The Berlin Institute for the Foundations of Learning and Data

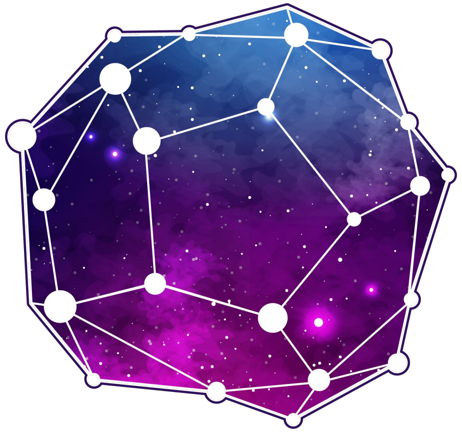
German Research Center for Artificial Intelligence

NebulaStream: Complex Analytics Beyond the Cloud

Steffen Zeuch, Eleni Tzirita Zacharatou, Shuhao Zhang, Xenofon Chatziliadis, Ankit Chaudhary, Bonaventura Del Monte, Dimitrios Giouroukis, Philipp M. Grulich, Ariane Ziehn , Volker Mark

What is this paper about

Core features enable the next generation of IoT applications but are not yet supported by state-of-the-art systems.



A general-purpose, end-to-end data management system for the IoT.

Domain specific features enable a richer set of applications over an IoT data management platform such as NES.

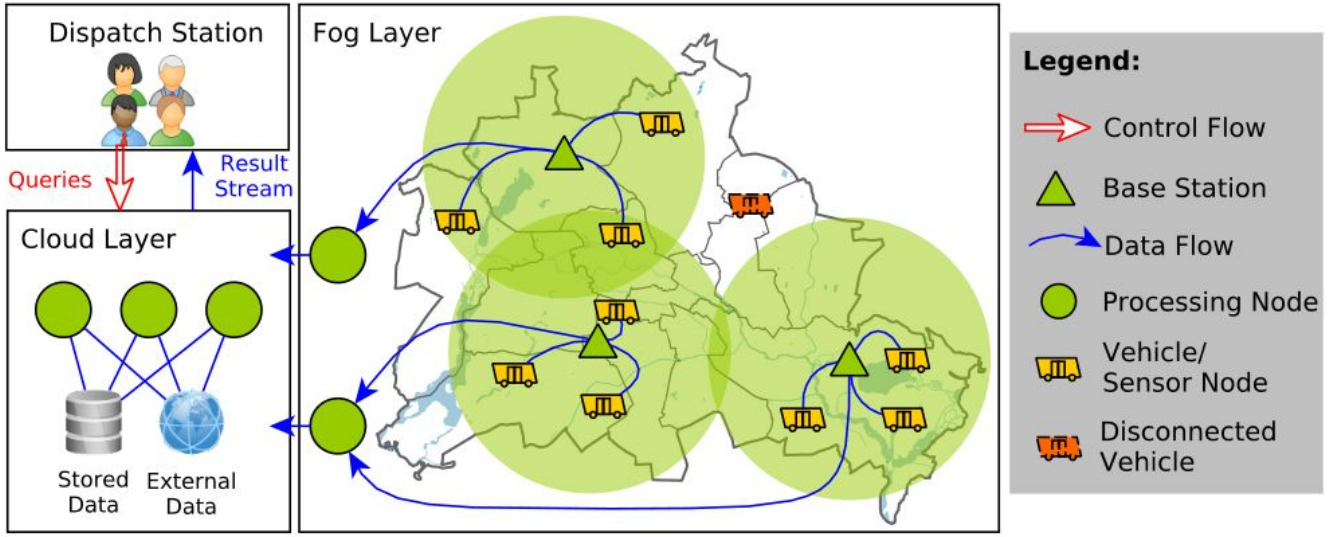
Smart Cities

- Each IoT device creates a data stream
- Example city: Berlin
  - Street lights **~200.000**
  - Traffic lights **~2.000**
  - Traffic sensors **~110.000**
  - Sensors in vehicles **~1.200.000**
  - Smartphones **~3.770.000**
  - …

