Методы инспекции пакетов и анализа трафика

Лекция 6
Принцип работы DPI

Фицов Вадим Владленович

ст.преп. кафедры ИКС

Содержание лекции:

- Потоки трафика

- Виды глубокого анализа трафика

- Сигнатурный анализа
- Поведенческий анализ
- Статистический анализ
- Эврестический анализ
- DataMining анализ

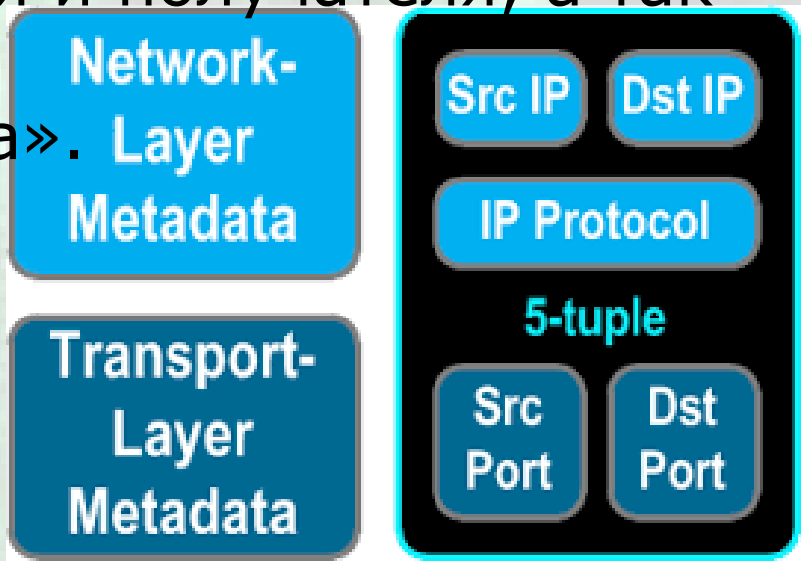
Потоки трафика

7 tuples (кортеж 7)

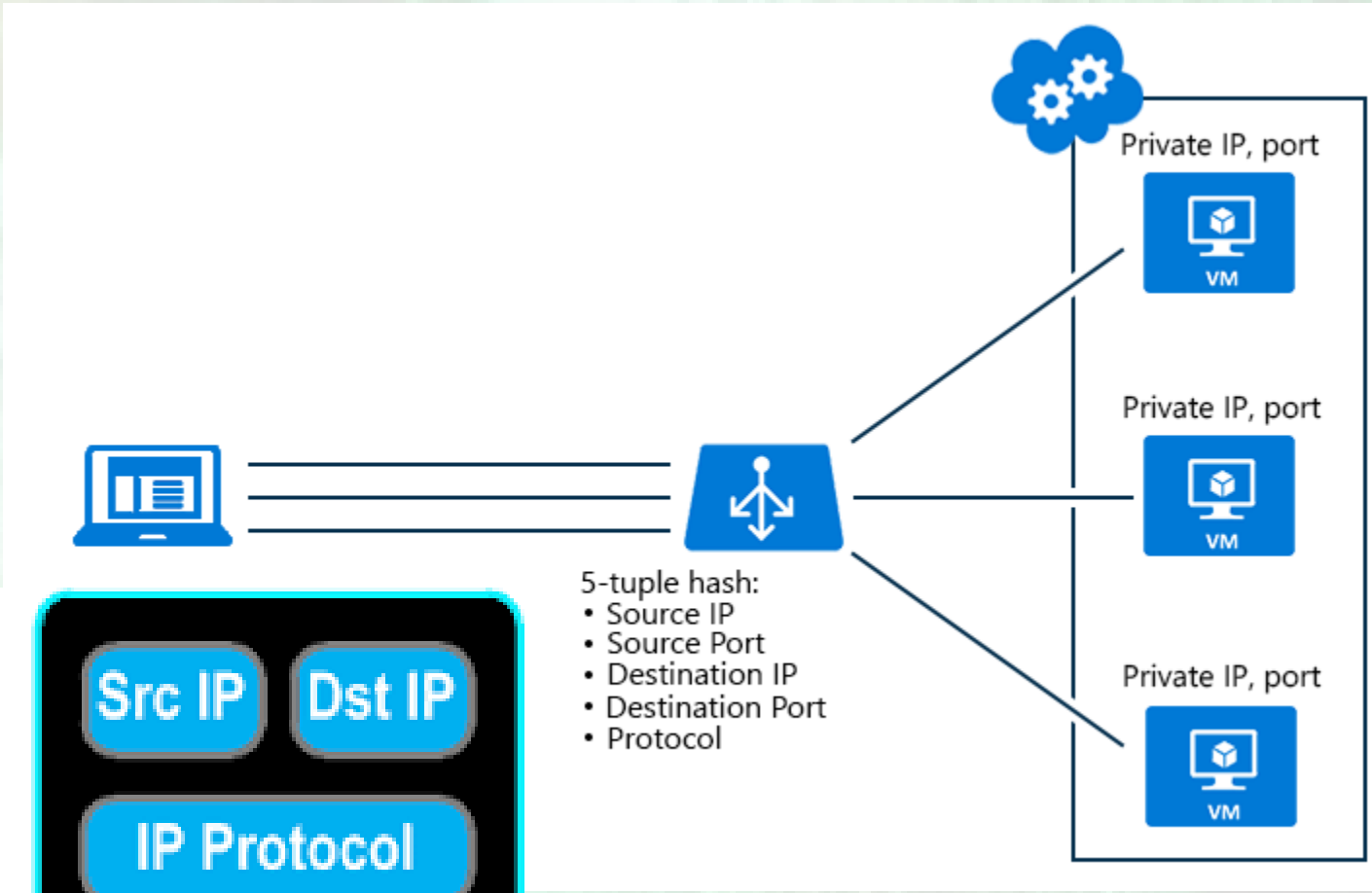
Поток идентифицируется с помощью адресной информации 2–4 уровней.

Критерии, по которым пакет относят к потоку, называют кортежем (tuple).

В кортеж может входить разное число критериев, обычно от 3 до 7: MAC-адреса отправителя и получателя, IP-адреса отправителя и получателя, транспортные порты отправителя и получателя, а так же значение
служебного поля «Тип протокола».

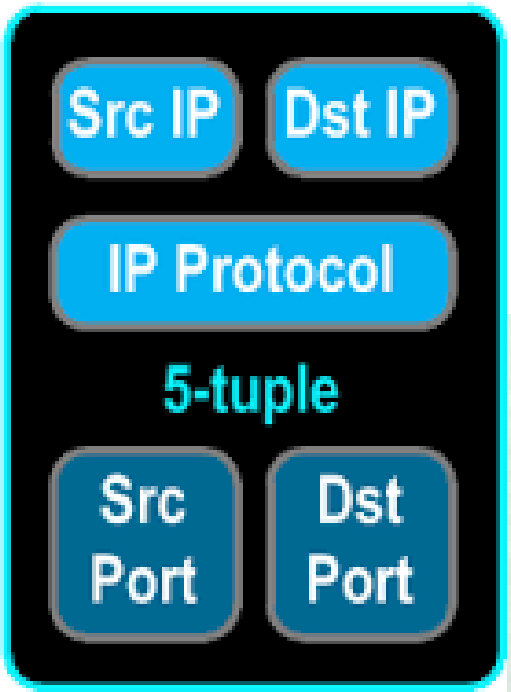


7 tuples (кортеж 7)



Network-Layer Metadata

Transport-Layer Metadata



Анализ:

- распределение по типам протоколов

(%, кол-во пакетов, кол-во байт, скорость)

- распределение по длине кадров
- расчет скорости для каждого протокола
- построение графиков (пакетов/с, бит/с)
- список сайтов к которым обращаются HTTP запросы
- распределение количества пакетов по IP адресам
- обнаружение udp multicast

6

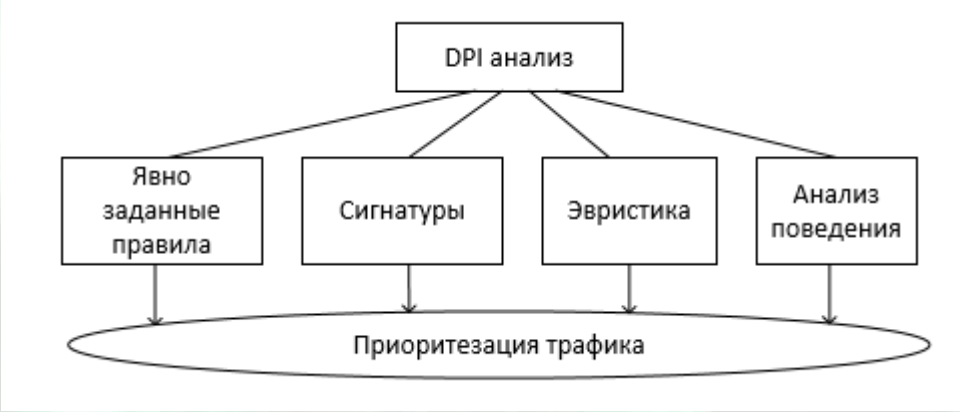
Виды глубокого анализа трафика

DPI средства управления трафиком

Что происходит при DPI?

DPI анализирует

- первые пакеты потока трафика
- или все проходящие через нее пакеты.



- Применяются:

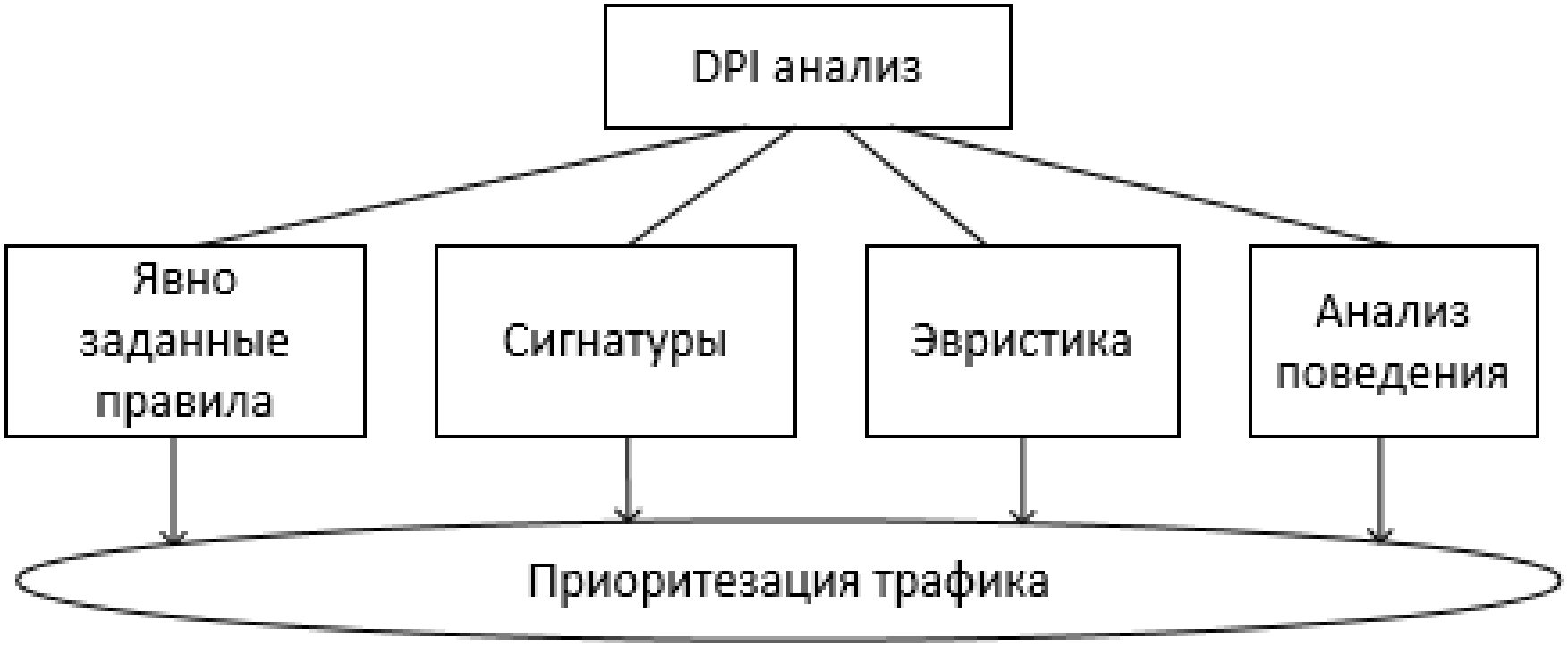
  - сигнатурный анализ,

  - статистические методы слежения за характеристиками пакетов,

  - поведенческий анализ и др.

