Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

Building Adaptive Data Mining Models on Streaming Data in Real-Time, an Outlook on Challenges, Approaches and Ongoing Research

Marine Perception Research Department of the German Research Center for Artificial Intelligence (DFKI)

Frederic Stahl

www.dfki.de/map

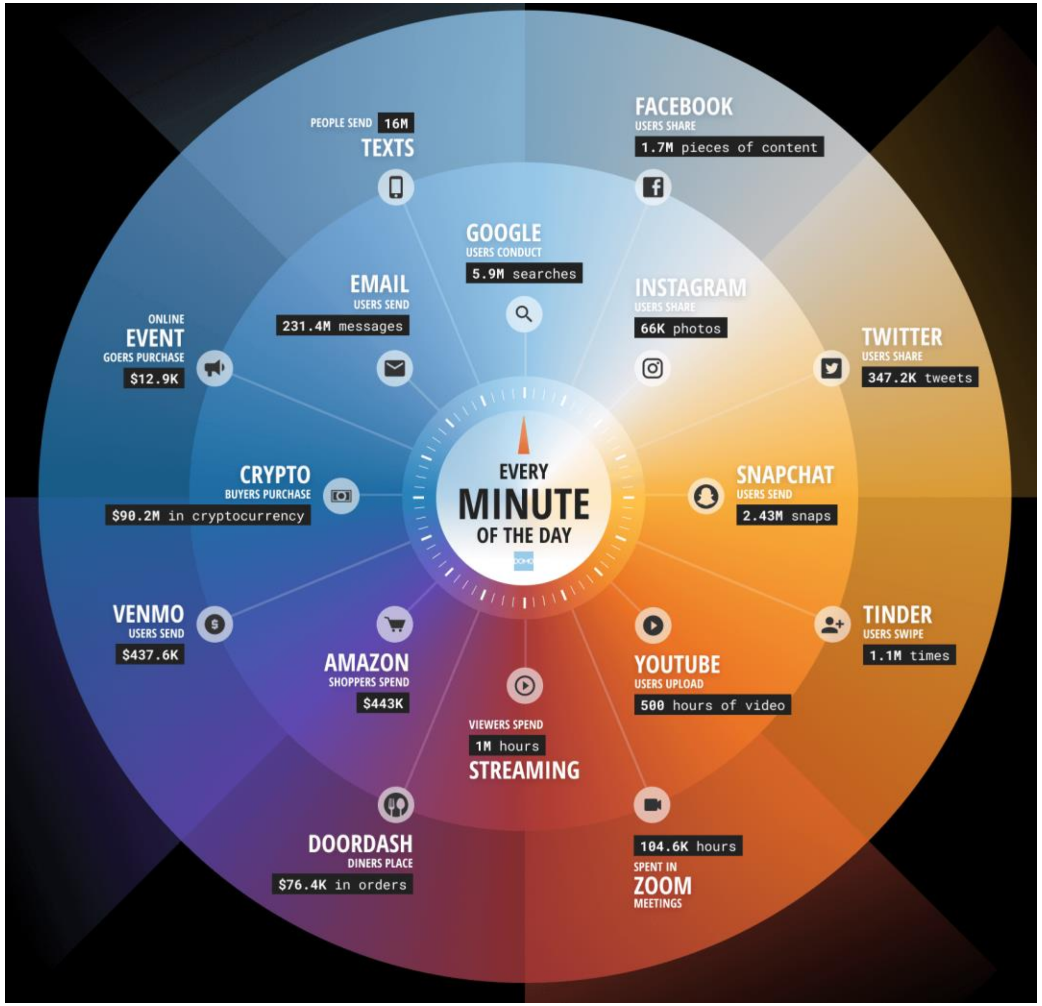
Marie-Curie-Str. 1
26129 Oldenburg
Germany

map-info@dfki.de

Data never sleeps!

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- Forbes: 2.5 quintillion bytes of data created every day.
- That’s about 100 million Blue-ray discs or about 530 million DVD discs.



Sources:
https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/
https://www.theregister.co.uk/2008/01/23/us_hd_player_sales/
https://www.domo.com/data-never-sleeps

How much Data Is created Every day?

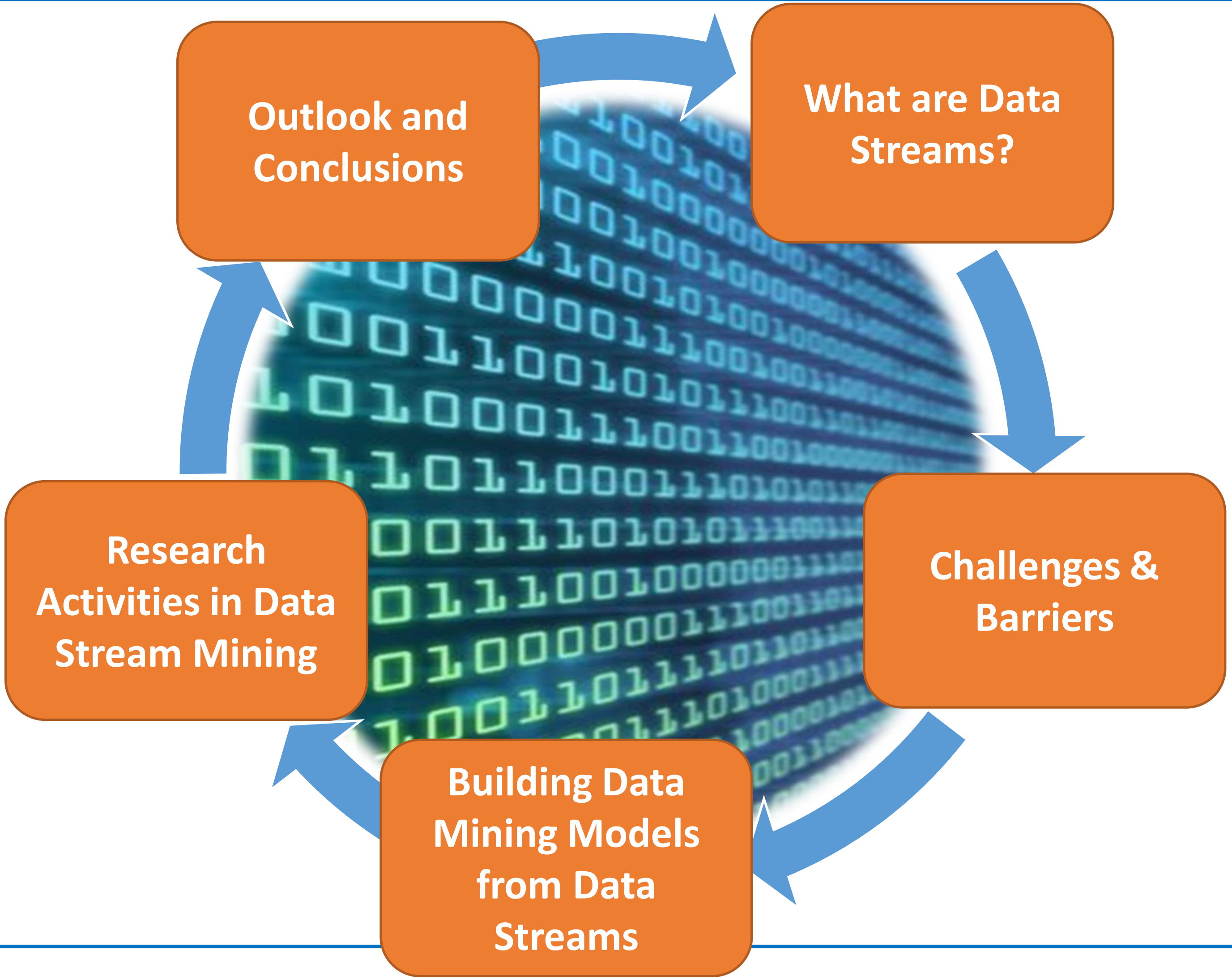
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

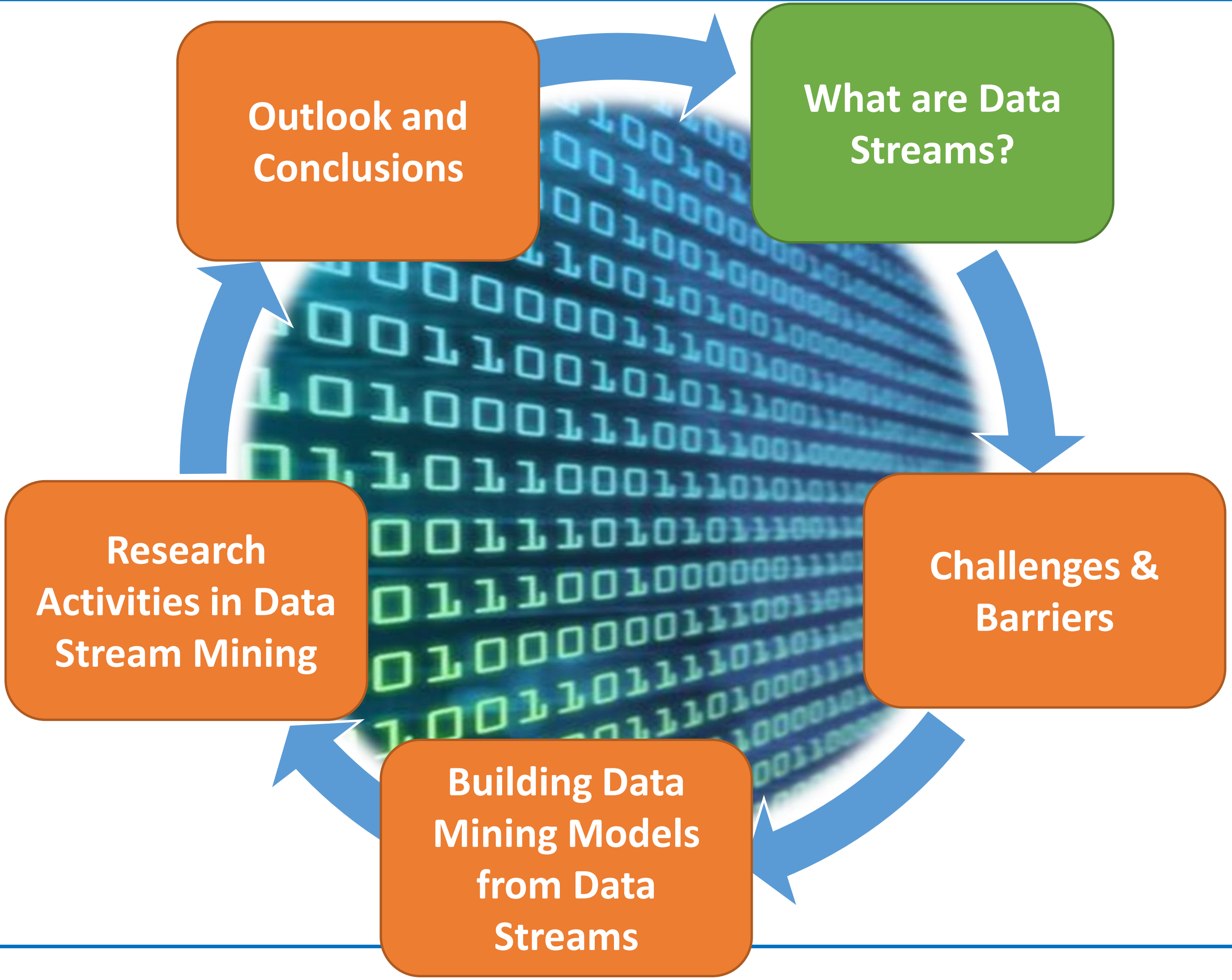
- That’s about 100 million Blue-ray each 25 GB discs.
- Each disc is 1.2mm thick
  - ⇒ This stacks to **120 km!**
  - ⇒ Distance Oldenburg to Hamburg!

- Or in DVDs (4.7 GB each disc)
- Each disc is 1.2mm thick
  - ⇒ This stacks to **630 km!**
  - ⇒ Distance Oldenburg to London/Reading!

To make sense of this real-time data, analytics methods that never sleep are required!

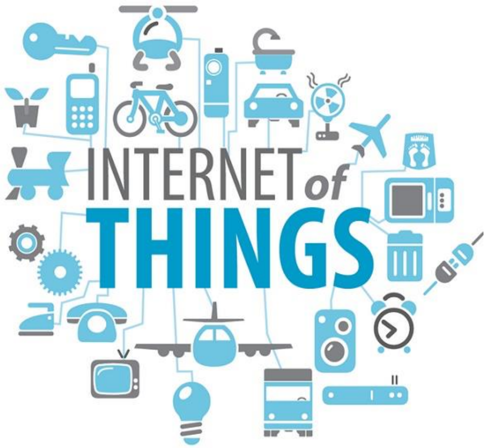
Outline


What are Data Streams?
Challenges & Barriers
Building Data Mining Models from Data Streams
Research Activities in Data Stream Mining
Outlook and Conclusions


What are Data Streams?
Challenges & Barriers
Building Data Mining Models from Data Streams
Research Activities in Data Stream Mining
Outlook and Conclusions

Sources of Data Streams

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

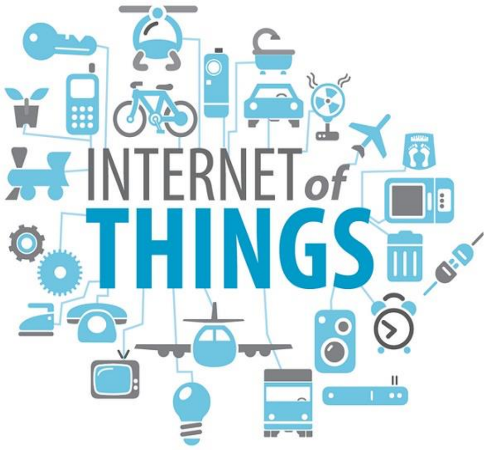


Internet of Things

- By year-end 2039, IoT devices worldwide are forecasted to almost triple from 9.7 billion in 2020 to 29 billion in 2030 [1]

