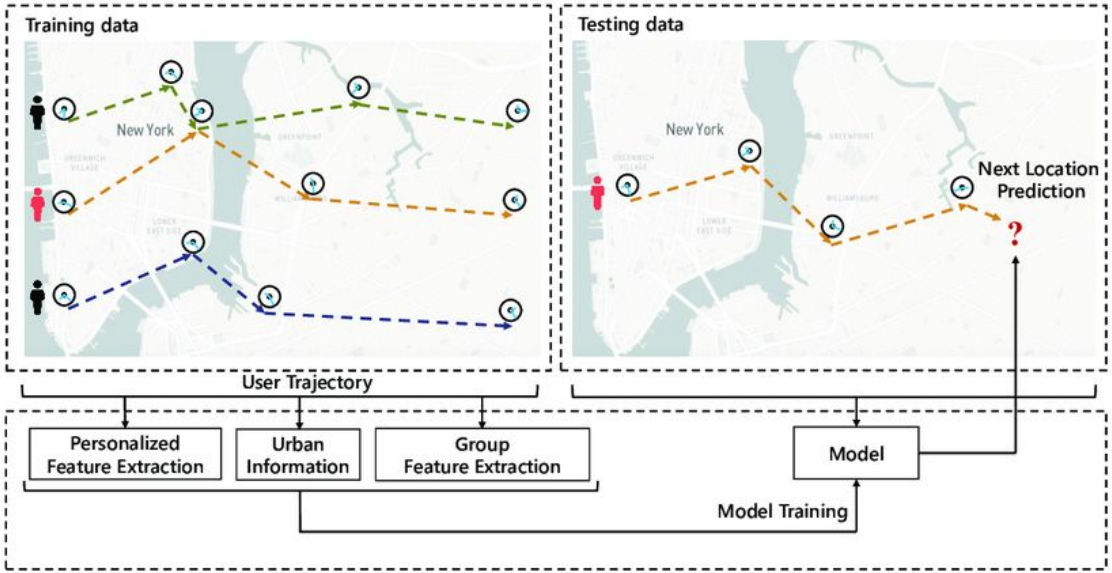

Human mobility research
Reproducibility

# Challenge 1: Inconsistent terminology

- Different methodologies developed in different fields, e.g. transportation vs mobility research [1]
  - Different terms are used to describe the same concept
  - The same term refers to different concepts in different fields

[1] Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation research part C: emerging technologies, 68, 285-299.

# Example: next location prediction




Li, Huifeng, et al. "PG^2 Net: Personalized and Group Preferences Guided Network for Next Place Prediction." *arXiv preprint arXiv:2110.08266* (2021).

What is a location?

→ a (GPS) track point

Feng, J., Y. Li, and C. Zhang. "Predicting human mobility with attentional recurrent networks." *Proceedings of the 2018 World Wide Web Conference on World Wide Web*.

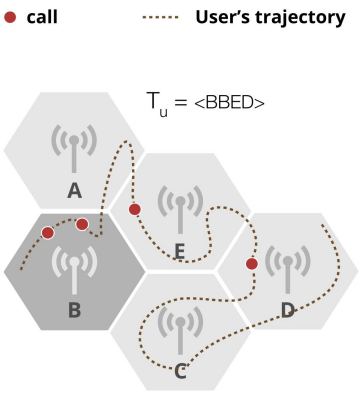
→ a cluster of points (k-means)

Urner, Jorim, et al. "Assessing the influence of spatio-temporal context for next place prediction using different machine learning approaches." *ISPRS International Journal of Geo-Information* 7.5 (2018): 166.

→ a cluster of points (mean-shift)

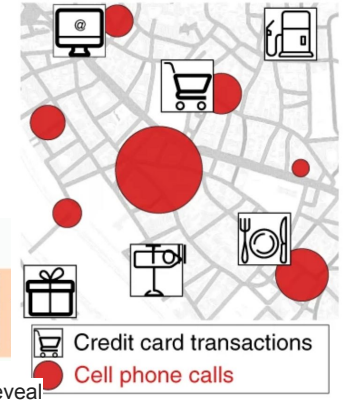
Solomon, Adir, et al. "Analyzing movement predictability using human attributes and behavioral patterns." *Computers, Environment and Urban Systems* 87 (2021): 101596.

# Challenge 2: diverse data sources and types

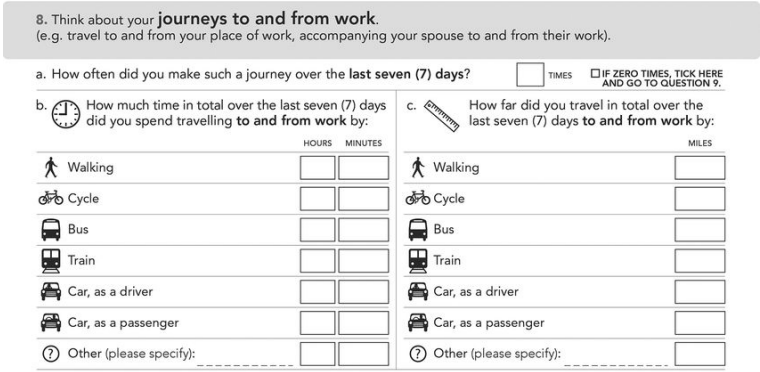



**CDRs**

Pappalardo, Luca, et al. "Evaluation of home detection algorithms on mobile phone data using individual-level ground truth." *EPJ data science* 10.1 (2021): 29.




Di Clemente, Riccardo, et al. "Sequences of purchases in credit card data reveal lifestyles in urban populations." *Nature communications* 9.1 (2018): 3330.



8. Think about your **journeys to and from work**.
(e.g. travel to and from your place of work, accompanying your spouse to and from their work).

a. How often did you make such a journey over the **last seven (7) days**? ☐ TIMES ☐ IF ZERO TIMES, TICK HERE AND GO TO QUESTION 9.

b. How much time in total over the last seven (7) days did you spend travelling **to and from work** by:

| | HOURS | MINUTES |
|---|---|---|
| Walking | | |
| Cycle | | |
| Bus | | |
| Train | | |
| Car, as a driver | | |
| Car, as a passenger | | |
| Other (please specify): | | |

c. How far did you travel in total over the last seven (7) days **to and from work** by:

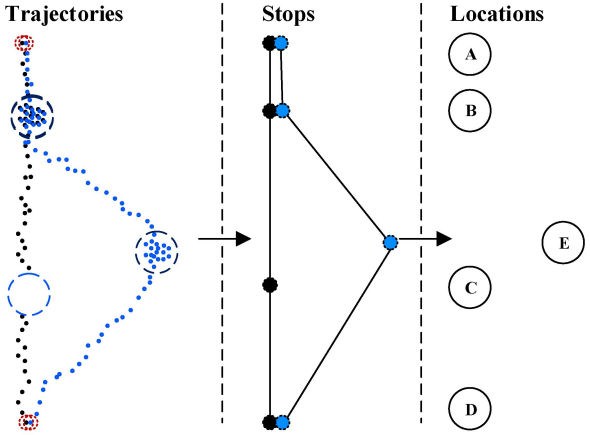
| | MILES |
|---|---|
| Walking | |
| Cycle | |
| Bus | |
| Train | |
| Car, as a driver | |
| Car, as a passenger | |
| Other (please specify): | |

Song, Yena, John M. Preston, and Christian Brand. "What explains active travel behaviour? Evidence from case studies in the UK." *Environment and Planning A* 45.12 (2013): 2980-2998.

