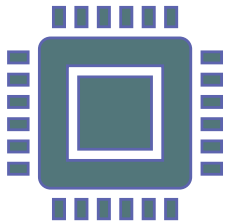


On Timeseries Downsampling in Stream Processing Engines

Iain Dixon (i.g.dixon@ncl.ac.uk)

Dr. Matthew Forshaw (mathew.forshaw@ncl.ac.uk)

Dr. Joe Matthews (joe.matthews@ncl.ac.uk)

UK Systems 2023

Ramside Hall, Carrville, Durham

27th April 2023

Challenges of large-scale timeseries processing

Problem: Sensors produce more data than we can store or transmit across a network

Solution: Limit sampling rate or take subset of sensors

Issue: Lose data arbitrarily

Problem: Offered load exceeds our SPE

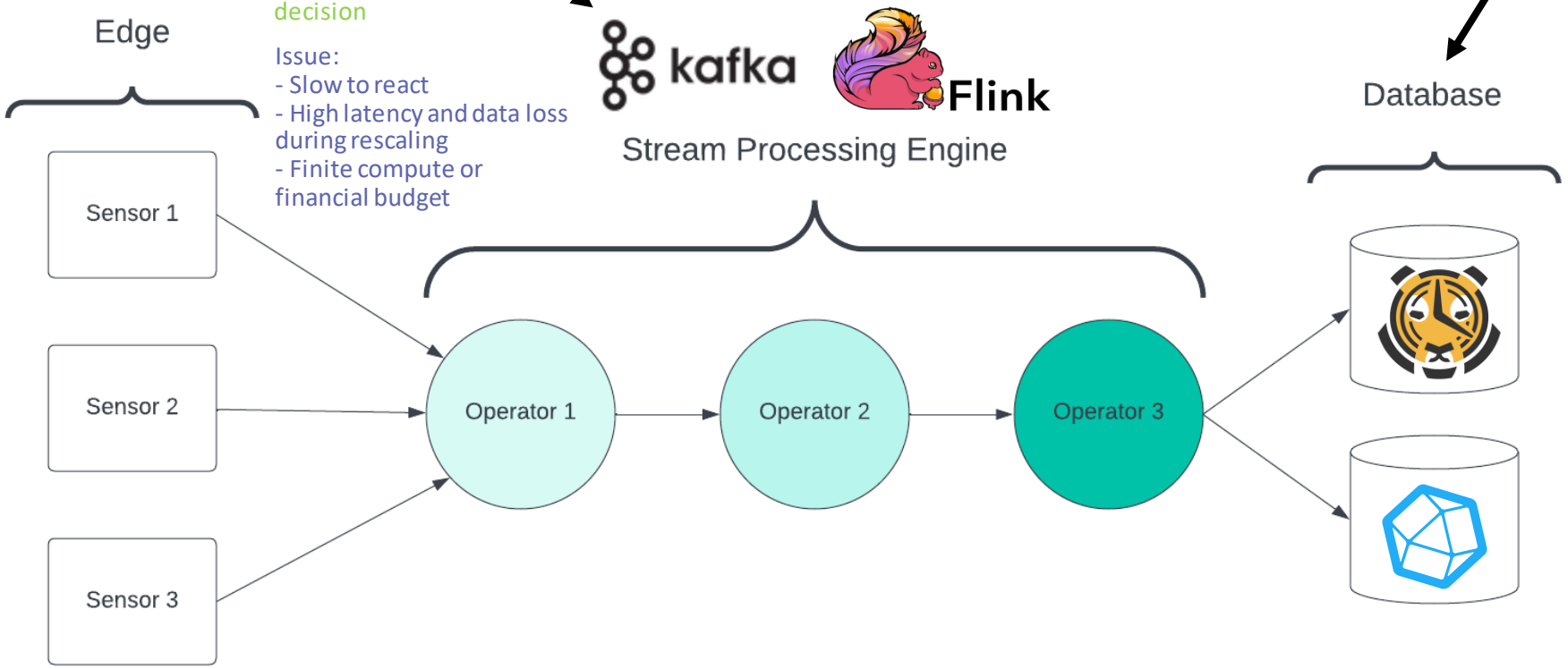
Solution: (Auto)scaling decision

Issue:
- Slow to react
- High latency and data loss during rescaling
- Finite compute or financial budget

