STREAMLINE: Improving Competitiveness of European Enterprises through Streamlined Analysis of Data at Rest and Data in Motion

MTA SZTAKI

LEARNING FROM DATA STREAMS THEORY AND PRACTICE

ANDRÁS BENCZÚR

INSTITUTE FOR COMPUTER SCIENCE AND CONTROL

HUNGARIAN ACADEMY OF SCIENCES (MTA SZTAKI)

JOINT WORK WITH

- **DOMOKOS KELEN, DANIEL BERECZ** (FLINK PARAMETER SERVER)
- **ROBERT PALOVICS** (RECOMMENDERS, NOW AT STANFORD)
- **LEVENTE KOCSIS** (REINFORCEMENT LEARNING – ECML TEST-OF-TIME PRIZE 2016 / BANDIT BASED MONTE-CARLO PLANNING 2006 W/ SZEPESVÁRI)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688191

NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS HIVATAL

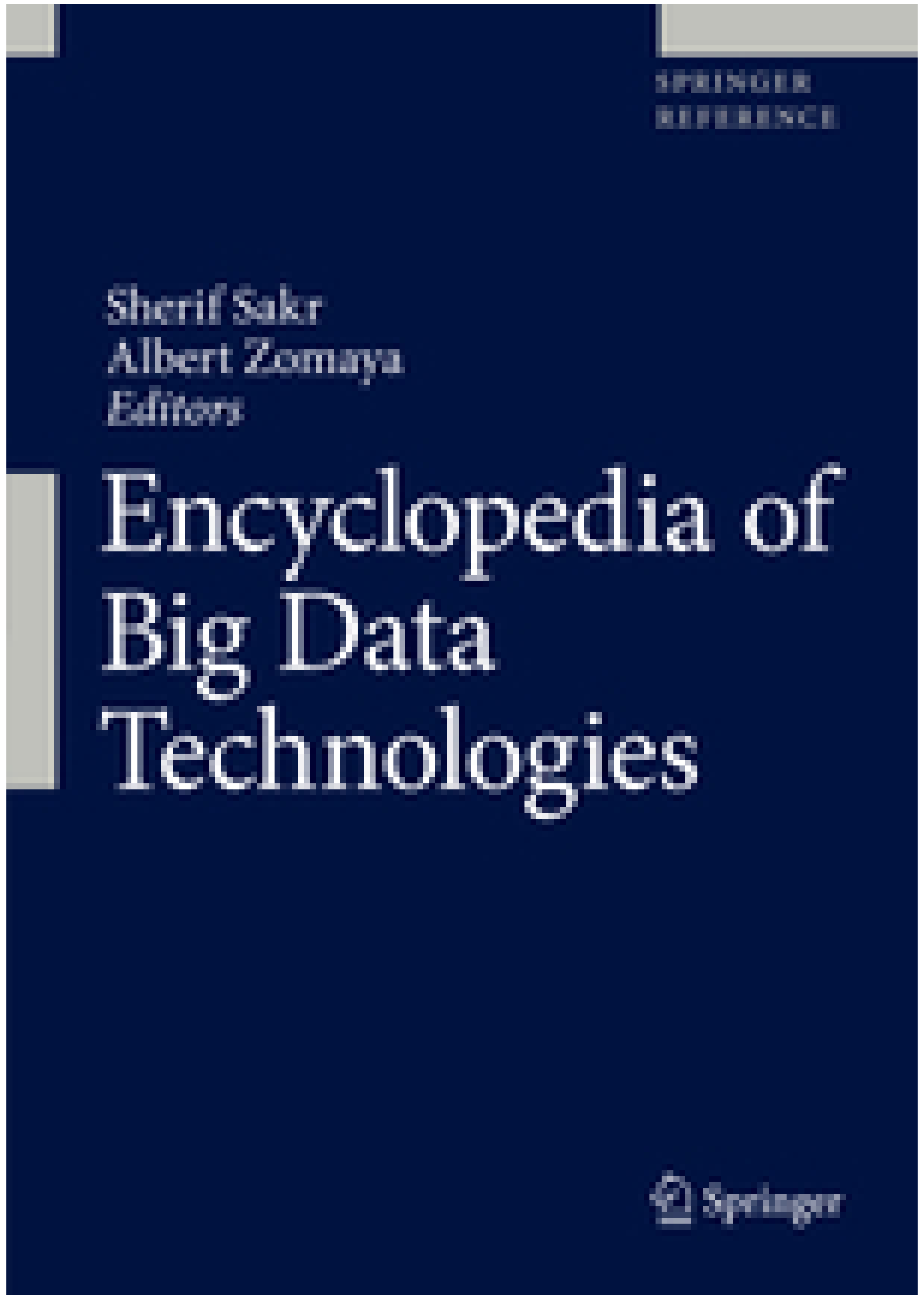
AZ INNOVÁCIÓ LENDÜLETE

AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

AIME 29/10/2018

Presentation based on four chapters on Online Machine Learning in Big Data

1. Requirements and distributed systems
   - The Parameter Server
2. Classification
   - Linear methods, Neural networks; Trees
3. Recommender Systems
4. Other online learning methods
   - Reinforcement learning
   - Unsupervised
   - Concept drift



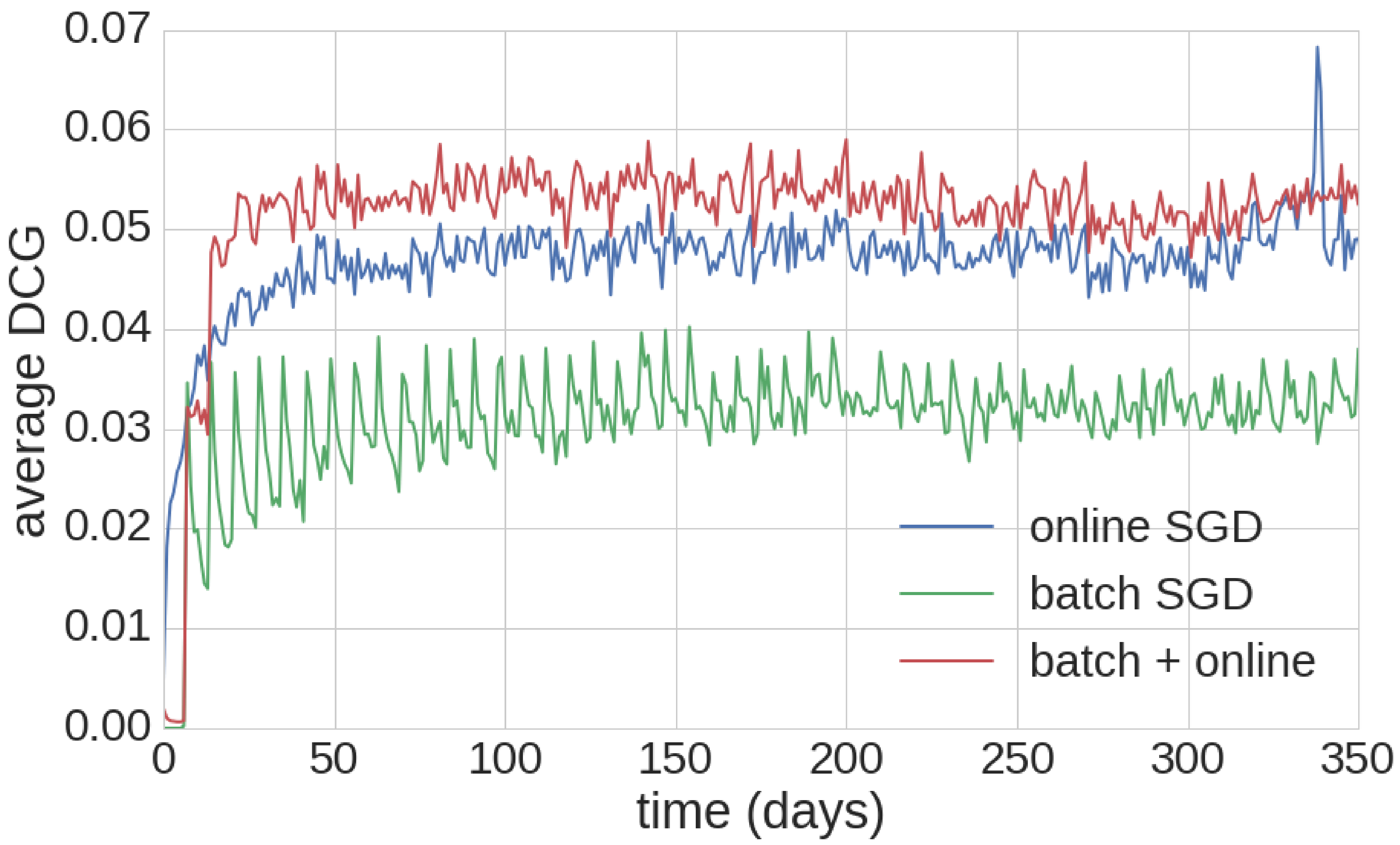
Due: February 14, 2019
Chapters should be available online
All in one ArXiv: 1802.05872