# Challenge 3: Intransparent preprocessing

- Tracking data is inherently noisy
  - GNSS inaccuracy
  - Loss of signal → missing data
  - Generally low tracking coverage

- Many parameters that sometimes are not reported, e.g.
  - distance threshold before detecting a new activity
  - minimum duration of missing records before labelling as a gap




Fu, Zhongliang, et al. "A two-step clustering approach to extract locations from individual GPS trajectory data." *ISPRS International Journal of Geo-Information* 5.10 (2016): 166.

Question:

During your research projects, have you tried to reproduce prior work, and did you experience any difficulties?

# Existing software

- General tools and libraries such as ArcGIS and Geopandas
- Tracking data processing:
  - Movingpandas
  - PyMove
  - Tracktable
  - Traja
  - MEOS
  - MoveTK
  - **Trackintel**
- Mobility analysis:
  - Scikit-mobility
  - MovinPy
  - HuMobi

ETH zürich

# Trackintel

# Tracking data

Large-scale datasets recording human digital traces

e.g., from taxi rides

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| id | latitude | longitude | elevation | tracked_at | user_id |
| 0 | 39.984702 | 116.318417 | 149.9616 | 2008-10-23 02:53:04+00:00 | 0 |
| 1 | 39.984683 | 116.31845 | 149.9616 | 2008-10-23 02:53:10+00:00 | 0 |
| 2 | 39.984686 | 116.318417 | 149.9616 | 2008-10-23 02:53:15+00:00 | 0 |
| 3 | 39.984688 | 116.318385 | 149.9616 | 2008-10-23 02:53:20+00:00 | 0 |
| 4 | 39.984655 | 116.318263 | 149.9616 | 2008-10-23 02:53:25+00:00 | 0 |
| 5 | 39.984611 | 116.318026 | 150.2664 | 2008-10-23 02:53:30+00:00 | 0 |
| 6 | 39.984608 | 116.317761 | 150.2664 | 2008-10-23 02:53:35+00:00 | 0 |
| 7 | 39.984563 | 116.317517 | 151.1808 | 2008-10-23 02:53:40+00:00 | 0 |
| 8 | 39.984539 | 116.317294 | 152.4 | 2008-10-23 02:53:45+00:00 | 0 |
| 9 | 39.984606 | 116.317065 | 153.924 | 2008-10-23 02:53:50+00:00 | 0 |
| 10 | 39.984568 | 116.316911 | 155.448 | 2008-10-23 02:53:55+00:00 | 0 |
| 11 | 39.984586 | 116.316716 | 156.972 | 2008-10-23 02:54:00+00:00 | 0 |
| 12 | 39.984561 | 116.316527 | 158.496 | 2008-10-23 02:54:05+00:00 | 0 |
| 13 | 39.984536 | 116.316354 | 160.02 | 2008-10-23 02:54:10+00:00 | 0 |
| 14 | 39.984523 | 116.316188 | 161.8488 | 2008-10-23 02:54:15+00:00 | 0 |

# Travel behaviour and human mobility research

**Travel behaviour research:**

- Traditionally based on travel survey data
- Rigorous definition of movement and activity
- Labour-intensive

**Human mobility analysis:**

- Efficient collection of “big data”
- Usually of poor quality
- No established data model

**Trackintel**

# Trackintel

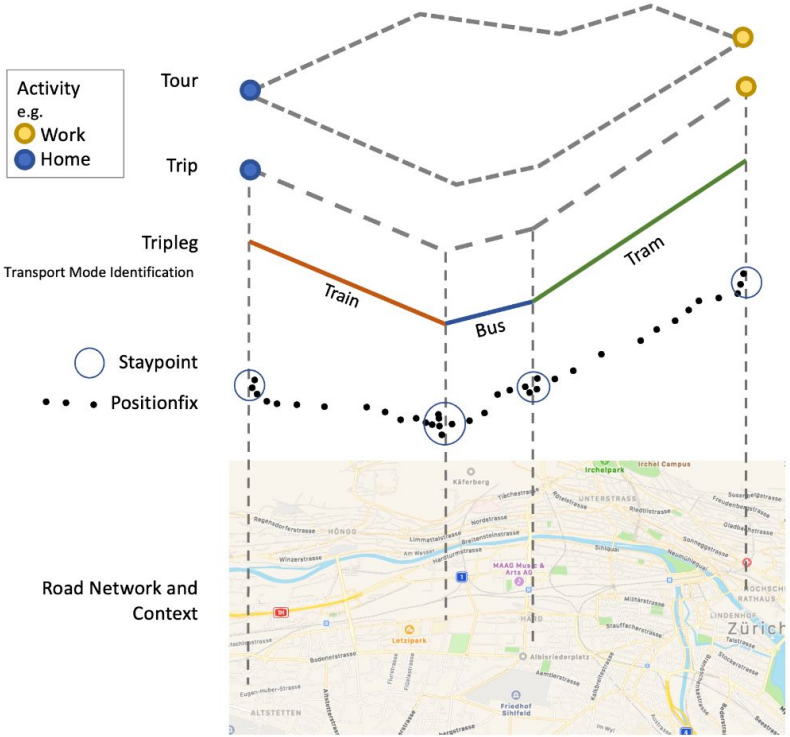
- Focus on individual mobility
- Implement methods to generate movement data models
- Provide functions for semantic enrichment and movement analysis

- Based on Pandas and GeoPandas
- Efficient implementation
- Robust (high test coverage)

## The trackintel framework

pypi package 1.2.2 | conda-forge v1.2.3 | Tests passing | docs passing | codecov 93% | code style black | benchmarked by asv | downloads 34k

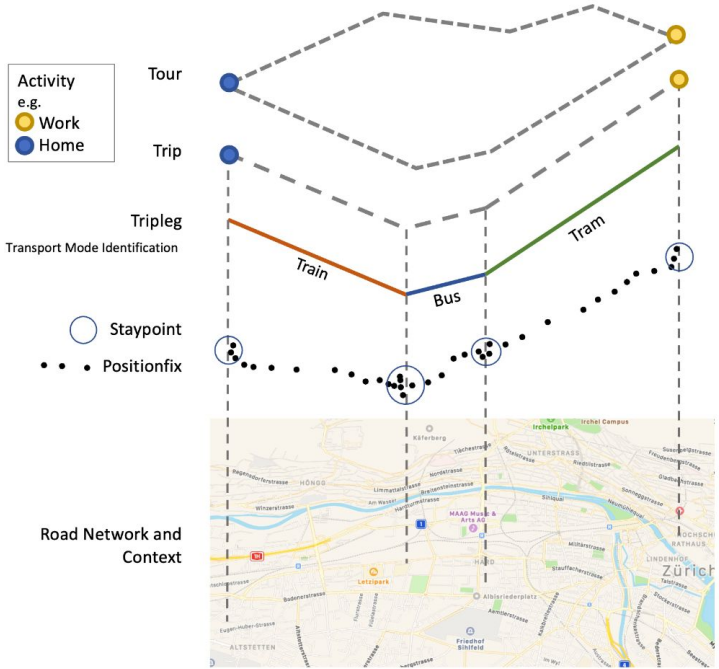
# Trackintel data model




- Established data model in transportation research [1]
- Hierarchy starting with *Positionfixes* (raw trackpoints)

[1] Axhausen, K.W. (2007), "Definition Of Movement and Activity For Transport Modelling", Hensher, D.A, and Button, K.J. (Ed.) *Handbook of Transport Modelling* (, *Vol. 1*), Emerald Group Publishing Limited, Bingley, pp. 329-343.

