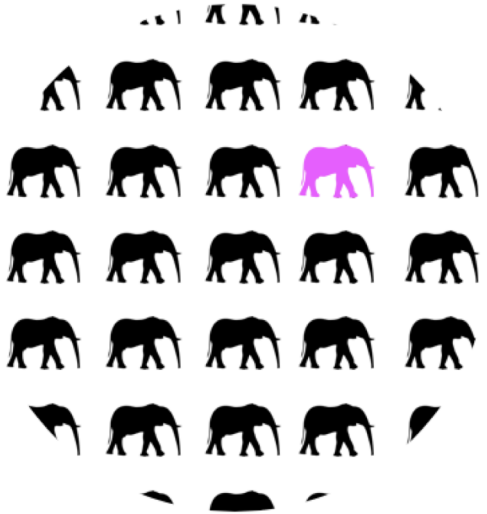
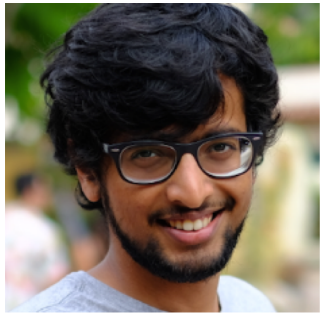
# PIDForest: Anomaly Detection via Partial Identification




**Parikshit Gopalan**

VMware Research



**Vatsal Sharan**

Stanford

**Udi Wieder**

VMware Research

# Anomaly/Outlier detection

**Problem**: Given an *unlabeled* dataset, find points that deviate from normal.

- Ubiquitous problem in unsupervised learning with several applications.

# Anomaly/Outlier detection

**Problem**: Given an *unlabeled* dataset, find points that deviate from normal.

- Ubiquitous problem in unsupervised learning with several applications.
- DNNs very effective for text, image, video.
- Our focus: tabular data, time-series.

# Anomaly/Outlier detection

**Problem**: Given an *unlabeled* dataset, find points that deviate from normal.

- Motivating setting: Get parameters regarding health of data center every second, want to detect unusual behavior.

| Time-stamp | CPU usage | Memory | Band-width | OS | … |
|---|---|---|---|---|---|
| 1 | 30% | 10GB | 10Gbps | Linux | |
| 2 | 35% | 11GB | 8Gbps | Mac | |
| 3 | 35% | 15GB | 8Gbps | Windows | |
| … | | | | | |

# Anomaly/Outlier detection

**Problem**: Given an *unlabeled* dataset, find points that deviate from normal.

- Motivating setting: Get parameters regarding health of data center every second, want to detect unusual behavior.

***Challenges:***

- ❑ *High-dimensional, irrelevant attributes*
- ❑ *Heterogeneous attributes*
- ❑ *Interpretability*

| Time-stamp | CPU usage | Memory | Band-width | OS | … |
|---|---|---|---|---|---|
| 1 | 30% | 10GB | 10Gbps | Linux | |
| 2 | 35% | 11GB | 8Gbps | Mac | |
| 3 | 35% | 15GB | 8Gbps | Windows | |
| … | | | | | |

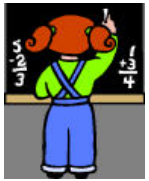
# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

## The ID game:

1. Alice and Bob share a table of all living animals.

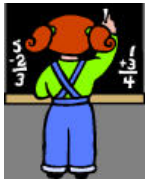


| No | Species | Color | Weight | Age | |
|---|---|---|---|---|---|
| 1 | Panther | Black | 95 | 53 | |
| 2 | Adder | Black | 30 | 36 | |
| 3 | Hornet | Green | 2 | 83 | |
| 4 | Spider | Scarlet | 1 | 2 | |
| 5 | | | | | |

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.

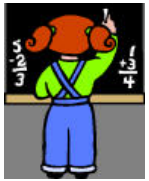


| No | Species | Color | Weight | Age | |
|---|---|---|---|---|---|
| 1 | Panther | Black | 95 | 53 | |
| 2 | Adder | Black | 30 | 36 | |
| 3 | Hornet | Green | 2 | 83 | |
| 4 | Spider | Scarlet | 1 | 2 | |
| 5 | | | | | |

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.

