Active discovery and classification with applications on activity modelling

Tao Xiang

Queen Mary, University of London

With Tim Hospedales, Chen Change Loy and Tom Haines

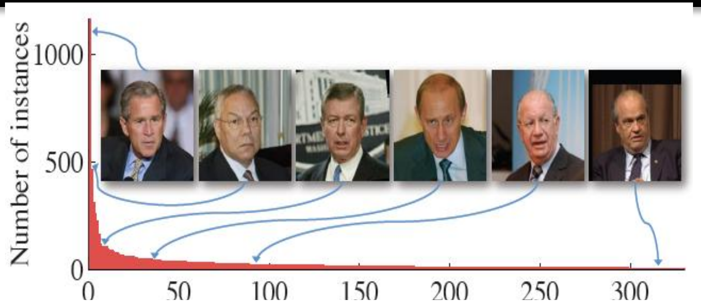
Outline

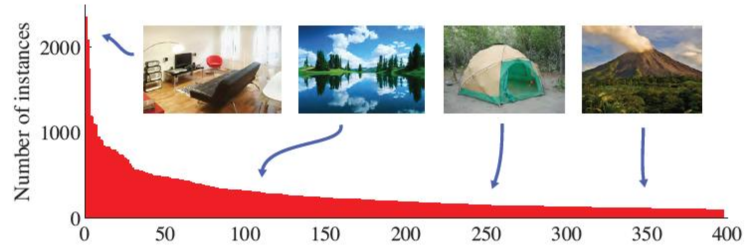
- The problem: modelling rare classes by active discovery and classification
- Pool-based:
  - Adapting generative and discriminative models
  - Misclassification criterion using Dirichlet process
- Stream-based: online active learning criterion selection
- Weakly-supervised learning of rare events

The problem

Problem: Joint (active) discovery and learning to classify rare categories

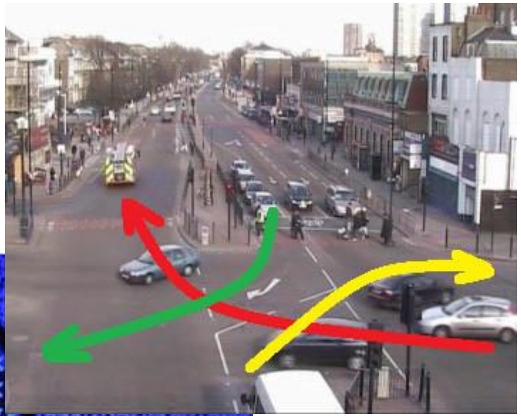
- Computer network intrusion detection
- Financial transaction monitoring
- Video surveillance





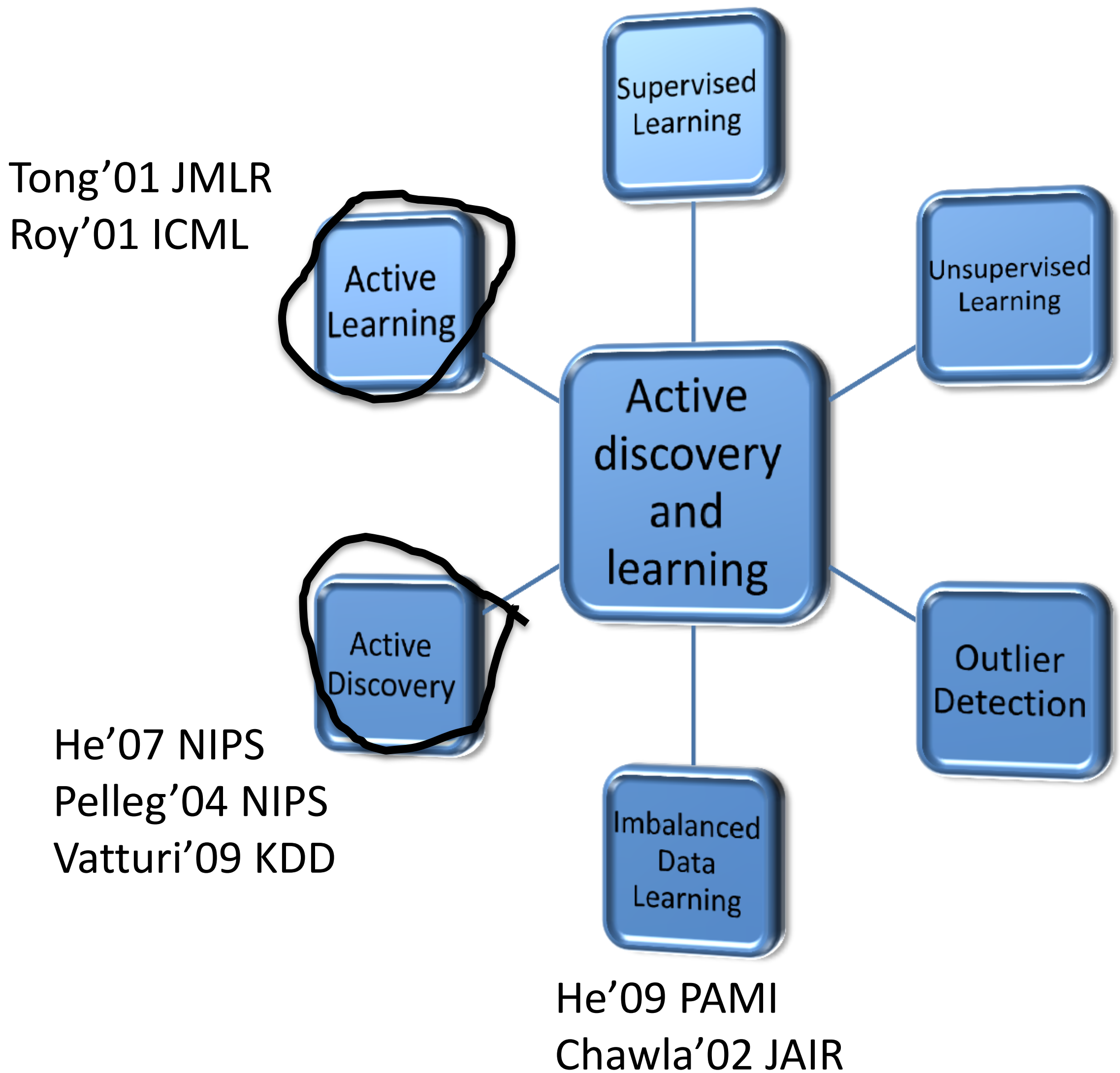
Characteristics:

- Large data volume: exhaustive labelling impossible
- Unbalanced classes
- Rare classes are unknown



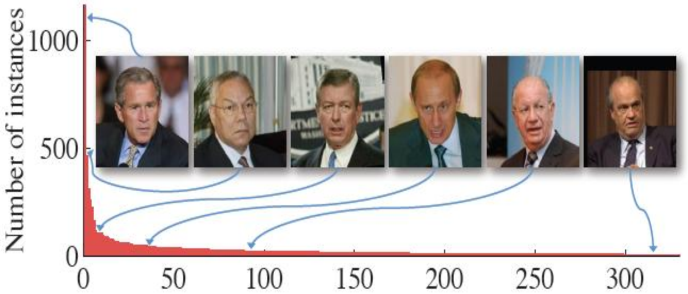
Background

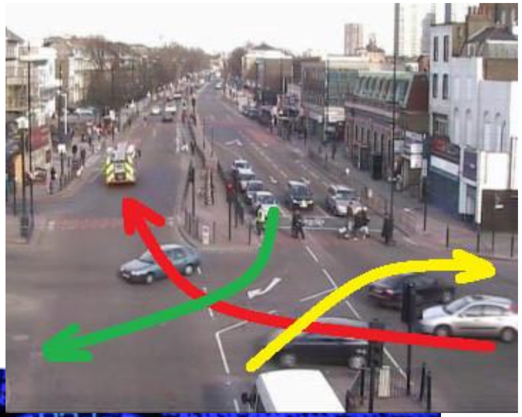
- Joint discovery and classification via active learning



Challenges

- Limited supervision
- Limited rare class data
- Joint detect & classify





State of the arts

Existing active learning methods

- Single objective: discovery or classification
- Mostly single criterion
  - Different criteria are needed for different objectives
- Single classifier
  - Different classifiers are more suitable for different data and different amount of supervision

What we need:

- Joint discovery and classification
- Adaptive multi-criteria weighting
- Classifier fusion

POOL-BASED ACTIVE DISCOVERY AND LEARNING

Criteria Selection: The Problem

Active learning query criteria:

- Discovery: typically likelihood (outlier detection)

