Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

MASTER OF SCIENCE IN ENGINEERING

Multimodal Processing, Recognition and Interaction

CASE STUDY

Bike Sharing Usage Prediction

Simon Ruffieux

Elena Mugellini, Jean Hennebert, Stefano Carrino

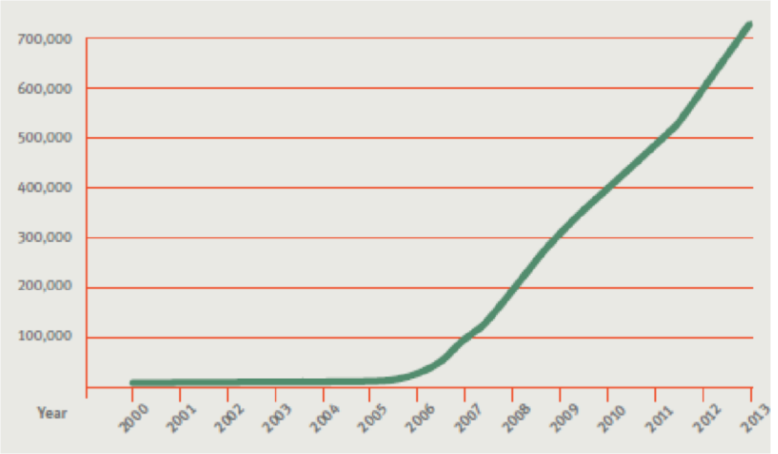
Summary

Introduction

- Project: GeVeLiSP
  - **Ge**stion de flotte de **Vé**hicules en **Li**bre-service basé sur des **S**ystèmes **P**rédictifs
- Project:
  - InnoSuisse - 18 months
- Partners
  - Intermobility SA (Velospot)
  - HEIA-FR - HumanTech
- Goal
  - Improve Bike Sharing Systems
    - Temporal predictions for bike usage
    - Optimization of rebalancing operations

Bike Sharing Systems - Overview

- Stations
  - Fixed locations
  - Bike return mechanism
    - Slots to return bikes
    - Zone to return bikes
- Bikes
  - Mostly standard bikes
  - Some electric bikes
- Redistribution
  - Need to (manually) redistribute bikes amongst stations



Company’s Goals

Ensure constant availability of bikes at stations

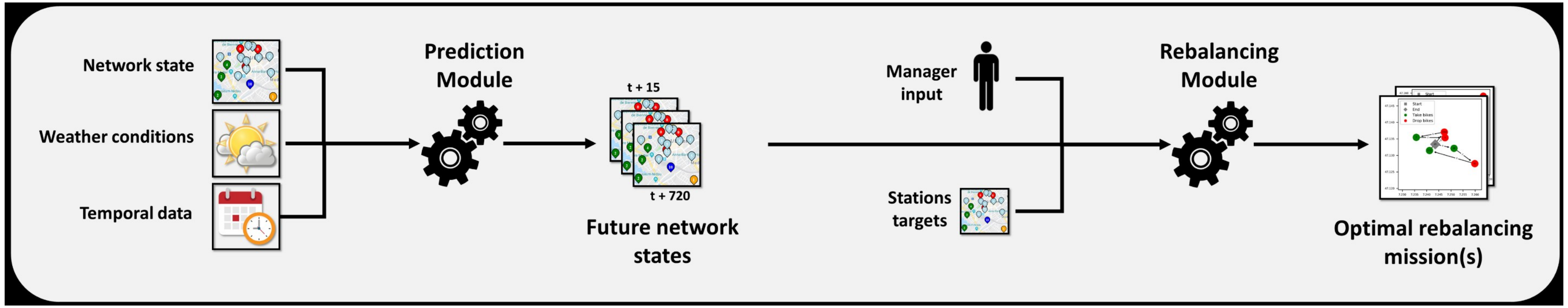
→Ensure there is constantly enough bikes at stations for customer satisfaction !

