



HAL
open science

AIDEme: An active learning based system for interactive exploration of large datasets

Enhui Huang, Luciano Di Palma, Laurent Cetinsoy, Yanlei Diao, Anna Liu

► **To cite this version:**

Enhui Huang, Luciano Di Palma, Laurent Cetinsoy, Yanlei Diao, Anna Liu. AIDEme: An active learning based system for interactive exploration of large datasets. 2019. hal-02430750

HAL Id: hal-02430750

<https://inria.hal.science/hal-02430750v1>

Submitted on 7 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AIDEme: An active learning based system for interactive exploration of large datasets

Enhui Huang, Luciano Di Palma, Laurent Cetinsoy, Yanlei Diao, Anna Liu

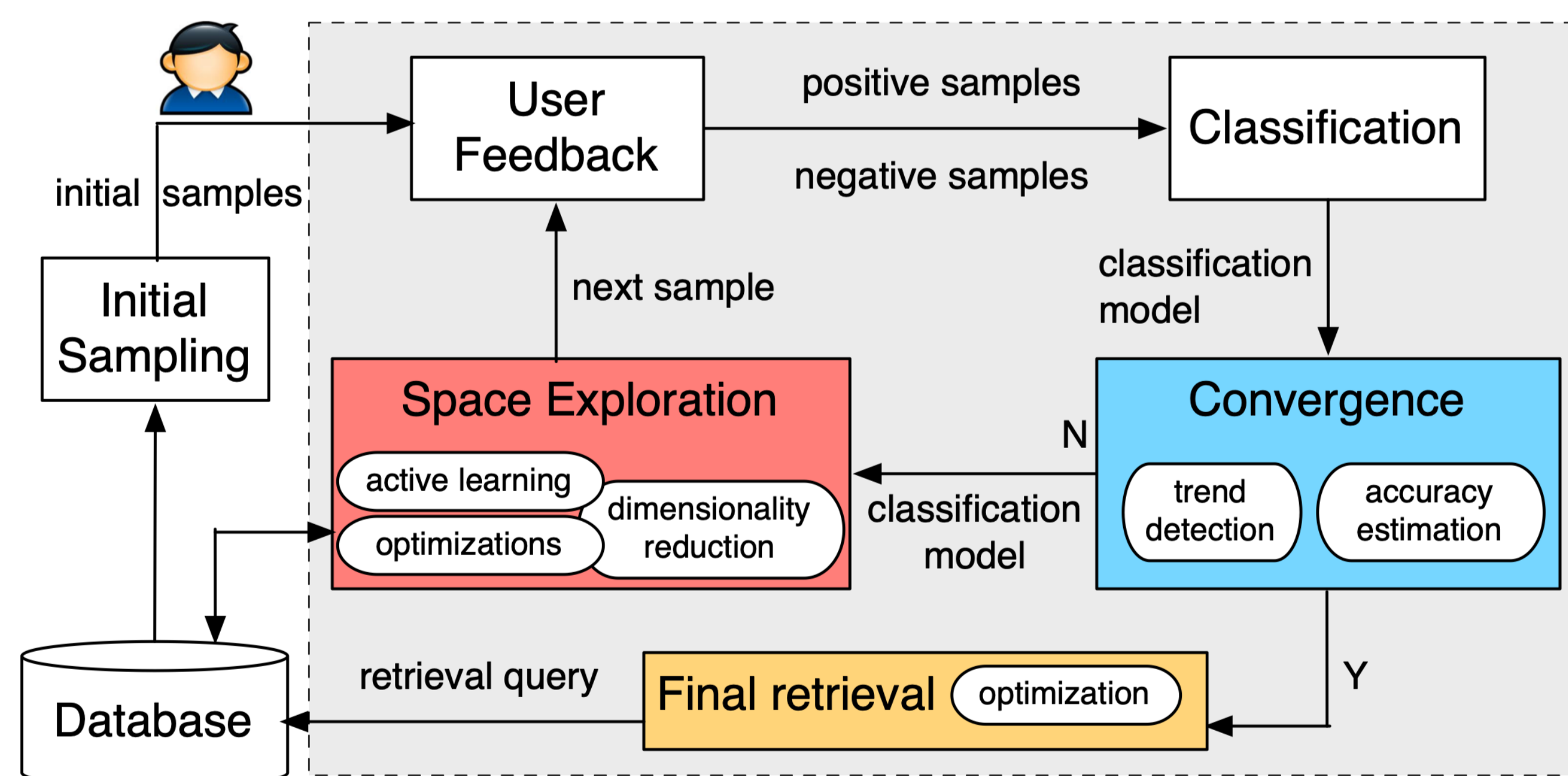
AIDEme is a scalable interactive data exploration system for efficiently learning a user interest pattern over a large dataset