$$p_l(i) \propto \exp\left(-\beta \max_{y_i} p(x_i|y_i)\right)$$

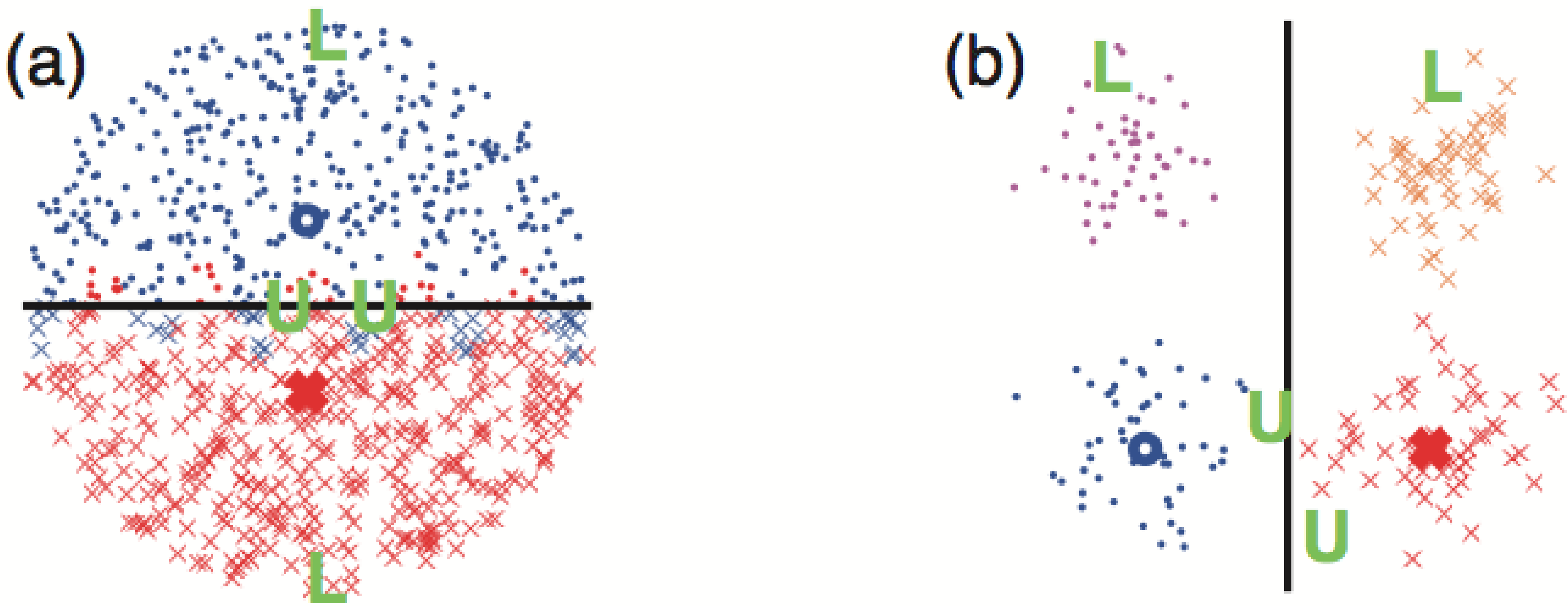
- Classification: typically uncertainty

$$p_u(i) \propto \exp\left(\beta \sum_{y_i} p(y_i|x_i) \log p(y_i|x_i)\right)$$

➢ The two problems are dependent

➢ How to balance different criteria?

Criteria Selection: Illustration



✔ Uncertainty
✖ Likelihood

✖ Uncertainty
✔ Likelihood

Solution: Adaptive Weighting

Choose criteria by adaptive weighting

- Select either uncertainty or likelihood
- Sample a multinomial distribution
- Two weights control the sampling, one for each criterion
- After each query, predict the classification performance via entropy of the classifier

$$H = -\sum_{y=1}^{n_y} \frac{\sum_i \mathrm{I}(f(\mathbf{x}_i) = y)}{|\mathcal{U}|} \log_{n_y} \frac{\sum_i \mathrm{I}(f(\mathbf{x}_i) = y)}{|\mathcal{U}|}$$

Solution: Adaptive Weighting

- Update the weights

$$w_{t+1,k}(q) \propto \lambda w_{t,k} + (1-\lambda)\phi_t(i)\frac{p_k(i)}{p(i)} + \epsilon.$$

- Where we define a reward function for Discovery and classification performance

$$\phi_t(i) = \alpha \mathrm{I}(y_i \notin \mathcal{L}) + (1-\alpha)\left((e^{H_t} - e^{H_{t-1}}\right)$$

Rewards discovery

Rewards increase in classification performance

Model Selection: The Problem

- Effective model/classifier types vary with data quantity. E.g. for a given generative-discriminative pair:
  - Low data: Generative better ✔
  - High data: Discriminative better ✔

GMM, Naïve Bayes

SVM, Logistic Regression

- How much is "sufficient" data?
  - Varies with dataset
  - May be crossed during active learning

➢ Need to select model online

Solution: Model Switching

- Solution: online classifier switching:
  - between GMM:

$$p(\mathbf{x}|y) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \sum_{n=1}^{N} \omega_n \exp -\frac{1}{2}\left((\mathbf{x}-\mathbf{x}_n)^T \Sigma^{-1} (\mathbf{x}-\mathbf{x}_n)\right)$$

  - ...and SVM:

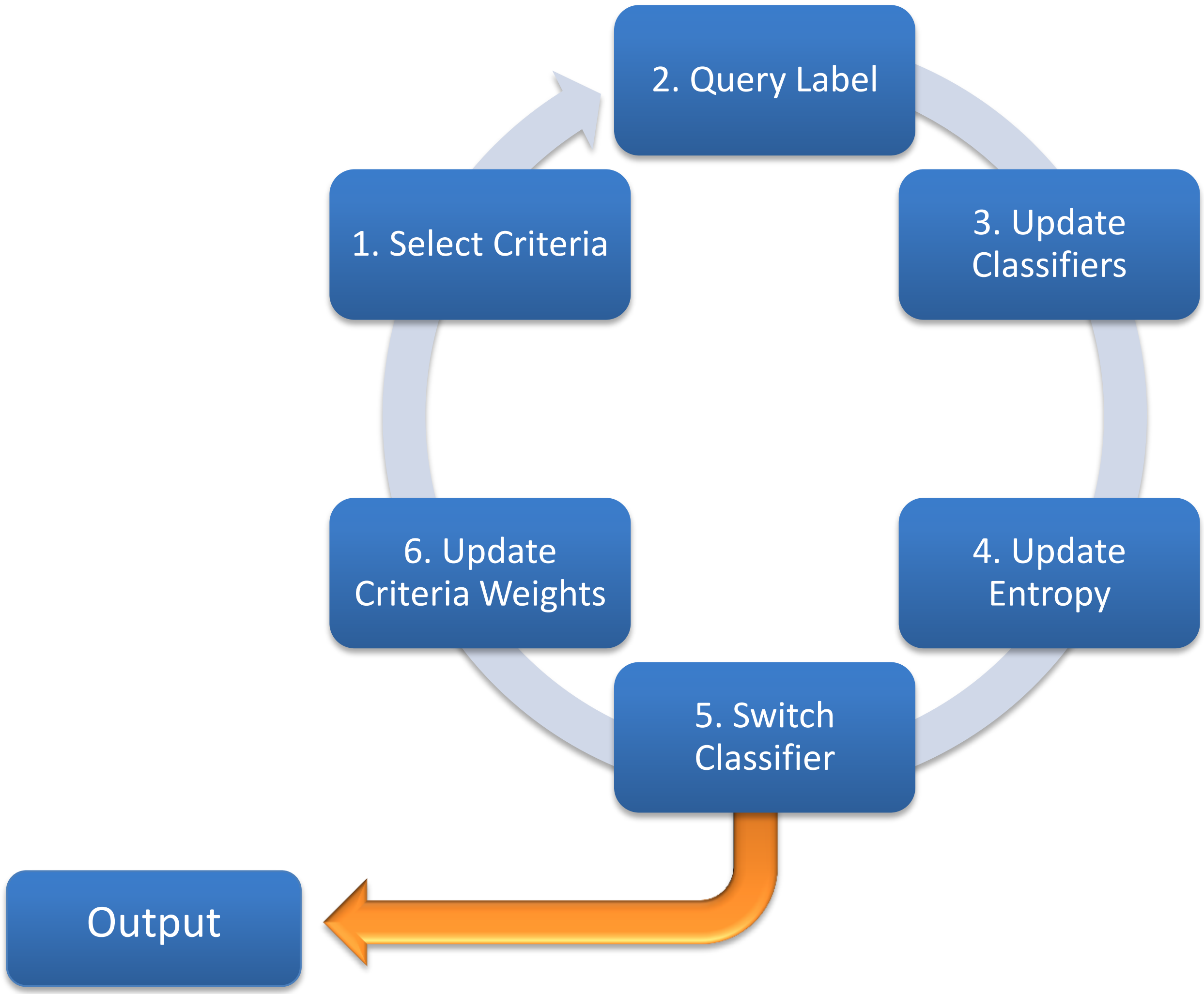
$$f_{svm}(\mathbf{x}) = \operatorname*{argmax}_{y} \left(\sum_{\mathbf{v}_i \in SV_y} \alpha_{ki} \mathcal{N}(\mathbf{x}; \mathbf{v}_i) + \alpha_{k0} \right)$$

  - According to classification performance (Entropy):

$$H = -\sum_{y=1}^{n_y} \frac{\sum_i \mathrm{I}(f(\mathbf{x}_i) = y)}{|\mathcal{U}|} \log_{n_y} \frac{\sum_i \mathrm{I}(f(\mathbf{x}_i) = y)}{|\mathcal{U}|}$$

Algorithm Summary

- T. Hospedales, S. Gong and T. Xiang, "Finding Rare Classes: Active Learning with Generative and Discriminative Models", *IEEE Transactions on Knowledge and Data Engineering (TKDE), 2012*



Algorithm Summary

Algorithm 1 Active Learning for Discovery and Classification

Active Learning

Input: Initial labeled $\mathcal{L}$ and unlabeled $\mathcal{U}$ samples. Classifiers $\{f_c\}$, query criteria $\{Q_k\}$, weights w.

