Mining logs to predict system errors

prof. Barbara Russo
Free University of Bozen-Bolzano, Italy
SwSE - Software and Systems Engineering

Masaryk University, November 10th 2016

Today systems

- Computational power increases
  - (2015) Tianhe-2, China - 3,120,000 cores
- The larger the system, the more frequent critical events
  - Lower overall system utilisation
  - Hardware **failure**, software **failure**, and user **errors**

Manage failures ...

- Crashes
  - immediately stop the system
  - easily identifiable (e.g., disk failure)
  - but can originate a large number of events spread across components
- Deviations from the expected output
  - let the system run
  - reveal only at completion of system tasks

To better manage failures …

… we need information on **system behaviour**
and make failure predictions

Systems generate big data

- Part of such data traces the **change in behaviour** of the system and its sub-components
- Logging services store state changes of a system in archives, **logs**

Mining logs

- How can we exploit log data to model and predict system behaviour?

Log events

- A log event represents a **change in a system state**

xml log event

```
<Info
   TimeStamp="2015-04-08T07:32:37.345"
   File="XXX.Control.ObservingModes.ObservingModeBaseImpl"
   Line="231" Routine="beginSubscan"
   Host="XXX01"
   Process="XXX/javaContainer"
   SourceObject="XXX/Array005"
   Thread="RequestProcessor-35023"
   LogId="343355"
   Audience="Operator">
<![CDATA[Text message … here]]>
</Info>
```

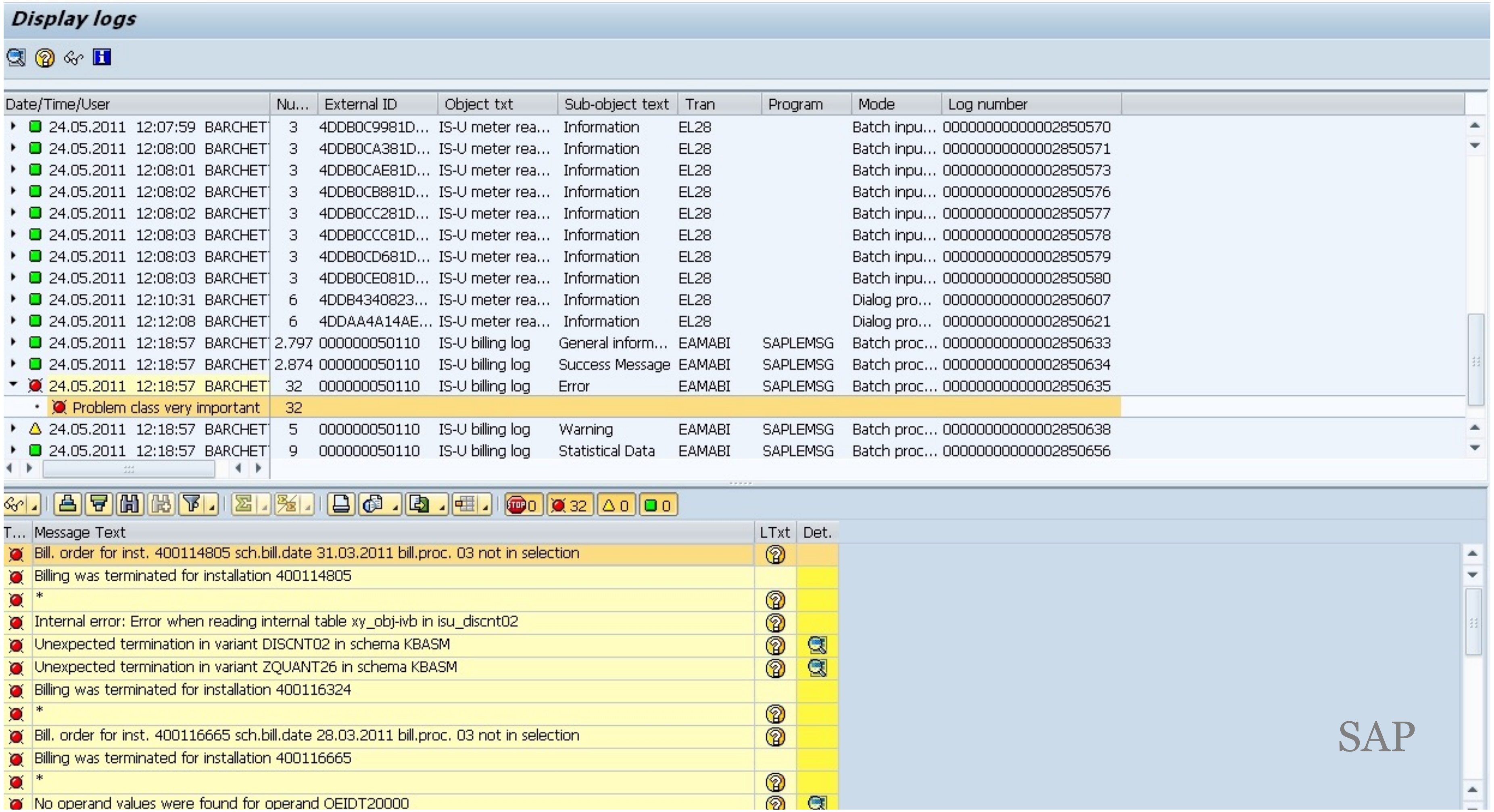
System misbehaviour

- Some events can tell about undesirable system behaviour

Error events act as alerts

- Events in error state (error events) act as **alerts of system failures**:
  - Interpretation of event data might be hard
  - Originated from a series of preceding events

