THE UNIVERSITY OF TEXAS AT DALLAS

IEEE ICDE 2020

April 20th -24th, 2020 - Dallas, Texas, USA

Essence

University of Glasgow

School of Computing Science

# SuRF: Identification of Interesting Data Regions with Surrogate Models

**Fotis Savva,** Christos Anagnostopoulos, Peter Triantafillou

**School of Computing Science, University of Glasgow**

**Essence: Pervasive & Distributed Intelligence**

**E-mail:** f.savva.1@research.gla.ac.uk

# Outline

Essence

University of Glasgow

School of Computing Science

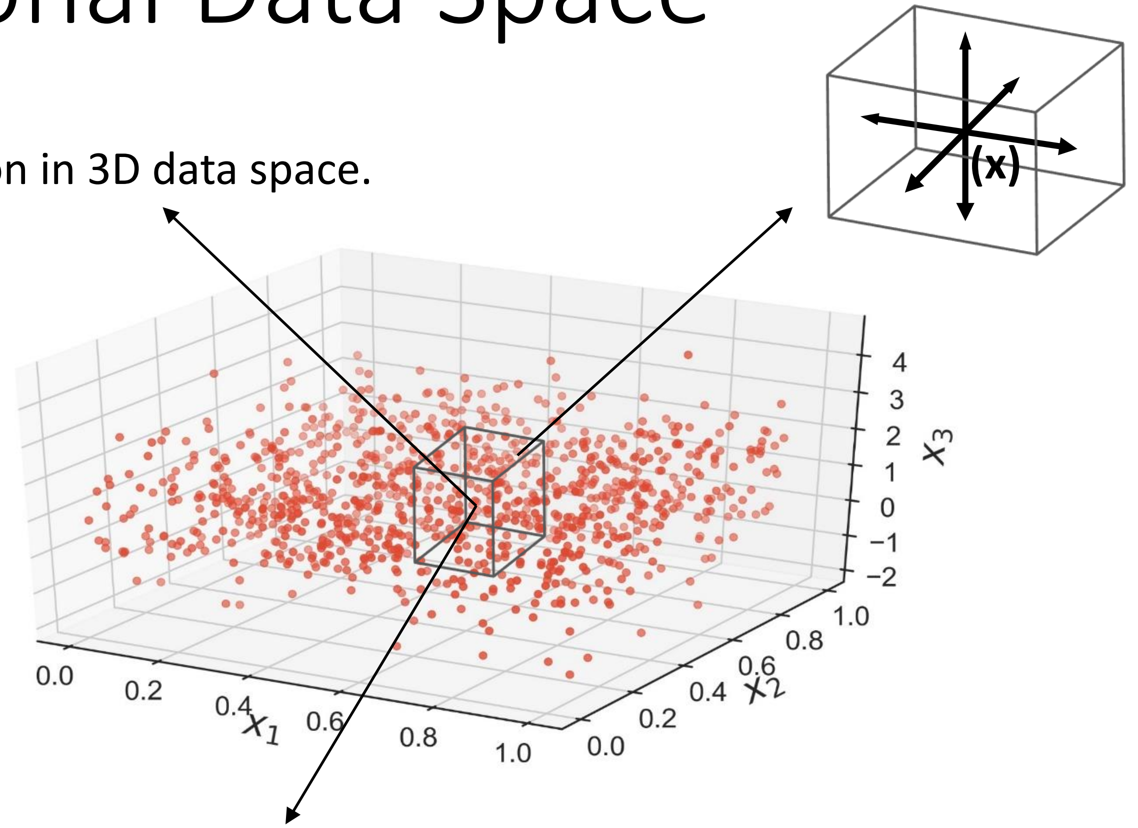
- Introduce problem of identifying sub-regions (examples and then formulation – aggregate function as a function mapping parameters to aggregate values)
- Baseline solution (visually) and complexity of baseline
- Introduce optimization problem – why multi-modal
  - Glow-worm optimization
  - Back-end data analytics system as the bottleneck
- Surrogate functions to learn back-end analytics system
- Experimental Evaluation – Accuracy, Efficiency, Example with Crimes

Essence | University of Glasgow | School of Computing Science

# Regions in Multi-dimensional Data Space

- **Region:** hyper-rectangle in multi-dimensional space
- Formally: A region has a center point $\boldsymbol{x} \in \mathbb{R}^d$ with side lengths $\boldsymbol{l} \in \mathbb{R}^d$
- Region's **interestingness** is inferred by an aggregate statistic over the retrieved data points.