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM
AZ INNOVÁCIÓ LENDÜLETE

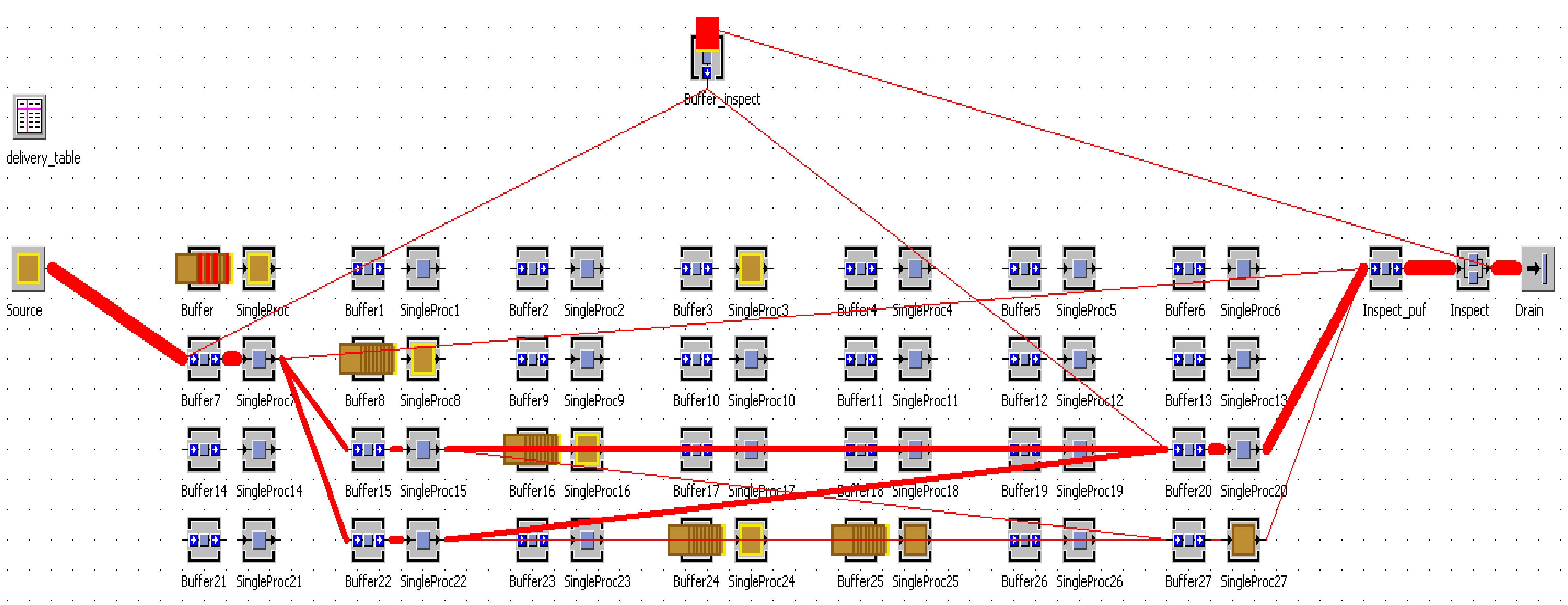
I will show two applications

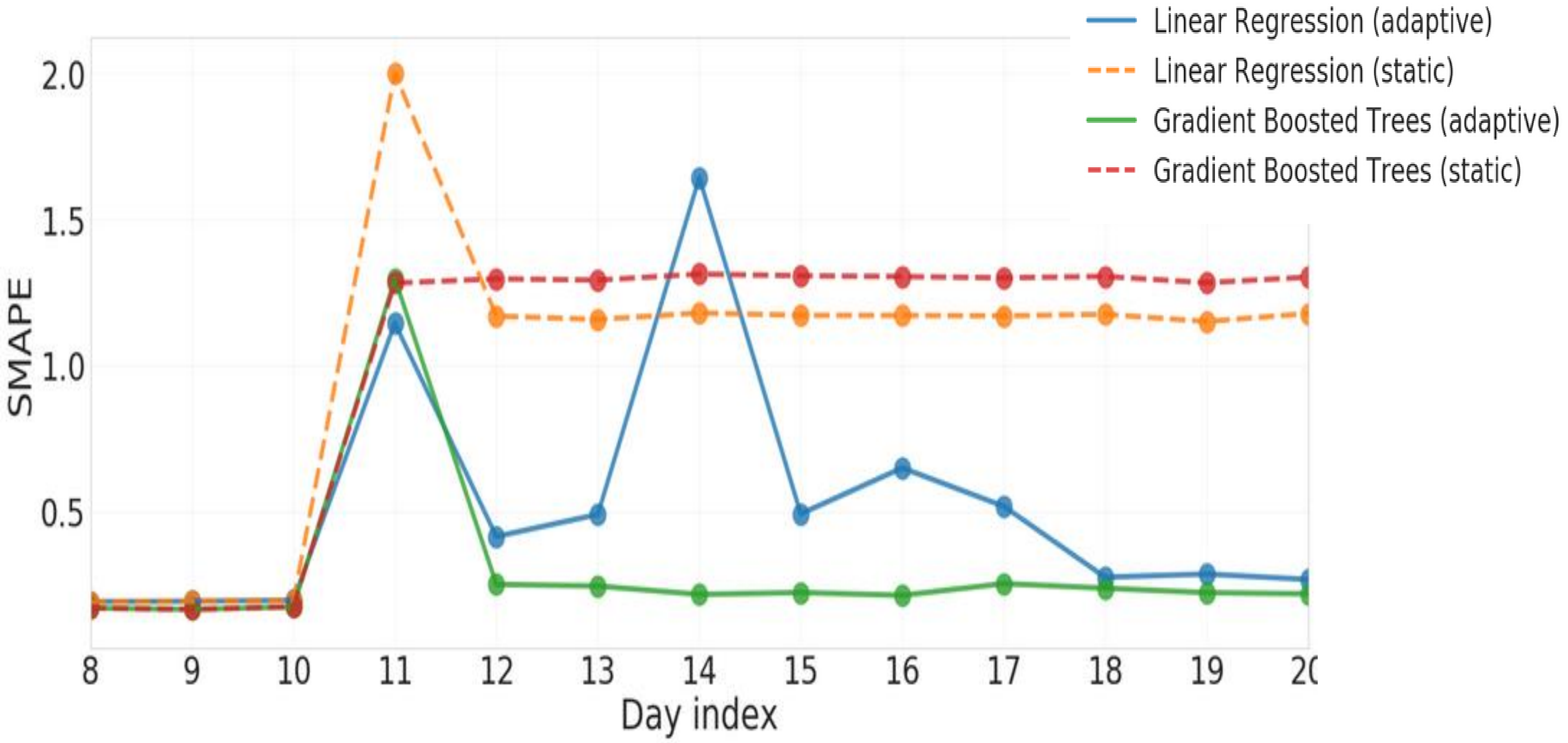
Recommenders

- Surprisingly, reading the data only once and forgetting helps!
- Our first main observation from back in 2013



Concept Drift in industrial IoT time series





Terminology – all depend on data scale („Big Data”)

MTA SZTAKI

Batch

- Repeatedly read all training data multiple times
- Stochastic gradient descent: use multiple times in random order
- Elaborate optimization procedures, e.g. SVM, gradient boosting

+ More accurate (?)

+ Easy to implement (?)

Online learning

- Update immediately, e.g. with large learning rate

Data streaming

- Read training/testing data only once, no chance to store

Real time / Interactive

+ More timely, adapts fast

- Challenging to implement

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
AZ INNOVÁCIÓ LENDÜLETE
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

Online learning in Big Data: Overview of the main issues

1. Limitations of the data steam model. Limited memory.
2. Certain evaluation methods fail since model can change during evaluation.
3. Concept Drift can occur.
4. Scalability, distributed processing.

Issue 1: Algorithmic limitations

MTA SZTAKI

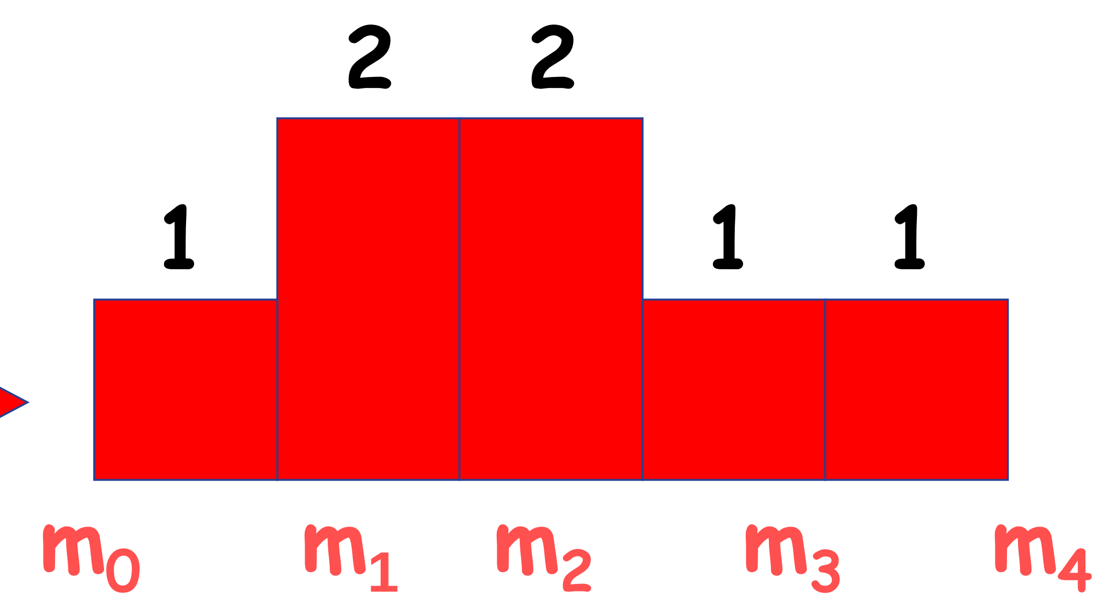
- We have to update the model after each data instance
- No access to (majority of) past data
  - No stochastic gradient descent: we cannot iterate over the data
    - Online gradient descent is possible
  - No decision trees: after deciding about a split, we cannot use data inside the new nodes for a next split
    - If confident about a split seeing some data, build further splits by the new data
    - Use concept drift statistics to rebuild certain branches
- Data streaming computational model

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
AZ INNOVÁCIÓ LENDÜLETE
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

The Data Streaming Computational Model
Illustration: number of different elements in data

MTA SZTAKI

Data Stream: 2, 0, 1, 3, 1, 2, 4, . .



- In-Memory
  - Hash tables
- On disk
  - Sort (mergesort)
- Distributed
  - Map-reduce – basically sorting again

- Streams?
  - No exact algorithm without storing all data – Proof:
    - Trivial information bound to decide if the next element is new or already present earlier in the stream
    - We may reconstruct the entire past stream as a set by probing with next elements
  - Random sampling fails very bad on rare elements
    - Assume sample consist of identical elements only
    - A likely answer is that there are no other elements
    - But we may have, with large probability, a large number of other elements that appear only once
    - Numerical example: 20% random sample, for any guess there is a data stream where relative error on count > 20%
  - Distinct sampling: read all data, make adaptive decisions

Muthukrishnan, S., et al.: Data streams: Algorithms and applications. Foundations and Trends in Theoretical Computer Science1(2), 117–236 (2005)

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM
AZ INNOVÁCIÓ LENDÜLETE

Issue 2: Difficulty in evaluation

- Model changes right after prediction is made
  - Precision, Recall and many other metrics compare a SET of items consumed against recommended
  - But for the next item consumed, a new model may potentially recommend completely different items
- Natural evaluation metric is clickthrough rate
  - Equivalent of the “Precision” of a single item
- AUC for classification is also a problem

- Prequential (predictive sequential) evaluation
  1. Give a prediction for the next data point
  2. Read its label, compare to the prediction and update quality metrics
  3. Update the model to be used for the next data point
- Slightly modified metrics are needed

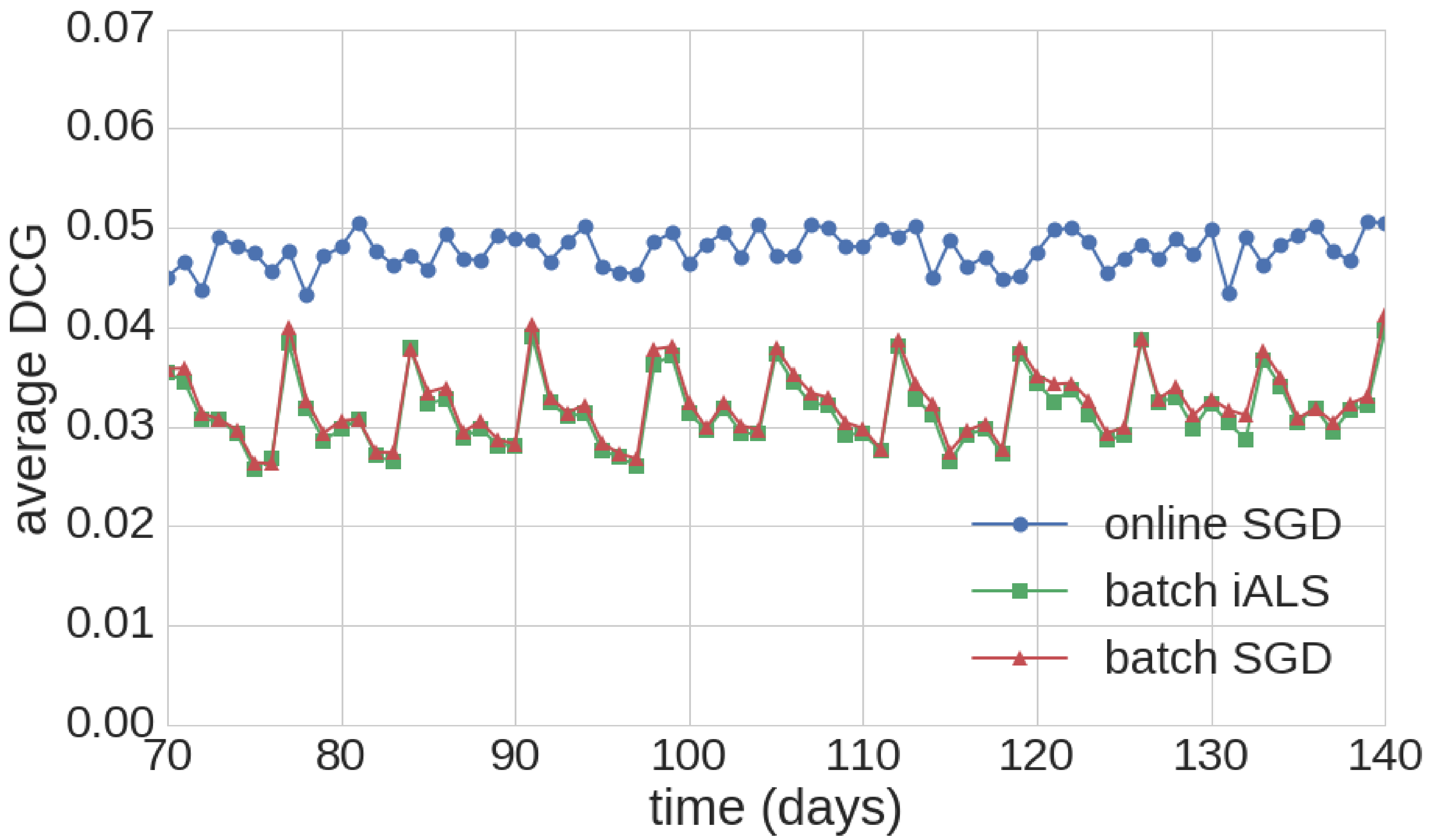
Dawid, A.P.: Present position and potential developments: Some personal views: Statistical theory:The prequential approach. Journal of the Royal Statistical Society. Series A (General) pp. 278–292 (**1984**)

