Isolation Forest for Anomaly Detection

Sahand Hariri
PhD Student, MechSE UIUC

Matias Carrasco Kind
Senior Research Scientist, NCSA

LSST Workshop 2018, June 21, NCSA, UIUC

Overview

Goal: Build a resilient scalable anomaly detection service.

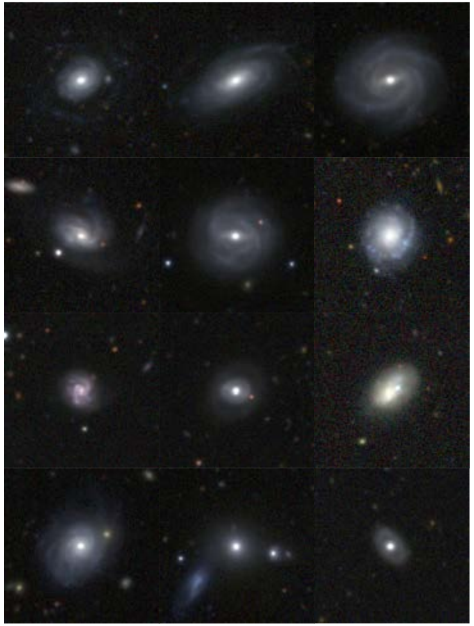
Motivation: Astronomical data (both literal and figurative)

Algorithm: Extended Isolation Forest

Infrastructure: Kubernetes cluster

Mapreduce package: Spark

Part of the Motivation



Astronomy is just one example where data exploration needs to be automated.

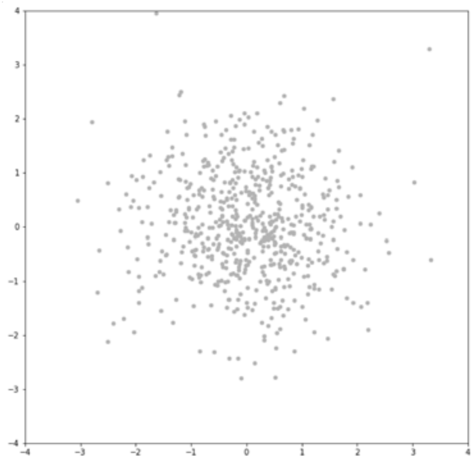
Large catalogs, Large number of images, many unexpected objects/problems → <u>Anomaly detection</u>

DARK ENERGY SURVEY

LSST
Large Synoptic Survey Telescope

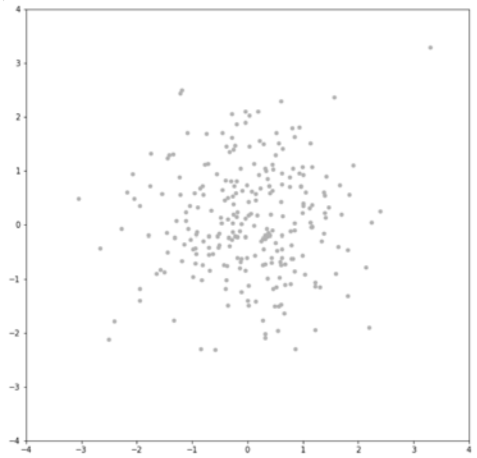
Isolation Forest (Liu et al. 2008 IEEE on Data Mining)

- Few and different to be isolated quicker



Isolation Forest

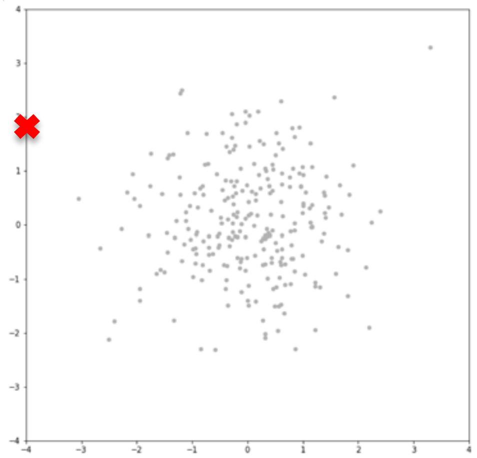
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data



Isolation Forest

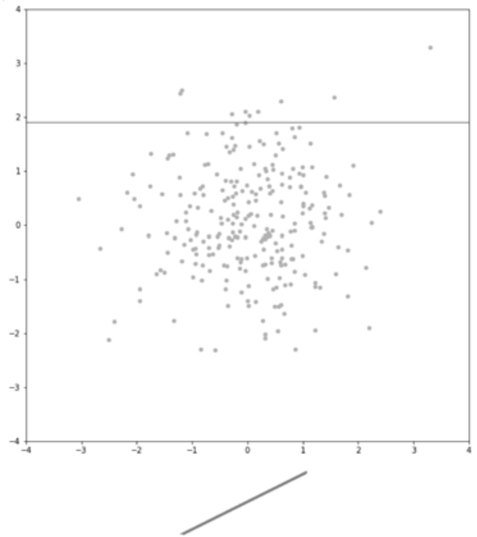
- Few and different to be isolated quicker
- For each tree:
    - Get a sample of the data
    - Randomly select a dimension
    - Randomly pick a value in that dimension

y



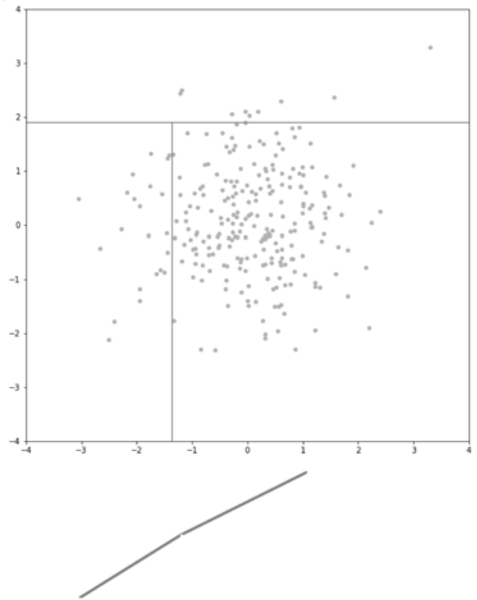
Isolation Forest

- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a dimension
  - Randomly pick a value in that dimension
  - Draw a straight line through the data at that value and split data



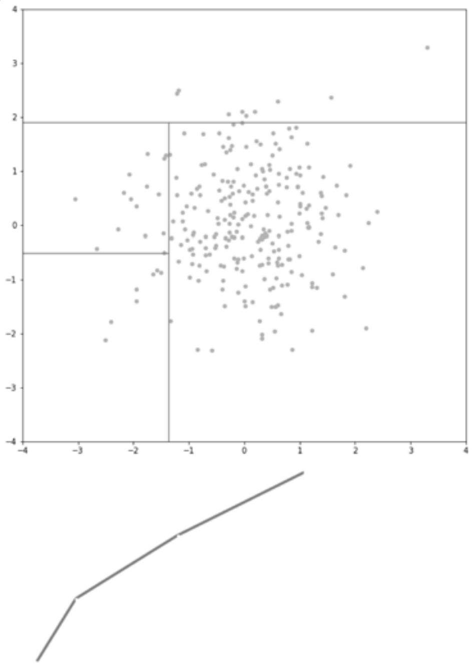
Isolation Forest

- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a dimension
  - Randomly pick a value in that dimension
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete



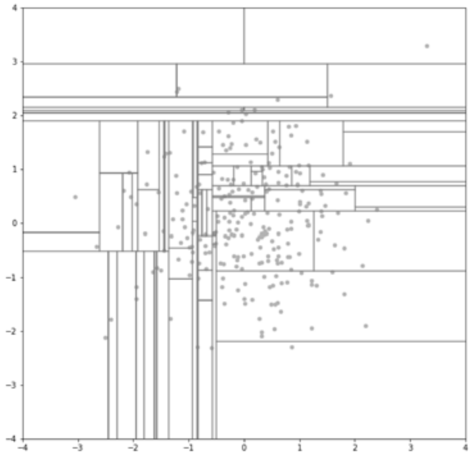
Isolation Forest

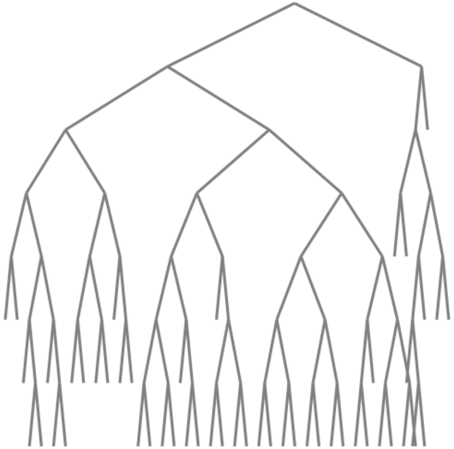
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a dimension
  - Randomly pick a value in that dimension
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete



Isolation Forest

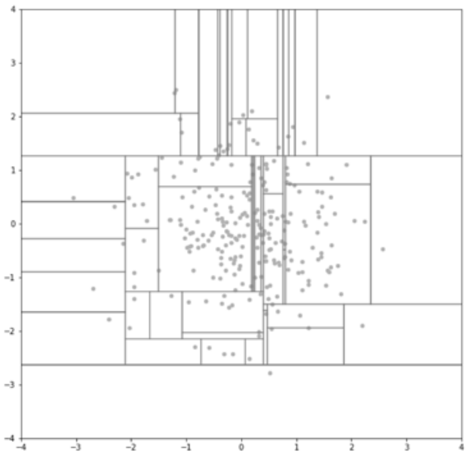
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a dimension
  - Randomly pick a value in that dimension
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete





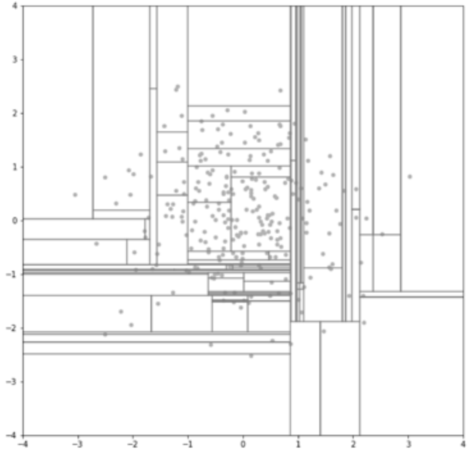
Isolation Forest

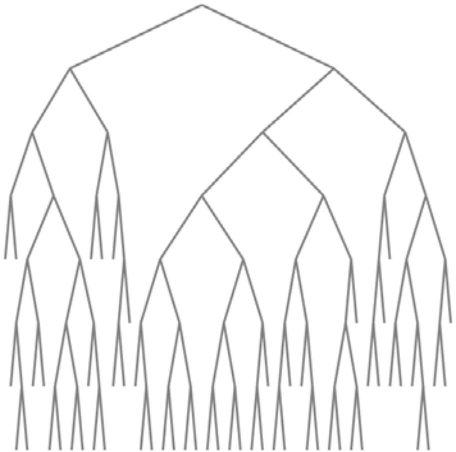
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a dimension
  - Randomly pick a value in that dimension
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete
- Generate multiple trees → forest



