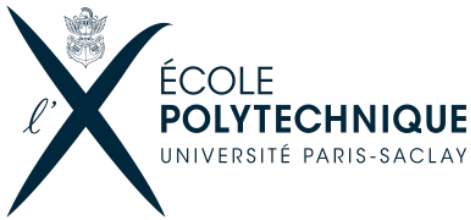
Optimization for Active Learning-based Interactive Database Exploration

VLDB 2019, Los Angeles, USA

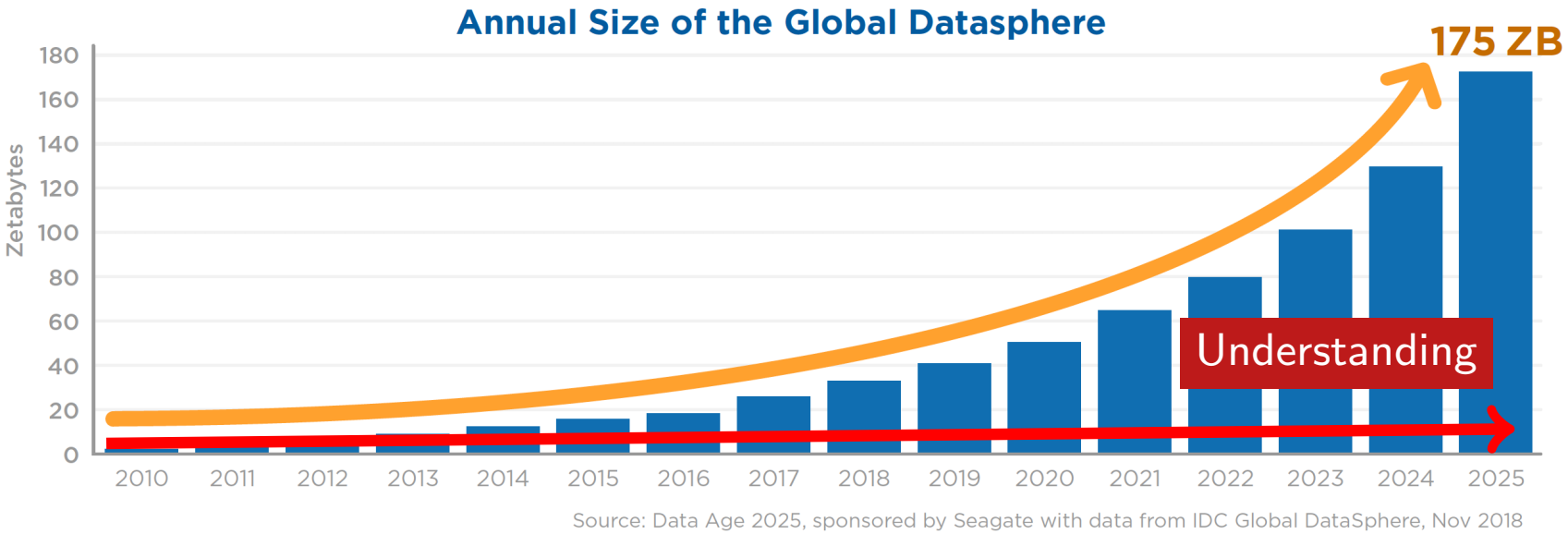
Enhui Huang[†], Liping Peng[*], Luciano Di Palma[†], Ahmed Abdelkafi[†], Anna Liu[*], Yanlei Diao[*,†]

[†]École Polytechnique, France
[*]University of Massachusetts Amherst, USA



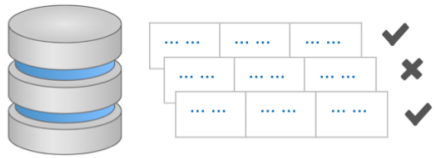
Motivation

- Data is growing extremely fast
- Human ability to comprehend data remains limited



Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



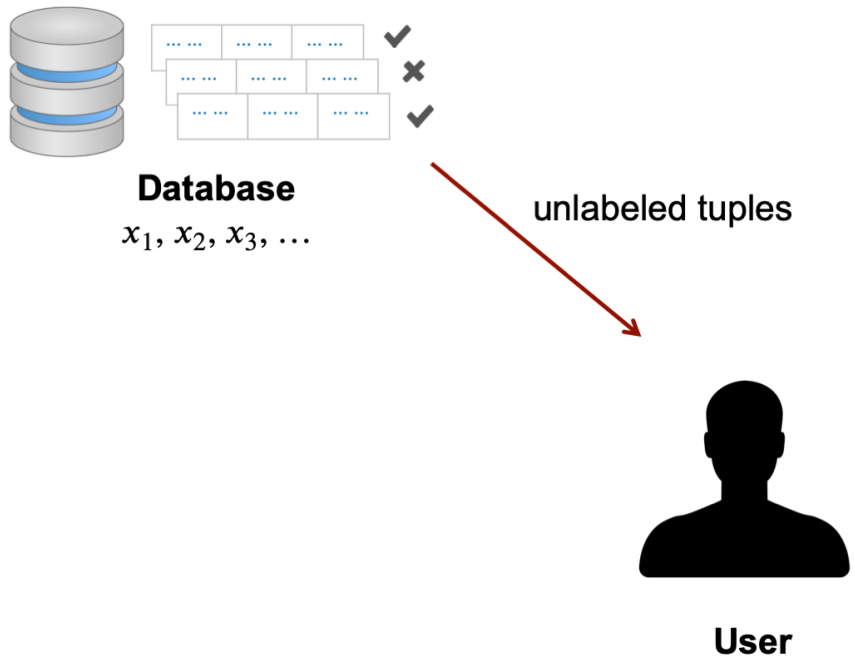
Database

$x_1, x_2, x_3, \ldots$

User

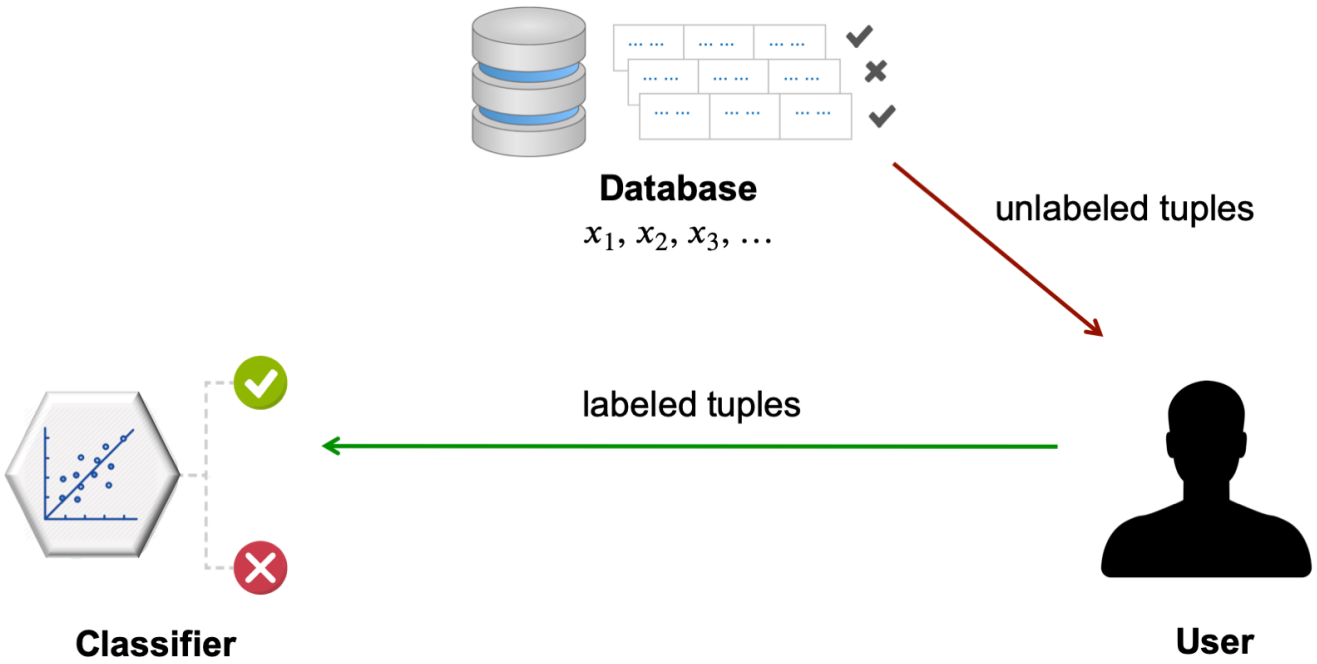
Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



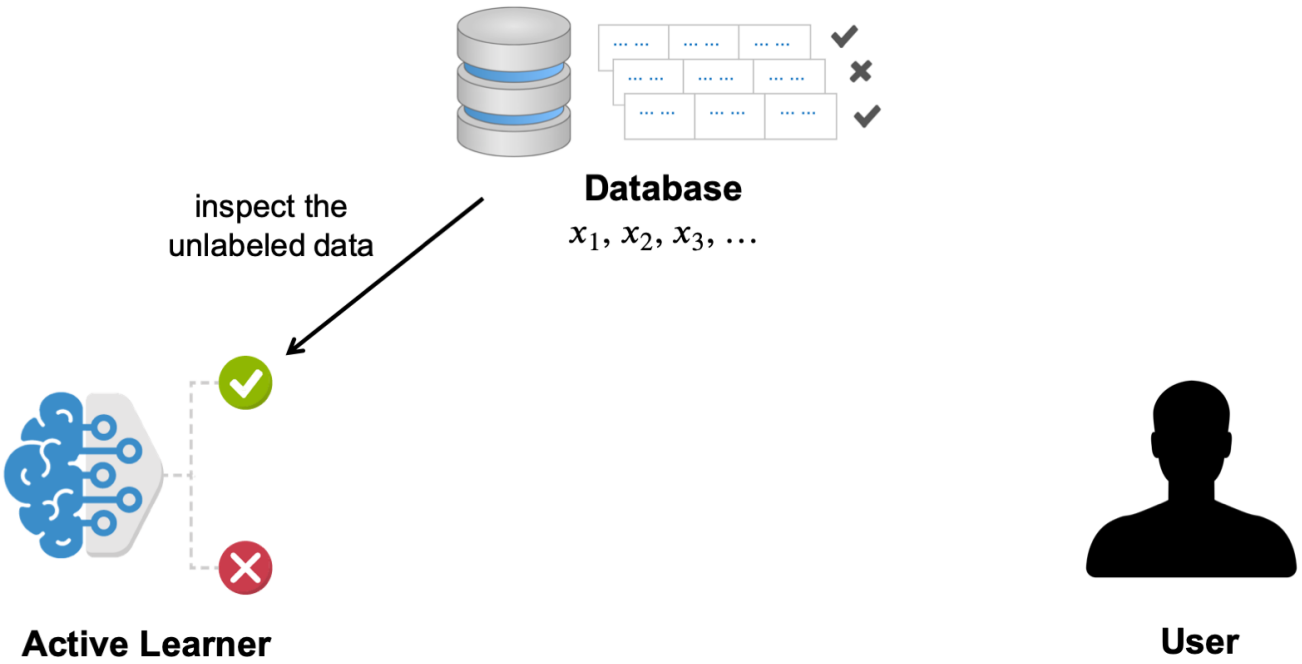
Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



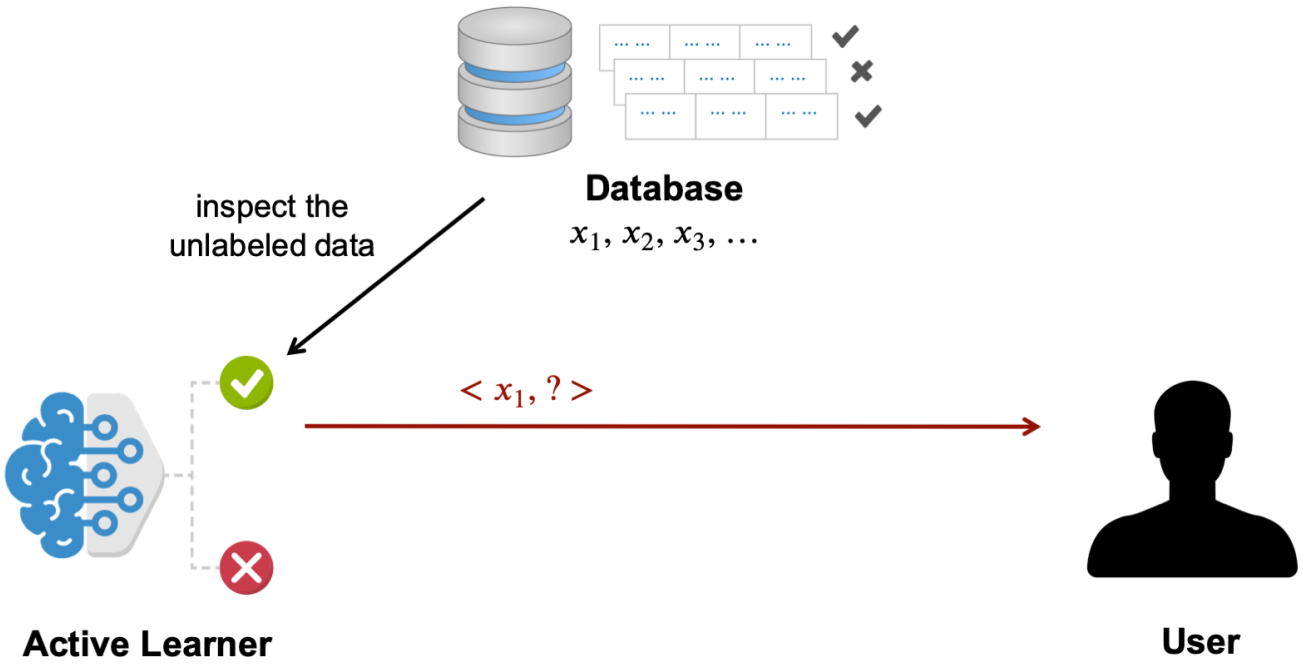
Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



