ASK-the-Expert: Active learning based knowledge discovery using the expert

Kamalika Das

Data Science Group

NASA Ames Research Center

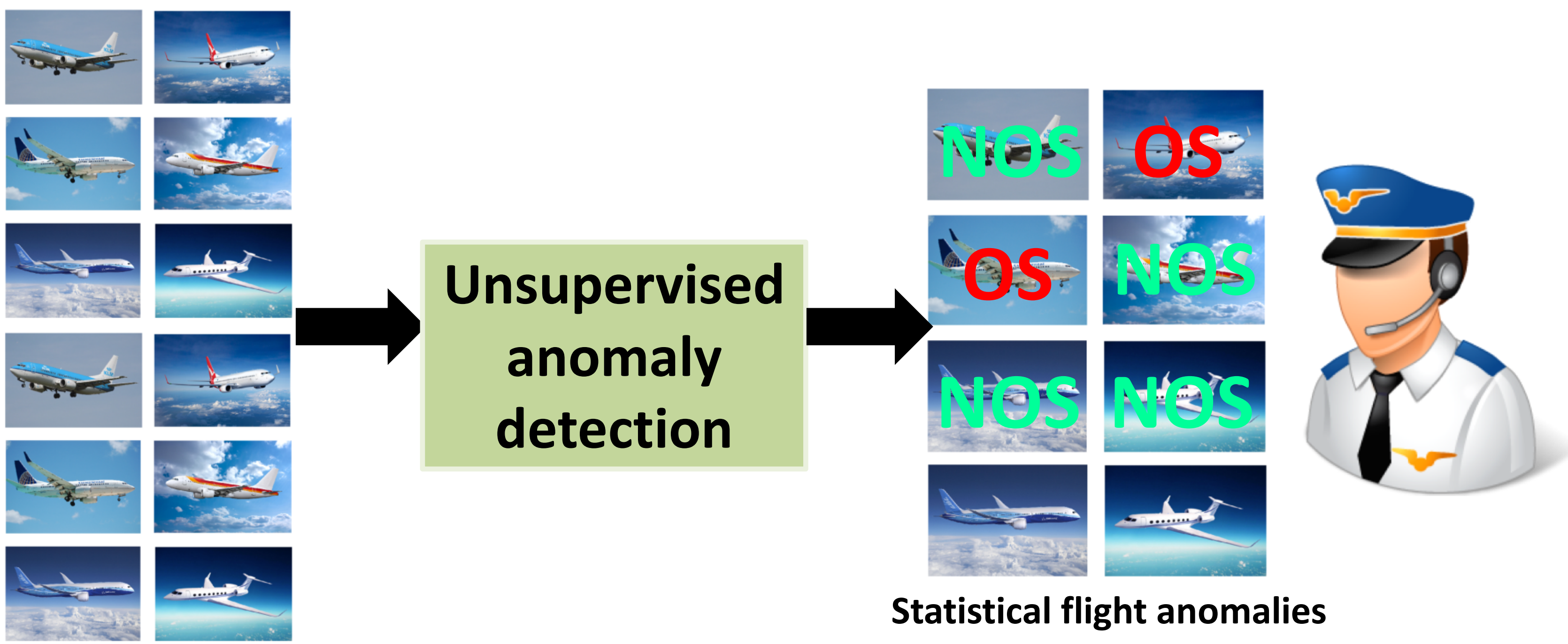
ML Workshop, August 2017

Roadmap

- Problem description
- State-of-the-art
- Proposed framework
- Tool description
- Algorithms
- Performance analysis
- Summary

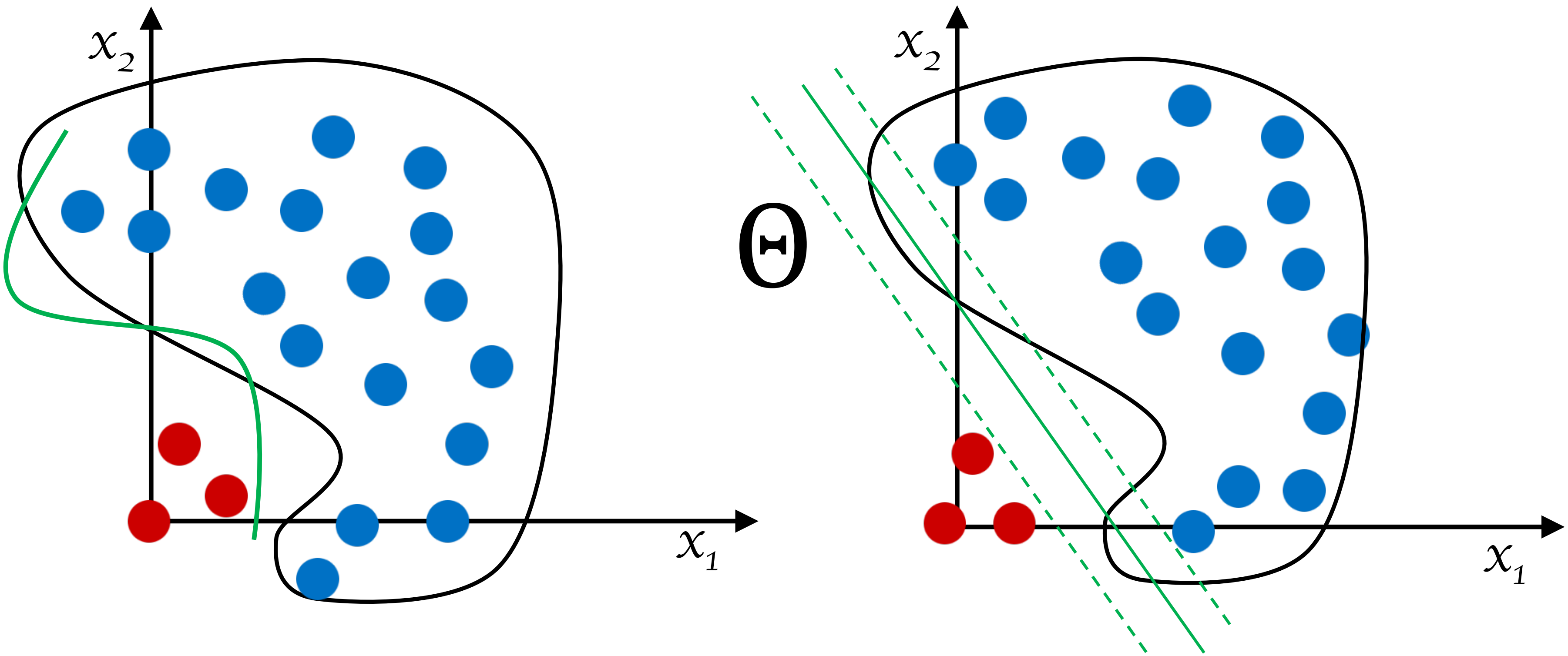
Problem

- Identify safety events in flight operational data
- Unsupervised anomaly detection
- SME review of anomalies