Isolation Forest

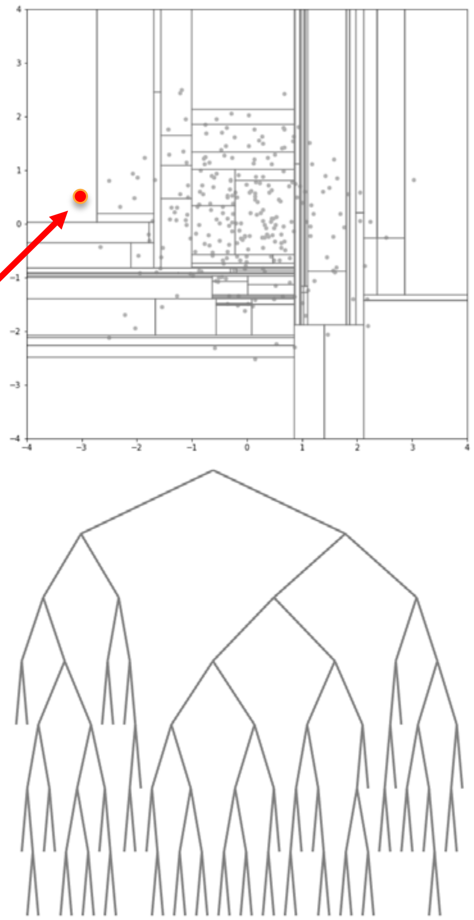
- Few and different to be isolated quicker
- For each tree:
    - Get a sample of the data
    - Randomly select a dimension
    - Randomly pick a value in that dimension
    - Draw a straight line through the data at that value and split data
    - Repeat until tree is complete
- Generate multiple trees → forest





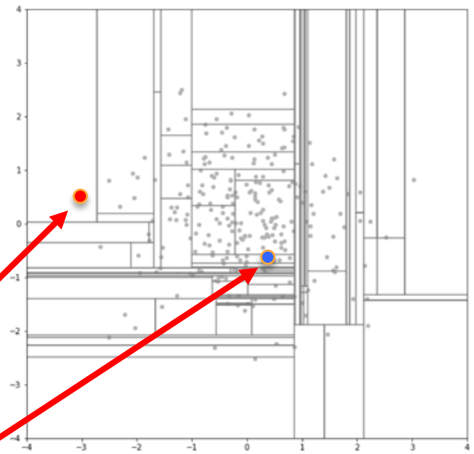
Isolation Forest

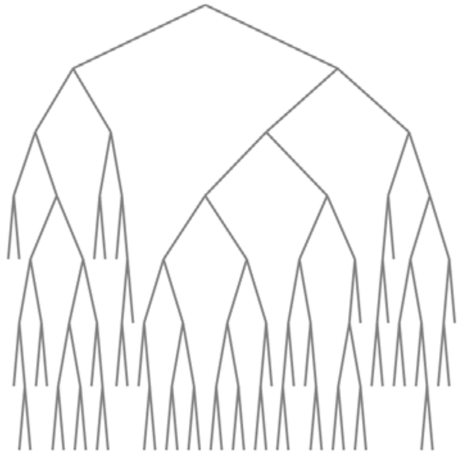
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a dimension
  - Randomly pick a value in that dimension
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete
- Generate multiple trees → forest
- Anomalies will be isolated in only a few steps



Isolation Forest

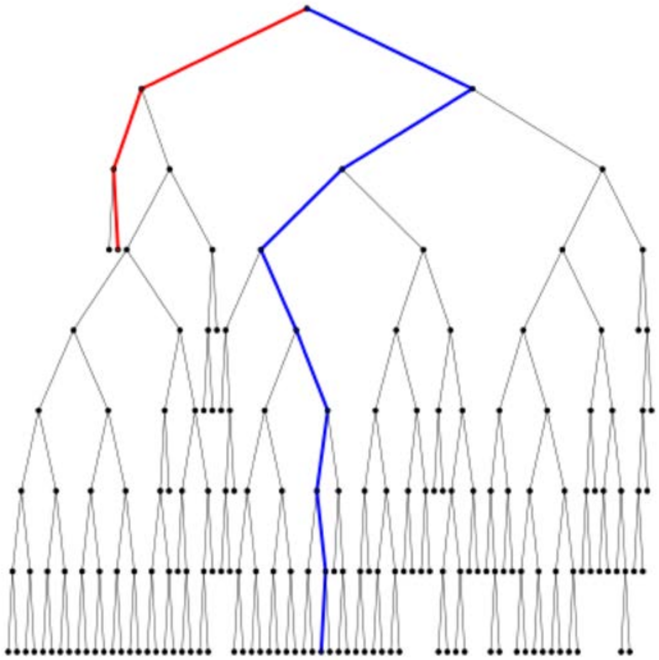
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a dimension
  - Randomly pick a value in that dimension
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete
- Generate multiple trees → forest
- Anomalies will be isolated in only a few steps
- Nominal points in more





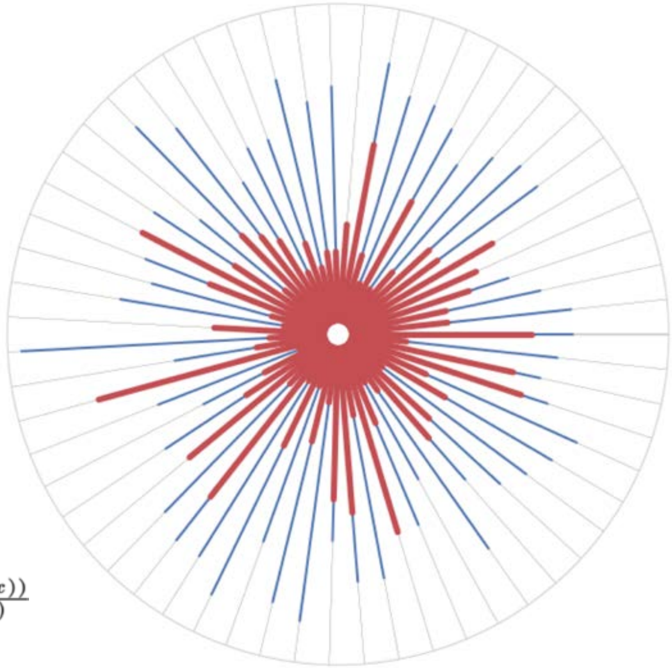
Isolation Forest

Single Tree scores for
anomaly and nominal points

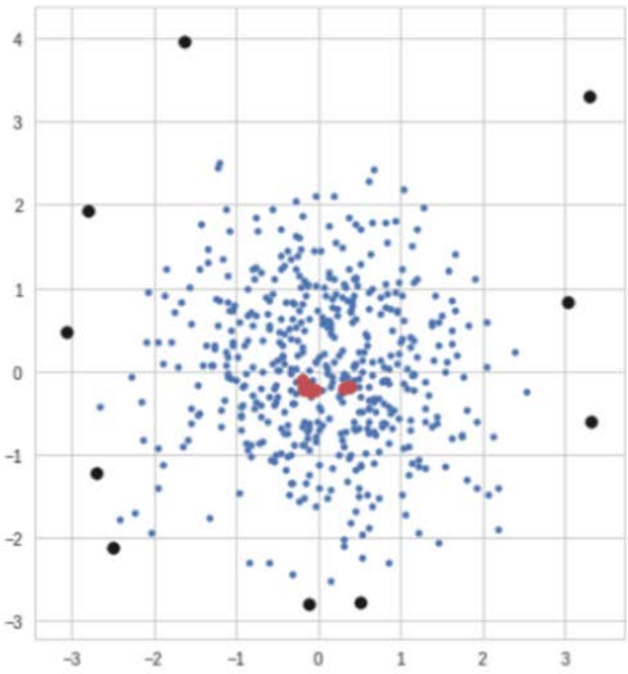


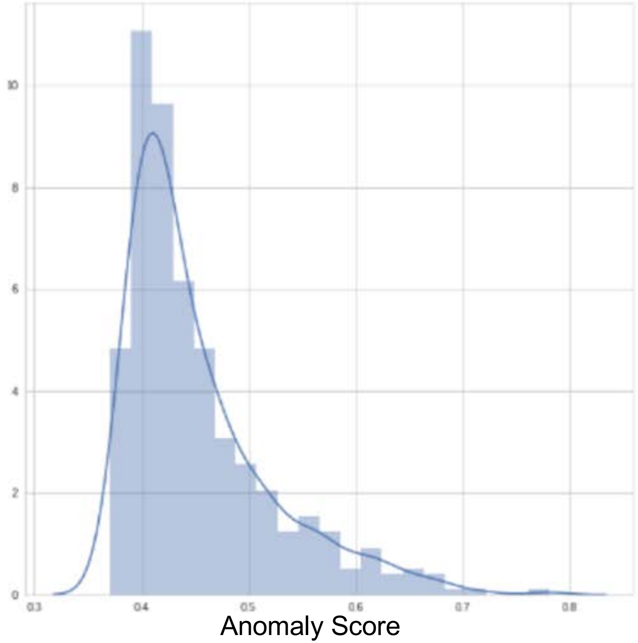
$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Forest plotted radially.
Scores for anomaly and
nominal shown as lines



Anomaly Detection with Isolation Forest





Anomaly Detection with Isolation Forest

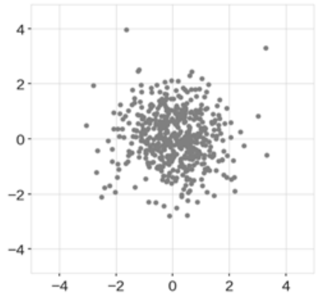
Isolation Forest:

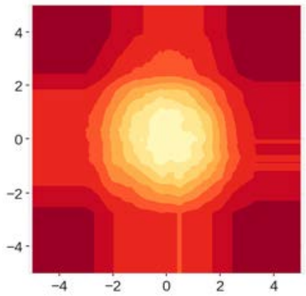
- Model free
- Computationally efficient
- Readily applicable to parallelization
- Readily applicable to high dimensional data
- Inconsistent scoring observed in score maps

Anomaly Detection with Isolation Forest

Isolation Forest:

- Model free
- Computationally efficient
- Readily applicable to parallelization
- Readily applicable to high dimensional data
- Inconsistent scoring observed in score maps

