

# Analysis of Process and biological data using support vector machines



Sankar Mahadevan, PhD student

Supervisor : Dr. Sirish Shah  
Department of Chemical and Materials Engineering  
University of Alberta

# Outline

---

1. Introduction to SVM
2. Fault detection and diagnosis in process data
3. Case studies
  - a) Tennessee Eastman Process
  - b) Rub Detection in rotating machineries
4. Classification of biological data
5. Biomarker identification (Feature selection)
6. Case Studies
7. Concluding Remarks

# Support Vector Machines

---

- SVM is a classifier derived from statistical learning theory by Vapnik and Chervonenkis
- SVMs introduced by Boser, Guyon, Vapnik in 1992
- SVMs are learning systems
  - based on margin maximization: linear SVM
  - in a high dimensional feature space — Kernel function
  - trained with a learning algorithm from optimization theory - Lagrange
  - have a very good generalization performance

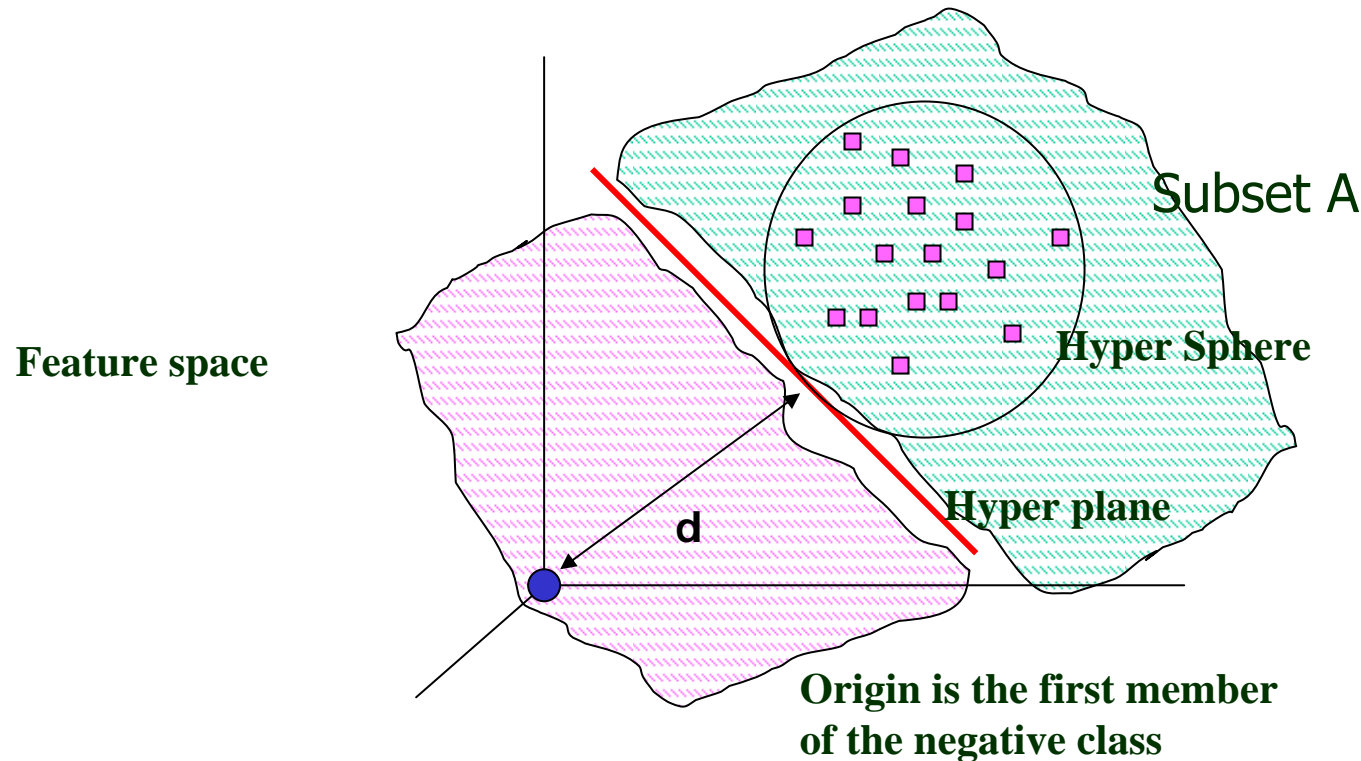
# One-class SVM

---

- One-class classification
  - Samples from just one class available
  - Boundary or 'support' of these samples computed
  - Detect outliers and/or faults: Samples lying outside this boundary
- Parameters involved
  - $\nu$ : measure of fraction of false alarms
  - $\sigma$ : rbf kernel parameter
- Fault detection model
  - High sensitivity (high fault detection rate)
  - High specificity (low false alarms)
  - Minimize detection latency
  - Generalize well to any test data
  - Sensitive to all possible faults

# One-class SVM Objectives

- Estimate the mapping function  $F$  in feature space which returns a positive value on subset  $A$  and a negative value for any value outside the subset.
- Maximize distance of hyperplane from origin

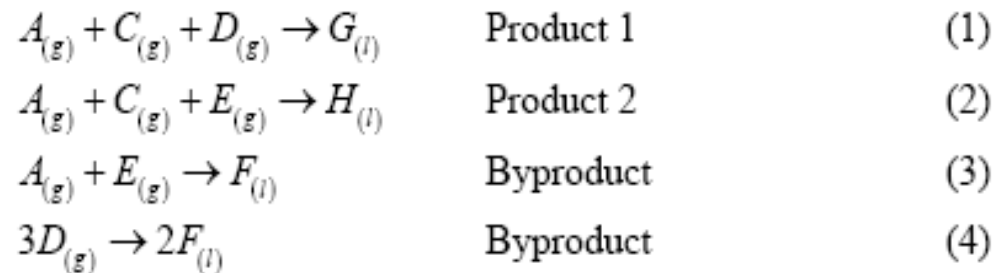
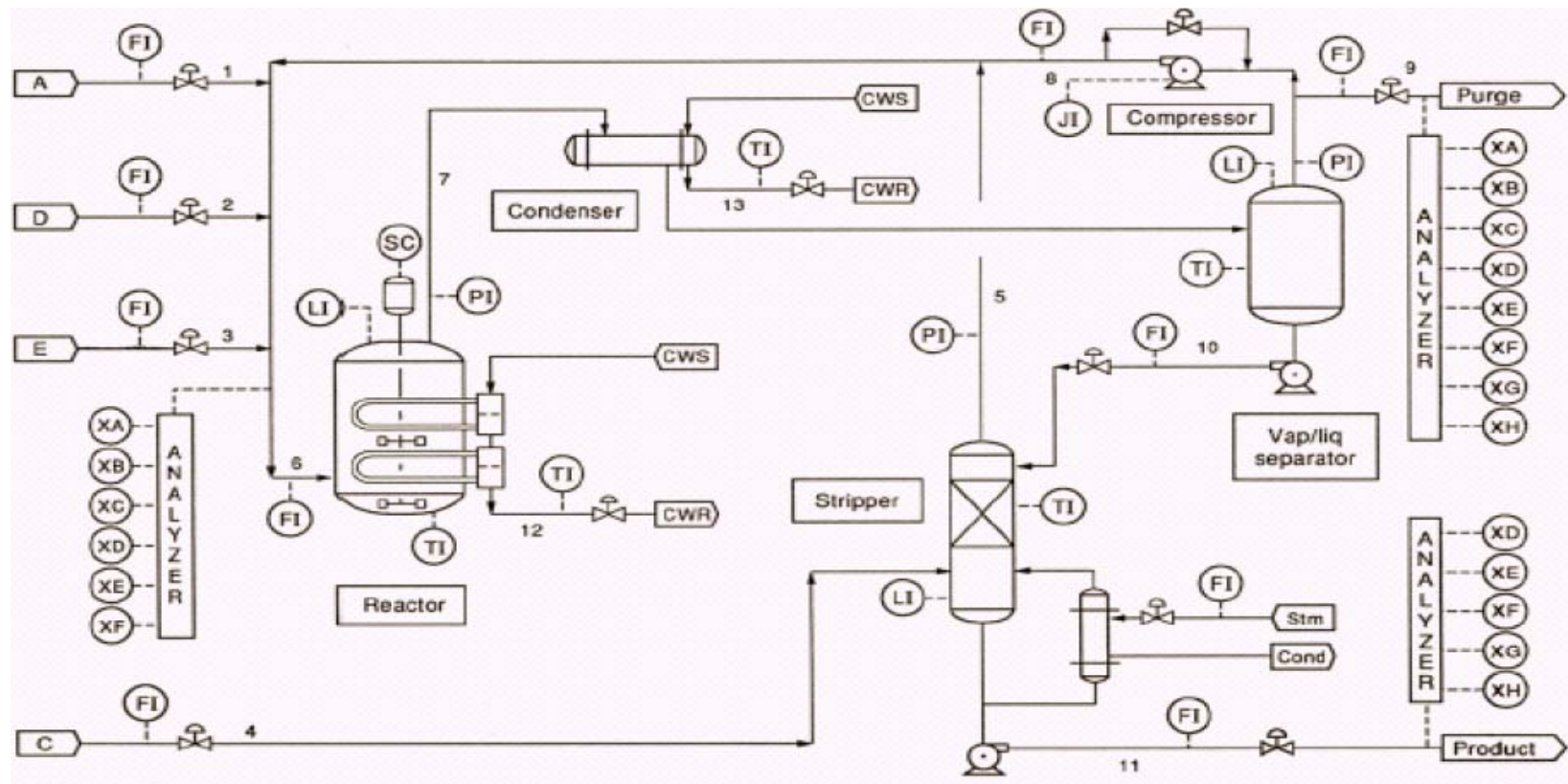