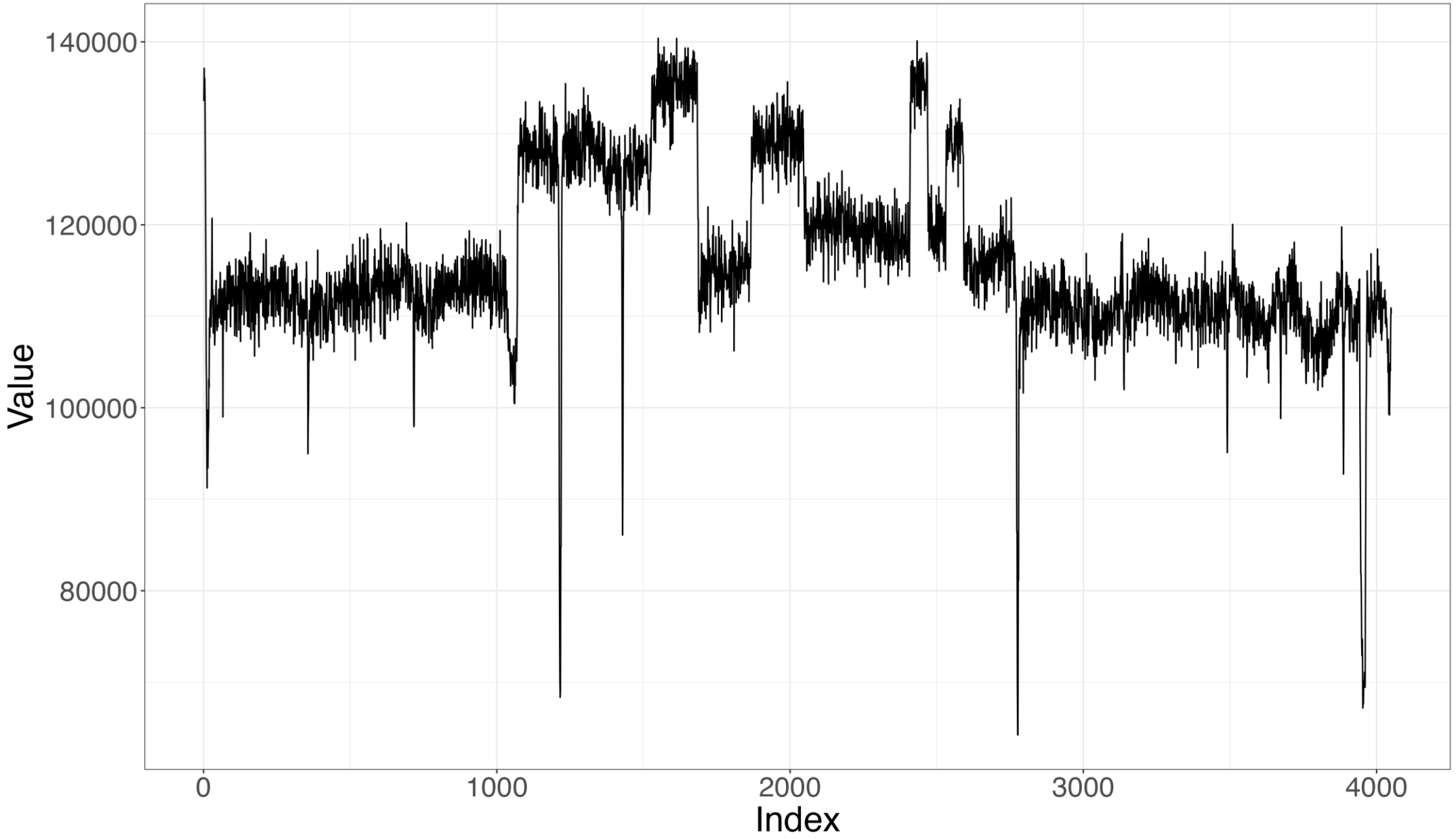
Problem: Storage capacity

Solution: Roll up aggregation

Issue: Lossy, potentially losing interesting data



Question: are all timeseries data points equally valuable?