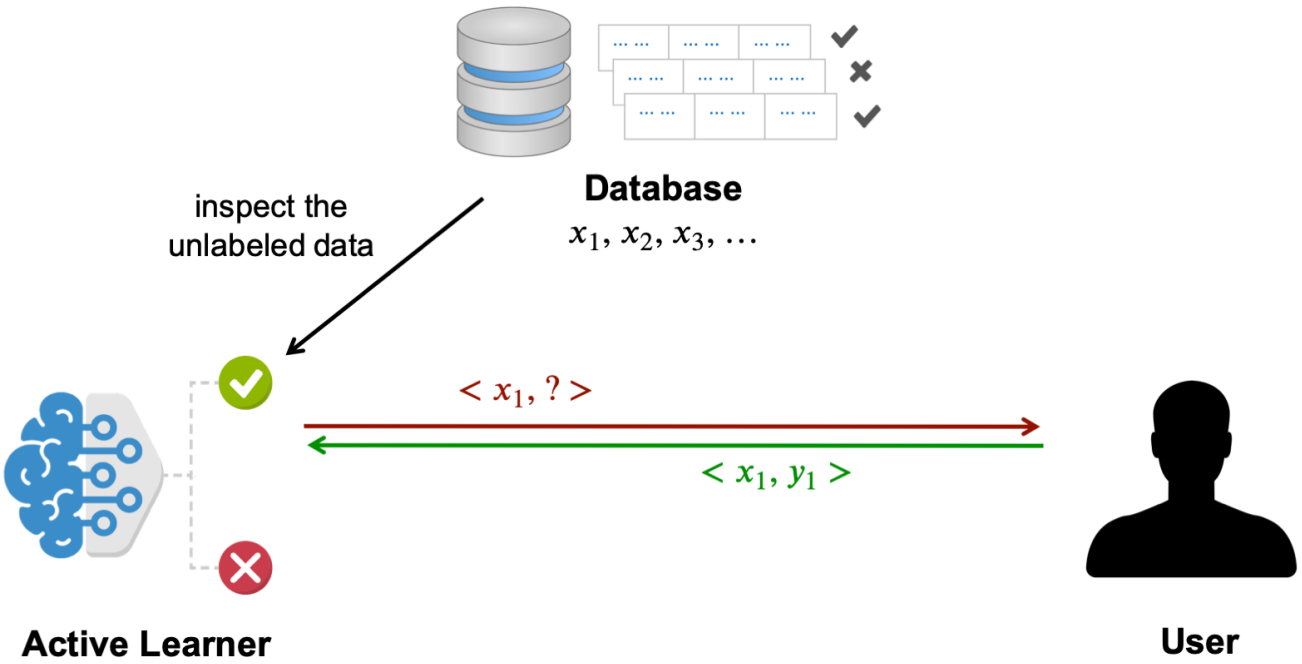
Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



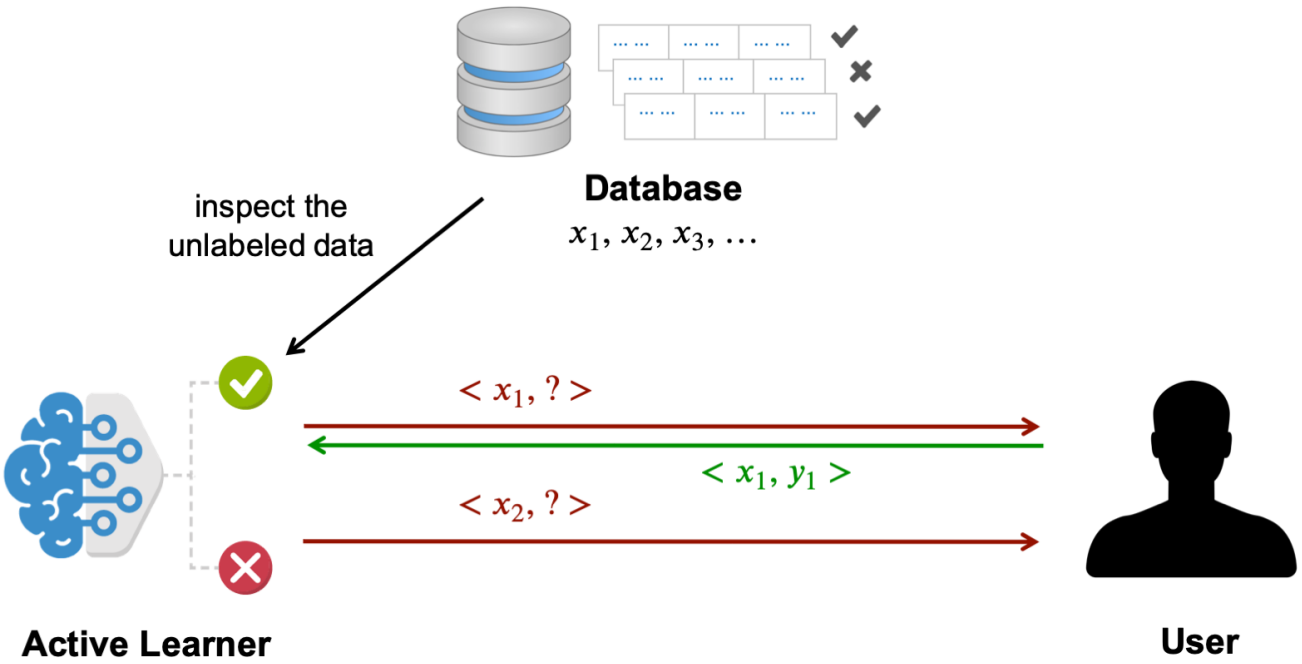
Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



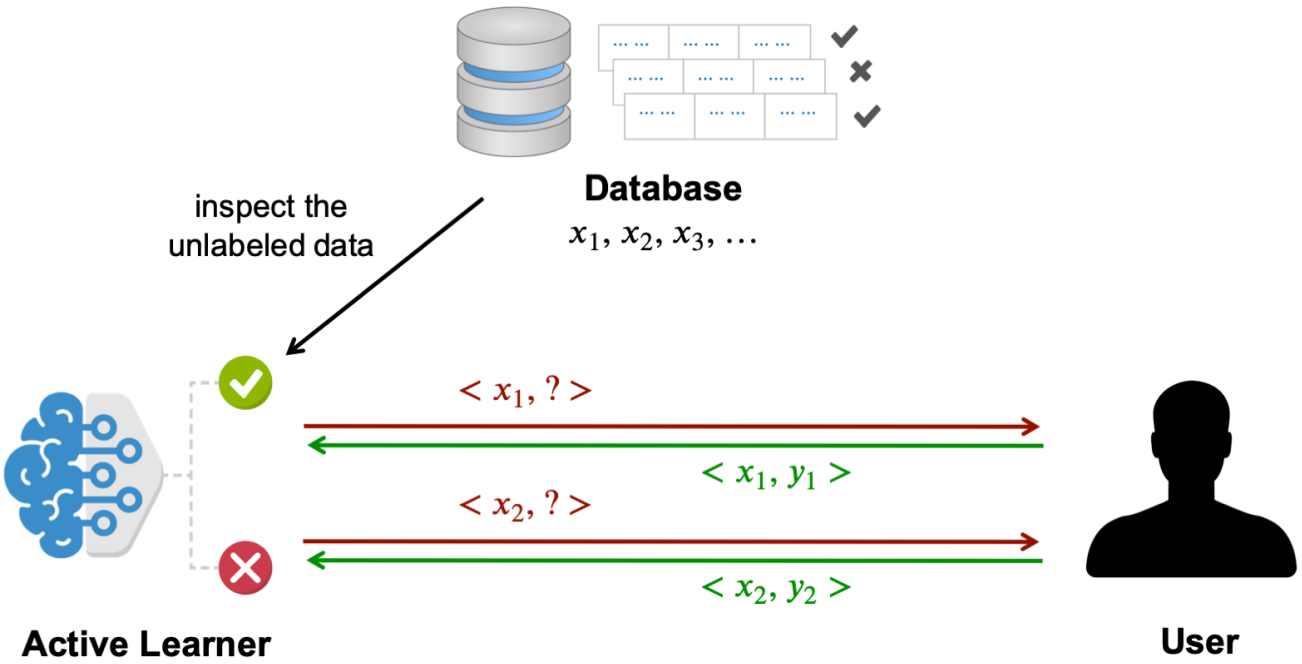
Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



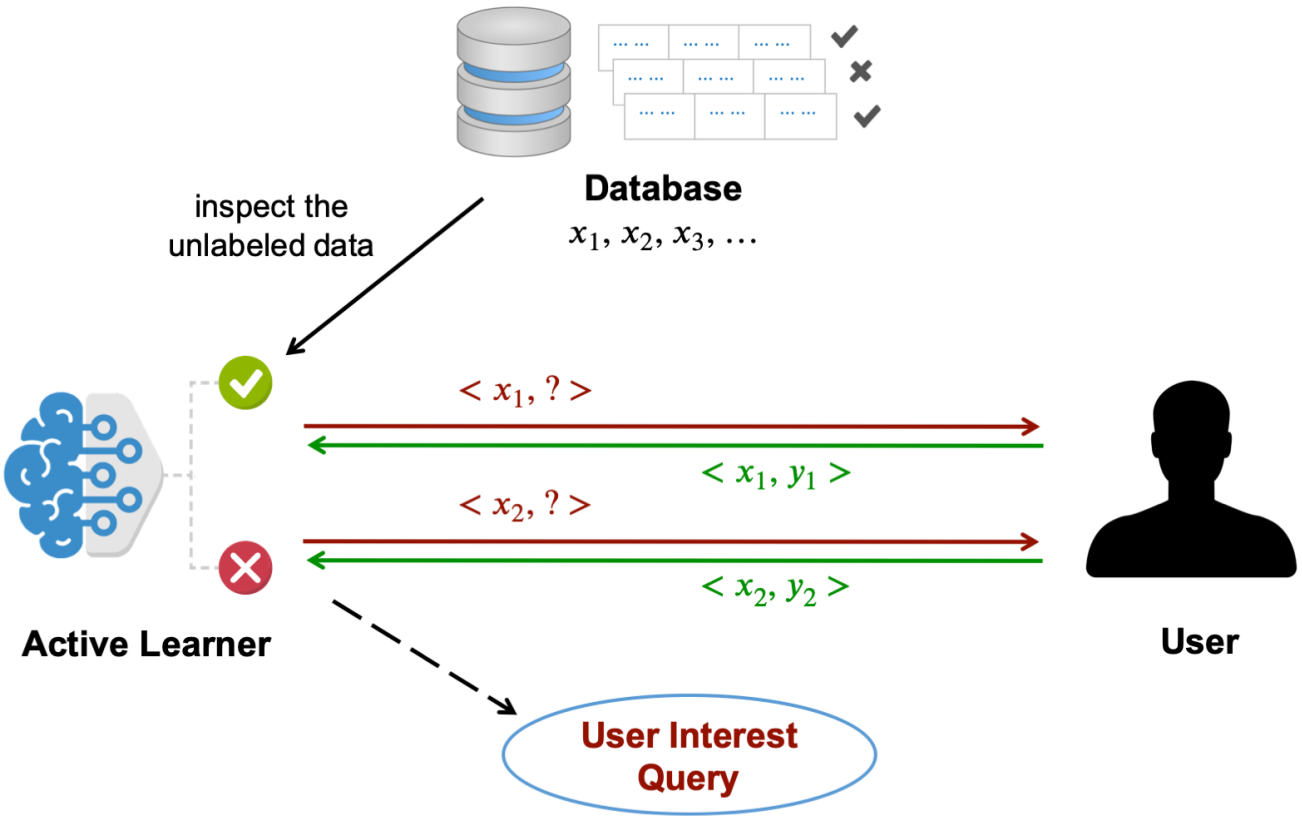
Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework



Active Learning-based Interactive Data Exploration (IDE)

