CERN openlab Major Review Meeting

29th January 2009

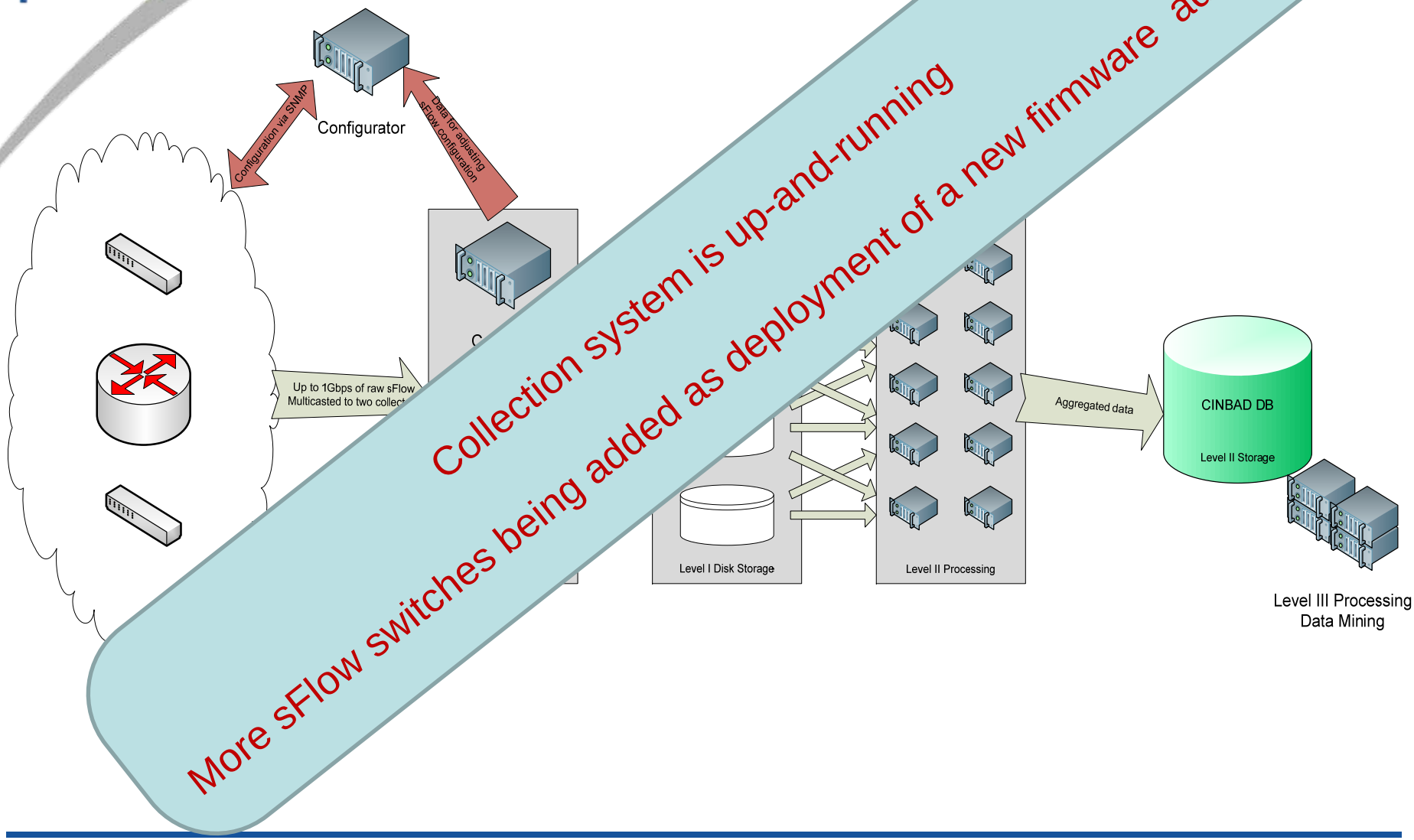
Milosz Marian Hulboj - CERN/Procurve

Ryszard Erazm Jurga - CERN/Procurve

Agenda

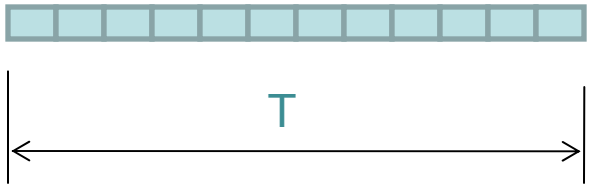
- Flow analysis
  - Statistical analysis methods
  - What about the signatures?
  - Results
- Time series mining – preliminary stage
  - Introduction
  - Future plans
- Conclusions

