Netherlands Organisation for Scientific Research

NL Agency
Ministry of Economic Affairs

Technische Universiteit
Eindhoven
University of Technology

Predictive Analytics on Evolving Data Streams:

Anticipating and adapting to changes in known and unknown contexts

Mykola Pechenizkiy
http://www.win.tue.nl/~mpechen/

Connected Intelligence Center, UTS, Sydney
13 February 2017

Outline

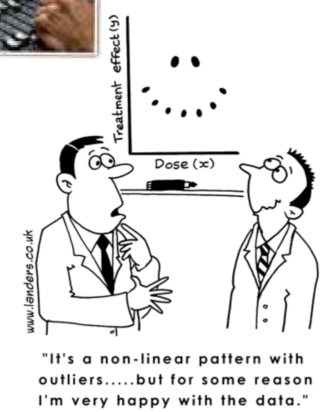
- What is predictive analytics?
  - Applications on streaming data
- Evolving data: known vs. hidden contexts
  - Context-awareness and concept drift handling
- The main-stream approaches and recent development for handling concept drift
  - Back to context awareness
- Outlook and take home messages

Predictive Modeling Tasks

Use some variables to predict unknown or future values of other variables (labeled data needed)

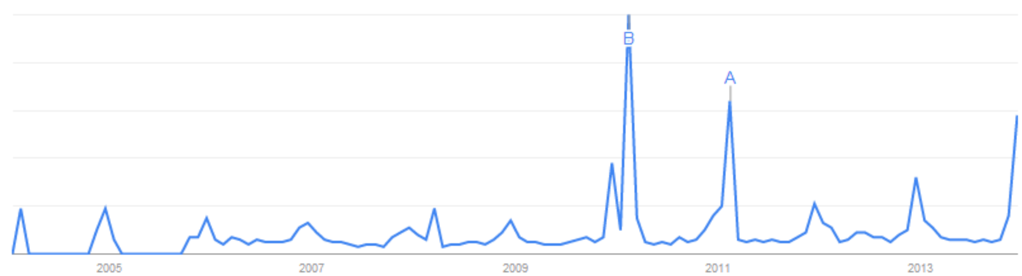
- Classification
  - expert or novice user?
  - information need: navigational or explorative?

- Regression
  - What score will this user give to this product?
  - How much is he ready to pay for it?

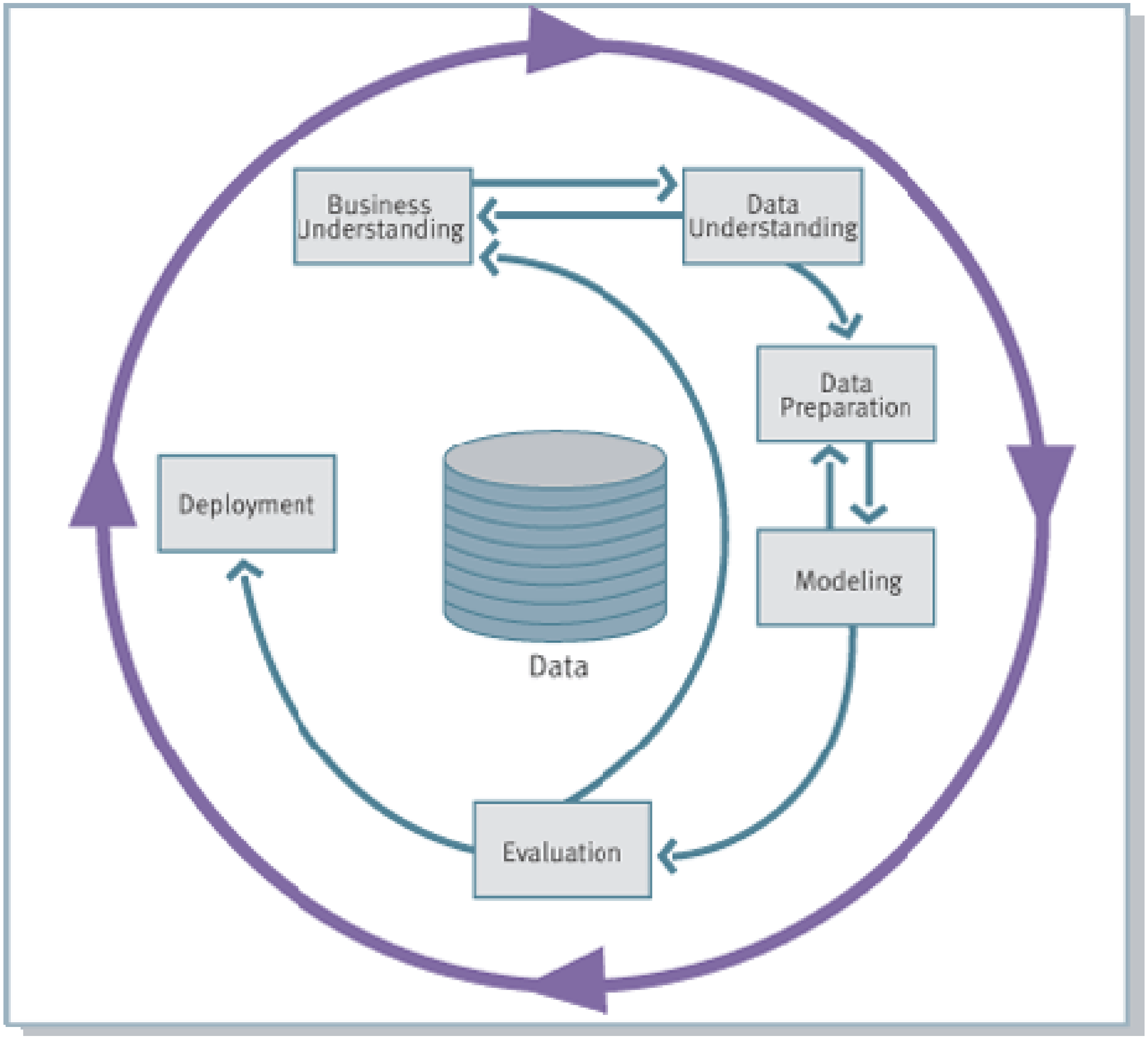


- Ranking and preference learning
  - What is more important/interesting for a user?

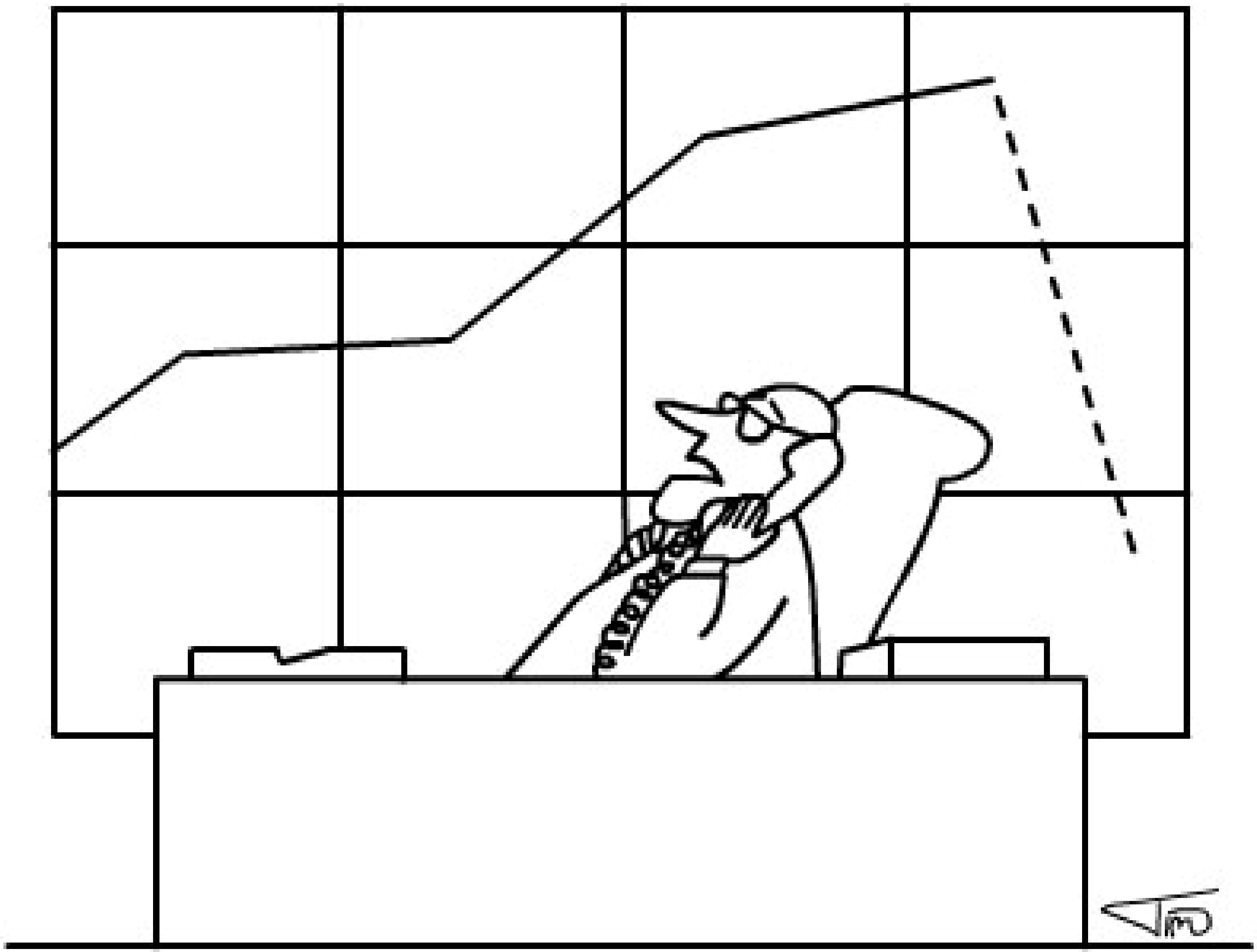
- Timeseries prediction
  - What would be CTR for a news item next day?



Predictive Analytics: CRISP-DM 1.0

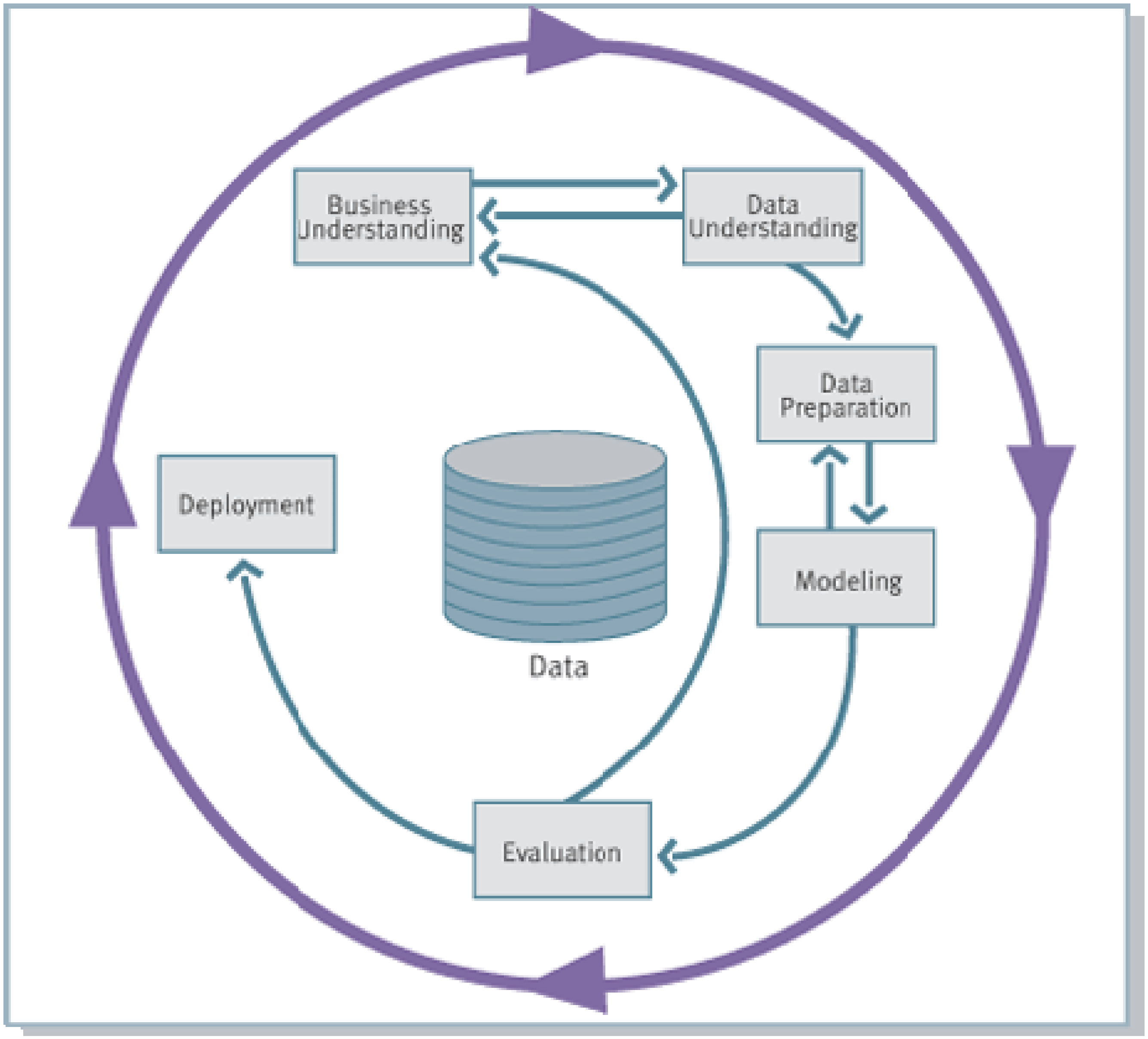


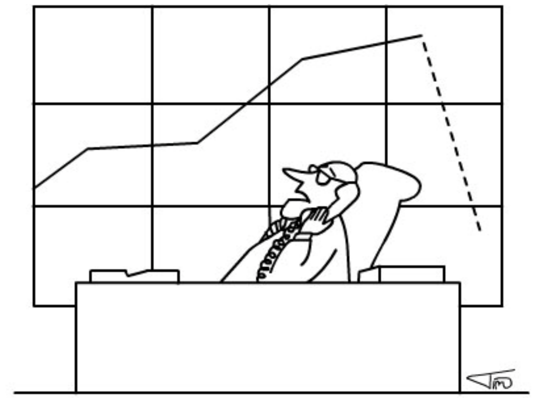
Predictive Analytics: CRISP-DM 1.0



"BI tech support? The predictive analysis system is giving the wrong answer again—can you please fix it?... "