[1] statistica. (2020). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030

Sources of Data Streams



Internet of Things

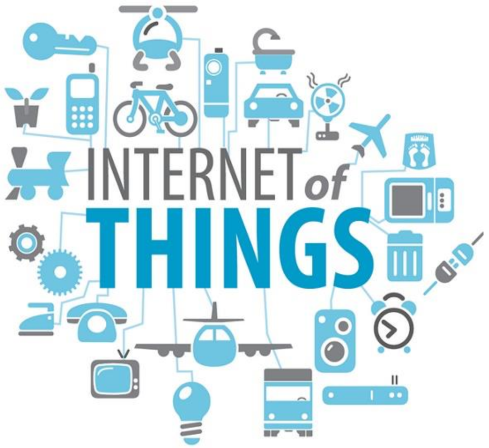
- By year-end 2039, IoT devices worldwide are forecasted to almost triple from 9.7 billion in 2020 to 29 billion in 2030 [1]

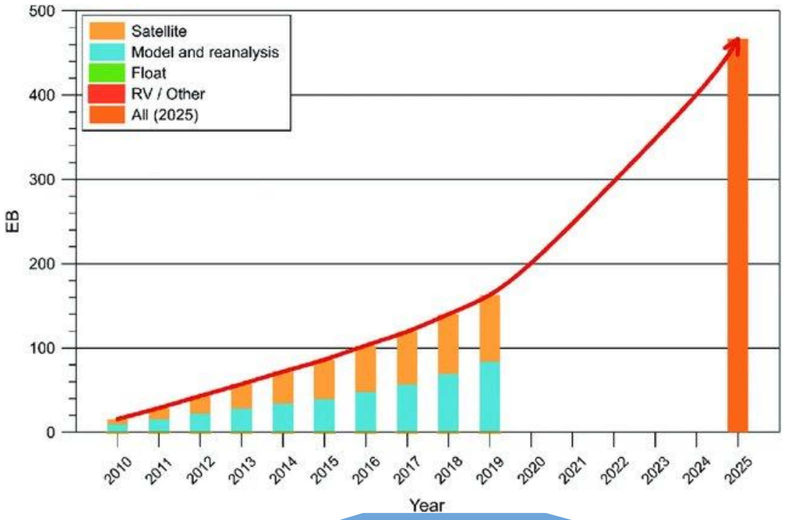
Personalisation

- Facebook:
  - 1.91 billion active users every day [2]
  - 4.75 billion pieces of content shared

[1] statistica. (2020). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030
[2] Noyes, A. and Noyes, D. (2014). The Top 20 Valuable Facebook Statistics - Updated October 2014 – Zephoria Inc.. [online] Zephoria Inc. Available at: https://zephoria.com/social-media/top-15-valuable-facebook-statistics/ [Accessed 2022].

Sources of Data Streams





Internet of Things

- By year-end 2039, IoT devices worldwide are forecasted to almost triple from 9.7 billion in 2020 to 29 billion in 2030 [1]

Personalisation

- Facebook:
  - 1.91 billion active users every day [2]
  - 4.75 billion pieces of content shared

Marine Sciences

- Distribution of ocean science data acquired in the past decade, based on publicly available data from the internet (CC BY 4.0) [3]
- Expected to reach almost 500 Exabytes by the year 2025

[1] statistica. (2020). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030
[2] Noyes, A. and Noyes, D. (2014). The Top 20 Valuable Facebook Statistics - Updated October 2014 – Zephoria Inc.. [online] Zephoria Inc. Available at: https://zephoria.com/social-media/top-15-valuable-facebook-statistics/ [Accessed 2022].
[3] Qian, C., Huang, B., Yang, X. and Chen, G., 2022. Data science for oceanography: From small data to big data. Big Earth Data, 6(2), pp.236-250.

Static versus Streaming Data

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

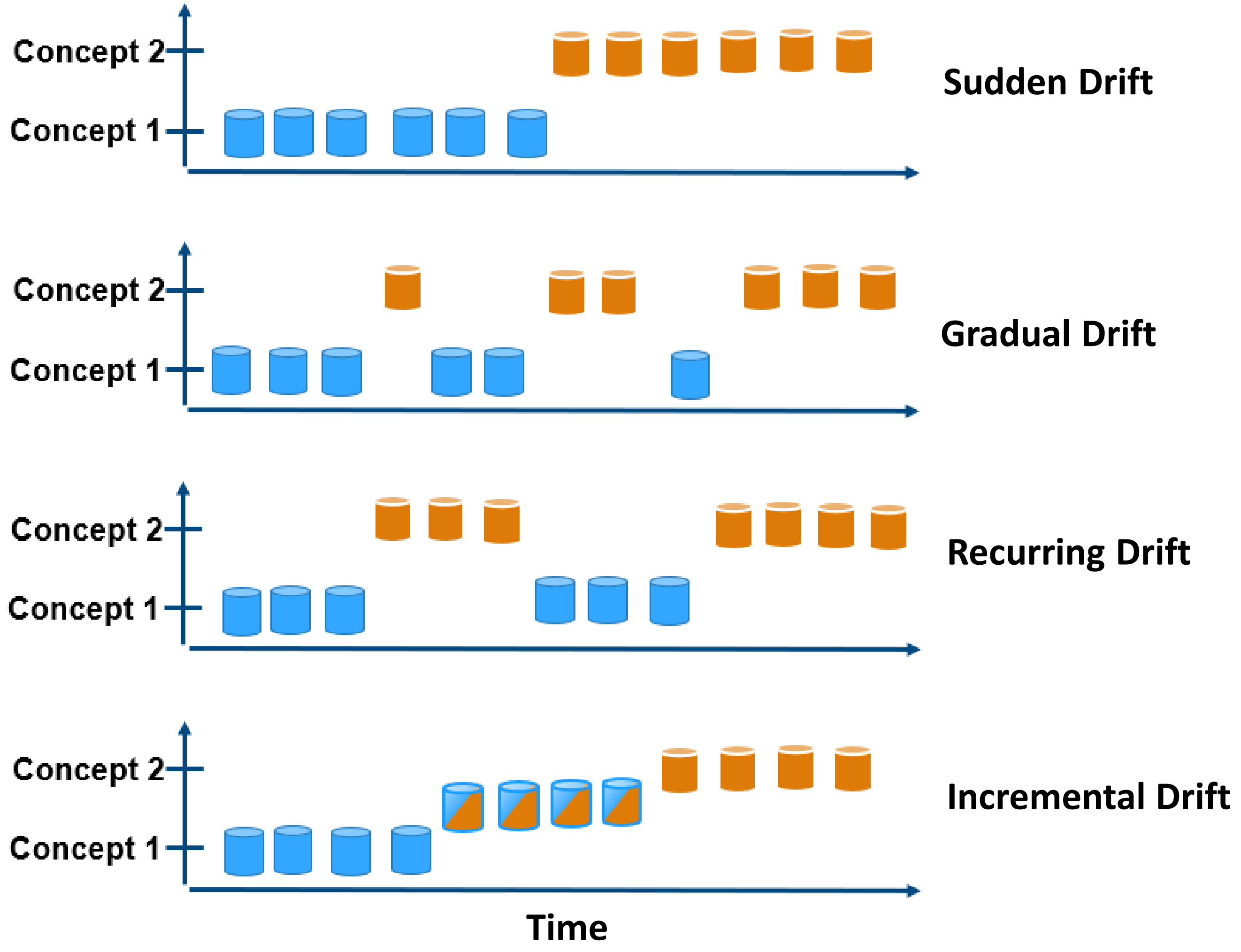
A data stream is *a continuous, rapid flow of data that challenges our state-of-the-art processing and communication infrastructure.*

Static Data	Streaming Data
• Historical data	• Often live, real-time data feed
• Randomly accessible	• Sequentially accessed
• Secondary storage	• Limited memory requirements
• No/low processing latency criticality	• High processing latency criticality
• Assumption of pre-processed dataset	• Assumption of inaccurate raw data

Volume and **Velocity**

Concept Drift

- Underlying concept defining the knowledge being learned, begins to shift over time.
- Concept change is unforeseen and unpredictable.
- Concepts from the past may re-occur in the future.
- Concept drift exists in real-life problems:
  - Seasonal weather
  - Stock market rallies because of breaking news
  - etc.

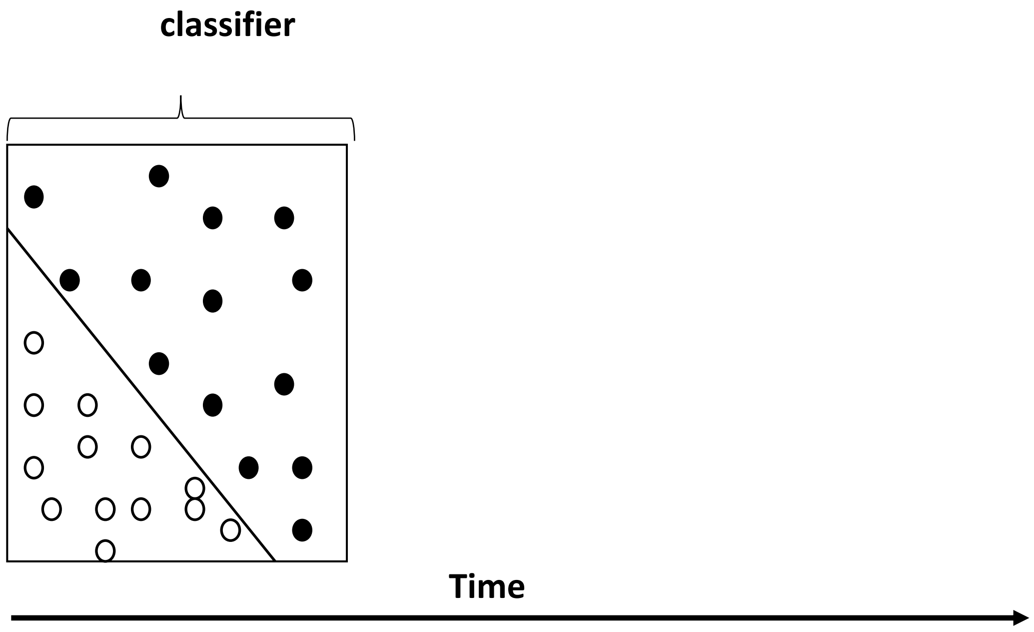


Concept Drift (cont.)

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Concept shift/drift: changes mining set statistics

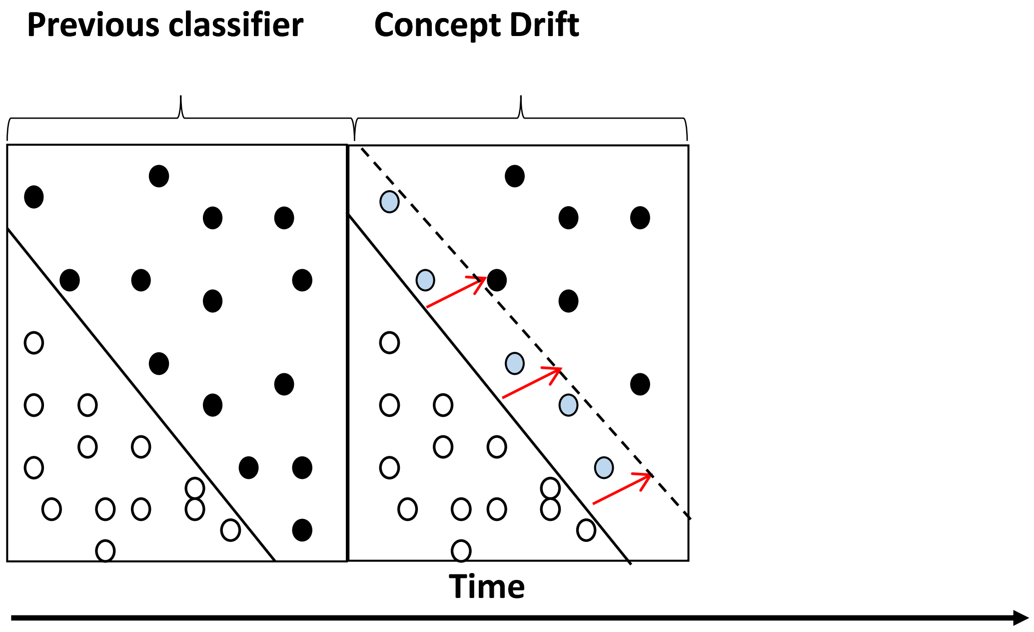
- A model should always reflect the time-changing concept.
- Render previously learned models inaccurate or invalid.
- Robustness and adaptability: quickly recover/adjust after concept changes.



Concept Drift (cont.)

Concept shift/drift: changes mining set statistics

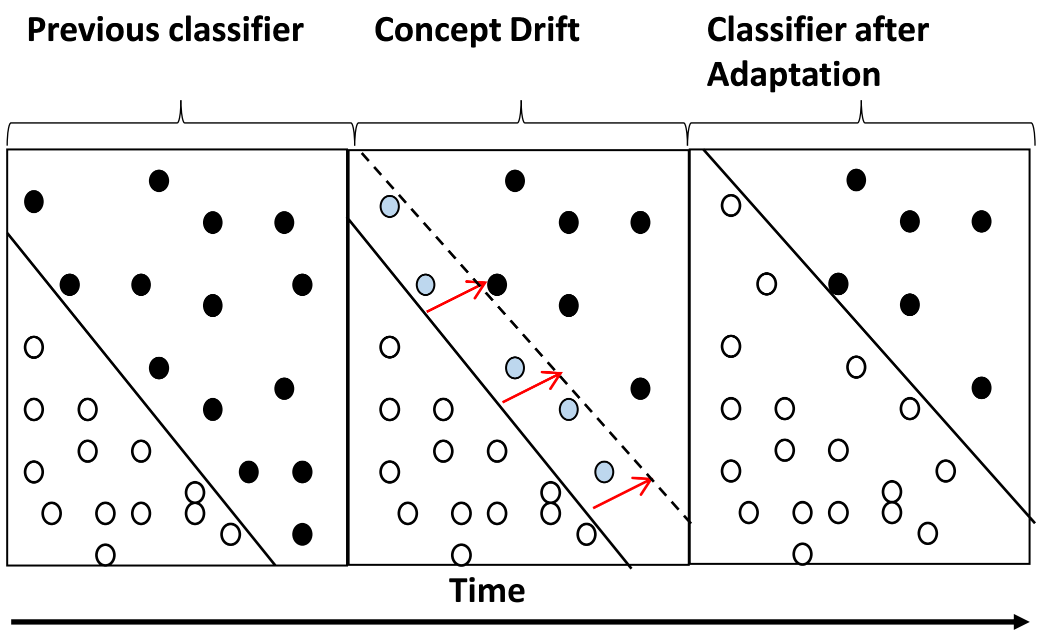
- A model should always reflect the time-changing concept.
- Render previously learned models inaccurate or invalid.
- Robustness and adaptability: quickly recover/adjust after concept changes.