Collector Impl ion



Statistical Flow Based Analysis (1)

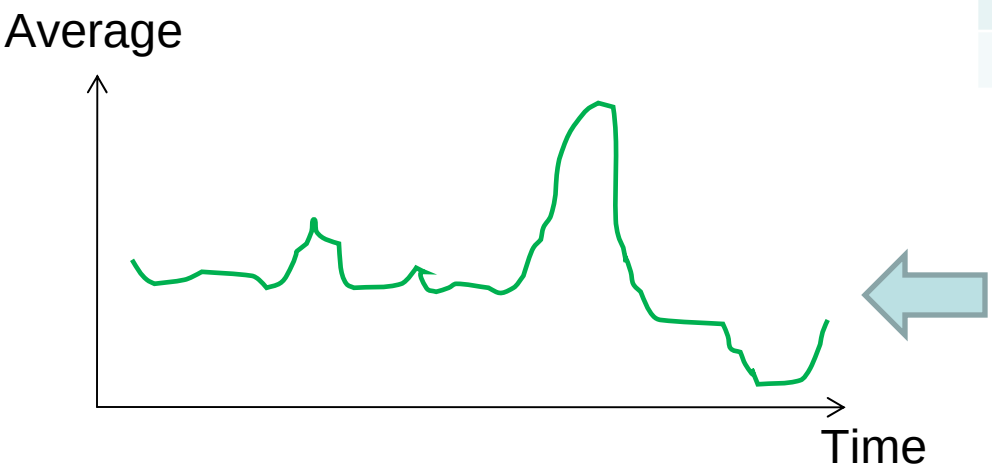
sampled packets



Sampled Flow Table

Time	Proto	Src IP	Dst IP	Counter
T1	UDP	IP1	IP2	2
T1	UDP	IP1	IP3	3
T1	UDP	IP1	IP4	5
T1	UDP	IP2	IP8	8
T1	UDP	IP2	IP7	1
T2	TCP	IP1	IP2	2
T2	UDP	IP3	IP3	3
T2	UDP	IP3	IP4	5
T2	UDP	IP4	IP8	8
T2	UDP	IP4	IP7	1