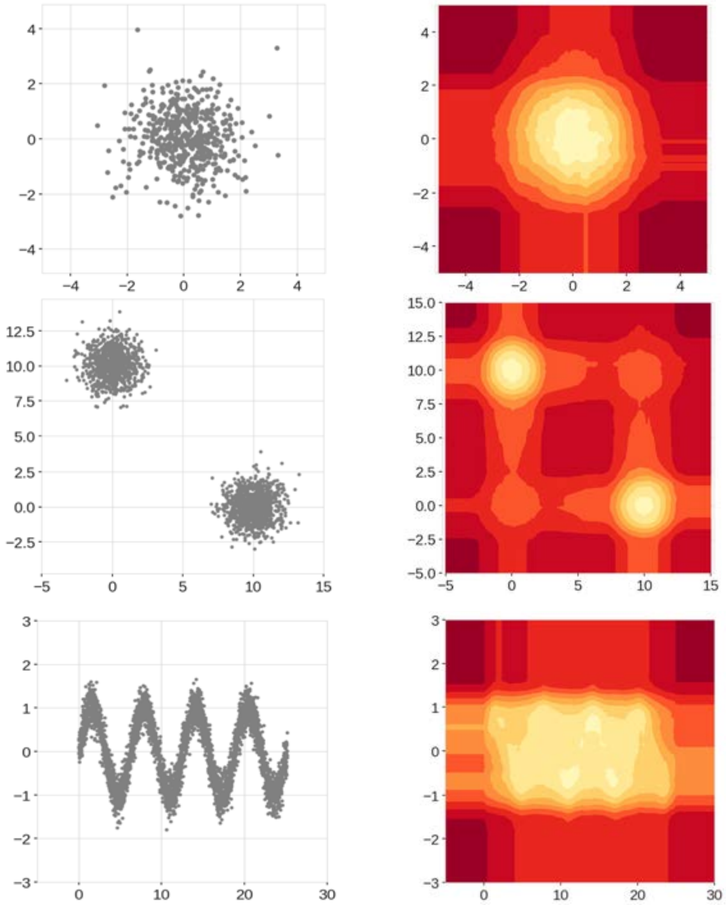




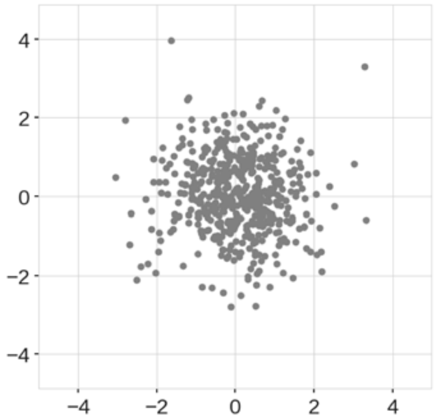
Anomaly Detection with Isolation Forest

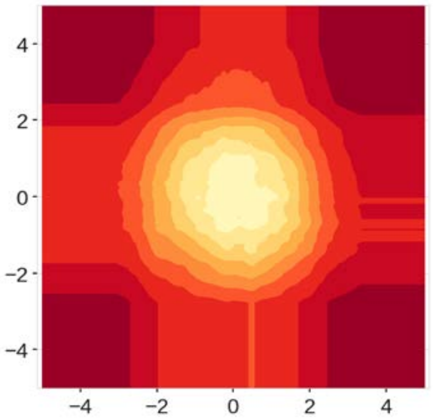
Isolation Forest:

- Model free
- Computationally efficient
- Readily applicable to parallelization
- Readily applicable to high dimensional data
- Inconsistent scoring observed in score maps