Motivation



- An increasing gap between fast growth of data and limited human ability to comprehend data.
- A growing demand of data analytics tools that can bridge this gap and help the user retrieve high-value content from data more effectively.

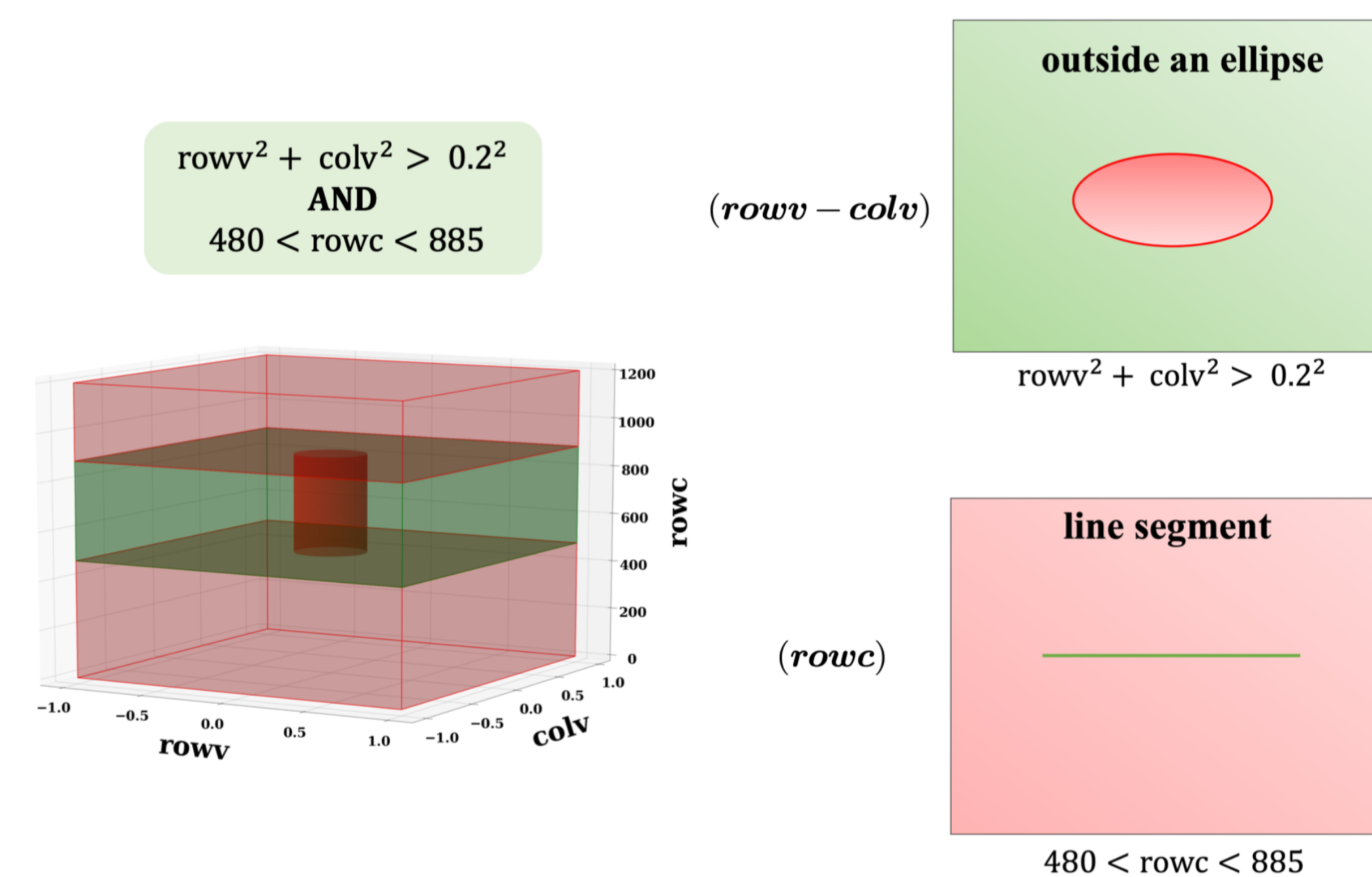
System Overview



- Consider the data content as a set of records, and the user is interested in some of them but not all.
- In each iteration,
 - the user labels a record as "interesting" or "not interesting",
 - a classification model is built,
 - active learning techniques are employed to select a new record from the unlabeled data source.
- Construct an increasingly-more-accurate model of the user interest.
- Upon convergence, the model is run through the entire data source to retrieve all relevant records.

Key Techniques

- Challenge: **Slow Convergence**
- Novel techniques in AIDEme:
 1. Factorization



– Factorized Version Space

Version Space V (size = 16)			Version Space V (size = 8)		
Color (B/R)	Size (S/L)	Label	Color (B/R)	Size (S/L)	Label
B(lack)	L(arge)	{-, +}	B(lack)	L(arge)	{-}
B(lack)	S(mall)	{-, +}	B(lack)	S(mall)	{-, +}