Upcoming IoT applications

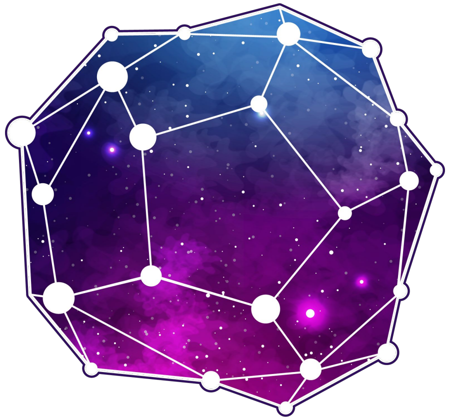
Highly Dynamic Execution Environment

Continuous Operation under Constant Evolution



Heterogeneity, Distribution, and Volume of Data and Compute

Core Features



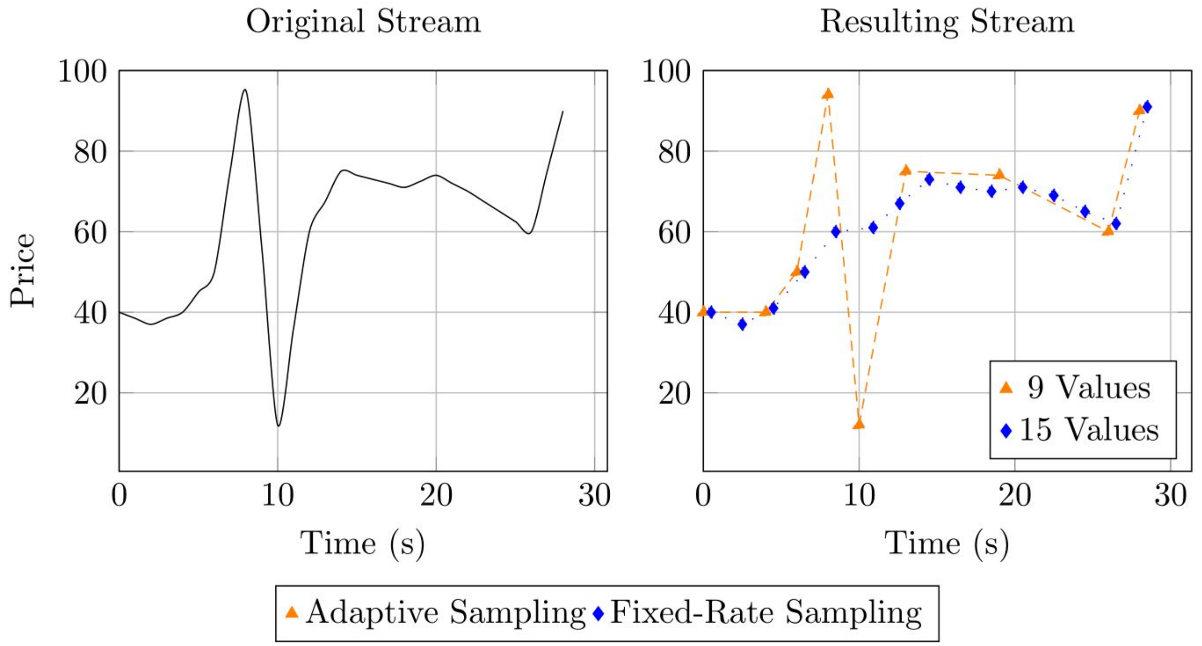
Core features enable the next generation of IoT applications but are not yet supported by state-of-the-art systems.

Adaptive Handling of Sensor Data Streams

- **General Description:** Adaptive sensor data handling allows scaling to large number of nodes and sensors while avoiding resource misuse.

- **State-of-the-Art Systems and Their Limitations:**
  - Assumes homogeneous hardware
  - Focuses on disseminating a single query
  - Does not exploit all sensor node capabilities

- **Enabling Emerging IoT Applications:**
  - Treat sensor nodes as first class components
  - Keep in check the dynamicity of the data while retaining high-quality representation of results
  - Enable more precise sampling for IoT applications