Logs can be cryptic



Display logs

Date/Time/User	Nu...	External ID	Object txt	Sub-object text	Tran	Program	Mode	Log number
24.05.2011 12:07:59 BARCHET	3	4DDB0C9981D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850570
24.05.2011 12:08:00 BARCHET	3	4DDB0CA381D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850571
24.05.2011 12:08:01 BARCHET	3	4DDB0CAE81D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850573
24.05.2011 12:08:02 BARCHET	3	4DDB0CB881D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850576
24.05.2011 12:08:02 BARCHET	3	4DDB0CC281D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850577
24.05.2011 12:08:03 BARCHET	3	4DDB0CCC81D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850578
24.05.2011 12:08:03 BARCHET	3	4DDB0CD681D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850579
24.05.2011 12:08:03 BARCHET	3	4DDB0CE081D...	IS-U meter rea...	Information	EL28		Batch inpu...	00000000000002850580
24.05.2011 12:10:31 BARCHET	6	4DDB4340823...	IS-U meter rea...	Information	EL28		Dialog pro...	00000000000002850607
24.05.2011 12:12:08 BARCHET	6	4DDAA4A14AE...	IS-U meter rea...	Information	EL28		Dialog pro...	00000000000002850621
24.05.2011 12:18:57 BARCHET	2.797	000000050110	IS-U billing log	General inform...	EAMABI	SAPLEMSG	Batch proc...	00000000000002850633
24.05.2011 12:18:57 BARCHET	2.874	000000050110	IS-U billing log	Success Message	EAMABI	SAPLEMSG	Batch proc...	00000000000002850634
24.05.2011 12:18:57 BARCHET	32	000000050110	IS-U billing log	Error	EAMABI	SAPLEMSG	Batch proc...	00000000000002850635
Problem class very important	32							
24.05.2011 12:18:57 BARCHET	5	000000050110	IS-U billing log	Warning	EAMABI	SAPLEMSG	Batch proc...	00000000000002850638
24.05.2011 12:18:57 BARCHET	9	000000050110	IS-U billing log	Statistical Data	EAMABI	SAPLEMSG	Batch proc...	00000000000002850656

0 32 0 0

T...	Message Text	LTxt	Det.
	Bill. order for inst. 400114805 sch.bill.date 31.03.2011 bill.proc. 03 not in selection		
	Billing was terminated for installation 400114805		
	*		
	Internal error: Error when reading internal table xy_obj-ivb in isu_discnt02		
	Unexpected termination in variant DISCNT02 in schema KBASM		
	Unexpected termination in variant ZQUANT26 in schema KBASM		
	Billing was terminated for installation 400116324		
	*		
	Bill. order for inst. 400116665 sch.bill.date 28.03.2011 bill.proc. 03 not in selection		
	Billing was terminated for installation 400116665		
	*		
	No operand values were found for operand OEIDT20000		

SAP

Interpretation

YY-MM-DD-HH:MM:SS NULL ZZZ MYHOST FAILURE xxx exited normally with exit code 0

- Failure, but the program exited cleanly

Interpretation

- If the system administrator was doing **maintenance** on the machine, this message is a harmless artefact of his actions
- If it was generated during **normal machine operation**, this message indicates that all running jobs on the computer were undesirably killed

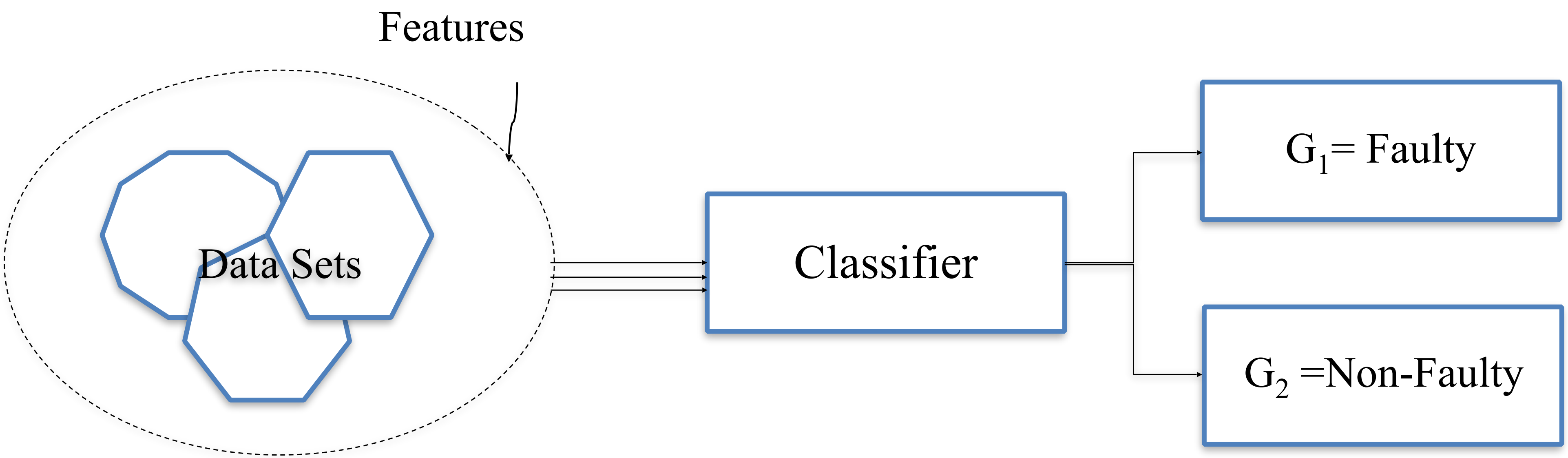
Operational context

- We need to understand the operational context

Originated from a series of preceding events

- System changes can have introduced errors **much earlier** than an error manifests in logs

The classification problem



Sequences

- **Event sequence:** set of events ordered by their timestamp occurring within a given time window
- A **sequence abstraction** is a representation of such sequence (e.g., vector) that can be used to feed classifiers (**features**)

Building features

A. Isolating sequences
   - Identify sequence length
   - Characterise sequence information
B. Build sequence abstraction (e.g., a vector)
C. Build features

Isolating sequences

Date	Server ID	PC ID	User ID	State	Type	Event Description
2009-03-02 07:05:45	1472	36248	26209	Information	Log In	Application LogOn
2009-03-02 07:05:46	1472	36248	26209	Timer	Systems	Application Connection Init.
2009-03-02 07:06:45	1472	26210	1863	Information	Log In	Time Stamp
2009-03-02 07:06:45	1472	26210	1863	Information	General	Generic Information
2009-03-02 07:10:20	1472	5776	19039	Error	General	Generic Error
2009-03-02 07:14:58	1472	5776	19039	Error	Performance	Generic Error

Seq.	Ev.	Date	User	State	Type
s_1	e_1	2009-03-02 07:05:45	26209	Information	Log In
	e_2	2009-03-02 07:05:46	26209	Timer	Systems
s_2	f_1	2009-03-02 07:06:45	1863	Information	Log In
	f_2	2009-03-02 07:06:45	1863	Information	General
	f_3	2009-03-02 07:10:20	19039	Error	General
	f_4	2009-03-02 07:14:58	19039	Error	Perform.

