ICT Support for Adaptiveness and (Cyber)Security in the Smart Grid DAT300

An overview of Data Streaming

Vincenzo Gulisano

vinmas@chalmers.se (room 5124A)

Chalmers University of technology

Distributed Computing and Systems
Chalmers university of technology

Agenda

- Motivation
- The data streaming philosophy
- System Model
- Sample Data Streaming application
- Evolution of Stream Processing Engines
- Challenges in the context of Smart Grids

Agenda

- Motivation
- The data streaming philosophy
- System Model
- Sample Data Streaming application
- Evolution of Stream Processing Engines
- Challenges in the context of Smart Grids

Geekonomics
How big data put Obama back in the White House | Pages 8 and 9
The privacy debate
A scramble for information threatens individual freedoms | Page 5
BusinessTechnology
The power behind decisions
natu
BIG DATA
A Revolution That Will Transform How We Live, Work and Think
BIG DATA
Viktor Mayer-Schönberger and Kenneth Cukier
The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
Making Everything Easier!
Big Data
FOR
DUMMIES
A Wiley Brand
Learn to:
Judith Hurwitz
Alan Nugent
Dr. Fern Halper
Marcia Kaufman
COMPUTERWORLD
WHY BIG DATA IS A BIG DEAL
COVER STORY | BI & ANALYTICS

Motivation

- Applications such as:
  - Sensor networks
  - Network Traffic Analysis
  - Financial tickers
  - Transaction Log Analysis
  - Fraud Detection

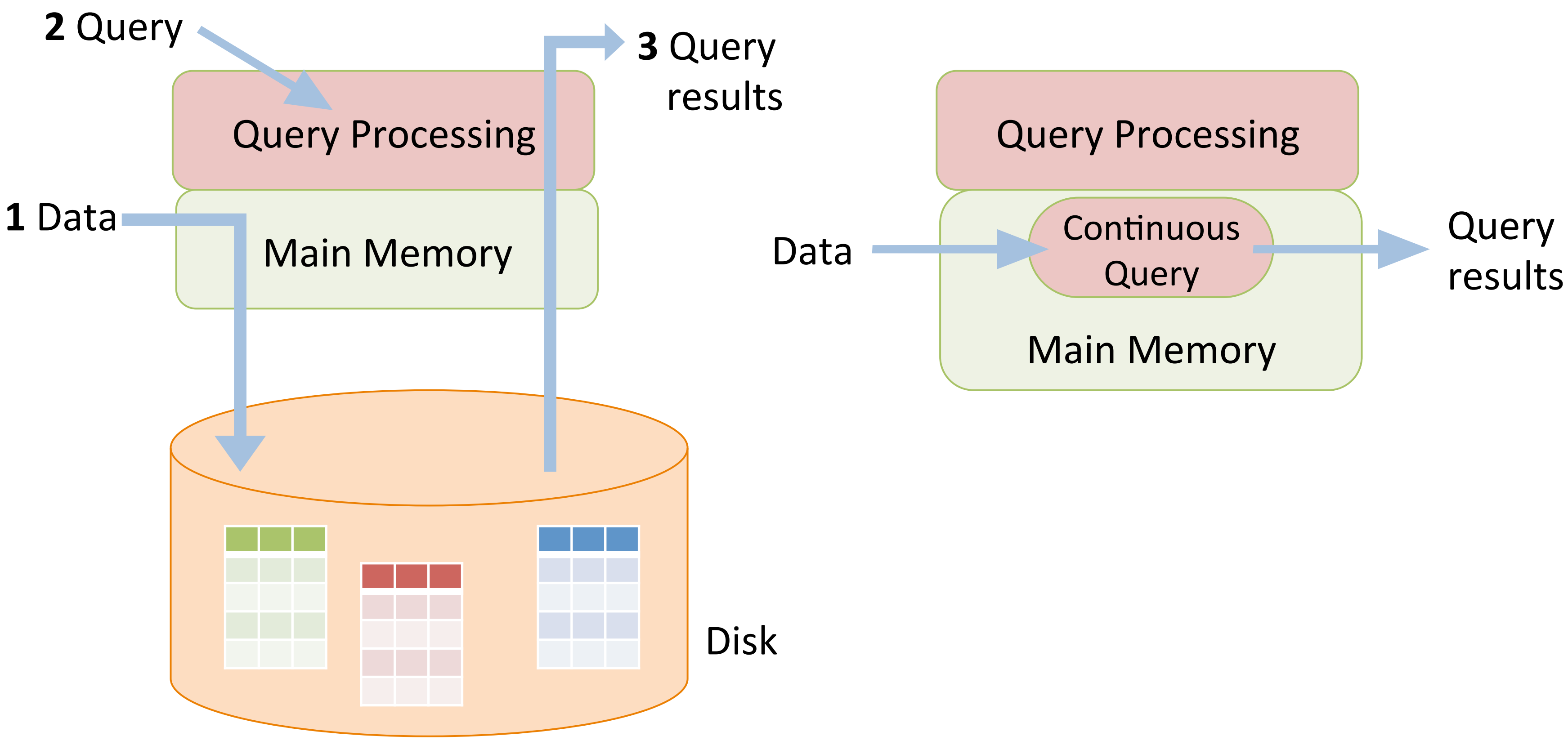
- Require:
  - Continuous processing of data streams
  - Real Time Fashion

Motivation

- Store and process is not feasible
  - high-speed networks, nanoseconds to handle a packet
  - ISP router: gigabytes of headers every hour,…
- Data Streaming:
  - In memory
  - Bounded resources
  - Efficient one-pass analysis

Motivation

- DBMS vs. DSMS



Agenda

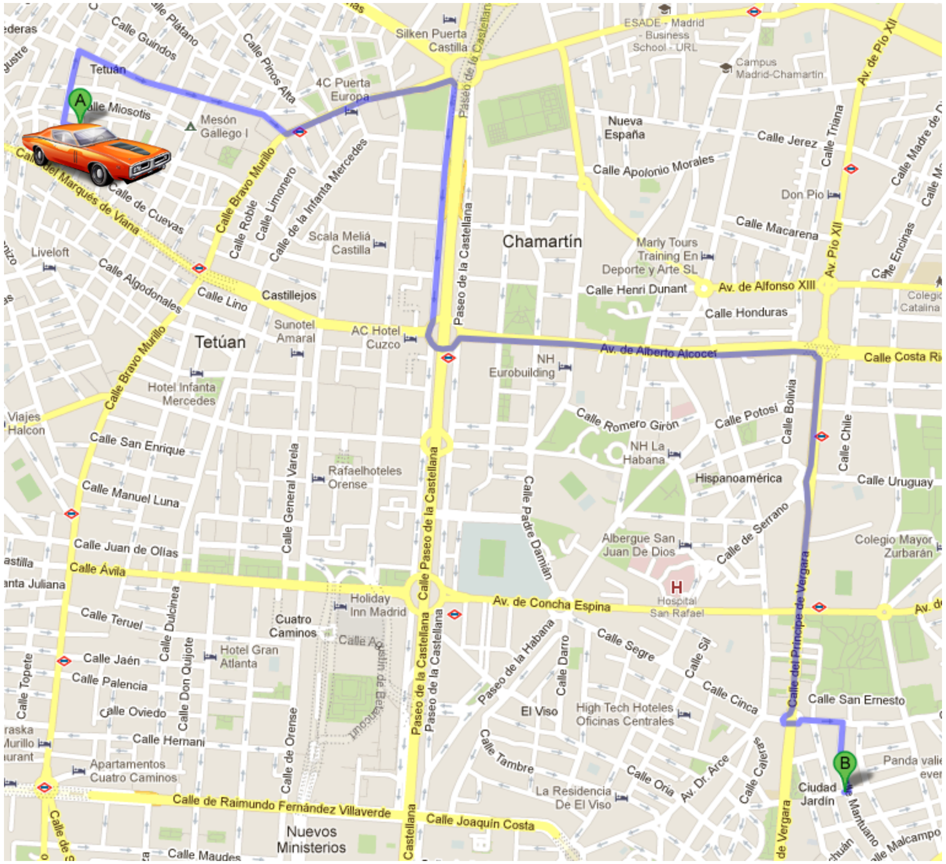
- Motivation
- The data streaming philosophy
- System Model
- Sample Data Streaming application
- Evolution of Stream Processing Engines
- Challenges in the context of Smart Grids

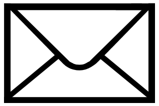
Database vs. Data Streaming

- Problem:
  - James travels by car from A to B
  - Mark is worried, he wants to know if he exceeds the speed limit

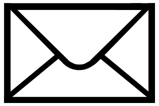
- How will a “database” and “data streaming” approach this?