– in an **"explore-by-example"** framework

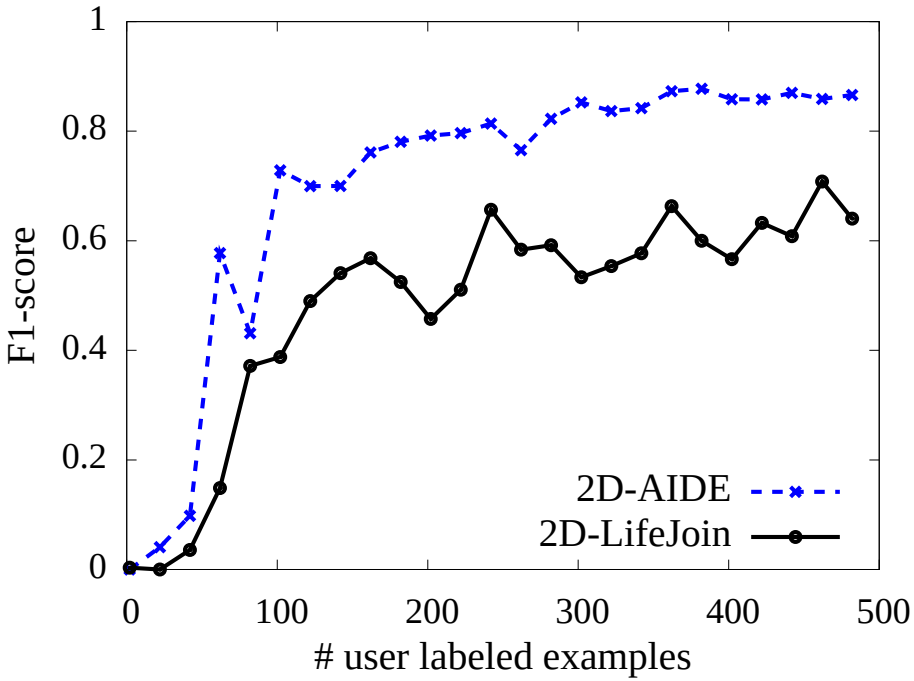


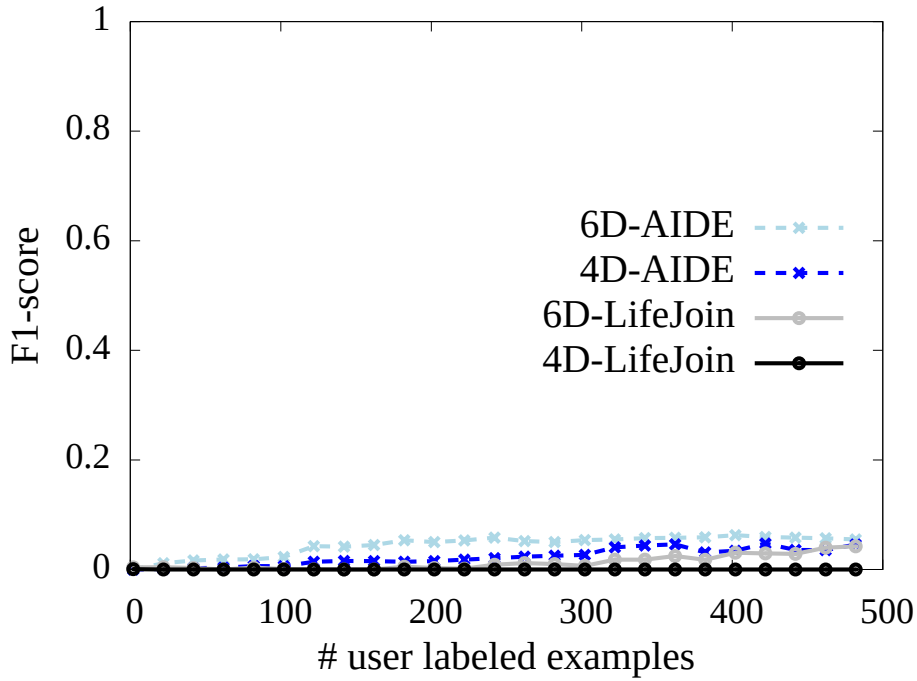
Active learning-based framework requires fewer labeled examples than a traditional learner

Slow convergence on large databases

Slow convergence problem of LifeJoin[1] and AIDE[2]

- a large number of labeled examples needed to achieve high accuracy
- exacerbated when the user interest query has **low selectivity** or **high dimensionality**





[1] A. Cheung, A. Solar-Lezama, and S. Madden. Using program synthesis for social recommendations. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM'12, pages 1732–1736, New York, NY, USA, 2012. ACM

[2] K. Dimitriadou, O. Papaemmanouil, and Y. Diao. Aide: an active learning-based approach for interactive data exploration. IEEE Transactions on Knowledge and Data Engineering, 28(11):2842-2856, 2016.

Our explore-by-example system

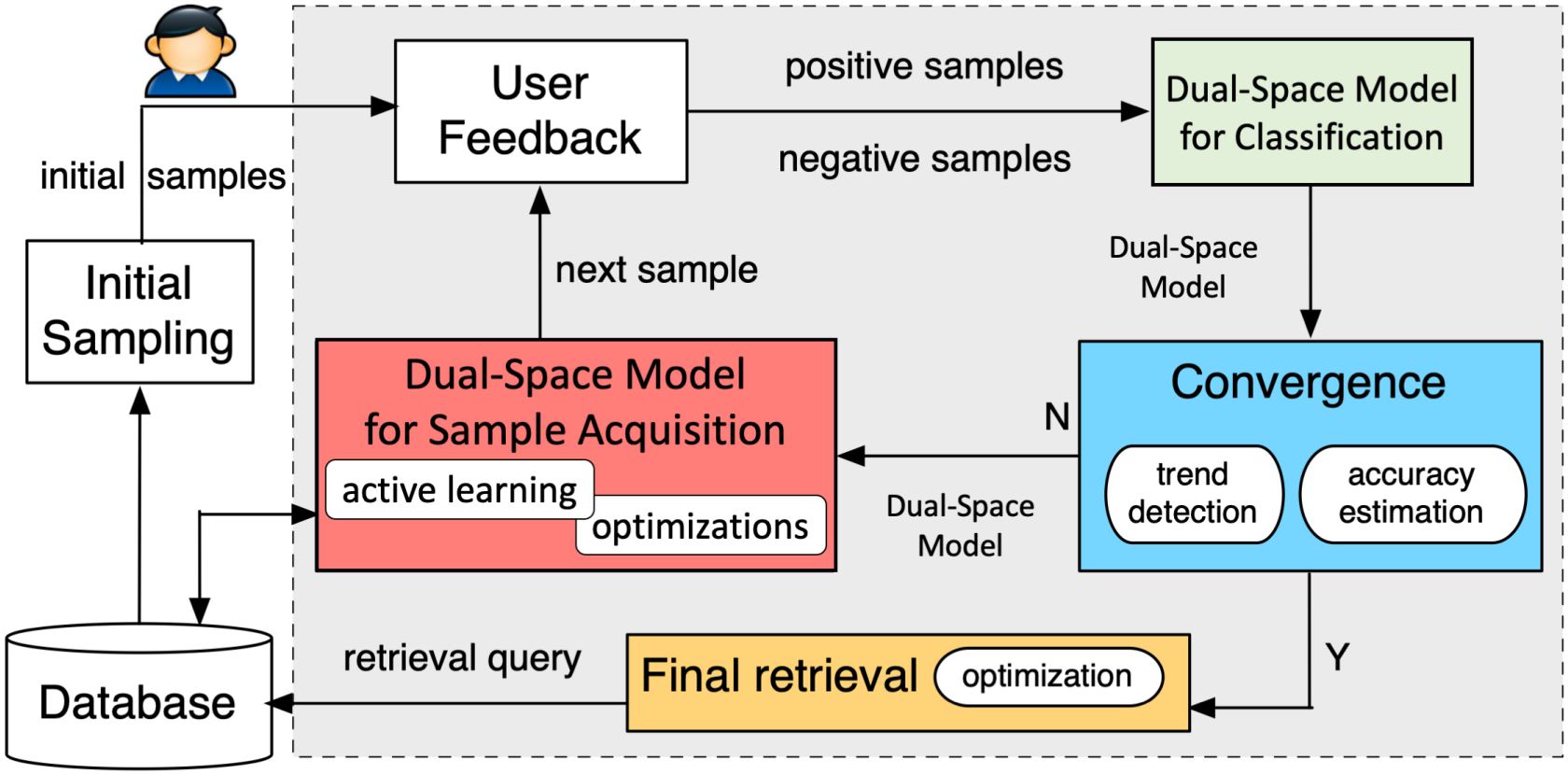
Objective

Leverage **query and data properties** from the database to overcome or alleviate the **slow convergence** problem when data exploration is performed with **high dimensionality** and **low selectivity** of the user interest query.

Main contributions

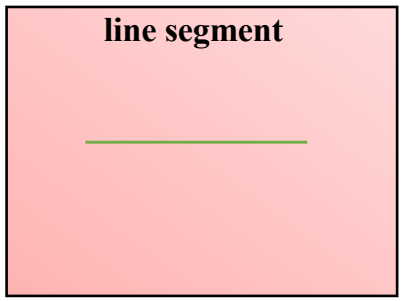
1. **Dual-Space Model** $\Longleftarrow$ **Subspatial Convexity**
2. **Factorization** for High-dimensional Exploration $\Longleftarrow$ **Conjunctivity**
3. Evaluation

Our explore-by-example system

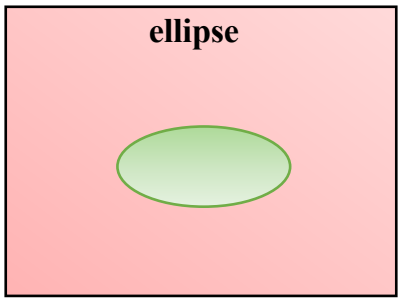


Subspatial Convexity

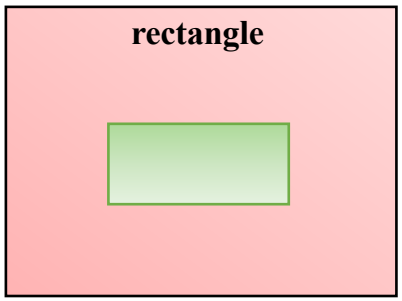
- In some lower-dimensional subspaces, the projected **user interest region** or its complement is convex.



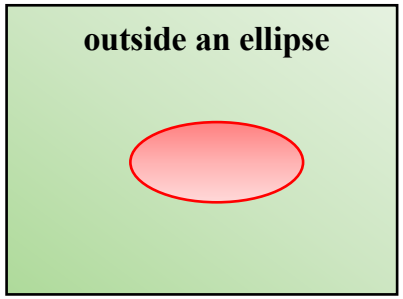
(a) rowc > 480 and rowc < 885



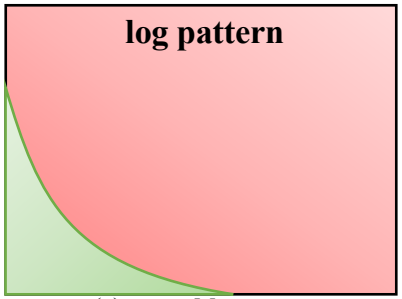
(b) $(\text{rowc} - 682.5)^2 + (\text{colc} - 1022.5)^2 < 90^2$



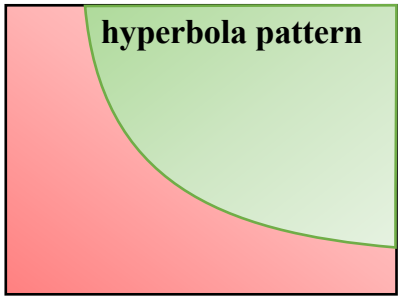
(c) rowc > 617.5 and rowc < 747.5 and colc > 925 and colc < 1120



