Smart Data Fusion

Probabilistic Record Linkage adapted to merge two trajectories from different sources

SIDs 2018 - Salzburg

Artificial Intelligence and Machine Learning in aviation...

"I want to build predictive model "

"I need a lot of data" → Not just in volume but also in **variety**

"OK, I will get all this data from others..."

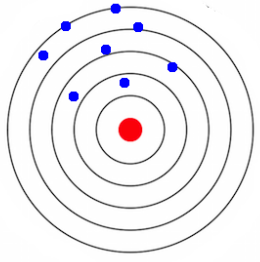
Different independent systems

Lack of consolidation

*Also... no one is sharing data with you because they **don't trust you**...*

"No problem, I will train the algorithms using only my data..."

Your model
is
highly biased
and
incomplete



Your model **performs poorly** and **can't be used**

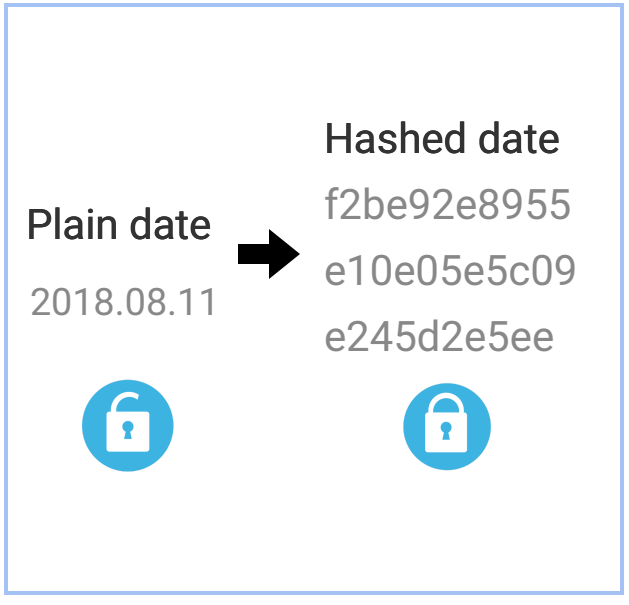
To train robust Machine Learning models...

Different data sources need to be **merged**.

To share data

Data needs to be **protected**

De-Identified datasets



We need **data merging techniques** that are able to match **de-identified data**

SMART DATA FUSION

Deterministic vs. Probabilistic Record Linkage

Record Linkage

Recognising **records which represent the same thing** in two datasets.

Deterministic Record Linkage

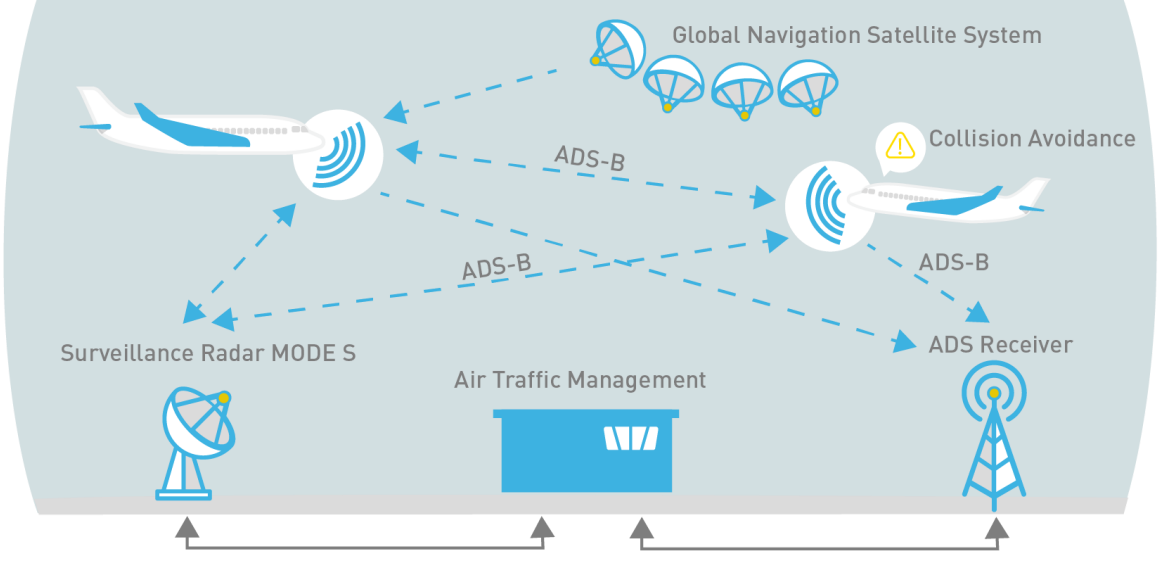
- **Sort-Merge operations** that find the exact match.
- Works only when the identifiers **don't contain errors**.
- Identifiers need to be **always present** in the data.