A region in 3D data space.




Using only the included data points, compute a statistic such as **COUNT/AVG/SUM/VAR**

**Example :** $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$ are the values (X, Y, Z) obtained from an accelerometer and the aggregate statistic the ratio of a specific activity (ie '*sitting'*).

$\max_{x,l} f(x,l) - y_R$

$f : \mathcal{R}_{x,l} \to \mathbb{R}$

Essence

# Identifying Interesting Regions

- Identify regions $\mathcal{R}_{x,l}$ which are greater/lower than threshold $y_R \in \mathbb{R}$

e.g., the ratio for activity '*sitting*' is over 0.3 – meaning >30% of data points were generated while performing activity '*sitting*'.)

$$\max_{\boldsymbol{x},\boldsymbol{l}} f(\boldsymbol{x},\boldsymbol{l}) - y_R$$

$f : \mathcal{R}_{\boldsymbol{x},\boldsymbol{l}} \to \mathbb{R}$

**Aggregate Function**

$$\max_{\boldsymbol{x},\boldsymbol{l}} \frac{f(\boldsymbol{x},\boldsymbol{l}) - y_R}{\left(\prod_{i=1}^{d} l_i\right)^c}$$

$c$ → Adjusting tolerance

**To control the size of a region, we introduce a penalizing factor on its lengths**

Essence
University of Glasgow
School of Computing Science

# Baseline Method

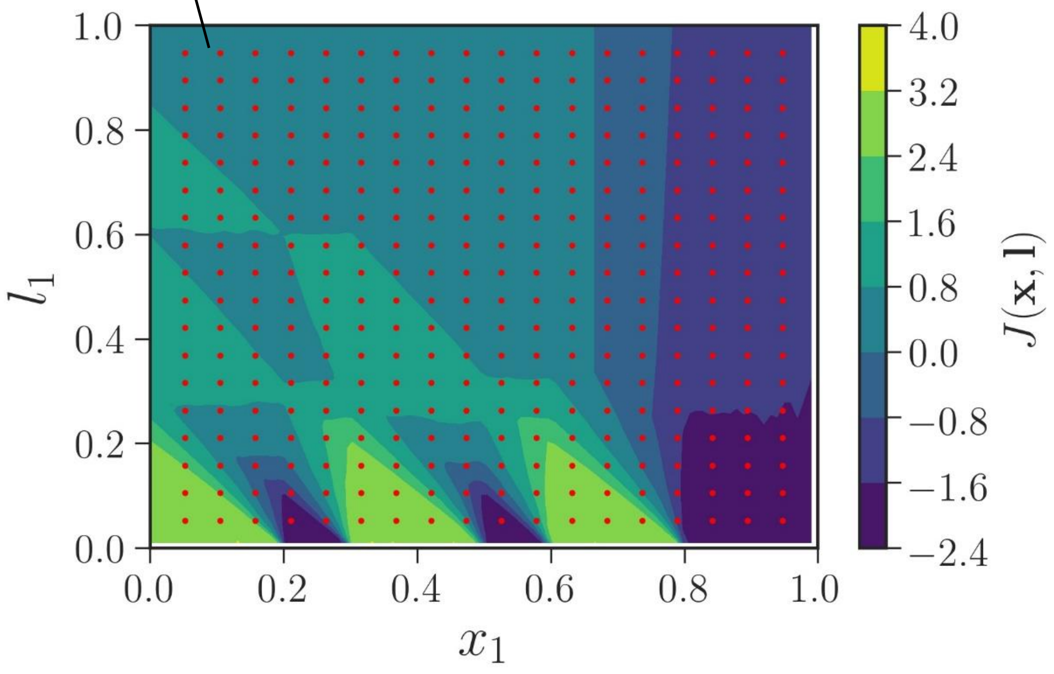
- Let us find a region $R_{x,l}$ for a single dimension.
- Discretise the space such that:
  - $x \in X$ and $l \in L$
  - Granularity : $\mathrm{n} = |X|, \mathrm{m} = |\mathrm{L}|$
- Evaluate the objective function

$$\max_{x \in X\ , l \in L} J(x, l)$$

to identify interesting regions

Each point $(x, l)$ has to be evaluated. Resulting in $n \times m$ evaluations for 1D