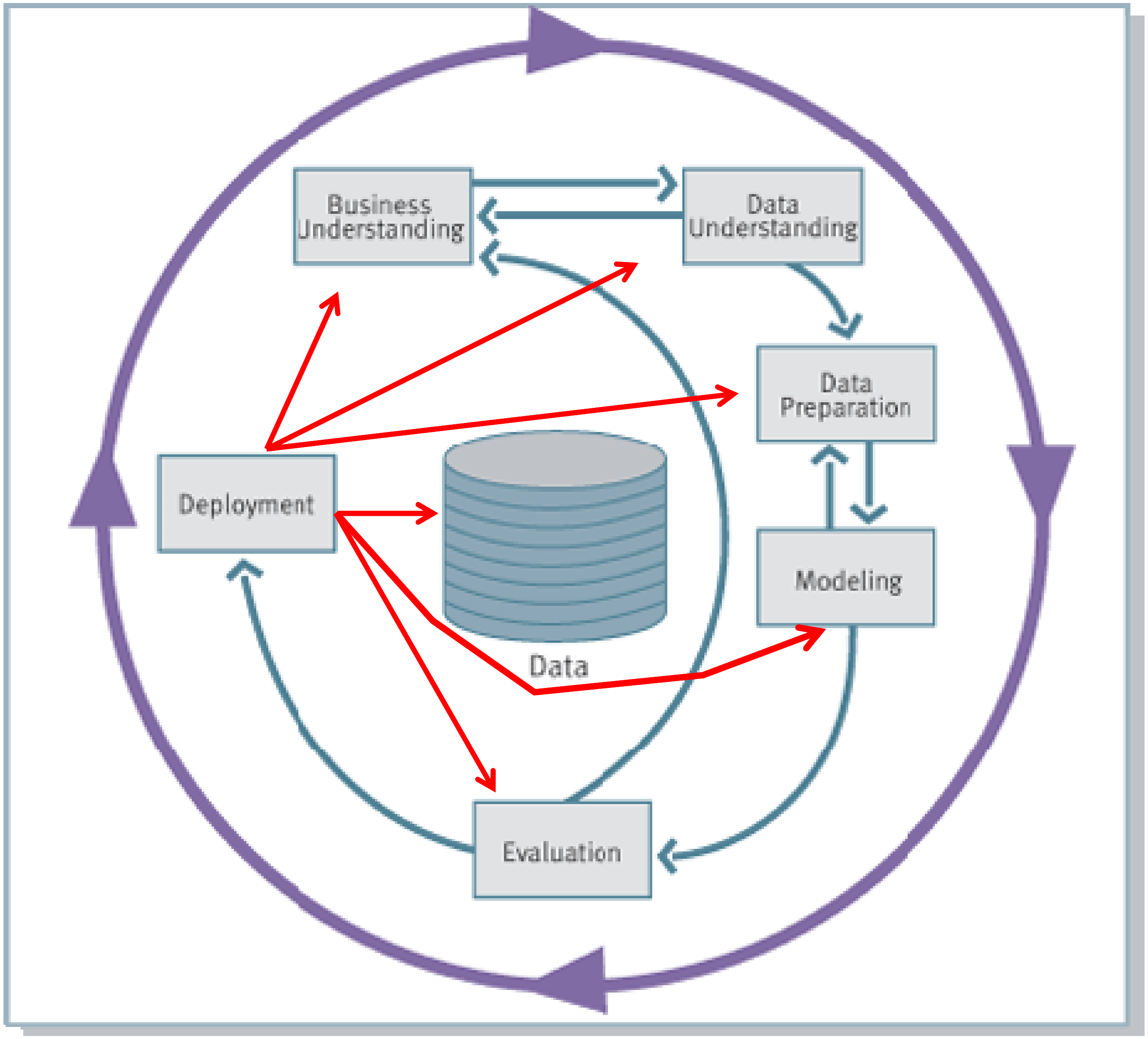
Predictive Analytics: CRISP-DM 1.0





"BI tech support? The predictive analysis system is giving the wrong answer again—can you please fix it?... "

Predictive Analytics: CRISP-DM 2.0



Evolving data

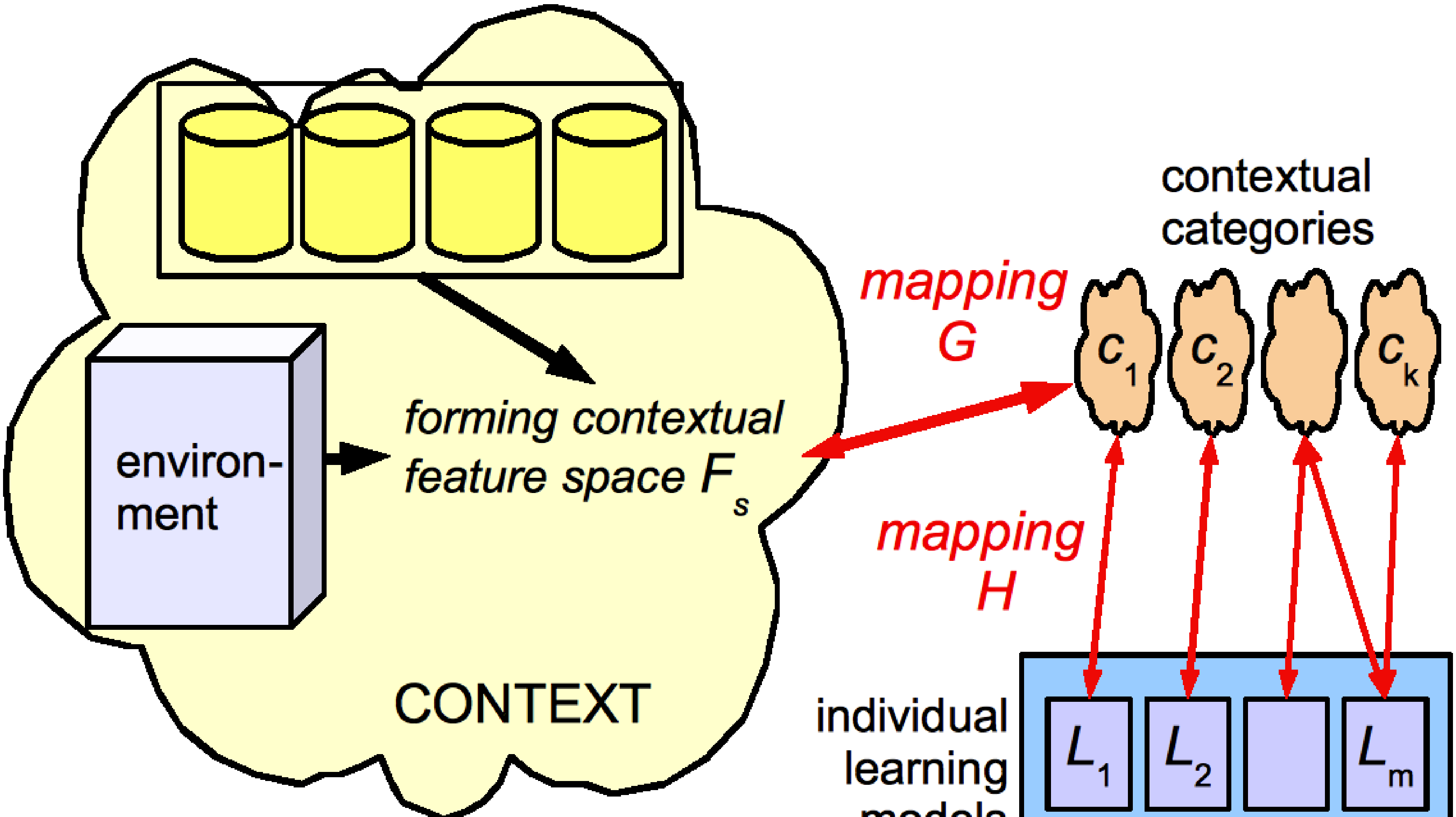
Performance monitoring

Model adaptation

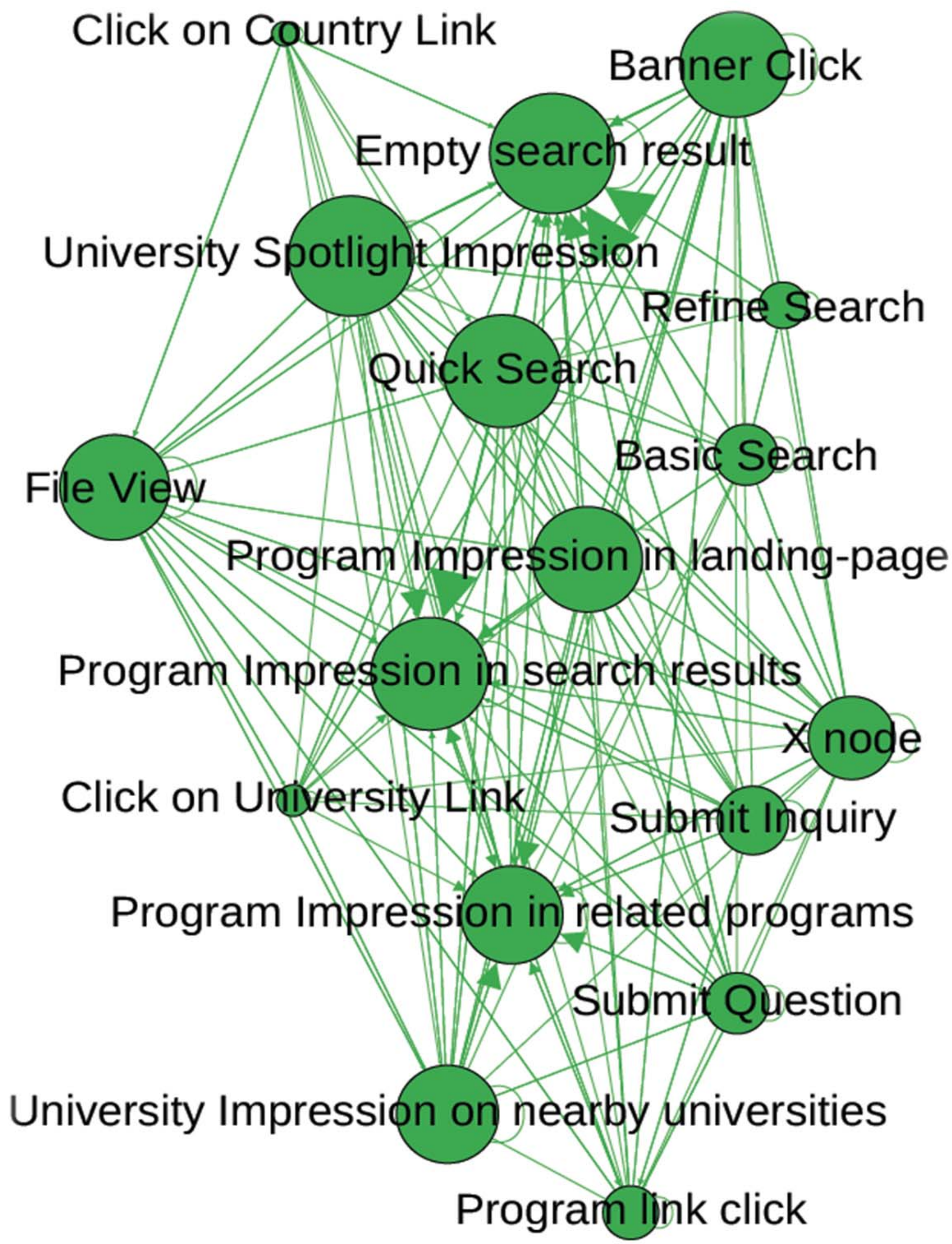
Context-awareness

Handling concept drift

Design: (Re-)Learning Classifiers & Context



User Navigation Graph

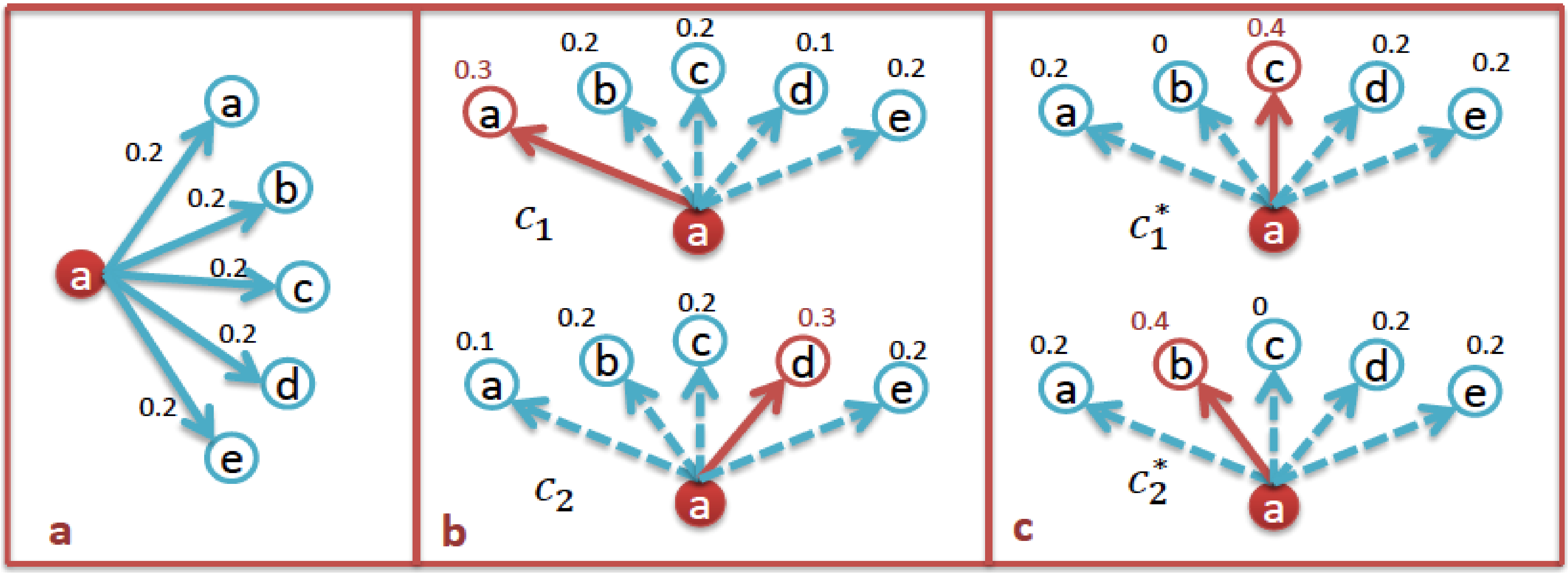


Motivation for Contextual Markov Models

Useful Contexts:

$E[M] < p_{c1} * E[M_{c1}] + p_{c2} * E[M_{c2}]$

Why should it help?

