

# Anomaly Detection and Failure Prediction

Evgeny Burnaev

Skoltech, Moscow, Russia



Skolkovo Institute of Science and Technology

- 1 Challenges
- 2 Examples of Projects
- 3 Methodology Principles
- 4 Anomaly Detection
- 5 One-Class SVM with Privileged Information

## 1 Challenges

## 2 Examples of Projects

## 3 Methodology Principles

## 4 Anomaly Detection

## 5 One-Class SVM with Privileged Information



Heat & power plant -  $10^3$  observations/ms



Rolled metal production -  $10^4$  observations/ms



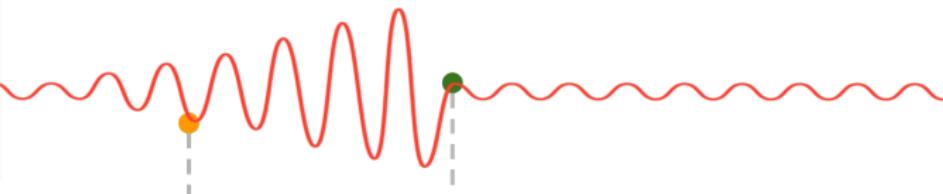
Modern aircraft – up to **0.5 Tb for one flight**

- Less than 1% of data is used, most of the data is not stored and used
- The next generation of Pratt & Whitney engines will produce up to **10 Gb/sec**



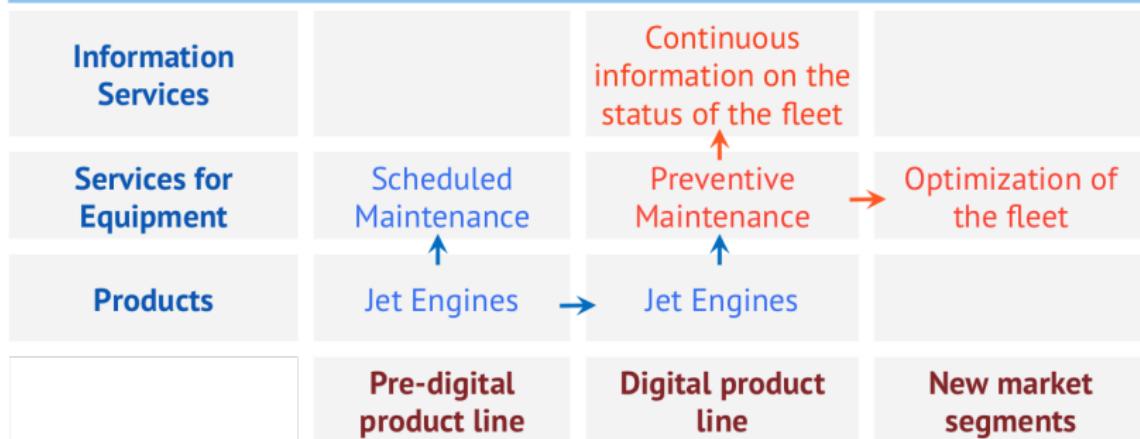
Self-driving Google car – up to **1 Gb/sec**

- Formula-1 car - **1.2 Gb/sec**



Exploitation	Time to Failure				Failure
Years/months	Months	Weeks	Days		
Optimal condition	Vibrations	Signs of wear	Decrease of main KPIs	Noise	Overheating
Intelligent Monitoring and Predictive Maintenance	First detection of a degradation		Corrective Maintenance		Failure is prevented
Other approaches to Maintenance	Scheduled Maintenance. Collected data is not used				Emergency Repair

Example from aviation industry on construction of brand new services based on predictive analytics of already available data



- Often when selecting a strategy it is necessary to optimize multiple targets
- Data Analytics can help to balance the “contradicting” targets



## Example from Aviation

When performing technical maintenance it is necessary to

- Increase the availability of an aircraft
- Reduce maintenance costs
- Minimize a number of operational interruptions
- Increase safety

- Often when selecting a strategy it is necessary to optimize multiple targets
- Data Analytics can help to balance the “contradicting” targets



## Example from Aviation

When performing technical maintenance it is necessary to

- Increase the availability of an aircraft
- Reduce maintenance costs
- Minimize a number of operational interruptions
- Increase safety

- Often when selecting a strategy it is necessary to optimize multiple targets
- Data Analytics can help to balance the “contradicting” targets



## Example from Aviation

When performing technical maintenance it is necessary to

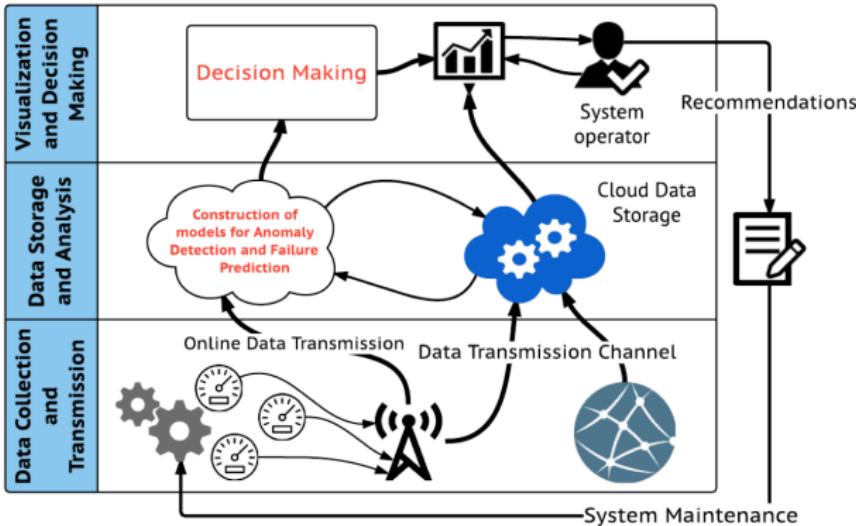
- Increase the availability of an aircraft
- Reduce maintenance costs
- Minimize a number of operational interruptions
- Increase safety

# Intelligent Maintenance System

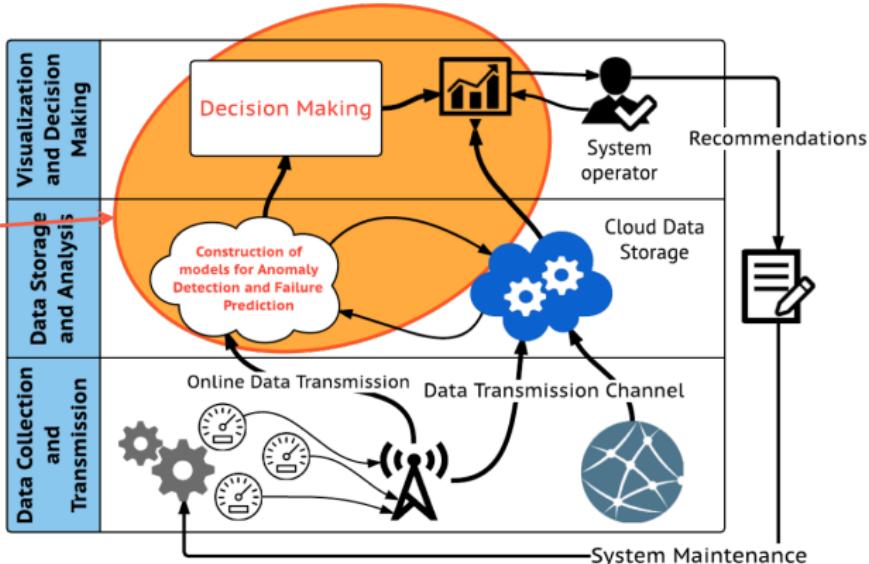
Specific for a lifecycle management system

Can be unified

Industry Specific



Our competences



## 1 Challenges

## 2 Examples of Projects

## 3 Methodology Principles

## 4 Anomaly Detection

## 5 One-Class SVM with Privileged Information

- **Objective:** optimize maintenance of the cooling system

- **Subtasks:**

- Quick leakage detection
- Determine whether a current refrigerant level is critical
- Prediction of time until achieving a critical level

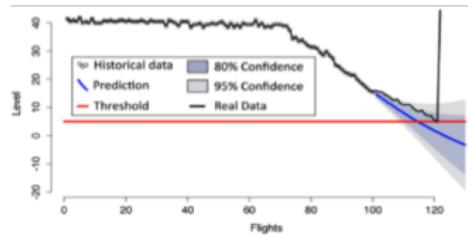
- **Data:**

- Time-series of levels of refrigerant
- 17 aircrafts, ~400 flights of each

- **Results:**

- False positive rate is < 1%
- Rate of correct detection is > 99%
- Average error of prediction before 10 flights until achieving a critical level is < 1 flight

- **Objective:** optimize maintenance of the cooling system
- **Subtasks:**
  - Quick leakage detection
  - Determine whether a current refrigerant level is critical
  - Prediction of time until achieving a critical level



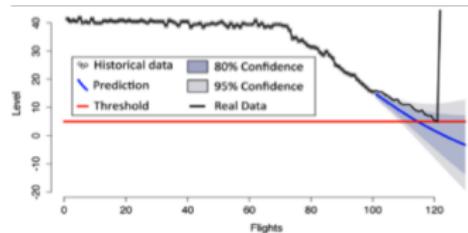
- **Data:**

- Time-series of levels of refrigerant
- 17 aircrafts, ~400 flights of each

- **Results:**

- False positive rate is < 1%
- Rate of correct detection is > 99%
- Average error of prediction before 10 flights until achieving a critical level is < 1 flight

- **Objective:** optimize maintenance of the cooling system
- **Subtasks:**
  - Quick leakage detection
  - Determine whether a current refrigerant level is critical
  - Prediction of time until achieving a critical level



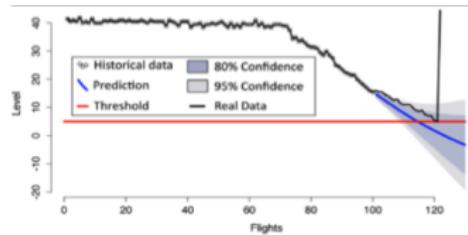
- **Data:**

- Time-series of levels of refrigerant
- 17 aircrafts, ~400 flights of each

- **Results:**

- False positive rate is < 1%
- Rate of correct detection is > 99%
- Average error of prediction before 10 flights until achieving a critical level is < 1 flight

- **Objective:** optimize maintenance of the cooling system
- **Subtasks:**
  - Quick leakage detection
  - Determine whether a current refrigerant level is critical
  - Prediction of time until achieving a critical level



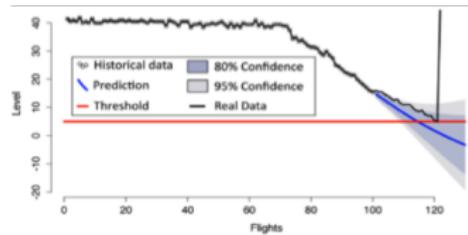
- **Data:**

- Time-series of levels of refrigerant
- 17 aircrafts,  $\sim 400$  flights of each

- **Results:**

- False positive rate is  $< 1\%$
- Rate of correct detection is  $> 99\%$
- Average error of prediction before 10 flights until achieving a critical level is  $< 1$  flight

- **Objective:** optimize maintenance of the cooling system
- **Subtasks:**
  - Quick leakage detection
  - Determine whether a current refrigerant level is critical
  - Prediction of time until achieving a critical level



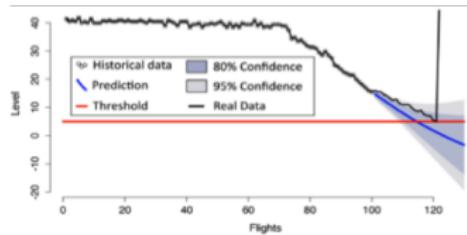
- **Data:**

- Time-series of levels of refrigerant
- 17 aircrafts,  $\sim 400$  flights of each

- **Results:**

- False positive rate is  $< 1\%$
- Rate of correct detection is  $> 99\%$
- Average error of prediction before 10 flights until achieving a critical level is  $< 1$  flight

- **Objective:** optimize maintenance of the cooling system
- **Subtasks:**
  - Quick leakage detection
  - Determine whether a current refrigerant level is critical
  - Prediction of time until achieving a critical level