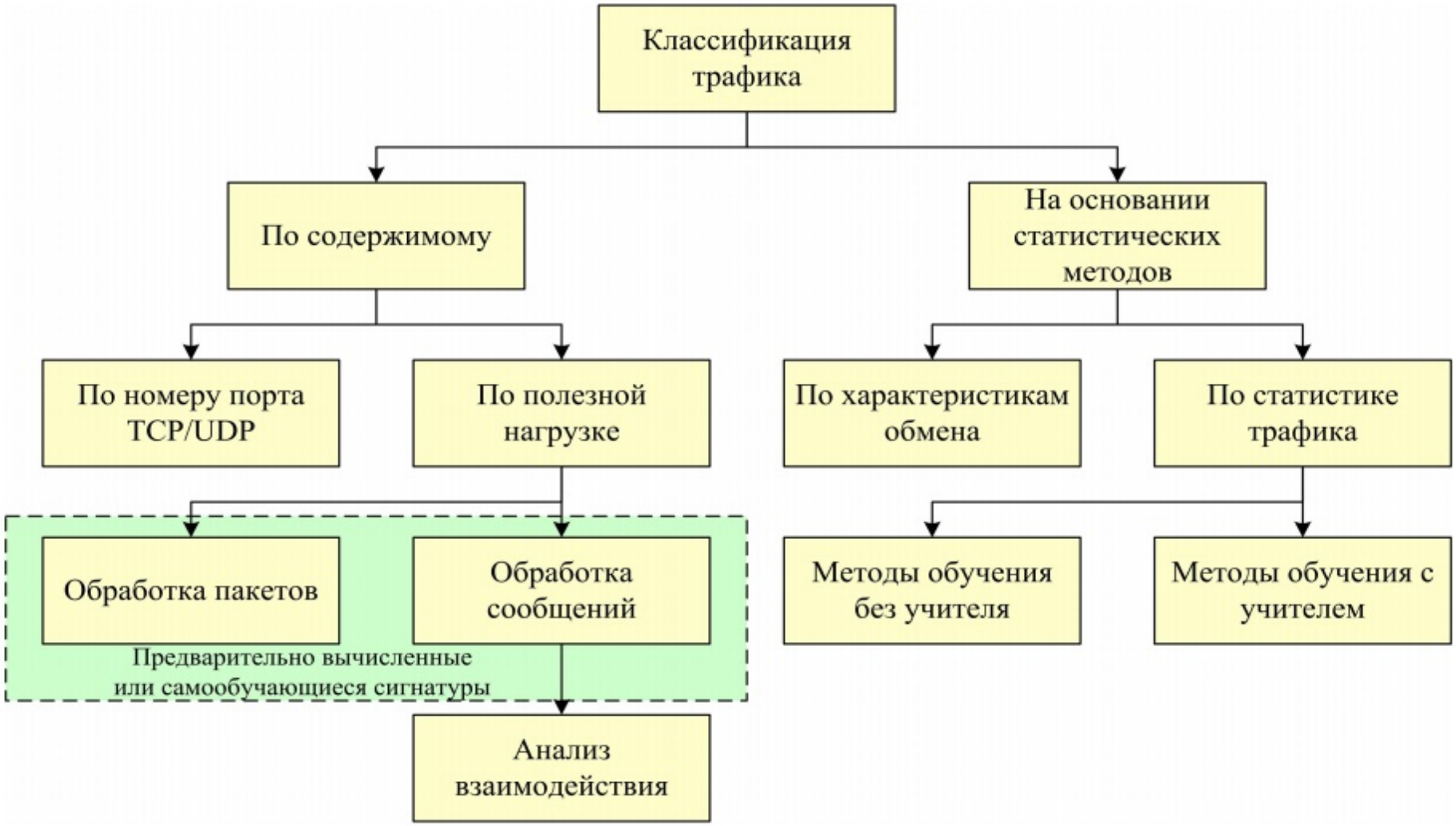
9

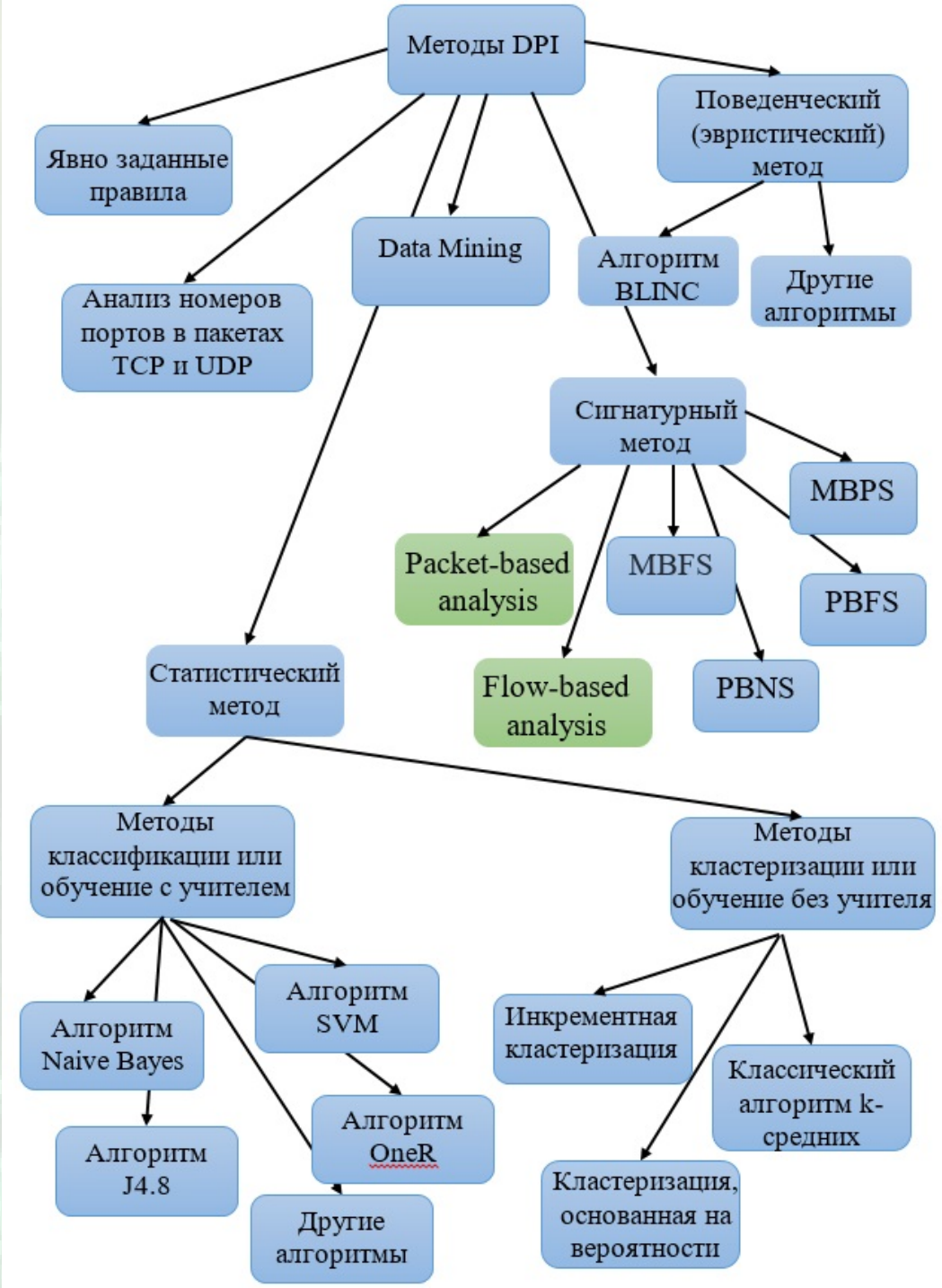
Что происходит при DPI?



10

Методы классификации сетевого трафика




Методы DPI
Явно заданные правила
Поведенческий (эвристический) метод
Анализ номеров портов в пакетах TCP и UDP
Data Mining
Алгоритм BLINC
Другие алгоритмы
Сигнатурный метод
MBPS
Packet-based analysis
MBFS
PBFS
Статистический метод
Flow-based analysis
PBNS
Методы классификации или обучение с учителем
Методы кластеризации или обучение без учителя
Алгоритм Naive Bayes
Алгоритм SVM
Инкрементная кластеризация
Классический алгоритм k-средних
Алгоритм OneR
Алгоритм J4.8
Кластеризация, основанная на вероятности
Другие алгоритмы

12

Сигнатурный анализ

13

Сигнатурный анализ

Сигнатура – это набор байтов в пакете или файле, позволяющий однозначно определить, к какому приложению, протоколу относится трафик, и классифицировать его.

Сигнатуры разрабатываются и распространяются вендором.

Базу сигнатур необходимо обновлять автоматически или вручную.

14

Сигнатурный анализ

стандартные паттерны (например, HEX),

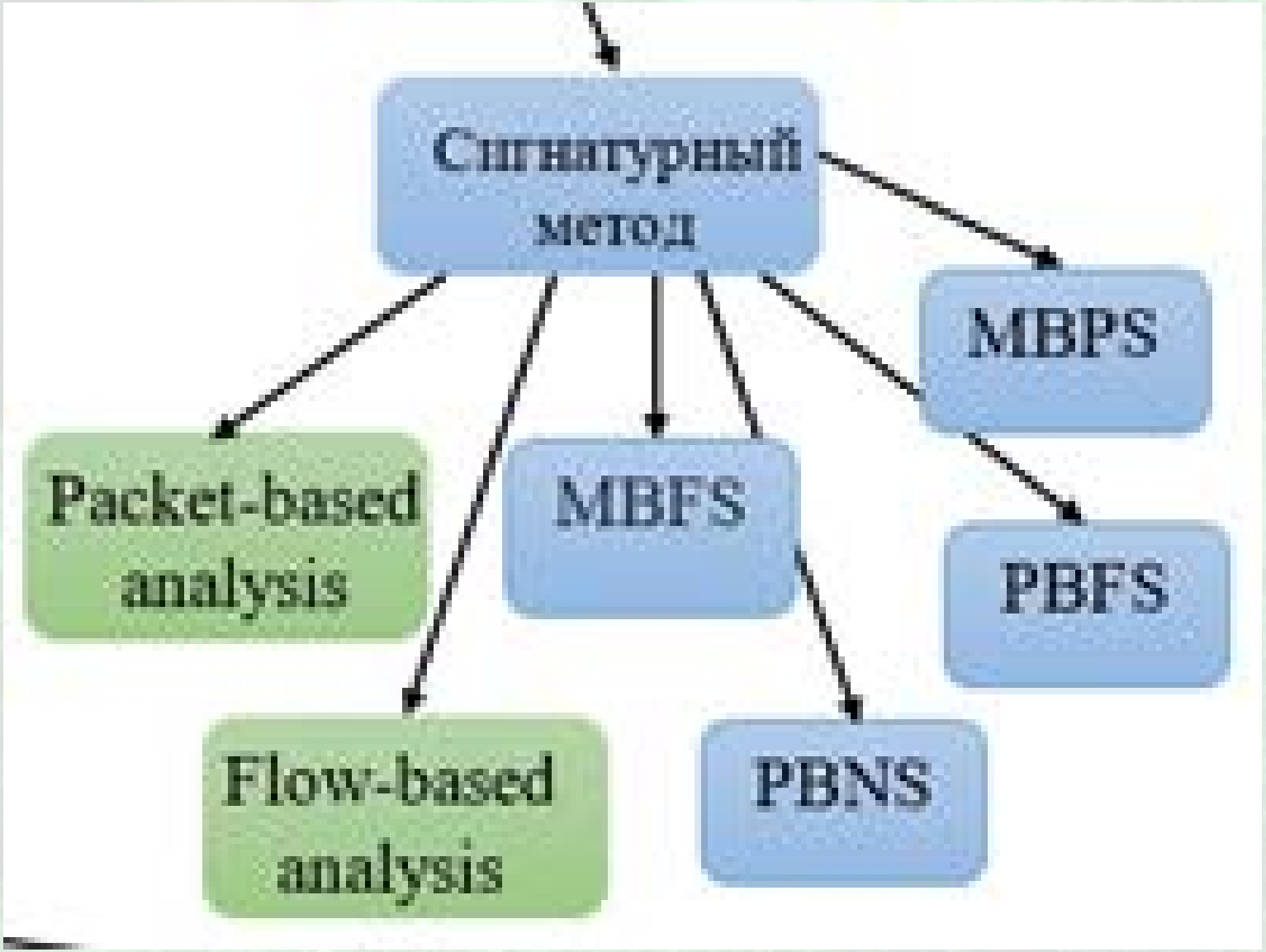
по которым можно однозначно определить принадлежность пакета определенному приложению.

Например:

- по формату заголовков,
- номерам портов и т. п.