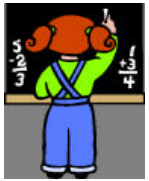


| No | Species | Color | Weight | Age | |
|---|---|---|---|---|---|
| 1 | Panther | Black | 95 | 53 | |
| 2 | Adder | Black | 30 | 36 | |
| 3 | Hornet | Green | 2 | 83 | |
| 4 | Spider | Scarlet | 1 | 2 | |
| 5 | | | | | |

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



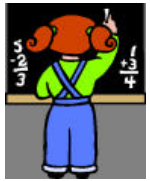
Goal: Minimize message length

| No | Species | Color | Weight | Age | |
|---|---|---|---|---|---|
| 1 | Panther | Black | 95 | 53 | |
| 2 | Adder | Black | 30 | 36 | |
| 3 | Hornet | Green | 2 | 83 | |
| 4 | Spider | Scarlet | 1 | 2 | |
| 5 | | | | | |

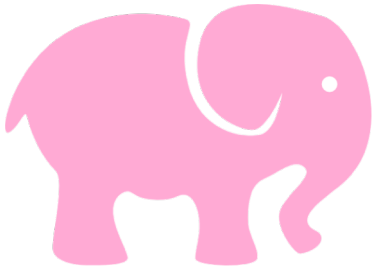
# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



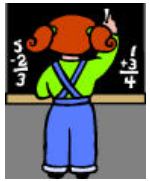
Goal: Minimize message length



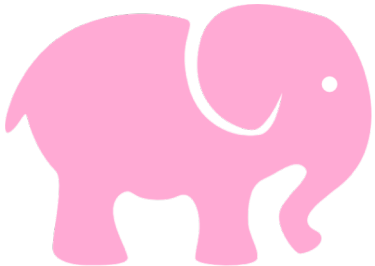
# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



Goal: Minimize message length

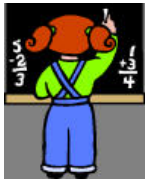


Alice: color = white, species = elephant

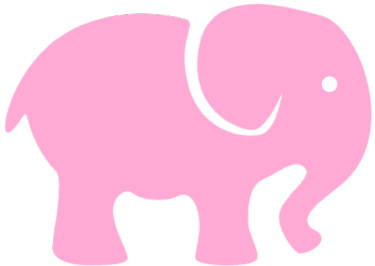
# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



Goal: Minimize message length



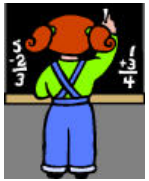
Alice: color = white, species = elephant

*Returns a list of 50.*

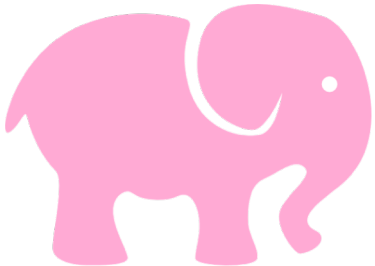
# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



Goal: Minimize message length



Alice: color = white, species = elephant

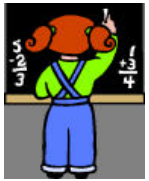
*Returns a list of 50.*

Alice: index = 30.

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.

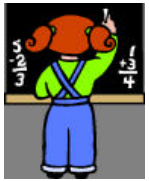


Goal: Minimize message length

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



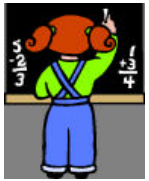
Goal: Minimize message length

Alice: color = white, species = rabbit

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



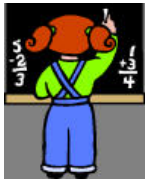
Goal: Minimize message length

Alice: color = white, species = rabbit
*Returns a list of 10000000000.*

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.



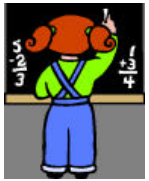
Goal: Minimize message length

Alice: color = white, species = rabbit
*Returns a list of 10000000000.*
Alice: weight = 5-6 lbs, age = 2-3 years, ...

# Partial Identification (PID) [Gopalan-S.-Wieder'2019]

The ID game:

1. Alice and Bob share a table of all living animals.
2. Row number, followed by attribute information.
3. Alice wants to tell Bob what she sees.

