Xi'an, Shaanxi, China
SIGMOD/PODS 2021

IDENTIFYING INSUFFICIENT DATA COVERAGE FOR ORDINAL CONTINUOUS-VALUED ATTRIBUTES

Abolfazl Asudeh, Nima Shahbazi

UIC UNIVERSITY OF ILLINOIS CHICAGO

InDeX Lab
Innovative
Data
eXploration
Laboratory

Zhongjun Jin, H. V. Jagadish

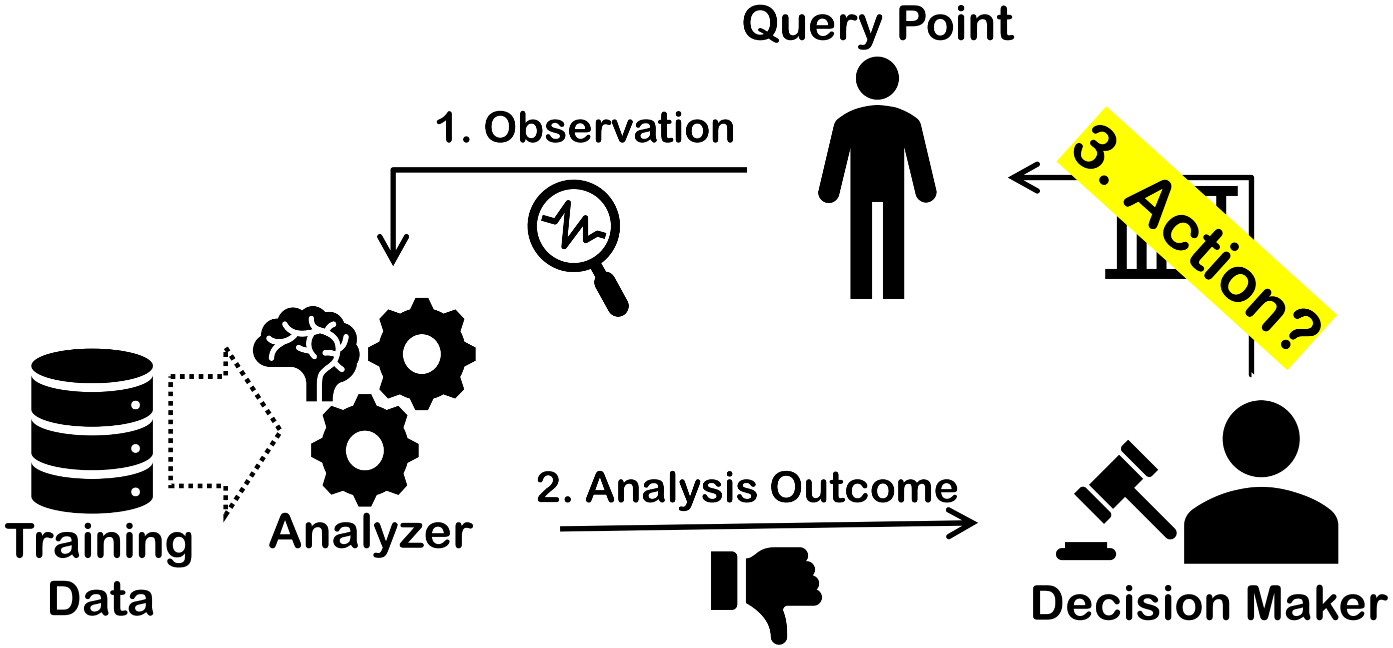
UNIVERSITY OF MICHIGAN

DATABASE RESEARCH GROUP
UNIVERSITY OF MICHIGAN

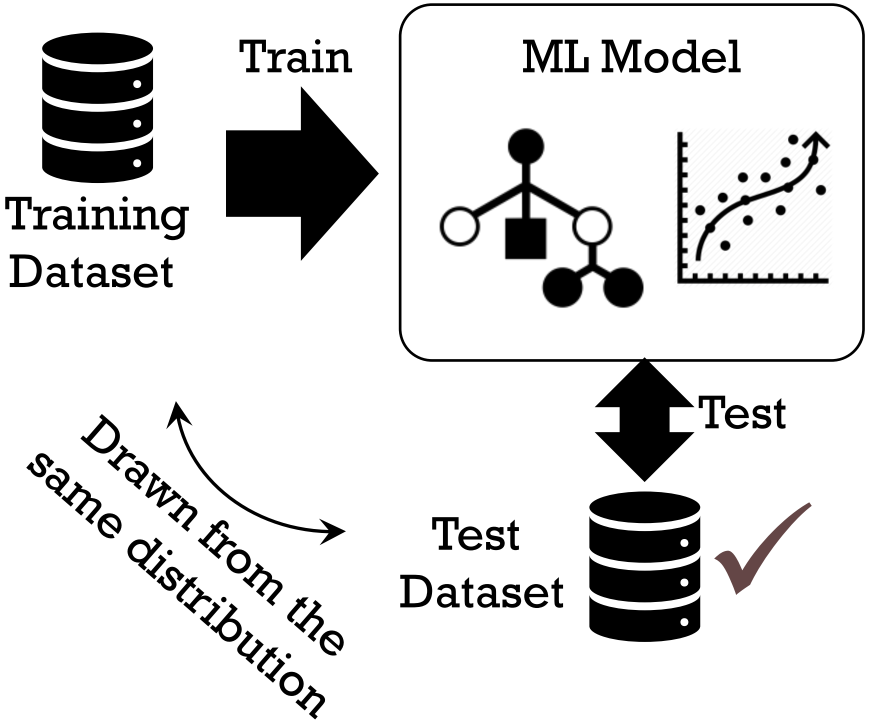
OUTLINE

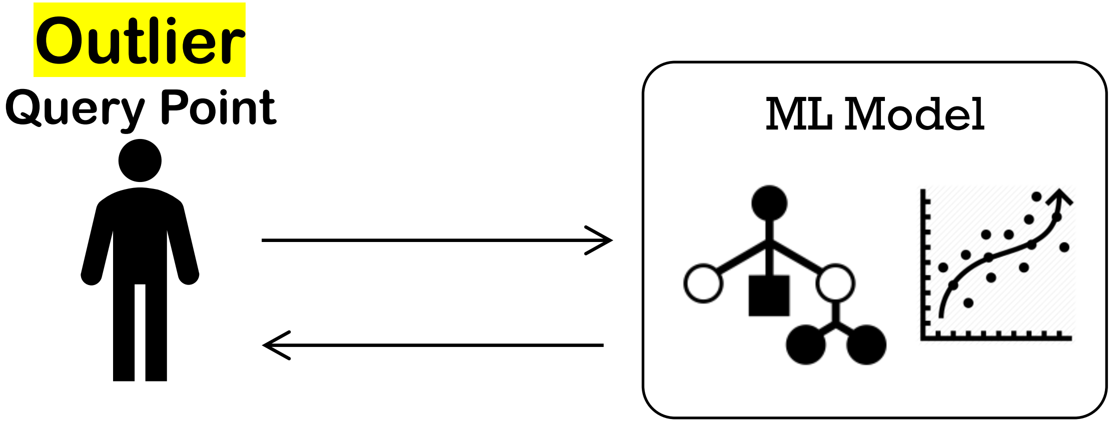
- Motivation
- Coverage
- Coverage in 2D
- Coverage in MD
- Experiments

MOTIVATION


Query Point
1. Observation
3. Action?
Training Data
Analyzer
2. Analysis Outcome
Decision Maker