(d) $\text{rowv}^2 + \text{colv}^2 > 0.2^2$

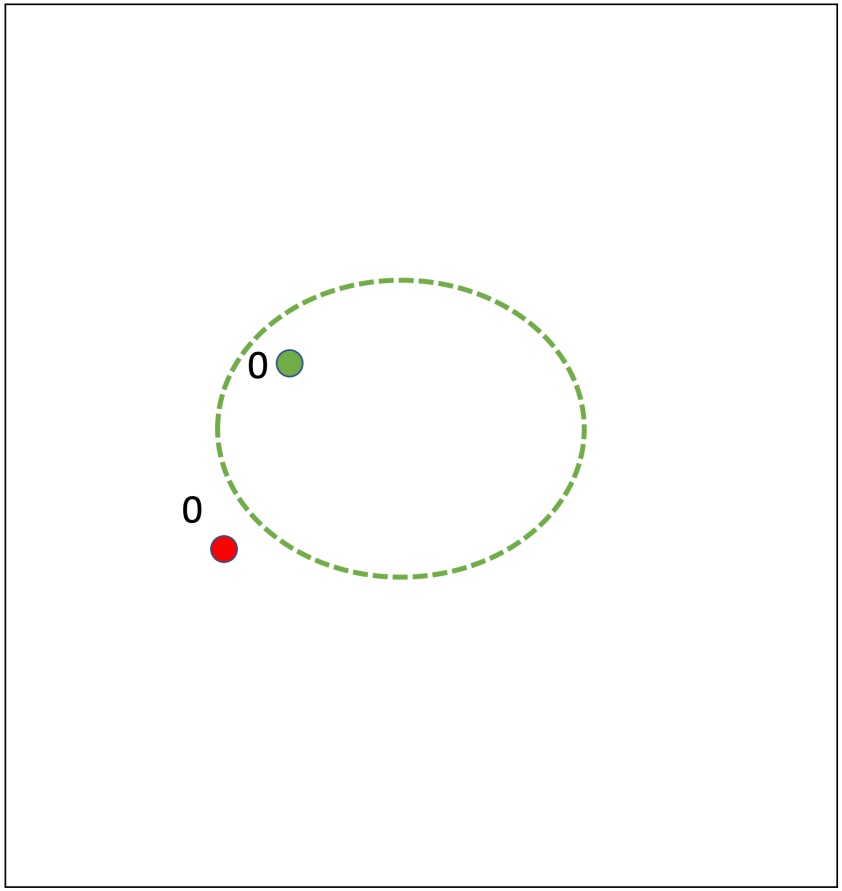


(e) $\text{petroMag\_ext\_r} + 2.5 * \log_{10}(\text{petroR50\_r2}) < 23.3$



(f) length * width >= 10.1

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

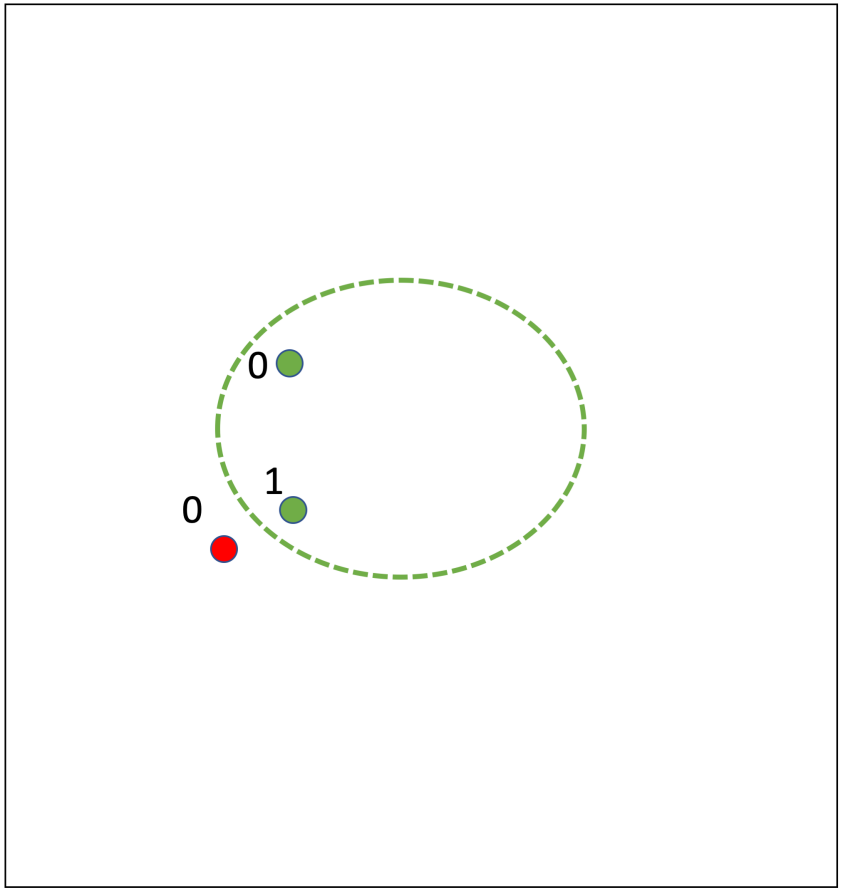
Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

Uncertain region (R^u)

$R^u = \mathbb{R}^d - R^+ - R^-$

$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

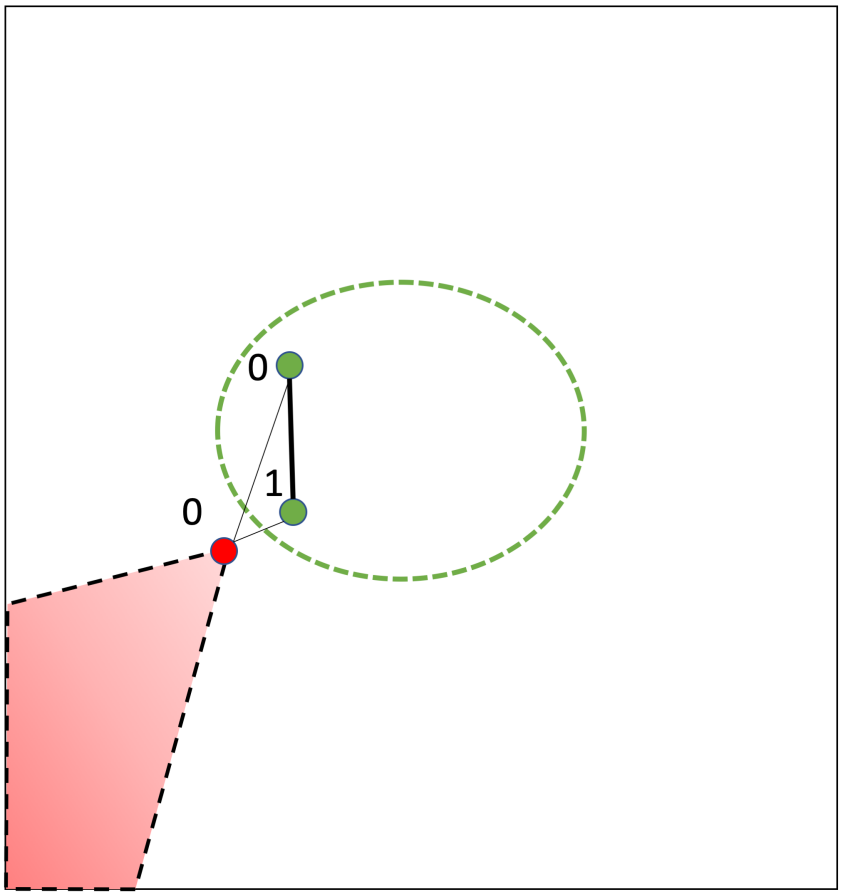
Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

Uncertain region (R^u)

$R^u = \mathbb{R}^d - R^+ - R^-$

$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

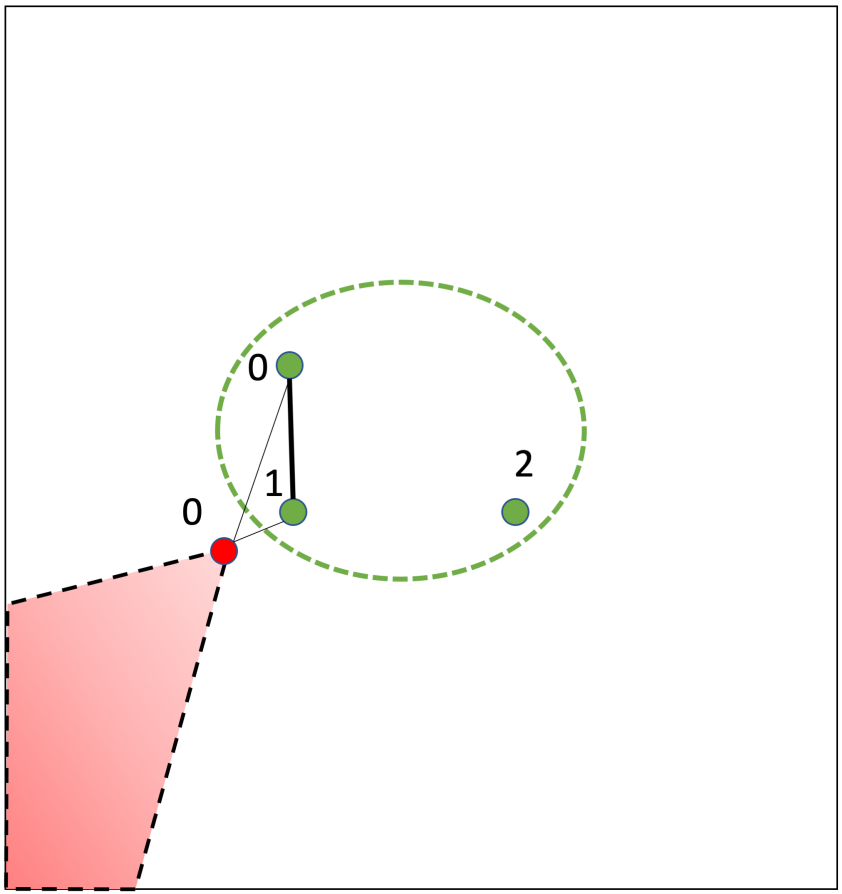
Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

Uncertain region (R^u)

$R^u = \mathbb{R}^d - R^+ - R^-$

$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

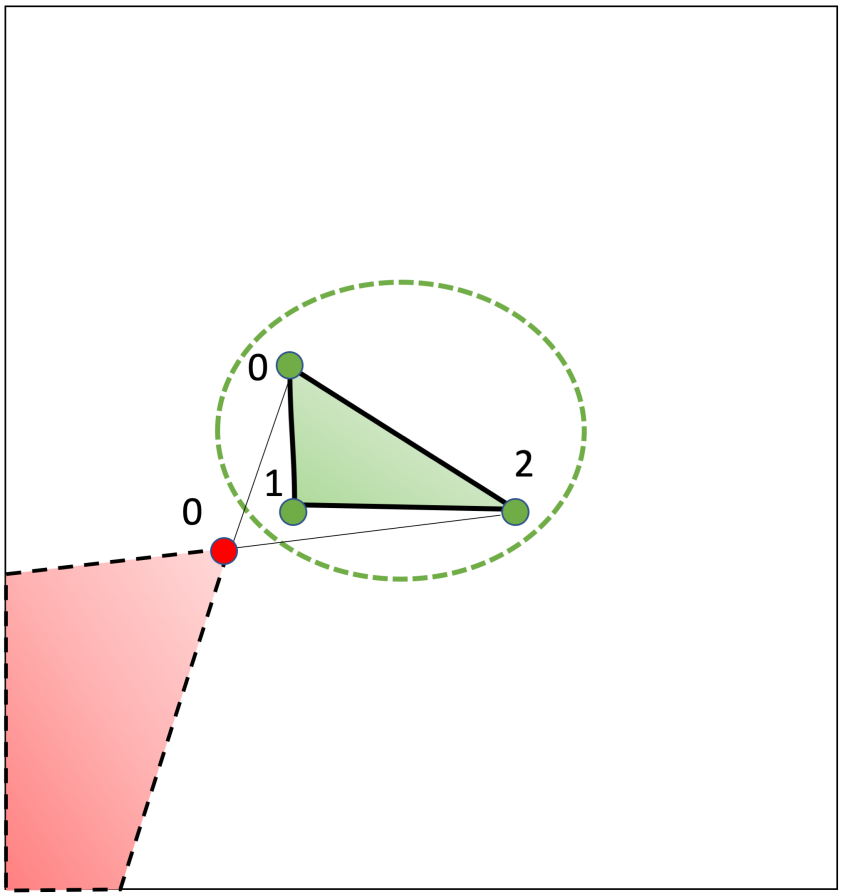
Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

Uncertain region (R^u)

$R^u = \mathbb{R}^d - R^+ - R^-$

$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

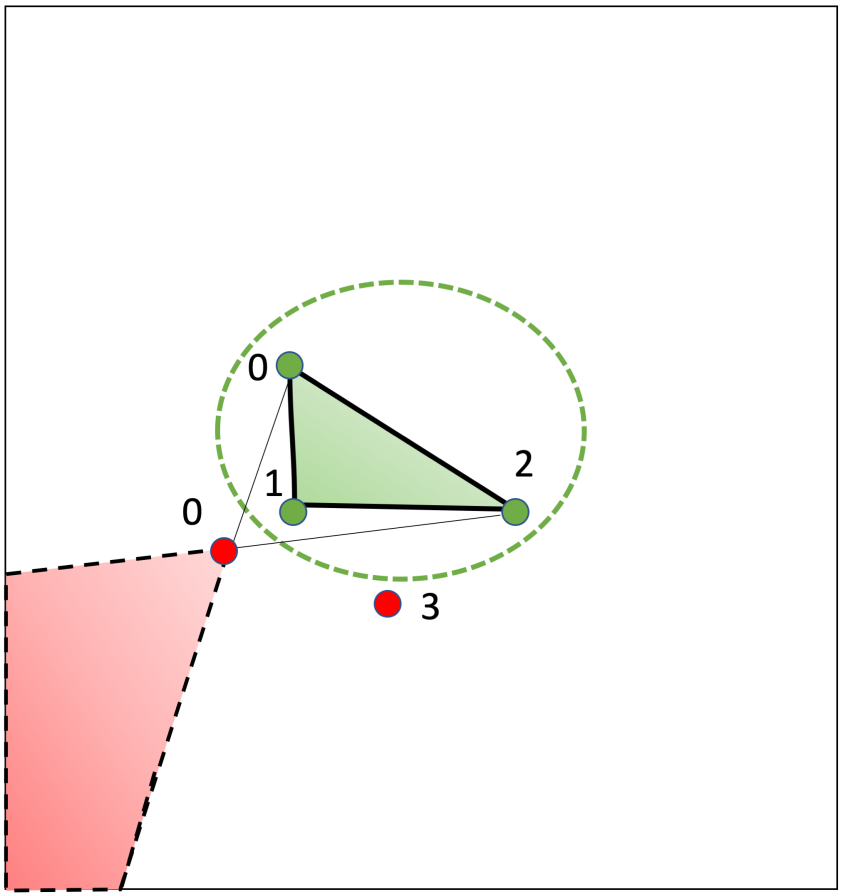
Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