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
AZ INNOVÁCIÓ LENDÜLETE
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

Issue 3: Concept drift

MTA SZTAKI

- Model performance very often deteriorates in time
- Observe the weekly retraining periods in performance on the right
- Concept drift detection either
  - Detects sharp changes in distribution, e.g. failures, or
  - Measures deterioration to schedule retraining
- Auto-forgetting old events can be an option
  - Gradient descent with negative sample generation
- Remark about recommenders
  - Items have stable characteristics in time, maybe novelty peaks and decays later, both easily described by item popularity
  - Users frequently change taste – session-based effect

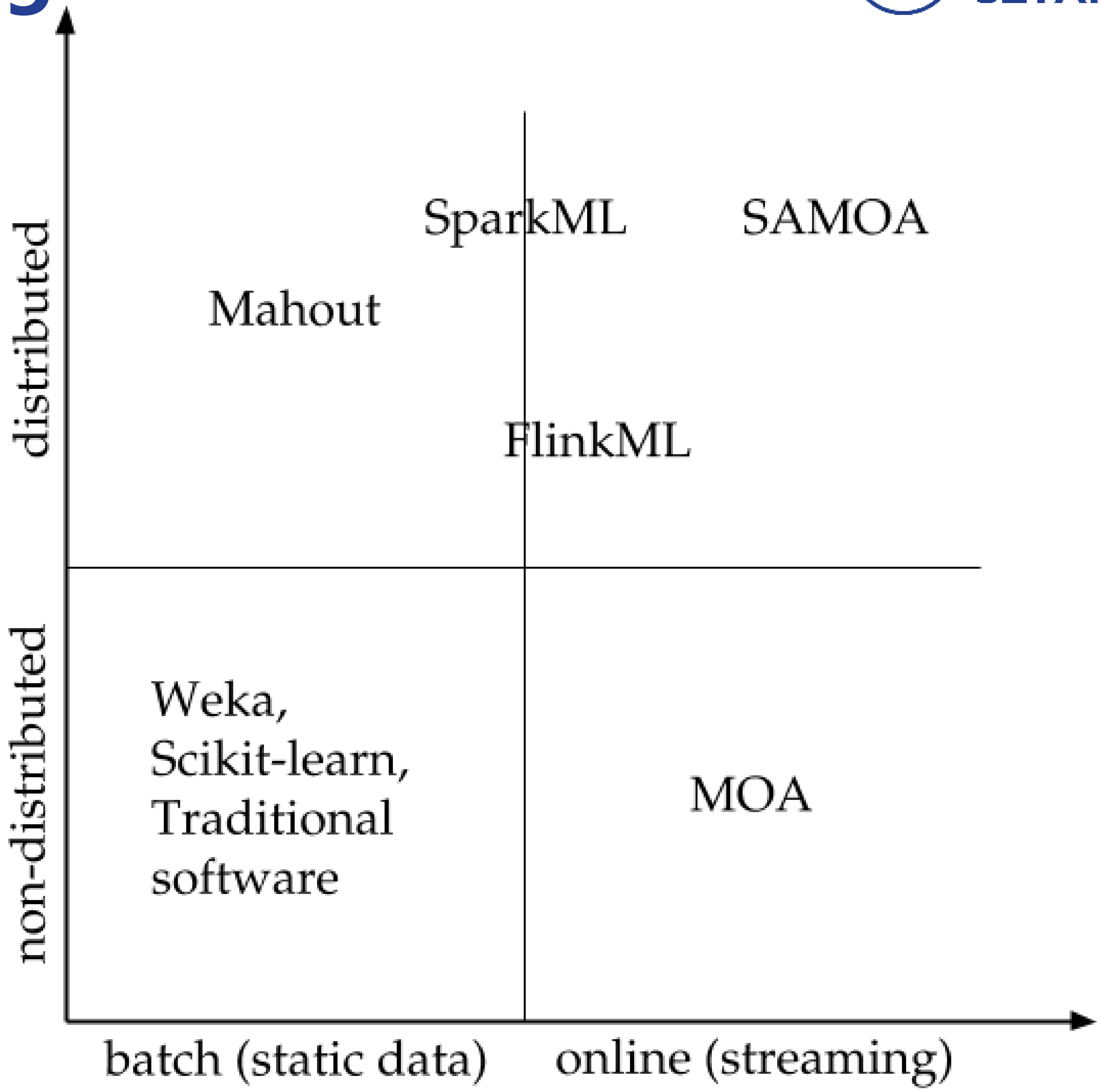


Last.fm "30M" Music listening dataset crawled by the CrowdRec team

NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS HIVATAL
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM
AZ INNOVÁCIÓ LENDÜLETE

Issue 4: Distributed stream processing architectures

- Easiest and typical way to train models is a single-server (maybe multicore or GPU but still in-memory) using static data
- Distributed is much less convenient
  - Install, optimize performance
  - Training labels are rarely available in huge quantities
- Online learning seems like a small niche application now
- Distributed and streaming
  - SAMOA: standalone library (Hoeffding trees, bagging, boosting, clustering) with connectors to many stream processing engines
  - SparkML: although Spark can process streams in mini-batches, SparkML methods are batch. Several external „parameter server" implementations
  - FlinkML under (also our) development with low community support – core system elements missing