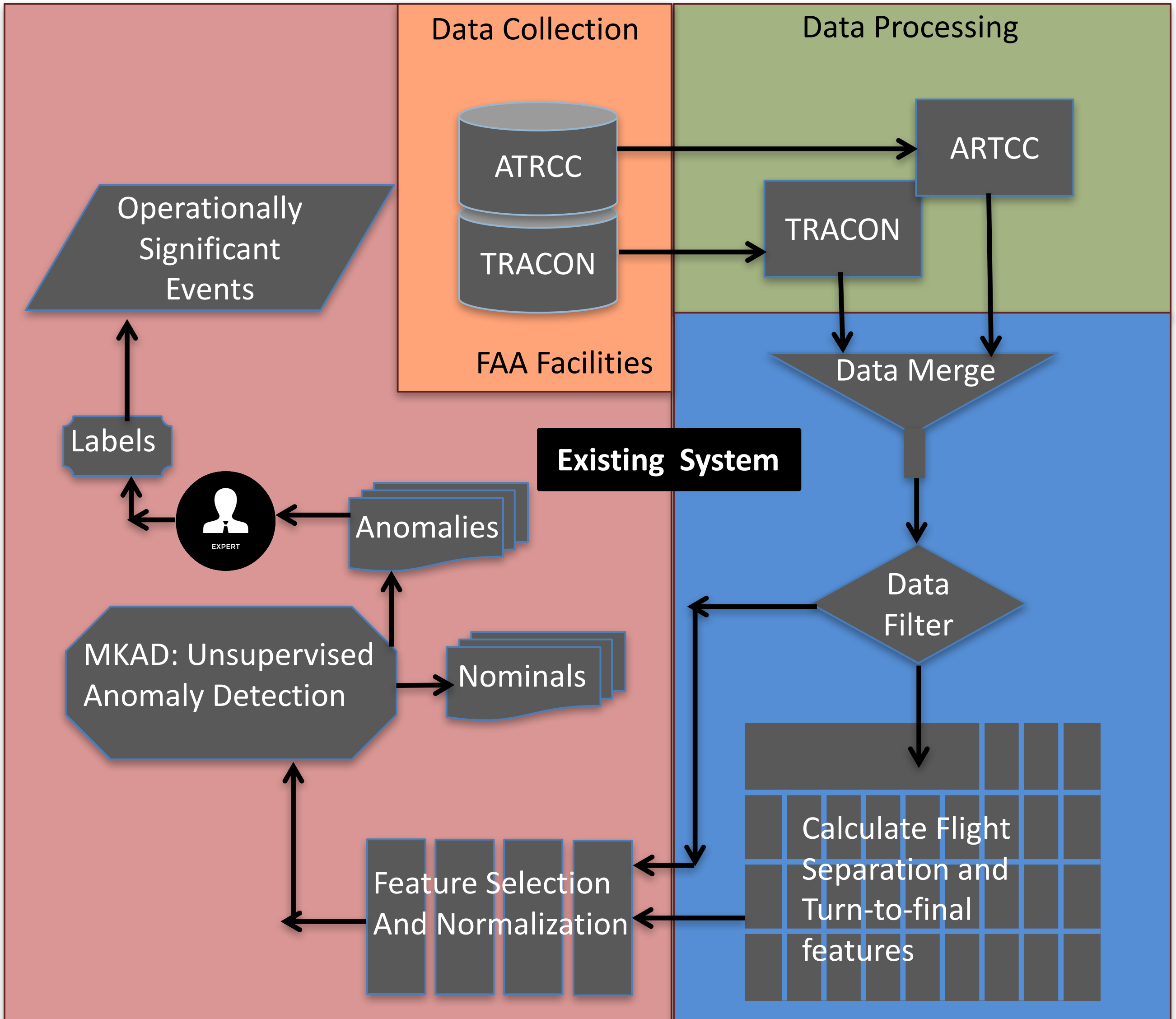
Unsupervised anomaly detection

- Lack of definition of ‘safety’ incident
- One-class SVM based anomaly detection

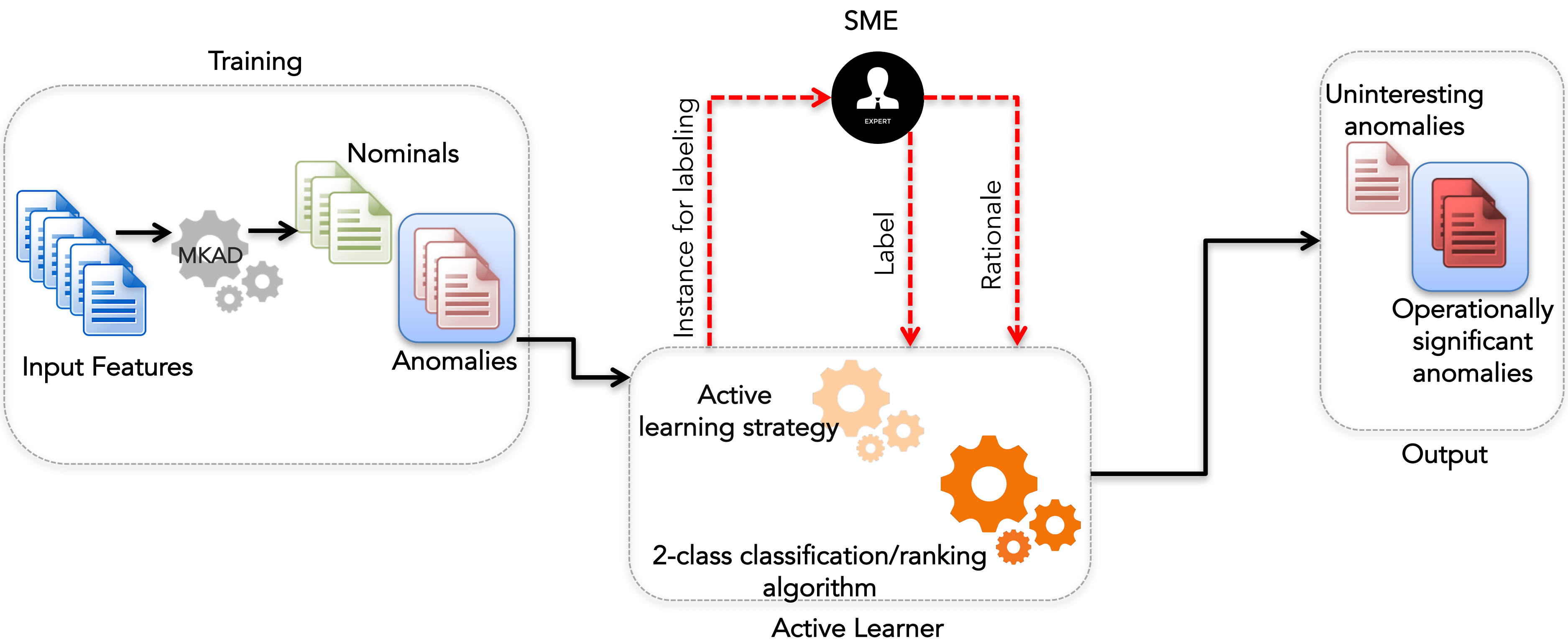


[+]S. Das, B. Matthews, A. Srivastava, N Oza. 2010. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. In Proceedings of the 16th ACM SIGKDD (KDD '10). 47-56.