R(ed)	L(arge)	{-, +}	R(ed)	L(arge)	{-, +}
R(ed)	S(mall)	{-, +}	R(ed)	S(mall)	{-, +}

(a) Without factorization

Version Space V (size = 16)				Version Space V (size = 4)			
Color (B/R)	Label	Size (S/L)	Label	Color (B/R)	Label	Size (S/L)	Label
B(lack)	{-, +}	L(arge)	{-, +}	B(lack)	{-}	L(arge)	{+}
R(ed)	{-, +}	S(mall)	{-, +}	R(ed)	{-, +}	S(mall)	{-, +}

(b) With factorization

2. Formal results on convergence

- Theoretical results on the convergence of our proposed techniques.
- Detect convergence and terminate the exploration process.

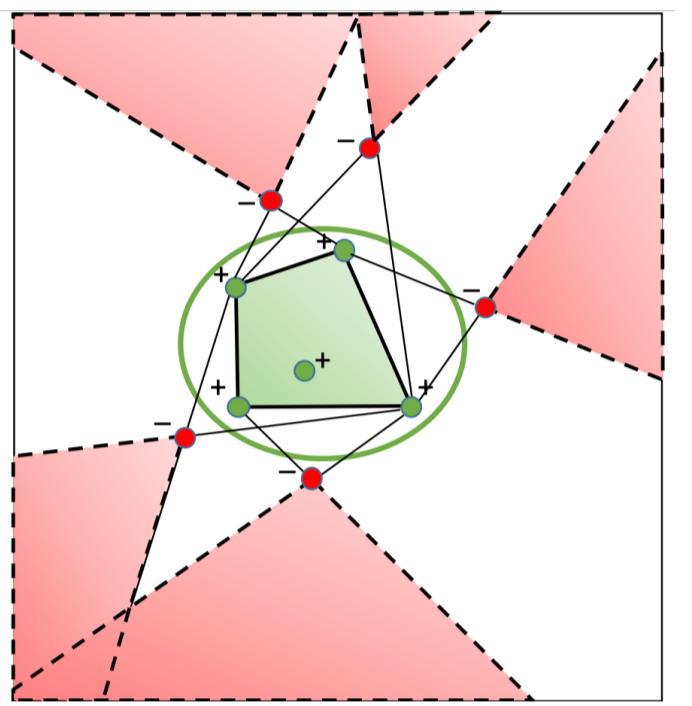
3. Scaling to large datasets

- Subsampling procedures
- Provide provable results that guarantee the performance of the model learned from the sample over the entire data source

4. Optimization using Class Distribution

- **Subspatial convex property:** the user interest pattern projected onto a subspace often entails a convex object.
- When the subspatial convex property holds, we introduce a Dual-Space Model (DSM).

$$\text{DSM} \xleftarrow[\text{sample next}]{\text{predict}} \text{Classifier} + \text{Polytope model.}$$