15

Сигнатурный анализ



16

Сигнатурный анализ

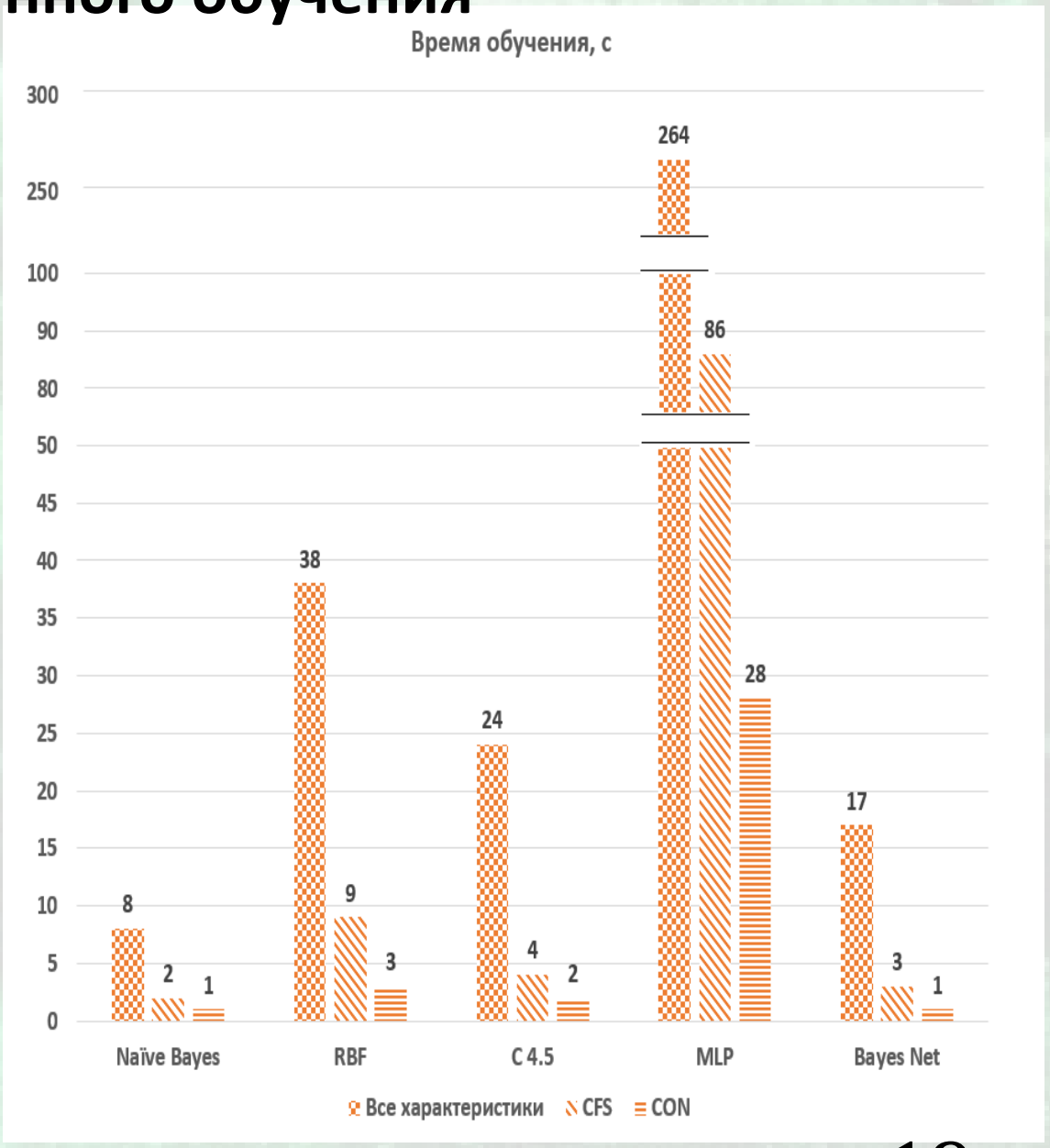
	Скорость сравнения в зависимости от синтаксиса сигнатур, с.		
Алгоритм	Фиксированный фрагмент	Плавающий фрагмент	Регулярные выражения
Rabin-Karp	0,03	1,28	3,45
DFA	0,05	0,32	0,19
NFA Full	0,08	0,42	0,16
NFA Particial		0,90	0,08

Сигнатурный анализ

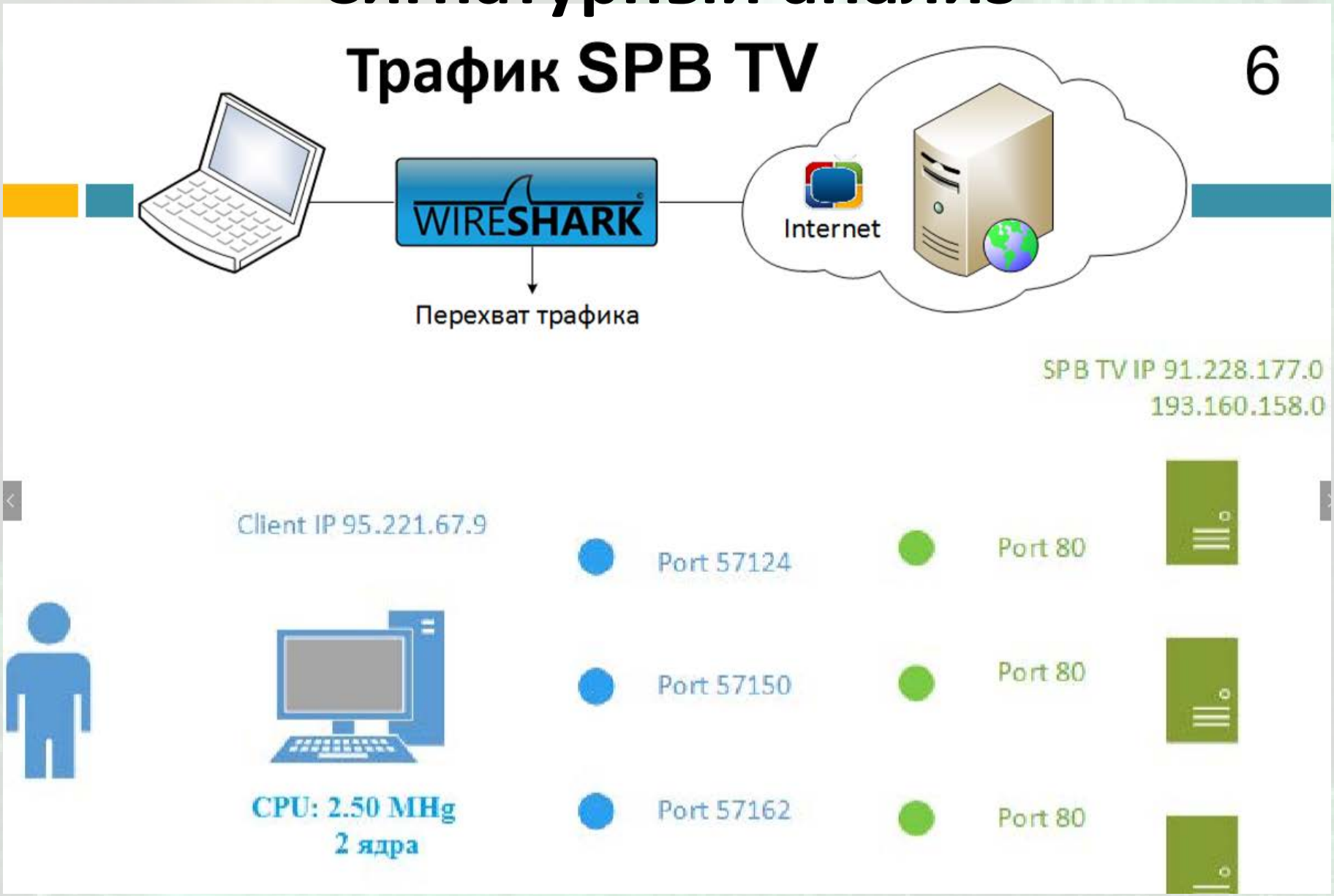
Тип сигнатуры	Сигнатура	Время сравнения при DFA Full, с.	Время сравнения при NFA Partial, с.
SSSA	^GET.*NateOn.*	0,23	0,05
	^GET.*nateon\.nate\.com.*	0,22	0,06
	^GET.*adimg\.nate\.com.*	0,18	0,11
	^GET.*cyad\.nate\.com.*	0,25	0,15
	^GET.*nateonipml\.nate\.com.*	0,21	0,10
MSSA	^GET.*(NateOn)|(nateon\.nate\.com)|(adimg\.nate\.com)|(cyad\.nate\.com)|(nateonipml\.nate\.com).*	0,22	9,24

18

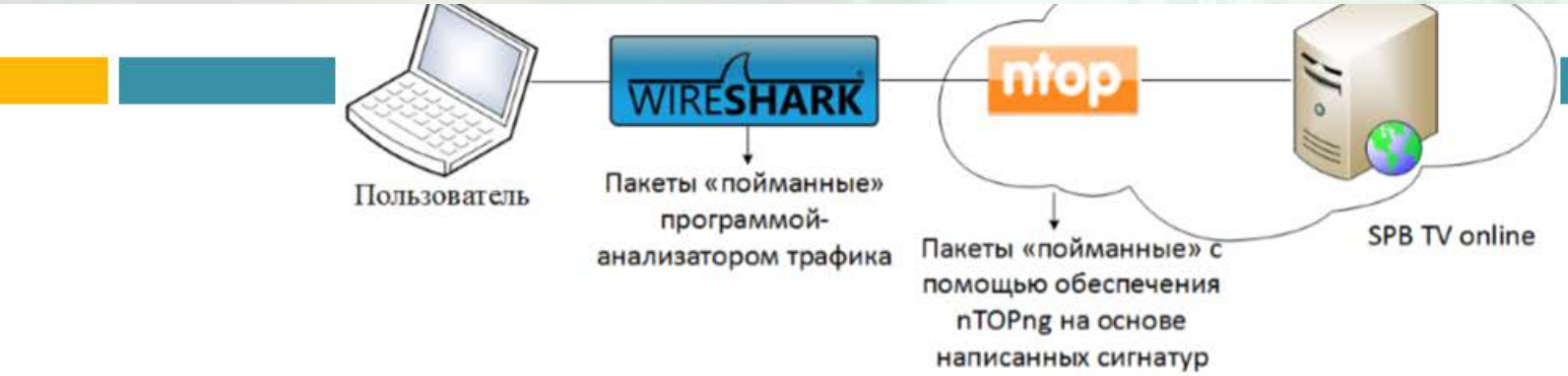
Анализ результатов исследования алгоритмов машинного обучения



Сигнатурный анализ



Сигнатурный анализ

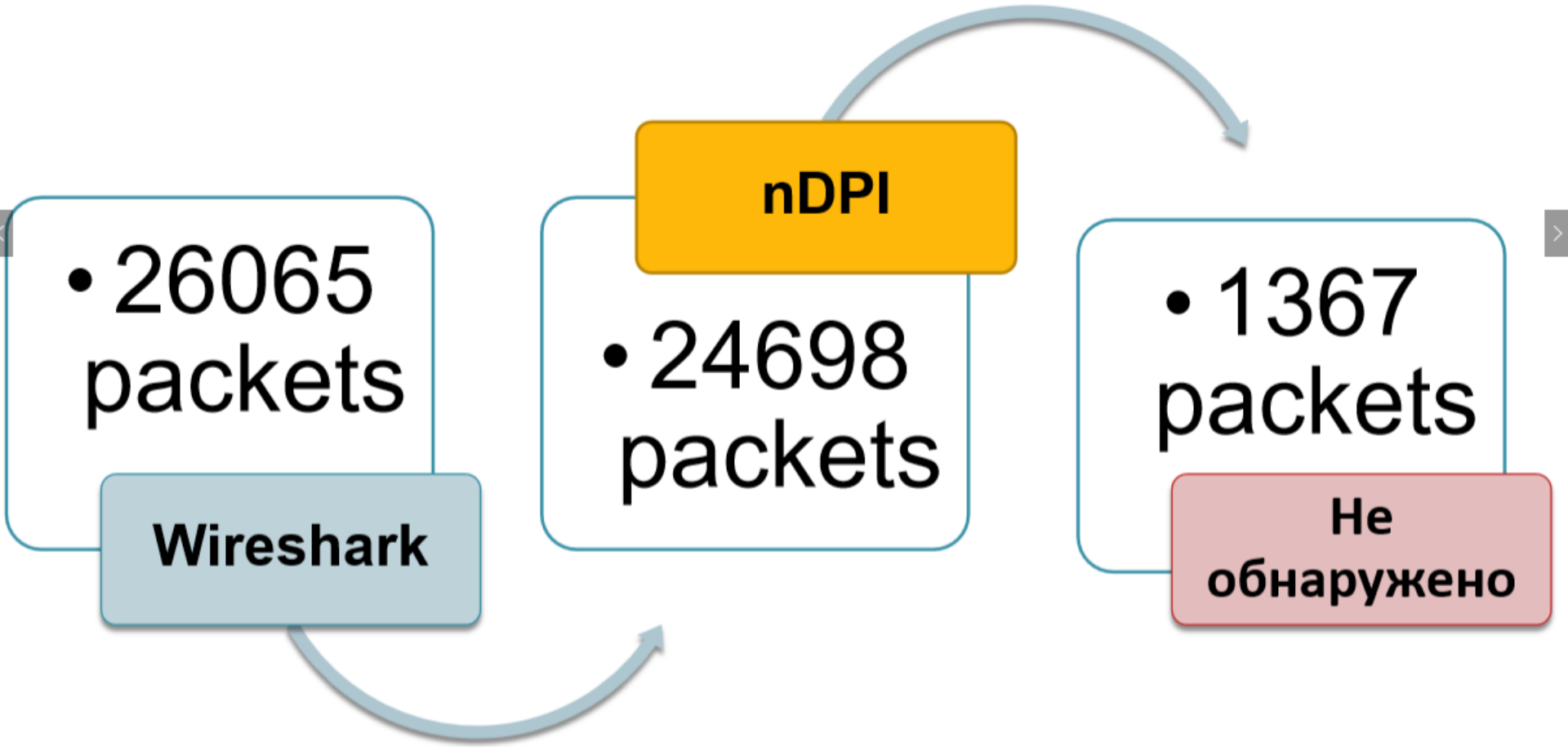


ntop | Flows | Hosts | Devices | Interfaces | Search Host

Host: 91.228.177.209 | Traffic | Packets | Ports | Peers | Protocols | Activity | Flows | Talkers