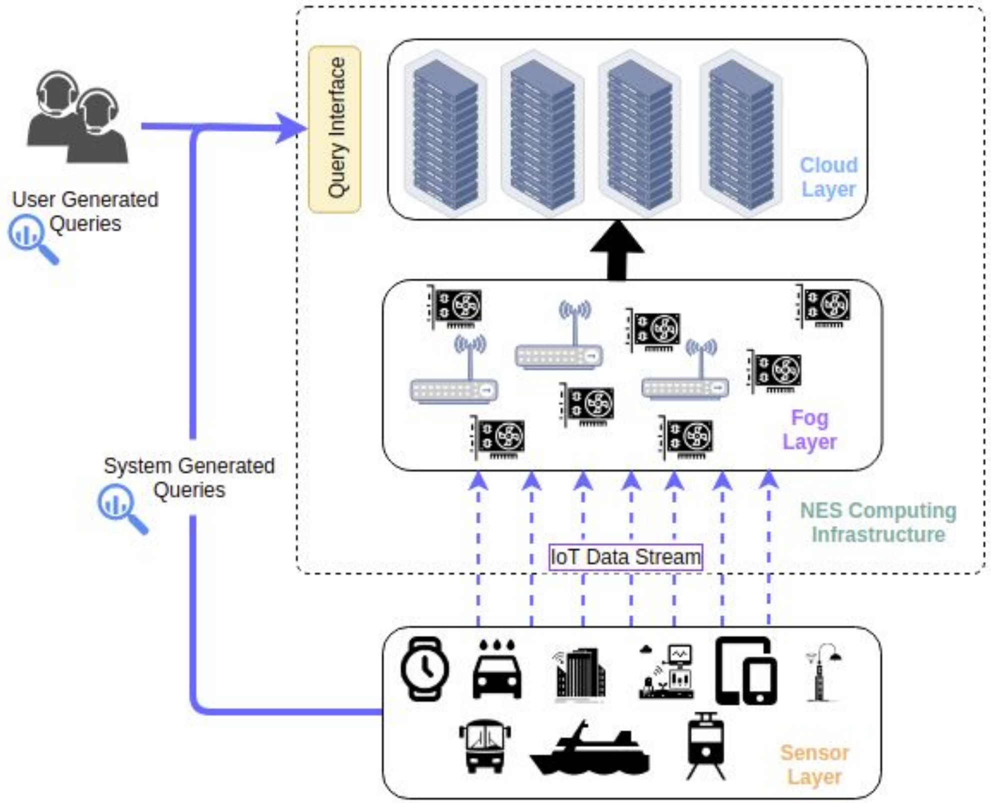


Massive Scalability

- **General Description:** Support thousands of queries on millions of heterogeneous and distributed data streams.

- **State-of-the-Art Systems and Their Limitations:**
  - Cloud-based systems are limited by the amount of data it can receive from IoT devices into cloud.
  - Fog-based systems process data closer to IoT devices but have limited computation resources.
  - Sensor-based systems provide only minimal functionality for data analytics.

- **Enabling Emerging IoT Applications:**
  - Large scale, real-time applications should leveraging sensors, fog, and cloud resources to process massive amount of geo-distributed IoT data streams, e.g., connected cars, smart cities.



Support for Heterogeneous Devices

- **General Description:** IoT environments consist of a wide range of diverse processing devices. Resource utilization is crucial for efficiency.

- **State-of-the-Art Systems and Their Limitations:**
  - Current systems are either hardware-oblivious or build for one specific hardware.
  - No system exploit the heterogeneous devices efficiently.
  - IoT environments introduce new challenges, e.g., device diversity and limited energy budget.

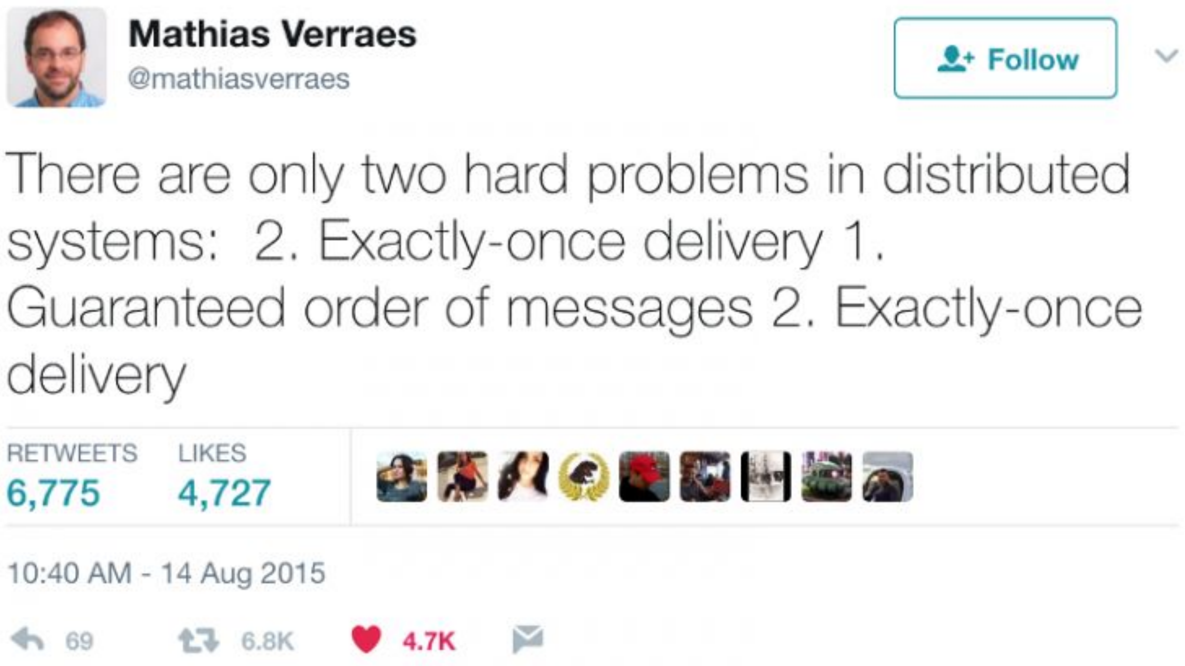
- **Enabling Emerging IoT Applications:**
  - Process data most efficiently and where it is generated to improve the overall IoT system efficiency. Thus, enabling larger IoT infrastructures such as smart cities.

