ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Using Cloud to Streamline R&D Workflow

Predicting Train Delays

Finnish Meteorological Institute

Roope Tervo
Laila Daniel

Photo by Kalevi Lehtonen 1955. Not published until Commons in 2014.
https://fi.wikipedia.org/wiki/Tiedosto:Finnish_class_Dm4_locomotive_number_1607_in_the_year_1955.jpg

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Image: Solita Oy

We aim to predict disruption of rail traffic caused by weather

Operation center can take several actions:

- Get more workforce
- Reduce train shifts
- Communicate

Project timeline: 01/2018-10/2018
Project partners: IL, LiVi, Trafi, VR
Area: Finland
Time range: 2 days ahead
Time step: 1 hour

Image: Solita Oy

Two goals

1) Predict train delays

2) Evaluate Google Cloud Services

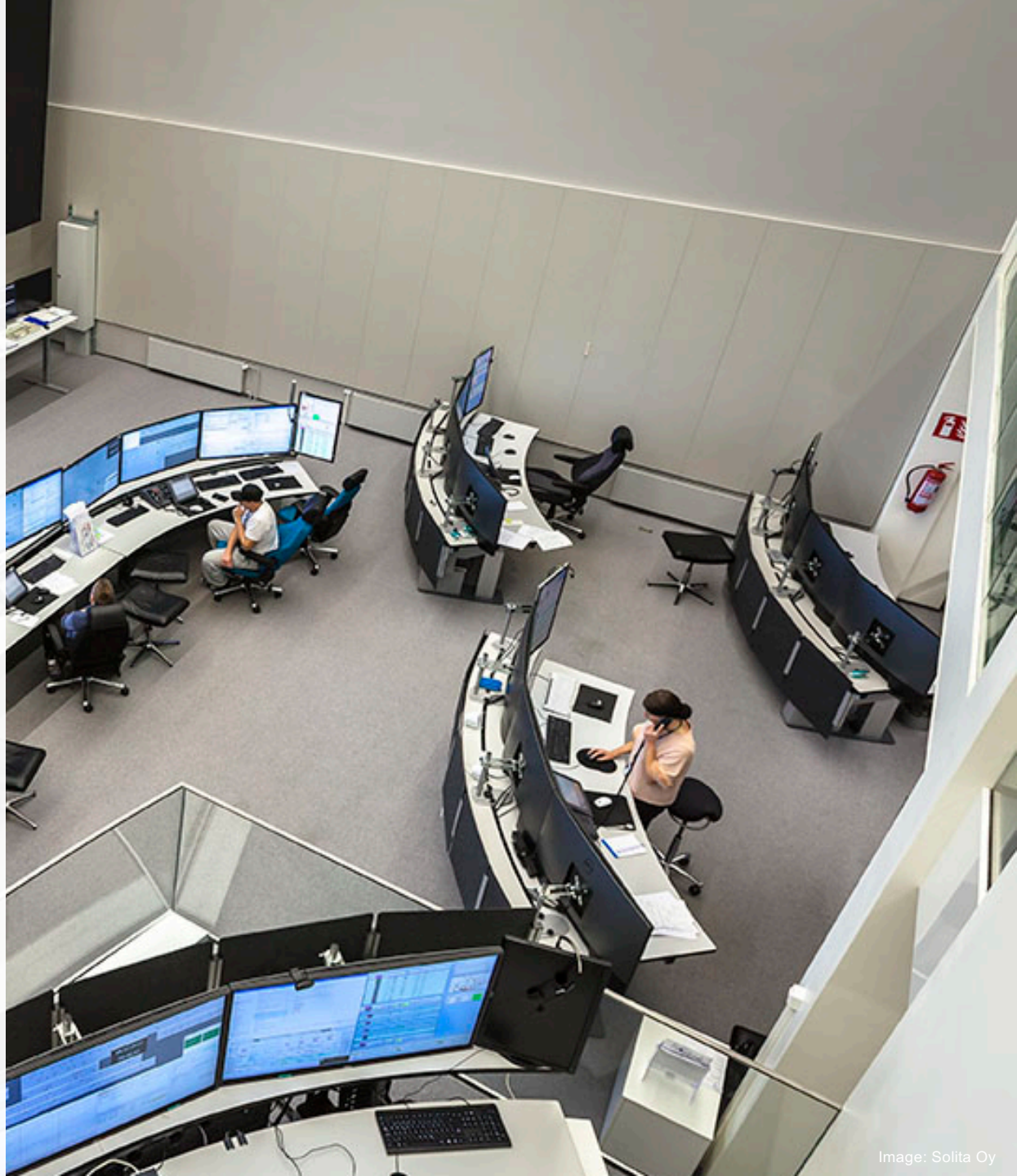
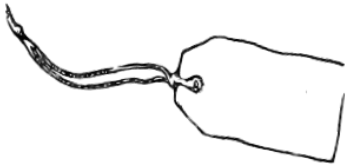
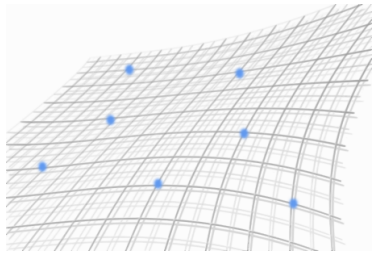


