Predicting Flight Arrival Times with a Multistage Model

Gábor Takács

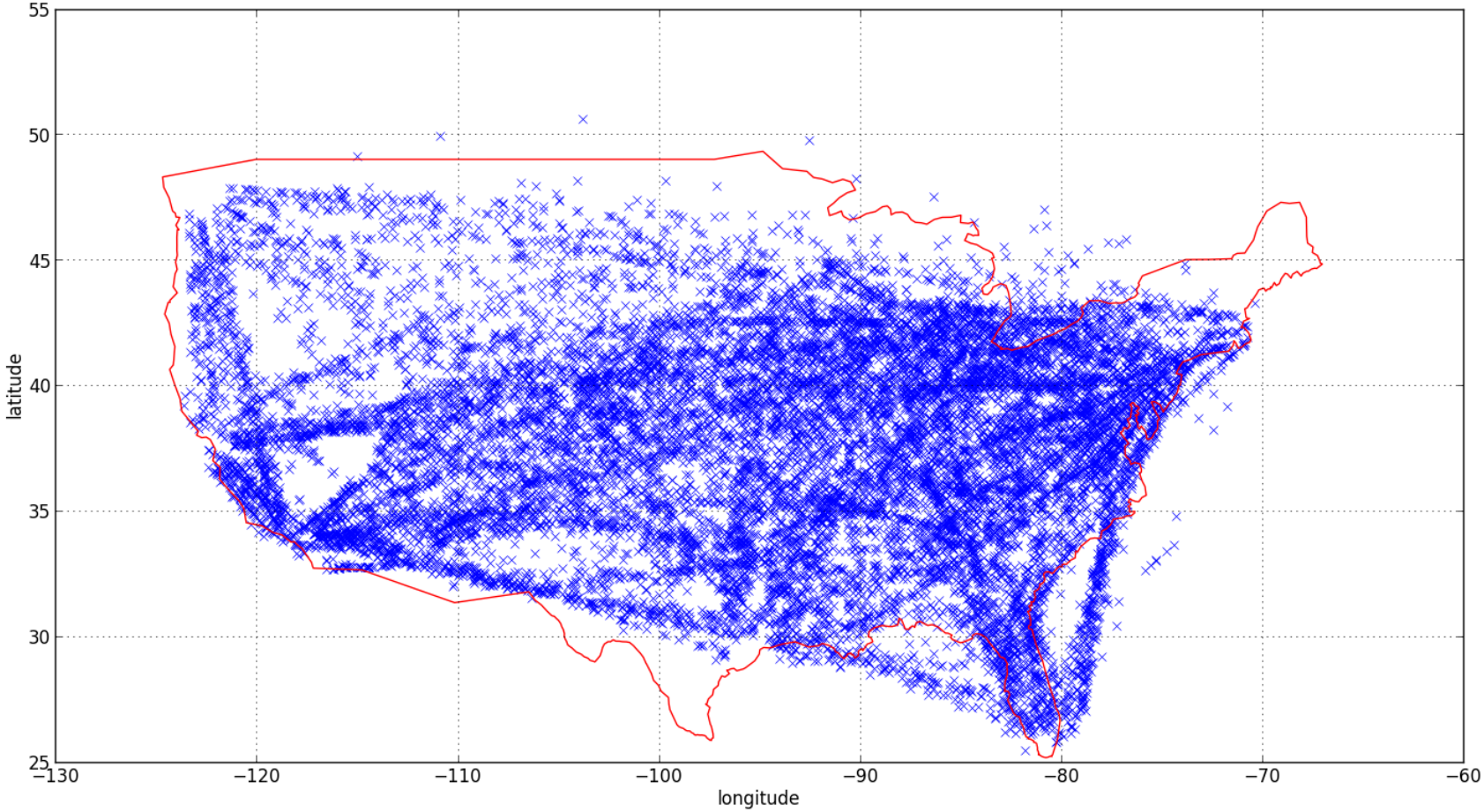
Department of Mathematics and Computer Science,
Széchenyi István University, Győr, Hungary

2014 IEEE Conference on BigData
Workshop on Large Data Analytics in Transportation Engineering
Washington DC, October 27, 2014