# Data model: Staypoints & Triplegs

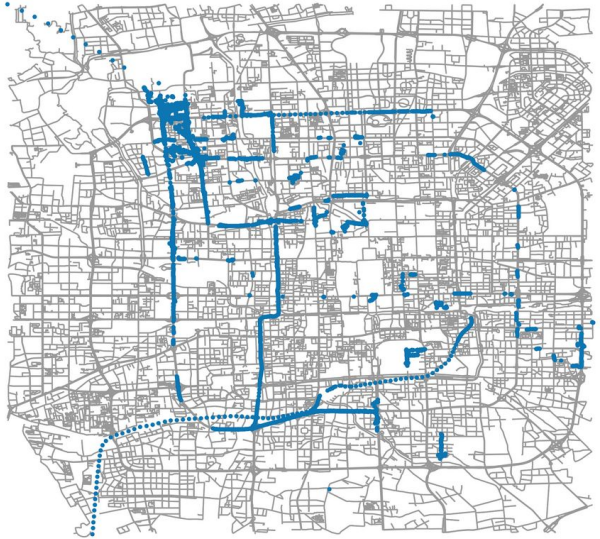



- Staypoints
  *trackintel.preprocessing.positionfixes.generate_staypoints()*
  - Aggregated positionfixes where a user spent time without moving
  - Implement the “sliding” method [1]
- Triplegs
  *trackintel.preprocessing.positionfixes.generate_triplegs()*
  - Continuous movement without changing travel mode, vehicle or stopping for too long
  - Positionfixes that do not belong to a staypoint

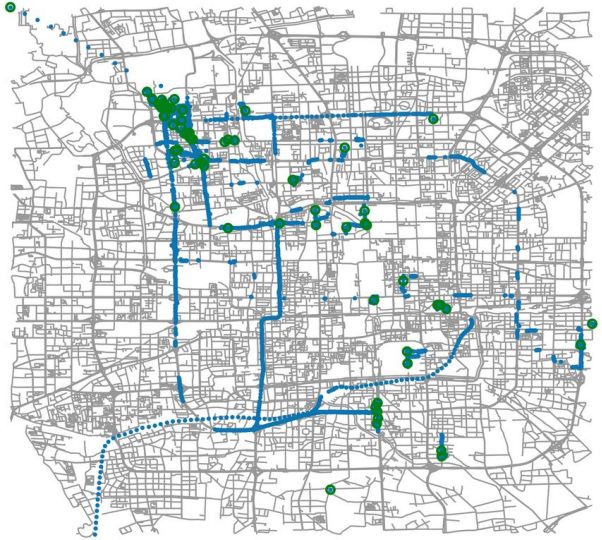
[1] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W. Y. (2008, November). Mining user similarity based on location history. In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems (p. 34). ACM.

# Geolife - from positionfixes to staypoints and triplegs

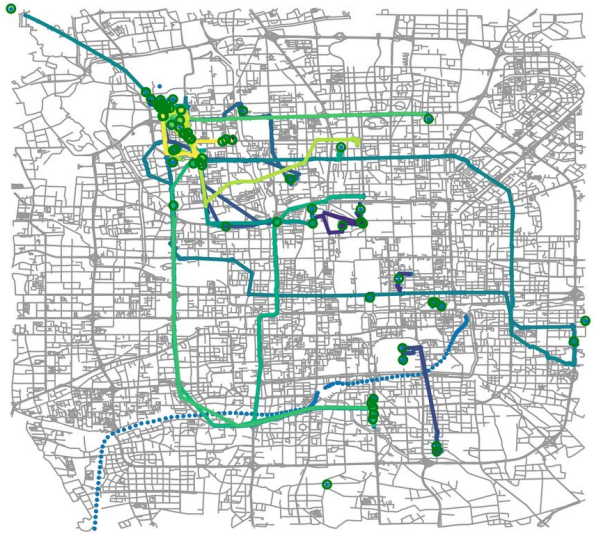
Positionfixes



Staypoints



Triplegs

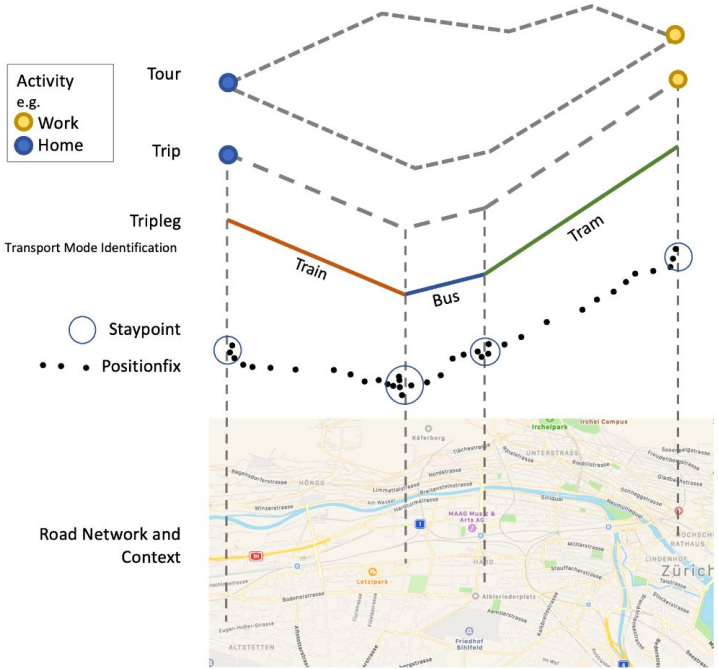


# Data model: Locations

- Locations
  *trackintel.preprocessing.staypoints.generate_locations()*
  - Important places that are visited more than once, e.g., home or work location
  - Spatially aggregated staypoints
  - Adopts the DBSCAN algorithm [1]

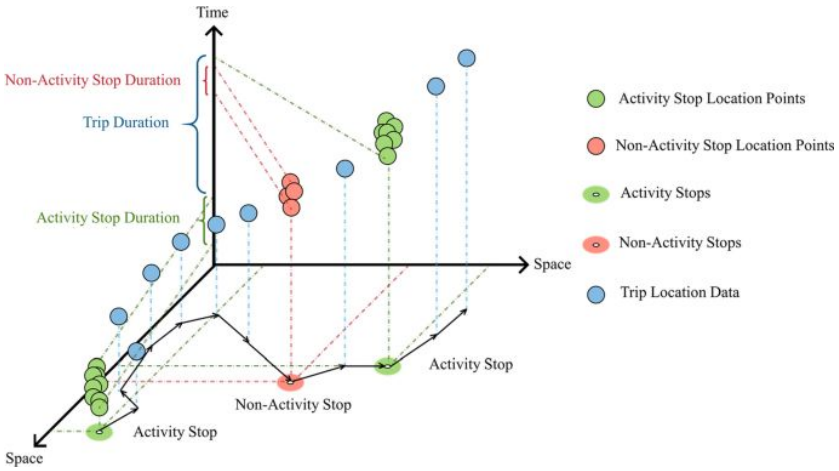
[1] Hariharan, Ramaswamy, and Kentaro Toyama. "Project Lachesis: parsing and modeling location histories." International Conference on Geographic Information Science. Springer, Berlin, Heidelberg, 2004.