Different length, different types

Sequence abstraction

- μ_i – number of the events of type i in a sequence (multiplicity)
- $sv=[\mu_1, \ldots, \mu_n]$ – vector of event multiplicities

Example – sequence abstraction

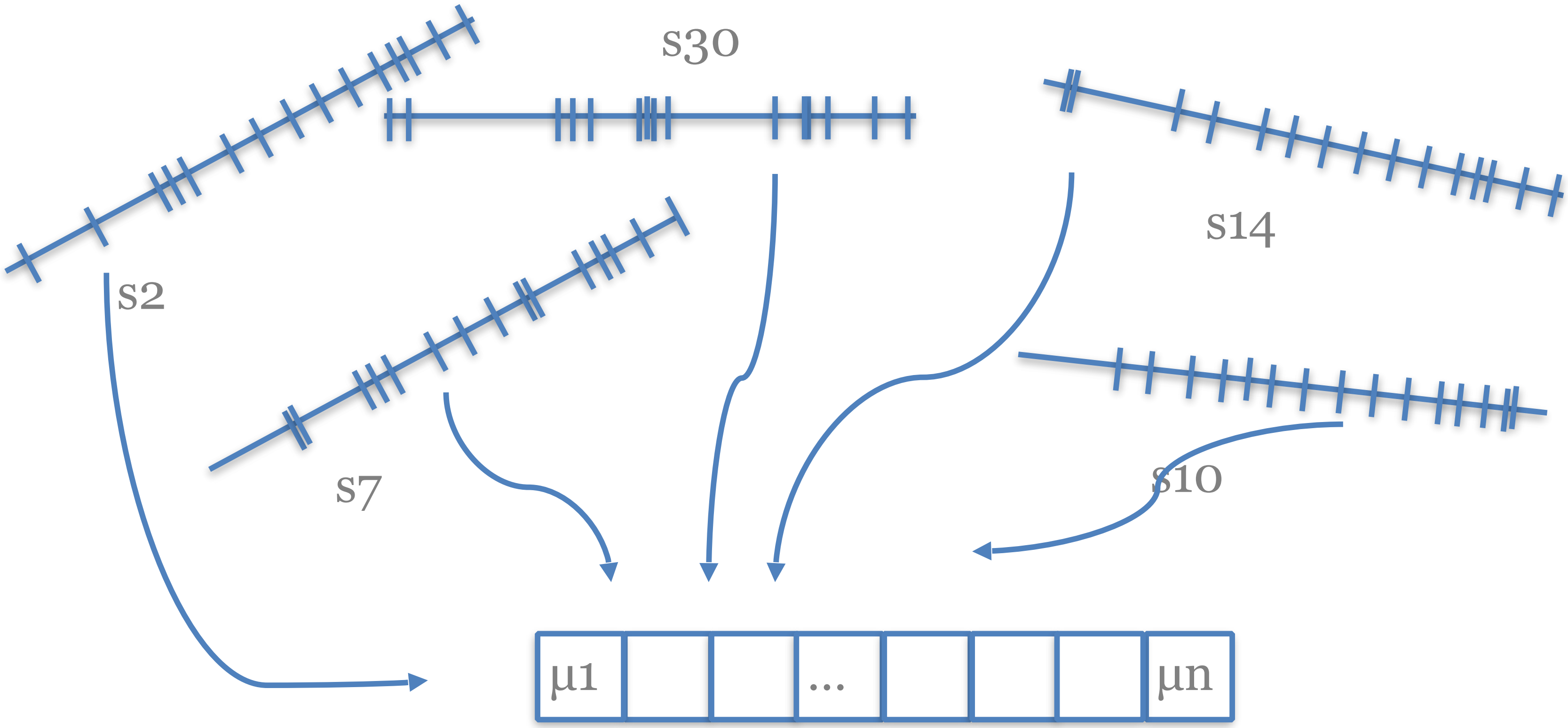
{General, Log In, Performance, Systems}

sv1=[0,1,0,1]

sv2=[2,1,1,0]

Seq.	Ev.	Date	User	State	Type
s_1	e_1	2009-03-02 07:05:45	26209	Information	Log In
	e_2	2009-03-02 07:05:46	26209	Timer	Systems
s_2	f_1	2009-03-02 07:06:45	1863	Information	Log In
	f_2	2009-03-02 07:06:45	1863	Information	General
	f_3	2009-03-02 07:10:20	19039	Error	General
	f_4	2009-03-02 07:14:58	19039	Error	Perform.

Multiple sequences and users



Same length, same types

Features

- **v= [sv, μ(sv), ν(sv)]** – feature
  - μ(sv) = **# sequences** mapping onto sv
  - ν(sv) = **average # of users** in sequences mapping onto sv
- **ρ(sv)** = number of **errors** in sequences mapping onto sv
- v is an **faulty feature** if at least one event in one sequence is in error state, **ρ(sv)>0**.

