UNIVERSITY OF WARSAW

WYDZIAŁ MATEMATYKI, INFORMATYKI I MECHANIKI

MIM

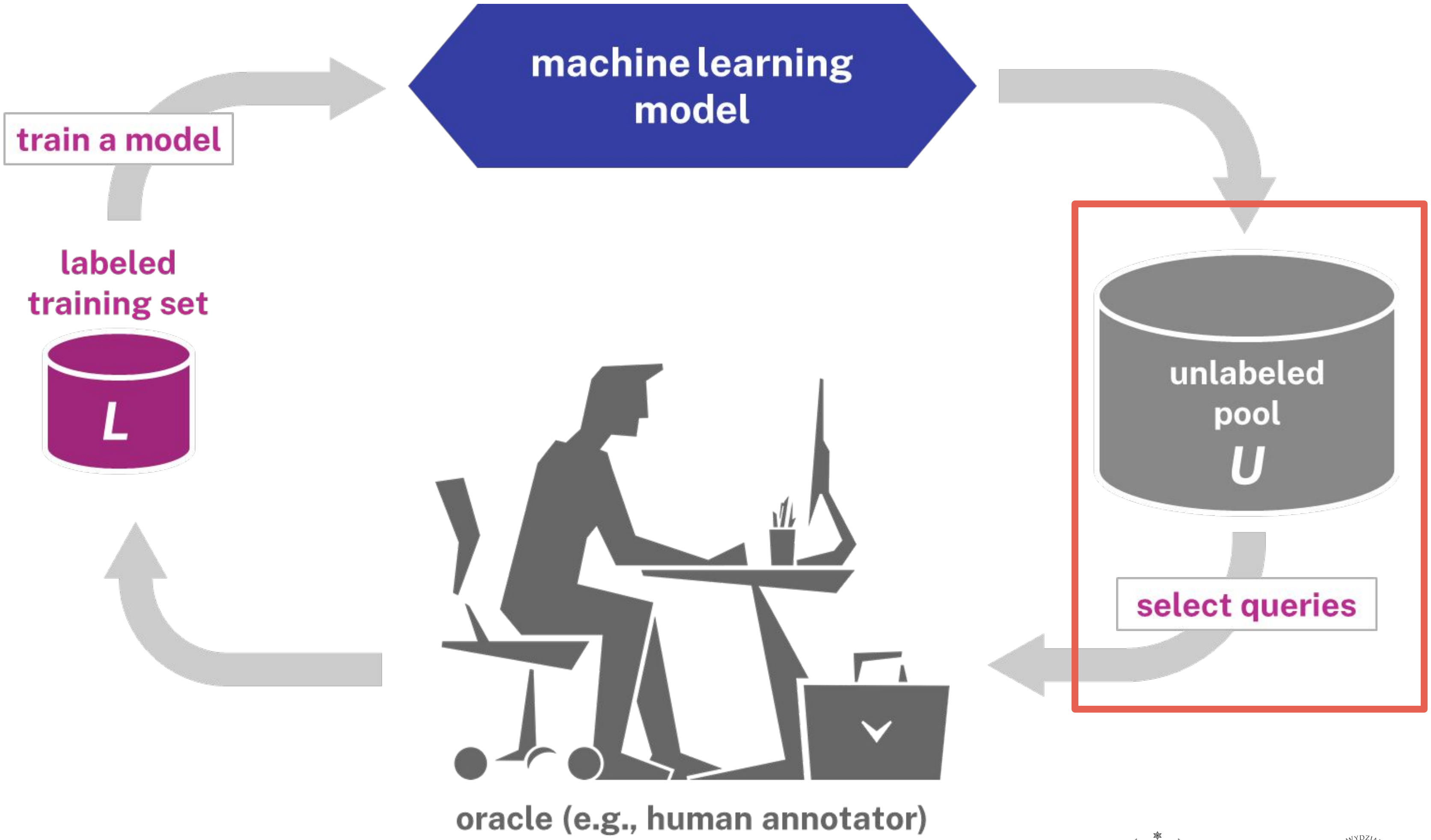
Active Learning - the pool-based selective sampling (part 2)

Andrzej Janusz
Daniel Kałuża

THE PLAN

- A recap of the previous lecture.
- Representativeness.
- Batch diversity.
- Selection of the initial batch.
- Evaluation of active learning results.
- Exemplary algorithms and use-cases.
- Summary.