1) Build unconditional GMM $p(\mathrm{x})$ from $\mathcal{L} \cup \mathcal{U}$ (8)-(12)
2) Estimate σ by cross-validation on $p(\mathrm{x})$ (13)
3) Train initial GMM and SVM classifiers on $\mathcal{L}$

Repeat as training budget allows:

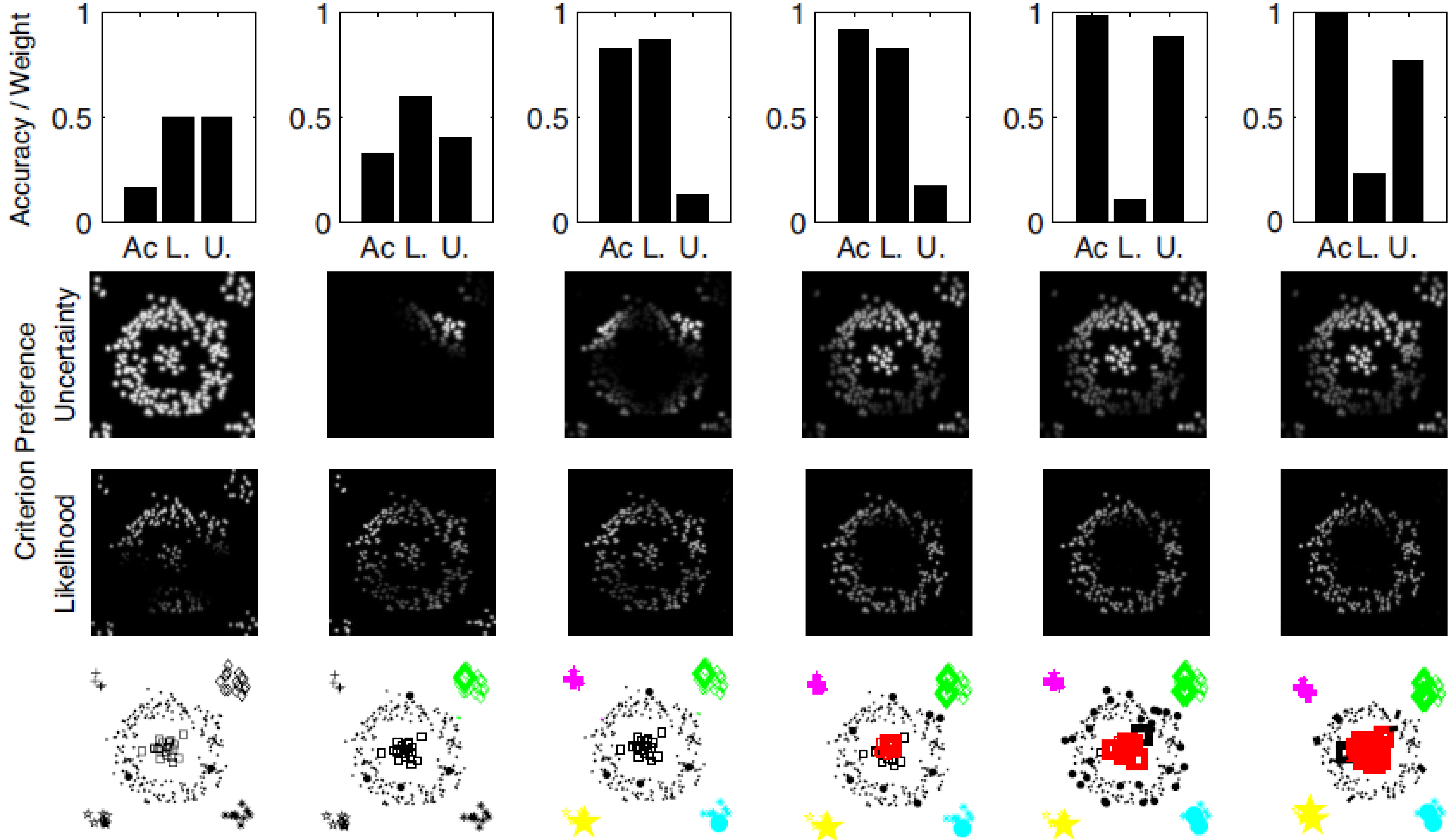
1) Compute query criteria $p_{lik}(i)$ (5) and $p_{unc}(i)$ (3)
2) Sample query criteria to use $k \sim \mathrm{Multi}(\mathrm{w})$
3) Query point $i^* \sim p_k(i)$, add $(\mathrm{x}_{i^*}, y_{i^*})$ to $\mathcal{L}$
4) Update classifiers with label i^* (14) and (15)
5) Update query criteria weights w (17) and (18)
6) Compute entropies H_{gmm} and H_{svm} (16)
7) If $H_{gmm} > H_{svm}$: select classifier $f_{gmm}(\mathrm{x})$ (19)
8) Else: select $f_{svm}(\mathrm{x})$ (19)

Testing

Input: Testing samples $\mathcal{U}^*$, selected classifier c.

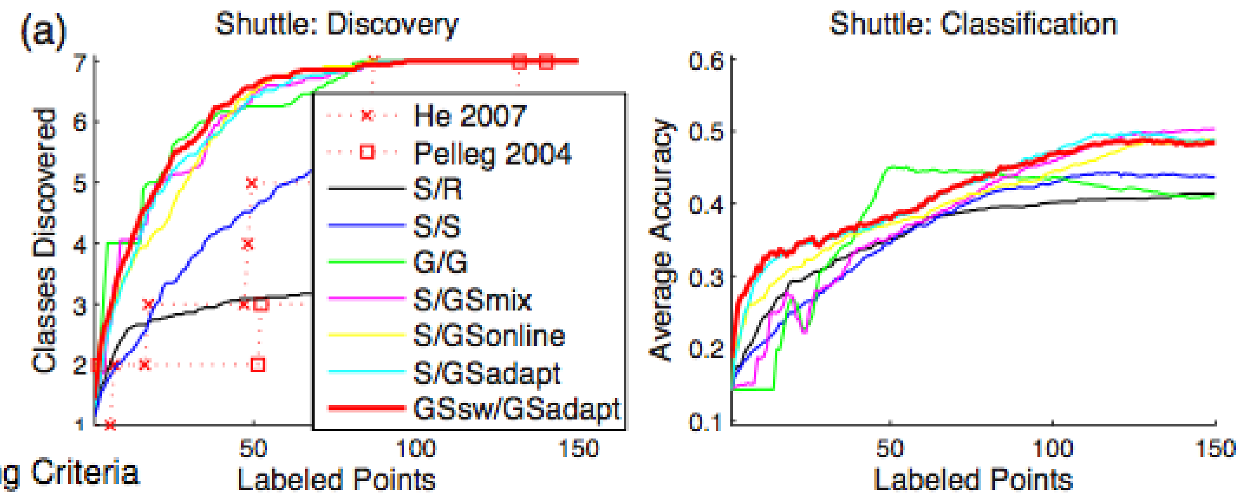
1) Classify $x \in \mathcal{U}^*$ with $f_c(x)$ ((14) or (15))

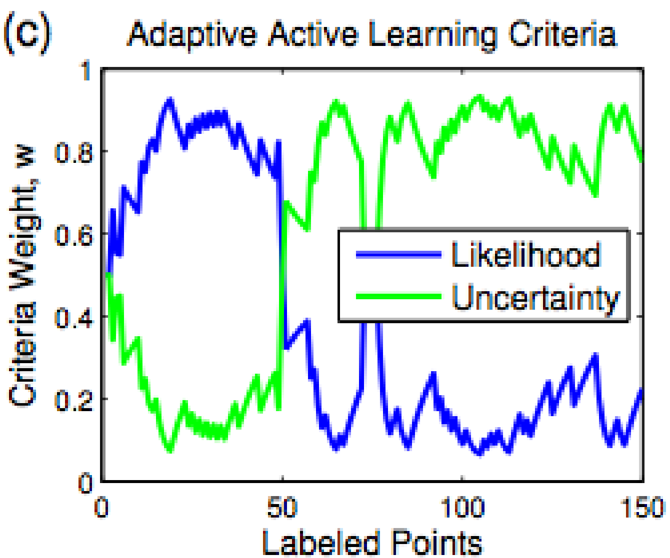
Synthetic data



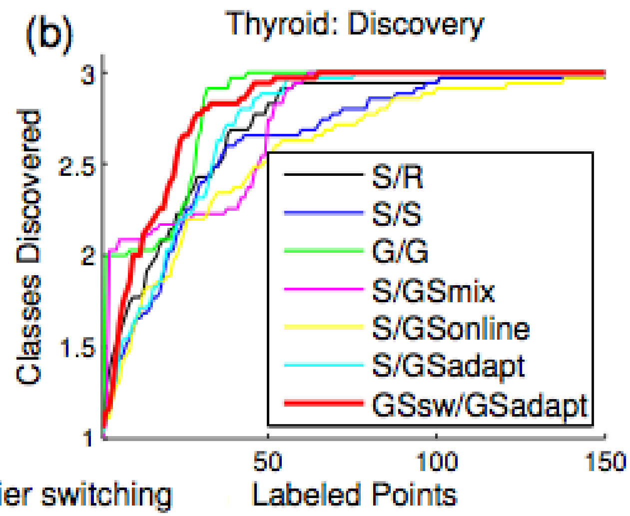
(a). Iter: 1 (b). Iter: 4 (c). Iter: 7 (d). Iter: 13 (e). Iter: 40 (f). Iter: 64

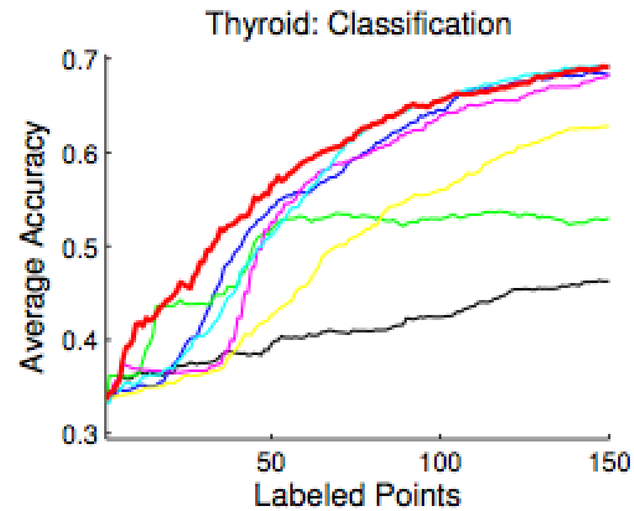
UCI data

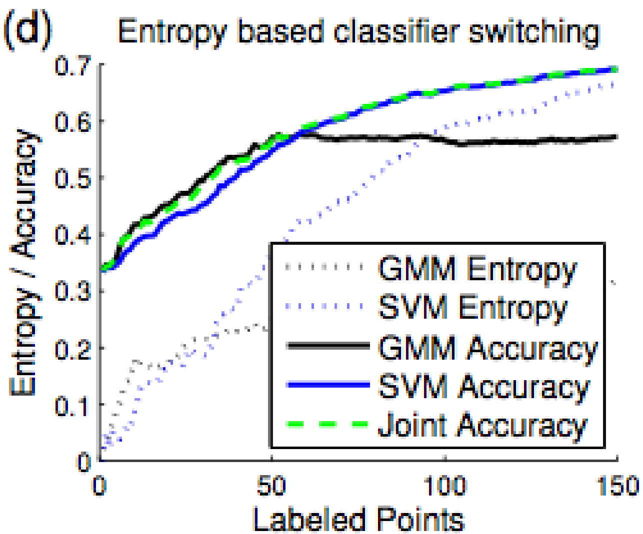

(a)
Shuttle: Discovery
Classes Discovered
Labeled Points
He 2007
Pelleg 2004
S/R
S/S
G/G
S/GSmix
S/GSonline
S/GSadapt
GSsw/GSadapt
Shuttle: Classification
Average Accuracy
Labeled Points