Example - features

$v_1 = [0,1,0,1;1,1]$, $sv_1 = [0,1,0,1]$

$\mu(sv_1) = 1$, $\nu(sv_1) = 1$, $\rho(sv_1) = 0$

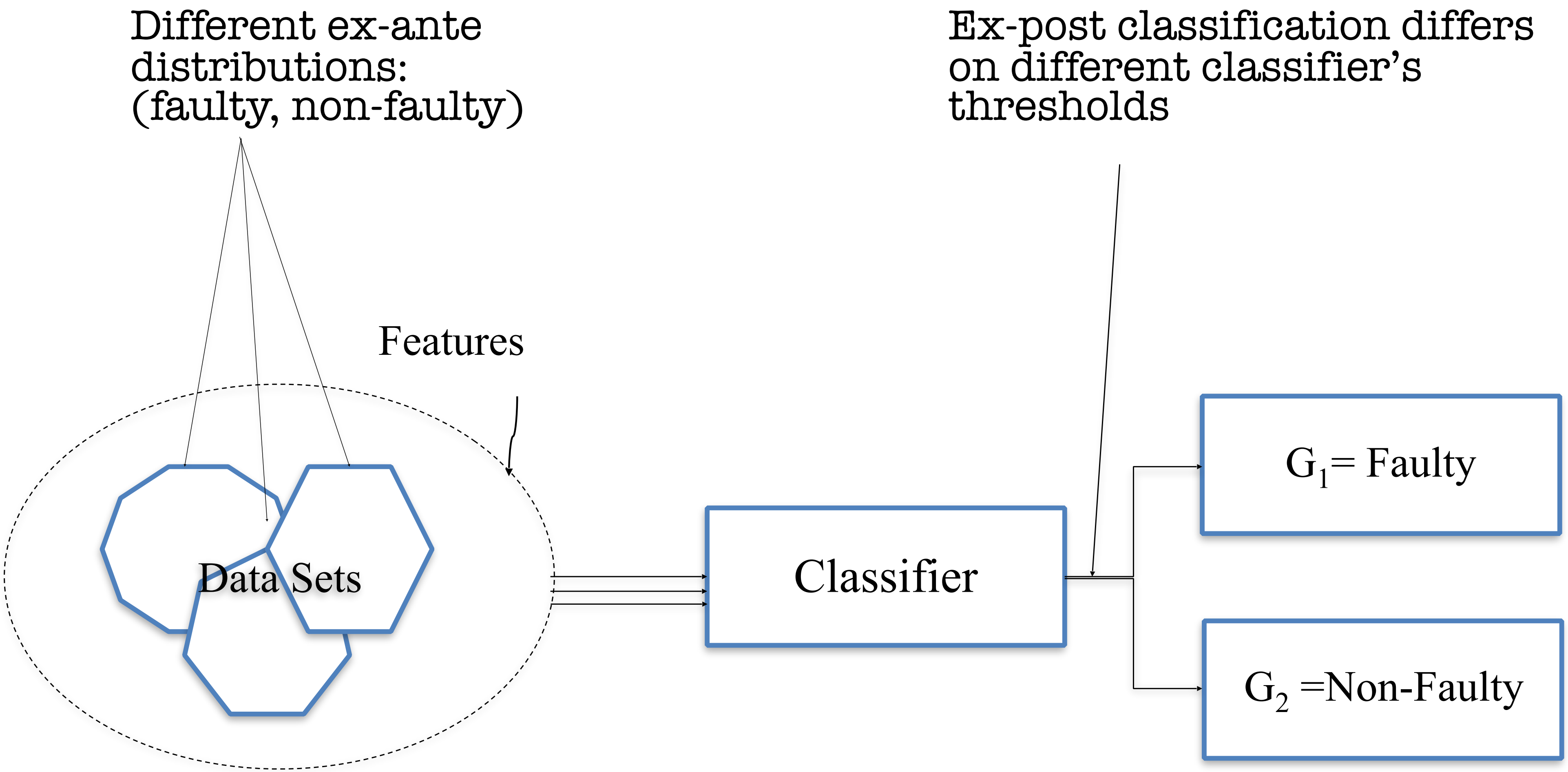
$v_2 = [2,1,1,0;1,2]$, $sv_2 = [2,1,1,0]$

$\mu(sv_2) = 1$, $\nu(sv_2) = 2$, $\rho(sv_2) = 2$

Seq.	Ev.	Date	User	State	Type
s_1	e_1	2009-03-02 07:05:45	26209	Information	Log In
	e_2	2009-03-02 07:05:46	26209	Timer	Systems
s_2	f_1	2009-03-02 07:06:45	1863	Information	Log In
	f_2	2009-03-02 07:06:45	1863	Information	General
	f_3	2009-03-02 07:10:20	19039	Error	General
	f_4	2009-03-02 07:14:58	19039	Error	Perform.

Which models do we use to predict system behaviour?