State of the art

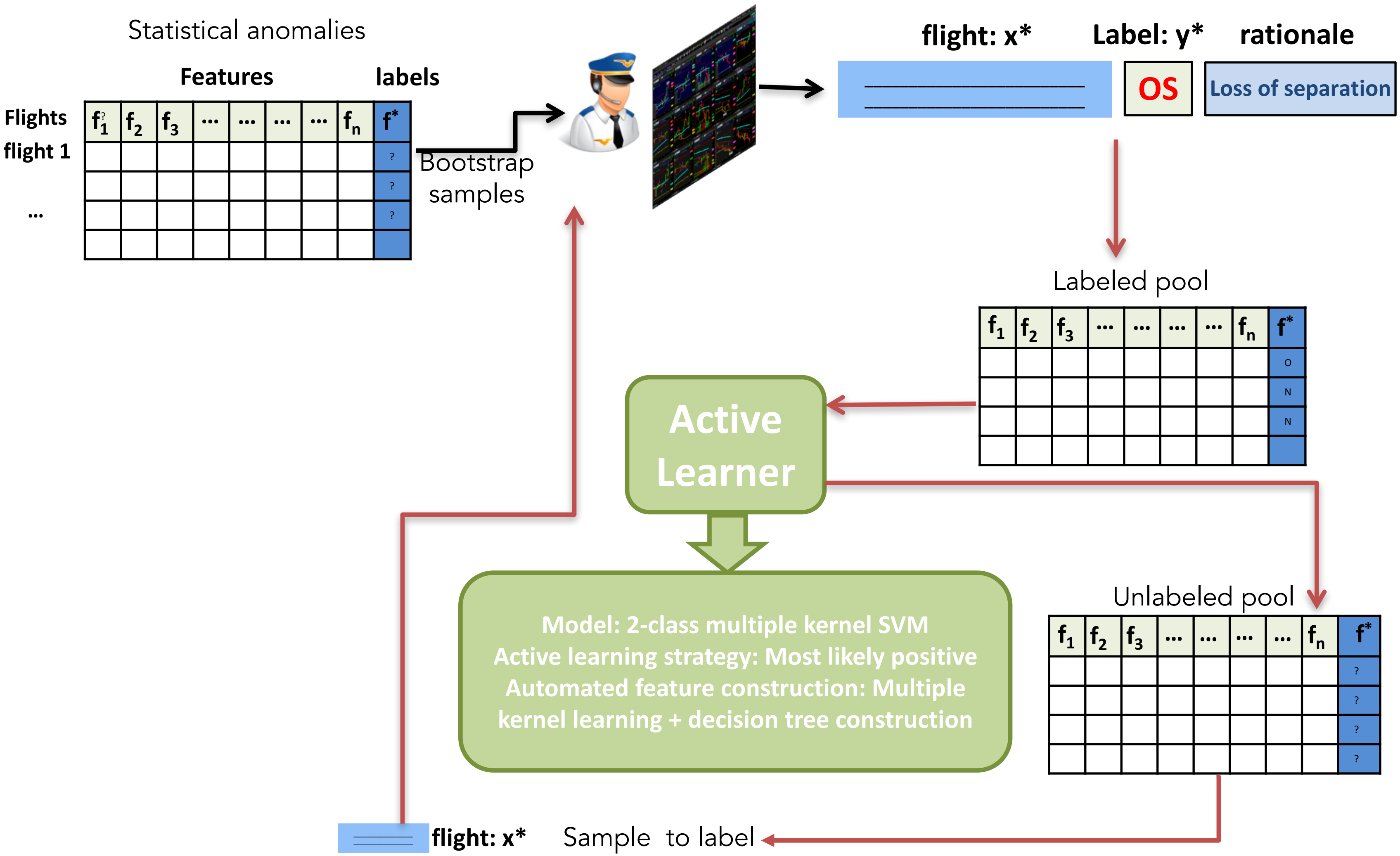


Proposed approach

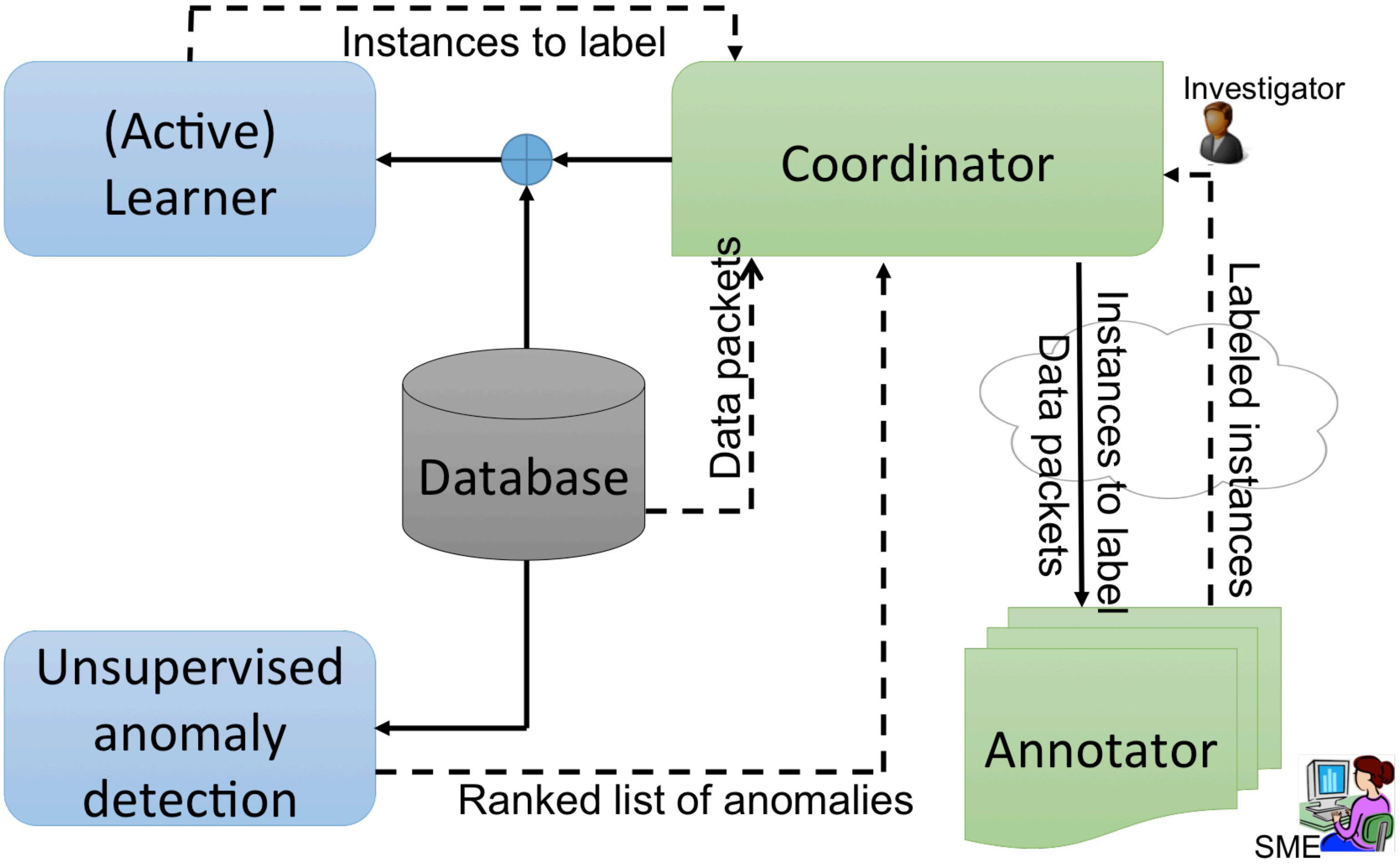


Active learning with rationales framework

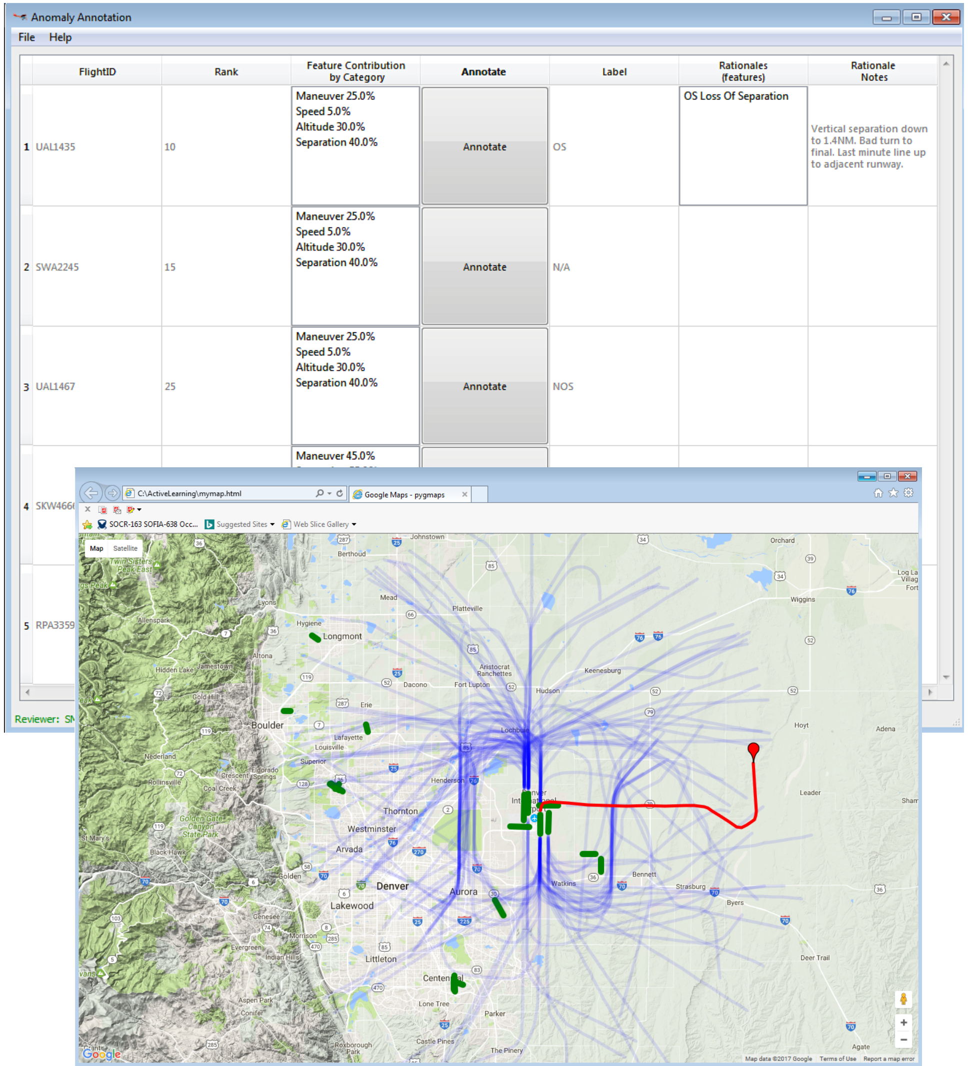
Active learning framework

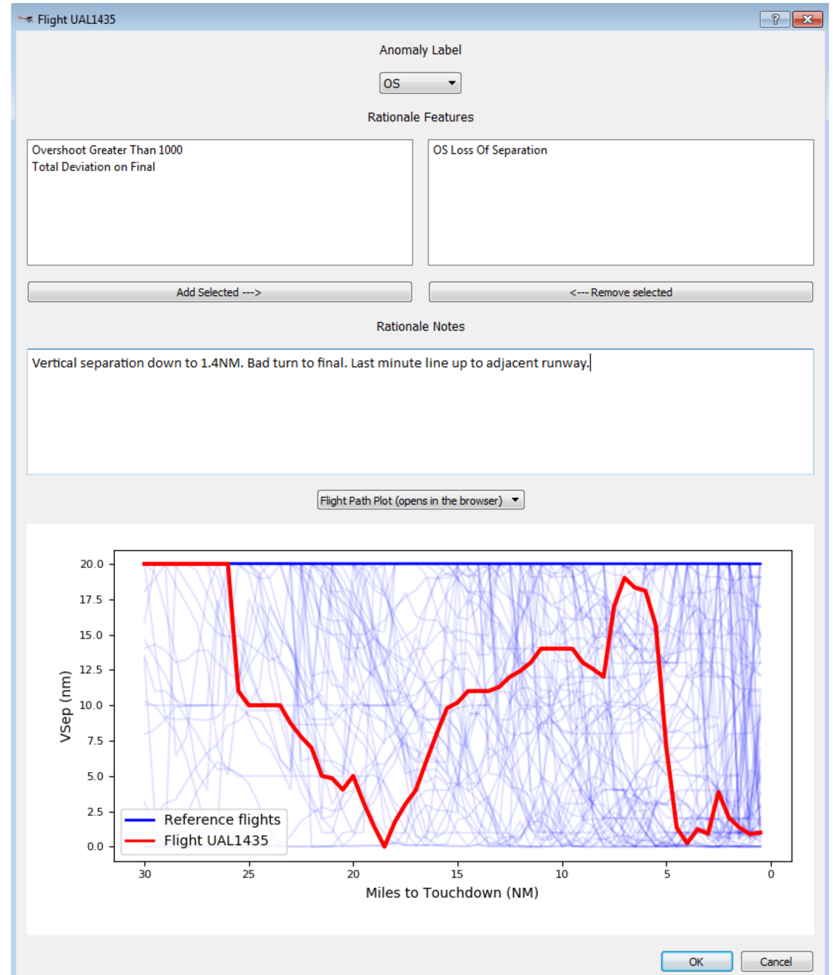


ASK-the-Expert tool: architecture