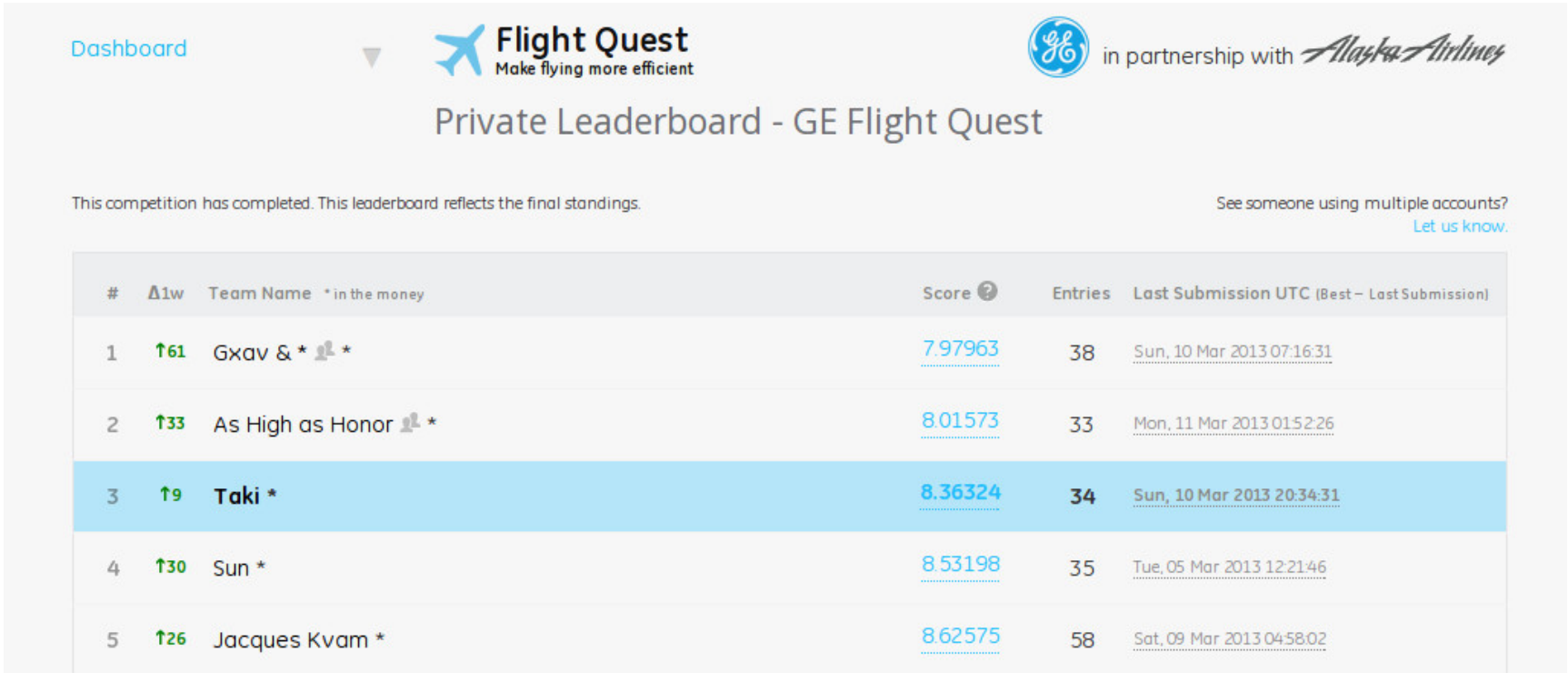
Győr, Hungary

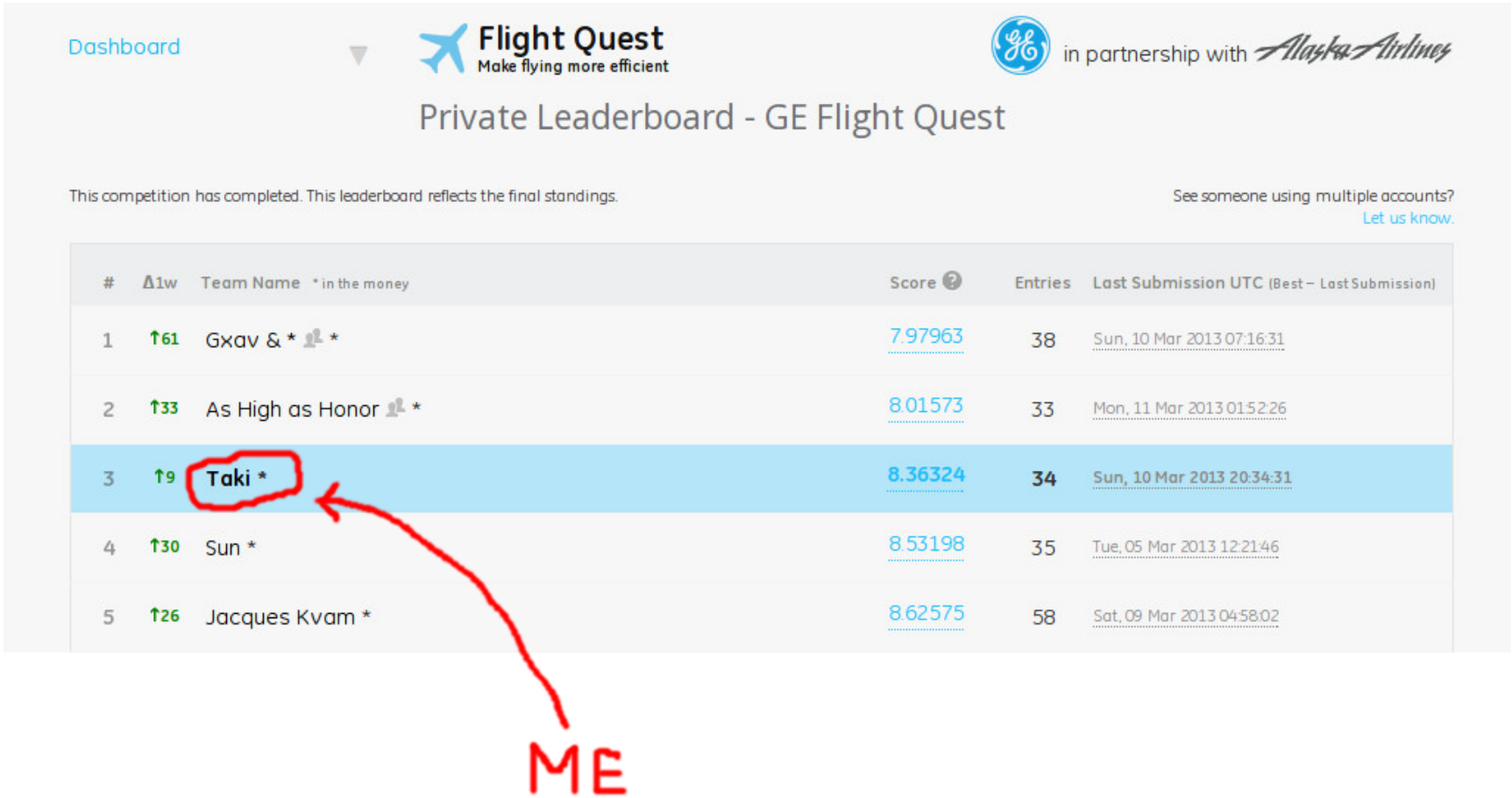
GE Flight Quest

- **GE Flight Quest** was a data science contest in 2012/13, organized by Kaggle.
  - Sponsor: GE (in partnership with Alaska Airlines).
  - Prize Pool: **250,000** $.
- The goal was to **improve the accuracy of flight arrival time estimates**, based on flight history, flight plan, weather and air traffic control data.



Flights corresponding to one test day.