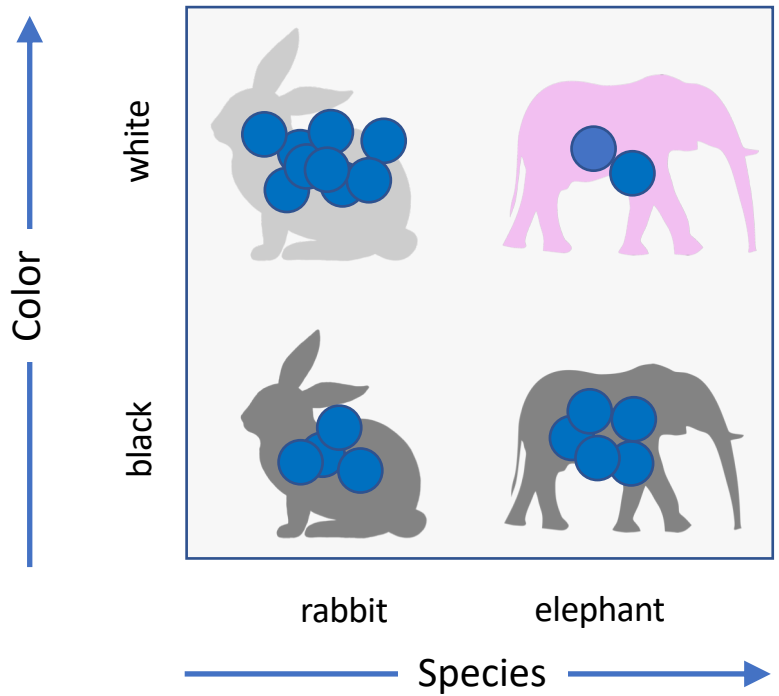


Goal: Minimize message length

**Takeaway**: Anomalies are easy to partially identify.

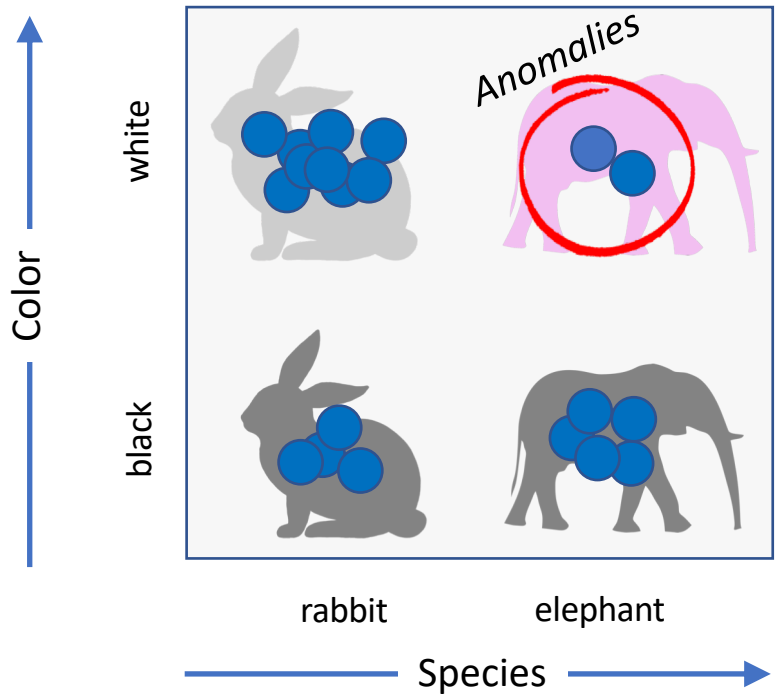
# Anomaly Score from Partial Identification (PIDScore)

Idea: Anomalies are points lying in relatively sparse regions of space.



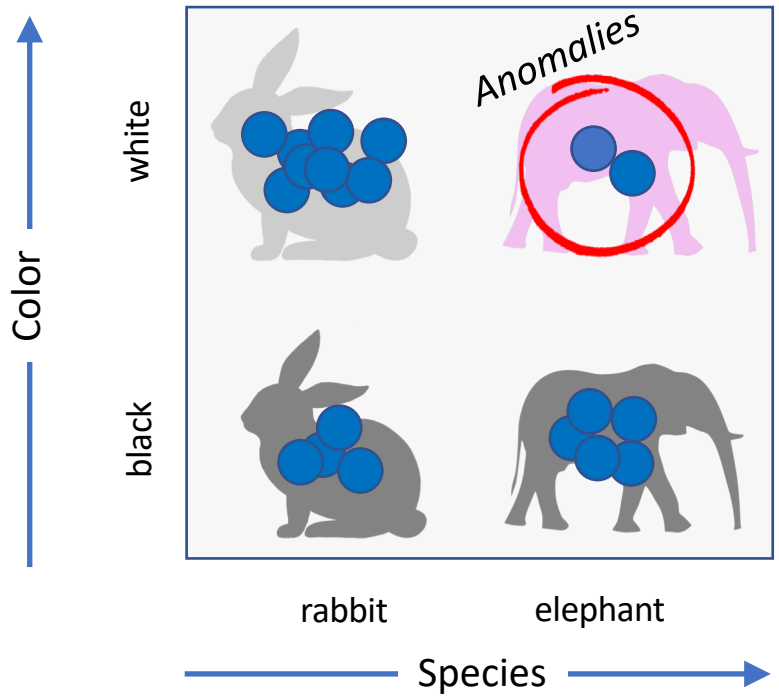
# Anomaly Score from Partial Identification (PIDScore)

Idea: Anomalies are points lying in relatively sparse regions of space.