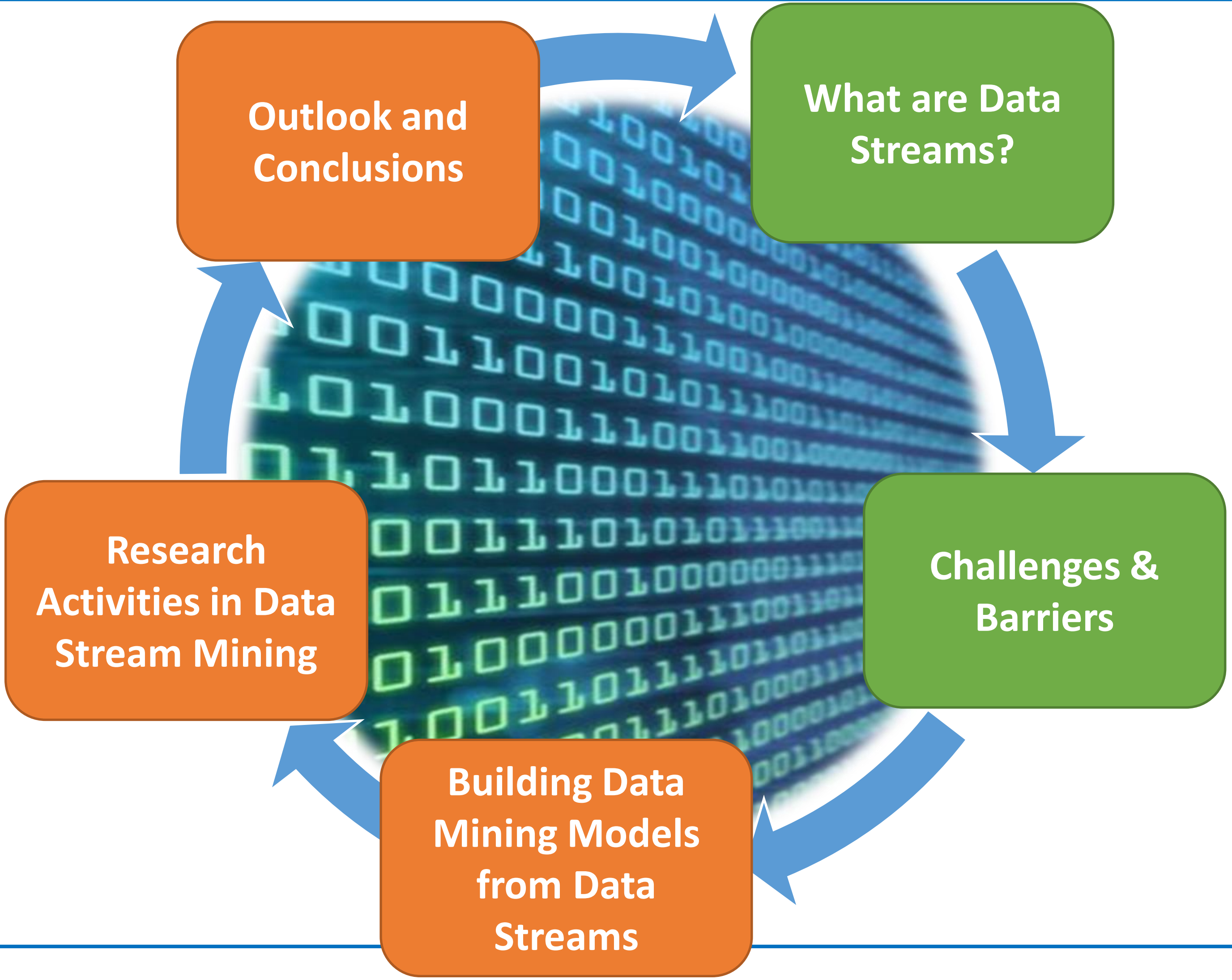


Concept Drift (cont.)

Concept shift/drift: changes mining set statistics

- A model should always reflect the time-changing concept.
- Render previously learned models inaccurate or invalid.
- Robustness and adaptability: quickly recover/adjust after concept changes.




What are Data Streams?
Challenges & Barriers
Building Data Mining Models from Data Streams
Research Activities in Data Stream Mining
Outlook and Conclusions

The Data Tsunami

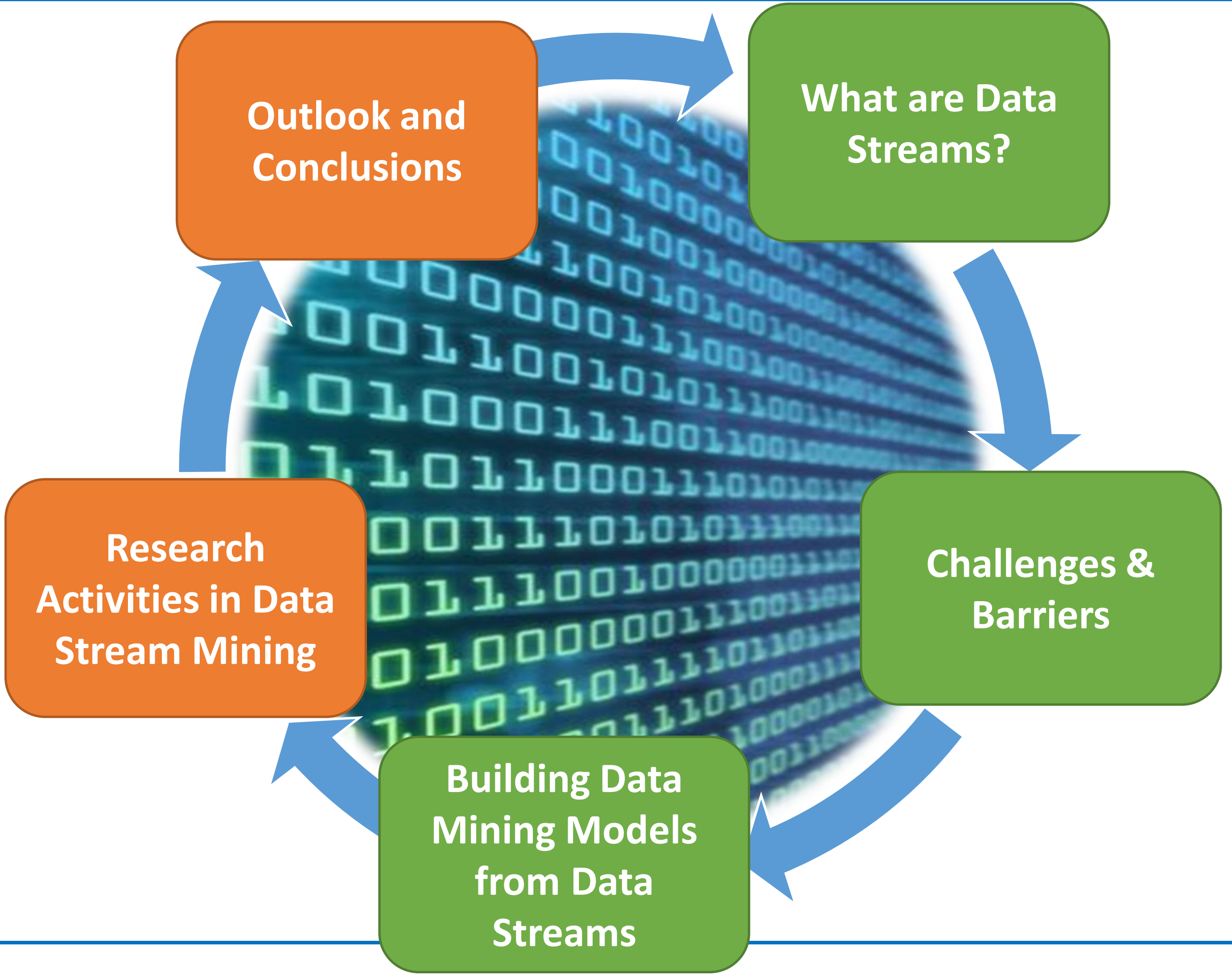
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Challenges

1) Data generated at a fast rate (Velocity), at potentially large and unknown quantities (Volume)
2) Concept Drift (changes of pattern encoded in in the data over time)
3) Modelling real-time analytics workflows from streaming data
4) Multi-modality of data sources (text, video/images, unstructured)
5) Class label sparsity: adapting predictive models
6) Explaining Concept Drift

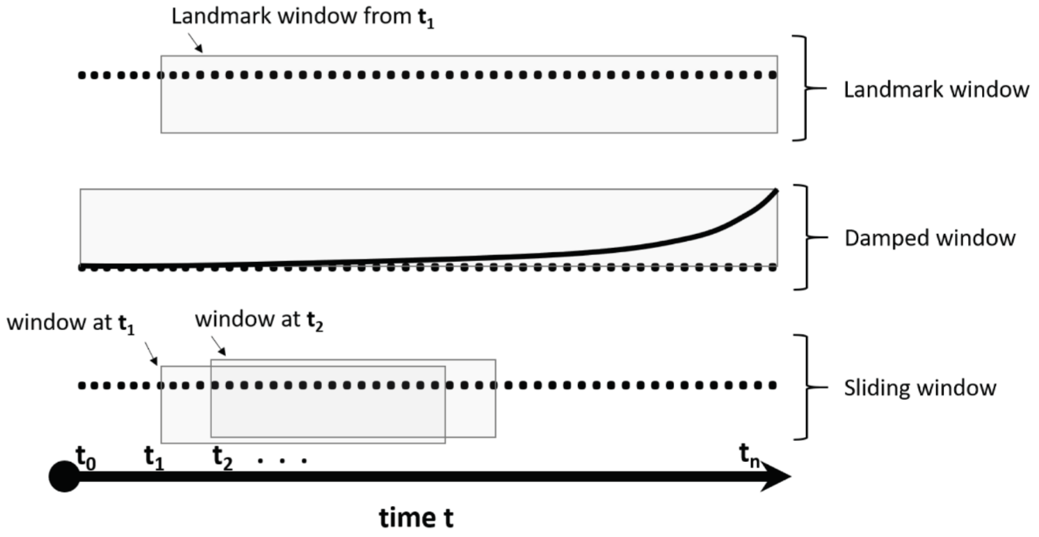
Barriers

1) Limited scalable (parallel) real-time high throughput data stream mining algorithms
2) Different and changing types of concept drift
3) Lack of customisable pre-processing techniques
4) Different time stamps but co-occurring data items
5) Supervised algorithms not applicable in many cases
6) Lack of drift detectors explaining concept drift


What are Data Streams?
Challenges & Barriers
Building Data Mining Models from Data Streams
Research Activities in Data Stream Mining
Outlook and Conclusions

Methods: Windowing approaches to induce data mining models

1) Create time windows

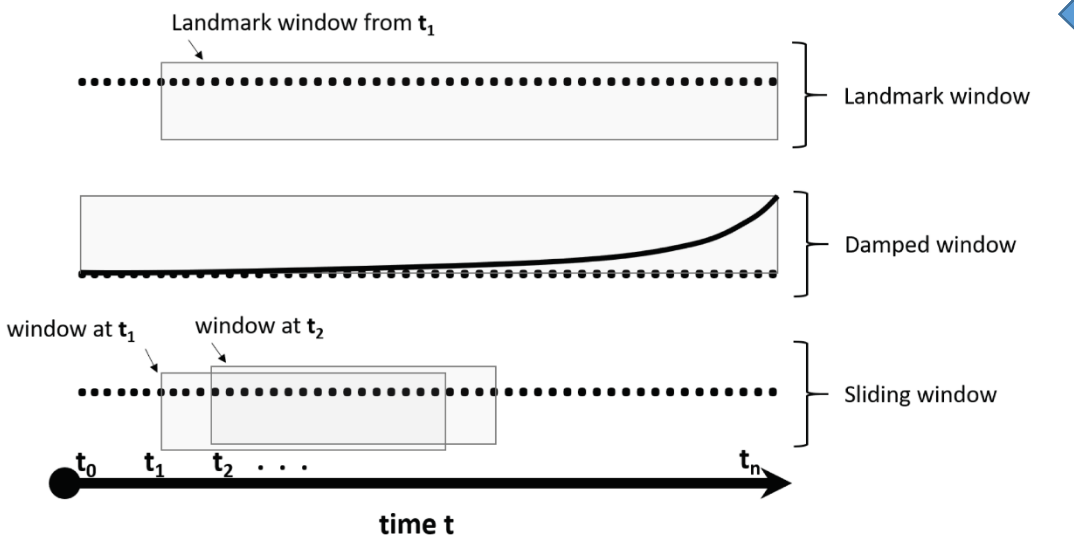


Source: Stahl, F., Le, T., Badii, A., Gaber, M.M. (2021) A frequent pattern conjunction Heuristic for rule generation in data streams. Information 12(1) (2021), ISSN 2078-2489, doi: 10.3390/info12010024

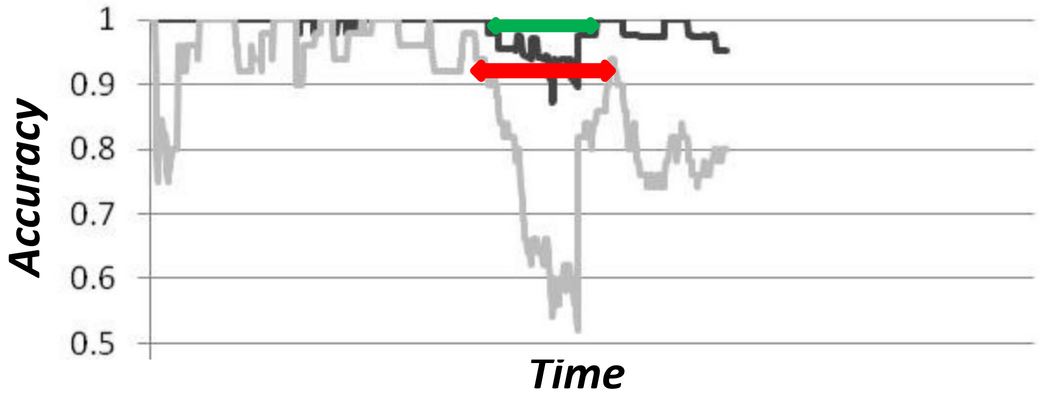
Methods: Windowing approaches to induce data mining models