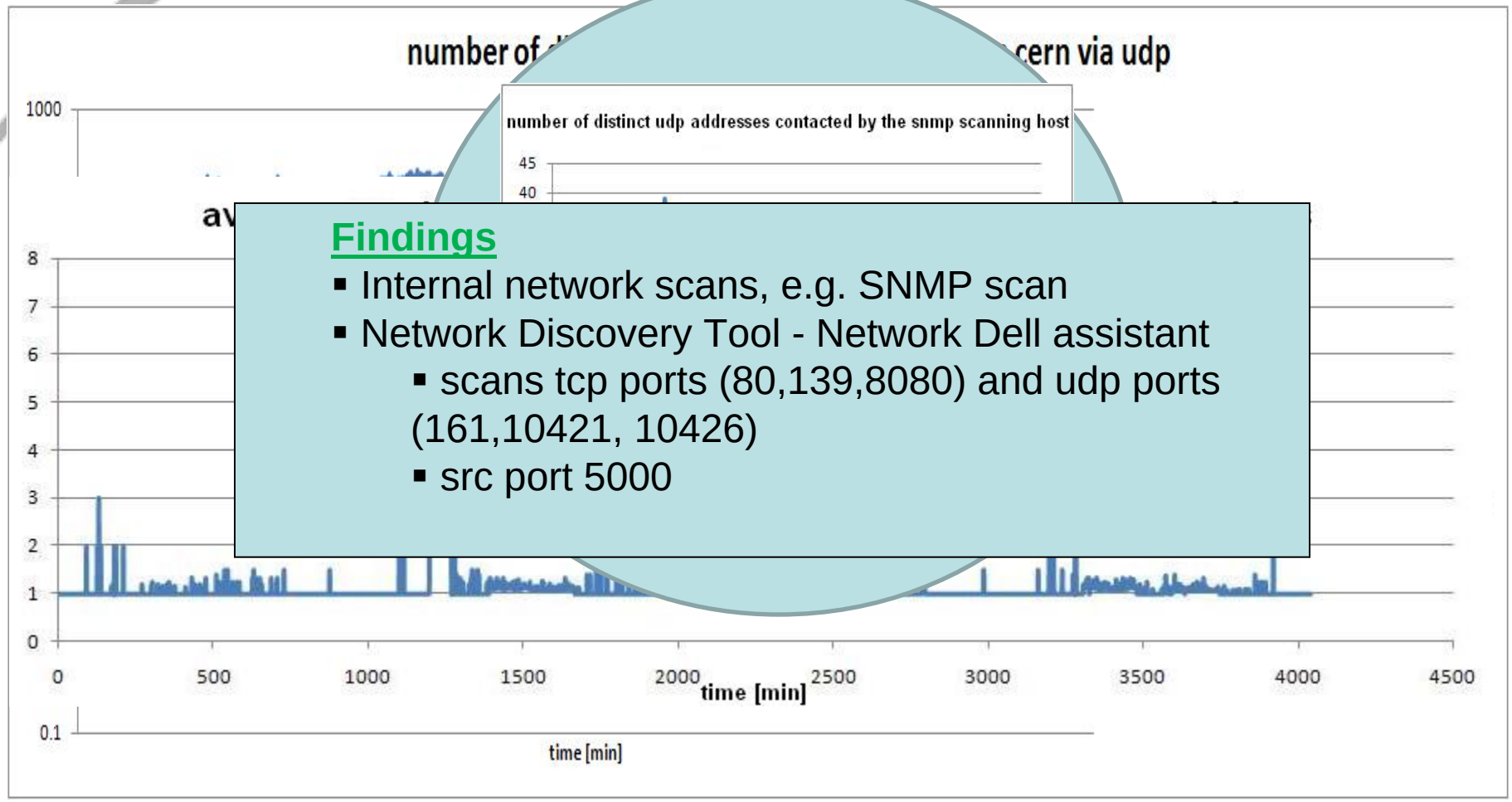
Time	Proto	Src IP	#Dst IP	
T1	UDP	IP1	3	
T1	UDP	IP2	2	Average
T2	UDP	IP3	2	
T2	UDP	IP4	2	Average



Statistical flow based analysis (2)

- Moving Average approach
  - Monitor a certain parameter and report hosts violating a given threshold, e.g:
    - Monitoring UDP connections within CERN from portable hosts
    - Measuring the number of different destination addresses contacted by a given source address
- Isolating anomalous traffic is challenging:
  - Very noisy traffic (many different protocols):
    - exclude flows from ‘well-known’ services (e.g. dhcp, ldap, AFS, etc)
    - the excluded ‘well-known’ flows should be analysed separately
  - Payload inspection might be necessary

Example of noise reduction



Findings

- Internal network scans, e.g. SNMP scan
- Network Discovery Tool - Network Dell assistant
  - scans tcp ports (80,139,8080) and udp ports (161,10421, 10426)
  - src port 5000

Top flow analysis

- T
- D

#	SIG_NAME	PAYLOAD
606	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
607	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
608	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
609	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
610	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
611	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
612	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
613	CINBAD BitTorrent 2	64313A7264323A696432303A939F70DF1D5C16A4EB9A909EA844276F42
614	CINBAD BitTorrent 2	64313A7264323A696432303A939F70DF1D5C16A4EB9A909EA844276F42
615	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
616	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
617	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
618	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
619	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
620	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
621	CINBAD BitTorrent 2	64313A7264323A696432303AA7F3318F65FF036B3549023F311B6E2934E
622	CINBAD BitTorrent 2	64313A7264323A696432303AA7F3318F65FF036B3549023F311B6E2934E
623	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
624	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
625	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
626	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
627	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
628	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
629	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F
630	CINBAD BitTorrent 1	64313A6164323A696432303A900B3E8310D2A9C1C582173EAA1ADF6D3F