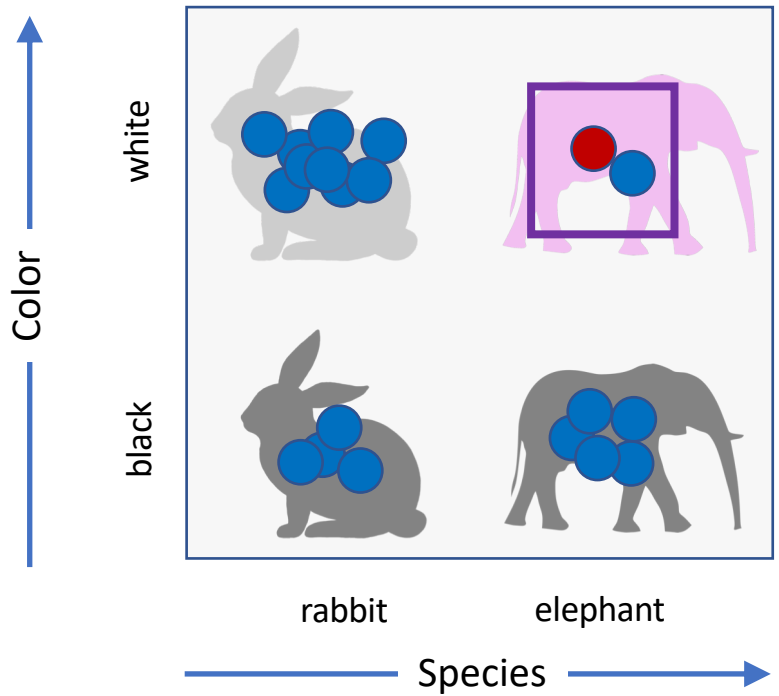
# Anomaly Score from Partial Identification (PIDScore)

Idea: Anomalies are points lying in relatively sparse regions of space.



# Anomaly Score from Partial Identification (PIDScore)

PIDScore: Look for sparse subcubes containing the point

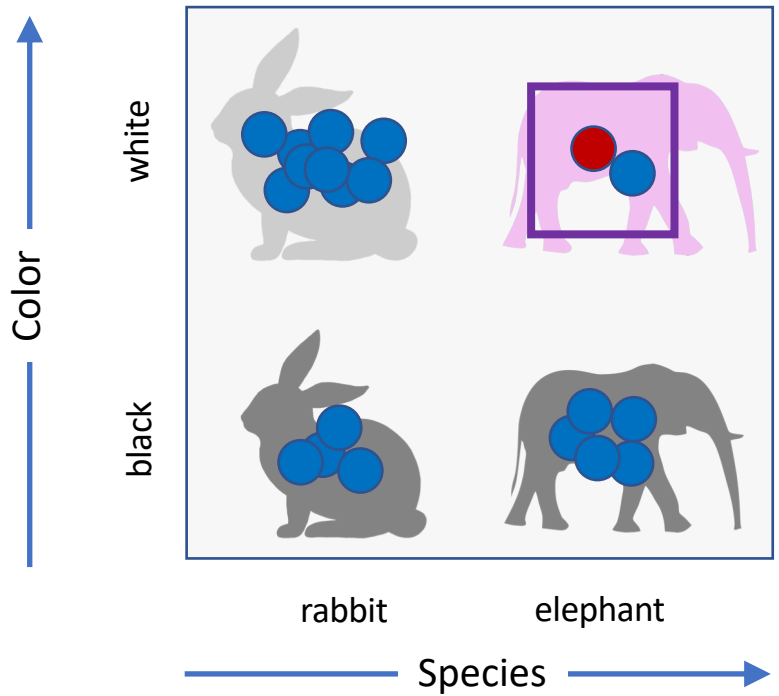



Sparsity of a subcube

$$C = \frac{Vol(C)}{\#data\ points\ in\ C}$$

# Anomaly Score from Partial Identification (PIDScore)

For any $x$, $PIDScore(x)$ is maximum sparsity over all subcubes containing $x$.