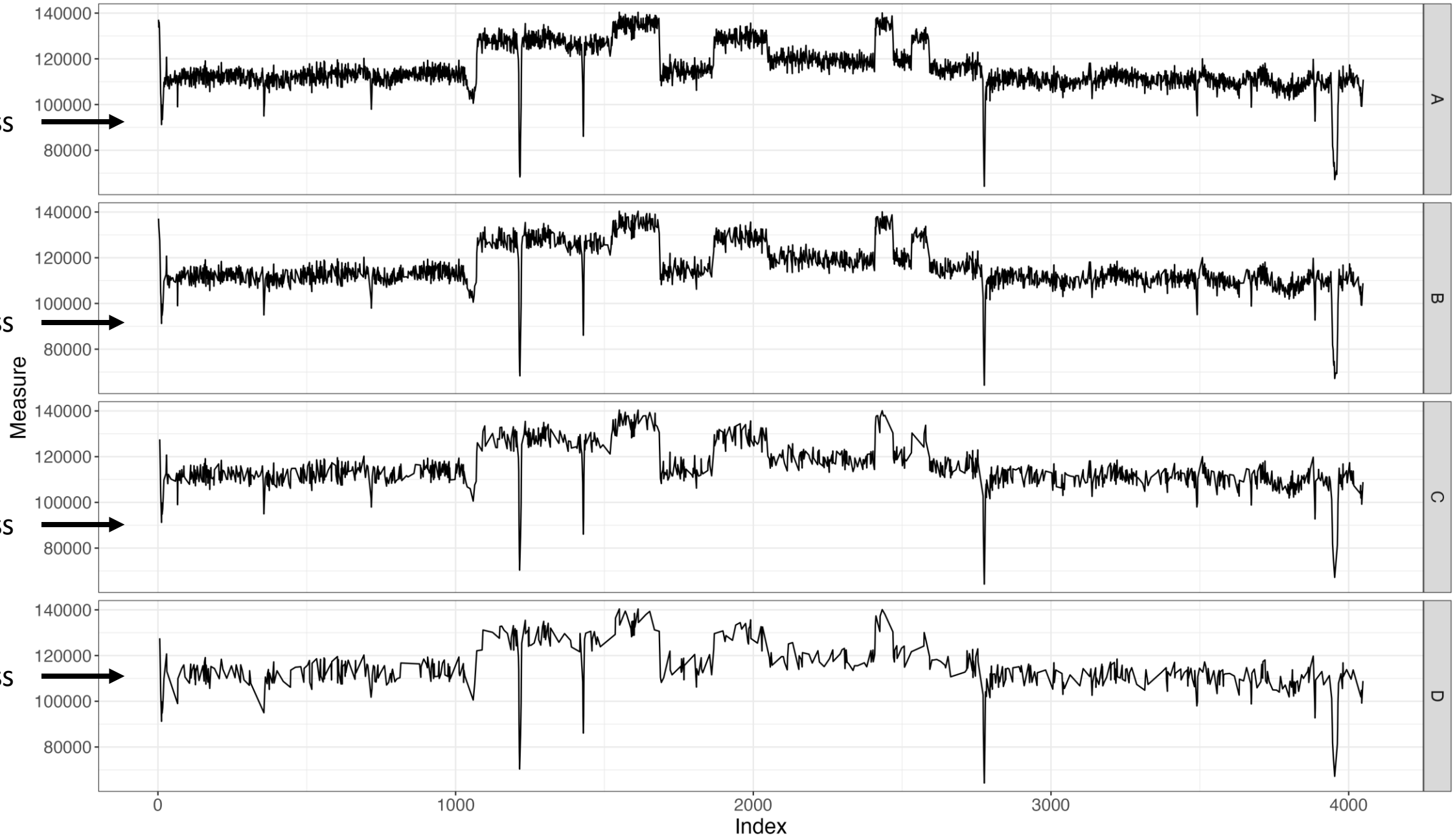

140000
120000
100000
80000
Value
0
1000
2000
3000
4000
Index

30% Dataloss

50% Dataloss

70% Dataloss

85% Dataloss

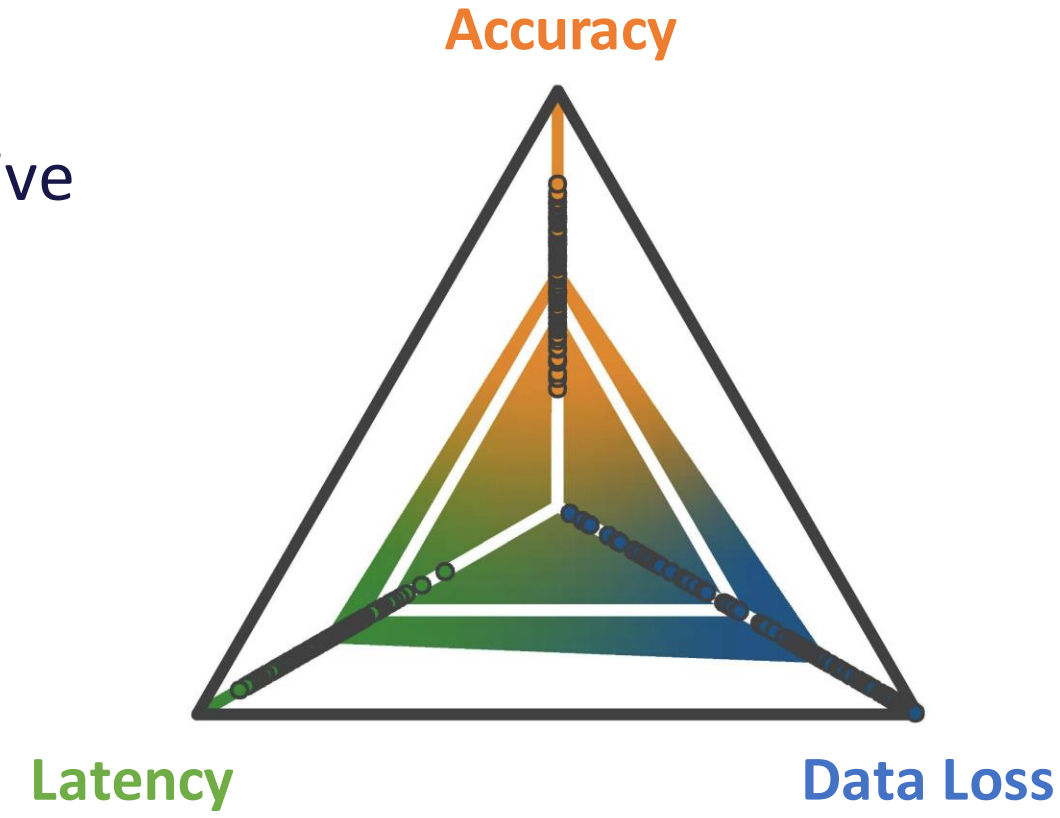


Existing Downsampling Approaches

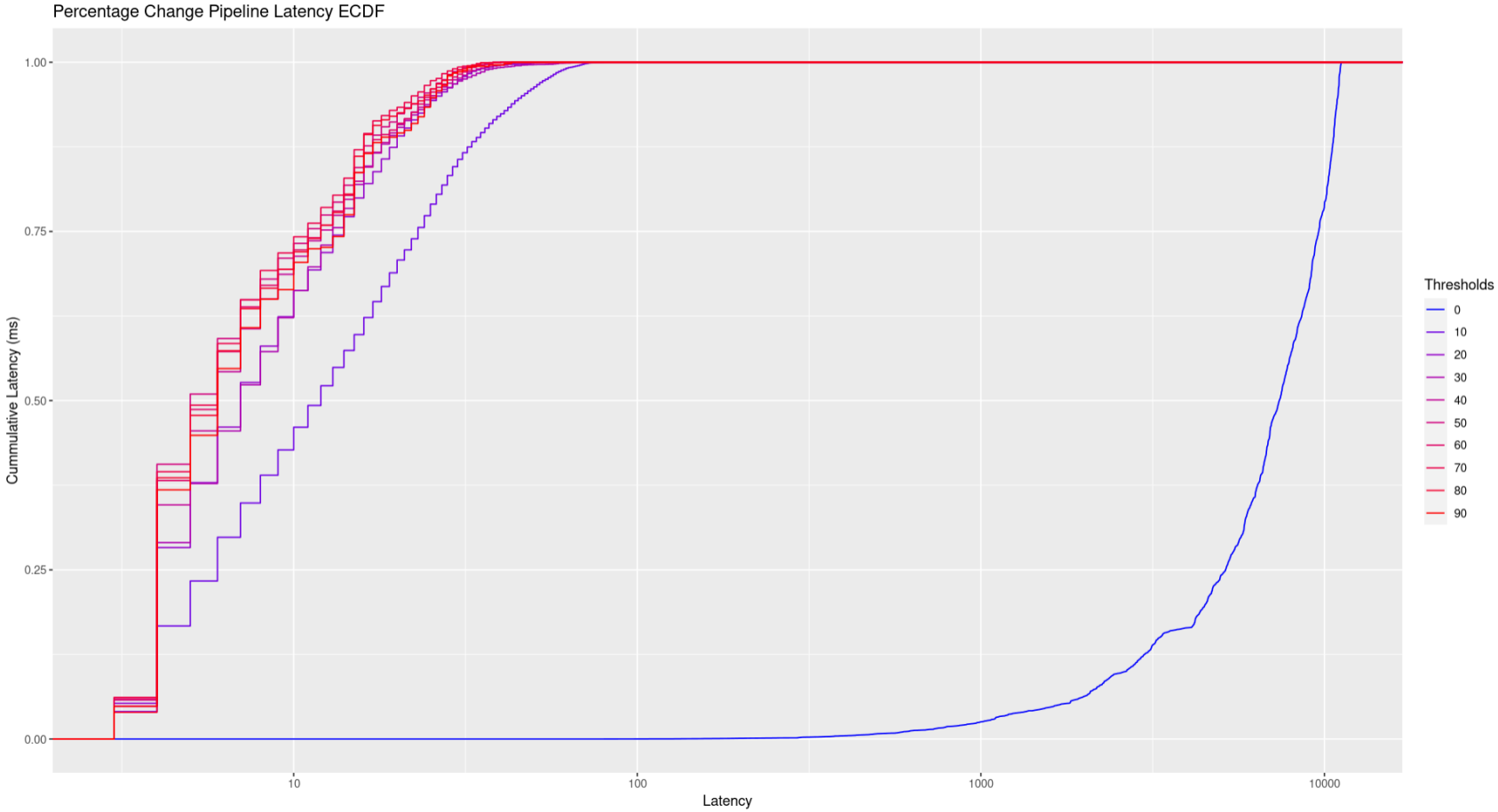
Approach	Data-Agnostic	Value Preserving Aggregation	Time-based Aggregation
Examples	Simple Decimation/EveryN, Reducing sensor sampling rate (Hz)	LTTB, M4, Percentage Change	Summary statistics (e.g. median) over Sliding/Tumbling windows
Characteristics	Random data loss	Selection of existing datapoints from the TS	Aggregates data in chunks
Pros	Cheap, Fast	Trust with users	Reduces data cardinality and
Cons	Poor Accuracy	Need for global knowledge	High latency across window

The Streaming Downsampling Trilemma

- Tradeoff between **accuracy** and **data loss**.
- Longer windows support more effective downsampling **(accuracy)**, at the expense of per-object **latency**.
- Application of downsampling in compute-, memory- or storage-constrained contexts.



Emerging results: downsampling highly effective in reducing per-object latency.



Next Steps

- Opportunities of Streaming Downsampling
  - **Decoupling Offered Load and Scalability**: Allows system operators to make a principled decision between deployment size (and cost) and accuracy.
  - **Smooth growing pains**: Provides an opportunity to handle resource contention while rescaling is being enacted.
  - **Adaptive approaches:** based on characteristics of the data. (Note: e.g. variable percentage change)
  - **Novel downsampling approaches:** leveraging streaming semantics.
- We welcome your feedback and recommendations.
  - **Application Areas:** At the start of the presentation we highlight opportuntiies for downsampling a) at the edge, b) in the SPE, and c) in storage. Are you aware of other areas where this could be applied?
  - **Case Studies and Datasets:** We are keen to test our approach against realistic challenges – we would like to make friends.

i.g.dixon1@newcastle.ac.uk