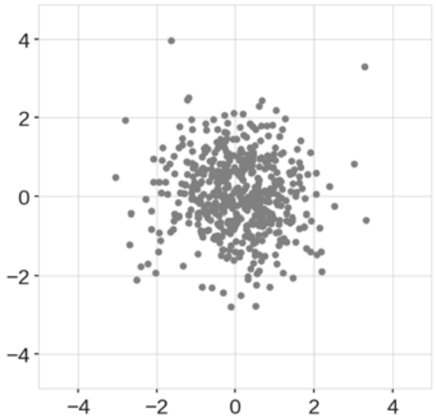


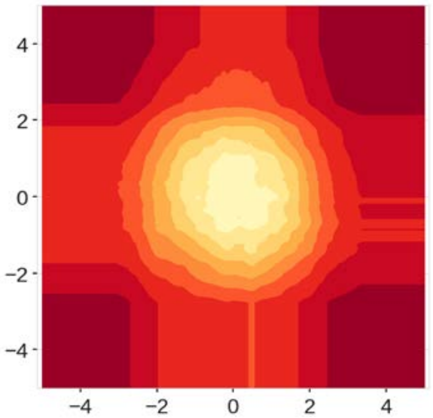
Anomaly Detection with Isolation Forest

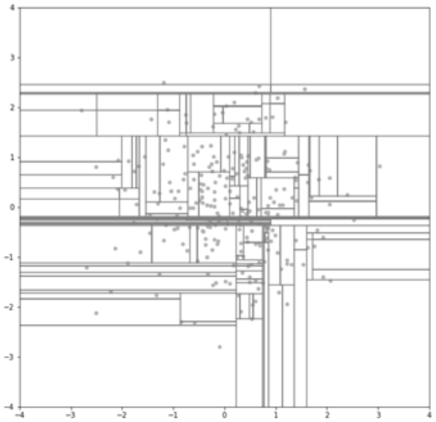




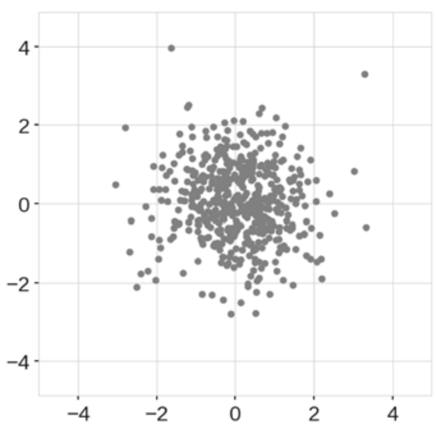
Anomaly Detection with Isolation Forest

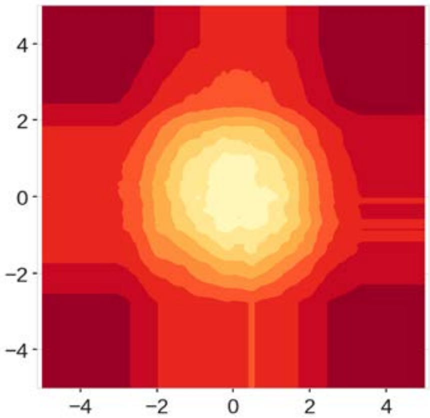


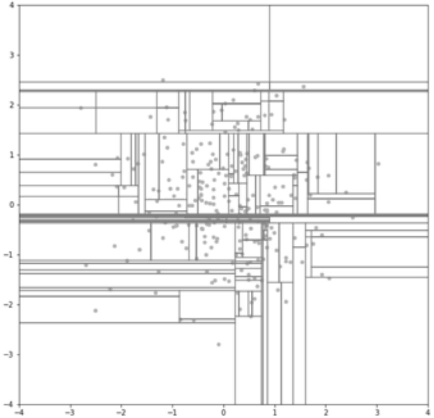


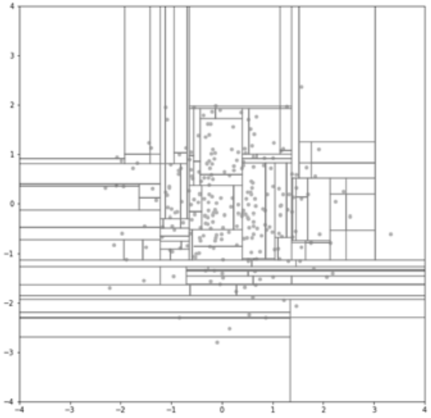


Anomaly Detection with Isolation Forest

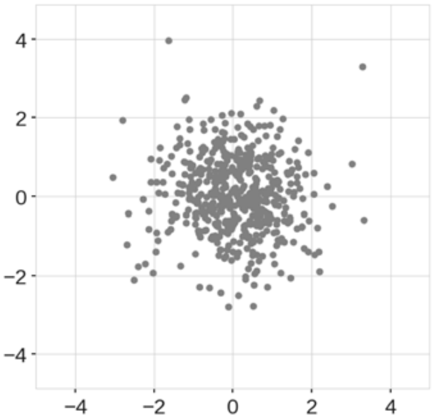


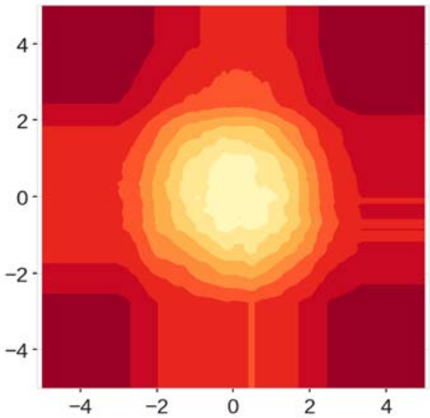


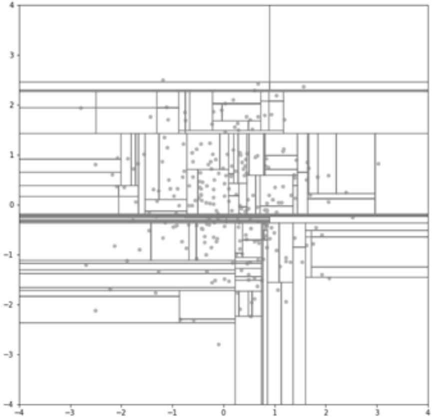


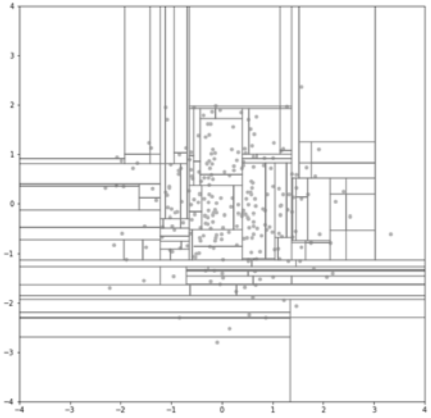


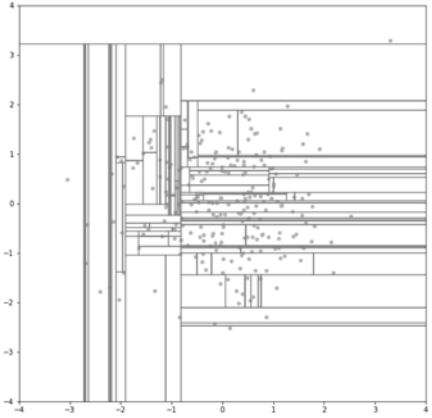
Anomaly Detection with Isolation Forest



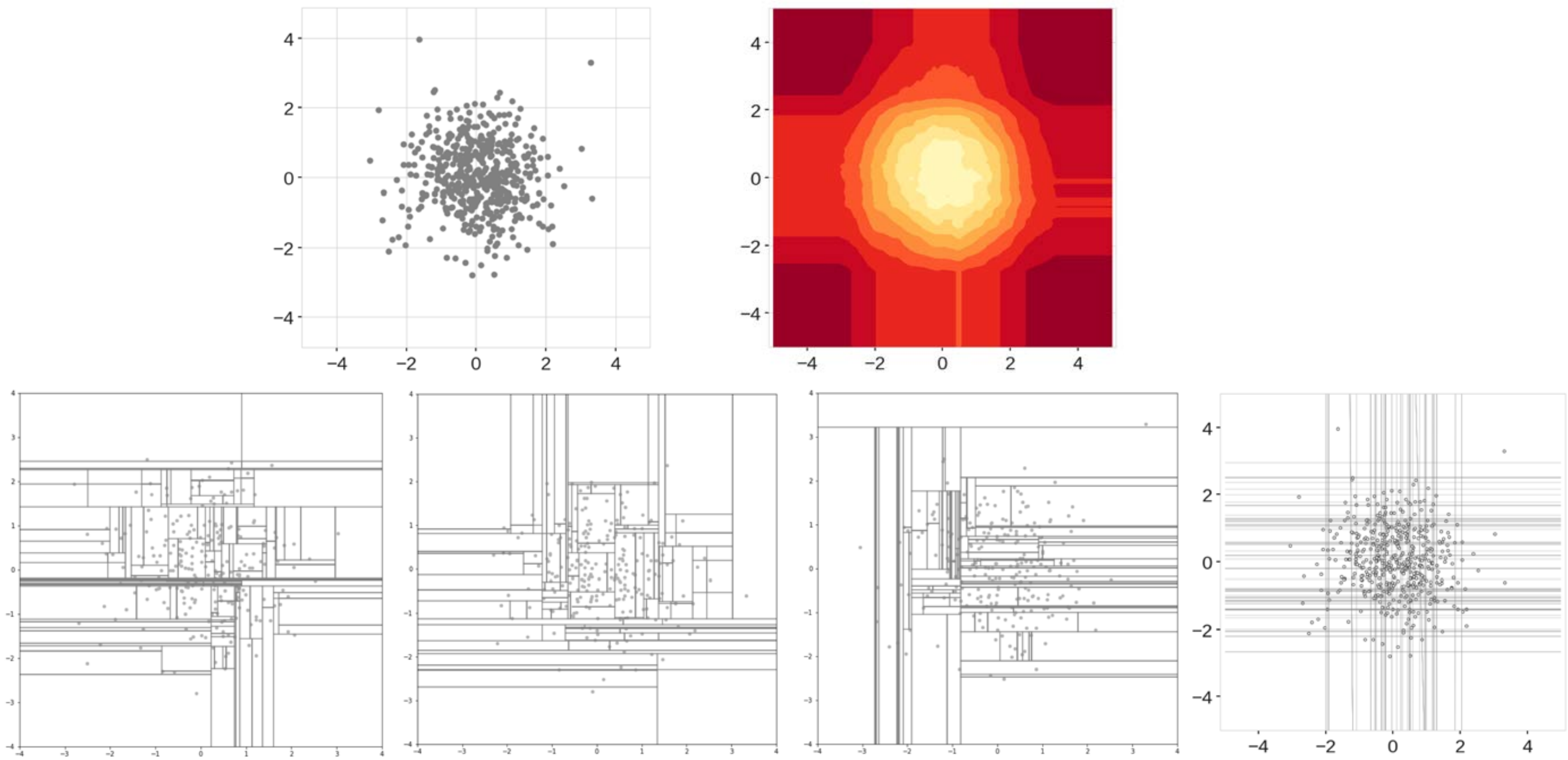






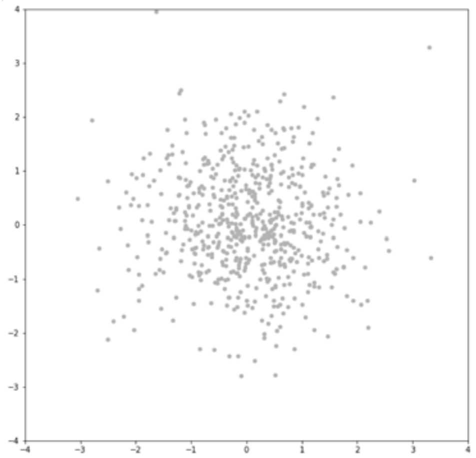


Anomaly Detection with Isolation Forest



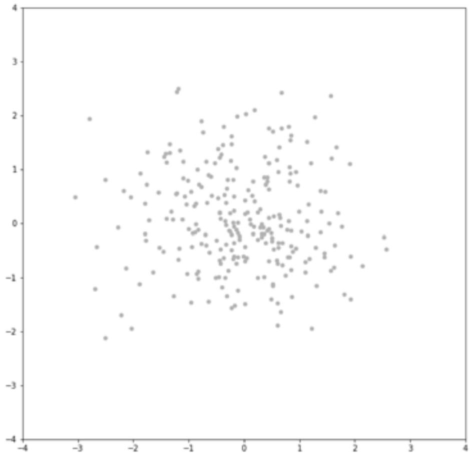
Anomaly Detection with Extended Isolation Forest

- Few and different to be isolated quicker



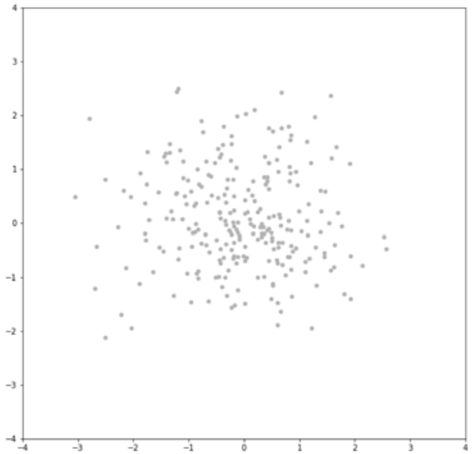
Anomaly Detection with Extended Isolation Forest

- Few and different to be isolated quicker
- For each tree:
    - Get a sample of the data



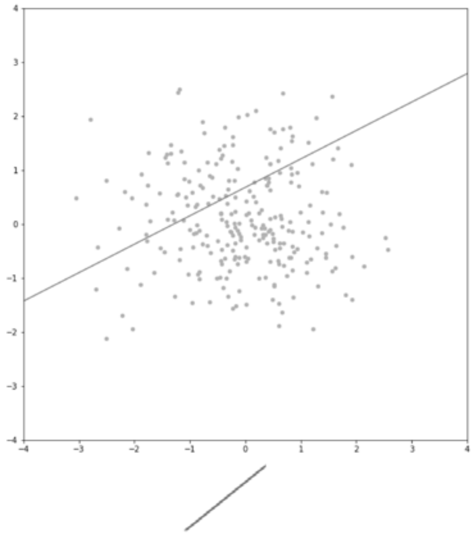
Anomaly Detection with Extended Isolation Forest

- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a normal vector
  - Randomly select an intercept



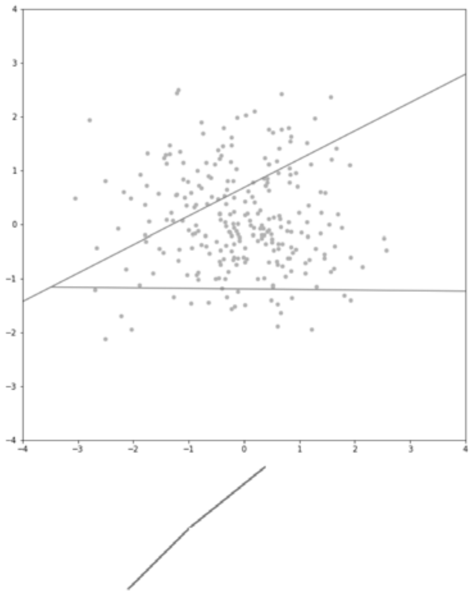
Anomaly Detection with Extended Isolation Forest

- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a normal vector
  - Randomly select an intercept
  - Draw a straight line through the data at that value and split data



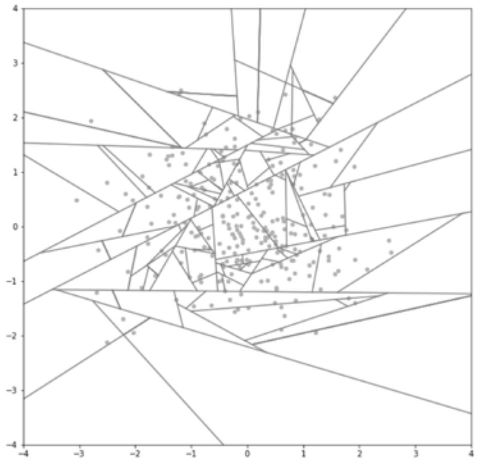
Anomaly Detection with Extended Isolation Forest

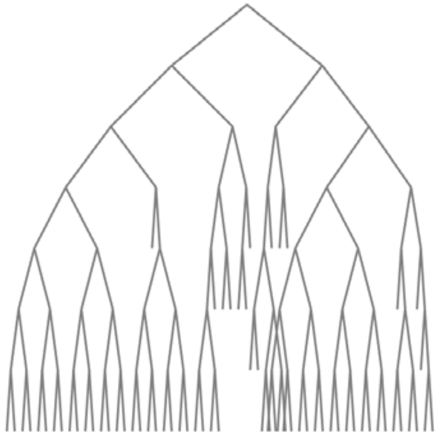
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a normal vector
  - Randomly select an intercept
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete



Anomaly Detection with Extended Isolation Forest

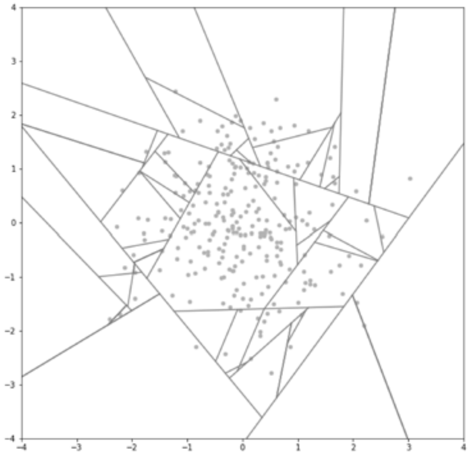
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a normal vector
  - Randomly select an intercept
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete





Anomaly Detection with Extended Isolation Forest

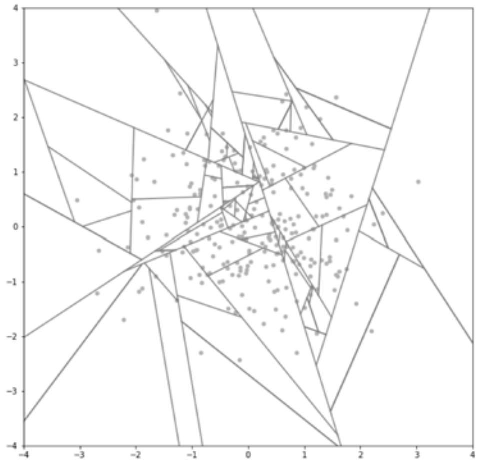
- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a normal vector
  - Randomly select an intercept
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete
- Generate multiple trees → forest

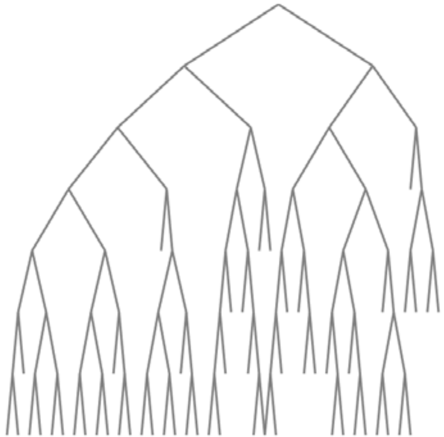


Anomaly Detection with Extended Isolation Forest

- Few and different to be isolated quicker
- For each tree:
  - Get a sample of the data
  - Randomly select a normal vector
  - Randomly select an intercept
  - Draw a straight line through the data at that value and split data
  - Repeat until tree is complete
- Generate multiple trees ⟶ forest
- No artificial extra slicing
- Same rules about scoring apply
- Checking for which side of the line the point lies:

$$\left(\vec{x} - \vec{p}\right) \cdot \vec{n} \leq 0$$





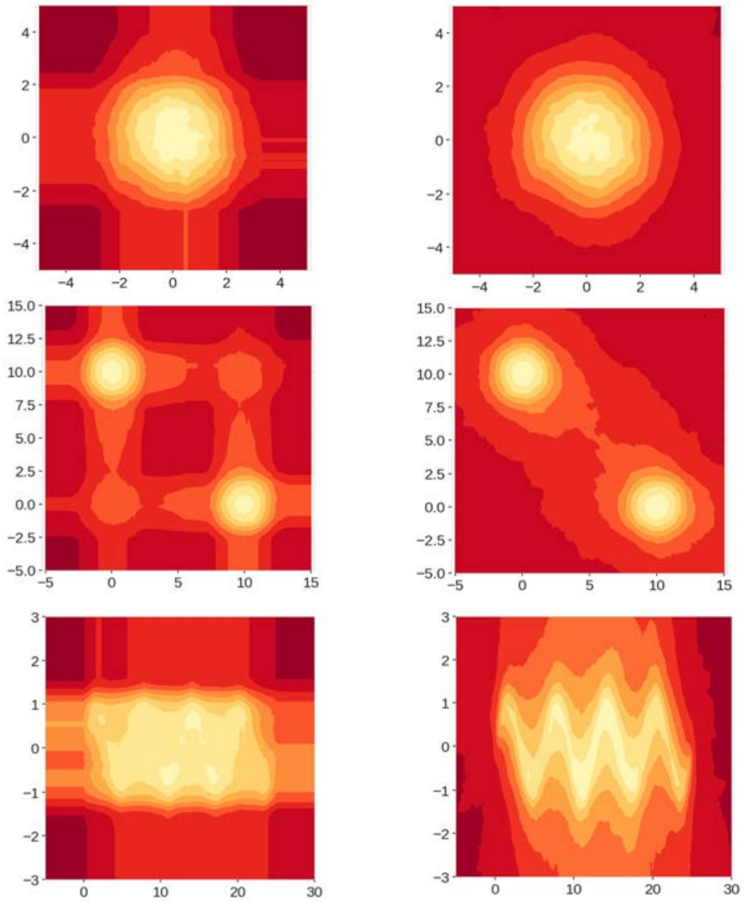
Anomaly Detection with Extended Isolation Forest

Isolation Forest:

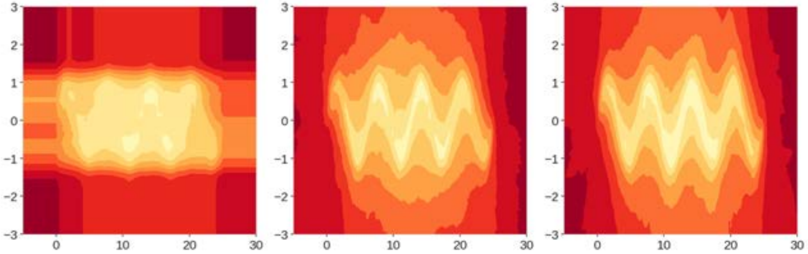
- Model free
- Computationally efficient
- Readily applicable to parallelization
- Readily applicable to high dimensional data
- Inconsistent scoring observed in score maps

Extended Isolation Forest:

- Model free
- Computationally efficient
- Readily applicable to parallelization
- Readily applicable to high dimensional data
- Consistent scoring



Anomaly Detection with Extended Isolation Forest



(a) Standard IF (b) Rotated IF (c) Extended IF

Algorithm 2 $iTree(X, e, l)$

Require: X - input data, e - current tree height, l - height limit

Ensure: an iTree

if $e \geq l$ or $|X| \leq 1$ **then**

  return $exNode\{Size \leftarrow |X|\}$

else

  randomly select a normal vector $n \in \mathbb{R}^{|X|}$ by drawing each coordinate of $\vec{n}$ from a uniform distribution.

  randomly select an intercept point $p \in \mathbb{R}^{|X|}$ in the range of X

  set coordinates of n to zero according to extension level

  $X_l \leftarrow filter(X, (X - p) \cdot n \leq 0)$

  $X_r \leftarrow filter(X, (X - p) \cdot n > 0)$

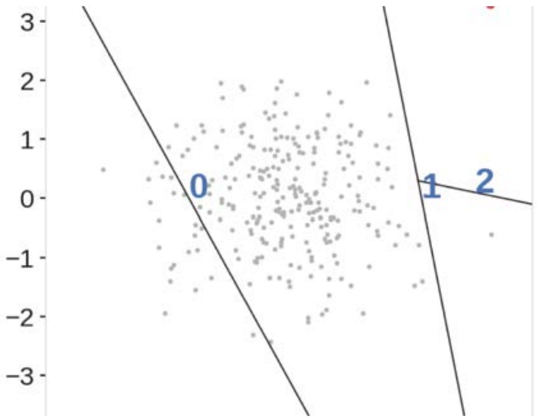
  return inNode{ $Left \leftarrow iTree(X_l, e + 1, l)$,

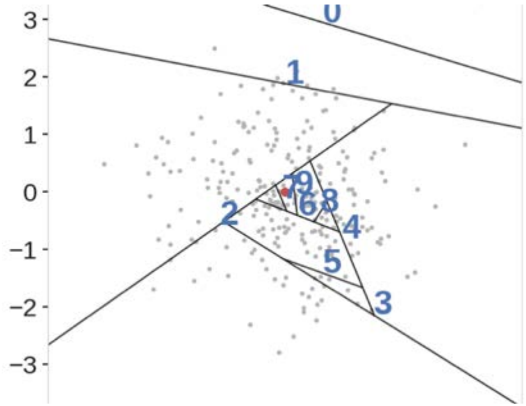
  $Right \leftarrow iTree(X_r, e + 1, l)$,

  $Normal \leftarrow n$,

  $Intercept \leftarrow p$}

end if



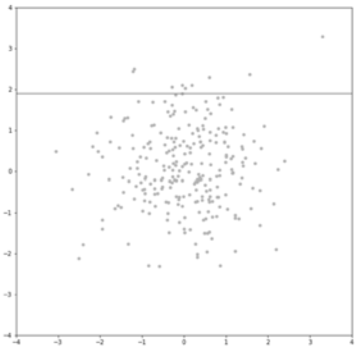


Multi-Dimensional Data

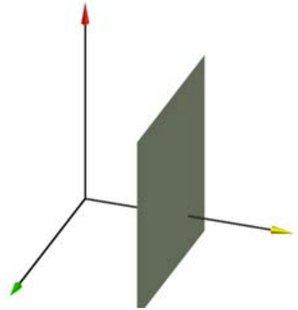
- For N dimensional data, the “line” becomes an N-1 dimensional hyperplanes

Standard Isolation Forest

2-D



3-D

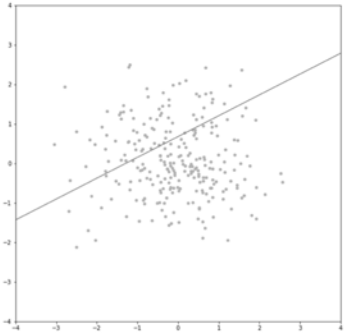


Multi-Dimensional Data

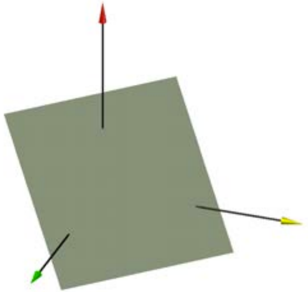
- For N dimensional data, the “line” becomes an N-1 dimensional hyperplanes

Extended Isolation Forest

2-D



3-D

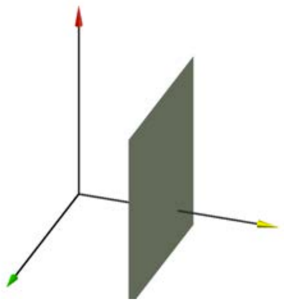


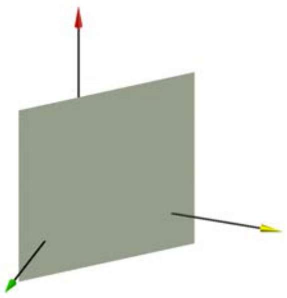
Multi-Dimensional Data

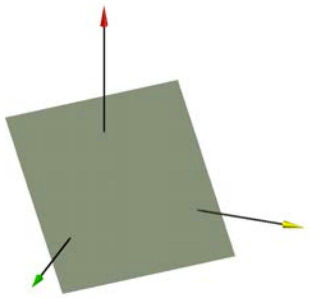
- For N dimensional data, the “line” becomes an N-1 dimensional hyperplanes
- With Extended Isolation Forest, there are extension levels