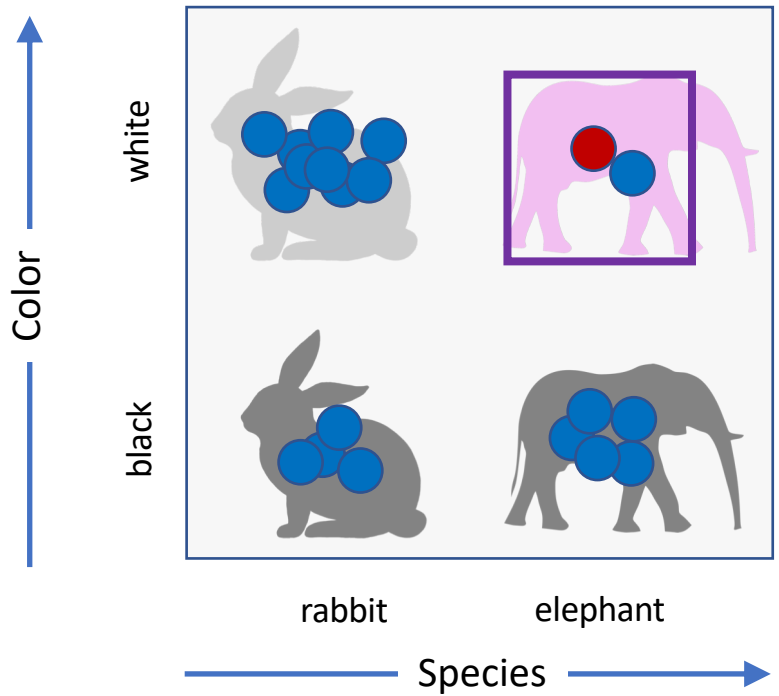



Sparsity of a subcube

$$C = \frac{Vol(C)}{\#data\ points\ in\ C}$$

# Anomaly Score from Partial Identification (PIDScore)

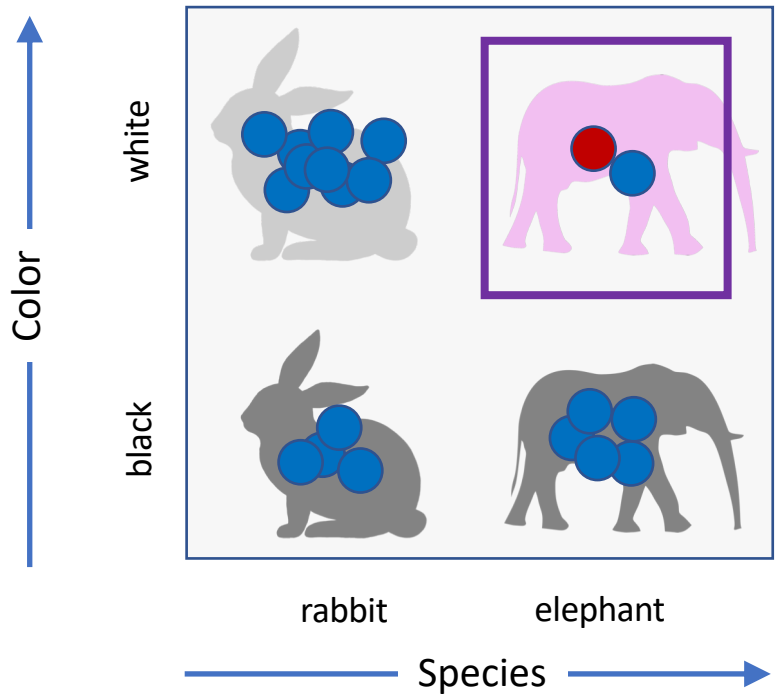
For any $x$, $PIDScore(x)$ is maximum sparsity over all subcubes containing $x$.




Sparsity of a subcube $C = \frac{Vol(C)}{\#data\ points\ in\ C}$

# Anomaly Score from Partial Identification (PIDScore)

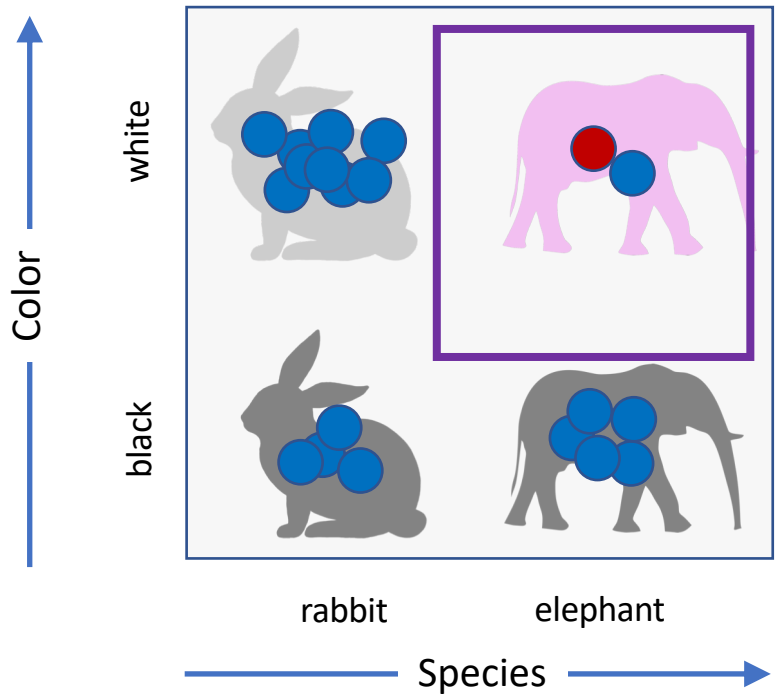
For any $x$, $PIDScore(x)$ is maximum sparsity over all subcubes containing $x$.




Sparsity of a subcube $C = \frac{Vol(C)}{\#data\ points\ in\ C}$

# Anomaly Score from Partial Identification (PIDScore)

For any $x$, $PIDScore(x)$ is maximum sparsity over all subcubes containing $x$.