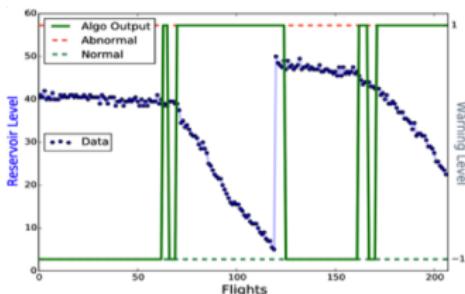


## ● Data:

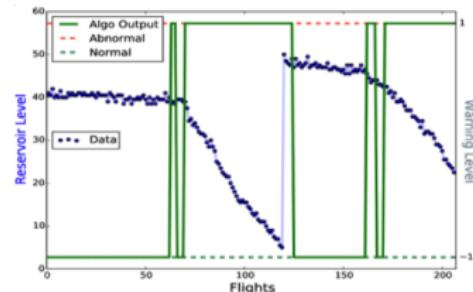
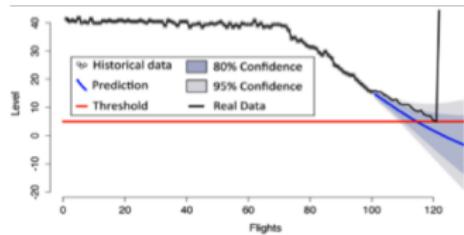
- Time-series of levels of refrigerant
- 17 aircrafts,  $\sim 400$  flights of each

## ● Results:

- False positive rate is  $< 1\%$
- Rate of correct detection is  $> 99\%$
- Average error of prediction before 10 flights until achieving a critical level is  $< 1$  flight



- **Objective:** optimize maintenance of the cooling system
- **Subtasks:**
  - Quick leakage detection
  - Determine whether a current refrigerant level is critical
  - Prediction of time until achieving a critical level



- **Data:**

- Time-series of levels of refrigerant
- 17 aircrafts,  $\sim 400$  flights of each

- **Results:**

- False positive rate is  $< 1\%$
- Rate of correct detection is  $> 99\%$
- Average error of prediction before 10 flights until achieving a critical level is  $< 1$  flight

# Auxiliary Power Unit Failures Prediction

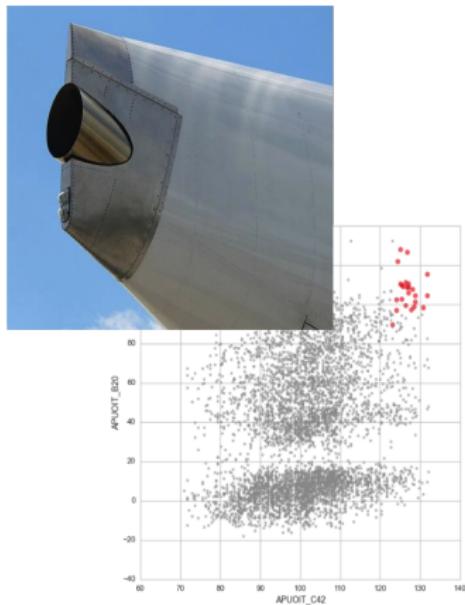
- **Objective:** to reduce costs associated with aircraft downtime due to unforeseen APU failures

- **Data:**

- 30 aircrafts, 200+ parameters for each aircraft
- Learning: 3 years (400 flights per year)
- Testing: 1/2 year

- **Conclusion/Benefits:**

- Early warnings about some types of failures
- Coverage (detected failures) ~ 34%
- Accuracy of Failures Detection is ~ 90% (for 9 correctly predicted failures on average we get only 1 false alarm)



# Auxiliary Power Unit Failures Prediction

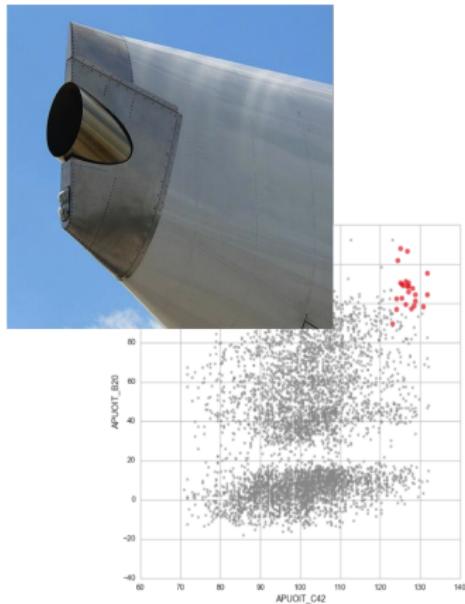
- **Objective:** to reduce costs associated with aircraft downtime due to unforeseen APU failures

- **Data:**

- 30 aircrafts, 200+ parameters for each aircraft
- Learning: 3 years (400 flights per year)
- Testing: 1/2 year

- **Conclusion/Benefits:**

- Early warnings about some types of failures
- Coverage (detected failures) ~ 34%
- Accuracy of Failures Detection is ~ 90% (for 9 correctly predicted failures on average we get only 1 false alarm)



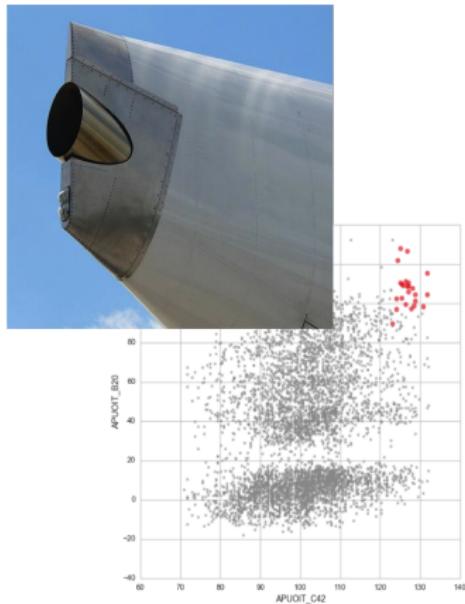
- **Objective:** to reduce costs associated with aircraft downtime due to unforeseen APU failures

- **Data:**

- 30 aircrafts, 200+ parameters for each aircraft
- Learning: 3 years (400 flights per year)
- Testing: 1/2 year

- **Conclusion/Benefits:**

- Early warnings about some types of failures
- Coverage (detected failures) ~ 34%
- Accuracy of Failures Detection is ~ 90% (for 9 correctly predicted failures on average we get only 1 false alarm)



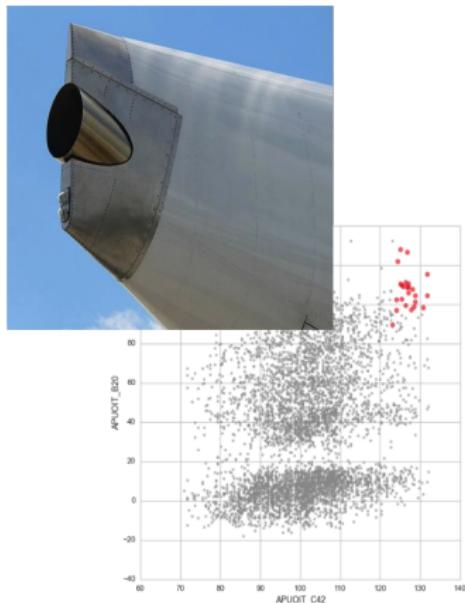
- **Objective:** to reduce costs associated with aircraft downtime due to unforeseen APU failures

- **Data:**

- 30 aircrafts, 200+ parameters for each aircraft
- Learning: 3 years (400 flights per year)
- Testing: 1/2 year

- **Conclusion/Benefits:**

- Early warnings about some types of failures
- Coverage (detected failures) ~ 34%
- Accuracy of Failures Detection is ~ 90% (for 9 correctly predicted failures on average we get only 1 false alarm)



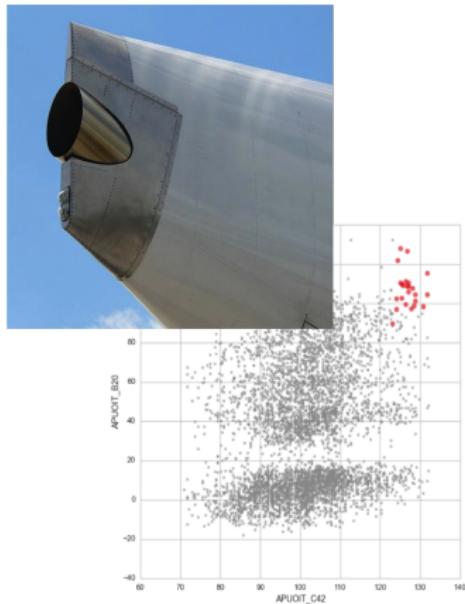
- **Objective:** to reduce costs associated with aircraft downtime due to unforeseen APU failures

- **Data:**

- 30 aircrafts, 200+ parameters for each aircraft
- Learning: 3 years (400 flights per year)
- Testing: 1/2 year

- **Conclusion/Benefits:**

- Early warnings about some types of failures
- Coverage (detected failures)  
~ 34%
- Accuracy of Failures Detection is  
~ 90% (for 9 correctly predicted failures on average we get only 1 false alarm)



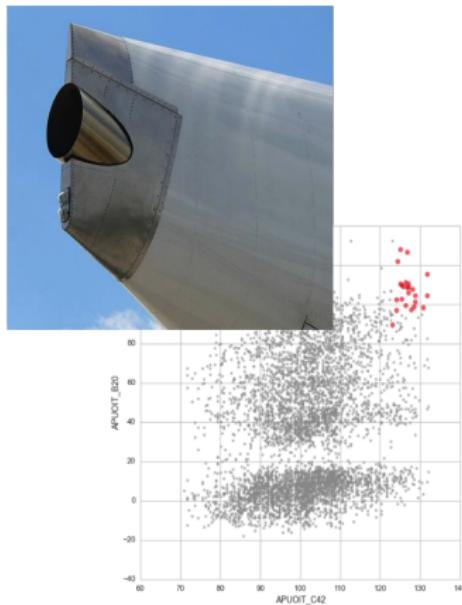
- **Objective:** to reduce costs associated with aircraft downtime due to unforeseen APU failures

- **Data:**

- 30 aircrafts, 200+ parameters for each aircraft
- Learning: 3 years (400 flights per year)
- Testing: 1/2 year

- **Conclusion/Benefits:**

- Early warnings about some types of failures
- **Coverage** (detected failures) ~ 34%
- **Accuracy of Failures Detection** is ~ 90% (for 9 correctly predicted failures on average we get only 1 false alarm)



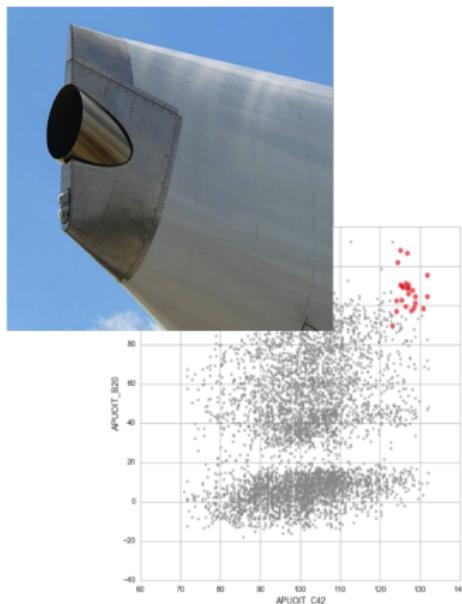
- **Objective:** to reduce costs associated with aircraft downtime due to unforeseen APU failures

- **Data:**

- 30 aircrafts, 200+ parameters for each aircraft
- Learning: 3 years (400 flights per year)
- Testing: 1/2 year

- **Conclusion/Benefits:**

- Early warnings about some types of failures
- **Coverage** (detected failures) ~ 34%
- **Accuracy** of Failures Detection is ~ 90% (for 9 correctly predicted failures on average we get only 1 false alarm)

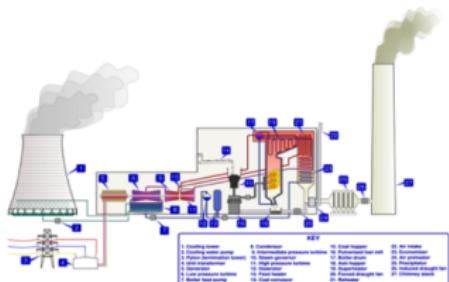


- **Objective:**

1. Detect Power Losses in the system
2. Localize origins of power losses

- **Subtasks:**

- Construct a model of a system in normal regime
- Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)



- **Data:** 200+ dimensional time-series, one-observation per 10 min.
- **Results:**
  - Detection of power losses in the system
  - Localization of the power losses origins
  - Results were confirmed by experts

- **Objective:**

1. Detect Power Losses in the system
2. Localize origins of power losses

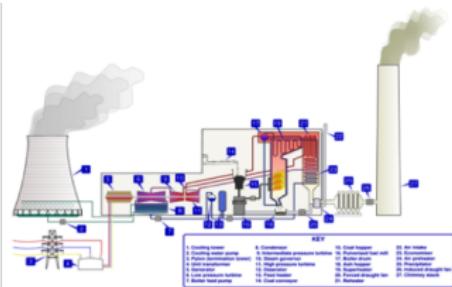
- **Subtasks:**

- Construct a model of a system in normal regime
- Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)

- **Data:** 200+ dimensional time-series, one-observation per 10 min.