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

1) Create time windows



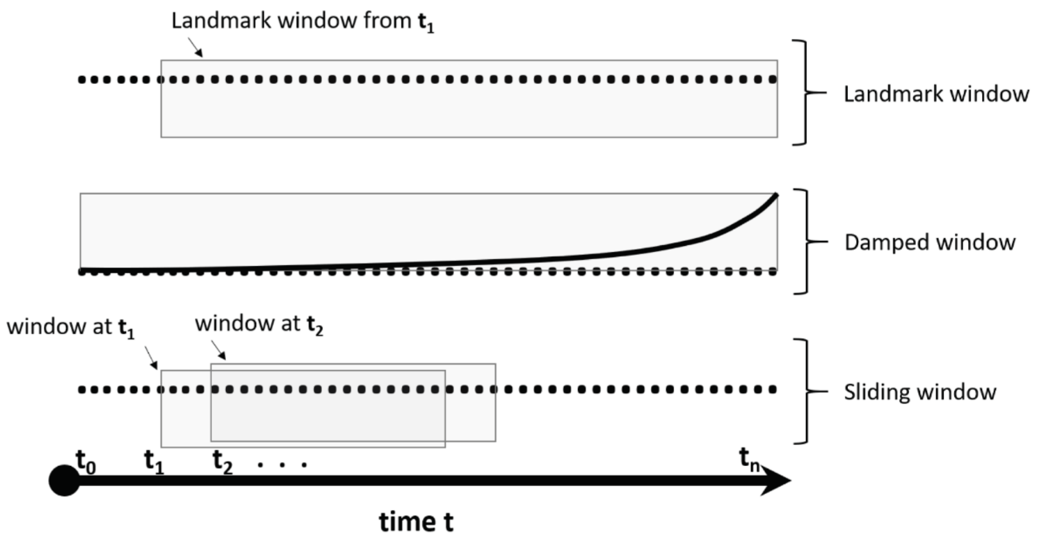
2) Detect concept drift



Source: Stahl, F., Le, T., Badii, A., Gaber, M.M. (2021) A frequent pattern conjunction Heuristic for rule generation in data streams. Information 12(1) (2021), ISSN 2078-2489, doi: 10.3390/info12010024

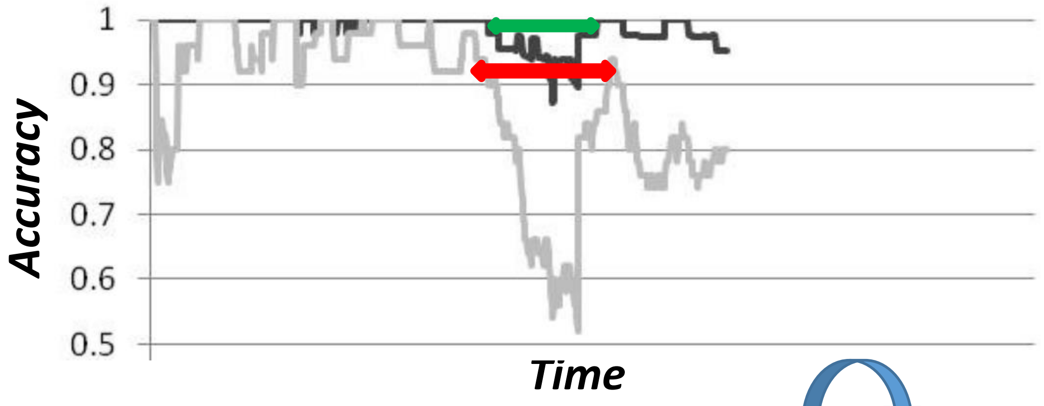
Methods: Windowing approaches to induce data mining models

1) Create time windows

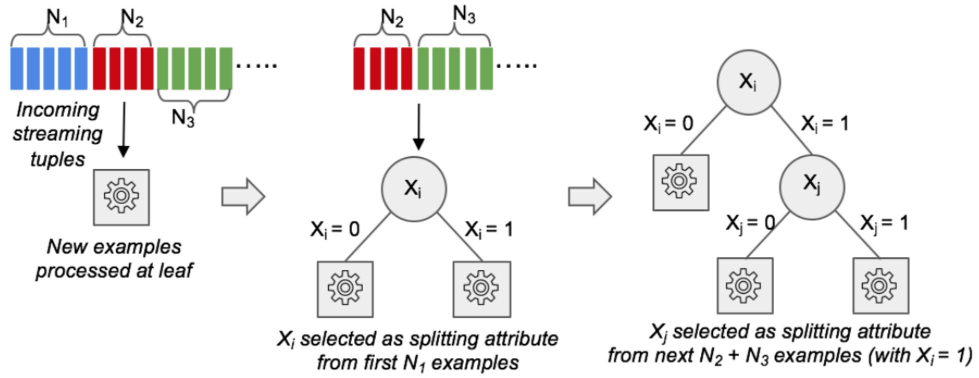


Source: Stahl, F., Le, T., Badii, A., Gaber, M.M. (2021) A frequent pattern conjunction Heuristic for rule generation in data streams. Information 12(1) (2021), ISSN 2078-2489, doi: 10.3390/info12010024

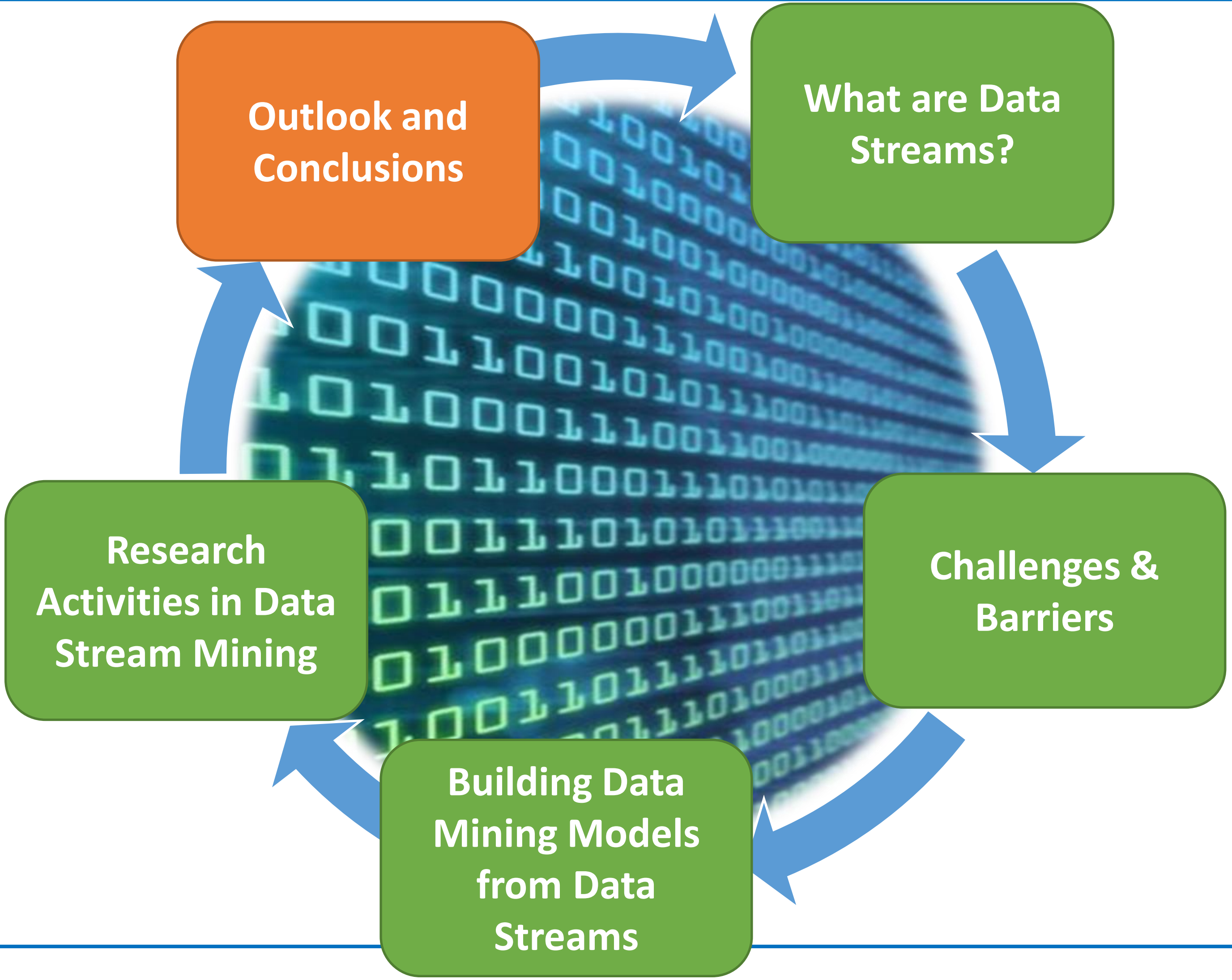
2) Detect concept drift



3) Adapt data mining model



Source: Domingos and Hulten, 2000] Pedro M. Domingos and Geoff Hulten. Mining high-speed data streams. In SIGKDD, pages 71–80, 2000


Outlook and Conclusions
What are Data Streams?
Challenges & Barriers
Building Data Mining Models from Data Streams
Research Activities in Data Stream Mining

MC-NN: Micro Cluster based Neareast Neighbour, basic Approach

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

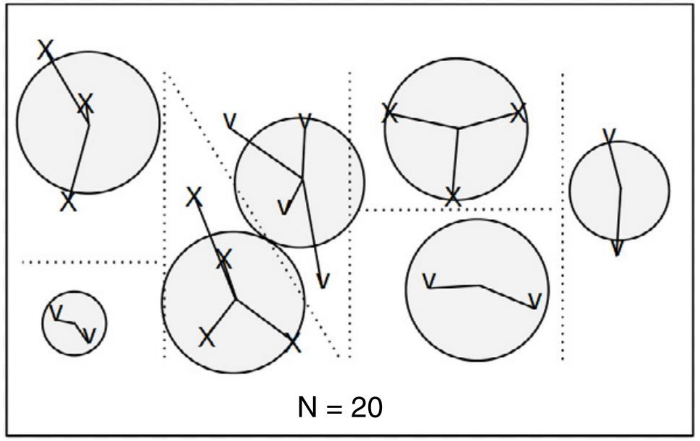
Objective: Develop a scalable predictive Data Stream classification

EPSRC
Engineering and Physical Sciences Research Council

MC-NN: Micro Cluster based Neareast Neighbour, basic Approach

Objective: Develop a scalable predictive Data Stream classification

1) Initialising Micro-Clusters and maintenance statistics



$$< CF2^x, CF1^x, CF1^t, n, CL, \epsilon, \Theta, \alpha, \Omega >$$

$$centroid(x) = \frac{CF1^x}{n}$$

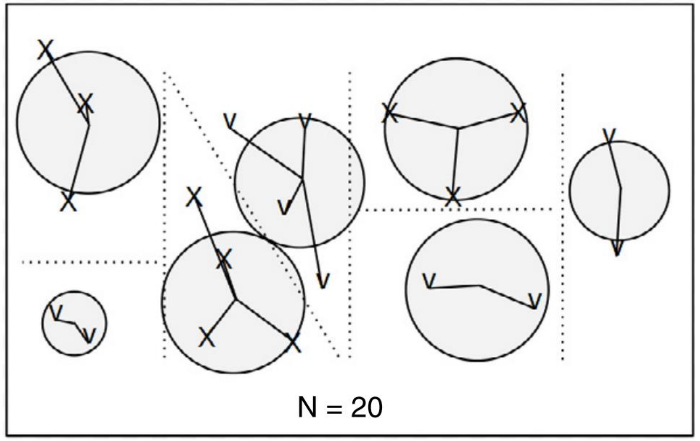
$$Variance[x] = \sqrt{\left(\frac{CF2^x}{n}\right) - \left(\frac{CF1^x}{n}\right)^2}$$

- Initially a fixed number of Micro-Clusters is randomly initialised.
- Only components outlined in the table are stored.
- These can be used to calculate the clusters centroid and boundary (variance).

MC-NN: Micro Cluster based Neareast Neighbour, basic Approach

Objective: Develop a scalable predictive Data Stream classification

1) Initialising Micro-Clusters and maintenance statistics



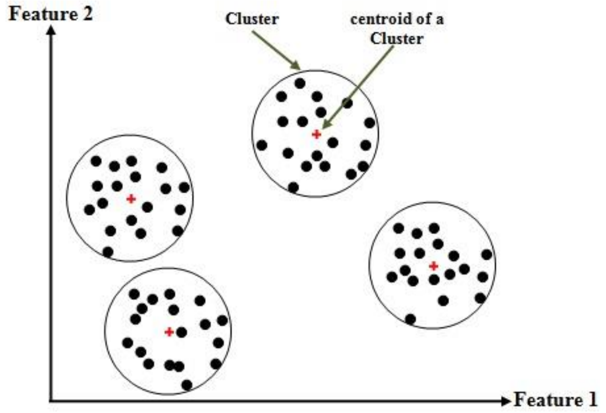
$$< CF2^x, CF1^x, CF1^t, n, CL, \epsilon, \Theta, \alpha, \Omega >$$

$$centroid(x) = \frac{CF1^x}{n}$$

$$Variance[x] = \sqrt{\left(\frac{CF2^x}{n}\right) - \left(\frac{CF1^x}{n}\right)^2}$$

- Initially a fixed number of Micro-Clusters is randomly initialised.
- Only components outlined in the table are stored.
- These can be used to calculate the clusters centroid and boundary (variance).