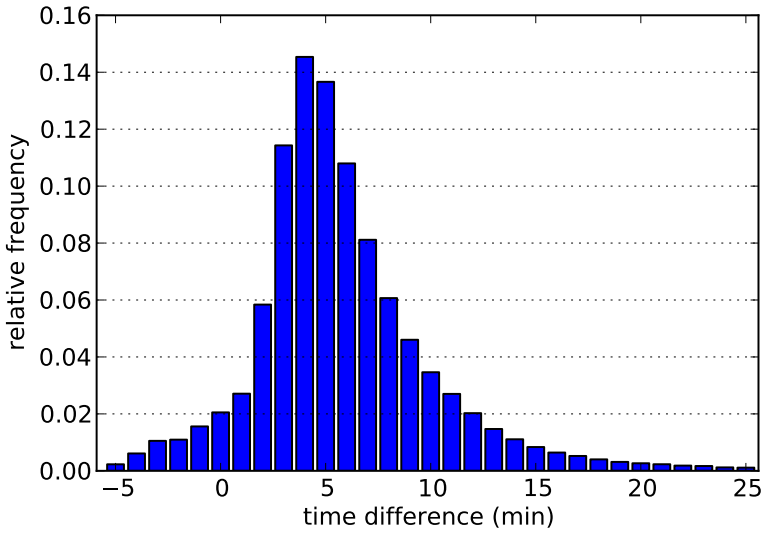

Dashboard
Flight Quest
Make flying more efficient
in partnership with Alaska Airlines
Private Leaderboard - GE Flight Quest
This competition has completed. This leaderboard reflects the final standings.
See someone using multiple accounts?
Let us know.
Δ1w Team Name * in the money Score Entries Last Submission UTC (Best – Last Submission)
1 ↑61 Gxav & * * 7.97963 38 Sun, 10 Mar 2013 07:16:31
2 ↑33 As High as Honor * 8.01573 33 Mon, 11 Mar 2013 01:52:26
3 ↑9 Taki * 8.36324 34 Sun, 10 Mar 2013 20:34:31
4 ↑30 Sun * 8.53198 35 Tue, 05 Mar 2013 12:21:46
5 ↑26 Jacques Kvam * 8.62575 58 Sat, 09 Mar 2013 04:58:02


Dashboard
Flight Quest
Make flying more efficient
in partnership with Alaska Airlines
Private Leaderboard - GE Flight Quest
This competition has completed. This leaderboard reflects the final standings.
See someone using multiple accounts?
Let us know.
Δ1w Team Name * in the money Score Entries Last Submission UTC (Best – Last Submission)
1 ↑61 Gxav & * * 7.97963 38 Sun, 10 Mar 2013 07:16:31
2 ↑33 As High as Honor * 8.01573 33 Mon, 11 Mar 2013 01:52:26
3 ↑9 Taki * 8.36324 34 Sun, 10 Mar 2013 20:34:31
4 ↑30 Sun * 8.53198 35 Tue, 05 Mar 2013 12:21:46
5 ↑26 Jacques Kvam * 8.62575 58 Sat, 09 Mar 2013 04:58:02
ME

Subtasks

Both the **runway** and the **gate** arrival time had to be predicted. The evaluation metric was

$$\mathrm{RMSE} = \frac{1}{4}\mathrm{RMSE}_{\mathrm{runway}} + \frac{3}{4}\mathrm{RMSE}_{\mathrm{gate}}.$$