Uncertain region (R^u)

$R^u = \mathbb{R}^d - R^+ - R^-$

$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

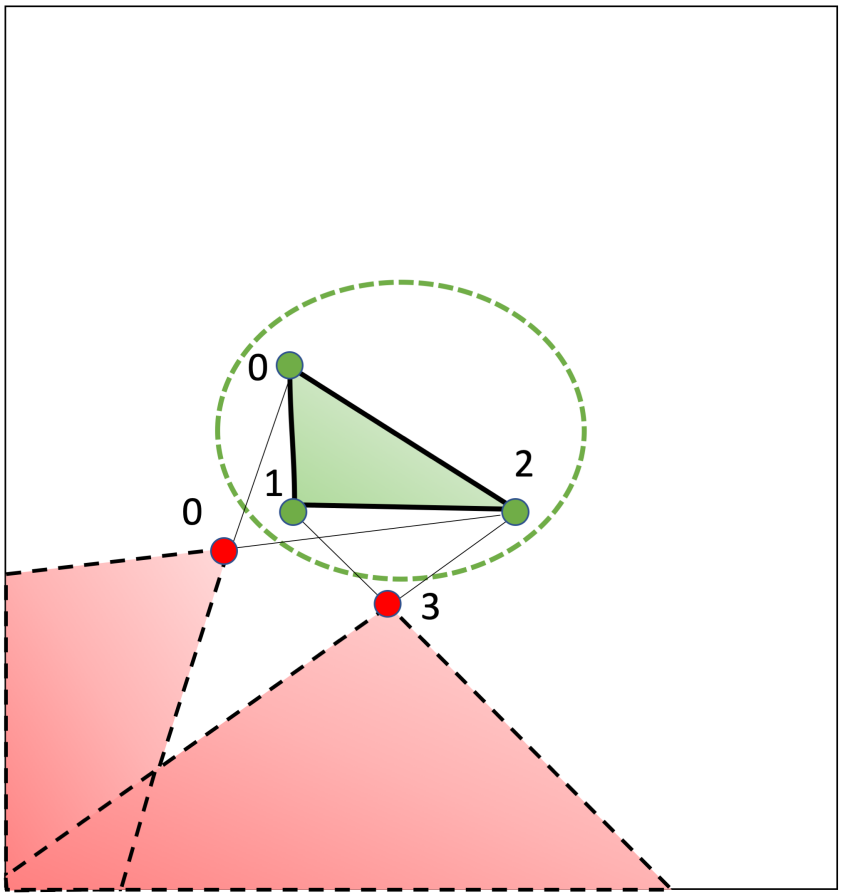
Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

Uncertain region (R^u)

$R^u = \mathbb{R}^d - R^+ - R^-$

$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

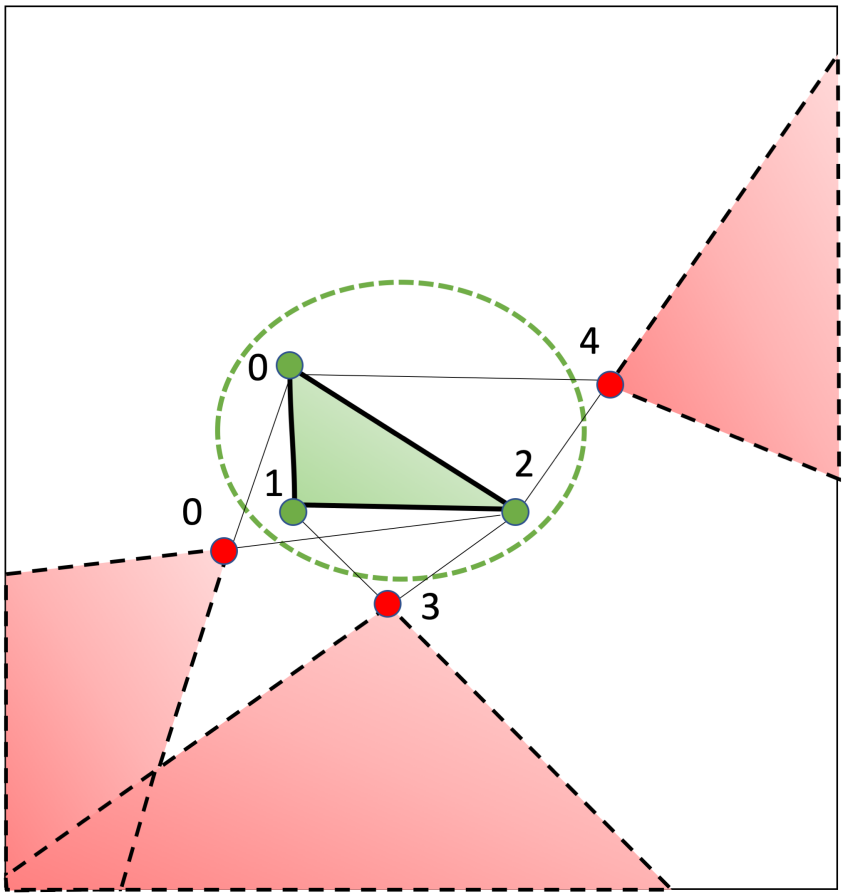
Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

Uncertain region (R^u)

$R^u = \mathbb{R}^d - R^+ - R^-$

$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Data Space Model ($F_{\mathcal{D}}$)



Positive Region (R^+)

The convex hull of positive examples

Negative Region (R^-)

Given R^+ and a negative example $\boldsymbol{e}_i^-$, a convex cone can be built by $R_i^- = \{\boldsymbol{x} | \overline{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ = \emptyset \wedge \overrightarrow{\boldsymbol{x}\boldsymbol{e}_i^-} \cap R^+ \neq \emptyset\}$. Given n^- negative examples, the negative region is $R^- = \cup_{i=1}^{n^-} R_i^-$.

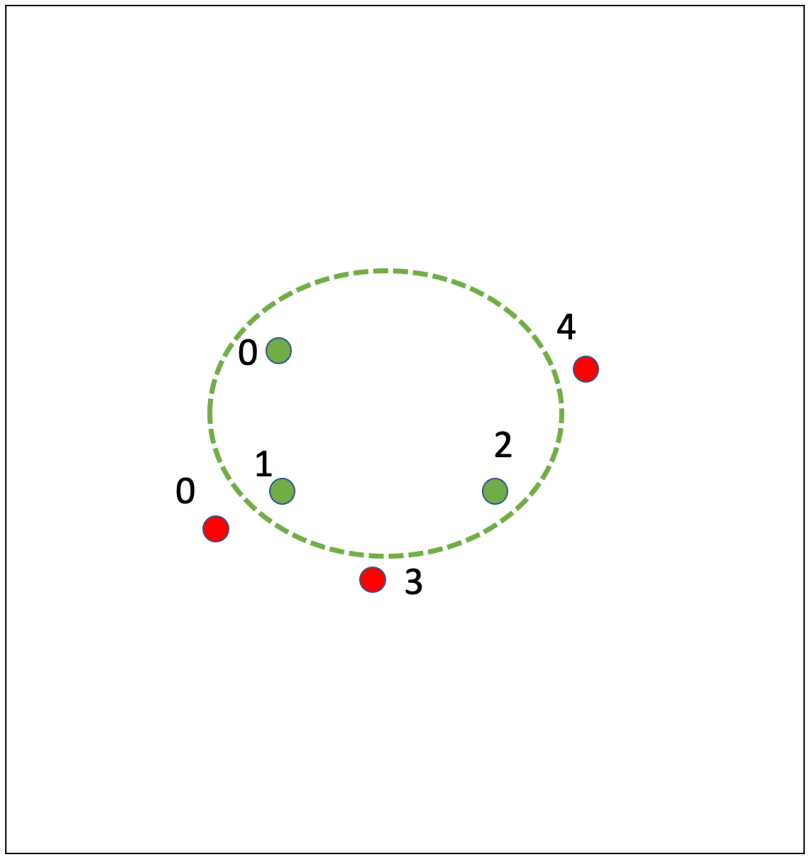
Uncertain region (R^u)

$$R^u = \mathbb{R}^d - R^+ - R^-$$

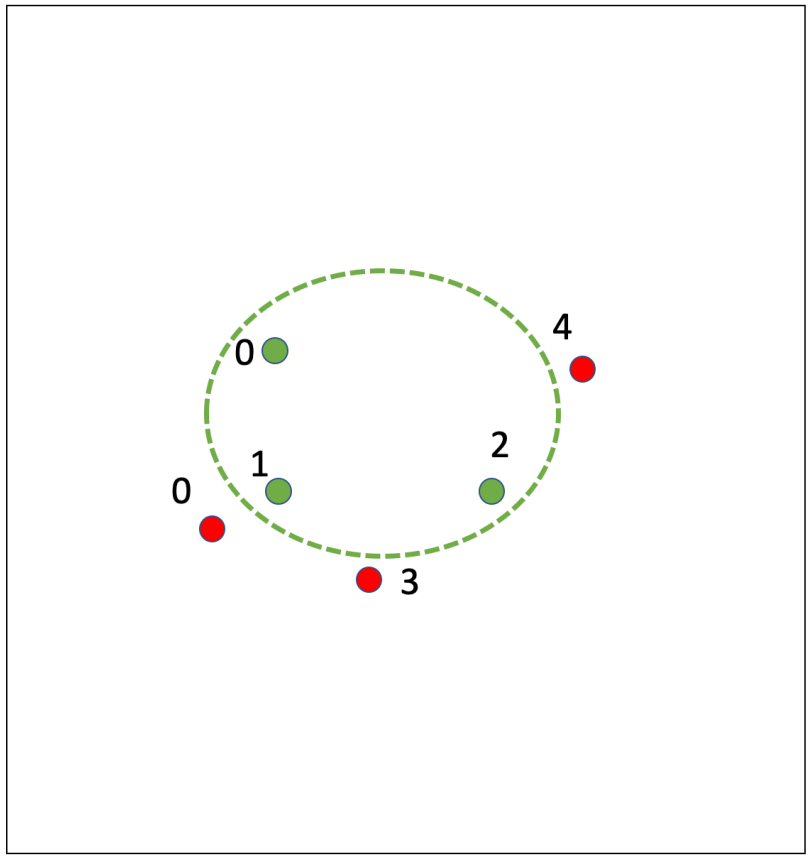
$$F_{\mathcal{D}}(\boldsymbol{x}) = 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^+) - 1 \cdot \mathbb{1}(\boldsymbol{x} \in R^-)$$

Dual-Space Model (DSM) for Classification

Data Space Model ($F_{\mathcal{D}}$)

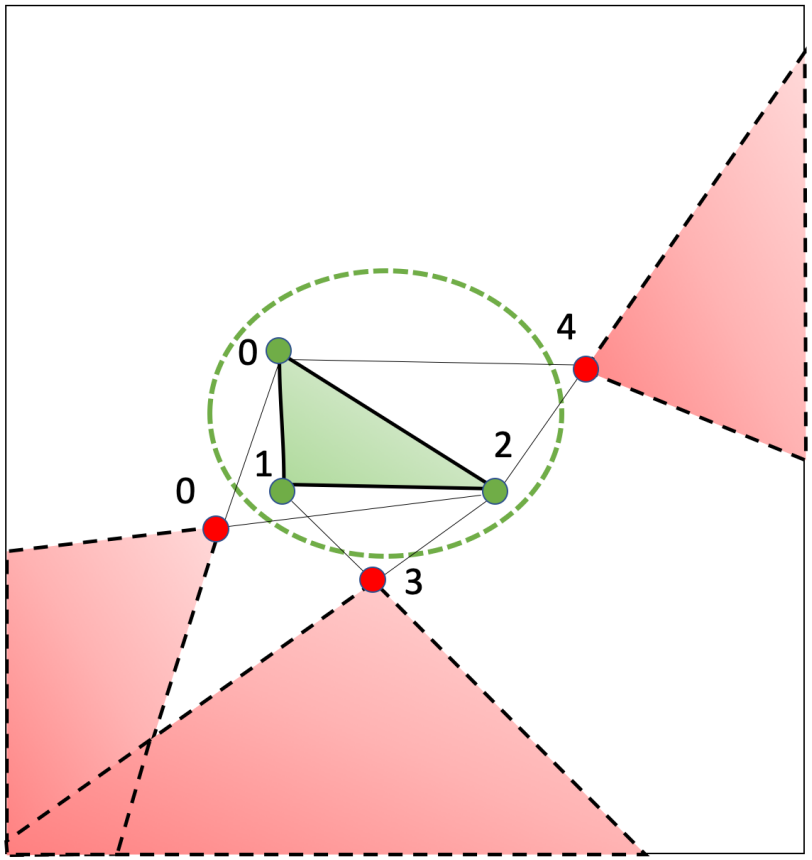


Classification Model ($F_{\mathcal{V}}$)