MOTIVATION





(Lucky): Predictable by non-outlier points → 👍

(Unlucky): Not Predictable → 👎

COVERAGE

- We may not trust the outcome, if the query point is an outlier.

- The query point **q** is covered by training data, if
  - there are at least **k** (training) points in neighborhood

$$Cov_{\rho,k}(q, \mathcal{D}) = \begin{cases} true & if\ |\{t \in \mathcal{D} \mid \Delta(t, q) \leq \rho\}| \geq k \\ false & otherwise \end{cases}$$

- w/o loss of generality, we use ℓ_2 norm for the distance function

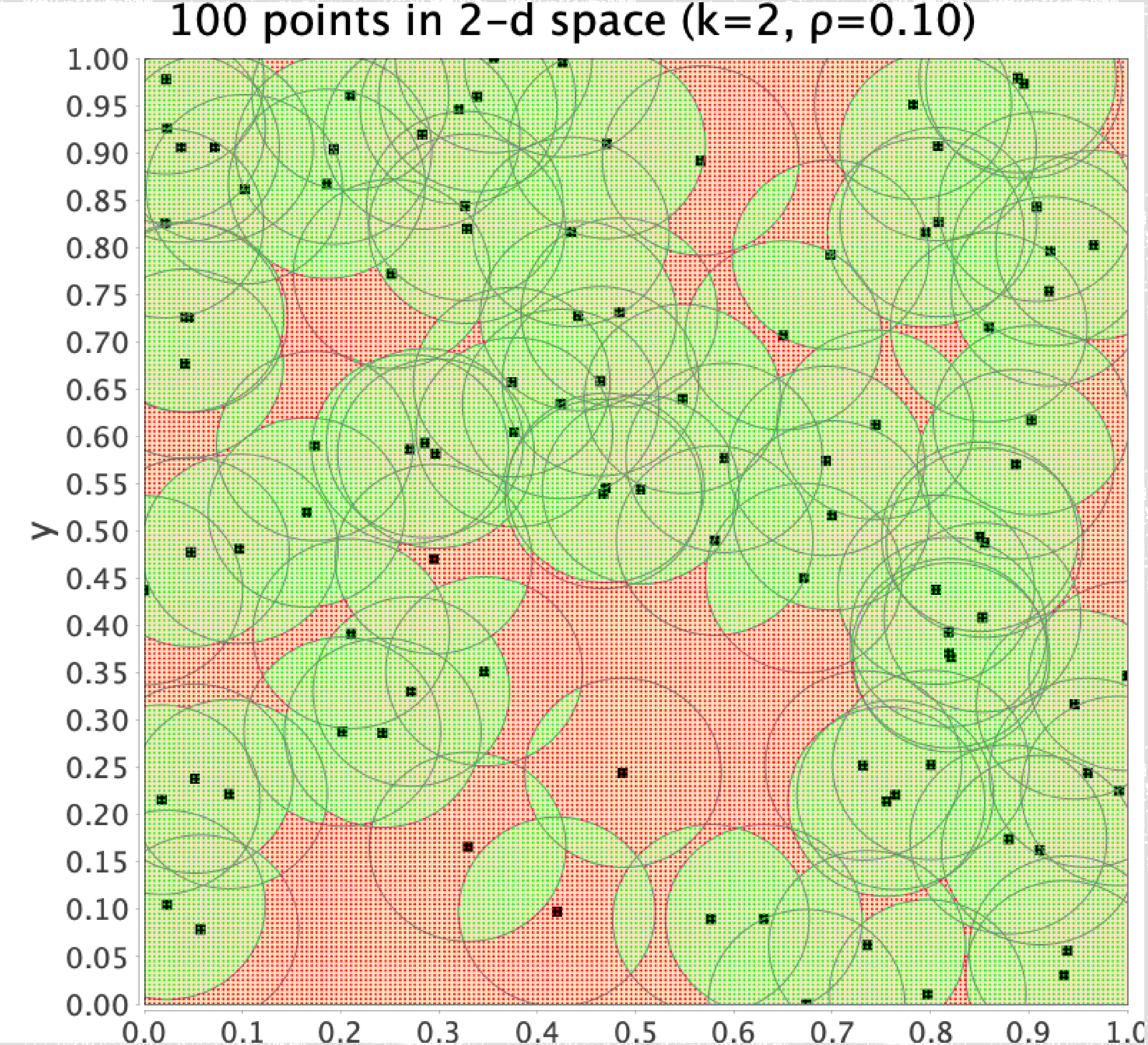
UNCOVERED REGION

- The collection of all uncovered points – any query point in this region is uncovered