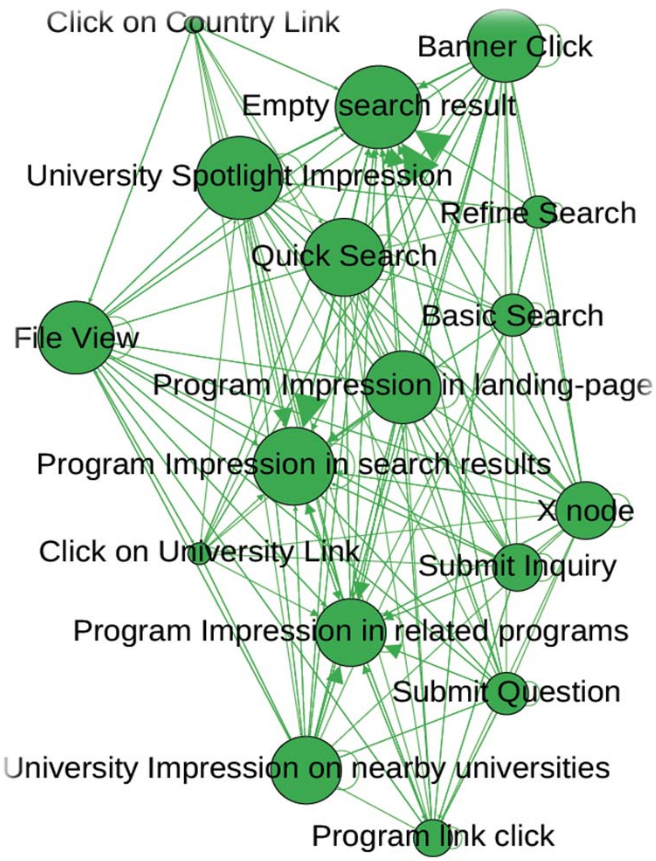


Explicit contexts (user location)

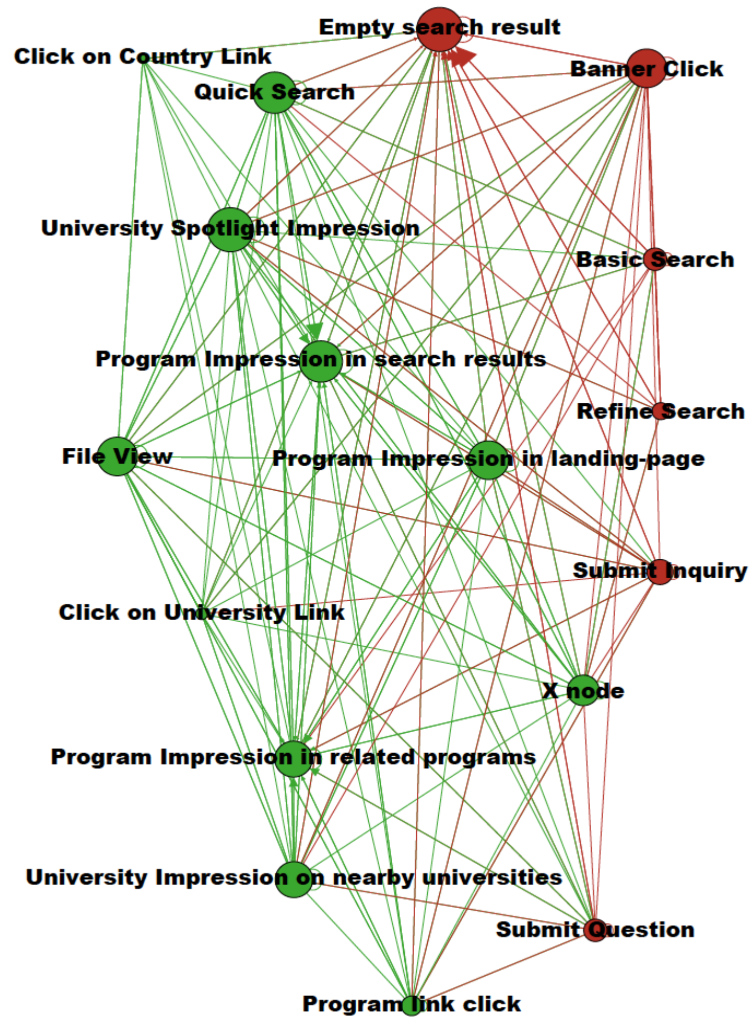
Implicit contexts (inferred from clickstream)

Implicit Context

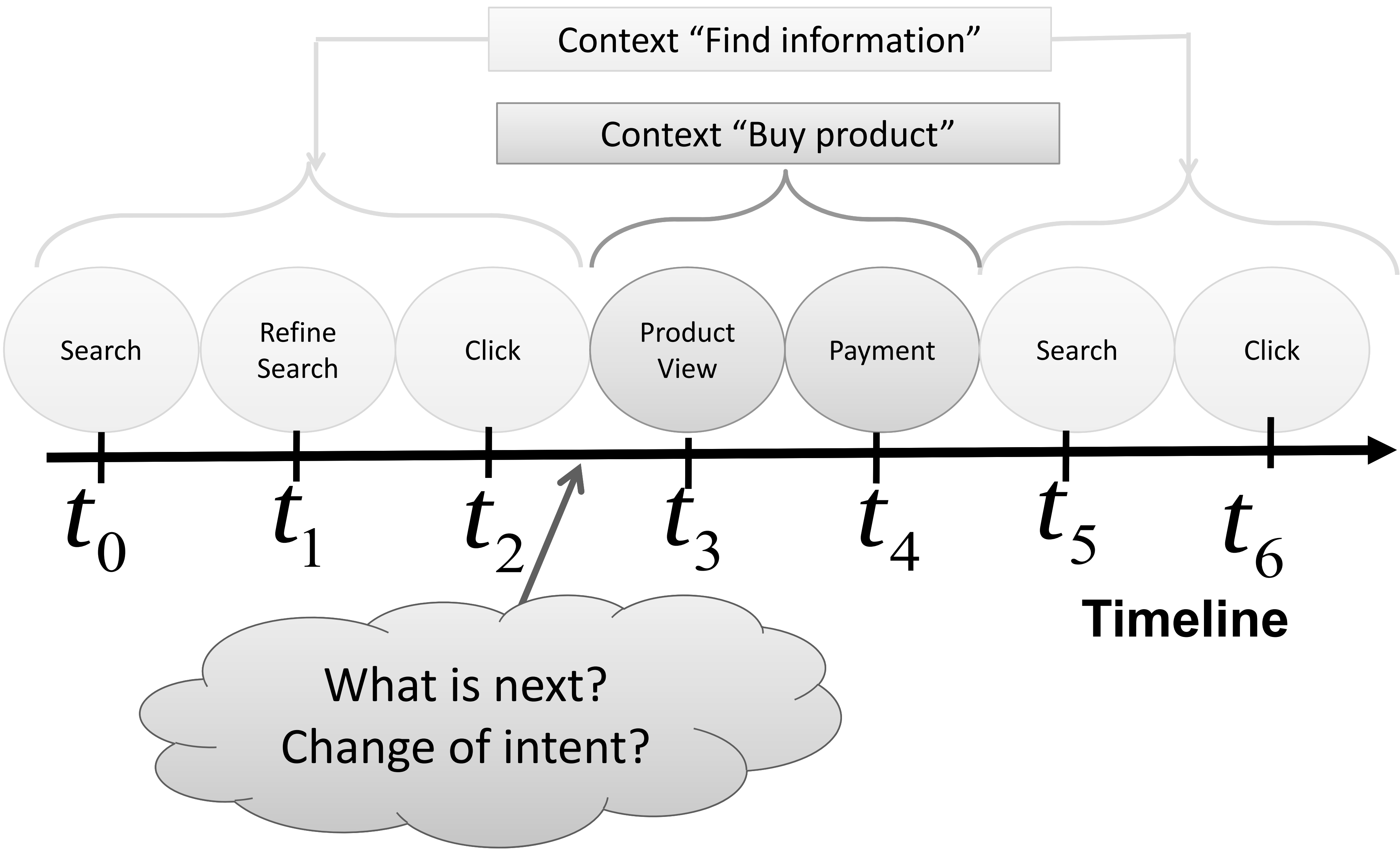
C = user type	c_1 = Novice users	c_1 = Experienced users



Discover clusters in the graph using community detection algorithm

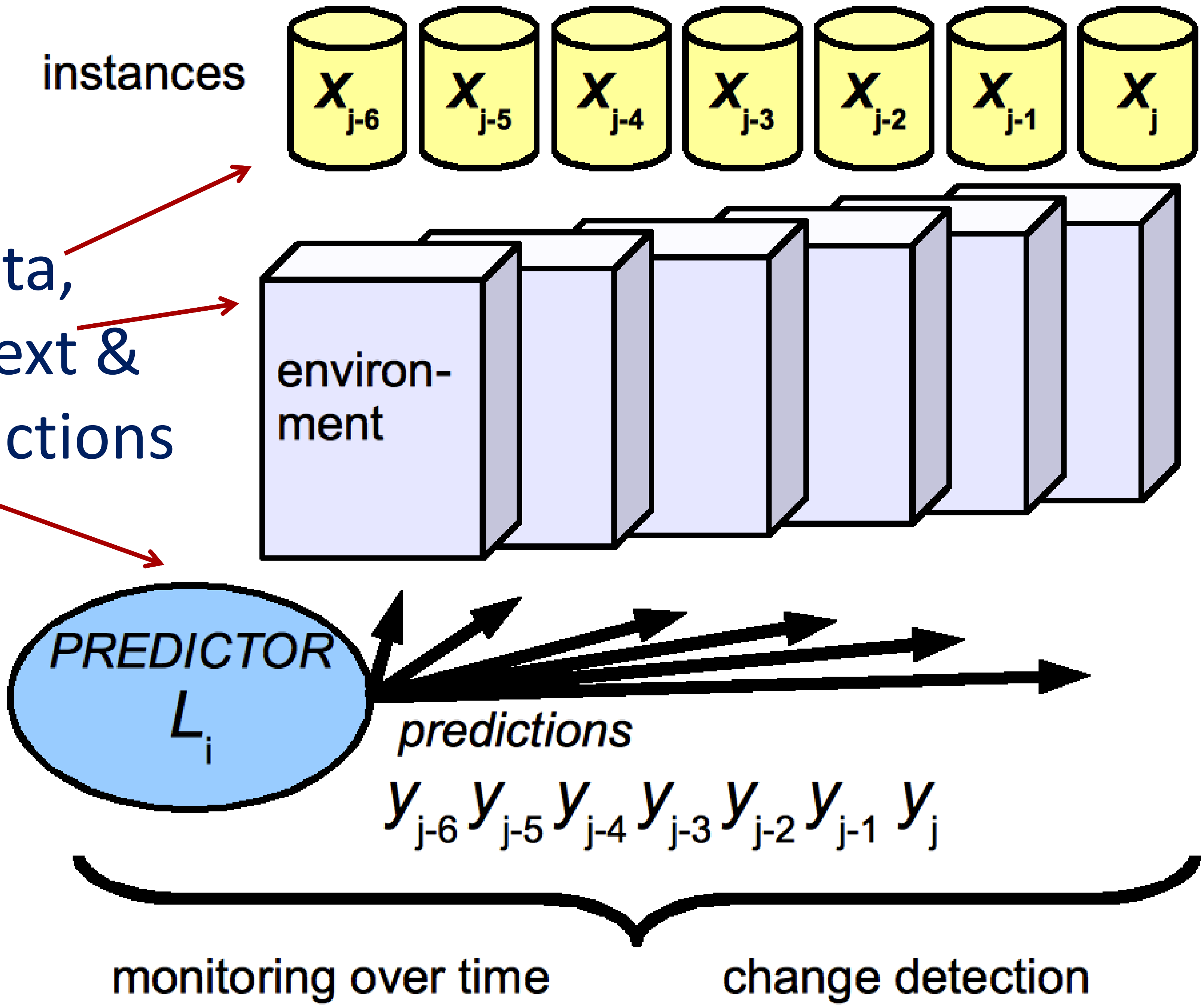


Change of Intent as Context Switch



Evolving data: Monitoring for Changes

in Data,
Context &
Predictions



Reactive vs. Proactive methods

Monitoring own recent performance

Monitoring for *recurrent contexts*

Monitoring *performance of peers*

Prediction under Concept Drift

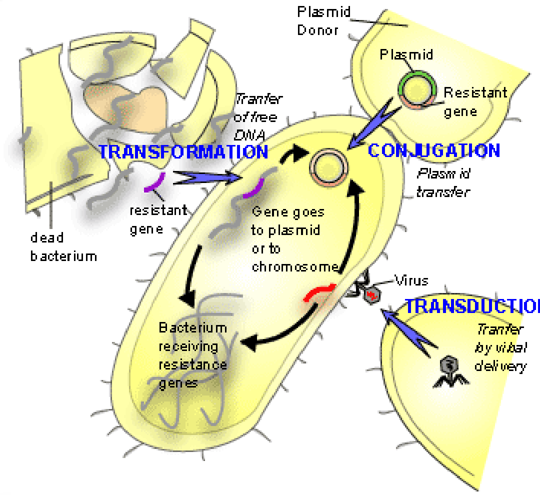
Antibiotic Resistance Prediction:

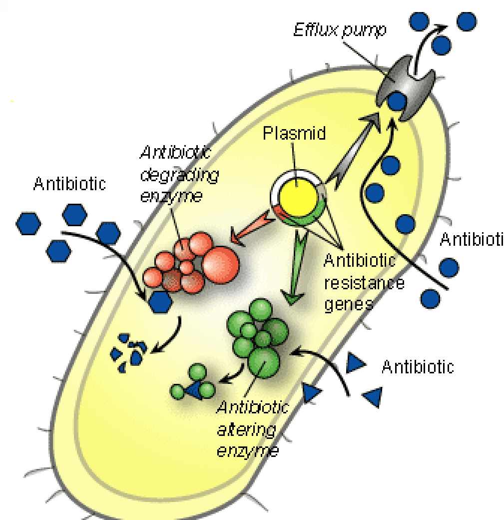
predict the sensitivity of a pathogen to an antibiotic based on data about the antibiotic, the isolated pathogen, and the demographic and clinical features of the patient.

date	sex	age	isNew	days_total	days_ICU	main_dept	pathogen data	antibiotic data	sensitivity
22.1.2002	m	25	1	171	81	9	...	...	3
22.1.2002	m	25	1	171	81	9	...	...	3
22.1.2002	m	25	1	171	81	9	...	...	3
22.1.2002	m	25	1	171	81	9	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	1
28.1.2002	f	61	0	261	52	3	...	...	1
28.1.2002	m	25	1	171	81	9	...	...	3
28.1.2002	m	25	1	171	81	9	...	...	3
30.1.2002	m	25	1	171	81	9	...	...	3
8.2.2002	m	30	0	209	209	9	...	...	3
8.2.2002	m	30	0	209	209	9	...	...	1
8.2.2002	m	30	0	209	209	9	...	...	1
11.2.2002	f	0	0	18	0	2	...	...	1
11.2.2002	f	0	0	18	0	2	...	...	1
11.2.2002	f	0	0	18	0	2	...	...	1
new data	...	...	...	...	...	...	...	...	?
new data	...	...	...	...	...	...	...	...	?
new data	...	...	...	...	...	...	...	...	?

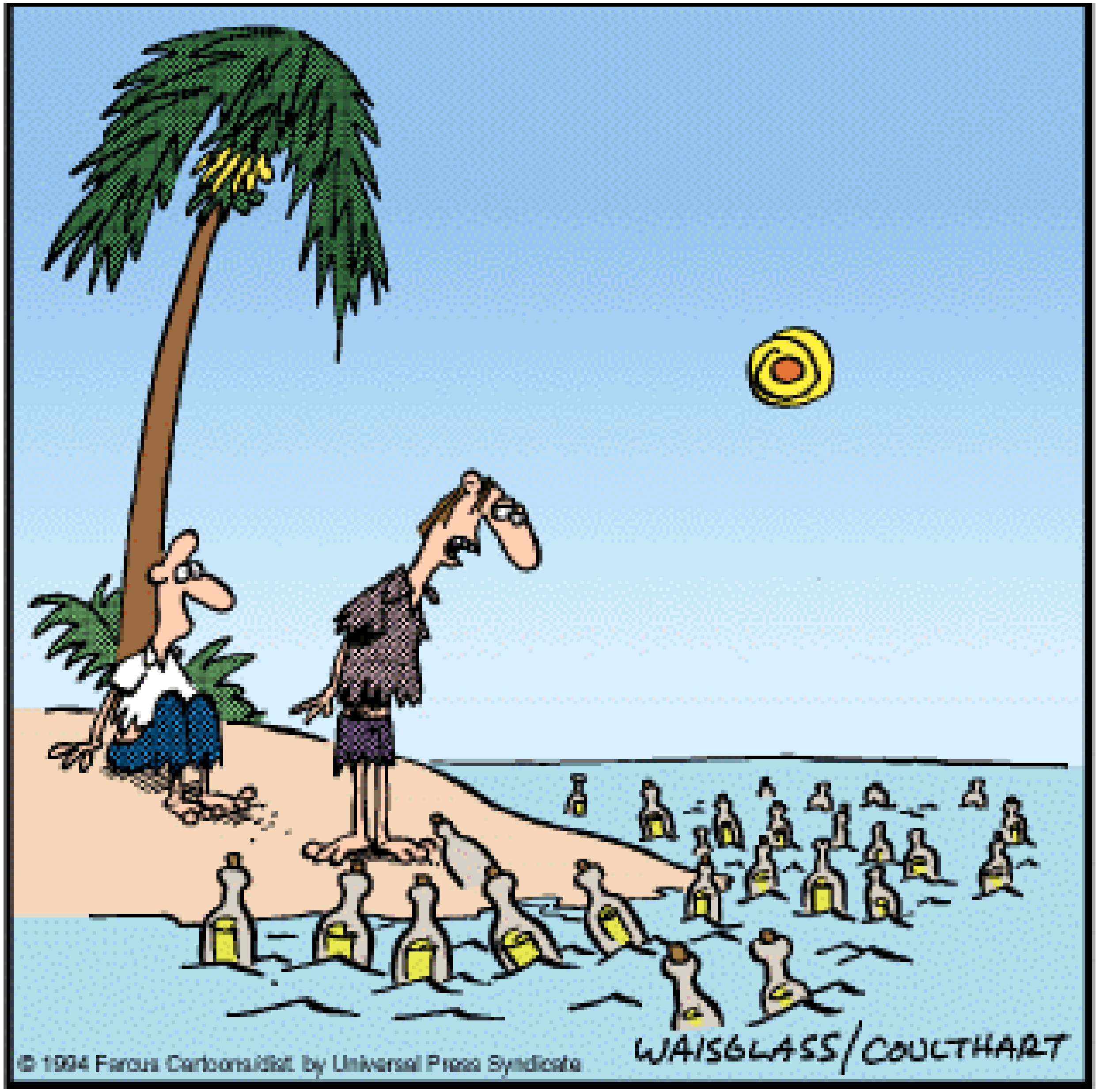
How Antibiotic Resistance Happens

It was on a short-cut through the hospital kitchens that Albert was first approached by a member of the Antibiotic Resistance.