Multi-Dimensional Data

- For N dimensional data, the “line” becomes an N-1 dimensional hyperplanes
- With Extended Isolation Forest, there are extension levels
- Standard Isolation Forest is recovered
- Extended Isolation Forest is a natural generalization of the original algorithm



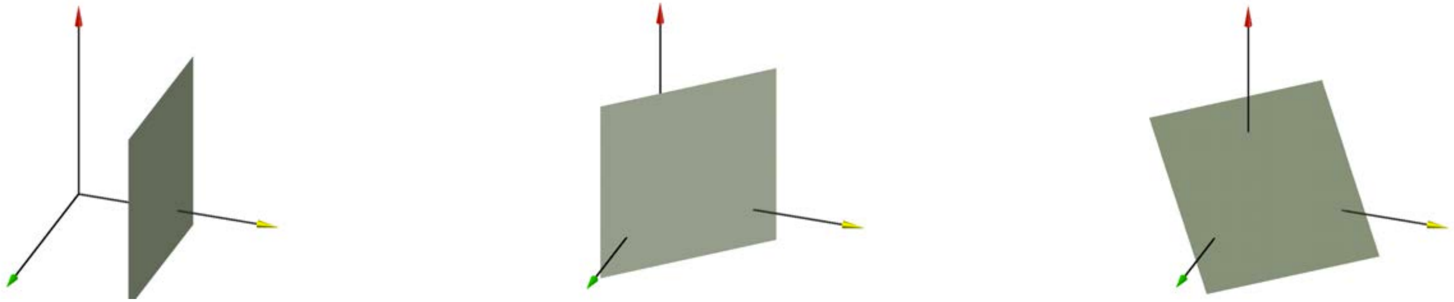




Multi-Dimensional Data

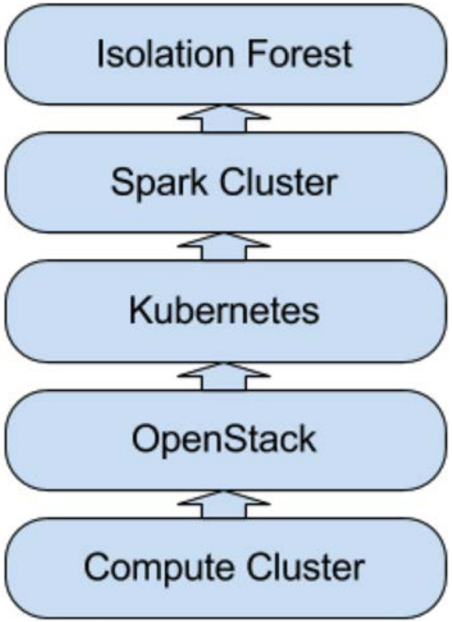
- For N dimensional data, the “line” becomes an N-1 dimensional hyperplanes
- With Extended Isolation Forest, there are extension levels
- Standard Isolation Forest is recovered
- Extended Isolation Forest is a natural generalization of the original algorithm

$$\left(\vec{x}-\vec{p}\right)\cdot\vec{n}\leq 0$$



Technology Stack For Anomaly Service

- Use Extended Isolation Forest as core algorithm
- Use Spark to parallelize trees and scoring
- Use Redis as a broker communicator
- To easily deploy in any environment, use Docker
- For orchestration of Docker containers, use Kubernetes
- Kubernetes cluster built on top of OpenStack, but it can be deployed also in AWS, GKE, etc.

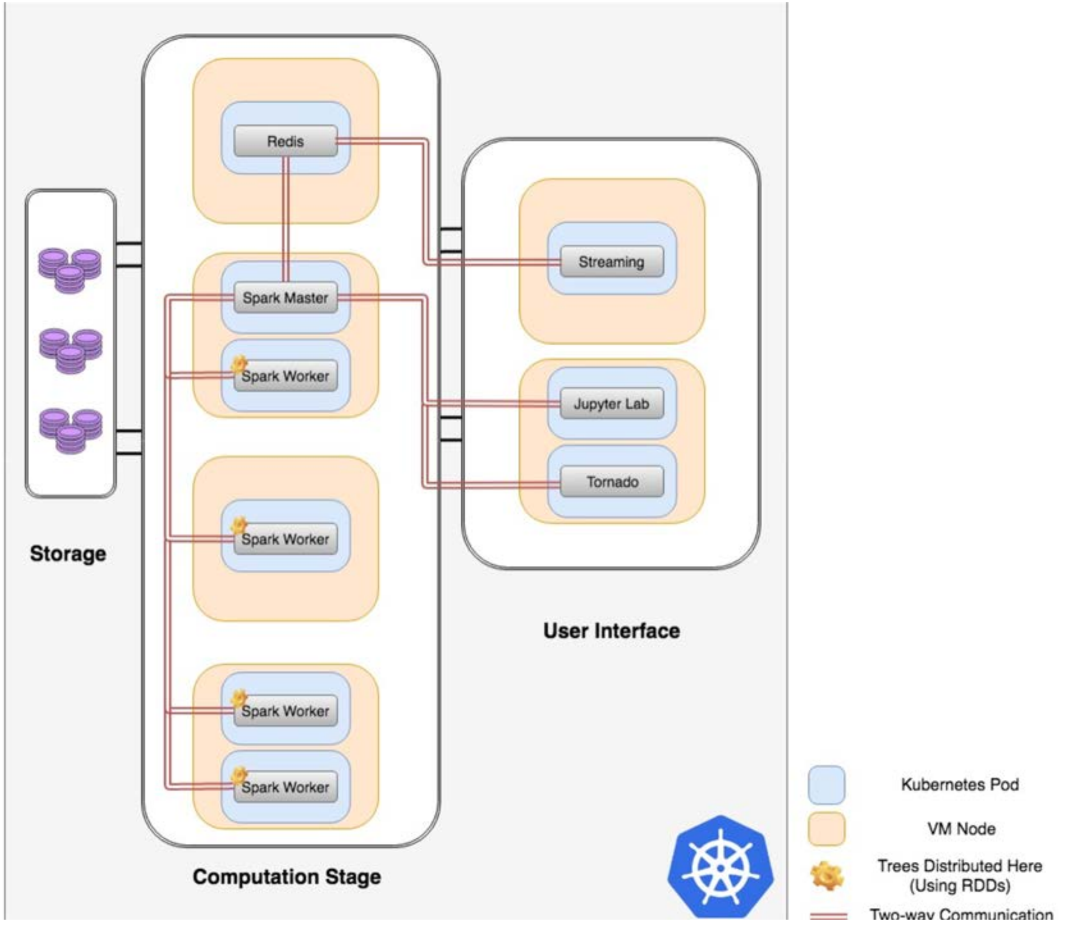


Spark openstack™

Framework Architecture

There are three main components:

1. Storage
2. Computation Stage
3. User Interface / Streaming



Framework Architecture

Storage:

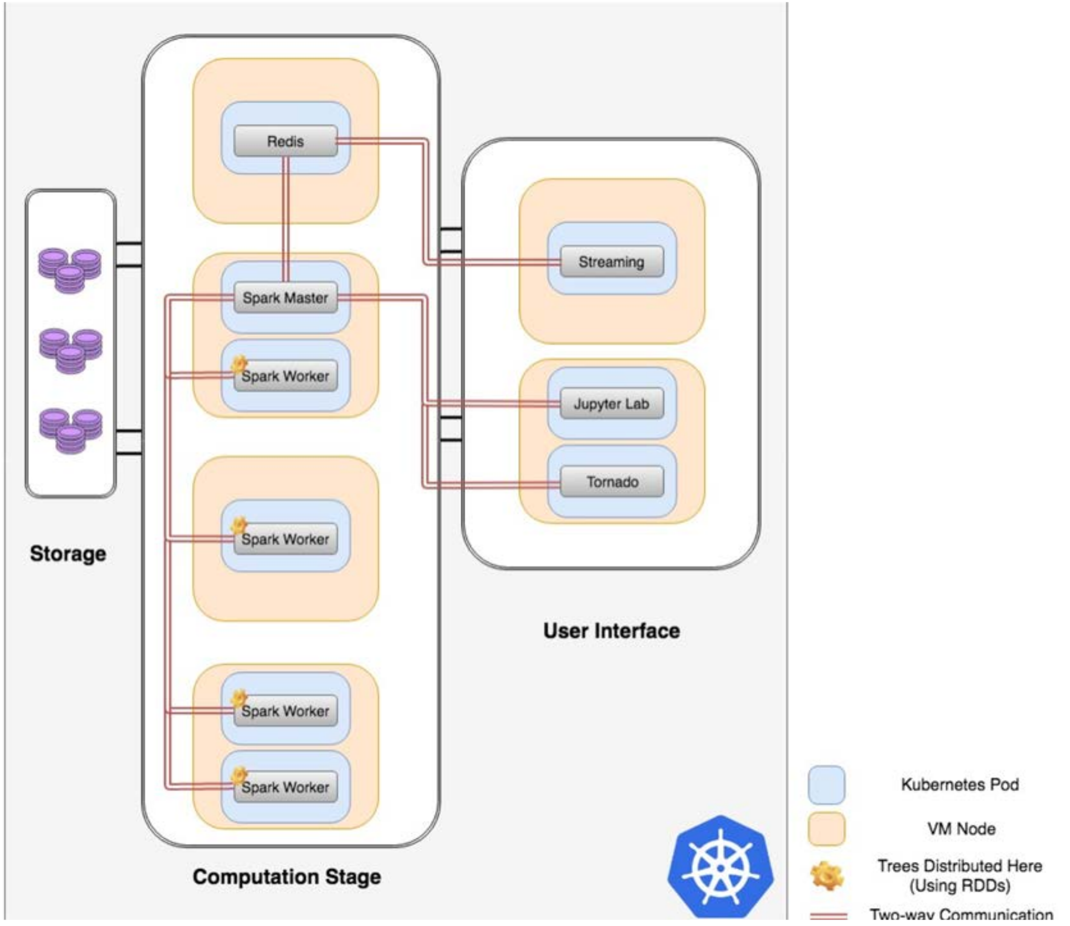
- NFS (Kubernetes PV/PVC)
- Redis
- RDD for Trees and Spark

User Interface:

- Jupyter notebooks
- Interactive web app for submitting jobs
- Streaming service

Computation Stage:

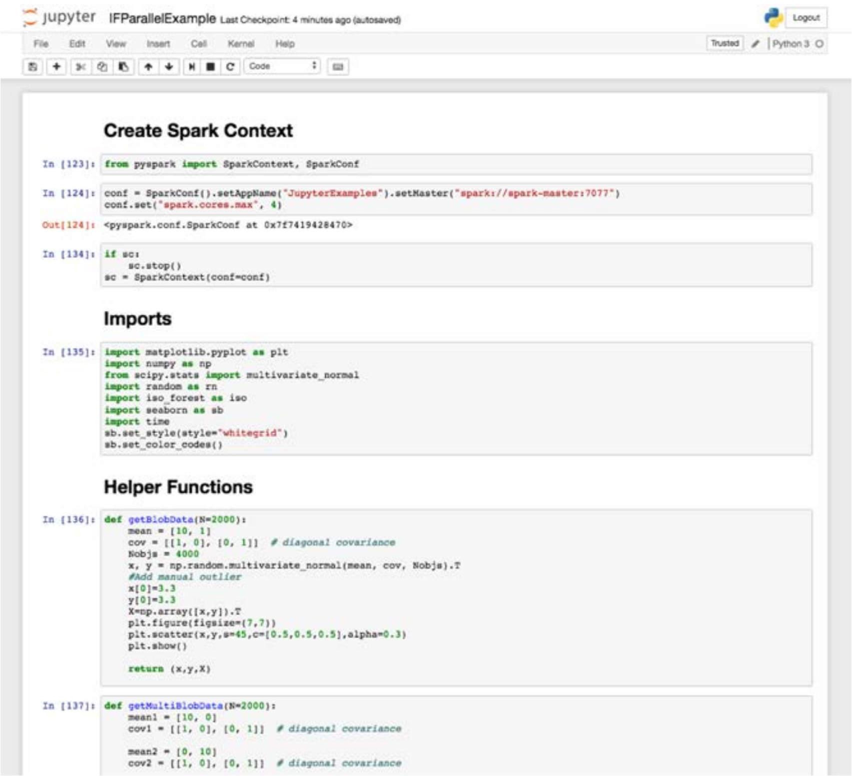
- Spark Master and Workers
- Communicator with Spark Master
- Subscription

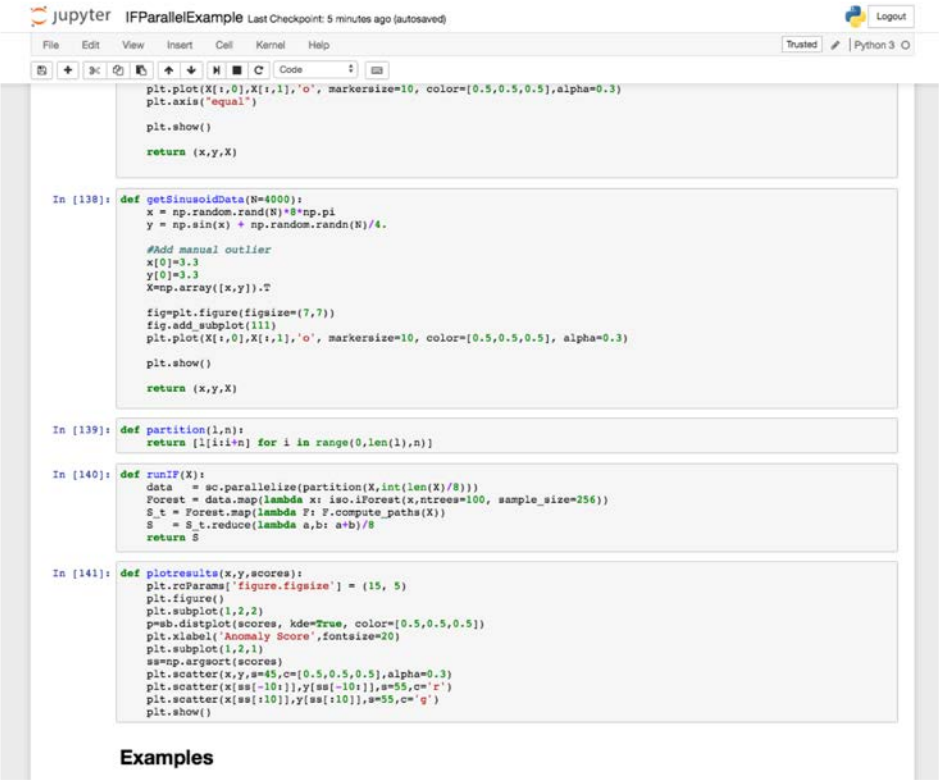


Deployment

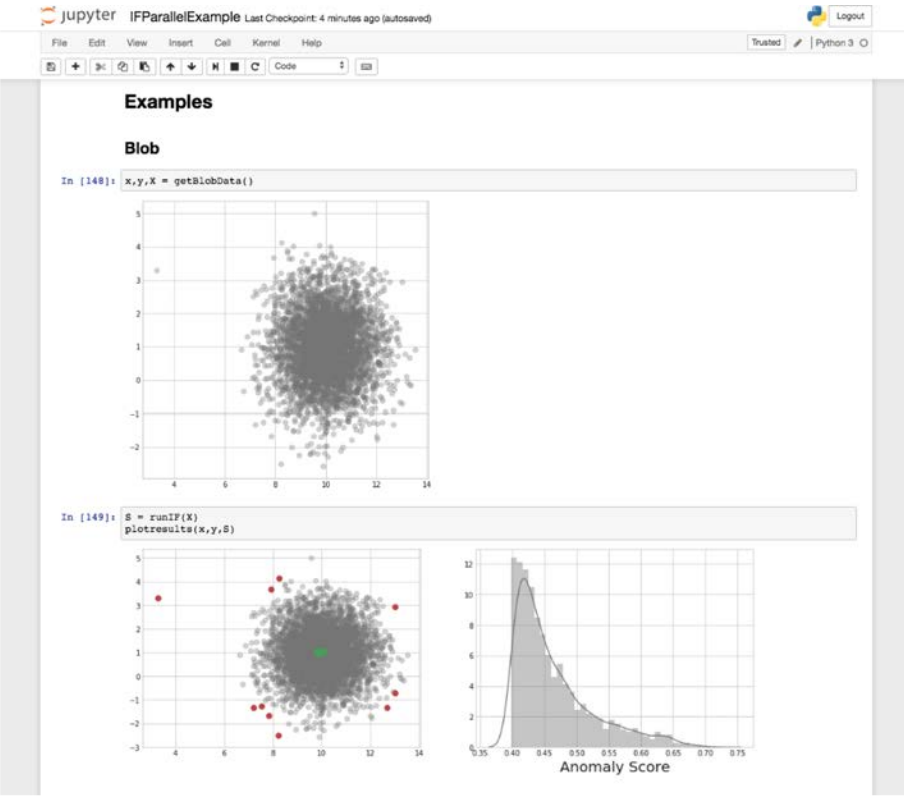
- Kubernetes allows very easy deployment, orchestration, scalability, resilience, replication, workloads and more
- Federation of services and Jobs
- From 0 to anomaly service → in minutes and config files
- Scale up/down (spark cluster and front-end) → Auto-scaling as an option
- Prototype support multiple users/projects, batch and streaming process
- Fault tolerant, disaster recovery