order

s in the

world

ports

same

How to use signatures

- simple_filter
  - CINBAD tool based on libpcap for filtering collected data
  - Signatures can be written as pcap filter strings
- Snort
  - Rule based network traffic monitoring system
  - Rules for detecting numerous anomalies available
  - Does not work with sampled traffic out-of-the box

Snort customisation

- Porting to work with sampled data
  - workaround for truncated payloads
  - snort rules translated into stateless ones (if applicable)
- Results of the analysis logged to Oracle
- 7000+ rules with daily updates
  - cinbad rules
    - e.g. bittorent, zatto, QQ …
  - http://www.emergingthreats.net
- Campus and Internet traffic analysis

Some signature analysis results

- ~45% alerts compared to the Snort analysis on the central firewall
  - we expect this ratio to increase when we add more switches
  - traffic rate at the firewall is high and Snort cannot process every packet
- internal and external traffic inspected
  - more p2p applications, instant messengers
  - two trojan likely infections
    - Password Stealer
    - Win32/Alureon.gen!J

Results are promising, but…

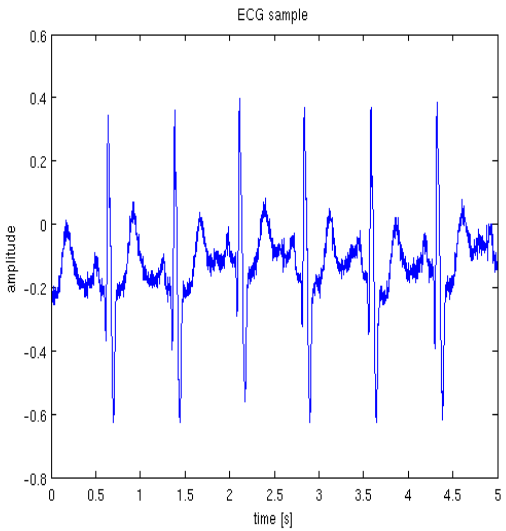
Presented methods require a significant amount of manual work

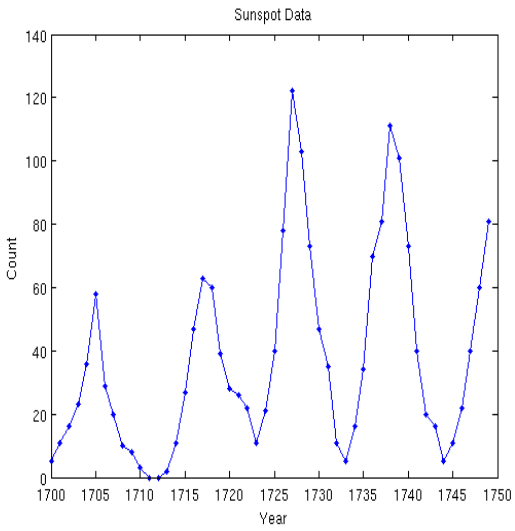
That is why we want to look at the Time Series Data Mining techniques…

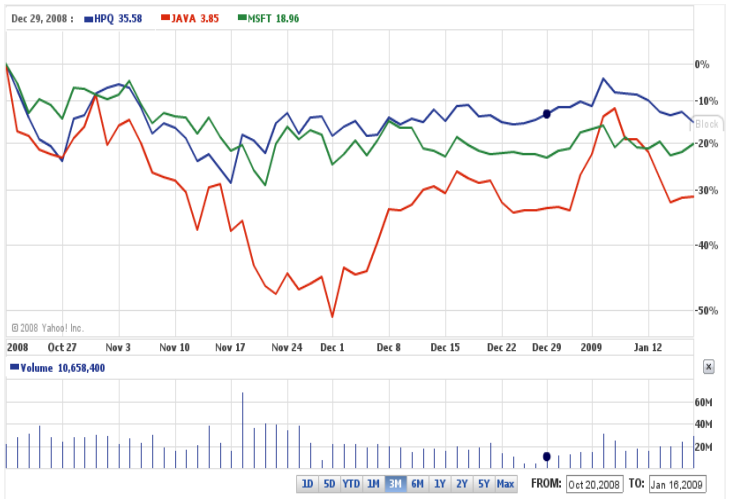
Which hopefully will allow to increase the automatisation

Time series and why do we care?