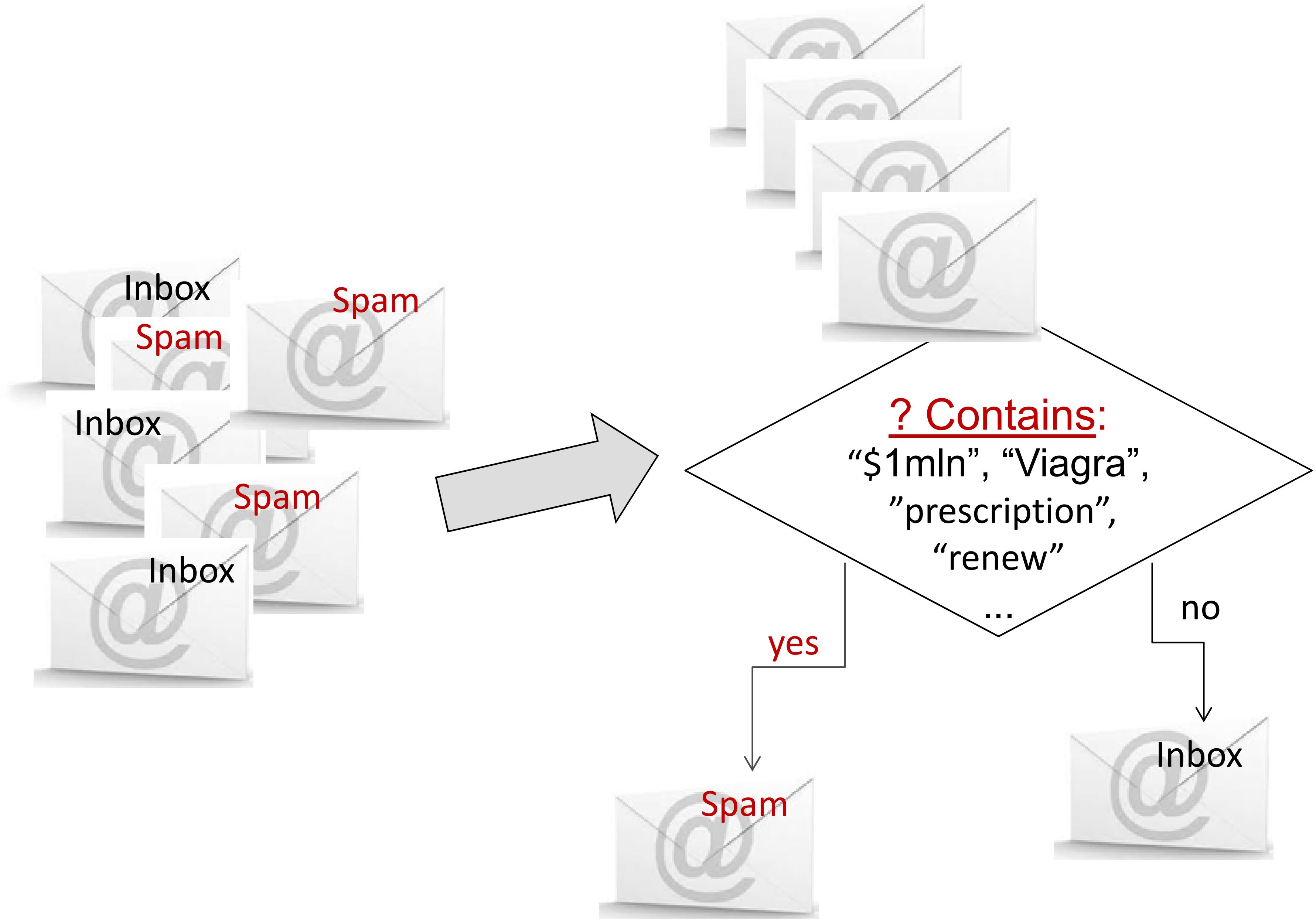


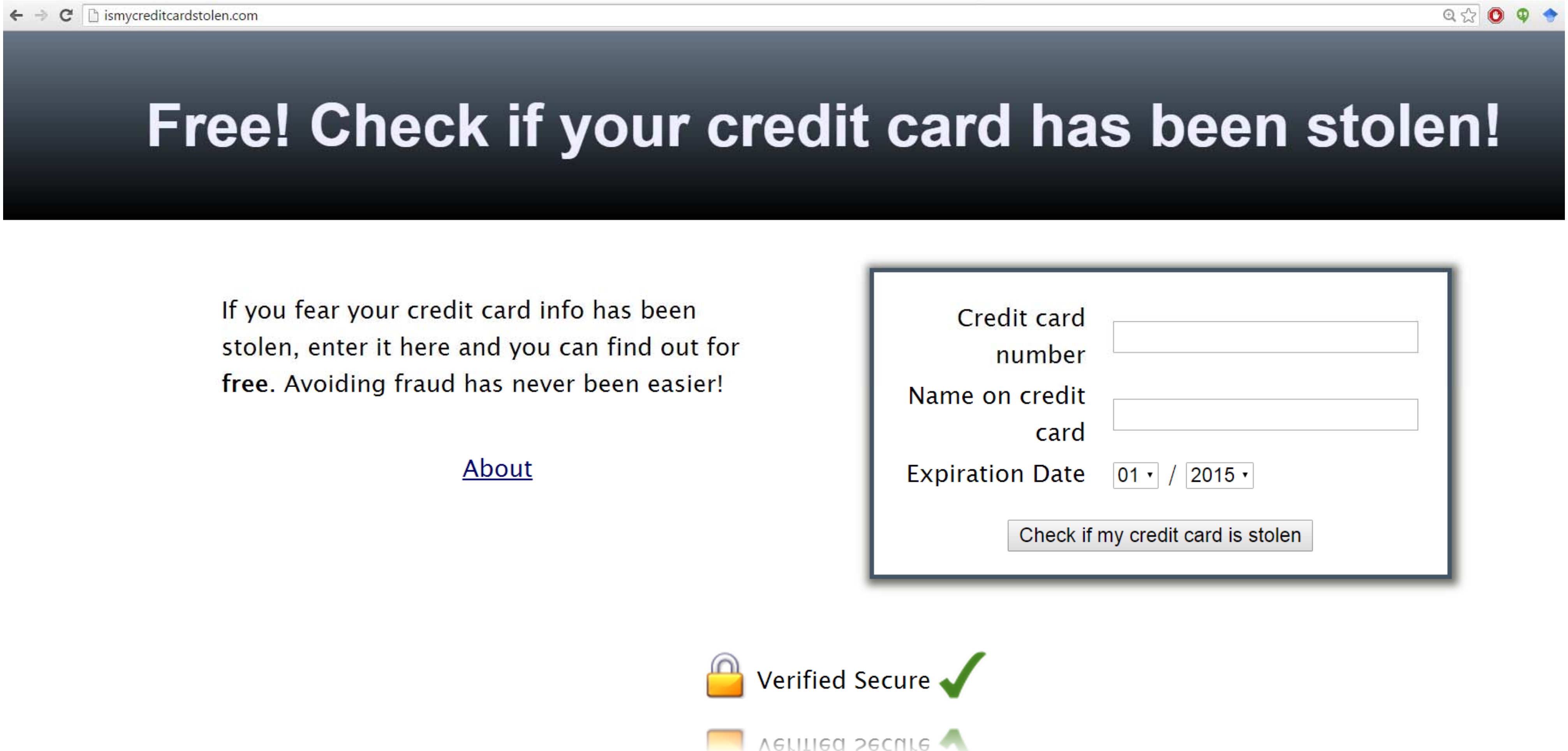
Identifying Worthy Content



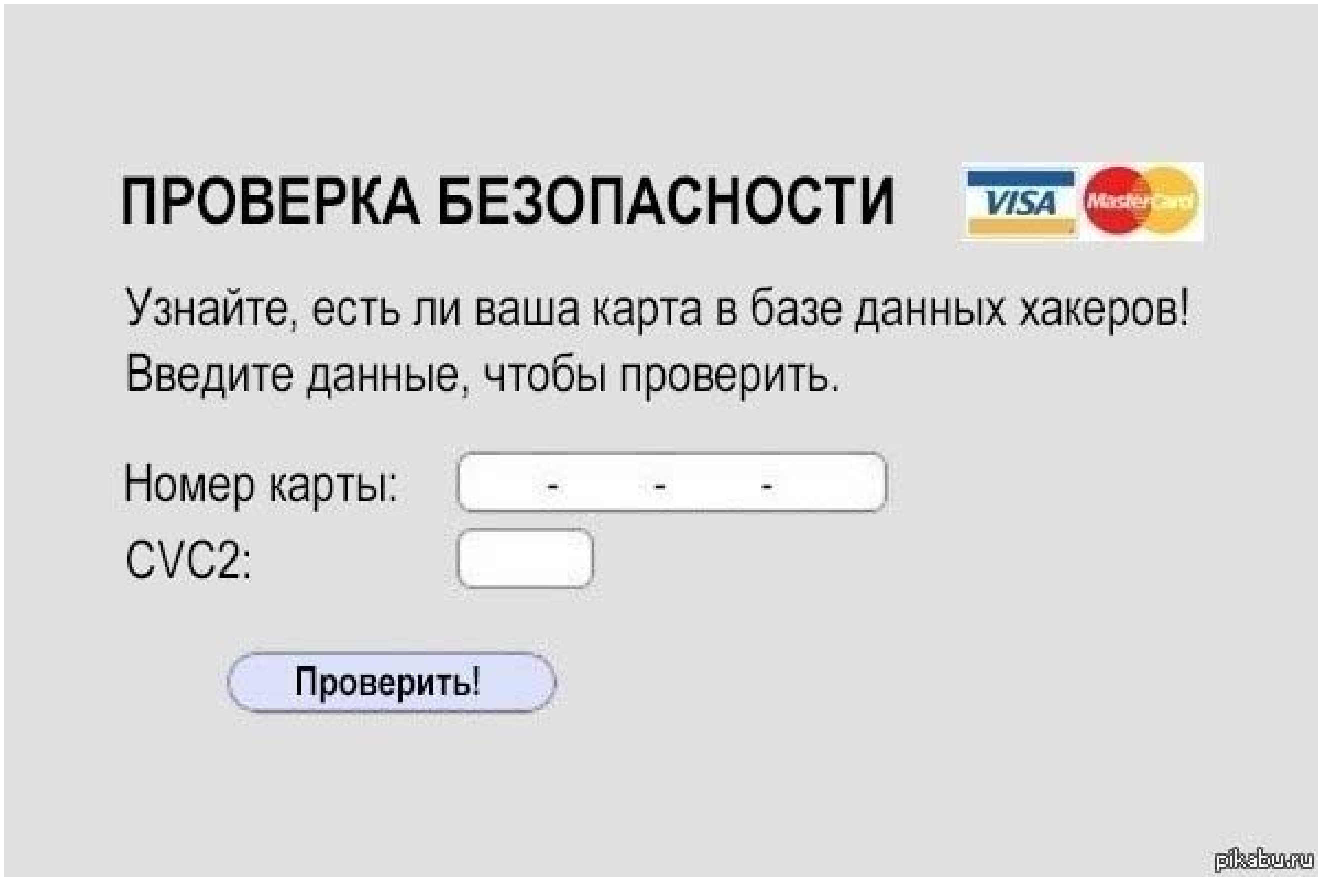
"We must be on a mailing list."

Classify e-mails into “Spam” vs. “Inbox”




ismycreditcardstolen.com
Free! Check if your credit card has been stolen!
If you fear your credit card info has been stolen, enter it here and you can find out for free. Avoiding fraud has never been easier!
About
Credit card number
Name on credit card
Expiration Date
01
/
2015
Check if my credit card is stolen
Verified Secure

Adversary activities



Predictive Analytics on Evolving Data

- Prediction systems need to be adaptive to changes over time to be up to date and useful

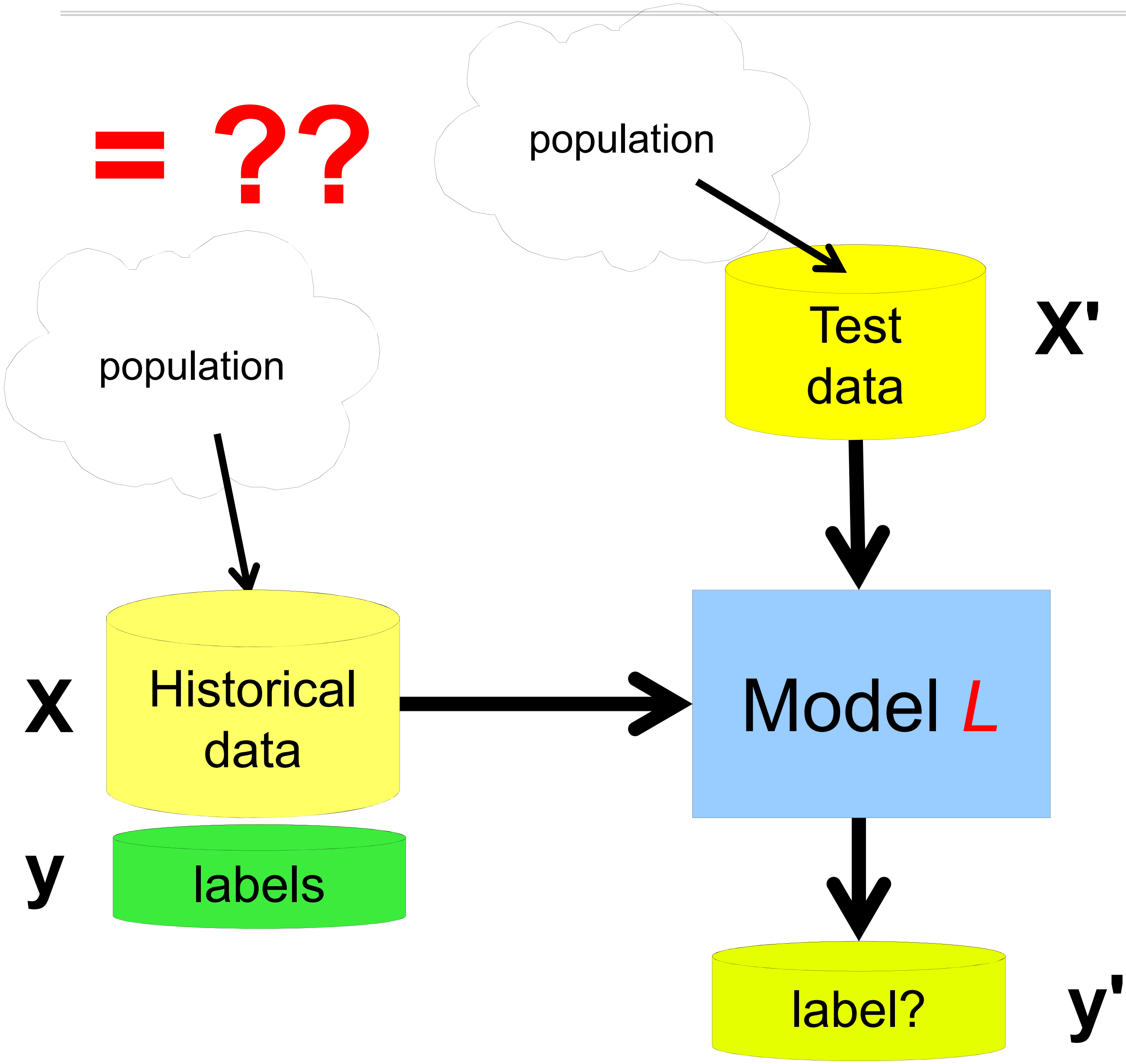
Changes in personal interests or in population characteristics (adaptive news access)