- **Results:**

- Detection of power losses in the system
- Localization of the power losses origins
- Results were confirmed by experts

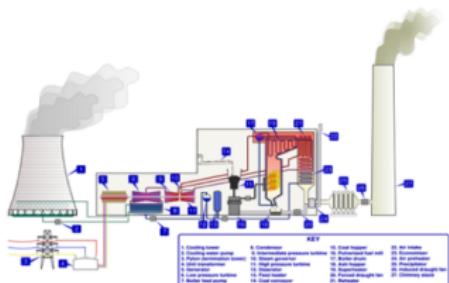


- **Objective:**

1. Detect Power Losses in the system
2. Localize origins of power losses

- **Subtasks:**

- Construct a model of a system in normal regime
- Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)



- **Data:** 200+ dimensional time-series, one-observation per 10 min.

- **Results:**

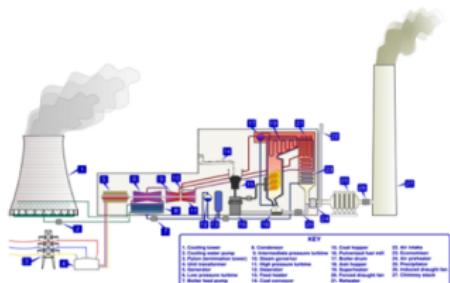
- Detection of power losses in the system
- Localization of the power losses origins
- Results were confirmed by experts

- **Objective:**

1. Detect Power Losses in the system
2. Localize origins of power losses

- **Subtasks:**

- Construct a model of a system in normal regime
- Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)



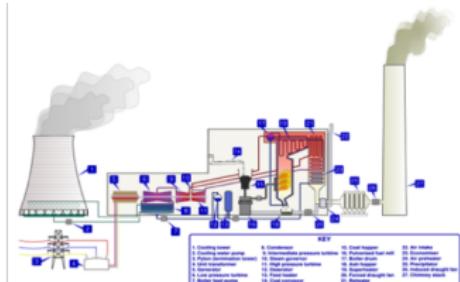
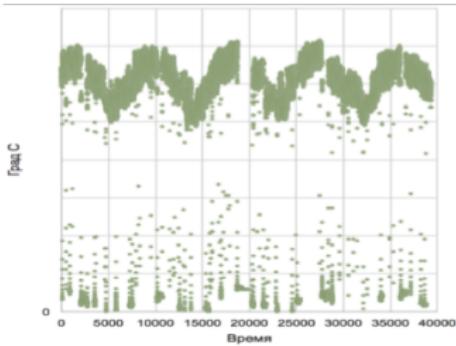
- **Data:** 200+ dimensional time-series, one-observation per 10 min.
- **Results:**
  - Detection of power losses in the system
  - Localization of the power losses origins
  - Results were confirmed by experts

- **Objective:**

1. Detect Power Losses in the system
2. Localize origins of power losses

- **Subtasks:**

- Construct a model of a system in normal regime
- Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)



- **Data:** 200+ dimensional time-series, one-observation per 10 min.

- **Results:**

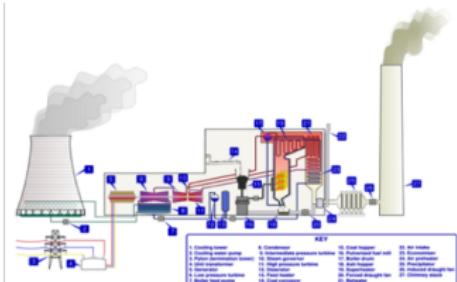
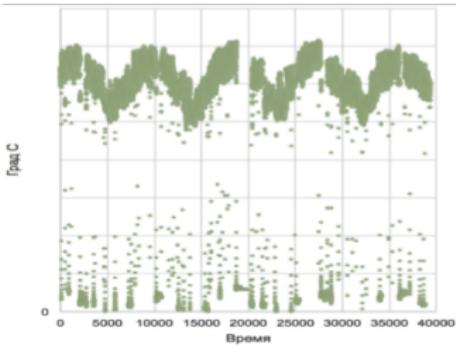
- Detection of power losses in the system
- Localization of the power losses origins
- Results were confirmed by experts

- **Objective:**

1. Detect Power Losses in the system
  2. Localize origins of power losses

- Subtasks:

- Construct a model of a system in normal regime
  - Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)



- **Data:** 200+ dimensional time-series, one-observation per 10 min.
  - **Results:**
    - Detection of power losses in the system
    - Localization of the power losses origins
    - Results were confirmed by experts

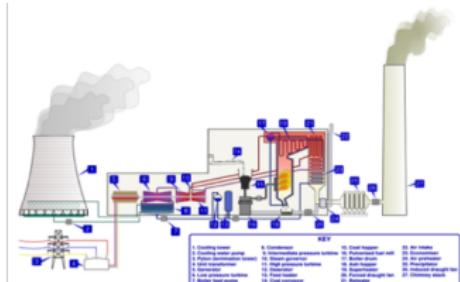
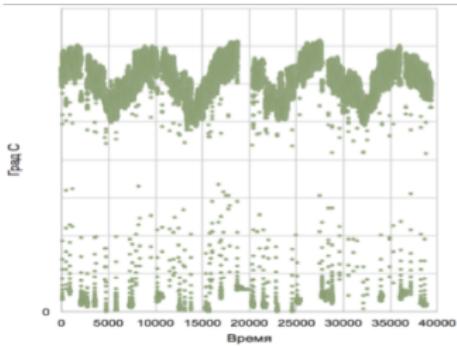
In collaboration with Datadvance LLC

- **Objective:**

1. Detect Power Losses in the system
2. Localize origins of power losses

- **Subtasks:**

- Construct a model of a system in normal regime
- Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)



- **Data:** 200+ dimensional time-series, one-observation per 10 min.

- **Results:**

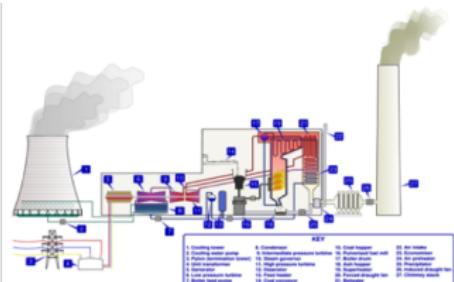
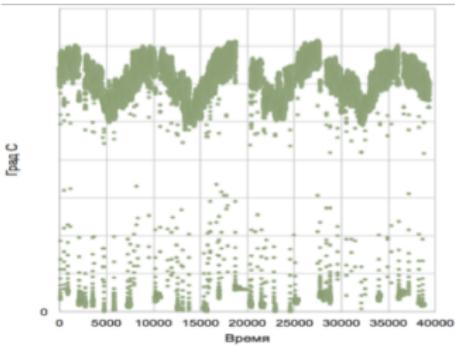
- Detection of power losses in the system
- Localization of the power losses origins
- Results were confirmed by experts

- **Objective:**

1. Detect Power Losses in the system
2. Localize origins of power losses

- **Subtasks:**

- Construct a model of a system in normal regime
- Model sensitivity analysis w.r.t. a change in a system behavior (potential failures)



- **Data:** 200+ dimensional time-series, one-observation per 10 min.

- **Results:**

- Detection of power losses in the system
- Localization of the power losses origins
- Results were confirmed by experts

## 1 Challenges

## 2 Examples of Projects

## 3 Methodology Principles

## 4 Anomaly Detection

## 5 One-Class SVM with Privileged Information

- Multistream/multichannel scenarios with unstructured hypotheses/patterns
- High-dimensionality and Large data volumes
- Composite hypotheses
- Stochastic models with non-stationary and dependent observations of a very general structure
- Prior uncertainty with respect to pre- and post-change distributions
- Parametric assumptions are inefficient
- Imbalanced learning samples

- Multistream/multichannel scenarios with unstructured hypotheses/patterns
- High-dimensionality and Large data volumes
- Composite hypotheses
- Stochastic models with non-stationary and dependent observations of a very general structure
- Prior uncertainty with respect to pre- and post-change distributions
- Parametric assumptions are inefficient
- Imbalanced learning samples

- Multistream/multichannel scenarios with unstructured hypotheses/patterns
- High-dimensionality and Large data volumes
- Composite hypotheses
- Stochastic models with non-stationary and dependent observations of a very general structure
- Prior uncertainty with respect to pre- and post-change distributions
- Parametric assumptions are inefficient
- Imbalanced learning samples

- Multistream/multichannel scenarios with unstructured hypotheses/patterns
- High-dimensionality and Large data volumes
- Composite hypotheses
- Stochastic models with non-stationary and dependent observations of a very general structure
- Prior uncertainty with respect to pre- and post-change distributions
- Parametric assumptions are inefficient
- Imbalanced learning samples

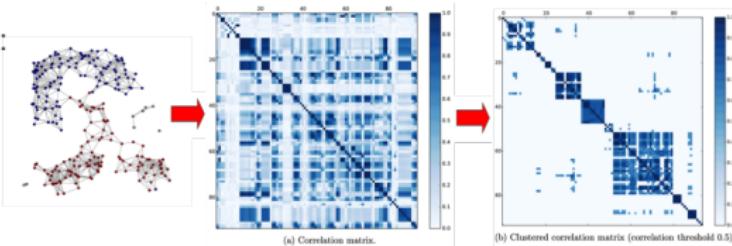
- Multistream/multichannel scenarios with unstructured hypotheses/patterns
- High-dimensionality and Large data volumes
- Composite hypotheses
- Stochastic models with non-stationary and dependent observations of a very general structure
- Prior uncertainty with respect to pre- and post-change distributions
- Parametric assumptions are inefficient
- Imbalanced learning samples

- Multistream/multichannel scenarios with unstructured hypotheses/patterns
- High-dimensionality and Large data volumes
- Composite hypotheses
- Stochastic models with non-stationary and dependent observations of a very general structure
- Prior uncertainty with respect to pre- and post-change distributions
- Parametric assumptions are inefficient
- Imbalanced learning samples

- Multistream/multichannel scenarios with unstructured hypotheses/patterns
- High-dimensionality and Large data volumes
- Composite hypotheses
- Stochastic models with non-stationary and dependent observations of a very general structure
- Prior uncertainty with respect to pre- and post-change distributions
- Parametric assumptions are inefficient
- Imbalanced learning samples

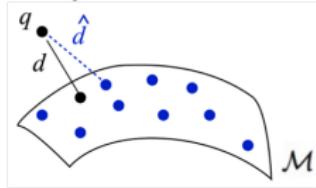
## 1. Subsystems Identification:

Identification of groups of dependent parameters, corresponding to different subsystems



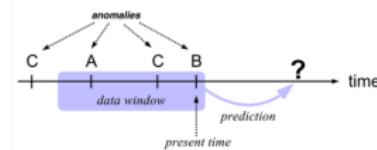
## 2. Anomalies Detection:

Detection of Anomalies based e.g. on Manifold Learning for identified subsystems



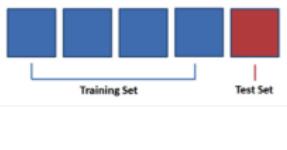
## 3. Events Matching:

Statistical techniques used to identify best anomalies preceding warnings (and not happening anywhere else).