# Data model: Trips & Tours

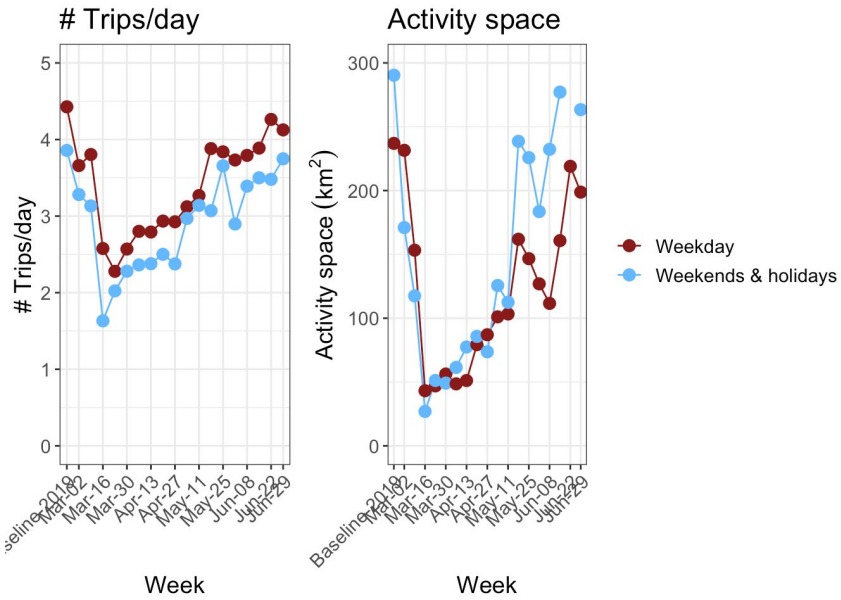



- Trips
  *trackintel.preprocessing.triplegs.generate_trips()*
  - All travels between two activities
- Tours
  *trackintel.preprocessing.trips.generate_tours()*
  - Sequence of trips that starts and ends in the same location
  - Based on duration and distance thresholds
  - Nested tours are possible

# The data model is widely used in transportation research

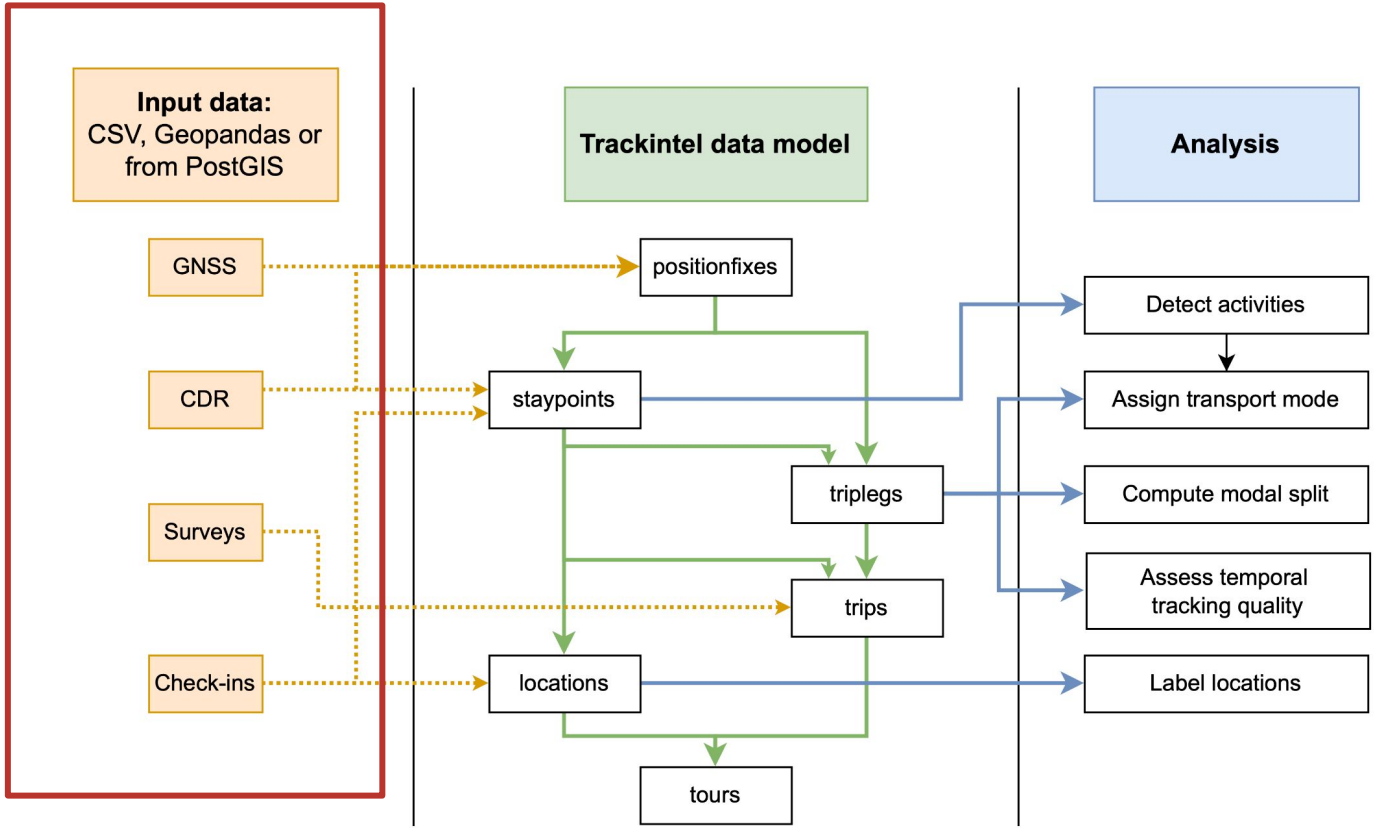



Yang, M., Pan, Y., Darzi, A. *et al.* A data-driven travel mode share estimation framework based on mobile device location data. *Transportation* **49**, 1339–1383 (2022).