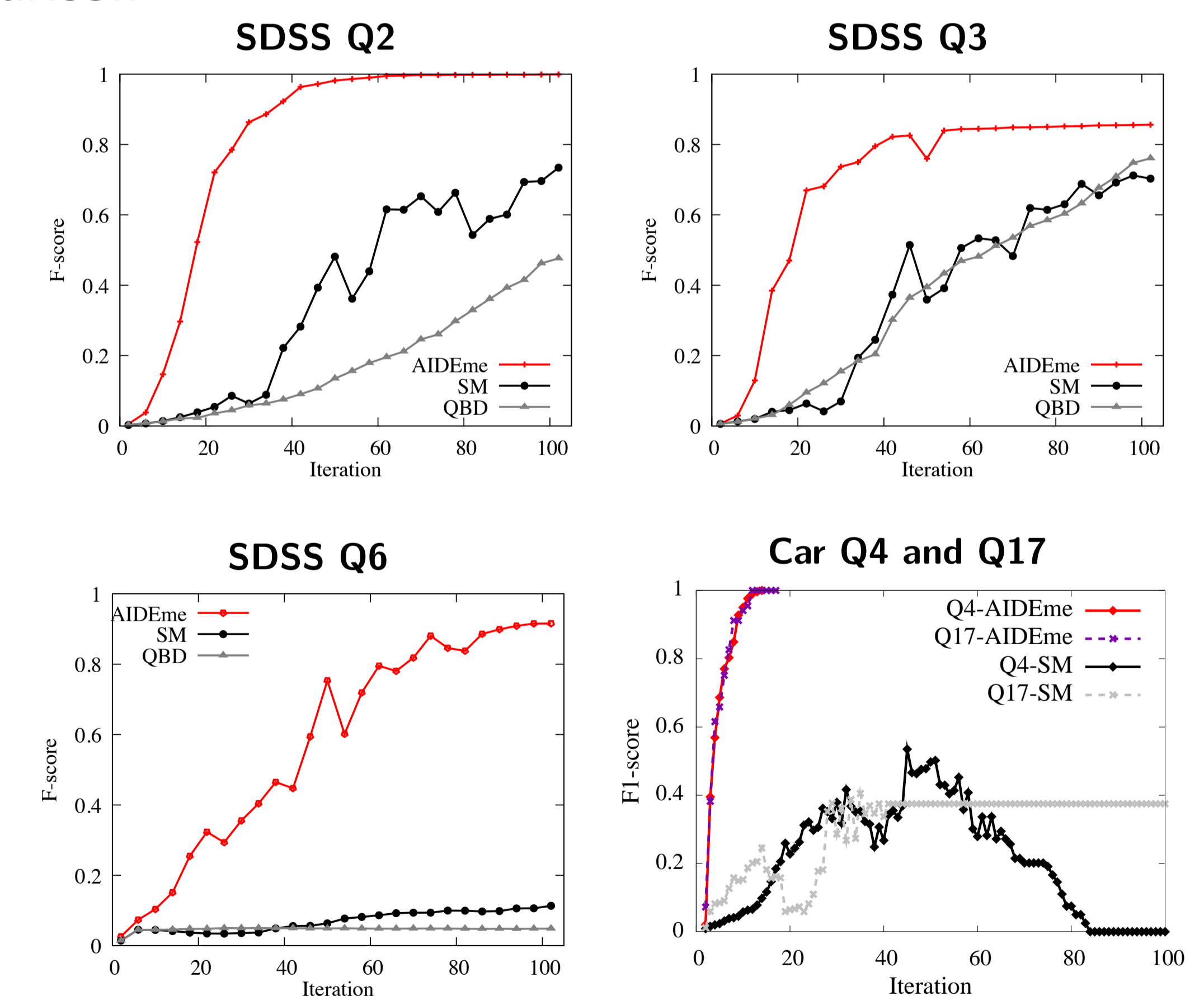


Demonstration

• Demonstration

Setup \Rightarrow Iterative Exploration \Rightarrow Final retrieval \Rightarrow Comparison

• Comparison



References

1. Huang, E., Peng, L., Di Palma, L., Abdelkafi, A., Liu, A. & Diao, Y. Optimization for active learning-based interactive database exploration. *Proceedings of the VLDB Endowment (PVLDB)*, 12(1), 71-84, September 2018.
2. Di Palma, L., Diao, Y. & Liu, A. A Factorized Version Space Algorithm for "Human-In-the-Loop" Data Exploration. *IEEE International Conference on Data Mining (ICDM)*, November 2019.