(c)
Adaptive Active Learning Criteria
Criteria Weight, w
Likelihood
Uncertainty
Labeled Points

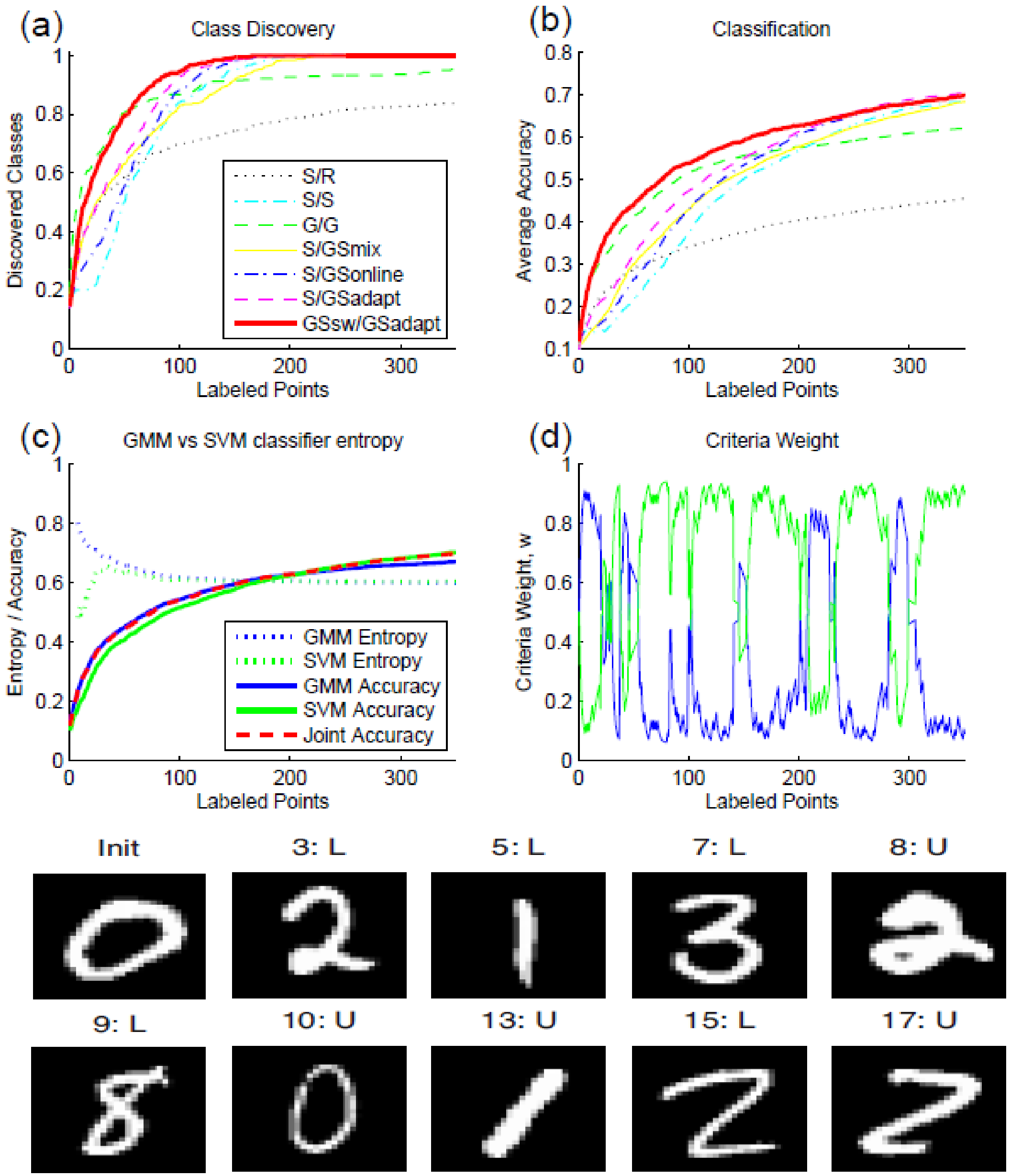
UCI Results


(b)
Thyroid: Discovery
Classes Discovered
3
2.5
2
1.5
1
50
100
150
Labeled Points
S/R
S/S
G/G
S/GSmix
S/GSonline
S/GSadapt
GSsw/GSadapt


Thyroid: Classification
Average Accuracy
0.7
0.6
0.5
0.4
0.3
50
100
150
Labeled Points


(d)
Entropy based classifier switching
Entropy / Accuracy
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
Labeled Points
GMM Entropy
SVM Entropy
GMM Accuracy
SVM Accuracy
Joint Accuracy

Handwriting Digits



Summary

- Joint discover & classify: Under-studied
- How to balance discovery & classification?
  - ➢ **ADAPTIVE CRITERIA SELECTION**
- How to generalise datasets and volume?
  - ➢ **SWITCHING GEN. & DISCR. CLASSIFIER MODELS**

Limitation:

- Adaption model is fairly heuristic
- Pool based setting only

A MORE PRINCIPLED WAY FOR JOINT DISCOVERY AND CLASSIFICATION

Discovery & Classification

- **Discovery** is when not all classes are known, and need to be found.
- **Classification** is where the classes are considered to be known but the boundaries between them need to be refined.
- We tackle both problems simultaneously, with the express purpose of *maximising classification* performance.

Assumptions

- **Assumption 1**: That the item with the greatest probability of being misclassified should be selected.
- **Assumption 2**: That the classes have been drawn from a Dirichlet process. This is equivalent to assuming the items in the pool come from a Dirichlet process mixture model.

Illustration of Dirichlet Process

(a)

(b)

The Algorithm

Class assignment that the classifier, which cannot consider new classes, gives:

$$\text{cc} = \underset{c \in C}{\text{argmax}}\, P_c(c|\text{data})$$

Class assignment probability, including the possibility of a new class:

$$P_n(c \in C \cup \{\text{new}\}|\text{data}) \propto \begin{cases} \frac{m_c}{\sum_{k \in C} m_k + \alpha} P_c(\text{data}|c) & \text{if } c \in C \\ \frac{\alpha}{\sum_{k \in C} m_k + \alpha} P(\text{data}) & \text{if } c = \text{new} \end{cases}$$

Probability of misclassification:

$$P(\text{wrong}|\text{data}) = 1 - P_n(\text{cc}|\text{data})$$

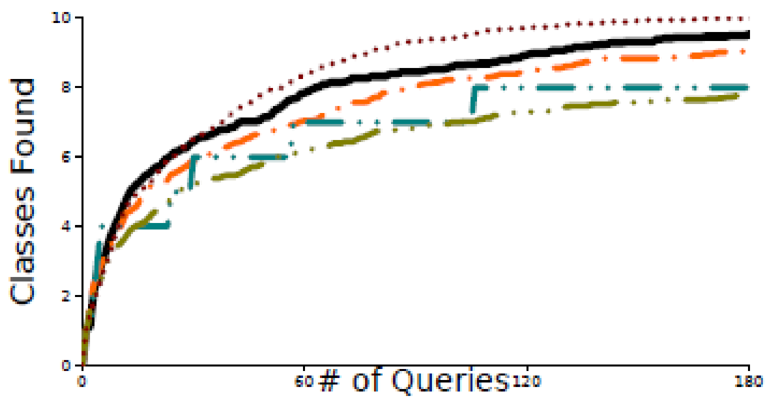
T. Haines and T. Xiang, "Active Learning using Dirichlet Processes for Rare Class Discovery and Classification", in *British Machine Vision Conference*, 2011.

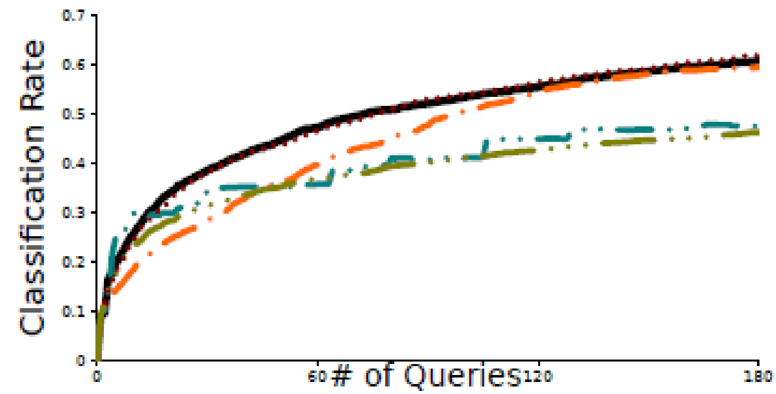
The Algorithm

- Note that the misclassification probability (P(wrong))
  - Is different from uncertainty due to the unknown new classes considered (P(new))
  - If P(new) is high, P(wrong) is high – encourage discovery of new classes
  - If the classifier is uncertain about existing classes, P(wrong) is high too
  - These two factors are dominated by the concentration parameter of the DP model – learned from data

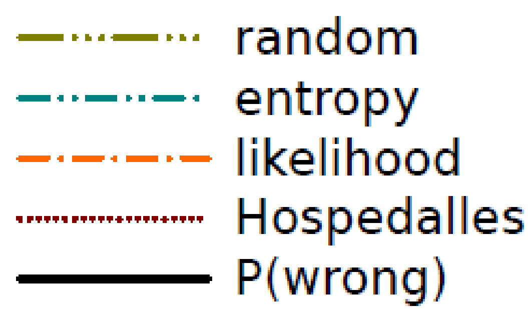
Handwriting digits data

- Digits problem: Recognising the ten handwritten digits.