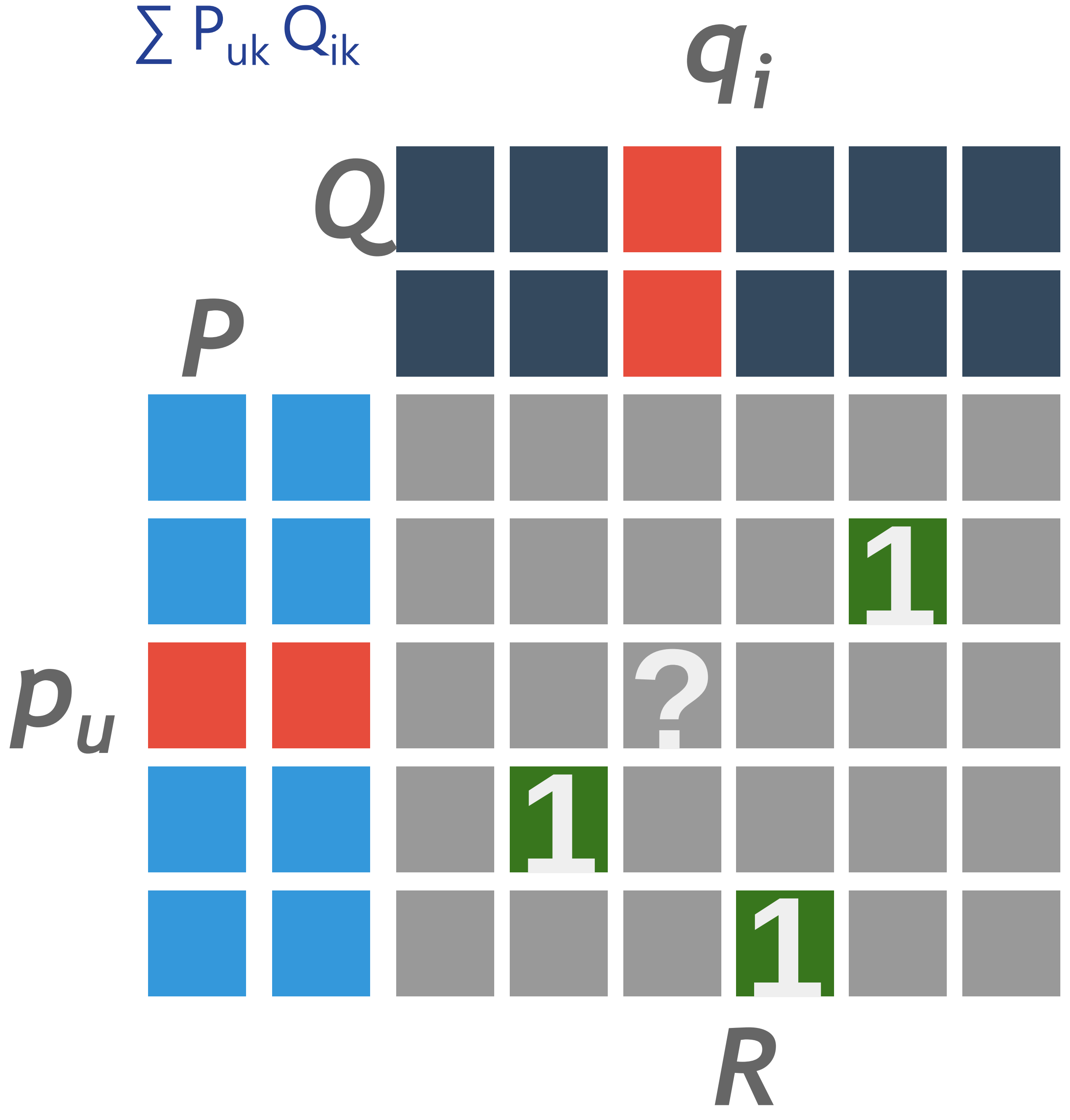


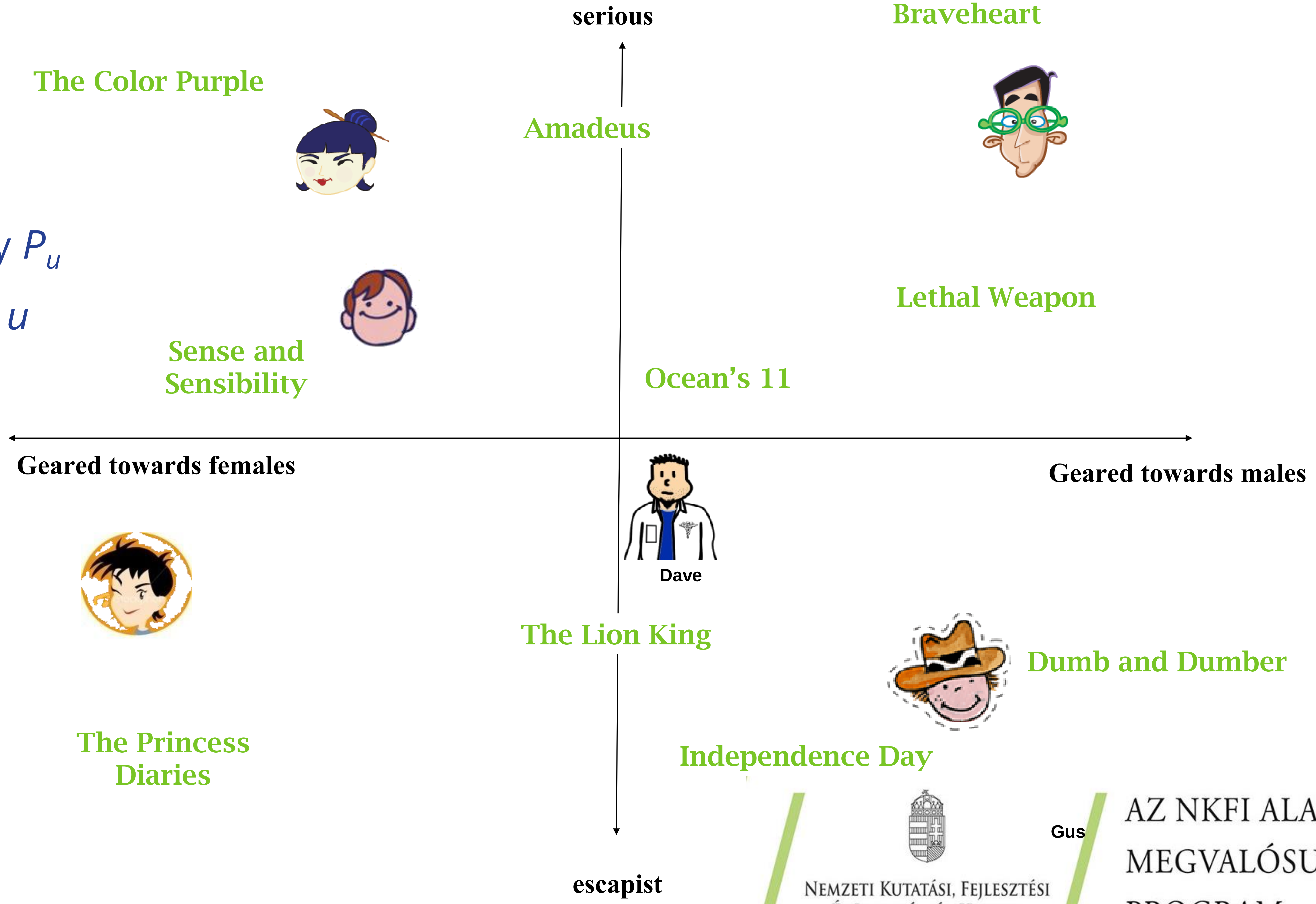
NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS HIVATAL
AZ INNOVÁCIÓ LENDÜLETE
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

Recommender Systems – illustration by Yehuda Koren

- Items and users described by unobserved factors
- Each item is summarized by a *d*-dimensional vector Q_i
- Similarly, each user summarized by P_u
- Predicted rating for Item *i* by User *u*
  - Inner product of Q_i and P_u

    $\sum P_{uk} Q_{ik}$



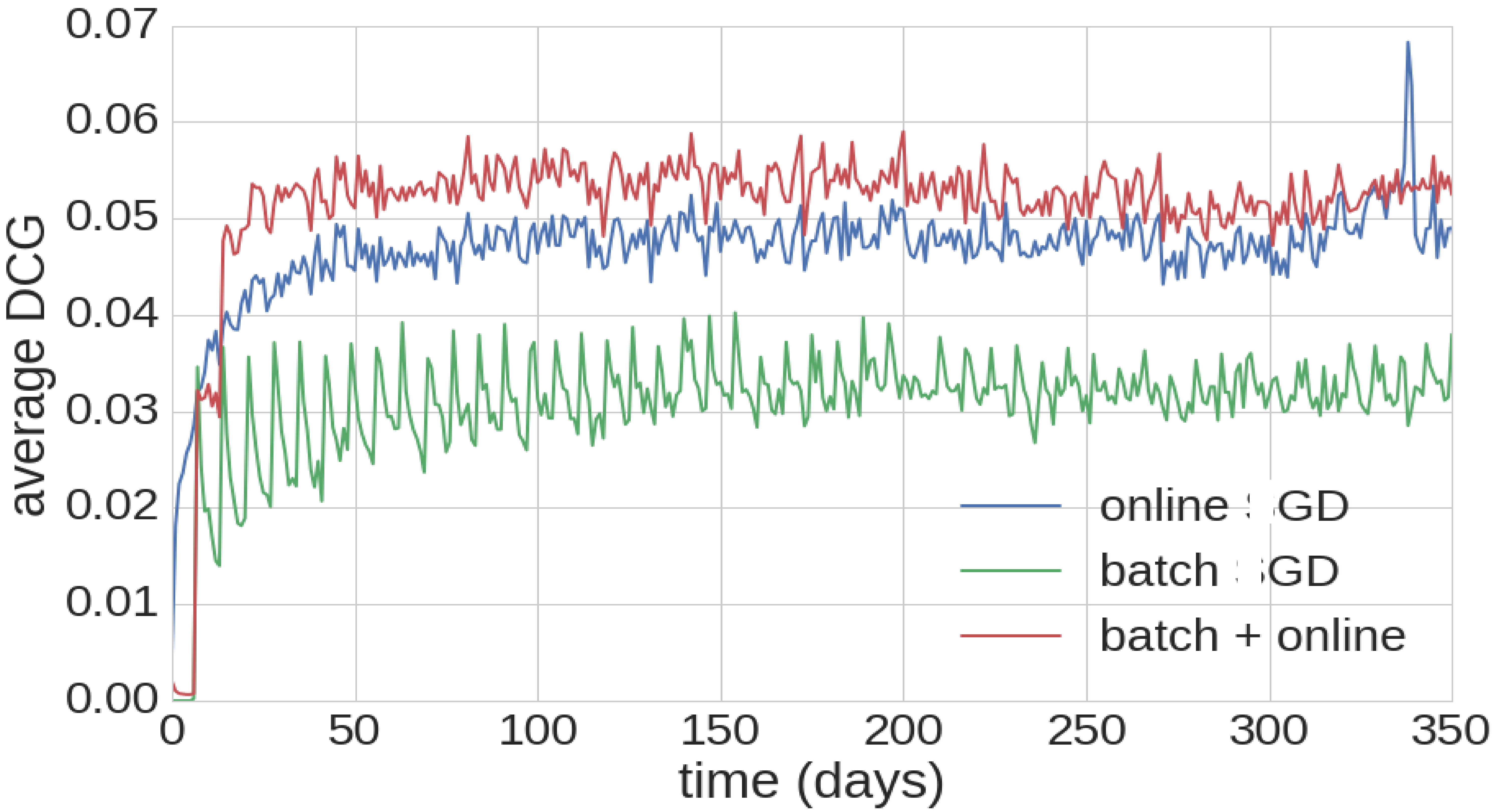


Nemzeti Kutatási, Fejlesztési És Innovációs Hivatal
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM
AZ INNOVÁCIÓ LENDÜLETE

Batch and then data streaming gradient descent recommendation

Pálovics, Benczúr: Temporal influence over the Last.fm social network. ASONAM 2013

MTA SZTAKI



- Surprisingly, reading the data only once and forgetting helps!
- Our first main observation

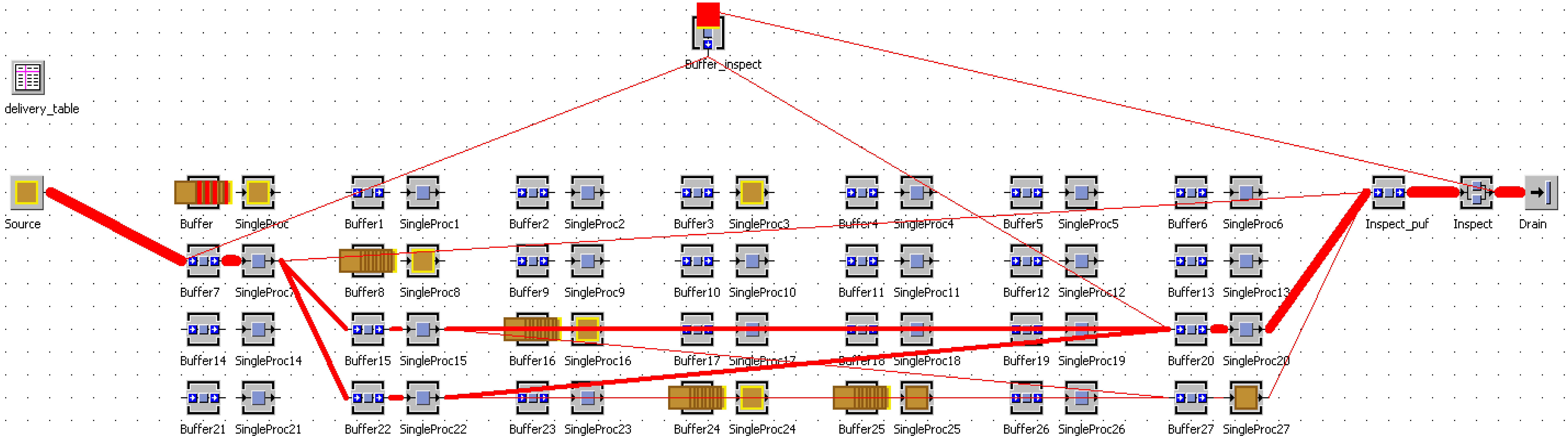
- Gradient descent is most commonly used optimization [LeCun et al. 1998, etc.]
- Natural online method [Juang et al. 1998]
- Traditional gradient descent is impractical for very large neural networks
- Downpour SGD: scalable online distributed version [Dean et al. 2011]
  - asynchronous updates
  - parameter server

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal

AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

AZ INNOVÁCIÓ LENDÜLETE

Preliminary application idea: Manufacturing lead time prediction



Material flow of product B visualized on a Sankey diagram (generated by simulator)

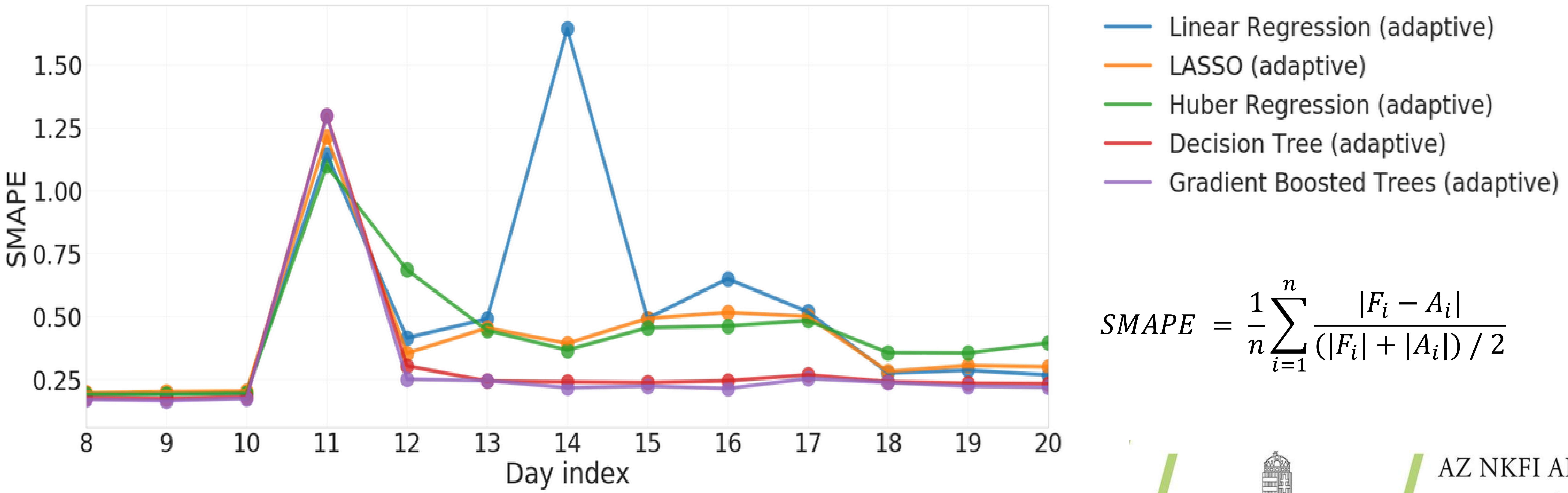
Szaller, Béres, Piller, Gyulai, Pfeiffer, Benczúr