Molloy, Joseph, et al. "MOBIS-COVID19/01: results as of 06/04/2020." Arbeitsberichte Verkehrs-und Raumplanung 1489 (2020). (https://ivtmobis.ethz.ch/mobis/covid19/reports/mobis_covid19_report_en_2020-07-06.html)

# Trackintel overview



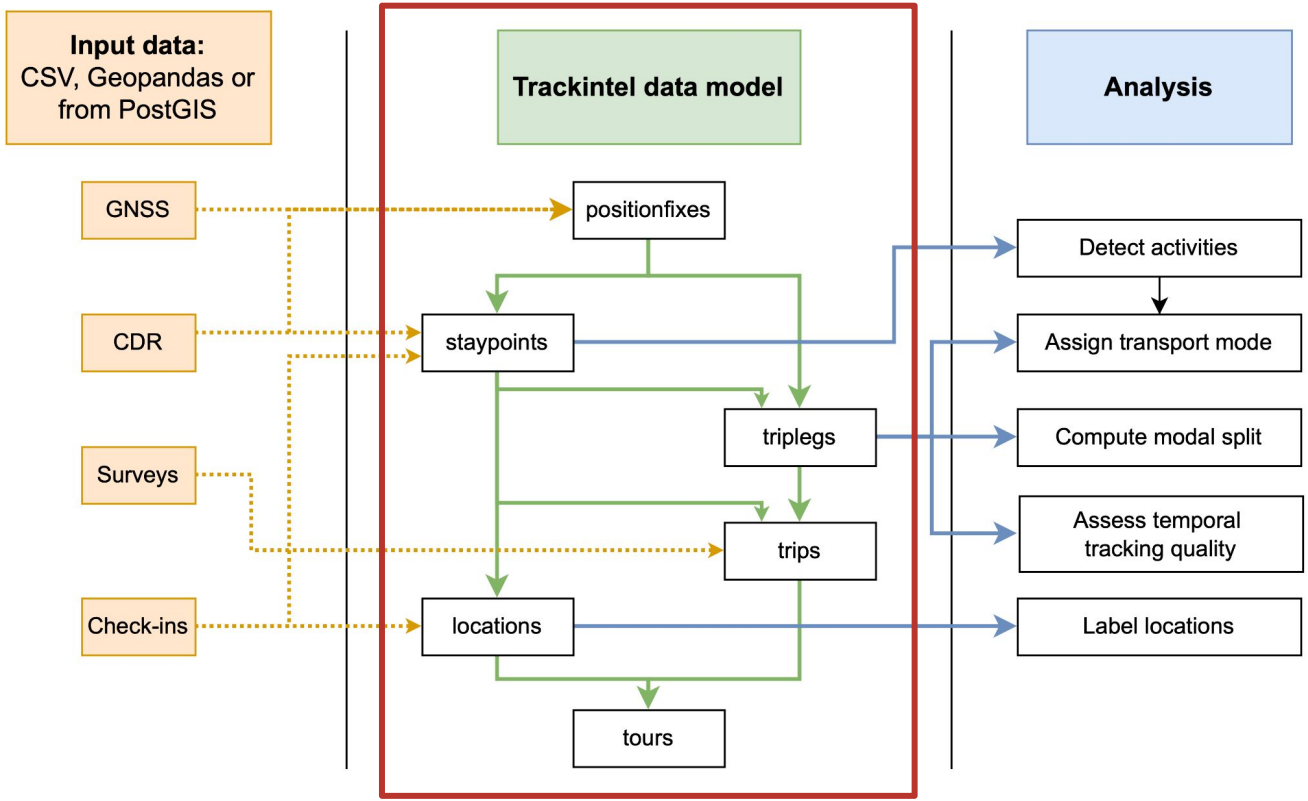
# Data compatibility

- Focus on GNSS data, but was also used with LBSN and travel survey data
- Geopandas-based interface
  - from GeoDataFrame
  - from csv
  - from PostGIS
- Dataset readers
  - Geolife dataset reader
  - gpx file reader
  - Mobility microcensus reader (travel surveys)


```
<trkseg>
 <trkpt lat="47.3109430" lon="9.5068150">
  <ele>1118.0</ele>
  <time>2023-10-22T10:47:48Z</time>
  <extensions>
   <gpxtpx:TrackPointExtension>
    <gpxtpx:atemp>29</gpxtpx:atemp>
   </gpxtpx:TrackPointExtension>
  </extensions>
 </trkpt>
 <trkpt lat="47.3109300" lon="9.5068030">
  <ele>1118.4</ele>
  <time>2023-10-22T10:47:49Z</time>
  <extensions>
   <gpxtpx:TrackPointExtension>
    <gpxtpx:atemp>29</gpxtpx:atemp>
   </gpxtpx:TrackPointExtension>
  </extensions>
 </trkpt>
 <trkpt lat="47.3109090" lon="9.5067320">
  <ele>1118.6</ele>
  <time>2023-10-22T10:47:55Z</time>
  <extensions>
   <gpxtpx:TrackPointExtension>
    <gpxtpx:atemp>29</gpxtpx:atemp>
   </gpxtpx:TrackPointExtension>
  </extensions>
 </trkpt>
```

Taken from: Nawaz, Asif, et al. "GPS trajectory completion using end-to-end bidirectional convolutional recurrent encoder-decoder architecture with attention mechanism." *Sensors* 20.18 (2020): 5143.