# Fault detection in Process data

---

- Fault detection
  - Detection of abnormal process behaviour
  - Done using a one-class SVM
    - Analogous to the PCA  $T^2$  and Q statistic
  - Can also be done using SVM models in the presence of fault database
- Fault identification
  - Identification of variables relevant to fault
  - Can be done using Support Vector machine-Recursive feature elimination (SVM-RFE)
    - Analogous to contribution chart (Miller)
- Root cause diagnosis
- Process recovery

# Case study 1: Tennessee Eastman process



# Tennessee Eastman process contd...

---

- Each sample consists of 52 features
  - 22 continuous process measurements
  - 11 manipulated variables
  - 19 composition measurements
- Simulated data collected at 3 min sampling interval
- 21 types of pre-defined programmed faults
- Training and Test data for normal operation and faults
  - Training set: 480 samples
  - Test set: 960 samples
    - Fault induced after 160 samples

# TEP: 21 different faults

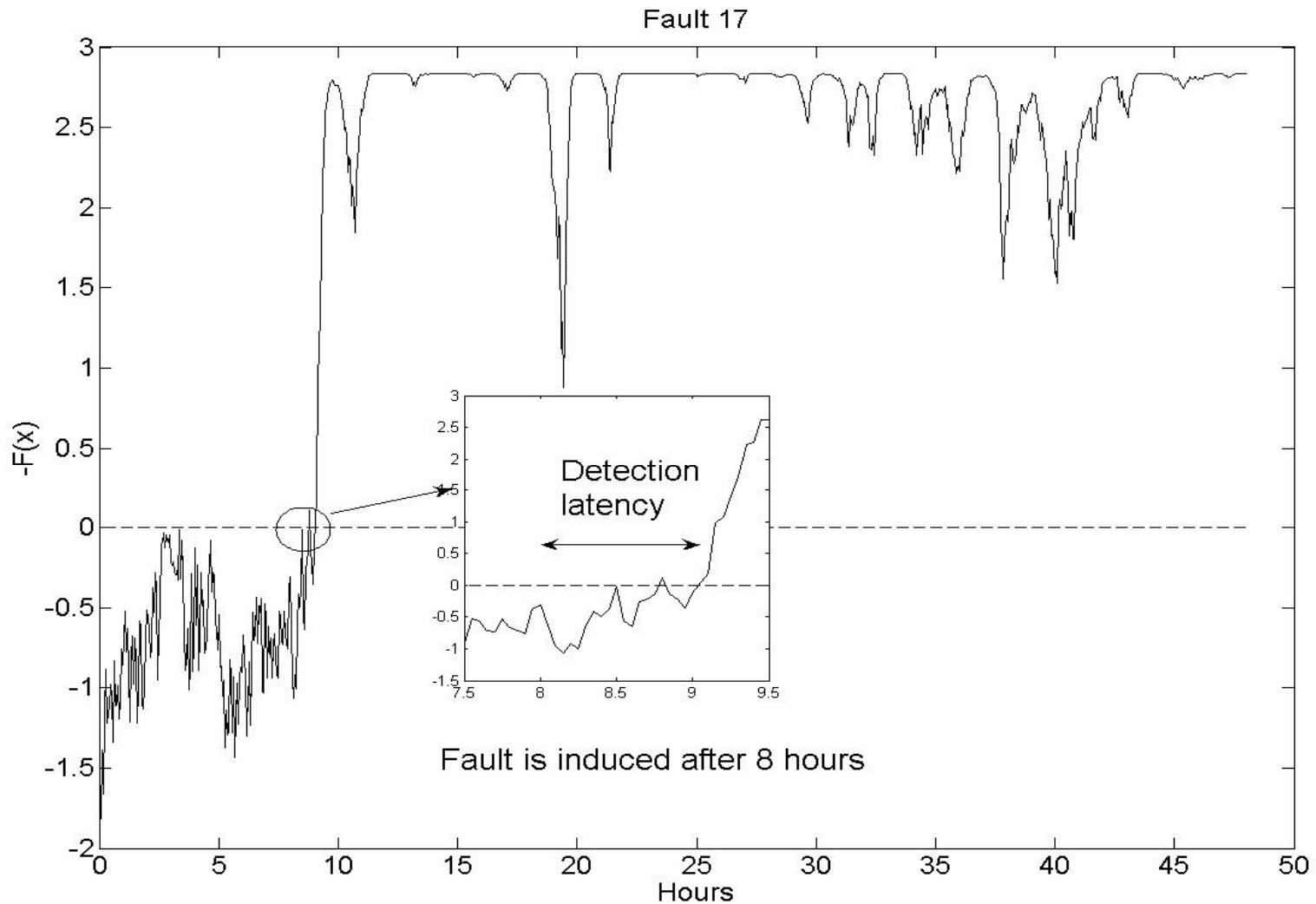
No.	Description	Type
1	<i>A/C</i> feed ratio, <i>B</i> composition constant (stream 4)	Step
2	<i>B</i> composition, <i>A/C</i> ratio constant (stream 4)	Step
3	<i>D</i> feed temperature (stream 2)	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser cooling water inlet temperature	Step
6	<i>A</i> feed loss (stream 1)	Step
7	<i>C</i> header pressure loss – reduced availability (stream 4)	Step
8	<i>A</i> , <i>B</i> , <i>C</i> feed composition (stream 4)	Random variation
9	<i>D</i> feed temperature (stream 2)	Random variation
10	<i>C</i> feed temperature (stream 4)	Random variation
11	Reactor cooling water inlet temperature	Random variation
12	Condenser cooling water inlet temperature	Random variation
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16	Unknown	
17	Unknown	
18	Unknown	
19	Unknown	
20	Unknown	
21	The valve for Stream 4 was fixed at the steady-state position	Constant Position

# Tennessee Eastman process contd...

---

- Algorithm compared with conventional techniques: PCA and DPCA
- Fault detection: performance measures
  - Detection delay
  - False alarm rates
    - Computed on the normal test data
  - Fault detection rates
    - Fraction of the faulty test data exceeding the threshold
- 1-class SVM decision function used as threshold
  - $w \cdot x + b = F(x) > 0$  : sample belongs to normal operating condition
  - $< 0$  : sample is outside normal regime