The classification problem

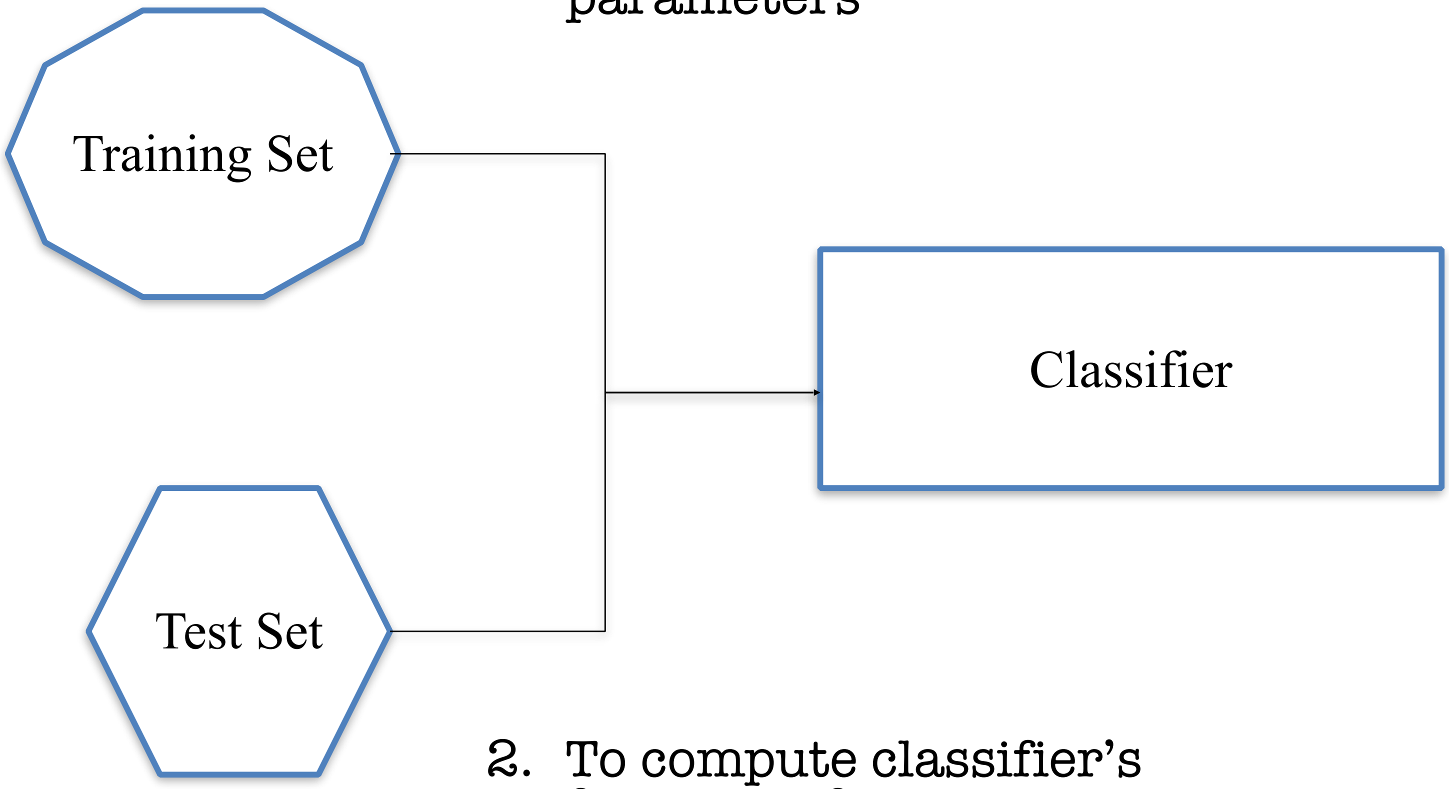


Parametric classification

- The problem varies depending on how many errors we allow in the system
- c – cut-off value, i.e., number of errors in a feature
- Categories:
  - $G_1(c)=\{v = [sv, \mu(sv),\nu(sv)] \mid \rho(sv)\geq c\}$ - faulty
  - $G_2(c)=\{v = [sv, \mu(sv),\nu(sv)] \mid \rho(sv)<c\}$ - non-faulty