- Given a dataset D with d attributes (features) $x_1 \ldots x_d$, a distance function $\Delta: R^d \times R^d \rightarrow R$, a vicinity value ρ, and a threshold value k, the uncovered region U is the set of points (value combinations) that are not covered by D. Formally:

$$U = \{q \in [0,1]^d \mid Cov(q, D) = false\}$$

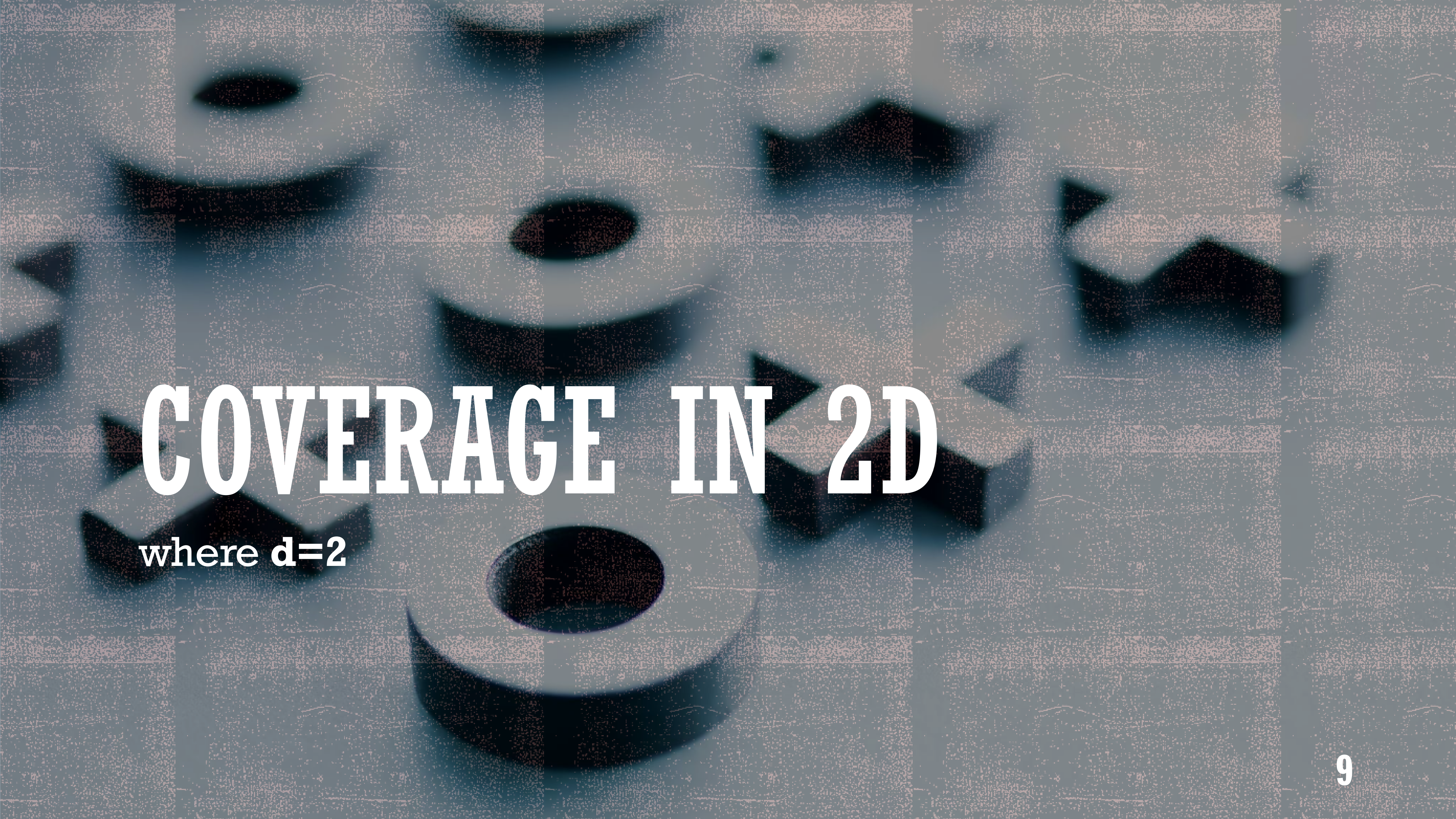
UNCOVERED REGION EXAMPLE



PROBLEM FORMULATION

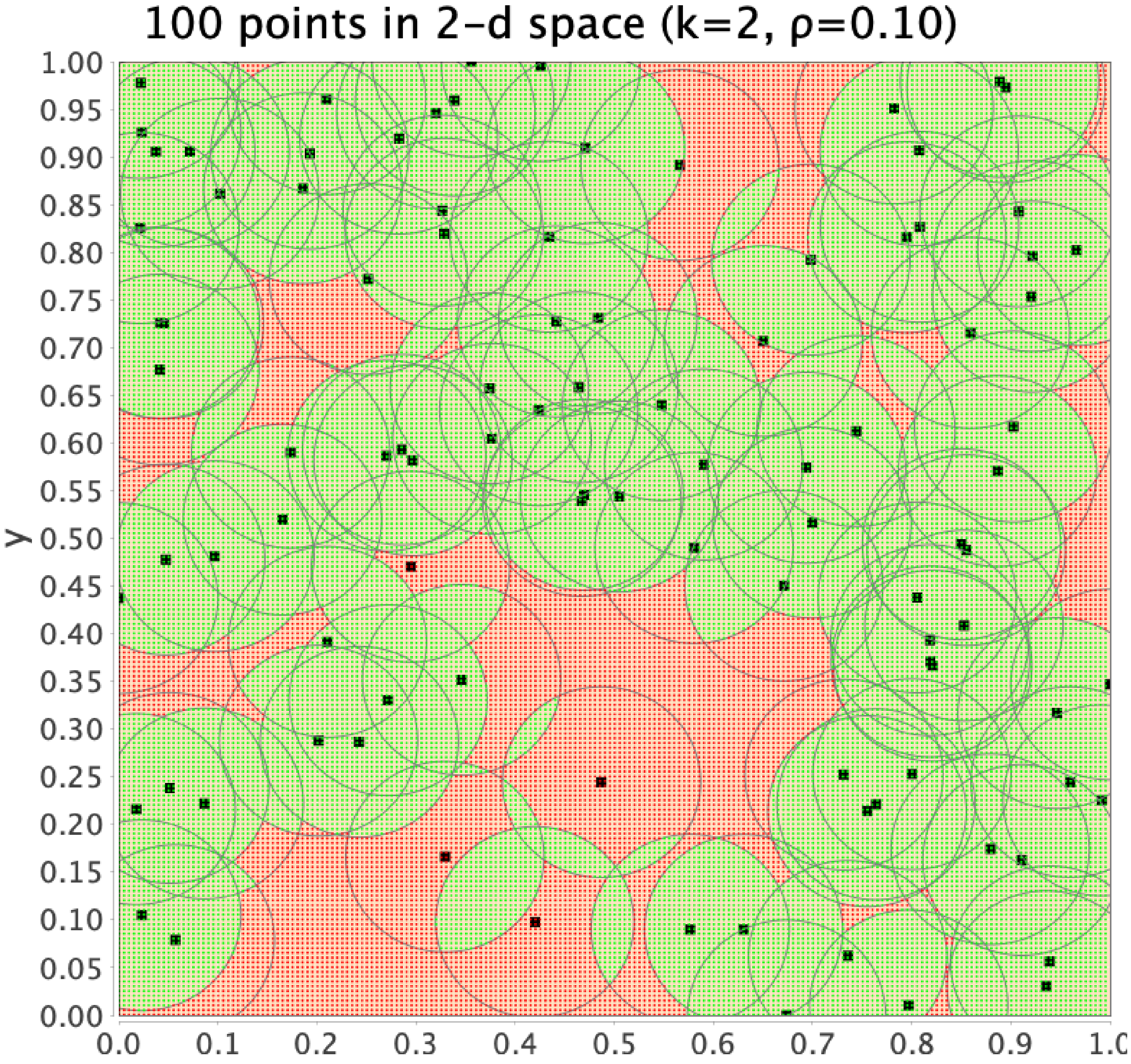