Colorbar denotes the value of the objective function



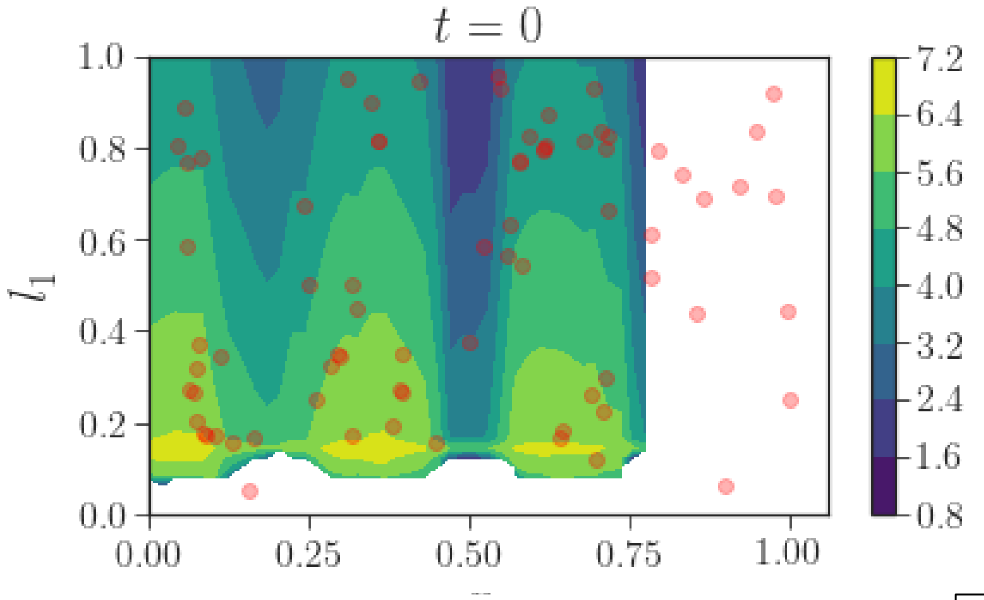

The number of evaluations is exponential w.r.t the dimensions $(n \times m)^d$

For 1D : $(20 \times 20)^1 = 400$
For 3D : $(20 \times 20)^3 =$ 64,000,000 – Function Evaluations !

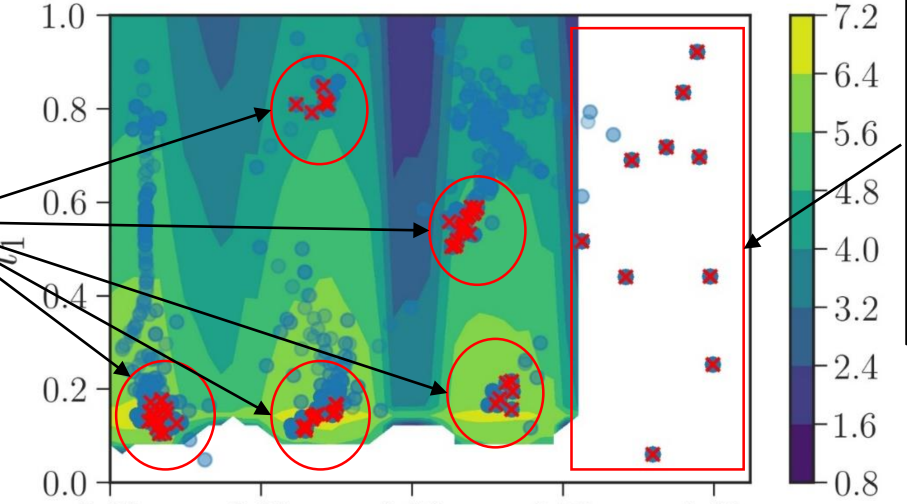
University of Glasgow
School of Computing Science
Essence

# Introducing Multi-modal Optimization

- **Objective function is multi-modal** as there could be many regions of interest.
- Function $f$ is **unknown**
- Adopt GlowWorm Swarm Optimization (GSO) to **locate** regions of interest.