# TEP: Detection latency



# Comparison of detection latency (minutes)

Fault	PCA-T <sup>2</sup>	PCA-Q	DPCA-T <sup>2</sup>	DPCA-Q	1-class SVM
1	21	9	18	15	6
2	51	36	48	39	33
4	-	9	453	3	3
5	48	3	6	6	3
6	30	3	33	3	3
7	3	3	3	3	3
8	69	60	69	63	63
10	288	147	303	150	33

# Comparison of False alarm rates

Method	Training set (%)	Test set (%)
PCA $T^2$	0.2	1.4
PCA Q	0.4	1.6
DPCA $T^2$	0.2	0.6
DPCA Q	0.4	28.1
1-class SVM	0.0	0.0

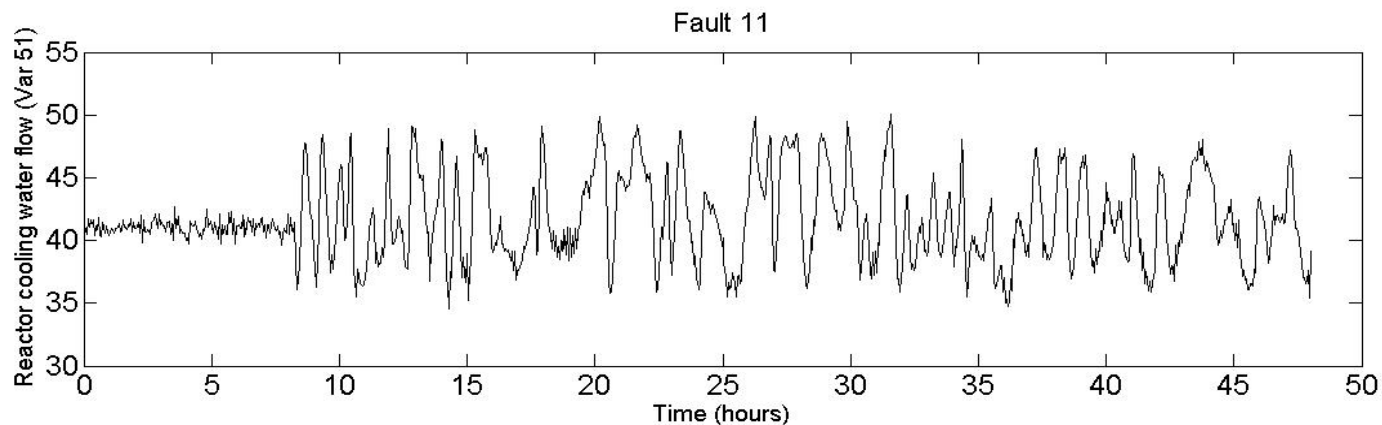
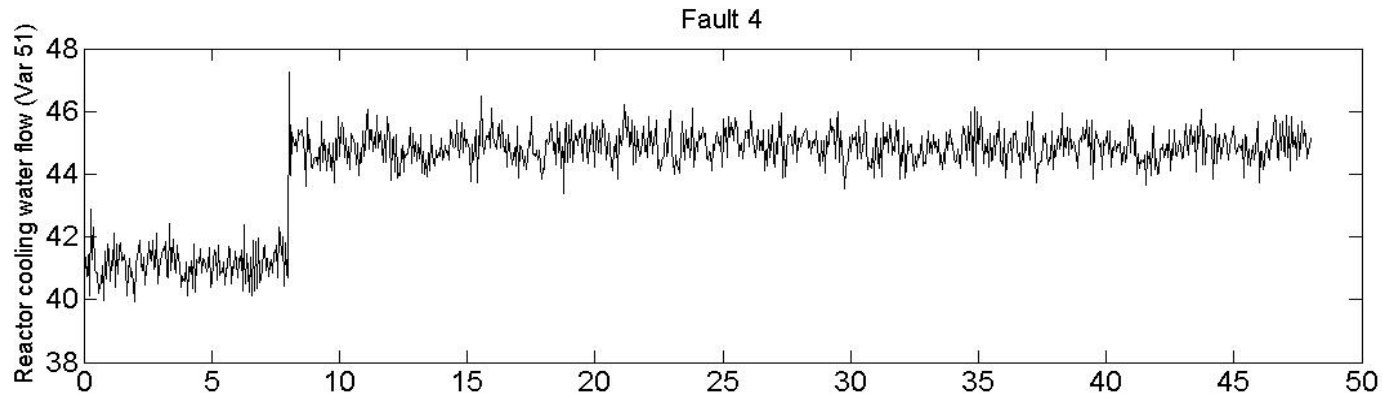
# Comparison of Fault detection rates

Fault	PCA-T <sup>2</sup> (%)	PCA-Q (%)	DPCA-T <sup>2</sup> (%)	DPCA-Q (%)	1-class SVM(%)
1	99.2	99.7	99.4	99.5	99.9
2	98	98.6	98.1	98.5	98.8
4	4.4	96.2	6.1	100	98.4
5	22.5	25.4	24.2	25.2	100
6	98.9	100	98.7	100	100
7	91.5	100	84.1	100	100
8	96.6	97.6	97.2	97.5	97.8
10	33.4	34.1	42	33.5	68.8

- Both normal and faulty data are required
- Diagnosis can be done using SVM-RFE
  - Identifying the best subset of features that results in maximum classification accuracy rate
  - Essentially feature selection process
  - Similar to contribution plots of PCA, PLS

# TEP: Fault diagnosis contd...

- Two faults with similar root cause:
  - **Fault 4:** Step disturbance in reactor cooling water inlet temperature
  - **Fault 11:** Random disturbance in reactor cooling water inlet temperature

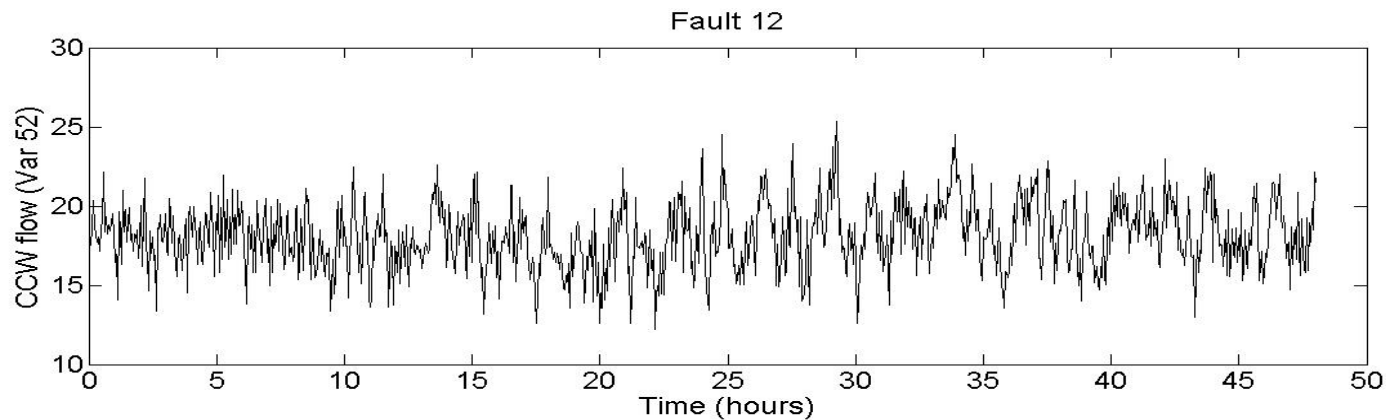
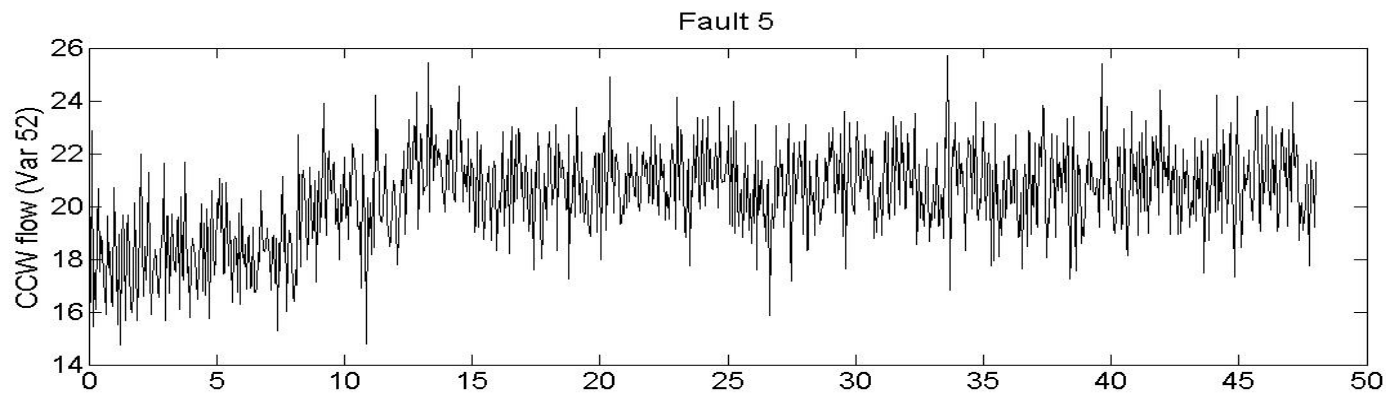