Delivery guarantees

- **General Description:** IoT may require new forms of delivery guarantees beyond at-least-once or exactly-once delivery.

- **State-of-the-Art Systems and Their Limitations:**
  - Cloud-tailored solutions tailored with persistent storage, e.g., Kafka
  - Specific solutions for certain IoT scenarios

- **Enabling Emerging IoT Applications:**
  - Trade-off consistency, availability, and resource consumption
    - Temperature monitoring: at-most-once
    - Accident detection: at-least-once
    - Smart purchases: exactly-once



Secure & Private Stream Processing

- **General Description:** IoT devices are vulnerable to being attacked leading to secure and private issues.

- **State-of-the-Art Systems and Their Limitations:**
  - Prior systems (e.g., StreamBox-TZ, TimeCrypt) are not designed for IoT environments and their specific characteristics.

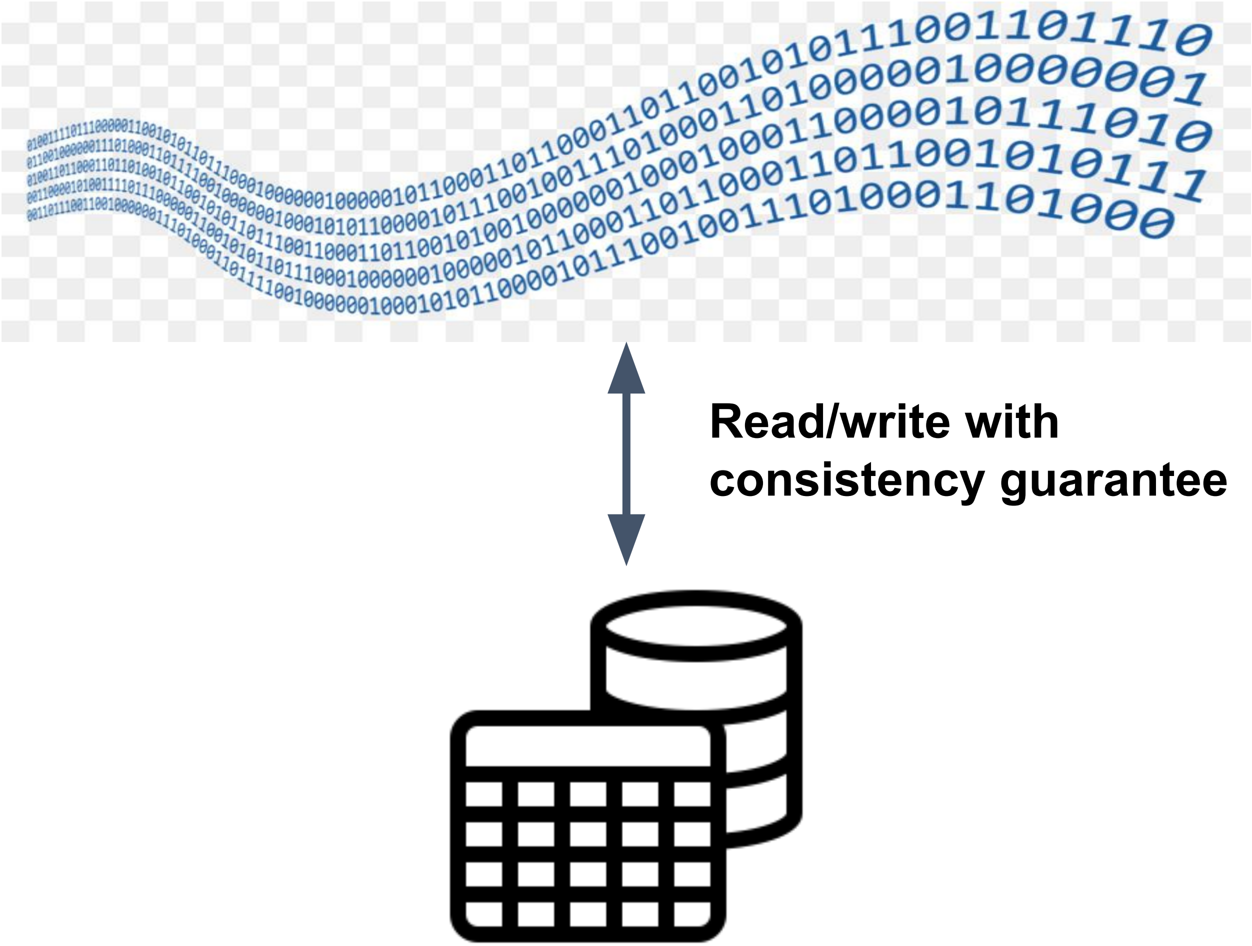
- **Enabling Emerging IoT Applications:**
  - Process confidential data by providing security and integrity guarantees in sensitive areas of smart city such as smart medication.