Adversary activities (avoiding spam filters; credit card fraud)

Changes in population characteristics (credit scoring)

Complexity of the environment (driverless cars)

Supervised Learning under Concept Drift



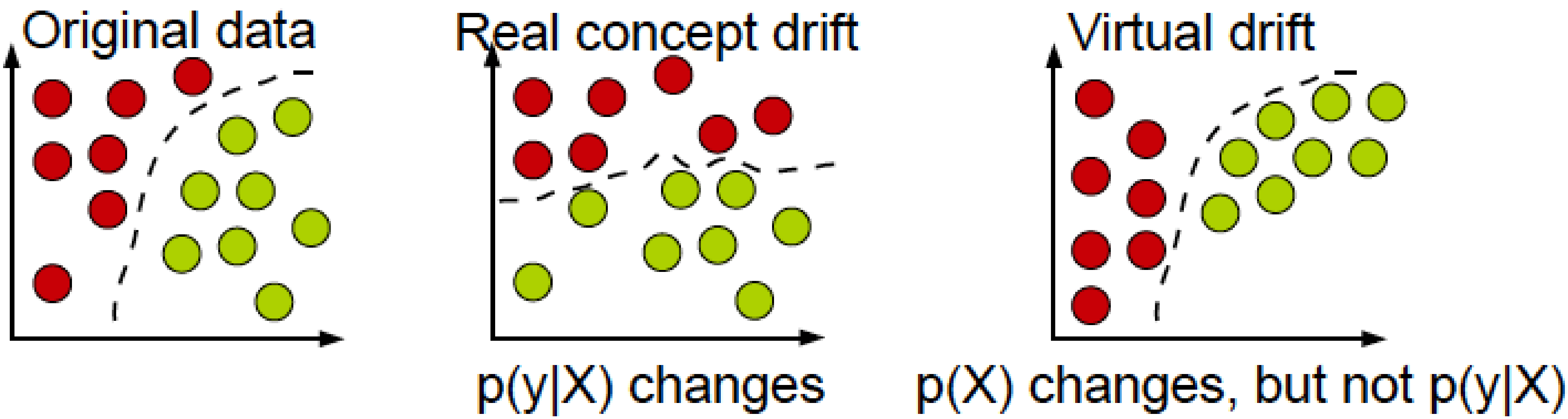
Training:

$$y = L\,(X)$$

Application:

$$y' = L??\,(X')$$

Real vs. Virtual Concept Drifts

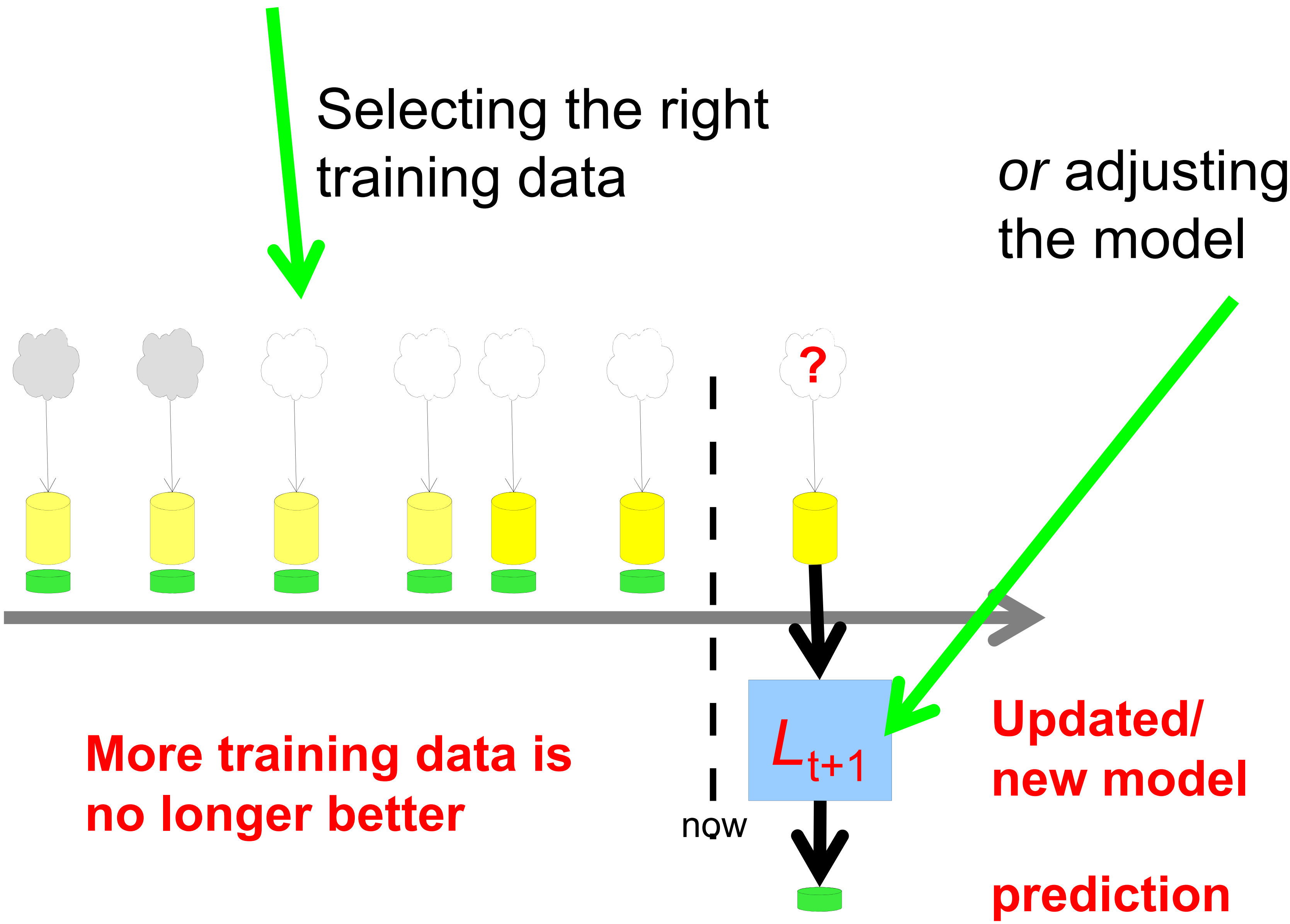


- circles represent instances (X),
- different colors represent different classes (y)

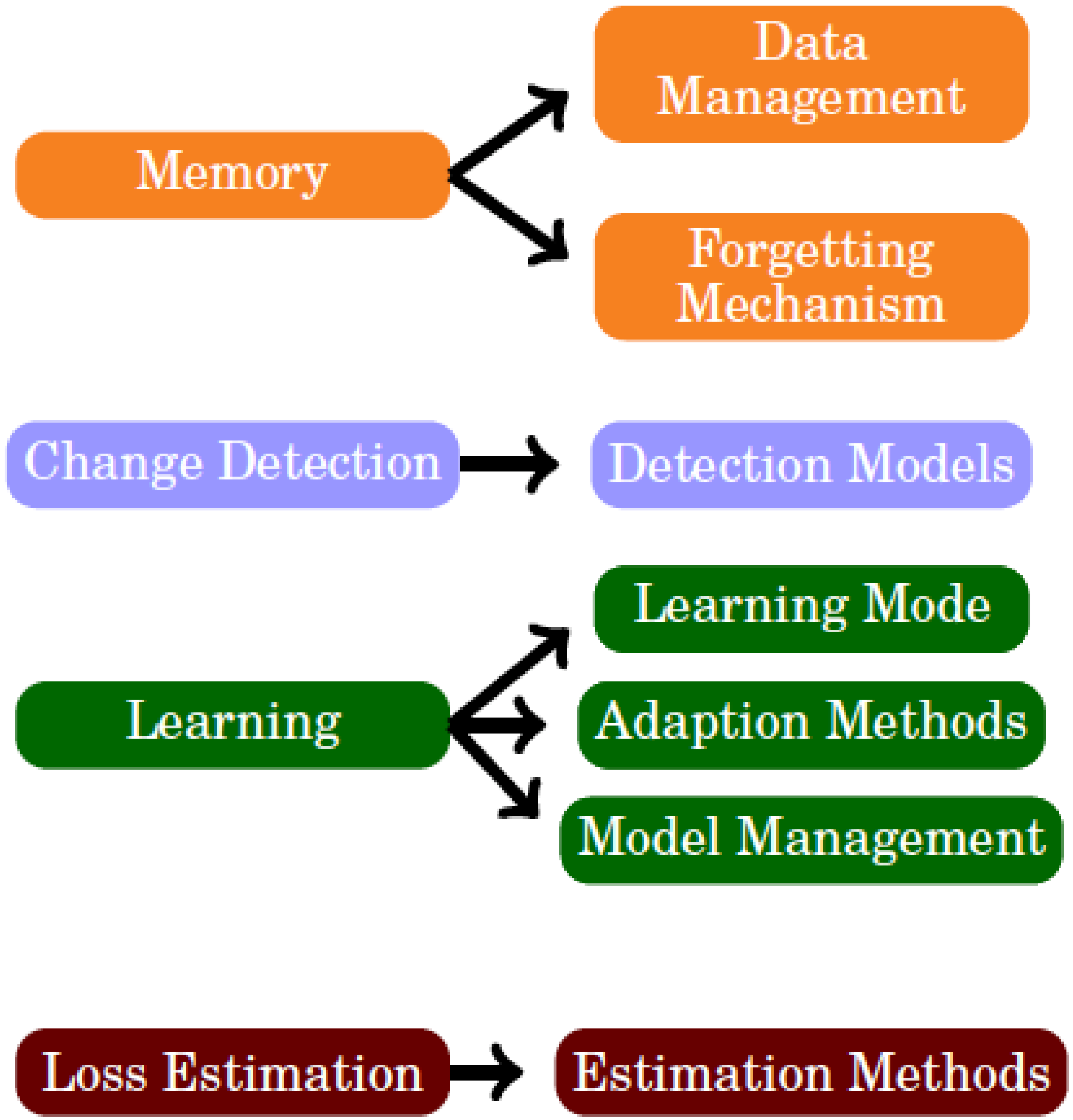
concept drift between t_0 and t_1: $\exists X : p_{t_0}(X, y) \neq p_{t_1}(X, y)$

changes that affect the prediction decision require adaptation

Adaptive Learning Strategies



Categorization of Approaches for HCD



Gama et al., A. (2014) *A Survey on Concept Drift Adaptation*, ACM Computing Surveys, 46(4)