Database vs. Data Streaming





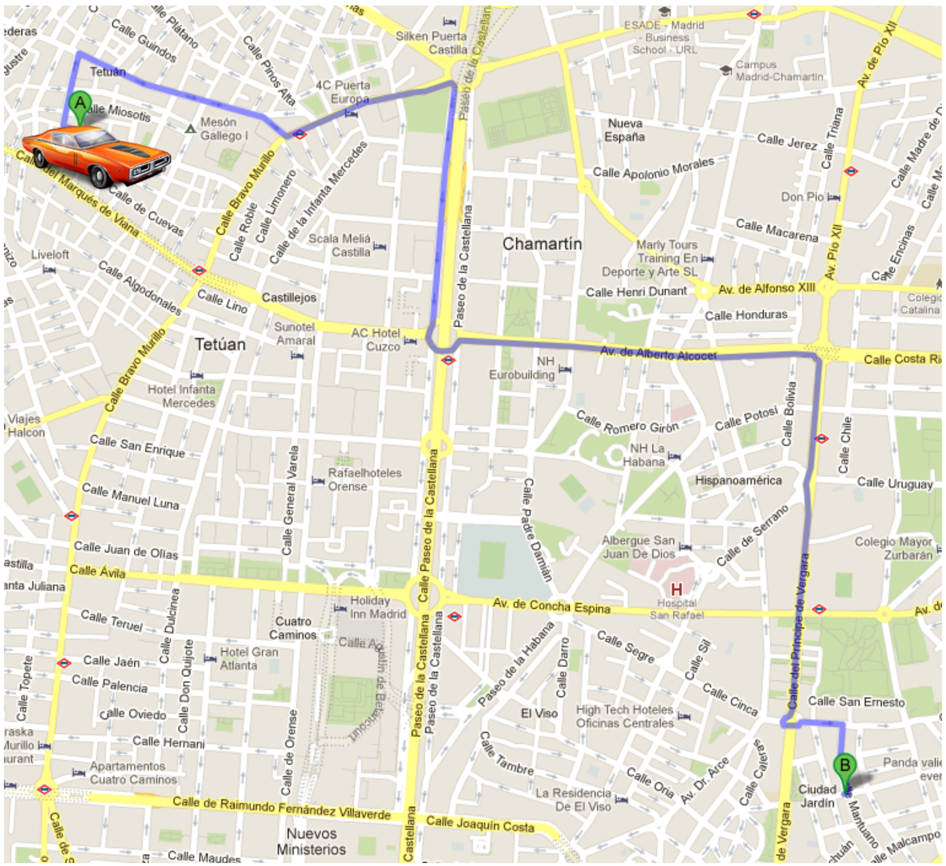
Start time
Position A



End time
Position B

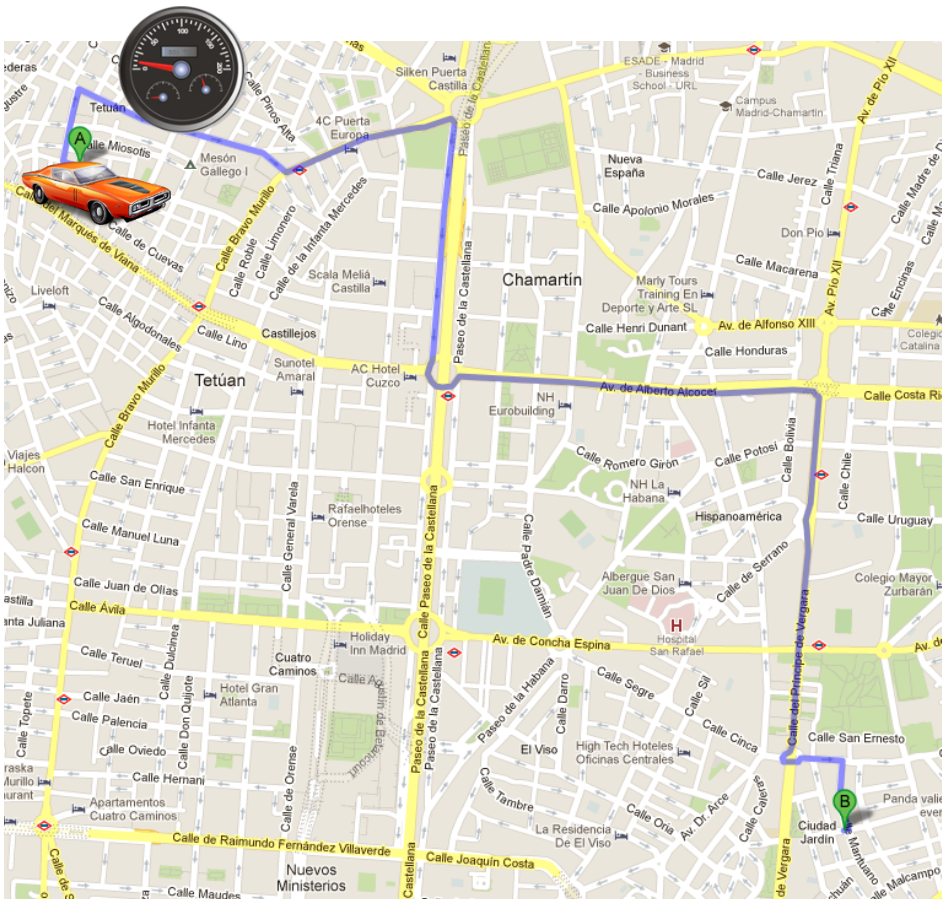
$$\frac{distance(A,B)}{End\ time - Start\ Time}$$

<u>Database</u> vs. Data Streaming



1. First the data, then the query
2. Precise result
3. Need to store information

Database vs. Data Streaming



1. First the query, then the data
2. “Continuous” result
3. No need to store information

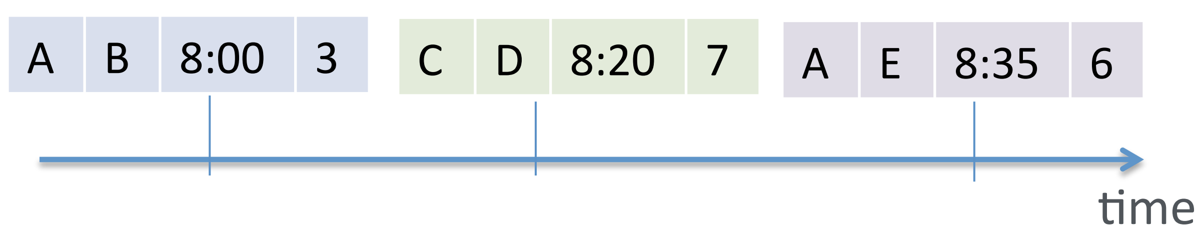
Agenda

- Motivation
- The data streaming philosophy
- System Model
- Sample Data Streaming application
- Evolution of Stream Processing Engines
- Challenges in the context of Smart Grids

System Model

- Data Stream: unbounded sequence of tuples
  - Example: Call Description Record (CDR)

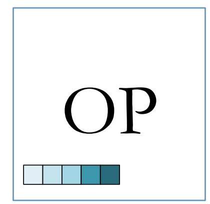
Field	Field
Caller	text
Callee	text
Time (secs)	int
Price (€)	double



System Model

- Operators:

Stateless
1 input tuple
1 output tuple



Stateful
1+ input tuple(s)
1 output tuple

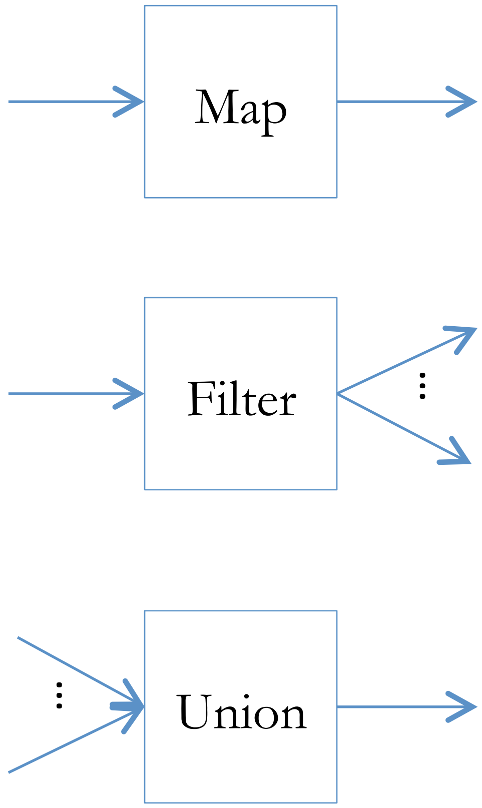
System Model

Stateless Operators

Map: transform tuples schema
Example: convert price € → $

Filter: discard / route tuples
Example: route depending on price

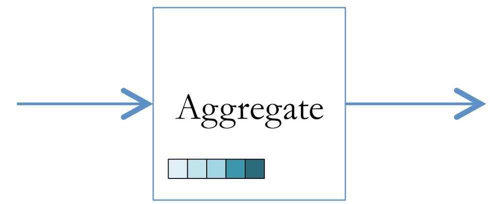
Union: merge multiple streams
(sharing the same schema)
Example: merge CDRs from
different sources



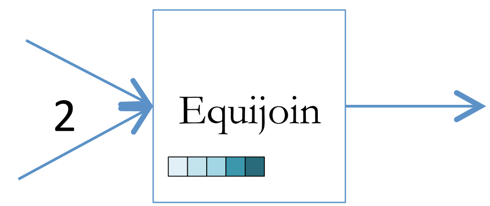
System Model

Stateful Operators

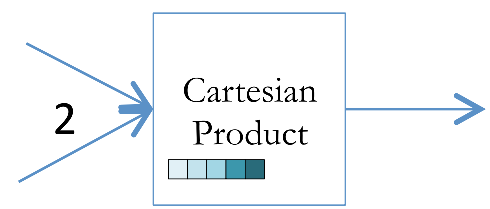
Aggregate: compute aggregate functions (group-by)
Example: compute avg. call duration



Equijoin: match tuples from 2 streams (equality predicate)
Example: match CDRs with same price