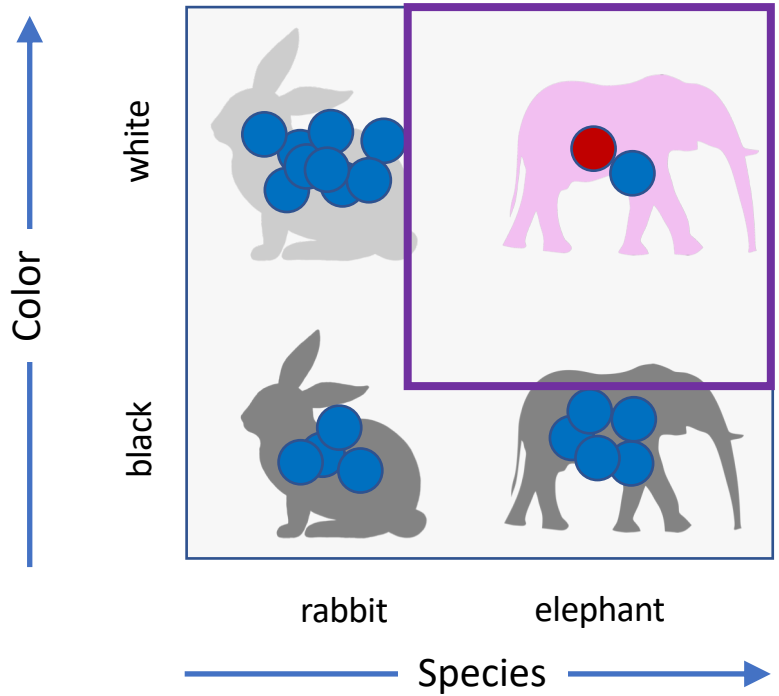



Sparsity of a subcube

$$C = \frac{Vol(C)}{\#data\ points\ in\ C}$$

# Anomaly Score from Partial Identification (PIDScore)

For any $x$, $PIDScore(x)$ is maximum sparsity over all subcubes containing $x$.

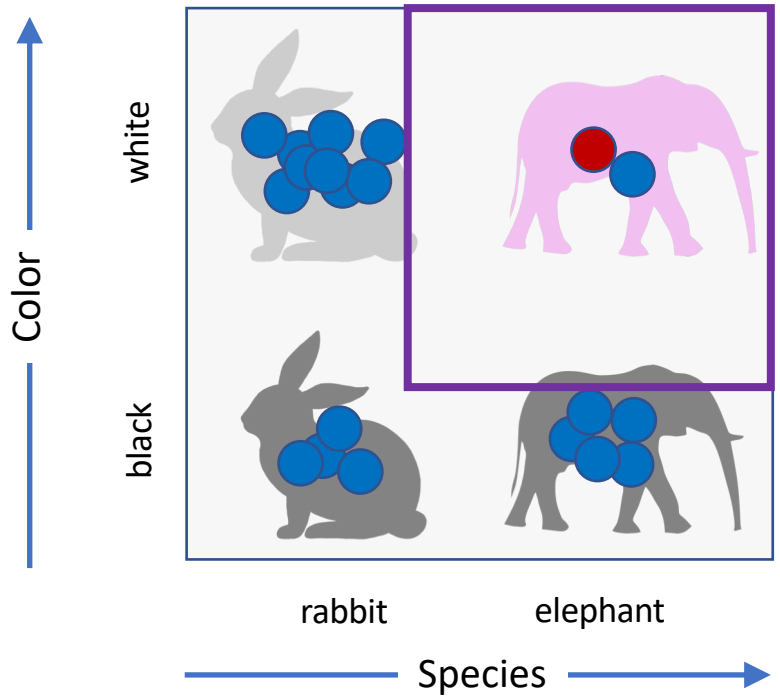



Sparsity of a subcube

$$C = \frac{Vol(C)}{\#data\ points\ in\ C}$$

# Anomaly Score from Partial Identification (PIDScore)

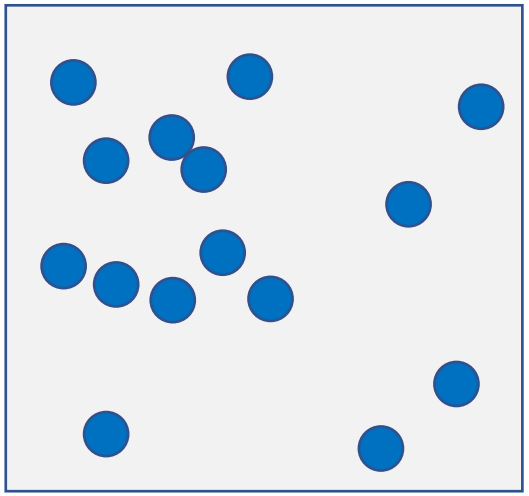
$\log(\ PIDScore(x)\ ) =$ cost of Partial Identification