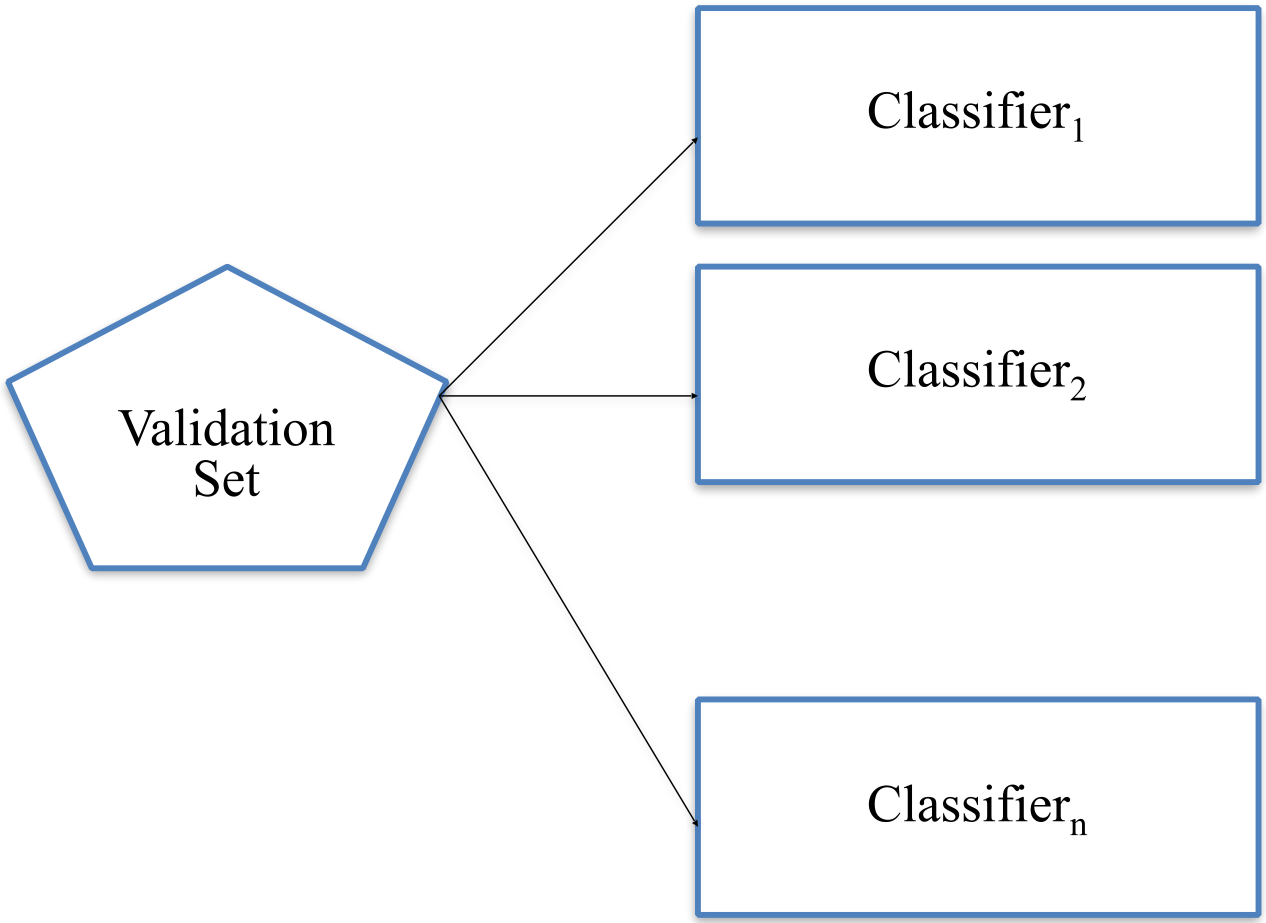
Build classifiers on historical data

1. To tune classifier's parameters



2. To compute classifier's fitting performance

Compare prediction performance



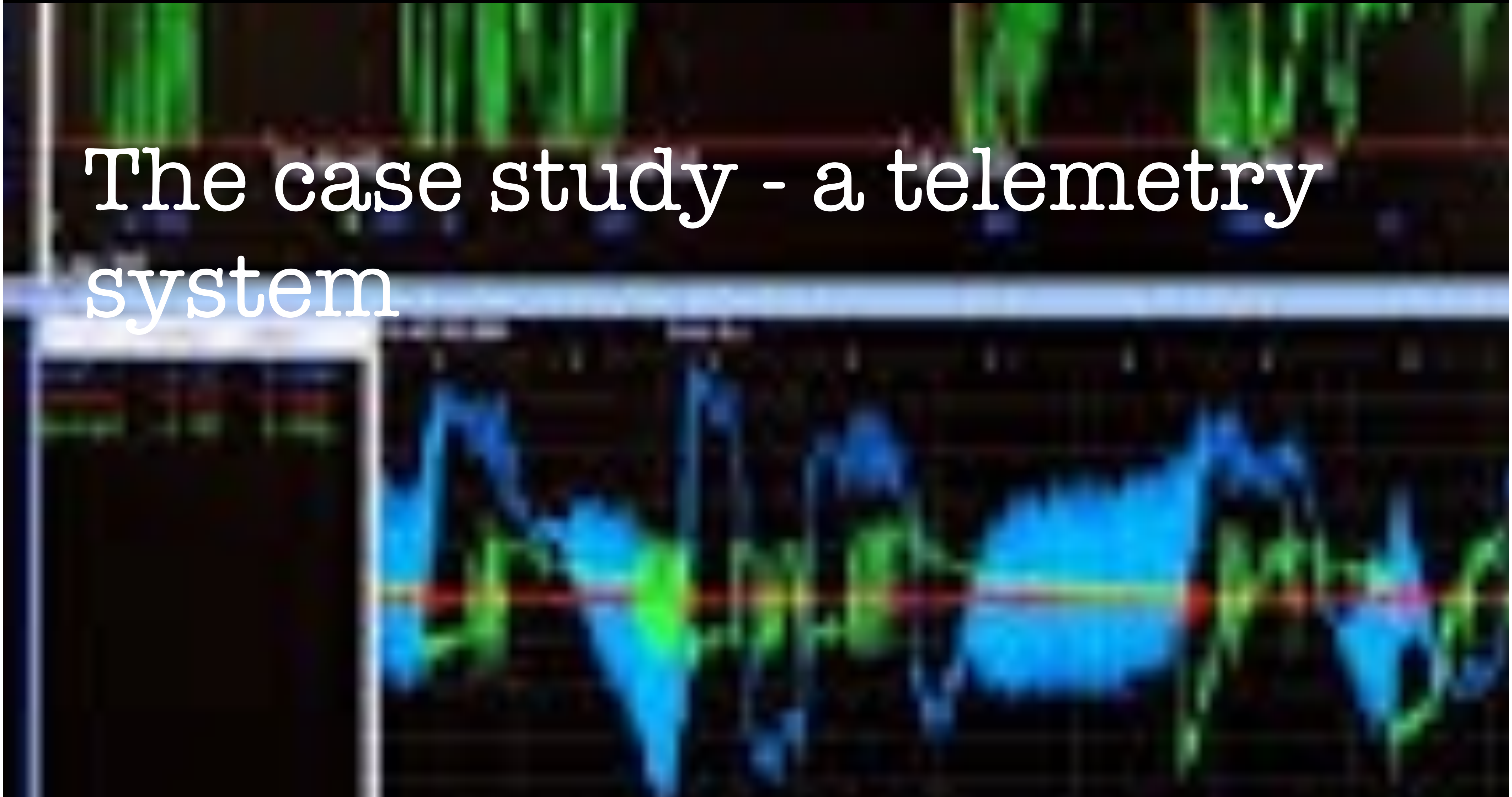
Information Gain

- Did we put too much information in our features?
  - Information Gain selects **feature attributes** that most contribute to the information of a given classification category