(Router/AccessPoint) MAC Address	CiscoInc_27:A2:40 (44:03:A7:27:A2:40)	
IP Address	91.228.177.209 [Moscow]	Host Pool: Not Assigned
ASN	SPB TV Telecom LLC ASN 197888]	Whois Lookup
Name	91.228.177.209 Remote	
First / Last Seen	05/06/2017 23:52:07 [3 min, 16 sec ago]	05/06/2017 23:55:19 [4 sec ago]
Sent vs Received Traffic Breakdown	Sent / Rcvd	
Traffic Sent / Received	24,698 Pkts / 32.91 MB	14,910 Pkts / 1.13 MB
Active Flows / Total Active / Low Goodput	'As Client'	'As Server'
	0 / 2 / 0	2 / 14 / 0
TCP Packets Sent Analysis	Retransmissions	0 Pkts
	Out of Order	81 Pkts
	Lost	38 Pkts
Further Host Names/Information	91.228.177.209	
JSON	Download	

Сигнатурный анализ



Сигнатурный анализ

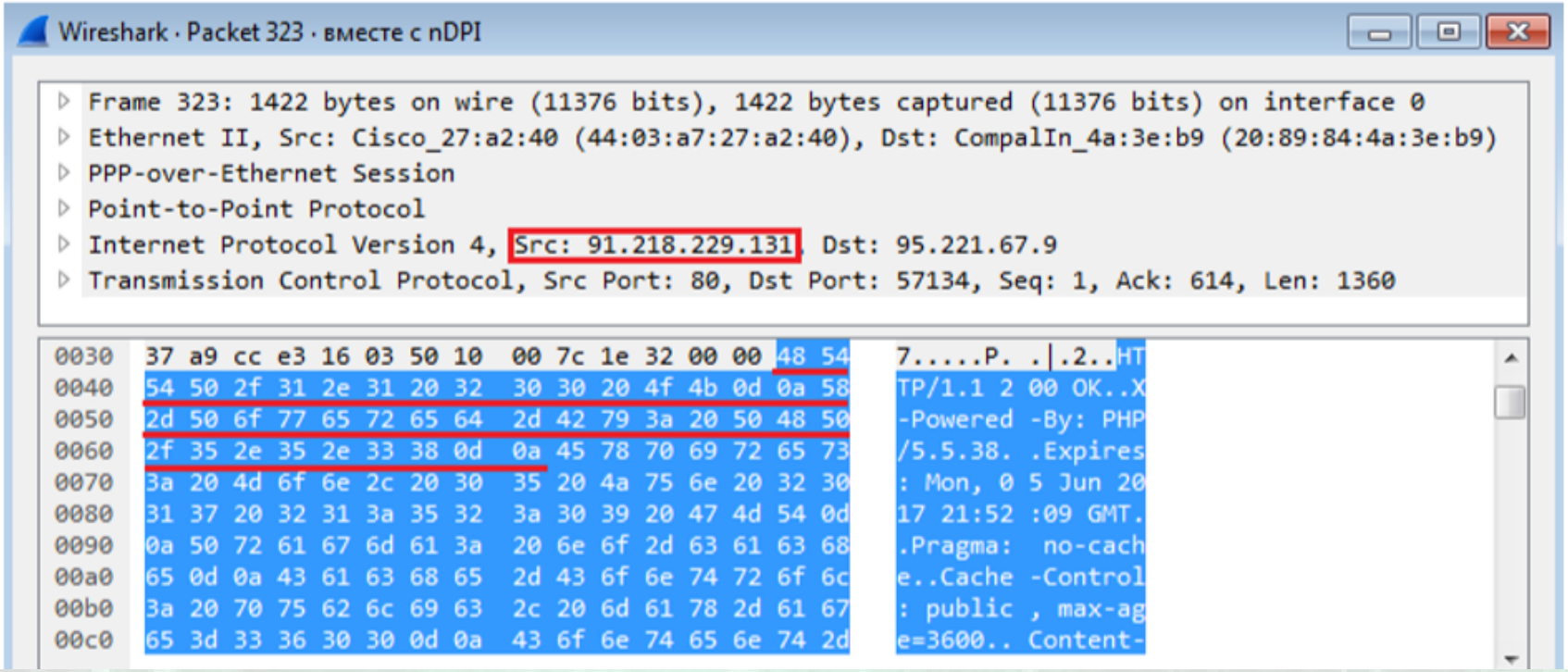
No.	Time	Source	Destination	Protocol	Length	Info
361	0.000436	95.221.67.9	81.19.88.103	HTTP	517	GET /top100.jcn?2815224 HTTP/1.1
362	0.002055	95.221.67.9	178.154.131.215	TCP	62	57143 → 443 [ACK] Seq=328 Ack=6303 Win=65776 Len=0
363	0.010957	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
364	0.000364	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
365	0.000005	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
366	0.000004	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
367	0.000131	95.221.67.9	81.19.88.103	TCP	62	57146 → 80 [ACK] Seq=456 Ack=2721 Win=65280 Len=0
368	0.000109	95.221.67.9	81.19.88.103	TCP	62	57146 → 80 [ACK] Seq=456 Ack=5441 Win=65280 Len=0
369	0.000139	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
370	0.000004	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
371	0.000010	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
372	0.000016	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
373	0.000098	95.221.67.9	81.19.88.103	TCP	62	57146 → 80 [ACK] Seq=456 Ack=8161 Win=65280 Len=0
374	0.000098	95.221.67.9	81.19.88.103	TCP	62	57146 → 80 [ACK] Seq=456 Ack=10881 Win=65280 Len=0
375	0.001072	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]
376	0.000005	81.19.88.103	95.221.67.9	TCP	1422	[TCP segment of a reassembled PDU]

(Router/AccessPoint) MAC Address	ZyxelCom_94:D6:00 (E8:37:7A:94:D6:00)	
IP Address	91.228.177.207 [Moscow]	Host Pool: Not Assigned
ASN	SPB TV Telecom LLC [ASN 197888]	Whois Lookup
Name	91.228.177.207 Remote	
First / Last Seen	15/06/2017 12:10:56 [1 min, 1 sec ago]	15/06/2017 12:11:52 [5 sec ago]
Sent vs Received Traffic Breakdown	Sent	
Traffic Sent / Received	2,755 Pkts / 3.65 MB	971 Pkts

23

Сигнатурный анализ

Последовательность в HEX: 48 54 54 50 2f 31 2e 31 20 32 30 30 20 4f 4b 0d 0a 58 2d 50 6f 77 65 72 65 64 2d 42 79 3a 20 50 48 50 2f 35 2e 35 2e 33 38 0d 0a



24

Сигнатурный анализ

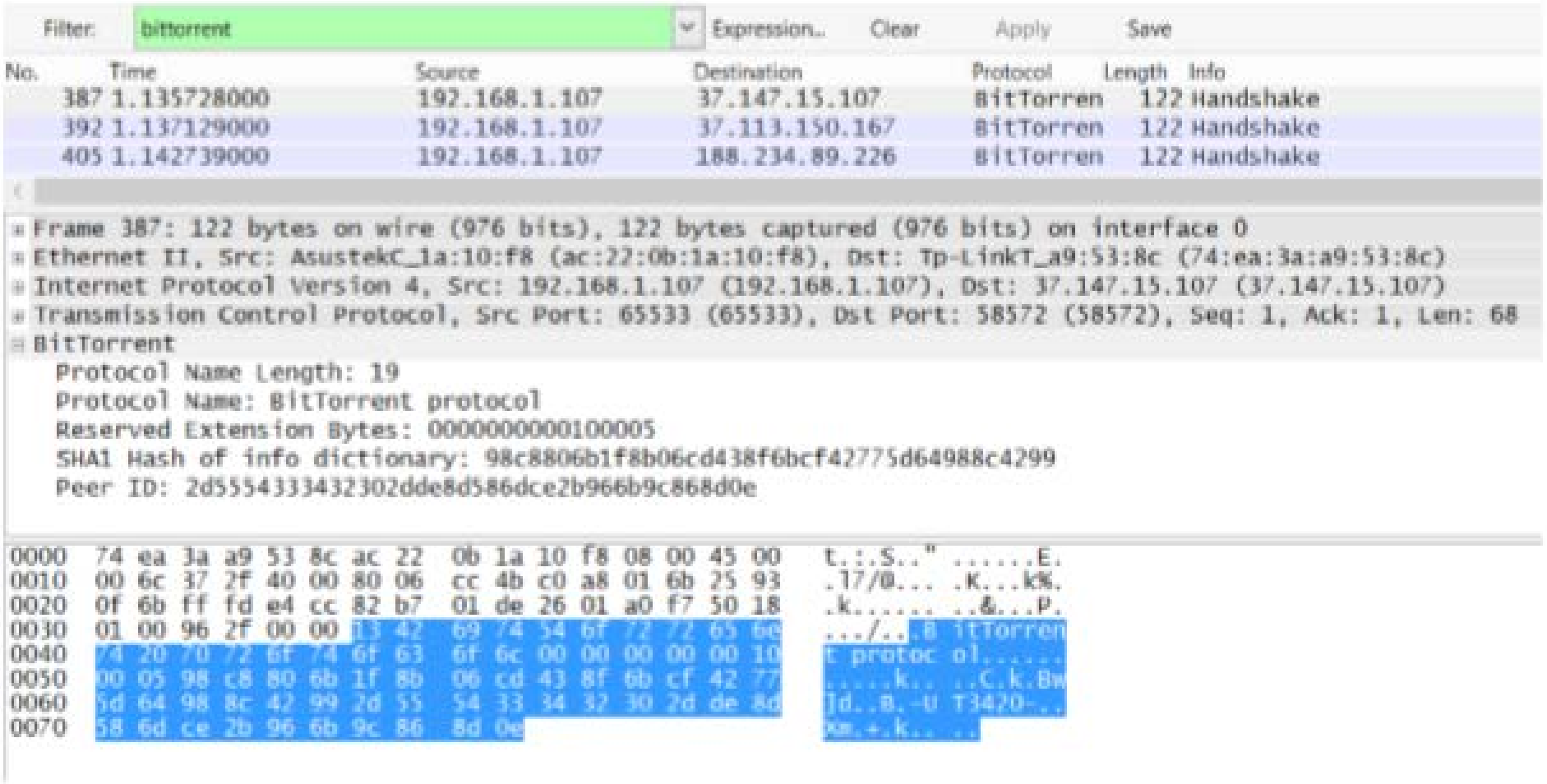


Рис. 2.10. Пакет BitTorrent, полученный с помощью Wireshark

25

Сигнатурный анализ

Заголовок BitTorrent при рукопожатии имеет следующий формат:
<acharacter (1 byte)><astring (19 byte)>.

Первый байт всегда фиксирован и имеет значение «19»,
а значение строки – «BitTorrent protocol».

Исходя из анализа поля данных, происходит идентификация BitTorrent-трафика:

1) первый байт в полезной нагрузке TCP-9 (0x13);
2) следующие 19 байт соответствуют строке «BitTorrent protocol».