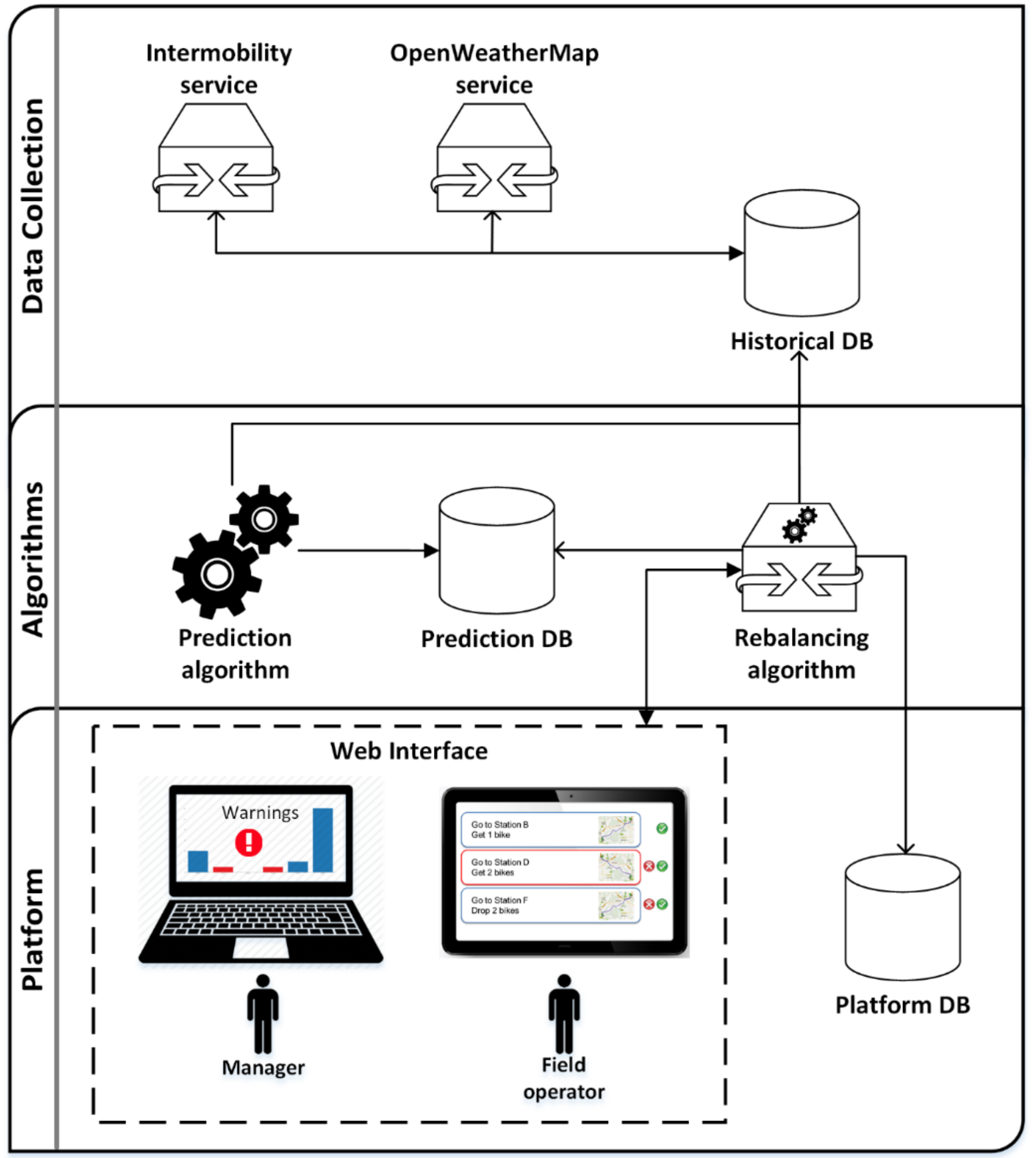
→The company must constantly rebalance bikes between the stations

Goal of predictions

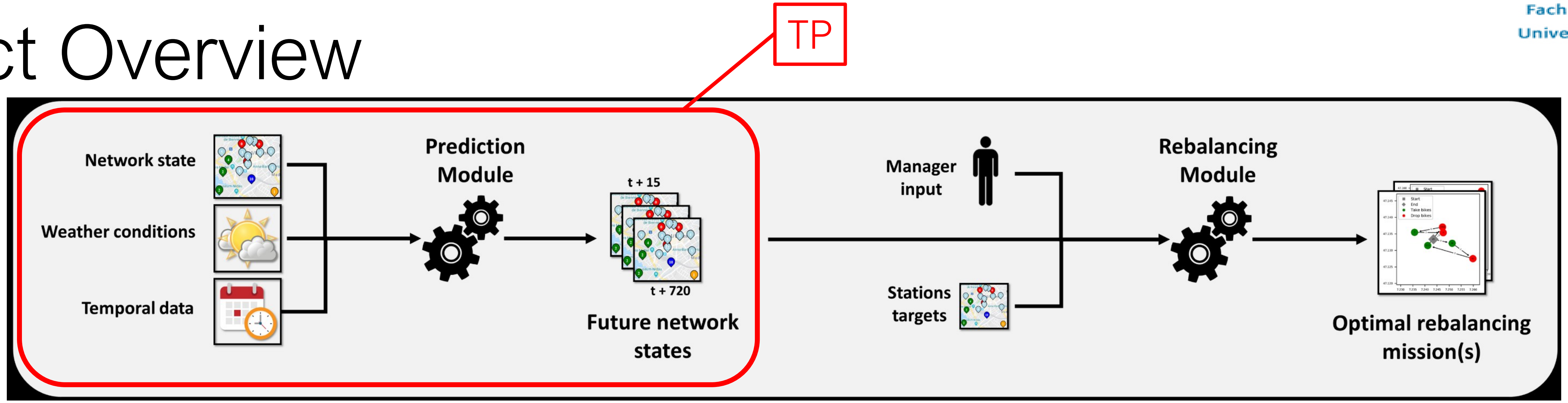
- Why predict the number of free bikes or free slots at a station ?
  - For users
    - Ensure there is a bike available
    - Ensure there is a free slot to return the bike
    - Better plan trips
    - Provide trip advices
  - For companies
    - Improve management of bike fleet
      - Improve rebalancing mechanisms
      - Receive warning before a station gets empty
    - Improve user satisfaction