Image: Solita Oy



+

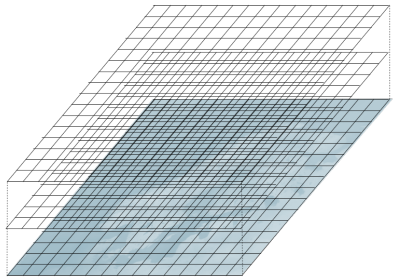


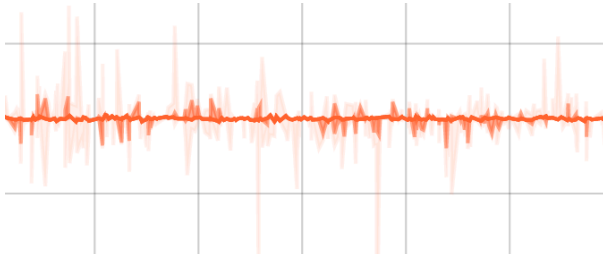
Label data

Feature data

Method

Results





NWP

Method

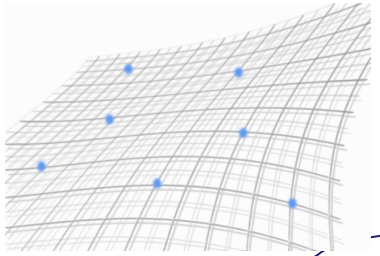
Prediction

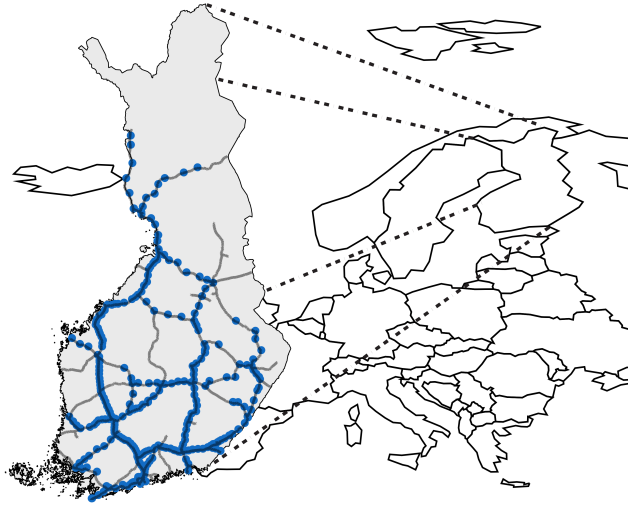
ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Data consist of train delays and corresponding weather observations

Delay between stations
- Passenger trains
- 514 stations

Weather observations
- 19 parameters





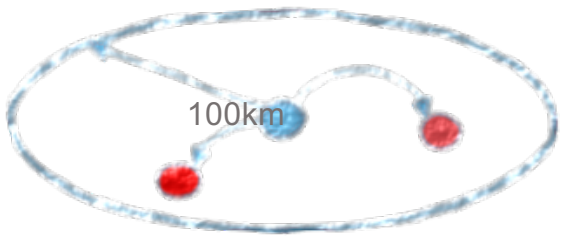
Data Liikennevirasto (CC4)