Particles in GSO only influence other particles within a neighbourhood




Infeasible solutions become stranded and are pruned away

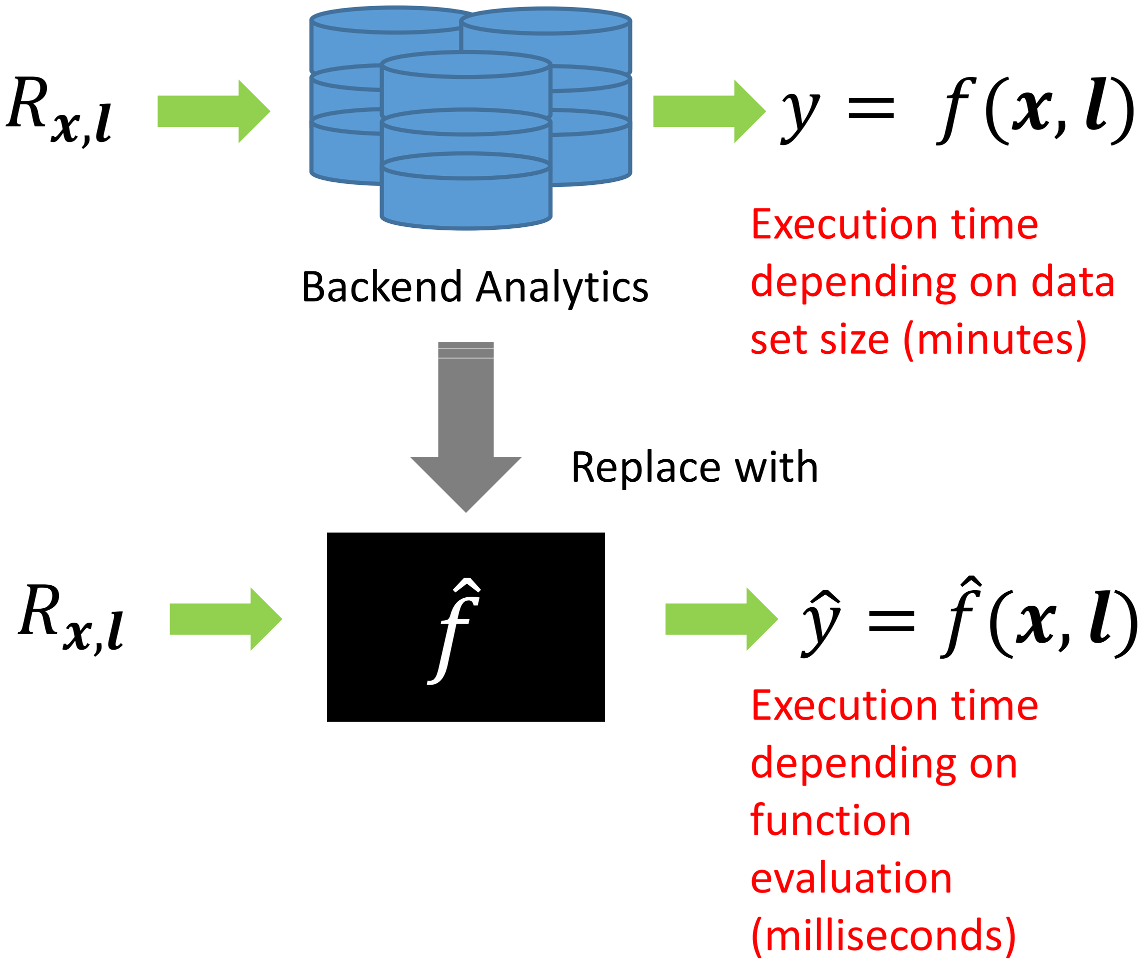
Essence University of Glasgow School of Computing Science

# Backend Analytics System is a ‘bottleneck’

- GSO **effectively** reduced the number of function evaluations to $O(TL)$.
- $T = iterations;$
- $L = number\ of\ particles$

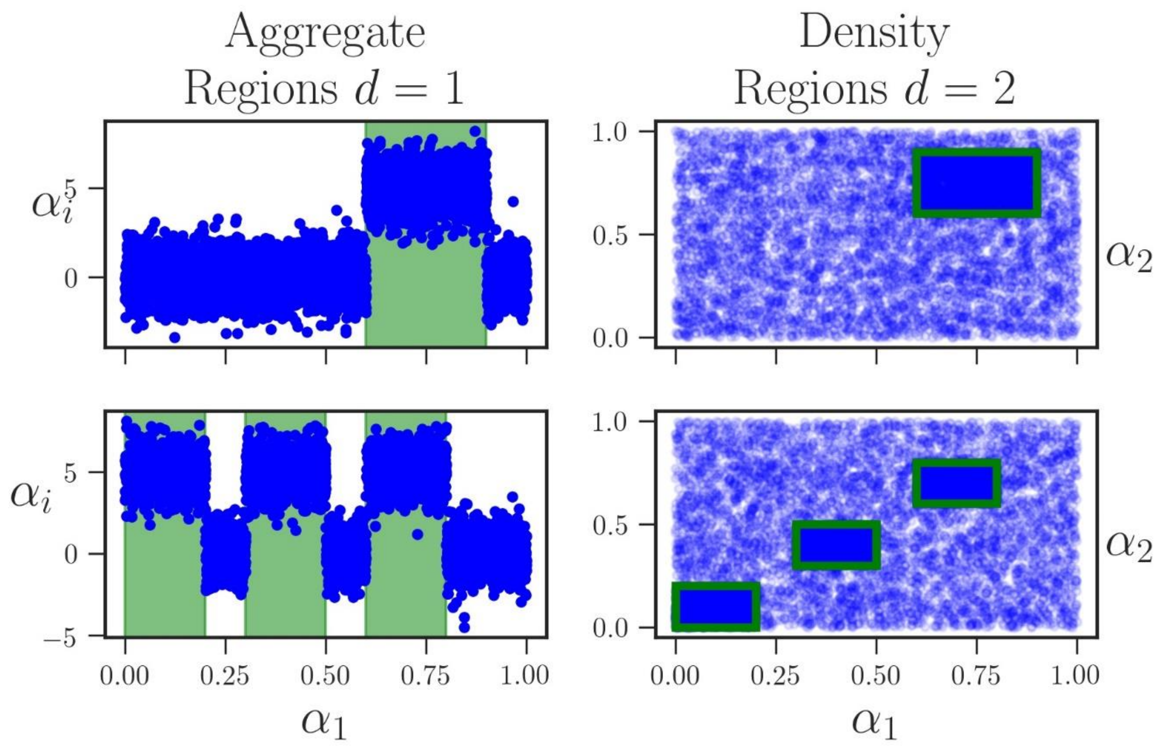
- Still have to compute $f(\boldsymbol{x}, \boldsymbol{l})$ over large data sets.

- **Solution:** Treat $f$ as a black-box function -> approximate using $\hat{f}$
  - A surrogate function trained using past function evaluations



# Experimental Evaluation - Task

- Artificial Continuous Region(s) :
  a) where the aggregate statistic is larger
  b) highly concentrated number of data points
- **Task**: identify $R_{\boldsymbol{x},\boldsymbol{l}}$ , i.e., boundaries of such sub-regions
- **Accuracy:** *Jaccard Similarity Index – (Intersection Over Union)*
  - *How much the predicted region overlaps with the pre-defined region*