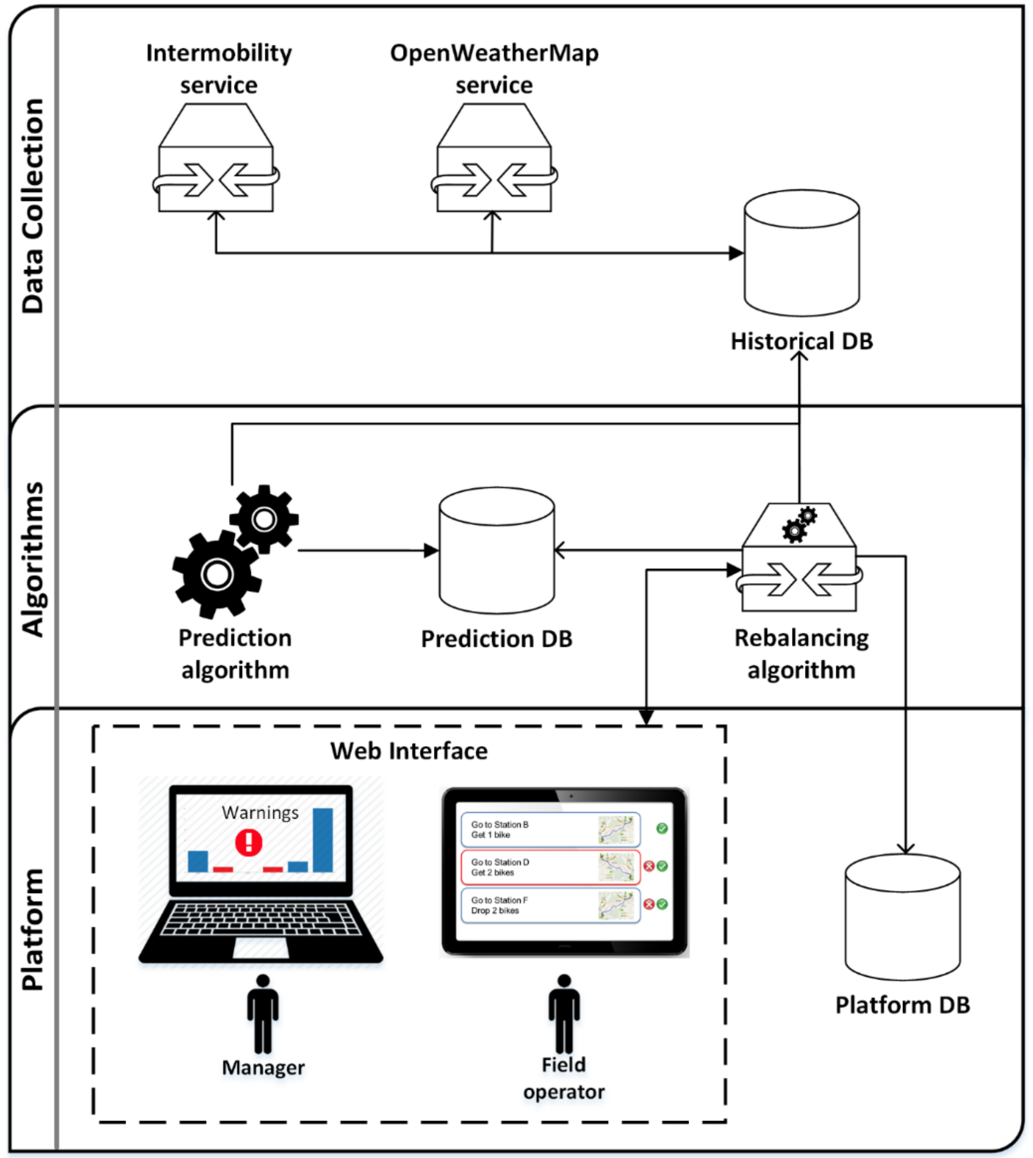
Project Overview





Project Overview





Input Data – Needs

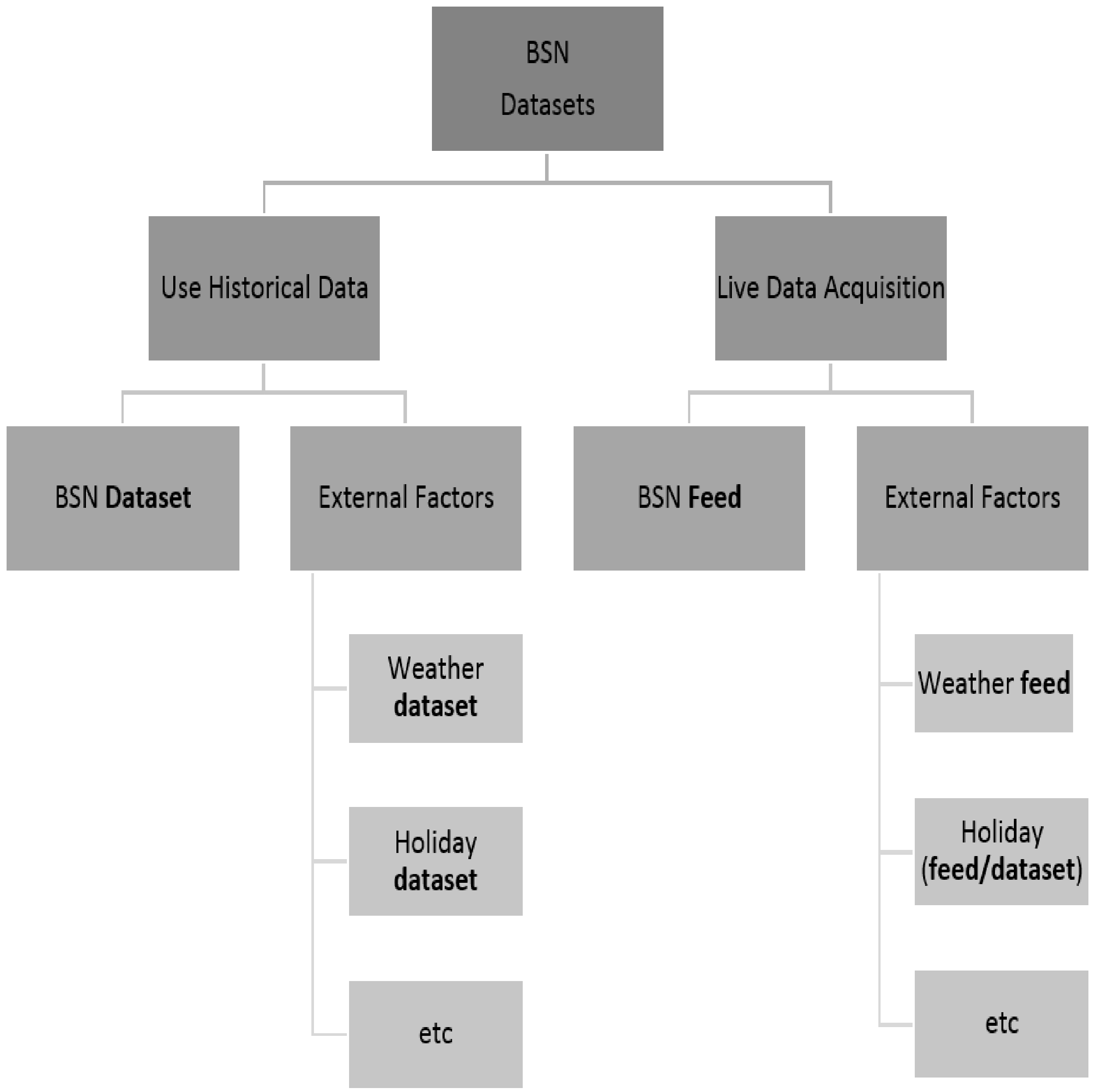
- What do we need to develop/test our algorithms ?

  → REAL DATA over a long period of time !
- What type of data (time series)?
  - Bike Sharing Systems data
    - Static: Bike Network Topology