- Data from 2010 – 2018
- 30 M rows | 5.5 GB data

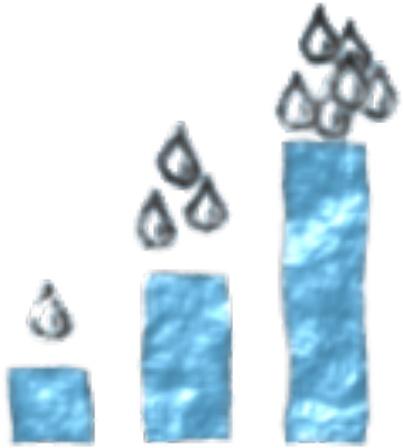
ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Various pre-processing methods used

Observations fetched with 100 km radius from train station using aggregation



Calculated 3h and 6h precipitation accumulation sums



Tried with and without imputation and normalisation

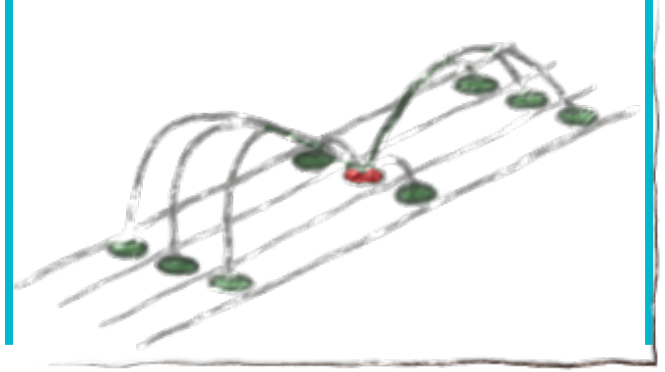


Image: Roope Tervo 2018. Original image: Mjanson srf. License: CC-BY-SA-3.0.

Tried PCA, ICA and K-Means clustering

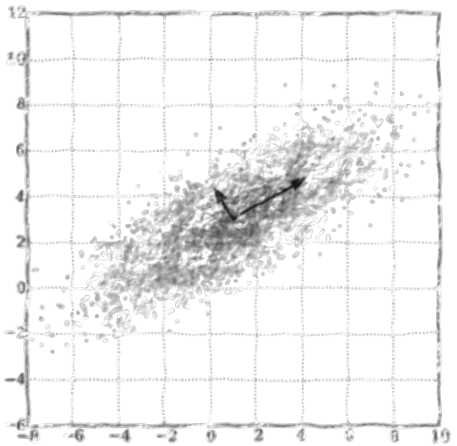


Image: Roope Tervo 2018.
Original image: Nicoguaro. License: CC4-BY.

