REMI:
Reusable Elements for Multi-Level Information Availability

Avigdor Gal
Nicolo Rivetti
Arik Senderovich
Technion - Israel Institute of Technology
[avigal|nriveti|sariks]@technion.ac.il

Dimitrios Gunopulos
Ioannis Katakis
Nikolaos Panagiotou
University of Athens
[dg|katak|n.panagiotou]@di.uoa.gr

Vana Kalogeraki
Athens University of Economics and Business
vana@aueb.gr

Introduction

Smart City

- Smart Cities are an enticing use-case of Big Data event-based methods and technologies.
- Urban data has many sources and can be leveraged to:
  - detect disasters;
  - monitor special;
  - improve the city efficiency.
- These **issues are common** to most medium-to-large sized cities.
- A solution should be as **flexible** as possible in terms of the **available data** to allow **portability** despite the **heterogeneity** of different city ecosystems.

- A major obstacle to a high portability involves **the differences in the levels of available information**, given the variety of data sources that are available in a particular city.
- We present **REMI**, a **reusable elements** framework to handle **varying degrees of information availability** by design from two complementary angles:
  - Graceful Degradation (**GRADE**) and;
  - Data Enrichment (**DARE**).

Motivation

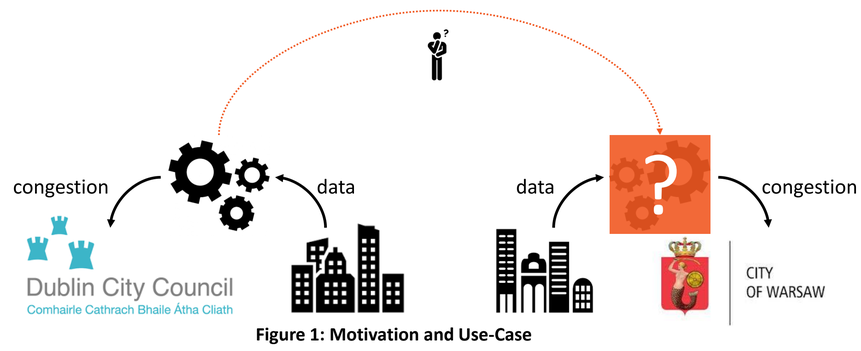

Figure 1: Motivation and Use-Case

- As an example of a Smart City application, we focus on predicting congestions in cities.
- Our experience in the INSIGHT[1] and VaVeL[2] European projects shows that deploying an algorithm that was developed in one city (Dublin in this case), in another (Warsaw), becomes at times impossible due to different levels of data availability.
- Therefore, it is clear that not taking into account the information availability may most likely jeopardize the generalization of any Big Data event-based systems.
- Even across different areas of the same city we may find different levels of availability. For instance, some sensors (measuring traffic, capturing video, etc.) may be accessible in a suburb while not being available in another, due to either cost constraints or infrastructure failures.

[1] http://www.insight-ict.eu/
[2] http://www.vavel-project.eu/

REMI

- Reusable elements should be **designed to adapt** to **various information levels**, depending on the **available data sources**.
- Prior to algorithmic solutions, we consider the **layering of possible information availability**.
  - This involves **identifying the minimum data requirements**, denoted layer L_0, which **is necessary to provide a meaningful answer** to the targeted problem and that can **be safely assumed to be available** in any deployment setting.
  - On top of layer L_0 , one can design levels of **increased information availability** (layers L_i ,$i > 0$), adding new data sources that may be **less available** or, alternatively, **more expensive**.
- Figure 2 shows an implementation of the aforementioned information layering approach, as applied to our transportation use-case.
  - We assume that the basic information needed is a stream of trams' GPS positioning, without which it would be impossible to perform on-line spatio-temporal analysis.
- Constructed on top of the information layering approach, the REMI allows to seamlessly analyze congestion given various information levels using **graceful degradation** and/or **data enrichment**.

Data Source Layering

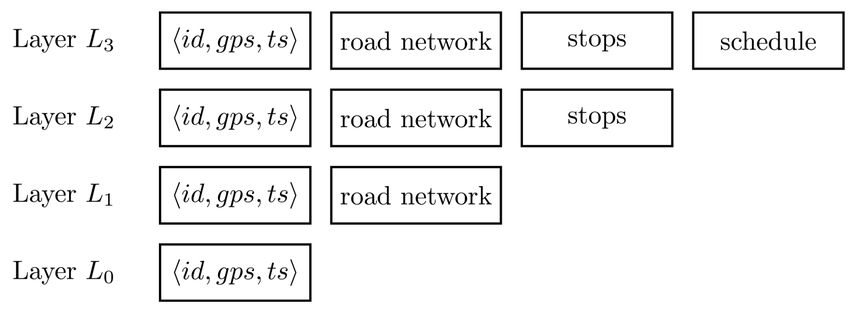

Figure 2: Layers of information availability.

- The data sources for the congestion analysis are modelled as follows:
  - The most basic layer of information (Layer L_0) contains the GPS sensor readings from buses (Dublin) and trams (Warsaw).
  - The next layer (Layer L_1) adds the network of roads in Warsaw and Dublin
  - The third layer of information (Layer L_2) contains the positions of the trams stops across both cities.
  - The top most layer (Layer L_3) adds the scheduled arrival and departure time for each tram line.

Grade and Dare Architecture

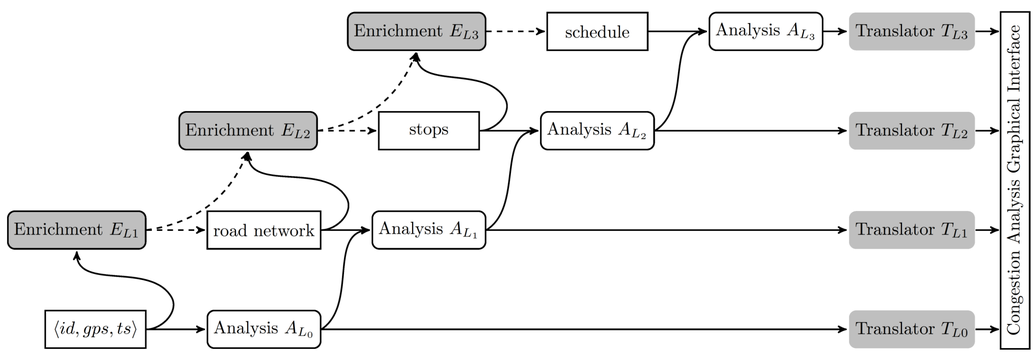

Figure 3: Reusable Elements: GRADE vs. DARE.

- The **graceful degradation** (GRADE) approach involves **decreasing the accuracy** of the output to **cope with the unavailability** of some data sources.
  - Algorithm A_{L_0} is unaware of the roads. Hence, it considers slowdowns as congestion.
  - Adding the road network on-top allows for a proper interpolation of these locations into congested road segments.
- The **data enrichment** (DARE) approach uses the **available data** in L_i to **infer the next (missing) layer** L_{i+1}. Unlike GRADE, the error when analyzing congestion with DARE comes from the inherent inference inaccuracies.
  - A_{L_1} provides a better congestion analysis, however it requires the road network knowledge.
  - To move from L_0 to in L_1 we infer an approximate road map from the GPS readings.

System Architecture

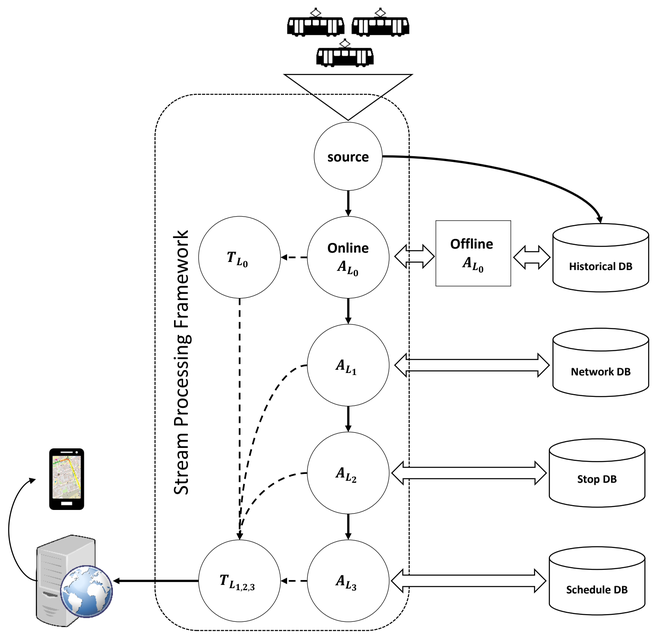

Figure 4: System Architecture.

- The additional data for $L_{i \geq 1}$ is stored in databases, namely: Network DB , Stop DB and Schedule DB .
- The stream σ of GPS readings (L_0) is collected and pushed into Apache Flink.
- The source duplicates the stream that is
  - stored in the Historical DB
  - and forwarded into the first layer analysis algorithm A_{L_0}.
- The analyzing algorithm for the following layers are simply pipelined and their output is processed by the translation operator before reaching the Web server.
- Finally, the GUI is based on a Webpage, displaying the cities maps with the results of the congestion analysis.

ACKNOWLEDGEMENT
This project received funding from the European Union Horizon 2020 Programme (Horizon2020/2014-2020), under grant agreement 688380