## TEP: Fault diagnosis contd...

Fault	PCA-contribution	SVM-RFE
4	51 (RCW flow) 21 (RCW Temp) 9 (React. Temp)	51 9
11	51 21	51 17 (Product flow)

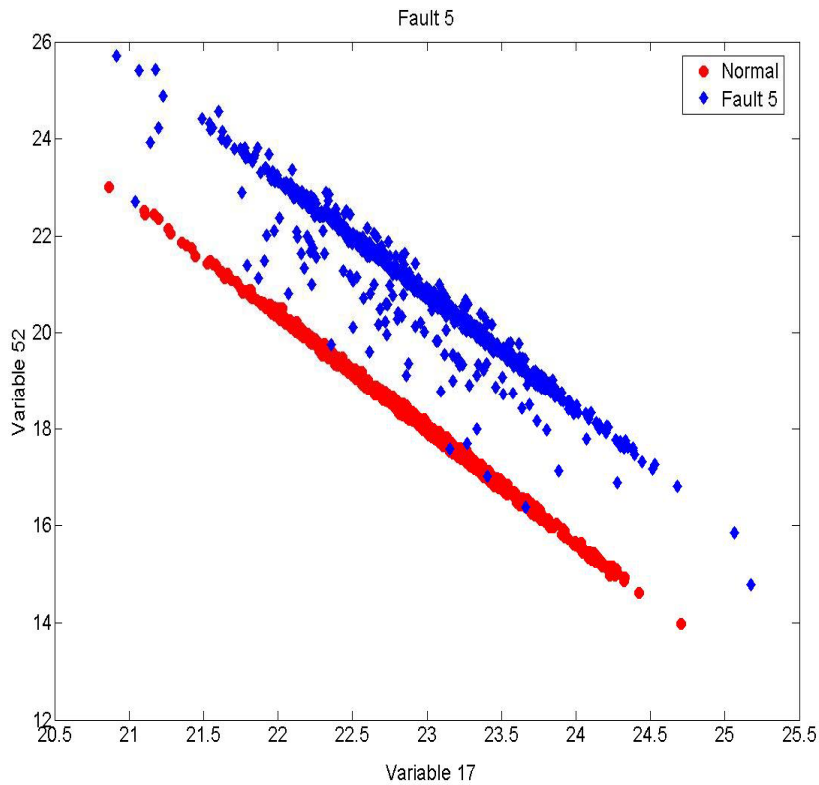
# TEP: Fault diagnosis contd...

- Two faults with similar root cause:
  - **Fault 5:** Step disturbance in condenser cooling water inlet temperature
  - **Fault 12:** Random disturbance in condenser cooling water inlet temperature

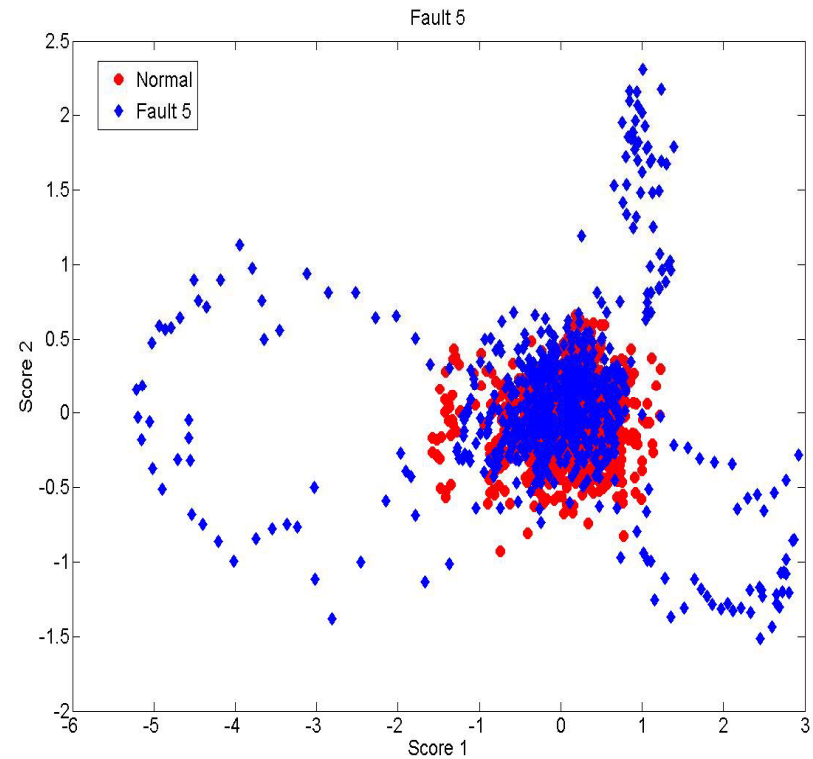


## TEP: Fault diagnosis contd...

Fault	PCA-contribution	SVM-RFE
5	11, 9, 35, 22, 18	52 and 17
12	11, 37, 22, 4	20, 46, 18, 50