Techniques to Handle Concept Drift

Triggering

Evolving

change detection and a follow up reaction

adapting every step

Single classifier

Detectors

variable windows

Forgetting

fixed windows,
Instance weighting

Ensemble

Contextual

dynamic integration,
meta learning

Dynamic ensemble

adaptive
fusion rules

Techniques to Handle Concept Drift

reactive, forgetting

Triggering

Evolving

Single classifier

Detectors

Forgetting

variable windows

fixed windows,
Instance weighting

Ensemble

Contextual

Dynamic ensemble

dynamic integration,
meta learning

adaptive
fusion rules

maintain some memory

Closer Look

Triggering

Evolving

Forget old data and retrain at a fixed rate

Single classifier

Detectors

Forgetting

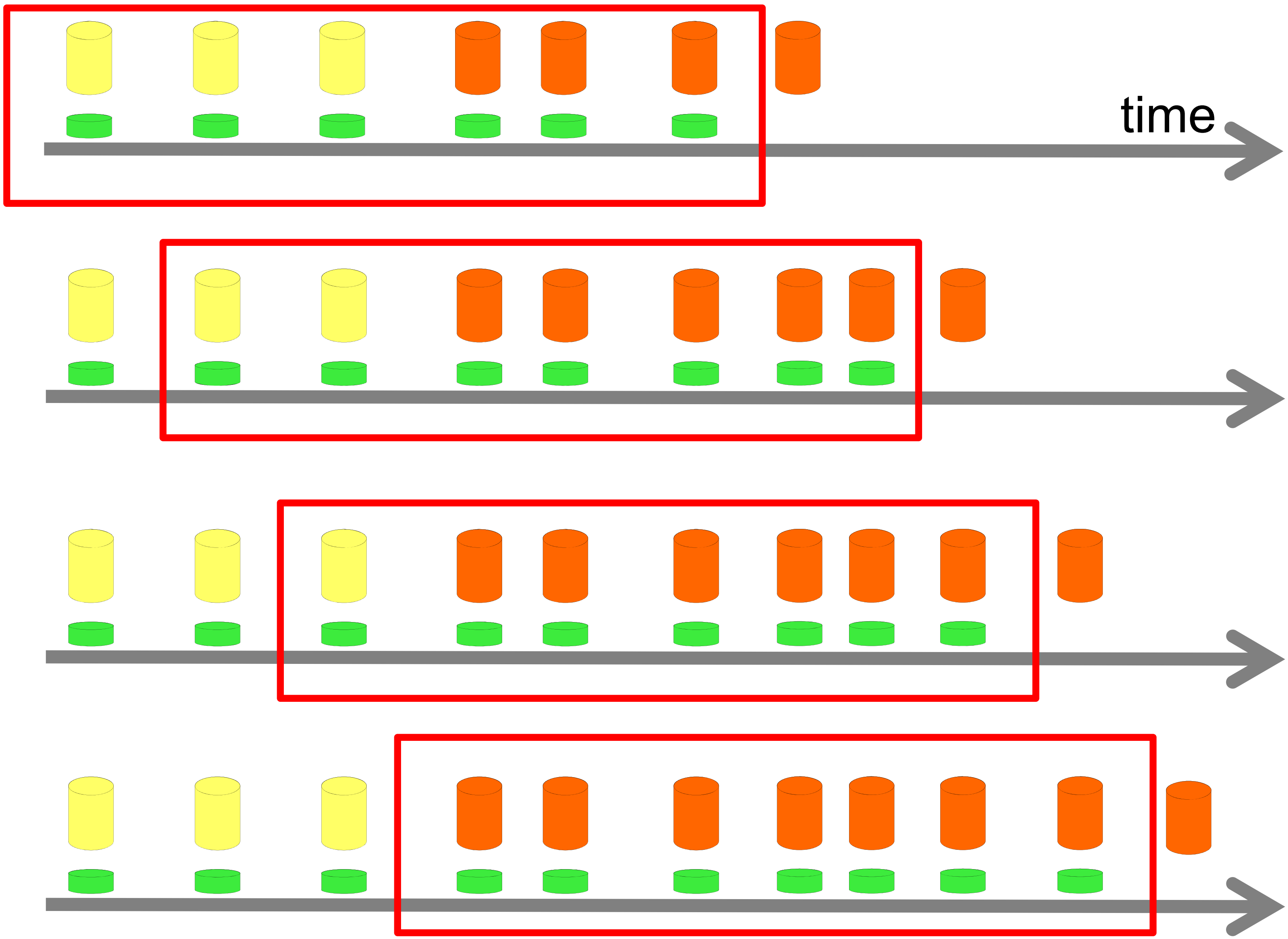
fixed windows,
Instance weighting

Ensemble

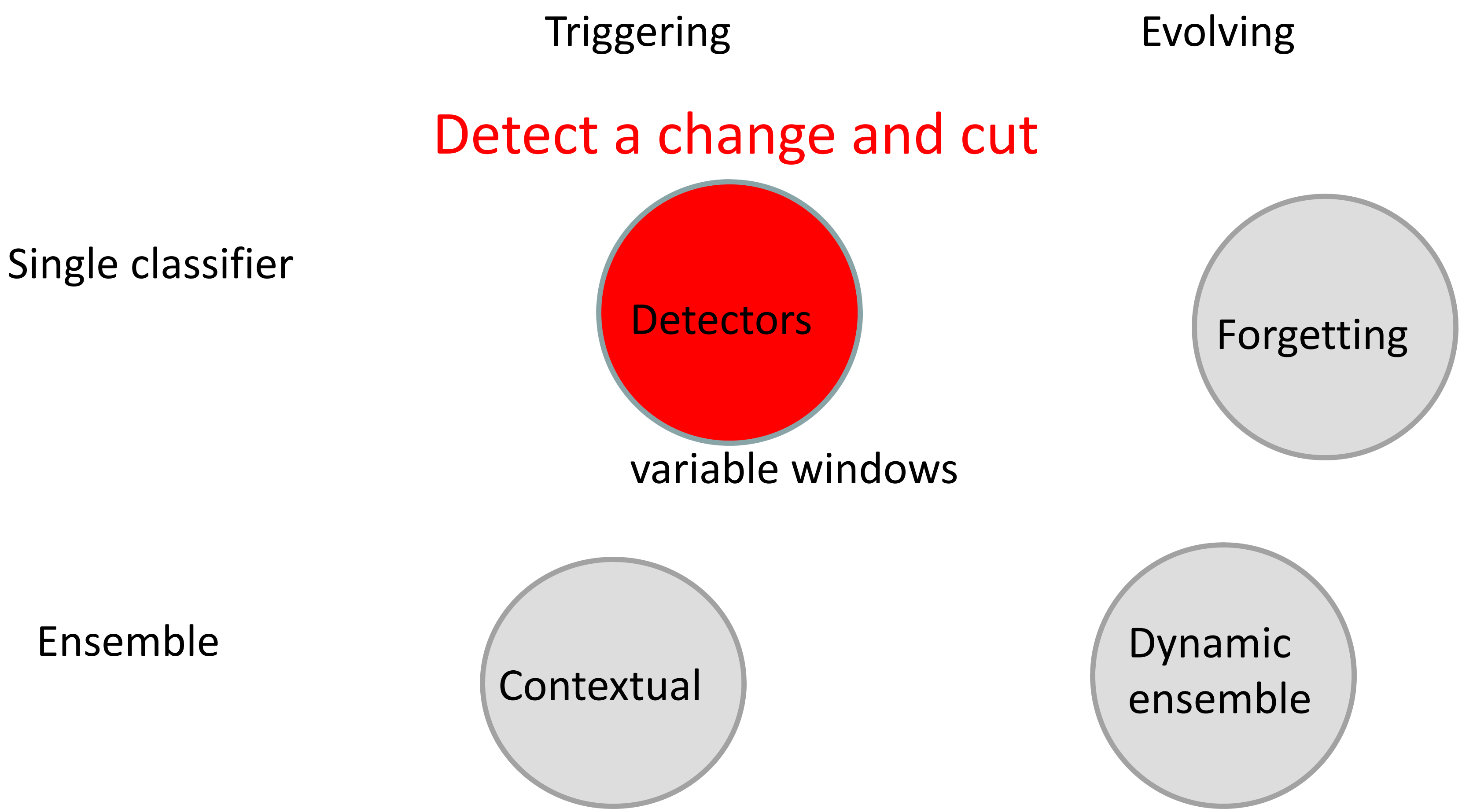
Contextual

Dynamic ensemble

Fixed Training Window

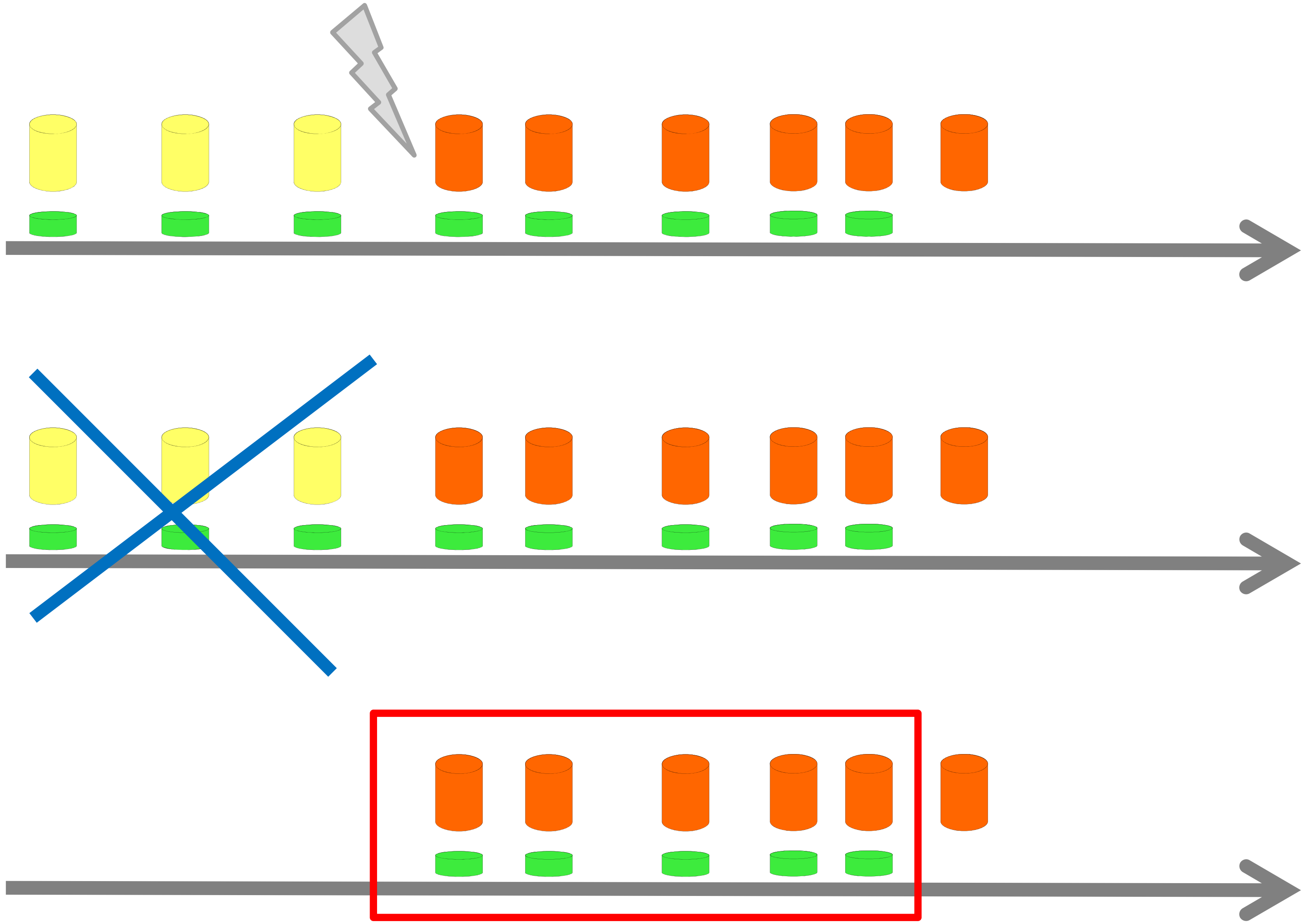


Closer Look



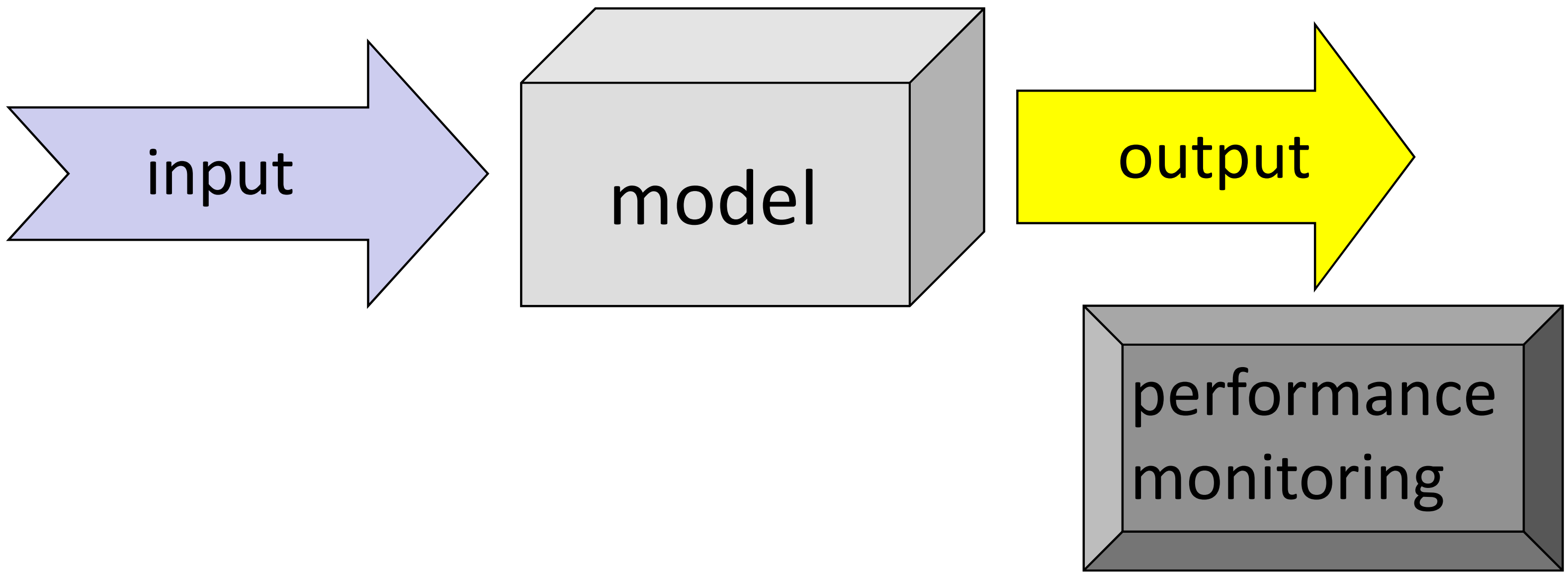
Triggering
Evolving
Detect a change and cut
Single classifier
Detectors
Forgetting
variable windows
Ensemble
Contextual
Dynamic ensemble

Variable Training Window



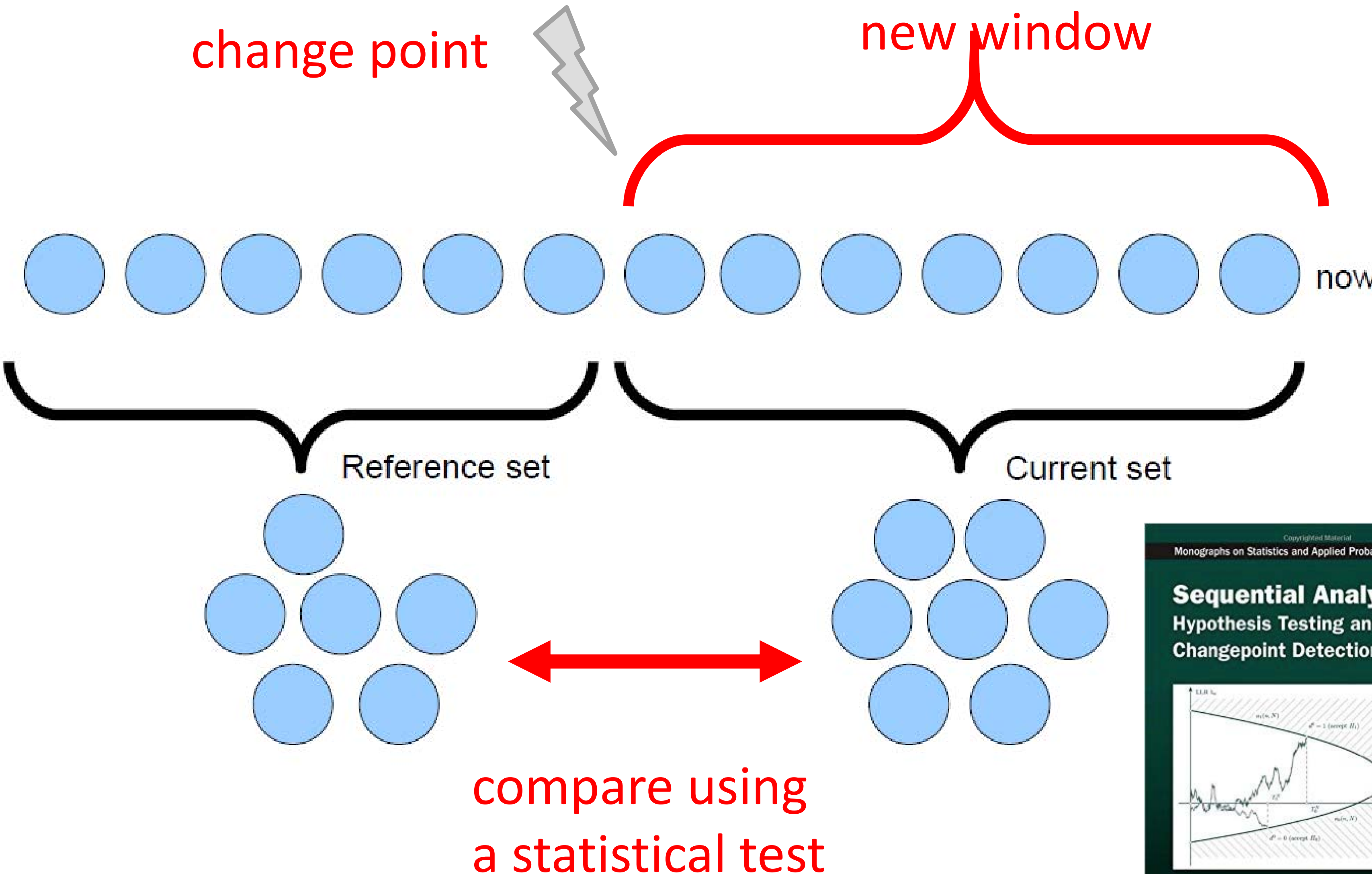
Change Detection

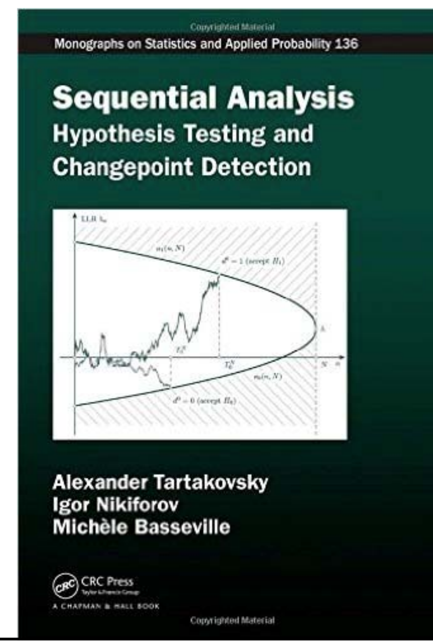
- Where to look for a change?