	digits	
	discovery	classification
random	915.2	54.6
entropy	974.0	57.1
likelihood	1060.2	61.9
Hospedales	1207.4	69.5
P(wrong)	1133.6	69.7



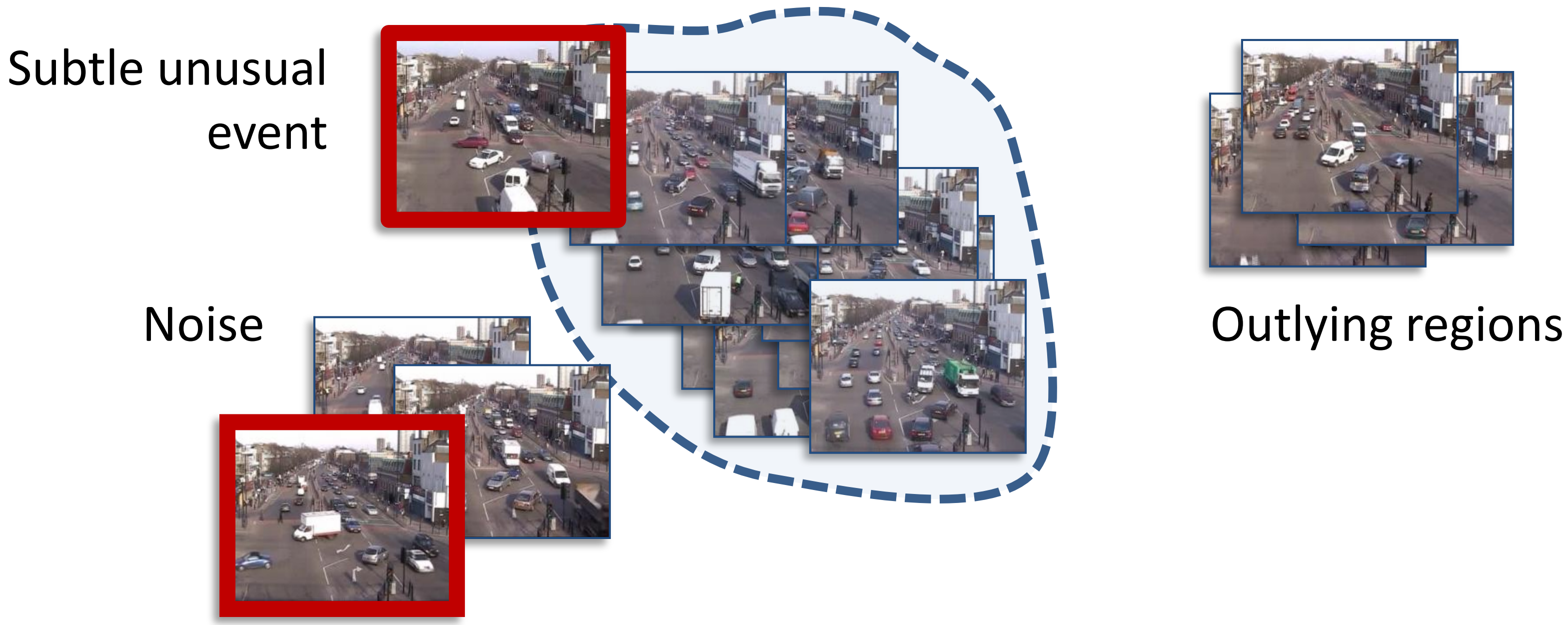
STEAM-BASED ACTIVE DISCOVERY AND LEARNING FOR VIDEO SURVEILLANCE

Problem

- Detect unusual event on-the-fly with limited data

State of the arts

- Unsupervised 1-class learning strategy
  [Mehran-CVPR09, Wang-TPAMI09, Kim-CVPR09]
  - Hard to detect visually subtle and ambiguous events
  - Confused between noise and genuine unusual event
  - Outlying normal regions causing false alarms



State of the arts

- Learning from human supervision
  - Resolving ambiguities
  - Arbitrating false alarms

- Fully supervised learning
  - Exhaustive annotation are time consuming
  - Unusual events not known *a priori*
  - Not all samples are critical for learning

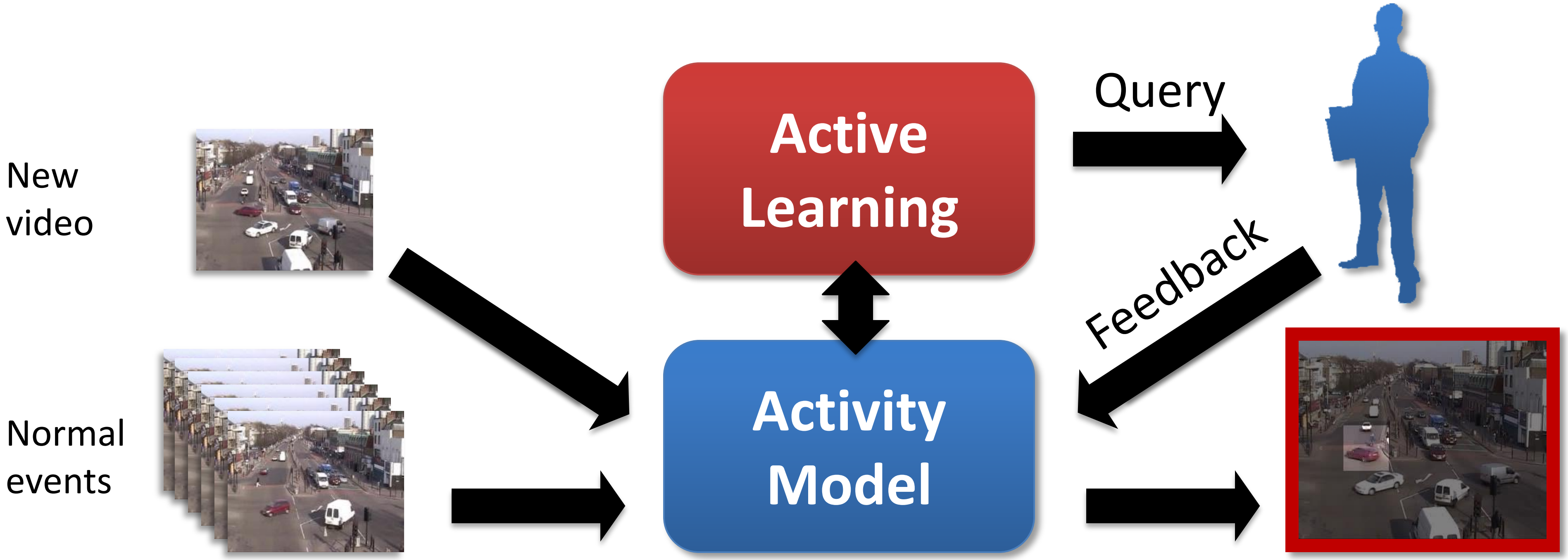
Unsupervised learning

Supervised learning

Exploit minimal human supervision without compromising the performance?

Motivation

- Some samples are more informative than others
- Select critical and informative queries for labelling based on predefined criteria

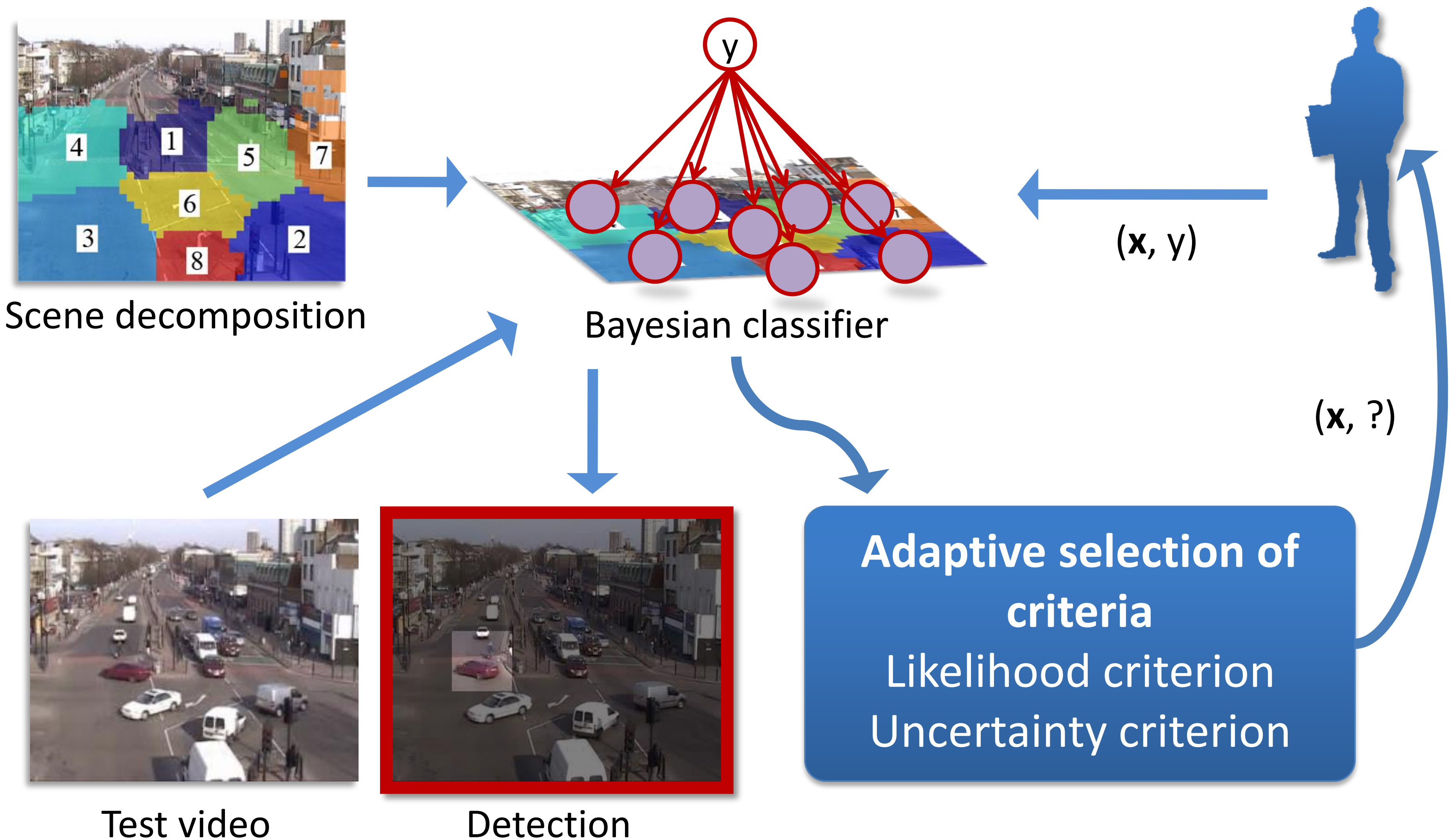


Why is it difficult?