Real-time prediction of manufacturing lead times in complex production environments

25th Conf of European Operations Management Association. 2018

Comparison of methods that only use last 7 days

- In general, Gradient Boosted Tree is the best method
- Trees adapt faster than regression



$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|F_i - A_i|}{(|F_i| + |A_i|)\ /\ 2}$$

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM
AZ INNOVÁCIÓ LENDÜLETE

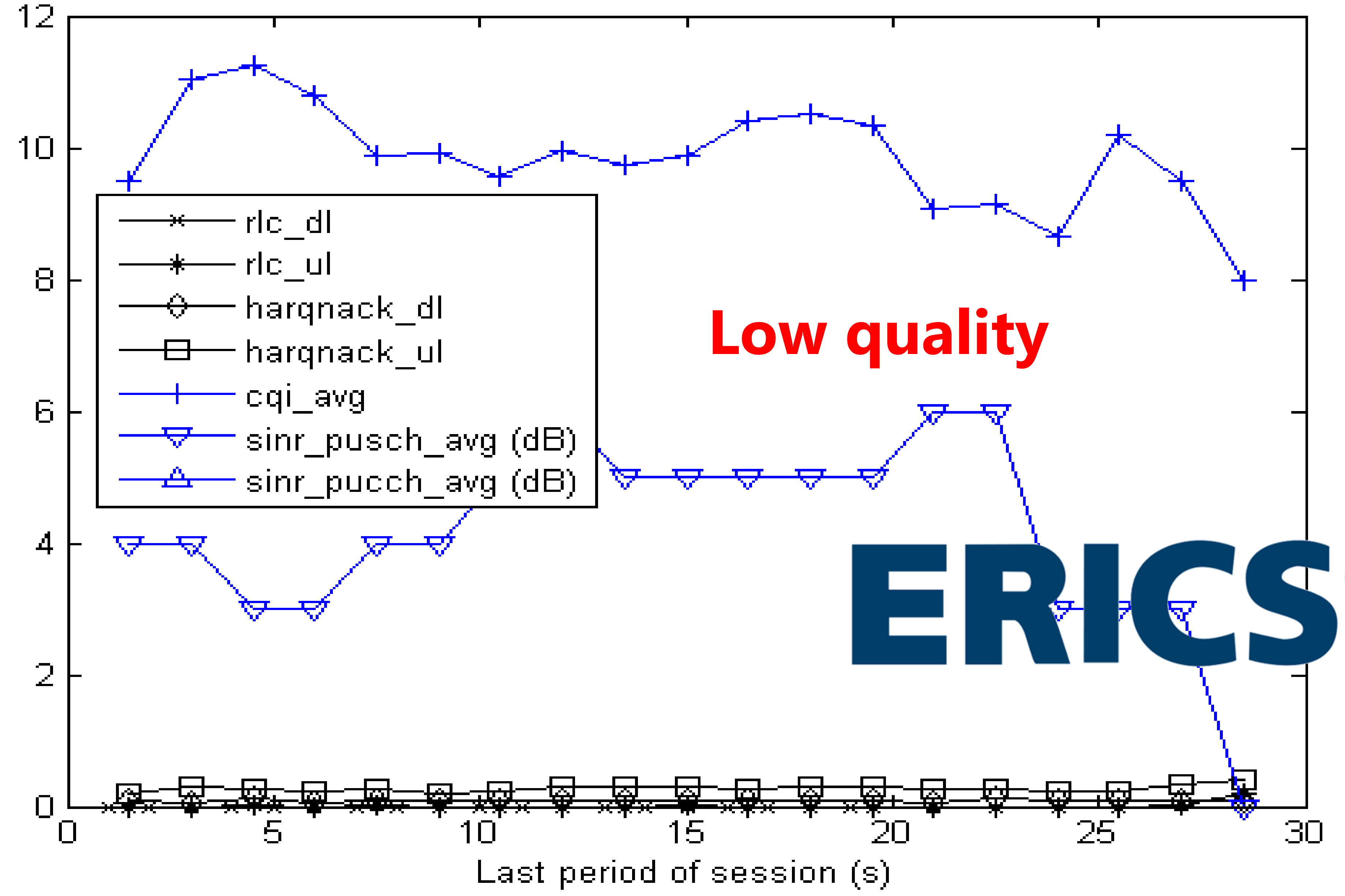
MTA SZTAKI
Hungarian Academy of Sciences
Institute for Computer Science and Control

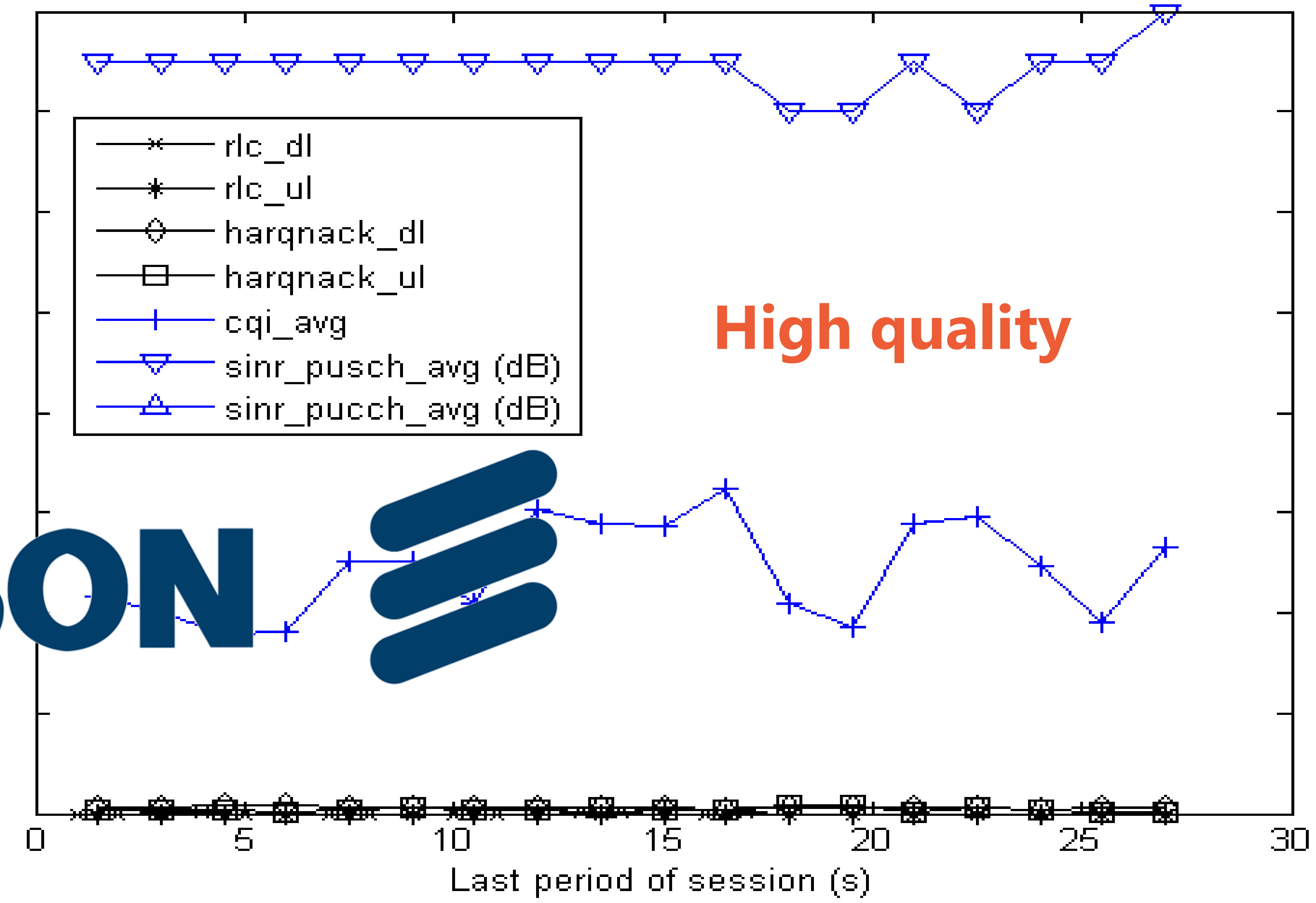
More Industrial IoT Prediction tasks

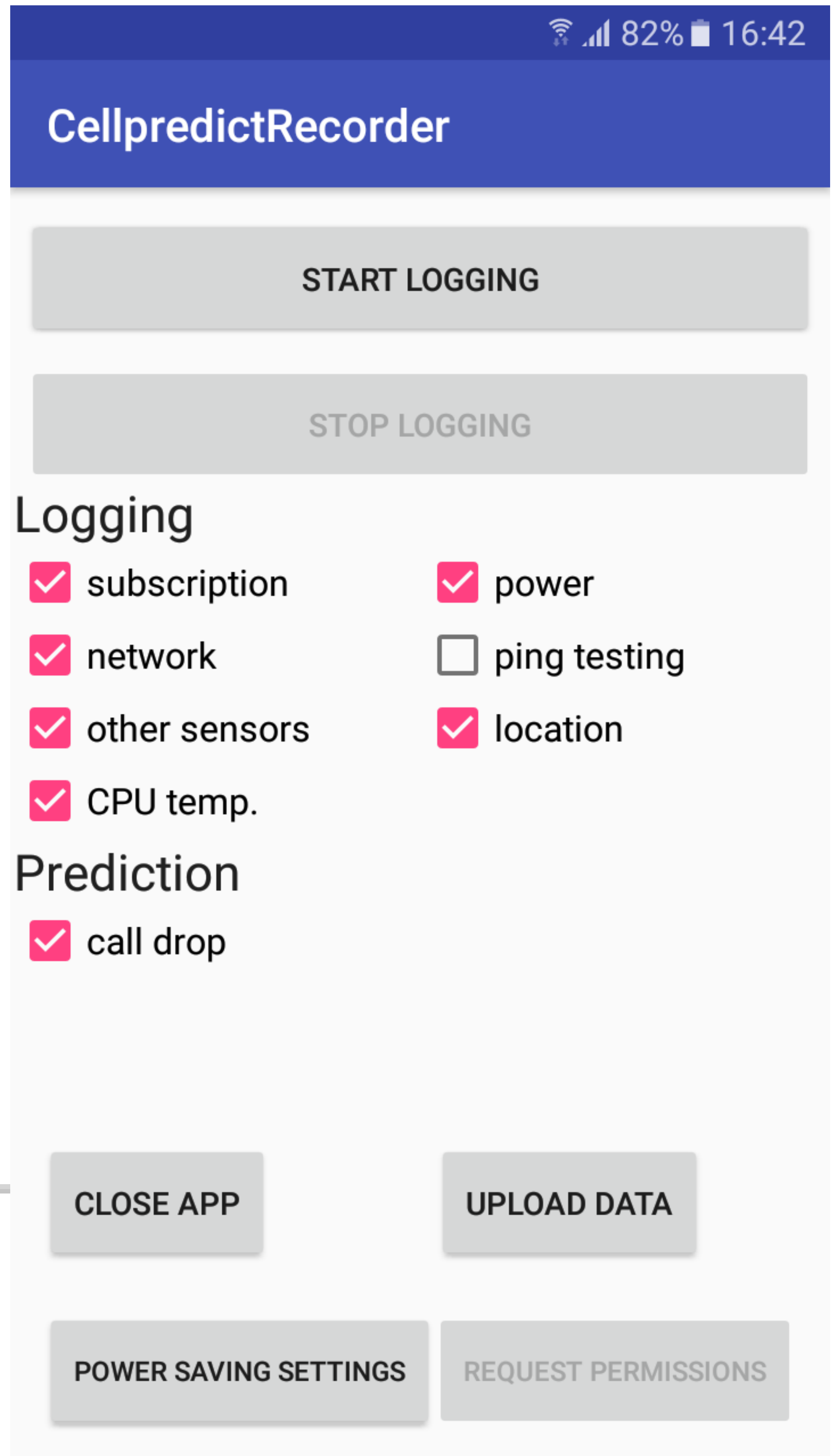
Radio connection loss
Contamination in transfer molding
Magnetron sputtering (glass coating) setpoint determination