Probabilistic Record Linkage

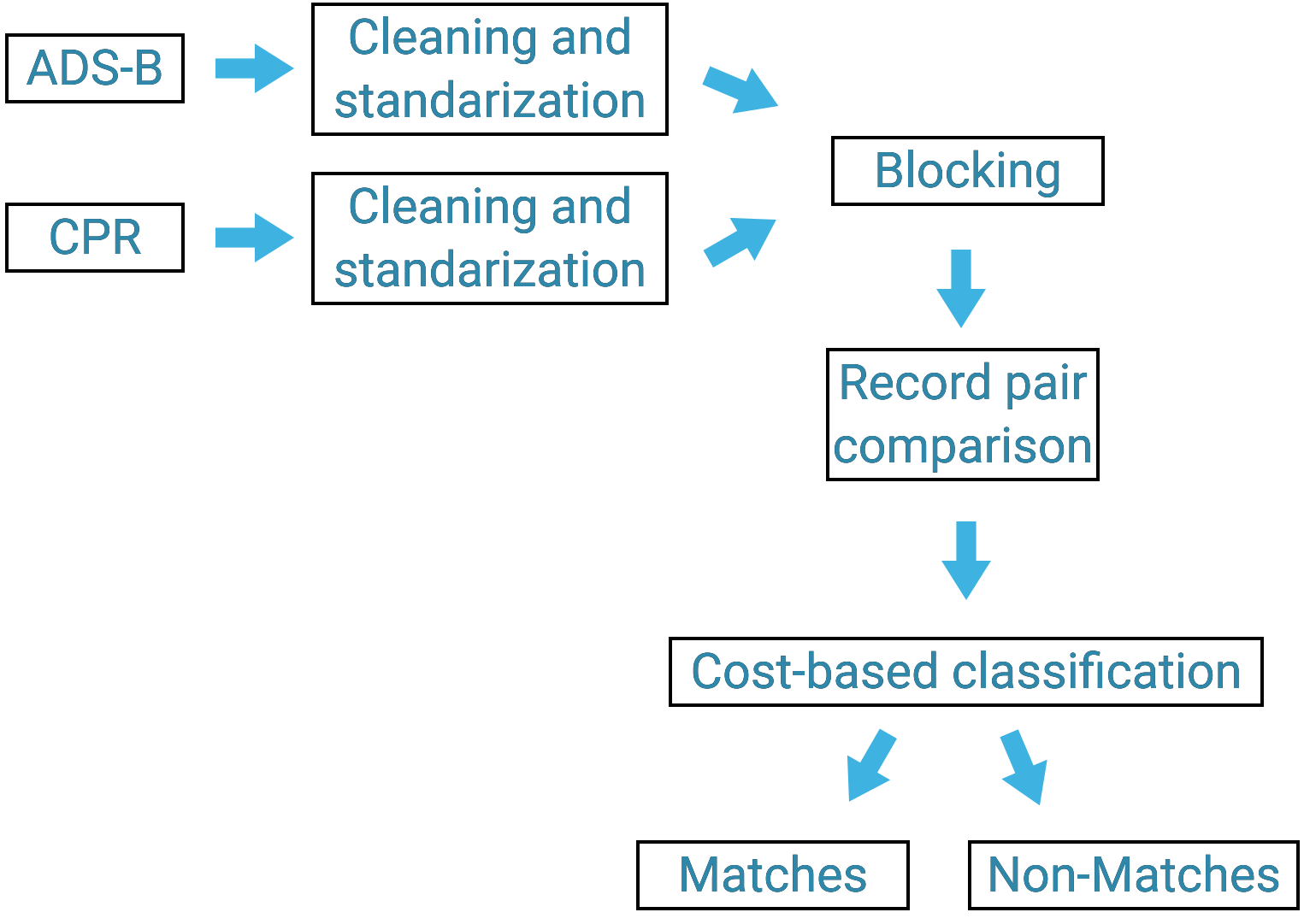
- Matches by comparing a **wider range of identifiers**.
- Computes a **weight** for each identifier.
- Uses the weights to calculate the total **probability** that two given records refer to the same entity.
- Proposed by Fellegi and Sunter (1969).

