

Streaming Data Analysis on the Wire

STREAM 2016

Dimitrios Katramatos
Shinjae Yoo
Meng Yue
Kerstin Kleese van Dam

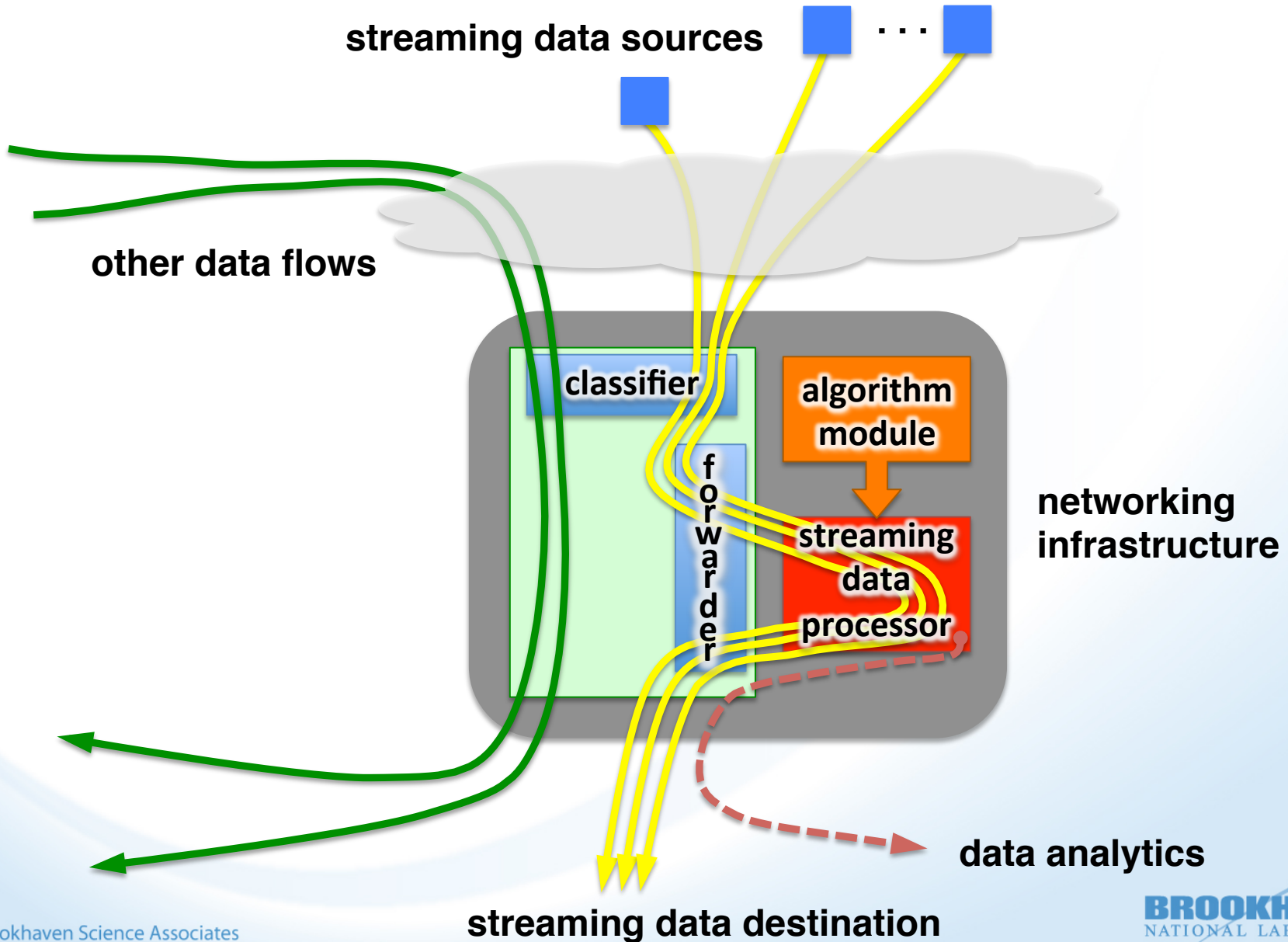
Motivation

- In the Big Data era, lots and lots of data can be found at any moment in transit, potentially more than what is in storage
- Is it feasible to devise a framework for data analytics on the wire, i.e., utilizing capabilities of the network infrastructure?
- Early processing provides real-time/near real-time information that can be used to speed up the decision processes

Concept

- Network infrastructure includes mechanisms that can be programmed to recognize specific data flows based on given criteria
- Flows are then intercepted and transparently forwarded to processing subsystem(s) where data is subjected to desired processing and information is extracted
- Processed data (original or transformed) is ultimately forwarded back to its original destination

Concept cont.



Potential Use Cases

- Scientific
 - Simple streaming data transformations
 - Statistical information extraction
 - Pre-processing during/after data acquisition
 - Sensor network data analysis (e.g., distributed solar irradiance prediction, Smart Grid Phasor Measurement Unit and Smart Meter data, DARPA SIGMA security sensor networks)
 - Internet of Things (IoT)

Potential Use Cases cont.

- Commercial
 - Real time processing before arrival at company's data center for decision making
 - Network provider offers valued-added processing services to subscribers lacking such capabilities
 - IoT

Current Solutions

- Mostly business oriented:
 - Cybersecurity
 - Firewalls
 - Deep Packet Inspection / Processing (DPI/DPP)
 - Intrusion Detection
 - Customer data analytics
- Special hardware and software
 - Vendor-specific
- Software Defined Approaches
 - Also mainly business-oriented
 - Interesting concepts
 - Network Function Virtualization (NFV)
 - Service Function Chaining (SFC)
 - SDN controller software maturity level?

Challenges

- Special-purpose vs. general-purpose
 - User-defined processing
 - Hardware/software is expected to need modifications
- Performance – how much penalty?
 - Additional processing adds overhead
 - Hardware limitations affect both special hardware and SDN solutions
- Algorithms – what can be run?
 - Streaming algorithms with low overhead
 - Loading/distribution – static/dynamic programmable
 - Coordination – single vs. group of network devices

Research Directions

Two directions with equal weight

- Networking

- Vendor hardware

- Deep Packet Inspection (DPI) and Deep Packet Processing (DPP)
 - Big IP F5 systems
 - Tap and redirect
 - Can we influence future vendor designs?

- Prime case for SDN

- Network Function Virtualization (NFV) and Service Function Chaining (SFC) – modify/augment SDN controllers?
 - Augment virtual switch capabilities?
 - Pica8 white box switches

Research Directions cont.

- Algorithms
 - Extract streaming data analytics, and/or
 - Transform streaming data
 - Low overhead to match capabilities
 - Examples
 - Outlier detection
 - Approximated summary statistics
 - Lightweight dimensionality reduction using problem characteristics
 - Batch supervised/unsupervised learning
 - Adaptive supervised/unsupervised learning

Questions?