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Three ML methods considered

LR

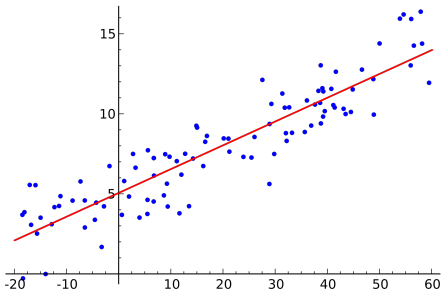


Image: CC0

RFR

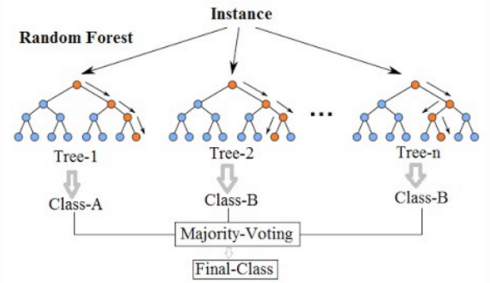


Image: Venkata Jagannath. License: CC4-BY

LSTM

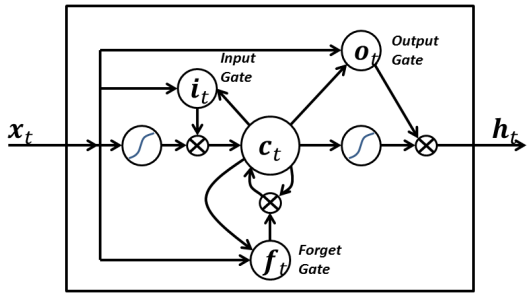
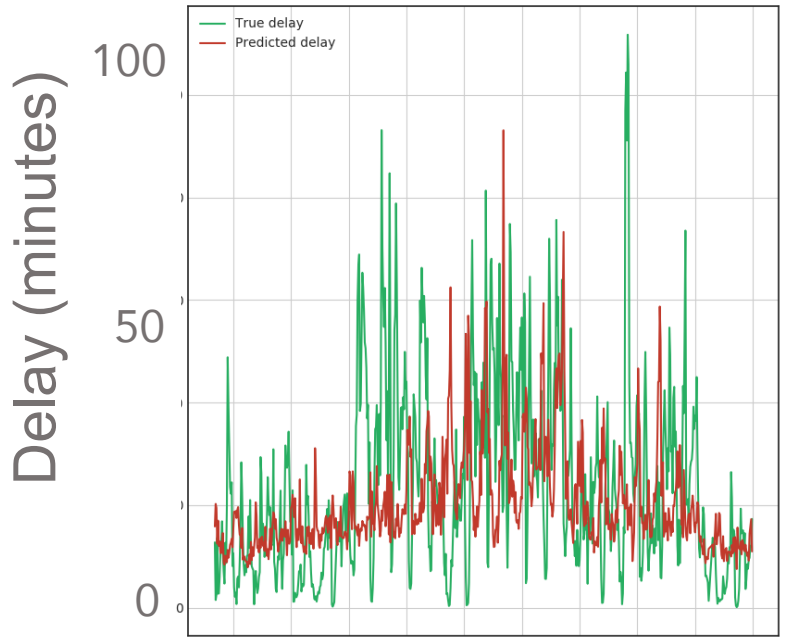


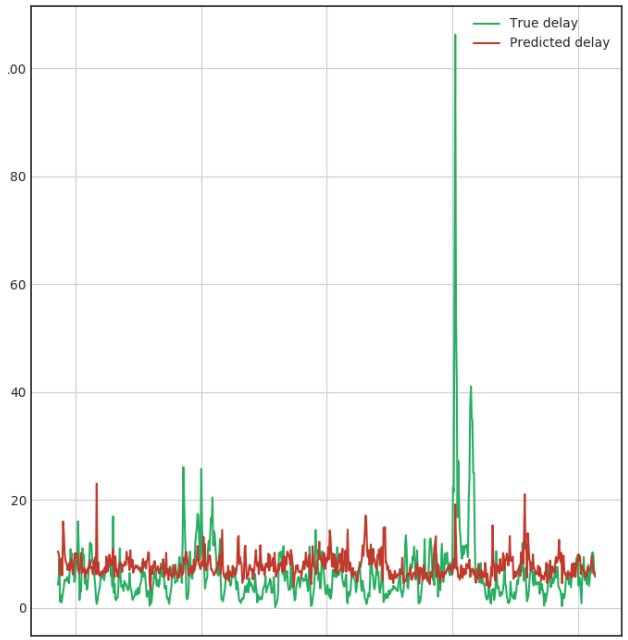
Image: BiObserver. License: CC4-BY

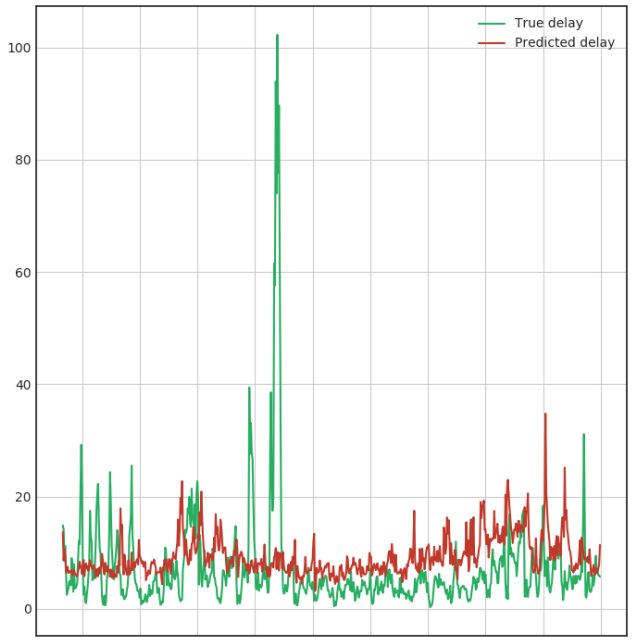
Random search used for finding optimal hyper parameters of LR and RFR