$$\mathrm{IG(X)} = H(C) - H(C|X)$$

$$H(C) = -\sum_{o \in C} -p(o) log_2 p(o)$$

The case study - a telemetry system

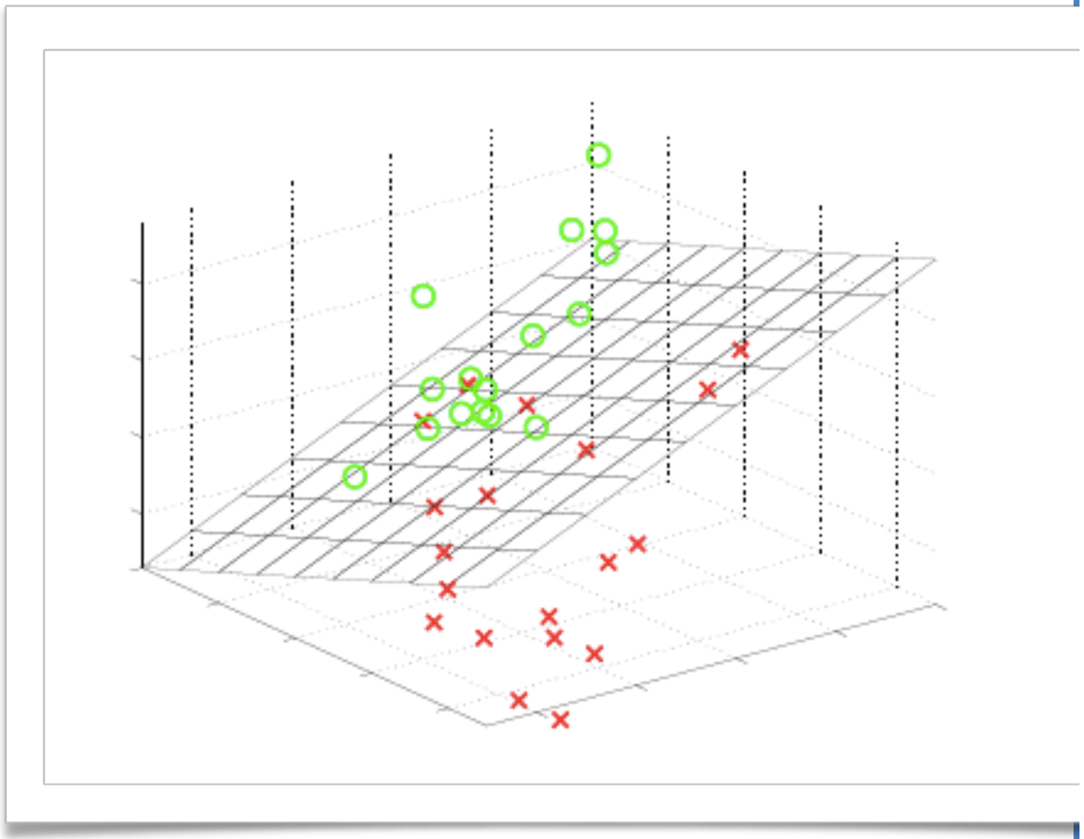


System applications

ID	Application Type
ST_1	Telemetry Module
ST_2	Telemetry Module
ST_3	Telemetry Module
ST_5	Sw. Resources Mgmt.
ST_6	Product Sw. Tools Mgmt.
ST_7	Procurement Sys. Module
ST_{10}	Telemetry Module
ST_{11}	Product Data Mgmt.
ST_{12}	Chain Supply Mgmt. Sys.
ST_{13}	Procurement Sys. Module
ST_{14}	Procurement Sys. Module
ST_{15}	Data Transfer Module
ST_{17}	Product Sensors Mgmt.
ST_{18}	Telemetry Module
ST_{19}	Secondary DB
ST_{21}	Virtual Disk Service Module
ST_{23}	Manufacturing Execution Sys.
ST_{25}	Virtual Disk Service

Support Vector Machines

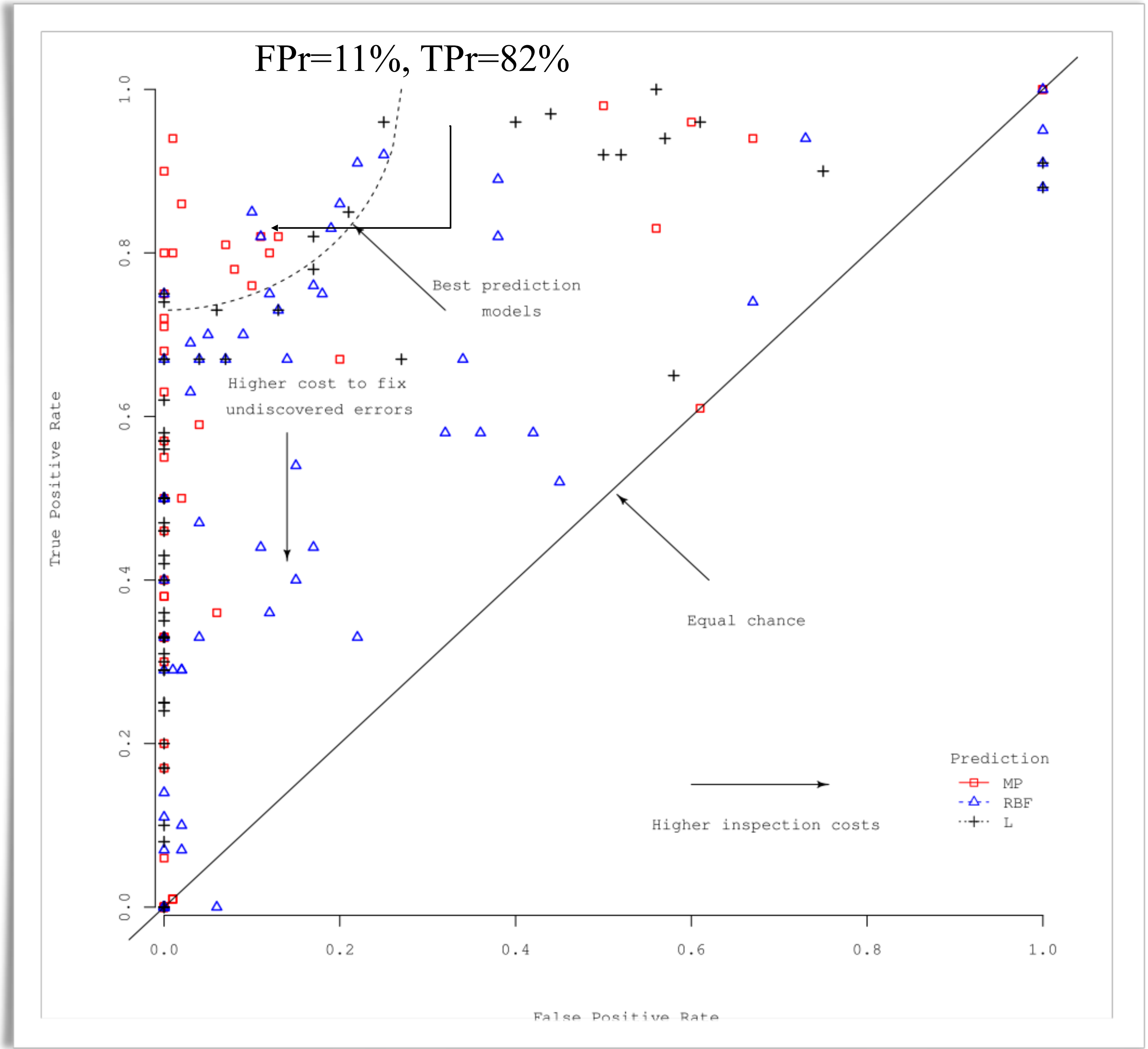
- Different kernels
  - Multilayer perceptron
  - Linear
  - Radial Basis Function



Findings

- Best performance at individual application (MP, c=3):
  - **1% false positive rate**, **94% true positive rate**, and **95% precision**
- Best performance across applications averaged over models for c=2,
  - **9% false positive rate**, **78% true positive rate**, and **95% precision**,

What can predictions tell administrators?