Sparsity of a subcube

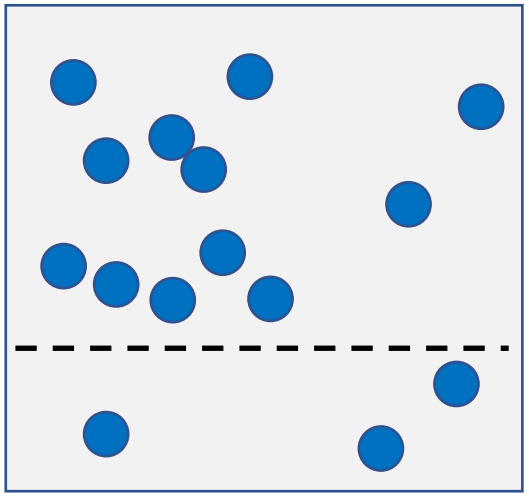
$C = \frac{Vol(C)}{\#data\ points\ in\ C}$

# PIDForest: an anomaly detection algorithm



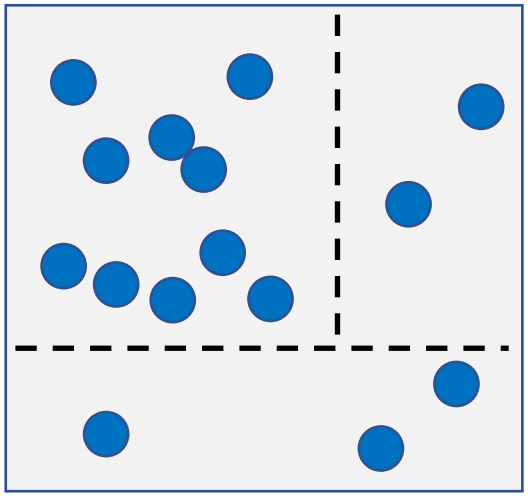
**Random forest based algorithm to partition space into sparse and dense regions**

# PIDForest: an anomaly detection algorithm



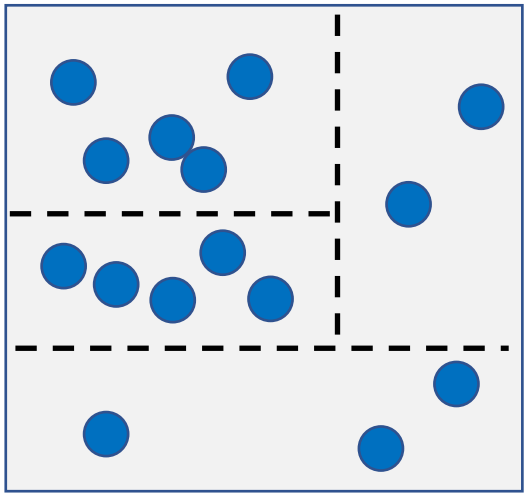
**Random forest based algorithm to partition space into sparse and dense regions**

# PIDForest: an anomaly detection algorithm



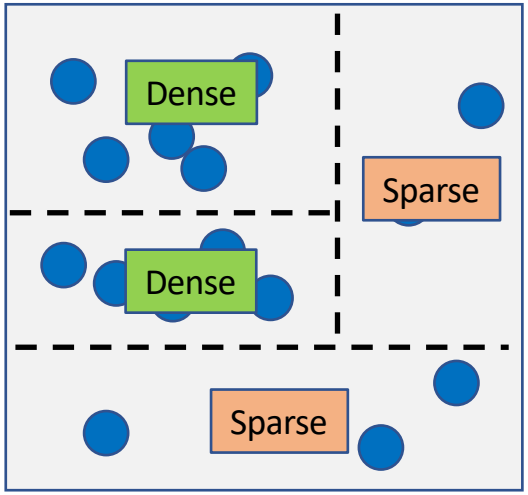
**Random forest based algorithm to partition space into sparse and dense regions**

# PIDForest: an anomaly detection algorithm

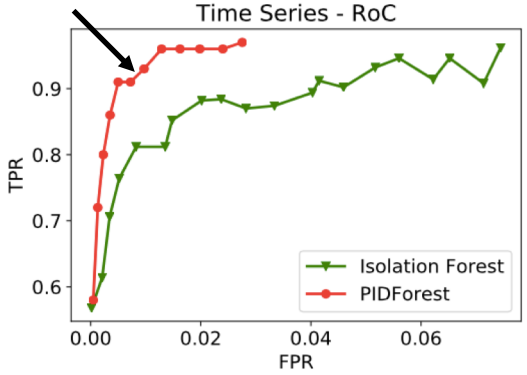


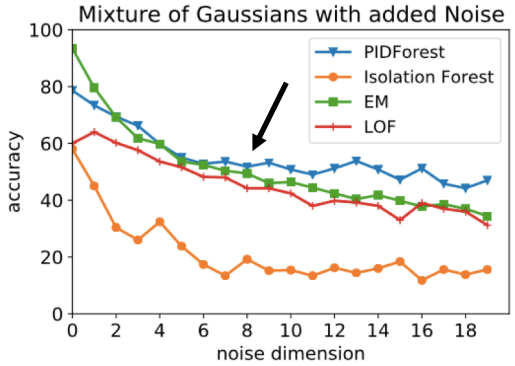
**Random forest based algorithm to partition space into sparse and dense regions**