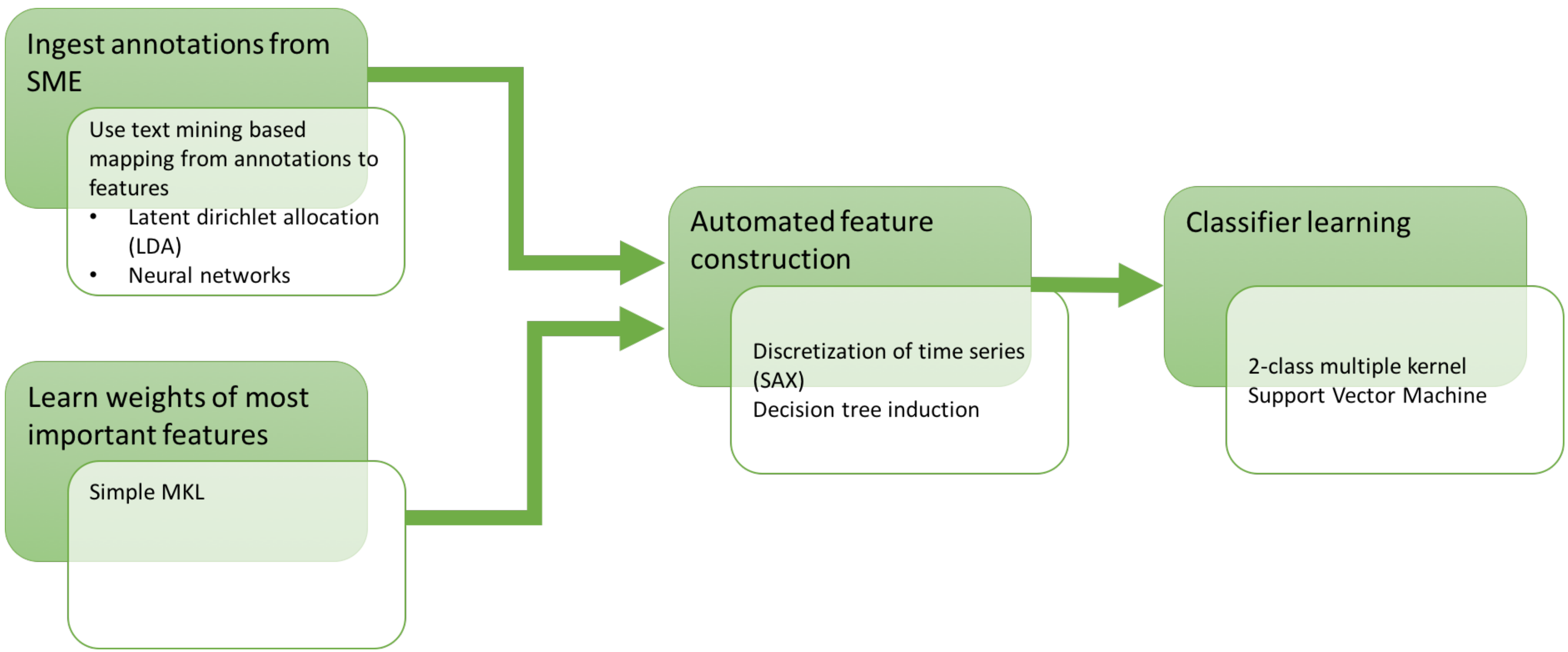


Annotator component



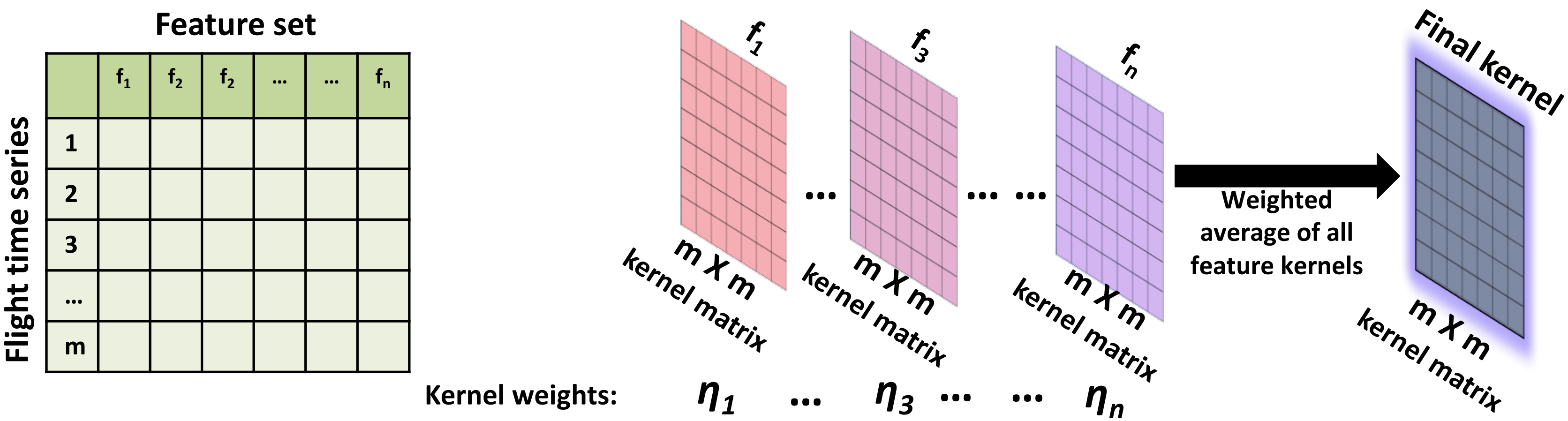


Coordinator component



Multiple kernel support vector machine

- Multiple kernel 2 class SVM: classifying between operationally significant (OS) and uninteresting (NOS) flights



- 2-class SVM objective: $$\min_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i,j}\alpha_i\,\alpha_j\,\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j) - \sum_i y_i\,\alpha_i \quad \text{s.t.} \begin{cases} \sum_i \alpha_i = 0 \\ 0 \le y_i\,\alpha_i \le C \end{cases}$$