FPr=11%, TPr=82%
Best prediction models
Higher cost to fix undiscovered errors
Equal chance
Higher inspection costs
Prediction
MP
RBF
L
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Example

ID	Application Type
ST_1	Telemetry Module
ST_2	Telemetry Module
ST_3	Telemetry Module
ST_5	Sw. Resources Mgmt.
ST_6	Product Sw. Tools Mgmt
ST_7	Procurement Sys. Module
ST_{10}	Telemetry Module
ST_{11}	Product Data Mgmt.
ST_{12}	Chain Supply Mgmt. Sys.
ST_{13}	Procurement Sys. Module
ST_{14}	Procurement Sys. Module
ST_{15}	Data Transfer Module
ST_{17}	Product Sensors Mgmt.
ST_{18}	Telemetry Module
ST_{19}	Secondary DB
ST_{21}	Virtual Disk Service Module
ST_{23}	Manufacturing Execution Sys.
ST_{25}	Virtual Disk Service

Example - ST6

- Application that manages software tools of cars
  - Pervasive in the telemetry system
- **106** distinct sequences of **10** different event types, **18%** multiple sequences, and **89%** with more than one user

ST6 - Analysis

- c=**1**
  - $G_1(\mathbf{1})=\{v = [sv, \mu(sv),\nu(sv)] \mid \rho(sv)\geq \mathbf{1}\}$
  - $G_2(\mathbf{1})=\{v = [sv, \mu(sv),\nu(sv)] \mid \rho(sv)< \mathbf{1}\}$
- IG reduction from 12 to 7 still including μ and ν

ST_6	Pos% test	Pos% validation	Model
t=1/2	0.4	0.45	MP
t=1/3	0.39	0.49	MP
t=1/4	0.49	0.23	MP
t=1/5	0.45	0.33	MP
avg.	**0.43**	**0.38**	**MP**

Confusion matrix - MP pred.

	Pred. Pos	Pred. Neg	Total
Pos	14 82%	3 18%	17 100%
Neg	2 11%	16 89%	18 100%
Total Percent	16 45%	19 54%	35 100%

TPr

FPr

Prediction - assumptions

- Behaviour is the same in next three months
- 1000 sequences
- Category balance in future sets is the one of the test set (39%)
  - 390 faulty sequences and 610 non- faulty sequences

In numbers

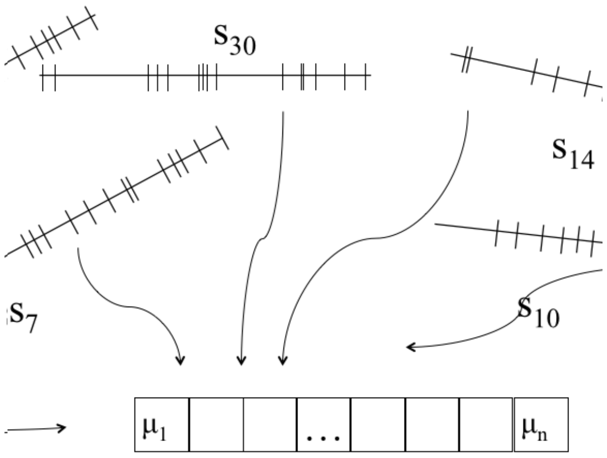
- 450 (45%* 1000) predicted faulty sequences
- Predicted faulty sequences that have no errors:
  - 67 = 11%*610
- Predicted non-faulty sequences that have an error
  - 70 =18%*390

	Pred pos	Pred neg	Total
Pos	*82%*	*18%*	*100%*
Neg	*11%*	*89%*	*100%*
Total	*45%*	*54%*	*100%*

Cost of prediction

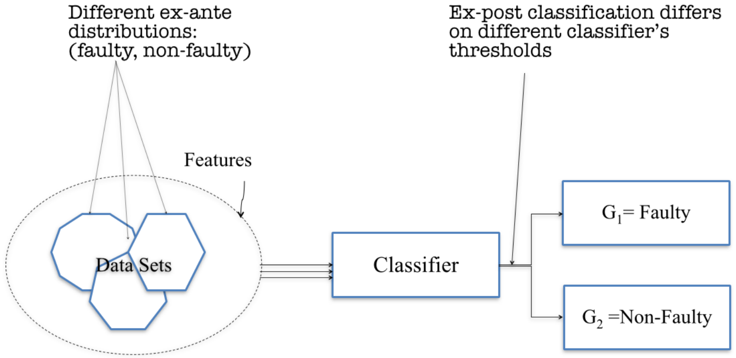
- *Inspection cost.*
  - Wasting time ≥ 67 * average cost to fix one error
- *Cost for undiscovered errors.*
  - Defect slippage ≥ 70

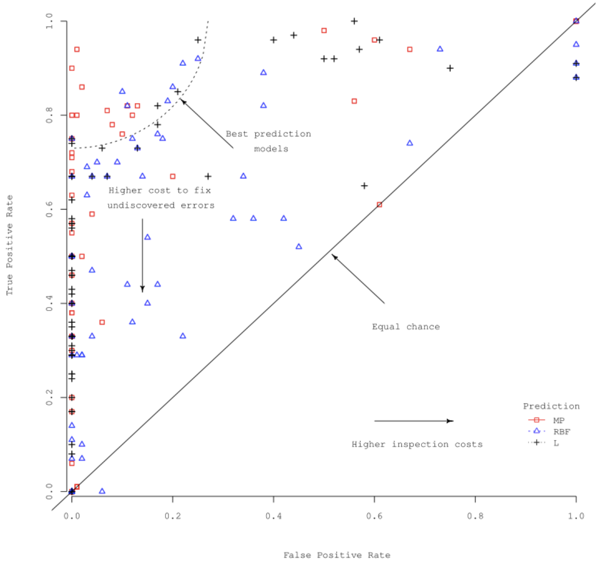
Recapitulation



$$IG(X) = H(C) - H(C|X)$$

$$H(C) = -\sum_{o\in C} -p(o)log_2p(o)$$





Sequences to model system changes

Classifiers to model and predict system behaviour

Accuracy to measure costs in prediction

Thank you