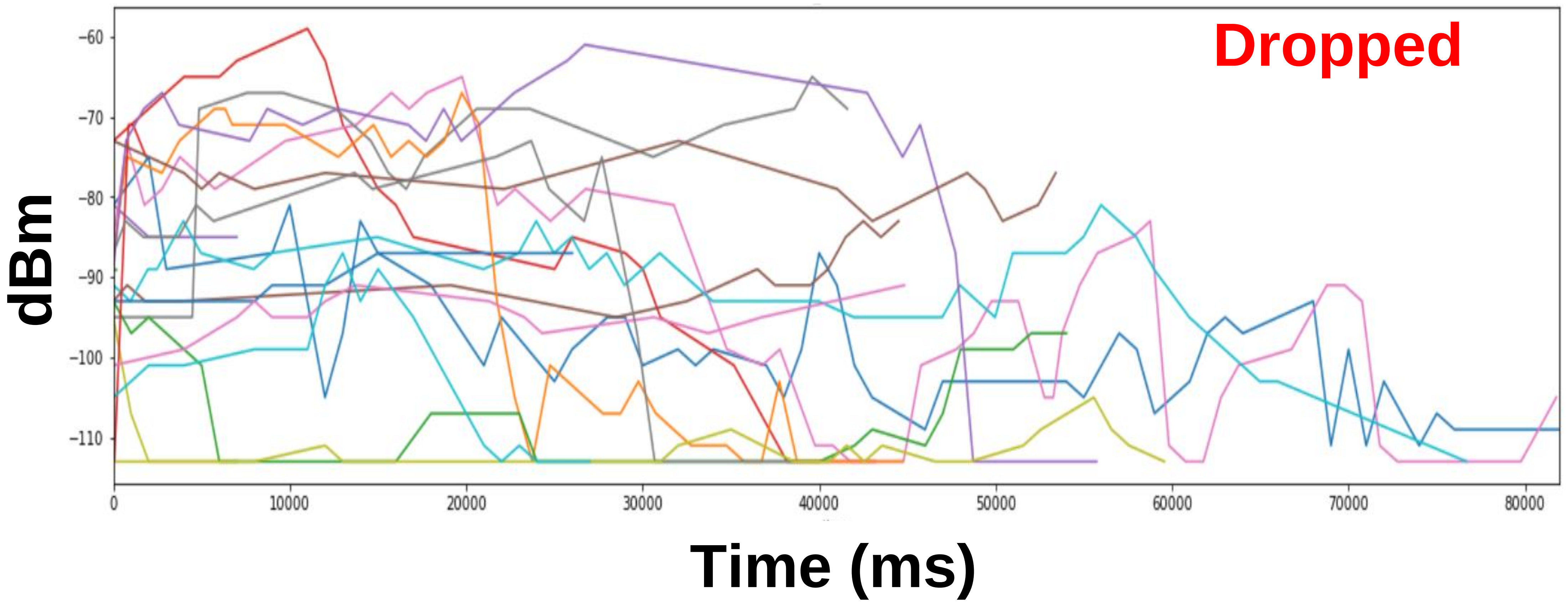
Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
Az innováció lendülete
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

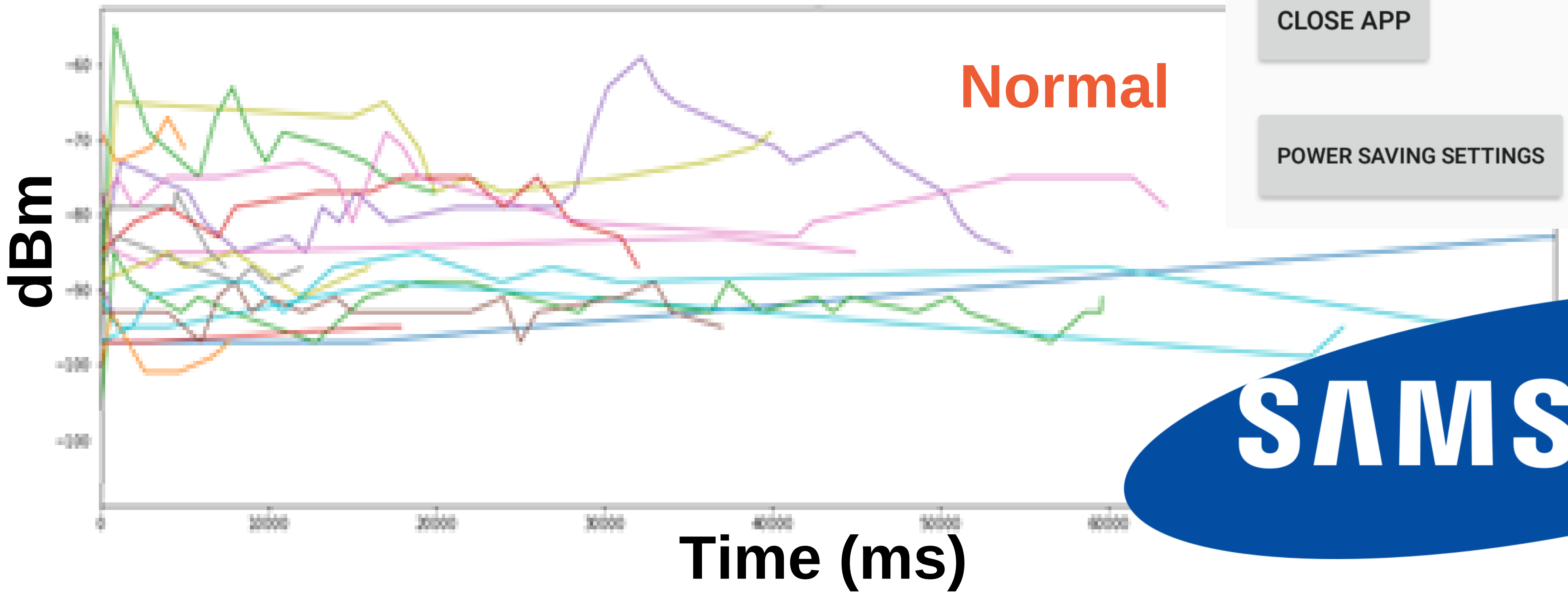
Radio connectivity - examples of low quality and loss