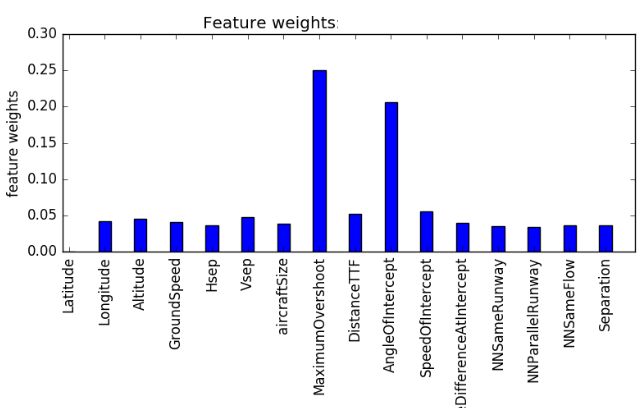
- Decision function: $$f(x) = \sum_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

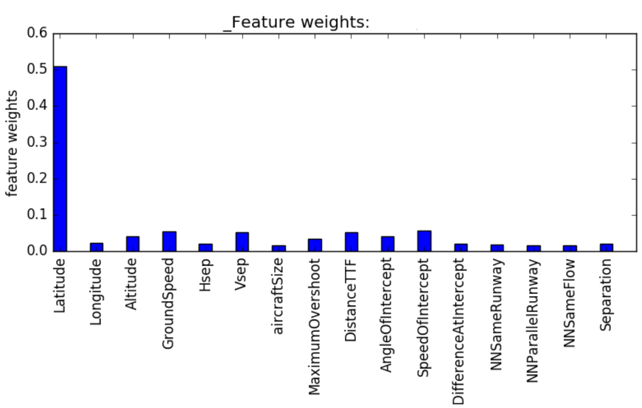
Rationale feature construction

- How to set weights: $\eta_1, \eta_2, \ldots, \eta_n$

$$K_\eta = \sum_{m=1}^{P} \eta_m k_m\left(x_i^m, x_j^m\right) \qquad s.t.\, \eta_m >= 0 \;\&\; \sum \eta_m = 1$$

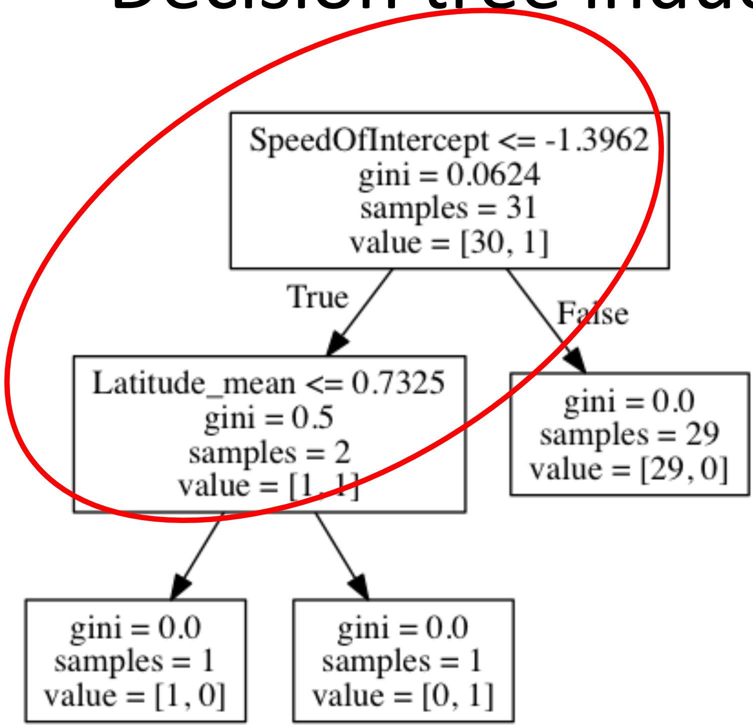
- Simple MKL algorithm
  – Modified objective function
  – Alternates between optimizing classifier margin and weights of kernels