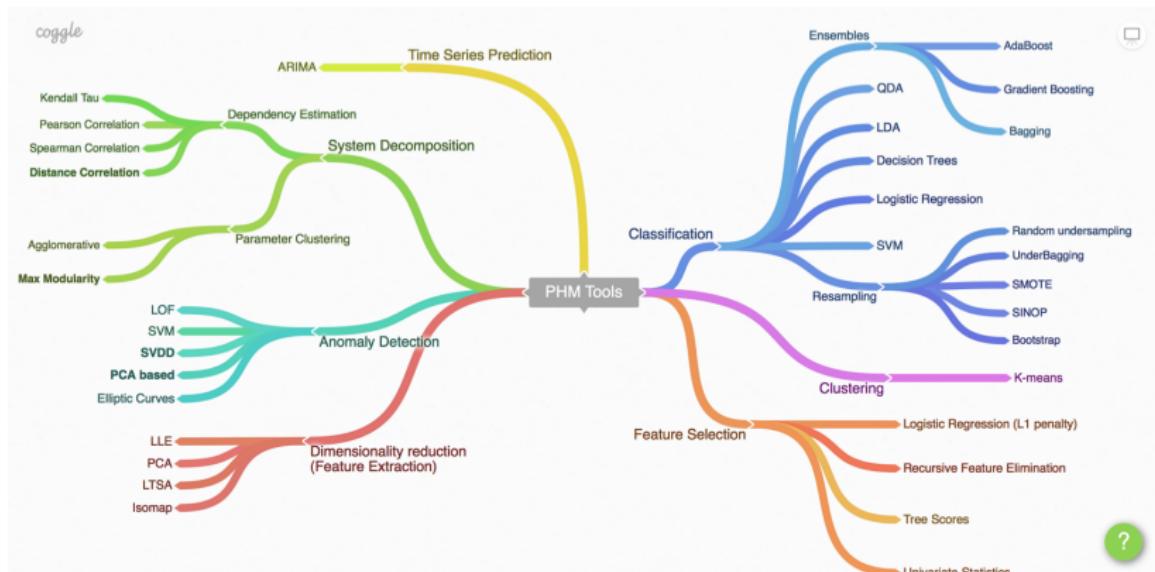


## 4. Validation:

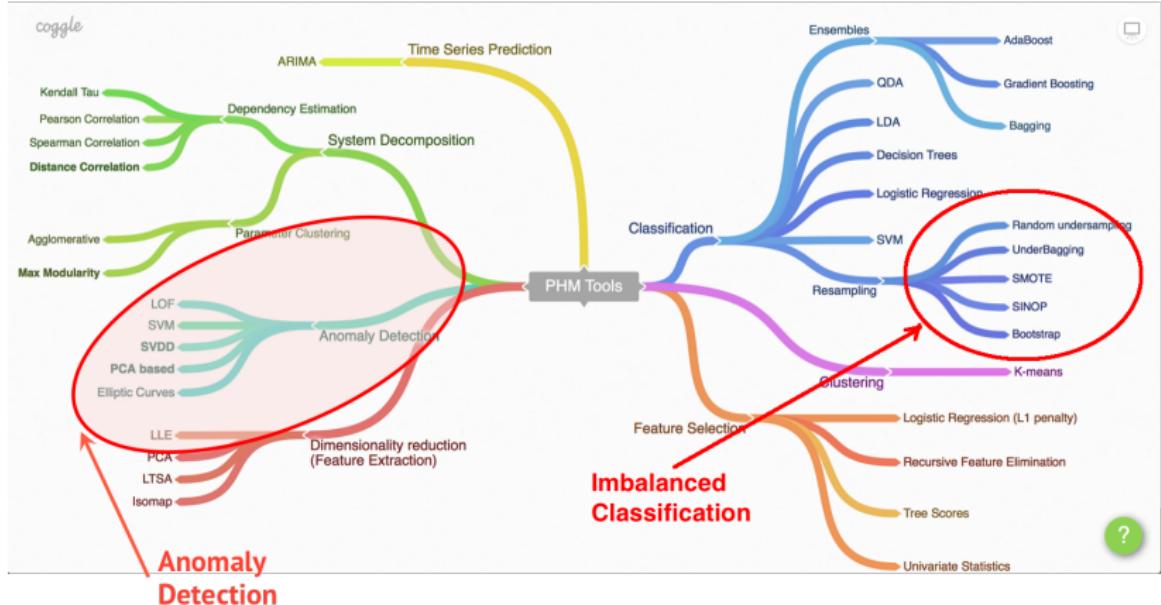
Apply anomalies detection logics in a new airplane (left apart) to calculate the prediction performances.



# Methods



# Anomaly Detection



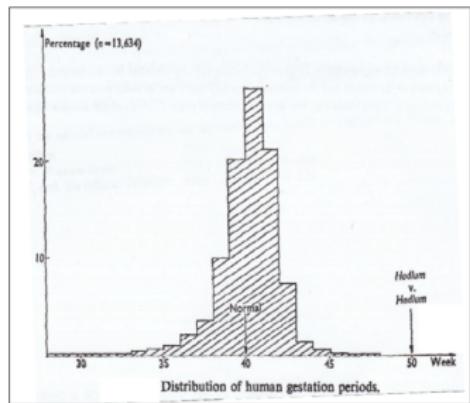
## 1 Challenges

## 2 Examples of Projects

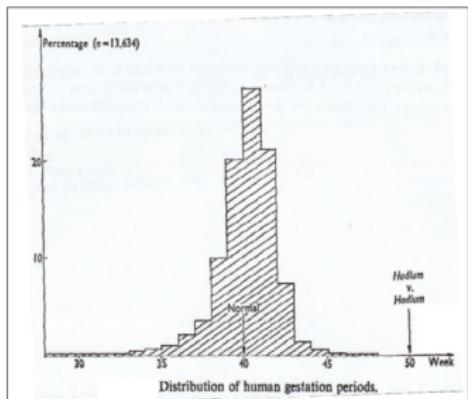
## 3 Methodology Principles

## 4 Anomaly Detection

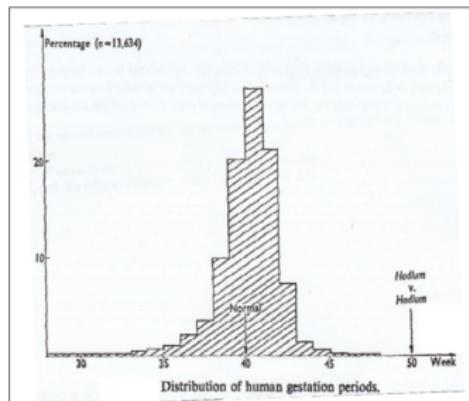
## 5 One-Class SVM with Privileged Information



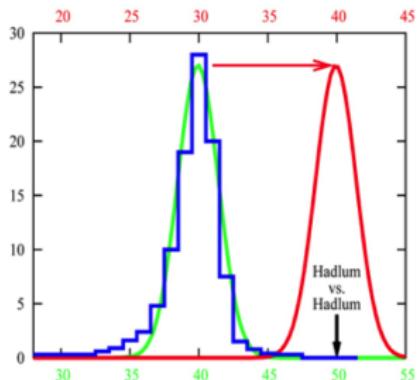
- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service
- Average human pregnancy period is 280 days (40 weeks)
- Statistically, 349 days is an outlier



- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service
- Average human pregnancy period is 280 days (40 weeks)
- Statistically, 349 days is an outlier



- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service
- Average human pregnancy period is 280 days (40 weeks)
- Statistically, 349 days is an outlier



- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service
- Average human pregnancy period is 280 days (40 weeks)
- Statistically, 349 days is an outlier

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”

- Data is multivariate

- There is usually more than one generating mechanism/statistical process underlying the “normal” data
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
- Domain specific definition of what to count as anomaly
- Normal behavior keeps evolving
- Malicious adversaries
- Availability of labeled data for training/validation

- Data is multivariate
- There is usually more than one generating mechanism/statistical process underlying the “normal” data
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
- Domain specific definition of what to count as anomaly
- Normal behavior keeps evolving
- Malicious adversaries
- Availability of labeled data for training/validation

- Data is multivariate
- There is usually more than one generating mechanism/statistical process underlying the “normal” data
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
- Domain specific definition of what to count as anomaly
- Normal behavior keeps evolving
- Malicious adversaries
- Availability of labeled data for training/validation

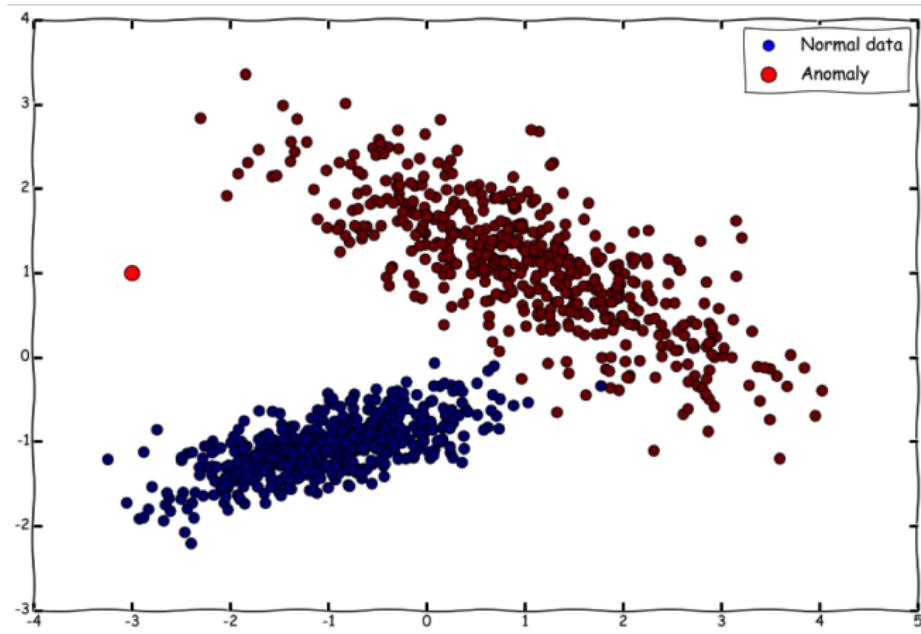
- Data is multivariate
- There is usually more than one generating mechanism/statistical process underlying the “normal” data
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
- Domain specific definition of what to count as anomaly
  - Normal behavior keeps evolving
  - Malicious adversaries
  - Availability of labeled data for training/validation

- Data is multivariate
- There is usually more than one generating mechanism/statistical process underlying the “normal” data
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
- Domain specific definition of what to count as anomaly
- Normal behavior keeps evolving
- Malicious adversaries
- Availability of labeled data for training/validation

- Data is multivariate
- There is usually more than one generating mechanism/statistical process underlying the “normal” data
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
- Domain specific definition of what to count as anomaly
- Normal behavior keeps evolving
- Malicious adversaries
- Availability of labeled data for training/validation

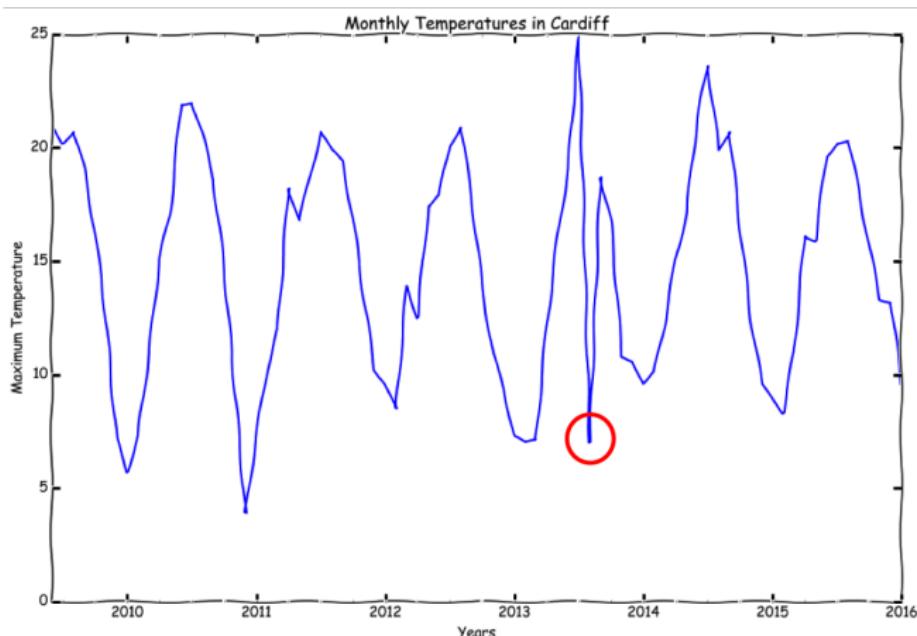
- Data is multivariate
- There is usually more than one generating mechanism/statistical process underlying the “normal” data
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
- Domain specific definition of what to count as anomaly
- Normal behavior keeps evolving
- Malicious adversaries
- Availability of labeled data for training/validation