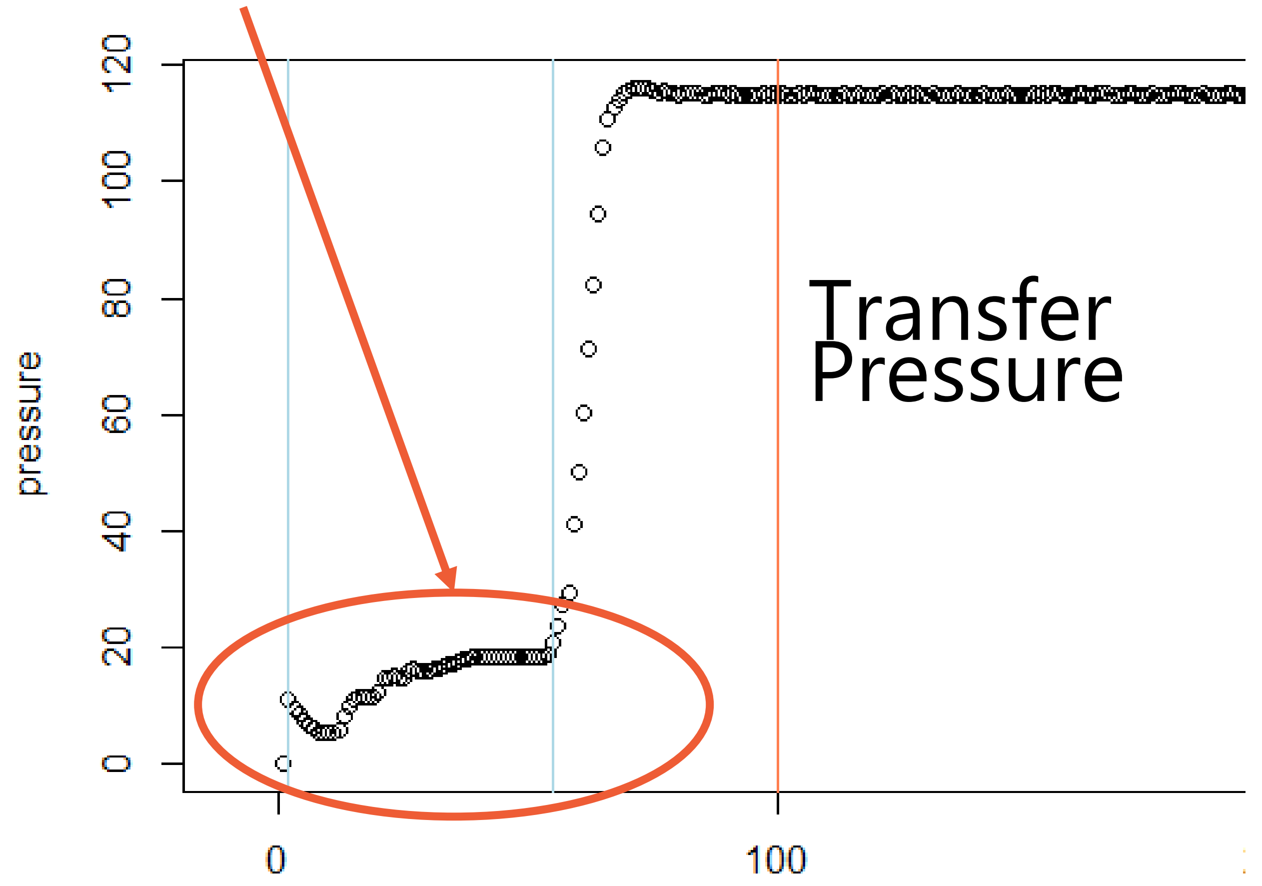




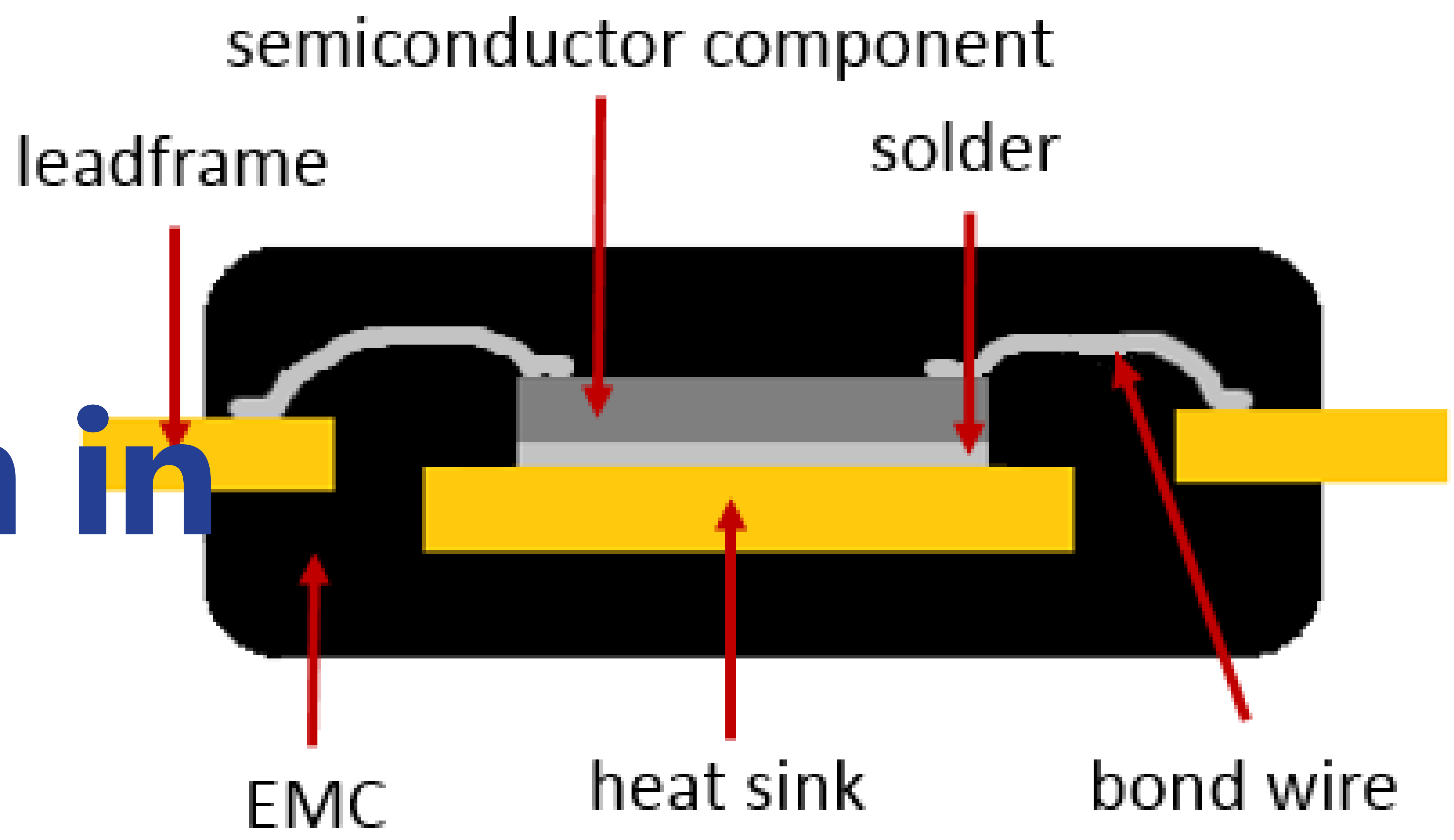


Daróczy, Bálint, Péter Vaderna, and András Benczúr. "Machine learning based session drop prediction in LTE networks and its SON aspects." *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st*. IEEE, 2015.

Filling pressure – shape indicates contamination



MTA SZTAKI
Hungarian Academy of Sciences
Institute for Computer Science and Control



Scrap rate prediction in transfer molding

BOSCH

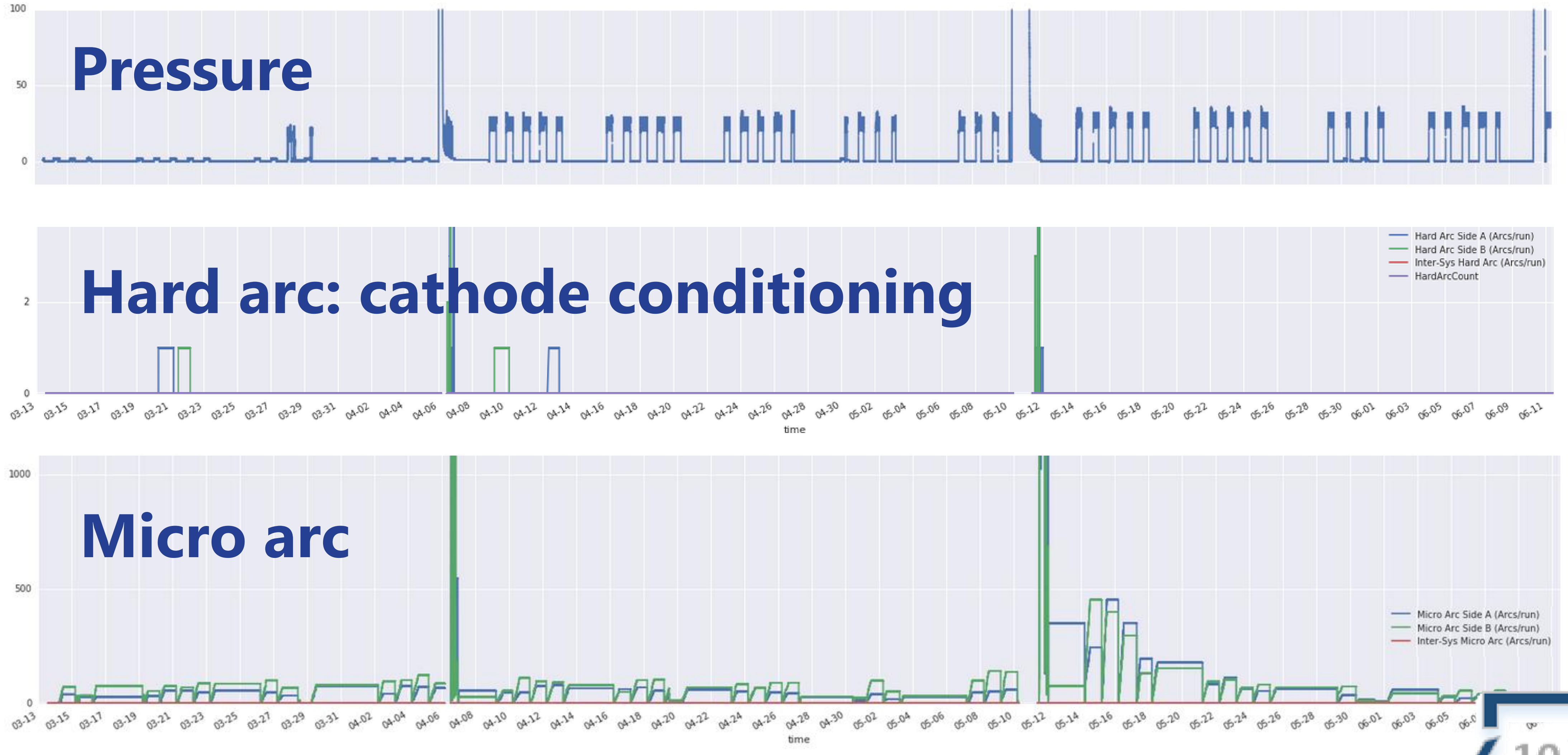
Tomorrow 16:45-17:20 Failure root-cause analysis by data analytics: concept and a case study – *László Milán Molnár*

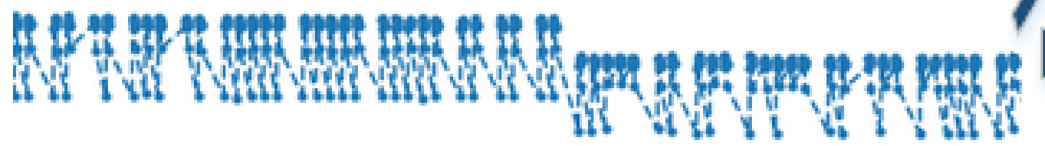
Mándli, Anna, Róbert Pálovics, Mátyás Susits, and András A. Benczúr. "Time Series Classification for Scrap Rate Prediction in Transfer Molding." 3rd SIGKDD Workshop on Mining and Learning from Time Series Held in conjunction with [KDD'17] Aug 14, 2017 - Halifax, Nova Scotia, Canada

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal
AZ INNOVÁCIÓ LENDÜLETE
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM

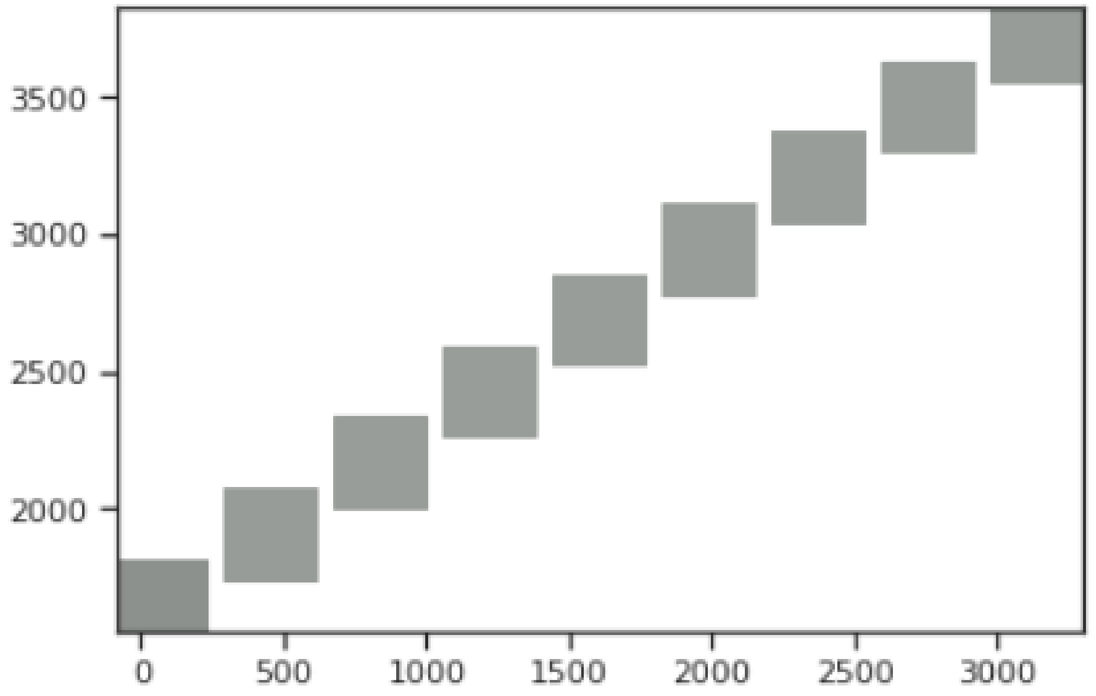
Magnetron Sputtering (glass coating technology): find the optimal power and N_2, Ar, … pressure as cathode wears out

