Fachhochschule Gelsenkirchen

Internet Analysis System → IAS

Prof. Dr.
Norbert Pohlmann

Institute for Internet Security - if(is)
University of Applied Sciences Gelsenkirchen
http://www.internet-sicherheit.de

if(is) internet security.

Content

- Aim and outcomes of this lecture
- Idea of the Internet Analysis System
- Knowledge Base
- Outline of the Current State
- Detection of Attacks and Deflection
- Forecast of Patterns and Attacks
- Summary

Content

- **Aim and outcomes of this lecture**
- Idea of the Internet Analysis System
- Knowledge Base
- Outline of the Current State
- Detection of Attacks and Deflection
- Forecast of Patterns and Attacks
- Summary

Internet Analysis System (IAS)
→ Aims and outcomes of this lecture

Aims

- To introduce an Internet Early Warning System with a statistical approach
- To explore the structure of the Internet Analysis System
- To analyze the results of the Internet Analysis System
- To assess the value the Internet Analysis System

At the end of this lecture you will be able to:

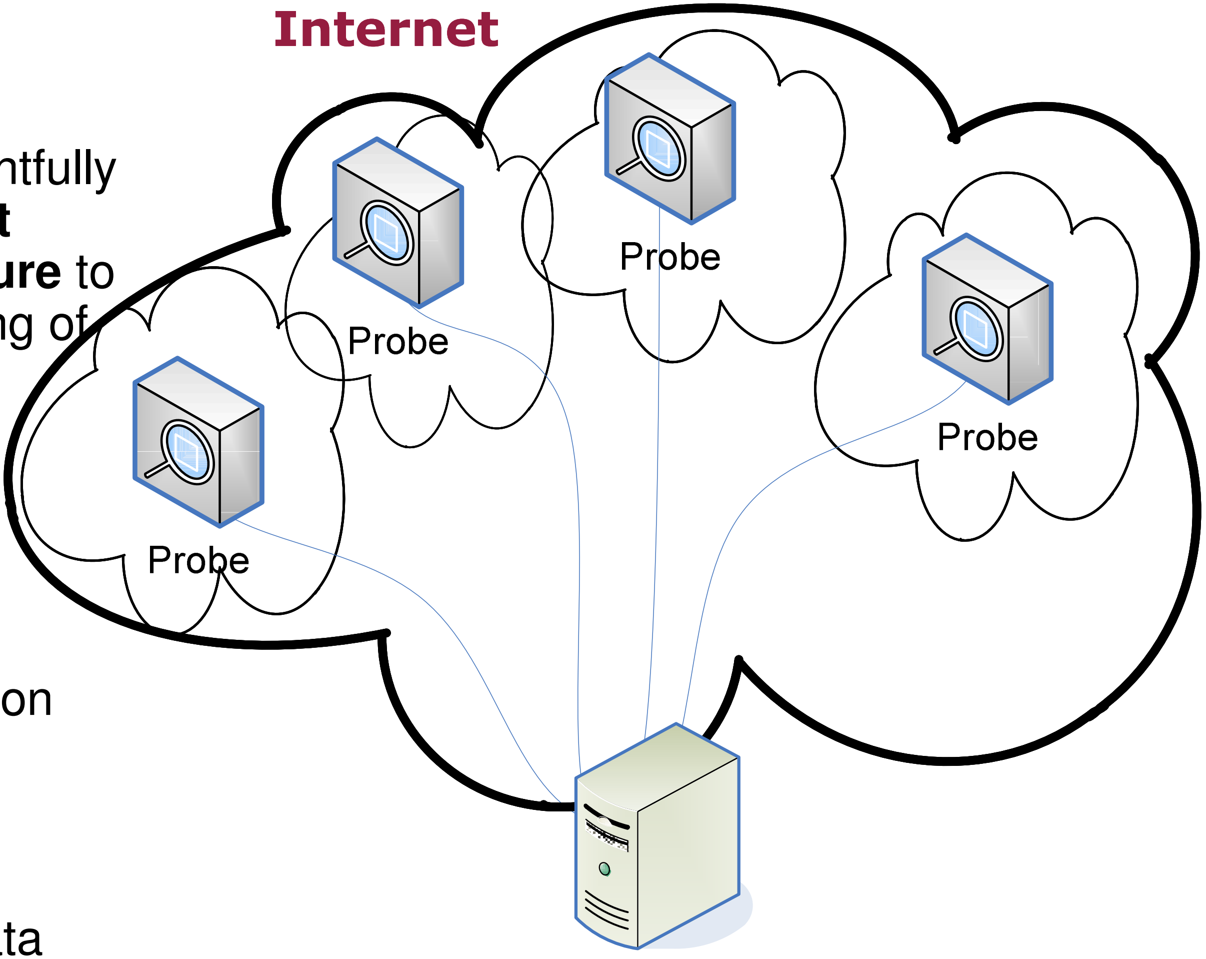
- Understand what is meant by the Internet Analysis System.
- Know something of the structure of the Internet Analysis System.
- Know what the results of the Internet Analysis System could be.
- Understand the capabilities and limitations of the Internet Analysis System.

Content

- Aim and outcomes of this lecture
- **Idea of the Internet Analysis System**
- Knowledge Base
- Outline of the Current State
- Detection of Attacks and Deflection
- Forecast of Patterns and Attacks
- Summary

Internet Analysis System (IAS)
→ Idea

- Observation of the critical infrastructure **„Internet“.**
- **Probes** are placed in thoughtfully selected spots of the **internet communication infrastructure** to gather the raw data, consisting of counted header information.
- Only header information is counted, which is **not considered as data privacy relevant**.
- The system gathers information over a **great period of time**!
- A centrally managed **Evaluation System** is used to analyze the raw data and to display the detailed results in an intuitive manner.

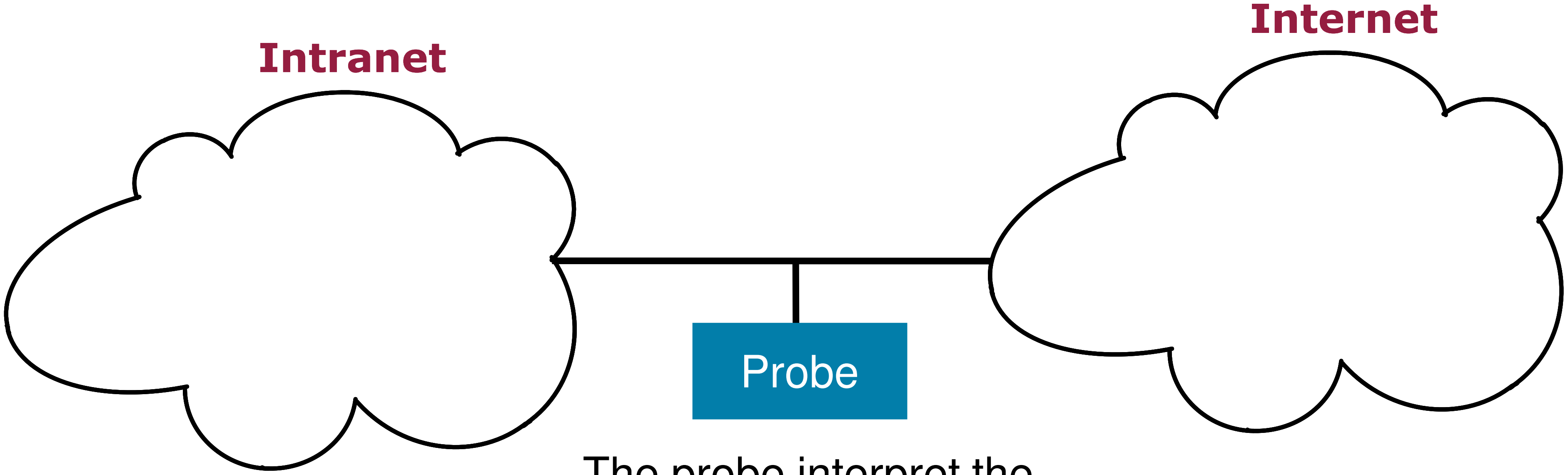


Evaluation System

IAS

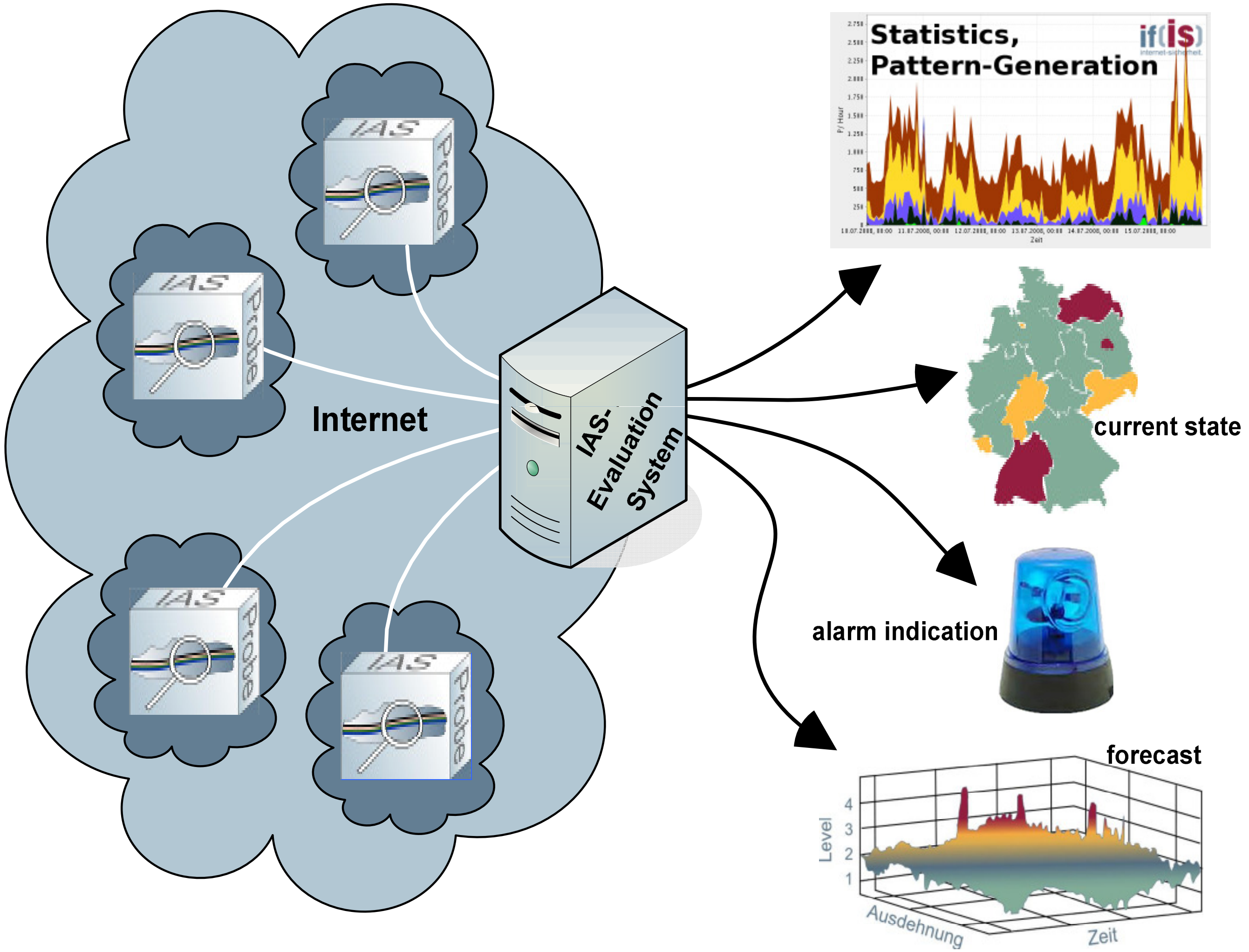
© Prof. Dr. Norbert Pohlmann, Institute for Internet Security - if(is), University of Applied Sciences Gelsenkirchen, Germany

Probe
→ Fundamental approach with the IAS



The probe interpret the
behavior in communication
between an intranet (e. g. corporate network) and the internet

Internet Analysis System (IAS)
→ Targets



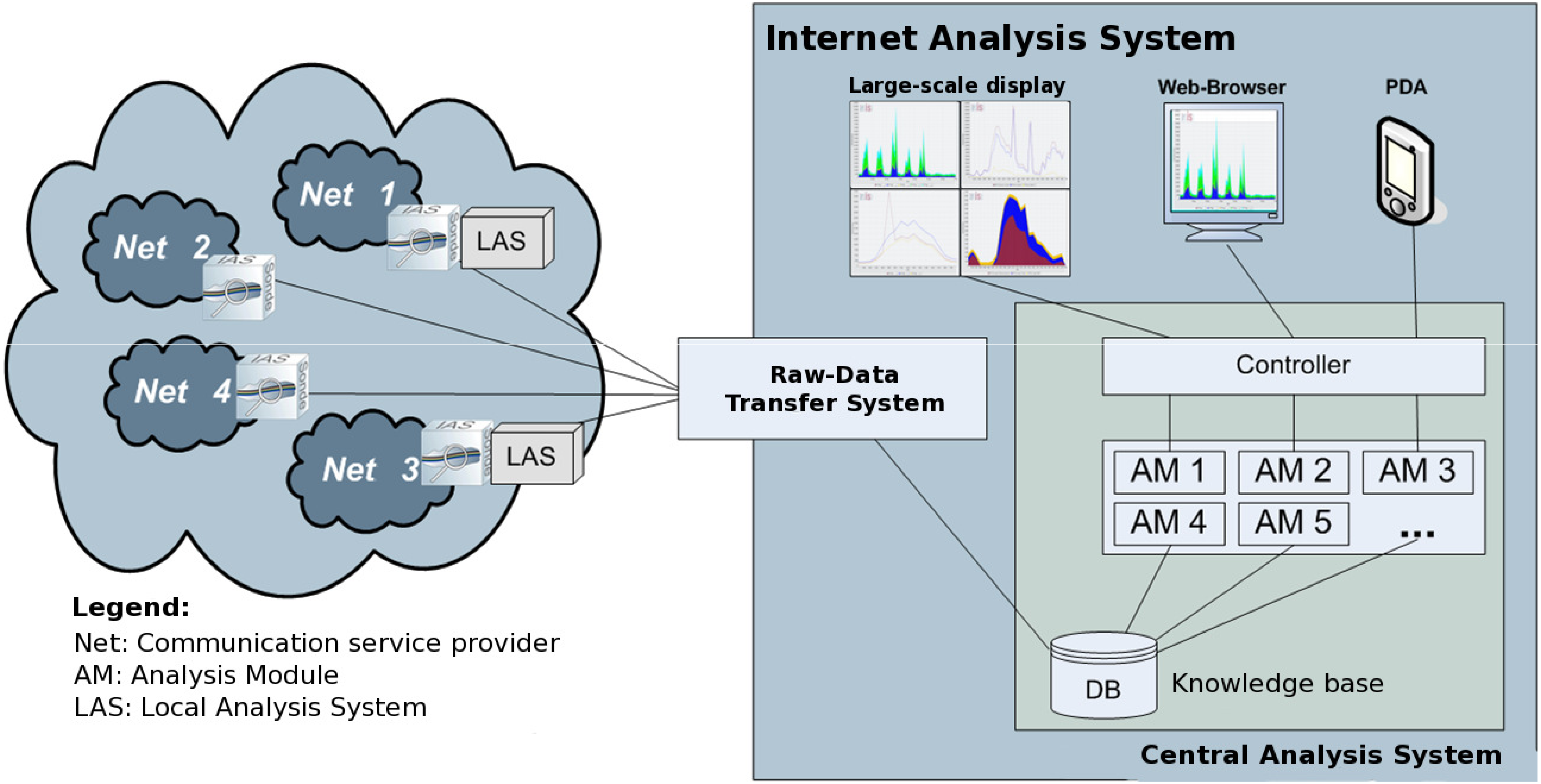
Description of profiles, patterns and coherences, creation of a **knowledge base.**

Outline of the **current state** of the internet.

Detection of attacks and of deflections.

Forecast of patterns and attacks.

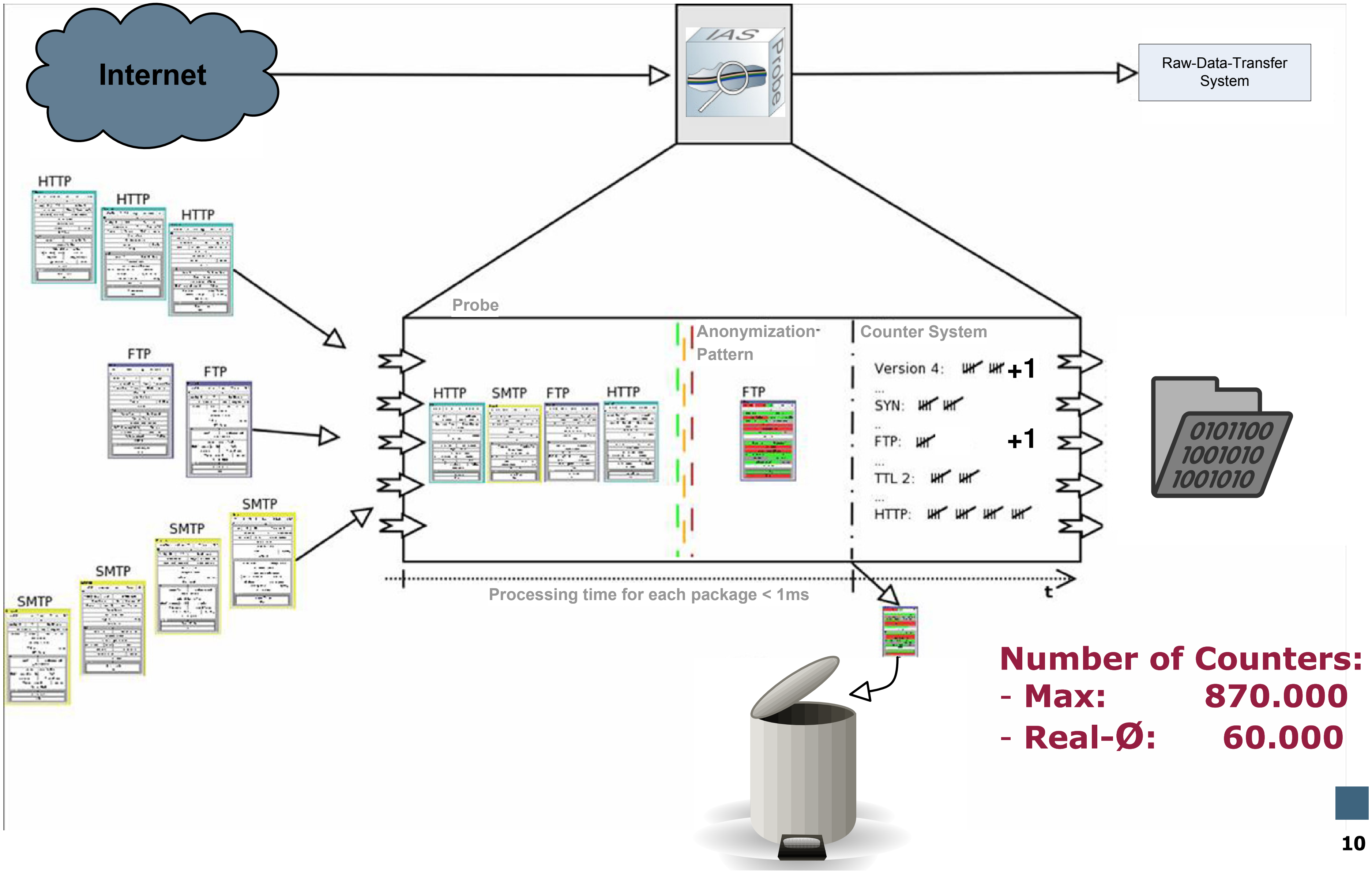
Implementation of the IAS
→ Overview



Legend:

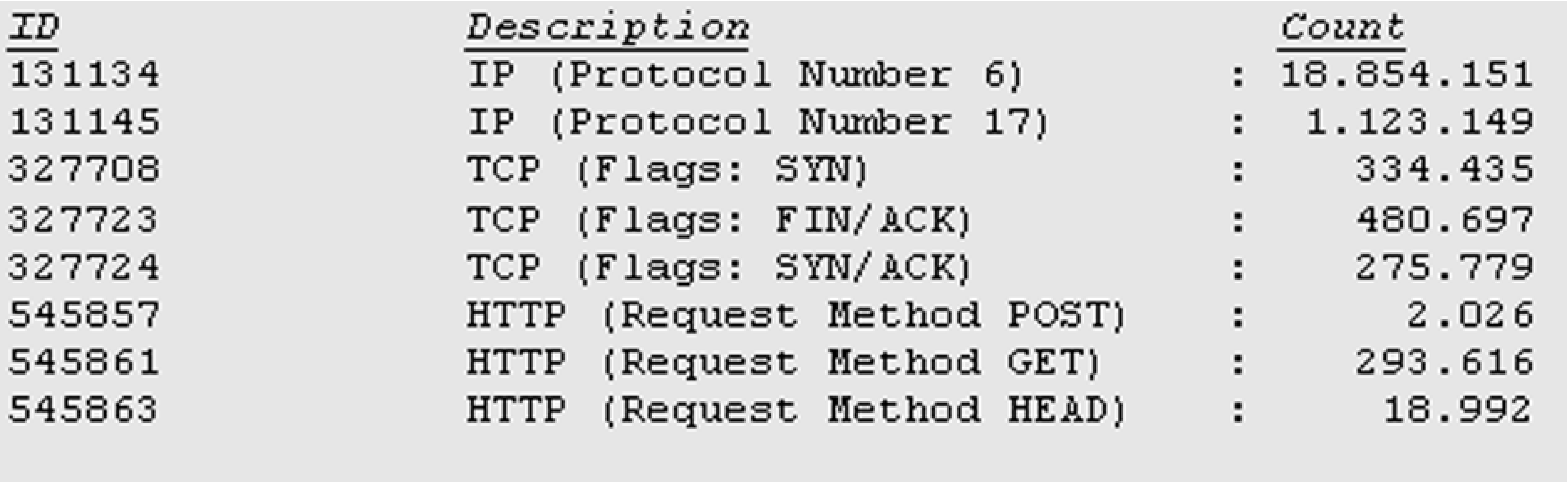
Net: Communication service provider
AM: Analysis Module
LAS: Local Analysis System

Internet Analysis System (IAS)
→ Counting of header information (1/2)



Number of Counters:
- Max: 870.000
- Real-Ø: 60.000

Internet Analysis System (IAS)
→ Counting of header information (2/2)

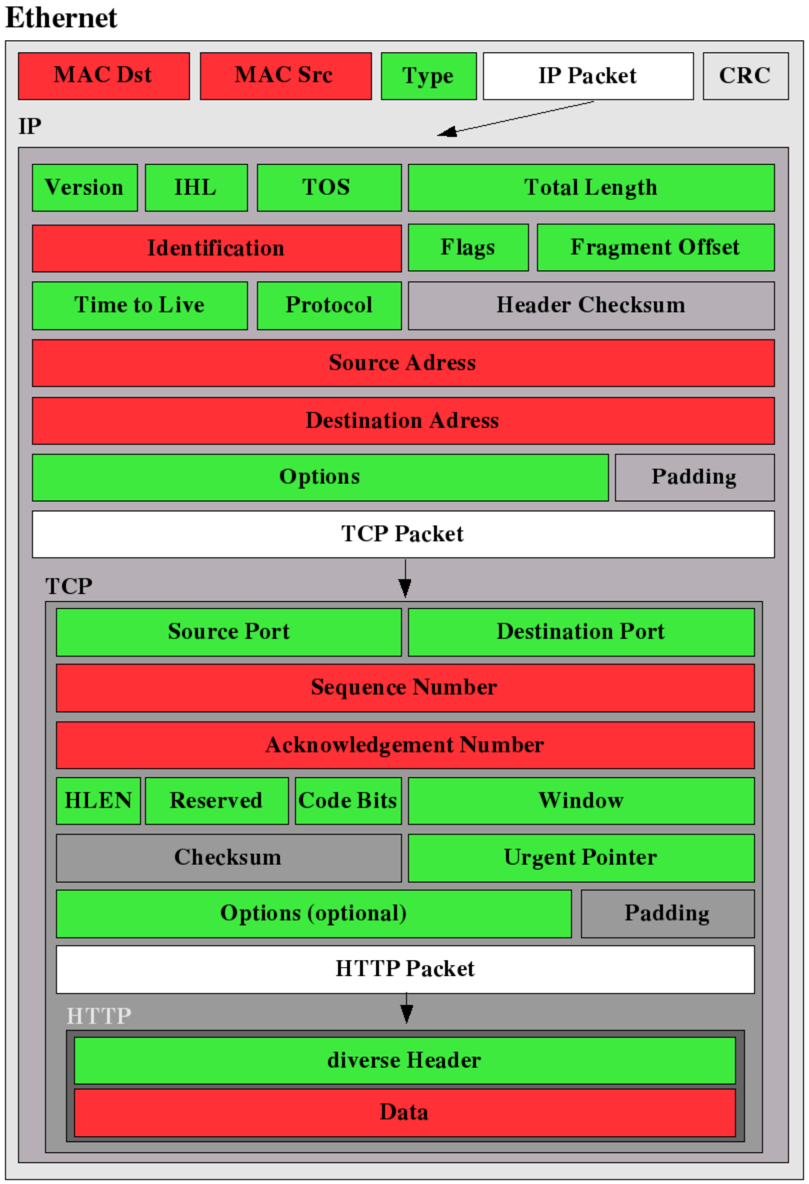


```
ID                  Description                                 Count
131134              IP  (Protocol Number  6)                 : 18.854.151
131145              IP  (Protocol Number  17)                :  1.123.149
327708              TCP  (Flags:  SYN)                       :    334.435
327723              TCP  (Flags:  FIN/ACK)                   :    480.697
327724              TCP  (Flags:  SYN/ACK)                   :    275.779
545857              HTTP  (Request  Method POST)             :      2.026
545861              HTTP  (Request  Method GET)              :    293.616
545863              HTTP  (Request  Method HEAD)             :     18.992
```

- On the right behind the colon character are the **counter values** for each parameter specified on the left.
- Each line stands for one counter.
- For example, line 2 indicates that 1,123,149 packets with the IP protocol number 17 (UDP) appeared in the prescribed time interval.
- All of this information is completely anonymous!

Principle of raw data collection
→ Protocol stack (1/2)

- Ethernet
  - Type: Type of the nested packets, in this case: 0x0800 (IP)
  - Checksum (CRC) irrelevant
- Internet Protocol (IP)
  - e.g.: Total Length of the packet
  - Protocol: Type of the nested Packet, in this case: 6 (TCP)
  - Source- and destination address privacy critical



Ethernet

MAC Dst | MAC Src | Type | IP Packet | CRC

IP

Version | IHL | TOS | Total Length

Identification | Flags | Fragment Offset

Time to Live | Protocol | Header Checksum

Source Adress

Destination Adress

Options | Padding

TCP Packet

TCP

Source Port | Destination Port

Sequence Number

Acknowledgement Number

HLEN | Reserved | Code Bits | Window

Checksum | Urgent Pointer

Options (optional) | Padding

HTTP Packet

HTTP

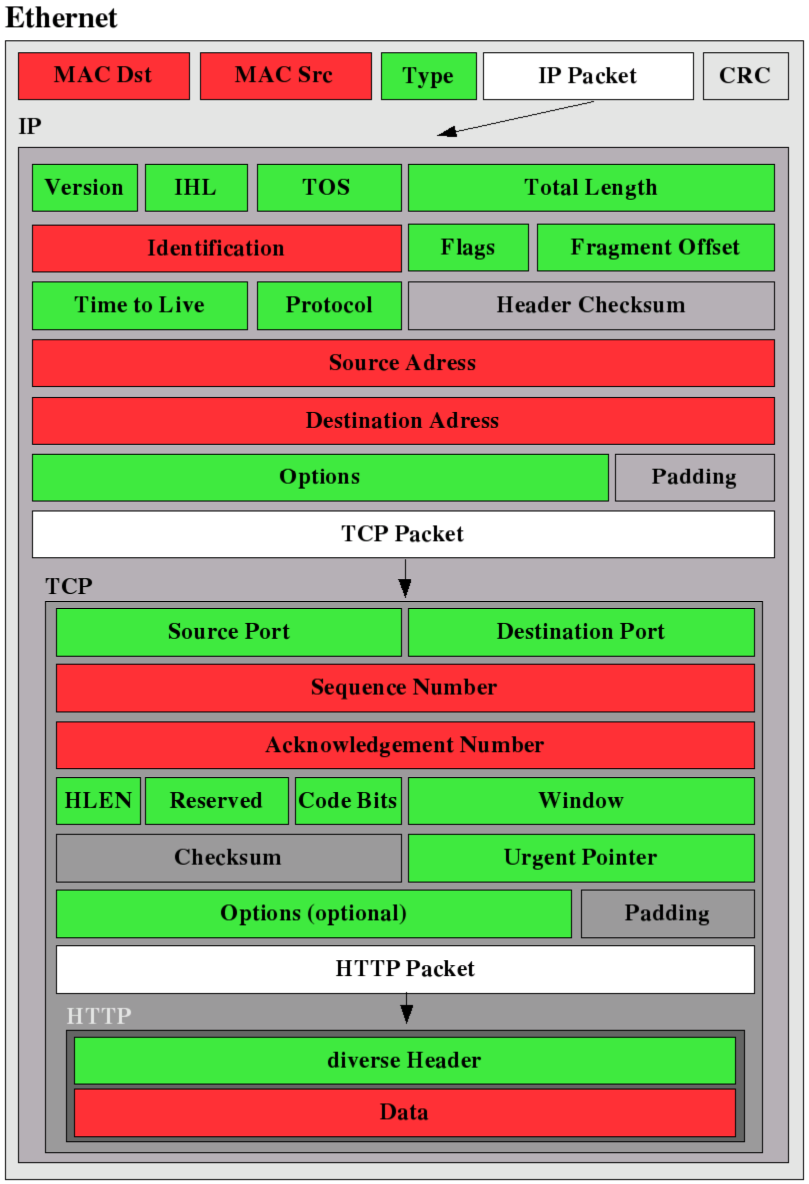
diverse Header

Data

Principle of raw data collection
→ Protocol stack (2/2)

- Transmission Control Protocol (TCP)
  - Port: end point of the connection
    - HTTP: 80 (WWW)
    - Others e.g.:
      SMTP (25), HTTPS (443)
  - Code Bits
    - Information about the connection establishment and shut down

- Hypertext Transfer Protocol (HTTP)
  - Header:
    - e.g.: User Agent:
      describes the user's browser
  - User data (DATA)
    e.g.: content of a web site



Ethernet

MAC Dst | MAC Src | Type | IP Packet | CRC

IP

Version | IHL | TOS | Total Length

Identification | Flags | Fragment Offset

Time to Live | Protocol | Header Checksum

Source Adress

Destination Adress

Options | Padding

TCP Packet

TCP

Source Port | Destination Port

Sequence Number

Acknowledgement Number

HLEN | Reserved | Code Bits | Window

Checksum | Urgent Pointer

Options (optional) | Padding

HTTP Packet

HTTP

diverse Header

Data

Internet Analysis System (IAS)
→ Counting of header information (1/3)

- Description of the network traffic
  - Sequence of packets on the line

$$S = \left\langle P_1, P_2, \ldots, P_N \right\rangle$$

- A network packet (P) consists of

$$P = \left\langle H, PL \right\rangle$$

  - H:= Header := $\left\langle h_1, h_2, \ldots, h_k \right\rangle$
  - PL:= Payload := $\left\langle b_1, b_2, \ldots, b_l \right\rangle$ (the payload could be empty)
  - Header fields can belong to different protocols