SVM-RFE



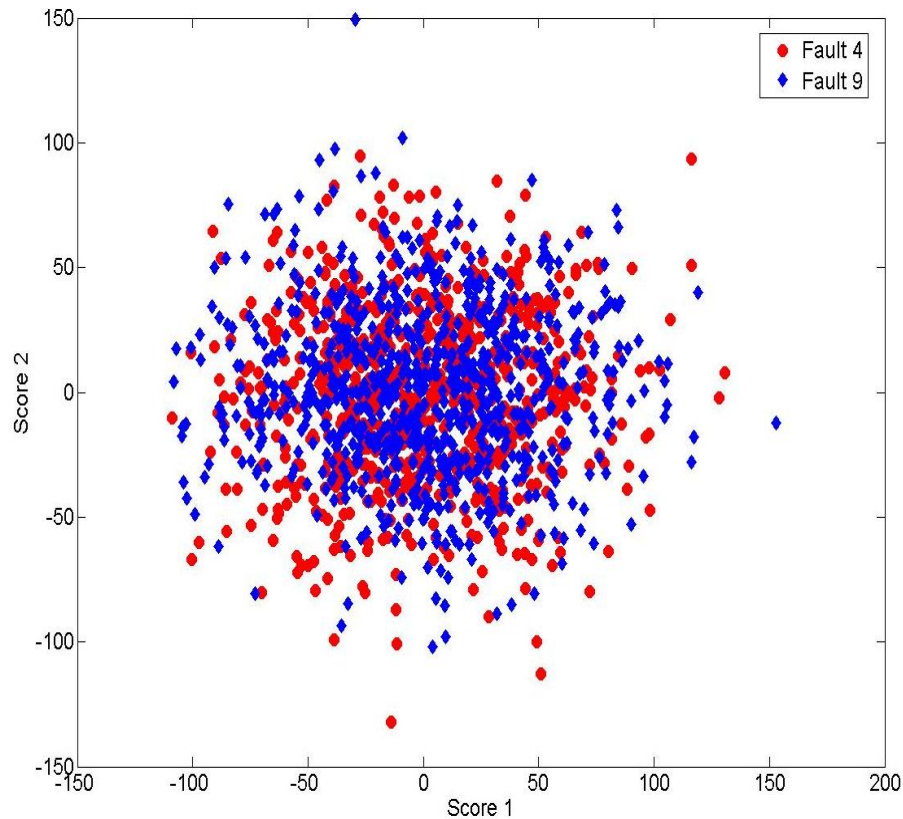
PCA-contribution

# TEP: Classification of faults

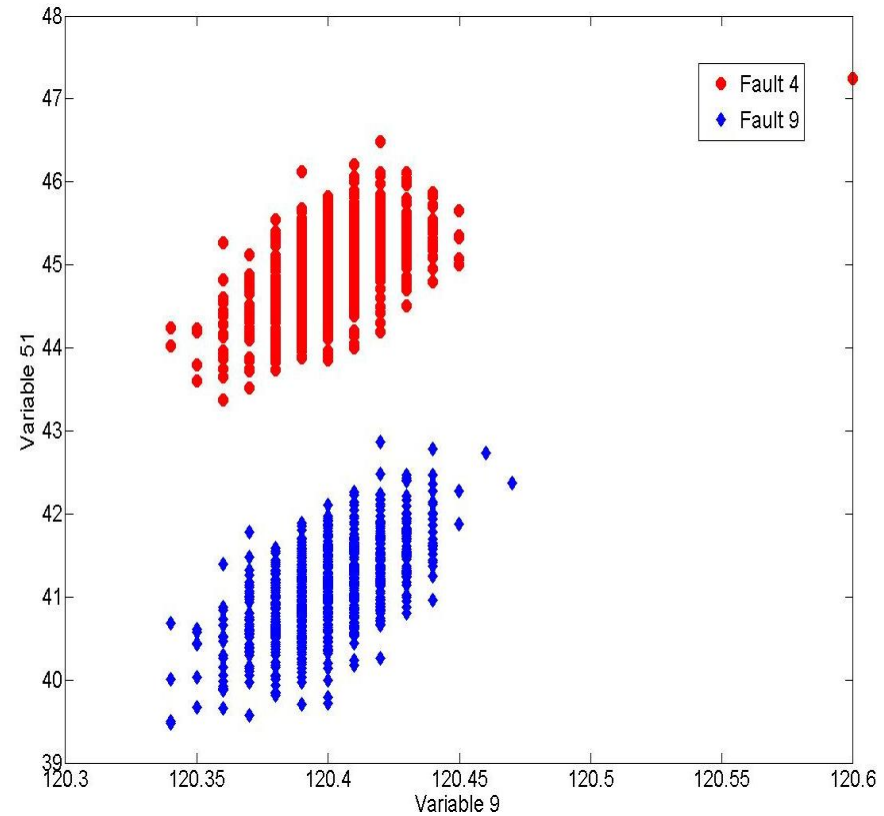
---

- SVMs can be used to build models for different faults
- Classification of faults 4 and 9
  - Fault 4: Step change in reactor cooling water inlet temperature
  - Fault 9: Random variation in D feed temperature
- 480 training data for each class
- 800 test data for each class
- 100% accuracy rate achieved using a rbf kernel
- Fault diagnosis done using SVM-RFE
  - Variables 9 and 51 were found to be important

# TEP: Classification of faults contd...

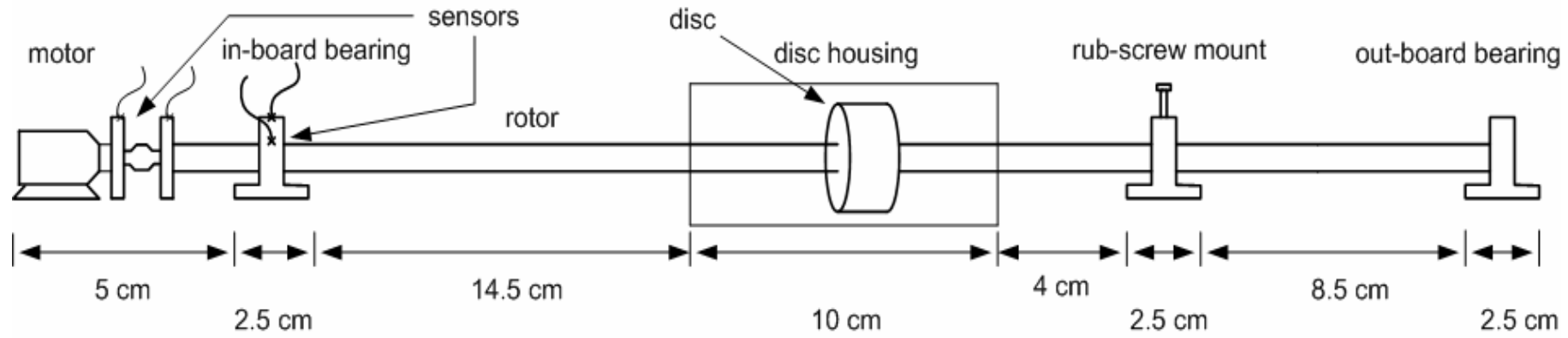


Before feature selection

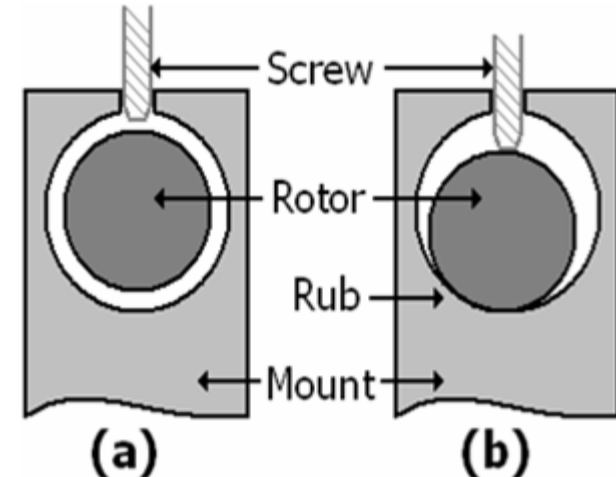


After feature selection

# Detection of rub in rotating machineries



- Detection of rub between rotor and stator
- Simulated rub-impact data from jeffcott rotor model
- Displacement of disc center in x and y direction measured
- 3 levels of rub
  - None :  $\omega = 500$  rpm
  - Mild :  $\omega = 562.5$  rpm
  - Severe :  $\omega = 625$  rpm