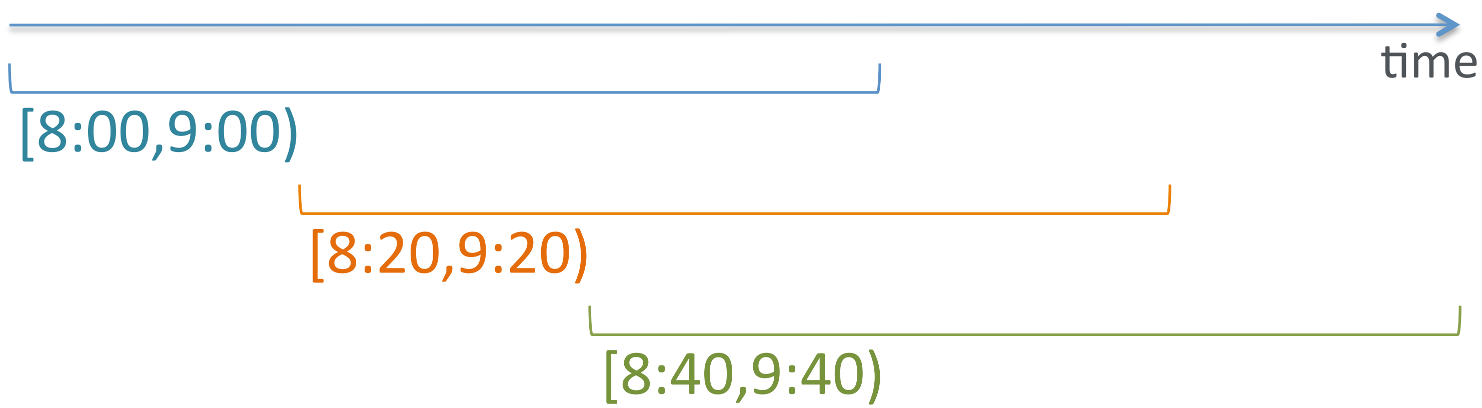


Cartesian Product: merge tuples from 2 streams (arbitrary predicate)
Example: match CDRs with prices in the same range



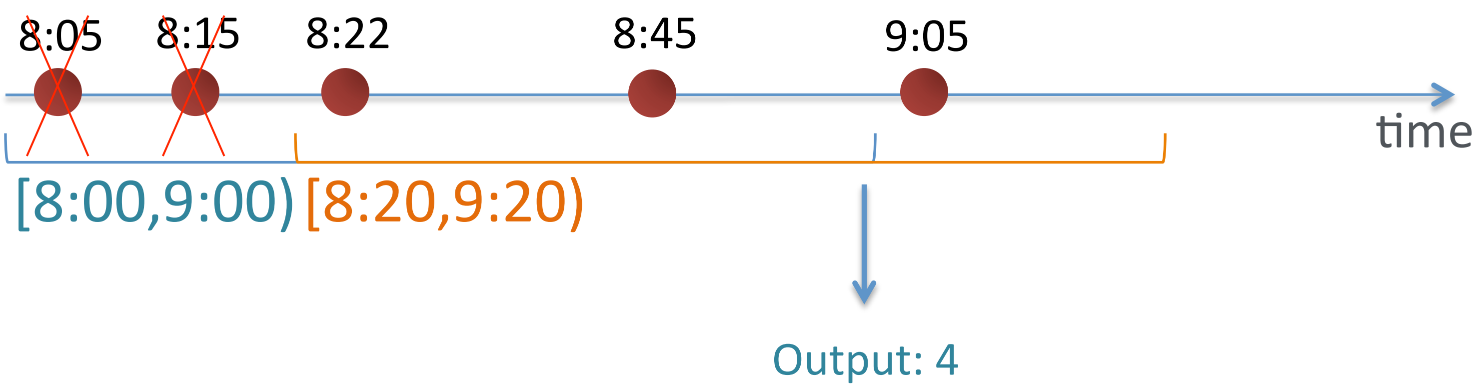
System Model

- Infinite sequence of tuples / bounded memory
  → windows
- Example: 1 hour windows



System Model

- Infinite sequence of tuples / bounded memory
  → windows
- Example: count tuples - 1 hour windows



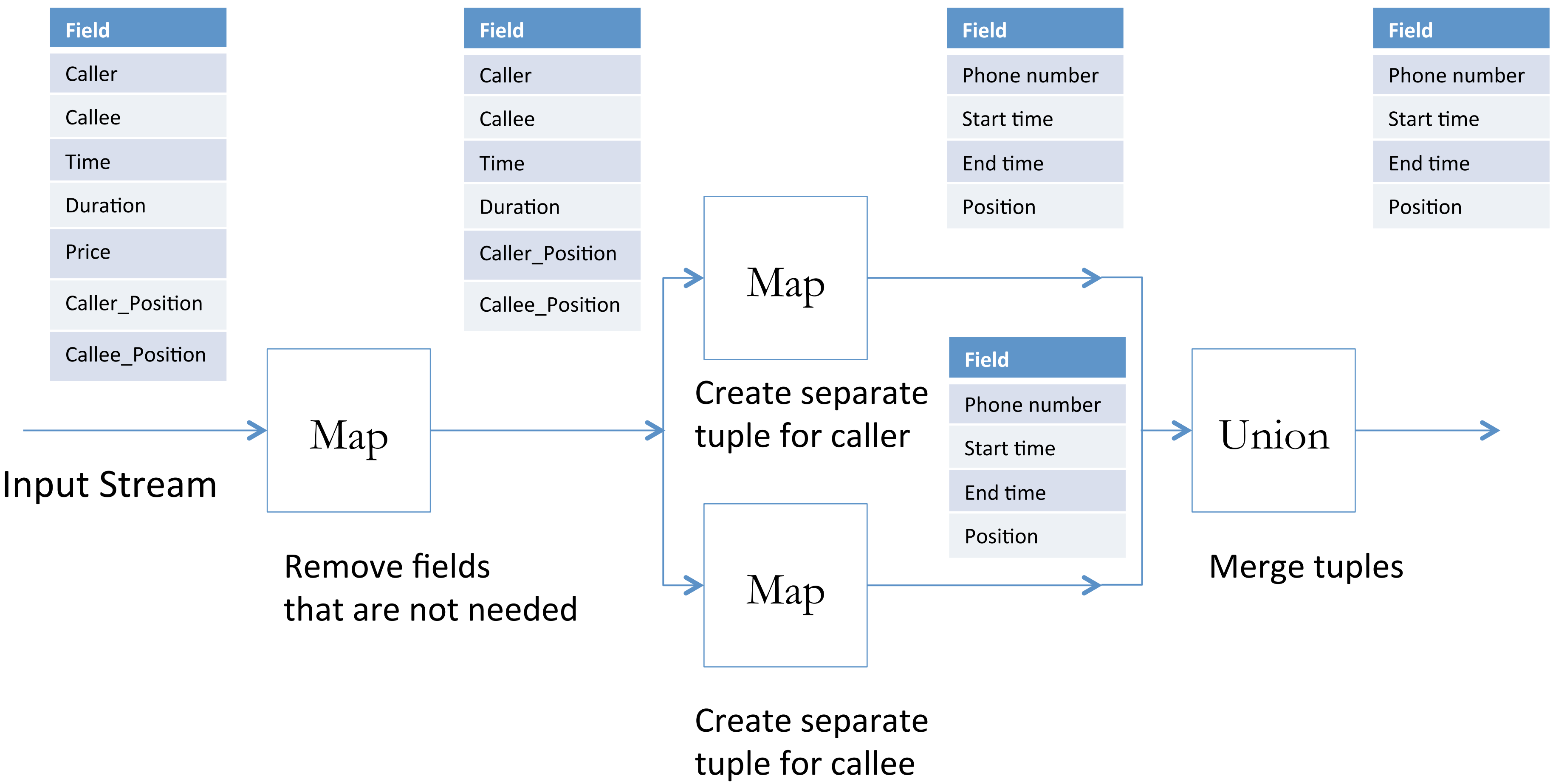
Agenda

- Motivation
- The data streaming philosophy
- System Model
- Sample Data Streaming application
- Evolution of Stream Processing Engines
- Challenges in the context of Smart Grids

Continuous Query Example

- Fraud detection, High Mobility
  - Spot mobile phone whose space and time distance between two consecutive calls is suspicious

High Mobility Continuous Query (1/2)

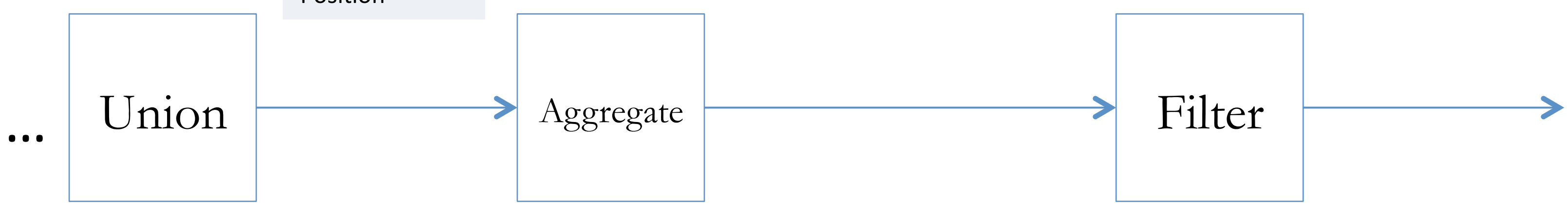


High Mobility Continuous Query (2/2)