RFR works relatively well

Predicted vs. true delay, average over all stations, average over all stations

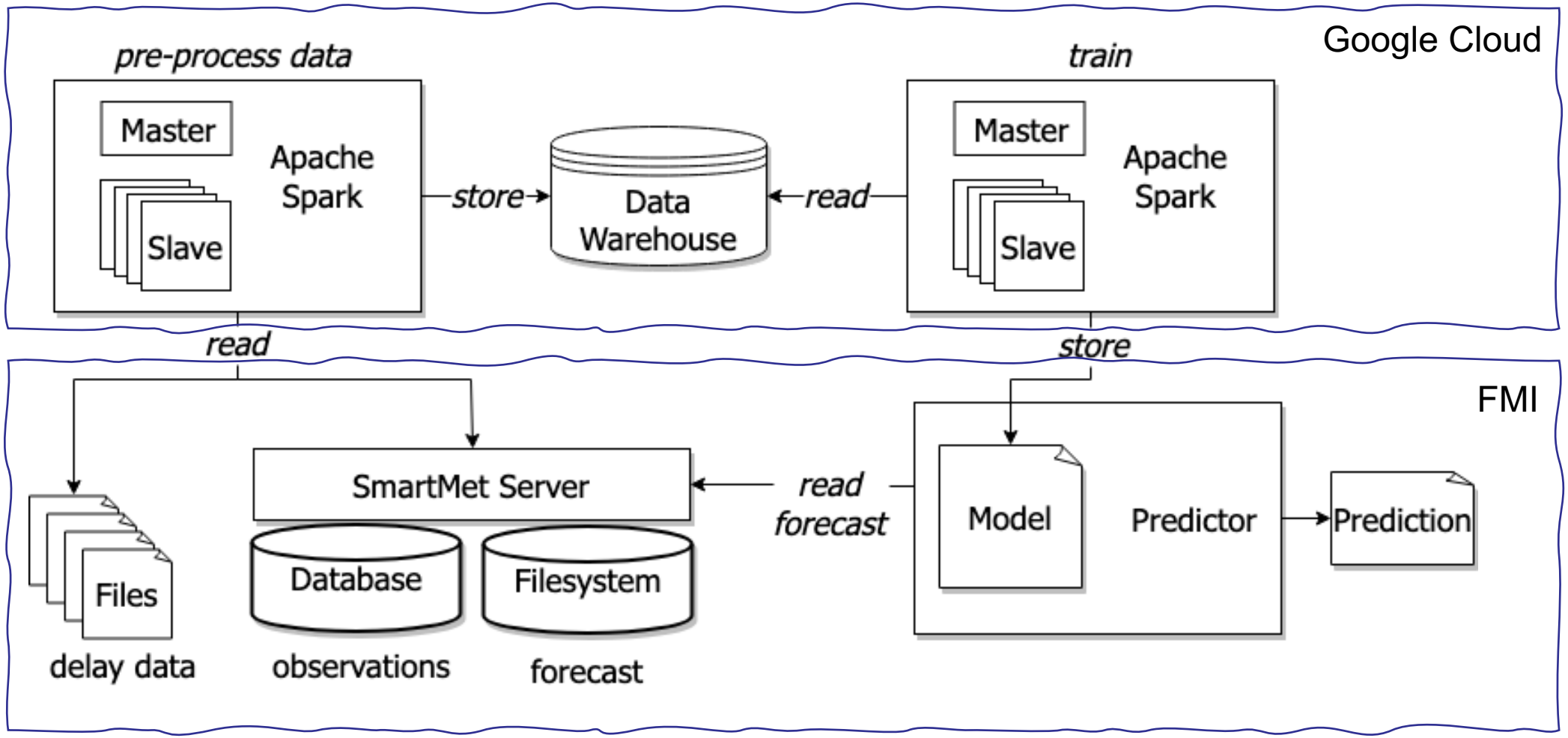




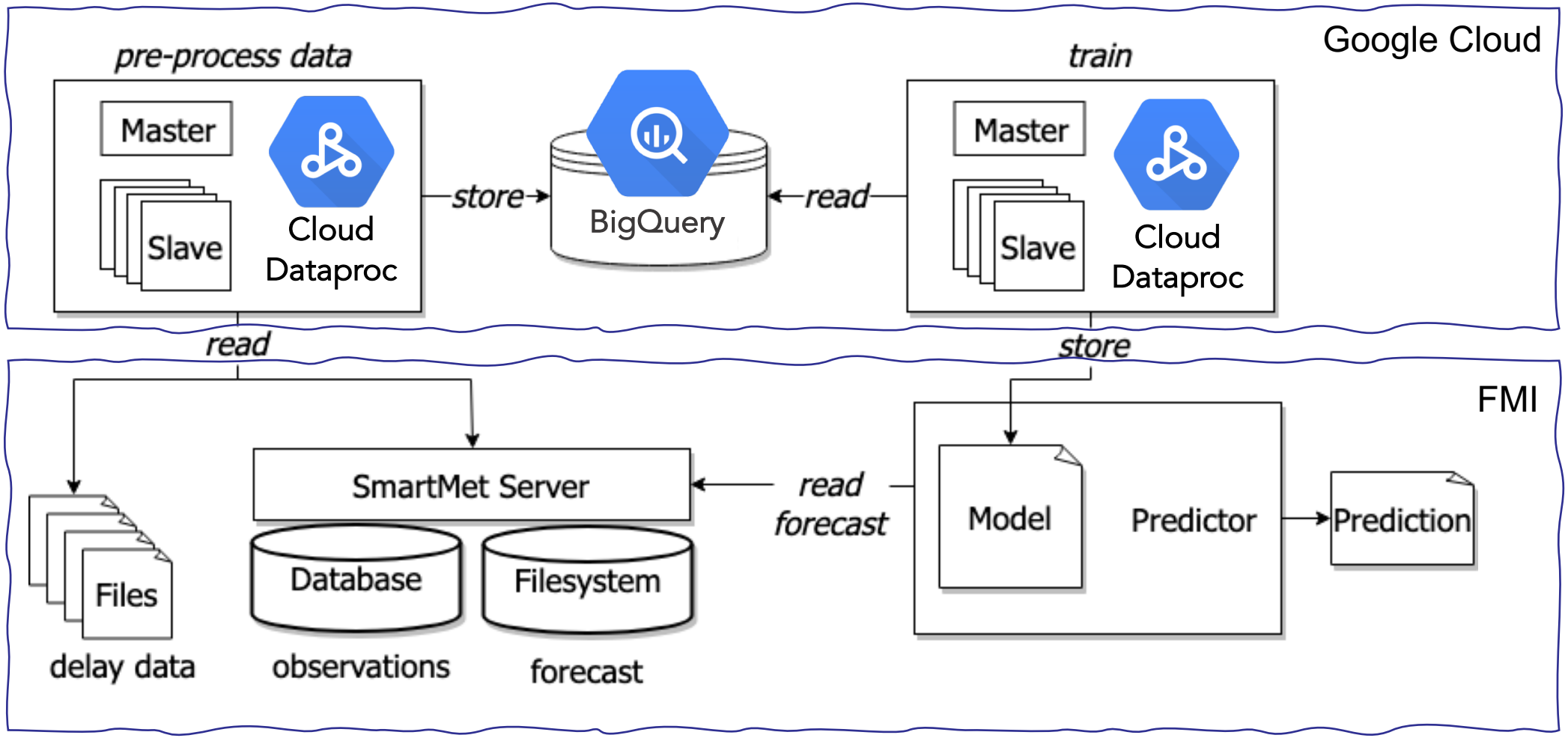


ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Platform consisted on-premise and Google Cloud components



DataProc and BigQuery the most important products



Apache Spark

Rough estimates of run times

Trains project using Spark framework	Elapsed time usin single server	Elapsed time using Spark
	8 cores, 64 Gi memory	8 workers, 2 cores, 7.5 Gi memory
Data retrieval from SmartMet (5 GB data, 3 million rows, 19 variables)	2 weeks	3 days
Random Forest Regression	1 day	8 hours
Keras LSTM	?	?

Apache Spark
Cloud benefits

Cloud benefits (versus on-premise)

- Easy to get resources when needed and scale down again
- Self-service
- Environment scriptable
- Fast data loading and storing from BigQuery and to buckets

Big Query

Data warehouse to handle large relational data

PostGIS

- Careful design needed
  - Table structure design
  - Indexing
  - Partition
- Few days of work