26

Сигнатурный анализ

Таблица 2.1

Формат правила политики для трафика протокола BitTorrent

Правило политики для трафика BitTorrent	
Условие	Действие
ЕСЛИ: У1: «L4+Тип протокола = TCP» И У2: «Первый байт в поле данных TCP = 0x13» И У3: «Следующие 19 байт в поле данных TCP = «BitTorrent protocol»»	ТОГДА: Д1: «Присвоить низший приоритет»

Сигнатурный анализ

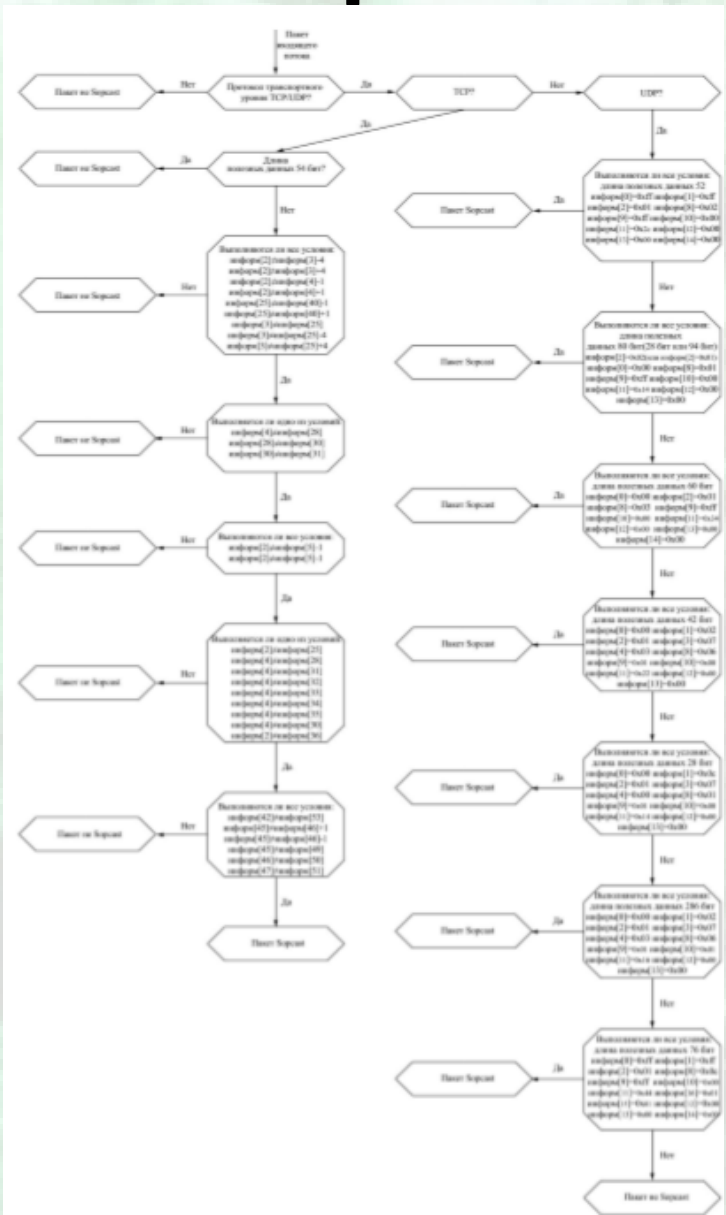
Таблица 2.2

Формат правила политики для трафика протокола HTTP

Правило политики для трафика HTTP	
Условие	Действие
ЕСЛИ: У1: «L4+Тип протокола = TCP» И У2: «Поле данных TCP = «GET» *или* «POST» *или* «HEAD» + «HTTP/»»	ТОГДА: Д1: «Присвоить приоритет»

28

Сигнатурный анализ Sopcast



Сигнатурный анализ SSH

```
if (flow->l4.tcp.ssh_stage == 0) {
        if (packet->payload_packet_len > 7 && packet->payload_packet_len < 100
                && memcmp(packet->payload, "SSH-", 4) == 0) {
                NDPI_LOG(NDPI_PROTOCOL_SSH, ndpi_struct, NDPI_LOG_DEBUG, "ssh stage 0 passed\n");
                flow->l4.tcp.ssh_stage = 1 + packet->packet_direction;
                return;
        }
} else if (flow->l4.tcp.ssh_stage == (2 - packet->packet_direction)) {
        if (packet->payload_packet_len > 7 && packet->payload_packet_len < 100
                && memcmp(packet->payload, "SSH-", 4) == 0) {
                NDPI_LOG(NDPI_PROTOCOL_SSH, ndpi_struct, NDPI_LOG_DEBUG, "found ssh\n");
                ndpi_int_ssh_add_connection(ndpi_struct, flow);
                return;
```


30

Поведенческий анализ

31

Поведенческий анализ

распознает приложения, без заранее известных заголовков и структуры данных.

Например – BitTorrent.

Применяется анализ последовательности пакетов, принадлежащих к одному потоку.

32

Поведенческий анализ

- транспортные порты отправителя и получателя,
- размер пакета,
- частота открытия новых сессий в единицу времени…

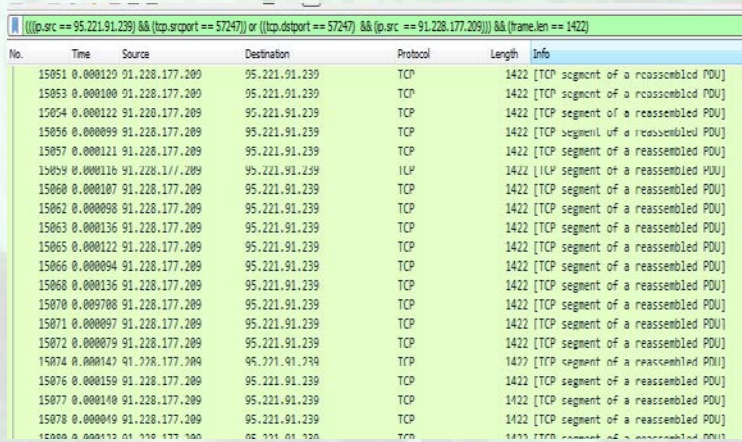
Существует множество поведенческих моделей соответствующих протоколов и приложений.

Точность определения различается.

Поведенческий анализ

Поведенческий анализ

Фрагмент видео



Поведенческий анализ

36

Поведенческий анализ

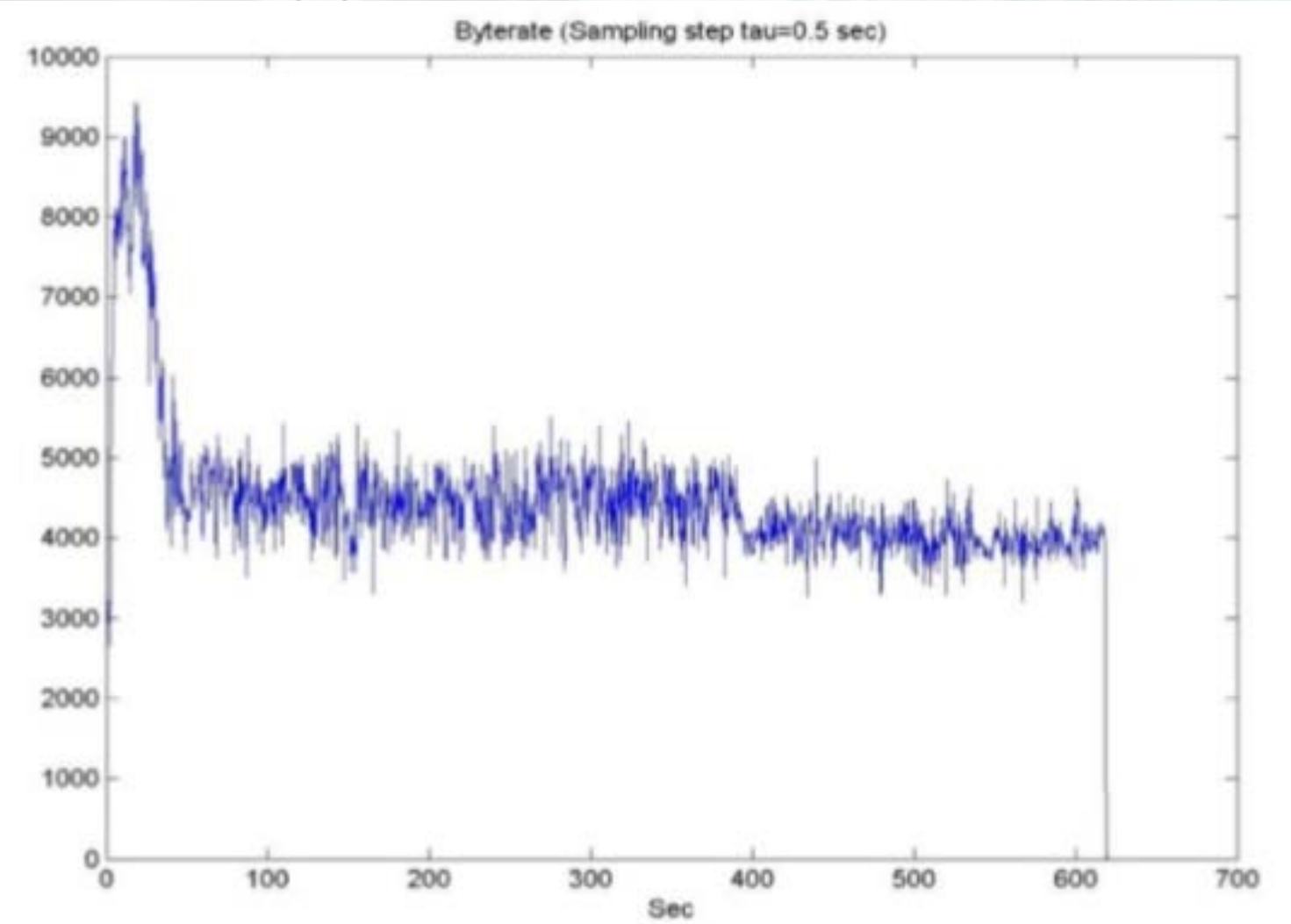


Рис. 1.15. Битовая скорость Skype трафика

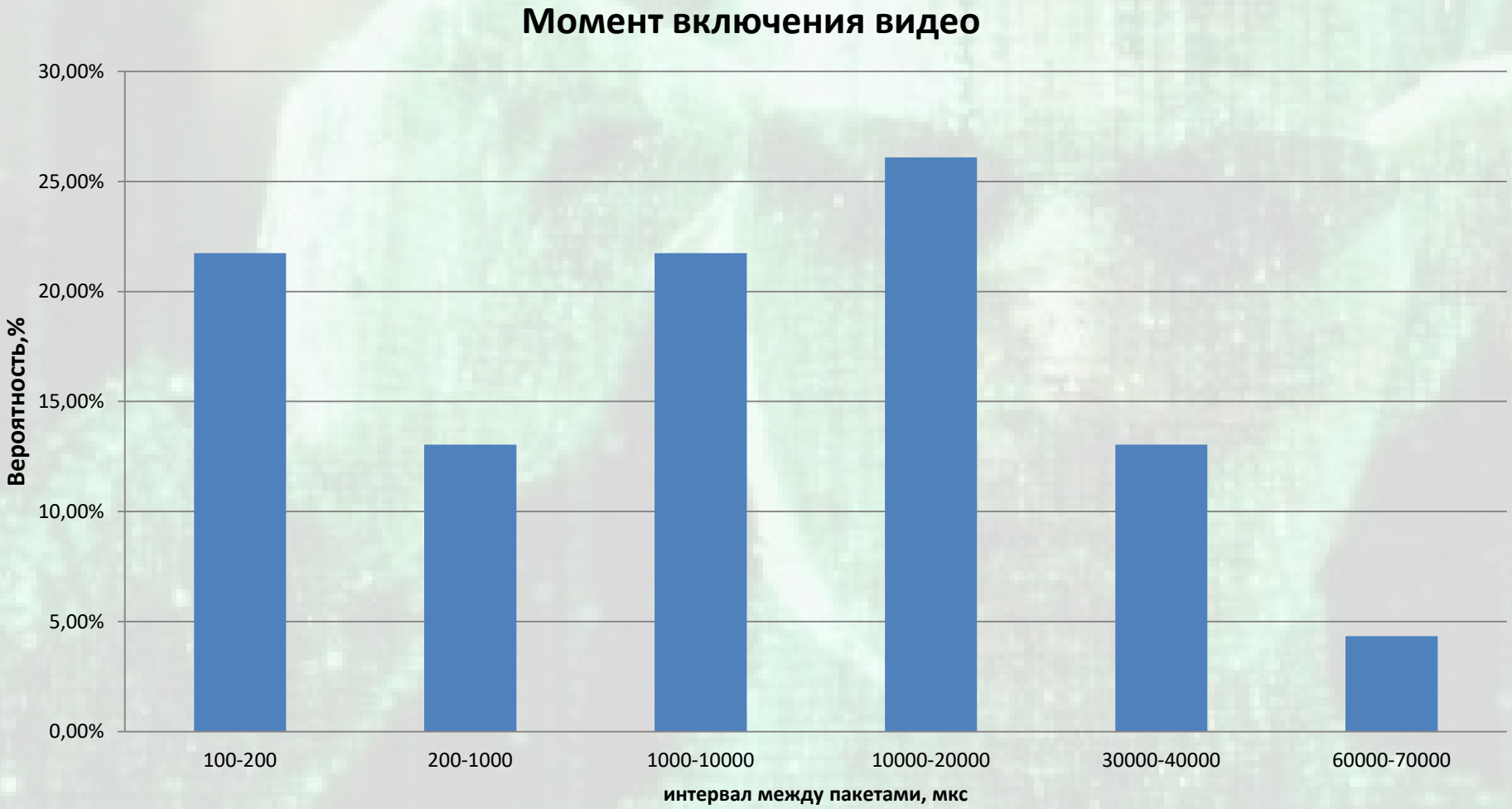
37

Статистический анализ

Статистический анализ

39

Статистический анализ



40

Статистический анализ

41

Статистический анализ (200 пакетов)

Статистический анализ (450 пакетов)

43

Эврестический анализ

44

Идентификация трафика в DPI

Эвристика

Эвристический алгоритм — это алгоритм решения задачи, правильность которого для всех возможных случаев не доказана, но про который известно, что он даёт достаточно хорошее решение в большинстве случаев.

Эвристический анализ - это технология обнаружения по признакам (без гарантированной точности). Используется, когда невозможно определить трафик с помощью сигнатурного анализа, то есть с помощью поиска и сравнения по базе сигнатур.