2) Absorbing new data stream instances

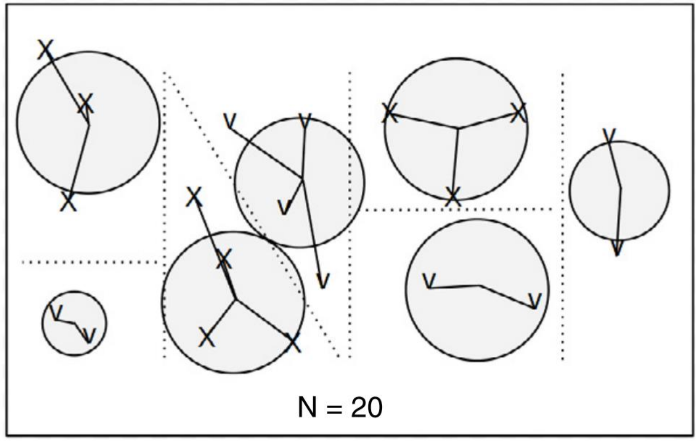


MC-NN: Micro Cluster based Neareast Neighbour, basic Approach

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Objective: Develop a scalable predictive Data Stream classification

1) Initialising Micro-Clusters and maintenance statistics



$$< CF2^x, CF1^x, CF1^t, n, CL, \epsilon, \Theta, \alpha, \Omega >$$

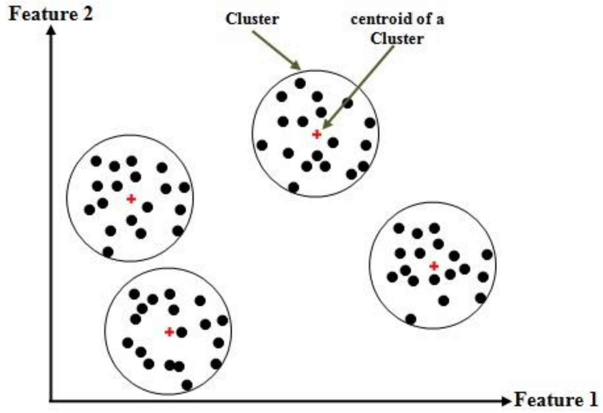
$$centroid(x) = \frac{CF1^x}{n}$$

$$Variance[x] = \sqrt{\left(\frac{CF2^x}{n}\right) - \left(\frac{CF1^x}{n}\right)^2}$$

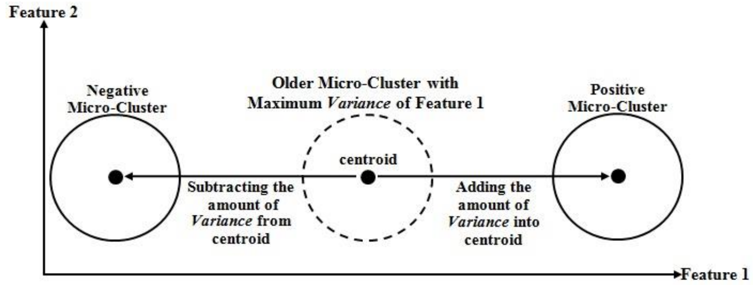
- Initially a fixed number of Micro-Clusters is randomly initialised.
- Only components outlined in the table are stored.
- These can be used to calculate the clusters centroid and boundary (variance).

2) Absorbing new data stream instances

EPSRC
Engineering and Physical Sciences Research Council



3) Splitting and removing of Micro-Clusters



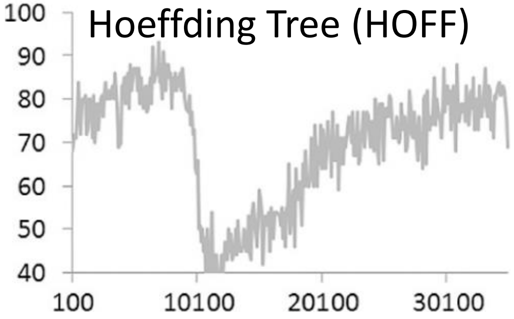
Published in: Tennant, M., Stahl, F., and Gomes, J.B. (2015), Fast Adaptive Real-Time Classification for Data Streams with Concept Drift, Proceedings of 8th International Conference on Internet and Distributed Computing Systems, Windsor, England, Springer LNCS, pp 265-272.

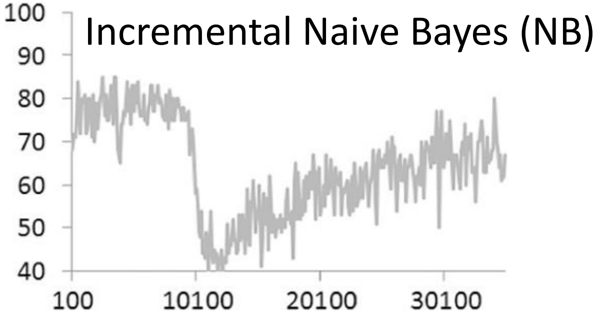
MC-NN: Results

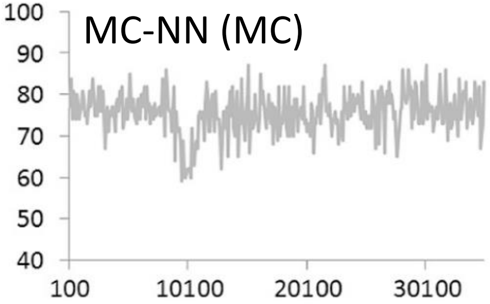
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

EPSRC
Engineering and Physical Sciences Research Council

Adaptating to Concept Drift

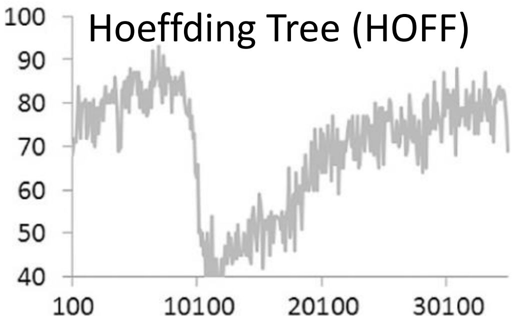


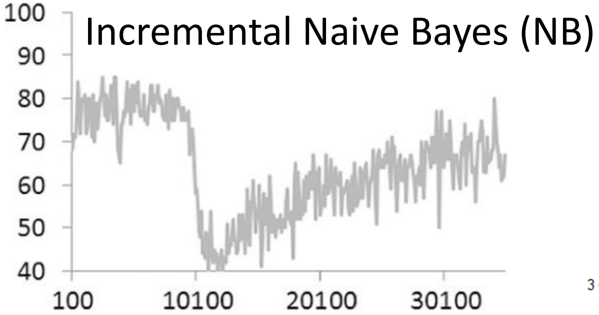


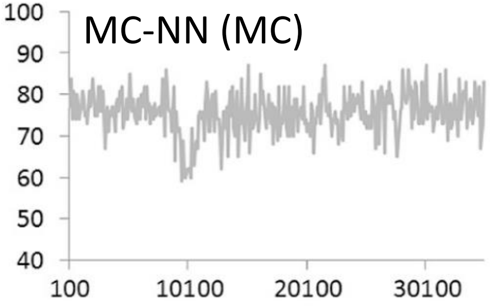


MC-NN: Results

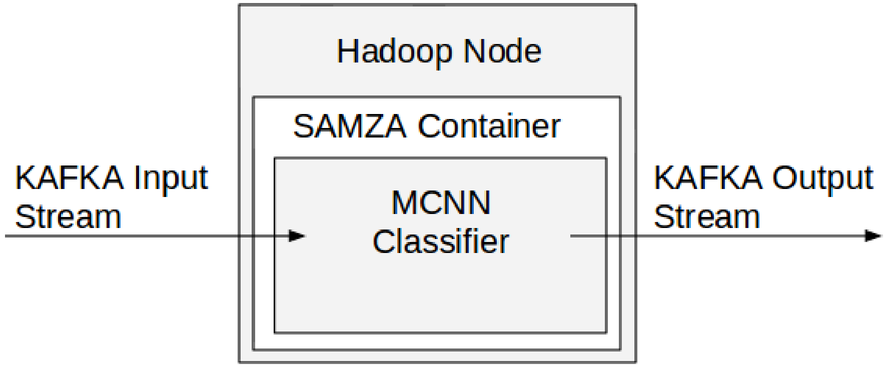
Adaptating to Concept Drift

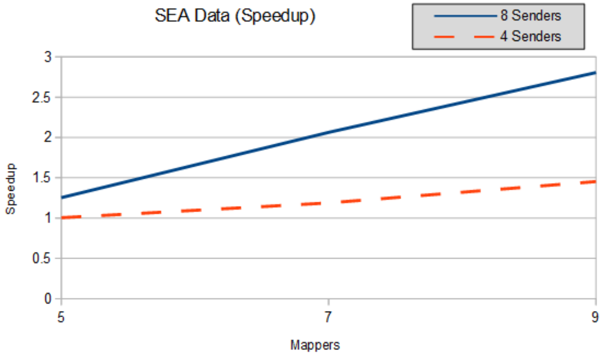


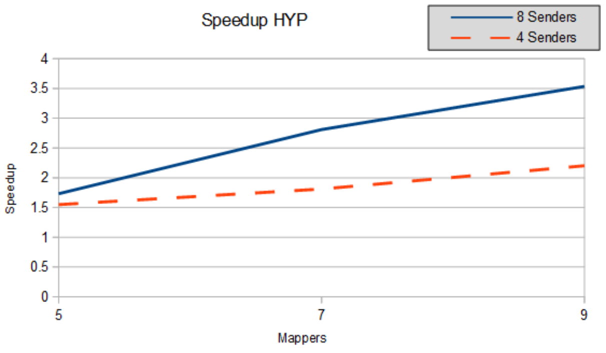




Scalability through parallelisation



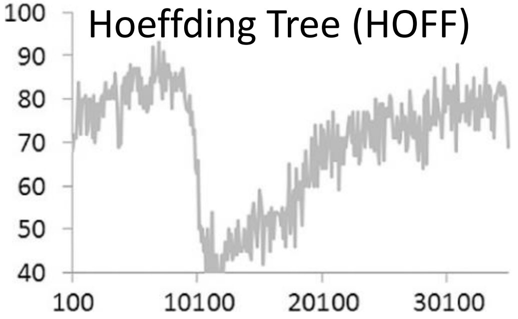


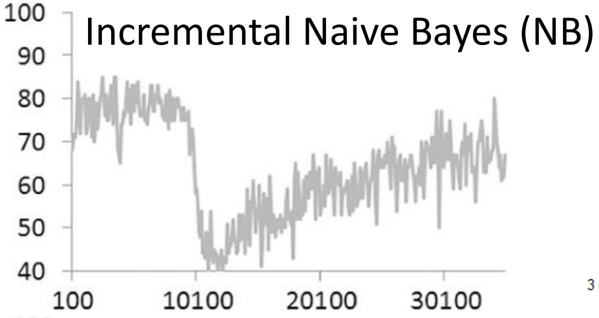


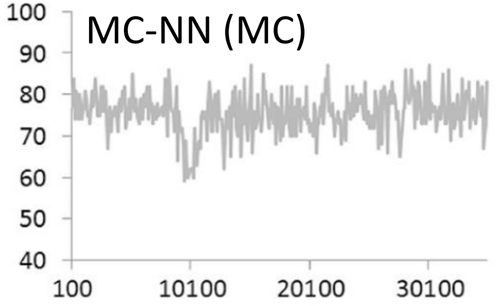
MC-NN: Results

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

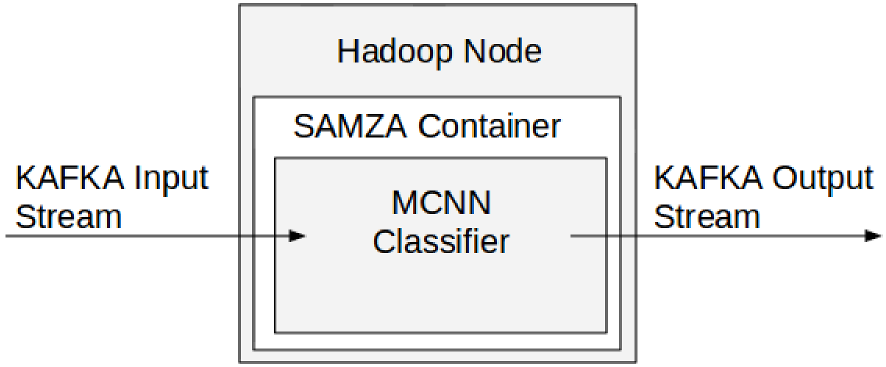
Adaptating to Concept Drift

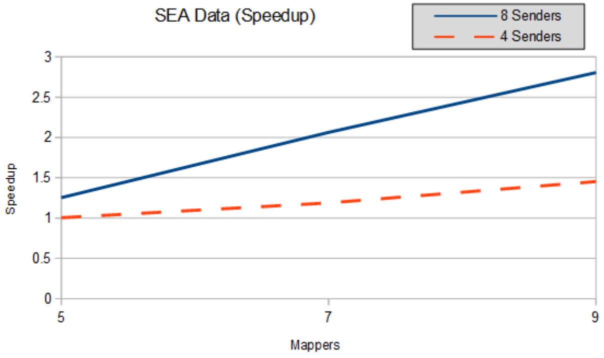


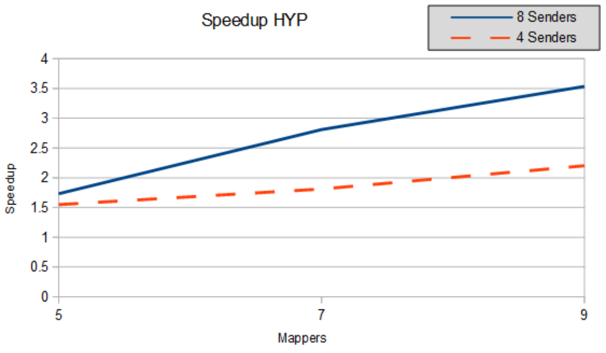




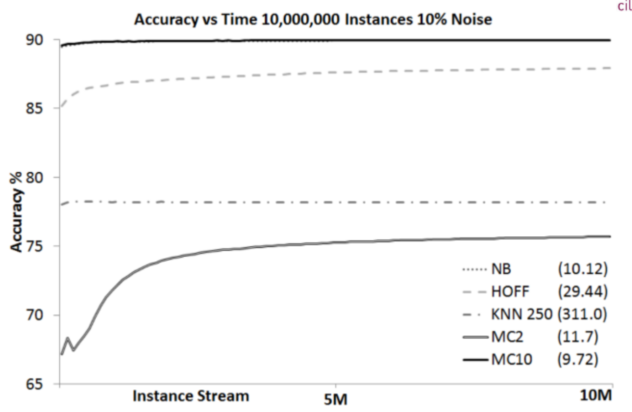
Scalability through parallelisation

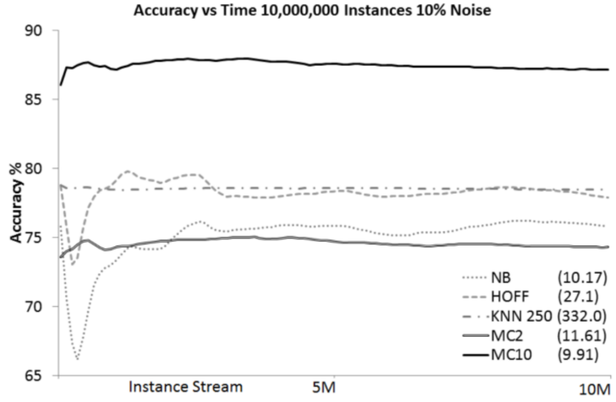






Long term adaptation



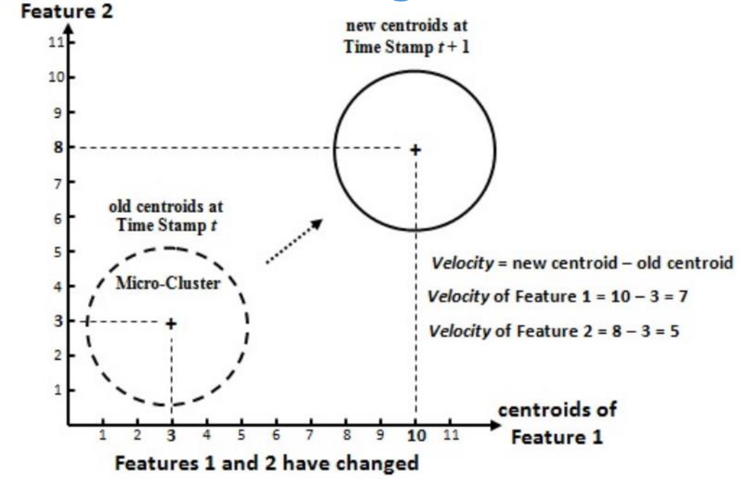


Published in: Tennant, M., Stahl, F., Rana, O. and Gomes,J.B., (2017) Scalable real-time classification of data streams with concept drift, Future Generation Computer Systems, Elsevier, 75, pp. 187-199, ISSN 0167-739X doi: 10.1016/j.future.2017.03.026