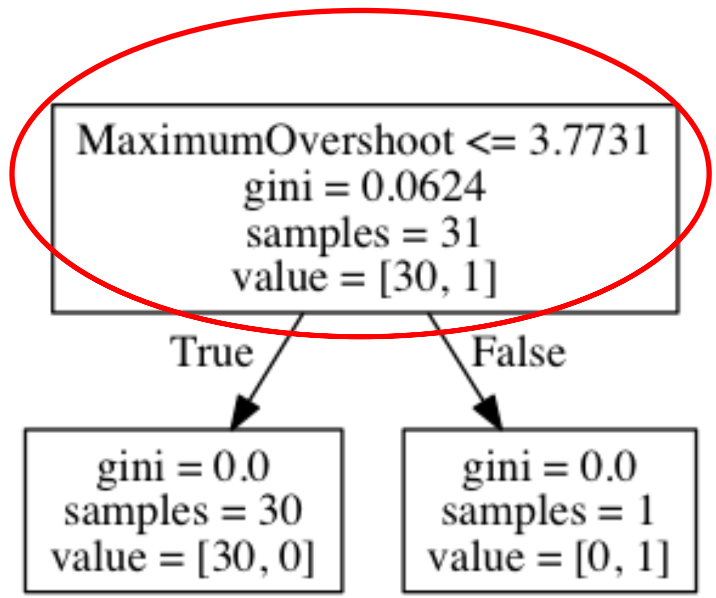




Rationale feature construction

- Decision tree induction

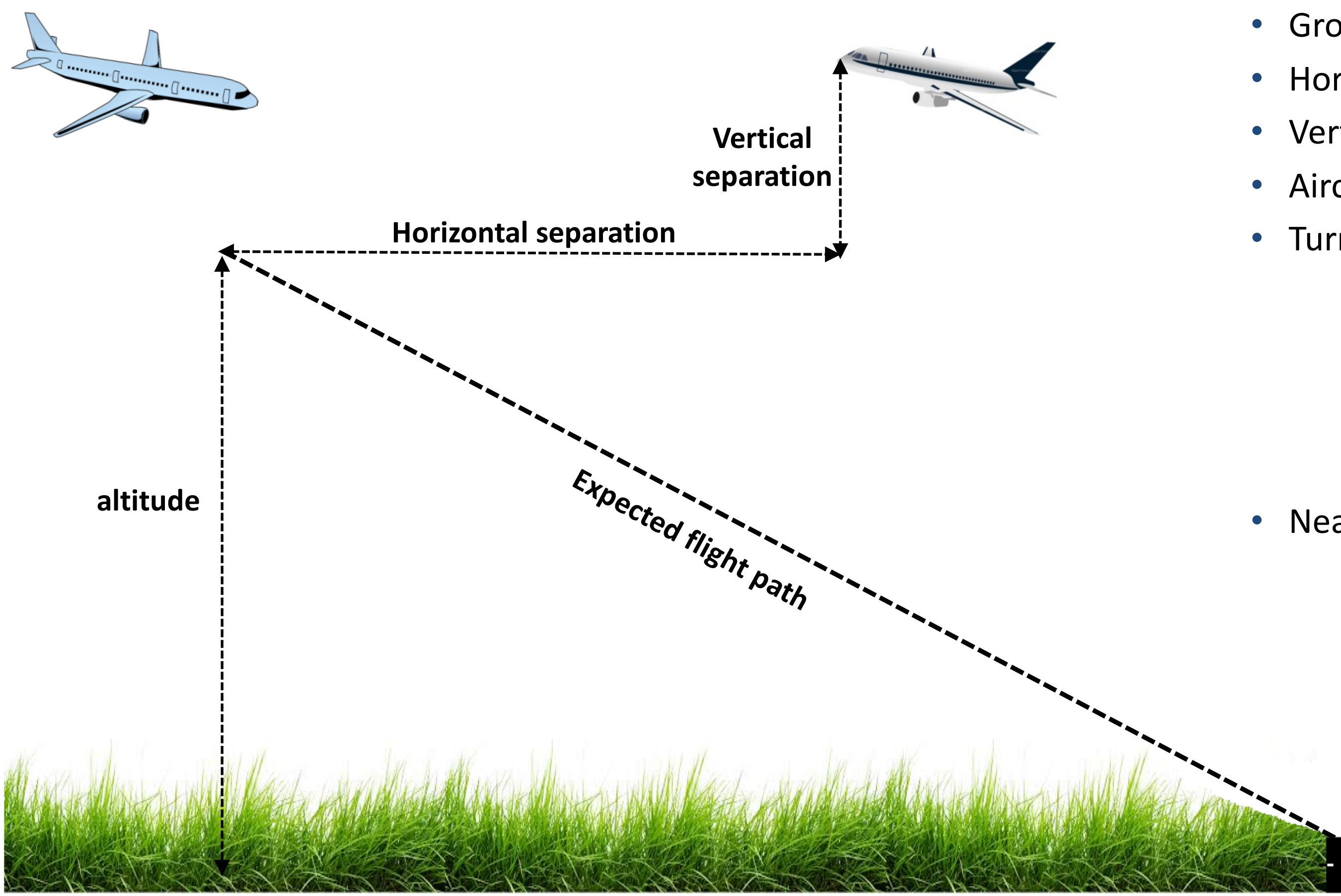




Data

ORIGINAL FEATURES

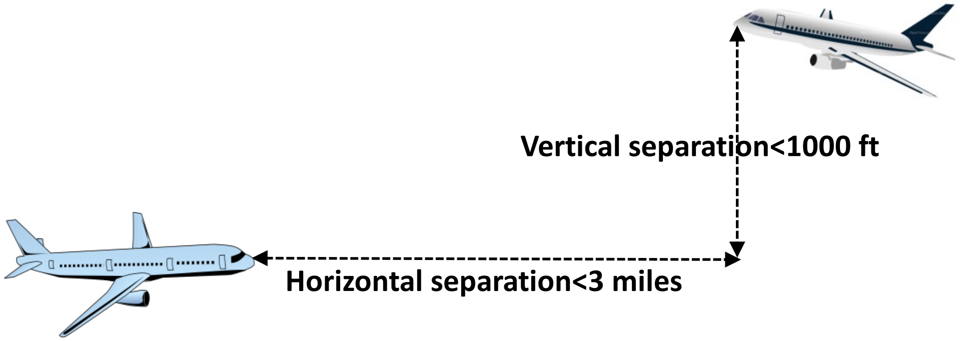
- Latitude
- Longitude
- Altitude
- Ground speed
- Horizontal separation
- Vertical separation
- Aircraft size
- Turn-to-final (TTF) parameters:
    - Maximum overshoot
    - Speed at TTF
    - Distance at TTF
    - Angle at TTF
    - Altitude difference at TTF
- Nearest neighboring (NN) flight info:
    - NN flight on same runway
    - NN flight on parallel runway
    - NN flight part of the same flow



Rationale features

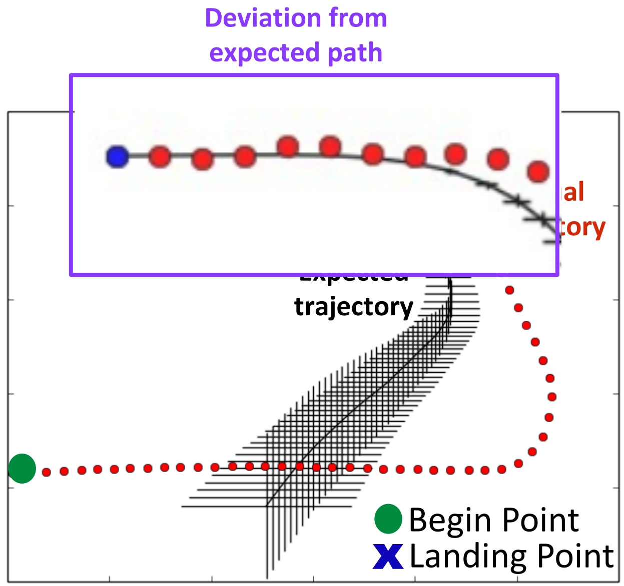
“Loss of separation”

- Horizontal separation < 3 miles AND Vertical separation < 1000 ft AND nearest neighboring flight is not on parallel runways and not part of the same flow



“Large overshoot”

- Maximum overshoot is greater than a threshold based on values of flights with positive labels



“Unusual flight path”

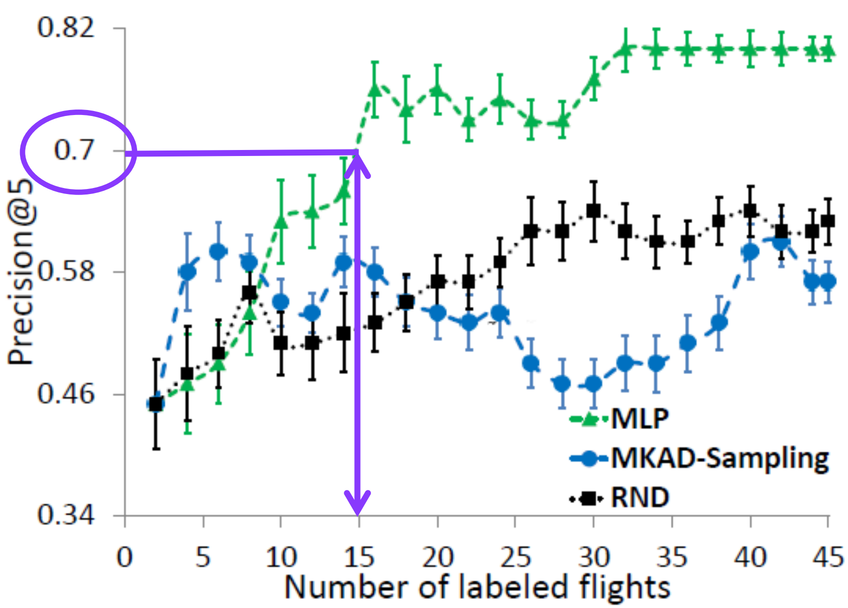
- Overall deviation from expected (average) trajectory of all landing flights on that runway