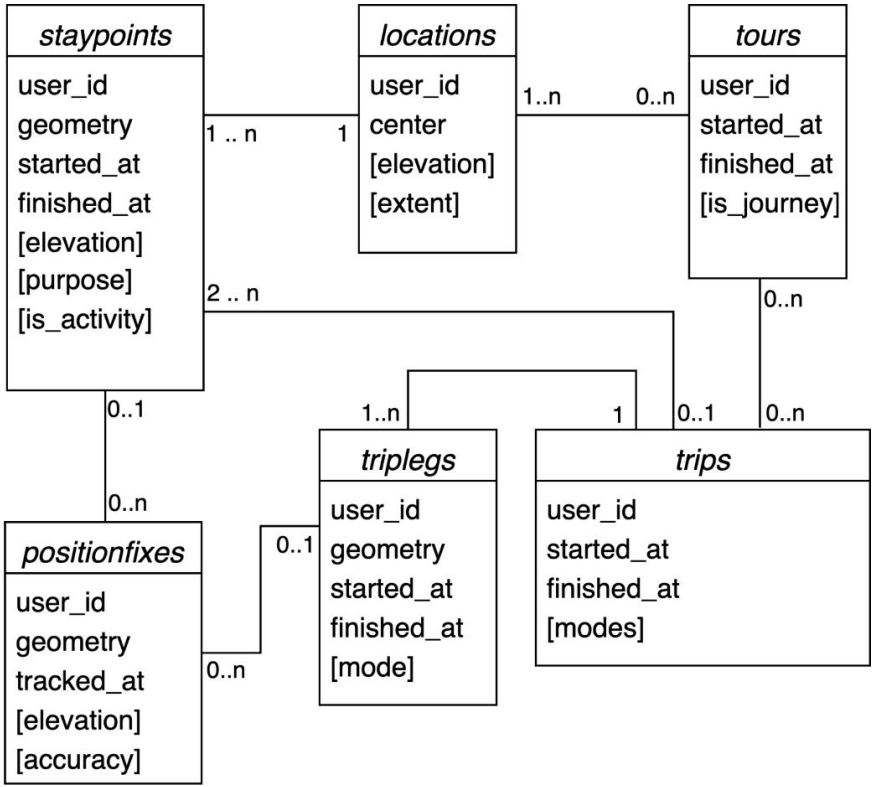
# Trackintel overview



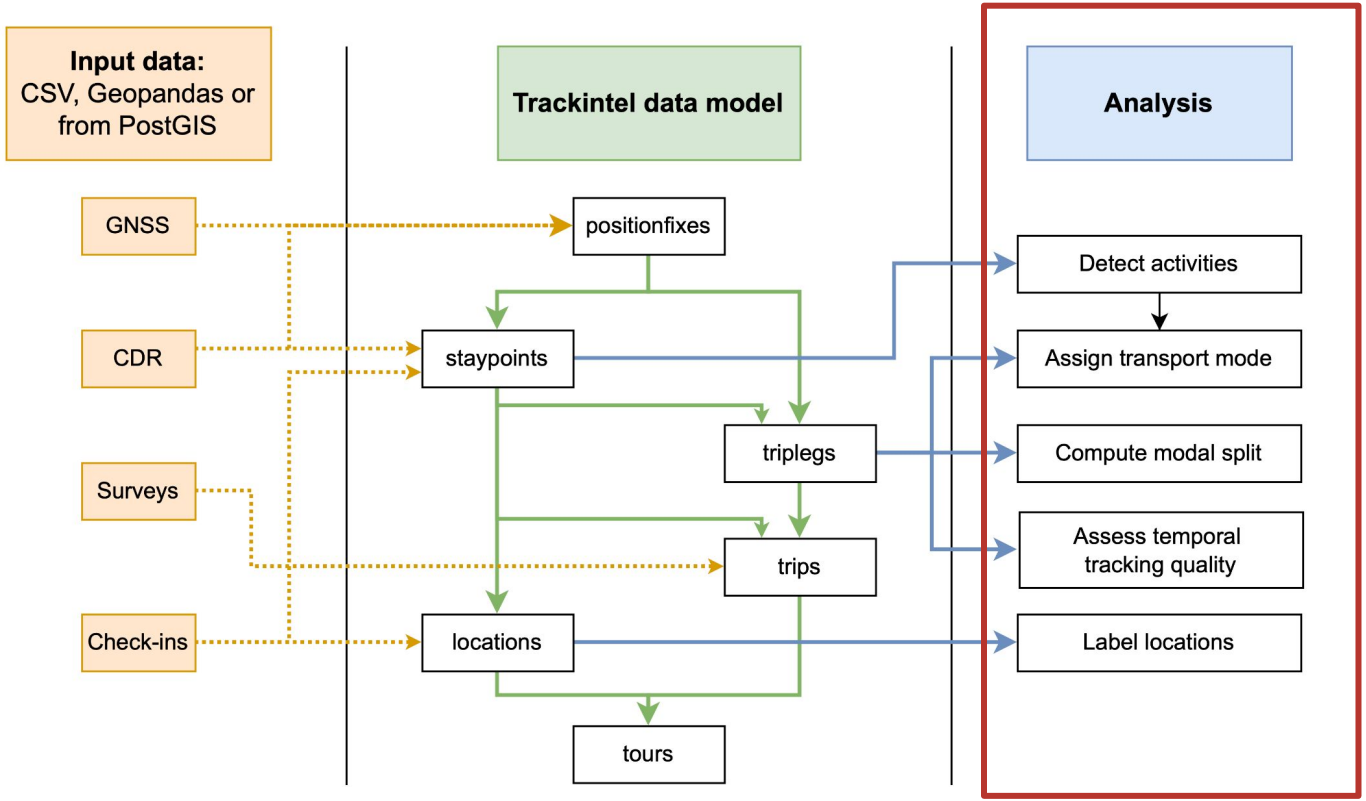
# Implementation in Trackintel

Classes and methods to move up the hierarchy:

- Positionfixes → Staypoints & Triplegs
- Staypoints → Locations
- Staypoint & Triplegs → Trips
- Trips & locations → Tours



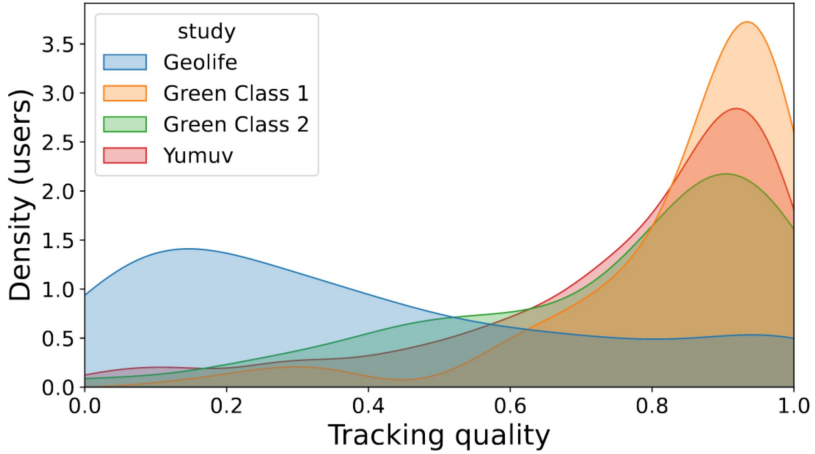
# Trackintel overview



# Analysis & visualization modules

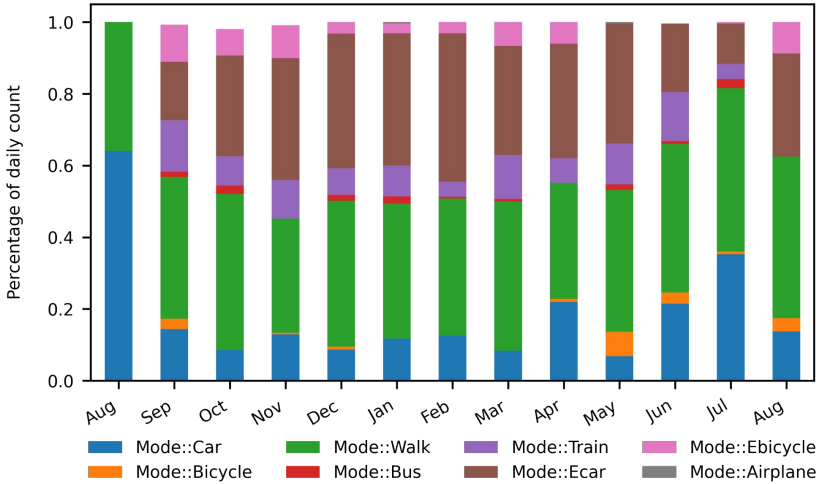
## Temporal tracking data quality

Proportion of time that a user's location is recorded



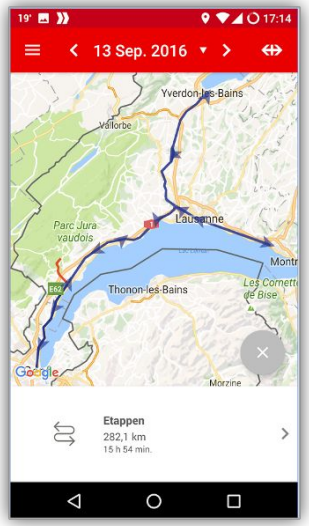

## Modal split

Percentage of travel duration / distance / count by mode

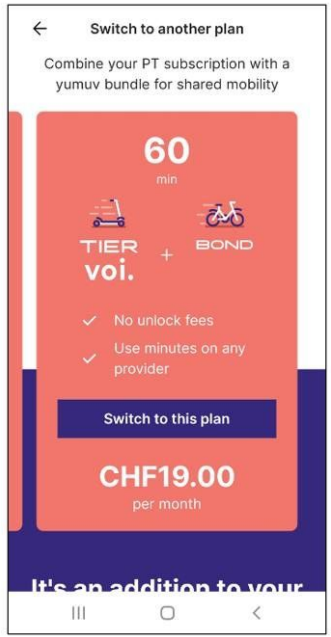


# Trackintel usage in our projects

SBB CFF FFS




yumuv



ETH zürich

# Interactive section

# Tasks overview

- replicate vs reproduce

- 1.Tracking quality assessment
- 2. Travel mode split
- Reproduce mobility scaling laws
  - 3. Jump length
  - 4. Radius of gyration
- 5. Context integration with OSM data