- Each header field (h_i) can consist of number of values (w_j)

- For each of these values a counter is defined $z_i \in \aleph$ which indicates, how often a specific value of a header field has already occurred

$$h_i \in \{w_1, w_2, \ldots, w_l\}$$

(see examples)

Internet Analysis System (IAS)
→ Counting of header information (2/3)

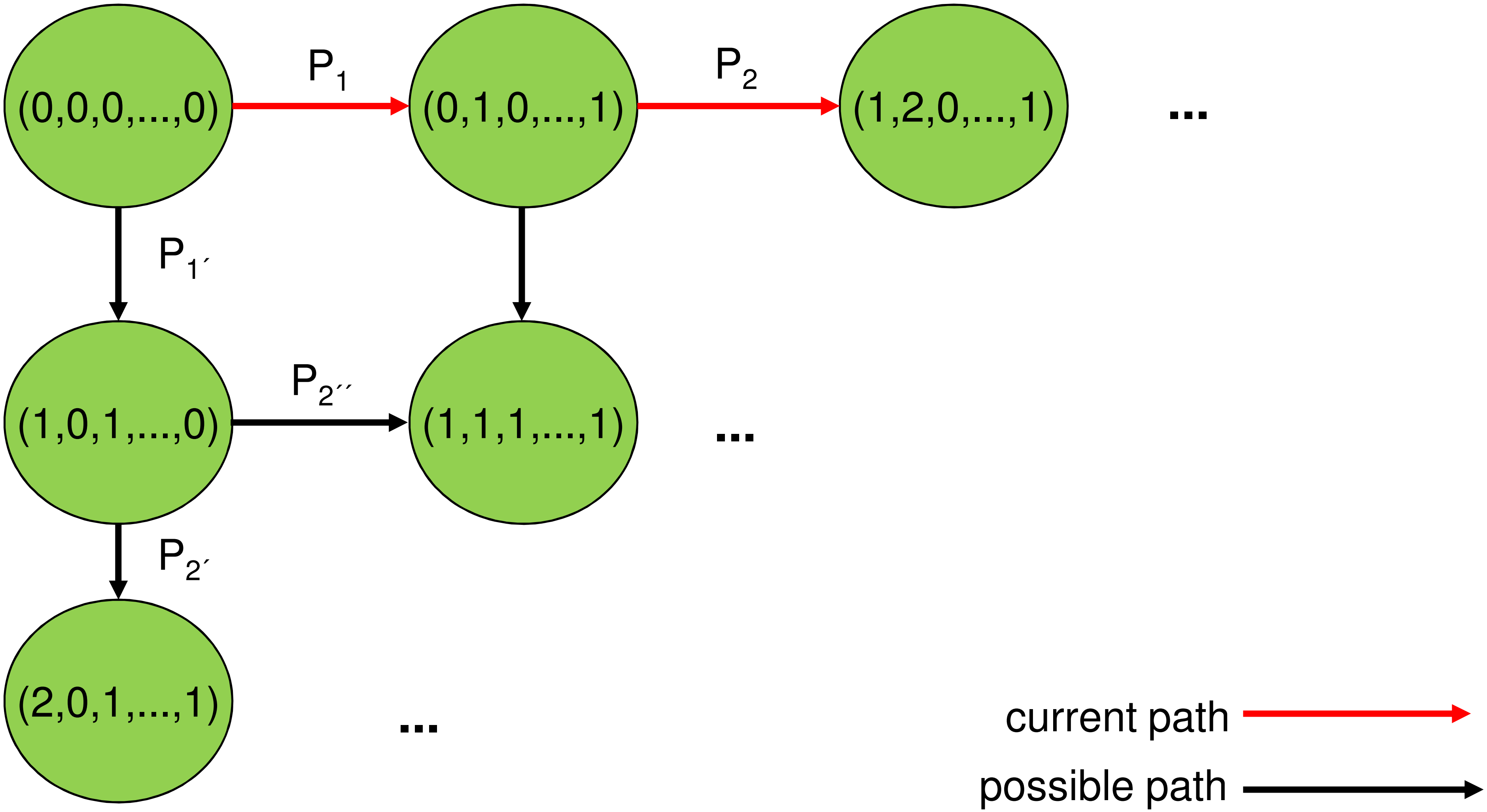
- This current setting of values of the set of counters can be understood as the current state (Z)

$$Z =< z_1, z_2, \ldots, z_P >$$

- With each packet the current state changes

=> **In principle a state machine, for which the counter values have been defined as states, runs from state to state as it processes the network packets as input data (for each of the time intervals)**

Internet Analysis System (IAS)
→ Counting of header information (3/3)



Internet Early Warning System
→ Evaluation counter (1/5)

Protocol	Number	Protocol	Number
DNS	9.458	EDONKEY	53
EMULE	19	Ethernet II	6
FTP	103	HTTP	1.123
HTTPS	179	ICMP	318
IKEv2	10.764	IMAP	40
IMAPS	179	IP	9.089
IPCO	4	IPSEC-AH	513
IRC	499	ISAKMP	4.912
META	14	P2P	6
POP	1.015	POPS	179

Internet Early Warning System
→ Evaluation counter (2/5)

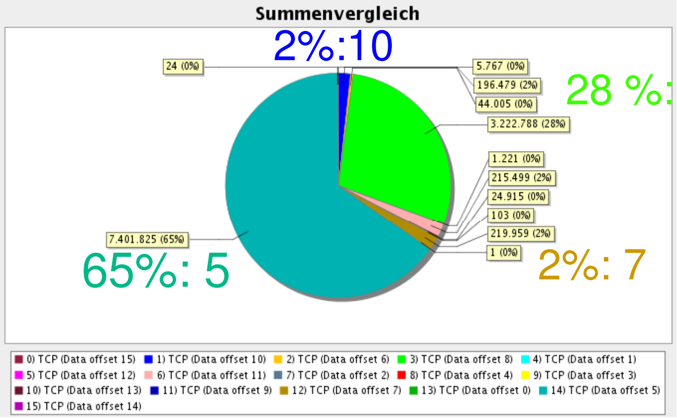
Protocol	Number	Protocol	Number
RTP	37	SIP	138
Skype	1	SMTP	1.624
SMTPS	179	TBURL	23.986
TCP	678.614	TFTP	17
UDP	131.590		
Sum	**876.596**		

Internet Early Warning System
→ Evaluation counter (3/5)

TCP	**678.614**	
reserved	1	
High ports (P2P definition)	4	
ecn	8	
HLEN	16	
TCP Flags (Code Bits)	66	
Window (size)	255	
options	522	
Well-known P-D (1024) + TTL (8)	8.192	(1.024 * 8)
Well-known P-S/D (1024) + Flags combi (8)	16.384	(2 * 8.192)
Port (Source/Destination – S/D)	131.072	(2 * 2^{16})
Well-known P-S/D (1024) + total length (255)	522.240	(2 * 261.120)

Internet Early Warning System
→ Evaluation counter (4/5) – TCP - HLEN

- **TCP – HLEN** (16)
  - Specifies the size of the TCP header in 32-bit words. The minimum size header is 5 and the maximum is 15 words.
  - Data Offset 0
  - Data Offset 1
  - Data Offset 2
  - Data Offset 3
  - Data Offset 4 (standard **HLEN 5**)
  - Data Offset 5 (+ option)
  - Data Offset 6 (+ option **HLEN: 7**)
  - Data Offset 7 (+ option **HLEN: 8**)
  - …
  - Data Offset 15 (+ option)



HLEN: 7
MSS (4 Byte)
NOP (2 Byte)
SACK permitted (2 Byte)

HLEN: 8
NOP (2 Byte)
Timestamp (10 Byte)

HLEN: 10
NOP (2 Byte)
SACK (18 Byte)
oder
MSS (4 Byte)
SACK permitted (2 Byte)
NOP (1 Byte)
WScale (3 Byte)

Internet Early Warning System
→ Evaluation counter (5/5) – P2P

- **P2P Counter**
  - UDP Port 4672
  - Source >= 1024 and Destination >= 1024 (« P2P »)
    - If both the source- and the destination-port are greater than or equal to 1024, this is an approximate estimate of the client-to-client communication, as for that only ports in the upper part are elected.
  - Source < 1024 and Destination < 1024 (« B2B »)
    - If both the source- and the destination-port are lower than 1024, this is an approximate estimate of the server-to-server communication, as for that only ports in the lower part are elected.
  - Source >= 1024 and Destination < 1024 (« P2B »)
  - Source < 1024 and Destination >= 1024 (« B2P »)

Internet Early Warning System
→ Evaluation counter (4/4)

- Distribution of the counters

	UNI Santa Maria	Computer Science Department	Dt. Messe AG	Dr. Buelow & Masiak
Max	115.343	16.462	26.167	75.595
Min	45.889	7.474	4.343	26.120
average	72.795	11.264	12.263	48.376
Total count of packets (P2P)	555.555 (150.00)	36.111	61.361	154.400
Union	276.287	125.696	148.291	267.840

Calculation of used parameters
→ Example (1/2)

- **Computer Science Department :**
  - Assumption: 100 users at the same time
  - 15 TCP connections per user → 15 * 100 * 2 ports → 3.000 parameters
  - 20 UDP connections per user → 20 * 100 * 2 ports → 4.000 parameters
    - Rough: 7.000 parameter
    - Real: average 7.474 parameter

- **Deutsche Messe AG: 700 employees in Germany**
  - Assumption: 200 users at the same time
  - 15 TCP connections per user → 15 * 200 * 2 ports → 6.000 parameters
    - Rough: 6.000 parameter
    - Real: average 4.343 parameter

Calculation of used parameters → Example (2/2)

- **Santa Maria: up to 19.000 people on campus**
  - Assumption: 1.000 users + 250 P2P users (25%) at the same time
  - 15 TCP connections per user → 15 * 1.000 * 2 ports → 30.000 parameters
  - 10 UDP connections per user → 10 * 1.000 * 2 ports → 20.000 parameters
  - 20 P2P-TCP connections per user → 20 * 250 * 2 ports →10.000 parameters

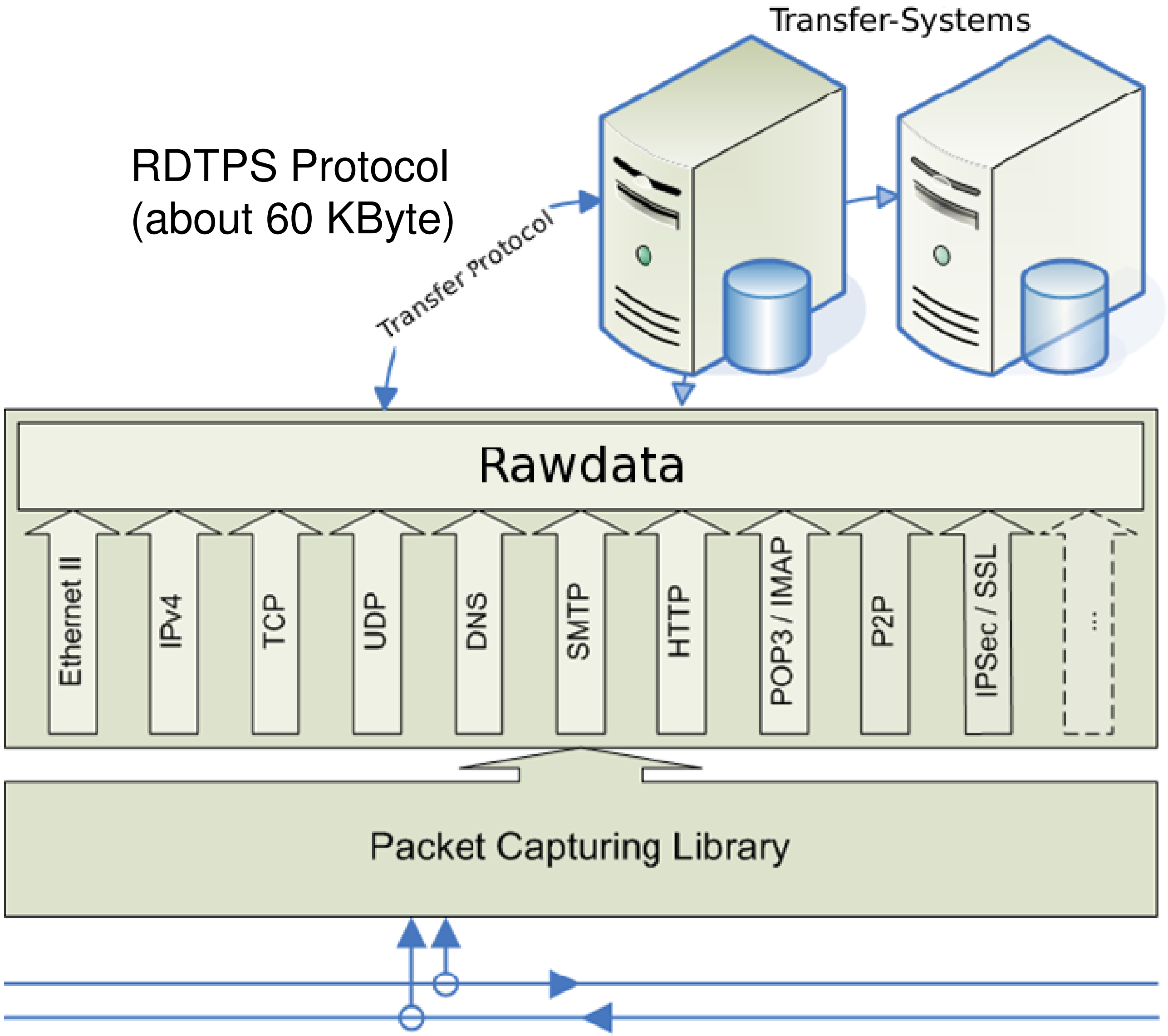
Rough:	60.000 parameter
Real: average	45.000 parameter

Principle of raw data collection
→ Why do we only use "tally sheets"?

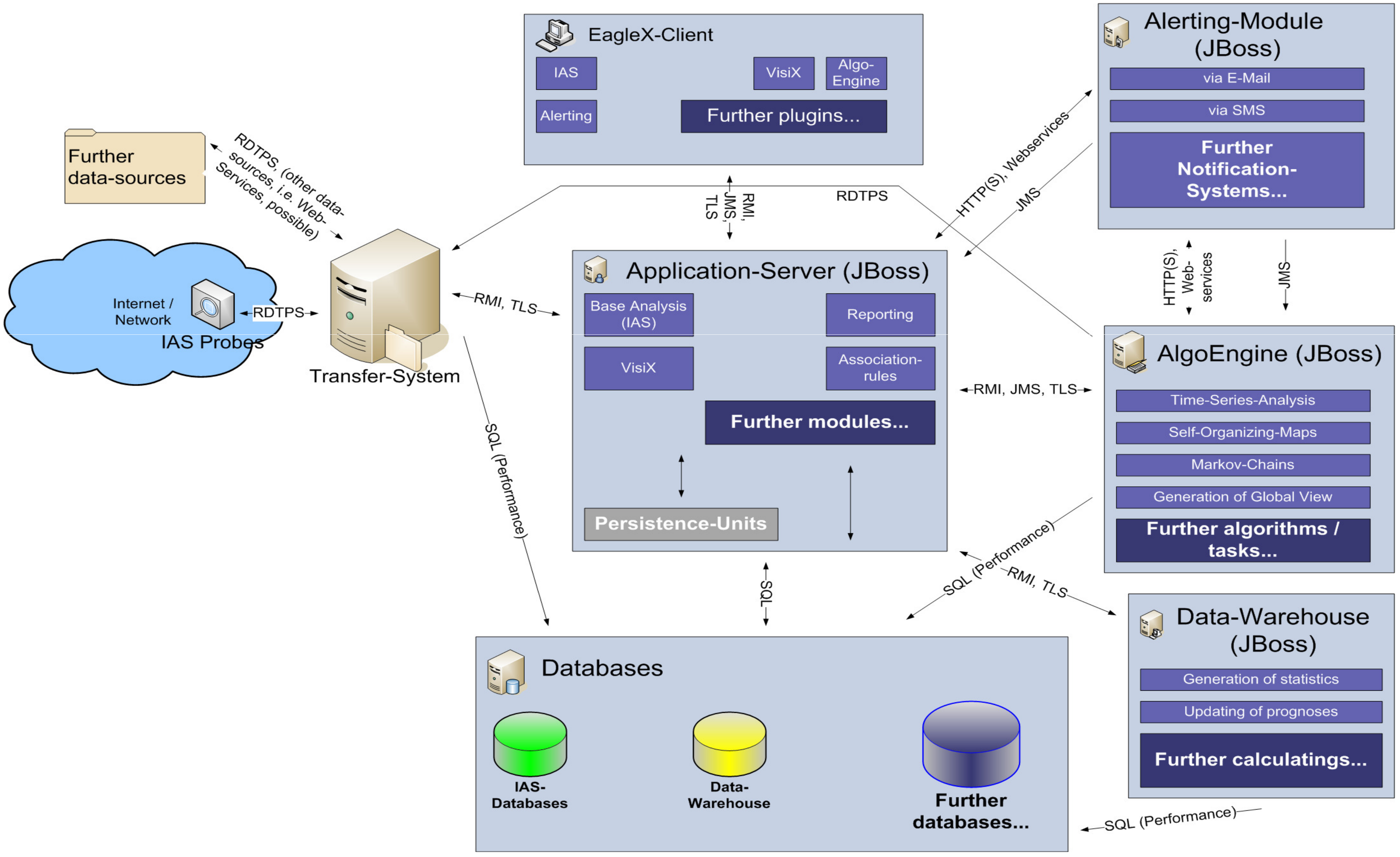
- **Enhances performance**
  - No tracking of connections or sessions
  - Irrelevant information can be ignored
    - e.g.: Checksums
- **Protection of critical information**
  - Since the connections and sessions cannot be put together again
  - Since critical content is left out from further processing
    - IP/ MAC addresses
    - User data
  - **Anonymization by design**

IAS: Current State of Development

→ Probe



IAS: Current State of Development
→ Architecture of the system



© Prof. Dr. Norbert Pohlmann, Institute for Internet Security - if(is), University of Applied Sciences Gelsenkirchen, Germany

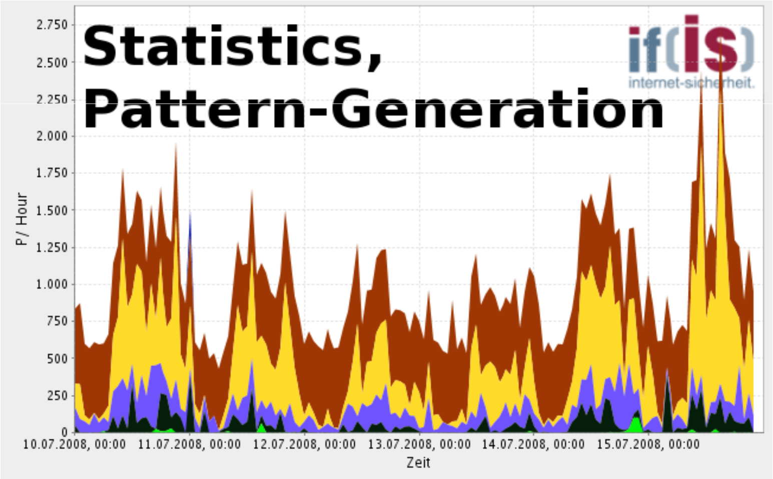
Content

- Aim and outcomes of this lecture
- Idea of the Internet Analysis System
- **Knowledge Base**
- Outline of the Current State
- Detection of Attacks and Deflection
- Forecast of Patterns and Attacks
- Summary

Internet Analysis System (IAS)
→ Defined targets

Target 1

- Description of profiles, patterns and coherences
- Creation of a knowledge base.



We want to create a knowledge base which we can use to understand the functioning of the internet from the “communication behaviour” point of view. The main task here is the support in analysing communication parameters – our row data - with the aim of identifying a pattern in the profiles, technological trends and correlations.

Internet Analysis System (IAS)
→ Process target 1

- Counting of communication parameters by the probe
- Transmitting of the counter readings (raw data) to the transfer system
- Long term storage in a database

Establishment of a knowledge base

- Preservation of the raw data in a database
- Gaining on experience and collection of events / incidents

Description of profiles, patterns, technology trends and other coherences.

- Analyzing of the raw data with the self-developed „***EagleX Analysis Client***“
  - Expert tool, which can be used to analyze raw data manually
- (automated) generation of reports

Knowledge base - IAS
→ Coherences (1/2)

- **Coherences in architectural matters**
  - When http is detected, then also **TCP and IP**
- **Coherences in protocol matters**
  - When we detect a http **request**, then we should detect a http **response** as well
- **Coherences due to system matters**
  - When we detect **http traffic**, then in most cases we have also recorded **DNS traffic**
- **Coherences coming form behavior**
  - E.g. When we detect **http**, then we also see **SMTP**, this means, when we **surf** online we also **write e-mails**

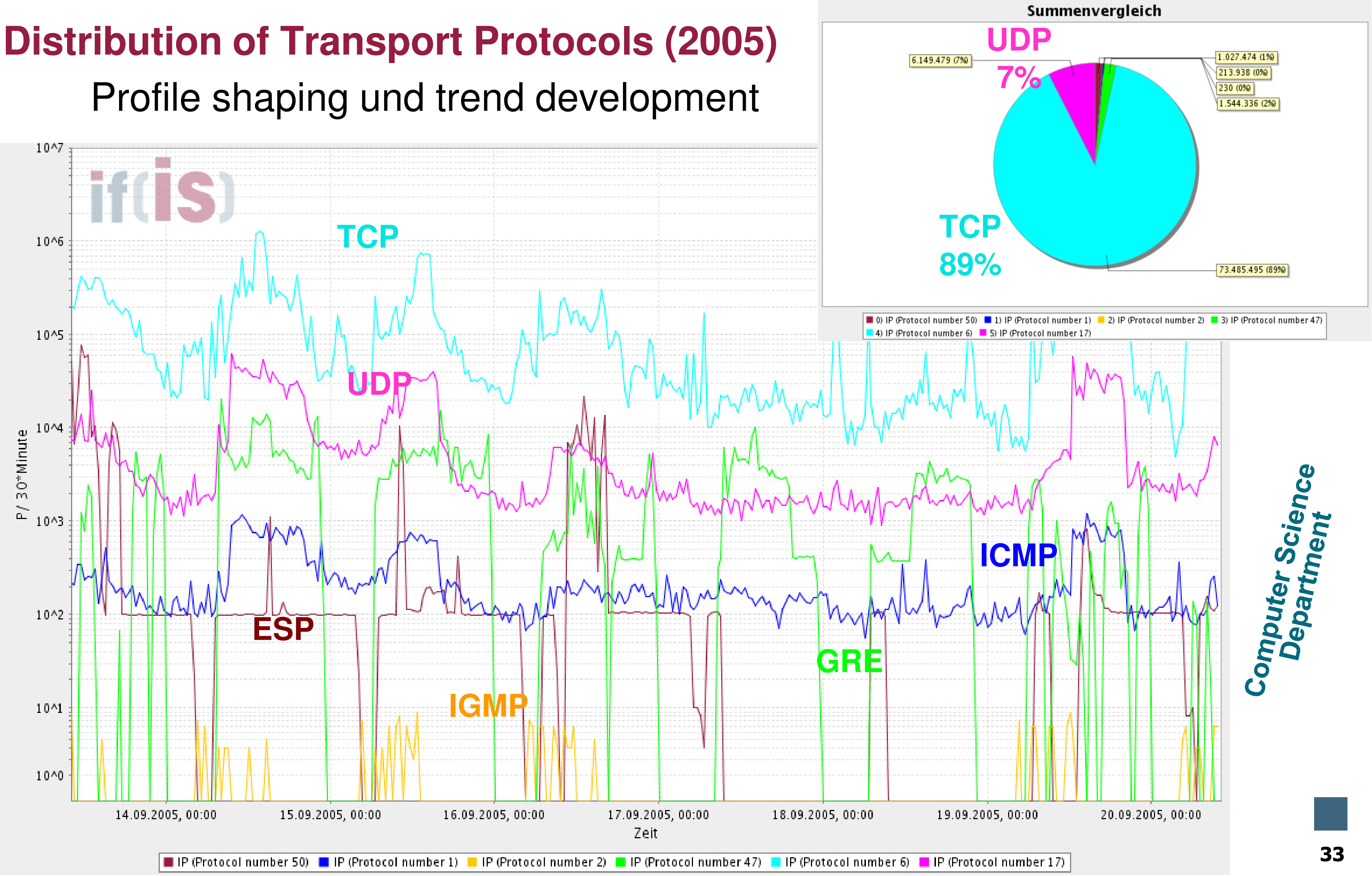
Knowledge base - IAS
→ Coherences (2/2)

- **Coherences due to situations**
  - When a news with an **important impact** is broadcasted, e.g. an act of terror, then we can see a lot more **Internet traffic**
- **Coherences because of the location of the probe and because of certain applications**
  - DSL provider, content provider and business user have very different colored Internet traffic depending on the services and applications used
  - For instance we can detect a lot more p2p traffic in the network of a DSL provider than in the network of a business user

Knowledge base - IAS
→ Result: Distribution transport protocol

- ## Distribution of Transport Protocols (2005)

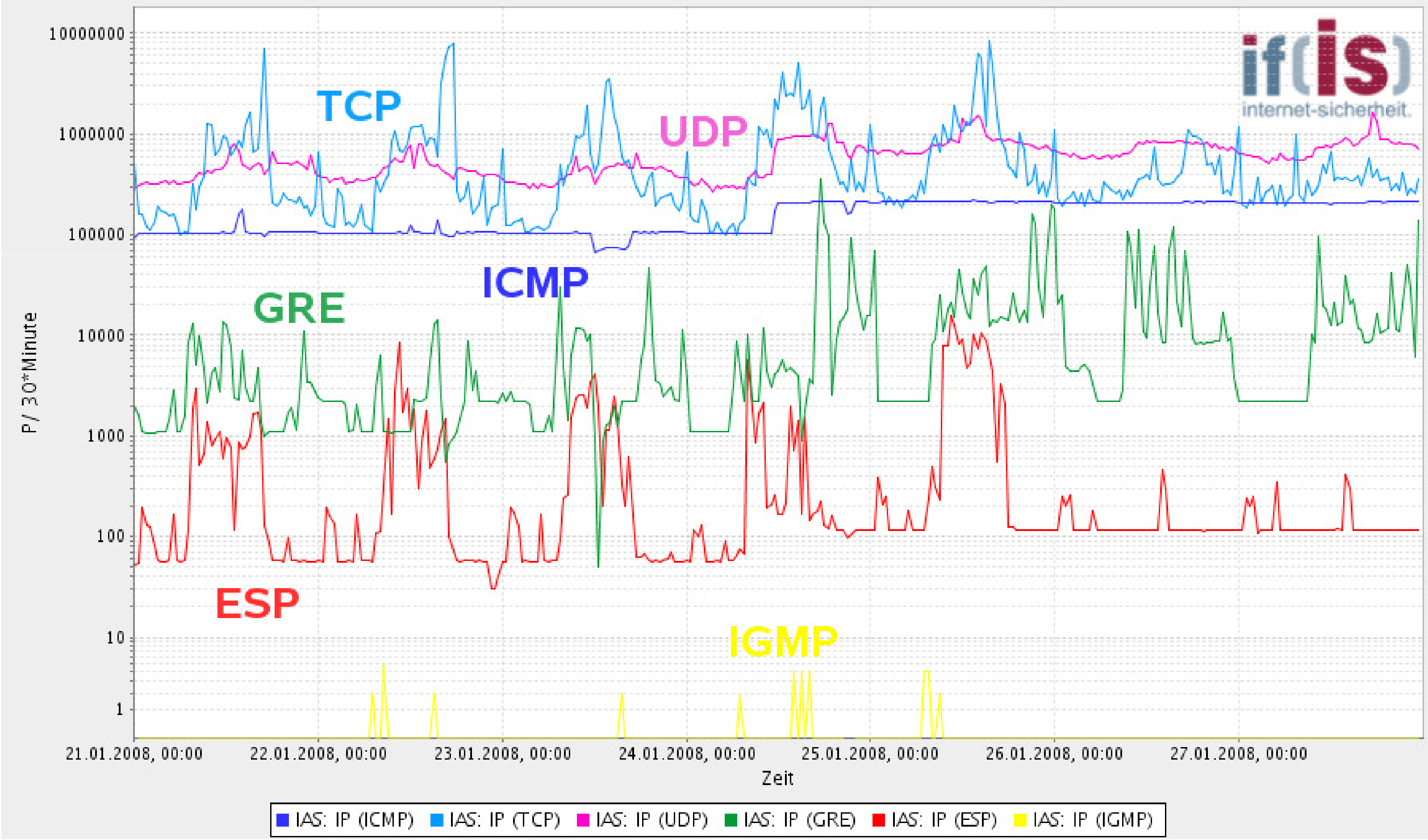
Profile shaping und trend development

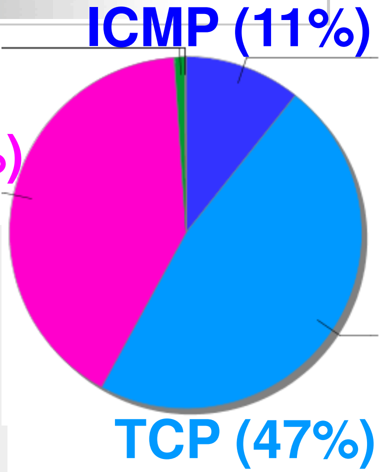


Knowledge base - IAS
→ Result: Distribution transport protocol

- **Distribution of Transport Protocols (2008)**

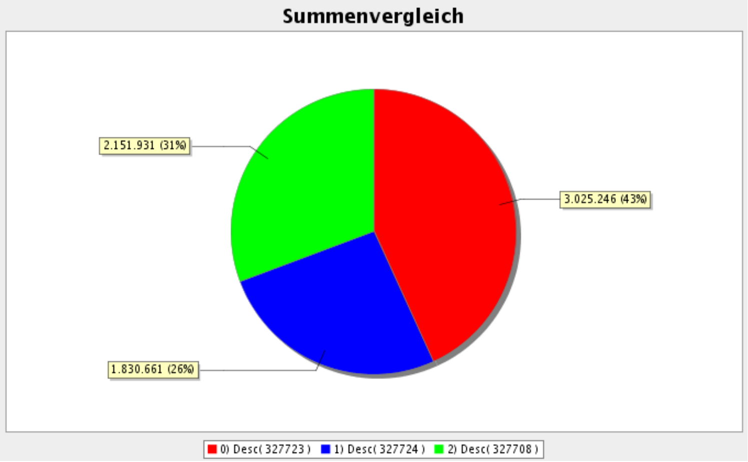
Profile shaping und trend development





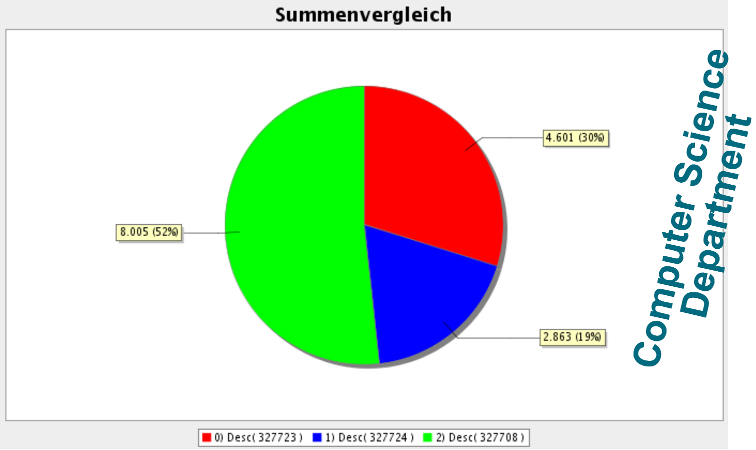
Knowledge base - IAS
→ Result: Expected distribution