- An individual data instance is anomalous w.r.t. the data



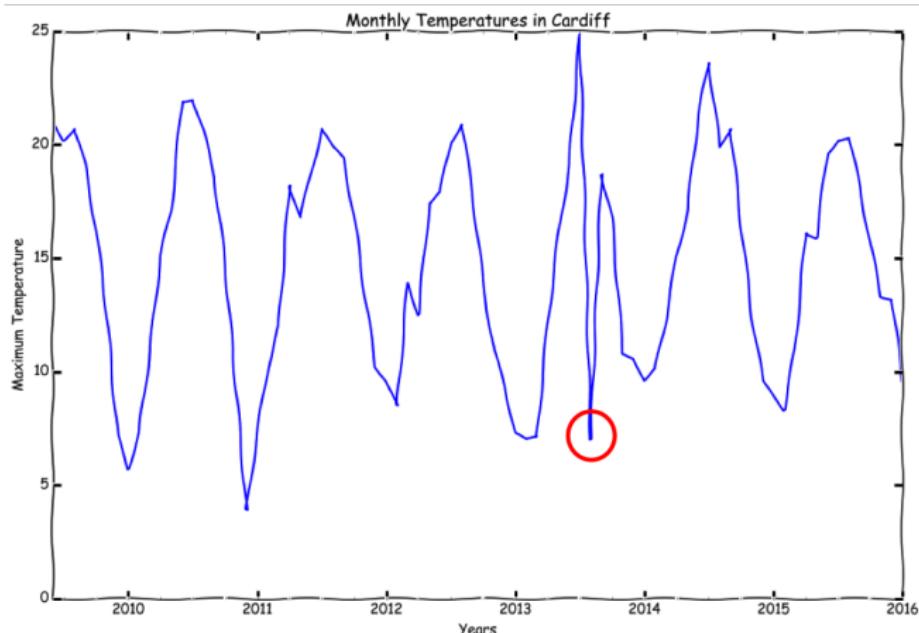
# Anomaly taxonomy: Contextual Anomalies

- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies



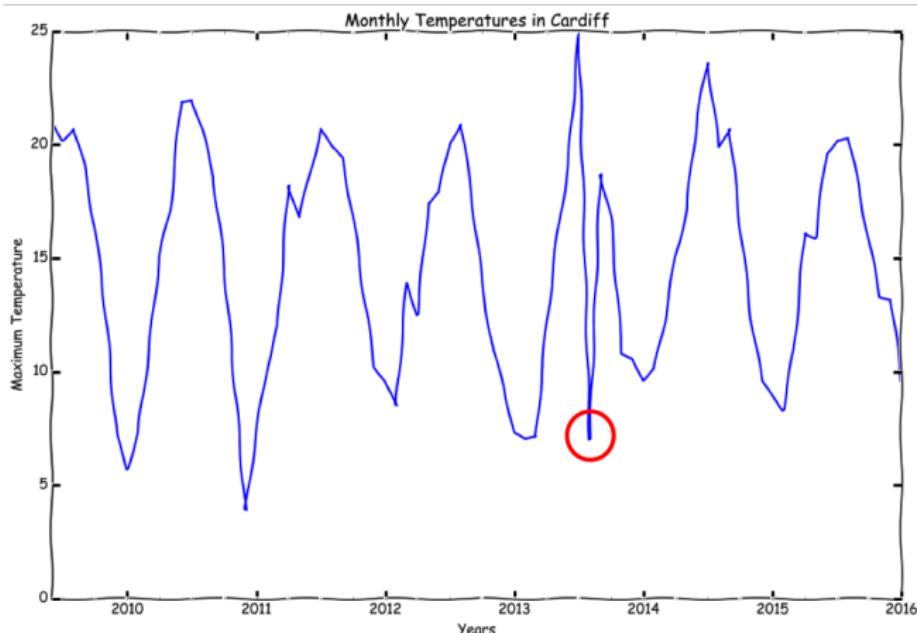
# Anomaly taxonomy: Contextual Anomalies

- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies

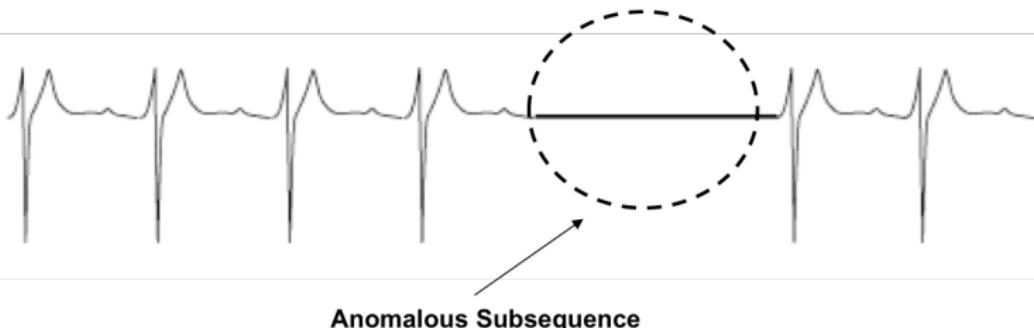


# Anomaly taxonomy: Contextual Anomalies

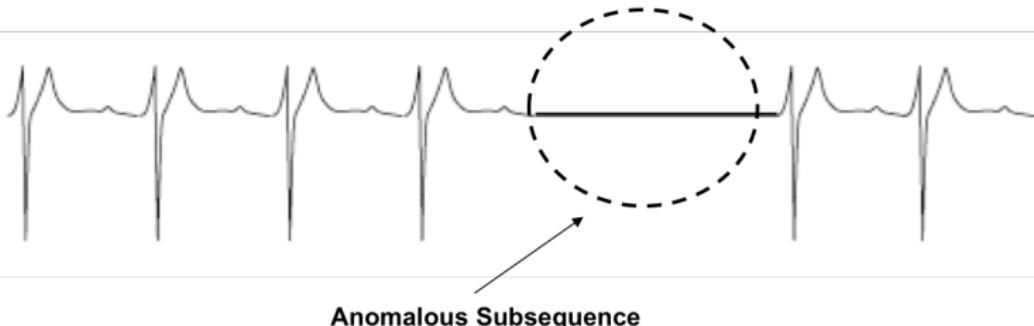
- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies



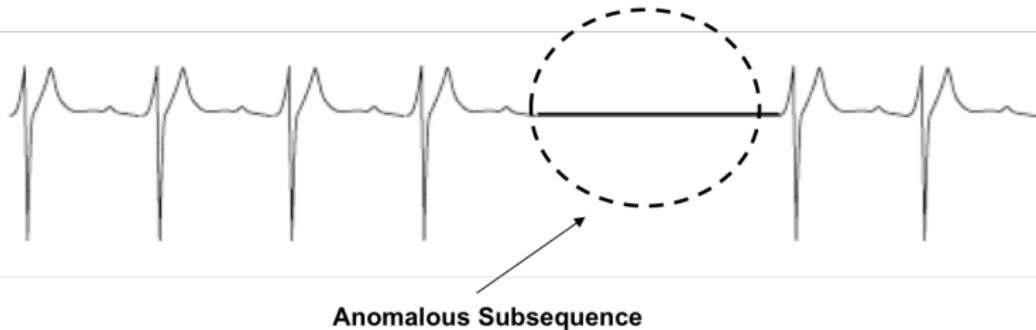
- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a set of causal anomalies are not anomalous by themselves



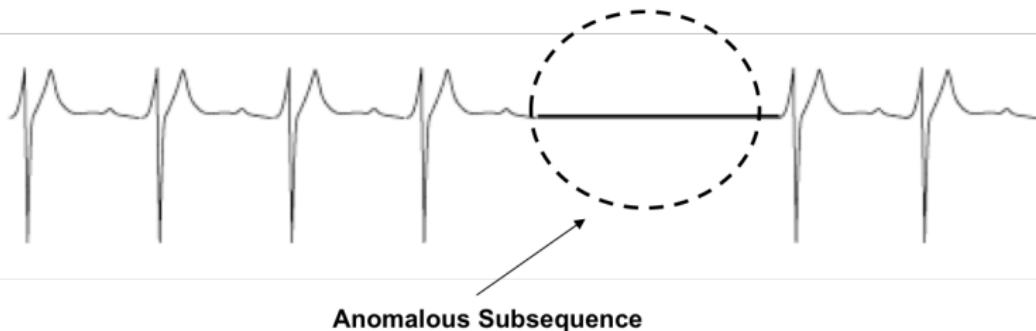
- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a set of causal anomalies are not anomalous by themselves



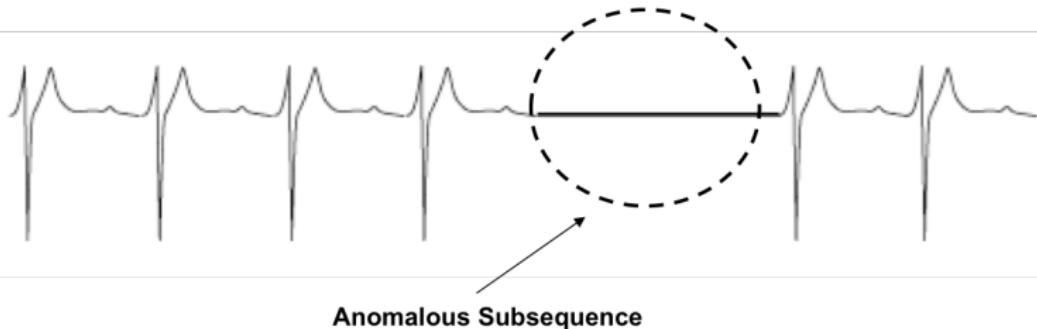
- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a set of causal anomalies are not anomalous by themselves

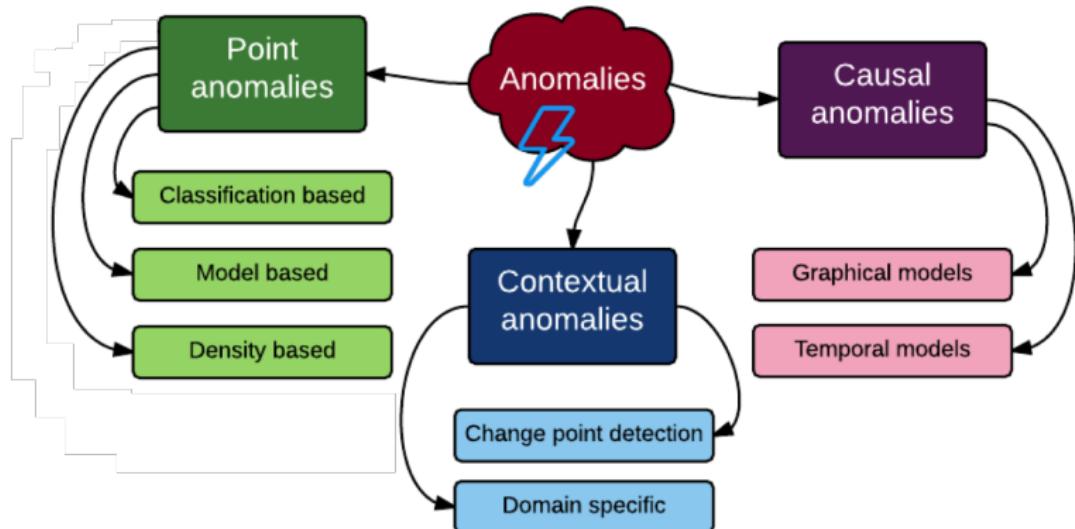


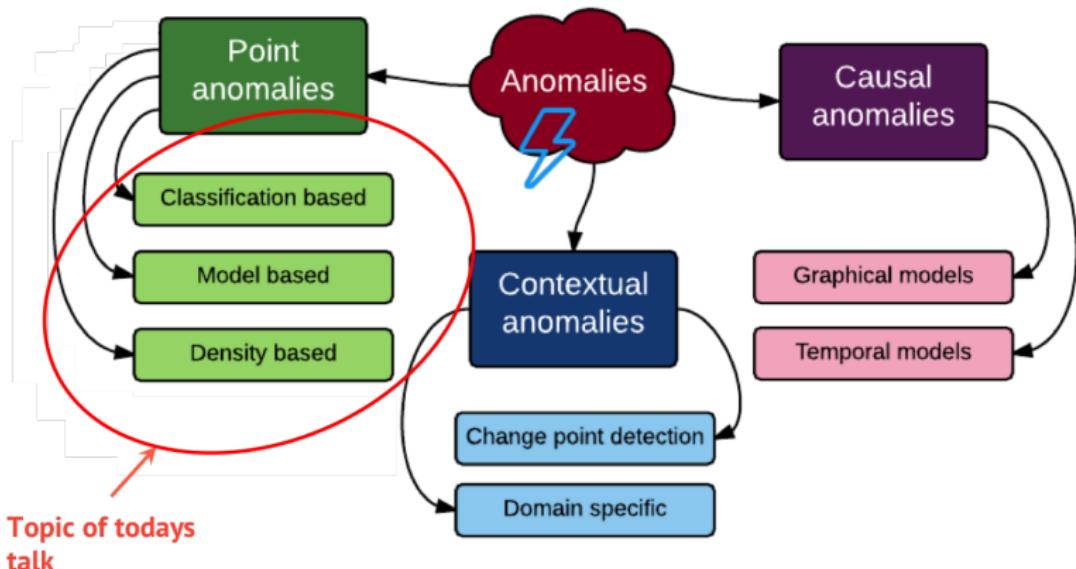
- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a set of causal anomalies are not anomalous by themselves



- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a set of causal anomalies are not anomalous by themselves







- Weights for classes
  - Proved not to be helpful in most cases
- Resampling methods
  - Oversampling (Bootstrap, SMOTE, etc.)
  - Undersampling
- How to choose which method to use?
- How to choose resampling parameter?

