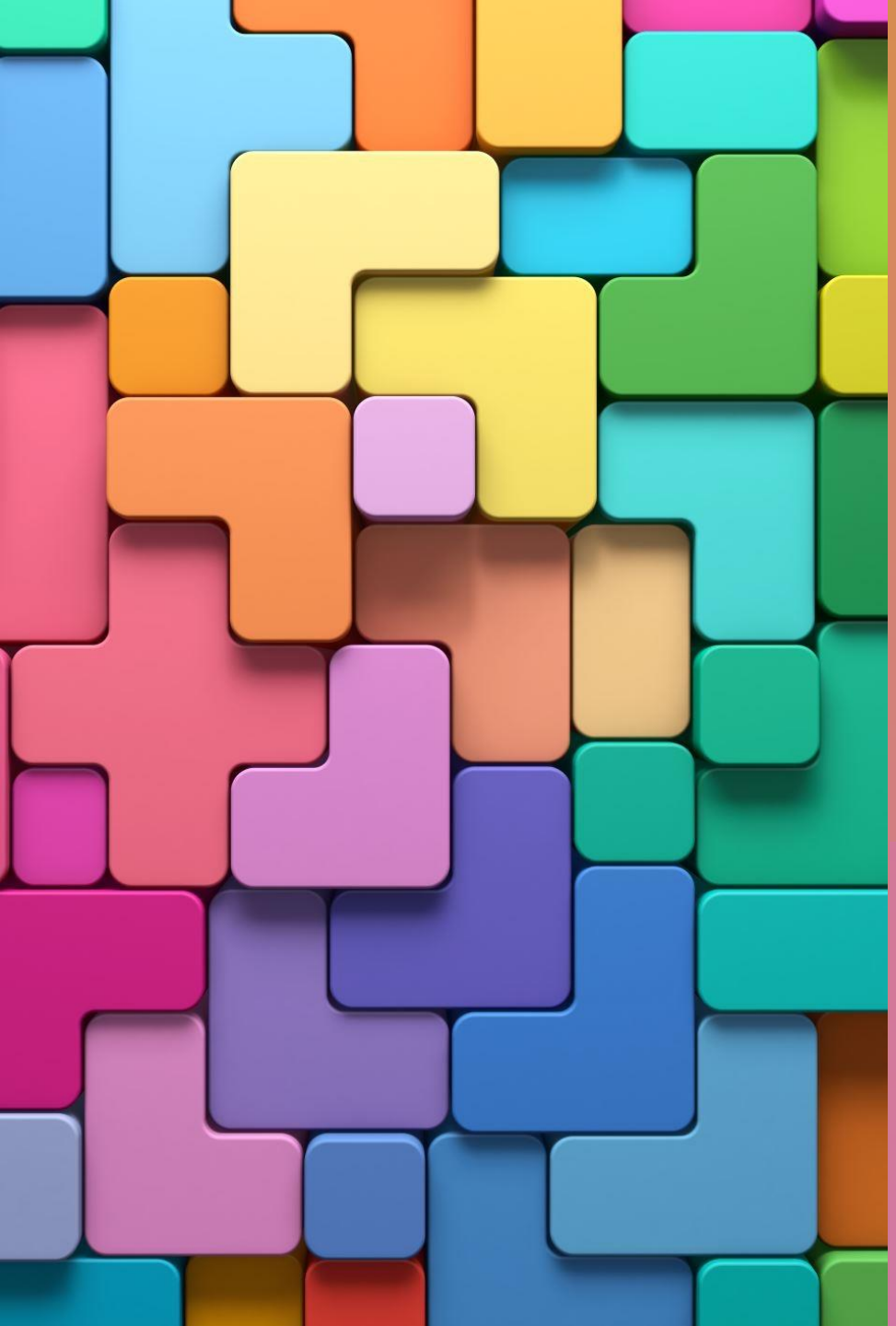
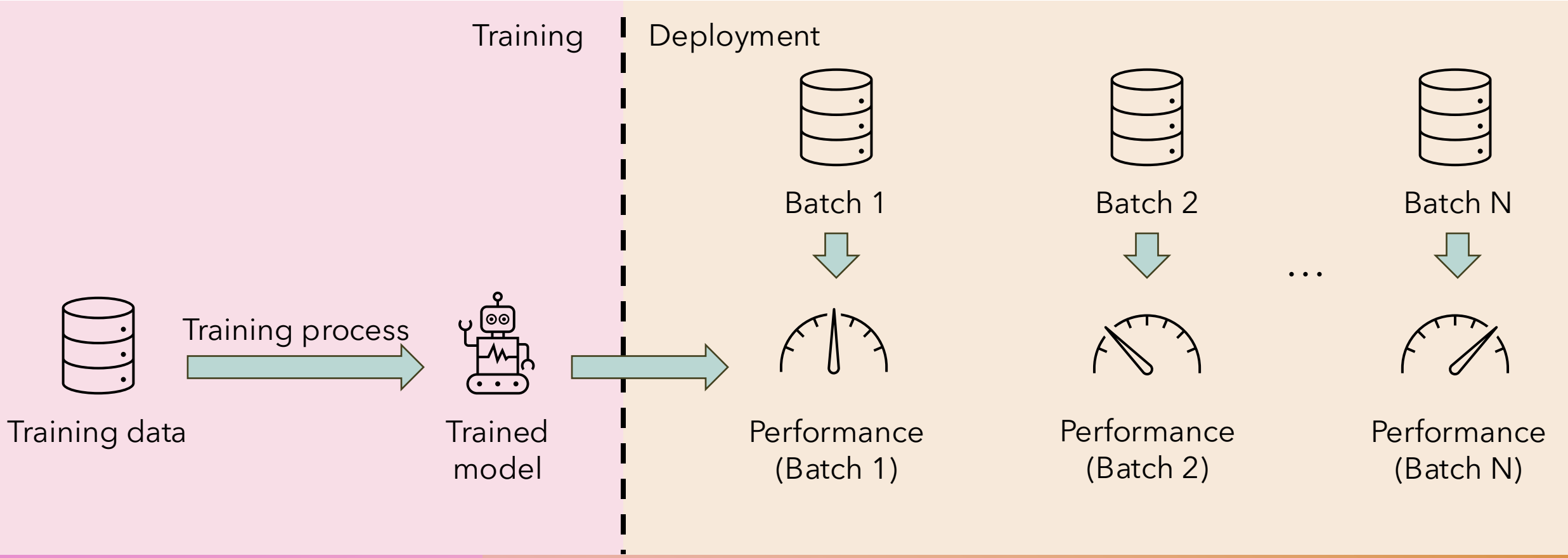


A Synthetic Benchmark to Explore Limitations of Localized Drift Detections

Flavio Giobergia, Eliana Pastor, Luca De Alfaro, Elena Baralis

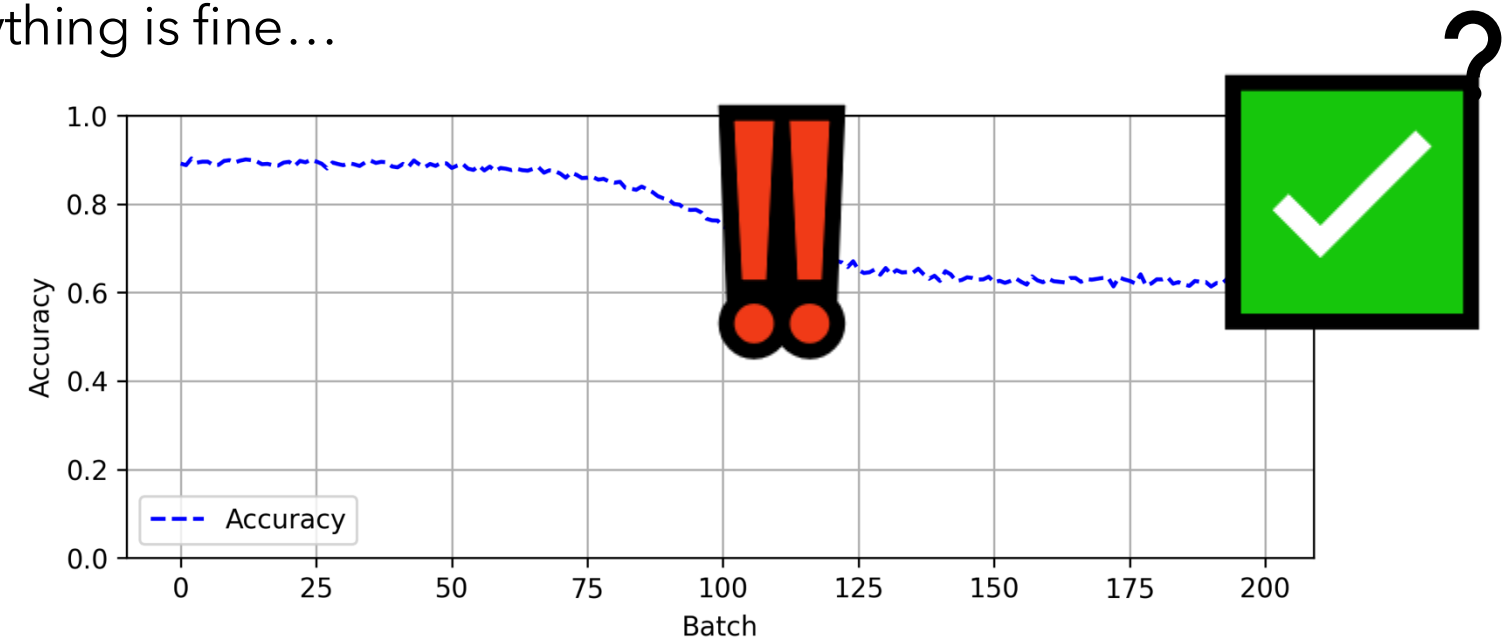
DELTA Workshop @ KDD 2024
Barcelona, Spain
August 26, 2024

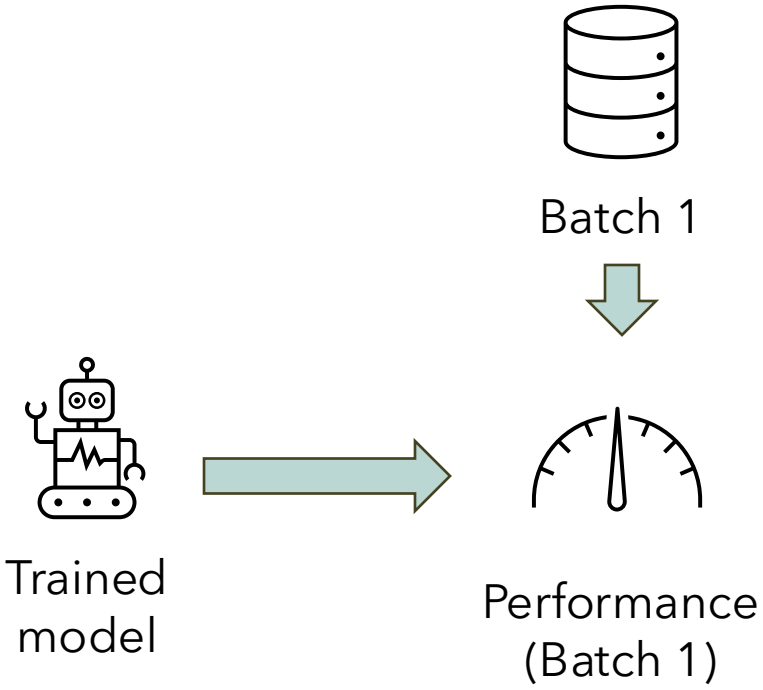
A CLASSIC SCENARIO

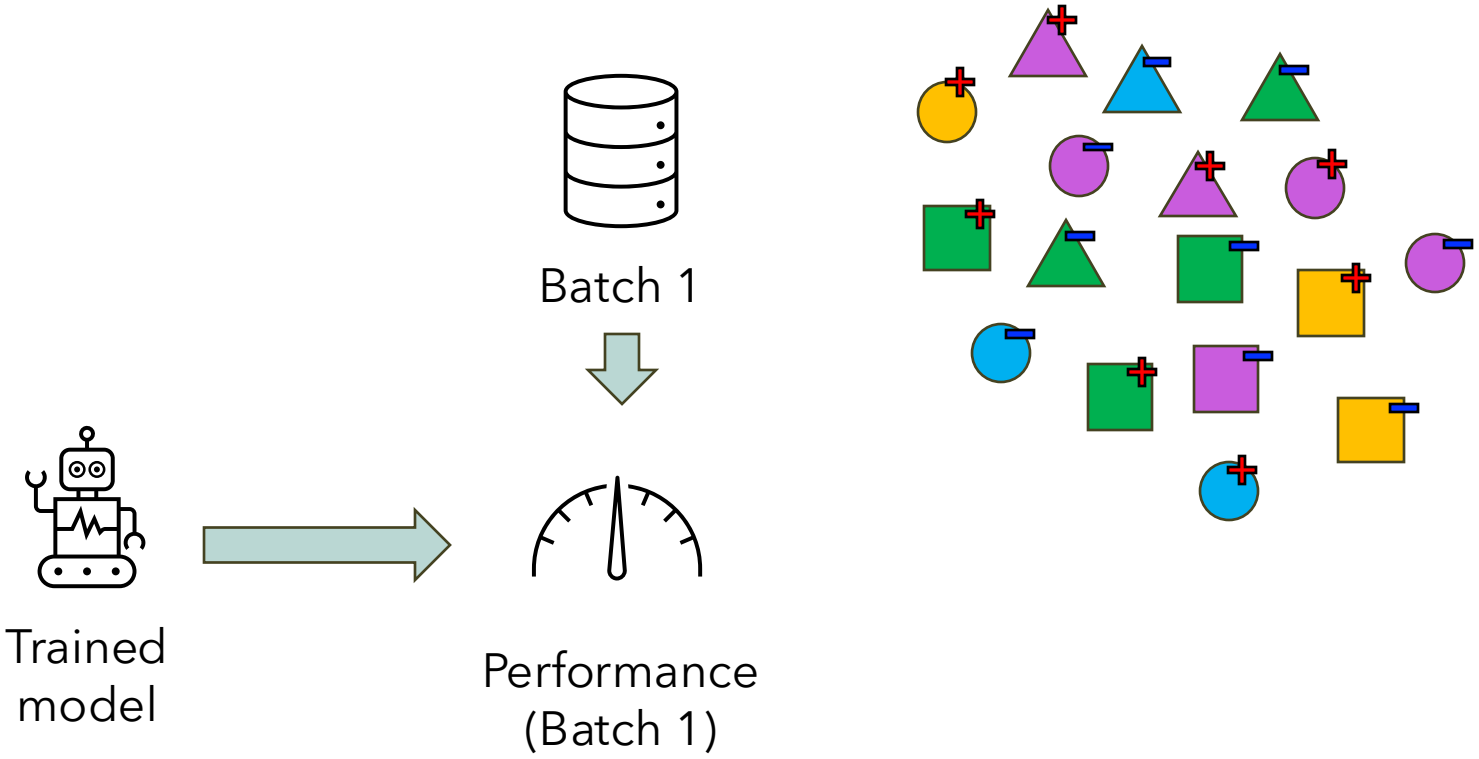


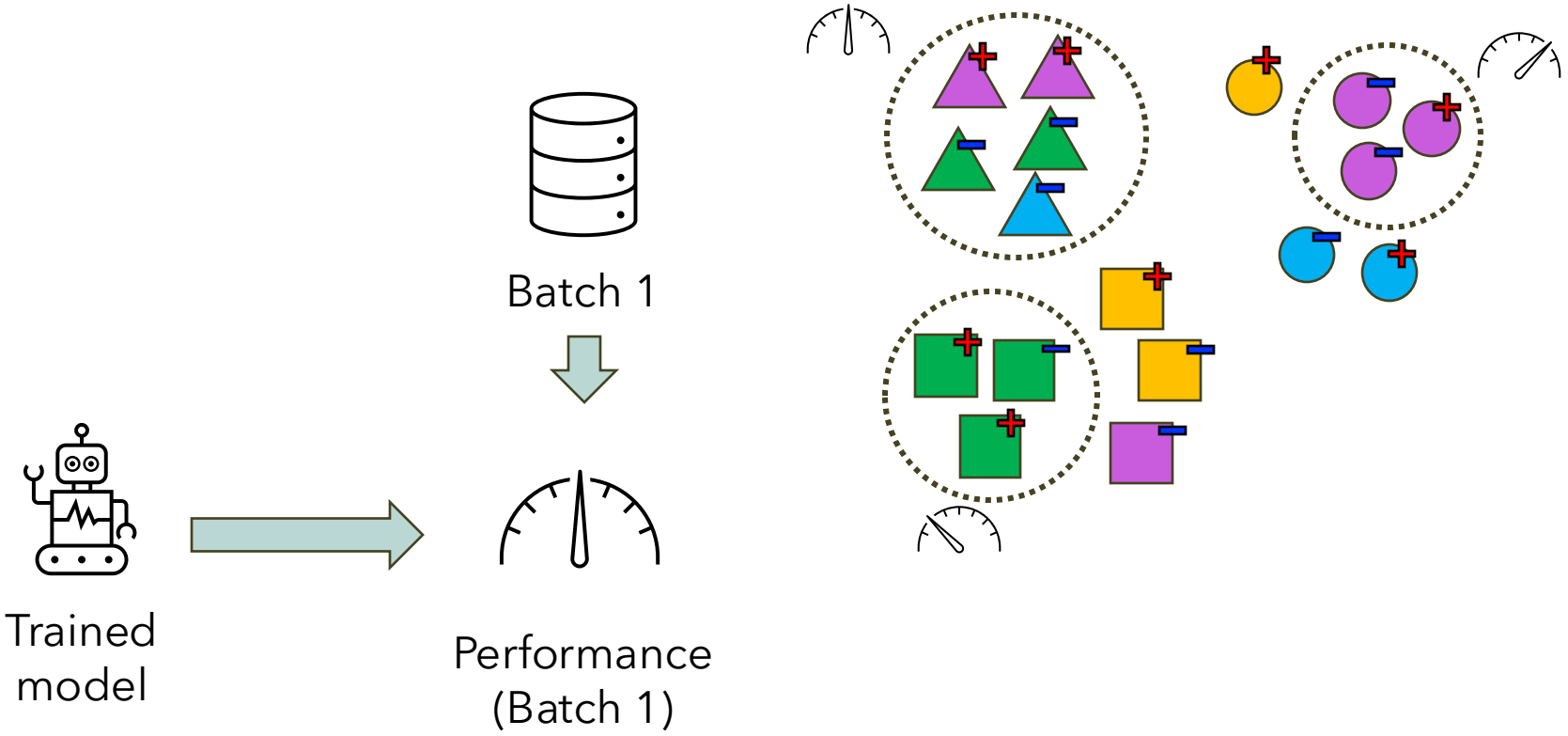
MODEL PERFORMANCE

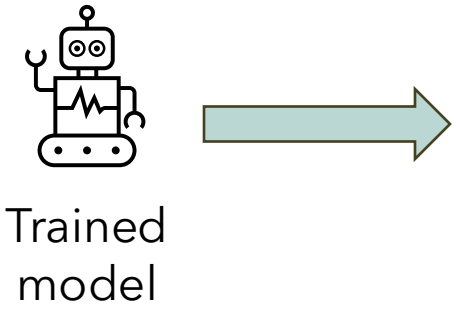
- When a drift occurs, the performance of a model will be affected over time
- If a drift occurs, we'd like to notice & take action
- If no drift is detected, everything is fine…
  - right?




Batch 1
Trained model
Performance (Batch 1)

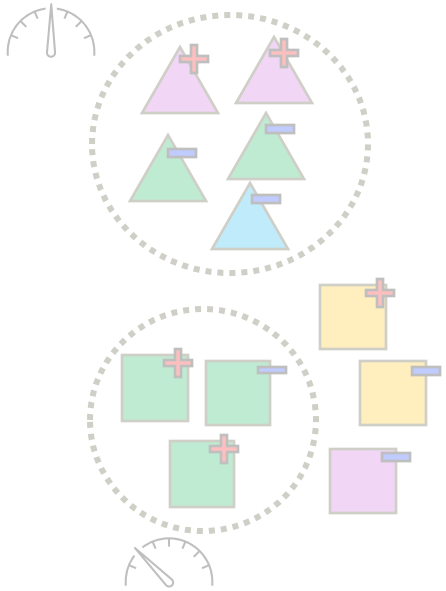

Batch 1
Trained model
Performance (Batch 1)

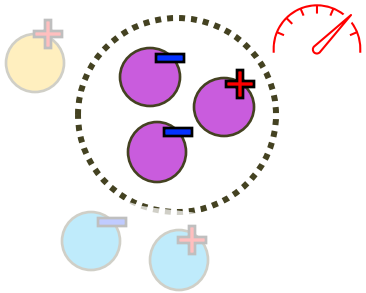

Batch 1
Trained model
Performance (Batch 1)

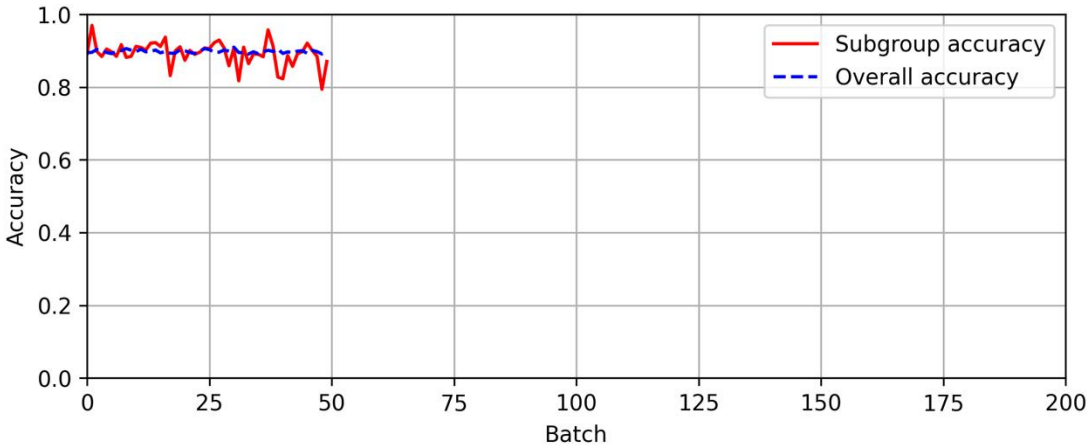

Trained
model

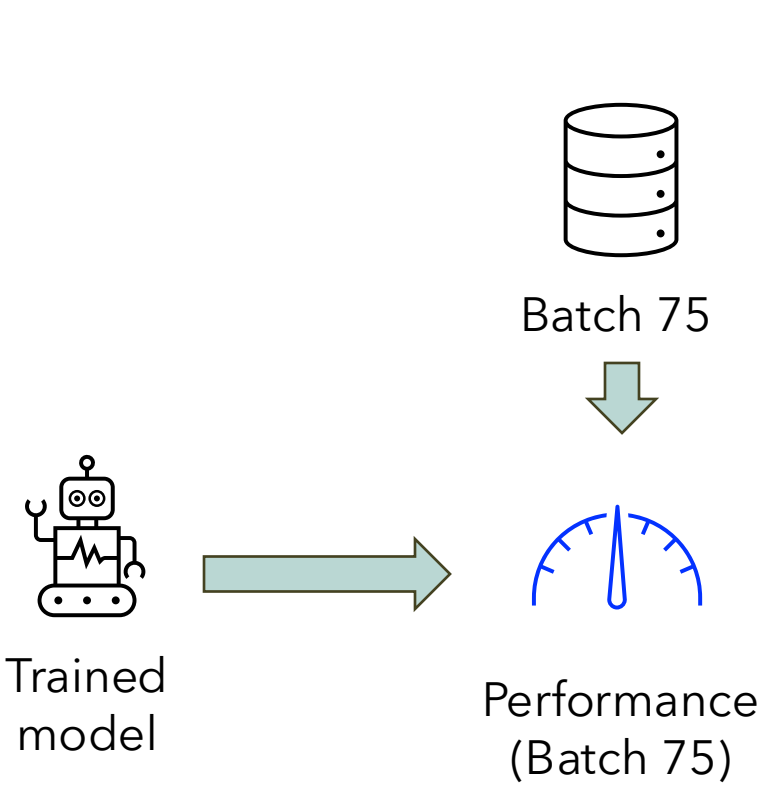

Performance
(Batch 1)

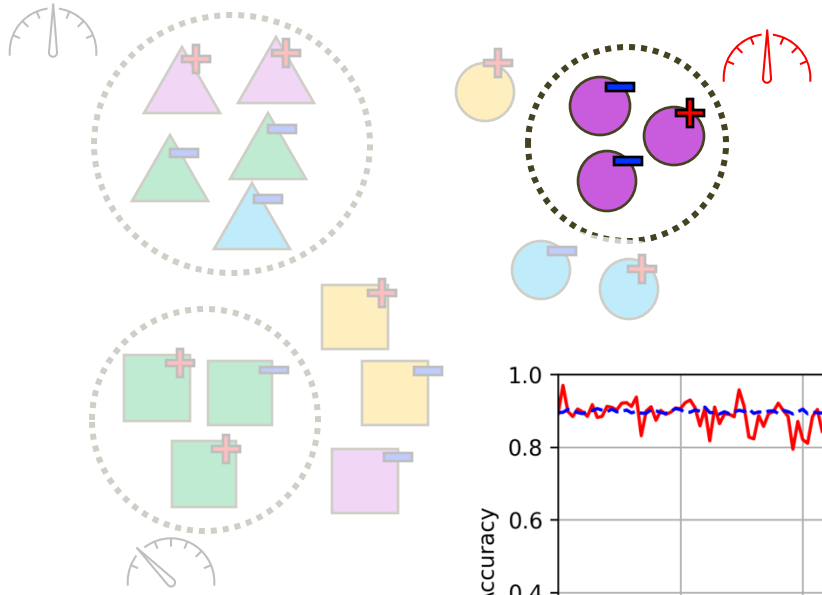

Batch 1

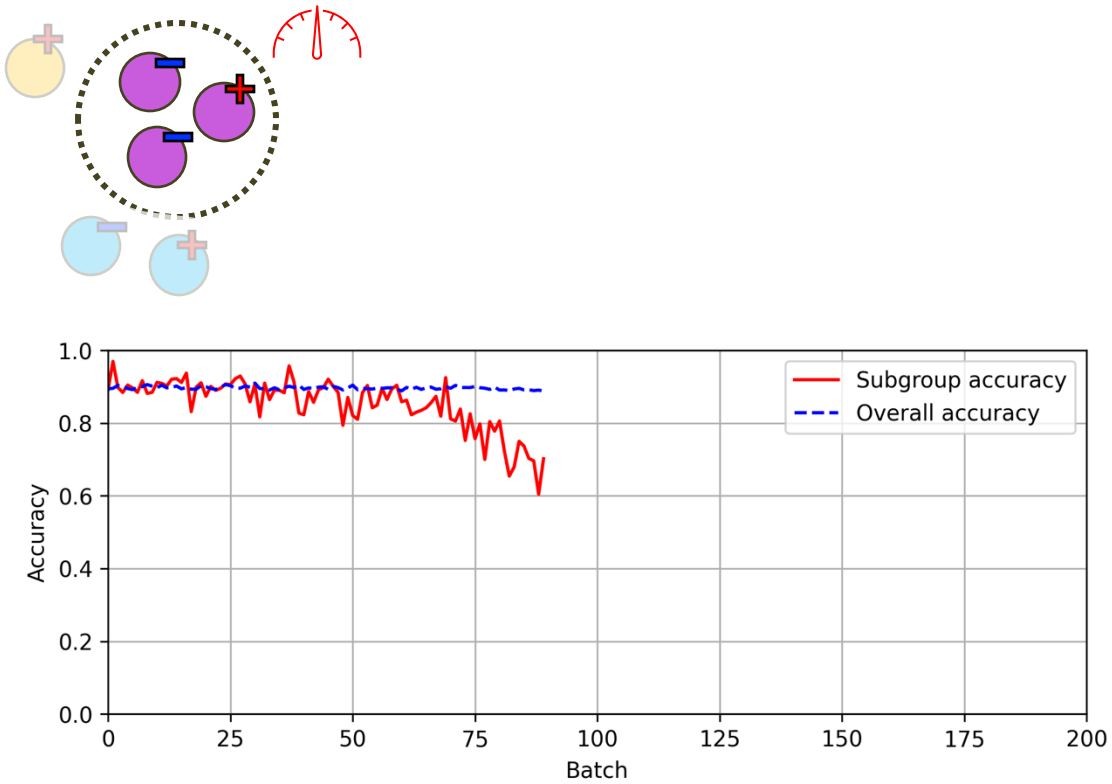


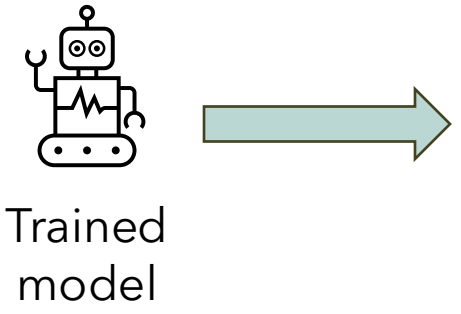



Subgroup accuracy
Overall accuracy
Accuracy
Batch
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200


Batch 75
Trained model
Performance (Batch 75)

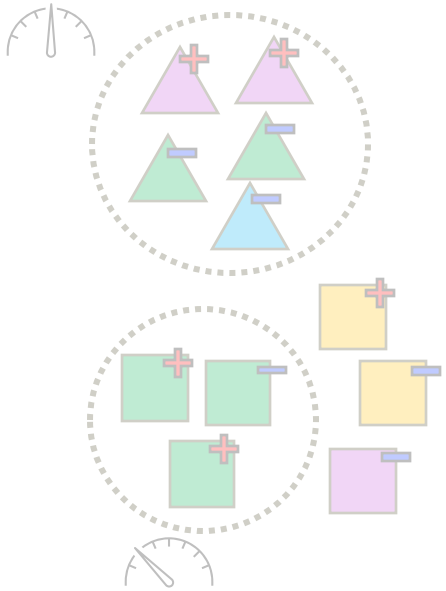


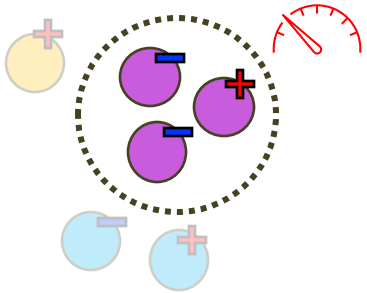

Subgroup accuracy
Overall accuracy
Accuracy
Batch
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200

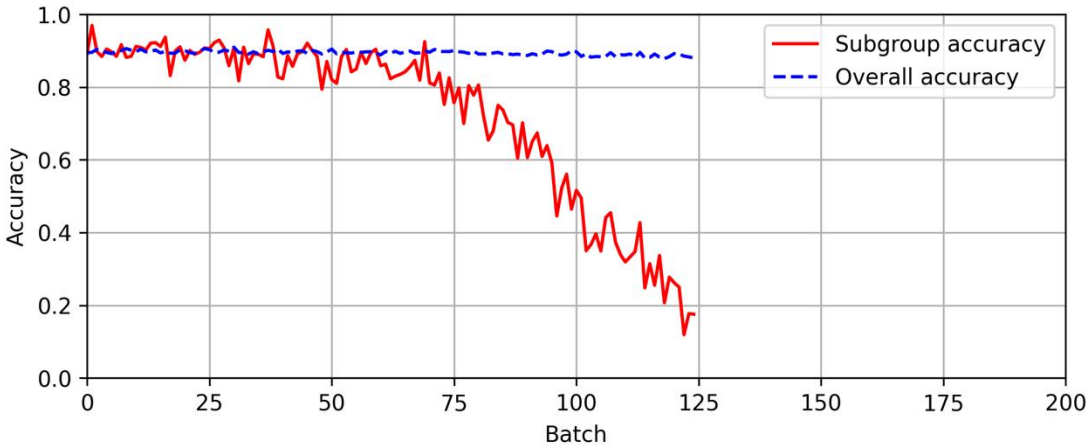

Trained
model


Performance
(Batch 150)


Batch 150



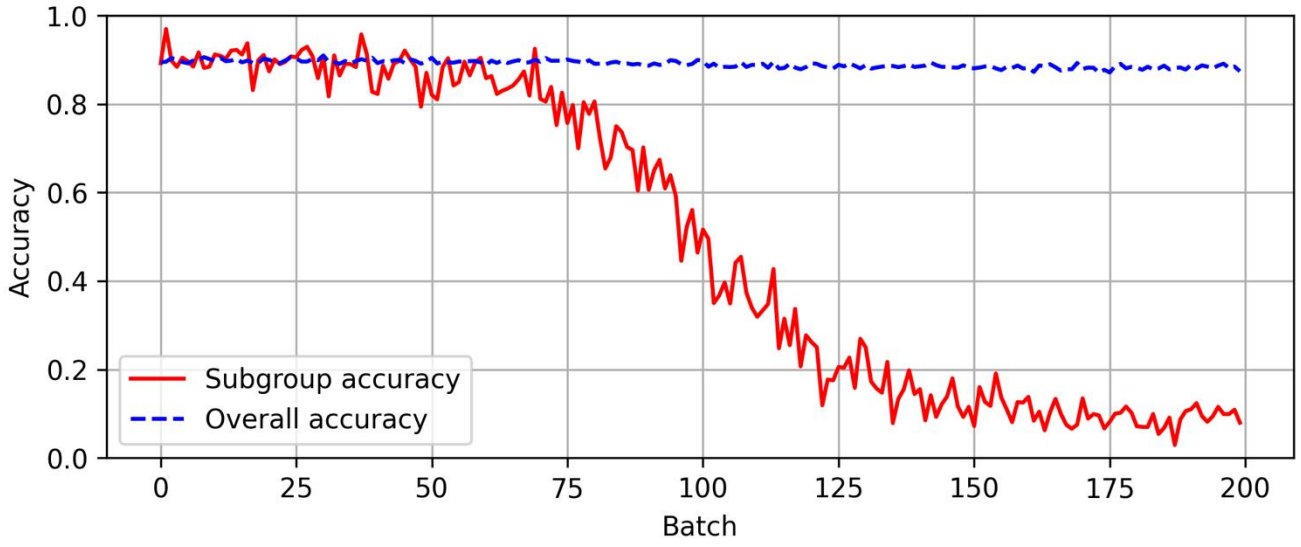



Subgroup accuracy
Overall accuracy
Accuracy
Batch
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200

LOCAL DRIFTS MAY GO UNNOTICED!

- A *small enough subgroup* of points may drift and *not have a significant effect* on the overall performance!

- If the drift goes *undetected*, the *subgroup* will be *affected disproportionatly* and nobody will even know



CREATING A LOCALIZED DRIFT BENCHMARK

- We set out to create a *controlled benchmark*, with *localized drifts* injected into it.

- Based on this dataset, we'd like to *quantify* the extent to which existing *drift detectors can find localized drifts*.

AGRAWAL DATASET

- We base the benchmark on the Agrawal stream generator [1] , a commonly adopted synthetic stream

- Each generated sample is a point (person) characterized by various features:
  - E.g. Age, Salary, Education level

- Concept Drift is injected by using different classification functions to generate target labels

[1] Rakesh Agrawal, Tomasz Imielinksi, and Arun Swami. “Database Mining: A Performance Perspective”, IEEE Transactions on Knowledge and Data Engineering, 5(6), December 1993.

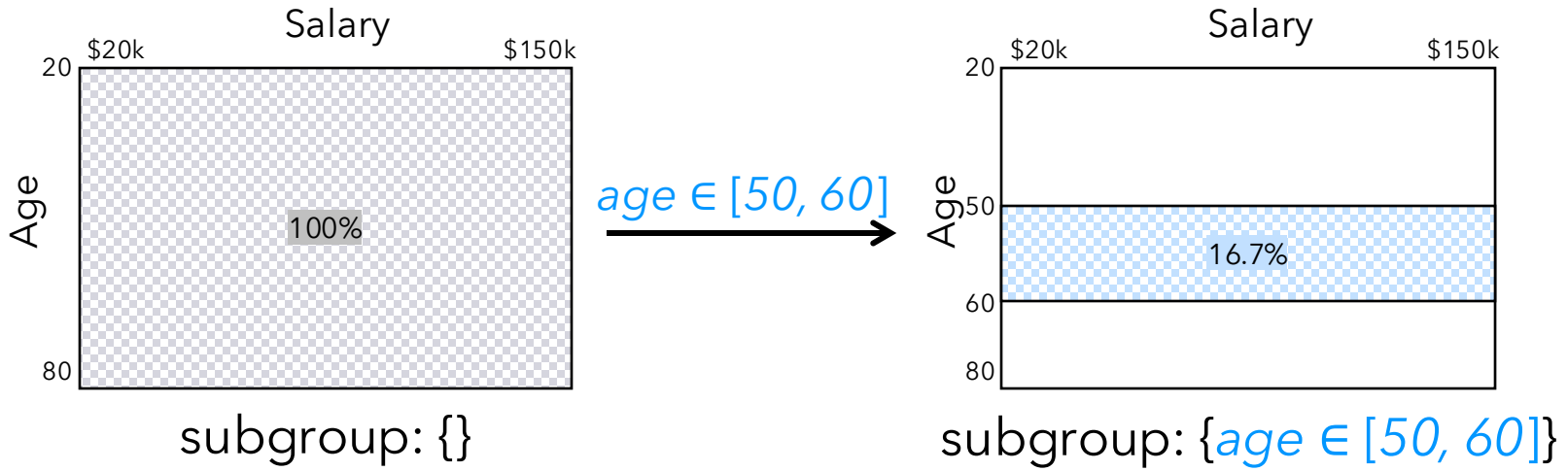
DRIFTING SUBGROUP(S)

- We want to target one specific subpopulation
  - E.g., “purple circles”
  - This simulates a subgroup that starts acting differently

- Desiderata for *Subgroup Agrawal Drift* (SAD):
  - Injected subgroups of different sizes
  - Subgroups defined in a procedural manner

GREEDY SUBGROUP DEFINITION

🎯 target support: 10%



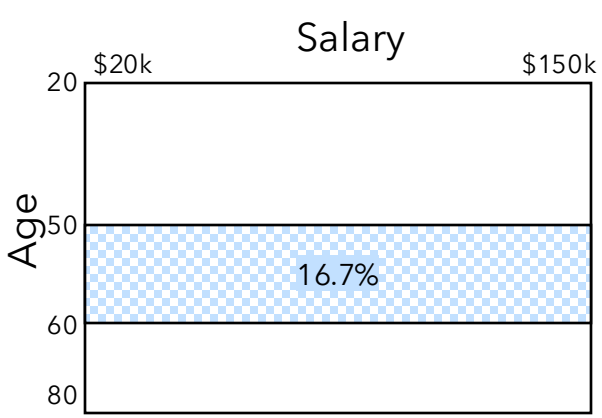
subgroup: {}

subgroup: {*age ∈ [50, 60]*}

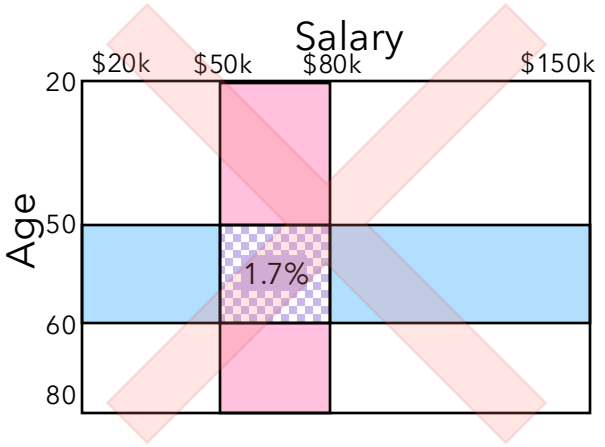
GREEDY SUBGROUP DEFINITION

🎯 target support: 10%

‼️ |0.017 – 0.1| > |0.167 – 0.1|



salary ∈ [$50k, $63k]

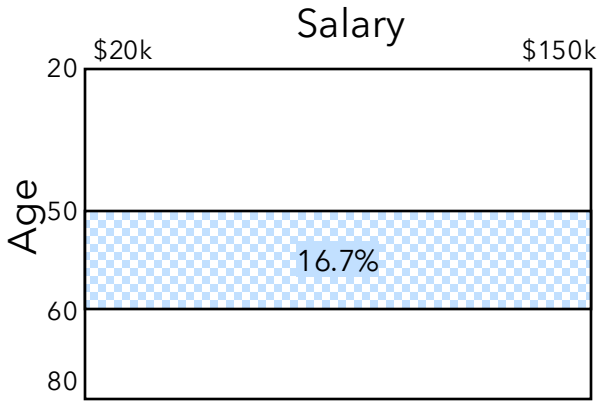


subgroup: {*age ∈ [50, 60]*}

GREEDY SUBGROUP DEFINITION

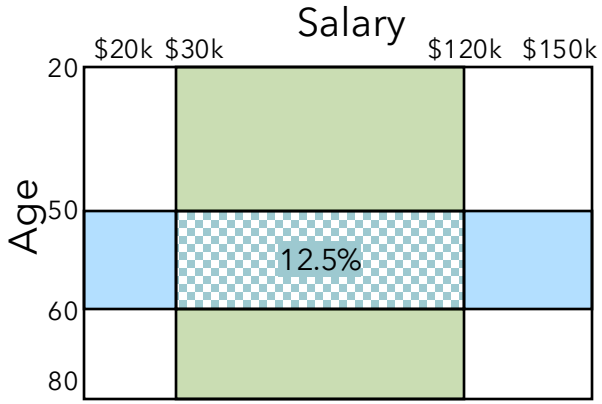
🎯 target support: 10%

✅ |0.125 – 0.1| < |0.167 – 0.1|



subgroup: {*age* ∈ [*50, 60*]}

salary ∈ [*$30k, $120k*]



subgroup: {*age* ∈ [*50, 60*], *salary* ∈ [*$30k, $120k*]}

HOW DO DETECTORS PERFORM?

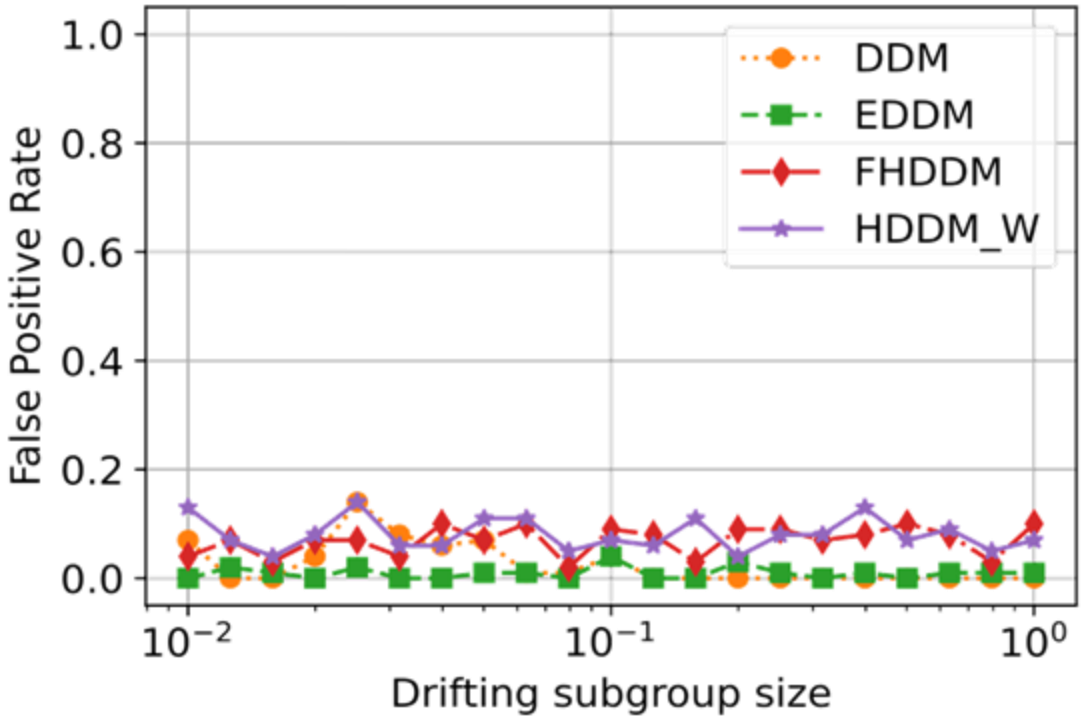
- We injected drift in subgroups of different sizes:
  - from 1% -- very *small subpopulations*,
  - to 100% (i.e., the *entire population* is affected by drift).

- We evaluate the results in terms of FNR, FPR, F1 score, accuracy for various drift detectors

FALSE POSITIVE RATE

- For all considered methods, the False Positive Rate is fairly constant, regardless of subgroup size, and low

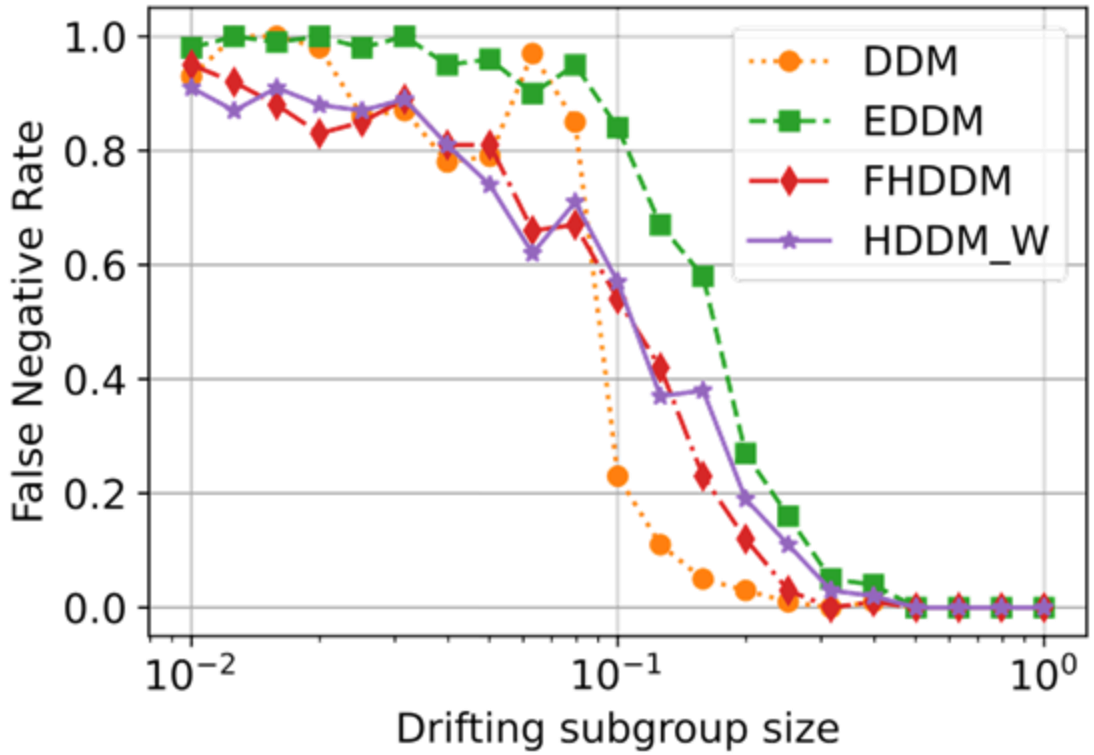
- In other words, the methods rarely fire "positive" predictions when no drift is occurring



FALSE NEGATIVE RATE

- By contrast (and, as expected), the FNR is heavily *affected by the subgroup size*.

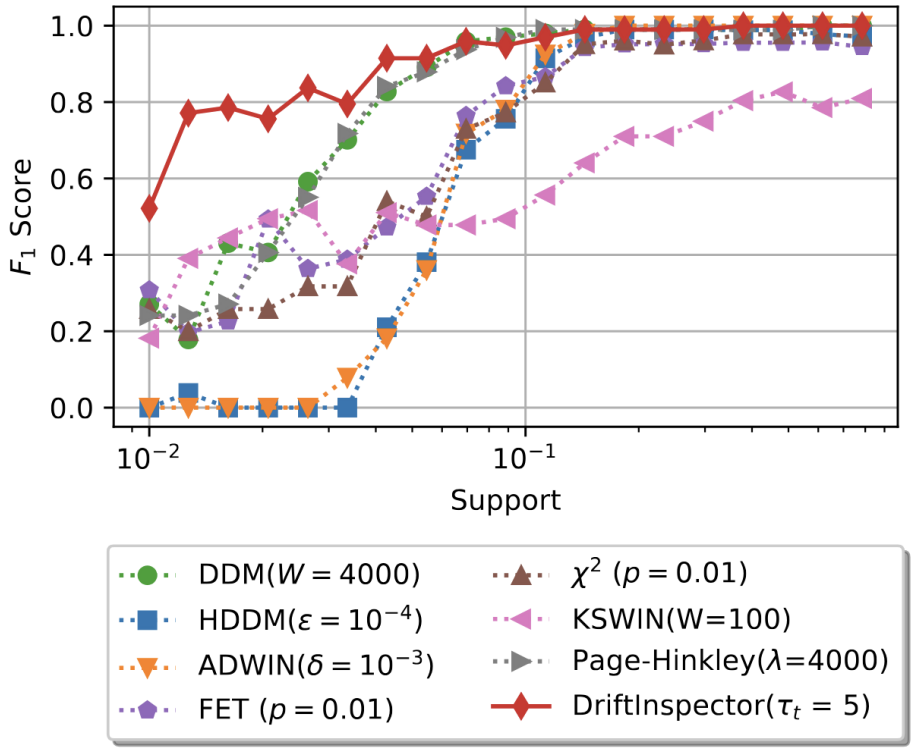
- When a smaller subgroup drifts, all methods struggle to detect drift
  - Even though, remember, the entire subgroup is drifting!
- For subgroups larger than ~10% of the population, all methods get better



WE'RE WORKING ON IT!

- We have addressed this problem in a recent work
  - With very promising preliminary results :)

- Pre-print available at https://bit.ly/DriftInspector



CONCLUSIONS

- We argue that drift detectors should be able to detect localized drifts
- We introduce Subgroup Agrawal Drift, a synthetic benchmark with local drift injections
- We show that various drift detectors struggle to detect drifts
- We hope for this to spark some interest in future efforts :)
  - (We are already onto that!)

THANK YOU :)

💁 Flavio Giobergia

✉ flavio.giobergia@polito.it

𝕏 @fgiobergia

https://bit.ly/SAD-repo

https://bit.ly/SAD-drift

https://bit.ly/DriftInspector

AGRAWAL DATASET

- We base the benchmark on the Agrawal stream generator [1]
- Each generated sample is a point (person) characterized by various features:
  - Salary
  - Commission
  - Age
  - Education level
  - Car maker
  - Zip code of the town
  - Value of the house
  - Years house owned
  - Total loan amount

[1] Rakesh Agrawal, Tomasz Imielinksi, and Arun Swami. "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, 5(6), December 1993.

DRIFTING AGRAWAL

- Each point is assigned a binary label (whether a loan should be approved)

- 10 classification functions $f_i : X \rightarrow \{0,1\}$ exist to assign each point to its ground truth
  - e.g., $f_8(x) = (0.67 \text{ x (salary + commission) - 5000 x elevel - 20K)} > 0$

- Various works introduce concept drift by gradually shifting from f_i to f_j $(i \neq j)$
  - E.g., $p(f = fi) = 1/(1 + \exp(-4(t-p)/w)$
  - Uses a sigmoid to parametrize when the drift occurs (p) and how long the transitory is (w)

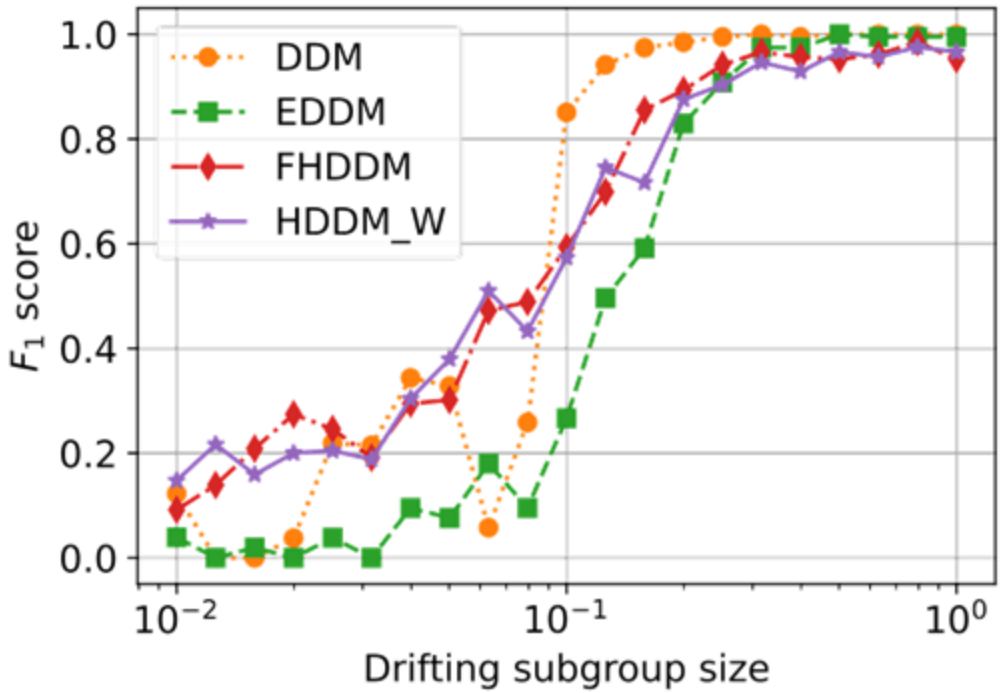
EXAMPLES OF SUBGROUPS

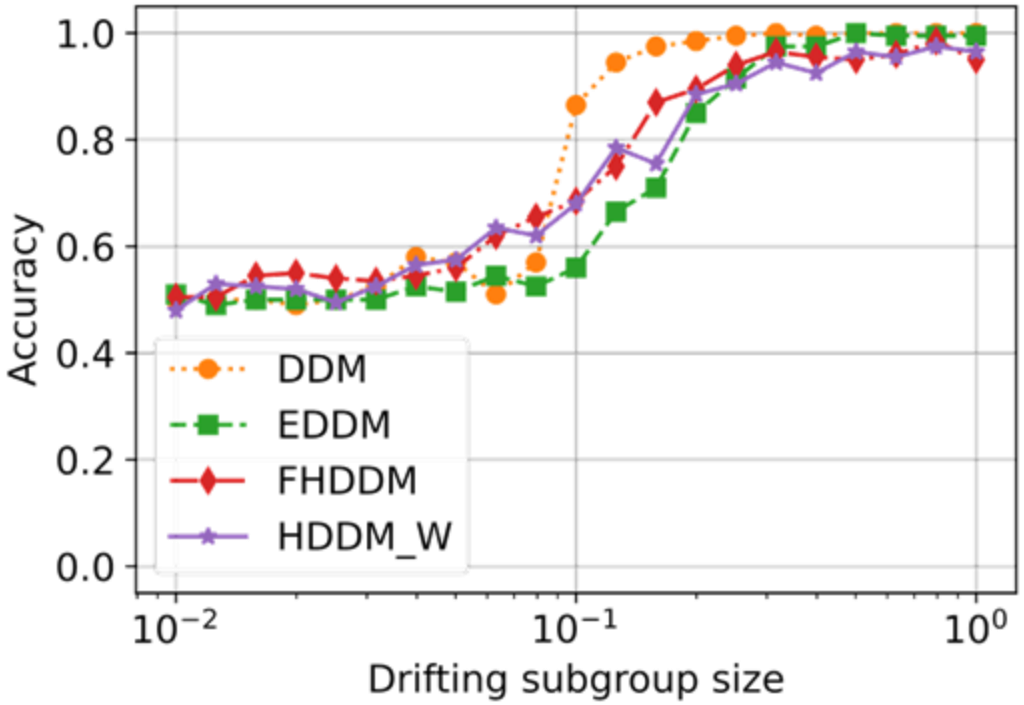
- So, we can generate subgroup of (approximately) any target size, and have that subgroup drift!
- Time to test some techniques!

Generated subgroup	Target size	Computed size	Actual size
$\{\ elevel \in [0,3) \wedge$ $zipcode \in [6,7) \wedge$ $age \in [29,78)\ \}$	0.05	0.0536	0.0552
$\{\ car \in [15,19) \wedge$ $salary \in [39000,116000) \wedge$ $zipcode \in [0,8)\ \}$	0.1	0.1045	0.107
$\{\ zipcode \in [2,5) \wedge$ $salary \in [30000,139000) \wedge$ $age \in [22,80) \wedge$ $car \in [1,20)\ \}$	0.25	0.2505	0.2527
$\{\ elevel \in [1,4) \wedge$ $age \in [20,78) \wedge$ $salary \in [21000,140000) \wedge$ $hyears \in [1,30)\ \}$	0.5	0.501	0.4965

Some examples of generated subgroups. Note that SG sizes may (slightly) differ from the requested one

F1, ACCURACY


DDM
EDDM
FHDDM
HDDM_W
F_1 score
1.0
0.8
0.6
0.4
0.2
0.0
10^{-2}
10^{-1}
10^{0}
Drifting subgroup size


DDM
EDDM
FHDDM
HDDM_W
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
10^{-2}
10^{-1}
10^{0}
Drifting subgroup size