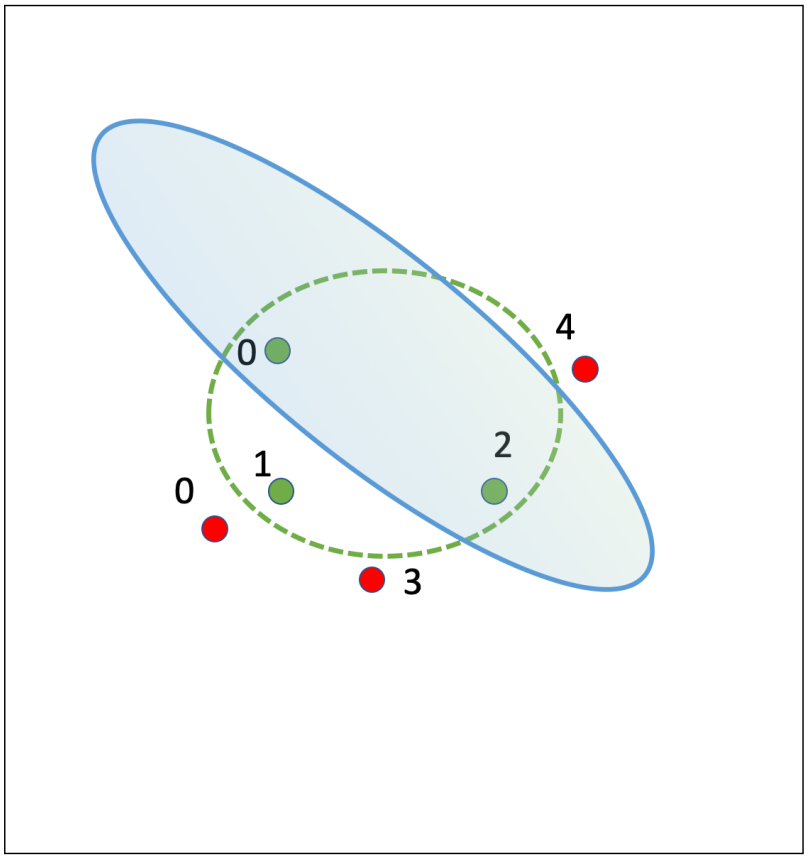


Dual-Space Model (DSM) for Classification

Data Space Model ($F_{\mathcal{D}}$)

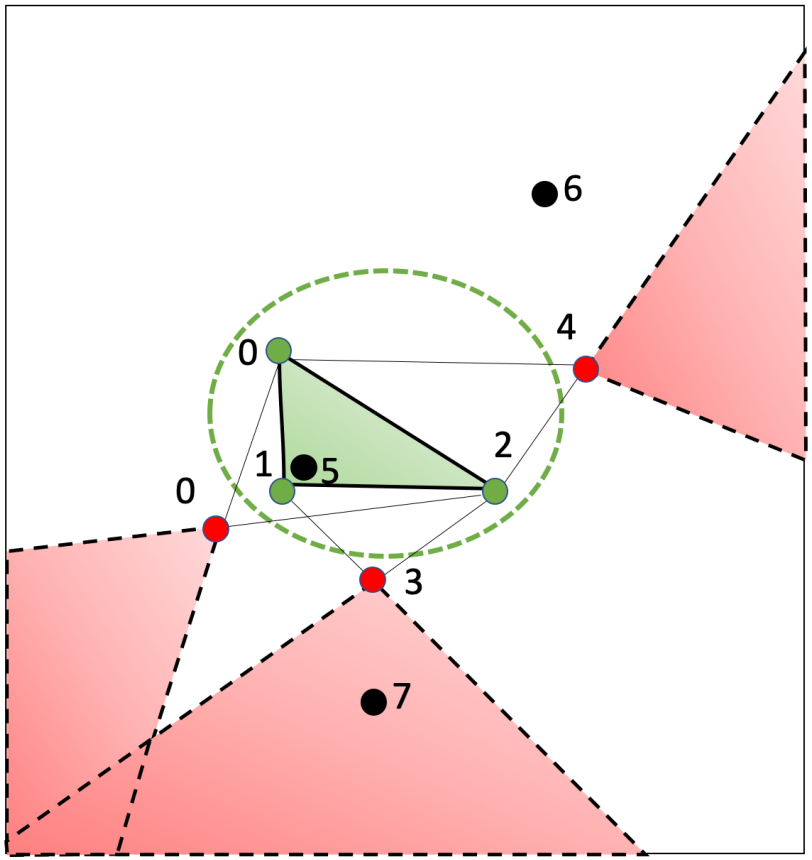


Classification Model ($F_{\mathcal{V}}$)

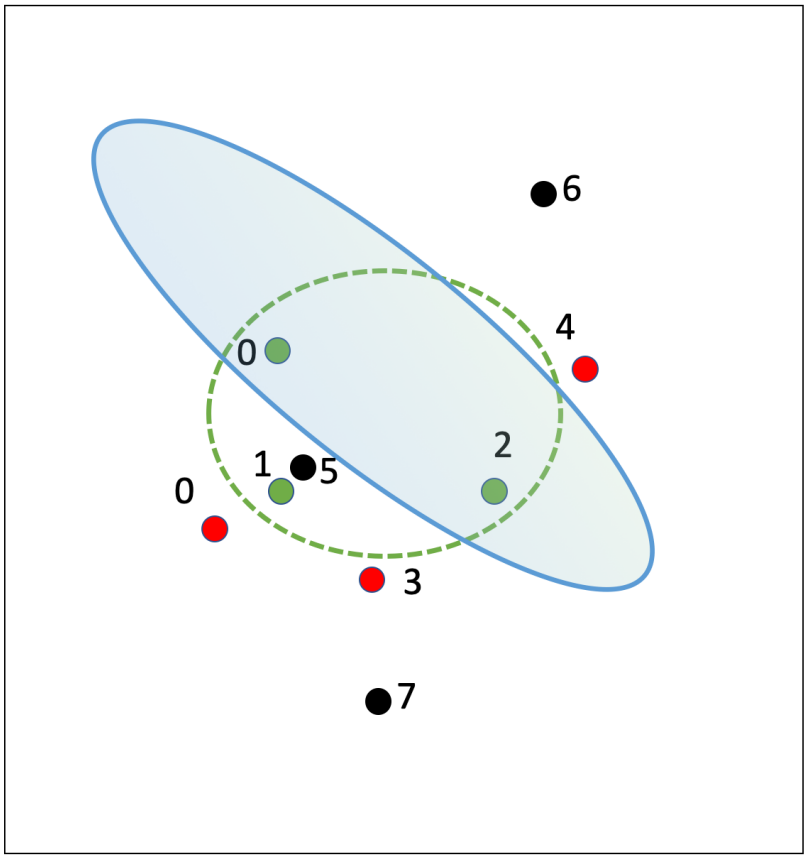


Dual-Space Model (DSM) for Classification

Data Space Model ($F_{\mathcal{D}}$)

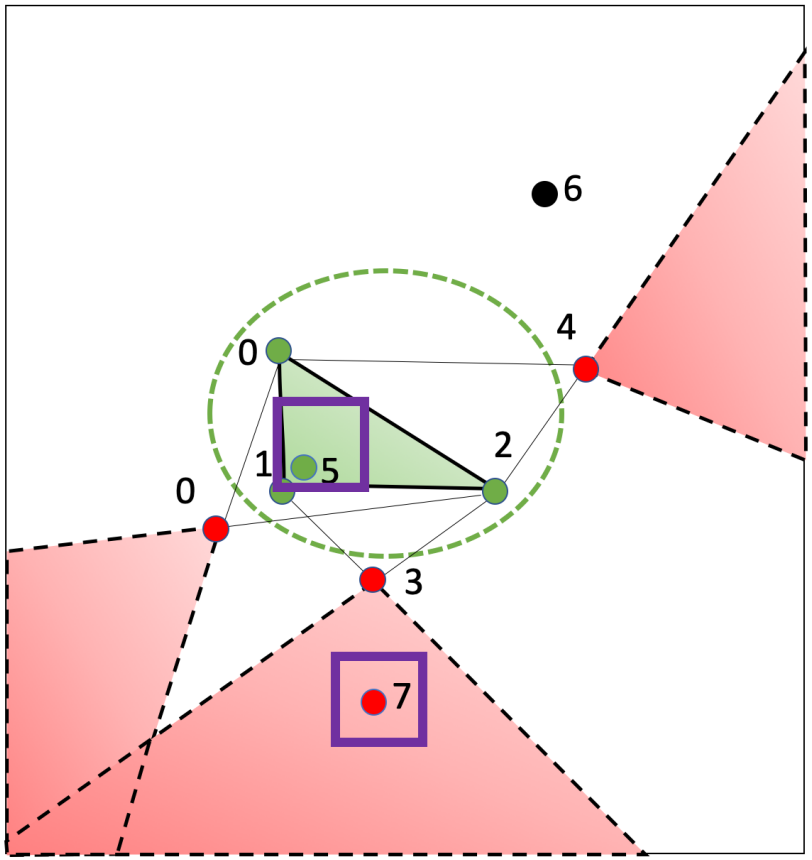


Classification Model ($F_{\mathcal{V}}$)

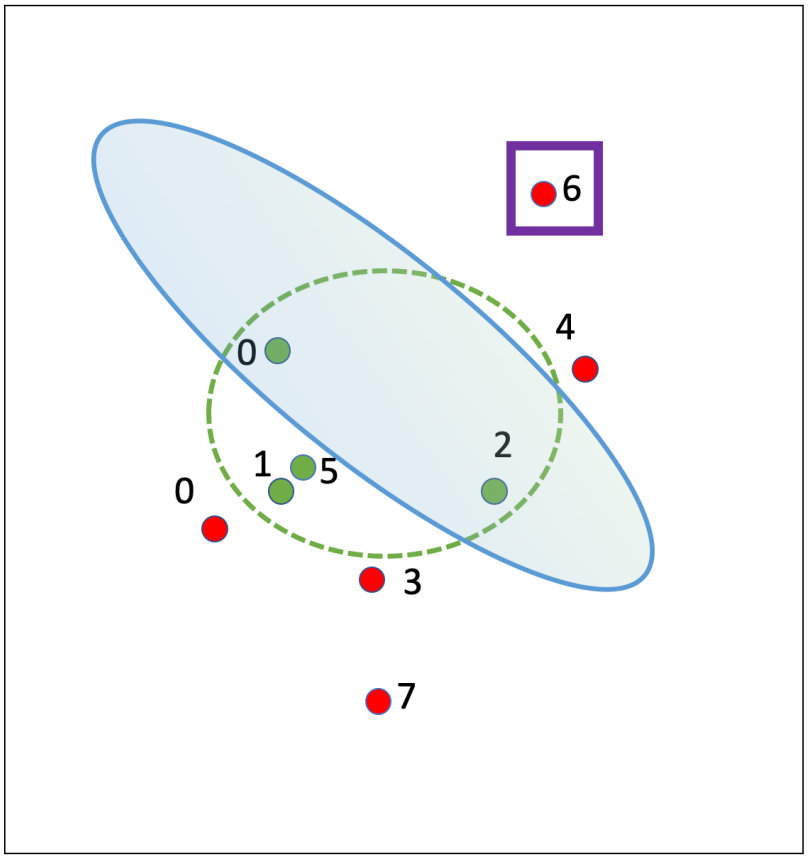


Dual-Space Model (DSM) for Classification

Data Space Model ($F_{\mathcal{D}}$)