- Weights for classes
  - Proved not to be helpful in most cases
- Resampling methods
  - Oversampling (Bootstrap, SMOTE, etc.)
  - Undersampling
- How to choose which method to use?
- How to choose resampling parameter?

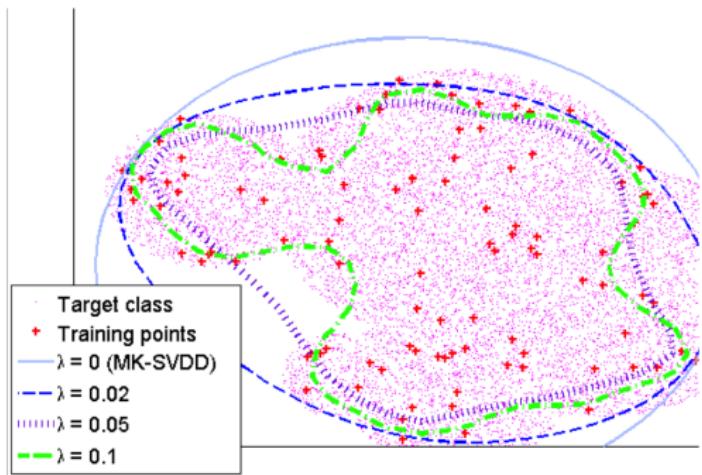
- Weights for classes
  - Proved not to be helpful in most cases
- Resampling methods
  - Oversampling (Bootstrap, SMOTE, etc.)
  - Undersampling
- How to choose which method to use?
- How to choose resampling parameter?

- Weights for classes
  - Proved not to be helpful in most cases
- Resampling methods
  - Oversampling (Bootstrap, SMOTE, etc.)
  - Undersampling
- How to choose which method to use?
- How to choose resampling parameter?

- Assumption on normal data generation procedure (e.g. Gaussian distribution, etc.)
- PCA is a method commonly used to extract most variant combinations in data
- PCA based anomaly detection is good for highly correlated environments

- Assumption on normal data generation procedure (e.g. Gaussian distribution, etc.)
- PCA is a method commonly used to extract most variant combinations in data
- PCA based anomaly detection is good for highly correlated environments

- Assumption on normal data generation procedure (e.g. Gaussian distribution, etc.)
- PCA is a method commonly used to extract most variant combinations in data
- PCA based anomaly detection is good for highly correlated environments



- SVM-based and nearest neighbours based
- How to choose best kernel parameter?

## 1 Challenges

## 2 Examples of Projects

## 3 Methodology Principles

## 4 Anomaly Detection

## 5 One-Class SVM with Privileged Information

1. Consider Anomaly Detection with Privileged Information
2. Modify accordingly One-Class Classification approach
3. Test Vapnik's methodology in some real-life application

- Let  $S = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ ,  $\mathbf{x}_i \in \mathbb{R}^p$  be an unlabeled sample (possibly containing some anomalies)
- We want to learn  $f : \mathbf{x} \rightarrow \{-1, 1\}$  using the sample

$$\mathbf{x} = \begin{cases} \text{normal, if } f(\mathbf{x}) = +1, \\ \text{anomaly, if } f(\mathbf{x}) = -1, \end{cases}$$

- Let  $S = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ ,  $\mathbf{x}_i \in \mathbb{R}^p$  be an unlabeled sample (possibly containing some anomalies)
- We want to learn  $f : \mathbf{x} \rightarrow \{-1, 1\}$  using the sample

$$\mathbf{x} = \begin{cases} \text{normal, if } f(\mathbf{x}) = +1, \\ \text{anomaly, if } f(\mathbf{x}) = -1, \end{cases}$$

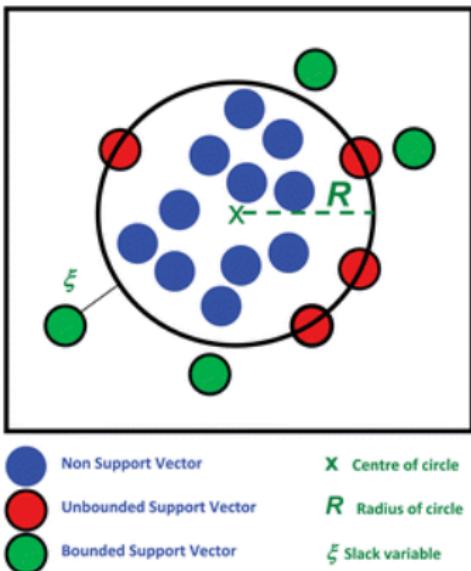
$$R + \frac{1}{m\nu} \sum_{i=1}^m \xi_i \rightarrow \min_{R, a, \xi}$$

$$\text{s.t. } \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + \xi_i$$

$$\xi_i \geq 0$$

$$R \geq 0$$

- $\nu$  is an upper bound on the fraction of anomalous patterns in the sample  $S$
- $\phi(\mathbf{x}_i)$  is the mapping to a high dimensional space



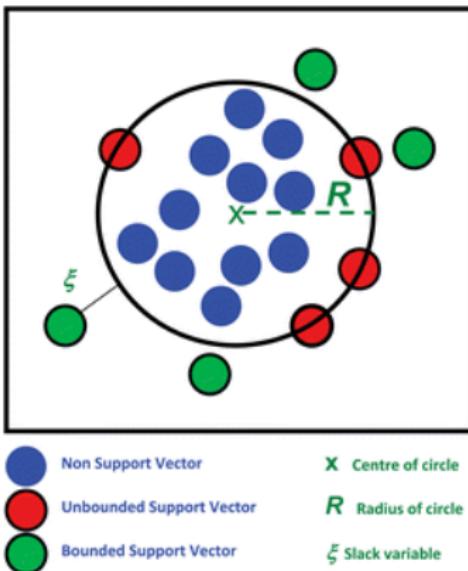
$$R + \frac{1}{m\nu} \sum_{i=1}^m \xi_i \rightarrow \min_{R, a, \xi}$$

$$\text{s.t. } \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + \xi_i$$

$$\xi_i \geq 0$$

$$R \geq 0$$

- $\nu$  is an upper bound on the fraction of anomalous patterns in the sample  $S$
- $\phi(\mathbf{x}_i)$  is the mapping to a high dimensional space



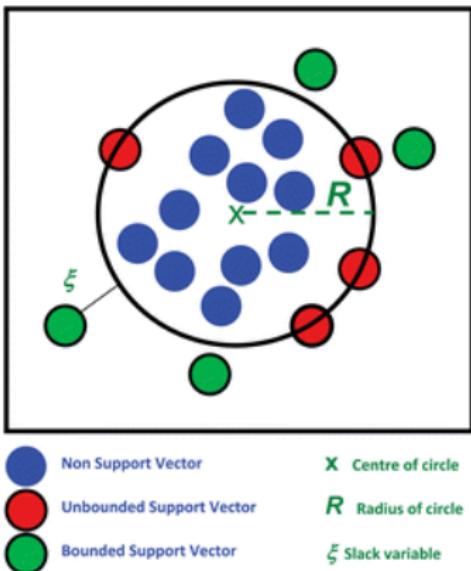
$$R + \frac{1}{m\nu} \sum_{i=1}^m \xi_i \rightarrow \min_{R, a, \xi}$$

$$\text{s.t. } \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + \xi_i$$

$$\xi_i \geq 0$$

$$R \geq 0$$

- $\nu$  is an upper bound on the fraction of anomalous patterns in the sample  $S$
- $\phi(\mathbf{x}_i)$  is the mapping to a high dimensional space



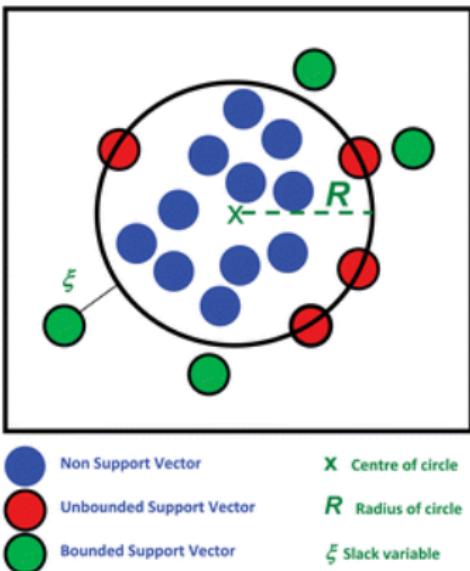
$$R + \frac{1}{m\nu} \sum_{i=1}^m \xi_i \rightarrow \min_{R, a, \xi}$$

$$\text{s.t. } \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + \xi_i$$

$$\xi_i \geq 0$$

$$R \geq 0$$

- $\nu$  is an upper bound on the fraction of anomalous patterns in the sample  $S$
- $\phi(\mathbf{x}_i)$  is the mapping to a high dimensional space



We consider the dual problem

$$\sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_i)^\top - \sum_{i,j=1}^m \alpha_i \alpha_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)^\top \rightarrow \max_{\alpha}$$

$$s.t. \quad \sum_{i=1}^m \alpha_i = 1$$

$$0 \leq \alpha_i \leq \frac{1}{m\nu}, \quad i = 1, \dots, m$$

We don't need to use explicit expression for  $\phi(\cdot)$ , we need only a definition of a dot product. We can use a kernel trick

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')^\top$$

E.g.  $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma^2)$

We consider the dual problem

$$\sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_i)^\top - \sum_{i,j=1}^m \alpha_i \alpha_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)^\top \rightarrow \max_{\alpha}$$

$$s.t. \quad \sum_{i=1}^m \alpha_i = 1$$

$$0 \leq \alpha_i \leq \frac{1}{m\nu}, \quad i = 1, \dots, m$$

We don't need to use explicit expression for  $\phi(\cdot)$ , we need only a definition of a dot product. We can use a kernel trick

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')^\top$$

E.g.  $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma^2)$

We consider the dual problem

$$\sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_i)^\top - \sum_{i,j=1}^m \alpha_i \alpha_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)^\top \rightarrow \max_{\alpha}$$

$$s.t. \quad \sum_{i=1}^m \alpha_i = 1$$

$$0 \leq \alpha_i \leq \frac{1}{m\nu}, \quad i = 1, \dots, m$$

We don't need to use explicit expression for  $\phi(\cdot)$ , we need only a definition of a dot product. We can use a kernel trick

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')^\top$$

E.g.  $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma^2)$

- We can write out the solution of the primal problem using the solution of the dual problem

$$a = \sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i), \quad R = \|\phi(\mathbf{x}_j)\|_2^2 - 2(a \cdot \phi(\mathbf{x}_j)^\top) + \|a\|_2^2,$$

where we can use any  $\mathbf{x}_j$ , such that  $\alpha_j > 0$

- Here

$$\|\phi(\mathbf{x})\|_2^2 = K(\mathbf{x}, \mathbf{x}),$$

$$(a \cdot \phi(\mathbf{x})^\top) = \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \mathbf{x}),$$

$$\|a\|_2^2 = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

- The decision function has the form

$$f(\mathbf{x}) = \text{sign} \left\{ R - K(\mathbf{x}, \mathbf{x}) + 2 \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) - \|a\|_2^2 \right\}$$

- We can write out the solution of the primal problem using the solution of the dual problem

$$a = \sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i), \quad R = \|\phi(\mathbf{x}_j)\|_2^2 - 2(a \cdot \phi(\mathbf{x}_j)^\top) + \|a\|_2^2,$$

where we can use any  $\mathbf{x}_j$ , such that  $\alpha_j > 0$

- Here

$$\|\phi(\mathbf{x})\|_2^2 = K(\mathbf{x}, \mathbf{x}),$$

$$(a \cdot \phi(\mathbf{x})^\top) = \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \mathbf{x}),$$

$$\|a\|_2^2 = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

- The decision function has the form

$$f(\mathbf{x}) = \text{sign} \left\{ R - K(\mathbf{x}, \mathbf{x}) + 2 \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) - \|a\|_2^2 \right\}$$