- **SYN-Scan (Potential Attack)**
  - Comparison between different periods
    - Expected: SYN > SYN/ACK > 2xFIN/ACK (TCP teardown handshake)
  - Gap between expected spreading and spreading in case of an attack → Detection of attacks



Expected Distribution

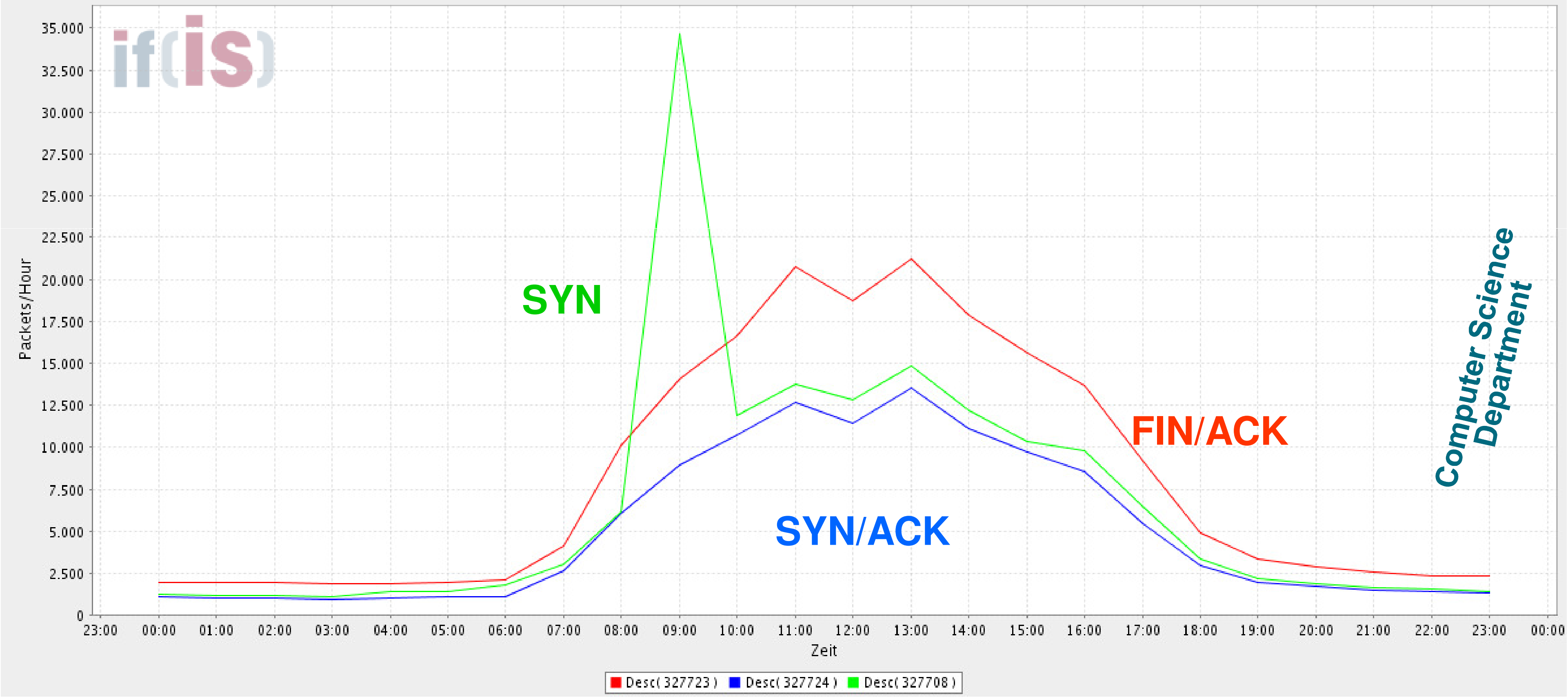
SYN
(31% - 52%)
SYN/ACK
(26% - 19%)
FIN/ACK
(43% - 30%)



Unexpected Distribution

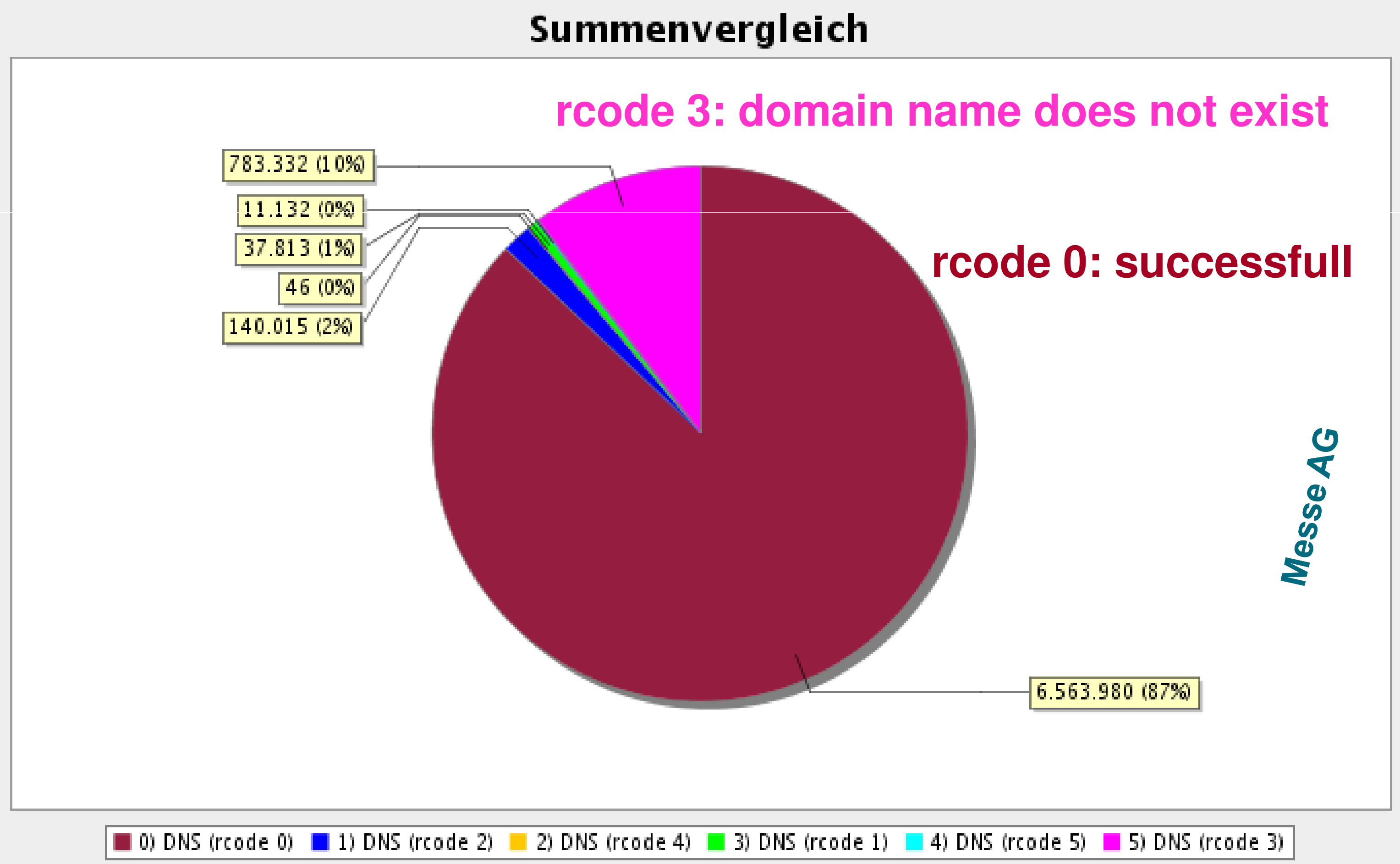
Knowledge base - IAS
→ Result: Detection of attacks

- **SYN-Scan (Potential Attack)**
  - Period of SYN scan can easily be detected



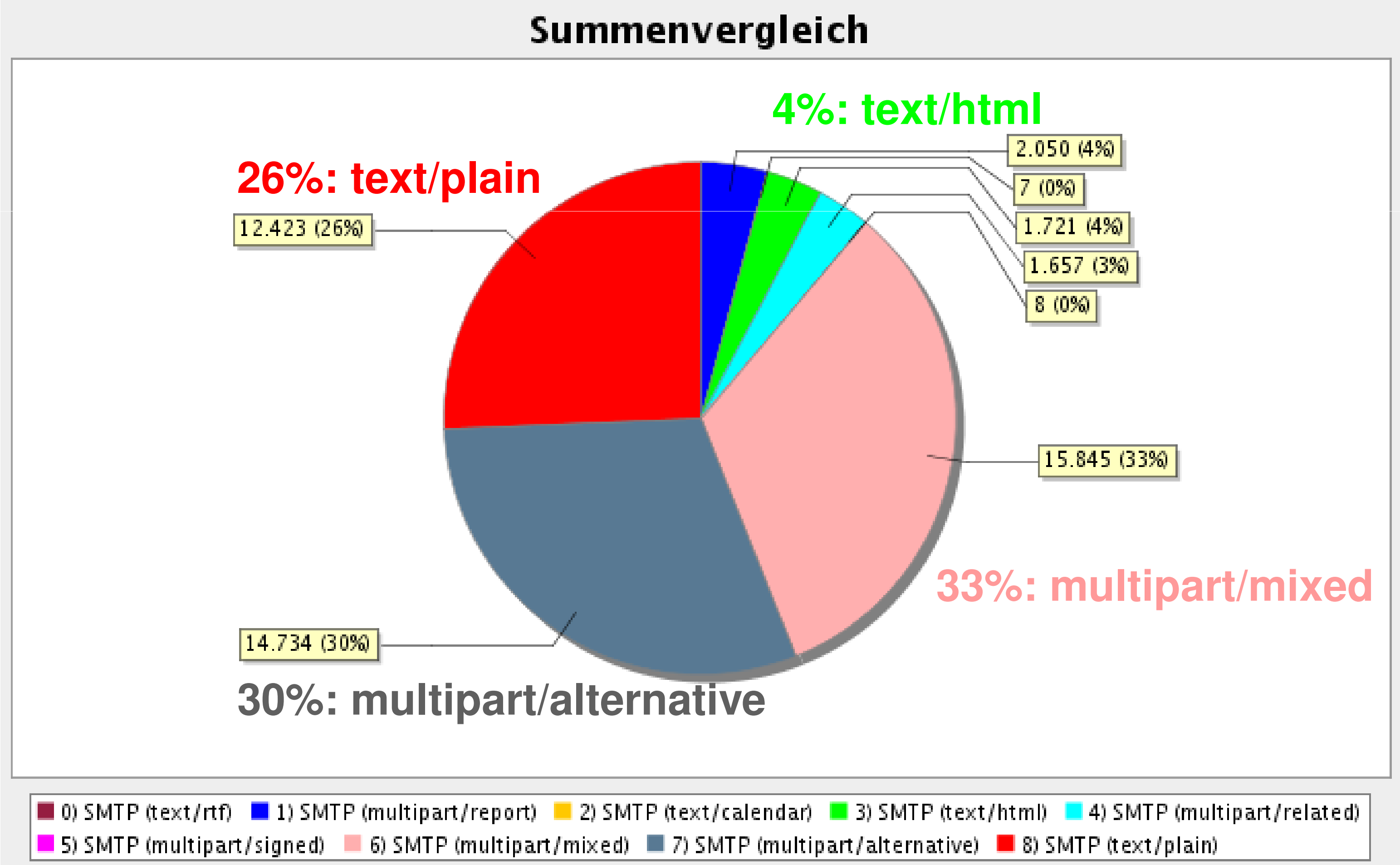
Knowledge base - IAS
→ Result: Examples

- **DNS Server Return Codes**
  - Normal distribution: Everything Ok
  - About 10%: Domain name not found



Knowledge base - IAS
→ Result: Distribution SMTP Content Type

- **SMTP Content Type**
  - 60% “text” Mails
  - 33 % “attachments”



4%: text/html

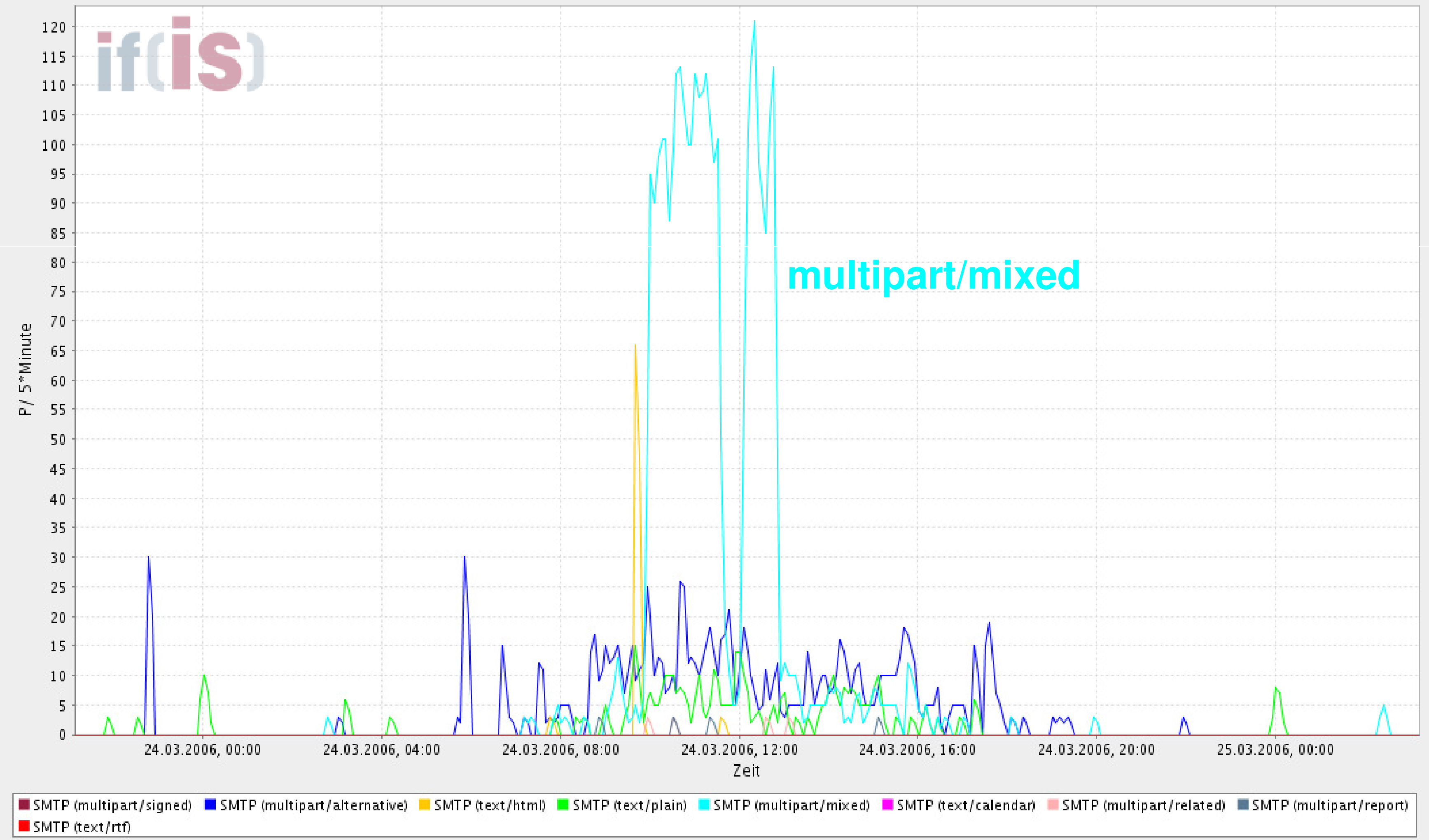
26%: text/plain

33%: multipart/mixed

30%: multipart/alternative

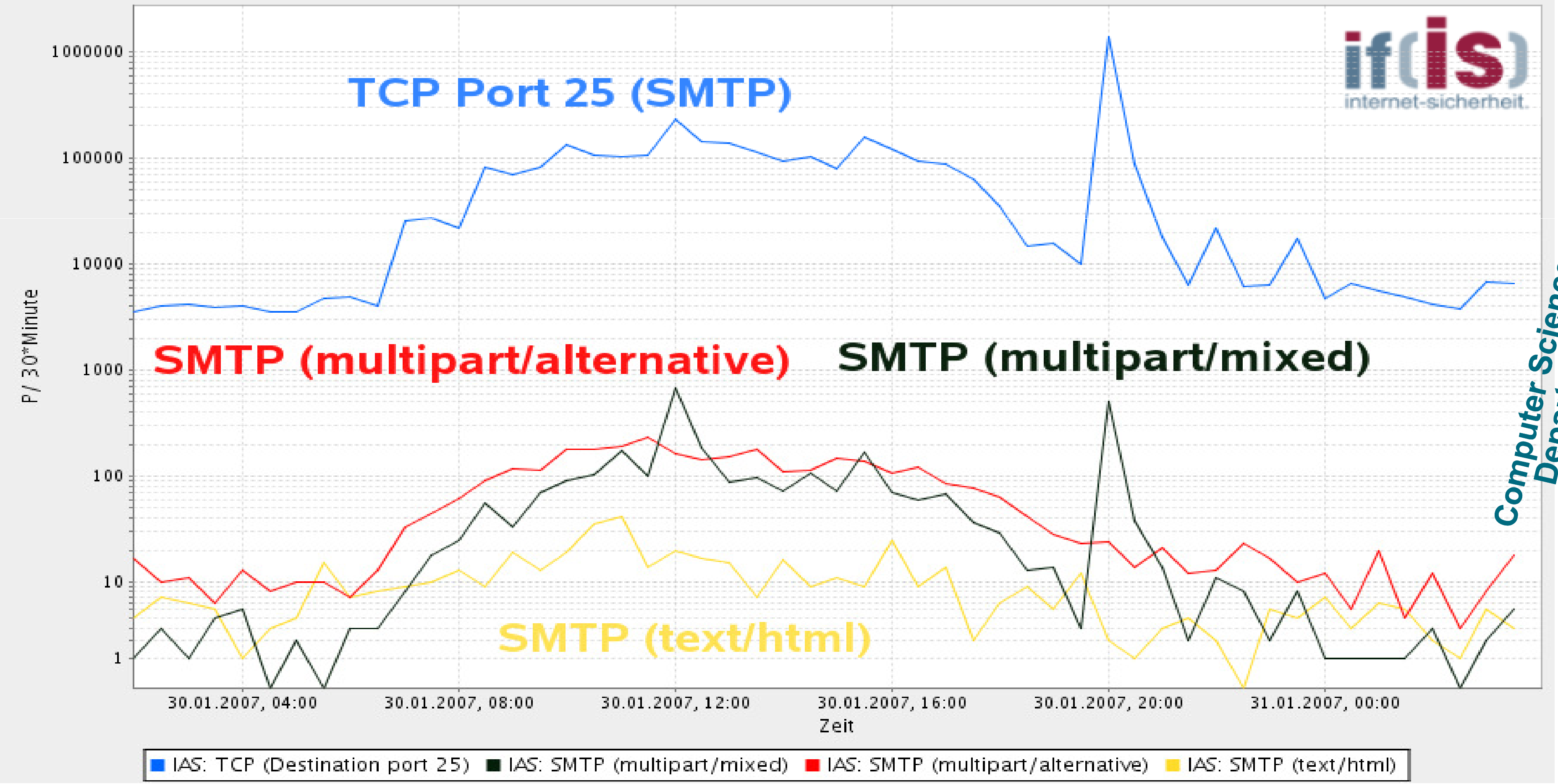
Knowledge base - IAS
→ Result: Detection of attacks (1/2)

- **SMTP Content Type**
  - Temporarily more e-mails witch attachments -> Mail-Virus!



Knowledge base - IAS
→ Result: Detection of attacks (2/2)

- **BKW worm (Sober.Z)**
  - The waves were transmitted in January 2007 concentrated at 3 pm and/or 8 pm.

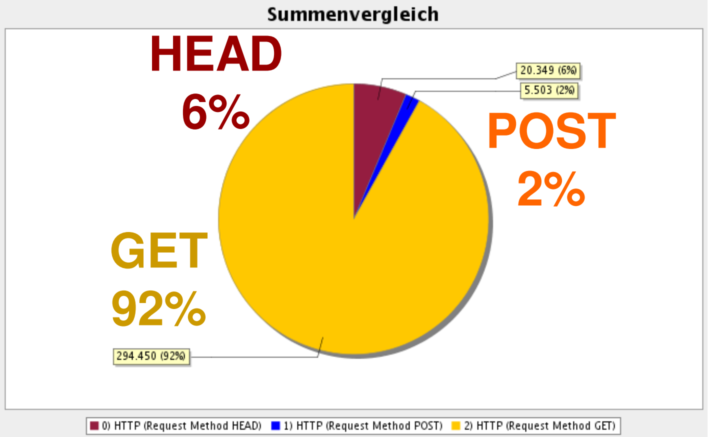


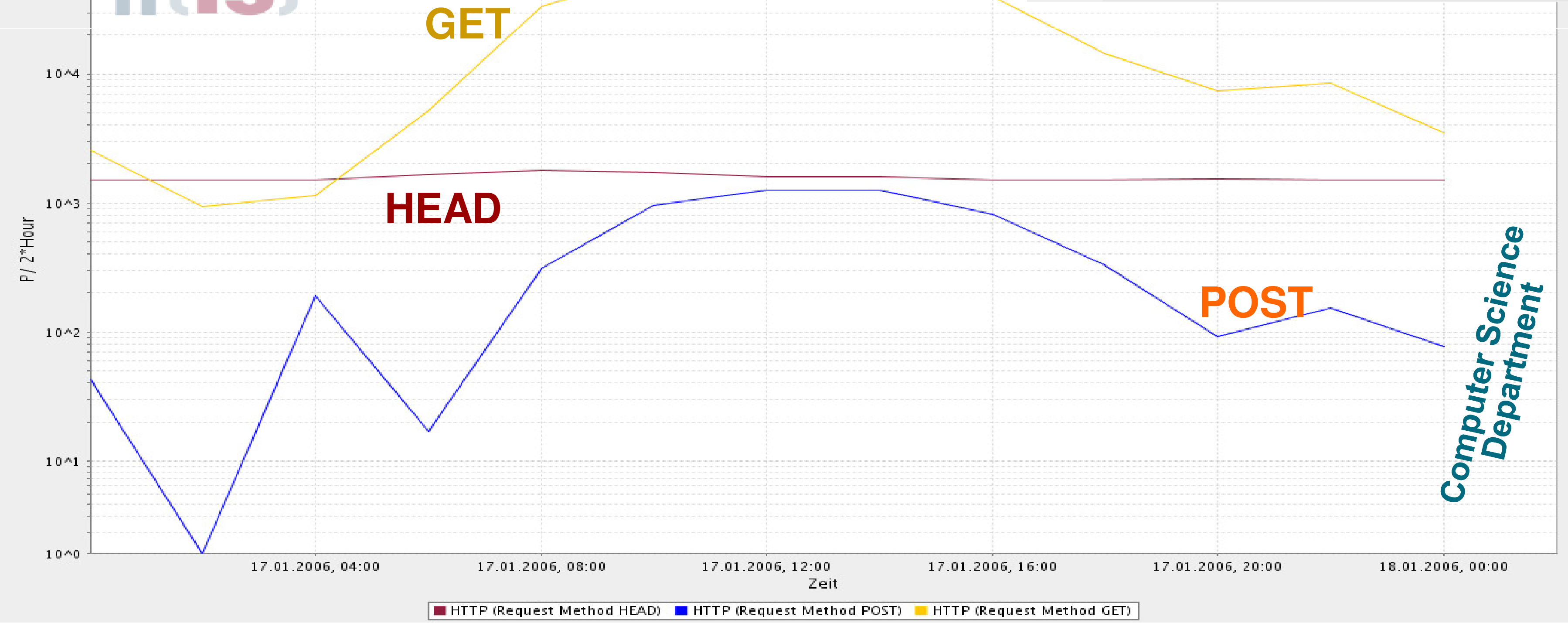
Computer Science Department

Knowledge base - IAS
→ Result: Distribution HTTP Methods

if(is) internet security.

- **HTTP Methods**
  - Diurnal rhythm
    - HEAD used by automated processes
    - GET und POST usually used by human users

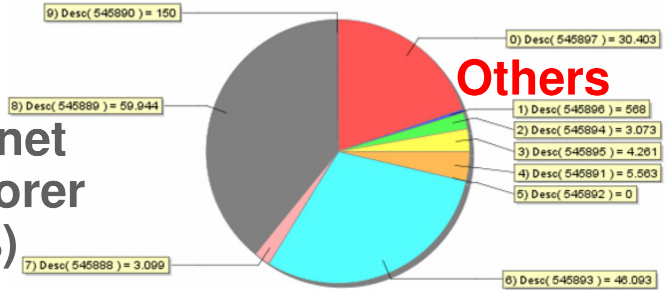




Computer Science Department

Knowledge base - IAS
→ Result: Technology trend (1/2)

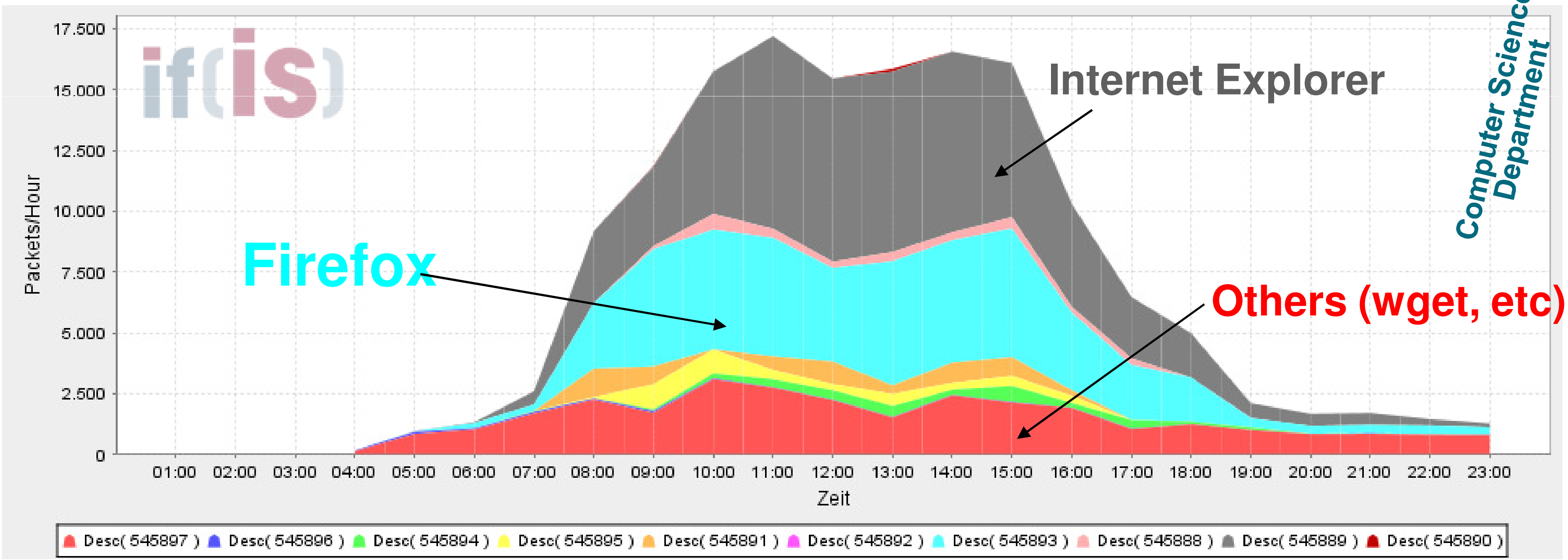
- **Distribution of browsers (2005)**
  - Diurnal profile
  - Differences between manual use (e.g. Internet Explorer und Firefox) and automated use (z.B. wget) are detectable.



Others

Internet Explorer (39%)

Firefox (30%)



Internet Explorer

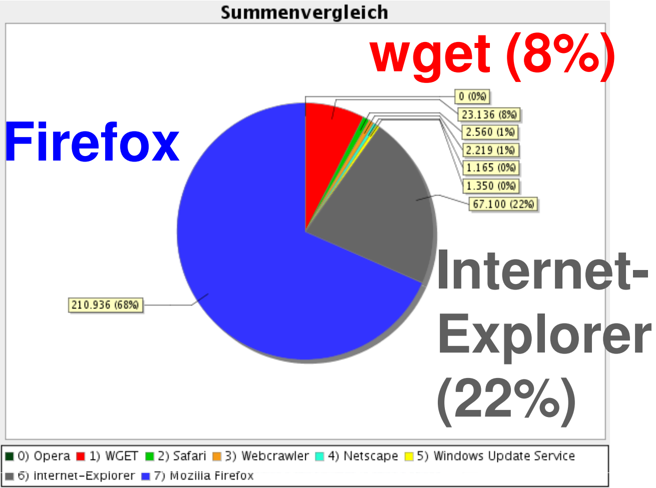
Firefox

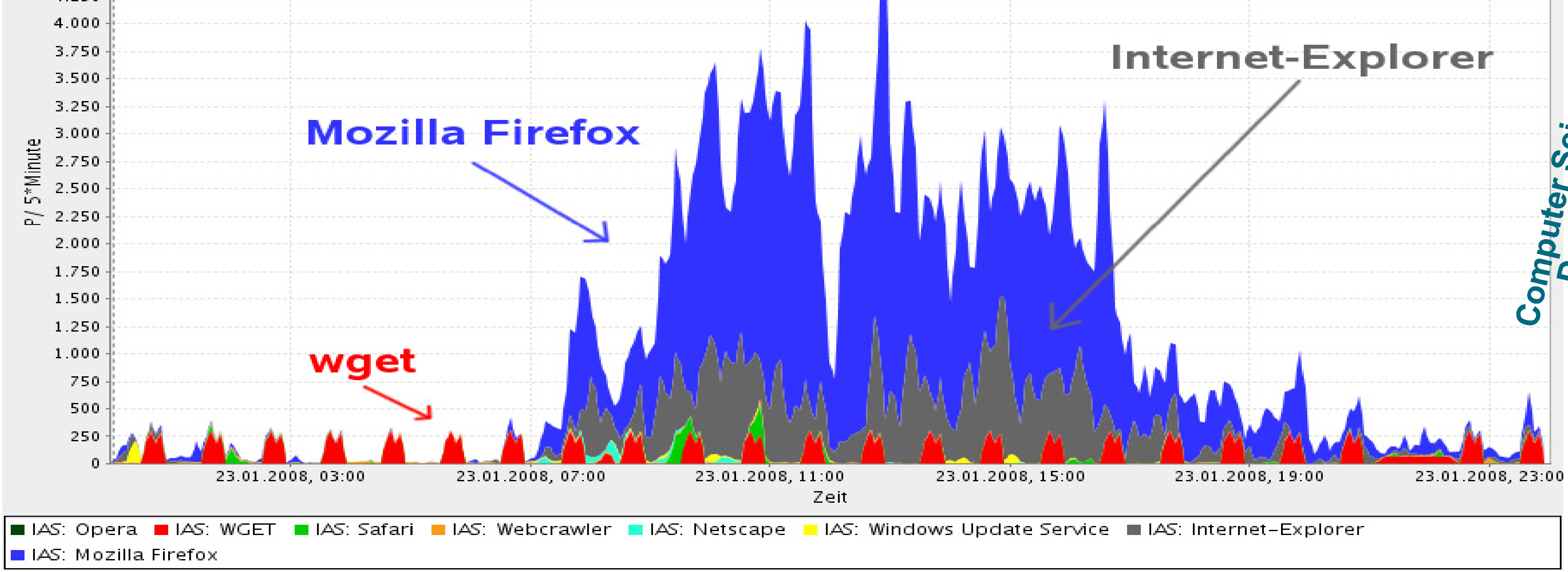
Others (wget, etc)

Computer Science Department

Knowledge base - IAS
→ Result: Technology trend (2/2)

- **Distribution of browsers (2008)**
  - Diurnal profile
  - Differences between manual use (e.g. Internet Explorer und Firefox) and automated use (z.B. wget) are detectable.