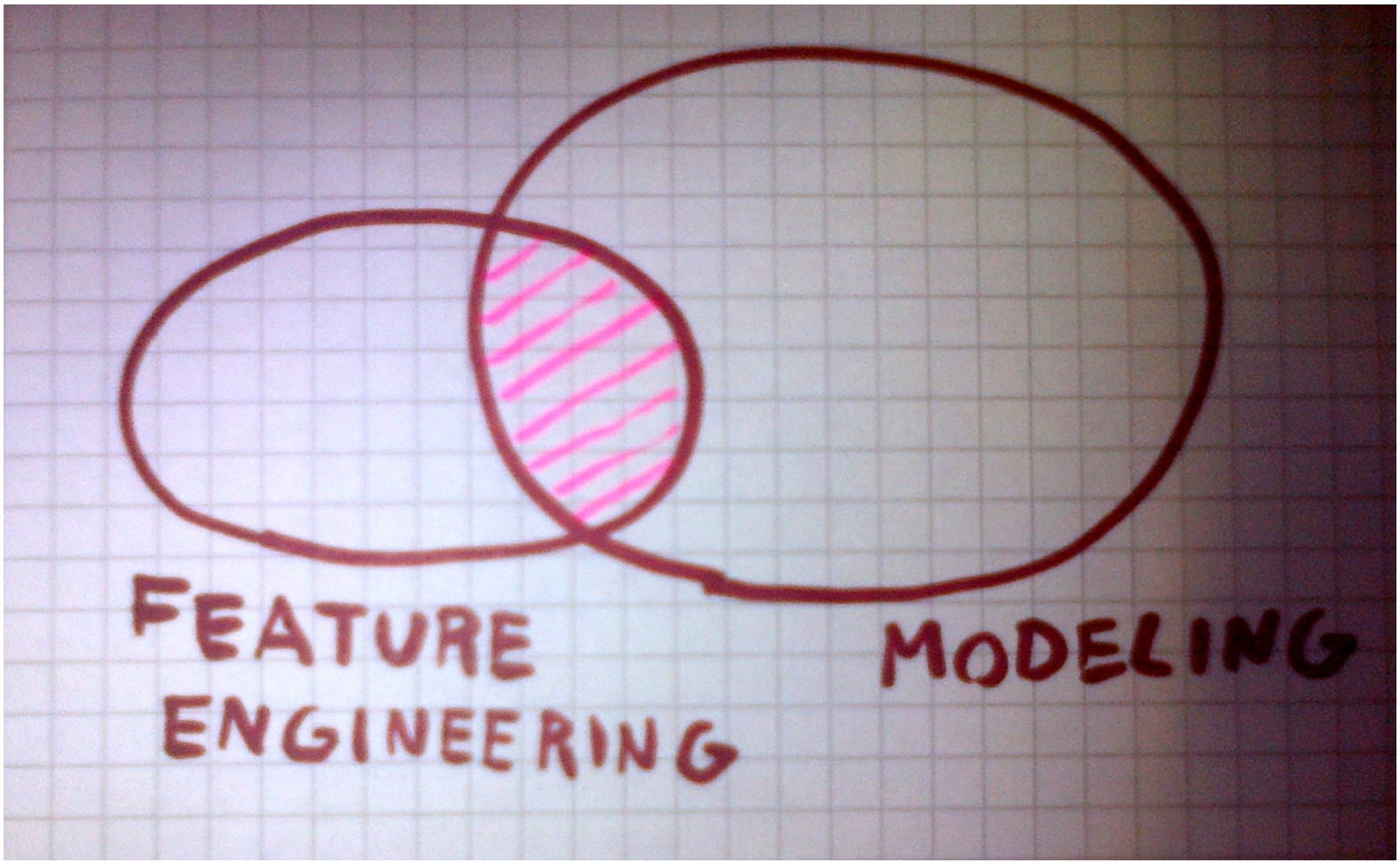


The distribution of taxi time (gate minus runway arrival time).

The Data Set

- The data set contained **252 columns** arranged in **34 tables**, with a lot of **missing** or **noisy** data.
- Size: **128 GB**.
- Time span: 109 days (the last 14 was the test set).
- Table groups:
  - **FlightHistory**: Direct information about the flights (e.g. departure and arrival location and time).
  - **ASDI**: Planned and actual travel path of the flights.
  - **ATSCC** Air traffic control events.
  - **Metar**: Actual weather data.
  - **OtherWeather**: Weather forecast data.