Объектам, обнаруженным с помощью эвристического анализа, присваивается вероятность соответствия, к примеру - 85%.

Совместное использование с другими методами анализа позволяет увеличить точность общей идентификации трафика.

DPI средства управления трафиком

Суммарное использование различных методов анализа позволяет значительно увеличить точность идентификации трафика

DPI
Анализ

Явно заданные правила

Сигнатуры

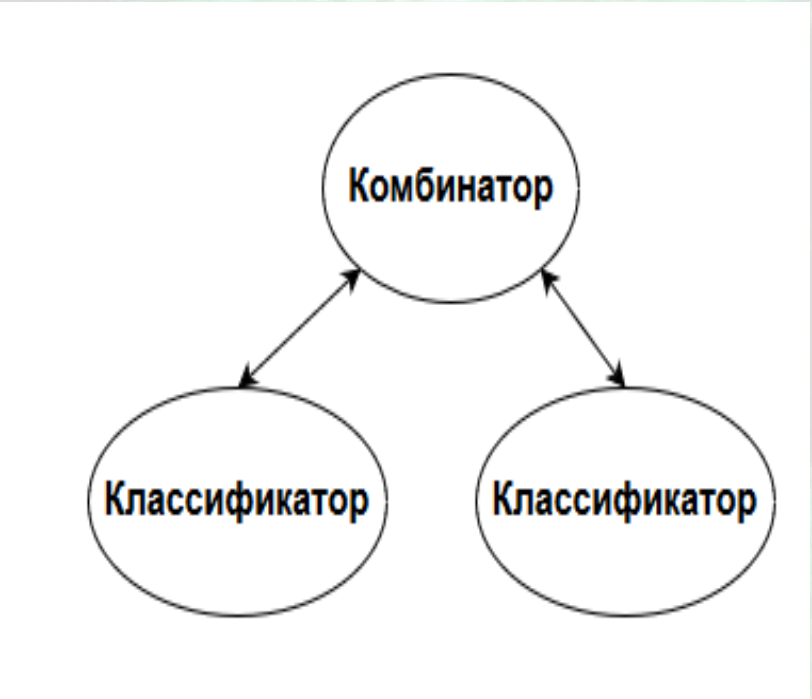
Эвристика

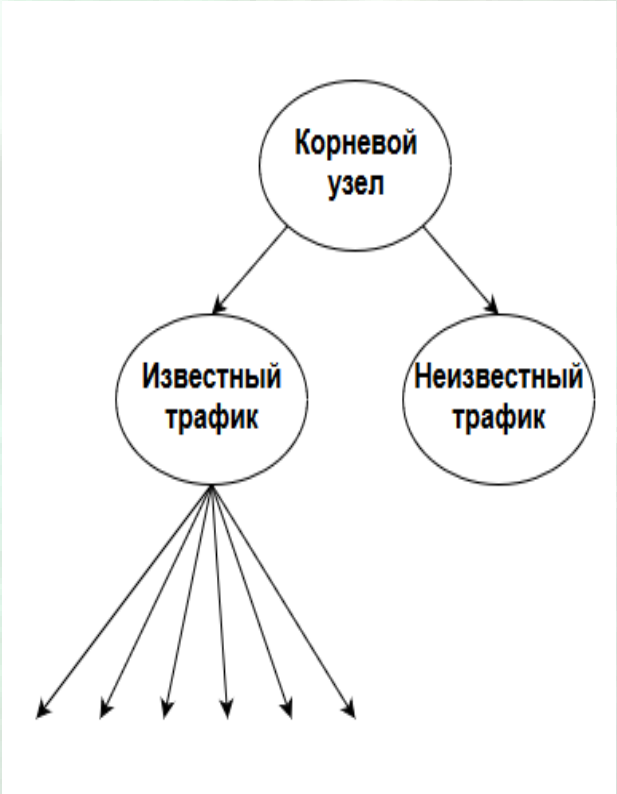
Анализ поведения

Классификация трафика

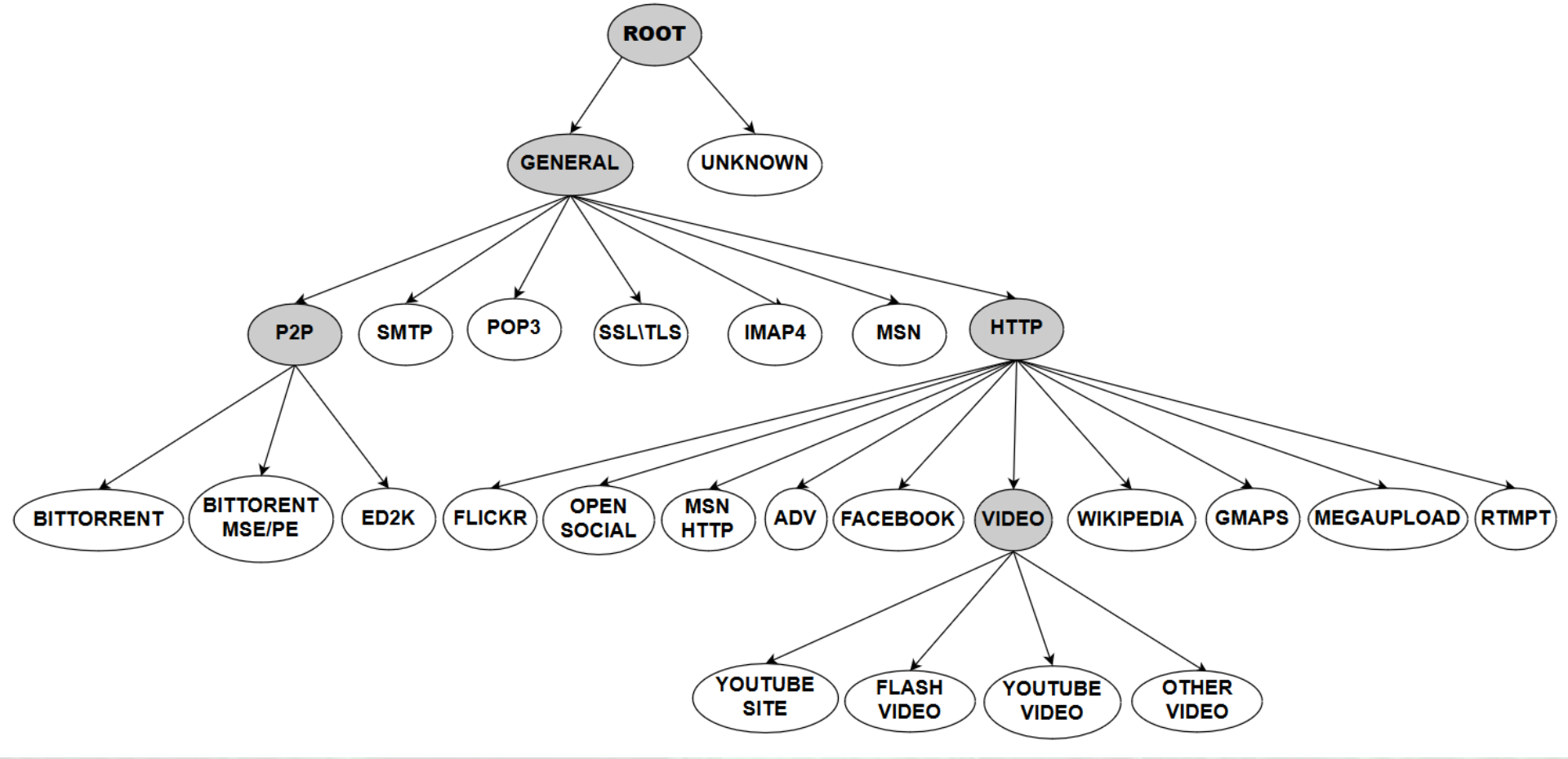
Комбинаторы и иерархический анализ

- Комбинаторы увеличивают точность классификации
- Иерархические системы улучшают эффективность и точность классификации





Комбинаторы и иерархический анализ



Комбинаторы

Существуют следующие виды алгоритмов комбинирования:

- **Majority Voting (Голосование по большинству),**
- **Weighted Majority Voting (Взвешенное голосование большинства),**
- **Naive Bayes,**
- **Dempster-Shafer combiner,**
- **Behavior-Knowledge Space (BKS) method,**
- **Wernecke's method,**
- **Oracle.**

Анализ применения комбинаторов

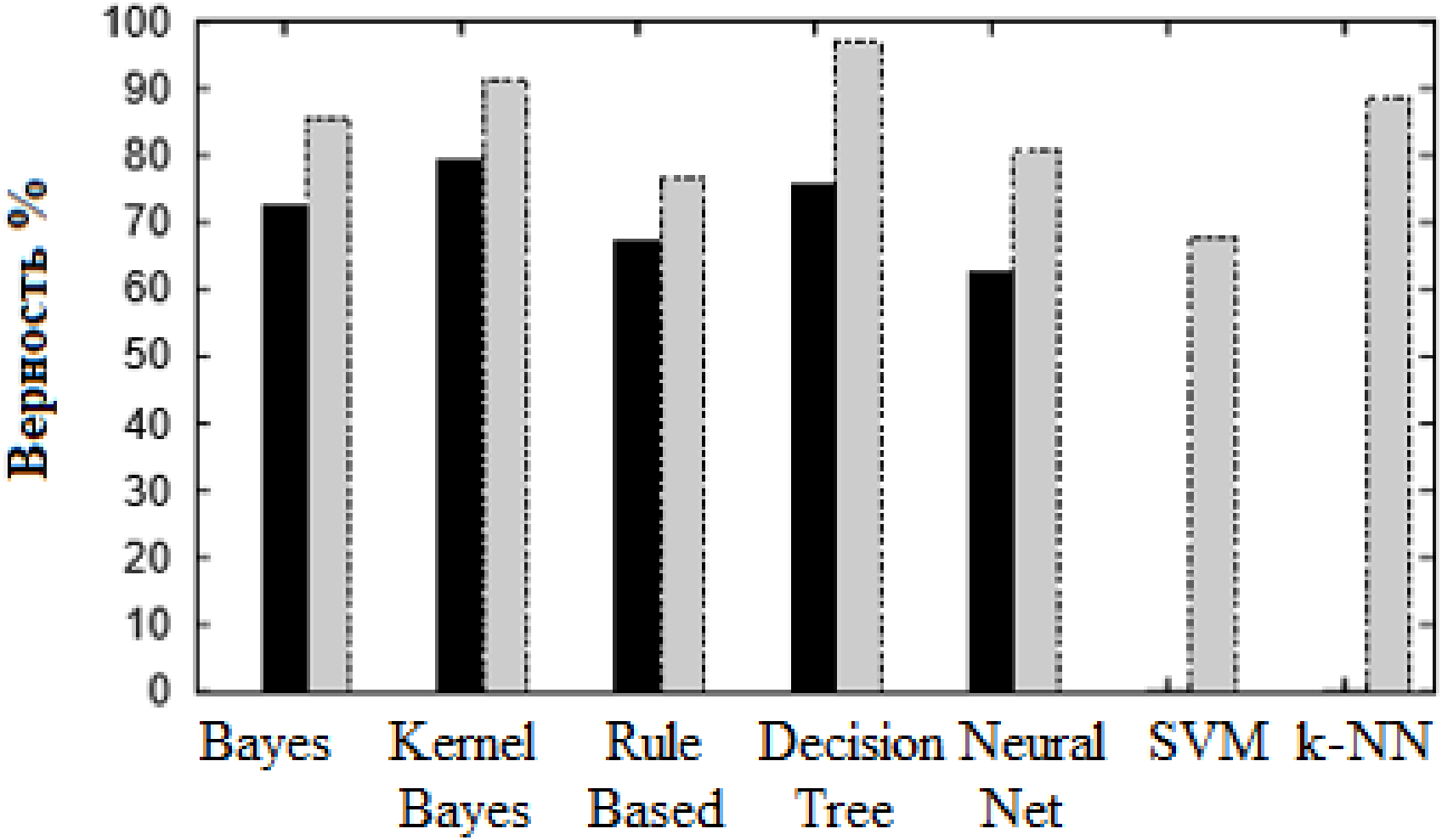
	Количество пакетов									
Классификатор	1	2	3	4	5	6	7	8	9	10
J48	62.1	94.6	95.9	96	96.8	97.1	97.2	97.2	97.2	97.2
K-NN	62.4	91.5	92.8	95	94.9	94.9	95.4	95.7	95.6	95.9
R-TR	72.7	93.4	93.6	94.9	95.3	96.8	96	96	96.1	96.2
RIP	69.5	93.7	94.7	96.2	96.1	96.5	96.7	96.9	96.9	96.9
MLP	43.5	71.7	81	82.3	82.3	82.3	82.3	82.3	82.3	82.3
NBAY	31.5	39.9	42.6	43.7	43.7	43.7	43.7	43.7	43.7	43.7
PL	76.2	83.7	83.7	83.7	83.7	83.7	83.7	83.7	83.7	83.7
PORT	15.6	15.6	15.6	15.6	15.6	15.6	15.6	15.6	15.6	15.6