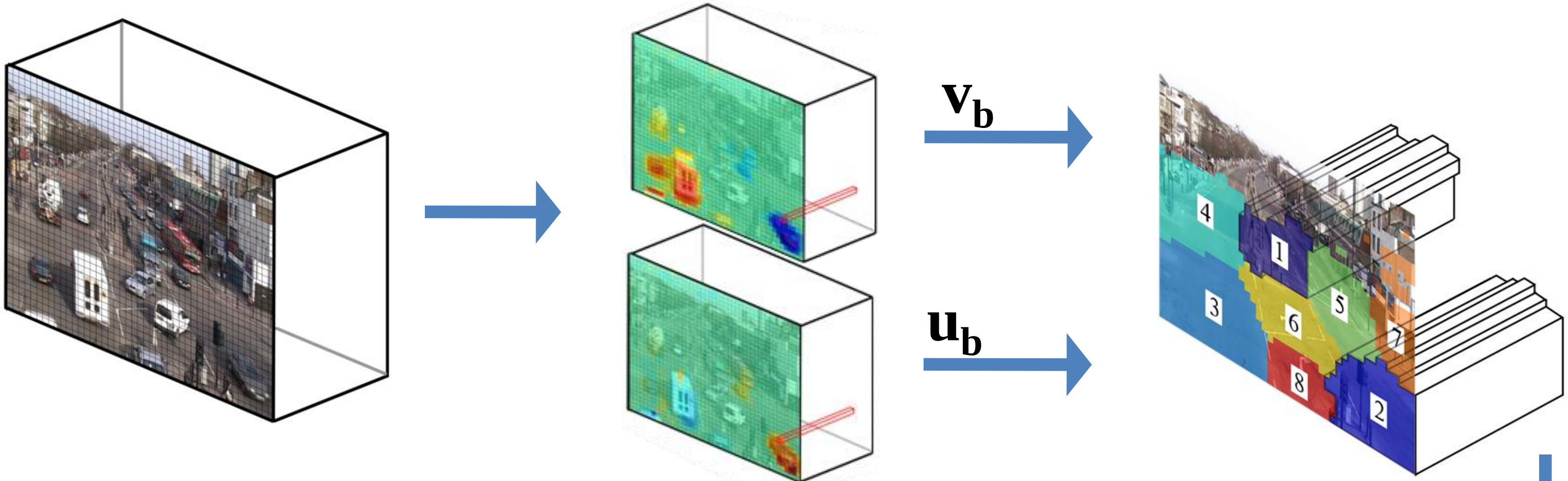
- Joint discovery of unknown events and refinement of classification boundary
- Stream-based observations demand on-the-fly decision

Overview of the approach

Stream-based Active Unusual Detection



Activity representation



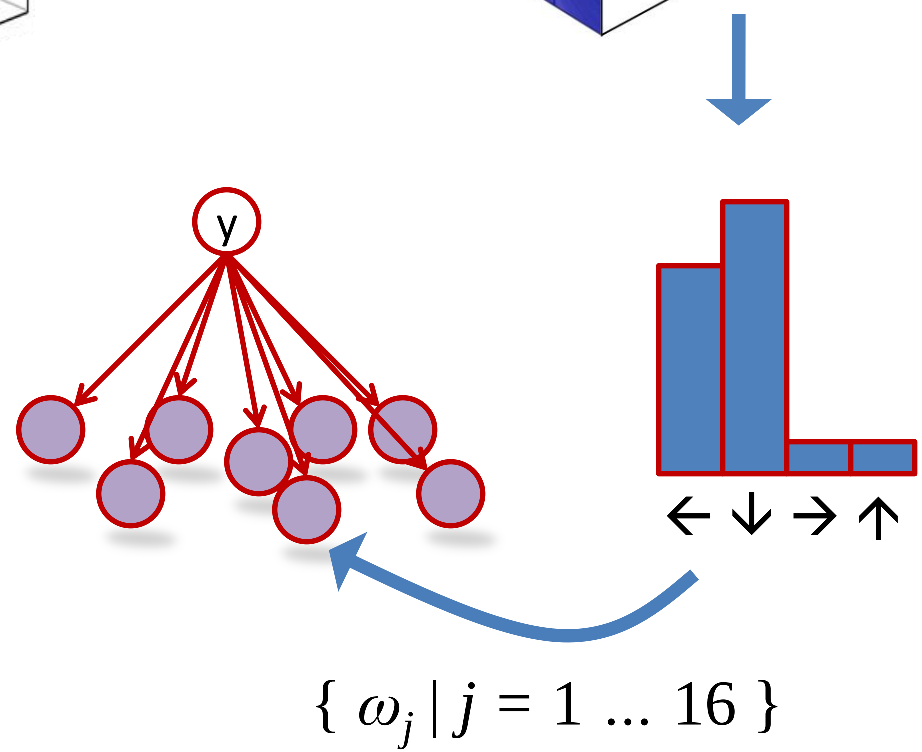
Optical flow represented as time series

Block clustering:

- spatial proximity
- activity correlation

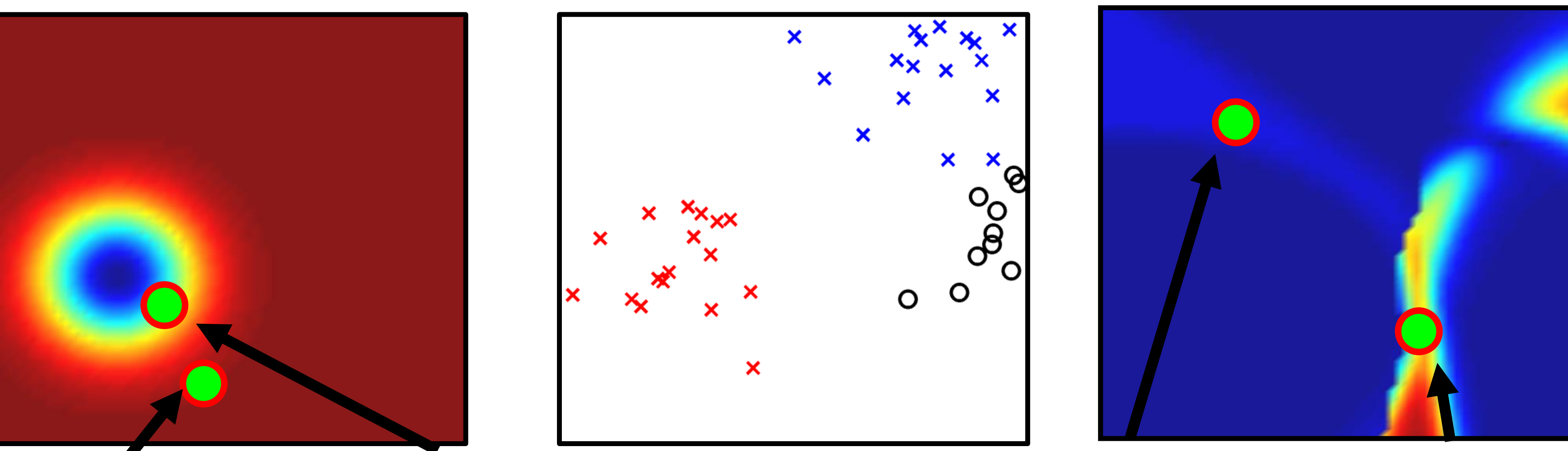
Motion features as words

Regions as graph nodes



Query Criteria

- Likelihood criterion :
  - Favour low-likelihood points
  - Discover unknown events
- Uncertainty criterion :
  - Favour ambiguous points
  - Refine classification boundary
  - Reformulate Query-by-Committee [Seung-COLT92, Engelson-JAIR99]



Adaptive criteria selection

- Best suited criterion for specific dataset at different phases of learning are not known *a priori*
- Favour criterion that returns query that brings more influence to a model

$\mathrm{Mult}(\mathbf{w}_t)$

Likelihood criterion

Uncertainty criterion

$$w_{a,t} = \beta w_{a,t-1} + (1-\beta)\frac{\overline{\mathcal{KL}}_a(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}})}{\sum_{a=1}^{A} \overline{\mathcal{KL}}_a(\boldsymbol{\theta} \parallel \tilde{\boldsymbol{\theta}})}$$

Controls updating rate

Weight of a criterion

Kullback–Leibler divergence of a model before and after it is trained using a sample

Kullback–Leibler divergence

Tendency to be selected

Algorithm Summary

Algorithm 1: Stream-based active unusual event detection.