- We can write out the solution of the primal problem using the solution of the dual problem

$$a = \sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i), \quad R = \|\phi(\mathbf{x}_j)\|_2^2 - 2(a \cdot \phi(\mathbf{x}_j)^\top) + \|a\|_2^2,$$

where we can use any  $\mathbf{x}_j$ , such that  $\alpha_j > 0$

- Here

$$\|\phi(\mathbf{x})\|_2^2 = K(\mathbf{x}, \mathbf{x}),$$

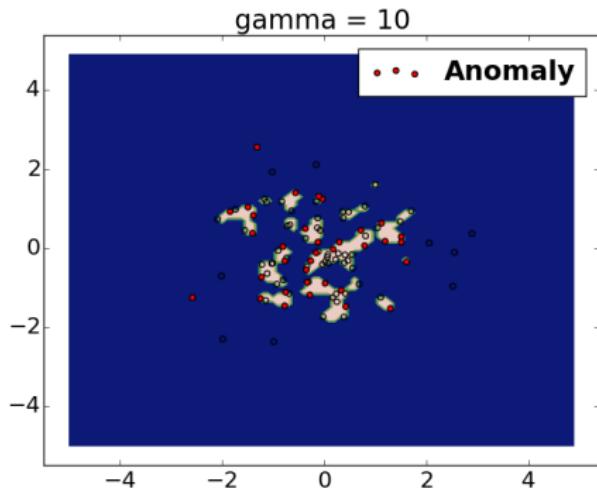
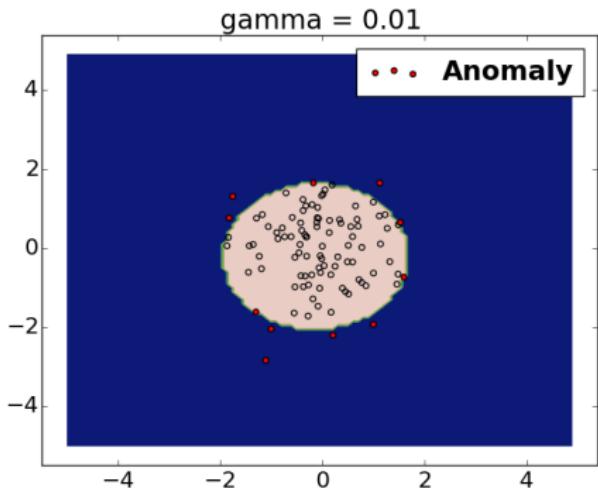
$$(a \cdot \phi(\mathbf{x})^\top) = \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \mathbf{x}),$$

$$\|a\|_2^2 = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

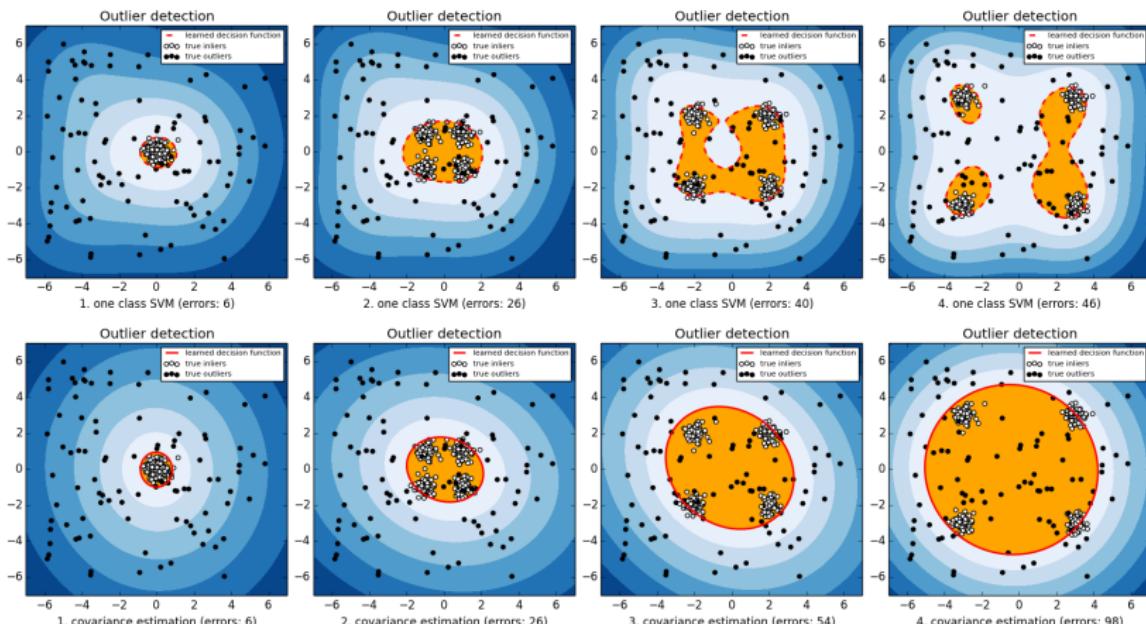
- The **decision function** has the form

$$f(\mathbf{x}) = \text{sign} \left\{ R - K(\mathbf{x}, \mathbf{x}) + 2 \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) - \|a\|_2^2 \right\}$$

Results can significantly depend on a kernel hyperparameters



# Example of the decision function



## 1. Supervised Learning

- Sample  $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S$

## 1. Supervised Learning

- Sample  $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S$

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
    - such information is not available during the test phase

## 2. Supervised Learning with Privileged Information (Vapnik, 2009)

- Sample  $S^* = \{(\mathbf{x}_1, \mathbf{x}_1^*, y_1), \dots, (\mathbf{x}_m, \mathbf{x}_m^*, y_m)\}$
- We want to learn  $f : \mathbf{x} \rightarrow y$  using the sample  $S^*$
- Privileged Information:
  - in the form of additional patterns  $\mathbf{x}^*$
  - is not available at the test time
- Example:
  - image classification problem
  - as the privileged information we can use a textual image description
  - such information is not available during the test phase

- Original patterns  $(\mathbf{x}_1, \dots, \mathbf{x}_m) \subset \mathbb{R}^p$
- Additional patterns  $(\mathbf{x}_1^*, \dots, \mathbf{x}_m^*) \subset \mathbb{R}^q$
- We train a decision rule on pairs of patterns  $\{(\mathbf{x}_i, \mathbf{x}_i^*)\}_{i=1}^m \in \mathbb{R}^{p+q}$ ,  
but when making decisions we can use only test patterns  $\mathbf{x} \in \mathbb{R}^p$

- Original patterns  $(\mathbf{x}_1, \dots, \mathbf{x}_m) \subset \mathbb{R}^p$
- Additional patterns  $(\mathbf{x}_1^*, \dots, \mathbf{x}_m^*) \subset \mathbb{R}^q$
- We train a decision rule on pairs of patterns  $\{(\mathbf{x}_i, \mathbf{x}_i^*)\}_{i=1}^m \in \mathbb{R}^{p+q}$ ,  
but when making decisions we can use only test patterns  $\mathbf{x} \in \mathbb{R}^p$

- Original patterns  $(\mathbf{x}_1, \dots, \mathbf{x}_m) \subset \mathbb{R}^p$
- Additional patterns  $(\mathbf{x}_1^*, \dots, \mathbf{x}_m^*) \subset \mathbb{R}^q$
- We train a decision rule on pairs of patterns  $\{(\mathbf{x}_i, \mathbf{x}_i^*)\}_{i=1}^m \in \mathbb{R}^{p+q}$ , but when making decisions we can use only test patterns  $\mathbf{x} \in \mathbb{R}^p$

$$\begin{aligned} R + \frac{1}{m\nu} \sum_{i=1}^m \xi_i &\rightarrow \min_{R, a, \xi} \\ \text{s.t. } \|\phi(\mathbf{x}_i) - a\|_2^2 &\leq R + \xi_i \\ \xi_i &\geq 0 \\ R &\geq 0 \end{aligned}$$

- The slack variables  $\xi_i$  characterizes the distance from the patterns  $\mathbf{x}_i$  to the separating boundary  $\|\phi(\mathbf{x}_i) - a\|_2$
- We assume that using the privileged patterns  $(\mathbf{x}_1^*, \dots, \mathbf{x}_m^*)$  we can refine the location of the separating boundary
- We model a slack variable  $\xi$  as

$$\xi = \xi(\mathbf{x}^*) = (\phi^*(\mathbf{x}^*) \cdot w^*) + b^*,$$

where  $\phi^*(\cdot)$  is a feature map in the space of privileged patterns

$$\begin{aligned} R + \frac{1}{m\nu} \sum_{i=1}^m \xi_i &\rightarrow \min_{R, a, \xi} \\ \text{s.t. } \|\phi(\mathbf{x}_i) - a\|_2^2 &\leq R + \xi_i \\ \xi_i &\geq 0 \\ R &\geq 0 \end{aligned}$$

- The slack variables  $\xi_i$  characterizes the distance from the patterns  $\mathbf{x}_i$  to the separating boundary  $\|\phi(\mathbf{x}_i) - a\|_2$
- We assume that using the privileged patterns  $(\mathbf{x}_1^*, \dots, \mathbf{x}_m^*)$  we can refine the location of the separating boundary
- We model a slack variable  $\xi$  as

$$\xi = \xi(\mathbf{x}^*) = (\phi^*(\mathbf{x}^*) \cdot w^*) + b^*,$$

where  $\phi^*(\cdot)$  is a feature map in the space of privileged patterns

$$\begin{aligned} R + \frac{1}{m\nu} \sum_{i=1}^m \xi_i &\rightarrow \min_{R, a, \xi} \\ \text{s.t. } \|\phi(\mathbf{x}_i) - a\|_2^2 &\leq R + \xi_i \\ \xi_i &\geq 0 \\ R &\geq 0 \end{aligned}$$

- The slack variables  $\xi_i$  characterizes the distance from the patterns  $\mathbf{x}_i$  to the separating boundary  $\|\phi(\mathbf{x}_i) - a\|_2$
- We assume that using the privileged patterns  $(\mathbf{x}_1^*, \dots, \mathbf{x}_m^*)$  we can refine the location of the separating boundary
- We model a slack variable  $\xi$  as

$$\xi = \xi(\mathbf{x}^*) = (\phi^*(\mathbf{x}^*) \cdot w^*) + b^*,$$

where  $\phi^*(\cdot)$  is a feature map in the space of privileged patterns

We incorporate the privileged information

$$\begin{aligned} & \nu m R + \frac{\gamma}{2} \|w^*\|_2^2 \\ & + \sum_{i=1}^m [(w^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i] \rightarrow \min_{R, a, w^*, b, \zeta} \\ & s.t. \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + [(w \cdot \phi^*(\mathbf{x}_i^*))^\top + b^*], \\ & (w^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i \geq 0, \quad \zeta_i \geq 0. \end{aligned}$$

We incorporate the privileged information

$$\begin{aligned} & \nu m R + \frac{\gamma}{2} \|w^*\|_2^2 \\ & + \sum_{i=1}^m [(w^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i] \rightarrow \min_{R, a, w^*, b, \zeta} \\ & s.t. \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + [(w \cdot \phi^*(\mathbf{x}_i^*))^\top + b^*], \\ & (w^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i \geq 0, \quad \zeta_i \geq 0. \end{aligned}$$

We incorporate the privileged information

$$\begin{aligned} & \nu m R + \frac{\gamma}{2} \|w^*\|_2^2 \\ & + \sum_{i=1}^m [(w^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i] \rightarrow \min_{R, a, w^*, b, \zeta} \\ & s.t. \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + [(w \cdot \phi^*(\mathbf{x}_i^*))^\top + b^*], \\ & (w^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i \geq 0, \quad \zeta_i \geq 0. \end{aligned}$$

We incorporate the privileged information

$$\begin{aligned} & \nu m R + \frac{\gamma}{2} \|w^*\|_2^2 \\ & + \sum_{i=1}^m [(\mathbf{w}^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i] \rightarrow \min_{R, a, w^*, b, \zeta} \\ & s.t. \|\phi(\mathbf{x}_i) - a\|_2^2 \leq R + [(\mathbf{w} \cdot \phi^*(\mathbf{x}_i^*))^\top + b^*], \\ & (\mathbf{w}^* \cdot \phi^*(\mathbf{x}_i^*))^\top + b^* + \zeta_i \geq 0, \quad \zeta_i \geq 0. \end{aligned}$$