Transactional Stream Processing

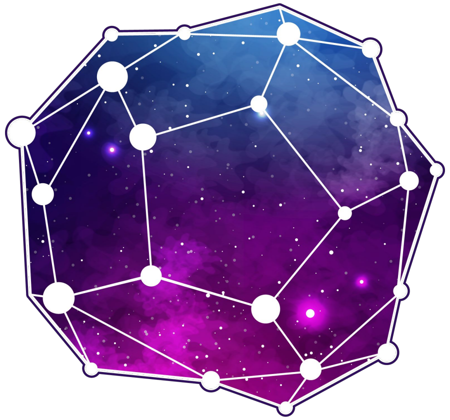
- **General Description:** SPEs with transactional state management would relieve the burden of managing state consistency from the users.

- **State-of-the-Art Systems and Their Limitations:**
  - Common problems in the IoT environment, e.g., transient node errors or unreliable connections, make existing solutions hardly applicable.

- **Enabling Emerging IoT Applications:**
  - Stream applications that require maintaining shared mutable states, e.g., self-driving vehicle monitoring as part of smart city development.



Domain Specific Features

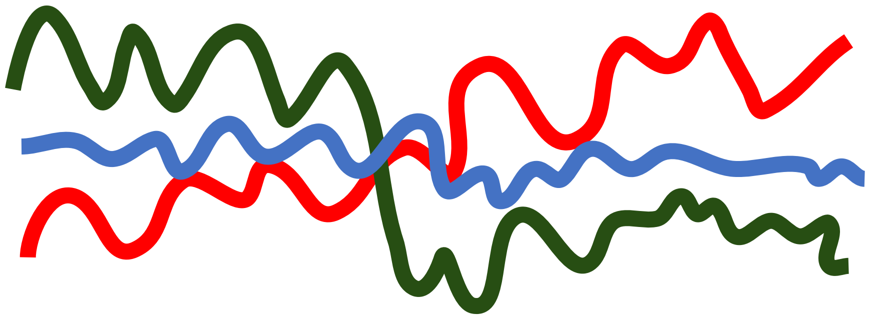


Domain specific features enable a richer set of applications over an IoT data management platform such as NES.

Digital Signal Processing (DSP)

- **General Description:** IoT applications mix relational and signal logic, e.g., filters, joins, group-by aggregates, interpolation for missing sensor values, noise reduction filters, FFT spectral analysis.

- **State-of-the-Art Systems and Their Limitations:**
  - No distributed SPE with support for DSP.
  - Memory and compute intensive DSP operators are unsuitable for low-end fog devices.
  - Non-commutative DSP operators require event-ordering, which is hard in the IoT.