Probabilistic Record Linkage in aviation datasets

- Aim:
  - Match de-identified flights in two radar sets.
  - Via a probabilistic linkage algorithm.
  - Identifiers: Date (de-id), callsign (de-id), latitude, longitude, flight level
- Data available:
  - 1 day of flights in EU area.
  - **ADS-B:** 32.484 flights
  - **Correlated Position Reports (CPR):** 32.673 flights



Methodology



Dealing with dimensionality

Problem:

1.000 million posible matches for 1 day of flights.

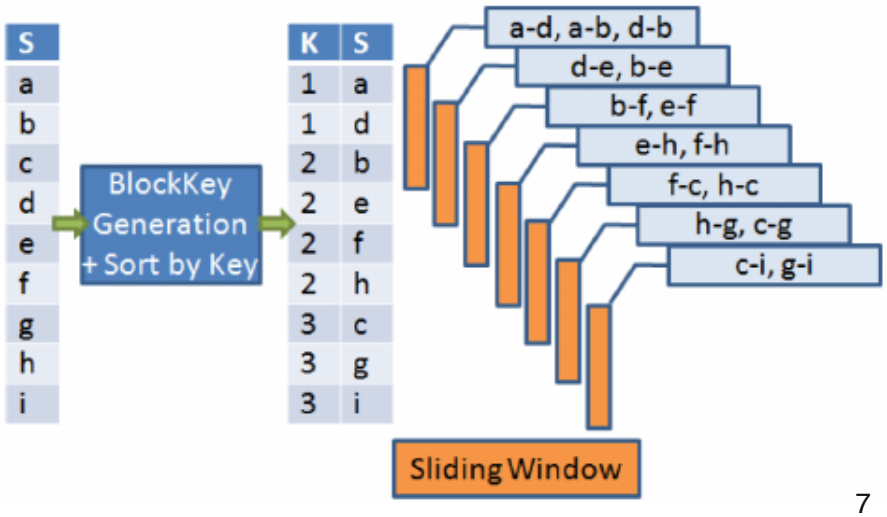
Solution: Blocking

Remove as many records pairs from the subset of non-matches that are obviously non-matches.

Multiple available techniques... But in this case, flights are **ordered by time** (sortable records).

"Sorted neighbourhood" blocking

1. A key for each record is created.
2. Sort the datasets by the created key, i.e. radar time.
3. **Fixed size window** is moved through the sequential list of records, i.e. 2 hours window



Cost-Based probabilistic linkage

1. Cost function

- Misclassification cost for each pair of trajectories.
- **Trajectory distances** as cost estimators.
  - Warping distances: DTW
  - Shape-Based: Haussdorff

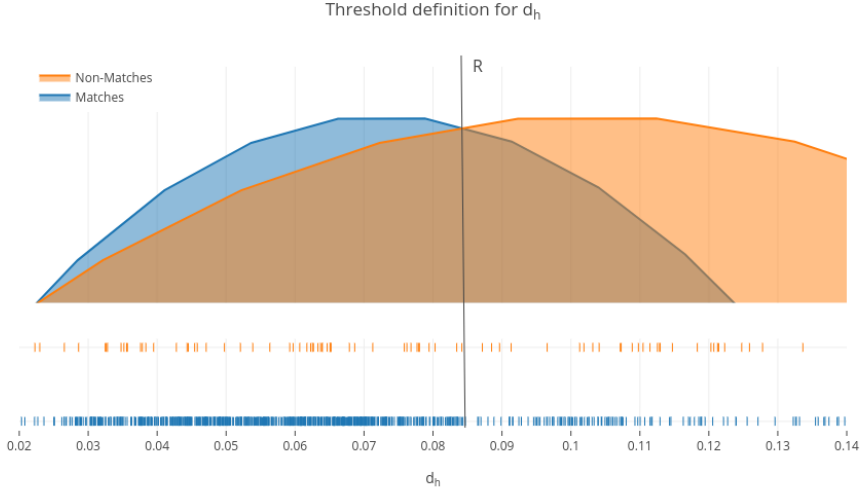
2. Weight computations and aggregation

$$w_1(a,b) = d(a(lon,lat), b(lon,lat))$$
$$w_2(a,b) = d(a(t,lat), b(t,lat))$$
$$w_3(a,b) = d(a(t,lon), b(t,lon))$$

$$c_{ab} = 0.7 \cdot w_1(a,b) + 0.15 \cdot w_2(a,b) + 0.15 \cdot w_3(a,b)$$

3. Threshold selection

- Following the definition of Felligi and Sunter estimating the probabilities require a fully matched dataset...
- Chosen empirically.