Field
Phone number
Start time
End time
Position

Field
Phone number
Time
Speed

Field
Phone number
Time
Speed



Merge tuples

For each consecutive pair of calls referring to the same number compute speed

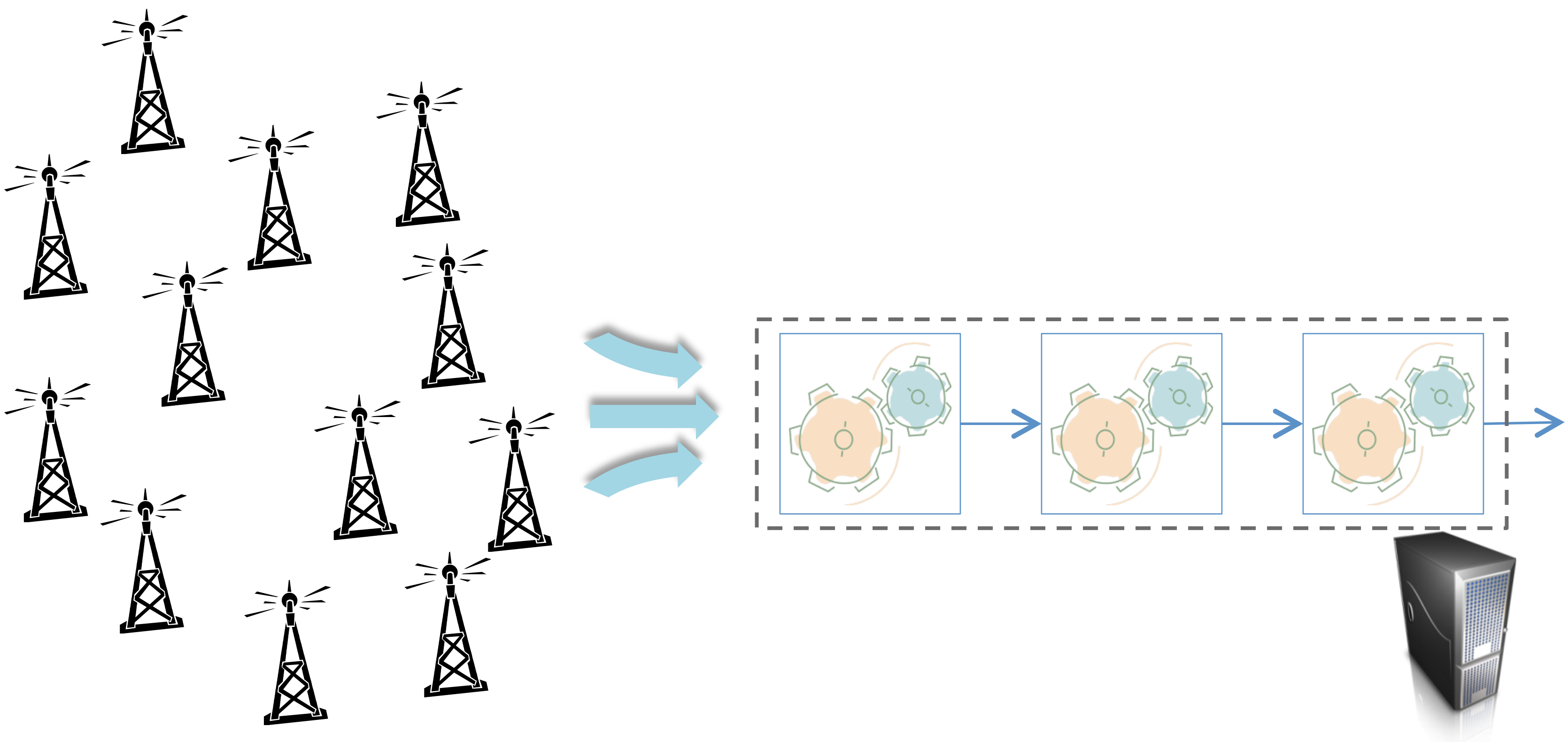
Window type: tuple based
Window size: 2
Window Advance: 1

Forward tuples with speed exceeding a given threshold

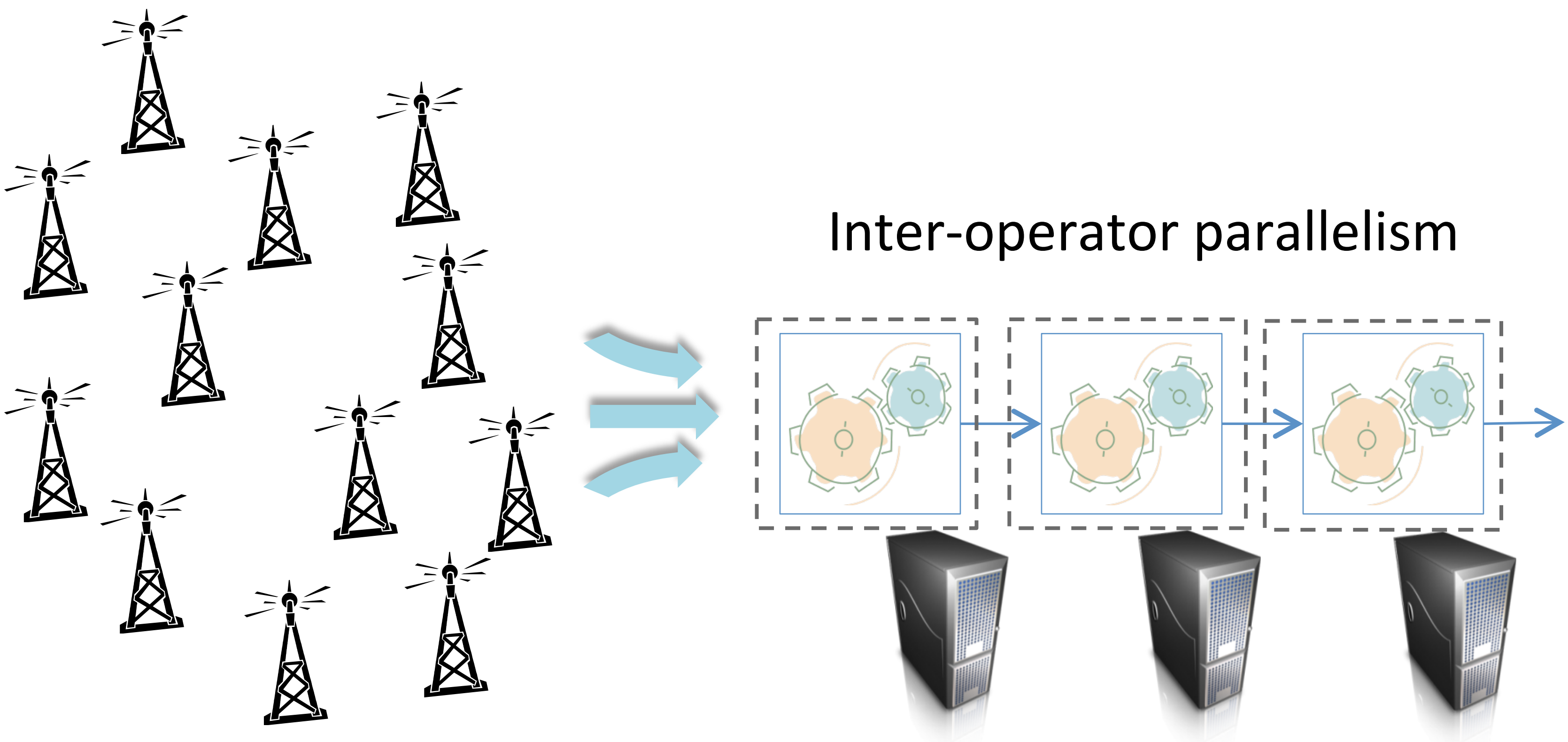
Agenda

- Motivation
- The data streaming philosophy
- System Model
- Sample Data Streaming application
- Evolution of Stream Processing Engines
- Challenges in the context of Smart Grids