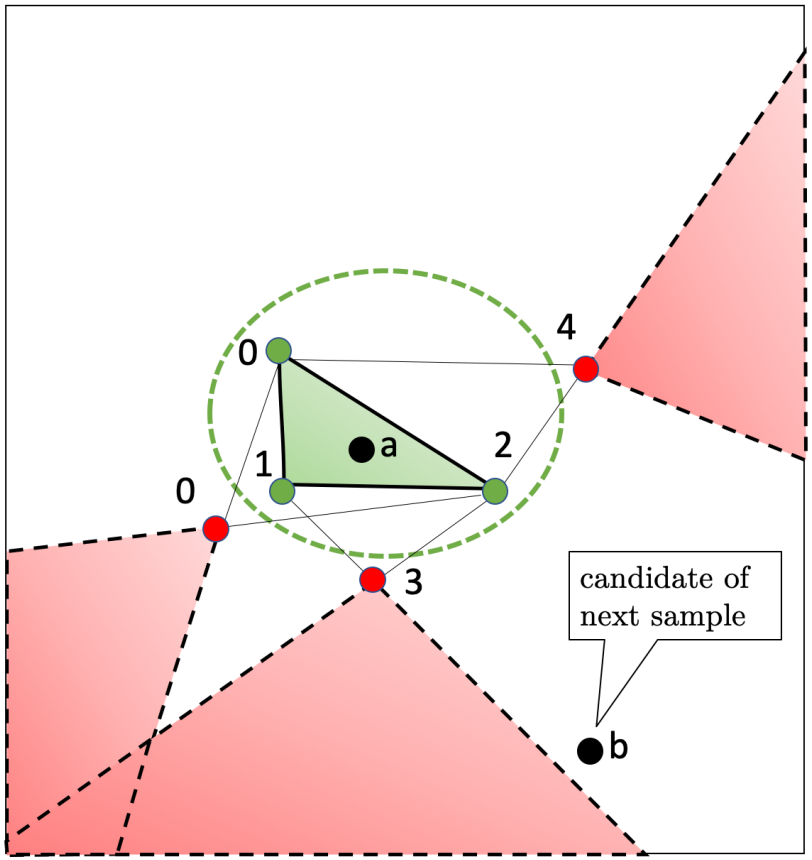


Classification Model ($F_{\mathcal{V}}$)

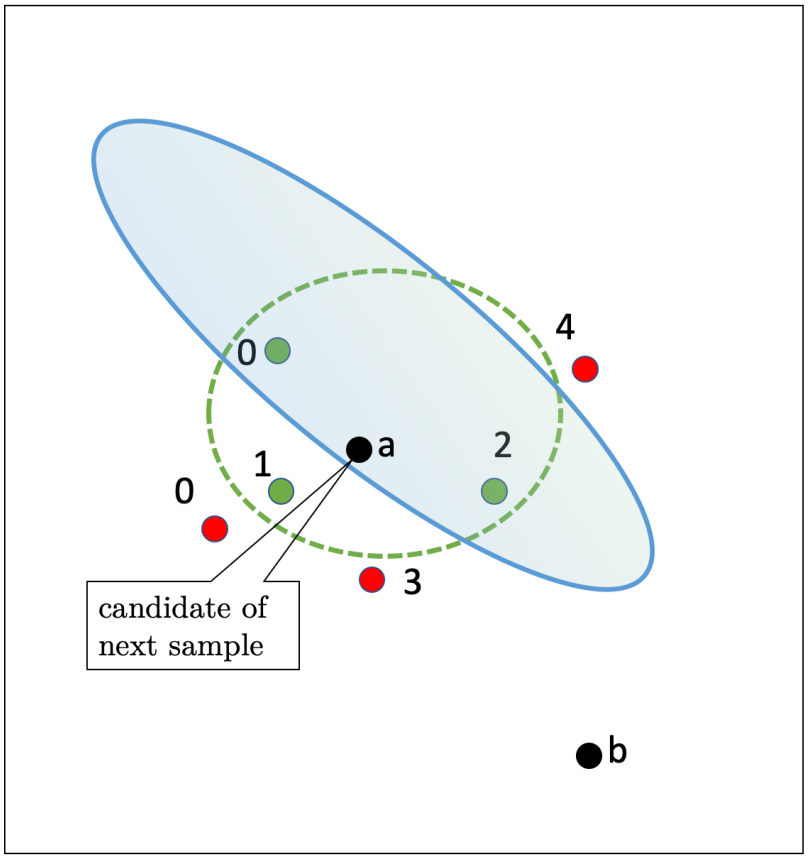


Dual-Space Model (DSM) for Sample Acquisition

Data Space Model ($F_{\mathcal{D}}$)



Classification Model ($F_{\mathcal{V}}$)



Convergence – Lower Bounds of $\mathbf{F_1}$ score

Evaluation Metric ($\mathbf{F_1}$ score)

$$\mathbf{F_1\ score} = \frac{2}{\mathbf{Precision}^{-1} + \mathbf{Recall}^{-1}}, \text{ where}$$

$$\mathbf{Precision} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{retrieved points}\}|}, \mathbf{Recall} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{positive points}\}|}$$

Convergence – Lower Bounds of $\mathbf{F_1}$ score

Evaluation Metric ($\mathbf{F_1}$ score)

$\mathbf{F_1\ score} = \frac{2}{\mathbf{Precision}^{-1} + \mathbf{Recall}^{-1}}$, where

$\mathbf{Precision} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{retrieved points}\}|}$, $\mathbf{Recall} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{positive points}\}|}$

Three-Set Metric

Let D_{eval} be the projection of D_{test} without labels. Denote $D^+ = D_{eval} \cap R^+$, $D^- = D_{eval} \cap R^-$, $D^u = D_{eval} \cap R^u$, and $|S|$ means the size of set S. At a specific iteration of exploration, the three-set metric is defined to be $\frac{|D^+|}{|D^+| + |D^u|}$.

Convergence – Lower Bounds of $\mathbf{F_1}$ score

Evaluation Metric ($\mathbf{F_1}$ score)

$\mathbf{F_1\ score} = \frac{2}{\mathbf{Precision}^{-1} + \mathbf{Recall}^{-1}}$, where

$\mathbf{Precision} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{retrieved points}\}|}, \mathbf{Recall} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{positive points}\}|}$

Three-Set Metric

Let D_{eval} be the projection of D_{test} without labels. Denote $D^+ = D_{eval} \cap R^+$, $D^- = D_{eval} \cap R^-$, $D^u = D_{eval} \cap R^u$, and $|S|$ means the size of set S. At a specific iteration of exploration, the three-set metric is defined to be $\frac{|D^+|}{|D^+| + |D^u|}$.

Exact Lower Bound

The Three-Set Metric evaluated on D_{eval} is a lower bound of the $\mathbf{F_1}$ score if $F_{\mathcal{D}}$ is evaluated on D_{test}.

Convergence – Lower Bounds of $\mathbf{F_1}$ score

Evaluation Metric ($\mathbf{F_1}$ score)

$\mathbf{F_1\ score} = \frac{2}{\mathbf{Precision}^{-1} + \mathbf{Recall}^{-1}}$, where

$\mathbf{Precision} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{retrieved points}\}|}, \mathbf{Recall} = \frac{|\{\text{positive points}\} \cap \{\text{retrieved points}\}|}{|\{\text{positive points}\}|}$

Three-Set Metric

Let D_{eval} be the projection of D_{test} without labels. Denote $D^+ = D_{eval} \cap R^+$, $D^- = D_{eval} \cap R^-$, $D^u = D_{eval} \cap R^u$, and $|S|$ means the size of set S. At a specific iteration of exploration, the three-set metric is defined to be $\frac{|D^+|}{|D^+|+|D^u|}$.

Exact Lower Bound

The Three-Set Metric evaluated on D_{eval} is a lower bound of the $\mathbf{F_1}$ score if $F_{\mathcal{D}}$ is evaluated on D_{test}.

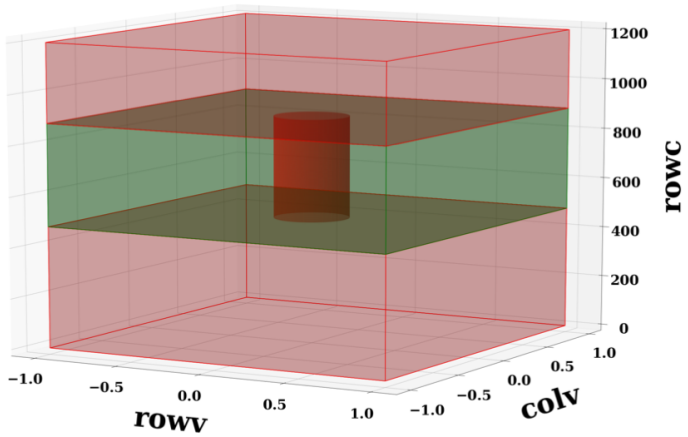
Approximate Lower Bound

More details in our paper.

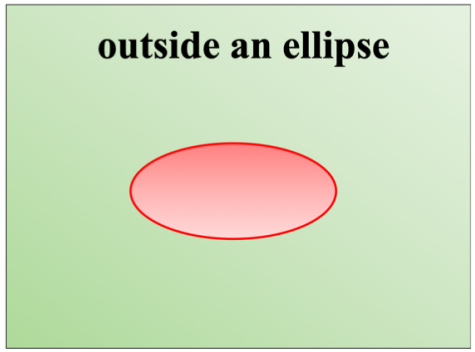
Data Space Model with factorization ($F_{\mathcal{D}_f}$)

- Conjunctive queries are a major class of database queries.

$\text{rowv}^2 + \text{colv}^2 > 0.2^2$
AND
$480 < \text{rowc} < 885$

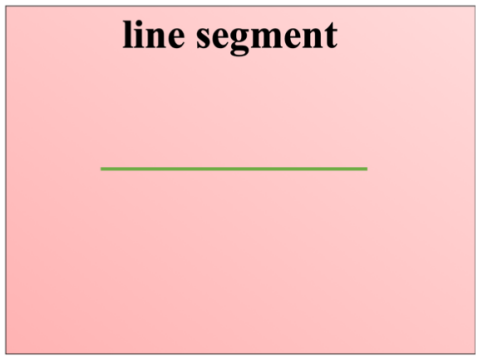


$(rowv - colv)$



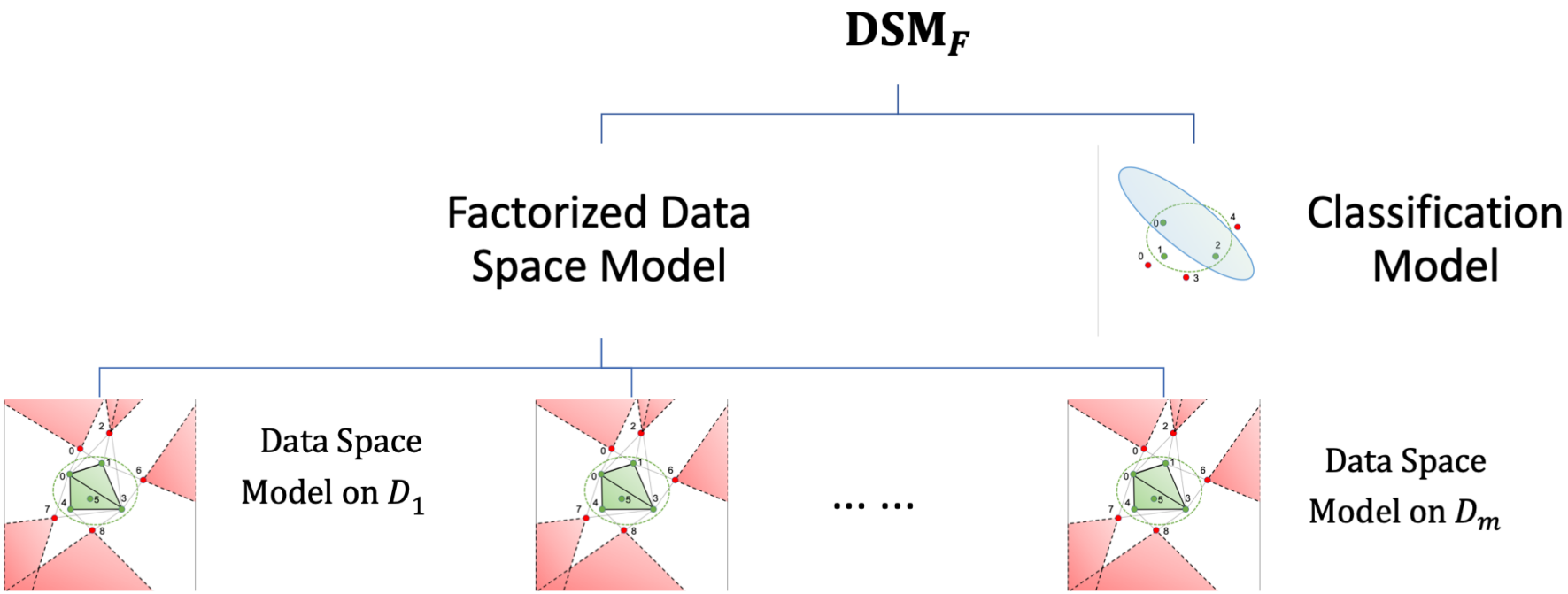
$\text{rowv}^2 + \text{colv}^2 > 0.2^2$

$(rowc)$



$480 < \text{rowc} < 885$

DSM with factorization (DSM_F)



Experimental Evaluation

- Datasets
  - Sloan Digital Sky Survey (SDSS)
    – a large sky survey database
    – 1% sample: 1.9 million tuples, 4.9GB
    – user interest queries from the SDSS query release[3]
  - Car database
    – extracted from http://www.teoalida.com/
    – 5,622 tuples
    – user interest queries from the user study

- Algorithms/Systems for comparison
  - DSM_F, DSM
  - AL-SVM(AL), $\mathrm{AL\text{-}KNN}^+$
  - AIDE, LifeJoin

[3] Sloan digital sky survey: Dr8 sample sql queries. http://skyserver.sdss.org/dr8/en/help/docs/realquery.asp.