The active learning cycle - revisited



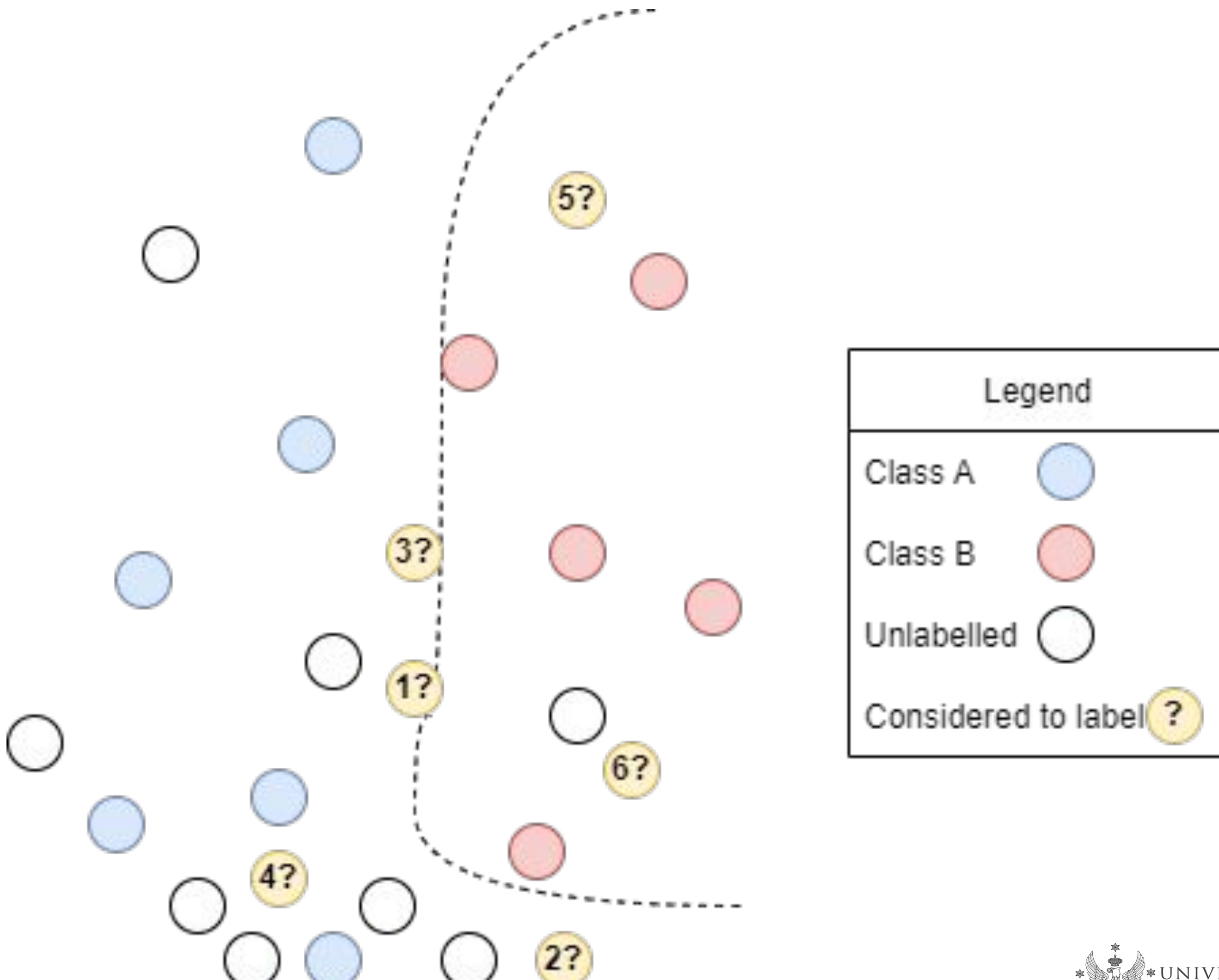
Active Learning as an optimization task

Formal task definition - we search for $U^* \subset DP$ such that:

$$U^* = \underset{U:|U|=K}{\arg\max}\ \mathbb{E}_{(X,Y)}[q(Y, f^U(X))]$$

where f^U is a model trained on a subset $U \subset DP$ whose size is *K* and *q* is a predefined quality metric.

Informativeness and uncertainty



UNIVERSITY OF WARSAW

Sample representativeness

- Not all ‘*uncertain*’ samples are equally informative
  - Learning from outliers may actually hinder the prediction performance.
  - We may want to prioritize learning from more probable inputs.
  - Labels for common examples might be more reliable or less expensive to obtain…
- Again, all we need is a good measure :-)
  - How can we measure the sample representativeness?
  - The efficiency is a serious consideration.
  - Alternatively, outlier filtering/clustering-based selection is also an option…

Representativeness measures

- Average sample-to-pool similarity:

$$R(u) = \frac{1}{|U|} \sum_{u' \in U} Sim(u, u')$$

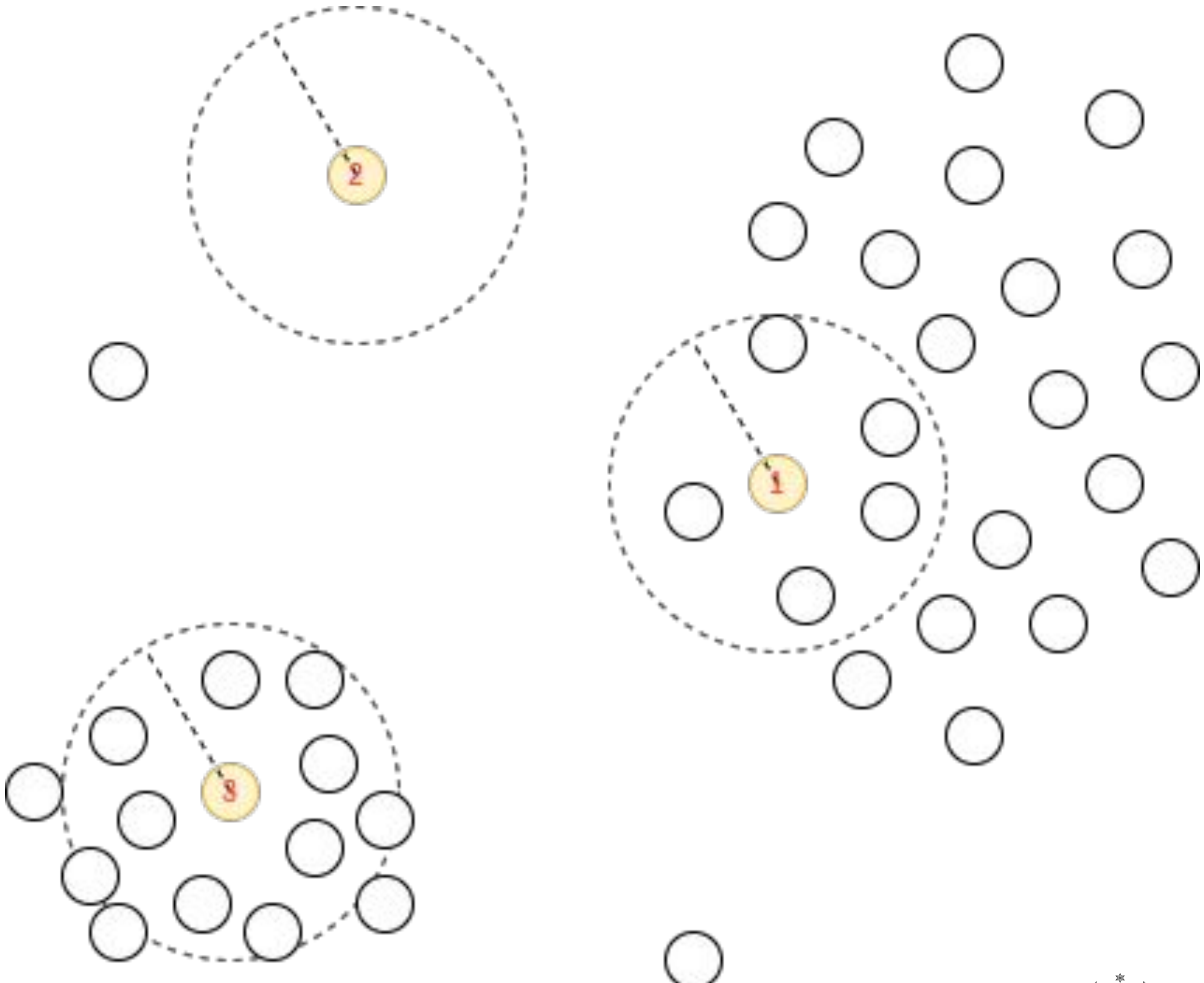
- Sum of similarities to K nearest neighbors:

$$R(u) = \sum_{u' \in NN_K(u)} Sim(u, u')$$

- Similarity to the corresponding cluster - *C(u)* - center (or the cluster medoid):

$$R(u) = Sim(u, \frac{1}{|C(u)|} \sum_{u' \in C(u)} u')$$

Sample representativeness - analysis



UNIVERSITY OF WARSAW

Uncertainty and representativeness

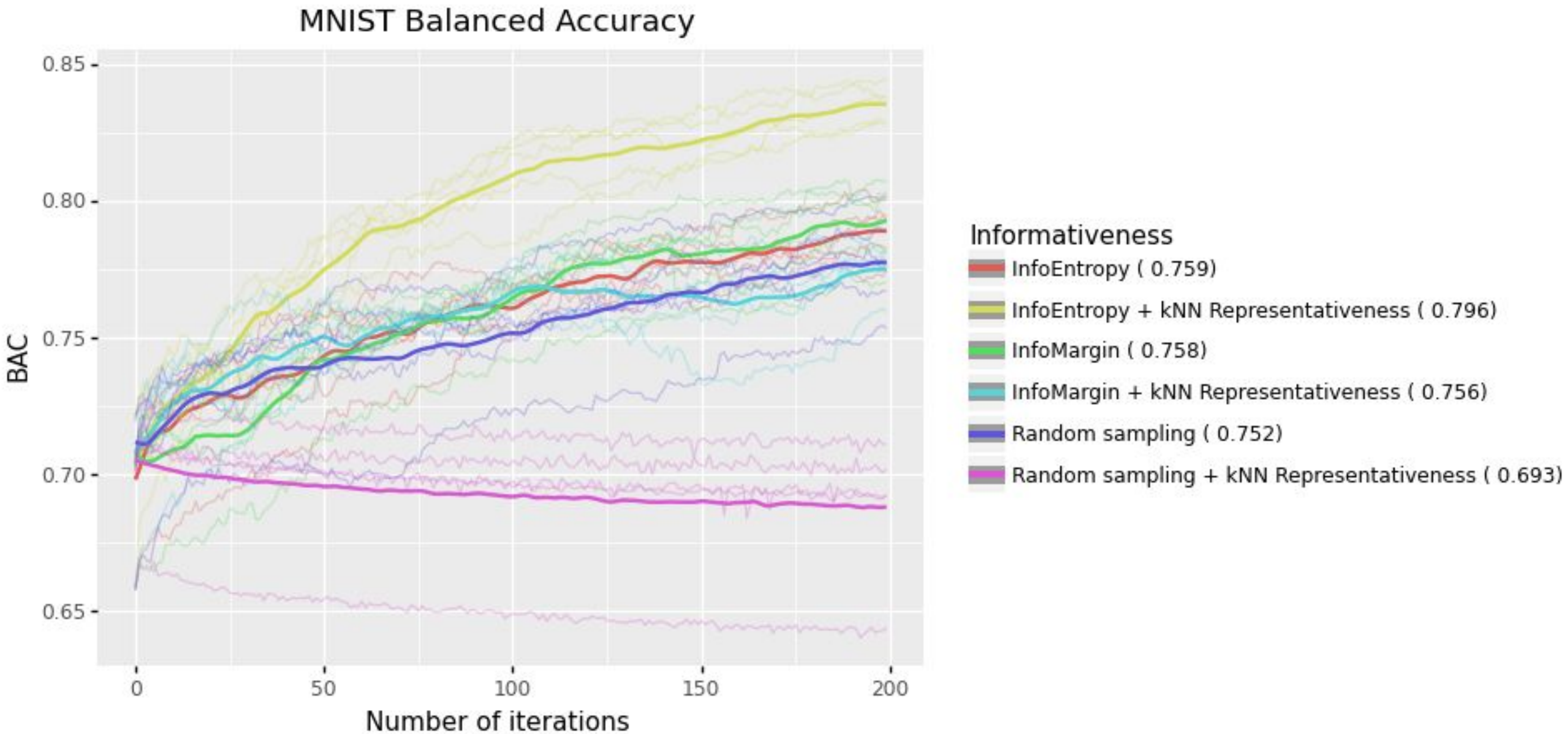
We have several options for combining uncertainty with representativeness:

- Select *m* most uncertain samples, then select the most representative one among them.
- Combine the uncertainty and representativeness into a single measure of informativeness, e.g.:

$$Info(u) = Unc(u) \times R(u)^{\alpha}$$

  - Query the sample with the highest informativeness.
- Other parametrizations of the informativeness function are possible.

Impact of the use of representativeness



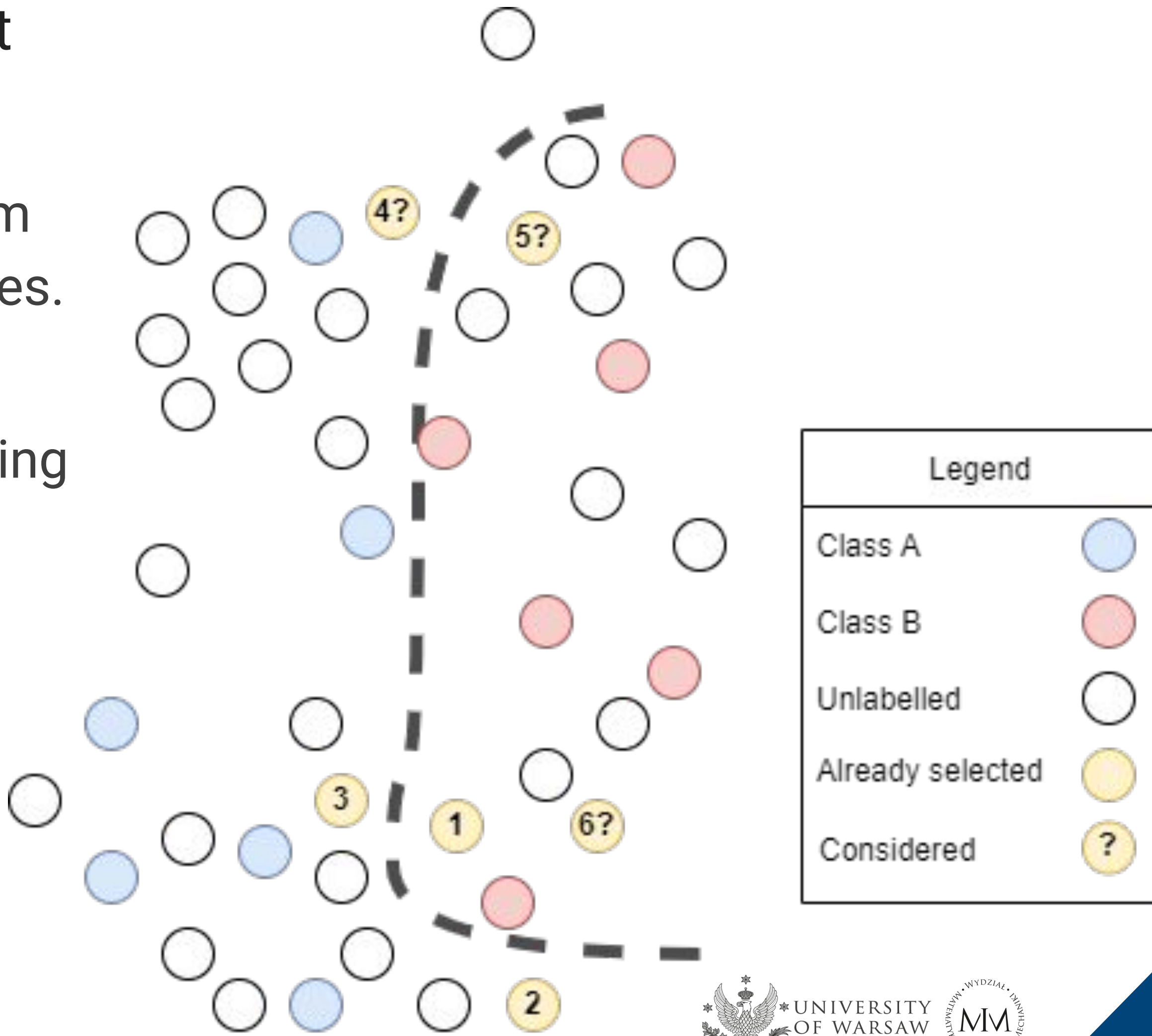
- An experiment on the MNIST data set.
  - A logistic regression learner.
  - Initial training data size is 0.33%, 200 iterations.
  - Various impact of the representativeness.

Time-constrained active learning?

- Updating the model and making predictions after each label query is time-consuming.
  - We don't want to waste the time of experts.
  - Several experts may work concurrently - we want to provide samples for labeling to all of them.
  - Experts work at different paces.
- It makes sense to buffer queries to the oracle.
  - At each iteration, we need to select a batch of queries.
  - A general rule - the smaller batch, the better (but the practical constraints have priority).

Active batch selection

- We don’t want to select similar queries.
  - Diminishing profits from labels of similar samples.
  - Experts may get irritated/bored by labeling similar cases.



Batch selection heuristics

- The greedy selection approach.
  - The informativeness function may incorporate the dissimilarity of samples selected for the batch:

$$Info(u, B) = Unc(u) \times \left(\frac{1}{|B|} \sum_{u' \in B} Dis(u, u') \right)^{\beta}$$

  - Set *B* contains samples that have been already selected for the current batch.
- Different optimization methods.
  - Use the genetic algorithm to select a subset of *q* most dissimilar queries among *m* most “uncertain” cases…
  - or directly optimize the above informativeness function.
  - Use unsupervised learning.

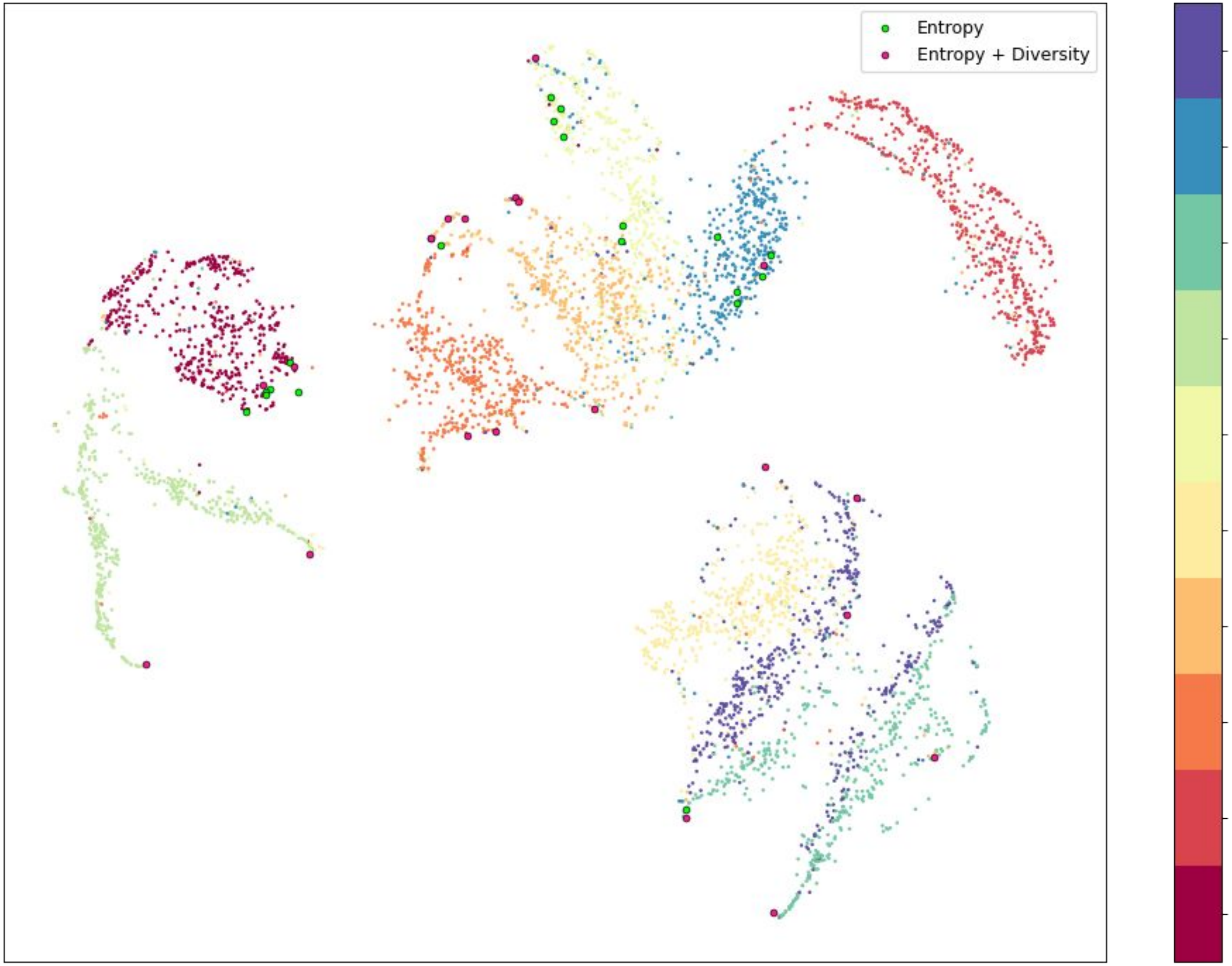
Batch selection - an alternative approach

- We may add random noise to our informativeness function:

$$Info(u) = Unc(u) + \beta \cdot Rand()$$

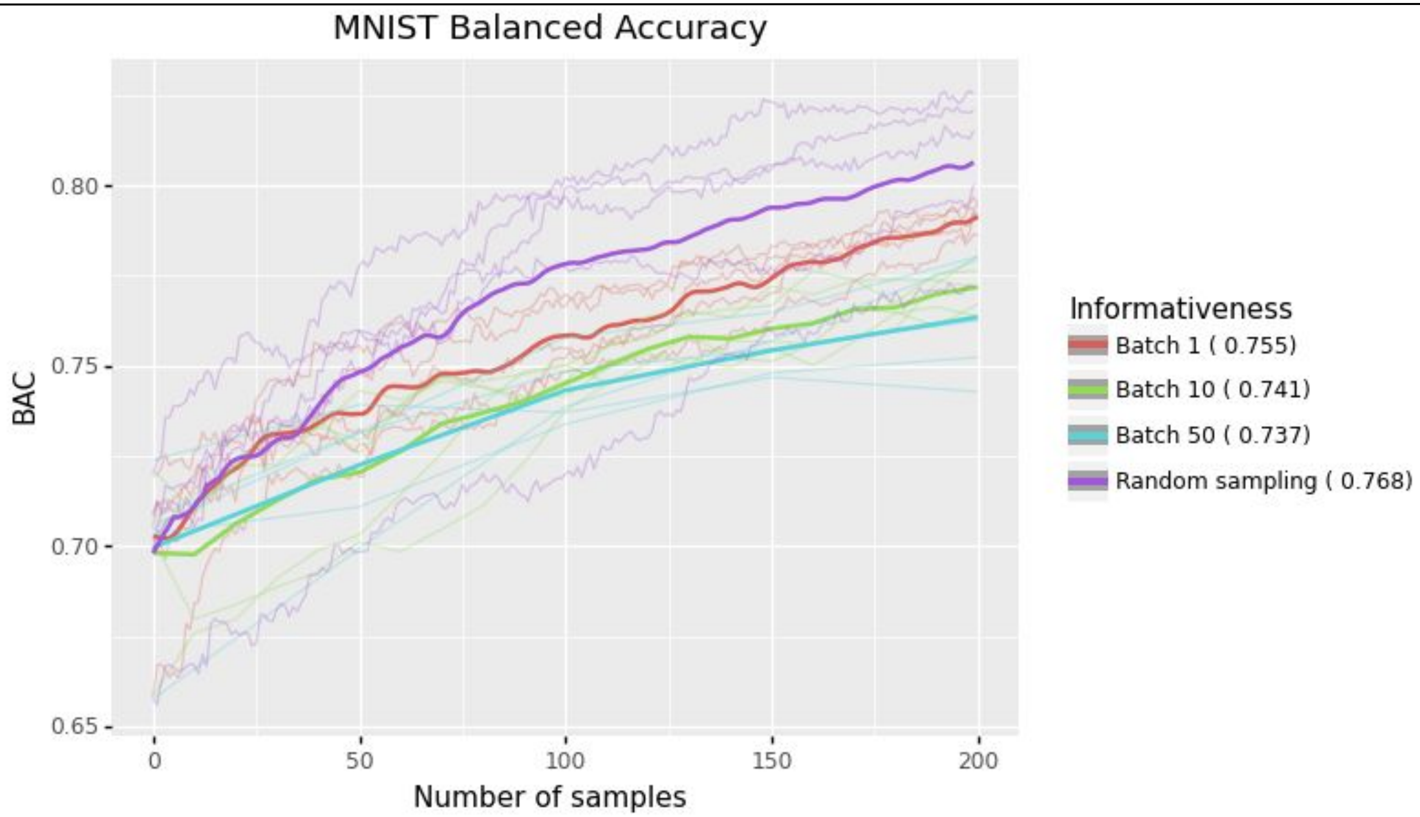
- Select *q* most informative samples (according to *Info(u)* defined above).
- The randomization diversifies selected queries.
  - But it may deteriorate the performance as well.
  - The efficiency is the obvious advantage.

Impact of the batch diversification



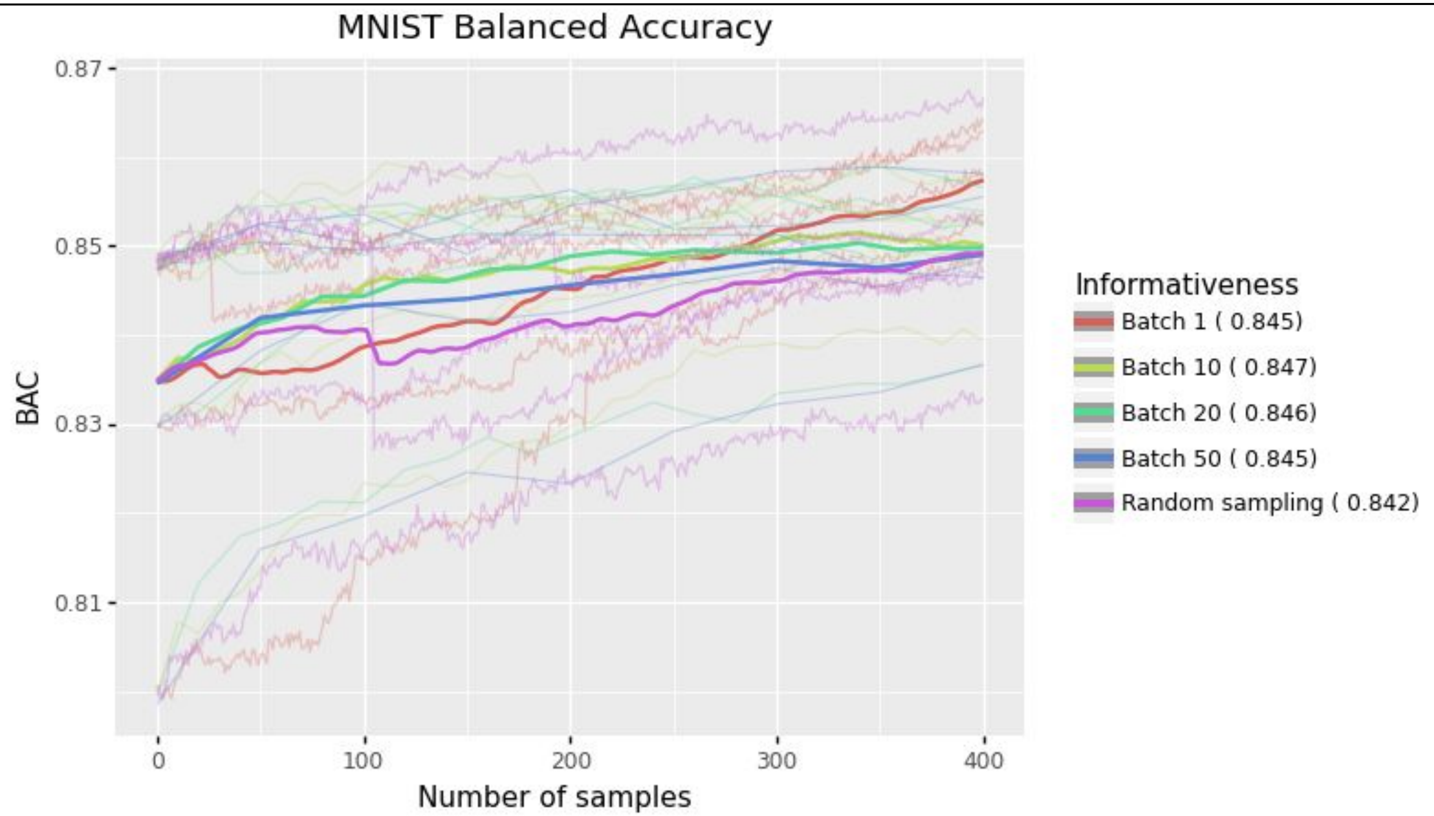
- UMAP visualization of the MNIST data set.
- Colors correspond to classes, the green and red dots represent queries selected using the entropy and the entropy+diversity sampling.
- Batch size = 20.

Impact of the batch size



- The MNIST data set again.
  - A logistic regression learner.
  - Initial training data size is 0.33%, 200 iterations.
  - The smaller the batch size is the better.

Impact of the batch size



- The MNIST data set - with larger initial batch sizes.
  - A logistic regression learner.
  - Initial training data size is 1%, 400 iterations.
  - Much smaller differences between the results.

Representativeness and diversity scaling

- If we want to combine the uncertainty, representativeness, and batch diversity, we need to make them comparable.
- Similarity/dissimilarity and uncertainty usually have different scales.
  - Linear scaling.
  - Rank scaling.
- In the end, we may combine all factors into a formula:

$$Info(u,B) = \frac{1}{c} \cdot Unc(u) \times \left(\frac{1}{r} \cdot R(u)\right)^{\alpha} \times \left(\frac{1}{d} \cdot Dis(u,B)\right)^{\beta}$$

- *c, r, d* above are scaling constants.

Selection of the initial batch

- The “chicken and egg” problem - how can we select the first batch of queries?
  - “*At random*” seems the obvious answer - but can we do any better?
  - It might be the cause of training instability and inconsistency of AL results.
- Alternative heuristics:
  - Iterative sampling.
  - Clustering-based sampling.
  - A hybrid approach.
- But we still need some samples for evaluation!

Iterative sampling heuristic

- We may rule out the uncertainty part from our informativeness function (or assume it's constant):

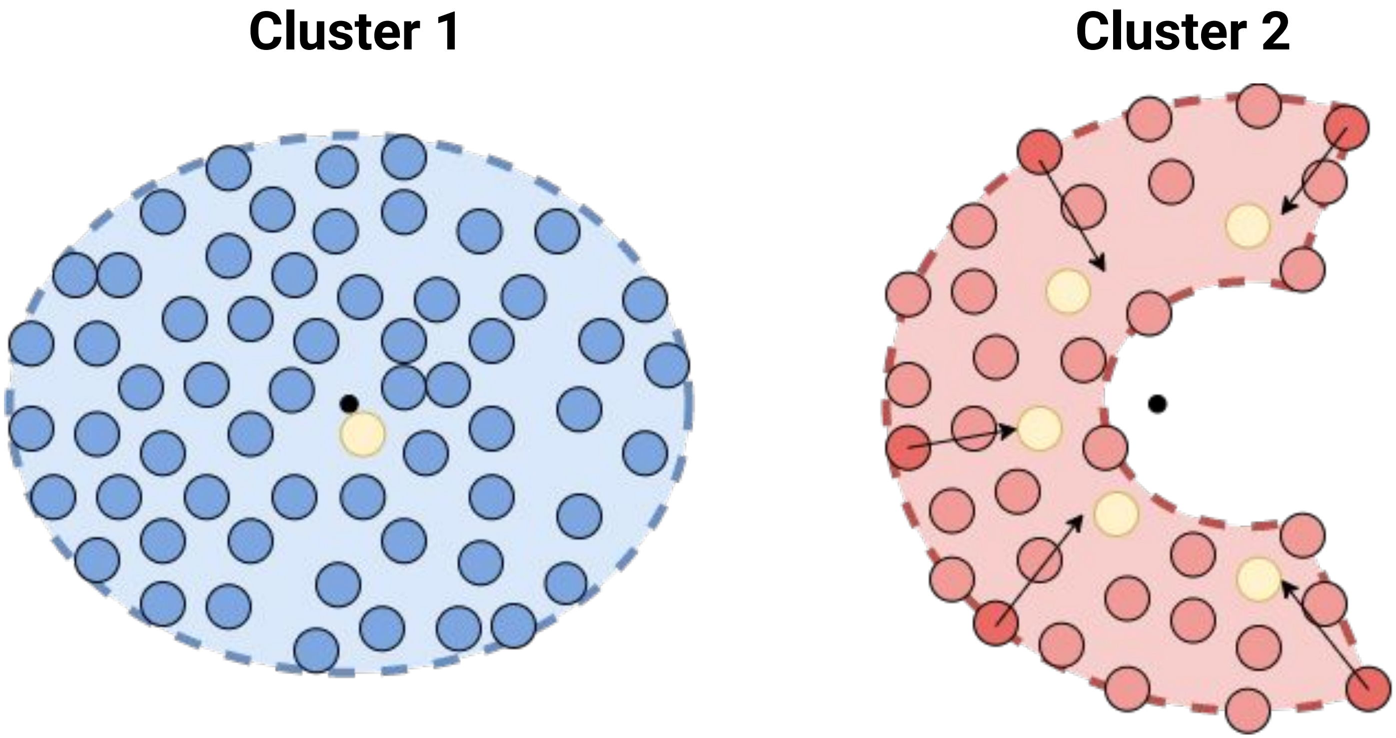
$$Info(u, B) = \left(\frac{1}{r} \cdot R(u)\right)^{\alpha} \times \left(\frac{1}{d} \cdot Dis(u, B)\right)^{\beta}$$

- We choose the initial batch as if it was any other iteration of the AL cycle.
  - Simplistic and general approach.
  - Consistent with later steps - if we use the "uncertainty-representativeness-diversity" approach.

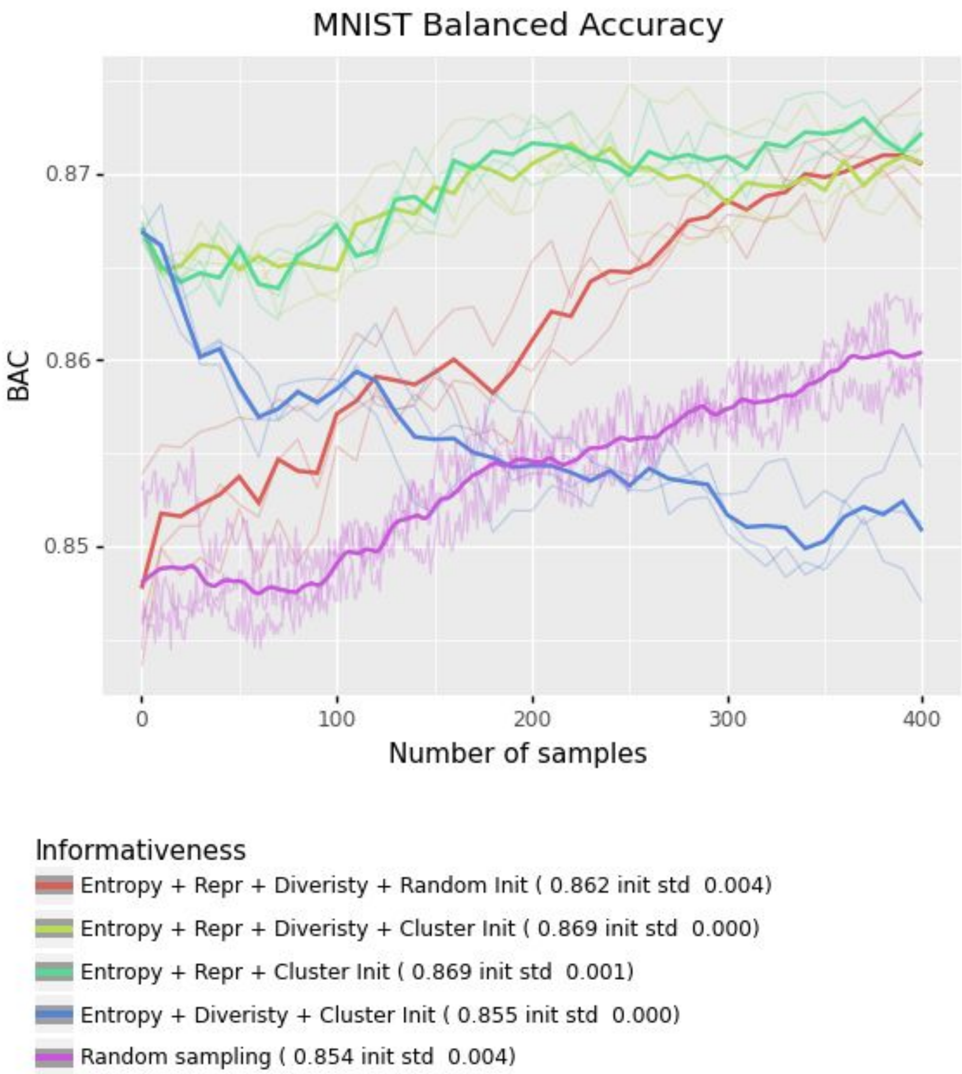
Clustering-based initial batch selection

- Available data pool can be divided into a number of subsets using a clustering algorithm.
  - Various algorithms can be used as long as they can process the data pool efficiently.
  - The number of clusters may depend on the required initial batch size.
- We independently sample the initial queries from each discovered data cluster.
  - We can select samples nearest cluster centers.
    - It may not work if clusters have irregular shapes.
- We aim to select representative, yet diverse cluster members - how can we do it?

Selection of cluster representatives



An example



- MNIST data with and without initial data batch smart-sampling.
- K-means clustering and near-center sampling used for the data batch selection.
- Informativeness = entropy + NN-representativeness + batch diversity.
- Clustering reduces the standard deviation of results.

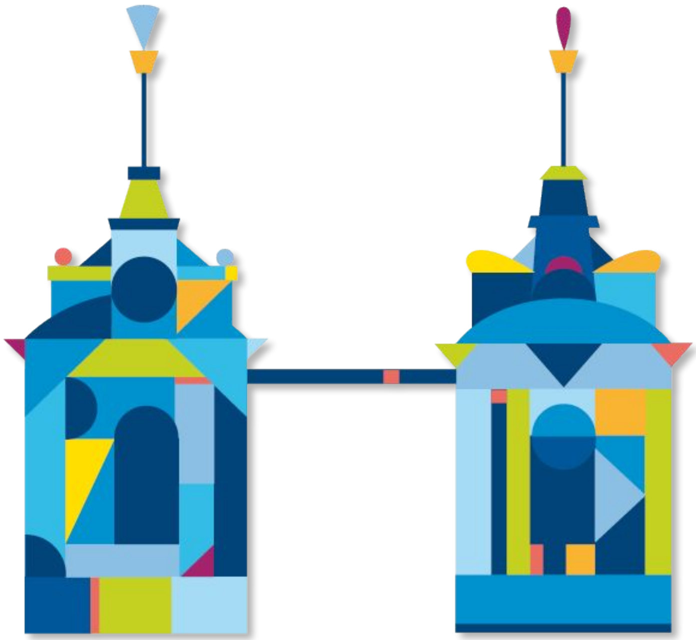
UNIVERSITY OF WARSAW

Summary

- We discussed selected methods of choosing queries in active learning.
- We considered the representativeness of samples and discuss its impact on the informativeness.
- We talked about selecting batches of queries - why is it needed and how can we assure it.
- We considered the problem of selecting the initial data batch.
- We analyzed an example in which we experimented on MNIST benchmark data set.

Literature:

1. B. Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, (2010).
2. J. Bosser, E. Sorstadius and M. Chehreghani, "Model-Centric and Data-Centric Aspects of Active Learning for Deep Neural Networks," in 2021 IEEE International Conference on Big Data (IEEE BigData), Orlando, FL, USA, 2021 pp. 5053-5062.
3. H. Yi et al., "AI Tool with Active Learning for Detection of Rural Roadside Safety Features," 2021 IEEE International Conference on Big Data (IEEE BigData), 2021, pp. 5317-5326
4. A. Janusz, Ł. Grad, M. Grzegorowski: Clash Royale Challenge: How to Select Training Decks for Win-rate Prediction. FedCSIS 2019: 3-6, (2019).
5. J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal: Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. ICLR 2020, (2020).
6. J. Azimi, A. Fern, X. Z. Fern, G. Borradaile, B. Heeringa: Batch Active Learning via Coordinated Matching. ICML 2012, (2012).
7. G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, S. Kumar: Batch Active Learning at Scale. 35th Conference on Neural Information Processing Systems (NeurIPS 2021), (2021).
8. W. Yuan, Y. Han, D. Guan, S. Lee, Y.-K. Lee: Initial training data selection for active learning. ICUIMC 2011: 5, (2011).



QUESTIONS OR COMMENTS?

a.janusz@mimuw.edu.pl

or

d.kaluza@mimuw.edu.pl

UNIVERSITY OF WARSAW