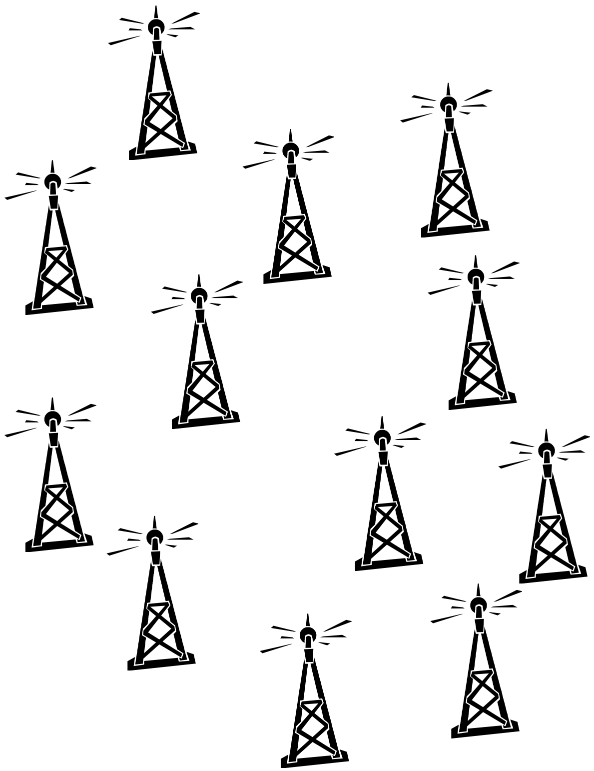
Centralized SPEs



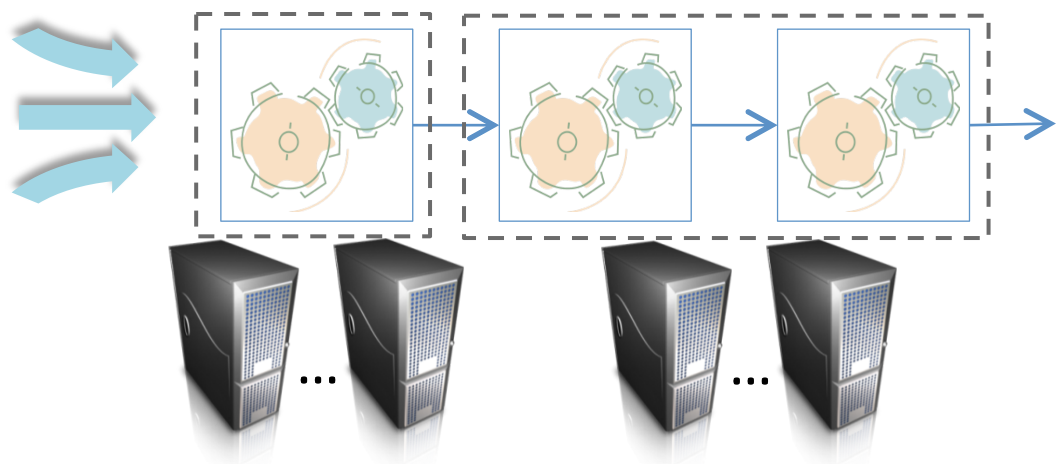
Distributed SPEs



Parallel SPEs



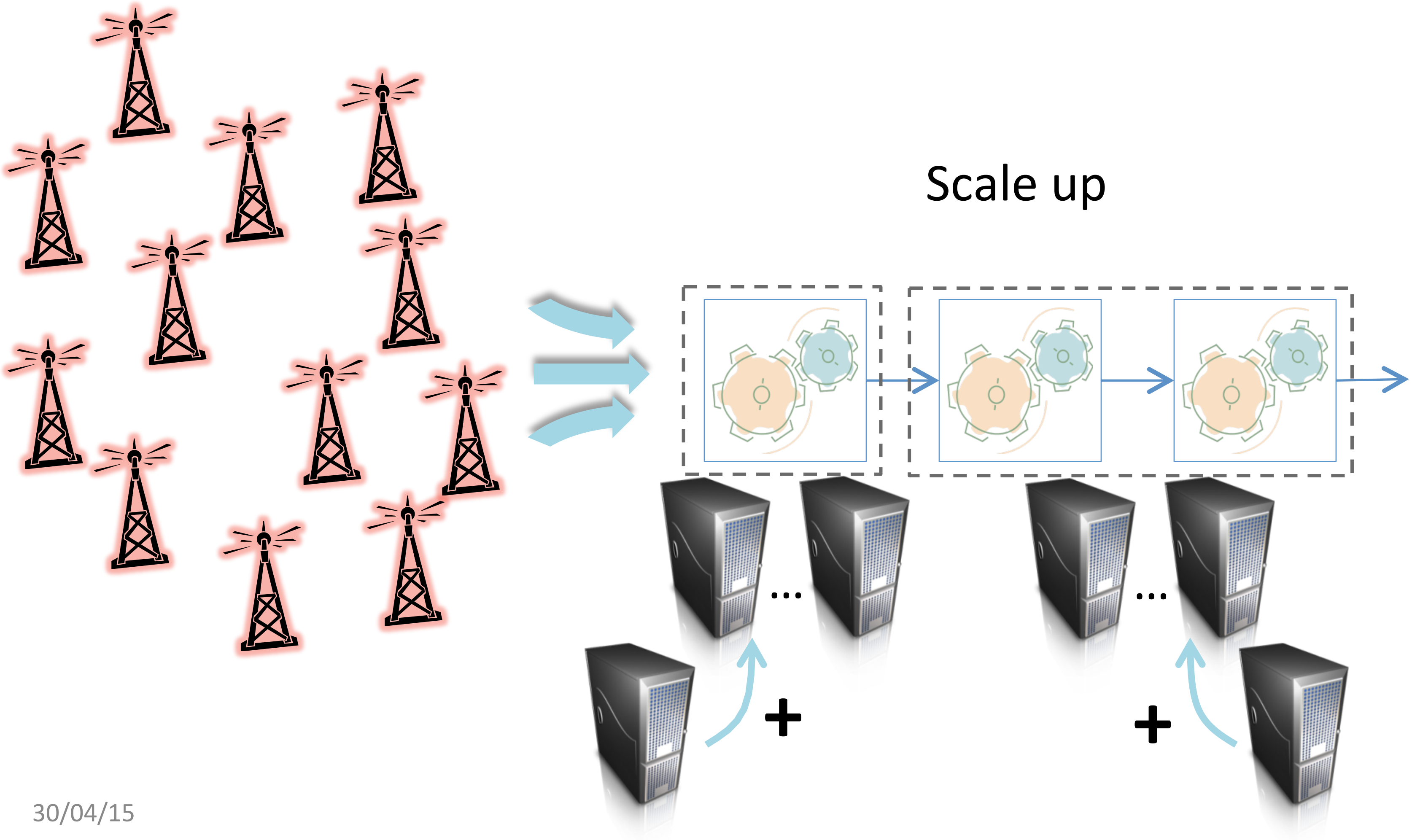
Intra-operator parallelism



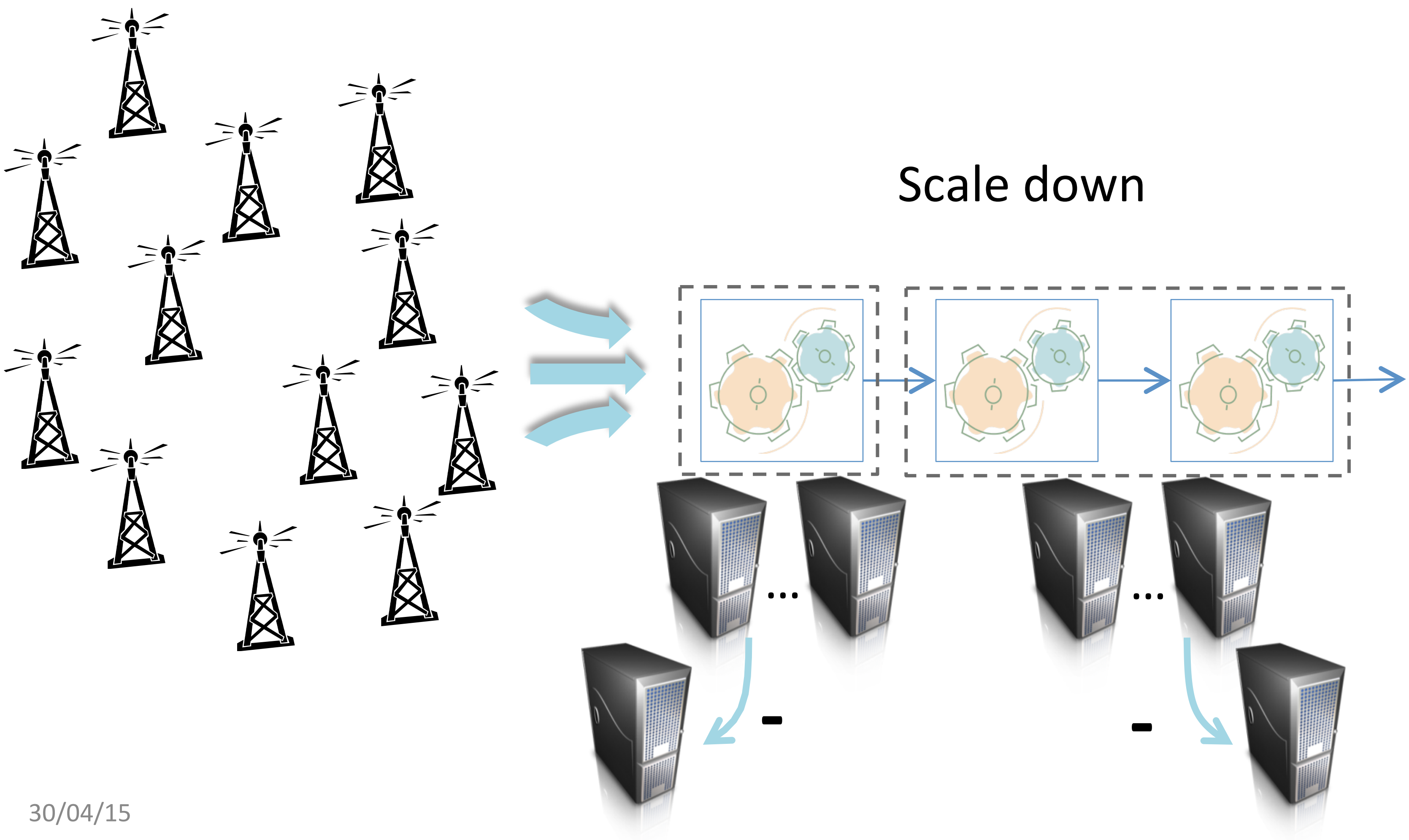
Over-provisioning or under-provisioning?