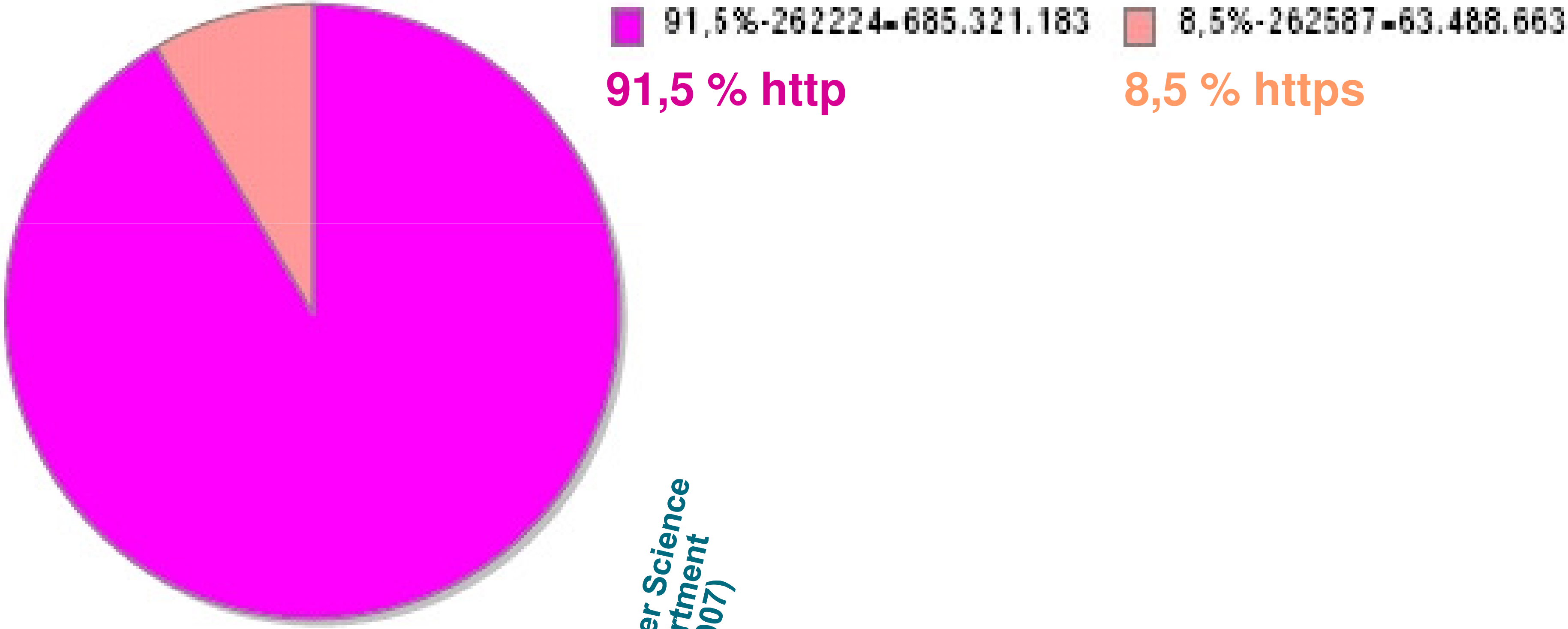




Computer Science Department

Knowledge base - IAS
→ Result: HTTP / HTTPS

- **Distribution of encrypted HTTP-Session**



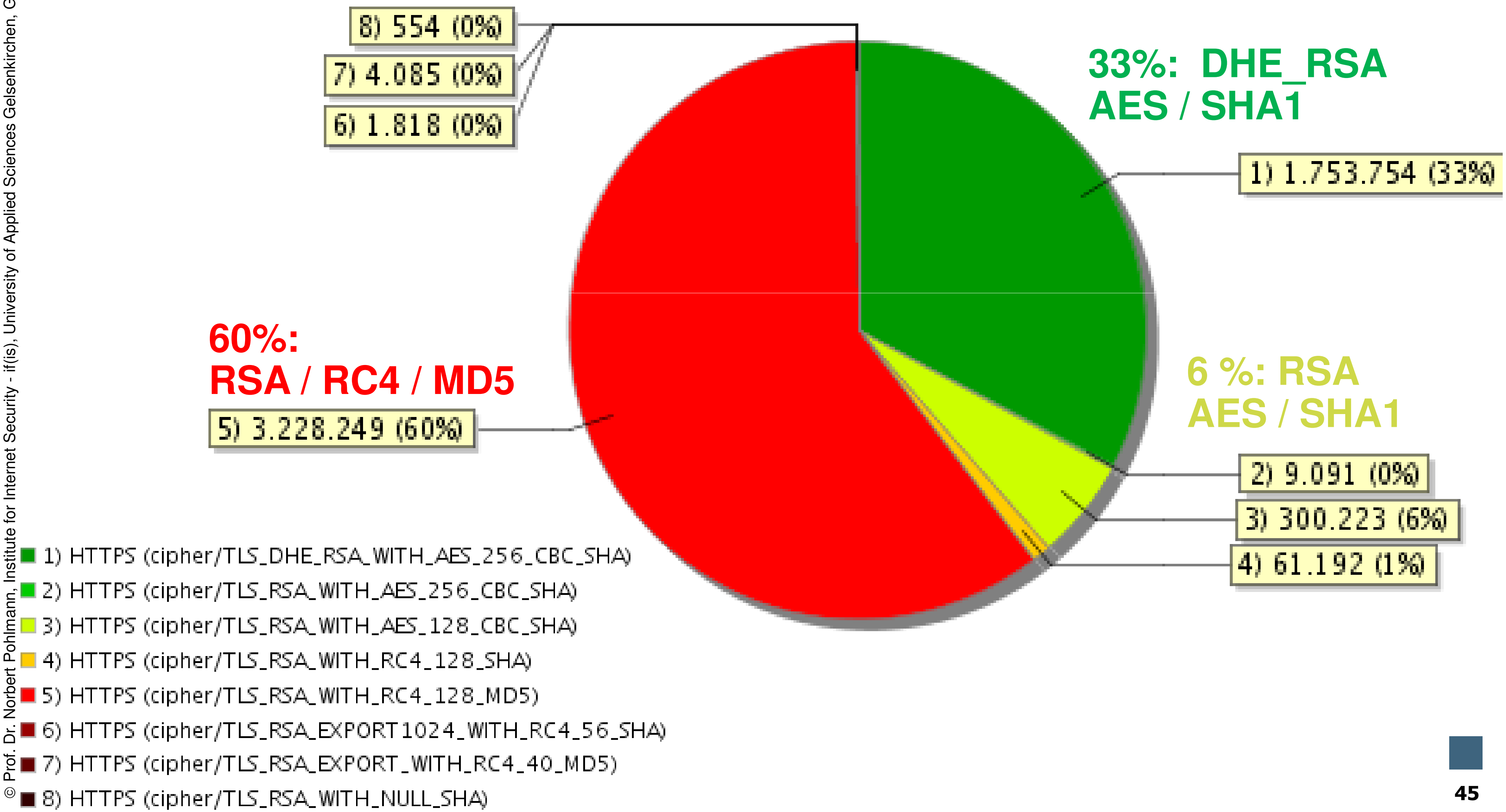
91,5 % http

8,5 % https

Computer Science Department (2007)

Knowledge base - IAS
→ Result: Awareness (Crypto used TLS)

!! 0.1 %: RSA / Export (40) / SHA1 and 0.01 %: RSA / NULL / SHA1 !!



33%: DHE_RSA AES / SHA1

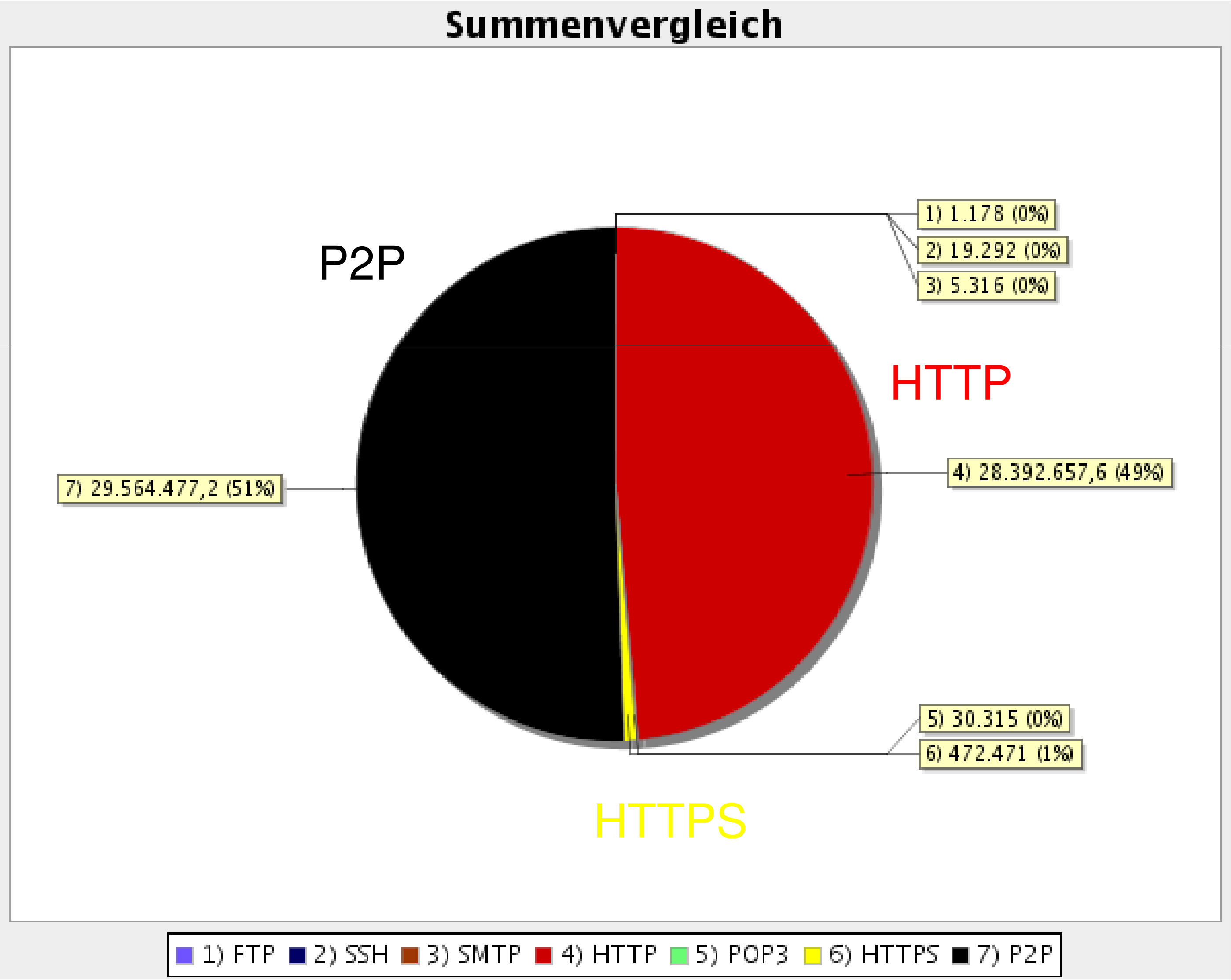
60%: RSA / RC4 / MD5

6 %: RSA AES / SHA1

- 🟩 1) HTTPS (cipher/TLS_DHE_RSA_WITH_AES_256_CBC_SHA)
- 🟩 2) HTTPS (cipher/TLS_RSA_WITH_AES_256_CBC_SHA)
- 🟩 3) HTTPS (cipher/TLS_RSA_WITH_AES_128_CBC_SHA)
- 🟧 4) HTTPS (cipher/TLS_RSA_WITH_RC4_128_SHA)
- 🟥 5) HTTPS (cipher/TLS_RSA_WITH_RC4_128_MD5)
- 🟥 6) HTTPS (cipher/TLS_RSA_EXPORT1024_WITH_RC4_56_SHA)
- 🟥 7) HTTPS (cipher/TLS_RSA_EXPORT_WITH_RC4_40_MD5)
- 🟥 8) HTTPS (cipher/TLS_RSA_WITH_NULL_SHA)

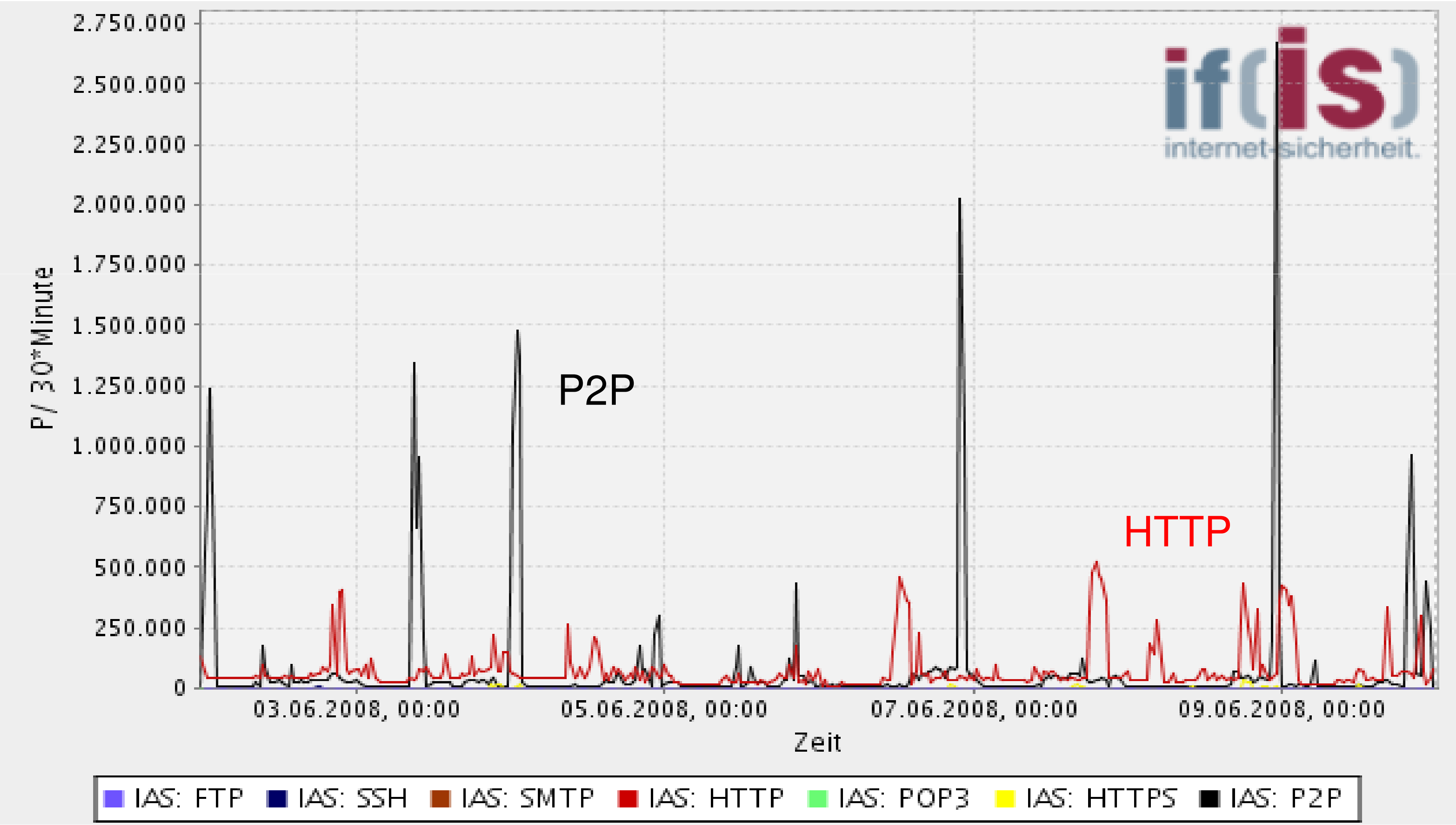
Knowledge base - IAS
→ Result: DSL-Connection (1/2)

- Distribution of protocols (sum)

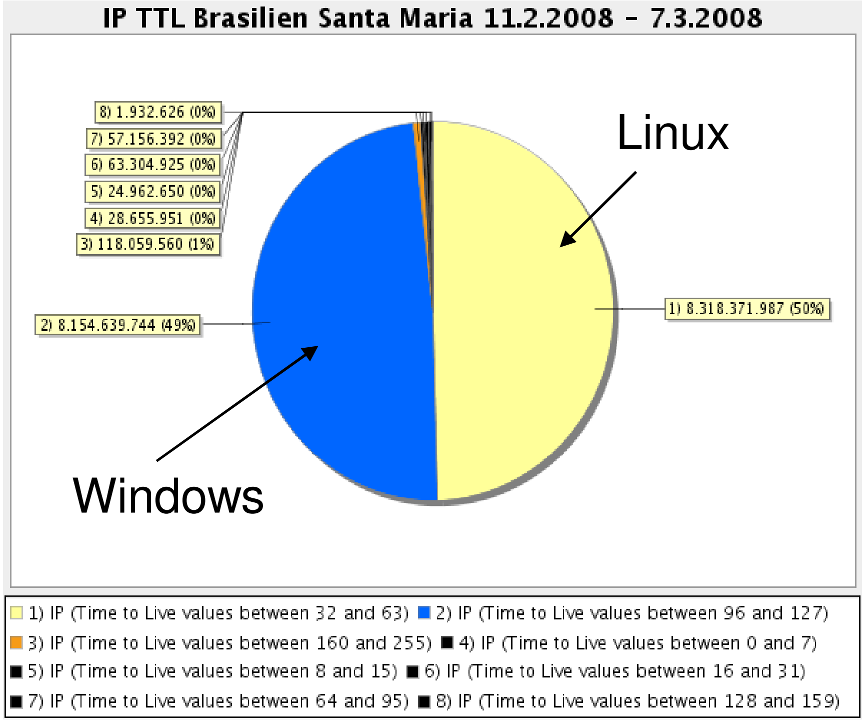


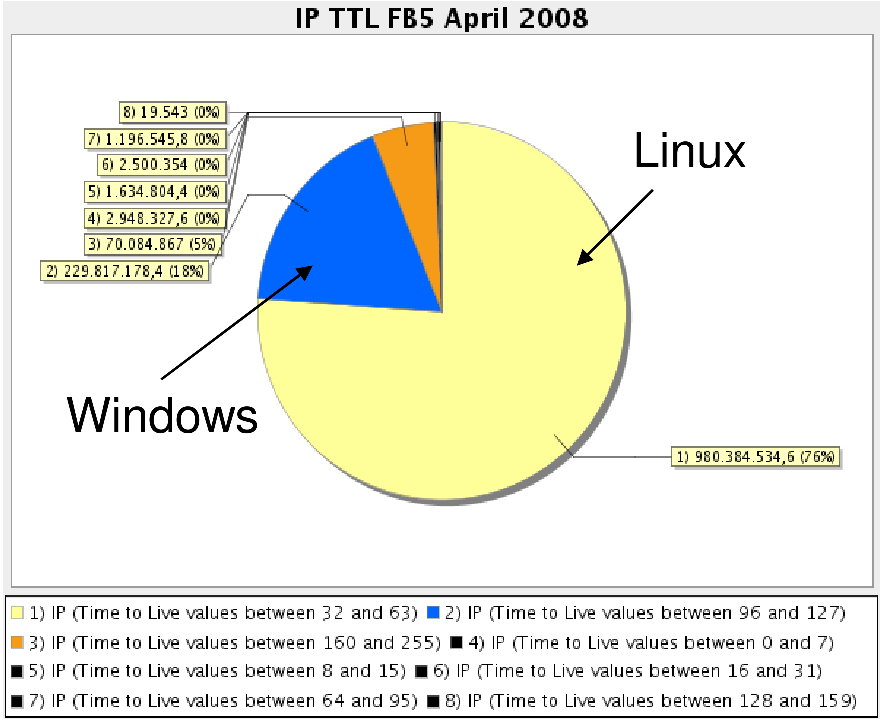
Knowledge base - IAS
→ Result: DSL-Connection (2/2)

- Distribution of protocols (over the time)



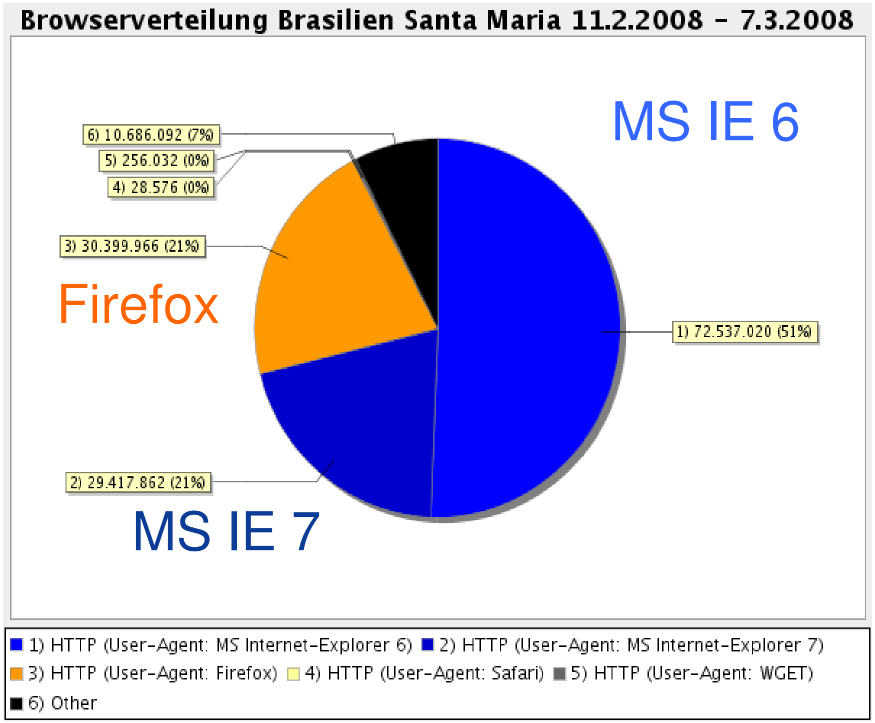
Knowledge base - IAS
→ Result: Different places (1/4)

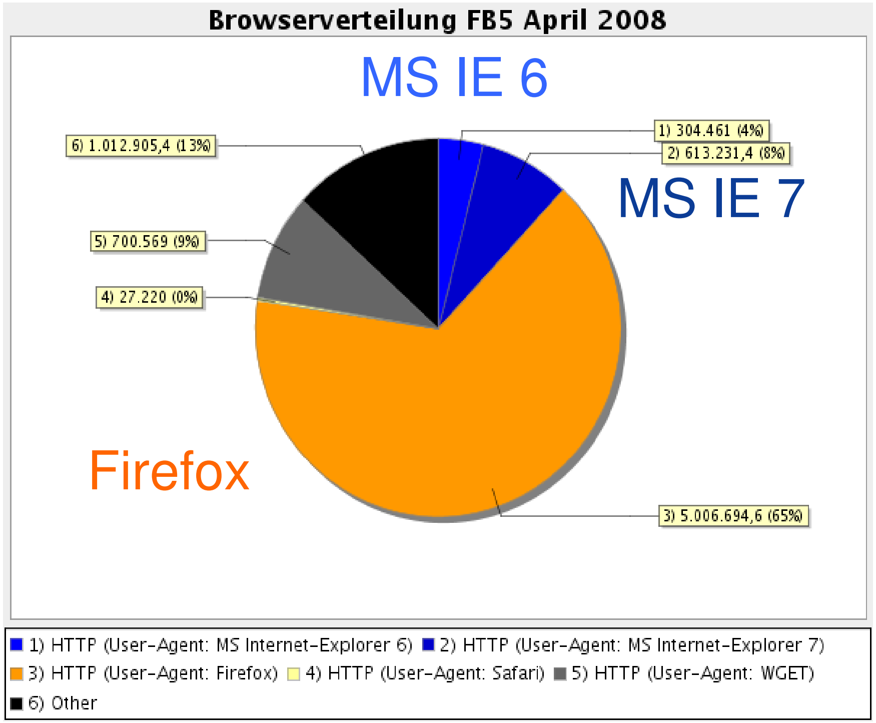




- TTL 64 value set by Linux.
- TTL 128 value set by Windows.
- TTL 255 value set by some Routers.
- Finding: a lot more Linux users at the department of computer science

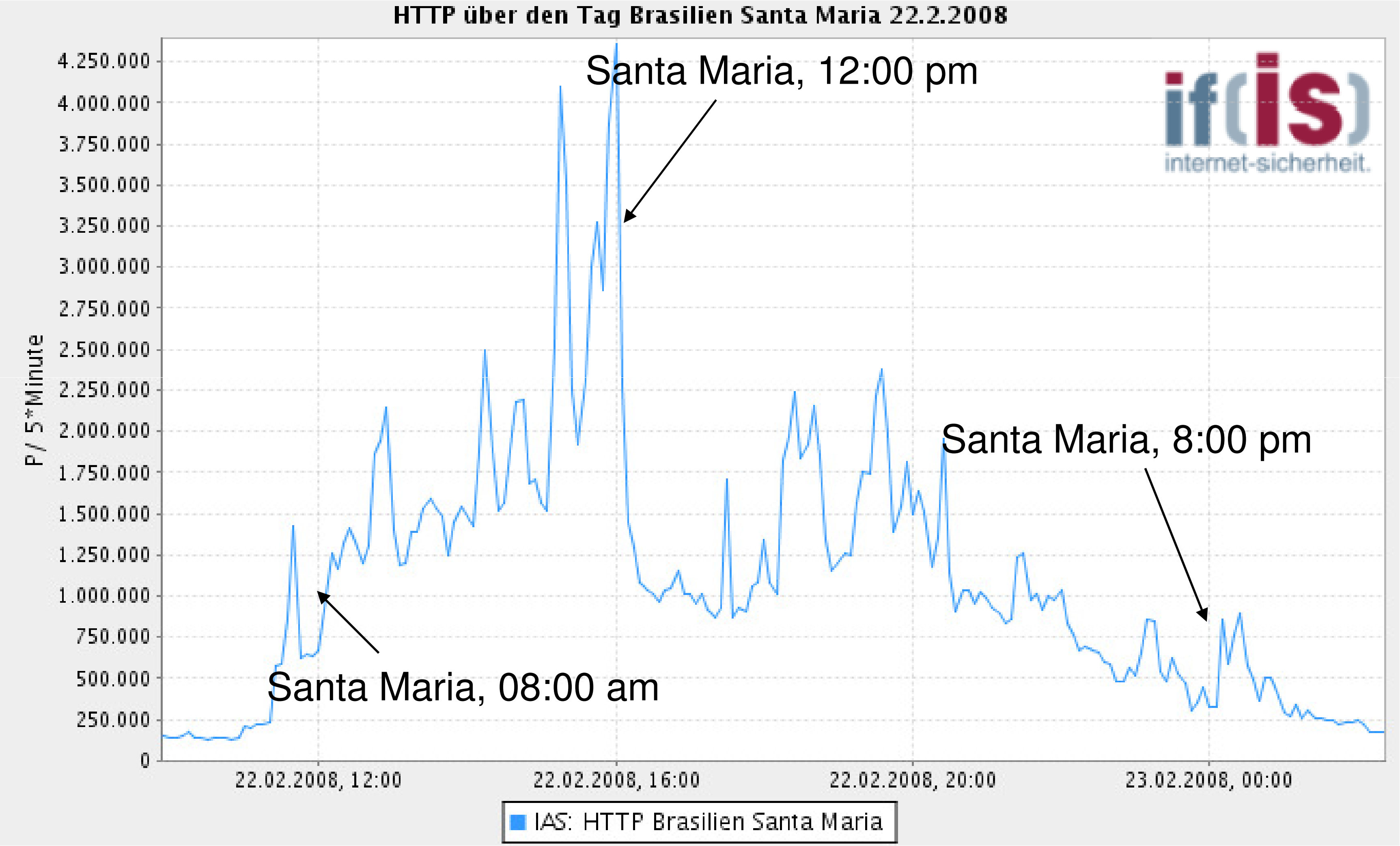
Knowledge base - IAS
→ Result: Different places (2/4)





- Department of computer science: large portion of Firefox users, even though windows is used as an operating system in the computer lab.
- Brazil: a lot of Internet Explorer, what requires a Microsoft operating system.