techniques that handle the real CD can also handle CDs that manifest in the input, but not the other way around

Detection


change point
new window
now
Reference set
Current set
compare using
a statistical test


Monographs on Statistics and Applied Probability 136
Sequential Analysis
Hypothesis Testing and
Changepoint Detection
Alexander Tartakovsky
Igor Nikiforov
Michèle Basseville
CRC Press

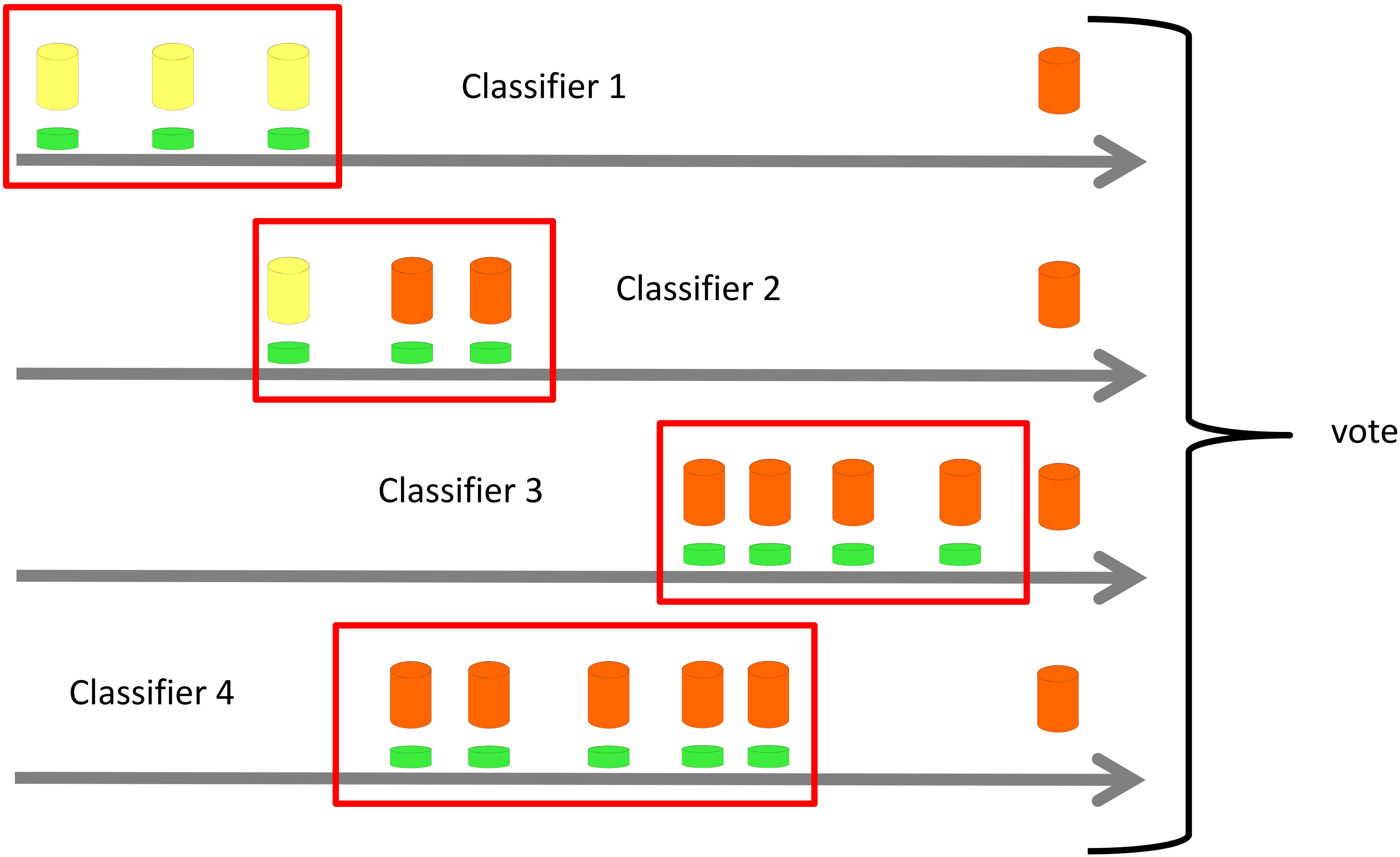
Closer Look

	Triggering	Evolving
Single classifier	Detectors	Forgetting
Ensemble	Contextual	Dynamic ensemble

adaptive fusion rules

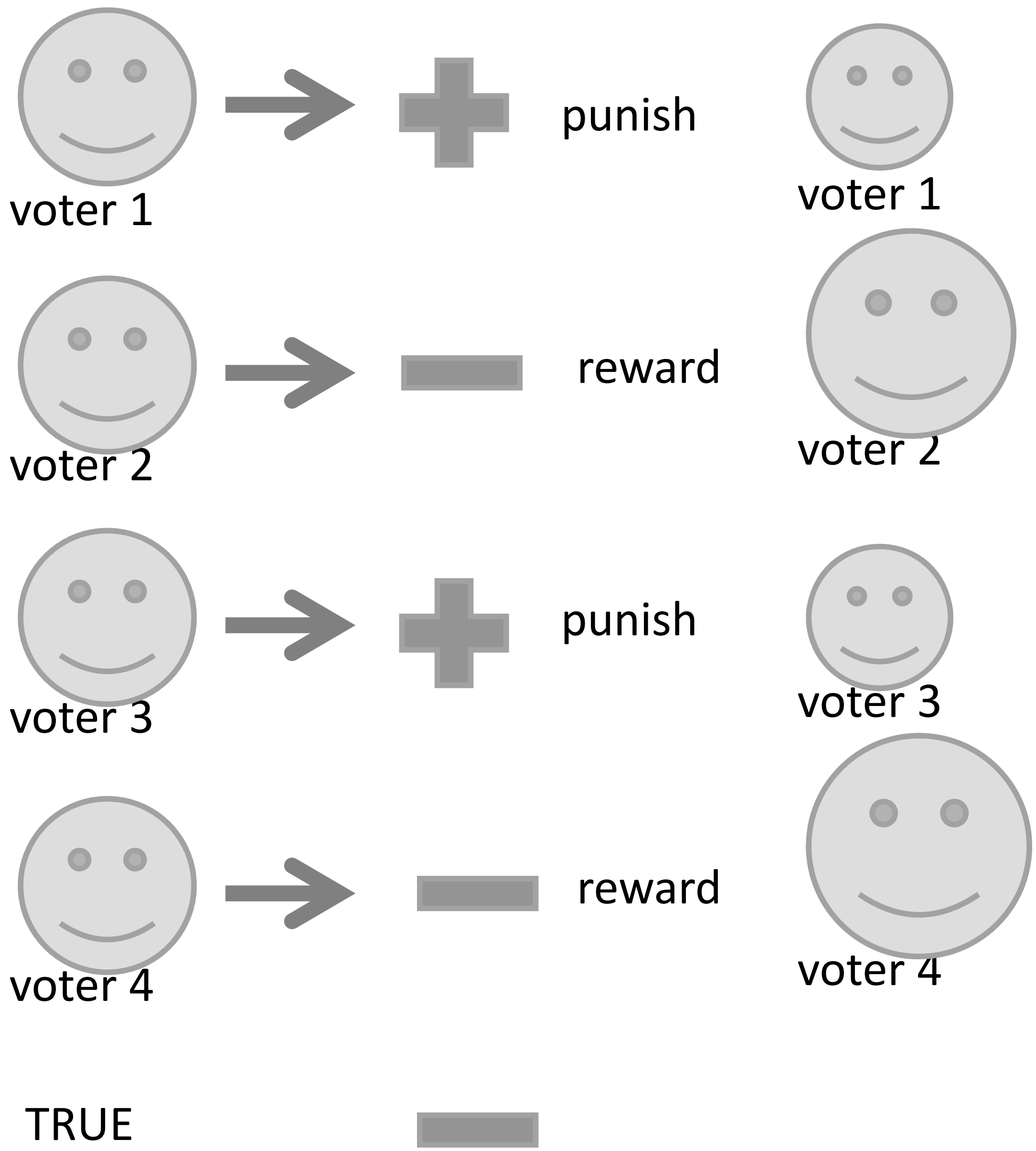
build many models, dynamically combine

Dynamic Ensemble

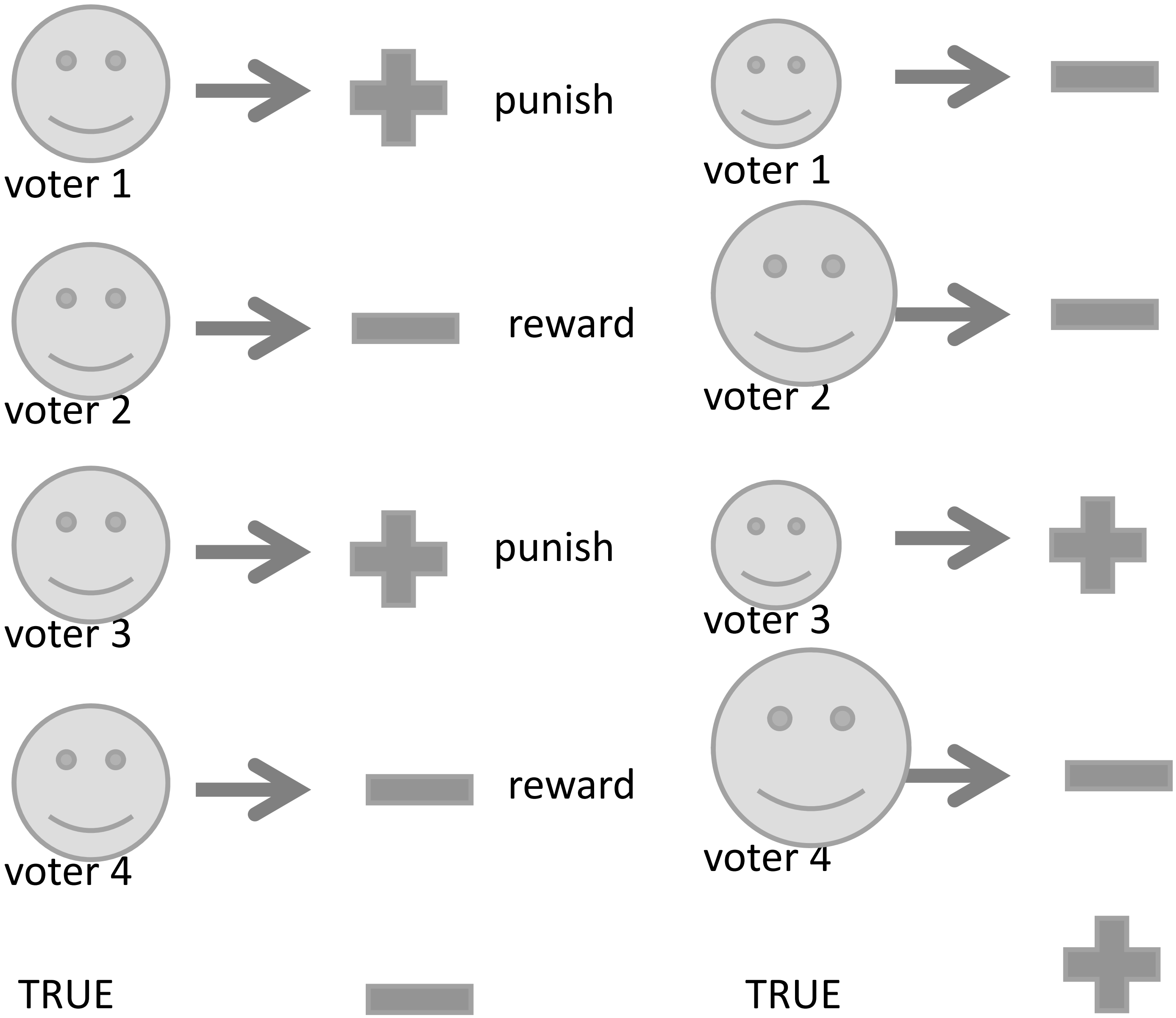


Dynamic Ensemble

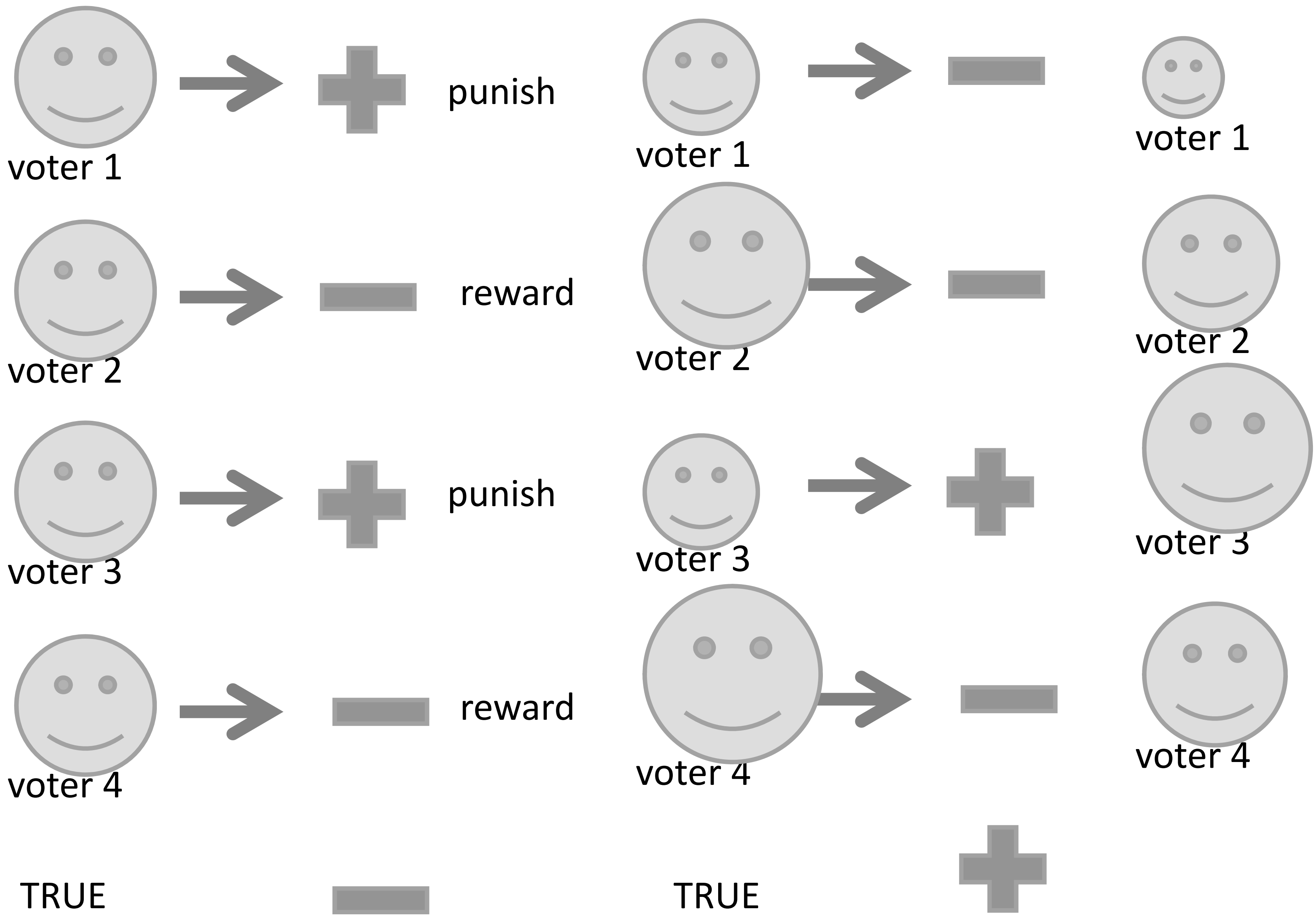
Dynamic Ensemble



Dynamic Ensemble



Dynamic Ensemble



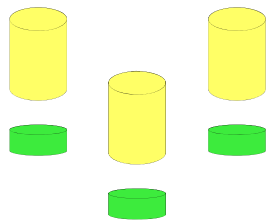
Closer Look

	Triggering	Evolving
Single classifier	Detectors	Forgetting
Ensemble	Contextual	Dynamic ensemble

dynamic integration,
meta learning

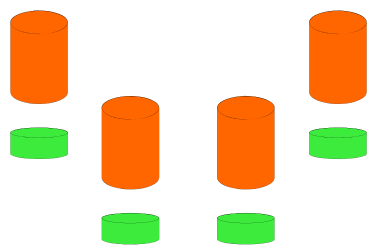
build many models,
switch models according to the
observed incoming data