	Количество пакетов									
Комбинатор	1	2	3	4	5	6	7	8	9	10
MV	57.8	93.9	94.4	95.6	95.9	96.2	96.3	96.3	96.4	96.4
D-S	83.1	96	96.9	97	97.4	97.4	96.4	96.5	96.5	96.5
BKS	97	98.4	98.3	98.4	98.4	98.4	98.4	98.4	98.4	98.4
WER	97	98.3	98.2	98.4	98.4	98.4	98.4	98.4	98.4	98.4

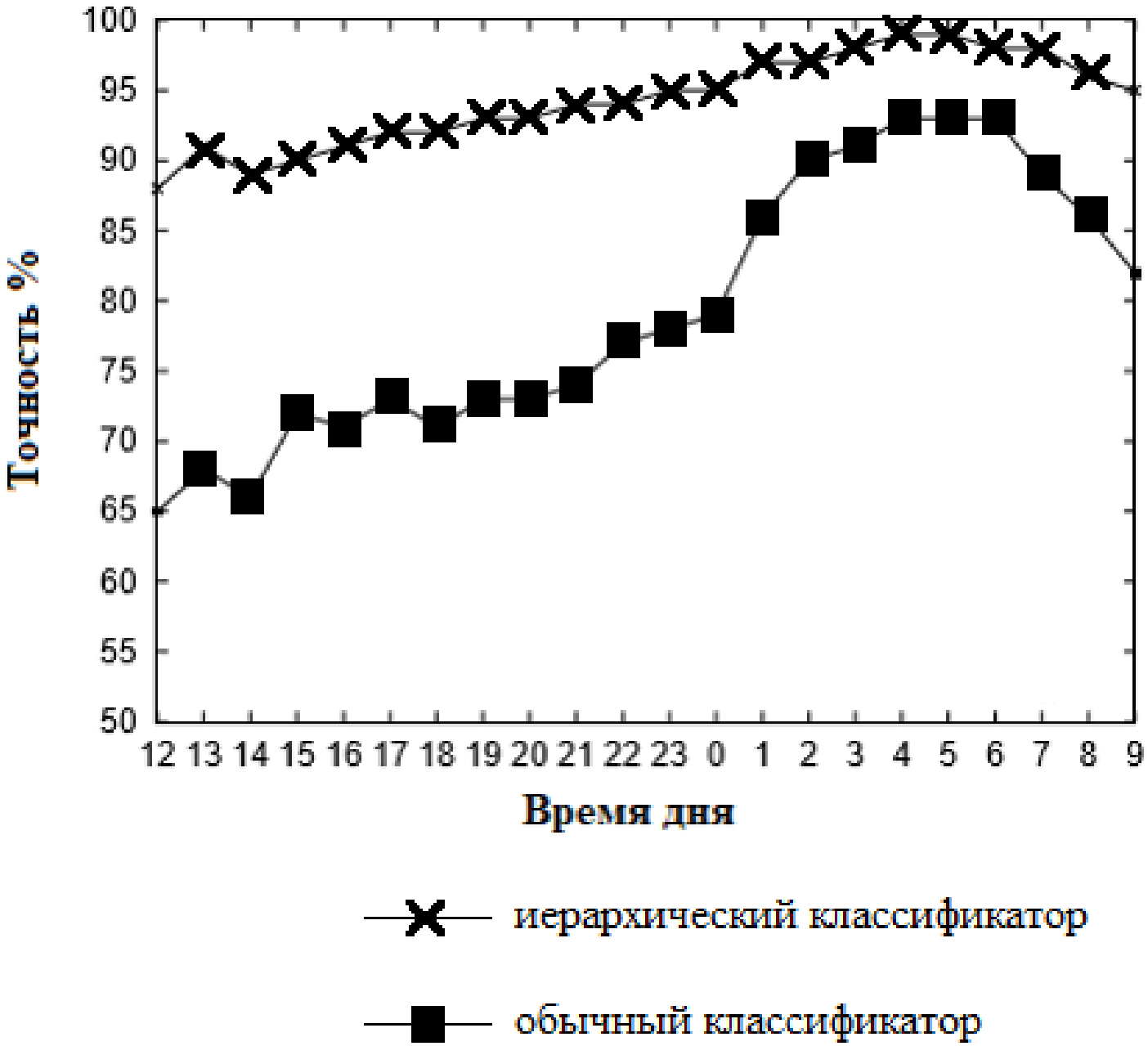
Анализ применения иерархического классификатора

черный – обычный классификатор

серый – иерархический классификатор



Анализ применения иерархического классификатора



DataMining анализ

53

DataMining
- это процесс выделения из анализируемых данных

неструктурированной информации и **метаданных**

и представления ее в виде, возможном для дальнейшего анализа и использования.

Основа **шаблоны - закономерности**, свойственные анализируемым данным, которые могут быть выражены в понятной человеку форме.

Цель поиска закономерностей - представление данных в виде, отображающем искомые процессы, а также построение моделей прогнозирования

54

DataMining
-использует аналитические методы,
большинство из которых было разработано в рамках теории **искусственного интеллекта**.

Непосредственное использование данных

Или выявление шаблонов-закономерностей

Непосредственное использование данных или сохранение данных.

При использовании данного метода исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях прогностического моделирования и/или анализа исключении.

Недостатком данного метода является возникновение трудностей при использовании больших баз данных.

Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии.

56

Непосредственное использование данных или сохранение данных.

Методы этой группы:

- **кластерный анализ,**
- **метод ближайшего соседа,**
- **метод k-ближайшего соседа,**
- **рассуждение по аналогии.**

57

Выявление и использование формализованных **закономерностей или дистилляция шаблонов**.

В рамках данной технологии на стадии свободного поиска из исходных данных **извлекается один образец** (шаблон)
и преобразуется в некие **формальные конструкции**, что является отличительной чертой данного метода.

Результаты стадии свободного поиска значительно компактнее самих баз данных, и они используются далее в прогностическом моделировании и анализе исключений .

Методы этой группы: логические методы; методы визуализации; методы кросс- табуляции; методы, основанные на уравнениях.

58

Выявление и использование формализованных **закономерностей или дистилляция шаблонов**.

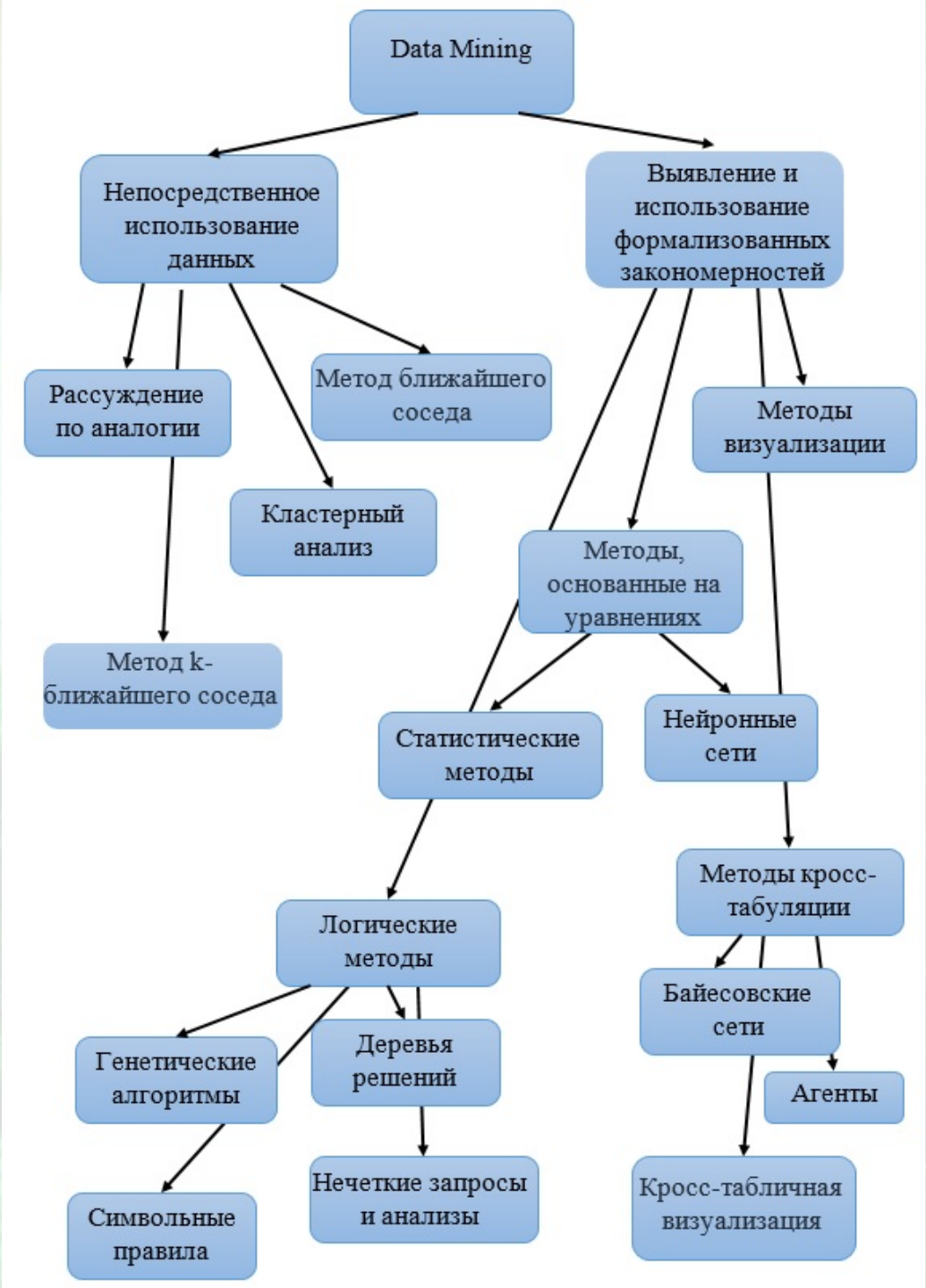
Методы этой группы:

- **логические методы;**
- **методы визуализации;**
- **методы кросс-табуляции;**
- **методы, основанные на уравнениях.**

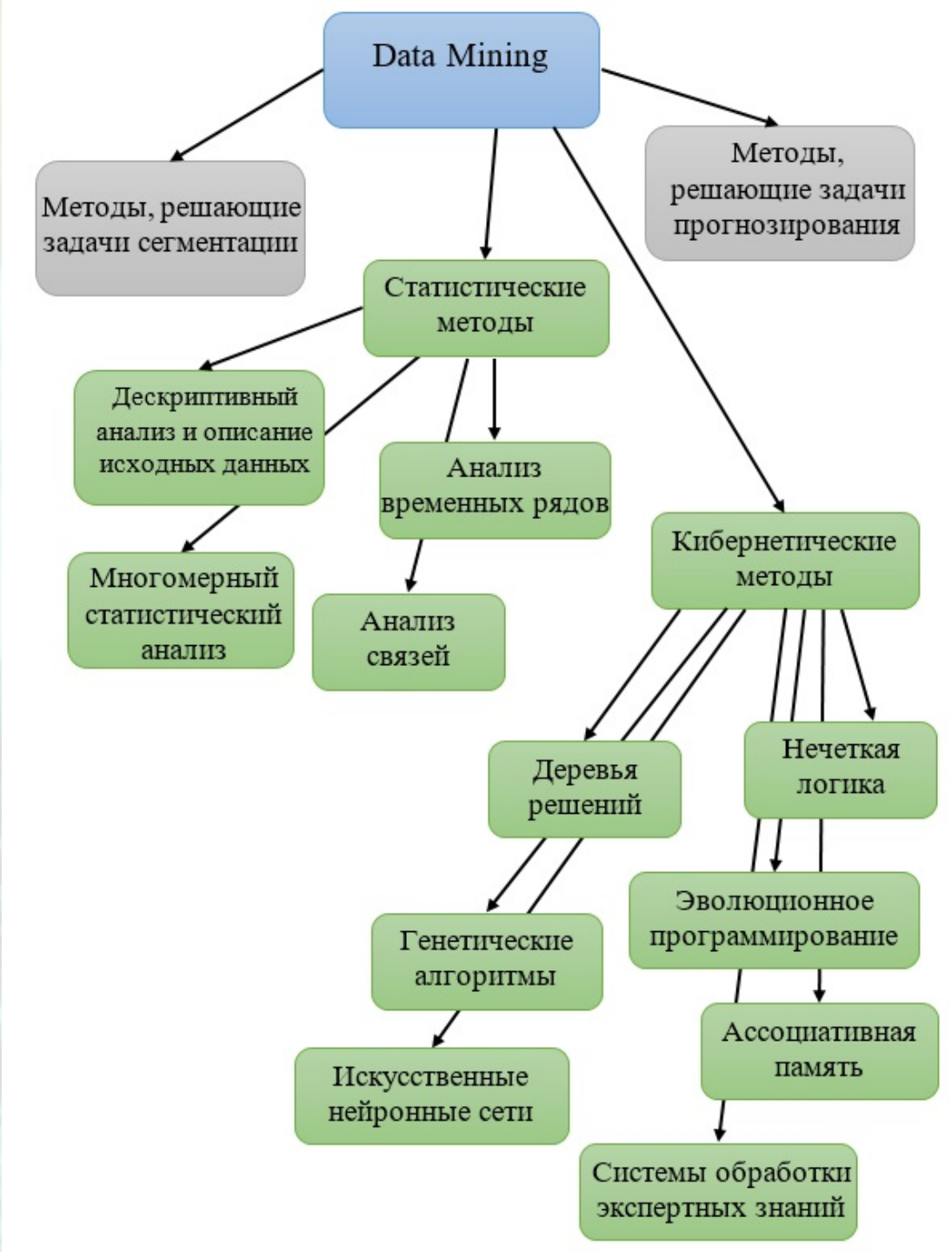
59

Методы DataMining

- Корреляционный, регрессионный и другие виды статистического анализа.
- Нечеткая логика
- Генетические алгоритмы
- Нейронные сети
- Кластеризация


