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

Enhui Huang, Luciano Di Palma, Laurent Cetinsay, 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
  2. Formal results on convergence
  3. Scaling to large datasets

---

**Factorized Version Space**

- Version Space $V$ (size = 16)
- Version Space $V$ (size = 8)

---

**References**