The Proposed Solution: Overview

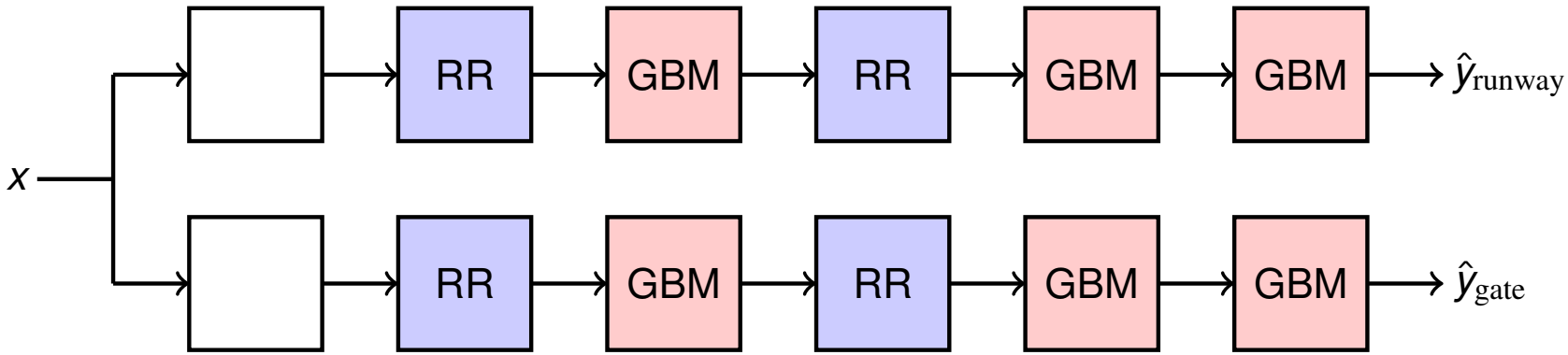


Feature Engineering

1. **Data cleaning**:
   - Flights with extreme delays were filtered out.
   - Intelligent missing value imputation for some variables.
2. **Extracting features** from the raw data:
   - Direct estimates (scheduled arrival times, most recent flight plan's estimates, etc.)
   - Sparse features (destination airport, airline, aircraft type, etc.)
   - Weather based features (e.g. is "Heavy Snow" present in the destination's weather report string?)
   - ...