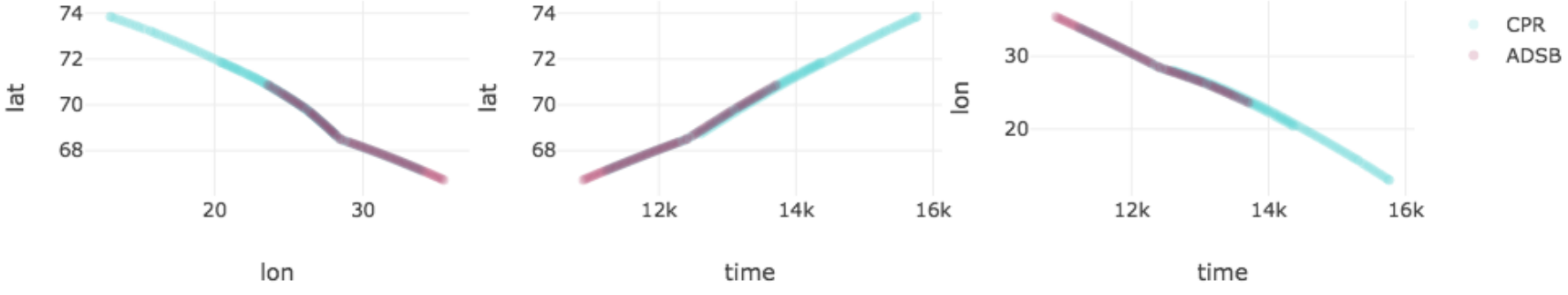


Results

Methodology / Classifier	Matched ADS-B	Matched CPR
Naive Inner Join using callsign as identifier.	39.5%	41.5%
Naive Inner Join using callsign as identifier: DTW	57.3%	59.3%
Cost-Based probabilistic record linkage: Haussdorff	63.7%	66.7%

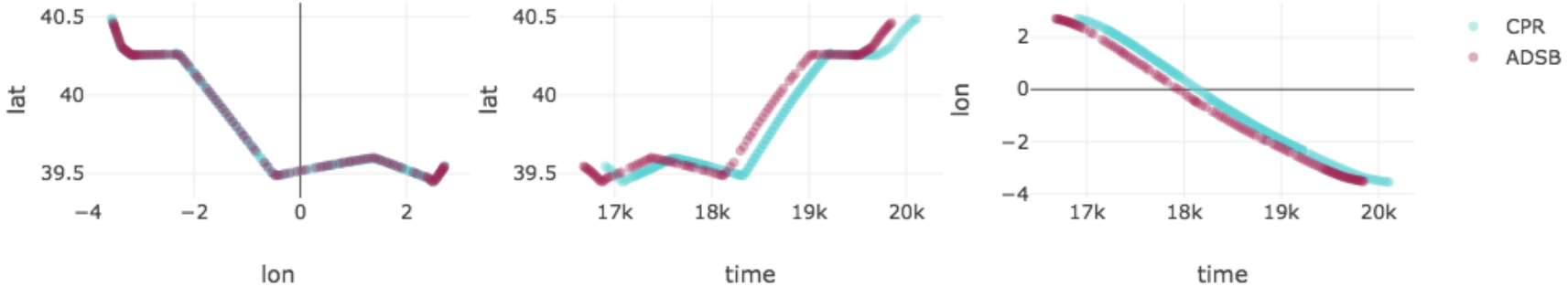
- **Shape-Based distance** performs better.
- Different trajectories may present the same "signal wrapping" parallel in latitude-longitude.
- **79%** matched data that were also matched by the sort-join linkage using the callsign as key.

False negative cases



- Incorrectly classified as non-matches.
- Due to missing data.
- Dut to errors in ADS-B data.
  - Range and frequency congestion of the antenna.
- Can be mitigated by pre-matching using a traditional join-sort methodology.

False positive cases



- Incorrectly classified as matches.
- Small time differences between flights with the same (or very similar) route.
- Further testing on weight distributions is required.
- Mitigating the effect of radar delay is crucial

Thank you!

Darío Martínez

dm@innaxis.org

www.linkedin.com/in/dmartr