Dynamic Integration



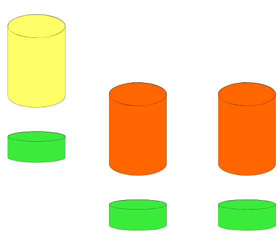
Group 1 = Classifier 1

- partition the training data
- build/select best classifiers for each partition



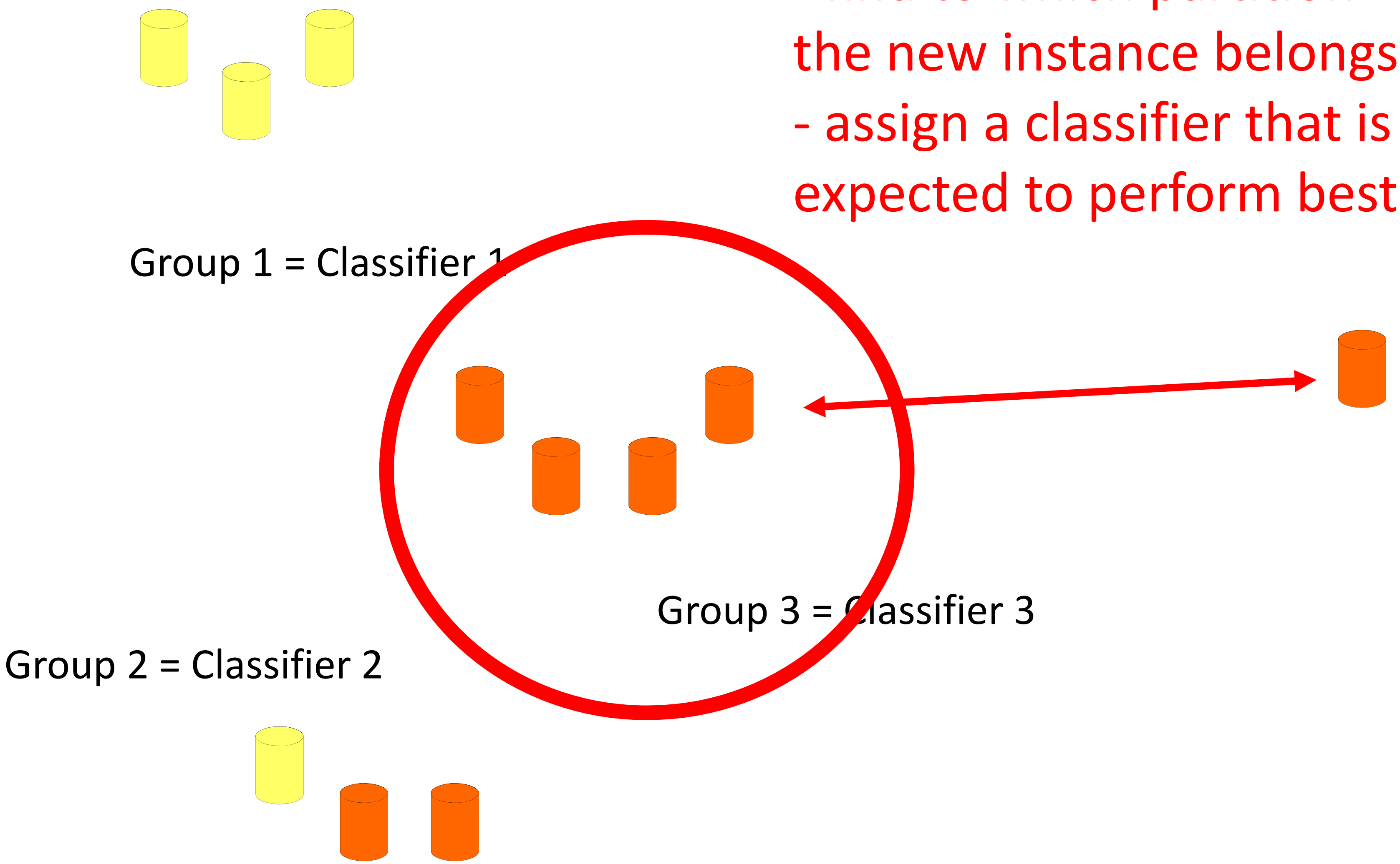
Group 3 = Classifier 3

Group 2 = Classifier 2



Dynamic Integration

- find to which partition the new instance belongs
- assign a classifier that is expected to perform best on it



Handling Concept Drift Summary

change source

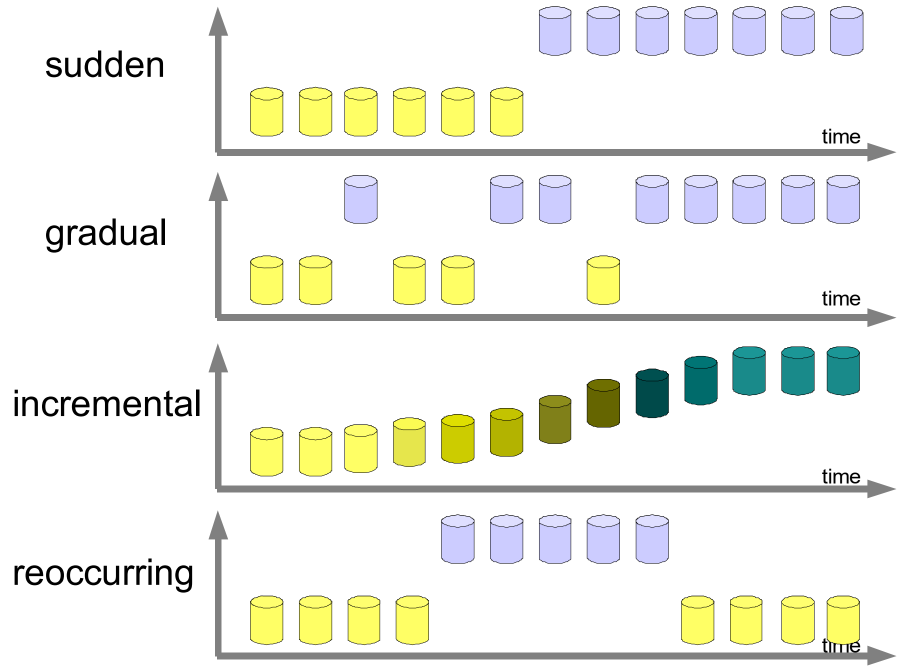
adversary
interests
population
complexity

expectations about changes

unpredictable
predictable
identifiable

expectations about desired action

keep the model uptodate
detect the change
identify/locate the change
explain the change



labels

real time
on demand
fixed lag
delay

ground truth labels

soft/hard

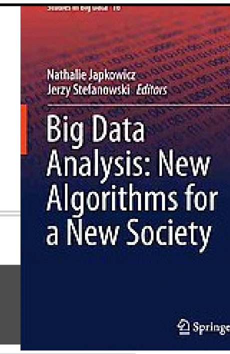
decision speed

real time
analytical

costs of mistakes

balanced/
unbalanced

Research vs. Practice



	Research	Practice
Change type	Sudden	Sudden, gradual/incremental, recurring Multiple types in the same application
Change expectation	Unpredictable, unexpected	Unpredictable, expected, predictable
Labels	Immediately available	Proxies for labels available, with some fixed/variable delay, never
Ground truth	Objective	Objective, subjective
Background knowledge	Not available	Available, not available
Evaluation	Simulation/log replay	Deployment and live traffic needed
Reoccurrence	Independent of each other, unexpected	Expected, predictable, explainable
Drifts in multiple objects	Independent of other objects	Affected by, predictable from other objects

Žliobaite et al. (2016) "*An overview of concept drift applications*", In Big Data Analysis: New Algorithms for a New Society, pp. 91-114. Springer.

Take Home Messages

- Data patterns change over time,
  - models need to be adaptive to maintain accuracy
- Four types of learning techniques
  - make different assumptions about the data and change
- Application tasks have different properties =>
  - pose different challenges,
  - require different handling techniques,
  - there is no “one size fits all” solution

Next Steps/Challenges

- Predictors should anticipate & adapt to changes
  - From reactive to proactive adaptation
  - context-awareness may become an answer
- Improve usability and trust
  - Integrate domain knowledge
  - Provide transparency, explanation and control for
    - how changes are detected
    - how changes are handled and models adapted
  - Visualization of drift, explanations, business logic
  - Semi-automation, i.e. interaction with an expert
- A system-oriented perspective is lacking

Outlook

- Changing the focus from blind adaptivity
  - to change/CD modeling and description
  - to recognizing & reusing similar situations from the past and from the peers
- Application driven concept drift problems, like
  - label unavailability or delay in availability
  - cost-benefit trade off of the model update
  - controlled adaptivity (due to adversaries)
  - lack of ground truth for training
- A CRISP-ADM reference framework and guidelines
  - for incorporating adaptivity in modeling
  - to be used in different application tasks

Thank you!

m.pechenizkiy@tue.nl

nl.linkedin.com/in/mpechen/

Many related topics (not covered)

- Online detection of recurrent changes
  - *Modelling recurrent events for improving online change detection, by* Maslov et al, SIAM SDM 2016
  - *BLPA: Bayesian Learn-Predict-Adjust Method for Online Detection of Recurrent Changepoints*, by Maslov et al., IJCNN 2017
- Concept drift in process mining
  - *Dealing with concept drifts in process mining*, by JC Bose et al. IEEE TNNLS 25(1), 2014.
- Delayed labeling settings
- Concept drift in unsupervised learning
  - Pattern mining and clustering
  - Mining temporal networks
    - Monitoring statistical properties, e.g. centralities
- Transfer learning

Bibliography – General Background

- *Characterizing Concept Drift.* Geoffrey I. Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, Francois Petitjean http://arxiv.org/abs/1511.03816
- *A Survey on Concept Drift Adaptation*, Gama, J., Žliobaite, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. ACM Computing Surveys, 2014
- *Detection of Abrupt Changes - Theory and Application*, M. Basseville, I. Nikiforov, Prentice-Hall, Inc. 1993.
- *Statistical Quality Control*, E. Grant, R. Leavenworth, McGraw-Hill, 1996.
- *Continuous Inspection Scheme*, E. Page. Biometrika 41 1954
- *Learning in the Presence of Concept Drift and Hidden Contexts*, G.Widmer, M. Kubat: Machine Learning 23(1): 69-101 (1996)
- *Learning drifting concepts: Example selection vs. example weighting*, R.Klinkenberg, IDA 2004
- *Adaptive Learning and Mining for Data Streams and Frequent Patterns,* Albert Bifet, PhD Thesis, 2009
- *Adaptive Training Set Formation*, Indre Žliobaitė, PhD Thesis, Vilnius University, Lithuania, 2010.

Bibliography - Approaches

- *Modelling recurrent events for improving online change detection, by* Maslov et al, SIAM SDM 2016
- *BLPA: Bayesian Learn-Predict-Adjust Method for Online Detection of Recurrent Changepoints*, by Maslov et al., IJCNN 2017
- *Predictive Handling of Asynchronous Concept Drifts in Distributed Environments*, by Ang, H., Gopalkrishnan, V., Žliobaite, I., Pechenizkiy, M. & Hoi, S., IEEE Transactions on Knowledge and Data Engineering (2013)
- *Mining Time-Changing Data Streams*, by G. Hulten, L. Spencer, P. Domingos, ACM SIGKDD, 2001.
- *Dynamic Weighted Majority: A New Ensemble Method for Tracking Concept Drift*, by J. Kolter, M. Maloof, ICDM 2003.
- Mining Concept Drifting Data Streams using Ensemble Classiers, by H.Wang, Wei Fan, P. Yu, J. Han, ACM SIGKDD 2003.
- *Decision Trees for Mining Data Streams*, by J. Gama, R. Fernandes, R.Rocha. Intelligent Data Analysis 10(1):23-45 (2006)
- *OLINDDA: A cluster-based approach for detecting novelty and concept drift in data streams.* Spinosa, E.J., Carvalho, A., and Gama, J. 22nd ACM SAC 2007: ACM Press.
- *An Ensemble of Classifiers for Coping with Recurring Contexts in Data Streams, by* Katakis, I., Tsoumakas, G., and Vlahavas, I.. in 18th ECAI. 2008, IOS
- *Dealing with concept drifts in process mining*, IEEE Trans. on Neural Networks and Learning Systems, by JC Bose, R., van der Aalst, W., Zliobaite, I. & Pechenizkiy, M. (2013)

Bibliography - Applications

- *An overview of concept drift applications, Žliobaite, I., Pechenizkiy, M. & Gama, J., In Big Data Analysis: New Algorithms for a New Society, Springer, 2016, http://www.win.tue.nl/~mpechen/publications/pubs/ZliobaiteCDApps2015.pdf*
- *Collaborative Filtering with Temporal Dynamics* Yehuda Koren, KDD 2009, ACM, 2009
- *MOA: Massive online analysis, a framework for stream classification and clustering*. Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl, HaCDAIS Workshop ECML-PKDD 2010
- *Online Mass Flow Prediction in CFB Boilers with Explicit Detection of Sudden Concept Drift*. Pechenizkiy, M., Bakker, J., Žliobaitė, I., Ivannikov, A., Karkkainen, T. SIGKDD Explorations 11(2), p. 109-116, 2009.
- *Dynamic Integration of Classifiers for Handling Concept Drift*, Tsymbal, A., Pechenizkiy, M., Cunningham, P. & Puuronen, S. Information Fusion, Special Issue on Applications of Ensemble Methods, 9(1), pp. 56-68, 2008.
- *Real-time algorithm for changes detection in depth of anesthesia signals*. R. Sebastião et al. Evolving Systems 4(1): 3-12 (2013)