Data Mining
Непосредственное использование данных
Выявление и использование формализованных закономерностей
Рассуждение по аналогии
Метод ближайшего соседа
Методы визуализации
Кластерный анализ
Методы, основанные на уравнениях
Метод k-ближайшего соседа
Статистические методы
Нейронные сети
Методы кросс-табуляции
Логические методы
Байесовские сети
Генетические алгоритмы
Деревья решений
Агенты
Нечеткие запросы и анализы
Кросс-табличная визуализация
Символьные правила

61


Data Mining
Методы, решающие задачи сегментации
Методы, решающие задачи прогнозирования
Статистические методы
Дескриптивный анализ и описание исходных данных
Анализ временных рядов
Многомерный статистический анализ
Анализ связей
Кибернетические методы
Деревья решений
Нечеткая логика
Генетические алгоритмы
Эволюционное программирование
Ассоциативная память
Искусственные нейронные сети
Системы обработки экспертных знаний

Логические методы

- **включают нечеткие запросы и анализы;**
- **символьные правила;**
- **деревья решений ;**
- **генетические алгоритмы.**

Методы данной группы являются наиболее интерпретируемыми:

они оформляют найденные закономерности, в большинстве случаев, в достаточно прозрачном виде с точки зрения пользователя.

Полученные правила могут включать непрерывные и дискретные переменные.

63

***Методы кросс-табуляции*:**

- **агенты,**
- **байесовские (доверительные) сети,**
- **кросс-табличная визуализация.**

64

При классификации с использованием **naive Bayes**

после проведенного машинного обучения на основе заранее известных потоков,

неизвестные потоки трафика соотносятся с выбранными категориями.

Например,

- www, email, bulk (большие объемы, ftp),
- serv (сервисы, протоколы X11, dns, ident, ldap, ntp),
- db (протоколы postgres, sqlnet oracle, ingres),
- p2p (KaZaA, BitTorrent, GnuTella),
- att (атаки, вирусы и черви),
- mmedia (мультимедиа, wmp, Real).

65

Методы на основе уравнений .

Методы данной группы выражают выявленные закономерности в виде уравнений .

Следовательно, они могут работать лишь с численными переменными, а другие типы переменных должны быть подвергнуты кодированию.

Это является ограничением применения методов данной группы, но несмотря на это, они широко используются при решении задач прогнозирования.

Основные методы данной группы:

- **статистические методы**
- **нейронные сети**

66

Статистические методы наиболее часто применяются для решения задач прогнозирования.

- **корреляционно-регрессионный анализ,**
- **корреляция рядов динамики,**
- **выявление тенденций динамических рядов,**
- **гармонический анализ.**

67

Также существует другой способ классификации методов DataMining

на статистические и кибернетические методы.

Этот способ разделения основан на различных подходах к обучению математических моделей.

Статистические методы основаны на использовании **усредненного накопленного опыта**, который отражен в ретроспективных данных.

Статистические методы DataMining включают в себя:
- **дескриптивный анализ и описание исх. данных,**
- **анализ связей,**
- **многомерный статистический анализ**
- **анализ временных рядов.**

68

Кибернетические методы включают в себя множество разнородных математических подходов.

Кибернетические методы можно разделить на:

- **искусственные нейронные сети,**
- **эволюционное программирование,**
- **генетические алгоритмы,**
- **ассоциативная память,**
- **нечеткая логика,**
- **деревья решений ,**
- **системы обработки экспертных знаний .**

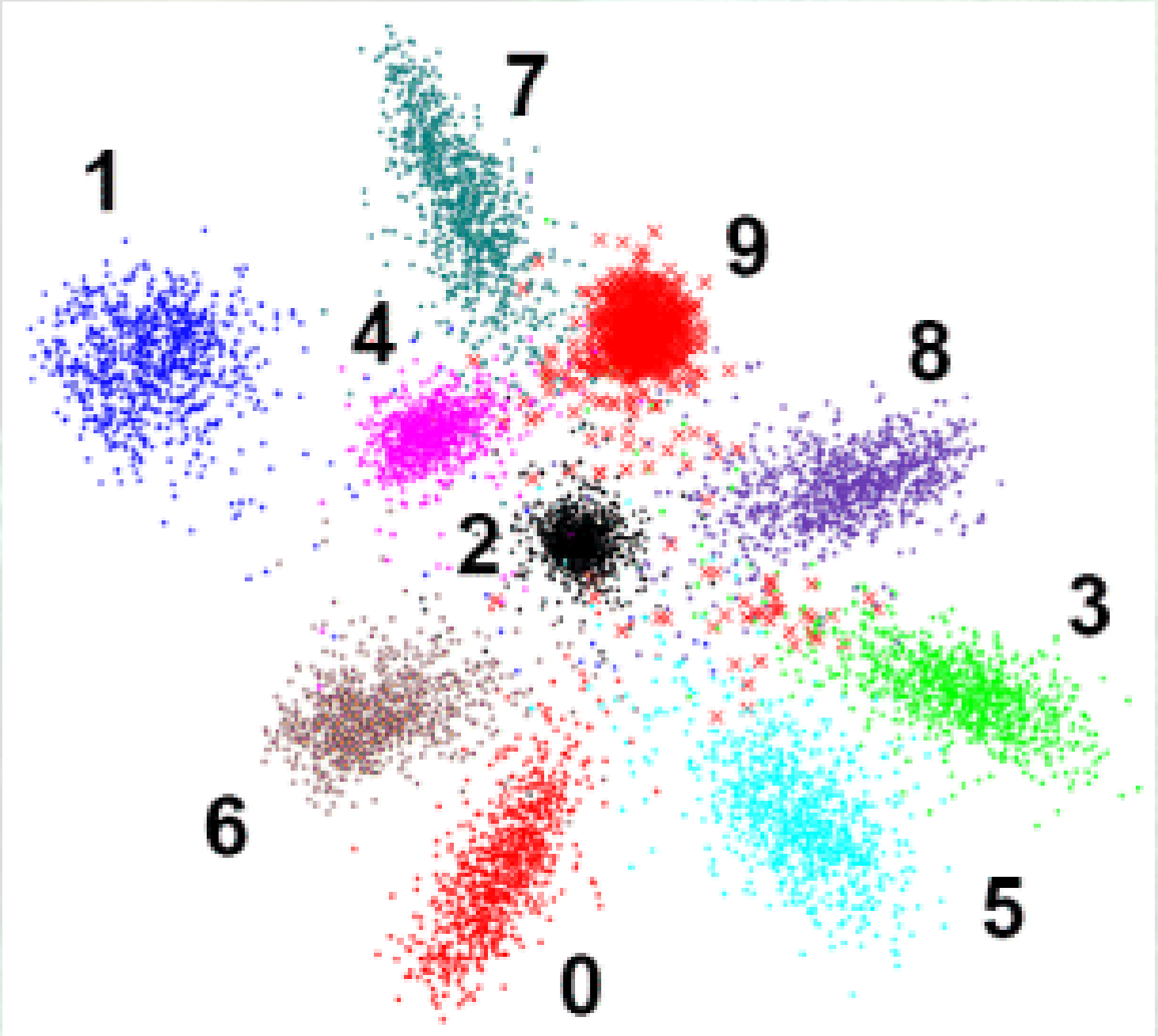
69

методы DataMining можно классифицировать по поставленным задачам: подразделение на решающие *задачи сегментации* (т.е. задачи классификации и кластеризации) К методам, направленным на получение **описательных результатов**, относятся:

- **алгоритм k-средних,**
- **k-медианы,**
- **иерархические методы кластерного анализа,**
- **самоорганизующиеся карты Кохонена,**
- **методы кросс-табличной визуализации,**
- **различные методы визуализации и другие.**

70

Кластеризация

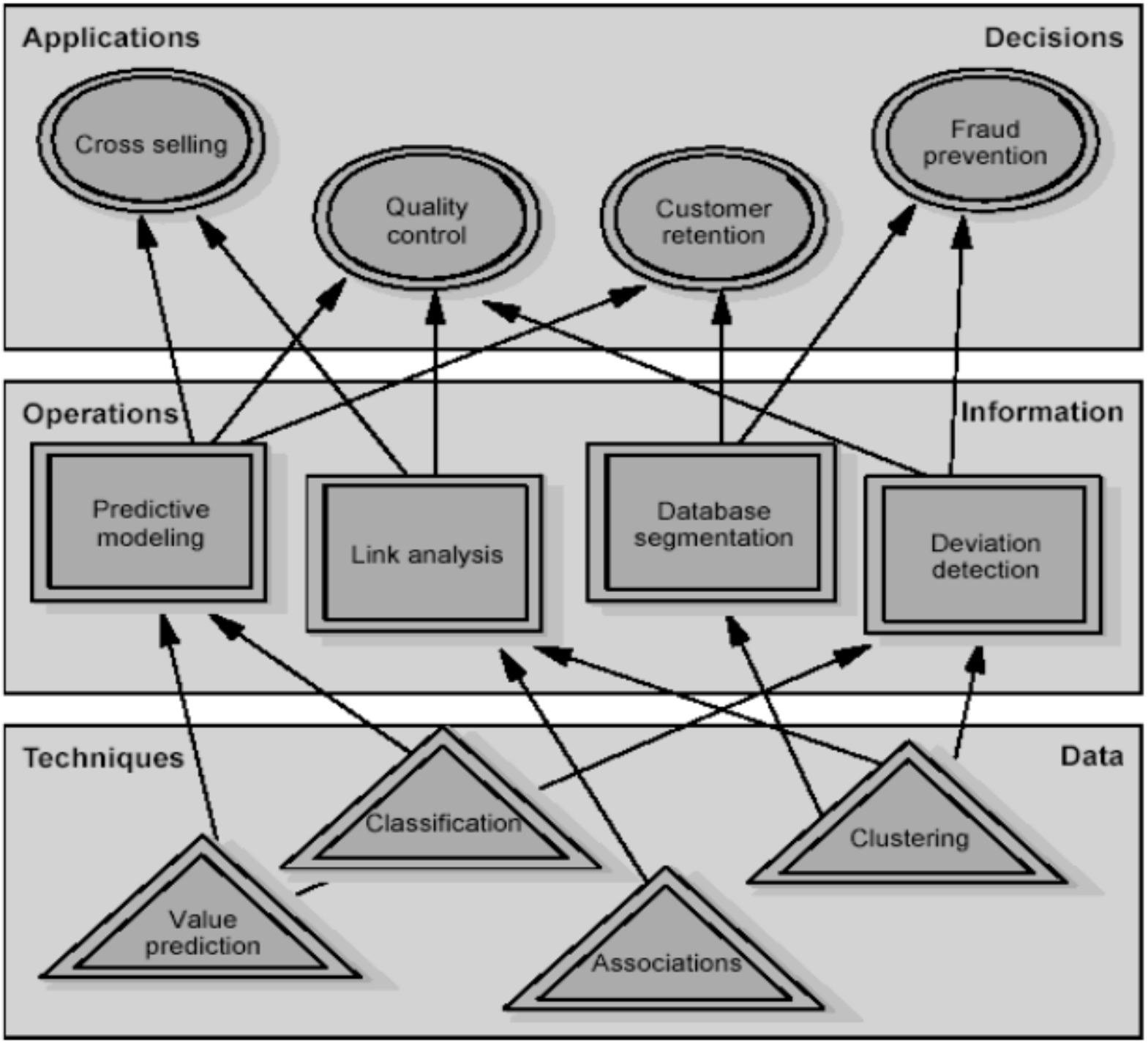


методы DataMining можно классифицировать по поставленным задачам: подразделение на решающие **задачи прогнозирования**.

Прогнозирующие методы используют значения одних переменных для прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.

К методам, направленным на получение прогнозирующих результатов, относятся такие методы:

- **нейронные сети,**
- **деревья решений ,**
- **линейная регрессия,**
- **метод ближайшего соседа,**
- **метод опорных векторов и другие.**


Applications
Decisions
Cross selling
Quality control
Customer retention
Fraud prevention
Operations
Information
Predictive modeling
Link analysis
Database segmentation
Deviation detection
Techniques
Data
Classification
Clustering
Value prediction
Associations

Спасибо за внимание.

Далее: Архитектура DPI.

Вопросы?

Ст. преп. каф. Инфокоммуникационных систем СПбГУТ,

Фицов Вадим Владленович,