Knowledge base - IAS
→ Result: Different places (3/4)

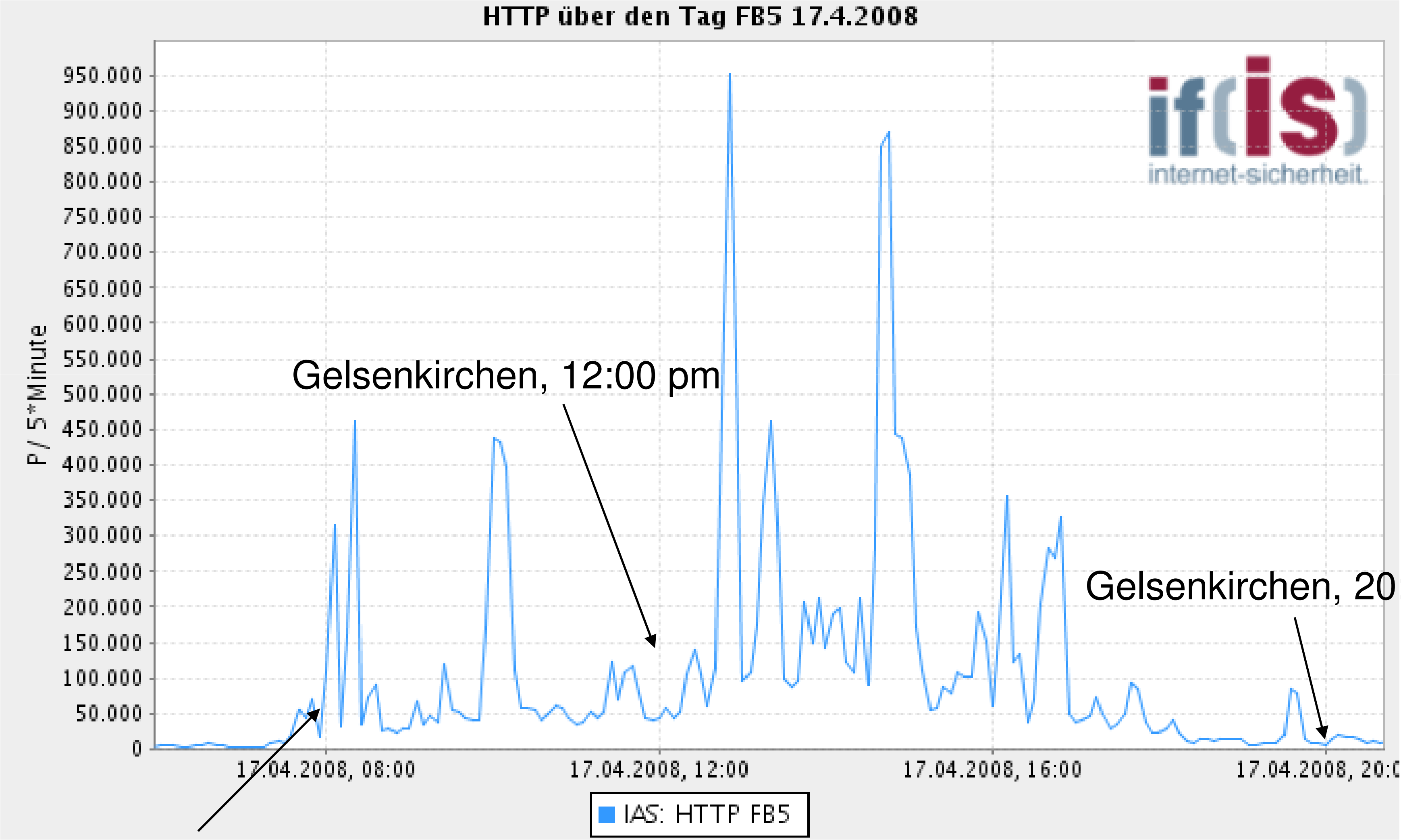


Santa Maria, 12:00 pm

Santa Maria, 8:00 pm

Santa Maria, 08:00 am

Knowledge base - IAS
→ Result: Different places (4/4)

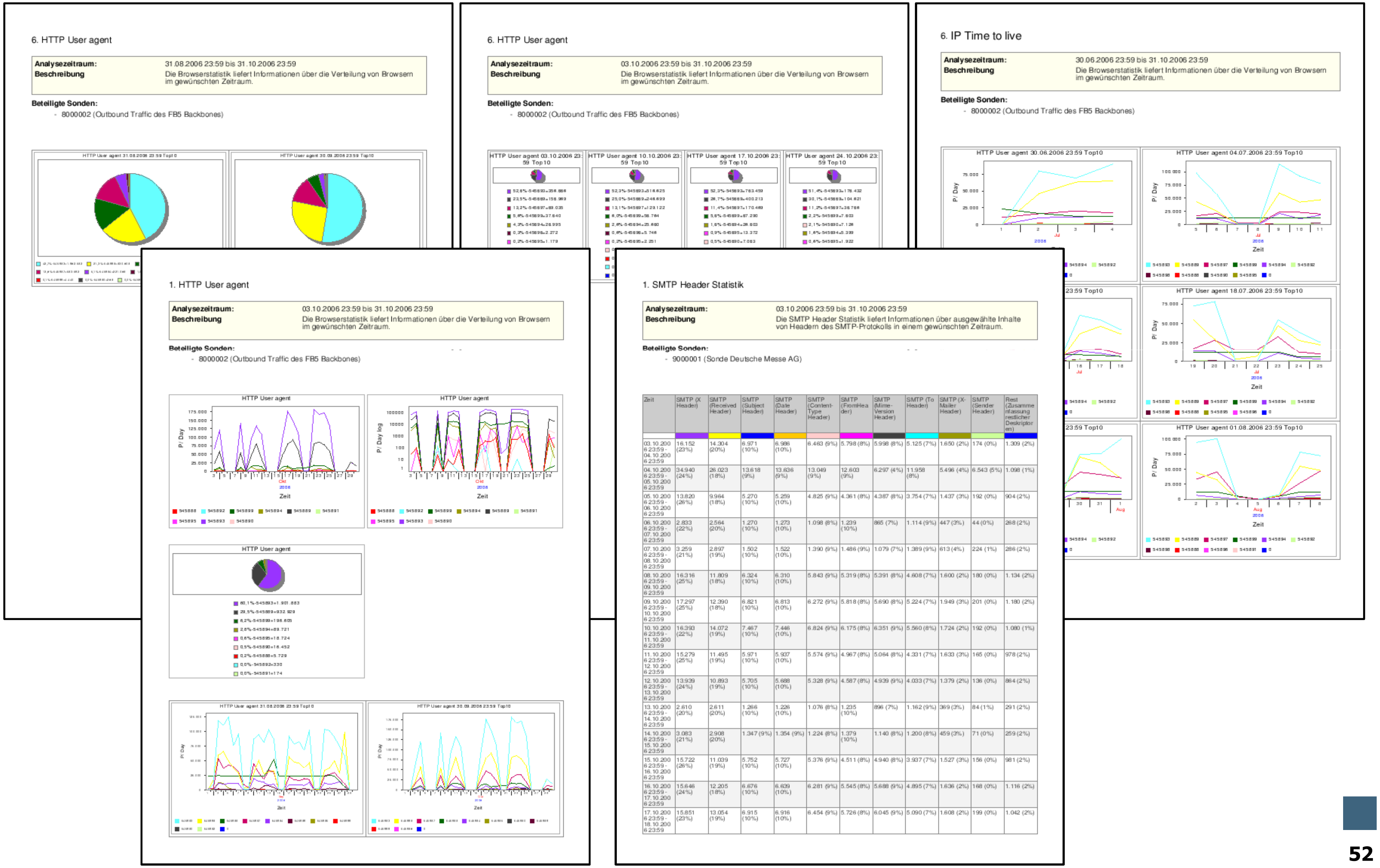


Gelsenkirchen, 12:00 pm

Gelsenkirchen, 20:00 pm

Gelsenkirchen, 08:00 am

IAS: Current State of Development
→ Situation Reports



6. HTTP User agent

Analysezeitraum: 31.08.2006 23:59 bis 31.10.2006 23:59
Beschreibung Die Browserstatistik liefert Informationen über die Verteilung von Browsern im gewünschten Zeitraum.

Beteiligte Sonden:
- 8000002 (Outbound Traffic des FB5 Backbones)

6. HTTP User agent

Analysezeitraum: 03.10.2006 23:59 bis 31.10.2006 23:59
Beschreibung Die Browserstatistik liefert Informationen über die Verteilung von Browsern im gewünschten Zeitraum.

Beteiligte Sonden:
- 8000002 (Outbound Traffic des FB5 Backbones)

6. IP Time to live

Analysezeitraum: 30.06.2006 23:59 bis 31.10.2006 23:59
Beschreibung Die Browserstatistik liefert Informationen über die Verteilung von Browsern im gewünschten Zeitraum.

Beteiligte Sonden:
- 8000002 (Outbound Traffic des FB5 Backbones)

1. HTTP User agent

Analysezeitraum: 03.10.2006 23:59 bis 31.10.2006 23:59
Beschreibung Die Browserstatistik liefert Informationen über die Verteilung von Browsern im gewünschten Zeitraum.

Beteiligte Sonden:
- 8000002 (Outbound Traffic des FB5 Backbones)

HTTP User agent
P/ Day
Zeit
HTTP User agent
P/ Day log
Zeit
HTTP User agent
HTTP User agent 31.08.2006 23:59 Top10
HTTP User agent 30.09.2006 23:59 Top10
HTTP User agent 03.10.2006 23:59 Top10
HTTP User agent 10.10.2006 23:59 Top10
HTTP User agent 17.10.2006 23:59 Top10
HTTP User agent 24.10.2006 23:59 Top10
HTTP User agent 30.06.2006 23:59 Top10
HTTP User agent 04.07.2006 23:59 Top10
HTTP User agent 18.07.2006 23:59 Top10
HTTP User agent 01.08.2006 23:59 Top10

1. SMTP Header Statistik

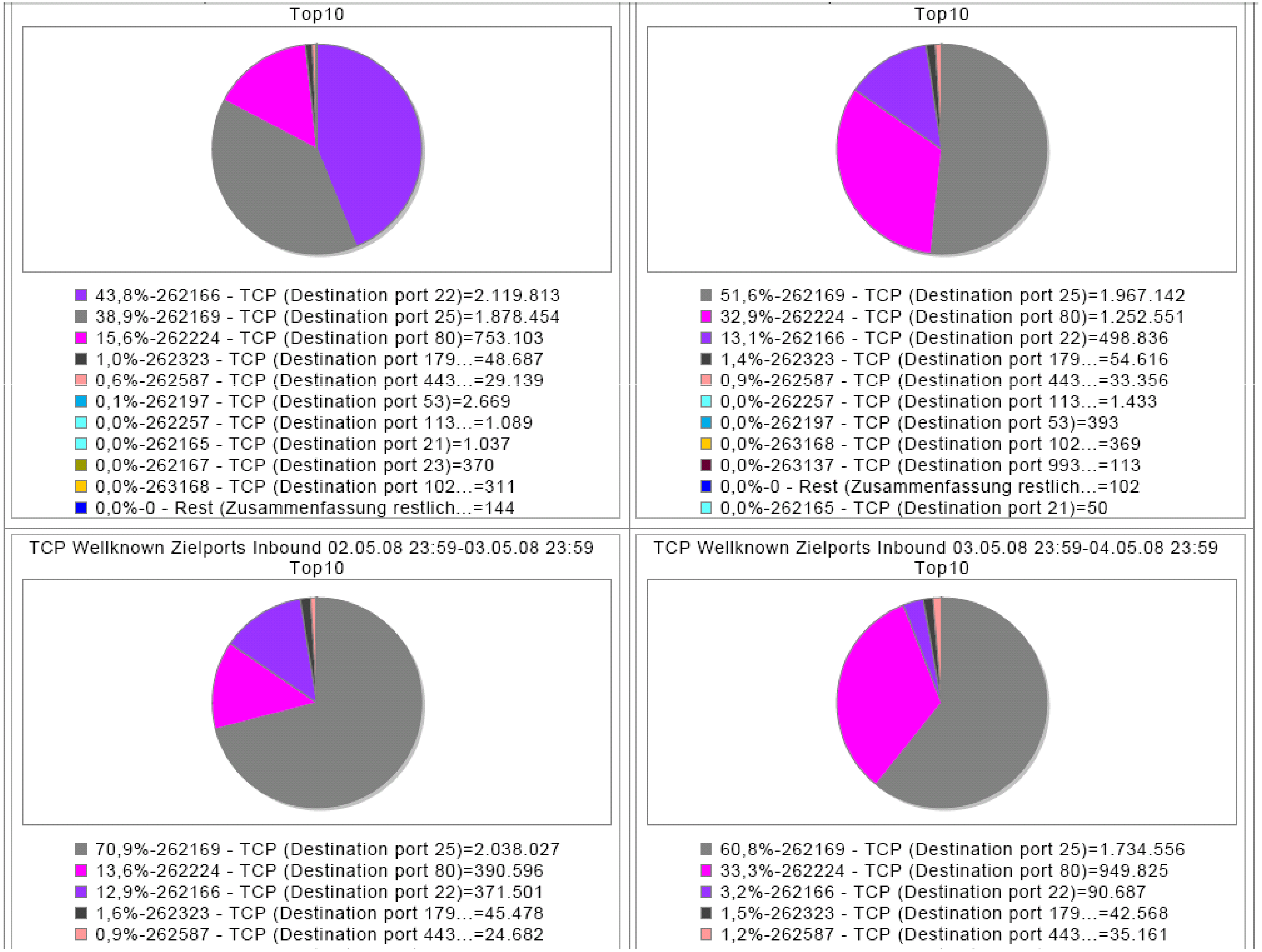
Analysezeitraum: 03.10.2006 23:59 bis 31.10.2006 23:59
Beschreibung Die SMTP Header Statistik liefert Informationen über ausgewählte Inhalte von Headern des SMTP-Protokolls in einem gewünschten Zeitraum.

Beteiligte Sonden:
- 9000001 (Sonde Deutsche Messe AG)

Zeit	SMTP (X Header)	SMTP (Received Header)	SMTP (Subject Header)	SMTP (Date Header)	SMTP (Content-Type Header)	SMTP (FromHeader)	SMTP (Mime-Version Header)	SMTP (To Header)	SMTP (X-Mailer Header)	SMTP (Sender Header)	Rest (Zusammenfassung restlicher Deskriptoren)
	🟪	🟨	🟦	🟧	▨	🟪	⬛	🟦	🟫	🟩	🟦
03.10.2006 23:59 - 04.10.2006 23:59	16.152 (23%)	14.304 (20%)	6.971 (10%)	6.986 (10%)	6.463 (9%)	5.798 (8%)	5.998 (8%)	5.125 (7%)	1.650 (2%)	174 (0%)	1.309 (2%)
04.10.2006 23:59 - 05.10.2006 23:59	34.940 (24%)	26.023 (18%)	13.618 (9%)	13.636 (9%)	13.049 (9%)	12.603 (9%)	6.297 (4%)	11.958 (8%)	5.496 (4%)	6.543 (5%)	1.098 (1%)
05.10.2006 23:59 - 06.10.2006 23:59	13.820 (26%)	9.964 (18%)	5.270 (10%)	5.259 (10%)	4.825 (9%)	4.361 (8%)	4.387 (8%)	3.754 (7%)	1.437 (3%)	192 (0%)	904 (2%)
06.10.2006 23:59 - 07.10.2006 23:59	2.833 (22%)	2.564 (20%)	1.270 (10%)	1.273 (10%)	1.098 (8%)	1.239 (10%)	865 (7%)	1.114 (9%)	447 (3%)	44 (0%)	268 (2%)
07.10.2006 23:59 - 08.10.2006 23:59	3.259 (21%)	2.897 (19%)	1.502 (10%)	1.522 (10%)	1.390 (9%)	1.486 (9%)	1.079 (7%)	1.389 (9%)	613 (4%)	224 (1%)	286 (2%)
08.10.2006 23:59 - 09.10.2006 23:59	16.316 (25%)	11.809 (18%)	6.324 (10%)	6.310 (10%)	5.843 (9%)	5.319 (8%)	5.391 (8%)	4.608 (7%)	1.600 (2%)	180 (0%)	1.134 (2%)
09.10.2006 23:59 - 10.10.2006 23:59	17.297 (25%)	12.390 (18%)	6.821 (10%)	6.813 (10%)	6.272 (9%)	5.818 (8%)	5.690 (8%)	5.224 (7%)	1.949 (3%)	201 (0%)	1.180 (2%)
10.10.2006 23:59 - 11.10.2006 23:59	16.393 (22%)	14.072 (19%)	7.467 (10%)	7.446 (10%)	6.824 (9%)	6.175 (8%)	6.351 (9%)	5.560 (8%)	1.724 (2%)	192 (0%)	1.080 (1%)
11.10.2006 23:59 - 12.10.2006 23:59	15.279 (25%)	11.495 (19%)	5.971 (10%)	5.937 (10%)	5.574 (9%)	4.967 (8%)	5.064 (8%)	4.331 (7%)	1.633 (3%)	165 (0%)	978 (2%)
12.10.2006 23:59 - 13.10.2006 23:59	13.939 (24%)	10.893 (19%)	5.705 (10%)	5.688 (10%)	5.328 (9%)	4.587 (8%)	4.939 (9%)	4.033 (7%)	1.379 (2%)	136 (0%)	864 (2%)
13.10.2006 23:59 - 14.10.2006 23:59	2.610 (20%)	2.611 (20%)	1.266 (10%)	1.226 (10%)	1.076 (8%)	1.235 (10%)	896 (7%)	1.162 (9%)	369 (3%)	84 (1%)	291 (2%)
14.10.2006 23:59 - 15.10.2006 23:59	3.083 (21%)	2.908 (20%)	1.347 (9%)	1.354 (9%)	1.224 (8%)	1.379 (10%)	1.140 (8%)	1.200 (8%)	459 (3%)	71 (0%)	259 (2%)
15.10.2006 23:59 - 16.10.2006 23:59	15.722 (26%)	11.039 (19%)	5.752 (10%)	5.727 (10%)	5.376 (9%)	4.511 (8%)	4.940 (8%)	3.937 (7%)	1.527 (3%)	156 (0%)	981 (2%)
16.10.2006 23:59 - 17.10.2006 23:59	15.646 (24%)	12.205 (18%)	6.676 (10%)	6.639 (10%)	6.281 (9%)	5.545 (8%)	5.688 (9%)	4.895 (7%)	1.636 (2%)	168 (0%)	1.116 (2%)
17.10.2006 23:59 - 18.10.2006 23:59	15.851 (23%)	13.054 (19%)	6.915 (10%)	6.916 (10%)	6.454 (9%)	5.726 (8%)	6.045 (9%)	5.090 (7%)	1.608 (2%)	199 (0%)	1.042 (2%)

Internet Analysis System (IAS)

→ Example for a fragment form the Reports



Top10

■ 43,8%-262166 - TCP (Destination port 22)=2.119.813
■ 38,9%-262169 - TCP (Destination port 25)=1.878.454
■ 15,6%-262224 - TCP (Destination port 80)=753.103
■ 1,0%-262323 - TCP (Destination port 179...=48.687
■ 0,6%-262587 - TCP (Destination port 443...=29.139
■ 0,1%-262197 - TCP (Destination port 53)=2.669
■ 0,0%-262257 - TCP (Destination port 113...=1.089
■ 0,0%-262165 - TCP (Destination port 21)=1.037
■ 0,0%-262167 - TCP (Destination port 23)=370
■ 0,0%-263168 - TCP (Destination port 102...=311
■ 0,0%-0 - Rest (Zusammenfassung restlich...=144

Top10

■ 51,6%-262169 - TCP (Destination port 25)=1.967.142
■ 32,9%-262224 - TCP (Destination port 80)=1.252.551
■ 13,1%-262166 - TCP (Destination port 22)=498.836
■ 1,4%-262323 - TCP (Destination port 179...=54.616
■ 0,9%-262587 - TCP (Destination port 443...=33.356
■ 0,0%-262257 - TCP (Destination port 113...=1.433
■ 0,0%-262197 - TCP (Destination port 53)=393
■ 0,0%-263168 - TCP (Destination port 102...=369
■ 0,0%-263137 - TCP (Destination port 993...=113
■ 0,0%-0 - Rest (Zusammenfassung restlich...=102
■ 0,0%-262165 - TCP (Destination port 21)=50

TCP Wellknown Zielports Inbound 02.05.08 23:59-03.05.08 23:59
Top10

■ 70,9%-262169 - TCP (Destination port 25)=2.038.027
■ 13,6%-262224 - TCP (Destination port 80)=390.596
■ 12,9%-262166 - TCP (Destination port 22)=371.501
■ 1,6%-262323 - TCP (Destination port 179...=45.478
■ 0,9%-262587 - TCP (Destination port 443...=24.682

TCP Wellknown Zielports Inbound 03.05.08 23:59-04.05.08 23:59
Top10

■ 60,8%-262169 - TCP (Destination port 25)=1.734.556
■ 33,3%-262224 - TCP (Destination port 80)=949.825
■ 3,2%-262166 - TCP (Destination port 22)=90.687
■ 1,5%-262323 - TCP (Destination port 179...=42.568
■ 1,2%-262587 - TCP (Destination port 443...=35.161

Internet Analysis System (IAS)
→ Status target 1

- At the moment we can analyze more than 870.000 different parameters
- We have collected a lot of data in your knowledge base, which helps us to define what we consider the normal state.
- The statistics help us, to understand the actual traffic on a data link and the also shows us, where theory deflects from reality. (distribution of SYN and SYN/ACK flags)
- With the help of the reports, we can receive aggregations with the most important results on a regular basis
  - Gives a great overview
  - These are very good information, to understand the normal behavior of an environment
  - The communication behavior stays under monitoring
  - Trends can be recognized at this stage
  - Abnormal behavior, which were left out of perspective during the manual analysis, can be detected with the help of these summaries

Internet Analysis System (IAS)
→ Further proceeding

- **Validating of the communication parameters**
  - Which are the once that are really used?
  - Which once are redundant?
  - How can we further reduce the amount of the collected data, for instance by using aggregation?

- **Identifying of new communication parameters**
  - Which protocols will gain in importance?
  - Which data is necessary to give a complete description of the Internet?

- **Working with / Analyzing of the knowledge base**
  - Use of data mining to find correlations and to better understand what we are dealing with

- **Find more partner**
  - Have more probes running at different pales

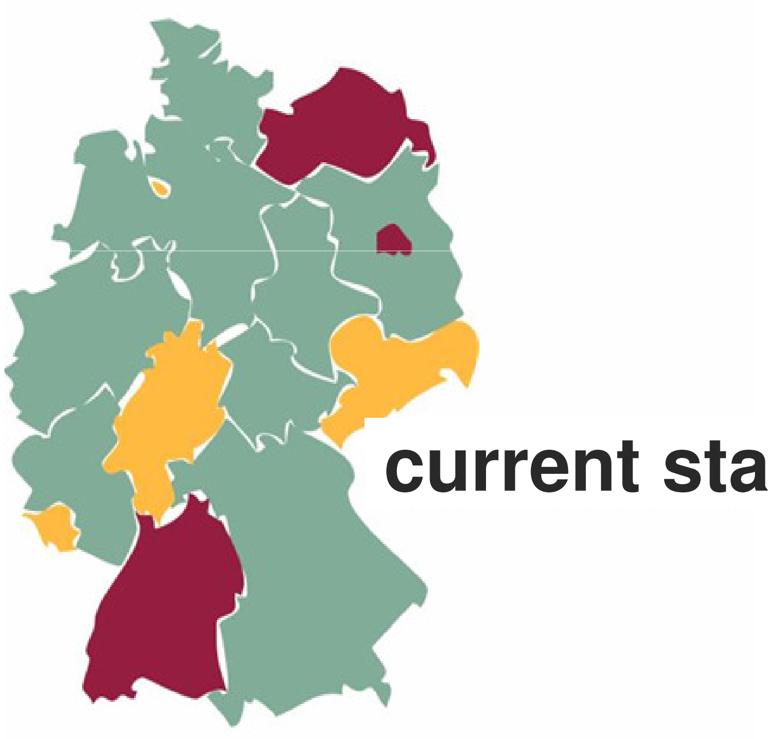
Content

- Aim and outcomes of this lecture
- Idea of the Internet Analysis System
- Knowledge Base
- **Outline of the Current State**
- Detection of Attacks and Deflection
- Forecast of Patterns and Attacks
- Summary

Internet Analysis System (IAS)
→ Defined targets

Target 2

- **Outline of the current state of the internet.**



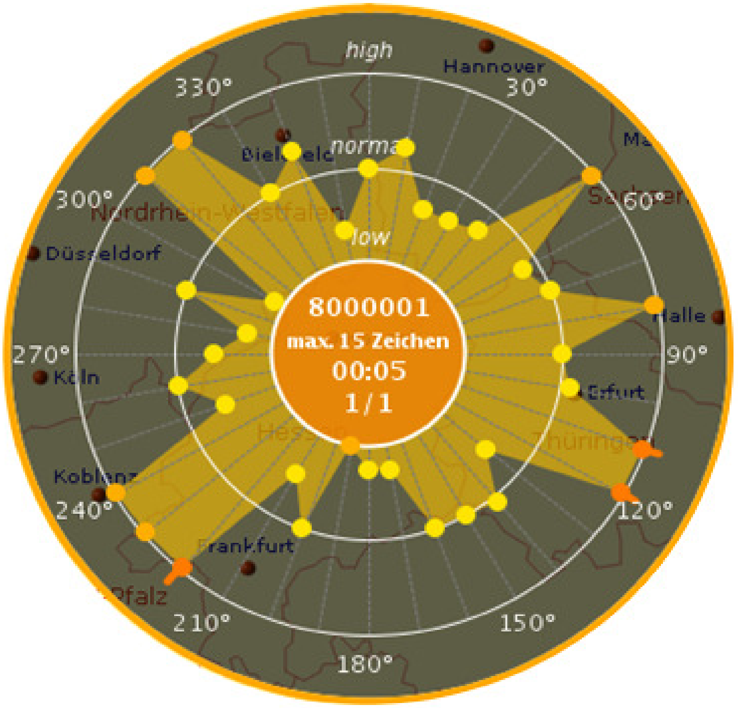
In this field an important function is the clear visual representation of the state of the internet, like traffic jam maps.

Current state - IAS
→ Target 2

- We need designs that help us determine the current state.

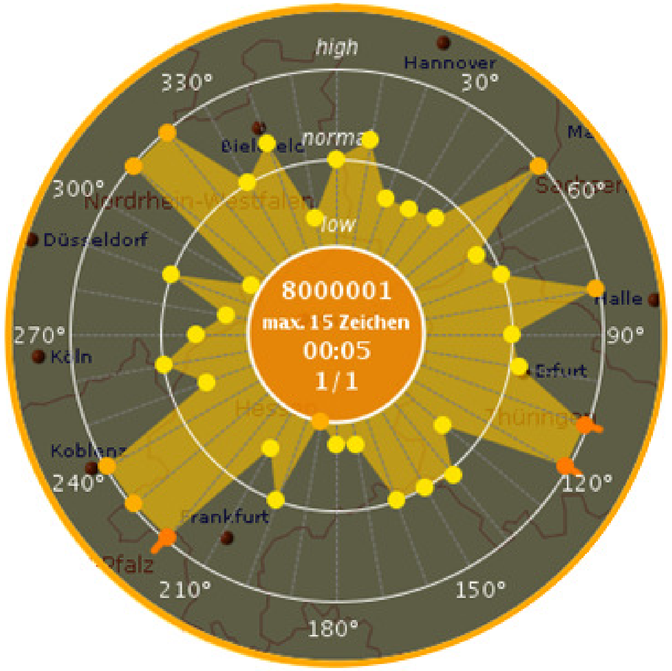
  Challenge: display enormous amount of data in an intuitive manner

- One example for a visualization tool we use to gain on experience is **VisiX.**
- VisiX: **V**isual **I**nternet **S**ensor **I**nformation
- Pre selected, **important Parameters**
- **Continuous updating**
- Alignment on the basis of fixed reference values
- selectable, **colored coding**



Current state - IAS → Target 2

- Visualization of the data of multiple probes at the same time using multiple diagrams in one visualization
- This allows to detect coherences between different probes
- For example:
  - probe X: extremely high level of http traffic
  - probe Y: extremely high as well
  - → external event like a Windows Update) or a possible attack



- VisiX allows the user to get to know the communication behavior of a network
- Continuous monitoring in the case of an alert
- Helps the user to initiate further measures
- Procedure: (i) Alert → (ii) VisiX → (iii) EagleX, ...

Current state - IAS
→ Special parameter: Overview

0 ° total packets (probe)

10° IP Counter SRC SYN

20° Σ TCP ports & Σ UDP ports

30° Σ TCP ports / TCP flag S

40° IP Protocols \ {ICMP, TCP, UDP}

50° ICMP types

60° UDP wellknown

70° UDP registered

80° UDP dynamic

90° TCP wellknown

100° TCP registered

110° TCP dynamic

120° TCP flag S

130° TCP flag R

140° TCP flag AF

150° TCP flag SA

160° DNS requests

170° DNS respones

180° DNS flag RA

190° DNS flag RD

200° HTTP traffic

210° HTTP GET

220° HTTP POST

230° HTTP HEAD

240° HTTP other

250° HTTP error

260° SMTP traffic

270° SMTP MAIL

280° email with attachment

290° email without attachment

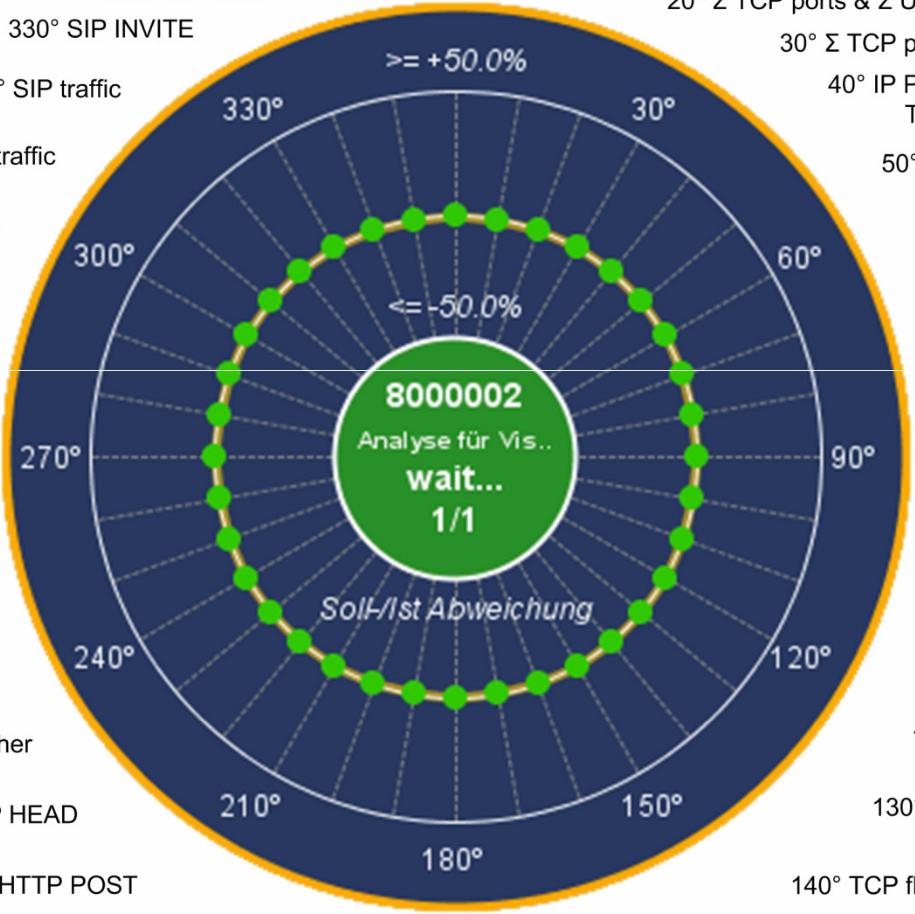
300° SMTP error

310° IMAP / POP traffic

320° SIP traffic

330° SIP INVITE

340° SIP error



Current state - IAS
→ Special communication parameter (1/6)

- 0 META: total Packets (probe)
- 10 IP Counter SRC + SynFlag
- 20 Total TCP and UDP Ports
- 30 Total Packet / (TCP Ports + UDP Ports) 7 / {na}
- 40 IP Protocols
  - ALL (without 1 , 6, 17)
- 50 ICMP Types
  - ICMP (Type 0 echo reply)
  - ICMP (Type 3 destination unreachable)
  - ICMP (Type 4 source quench)
  - ICMP (Type 5 redirect)
  - ICMP (Type 6 alternate host address)
  - ICMP (Type 8 echo request)
  - ICMP (Type 9 router advertisment)
  - ICMP (Type 10 router solicitation)
  - ICMP (Type 11 time exceeded)
  - ICMP (Type 12 parameter problem)

Current state - IAS
→ Special communication parameter (2/6)

- 60 UDP wellknown Ports
  - UDP SRC Port wellknown
  - UDP DST Port wellknown
- 70 UDP registered Ports
  - UDP SRC Port registered
  - UDP DST Port registered
- 80 UDP dynamic Ports
  - UDP SRC Port dynamic
  - UDP DST Port dynamic
- 90 TCP wellknown Ports
  - TCP SRC Port wellknown
  - TCP DST Port wellknown
- 100 TCP registered Ports
  - TCP SRC Port registered
  - TCP DST Port registered
- 110 TCP dynamic Ports
  - TCP SRC Port dynamic
  - TCP DST Port dynamic

Current state - IAS
→ Special communication parameter (3/6)

- 120 TCP Flag S
- 130 TCP Flag R
- 140 TCP Flag AF
- 150 TCP Flag SA
- 160 DNS Requests
- 170 DNS Responses
- 180 DNS Flag RA
- 190 DNS Flag RD
- 200 HTTP / HTTPS Traffic
  - TCP (Source port 80)
  - TCP (Source port 443)
  - TCP (Destination port 80)
  - TCP (Destination port 443)

Current state - IAS
→ Special communication parameter (4/6)

- 210 HTTP Request method GET
- 220 HTTP Request method POST
- 230 HTTP Request method HEAD
- 240 HTTP Request method OTHER
  - HTTP (Request Method PUT)
  - HTTP (Request Method DELETE)
  - HTTP (Request Method TRACE)
  - HTTP (Request Method OPTIONS)
  - HTTP (Request Method CONNECT)
- 250 HTTP Server response codes
  - 4xx
  - 5xx
- 260 SMTP / SMTPS Traffic
  - TCP (Source port 25)
  - TCP (Source port 465)
  - TCP (Source port 587)
  - TCP (Destination port 25)
  - TCP (Destination port 465)
  - TCP (Destination port 587)

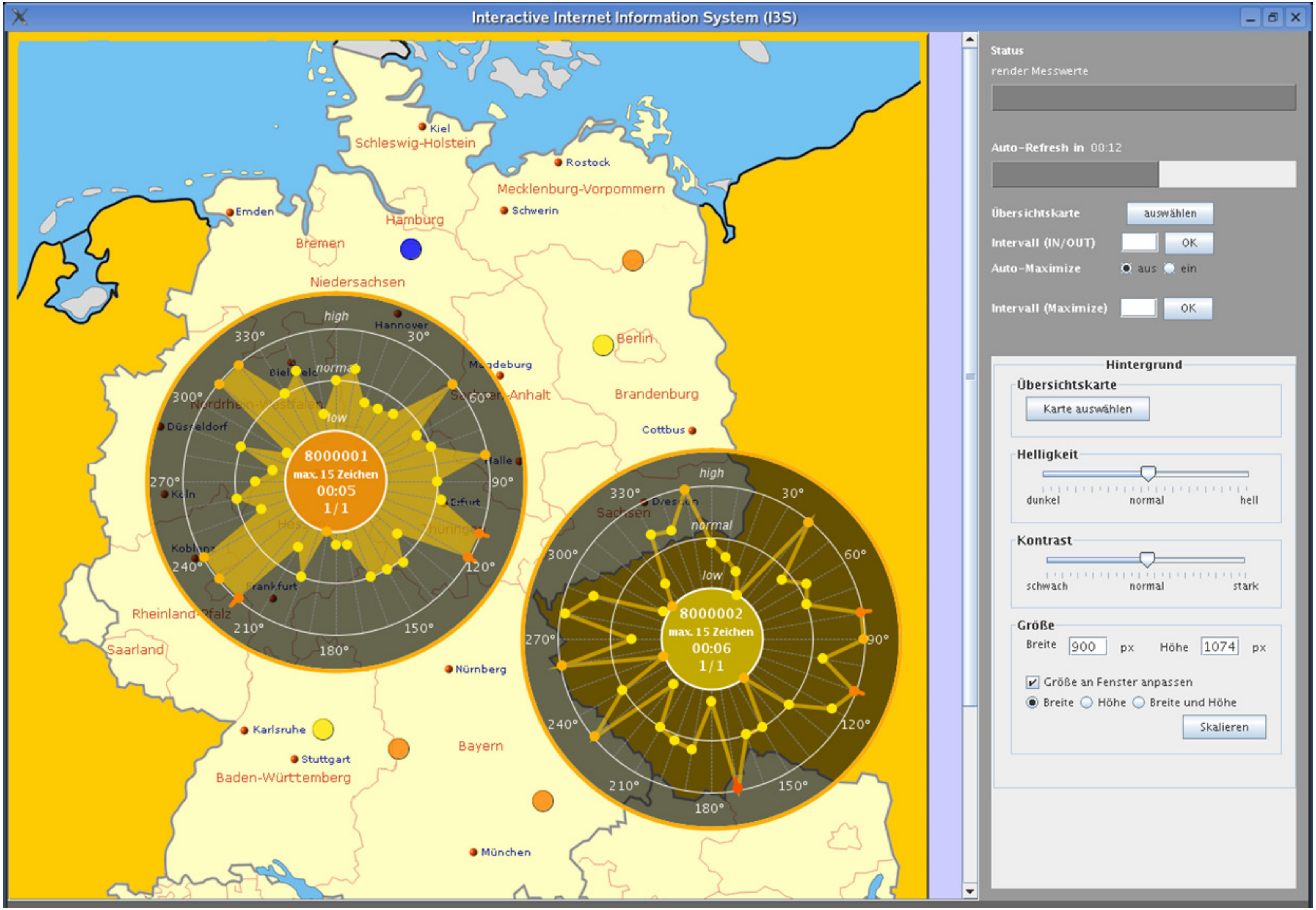
Current state - IAS
→ Special communication parameter (5/6)

- 270 SMTP MAIL
- 280 SMTP E-Mail with attachment
  - Client header multipart/mixed
- 290 SMTP E-Mail without attachment
  - client header text/plain
  - client header text/html
  - client header multipart/alternative
  - client header multipart/report
- 300 SMTP Server response codes
  - 4xx
  - 5xx
- 310 POP / POPS / IMAP / IMAPS
  - TCP (Source port 110)
  - TCP (Source port 143)
  - TCP (Source port 993)
  - TCP (Source port 995)
  - TCP (Destination port 110)
  - TCP (Destination port 143)
  - TCP (Destination port 993)
  - TCP (Destination port 995)

Current state - IAS
→ Special communication parameter (6/6)

- 320 SIP Invite
- 330 SIP Traffic
  - TCP (Source port 5060)
  - TCP (Destination port 5060)
  - UDP (Source port 5060)
  - UDP (Destination port 5060)
- 340 SIP error codes
  - 4xx
  - 5xx
  - 6xx

IAS: Current State of Development
→ Continuous situation awareness



Interactive Internet Information System (I3S)

Status

render Messwerte

Auto-Refresh in 00:12

Übersichtskarte auswählen

Intervall (IN/OUT) OK

Auto-Maximize ● aus ○ ein

Intervall (Maximize) OK

Hintergrund

Übersichtskarte

Karte auswählen

Helligkeit: dunkel – normal – hell

Kontrast: schwach – normal – stark

Größe

Breite 900 px Höhe 1074 px

☑ Größe an Fenster anpassen

● Breite ○ Höhe ○ Breite und Höhe

Skalieren

Detection of Anomalies
→ Principle idea (1/4)

- **The detection of anomalies** can be used to observe behavior deflecting from the normal state
- In contrary to **misuse detection**, using **patterns** to find behavior that is not permitted
  - Which does not allow the detection of threats that are so far **unknown**
- **Problem**: What can be defined as “normal”?
  - Description of the “normal state” is difficult
  - Requires adequate representation
- **Important:** the **detection of anomalies** does not generate alerts concerning attacks, but informs of **abnormal behavior**
  - After a successful alert further **analysis** is necessary
- The detection of anomalies allows so far **unknown** threats to be identified
- By doing this, the risk of false alarm also rises

Detection of Anomalies
→ Principle idea (2/4)

- Basis of the method to detect anomalies is the development of a **model** which describes the **normal state**
- **Problem**
  - What can be defined as “normal”?
    - Are regular port scans normal or should they be reported as an anomaly?
    - Isn’t an anomaly more likely to be a successful exploit after performing a port scan?
    - **Precise definition** of the normal state is required
  - Description of the **“normal state”** is difficult
    - Useful data needs to be collected, which help to describe the current state
    - Adequate **representation** of the data is necessary

Detection of Anomalies
→ Principle idea (3/4)

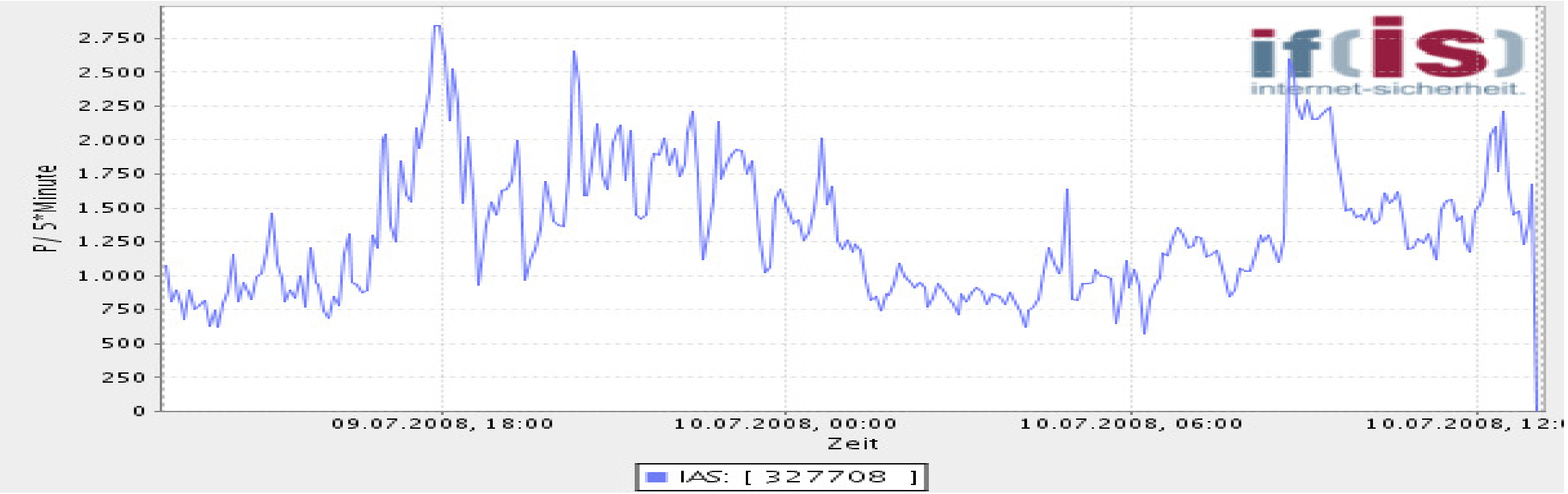
- **Problem**
  - The normal state changes over time
    - These changes need to be considered
    - In principle comparable to the misuse detection, where new patterns reflecting attacks need to be considered
    - The process of the detection of anomalies needs update itself continuously
  - The normal state is different at every location
    - Methods need to learn a different normal state at each location they are used

Detection of Anomalies
→ Principle idea (4/4)

- **A lot of** different methods from various disciplines
  - Signal processing, pattern detection, artificial intelligence, statistics
- Two different approaches for the detection of anomalies shall now be introduced exemplarily
  - Modeling of the descriptors as time series
  - Modeling of the data with the help of cluster models
- Further approaches are possible and can be looked up in scientific publications
  - *Adaptive thresholding for proactive network problem detection – Ji, Thottan - 1998*
  - *Anomalous payload-based network intrusion detection – Wang, Stolfo – 2004*
  - *Information-theoretic measures for anomaly detection – Lee, Xiang - 2001*

Time series approach

- IAS sensor records about 870.000 parameters
  - In reality there is only a part of the communication stream included
  - Parameters are referred to as descriptors
- **Idea:** consider descriptors independently
- Over time the descriptors make up a time series
  - A set of values which are displayed over time



Time series approach

- Time series can be described by mathematical models
  - Linear models
  - Non linear models

- One possible approach is given with ARMA models, which will be introduced in the following slides
  - Also possible: other time series models
  - ARMA models are already in use by [8] for the modeling of network traffic

Time series approach

- Linear models of the form

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=1}^{q} b_i \varepsilon_{t-j} + \varepsilon_t$$

with

Y_t := value of the times series at the time t

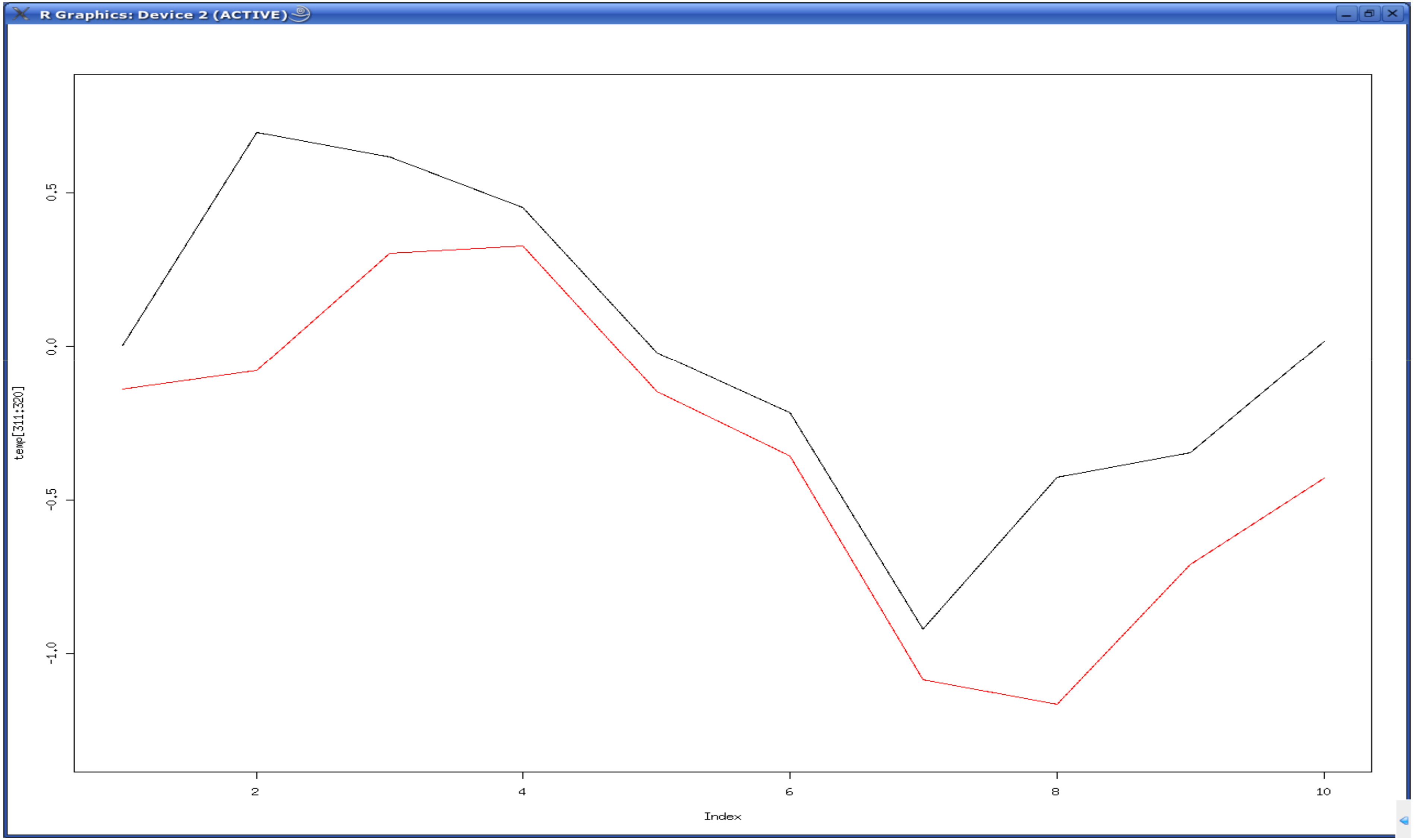
a_i, b_i := coefficients, which are weighting past values

ε_t := random influences

p, q := order of the model

- For the set up of the model the following parameters must be determined
  - Order of the model, that is to say how many past values have an influence on the current value
  - The coefficients (a_i, b_i)
  - Methods for the determination is not further described here, but can bee looked up in [9]
- The figure on the next slide shows the run of the descriptors (black) and the estimated run (red) (10 data points in the future)

Time series approach



Time series approach

- Based on this model the validity of the next measured data point in dependency to the previously measured data point can be determined
- For this the actual measured data point is compared to the estimated data point of the model
  - Determination of the random influences, which in principle is equivalent to the deflection from the model
- The variance of the estimates is included, which means the deflection of the data points estimated by the model to the actually measured data points
  - This is determined in relation to the random influence ε_t of the model $\varepsilon_t > n\sigma_\varepsilon$
- Put in the terms of mathematics: a deflection is proclaimed when with
  - σ_ε := standard deviation of the random influence
  - n := factor selectable by the user

Time series approach

- Solution to the mentioned problem from section 1:
  - What is normal?
    - Defined by the model of the times series
    - The model can be improved with an iterative approach
      1. Determine model
      2. Remove outliers
      3. Restart from point 1 until model is good enough
  - Description of the normal
    - With the time series model
  - Adjustment to changes
    - Re-determine model on a regular basis
  - Dependencies of the normal state depending on local influences
    - Determination of the time series model at each location

Time series approach

- **Drawbacks**
  - Time series must be determined for each descriptor
    - Very many models (about 870.000 per sensor)
    - Search for models is very complex
    - Update of the models is complex as well
  - It is about linear models
    - Possibly not sufficient to describe all the data which has been collected
  - Relations between descriptors are not considered

Cluster approach

- For the previously described approach the descriptors had to be processed one by one
- The relations between descriptors have not been considered
  - Example: relations between SYN, SYN/ACK, FIN/ACK flags
  - Information is lost
- Can be avoided, by
  - the creation of ratios
    - E. g. ratio between SYN to SYN/ACK modeled as a times series
  - using other times series approaches
  - Using a different form for the modeling
    - Idea: presentation as vectors

Cluster approach

- So far descriptors have been analyzed one by one
- This approach analyzes the certain descriptors as a combination
- The descriptor values are concentrated as vectors

  $$\vec{v} = (d_1, d_2, \ldots, d_n)$$

- For example for the descriptors SYN, SYN/ACK,FIN/ACK (figure on the next slide)
  - From the values of the descriptors, vectors of the following form are build

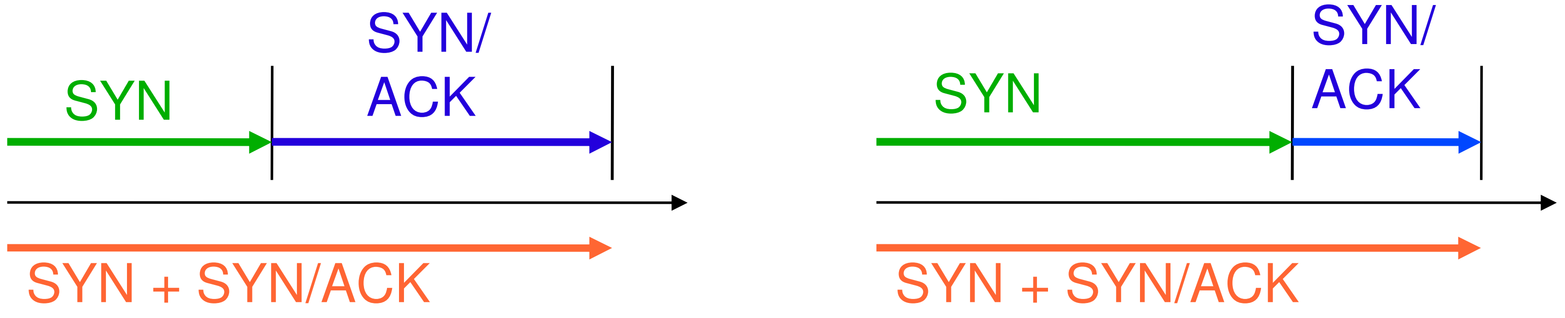
    $$\vec{v} = (SYN, SYN / ACK, FIN / ACK)$$

  - These vectors are scaled to the length of one
  - Given by specification of the protocol the ratio between SYN, SYN/ACK and FIN/ACK (1/1 respectively ½) the scaled vectors should comply with about the following vectors $\vec{v} = (\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}})$

Input for neural networks

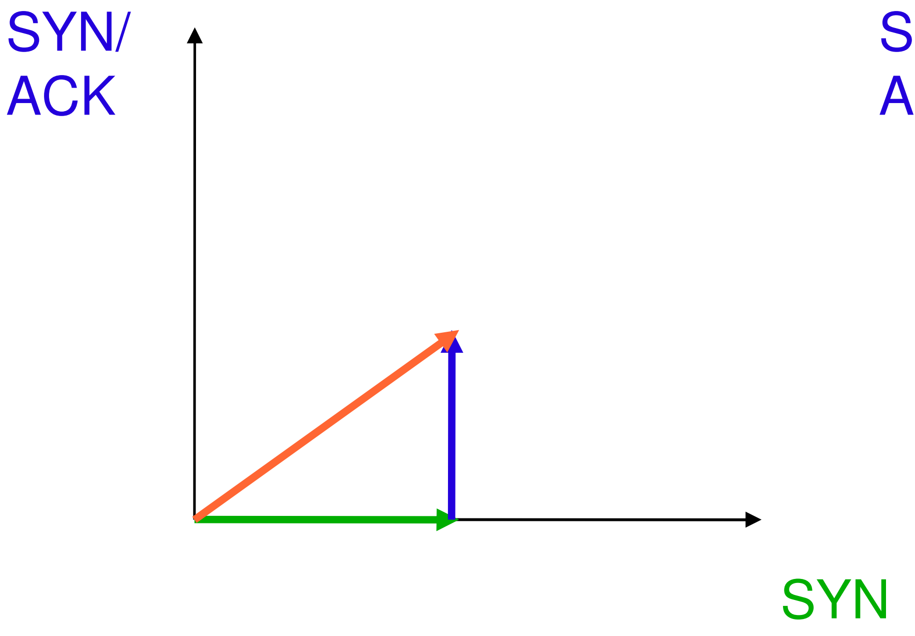
- „36“ parameters consisting of aggregation of different parameters
  - Pseudo parameter “10000009” consists of about 400 real parameters
  - In total these monitored „36“ parameter consists of about 1200 real parameters of the IAS which are monitored at the same time.

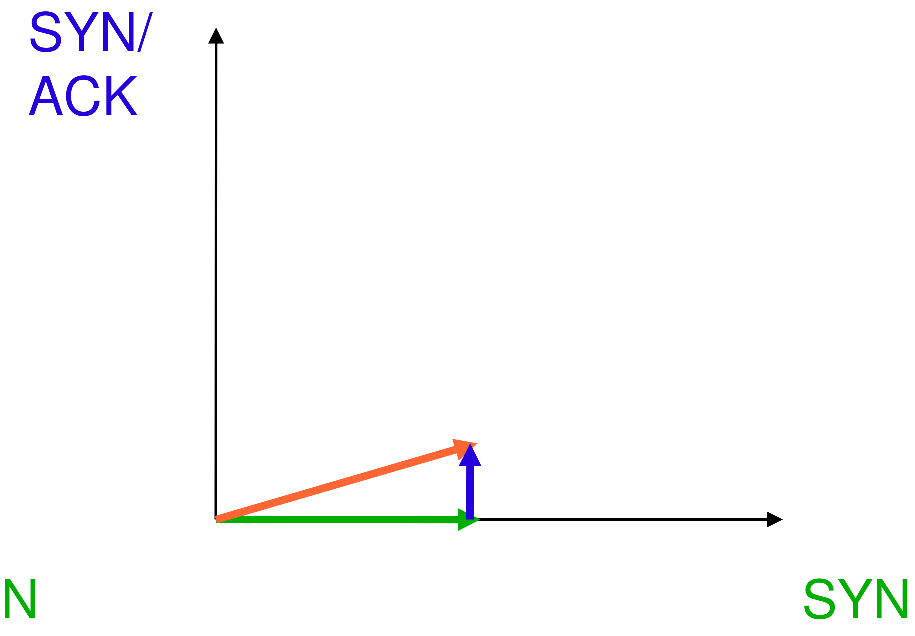
- Mapping of the parameters on the „36“ visualized parameters
  - Sum
    - Information about the relation of the measured values are lost
    - Example:



Input for neural networks

- Mapping of the parameters on the „36“ visualized parameters
  - Mapping as a vector
    - Real parameters form the components of a vector
    - Relation between parameters are contained
    - Example:

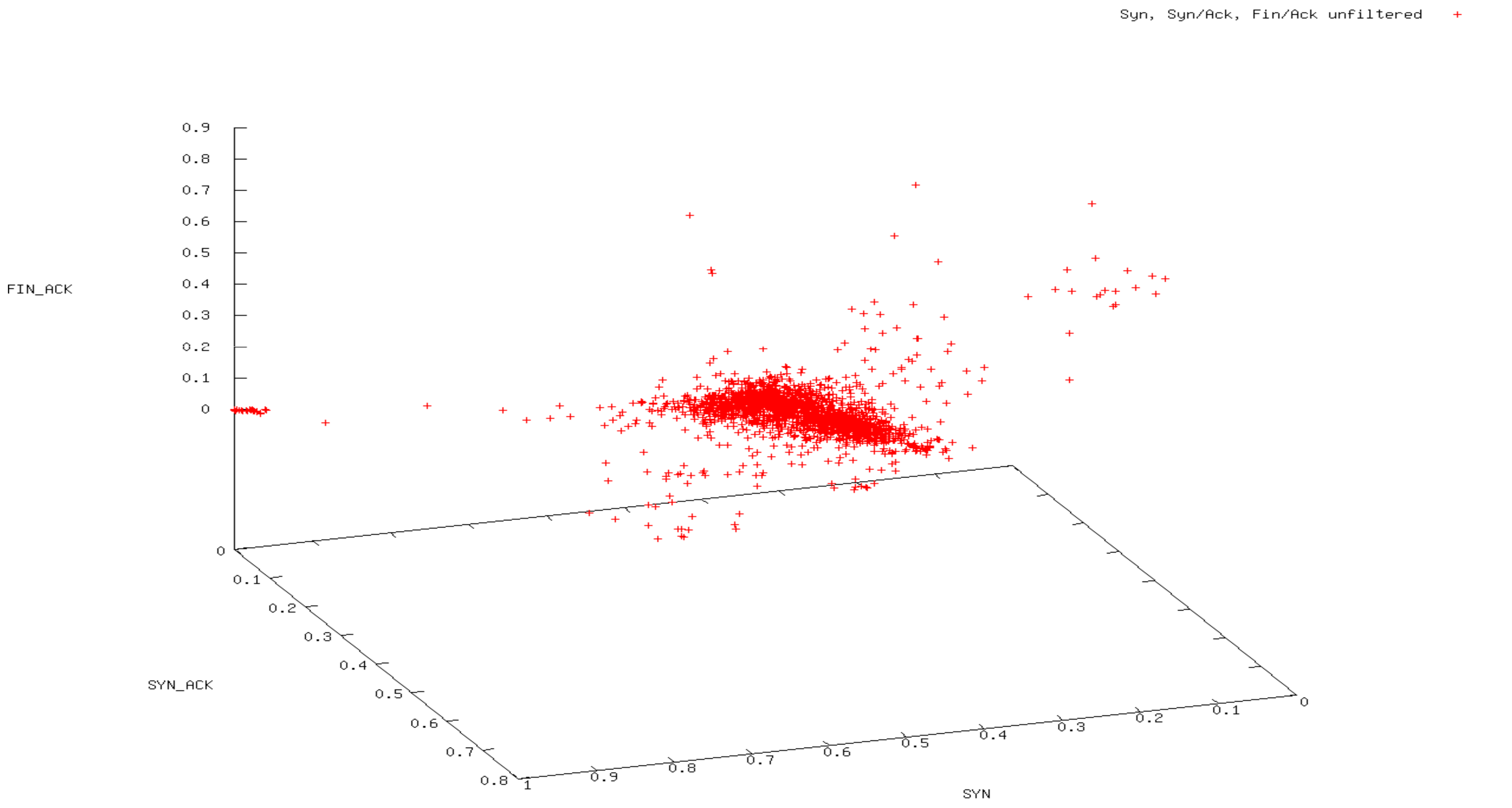




Input for neural networks

- Parameter for themselves are insufficient information
  - In relation to the time (current interval)
  - Information about the last values of the past
- Vectors are normalized prior to processing by the neural networks, to avoid dependencies of absolute values
  - Length of vectors are normalized to the value one

Self Organizing Map
→ Example: Syn, Syn/Ack, Fin/Ack



Cluster approach

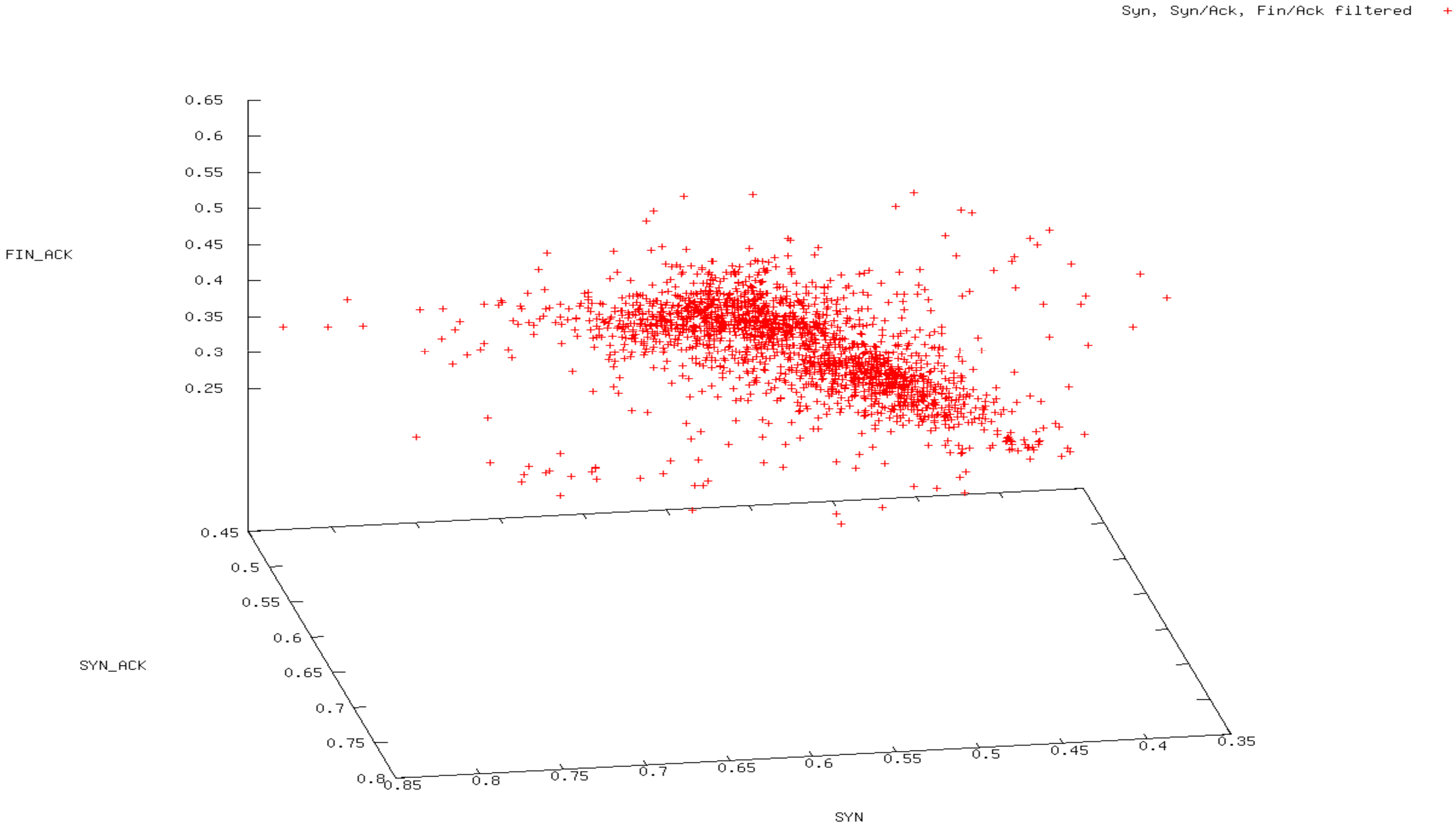
- Previous figure shows the distribution in space of the vectors. The normal behavior has not been filtered.