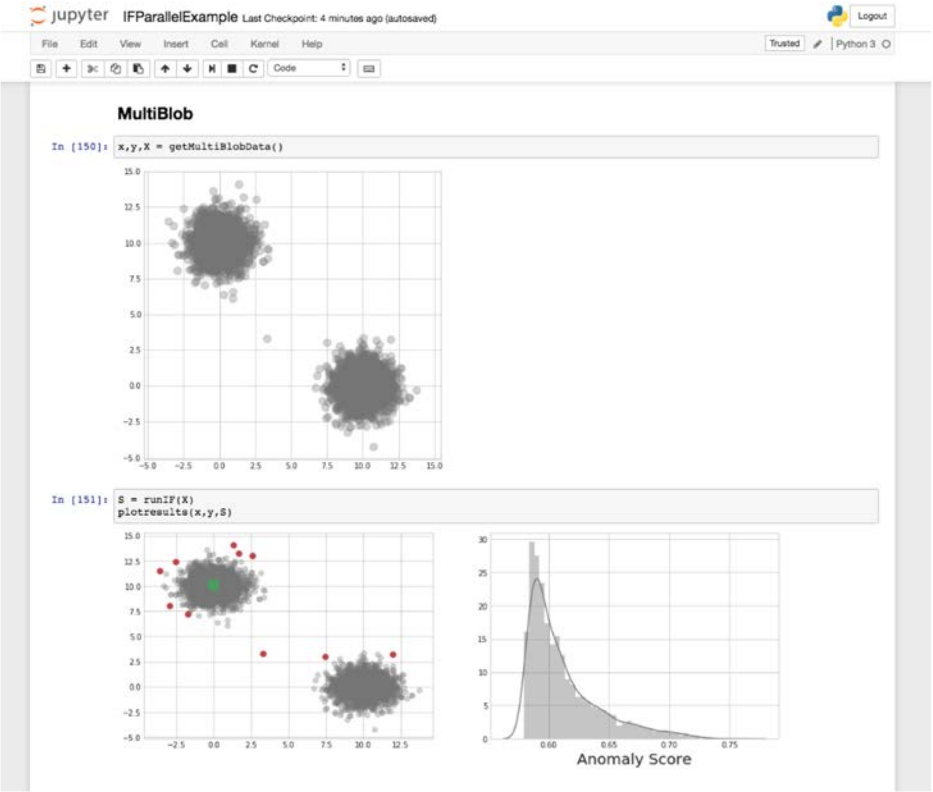
Example: Jupyter Notebooks



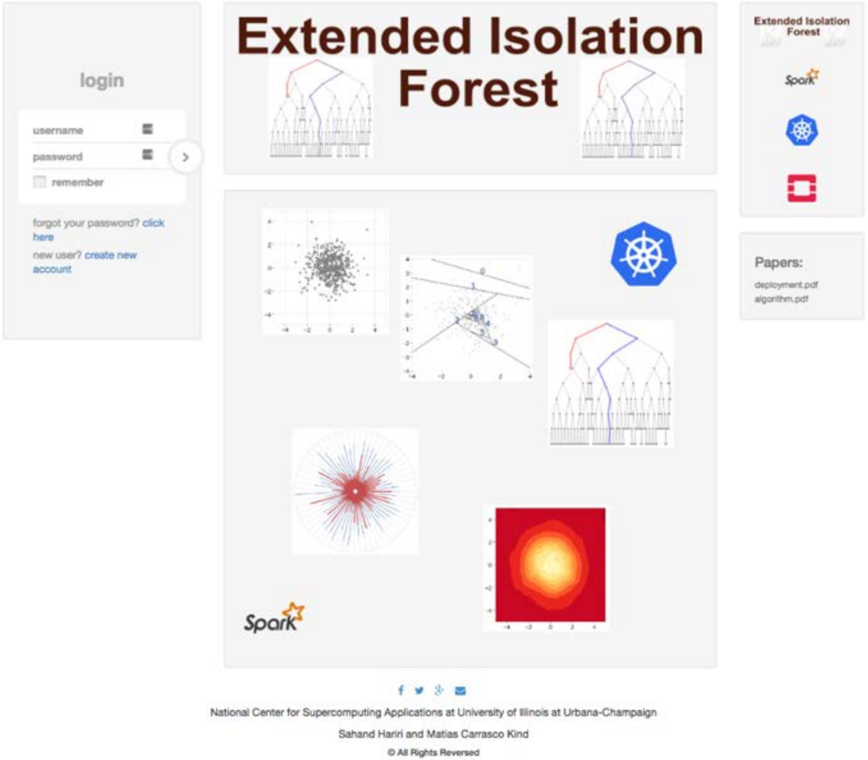


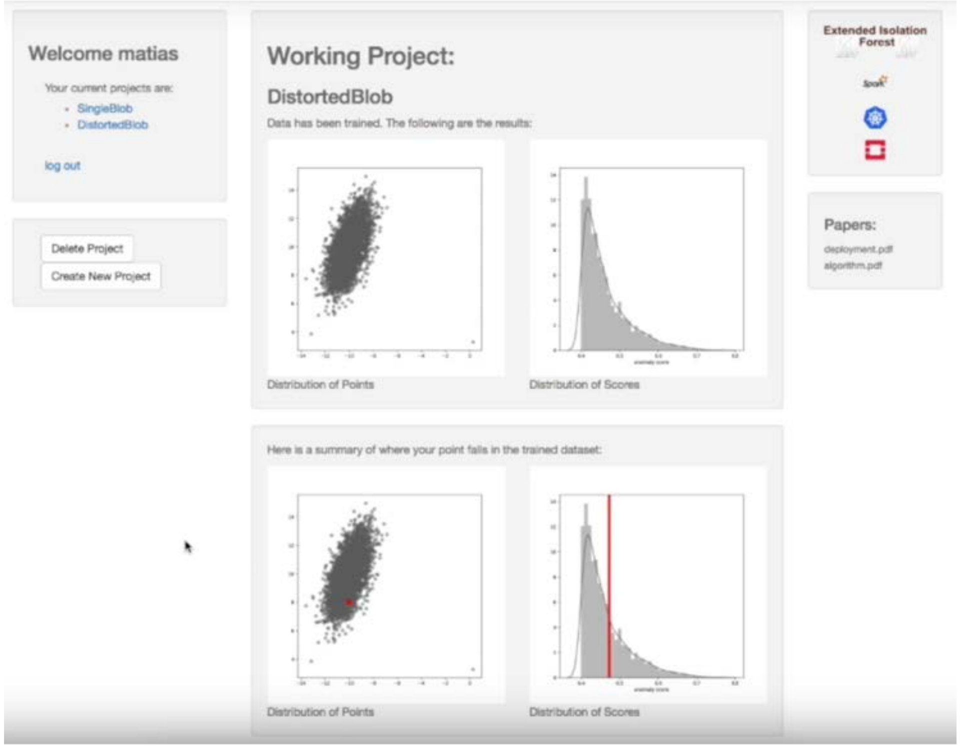
Example: Jupyter Notebooks





Examples: User interface





Conclusions

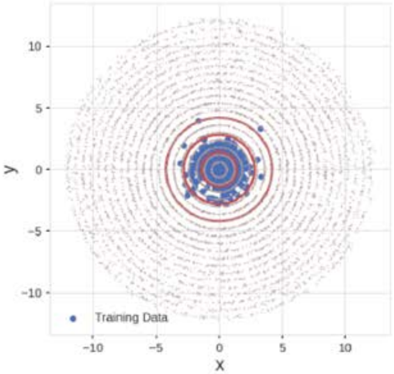
- Open source anomaly detection software package for scientific application using fast and efficient isolation forest
- Fault tolerant, robust, scalable deployment
- Train and scoring using Spark
- Ready-to-deploy infrastructure on Kubernetes
- Production services for large datasets

Thank you!

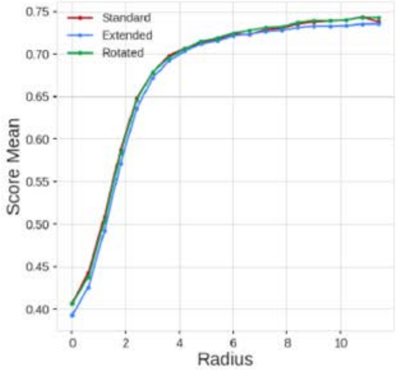
Questions?

Sahand Hariri -- NCSA
hariria2@illinois.edu
github.com/sahandha
sahandhariri.com

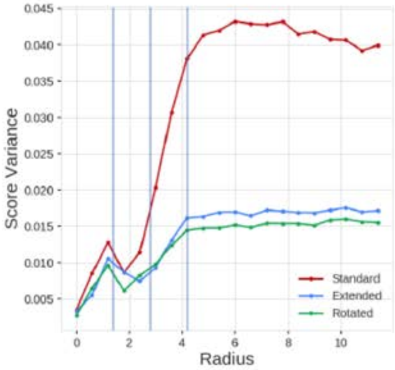
Variance



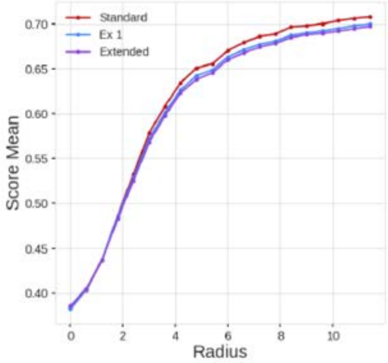
(a) Data



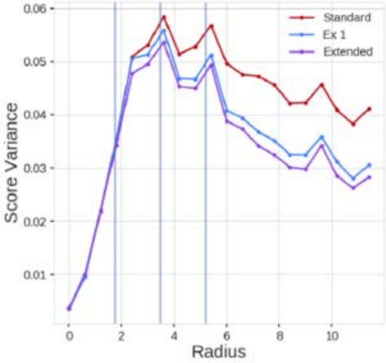
(b) Score Mean



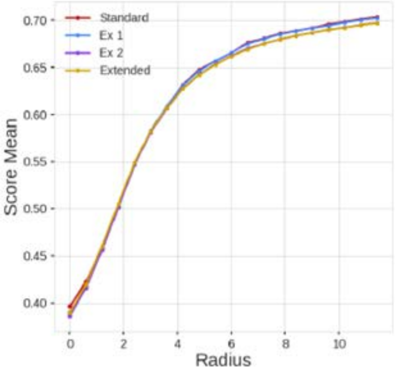
(c) Score Variance



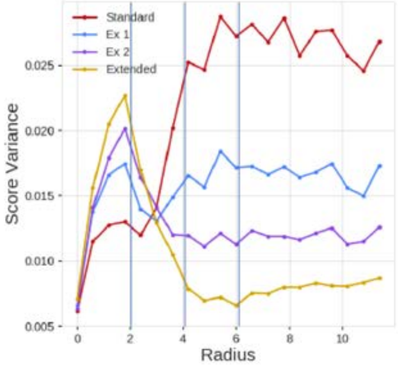
(a) 3-D Blob, mean of the scores



(b) 3-D Blob, variance of the scores

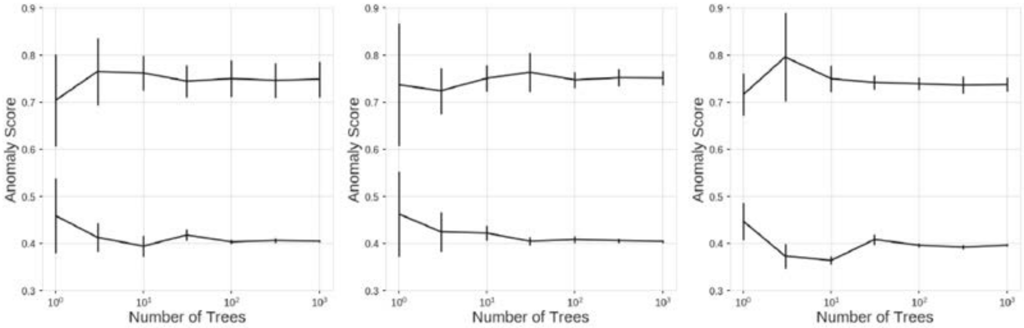


(a) 4-D Blob, mean of the scores

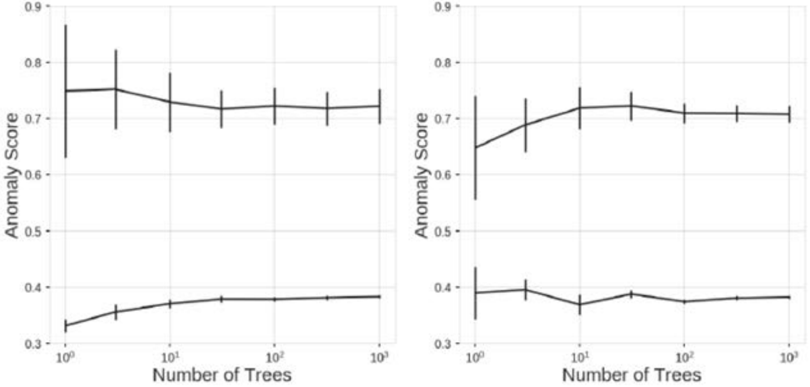


(b) 4-D Blob, variance of the scores

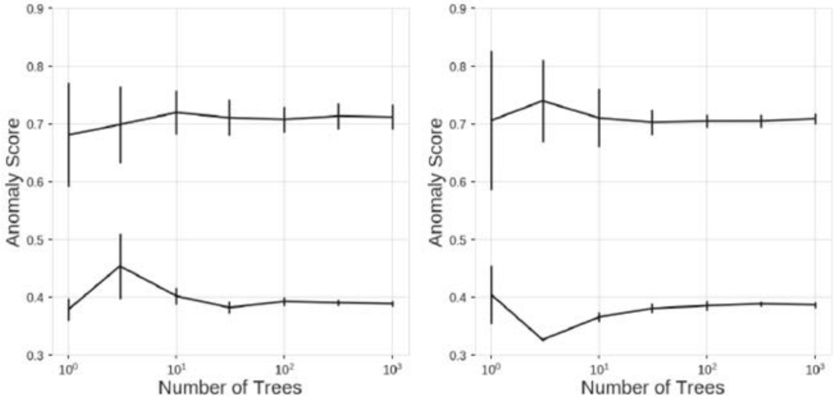
Convergence



(a) Standard Isolation Forest (b) Rotated Isolation Forest (c) Extended Isolation Forest



(a) Standard Isolation Forest (b) Extended Isolation forest



(a) Standard Isolation Forest (b) Extended Isolation forest

Streaming

- 2 cases: Time evolving data, Time accumulative data
- Streaming isolation forest exists, not extended
- We can adapt and retrain trees as new data is presented
- Replace trees one by one until whole forest is replaced
- Work with window size to retrain trees