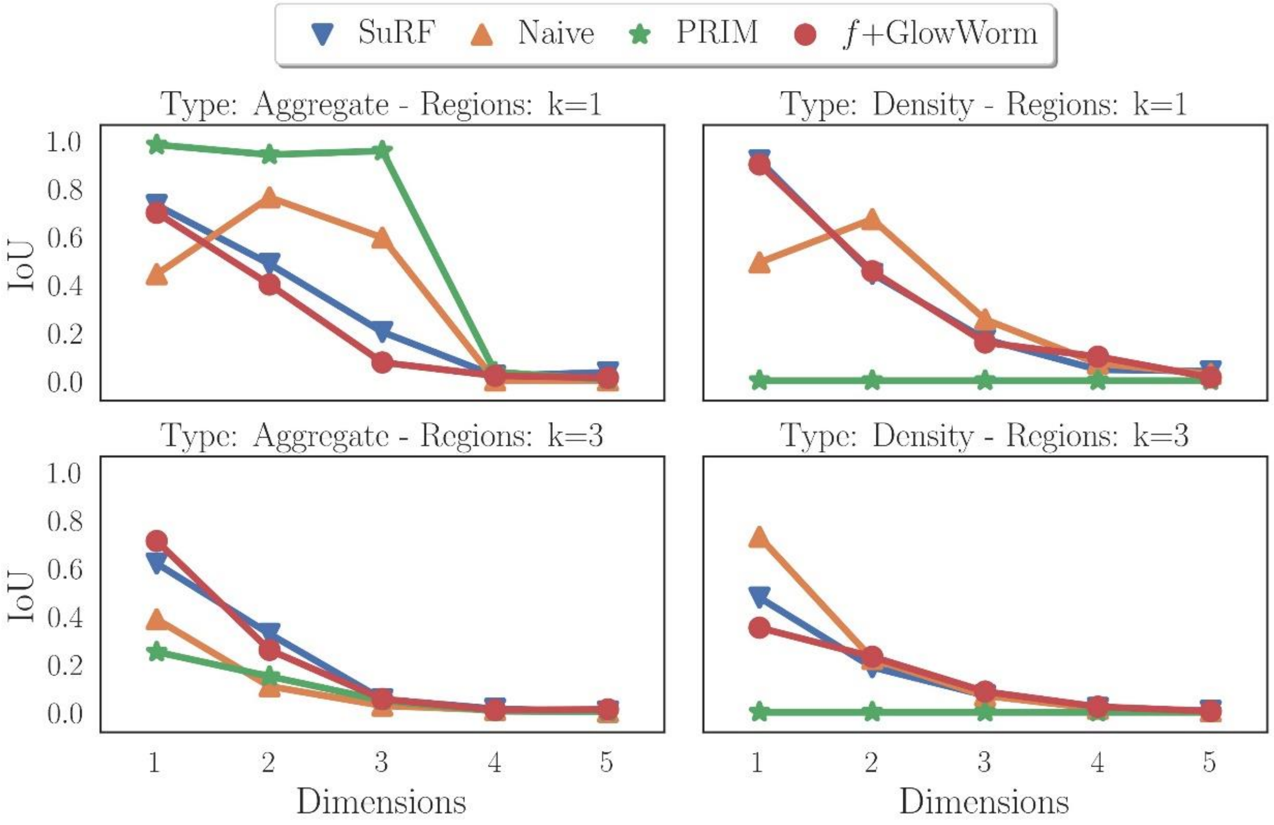



Essence
University of Glasgow
School of Computing Science

# Experimental Evaluation – Key Results (1)

Essence
University of Glasgow
School of Computing Science

- **SuRF** (proposed method)
- **Naïve:** discretisation process described in the beginning
- **PRIM** by Friedman & Fisher [1]
- ***f* + GlowWorm:** GSO with backend analytics




[1] Friedman, Jerome H., and Nicholas I. Fisher. "Bump hunting in high-dimensional data." *Statistics and Computing* 9.2 (1999): 123-143.

# Experimental Evaluation – Results (2)

Increase in data set size $N$

Increase in data set dimensionality

**Cuttoff time set at 5 minutes (3000 seconds)**

TABLE I
COMPARATIVE ASSESSMENT OF DIFFERENT METHODS.

| Method | Data size $N$ / $d$ dim. | $10^5$ | $10^6$ Time (sec) | $10^7$ |
|---|---|---|---|---|
| **SuRF** | 1 | 1.28 | 1.28 | 1.3 |
| | 2 | 1.4 | 1.4 | 1.4 |
| | 3 | 1.35 | 1.35 | 1.35 |
| | 4 | 1.63 | 1.63 | 1.64 |
| | 5 | 1.68 | 1.68 | 1.69 |
| Naive | 1 | 0.01 | 0.16 | 1.94 |
| | 2 | 3.22 | 33.72 | 341.7 |
| | 3 | 115.49 | 1221.6 | - (22%) |
| | 4 | - (66%) | - (6%) | - (0.5%) |
| | 5 | - (1%) | - (0.1%) | - (0.01%) |
| $f$+GlowWorm | 1 | 4.71 | 51.9 | 601.32 |
| | 2 | 26.7 | 280.14 | 2856.02 |
| | 3 | 26.46 | 289.5 | 2808.42 |
| | 4 | 27.1 | 293.62 | 2981.81 |
| | 5 | 30.21 | 320.03 | - |
| PRIM | 1 | 0.15 | 0.4 | 4.8 |
| | 2 | 0.2 | 1.9 | 32.2 |
| | 3 | 0.56 | 9.3 | 46.3 |
| | 4 | 0.9 | 9.5 | 160.5 |
| | 5 | 1.28 | 7.36 | 282.6 |

Relatively constant across all settings

As data set size and dimensionality increase method does not scale.

Similar to Naive

PRIM is highly scalable, starts to degrade as data set size increases

Essence

School of
Computing Science

# Thank you!

**Essence: Pervasive & Distributed Intelligence**
http://www.dcs.gla.ac.uk/essence/

**E-mail**: f.savva.1@research.gla.ac.uk