Same performance

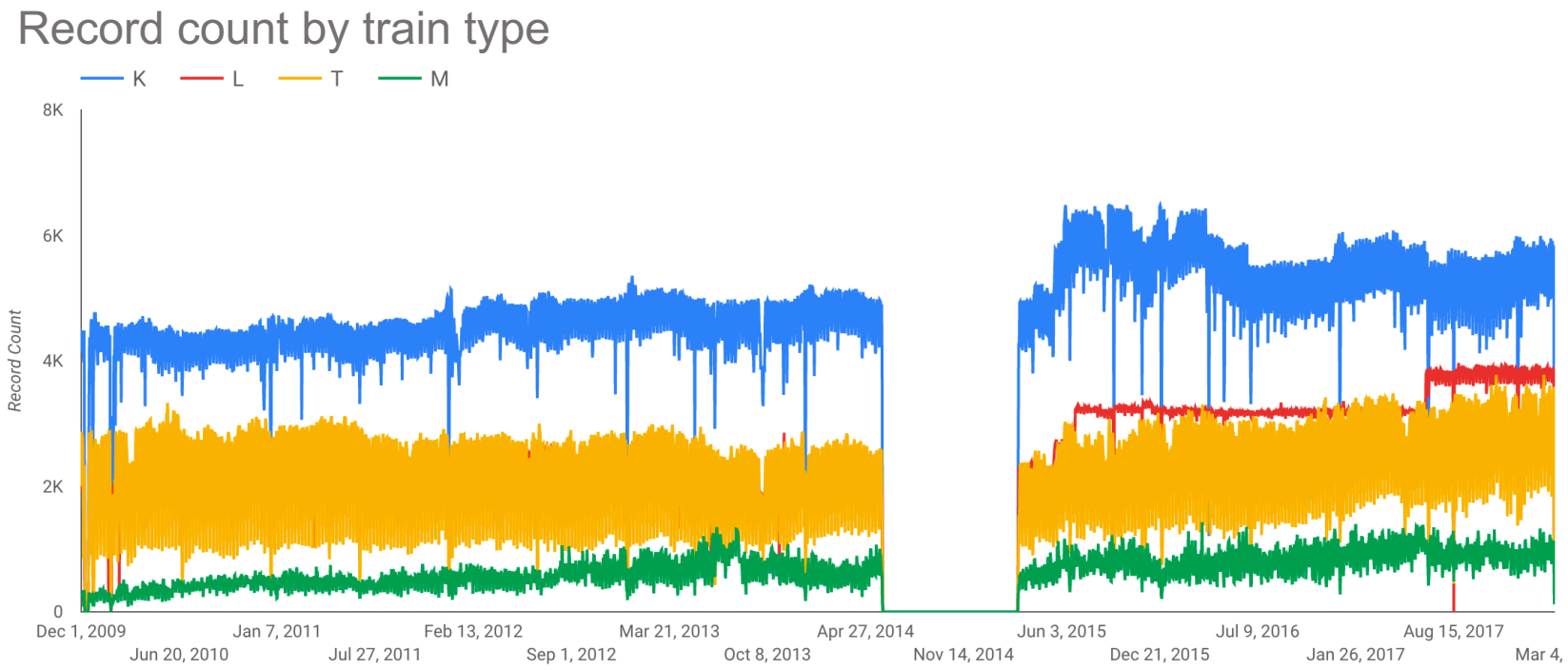
BigQuery

```
from google.cloud import bigquery

client = bigquery.Client()
dataset_ref = client.dataset('dataset_name')
table_ref = dataset_ref.table('table_name')
job = client.load_table_from_dataframe(df, table_ref)
job.result()
```

Data Studio

Very nice UI to quickly explore the data



General Notes

- Costs roughly $3000
  - Virtually all costs from using Spark
  - Ability to scale up and down is crucial
- Google Cloud easy to use
  - Especially command line tools convenient

Cloud resources improved the results

- Results would had been worse without cloud results
  - Faster run time enabled more evaluations
  - Cloud resources ease data handling
    → more time to develop prediction
  - BigQuery and Data Studio helped to reveal peculiarities in the data