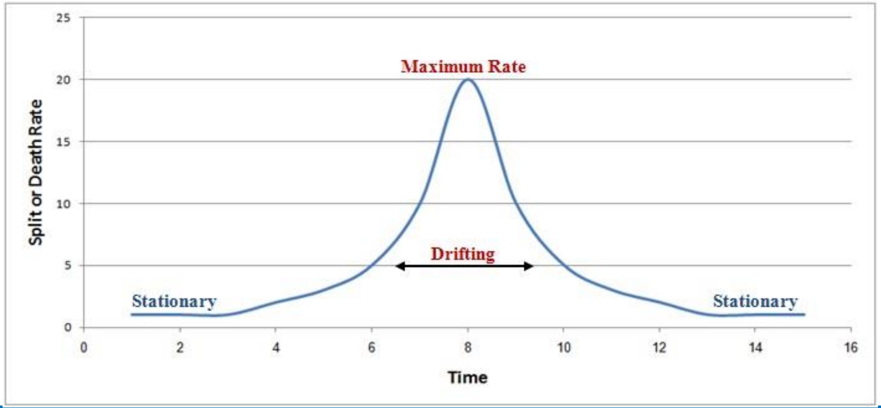
Using MC-NN for Explaining Concept Drift through Feature Tracking

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Measuring centroid velocity



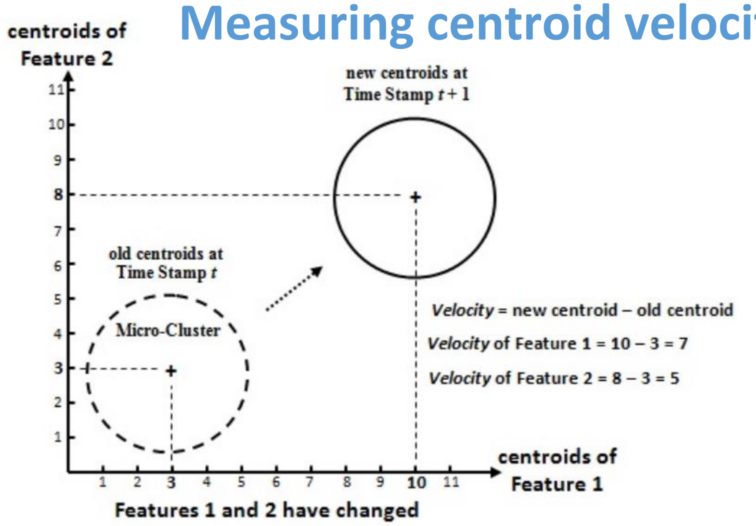
Measuring split & death rate



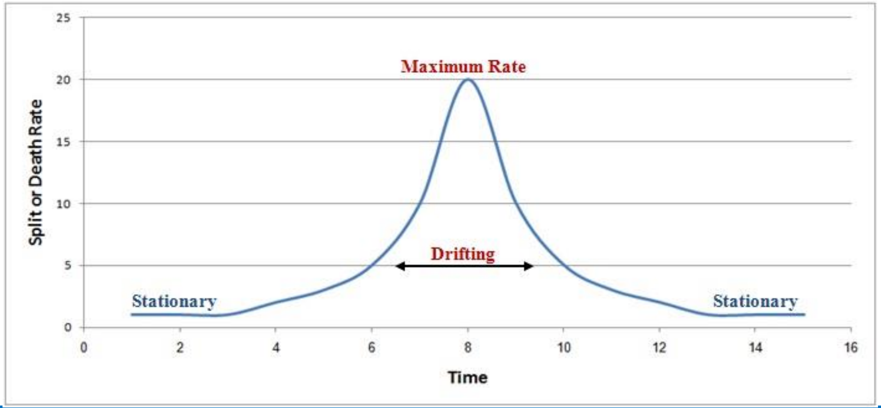
Using MC-NN for Explaining Concept Drift through Feature Tracking

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

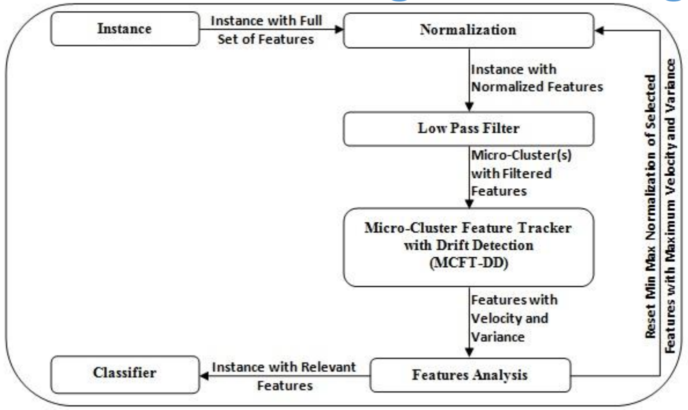
Measuring centroid velocity



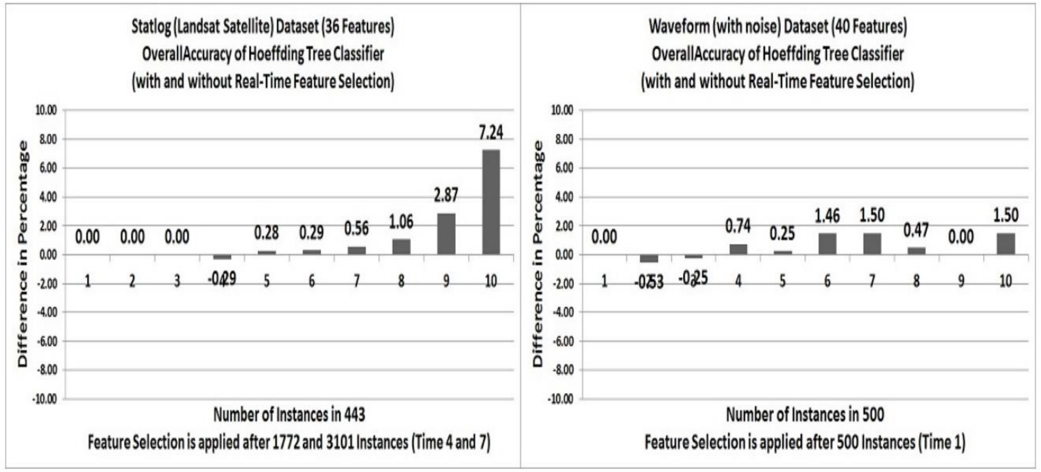
Measuring split & death rate



Feature tracking and ranking

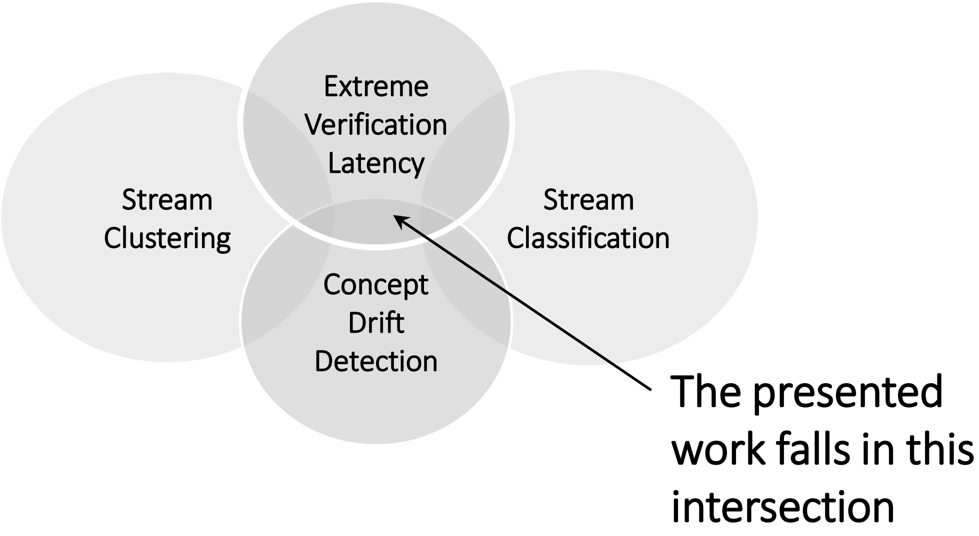


Results

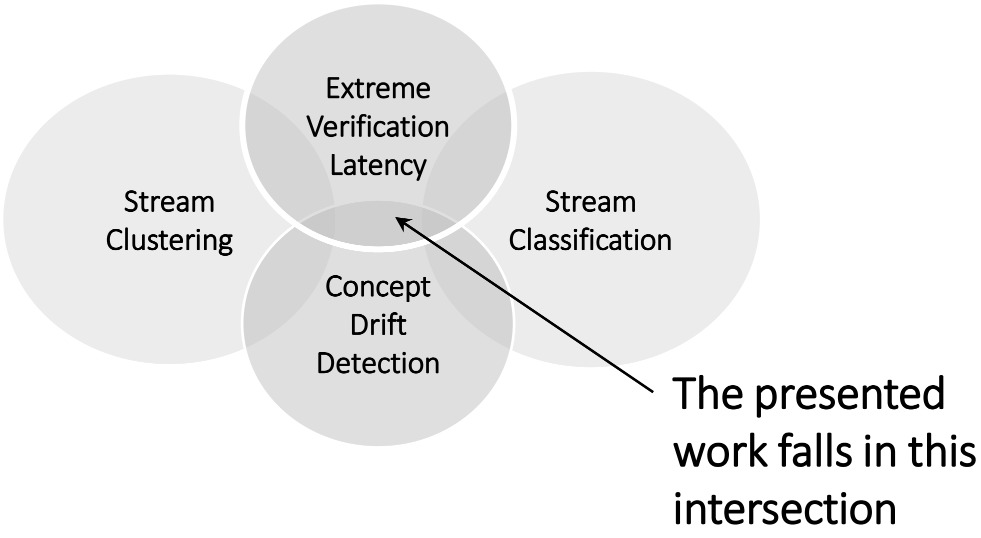


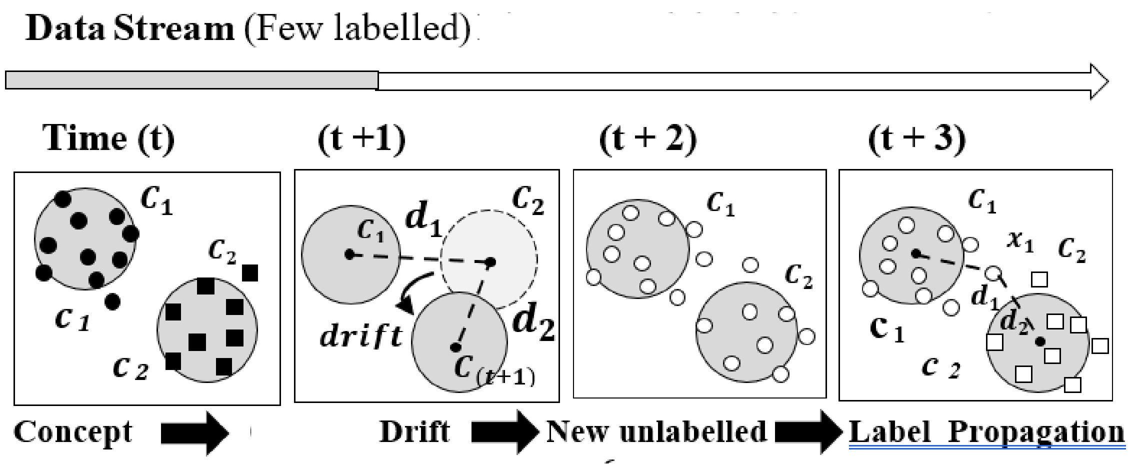
Published in: Hammoodi, M. S., Stahl, F. and Badii, A., (2018) Real-time feature selection technique with concept drift detection using Adaptive Micro-Clusters for data stream mining. Knowledge-Based Systems, Elsevier, 75, pp. 205-239.