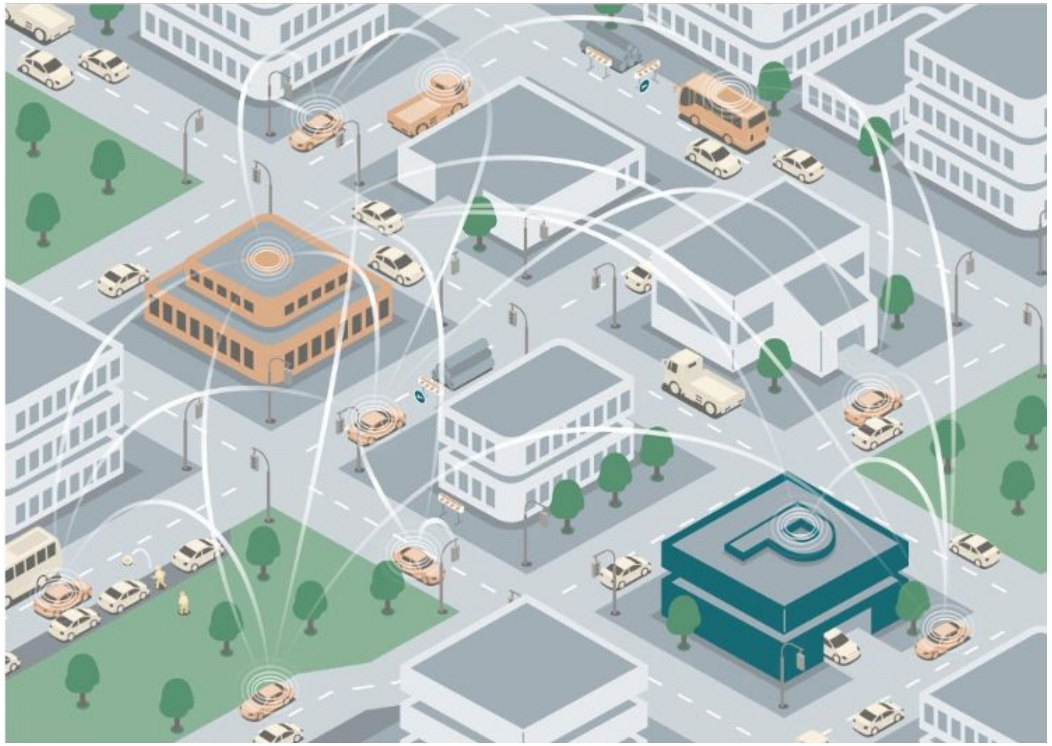
- **Enabling Emerging IoT Applications:**
  - Tightly integrating DSP operators in the execution engine would enable new IoT applications (e.g. gunshot detection from audio signals in a city)

Efficient Spatial Analytics

- **General Description:** Phenomena in the IoT are location-dependent and thus IoT applications should offer analytics on them.

- **State-of-the-Art Systems and Their Limitations:**
  - No, or very limited support for spatial queries
  - Cloud-tailored solutions
  - Inefficient use of resources

- **Enabling Emerging IoT Applications:**
  - Transportation: driverless vehicles and connected cars
  - Public safety: networks of connected cameras or acoustic sensors
  - Health: wearable health trackers