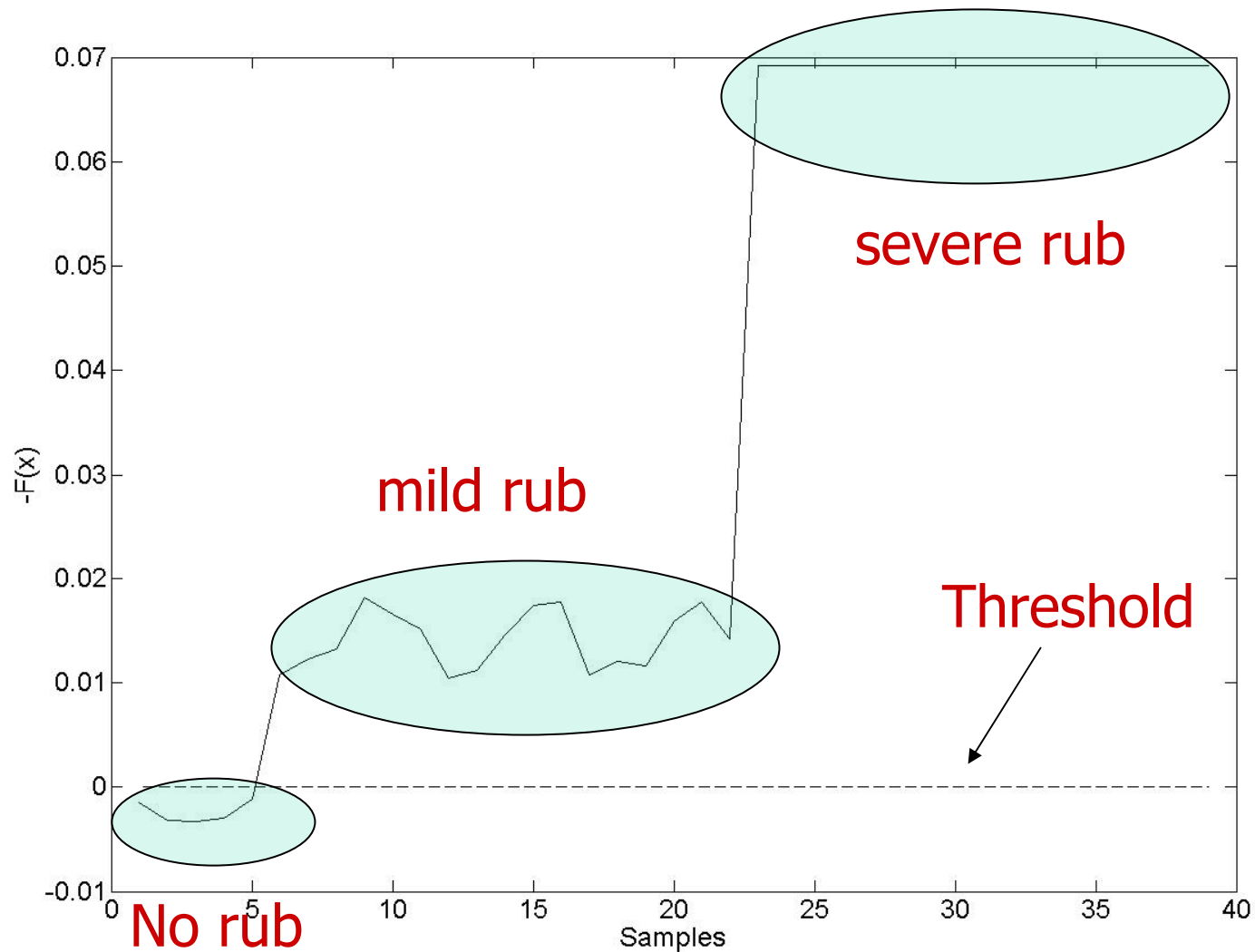


# Rub detection: 1- class SVM

---

- Simulation carried out for 3 minutes generating a 17000 sample time series for each kind of rub.
- The time series is split using non-overlapping window of 1000 points to generate 17 samples for each kind of rub data
- Temporal data converted to spectral domain using FFT
- Rub detection is done using 1-class SVM
  - Model is built based on the 'no rub' data
  - The 'mild rub' and 'severe rub' data is projected on to this model
- Training data
  - 12 'no rub' data
- Test data
  - 5 no rub
  - 17 mild rub
  - 17 severe rub