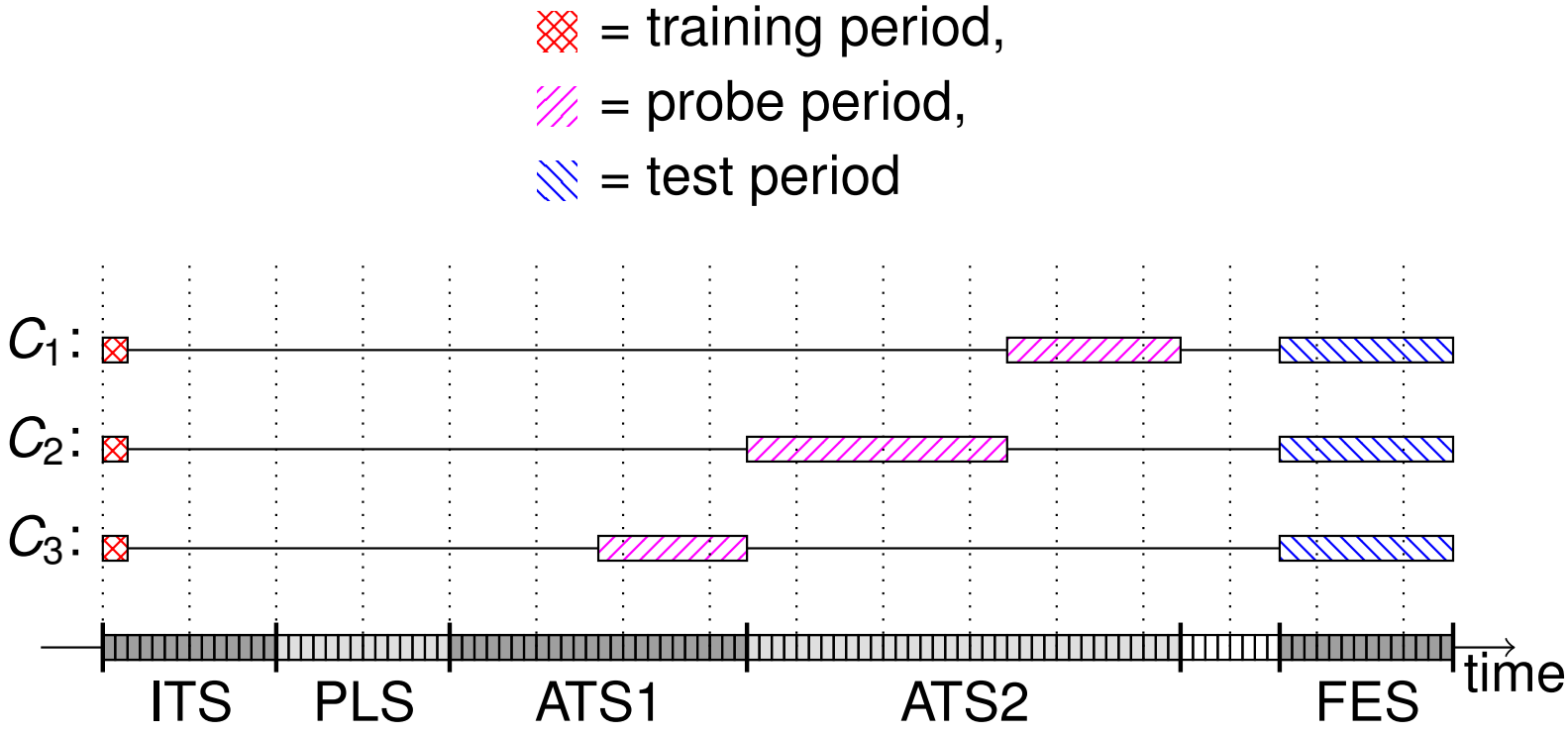
The Multistage Model

- A **6-stage model** of consecutive ridge regressions and gradient boosting machines was applied.
- The model uses **56 features** extracted from the raw data.
- The same feature set was used for runway and gate arrival time prediction, only the target was different.



The high-level structure of the solution.

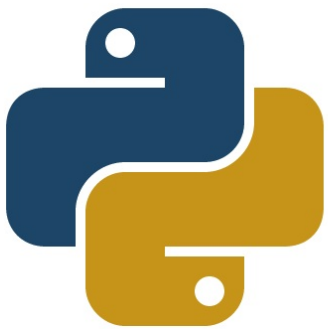
Partitioning



Configurations involved in generating the final submission.

Implementation Details

- The hardware I used during the competition was a **64 core** Linux server with **1 terabyte of memory**.
- Except some utility Bash scripts, the system is implemented in **Python** (with pandas and scikit-learn).
- The core of the system consists of **data processing nodes** that depend on each other.
  - Each node expects a set of files as input an produces another set of files as output.
  - An example node is for instance the conversion of raw FlightHistory table from CSV to binary format with parsed date values.



PYTHON IS AWESOME!

Per Stage Results

Stage	Alg.	Time (s)	RMSE
1		0	93.8241
2	RR	1,279	6.4736
3	GBM	145,243	5.7108
4	RR	167,355	5.4262
5	GBM	255,610	5.3932
6	GBM	**290,677**	**5.3818**

RMSE of the 6-stage model using two configurations.

Simplifications

Features	Alg.	Time (s)	RMSE
fh/cutoff		0	93.8241
Main	RR	1	6.8691
+ Sparse_1	RR	32	6.3515
+ Sparse_2	RR	55	6.2155
+ Sparse_3	RR	181	6.1051
+ Sparse_4	RR	254	6.0775
+ Metar	RR	422	5.9307
+ ATSCC	RR	453	5.9150
+ GroupBy	RR	525	5.9172
Main	GBM	895	5.8804
+ Delta	GBM	1,735	5.7748
+ Position	GBM	2,399	5.6857
+ Other	GBM	**3,641**	**5.6415**

RMSE of a 3-stage model with various feature groups.

Time & space requirement

- **Time requirement**:
  - The total training and prediction time of the 6-stage, 2-conf, 10-fold solution is **81 CPU hours**.
  - It can be reduced by **99 %** at the cost of **5 %** loss in the accuracy.

- **Space requirement**:
  - The total model size of the 6-stage, 2-conf, 10-fold solution is **400 MB**.
  - It is **25 MB** in the simplified solution.

Thank you for you attention!

This work was subsidized by TÁMOP-4.2.2.C-11/1/KONV-2012-0012/: “Smarter Transport” – IT for co-operative transport system – The Project is supported by the Hungarian Government and co-financed by the European Social Fund.