MC-NN: Unsupervised Classification



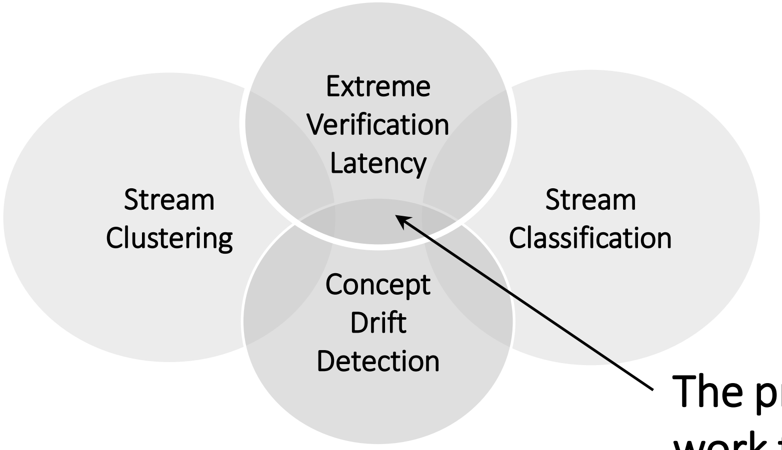
MC-NN: Unsupervised Classification



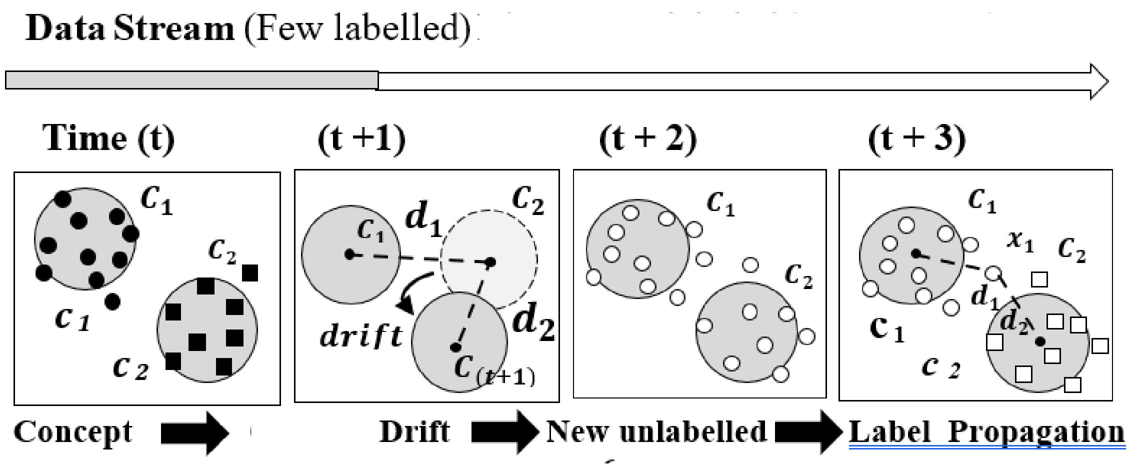


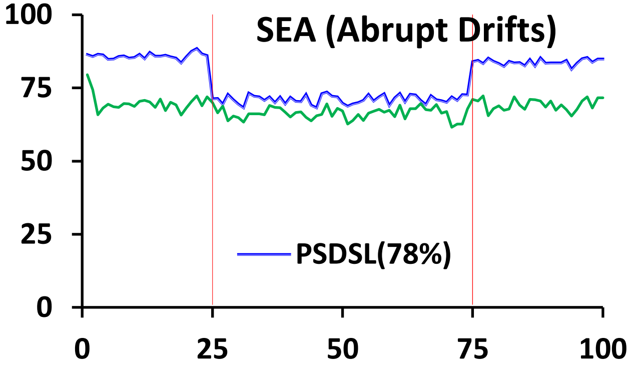
MC-NN: Unsupervised Classification

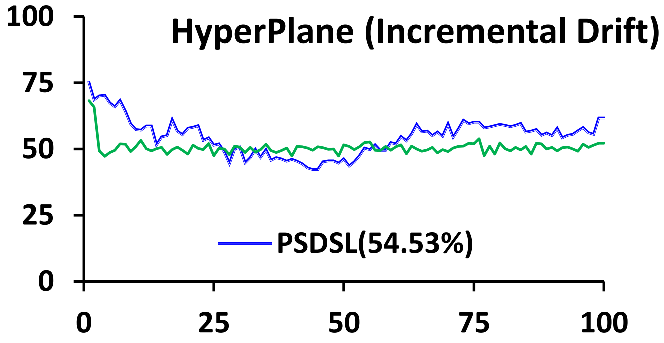
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

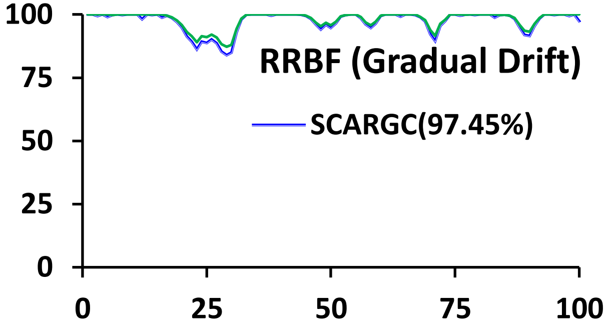


The presented work falls in this intersection









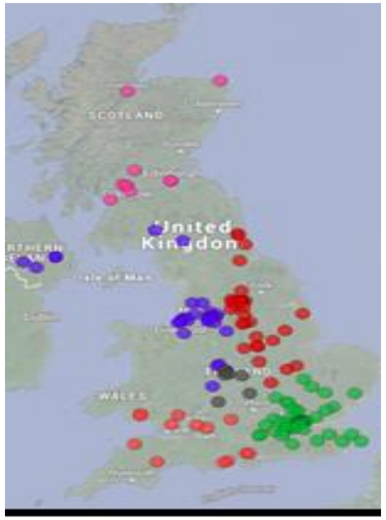
Published in: M. M. Idrees, F. Stahl and A. Badii, "Adaptive Learning with Extreme Verification Latency in Non-Stationary Environments," in IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3225225.

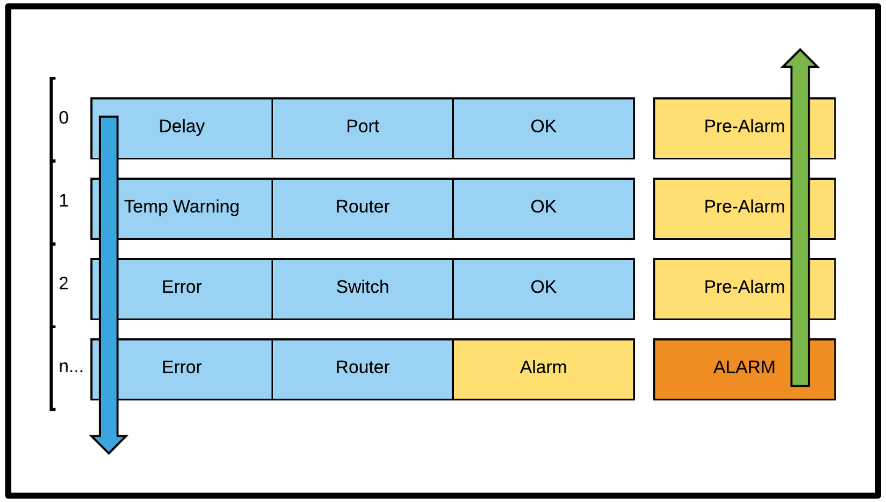
Applications: BT (completed)

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Example Problem: Real-time Network Alarm Forecasting

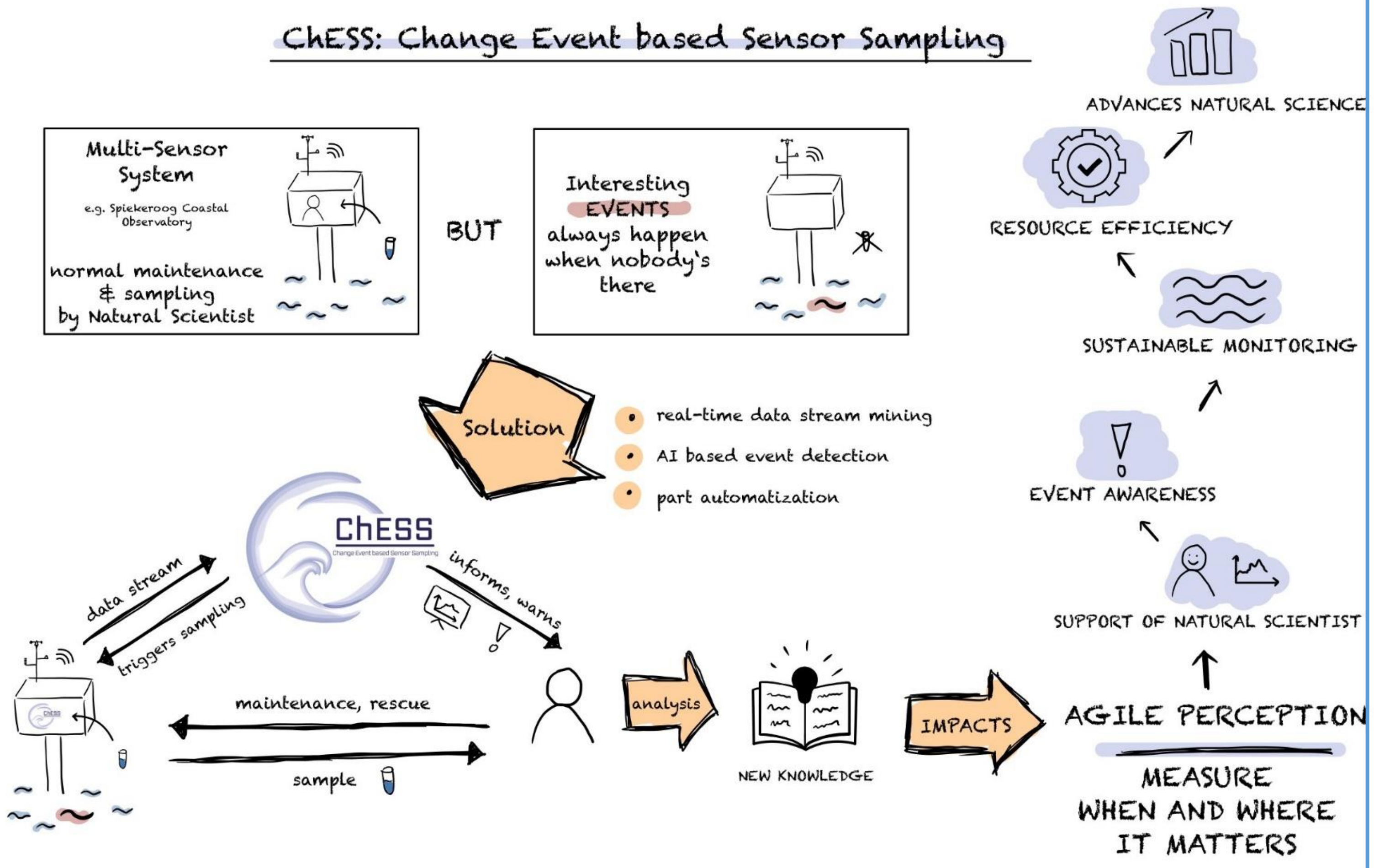
- Increasing reliance on Telecommunication services for business and personal use
- Telecommunication Networks have a great deal of redundancy (99.999% availability), however, the “last mile” is often a single point of failure
- Network devices emit different events data at different frequencies under different conditions. Yet they may be linked.





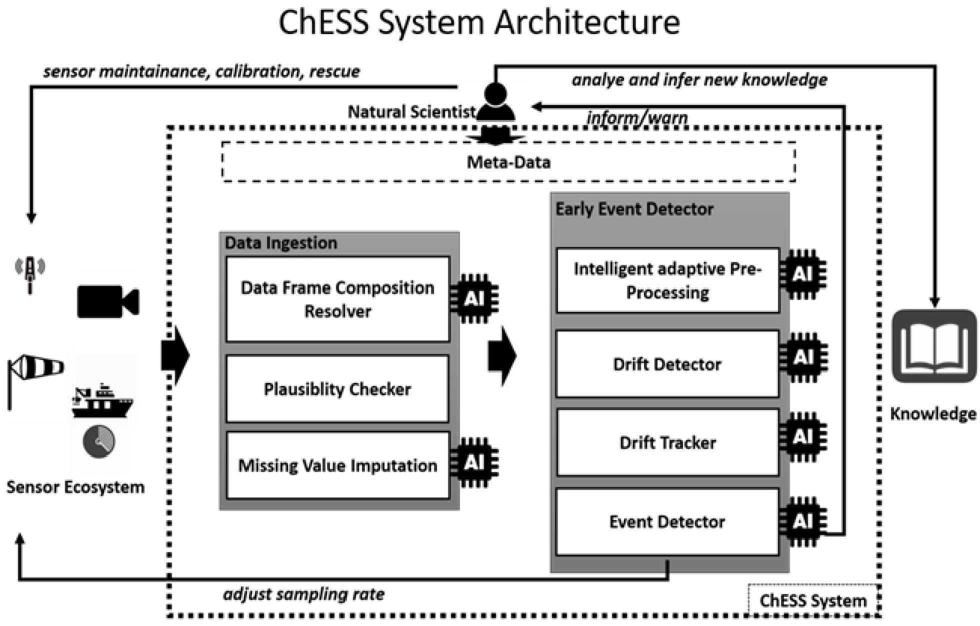
Published in: Wrench, C., Stahl, F., Di Fatta, G., Karthikeyan, V., Nauck, D. (2019) A rule induction approach to forecasting critical alarms in a telecommunication network. In: 2019 IEEE International Conference on Data Mining Workshops (ICDMW), 2019, Beijing, China.