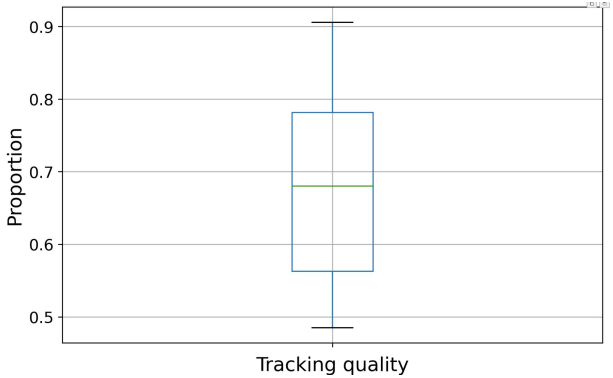
- Possibility to use your own tracking data set!

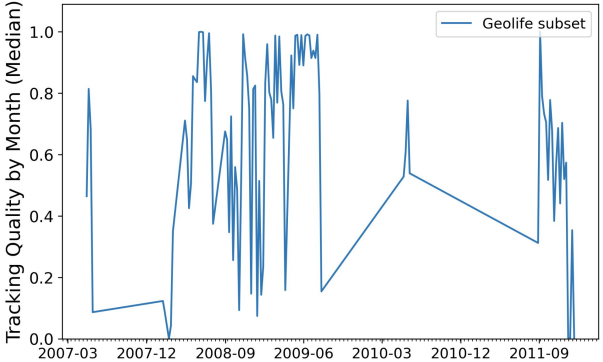
| | | Data | |
|---|---|---|---|
| | | Same | Different |
| Analysis | Same | Reproducible | Replicable |
| | Different | Robust | Generalisable |

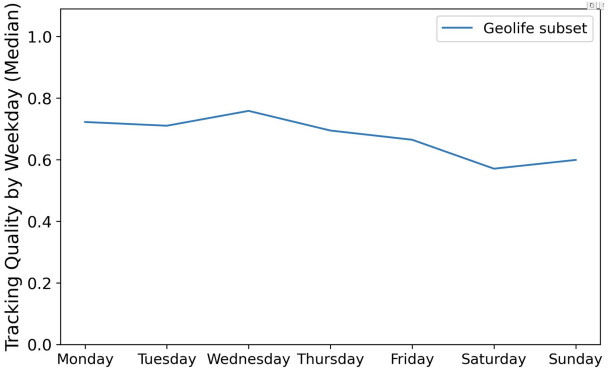
https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html

# Tracking quality - temporal tracking coverage

- The proportion of time the user's whereabouts are recorded
- `trackintel.analysis.temporal_tracking_quality`
  - Can be obtained with different “granularity”
  - "all", "day", "week", "weekday", "hour"

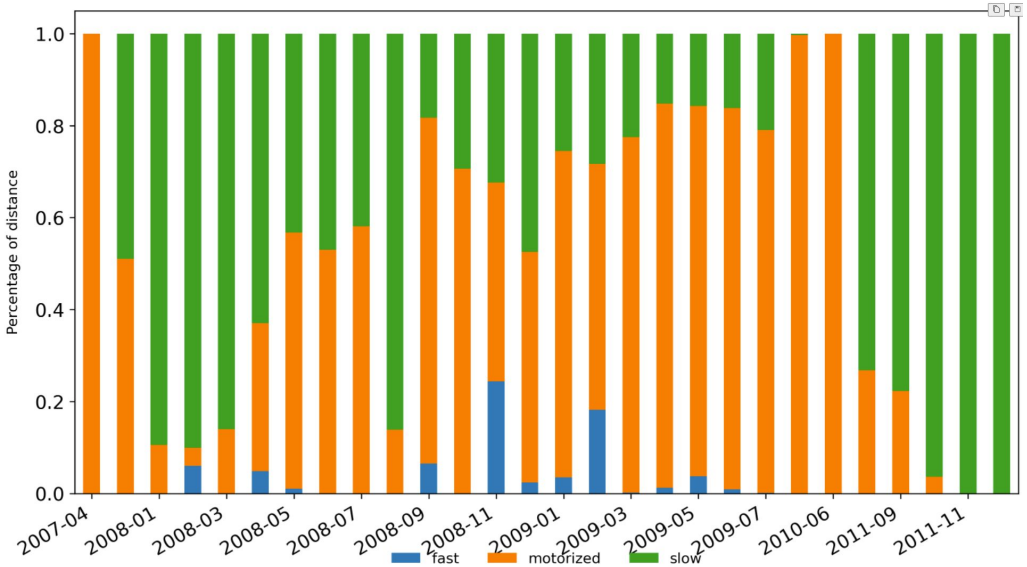


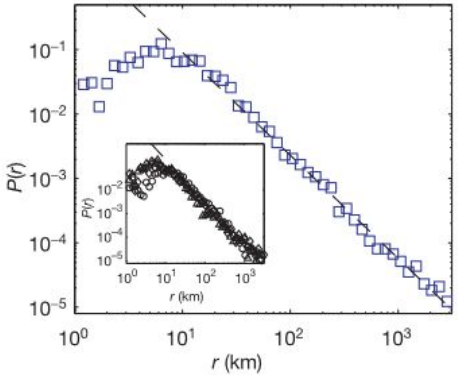
# Travel mode split

- The share of travel mode usage of users over time
- `trackintel.analysis.predict_transport_mode`
  - Mode detection based on speed
- `trackintel.analysis.calculate_modal_split`
  - Mode split for various durations, with criteria {'count', 'distance', 'duration'}
- `trackintel.plot_modal_split`
  - Mode split plotting



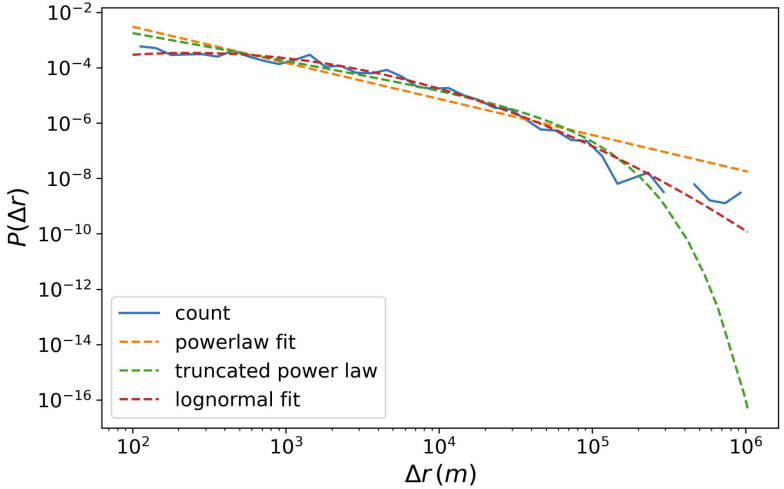
# Metric - Jump length

- Displacements between consecutive staypoints
- Characteristic of human mobility: scale-free random walks (Lévy flight).
- `trackintel.analysis.jump_length`