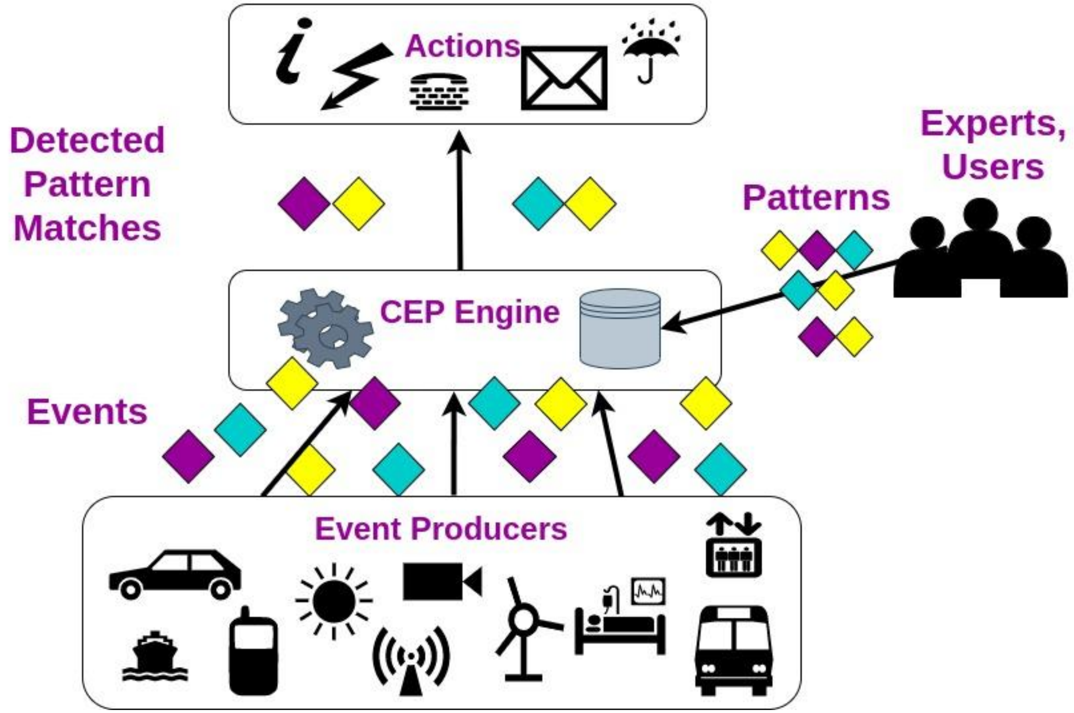


Complex Event Processing (CEP)

- **General Description:** CEP in the IoT would enable users to feed system with knowledge, i.e., patterns, and automate decision making.

- **State-of-the-Art Systems and Their limitations:**
  - Rely on central components and serial processing model which prevents large scale processing.

- **Enabling Emerging IoT Applications:**
  - Scalable CEP would enable future IoT applications such as:
    - Smart hospitals with private fogs can contribute to a public smart city query (COVID-19 cases)
    - Smart street lamps
    - Traffic flow management

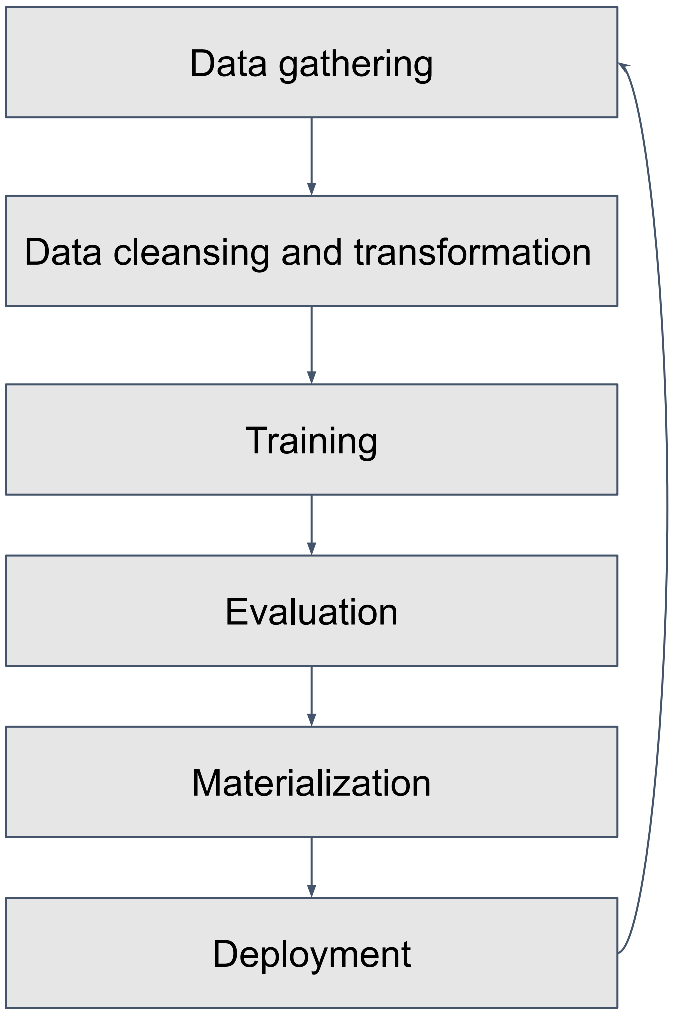


Machine Learning

- **General Description:** Complex machine learning (ML) tasks, e.g., classification, clustering, and prediction are key applications that would profit from being deployed on IoT devices.