- Problem 1 (Uncovered Region Discovery): Given a dataset **D**, identify the uncovered region
  - Dataset *Annotation*: shows *potential deficiencies* in the (training) data set.

- Problem 2 (Uncovered Query Answering): Given the uncovered region, identify if a query point **q** is uncovered.



COVERAGE IN 2D

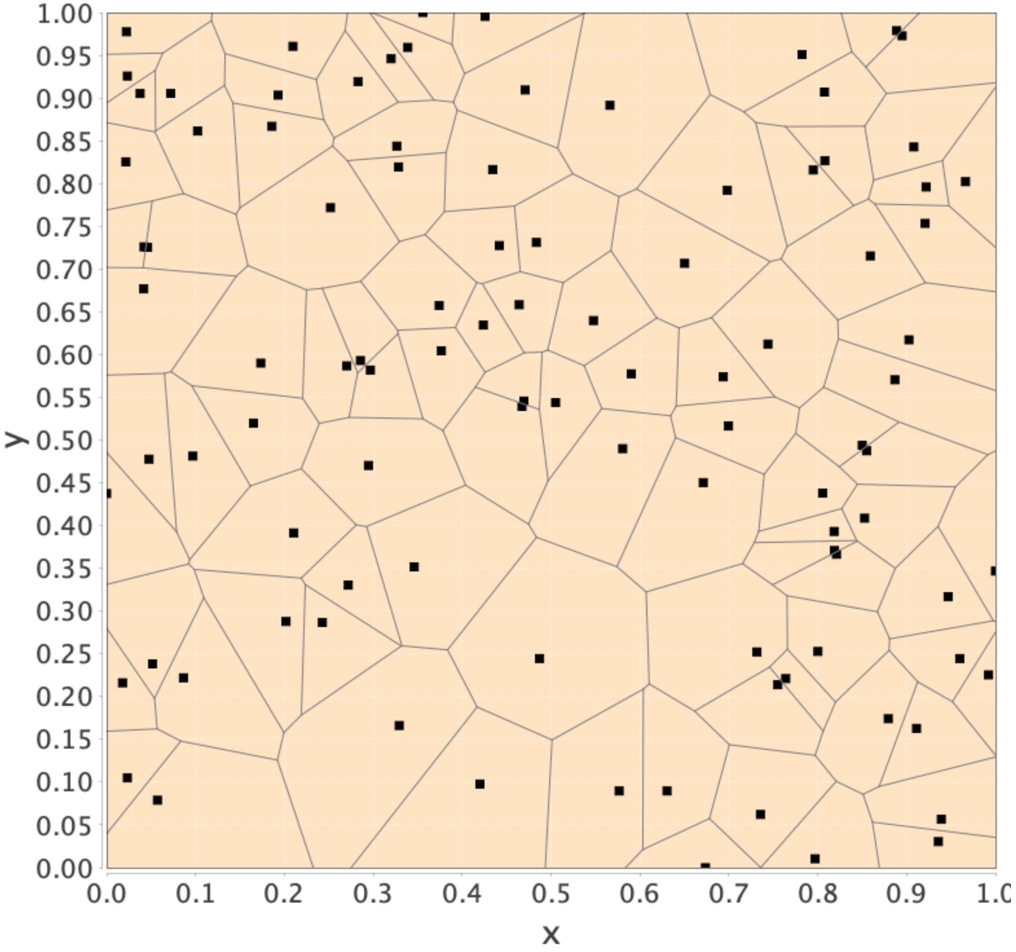
where **d=2**



UNCOVERED REGION EXAMPLE

(REVIEW): VORONOI DIAGRAMS

100 points in 2-d space (k=1, ρ=0.10)