- What are the Time Series?
  - A **time series** is a sequence of data points, measured at successive times
  - Time series are ubiquitous, more and more data is being measured and collected
- Examples:







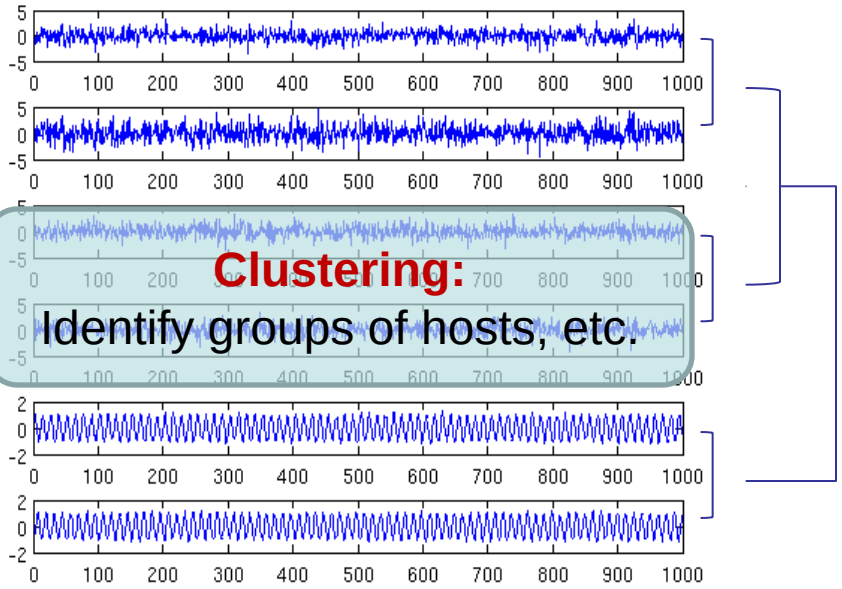
Time series analysis

- Time series carry much information about the generating processes, for example:
  - What kind of system generated the data?
  - Temporal behaviour of the phenomenon
  - What is likely to happen in the future?
  - Relation with other data sources
  - Occurrence of the anomaly (novelty/surprise)
  - …
- Unfortunately the information is usually buried deeply within the multivariate data

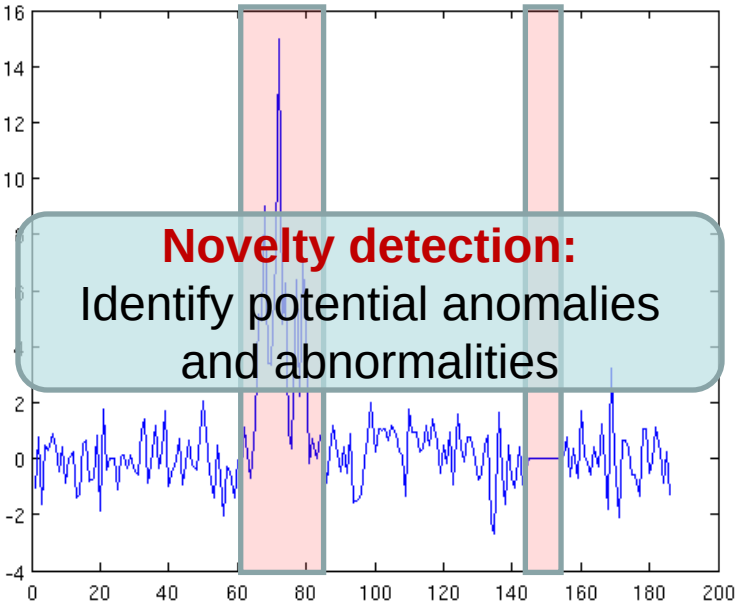
Time series in networking

- Time series are the most natural representation for the state of the network:
  - Amount of traffic
  - SNMP counters
  - Number of connections
  - Number of distinct hosts contacted
  - Number of different TTL values
  - Number of BitTorrent signatures
  - Average packet roundtrip
  - …