Input: Data stream $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_t, \dots)$, an initial classifier $\mathcal{C}_0$ trained with a small set of labelled samples from known classes

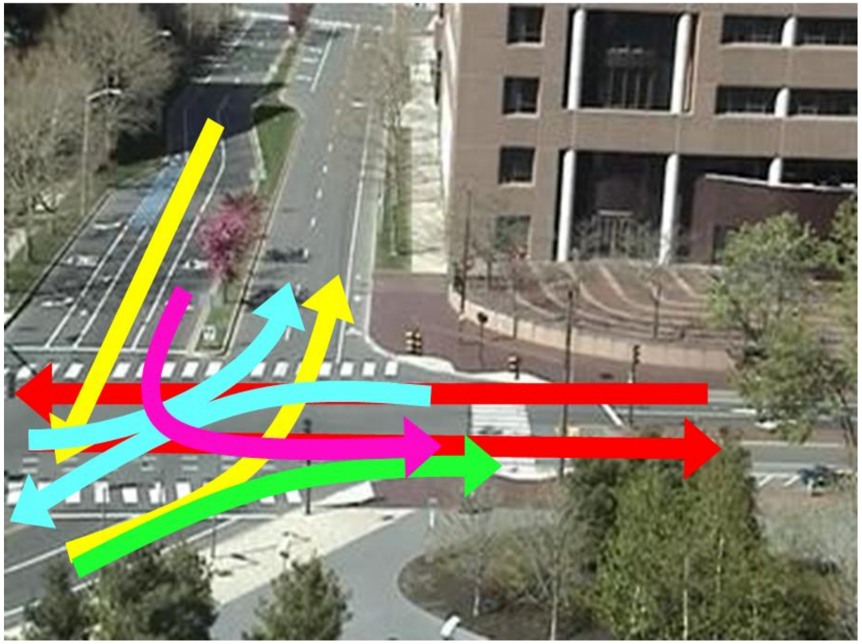
Output: A set of labelled samples $\mathcal{S}$ and a classifier $\mathcal{C}$ trained with $\mathcal{S}$

Set $\mathcal{S}_0$ = a small set of labelled samples from known classes ;
for t *from* $1, 2, \dots$ *until the data stream runs out* **do**
    Receive $\mathbf{x}_t$;
    Compute p_t^l (Eqn. (3)) ;
    Compute p_t^u (Eqn. (6)) ;
    Select query criterion by sampling $a \sim \mathrm{Mult}(\mathbf{w})$, assign p_t^{query} based on the selected criterion ;
    if $p_t^{\mathrm{query}} \geq \mathrm{Th}$ **then**
        Request y_t and set $\mathcal{S}_t = \mathcal{S}_{t-1} \bigcup \{(\mathbf{x}_t, y_t)\}$;
        Obtain classifier $\mathcal{C}_{t+1}$ by updating classifier $\mathcal{C}_t$ with $\{(\mathbf{x}_t, y_t)\}$;
        Update query criteria weights $\mathbf{w}$ (Eqn. (9)) ;
    else
        $\mathcal{S}_t = \mathcal{S}_{t-1}$;
    end
end
Unusual event is detected if $p(y = \mathrm{unusual}|\mathbf{x})$ is higher than $\mathrm{Th}_{\mathrm{unusual}}$;

C. Loy, T. Xiang and S. Gong, "Stream-based Active Anomaly Detection", in *Asian Conference on Computer Vision*, 2010.

Experiments

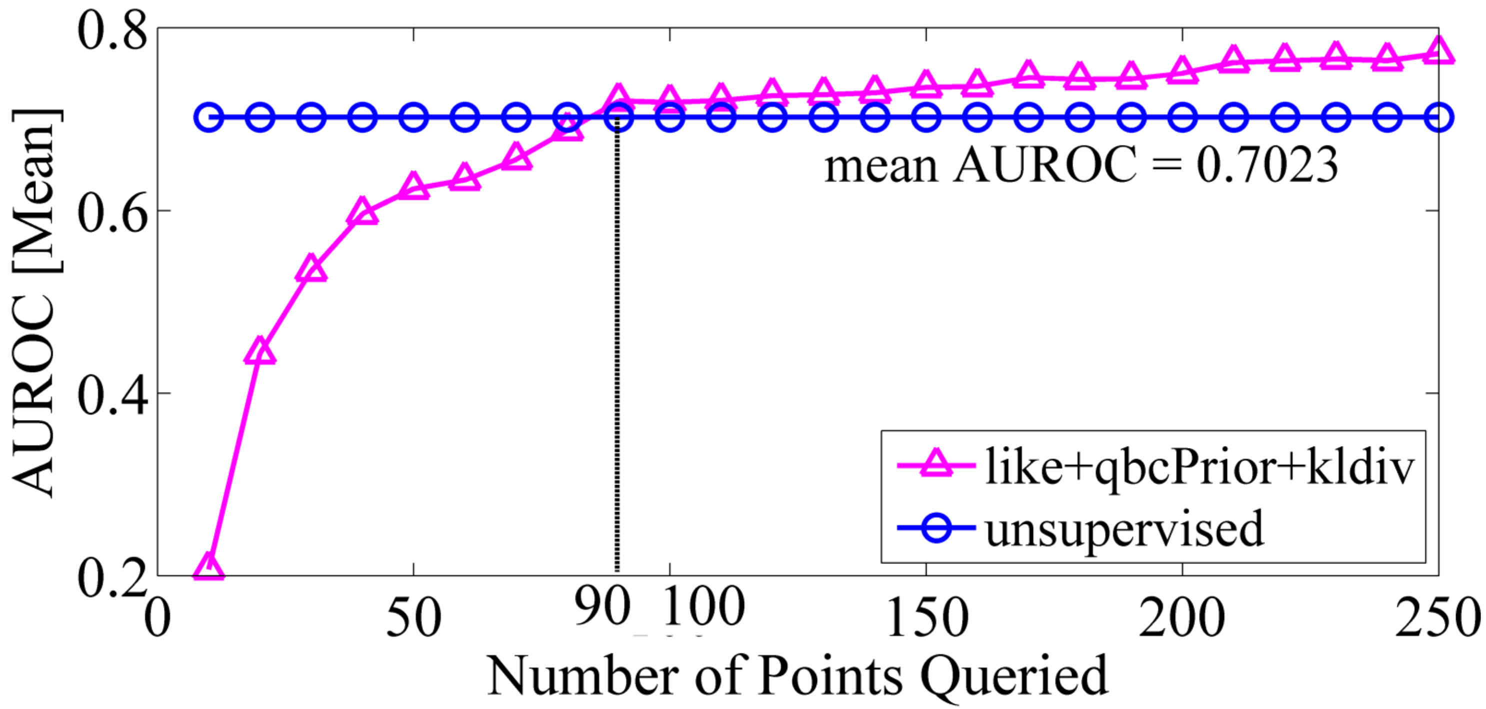
MIT Traffic Dataset



QMUL Junction Dataset

- Dominant traffic flows as normal classes
- Unusual events including illegal u-turns, improper lane usage etc.