- One can notice the accumulation of vectors in the area of ideal behavior

- In addition, you can see a dispersion around the ideal area
  - This can be explained with the fact, that the reality is not ideal

- You may also notice vectors that deflect very much from the ideal area
  - What is going on with these vectors?
  - When reviewing the data one should filter the data regarding its level of ideal behavior

Self Organizing Map
→ Example: Syn, Syn/Ack, Fin/Ack filtered



Cluster approach

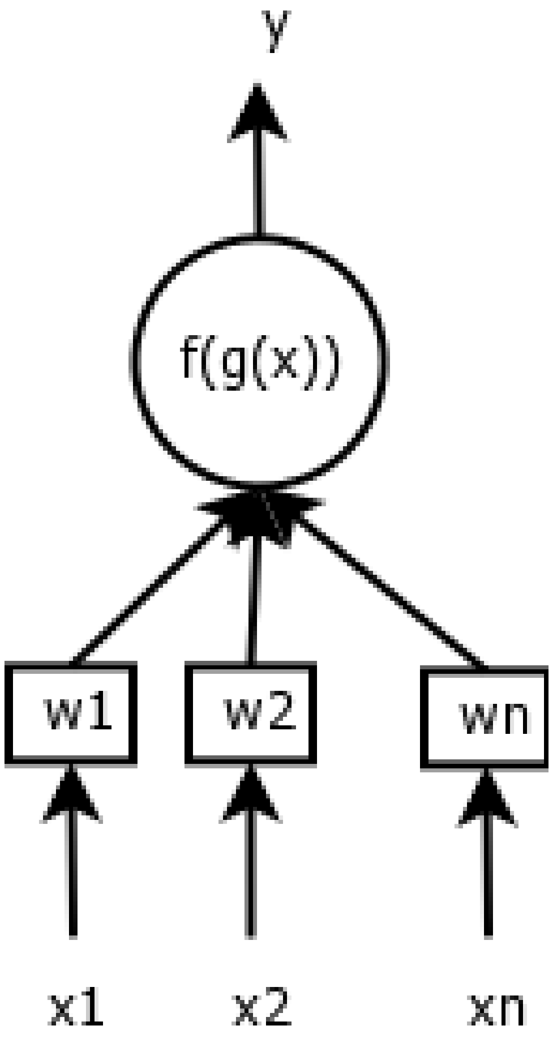
- Figure shows the filtered results (be aware: using a different scale!)
- The vectors which were having an extreme deflection were dropped in this presentation
- Leaving all data points reflecting normal behavior and aggregating in this special area in space.
- This example illustrates:
  - By means of vectors relations between different parameters can be considered
  - Data reflecting normal behavior aggregates in certain areas in space
  - Abnormal data show up in different areas in space and can be identified this way
- How can this be implemented with algorithms?

Cluster approach

- Observation: data points create a cluster in space
- This behavior can be learned
  - E. g. by the cluster approach
- Based on the learned clusters decisions are made concerning whether a data point is normal or abnormal
  - Can be done using the distance to the learned cluster
- Example: Self Organizing Map
  - Neuronal cluster method
  - This will now be introduced
  - Belongs to the area of unmonitored learning methods
  - Other cluster methods are also possible

Self Organizing Maps
→ Neurons (1/2)

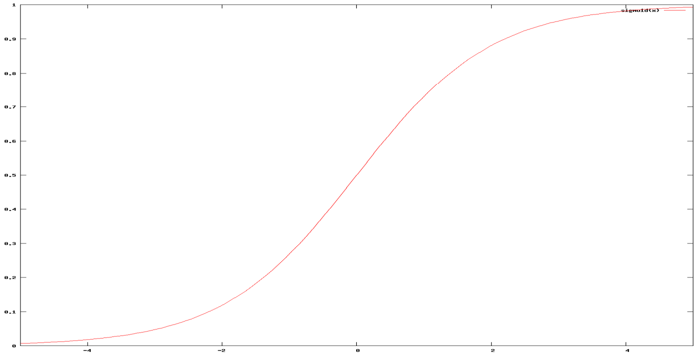
Neuron



- **Output**

$$y = f(g(\vec{x}))$$

- **Activation function**



$$f(x) = \frac{1}{1+e^{-x}}$$

- **Propagation function**

$$g(\vec{x}) = \sum_{i=1}^{n} w_i x_i$$

Self Organizing Maps
→ Neurons (2/2)

- Neuron gets input values consisting of a vector
  - E. g. values of the sensors

- Weighted average sums are build over the input value (propagating function)
  - The function of a neuron is defined through the weighting
  - Neuron is a black box

- Activation of the neuron is determined by the use of an activation function
  - Output value indicates how intensively the neuron has been activated by the input
  - Restricted to the top and bottom (take a look at the previous figure)

Self Organizing Map
→ Networks

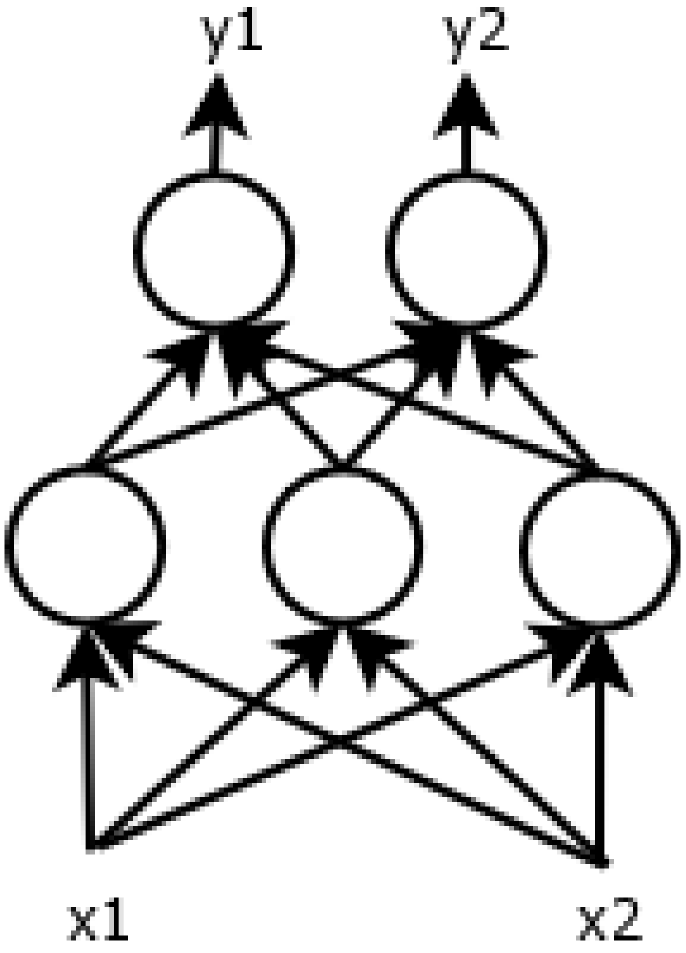
- The possibilities of expression of a neuron are limited

⇒ Interconnection of networks to lift limitations

- **Network** := linking of a number of neurons in

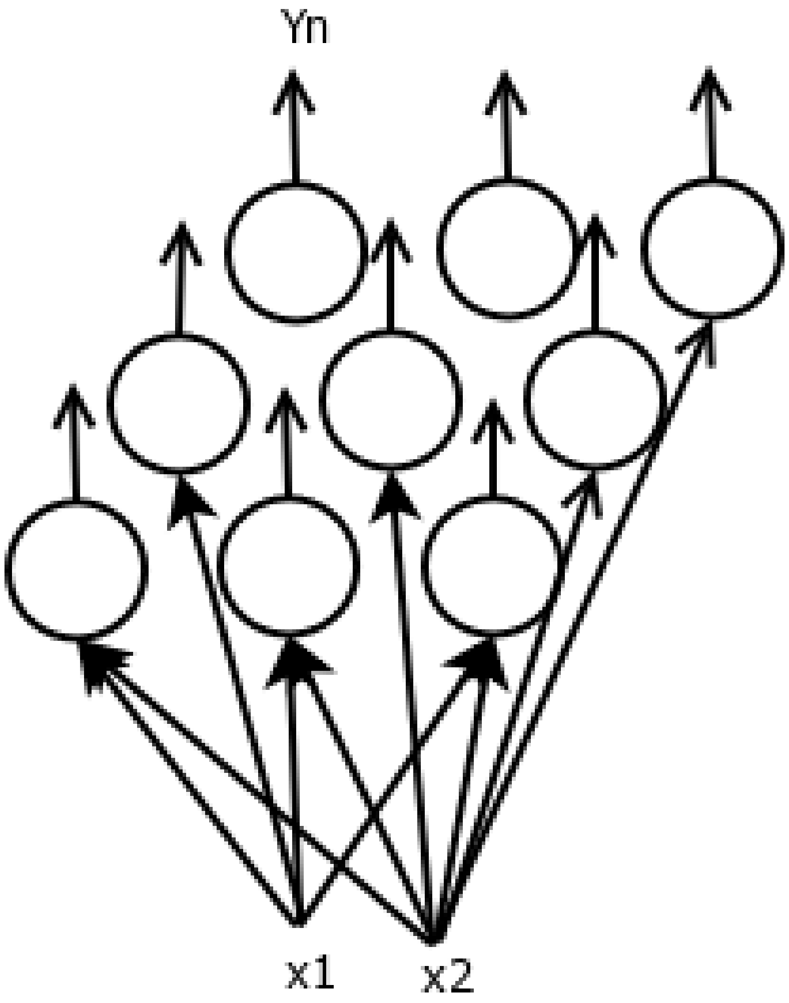
- Function of the network is determined by the weighting of the factors w

- Weighting is determined by learning
  - Presenting of examples
  - Based on the error determined by the distance to the expected value the weighting factor w is adjusted
  - If error is beneath a bound the learning process is finished



Self Organizing Map
→ Set-up

- Special characteristics of neuronal networks
- Network for the clustering of data
  → generates groups of the presented data
- Two dimensional arrangement of the neurons in one layer
- Input value is handled for each neuron
- Belongs to the class of unsupervised learning methods
- Abbreviation: SOM



Self Organizing Map
→ Learning method (1/3)

- Selection of a set of data for training

- Initialization of the weighting factor w

- Presentation of the pattern vector
for a set of data for training and estimation of the activation of neurons by the means of the Euclidean distance

$$y_k = \sum_{i=1}^{n} (x_i - w_{ik})^2 = (\vec{x} - \vec{w}_k)^T (\vec{x} - \vec{w}_k)$$

- Search for the minimal activation
within the neurons (neurons with the smallest distance to the current pattern vector)

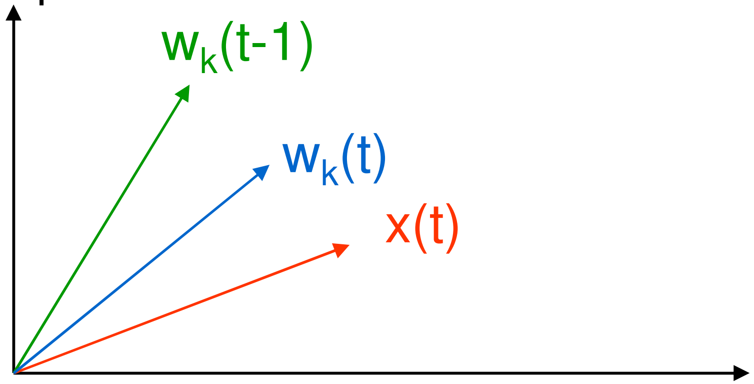
Self Organizing Map
→ Learning method (2/3)

- Adjusting of the weighting of the neurons with minimal activation by the means of

$$w_k(t) = (1 - \alpha(t))w_k(t-1) + \alpha(t)x(t)$$

$$\alpha(t) = \begin{cases} \alpha_0(1 - \frac{t}{t_1}) & 0 \le t \le t_1 \\ \alpha_2(1 - c_1 \frac{(t - t_1)}{(T - t_1)}) & t_1 < t \le T \end{cases}$$

  - Weighting factor w_k is pushed closer to the current vector for training in space

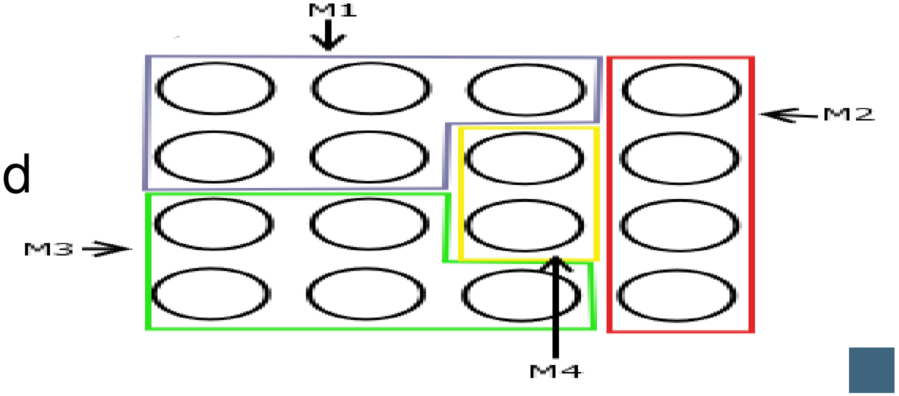


Self Organizing Map
→ Learning method (3/3)

- Adjusting is not just done for the neuron with the lowest activation, but also in the environment of the neuron based on

$$R(t) = \begin{cases} R_0 - (R_0 - R_1)\dfrac{t}{t_1} & 0 \le t \le t_1 \\ R_1 & t_1 < t \le K \end{cases}$$

- Resulting in a kind of data map
  - Neurons are grouped, each representing a cluster
  - When applying a vector those neurons belonging to the group of this vectors are especially activated
  - The figure illustrates the principle



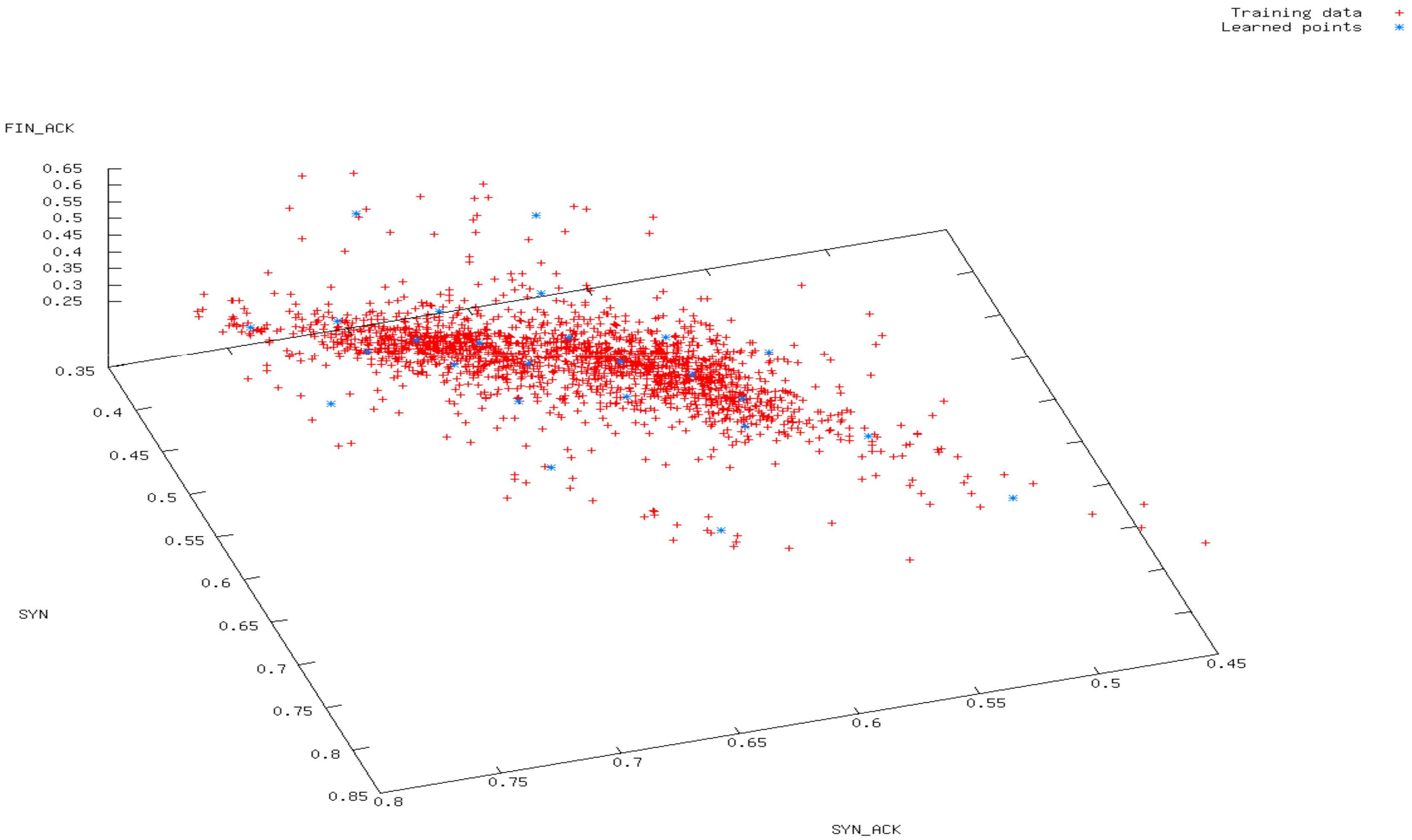
Self Organizing Map
→ Applying

- Initialization
  - Network is trained with data representing the normal state

  ⇒ Network learns the structure of the cluster of this data

- For example: the network is trained with the explained approach with the vectors, which are left over after the filtering, for SYN, SYN/ACK, FIN/ACK
  - In this example a network with 25 neurons has been used
  - Therefore, 25 representatives for the cluster have been learned, which are shall be referred to as c_j
  - The figure on the next slides illustrates the learned representatives (blue) in comparison to the trained data (red)

Self Organizing Map
→ learned representatives



Self Organizing Map
→ Use

- Comparison of the trained data to the learned vector

- Learned vectors (blue) are distributed in the area of the trained data (red)

  => network has learned coherences

- A classification of new vectors can be performed based on the learned vectors

- Therefore a vector is compared to a learned vector
  - Determination of the distance to all learned vectors
  - Determination of the distance to the cluster vector j by the means of the Euclidean distance

    $$d_j = \sqrt{\sum_{i=1}^{N} (x - c_{ji})^2}$$

  - In case max d_j > threshold, then generate alert

Self Organizing Map
→ Cluster approach

- **Solution to the mentioned problem from section 1:**
  - What is normal?
    - Defined by learned cluster
  - Description of the normal
    - With learned cluster vector
  - Adjustment to changes
    - Adaptation of the learned cluster on a regular basis (best with each new data point)
  - Dependencies of the normal state depending on local influences
    - Determination of the cluster model at each location

Self Organizing Map
→ Cluster approach

- **Drawback**
  - Training must be done with data reflecting the normal state
    - Identification is very complex
  - Adaptation is not yet included in the described approach
    - Must be completed
  - Determination of the threshold can be a real problem

Content

- Aim and outcomes of this lecture
- Idea of the Internet Analysis System
- Knowledge Base
- Outline of the Current State
- **Detection of Attacks and Deflection**
- Forecast of Patterns and Attacks
- Summary

Internet Analysis System (IAS)
→ Defined targets

Target 3

- **Detection of attacks and of deflections.**

With the knowledge of the current state and with the help of historical data, a warning will be issued if there are significant changes can be identified.
And this function helps us to reduce the damage in the Internet

Internet Analysis System (IAS)
→ Detection of attacks and of deflections

Attacks

- Definition of signatures, which we have determined by the use of the „**EagleX Analysis Client**“
- Cooperation with the University of Mannheim (Germany)
  - **CW-Sandbox**
    - **Exchange of patterns and “reload of URLs to the probe”**
    - **Determine communication profiles**

Deflections

- Core component of an Early Warning System
- We develop and use designs, which we can use to determine the deflection to the expected behavior

Anomaly
→ Overview

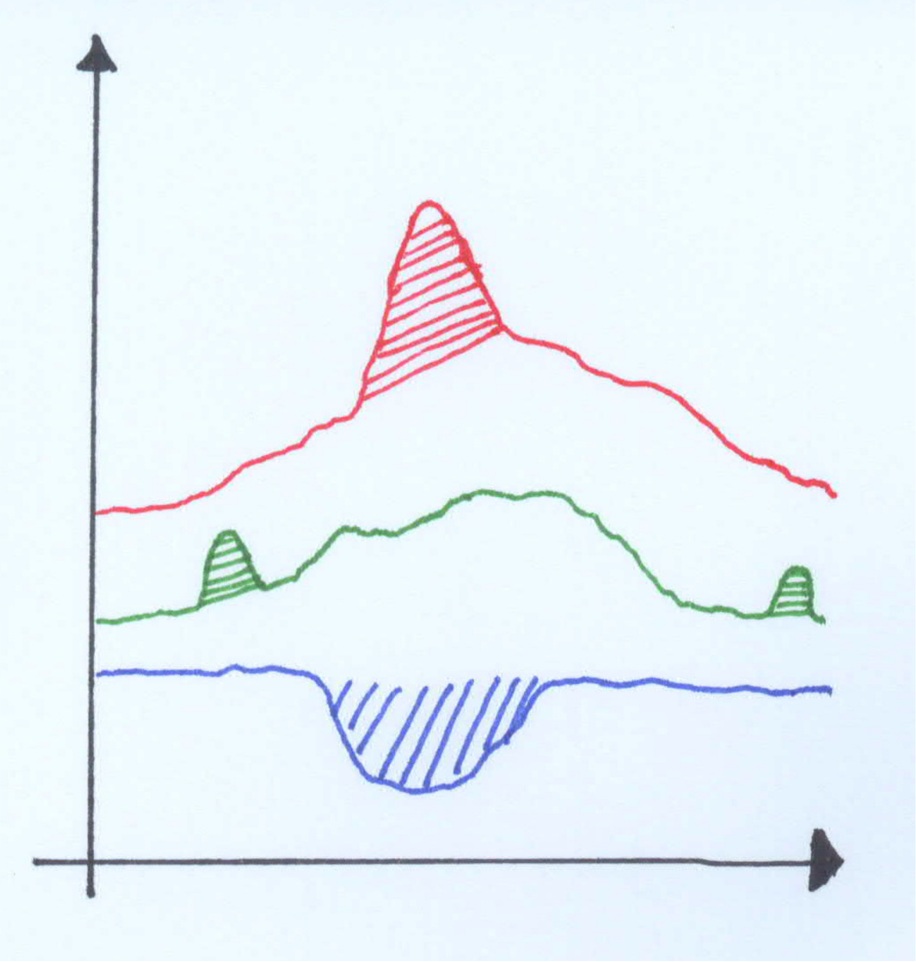
- Anomaly: Greek „*anomal*“ : „unsteady“, „irregular“
- In terms of the IAS: deflections in the measured network traffic
- Anomaly != attack
  attack implies an anomaly,
  but only an analysis can verify if an anomaly is an attack
- Two dimensional detection of anomaly
  - Based on location
  - Based on time

Anomaly
→ Detection of anomalies

- **Detection of anomalies based on location**
  - Different locations can not always be compared
    - Normal traffic at location A can be abnormal at location B.
    - Detection of anomalies is therefore always based on the location as well

- **Detection of anomalies based on time**
  - Point of measurement in a business environment: more data traffic during day time, a lot less during night time.
  - Point of measurement in a home environment: outside the office hours higher traffic load, power users have a high load 24h a day.

Anomaly
→ Detection of anomalies



- Requirement: location unchanged
- Capturing of deflections in parameters based on the time
- A data traffic that is not changing abruptly over time is defined as being normal.
- In relation to the time of the day smaller peaks can be normal too, because the common traffic is reduced as well.
- Data traffic is changing over time, an anomaly today can be normal tomorrow.

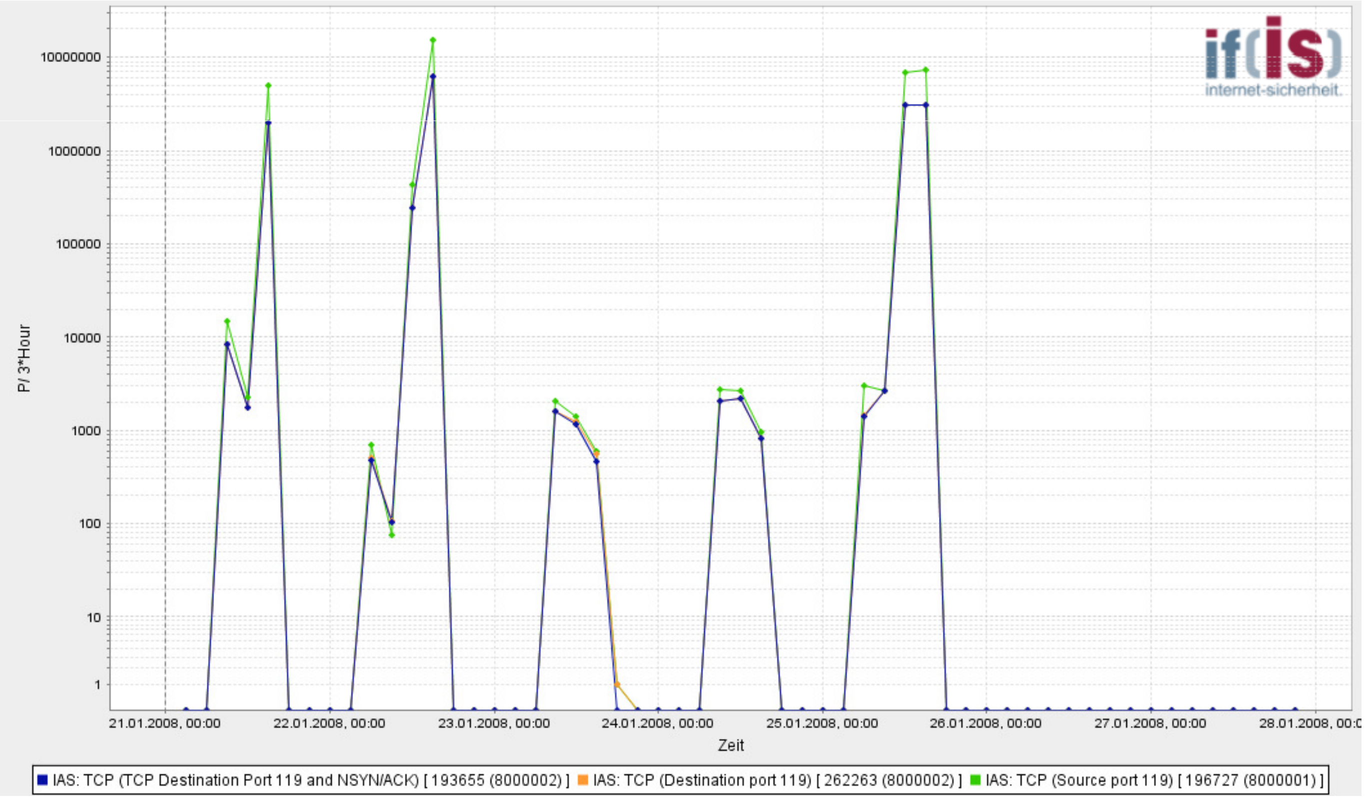
Anomaly
→ Detection of anomalies

- **One-Packet-Attacks**
  - Intentional and wanted manipulation of one single packet.
    - Country – Attack (Source IP = destination IP)
    - Ping of Death (oversized IP packet)
  - Detection is possible in principle
  - Early Warning impossible

- **Multi-Packet-Attacks**
  - Due to stateless design of the IAS no direct detection is possible
  - Detectable, if attacks don’t vanish in traffic noise
  - Early Warning possible

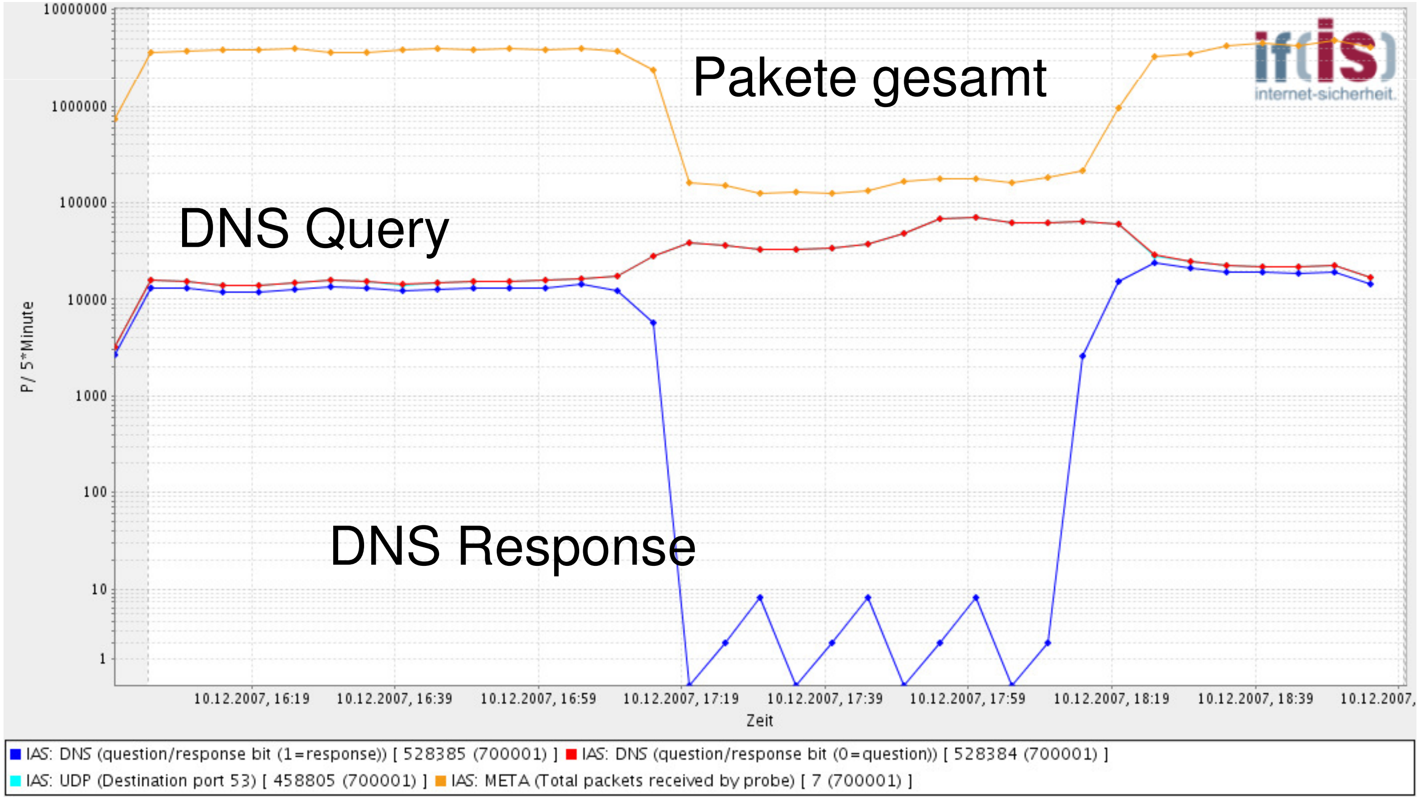
Anomaly
→ Different kinds of anomalies

- **Misusage of services**
  - Use of resources in an not intended manner like the NNTP in the department of computer science
  - Incoming data traffic
  - Traffic accumulated during Work days up to 4 GB



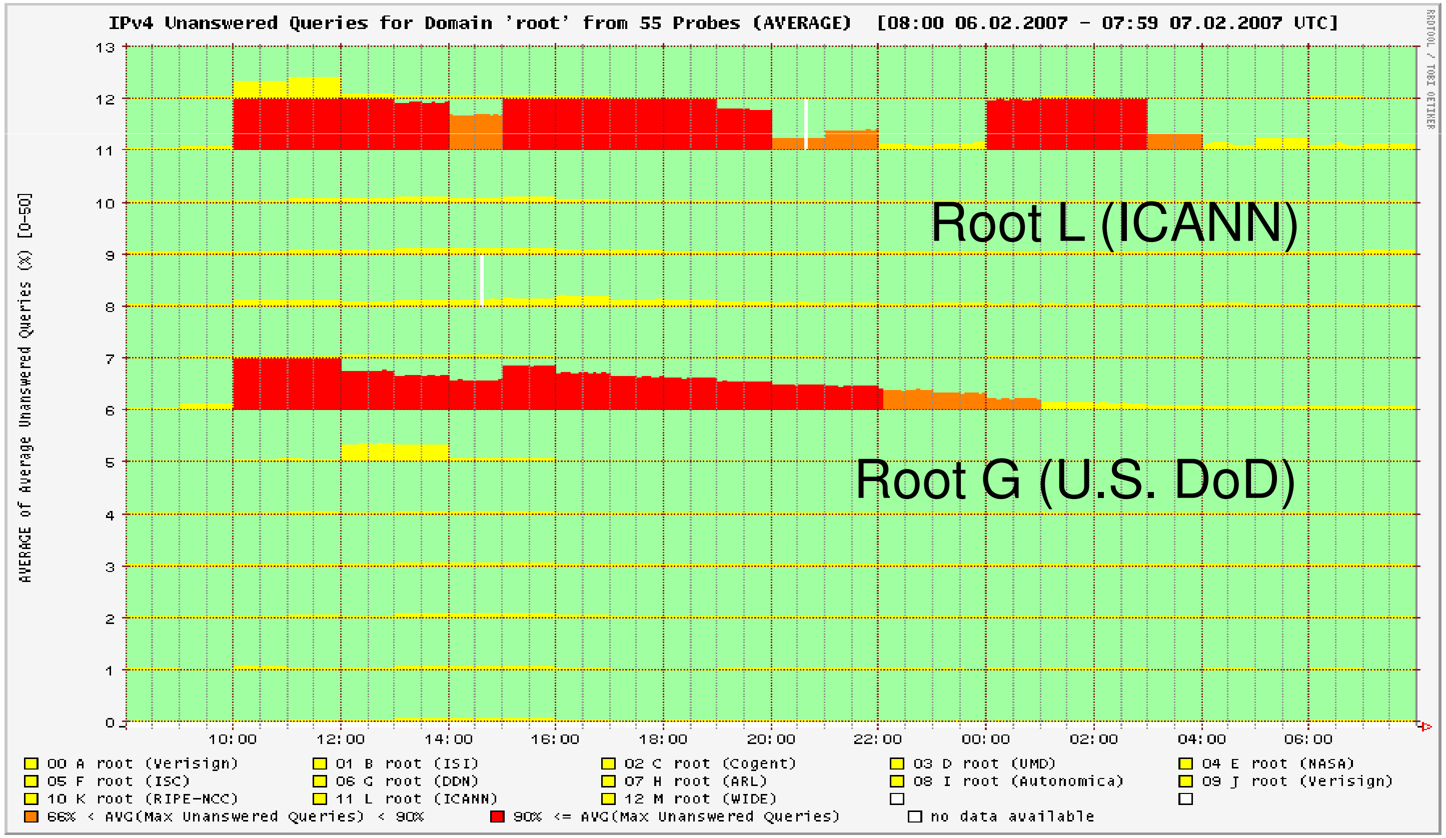
Anomaly
→ Different kinds of anomalies

- **System break down in local network (DNS server)**
  - Accumulated examination of the data traffic shows a massive increase, destination port, massive increase of data traffic on UDP port 53, source port: no traffic
  - separate examination of source and destination ports to avoid misinterpretation.



