**Probabilistic Predicates**

- **Key Ideas**
  - Complex query predicates: Large space of possible predicates, costly to train/store per predicate, indiv. PPs do not generalize
  - Challenge 1: Build PPs for arbitrary input?
    - Dimensionality, Sparsity, Linear separability?
    - Any classifier that fits $f(x) < \theta$ can work for PP
    - Pre-process: dim. reduction (PCA, feature hashing.) (Opt.)
    - Model Selection
  - Challenge 2: Use PPs for complex predicates w/o per-query training (i.e., small overhead)
    - 5 cols $\times$ 10 values $\Rightarrow \geq 10^5$ predicates

**Problem**

- Images $\rightarrow$ `YOLOv2` $\rightarrow$ $\sigma$ $\rightarrow$ Result

**Predicate pushdown? X**

- No harm on query accuracy, but less speedup

**Predicate pushdown? X**

- False Negative

**Predicate pushdown? X**

- True Negative

**PP Discards $f(x) \leq \theta$**

- Accuracy/reduction tradeoff: choose threshold ($\theta$)

**Images $\rightarrow$ `YOLOv2` $\rightarrow$ $\sigma$ $\rightarrow$ Result**

- Accuracy = 3/3, Reduction = 5/10

**Images $\rightarrow$ `YOLOv2` $\rightarrow$ $\sigma$ $\rightarrow$ Result**

- Accuracy = 2/3, Reduction = 7/10

**Technical problems addressed**

- Train PPs for diverse inputs - SVMs, KDE, shallow NNs - model selection.
- Avoid per-query training - train PP for simple preds., use QO to build PP comb.
- Sizable gains on images, video and text datasets

**Interactive Demo**