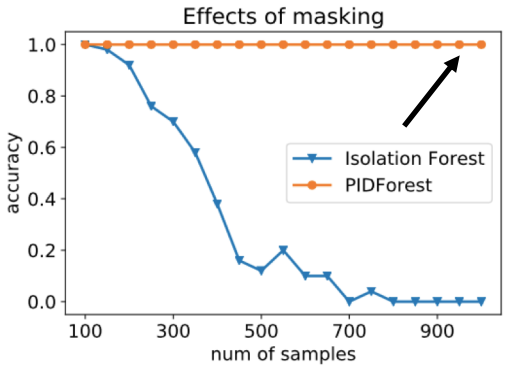
# PIDForest: an anomaly detection algorithm




**Random forest based algorithm to partition space into sparse and dense regions**




| Data set | PIDForest | iForest | RRCF | LOF | SVM | kNN | PCA |
|---|---|---|---|---|---|---|---|
| Thyroid | **0.876 ± 0.013** | 0.819 ± 0.013 | 0.739± 0.004 | 0.737 | 0.547 | 0.751 | 0.673 |
| Mammo. | 0.840 ± 0.010 | 0.862 ± 0.008 | 0.830 ± 0.002 | 0.720 | 0.872 | 0.839 | **0.886** |
| Siesmic | **0.733 ± 0.006** | 0.698 ± 0.004 | 0.701 ± 0.004 | 0.553 | 0.601 | **0.740** | 0.682 |
| Satimage | 0.987 ± 0.001 | **0.994 ± 0.001** | **0.991 ± 0.002** | 0.540 | 0.421 | 0.936 | 0.977 |
| Vowels | 0.741 ± 0.008 | 0.736 ± 0.026 | 0.813± 0.007 | 0.943 | 0.778 | **0.975** | 0.606 |
| Musk | **1.000 ± 0.000** | **0.998 ± 0.003** | 0.998 ± 0.000 | 0.416 | 0.573 | 0.373 | **1.000** |
| http | 0.986 ± 0.004 | **1.000 ± 0.000** | 0.993 ± 0.000 | 0.353 | 0.994 | 0.231 | 0.996 |
| smtp | **0.923 ± 0.003** | 0.908 ± 0.003 | 0.886 ± 0.017 | 0.905 | 0.841 | 0.895 | 0.823 |
| NYC | 0.564 ± 0.004 | 0.550 ± 0.005 | 0.543 ± 0.004 | 0.671 | 0.500 | **0.697** | 0.511 |
| A.T. | **0.810** ± 0.005 | 0.780 ± 0.006 | 0.695 ±0.004 | 0.563 | 0.670 | 0.634 | 0.792 |
| CPU | **0.935 ± 0.003** | 0.917 ± 0.002 | 0.785 ± 0.002 | 0.560 | 0.794 | 0.724 | 0.858 |
| M.T. | 0.813 ± 0.006 | 0.828 ± 0.002 | 0.7524 ± 0.003 | 0.501 | 0.796 | 0.759 | **0.834** |

Table 1: Results on real-world datasets. We bold the algorithm(s) which get the best AUC.

*Robust to noise.*
*Robust to irrelevant attributes.*
*Robust to choice of hyperparameters.*

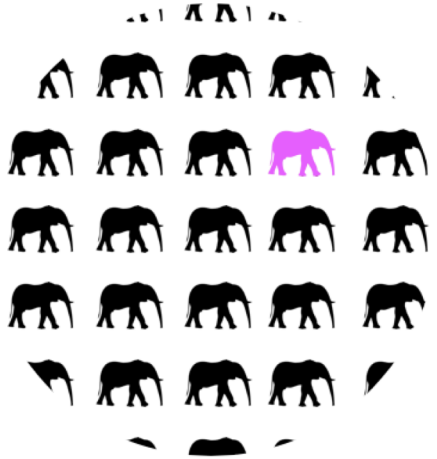
https://github.com/vatsalsharan/pidforest

# Summary

Partial Identification: Rigorous framework for anomaly detection, with minimal assumptions.

PIDScore: Definition of anomaly score based on Partial Identification.

PIDForest: Random forest-based anomaly detection.



- Outperforms commonly used anomaly detection algorithms.
- Code available to try out: https://github.com/vatsalsharan/pidforest

## Poster #60, now!! -- 12:45 PM

pgopalan@vmware.com, vsharan@stanford.edu, uwieder@vmware.com