    - Dynamic: #FreeBikes at each station
  - External factors
    - Weather (current, forecasts)
    - Type of day (weekday, holiday [school, university, …])
    - Events (concerts, manifestation, meetings, etc.)
    - …

Input Data – Creating datasets

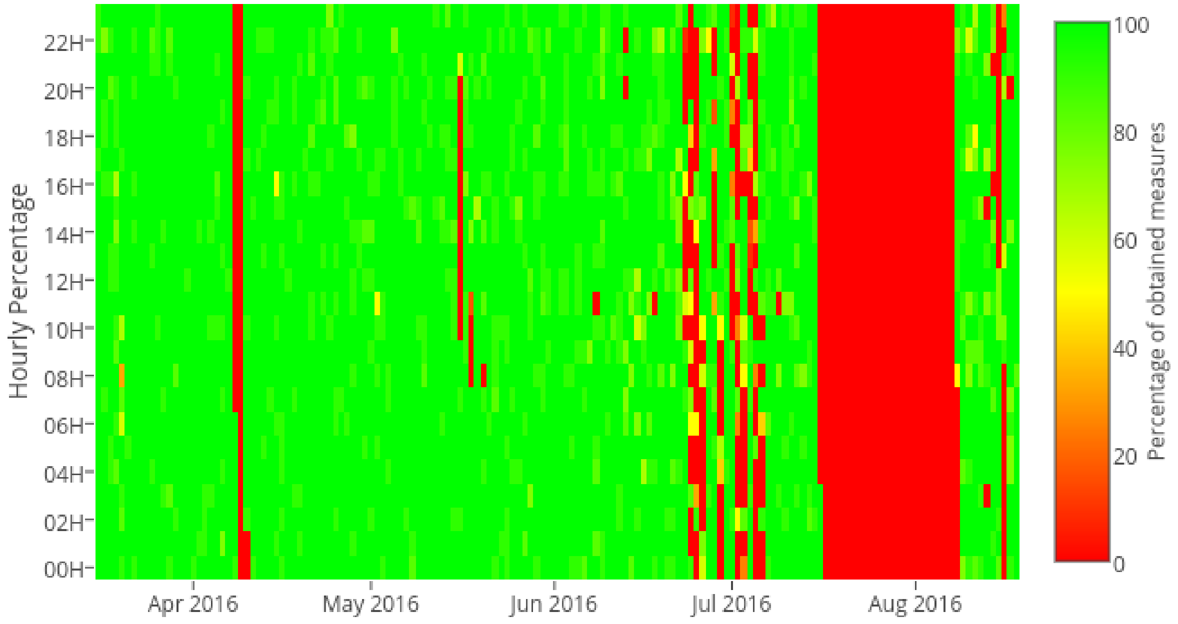
Two approaches:

1. Use existing **historical dataset** and augment it
   - Quickly obtain data over long period
   - Difficulty to obtain external factors
   - Limited BSN data availability
   - Need to fusion data (BSN + external)
2. Use live-data feeds to create our own dataset
   - Allow to choose frequency, data, etc
   - Provide live-data for prototypes and real demonstrator
   - Always augmenting dataset
   - Availability of data feeds
   - Require time to obtain enough data

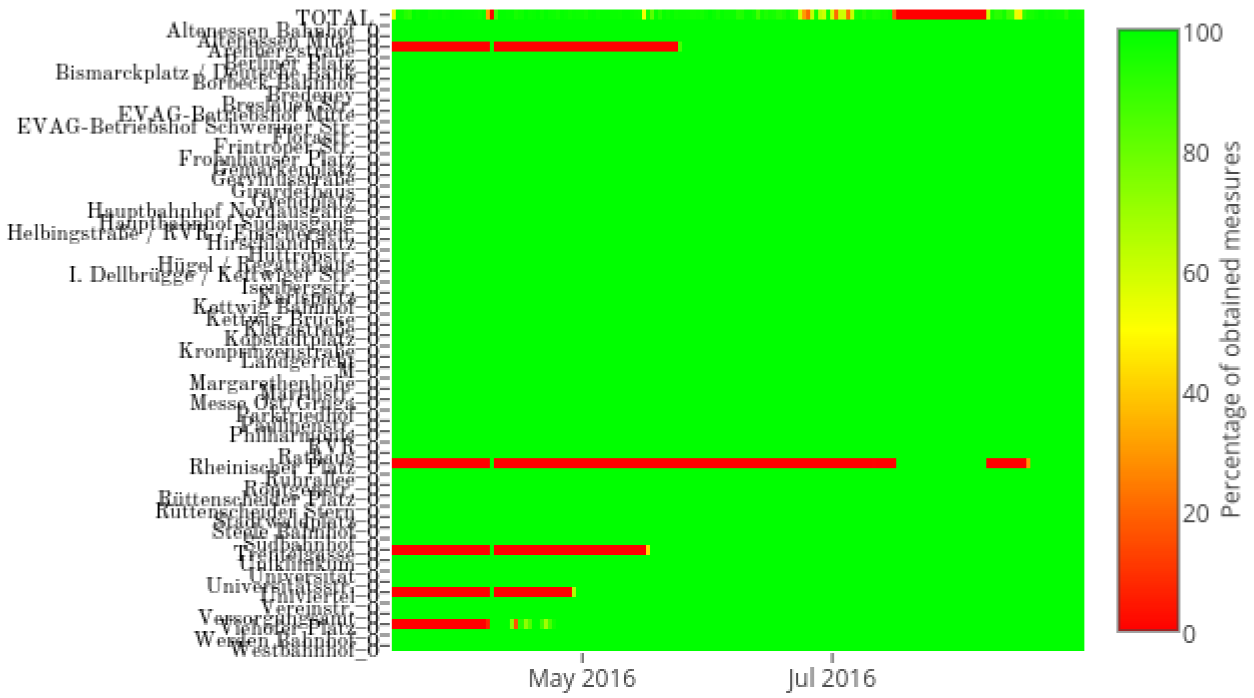


Input Data – Incomplete data (feeds)

- Data fetching problems:
  - Services are sometimes down
  - School server is sometimes restarted without warning
  - Internet connectivity problems



- Bike networks data problems
  - Stations are sometimes down
  - Stations are sometimes replaced



→ Pre-processing data !

Error Measures - Scoring

- What measure should we use for scoring ?
  - We have a regression problem, so we must quantify the error (no confusion matrix)
    - Mean Absolute Error

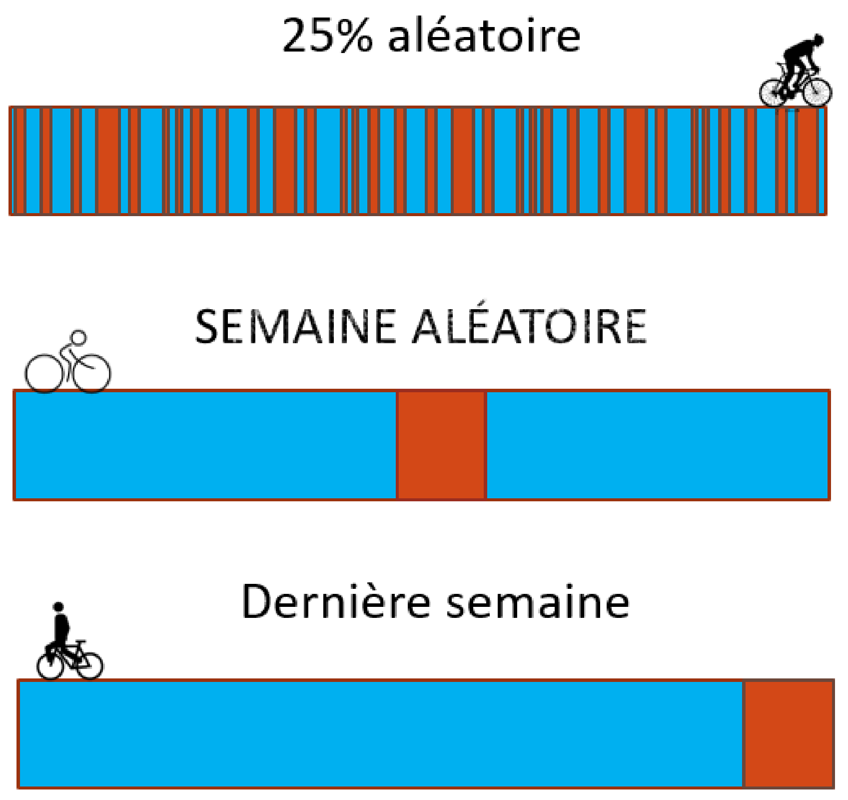
$$\rightarrow \boldsymbol{MAE} = \frac{\sum_1^n \ abs(Bike_{pred} - Bike_{real})}{N}$$

    - Root Mean Squared Error

$$\rightarrow \mathbf{R}\boldsymbol{MSE} = \sqrt{\frac{\sum_1^n (Bike_{pred} - Bike_{real})^2}{n}}$$

Error Measures - Partitioning

- How to split data into train/test sets ?
  - Randomly select samples ?
  - Select a week for test and the rest for train ?
  - Last week of the set ?
  - Weekly cross-validation ?



→ **In our case**, do you think it has an influence ?

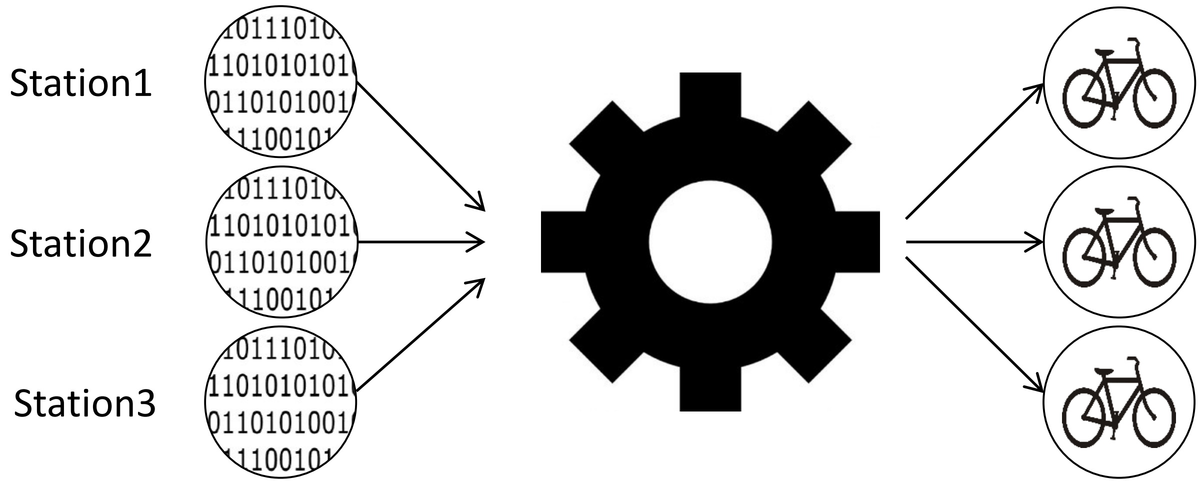
→ What are the advantage(s) of the “Weekly Cross-validation approach ?

Feature Processing – Model

- How do we model the problem ?
  - Whole Network ?
    - We assume the network as a whole and consider that all stations are linked
  - Independent Stations ?
    - We assume the stations are independent and consider each station separately
  - Partially Dependant Stations ?
    - We assume stations are partially linked and consider each station separately

Feature Processing – Model

- Whole Network Model (FLSM)
  - Only 1 algorithm model
  - INPUT - Data from all stations
  - OUTPUT – All bike predictions for every stations



Positive

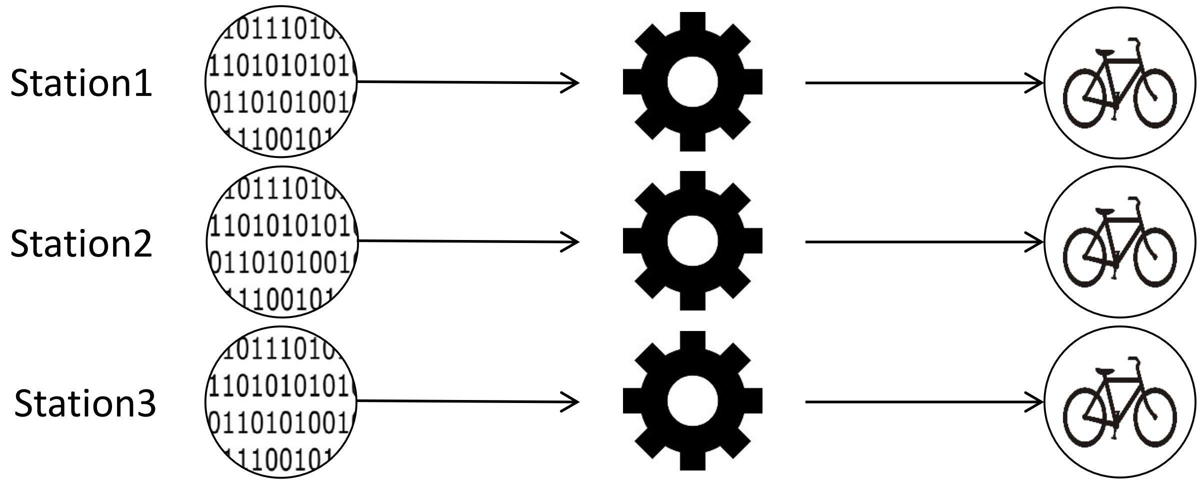
- 1 algorithm
- Algorithm may understand "connections" between stations

Negative

- Complexity of data
- Averaging over all stations error
- May not take into account station specifics

Feature Processing – Model

- Independent Stations Model (IIDSM)
  - 1 algorithm model per station
  - INPUT - Data from the station
  - OUTPUT – All bike predictions for the station



Positive

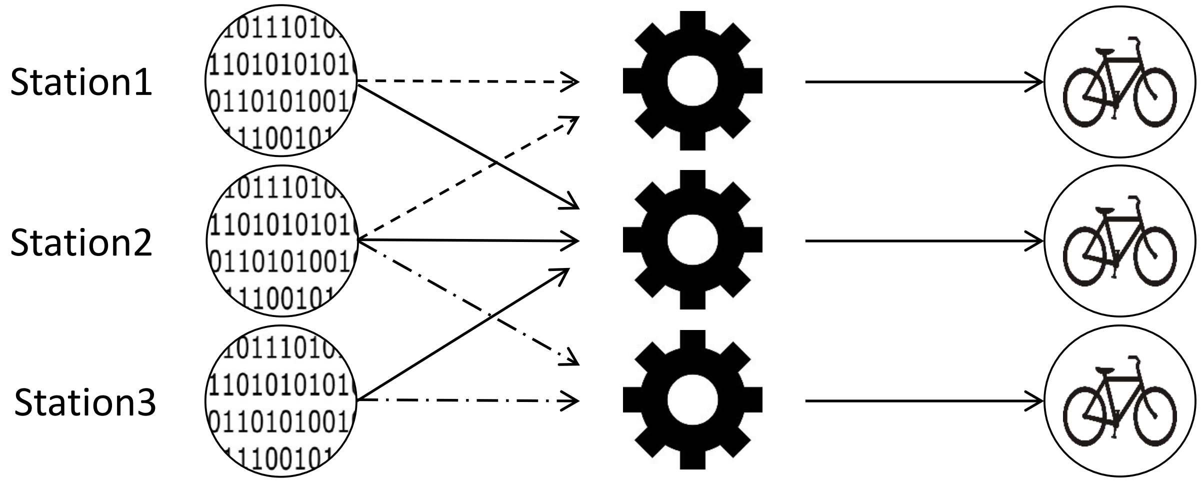
- Less complexity in data
- Take into account station specifics
- Parallelization friendly

Negative

- Lose information (dep. Between stations)
- Many algorithm models may take more time to train/predict

Feature Processing – Model

- Partially Dependent Stations (IPISM)
  - One algorithm model per station
  - INPUT – Data from the station and some data from surrounding stations
  - OUTPUT – All bike predictions for the station