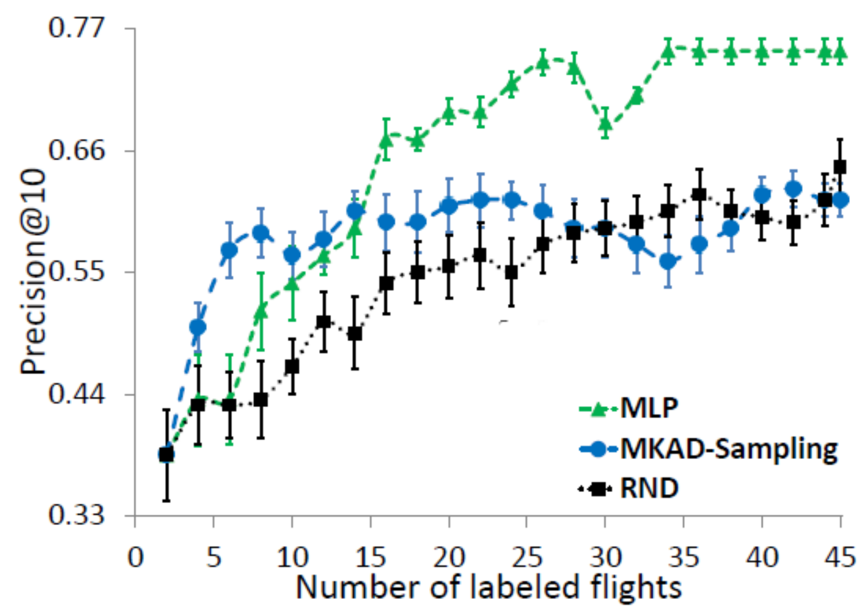
Experimental setup

- Data set: 30 NM airspace around Denver International Airport for Aug 2014
  - Training set: ~2400 flights
  - Statistical anomalies: 153
  - OS flights: 24
- 2 fold cross validation with 10 random bootstraps for each fold

Performance analysis

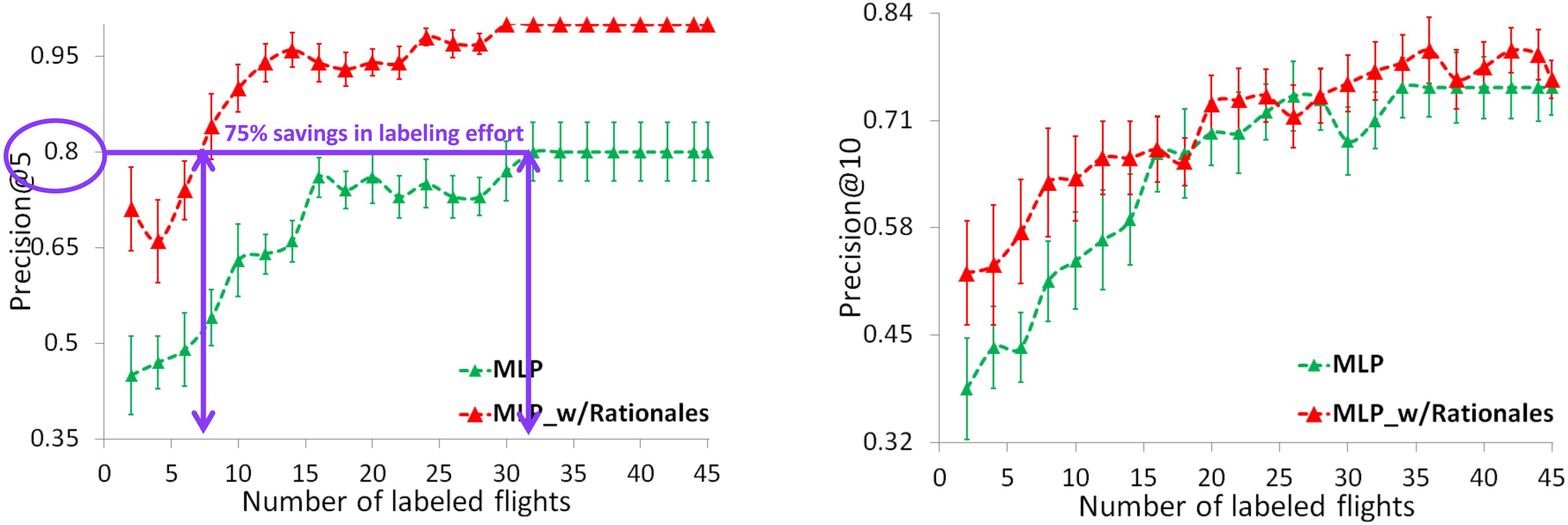
- Metrics: precision@5 and precision@10
- Most-likely positive strategy $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{U}} P_\theta(\hat{\mathbf{y}}^+|\mathbf{x})$





Learning curves for different active learning strategies

Performance analysis



Learning curves for most likely positive strategy with and without rationales

Performance analysis

	Target *precision*@5						Target *precision*@10					
Method	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**
RND	6	25	n/a	n/a	n/a	n/a	12	18	33	n/a	n/a	n/a
MKAD-Sampling	4	6	n/a	n/a	n/a	n/a	4	6	13	n/a	n/a	n/a
MLP	5	10	16	32	n/a	n/a	8	12	15	16	23	34
MLP_w/Rationales	2	2	2	8	10	29	2	5	7	11	19	29

Comparison of number of labeled flights required by various strategies to achieve a target performance measure. 'n/a' represents that the target performance cannot be achieved by a method even with 45 labeled flights.

Performance benefits

- Generalization
  - Two different test data sets: July 2014 and July 2015
  - Average improvement in precision@5: ~30%
  - Average improvement in precision @10: ~65%
- Review time
  - Up to 75% reduction in review time for same target performance

Summary

- Goal: to reduce SME review time of statistical anomalies identified using unsupervised anomaly detection
- Use active learning with rationales to learn 2-class classifier to distinguish between operationally significant and uninteresting anomalies
- Classifier generalizes to other data sets from the same domain
- Up to 75% reduction in SME review time

Acknowledgement

- This work is supported by Center Innovation Fund (CIF) 2017 award
- Team:
  - Nikunj Oza, NASA Ames Research Center
  - Bryan Matthews, SGT Inc.
  - Illya Avrekh, SGT Inc.
  - Manali Sharma, PhD Student, Illinois Institute of Technology
  - Sayeri Lala, Undergraduate Student, Massachusetts Institute of Technology

Thank You