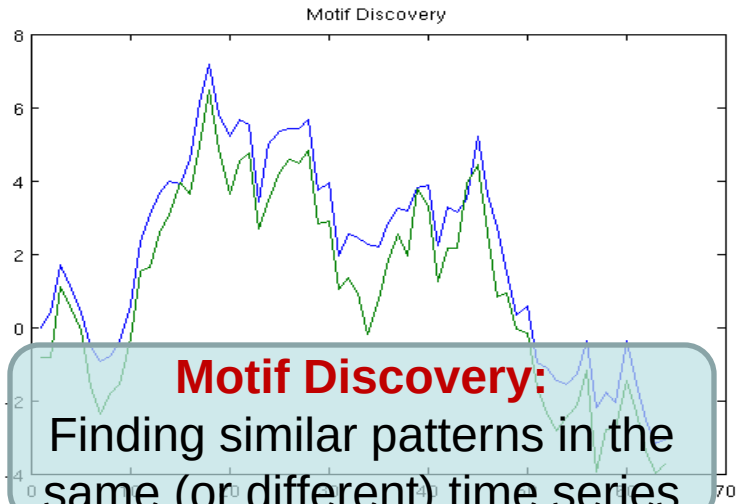
What can we find in time series data?



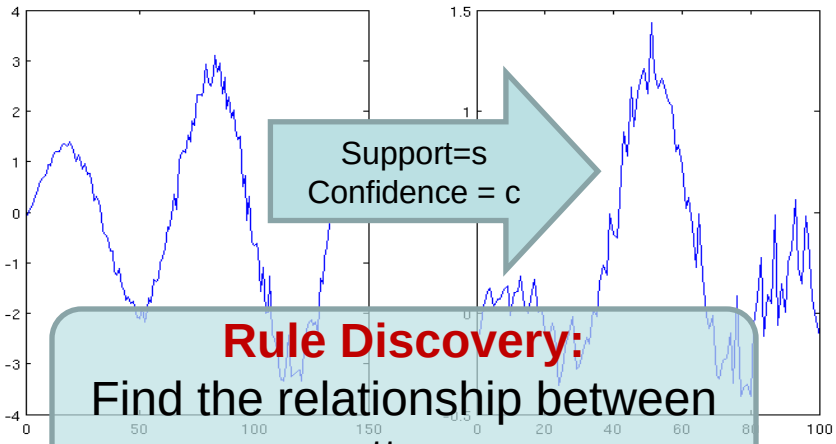
Clustering:
Identify groups of hosts, etc.



Novelty detection:
Identify potential anomalies and abnormalities



Motif Discovery:
Finding similar patterns in the same (or different) time series



Rule Discovery:
Find the relationship between patterns

Time series mining – plans

- Much ongoing research in the area of time series analysis, for example:
  - Financial data analysis (we all want to be rich…)
  - Bioinformatics, genomics (i.e. DNA analysis)
  - Medicine (i.e. attempts to build brain-computer interface)
  - …
  - Network traffic analysis (i.e. detecting traffic volume anomalies)
- Look at the current state of the time series data mining
- Develop methods useful for the CINBAD project

Time series mining – plans

- Try to combine time series data mining with automatic signature extraction:
  1. An anomaly (novelty) is being detected by the mining procedure
  2. Feed the relevant traffic to the extraction engine
  3. Attempt to identify common patterns (longest common substrings) in the sampled packets
  4. Manual assessment of usefulness of the rules seems unavoidable

Conclusions

- Starting with the statistical flow analysis we have discovered typical patterns and signatures
- We obtained encouraging initial results from SNORT signature analysis
- Very good feedback received from HP ProCurve and its Architects' Forum
- We need to investigate the time series mining techniques