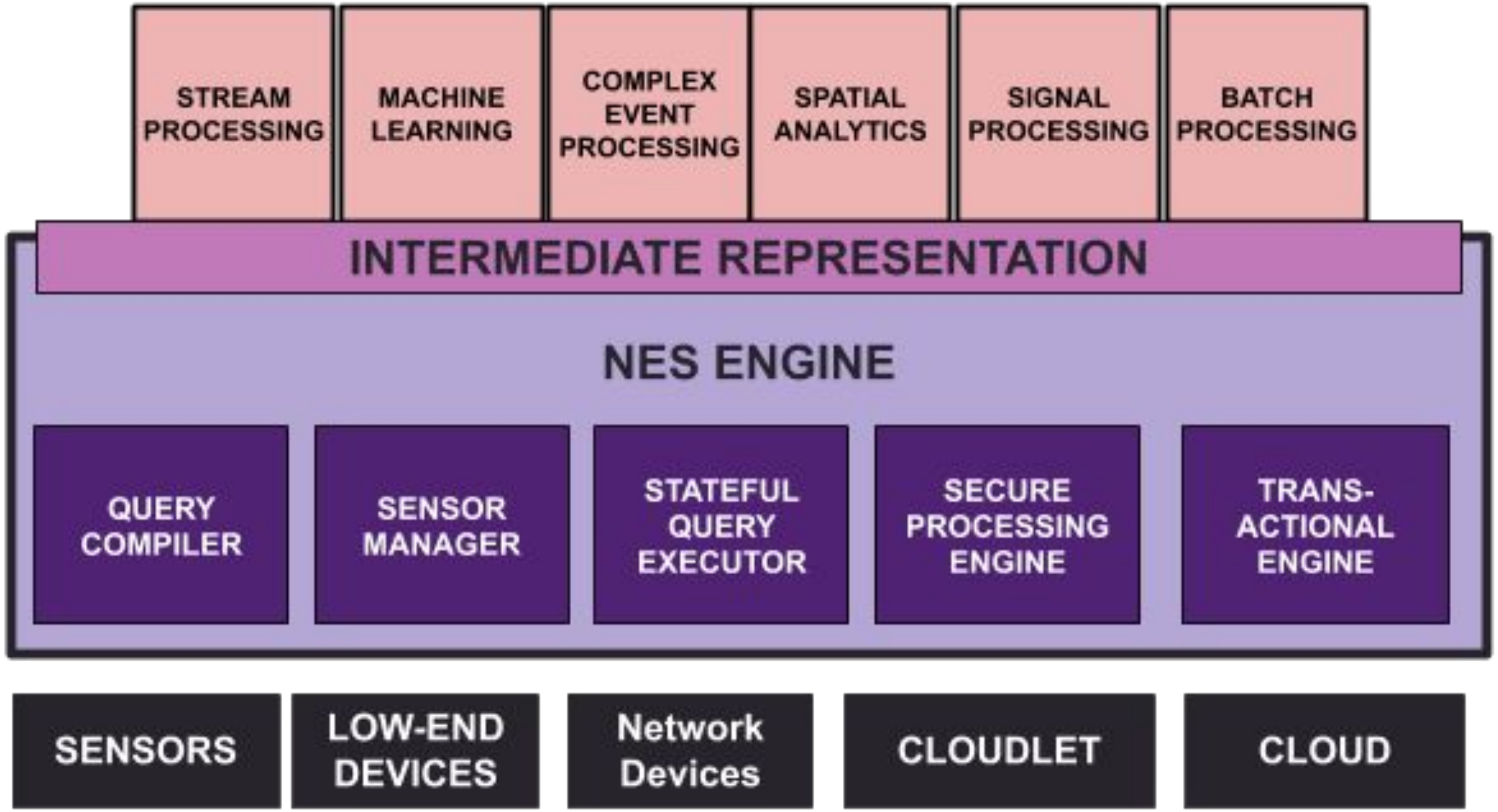
- **State-of-the-Art Systems and Their limitations:**
  - Customized solutions of batch- and stream processing systems (e.g. Flink, Spark)
  - ML frameworks (PyTorch, scikit-learn)
  - Inference in IoT environments are not supported by general purpose ML systems.

- **Enabling Emerging IoT Applications:**
  - Local and distributed models would allow for lower latency
  - Reduce network load and latency during inference



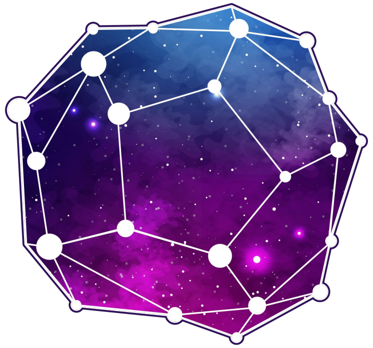
Stages of ML Pipelines

NebulaStream Stack



Summary

Core features enable the next generation of IoT applications but are not yet supported by state-of-the-art systems.



Domain specific features enable a richer set of applications over an IoT data management platform such as NES.

A general-purpose, end-to-end data management system for the IoT.

@NebulaStream www.nebula.stream