Systems Development: ChESS (ongoing)



Gefördert durch:

Niedersächsisches Ministerium für Wissenschaft und Kultur

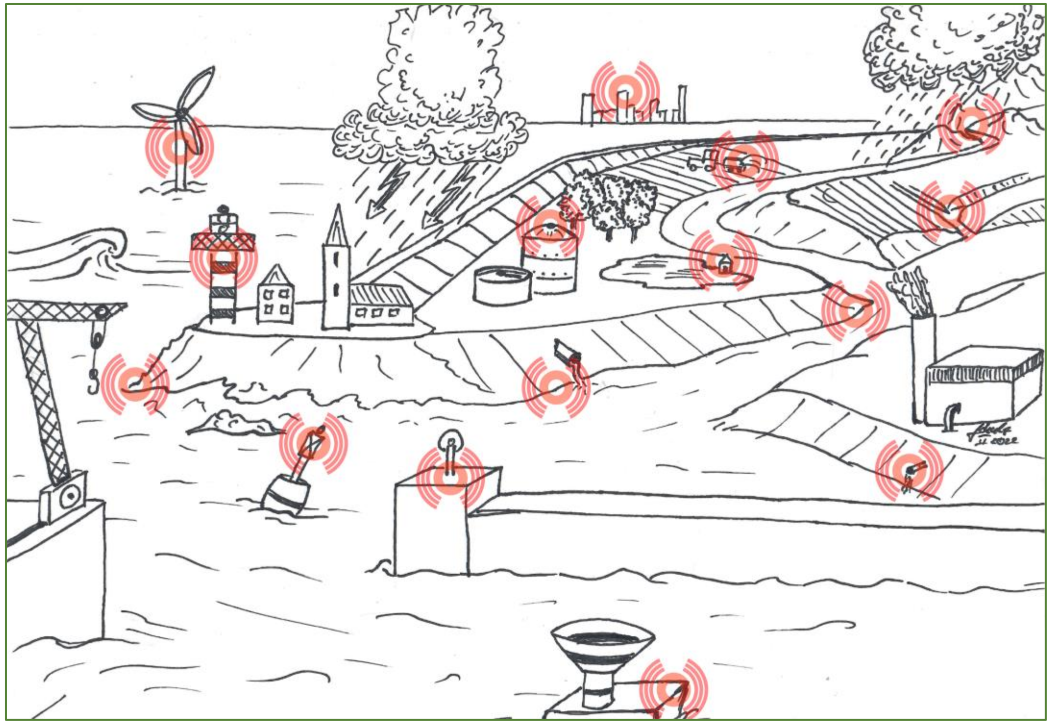


Published in: Lukats, D., Berghöfer, E., Stahl, F., Schneider, J., Pieck, D., Idrees, M.M., Nolle, L., Zielinski, O. (2021), Towards Concept Change Detection in Marine Ecosystems, In: IEEE Journal of Oceanic Engineering (OES) OCEANS 2021 San Diego – Porto Online Proceedings. OCEANS MTS/IEEE Conference (OCEANS-2021), pp. 1-10.

Applications: Intelligent Maintenance of Costal Environments (just starting)

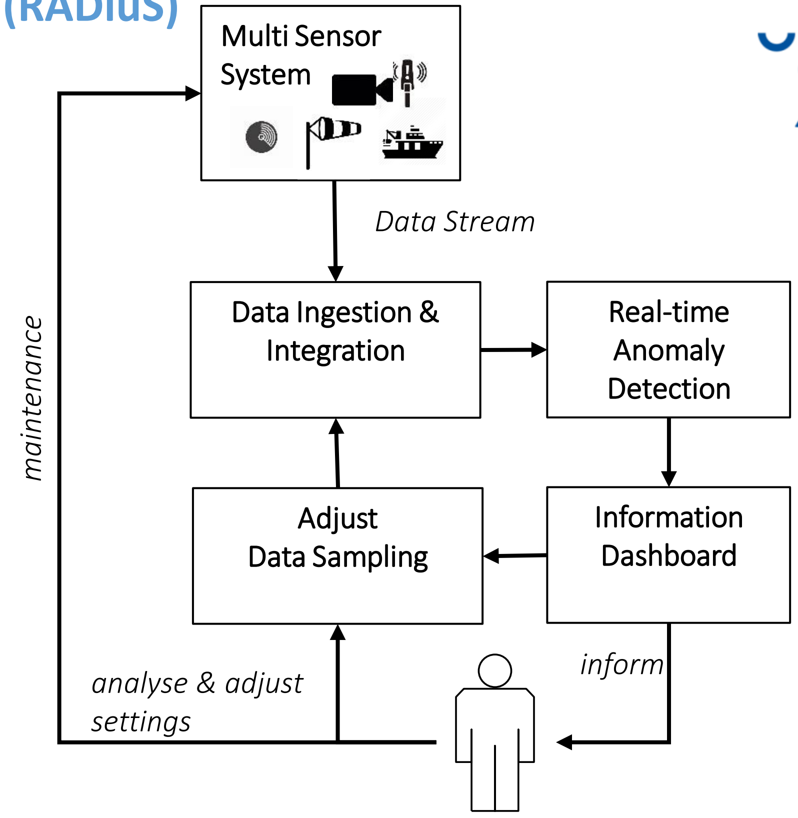
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Ad-hoc data acquisition mesh for enhanced versatile explorations of waters

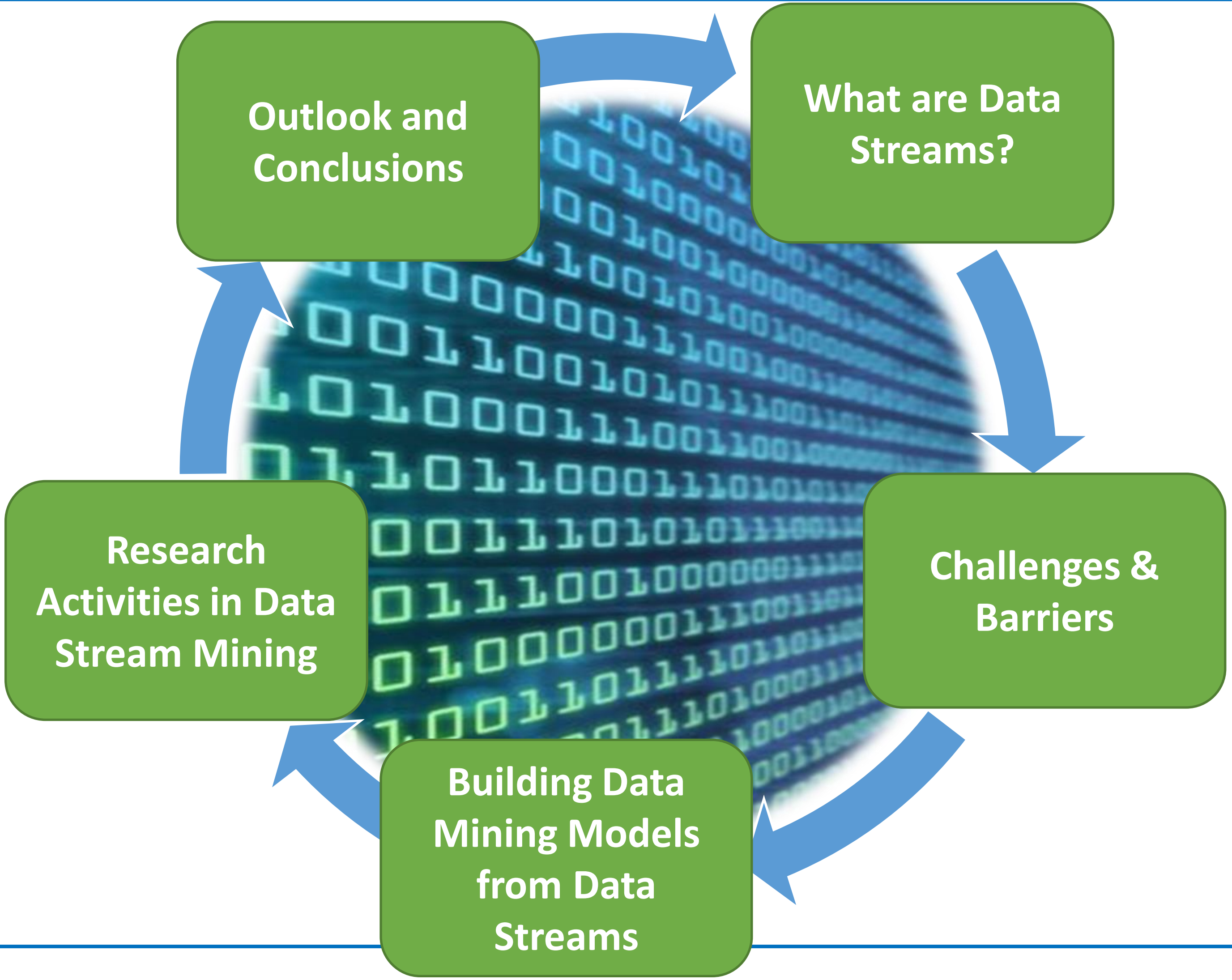


@ Prof. Jan Schulz

Real-time anomaly detection in aquatic system (RADiuS)



ZDIN Zentrum für digitale Innovationen Niedersachsen
ZUKUNFTSLABOR WASSER


What are Data Streams?
Challenges & Barriers
Building Data Mining Models from Data Streams
Research Activities in Data Stream Mining
Outlook and Conclusions

The Data Tsunami

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Challenges

1) Data generated at a fast rate (Velocity), at potentially large and unknown quantities (Volume)
2) Concept Drift (changes of pattern encoded in in the data over time)
3) Modelling real-time analytics workflows from streaming data
4) Multi-modality of data sources (text, video/images, unstructured)
5) Class label sparsity: adapting predictive models
6) Explaining Concept Drift

Barriers

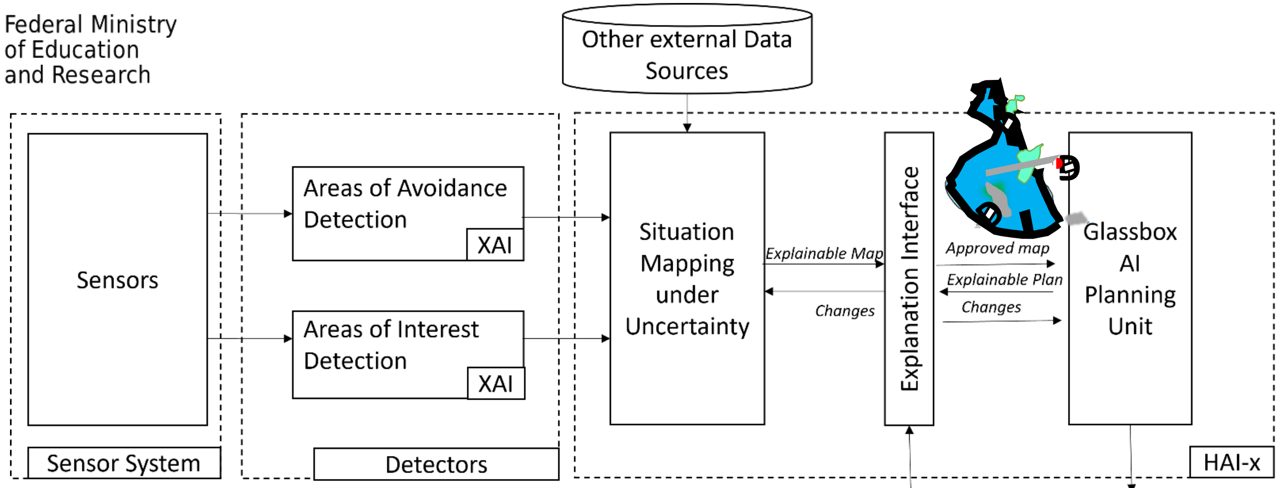
1) Limited scalable (parallel) real-time high throughput data stream mining algorithms
2) Different and changing types of concept drift
3) Lack of customisable pre-processing techniques
4) Different time stamps but co-occurring data items
5) Supervised algorithms not applicable in many cases
6) Lack of drift detectors explaining concept drift

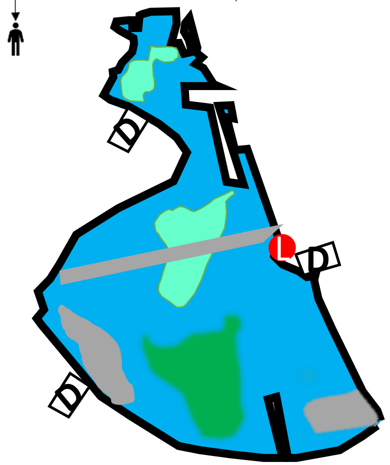
Future Directions

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Hybride AI Explainer (HAI-x): weed harvester scenario

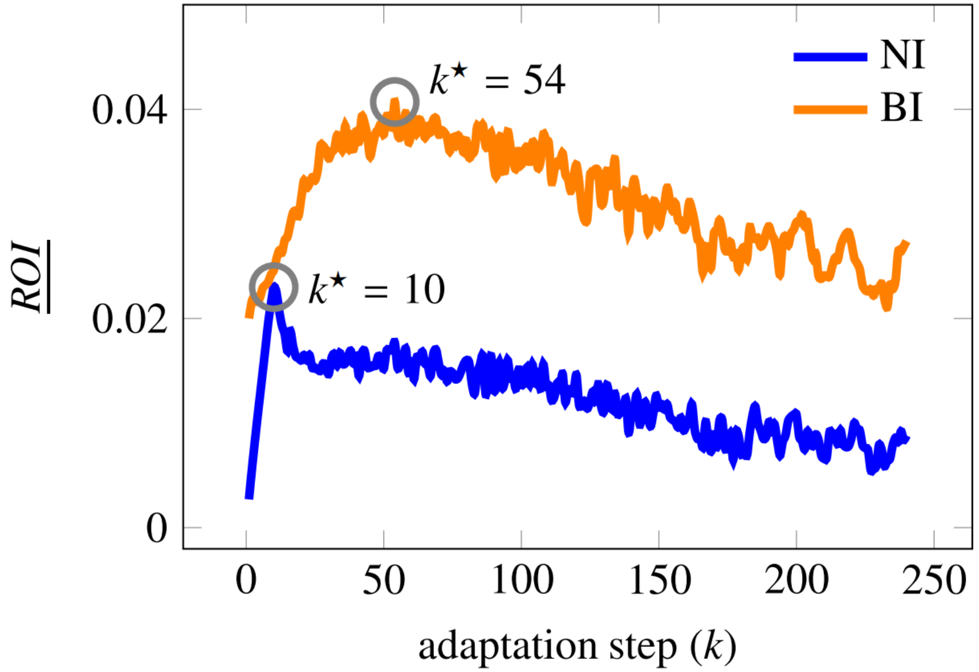
Federal Ministry of Education and Research





Utility of Adaptation: when is it worth updating your model?

ROI is return on employing an adaptive predictor as compared to keeping a fixed nonadaptive model



Source: Zliobaite, I., Budka, M. and Stahl, F. (2015) Towards cost-sensitive adaptation: When is it worth updating your predictive model? Neurocomputing, Elsevier, 150 (A), pp. 240-249. ISSN 0925-2312 doi: 10.1016/j.neucom.2014.05.084.

Thank you!

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- Marine Perception Team at DFKI
- Prof. Oliver Zielinski
- Prof. Lars Nolle
- The team of the Department of Computer Science at the University of Reading
- Prof. Giuseppe Di Fatta
- UK and ISG Teams
- Prof. Mohamed Gaber
- My wife Laura
- And others.

Questions?

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH