Active learning with evolving streaming data

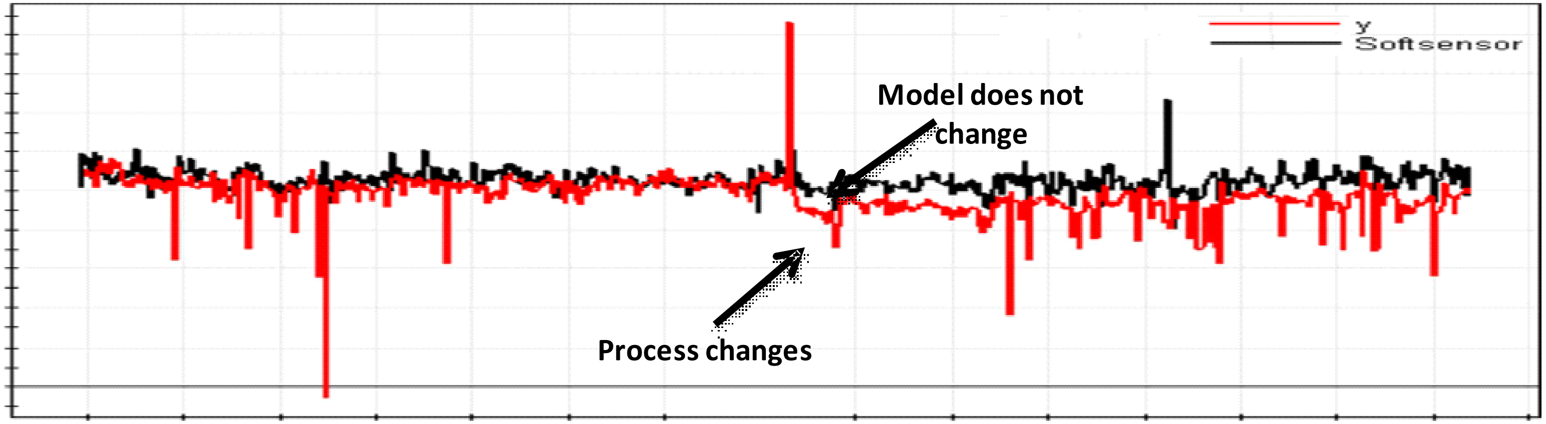
Indrė Žliobaitė, Albert Bifet,
Bernhard Pfahringer, Geoff Holmes

September 6, 2011

setting

Data stream mining

Chemical production plant
given sensor readings
predict the quality of the output
24/7 plant operation



source: Evonik Industries

Examples of data streams

Sensor data

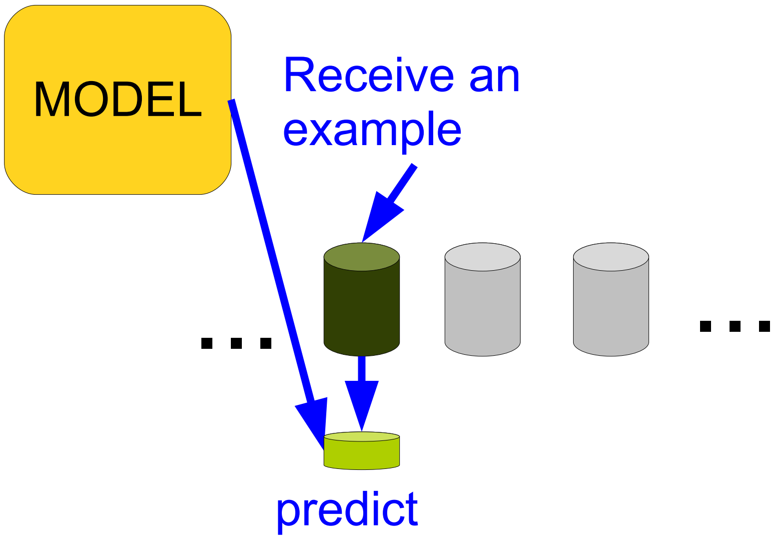
Web data
(logs,content)

Activity data

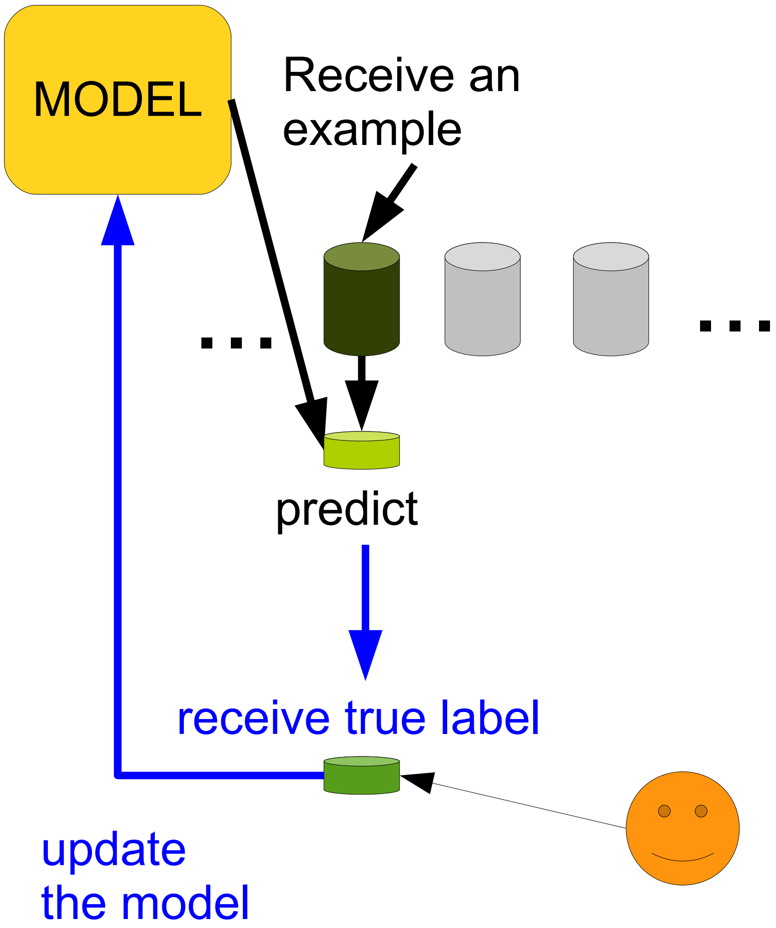
Mining data streams

- Data
  - arrives in real time, potentially infinitely
  - is changing over time
  - not possible to store everything, discard (or archive) after processing
- Requirements for predictive models
  - operate in less than example arrival time
  - fit into strictly limited memory
  - adapt to changing data (update/retrain online)
    - otherwise accuracy will degrade over time

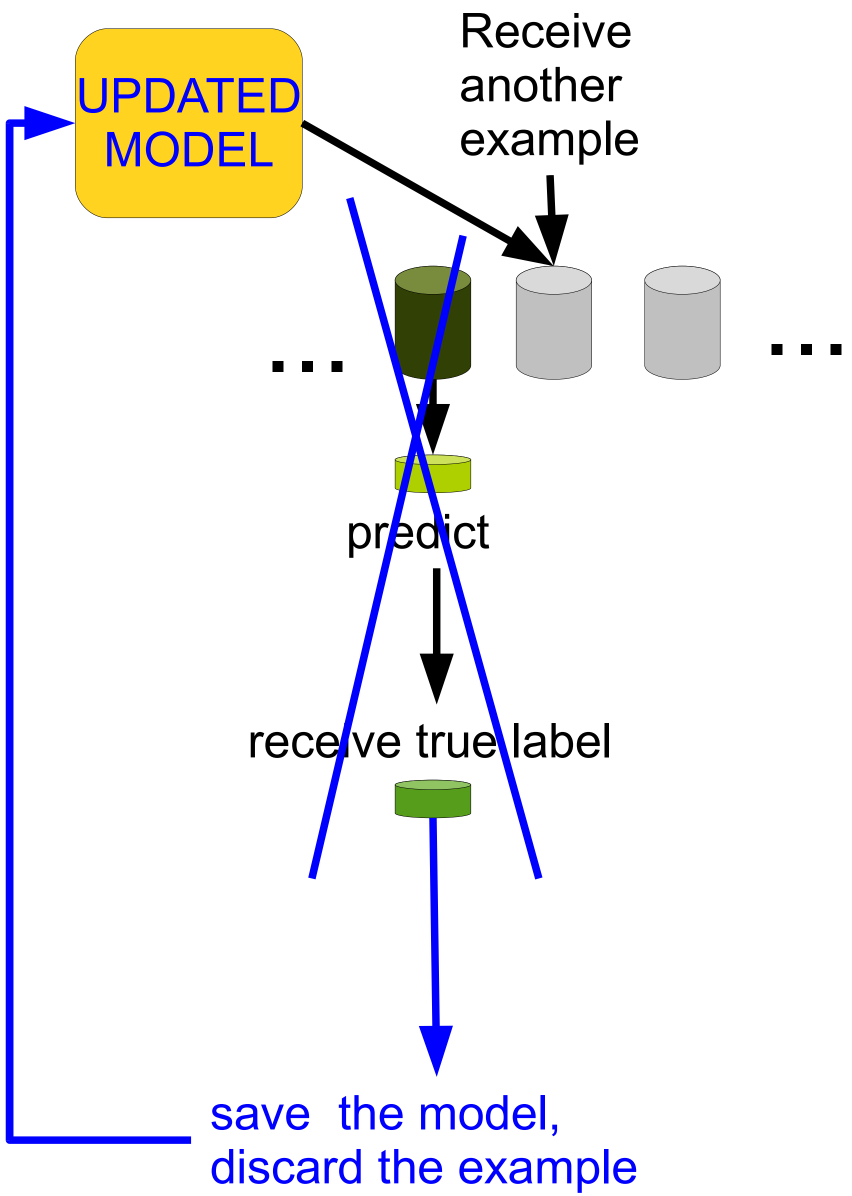
Predictive models for data streams



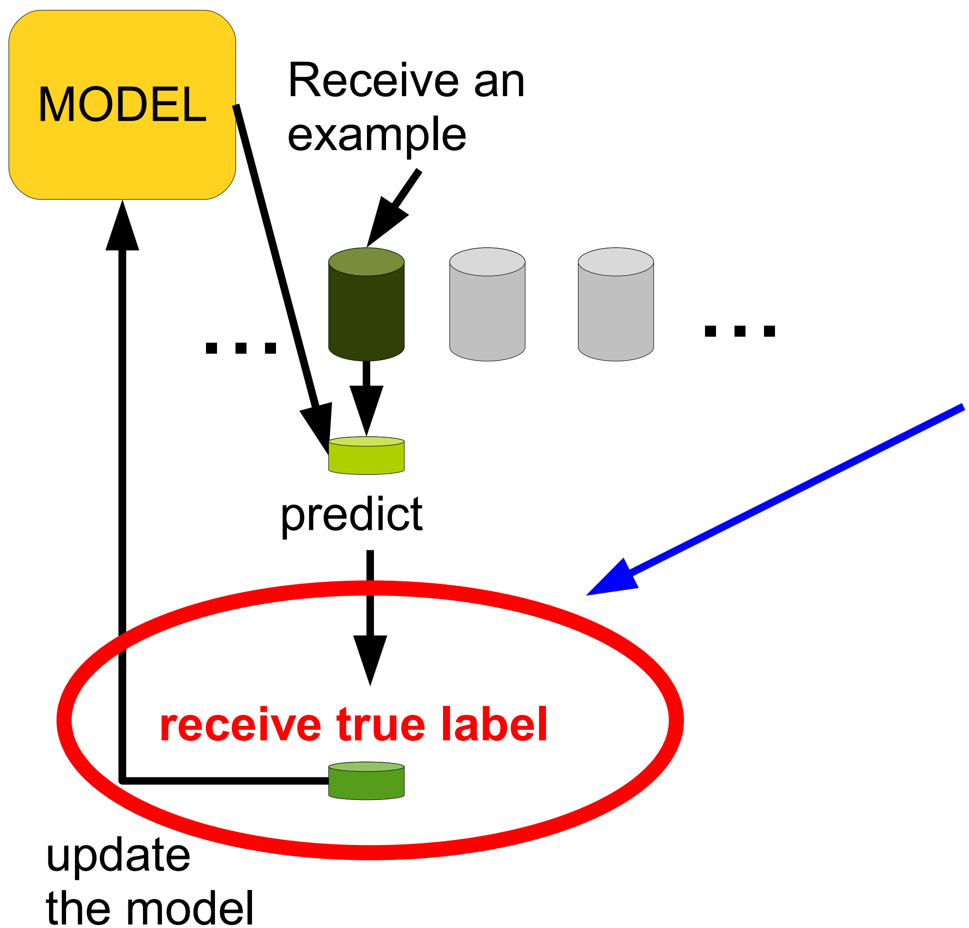
Predictive models for data streams



Predictive models for data streams



Predictive models for data streams

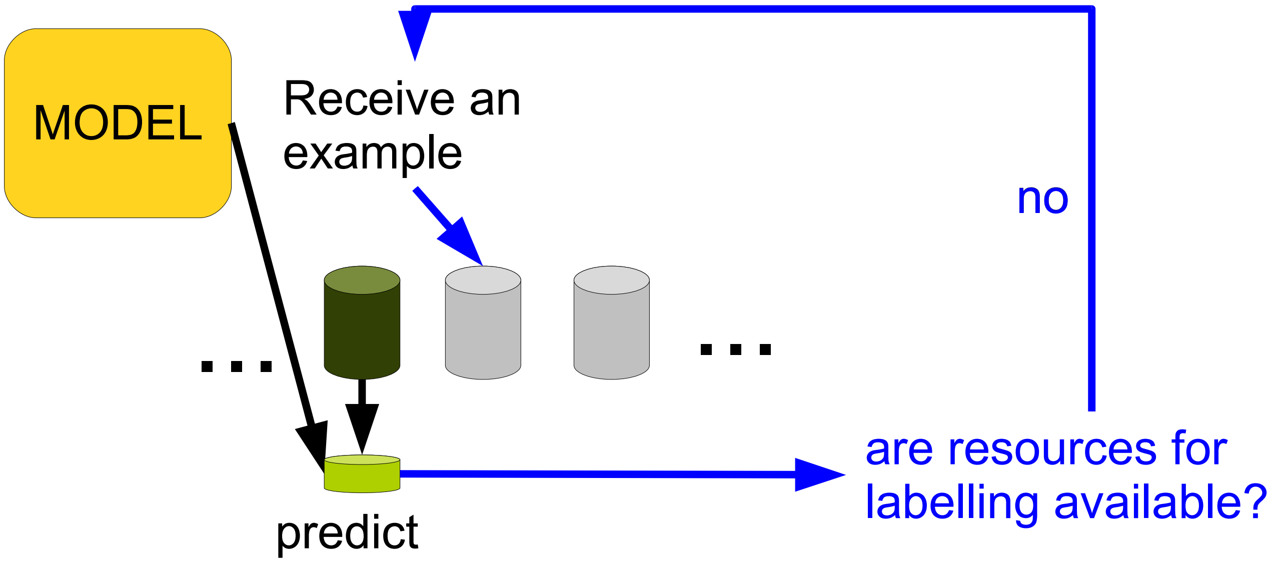


It is unreasonable
to ask for feedback at every iteration,

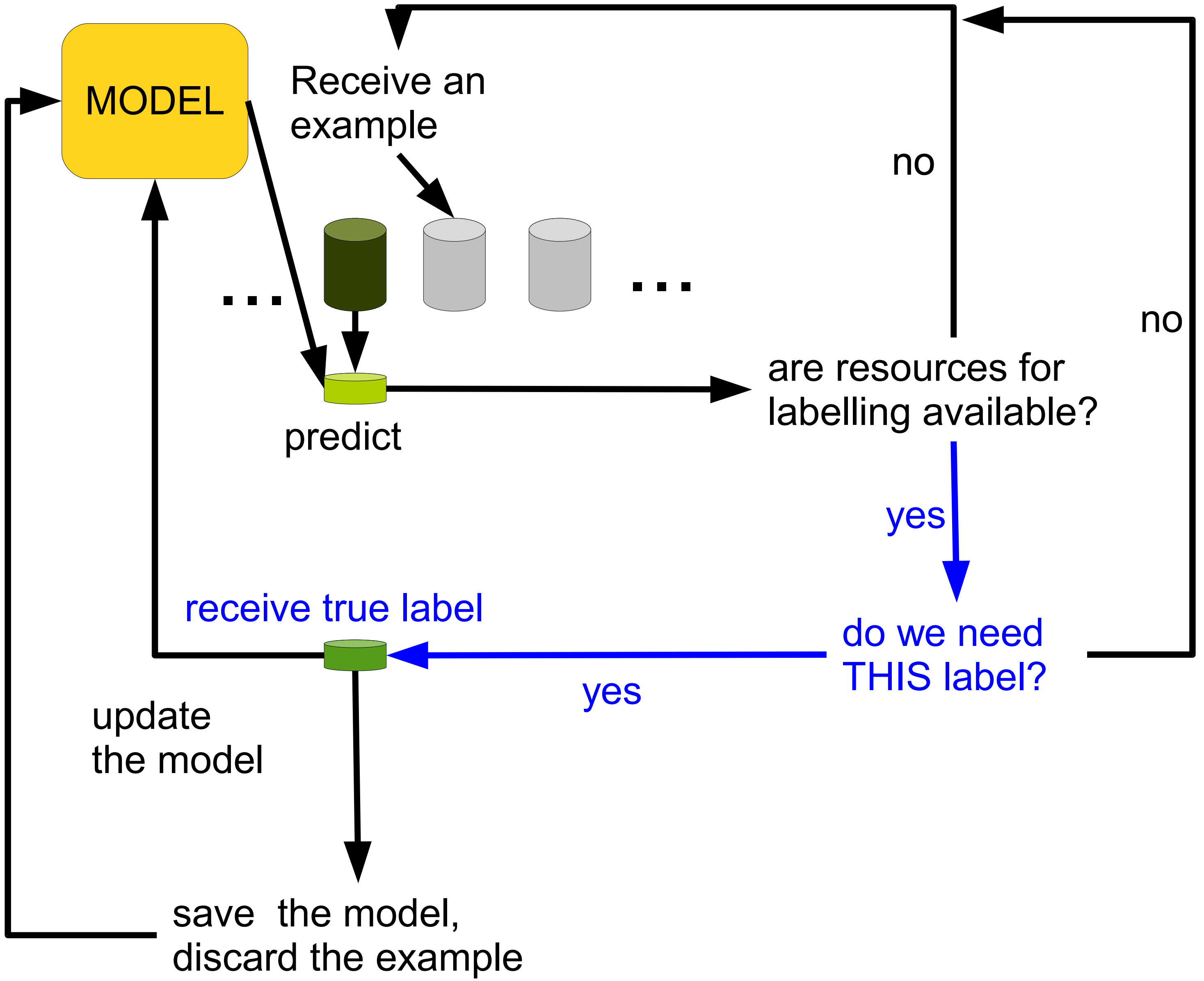
labels may be
costly or infeasible to obtain due to

- human labour (text, images)
- laboratory tests
- destructive tests

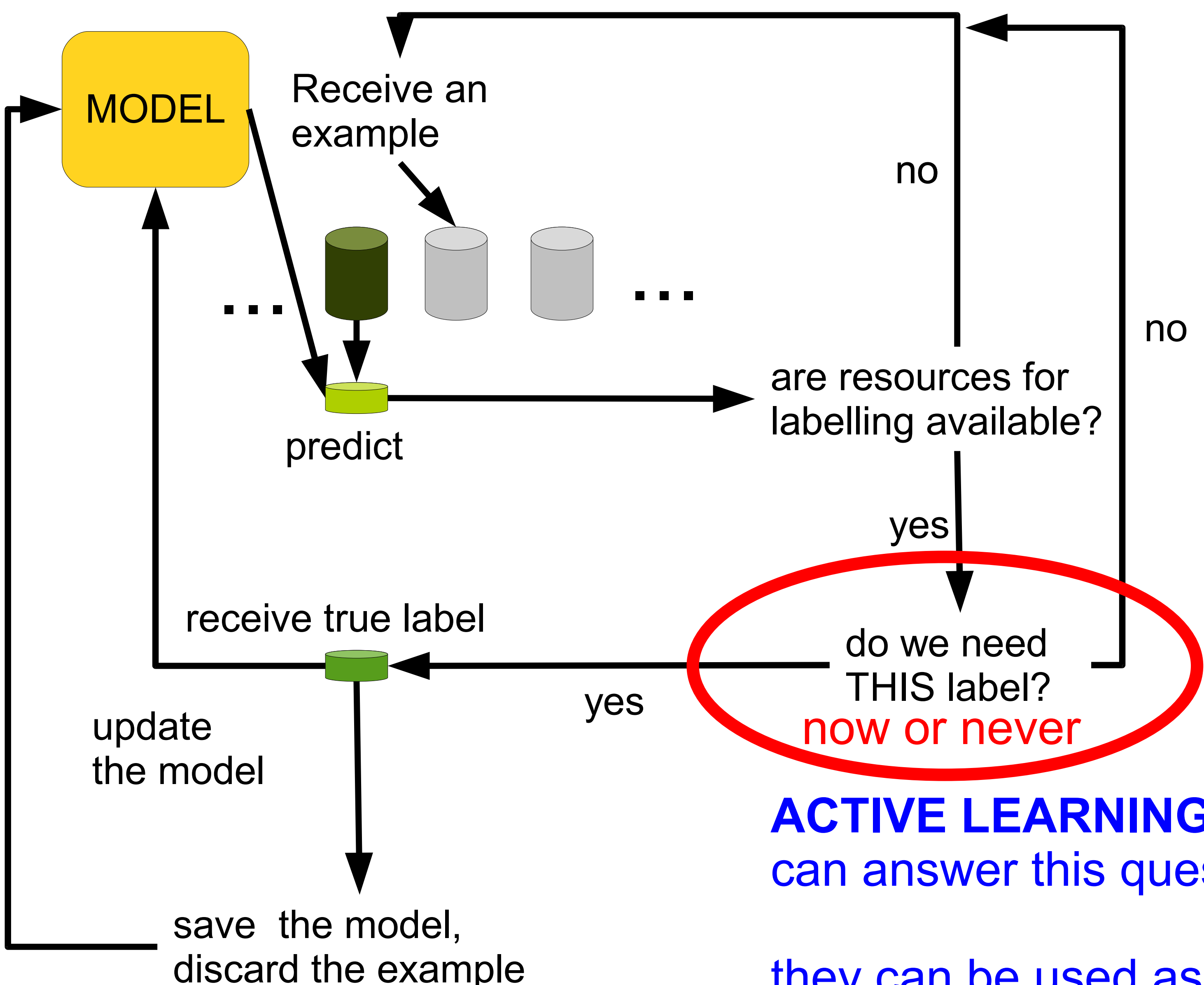
Active learning for data streams



Active learning for data streams



Active learning for data streams



ACTIVE LEARNING STRATEGIES
can answer this question,

they can be used as a wrapper
with a learning model of user's choice

Problem setting summary

- Supervised learning
- Evolving (changing) streaming data

- Models need to adapt to changes over time
- For adapting, feedback is needed

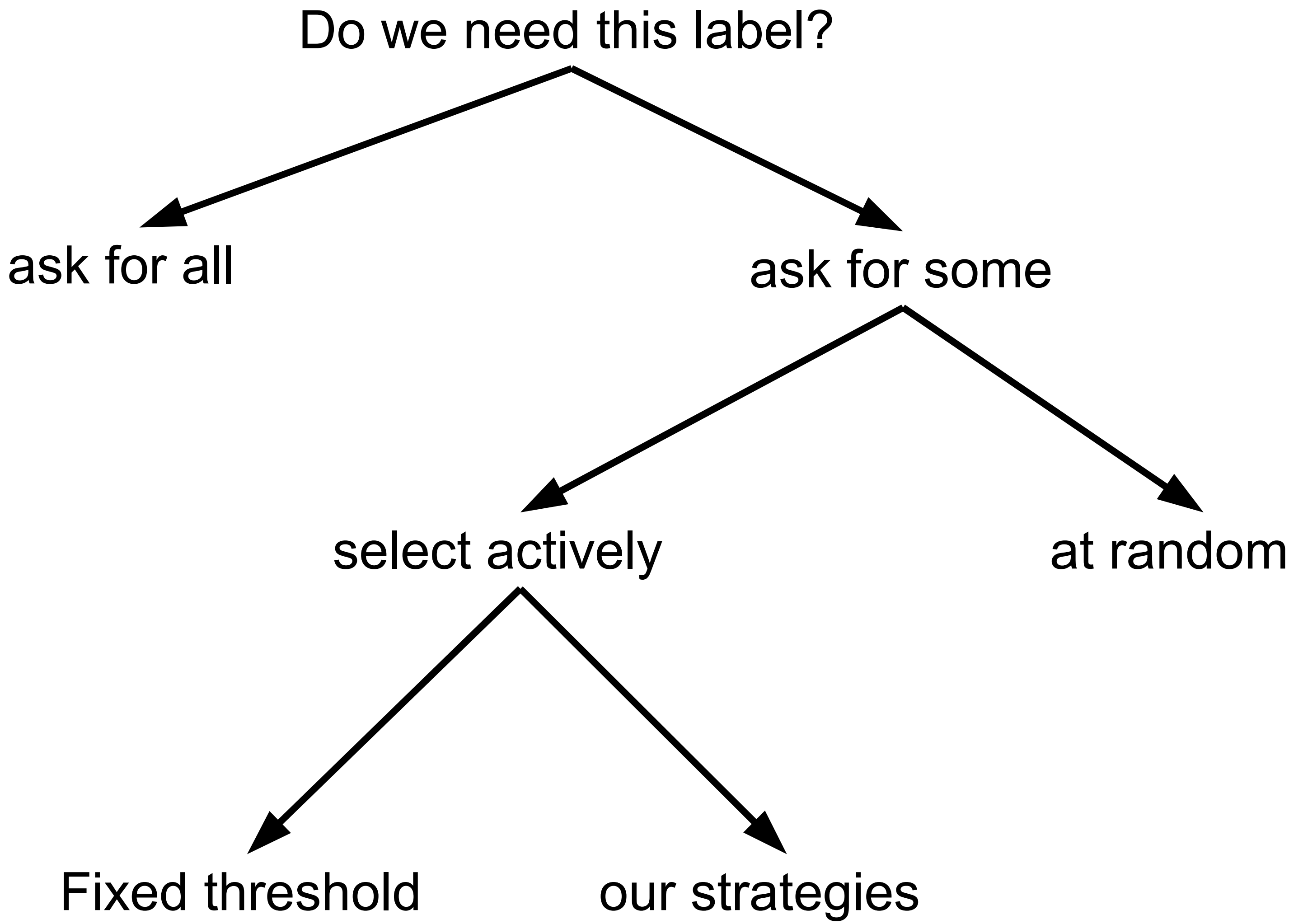
- True labels may be costly or infeasible to obtain
- We need to decide whether to ask for the true label for an example **now or never**

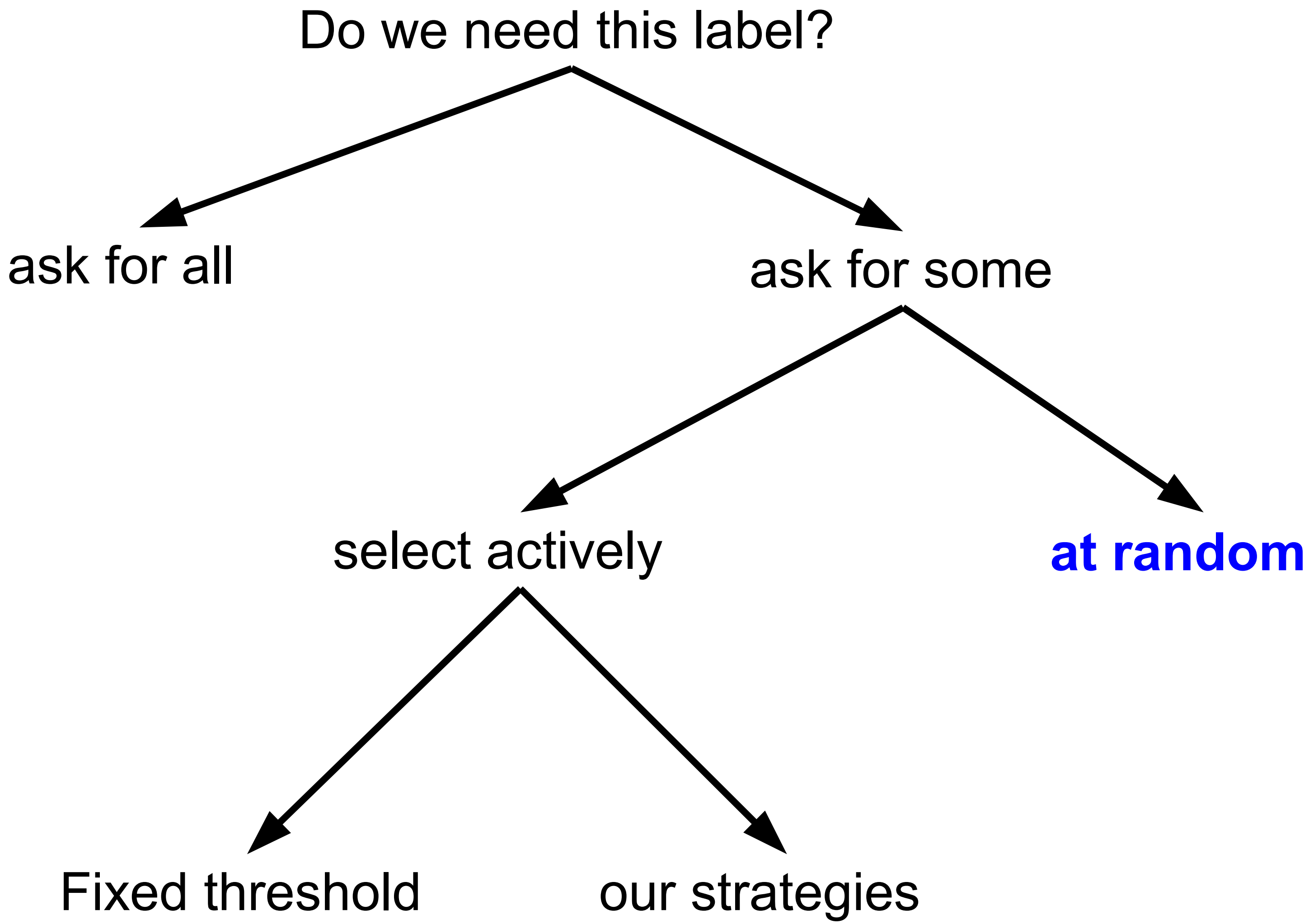
Contributions

- a framework
  - for active learning in the data stream setting
- specific requirements
  - for active learning strategies
- two corresponding active learning strategies
  - that can be integrated with an adaptive learning algorithm of a user's choice

active learning strategies for data streams

How to decide
whether to ask for the true label
for a given example?


Do we need this label?
ask for all
ask for some
select actively
at random
Fixed threshold
our strategies

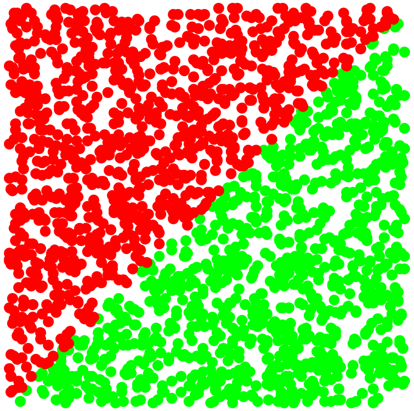


Do we need this label?
ask for all
ask for some
select actively
at random
Fixed threshold
our strategies

Random strategy (naive)

- Receive example X_t
- If $z<B$, where $z \sim U(0,1)$
  - ask for the *true label* y_t

original
(instance space)

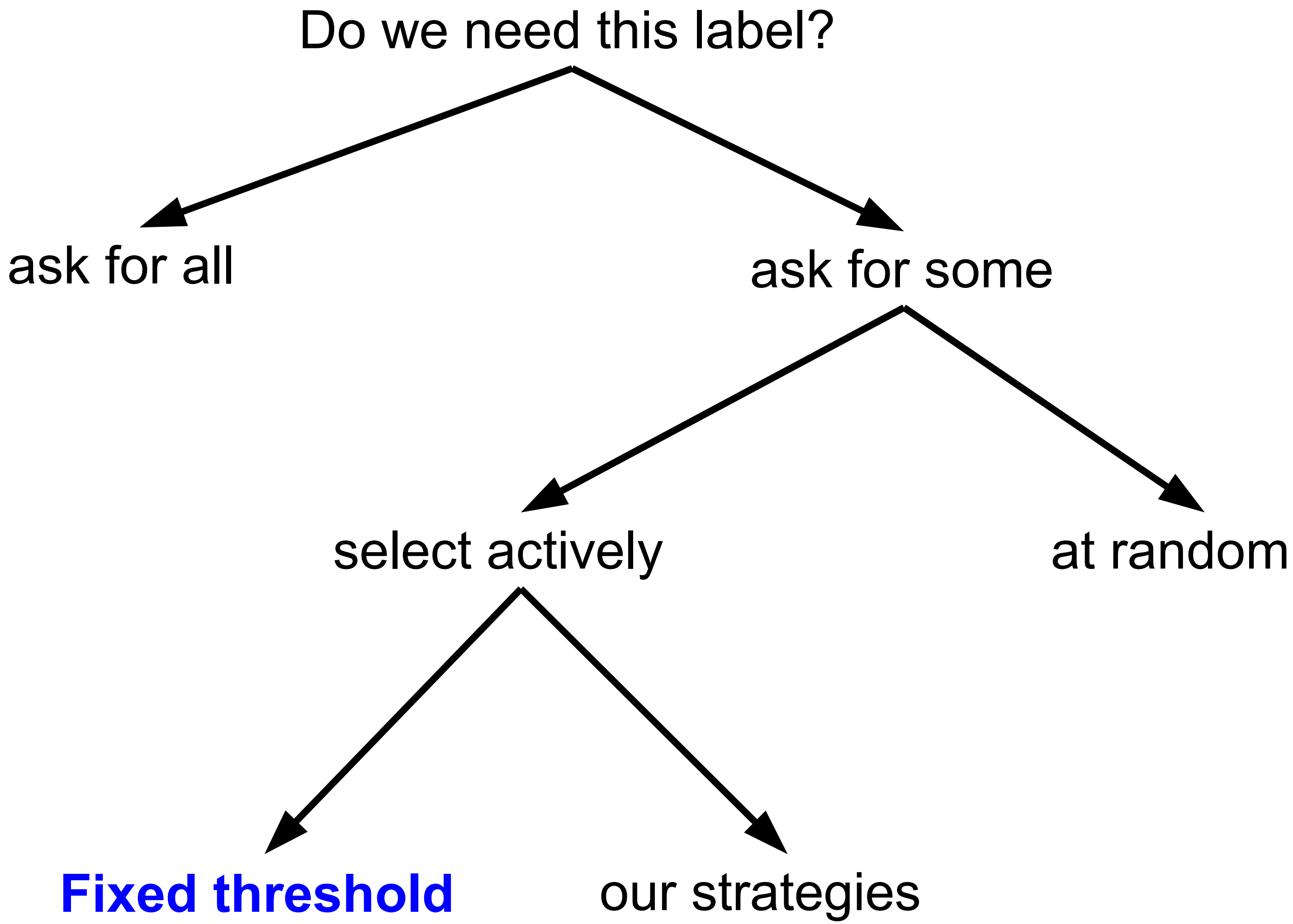


uniform
random sampling

Random strategy (naive)

- Receive example X_t
- If $z<B$, where $z \sim U(0,1)$
  - ask for the *true label* y_t

slow to learn


Do we need this label?
ask for all
ask for some
select actively
at random
Fixed threshold
our strategies

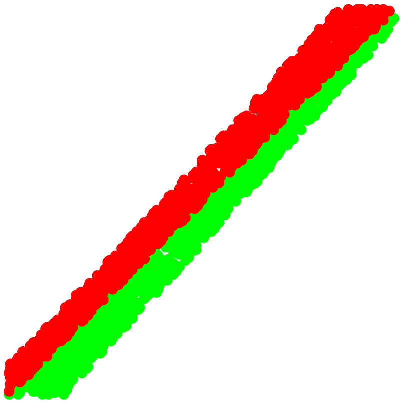
Online active learning in the data stream setting?

- Online setting
  - fix a threshold (e.g. uncertainty threshold)
  - check every incoming example against the threshold
  - if over the threshold, ask for the true label

Fixed uncertainty

- Receive example X_t and a prediction y^*_t
- If labelling *budget* is available [u/t < B]
  - If *uncertainty* of X_t is greater than threshold [$P(y^*_t|X_t) < K$]
    - ask for the *true label* y_t
    - update the model with (X_t,y_t), u=u+1

original



threshold

Online active learning in the data stream setting?

- Online setting
  - fix a threshold (e.g. uncertainty threshold)
  - check every incoming example against the threshold
  - if over the threshold, ask for the true label

PROBLEMS for streaming data

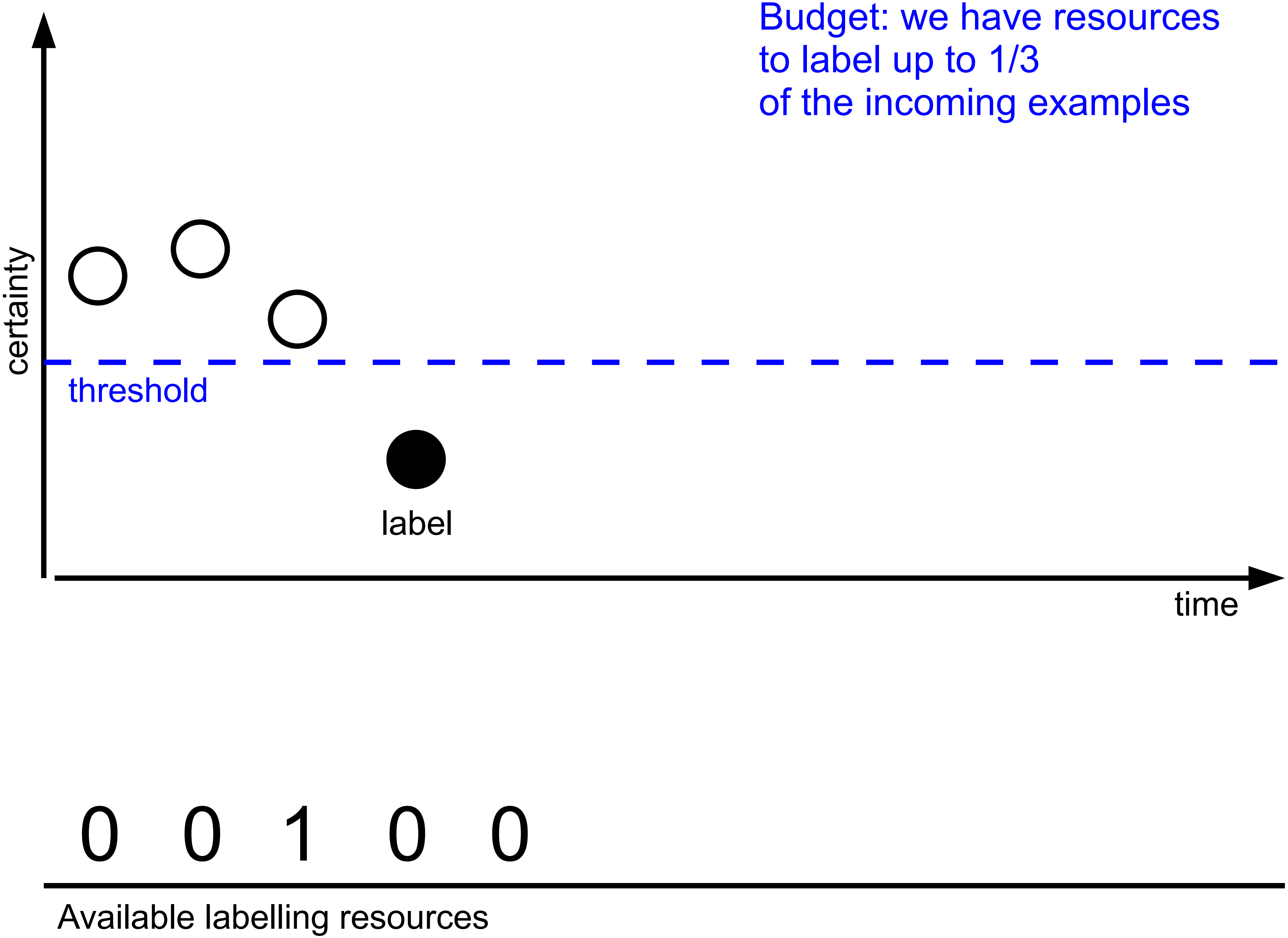
data is changing, models need to evolve

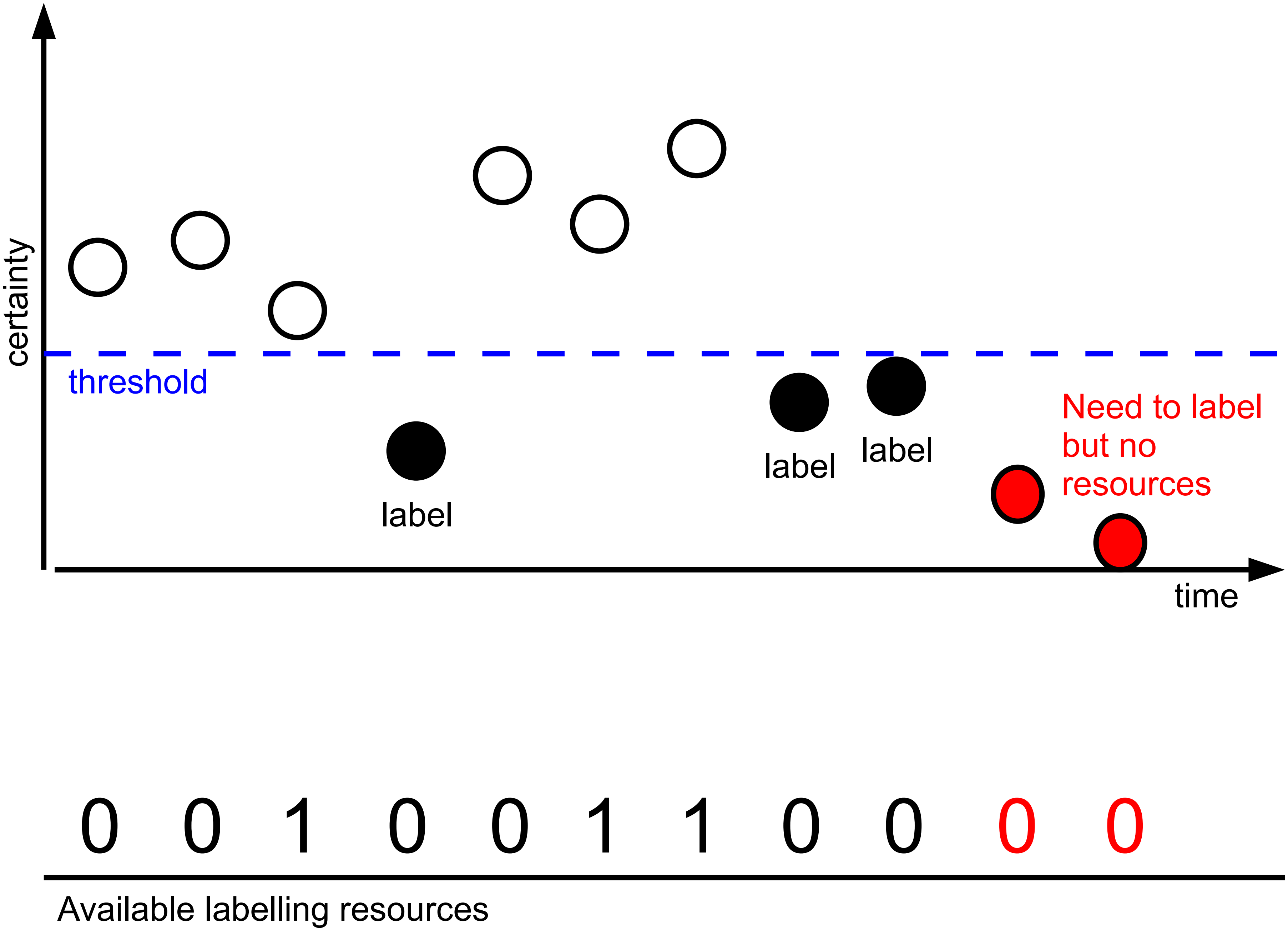
if the threshold is fixed,
model becomes confident, stops learning,
fails to notice changes and fails to adapt

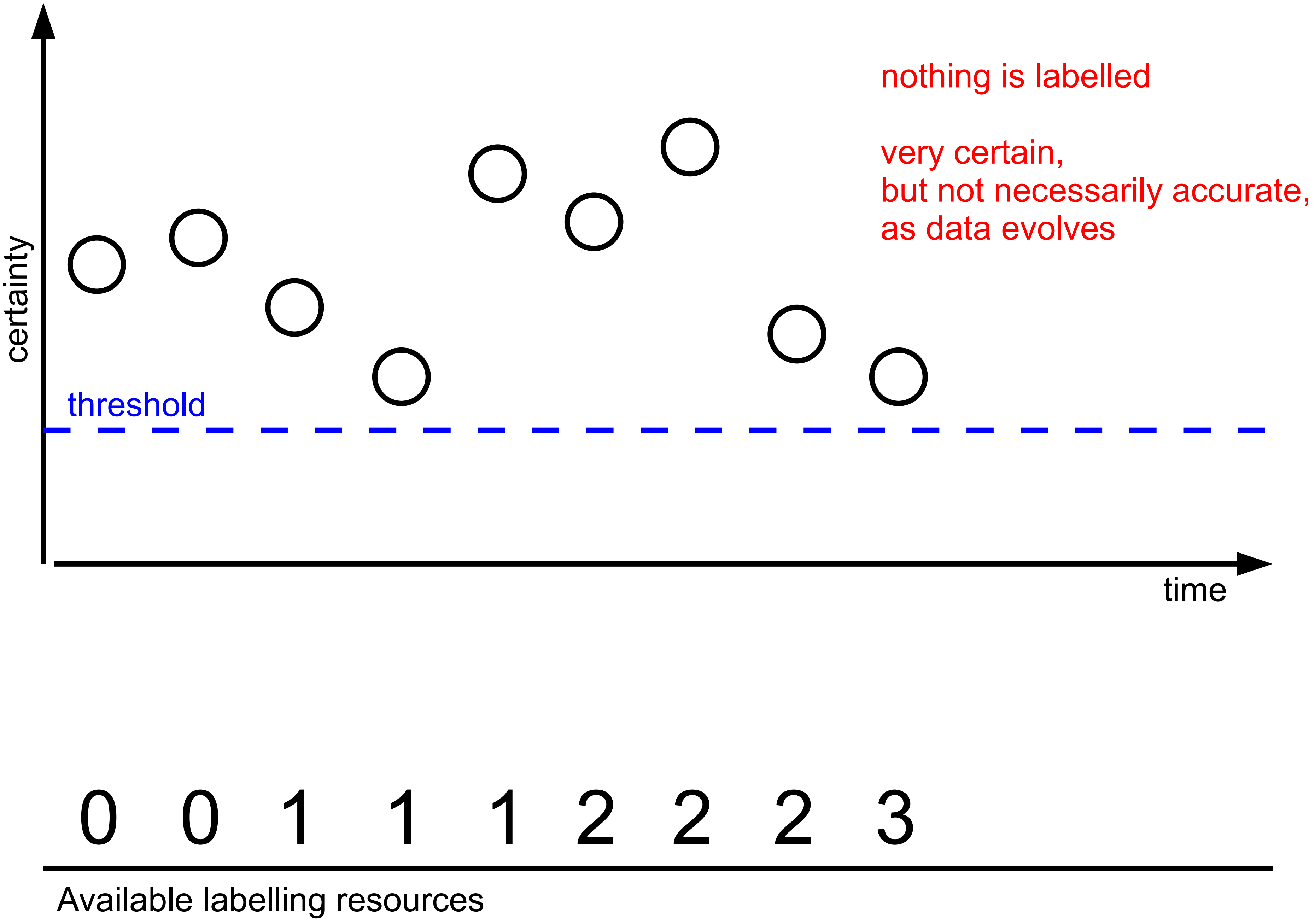
What is needed?

- In data streams
  - Changes may happen at **any time**
  - **Requirement 1**
    - we should ask for labels over time in a balanced way

Budget: we have resources
to label up to 1/3
of the incoming examples

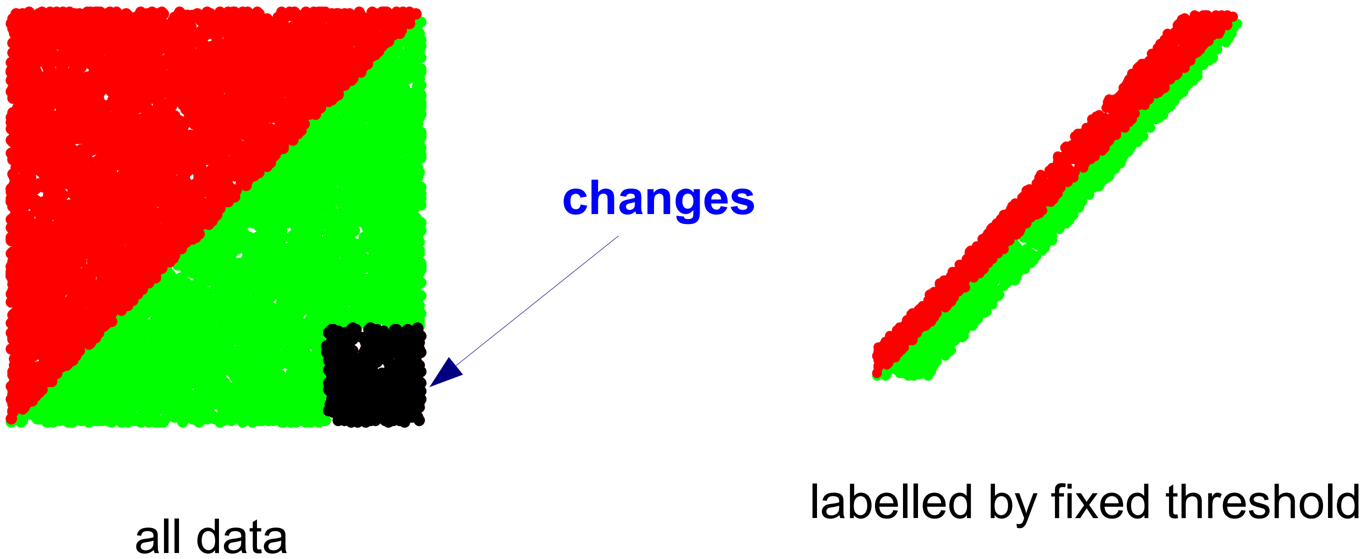



certainty
threshold
label
label
label
Need to label
but no
resources
time
0 0 1 0 0 1 1 0 0 0 0
Available labelling resources

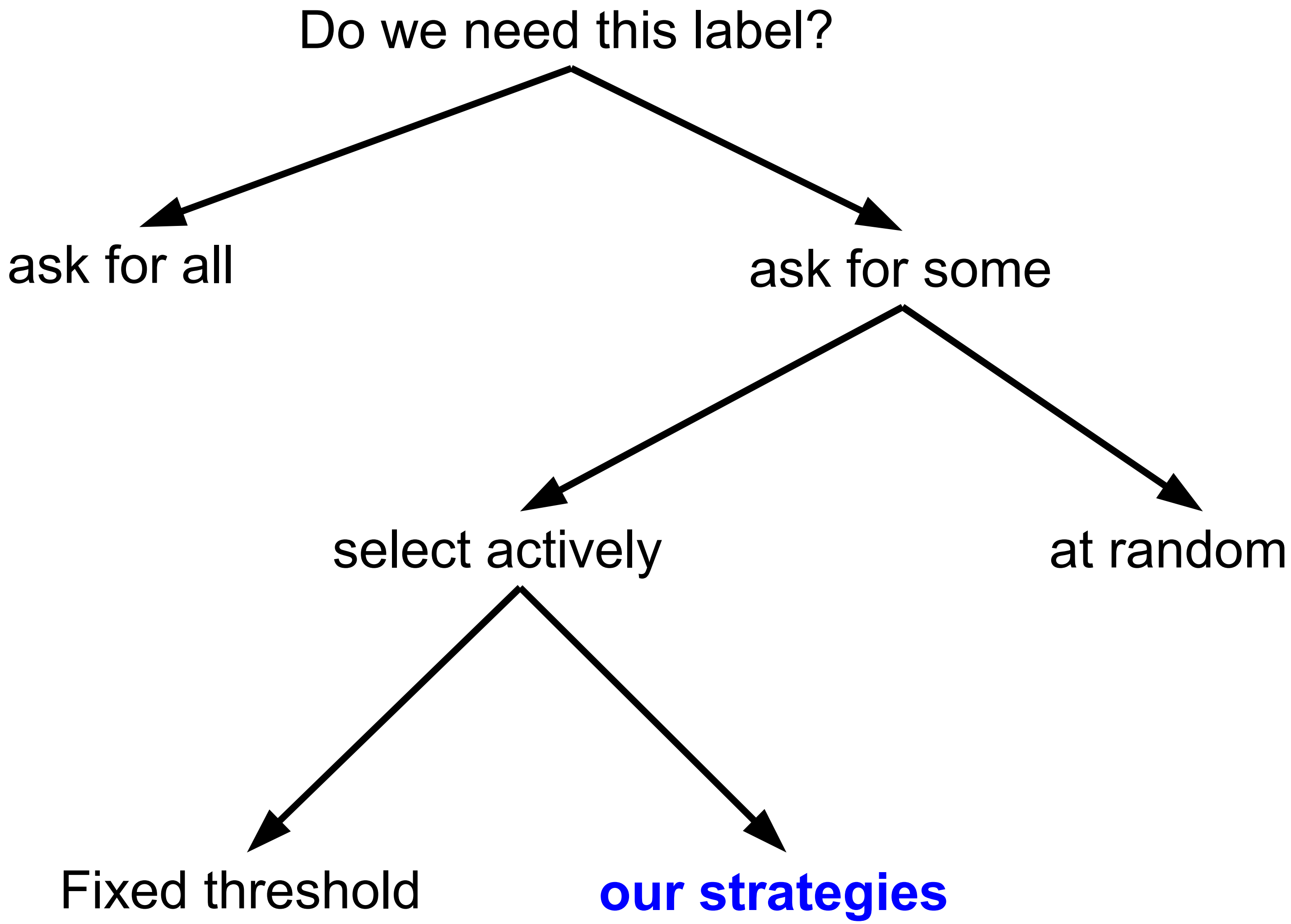

nothing is labelled
very certain,
but not necessarily accurate,
as data evolves
certainty
threshold
time
0 0 1 1 1 2 2 2 3
Available labelling resources

What is needed?

- In data streams
  - Changes may happen at **any time**
  - **Requirement 1**
    - we should ask for labels over time in a balanced way
- Changes may happen **anywhere**
  - **Requirement 2**
    - given enough time, we should ask label for any data point
  - otherwise, we may never detect changes in some regions, and model will never adapt



Changes in the regions where classifier is very certain should not be missed


Do we need this label?
ask for all
ask for some
select actively
at random
Fixed threshold
our strategies

What is needed?

- **Requirement 1**
  - we should ask for labels over time in a balanced way
- *we propose*: adaptive threshold

- **Requirement 2**
  - given enough time, we should ask label for any data point
- *we propose*: add randomization to the threshold

Adaptive uncertainty strategy

- Receive example X_t and a prediction y^*_t
- If labelling *budget* is available [u/t < B]
  - If *uncertainty* of X_t is greater than threshold [$P(y^*_t|X_t) < K$]
    - ask for the *true label* y_t
    - update the model with (X_t, y_t), increment budget counter u=u+1
    - shrink the threshold [K = K(1 – s)]
  - else
    - expand the threshold [K = K(1 + s)]

Requirement 1

balances labelling budget over infinite **time**

Randomized uncertainty

- Receive example X_t and a prediction y^*_t
- If labelling *budget* is available [u/t < B]
  - If *uncertainty* of X_t is greater than *randomized* threshold
    [$P(y^*_t|X_t) < K_{randomized}$, $K_{randomized} = Kv$, where $v \sim N(1,d)$]
    - ask for the *true label* y_t
    - update the predictive model with (X_t,y_t), u=u+1
    - shrink the threshold [K = K(1 – s)]
  - else
    - expand the threshold [K = K(1 + s)]

Requirement 2

balances labelling
to cover
the instance **space**

empirical results

MOA

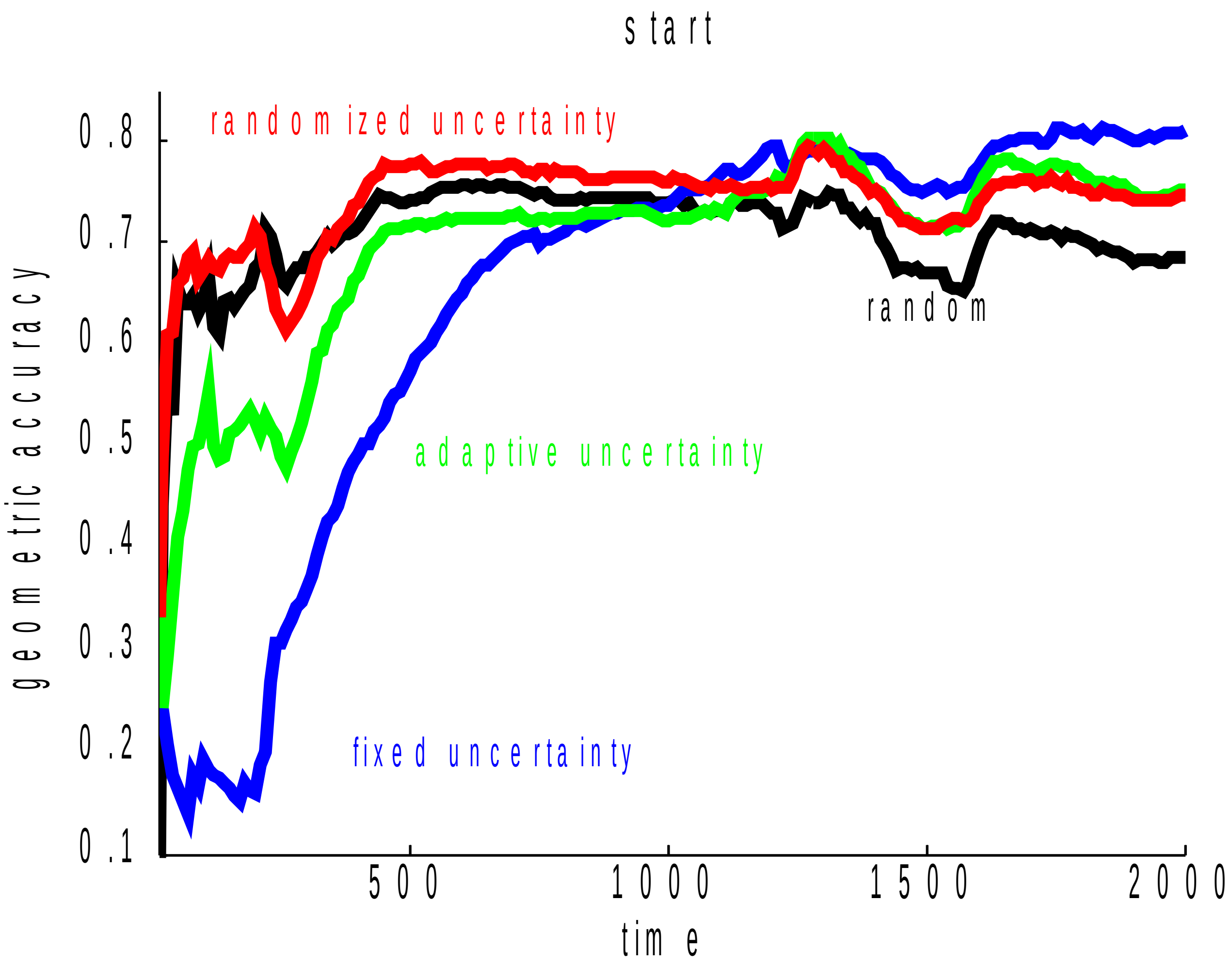
- {M}assive {O}nline {A}nalysis is a framework for online learning from data streams
- It is closely related to WEKA
- It includes a collection of online and offline algorithms and tools for evaluation
  - classification
  - clustering
- Easy to extend
- Easy to design and run experiments

Experimental evaluation

- Strategies
  - random sampling, fixed uncertainty, adaptive uncertainty, randomized (adaptive) uncertainty, selective uncertainty
- Adaptive learner: DDM (Gama et al, 2004)
- Evaluation: accuracy over a dataset, accuracy in time
- Datasets
  - synthetic (hyperplane)
  - real-life textual with our labels (IMDB-E, IMDB-D, Reuters)
  - real-life with expected changes (Electricity, Cover type, Airlines)

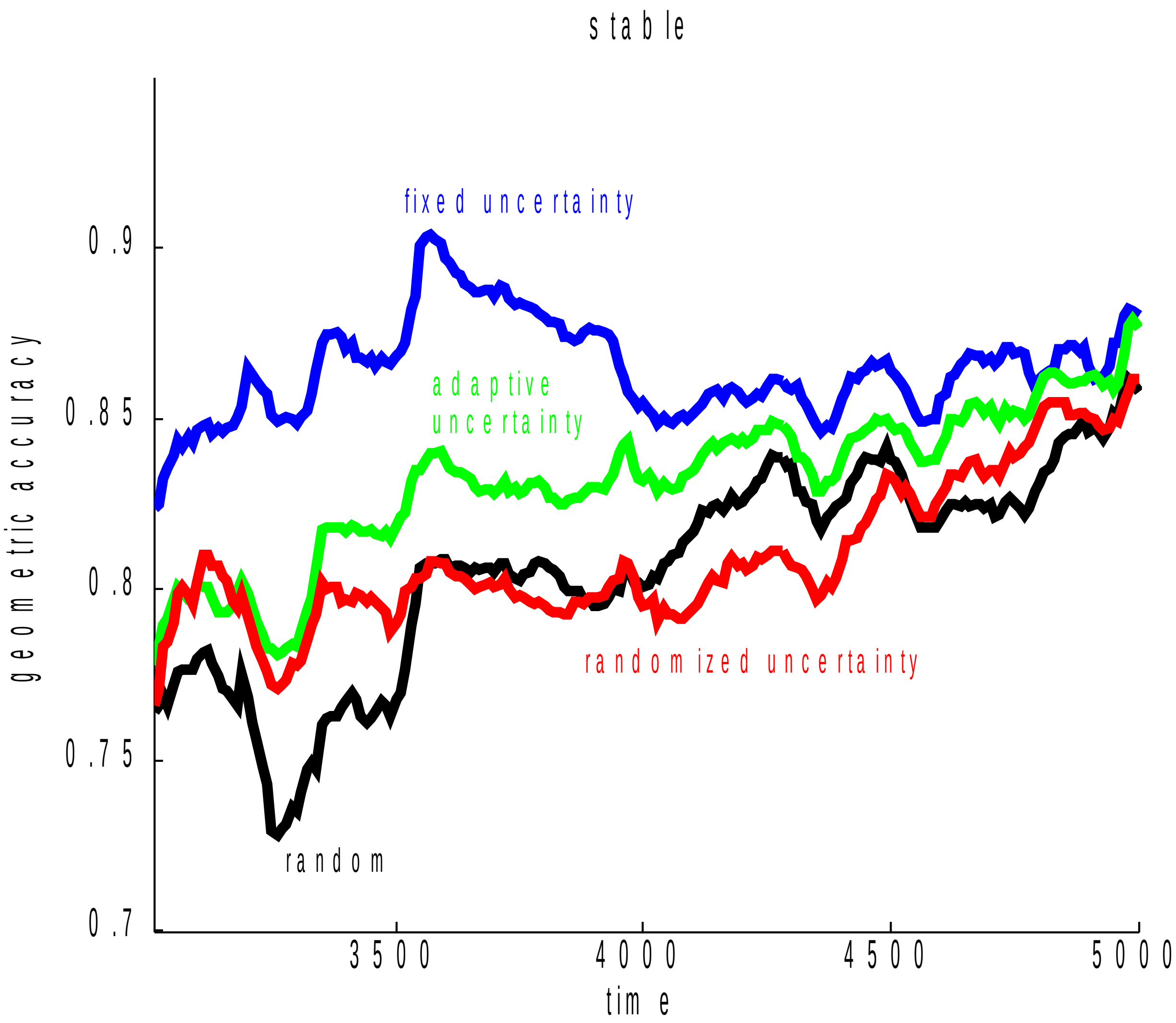
- The results demonstrate advantages of our strategies against fixed threshold and random sampling in the data stream settings where data is evolving

REUTERS data



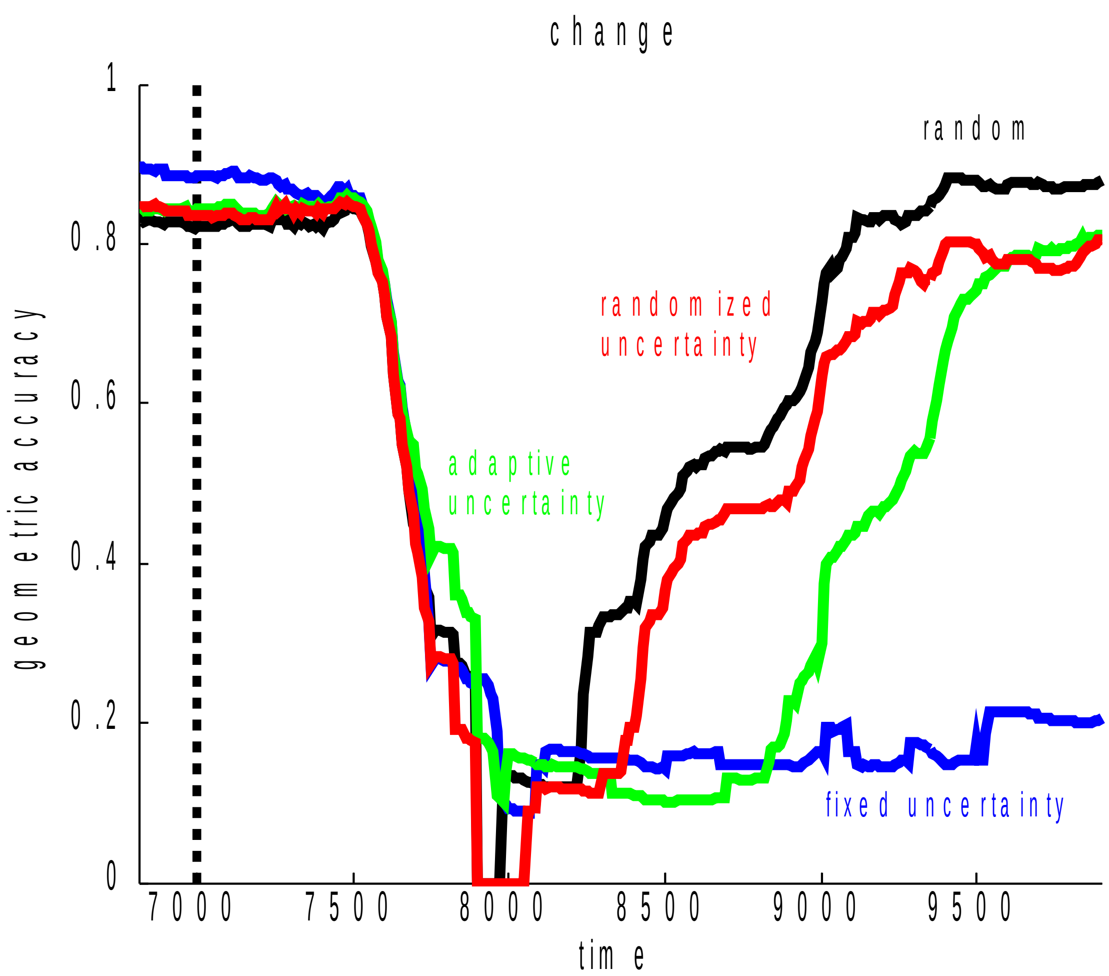
Fixed uncertainty becomes very confident in its predictions and adapts slowly

REUTERS data



Fixed uncertainty and adaptive uncertainty do not waste labelling budget for querying very certain examples, thus is more accurate when there are no changes in data

REUTERS data



Fixed uncertainty fails to adapt, strategies with randomization adapt faster conclusion

Conclusion

- We explore active learning in the strict data stream settings
- We equip active learning strategies with mechanisms to
  - control distribution of labelling budget over infinite time
  - trade off labelling some of the uncertain examples for labelling very confident examples in order to capture changes anywhere in the input space
- Empirical results suggest that our strategies
  - have an advantage *in accuracy* against fixed threshold and random sampling
  - in data stream settings where data evolves over time
- Adaptive uncertainty is preferred when mild changes are expected, randomized uncertainty if preferred for data with strong changes

Thanks!

Acknowledgements

Part of the research leading to these results has received funding from the EC within the Marie Curie Industry and Academia Partnerships and Pathways (IAPP) programme under grant agreement no. 251617.