Let us formulate the dual problem:

$$\begin{aligned} & \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - \frac{1}{2\nu m} \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ & - \sum_{i,j} \frac{1}{2\gamma} (\alpha_i - \delta_i) K^*(\mathbf{x}_i^*, \mathbf{x}_j^*) (\alpha_j - \delta_j) \rightarrow \max_{\alpha, \delta} \\ \text{s.t. } & \sum_{i=1}^m \alpha_i = \nu m, \quad \sum_{i=1}^m \delta_i = \nu m, \quad 0 \leq \delta_i \leq 1, \quad \alpha_i \geq 0. \end{aligned}$$

The **decision function** has again the same form

$$f(\mathbf{x}) = \text{sign} \left\{ R - K(\mathbf{x}, \mathbf{x}) + 2 \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) - \|a\|_2^2 \right\}$$

Let us formulate the dual problem:

$$\begin{aligned} & \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - \frac{1}{2\nu m} \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ & - \sum_{i,j} \frac{1}{2\gamma} (\alpha_i - \delta_i) K^*(\mathbf{x}_i^*, \mathbf{x}_j^*) (\alpha_j - \delta_j) \rightarrow \max_{\alpha, \delta} \\ \text{s.t. } & \sum_{i=1}^m \alpha_i = \nu m, \quad \sum_{i=1}^m \delta_i = \nu m, \quad 0 \leq \delta_i \leq 1, \quad \alpha_i \geq 0. \end{aligned}$$

The **decision function** has again the same form

$$f(\mathbf{x}) = \text{sign} \left\{ R - K(\mathbf{x}, \mathbf{x}) + 2 \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) - \|a\|_2^2 \right\}$$

- Synthetic data
- Real data from the Microsoft Malware Classification Challenge (BIG 2015)

Let us provide an example of synthetic data

- Let  $c_1 = (2, 2)$  and  $c_2 = (-2, -2)$
- $\mathbf{x} \in \mathbb{R}^2$  and is generated from:

$$\mathbf{x} \sim \phi \cdot N(c_1, I) + (1 - \phi) \cdot N(c_2, I), \quad \phi \sim Ber(0.5)$$

- As a privileged information we use coordinates of a pattern after subtracting the nearest mean vector, i.e.

$$\mathbf{x}^* = \mathbf{x} - \arg \min_{c_1, c_2} (\|\mathbf{x} - c_1\|, \|\mathbf{x} - c_2\|)$$

- A uniformly distributed anomalies are generated and added to the sample

Let us provide an example of synthetic data

- Let  $c_1 = (2, 2)$  and  $c_2 = (-2, -2)$
- $\mathbf{x} \in \mathbb{R}^2$  and is generated from:

$$\mathbf{x} \sim \phi \cdot N(c_1, I) + (1 - \phi) \cdot N(c_2, I), \quad \phi \sim Ber(0.5)$$

- As a privileged information we use coordinates of a pattern after subtracting the nearest mean vector, i.e.

$$\mathbf{x}^* = \mathbf{x} - \arg \min_{c_1, c_2} (\|\mathbf{x} - c_1\|, \|\mathbf{x} - c_2\|)$$

- A uniformly distributed anomalies are generated and added to the sample

Let us provide an example of synthetic data

- Let  $c_1 = (2, 2)$  and  $c_2 = (-2, -2)$
- $\mathbf{x} \in \mathbb{R}^2$  and is generated from:

$$\mathbf{x} \sim \phi \cdot N(c_1, I) + (1 - \phi) \cdot N(c_2, I), \quad \phi \sim Ber(0.5)$$

- As a privileged information we use coordinates of a pattern after subtracting the nearest mean vector, i.e.

$$\mathbf{x}^* = \mathbf{x} - \arg \min_{c_1, c_2} (\|\mathbf{x} - c_1\|, \|\mathbf{x} - c_2\|)$$

- A uniformly distributed anomalies are generated and added to the sample

Let us provide an example of synthetic data

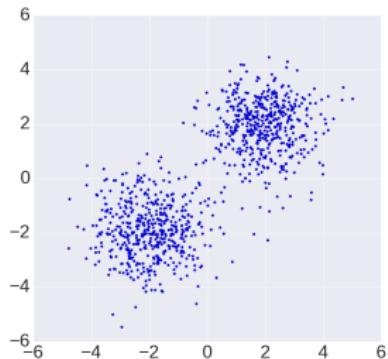
- Let  $c_1 = (2, 2)$  and  $c_2 = (-2, -2)$
- $\mathbf{x} \in \mathbb{R}^2$  and is generated from:

$$\mathbf{x} \sim \phi \cdot N(c_1, I) + (1 - \phi) \cdot N(c_2, I), \quad \phi \sim Ber(0.5)$$

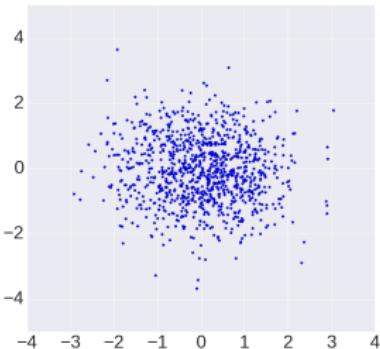
- As a privileged information we use coordinates of a pattern after subtracting the nearest mean vector, i.e.

$$\mathbf{x}^* = \mathbf{x} - \arg \min_{c_1, c_2} (\|\mathbf{x} - c_1\|, \|\mathbf{x} - c_2\|)$$

- A uniformly distributed anomalies are generated and added to the sample



Mixture of Gaussians



Mixture of Gaussians,  
privileged feature space

Accuracy:

- **SVDD**: 0.55
- **SVDD+**: 0.98

- Microsoft Malware Classification Challenge (2015)
  - Nine malware families
  - Raw data contains:
    - the hexadecimal representation of the file's binary content
    - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- Microsoft Malware Classification Challenge (2015)
- Nine malware families
- Raw data contains:
  - the hexadecimal representation of the file's binary content
  - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- Microsoft Malware Classification Challenge (2015)
- Nine malware families
- Raw data contains:
  - the hexadecimal representation of the file's binary content
  - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- Microsoft Malware Classification Challenge (2015)
- Nine malware families
- Raw data contains:
  - the hexadecimal representation of the file's binary content
  - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- Microsoft Malware Classification Challenge (2015)
- Nine malware families
- Raw data contains:
  - the hexadecimal representation of the file's binary content
  - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- Microsoft Malware Classification Challenge (2015)
- Nine malware families
- Raw data contains:
  - the hexadecimal representation of the file's binary content
  - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- Microsoft Malware Classification Challenge (2015)
- Nine malware families
- Raw data contains:
  - the hexadecimal representation of the file's binary content
  - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- Microsoft Malware Classification Challenge (2015)
- Nine malware families
- Raw data contains:
  - the hexadecimal representation of the file's binary content
  - logs containing various metadata information extracted from the binary, such as function calls, strings, etc.
- As the original features we use information, obtained from the binary files
  - frequencies of bytes
  - number of different four-grams, etc.

- As the privileged information we use features, obtained from the assembly code
  - frequencies of each command
  - a number of calls to external dll files
  - transformation of the assembly code to an image
  - we use a feature based on image texture which is commonly used in scene category classification such as coast, mountain, forest, street, etc.
  - Here, instead of scene categories, we have malware families

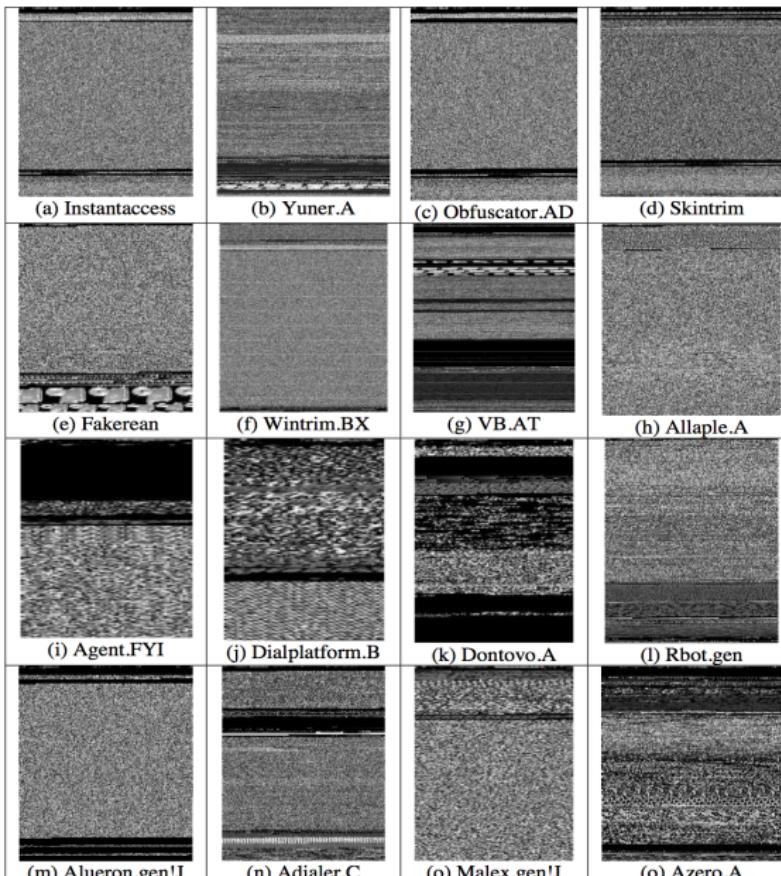
- As the privileged information we use features, obtained from the assembly code
  - frequencies of each command
  - a number of calls to external dll files
  - transformation of the assembly code to an image
  - we use a feature based on image texture which is commonly used in scene category classification such as coast, mountain, forest, street, etc.
- Here, instead of scene categories, we have malware families

- As the privileged information we use features, obtained from the assembly code
  - frequencies of each command
  - a number of calls to external dll files
  - transformation of the assembly code to an image
  - we use a feature based on image texture which is commonly used in scene category classification such as coast, mountain, forest, street, etc.
- Here, instead of scene categories, we have malware families

- As the privileged information we use features, obtained from the assembly code
  - frequencies of each command
  - a number of calls to external dll files
  - transformation of the assembly code to an image
  - we use a feature based on image texture which is commonly used in scene category classification such as coast, mountain, forest, street, etc.
- Here, instead of scene categories, we have malware families

- As the privileged information we use features, obtained from the assembly code
  - frequencies of each command
  - a number of calls to external dll files
  - transformation of the assembly code to an image
  - we use a feature based on image texture which is commonly used in scene category classification such as coast, mountain, forest, street, etc.
- Here, instead of scene categories, we have malware families

# Malware Images of Various Families



- We select one of the nine classes
- As a train set we use half of patterns from the selected class
- As a test set we use patterns from another half of the selected class, as well as patterns from other eight classes
- We consider patterns from the selected class to be “normal” and patterns from other classes as “abnormal”
- We use predictions on the test set to calculate the area under the precision/recall curve, representing the accuracy of anomaly detection

- We select one of the nine classes
- As a train set we use half of patterns from the selected class
- As a test set we use patterns from another half of the selected class, as well as patterns from other eight classes
- We consider patterns from the selected class to be “normal” and patterns from other classes as “abnormal”
- We use predictions on the test set to calculate the area under the precision/recall curve, representing the accuracy of anomaly detection

- We select one of the nine classes
- As a train set we use half of patterns from the selected class
- As a test set we use patterns from another half of the selected class, as well as patterns from other eight classes
- We consider patterns from the selected class to be “normal” and patterns from other classes as “abnormal”
- We use predictions on the test set to calculate the area under the precision/recall curve, representing the accuracy of anomaly detection

- We select one of the nine classes
- As a train set we use half of patterns from the selected class
- As a test set we use patterns from another half of the selected class, as well as patterns from other eight classes
- We consider patterns from the selected class to be “normal” and patterns from other classes as “abnormal”
- We use predictions on the test set to calculate the area under the precision/recall curve, representing the accuracy of anomaly detection

- We select one of the nine classes
- As a train set we use half of patterns from the selected class
- As a test set we use patterns from another half of the selected class, as well as patterns from other eight classes
- We consider patterns from the selected class to be “normal” and patterns from other classes as “abnormal”
- We use predictions on the test set to calculate the area under the precision/recall curve, representing the accuracy of anomaly detection

# Results on Malware Classification Challenge

---

Algorithm/Malware Class	1	2	3	4	5
SVDD	0.67	0.93	0.97	0.69	0.57
SVDD+	<b>0.81</b>	<b>0.95</b>	<b>0.99</b>	<b>0.72</b>	<b>0.57</b>

Algorithm/Malware Class	6	7	8	9
SVDD	0.80	0.82	0.84	0.60
SVDD+	<b>0.81</b>	<b>0.87</b>	<b>0.88</b>	<b>0.62</b>

Table – Malware Detection problem