Elastic SPEs

Scale up



Elastic SPEs



Agenda

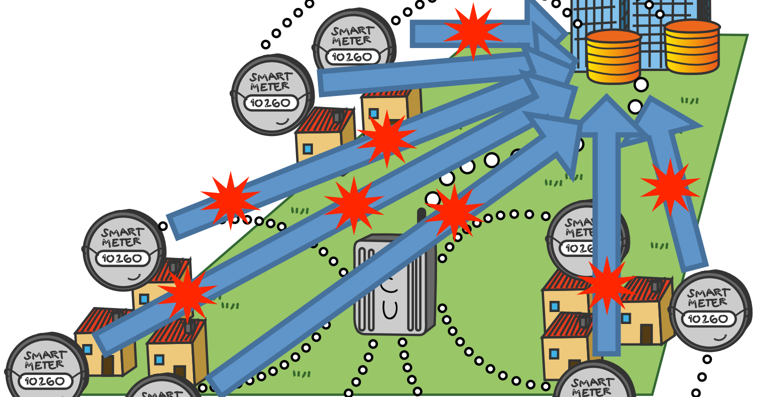
- Motivation
- The data streaming philosophy
- System Model
- Sample Data Streaming application
- Evolution of Stream Processing Engines
- Challenges in the context of Smart Grids

Online and Scalable Data Validation in Advanced Metering Infrastructures

Vincenzo Gulisano, Magnus Almgren and Marina Papatriantafilou

Chalmers University of technology

Online and Scalable Data Validation in Advanced Metering Infrastructures



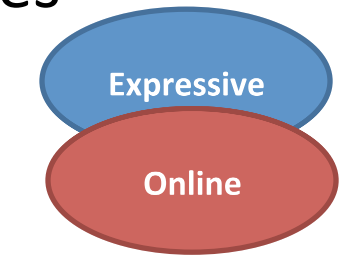
Noisy and Lossy data: bad-calibrated / faulty devices, lossy communication, malicious users, …

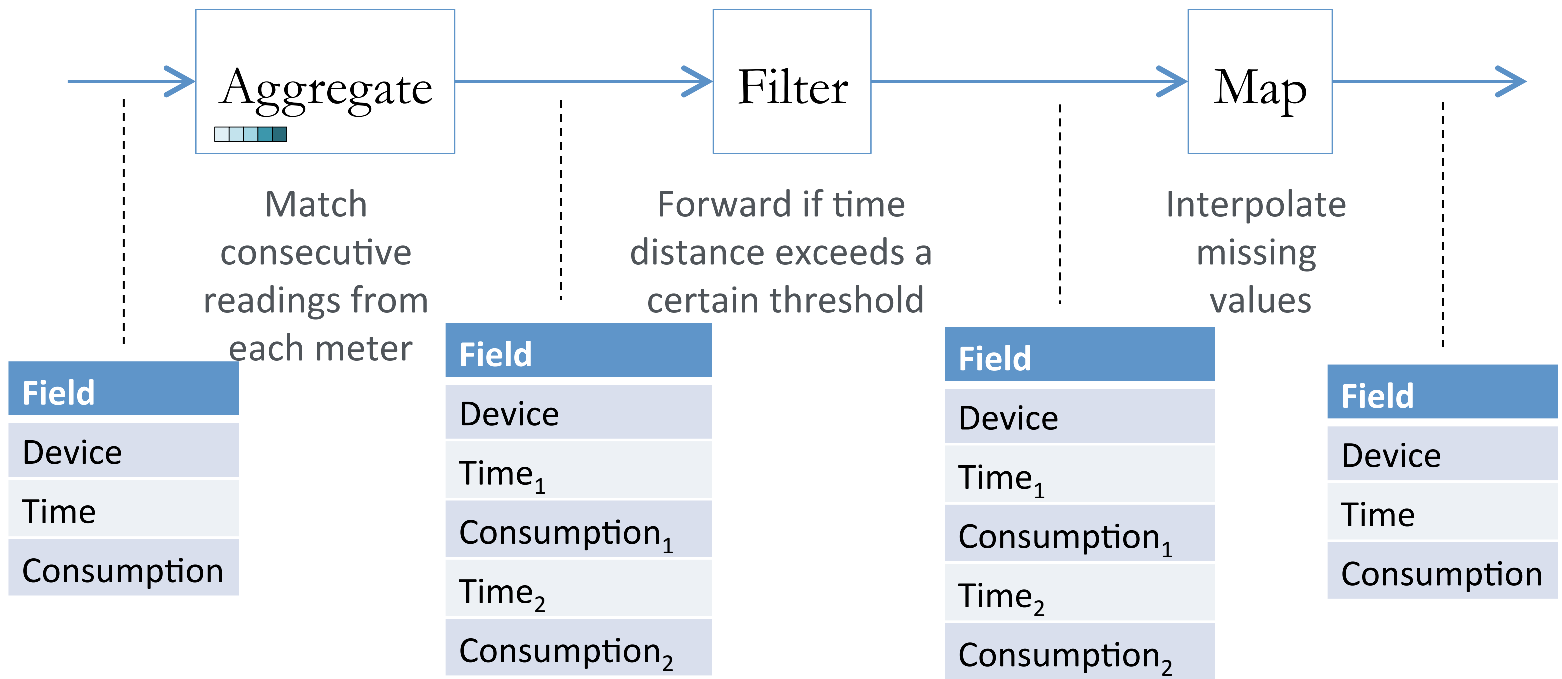
How can we validate data given that…

- There is a large volume of *continuous* data demanding for distributed and parallel analysis
- Validation rules depend on installation-specific features such as brands, devices, protocols, …
    - System experts should define installation-specific validation rules?

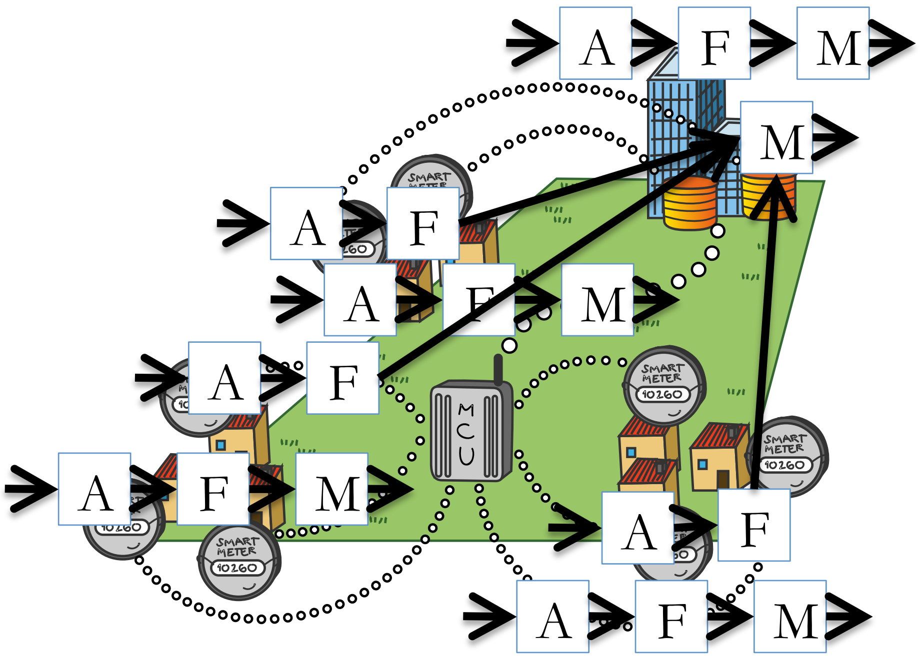
Why Streaming-based data Validation?

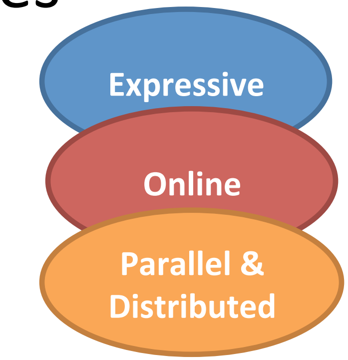
Sample Streaming-based Data Validation: Interpolate missing consumption values





Sample Streaming-based Data Validation: Interpolate missing consumption values





METIS: a Two-Tier Intrusion Detection System for Advanced Metering Infrastructures

Vincenzo Gulisano, Magnus Almgren and Marina Papatriantafilou

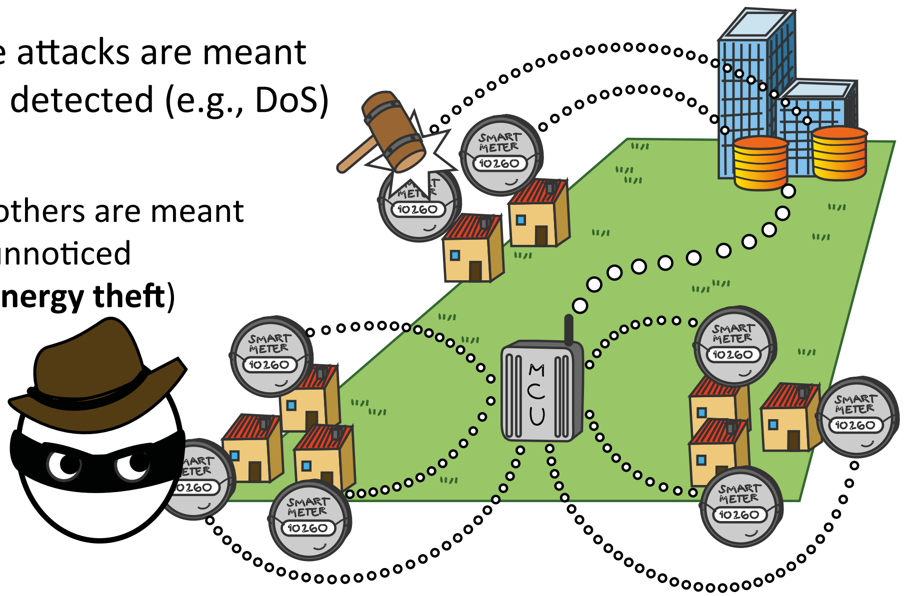
Chalmers University of technology

Why METIS?