Main advances are additional services

Additional services

- Benefits comes from hosted services, not virtual servers

Self-service

→ we were able to proceed in agile way

Scriptable environment really helps reproducibility

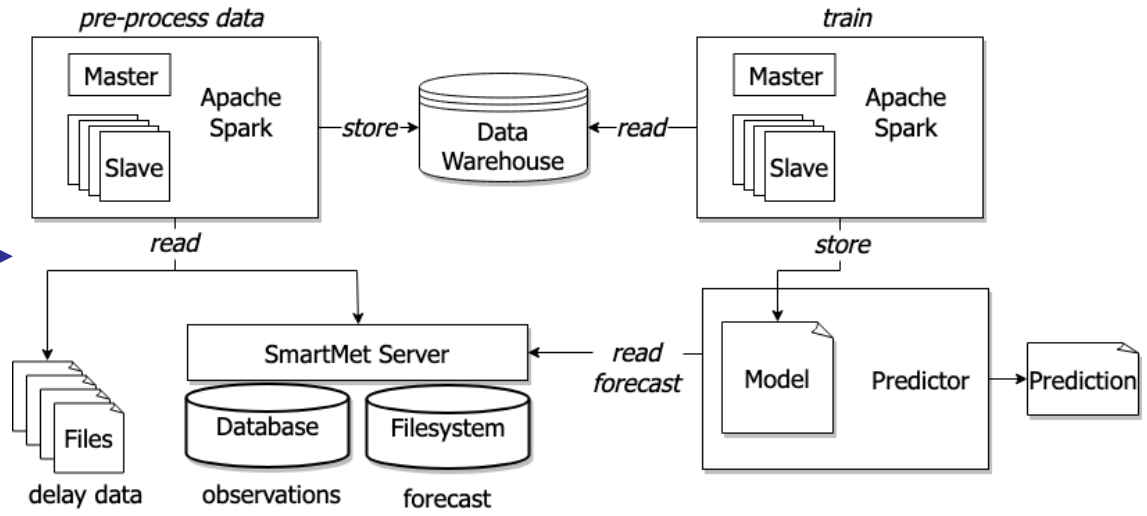
```
#!/bin/bash
# script to create a BigQuery dataset, create cluster,
# run the data retrieval program and finally delate the
cluster

echo "Creating bq dataset trains_all"

bq mk trains_all


echo "Creating cluster"
gcloud dataproc clusters create spark3 --master-boot-
disk-size 64 --image-version 1.2 --bucket trains-data --
master-machine-type n1-standard-4  --num-workers 2 --
worker-boot-disk-size 32 --worker-machine-type n1-
standard-2 --region europe-north1 --zone europe-north1-c
--properties spark:spark.executorEnv.PYTHONHASHSEED=0 --
initialization-actions gs://trains-data/bin/install-py3-
dataproc.sh


echo "running trains_get_surface_flash_obs_bq.py"
```



Scriptable environment can be moved to the data

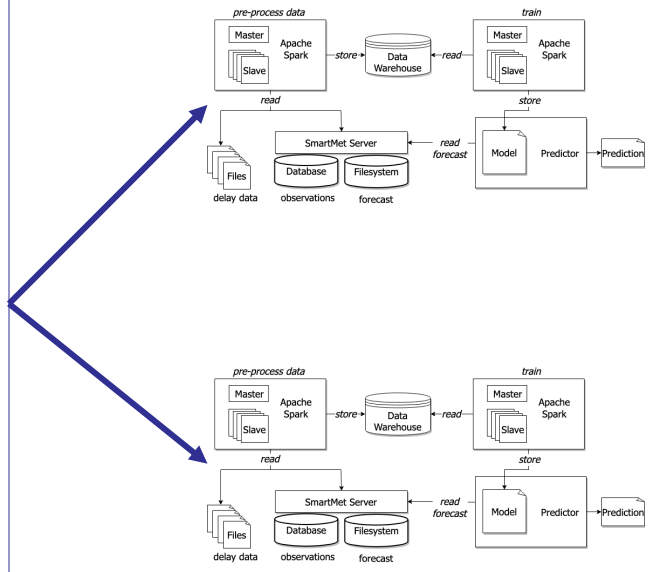
```
#!/bin/bash
# script to create a BigQuery dataset, create cluster,
# run the data retrieval program and finally delate the
cluster

echo "Creating bq dataset trains_all"

bq mk trains_all


echo "Creating cluster"
gcloud dataproc clusters create spark3 --master-boot-
disk-size 64 --image-version 1.2 --bucket trains-data --
master-machine-type n1-standard-4  --num-workers 2 --
worker-boot-disk-size 32 --worker-machine-type n1-
standard-2 --region europe-north1 --zone europe-north1-c
--properties spark:spark.executorEnv.PYTHONHASHSEED=0 --
initialization-actions gs://trains-data/bin/install-py3-
dataproc.sh


echo "running trains_get_surface_flash_obs_bq.py"
```



Summary

- Cloud services really helped us to improve the results
- Code defined environment is a significant part of reproducibility

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Questions

roope.tervo@fmi.fi