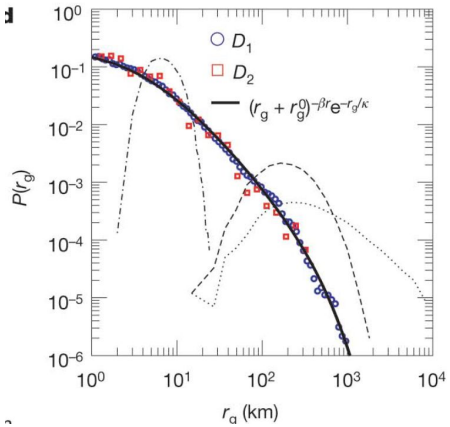

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel.

*Nature*, *439*(7075), 462–465. https://doi.org/10.1038/nature04292

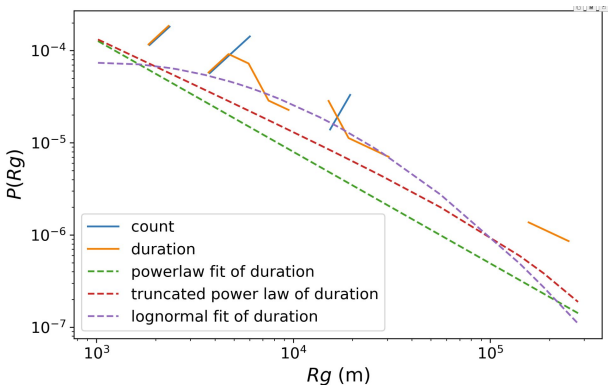


# Metric - Radius of gyration

- Characteristic distance travelled by users
- Showing population heterogeneity in movement
- **`trackintel.analysis.radius_gyration`**




```
user_id
1        2555.886949
2      141198.752644
7       13553.414520
9        1018.344049
11       4754.886298
12     186413.225495
14       8555.789344
16       7210.718648
19       6092.686357
20       1654.699671
22     279212.826582
30      25388.192564
35       2341.146492
39      17311.886386
41      28546.339318
103     15555.883922
112      3547.498514
113      5115.375339
154     55885.293697
169      5419.747608
Name: radius_gyration, dtype: float64
```

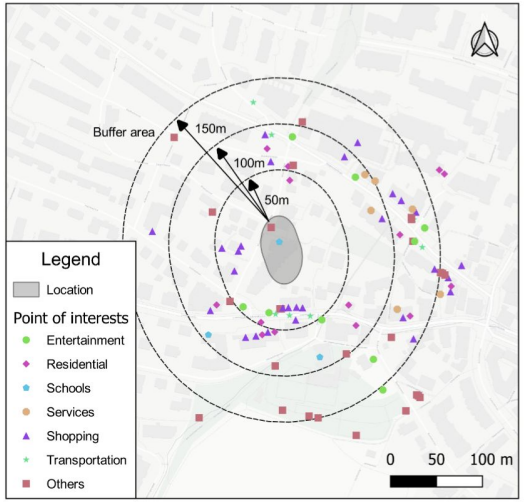


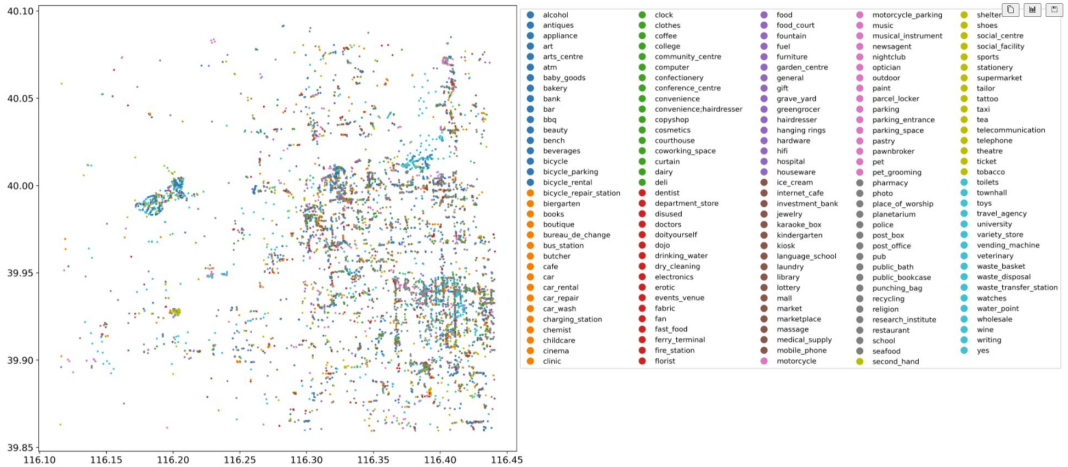

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns.

*Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958

# Analysis - Context integration with OSM data

- Human mobility is content dependent
- Activity prediction or location characterization
- Trackintel can easily integrate built environment context with the help of OSMnx




Hong, Y., Zhang, Y., Schindler, K., & Raubal, M. (2023). Context-aware multi-head self-attentional neural network model for next location prediction. *Transportation Research Part C: Emerging Technologies*, *156*, 104315. https://doi.org/10.1016/j.trc.2023.104315

ETH zürich

ETH zürich

# Discussion

Reproducibility: A technical problem or a problem of the right mindset?

What are important technical barriers for better reproducibility?

- ➢ How can we do replicability studies with privacy sensitive datasets?
- ➢ How to efficiently communicate the preprocessing steps/parameters with the final result?

How can tools such as Trackintel improve reproducibility?

What will you do in the future to make your research more reproducible?

# Contribute to Trackintel

All contributions, bug reports, bug fixes, documentation improvements, enhancements, and ideas are welcome!

## How to contribute

Thank you for your interest in the *trackintel* development. All contributions, bug reports, bug fixes, documentation improvements, enhancements, and ideas are welcome.

### Development

You can download the whole repository and install *trackintel* locally using `pip install .`. For quick testing, use `trackintel.print_version()`.

Testing is done using pytest. Simply run the tests using `pytest` in the top-level trackintel folder. In case you use `pipenv`, install *pytest* first (`pip install pytest`), then run *pytest* using this version: `python -m pytest`. The use of fixtures for data generation (e.g., trips and trackpoints) is still an open todo. As for now, there are some smaller datasets in the `tests` folder.

### Documentation

The documentation follws the pandas resp. numpy docstring standard. In particular, it uses Sphinx to create the documentation. You can install Sphinx using `pip install -U sphinx` or `conda install sphinx`.

If you use additional dependencies during development, do not forget to add them to `autodoc_mock_imports` in `docs/conf.py` for readthedocs.org to work properly.

You can then generate the documentation using `sphinx-build -b html docs docs.gen`. This will put the documentation in `docs.gen`, which is in `.gitignore`.

# Thank you for participating!

And many thanks to

Christof Leutenegger

Dominik Bucher