# Rub detection: 1- class SVM

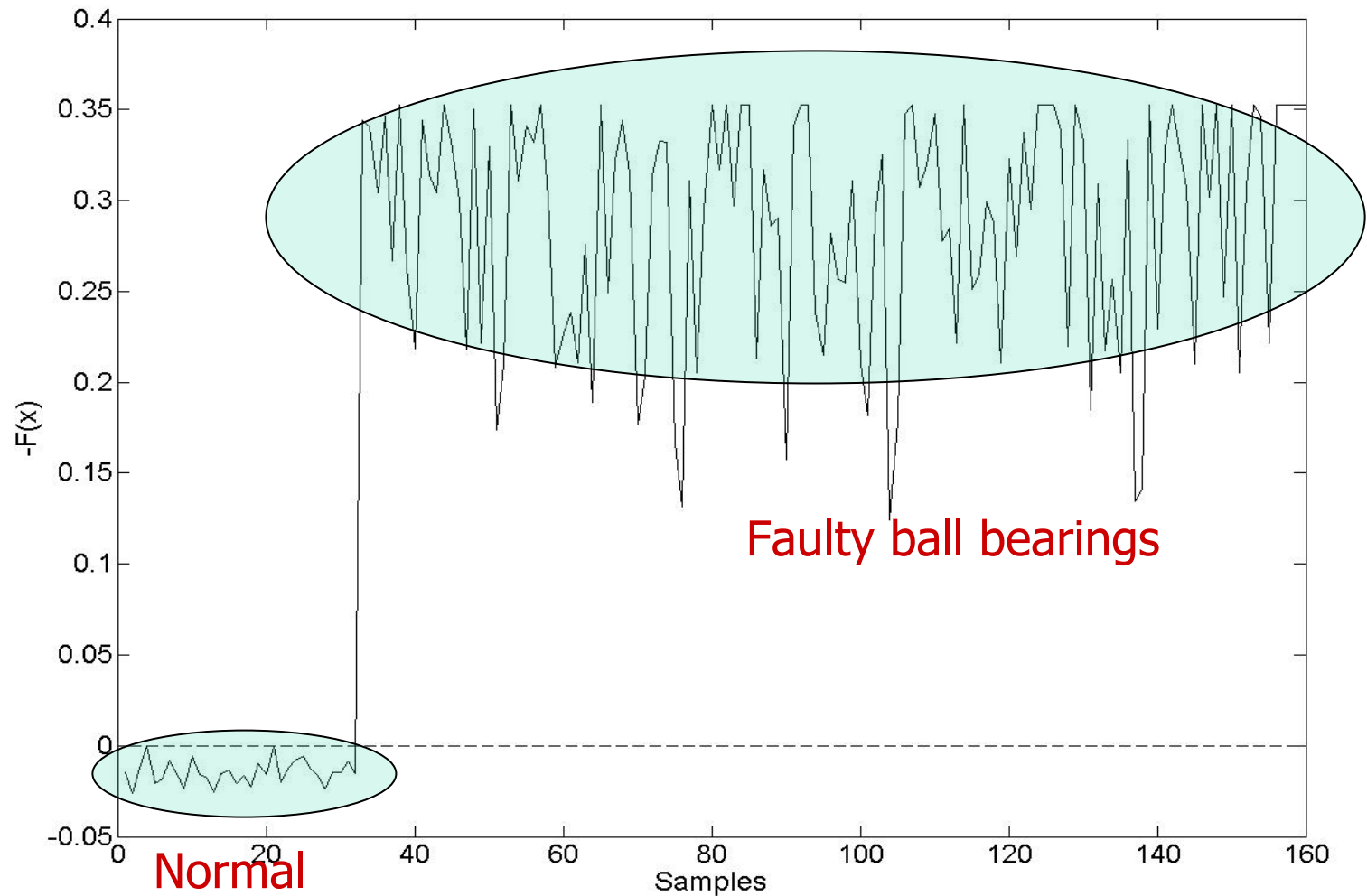


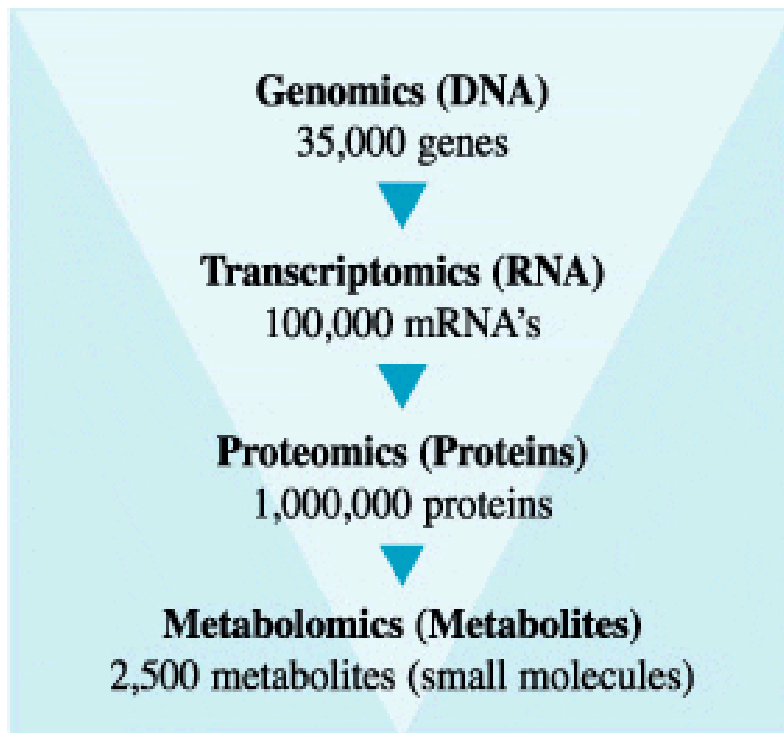
# Fault detection in Ball bearings

---

- Ball Bearing Dataset
  - 6 different categories
    - No fault
    - Outer cage completely broken
    - Broken cage with one loose element
    - Damaged cage
    - 4 loose elements
    - Badly worn ball bearing
  - Each sample has 26 features.
  - Training data
    - 56 'no fault' samples
  - Test data
    - 32 'no fault' samples
    - 128 samples belonging to any one of the 5 categories
- Vibrational time series data converted to power spectrum

# Fault detection using 1-class SVM

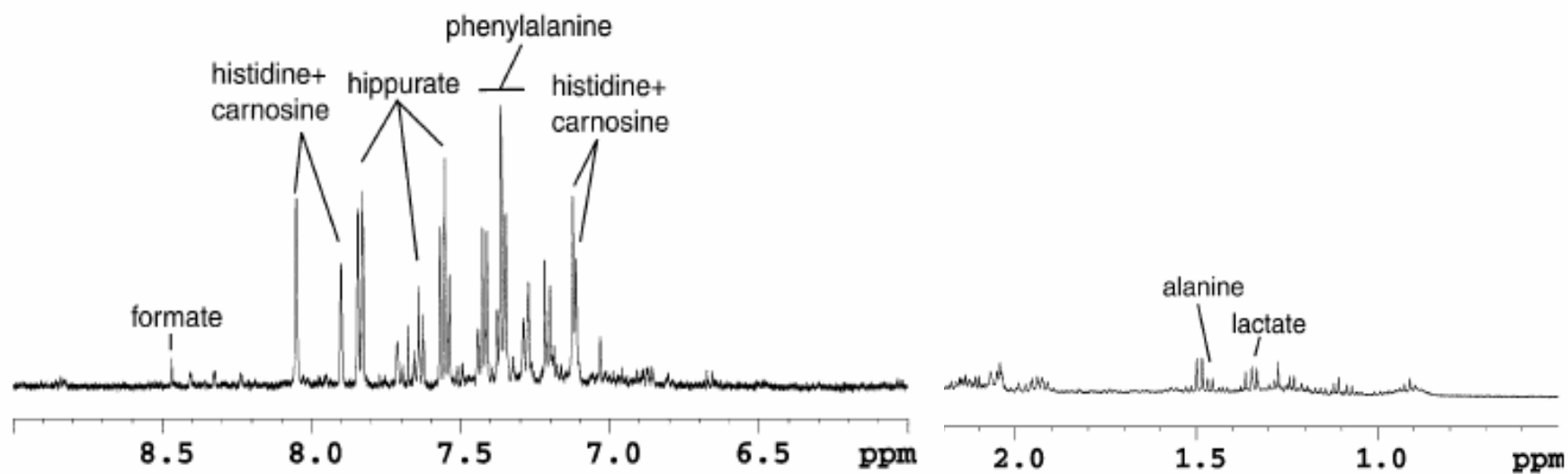
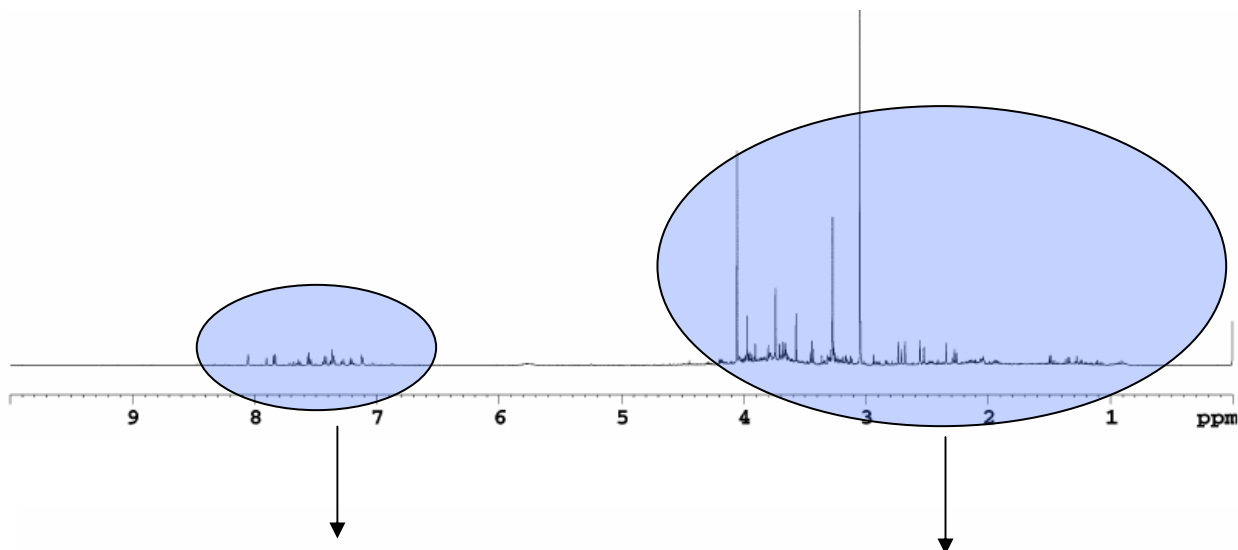




- Metabolomics: defined as 'quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification'.
- Metabolomic analysis aims at quantifying all metabolites at the organ, tissue, cellular or even at subcellular level.

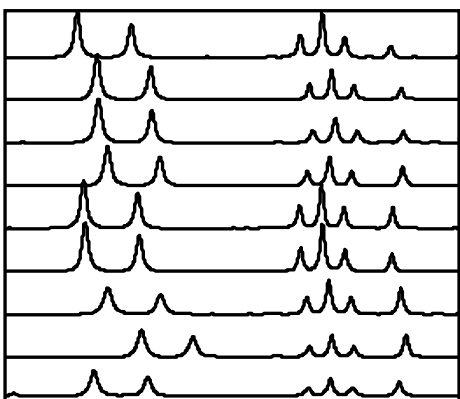
- Disease diagnostics

# NMR Spectroscopy



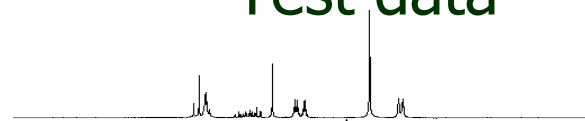
# NMR data analysis: Machine Learning

Training data



(learning machine)

Test data



Model

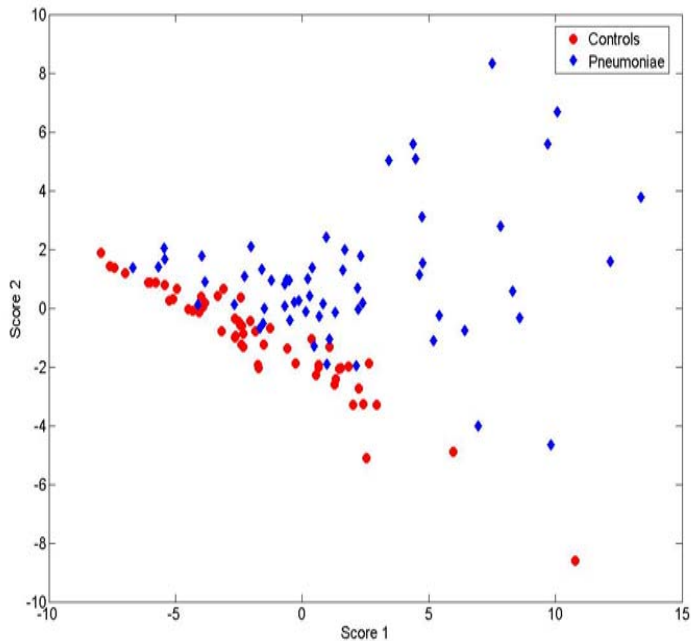
Prediction

Examples of learning algorithms are:

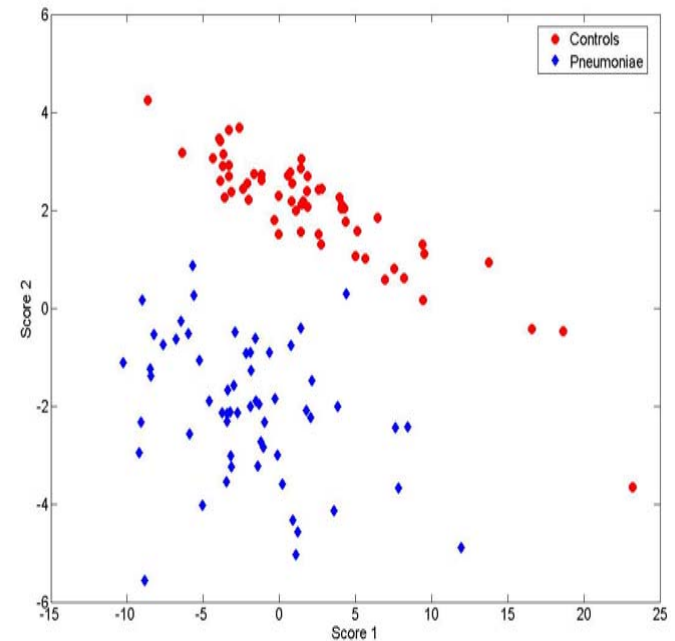
- Support Vector Machines (SVM)
- Artificial Neural Networks
- Hidden Markov Models

# Transformations

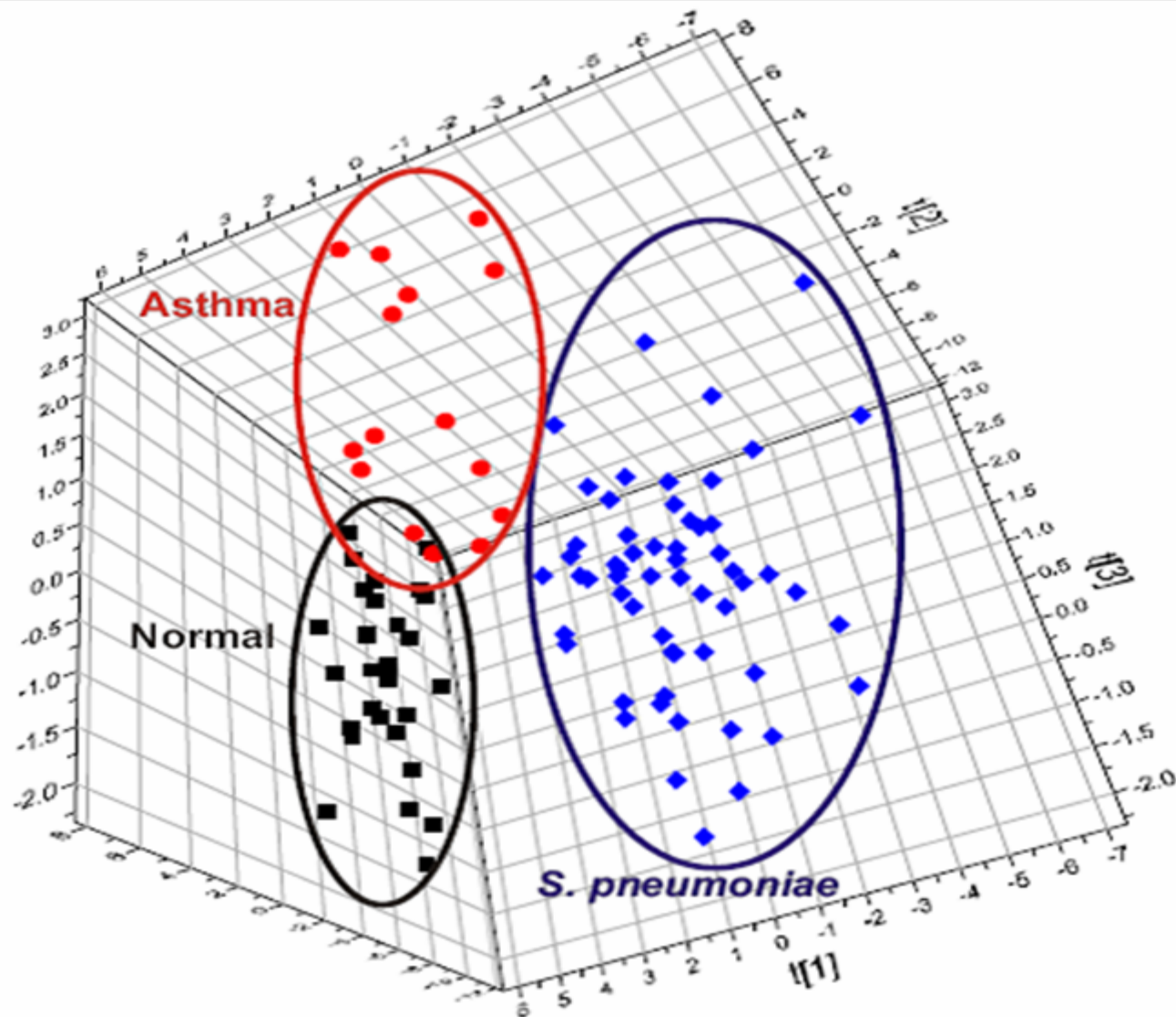
- Transformations can make the data more interpretable



log →



# 3-way classification of metabolomic data



# Normal-Pneumonia classification

Feature selection method	No. of features selected	Acc. Rate (4-fold)
Fisher Score	10	95.20
SVM-RFE	8	97.89

Feature index	Feature (metabolite)	Frequency (100 iterations)
31	Glutamate	87
52	O-acetylcarnitine	69
12	Acetoacetate	63
51	N-acetylglycine	63
24	Creatinine	59
22	Citrate	53
67	trigonelline	49
29	Fumarate	40

## Case study 2: gender based classification of normals

---

- Spectral Data consists of
  - 162 male, 194 female samples
  - 360 bins each 0.025 ppm width

Method	Kernel type	$\gamma$	C	d	Acc. Rate (4-fold)	Acc. Rate (LOOCV)
SVM	poly	$2^{-1}$	$2^{11}$	1	83.65	86.80
SVM	rbf	$2^{-9}$	$2^{15}$		83.22	86.23

# Biomarker identification

<b>Feature index</b>	<b>Feature (metabolite)</b>	<b>Frequency (100 iterations)</b>
19	Carnitine	98
22	Creatine	90
21	Citrate	88
23	Creatinine	63
51	Suberate	53
8	4-aminobutyrate	53
52	Succinate	52
61	Tryptophan	49

# Conclusions

---

1. Fault detection and diagnosis using 1-class SVM was demonstrated
2. Classification of faults can be done by building appropriate SVM models
3. SVM was effectively applied to detect Rub in rotating machineries and faults in ball bearings
4. Support vector machines have been effectively used for classifying metabolomic data to build predictive models which can potentially be used in disease diagnostics.
5. Distinguishing biomarkers are identified using appropriate feature selection to get a deeper insight into the physiological mechanism of the process.

# Acknowledgement

---

1. Dr. Sirish L. Shah
2. Dr. Carolyn Slupsky
3. CPC Group Members
4. NSERC