Active vs. unsupervised learning



	# training samples	mean AUROC
unsupervised	800	0.7153 ± 0.0085
like + qbcPrior + kldiv	250	0.7720 ± 0.0078

Active vs. unsupervised learning

Unsupervised Learning

Active Learning

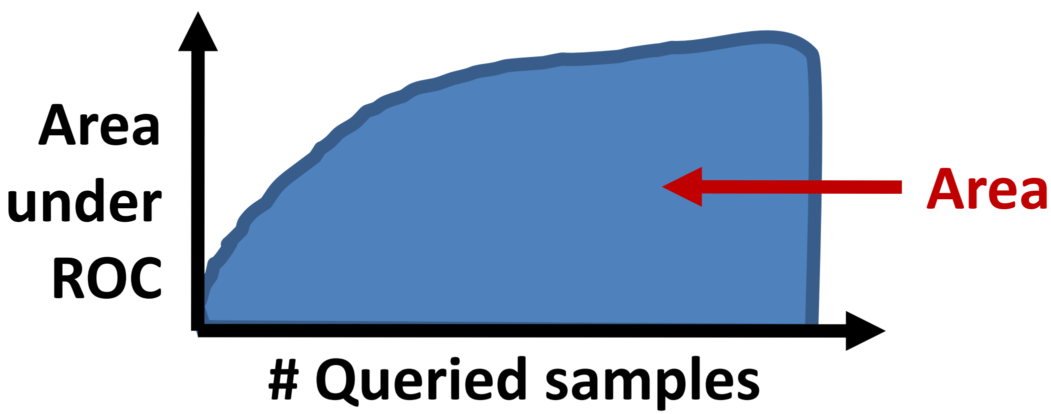
False positive

True positive

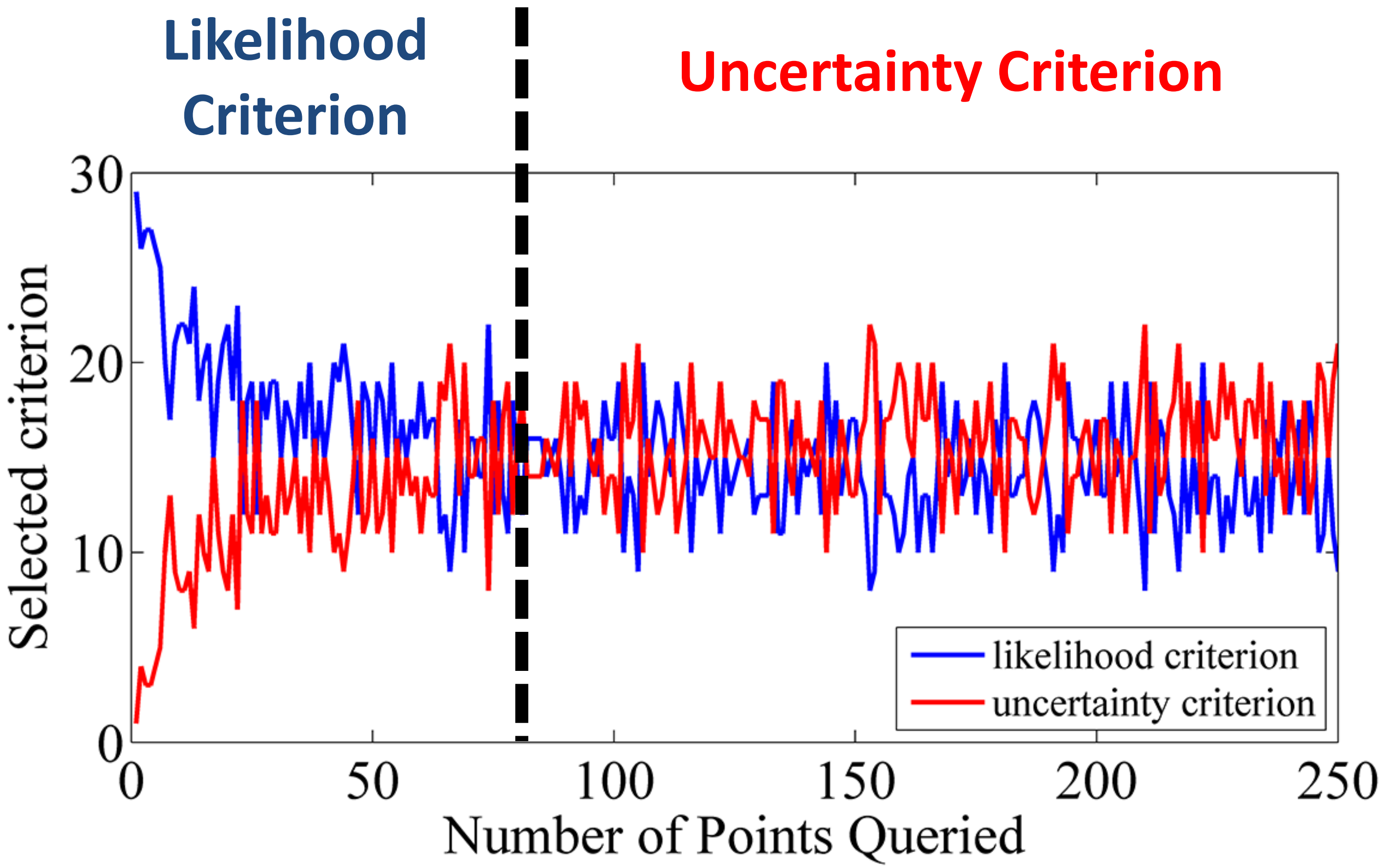
Comparison with other strategies

	MIT Traffic Dataset	QMUL Junction Dataset
Proposed method	**12.26**	**16.52**
Random sampling	11.75	15.22
Likelihood	11.87	16.52
QBC with Vote Entropy	11.83	16.37
QBC with Prior	11.90	16.40
Likelihood + QBC with interleave strategy	11.95	16.33

Performance metric



Adaptive criteria selection



Summary

- Learning from human feedback
  - Resolving ambiguity
  - Arbitrating false alarm
- On-the-fly decision using stream-based active learning
- Adaptive selection of different criteria
  - Discovering unknown classes and regions
  - Refining classification boundary
- Limitations
  - The naïve-Bayes based activity model is weak
  - How do we model a new class with:
    - A single sample
    - Weak label

LEARNING RARE EVENTS USING WEAKLY-SUPERVISED TOPIC MODEL

Rare Events: Weak Supervision

Challenge: Detect Events…

- Too visually subtle to be obvious anomalies
- Too rare to learn a traditional classifier
  - ➢ N-shot learning
- With only weak supervision
  - ➢ Important for practical use!

Rare Events: Weak Supervision

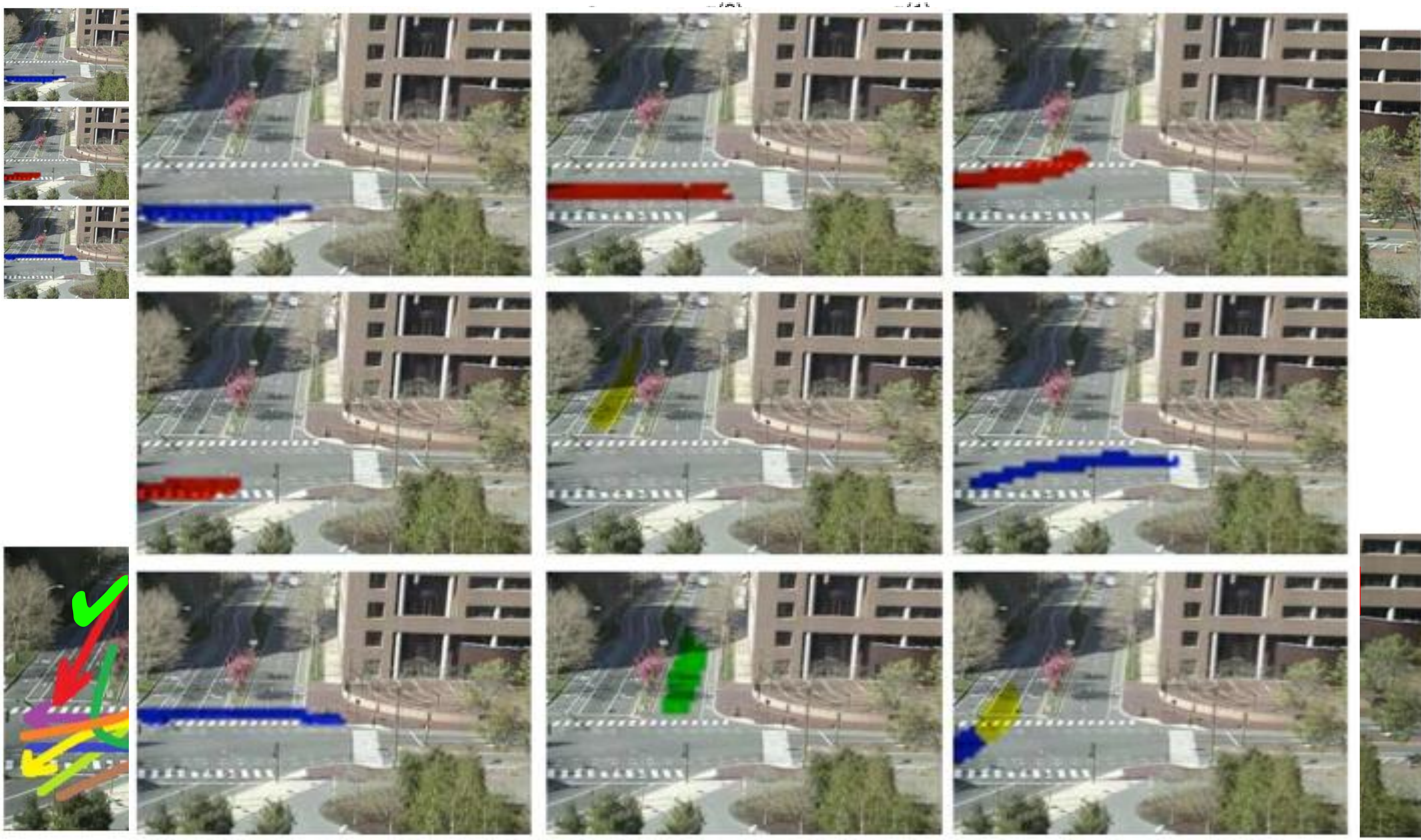
Weakly supervised joint topic model: Learning Example

1 Example Each

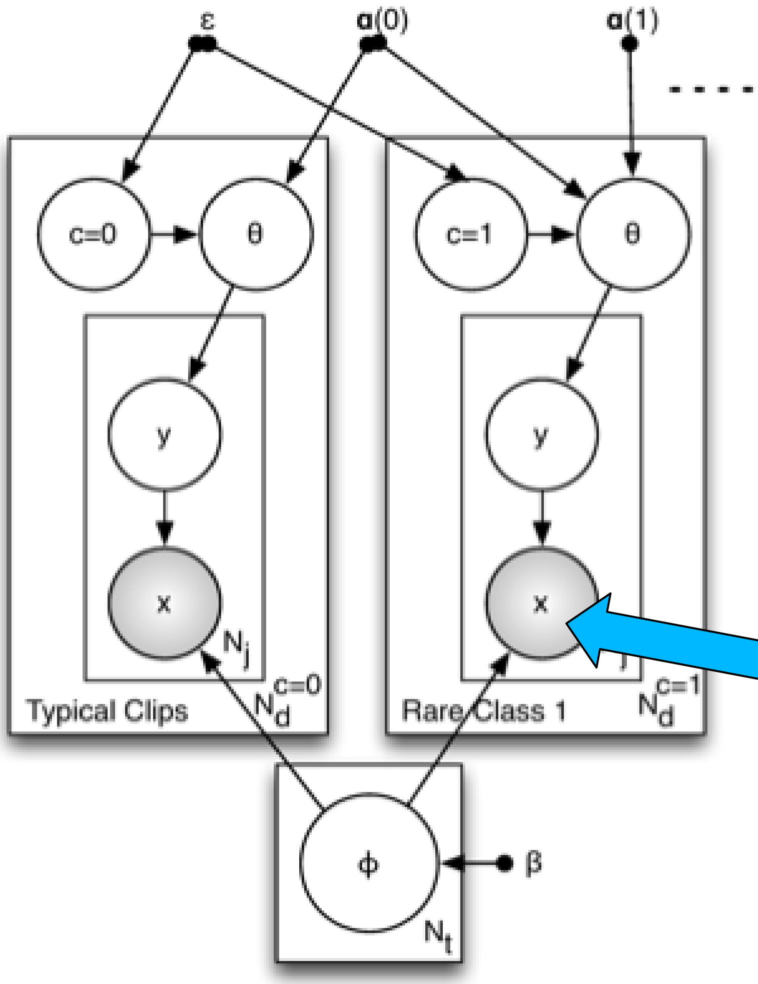
100 Examples

Rare Events: Weak Supervision

Weakly supervised joint topic model: Learning



WSJTM: Inference



Classify: Compute p(C|X)

- Bayesian Model Selection
  - Variational Importance Sampler

Locate: Infer p(Y|X,C)

- Gibbs

Behavior Profiling: Results

Rare Events: Weak Supervision

Summary:

- Learning: MCMC collapsed Gibbs sampling
  - Almost real time
- Inference: Model Selection by Variational Importance Sampler
  - >> Real-time.
- Weak (1-bit) supervision
  - Outperforms LDA, S-LDA, SVM, etc.
- Published in: T. Hospedales, J. Li, S. Gong and T. Xiang, "Identifying Rare and Subtle Behaviours: A Weakly Supervised Joint Topic Model", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2011.

Thank You