Advanced Metering Infrastructures (AMIs)

Some attacks are meant to be detected (e.g., DoS)

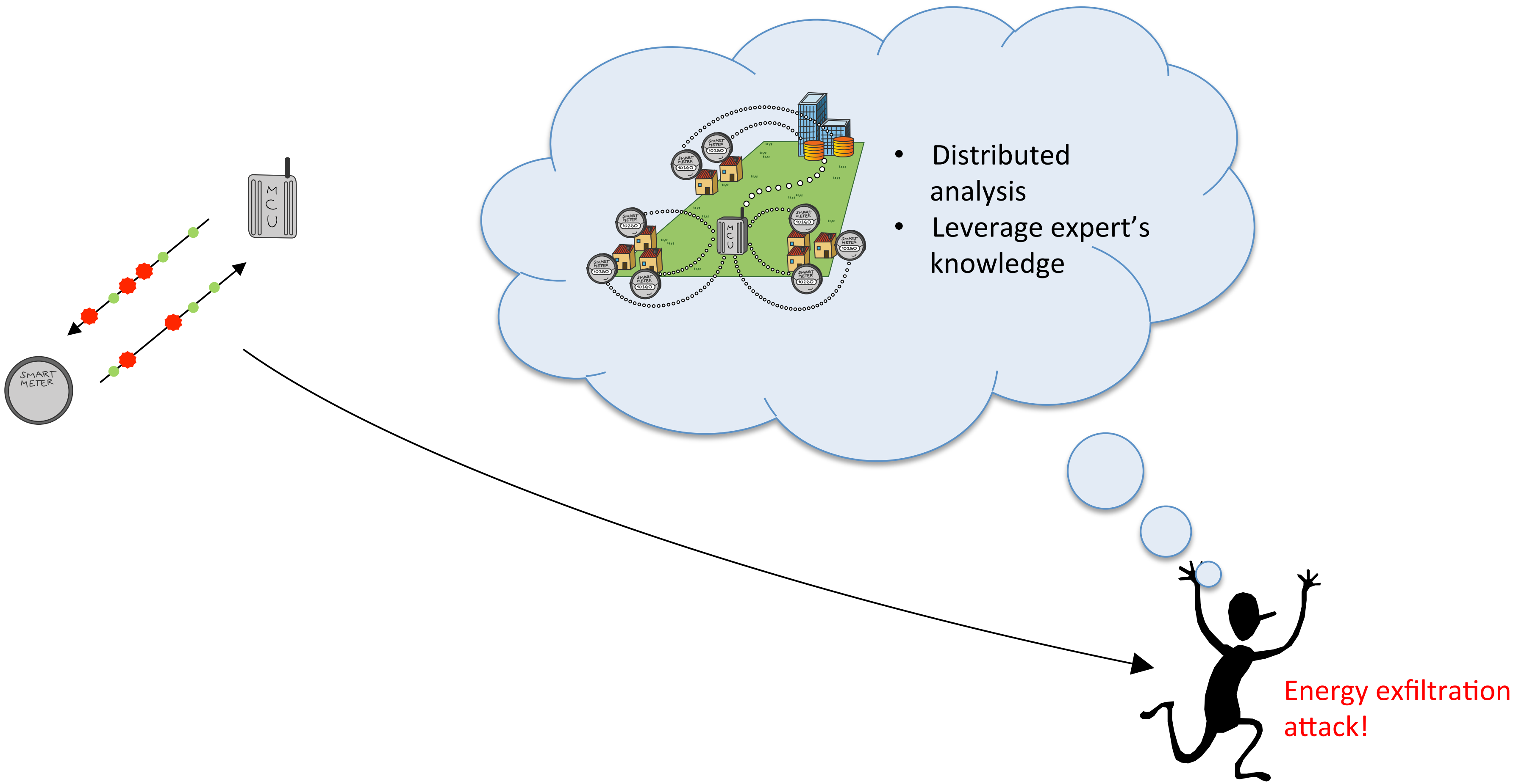
Some others are meant to go unnoticed (e.g., **energy theft**)



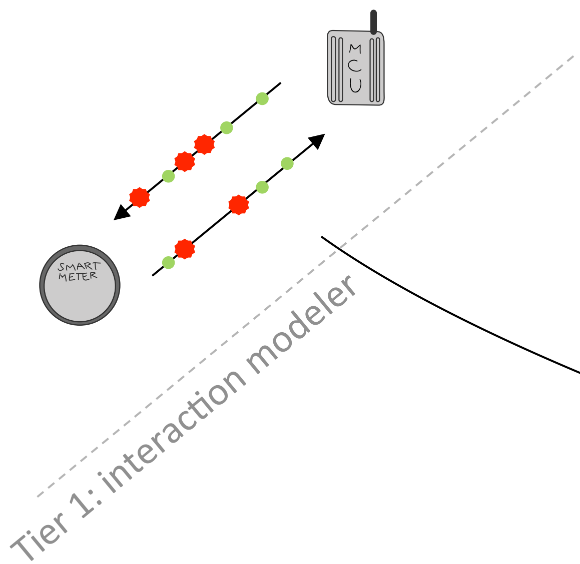
How can we detect them given that…

- … there is a large volume of *continuous* data demanding for distributed and parallel analysis
- … Most data is local to the devices
- … Such attacks are not documented
- … Each AMI relies on its brands, devices, protocols (i.e., system expert's knowledge plays a key role)

METIS overview

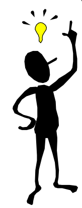


METIS overview



AMIs evolve slowly and have limited heterogeneity.
Let's learn which messages are expected and which are not!

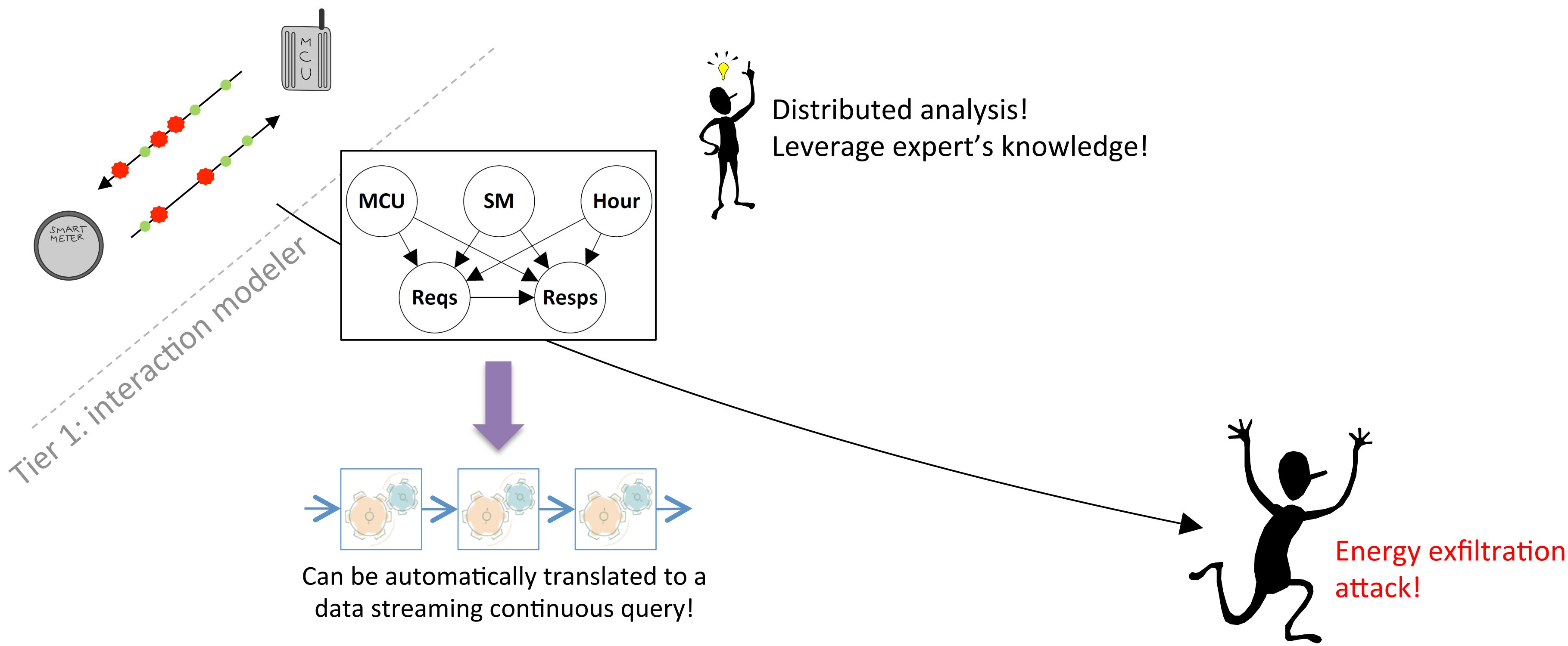
What influences an expected message?