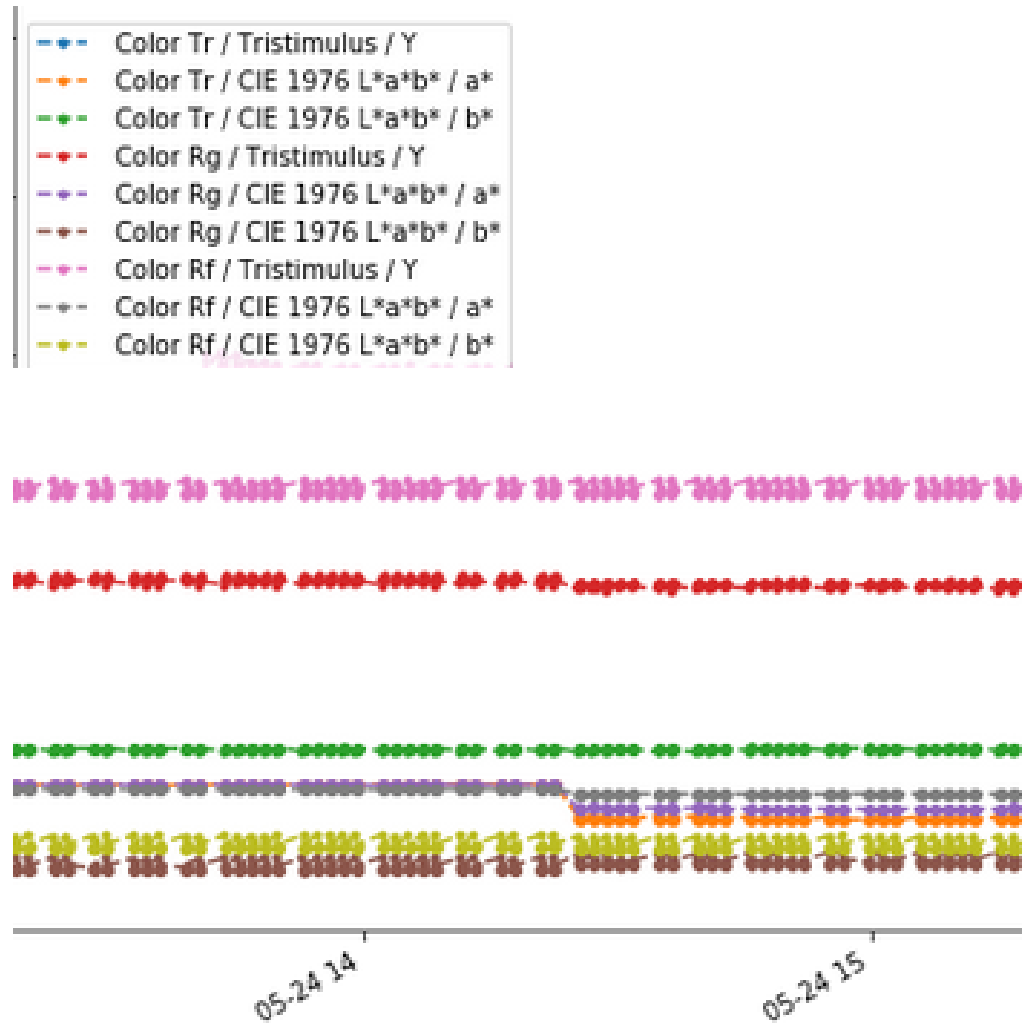
MTA SZTAKI





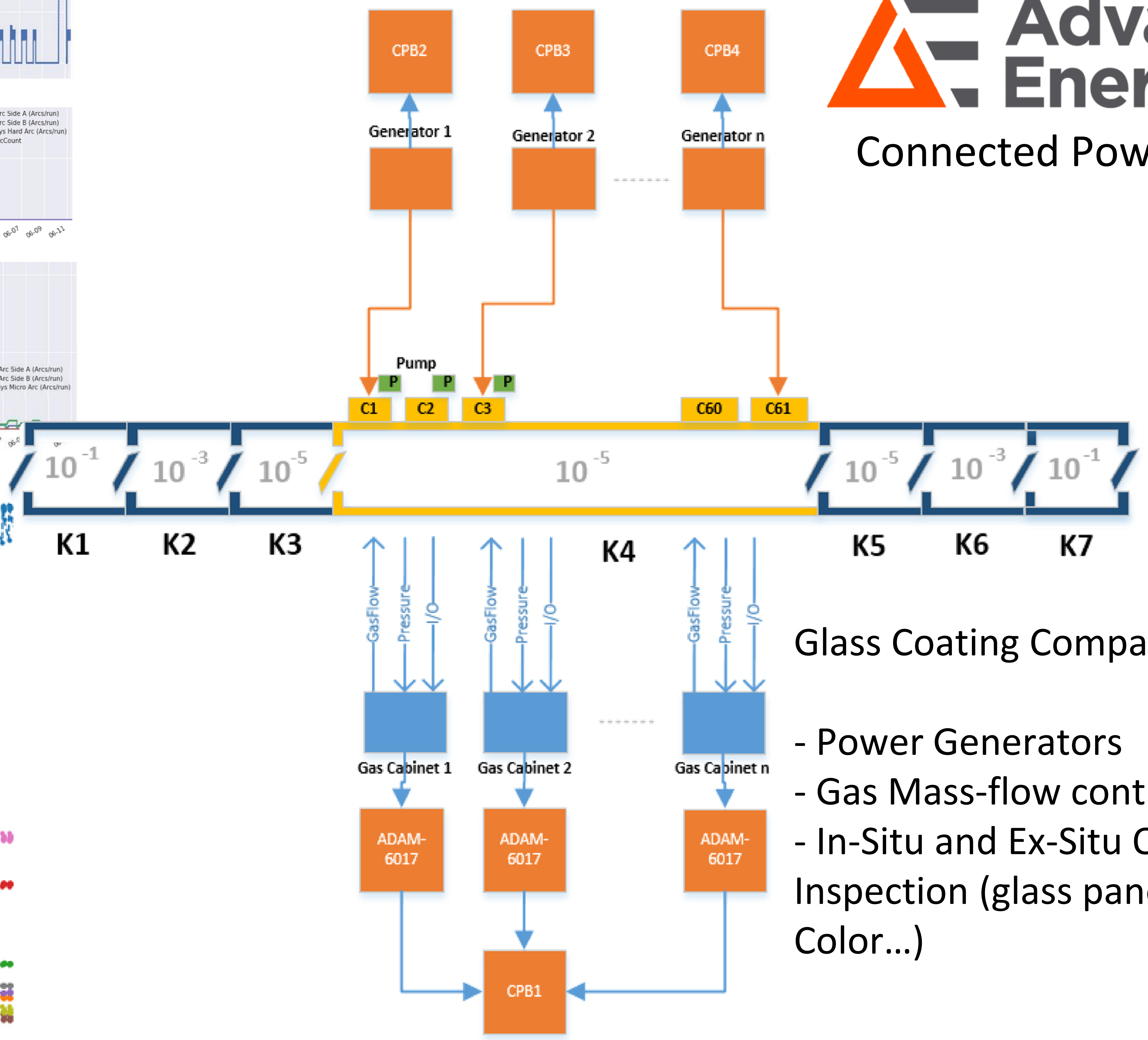
Glass panel Ex-Situ true color
(reflection and transmission, both sides)





Advanced Energy

Connected Power Boxes



Glass Coating Compartments:

- Power Generators
- Gas Mass-flow controllers
- In-Situ and Ex-Situ Quality Inspection (glass panel resistance, Color…)

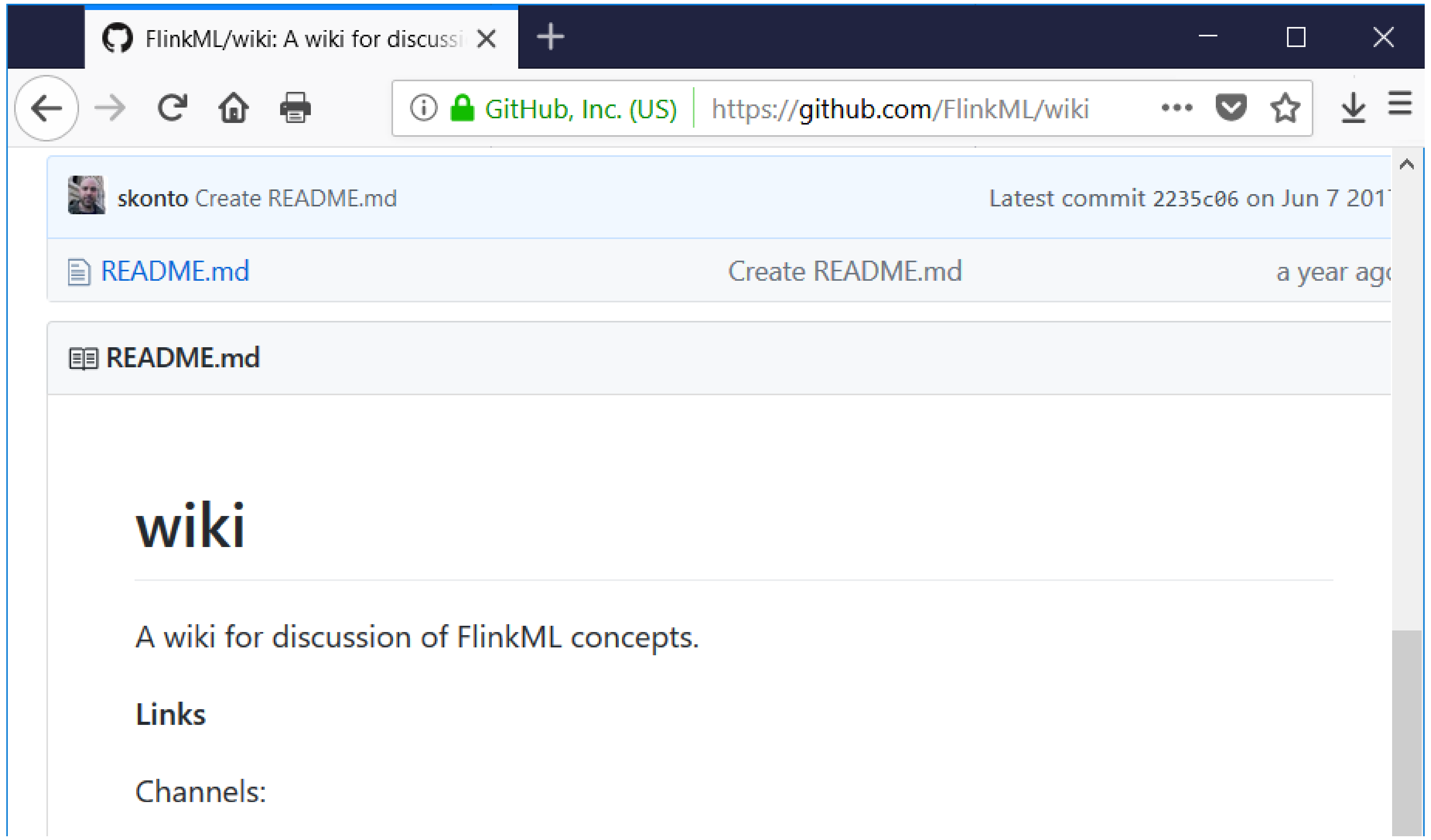
Future of Online Learning?
Future of Flink ML?

Theodore Vasiloudis Feb 21, 2017; 12:04pm

Re: [DISCUSS] Flink ML roadmap

"The idea of an online learning library for Flink has been broached before, and this semester I have one Master student working on exactly that. From my conversations with people in the industry however, almost nobody uses online learning in production, at best models are updated every 5 minutes. So the impact would probably not be very large."

- Will practitioners have really big training tasks? Or only model serving tasks?
- Will real recommender systems ever need distributed matrix factorization, or will they always work either multicore only, or use session-based methods?
- Will we find more convincing use cases?



MTA **SZTAKI**
Hungarian Academy of Sciences
Institute for Computer Science and Control

Questions?

András Benczúr,
head, Informatics Lab

Nemzeti Kutatási, Fejlesztési És Innovációs Hivatal
AZ INNOVÁCIÓ LENDÜLETE
AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM