Water4Cities Webinar 1:

Methodology transfer from energy/mobility related scenarios to water management domain

Klemen Kenda, Matej Senožetnik,
@SingularLogic, Athens, November 24th 2017

SingularLogic

Artificial Intelligence Laboratory
Jožef Stefan Institute

Agenda

- Motivation: Results from FP7/H2020 mobility/energy related ICT projects
- Machine learning / Stream Mining Overview
- QMiner – open-source tool for stream processing
- Demo 1: Water4Cities Data Gathering Infrastructure Prototype
- Hands-on: Machine learning with Data from Water4Cities Infrastructure
- Demo 2: Smart Energy Forecasting on a Stream

Water4Cities

Artificial Intelligence Laboratory
Jožef Stefan Institute

Agenda

- **Motivation: Results from FP7/H2020 mobility/energy related ICT projects**
- Machine learning / Stream Mining Overview
- QMiner – open-source tool for stream processing
- Demo 1: Water4Cities Data Gathering Infrastructure Prototype
- Hands-on: Machine learning with Data from Water4Cities Infrastructure
- Demo 2: Smart Energy Forecasting on a Stream

Water4Cities

Artificial Intelligence Laboratory
Jožef Stefan Institute

FP7: NRG4CAST

Input: Streaming Live Energy Data

Output: Forecasts of Energy Demand/Consumption

Aim: Reduce energy usage and price

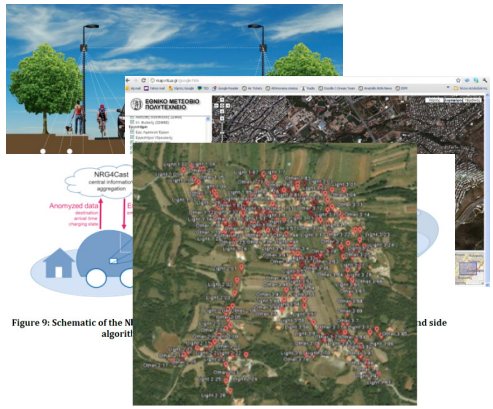


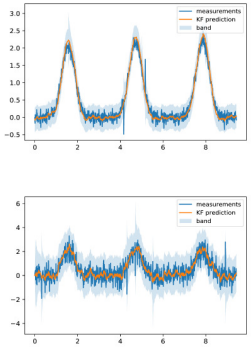
Figure 9: Schematic of the N… algorit… …d side

Artificial Intelligence Laboratory
Jožef Stefan Institute

Data Cleaning Approach

NRG4CAST & Water4Cities

- Autonomous and on-line – can be applied to real-world streaming scenario
- Using short-term prediction based on Kalman filtering
- Applied to artificial dataset, to NRG4CAST scenarios and to Water4Cities initial datasets
- Comparable to standard data cleaning approaches; behaves even better on datasets with low noise



Planned publication, related to W4C: *Business System Research Journal – K. Kenda, D. Mladenić: „Autonomous Sensor Data Cleaning in Stream Mining Setting" (2018).*

FP7: Sunseed

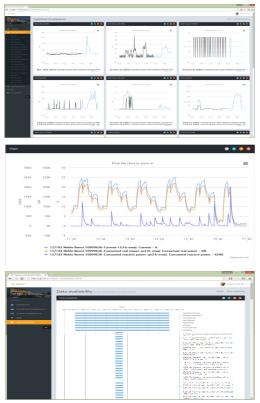
Prediction of consumption for different smart-meters.

Datasources:
- AMI meters
- Batched stream - once per day, resolution 15min

Model characteristics:
- Prediction horizon: 1h – 24h
- Algorithm: on-line linear regression

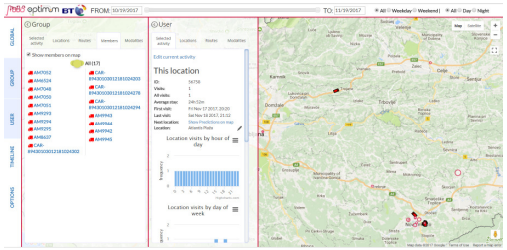
Result:
- Forecasted profiles for all AMI measurements
- At each new record, for the next 24 hours

Usage:
- Used by state estimation (SE) as quasy real-time measurements



H2020: Optimum

- Aim is to improve transit, freight transportation and traffic connectivity throughout Europe.
- Mobility patterns
- Data infrastructure



http://www.optimumproject.eu/

Agenda

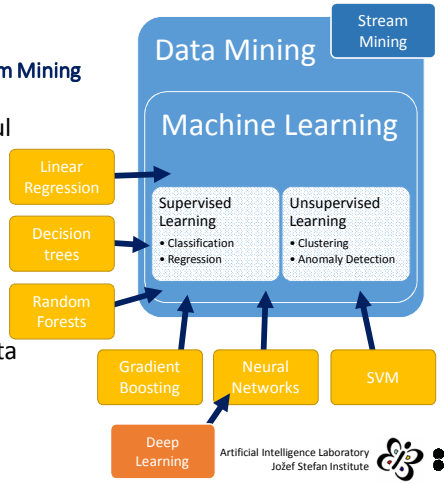
- Motivation: Results from FP7/H2020 mobility/energy related ICT projects
- **Machine learning / Stream Mining Overview**
- QMiner – open-source tool for stream processing
- Demo 1: Water4Cities Data Gathering Infrastructure Prototype
- Hands-on: Machine learning with Data from Water4Cities Infrastructure
- Demo 2: Smart Energy Forecasting on a Stream

Water4Cities

Definitions

Data Mining / Machine Learning / Stream Mining

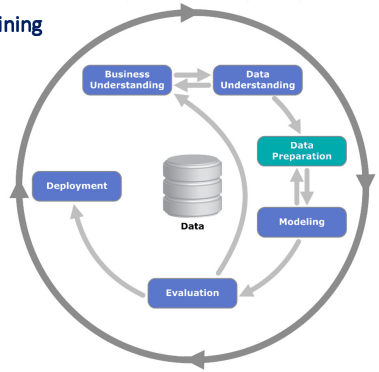
- Data Mining: Extraction of useful information from data
- Data Mining is application of Machine Learnig techniques to solve real-life data analysis problems
- Stream Mining is applied to streaming data (ideal for Big Data problems and for real-time systems)



Data Mining Process

Data Mining / Machine Learning / Stream Mining

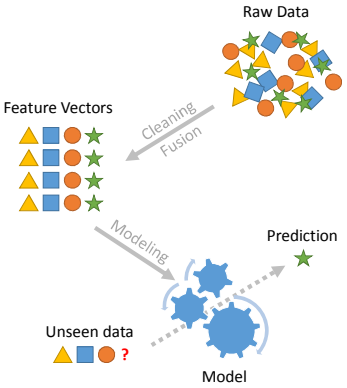
- Cross Industry Standard Process for Data Mining
- Holistic approach to data-driven modeling – useful for real-world applications
- From understanding of needs to deployment of models
- Data Preparation is the most time-consuming step



Supervised Learning

Data Mining / Machine Learning / Stream Mining

- **Model** that can predict continuous or nominal attributes
- Different than process-based models (!); underlying mechanisms are **not** important
- Based only on data
- Domain knowledge introduced through feature engineering



Agenda

- Motivation: Results from FP7/H2020 mobility/energy related ICT projects
- Machine learning / Stream Mining Overview
- **QMiner – open-source tool for stream processing**
- Demo 1: Water4Cities Data Gathering Infrastructure Prototype
- Hands-on: Machine learning with Data from Water4Cities Infrastructure
- Demo 2: Smart Energy Forecasting on a Stream

Water4Cities

QMiner – Data Analytics Platform

QMiner is an open-source data analytics engine for processing large-scale real-time streams containing structured and unstructured data.

- C++ library wrapped in Node.JS (Javascript) for fast development of integrated data analytics applications which can scale to mid-size data (TBs)
  - Scalable data structures (vectors, matrices, hash tables, graphs)
  - Data transformation (text, social networks, streams)
  - Numerical computation (linear algebra, machine learning)
- The basic open source engine QMiner was developed by Artificial Intelligence Laboratory at Jožef Stefan Institute: http://qminer.ijs.si/

QMiner - Background

History

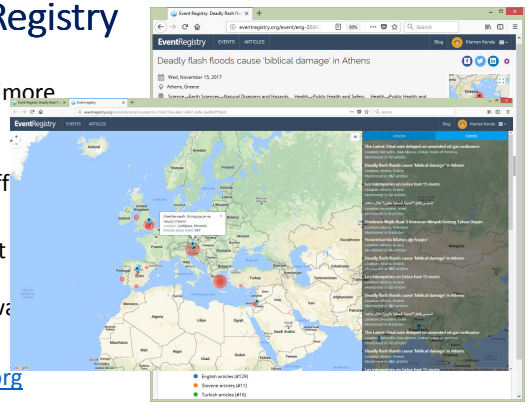
- Built on top of TextGarden and GLib C++ libraries
  Original use for text mining.
- First application for sensor data in 2011. Mostly used for environmental and energy-related data analysis.
- Stream Processing Engine (SPE) – working with real-time/real-world data
- Data fusion on a heterogeneous stream data
- Stream-mining techniques

When to use?

- QMiner covers storage and analytics
- Create industry-level solution in few lines of code
- Use-cases: Event Registry, Bloomberg Recommendation, Stream Story, Traffic…
- Covers all three big data Vs:
  - Variety: text, time series, social networks, structured data
  - Velocity: storage and analytics optimized for streaming scenarios
  - Volume: within limitation of non-distributed systems
- Most problems are mid-size data and not petabytes

QMiner – EventRegistry

- Monitoring world media (more than 100.000 sources)
- Discovering events
- Observing information diff
- Observe locations
- Clustering of events (what media focus on?)
- Custom topic pages (for wa example)



http://www.eventregistry.org

QMiner – Use it!

```
$ npm install qminer
```

For more information about basic QMiner engine, please visit:

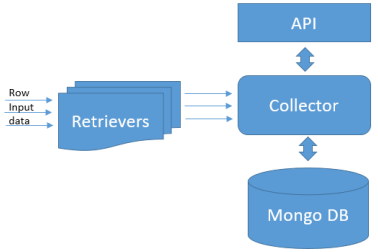
https://github.com/qminer

Agenda

- Motivation: Results from FP7/H2020 mobility/energy related ICT projects
- Machine learning / Stream Mining Overview
- QMiner – open-source tool for stream processing
- **Demo 1: Water4Cities Data Gathering Infrastructure Prototype**
- Hands-on: Machine learning with Data from Water4Cities Infrastructure
- Demo 2: Smart Energy Forecasting on a Stream

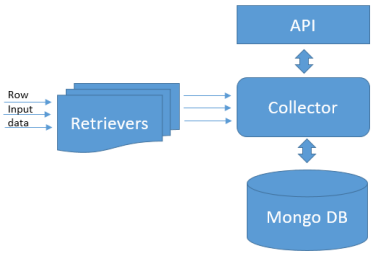
Data Collectors – Retrievers

- Services which function as data feed extractors by periodically polling an external data sources for updates
- Enable transforming pooled data into a common format
- End point is storing it by either writing to a file or forwarding it to a collection service
- Handle all data as JSON/XML objects



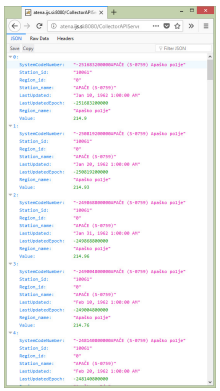
Collector

- Service that is backed by a Mongo database, containing real-time data, historical data
- Server uses Java servlets as API endpoints, through which data can be pushed to or queried from Mongo database.



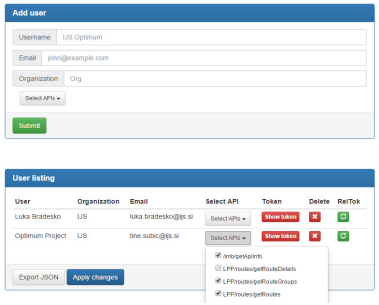
Data Retrievers - Demo

- 4 retrievers:
  - Underground water (Slovenia)
    - http://atena.ijs.si:8080/CollectorAPIServer/undergroundWater?region_id=0&station_id=10061&time_from=01/01/1962&time_to=01/01/1964
  - Pump sensors (Skiathos)
    - http://atena.ijs.si:8080/CollectorAPIServer/pumpSensors
  - Weather (Skiathos & Ljubljana)
    - http://atena.ijs.si:8080/CollectorAPIServer/weather?city=Ljubljana



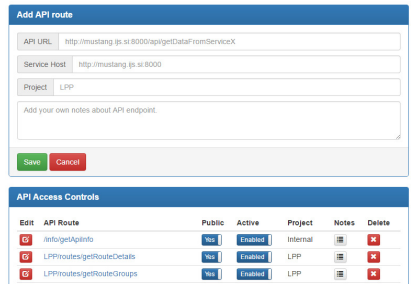
API Management (1)

- tool for managing access to API endpoints for end users
- MongoDB database for storing users and API endpoint data, graphical interface for managing users and APIs



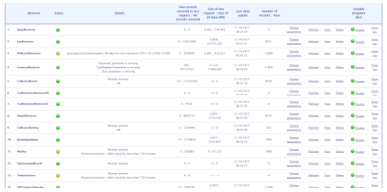
API Management (2)

- Users can view the list of APIs entered into the system
- Also possible to delete APIs, check notes or edit API properties
- Each API status can be set to “private”, meaning end users need authentication token and appropriate access rights to call this API



Watchdog

- Watchdog is a server with the purpose of having an overview of the numerous data retrievers
- When necessary, it (re)starts a retriever and keeps basic log of the retrievers activities in one place
- The aim is significantly lowering maintenance time and complexity of retrieving data from multiple sources



Agenda

- Motivation: Results from FP7/H2020 mobility/energy related ICT projects
- Machine learning / Stream Mining Overview
- QMiner – open-source tool for stream processing
- Demo 1: Water4Cities Data Gathering Infrastructure Prototype
- **Hands-on: Machine learning with Data from Water4Cities Infrastructure**
- Demo 2: Smart Energy Forecasting on a Stream

Water4Cities

Hands-On

Use:

git clone https://github.com/klemenkenda/w4c-webinar1

Or simply just visit the URL above.

You will get:

- slides
- Jupyter Notebook with example
- How to use Jupyter Notebook (in the README.md)

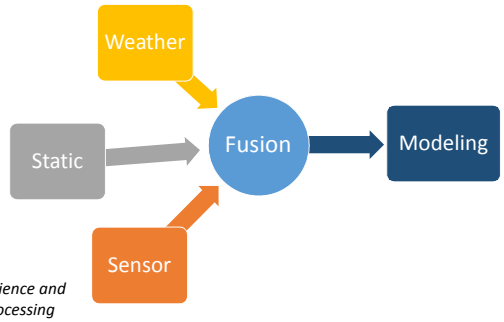
Agenda

- Motivation: Results from FP7/H2020 mobility/energy related ICT projects
- Machine learning / Stream Mining Overview
- QMiner – open-source tool for stream processing
- Demo 1: Water4Cities Data Gathering Infrastructure Prototype
- Hands-on: Machine learning with Data from Water4Cities Infrastructure
- **Demo 2: Smart Energy Forecasting on a Stream**

Smart Energy Forecasting on a Stream

Data Fusion

- heterogeneous data sources
- sensor data with different properties/delays
- static data (pre-calculated)
- weather forecasts and current weather data



Planned publication: *Journal of Computer Science and Technology – K. Kenda et al.: „Stream Pre-processing Framework for Heterogeneous Data Sources" (2018).*

Smart Energy Forecasting on a Stream

Architecture

- Big Data ready forecasting infrastructure
- Kafka for scalable dat exchange
- Loosely-coupled components architecture
- Modeling componets
- Reporting DB