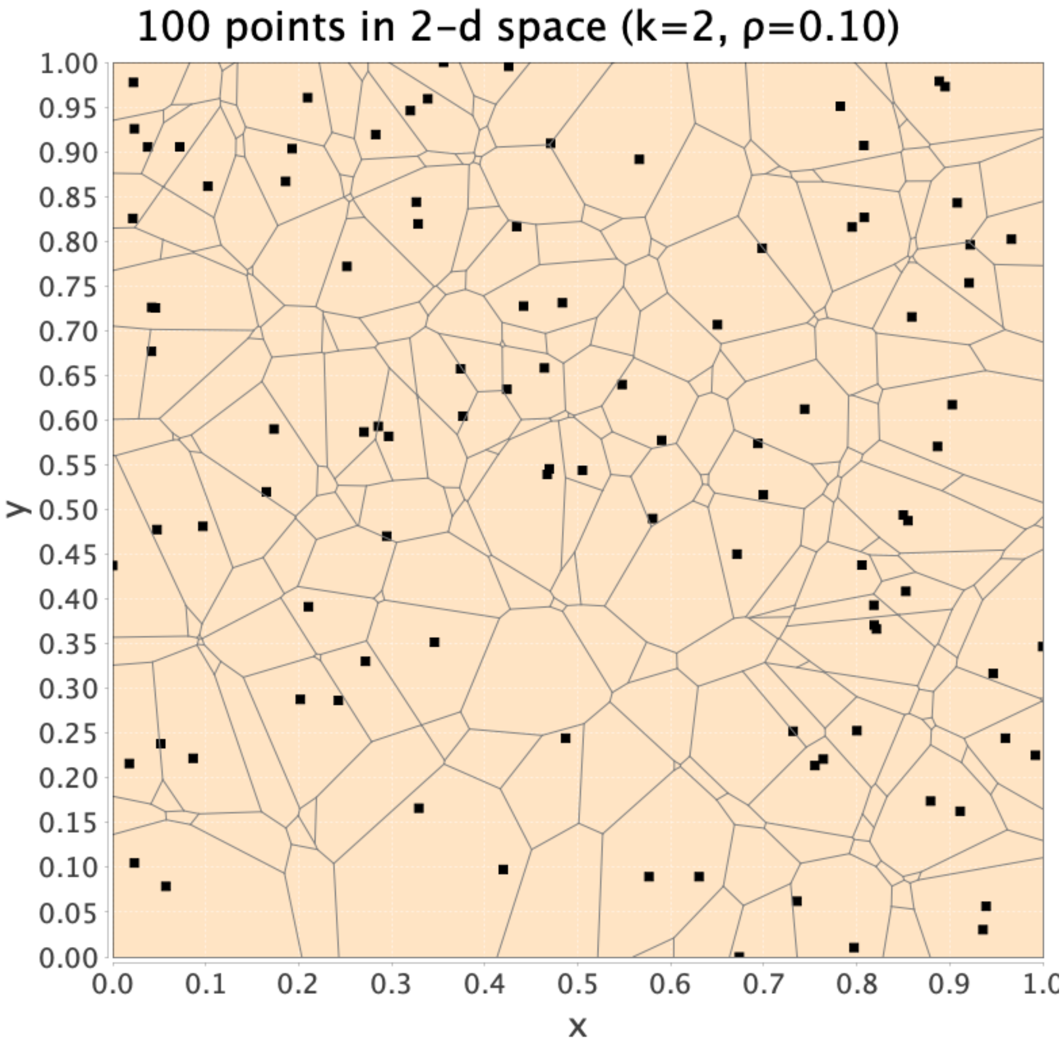


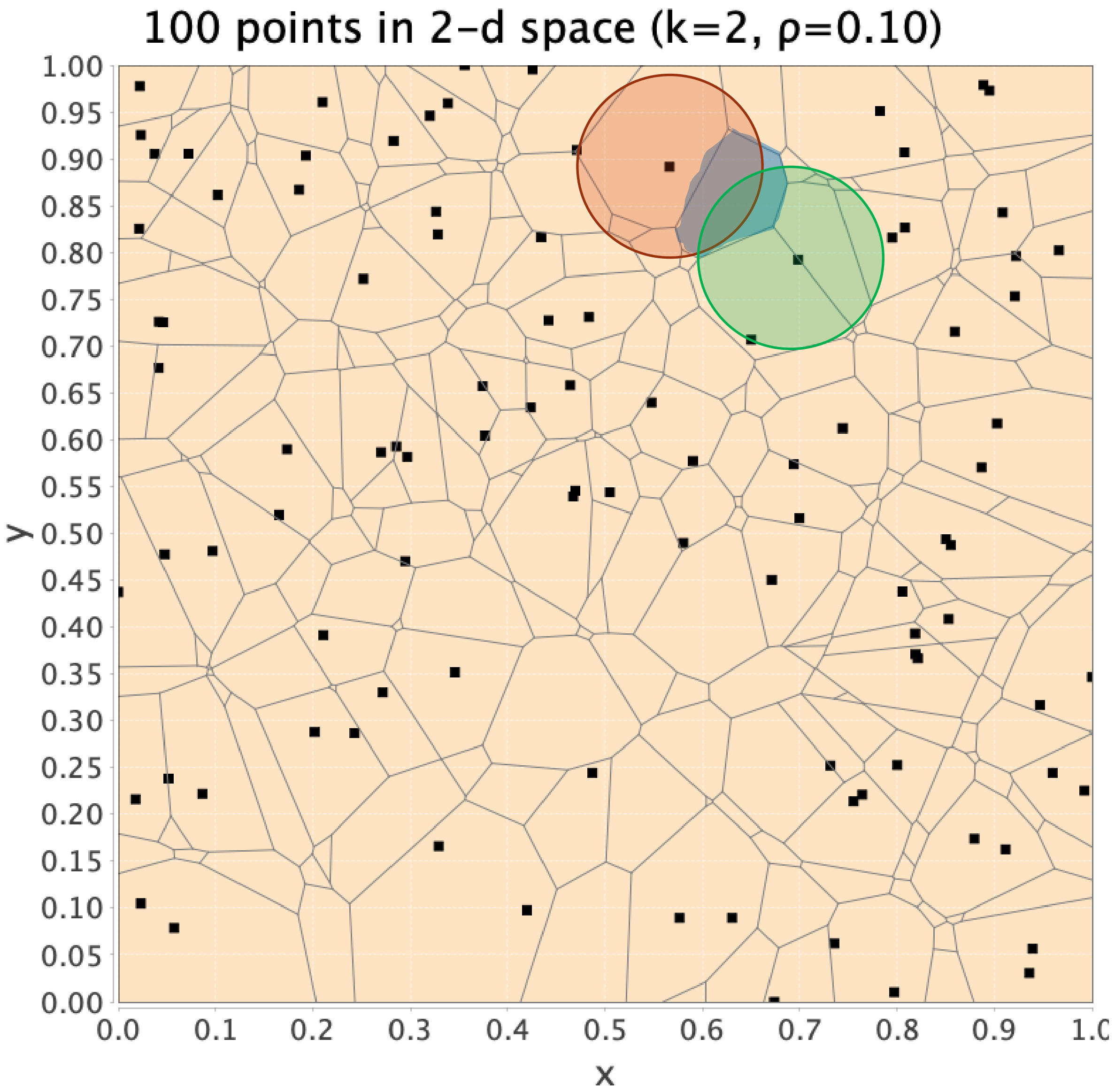
- Partition of a plane with **n** points into cells, such that all points in each cell have the same nearest point.

(REVIEW): K-VORONOI DIAGRAMS

- Extend the notion of Voronoi diagrams from nearest neighbor to k-nearest neighbor

- $O(k(n-k))$ cells
- Construction [D. T. Lee et al.]:
  - Time: $O(k^2\, n \log(n))$
  - Space: $O(k^2(n-k))$
- Query time:
  - $O(\log n)$





Connection to k-Voronoi Diagrams

- Uncovered Region Discovery :
  - Construct the k-Voronoi diagram
  - For every Voronoi cell $V(S)$:
    - Add the region outside the intersection $\cap\, O_t\ \forall\, t \in\ S$ to the uncovered region

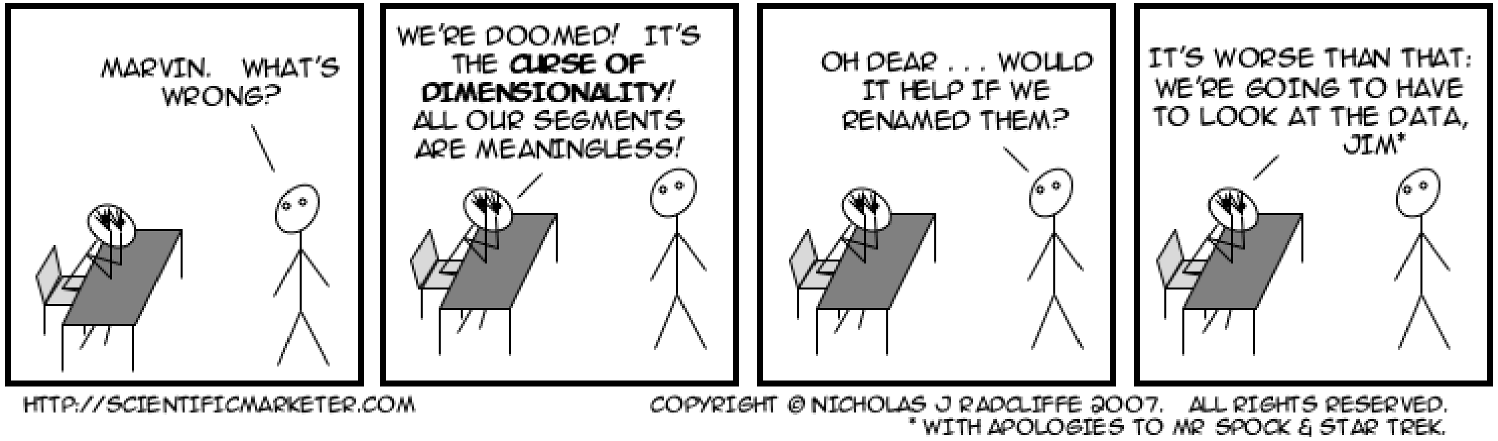
- Uncovered Query Answering:
  - Find the cell **V(S)** that **q** belongs to
  - return **uncovered** iff $\exists t \in S$ s.t. $\Delta(q,t) > \rho$

COVERAGE IN MD

where **d**≥**2**

EXTENDING 2D CASE TO MD

- **Theoretically**: Yes, but…
- **Practically**: No, due to the curse of dimensionality


MARVIN. WHAT'S WRONG?
WE'RE DOOMED! IT'S THE CURSE OF DIMENSIONALITY! ALL OUR SEGMENTS ARE MEANINGLESS!
OH DEAR . . . WOULD IT HELP IF WE RENAMED THEM?
IT'S WORSE THAN THAT: WE'RE GOING TO HAVE TO LOOK AT THE DATA, JIM*
HTTP://SCIENTIFICMARKETER.COM
COPYRIGHT © NICHOLAS J RADCLIFFE 2007. ALL RIGHTS RESERVED.
* WITH APOLOGIES TO MR SPOCK & STAR TREK.

LEARN THE UNCOVERED REGION *(APPROXIMATELY)*

- **High-level idea**:
  - Construct an $\boldsymbol{\epsilon}$**-net** by sampling "enough" query points:
    - A sample point is labeled as +1 if uncovered, -1 otherwise
  - Learn the uncovered region boundary using the $\boldsymbol{\epsilon}$**-net**

- **Negative result** (A theoretical upper-bound on the complexity of uncovered region)
  - In $\mathbf{R^d}$, the VC-dimension of the uncovered region is bounded by

$$\boldsymbol{O}(\,(\boldsymbol{d}+\mathbf{1})\,\boldsymbol{n}^{\left\lfloor \frac{\boldsymbol{d}}{2} \right\rfloor}\,\boldsymbol{k}^{\left\lceil \frac{\boldsymbol{d}}{2}+\mathbf{1} \right\rceil}\,)$$

LEARN THE UNCOVERED REGION *(APPROXIMATELY)*

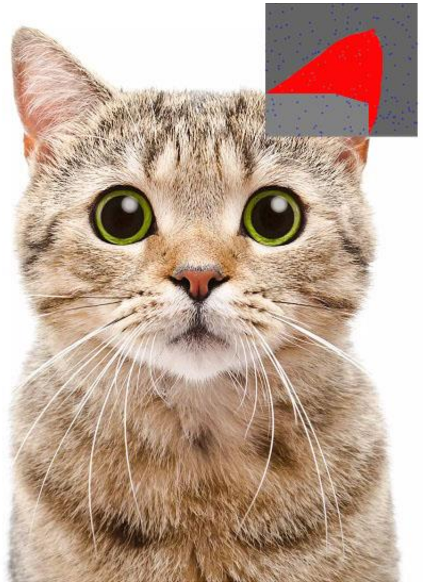
- **Practical Resolution**:
  - **Observation**: The boundary complexity depends on the number of arcs constructing it – which can be significantly less than the upper-bound
  - **High-level idea**: Apply an **exponential search** on the number of samples, until the result forms an ϵ**-net**

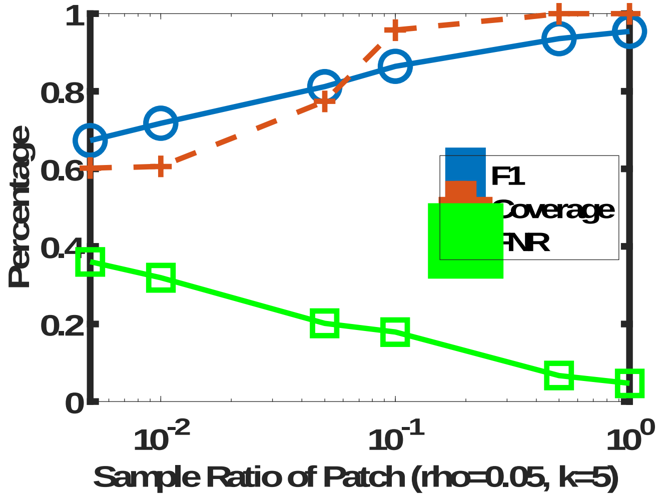
- **Uncovered Query Answering:**
  - Pass the query point **q** to the learned classifier.

EXPERIMENTS

PROOF OF CONCEPT: CLASSIFICATION

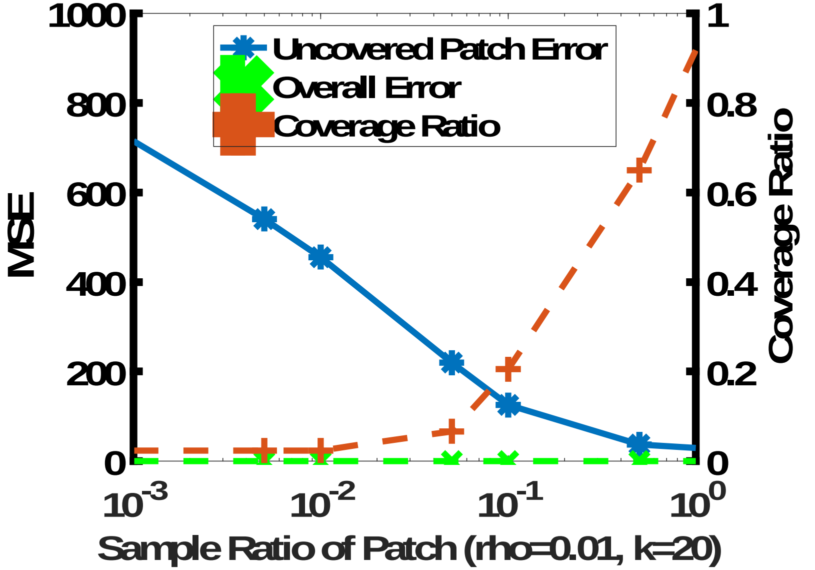
- **Goal**: Determine whether a query point belongs to the body of a cat image or background
- **Experiment**:
  - Removing the samples from the highlighted rectangle to make it uncovered
  - Overall F1 vs. Uncovered region's F1
  - False-Negatives in Red
  - Decision boundary in uncovered region
  - Effect of gradually adding points to the patch





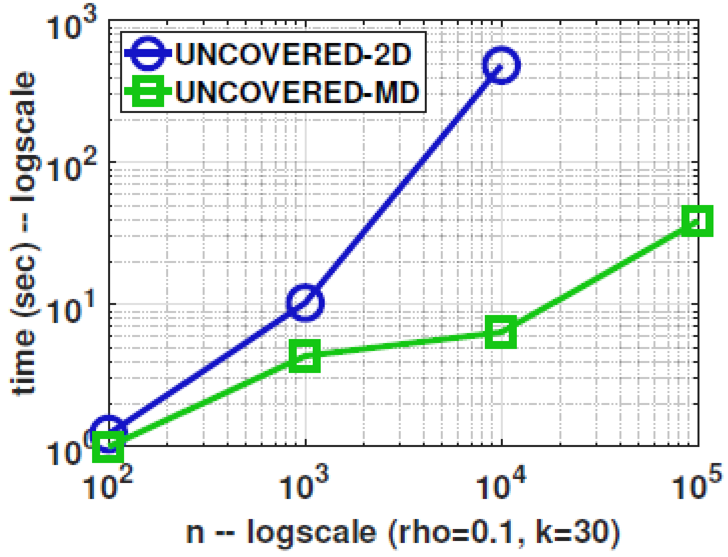
PROOF OF CONCEPT 2: REGRESSION

- **Goal:** Predict Altitude of a query point based on (Longitude, Latitude)
- **Experiment**:
  - RN dataset: (Longitude, Latitude, Altitude)
  - Removing samples from a cell in the range *10<Longitude<10.6* and *57.1<Latitude<57.6* with highly fluctuating *Altitudes* to make it uncovered
  - Overall prediction error vs. Uncovered region's prediction error
  - Effect of gradually adding points to the patch

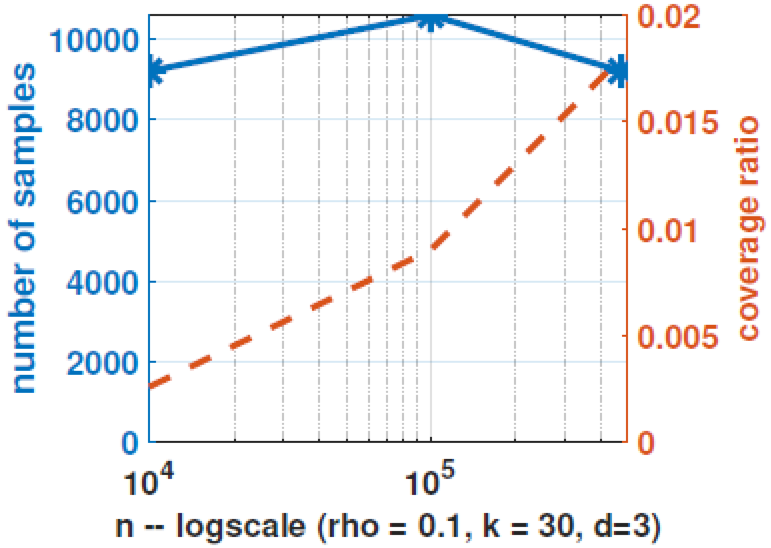


PERFORMANCE EVALUATION

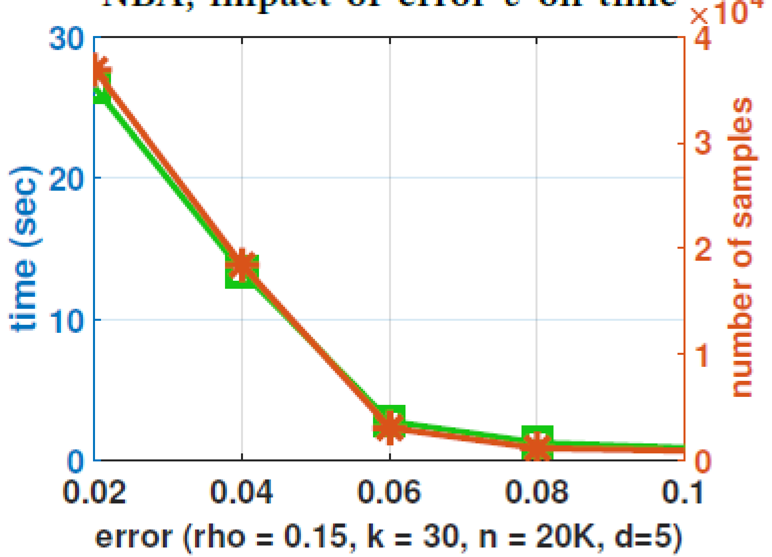
FI, impact of n on time



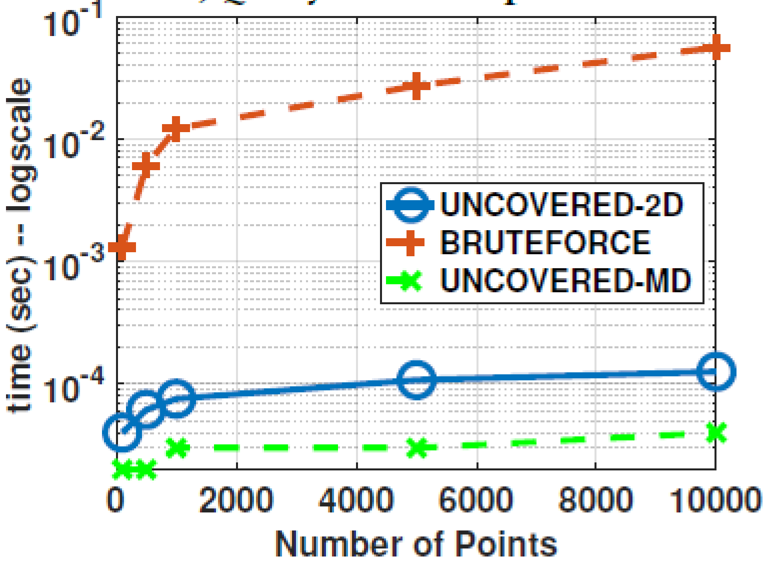
DOT, impact of n on sample size



NBA, impact of error e on time



RN, Query time comparison



THANK YOU

- Abolfazl Asudeh, asudeh@uic.edu, www.cs.uic.edu/~asudeh/

  @ab_asudeh

- Nima Shahbazi, nshahb3@uic.edu
- Zhongjun Jin, markjin@umich.edu, https://markjin1990.github.io/
- H. V. Jagadish, jag@umich.edu, web.eecs.umich.edu/~jag/