<u>Positive</u>

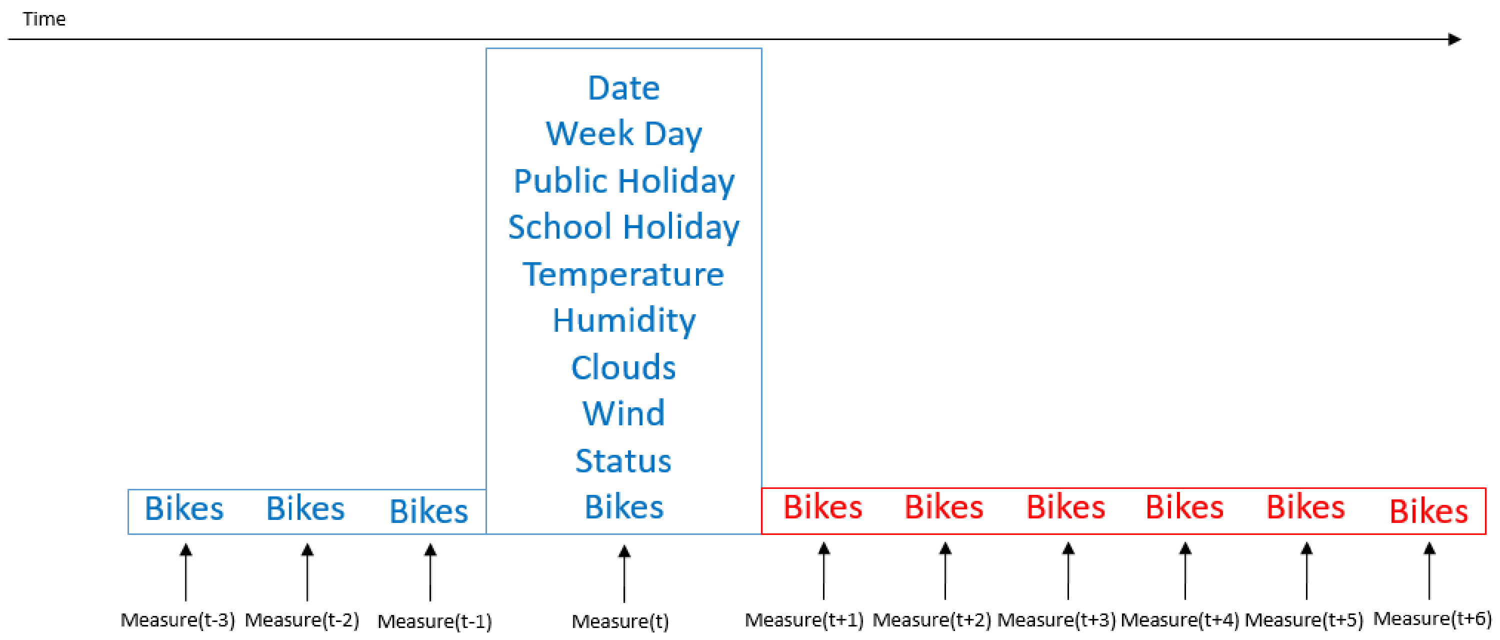
- Less complexity in data
- Take into account station specifics
- Parallelization friendly

<u>Negative</u>

- Many algorithm models may take more time to train/predict
- Only lose some dependency data

Feature Processing – What data ?

- Current and past data (in blue) to predict future data (in red)



Input data

- How far in past ?
- Pre-process data ?

Output data

- How far in future ?
  - What's the impact on accuracy ?

Data – Input dataframe (pandas df)

The input data vector is composed of the following features

X

- Date Information (index)
- Quantity of bike(s) available at station
  (t = 0, 5, 10, 15 and 20 minutes ago)
- Weather information
  (Temperature, Humidity, Cloudiness, Wind speed and 3h forecasts)
- Holiday information
  Integer representing the type of holiday (none, public, school)
- Quantity of bikes taken by a client at surrounding station<i>
  (t= 5, 10, 15 and 20 minutes ago)

y

- Future quantity of bike(s) available at station
  (t = 15, 30, 60 minutes in the future)

What You Should Know

- Historical & live datasets
  - Difference
  - Pros/cons
- Data partitioning for temporal predictions
- Feature processing
  - Explain the three models
    - Pros/cons of each model

- Practical work:
  - Random Forest Gini Coefficients

References

1. L. Chen, D. Zhang, G. Pan, X. Ma, D. Yang, K. Kushlev, W. Zhang, and S. Li. Bike sharing station placement leveraging heterogeneous urban open data. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, pages 571–575, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4. doi: 10.1145/2750858.2804291. URL http://doi.acm.org/10.1145/2750858.2804291.
2. P. Jiménez, M. Nogal, B. Caulfield, and F. Pilla. Perceptually important points of mobility patterns to characterize bike sharing systems: The Dublin case. Journal of Transport Geography, 54:228–239, June 2016.
3. M. Kaspi, T. Raviv, and M. Tzur. Bike sharing systems: User dissatisfaction in the presence of unusable bicycles. IIE Transactions, August 2016.
4. S. Wakamiya, Y. Kawai, H. Kawasaki, R. Lee, K. Sumiya, and T. Akiyama.Crowd-sourced prediction of pedestrian congestion for bike navigation systems. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS '14, pages 25–32, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3139-5. doi: 10.1145/2676552.2676562. URL http://doi.acm.org/10.1145/2676552.2676562
5. L. Caggiani and M. Ottomanelli, “A Dynamic Simulation based Model for Optimal Fleet Repositioning in Bike-sharing Systems,” *Procedia - Soc. Behav. Sci.*, vol. 87, pp. 203–210, 2013.
6. P. Aeschbach, X. Zhang, A. Georghiou, and J. Lygeros, “Balancing Bike Sharing Systems through Customer Cooperation – A Case Study on London ' s Barclays Cycle Hire,” *Conf. Decis. Control*, no. Cdc, pp. 4722–4727, 2015.
7. J. W. Yoon, F. Pinelli, and F. Calabrese, “Cityride: A Predictive Bike Sharing Journey Advisor,” 2012 IEEE 13th Int. Conf. Mob. Data Manag., pp. 306–311, 2012.