Experimental Evaluation

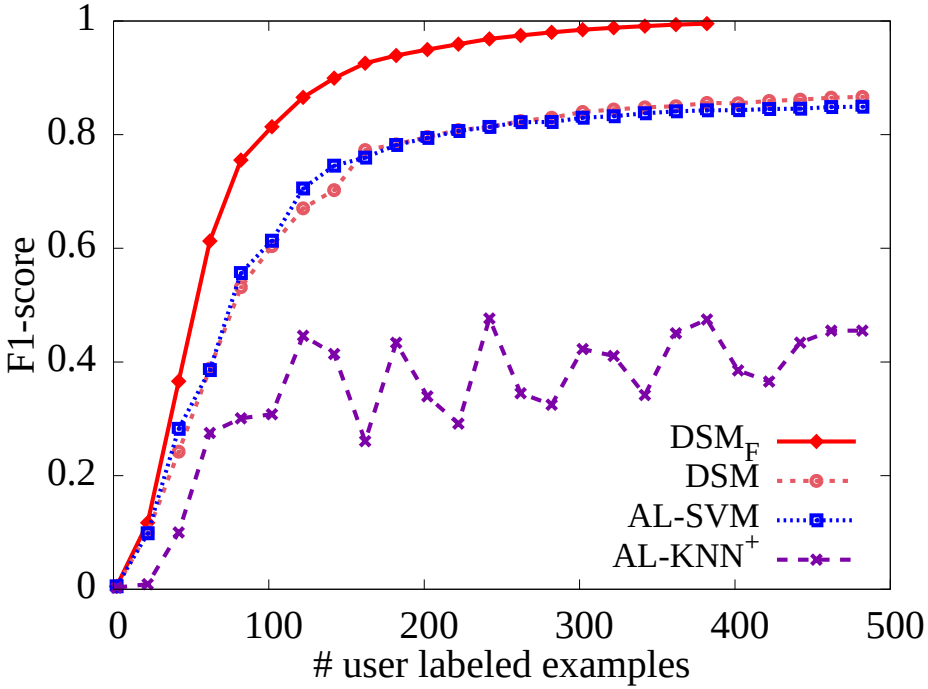


Figure: $\mathbf{F_1}$ score for 4D (0.1%) query

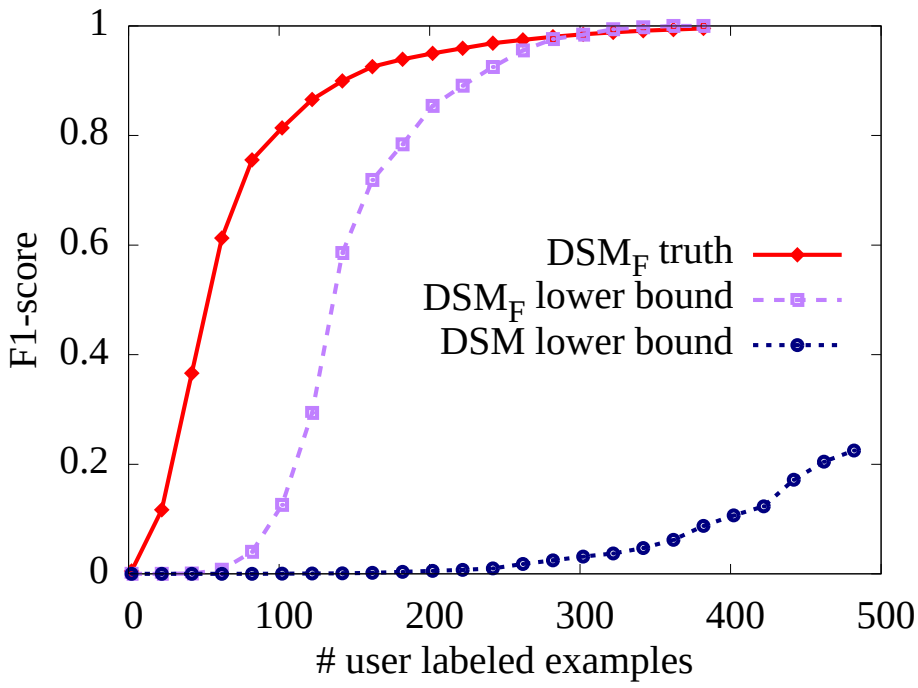


Figure: Lower bound for 4D (0.1%) query

- DSM_F outperforms other algorithms
- DSM_F improves the lower bound of $\mathbf{F_1}$ score

Experimental Evaluation

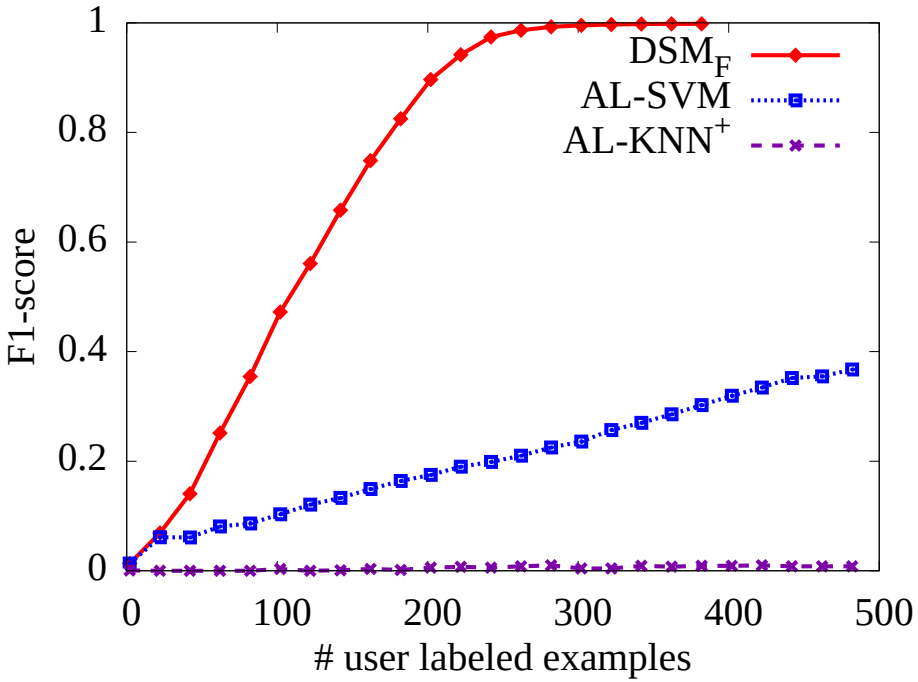


Figure: $\mathbf{F_1}$ score for 6D (0.01%) query

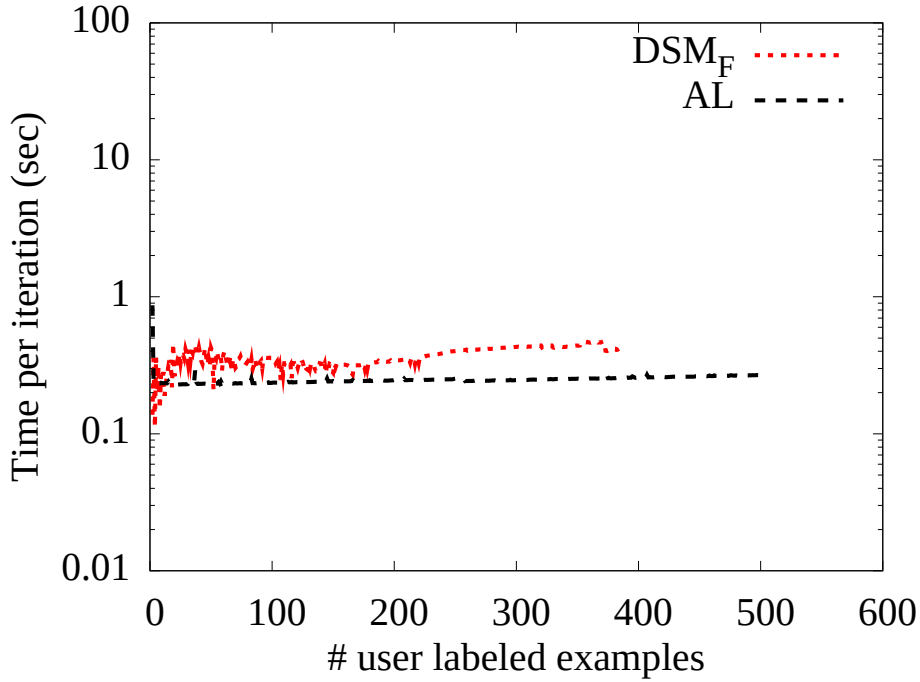


Figure: Time for 6D (0.01%) query

- DSM_F significantly outperforms the others for queries with lower selectivity and higher dimensionality
- interactive performance: within 1-2 seconds per iteration

Experimental Evaluation

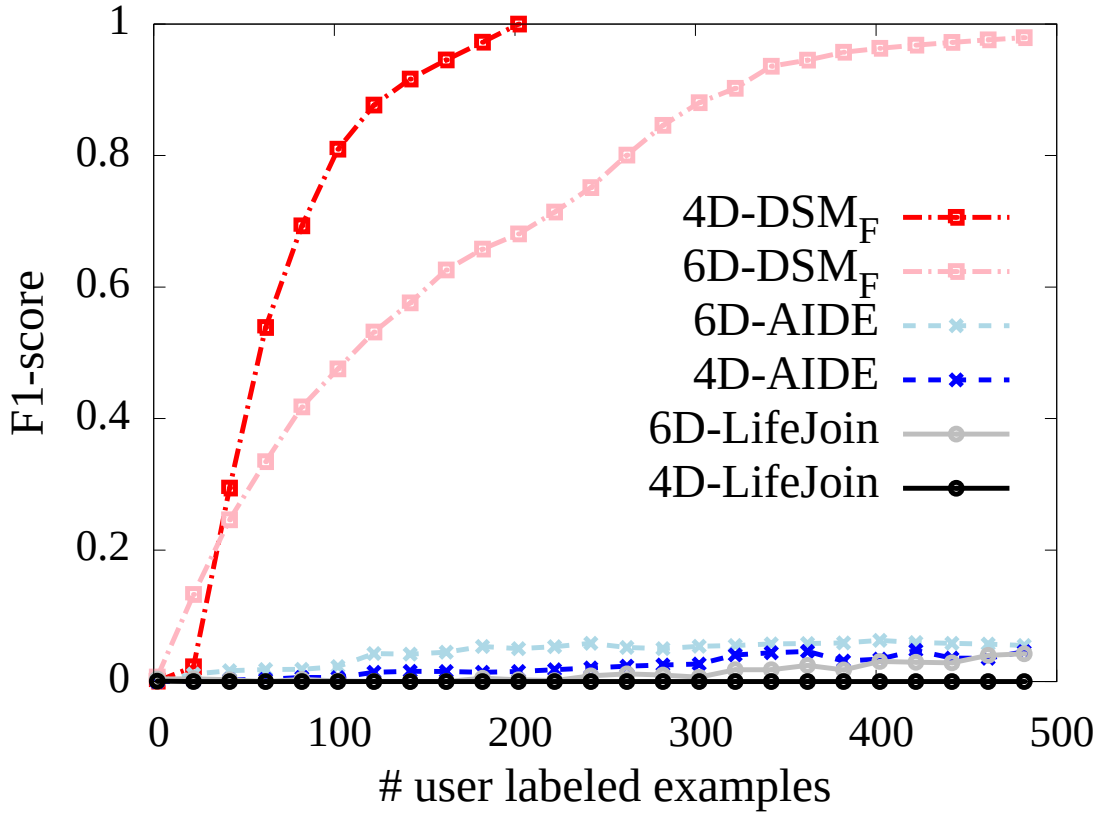


Figure: DSM_F vs. AIDE and LifeJoin

- DSM_F consistently wins against AIDE and LIfeJoin

Summary

- Take-away messages
  - design new algorithms for IDE in the active learning framework
  - leverage the **subspatial convex and conjunctive properties of database queries** to overcome the **slow convergence problem**
  - our DSM algorithm significantly outperforms the existing IDE systems in **accuracy** and **convergence speed**, while maintaining the per-iteration time within 1-2 seconds.

- Future work
  - address **inconsistent labeling** by extending our DSM model to a probabilistic model
  - extend query patterns to **multiple disjoint areas** using exploration versus exploitation
  - leverage **data properties** to further improve accuracy

Thank you!

Questions ?