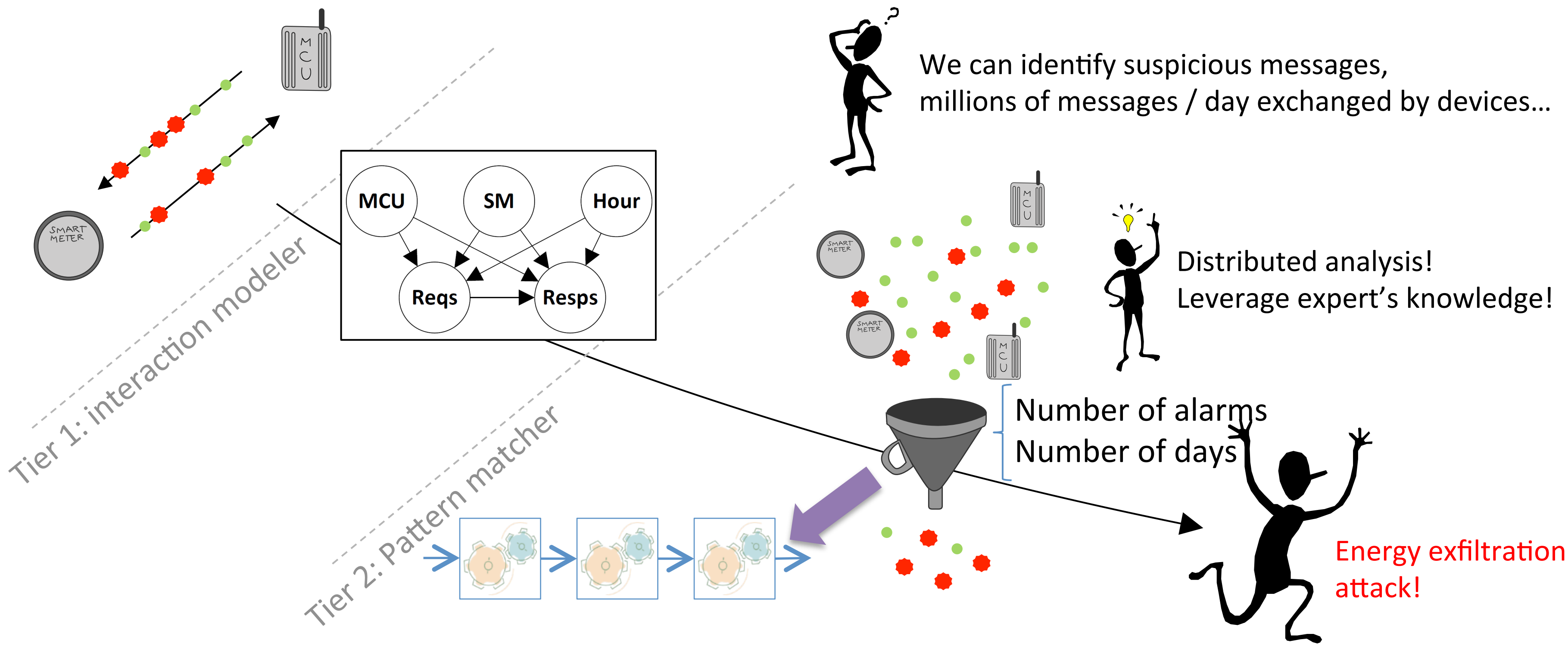
Let's use a Bayesian Network!

Energy exfiltration attack!

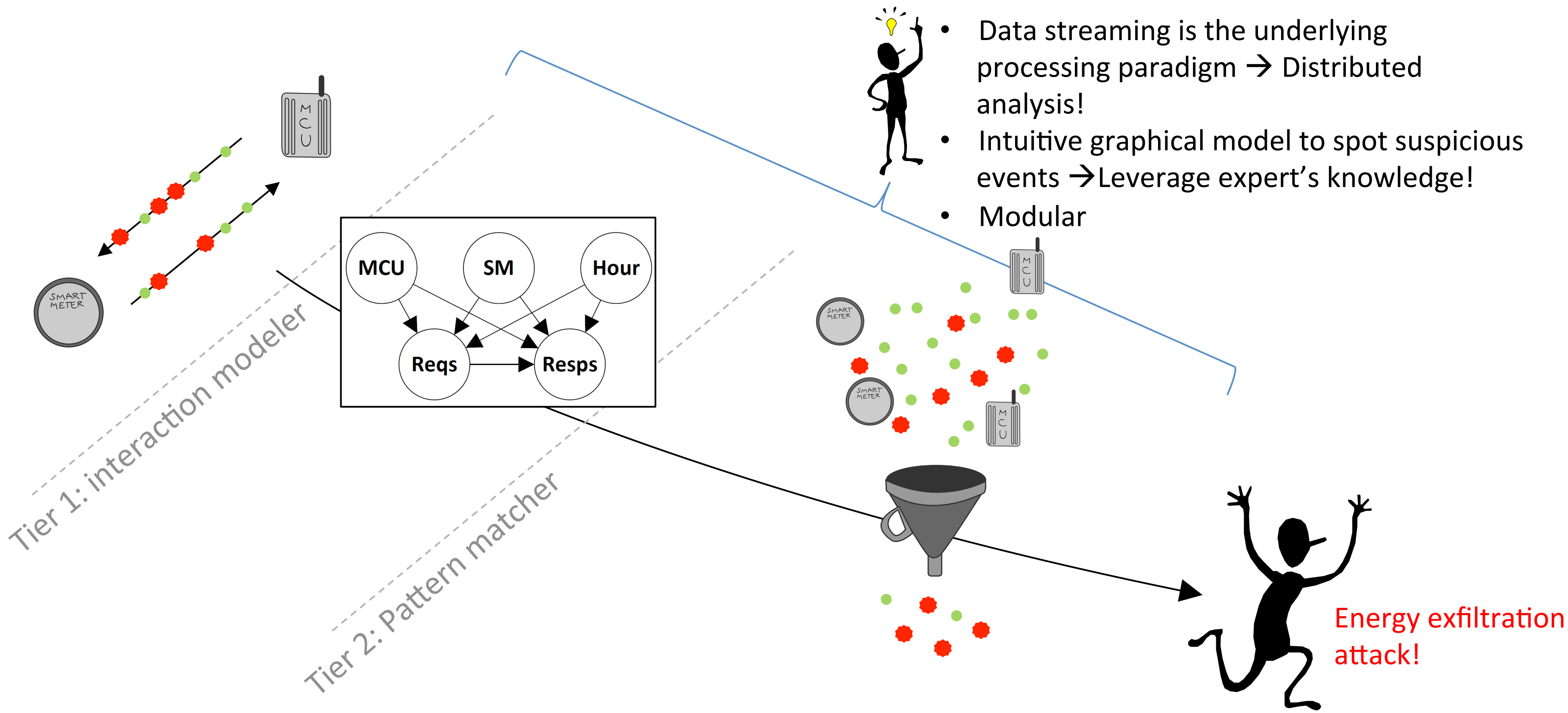
METIS overview



METIS overview



METIS overview



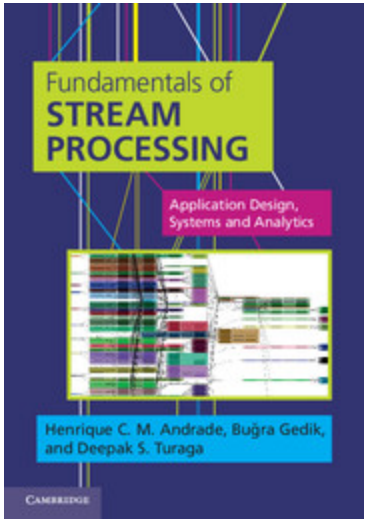
- Data streaming is the underlying processing paradigm → Distributed analysis!
- Intuitive graphical model to spot suspicious events →Leverage expert's knowledge!
- Modular

An overview of Data Streaming

Questions?

An overview of Data Streaming

- Something to read:



http://www.cambridge.org/se/academic/subjects/engineering/communications-and-signal-processing/fundamentals-stream-processing-application-design-systems-and-analytics

Bibliography

1. Online and Scalable Data Validation in Advanced Metering Infrastructures. Vincenzo Gulisano, Magnus Almgren, Marina Papatriantafilou. The 5th IEEE PES Innovative Smart Grid Technologies (ISGT) European 2014 Conference
2. METIS: a Two-Tier Intrusion Detection System for Advanced Metering Infrastructures. Vincenzo Gulisano, Magnus Almgren, Marina Papatriantafilou. 10th International Conference on Security and Privacy in Communication Networks (SecureComm) 2014
3. Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS ’02, New York, NY, USA, 2002. ACM.
4. Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS ’02, New York, NY, USA, 2002. ACM.
5. Michael Stonebraker, Uˇgur Çetintemel, and Stan Zdonik. The 8 requirements of realtime stream processing. SIGMOD Rec., 34(4), December 2005.
6. Nesime Tatbul. QoS-Driven load shedding on data streams. In Proceedings of the Worshops XMLDM, MDDE, and YRWS on XML-Based Data Management and Multimedia Engineering-Revised Papers, EDBT ’02, London, UK, UK, 2002. Springer-Verlag.
7. Arvind Arasu, Shivnath Babu, and Jennifer Widom. The CQL continuous query language: semantic foundations and query execution. The VLDB Journal, 15(2), June 2006.
8. Daniel J. Abadi, Don Carney, Ugur Cetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stan Zdonik. Aurora: a new model and architecture for data stream management. The VLDB Journal, 12(2), August 2003.
9. Arvind Arasu, Shivnath Babu, and Jennifer Widom. The CQL continuous query language: semantic foundations and query execution. The VLDB Journal, 15(2), June 2006.
10. Daniel J. Abadi, Don Carney, Ugur Cetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stan Zdonik. Aurora: a new model and architecture for data stream management. The VLDB Journal, 12(2), August 2003.

Bibliography

9. Vincenzo Gulisano, Ricardo Jiménez-Peris, Marta Patiño-Martínez, and Patrick Valduriez. Streamcloud: A large scale data streaming system. In ICDCS 2010: International Conference on Distributed Computing Systems, pages 126–137, June 2010.
10. Mehul Shah Joseph, Joseph M. Hellerstein, Sirish Ch, and Michael J. Franklin. Flux: An adaptive partitioning operator for continuous query systems. In In ICDE, 2002.
11. Vincenzo Gulisano, Ricardo Jimenez-Peris, Marta Patino-Martinez, Claudio Soriente, and Patrick Valduriez. Streamcloud: An elastic and scalable data streaming system. IEEE Transactions on Parallel and Distributed Systems, 99(PrePrints), 2012.
12. Thomas Heinze. Elastic complex event processing. In Proceedings of the 8th Middleware Doctoral Symposium, MDS ’11, New York, NY, USA, 2011. ACM.