Anomaly
→ Different kinds of anomalies

- Attack towards the network infrastructure
  - Local attack: DDOS against www.heise.de 2005
  - Global attack: DDOS against all DNS Root Server 2002 / 2006 / 2007



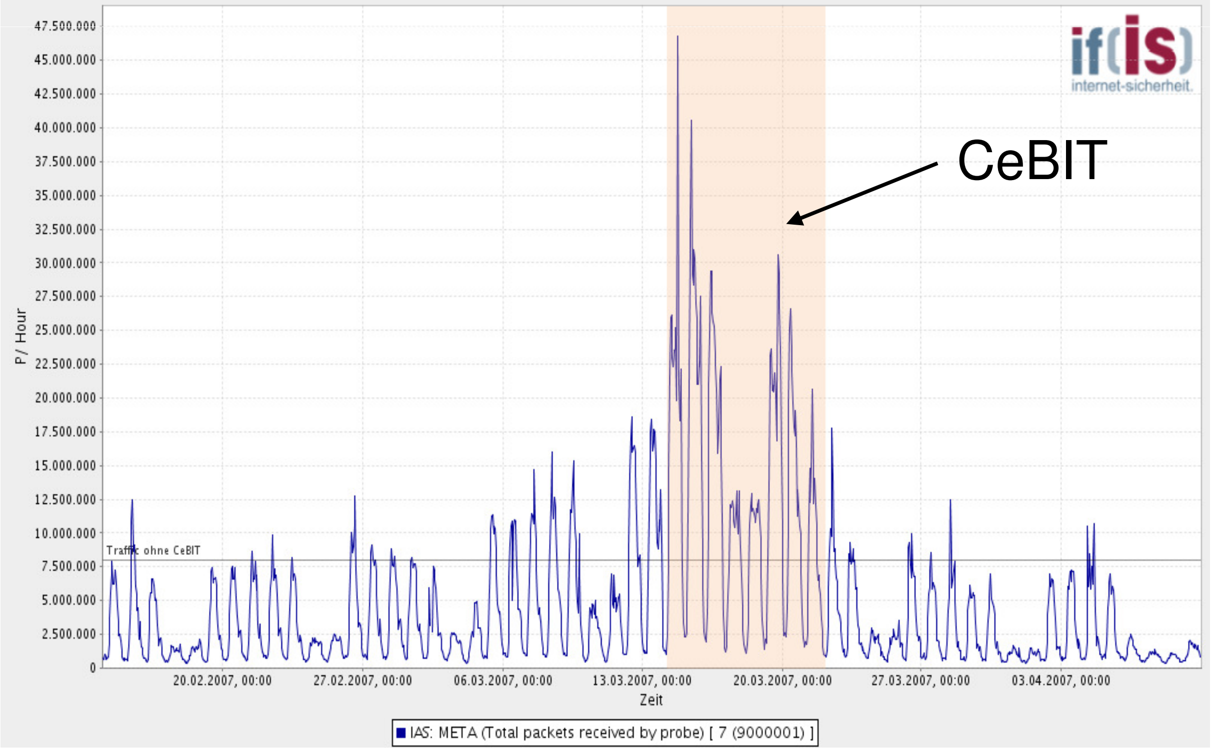
Root L (ICANN)

Root G (U.S. DoD)

Quelle: RIPE

Anomaly
→ Different kinds of anomalies

- **Results**
  - local: predictable or not
    - when predictable strictly speaking not an anomaly
  - Global: terror attacks (September 11th 2001 / March 11th 2004)
    - Strictly speaking no attacks but “collateral damage”

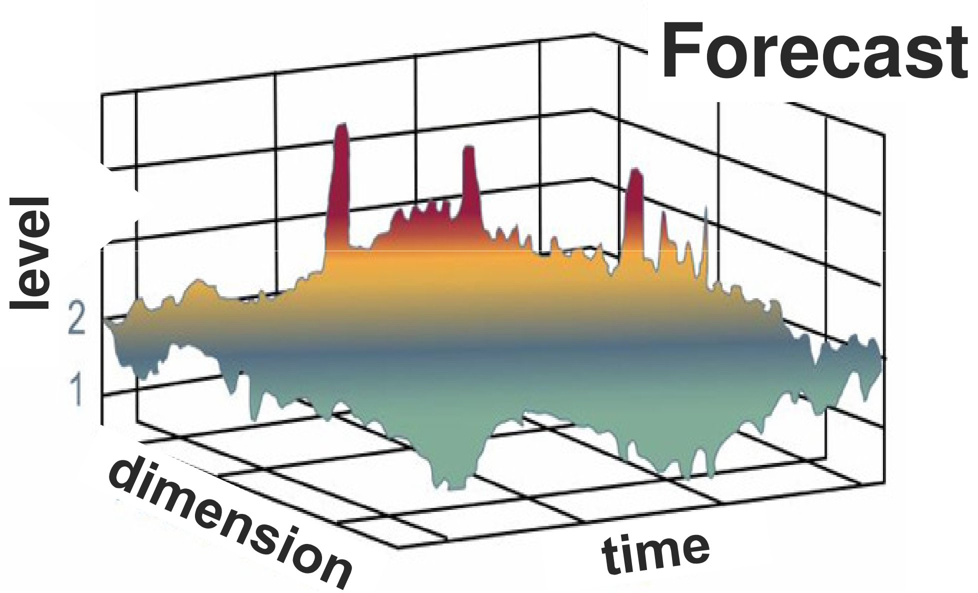


Content

- Aim and outcomes of this lecture
- Idea of the Internet Analysis System
- Knowledge Base
- Outline of the Current State
- Detection of Attacks and Deflection
- **Forecast of Patterns and Attacks**
- Summary

Target 4

- **Forecast of patterns and attacks.**



By investing and analyzing the extrapolated profiles, technological trends, correlations and patterns it is possible to make forecasts about the changes in the state of the internet by an evolution process of the findings.
In this way attacks and important changes can already be identified early and this helps us to avoid damage in the internet.

Internet Analysis System (IAS)
→ Forecast of patterns and attacks

- Two aspects are of relevance to secure the operability of the Internet:
  - The network has to be prepared for emerging technology
  - Attacks have to be detected in time and the distribution has to be prevented efficiently

=> 1. Technology trends have to be detected early enough!

2. The initial phase of attacks needs to be better understood and described!

3. The distribution of attacks needs to be understood!

4. Security mechanisms need to enhance secure cryptographic methods

=> Ability to forecast and to identify patterns

Forecast for the Internet Analysis System → Targets

- Assistance to generate forecasts, project measured values to the future

- Short time forecasts
  - Minutes up to days
  - Forecast and detection of deflections from the normal behavior, which can be used for Early Detection of attacks and anomalies

- Long term forecasts
  - Weeks up to months
  - Forecast and detection of technology trends

Characteristics of the measured values

- **The IAS collects measured values in the interval of (at the moment) 5 minutes**

- **Identical monitoring at the same location (IAS sensor)**

- **Parameters are discrete**
  - **=> analysis of time series**
    - Moving average (arithmetic mean)
    - Exponential smoothing (arithmetic mean including the loading with past values)
    - Linear regression (Trend +/- seasonal examination)
    - Holts-Winters-Method (Trend +/- seasonal examination including loading)

Component model of the time series

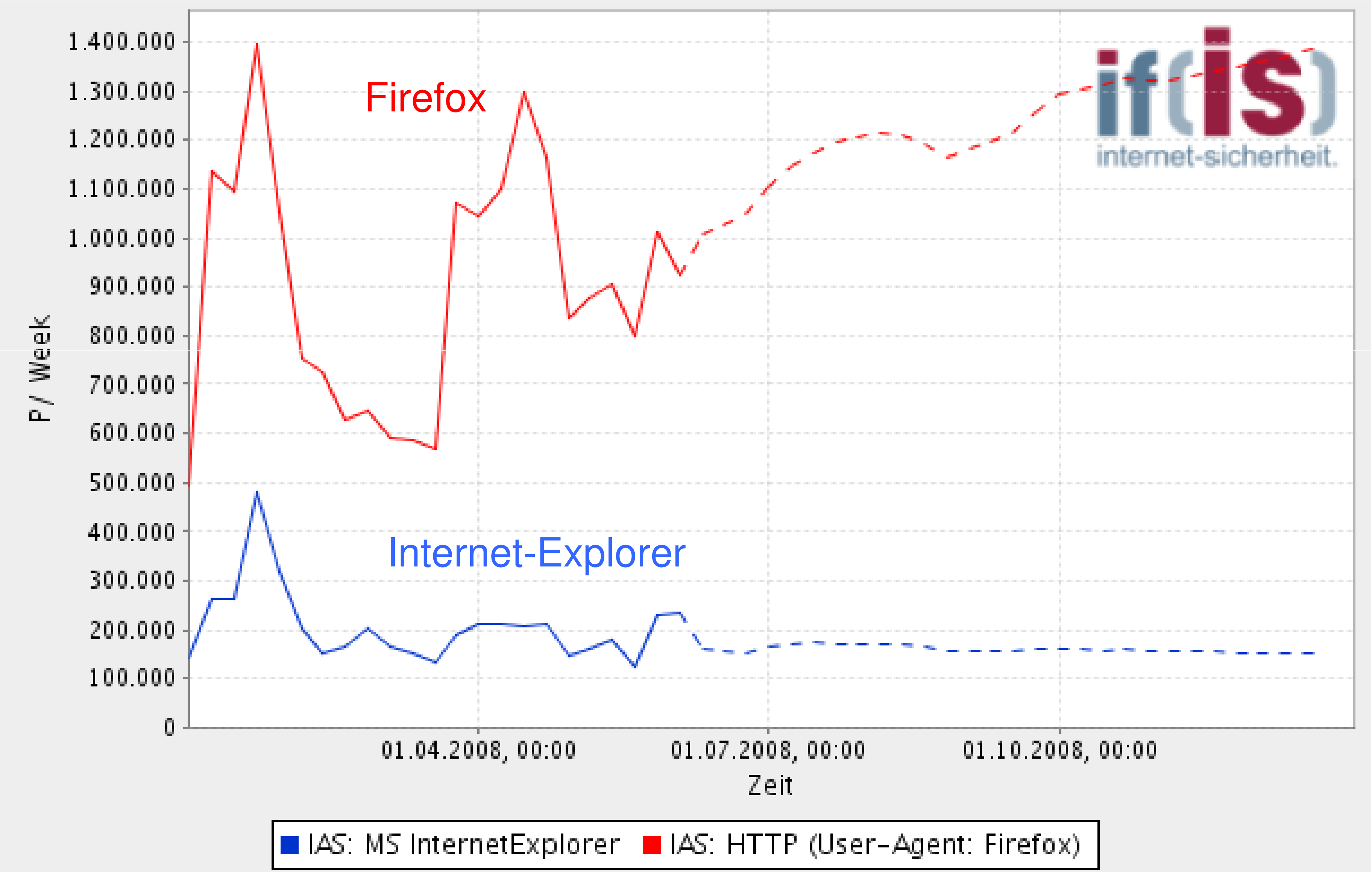
$$time\ series = Trend + economic\ cycle + seasonal + Rest$$

- Time series consists out of different components
  - Trend
    - Long-term changes of the average (direction)
  - Economic cycle
    - Short-term changes (local trends)
  - Seasonal
    - Variations due to day, night, weekends etc.
  - Rest
    - Unaccountable influences or breakdowns

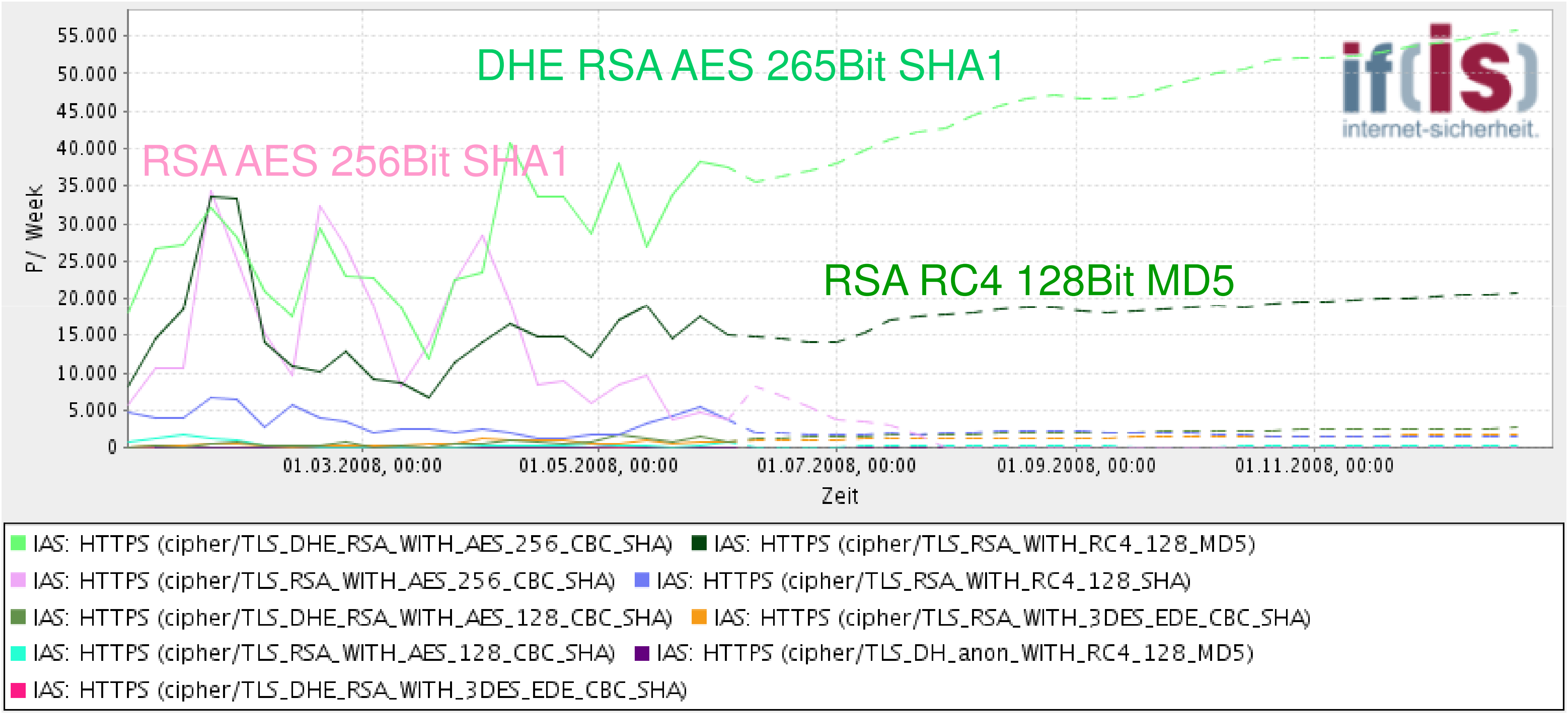
Live-Demo: EagleX Plotter

- Statistics of browsers used
- SSL (HTTPS Cipher)
- Operating systems by the means of IP TTL
- Overload on the mail system

Internet Analysis System (IAS)
→Technology trend (Firefox vs. IE)



Internet Analysis System (IAS)
→Technology trend: SSL (HTTPS Cipher)

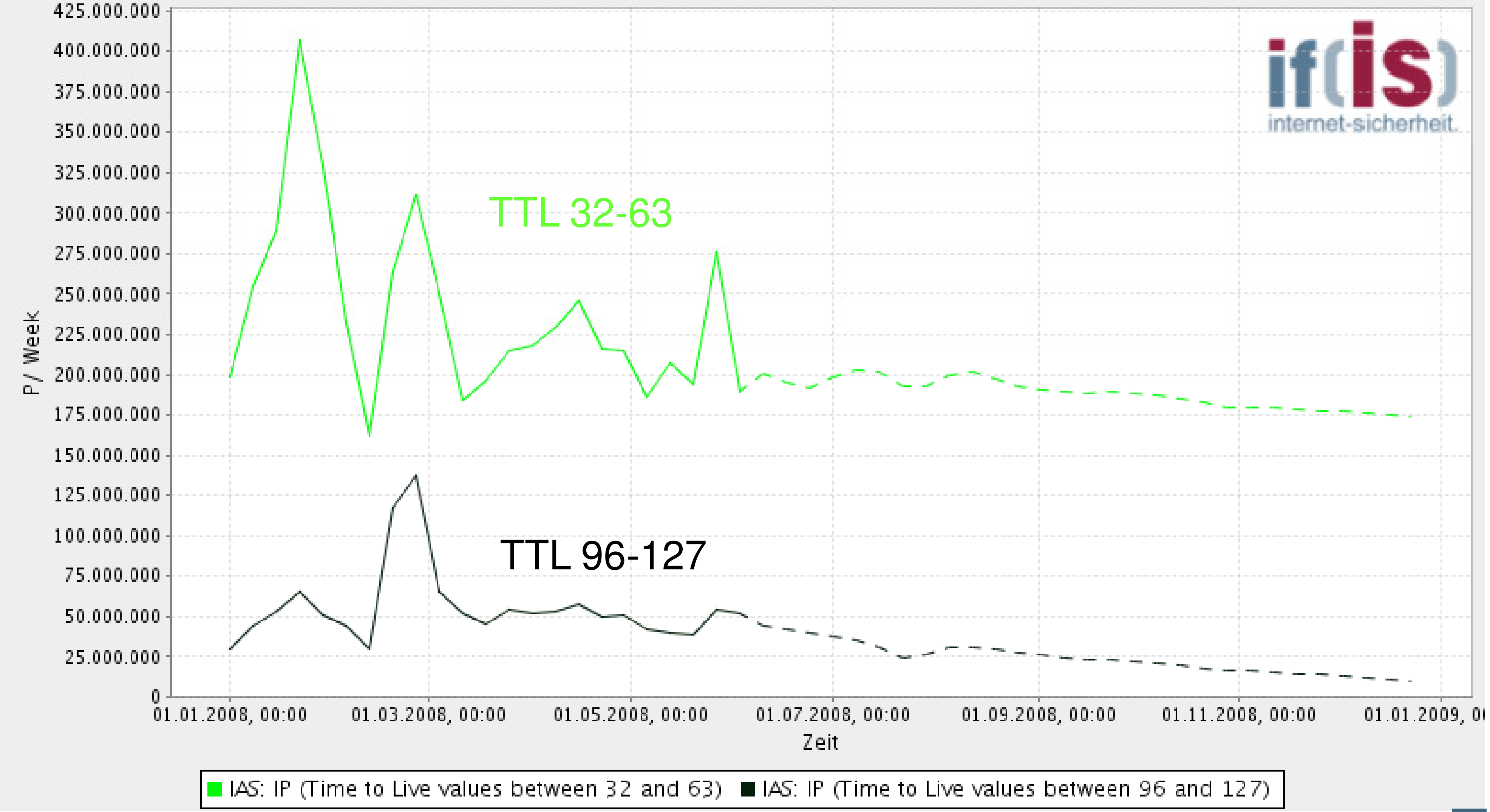


DHE RSA AES 265Bit SHA1

RSA AES 256Bit SHA1

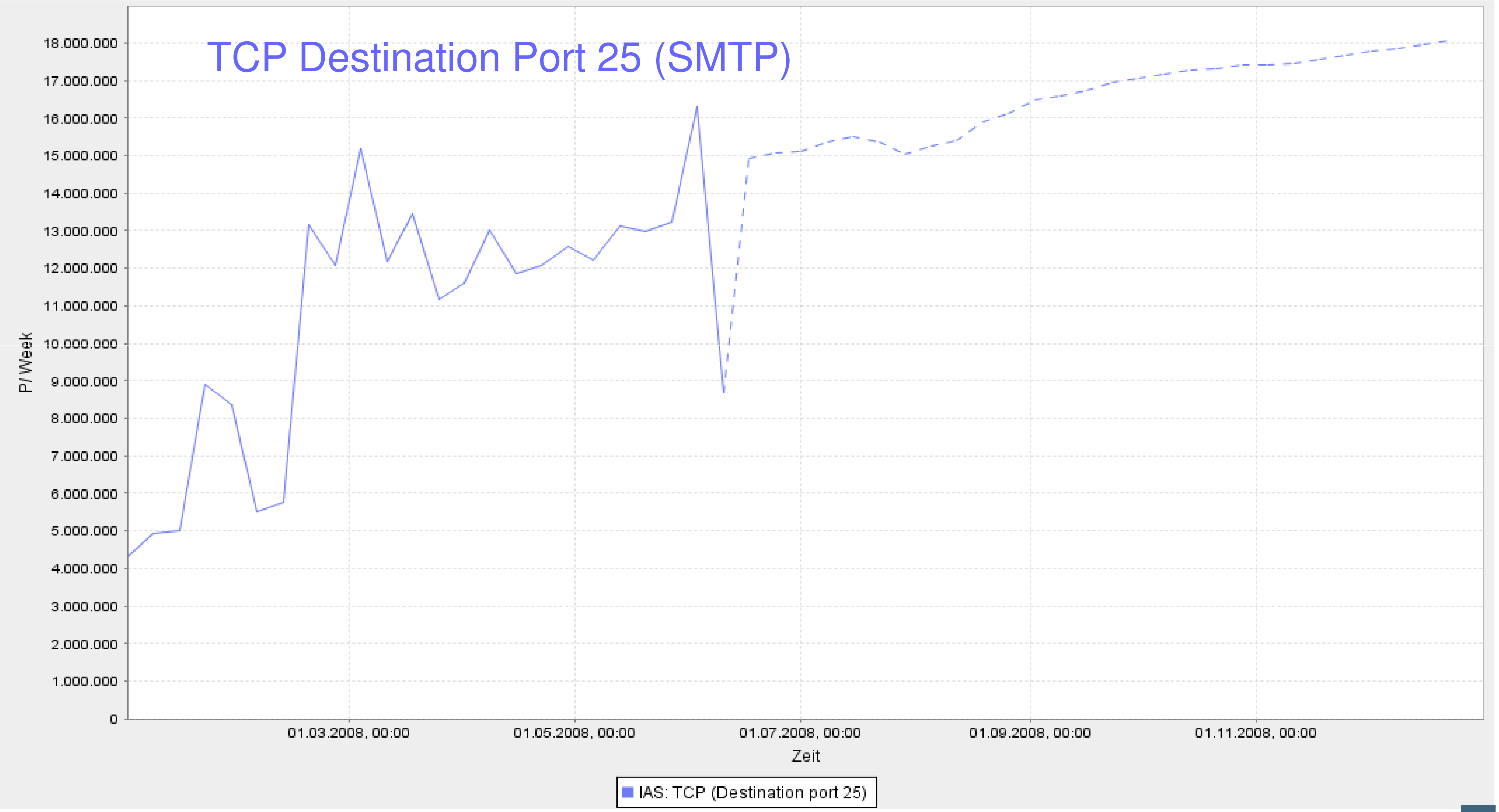
RSA RC4 128Bit MD5

Internet Analysis System (IAS)
→T-trend: Operation Systems (IP TTL)



- Default value: Linux: 64 and Windows: 128

Internet Analysis System (IAS)
→Technology trend (TCP Dst Port 25)



Results

- **Long-term forecast**
  - Possible with or without seasonal
    - Without season the trend is noticeable very clearly
  - Linear regression offers the greatest accuracy
  - In case of heavy noise methods of smoothing are more precise

- **Short-term forecast**
  - Seasonal examination is important
    - Day- and night changes, lunch break
    - Working day and holidays
  - When you break down to an interval of hours: Linear regression
  - In the interval of minutes: Holts-Winters-Method
  - The shorter the interval the more exact Holts-Winters turns out to be

Content

- Aim and outcomes of this lecture
- Idea of the Internet Analysis System
- Knowledge Base
- Outline of the Current State
- Detection of Attacks and Deflection
- Forecast of Patterns and Attacks
- **Summary**

Internet Analysis System (IAS)
→ Capabilities

- The use of counters enables a wide spectrum for utilization
  - No problem with laws
- All layers are being analyzed, especially the application layer as well
  - Trends in the use of technology
  - Detection of attacks even on the level of the application layer
  - Understanding of the detailed network traffic
  - Differences between theory and reality
  - ...
- Results of very many sensors can be stored for a long period of time, since very little amounts of storage are needed for archiving
- Ideal to be used for internet monitoring cooperations (global view)

Internet Analysis System (IAS)
→ Limitations

- Statistical conclusions
- No IP addresses (e-mail addresses, user data …)
- The payload is not being analyzed
- So far no conclusions on routing or similar things
- No stream-wise analyzing of the data traffic
- Pointed attacks towards single systems can not be detected (perishes in noise)
- So far the decision which protocol is expected to be found in the packet and which plug-in should be used for analyzing is done due to the used port
- Records cannot be used for forensics

Internet Analysis System (IAS)
→ Compensation of Limitations

- It can be determined with a certain likelihood, whether an attack is initiated by one or many computers (bots).
- If ports are used that so far have never been used, the IAS will detect this as an anomaly right away.
- If an attack has been detected the administrator of a domain can (by law) use other sources (like log files) to look at privacy relevant data (e. g. IP addresses etc.) to identify the attacker and to initiate counteractive measures.

Internet Analysis System (IAS)
→ Summary

- The IAS comes with a comprehensive knowledge base
- The IAS allows a continuous situation awareness to be recorded and displayed.
- Attacks can be detected
- Forecasts can be made and displayed
- The IAS is a great concept to build a global view (take a look at the lecture about global view)
- In combination with other systems a comprehensive monitoring tool is created which offers a great value to the user.

Fachhochschule Gelsenkirchen

Internet Analysis System → IAS

Thank you for your attention!
Questions?

Prof. Dr.
Norbert Pohlmann

Institute for Internet Security - if(is)
University of Applied Sciences Gelsenkirchen
http://www.internet-sicherheit.de

if(is) internet security.

Internet Analysis System (IAS)
→ Literature (1/2)

- [1] N. Pohlmann: "Internetstatistik" (statistics of the internet), Proceedings of CIP Europe Publisher, B.M. Hämmerli, 2005.
- [2] N. Pohlmann, M. Proest: „Internet Early Warning System: The Global View", in "Securing Electronic Business Processes - Highlights of the Information Security Solutions Europe 2006 Conference", Hrsg.: S. Paulus, N. Pohlmann, H. Reimer, Vieweg-Verlag, Wiesbaden 2006
- [3] N. Pohlmann: "Probe-based Internet Early Warning System", ENISA Quarterly Vol. 3, No. 1, Jan-Mar 2007
- [4] N. Pohlmann: „The global View of Security Situation in the Internet", ECN - European CIIP Newsletter, Volume 3, Brüssel 12/2007
- [5] Sebastian Spooren, Entwicklung eines profilgestützten Visualisierungssystems zur Darstellung von raum- & zeitbezogenen Soll-/Ist-Abweichungen (development of a visualization tool for the IAS), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2007.
- [6] Gianfranco Ricci, Betrachtung der vom IAS gesammelten Kommunikationsparameter auf Relevanz zur Anomalie und Angriffserkennung (evaluation of the relevance for the detection of abnormalities and attacks of the communication parameters collected by the internet analysis system), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2008

Internet Analysis System (IAS)
→ Literature (1/2)

- [7] Uwe van Heesch: Entwicklung eines Plugin basierten Analyse-Frameworks für das Internet-Analyse-System (development of a plugin-based analyzing framework for the Internet Analysis System), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2006.
- [8] Sabyasachi Basu, Amarnath Mukherjee, Steve Klivansky: Time Series Models For Internet Traffic, 1996
- [9] Peter J. Brockwell, Richard A. Davis: Introduction To Time Series and Forecasting, Springer, 2002

Links:

Institute for Internet Security:
http://www.internet-sicherheit.de/forschung/aktuelle-projekte/internet-frhwarnsysteme/